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 Institute of Allergy and Infectious Diseases, 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 or chaim.schramm@nih.gov

Abstract

Background:

Nonhuman primates are valuable 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, termed H1, H2, and H3. IGH haplotype H1 was observed in two of the homozygotes and 5 additional heterozygous animals, accounting for half of the assemblies recovered. H1 shares only 83% average sequence identity with the IGH locus of the rhesus macaque reference genome, in addition to numerous large structural variations. The other two homozygous haplotypes exhibited considerable variation, including a 60 kilobase (Kbp) deletion and 200 Kbp insertion relative to H1. Furthermore, we annotated the IG gene content from all complete MCM IGH assemblies and found 288 functional IGHV alleles, of which 94 (33%) were not in existing databases. We also identified 68 functional IGHD alleles, 11 functional IGHJ alleles, and 33 functional constant gene alleles across all 5 isotypes.

Conclusions:

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.

Keywords: Mauritian cynomolgus macaque, immunogenetics, IGH

Background

Cynomolgus (*Macaca fascicularis*) and rhesus (*Macaca mulatta*) macaques share a close evolutionary history with humans and are widely used as biomedical models of human health [1–7]. Mauritian-origin cynomolgus macaques (MCMs) descend from a small (~20) founding population introduced to the island of Mauritius in the 16th century. As a result of this bottleneck event, MCMs exhibit significantly constrained genetic diversity compared to other cynomolgus and rhesus macaques [8]. This makes today's population of more than 50,000 MCM [8,9]

excellent model organisms especially in the context of immunological studies where germline variation can significantly impact results [10–12].

Immune loci such as the major histocompatibility complex (MHC) [13–18], killer cell immunoglobulin-like receptors (KIR) [19,20], and Fc-gamma receptors (FcγR) [21] all have limited genetic diversity in MCM. Relatively little is known about the T cell receptor (TR) [11,22–24] and IG loci [11,25,26] in MCM. The high density of segmental duplications and repetitive elements, as well as the presence of large structural variations, make accurate assembly of these loci difficult [25,27,28]. The IGH locus is particularly challenging and requires specialized sequencing and assembly methods alongside purpose-built annotation tools [27–33].

B cells produce receptors (BCR) and antibodies (Ab) to neutralize or destroy viruses, bacteria, aberrant cells, and other antigenic material. The generation of a diverse Ab repertoire relies on somatic rearrangement of genomic DNA in an individual B cell to create a complete BCR/Ab sequence, a process called V(D)J rearrangement. In the germline, individual IG genes, including variable (V), diversity (D), and joining (J) genes, are selected and joined together permanently altering the resident B cell's genome. V(D)J rearrangement in the IG heavy chain locus (IGH) generate a VDJ sequence corresponding to the variable region of an Ab heavy chain. The light chain loci, kappa and lambda, do not contain D genes and generate VJ recombination sequences which encode the Ab light chain variable region [34]. The genetic content and organization of the IG loci impact which genes are used most frequently in functional repertoires, and thus are of great interest for immunological studies [35–39].

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 (AIRR-seq) [40–43]. The Karolinska Macaque database (KIMDB) [26,27] contains 615 and 14 functional IGHV and IGHJ alleles, respectively, inferred from 18 cynomolgus macaques of

Indonesian and Mauritian origin. Of these, 279 IGHV and 11 IGHJ alleles were found specifically in MCM. Another recent study added an additional 84 inferred IGHV alleles from 3 Chinese-origin cynomolgus macaques [11]. The large number of reported IGH alleles from a relatively small number of sequenced individuals suggests the macaque IGH loci harbor significant diversity, most of which has not yet been characterized as is true for many other species [35,44,45].

Despite these successes, AIRR-seq-derived inferences have limited ability to resolve V, D, and J genes due to the removal of germline nucleotides during V(D)J rearrangement from the 3' end of V genes, 3' and 5' end of D genes, and 5' end of J genes. Additionally, most AIRR-seq-derived sequences capture only a small fraction of the constant (C) gene and are unable to reveal C gene polymorphisms which can impact function [46,47]. Furthermore, large structural variations within the IGH locus have been observed in macaques [25,27,29], similar to other primates [28], and humans [35]. High quality genomic references are thus essential for elucidating the complexity of the IGH loci in MCMs and accurate annotation of rearranged IGs. Additionally, the power of the MCM as a model organism relies on a comprehensive understanding of genetic diversity within the constrained population.

While there are multiple high quality, annotated assemblies of the IG loci for rhesus macaques [25,27,29,48], the only genomic IG reference for cynomolgus macaques is an unlocalized scaffold from a 2013 draft genome primarily generated using 2x100 bp paired-end short reads [49]. A new telomere-to-telomere cynomolgus reference genome has recently become available [50] 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 [26].

To fill in this gap, we leveraged long-read whole-genome sequencing data from 13 MCMs [20] to generate curated and annotated assemblies of the MCM IGH locus. 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 IGH [28,30,51]. We previously generated whole genome sequencing of 13 MCMs using a combination of Oxford Nanopore Technology and PacBio HiFi sequencing [20]. We assembled the reads for each animal using Hifiasm [52,53] and extracted the IGH locus from contigs that either partially or completely spanned the region. We recovered 13 scaffolds from 10 animals that spanned the entire IGH locus from IGHA to the most distal IGHV7 gene (Table 1). Notably, four animals were homozygous in IGH (Fig 1a and Fig S1).

The most common haplotype ("H1") was present in two of the homozygotes (CY0568 and CY0695) with relatively uniform read coverage and an average depth of 7.4x and 10.3x, respectively, across the IGH locus (Fig 1a). Large spikes in coverage occurred over repetitive elements due to off-target reads mapping to these regions. We identified five additional H1 haplotypes from heterozygous animals based on average sequence identity >99% and no structural variants greater than 2 Kbp. Two of these heterozygous H1 haplotypes in subjects CY330 and CY0558 were fragmented into two and three contigs, respectively (Table 1). We combined the three contiguous H1 haplotypes from heterozygotes 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 (Mbp)

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. In addition, haplotypes H2 and H3 were identified from additional homozygous MCM, and haplotypes H4, H5, and H6 were identified from heterozygous MCM (Table 1).

We used ALIGaToR [54] to annotate the IG genes present in H1 (Fig 2a-b). As per the current recommendations of the T-cell Receptor and Immunoglobulin Nomenclature Sub-Committee (TR-IG NSC) of the International Union of Immunological Societies [55], we have assigned temporary labels to these genes using IgLabel [44] (see Methods). Correspondences between temporary labels and the names currently used by KIMDB are available in Table S1. We found 80 functional, 3 ORF, and 40 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 (73%, 66%, and 100%, respectively) were previously annotated in KIMDB [26]. The novel alleles were distributed throughout the locus and across IGHV and IGHD gene families (Fig 2c). Notably, when compared to the IGHV alleles present in the existing IMGT cynomolgus genomic reference, only 69 (83%) of the H1 IGHV alleles had close orthologs (≥95% sequence identity; Fig S3), highlighting the need for updated allele databases fully capture the immunogenetic diversity of this species.

Differences between MCM and the rhesus Mmul_10 genomic reference

We conducted a comparative analysis of the H1 haplotype with the IGH locus of the rhesus macaque reference assembly Mmul_10 [29]. The Mmul_10 IGH locus contains 79 IGHV genes, 44 IGHD, 7 IGHJ, and 8 IGHC genes [56]. The H1 haplotype has one more IGHV genes, six fewer IGHD genes, the same number of IGHC genes and the same number of IGHJ genes

though two are designated as ORFs by ALIGaToR (Fig 3). Only 5 IGHV alleles are shared exactly between Mmul_10 and H1, while 7/75 (9%) Mmul_10 IGHV alleles and 11/83 (13%) H1 IGHV alleles share less than 95% sequence identity with any allele in the other haplotype (Fig S4). In addition, there are large structural variations between these haplotypes, consistent with comparisons between other primate species [28]. These structural variations are localized to the IGHV region of the locus and include the insertion or deletion of multiple functional genes, resulting in the H1 haplotype being 103 Kbp larger than the Mmul_10 haplotype. There are also large differences in nucleotide sequence between the H1 and Mmul_10 haplotypes with only 83% identity even in regions that can be aligned between species. Divergence time between rhesus and cynomolgus macaque is estimated at 0.49 to 0.90 million years ago [57]. To compare divergence between rhesus and cynomolgus macaque we aligned whole genome contigs from CY0695 (H1) to Mmul_10 and computed the frequency of single nucleotide variant (SNV) events across all Mmul_10 covered bases. The average frequency of SNV events across the genome is 4.93 bases per 1 Kbp while in IGH only the average frequency is 12.38 bases per 1 Kbp. These results suggest the accelerated divergence of the IGH locus between rhesus macaque and MCM relative to other genomic regions [58] which is consistent with the IGH region in other ape species [28].

