A Broad Survey and Functional Analysis of Immunoglobulin Loci Variation in Rhesus Macaques

Ayelet Peres[1,2,*], Amit A. Upadhyay[3,*], Vered Klein[1,2,*], Swati Saha[4], Oscar L. Rodriguez[4], Zachary M. Vanwinkle[3], Lukas Granholm[3], Kirti Karunakaran[3], William Lauer[4], Mark C. Lin[3], Timothy Melton[3], Amanda Metz[3], Pazit Polak[1], Nagarajan Raju[3], Kaitlyn Shields[4], Steven Schultze[4], Thang Ton[3], Adam Ericsen[3], Stacey A. Lapp[3], Melissa L. Smith[4], William Lees[1,4], Corey T. Watson[4,†,§], Gur Yaari[1,2,5,†,§] and Steven E. Bosinger[3,6,†,§]

[1]Faculty of Engineering, Bar Ilan University, Ramat Gan, Israel

[2]Bar Ilan institute of nanotechnology and advanced materials, Bar Ilan university, Ramat Gan, Israel

[3]Emory National Primate Research Center, Atlanta, GA, USA

[4]Department of Biochemistry and Molecular Genetics, University of Louisville, Louisville, KY, USA

[5]Department of Pathology, Yale School of Medicine, New Haven, CT, USA

[6]Department of Pathology & Laboratory Medicine, Emory University School of Medicine, Atlanta, GA, USA

[*]These authors contributed equally to this work

[†]These authors jointly supervised this work

[§]Corresponding authors emails: corey.watson@louisville.edu, gur.yaari@yale.edu, steven.bosinger@emory.edu

January 7, 2025

Abstract

Rhesus macaques (RMs) are vital models for studying human disease, and are invaluable to pre-clinical pipelines for vaccine discovery and testing. Particularly in this regard, they are often used to study infection and vaccine-associated broadly neutralizing antibody responses. This has resulted in an increasing demand for improved genetic resources for the immunoglobulin (IG) loci, which harbor antibody-encoding genes. However, the highly polymorphic and structurally variable nature of these loci have them historically challenging to sequence and characterize at the level of both the genome and expressed repertoire. To address these challenges, we have developed a novel integrated analysis workflow for conducting the combined processing of B cell receptor repertoire sequencing data with matched whole-genome and targeted long-read genomic sequencing data. Using this novel approach, we have assembled the largest collection of IG germline alleles reported to date, amassed from 106 Indian origin RMs. Using a conservative annotation approach, requiring sample-level internal validation from both genomic and expressed datasets, we created a comprehensive resource that captures extensive diversity of IG heavy and light chain variable (V), diversity (D), and joining (J) alleles, as well as leader, intronic, and recombination signal sequences (RSSs). This publicly available, continually updated database will advance vaccine research for infectious disease, and provide a robust foundation for immunogenomics and future translational research.

1 Introduction

The extraordinary diversity of B cell receptors (BCRs) underpins the adaptive immune system's ability to recognize a vast range of foreign antigens and produce high-affinity, neutralizing antibodies [1]. In mammals, BCRs are encoded by an extensive family of genes that reside within three immunoglobulin (IG) regions: the heavy (IGH), and light chain kappa (IGK) and lambda (IGL) loci. Diversity in the expressed BCR repertoire is generated through the recombination of V, D (IGH only), and J gene segments, combined with imprecise joining of the segments during B cell development. This diversification process is further enhanced by somatic hypermutation (SHM) after antigen exposure [2]. Despite their fundamental role in forming the BCR repertoire, the study of germline IG gene diversity is in its infancy [3]. Historically, the IG loci have been difficult to sequence and characterize using standard high-throughput genomic approaches, largely due to their genomic complexity, represented by high repeat content and segmental duplication, as well as the fact that they can exhibit extensive genetic polymorphism between individuals. Recent advances in genomic and expressed BCR repertoire sequencing and analysis methods have begun to enable large-scale characterization of the IG loci, uncovering the first hints of the remarkable allelic and haplotype diversity present at the population level [4, 5, 6]. Importantly, there is a growing appreciation for the fact that the composition of an individual's BCR repertoire is shaped at multiple levels by heritable factors within the IG loci[7, 8, 9, 5], and even single nucleotide changes can alter the ability of an IG gene to encode protective IGs[10, 11, 12, 13].

Many efforts toward cataloging IG genes and alleles in Rhesus macaques (RM) have emerged over the past decade [14, 15, 16, 17, 18, 19, 20, 21, 22], providing initially valuable resources for immunogenetic analyses, and highlighting the immense diversity of the RM IG loci. Collectively these studies have utilized both genomic and expressed BCR sequencing data to characterize IG sequences. The analysis of both short-read and long-read sequencing based assemblies have revealed that the IG loci of RM are structured similar to those of human, each spanning large $\tilde{1}$MB regions of the genome, and harboring hundreds of functional/open reading frame variable (V), diversity (D), and joining (J) gene segments [16, 17, 20]. While these studies provided an initial impression of locus structure, they were based on assemblies from only a small number of individual RMs, and of variable and undefined quality. However, initial comparisons of these haplotypes revealed the presence of large structural variants (SVs), altering gene copy number, and highlighting the potential for significant haplotype variation [17, 16]. Furthermore, additional IG germlines have been identified from expressed BCR and targeted genomic sequencing [19, 15, 21, 22, 23] in additional animals, from which 100's of allelic variants have been curated. These studies have highlighted the tremendous IG genetic diversity present in RM, providing early findings that this diversity parallels (or even surpasses) that known for human.

While the RM IG genetic studies conducted to date have been instrumental in establishing essential foundational resources for the community, they currently face several limitations: (1) they are collectively heavily biased toward IGHV sequences, with limited coverage of light chain alleles; (2) they are derived from small subsets of animals, offering limited representation of colony and species diversity; (3) they have been sourced largely from BCR sequencing datasets, lacking information on locus position and annotation of elements, such as introns, leaders, and recombination signal sequences (RSSs). This last limitation, specifically, has impeded efforts to carry out comprehensive gene and allele assignments for the majority of known IG germlines; as a result, alleles identified across studies and available resources lack a unified nomenclature system.

The importance of continuing to build more robust IG germline resources is motivated by the fact that the RM remains a cornerstone model in biomedical and pre-clinical research [24, 25, 26]. Indian-origin RM, in particular, are valuable for the preclinical testing of candidate HIV vaccines [27]. This is highly relevant, as many of the most promising HIV vaccine candidates aim to elicit B cell responses from a predefined set of IG genes, a strategy known as "germline-targeting" (GT) [28]. This approach focuses on the elicitation of broadly neutralizing antibody (bNAb) precusors encoded by specific germline (unmutated) IG gene segments. In humans, it has been clearly demonstrated that IG genotype and germline BCR precursor frequencies in the pre-vaccination repertoire can significantly influence vaccination outcomes [29, 30]. This observation necessitates a clear understanding of putative RM homologs for these candidate GT vaccine targets in human. Thus, the paucity of comprehensive and well-curated IG genomic resources has the potential to constrain the utility of RMs for evaluating GT immunogens.

This potential limitation is illustrated by studies in which the GT immunogen eOD-GT8 successfully elicited VRC01-class B cell precursors in humans but failed to do so in RM, likely due to the absence of a direct homolog to the human IGHV1-2 gene [29, 31]. Conversely, GT immunogens designed for IG genes with closely related homologs

in both species have successfully elicited B cells using the appropriate IG gene segment in RM [32]. As additional GT immunogens advance through preclinical development [33], a comprehensive understanding of the IG loci in RM has become increasingly critical.

To better meet the growing needs of the community, we initiated the RM IGSeq Project, which aims to conduct population scale genotyping of the IGH, IGK, and IGL loci in RM from AIDS-designated colonies across the US. In order to overcome the technical hurdles associated with sequencing and annotation of these notoriously challenging regions, we implemented a novel approach of re-constructing full-locus IG haplotypes utilizing long-read sequencing of genomic DNA and pairing this data with expressed adaptive immune receptor repertoire sequencing (AIRR-Seq) analyses. Here, we report the first release of alleles from this project, resulting from paired analyses of 106 Indian origin RMs sourced from three different colonies. Curated datasets have been made publicly available, and we provide resources for implementing the dataset for single-cell RNA-Seq analyses. These data provide a significantly expanded resource for immunogenetic analyses of the RM, and lay the foundation for a unified nomenclature system of the IG loci in this species.

2 Results

2.1 Unprecedented Collection of IG Alleles Illuminates Surprisingly High Levels of Germline Variation

We curated an unprecedented collection of IG-encoding alleles in RM using a cohort of 106 subjects, enabling an in-depth exploration of IG germline variability. This dataset comprises targeted IG long-read genomic sequencing data matched with expressed IgM, IgK, and IgL BCR repertoires generated from peripheral blood mononuclear cells (PMBCs) (Supplementary Table 1), facilitating a comprehensive analysis at both the genomic and expressed levels (Figure 1).

We constructed a "seed" reference set for V, D, and J segments across the IGH, IGK, and IGL loci by integrating established sources (KIMDB [21], IMGT [20], and RhGLDB [19]) alongside newly inferred alleles derived from diverse sources (RhGLDB+) and from genomic data that will be presented in a forthcoming sister publication (the Trio set, see Methods section 4.4). Targeted IG long-read genomic data was generated using an optimized protocol based on those previously published for the human IG loci [34, 35, 36]. This protocol utilized a custom target oligo panel designed from known IG genomic data to enrich long-read single-molecule-real-time (SMRT) Pacific Biosciences sequencing libraries for IG locus DNA. Following this approach, sequencing libraries for multiple samples were multi-plexed, allowing for high-throughput generation of long-read IG genomic data. High-fidelity (HiFi) long-reads were assembled into highly accurate haplotype-specific contigs. Sample-level genomic assemblies were then annotated using DIgGER [37] to identify known and undocumented alleles, defining a genomic genotype (GG) for each subject. Matching AIRR-seq data were annotated using these personal GG sets, applying stringent criteria to identify high-confidence expressed alleles (Methods, Section 4.8). This conservative approach enabled the confident curation of high-fidelity alleles at both the genomic and expressed repertoire levels for each subject.

Our cohort exemplifies surprisingly high levels of IG germline variations in RM BCR repertoires (Figure 2A). The ratio of expressed-to-genomic alleles varied across loci: IGKV and IGLV had a median ratio of $\tilde{0}.74$, IGHV 0.62, and IGHD 0.61. J segments (IGHJ, IGKJ) showed near-complete overlap ($\tilde{1}.0$), while IGLJ had a lower ratio (0.6) due to three non-expressed alleles. These findings suggest that the IGH loci contain more genomic "passenger" alleles than the IGK/L loci.

