Systematic Profiling of Full-Length Ig and TCR Repertoire Diversity in Rhesus Macaque through Long Read Transcriptome Sequencing

Hayden N. Brochu,*,† Elizabeth Tseng,‡ Elise Smith,§ Matthew J. Thomas,§,¶ Aiden M. Jones,*,‖ Kayleigh R. Diveley,*,‖ Lynn Law,§,¶ Scott G. Hansen,# Louis J. Picker,# Michael Gale, Jr.,§,¶,** and Xinxia Peng*,†,††

The diversity of Ig and TCR repertoires is a focal point of immunological studies. Rhesus macaques (*Macaca mulatta*) are key for modeling human immune responses, placing critical importance on the accurate annotation and quantification of their Ig and TCR repertoires. However, because of incomplete reference resources, the coverage and accuracy of the traditional targeted amplification strategies for profiling rhesus Ig and TCR repertoires are largely unknown. In this study, using long read sequencing, we sequenced four Indian-origin rhesus macaque tissues and obtained high-quality, full-length sequences for over 6000 unique Ig and TCR transcripts, without the need for sequence assembly. We constructed, to our knowledge, the first complete reference set for the constant regions of all known isotypes and chain types of rhesus Ig and TCR repertoires. We show that sequence diversity exists across the entire variable regions of rhesus Ig and TCR transcripts. Consequently, existing strategies using targeted amplification of rearranged variable regions comprised of V(D)J gene segments miss a significant fraction (27–53% and 42–49%) of rhesus Ig/TCR diversity. To overcome these limitations, we designed new rhesus-specific assays that remove the need for primers conventionally targeting variable regions and allow single cell level Ig and TCR repertoire analysis. Our improved approach will enable future studies to fully capture rhesus Ig and TCR repertoire diversity and is applicable for improving annotations in any model organism. *The Journal of Immunology*, 2020, 204: 3434–3444.

R hesus macaque (*Macaca mulatta*) is one of the most commonly used and best-studied nonhuman primate (NHP) animal models. NHPs are key to studying human biology and human diseases because of their close phylogenetic relationship and similar physiology to humans (1–4). They are frequently used for vaccine development (5) and to model infection with human pathogens, such as *Mycobacterium tuberculosis* (6, 7), HIV (8, 9), influenza A virus (10), and Zika virus (11), among others (12, 13). Developing complete and accurate NHP genomic resources, especially for the immune system, is imperative for efficient translational interpretation (14, 15).

Critical for mounting adaptive immune responses, Ig and TCR repertoires house an enormous amount of diversity responsible for

recognizing a near limitless array of Ags presented through environmental exposures. Ig and TCR have two domains: a C region and a V region, which is comprised of a variable (V), joining (J), and in some cases, a diversity (D) gene segment. These gene segments are duplicated in several large loci in the genome, making their correct assembly a major technical challenge, one that has yet to be fully resolved in rhesus macaques despite significant improvements (16) and the recent release of a new genome assembly, rheMac10 (GCA_003339765.3). As demonstrated by a recent vaccine-related study in rhesus macaque (17), correct assembly of these complex regions requires much longer sequencing reads. Standard databases (e.g., the international ImMunoGeneTics information system [IMGT]) for these diverse sequences also

*Department of Molecular Biomedical Sciences, North Carolina State University College of Veterinary Medicine, Raleigh, NC 27607; †Bioinformatics Graduate Program, North Carolina State University, Raleigh, NC 27695; ‡Pacific Biosciences, Menlo Park, CA 94025; §Department of Immunology, University of Washington, Seattle, WA 98109; ¶Center for Innate Immunity and Immune Diseases, University of Washington, Seattle, WA 98109; ‖Genetics Graduate Program, North Carolina State University, Raleigh, NC 27695; #Vaccine and Gene Therapy Institute, Oregon Health & Science University, Beaverton, OR 97006; **Washington National Primate Research Center, University of Washington, Seattle, WA 98121; and ††Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695

ORCIDs: 0000-0002-9226-3910 (A.M.J.); 0000-0003-0176-7732 (K.R.D.); 0000-0003-2820-7428 (L.L.); 0000-0002-6332-7436 (M.G.); 0000-0002-0117-3582 (X.P.).

remain fairly meager relative to their human counterparts, although there are new tools developed to address these gaps (18). Because the design of available rhesus-specific assays for profiling Ig and TCR diversity has heavily relied on these limited rhesus reference resources, the coverage and accuracy of these assays still require unbiased assessment, and potentially improved approaches might be necessary. We begin by briefly summarizing the complexity of these immune repertoires in general and how this complexity has constrained the development of resources and assays for rhesus macaque.

In individual B and T cells, different gene segments of both variable and constant regions are combined at the DNA level to encode distinctive functional Ig and TCR genes through a process known as somatic recombination (reviewed in Ref. 19–21). This process accounts for the majority of the diversity within the V region domain, with the number of unique Ig and TCR V region domains estimated to be on the order of 10^{13} and 10^{18}, respectively (22). This diversity is further increased by chain pairing within Ig and TCR (23) and through somatic hypermutation (24).

The genetic diversity of the Ig and TCR loci present a unique challenge for accurate measurement. Traditionally, these immune repertoires are targeted for amplification either by multiplex PCR (MPCR) (25), RNA capture (26), or 5′ RACE (27). For rhesus macaque, many Ig repertoire sequencing efforts use a MPCR approach (28, 29), whereas some employ 5′ RACE (30). Typically, such PCR-based approaches are designed for individually sorted B or T cells (28) and facilitate cloning efforts (31). A more recent rhesus-specific MPCR design aimed to expand coverage of the Ig repertoire (32). There were also attempts to improve rhesus V and J germline gene annotation using 5′ RACE sequencing (RACE-seq) (18, 33). The recently developed IgDiscover tool (18) now makes it possible to leverage germline databases of related species to improve those of model organisms, for example using human germline databases to study Indian- and Chinese-origin rhesus macaques. Authors from the same laboratory also developed a strategy for targeting the 5′ untranslated region (UTR) of V genes, often conserved among these gene clusters, thus reducing the number of primers needed to target the V region in humans (34). However, rhesus MPCR amplification systems are inherently biased to the V and J gene segments they target because the primers are designed based on the consensus of a limited number of reference sequences (35).

RACE-seq has the advantage of only targeting the C region, where primer design is significantly easier given the low sequence variability. 5′ RACE in addition to MPCR have also been modified to incorporate unique molecular identifiers for repertoire sequencing (36, 37), mitigating issues arising from PCR bias (38). However, such protocols are optimized for applications in humans and mice (39) and have not yet been applied for Ig/TCR analysis in rhesus macaques. Moreover, even with MPCR or 5′ RACE, it is still difficult to capture complete Ig mRNA transcripts with the commonly available Illumina high-throughput sequencing instruments like the 2 × 250 bp HiSeq and the 2 × 300 bp MiSeq, in part because of the large length variability of recombined molecules (40) and the potential for noncanonical recombination events (41). Library preparation techniques have recently been improved to support this capability but only in humans (34). Despite these technical advances, it is still unclear how well these strategies perform in rhesus macaques, in which the relatively limited resources preclude similar benchmarking efforts previously done for human Ig repertoires (42) and TCR repertoires (43).

The technical robustness and unique advantages of single cell RNA sequencing (scRNA-seq) (44, 45) have revolutionized the analysis of human immune systems (46–50), resulting in an increasing number of studies for single cell–based Ig and TCR repertoire sequencing in human (51–54). Single cell protocols target repertoire sequences similarly to RACE-seq by only targeting the constant regions of sequences and by incorporating unique molecular identifiers. Interrogation of Ig/TCR repertoire and B/T cell transcriptomes in the same single cells has provided novel insights in humans, such as evaluating T cell clones within tumors (51), modeling of epitope specificities (52, 53), and assessment of host–microbiome immune homeostasis (54). However, to the best of our knowledge, for NHP models including rhesus macaques, equivalent assays for single cell Ig and TCR repertoire analysis are still not available.

In this study, we sought to address the technical limitations of existing immune repertoire sequencing protocols in rhesus macaques. To circumvent the highly difficult task of correctly assembling Ig/TCR transcripts from short sequencing reads, we performed long read transcriptome sequencing of rhesus tissue samples. We constructed, to our knowledge, the first complete reference set for the constant regions of all known isotypes and chain types of rhesus Ig and TCR repertoires from high-quality, full-length Ig and TCR transcript sequences. Our resource provides the multiple rhesus Ig and TCR constant regions still missing in the current IMGT database, avoiding complications due to incomplete assemblies of their respective genomic loci. Using this unbiased collection of full-length Ig and TCR transcript sequences, we were able to examine the diversities across the entire variable regions of rhesus Ig and TCR transcripts and the coverage of available assays for profiling rhesus Ig and TCR repertoires. Further, we designed new rhesus-specific assays that allow much broader rhesus Ig and TCR repertoire analysis, both at the single cell level and using 5′ RACE. These results furthermore demonstrate that immunoprofiling resources and assays can be developed for other organisms similarly without complete genomic assemblies of immune loci.