Repetitive content within MCM IGH

To evaluate highly repetitive regions across H1 we used the *high multiplicity repeat content* (HMRC) metric to compute the fraction of positions that are covered by repeats of minimum length L at least N times [59]. Positions of repeats were computed by self-alignment of H1 using YASS [60] with a minimum repeat length of 15 Kbp and a minimum coverage of 5. Previous studies have shown that HMRC in the IGH locus can vary significantly across species [59]

[28,61]. In H1 we found five HMRC regions in total covering 367 Kbp (18%) of the locus (Fig 4a). Four of the HMRC regions cover a total of 14 V genes (17%) (Fig 4c-d) and a single HMRC region covers 38 (100%) of D genes (Fig 4b). High repeat density across D genes is present in other ape species and human [28,62]. The four HMRC regions across V genes fell in two larger repetitive clusters (Fig 4c-d). The proximal region (Fig 4c) spans ~450 Kbp, contains 24 functional IGHV alleles and two HMRC regions. The two HMRC regions span 44 Kbp and 30 Kbp respectively; the first contains one IGHV gene and one IGHV pseudogene, while the second contains two IGHV genes which differ from each other by 19 bases (93.6%) and 25 bases (95.8%) in the V exon. The distal region (Fig 4d) is ~400 Kbp and contains 20 functional IGHV genes and two HMRC regions. The first HMRC region is 155 Kbp and contains eight IGHV genes while the second is 69 Kbp and contains three IGHV genes. 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 [28,35,63,64]. This demonstrates the power and necessity of long-read sequencing approaches like the one used here, as short-read technologies often fail at resolving complex and highly repetitive regions like the IGH locus [30,65].

Structural variation between MCM IGH haplotypes

We then compared the H1 haplotype to two additional haplotypes, H2 and H3, resolved from homozygous MCM. H2 was distributed over 3 scaffolded contigs, but spanned the entire locus, while H3 was comprised of a single full-length contig. Similar to other studies of immunogenetic regions in these animals [16,20], we found a high degree of haplotype diversity (Fig 5). The H1 IGH haplotype is the shortest at 2.07 Mbp, while the H2 and H3 haplotypes are 2.3 Mbp and 2.48 Mbp 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 of 95-97.5% sequence identity and two large structural variations approximately 600 Kbp and 1,400 Kbp into the locus (Fig 5). The distal end of H1 and H2 have large blocks of highly homologous sequence with >99% identity (Fig 5). 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 5). Among regions of structural similarity across the IGH locus, the H1 haplotype has a sequence identity of 99.06% and 98.14% to H2 and H3, respectively.

Relative to H1, the H3 haplotype contains a ~60 Kbp deletion and a large ~400 Kbp insertion characterized by multiple segmental duplications in the proximal highly repetitive region (Fig 6a). While the segmental duplications in this region share greater than 99.99% identity, we were able to identify multiple ONT reads that spanned the segmental duplication breakpoints. In this region, H1 contains 26 functional IGHV genes while H3 contains 52. Among these, H1 includes two genes absent in H3 while H3 contains 14 genes not found in H1, as well as a tandem duplication containing 14 genes (28 total). No genes shared between H1 and H3 have identical sequences. Comparing this region across haplotypes H1, H2, H3, and H6 (see below) reveal a potential progression of deletions and duplications that contribute to the divergence between H1 and H3 (Fig 6b). The H2 haplotype retains all 23 genes found in this segment of H1 and includes an additional ~200 Kbp insertion containing 14 genes that match those in H3. All genes shared between H1 and H2 are sequence identical, with overlapping regions having ≥99.85% sequence identity. Haplotype H6, like H2, contains the ~200 Kbp insertion absent in H1 but also features a ~60 Kbp deletion that contains two genes in H1. Of the 38 genes overlapping

between H6 and H2, only 10 sequences are identical, and all lie within the ~200 Kbp insertion region. The H3 haplotype contains an additional ~200 Kbp duplication that significantly overlaps the insertion found in H2 and H6. Notably, H3 and H6 have 99.88% sequence identity in this region and all 39 distinct genes have identical sequence. In fact, the coding alleles of H3 and H6 are identical, outside of the deletion. This suggests a recent duplication, similar to that observed for the MHC class I *Maga-AG*/*Mafa-G* region in these animals [16]. The H3 and H6 haplotypes are 99.19% identical to H2 and 99.03% identical to H1 with the lowest sequence identity found in the regions immediately flanking the 60 Kbp deletion. Structural variations in the IGH region of humans [35–37] and mice [38,66,67] can significantly impact IGHV gene usage and V(D)J rearrangement. These results suggest that MCMs would be an ideal NHP model to test the impact of structural variation on V(D)J rearrangement and downstream response in therapeutic and vaccine studies, eg by investigating difference in the expressed repertoires between H3- and H6-homozygous animals.

Mapping the range of IGH genetic diversity within MCM

In addition to the complete H1 and H3 haplotypes recovered from homozygous animals, we also recovered a full-length H2 assembly and three additional fully contiguous haplotypes (H4, H5, and H6) from four heterozygous animals, each of which representing a unique haplotype (Table 1, Fig S5). We found that plurality of IGHV alleles (123/288, 43%) are uniquely found in a single haplotype (Fig 7a). An additional 113 alleles (39%) were found in exactly two haplotypes, due to the large blocks of high sequence identity in the IGHV region shared between H1/H2 and H3/H6 (Fig 5, Fig 6b, Fig S6). By contrast, IGHD alleles were predominately seen across all six haplotypes (18/68, 26%) or unique to a haplotype (20/68, 29%). (Fig 7b). No clear pattern was observed for IGHJ alleles (Fig 7c).

The 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 [30,62,68]. Instead, we used PIgLET [31] 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 97 IGHV, 36 IGHD, and 7 IGHJ ACSs across all six haplotypes. In general, IGHV ASCs contain multiple alleles per ASC, with 35 IGHV ASCs (36%) having four or more alleles (Fig 7d). In contrast, IGHD ASCs predominantly contained one or two alleles, with only six IGHD ASCs (17%) having four or more alleles (Fig 7e). The IGHJ ACSs followed a similar pattern, with three of the seven 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. Importantly, many IGHD alleles are duplicated, such that there are 68 unique sequences but 81 total alleles. Moreover, some of these duplications are conserved, with all 6 haplotypes carrying at least two identical copies of IGHD22-W6OT*01 and IGHD3-XNAE*05. Notably, IGHD coding alleles appear to be strongly conserved across macaques, with a previous study having found 29/54 (54%) total observed IGHD alleles to be expressed by all animals in all subsets of both Rhesus and Cynomolgus species [26].

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 [26]. Of 373 total IGHV alleles across both studies, only 194 (52%) were present in both groups of animals (Fig 7g). The 94 alleles that were found exclusively in the current cohort were uniformly distributed across all 6 haplotypes (Fig S6a) and across the locus

within each haplotype (Fig 2c). These likely have low usage in the expressed repertoires, which would have prevented their recovery by [26]. To determine if the 85 IGHV alleles present only in KIMDB might represent unmapped haplotypes, we annotated 5 unscaffolded partial IGH contigs (Table 1). This suggested the existence of two additional haplotypes (H7 and H8) and yielded 72 further IGHV alleles, including 41 KIMDB alleles that had not been captured on H1-H6 (Fig S6b). Furthermore, 40/44 remaining KIMDB IGHV alleles not present in the genomic data can be accounted for by pairs of alleles that differ only at position 320 or beyond, which are known to be difficult to infer from rearranged sequences [69] (Fig S6c). The high concordance between the two cohorts provides further evidence that 12 animals are sufficient to capture most of the restricted genetic diversity of MCM even in IGH.

Low overlap was also observed for IGHD alleles, with only 39/76 (51%) observed in both cohorts (Fig 7h), although with significantly more unique alleles in the current study than in the prior cohort. This imbalance likely reflects the difficulty of reconstructing germline IGHD gene sequences from repertoire sequencing data [41,70] rather than true genetic differences between the cohorts. Notably, the addition of the two partial haplotypes did not lead to any further shared alleles between the two databases, consistent with the differences in IGHD not being due to sampling.

We also considered allelic diversity in the constant chains. Eight IGHC genes were annotated in the expected order on all 10 fully spanning assemblies (Table 2). However, IgG4 was not automatically annotated by ALIGaToR in haplotypes H3 and H6 due to loss of the M2 exon. The observation that the H3 and H6 haplotypes share identical allelic variants across the IGHC region suggest that this is not a sequencing or assembly artifact. Further study will be required however, to determine if this represents a functional allele with a truncated intracellular domain or the pseudogenization of IgG4 in these haplotypes.

Within IgG, subtypes were labeled by genomic position, rather than sequence homology to other species or inferred function [71]. Similar to previous reports [71] human and macaque IgG clustered separately, with all macaque subtypes being more similar to each other than to any human subtype (Fig 8). Within the macaques, cynomolgus IgG1, IgG2, and IgG4 each clustered together with their rhesus ortholog. Notably, however, IgG3 did not form a distinct cluster, with the MCM alleles instead falling between rhesus IgG3 and cynomolgus IgG4. Additionally, only IGHG3*01, on the H1 and H5 haplotypes, is similar to cynomolgus IgG3 alleles previous documented by IMGT. The other haplotypes carry IGHG3*02, which is in the same relative position as IGHG3*01 but has a shorter hinge region and is closer in sequence to the IGHG4 alleles. Indeed, the coding sequence of IGHG3*02 here is identical to that of cynomolgus IGHG4*02 in the in current IMGT database. It is unclear if these IgGs function more similarly to the cynomolgus IgG4 alleles or the IGHG3*01 allele. Moreover, the diversity in this part of the genome is limited to only 2 total alleles, compared to 4-5 alleles present at each of the other IgG subtype genes. Together, these data suggest a complex evolutionary history in this region.

Discussion

Overall, we have provided the first full-length, curated genomic reference sequences 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 assess the immunogenetic diversity within the MCM population, as well. Together with other, immune receptor genes such as MHC, KIR, NKGC, and FCGR, IGs play an outsized role in the course of infectious disease and the response to vaccination [72–76]. Despite this, controlling for genetic variation in these highly polymorphic loci is extremely challenging in outbred species like primates. Crucially, we found that a single dominant haplotype was present

in half of the animals studied, highlighting the potential of MCMs 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, natural killer group 2 (NKG2) receptors, and Fcγ receptors in MCM [16,20,76]. Here, we assembled and characterized 6 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 3 MCM in this study are homozygous for a common NKG2 haplotype (N1) with a frequency of 40% [20]. 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. Moreover, knowledge of the specific IgG alleles carried by MCM will provide a fuller understanding of their interactions and binding affinity to the Fcγ receptors present in this subspecies [76].

