# Successful exome capture and sequencing in lemurs using human baits.

Timothy H. Webster[1*], Elaine E. Guevara[2,3], Richard R. Lawler[4], Brenda J. Bradley[2]
[1]School of Life Sciences, Arizona State University, Tempe, AZ 85287
[2]Center for the Advanced Study of Human Paleobiology, The George Washington University, Washington, DC 20052
[3]Department of Anthropology, Yale University, New Haven, CT 06511
[4]Department of Sociology and Anthropology, James Madison University, Harrisonburg, VA 22807

Figures: 4
Tables: 1
Abbreviated title: Exome capture in strepsirrhines

[*]Correspondence to:
Timothy H. Webster
Arizona State University
School of Life Sciences
P.O. Box 874501
Tempe, AZ 85287
Email: timothy.h.webster@gmail.com

Grant sponsorship: This work was funded by Yale University, The George Washington University, the Nacey Maggioncalda Foundation, the Wenner-Gren Foundation, the National Science Foundation (NSF BCS-1455818), the Yale MacMillan Center, and the Yale Institute for Biospheric Studies.

## ABSTRACT

### Objectives

We assessed the efficacy of exome capture in lemurs using commercially available human baits.

### Materials and Methods

We used two human kits (Nimblegen SeqCap EZ Exome Probes v2.0; IDT xGen Exome Research Panel v1.0) to capture and sequence the exomes of wild Verreaux's sifakas (*Propithecus verreauxi,* n = 8), a lemur species distantly related to humans. For comparison, we also captured exomes of a primate species more closely related to humans (*Macaca mulatta,* n= 4). We mapped reads to both the human reference assembly and the most closely related reference for each species before calling variants. We used measures of mapping quality and read coverage to compare capture success.

### Results

We observed high and comparable mapping qualities for both species when mapped to their respective nearest-relative reference genomes. When investigating breadth of coverage, we found greater capture success in macaques than sifakas using both nearest-relative and human assemblies. Exome capture in sifakas was still highly successful with more than 90% of annotated coding sequence in the sifaka reference genome captured, and 80% sequenced to a depth greater than 7x using Nimblegen baits. However, this success depended on probe design: the use of IDT probes resulted in substantially less callable sequence at low-to-moderate depths.

### Discussion

Overall, we demonstrate successful exome capture in lemurs using human baits, though success differed between kits tested. These results indicate that exome capture is an effective and economical genomic method of broad utility to evolutionary primatologists working across the entire primate order.

**KEY WORDS:** genomics, strepsirrhines, primates, macaques, methods

## Introduction

Recent advances in next generation sequencing technology and the increasing availability of annotated reference genomes have made feasible the genomic study of nonmodel taxa (Ellegren, 2014; Goodwin, McPherson, & McCombie, 2016). Nonhuman catarrhines, in particular papionin monkeys (Bergey, Phillips-Conroy, Disotell, & Jolly, 2016; Gibbs et al., 2007; Lea, Altmann, Alberts, & Tung, 2016; Wall et al., 2016) and apes (Carbone et al., 2014; de Manuel et al., 2016; Locke et al., 2011; Perry et al., 2008; Prado-Martinez et al., 2013), have been the focus of intense genomic study because of their importance in understanding human evolutionary history (Jolly, 2001; Swedell & Plummer, 2012; Wrangham, 1987) and history of use as biomedical models (Carlsson, Schapiro, Farah, & Hau, 2004; Rogers & Gibbs, 2014; Varki, 2000). However, genomic data hold promise to enable vast insights into evolution, ecology, and behavior, as well as inform conservation management across the entire primate order.

Nevertheless, genomic analyses remain out-of-reach for many species. Even for species for which there is a draft genome available, population-scale whole genome sequencing and the concomitant data storage, management, and analyses often require prohibitively vast financial, computational, and bioinformatics resources. These conditions have fostered the development and wide adoption of reduced representation genomic sequencing methods, like restriction-associated DNA sequencing (RAD-seq; K. R. Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Baird et al., 2008). While RAD-seq and similar “genotyping-by-sequencing” methods have enabled the genomic study of a variety of nonmodel organisms, aspects of the data—particularly marker

78 sparseness and discontinuity—can be limiting for some research questions (Arnold,
79 Corbett-Detig, Hartl, & Bomblies, 2013; Lowry et al., 2017; Rubin, Ree, & Moreau,
80 2012).

81 In contrast, targeted capture involves the selective enrichment of genomic
82 regions before sequencing, allowing both for more continuous sequence and for control
83 over the density and identity of targets (Gnirke et al., 2009; Jones & Good, 2016).
84 Foremost among targeted capture techniques is exome capture and sequencing
85 (exome sequencing), which primarily targets all the protein coding regions of the
86 genome along with a number of untranslated regions, promoter regions, and miRNAs
87 (Clark et al., 2011). In total, these targets account for less than 2% of the genome,
88 making exome sequencing much more cost-effective than whole genome sequencing,
89 while still providing the majority of data often desired by those undertaking high
90 throughput sequencing. Numerous commercial exome capture kits based on the human
91 genome have been developed and widely adopted in clinical settings and for identifying
92 the underlying basis of human genetic disorders (Bamshad et al., 2011; Bilgüvar et al.,
93 2010; Ng et al., 2010).

94 Synthesizing custom high-quality oligonucleotide baits for targeted capture is
95 expensive and generally requires a high-quality reference genome (Jones & Good,
96 2016; but see Snyder-Mackler et al., 2016). Because of the close evolutionary, and thus
97 genetic, relationship between human and nonhuman primates, researchers studying
98 nonhuman primates are advantageously situated to potentially exploit the baits and
99 resources developed for human exome sequencing. In particular, human exome baits
100 have been successfully used in haplorrhine primates (Bataillon et al., 2015; George et

al., 2011; Hvilsom et al., 2012; Jin et al., 2012; Teixeira et al., 2015; Vallender, 2011). However, it is currently unclear how well human exome baits would work for more distantly related species (e.g., strepsirrhine primates).

To ascertain and quantify the utility of exome sequencing across the order Primates, we performed exome capture and sequencing of a distantly related strepsirrhine species, Verreaux's sifaka (*Propithecus verreauxi*), that diverged from humans over 60 million years ago (dos Reis et al., 2018). As a direct comparison to provide context for assessing the strepsirrhine results we also included rhesus macaques (*Macaca mulatta*), a catarrhine species for which the efficacy of exome capture using baits designed for humans has already been established (George et al., 2011; Vallender, 2011). Both species have closely-related reference genomes available (*P. coquereli, M. mulatta*). Our overall goal is to assess capture efficiency, mapping success, and variant calling using two commercially available human exome capture kits.

## MATERIALS AND METHODS

### *Samples*

We collected the Verreaux's sifaka samples from individuals living at Bezà Mahafaly Special Reserve (Bezà), located in southwestern Madagascar (Toliara province). As part of long-term research, research team members capture unmarked yearlings and recent immigrants annually to collect biometric data and give each individual a unique identifying collar and ear notch pattern (Richard, Dewar, Schwartz, & Ratsirarson, 2002).

For this study, we generated two different Verreaux's sifaka datasets (Sifaka1 and Sifaka2). Sifaka1 is the primary dataset we use throughout the study in comparison with the macaque samples (Macaque1). We generated the Sifaka2 dataset using different exome capture kit to explore any effects of bait design on capture success (described below). For Sifaka1, we extracted DNA from banked ear tissue biopsies as described in Lawler et al. (2001) from four sifakas: a mother-daughter pair and two unrelated males (Supporting Information Table S1). For Sifaka2, we extracted DNA from the ear tissue of two additional male and two additional female sifakas using the QIAgen DNeasy Blood and Tissue (Qiagen) kit following manufacturer instructions with an extended lysis step (Supporting Information Table S1).

