The Inference of Phased Haplotypes for the Immunoglobulin H Chain V Region Gene Loci by Analysis of VDJ Gene Rearrangements

Marie J. Kidd, Zhiliang Chen, Yan Wang, Katherine J. Jackson, Lyndon Zhang, Scott D. Boyd, Andrew Z. Fire, Mark M. Tanaka, Bruno A. Gaëta and Andrew M. Collins

*J Immunol* 2012; 188:1333-1340; Prepublished online 28 December 2011; doi: 10.4049/jimmunol.1102097
http://www.jimmunol.org/content/188/3/1333

Supplementary Material

http://www.jimmunol.org/content/suppl/2011/12/29/jimmunol.1102097.DC1.html

References

This article cites 42 articles, 14 of which you can access for free at:
http://www.jimmunol.org/content/188/3/1333.full#ref-list-1

Subscriptions

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

Permissions

Submit copyright permission requests at:
http://www.aai.org/ji/copyright.html

Email Alerts

Receive free email-alerts when new articles cite this article. Sign up at:
http://jimmunol.org/cgi/alerts/etoc

A major gap in our knowledge of the human genome is the lack of phased genomic data describing the chromosomal associations of different nucleotide sequences (1). Phased (or “haplotype”) data bring additional power to the investigation of species evolution (2) and to the exploration of relationships between human populations (3). Such data are of particular importance where the chromosomal location of genes could be of physiological consequence, as is the case, for example, with the rearrangeable genes that encode the highly variable receptors of B cells and T cells of the immune system.

The mammalian immune system has the ability to respond to almost any Ag to which it is exposed because of the incredible diversity of lymphocyte receptor molecules. BCR diversity is made possible by multiple sets of highly similar genes that recombine to form functional VDJ gene rearrangements encoding the Ig H chain and VJ gene rearrangements encoding the Ig L chain (4). Ig gene recombination is an intrachromosomal event, and knowledge of gene associations would facilitate understanding of the generation of diversity. Yet we have very little phased data relating to the Ig gene loci. Although both the HapMap project (5) and the 1000 Genomes Project (6) have identified numerous polymorphisms in these loci, the association of so many highly similar sequences on each chromosome prevents the inference of haplotypes using conventional linkage disequilibrium-based phasing methods.

The H chain variable gene (IGHV) locus was first mapped in the 1990s (7), but the difficulties associated with the sequencing of so many similar genes and pseudogenes has meant that a complete sequence of the locus has only been reported once (8), and it was this sequence that was incorporated into the first two published versions of the human genome (9, 10). The sequence is not a true haplotype, for it was assembled from contigs derived from three lymphoblastoid cell lines (8). A complete description of the H chain diversity gene (IGHD) locus has also only been published once (11), and a handful of early reports described a number of distinct haplotypes encoding the six genes of the shorter joining gene (IGHJ) locus (12–14). More recently, it has been suggested that some of the apparent allelic variants that defined these IGHJ haplotypes may have been misidentified as a consequence of sequencing errors (15).

High-throughput sequencing has recently been applied to the study of human Ig (16–18) and TCR genes (19). Although such technology can be used for the direct investigation of germline genes (18), in this study we report how the analysis of amplified H chain VDJ gene rearrangements can be used to efficiently infer phased data covering all rearrangeable genes in the H chain V region.

The existence of many highly similar genes in the lymphocyte receptor gene loci makes them difficult to investigate, and the determination of phased “haplotypes” has been particularly problematic. However, V(D)J gene rearrangements provide an opportunity to infer the association of Ig genes along the chromosomes. The chromosomal distribution of H chain genes in an Ig genotype can be inferred through analysis of VDJ rearrangements in individuals who are heterozygous at points within the IGH locus. We analyzed VDJ rearrangements from 44 individuals for whom sufficient unique rearrangements were available to allow comprehensive genotyping. Nine individuals were identified who were heterozygous at the IGHJ6 locus and for whom sufficient suitable VDJ rearrangements were available to allow comprehensive haplotyping. Each of the 18 resulting IGHV genes (18), in this study we report how the analysis of amplified H chain VDJ gene rearrangements can be used to efficiently infer phased data covering all rearrangeable genes in the H chain V region. The chromosomal distribution of H chain genes in an Ig genotype can be inferred through analysis of VDJ rearrangements in individuals who are heterozygous at points within the IGH locus. We analyzed VDJ rearrangements from 44 individuals for whom sufficient unique rearrangements were available to allow comprehensive genotyping. Nine individuals were identified who were heterozygous at the IGHJ6 locus and for whom sufficient suitable VDJ rearrangements were available to allow comprehensive haplotyping. Each of the 18 resulting IGHV genes (18), in this study we report how the analysis of amplified H chain VDJ gene rearrangements can be used to efficiently infer phased data covering all rearrangeable genes in the H chain V region.