Collectively, the haplotypes reported here identify a high total degree of sequence diversity moderated by the relative abundance of a common haplotype. Regions of high sequence homology between haplotypes can nonetheless contain structural variations (Fig 5), with both small (Fig 4) and large (Fig 6) segmental duplications in gene-rich regions. Haplotype diversity is even more extreme when compared to rhesus macaque (Fig 3), further signifying the importance of accurate genomic references to account for genetic variation. Structural variation in the IGH, IGK, and IGL loci of humans has been correlated with differential gene usage [35–37]. While SNV and structural variation in the IGH locus can significantly alter gene usage, the mechanisms and impact of genetic variation on the expressed repertoire is not well understood. The diverse haplotypes characterized from this genetically restricted population represent a powerful model to further explore the associations between germline and phenotype. For

example, a vaccine study involving two separate groups of genetically restricted MCMs can elucidate the impact of genetic features on vaccine response.

Our study is limited by the small number of animals sequenced and the limited sequencing depth achieved for some of the animals, especially for the ONT reads necessary to span large SVs and scaffold assemblies of the entire IGH locus. Thus, we fully assembled 13 loci, which, accounting for the 4 homozygous animals, cover 17/26 (65%) copies of the IGH locus. Analysis of partial assemblies and comparisons to an existing database indicated the presence of at least two additional haplotypes beyond the six analyzed here. In addition, 6/10 (60%) of the animals with at least one complete assembly are heterozygous, raising the possibility of assembly errors confounding our analyses [77]. Finally, we currently lack sufficient data to fully 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.

In addition to structural variation across haplotypes, more data are needed to distinguish between the likely ancestral IGH haplotypes of the MCM founder population and more recent recombinants. Previous work has identified 7, 8, 8 and 7 MCM haplotypes for MHC, KIR, FCGR and NKG2, respectively [76,78–81]. Each of these loci is on a different chromosome and therefore expected to segregate independently. Similarly, we currently describe 6 IGH haplotypes, with evidence for two additional haplotypes present in these animals (Table 1) and other MCM (Fig 7g, Fig S6b). Moreover, we found large blocks of high sequence homology between H1 and H2 (Fig 5, Fig 6b), as well as between H3 and H6 (Fig 6b). The restricted genetic diversity of this subspecies thus provides a tractable system for the analysis of genetic effects on the expressed repertoire and responses to vaccines and other immunological challenges.

Conclusions

In total, we annotated over 400 unique alleles across all IGH gene types, 132 (33%) of which are not present in a previous MCM database. Furthermore, 43% 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[35,45], 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 [20]) and assembled *de novo* with Hifiasm [52,53] using the command:

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

Contigs either partially or fully spanning the IGH locus were identified using BLAST+ with the Mmul_10 IGH locus as the query. Contigs spanning the centromeric edge of the IGH locus were trimmed using an anchor sequence ~2 Kbp centromeric of the IgA gene. To assess assembly

quality over the IGH locus HiFi and Nanopore reads were mapped back to the assembly contigs and IGH locus contigs were visually inspected [82]:

```
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 2022.1.1 (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 [83] 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 [54] was used to annotate the IGH assemblies starting from the IMGT [84] annotations of the Mmul_10 reference genome. Recombination signal sequences (RSS) were predicted based on RSS Information Content [85] using Dnagrep [86] in ALIGaToR. *M. fascicularis* IGHV, IGHD, and IGHJ coding sequences were downloaded from KIMDB [26] and filtered to only those identified as present in at least one MCM. These were combined with coding sequences of *M. fasciularis* IGHC from IMGT and used to identify shared 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
```

Piglet [31] was then use to separately cluster all unique coding alleles of V, D, and J genes across the 27 IGH contigs (Table 1) and the KIMDB MCM subset. V and J genes used `allele_cluster_threshold=95`, while D genes were clustered with `allele_cluster_threshold=90`. For V genes only, `trim_3prime_side=318` was used, while D and J had `trim_3prime_side=*NULL*`. For each ASC, IgLabel [44] was used to assign a single temporary label, with allele numbers added manually based on ASC membership. For IGHV and IGHD, family numbers are matched as much as possible to those reported in a recent immunogenetic survey of rhesus macaques [45], though individual ASCs were not screened for sequence homology to that database. For IGHC, allele numbers were assigned manually with *01 corresponding to the H1 haplotype and no screening for homology to alleles in the current IMGT database (of which only 3 were observed in these animals).

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.

Identifying High Multiplicity Repeat Content Regions

To identify regions of HMRC we first self-aligned the H1 haplotype to itself using YASS[60] using default settings and the -d 2 option for tabular blast format output. This table was then filtered to remove alignments less than L=15 Kbp and to remove self-alignments where query and target coordinates are equal. We then calculated the number of unique target alignments over every query region to determine the regions with coverage N=5 or greater and filtered out

the rest. Any overlapping regions were then merged. Due to the presence of telomeric sequence at the distal end of the H1 haplotype, 27,752 bp of telomeric sequence was masked from this analysis.

Structural Variation

Alignments between haplotypes were generated using minimap2 [82]:

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

SVbyEye [87] was used to visualize sequence alignments. A minimum alignment size cutoff of 9000, 13000, and 10000 were used to generate Fig 3, 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. The locus coordinates used for Fig 6 are H1:950,000-1,450,000, H2:989,287-1,697,206, H6:964,653-1,595,927, and H3:964,529-1,803,724. PatchWorkPlot [88] (https://github.com/yana-safonova/PatchWorkPlot) was used to generate dot plots of self-alignments (Fig 4). The command used was '`python PatchWorkPlot.py -i input_config.csv -o patchwork_output --cmap jet --reverse-cmap false`' with input fastas corresponding to the different haplotypes or regions of focus. StainedGlass [89] was used to generate dot plots of haplotype comparisons (Fig 6). In brief, sequences for comparison were indexed using samtools [83] and used as inputs for the StainedGlass snakemake pipeline using default parameters.

Declarations

Ethics approval and consent to participate

Animal selection PBMCs and splenocytes were obtained from 13 MCM housed at the Wisconsin National Primate Research Center during semiannual health checks or at necropsy. The animals were selected for analysis based on previously established MHC and KIR genotypes [78,81,90,91]. Sampling was performed in concordance with protocols approved by the University of Wisconsin-Madison Institutional Animal Care and Use Committee, as well as guidelines contained within the Animal Welfare Act, the Guide for Care and Use of Laboratory Animals, and the Weatherall report [92]

Consent for publication

Not applicable

Availability of data and materials

Genomic sequencing for this study was downloaded from NCBI BioProject with accession number PRJNA854973. Assembled IGH contigs and annotations have been deposited in Genbank with accession numbers PX411181-PX411201.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

This work was also supported in part by contract 75N93021C00006 to David O'Connor from the National Institute of Allergy and Infectious Disease, National Institutes of Health.

Authors' contributions

Conceptualization: RWW, DHO, DCD, CAS

Data curation: SO, WSG, TMP, CAS

Formal Analysis: SO, WSG, CAS

Investigation: TMP, JAK

Supervision: RWW, DHO, DCD, CAS

Visualization: SO, WSG, CAS

Writing – original draft: SO, WSG, CAS

Writing – review & editing: all authors

Acknowledgements

Not applicable

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. 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.

512 9. Osada N, Hettiarachchi N, Adeyemi Babarinde I, Saitou N, Blancher A. Whole-genome sequencing of
513 six Mauritian Cynomolgus macaques (Macaca fascicularis) reveals a genome-wide pattern of
514 polymorphisms under extreme population bottleneck. Genome Biol Evol. 2015;7:821–30.

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

518 11. Zhu Y, Tang H, Xie W, Chen S, Zeng H, Lan C, et al. The multilevel extensive diversity across the
519 cynomolgus macaque captured by ultra-deep adaptive immune receptor repertoire sequencing. Sci Adv.
520 2024;10:eadj5640.

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

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

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

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

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

534 17. de Groot NG, de Groot N, de Vos-Rouweler AJM, Louwerse A, Bruijnesteijn J, Bontrop RE. Dynamic
535 evolution of Mhc haplotypes in cynomolgus macaques of different geographic origins. Immunogenetics.
536 2022;74:409–29.

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

540 19. Bruijnesteijn J, de Groot N, van der Wiel MKH, Otting N, de Vos-Rouweler AJM, de Groot NG, et al.
541 Unparalleled Rapid Evolution of KIR Genes in Rhesus and Cynomolgus Macaque Populations. J Immunol.
542 2020;204:1770–86.

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

546 21. Conley HE, He MM, Easterhoff D, Kirshner HF, Cocklin SL, Meyer J, et al. Defining genetic diversity of
547 rhesus macaque Fcγ receptors with long-read RNA sequencing. Front Immunol [Internet]. 2024 [cited
548 2024 Nov 17];14. Available from:
549 https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2023.1306292/full

550 22. Fu L, Li X, Zhang W, Wang C, Wu J, Yang H, et al. A comprehensive profiling of T- and B-lymphocyte
551 receptor repertoires from a Chinese-origin rhesus macaque by high-throughput sequencing. PLOS ONE.
552 2017;12:e0182733.

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

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

559 25. Ramesh A, Darko S, Hua A, Overman G, Ransier A, Francica JR, et al. Structure and Diversity of the
560 Rhesus Macaque Immunoglobulin Loci through Multiple De Novo Genome Assemblies. Front Immunol.
561 2017;8:1407.

