Published in final edited form as:
J Immunol. 2020 June 15; 204(12): 3434–3444. doi:10.4049/jimmunol.1901256.

Systematic profiling of full-length immunoglobulin and T-cell receptor 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.**[§,¶,**], **Xinxia Peng**[*,†,††]

[*]Department of Molecular Biomedical Sciences, North Carolina State University College of Veterinary Medicine, Raleigh, NC 27607, USA

[†]Bioinformatics Graduate Program, North Carolina State University, Raleigh, NC 27695, USA

[‡]Pacific Biosciences, Menlo Park, CA 94025, USA

[§]Department of Immunology, University of Washington, Seattle, WA, USA

[¶]Center for Innate Immunity and Immune Diseases, University of Washington, WA, USA

[‖]Genetics Graduate Program, North Carolina State University, Raleigh, NC 27695

[#]Vaccine and Gene Therapy Institute, Oregon Health & Science University, Beaverton, OR 97006, USA

[**]Washington National Primate Research Center, University of Washington, Seattle, WA, USA

[††]Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695, USA

Abstract

The diversity of immunoglobulin (Ig) and T-cell receptor (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, due to incomplete reference resources, the coverage and accuracy of the traditional targeted amplification strategies for profiling rhesus Ig and TCR repertoires are largely unknown. Here, using long read sequencing, we sequenced four Indian-origin rhesus macaque tissues and obtained high quality, full-length sequences for over 6,000 unique Ig and TCR transcripts, without the need for sequence assembly. We constructed 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% to 53% and 42% to 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

Address correspondence and reprint requests to Dr. Xinxia Peng, NC State University, CVM Research Bldg Rm 476B, 1060 William Moore Dr, Raleigh, NC 27607; xpeng5@ncsu.edu; phone: (919) 515-4481; fax: (919)-513-6464.

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.

Introduction

Rhesus macaque (*Macaca mulatta*) is one of the most commonly used and best-studied non-human primate (NHP) animal models. NHPs are key to studying human biology and human diseases due to 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, immunoglobulin (Ig) and T-cell receptor (TCR) repertoires house an enormous amount of diversity responsible for recognizing a near limitless array of antigens presented through environmental exposures. Ig and TCR have two domains: a constant region and a variable 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 macaque 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 or IMGT) for these diverse sequences also remain fairly meager relative to their human counterparts, although there are new tools developed to address these gaps (18). Since 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 the 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 (19–21)). This process accounts for the majority of the diversity within the variable region domain, with the number of unique Ig and TCR variable 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 polymerase chain reaction (MPCR) (25), RNA capture (26), or 5’ rapid amplification of cDNA ends (5’ RACE) (27). For rhesus macaque, many Ig repertoire sequencing efforts use a MPCR approach (28, 29), while 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 lab also developed a strategy for targeting the 5’ UTR of V genes, often conserved among these gene clusters, thus reducing the number of primers needed to target the variable region in human (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 constant 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 (UMIs) for repertoire sequencing (36, 37), mitigating issues arising from PCR bias (38). However, such protocols are optimized for applications in human and mouse (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 sequencing instruments like the 2 × 250 bp HiSeq system and the 2 × 300 bp MiSeq system, in part due to the large length variability of recombined molecules (40) and the potential for non-canonical recombination events (41). Library preparation techniques have recently been improved to support this capability but only in human (34). Despite these technical advances, it is still unclear how well these strategies perform in rhesus macaque, where 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 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 UMIs. Interrogation of Ig/TCR repertoire and B/T cell transcriptomes in the same single cells has provided novel insights in human, 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 macaque, 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 macaque. 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 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 PacBio SMRT sequencing, we used previously isolated Indian-origin rhesus macaque RNA from four tissue types: lymph node (LN), peripheral blood mononuclear cells (PBMC), 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, while 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 lymph nodes. 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, >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-hour movie times at the University of Washington PacBio Sequencing Services site.

Data processing

Raw PacBio sequencing data was first run through the CCS protocol in SMRT Analysis 2.3.0 to generate Circular Consensus Sequence (CCS) reads. Each CCS read is the circular consensus sequence of a single molecule. CCS reads were classified as non-full-length and full-length, where the latter has all of the following detected: polyA tail, 5’ cDNA primer, and 3’ cDNA primer. Full-length CCS sequences were aligned to immunoglobulin (Ig) and T cell receptor (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 constant 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 constant region sequences as below. Using the J region reported by IgBLAST, we extracted the constant region sequence for each full-length sequence. Next, we clustered extracted constant 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 ten 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), where 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 constant 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 Iso-Seq dataset as input. CCS reads were separately globally aligned to the custom databases of consensus Ig and TCR constant 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 (>10bp) 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 (Supplementary Fig. 1B). Further, the final set of TCR transcripts only required CCS reads to be confirmed by the custom database (Supplementary Fig. 1A), as it was unclear if IMGT functional annotation for human 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 CIGAR 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 variable 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. Chi-square 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 variable region sequences was determined by first removing the constant region sequence determined by its alignment to the corresponding consensus constant 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 TCRA and TCRB sequences (66, 67) were tested via *in silico* PCR analysis using the UCSC 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 10x B cell and T cell V(D)J assays were also assessed via *in silico* PCR analysis (68) where 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 variable region (forward primer) and constant 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 nucleotides on the 3' end of the primer.

Design of rhesus-specific primers and PCR validation

Ig and TCR constant region inner and outer reverse primers were designed using NCBI primerBLAST (69) with target Tms consistent with current 10x V(D)J assay forward primers. For those targeting multiple consensus sequences, the search space was constrained to consensus regions. To validate these 10x optimized reverse primers, primerBLAST (69) was used to design standard PCR assays each with a forward primer compatible with its respective 10x optimized reverse primer and target reference sequence. This was only done for custom primers in the V(D)J assays, while primers adopted from Sundling et al (28) were assumed to be efficacious. 100 nmole HPLC purified DNA oligonucleotides were ordered from IDT. Reverse transcription was performed using Qiagen QuantiTect Reverse Transcription Kit (Cat. No. 205313, Lot 157037104) following the manufacturer's protocol with 1 ug rhesus macaque lymph node tissue RNA. The resulting cDNA was then amplified using Takara Titanium TaqPCR kit (Cat. 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 minute. 400 ng of each PCR reaction was then run on a 10% TBE polyacrylamide gel alongside Thermo 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 Circular Consensus Sequence (CCS) reads from four different rhesus macaque tissues (Supplementary Table 1). Each CCS read is the circular consensus sequence 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 polyA tail (Supplementary Table 1). 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 2,534 TCR putative transcript sequences, respectively accounting for 1.4% and 0.28% of the total number of full-length CCS reads (Supplementary Table 1). Only functional Ig transcripts (i.e. those in-frame and without premature stop codons according to IMGT germline annotation) were used in downstream analyses (5,666 / 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. Since 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 constant region consensus (i.e. including the 3’ UTR), while 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 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 gDNA 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% to 99.9%). To properly classify the TCRBC subclasses, we aligned the two TCRBC protein sequences to the two reference ORFs available in IMGT, yielding perfect matches (data not shown). We leveraged the available human ORFs in IMGT to determine subclasses of TCRGC, as reference sequences were not available for rhesus macaque. 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 2,534) 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 (Supplementary Fig. 1A). While 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% (5,415 / 5,666) of these full-length Ig transcript sequences. Those transcript sequences that failed to align tended to be other molecules from the immunoglobulin superfamily that contain V-set domains (i.e. a variable 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 5,384 full-length Ig transcript sequences (Table 1). Among four tissue samples sequenced here, the rectal biopsy and lymph node 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 rectal biopsy and lymph node 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 2), 50 of which (7%) we did not originally recover using the human IMGT reference database (Supplementary 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 rectal biopsy sample, while the whole blood, PBMC, and lymph node samples each had larger recoveries of all TCR isotypes (Table 2). TCRA was the most abundant in each tissue, representing

61% of all sequences recovered (455 / 741). TCRG2 had the second highest abundance with its strongest representation in PBMCs (Table 2).

