TITLE: Germline murine immunoglobulin IGHV genes in wild-derived and classical inbred strains: a comparison

Short running title: Germline IGHV genes in inbred mice

Corey T. Watson[1]*, Justin T. Kos[1], William S. Gibson[1], Christian E. Busse[2], Leah Newman[3,4], Gintaras Deikus[3,4], Melissa Laird Smith[3,4], Katherine J.L. Jackson[5], Andrew M. Collins[6]*

[1]Department of Biochemistry and Molecular Genetics, University of Louisville School of Medicine, Louisville, KY USA 40202; [2]Division of B Cell Immunology, German Cancer Research Center, 69120 Heidelberg, Germany; [3]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029; [4]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029; [5]Immunology Division, Garvan Institute of Medical Research, Darlinghurst, 2010 New South Wales, Australia; [6]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, 2052 New South Wales, Australia

*To whom correspondence should be addressed.

Corey T. Watson: corey.watson@louisville.edu

Andrew Collins: a.collins@unsw.edu.au

ACKNOWLEDGMENTS

ABSTRACT

To better understand the subspecies origin of antibody genes in classical inbred mouse strains, the IGH gene loci of four wild-derived mouse strains were explored by analysis of VDJ gene rearrangements. A total of 341 unique IGHV gene sequences were inferred in the wild-derived strains, including 247 sequences that have not previously been reported. The genes of the Non-Obese Diabetic (NOD) strain were also documented, and all but one of the 84 inferred NOD IGHV genes have previously been observed in C57BL/6 mice. This is surprising because the Swiss mouse-derived NOD strain and the C57BL/6 strain have no known shared ancestry. The relationships between the genes of the wild-derived inbred strains and of the C57BL/6, NOD and BALB/c classical inbred strain were then explored. The IGH loci of the C57BL/6 and the MSM/MsJ strains share many sequences, but analysis showed that few sequences are shared with wild-derived strains representing the three major subspecies of the house mouse. There were also few IGHV sequences that were shared by the BALB/c strain and any of the four wild-derived strains. The origins of IGHV genes in the C57BL/6, MSM/MsJ and BALB/c strains therefore remain unclear. These unexpected similarities

and differences highlight our lack of understanding of the antibody gene loci of the laboratory mouse, with implications for the interpretation of strain-specific differences in models of antibody-mediated diseases, and of Adaptive Immune Receptor Repertoire sequencing (AIRR-seq) data. These results also suggest that a position-based immunoglobulin gene nomenclature may be unworkable in the mouse.

Keywords: AIRR-seq; mouse immunoglobulin; IGHV; wild-derived; Non-Obese Diabetic

INTRODUCTION

Inbred mouse strains are critical to biomedical research, and many of the most important strains such as DBA, C57BL, C3H, CBA and BALB/c have now been in use for almost a century [1]. The C57BL and BALB/c strains have been particularly important for our understanding of the biochemistry and immunogenetics of immunoglobulins (IG) [2, 3, 4, 5]. This understanding was achieved despite a lack of detailed knowledge of the antibody genes of these and other inbred laboratory mouse strains, and until recently it was thought that the genes present in these different strains were likely to be highly similar [6].

Mouse antibody genes were first identified using cell lines derived from BALB/c mice because of the availability of mineral-oil induced plasmacytomas from this strain [7]. The cataloguing of BALB/c antibody genes effectively ceased with the emergence of the C57BL/6 strain as the workhorse of transgenic and genomic studies. The IG heavy chain variable region (IGHV) gene locus of the C57BL/6 strain was therefore the first to be sequenced and annotated [8, 9]. Part of the IGH locus of the 129S1 strain was also reported [10], before comprehensive genomic investigation of mouse germline IGHV genes essentially ceased. By this time, two databases had catalogued mouse IGHV genes and apparent allelic variants of these genes: the VBASE2 [11] and IMGT [12] databases. A positional nomenclature was then developed by IMGT, based upon the mouse genome reference sequence, while an alternative positional nomenclature was developed by Johnston and colleagues, based upon an alternative assembly of the

C57BL/6 genome [9]. A non-positional gene sequence identifier system was also developed by VBASE2 [11].

The study of IGHV gene variation in humans followed a similar trajectory to that of the mouse. Genes and likely allelic variants were reported over a twenty year period, starting in the late 1970s [13]. There was a sharp decline in the reporting of new sequences once the complete human IGH locus was published in 1998 [14], but the advent of high-throughput sequencing of human antibody genes reawakened interest in the documentation of allelic variants [15, 16]. A surprising level of antibody gene variation, including structural variation of the IGH locus, has since been shown within the human population [17, 18, 19, 20], and such variation can have important consequences for the development of a suitable protective antibody repertoire [21, 22].

Many recently discovered allelic variants have been identified from sets of VDJ gene rearrangements, using a process of inference. Rearranged VDJ sequences are often affected by somatic hypermutations, and such mutations are distributed throughout VDJ rearrangements. When the same mismatch to a known germline gene is repeatedly observed in data from a single subject, however, it is more likely that the nucleotide in question is a single nucleotide polymorphism (SNP), rather than being a nucleotide that has arisen by somatic hypermutation [23]. When such mismatches are repeatedly seen in a large set of VDJ rearrangements involving multiple IGHJ genes, having diverse CDR3 regions and amplified from a single individual, the inference of a previously undiscovered gene polymorphism may be made with confidence. The discovery of

allelic variants by inference is now a feature of most human repertoire studies, and this is facilitated by a number of recently developed utilities [24, 25, 26, 27] (https://arxiv.org/abs/1711.05843). When this approach was applied to the mouse, the outcome was quite unexpected.

Analysis of thousands of C57BL/6 VDJ rearrangements identified 99 of the 115 germline IGHV genes that have been reported to be functional in this strain [6]. It was concluded that the remaining 16 genes are either non-functional or are expressed at such low frequencies that they would only be detectable in larger studies. It was also concluded that most of the IGHV genes carried by any strain of inbred mice should be readily determinable by inference from VDJ gene datasets.