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

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

568 28. Yoo D, Rhie A, Hebbar P, Antonacci F, Logsdon GA, Solar SJ, et al. Complete sequencing of ape
569 genomes. Nature [Internet]. 2025 [cited 2025 Apr 10]; Available from:
570 https://www.nature.com/articles/s41586-025-08816-3

571 29. Warren WC, Harris RA, Haukness M, Fiddes IT, Murali SC, Fernandes J, et al. Sequence diversity
572 analyses of an improved rhesus macaque genome enhance its biomedical utility. Science.
573 2020;370:eabc6617.

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

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

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

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

584 34. Tonegawa S. Somatic generation of antibody diversity. Nature. 1983;302:575–81.

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

587 36. Rodriguez OL, Qiu X, Shields K, Dunn C, Singh A, Kaileh M, et al. Human genetic variation shapes the
588 antibody repertoire across B cell development [Internet]. Cold Spring Harbor Laboratory; 2025 [cited
589 2025 Jul 11]. Available from: http://biorxiv.org/lookup/doi/10.1101/2025.05.19.654982

590 37. Engelbrecht E, Rodriguez OL, Lees W, Vanwinkle Z, Shields K, Schultze S, et al. Germline
591 polymorphism in the immunoglobulin kappa and lambda loci explain variation in the expressed light
592 chain antibody repertoire [Internet]. Cold Spring Harbor Laboratory; 2025 [cited 2025 Jul 11]. Available
593 from: http://biorxiv.org/lookup/doi/10.1101/2025.05.28.656470

594 38. Bhat KH, Priyadarshi S, Naiyer S, Qu X, Farooq H, Kleiman E, et al. An Igh distal enhancer modulates
595 antigen receptor diversity by determining locus conformation. Nat Commun. 2023;14:1225.

596 39. Barajas-Mora EM, Lee L, Lu H, Valderrama JA, Bjanes E, Nizet V, et al. Enhancer-instructed epigenetic
597 landscape and chromatin compartmentalization dictate a primary antibody repertoire protective against
598 specific bacterial pathogens. Nat Immunol. 2023;24:320–36.

599 40. Corcoran MM, Phad GE, Bernat NV, Stahl-Hennig C, Sumida N, Persson MAA, et al. Production of
600 individualized V gene databases reveals high levels of immunoglobulin genetic diversity. Nat Commun.
601 2016;7:13642.

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

604 42. Gadala-Maria D, Gidoni M, Marquez S, Vander Heiden JA, Kos JT, Watson CT, et al. Identification of
605 Subject-Specific Immunoglobulin Alleles From Expressed Repertoire Sequencing Data. Front Immunol.
606 2019;10:129.

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

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

613 45. Peres A, Upadhyay AA, Klein V, Saha S, Rodriguez OL, Vanwinkle ZM, et al. A Broad Survey and
614 Functional Analysis of Immunoglobulin Loci Variation in Rhesus Macaques [Internet]. bioRxiv; 2025 [cited

615 2025 Jan 28]. p. 2025.01.07.631319. Available from:
616 https://www.biorxiv.org/content/10.1101/2025.01.07.631319v1

617 46. Ford EE, Tieri D, Rodriguez OL, Francoeur NJ, Soto J, Kos JT, et al. FLAIRR-Seq: A Method for Single-
618 Molecule Resolution of Near Full-Length Antibody H Chain Repertoires. J Immunol. 2023;ji2200825.

619 47. Ko S, Park S, Sohn MH, Jo M, Ko BJ, Na J-H, et al. An Fc variant with two mutations confers prolonged
620 serum half-life and enhanced effector functions on IgG antibodies. Exp Mol Med. 2022;54:1850–61.

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

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

625 50. Zhang S, Xu N, Fu L, Yang X, Li Y, Yang Z, et al. Comparative genomics of macaques and integrated
626 insights into genetic variation and population history [Internet]. bioRxiv; 2024 [cited 2024 Nov 27]. p.
627 2024.04.07.588379. Available from: https://www.biorxiv.org/content/10.1101/2024.04.07.588379v1

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

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

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

634 54. Schramm CA, Olubo S, Douek DC. ALIGaToR: A tool for the automated annotation of immunoglobulin
635 and T cell receptor genomic loci [Internet]. bioRxiv; 2025 [cited 2025 Jul 21]. p. 2025.05.02.651960.
636 Available from: https://www.biorxiv.org/content/10.1101/2025.05.02.651960v1

637 55. T-cell Receptor and Immunoglobulin Nomenclature Sub-Committee. Data Review and Naming
638 [Internet]. International Union of Immunological; 2023 [cited 2025 Jul 21]. Available from: https://wp-
639 iuis.s3.eu-west-
640 1.amazonaws.com/app/uploads/2024/10/04120932/Data.Review.and_.Naming.Dec2023.pdf

641 56. Nguefack Ngoune V, Bertignac M, Georga M, Papadaki A, Albani A, Folch G, et al. IMGT® Biocuration
642 and Analysis of the Rhesus Monkey IG Loci. Vaccines. 2022;10:394.

643 57. Ito T, Kanthaswamy S, Bunlungsup S, Oldt RF, Houghton P, Hamada Y, et al. Secondary contact and
644 genomic admixture between rhesus and long-tailed macaques in the Indochina Peninsula. J Evol Biol.
645 2020;33:1164–79.

646 58. Zhang S, Xu N, Fu L, Yang X, Ma K, Li Y, et al. Integrated analysis of the complete sequence of a
647 macaque genome. Nature. 2025;640:714–21.

648 59. Pospelova M, Voss K, Zamyatin A, Watson CT, Koepfli K-P, Bankevich A, et al. Comparative Analysis of
649 Mammalian Adaptive Immune Loci Revealed Spectacular Divergence and Common Genetic Patterns. Mol
650 Biol Evol. 2025;42:msaf152.

651 60. Noe L, Kucherov G. YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Res.
652 2005;33:W540–3.

653 61. Sepulveda MA, Garrett FE, Price-Whelan A, Birshtein BK. Comparative analysis of human and mouse
654 3′ *Igh* regulatory regions identifies distinctive structural features. Mol Immunol. 2005;42:605–15.

655 62. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, et al. Complete Haplotype
656 Sequence of the Human Immunoglobulin Heavy-Chain Variable, Diversity, and Joining Genes and
657 Characterization of Allelic and Copy-Number Variation. Am J Hum Genet. 2013;92:530–46.

658 63. Das S, Nozawa M, Klein J, Nei M. Evolutionary dynamics of the immunoglobulin heavy chain variable
659 region genes in vertebrates. Immunogenetics. 2008;60:47–55.

660 64. Pospelova M, Voss K, Zamyatin A, Watson CT, Koepfli K-P, Bankevich A, et al. Comparative analysis of
661 mammalian adaptive immune loci revealed spectacular divergence and common genetic patterns
662 [Internet]. 2025 [cited 2025 Apr 11]. Available from:
663 http://biorxiv.org/lookup/doi/10.1101/2025.04.01.646651

664 65. Watson CT, Breden F. The immunoglobulin heavy chain locus: genetic variation, missing data, and
665 implications for human disease. Genes Immun. 2012;13:363–73.

666 66. Rogers CH, Mielczarek O, Corcoran AE. Dynamic 3D Locus Organization and Its Drivers Underpin
667 Immunoglobulin Recombination. Front Immunol. 2021;11:633705.

668 67. Lin SG, Guo C, Su A, Zhang Y, Alt FW. CTCF-binding elements 1 and 2 in the *Igh* intergenic control
669 region cooperatively regulate V(D)J recombination. Proc Natl Acad Sci. 2015;112:1815–20.

670 68. Matsuda F, Ishii K, Bourvagnet P, Kuma K, Hayashida H, Miyata T, et al. The Complete Nucleotide
671 Sequence of the Human Immunoglobulin Heavy Chain Variable Region Locus. J Exp Med.
672 1998;188:2151–62.

673 69. Thörnqvist L, Ohlin M. The functional 3′-end of immunoglobulin heavy chain variable (IGHV) genes.
674 Mol Immunol. 2018;96:61–8.

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

677 71. Boesch AW, Osei-Owusu NY, Crowley AR, Chu TH, Chan YN, Weiner JA, et al. Biophysical and
678 Functional Characterization of Rhesus Macaque IgG Subclasses. Front Immunol [Internet]. 2016 [cited
679 2025 Jul 23];7. Available from:
680 https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2016.00589/full

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

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

686 74. Pushparaj P, Nicoletto A, Sheward DJ, Das H, Dopico XC, Vidakovics LP, et al. Immunoglobulin
687 germline gene polymorphisms influence the function of SARS-CoV-2 neutralizing antibodies. Immunity.
688 2023;56:193-206.e7.

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

691 76. Haj AK, Arbanas JM, Yamniuk AP, Karl JA, Bussan HE, Drinkwater KY, et al. Characterization of
692 Mauritian Cynomolgus Macaque FcγR Alleles Using Long-Read Sequencing. J Immunol. 2019;202:151–9.

693 77. Zhu Y, Watson C, Safonova Y, Pennell M, Bankevich A. CloseRead: a tool for assessing assembly errors
694 in immunoglobulin loci applied to vertebrate long-read genome assemblies. Genome Biol. 2025;26:131.