Figure 2B illustrates the overlap among three allele groups. The first group, GG, includes all unique alleles found in the personal genotypes of all RMs. The second group, the baseline reference, comprises previously known alleles, including those from the seed set (see Method Section 4.4) and external sources (VRC[38] and Guo et al. [22]). The third group represents alleles that were in the expressed RM repertoires and inferred to be present in their personal genotype. Our cohort significantly expanded the known allele set. For IGHV, 727 ($= 4 + 246 + 477$) alleles were identified as expressed in our cohort, with 250 previously undocumented. Of the 479 ($= 246 + 4 + 229$) de novo IGHV alleles, 52% were expressed. Similar trends were observed for IGKV and IGLV, with 352 and 255 de novo alleles, of which 70% and 62% were expressed, respectively.

We also observed substantial variability in allelic sets between individuals. Among the 727 IGHV alleles expressed

in at least one individual, 63% were detected in 10 or fewer animals, reflecting high inter-individual diversity (Figure 2C). Genomic and expressed allele-sharing patterns were highly similar, supporting the robustness of our approach and reflecting population-level variation.

2.2 USofA Extends the Breadth of All Previously Used Datasets

We created the final unified set of alleles (USofA) by integrating alleles from the baseline reference described above and previously undocumented alleles newly discovered in this study. This set provides a broader and more comprehensive dataset compared to previously established IG germline databases (Figure 3A). Alleles shared across multiple sources were more frequently observed in both genomic and expressed repertoires than alleles unique to a single source. For instance, within the IGHV locus, alleles found in KIMDB, Trios, RhGLDB+, and IMGT were consistently observed at higher frequencies as expressed in our cohort (70 out of 70) than those exclusive to either source (56 out of 421 for KIMDB, 50 of 73 for the Trios, 8 of 96 for RhGLDB+, and 7 of 137 for IMGT-exclusive alleles). This trend extended to other loci as well; in IGLV and IGKV, all alleles that were shared by more than one source were observed in our expressed cohort, as opposed to alleles that were present only in IMGT or only in RhGLDB+ (*e.g.*, 3 of 40 and 11 of 113 in IGKV and IGLV respectively). Similarly, for IGHD and IGHJ, alleles exclusive to KIMDB were rarely detected in our cohort (4 out of 9 for IGHD and none for IGHJ), while shared alleles between KIMDB and other sources were more represented (32 out of 34 for IGHD and 8 out of 8 for IGHJ, Supplementary Figure 1).

As anticipated, alleles derived from the trio dataset were more frequently represented in our cohort, both genomically and in the expressed repertoires, with 50 out of 73 alleles detected for IGHV, 54 out of 57 for IGLV, and 95 out of 115 for IGKV. This likely reflects the similarity in sequencing protocols and the shared origins of the animals used to infer these alleles, leading to a stronger overlap with this larger cohort.

Newly identified alleles from the USofA enrich the diversity of IG loci by addressing gaps in existing databases and adding alleles not previously described (Figure 3B). For IGHD alleles, a stringent filtration process was applied to overcome unique challenges in their reliable inference. The shorter length of D segments complicates their identification in genomic data, while extensive trimming during V(D)J recombination often obscures their full sequence in repertoire data. We took a conservative approach here, including only IGHD alleles expressed in the repertoire with a median ratio of observed sequence length to allele length above 0.5 (Supplementary Figure 2). This filtering reduced the total number of IGHD novel alleles from 260 to 9 (Figure 3B IGHD column crossed bar), ensuring that only high-confidence alleles were included.

2.3 Unsupervised Clustering of the USofA Based on Coding Region Sequences Distinguishes Between Expressed and Non-Expressed Clusters

Unsupervised clustering of the USofA was performed in two steps. First, complete linkage clustering was applied with a 75% similarity threshold to define "allele families". Next, a community detection algorithm grouped alleles into clusters referred to as Allele Similarity Clusters (ASCs) [39].

The ASCs varied in size, ranging from 1 to 68 alleles (Figure 4A). V segments showed the most diversity, forming clusters of up to 68 alleles in IGHV, 36 in IGKV, and 29 in IGLV. D and J segments were more constrained, with up to 8 alleles in IGHD and 3–4 in the J segments. To further explore the diversity of ASCs and the families they form, we examined the number of ASCs grouped into each family. Families of V segments contained between 1 and 63 ASCs for IGHV, 1 and 61 for IGKV, and 1 and 14 for IGLV. Families of D and J segments were smaller, containing between 1 and 5 ASCs for IGHD and only 1 to 3 ASCs for IGHJ, IGKJ, and IGLJ. Where possible, we maintained correspondence with IMGT family classifications to ensure consistency. To illustrate the distribution of alleles in these families, their sizes were categorized into groups based on allele counts (Figures 4B and 4C). Most ASCs contained only 1–4 alleles, while larger groups with more than 16 alleles occurred predominantly in V segments.

A comparison of the fractions of ASCs and families observed in AIRR-seq in our cohort for each gene type revealed notable differences (Figure 4D). The IGHD and IGHJ loci had the highest fractions of ASCs represented in AIRR-seq (approximately 87% and 88%, respectively). In contrast, only 37% of IGHV ASCs and 15% of IGHV families were observed in AIRR-seq, suggesting that a large portion of the IGHV alleles may be non-functional, very

lowly expressed, or restricted to a subpopulation that is not present in our cohort. For the IGKJ and IGLJ loci, 67% and 50% of families were represented, respectively, while the IGKV and IGLV loci showed moderate representation, with 31% and 55% of their families observed in repertoire data of current cohort.

To further explore IG gene diversity, we assessed the number of alleles observed for each ASC in each animal, and across the cohort as a whole. Although ASCs are not by definition "genes", this assessment could approximate zygosity states (homozygous vs. heterozygous) across individuals. Zygosity profiles revealed extensive allelic variability across the population (Figure 5, and Supplementary Figure 3). For approximately three quarters of the ASCs, we observed 0, 1 or 2 distinct allele sequences in each individual, most likely representing gene deletions, homozygotes/hemizygotes, and heterozygotes, respectively. For the remaining ASCs, we observed animals in which the allele number was > 2, indicating that these ASCs were likely to represent gene duplication events resulting in gene copy number variants (CNVs). We observed 7 alleles for one ASC in the IGKV locus. Relative to IGHV and IGKV, IGLV had fewer ASCs with > 3 alleles, indicating that there may be fewer CNVs/gene duplication expansions in IGL at the population level.

We next examined the genomic assembly data to garner additional support for the presence of CNV. Assessments of assembly contigs harboring the ASC, IGHV3-GYPG (Figure 5D), indeed revealed that individuals carrying > 3 alleles had haplotype-resolved contigs with > 1 IGHV3-GYPG alleles present. The contigs characterized provide direct evidence for a CNV, with IGHV3-GYPG copy number ranging from 1 to 4 per contig.

2.4 Analysis of Recombination Signal Sequences of Expressed Alleles Reveals Striking Patterns of Diversity

RSSs are critical genomic elements for the proper functioning of V(D)J recombination, and are common motifs for defining "functional" categories of IG genes. However, few studies have had the opportunity to assess variation in RSSs among IG alleles with solid evidence of expression in the repertoire. Here, we curated RSSs from genomic assembly data for all IG alleles identified in our cohort, including heptamer, nonamer, and spacer elements. We limited our analysis to alleles with high-confidence RSSs, defined as cases where a given RSS was the only non-coding variation for an allele expressed in at least one subject (Figure 6A). We found that RSSs exhibit substantial variability among expressed alleles, gene segment types and loci (Figure 6). In total, we identified 328 RSSs across loci: 78 in IGHV, 8 in IGHJ, 35 in 5' IGHD, 48 in 3' IGHD, 70 in IGKV, 5 in IGKJ, 76 in IGLV, and 8 in IGLJ.

In addition to assessing sequencing variation within nonamers and heptamers, we also identified variation in spacer lengths among expressed alleles, including non-canonical lengths. For example we observed the space length of 22 nucleotides in IGHV was associated with four unique heptamer-nonamer pairs, representing 41 alleles. Similarly, we observed an 11-nucleotide spacer in the 3' IGHD, observed for one RSS for a single allele, found in 81 subjects. While most RSS heptamer-nonamer pairs were shared with those in humans, a small subset were unique to RM. For instance, IGHV contains 62 pairs shared with human data from VDJbase, and 16 that were only found in our RM dataset. In IGKV, we identified 51 shared pairs and 19 that were only in RM. Notably, all IGKJ pairs identified in this study were observed in humans.

We found that the RSS spacer exhibited greater variability between loci, relative to intra-locus differences observed for the heptamer and nonamer regions (Figure 6B). Relative to the number of unique IGHV alleles observed, we observed relatively low variability in RSSs. In contrast, the J and D segments exhibited a nearly one-to-one ratio of alleles to unique RSSs. This suggested that RSS specificity is likely critical to guiding the precise recombination and pairing of particular V's, D's, and J's during V(D)J recombination, whereas RSS variation in IGHV may play a less prominent role to dictating variation in V gene selection by recombination machinery.

The locus-specific nature of RSSs was further emphasized by patterns of sharing of heptamer-nonamer pairs across loci (Figure 6C). Most RSSs were exclusive to a single locus. However, IGKV and IGLV displayed an unexpected degree of sharing, with 8 and 9 shared heptamer-nonamer pairs, respectively, across other loci.

These results highlight the substantial variability and locus specificity of RSSs in the RM population. While the majority of RSS's heptamer-nonamer pairs in our dataset were shared with those in human, the observed diversity with our dataset, particularly in the J and D segments, underscores the importance of RSS uniqueness in recombination efficiency and the generation of a diverse immune repertoire.

2.5 Allele Browser

We have developed an interactive web server, accessible at https://vdjbase.org/reference_book/Rhesus%20Macaque, that provides comprehensive insights into RMs IG data. For each IG chain, the platform offers an Allele Visual Summary, presenting a detailed overview of alleles categorized by gene types, families, and ASCs. Accompanying this summary is a matching Allele Table that includes information on the coding region, such as nucleotide sequences, IgLabel [40], and the other sources where the allele appears, as well as details on non-coding regions, including RSS and leader sequences. Additionally, the web server presents ASCs and the absolute frequencies of alleles within the dataset, along with the chosen allele-specific thresholds. Interactive graphs enable users to explore available group information, thresholds, frequencies, genotype, and zygosity data by selecting individual data points. This user-friendly interface enhances the analysis of RM IG data, allowing for detailed exploration of allele distributions and relationships.