An analysis of BALB/c VDJ rearrangements was then performed, and 163 BALB/c IGHV gene sequences were identified [6]. Only half of the identified BALB/c genes were present in the IMGT database [12]. The differences between the BALB/c and the C57BL/6 strain were so profound that it was proposed that the IGH loci of these strains may have had their origins in different subspecies of the house mouse, for three major subspecies of the house mouse have been described: *Mus musculus musculus*, *M. m. domesticus* and *M. m. castaneus*. This hypothesis was subsequently supported by the analysis of genome-wide SNP data [28] (see Supplementary Figure 1), and to test the hypothesis, we have now investigated the IGHV gene repertoires of a number of wild-derived inbred mouse strains. These strains were developed in the 1970s from pairs of wild mice from known locations, with the intention that each inbred strain would carry a genome that

was derived from a single subspecies of the house mouse. To explore the differences between strains that appear, based upon preliminary SNP analysis (Supplementary Figure 1), to be derived from the same subspecies of the house mouse, we also investigated the NOD/ShiLtJ inbred strain.

The divergence that was seen between the strains in this study suggests that the IGH loci of inbred mouse strains are likely to harbour so much genetic variation that if the antibody responses of these mouse strains are to be properly understood, it will be essential to document this diversity. The divergence between the strains also suggests that a single positional mouse IGHV gene nomenclature based upon the C57BL/6 genome reference sequence may fail to properly represent the genes of many important inbred mouse strains.

METHODS

Antibody gene repertoire sequencing

Whole dissected spleens, preserved in RNAlater, were obtained from female mice from Jackson Laboratories (https://www.jax.org) for five inbred strains (CAST/EiJ [JAX stock #000928], n=1; LEWES/EiJ [JAX stock #002798], n=1; PWD/PhJ [JAX stock #004660], n=1; MSM/MsJ [JAX stock #003719], n=1; NOD/ShiLtJ [JAX stock #001976], n=1). Total RNA was extracted from a section of each spleen using the RNeasy Mini kit (Qiagen, Cat No. 74104). For each strain, 5' RACE first-strand cDNA synthesis was conducted using the SMARTer RACE cDNA Amplification Kit (Clontech), with an input amount of 1 µg of RNA per sample. Rearranged VDJ-IgM amplicons were generated

using a single IgM oligo positioned in the CH3 region of the mouse IGHM gene (5'-CAGATCCCTGTGAGTCACAGTACAC-3'; 10 µM), paired with a universal primer (Clontech; 5'-AAGCAGTGGTATCAACGCAGAGT-3'; 10 µM). First-strand cDNA for each strain was amplified using Thermo Fisher Phusion HF Buffer (Cat No. F530S) for 30 PCR cycles. Amplicons were run on 2% agarose gel for size confirmation. Final amplicons were used to generate SMRTbell template libraries and each library was sequenced across 2 SMRTcells on a Pacific Biosciences RSII system using P6/C4 chemistry and 240 minute movies.

Data processing and germline gene inference

Reads from each RSII run were combined and processed using the circular consensus sequencing algorithm (CCS2). CCS2-processed reads of Q30 or greater were processed with pRESTO v0.5.1 as follows: 1) Reads without a universal primer were removed using a maximum primer match error rate of 0.2 and a maximum template-switch error rate of 0.5 to de-duplicate the reads; 2) Reads without an IgM primer were removed using a maximum primer match error rate of 0.2 and a maximum template-switch error rate of 0.5; 3) Duplicate sequences from the FASTQ files were removed using the default value of a maximum of 20 ambiguous nucleotides. Following pRESTO [29] processing, IGHV gene assignments were noted after mapping reads to the IMGT germline database using IMGT/HighV-QUEST version 1.6.0 (20 June 2018). Resulting IMGT summary output was processed using Change-O v0.4.3 [30]. Reads representing incomplete or non-productive VDJ rearrangements were discarded. All sequences in each sample were assigned to clones using the distToNearest function in the SHazaM

R package and the DefineClones function in Change-O [30]. A single member of each clonal group was used for downstream analyses.

For germline gene inference for each strain, the sequences representing the strain's clonal groups were first clustered based on the IGHV gene assignment and associated percent identity of the alignment to the nearest IMGT reference directory gene sequence. Sequences shorter than the 138 base pair (bp) length of the shortest reported functional IGHV sequence in IMGT were excluded from the analysis. Consensus IGHV gene sequences were then determined for each cluster using CD-HIT (cd-hit-est v4.6.8) [31] requiring that a given inference be represented by at least 0.1% of total clonal groups identified in the strain's dataset. To be conservative, consensus sequences for each inferred germline were trimmed to the shortest sequence representative in the identified cluster. If the percent match identity to the closest IMGT germline gene sequence was <95%, then the alignments were manually inspected for evidence of chimeric PCR amplification [32].

An inferred germline IGHV sequence reference dataset was generated for each strain. Predicted germline IGHV gene databases from Q30 datasets were assessed using one iteration of IgDiscover v0.10 [25]. For the analysis of each strain, inferred germline IGHV sequences generated from our clustering method were used as the IGHV gene database for IgDiscover. Changes made to the default configuration were as follows: 1) 'race_g' set to 'true' to account for a run of G nucleotides at the beginning of the sequence 2) 'stranded' set to 'false' since forward primer was not always at 5' end 3)

'ignore_j' set to 'true' to ignore whether or not a joining (J) gene had been assigned for a newly discovered IGHV gene.

All inferred germline sequences were compared, using BLAST (https://blast.ncbi.nlm.nih.gov/) [33], to sequences in the following databases: IMGT (http://www.imgt.org), VBASE2 (http://www.vbase2.org/), and the NCBI non-redundant nucleotide sequence collection. Germline sets were compared across strains using BLAT [34]. Perfect matches between sequences of different strains required full length alignments of 100% identity. Calculations of mean IGHV sequence identities between strains were computed by taking the mean of the sequence identities for the best pair-wise hits of all genes in the smaller germline set of two strains being compared. For the comparison between IGHV sequences of NOD/ShiLtJ and MSM/MsJ, which had an equal number of inferred germline genes, we opted to present the mean sequence identity using the NOD/ShiLtJ germline set.

RESULTS

Defining inferred IGH germline gene sets in diverse inbred mouse strains

To ensure the highest quality input data, prior to VDJ assignment, we leveraged the PacBio CCS2 algorithm to generate high quality circular consensus reads for each sample. The average read length across the libraries sequenced was 22.5 Kb (Supplementary Table I). The long read lengths paired with the target amplicon library size of ~1200 bp, resulted in a mean of 23.9 circular consensus passes per amplicon

(Supplementary Table I). Finally, we applied a Q30 cutoff to data from each library, resulting in a total of 36782, 43522, 43173, 28136 and 43044 pre-mapped reads for CAST/EiJ, LEWES/EiJ, MSM/MsJ, PWD/PhJ, and NOD/ShiLtJ, respectively, with mean CCS read scores of 1 (Supplementary Table I). Each of these high quality read datasets were used for gene and clonal assignment, and IGHV germline inference. Read data for each strain have been submitted to the Sequence Read Archive (SRA) under the BioProject ID PRJNA533312.