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

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

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

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

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

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

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

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

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

717 87. Porubsky D, Guitart X, Yoo D, Dishuck PC, Harvey WT, Eichler EE. SVbyEye: A visual tool to
718 characterize structural variation among whole-genome assemblies [Internet]. bioRxiv; 2024 [cited 2024
719 Nov 27]. p. 2024.09.11.612418. Available from:
720 https://www.biorxiv.org/content/10.1101/2024.09.11.612418v2

721 88. Pospelova M, Safonova Y. PatchWorkPlot: simultaneous visualization of local alignments across
722 multiple sequences [Internet]. arXiv; 2025 [cited 2025 Apr 28]. Available from:
723 http://arxiv.org/abs/2503.20766

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

726 90. O'Connor SL, Lhost JJ, Becker EA, Detmer AM, Johnson RC, Macnair CE, et al. MHC heterozygote
727 advantage in simian immunodeficiency virus-infected Mauritian cynomolgus macaques. Sci Transl Med.
728 2010;2:22ra18.

729 91. Ericsen AJ, Starrett GJ, Greene JM, Lauck M, Raveendran M, Deiros DR, et al. Whole genome
730 sequencing of SIV-infected macaques identifies candidate loci that may contribute to host control of
731 virus replication. Genome Biol. 2014;15:478.

732 92. Weatherall D. The use of non-human primates in research [Internet]. London: Academy of Medical
733 Sciences; 2006 [cited 2025 Jul 23]. Available from: https://royalsociety.org/-
734 /media/policy/publications/2006/weatherall-report.pdf

Figure Legends

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.

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. Correspondences between temporary labels and names in KIMdb can be found in Table S1.

Figure 3. 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 4. Regions enriched for duplications/repeats within H1. (A) Locus track of the H1 IGH locus. The left-most track are all functional genes outside of HMRC (high multiplicity repeat content) regions (yellow). The right-most track are all genes within HMRC regions. (B) Dot plot of the HMRC region spanning all 38 functional IGHD genes. The dot plot colors correspond to sequence identity. (C) Dot plot of the proximal repetitive region containing two HMRC blocks.

The dot plot covers 24 functional IGHV alleles in this region with 3 inside HMRC regions highlighted by a yellow bar. (D) Dot plot of a more distal region with larger homologous blocks and HMRC regions. The dot plot covers 20 functional IGHV alleles in this region with 11 inside HMRC regions highlighted by a yellow bar.

Figure 5. 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 homology. The underlined region of the H3 haplotype represents a segmental duplication relative to H2.

Figure 6. Haplotype H3 contains a large duplications compared to H1. (A) Dot plot comparing H1 (horizontal axis) to H3 (vertical axis). Box (H1:950,000-1,450,000 bp) highlights a region of segmental duplication in H3, just distal of the H1 repeat region show in Fig 3a. The histogram shows the frequency of bins at each identity level, with colors corresponding to colors in the dot plot. (B) Dot plots (left) and Miro plot (right) comparisons of highlighted segmental duplication in haplotypes H1, H2, H3, H6. From top to bottom; H1 is compared to H2, H2 is compared to H6, and H6 is compared to H3. For Dot plots, the top comparison is on the Y-axis and the bottom comparison is on the X-axis. For Miro plots, genes for each haplotype are marked by lines and colored based on sequence identity to comparison haplotype. Blocks of structural variation are highlighted with dotted lines demarcating region of duplication in H3.

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 between two

haplotypes is due to extended blocks of shared sequence between H1 and H2, as well as between H3 and H6 (see Fig 6b). (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) Overlap between the IGHV alleles observed in this cohort and those from a previous cohort [26]. See also Fig S6. (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.

Figure 8. Diversity of MCM IgG constant region subtypes compared to rhesus macaque and humans. Neighbor-joining tree for full-length alleles (CH1 through CHS) of all IgG subtypes found in this study, together with representative alleles from human (purple) and rhesus (green) IgG subtypes. MCM alleles present in H1 are highlighted in red. All other MCM alleles are shown in black.

Tables

Table 1. IGH contigs assembled from long-read whole genome sequencing of 13 Mauritian-origin cynomolgus macaques. Asterisk denotes homozygous contig that completely spans IGH. Partial assemblies for a single haplotype are ordered from centromeric to telomeric. Parentheses indicate unscaffolded partial assemblies. Dash denotes unclassified or unresolved haplotype. IGH contig coverage is the mean read depth for each contig, contig coverage is the mean read depth across all contigs, and whole genome coverage is calculated as total base pairs mapped divided by 3 Gbp (the approximate size of the MCM genome).

Subject	Haplotype	# of IGH contigs	Genbank Accession	IGH contig length (Kbp)	Mean IGH contig HIFI coverage	Mean genome contig HIFI coverage	Mean whole genome HIFI coverage
CY0568	H1*	1		2,067	7.4	8.2	17.9
CY0695	H1*	1		2,015	10.3	9.9	20.3
CY0973	H3*	1		2,483	12.9	11.4	26.3
CY0692	H2	3		310 1,035 906	5.6 4.2 4.4	4.8	8.0
CY0161	H2 H4	1 1		2,310 1,896	7.6 7.4	7.2	15.3
CY0424	H1 H5	1 1		2,049 1,843	15.3 15.3	13.8	30.5
CY0333	H1 H6	1 1		2,060 2,280	14.1 14.2	12.8	28.5
CY0558	H1 (H7)	3 1		630 1,264 105 1,962	7.4 6.9 4.2 6.9	7.1	15.5
CY0390	(H2) (H7)	1 2		1,851 911 264	15.6 10.3 15.4	14.1	30.9
CY0322	H1 (H8)	1 2		2,015 1,458 163	11.1 11.0 8.7	10.3	21.4
CY0330	H1 (H8)	2 2		1,791 263 426 1,197	13.3 12.6 13.4 15.0	14.2	29.8
CY0325	-	3		2,047 625 481	10.6 7.7 10.5	11.0	21.1

			283	8.0			
			1,504	7.7			
CY0355	-	5	1,520	7.4	7.5	16.2	
			457	8.0			
			257	6.3			

Table 2. IG constant region alleles identified on 10 fully spanning contigs. All eight expected constant alleles were detected in all fully spanning assemblies. Text color indicates the allele number of each gene. For haplotypes H3 and H6, IgG4 was manually annotated (green shading) after failing to be detected by ALIGaToR due to missing M2 exons. Bold font indicates homozygous animals. See also Table 1.

Animal	Hap	IgM	IgD	IgG1	IgG2	IgG3	IgG4	IgE	IgA
CY0568	**H1**	**IGHM*01**	**IGHD*01**	**IGHG1*01**	**IGHG2*01**	**IGHG3*01**	**IGHG4*01**	**IGHE*01**	**IGHA*01**
CY0695	**H1**	**IGHM*01**	**IGHD*01**	**IGHG1*01**	**IGHG2*01**	**IGHG3*01**	**IGHG4*01**	**IGHE*01**	**IGHA*01**
CY0322	H1	IGHM*01	IGHD*01	IGHG1*01	IGHG2*01	IGHG3*01	IGHG4*01	IGHE*01	IGHA*01
CY0333	H1	IGHM*01	IGHD*01	IGHG1*01	IGHG2*01	IGHG3*01	IGHG4*01	IGHE*01	IGHA*01
CY0424	H1	IGHM*01	IGHD*01	IGHG1*01	IGHG2*01	IGHG3*01	IGHG4*01	IGHE*01	IGHA*01
CY0161	H2	IGHM*01	IGHD*01	IGHG1*02	IGHG2*02	IGHG3*02	IGHG4*02	IGHE*01	IGHA*01
CY0973	**H3**	**IGHM*02**	**IGHD*02**	**IGHG1*03**	**IGHG2*03**	**IGHG3*02**	**IGH4*05p**	**IGHE*02**	**IGHA*02**
CY0161	H4	IGHM*03	IGHD*03	IGHG1*04	IGHG2*05	IGHG3*02	IGHG4*03	IGHE*03	IGHA*03
CY0424	H5	IGHM*04	IGHD*04	IGHG1*05	IGHG2*04	IGHG3*01	IGHG4*04	IGHE*04	IGHA*04
CY0333	H6	IGHM*02	IGHD*02	IGHG1*03	IGHG2*03	IGHG3*02	IGH4*05p	IGHE*02	IGHA*02

Supplemental Information

Figure S1. Comparison of read coverage for IGH assemblies from homozygous animals. (A) CY0692 and (B) CY0973 versus (C) both assemblies from heterozygous animal CY0161. The assembly for CY092 is fragmented into 3 contigs. Dashed red lines show average genomic coverage for each animal. and plots, bin size of 10,000bp. Fragmented homozygous haplotype CY0692.

Figure S2. Read support for each of the 20 H1 SNVs relative to consensus shown in Fig 1b.

Figure S3. Current databases do not adequately capture the immunogenetic diversity of cynomolgus macaques. Heatmap showing nucleotide sequence identity between MCM IGHV alleles observed on haplotype H1 (rows) and known cynomolgus macaque IGHV alleles from IMGT (columns). Although H1 is a commonly observed haplotype in our cohort, only 69/82 (84%) of H1 IGHV alleles have at least 95% sequence identity to at least one allele in the IMGT database.

Figure S4. Extensive allelic differences between species. Heatmap showing nucleotide sequence identity between MCM IGHV alleles observed on haplotype H1 (rows) and IGHV alleles observed on the rhesus macaque Mmul_10 reference genome (columns). Only 5 alleles are shared exactly between species and 11/82 (13%) of H1 IGHV alleles have no counterpart in Mmul_10 with at least 95% sequence identity.