The existence of many highly similar genes in the lymphocyte receptor gene loci makes them difficult to investigate, and the determination of phased “haplotypes” has been particularly problematic. However, V(D)J gene rearrangements provide an opportunity to infer the association of Ig genes along the chromosomes. The chromosomal distribution of H chain genes in an Ig genotype can be inferred through analysis of VDJ rearrangements in individuals who are heterozygous at points within the IGH locus. We analyzed VDJ rearrangements from 44 individuals for whom sufficient unique rearrangements were available to allow comprehensive genotyping. Nine individuals were identified who were heterozygous at the IGHJ6 locus and for whom sufficient suitable VDJ rearrangements were available to allow comprehensive haplotyping. Each of the 18 resulting IGHV genes (18), in this study we report how the analysis of amplified H chain VDJ gene rearrangements can be used to efficiently infer phased data covering all rearrangeable genes in the H chain V region.
region gene loci. Despite the challenges involved in the identifi-
cation of the germline genes that contribute to some VDJ rear-
rangements, in large data sets each rearrangement provides in-
dependent evidence of the presence of particular IGHV, IGHD,
and IGJH genes on one chromosome. Together, thousands of
rearrangements can also provide evidence for the absence of
particular genes, for deletion polymorphisms have been reported
to occur in the IGHV (20) and IGHD (21) loci. Ig gene haplotypes,
including deletion polymorphisms, can therefore be inferred, and in
this study we describe the inferred diplotypes of nine individuals.
The IgH gene locus is revealed as a locus of considerable
variability. Not only are the 18 haplotypes described in this study all
unique, but the extent of the differences between the haplotypes
suggests the existence of many additional haplotypes. This vari-
ation is likely to result in substantial individual differences in
repertoires of available Abs.

Materials and Methods

Sequence generation and data set preparation

Ig gene sequence data sets were generated from samples of human pe-
ripheral blood that were obtained under a Stanford University Institutional
Review Board-approved protocol. Donors were of unknown ethnicity and
were recruited from the San Francisco Bay Area. PBMCs were isolated
by density gradient of diluted blood layered over Hypaque 1077 (Sigma-
Aldrich), and purified DNA template was prepared either by column pu-
rification (Qiagen, Valencia, CA) or by magnetic bead-based isolation
(Magnapure, Roche Diagnostics Corporation, Indianapolis, IN). PCR was
performed using BIOMED-2 primers to which were added 5’ sequencing
elements, and unique 6-, 7-, or 10-nucleotide sample “barcodes” were
added to the IGHJ primers as previously described (17). To facilitate pool-
ing DNA, PCR products were purified, and later determination of the sample source
of each sequence, production of amplicons for later sequencing using the 454
Titanium sequencer also used additional 10-nucleotide sample barcodes
that were added to the IGHV primers (17). Analysis of many of the re-
sulting amplicons has previously been described in two separate reports
(16, 17). For this study, longer amplicons were also produced using the
same IGHJ primer in conjunction with BIOMED2 IGHV framework 1
(FR1) primers as reported in Table I (22). PCR amplifications were per-
formed, then amplicons were pooled and purified as previously described
(17). High-throughput amplicon pyrosequencing data were then generated
using either 454 FLX chemistry or Titanium chemistry (Roche, Branford, CT).

Sequences were sorted into individual data sets based on the presence of
perfect matches to sample barcodes, IGHV primers, and to the first three
bases of the IGHJ common primer.

Determination of genotypes

The sets of germline IGHV, IGHD, and IGJH genes that collectively define
an individual’s genotype were determined by analysis of VDJ gene data
sets. The sets of sequences first aligned to a set of germline genes
using Vmatch, a general sequence alignment utility (23). The gene set
included both the UNSWig IGHV germline repertoire (24) (http://www.
immune.unsw.edu.au UNSWig.php) and the IMGT repertoire of functional
genes and pseudogenes (25). Levels of mutations in the IGHV genes were
reviewed, and where large numbers of alignments to a particular gene
were carried forward, further investigations were carried out to
identify putative unreported polymorphisms, as previously described (16).
For each 454 sequence, the five germline IGHV sequences or putative
IGHV polymorphisms that aligned with the least mismatches and longest
aligned fragment length were selected as candidate progenitor sequences for
the rearrangement. In some cases, it was not possible to distinguish between
two possible contributing alleles, because the critical nucleotides that
distinguish them are located upstream of the IGHV primers that had been
used for sequence amplification. In such cases, if a containing allele had
been highlighted previously as being of dubious provenance (24), it was
excluded from further consideration. Maximum likelihood genotyping was
then carried out to find the set of alleles that had the highest probability of
resulting in the observed sequence set. Where a single allele provided the
best alignment for all instances of that gene, it was immediately assigned
to the final genotype.

We defined $G$ as a genotype for a specific gene and $S (S = \{ s_1, s_2, \ldots, s_l \})$ as the set of sequences that were derived from this specific gene. For each
gene, the maximum likelihood genotype was defined as

$$\arg\max_G \Pr(S|G),$$

where $\Pr(S|G)$ is the probability of the sequence set $S$ given a genotype $G$.

Most genes are normally present as one or two alleles in an individual
genotype; however, in cases of gene duplication, there may appear to be
three or even four alleles of the same gene. For each IGHV gene, the potential
genus of most likely to have been incorporated into each rearrangement.
resulting amplicons has previously been described in two separate reports
as previously implemented in iHMMune-align (26).