Materials and Methods

Tissue sample preparation and transcriptome sequencing

To construct cDNA libraries for Pacific Biosciences (PacBio) single molecule real-time (SMRT) sequencing, we used previously isolated Indian-origin rhesus macaque RNA from four tissue types: lymph node (LN), PBMCs, whole blood (WB), and rectal biopsy (RB). To maximize the coverage of transcript diversity, we pooled RNA from multiple Indian-origin rhesus macaques. The pooled WB RNA was from six healthy, uninfected macaques, whereas the pooled LN, PBMC, and RB RNA were from 18 macaques infected with SIVmac251. The LN RNA came from a combination of peripheral and proximal mesenteric LNs. Using a Clontech SMARTer kit, cDNA was produced from each pooled RNA sample followed by PCR amplification. The amplified cDNA samples were size selected using BluePippin (Sage Science) into four size fractions: 1–2 kb, 2–3 kb, 3–6 kb, and >6 kb. Libraries were prepared using the SMRTbell Template Prep Kit 1.0 and sequenced on the PacBio RS II platform with P6-C4 chemistry and 4-h movie times at the University of Washington PacBio Sequencing Services site.

Data processing

Raw PacBio sequencing data were first run through the circular consensus sequence (CCS) protocol in SMRT Analysis 2.3.0 to generate CCS reads. Each CCS read is the CCS of a single molecule. CCS reads were classified as non–full length and full length, in which the latter has all of the following detected: polyadenylation tail, 5′ cDNA primer, and 3′ cDNA primer. Full-length CCS sequences were aligned to Ig and TCR databases using IgBLAST v1.8.0 (55) with an e-value cutoff of 0.001 and IMGT/V-QUEST annotation release 201948-5 (56, 57). Separate searches were carried out using either human or rhesus macaque germline V, D, and J sequences where available (both for Ig, only human for TCR). Full-length Ig hits were kept for downstream analyses when they were deemed functional by IgBLAST output, meaning they lacked premature stop codons and were in-frame. When Ig hits had both rhesus and human database hits, the rhesus annotation was used in all subsequent analyses.

Generation of consensus C region sequences

To validate the chain types reported by IgBLAST and gather additional information about rhesus Ig/TCR isotypes and their subclasses, we further processed full-length Ig and TCR C region sequences as below. Using the J region reported by IgBLAST, we extracted the C region sequence for each full-length sequence. Next, we clustered extracted C region sequences using CD-HIT v4.6.6 (58) with the following parameters: -c 0.97 -G 0 -aL 0.95 -AL 100 -aS 0.99 -AS 30 (https://github.com/Magdoll/cDNA_Cupcake/wiki/Tutorial:-Collapse-redundant-isoforms-without-genome). Ig and TCR sequences were clustered separately. We only kept clusters with at least three constituent sequences for downstream consensus sequence analysis. For those clusters with at least 10 sequences, we also performed a second round of more stringent clustering using a 99% identity cutoff (-c 0.99 with CD-HIT) to separate quality sequences from those with multiple indels and to identify potential allotypes. From the second round of more stringent clustering, consensus sequences were generated using only the sequences from the largest remaining clusters, with the goal of removing low quality sequences that formed many small clusters. When this second round of clustering resulted in multiple large clusters of comparable size, a consensus sequence was generated for each cluster. We used the cons tool from the EMBOSS v6.6.0 suite (59) to generate consensus sequences, identifying and removing any remaining insertions at predominantly gapped locations. Protein sequences were derived from the generated consensus sequences using the EMBOSS transeq tool (59), in which the frame with the longest protein sequence was kept. We verified their identities via global alignments of the coding regions of cDNA sequences to a database of IMGT human C region nucleotide coding regions (60) using the usearch_global tool in the USEARCH suite (61), considering only the best alignment. We aligned consensus sequences belonging to the same chain type (and isotype if applicable) using Clustal Omega (62) and removed any remaining redundancies in the consensus sequences after manual visualization and curation in UGENE (63).

Next, we used these databases to systematically classify putative Ig and TCR sequences using all CCS reads from the complete full-length isoform sequencing (Iso-Seq) dataset as input. CCS reads were separately globally aligned to the custom databases of consensus Ig and TCR C region sequences using the usearch_global tool (61) with a 90% identity threshold to ascertain chain type and isotype/subclass where appropriate. Sequences without large gaps (>10 bp) in their alignment were kept in the final assignment of Ig and TCR sequences, as it was unclear if those sequences with large gaps were rare alternatively spliced transcripts or the result of sequencing errors. We then compared the CCS reads confirmed as Ig/TCR by our custom databases with the CCS reads identified by the initial IgBLAST search. Based on these comparisons, the final set of Ig transcripts were filtered to only include CCS reads that were confirmed by the custom database and also IgBLAST hits deemed functional by rhesus IMGT annotation (Supplemental Fig. 1B). Further, the final set of TCR transcripts only required CCS reads to be confirmed by the custom database (Supplemental Fig. 1A), as it was unclear if IMGT functional annotation for humans would be informative for rhesus-derived transcripts. We quantified the rates of insertions and deletions in the constant regions of confirmed Ig and TCR sequences using the Concise Idiosyncratic Gapped Alignment Report strings in the global alignments of CCS reads against this custom constant reference database by usearch_global (61). Non-Ig/TCR CCS reads of this complete Iso-Seq dataset are being analyzed in a separate study.

Analysis of V–J usage and V region diversity

The frequencies of V–J combinations were measured using the IMGT annotation reported by IgBLAST (55, 60) for all final Ig transcript sequences ascertained in the above analysis. χ^2 Test for independence based on V–J combination frequencies was performed in R (64) and by requiring that at least 50% of elements in each row and column have at least 10 counts. The diversity of the V region sequences was determined by first removing the C region sequence determined by its alignment to the corresponding consensus C region sequences. Separate multiple sequence alignments of the most common V–J regions were performed using Clustal Omega (62), and the consensus profiles were visualized in R with ggplot2 (64, 65).

In silico PCR analysis

Common rhesus-specific MPCR assays for amplifying IgH, IgK, and IgL sequences (28, 29, 32) and TCR α (TCRA) and TCR β (TCRB) sequences (66, 67) were tested via in silico PCR analysis using the University of California, Santa Cruz isPcr tool (68). For those primer sets with a nested design (28, 29), only the inner primers were used for the analysis.

The standard inner and outer reverse primers used for the 10× B cell and T cell V(D)J assays were also assessed via in silico PCR analysis (68) in which a dummy adapter sequence was prepended to the 5′ end of sequences to enable analysis with the isPcr tool and only test reverse primer efficiencies. The efficiencies of rhesus-specific MPCR assays (28, 29, 32) in amplifying the V region (forward primer) and C region (reverse primer) were similarly tested by either appending or prepending a dummy adapter sequence, respectively. In all uses of the isPcr tool, default parameters were used, requiring a perfect match for the first 15 nt on the 3′ end of the primer.

Design of rhesus-specific primers and PCR validation

Ig and TCR C region inner and outer reverse primers were designed using National Center for Biotechnology Information primerBLAST (69) with target melting temperatures consistent with current 10× V(D)J assay forward primers. For those targeting multiple consensus sequences, the search space was constrained to consensus regions. To validate these 10× optimized reverse primers, primerBLAST (69) was used to design standard PCR assays each with a forward primer compatible with its respective 10× optimized reverse primer and target reference sequence. This was only done for custom primers in the V(D)J assays, whereas primers adopted from Sundling et al. (28) were assumed to be efficacious. One hundred nanomoles HPLC purified DNA oligonucleotides were ordered from Integrated DNA Technologies. Reverse transcription was performed using QIAGEN QuantiTect Reverse Transcription Kit (catalog no. 205313, Lot 157037104) following the manufacturer's protocol with 1 μg rhesus macaque LN tissue RNA. The resulting cDNA was then amplified using Takara Bio Titanium Taq PCR Kit (catalog no. 639210, Lot 1805465A) following the manufacturer's protocol (https://www.takarabio.com/assets/documents/User%20Manual/PT3304-2.pdf), altering the recommended cycling conditions with 30 cycles and an annealing temperature and time of 50°C and 1 min. Four hundred nanograms of each PCR was then run on a 10% TBE polyacrylamide gel alongside Thermo Fisher Scientific 100-bp GeneRuler to confirm product size from primer pairs.

Results

Generation of a complete reference collection for rhesus macaque Ig and TCR constant regions

Using PacBio transcriptome sequencing (the Iso-Seq method), we obtained over 2.8 million CCS reads from four different rhesus macaque tissues (Supplemental Table I). Each CCS read is the CCS of a single transcript molecule (70). About 33% of these CCS reads were full length (i.e., contained the 5′ cDNA primer, 3′ cDNA primer, and polyadenylation tail) (Supplemental Table I). Using the available IMGT annotation of the variable regions of Ig (human and rhesus) and TCR (human) germline sequences (60) and IgBLAST (55), we recovered 13,118 Ig and 2534 TCR putative transcript sequences, respectively, accounting for 1.4 and 0. 28% of the total number of full-length CCS reads (Supplemental Table I). Only functional Ig transcripts (i.e., those in-frame and without premature stop codons according to IMGT germline annotation) were used in downstream analyses (5666/13,118 or 43% of full-length Ig transcripts). We kept all full-length TCR transcripts at this step as it was unclear if their functionalities could be correctly determined with the use of human germline TCR reference sequences.

