CrossMark
click for updates

Research

Cite this article: Collins AM, Wang Y, Roskin KM, Marquis CP, Jackson KJL. 2015 The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Phil. Trans. R. Soc. B* **370**: 20140236. <http://dx.doi.org/10.1098/rstb.2014.0236>

Accepted: 22 May 2015

One contribution of 13 to a theme issue
'The dynamics of antibody repertoires'.

Subject Areas:

immunology

Keywords:

IGHV, IGHD, IGHJ, BALB/c, C57BL/6,
immunoglobulin repertoire

Author for correspondence:

Andrew M. Collins

e-mail: a.collins@unsw.edu.au

Electronic supplementary material is available
at <http://dx.doi.org/10.1098/rstb.2014.0236> or
via <http://rstb.royalsocietypublishing.org>.

The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains

Andrew M. Collins¹, Yan Wang¹, Krishna M. Roskin², Christopher P. Marquis¹
and Katherine J. L. Jackson^{1,2}

¹School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, 2052 NSW, Australia

²Department of Pathology, School of Medicine, Stanford University, Stanford, CA 94305-5324, USA

The human and mouse antibody repertoires are formed by identical processes, but like all small animals, mice only have sufficient lymphocytes to express a small part of the potential antibody repertoire. In this study, we determined how the heavy chain repertoires of two mouse strains are generated. Analysis of IgM- and IgG-associated VDJ rearrangements generated by high-throughput sequencing confirmed the presence of 99 functional immunoglobulin heavy chain variable (IGHV) genes in the C57BL/6 genome, and inferred the presence of 164 IGHV genes in the BALB/c genome. Remarkably, only five IGHV sequences were common to both strains. Compared with humans, little N nucleotide addition was seen in the junctions of mouse VDJ genes. Germline human IgG-associated IGHV genes are rare, but many murine IgG-associated IGHV genes were unmutated. Together these results suggest that the expressed mouse repertoire is more germline-focused than the human repertoire. The apparently divergent germline repertoires of the mouse strains are discussed with reference to reports that inbred mouse strains carry blocks of genes derived from each of the three subspecies of the house mouse. We hypothesize that the germline genes of BALB/c and C57BL/6 mice may originally have evolved to generate distinct germline-focused antibody repertoires in the different mouse subspecies.

1. Introduction

The mammalian immune system has the ability to respond to almost any pathogen-associated macromolecule because of the incredible diversity of lymphocyte receptor molecules. B cells not only express receptor molecules on their surface, but also upon activation they release soluble receptor molecules called antibodies. These antibodies then initiate protective effector functions by binding to the foreign antigens that triggered their production [1]. The diversity of B-cell receptors and antibodies is made possible by the existence of multiple sets of highly similar genes that recombine to form the functional genes of the two polypeptides, called the heavy and light chains, that make up antibody molecules [2]. The rearranged heavy chain genes, called VDJ genes, are the products of recombining IGHV, IGHD and IGHJ genes. Light chain genes are similarly the products of recombination of light chain V and J genes. Antibody diversity is expanded still further by junctional diversification arising from the incorporation of P nucleotides derived from the opening of hairpin loops that form at gene ends as part of the rearrangement process, from exonuclease trimming of the recombining gene ends, and from the essentially random addition of N nucleotides between the recombining genes [3]. The permutations of recombining gene segments, the generation of junctional diversity and the permutations of associating heavy and light chain pairs together give the immune system the capacity to produce billions of different antibody molecules [4].

Variations in the repertoire of human germline genes can lead individuals and population groups to vary in their susceptibility to particular infections.

For example, differing susceptibility to infection with the important human pathogen *Haemophilus influenzae* has been linked to different alleles of the IGKV2D-29 and IGHV3-23 genes. The increased incidence of *H. influenzae* infection among the Navajo and other Native American populations has been linked to the high frequency of the IGKV2D-29*02 allele in these populations [5]. The IGKV2D-29 gene is critical for production of high affinity antibodies that target the *H. influenzae* capsule [6], but the IGKV2D-29*02 allele is unable to recombine efficiently because of a defective recombination signal sequence (RSS) [7]. IGHV3-23-encoded antibodies also target the polysaccharide capsule of *H. influenzae*, but while the IGHV3-23*01 gene encodes high affinity antibodies, the IGHV3-23*03 allele encodes antibodies with much reduced affinity [8], and carriage of the IGHV3-23*03 allele is likely to be associated with increased susceptibility to the infection. Similarly, it has recently been shown that some but not all IGHV1-69 alleles are readily able to form broadly neutralising anti-influenza antibodies [9], and an individual's susceptibility to influenza is therefore also likely to reflect the IGHV1-69 alleles that they carry.

Investigations of associations between particular antibody genes and disease susceptibilities have been hampered by the technical difficulties involved in the documentation of an individual's germline antibody genes. In recent years, the advent of high-throughput sequencing has made this much easier, leading to major advances in our understanding of individual variation in the repertoires of available human germline variable region genes. Many new allelic variations of IGHV genes have been identified [10–12] and numerous deletion polymorphisms have also been found in the IGHV [11,13] and the IGHD loci [13].

In contrast to the wealth of recent human studies, there has been surprisingly little application of high-throughput sequencing to the study of murine antibody genes, and differences in the germline genes that are available for antibody production in different mouse strains have received little attention. This is surprising as different strains of inbred mice have differing susceptibilities to infectious diseases [14–16] and to antibody-mediated pathologies [17]. In the absence of such studies, the recognized repertoire of murine germline genes has remained essentially unchanged for many years. This repertoire is dominated by sequences derived from the BALB/c and C57BL/6 strains. BALB/c mice were the early focus of immunogenetic studies, because of the availability of mineral-oil induced plasmacytomas from this strain [18]. The sequencing of the C57BL/6 genome subsequently led to two reports describing the assembly of nucleotide sequences of the heavy chain locus, and the identification of almost 200 IGHV genes and pseudo-genes [19,20]. The sequence of Riblet [19] provided the mapped genes that were then used to develop the ImMunoGeneTics (IMGT) murine gene nomenclature, based upon gene family names and gene positions within the locus [21]. Apparent allelic variants from other strains were assigned by IMGT to mapped genes by comparison with this C57BL/6 sequence (see <http://www.imgt.org/IMGTrepertoire/index.php>). Subsequently, half of the heavy chain IGHV gene locus of the 129/Sv mouse strain was sequenced, providing additional allelic variants [22]. Studies to further clarify the immunoglobulin genes of different mouse strains then appear to have stopped. Certainly, no new allelic variants have been added to the IMGT repertoire since 2007.

This study was conducted to infer the complete BALB/c genotype of rearrangeable heavy chain variable region genes from an analysis of VDJ rearrangements, and to study how these genes rearrange to generate the murine heavy chain repertoire. To perform this analysis, we first demonstrated the utility of our genotyping approach in an analysis of VDJ rearrangements from C57BL/6 mice, allowing us to compare our inferences with the reported C57BL/6 immunogenotype. The BALB/c immunogenotype was then determined, and the IGHV locus of the BALB/c strain shows an extraordinary divergence from the IGHV locus of the C57BL/6 strain. Results are presented demonstrating that the VDJ repertoires formed from these genes are germline-focused and surprisingly restricted. The number of available IGHD genes is small, many of them are highly similar, and there are strong biases towards particular IGHD gene reading frames (RFs). Mechanisms that might generate diversity such as D–D fusion and the use of IGHD genes in inverted orientations are rare events, if they occur at all. P nucleotide additions are more common, but their contribution to diversity can also be considered to be germline-derived. On the other hand, the stochastic process of N addition contributes relatively little to the diversity of the naive repertoire. Finally, there is little diversification of the naive repertoire through the process of somatic point mutation, because even IgG-associated murine VDJ genes carry relatively few mutations in these laboratory mice.

The implied importance of germline sequences and the apparent divergence of genes between the two mouse strains suggest that the loci have evolved under very strong selection pressures. These results are discussed with respect to recent reports highlighting the mosaic structure of the genomes of classical inbred laboratory mouse strains, and we conclude that it is likely that the heavy chain variable region genes of the BALB/c and C57BL/6 strains are derived from different subspecies of the house mouse.

