MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2014) doi: 10.1111/1755-0998.12274

Next-generation approaches to advancing eco-immunogenomic research in critically endangered primates

P. A. LARSEN, C. R. CAMPBELL and A. D. YODER
Department of Biology, Box 90338, Duke University, Durham, NC 27708, USA

Abstract

High-throughput sequencing platforms are generating massive amounts of genomic data from nonmodel species, and these data sets are valuable resources that can be mined to advance a number of research areas. An example is the growing amount of transcriptome data that allow for examination of gene expression in nonmodel species. Here, we show how publicly available transcriptome data from nonmodel primates can be used to design novel research focused on immunogenomics. We mined transcriptome data from the world's most endangered group of primates, the lemurs of Madagascar, for sequences corresponding to immunoglobulins. Our results confirmed homology between strepsirrhine and haplorrhine primate immunoglobulins and allowed for high-throughput sequencing of expressed antibodies (Ig-seq) in Coquerel's sifaka (*Propithecus coquereli*). Using both Pacific Biosciences *RS* and Ion Torrent PGM sequencing, we performed Ig-seq on two individuals of Coquerel's sifaka. We generated over 150 000 sequences of expressed antibodies, allowing for molecular characterization of the antigen-binding region. Our analyses suggest that similar *VDJ* expression patterns exist across all primates, with sequences closely related to the human V_H3 immunoglobulin family being heavily represented in sifaka antibodies. Moreover, the antigen-binding region of sifaka antibodies exhibited similar amino acid variation with respect to haplorrhine primates. Our study represents the first attempt to characterize sequence diversity of the expressed antibody repertoire in a species of lemur. We anticipate that methods similar to ours will provide the framework for investigating the adaptive immune response in wild populations of other nonmodel organisms and can be used to advance the burgeoning field of eco-immunology.

Keywords: adaptive immune system, antibody repertoire, antibodyome, eco-immunology, Ig-seq, immunogenomics, *Propithecus coquereli*

Received 25 February 2014; revision received 1 May 2014; accepted 5 May 2014

Introduction

Genomic data from nonmodel organisms are flooding public repositories. And although it is indisputable that these data will increasingly serve as valuable tools for developing studies novel to these species (Hudson 2008; Ekblom & Galindo 2011), many investigators are at a loss as how best to utilize these genomic resources. In particular, the growing number of transcriptome databases provides important insight into the functional genomics and gene expression profiles of species not typically targeted for investigation (Vera *et al.* 2008; Renaut *et al.* 2010; Pipes *et al.* 2013; Ekblom *et al.* 2014). These data are readily available and provide as-yet-untapped opportunities for elucidating the molecular underpinnings of the interaction between species and their environment, which in turn can be used to develop novel research areas. An example can be drawn from eco-immunology, a growing field of study that aims to better understand the interaction between natural populations and their associated pathogens, especially disease ecology and host immune system function (Demas & Nelson 2011; Hawley & Altizer 2011; Pedersen & Babayan 2011). The interplay among pathogens, hosts and their respective genomes is inherently complex and thus has made the wide application of traditional immunological research techniques to nonmodel species difficult (Pedersen & Babayan 2011). High-throughput sequencing applications have created opportunities for eco-immunological studies and are allowing researchers to study the molecular interplay between the immune systems of nonmodel organisms and their pathogens. In particular, the high-throughput sequencing of expressed immunoglobulin

Correspondence: Peter A. Larsen, Fax: 919-660-7293; E-mail: peter.larsen@duke.edu

P. A. Larsen and C. R. Campbell contributed equally to this work

genes (i.e. antibodies; Ig-seq *sensu* Georgiou *et al.* (2014)) is a novel transcriptomic approach that holds great promise for eco-immunology research.

Antibodies are an important component of the vertebrate adaptive immune response as they actively neutralize antigens by binding to epitopes present on the antigen surface. Antibody molecules primarily consist of two heavy chains and two light chains, each with variable (V) and constant (C) domains (Fig. 1). The heavy-chain V region accounts for a large percentage of physical interaction with antigens and is formed by the recombination of numerous germline variable (*V*), diversity (*D*) and joining (*J*) gene segments (e.g. the human genome contains approximately 47 *V*, 23 *D* and 6 *J* functional heavy-chain gene segments). This recombination is accompanied by a complex series of combination and hypermutation events which generate exceptional amounts of antibody diversity [e.g. ~1×10^{13} potential unique antibodies in humans (Janeway *et al.* 2004; Schroeder 2006)]. The abundance of unique antibodies arising from this complex process allows for the neutralization of a plethora of invading pathogens and plays a central role in the adaptive immune response. Thus, Ig-seq studies are of great interest because they provide the foundation for the development of new disease surveillance methods, antibody manufacturing and vaccine design (Benichou *et al.* 2012; Robins 2013; Vollmers *et al.* 2013; Zhu *et al.* 2013a; Georgiou *et al.* 2014).

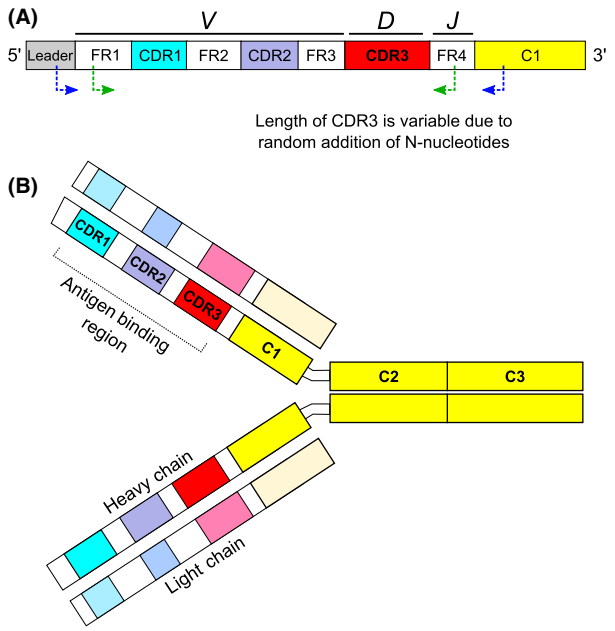

Fig. 1 Diagrams of immunoglobulin heavy-chain mRNA (A) and IgG structure (B). (A) Heavy-chain mRNA is the result of somatic recombination of germline variable (*V*), diversity (*D*) and joining (*J*) segments. The mRNA molecule consists of framework (FR1-4), complementarity determining regions (CDR1-3) and immunoglobulin (IgG, IgM, IgD, etc.) constant regions (C1 to C4; IgG C1 is shown). Exploratory sequencing was performed using PacBio *RS* circular consensus sequencing of amplicons generated using primers specific to a conserved region within the 5′ leader and IgG constant region 1 (C1, exterior blue arrows). Additional sequencing using the Ion Torrent PGM was performed on amplicons generated using primers targeting conserved regions within FR1 and a sifaka-specific J primer (interior green arrows). (B) A generic IgG antibody molecule consists of two heavy chains, two light chains and two constant regions. The region that physically interacts with pathogens is commonly referred to as the antigen-binding region and heavy-chain CDRs 1-3 account for the majority of this physical interaction.