We first sought to characterize the constant regions of these full-length Ig and TCR transcript sequences. Because constant regions are significantly less variable, this would allow us to compile a complete reference set and accurately classify the different chain types and isotypes. Using the V(D)J annotation provided by IMGT (60), we extracted constant regions from the full-length Ig and TCR transcript sequences. We then clustered and curated these sequences to generate complete cDNA and coding sequences (*Materials and Methods*). Complete cDNA sequences represent the entire C region consensus (i.e., including the 3′ UTR), although the coding regions were generated at both the cDNA and protein level. We obtained the complete consensus sequences for the following chain types and isotypes with multiple references

Table I. Number of unique full-length rhesus Ig transcript sequences identified in each tissue

	RB	WB	PBMC	LN	Total
IgHA*01	1411 (8)	67 (0)	21 (2)	224 (0)	1723 (10)
IgHA*02	705 (4)	7 (0)	21 (1)	275 (6)	1008 (11)
IgHA*03	101	0	5	9	115
IgHD	0	12	4	3	19
IgHE	1	1	0	4	6
IgHG1	78	21	21	1030	1150
IgHG2	9	9	4	311	333
IgHG3	0	1	0	30	31
IgHG4	1	0	0	13	14
IgHM	129 (3)	20 (21)	18 (14)	169 (8)	336 (46)
IgK	97	30	12	155	294
IgL	100	30	18	140	288
Total	2632 (15)	198 (21)	124 (17)	2363 (14)	5317 (67)

Counts for secreted H chain Ig sequences (IgH A, D, E, G1-4, and M) are shown with membrane-bound counts in parentheses where applicable. IgHA is further classified by allotype, conventionally denoted by an asterisk followed by a number.

noted in parentheses: IgHA (5), IgHD, IgHE, IgHG1, IgHG2, IgHG3, IgHG4, IgHM (2), IgLC, IgKC, TCRAC, TCRBC1, TCRBC2, TCRDC, TCRGC1, and TCRGC2. We thus recovered complete reference sequences for all known secreted IgH isotypes and additionally two IgHA and one IgHM membrane-bound reference sequences. We numbered membrane-bound rhesus Ig reference sequences to be consistent with their corresponding secreted forms.

We identified three unique IgHA consensus sequences: two in both secreted and bound form and one in only secreted form. Much of the variation among these sequences was confined to the hinge region, which has previously been reported to be highly heterogeneous among rhesus macaques (71). In fact, we recovered three of the eight unique hinge regions previously identified by (72) (data not shown). We numbered these unique IgHA sequences based on the relative abundances we later determined within our samples (i.e., IgHA*01 was the most abundant), following conventional naming for allelic variants.

We also identified all known subclasses of IgHG, TCRBC, and TCRGC. We aligned the four IgHG cDNA coding sequences we recovered to the available rhesus IgHG genomic DNA in IMGT to determine their subclasses. Each cDNA coding sequence reference corresponded to one of the four IgHG subclasses with high sequence identities (98.6–99.9%). To properly classify the TCRBC subclasses, we aligned the two TCRBC protein sequences to the two reference open reading frames available in IMGT, yielding perfect matches (data not shown). We leveraged the available human open reading frames in IMGT to determine subclasses of TCRGC as reference sequences were not available for rhesus macaques. Using a multiple sequence alignment, we assigned orthologous subclass naming to the rhesus references based on the human reference with the highest structural similarity (data not shown).

Interestingly, 39% (979 of 2534) of putative TCR transcript sequences identified by the initial IgBLAST search were in fact Ig transcripts, evidenced by recapitulation of several Ig consensus sequences among the set of TCR clusters and by successful alignment of such sequences to the set of Ig reference sequences (Supplemental Fig. 1A). Although this did not prevent our complete recovery of TCR consensus sequences, it suggests that commonly used human reference–based IgBLAST searches could be inaccurate for rhesus TCR repertoire analysis.

Classification of rhesus Ig and TCR transcript sequences

Next, we used these consensus sequences of rhesus constant regions to classify the full-length rhesus Ig and TCR transcript sequences. When we aligned the constant regions of full-length

functional Ig transcripts to this newly constructed database of rhesus Ig cDNA consensus sequences, we obtained successful assignments for 96% (5415/5666) of these full-length Ig transcript sequences. Those transcript sequences that failed to align tended to be other molecules from the Ig superfamily that contain V-set domains (i.e., a V region), representing false positives from our initial IgBLAST screen (data not shown). Among those aligned, a small fraction (1%) of sequences had large structural differences, indicating infrequent alternative splicing events (data not shown). To simplify the downstream analysis, we removed these transcripts from further processing, yielding a final set of 5384 full-length Ig transcript sequences (Table I). Among four tissue samples sequenced in this study, the RB and LN had the largest number of unique Ig transcript sequences overall. Secreted IgHA and IgHG transcripts had the highest relative abundance and were most prevalent in the RB and LN samples. IgK and IgL sequences were significantly less abundant than IgH isotypes in general, but they were still detected across each tissue.

Similarly, we identified rhesus TCR transcript sequences by aligning raw full-length CCS reads directly to this custom collection of rhesus TCR cDNA consensus sequences. This precluded any bias against rhesus TCR transcript sequences that contained V, D, and J genes with low similarity to those in the human IMGT reference. In total, we assigned 741 full-length CCS reads as rhesus TCR transcript sequences (Table II), 50 of which (7%) we did not originally recover using the human IMGT reference database (Supplemental Fig. 1A). This suggests there was a significant divergence between human and rhesus TCR germline V(D)J genes, indicating that a species-specific germline database is necessary for complete repertoire recovery. We detected the fewest TCR transcript sequences in the RB sample, whereas the WB, PBMC, and LN samples each had larger recoveries of all TCR isotypes (Table II). TCRA was the most abundant in each

Table II. Number of unique full-length rhesus TCR transcript sequences identified in each tissue

	RB	WB	PBMC	LN	Total
TCRA	14	71	103	267	455
TCRB1	1	9	11	26	47
TCRB2	0	13	16	44	73
TCRD	1	1	27	10	39
TCRG1	2	3	6	5	16
TCRG2	3	11	75	22	111
Total	21	108	238	374	741

tissue, representing 61% of all sequences recovered (455/741). TCR γ (TCRG) 2 had the second highest abundance with its strongest representation in PBMCs (Table II).

We then assessed the sequence quality of these Ig and TCR transcripts by quantifying insertion and deletion events within the alignments of their C region sequences and the corresponding reference cDNA sequences. We elected not to evaluate the rate of substitutions as they are difficult to discern from allelic variation given our sequencing depth and are generally less common than insertions and deletions in CCS reads (33). The rate of insertions in Ig C region sequences was very low, with a mean of 0.11% and median of 0.06%. The mean and median deletion rates were slightly lower at 0.04 and 0%, respectively. Error rates within the TCR constant regions were comparable to those of Ig constant regions, with mean insertion and deletion rates of 0.15 and 0.07% and median rates of 0.07 and 0%, respectively. These low error rates reflected the high quality of the final set of full-length rhesus Ig and TCR transcript sequences we obtained in this study.

We found the recovery of TCR transcripts to be significantly more accurate and marginally more sensitive when using our custom TCR C region database for identification instead of the traditional IgBLAST approach with human germline annotation (Supplemental Fig. 1A). We thus elected to align CCS reads directly to our custom Ig C region database to make a similar comparison. We recovered 11,451 Ig transcripts using this strategy; of these, 5384 (47%) were functional IgBLAST hits (see Table I), 5858 (51%) were nonfunctional IgBLAST hits, and 209 (2%) were not detected by IgBLAST (Supplemental Fig. 1B). B cells that contain nonfunctional receptor sequences are known to apoptose early in their development; therefore, the proportion of nonfunctional IgBLAST hits observed in this study may be a reflection of the B cell population captured with these full-length transcripts. Many of the full-length Ig transcripts that eluded IgBLAST detection had enough sequence upstream of the C region to harbor a complete V region (127 or 1% overall) (Supplemental Fig. 1B), indicating that some V region genes may be significantly diverged from those annotated in IMGT.