We took an integrated approach for defining inferred sets of germline sequences from each strain (see Materials and Methods). We modeled the initial discovery stage of our inference pipeline after that used previously [6], followed by confirmatory analysis using IgDiscover [25]. Clonal assignment and clustering resulted in a total of 4948 (CAST/EiJ), 4894 (LEWES/EiJ), 4149 (MSM/MsJ), 3722 (PWD/PhJ), and 3525 (NOD/ShiLtJ) unique clones; a single representative sequence from each of the identified clones was then used for germline inference. With these sequences as input, the first stage of our approach yielded a total of 87, 78, 84, 92, and 84 putative inferred germline sequences for CAST/EiJ, LEWES/EiJ, and MSM/MsJ, PWD/PhJ, and NOD/ShiLtJ, respectively (Figure 1A; Supplementary Table II). Each inferred germline sequence was represented by at least 0.1% of the total clones observed in a given strain; the numbers of clones representing each inference are provided in Supplementary Table II and plotted in Supplementary Figure 2. These initial germline sets were then used as starting germline databases in IgDiscover [25] to gain further support for each inference using a secondary method. In summary, IgDiscover provided additional support of 82/87 (94%), 68/78

(87%), 76/84 (90%), 86/92 (93%), and 84/84 (100%) inferred sequences for CAST/EiJ, LEWES/EiJ, and MSM/MsJ, PWD/PhJ, and NOD/ShiLtJ, respectively. Inferences in the initial germline discovery sets that were not subsequently confirmed by IgDiscover tended to be those represented by fewer clones on average (Supplementary Table II and Supplementary Figure 3). The validity of many of these genes was however bolstered by secondary evidence from public databases (Supplementary Table II). Of special note, all but one of the NOD/ShiLtJ genes have previously been reported as C57BL/6 sequences.

The majority of inferred germlines from wild-derived strains are not curated in IMGT

To enable broad comparisons between strains, we opted to use the full set of inferred genes from each strain, rather than restricting downstream analyses to only those supported by IgDiscover [25]. Across strains, we observed varying numbers of exact matches between inferred germline sequences and those found in the mouse IMGT reference directory (http://www.imgt.org) (Figure 1A). This included 22 non-IMGT sequences in MSM/MsJ, 63 in LEWES/EiJ, 80 in PWD/PhJ and 82 in CAST/EiJ. Alleles inferred from the wild-derived strains were therefore dominated by non-IMGT alleles (247/341, 72%; Figure 1A). For many of these (n=44), additional supporting evidence was found in other public sequence repositories (Supplementary Table II). Two non-IMGT sequences were found in the MSM/MsJ strain with perfect matches to sequences reported in either VBASE2 (http://www.vbase2.org/) or the NCBI reference set. This was also true for 2, 29, and 11 non-IMGT sequences inferred from CAST/EiJ, LEWES/EiJ,

and PWD/PhJ, respectively. In some cases, novel sequences from wild-derived strains were quite divergent from published IGHV sequences, varying from 99.66% sequence identity to 87.02%. This was also dependent on strain in that, among the wild-derived strains, inferred germlines from CAST/EiJ, LEWES/EiJ, and PWD/PhJ exhibited much greater sequence divergence from IMGT alleles than sequences inferred in MSM/MsJ (Figure 1B).

The count of genes within the different IGHV families were generally comparable across the five strains (Figure 1C), with some exceptions. The germline repertoires of all strains were clearly dominated by the IGHV1 family. However, the numbers of IGHV genes in other subgroups were more variable. For example, the repertoire of CAST/EiJ harbored fewer IGHV2 genes relative to the other strains, but greater numbers of IGHV3, IGHV6, IGHV9, and IGHV14 genes. At least one representative germline sequence of subfamilies IGHV1-IGHV3, IGHV5-IGHV10, and IGHV14 was inferred from all strains; in contrast, sequences of the remaining subfamilies, which are all small subfamilies in the C57BL/6 strain, were absent in at least one strain. IGHV13 and IGHV15 sequences were only observed in two of the five strains. Whether this represents a genuine lack of functional IGHV subfamily sequences in these strains (e.g., as a result of pseudogenization or genomic deletion), or whether this is due to undersampling in these repertoires is not clear. Deeper sequencing across additional individuals in each strain will be needed to fully assess this. Consistent with the general subfamily distributions observed, the majority of non-IMGT sequences in CAST/EiJ (n=32), LEWES/EiJ (n=32), and PWD/PhJ (n=39) represented IGHV1 genes (Figure 1D). In contrast, however,

although the MSM/MsJ repertoire was also dominated by IGHV1, the majority of non-IMGT sequences identified in that strain were from IGHV2 (n=9) and IGHV5 (n=9) (Figure 1D).

Extensive IGHV germline diversity and limited overlap between strains

We next investigated the extent of overlap of IGHV sequences among the surveyed strains. The sets of germline genes inferred from each inbred strain were compared to sequences identified in all other strains to determine how many sequences were identical between strains. Comparisons were additionally made with previously published inferences from BALB/c mice (n=163) [6], and with the IMGT repertoire of functional C57BL/6 sequences (n=114). Surprisingly little overlap was observed between strains, and the majority of inferred germline sequences were unique to a single strain (Figure 2A). Among the wild-derived strains surveyed, CAST/EiJ had the highest number of unique germline sequences (76/87 sequences). On the other hand, 83 sequences were observed in both NOD/ShiLtJ and C57BL/6, with 48 of these 83 sequences being additionally shared by MSM/MsJ. A single sequence was identified in five different strains (LEWES/EiJ, NOD/ShiLtJ, CAST/EiJ, PWD/PhJ, and C57BL/6).