2. Material and methods

(a) Sample collection and sequence generation

Splenocytes were isolated from eight C57BL/6 and eight BALB/c mice using Ficoll-Paque PREMIUM 1.084 (GE Healthcare). Total RNA was extracted from each sample using AllPrep® DNA/RNA/miRNA (QIAGEN). All eight samples from each strain were pooled equally by their RNA concentration and mRNA was extracted from pooled RNA by magnetic bead separation using the Dynabeads® mRNA DIRECT™ Kit (Life Technology). 5' Rapid amplification of cDNA ends (RACE) was performed with first strand cDNA synthesis on the mRNA samples by the SMARTer™ RACE cDNA Amplification Kit (Clontech). VDJ sequences were then amplified by polymerase chain reaction (PCR). The forward primers incorporated the Nested Universal Primer A of the SMARTer RACE cDNA Amplification Kit (5'–AAGCAGTGGTATCAACGCAGAGT–3'). Reverse primers were designed for both mice strains based upon the 5' end of CH1 region of IgG1, IgG2a, IgG2b and IgG2c (5'–CASABMCAGGGGCCAGTGGATAGAC–3'), IgG3 (5'–TGCAGCCAGGGACCAAGGGATAGAC–3') and IgM (5'–GGGAAGACATTTGGGAAGGACTGAC–3'). The 454 Lib-L Primer A with Multiplex Identifier sequences (MIDs) and Primer B were added to the reverse and forward primers, using standard 454 methods. PCRs were performed using the FastStart High Fidelity PCR System (Roche), with 0.4 μM of the forward primer and 0.4 μM of the reverse primer. PCR was initiated with 3 min at 95°C, followed

by 32 cycles of 30 s at 95°C, 30 s at 65°C and 42 s at 72°C, and ended with a final extension of 2 min at 72°C. The amplified products were first purified using QIAquick PCR Purification Kits (QIAGEN) and then further purified by gel extraction using QIAquick Gel Extraction Kits (QIAGEN). The purified PCR products were then sent to the Australian Genome Research Facility (QLD, Australia) and the Institute for Immunology & Infectious Diseases, Murdoch University (WA, Australia) for 454 sequencing. We have previously calculated the error rate for PCR amplification and 454 sequencing of immunoglobulin sequences as 0.1% [23].

(b) Analysis of VDJ sequences

The isotype of each read was determined from the presence of sequence with complete identity to the CH1 region of the isotype-specific reference sequences, upstream from the IGHC primer. Reads that could not be assigned to an isotype were discarded. Sequences were then aligned against the IMGT databases of germline murine IGHV, IGHD and IGHJ genes [21] using the IgBLAST program [24], and the most closely matching germline IGHV, IGHD and IGHJ genes were recorded. Nucleotide mismatch counts were also recorded for each sequence, as were non-template encoded nucleotide additions at the IGHV-D (N1) and IGHD-J (N2) joins. Non-productive sequences, with either out-of-frame IGHJ or which included stop codons, and duplicate sequences were removed from the dataset. Clonally related sequences were identified on the basis of shared IGHV and IGHJ genes and CDR3 nucleotide sequences that clustered by centroid clustering with a 90% identity threshold [25]. A representative sequence for each isotype was selected from each clone set, and other sequences were excluded from further analysis.

An additional dataset of publicly available BALB/c IgG sequences was accessed. These VDJ rearrangements were amplified from cDNA using fifteen IGHV gene family-specific forward primers and an IgG-specific reverse primer, and were sequenced using the Illumina platform [26]. IgG subclasses were assigned and the datasets were processed in the same manner as for the 454 sequenced datasets, with the isotype-specific reference sequences adjusted for the study's IgG primer.

(c) Immunogenotyping of mouse strains

The sets of germline variable region IGHV, IGHD and IGHJ genes that are carried by each mouse strain were investigated by analysis of the 454 VDJ datasets, using a method previously developed for analysis of human immunogenotypes [27]. IgM sequences were analysed to detect the presence of genes of the IMGT murine IGHV, IGHD and IGHJ gene repertoires in the VDJ rearrangements. The majority of BALB/c VDJ rearrangements did not align perfectly to any IGHV gene in the IMGT repertoire and the BALB/c dataset was realigned against a composite repertoire made up of all unique IGHV genes in the IMGT and VBASE2 [28] repertoires, as well as sequences associated with the NCBI IgBLAST utility [24]. Rearranged sequences that aligned perfectly to a germline IGHV gene were accepted as evidence of the presence of that gene in the repertoire of the mouse strain, provided that later analysis did not reveal the likelihood that a commonly rearranged gene had given rise to a small number of apparent alignments to an alternative gene, through the process of somatic point mutation. Alignments to each IGHV gene were analysed for the number of mismatches seen between the VDJ sequences and the IGHV gene, leading to the identification of sets of sequences that aligned to particular genes with a shared number of mismatches. These VDJ rearrangements were manually reviewed, and where the set of sequences included diverse IGHD genes, IGHJ genes and N regions, as well as shared IGHV gene mismatches, a probable new IGHV sequence was inferred. The Illumina IgG dataset was then searched, and if the same sequence was identified

within this independent dataset, it was confirmed as a putative IGHV gene. Analysis of the Illumina dataset showed biases in the amplification of sequences containing different IGHV gene families, with an almost total absence of sequences using the IGHV9 gene family. In the absence of Illumina dataset alignments to a putative IGHV9-family gene that was abundant in the IgM 454 dataset, an additional search for confirming sequences was made among the IgG 454 dataset sequences. The presence of multiple identical IGHV sequences in the IgG-associated VDJ 454 gene dataset was accepted as confirmation of the sequence as a putative BALB/c IGHV gene.

Because there are so few reported IGHJ genes in the mouse, the presence or absence of IGHJ genes in the genomes of the two mouse strains was simply determined by the abundance of sequences containing each reported IGHJ gene, after confirming the absence of additional alleles through mutation analysis. The identification of IGHD genes in each genome from VDJ rearrangements was more challenging, because IGHD genes are short and highly similar. Their lengths are also usually reduced by exonuclease activity. To complicate matters, some IGHD genes are identical, while other genes differ from one another by 1 or 2 nt at the 3' or 5' ends of the genes. These nucleotides are often removed by exonuclease activity, which means that random N nucleotide addition can give one gene in a VDJ rearrangement the appearance of another. To confirm the presence or absence of reported IGHD genes in the BALB/c and C57BL/6 genomes, the lengths of each identified IGHD gene within the VDJ dataset were analysed. The IGHD gene repertoire, as reported by IMGT, was first determined. These IGHD genes were then confirmed as being present within the genomes of the two strains if abundant full-length IGHD sequences were seen within the datasets of VDJ rearrangements.

Once the apparent germline repertoires of the IGHV, IGHD and IGHJ genes of both strains were determined, the datasets were re-aligned against strain-specific repertoires to determine the rearrangement frequencies of each gene. The length of each expressed IGHD gene was then determined, following filtering of the VDJ alignments to improve accuracy of IGHD alignments. The filtering required any D-REGION to be more than 7 nt in length and to include at most two mismatches if the length of the D-REGION was 10 nt, a single mismatch if it was 9 nt and no mismatches if it was 8 nt. The nucleotides of the V-D and D-J junctions were identified based on the IgBLAST output, and the lengths of the N1 and N2 regions were determined. Finally, the 454 IgG-associated VDJ sequences were also analysed, and the number of mutations in each IGHV gene was recorded.

(d) Identification of additional contributions to CDR3 diversity

Molecular mechanisms that further diversify VDJ rearrangements include variable exonuclease processing of gene segment ends, palindromic (P) nucleotide inclusions, the presence of multiple IGHD segments within a single rearrangement and the utilization of IGHD segments in different RFs and orientations. To examine how these mechanisms contribute to the C57BL/6 and BALB/c repertoires, datasets were reduced to include a single representative for each clone lineage across all IgG subclasses. Exonuclease trimming for each gene end was calculated by comparison of the rearranged genes ends with the full-length germline sequence. In the absence of nucleotide loss, putative P nucleotide motifs were identified in the relevant downstream (IGHV, 3' IGHD) or upstream (5' IGHD, IGHJ) junctional nucleotides. The frequency of these observed motifs were compared with their expected occurrence as a consequence of N addition. The probability of N addition adding an A or T base was taken to be 0.15 and of adding a G or C base was 0.35. The approach of Meier and Lewis [29], based on the

binomial distribution, was then used to determine if any over-represented putative P nucleotide motifs were significantly more abundant than could be expected to have arisen through N addition.

IGHD RFs were determined based upon whether the amino acids encoded by the D-REGION contributions to a VDJ were the result of translating the germline IGHD from the RF indexed from the first (RF1), second (RF2) or third (RF3) nucleotide of the sequence. All sequences were considered for possible alternative IGHD alignments to inverted orientation IGHDs. In the absence of an original IGHD assignment, inverted IGHDs were allowed using the same rules as the regular orientation, however, if both regular and inverted alignments were found, the longer, less mismatched IGHD alignment was selected. Secondary IGHD segments within a single rearrangement were sought within N1- and N2-REGIONS of at least 8 nt, and required alignments meeting the D-REGION filtering criteria and with the gene order being consistent with the order of genes within the IGHD locus.

3. Results

After removal of duplicate sequences and clonally related sequences, 15 103 unique BALB/c and 20 928 unique C57BL/6 IgM-associated VDJ sequences and 1836 unique BALB/c and 10 241 unique C57BL/6 IgG-associated VDJ sequences remained. After similar removal of duplicate sequences and clonally related sequences, 47 910 unique IgG-associated BALB/c sequences were identified in the Illumina dataset. Raw reads of the complete dataset are available from the European Nucleotide Archive under project accession number PRJEB8745 (www.ebi.ac.uk/ena).