Usage of known V, D, and J genes in Ig transcripts indicates broad coverage of rhesus Ig repertoire

To assess our overall coverage of the rhesus Ig repertoire, we examined the usage of individual V, D, and J genes among these full-length rhesus Ig sequences based on the alignment of their variable regions to rhesus Ig germline annotations in IMGT (60). We detected all known gene families of rhesus IgHV, IgHD, and IgHJ genes among the four tissue samples (Fig. 1). The majority of IgHV and IgHJ genes detected were from the IgHV4 and IgHJ4 families, respectively, although there was broad coverage of IgHD gene families. Among L chain gene families, IgKV1 and 2 as well as IgLV1, 2, and 3 were the most frequent. IgKJ families were relatively less skewed in frequency, although IgLJ1 was in a slightly higher frequency than other IgLJ families. The only known gene families we did not detect in these tissue samples were IgKV5, IgLJ4, and IgLJ5, likely because of the overall lower abundance of L chain Ig sequences we observed (Table I). Interestingly, relative frequencies of these different gene families were highly consistent across the four tissues analyzed in this study (Supplemental Fig. 2), despite drastically different sampling depth (Fig. 1). Only L chain gene families within PBMCs and WB deviated from this trend. Because we had significantly less detection of IgK and IgL sequences in these tissues (Table I), their observed deviations require further investigation.
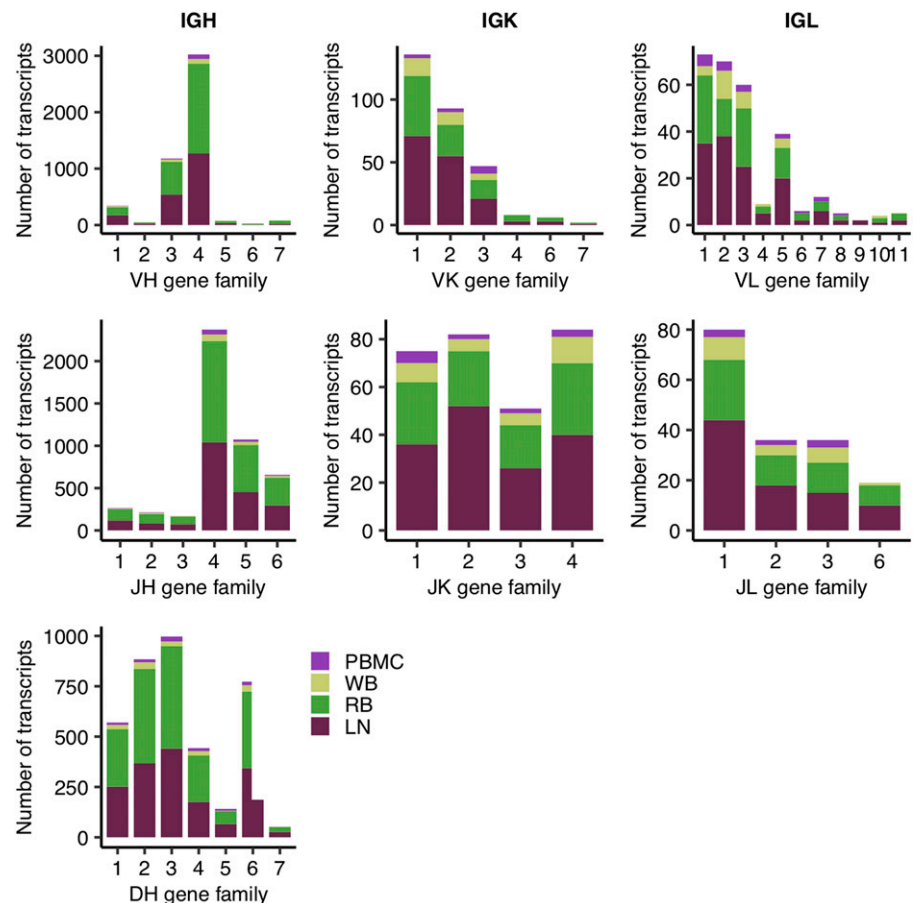
FIGURE 1. Stacked bar plot of absolute transcript abundance of V, D, and J genes for each Ig chain type, where colors of the bar indicate the tissue source of the gene. Variable gene abundance was computed using full-length Ig transcripts reported in Table I. Each individual plot shows the various known gene families of a particular gene segment type (e.g., VH), in which gene families are clusters of highly similar gene segments within that gene segment type.
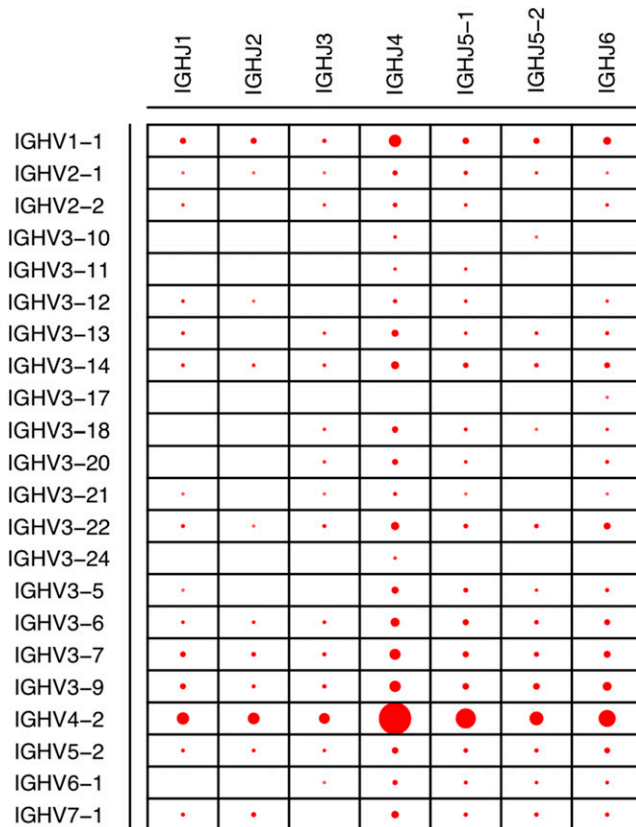
FIGURE 2. Dot plot of various IgHV and IgHJ combinations. The sizes of dots are proportional to the relative frequencies among the full-length Ig transcript sequences. The gene segments are conventionally named according to gene family (e.g., IgHV1) followed by the specific gene segment number. Gene segments simply have the family gene name (e.g., IgHJ1) when there is only one known gene segment in the family.

FIGURE 3. Consensus profiles of the most abundant IgHV genes detected. Panels (**A–C**) respectively show the percentage of nucleotides at each position that match the consensus (most abundant) nucleotide for IgHV4-2, IgHV1-1, and IgHV3-9. The IgHV4-2 consensus was constructed from 1383 sequences, whereas IgHV1-1 was from 163 sequences and IgHV3-9 was from 128 sequences. The regions targeted by primers from Sundling et al. (28), Wiehe et al. (29), and Rosenfeld et al. (32), respectively, have either a yellow, red, or green bar above.

Given the sufficient depth of coverage observed for IgH V, D, and J gene families, we calculated the combination frequencies of these V and J genes. Consistent with V and J gene frequencies observed in Fig. 1, the majority of V–J combination events contained IgHV4-2 and/or IgHJ4 (Fig. 2). Using the most abundant V and J genes, we tested the independence of V–J recombination events. We observed a significant nonrandom distribution of V–J recombination frequencies ($\chi^2 = 77.6$, df $= 15$, $p = 1.9e-10$). Furthermore, we discovered a strong positive correlation between the V–J recombination rates in the LN and RB samples (Pearson correlation $= 0.86$). Because these two samples had highly different compositions of IgH isotypes (Table I), we also reasoned that there might be no discernable association between the variable regions and constant regions in general.

Sequence diversity existing across the entire V region shows that MPCR amplification of rhesus immune repertoires is inherently biased

The collection of unbiased full-length Ig and TCR transcript sequences also provided a unique opportunity for exploring the sequence diversity across the entire V region, the target of current repertoire amplification efforts. We assessed the efficiencies of three commonly used rhesus-specific Ig MPCR strategies (28, 29, 32) by examining the sites targeted by their primers in the context of their respective Ig V genes. We selected the three most abundant IgHV genes represented among the unique Ig transcripts: IgHV4-2 (1383 sequences), IgHV1-1 (163 sequences), and IgHV3-9 (128 sequences) and generated consensus profiles for each. We discovered
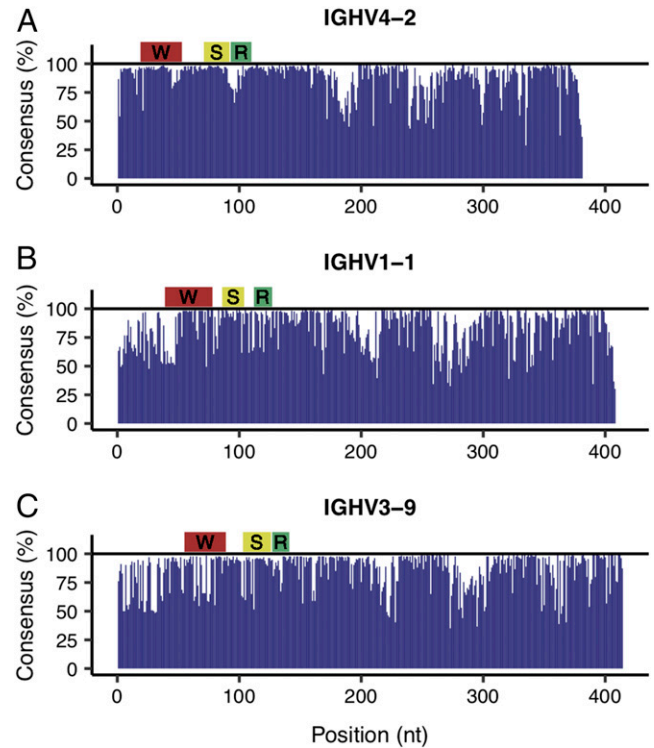
that the rhesus primers designed to target these genes all locate in regions rich in sequence diversity (Fig. 3). For example, Sundling et al. (28) targeted well-conserved regions of IgHV4-1 and IgHV1-1 (Fig. 3A, 3B) yet appeared to target a region in IgHV3-9 that had a low percentage of sequences sharing the same consensus nucleotides (Fig. 3C). The low consensus observed in all three profiles at the 5′ end of the cDNA sequence (left side of profiles) was likely a result of the lack of guaranteed 5′ capture among CCS reads. However, it was confined to regions upstream of the primer target sites (Fig. 3A–C) and thus did not affect our assessment of available primers.