We further explored interstrain IGHV sequence relationships by estimating the average sequence similarities of IGHV sets between strains. Consistent with sequence overlaps presented in Figure 2A, we noted a range of mean pairwise sequence identities, depending on the strains in question. For example, sequences in MSM/MsJ, C57BL/6, and NOD/ShiLtJ, strains which share the most identical sequences with one another

(Figure 2A), also have high average pairwise sequence identities (>99%; Figure 2B; see also Supplementary Figure 4 for full pairwise comparisons). This is in contrast to mean identities observed for all other pairwise strain comparisons, which ranged from 94.8% to 97.1%. These levels of identity also generally matched the number of shared sequences between strains (see Figure 2A). For example, LEWES/EiJ shared the most identical sequences with BALB/c, and among all pairwise sequence comparisons between LEWES/EiJ and other strains, the highest mean sequence identity was with the BALB/c germline set (97.1%; Figure 2B).

DISCUSSION

This study was undertaken to investigate the hypothesis that the antibody genes of subspecies of the house mouse are highly divergent. Since the mice that were used to establish the classical inbred strains of laboratory mice came from diverse and usually undocumented sources, this could explain the marked differences that are seen between the sets of IGHV genes found in different strains of mice. To test this hypothesis, we inferred the IGHV germline genes of wild-derived strains representing each of the three major subspecies of the house mouse (CAST/EiJ: *M. m. castaneus*; PWD/PhJ: *M. m. musculus*; LEWES/EiJ: *M. m. domesticus*), and of a wild-derived strain (MSM/MsJ) that originated from *M. m. molossinus* mice that are generally considered to be hybrids of *M. m. musculus* and *M. m. castaneus*. Whereas SNP-inferred haplotypes reported by Yang and colleagues [35] supported the suspected subspecies origins of

CAST/EiJ, PWD/PhJ, and LEWES/EiJ, this genomic data suggested that the MSM/MsJ IGH locus is *M. m. musculus*-derived, with a SNP profile that is little different to that of the C57BL/6 mouse (see Supplementary Figure 1). The NOD/ShiLtJ strain was also included in the study, because SNP analysis suggested that it too carries a C57BL/6-like haplotype. It was hoped that analysis of these strains would therefore provide some insights into genetic variation within the *M. m. musculus* subspecies.

In general, among the wild-derived strains surveyed in this study we observed surprisingly little overlap between IGHV sequences. While this was expected in comparisons of strains predicted to carry IGH loci originating from different subspecies, as postulated above, intriguingly there were notable differences in IGHV sequence sets between strains carrying loci of the same predicted subspecies. For example, PWD/PhJ mice only shared 13 (~14%) IGHV gene sequences with the C57BL/6, NOD/ShiLtJ, or MSM/MsJ strains, despite the fact that all of these strains were predicted to have IGH loci of shared *M. m. musculus* origin. In fact, the PWD/PhJ strain shared almost the same number of genes with the CAST/EiJ, LEWES/EiJ, and BALB/c strains (predicted to represent the *M. m. castaneus, M. m. domesticus* and *M. m. domesticus* subspecies respectively), and PWD/PhJ IGHV sequences were no more similar to sequences of other *M. m. musculus*-derived strains than to *M. m. castaneus*- or *M. m. domesticus*-derived strains. Similarly SNP analysis suggested that BALB/c and LEWES/EiJ mice both carry *M. m. domesticus*-derived IGH loci, but only 13 (~16%) of the identified LEWES/EiJ IGHV sequences are shared with the BALB/c strain. LEWES/EiJ IGHV sequences were collectively most similar to BALB/c genes, relative to the other strains

sequenced here, but it is difficult to believe that the IGH loci of the two strains are derived in their entirety from *M. m. domesticus* ancestors.

The IGHV genes of the MSM/MsJ strain were particularly surprising. Few of the 84 MSM IGHV sequences were seen in any of the other three wild-derived strains. As expected, none of these sequences were amongst the 78 IGHV genes identified in the *M. m. domesticus*-derived LEWES/EiJ strain; however, there was also little identity with sequences identified in either of the subspecies that are said to have given rise to the hybrid *M. m. molossinus* mice. Only 10 of the 84 MSM/MsJ IGHV sequences matched those seen in the *M. m. musculus*-derived PWD/PhJ strain, and only 4 sequences matched those in the *M. m. castaneus*-derived CAST/EiJ strain. Instead, substantial identity was seen between MSM/MsJ mice and inbred C57BL/6 and NOD/ShiLtJ mice, that SNP analysis suggests are both *M. m. musculus*-derived (Supplementary Figure 1).

The demonstration that NOD/ShiLtJ mice carry an IGH locus that is so closely related to that of the C57BL/6 strain is intriguing and also quite unexpected, because no direct relationship between these strains has been reported. The NOD/ShiLtJ mouse was derived in the 1970s from cataract-prone CTS mice, which in turn were developed in the 1960s from outbred Swiss mice [36, 37]. C57BL/6 mice, on the other hand, were developed in the 1920s by Clarence Little from the progeny of a pair of 'fancy mice' [38, 39]. It is difficult to believe that the NOD and C57BL/6 loci could have arisen independently by the chance selection of unrelated outbred founder pairs. The nearly identical sets of IGHV genes in the C57BL/6 and NOD/ShiLtJ mice are more suggestive of introgression,

with a C57BL/6-like locus being introduced into the ancestors of the modern NOD strain by outcrossing. This notion is further supported by the observation that NOD clusters together with C57BL/6 based on mitochondrial SNP but is rather distant from the strain based on chromosome Y SNP [35], hinting towards a contamination via the maternal line. Of note, this would not be the first reported breeding accident involving the NOD lineage (e.g., see [40]).

If the IGHV genes of the LEWES/EiJ, PWD/PhJ and CAST/EiJ mice are accepted as being broadly representative of the three major subspecies of the house mouse, then neither the BALB/c strain nor the C57BL/6 strain can be unequivocally linked with one or other of the three major subspecies of the house mouse. It may be that the IGH loci of these and other classical inbred strains have a mosaic structure, being made up of many blocks of genes from the three mouse subspecies. It is also possible that the IGH loci of these strains include haplotype blocks derived from other lineages of the house mouse, or even from other Mus species such as *M. spretus*. Other subspecies of *M. m. musculus* probably exist, and a number have been proposed, such as *M. m. bactrianus* and *M. m. gentilulus* [41, 42, 43].