3 Discussion

This study represents the first large-scale population resource for characterizing IG-encoding variants in RMs, integrating both sample-level comprehensive locus-wide high-quality genomic data with expressed repertoires to facilitate an unprecedented characterization of IG germline diversity in RMs, including both heavy and light chain loci. By analyzing 106 Indian-origin RMs from multiple colonies, a significant number of "novel" alleles were identified across seven IG gene types and loci. Despite the breadth of this dataset, it is evident that our current collections of alleles have not reached saturation, and new genes and alleles are likely to be discovered with the inclusion of additional samples. This observation applies not only to IG-encoding regions but also to associated leader sequences and RSSs. A unique strength of this cohort is its capacity to explore in depth how genetic variations influence the expressed repertoire. This analysis extends to IG non-coding regions, leader sequences, and RSSs, providing valuable insights into their combined roles in shaping BCR repertoire diversity. Notably, we have substantially increased the documentation of leader sequences, and were able to confirm proper delineation of L-PART1 and L-PART2 by verifying the expressed leader sequence in AIRR-seq. Importantly, these gene elements are essential for the proper functioning of IG sequence assembly tools, such as 10X Genomics' Cell Ranger software, and for the design of comprehensive and efficient PCR amplification primers.

All supporting evidence for the observed IG alleles in our cohort are accessible, including genomic and expressed repertoire data from the cohort. For comparative purposes, germline alleles from external databases were incorporated into a comprehensive table, with annotations indicating their presence in other resources, where applicable. While users of this resource may download the full table as a reference for aligning new datasets, users may also choose to apply various filter criteria to tailor the resource to their specific needs. For example, a user can prioritize alleles with demonstrated expression in repertoires or those with matched RSS sequences. It should be noted that the table contains alleles which were not observed in our samples (and are annotated as such in the table). The evidential standards applied to these alleles depend upon the sources from which they are taken. Recognizing the need for comprehensive germline sets that are easy for researchers to adopt and apply, we plan to collaborate with other groups within the AIRR Community to develop such sets and ensure the development of appropriate evidential standards [40].

When examining shared alleles across the cohort, baseline alleles collected from external sources and identified within our cohort exhibited consistent behavior whether observed in genomic or expressed repertoire data. In contrast, previously undocumented alleles were more frequently detected only in genomic data. This discrepancy underscores the critical role of genomic data in studying evolutionary features and unraveling the complex genetic mechanisms underlying IG variability in RMs.

While this study has significantly expanded the known repertoire of IG germline alleles, challenges persist in developing a unified nomenclature system that is independent of presumed genomic positions. The ASC approach employed here reveals that, while most IG ASCs in RM harbor 1-2 alleles per animal, approximately one-quarter of ASCs represent expanded sets of duplicated genes, which vary in copy number between individuals. Without the utilization of comprehensive genomic datasets, such genes will be difficult to categorize and name using traditional nomenclature approaches. However, if we are to ultimately understand what role these types of genomic variants play in determining characteristics of antibodies and circulating repertoires, it will be crucial for gene and allele

names to properly represent biological and genetic concepts. Our data clearly demonstrate that the RM IG genomic community faces new challenges that will have to be tackled head on. Designing a nomenclature system robust enough to accommodate the addition of numerous new alleles curated from various forms of data, particularly while lacking genomic context will be a non-trivial task. This challenge will require a community-wide effort to facilitate efficient communication about curated IG germline variation.

Future efforts will focus on expanding this resource to include more diverse populations of RMs, as genetic diversity across colonies and geographic origins remains under-explored. Our data clearly demonstrate that drastic differences between NHP colonies are likely to exist. We plan to add additional datasets from a more diverse set of RMs currently being generated to this resource, expanding sampling breadth in subsequent versions. Collaboration across research groups and NHP colonies will be critical to ensure that we effectively incorporate additional datasets that better represent RM IG diversity, and ultimately enhancing this resource for the community. Finally, as we build out the utility of this resource, it will be important to begin integratingfunctional data alongside curated genes and alleles, particularly as we gain knowledge of what IG variants impact antibody structure and repertoire composition. These links will be key for leveraging knowledge of IG diversity to inform our understanding of disease, and the development of better therapeutics and vaccines.

In conclusion, this study provides a robust and dynamic resource for immunogenetic research in RMs, addressing key limitations in existing IG germline datasets while offering novel insights into IG genetic diversity. As additional data are incorporated, this resource will continue to evolve, ensuring its relevance for advancing our understanding of antibody diversity and its implications for immunology and vaccinology.

4 Methods

4.1 Sample collection

Blood was collected from Rhesus Macaques (RM) under ketamine anesthesia via the femoral artery, following Emory National Primate Research Center's IACUC policy 355 "Blood Collection". PBMCs were isolated using BD® Vacutainer™ CPT™ Mononuclear Cell Preparation Tubes (Cat no. 362753) per the manufacturer's instructions. The isolated PBMCs were washed with DPBS (Cat no. 21-031-CM) in 50mL conical tubes and resuspended in ACK Lysis Buffer (Cat no. 118-156-101) for 10 minutes to lyse red blood cells, followed by additional DPBS washes.

PBMCs were aliquoted at 5 million cells per tube and lysed in 350 µL RLT/BME for RNA extraction or flash-frozen in an isopropanol/dry ice bath for later DNA extraction. RNA was extracted using the RNEasy Mini Kit (Cat no. 74106) with DNAse (Cat no. 79256) treatment, following the manufacturer's protocol for total RNA extraction. RNA concentration was measured using the RNA BR Qubit Assay (Cat no. Q10210), and quality and RINe scores were assessed using the 4200 TapeStation RNA Screentape.

4.2 AIRR-seq Library preparation and sequencing

PBMCs were isolated from blood samples using density gradient centrifugation with Ficoll-paque. The procedure for library preparation was based on protocols provided by Dr. Daniel Douek (NIAID/VRC) that have been previously published [41, 42]. The IgM constant primers were based on the primer sequence as previously described [15]. The extraction of RNA from PBMCs was performed using the RNeasy Micro-DNase Digest protocol (QIAGEN) on QIAcube automation platforms (Valencia, CA) into 350 µL QIAGEN RLT buffer. The reverse transcription step was performed using Clontech SMARTer cDNA template switching.

A total of 1 µL of 12 µM 5' CDS oligo(dT) primer was combined with 8 µL of RNA for 3 seconds. After incubation at 72°C for 3 minutes and cooling to 4°C for at least 1 minute, 8.5 µL of master mix was added, comprising 3.5 µL of 5x RT Buffer (250 mM Tris-HCl (pH 8.3), 375 mM KCl, 30 mM $MgCl_2$), 1 µL Dithiothreitol (DTT, 20 mM), 1 µL dNTP Mix (10 mM), 1 µL RNAse Out (40 U/µL), 1 µL SMARTer II A Oligo (12 µM), and 1 µL Superscript II RT (200 U/µL).

The plates were sealed and spun briefly, followed by incubation at 42°C for 90 minutes and 70°C for 10 minutes. The first-strand cDNA thus obtained was purified using AMPure XP beads.

To amplify the cDNA, 19.3 µL was added to 30.7 µL of master mix containing 25 µL of 2x KAPA Real-Time Library Amplification Kit (catalog# KK2702), 0.7 µL of 5PIIA (12 µM) forward primer, and 5 µL of IgM primer

(2 µM) reverse primer. The mixture was vortexed for 5 seconds, centrifuged at 2000 RCF for 1 minute, and monitored using real-time PCR. After amplification terminated in the exponential phase, AMPure XP beads were used for purification.

The addition of barcodes and adapters for sequencing required two rounds of PCR amplification. In the first step, 2 µL of 1:10 diluted amplified rearranged b cell receptor was combined with 46 µL of master mix comprising 25 µL of KAPA HotStart ReadyMix (catalog# KK2602), 2.5 µL of SYBR Green (1:10K), 18.5 µL of nuclease-free water, 1 µL of P5_Seq BC_XX 5PIIA (10 µM), and 1 µL of P7_i7_XX IgM (10 µM). The mixture was vortexed, centrifuged, and subjected to real-time PCR amplification and purification as described previously.

In the second step, 5 µL of the amplified library was combined with 44 µL of master mix comprising 25 µL of KAPA HotStart ReadyMix (catalog# KK2602), 1 µL of P5_Graft P5_seq (10 µM) forward primer, 18 µL of nuclease-free water, and 1 µL of P7_i7_XX IgM (10 µM) reverse primer.

A final round of purification with AMPure XP beads was performed. The quality of libraries was assessed using an Agilent Bioanalyzer. The libraries were pooled and sequenced as 309-base paired-end runs on an Illumina MiSeq.

4.3 Genomic library preparation and seuqnecing

Targeted long-read sequencing of the immunoglobulin (IG) loci from 100+ rhesus macaques was performed using Pacific Biosciences single-molecule real-time (SMRT) technology. Sequencing libraries were generated by adapting a published protocol for targeted enrichment of human immunoglobulin loci. A custom oligonucleotide probe panel ('HyperExplore', Roche) was designed to target the immunoglobulin heavy chain (IGH), kappa (IGK), and lambda (IGL) genomic regions, utilizing sequences from the Rhesus macaque genome reference (RheMac10) [43] and alternative haplotype assemblies [17].

High-molecular-weight genomic DNA was extracted from peripheral blood mononuclear cells (PBMCs) using the DNeasy Blood & Tissue Kit (Qiagen). DNA (1–2 µg) was sheared using g-TUBEs (Covaris) and size-selected with a BluePippin system (Sage Science). The size-selected DNA was subjected to end repair and A-tailing following the KAPA HyperPrep library preparation protocol (Roche) and ligated to sample-specific barcodes and universal adapters. PCR amplification was conducted using PrimeSTAR GXL DNA Polymerase (Takara) with 8–9 cycles. Amplicons were further purified and size-selected using 0.7× AMPure PB beads (Pacific Biosciences). Target enrichment was achieved through hybridization with IGH, IGK, and IGL-specific oligonucleotide probes (Roche), followed by recovery with streptavidin-coated magnetic beads (Life Technologies). A secondary PCR amplification was performed with 16–18 cycles using PrimeSTAR GXL Polymerase. SMRTbell libraries were constructed using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences), incorporating DNA damage repair and end repair steps. Following A-tailing and ligation of SMRTbell adapters, libraries were treated with a nuclease cocktail to remove unligated fragments and purified using 0.45× AMPure PB beads. Final libraries were pooled and sequenced on the Sequel IIe platform (2.0 chemistry) with 30-hour movie acquisitions to generate high-fidelity (HiFi) reads, with an average read accuracy of 99.8.