To illustrate the limitation that the observed diversity could reduce the overall coverage of these MPCR approaches, we evaluated these primer sets via in silico PCR analysis with the full-length Ig transcript sequences as the source of potential template transcripts (Fig. 4). Similarly, we also leveraged known rhesus-specific MPCR strategies for TCR (66, 67) in this analysis (Fig. 4). We first tested only the forward (V gene) primers from these Ig and TCR primer sets (28, 29, 32, 66, 67) and found that IgH, IgK, and IgL amplification rates did not exceed 87% and were as low as 45% (Fig. 4A). Amplification rates for TCRA and TCRB were similar, ranging from 53 to 73% (Fig. 4A). The Ig primer set designed earliest by Sundling et al. (28) in 2012 had the highest recovery of IgH sequences (76%), whereas the most recent Ig primer set designed by Rosenfeld et al. (32) in 2019 had the highest recovery of IgK (87%) and IgL (71%). The TCRB-specific primers designed by Li et al. (67) recovered 73% of the TCRB sequences, whereas Chen et al. (66) recovered 53% of TCRA sequences.
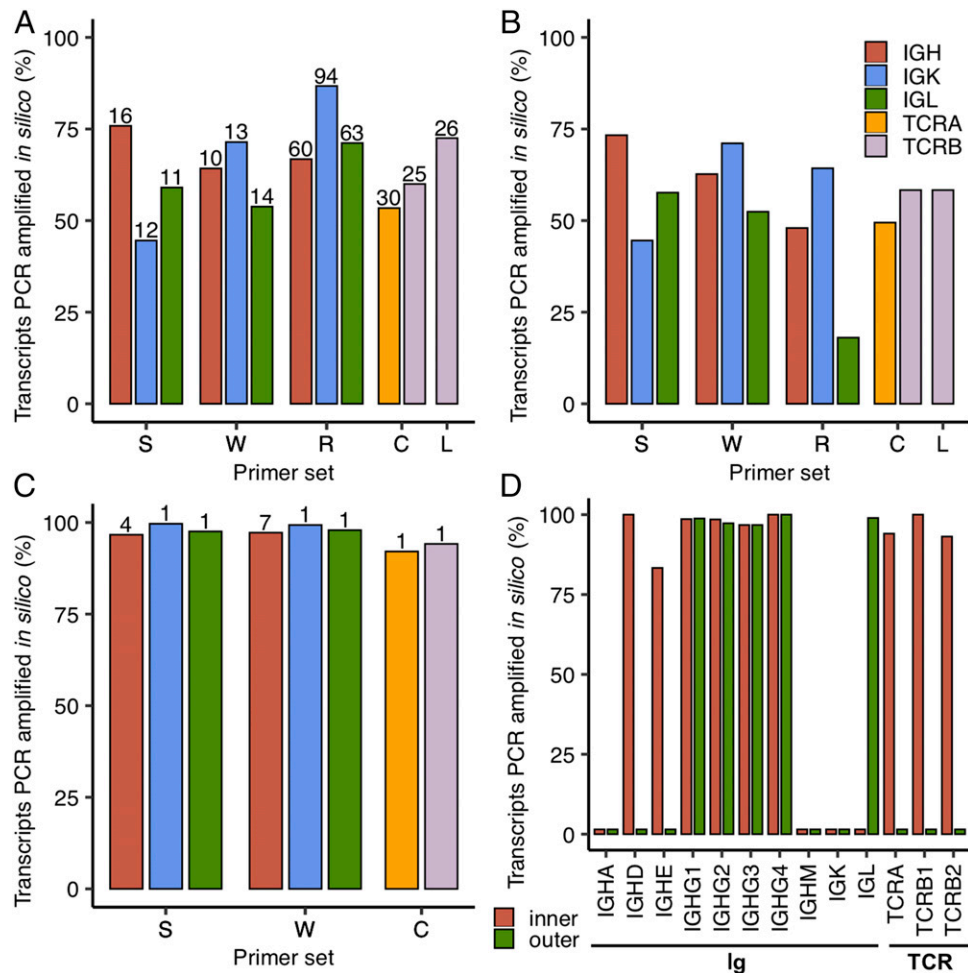
FIGURE 4. In silico PCR analysis of rhesus-specific primers from Sundling et al. (S) (28), Wiehe et al. (W) (29), Rosenfeld et al. (R) (32), Chen et al. (C) (66), and Li et al. (L) (67) as well as human 10× B and T cell V(D)J reverse primers (68). Only the inner primers were tested for S and W, whereas both inner and outer 10× primers were tested. The primers from C target both TCRA and TCRB, whereas those from L only target TCRB as indicated by the bars shown in (**A–C**). The percentage of transcripts amplified in silico are shown as bars for different primer sets and types of transcripts. (A) Amplification in which only primers targeting the V gene segment within the V region from S, W, R, C, and L are tested with the total number of primers (after enumerating primers with degenerate nucleotides) shown above the bars. (B) Amplification using forward (V gene) and reverse (J gene segment or C region) primers from S, W, R, C, and L. (C) Amplification in which only primers targeting the C region from S, W, and C are tested with the total number of primers (after enumerating primers with degenerate nucleotides) shown above the bars. R and L are not shown here as their reverse primers target J gene segments. (**D**) Analysis of 10× B and T cell V(D)J inner and outer reverse primers. The different groups of transcripts (Ig and TCR) are noted below the bar plot.

As expected, when we assessed both forward and reverse primers together, the low percentages of sequences amplified largely remained the same (Fig. 4B). Notably, the percentages for Rosenfeld et al. (32) and the percentage for Li et al. (67) were even lower, as their reverse primers targeted J gene segments instead of the C region. Indeed, when we tested the C region primers alone (Fig. 4C), we found that Sundling et al. (28), Wiehe et al. (29), and Chen et al. (66) had near perfect amplification rates, which was expected.

Design of rhesus-specific B and T cell V(D)J assays for single cell analysis

Both our consensus profiling (Fig. 3) and in silico PCR analysis (Fig. 4A–C) showed that primers targeting variable regions are the source of bias in targeted amplification of rhesus Ig and TCR repertoires. Other methods, particularly RACE-seq and those using scRNA-seq, avoid this problem by only using primers targeting the C region. Because there are no known single cell repertoire analysis assays for rhesus macaques, we first checked the efficiency of human B and T cell V(D)J assays developed by 10× Genomics (68) against our collection of Ig and TCR

transcript sequences. We found that neither assay was able to fully capture this set of rhesus Ig and TCR transcript sequences (Fig. 4D). The human inner primers successfully amplified IgHD (100%), IgHE (83%), IgHG1 (99%), IgHG2 (98%), IgHG3 (97%), IgHG4 (100%), TCRA (94%), TCRB1 (100%), and TCRB2 (93%). Meanwhile, the outer primers only amplified IgHG1 (99%), IgHG2 (97%), IgHG3 (97%), IgHG4 (100%), and IgL (99%). Notably, TCR δ (TCRD), TCRG1, and TCRG2 were not included in this analysis as 10× Genomics assays do not cover these transcripts.

We further assessed sequence homology between human and rhesus reference sequences for the constant regions. For each Ig and TCR isotype, we aligned our reference cDNA coding sequences for rhesus macaque with those available for humans and evaluated the compatibility of their respective 10× human primers where relevant. For example, we aligned human IgHA1 and IgHA2 to the five rhesus IgHA consensus sequences (three secreted, two membrane bound) and examined the regions targeted by 10× human primers (Fig. 5A). Overall, the two human cDNA coding sequences had a high similarity to the rhesus consensus sequences, averaging 89.7% identity. However, multiple nucleotide differences existed in the regions of rhesus reference sequences targeted
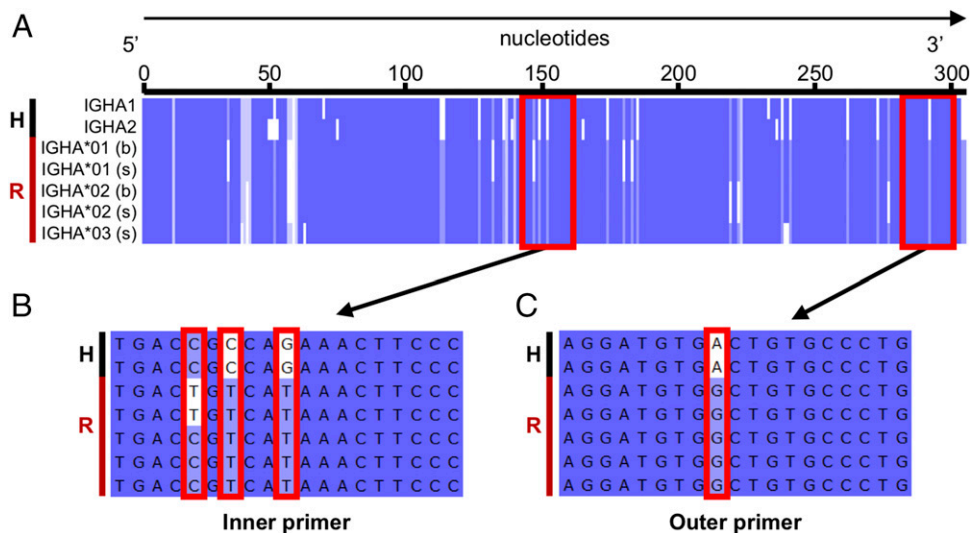
FIGURE 5. Comparison of human and rhesus IgHA reference sequences and illustration of human 10× primer target regions. A black H and red R, respectively, highlight human and rhesus sequences on the left side of each panel. (**A**) Multiple sequence alignment of putative human IgHA1 (J00220) and IgHA2 (J00221) consensus domain sequences with rhesus consensus cDNA coding sequences recovered (IgHA*01, IgHA*02, and IgHA*03) in either secreted (s) or membrane-bound (b) form. The target sites of human 10× primers are denoted by red boxes. (**B** and **C**) Zoom in of the inner and outer primer target sites, where mismatches between the primer and rhesus consensus sequence(s) are denoted by red boxes.

by the 10× human outer and inner primers (Fig. 5A–C). These primer inconsistencies were not specific to IgHA as all other human primers with the exception of the IgHG outer primer had mismatches with at least one of its corresponding rhesus reference sequences (data not shown). These results clearly indicated that rhesus-specific primers are necessary to perform comparable single cell analysis for rhesus macaques.