We posit that the characterization of antibody repertoires in wild populations will be a fundamental contribution to the field of eco-immunology and will aid in the development of novel disease surveillance methodologies as well as contribute to advanced conservation approaches. Indeed, Ig-seq is providing key insights into the variation of the adaptive immune system and into the immune response to vaccines and infections in human subjects and model species (Glanville *et al.* 2009; Arnaout *et al.* 2011; Racanelli *et al.* 2011; Briney *et al.* 2012; Sundling *et al.* 2012, 2014; Parameswaran *et al.* 2013; Zhu *et al.* 2013a,b), although very few studies have utilized Ig-seq in nonmodel species (Larsen & Smith 2012; Wu *et al.* 2012; Castro *et al.* 2013). Until recently, the absence of high-quality genomes and immunogenetic data from nonmodel species has hindered the implementation of advanced approaches such as Ig-seq. Now, however, *de novo* transcriptome assemblies and databases of raw transcriptome data are readily available resources that can be used to advance eco-immunogenetic studies. Here, we show how recently available transcriptome data originating from a study of endangered primates (Perry *et al.* 2012) can be used to design an Ig-seq experiment in order to produce meaningful immunogenomic information.

We focus our study on the world's most endangered group of primates, the lemurs of Madagascar. Lemurs have evolved in isolation on the island of Madagascar for at least 60 million years (Yoder & Yang 2004). They are exceptionally diverse (Mittermeier *et al.* 2010) and highly endangered. Increasingly, natural populations are becoming fragmented due to ongoing destruction of their natural habitat as well as to the effects of global climate change. This habitat loss has created a challenging and fragile environment that threatens the health of extant lemur populations as well as prospects for their long-term survival (Barrett *et al.* 2013). Accordingly, eco-immunogenetic research utilizing advanced methods (i.e. Ig-seq and other gene expression studies) is of vital importance for assessing the current health status of wild

lemur populations. This approach allows for the direct examination of gene expression profiles in response to stress as well as to parasite and pathogen infections (Hawley & Altizer 2011; Pedersen & Babayan 2011; Tung *et al.* 2012). To this end, we mined publicly available lemur transcriptome data to target expressed lemur antibodies for high-throughput sequencing. Prior to our study, no expressed antibody sequence data existed for lemurs. Even so, previous studies focused on primate evolution and biomedical applications have identified sequence homology among the genes underlying the formation of antibodies (Meek *et al.* 1991; Helmuth *et al.* 2000; Link *et al.* 2002; Sundling *et al.* 2012). This information motivated us to map lemur transcriptome data to the human heavy-chain locus, a region approximately 1.27 Mb long that encodes the heavy-chain immunoglobulin genes (Matsuda *et al.* 1998).

In the light of our mapping results, we used a combination of Pacific Biosciences (PacBio) *RS* single-molecule real-time (SMRT) circular consensus sequencing (CCS) and Ion Torrent PGM sequencing to perform Ig-seq in two individuals of Coquerel's sifaka (*Propithecus coquereli*). PacBio SMRT CCS technology provides high-quality consensus sequences for individual templates of varying lengths (e.g. 500–2500 bp) and was selected based on the absence of systematic error profiles. Moreover, PacBio CCS allowed for the sequencing of relatively long templates that can vary in length (i.e. due to complementarity determining region 3 (CDR3) length variability; see Fig. 1), which is crucial given the potential for extremely long antibody sequences (see Larsen & Smith 2012). Whereas PacBio CCS was useful for exploratory and relatively low-throughput sequencing, the Ion Torrent PGM allowed for greater throughput of templates up to 400 bp in length. Our molecular analyses focused largely on the CDR3 region within expressed sifaka antibodies. This region accounts for the majority of the physical interaction between antibodies and antigens (Wedemayer *et al.* 1997; Xu & Davis 2000), making its characterization a useful first step towards elucidating the evolution of the lemur adaptive immune system and serving to inform future disease-related lemur conservation efforts as they relate to the transmission and abundance of environmental pathogens. More broadly, in combination with existing or newly generated transcriptome data, our experimental design can be modified and adapted to almost any vertebrate species.

Materials and methods

Quality filtering, mapping and primer design

Raw Illumina transcriptome sequence data for four individuals each of humans (*Homo sapiens*), chimpanzees (*Pan troglodytes*) and four species of lemurs (*Propithecus coquereli*, *Daubentonia madagascarensis*, *Varecia variegata* and *Eulemur mongoz*) generated by Perry *et al.* (2012) were gathered from the NCBI short-read archive (SRP008743). Species-specific sequence data were separated by individual, and all reads were quality trimmed using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) with a minimum phred score of q20 and minimum length of 24 base pairs for any individual read (Table S1, Supporting information). Resulting reads were then mapped to the human heavy-chain locus (~1.27 Mb on chromosome 14; RefSeq Accession NG_001019.5) using the multiseed alignment software BOWTIE 2 version 2.1.0 (Langmead *et al.* 2009) with the preset option – very sensitive local. We further assessed species-by-species coverage within relatively conserved exons of main immunoglobulin classes, including IgA, IgG isotypes 1–4 and IgM (Table S2, Supporting information).

Resulting alignments were used to confirm homology between human and lemur IgH loci and select/design primers that would amplify the antigen-binding region of expressed antibodies within sifakas by binding to *V* segments (forward primers) and IgG exons (reverse primers). All primers used to generate molecular data herein are presented in Table S3 (Supporting information). A sequencing run targeting *V* segment family 3 (V_H3) and IgG-specific antibodies was performed using PacBio SMRT CCS (see below). The resulting sequence data were used to design a sifaka-specific *J* segment primer, which allowed for global immunoglobulin heavy (IgH) expression analyses using the 400-bp Ion Torrent 314 chip.

RNA extraction and cDNA synthesis

Animal procedures were reviewed and approved by the Duke University IACUC (protocol number A294-12-11) and Duke Lemur Center (DLC) Research committees. Peripheral blood samples (2 cc) were collected from 2 sifakas (sifaka 1: DLC 6884, 6-year-old female; sifaka 2: DLC 6583, 19-year-old male). Both individuals were deemed healthy by DLC veterinarian staff and exhibited no signs of infection/illness during, and in the weeks prior to, the collection of blood samples. Whole blood was centrifuged at 1057 ***g*** for 15 min at room temperature, and leukocytes were collected and stored at −80 °C. Total RNA was isolated from leukocyte-enriched samples using the RiboPure™-Blood Kit (Ambion by Life Technologies), and OD260/280 measurements were taken to quantify each sample. cDNA of full length mRNA was synthesized using the SuperScript™ III First-Strand Synthesis System (Invitrogen by Life Technologies) and the included poly-T oligo.