4.4 Seed reference set curation

The seed reference set used to annotate genomic sequences was derived from a database curated by a Working Group of the AIRR Community (all authors are members of the Working Group). The database included sequences from KIMDB [21], IMGT [20], and RhGLDB [19]. As part of this activity, the authors of RhGLDB contributed additional sequences inferred from repertoires, including light chain sequences and VH alleles. In this manuscript, we refer to the augmented RhGLDB database as RhGLDB+. In addition to alleles from the Working Group database, functional alleles were incorporated from whole-genome assemblies (WGS) of nine rhesus macaques, denoted the Trio set. The seed reference set included 1054 IGHV, 65 IGHD, and 15 IGHJ alleles for the IGH locus; 523 IGKV and 6 IGKJ alleles for the IGK locus; and 342 IGLV and 11 IGLJ alleles for the IGL locus. KIMDB covered the IGH locus only, whereas other sets contained sequences from all three IG loci.

4.5 Genomic Sequencing Processing and Personal Reference Set Construction

4.5.1 Genomic Data Assembly and Filtering

Following genomic data acquisition, de novo assembly was performed using hifiasm [44]. The resulting contigs were filtered based on the following criteria: contigs aligning to the RheMac10 IG loci within chromosomal coordinates using minimap2 [45] (IGH: chr7:167,907,503–169,865,360; IGK: chr13:16,797,737–18,140,859; IGL: chr10:29,621,424–30,895,807), contigs showing significant alignment to validated, high-accuracy long-read assemblies from the Emory rhesus macaque breeding colony, and contigs exhibiting greater than 70% sequence similarity to the seed set using BLAST [46]. Filtered contigs were categorized into IGH, IGK, and IGL loci groups for further annotation.

4.5.2 Contig Annotation and Allele Identification

Annotation of the filtered contigs was performed using DIgGER [37], with the seed set as reference (see Methods, Section 4.4). DIgGER identified de novo alleles and extracted recombination signal sequences (RSS) heptamer and nonamer elements using predefined position-weighted matrices (PWMs) based on known sequence motifs. Spacer sequences, located between the heptamer and nonamer, were extracted directly from the assembled contigs based on DIgGER-reported coordinates. For antisense strand sequences, the reverse complement of the contig was used to identify spacer positions.

4.5.3 Spacer Length Handling

Spacer lengths vary by gene segment. IGHV, IGHJ, IGLV, and IGKJ segments typically use 22–23 bp spacers, IGKV and IGLJ use 11–12 bp spacers, and IGHD uses a 12 bp spacer. To account for non-canonical spacer lengths, DIgGER was run twice. The first run used canonical spacer lengths, while the second run employed expanded spacer ranges of 10–13 bp and 21–24 bp for V and J segments, and 11–13 bp for D segments. Outputs from both runs were combined, with priority given to alleles identified with default spacer lengths in cases of overlap.

4.5.4 Personal Reference Set Construction

Personal reference sets were constructed from DIgGER-inferred alleles. Alleles were included if they had more than 10 fully spanning reads at 100% sequence identity. Under relaxed conditions, alleles were also included if they were part of the seed reference set or observed in another subject with more than 10 fully spanning reads at 100% sequence identity. Alleles classified as pseudogenes or those lacking recognizable heptamer or nonamer sequences were excluded to maintain accuracy and reliability.

4.5.5 IGHD Allele-Specific Criteria

Additional criteria were applied to ensure robust IGHD allele identification. IGHD alleles were retained only if they were located adjacent to the common allele similarity cluster IGHV6-UXXO near the D region if two functional alleles were positioned adjacent to each other in the contig, or if they appeared in a sequence of at least three consecutive IGHD alleles, provided at least one allele was part of the seed reference set.

4.6 AIRR-seq Data Preprocessing and Quality Control

The AIRR-seq data pre-processing pipeline consisted of two main steps using the pRESTO toolkit [47]. First, paired-end reads were assembled using `AssemblePairs` with the align function and default parameters. The assembled sequences were then filtered for quality control using `FilterSeq` with a Phred score threshold of 20.

For each locus in each subject, pre-processed sequence files from technical replicates were merged and aligned against the subject's personalized reference set using IgBLAST [48]. To ensure unique representation, duplicate sequences were collapsed, retaining only those with a duplicate count of at least 2.

4.7 Novel Allele Inference and Alignment

Novel V alleles were inferred using TIgGER [49], enabling the detection of sequence variation up to nucleotide position 318 in the heavy chain and position 330 in the light chain. Similarly, novel J alleles were inferred using TIgGER with modified parameters, detecting sequence variation up to nucleotide position 46 in the heavy chain and position 38 in the light chain.

The sequences were then re-aligned to a germline reference set updated to include any newly inferred alleles. Clone identification was performed using Change-O's `DefinedClone` [50], and a single representative sequence with the fewest mutations was selected for each clone to reduce bias during heavy chain genotype inference.

4.8 Subject Filtration and Genotype Inference

To ensure high-quality data for genotype inference and constructing the USofA, a subject filtration process was applied. Subjects were included only if they had both repertoire and genomic datasets available for all three Ig loci, resulting in an initial set of 132 subjects. Subjects with at least 1,000 annotated reads per chain in the repertoire were retained, reducing the dataset to 125 subjects. Additionally, subjects with low genomic inference coverage or significant discrepancies between genomic and repertoire data were excluded, resulting in a final set of 106 subjects used for genotype inference and USofA construction.

Genotype inference was performed to identify alleles present in each subject's naive repertoire using stringent filtering criteria. Sequences were required to have no mutations within the V region, a single V allele assignment, and alignment starting at position one of the V germline. For IGHD genotype inference, only sequences with no mutations in the D region and a single assigned D allele were included.

The allele-based method from Piglet [39] was used to determine the presence of specific alleles by comparing allele usage against population-derived thresholds. Building on this, an optimized method detailed in a forthcoming sister publication introduced Z-score-based confidence criteria to refine thresholds and maximize the balance between true positives and false positives. Z scores assess the significance of an allele's presence by measuring how its usage deviates from the expected threshold. Alleles with Z scores above 0 were included, indicating a high likelihood of true presence in the repertoire.

4.9 USofA Construction and Allele Clustering

The Unified Set of Alleles (USofA) was constructed by integrating alleles from the seed set, two external sources (VRC [38] and Guo et al. [22]), and de novo inferred alleles from both genomic and repertoire data. Genomic-derived de novo alleles were included only if they were part of a subject's "personal reference," while repertoire-derived de novo alleles were included only if they met the genotype inference criterion of a Z-score above 0.

To address the unique challenges posed by IGHD segments due to their short length and extensive modifications during V(D)J recombination, a stringent filtering process was applied to de novo inferred IGHD alleles. The ratio of observed sequence length to full allele length in repertoire data was calculated for D annotations matching alleles in the subject's genotype. Alleles with median ratios below 0.5 were excluded, with this threshold validated against baseline alleles (Supplementary Figure 2).

Novel alleles matching reference alleles from the seed set were replaced with their corresponding references. For alleles detected in both repertoire and genomic data, genomic alleles were prioritized for accuracy.

An Allele Similarity Clustering (ASC) method utilizing a community detection algorithm was applied to group similar alleles (detailed in a forthcoming sister publication). A dissimilarity matrix was constructed using Hamming distance for aligned V allele sequences and Levenshtein distance for unaligned D and J sequences. After normalization and a negative logarithmic transformation, the Leiden community detection algorithm, combined with the Constant Potts Model (CPM), was used to identify ASCs. This approach was found to provide better separation in this large dataset than the single-linkage approach used in [39].

4.10 Incorporating RSS Variation

After defining the USofA, variations in RSS were analyzed and integrated. Each allele was assigned an RSS-allele pair and categorized as "sure" or "non-sure" based on genomic data. If a single RSS-allele pair was identified in a

subject, it was included and categorized as "sure," regardless of whether the spacer was canonical or non-canonical. If multiple RSS-allele pairs were identified for the same allele in a subject, and all pairs had either canonical spacers or all non-canonical spacers, all pairs were included, but the allele was marked as "non-sure." If an allele appeared with both canonical and non-canonical spacers within a subject, the pair with the canonical spacer was prioritized.

5 Funding

The study was supported by the National Institutes of Health [grant number R24AI162317]. Funding for sample collection: Emory National Primate Research Center [P51OD011132]; Next generation sequencing services were provided by the Emory NPRC Genomics Core which is supported in part by NIH P51 OD011132 and NIH S10 OD026799.