The divergence of the IGH loci of the mouse strains reported here can be contrasted with the allelic diversity that has been reported in humans. Although the population genetics of the human IGH locus remains relatively uncertain [18, 21, 22], some indirect measures of diversity are available to us. One measure of diversity is provided by consideration of heterozygous loci in individuals who have been genotyped by the

analysis of VDJ rearrangements. When heterozygosity was explored at 50 IGHV gene loci in 98 individuals, only 5 genes were heterozygous in more than 50% of individuals, while 19 genes were heterozygous in more than 20% of individuals [20]. Homozygosity was a conspicuous feature of individual genotypes, with heterozygosity being seen at fewer than ⅓ of genes with a defined genotype in all but a handful of subjects [20]. A further measure of human diversity comes from reported allele usage in geographically distant communities. One study has reported many previously unreported IGHV alleles in individuals from southern Africa, however it is equally noteworthy that most sequences identified in this study have been repeatedly reported in studies from Europe and America [19]. Similarly, a study of 10 individuals from Papua New Guinea identified 17 previously unreported IGHV sequences, but in each individual, all but two or three sequences had previously been reported from studies in Europe, America and Australia [17]. Taken together these studies suggest that there may be less variability in the human population than we report here from a handful of inbred mouse strains. Investigation of many mouse strains will be required to determine the relative contributions of allelic variation and structural variation to this variability.

If we are to properly understand the antibody repertoires of the major laboratory strains, their IGH loci will all need to be separately investigated, and in the short term this is likely to be done by the inference of genes from AIRR-seq data rather than by genomic studies. The resulting lack of knowledge of non-coding regions, and the lack of positional data, may make it difficult to decipher relationships between some strains, but it should provide the basic information regarding germline genes that is needed for

accurate repertoire studies. It should also help us better understand the genetic basis for strain-related differences in mouse models of human disease. Allelic variants have been associated with differences in disease susceptibility of rats [44] and humans [21]. IGHV sequence variability might also contribute to the differences that have been reported in the susceptibility of inbred mouse strains to both infectious [45, 46, 47] and autoimmune diseases [48, 49].

The discovery of striking differences in the number of apparently functional IGHV genes between the C57BL/6 and the BALB/c strains first raised the possibility that the continued use of a positional nomenclature system for IGHV genes could be problematic [6]. The results presented here confirm that this is the case. The C57BL/6 locus is unable to serve as a map for IGHV genes from other strains, and this appears to be the case even for inbred strains that were shown in earlier SNP analysis to carry *M. m. musculus*-derived IGH loci. A non-positional nomenclature should therefore be developed. Attention should also be paid to the light chain loci carried by different mouse strains. We have shown by SNP analysis that the three major subspecies of the house mouse may all have contributed to the kappa loci (IGK) of the major strains of inbred laboratory mice [50]. If the IGK and IGL loci of laboratory mice are shown to have the same kind of strain to strain variability as we have shown here for the IGH locus, then a new non-positional nomenclature will be required for all the genes of the immunoglobulin loci.

FIGURE LEGENDS

Figure 1. Comparisons of inferred germline IGHV sequences to those represented in the mouse IMGT database (imgt.org). (A) Donut plots depicting the proportion of inferred germline sequences from each strain that align to a known IMGT allele with 100% match identity. (B) Boxplots depicting the sequence similarities of inferred germline sequences from each strain when compared to the closest known IMGT allele. (C) The count of identified germline sequences from each strain representing known mouse IGHV gene subfamilies. (D) The count of non-IMGT inferred germline sequences in each strain, partitioned by IGHV gene subfamily.

Figure 2. Relationships of IGHV inferred germline sequences among mouse strains. (A) Upset plot depicting the size of the germline set from each of the analyzed strains (left), as well as the numbers of sequences either unique to a given strain or shared among strains (identical sequences). (B) Heatmap depicting the mean percent sequence match identities among inferred IGHV germline sets for each pair-wise strain comparison (see also Supplementary Figure 4).

Supplementary Figure 1. SNP data from the IGHV gene region (chr12:114700000-117270000) suggest common subspecies origins for IGHV genomic haplotypes of inbred and wild-derived laboratory mouse strains. This figure depicts the predicted subspecies origins (*Mus musculus domesticus*; *M. m. musculus*; *M. m. castaneous*) of IGHV haplotypes in the six strains analyzed in the present study. Here, we consider the relationships between inferred germline IGHV gene sets of these strains in the context of these predicted subspecific origins. Data presented in this figure were obtained from the Mouse Phylogeny Viewer [1] (https://msub.csbio.unc.edu/, based on previously published whole-genome SNP data [2].

Supplementary Figure 2. Counts of identified clones representing inferred germline sequences from each strain sequenced in this study. Additional information, including the nucleotide sequences of each inferred germline presented in these plots, can be found in Supplemental Table II.

Supplementary Figure 3. Boxplots depicting the numbers of representative clones for inferred IGHV germline sequence identified in each strain, partitioned by whether the inferred sequence was supported by IgDiscover.

Supplementary Figure 4. Violin plots depicting the best percent match identities among inferred IGHV germline sets for each pair-wise strain comparison. Individual best-hit alignments from each pair-wise sequence set comparison are displayed as individual points within each violin plot.

Supplementary Table I. Per-strain AIRR-seq library summary statistics for CAST/EiJ, LEWES/EiJ, MSM/MsJ, PWD/PhJ, and NOD/ShiLtJ.

Supplementary Table II. Complete database of inferred germline IGHV sequences from CAST/EiJ, LEWES/EiJ, MSM/MsJ, PWD/PhJ, and NOD/ShiLtJ.

REFERENCES

1. Morse HC. *Origins of Inbred Mice*. Academic Press: New York, 1978.

2. Svasti J, Milstein C. The complete amino acid sequence of a mouse kappa light chain. *Biochem J* 1972, **128**(2)**:** 427-444.

3. Leder P, Honjo T, Packman S, Swan D, Nau M, Norman B. The organization and diversity of immunoglobulin genes. *Proc Natl Acad Sci USA* 1974, **71**(12)**:** 5109-5115.

4. Alt FW, Baltimore D. Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-JH fusions. *Proc Natl Acad Sci USA* 1982, **79**(13)**:** 4118-4122.

5. Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* 2000, **102**(5)**:** 553-563.

6. Collins AM, Wang Y, Roskin KM, Marquis CP, Jackson KJ. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Philos Trans R Soc Lond B Biol Sci* 2015, **370**(1676).

7. Potter M. Antigen-binding myeloma proteins of mice. *Adv Immunol* 1977, **25:** 141-211.

8. Riblet R. Immunoglobulin heavy chain genes in the mouse. In: Honjo T, Alt FW, Neuberger M (eds). *Molecular Biology of B cells*. Elsevier Academic Press: London, 2004, pp 19-26.