PacBio SMRT CCS: IgG specific

Based on our transcriptome mining results, we selected a forward primer targeting a conserved leader sequence of the V_H3 region and designed a sifaka-specific reverse primer targeting IgG constant region 1 (Table S3, Supporting information), the combination of which allowed for PCR amplification of the antigen-binding region of expressed IgG antibodies within sifaka 1 (DLC 6884). Amplicons were obtained using a high-fidelity Taq DNA polymerase (Phusion; New England Biolabs) in 50-μL reactions and the following thermal profile: initial denaturation at 98 °C for 30 s followed by 35 cycles of 98 °C for 10 s, 69 °C for 30 s and 72 °C for 20 s with a final extension step of 72 °C for 5 min. Amplicons from two PCRs were pooled and gel-purified (1.5% agarose, SYBR green stain), and ~800 ng was provided to the Duke IGSP sequencing core facility for SMRT CCS. Sequencing library preparation followed existing protocols for small-insert library preparation (Pacific Biosciences), and sequencing was performed using a Pacific Biosciences *RS* sequencer with two SMRT cells (120 min movie run-times). The resulting ccs.fastq files (one per SMRT cell) were concatenated and quality filtered (minimum q20 for 90% of bases averaged per read) using the FASTX-Toolkit [as implemented in the Galaxy platform (Blankenberg *et al.* 2010)]. Quality filtered data were aligned to the V_H3 and IgG primers used for amplification using GENEIOUS version 7.0 software (http://www.geneious.com/). Reads successfully mapping to both primers were retained for downstream analysis using the International ImMunoGeneTics information system [IMGT; (Lefranc *et al.* 2009)]. IMGT/HighV-Quest (Alamyar *et al.* 2012) was used to map sifaka IgG reads to the human heavy-chain immunoglobulin database, and downstream analyses of IMGT mapping results were performed using the Galaxy platform.

Ion Torrent: global IgH

Using sequence data generated from SMRT CCS, we identified a highly conserved region corresponding to the sifaka *J* segment(s) and created an additional reverse primer. PCRs were performed in 50 μL reactions using this newly designed primer and forward primers targeting conserved regions of human V_H families 1–6 (Table S3, Supporting information) resulting in the amplification of an approximately 380 bp region of the antigen-binding region of the global circulating IgH repertoire within sifaka 2 (DLC 6583). Amplicons were obtained using a high-fidelity Taq DNA polymerase (Phusion; New England Biolabs) and the following touchdown thermal profile: initial denaturation at 98 °C for 30 s, 10 cycles of 98 °C for 10 s, 69 °C for 30 s (decreasing 0.5 °C per cycle), and 72 °C for 20 s followed by 24 cycles of 98 °C for 10 s, 64 °C for 30 s, and 72 °C for 20 s with a final extension step of 72 °C for 5 min. Amplicons (two 50 μL reactions per V_H family) were gel-purified (1.5% agarose; SYBR green stained) and were pooled to 2 μg total product. This library was then sequenced by the Duke IGSP sequencing core using an Ion Torrent PGM (314 chip with 400 bp sequencing kit) following standard manufacturers procedures. Quality filtering, primer trimming and IMGT analyses of Ion Torrent data followed those described above for analyses of PacBio CCS data. χ^2 statistics for amino acid distributions (i.e. IgG PacBio data vs. global IgH Ion Torrent data) were performed following Collis *et al.* (2003). CIRCOS software (Krzywinski *et al.* 2009) was utilized to visualize IMGT mapping results of the Ion Torrent data presented herein.

Results

Transcriptome mapping

Transcriptome sequence data of six primate species, generated by Perry *et al.* (2012), were mapped to the human heavy-chain locus. Of the reads downloaded for *Homo sapiens*, 97.6% were retained after quality filtering, leaving 106 796 154 reads of which 3.42% (3667 309 reads) mapped to the heavy-chain locus. Using the same filtering and mapping criteria, 96.4% (112 862 480 reads) of raw sequence data from *P. coquereli* was retained post-filtering, and of these, 0.71% (805 595 reads) mapped to the human heavy-chain locus. Mapping percentages for the remaining species varied from 2.92% to 3.89% (Table 1).

We further analysed these results to assess the amount of the transcriptomic data that mapped to

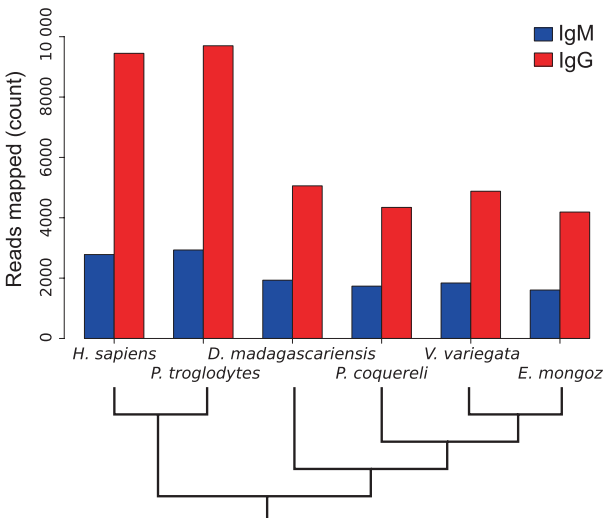

Fig. 2 Comparison of primate transcriptome reads (Perry *et al.* 2012) that successfully mapped to human IgG and IgM exons (see Methods). Phylogenetic relationships among species included in the analysis appear below the graph.

Table 1 Summary statistics for primate transcriptome data gathered from Perry *et al.* (2012) and BOWTIE 2 alignment to the human heavy-chain locus (see Materials and methods)

Species	Individuals (count)	Reads (count)	Filtered (per cent)	Reads mapped to IgH (per cent)	Reads mapped to IgH (count)
Daubentonia madagascariensis	4	117 264 958	3.71%	0.62%	699 199
Pan troglodytes	4	106 174 966	3.80%	3.22%	3293 852
Homo sapiens	4	109 456 478	2.43%	3.43%	3667 309
Eulemur mongoz	4	112 691 070	4.59%	0.68%	729 779
Propithecus coquereli	4	117 127 452	3.64%	0.71%	805 595
Varecia variegata	4	104 289 004	3.46%	1.12%	1132 143

Table 2 Summary of PacBio CCS and Ion Torrent sequencing and IMGT results

	PacBio	Ion Torrent
Raw reads	34 074	236 484
Filtered reads	15 282	144 441
IMGT annotated	12 827	139 159
IMGT no results	2455	2093
Complete CDR annotation	7023	98 783
Unique CDR3	2032	9973

specific coding regions within the heavy-chain locus, including the exons coding for specific antibody isotypes IgA, IgD, IgG (1–4) and IgM. Collectively, the *H. sapiens* transcriptomes had 9499 reads map to IgG (1–4) exons and 2781 reads map to IgM, while *P. coquereli* had 4343 and 1734 reads map to these same exons, respectively (Fig. 2, Table S2, Supporting information). Mapping results for all species are found in Table 1.