When the 454 IgM-associated sequences from C57BL/6 mice were aligned against the IMGT repertoire of germline murine IGHV, IGHD and IGHJ genes, perfect matches were seen to 97 IGHV sequences in the IMGT database. These 97 sequences correspond to 101 unique IGHV genes, for IGHV7-1*01 and IGHV7-1*03 are identical, as are IGHV2-6*01 and IGHV2-6-8*01, IGHV1-62-2*01 and IGHV1-71*01, and IGHV5-6*01 and IGHV5-6-1*01. All but two of these 101 genes are defined by IMGT as functional genes. IGHV1-62-3*01 and IGHV1-23*01 are defined as IMGT Open Reading Frames rather than functional genes. IGHV1-62-3*01 was seen in seven unique VDJ rearrangements, while IGHV1-23*01 was seen as a single rearrangement. Two additional germline genes were identified that are absent from the IMGT repertoire but are included in both the VBASE2 and NCBI databases, and were identified in the locus assembly compiled by Johnston *et al.* [20]. *musIGHV211* (Q52.9.59) was seen in 275 unique VDJ rearrangements, associating with a wide range of IGHD and IGHJ gene combinations. *musIGHV269* (J558.1.85) was seen in 13 unique VDJ rearrangements made up with various IGHD and IGHJ. The sequence is defined as a pseudogene by VBASE2 because the 3' terminal nucleotides encode a stop codon. All 13 sequences lacked the 3' stop codon because of exonuclease removal of nucleotides. The complete repertoire of rearrangeable C57BL/6 IGHV genes identified in this study and the rearrangement frequencies of the genes are shown in table 1.

Fourteen sequences that have been reported as present in the C57BL/6 genome and that are defined as functional C57BL/6 genes by IMGT were missing from the dataset of VDJ rearrangements. If they exist, they may be incapable of

rearrangement. If, however, they are functional, they make a trivial contribution to the C57BL/6 heavy chain repertoire. No alignments were seen to 39 IGHV sequences that IMGT reports as functional C57BL/6 'genes of uncertain origin'. Their absence from our large dataset of rearrangements makes it unlikely that any of these sequences are real IGHV genes.

When the BALB/c-derived 454 sequences were aligned against the IMGT repertoire, it was immediately apparent that IGHV sequences were present in the rearrangements that are not present in the IMGT database. As a consequence, alignments by IgBLAST were incorrectly made to similar genes, resulting in just 57% of BALB/c IgM-associated VDJ rearrangements aligning to IMGT IGHV genes without mismatches. Such perfect alignments had been seen in 81% of the C57BL/6 IgM-associated VDJ rearrangements. Analysis of the frequency distribution of mismatches in sets of rearrangements of each identified IGHV gene showed conspicuous clusters of BALB/c sequences with shared mismatch distance from the most closely matched IMGT gene (data not shown). For example, while there were no perfect alignments to IGHV1-5*01, and two sequences with four mismatches were the 'best' alignments seen, there were 123 alignments with five mismatches. A review of these sequences confirmed that the IGHD and IGHJ gene usage was varied, and that all sequences shared the same mismatches. This almost certainly is the result of the presence in the BALB/c genome of an IGHV gene that is absent from the IMGT repertoire and that differs from the IGHV1-5*01 sequence at five nucleotide positions. This kind of approach to the identification of putative polymorphisms is now well established for human antibody genes [30].

Further investigation of sequence clusters led to the discovery that some of the putative IGHV sequences that had been identified are present in the VBASE2 repertoires or are listed in association with the NCBI IgBLAST utility. A new repertoire of germline genes, including all murine sequences from the three sources was therefore compiled and used to realign the BALB/c IgM-associated VDJ sequences.

Realignment of the sequences against the combined IMGT/VBASE2/NCBI repertoire of germline IGHV genes led to the identification of perfect alignments to many IGHV sequences that are found in the VBASE2 and NCBI databases, but are not present in the IMGT database. A few of these putative IGHV genes were present as a single alignment, and some others were present at low frequency (less than 10 alignments). To confirm such low abundance IGHV genes as putative BALB/c IGHV genes, we investigated the presence of these sequences in the Illumina database of IgG-associated VDJ rearrangements. Where perfect alignments were seen to such IGHV genes, the presence of the genes in the BALB/c genome was provisionally accepted. These provisional genes were then compared with other similar genes in the BALB/c genome. If a review could exclude the likelihood that mutations of abundant rearrangements of one IGHV gene gave rise to small numbers of rearrangements that appeared to be a different IGHV gene, the sequence was accepted as a putative BALB/c gene.

The frequency distribution of mutations in each identified IGHV gene was again analysed, and some conspicuous clusters of sequences remained in the BALB/c dataset. Groups of sequences that used the same IGHV gene and which shared unexpected numbers of mismatches were identified. The sequences were manually reviewed to identify groups that

Table 1. IGHV genes and their rearrangement frequencies, in a dataset of 20 928 C57BL/6 IgM-associated VDJ rearrangements.

IMGT ^a	VBASE2 ^b	Johnston <i>et al.</i> [20] ^c	frequency (%)
IGHV1–64*01	musIGHV346	J558.67.166	10.38
IGHV3–6*01	musIGHV253	36–60.6.70	5.33
IGHV1–80*01	musIGHV362	J558.83.189	5.08
IGHV1–69*01	musIGHV353	J558.72.173	4.71
IGHV1–72*01	musIGHV057	J558.75.177	4.47
IGHV1–53*01	musIGHV329	J558.53.146	4.15
IGHV1–55*01	musIGHV332	J558.55.149	4.08
IGHV9–3*01	musIGHV244	VGAM3.8.3.61	3.82
IGHV1–59*01	musIGHV338	J558.59.155	3.40
IGHV1–82*01	musIGHV364	J558.85.191	3.22
IGHV1–26*01	musIGHV301	J558.26.116	3.01
IGHV14–2*01	musIGHV231	SM7.2.49	3.01
IGHV6–3*01	musIGHV262	J606.1.79	2.23
IGHV2–2*01	musIGHV186	Q52.2.4	2.21
IGHV1–85*01	musIGHV367	J558.88.194	2.14
IGHV14–4*01	musIGHV247	SM7.4.63	2.12
IGHV1–61*01	musIGHV340	J558.61.157	1.94
IGHV1–78*01	musIGHV360	J558.81.187	1.93
IGHV8–12*01	musIGHV354	3609.12.174	1.89
	musIGHV211	Q52.9.29	1.62
IGHV2–3*01	musIGHV190	Q52.3.8	1.59
IGHV14–3*01	musIGHV236	SM7.3.54	1.55
IGHV7–3*01	musIGHV245	S107.3.62	1.46
IGHV1–15*01	musIGHV287	J558.12.102	1.31
IGHV1–52*01	musIGHV328	J558.52.145	1.25
IGHV1–66*01	musIGHV350	J558.69.170	1.20
IGHV1–9*01	musIGHV281	J558.6.96	1.18
IGHV4–1*01	musIGHV227	X24.1pg.45 (P) ^d	1.16
IGHV1–81*01	musIGHV363	J558.84.190	1.11
IGHV2–5*01	musIGHV201	Q52.7.18	0.90
IGHV1–74*01	musIGHV110	J558.77.180	0.88
IGHV3–1*01	musIGHV228	36–60.1.46	0.88
IGHV14–1*01	musIGHV226	SM7.1.44	0.80
IGHV10–1*01	musIGHV270	VH10.1.86	0.75
IGHV1–50*01	musIGHV326	J558.50.143	0.70
IGHV5–17*01	musIGHV219	7183.20.37	0.68
IGHV8–8*01	musIGHV336	3609.7.153	0.66
IGHV2–6*03	musIGHV205	Q52.8.22	0.65
IGHV1–58*01	musIGHV337	J558.58.154	0.63
IGHV9–1*01	musIGHV240	VGAM3.8.1.57	0.60
IGHV5–9–1*02	musIGHV207	7183.14.25	0.56
IGHV1–75*01	musIGHV094	J558.78.182	0.56
IGHV6–6*01	musIGHV265	J606.4.82	0.56
IGHV1–12*01	musIGHV284	J558.9.99	0.56
IGHV1–77*01	musIGHV359	J558.80.186	0.53
IGHV1–76*01	musIGHV435	J558.79.184	0.46

(Continued.)

Table 1. (Continued.)

