Multiple full-length homozygous IGH haplotypes from Mauritian cynomolgus macaques

Simone Olubo[1]*, William S. Gibson[1]*, Trent M. Prall[2], Julie A. Karl[2], Roger W. Wiseman[2,3], David H. O'Connor[2,3], Daniel C. Douek[1#], Chaim A. Schramm[1#]

[1]Vaccine Research Center, National Institutes of Health, Bethesda, MD, 20892, USA.

[2]Department of Pathology and Laboratory Medicine, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, Wisconsin 53705, USA.

[3]Wisconsin National Primate Research Center, University of Wisconsin-Madison, Madison, Wisconsin 53715, USA.

*These authors contributed equally to this work

[#]Correspondence to ddouek@mail.nih.gov and chaim.schramm@nih.gov

Abstract

Background:

Nonhuman primates are frequent experimental models for human disease pathology and vaccine design. However, the vast and mostly uncatalogued immunogenomic diversity of typical species adds complexity to the interpretation of experiments and hinders reproducibility. Mauritian cynomolgus macaques (MCM) offer a unique opportunity to circumvent these difficulties, due to their restricted genetic diversity.

Results:

We assembled high-quality immunoglobulin heavy chain (IGH) haplotypes from long-read genomic sequencing of 13 MCM. Four animals were homozygous for IGH, yielding 3 distinct haplotypes. IGH haplotype H1 was observed in two of the homozygotes and 5 additional heterozygous animals, accounting for half of the assemblies recovered. The 3 homozygous haplotypes exhibited considerable variation, including a 125 kilobase region that was duplicated twice in H3. Furthermore, H1 shares only 83% average sequence identity with the IGH locus of the rhesus macaque reference genome, in addition to numerous large structural variations. We annotated the IG gene content from all complete MCM IGH assemblies and found 298 functional IGHV alleles, of which 94 (32%) were not in existing databases. We also identified 69 functional IGHD alleles, 11 functional IGHJ alleles, and 38 functional constant gene alleles across all 5 isotypes.

Conclusions:

In total, we identified multiple common and genetically diverse IGH haplotypes within MCM and provide high-quality reference assemblies and annotations for these to facilitate future work with this important animal model.

Background

Rhesus (*Macaca mulatta*) and cynomolgus (*Macaca fascicularis*) macaques are non-human primates (NHP) that share a close evolutionary history with humans and are widely used as biomedical models of human health for infectious disease and transplantation research [1–7]. However, their increasing cost and the logistical complexity of animal research protocols often limit studies to small sample sizes [8,9]. This means that NHP studies are typically underpowered to determine the magnitude of genetic effects on traits of interest [10], a problem compounded by unavailable or incorrect animal pedigrees [8].

Studies of rhesus and cynomolgus genetics reveal extensive genetic diversity inter- and intra-species with rhesus macaques being more than twice as diverse per individual compared to humans [11]. In this context, Mauritian-origin cynomolgus macaques (MCMs) provide a particularly valuable advantage as model organisms compared to other genetically diverse NHP. MCMs descended from a small (~20) founding population introduced to the island of Mauritius in the 16th century, which expanded to the current population of ~50,000 animals [12,13]. As a result of this bottleneck event, MCMs exhibit significantly constrained genetic diversity compared to other cynomolgus and rhesus macaques [13]. This makes them excellent model organisms especially in the context of immunological studies where germline variation can significantly impact results [14–16].

Genetic diversity is exceptionally pronounced in key immune loci such as the major histocompatibility complex (MHC) [17–22], killer cell immunoglobulin-like receptors(KIR) [23,24], and Fc-gamma receptors (FcγR) [25]. In comparison, relatively little is known about the T cell receptor (TR) [15,26–28] and immunoglobulin (IG) loci [15,29,30] in macaques, despite their crucial roles in the adaptive immune system. The high density of segmental duplications and repetitive elements, as well as the presence of large structural variations, make accurate assembly of the IGH locus difficult [29,31,32]. Specialized sequencing and assembly methods alongside purpose-built annotation tools are required to resolve this region [11,31–36].

As an alternative to genomic sequencing, the germline sequences of variable (V), diversity (D), and joining (J) genes can be inferred from adaptive immune receptor repertoire sequencing [37–40] (AIRR-seq). The Karolinska Macaque database [30,31] contains 774 functional IGHV alleles and 14 functional IGHJ alleles inferred from 28 rhesus macaques, as well as 615 and 14 functional IGHV and IGHJ alleles, respectively, inferred from 18 cynomolgus macaques. Another recent study added an additional 84 inferred cynomolgus IGHV alleles [15]. The large number of reported IGH alleles from a relatively small number of sequenced individuals suggests the macaque IGH loci harbor significant diversity, similar to observations made in humans [41,42].

Despite these successes, AIRR-seq-derived inferences have limited ability to resolve D and constant (C) genetics. Furthermore, large structural variations within the IGH locus have been observed in macaques [11,29,31], similar to other primates [32], and humans [42]. High quality genomic references are thus essential for elucidating the complexity of the IGH loci in MCMs and as robust frameworks for inference-based approaches. Additionally, the power of the MCM as a model organism relies on a comprehensive understanding of genetic diversity within the constrained population. This enables explicitly designed studies that either limit or specifically target the impact of genetic diversity on the trait of interest.

While there are multiple high quality, annotated assemblies of the IG loci for rhesus macaques [11,29,31,43], the only genomic IG reference for cynomolgus macaques is an unlocalized scaffold from a 2013 draft genome primarily generated using 2x100 paired-end short reads [44]. A new telomere-to-telomere cynomolgus reference genome has recently become available [45], but IG genes have not been annotated. In addition, there are known to be differences in IGH germline alleles and gene usage between MCM and other cynomolgus populations [30].

To fill in this gap, we leveraged long-read whole-genome sequencing data from 13 MCMs [24] to generate curated and annotated assemblies of the MCM IGH locus for use as a genomic resource. MCMs represent a unique opportunity to account for genetic background in a highly utilized NHP model organism. Despite the genetic restriction of MCMs, we identified significant allelic and structural variation between haplotypes further demonstrating the potential of MCMs as a powerful model for immunogenetic investigation and reproducibility.

Results

Efficient assembly of IGH from whole genome long read sequencing

Long read technologies enable the assembly of highly repetitive regions and the resolution of structural variants, including in IGH [32,33,46]. We have previously generated whole genome sequencing of 13 MCMs using a combination of Oxford Nanopore Technology and PacBio HiFi sequencing [24]. We assembled the reads for each animal using Hifiasm [47,48] and extracted the IGH locus (Table S1). We were able to recover 14 scaffolds from 11 animals that spanned the entire IGH locus (Table 1). Notably, four animals were homozygous in IGH (Fig 1a and Fig S1) with the rest being heterozygous.

The most common haplotype ("H1") was present in two of the homozygotes with relatively uniform read coverage of 7.3x and 10.3x, respectively, across the IGH locus (Fig 1a). Large spikes in coverage occurred over repetitive elements as a result of off-target reads mapping to these regions. We identified five additional H1 haplotypes from heterozygous animals. Two of these heterozygous H1 haplotypes were fragmented into two and three contigs, respectively (Table 1). We combined the three remaining contiguous H1 haplotypes with the two homozygous haplotypes to generate an H1 consensus for further analysis (Fig 1b). Remarkably, the five H1 haplotypes were 99.999% identical, differing at only 20 positions across the 2 megabase (Mb) H1 haplotype. All variable positions fell within intergenic regions (Fig 1, Fig S2). Overall, this suggests that the H1 consensus is representative of the common H1 haplotype and can serve as a reliable genetic reference for analysis of MCM IGH.

We used ALIGaToR (https://github.com/scharch/aligator) to annotate the IG genes present in H1 (Fig 2a-b). We found 81 functional, 1 ORF, and 30 pseudo IGHV genes; 38 functional and 6 ORF IGHD genes; 5 functional and 2 ORF IGHJ genes; and 8 functional IGHC genes. Of the functional and ORF alleles, 61 IGHV, 29 IGHD, and 7 IGHJ alleles were previously annotated in KIMDB [30]. The novel alleles were distributed throughout the locus and across IGHV and IGHD gene families (Fig 2b). Notably, when compared to the IGHV alleles present in the existing cynomolgus genomic reference, only 69 (84%) of the H1 IGHV alleles had close orthologs (≥95% sequence identity; Fig S3), highlighting the need for updated resources that fully capture the immunogenetic diversity of this species.