Therefore, $Pr(s_i|g_k)G)$ was estimated as:

$$\sum_{g \in G} \Pr(s_i|g_k)Pr(g_k|G).$$

Assuming, initially, that all alleles of a given gene are equally likely to
appear in a genotype,

$$Pr(s_i|G) = \frac{1}{N} \sum_{g \in G} \Pr(s_i|g_k).$$

Therefore according to Eqs. 1, 2, and 3,

$$\arg\max_G \Pr(S|G) = \arg\max_G \prod_{s \in S} \sum_{g \in G} \Pr(s_i|g_k).$$

A preliminary analysis of the IGH sequences was then performed using
the partitioning utility iHMMune-align (26) to identify the germline IGHD
sequences. In such cases, if a contending allele had
to ambiguous sequences using a simplified estimation of $Pr(g_k|G)$,

$$Pr(s_i|g_k)G) = \sum_{g \in G} \Pr(s_i|g_k)Pr(g_k|G).$$

A preliminary analysis of the IGH sequences was then performed using
the partitioning utility iHMMune-align (26) to identify the germline IGHD
sequences. For each individual, the IGHD genotype was then defined by the
presence or absence of each gene in the iHMMune-
align partitioning results using IGHD gene-dependent rules. Three mem-
bers of the IGHD1 gene family (IGHD1-1, IGHD1-7, and IGHD1-20)
were not included in the genotypes because these genes are short and
highly similar and therefore cannot be identified with sufficient certainty.
Similarly, the IGHD7-27 gene was excluded because of its very short
length. The IGHD2-2 gene was included in the genotyping, but no attempt
was made to identify the three rarely distinguishable IGHD2-2 alleles. No
attempt was made to distinguish between IGHD5-5 and IGHD5-18, as
these genes share identical coding regions. Finally, the IGHD4-11,
IGHD1-14, and IGHD6-25 genes were excluded as they are apparently
not rearrangeable (27). Rules based on specific sequence motifs were used
to discriminate between alleles for the IGHD2-8, IGHD2-21, IGHD3-3,
IGHD3-10, and IGHD3-16 genes. For other IGHD genes, inclusion in the
final genotype was based on the presence or absence of satisfactory
alignments in the iHMMune-align output.

Genotyping of IGHJ genes was performed using a strategy similar to that
of IGHV genotyping. The 3’ end of each VDJ sequence was aligned against
the IGHJ repertoire using Vmatch (23). After determination of the draft
IGHJ genotype, the maximum likelihood genotyping method was applied
to ambiguous sequences using a simplified estimation of $Pr(s_i|g_k)$ based on
the number of mutations observed in their IGHV region alignment.

Inference of haplotypes

Genotypes were examined, and individuals who were heterozygous at the
IGHJ6 loci were identified. Data sets for each donor were checked to ensure
that they contained the minimum number of sequences required to give
a 95% probability of being able to detect the presence of an IGHV or IGHD
genome within a haplotype with an average rearrangement frequency of 0.5%.
This was determined as follows:

$$N = \frac{\ln(1-0.95)}{\ln(1-f)}$$

where $N$ is the required number of sequences, $0$ is the desired probability
of observing the gene, and $f = 0.005 \times \left(\frac{\text{the proportion of the total data set}}{\text{utilizing the least frequently observed IGJH6 allele}}\right)$.

For suitable individuals, $N$ varied between 4,289 and 10,272. VDJ sequences from these individuals were repartitioned with iHMMune-align
using customized germline repertoires based on the genotype of each
individual. The two IGHJ haplotypes of each individual were then determined
by examining the number of instances in which an IGHJ or IGHD gene
was associated with each of the $G$ alleles of the heterozygous IGHJ6 locus. Genes
and allelic variants were only included in this analysis if at least 10 VDJ rearrangements were identified that appeared to include that gene or allele. On the assumption that a proportion of the apparent IGHV and IGHD gene sequences identified within VDJ rearrangements would be incorrectly associated with an IGHJ6 allele, two hypotheses were considered: that each germline gene or allelic variant was either found on one or on both chromosomes. Where x sequences were assigned to chromosome A, y sequences were assigned to chromosome B, and \( x \approx y \), the probabilities associated with each hypothesis were calculated as follows:

\[
Pr\{x, y|\text{Only A}\} = \left(\frac{n}{x}\right) e^x (1 - e)^{n-x}
\]

(6)

\[
Pr\{x, y|\text{A and B}\} = \left(\frac{n}{x}\right) p^x (1 - p)^{n-x},
\]

(7)

where \( n = x + y \) is the total number of alignments to the gene, \( e \) is the error rate of assignments to chromosome A, and \( p \) is the overall proportion of sequences in the data set involving rearrangement of chromosome A.