We then assessed the sequence quality of these Ig and TCR transcripts, by quantifying insertion and deletion events within the alignments of their constant 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 constant 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 constant region database for identification instead of the traditional IgBLAST approach with human germline annotation (Supplementary Fig. 1A). We thus elected to align CCS reads directly to our custom Ig constant region database to make a similar comparison. We recovered 11,451 Ig transcripts using this strategy; of these, 5,384 (47%) were functional IgBLAST hits (see Table 1), 5,858 (51%) were non-functional IgBLAST hits, and 209 (2%) were not detected by IgBLAST (Supplementary Fig. 1B). B cells that contain non-functional receptor sequences are known to apoptose early in their development; therefore, the proportion of non-functional IgBLAST hits observed here 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 constant region to harbor a complete variable region (127 or 1% overall) (Supplementary Fig. 1B), indicating that some variable 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, while there was broad coverage of IGHD gene families. Among light 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, though IGLJ1 was in 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 due to the overall lower abundance of light chain Ig sequences we observed (Table 1). Interestingly, relative frequencies of these different gene families were highly consistent across the four tissues analyzed here (Supplementary Fig. 2), despite drastically different sampling depth

(Fig. 1). Only light chain gene families within PBMCs and whole blood deviated from this trend. Since we had significantly less detection of IGK and IGL sequences in these tissues (Table 1), their observed deviations require further investigation.

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 square = 77.6, df = 15, p = 1.9e-10). Furthermore, we discovered a strong positive correlation between the V-J recombination rates in the lymph node and rectal biopsy samples (Pearson correlation = 0.86). Since these two samples had highly different compositions of IGH isotypes (Table 1), 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 variable region shows 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 variable 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 (1,383 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–B), 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 in 2012 (28) had the highest recovery of IGH sequences (76%), while the most recent Ig primer set designed by Rosenfeld et al in 2019 (32) 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, while Chen et al (66) recovered 53% of TCRA sequences.

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 constant region. Indeed, when we tested the constant 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 constant region. Since there are no known single cell repertoire analysis assays for rhesus macaque, we first checked the efficiency of human B and T cell V(D)J assays developed by 10x 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, TCRD, TCRG1, and TCRG2 were not included in this analysis, as 10x 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 human and evaluated the compatibility of their respective 10x human primers where relevant. For example, we aligned human IGHA1 and IGHA2 to the five rhesus IGHA consensus sequences (3 secreted, 2 membrane-bound) and examined the regions targeted by 10x human primers (Fig. 5A). Overall, the two human cDNA coding sequences had high similarity to the rhesus consensus sequences, averaging 89.7% identity. However, multiple nucleotide differences existed in the regions of rhesus reference sequences targeted by the 10x 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 rhesus-specific primers are necessary to perform comparable single cell analysis for rhesus macaques.

To properly design rhesus-specific Ig and TCR constant region primers that are compatible with respective 10x human V(D)J assay designs, we first selected existing primers that were free of mismatches, which included the 10x human IGHG outer primer and rhesus primers designed by Sundling et al (28) (Supplementary Table 2). 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 (Tms) compatible with that of the 10x forward primer (Supplementary Table 2). Similar to the Ig primers, 10x 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 macaque (Supplementary Table 2). 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 10x assays (68).

To ensure 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 constant region (Supplementary Table 2). 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 (Supplementary 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 (7 of 8 outer primers and 2 of 5 inner primers tested). Many of those additional bands were larger than the desired product size (Supplementary 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 10x 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. Since 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 the first complete reference set for the constant regions of all expected rhesus Ig and TCR isotypes and chain types based on homology with human, including three newly identified reference sequences. While most of publicly available rhesus Ig and TCR constant 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 constant region reference is improved accuracy of rhesus Ig/TCR transcripts identification. For example, we found that about 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 constant region portion against these reference sequences. This miss-assignment could be due to several reasons. First, the current IMGT database has a limited number of TCR variable region germline sequences for rhesus macaque. Second, the alignment of rhesus

Ig/TCR transcripts to the human variable 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 macaque, 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 multi-exonic regions. As the genomic assembly and annotation of these constant region loci improves, long read sequencing may enable discovery of similar alternative exon usage as described for human Ig heavy chain constant 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 antigen 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 alpha constant 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.

Since 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, since available assays for profiling rhesus Ig/TCR repertoires have relied on the consensus primers derived from a limited number of available rhesus variable region sequences. Previous benchmarking analyses for targeted repertoire sequencing in human 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 variable 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 variable 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 Rep-Seq (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 constant region primers from commercially available assays were not compatible for use in rhesus macaque. For these reasons, we elected to design the first single cell based B- and T-cell V(D)J assays for rhesus macaque. We leveraged the complete set of rhesus constant region reference sequences compiled here 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 10x 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. Since Ig/TCR loci are among the most complex regions in the genome due to 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 human and mouse; meanwhile, these loci still do not have complete assemblies for rhesus macaque despite recent efforts (16). The PacBio Iso-Seq approach we used here 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 to the whole tissue sample used here, 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 due to both the insufficient sequencing depth and the additional genotypic diversity introduced due to having each tissue sample pooled from multiple animals. However, we anticipate future Iso-Seq studies tailored for immune repertoire analysis will be able to support application of allele discovery tools, such as IgDiscover (18) and TIgGER (76). The custom computational solution we developed and describe here can be modified for other species by replacing constant 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 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This project has been funded in part with funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, R21AI120713 (XP). This project has been funded in part with funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272201300010C and HHSN272201800008C (MG) and the National Institutes of Health, Office of the Director P51OD010425 (MG).

Raw PacBio movie files used to generate the results in this study are available under NCBI BioProject (https://www.ncbi.nlm.nih.gov/bioproject/) accession number PRJNA389440. The full-length CCS reads used directly in this analysis are available on Zenodo (doi: 10.5281/zenodo.3634898). The curated constant region sequences generated in this study have been submitted to GenBank (https://www.ncbi.nlm.nih.gov/genbank/) under accession numbers MT043311-MT043331.