We were also able to resolve highly repetitive regions of H1 that were enriched for homologous sequence (Fig 3). The proximal region (Fig 3a) spans ~450 KB and contains 24 functional IGHV alleles. Gene-rich homology blocks ~10-15 KB in size are interspersed in the region suggesting a history of sequence duplications and rearrangements. The distal region (Fig 3b) is ~400 KB and contains 20 functional IGHV alleles arranged in larger homologous blocks, indicating more recent segmental duplication events. These patterns of highly homologous sequence clusters are a common feature of the immunoglobulin loci across different species and are frequently hotspots of structural variation within a species.

Structural variation within MCM IGH

We then compared the H1 haplotype to two additional haplotypes, H2 and H3, resolved from homozygous animals. Similar to other studies of immunogenetic regions in these animals [20,24], we found a high degree of haplotype diversity. The H1 IGH haplotype is the shortest at 2.07 Mb, while the H2 and H3 haplotypes are 2.3 Mb and 2.48 Mb respectively. The sequence disparity between haplotypes is primarily due to large gene-rich structural variations. The IGHC, IGHJ, and IGHD regions contain little or no structural variation though the sequence identity in these regions varies from ~92% to >99%. In contrast, the IGHV region contains significant structural variation which primarily occur in or directly adjacent to highly repetitive gene-rich regions. Haplotypes H1 and H2 primarily differ at the proximal end of the IGHV region with interspersed regions 95-97.5% sequence identity and two large structural variations approximately 600 Kb and 1,400 Kb into the locus (Fig 4). The distal end of H1 and H2 have large blocks of highly homologous sequence with >99% identity (Fig 4). Comparatively, the H3 haplotype has a high degree of sequence variation across the entirety of the IGHV region relative to both H1 and H2 (Fig 4).

Relative to H1, the H3 haplotype contains a large ~200 Kb segmental duplication in the distal highly repetitive region. Each duplication block contains 14 functional IGHV genes, is repeated two times, and has an overall sequence similarity of >99.6% between the blocks. This large segmental duplication is responsible for the majority of the size difference between the H1 and H3 IGH locus. While the segmental duplication is highly sequence similar, the IGHV genes within each duplication block are not all identical to their counterparts. Across the three duplication blocks, there are 27 distinct functional alleles with 13 appearing once, 13 appearing twice, and one appearing in all three blocks. This degree of allelic differentiation in an otherwise highly homologous segmental duplication may be a result of selective pressure driving diversity.

Differences between MCM and the rhesus genomic reference

We conducted a comparative analysis of the H1 haplotype with the IGH locus of the rhesus macaque reference assembly Mmul_10. The Mmul_10 IGH locus contains 27 fewer functional genes overall, missing one IGHC gene, 25 IGHD alleles, and three IGHV alleles while having seven functional IGHJ alleles compared to H1's five functional and two ORF IGHJ alleles (Fig 6). A large proportion of the IGHV allele sequences are not shared between the two haplotypes (Fig S3). In addition, there are large structural variations between these haplotypes, consistent with comparisons between other primate species [32]. These structural variations result in the H1 haplotype being 103 Kb larger than the Mmul_10 haplotype, are localized to the IGHV region of the locus, and result in the insertion or deletion of multiple functional genes. There are also large differences in nucleotide sequence between the H1 and Mmul_10 haplotypes with only 83% homology.

Mapping the range of IGH genetic diversity within MCM

In addition to the three complete haplotypes recovered from homozygous animals, four additional fully contiguous haplotypes were resolved from four heterozygous animals, each of which representing a unique haplotype (Fig S4). For H4, we were able to identify the full complement of IGHV and IGHC alleles, but only 1 IGHJ and 7 IGHD alleles, possibly due to assembly artifacts (Table 1). We found that plurality of IGHV alleles (132/298, 44%) are uniquely

found in a single haplotype (Fig 7a). An additional 78 alleles (26%) were found in exactly three haplotypes, due to several large blocks of high sequence identity in the IGHV region shared between H1, H2, and H4 (Fig 4, S5). By contrast, IGHD alleles were predominately seen across at least six haplotypes (19/69, 28%) or unique to a haplotype (20/69, 29%). (Fig 7b). No clear pattern was observed for IGHJ alleles (Fig 7c).

The immense structural variation among macaques in immune loci like IGH makes it difficult to align genomes with sufficient resolution to identify sets of alleles that originate from the same gene [33,49,50]. Instead, we used PIgLET [34] to group coding sequences into Allele Similarity Clusters (ASCs), using an identity cutoff of 95% (90% for IGHD alleles) to approximate the similarity among alleles of a single gene. We found 109 IGHV ASCs, 36 IGHD ASCs, and 7 IGHJ ACSs across all seven haplotypes. In general, IGHV ASCs contain multiple alleles per ASC, with 35 IGHV ASCs (32%) having four or more distinct alleles (Fig 7d). In contrast, IGHD ASCs predominantly contained one or two alleles, with only three IGHD ASCs (8%) having four or more alleles (Fig 7e). The IGHJ ACSs followed a similar pattern, with five of the eight ASCs having a single allele with the remainder having two alleles per ASC (Fig 7f). While the shorter length of IGHD and IGHJ alleles compared to IGHV alleles would require a smaller number of nucleotide variants to separate sequences into distinct ASC clusters, the conservation of IGHJ allele sequences and the bimodal frequency of unique and common IGHD alleles across all haplotypes suggest multiple factors influencing the diversity of these genes.

To assess the extent of MCM immunogenetic diversity, we compared the alleles recovered from the 13 animals in this study to those inferred from repertoire sequencing of an independent group of 12 MCM [30]. Of 376 total IGHV alleles across both studies, 204 (54%) were present in both groups of animals (Fig 7g). Given the propensity of IGHV alleles to segregate by haplotype (Fig 7a), this offers further evidence that a small number of animals is sufficient to capture most of the genetic diversity of IGH in MCM. Similarly, 40/77 (52%) of all IGHD alleles were observed in both cohorts (Fig 7h), although with significantly more unique alleles in the current study. This may reflect the difficulty of reconstructing germline IGHD gene sequences from repertoire sequencing data [38,51].

We also considered allelic diversity in the constant chains. 12 of the 14 assemblies had all 8 expected isotypes annotated in the expected order, while IgG4 was not annotated in haplotype H3 and the related H7 due to an alignment error in ALIGaToR. Notably, we found IgE on all 14 assemblies, with a total of 5 unique alleles, though IgE is not annotated in the existing cynomolgus genomic reference. Within IgG, subtypes IgG1 and IgG2 are the most divergent (Fig 8a), while IgG3 and IgG4 were more similar to each other. Notably, the IgG3 alleles present in haplotypes H2, H3, and H7 had a short, IgG4-like hinge exon between the CH1 and CH2 domains, while the alleles from haplotype H1 and the remaining haplotypes have a longer hinge exon (Fig 8b). A longer hinge region may increase flexibility between antibody variable segments and potentially improve the ability to simultaneously bind multiple epitopes. Longer or shorter hinge regions may also impact effector functions and stability of the antibody.

Overall, we have provided the first full-length, curated genomic reference for IGH in cynomolgus macaques. In addition, by analyzing sequences from 13 animals, we were able to not only characterize inter-species variation between cynomolgus and rhesus macaques, but also asses the immunogenetic diversity within the MCM population, as well. Our results will facilitate future research with this important model species.

Discussion

Genomic resources are critical for fully unlocking the translational potential of studies using animal models for biomedical research. In particular, immune receptor genes such as MHC, KIR, TR, and IG play an outsized role in the course of infectious disease and the response to vaccination [52–55]. Despite this, controlling for genetic variation in these highly polymorphic loci is extremely challenging in outbred species like primates. MCMs offer a powerful tool for disentangling and assessing the role of genetic factors, due to the restricted genetic diversity in this population.