For a catarrhine comparison, we used DNA derived from blood samples from four unrelated—two male and two female—captive Indian rhesus macaques (Macaque1) from the Wisconsin National Primate Research Center (Supporting Information Table S1).

*DNA extraction, library preparation, and sequencing*

We sent extracted DNA to the Yale Center for Genome Analysis (YCGA) for exome capture, library preparation, and multiplexed sequencing following their standard protocols, described as follows. For all three datasets (Sifaka1, Sifaka2, and Macaque1), genomic DNA was sheared to a mean fragment length of 140 bp and adapters were ligated onto both ends of fragments. Fragments were then PCR amplified, during which a 6 bp barcode was inserted at one end of each fragment. Libraries were hybridized with baits from two different kits: Nimblegen baits (Nimblegen

147 SeqCap EZ Exome version 2) were used for Sifaka1 and Macaque1, and IDT xGen
148 baits (IDT xGen Exome Research Panel 1.0) were used for Sifaka2. Fragments were
149 then mixed with streptavidin-coated beads and washed to remove unbound fragments.
150 Captured fragments were then PCR amplified and purified with AMPure XP beads.
151 Libraries from Sifaka1 and Macaque1 were multiplexed (all four sifaka samples in one
152 lane, and the four macaques in another lane with two other samples) and sequenced
153 using 75 bp paired-end reads on a single lane of an Illumina HiSeq 2000 using Illumina
154 protocols. Sifaka2 libraries were sequenced using 100 bp paired-end reads on a single
155 Illumina HiSeq 4000 lane and multiplexed with eight other samples (12 total samples
156 per lane, but only four are included in this study).
157
158 *Exome assembly*
159 We assessed read quality pre- and post-trimming using FastQC (S. Andrews,
160 2018) and MultiQC (Ewels, Magnusson, Lundin, & Käller, 2016). We used BBDuk
161 (Bushnell, 2018) to remove adapters and perform quality trimming using the parameters
162 “ktrim=r k=21 mink=11 hdist=2 tbo tpe qtrim=rl trimq=10”. We then mapped reads from
163 sifaka samples (Sifaka1 and Sifaka2) to the *Propithecus coquereli* draft genome
164 (Pcoq_1.0; Baylor College of Medicine; https://www.ncbi.nlm.nih.gov/genome/24390).
165 *P. verreauxi* and *P. coquereli* share a common ancestor 3-8 million years ago (Herrera
166 & Dávalos, 2016; Springer et al., 2012). We mapped macaque samples (Macaque1) to
167 the Indian rhesus macaque draft genome (Mmul_8.0.1; *Macaca mulatta* Genome
168 Sequencing Consortium;
169 https://www.ncbi.nlm.nih.gov/genome/215?genome_assembly_id=259055). Finally, we

mapped reads from both species to the human reference genome (hg38; Genome Reference Consortium, Dec 2013). For the rest of the manuscript, we refer to Mmul_8.0.1 as mmul8, proCoq_1.0 as pcoq1, and hg38 as hg38. In all cases, we mapped reads using BWA MEM (Li, 2013) using default parameters except for "-t 4" and "-R" to add read group information. We marked duplicates with SAMBLASTER (Faust & Hall, 2014). We then used SAMtools (Li et al., 2009) to fix read pairing, and sort and index BAM files.

To enable a direct comparison of exome capture success between species for which we had different numbers of raw reads and different duplication rates, we conducted all downstream analyses on downsampled BAM files (containing the same number of reads for each individual). To downsample BAM files, we first used the "stats" tool in SAMtools (Li et al., 2009) to count the total number of reads and number of duplicate reads in each BAM file. We then used the "view" tool in SAMtools (Li et al., 2009) with the parameters "-F 1024 –s 0.<subsample_fraction>" to subsample approximately 50 million reads, where <subsample_fraction> is equal to 50 million divided by the total number of nonduplicate reads. The flag "-F 1024" removes reads flagged as duplicate.

*Variant calling*

We jointly called variants for each dataset using both GATK's HaplotypeCaller (Poplin et al., 2018) and Freebayes (Garrison & Marth, 2012). To speed up processing, we input BED files containing minimally callable sites—depth greater than 3, mapping quality greater than 19, and base quality greater than 29—generated using CallableLoci

193 in GATK (McKenna et al., 2010). Finally, we filtered variants for site quality (minimum of
194 30), sample depth (minimum of 8), sample genotype quality (minimum of 30), allele
195 support (minimum of 3 reads), and number of passing samples (minimum of 4) with a
196 Python script built using the cyvcf2 library (Pedersen & Quinlan, 2017).

197 We functionally annotated filtered variants using Ensembl's Variant Effect
198 Predictor (McLaren et al., 2016) tool with annotations derived from the NCBI gene
199 format files corresponding to the respective reference genomes for the rhesus and
200 sifaka references (NCBI *Macaca mulatta* Annotation Release 102 [GCF_000772875.2]
201 and *Propithecus coquereli* Annotation Release 100 [GCA_000956105.1]), and
202 Ensembl's cache for the human reference (hg38). We also obtained NCBI's annotation
203 for the human reference (GCF_00001405.37) for use in our coverage analyses (see
204 below). Using these annotation files, we intersected various regions (exon, intron, and
205 intergenic) with filtered variants using bedtools "intersect" (Quinlan & Hall, 2010) and
206 then used the "stats" module of BCFtools (Li, 2011) to tally variants in each region.
207
208 *Coverage analysis*

209 We calculated the mean and standard deviation of mapping quality (MAPQ) of
210 reads within each BAM file using a custom program written in Go (*"mapqs.go"*) using
211 packages in bíogo/hts (Kortschak, Pedersen, & Adelson, 2017). BWA MEM's (Li, 2013)
212 MAPQ scores are PHRED-scaled and can range from 0-60, with higher values
213 indicating increased confidence in mapping accuracy.

214 We counted the number of callable sites across a variety of depths and genomic
215 regions by first using SAMtools (Li et al., 2009) "view" to remove duplicates and reads

with a mapping quality less than 20 with the flags "-F 1024 -q 20", and then calculating per site depths with *genomecov* in bedtools (Quinlan & Hall, 2010), outputting in bedgraph format ("-bg"). We then processed bed files, including intersecting with genomic regions derived from the NCBI annotation described above using bedtools (Quinlan & Hall, 2010), BEDOPS (Neph et al., 2012), and a custom Python script ("*Compute_histogram_from_bed.py"*). Finally, we used the *coverage* module in bedtools with default parameters (Quinlan & Hall, 2010) to calculate the fraction of each coding region with coverage. Note that for all region-based analyses, we merged regions in the NCBI GFF annotations during processing because many, but not all, regions were present multiple times.

*Exome capture kit comparison*

We used the sifaka datasets (Sifaka1 and Sifaka2) for a direct comparison of capture success using the two different capture kits (NimbleGen SeqCap EZ Exome version 2 for Sifaka1 and IDT xGen Exome Research Panel 1.0 for Sifaka2). We ran both datasets through identical exome assembly and coverage analysis steps as described above.

*Data Availability*

We deposited raw sequencing reads in NCBI's Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under BioProject PRJNA417716. We provide SRA accession numbers in Supporting Information Table S1.

We built all analyses into a reproducible pipeline using Snakemake (Köster & Rahmann, 2012), Bioconda (Grüning et al., 2018). The entire pipeline—including all scripts, environment files, and software versions—is available on Github (https://github.com/thw17/Sifaka_assembly).

*Ethics Statement*

We report no conflict of interest. All research conformed to institutional and national guidelines, and complied with the American Association of Physical Anthropologists Code of Ethics. This protocol is approved by the James Madison University Institutional Animal Care and Use Committee (protocol numbers A03-14 and A18-04) and permission to conduct research at Bezà was granted by the Malagasy Ministry of the Environment.