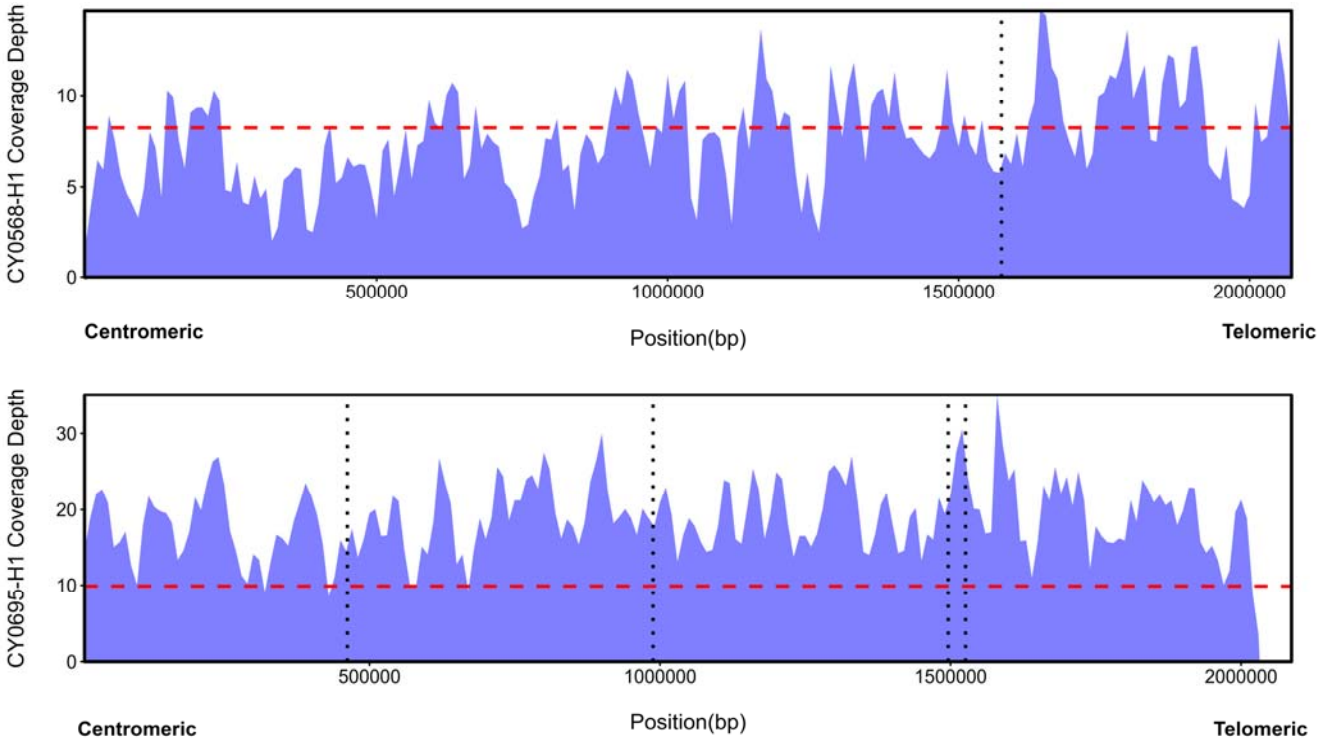
Figure S5. Significant diversity within MCM IGH haplotypes. Dotplots of H1, H2, H3, H4, H5, and H6 visualized using PatchWorkPlot [88].

Figure S6. Differences between KIMDB and the current study. (A) Each haplotype identified in this study includes 21-33 IGHV alleles that were not captured in KIMDB. See also Fig 2c. (B) When alleles from two partial additional haplotypes (Table 1) are included, 41 alleles previously found only in KIMDB are now identified as shared alleles. (C) When differences at or after position 320 are ignored, 40 KIMDB-only alleles collapse onto 38 alleles identified in the current study, including 10 other KIMDB alleles. Together with the two partial haplotypes, this eliminates almost all potential false negatives in the current study. (D) Only 3 additional IGHD alleles are identified in the partial haplotypes, none of which were previously documented in KIMDB.

Table S1. Sequences of alleles reported with corresponding KIMDB labels.

Fig 1

A



B

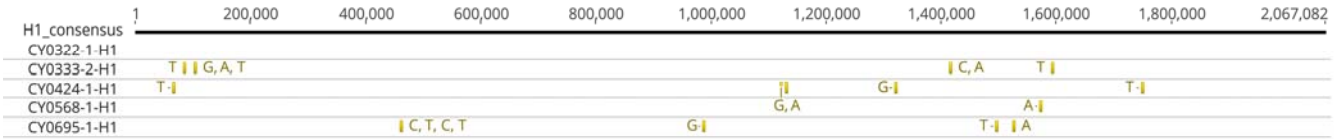


Figure 1. H1 is the predominant IGH haplotype found. (A) Read coverage plots of H1 assemblies from two homozygous animals. Dashed red lines show the average genomic read coverage for each animal. Vertical dotted lines indicate the positions of SNVs shown in panel B. (B) Five full-length, contiguous H1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

Fig 2

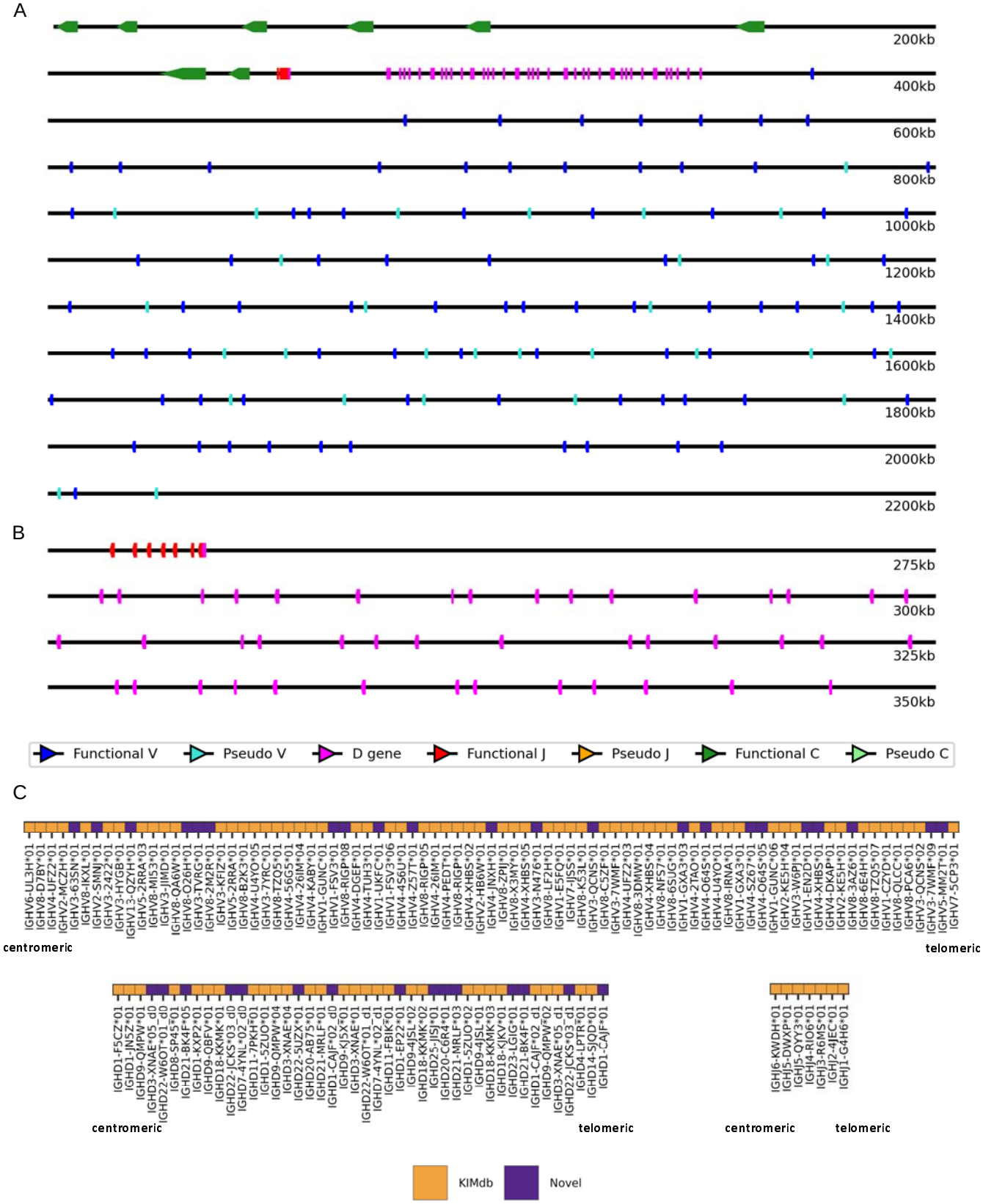


Figure 2. IGH haplotype H1 gene content. (A) Schematic representation of H1 showing the genomic positions of genes annotated by ALIGaToR. (B) Enlarged schematic view of the D-J cluster. (C) H1 V, D, and J genes annotated as functional colored by whether they correspond to alleles represented in the existing KIMdb database. Correspondences between temporary labels and names in KIMdb can be found in Table S1.