To properly design rhesus-specific Ig and TCR C region primers that are compatible with respective 10× human V(D)J assay designs, we first selected existing primers that were free of mismatches, which included the 10× human IgHG outer primer and rhesus primers designed by Sundling et al. (28) (Supplemental Table II). Notably, some primers that successfully in silico PCR amplified Ig/TCR transcripts contained one or more mismatches, thus requiring updated designs (*Materials and Methods*). For isotypes that lacked an inner and/or outer primer, we designed new primers with melting temperatures compatible with that of the 10× forward primer (Supplemental Table II). Similar to the Ig primers, 10× human TCR primers were also largely incompatible with the rhesus consensus sequences. We found that only the inner primer for TCRB had a perfect match and thus designed all other primers for rhesus macaques (Supplemental Table II). We also included primers that target TCRD and TCRG in the final set of our rhesus-specific T cell V(D)J primers, which are not included in the human 10× assays (68).

To ensure that these newly designed primers can amplify rhesus transcripts experimentally, for each of them, we designed a corresponding forward primer that was upstream but still within the C region (Supplemental Table II). Then, we tested all these custom inner and outer primers by pairing with their forward primers via PCR (*Materials and Methods*). We obtained the desired product size for each assay (Supplemental Fig. 3), confirming that these rhesus-specific primers newly designed for single cell–based V(D)J assays targeted rhesus Ig/TCR transcripts as expected. In some assays, we also observed additional bands, but the desired bands tended to be visibly the most prominent (seven of eight outer primers and two of five inner primers tested). Many of those additional bands were larger than the desired product size (Supplemental Fig. 3), suggesting some forward primers might have shared certain

sequence similarities with the highly diverse upstream variable regions. The actual single cell assays use nested template switch PCR; therefore, these forward primers are not needed, and such off-target amplification would not be a concern. Together, these results indicate that the primers we designed will be useful for single cell assays and compatible with the 10× Genomics platform.

Discussion

In this study, we profiled Ig and TCR repertoires using long read transcript sequencing of tissues collected from Indian-origin rhesus macaques, one of the most widely used NHP models. Because we obtained high-quality, full-length sequences for individual Ig and TCR transcripts, no sequence assembly was necessary. This unbiased profiling analysis allowed us to compile, to our knowledge, the first complete reference set for the constant regions of all expected rhesus Ig and TCR isotypes and chain types based on homology with humans, including three newly identified reference sequences. Whereas most of publicly available rhesus Ig and TCR C region sequences were computationally predicted from genomic sequences and/or partial sequences, our reference sequences were from directly sequenced transcripts and full length. We were able to define the coding regions for these constant regions, yielding full coding sequences as well as cDNA sequences containing 3′ UTRs.

An immediate benefit of having this complete C region reference is improved accuracy of rhesus Ig/TCR transcripts identification. For example, we found that ~39% of rhesus TCR transcripts that we initially identified by the commonly used combination of IgBLAST and IMGT databases were actually Ig transcripts, based on the alignment of their C region portion against these reference sequences. This misassignment could be because of several reasons. First, the current IMGT database has a limited number of TCR V region germline sequences for rhesus macaques. Second, the alignment of rhesus Ig/TCR transcripts to the human V region reference is not sufficiently reliable. The incomplete 5′ end of some rhesus Ig/TCR transcript sequences was also a contributing factor. In the future, we expect that our reference sequences will improve identification of novel germline variable genes extracted from recombined Ig and TCR sequences in rhesus macaques as the number of false positives will be substantially abrogated.

These full-length Ig/TCR transcript sequences offered insights into the rhesus immune system. For the Ig constant regions, we identified both cell membrane–bound and secreted forms that we differentiated by distinct 3′ end splicing. This demonstrates our ability to detect the alternative splicing events among these multiexonic regions. As the genomic assembly and annotation of these C region loci improves, long read sequencing may enable the discovery of similar alternative exon usage as described for human IgH C region genes (73). In addition, we recovered three different allotypes for IgHA with significant variation in the hinge region. When this hinge region is longer, the avidity for Ag interactions is increased at the expense of increased susceptibility to pathogen-derived proteases known to target the hinge region (74). We suspect that our ability to identify hinge region variation was driven by the high heterogeneity of the α C region (71) as well as its highest frequency among the Ig isotypes we recovered. Given the interspecies variability of IgHG between macaques and humans (75) and the heterogeneity of rhesus IgHC genes in general (16), it is imperative that future rhesus Ig gene analyses also consider this allotypic diversity to improve the translatability of NHP models.

Because we did not use any targeted experimental approaches to obtain these full-length Ig/TCR transcripts, we were able to examine the diversities of rhesus Ig/TCR variable regions in an unbiased manner. We found that large diversities existed across the entire variable regions of rhesus Ig and TCR transcripts. This is especially relevant because available assays for profiling rhesus Ig/TCR repertoires have relied on the consensus primers derived from a limited number of available rhesus V region sequences. Previous benchmarking analyses for targeted repertoire sequencing in humans used either spiked-in samples (43) or samples of unknown composition (42), motivating us to leverage our collected sequences for a truly unbiased assessment of conventional rhesus-specific MPCR designs for Ig (28, 29, 32) and TCR (66, 67). MPCR targets specific V gene segments (forward primer) and either specific J gene segments or constant regions (reverse primer) to generate amplicon libraries. Not surprisingly, we found that all of these assays failed to cover a significant percentage of these rhesus Ig/TCR transcripts. We do not contend that this method be avoided altogether moving forward, but additional improvements will be necessary. For example, databases of V region gene segments can be improved using the IgDiscover tool (18). Targeting the more conserved 5′ UTR of V genes may reduce the amount of primers needed in MPCR designs and yield amplicons covering the entire V region but still need to be designed for rhesus macaque (34). New algorithms such as the recently described version of the tool for Ig Genotype Elucidation via repertoire sequencing (TIgGER) for identifying human Ig (76) can also improve germline allele assignment if adapted for rhesus macaque.

Alternatively, scRNA-seq and RACE-seq only use primers that target the constant regions of these Ig/TCR transcripts. In particular, scRNA-seq combines repertoire analysis and transcriptome profiling in single B and T cells (51–54), and it can identify chain pairs within single cells (77). Our analysis showed that human-specific C region primers from commercially available assays were not compatible for use in rhesus macaques. For these reasons, we elected to design the first single cell–based B and T cell V(D)J assays for rhesus macaques. We leveraged the complete set of rhesus C region reference sequences compiled in this study to ensure a broad coverage of rhesus Ig/TCR diversity and experimentally validated the newly designed rhesus-specific primers. Our assays are fully compatible with the standard protocols from 10× Genomics, currently a popular platform for single cell analysis. These primers can also be used or modified for RACE-seq applications, considering similar library preparation techniques.

More broadly, this study also demonstrated an accessible strategy for developing immune repertoire resources and assays for less studied organisms. Traditionally, the development of Ig/TCR resources would start with collecting large numbers of species-specific transcript and germline sequences. Because Ig/TCR loci are among the most complex regions in the genome because of their duplicated and polymorphic nature, proper sequencing and assembly of these genomic regions requires sophisticated efforts (78). Consequently, complete assemblies are only available for selected model organisms such as humans and mice; meanwhile, these loci still do not have complete assemblies for rhesus macaques despite recent efforts (16). The PacBio Iso-Seq approach we used in this study sequences full-length Ig/TCR transcripts directly and in a high-throughput manner, bypassing transcript assembly issues. The performance of this strategy could be further improved as well. For example, compared with the whole tissue sample used in this study, use of isolated B and T cells may significantly improve the recovery of these repertoire sequences. In this study, we did not infer germline alleles because of both the insufficient sequencing depth and the additional genotypic diversity introduced due to having each tissue sample pooled from multiple animals. However, we anticipate that future Iso-Seq studies tailored for immune repertoire analysis will be able to support the application of allele discovery tools, such as IgDiscover (18) and TIgGER (76). The custom computational solution we developed and describe in this study can be modified for other species by replacing C region references from a related species to identify putative species-specific Ig/TCR transcript sequences. However, additional information like Ig/TCR germline allele assignment, the genomic order of individual genes, and copy number variations will still require complete genomic sequencing and assembly.