IMGT ^a	VBASE2 ^b	Johnston <i>et al.</i> [20] ^c	frequency (%)
IGHV10–3*01	musIGHV275	VH10.3.91	0.45
IGHV5–6*01 or IGHV5–6–1*01	musIGHV192	7183.7.10	0.38
IGHV9–4*01	musIGHV254	VGAM3.8.4.71	0.37
IGHV9–2*01	musIGHV242	VGAM3.8.2.59	0.36
IGHV5–4*01	musIGHV188	7183.4.6	0.30
IGHV1–63*01	musIGHV345	J558.66.165	0.26
IGHV1–62–2*01 or IGHV1–71*01	musIGHV357	J558.64.162 or J558.74.176	0.25
IGHV7–1*01 or IGHV7–1*03	musIGHV224	S107.1.42	0.28
IGHV2–6–8*01 or IGHV2–6*01	musIGHV214	Q52.10.33	0.24
IGHV1–19*01	musIGHV293	J558.18.108	0.22
IGHV5–16*01	musIGHV218	7183.19.36	0.20
IGHV1–7*01	musIGHV277	J558.4.93	0.20
IGHV1–20*01	musIGHV294	J558.19.109	0.19
IGHV1–67*01	musIGHV351	J558.70pg.171 (P)	0.19
IGHV13–2*01	musIGHV260	3609N.2.77	0.16
IGHV11–2*01	musIGHV235	VH11.2.53	0.15
IGHV1–39*01	musIGHV313	J558.39.129	0.14
IGHV1–49*01	musIGHV324	J558.49.141	0.14
IGHV1–5*01	musIGHV274	J558.3.90	0.14
IGHV1–4*01	musIGHV272	J558.2.88	0.10
IGHV1–47*01	musIGHV320	J558.47.137	0.09
IGHV8–5*01	musIGHV325	3609.4.142	0.08
IGHV15–2*01	musIGHV280	VH15.1.95	0.08
IGHV5–12*01	musIGHV203	7183.12.20	0.14
IGHV7–4*01	musIGHV249	S107.4.65	0.08
	musIGHV269 (P)	J558.1.85	0.08
IGHV2–9*01	musIGHV222	Q52.13.40	0.08
IGHV2–4*01	musIGHV195	Q52.5.13	0.08
IGHV1–42*01	musIGHV315	J558.42.132	0.06
IGHV1–54*01	musIGHV331	J558.54.148	0.06
IGHV1–36*01	musIGHV310	J558.36.126	0.06
IGHV5–9*01	musIGHV198	7183.9.15	0.05
IGHV3–8*01	musIGHV257	36–60.8.74	0.05
IGHV1–62–3*01 ORF	musIGHV343	J558.65.163	0.04
IGHV1–22*01	musIGHV297	J558.22.112	0.04
IGHV12–3*01	musIGHV261	VH12.1.78	0.04
IGHV1–34*01	musIGHV308	J558.34.124	0.03
IGHV8–4*01	musIGHV322	3609.3.139	0.03
IGHV5–2*01	musIGHV185	7183.2.3	0.03
IGHV1–84*01	musIGHV366	J558.87.193	0.02
IGHV5–15*01	musIGHV217	7183.18.35	0.02
IGHV2–7*01	musIGHV216	Q52.11.34	0.02
IGHV8–6*01	musIGHV330	3609.5.147	0.01
IGHV1–18*01	musIGHV291	J558.16.106	0.01
IGHV1–31*01	musIGHV305	J558.31.121	0.01

(Continued.)

Table 1. (Continued.)

IMGT ^a	VBASE2 ^b	Johnston <i>et al.</i> [20] ^c	frequency (%)
IGHV1–43*01	musIGHV316	J558.43.133	0.01
IGHV6–5*01	musIGHV264	J606.3.81	0.01
IGHV1–37*01	musIGHV311	J558.37.127	0.01
IGHV1–74*04	musIGHV062		0.01
IGHV3–3*01	musIGHV248	36–60.3.64	0.01
IGHV3–5*01	musIGHV251	36–60.5.67	0.01
IGHV1–23*01 ORF	musIGHV298	J558.23.113	0.01
IGHV11–1*01	musIGHV230	VH11.1.48	0.01

^aThe nomenclature of the ImMunoGeneTics group [21].^bThe nomenclature of the VBASE2 database [26].^cThe nomenclature of Johnston *et al.* [20].^dSequences that have been reported to be pseudogenes in one or other of the datasets associated with the three nomenclatures are indicated with (P).

used disparate IGHD and IGHJ genes, and shared all mismatches to the aligned germline IGHV gene. The Illumina database of IgG-associated VDJ rearrangements was then searched, and where sequences were found that shared the same mismatches, the sequences were considered to be putative, unreported IGHV genes. Electronic supplementary material, table S1, shows the 34 newly identified putative IGHV sequences in FASTA format. Eleven of these sequences differed from previously reported IGHV genes by a single nucleotide, but seven sequences differed by 10 or more nucleotides, with one sequence differing by 16 nt.

The final tally of unique rearrangeable putative IGHV germline gene sequences in the BALB/c genome was 164, and these were used at frequencies ranging from 0.01 to 5.94%. The putative genes and their utilization frequencies are listed in table 2. Eighty-two of the unique IGHV sequences match 86 IGHV genes in the IMGT database. Four sequences could be derived from either or both of four identical gene pairs: IGHV5–6–1*01 and IGHV5–6*01; IGHV5–9–1*01 and IGHV5S9*01; IGHV5–9–2*01 and IGHV5S4*02; IGHV5–9*03 and IGHV5–9–5*01. A further 48 sequences are not present in the IMGT database, but are present in either the VBASE2 or NCBI databases.

Only five IGHV sequences were identified in both C57BL/6 and BALB/c mice. Both strains appear to carry IGHV5–6*01 and/or the identical IGHV5–6–1*01. They also appear to carry IGHV1–69*01, IGHV2–3*01, IGHV2–5*01 and IGHV5–2*01. Interestingly, the utilization frequencies of these genes vary substantially between the strains. In particular, the IGHV1–69*01 gene is used by 4.71% of all C57BL/6 VDJ sequences but is used by just 0.62% of BALB/c VDJ sequences.

The IGHD loci of the two strains are more similar. We identified the presence of eight unique IGHD sequences in C57BL/6 mice, and 10 unique sequences in BALB/c mice. The eight unique C57BL/6 IGHD sequences are likely to correspond to nine IGHD genes, for IGHD2–5*01 and IGHD2–6*01 are identical, and both have been reported in C57BL/6 mice [32]. Similarly the 10 unique BALB/c IGHD sequences are likely to correspond to 12 IGHD genes, for IGHD2–2*01 and IGHD2–7*01 are identical, as are IGHD2–1*01 and IGHD2–8*01. These duplicate genes have been reported in BALB/c mice [33]. The IMGT IGHD gene repertoires of the

two strains and the genes that were confirmed in this study are shown in table 3. Six of the IGHD sequences were carried by both strains, and different allelic variants of the IGHD 3–2 gene are also carried by the two strains. The genes identified for each strain were in general agreement with those reported by IMGT, however, the reported C57BL/6 genes IGHD1–3*01 and IGHD3–1*01 could not be confirmed as present and functional. Similarly the BALB/c genes IGHD2–9*01, IGHD2–11*01 and IGHD4–1*02 could not be confirmed. Small numbers of alignments were seen to IGHD4–1*02. This 10 nt sequence is little different to the 11 nt IGHD4–1*01 sequence, and both alleles are listed by IMGT in the BALB/c genome. The few BALB/c alignments that were seen to IGHD4–1*02 are most probably the result of exonuclease trimming and N nucleotide additions to the IGHD4–1*01 gene. Both strains carry the IGHJ2*01, IGHJ3*01 and IGHJ4*01 genes. BALB/c mice additionally carry the IGHJ1*01 gene while C57BL/6 mice carry the IGHJ1*03 gene.

After the determination of the apparent germline gene repertoires of both strains, the VDJ rearrangements were re-aligned against strain-specific germline gene sets, and the rearrangement frequencies of each of the genes and putative genes were determined. The IGHV gene rearrangement frequencies are included in tables 1 and 2, the IGHD frequencies are shown in table 3, and the IGHJ frequencies are shown in table 4. The average lengths of IGHD genes within the VDJ rearrangements were also determined, and are included in table 3. Calculation of the average length of each gene within VDJ rearrangements is complicated by the fact that some rearrangements include very short IGHD genes. It is impossible to be confident of any alignment of a human IGHD gene that is shorter than 8 nt [34], and alignments of 8 or more nucleotides were therefore used to gauge the extent of exonuclease activity within the mouse. Such IGHD alignments were seen in 63.2% of BALB/c and 62.7% of C57BL/6 VDJ. The average lengths of these IGHD genes within VDJ rearrangements are also shown in table 3, and this shows that the loss of 3 or 4 nt from each end of the IGHD gene is common. The majority of IGHDs were found in all three RFs. Although RFs were not equally used in the productive repertoires of C57BL/6 (RF1 15.4%, RF2 7.7%, RF3 76.9%) or BALB/c (RF1 15.6%, RF2 8.9%, RF3 75.5%), the usage frequencies were highly consistent between the two strains. Each IGHD

Table 2. IGHV genes and their rearrangement frequencies in a dataset of 15 103 BALB/c IgM-associated VDJ rearrangements. Previously reported genes are listed using three different nomenclatures, and additional previously unreported putative IGHV genes are also shown.