The error rate was set by reference to published evaluations of the performance of the iHMmune-align utility (28). Where a single allele was present in the genotype, \( e \) was set as 0.02, and where multiple alleles were present, \( e \) was set as 0.1. The likelihood ratio (LR) was then calculated to determine the most likely explanation of the results, as follows:

\[
LR_1 = \frac{Pr(\text{More likely hypothesis})}{Pr(\text{Less likely hypothesis})}
\]

(8)

where LR1 was >20, the more likely hypothesis was confirmed. Because the presence of the gene within the individual’s genotype had been independently determined, where LR1 was <20, it was accepted that the gene would likely be present on at least one chromosome. The likelihood ratios of the probabilities that the gene was either present on both chromosomes or only on chromosome B was then calculated as follows:

\[
LR_2 = \frac{Pr\{x, y|\text{A and B}\}}{Pr\{x, y|\text{Only B}\}}
\]

(9)

where the value of this ratio was >20, the presence of the gene on chromosome A was confirmed, while the presence or absence of the gene on chromosome B remained uncertain.

Unrearranged IGH locus genomic DNA sequencing

Genomic DNA was prepared from PBMC samples using column purification kits (Qiagen) and was treated with RNase A prior to genomic library construction. DNA was fragmented using fragmentase enzyme (New England Biolabs) and size selected with agarose gel electrophoresis and gel purification column kits (Qiagen). Fragmented DNA was end-repaired to form blunt-ended fragments using Klenow DNA Polymerase I, T4 DNA Polymerase, and T4 Polynucleotide Kinase (New England Biolabs), then linked with Illumina linkers for single-end nondirectional sequencing. Agarose gels were used for size selection of linkeder fragments of 150- to 400-bp size, and fragments were purified using a gel purification column kit (Qiagen). Twenty-five cycles of PCR were carried out on the linkeder libraries with linker-specific primers and Phusion polymerase (New England Biolabs), and the final library was gel purified. Sequencing of each genomic fragment library with single-end 36 base reads was carried out in a single lane of an Illumina Genome Analyzer Ix sequencer.

Genomic DNA sequence analysis

Single-end 36 base genomic DNA reads were aligned to the HG19 version of the reference human genome using the Needleman–Wunsch algorithm implemented in the Novoalign alignment software (Novocraft). Reads aligning to multiple sites in the genome were excluded from further analysis. Reads aligning to unique positions in the genome were evaluated further, after collapsing identical reads to minimize bias from PCR amplification of the genomic libraries. Collapsed reads aligning to the predicted region of D segment deletion (spanning from IGHD3-3 to IGHD2-8) were counted and were compared with the total count of reads aligning to the rest of the genome. Statistical testing for the significance of the proportion of reads in the region of putative deletion compared with the number of reads over the rest of the genome, comparing the sample SBBKn14 (predicted to contain the deletion) with a control sample, was calculated using \( \chi^2 \) testing and Fisher’s exact test.

Results

Sets of VDJ rearrangements were considered from 44 individuals for whom there were sufficient unique sequences to allow comprehensive IGHV, IGHD, and IGHJ genotyping. The sequence data sets for these individuals averaged 4240 VDJ rearrangements. Sequences containing no mismatches to germline genes composed 47.5% of the data sets, and most of these sequences were therefore likely derived from IgM/IgD naive B cells. This general lack of mutation aided genotype determination. The defined genotypes included between 39 and 55 distinct, functional germline IGHV sequences. A number of individuals were seen who appear to carry homozygous deletion polymorphisms involving IGHV genes, including IGHV4-30-2, IGHV3-30-3, IGHV4-30-4, IGHV4-39, IGHV4-b, and IGHV5-a. Deletions of the contiguous genes IGHV1-8 and IGHV3-9 were also inferred in several individuals. As many as seven rearrangeable pseudogenes were seen in a single genotype, and rearrangeable pseudogenes included humIGHV177, humIGHV181, IGHV1-14, IGHV1-17, IGHV3-19, IGHV3-22, IGHSV3-41, IGHSV3-47, IGHSV3-52, and IGHSV4-55. IGHJ genotypes included between 15 and 23 distinct sequences, and apparent homozygous IGHD deletion polymorphisms were also seen in the IGHD locus. The partitioning of ~5% of sequences failed to identify a candidate IGHD gene that met the stringent requirements of iHMmune-align for IGHD gene identification (26).

Little diversity was seen among the IGHJ genotypes, although 15 individuals were heterozygous at the IGHD6 locus. Although sequence data sets from every such individual can be used to define at least partial haplotypes, the inference of phased haplotypes was restricted to data sets that included at least 4200 unique rearrangements. This ensured that resulting haplotypes included many loci and that these loci could all be analyzed using multiple sequence reads. Nine individuals had sufficient IGHJ6-utilizing sequences for such comprehensive haplotyping. In these individuals, IGHJ6-utilizing sequences accounted for between 29.7 and 50.1% of all sequences.