9. Johnston CM, Wood AL, Bolland DJ, Corcoran AE. Complete sequence assembly and characterization of the C57BL/6 mouse Ig heavy chain V region. *J Immunol* 2006, **176**(7)**:** 4221-4234.

10. Retter I, Chevillard C, Scharfe M, Conrad A, Hafner M, Im TH*, et al.* Sequence and characterization of the Ig heavy chain constant and partial variable region of the mouse strain 129S1. *J Immunol* 2007, **179**(4)**:** 2419-2427.

11. Retter I, Althaus HH, Munch R, Muller W. VBASE2, an integrative V gene database. *Nucleic Acids Res* 2005, **33**(Database issue)**:** D671-674.

12. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D*, et al.* IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res* 2006, **34**(Database issue)**:** D781-784.

13. Matthyssens G, Rabbitts TH. Structure and multiplicity of genes for the human immunoglobulin heavy chain variable region. *Proc Natl Acad Sci USA* 1980, **77**(11)**:** 6561-6565.

14. Matsuda F, Ishii K, Bourvagnet P, Kuma K, Hayashida H, Miyata T*, et al.* The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J Exp Med* 1998, **188**(11)**:** 2151-2162.

15. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR*, et al.* Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc Natl Acad Sci USA* 2009, **106**(48)**:** 20216-20221.

16. Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD*, et al.* Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol* 2010, **184**(12)**:** 6986-6992.

17. Wang Y, Jackson KJ, Gaeta B, Pomat W, Siba P, Sewell WA*, et al.* Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. *Immunogenetics* 2011, **63**(5)**:** 259-265.

18. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J*, et al.* 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* 2013, **92**(4)**:** 530-546.

19. Scheepers C, Shrestha RK, Lambson BE, Jackson KJ, Wright IA, Naicker D*, et al.* Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire. *J Immunol* 2015, **194**(9)**:** 4371-4378.

20. Gidoni M, Snir O, Peres A, Polak P, Lindeman I, Mikocziova I*, et al.* Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nature communications* 2019, **10**(1)**:** 628.

21. Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS*, et al.* IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Sci Rep* 2016, **6:** 20842.

22. Watson CT, Glanville J, Marasco WA. The Individual and Population Genetics of Antibody Immunity. *Trends Immunol* 2017, **38**(7)**:** 459-470.

23. Ohlin M, Scheepers C, Corcoran M, Lees WD, Busse CE, Bagnara D*, et al.* Inferred Allelic Variants of Immunoglobulin Receptor Genes: A System for Their Evaluation, Documentation, and Naming. *Front Immunol* 2019, **10:** 435.

24. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. 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* 2015, **112**(8)**:** E862-870.

25. Corcoran MM, Phad GE, Vazquez Bernat N, Stahl-Hennig C, Sumida N, Persson MA*, et al.* Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nature communications* 2016, **7:** 13642.

26. Zhang W, Wang IM, Wang C, Lin L, Chai X, Wu J*, et al.* IMPre: An Accurate and Efficient Software for Prediction of T- and B-Cell Receptor Germline Genes and Alleles from Rearranged Repertoire Data. *Front Immunol* 2016, **7:** 457.

27. Gadala-Maria D, Gidoni M, Marquez S, Vander Heiden JA, Kos JT, Watson CT*, et al.* Identification of Subject-Specific Immunoglobulin Alleles From Expressed Repertoire Sequencing Data. *Front Immunol* 2019, **10:** 129.

28. Collins AM, Jackson KJL. On being the right size: antibody repertoire formation in the mouse and human. *Immunogenetics* 2018, **70**(3)**:** 143-158.

29. Vander Heiden JA, Yaari G, Uduman M, Stern JN, O'Connor KC, Hafler DA*, et al.* pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* 2014, **30**(13)**:** 1930-1932.

30. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* 2015, **31**(20)**:** 3356-3358.

31. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010, **26**(5)**:** 680-682.

32. Meyerhans A, Vartanian JP, Wain-Hobson S. DNA recombination during PCR. *Nucleic Acids Res* 1990, **18**(7)**:** 1687-1691.

33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990, **215**(3)**:** 403-410.

34. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res* 2002, **12**(4)**:** 656-664.

35. Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE*, et al.* Subspecific origin and haplotype diversity in the laboratory mouse. *Nat Genet* 2011, **43**(7)**:** 648-655.

36. Makino S, Kunimoto K, Muraoka Y, Mizushima Y, Katagiri K, Tochino Y. Breeding of a non-obese, diabetic strain of mice. *Jikken Dobutsu* 1980, **29**(1)**:** 1-13.

37. Mullen Y. Development of the Nonobese Diabetic Mouse and Contribution of Animal Models for Understanding Type 1 Diabetes. *Pancreas* 2017, **46**(4)**:** 455-466.

38. Staats J. The laboratory mouse. In: Green EL (ed). *Biology of the Laboratory Mouse*. McGraw-Hill: New York, 1966, pp 1-9.

39. Beck JA, Lloyd S, Hafezparast M, Lennon-Pierce M, Eppig JT, Festing MF*, et al.* Genealogies of mouse inbred strains. *Nat Genet* 2000, **24**(1)**:** 23-25.

40. Prochazka M, Serreze DV, Frankel WN, Leiter EH. NOR/Lt mice: MHC-matched diabetes-resistant control strain for NOD mice. *Diabetes* 1992, **41**(1)**:** 98-106.

41. Musser GG, Carleton MD. Superfamily Muroidea. In: Wilson DE, Reeder DM (eds). *Mammal Species of the World: A Taxonomic and Geographic Reference*, 3rd edition edn. Johns Hopkins University Press: Baltimore, 2005, pp 894–1531.

42. Suzuki H, Nunome M, Kinoshita G, Aplin KP, Vogel P, Kryukov AP*, et al.* Evolutionary and dispersal history of Eurasian house mice Mus musculus clarified by more extensive geographic sampling of mitochondrial DNA. *Heredity (Edinb)* 2013, **111**(5)**:** 375-390.

43. Suzuki H, Yakimenko LV, Usuda D, Frisman LV. Tracing the eastward dispersal of the house mouse, Mus musculus. *Genes and environment* 2015, **37:** 20.