We have previously used long-read technology to resolve complete haplotypes of the genetic loci for MHC, KIR, and natural killer group 2 (NKG2) receptors in MCM [20,24]. Here, we assembled and characterized 7 full-length haplotypes of MCM IGH, including three from homozygous animals. The predominant haplotype, H1, accounted for half of the full-length IGH assemblies recovered. These results are consistent with our previous analyses of these animals, which found that 2 and 3 are homozygous for KIR and NKG2, respectively, and a dominant NKG2 haplotype N1 with a frequency of 40% [24]. As these are the first full-length curated assemblies of the IGH region in cynomolgus macaques, this represents a significant advance in the genomic resources available for this species.

Collectively, the haplotypes reported here identify a high degree of sequence diversity across individuals, despite the abundance of common haplotypes. Regions of high sequence homology between haplotypes nonetheless contain structural variations (Fig 4), with both small (Fig 3) and large (Fig 5) segmental duplications in gene-rich regions. Haplotype diversity is even more extreme when compared to rhesus macaque (Fig 6), further signifying the importance of accurate genomic references to account for genetic variation and to further explore the associations between germline and phenotype.

Our study is limited by the small number of animals sequenced and the limited sequencing depth achieved. We thus only assembled 14/26 (54%) of possible haplotypes, potentially biasing our assessment of MCM genetic diversity within IGH. In addition, 7/11 (64%) of the animals with at least one complete assembly are heterozygous, raising the possibility of assembly errors confounding our analyses [56]. Finally, we currently lack sufficient data to trace the history of complex structural variations across haplotypes and confidently link alleles from orthologous genes. Instead, we have relied on ASCs, which collapse evolutionarily recent duplications that still maintain high sequence homology (Fig 5).

In addition to structural variation across haplotypes, more data are needed to distinguish between the likely ancestral IGH haplotypes of the MCM population and more recent recombinants. Previous work has identified 7, 8, and 7 MCM haplotypes for MHC, KIR, and NKG2, respectively [57–60]. Each of these loci is on a different chromosome and therefore expected to segregate independently. By contrast, we currently describe 7 IGH haplotypes, with significant additional diversity still present both in these animals (Table 1) and other MCM (Fig 7g-h). Moreover, we found large blocks of high sequence homology across H1, H2, and H4 (Fig S5a) and between H3 and H7 (Fig S5b) that may be indicative of recent crossover events. Characterizing the evolutionary history of IGH since MCM diverged from other cynomolgus macaques will provide a valuable window into how diversity is generated and selected in this important locus.

In total, we annotated over 400 unique alleles across all IGH gene types, 160 (30%) of which are not present in a previous MCM database. Furthermore, 44% of IGHV alleles and 29% of IGHD alleles were unique to a specific haplotype (Fig 7 a-b). The large number of novel alleles and the diversity in gene and allele content between haplotypes demonstrates the necessity of genotyping study animals and continuing to add to databases. Future work will link genomic analysis with AIRR-seq data to better understand the functional effects of genetic diversity [42], with the potential to provide insights into the biological mechanisms underpinning different immune responses. Overall, the extensive analysis of the IGH region here has revealed significant biological implications for the usage of MCMs and cynomolgus macaques in biomedical research.

Methods

Assembly and mapping

Whole genome sequencing using PacBio and Oxford Nanopore Technology for 13 MCM were retrieved from NCBI (PRJNA854973 [24]) and assembled *de novo* with Hifiasm [47,48] (Cheng 2021; 2022) using the command:

```
hifiasm -o {output.asm} -t16 --ul {Nanopore.fq.gz} {PacBio.fq.gz}
```

BLAST+ was used to identify each IGH locus and the assemblies were then trimmed using an anchor sequence that fell between the constant and centromeric region of the IGH locus. Animal-specific genome references were created using the Mmul_10 IGH locus and the contigs generated by Hifiasm and used as the input for minimap2 [61]:

```
minimap2 -ax map-pb {reference}.fa {reads.fastq} > {output.sam}
```

Coverage was calculated using `samtools coverage` and plotted in R using `ggplot2`.

The H1 consensus was generated in Geneious Prime version 2023.0.4 (Biomatters Ltd.) and minimap2 was used to map each full-length H1 contig against the consensus:

```
minimap2 -ax asm20 {H1-cons.fasta} {input.fasta} > {output.sam}
```

and bcftools (Danecek2021) was used to call variants:

```
bcftools mpileup -f {H1-cons.fasta} {input.bam} |
      bcftools call -mv -o {input.vcf}
```

The SNVs were then visualized in Geneious.

Gene Annotation

ALIGaToR (https://github.com/scharch/aligator) was used to annotate the IGH assemblies starting from the IMGT [62] annotations of the Mmul_10 reference genome. Recombination signal sequences (RSS) were predicted based on RSS Information Content [63] using Dnagrep [64] in ALIGaToR. *M. fascicularis* IGHV, IGHD, and IGHJ coding sequences were downloaded from KIMDB (bernat2021) and combined with coding sequences of *M. fasciularis* IGHC from IMGT to use for naming discovered alleles. The final ALIGaToR command was:

```
aligator annotate contig.fasta contig.RSS12.bed contig.RSS23.bed \
      IGH IMGT_reference.fasta IMGT_reference.bed --outgff contig.gff \
      --outfasta functional_genes.fa \
      --alleledb KIMdb_IMGT_reference_sequences.fa \
      --psfasta pseudo_genes.fa
```

The annotations were visualized as a number line schematic using `matplotlib`. Tile plots and bar graphs were visualized using `ggplot2`. Venn diagrams were created using the `eulerr` package in R. The neighbor-joining tree was generated using the `ape` package and visualized using `color.plot.phylo` from the `picante` package.

Structural Variation

Alignments between haplotypes were generated using minimap2 [61]:

```
minimap2 -x asm20 -c -eqx -secondary=no {input_1.fasta} {input_2.fasta} >
{output.paf}
```

SVbyEye [65] was used to visualize sequence alignments. A minimum alignment size cutoff of 13000, 10000, and 9000 were used to generate Fig 4, Fig 5, and Fig 6 respectively. To filter out alignments, the `breakPaf` or `breakPafAlignment` commands were used to expand PAF alignments which were then subset to the specified cutoffs. In addition, the SVbyEye commands `addAnnotation` and `plotSelf` were modified to replace default coloring and shape parameters with custom coloring and shape parameters. StainedGlass [66] was used to generate dot plots of self-alignments (Fig 3) and haplotype comparisons (Fig 5). In brief, sequences for comparison were indexed using samtools [67] and used as inputs for the StainedGlass snakemake pipeline using default parameters. .

References

1. Anderson DJ, Kirk AD. Primate Models in Organ Transplantation. Cold Spring Harb Perspect Med. 2013;3:a015503.

2. Burwitz BJ, Wu HL, Abdulhaqq S, Shriver-Munsch C, Swanson T, Legasse AW, et al. Allogeneic stem cell transplantation in fully MHC-matched Mauritian cynomolgus macaques recapitulates diverse human clinical outcomes. Nat Commun. 2017;8:1418.

3. Knechtle SJ, Shaw JM, Hering BJ, Kraemer K, Madsen JC. Translational impact of NIH-funded nonhuman primate research in transplantation. Sci Transl Med. 2019;11:eaau0143.

4. Shiina T, Blancher A. The Cynomolgus Macaque MHC Polymorphism in Experimental Medicine. Cells. 2019;8:978.

5. D'Souza SS, Kumar A, Maufort J, Weinfurter JT, Raymond M, Strelchenko NS, et al. Assessment of safety and immunogenicity of MHC homozygous iPSC-derived CD34+ hematopoietic progenitors in an NHP model. Blood Adv. 2022;6:5267–78.

6. Gardner MB, Luciw PA. Macaque models of human infectious disease. ILAR J. 2008;49:220–55.

7. Shedlock DJ, Silvestri G, Weiner DB. Monkeying around with HIV vaccines: using rhesus macaques to define "gatekeepers" for clinical trials. Nat Rev Immunol. 2009;9:717–28.

8. Supply of monkeys for research is at a crisis point, U.S. government report concludes [Internet]. [cited 2024 Nov 27]. Available from: https://www.science.org/content/article/supply-monkeys-research-crisis-point-u-s-government-report-concludes

9. Record number of monkeys being used in U.S. research [Internet]. [cited 2024 Nov 27]. Available from: https://www.science.org/content/article/record-number-monkeys-being-used-us-research

10. Bliss-Moreau E, Amara RR, Buffalo EA, Colman RJ, Embers ME, Morrison JH, et al. Improving rigor and reproducibility in nonhuman primate research. Am J Primatol. 2021;83:e23331.