Fig 3

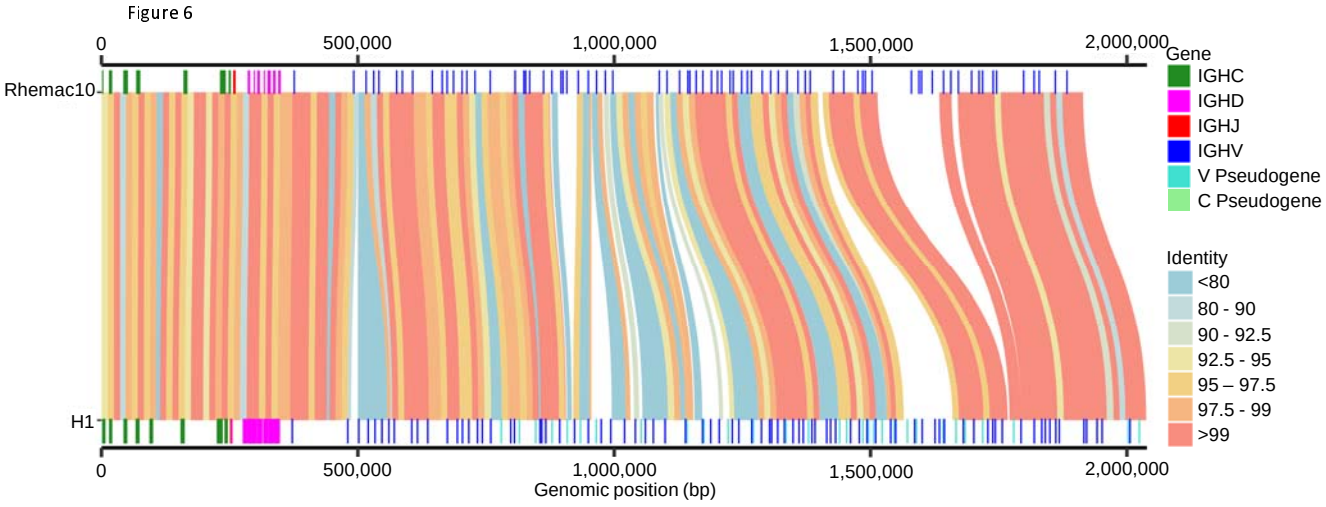


Figure 3. Comparison of IGH from MCM haplotype H1 to M_mul10 reference genome. H1 has approximately 100 kilobases of sequence not found in M_mul10, and overall sequence homology is only 83%.

Fig 4

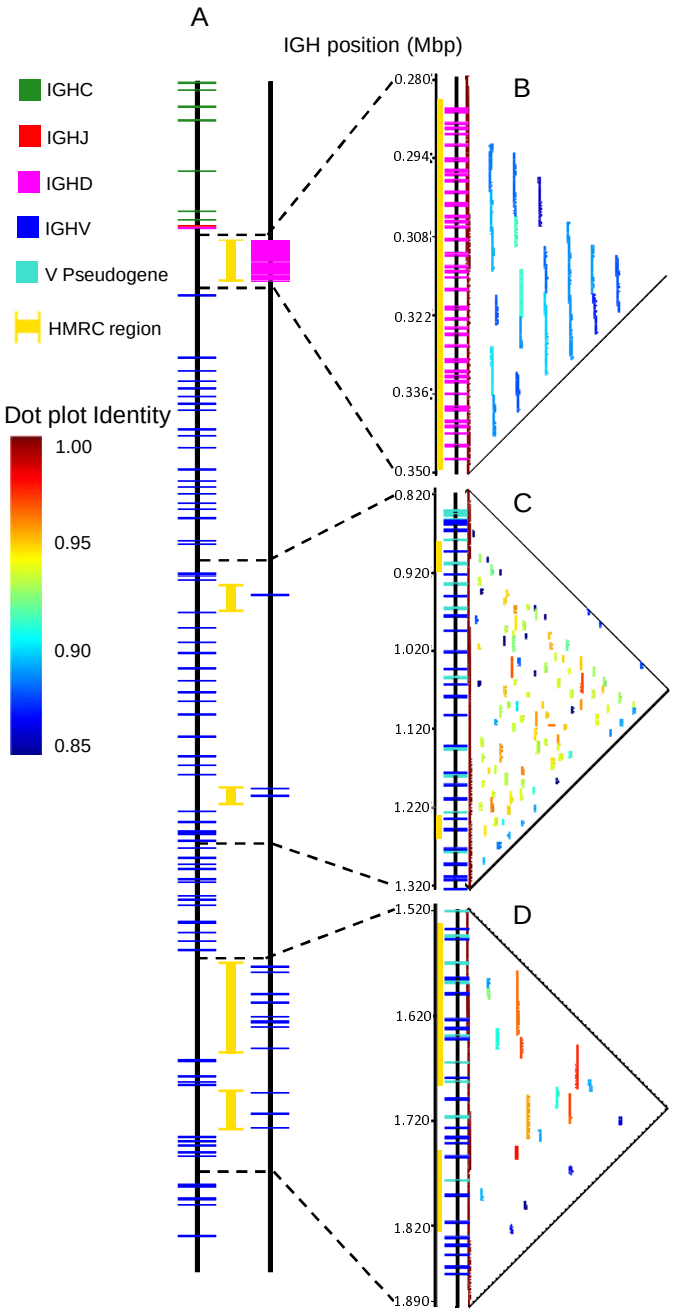


Figure 4. Regions enriched for duplications/repeats within H1. (A) Locus track of the H1 IGH locus. The left-most track are all functional genes outside of HMRC (high multiplicity repeat content) regions (yellow). The right-most track are all genes within HMRC regions. (B) Dot plot of the HMRC region spanning all 38 functional IGHD genes. The dot plot colors correspond to sequence identity. (C) Dot plot of the proximal repetitive region containing two HMRC blocks. The dot plot covers 24 functional IGHV alleles in this region with 3 inside HMRC regions highlighted by a yellow bar. (D) Dot plot of a more distal region with larger homologous blocks and HMRC regions. The dot plot covers 20 functional IGHV alleles in this region with 11 inside HMRC regions highlighted by a yellow bar.

Fig 5

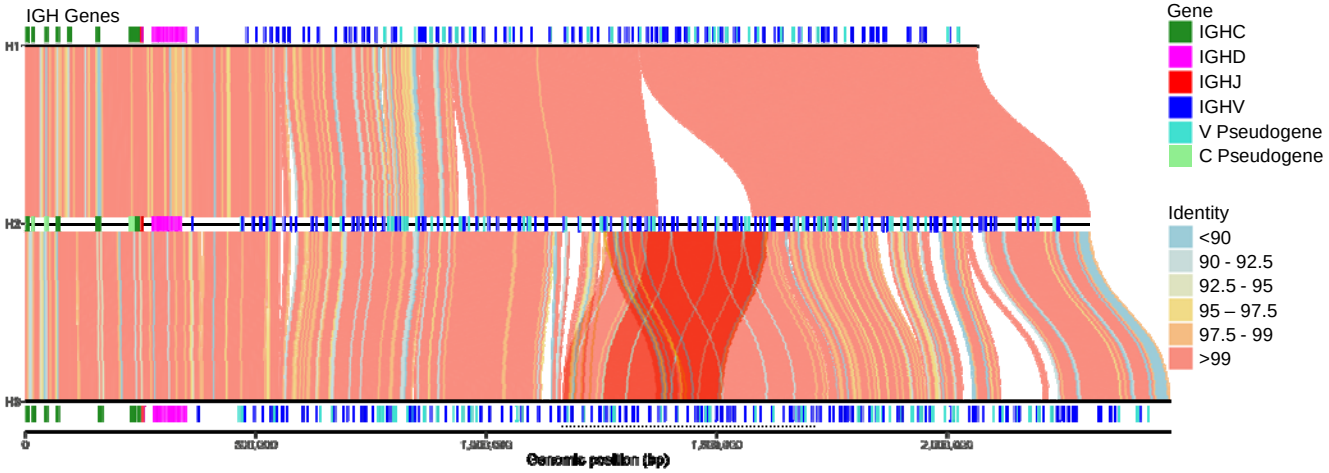


Figure 5. Significant diversity within MCM IGH haplotypes. Alignments of H1, H2, and H3 visualized using SVbyEye. Little structural variation is observed in the IGHC and IGHJ regions (left) with more complex events interspersed with regions of high homology. The underlined region of the H3 haplotype represents a segmental duplication relative to H2.