In this study, we systematically profiled the diversity of rhesus Ig and TCR genes using full-length transcriptome sequencing. We performed benchmark analysis of common MPCR-based targeted sequencing strategies for rhesus macaque and demonstrated that available MPCR assays do not cover the rhesus repertoire diversity adequately. Our construction of a complete reference set for rhesus macaque Ig/TCR constant regions enabled the design, to our knowledge, of the first rhesus-specific single cell–based V(D)J assays, addressing many of the current technical limitations of rhesus macaque repertoire analysis. Furthermore, our approach circumvents the computational challenges concomitant with the genomic assembly of these complex immune genes by combining long read sequencing and new computational solutions. This study thus represents a new direction for developing nascent immune resources.

Disclosures

References

1. Williams, K., A. Lackner, and J. Mallard. 2016. Non-human primate models of SIV infection and CNS neuropathology. *Curr. Opin. Virol.* 19: 92–98.
2. Zhao, H., Y. H. Jiang, and Y. Q. Zhang. 2018. Modeling autism in non-human primates: opportunities and challenges. *Autism Res.* 11: 686–694.
3. Flynn, J. L., H. P. Gideon, J. T. Mattila, and P. L. Lin. 2015. Immunology studies in non-human primate models of tuberculosis. *Immunol. Rev.* 264: 60–73.
4. Reyes, L. F., M. I. Restrepo, C. A. Hinojosa, N. J. Soni, A. T. Shenoy, R. P. Gilley, N. Gonzalez-Juarbe, J. R. Noda, V. T. Winter, M. A. de la Garza, et al. 2016. A non-human primate model of severe pneumococcal pneumonia. *PLoS One* 11: e0166092.
5. Estes, J. D., S. W. Wong, and J. M. Brenchley. 2018. Nonhuman primate models of human viral infections. *Nat. Rev. Immunol.* 18: 390–404.
6. Hansen, S. G., D. E. Zak, G. Xu, J. C. Ford, E. E. Marshall, D. Malouli, R. M. Gilbride, C. M. Hughes, A. B. Ventura, E. Ainslie, et al. 2018. Prevention

of tuberculosis in rhesus macaques by a cytomegalovirus-based vaccine. *Nat. Med.* 24: 130–143.

7. Carpenter, S. M., and S. M. Behar. 2018. A new vaccine for tuberculosis in rhesus macaques. *Nat. Med.* 24: 124–126.

8. Hansen, S. G., J. C. Ford, M. S. Lewis, A. B. Ventura, C. M. Hughes, L. Coyne-Johnson, N. Whizin, K. Oswald, R. Shoemaker, T. Swanson, et al. 2011. Profound early control of highly pathogenic SIV by an effector memory T-cell vaccine. *Nature* 473: 523–527.

9. Hansen, S. G., M. Piatak, Jr., A. B. Ventura, C. M. Hughes, R. M. Gilbride, J. C. Ford, K. Oswald, R. Shoemaker, Y. Li, M. S. Lewis, et al. 2013. Immune clearance of highly pathogenic SIV infection. [Published erratum appears in 2014 Nature 514: 654.; Published erratum appears in 2017 *Nature* 547: 123–124.] *Nature* 502: 100–104.

10. Carroll, T. D., S. R. Matzinger, P. A. Barry, M. B. McChesney, J. Fairman, and C. J. Miller. 2014. Efficacy of influenza vaccination of elderly rhesus macaques is dramatically improved by addition of a cationic lipid/DNA adjuvant. *J. Infect. Dis.* 209: 24–33.

11. Adams Waldorf, K. M., J. E. Stencel-Baerenwald, R. P. Kapur, C. Studholme, E. Boldenow, J. Vornhagen, A. Baldessari, M. K. Dighe, J. Thiel, S. Merillat, et al. 2016. Fetal brain lesions after subcutaneous inoculation of Zika virus in a pregnant nonhuman primate. *Nat. Med.* 22: 1256–1259.

12. Clement, K. H., T. L. Rudge, Jr., H. J. Mayfield, L. A. Carlton, A. Hester, N. A. Niemuth, C. L. Sabourin, A. M. Brys, and C. P. Quinn. 2010. Vaccination of rhesus macaques with the anthrax vaccine adsorbed vaccine produces a serum antibody response that effectively neutralizes receptor-bound protective antigen in vitro. *Clin. Vaccine Immunol.* 17: 1753–1762.

13. Awasthi, S., L. M. Hook, C. E. Shaw, B. Pahar, J. A. Stagray, D. Liu, R. S. Veazey, and H. M. Friedman. 2017. An HSV-2 trivalent vaccine is immunogenic in rhesus macaques and highly efficacious in guinea pigs. *PLoS Pathog.* 13: e1006141.

14. Zimin, A. V., A. S. Cornish, M. D. Maudhoo, R. M. Gibbs, X. Zhang, S. Pandey, D. T. Meehan, K. Wipfler, S. E. Bosinger, Z. P. Johnson, et al. 2014. A new rhesus macaque assembly and annotation for next-generation sequencing analyses. *Biol. Direct* 9: 20.

15. Harding, J. D. 2017. Nonhuman primates and translational research: progress, opportunities, and challenges. *ILAR J.* 58: 141–150.

16. Ramesh, A., S. Darko, A. Hua, G. Overman, A. Ransier, J. R. Francica, A. Trama, G. D. Tomaras, B. F. Haynes, D. C. Douek, and T. B. Kepler. 2017. Structure and diversity of the rhesus macaque immunoglobulin loci through multiple *de novo* genome assemblies. *Front. Immunol.* 8: 1407.

17. Cirelli, K. M., D. G. Carnathan, B. Nogal, J. T. Martin, O. L. Rodriguez, A. A. Upadhyay, C. A. Enemuo, E. H. Gebru, Y. Choe, F. Viviano, et al. 2019. Slow delivery immunization enhances HIV neutralizing antibody and germinal center responses via modulation of immunodominance. [Published erratum appears in 2020 Cell 180: 206.] *Cell* 177: 1153–1171.e28.

18. Corcoran, M. M., G. E. Phad, N. Vázquez Bernat, C. Stahl-Hennig, N. Sumida, M. A. A. Persson, M. Martin, and G. B. Karlsson Hedestam. 2016. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat. Commun.* 7: 13642.

19. Nishana, M., and S. C. Raghavan. 2012. Role of recombination activating genes in the generation of antigen receptor diversity and beyond. *Immunology* 137: 271–281.

20. Little, A. J., A. Matthews, M. Oettinger, D. B. Roth, and D. G. Schatz. 2015. The mechanism of V(D)J recombination. In *Molecular Biology of B Cells*, 2nd Ed., F. W. Alt, T. Honjo, A. Radbruch, and M. Reth, eds. Academic Press, London, p. 13–34.

21. Hesslein, D. G. T., and D. G. Schatz. 2001. Factors and forces controlling V(D)J recombination. In *Advances in Immunology*. F. J. Dixon, ed. Academic Press, San Diego, CA, p. 169–232.

22. Murphy, K., P. Travers, and M. Walport. 2017. The generation of lymphocyte antigen receptors. In *Janeway's Immunobiology*, Chapter 5, 7th Ed. Garland Science, New York, p. 188–189.

23. Manis, J. P., M. Tian, and F. W. Alt. 2002. Mechanism and control of class-switch recombination. *Trends Immunol.* 23: 31–39.

24. Lucas, J. S., C. Murre, A. J. Feeney, and R. Riblet. 2015. The structure and regulation of the immunoglobulin loci. In *Molecular Biology of B Cells*, 2nd Ed., F. W. Alt, T. Honjo, A. Radbruch, and M. Reth, eds. Academic Press, London, p. 1–11.

25. Markoulatos, P., N. Siafakas, and M. Moncany. 2002. Multiplex polymerase chain reaction: a practical approach. *J. Clin. Lab. Anal.* 16: 47–51.

26. Hodges, E., Z. Xuan, V. Balija, M. Kramer, M. N. Molla, S. W. Smith, C. M. Middle, M. J. Rodesch, T. J. Albert, G. J. Hannon, and W. R. McCombie. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* 39: 1522–1527.

27. Yeku, O., and M. A. Frohman. 2011. Rapid amplification of cDNA ends (RACE). In *RNA: Methods and Protocols*. H. Nielsen, ed. Humana Press, Totowa, NJ, p. 107–122.

28. Sundling, C., G. Phad, I. Douagi, M. Navis, and G. B. Karlsson Hedestam. 2012. Isolation of antibody V(D)J sequences from single cell sorted rhesus macaque B cells. *J. Immunol. Methods* 386: 85–93.

29. Wiehe, K., D. Easterhoff, K. Luo, N. I. Nicely, T. Bradley, F. H. Jaeger, S. M. Dennison, R. Zhang, K. E. Lloyd, C. Stolarchuk, et al. 2014. Antibody light-chain-restricted recognition of the site of immune pressure in the RV144 HIV-1 vaccine trial is phylogenetically conserved. *Immunity* 41: 909–918.

30. Fu, L., X. Li, W. Zhang, C. Wang, J. Wu, H. Yang, J. Wang, and X. Liu. 2017. A comprehensive profiling of T- and B-lymphocyte receptor repertoires from a Chinese-origin rhesus macaque by high-throughput sequencing. *PLoS One* 12: e0182733.