Rearranged VDJ sequences were analyzed to determine the most likely haplotypes. A number of likely IGHD gene deletion polymorphisms were highlighted. Because of the difficulty of aligning rearranged sequences against the germline IGHD repertoire, the data sets that included these likely deletions, as well as any other rare IGHV/IGHD pairings, were carefully reviewed. In one individual (AL13), a handful of sequences suggested the presence of the IGHD4-4 gene in a haplotype. In other respects, this region of the haplotype matched two other haplotypes that shared a deletion of the IGHD3-3, IGHD4-4, IGHD6-6, IGHD1-7, and IGHD2-8 genes. (The presence or absence of IGHD5-5 could not be determined because its sequence is identical to the IGHD5-18 gene.) Review of the IGHD4-4 alignments suggests that they were all in error, for they were all short and mutated. Apparent deletion polymorphisms involving single genes (IGHD2-8, IGHD3-9, and IGHD5-24) were also reviewed and confirmed.

The absence of D segments in sets of rearranged IGH VDJ sequences could, in theory, be the result of sequence variants preventing rearrangement of these segments, or could be due to deletion removing these regions of the genome entirely. To evalu-

Table I. IGHV FR1-region family-specific forward primers used in PCR amplifications of rearranged Ig gene sequences

<table>
<thead>
<tr>
<th>IGHV Family</th>
<th>Sequences (5’ to 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGHV1</td>
<td>GGCCTCAGCTGAAAGTCTCTCTGCAAG</td>
</tr>
<tr>
<td>IGHV2</td>
<td>GTCCTGTCCTCTACTGCTTGAGAACC</td>
</tr>
<tr>
<td>IGHV3</td>
<td>CTGGGGGCTCCCCCTGAGCTTCTCTG</td>
</tr>
<tr>
<td>IGHV4</td>
<td>CTTCTGAAGACTCTGTGCTTCCCTTCT</td>
</tr>
<tr>
<td>IGHV5</td>
<td>CGGGGAGCTTGAAGACTCTCTTCTG</td>
</tr>
<tr>
<td>IGHV6</td>
<td>TCCAGACCTCTTCATCTCCTGCTG</td>
</tr>
</tbody>
</table>
ate further for the presence of a homozygous germline genomic deletion of the region of the IGH locus containing D segments IGH3-3 to IGH2-8, whole genome shotgun sequencing was performed on DNA from an individual predicted to have the D3-3 to D2-8 deletion and on a control sample. Supplemental Fig. 1 shows that in subject SBKN14, zero reads aligned to the region of predicted deletion, whereas multiple reads aligned to this region in the control sample.

\[ \chi^2 \text{ testing and Fisher's exact test for the significance of the proportion of reads in the region of predicted deletion compared with the rest of the genome gave } p \text{ values of } 2 \times 10^{-6} (\chi^2) \text{ and } 7 \times 10^{-7} \text{ (Fisher's exact test).} \]

The diversity of the 18 IGHD haplotypes is represented in Fig. 1. Without consideration of loci where the allelic variant of the gene could not be determined with certainty, seven different inferred haplotypes were seen, including one haplotype that was identified nine times. IGHD haplotype analysis for one individual is presented in Table II.

### Table II. Sequence counts of IGHD genes and allelic variants associated with IGHJ6 alleles in VDJ rearrangements, and LRs for the inference of the presence of IGHD genes within the diplotypes of a representative individual

<table>
<thead>
<tr>
<th>Sequence Counts</th>
<th>Presence of IGHD Genes on Chromosomes as Determined by LRs</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGHD6*02</td>
<td>IGHD6*03</td>
</tr>
<tr>
<td>IGHD2-2*02</td>
<td>97</td>
</tr>
<tr>
<td>IGHD3-3*01</td>
<td>124</td>
</tr>
<tr>
<td>IGHD4-4<em>11</em>01</td>
<td>8</td>
</tr>
<tr>
<td>IGHD5-5<em>18</em>01</td>
<td>9</td>
</tr>
<tr>
<td>IGHD6-6*01</td>
<td>51</td>
</tr>
<tr>
<td>IGHD2-8*01</td>
<td>0</td>
</tr>
<tr>
<td>IGHD2-8*02</td>
<td>9</td>
</tr>
<tr>
<td>IGHD3-9*01</td>
<td>26</td>
</tr>
<tr>
<td>IGHD3-10*01</td>
<td>3</td>
</tr>
<tr>
<td>IGHD3-10*p03</td>
<td>82</td>
</tr>
<tr>
<td>IGHD5-12*01</td>
<td>30</td>
</tr>
<tr>
<td>IGHD6-13*01</td>
<td>92</td>
</tr>
<tr>
<td>IGHD2-15*01</td>
<td>84</td>
</tr>
<tr>
<td>IGHD3-16*02</td>
<td>0</td>
</tr>
<tr>
<td>IGHD3-16*p03</td>
<td>28</td>
</tr>
<tr>
<td>IGHD4-17*01</td>
<td>48</td>
</tr>
<tr>
<td>IGHD6-19*01</td>
<td>38</td>
</tr>
<tr>
<td>IGHD2-21*01</td>
<td>30</td>
</tr>
<tr>
<td>IGHD3-22*01</td>
<td>65</td>
</tr>
<tr>
<td>IGHD5-24*01</td>
<td>6</td>
</tr>
</tbody>
</table>

*a Chromosome 1 is the IGHD6*02-defined chromosome.