## 3. Results

We generated the following mean numbers of raw sequencing reads per sample: 96,358,883 for Sifaka1 (range 80,434,390–108,538,766), 65,460,924 for Macaque1 (range 59,422,700–70,596,638), and 75,441,014 for Sifaka2 (range 70,264,610–79,520,110) (Supporting Information Table S2). After trimming, duplicate removal, and quality control, 83-86% of Sifaka1 reads, 89-92% of Macaque1 reads, and 68-70% of Sifaka2 reads passed all filters, with differences among datasets largely driven by duplication rates (Supporting Information Table S2). To account for these differences in duplication rates and raw sequences generated, we downsampled reads for all samples

to approximately 50 million nonduplicate reads (Supporting Information Table S2). We only included the downsampled datasets in downstream analyses.

Mapping qualities were very similar when mapping samples to their most closely related reference genomes (pcoq1 for sifakas; mmul8 for macaques). Across datasets, we observed mean mapping qualities of approximately 56 (out of a maximum of 60), with standard deviations ranging between 11 and 14 (Figure 1). However, when mapping to the human reference genome (hg38), mapping qualities decreased substantially—dropping to approximately 52 in Macaque1, 45 in Sifaka1, and 48 in Sifaka2—and the standard deviation increased (Figure 1).

We measured the number of sites in coding (CDS), intergenic, intronic, and untranslated (UTR) regions at four different depth thresholds (1x, 4x, 8x, and 12x), counting only nonduplicate reads with a minimum mapping quality of 20, which we term "callable sites." Across all regions and in both datasets (Sifaka1 and Macaque1), we observed a decrease in the number of callable sites as we increased minimum depth of coverage (Figure 2). This decrease was minor for CDS and UTR, while intronic and intergenic regions exhibited a disproportionate drop moving from 1x to 4x thresholds (Figure 2). We observed taxon differences as well. Specifically, the Macaque1 samples exhibited more callable sites in each region than those in Sifaka1 for all reference genomes. Moreover, we found little difference between callable sites in mmul8 and hg38 for each region in Macaque1, in contrast to Sifaka1, for which we observed a decrease in callable sites across regions when moving from pcoq1 to hg38 (Figure 2).

Because the primary goal of exome sequencing is to target coding sequence, we explored CDS in more detail (Figure 3; Figure 4). For both Sifaka1 and Macaque 1,

283 when mapping to the most closely related reference genome (pcoq1 and mmul8,
284 respectively), we found that more than 90% of annotated CDS had one or more reads
285 mapped to it (Figure 3; Sifaka1 mean = 90.9%, Macaque1 mean = 92.8%). However, as
286 the minimum depth threshold increased, we observed a steeper decline in Sifaka1 than
287 Macaque 1 until approximately 20x coverage. For example, Sifaka1 had means of
288 84.1% (4x), 78.7% (8x), 74.0% (12x), 69.9% (16x) and 66.1% (20x) of CDS covered at
289 increasing thresholds, while Macaque1 had broader coverage at each threshold: 89.1%
290 (4x), 85.1% (8x), 80.7% (12x), 75.8% (16x), and 70.6% (20x) of CDS covered (Figure
291 3). This pattern was far more pronounced when the two datasets were mapped to hg38.
292 Across the same depth thresholds, Sifaka1 had approximately 10-14% fewer bases
293 covered when mapping to pcoq1 to hg38 (Figure 3), while Macaque1 only exhibited a 3-
294 5% decrease per threshold moving from mmul8 to hg38 (Figure 3).

295 We also tested to see if exome capture success in strepsirrhines was consistent
296 across two commonly used commercially available human kits: NimbleGen (Sifaka1)
297 and IDT (Sifaka2). At lower minimum depth thresholds typically used in genomic
298 analyses (e.g., 8x and 12x), the NimbleGen kit recovered more than 20% more CDS in
299 pcoq1 and 15% more CDS in hg38 than IDT (Figure 4). This difference was significant
300 across depths less than 50x ($U= 31378$, $p < 2.2 \times 10^{-16}$). Interestingly, because
301 NimbleGen and IDT exhibit different slopes, they intersect at approximately 50x
302 coverage (Figure 4). While NimbleGen probes still recover significantly more CDS at
303 depth thresholds between 50x and 100x ($U=38416$, $p < 2.2 \times 10^{-16}$), the proportion of
304 bases with *X* or more coverage exhibits the opposite pattern in this interval, with IDT

displaying higher values (Figure 4). This pattern is consistent with IDT capturing less sequence, but at greater depths (i.e., depths greater than 100x).

To further explore the difference in capture success between NimbleGen and IDT, we calculated the breadth of coverage across coding regions in pcoq1. NimbleGen probes captured a significantly greater mean fraction of coding regions, measured as the mean fraction of each coding sequence covered by at least one read (185,162 regions; NimbleGen mean = 0.91, IDT mean = 0.63; $U=1.89 \times 10^{11}$, $p < 2.2 \times 10^{-16}$). Upon closer examination, this difference was primarily driven by IDT completely missing more coding regions. Among 185,162 coding regions in pcoq1, 33.8% of regions lacked coverage in the IDT data (range 33.6-34%), while only 7.1% completely lacked coverage in the NimbleGen dataset (range = 6.6-7.3%). When we excluded these regions with no coverage, the difference in mean fraction of coding regions captured decreased substantially, though NimbleGen still captured significantly more (NimbleGen mean = 0.98, IDT mean = 0.95; $U=1.57 \times 10^{11}$, $p < 2.2 \times 10^{-16}$).

We used two variant callers, GATK's HaplotypeCaller and Freebayes, to genotype Sifaka1 and Macaque1 when mapped to hg38 and the closest reference (pcoq1 for Sifaka1, and mmul8 for Macaque1), for a total of eight sets of variant calls (Supporting Information Table 3). In both datasets, variant call sets for the most closely related genome were broadly similar between HaplotypeCaller and Freebayes in terms of number of variants identified and genes overlapped (Supporting Information Table 3). However, when mapped to hg38, the datasets showed opposite patterns: for Sifaka1, HaplotypeCaller identified approximately four times as many variants as Freebayes (HaplotypeCaller = 250,389, Freebayes = 62,793), while in Macaque1 Freebayes

identified more than 56% more variants (HaplotypeCaller = 97,709, Freebayes = 152,768; Supporting Information Table 3).

While the number of variants identified across call sets differed substantially, within call sets, proportions of variant types were broadly similar (Supporting Information Table 3). Most variants identified across call sets were single nucleotide variants (SNVs; 71.6-81.2%), though proportions of multiple nucleotide variants (MNVs), insertions, and deletions increased when mapping to hg38. Similarly, the relative numbers of nonsynonymous, frameshift, and stop gained variants in exons were much higher when mapping to hg38 (Table 1). Variants were not limited to exons however, as most variants were intronic (Supporting Information Table 3).

## 4. Discussion

In this study we demonstrate, for the first time, that human baits can be used to successfully capture high-coverage exomic data for strepsirrhines. While previous studies have established that human baits are effective in anthropoid primates (George et al., 2011; Jin et al., 2012; Vallender, 2011), our results extend the cross-species application of baits to lineages diverged over 60 million years ago (dos Reis et al., 2018) and indicate that human baits are likely viable options for genomic analyses across the entire primate order.

We found that a mean of 90.9% of annotated coding sequence (CDS) in the draft *P. coquereli* genome was covered by one or more reads in our *P. verreauxi* samples. As we increased the minimum depth of coverage thresholds to match common filter values (e.g., 4x, 8x, and 12x coverage), we observed a steady, curvilinear decline in breadth of

CDS coverage (Figure 3). This pattern indicates that coverage is not uniform across CDS, consistent with predictions for next-generation sequencing (Lander & Waterman, 1988), particularly those for targeted capture (Clark et al., 2011; Sims, Sudbery, Ilott, Heger, & Ponting, 2014). In particular, in targeted sequencing, there are expected position-based sampling biases that lead to greater coverage towards the middle of targets (Wendl & Barbazuk, 2005). However, despite the fact that increasing sequencing effort will increase depth nonuniformly across targets, clearly any CDS base with coverage has been successfully captured. Therefore, increasing sequencing effort—we used 50 million nonduplicate reads in this study—should increase the fraction of callable CDS at various coverage thresholds up to at least 90.9%, the amount of CDS we observed covered by at least one read in this study.