31. Sundling, C., Z. Zhang, G. E. Phad, Z. Sheng, Y. Wang, J. R. Mascola, Y. Li, R. T. Wyatt, L. Shapiro, and G. B. Karlsson Hedestam. 2014. Single-cell and deep sequencing of IgG-switched macaque B cells reveal a diverse Ig repertoire following immunization. *J. Immunol.* 192: 3637–3644.

32. Rosenfeld, R., A. Zvi, E. Winter, R. Hope, O. Israeli, O. Mazor, and G. Yaari. 2019. A primer set for comprehensive amplification of V-genes from rhesus macaque origin based on repertoire sequencing. *J. Immunol. Methods* 465: 67–71.

33. Zhang, W., X. Li, L. Wang, J. Deng, L. Lin, L. Tian, J. Wu, C. Tang, H. Yang, J. Wang, et al. 2019. Identification of variable and joining germline genes and alleles for rhesus macaque from B cell receptor repertoires. *J. Immunol.* 202: 1612–1622.

34. Vázquez Bernat, N., M. Corcoran, U. Hardt, M. Kaduk, G. E. Phad, M. Martin, and G. B. Karlsson Hedestam. 2019. High-quality library preparation for NGS-based immunoglobulin germline gene inference and repertoire expression analysis. *Front. Immunol.* 10: 660.

35. Benichou, J., R. Ben-Hamo, Y. Louzoun, and S. Efroni. 2012. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* 135: 183–191.

36. Mamedov, I. Z., O. V. Britanova, I. V. Zvyagin, M. A. Turchaninova, D. A. Bolotin, E. V. Putintseva, Y. B. Lebedev, and D. M. Chudakov. 2013. Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. *Front. Immunol.* 4: 456.

37. He, L., D. Sok, P. Azadnia, J. Hsueh, E. Landais, M. Simek, W. C. Koff, P. Poignard, D. R. Burton, and J. Zhu. 2014. Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. *Sci. Rep.* 4: 6778.

38. Fu, G. K., W. Xu, J. Wilhelmy, M. N. Mindrinos, R. W. Davis, W. Xiao, and S. P. Fodor. 2014. Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. *Proc. Natl. Acad. Sci. USA* 111: 1891–1896.

39. Turchaninova, M. A., A. Davydov, O. V. Britanova, M. Shugay, V. Bikos, E. S. Egorov, V. I. Kirgizova, E. M. Merzlyak, D. B. Staroverov, D. A. Bolotin, et al. 2016. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat. Protoc.* 11: 1599–1616.

40. Rock, E. P., P. R. Sibbald, M. M. Davis, and Y. H. Chien. 1994. CDR3 length in antigen-specific immune receptors. *J. Exp. Med.* 179: 323–328.

41. Safonova, Y., and P. A. Pevzner. 2019. *De novo* inference of diversity genes and analysis of non-canonical V(DD)J recombination in immunoglobulins. *Front. Immunol.* 10: 987.

42. Bashford-Rogers, R. J., A. L. Palser, S. F. Idris, L. Carter, M. Epstein, R. E. Callard, D. C. Douek, G. S. Vassiliou, G. A. Follows, M. Hubank, and P. Kellam. 2014. Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. *BMC Immunol.* 15: 29.

43. Liu, X., W. Zhang, X. Zeng, R. Zhang, Y. Du, X. Hong, H. Cao, Z. Su, C. Wang, J. Wu, et al. 2016. Systematic comparative evaluation of methods for investigating the TCRβ repertoire. *PLoS One* 11: e0152464.

44. Macosko, E. Z., A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161: 1202–1214.

45. Zheng, G. X. Y., J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, et al. 2017. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8: 14049.

46. Chattopadhyay, P. K., T. M. Gierahn, M. Roederer, and J. C. Love. 2014. Single-cell technologies for monitoring immune systems. *Nat. Immunol.* 15: 128–135.

47. Proserpio, V., and B. Mahata. 2016. Single-cell technologies to study the immune system. *Immunology* 147: 133–140.

48. Papalexi, E., and R. Satija. 2018. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* 18: 35–45.

49. Stubbington, M. J. T., O. Rozenblatt-Rosen, A. Regev, and S. A. Teichmann. 2017. Single-cell transcriptomics to explore the immune system in health and disease. *Science* 358: 58–63.

50. Chen, H., F. Ye, and G. Guo. 2019. Revolutionizing immunology with single-cell RNA sequencing. *Cell. Mol. Immunol.* 16: 242–249.

51. Neal, J. T., X. Li, J. Zhu, V. Giangarra, C. L. Grzeskowiak, J. Ju, I. H. Liu, S.-H. Chiou, A. A. Salahudeen, A. R. Smith, et al. 2018. Organoid modeling of the tumor immune microenvironment. *Cell* 175: 1972–1988.e16.

52. Glanville, J., H. Huang, A. Nau, O. Hatton, L. E. Wagar, F. Rubelt, X. Ji, A. Han, S. M. Krams, C. Pettus, et al. 2017. Identifying specificity groups in the T cell receptor repertoire. *Nature* 547: 94–98.

53. Dash, P., A. J. Fiore-Gartland, T. Hertz, G. C. Wang, S. Sharma, A. Souquette, J. C. Crawford, E. B. Clemens, T. H. O. Nguyen, K. Kedzierska, et al. 2017. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 547: 89–93.

54. Ichinohe, T., T. Miyama, T. Kawase, Y. Honjo, K. Kitaura, H. Sato, T. Shin-I, and R. Suzuki. 2018. Next-generation immune repertoire sequencing as a clue to elucidate the landscape of immune modulation by host-gut microbiome interactions. *Front. Immunol.* 9: 668.

55. Ye, J., N. Ma, T. L. Madden, and J. M. Ostell. 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* 41(W1): W34–W40.

56. Brochet, X., M. P. Lefranc, and V. Giudicelli. 2008. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* 36(Web Server): W503–W508.

57. Giudicelli, V., X. Brochet, and M. P. Lefranc. 2011. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb. Protoc.* 2011: 695–715.

58. Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28: 3150–3152.

59. Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16: 276–277.

60. Lefranc, M.-P., V. Giudicelli, C. Ginestoux, J. Bodmer, W. Müller, R. Bontrop, M. Lemaitre, A. Malik, V. Barbié, and D. Chaume. 1999. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.* 27: 209–212.

61. Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461.

62. Chojnacki, S., A. Cowley, J. Lee, A. Foix, and R. Lopez. 2017. Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. *Nucleic Acids Res.* 45(W1): W550–W553.

63. Okonechnikov, K., O. Golosova, and M. Fursov, UGENE team. 2012. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28: 1166–1167.

64. R Development Core Team. 2010. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

65. Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag, New York.

66. Chen, Z. W., Z. C. Kou, L. Shen, K. A. Reimann, and N. L. Letvin. 1993. Conserved T-cell receptor repertoire in simian immunodeficiency virus-infected rhesus monkeys. *J. Immunol.* 151: 2177–2187.

67. Li, Z., G. Liu, Y. Tong, M. Zhang, Y. Xu, L. Qin, Z. Wang, X. Chen, and J. He. 2015. Comprehensive analysis of the T-cell receptor beta chain gene in rhesus monkey by high throughput sequencing. *Sci. Rep.* 5: 10092.

68. Kuhn, R. M., D. Haussler, and W. J. Kent. 2013. The UCSC genome browser and associated tools. *Brief. Bioinform.* 14: 144–161.

69. Ye, J., G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen, and T. L. Madden. 2012. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13: 134.

70. Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* 323: 133–138.

71. Scinicariello, F., and R. Attanasio. 2001. Intraspecies heterogeneity of immunoglobulin alpha-chain constant region genes in rhesus macaques. *Immunology* 103: 441–448.

72. Rogers, K. A., L. Jayashankar, F. Scinicariello, and R. Attanasio. 2008. Nonhuman primate IgA: genetic heterogeneity and interactions with CD89. *J. Immunol.* 180: 4816–4824.

73. Vollmers, C., L. Penland, J. N. Kanbar, and S. R. Quake. 2015. Novel exons and splice variants in the human antibody heavy chain identified by single cell and single molecule sequencing. *PLoS One* 10: e0117050.

74. Woof, J. M., and M. A. Kerr. 2004. IgA function--variations on a theme. *Immunology* 113: 175–177.

75. Crowley, A. R., and M. E. Ackerman. 2019. Mind the gap: how interspecies variability in IgG and its receptors may complicate comparisons of human and non-human primate effector function. *Front. Immunol.* 10: 697.

76. Gadala-Maria, D., M. Gidoni, S. Marquez, J. A. Vander Heiden, J. T. Kos, C. T. Watson, K. C. O'Connor, G. Yaari, and S. H. Kleinstein. 2019. Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. *Front. Immunol.* 10: 129.

77. Carter, J. A., J. B. Preall, K. Grigaityte, S. J. Goldfless, E. Jeffery, A. W. Briggs, F. Vigneault, and G. S. Atwal. 2019. Single T cell sequencing demonstrates the functional role of $\alpha\beta$ TCR pairing in cell lineage and antigen specificity. *Front. Immunol.* 10: 1516.

78. Watson, C. T., K. M. Steinberg, J. Huddleston, R. L. Warren, M. Malig, J. Schein, A. J. Willsey, J. B. Joy, J. K. Scott, T. A. Graves, et al. 2013. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am. J. Hum. Genet.* 92: 530–546.