IMGT	VBASE2	Haines <i>et al.</i> [31]	putative IGHV genes	frequency (%)
IGHV3 – 2*02	musIGHV126			5.84
	musIGHV386			4.00
IGHV9 – 2 – 1*01	musIGHV157			3.29
IGHV9 – 3*03	musIGHV155			3.15
IGHV5 – 4*02	musIGHV181			2.65
IGHV5S12*01	musIGHV176			2.40
IGHV7 – 3*02	musIGHV158			2.22
IGHV9 – 3 – 1*01	musIGHV156			2.09
IGHV14 – 3*02	musIGHV125			2.08
IGHV14 – 1*02	musIGHV141			2.08
IGHV6 – 6*02	musIGHV114			1.96
IGHV3 – 6*02	musIGHV120			1.88
IGHV5 – 17*02	musIGHV139			1.82
IGHV5 – 12 – 2*01	musIGHV177			1.81
IGHV4 – 1*02	musIGHV142			1.74
IGHV5 – 12 – 1*01	musIGHV160			1.57
	musIGHV398			1.56
	musIGHV585			1.53
	musIGHV702			1.50
IGHV1S82*01				1.26
	musIGHV023			1.25
IGHV2 – 5*01	musIGHV201			1.24
	musIGHV532			1.22
IGHV9 – 4*02	musIGHV118			1.21
IGHV5 – 9 – 4*01	musIGHV122			1.19
IGHV14 – 4*02	musIGHV154			1.19
IGHV2 – 2*02	musIGHV183			1.15
IGHV2 – 6 – 7*01	musIGHV134			1.10
			balbcdGHV023	1.10
			balbcdGHV027	1.06
	musIGHV671			1.06
IGHV10 – 1*02	musIGHV028			1.04
	musIGHV021			1.03
	musIGHV657			1.02
	musIGHV559			0.96
	musIGHV560			0.93
			balbcdGHV022	0.93
IGHV2 – 6*02	musIGHV132			0.93
IGHV5 – 6 – 4*01	musIGHV178			0.89
	musIGHV629			0.88
	musIGHV655			0.87
			balbcdGHV020	0.84
IGHV9 – 1*02	musIGHV151			0.78
IGHV2 – 3*01	musIGHV190			0.77
IGHV1S53*02	musIGHV058			0.69

(Continued.)

Table 2. (Continued.)

IMGT	VBASE2	Haines <i>et al.</i> [31]	putative IGHV genes	frequency (%)
IGHV1S33*01	musIGHV049			0.68
			balbcdGHV021	0.66
IGHV2–6–1*01	musIGHV146			0.65
			balbcdGHV008	0.64
IGHV10S3*01	musIGHV029			0.64
IGHV1–84*02	musIGHV034			0.62
IGHV2–6–4*01	musIGHV175			0.62
IGHV1–69*02	musIGHV045			0.60
IGHV1–69*01	musIGHV707			0.59
			balbcdGHV026	0.59
	musIGHV588			0.59
IGHV5–9–3*01	musIGHV135			0.55
	musIGHV014			0.54
		J558.22		0.53
IGHV4–2*02	musIGHV149			0.53
IGHV2–9–1*01	musIGHV171			0.52
IGHV1–4*02	musIGHV074			0.50
			balbcdGHV025	0.50
			balbcdGHV015	0.47
	musIGHV528			0.46
IGHV11–2*02	musIGHV153			0.46
	musIGHV433			0.45
			balbcdGHV009	0.45
	musIGHV710			0.42
			balbcdGHV006	0.40
IGHV5–12*02	musIGHV148			0.40
IGHV1S68*02				0.40
		J558.13		0.37
IGHV3–1*02	musIGHV138			0.37
IGHV1S72*01	musIGHV518			0.37
		J558.32		0.35
	musIGHV480			0.35
IGHV6–7*02	musIGHV115			0.33
	musIGHV483			0.33
IGHV1S122				0.32
IGHV15–2*02				0.30
IGHV1–63*02	musIGHV066			0.30
IGHV5–9–1*01; IGHV5S9*01	musIGHV147			0.30
IGHV1–20*02	musIGHV479			0.30
IGHV1S123*01	musIGHV683			0.30
			balbcdGHV011	0.29
			balbcdGHV018	0.27
			balbcdGHV007	0.27
IGHV5–6*01; IGHV5–6–1*01	musIGHV192			0.27
IGHV1–54*03				0.26

(Continued.)

Table 2. (Continued.)

IMGT	VBASE2	Haines <i>et al.</i> [31]	putative IGHV genes	frequency (%)
			balbcdGHV028	0.26
IGHV7 – 1*02	musIGHV168			0.26
IGHV5 – 15*02	musIGHV166			0.25
	musIGHV616			0.25
		J558.52		0.24
	musIGHV612			0.24
			balbcdGHV019	0.23
IGHV5 – 6 – 3*01	musIGHV165			0.23
	musIGHV396			0.23
	musIGHV578			0.23
IGHV2 – 6 – 2*01	musIGHV162			0.22
	musIGHV592			0.22
		J558.44		0.21
			balbcdGHV032	0.21
			balbcdGHV002	0.19
			balbcdGHV024	0.19
			balbcdGHV005	0.18
		J558.18		0.18
			balbcdGHV014	0.18
			balbcdGHV010	0.16
	musIGHV672			0.16
			balbcdGHV001	0.15
IGHV1S12*01	musIGHV053			0.15
IGHV5 – 9 – 2*01; IGHV5S4*02	musIGHV145			0.14
	musIGHV667			0.14
			balbcdGHV012	0.12
			balbcdGHV013	0.11
			balbcdGHV003	0.11
	musIGHV668			0.11
	musIGHV710			0.11
IGHV13 – 2*02	musIGHV116			0.10
IGHV1S34*01				0.09
IGHV9S8*01				0.09
			balbcdGHV029	0.09
IGHV2 – 5 – 1*01	musIGHV159			0.09
IGHV5 – 2*01	musIGHV185			0.09
	musIGHV042			0.08
IGHV2 – 9*02	musIGHV137			0.08
IGHV9 – 2*02	musIGHV152			0.08
			balbcdGHV031	0.08
	musIGHV394			0.08
	musIGHV555			0.08
	musIGHV560			0.08
			balbcdGHV030	0.08
IGHV12 – 3*02				0.08

(Continued.)

Table 2. (Continued.)

IMGT	VBASE2	Haines <i>et al.</i> [31]	putative IGHV genes	frequency (%)
		J558.29		0.08
		J558.36		0.08
			balbcdGHV004	0.07
			balbcdGHV017	0.07
		J558.20		0.07
IGHV2–6–5*01	musIGHV174			0.07
IGHV1S135*01				0.06
IGHV5–6–2*01	musIGHV163			0.06
IGHV1S22*01	musIGHV622			0.06
IGHV2–4–1*01	musIGHV124			0.05
	musIGHV591			0.05
			balbcdGHV033	0.05
IGHV1S136*01	musIGHV689			0.05
		J558.34		0.04
			balbcdGHV016	0.04
			balbcdGHV034	0.04
	musIGHV577			0.03
IGHV1S75*01	musIGHV579			0.03
	musIGHV692			0.03
IGHV3–8*02	musIGHV128			0.02
IGHV5–9*02	musIGHV131			0.02
IGHV5–9*03; IGHV5–9–5*01	musIGHV161			0.02
IGHV1S132*01	musIGHV706			0.02
		J558.6		0.02
IGHV2–2–2*01	musIGHV173			0.02
		J558.27		0.01
IGHV2–4*02	musIGHV179			0.01
IGHV3–5*02	musIGHV437			0.01
				0.01
IGHV1S113*01	musIGHV681			0.01

gene displayed a preferential RF. In most but not all cases, this was RF3 (electronic supplementary material, figure S1).

IgBLAST did not consider the possible contributions to junctional diversity of multiple IGHD genes or the utilization of the IGHDs in inverted orientation, but a small number of examples of possible usage of multiple IGHDs were identified in both C57BL/6 and BALB/c VDJs. The 22 BALB/c and the 18 C57BL/6 sequences represented just 0.14% and 0.10% of total lineages, respectively (electronic supplementary material, table S2). The mean length of these 40 secondary IGHDs was 9.0 nt compared with 9.7 nt for the primary IGHD within the 40 VDJs. The maximum length for a secondary IGHD was 13 nt with no mismatches for BALB/c and 13 nt with a single mismatch for C57BL/6. Not all secondary IGHDs could be confirmed as permissible IGHD rearrangement events as not all IGHDs have had their genomic positions mapped within the locus.

Inverted IGHDs are IGHDs that are said to appear in rearrangements in an orientation opposite to their genomic

state. IgBLAST does not attempt to identify inverted IGHDs within VDJ rearrangements. Rearrangements were therefore re-evaluated to identify such IGHDs among rearrangements both with and without primary IgBLAST IGHD assignments. Inverted IGHDs were accepted only where a primary IGHD was absent or the inverted segment represented a longer and less mismatched segment when compared with the IgBLAST IGHD (electronic supplementary material, table S3). Forty-eight apparent inverted IGHDs were identified for BALB/c (0.32% of total lineages) and 80 for C57BL/6 (0.44%), of which all but one were identified in the absence of a primary IGHD being found by IgBLAST. The average length of apparent inverted IGHD in the absence of a primary IGHD alignment was just 8.8 nt.