44. Dhande IS, Cranford SM, Zhu Y, Kneedler SC, Hicks MJ, Wenderfer SE*, et al.* Susceptibility to Hypertensive Renal Disease in the Spontaneously Hypertensive Rat Is Influenced by 2 Loci Affecting Blood Pressure and Immunoglobulin Repertoire. *Hypertension* 2018, **71**(4)**:** 700-708.

45. Swihart K, Fruth U, Messmer N, Hug K, Behin R, Huang S*, et al.* Mice from a genetically resistant background lacking the interferon gamma receptor are susceptible to infection with Leishmania major but mount a polarized T helper cell 1-type CD4+ T cell response. *J Exp Med* 1995, **181**(3)**:** 961-971.

46. Caron J, Loredo-Osti JC, Laroche L, Skamene E, Morgan K, Malo D. Identification of genetic loci controlling bacterial clearance in experimental Salmonella enteritidis infection: an unexpected role of Nramp1 (Slc11a1) in the persistence of infection in mice. *Genes Immun* 2002, **3**(4)**:** 196-204.

47. Fortier A, Min-Oo G, Forbes J, Lam-Yuk-Tseung S, Gros P. Single gene effects in mouse models of host: pathogen interactions. *J Leukoc Biol* 2005, **77**(6)**:** 868-877.

48. Hannestad K, Scott H. The MHC haplotype H2b converts two pure nonlupus mouse strains to producers of antinuclear antibodies. *J Immunol* 2009, **183**(5)**:** 3542-3550.

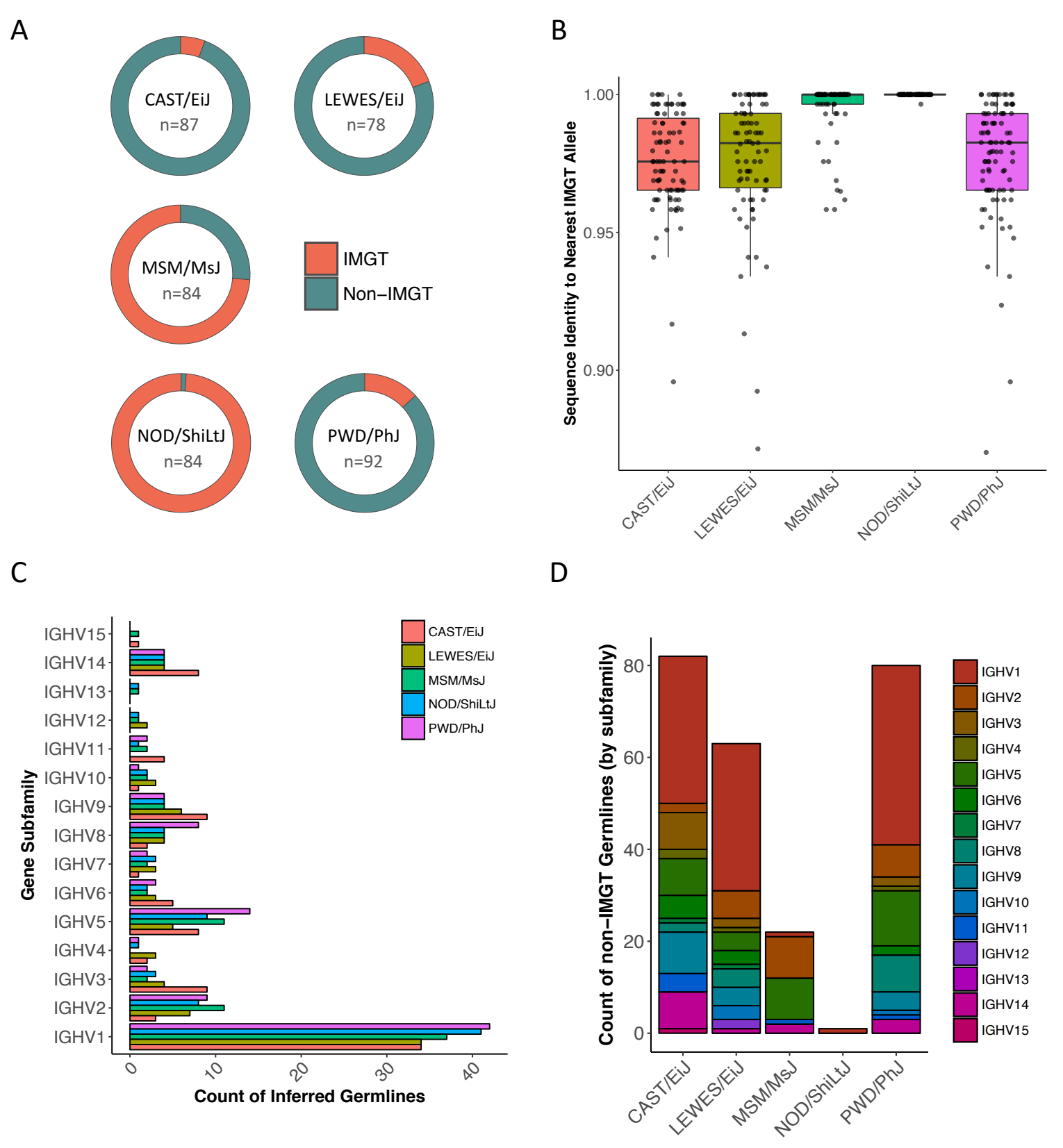
49. Koh AE, Njoroge SW, Feliu M, Cook A, Selig MK, Latchman YE*, et al.* The SLAM family member CD48 (Slamf2) protects lupus-prone mice from autoimmune nephritis. *J Autoimmun* 2011, **37**(1)**:** 48-57.

50. Collins AM, Watson CT. Immunoglobulin Light Chain Gene Rearrangements, Receptor Editing and the Development of a Self-Tolerant Antibody Repertoire. *Front Immunol* 2018, **9** (2249).