Surprisingly, the fraction of captured CDS in sifakas (90.9%) was very similar to, albeit slightly smaller than, the fraction captured in rhesus macaques (92.8%), even though macaques share a much more recent common ancestor with humans (30-35 million years; dos Reis et al., 2018). However, the macaques exhibited a slower decrease in breadth of CDS coverage at increasing minimum depth thresholds, particularly across thresholds most commonly used (Figure 3). Thus, while exome capture is certainly highly successful in sifakas, there is a decrease in efficiency compared to lineages more closely related to humans. This pattern holds both when mapping to the nearest reference genomes (pcoq1 for the sifakas and mmul8 for the macaques) and when mapping back to the human reference, and therefore appears to be driven by capture success, rather than assembly methods (e.g, mapping).

373 Similar to our results, previous studies have noted a decrease in capture
374 efficiency across increasing evolutionary distances within catarrhine primates (Jin et al.,
375 2012; Vallender, 2011). In this study, however, we found that this effect was much less
376 pronounced even though we sampled much greater evolutionary distances. This is likely
377 driven by differences in mapping strategies. Previously, assessments of capture
378 efficiency involved mapping back to the human reference genome (George et al., 2011;
379 Jin et al., 2012; Vallender, 2011). In this study, we found that mapping across large
380 evolutionary distances appears to reduce both breadth and depth of coverage (Figure
381 3), an effect likely caused by the greater number of differences between reads and the
382 reference sequence, which substantially impacts mapping quality (Figure 1). In fact, the
383 Indian rhesus macaques used in this study were much more closely related to their
384 nearest reference (same population and species) than the Verreaux's sifakas (about 6
385 million years diverged from *P. coquereli*; dos Reis et al., 2018), which might account for
386 some of the difference in our observed capture success between the two species,
387 though this requires further study. In addition, while most protein-coding genes in
388 sifakas and macaques are expected to have homologues in humans, gene content is
389 not identical across primates (Rogers & Gibbs, 2014). It is therefore possible that our
390 results were also influenced by the presence of more sifaka-specific gene content than
391 macaque-specific gene content.
392 While exome capture was successful in the sifakas, the degree of success
393 depended on the capture baits used. Specifically, while the NimbleGen probes captured
394 an average of more than 90% of pcoq1 CDS and only completely missed 6-7% of
395 coding regions, the IDT probes captured less than 70% of pcoq1 CDS and completely

missed approximately one-third of coding regions. When we excluded missed regions, the difference in coverage reduced substantially, with both baits covering more than 95% of CDS in regions with any coverage. Taken together, the difference between baits is primarily driven by IDT baits completely missing entire coding regions, rather than the failure of IDT baits to capture entire targets. Commercially available human exome capture kits differ markedly in design, with different targets, bait lengths, and bait overlap (Clark et al., 2011). Even in human samples, for which the baits are designed, these differences in bait design affect capture efficiency and the number and location of variants detected (Clark et al., 2011; Sulonen et al., 2011).

Compared to other reduced representation methods (e.g., RAD-seq), exome capture's primary strength is that it aims to capture all protein coding regions of the genome—the regions frequently of most interest from a functional standpoint. To this end, exome capture and sequencing, particularly with the NimbleGen probes, was highly successful in our samples, capturing the vast majority of CDS and leading to the identification of a rich suite of variants. However, exome capture's utility is not limited to these regions, and it can generate high-quality data in regulatory and untranslated regions (UTRs), as well as other intronic and intergenic regions (Samuels et al., 2013). In our data, we identified tens of millions of base pairs of sequence outside of coding regions (Figure 2); in fact, more variants were identified in introns than any other sequence class. Thus, exome capture across nonhuman primates holds great promise for not only recovering coding regions across the genome, but also recovering putatively neutral sequences (introns, intergenic regions, and four-fold degenerate sites) that can

418 be applied to traditional questions in molecular ecology regarding kinship, geneflow and
419 demographic history.
420

421 **Acknowledgments**

422 We thank Sibien Mahereza, Enafa Jaonarisoa, Elahavelo Efitroarane, Efitiria, Edouard
423 Ramahatratra, Alison Richard, Roshna Wunderlich, Jeannin Ranaivonasy, and Joel
424 Ratsirarson for helping with sample collection at Bezà Mahafaly; Roger Wiseman
425 (Wisconsin National Primate Research Center) for providing macaque samples; and
426 Gary Aronsen for lab assistance. We are grateful to the Nacey Maggioncalda
427 Foundation (THW); National Science Foundation (NSF BCS-1455818; THW); the Yale
428 Institute for Biospheric Studies (THW), Yale University (BJB); The George Washington
429 University (EEG, BJB); the Yale MacMillan Center (EEG), and the Wenner-Gren
430 Foundation (EEG) for generous financial support. Many analyses were conducted on
431 the Louise and Ruddle High Performance Computing Clusters at Yale University
432 (supported by NIH grants RR19895 and RR029676-01) and the Arizona State
433 University High Performance Computing Cluster.
434

435 **References**

436 Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016).
437 Harnessing the power of RADseq for ecological and evolutionary genomics.
438 *Nature Reviews Genetics*, *17*(2), 81–92. https://doi.org/10.1038/nrg.2015.28
439 Andrews, S. (2018). *FastQC: A Quality Control tool for High Throughput Sequence*
440 *Data*. Retrieved from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

441 Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq
442 underestimates diversity and introduces genealogical biases due to nonrandom
443 haplotype sampling. *Molecular Ecology*, *22*(11), 3179–3190.
444 https://doi.org/10.1111/mec.12276
445 Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., …
446 Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using
447 Sequenced RAD Markers. *PLoS ONE*, *3*(10), e3376.
448 https://doi.org/10.1371/journal.pone.0003376
449 Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A.,
450 & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene
451 discovery. *Nature Reviews Genetics*, *12*(11), 745–755.
452 https://doi.org/10.1038/nrg3031
453 Bataillon, T., Duan, J., Hvilsom, C., Jin, X., Li, Y., Skov, L., … Schierup, M. H. (2015).
454 Inference of Purifying and Positive Selection in Three Subspecies of
455 Chimpanzees (*Pan troglodytes*) from Exome Sequencing. *Genome Biology and*
456 *Evolution*, *7*(4), 1122–1132. https://doi.org/10.1093/gbe/evv058
457 Bergey, C. M., Phillips-Conroy, J. E., Disotell, T. R., & Jolly, C. J. (2016). Dopamine
458 pathway is highly diverged in primate species that differ markedly in social
459 behavior. *Proceedings of the National Academy of Sciences*, *113*(22), 6178–
460 6181. https://doi.org/10.1073/pnas.1525530113
461 Bilgüvar, K., Öztürk, A. K., Louvi, A., Kwan, K. Y., Choi, M., Tatlı, B., … Günel, M.
462 (2010). Whole-exome sequencing identifies recessive WDR62 mutations in

