Received: 18 September 2018 | Revised: 14 December 2018 | Accepted: 8 January 2019

DOI: 10.1111/1755-0998.12993

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

A roadmap for high-throughput sequencing studies of wild animal populations using noninvasive samples and hybridization capture

Lauren C. White[1] | Claudia Fontsere[2] | Esther Lizano[2] | David A. Hughes[3,4] | Samuel Angedakin[1] | Mimi Arandjelovic[1] | Anne-Céline Granjon[1] | Jörg B. Hans[1] | Jack D. Lester[1] | M. Timothy Rabanus-Wallace[5] | Carolyn Rowney[1] | Veronika Städele[1] | Tomas Marques-Bonet[2,6,7,8] | Kevin E. Langergraber[9,10] | Linda Vigilant[1]

[1]Department of Primatology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[2]Institut de Biologia Evolutiva (Consejo Superior de Investigaciones Científicas-Universitat Pompeu Fabra), Barcelona Biomedical Research Park, Barcelona, Spain

[3]MRC Integrative Epidemiology Unit at University of Bristol, Bristol, UK

[4]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[5]Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK), Seeland, Germany

[6]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain

[7]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain

[8]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Barcelona, Spain

[9]School of Human Evolution and Social Change, Arizona State University, Tempe, Arizona

[10]Institute of Human Origins, Arizona State University, Tempe, Arizona

Correspondence
Lauren C. White, Department of Primatology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.
Email: lauren_white@eva.mpg.de

Funding information
Max Planck Society; President's Strategic Initiative Fund of ASU

Abstract

Large-scale genomic studies of wild animal populations are often limited by access to high-quality DNA. Although noninvasive samples, such as faeces, can be readily collected, DNA from the sample producers is usually present in low quantities, fragmented, and contaminated by microorganism and dietary DNAs. Hybridization capture can help to overcome these impediments by increasing the proportion of subject DNA prior to high-throughput sequencing. Here we evaluate a key design variable for hybridization capture, the number of rounds of capture, by testing whether one or two rounds are most appropriate, given varying sample quality (as measured by the ratios of subject to total DNA). We used a set of 1,780 quality-assessed wild chimpanzee (*Pan troglodytes schweinfurthii*) faecal samples and chose 110 samples of varying quality for exome capture and sequencing. We used multiple regression to assess the effects of the ratio of subject to total DNA (sample quality), rounds of capture and sequencing effort on the number of unique exome reads sequenced. We not only show that one round of capture is preferable when the proportion of subject DNA in a sample is above ~2%–3%, but also explore various types of bias introduced by capture, and develop a model that predicts the sequencing effort necessary for a desired data yield from samples of a given quality. Thus, our results provide a useful guide and pave a methodological way forward for researchers wishing to plan similar hybridization capture studies.

KEYWORDS
chimpanzees, conservation genomics, faecal samples, population genomics, target enrichment

1 | INTRODUCTION

The dynamics of wild animal populations may be effectively revealed through genetic analyses. Microsatellites have been the backbone of such population genetic studies since the early 1990s, enabling study of parentage assignment, individual discrimination, abundance estimation and demographic inferences, among other topics (Ashley & Dow, 1994; Morin et al., 1994; Paetkau & Strobeck, 1994). However, by today's standards, microsatellites represent a very small amount of genetic data and provide relatively modest power, making questions such as the reliable assessment of familial kin relationships or genetic signals of local adaptation inaccessible (Gienapp et al., 2017; Städele & Vigilant, 2016). The availability and use of high-throughput sequencing technology has exploded over the last two decades and has been applied to various wild animal populations, such as in the study of admixture in baboons (Wall et al., 2016), genomic signals of adaptation in finless dolphins (Zhou et al., 2018) and recombination rate variation in wild red deer (Johnston, Huisman, Ellis, & Pemberton, 2017). However, the trapping or darting of animals necessary for obtaining high-quality DNA samples is often impractical for ethical and logistical reasons. Although noninvasive samples such as faeces can typically be readily collected, DNAs isolated from such samples are often fragmented, present in low quantities, and contaminated by microorganism and dietary DNA, rendering standard high-throughput sequencing from more than a few individuals cost-prohibitive.