PacBio CCS results: sifaka IgG

SMRT CCS resulted in 34 074 reads, and of these, 21 752 (63.8%) were retained for downstream analyses based on our quality filtering criteria. A total of 15 282 reads successfully mapped to both V_H3 and IgG amplification primers (Table 2). A bimodal distribution of sequence lengths was observed with peaks at ~390 and ~545 bp (Fig. S1, Supporting information). IMGT analyses of the 15 282 sequences resulted in 13 073 ranging between 44% and 92% similarity with human V_H3 segments (Fig. S2, Supporting information). Human *V* segments V3-23, V3-74, V3-9 and V3-71 accounted for ~67% of the matches between the sifaka IgG repertoire and human heavy-chain locus. Approximately 13% of all sifaka IgG sequences were identified as sharing homology with multiple human V_H3 segments (Fig. S2, Supporting information). IMGT identified 11 810 sifaka IgG sequences that shared between 60 and 80% sequence homology with human *J* segments and human J_H3, J_H4

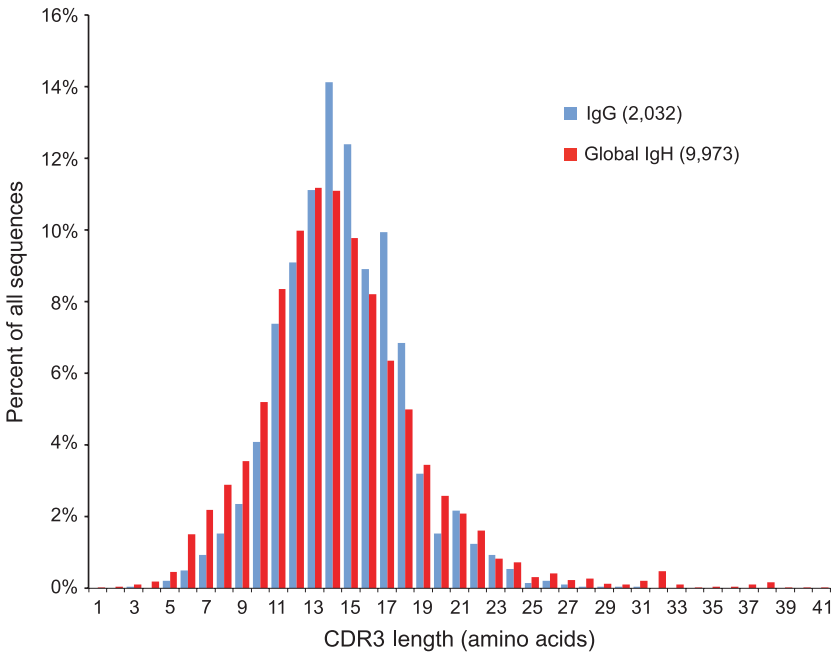

Fig. 3 Length distribution (in amino acids) of unique functional sifaka CDR3 sequences identified by IMGT analyses. Sequences generated using PacBio CCS targeted IgG, while those generated using Ion Torrent technology targeted global IgH.

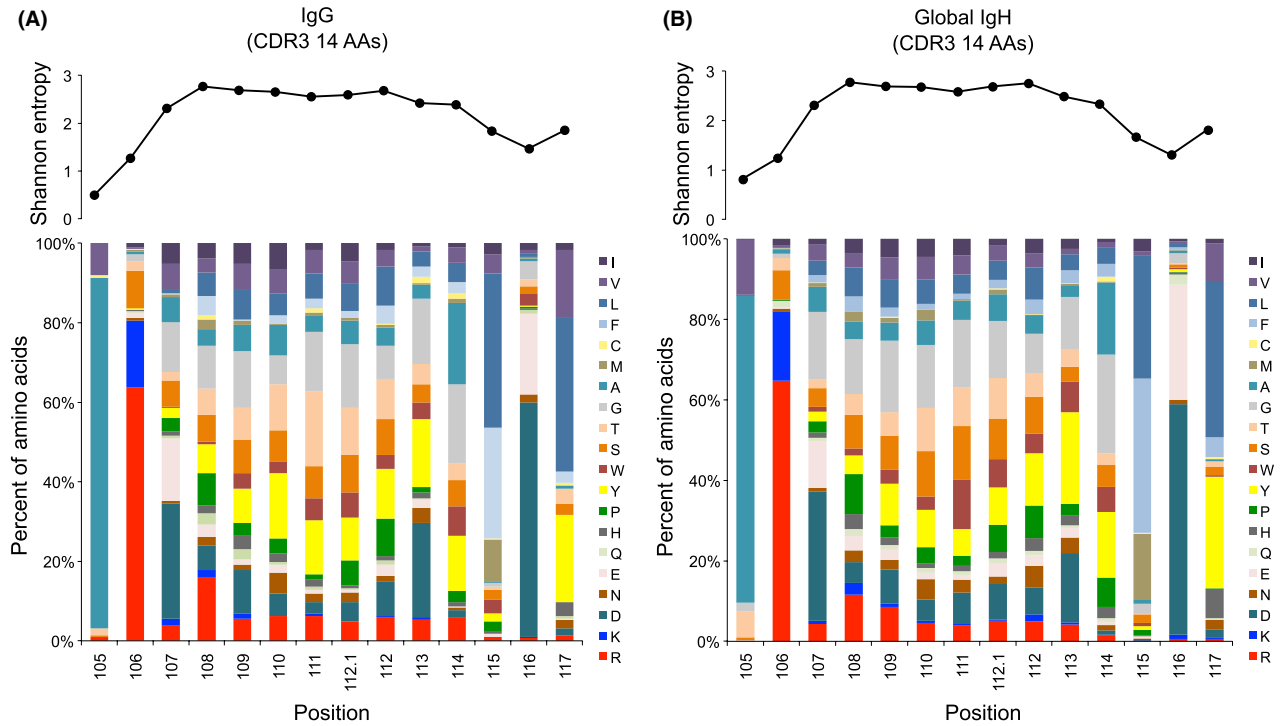

Fig. 4 Shannon entropy values (top panels) and amino acid frequencies (bottom panels) of unique CDR3 sequences identified by IMGT analyses of sifaka IgG (A; n = 287) and global IgH (B; n = 1106) sequence data. CDR3 lengths of 14 amino acids are shown, and positions follow the IMGT numbering convention. Each bar represents 100% of the amino acids identified at that position. Amino acids are grouped according to hydrophobicity with charged amino acids at the bottom and hydrophobic amino acids at the top.

and J_H6 segments accounted for the majority (Fig. S3, Supporting information). Our analyses resulted in the identification of 7023 complete CDR1, CDR2 and CDR3 regions within the sifaka IgG sequence data. Mean lengths consisted of 7.3 (±1.5), 7.8 (±0.8) and 14.7 (±3.7) amino acids for CDR1, CDR2 and CDR3, respectively. We further investigated sequence length of the CDR3 region from sifaka IgG sequences and identified 2032 unique IgG CDR3s ranging from 3 to 31 amino acids in length (Fig. 3). Analyses of amino acid content for the most common CDR3 size class, 14 amino acids (n = 287), revealed conserved amino acids at IMGT positions 105, 106 and 116 (Fig. 4). Overall, analyses of amino acid content of unique CDR3 sequences revealed common usage of alanine, glycine, aspartic acid and tyrosine (Fig. 5). We identified 2455 sifaka IgG sequences as having no IMGT results; however, 1870 (76%) of these were identified as putative immunoglobulins based on local BLAST queries of the NCBI immunoglobulin database (http://www.ncbi.nlm.nih.gov/igblast/).

IonTorrent results: sifaka IgH

A total of 236 484 Ion Torrent reads were within the expected size distribution of our IgH amplicons (~250–450 bp; Table 2). Quality filtering retained 144 441 sequences for analyses that had an average sequence length of 328 bp with a unimodal length distribution (Fig. S4, Supporting information). IMGT analysis resulted in 139 159 sifaka IgH sequences ranging from 45% to 89% sequence similarity to human V segments (Fig. 6). Approximately 92% (129 214 sequences) of all sifaka IgH sequences shared homology with the human V_H3 family and 7% (9900) shared homology with the human V_H4 family (as identified by IMGT analyses; Fig. 6), whereas only 42 and 3 sifaka IgH reads were identified by IMGT as sharing homology with human V_H1 and V_H2 families, respectively. Low sequence similarity caused IMGT to assign multiple human V and J segments for individual reads (Fig. S5, Supporting information). No sequences were identified as sharing homology with human V_H5 or V_H6 families.