463 severe brain malformations. *Nature*, *467*(7312), 207–210.
464 https://doi.org/10.1038/nature09327
465 Bushnell, B. (2018). *BBTools*. Retrieved from https://sourceforge.net/projects/bbmap/
466 Carbone, L., Harris, R. A., Gnerre, S., Veeramah, K. R., Lorente-Galdos, B.,
467 Huddleston, J., … Gibbs, R. A. (2014). Gibbon genome and the fast karyotype
468 evolution of small apes. *Nature*, *513*(7517), 195–201.
469 https://doi.org/10.1038/nature13679
470 Carlsson, H.-E., Schapiro, S. J., Farah, I., & Hau, J. (2004). Use of primates in
471 research: A global overview. *American Journal of Primatology*, *63*(4), 225–237.
472 https://doi.org/10.1002/ajp.20054
473 Clark, M. J., Chen, R., Lam, H. Y. K., Karczewski, K. J., Chen, R., Euskirchen, G., …
474 Snyder, M. (2011). Performance comparison of exome DNA sequencing
475 technologies. *Nature Biotechnology*, *29*(10), 908–914.
476 https://doi.org/10.1038/nbt.1975
477 de Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J.,
478 … Marques-Bonet, T. (2016). Chimpanzee genomic diversity reveals ancient
479 admixture with bonobos. *Science*, *354*(6311), 477–481.
480 https://doi.org/10.1126/science.aag2602
481 dos Reis, M., Gunnell, G. F., Barba-Montoya, J., Wilkins, A., Yang, Z., & Yoder, A. D.
482 (2018). Using Phylogenomic Data to Explore the Effects of Relaxed Clocks and
483 Calibration Strategies on Divergence Time Estimation: Primates as a Test Case.
484 *Systematic Biology*, *67*(4), 594–615. https://doi.org/10.1093/sysbio/syy001

485 Ellegren, H. (2014). Genome sequencing and population genomics in non-model
486 organisms. *Trends in Ecology & Evolution*, *29*(1), 51–63.
487 https://doi.org/10.1016/j.tree.2013.09.008
488 Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize
489 analysis results for multiple tools and samples in a single report. *Bioinformatics*,
490 *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354
491 Faust, G. G., & Hall, I. M. (2014). SAMBLASTER: fast duplicate marking and structural
492 variant read extraction. *Bioinformatics*, *30*(17), 2503–2505.
493 https://doi.org/10.1093/bioinformatics/btu314
494 Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read
495 sequencing. *ArXiv:1207.3907 [q-Bio.GN]*. Retrieved from
496 http://arxiv.org/abs/1207.3907
497 George, R. D., McVicker, G., Diederich, R., Ng, S. B., MacKenzie, A. P., Swanson, W.
498 J., … Thomas, J. H. (2011). Trans genomic capture and sequencing of primate
499 exomes reveals new targets of positive selection. *Genome Research*, *21*(10),
500 1686–1694. https://doi.org/10.1101/gr.121327.111
501 Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M., Mardis, E. R.,
502 … Zwieg, A. S. (2007). Evolutionary and Biomedical Insights from the Rhesus
503 Macaque Genome. *Science*, *316*(5822), 222–234.
504 https://doi.org/10.1126/science.1139247
505 Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., …
506 Nusbaum, C. (2009). Solution hybrid selection with ultra-long oligonucleotides for

507 massively parallel targeted sequencing. *Nature Biotechnology*, *27*(2), 182–189.
508 https://doi.org/10.1038/nbt.1523
509 Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years
510 of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6),
511 333–351. https://doi.org/10.1038/nrg.2016.49
512 Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., …
513 Köster, J. (2018). Bioconda: sustainable and comprehensive software distribution
514 for the life sciences. *Nature Methods*, *15*(7), 475–476.
515 https://doi.org/10.1038/s41592-018-0046-7
516 Herrera, J. P., & Dávalos, L. M. (2016). Phylogeny and Divergence Times of Lemurs
517 Inferred with Recent and Ancient Fossils in the Tree. *Systematic Biology*, *65*(5),
518 772–791. https://doi.org/10.1093/sysbio/syw035
519 Hvilsom, C., Qian, Y., Bataillon, T., Li, Y., Mailund, T., Sallé, B., … Schierup, M. H.
520 (2012). Extensive X-linked adaptive evolution in central chimpanzees.
521 *Proceedings of the National Academy of Sciences*, *109*(6), 2054–2059.
522 https://doi.org/10.1073/pnas.1106877109
523 Jin, X., He, M., Ferguson, B., Meng, Y., Ouyang, L., Ren, J., … Wang, X. (2012). An
524 Effort to Use Human-Based Exome Capture Methods to Analyze Chimpanzee
525 and Macaque Exomes. *PLoS ONE*, *7*(7), e40637.
526 https://doi.org/10.1371/journal.pone.0040637
527 Jolly, C. J. (2001). A proper study for mankind: Analogies from the Papionin monkeys
528 and their implications for human evolution. *American Journal of Physical*
529 *Anthropology*, *116*(S33), 177–204. https://doi.org/10.1002/ajpa.10021

530 Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological
531 genomics. *Molecular Ecology*, *25*(1), 185–202.
532 https://doi.org/10.1111/mec.13304
533 Kortschak, R. D., Pedersen, B. S., & Adelson, D. L. (2017). bíogo/hts: high throughput
534 sequence handling for the Go language. *The Journal of Open Source Software*,
535 *2*(10), 168. https://doi.org/10.21105/joss.00168
536 Köster, J., & Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow
537 engine. *Bioinformatics*, *28*(19), 2520–2522.
538 https://doi.org/10.1093/bioinformatics/bts480
539 Lander, E. S., & Waterman, M. S. (1988). Genomic mapping by fingerprinting random
540 clones: A mathematical analysis. *Genomics*, *2*(3), 231–239.
541 https://doi.org/10.1016/0888-7543(88)90007-9
542 Lawler, R. R., Richard, A. F., & Riley, M. A. (2001). Characterization and screening of
543 microsatellite loci in a wild lemur population (Propithecus verreauxi verreauxi).
544 *American Journal of Primatology*, *55*(4), 253–259.
545 https://doi.org/10.1002/ajp.1058
546 Lea, A. J., Altmann, J., Alberts, S. C., & Tung, J. (2016). Resource base influences
547 genome-wide DNA methylation levels in wild baboons (*Papio cynocephalus*).
548 *Molecular Ecology*, *25*(8), 1681–1696. https://doi.org/10.1111/mec.13436
549 Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association
550 mapping and population genetical parameter estimation from sequencing data.
551 *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

552 Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with
553 BWA-MEM. *ArXiv*, *1303.3997*.
554 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome
555 Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format
556 and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.
557 https://doi.org/10.1093/bioinformatics/btp352
558 Locke, D. P., Hillier, L. W., Warren, W. C., Worley, K. C., Nazareth, L. V., Muzny, D. M.,
559 … Wilson, R. K. (2011). Comparative and demographic analysis of orang-utan
560 genomes. *Nature*, *469*(7331), 529–533. https://doi.org/10.1038/nature09687
561 Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., &
562 Storfer, A. (2017). Breaking RAD: an evaluation of the utility of restriction site-
563 associated DNA sequencing for genome scans of adaptation. *Molecular Ecology*
564 *Resources*, *17*(2), 142–152. https://doi.org/10.1111/1755-0998.12635
565 McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., …
566 DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework
567 for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9),
568 1297–1303. https://doi.org/10.1101/gr.107524.110
569 McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., …
570 Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*,
571 *17*(1). https://doi.org/10.1186/s13059-016-0974-4
572 Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K.,
573 … Stamatoyannopoulos, J. A. (2012). BEDOPS: high-performance genomic