To reduce sequencing costs and improve data quality, researchers have begun to use hybridization capture (also termed enrichment) to increase the proportion of subject DNA in sequencing libraries prepared from noninvasive samples (Hernandez-Rodriguez et al., 2018; Perry, Marioni, Melsted, & Gilad, 2010; Snyder-Mackler et al., 2016; van der Valk, Durazo, Dalén, & Guschanski, 2017). These methods involve first hybridizing target DNA to DNA or RNA baits, then immobilizing the target/bait complex, and washing away nontarget DNA fragments before sequencing. Optimization experiments have shown that bait characteristics, sample characteristics and experimental conditions all impact capture efficiency (Ávila-Arcos et al., 2015; Bragg, Potter, Bi, & Moritz, 2016; Carpenter et al., 2013; Cruz-Dávalos et al., 2017; Enk et al., 2014; Hernandez-Rodriguez et al., 2018; Mason, Li, Helgen, & Murphy, 2011; Paijmans, Fickel, Courtiol, Hofreiter, & Förster, 2016; Portik, Smith, & Bi, 2016; Schott et al., 2017).

Despite such efforts towards the optimization of hybridization capture, several factors hamper the cost-effective use of the large number of noninvasive samples needed for population-level studies. For example, although conducting successive rounds of capture can increase the proportion of nontarget DNA that is removed (thus increasing library "specificity"), each round of capture necessitates more cycles of PCR, thereby increasing the proportion of duplicate (i.e., redundant) reads in the library (thus decreasing library "complexity"). The impact of low library complexity may only become apparent as sequencing effort increases, because as more and more sequences are produced from a given library, the probability of a new read representing a PCR duplicate theoretically increases until every new read represents redundant information. At this point, a DNA library made from the sample is effectively saturated, and no new information can be derived from increased sequencing. Libraries of lower complexity reach saturation earlier and require more sequencing effort to reach similar data yields compared to libraries of equal specificity, but greater complexity.

This trade-off between library specificity and complexity leads to different recommendations for samples of high and low quality. For samples of higher quality, one round of capture may suffice to provide adequate library specificity, but for lower quality samples, two rounds of capture are generally recommended, despite the resulting decrease in library complexity. However, because hybridization studies have generally focused either on very high-quality (e.g., tissue; Bragg et al., 2016) or low-quality (e.g., ancient DNA; Carpenter et al., 2013) samples, it is unclear at what level of sample quality (i.e., proportion of subject DNA in an extract) two capture rounds provides greater sequencing efficiency than one. This is especially important for studies of noninvasive samples, which typically exhibit high variation in quality (Taberlet & Luikart, 1999). Note that throughout this paper we refer to sample DNA "quality" as the proportion of subject DNA to total DNA, a readily assessed measure. Other aspects of sample quality, such as fragmentation and damage-induced misincorporations, are less quantifiable prior to sequencing and are not addressed here.

We use a large set of noninvasive samples from wild eastern chimpanzees (*Pan troglodytes schweinfurthii*) to explicitly examine how the proportion of subject DNA in the extracts and the number of rounds of capture interact to influence the efficiency of hybridization capture and high-throughput sequencing (as measured by the ratio of unique on-target reads to total reads sequenced). We captured chimpanzee exome DNA in one or two rounds and used multiple regression analysis to determine the threshold of sample quality at which two rounds of capture confers a greater ratio of unique reads mapping to the exome to total reads sequenced.

Because the standard method of pooling samples for capture according to molarity may lead to sequencing bias across samples within a pool (Hernandez-Rodriguez et al., 2018), we conducted shotgun sequencing to re-estimate the proportion of subject DNA in each sequencing library. This enables us to compare two measures (qPCR/Fragment Analyzer vs. shotgun) of percentage subject DNA in each library, and to construct pools of samples with similar proportions of subject DNA prior to capture, a practice recommended by Hernandez-Rodriguez et al. (2018) to minimize sequencing bias. We also explored within-sample bias introduced by successive rounds of capture by examining evenness of capture across the target space and three possible drivers of bias: GC content, fragment length, and divergence between the bait-design species (human) and the target species (chimpanzee).

By incorporating sequencing effort in our multiple regression model, we are able to show how the expected data yield (i.e., the number of uniquely mapped reads) varies with sample quality and

sequencing effort, or alternatively, the extent of sequencing required to reach a desired data yield for samples of a given quality, thus providing a useful guide to the feasibility of hybridization capture. Finally, to further facilitate future research we present detailed protocols and budget summaries. Our results provide a methodological way forward for large-scale hybridization capture-based studies of noninvasive samples, and highlight the importance of explicit consideration of sample availability and quality when project planning.