Fig 6

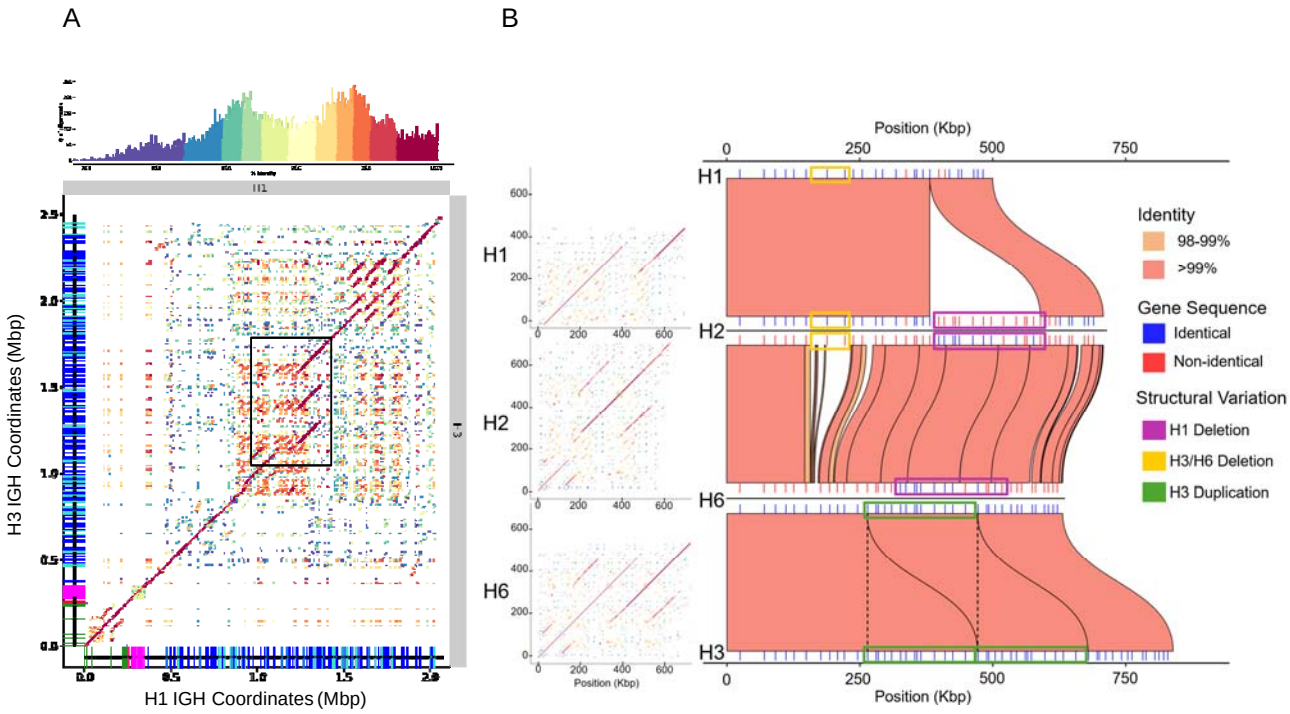


Figure 6. Haplotype H3 contains a large duplications compared to H1. (A) Dot plot comparing H1 (horizontal axis) to H3 (vertical axis). Box (H1:950,000-1,450,000 bp) highlights a region of segmental duplication in H3, just distal of the H1 repeat region show in Fig 3a. The histogram shows the frequency of bins at each identity level, with colors corresponding to colors in the dot plot. (B) Dot plots (left) and Miro plot (right) comparisons of highlighted segmental duplication in haplotypes H1, H2, H3, H6. From top to bottom; H1 is compared to H2, H2 is compared to H6, and H6 is compared to H3. For Dot plots, the top comparison is on the Y-axis and the bottom comparison is on the X-axis. For Miro plots, genes for each haplotype are marked by lines and colored based on sequence identity to comparison haplotype. Blocks of structural variation are highlighted with dotted lines demarcating region of duplication in H3.

Fig 7

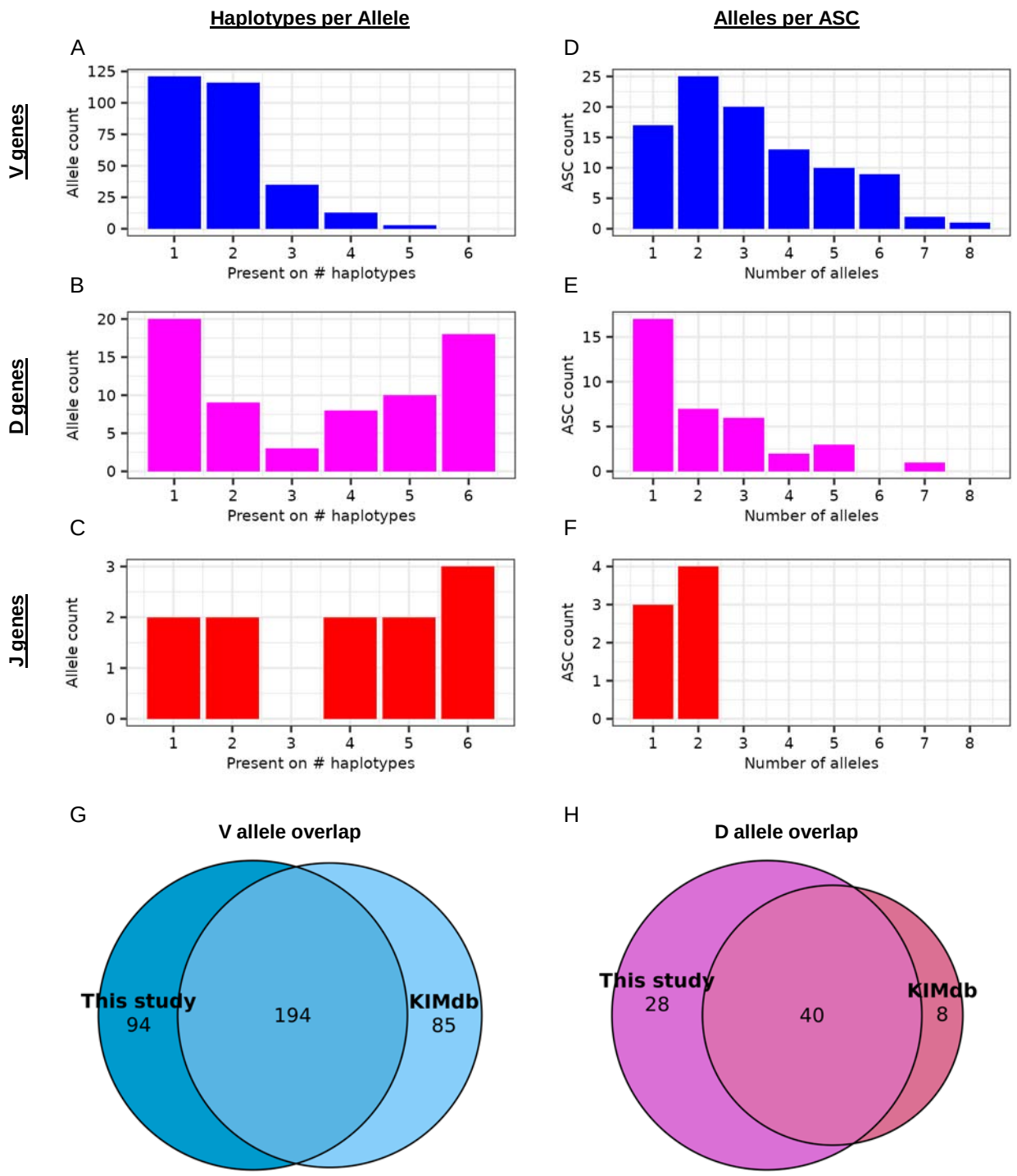


Figure 7. Genetic diversity among MCM haplotypes. (A) The plurality of IGHV alleles appear uniquely in a single haplotype, while (B) IGHD and especially (C) IGHJ alleles are more likely to be shared across most or all haplotypes. The large number of IGHV alleles shared between two haplotypes is due to extended blocks of shared sequence between H1 and H2, as well as between H3 and H6 (see Fig 6b). (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) Overlap between the IGHV alleles observed in this cohort and those from a previous cohort [CITE]. See also Fig S6. (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.

Fig 8

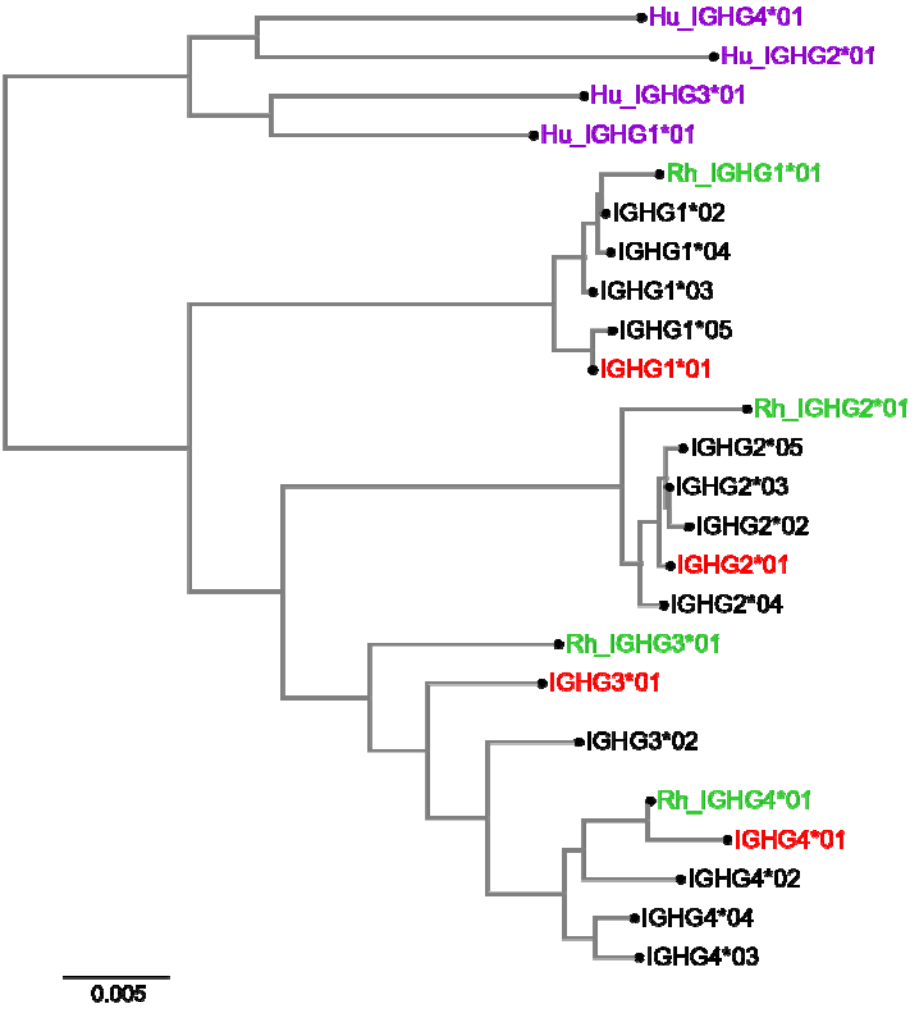


Figure 8. Diversity of MCM IgG constant region subtypes compared to rhesus macaque and humans. Neighbor-joining tree for full-length alleles (CH1 through CHS) of all IgG subtypes found in this study, together with representative alleles from human (purple) and rhesus (green) IgG subtypes. MCM alleles present in H1 are highlighted in red. All other MCM alleles are shown in black.