574 feature operations. *Bioinformatics*, *28*(14), 1919–1920.
575 https://doi.org/10.1093/bioinformatics/bts277
576 Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., …
577 Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian
578 disorder. *Nature Genetics*, *42*(1), 30–35. https://doi.org/10.1038/ng.499
579 Pedersen, B. S., & Quinlan, A. R. (2017). cyvcf2: fast, flexible variant analysis with
580 Python. *Bioinformatics*, *33*(12), 1867–1869.
581 https://doi.org/10.1093/bioinformatics/btx057
582 Perry, G. H., Yang, F., Marques-Bonet, T., Murphy, C., Fitzgerald, T., Lee, A. S., …
583 Redon, R. (2008). Copy number variation and evolution in humans and
584 chimpanzees. *Genome Research*, *18*(11), 1698–1710.
585 https://doi.org/10.1101/gr.082016.108
586 Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der
587 Auwera, G. A., … Banks, E. (2018). Scaling accurate genetic variant discovery to
588 tens of thousands of samples. *BioRxiv*, *201178*. https://doi.org/10.1101/201178
589 Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B.,
590 … Marques-Bonet, T. (2013). Great ape genetic diversity and population history.
591 *Nature*, *499*(7459), 471–475. https://doi.org/10.1038/nature12228
592 Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing
593 genomic features. *Bioinformatics*, *26*(6), 841–842.
594 https://doi.org/10.1093/bioinformatics/btq033
595 Richard, A. F., Dewar, R. E., Schwartz, M., & Ratsirarson, J. (2002). Life in the slow
596 lane? Demography and life histories of male and female sifaka (*Propithecus*

*verreauxi verreauxi*): Demography and life histories of male and female sifaka. *Journal of Zoology*, *256*(4), 421–436. https://doi.org/10.1017/S0952836902000468

Rogers, J., & Gibbs, R. A. (2014). Comparative primate genomics: emerging patterns of genome content and dynamics. *Nature Reviews Genetics*, *15*(5), 347–359. https://doi.org/10.1038/nrg3707

Rubin, B. E. R., Ree, R. H., & Moreau, C. S. (2012). Inferring Phylogenies from RAD Sequence Data. *PLoS ONE*, *7*(4), e33394. https://doi.org/10.1371/journal.pone.0033394

Samuels, D. C., Han, L., Li, J., Quanghu, S., Clark, T. A., Shyr, Y., & Guo, Y. (2013). Finding the lost treasures in exome sequencing data. *Trends in Genetics*, *29*(10), 593–599. https://doi.org/10.1016/j.tig.2013.07.006

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, *15*(2), 121–132. https://doi.org/10.1038/nrg3642

Snyder-Mackler, N., Majoros, W. H., Yuan, M. L., Shaver, A. O., Gordon, J. B., Kopp, G. H., … Tung, J. (2016). Efficient Genome-Wide Sequencing and Low-Coverage Pedigree Analysis from Noninvasively Collected Samples. *Genetics*, *203*(2), 699–714. https://doi.org/10.1534/genetics.116.187492

Springer, M. S., Meredith, R. W., Gatesy, J., Emerling, C. A., Park, J., Rabosky, D. L., … Murphy, W. J. (2012). Macroevolutionary Dynamics and Historical Biogeography of Primate Diversification Inferred from a Species Supermatrix. *PLoS ONE*, *7*(11), e49521. https://doi.org/10.1371/journal.pone.0049521

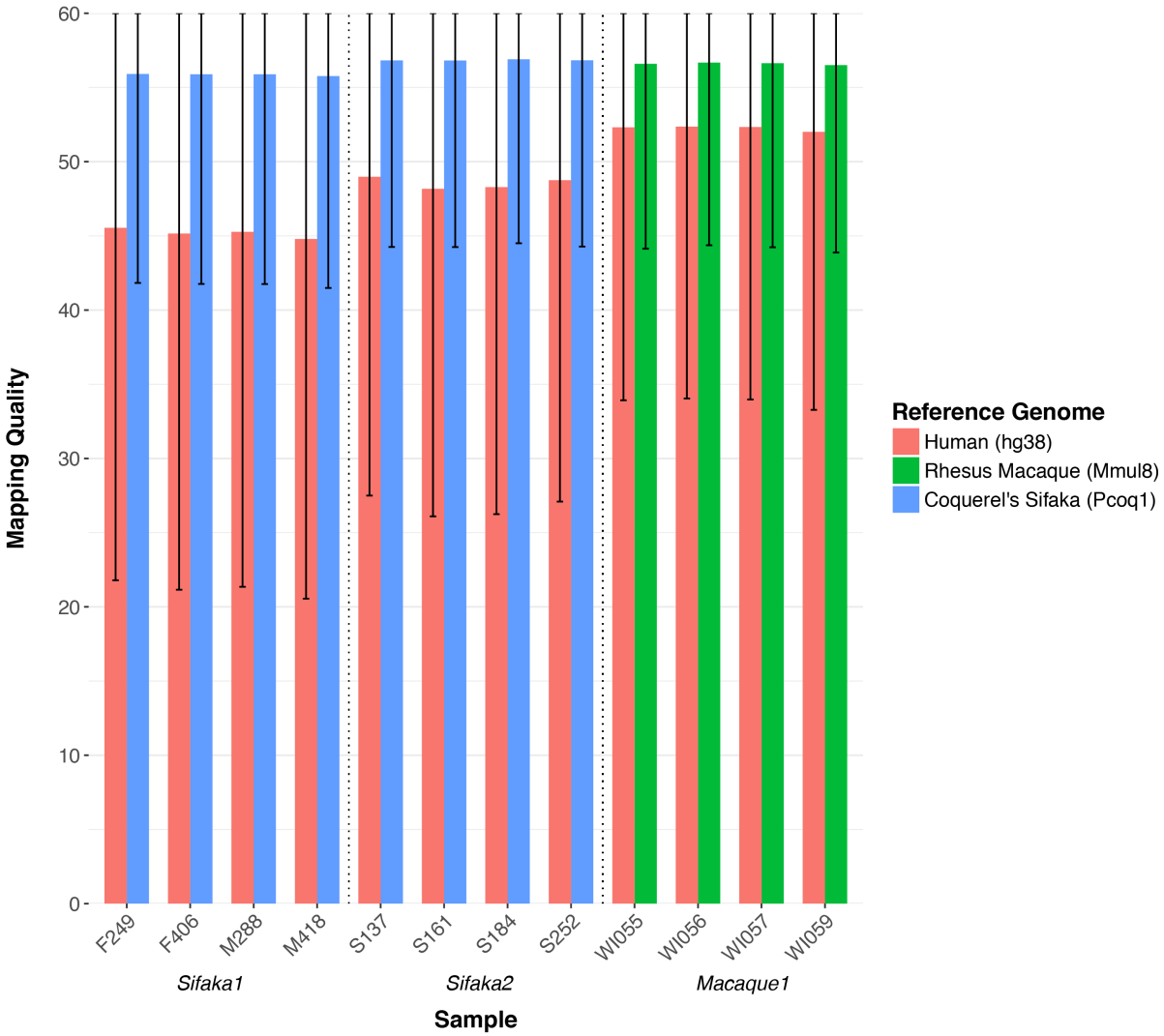
620 Sulonen, A.-M., Ellonen, P., Almusa, H., Lepistö, M., Eldfors, S., Hannula, S., …
621 Saarela, J. (2011). Comparison of solution-based exome capture methods for
622 next generation sequencing. *Genome Biology*, *12*(9), R94.
623 https://doi.org/10.1186/gb-2011-12-9-r94
624 Swedell, L., & Plummer, T. (2012). A Papionin Multilevel Society as a Model for Hominin
625 Social Evolution. *International Journal of Primatology*, *33*(5), 1165–1193.
626 https://doi.org/10.1007/s10764-012-9600-9
627 Teixeira, J. C., de Filippo, C., Weihmann, A., Meneu, J. R., Racimo, F., Dannemann,
628 M., … Andrés, A. M. (2015). Long-Term Balancing Selection in LAD1 Maintains a
629 Missense Trans-Species Polymorphism in Humans, Chimpanzees, and
630 Bonobos. *Molecular Biology and Evolution*, *32*(5), 1186–1196.
631 https://doi.org/10.1093/molbev/msv007
632 Vallender, E. J. (2011). Expanding whole exome resequencing into non-human
633 primates. *Genome Biology*, *12*(9), R87. https://doi.org/10.1186/gb-2011-12-9-r87
634 Varki, A. (2000). A Chimpanzee Genome Project Is a Biomedical Imperative. *Genome*
635 *Research*, *10*(8), 1065–1070. https://doi.org/10.1101/gr.10.8.1065
636 Wall, J. D., Schlebusch, S. A., Alberts, S. C., Cox, L. A., Snyder-Mackler, N., Nevonen,
637 K. A., … Tung, J. (2016). Genomewide ancestry and divergence patterns from
638 low-coverage sequencing data reveal a complex history of admixture in wild
639 baboons. *Molecular Ecology*, *25*(14), 3469–3483.
640 https://doi.org/10.1111/mec.13684

641 Wendl, M. C., & Barbazuk, W. B. (2005). Extension of Lander-Waterman theory for
642 sequencing filtered DNA libraries. *BMC Bioinformatics*, *6*, 245.
643 https://doi.org/10.1186/1471-2105-6-245
644 Wrangham, R. W. (1987). The significance of African apes for reconstructing human
645 social evolution. In W. G. Kinzey (Ed.), *The evolution of human behavior: primate*
646 *models* (pp. 51–71). Albany, NY: SUNY Press.