Abbreviations

CCS	circular consensus sequence
NHP	non-human primate
Ig	immunoglobulin
IGH	Ig heavy chain
IGHA	IGH alpha
IGHD	IGH delta
IGHE	IGH epsilon
IGHG	IGH gamma
IGHM	IGH mu
IGK	Ig kappa chain
IGL	Ig lambda chain
TCR	T cell receptor
TCRA	TCR alpha
TCRB	TCR beta
TCRD	TCR delta
TCRG	TCR gamma
IMGT	the international ImMunoGeneTics information system
MPCR	multiplex PCR
RACE	rapid amplification of cDNA ends

References

1. Williams K, Lackner A, and Mallard J. 2016 Non-human primate models of SIV infection and CNS neuropathology. Current Opinion in Virology 19: 92–98. [PubMed: 27544476]
2. Zhao H, Jiang YH, and Zhang YQ. 2018 Modeling autism in non-human primates: Opportunities and challenges. Autism research : official journal of the International Society for Autism Research 11: 686–694. [PubMed: 29573234]
3. Flynn JL, Gideon HP, Mattila JT, and Lin PL. 2015 Immunology studies in non-human primate models of tuberculosis. Immunological reviews 264: 60–73. [PubMed: 25703552]
4. Reyes LF, Restrepo MI, Hinojosa CA, Soni NJ, Shenoy AT, Gilley RP, Gonzalez-Juarbe N, Noda JR, Winter VT, de la Garza MA, Shade RE, Coalson JJ, Giavedoni LD, Anzueto A, and Orihuela CJ. 2016 A Non-Human Primate Model of Severe Pneumococcal Pneumonia. PLOS ONE 11: e0166092. [PubMed: 27855182]
5. Estes JD, Wong SW, and Brenchley JM. 2018 Nonhuman primate models of human viral infections. Nat Rev Immunol 18: 390–404. [PubMed: 29556017]
6. Hansen SG, Zak DE, Xu G, Ford JC, Marshall EE, Malouli D, Gilbride RM, Hughes CM, Ventura AB, Ainslie E, Randall KT, Selseth AN, Rundstrom P, Herlache L, Lewis MS, Park H, Planer SL, Turner JM, Fischer M, Armstrong C, Zweig RC, Valvo J, Braun JM, Shankar S, Lu L, Sylwester AW, Legasse AW, Messerle M, Jarvis MA, Amon LM, Aderem A, Alter G, Laddy DJ, Stone M, Bonavia A, Evans TG, Axthelm MK, Früh K, Edlefsen PT, and Picker LJ. 2018 Prevention of tuberculosis in rhesus macaques by a cytomegalovirus-based vaccine. Nature Medicine 24: 130–143.
7. Carpenter SM, and Behar SM. 2018 A new vaccine for tuberculosis in rhesus macaques. Nature Medicine 24: 124–126.
8. Hansen SG, Ford JC, Lewis MS, Ventura AB, Hughes CM, Coyne-Johnson L, Whizin N, Oswald K, Shoemaker R, Swanson T, Legasse AW, Chiuchiolo MJ, Parks CL, Axthelm MK, Nelson JA, Jarvis MA, Piatak M, Lifson JD, and Picker LJ. 2011 Profound early control of highly pathogenic SIV by an effector memory T-cell vaccine. Nature 473: 523–527. [PubMed: 21562493]
9. Hansen SG, Jr MP, Ventura AB, Hughes CM, Gilbride RM, Ford JC, Oswald K, Shoemaker R, Li Y, Lewis MS, Gilliam AN, Xu G, Whizin N, Burwitz BJ, Planer SL, Turner JM, Legasse AW, Axthelm MK, Nelson JA, Früh K, Sacha JB, Estes JD, Keele BF, Edlefsen PT, Lifson JD, and Picker LJ. 2013 Immune clearance of highly pathogenic SIV infection. Nature 502: 100–104. [PubMed: 24025770]
10. Carroll TD, Matzinger SR, Barry PA, McChesney MB, Fairman J, and Miller CJ. 2013 Efficacy of Influenza Vaccination of Elderly Rhesus Macaques Is Dramatically Improved by Addition of a Cationic Lipid/DNA Adjuvant. The Journal of Infectious Diseases 209: 24–33. [PubMed: 24141979]
11. Adams Waldorf KM, Stencel-Baerenwald JE, Kapur RP, Studholme C, Boldenow E, Vornhagen J, Baldessari A, Dighe MK, Thiel J, Merillat S, Armistead B, Tisoncik-Go J, Green RR, Davis MA, Dewey EC, Fairgrieve MR, Gatenby JC, Richards T, Garden GA, Diamond MS, Juul SE, Grant RF, Kuller L, Shaw DW, Ogle J, Gough GM, Lee W, English C, Hevner RF, Dobyns WB, Gale M Jr., and Rajagopal L. 2016 Fetal brain lesions after subcutaneous inoculation of Zika virus in a pregnant nonhuman primate. Nature medicine 22: 1256–1259.
12. Clement KH, Rudge TL, Mayfield HJ, Carlton LA, Hester A, Niemuth NA, Sabourin CL, Brys AM, and Quinn CP. 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. Clinical and Vaccine Immunology 17: 1753–1762. [PubMed: 20739500]
13. Awasthi S, Hook LM, Shaw CE, Pahar B, Stagray JA, Liu D, Veazey RS, and Friedman HM. 2017 An HSV-2 Trivalent Vaccine Is Immunogenic in Rhesus Macaques and Highly Efficacious in Guinea Pigs. PLoS pathogens 13: e1006141. [PubMed: 28103319]
14. Zimin AV, Cornish AS, Maudhoo MD, Gibbs RM, Zhang X, Pandey S, Meehan DT, Wipfler K, Bosinger SE, Johnson ZP, Tharp GK, Marcais G, Roberts M, Ferguson B, Fox HS, Treangen T, Salzberg SL, Yorke JA, and Norgren RB Jr. 2014 A new rhesus macaque assembly and annotation for next-generation sequencing analyses. Biology direct 9: 20. [PubMed: 25319552]

15. Harding JD 2017 Nonhuman Primates and Translational Research: Progress, Opportunities, and Challenges. ILAR Journal 58: 141–150. [PubMed: 29253273]

16. Ramesh A, Darko S, Hua A, Overman G, Ransier A, Francica JR, Trama A, Tomaras GD, Haynes BF, Douek DC, and Kepler TB. 2017 Structure and Diversity of the Rhesus Macaque Immunoglobulin Loci through Multiple De Novo Genome Assemblies. Frontiers in immunology 8: 1407. [PubMed: 29163486]

17. Cirelli KM, Carnathan DG, Nogal B, Martin JT, Rodriguez OL, Upadhyay AA, Enemuo CA, Gebru EH, Choe Y, Viviano F, Nakao C, Pauthner MG, Reiss S, Cottrell CA, Smith ML, Bastidas R, Gibson W, Wolabaugh AN, Melo MB, Cossette B, Kumar V, Patel NB, Tokatlian T, Menis S, Kulp DW, Burton DR, Murrell B, Schief WR, Bosinger SE, Ward AB, Watson CT, Silvestri G, Irvine DJ, and Crotty S. 2019 Slow Delivery Immunization Enhances HIV Neutralizing Antibody and Germinal Center Responses via Modulation of Immunodominance. Cell 177: 1153–1171.e1128. [PubMed: 31080066]

18. Corcoran MM, Phad GE, Bernat NV, Stahl-Hennig C, Sumida N, Persson MAA, Martin M, and Hedestam GBK. 2016 Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. Nature Communications 7: 13642.