11. Warren WC, Harris RA, Haukness M, Fiddes IT, Murali SC, Fernandes J, et al. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. Science. 2020;370:eabc6617.

12. Osada N, Hettiarachchi N, Adeyemi Babarinde I, Saitou N, Blancher A. Whole-Genome Sequencing of Six Mauritian Cynomolgus Macaques (Macaca fascicularis) Reveals a Genome-Wide Pattern of Polymorphisms under Extreme Population Bottleneck. Genome Biol Evol. 2015;7:821–30.

13. Sussman RW, Tattersall I. Distribution, Abundance, and Putative Ecological Strategy of Macaca fascicularis on the Island of Mauritius, Southwestern Indian Ocean. Folia Primatol (Basel). 1986;46:28–43.

14. Mikocziova I, Greiff V, Sollid LM. Immunoglobulin germline gene variation and its impact on human disease. Genes Immun [Internet]. 2021 [cited 2021 Jul 8]; Available from: http://www.nature.com/articles/s41435-021-00145-5

15. Zhu Y, Tang H, Xie W, Chen S, Zeng H, Lan C, et al. The multilevel extensive diversity across the cynomolgus macaque captured by ultra-deep adaptive immune receptor repertoire sequencing. Sci Adv. 2024;10:eadj5640.

16. Colman K. Impact of the Genetics and Source of Preclinical Safety Animal Models on Study Design, Results, and Interpretation. Toxicol Pathol. 2017;45:94–106.

17. Doxiadis GGM, de Groot N, Otting N, de Vos-Rouweler AJM, Bolijn MJ, Heijmans CMC, et al. Haplotype diversity generated by ancient recombination-like events in the MHC of Indian rhesus macaques. Immunogenetics. 2013;65:569–84.

18. Karl JA, Bohn PS, Wiseman RW, Nimityongskul FA, Lank SM, Starrett GJ, et al. Major Histocompatibility Complex Class I Haplotype Diversity in Chinese Rhesus Macaques. G3 GenesGenomesGenetics. 2013;3:1195–201.

19. Karl JA, Graham ME, Wiseman RW, Heimbruch KE, Gieger SM, Doxiadis GGM, et al. Major histocompatibility complex haplotyping and long-amplicon allele discovery in cynomolgus macaques from Chinese breeding facilities. Immunogenetics. 2017;69:211–29.

20. Karl JA, Prall TM, Bussan HE, Varghese JM, Pal A, Wiseman RW, et al. Complete sequencing of a cynomolgus macaque major histocompatibility complex haplotype. Genome Res. 2023;33:448–62.

21. de Groot NG, de Groot N, de Vos-Rouweler AJM, Louwerse A, Bruijnesteijn J, Bontrop RE. Dynamic evolution of Mhc haplotypes in cynomolgus macaques of different geographic origins. Immunogenetics. 2022;74:409–29.

22. Shortreed CG, Wiseman RW, Karl JA, Bussan HE, Baker DA, Prall TM, et al. Characterization of 100 extended major histocompatibility complex haplotypes in Indonesian cynomolgus macaques. Immunogenetics. 2020;72:225–39.

23. Bruijnesteijn J, de Groot N, van der Wiel MKH, Otting N, de Vos-Rouweler AJM, de Groot NG, et al. Unparalleled Rapid Evolution of KIR Genes in Rhesus and Cynomolgus Macaque Populations. J Immunol. 2020;204:1770–86.

24. Prall TM, Karl JA, Varghese JM, Baker DA, Minor NR, Raveendran M, et al. Complete Genomic Assembly of Mauritian Cynomolgus Macaque Killer Ig-like Receptor and Natural Killer Group 2 Haplotypes. J Immunol. 2024;212:1754–65.

25. Conley HE, He MM, Easterhoff D, Kirshner HF, Cocklin SL, Meyer J, et al. Defining genetic diversity of rhesus macaque Fcγ receptors with long-read RNA sequencing. Front Immunol [Internet]. 2024 [cited 2024 Nov 17];14. Available from: https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2023.1306292/full

26. Fu L, Li X, Zhang W, Wang C, Wu J, Yang H, et al. A comprehensive profiling of T- and B-lymphocyte receptor repertoires from a Chinese-origin rhesus macaque by high-throughput sequencing. PLOS ONE. 2017;12:e0182733.

27. Brochu HN, Tseng E, Smith E, Thomas MJ, Jones AM, Diveley KR, et al. Systematic Profiling of Full-Length Ig and TCR Repertoire Diversity in Rhesus Macaque through Long Read Transcriptome Sequencing. J Immunol. 2020;204:3434–44.

28. Jaiswal S, Nyquist SK, Boyce S, Jivanjee T, Ibrahim S, Bromley JD, et al. Identification and characterization of the T cell receptor (TCR) repertoire of the cynomolgus macaque (Macaca Fascicularis). BMC Genomics. 2022;23:647.

29. Ramesh A, Darko S, Hua A, Overman G, Ransier A, Francica JR, et al. Structure and Diversity of the Rhesus Macaque Immunoglobulin Loci through Multiple De Novo Genome Assemblies. Front Immunol. 2017;8:1407.

30. Bernat NV, Corcoran M, Nowak I, Kaduk M, Dopico XC, Narang S, et al. Rhesus and cynomolgus macaque immunoglobulin heavy-chain genotyping yields comprehensive databases of germline VDJ alleles. Immunity. 2021;54:355-366.e4.

31. Cirelli KM, Carnathan DG, Nogal B, Martin JT, Rodriguez OL, Upadhyay AA, et al. Slow Delivery Immunization Enhances HIV Neutralizing Antibody and Germinal Center Responses via Modulation of Immunodominance. Cell. 2019;177:1153-1171.e28.

32. Yoo D, Rhie A, Hebbar P, Antonacci F, Logsdon GA, Solar SJ, et al. Complete sequencing of ape genomes [Internet]. bioRxiv; 2024 [cited 2024 Nov 17]. p. 2024.07.31.605654. Available from: https://www.biorxiv.org/content/10.1101/2024.07.31.605654v2

33. Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M, et al. A Novel Framework for Characterizing Genomic Haplotype Diversity in the Human Immunoglobulin Heavy Chain Locus. Front Immunol. 2020;11:2136.

34. Peres A, Lees WD, Rodriguez OL, Lee NY, Polak P, Hope R, et al. IGHV allele similarity clustering improves genotype inference from adaptive immune receptor repertoire sequencing data. Nucleic Acids Res. 2023;51:e86.

35. Lin M-J, Langmead B, Safonova Y. IGLoo: Profiling the Immunoglobulin Heavy chain locus in Lymphoblastoid Cell Lines with PacBio High-Fidelity Sequencing reads. bioRxiv. 2024;2024.07.20.604421.

36. Lees WD, Saha S, Yaari G, Watson CT. Digger: directed annotation of immunoglobulin and T cell receptor V, D, and J gene sequences and assemblies. Bioinformatics. 2024;40:btae144.

37. Corcoran MM, Phad GE, Bernat NV, Stahl-Hennig C, Sumida N, Persson MAA, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. Nat Commun. 2016;7:13642.

38. Safonova Y, Pevzner PA. De novo Inference of Diversity Genes and Analysis of Non-canonical V(DD)J Recombination in Immunoglobulins. Front Immunol. 2019;10:987.

39. 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.

40. Mikelov A, Nefedev G, Tashkeev A, Rodriguez OL, Ortmans DA, Skatova V, et al. Ultrasensitive allele inference from immune repertoire sequencing data with MiXCR. Genome Res. 2024;gr.278775.123.

41. Lees WD, Christley S, Peres A, Kos JT, Corrie B, Ralph D, et al. AIRR community curation and standardised representation for immunoglobulin and T cell receptor germline sets. ImmunoInformatics [Internet]. 2023 [cited 2024 Nov 17];10. Available from: https://www.immunoinformaticsjournal.com/article/S2667-1190(23)00005-8/fulltext

42. Rodriguez OL, Safonova Y, Silver CA, Shields K, Gibson WS, Kos JT, et al. Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. Nat Commun. 2023;14:4419.