6 Figures

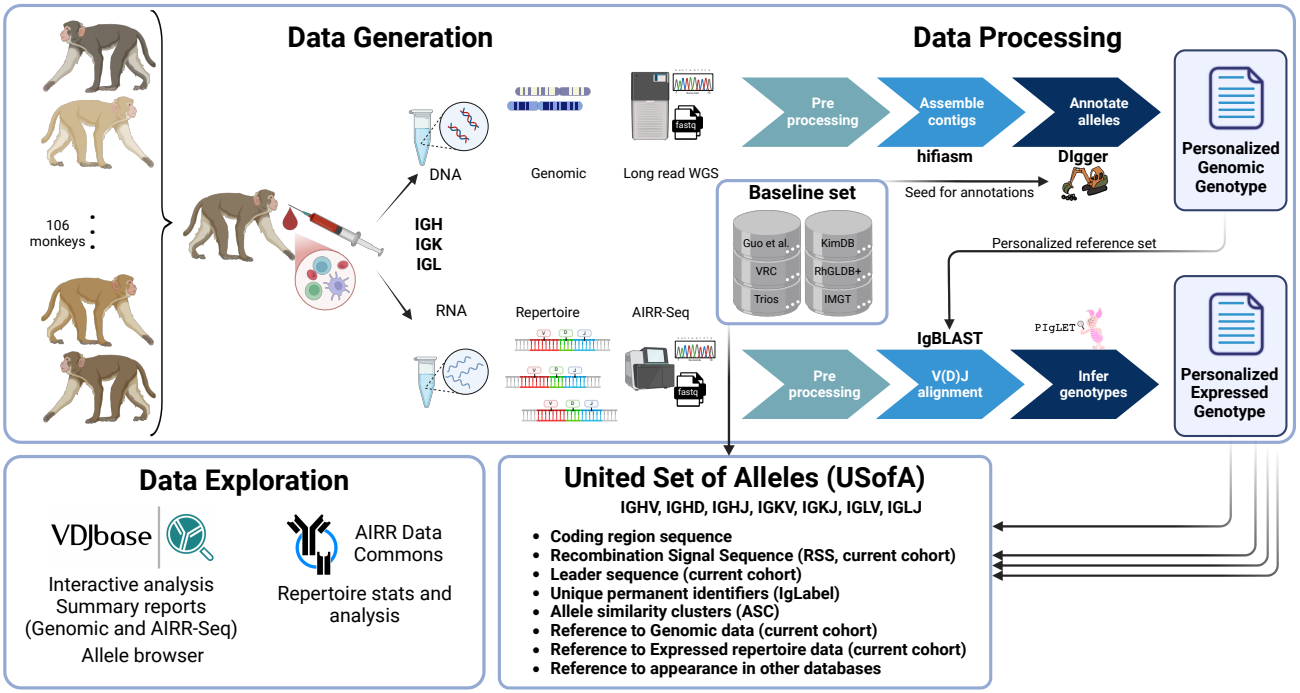

Figure 1: **Workflow** PBMCs were collected from 106 RMs for RNA and DNA extraction. RNA and genomic DNA were extracted, followed by sequencing of expressed B cell receptors and genomic IG-encoding loci. The resulting data were analyzed to identify personalized genotypes and expressed repertoires for each animal. These data were integrated with existing datasets of IG alleles to form the unified set of alleles (USofA). The exploration phase includes summary reports and interactive visualizations for comparing genomic and repertoire data.

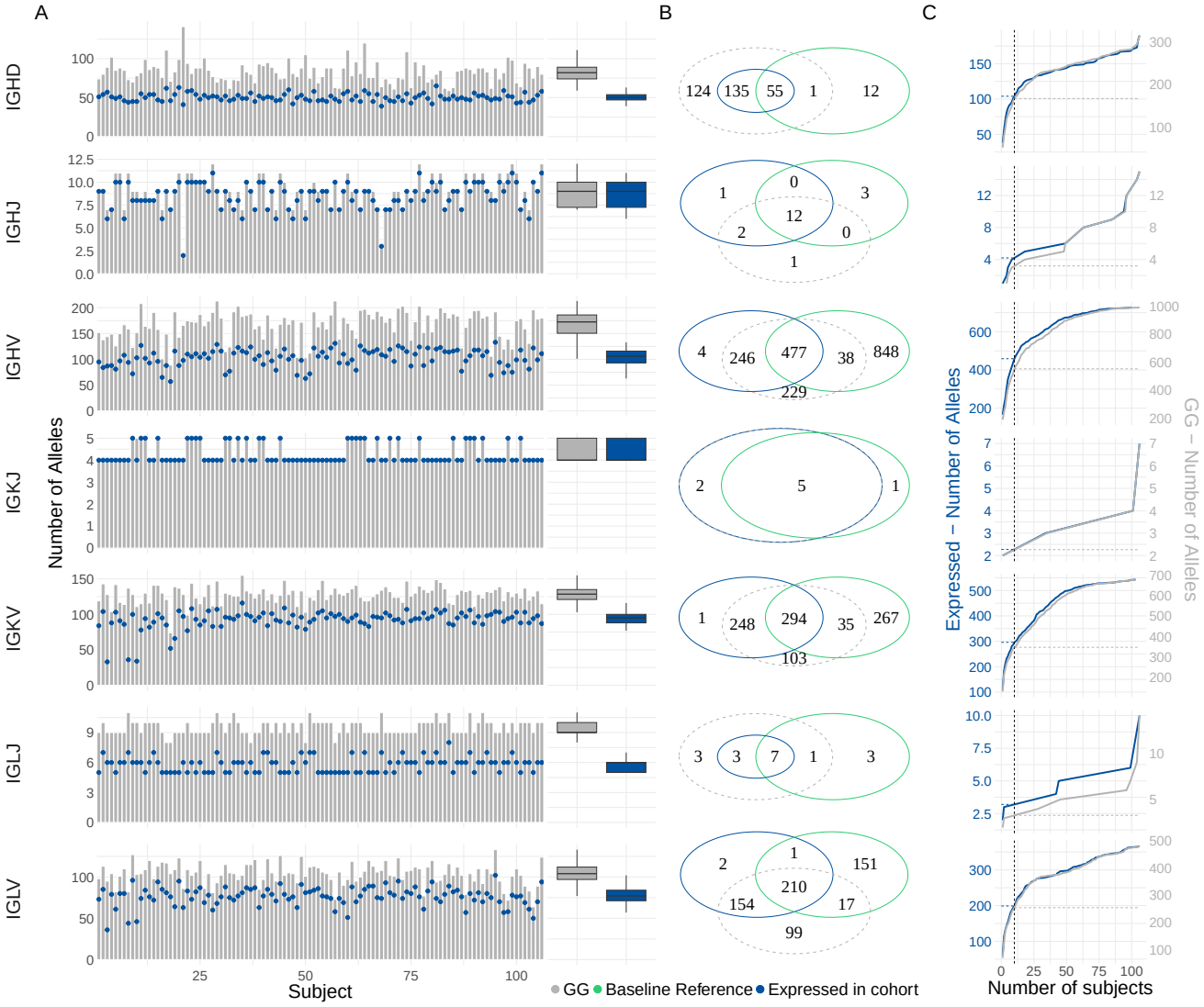

Figure 2: **Baseline and De Novo Allele Distribution in Genomic and Repertoire Data of the Current Cohort** (A) The left panel displays allele counts per sample, with the x-axis representing samples and the y-axis showing allele totals. Bars represent allele counts from the Genomic Genotype (GG), while blue points indicate repertoire genotype alleles identified using the personal GG reference. The right panel summarizes the distribution of allele counts across the two reference categories. (B) Venn diagrams showing the relationships between the set of alleles in the Baseline reference (green), the combined alleles from the GG references (gray), and the expressed alleles present in repertoire genotypes annotated using the personal GG reference(blue). (C) Cumulative distribution function curves illustrating allele sharing among subjects for each locus. The x-axis indicates the number of subjects sharing a given allele, while the y-axis represents the number of alleles shared by X or fewer subjects. The gray curve represents alleles found in the GGs (left axis scale), and the blue curve represents expressed alleles from repertoire genotypes annotated with the personal GG reference (right axis scale).

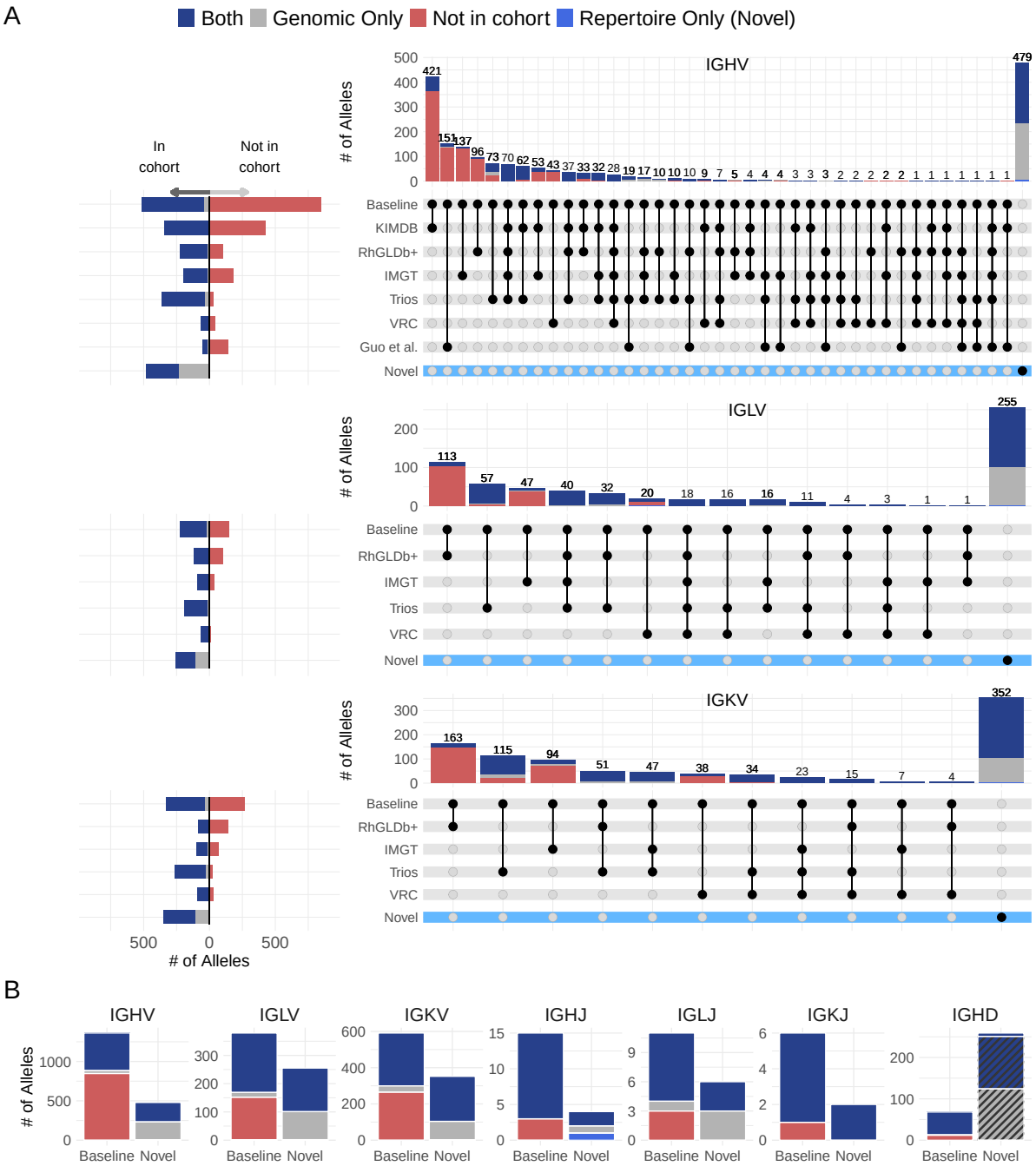

Figure 3: **Unified Set of Alleles Categorized by Source.** (A) UpSet plots showing intersections of IGHV, IGLV, and IGKV allele sets derived from various sources. The left-side horizontal bars indicate the number of alleles observed (left) or not observed (right) in the current cohort for each source. The top vertical bars represent the number of alleles unique to or shared among sources, as indicated by the black dots in each column. (B) Summary of the left-side data from the UpSet plots shown here (IGHV, IGLV, IGKV) and in Supplementary Figure 2 (IGHJ, IGLJ, IGKJ, and IGHD). Alleles are stratified based on their presence in the baseline reference used for cohort annotation (left) or inferred de novo from the cohort data (right). The bar color scheme for both panels is described in the legend at the top of the figure. The cross lines indicate alleles that were filtered out.