*b Data are shown for all readily identifiable and rearrangeable IGHD genes. No attempt was made to distinguish between the highly similar IGHD2-2 alleles.

*c No attempt was made to assign a gene or allelic variant to a chromosome where fewer than 10 VDJ rearrangements appeared to include that sequence.

\[ x \text{, Absent; } ✓, \text{ present.} \]
sent as Table II, and analysis for all individuals is available as Supplemental Table I.

Analysis of IgH rearrangements as described earlier was then used to define the presence or absence of IGHV genes in each haplotype. A new putative variant of IGHV1-18 was identified. We have named this putative allele IGHV1-18*p05, and it is presented as Supplemental Table II. Subsequent investigation of dbSNP showed that the polymorphism has been reported by the 1000 Genomes Project (rs72695948). The diploidy of one individual is presented for the 3′ region of the genome, from IGHV6-1 to IGHV1-18, as Table III. One of the haplotypes shown appears to lack the contiguous IGHV8-8 and IGHV3-9 genes. The diversity of the 18 unique IGHV haplotypes is illustrated in the partial haplotypes of Fig. 2. Statistical analysis could not determine the presence or absence of some genes with certainty for all haplotypes, and these uncertainties are also indicated in Fig. 2. Additional associations across the IGHV locus are also available for all individuals as Supplemental Table III. These haplotypes include between 35 and 46 functional genes. In total, 54 IGHV loci were included in the inferred haplotypes, including seven loci where the data were indicative of duplications of previously reported genes. Such apparent duplications were seen for IGHV1-2, IGHV3-11, IGHV3-30, IGHV1-46, IGHV4-59, IGHV3-64, and IGHV1-69.

Apparent deletion polymorphisms were inferred at the following IGHV loci: IGHV1-8, IGHV3-9, IGHV3-30, IGHV4-30-2, IGHV3-30-3, IGHV4-30-4, IGHV4-31, IGHV3-33, and IGHV4-39. In most cases, deletion of IGHV1-8 was associated with deletion of IGHV3-9, and this polymorphism was seen in 7 of the 18 haplotypes. Twelve different haplotypes included deletions of between one and four of the six contiguous functional genes from IGHV3-30 to IGHV3-33. Without consideration of uncertainties, eight distinct patterns of deletions were seen. Unmapped genes are generally rearranged at frequencies that preclude haplotyping, but IGHV4-b and IGHV5-a rearrange at higher frequencies and were also seen to be absent from many inferred haplotypes.

In many cases, although likelihood ratios suggested the absence of a gene from a chromosome, small numbers of sequences were seen that appeared to associate the “deleted” gene with that chromosome. Review of these data sets showed that these aberrant sequences were typically highly mutated. It is likely that these mutations led to a misidentification of the IGHV gene or allele in the rearranged sequences. For example, the data set generated from one individual included 120 sequences that used the IGHV3-30 gene and could be associated with the chromosome bearing IGHJ6*02. A further 13 sequences appeared to link the IGHV3-30 gene with the alternate chromosome. These sequences were all highly mutated. It is likely that these rearrangements do not involve the IGHV3-30 gene, but rather involve the highly similar IGHV3-33 gene. Mutations made some of the many IGHV3-33 sequences align more closely to the germline IGHV3-30 gene.

Inferred gene duplications were also supported by mutation analysis. If an apparent duplication was inferred as a consequence of misidentification of IGHV alleles, we would expect to see that a high proportion of these sequences were mutated. This was not the case.

Rearrangement frequencies were analyzed, and IGHD gene rearrangement frequencies were particularly variable. For example, IGHD5-12 was present in 0.4–5.0% and IGHD2-2 in 4.5–28.0% of all rearrangements of single chromosomes. IGHV gene rearrangement frequencies were less variable, and these genes were usually present in between 0.5 and 2% of all rearrangements. In contrast, a single allele of the IGHV1-69 gene was present in 18.0% of rearrangements of one chromosome, and two IGHV1-69 alleles were responsible for 20.6% of all rearrangements of another chromosome that carried an evident IGHV1-69 gene duplication. Much of the variability that was seen appears to result from differing rearrangement frequencies of the different allelic variants of the IGHV genes. This was most clearly seen in the case of the IGHV7-4-1 and IGHV1-3 genes. Although the IGHV7-4-1*01 allele was detected in the genotype of six individuals, it was only present at a high enough frequency to allow haplotyping in one individual, where this allele was present in 1.1% of all rearrangements of a single chromosome. The alternative IGHV7-4-1*02 allele was detected in five haplotypes, and it was seen in between 2.3 and 7.5% (mean: 3.9%) of rearrangements of those chromosomes. Similarly, the IGHV1-3*01 allele was present in 14 haplotypes and was seen in between 2.5 and 5.6% (mean: 3.7%) of

<table>
<thead>
<tr>
<th>IGHV6-1*01</th>
<th>IGHV1-2*02</th>
<th>IGHV1-2*04</th>
<th>IGHV1-3*01</th>
<th>IGHV4-4*02</th>
<th>IGHV4-4*02</th>
<th>IGHV2-5*01</th>
<th>IGHV2-5*01</th>
<th>IGHV3-7*01</th>
<th>IGHV1-8*01</th>
<th>IGHV3-9*01</th>
<th>IGHV3-11*01</th>
<th>IGHV3-15*01/02</th>
<th>IGHV3-15*07</th>
<th>IGHV1-18*01/p05</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>15</td>
<td>20</td>
<td>2</td>
<td>42</td>
<td>42</td>
<td>10</td>
<td>2</td>
<td>18</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td>26</td>
<td>29</td>
</tr>
</tbody>
</table>

*Chromosome 1 refers to the IGH6*02-defined chromosome.