The number of IgM- and IgG-associated sequences that lacked exonuclease removals from the IGHV was similar for both the C57BL/6 (35.5%) and BALB/c strains (35.0%). Among these sequences, 31.2% of C57BL/6 sequences and 40.3% of BALB/c sequences were associated with possible

Table 3. Functional IGHD genes present (✓) in the IMGT repertoire of the C57BL/6 and BALB/c strains, confirmation of their presence and functionality in the strains (✓) from analysis of VDJ rearrangements, and the average nucleotide length of each IGHD gene within the rearrangements.

	C57BL/6			BALB/c		
	IMGT	confirmed	average nt length ^a	IMGT	confirmed	average nt length ^a
IGHD1–1*01	✓	✓	14.6 (23)	✓	✓	14.4 (23)
IGHD1–2*01				✓	✓	11.6 (17)
IGHD1–3*01	✓	✗				
IGHD2–1*01				✓	✓ ^d	11.1 (17)
IGHD2–2*01				✓	✓ ^c	11.2 (17)
IGHD2–3*01	✓	✓	11.7 (17)	✓	✓	11.7 (17)
IGHD2–4*01	✓	✓	11.5 (17)	✓	✓	11.4 (17)
IGHD2–5*01	✓	✓ ^b	12.1 (17)			
IGHD2–6*01	✓	✓ ^b	12.1 (17)			
IGHD2–7*01	✓	✓ ^c	11.7 (17)	✓	✓ ^c	11.2 (17)
IGHD2–8*01	✓	✓ ^d	11.3 (17)	✓	✓ ^d	11.1 (17)
IGHD2–9*01				✓	✗	
IGHD2–10*01				✓	✓	14.7 ^e (17)
IGHD2–11*01				✓	✗	
IGHD2–14*01				✓	✓	11.4 (17)
IGHD3–1*01	✓	✗				
IGHD3–2*01				✓	✓	14.2 (16)
IGHD3–2*02	✓	✓	12.4 (17)			
IGHD4–1*01	✓	✓	9.1 (11)	✓	✓	9.0 (11)
IGHD4–1*02				✓	✗	

^aThe full length of each IGHD gene is in parenthesis.

^bIGHD2–5*01 and IGH2–6*01 are identical sequences, and cannot be distinguished in VDJ rearrangements.

^cIGHD2–2*01 and IGH2–7*01 are identical sequences, and cannot be distinguished in VDJ rearrangements.

^dIGHD2–1*01 and IGH2–8*01 are identical sequences, and cannot be distinguished in VDJ rearrangements.

^eIGHD2–10*01 shares all but its most 5' nt with IGH2–1*01. Short IGH2–10*01 genes are probably assigned to IGH2–1*01 by IgBLAST.

P nucleotides. Six C57BL/6 motifs ranged in length from 2 to 6 nt and seven BALB/c motifs ranged from 2 to 5 nt (electronic supplementary material, table S4). Three of the six C57BL/6 motifs and four of the seven BALB/c motifs were observed at frequencies that were significantly higher than can be accounted for by the chance occurrence of palindromic repeat motifs among N-nucleotide additions. The proportion of IGHDs without nucleotide trimming was 7.6% (5') and 27.4% (3') for C57BL/6 IGHDs. BALB/c IGHDs were untrimmed in 7.8% (5') and 30.7% (3') of sequences. Overabundant motifs that are likely to be P inclusions were associated with untrimmed 5' IGHD ends: 'aga' and 'ga' for BALB/c and five 'a'-rich motifs for C57BL/6. The same three overabundant motifs were observed for both strains at untrimmed 3' IGHD ends: 'g', 'gt' and 'gtag'; 12.3% (C57BL/6) and 19.8% (BALB/c) of sequences lacked IGHJ removals and possible P nucleotide inclusions ranging in length from 1 to 7 nt were associated with 31.6% of C57BL/6 IGHJ and 24.4% of untrimmed BALB/c IGHJ. This was largely because of the over-representation of 't' upstream from untrimmed IGHJs in both strains.

The extent of N addition within the V–D and D–J junctions was then investigated, with possible P nucleotides being included in the N regions. The results of this analysis

are shown as figure 1. Small but significant differences were seen between the strains. The N1 regions of the V–D junction had an average of 3.4 and 4.0 nt for the BALB/c and C57BL/6 strains, respectively. The N2 regions of the D–J junction had an average of 2.7 and 2.9 nt, respectively. To explore the extent of the resulting junctional diversity, the nucleotide sequences were translated and the amino acid sequences of the complementarity determining region 3 (CDR3) were compared between the two datasets of IgM-associated VDJ rearrangements. The CDR3 spans the VDJ junction region, and includes the 3' end of the IGHV gene and the 5' end of the IGHJ gene; 560 unique CDR3 sequences were seen that were shared by the two strains, representing 4.5% of the C57BL/6 CDR3s and 5.9% of the BALB/c CDR3s.

The extent of somatic mutation of IGHV genes within VDJ rearrangements was finally determined. The percentage of sequences associated with the different isotypes that were unmutated is shown in figure 2a and the average number of mutations in the sequences is shown in figure 2b. The percentage of unmutated IgG-associated IGHV genes was unexpectedly high, ranging from almost 20% (BALB/c IgG1) to over 40% (C57BL/6 IgG2C). The average number of mutations seen was correspondingly low, and ranged from 2.4 (IgG2A BALB/c) to 4.6 (C57BL/6 IgG1). In both strains

Table 4. Functional IGHJ genes present (✓) in the IMGT repertoire of the C57BL/6 and BALB/c strains, confirmation of their presence and functionality in the strains (✓) from analysis of VDJ rearrangements, and the rearrangement frequency of each gene.

	C57BL/6			BALB/c		
	IMGT	confirmed	frequency (%)	IMGT	confirmed	frequency (%)
IGHJ1*01				✓	✓	8.6
IGHJ1*03	✓	✓	15.8			
IGHJ2*01	✓	✓	32.5	✓	✓	25.3
IGHJ3*01	✓	✓	21.8	✓	✓	29.2
IGHJ4*01	✓	✓	29.7	✓	✓	36.8

Figure 1. The average nucleotide lengths of the N1 regions of the VD junctions and the N2 regions of the DJ junctions, as determined from datasets of IgM-associated VDJ rearrangements of BALB/c and C57BL/6 mice.

there was a weak, but significant, positive correlation between the maximum mutation level of a clone lineage and the total number of lineage members detected by repertoire sequencing: C57BL/6 0.2717 and BALB/c 0.2246 (both $p < 0.0001$, Kendall's Tau).

4. Discussion

A complete and accurate inventory of the germline immunoglobulin genes of a species is necessary if the expressed antibody repertoire is to be understood, and if somatic point mutations are to be correctly identified in rearranged V(D)J genes. The aim of this study was therefore to first infer the germline heavy chain variable region genes of the BALB/c and C57BL/6 mouse strains, and then to define the processes and biases that shape their VDJ repertoires. Unexpectedly, this showed that the IGHV gene repertoires of the two strains are almost entirely non-overlapping. This striking observation stands in contrast to observations in the human. Although human studies in recent years have highlighted substantial differences in the germline repertoires of different individuals [10,13,27], these studies also show many commonly shared germline sequences. Numerous IGHV genes have been seen in all individuals studied. The human genome usually includes 40–46 different IGHV genes, and in a study of 14 individuals, we reported that individuals appeared on average to be homozygous at around 40 of the loci [27]. Although it is now clear that deletion polymorphisms are relatively common in the human IGHV locus [13], and homozygosity is therefore probably not quite as high as we reported, it remains true that many IGHV sequences are carried at high frequency within the human population.

That so few IGHV genes are shared by the BALB/c and C57BL/6 mouse strains was surprising, but other differences between the immunoglobulin genes of the BALB/c and C57BL/6 mice have been known for decades. Constant region gene variation between strains was identified in the 1970s [35], leading to the development of the *Igh^a* and *Igh^b* nomenclature for immunoglobulin heavy chain constant region gene haplotypes. The *Igh^a* haplotype of the BALB/c, 129/Sv and many other mouse strains includes the *IGHG2A* gene but not the *IGHG2C* gene, while the *Igh^b* haplotype of the C57BL/6, NOD and SJL strains includes the *IGHG2C* gene but not the *IGHG2A* gene [36,37]. Southern blotting was used to explore *Igh* haplotype-associated strain differences within the IGHV locus. This suggested that the IGHV loci of strains that carry the *Igh^a* haplotype are similar to one another, and different to those of *Igh^b* strains [38]. Half of the IGHV locus of the 129/Sv strain has now been sequenced, and the IGHV genes of the 129/Sv strain are quite different to those of C57BL/6 mice [22]. This study now confirms that most reported 129/Sv IGHV sequences are present in the BALB/c strain (data not shown). The partial IGHV locus map for the 129/Sv strain, reported by Retter *et al.* [22], describes IGHV genes belonging to 12 of the 15 IGHV gene families. It does not report genes of the very large J558/IGHV1 gene family, or genes of the IGHV8 and IGHV10 families. The 164 BALB/c IGHV genes identified in this study include 54 sequences reported in the 129/Sv strain [22].