19. Nishana M, and Raghavan SC. 2012 Role of recombination activating genes in the generation of antigen receptor diversity and beyond. Immunology 137: 271–281. [PubMed: 23039142]

20. Little AJ, Matthews A, Oettinger M, Roth DB, and Schatz DG. 2015 Chapter 2 - The Mechanism of V(D)J Recombination In Molecular Biology of B Cells (Second Edition). Alt FW, Honjo T, Radbruch A, and Reth M, eds. Academic Press, London 13–34.

21. Hesslein DGT, and Schatz DG. 2001 Factors and Forces Controlling V(D)J Recombination In Advances in Immunology. Dixon FJ, ed. Academic Press 169–232.

22. Murphy K, and Weaver C. 2017 Chapter 5 - The Generation of Lymphocyte Antigen Receptors. In Janeway’s Immunobiology, 7th ed Garland Science. 188–189.

23. Manis JP, Tian M, and Alt FW. 2002 Mechanism and control of class-switch recombination. Trends in Immunology 23: 31–39. [PubMed: 11801452]

24. Lucas JS, Murre C, Feeney AJ, and Riblet R. 2015 Chapter 1 - The Structure and Regulation of the Immunoglobulin Loci In Molecular Biology of B Cells (Second Edition). Alt FW, Honjo T, A. Radbruch, and Reth M, eds. Academic Press, London 1–11.

25. Markoulatos P, Siafakas N, and Moncany M. 2002 Multiplex polymerase chain reaction: a practical approach. Journal of clinical laboratory analysis 16: 47–51. [PubMed: 11835531]

26. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, and McCombie WR. 2007 Genome-wide in situ exon capture for selective resequencing. Nature genetics 39: 1522–1527. [PubMed: 17982454]

27. Yeku O, and Frohman MA. 2011 Rapid Amplification of cDNA Ends (RACE) In RNA: Methods and Protocols. Nielsen H, ed. Humana Press, Totowa, NJ 107–122.

28. Sundling C, Phad G, Douagi I, Navis M, and Karlsson Hedestam GB. 2012 Isolation of antibody V(D)J sequences from single cell sorted rhesus macaque B cells. Journal of immunological methods 386: 85–93. [PubMed: 22989932]

29. Wiehe K, Easterhoff D, Luo K, Nicely NI, Bradley T, Jaeger FH, Dennison SM, Zhang R, Lloyd KE, Stolarchuk C, Parks R, Sutherland LL, Scearce RM, Morris L, Kaewkungwal J, Nitayaphan S, Pitisuttithum P, Rerks-Ngarm S, Sinangil F, Phogat S, Michael NL, Kim JH, Kelsoe G, Montefiori DC, Tomaras GD, Bonsignori M, Santra S, Kepler TB, Alam SM, Moody MA, Liao HX, and Haynes BF. 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. [PubMed: 25526306]

30. Fu L, Li X, Zhang W, Wang C, Wu J, Yang H, Wang J, and Liu X. 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. [PubMed: 28813462]

31. Sundling C, Zhang Z, Phad GE, Sheng Z, Wang Y, Mascola JR, Li Y, Wyatt RT, Shapiro L, and Karlsson Hedestam GB. 2014 Single-cell and deep sequencing of IgG-switched macaque B cells reveal a diverse Ig repertoire following immunization. J Immunol 192: 3637–3644. [PubMed: 24623130]

32. Rosenfeld R, Zvi A, Winter E, Hope R, Israeli O, Mazor O, and Yaari G. 2019 A primer set for comprehensive amplification of V-genes from rhesus macaque origin based on repertoire sequencing. Journal of immunological methods 465: 67–71. [PubMed: 30471299]

33. Zhang W, Li X, Wang L, Deng J, Lin L, Tian L, Wu J, Tang C, Yang H, Wang J, Qiu P, Fu T-M, Saksena NK, Wang I-M, and Liu X. 2019 Identification of Variable and Joining Germline Genes and Alleles for Rhesus Macaque from B Cell Receptor Repertoires. The Journal of Immunology 202: 1612–1622. [PubMed: 30700589]

34. Vázquez Bernat N, Corcoran M, Hardt U, Kaduk M, Phad GE, Martin M, and Karlsson Hedestam GB. 2019 High-Quality Library Preparation for NGS-Based Immunoglobulin Germline Gene Inference and Repertoire Expression Analysis. Frontiers in Immunology 10: 660. [PubMed: 31024532]

35. Benichou J, Ben-Hamo R, Louzoun Y, and Efroni S. 2012 Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. Immunology 135: 183–191. [PubMed: 22043864]

36. Mamedov IZ, Britanova OV, Zvyagin IV, Turchaninova MA, Bolotin DA, Putintseva EV, Lebedev YB, and Chudakov DM. 2013 Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. Frontiers in immunology 4: 456. [PubMed: 24391640]

37. He L, Sok D, Azadnia P, Hsueh J, Landais E, Simek M, Koff WC, Poignard P, Burton DR, and Zhu J. 2014 Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. Scientific reports 4: 6778. [PubMed: 25345460]

38. Fu GK, Xu W, Wilhelmy J, Mindrinos MN, Davis RW, Xiao W, and Fodor SP. 2014 Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. Proceedings of the National Academy of Sciences of the United States of America 111: 1891–1896. [PubMed: 24449890]

39. Turchaninova MA, Davydov A, Britanova OV, Shugay M, Bikos V, Egorov ES, Kirgizova VI, Merzlyak EM, Staroverov DB, Bolotin DA, Mamedov IZ, Izraelson M, Logacheva MD, Kladova O, Plevova K, Pospisilova S, and Chudakov DM. 2016 High-quality full-length immunoglobulin profiling with unique molecular barcoding. Nature protocols 11: 1599–1616. [PubMed: 27490633]

40. Rock EP, Sibbald PR, Davis MM, and Chien YH. 1994 CDR3 length in antigen-specific immune receptors. J Exp Med 179: 323–328. [PubMed: 8270877]

41. Safonova Y, and Pevzner PA. 2019 De novo Inference of Diversity Genes and Analysis of Non-canonical V(DD)J Recombination in Immunoglobulins. Frontiers in Immunology 10: 987. [PubMed: 31134072]

42. Bashford-Rogers RJ, Palser AL, Idris SF, Carter L, Epstein M, Callard RE, Douek DC, Vassiliou GS, Follows GA, Hubank M, and Kellam P. 2014 Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. BMC Immunology 15: 29. [PubMed: 25189176]

43. Liu X, Zhang W, Zeng X, Zhang R, Du Y, Hong X, Cao H, Su Z, Wang C, Wu J, Nie C, Xu X, and Kristiansen K. 2016 Systematic Comparative Evaluation of Methods for Investigating the TCRβ Repertoire. PLOS ONE 11: e0152464. [PubMed: 27019362]

44. Macosko Evan Z., Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas Allison R., Kamitaki N, Martersteck Emily M., Trombetta John J., Weitz David A., Sanes Joshua R., Shalek Alex K., Regev A, and McCarroll Steven A.. 2015 Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161: 1202–1214. [PubMed: 26000488]

45. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, and Bielas JH. 2017 Massively parallel digital transcriptional profiling of single cells. Nature Communications 8: 14049.