43. He Y, Luo X, Zhou B, Hu T, Meng X, Audano PA, et al. Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants. Nat Commun. 2019;10:4233.

44. Yu G-Y, Mate S, Garcia K, Ward MD, Brueggemann E, Hall M, et al. Cynomolgus macaque (Macaca fascicularis) immunoglobulin heavy chain locus description. Immunogenetics. 2016;68:417–28.

45. Zhang S, Xu N, Fu L, Yang X, Li Y, Yang Z, et al. Comparative genomics of macaques and integrated insights into genetic variation and population history [Internet]. bioRxiv; 2024 [cited 2024 Nov 27]. p. 2024.04.07.588379. Available from: https://www.biorxiv.org/content/10.1101/2024.04.07.588379v1

46. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376:44–53.

47. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5.

48. Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, et al. Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol. 2022;40:1332–5.

49. 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;188:2151–62.

50. 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;92:530–46.

51. Ralph DK, Iv FAM. Consistency of VDJ Rearrangement and Substitution Parameters Enables Accurate B Cell Receptor Sequence Annotation. PLOS Comput Biol. 2016;12:e1004409.

52. MacDonald KS, Fowke KR, Kimani J, Dunand VA, Nagelkerke NJD, Blake Ball T, et al. Influence of HLA Supertypes on Susceptibility and Resistance to Human Immunodeficiency Virus Type 1 Infection. J Infect Dis. 2000;181:1581–9.

53. Gras S, Chen Z, Miles JJ, Liu YC, Bell MJ, Sullivan LC, et al. Allelic polymorphism in the T cell receptor and its impact on immune responses. J Exp Med. 2010;207:1555–67.

54. Pushparaj P, Nicoletto A, Sheward DJ, Das H, Dopico XC, Vidakovics LP, et al. Immunoglobulin germline gene polymorphisms influence the function of SARS-CoV-2 neutralizing antibodies. Immunity. 2023;56:193-206.e7.

55. deCamp AC, Corcoran MM, Fulp WJ, Willis JR, Cottrell CA, Bader DLV, et al. Human immunoglobulin gene allelic variation impacts germline-targeting vaccine priming. Npj Vaccines. 2024;9:1–13.

56. Zhu Y, Watson C, Safonova Y, Pennell M, Bankevich A. Assessing Assembly Errors in Immunoglobulin Loci: A Comprehensive Evaluation of Long-read Genome Assemblies Across Vertebrates [Internet]. bioRxiv; 2024 [cited 2024 Nov 27]. p. 2024.07.19.604360. Available from: https://www.biorxiv.org/content/10.1101/2024.07.19.604360v2

57. Bimber BN, Moreland AJ, Wiseman RW, Hughes AL, O'Connor DH. Complete characterization of killer Ig-like receptor (KIR) haplotypes in Mauritian cynomolgus macaques: novel insights into nonhuman primate KIR gene content and organization. J Immunol Baltim Md 1950. 2008;181:6301–8.

58. Budde ML, Lhost JJ, Burwitz BJ, Becker EA, Burns CM, O'Connor SL, et al. Transcriptionally abundant major histocompatibility complex class I alleles are fundamental to nonhuman primate simian immunodeficiency virus-specific CD8+ T cell responses. J Virol. 2011;85:3250–61.

59. Wiseman RW, Karl JA, Bohn PS, Nimityongskul FA, Starrett GJ, O'Connor DH. Haplessly hoping: macaque major histocompatibility complex made easy. ILAR J. 2013;54:196–210.

60. Prall TM, Graham ME, Karl JA, Wiseman RW, Ericsen AJ, Raveendran M, et al. Improved full-length killer cell immunoglobulin-like receptor transcript discovery in Mauritian cynomolgus macaques. Immunogenetics. 2017;69:325–39.

61. Li H. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 2021;37:4572–4.

62. Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, et al. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. Dev Comp Immunol. 2005;29:185–203.

63. Cowell LG, Davila M, Kepler TB, Kelsoe G. Identification and utilization of arbitrary correlations in models of recombination signal sequences. Genome Biol. 2002;3:research0072.1.

64. Merelli I, Guffanti A, Fabbri M, Cocito A, Furia L, Grazini U, et al. RSSsite: a reference database and prediction tool for the identification of cryptic Recombination Signal Sequences in human and murine genomes. Nucleic Acids Res. 2010;38:W262–7.

65. Porubsky D, Guitart X, Yoo D, Dishuck PC, Harvey WT, Eichler EE. SVbyEye: A visual tool to characterize structural variation among whole-genome assemblies [Internet]. bioRxiv; 2024

[cited 2024 Nov 27]. p. 2024.09.11.612418. Available from: https://www.biorxiv.org/content/10.1101/2024.09.11.612418v2

66. Vollger MR, Kerpedjiev P, Phillippy AM, Eichler EE. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. Bioinformatics. 2022;38:2049–51.

67. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab008.

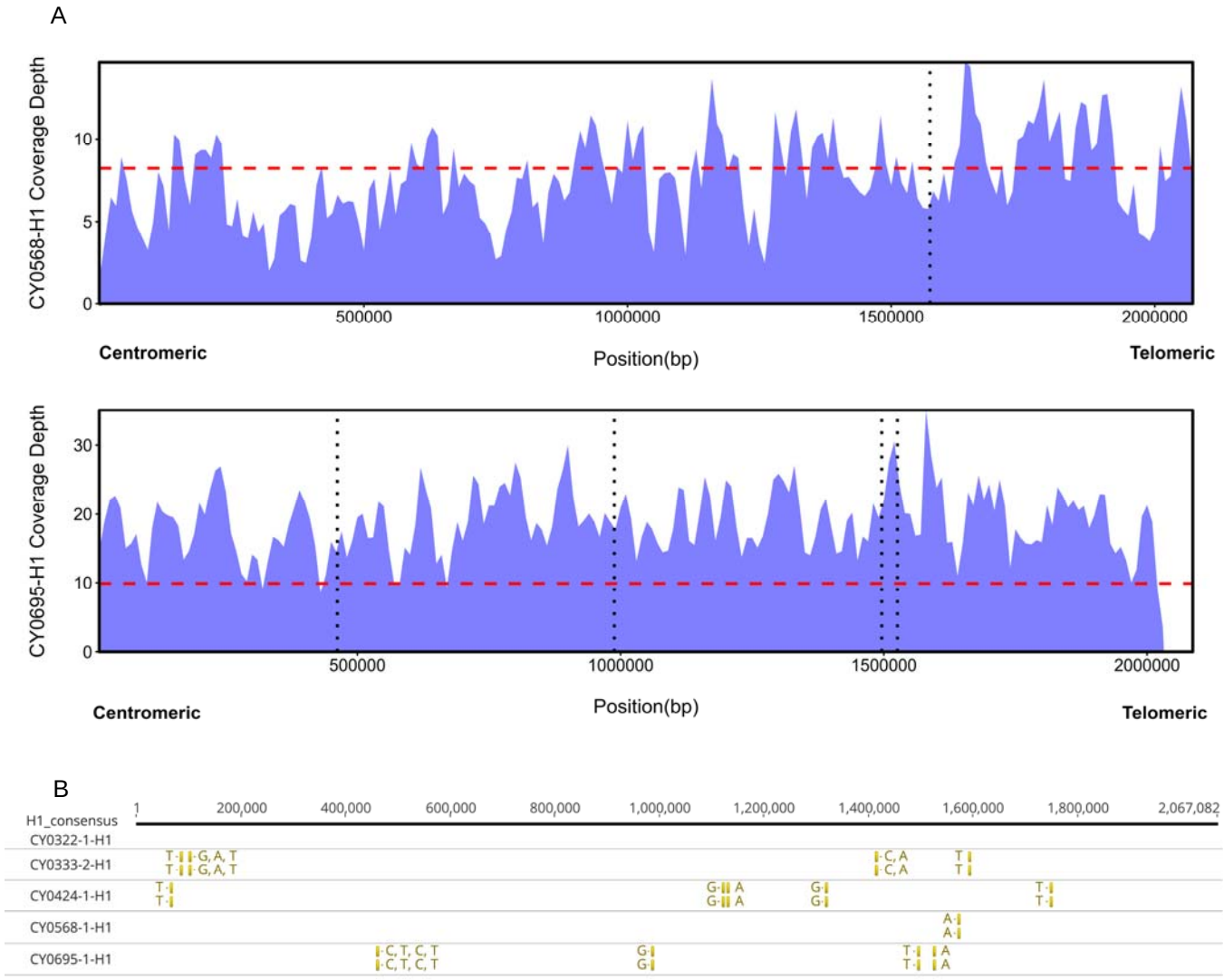


Figure 1. H1 is the predominant IGH haplotype found. (A) Read coverage plots of H1 assemblies from two homozygous animals. Dashed red lines show the average genomic read coverage for each animal. Vertical dotted lines indicate the positions of SNVs shown in panel B. (B) Five full-length, contiguous H1 assemblies differ at a total of just 20 sites across 2 megabases of sequence.