Data are shown for functional IGHV genes at the 3′ end of the IGHV locus.

No attempt was made to assign a gene or allelic variant to a chromosome where fewer than 10 VDJ rearrangements appeared to include that sequence.

x, Absent; ✓, present.
rearrangements of those chromosomes. The alternative IGHV1-3*02 allele was present in three genotypes, but was only present at a high enough frequency to be identified within one haplotype, where it was associated with 0.7% of rearrangements of that chromosome. No significant differences in rearrangement frequencies were detected between unmutated and highly mutated sequences. Additional variability appeared to be associated with partial haplotypes, and this was particularly evident for genes of the IGHD locus. For example, as shown in Fig. 3, the frequency of rearrangements utilizing the IGHD3-9*01 and IGHD3-10*01 genes was significantly higher on chromosomes that lacked the genes of the IGHD3-3 to IGHD2-8 locus (Mann–Whitney U test: \( p < 0.01 \) in both cases).

Discussion

The IgH V region gene locus contains in the order of 120 highly similar functional genes and pseudogenes interspersed among >700 repetitive elements that make up almost half of the locus (8). This has made the study of the chromosomal associations of these genes very difficult. The HapMap project and the 1000 Genomes Project have identified many variant sequences, but neither project is capable of generating comprehensive phased data for the locus. A recent report on the 1000 Genomes Project shows that the sequencing is either too short or too shallow for reliable reconstruction of the IGHV locus (6). In addition, both projects use libraries prepared from EBV-immortalized lymphoblastoid cell lines. By definition, these cells have lost genes from the Ig loci through gene rearrangement. Analysis of data from the HapMap project, for the identification of deletion polymorphisms, did not include the IGH, IGK, and IGL loci for this reason (5). However, although assembly of contigs for the locus still remains problematic, long-read 454 pyrosequencing is now being applied to the study of the Ab repertoire (17).
clarity that comes from haplotype analysis has allowed us to confirm and extend these initial findings. Although the size of the data sets and the rate of errors generated during pyrosequencing precluded proper analysis of those Ig genes that are rearranged at very low frequency, most genes could be included in the analysis. Those genes that could not be properly analyzed make a relatively small contribution to the overall Ab repertoires of these individuals. As sequencing costs and pyrosequencing error rates are both falling, the future generation of larger relatively error-free data sets will likely allow the confident inference, within a haplotype, of even these rarely used genes.

Our earlier study led to the inference of the existence of 14 previously unreported IGHV polymorphisms, as well as a new IGHD allele and a new IGHJ allele (16). In this study, using IGHV FR1 primers, we have been able to identify an additional putative polymorphism. These sequences have all been named as putative alleles by the inclusion of the descriptor “p” in their proposed allele names. This unofficial naming and reporting is important, for the existence of these alleles will not be accepted by the WHO/IUIS/IMGT Nomenclature Subcommittee for Igs (IG) and TCRs (TR), who stipulate that new allelic variants must be identified as unarranged genomic sequences. Thousands of VDJ rearrangements now stand as evidence of the existence of these putative alleles, and data from the dbSNP provide additional evidence in support of their existence. The principles governing the acceptance of allelic variants by the nomenclature subcommittee may therefore require revision, given that rapid confirmation of the existence of these new polymorphisms by the stipulated genomic screening is unlikely.

The common heterozygosity that we see at Ig gene loci undoubtedly contributes to repertoire diversity. Although many allelic variants differ from one another by just a single amino acid, even such small differences can give rise to Abs with quite different binding properties (29). Susceptibility to Haemophilus influenzae type b disease, for example, has been associated with allelic variation in the κ L chain genes. It is also now clear that the apparent heterozygosity that can be seen in genotypes is often a consequence of the carriage of multiple “alleles” on a single chromosome. Such duplication of Ig genes has been reported previously. By employing RFLP analysis with sequence-specific oligonucleotide probes, Sasso and colleagues (30) identified two separate loci for sequences that are now identified as IGHV3-30 and IGHV4-28, as well as for IGHV1-69 (31). They also claimed there can be multiple copies of the IGHV3-23 sequence on a single chromosome (32), and others have reported duplication of IGHV4-31 (33). Although it has not been possible to confirm duplication of IGHV3-23 in this study, as all sequences seen were identical to the IGHV3-23*01 sequence, duplications were seen for IGHV3-30 and IGHV1-69. The apparent duplicated IGHV3-30 sequences may actually be amplifications of the IGHV3-30-5 locus, as sequences that map to this locus have been shown to be identical to IGHV3-30 sequences (34). Duplications of IGHV1-2, IGHV3-11, and IGHV4-59, which have not previously been reported, were also seen.