46. Chattopadhyay PK, Gierahn TM, Roederer M, and Love JC. 2014 Single-cell technologies for monitoring immune systems. Nature Immunology 15: 128–135. [PubMed: 24448570]

47. Proserpio V, and Mahata B. 2016 Single-cell technologies to study the immune system. Immunology 147: 133–140. [PubMed: 26551575]

48. Papalexi E, and Satija R. 2017 Single-cell RNA sequencing to explore immune cell heterogeneity. Nature Reviews Immunology 18: 35–45.

49. Stubbington MJT, Rozenblatt-Rosen O, Regev A, and Teichmann SA. 2017 Single-cell transcriptomics to explore the immune system in health and disease. Science 358: 58–63. [PubMed: 28983043]

50. Chen H, Ye F, and Guo G. 2019 Revolutionizing immunology with single-cell RNA sequencing. Cellular & Molecular Immunology 16: 242–249. [PubMed: 30796351]

51. Neal JT, Li X, Zhu J, Giangarra V, Grzeskowiak CL, Ju J, Liu IH, Chiou S-H, Salahudeen AA, Smith AR, Deutsch BC, Liao L, Zemek AJ, Zhao F, Karlsson K, Schultz LM, Metzner TJ, Nadauld LD, Tseng Y-Y, Alkhairy S, Oh C, Keskula P, Mendoza-Villanueva D, De La Vega FM, Kunz PL, Liao JC, Leppert JT, Sunwoo JB, Sabatti C, Boehm JS, Hahn WC, Zheng GXY, Davis MM, and Kuo CJ. 2018 Organoid Modeling of the Tumor Immune Microenvironment. Cell 175: 1972–1988.e1916. [PubMed: 30550791]

52. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, Ji X, Han A, Krams SM, Pettus C, Haas N, Arlehamn CSL, Sette A, Boyd SD, Scriba TJ, Martinez OM, and Davis MM. 2017 Identifying specificity groups in the T cell receptor repertoire. Nature 547: 94–98. [PubMed: 28636589]

53. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB, Nguyen THO, Kedzierska K, La Gruta NL, Bradley P, and Thomas PG. 2017 Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature 547: 89–93. [PubMed: 28636592]

54. Ichinohe T, Miyama T, Kawase T, Honjo Y, Kitaura K, Sato H, Shin-I T, and Suzuki R. 2018 Next-Generation Immune Repertoire Sequencing as a Clue to Elucidate the Landscape of Immune Modulation by Host–Gut Microbiome Interactions. Frontiers in Immunology 9: 668. [PubMed: 29666626]

55. Ye J, Ma N, Madden TL, and Ostell JM. 2013 IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic acids research 41: W34–40. [PubMed: 23671333]

56. Brochet X, Lefranc MP, and Giudicelli V. 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 research 36: W503–508. [PubMed: 18503082]

57. Giudicelli V, Brochet X, and Lefranc MP. 2011 IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. Cold Spring Harbor protocols 2011: 695–715. [PubMed: 21632778]

58. Fu L, Niu B, Zhu Z, Wu S, and Li W. 2012 CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28: 3150–3152. [PubMed: 23060610]

59. Rice P, Longden I, and Bleasby A. 2000 EMBOSS: the European Molecular Biology Open Software Suite. Trends in genetics : TIG 16: 276–277. [PubMed: 10827456]

60. Lefranc M-P, Giudicelli V, Ginestoux C, Bodmer J, Müller W, Bontrop R, Lemaitre M, Malik A, Barbié V, and Chaume D. 1999 IMGT, the international ImMunoGeneTics database. Nucleic Acids Research 27: 209–212. [PubMed: 9847182]

61. Edgar RC 2010 Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460–2461. [PubMed: 20709691]

62. Chojnacki S, Cowley A, Lee J, Foix A, and Lopez R. 2017 Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. Nucleic acids research 45: W550–w553. [PubMed: 28431173]

63. Okonechnikov K, Fursov M, Golosova O, and U. t. team. 2012 Unipro UGENE: a unified bioinformatics toolkit Bioinformatics (Oxford, England) 28: 1166–1167. [PubMed: 22368248]

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 ZW, Kou ZC, Shen L, Reimann KA, and Letvin NL. 1993 Conserved T-cell receptor repertoire in simian immunodeficiency virus-infected rhesus monkeys Journal of immunology (Baltimore, Md : 1950) 151: 2177–2187.

67. Li Z, Liu G, Tong Y, Zhang M, Xu Y, Qin L, Wang Z, Chen X, and He J. 2015 Comprehensive analysis of the T-cell receptor beta chain gene in rhesus monkey by high throughput sequencing. Scientific Reports 5: 10092. [PubMed: 25961410]

68. Kuhn RM, Haussler D, and Kent WJ. 2013 The UCSC genome browser and associated tools. Briefings in bioinformatics 14: 144–161. [PubMed: 22908213]

69. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, and Madden TL. 2012 Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC bioinformatics 13: 134. [PubMed: 22708584]

70. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, and Turner S. 2009 Real-time DNA sequencing from single polymerase molecules. Science 323: 133–138. [PubMed: 19023044]

71. Scinicariello F, and Attanasio R. 2001 Intraspecies heterogeneity of immunoglobulin alpha-chain constant region genes in rhesus macaques. Immunology 103: 441–448. [PubMed: 11529934]

72. Rogers KA, Jayashankar L, Scinicariello F, and Attanasio R. 2008 Nonhuman Primate IgA: Genetic Heterogeneity and Interactions with CD89. The Journal of Immunology 180: 4816–4824. [PubMed: 18354205]

73. Vollmers C, Penland L, Kanbar JN, and Quake SR. 2015 Novel Exons and Splice Variants in the Human Antibody Heavy Chain Identified by Single Cell and Single Molecule Sequencing. PLOS ONE 10: e0117050. [PubMed: 25611855]

74. Woof JM, and Kerr MA. 2004 IgA function--variations on a theme. Immunology 113: 175–177. [PubMed: 15379977]

75. Crowley AR, and Ackerman ME. 2019 Mind the Gap: How Interspecies Variability in IgG and Its Receptors May Complicate Comparisons of Human and Non-human Primate Effector Function. Frontiers in Immunology 10: 697. [PubMed: 31024542]

76. Gadala-Maria D, Gidoni M, Marquez S, Vander Heiden JA, Kos JT, Watson CT, O'Connor KC, Yaari G, and Kleinstein SH. 2019 Identification of Subject-Specific Immunoglobulin Alleles From Expressed Repertoire Sequencing Data. Frontiers in Immunology 10: 129. [PubMed: 30814994]

77. Carter JA, Preall JB, Grigaityte K, Goldfless SJ, Jeffery E, Briggs AW, Vigneault F, and Atwal GS. 2019 Single T Cell Sequencing Demonstrates the Functional Role of αβ TCR Pairing in Cell Lineage and Antigen Specificity. Frontiers in Immunology 10: 1516. [PubMed: 31417541]

78. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, Willsey AJ, Joy JB, Scott JK, Graves TA, Wilson RK, Holt RA, Eichler EE, and Breden F. 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. [PubMed: 23541343]