2 | METHODS

For easy reference, a simple schematic detailing the various steps of our protocol is provided in Supporting Information Figure S1.

2.1 | Sample collection and screening

Chimpanzee faecal samples were collected opportunistically during routine surveys for the removal of illegal snares at Kibale National Park, Uganda, from 2011 to 2016 and were stored using a two-step ethanol–silica preservation method (Nsubuga et al., 2004). As part of an on-going population size monitoring project, DNA was extracted from these faecal samples using either the GeneMATRIX Stool DNA Purification Kit (Roboklon) according to the manufacturer's instructions or QIAmp Stool kit (Qiagen) with slight modifications from the manufacturer's protocol (Nsubuga et al., 2004). Microsatellite genotyping was performed to establish sex and individual identity as previously described (Arandjelovic et al., 2009; Granjon, Rowney, Vigilant, & Langergraber, 2017). We considered only extracts that were successfully genotyped at enough loci to confidently assign an ID (for details see Granjon et al., 2017) for exome sequencing. Additionally, two tissue samples were collected from two deceased chimpanzees found within Kibale. These samples were collected in RNAlater (Ambion), extracted using the DNeasy Blood and Tissue kit (Qiagen) and also microsatellite genotyped.

Because the DNA present in each faecal extract is expected to derive from bacterial, fungal and dietary sources as well as the chimpanzee itself, we estimated the concentration of amplifiable chimpanzee DNA in each extract using a qPCR assay designed by Morin, Chambers, Boesch, and Vigilant (2001) with some modifications. Reactions were performed in triplicate using 1 μl of extract or DNA standard, 1× Maxima SYBR Green Mastermix (Thermo Scientific) and 0.3 μM forward and reverse primer. All qPCRs included no-template controls and were performed on a Bio-Rad CFX96 instrument with the following cycling conditions: 95°C for 15 min, followed by 40 cycles of 94°C for 30 s, 59°C for 30 s and 72°C for 30 s with a plate read after every cycle. We then measured the total DNA content of each extract using the Fragment Analyzer System (Large Fragment Standard Sensitivity Kit; Advanced Analytical). We define the estimated fraction of subject DNA in each extract as the estimated chimpanzee DNA concentration divided by the total DNA concentration.

2.2 | Library preparation

Along with the two tissue extracts which serve as high-quality DNA controls, we chose 110 faecal extracts, each representing a unique individual, for exome sequencing based on their estimated percentage of subject DNA relative to total DNA concentration. Because studies of ancient DNA found that hybridization capture of samples with <1% subject DNA yielded only small amounts of useable data (Ávila-Arcos et al., 2015; Cruz-Dávalos et al., 2018), we conservatively chose extracts with more than 2% subject DNA. In addition, we only used samples with total DNA concentrations >6 ng/μl so that we could remove 200 ng for library preparation without needing to concentrate the extract.

Recent published works, and our own Fragment Analyzer results (chromatograms, data not shown), show that fragment length distributions of faecal sample extracts are extremely varied, both across and within samples, and that such samples require shearing to acquire shorter and more normally distributed fragment lengths for library preparation (Hernandez-Rodriguez et al., 2018; van der Valk et al., 2017). Hernandez-Rodriguez et al. (2018) found that shorter fragments already below the target size were not further fragmented through sonication (which may have led to a loss of useable data), probably because of the exponentially higher amount of energy required to shear shorter fragments. Thus, our selected extract DNAs were sheared to 200-bp fragments using the Covaris S2 ultrasonicator (Covaris) under the following settings: intensity 5, duty cycle 10%, cycles per burst 200, treatment time 120 s, temperature 7°C and water level 14.

We converted 200 ng of sheared extract into Illumina sequencing libraries using the "BEST" method described by Carøe et al. (2018), with some modifications described in detail in the Supporting Information. After library preparation, two PCRs were performed in parallel using the purified nick-repaired products. One PCR, conducted to prepare the libraries for shotgun sequencing, used adapter-targeted primers with a 5′ tail incorporating the full indexing adapter (P5/P7 Indexing Primers, Supporting Information Table S1). The other PCR, done to prepare the libraries for hybridization capture, used adapter-targeted primers without the extended sequence (i.e., without the index and the remainder of the Illumina adapter, P5/P7 PreHyb Primers, Supporting Information Table S1). PCRs were performed in 50-μl volumes using 5 μl of template, 1× Phusion Mastermix (Thermo Scientific) and 0.5 μM forward and reverse primer under the following conditions: 98°C for 30 s, followed by 8–10 cycles (depending on qPCR results, Supporting Information) of 98°C for 10 s, 60°C for 30 s and 72°C for 30 s, followed by 72°C for 5 min.