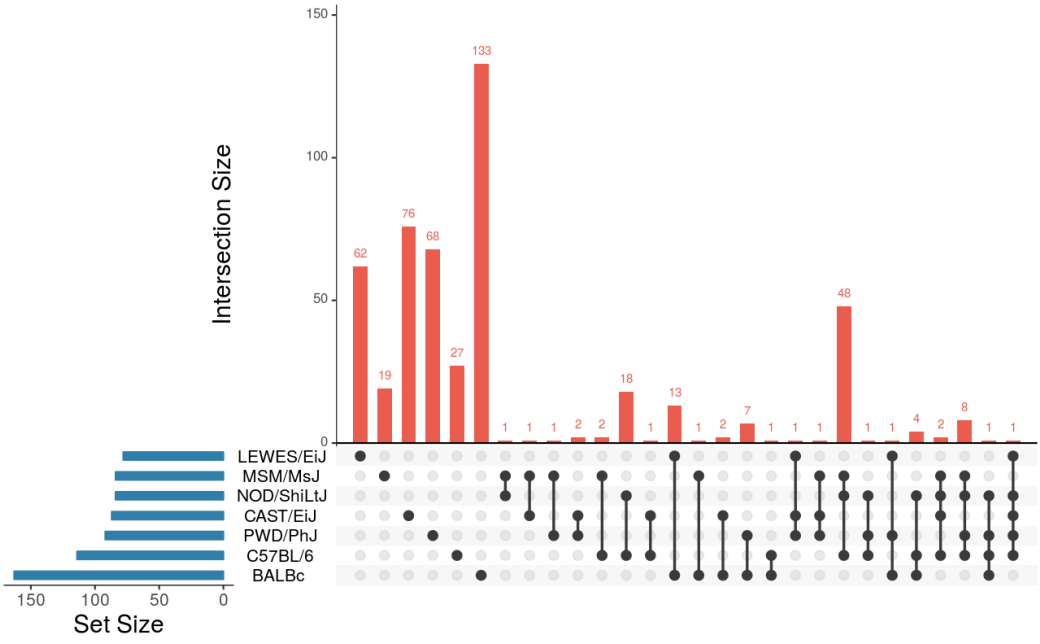
Additional References for Supplementary Figures

1. Wang JR, de Villena FP, McMillan L. Comparative analysis and visualization of multiple collinear genomes. *BMC Bioinformatics* 2012, **13 Suppl 3:** S13.

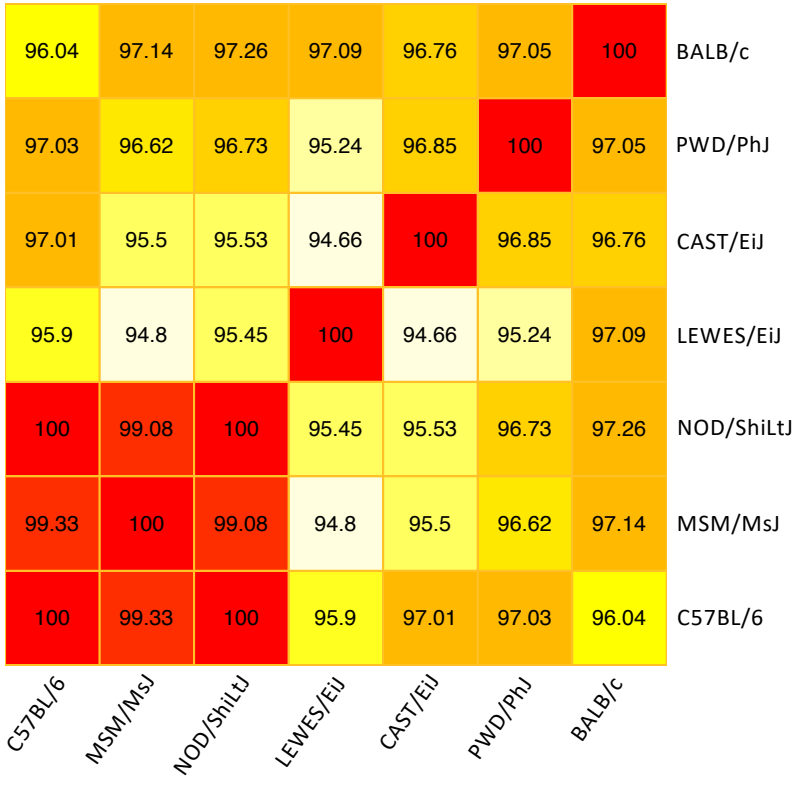
2. Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE*, et al.* Subspecific origin and haplotype diversity in the laboratory mouse. *Nat Genet* 2011, **43**(7)**:** 648-655.


A
CAST/EiJ
n=87
LEWES/EiJ
n=78
MSM/MsJ
n=84
NOD/ShiLtJ
n=84
PWD/PhJ
n=92
IMGT
Non−IMGT
B
Sequence Identity to Nearest IMGT Allele
1.00
0.95
0.90
CAST/EiJ
LEWES/EiJ
MSM/MsJ
NOD/ShiLtJ
PWD/PhJ
C
Gene Subfamily
IGHV15
IGHV14
IGHV13
IGHV12
IGHV11
IGHV10
IGHV9
IGHV8
IGHV7
IGHV6
IGHV5
IGHV4
IGHV3
IGHV2
IGHV1
0
10
20
30
40
Count of Inferred Germlines
CAST/EiJ
LEWES/EiJ
MSM/MsJ
NOD/ShiLtJ
PWD/PhJ
D
Count of non−IMGT Germlines (by subfamily)
0
20
40
60
80
CAST/EiJ
LEWES/EiJ
MSM/MsJ
NOD/ShiLtJ
PWD/PhJ
IGHV1
IGHV2
IGHV3
IGHV4
IGHV5
IGHV6
IGHV7
IGHV8
IGHV9
IGHV10
IGHV11
IGHV12
IGHV13
IGHV14
IGHV15

A


Intersection Size
150
100
50
0
62
19
76
68
27
133
1
1
1
2
2
18
1
13
1
2
7
1
1
1
48
1
1
4
2
8
1
1
LEWES/EiJ
MSM/MsJ
NOD/ShiLtJ
CAST/EiJ
PWD/PhJ
C57BL/6
BALBc
150
100
50
0
Set Size

B


96.04 97.14 97.26 97.09 96.76 97.05 100 BALB/c
97.03 96.62 96.73 95.24 96.85 100 97.05 PWD/PhJ
97.01 95.5 95.53 94.66 100 96.85 96.76 CAST/EiJ
95.9 94.8 95.45 100 94.66 95.24 97.09 LEWES/EiJ
100 99.08 100 95.45 95.53 96.73 97.26 NOD/ShiLtJ
99.33 100 99.08 94.8 95.5 96.62 97.14 MSM/MsJ
100 99.33 100 95.9 97.01 97.03 96.04 C57BL/6
C57BL/6
MSM/MsJ
NOD/ShiLtJ
LEWES/EiJ
CAST/EiJ
PWD/PhJ
BALB/c