In addition to its clarification of the genes of the murine heavy chain gene locus, this transcriptome-based study also clarifies reports of the functionality of those genes. Riblet [19] reported most but not all of the IGHV locus of the C57BL/6 strain and identified 69 pseudogenes and 101 apparently functional genes, suggesting a further 20–30 genes and pseudogenes remained to be identified at the 5' end of the locus. Functionality was based upon the presence or absence of functional RSS and other control elements. The IMGT germline IGHV repertoire is based upon the Riblet sequence, and upon unpublished sequences from the 5' end of the locus. It includes 14 additional genes that are identified as functional. The functionality of the IMGT genes is based upon an independent assessment by the IMGT group of the control elements associated with each gene. In total, IMGT recognizes 113 functional genes in the C57BL/6 genome. A separate assembly of the complete sequence of the locus identified 110 functional genes [20] which largely overlap with the IMGT repertoire. A small number of genes are absent from one or other repertoire. The present study, on

Figure 2. (a) The percentage of unmutated IGHV genes and (b) the average number of mutations in IGHV genes that were seen in association with different isotypes in VDJ rearrangements of BALB/c and C57BL/6 mice.

the other hand, identified just 99 unique, functional C57BL/6 IGHV sequences. It also identified 164 functional putative BALB/c IGHV sequences. The utilization frequencies of some of these genes and putative genes were low, and it is possible that additional functional IGHV genes with very low utilization frequencies are present within the genomes of these strains. It is unlikely, however, that any gene that makes a significant contribution to the murine repertoire was overlooked.

The mouse genome carries a relatively high number of IGHV genes, but in comparison to the human repertoire, the murine VDJ repertoire is constrained by the small number of available IGHD genes, and by similarities between those genes. The nine C57BL/6 IGHD genes include six genes of the IGHD2 family that differ from one another by no more than 2 nt. The BALB/c repertoire of IGHD genes is only slightly more diverse, having 12 genes of which eight are genes of the IGHD2 family. Four of the IGHD2 genes, as well as IGHD1-1*01 and IGHD4-1*01, are carried by both mouse strains.

The diversity of the murine repertoire is not expanded by the use of unconventional IGHD rearrangements. Rearrangements containing apparent inverted IGHD genes and D-D fusions were observed at extremely low frequencies within the productive repertoires of both mouse strains in this study. This is consistent with observations made of the human repertoire. If inverted IGHD genes and D-D fusions make any contribution to human repertoire diversity, it is an insignificant one [39].

The diversity of the murine repertoire may also be restricted by biases in the use of the different IGHD RFs. Both strains use IGHDs in each RF and the overall RF preference across all IGHDs is essentially identical between the strains. In most cases, RF3 is dominant, and the use of the alternative RFs is particularly rare for the IGHD2-family genes. This RF bias is therefore much stronger than biases seen in the human, where there may be dominant RFs but each RF of each gene is relatively commonly used [40].

The VD and DJ junctions are regions of variability, but because of the nature of P addition, and because of the paucity of N addition in mouse VDJ rearrangement, the CDR3 regions of mouse heavy chain genes are largely germline-derived. Germline-encoded junctional diversification comes from the presence of P nucleotide inclusions in murine sequences, with P nucleotides being observed at much higher frequencies than was previously reported from analysis of human VDJ sequences [41]. By contrast, the diversity of murine junction regions is constrained by low levels of N nucleotide addition. The 3 or 4 nt that are typically added to the murine VD and DJ junctions contrast with the 7.7 and 6.5 nt that are added on average to human VD and DJ junctions [34]. The short N regions that are typical in the mouse often fail to introduce non-template encoded amino acids. Many short N regions simply contribute nucleotides for the completion of codons that are partially template encoded, and the diversity of the resulting amino acids is constrained by the genetic code. Redundancy in the genetic code

means that this is particularly true in the case of amino acids found at the 3' ends of the IGHV and IGHD genes.

Given the lack of N addition in each strain, and the shared IGHD genes, it is unsurprising that identical amino acid junctions were seen when VDJ genes from both strains were translated. Shared CDR3 regions represent 4–6% of all clones in the datasets, which is a frequency that is far higher than that seen in humans. When thousands of human IgM sequences from different individuals are similarly compared, including between monozygotic twins, it is unusual to identify even a single shared CDR3 region [42]. This difference between humans and mice can be viewed as a consequence of the greater germline focus and therefore smaller size of the murine repertoire.

The diversity of human sequences in the naive B-cell repertoire is expanded considerably during an immune response through the process of somatic point mutation. Human IgG-associated IGHV genes accumulate mutations at frequencies that vary between IgG subclasses, and average mutation numbers range from 16.5 (IgG3) to 21.8 (IgG4) [23]. Very few murine sequences accumulate so many mutations, and the percentage of germline (unmutated) IgG-associated IGHV genes in this study ranged between 19.4% (IgG1 BALB/c) and 43.6% (IgG2C C57BL/6). Human IgG3-associated IGHV genes have the highest percentage of germline sequences, with just 5.7% of sequences being unmutated [23]. Not a single germline IGHV gene was detected among 288 human IgG4 sequences [23]. This difference between humans and laboratory mice also suggests a greater germline focus of the murine repertoire.

The view that the heavy chain repertoire is germline-focused seems at first to be at odds with our observation of the divergence of the IGHV loci of the BALB/c and C57BL/6 strains. In fact, differences between the two strains could be in accord with this view if the divergence can be explained by the breeding histories of the strains. It has recently been shown that the genomes of the common inbred strains of laboratory mice are predominantly derived from the *Mus musculus domesticus* subspecies of the house mouse, which is found in western Europe and the Americas [43]. As much as 10% of the genomes of different strains are derived from the *M. m. musculus* subspecies, which is found from eastern Europe to North China. Smaller components of the genomes appear to be derived from *M. m. castaneus*, which is found in southern China, Japan and southeast Asia [43]. This genetic mosaic is likely the result of mating

M. m. domesticus mice with *M. m. molossinus* mice, which are natural hybrids of *M. m. castaneus* and *M. m. musculus*. *Mus m. molossinus* mice were available to the American developers of the inbred strains as 'Japanese fancy mice' [43].

It is possible that the IGHV genes of the BALB/c and C57BL/6 strains are derived from different mouse subspecies, and that the loci diverged through evolution as a consequence of the different selection pressures that acted in the different geographical ranges of the subspecies. Evolution may well have led to different sets of IGHV genes that provide protective antibodies against major pathogens in their germline configuration. Because of its size, a mouse will be rapidly killed by any serious infection. It requires antibody-mediated protection that is fast, yet suitably specific for invading pathogens. The germline focus of the murine antibody heavy chain repertoire could help ensure such protection, if the sets of murine germline genes have evolved under strong selection pressure.

Future studies will need to confirm inferred BALB/c genes by genomic sequencing of unrearranged genes. Studies will also need to confirm that observations reported here from the study of laboratory mice are to be seen in wild mice. It will also be important to determine how well the IGHV loci of inbred laboratory strains reflect the loci of wild mice, for it has been suggested that the inbreeding of the BALB/c strain may have led to a loss of sequence diversity within the lambda light chain IGLV gene locus [44,45]. It is conceivable that there has also been a loss of diversity within the heavy chain locus, as a result of inbreeding. Additional studies in other species will then be required to determine whether or not the germline focus of the heavy chain gene repertoire of laboratory strains of mice is a general feature of the humoral immune systems of small mammals.

Ethics. This study was conducted with the approval of the University of New South Wales Animal Care and Ethics Committee (Approval no. 13/98B).

Data accessibility. Data are available via the European Nucleotide Archive at <http://www.ebi.ac.uk/ena/data/view/PRJEB8745>.

Authors' contributions. A.C. designed the project and prepared the manuscript, C.M. and Y.W. conducted the laboratory work and K.J. and K.R. performed the bioinformatics analysis.

Competing interests. The authors declare they have no competing interests.

Funding. This work was supported by a grant from the National Health and Medical Research Council.