monkey	# of contigs	Length of contigs	IGH contig Coverage	Global coverage	Haplotypes
CY0161	2	2310648bp 1896742bp	7.62 7.36	7.20	H2/H5
CY0322	3	2015358bp 1458317bp 163458bp	11.07 11.03 8.72	10.31	H1/-
CY0325	3	2047163bp 625213bp 481108bp	10.58 7.70 10.51	11.04	H4/-
CY0330	4	425556bp 1197049bp 1791572bp 263190bp	13.42 15.02 13.31 12.64	14.21	H1/-
CY0333	2	2279552bp 2059666bp	14.18 14.13	12.84	H1/H7
CY0355	5	282686bp 1503951bp 1520213bp 456561bp 256504bp	8.02 7.69 7.35 7.97 6.31	7.49	-/-
CY0390	4	1850858bp 643229bp 264023bp 911198bp	15.56 13.44 15.41 10.26	14.11	-/-
CY0424	2	2048509bp 1843096bp	15.31 15.33	13.77	H1/H6
CY0558	4	629615bp 1264486bp 105347bp 1961649bp	7.36 6.90 4.19 6.93	7.07	H1/-
CY0568	1	2067082bp	7.35	8.22	H1/H1
CY0692	3	1035266bp 310717bp 905662bp	4.19 5.64 4.37	4.76	H2/H2
CY0695	1	2015353bp	10.28	9.89	H1/H1
CY0973	1	2483500bp	12.85	11.39	H3/H3

Table 1: IGH contigs assembled from long-read whole genome sequencing of 13 Mauritian-origin cynomolgus macaques.

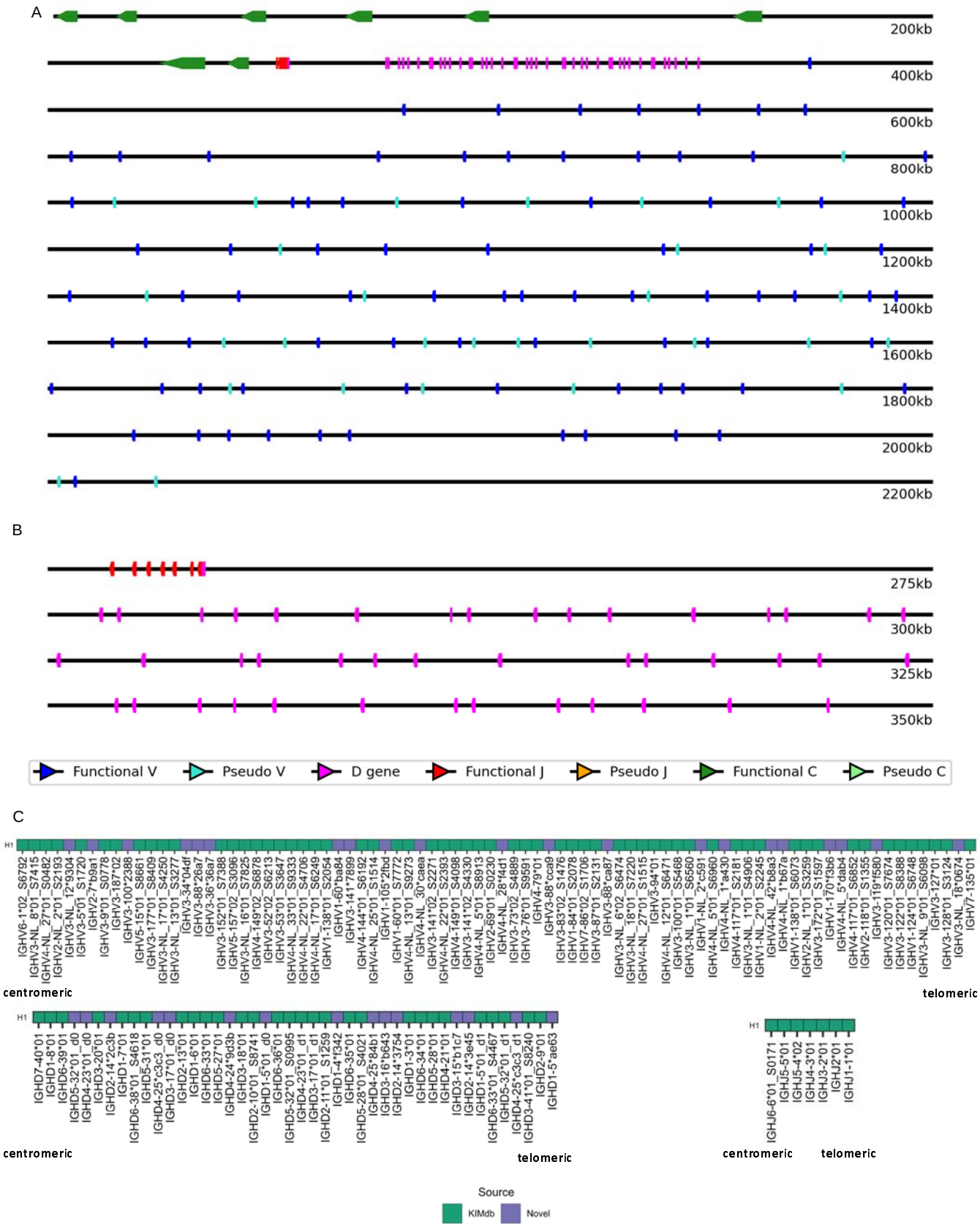


Figure 2. IGH haplotype H1 gene content. (A) Schematic representation of H1 showing the genomic positions of genes annotated by ALIGaToR. (B) Enlarged schematic view of the D-J cluster. (C) H1 V, D, and J genes annotated as functional colored by whether they correspond to alleles represented in the existing KIMdb database.