IMGT identified complete CDR1, CDR2 and CDR3 amino acid sequences for 98 783 sequences. Global CDR1, CDR2 and CDR3 amino acid lengths averaged 7.9 (±0.5), 7.7 (±0.9) and 13.7 (±3.9) amino acids, respectively. We identified 9973 unique CDR3 sequences, and these ranged in length from 1 to 41 amino acids (Fig. 3). Amino acid content for CDR3 sequences of 14 amino acids in length is presented in Fig. 4. Global IgH (Ion Torrent) amino acid content of unique CDR3 sequences was similar to the amino acid content for IgG (PacBio)

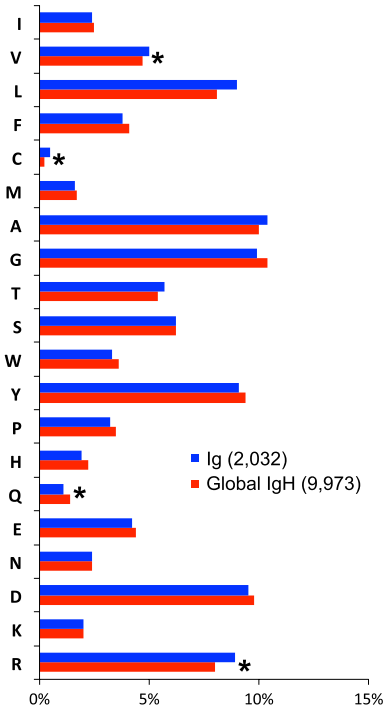

Fig. 5 Amino acid frequencies of unique CDR3 sequences identified by IMGT analyses of expressed sifaka antibodies using PacBio CCS (IgG) and Ion Torrent (global IgH) sequencing. Asterisks identify significant differences in amino acid frequency between the two data sets based on χ^2 tests.

sequences; however, χ^2 tests identified significant differences in valine, cysteine, glutamine and arginine usage between the two data sets ($P < 0.0005$; Fig. 5). We combined unique CDR3 sequences originating from Ion Torrent sequencing (IgH, sifaka 2) with unique PacBio CDR3 sequences (IgG, sifaka 1) and identified 6 amino acid sequences shared between individuals (~0.049% of 12 005 sequences). We identified 2093 sequences with no IMGT results; however, 348 of these (16.6%) were identified as putative immunoglobulins based on local BLAST queries of the NCBI immunoglobulin database.

Discussion

Our data mining and Ig-seq approach resulted in the generation of sequence data from over 150 000 expressed *P. coquereli* antibodies. The success of these methods demonstrates that widely available transcriptome data can be mined to develop novel immunogenetic research in nonmodel primates. Our transcriptome mapping identified homology within immunoglobulin sequence variation across all six primate species (Fig. 2), a result largely congruent with previous studies of haplorrhine primates (Helmuth *et al.* 2000; Von Büdingen *et al.* 2001; Sundling *et al.* 2012). We therefore hypothesize that several immunoglobulin heavy-chain *V* segment families are conserved across all primates. Seven V_H families have been identified in primates (V_H1–V_H7; Matsuda *et al.* 1998; Sundling *et al.* 2012), and the V_H3 and V_H4 families exhibited the greatest depth of coverage among the six primate species examined herein (Table S4, Supporting information), while the majority (~92%) of our global IgH Ion Torrent sequence data shared homology with the human V_H3 family. These results are not surprising inasmuch as the V_H3 gene family is relatively conserved across mammals (Tutter & Riblet 1989), and both the V_H3 and V_H4 families are abundant in human and macaque antibody repertoires (Helmuth *et al.* 2000; Arnaout *et al.* 2011; Briney *et al.* 2012; Sundling *et al.* 2014). Our Ion Torrent data reinforce this observation and suggest that sifaka antibody repertories heavily utilize *V* segments closely related to human *V* segment 3-23, a commonality shared with similar Ig-seq data from humans [Fig. 6; Briney *et al.* (2012)].

Previous research has shown conserved *V* family sequence variation in haplorrhines (Sundling *et al.* 2012); therefore, the paucity of sifaka immunoglobulin sequences sharing homology with human V_H1, V_H2, V_H5 and V_H6 families within our data warrants further inspection and may have several explanations. For example, human V_H5 and V_H6 families are comprised of 2 and 1 genes, respectively, and are depauperate with respect to the expansive V_H3 family (~20 functional genes). Thus, sifaka antibodies utilizing *V* segments related to these two families may be relatively rare in the circulating repertoire assuming the primate heavy-chain locus is evolutionarily conserved. Alternatively, lemurs have evolved in isolation for at least 50 million years, and thus, it is possible that there are novel aspects of the lemur adaptive immune system with respect to other primates. Our IMGT analysis resulted in *V* segment sequence homology ranging from 40% to 80%, values that suggest the possibility of uncharacterized *V* segment diversity within the lemuriform lineage. Nevertheless, the amino acid variation observed within the antigen-binding region was consistent with what is currently known of primate immunoglobulin diversity. Additional progress towards the assembly and annotation of an exemplar lemur genome, and more specifically the immunoglobulin heavy-chain locus, is required to make meaningful comparisons to haplorrhine primates and to formally test hypotheses regarding evolutionary patterns of lemur immunoglobulin families.

Variability of the sifaka antigen-binding region

Molecular characterization of the antigen-binding region of expressed antibodies is critical to our understanding of the adaptive immune response in nonmodel species.