647
648
649

650

# Figures




651

652 Figure 1. Mean mapping quality (MAPQ) for samples mapped to their most closely
653 related reference genome (pcoq1 for sifakas and mmul8 for macaques) and the human
654 reference genome (hg38). Samples are organized by dataset membership, defined by
655 species and capture kit. Sifaka1 and Macaque1 were processed using NimbleGen baits
656 and Sifaka2 was processed using IDT baits. Error bars denote plus/minus one standard
657 deviation. Note that maximum mapping quality is 60.

658

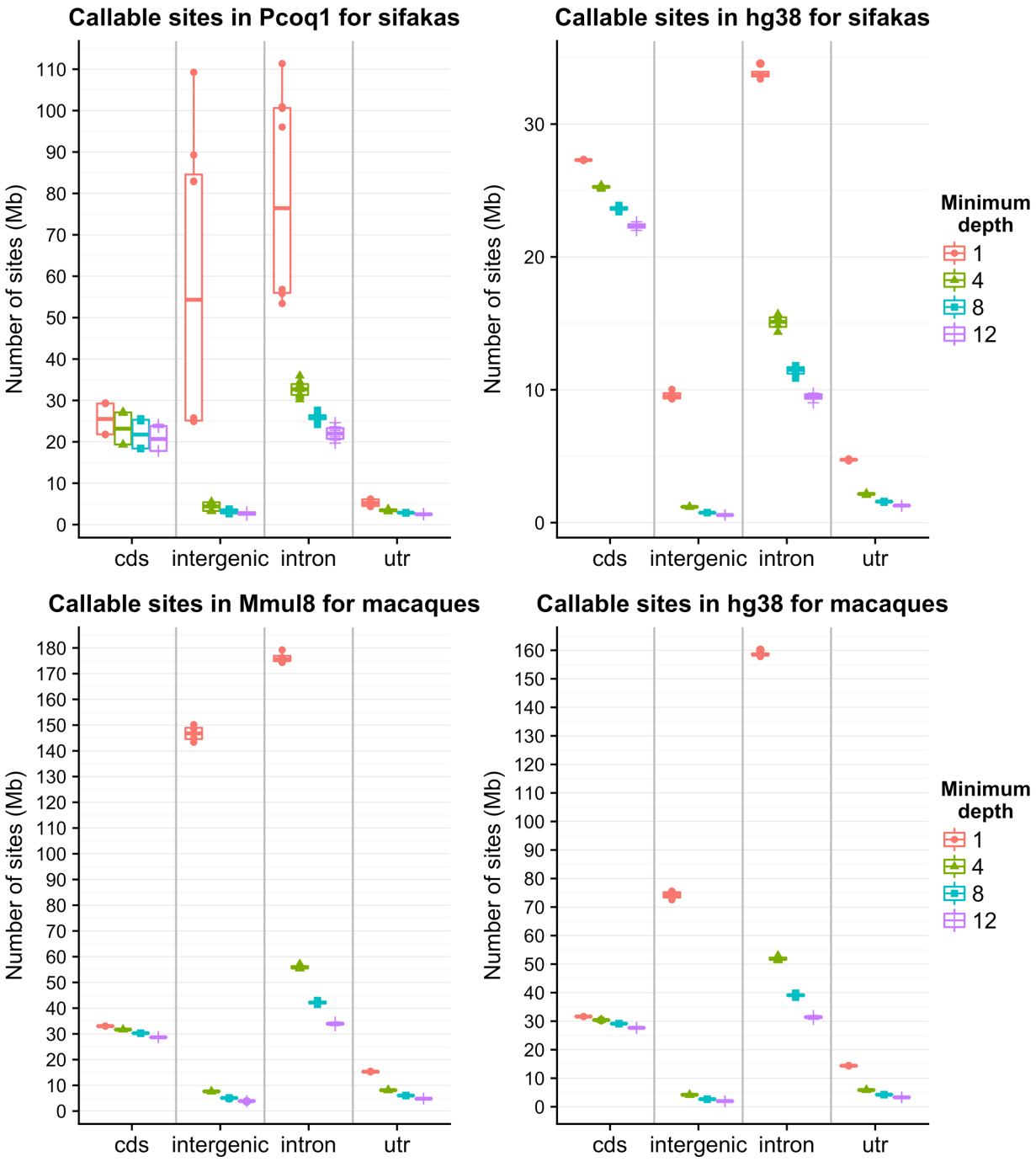



Figure 2. The effect of minimum depth on the number of callable sites across genomic regions. Samples are mapped to their most closely related reference genome (pcoq1 for sifakas and mmul8 for macaques) and the human reference genome (hg38). Minimum

663 depth thresholds were 1, 4, 8, and 12 nonduplicate reads per site with MAPQ greater

664 than or equal to 20.