Libraries prepared for shotgun sequencing were pooled in equimolar concentrations, and 2 μl of each library blank was added to the pool. The pool was then sequenced on one lane of an Illumina HiSeq 2500 (125-bp read lengths, paired-end).

2.3 | Hybridization capture

Our experimental design was constructed to assess the suitability of one versus two rounds of capture for libraries with varied

subject DNA content. A recent study on noninvasive sample hybridization capture found a high correlation between subject DNA content and total number of reads sequenced across pooled libraries (Hernandez-Rodriguez et al., 2018). The authors suggest an approach to reduce sequencing bias in which all samples are pooled in equitable ratios with respect to subject DNA prior to capture (in contrast to the standard equimolar ratios). As such, it is important to accurately estimate the percentage of subject DNA for each library. Therefore, following data processing and read mapping (see below), we used the shotgun sequencing data to re-estimate the percentage subject DNA (as the number of unique mapped reads divided by total reads sequenced) prior to hybridization capture.

We then calculated the subject DNA molarity of each library by multiplying total molarity by percentage subject DNA as estimated from shotgun sequencing. This value was used to pool faecal sample libraries equitably into 11 groups of 10 libraries (Table 1). The two tissue sample libraries were pooled separately, captured once and used as positive controls. Five faecal sample pools were captured once; the remaining six were captured in two rounds.

Many capture experiments sequence the same library after each round of capture to have a matched comparison of one versus two rounds (Hernandez-Rodriguez et al., 2018; van der Valk et al., 2017). We chose to eschew such a design for two reasons. First, our large sample size allows meaningful analysis without matched sampling. Second, removing a portion of the library after the first round of capture will reduce the total amount of DNA (already significantly reduced during the first round of capture) available for the second round of capture, potentially biasing the results.

We chose SureSelect XT Human All Exon version 6 RNA Library baits (target space ~60 Mb, Agilent) for the enrichment of libraries because human exome baits have been used to capture chimpanzee exomes successfully in the past (Jin et al., 2012; Vallender, 2011). We followed a modified version of the manufacturer's protocol using homemade buffers and custom xGen blocking oligos (P5/P7 Blocking Oligos, IDT), which use proprietary modifications to block the barcode sequences and increase the melting temperature of blocker–adapter duplexes. Our protocol is provided in detail in the Supporting Information. After enrichment via one or two rounds of capture (Table 1), all library-pools were then pooled at equimolar concentrations and the pool was sequenced on two lanes on the Illumina HiSeq 2500 using paired-end, 125-bp read lengths.

2.4 | Sequence processing

Raw reads were demultiplexed and internal barcodes were removed using SABRE (https://github.com/najoshi/sabre) allowing for a single mismatch (−*m* = 1). Adapter contamination was trimmed using BBDUK (part of the BBTOOLS software suite: http://jgi.doe.gov/data-and-tools/bbtools/) and reads overlapping by 10 bp or more were merged using FLASH (Magoč & Salzberg, 2011). We mapped reads to the chimpanzee reference genome (panTro4) using BBMAP with default settings. Collapsed and paired reads were mapped separately and the resultant bam files were merged and sorted using SAMTOOLS (Li et al., 2009). Duplicate reads were identified and removed using PICARDTOOLS (http://broadinstitute.github.io/picard/) and reads were filtered by minimum length (35 bp) and a minimum mapping quality score of 30 using the *reformat* option of BBTOOLS. Target regions (as provided by Agilent) were translated from hg19 coordinates to panTro4 using the liftover utility from the UCSC (University of California Santa Cruz) Genome Browser with default parameters (Kuhn, Haussler, & Kent, 2013). Finally, reads mapping to the target region were extracted using BEDTOOLS *intersectbed* (Quinlan & Hall, 2010). From the resulting bam files we extracted read counts and coverage information using a combination of SAMTOOLS and BEDTOOLS.