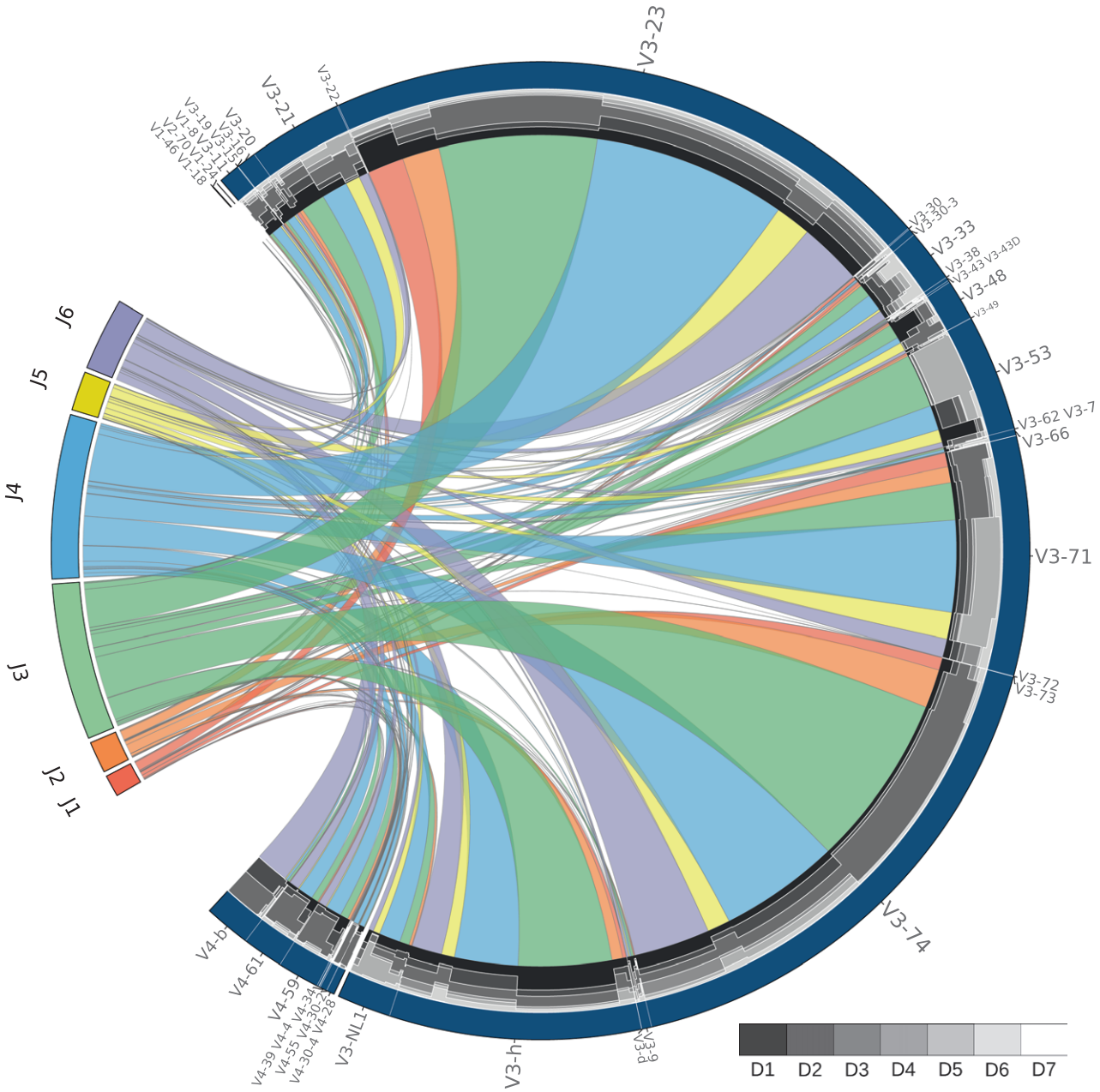

Fig. 6 Circos diagram showing sequence diversity of expressed sifaka antibodies (131 829 sequences, generated using an Ion Torrent PGM) based on their similarity to human *V*, *D* and *J* segments. Human *V* segments are shown in blue along the right outer edge of the diagram. The width of each *V* segment is representative of the relative frequency of the number of sifaka antibody sequences sharing homology with that particular segment (e.g. the majority of sifaka antibodies shared sequence similarity with V3-23, and very few were identified as closely related to V3-22). Human *J* segments are shown on the left outer edge, width again demonstrating relative frequency of homologous regions within sifaka antibodies. Ribbons are coloured according to *J* segment usage and connect *V* and *J* segments, thus displaying relative proportion that each *V*-*J* combination was identified in the circulating sifaka antibody repertoire. Relative *D* segment usage is shown in the histogram underlying the *V* segments.

In particular, CDR3 accounts for the majority of the physical interaction between antibodies and antigens (Wedemayer *et al*. 1997; Xu & Davis 2000), thus characterizing the amino acid variation of this region in expressed antibody repertoires provides important insight into the sifaka adaptive immune response. A growing number of studies have documented distinct lineage-specific patterns of amino acid variation and overall length of the CDR3 region in a variety of mammals [e.g. cattle (Larsen & Smith 2012), camels (Nguyen

et al. 2000), platypus (Johansson *et al.* 2002)]. However, relatively few studies have examined antibody diversity of non-human primates, with those that have focusing entirely on primates of biomedical research interest (Von Büdingen *et al.* 2001; Druar *et al.* 2005; Sundling *et al.* 2012, 2014). When comparing complete CDR3 sifaka sequences (n = 12 005) with existing primate data, our results document CDR3 properties within *P. coquereli* that are probably conserved across all primates, including CDR3 length variation and general amino acid content (Figs 3–5). The average length and the amino acid composition of the sifaka antibodies were comparable to those of humans and other primates [Figs 3–5; (Druar *et al.* 2005; Von Büdingen *et al.* 2001; Wu *et al.* 2012; Zemlin *et al.* 2003)]. Additionally, approximately 0.04% of CDR3 sequences identified by IMGT were shared between the two sifaka individuals, similar to findings noted in the antibody repertoires of humans [e.g. 0.025%; Vollmers *et al.* (2013)]. Evolutionary conservation of amino acid composition also appears to hold, with vertebrate antibody repertoires having CDR3s that tend to contain an abundance of neutral or hydrophilic amino acids (Zemlin *et al.* 2003; Schroeder 2006; Wu *et al.* 2012).

Lineage-specific characteristics of CDR3 have been hypothesized to be indicative of differing evolutionary pressures on the adaptive immune system. For example, the range of CDR3 lengths in human repertoires is greater than that of the mouse, suggesting higher potential immunoglobulin diversity in the human Ig repertoire (Zemlin *et al.* 2003). Other mammalian species, such as cattle, have excessively long CDR3 lengths which theoretically evolved to compensate for limited V_H variability in the germline (Berens *et al.* 1997; Larsen & Smith 2012). We find no evidence of such changes (i.e. remarkable CDR3 lengths) between the sifaka and the more heavily studied haplorrhine primates (human and macaque) thus establishing baseline similarity between the adaptive immune systems of the two main clades of primates. These findings are relevant not only for future lemur eco-immunogenomic research and preservation efforts, but also for lemur biomedical research (e.g. *Microcebus murinus*; Bons *et al.* 2006; Languille *et al.* 2012).

New approaches to ecological immunology research

We used two recently developed sequencing technologies to characterize the antigen-binding region of expressed antibodies within *P. coquereli*. The sequencing approaches we pursued provided flexibility in our experimental design as they allowed us to target IgG-specific antibodies of lengths that varied up to 200 bp (~350–550 bp) and eliminated the necessity for downstream sequence assembly that is difficult given the dynamic nature of antibody formation. Despite low throughput, the long-read aspect of PacBio SMRT sequencing technology is useful for antibody discovery, especially given the potential for extremely long antibody sequences. We anticipate that this approach will be useful exploratory sequencing of antibody repertoires in nonmodel species, wherein patterns of antibody formation may be poorly characterized or unknown (e.g. Larsen & Smith (2012) discovered *Bos taurus* IgG CDR3 regions up to 62 amino acids long using PacBio CCS). Recent advancements in the sequence read lengths of next-generation technologies allow for more flexibility in the choice of sequencing platform and much greater sequencing depth of antibody repertoires. Importantly, targeted approaches that utilize a combination of long-read PacBio sequencing (i.e. >8 kb) and short-read sequencing can be used to provide accurate assemblies of both heavy-chain and light-chain germline loci in nonmodel species. The assembly of these repetitive genomic regions will be an important step towards better understanding the evolution of the primate adaptive immune system. With respect to this study, the utilization of the human heavy-chain locus was useful for general comparisons and primer design, but it is likely that we have only captured the portion of the sifaka antibody repertoire that shares homology across all primates. Future work will focus on *de novo* blood transcriptome assemblies and the assembly and annotation of regions of the lemur genome that are important to immune function. Such information will allow for broader investigation of the lemur adaptive immune response.