References

- Schroeder Jr HW, Cavacini L. 2010 Structure and function of immunoglobulins. *J. Allergy Clin. Immunol.* **125**, S41–S52. (doi:10.1016/j.jaci.2009.09.046)
- Tonegawa S. 1983 Somatic generation of antibody diversity. *Nature* **302**, 575–581. (doi:10.1038/302575a0)
- Alt FW, Baltimore D. 1982 Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-JH fusions. *Proc. Natl Acad. Sci. USA* **79**, 4118–4122. (doi:10.1073/pnas.79.13.4118)
- Jackson KJ, Kidd MJ, Wang Y, Collins AM. 2013 The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor. *Front. Immunol.* **4**, 263. (doi:10.3389/fimmu.2013.00263)
- Atkinson MJ, Cowan MJ, Feeney AJ. 1996 New alleles of IGKV genes A2 and A18 suggest significant human IGKV locus polymorphism. *Immunogenetics* **44**, 115–120. (doi:10.1007/s002510050098)
- Granoff DM, Shackelford PG, Holmes SJ, Lucas AH. 1993 Variable region expression in the antibody responses of infants vaccinated with *Haemophilus influenzae* type B polysaccharide-protein conjugates. Description of a new lambda light chain-associated idiotype and the relation between idiotype expression, avidity, and vaccine formulation. The Collaborative Vaccine Study Group. *J. Clin. Invest.* **91**, 788–796. (doi:10.1172/JCI116298)
- Nadel B *et al.* 1998 Decreased frequency of rearrangement due to the synergistic effect of nucleotide changes in the heptamer and nonamer of the recombination signal sequence of the V kappa gene A2b, which is associated with increased susceptibility of Navajos to *Haemophilus*

- influenzae* type b disease. *J. Immunol.* **161**, 6068–6073.
8. Liu L, Lucas AH. 2003 IGH V3–23*01 and its allele V3–23*03 differ in their capacity to form the canonical human antibody combining site specific for the capsular polysaccharide of *Haemophilus influenzae* type B. *Immunogenetics* **55**, 336–338. (doi:10.1007/s00251-003-0583-8)
 9. Pappas L *et al.* 2014 Rapid development of broadly influenza neutralizing antibodies through redundant mutations. *Nature* **516**, 418–422. (doi:10.1038/nature13764)
 10. Wang Y, Jackson KJ, Gaeta B, Pomat W, Siba P, Sewell WA, Collins AM. 2011 Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. *Immunogenetics* **63**, 259–265. (doi:10.1007/s00251-010-0510-8)
 11. Watson CT *et al.* 2013 Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am. J. Hum. Genet.* **92**, 530–546. (doi:10.1016/j.ajhg.2013.03.004)
 12. Scheepers C *et al.* 2015 Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire. *J. Immunol.* **194**, 4371–4378. (doi:10.4049/jimmunol.1500118)
 13. Kidd MJ *et al.* 2012 The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. *J. Immunol.* **188**, 1333–1340. (doi:10.4049/jimmunol.1102097)
 14. Houpt ER *et al.* 2002 The mouse model of amebic colitis reveals mouse strain susceptibility to infection and exacerbation of disease by CD4⁺ T cells. *J. Immunol.* **169**, 4496–4503. (doi:10.4049/jimmunol.169.8.4496)
 15. Breitbach K, Wongprompitak P, Steinmetz I. 2011 Distinct roles for nitric oxide in resistant C57BL/6 and susceptible BALB/c mice to control *Burkholderia pseudomallei* infection. *BMC Immunol.* **12**, 20. (doi:10.1186/1471-2172-12-20)
 16. Peltz G *et al.* 2011 Next-generation computational genetic analysis: multiple complement alleles control survival after *Candida albicans* infection. *Infect. Immunity* **79**, 4472–4479. (doi:10.1128/IAI.05666-11)
 17. Marino E, Grey ST. 2012 B cells as effectors and regulators of autoimmunity. *Autoimmunity* **45**, 377–387. (doi:10.3109/08916934.2012.665527)
 18. Potter M. 1977 Antigen-binding myeloma proteins of mice. *Adv. Immunol.* **25**, 141–211.
 19. Riblet R. 2003 Immunoglobulin heavy chain genes in the mouse. In *Molecular biology of B cells* (eds T Honjo, FW Alt, M Neuberger), pp. 19–26. London, UK: Elsevier Academic Press.
 20. Johnston CM *et al.* 2006 Complete sequence assembly and characterization of the C57BL/6 mouse Ig heavy chain V region. *J. Immunol.* **176**, 4221–4234. (doi:10.4049/jimmunol.176.7.4221)
 21. Lefranc MP *et al.* 2009 IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* **37**, D1006–D1012. (doi:10.1093/nar/gkn838)
 22. Retter I *et al.* 2007 Sequence and characterization of the Ig heavy chain constant and partial variable region of the mouse strain 129S1. *J. Immunol.* **179**, 2419–2427. (doi:10.4049/jimmunol.179.4.2419)
 23. Jackson KJ, Wang Y, Collins AM. 2014 Human immunoglobulin classes and subclasses show variability in VDJ gene mutation levels. *Immunol. Cell Biol.* **92**, 729–733. (doi:10.1038/icb.2014.44)
 24. Ye J, Ma N, Madden TL, Ostell JM. 2013 IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34–W40. (doi:10.1093/nar/gkt382)
 25. Edgar RC. 2010 Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461. (doi:10.1093/bioinformatics/btq461)
 26. Greiff V, Menzel U, Haessler U, Cook SC, Friedensohn S, Khan TA, Pogson M, Hellmann I, Reddy ST. 2014 Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunol.* **15**, 40. (doi:10.1186/s12865-014-0040-5)
 27. Boyd SD *et al.* 2010 Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J. Immunol.* **184**, 6986–6992. (doi:10.4049/jimmunol.1000445)
 28. Retter I *et al.* 2005 VBASE2, an integrative V gene database. *Nucleic Acids Res.* **33**, D671–D674. (doi:10.1093/nar/gki088)
 29. Meier JT, Lewis SM. 1993 P nucleotides in V(D)J recombination: a fine-structure analysis. *Mol. Cell Biol.* **13**, 1078–1092.
 30. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. 2015 Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc. Natl Acad. Sci. USA* **112**, E862–E870. (doi:10.1073/pnas.1417683112)
 31. Haines BB, Angeles CV, Parmelee AP, McLean PA, Brodeur PH. 2001 Germline diversity of the expressed BALB/c VhJ558 gene family. *Mol. Immunol.* **38**, 9–18. (doi:10.1016/S0161-5890(01)00049-9)
 32. Ye J. 2004 The immunoglobulin IGHD gene locus in C57BL/6 mice. *Immunogenetics* **56**, 399–404. (doi:10.1007/s00251-004-0712-z)
 33. Feeney AJ, Riblet R. 1993 DST4: a new, and probably the last, functional DH gene in the BALB/c mouse. *Immunogenetics* **37**, 217–221. (doi:10.1007/BF00191888)
 34. Jackson KJ, Gaeta BA, Collins AM. 2007 Identifying highly mutated IGHD genes in the junctions of rearranged human immunoglobulin heavy chain genes. *J. Immunol. Methods* **324**, 26–37. (doi:10.1016/j.jim.2007.04.011)
 35. Blomberg B, Geckeler WR, Weigert M. 1972 Genetics of the antibody response to dextran in mice. *Science* **177**, 178–180. (doi:10.1126/science.177.4044.178)
 36. Jouvin-Marche E *et al.* 1989 The mouse *Igh-1a* and *Igh-1b* H chain constant regions are derived from two distinct isotypic genes. *Immunogenetics* **29**, 92–97. (doi:10.1007/BF00395856)
 37. Morgado MG *et al.* 1989 Further evidence that BALB/c and C57BL/6 gamma 2a genes originate from two distinct isotypes. *EMBO J.* **8**, 3245–3251.
 38. Brodeur PH, Riblet R. 1984 The immunoglobulin heavy-chain variable region (Igh-V) locus in the mouse. I. One hundred Igh-V genes comprise seven families of homologous genes. *Eur. J. Immunol.* **14**, 2385–2388. (doi:10.1002/eji.1830141012)
 39. Collins AM, Ikutani M, Puiu D, Buck GA, Nadkarni A, Gaeta B. 2004 Partitioning of rearranged Ig genes by mutation analysis demonstrates D-D fusion and V gene replacement in the expressed human repertoire. *J. Immunol.* **172**, 340–348. (doi:10.4049/jimmunol.172.1.340)
 40. Larimore K, McCormick MW, Robins HS, Greenberg PD. 2012 Shaping of human germline IgH repertoires revealed by deep sequencing. *J. Immunol.* **189**, 3221–3230. (doi:10.4049/jimmunol.1201303)
 41. Jackson KJ, Gaeta B, Sewell W, Collins AM. 2004 Exonuclease activity and P nucleotide addition in the generation of the expressed immunoglobulin repertoire. *BMC Immunol.* **5**, 19. (doi:10.1186/1471-2172-5-19)
 42. Glanville J *et al.* 2011 Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl Acad. Sci. USA* **108**, 20 066–20 071. (doi:10.1073/pnas.1107498108)
 43. Yang H *et al.* 2011 Subspecific origin and haplotype diversity in the laboratory mouse. *Nat. Genet.* **43**, 648–655. (doi:10.1038/ng.847)
 44. Scott CL, Potter M. 1984 Polymorphism of C lambda genes and units of duplication in the genus *Mus*. *J. Immunol.* **132**, 2630–2637.
 45. Reidl LS, Kinoshita CM, Steiner LA. 1992 Wild mice express an Ig V lambda gene that differs from any V lambda in BALB/c but resembles a human V lambda subgroup. *J. Immunol.* **149**, 471–480.