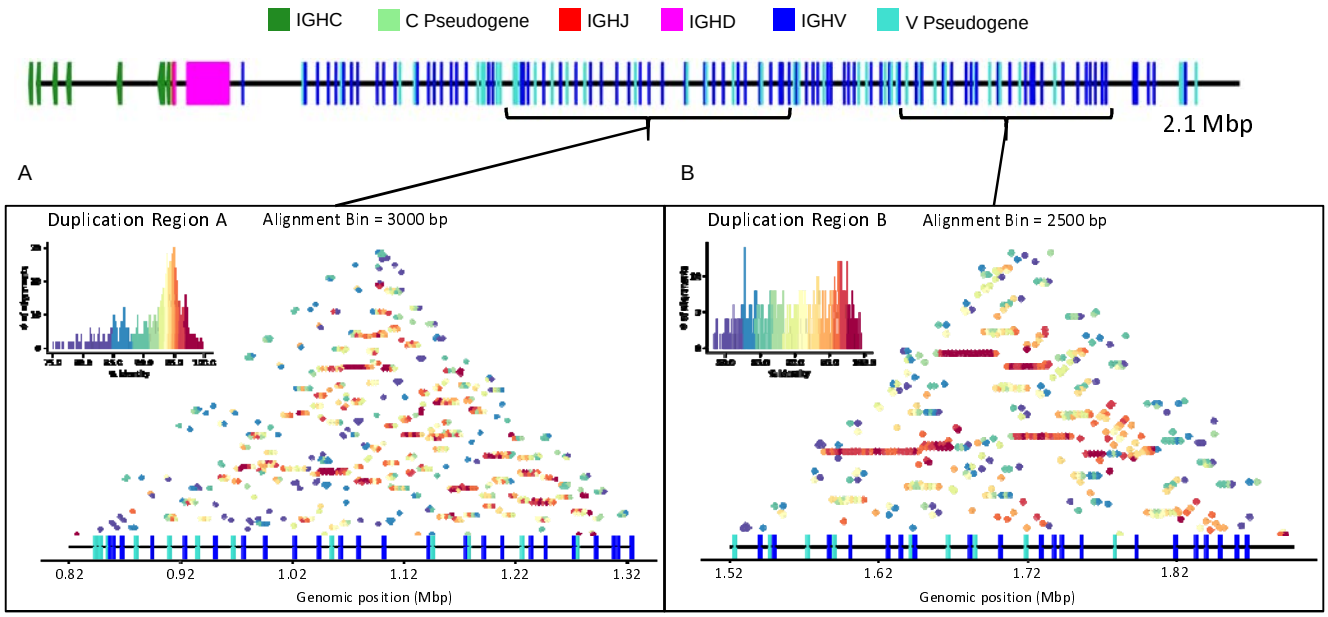


Figure 3. Regions enriched for duplications/repeats within H1. (A) Dot plot showing blocks of 10-15 kilobases with high homology. Gene track on the bottom shows 24 functional IGHV alleles in this region, many of which fall within duplicated block. (B) Dot plot of a more distal region with larger homologous blocks indicating more recent duplications. Gene track shows 20 functional IGHV alleles in the region. Insets show histogram of self-identity values within repeat regions.

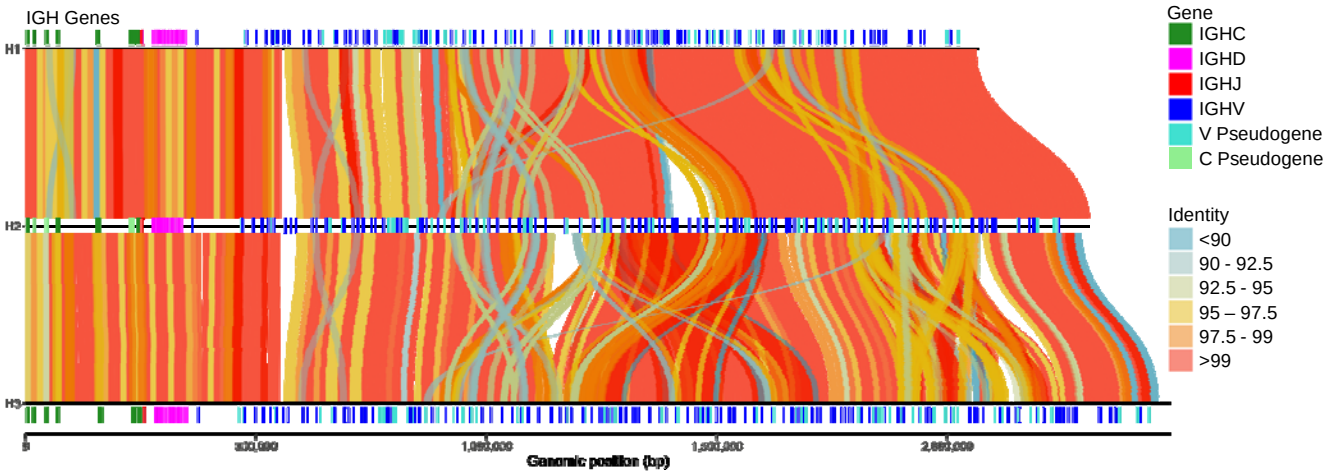


Figure 4. Significant diversity within MCM IGH haplotypes. Alignments of H1, H2, and H3 visualized using SVbyEye. Little structural variation is observed in the IGHC and IGHJ regions (left) with more complex events interspersed with regions of high homologoy.

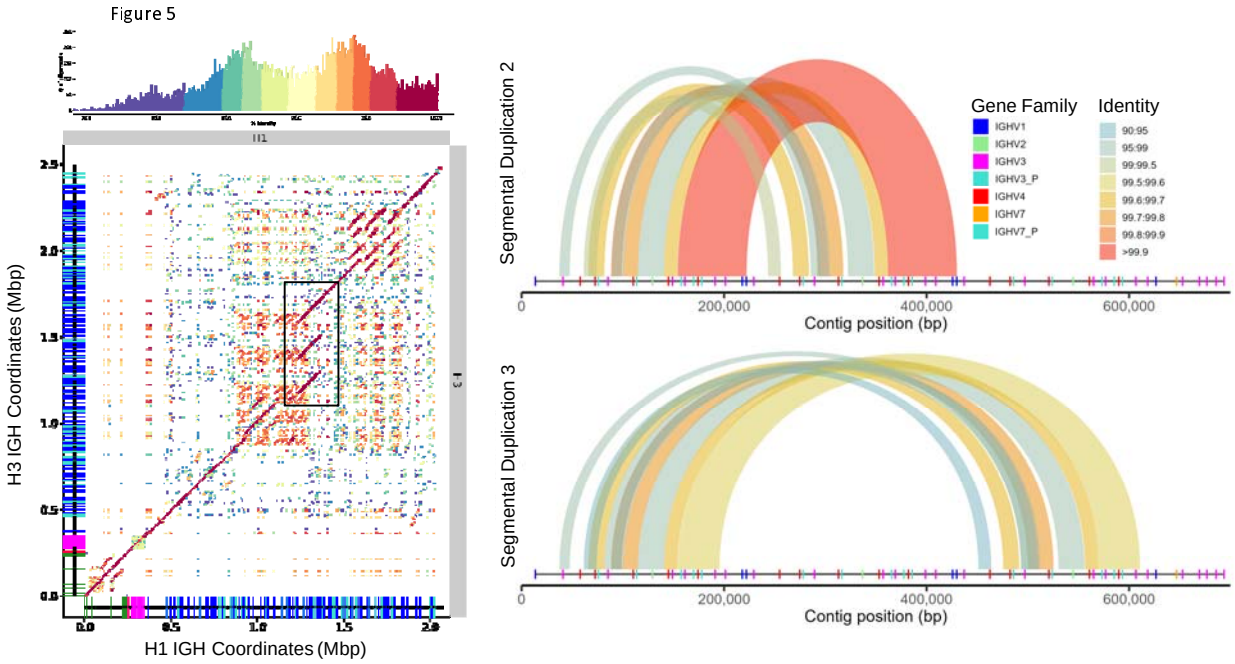


Figure 5. Haplotype H3 contains a large duplications compared to H1. (A) Dot plot comparing H1 (horizontal axis) to H3 (vertical axis). Box highlights a region of segmental duplication in H3, just distal of the H1 repeat region show in Fig 3a. (B) Arcs connect regions of high homology, as indicated by color.

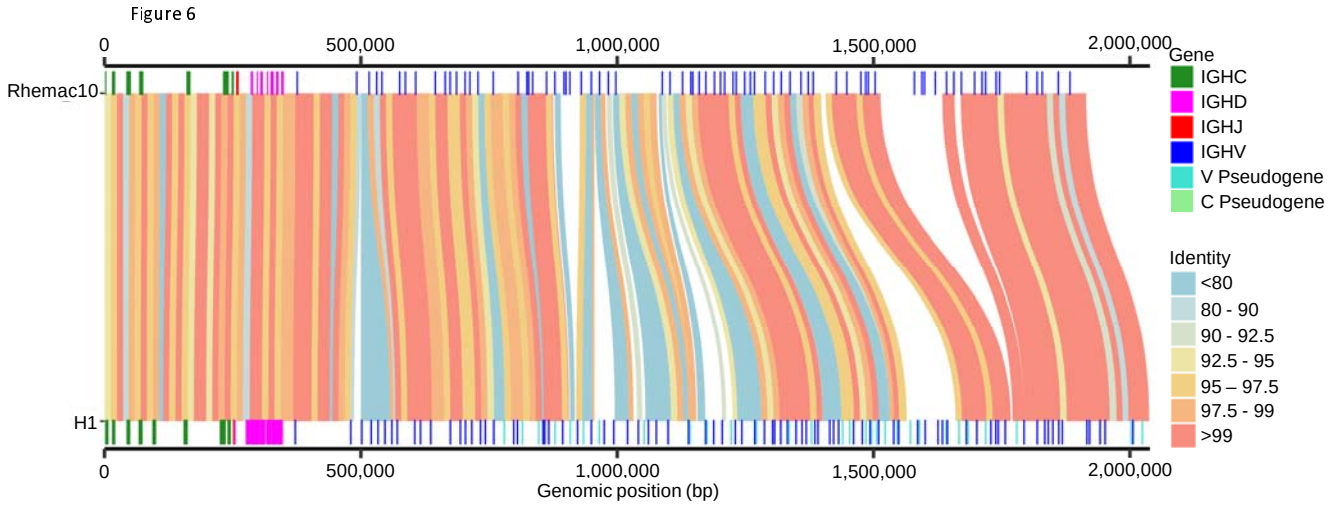


Figure 6. Comparison of IGH from MCM haplotype H1 to M_mul10 reference genome. H1 has approximately 100 kilobases of sequence not found in M_mul10, and overall sequence homology is only 83%.