High-throughput sequencing of expressed antibody repertories is a major advancement towards understanding the evolution of the adaptive immune system. To date, this technique has been used for only a few organisms (e.g. humans, macaques, cattle, zebrafish), and data generated from humans have provided new insights into a variety of important areas including personalized medicine, cancer and HIV research, and vaccine development (Parameswaran *et al.* 2013; Robins 2013; Vollmers *et al.* 2013; Zhu *et al.* 2013a,b). Thus, Ig-seq studies of nonmodel species may spur the development of vaccines for wild populations, an approach that holds promise from both conservation and human health perspectives (Tsao *et al.* 2004; Abbott *et al.* 2012; Fausther-Bovendo *et al.* 2012; Carne *et al.* 2013; Monath 2013; Richer *et al.* 2014). Ecological approaches to disease prevention would benefit greatly from an increased understanding of species-specific adaptive immune system variation, and advanced methodologies such as Ig-seq provide this information through the unprecedented measurement and characterization of naturally circulating antibody variation. Moreover, ecological and conservation research initiatives can benefit from Ig-seq through the creation of diagnostic

tools for wild populations. Parameswaran *et al.* (2013) used Ig-seq to identify convergent CDR1, CDR2 and CDR3 clusters in the antibody repertoires of human subjects experiencing Dengue fever. The significance of this observation is that the development of immunoglobulin expression databases can potentially be used to rapidly screen blood samples for expression patterns indicative of the presence of known pathogens. With respect to nonmodel species, such approaches could be used for a variety of applications ranging from the monitoring of disease status of wild populations to addressing biosecurity concerns of imported/exported wildlife.

Our study represents the first attempt to characterize sequence diversity of expressed antibodies in an endangered nonmodel primate. The methods described herein therefore provide a framework for investigating the adaptive immune response in wild populations of lemurs and other endangered vertebrates. More generally, we anticipate that these methods can be adapted to a number of nonmodel species and thus advance eco-immunogenomic studies of wild populations that will aid conservation efforts and identify novel solutions for treating zoonotic pathogens of concern to human health (Koff *et al.* 2013).

Acknowledgements

We thank the Duke Lemur Center staff, especially Erin Ehmke, Cathy Williams and Bobby Schopler for logistical support of our research. The Duke IGSP Genome Sequencing and Analysis Core Resource generated the sequence data reported herein, and we thank Olivier Fredrigo, Graham Alexander and Nick Hoang for their assistance. Grants in Aid of research were generously provided to CRC by Sigma Xi and the Duke Chapter of Sigma Xi. PAL thanks the American Society of Mammalogists for financial support. This project was funded by Duke University start-up funds to ADY. This is Duke Lemur Center publication number 1267.

References

Abbott RC, Osorio JE, Bunck CM, Rocke TE (2012) Sylvatic plague vaccine: a new tool for conservation of threatened and endangered species? *EcoHealth*, **9**, 243–250.

Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc M-P (2012) IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Research*, **8**, 1–15.

Arnaout R, Lee W, Cahill P *et al.* (2011) High-resolution description of antibody heavy-chain repertoires in humans. *PLoS ONE*, **6**, e22365.

Barrett MA, Brown JL, Junge RE, Yoder AD (2013) Climate change, predictive modeling and lemur health: assessing impacts of changing climate on health and conservation in Madagascar. *Biological Conservation*, **157**, 409–422.

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

Berens SJ, Wylie DE, Lopez OJ (1997) Use of a single VH family and long CDR3s in the variable region of cattle Ig heavy chains. *International immunology*, **9**, 189–199.

Blankenberg D, Kuster GV, Coraor N *et al.* (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology*, **19**, 19.10.11–19.10.21.

Bons N, Rieger F, Prudhomme D, Fisher A, Krause KH (2006) *Microcebus murinus*: a useful primate model for human cerebral aging and Alzheimer's disease?

Briney B, Willis J, McKinney B, Crowe J (2012) High-throughput antibody sequencing reveals genetic evidence of global regulation of the naive and memory repertoires that extends across individuals. *Genes and Immunity*, **13**, 469–473.

Carne C, Semple S, Morrogh-Bernard H *et al.* (2013) Predicting the vulnerability of great apes to disease: the role of superspreaders and their potential vaccination. *PLoS ONE*, **8**, e84642.

Castro R, Jouneau L, Pham H-P *et al.* (2013) Teleost fish mount complex clonal IgM and IgT responses in spleen upon systemic viral infection. *PLoS pathogens*, **9**, e1003098.

Collis AV, Brouwer AP, Martin AC (2003) Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *Journal of molecular biology*, **325**, 337–354.

Demas GE, Nelson RJ (2011) *Ecoimmunology*. Oxford University Press, New York.

Druar C, Saini SS, Cossitt MA *et al.* (2005) Analysis of the expressed heavy chain variable-region genes of *Macaca fascicularis* and isolation of monoclonal antibodies specific for the Ebola virus' soluble glycoprotein. *Immunogenetics*, **57**, 730–738.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Ekblom R, Wennekes P, Horsburgh GJ, Burke T (2014) Characterization of the house sparrow (*Passer domesticus*) transcriptome: a resource for molecular ecology and immunogenetics. *Molecular Ecology Resources*, **14**, 636–646.

Fausther-Bovendo H, Mulangu S, Sullivan NJ (2012) Ebolavirus vaccines for humans and apes. *Current Opinion in Virology*, **2**, 324–329.

Georgiou G, Ippolito GC, Beausang J *et al.* (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology*, **32**, 158–168.

Glanville J, Zhai W, Berka J *et al.* (2009) Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences*, **106**, 20216–20221.

Hawley DM, Altizer SM (2011) Disease ecology meets ecological immunology: understanding the links between organismal immunity and winfection dynamics in natural populations. *Functional Ecology*, **25**, 48–60.

Helmuth EF, Letvin NL, Margolin DH (2000) Germline repertoire of the immunoglobulin V H 3 family in rhesus monkeys. *Immunogenetics*, **51**, 519–527.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

Janeway C, Travers P, Walport M, Shlomchik M (2004) *Immunobiology*. Garland Science, New York.

Johansson J, Aveskogh M, Munday B, Hellman L (2002) Heavy chain V region diversity in the duck-billed platypus (*Ornithorhynchus anatinus*): long and highly variable complementarity-determining region 3 compensates for limited germline diversity. *Journal of Immunology*, **168**, 5155–5162.

Koff WC, Burton DR, Johnson PR *et al.* (2013) Accelerating next-generation vaccine development for global disease prevention. *Science*, **340**, 1064–1071.

Krzywinski M, Schein J, Birol İ *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome research*, **19**, 1639–1645.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Languille S, Blanc S, Blin O *et al.* (2012) The grey mouse lemur: a non-human primate model for ageing studies. *Ageing Research Reviews*, **11**, 150–162.