Gene deletion polymorphisms were also inferred in the current study, and in many cases, homozygous deletions were inferred that were supported by the analysis of genotypes. In our earlier genotyping study, we were able to infer that one individual carried a homozygous deletion polymorphism of a number of contiguous IGHD genes (16). In the absence of phased data, however, single copy deletion polymorphisms are impossible to detect by such analysis. Very limited phased data for the IGHV locus were generated in the early 1990s, leading to the inference of an IGHV4-4 deletion polymorphism (35). The reporting of a complete sequence of the IGHV locus in 1998 also revealed the existence of indels involving the unmapped genes IGHV1-f, IGHV4-b, and IGHV5-a (8).

Additional IGHV deletion polymorphisms have recently been revealed. Chime and colleagues (20) have constructed IGHV haplotypes from single sperm, using multiplex PCR followed by microarray detection using IGHV gene-specific probes. They reported that two individuals of five carried heterozygous deletions of the IGHV4-39 gene. They have also reported deletions of IGHV4-61 (36) and of IGHV7-4-1, IGHV1-8, and IGHV3-9 (37). Recently, they amplified DNA sequence tags to define haplotypes for the complex region from IGHV3-30 to IGHV3-33, where a five-gene indel has been reported (34), and confirmed common deletions within this locus (33). However, a limitation of these studies is that they are only able to define haplotypes by the presence or absence of genes.

The haplotyping technique reported in the current study now reveals the full extent of diversity within the IGH locus because of its ability to define haplotypes that include the identification of allelic variants of each gene. This is perhaps seen most strikingly in our ability to define the haplotypes of the IGHV3-30 to IGHV3-33 region. Twelve different partial haplotypes were seen for this section of the IGHV locus, and these involved as many as four apparent gene deletion polymorphisms. We inferred these and other likely deletion polymorphisms from the absence of certain VDJ rearrangements. It is possible that the data reflect an incompatibility of some IGHV and IGHJ allelic combinations, rather than the absence of particular genes. It is also possible that epigenetic variation blocks recombination of some genes in some individuals. Whereas this study reports “functional haplotypes,” the fact that identical deletions were inferred in different individuals and the fact that most of these apparent deletion polymorphisms have been previously reported lead us to the working hypothesis that the deletions are not merely functional but are physical. In the case of the inferred deletion of the contiguous IGHJ4 genes from IGHD3-3 to IGHV2-8, whole genome shotgun sequencing demonstrates that this dramatic deletion is a physical one.

The value of haplotyping goes beyond the deliverable of a more complete description of the genes of the Ig gene loci. Haplotyping also brings clarity to the study of Ig gene rearrangement frequencies. Variability in recombination frequencies of different genes has been attributed either to variations in recombination signal sequences (38) or to variation in the accessibility of genes through chromatin remodeling (39). Although variation in RSS sequences undoubtedly contributes to unequal utilization of IGHV genes (40), the hypothesis that recombination is affected by chromosomal position and context is also well supported (41).

In this study, rearrangement frequencies were shown to vary between alleles. This variation was seen even where analysis of genomic sequencing data, including the original report of the locus (8), shows such alleles to be associated with identical RSS (data not shown). Rearrangement frequencies also appear to vary according to the genomic context of a gene, such that rearrangement frequencies were most similar among haplotypes that were most similar. This was most dramatically seen in the very high rearrangement frequencies for IGHD3-9 and IGHD3-10 on chromosomes that appear to lack the IGHD3-3 to IGHD2-8 genes. The IGHD3-9 and IGHD3-10 genes are immediately downstream of this deletion. This elevation in the rearrangement frequencies would not be expected to “plug” a hole in the repertoire, created by the absence of IGHD3-3, for although these genes are from the same IGHD gene family, they encode strikingly different amino acid sequences. It therefore may be more likely to be a reflection of changes in the noncoding sequences that flank these genes, as
a consequence of the gene deletions, and of the roles of these noncoding sequences in the regulation of gene rearrangement.

Before the development of modern sequencing technologies, when investigations of genomic differences were still very difficult, it was suggested that enormous variability of the Ig gene loci could exist within the human population (42). Analysis of thousands of VDJ rearrangements in this study now clearly demonstrates that this is true. This study has shown that the functional IGH haplotypes that encode the Ig H chain are incredibly diverse, and that this diversity includes a surprisingly high frequency of deletion and duplication polymorphisms. Where genes are present, dramatic variations are also seen in rearrangement frequencies, particularly with the genes of the IGHD locus. Together, this variation suggests that the repertoire of VDJ rearrangements may vary substantially between individuals. It may well be that individual variation in the genes of the Ig loci, and resulting repertoire variation, emerge as important contributors to individual variation in immunocompetency and individual susceptibility to Ab-mediated immunopathology.

If the contribution of the locus to health and disease is to be understood, haplotype variation and repertoire variation within the population must now be explored.

Disclosures

The authors have no financial conflicts of interest.

References