Key points

1. PacBio Iso-Seq enables generation of immune repertoire constant region references
2. Full-length CCS reads benchmark targeted immune repertoire assay efficiencies

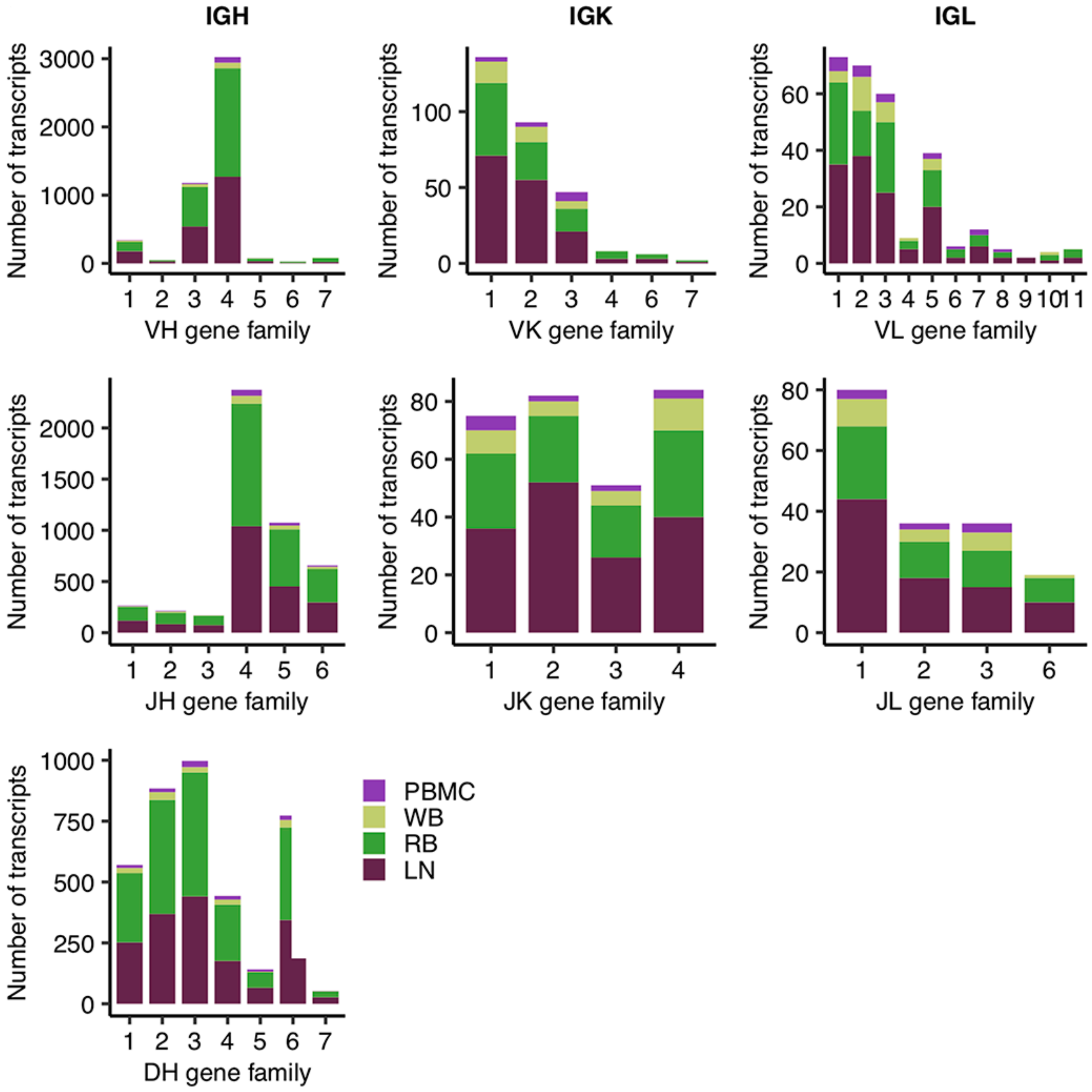

Fig 1.

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 1. Each individual plot shows the various known gene families of a particular gene segment type (e.g. VH), where gene families are clusters of highly similar gene segments within that gene segment type.

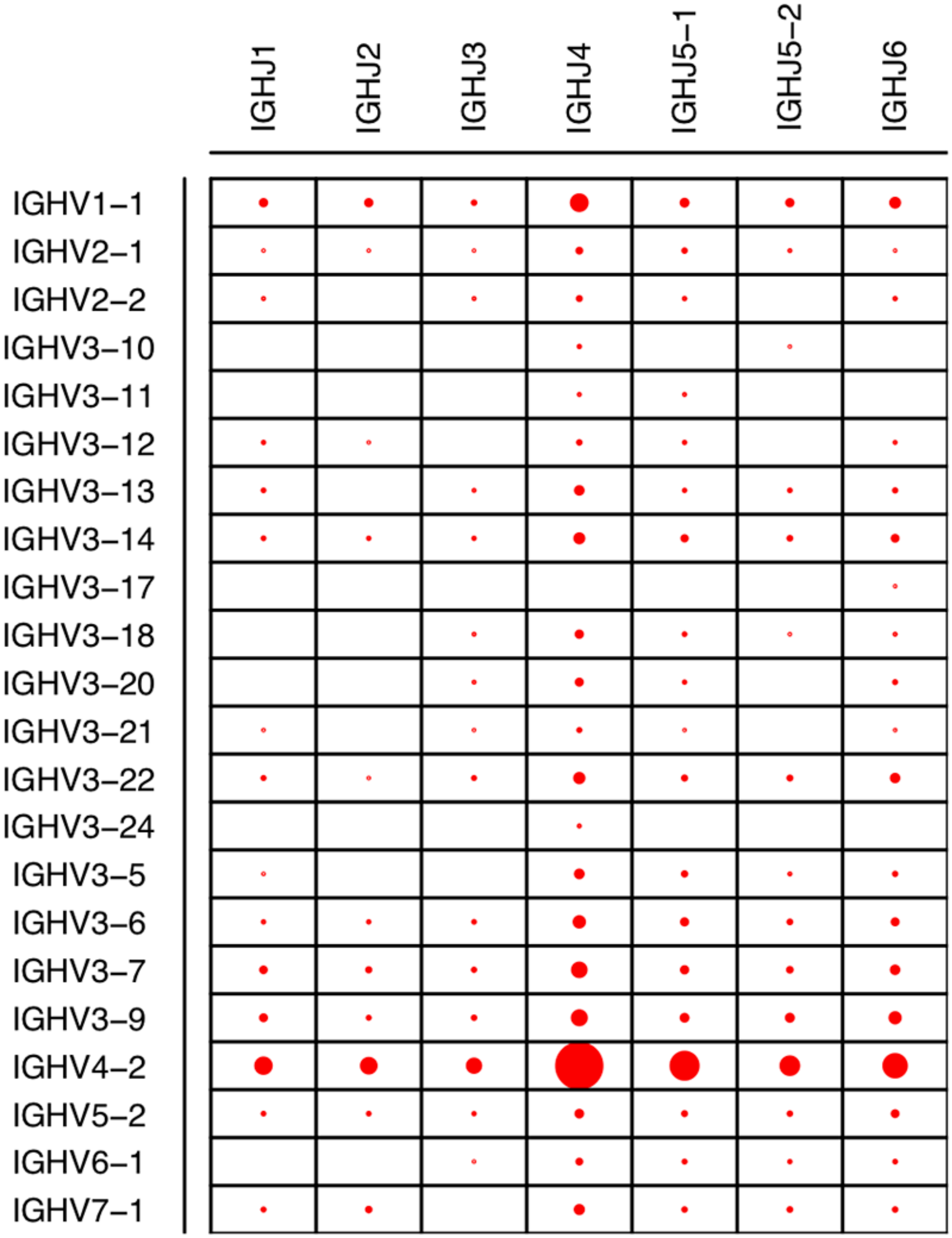

Fig 2.