Larsen P, Smith T (2012) Application of circular consensus sequencing and network analysis to characterize the bovine IgG repertoire. *BMC Immunology*, **13**, 52.

Lefranc M-P, Giudicelli V, Ginestoux C *et al.* (2009) IMGT®, the international ImMunoGeneTics information system®. *Nucleic acids research*, **37**, D1006–D1012.

Link JM, Hellinger MA, Schroeder HW (2002) The Rhesus monkey immunoglobulin IGHD and IGHJ germline repertoire. *Immunogenetics*, **54**, 240–250.

Matsuda F, Ishii K, Bourvagnet P *et al.* (1998) The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *The Journal of experimental medicine*, **188**, 2151–2162.

Meek K, Eversole T, Capra J (1991) Conservation of the most JH proximal Ig VH gene segment (VHVI) throughout primate evolution. *The Journal of immunology*, **146**, 2434–2438.

Mittermeier R, Louis E, Richardson M *et al.* (2010) *Lemurs of Madagascar*, 3rd edn. Tropical field guide series. Conservation International, Arlington, VA.

Monath TP (2013) Vaccines against diseases transmitted from animals to humans: a one health paradigm. *Vaccine*, **31**, 5321–5338.

Nguyen VK, Hamers R, Wyns L, Muyldermans S (2000) Camel heavy-chain antibodies: diverse germline V_HH and specific mechanisms enlarge the antigen-binding repertoire. *EMBO Journal*, **19**, 921–930.

Parameswaran P, Liu Y, Roskin KM *et al.* (2013) Convergent antibody signatures in human dengue. *Cell Host & Microbe*, **13**, 691–700.

Pedersen AB, Babayan SA (2011) Wild immunology. *Molecular Ecology*, **20**, 872–880.

Perry GH, Melsted P, Marioni JC *et al.* (2012) Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome research*, **22**, 602–610.

Pipes L, Li S, Bozinoski M *et al.* (2013) The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. *Nucleic acids research*, **41**, D906–D914.

Racanelli V, Brunetti C, De Re V *et al.* (2011) Antibody Vh repertoire differences between resolving and chronically evolving hepatitis C virus infections. *PLoS ONE*, **6**, e25606.

Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. *Salmonidae*). *Molecular Ecology*, **19**, 115–131.

Richer LM, Brisson D, Melo R, Ostfeld RS, Zeidner N, Gomes-Solecki M (2014) Reservoir targeted vaccine against *Borrelia burgdorferi*: a new strategy to prevent lyme disease transmission. *Journal of Infectious Diseases*. doi:10.1093/infdis/jiu005 (in press).

Robins H (2013) Immunosequencing: applications of immune repertoire deep sequencing. *Current Opinion in Immunology*, **25**, 646–652.

Schroeder HWJ (2006) Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Developmental & Comparative Immunology*, **30**, 119–135.

Sundling C, Li Y, Huynh N *et al.* (2012) High-resolution definition of vaccine-elicited B cell responses against the HIV primary receptor binding site. *Science Translational Medicine*, **4**, 1–12.

Sundling C, Zhang Z, Phad GE *et al.* (2014) Single-Cell and deep sequencing of IgG-switched macaque B cells reveal a diverse Ig repertoire following immunization. *The Journal of Immunology*, **192**, 367–3644.

Tsao JI, Wootton JT, Bunikis J *et al.* (2004) An ecological approach to preventing human infection: vaccinating wild mouse reservoirs intervenes in the Lyme disease cycle. *Proceedings of the National Academy of Sciences*, **101**, 18159–18164.

Tung J, Barreiro LB, Johnson ZP *et al.* (2012) Social environment is associated with gene regulatory variation in the rhesus macaque immune system. *Proceedings of the National Academy of Sciences*, **109**, 6490–6495.

Tutter A, Riblet R (1989) Conservation of an immunoglobulin variable-region gene family indicates a specific, noncoding function. *Proceedings of the National Academy of Sciences*, **86**, 7460–7464.

Vera JC, Wheat CW, Fescemyer HW *et al.* (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR (2013) Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences*, **110**, 13463–13468.

Von Büdingen HC, Hauser SL, Nabavi CB, Genain CP (2001) Characterization of the expressed immunoglobulin IGHV repertoire in the New World marmoset *Callithrix jacchus*. *Immunogenetics*, **53**, 557–563.

Wedemayer GJ, Patten PA, Wang LH, Schultz PG, Stevens RC (1997) Structural insights into the evolution of an antibody combining site. *Science*, **276**, 1665–1669.

Wu L, Oficjalska K, Lambert M *et al.* (2012) Fundamental characteristics of the immunoglobulin VH repertoire of chickens in comparison with those of humans, mice, and camelids. *The Journal of immunology*, **188**, 322–333.

Xu JL, Davis MM (2000) Diversity in the CDR3 region of V_H is sufficient for most antibody specificities. *Immunity*, **13**, 37–45.

Yoder AD, Yang Z (2004) Divergence dates for Malagasy lemurs estimated from multiple gene loci: geological and evolutionary context. *Molecular Ecology*, **13**, 757–773.

Zemlin M, Klinger M, Link J *et al.* (2003) Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *Journal of molecular biology*, **334**, 733–749.

Zhu J, Ofek G, Yang Y *et al.* (2013a) Mining the antibodyome for HIV-1–neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proceedings of the National Academy of Sciences*, **110**, 6470–6475.

Zhu J, Wu X, Zhang B *et al.* (2013b) De novo identification of VRC01 class HIV-1–neutralizing antibodies by next-generation sequencing of B-cell transcripts. *Proceedings of the National Academy of Sciences*, **110**, E4088–E4097.

P.A.L. conceived the study. P.A.L. and C.R.C. designed the experiments and analyzed the data. All authors contributed to writing the manuscript.

Data Accessibility

All raw sequence data associated with this manuscript can be found on the NCBI Sequence Read Archive under project number PRJNA239248. Quality filtered fastq files for both PacBio SMRT CCS and Ion Torrent PGM data are available on the DRYAD digital repository (doi:10.5061/dryad.8167k).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1 Sequence length of quality filtered IgG PacBio sequence data (n = 15 282).

Figure S2 IMGT human *V* segment classification of 13 073 expressed IgG sequences from the sifaka (PacBio CCS data).

Figure S3 Frequencies of *J* segments within expressed sifaka IgG antibodies (from 11 810 IMGT categorized sequences, generated using PacBio CCS) based on their closest match with human *J* segments 1-6.

Figure S4 Sequence length of quality filtered global IgH Ion Torrent sequence data (n = 144 441).

Figure S5 Circos diagram showing common IMGT identification of expressed sifaka IgG antibodies (from 131 829 Ion Torrent PGM sequences) based on their similarity to human *V* and *J* segments.

Table S1 Raw reads, filtered reads, and reads mapped to the human heavy-chain locus (see Methods) per species examined by Perry *et al.* (2012)

Table S2 Results of mapping Perry *et al.* (2012) transcriptome data to the human heavy-chain locus (a) raw reads. (b) total bases

Table S3 Primers used for successful amplification of IgG and global IgH within Propithecus coquereli

Table S4 Bowtie2 mapping results of Perry *et al.* (2012) transcriptome data to human heavy-chain V segments