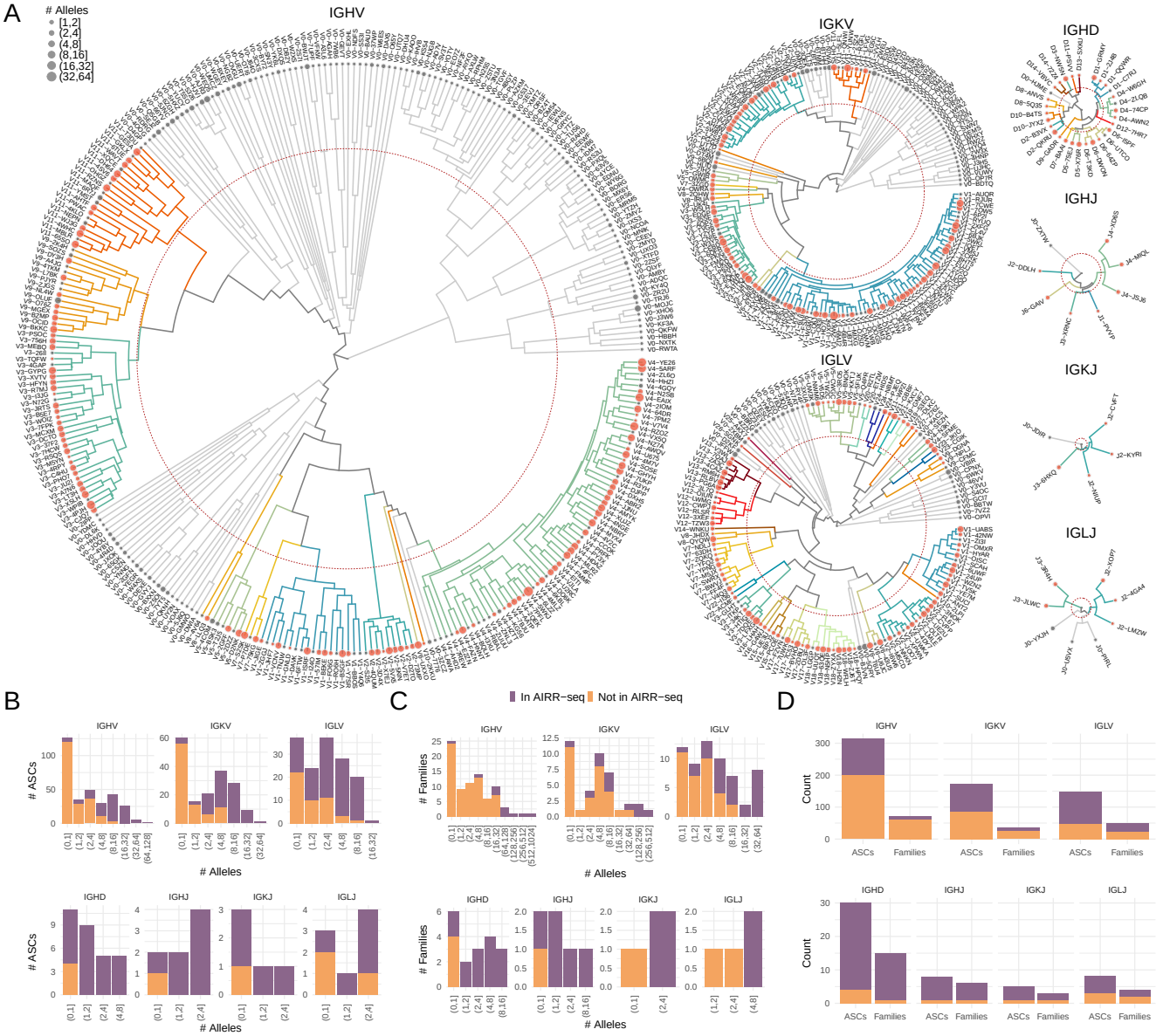

Figure 4: **Allele Similarity Clusters (ASCs) within the Unified Set of Alleles (USofA).** (A) Circular dendrograms illustrate the ASCs comprising the USofA. Each node represents an ASC, with the dot size indicating the number of alleles in the cluster and the color signifying whether a representative allele of the ASC is present (red) or absent (gray) in at least one repertoire within the cohort. The red dashed circular line marks the 75% similarity threshold used for family definition. Branch colors denote different families, while gray branches represent ASCs not observed in any repertoire and therefore lack family assignments. (B-C) Summary bar plots of the number of alleles per ASC (B) and per Family (C) for each immunoglobulin gene type. The x-axis represents the number of alleles grouped into logarithmic bins, indicating the size of each group in terms of allele counts. The y-axis represents the number of ASCs or Families corresponding to each allele bin. The color represents whether the ASC or Family was seen in the AIRR-seq data (color-coded in the legend). (D) Total ASCs and Families in each Ig gene type. The y-axis represents the total count, while the x-axis represents the categories. The color indicates whether the ASC or Family was observed in the AIRR-seq data (color-coded in the legend).

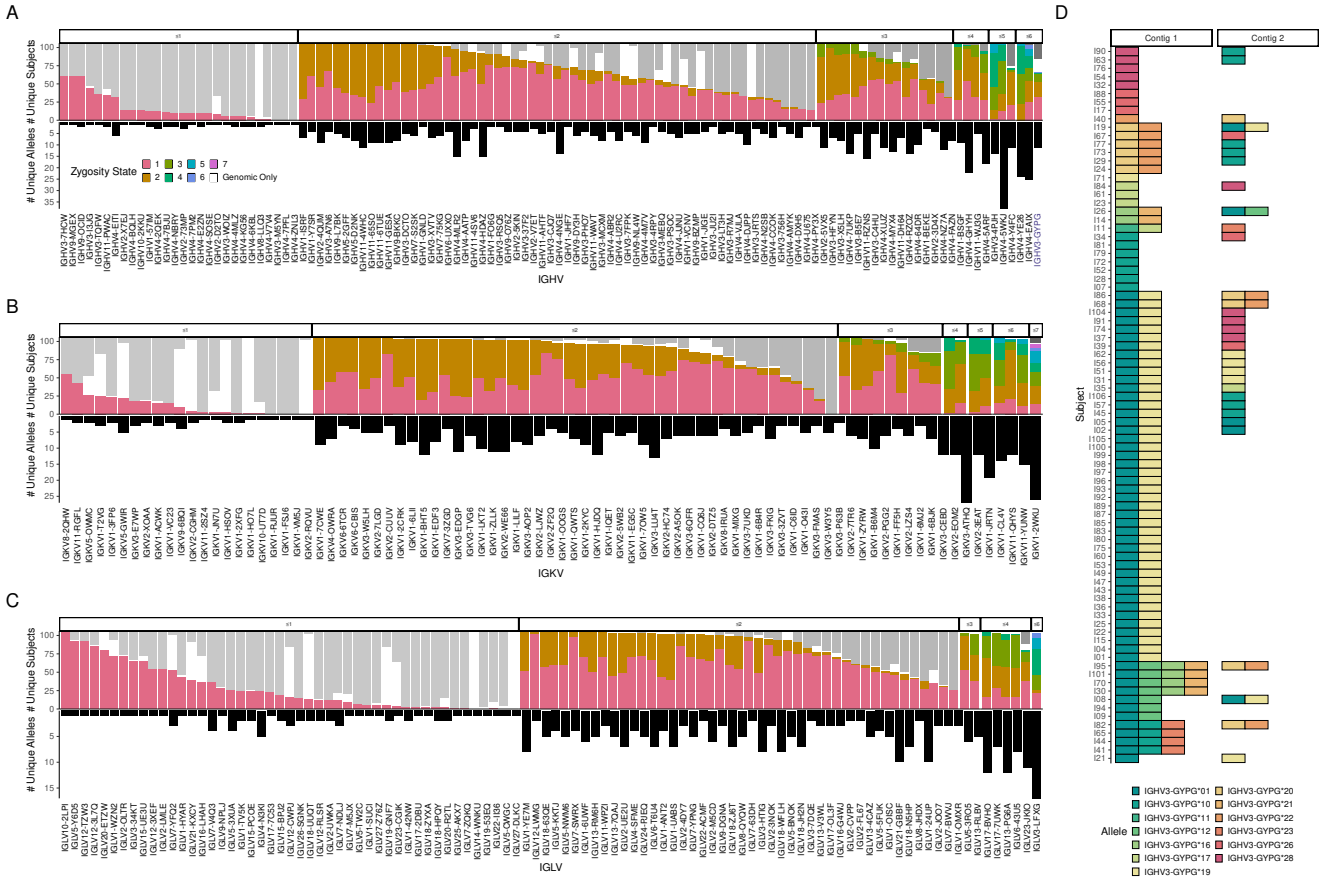

Figure 5: **Zygosity Events of Allele Similarity Clusters (ASCs) Across the Population.** (A-C) Bar plots illustrating zygosity events for the IG gene types IGHV, IGLV, and IGKV. The x-axis represents different ASCs for each IG gene type, while the top y-axis shows the number of subjects associated with each zygosity event, and the bottom y-axis displays the number of unique alleles detected in AIRR-seq data. The bar colors correspond to zygosity events, as indicated in the legend. (D) A tile plot showing the distribution of alleles within the ASC IGHV3-GYPG (Highlighted in blue in panel A) in genomic contigs across the population. The y-axis represents individual subjects, with each tile indicating a detected allele. The tile colors correspond to specific alleles, as denoted in the legend.