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.

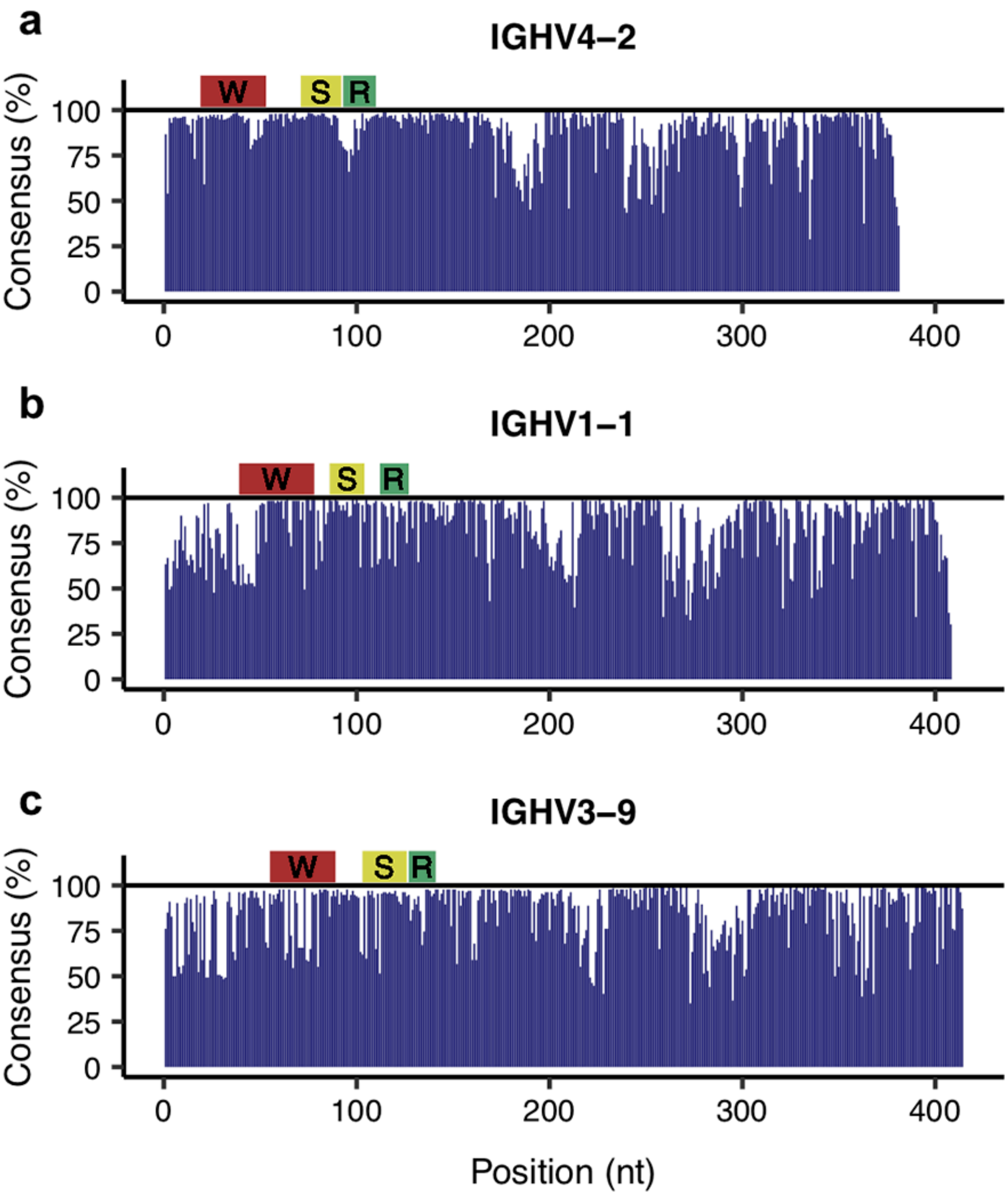

Fig 3.

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 1,383 sequences, while IGHV1-1 was from 163 sequences and IGHV3-9 was from 128 sequences. The regions targeted by primers from (S)undling et al (28), (W)iehe et al (29), and (R)osenfeld et al (32) respectively have either a yellow, red, or green bar above.