665

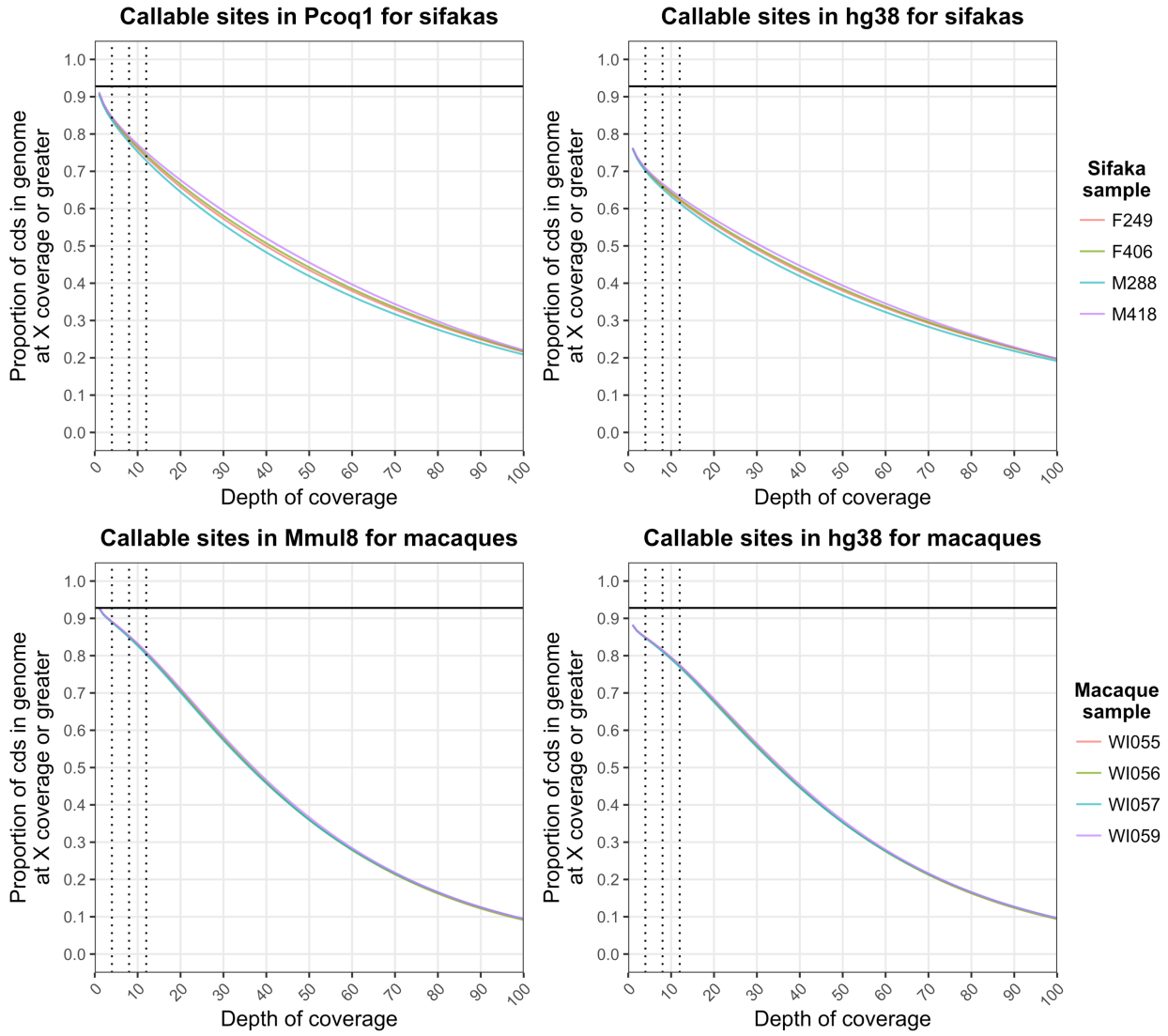


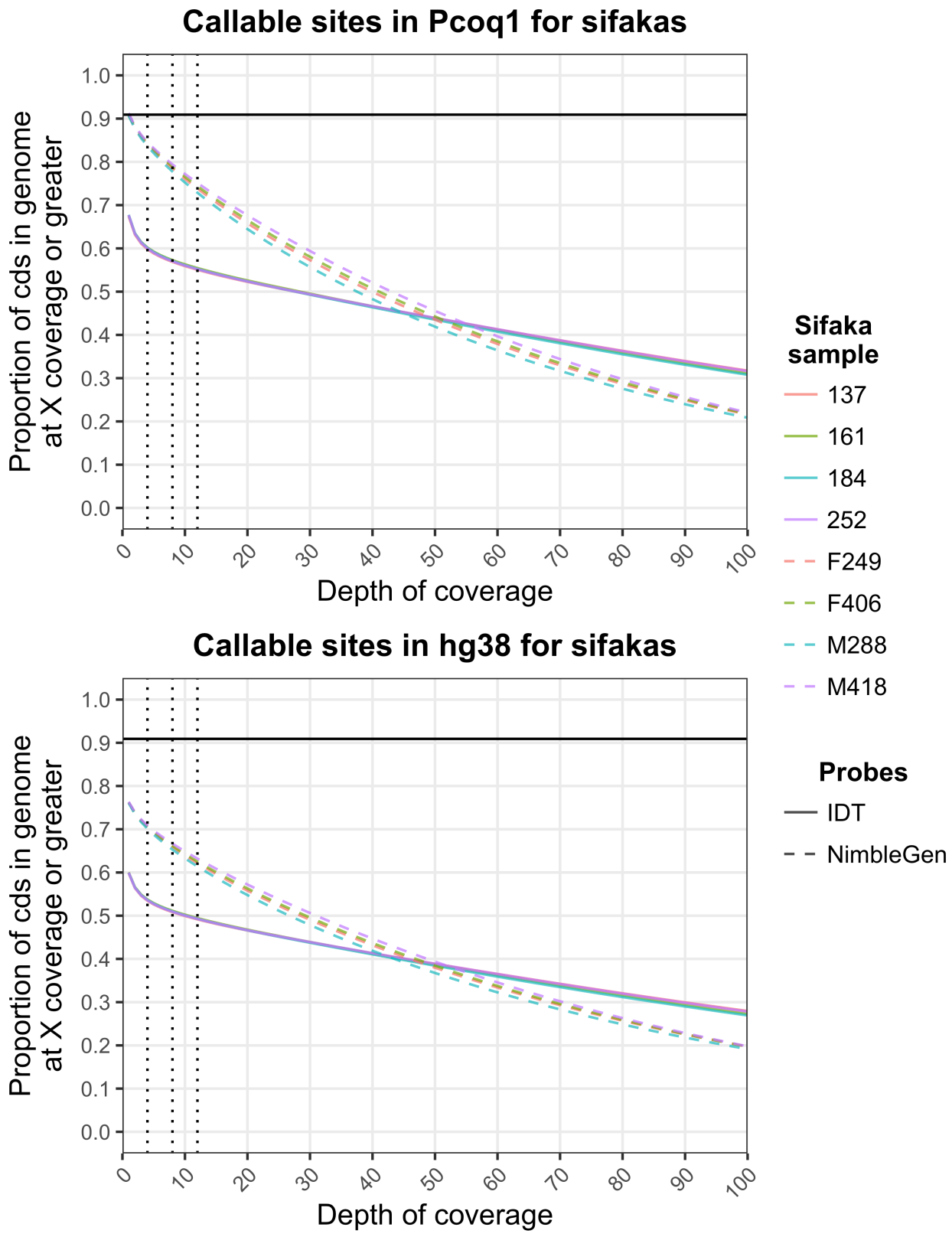

Figure 3. Depth of coverage across the coding regions of the genome. Samples are mapped to their most closely related reference genome (pcoq1 for sifakas and mmul8 for macaques) and the human reference genome (hg38). The x-axis presents depth of coverage, measured as the number of nonduplicate reads with MAPQ >= 20. The y-axis presents the proportion of coding sequence in the genome with X or greater coverage, where X is the value on the x-axis. The vertical dotted lines highlight three common filter values: 4x or greater coverage, 8x or greater coverage, and 12x or greater coverage.

674 The solid horizontal line marks the fraction of the genome covered by one or more

675 reads for macaque samples mapped to mmul8.

676

677




678

679 Figure 4. A comparison of depth of coverage across the coding regions of the genome
680 for NimbleGen and IDT baits. Samples are mapped to pcoq1 and hg38. Samples
681 captured with NimbleGen baits are represented by dashed lines, while those captured
682 with IDT baits are represented by solid lines. The x-axis presents depth of coverage,
683 measured as the number of nonduplicate reads with MAPQ >= 20. The y-axis presents
684 the proportion of coding sequence in the genome with X or greater coverage, where X is
685 the value on the x-axis. The vertical dotted lines highlight three common filter values: 4x
686 or greater coverage, 8x or greater coverage, and 12x or greater coverage. The solid
687 horizontal line marks the fraction of the genome covered by one or more reads for
688 NimbleGen samples mapped to pcoq1.
689

690 # Tables

691 **Table 1. Coding variants identified.[a]**

| Coding variant type | Sifaka1 *(Pcoq1)* | Sifaka1 *(hg38)* | Macaque1 *(Mmul8)* | Macaque1 *(hg38)* |
|---|---|---|---|---|
| Synonymous | 62,201 | 114,951 | 96,650 | 95,982 |
| Nonsynonymous | 29,079 | 102,711 | 51,743 | 75,499 |
| Frameshift variant | 1,230 | 9,648 | 1,044 | 13,171 |
| Stop | 409 | 3,563 | 408 | 3,231 |

692 [a]Values are counts of variants identified for each dataset-reference genome

693 combination.

694