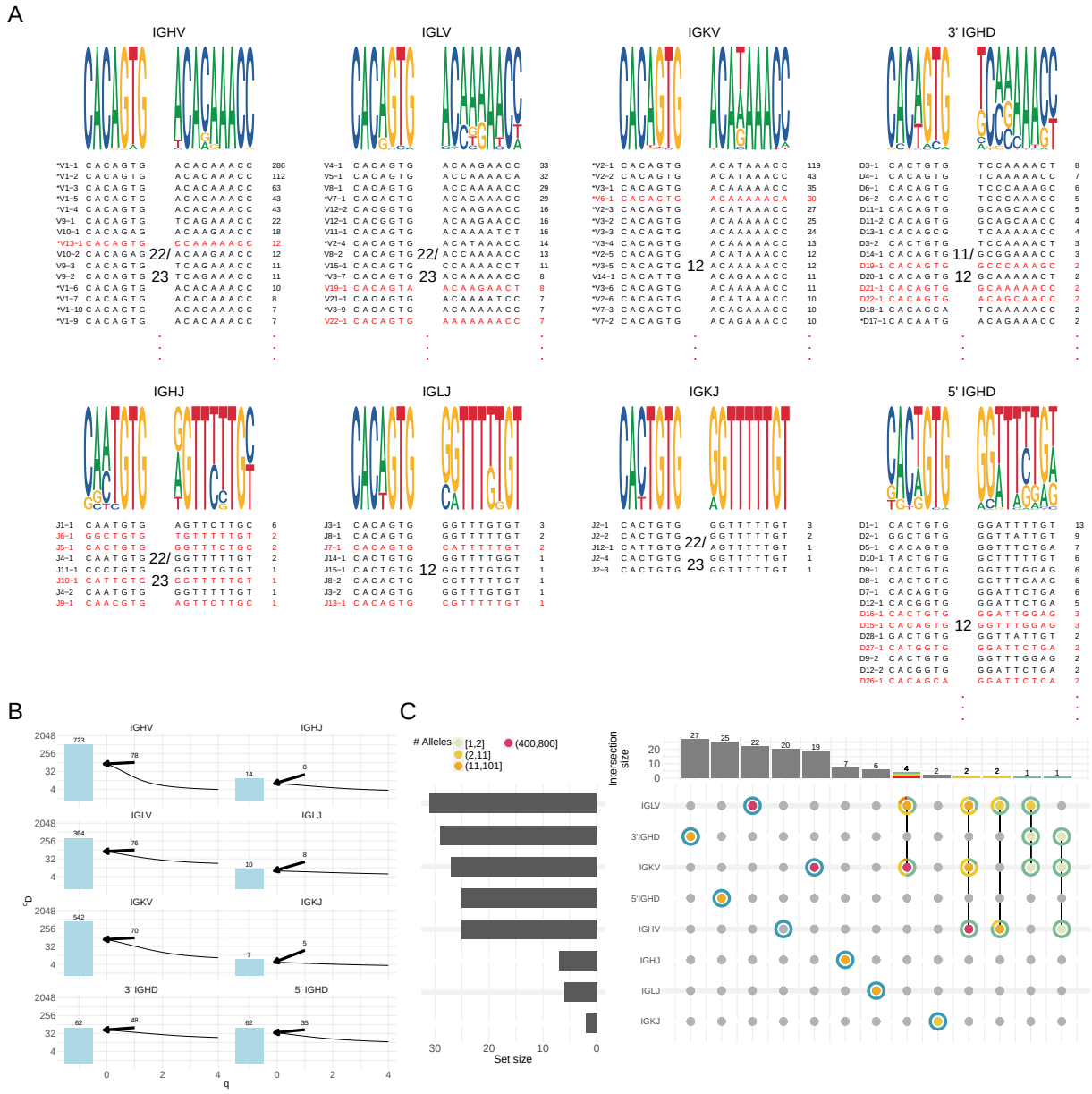

Figure 6: **Variability in Recombination Signal Sequence (RSS) Heptamer and Nonamer.** (A) Sequence logos (top) represent the variability of RSS heptamers (left) and nonamers (right) for each gene segment, weighted by the number of unique allele-RSS sequence pairs. The degeneracy of each RSS (bottom), measured by the count of unique allele-RSS sequence pairs, is displayed to the right of each heptamer-nonamer sequence. Each RSS is labeled on the left, with an asterisk indicating that the heptamer-nonamer sequence also occurs in other loci. Heptamer-nonamer sequences not observed in humans are highlighted in red. (B) Plots illustrate the number of unique alleles (bar) and the Hill diversity index (line) of the RSS groups for each gene segment. The y-axis uses a logarithmic scale, while the x-axis represents the Hill diversity index parameter q. Labels denote RSS richness (unique counts) at $q = 0$. (C) UpSet plots depicting intersections of shared heptamer-nonamer pairs among various loci. The horizontal bars on the left represent the number of unique heptamer-nonamer pairs for each locus. The vertical bars at the top illustrate the count of shared heptamer-nonamer pairs among loci, as indicated by the colored dots within each column. The color of the dots corresponds to the number of alleles carrying the shared heptamer-nonamer pair. Surrounding the colored dots, pie charts provide a visual representation of the fraction of alleles from each locus that contain the shared heptamer-nonamer pair.

References

[1] Kenneth Murphy and Casey Weaver. *Janeway's immunobiology*. Garland science, 2016.

[2] Valerie H Odegard and David G Schatz. Targeting of somatic hypermutation. *Nature Reviews Immunology*, 6(8):573–583, 2006.

[3] Andrew M Collins, Gur Yaari, Adrian J Shepherd, William Lees, and Corey T Watson. Germline immunoglobulin genes: disease susceptibility genes hidden in plain sight? *Current Opinion in Systems Biology*, 24:100–108, 2020.

[4] Moriah Gidoni, Omri Snir, Ayelet Peres, Pazit Polak, Ida Lindeman, Ivana Mikocziova, Vikas Kumar Sarna, Knut EA Lundin, Christopher Clouser, Francois Vigneault, et al. Mosaic deletion patterns of the human antibody heavy chain gene locus shown by bayesian haplotyping. *Nature communications*, 10(1):628, 2019.

[5] Oscar L Rodriguez, Yana Safonova, Catherine A Silver, Kaitlyn Shields, William S Gibson, Justin T Kos, David Tieri, Hanzhong Ke, Katherine JL Jackson, Scott D Boyd, et al. Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. *Nature communications*, 14(1):4419, 2023.

[6] Kerui Peng, Yana Safonova, Mikhail Shugay, Alice B Popejoy, Oscar L Rodriguez, Felix Breden, Petter Brodin, Amanda M Burkhardt, Carlos Bustamante, Van-Mai Cao-Lormeau, et al. Diversity in immunogenomics: the value and the challenge. *Nature methods*, 18(6):588–591, 2021.

[7] Jacob Glanville, Tracy C Kuo, H-Christian von Büdingen, Lin Guey, Jan Berka, Purnima D Sundar, Gabriella Huerta, Gautam R Mehta, Jorge R Oksenberg, Stephen L Hauser, et al. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proceedings of the National Academy of Sciences*, 108(50):20066–20071, 2011.

[8] Victor Greiff, Cédric R Weber, Johannes Palme, Ulrich Bodenhofer, Enkelejda Miho, Ulrike Menzel, and Sai T Reddy. Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. *The Journal of Immunology*, 199(8):2985–2997, 2017.

[9] Giulio Isacchini, Aleksandra M Walczak, Thierry Mora, and Armita Nourmohammad. Deep generative selection models of t and b cell receptor repertoires with sonnia. *Proceedings of the National Academy of Sciences*, 118(14):e2023141118, 2021.

[10] Yuval Avnir, Corey T Watson, Jacob Glanville, Eric C Peterson, Aimee S Tallarico, Andrew S Bennett, Kun Qin, Ying Fu, Chiung-Yu Huang, John H Beigel, et al. Ighv1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with ighv utilization shifts and varies by ethnicity. *Scientific reports*, 6(1):20842, 2016.

[11] Christina Yacoob, Marie Pancera, Vladimir Vigdorovich, Brian G Oliver, Jolene A Glenn, Junli Feng, D Noah Sather, Andrew T McGuire, and Leonidas Stamatatos. Differences in allelic frequency and cdrh3 region limit the engagement of hiv env immunogens by putative vrc01 neutralizing antibody precursors. *Cell reports*, 17(6):1560–1570, 2016.

[12] Todd A Johnson, Yoichi Mashimo, Jer-Yuarn Wu, Dankyu Yoon, Akira Hata, Michiaki Kubo, Atsushi Takahashi, Tatsuhiko Tsunoda, Kouichi Ozaki, Toshihiro Tanaka, et al. Association of an ighv3-66 gene variant with kawasaki disease. *Journal of human genetics*, 66(5):475–489, 2021.

[13] Pradeepa Pushparaj, Andrea Nicoletto, Daniel J Sheward, Hrishikesh Das, Xaquin Castro Dopico, Laura Perez Vidakovics, Leo Hanke, Mark Chernyshev, Sanjana Narang, Sungyong Kim, et al. Immunoglobulin germline gene polymorphisms influence the function of sars-cov-2 neutralizing antibodies. *Immunity*, 56(1):193–206, 2023.

[14] Christopher Sundling, Ganesh Phad, Iyadh Douagi, Marjon Navis, and Gunilla B Karlsson Hedestam. Isolation of antibody v (d) j sequences from single cell sorted rhesus macaque b cells. *Journal of immunological methods*, 386(1-2):85–93, 2012.

[15] Martin M Corcoran, Ganesh E Phad, Néstor Vázquez Bernat, Christiane Stahl-Hennig, Noriyuki Sumida, Mats AA Persson, Marcel Martin, and Gunilla B Karlsson Hedestam. Production of individualized v gene databases reveals high levels of immunoglobulin genetic diversity. *Nature communications*, 7(1):13642, 2016.

[16] Akshaya Ramesh, Sam Darko, Axin Hua, Glenn Overman, Amy Ransier, Joseph R Francica, Ashley Trama, Georgia D Tomaras, Barton F Haynes, Daniel C Douek, et al. Structure and diversity of the rhesus macaque immunoglobulin loci through multiple de novo genome assemblies. *Frontiers in immunology*, 8:1407, 2017.

[17] Kimberly M Cirelli, Diane G Carnathan, Bartek Nogal, Jacob T Martin, Oscar L Rodriguez, Amit A Upadhyay, Chiamaka A Enemuo, Etse H Gebru, Yury Choe, Federico Viviano, et al. Slow delivery immunization enhances hiv neutralizing antibody and germinal center responses via modulation of immunodominance. *Cell*, 177(5):1153–1171, 2019.

[18] Ronit Rosenfeld, Anat Zvi, Eitan Winter, Ronen Hope, Ofir Israeli, Ohad Mazor, and Gur Yaari. A primer set for comprehensive amplification of v-genes from rhesus macaque origin based on repertoire sequencing. *Journal of immunological methods*, 465:67–71, 2019.