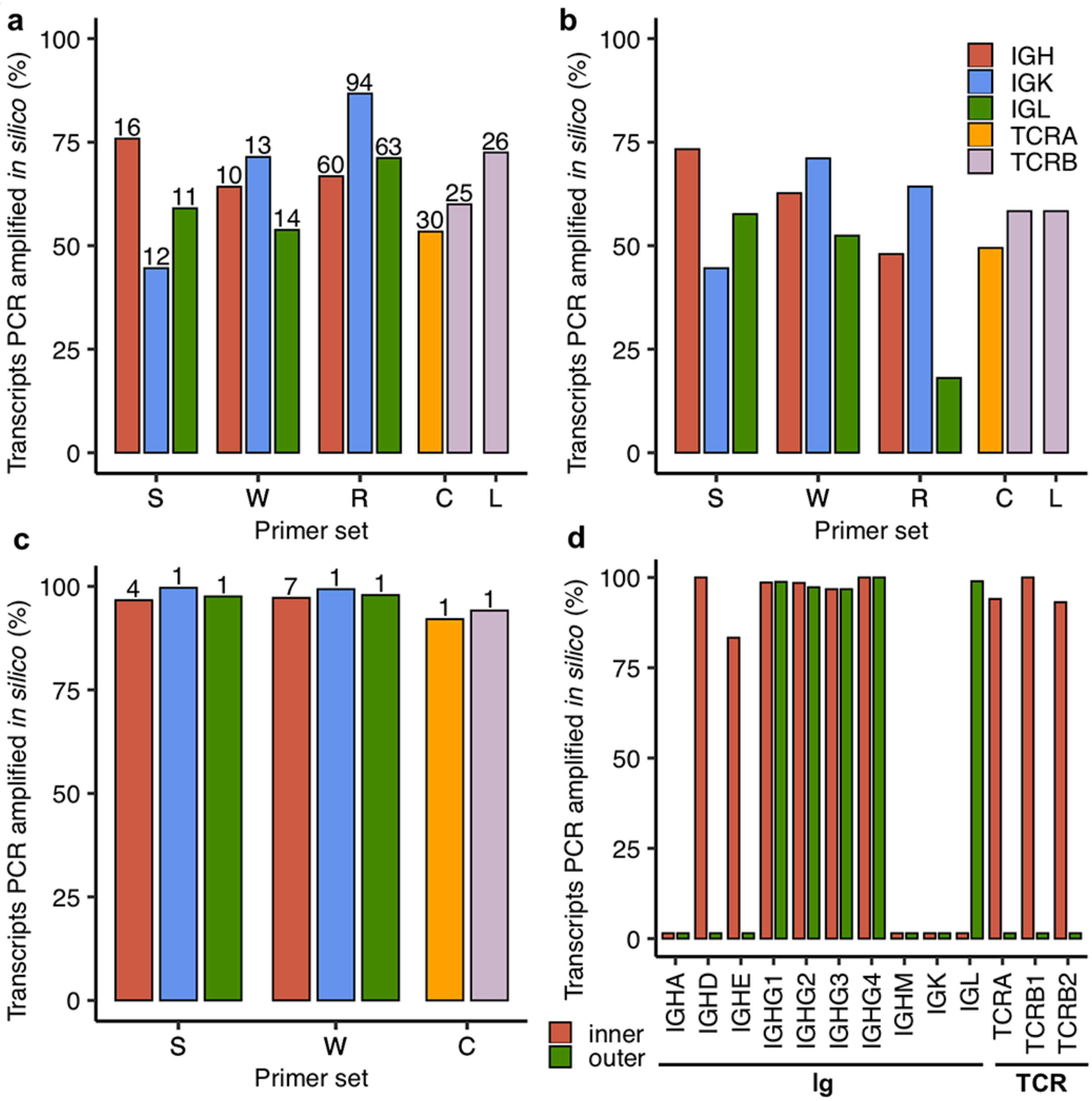

Fig 4.

In silico PCR analysis of rhesus-specific primers from (S)undling et al (28), (W)iehe et al (29), (R)osenfeld et al (32), (C)hen et al (66), and (L)i et al (67) as well as human 10x B and T cell V(D)J reverse primers (68). Only the inner primers were tested for S, W, while both inner and outer 10x primers were tested. The primers from C target both TCRA and TCRB, while 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 where only primers targeting the V gene segment within the variable 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 constant region) primers from S, W, R, C, and L. **C)** Amplification where only primers targeting the constant 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 10x 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.

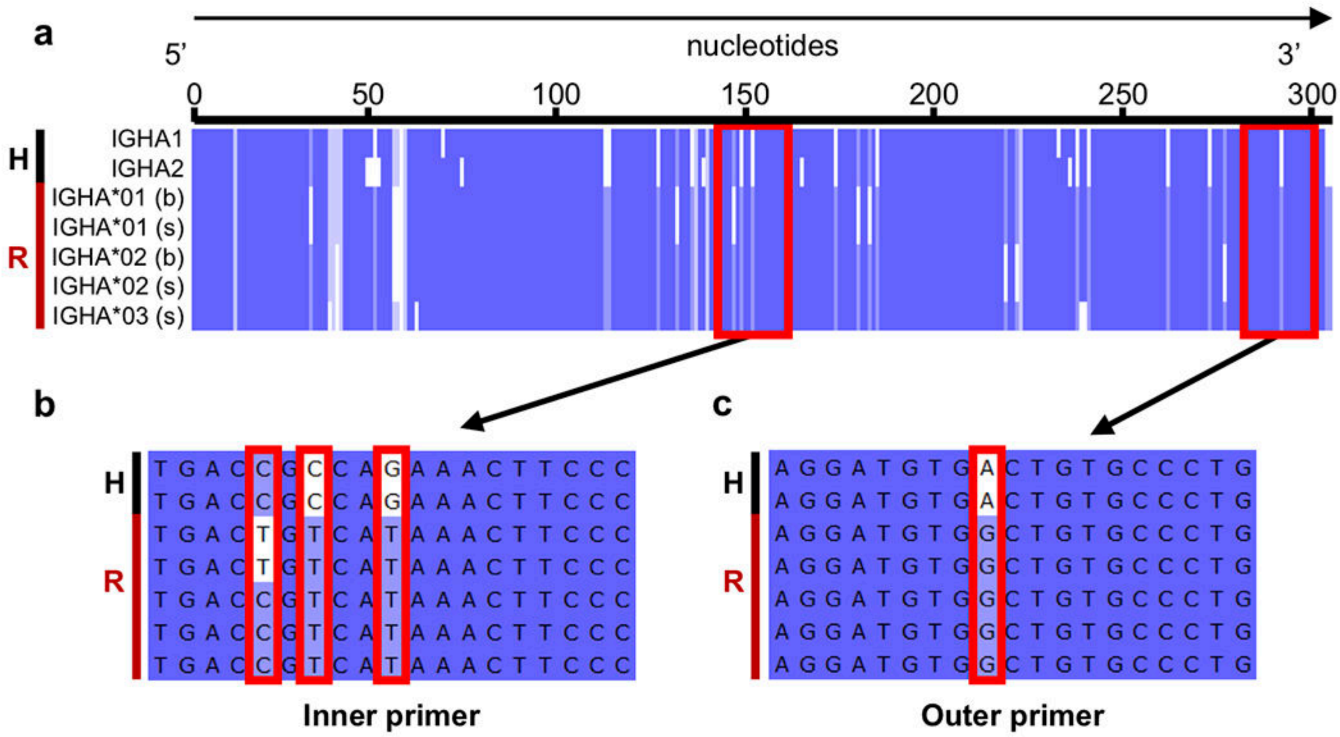
a
nucleotides
5'
3'
0
50
100
150
200
250
300
H
R
IGHA1
IGHA2
IGHA*01 (b)
IGHA*01 (s)
IGHA*02 (b)
IGHA*02 (s)
IGHA*03 (s)
b
H
R
T G A C C G C C A G A A A C T T C C C
T G A C C G C C A G A A A C T T C C C
T G A C T G T C A T A A A C T T C C C
T G A C T G T C A T A A A C T T C C C
T G A C C G T C A T A A A C T T C C C
T G A C C G T C A T A A A C T T C C C
T G A C C G T C A T A A A C T T C C C
Inner primer
c
H
R
A G G A T G T G A C T G T G C C C T G
A G G A T G T G A C T G T G C C C T G
A G G A T G T G G C T G T G C C C T G
A G G A T G T G G C T G T G C C C T G
A G G A T G T G G C T G T G C C C T G
A G G A T G T G G C T G T G C C C T G
A G G A T G T G G C T G T G C C C T G
Outer primer

Fig 5.

Comparison of human and rhesus IGHA reference sequences and illustration of human 10x 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) CDSs 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 10x primers are denoted by red boxes. **B-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.

Table 1.

Number of unique FL rhesus IG transcript sequences identified in each tissue. Counts for secreted heavy chain IG sequences (IGH A, D, E, G1-4, and M) are shown with membrane-bound counts in parenthesis where applicable. IGHA is further classified by allotype, conventionally denoted by an asterisk followed by a number.

	Rectal Biopsy	Whole Blood	PBMC	Lymph Node	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)

Table 2.

Number of unique FL rhesus TCR transcript sequences identified in each tissue.

	Rectal Biopsy	Whole Blood	PBMC	Lymph Node	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