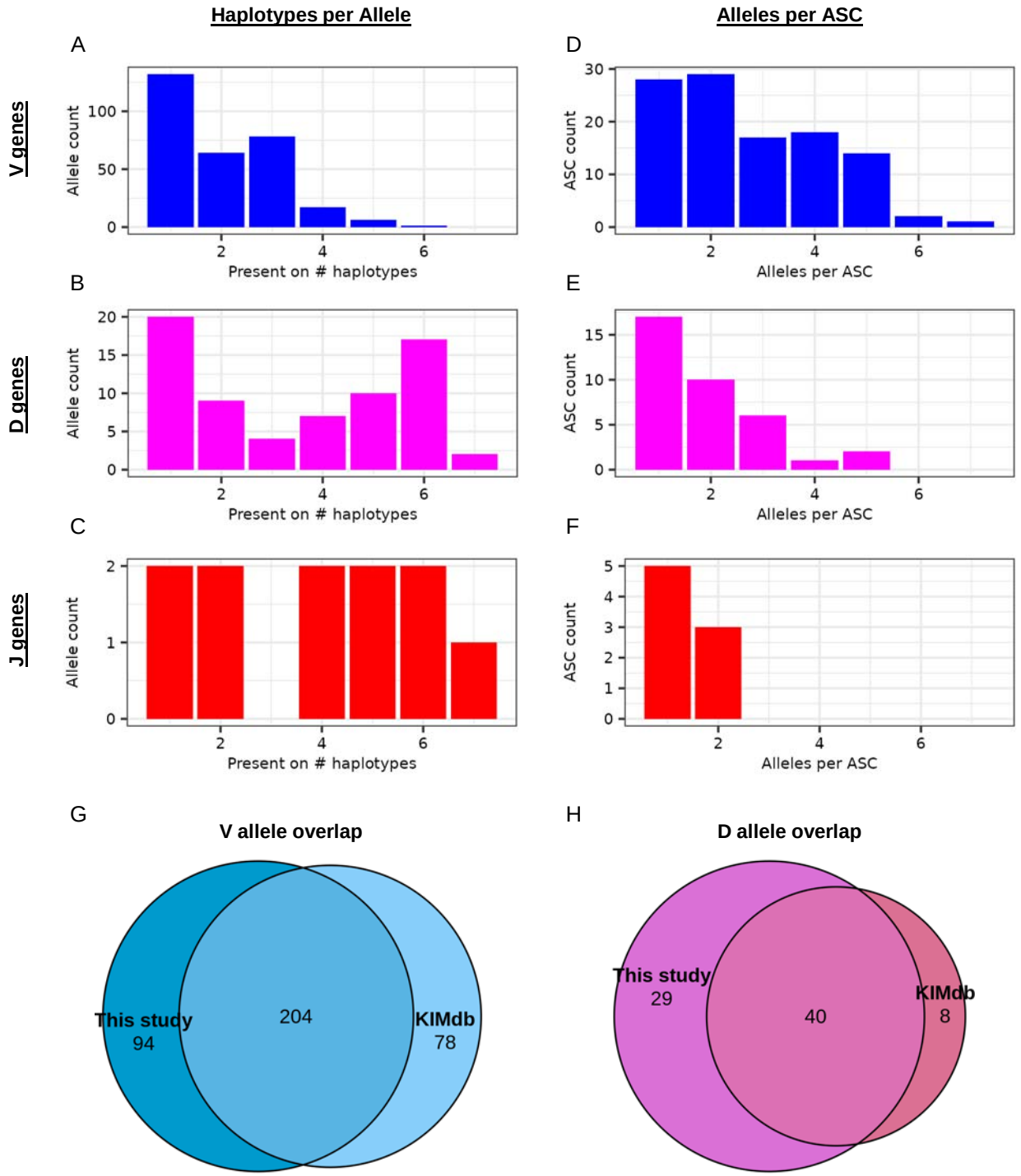
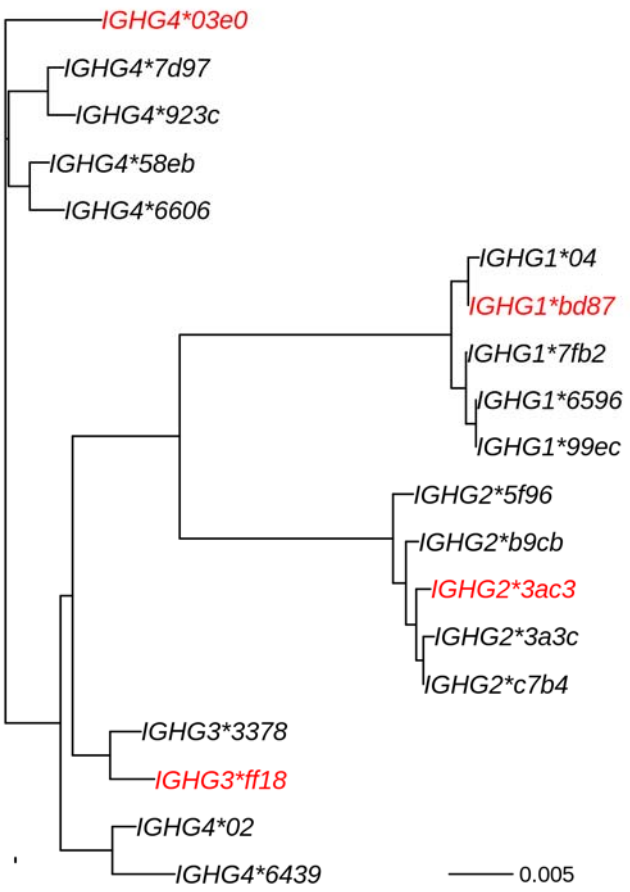


Figure 7. Genetic diversity among MCM haplotypes. (A) The plurality of IGHV alleles appear uniquely in a single haplotype, while (B) IGHD and especially (C) IGHJ alleles are more likely to be shared across most or all haplotypes. The large number of IGHV alleles shared across three haplotypes is due to extended blocks of shared sequence between H1, H2, and H4. The assembly for H7 does not cover most of the D-J cluster, leaving a maximum of six possible haplotypes over most of those regions. (D) IGHV genes are the most diverse, with up to eight alleles per ASC observed in this cohort. (E) IGHD genes and (F) IGHJ genes are less allelically diverse. (G) The IGHV alleles observed in this cohort largely overlap with those in existing databases, suggesting that 12 animals is sufficient to capture much of the immunogenetic diversity within MCM. (H) A greater proportion of IGHD alleles are unique to this study due to the difficulty of recovering germline IGHD gene sequences from rearranged IGs.

A



B

```
IGHG3*3378_H  EFTRPCDDTTPPCPPCP
IGHG3*ff18_H  EVTRPCDDTTPPCPPCP
IGHG4*02_H    EF-------TPPCPPCP
IGHG4*6439_H  EF-------TPPCPPCP
IGHG4*03e0_H  EF-------TPPCPPCP
IGHG4*58eb_H  EF-------TPPCPPCP
IGHG4*6606_H  EF-------TPPCPPCP
IGHG4*7d97_H  EF-------TPPCPPCP
IGHG4*923c_H  EF-------TPPCPPCP
```

Figure 8. Diversity within MCM IGH constant region isotypes. (A) Neighbor-joining tree for full-length alleles (CH1 through CHS) of all IgG subtypes found in this study, with the alleles present in H1 highlighted in red. (B) Two IgG3 alleles are named as IgG4 alleles by ALIGaToR due to short IgG4-like hinge exons.