[19] Christopher A Cottrell, Jelle van Schooten, Charles A Bowman, Meng Yuan, David Oyen, Mia Shin, Robert Morpurgo, Patricia van der Woude, Marielle van Breemen, Jonathan L Torres, et al. Mapping the immunogenic landscape of near-native hiv-1 envelope trimers in non-human primates. *PLoS pathogens*, 16(8):e1008753, 2020.

[20] Viviane Nguefack Ngoune, Morgane Bertignac, Maria Georga, Ariadni Papadaki, Alexandre Albani, Géraldine Folch, Joumana Jabado-Michaloud, Véronique Giudicelli, Patrice Duroux, Marie-Paule Lefranc, et al. Imgt® biocuration and analysis of the rhesus monkey ig loci. *Vaccines*, 10(3):394, 2022.

[21] Néstor Vázquez Bernat, Martin Corcoran, Izabela Nowak, Mateusz Kaduk, Xaquin Castro Dopico, Sanjana Narang, Pauline Maisonasse, Nathalie Dereuddre-Bosquet, Ben Murrell, and Gunilla B Karlsson Hedestam. Rhesus and cynomolgus macaque immunoglobulin heavy-chain genotyping yields comprehensive databases of germline vdj alleles. *Immunity*, 54(2):355–366, 2021.

[22] Yicheng Guo, Eric Waltari, Hong Lu, Zizhang Sheng, and Xueling Wu. Novel rhesus macaque immunoglobulin germline genes identified by three sequencing approaches. *Frontiers in Immunology*, 15:1506348, 2024.

[23] Christopher Sundling, Zhenhai Zhang, Ganesh E Phad, Zizhang Sheng, Yimeng Wang, John R Mascola, Yuxing Li, Richard T Wyatt, Lawrence Shapiro, and Gunilla B Karlsson Hedestam. Single-cell and deep sequencing of igg-switched macaque b cells reveal a diverse ig repertoire following immunization. *The Journal of Immunology*, 192(8):3637–3644, 2014.

[24] Richard A Gibbs, Jeffrey Rogers, Michael G Katze, Roger Bumgarner, George M Weinstock, Elaine R Mardis, Karin A Remington, Robert L Strausberg, J Craig Venter, Richard K Wilson, et al. Evolutionary and biomedical insights from the rhesus macaque genome. *science*, 316(5822):222–234, 2007.

[25] Aleksey V Zimin, Adam S Cornish, Mnirnal D Maudhoo, Robert M Gibbs, Xiongfei Zhang, Sanjit Pandey, Daniel T Meehan, Kristin Wipfler, Steven E Bosinger, Zachary P Johnson, et al. A new rhesus macaque assembly and annotation for next-generation sequencing analyses. *Biology direct*, 9:1–15, 2014.

[26] Wesley C Warren, R Alan Harris, Marina Haukness, Ian T Fiddes, Shwetha C Murali, Jason Fernandes, Philip C Dishuck, Jessica M Storer, Muthuswamy Raveendran, LaDeana W Hillier, et al. Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science*, 370(6523):eabc6617, 2020.

[27] Gregory Q Del Prete, Jeffrey D Lifson, and Brandon F Keele. Nonhuman primate models for the evaluation of hiv-1 preventive vaccine strategies: model parameter considerations and consequences. *Current opinion in HIV and AIDS*, 11(6):546–554, 2016.

[28] Barton F Haynes, Kevin Wiehe, Persephone Borrow, Kevin O Saunders, Bette Korber, Kshitij Wagh, Andrew J McMichael, Garnett Kelsoe, Beatrice H Hahn, Frederick Alt, et al. Strategies for hiv-1 vaccines that induce broadly neutralizing antibodies. *Nature Reviews Immunology*, 23(3):142–158, 2023.

[29] David J Leggat, Kristen W Cohen, Jordan R Willis, William J Fulp, Allan C Decamp, Oleksandr Kalyuzhniy, Christopher A Cottrell, Sergey Menis, Greg Finak, Lamar Ballweber-Fleming, et al. Vaccination induces hiv broadly neutralizing antibody precursors in humans. *Science*, 378(6623):eadd6502, 2022.

[30] Allan C Decamp, Martin M Corcoran, William J Fulp, Jordan R Willis, Christopher A Cottrell, Daniel LV Bader, Oleksandr Kalyuzhniy, David J Leggat, Kristen W Cohen, Ollivier Hyrien, et al. Human immunoglobulin gene allelic variation impacts germline-targeting vaccine priming. *npj Vaccines*, 9(1):58, 2024.

[31] Colin Havenar-Daughton, Diane G Carnathan, Archana V Boopathy, Amit A Upadhyay, Ben Murrell, Samantha M Reiss, Chiamaka A Enemuo, Etse H Gebru, Yury Choe, Pallavi Dhadvai, et al. Rapid germinal center and antibody responses in non-human primates after a single nanoparticle vaccine immunization. *Cell reports*, 29(7):1756–1766, 2019.

[32] Jon M Steichen, Ivy Phung, Eugenia Salcedo, Gabriel Ozorowski, Jordan R Willis, Sabyasachi Baboo, Alessia Liguori, Christopher A Cottrell, Jonathan L Torres, Patrick J Madden, et al. Vaccine priming of rare hiv broadly neutralizing antibody precursors in nonhuman primates. *Science*, 384(6697):eadj8321, 2024.

[33] Kevin O Saunders, James Counts, Bhishem Thakur, Victoria Stalls, Robert Edwards, Kartik Manne, Xiaozhi Lu, Katayoun Mansouri, Yue Chen, Rob Parks, et al. Vaccine induction of cd4-mimicking hiv-1 broadly neutralizing antibody precursors in macaques. *Cell*, 187(1):79–94, 2024.

[34] Oscar L. Rodriguez, William S. Gibson, Tom Parks, Matthew Emery, James Powell, Maya Strahl, Gintaras Deikus, Kathryn Auckland, Evan E. Eichler, Wayne A. Marasco, Robert Sebra, Andrew J. Sharp, Melissa L. Smith, Ali Bashir, and Corey T. Watson. A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. *Frontiers in Immunology*, 11:2136, 2020.

[35] William S Gibson, Oscar L Rodriguez, Kaitlyn Shields, Catherine A Silver, Abdullah Dorgham, Matthew Emery, Gintaras Deikus, Robert Sebra, Evan E Eichler, Ali Bashir, et al. Characterization of the immunoglobulin lambda chain locus from diverse populations reveals extensive genetic variation. *Genes & Immunity*, 24(1):21–31, 2023.

[36] Eric Engelbrecht, Oscar L Rodriguez, Kaitlyn Shields, Steven Schultze, David Tieri, Uddalok Jana, Gur Yaari, William D Lees, Melissa L Smith, and Corey T Watson. Resolving haplotype variation and complex genetic architecture in the human immunoglobulin kappa chain locus in individuals of diverse ancestry. *Genes & Immunity*, pages 1–10, 2024.

[37] William D Lees, Swati Saha, Gur Yaari, and Corey T Watson. Digger: directed annotation of immunoglobulin and t cell receptor v, d, and j gene sequences and assemblies. *Bioinformatics*, 40(3):btae144, 2024.

[38] Rui Kong, Hongying Duan, Zizhang Sheng, Kai Xu, Priyamvada Acharya, Xuejun Chen, Cheng Cheng, Adam S Dingens, Jason Gorman, Mallika Sastry, et al. Antibody lineages with vaccine-induced antigen-binding hotspots develop broad hiv neutralization. *Cell*, 178(3):567–584, 2019.

[39] Ayelet Peres, William D Lees, Oscar L Rodriguez, Noah Y Lee, Pazit Polak, Ronen Hope, Meirav Kedmi, Andrew M Collins, Mats Ohlin, Steven H Kleinstein, Corey T Watson, and Gur Yaari. IGHV allele similarity clustering improves genotype inference from adaptive immune receptor repertoire sequencing data. *Nucleic Acids Research*, page gkad603, 08 2023.

[40] William D Lees, Scott Christley, Ayelet Peres, Justin T Kos, Brian Corrie, Duncan Ralph, Felix Breden, Lindsay G Cowell, Gur Yaari, Martin Corcoran, et al. Airr community curation and standardised representation for immunoglobulin and t cell receptor germline sets. *ImmunoInformatics*, 10:100025, 2023.

[41] Jinghe Huang, Byong H Kang, Elise Ishida, Tongqing Zhou, Trevor Griesman, Zizhang Sheng, Fan Wu, Nicole A Doria-Rose, Baoshan Zhang, Krisha McKee, et al. Identification of a cd4-binding-site antibody to hiv that evolved near-pan neutralization breadth. *Immunity*, 45(5):1108–1121, 2016.

[42] Joseph R Francica, Zizhang Sheng, Zhenhai Zhang, Yoshiaki Nishimura, Masashi Shingai, Akshaya Ramesh, Brandon F Keele, Stephen D Schmidt, Barbara J Flynn, Sam Darko, et al. Analysis of immunoglobulin transcripts and hypermutation following shivad8 infection and protein-plus-adjuvant immunization. *Nature communications*, 6(1):6565, 2015.

[43] Jeffrey Rogers. Genomic resources for rhesus macaques (macaca mulatta). *Mammalian Genome*, 33(1):91–99, 2022.

[44] Haoyu Cheng, Gregory T Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods*, 18(2):170–175, 2021.

[45] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[46] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[47] Jason A Vander Heiden, Gur Yaari, Mohamed Uduman, Joel NH Stern, Kevin C O'connor, David A Hafler, Francois Vigneault, and Steven H Kleinstein. presto: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, 30(13):1930–1932, 2014.

[48] Jian Ye, Ning Ma, Thomas L. Madden, and James M. Ostell. Igblast: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, 41(W1):W34–W40, May 2013.

[49] Daniel Gadala-Maria, Gur Yaari, Mohamed Uduman, and Steven H. Kleinstein. Automated analysis of high-throughput b-cell sequencing data reveals a high frequency of novel immunoglobulin v gene segment alleles. *Proceedings of the National Academy of Sciences*, 112(8):E862–E870, 2015.

[50] Namita T Gupta, Jason A Vander Heiden, Mohamed Uduman, Daniel Gadala-Maria, Gur Yaari, and Steven H Kleinstein. Change-o: a toolkit for analyzing large-scale b cell immunoglobulin repertoire sequencing data. *Bioinformatics*, 31(20):3356–3358, 2015.