We estimated library complexity and saturation curves for each library individually using the *lc_extrap* option of PRESEQ (Daley & Smith, 2013). PRESEQ uses duplication rate information from shallow sequencing experiments to predict the gain in unique reads from increased sequencing effort. However, PRESEQ was not designed for capture experiments and ignores off-target reads. We therefore modified the

TABLE 1 Experimental design to assess the outcome of one versus two rounds of capture

Pool	Number of libraries	Percentage subject DNA in libraries[a]	Rounds of capture
1	10	22.7–37.4	1
2	10	17.6–21.9	2
3	10	15.3–17.5	1
4	10	13.4–15.1	2
5	10	10.3–12.5	1
6	10	9.3–10.3	2
7	10	8.2–9.2	1
8	10	6.3–8.0	2
9	10	5.5–6.3	1
10	10	3.8–5.2	2
11	10	1.7–3.0	2
12[b]	2	79.0–87.0	1

[a]Estimated using shotgun sequencing. [b]Pool 12 is the tissue sample library pool.

PRESEQ output to correct the amount of required sequencing by the fraction of on-target reads in the libraries. This modification results in varying "sequenced reads" values for each library analysed. To allow valid comparisons to be made across libraries, we used linear interpolation to extract the predicted number of unique reads for a given number of sequenced reads for all samples. R-code for the PRESEQ output modification and linear interpolation are available on github (https://github.com/mtrw/white_etal_preseq_hybcap).

2.5 | Analysis of capture rounds

Our measure of capture efficiency is the number of unique (i.e., non-redundant) reads that map to the target region (the exome) out of the total number of reads sequenced (proportion of unique, on-target reads). We aimed to assess how the proportion of unique, on-target reads is impacted by the number of rounds of capture, in interaction with the ratio of subject to total DNA (sample quality). However, the proportion of unique, on-target reads is not static in regard to sequencing effort because, as sequencing effort increases, DNA libraries will approach the saturation point at variable rates, depending on complexity. When saturation is reached, all unique reads have been exhausted and all new reads represent duplicate (i.e., redundant) information. We therefore also included sequencing effort in our multiple regression model by measuring the number of unique on-target reads at various levels of sequencing effort as provided by the modified PRESEQ output.

Specifically, after excluding the tissue-sample control libraries, we had 110 faecal sample libraries measured at eight levels of sequencing effort (i.e., 2, 5, 8, 11, 14, 15, 17 and 20 million total reads sequenced, $n = 8 \times 110 = 880$). We evaluated the predicted number of unique on-target reads (response variable) for various values of total reads sequenced (modified from the original PRESEQ output as above). This allowed us to include and standardize sequencing effort as a predictor in our model, as well as examine the probable impact of the number of rounds of capture at higher sequencing depths than we actually produced. Our other predictors were the number of rounds of capture (one or two) and the precapture percentage of subject DNA in each library (estimated from shotgun sequencing results). We included the three-way interaction of our predictors and all lower terms this encompassed.

Prior to fitting the model we examined the distribution of each predictor, and log-transformed the precapture percentage of subject DNA to achieve a more symmetrical distribution. We then z-transformed the two covariate predictors to yield comparable estimates and a more easily interpretable model (Schielzeth, 2010). To verify the assumptions of normality and homogeneously distributed residuals we visually inspected a qqplot and scatterplot of the residuals plotted against fitted values. These and diagnostics of model stability (leverage, dffits and Cook's distance) suggested some potentially influential cases (Quinn & Keough, 2002). However, fitting the model with these data points excluded revealed essentially identical results as the model based on all data, and hence we report the results obtained from the complete data set. To test for collinearity we checked variance inflation factors (VIFs; Zuur, Ieno, & Elphick, 2010) of the model excluding interactions, which indicated that collinearity was not an issue (largest VIF = 1.13). We fit the model in R (version 3.4.2; R Core Team, 2017) using the function *lm*, and first compared the full model to the null model (comprising just the intercept) using an *F*-test, before examining individual effects including the interactions (Forstmeier & Schielzeth, 2011). R-code for our model implementation is available on github (https://github.com/mtrw/white_etal_preseq_hybcap).

2.6 | Exploration of capture bias

We assessed bias introduced by successive rounds of capture by first examining whether the amount of drop-out and/or highly sequenced targets increased with rounds of capture. We extracted average depth of coverage per target region (from the bed file provided by Agilent with overlapping and book-ended regions merged), per sample ($n = 112$) using the BEDTOOLS *coverageBed* option, excluding the X and Y chromosomes. Using multiple regression, we tested whether the average number of drop-out regions (depth = 0) or highly sequenced regions (depth > 10) was significantly different between samples captured in one or two rounds, whilst controlling for sequencing effort and capture success by including the number of unique exome reads per sample. As above, we fit the models in R using the function *lm*, and ensured that assumptions were met. We again found some evidence of influential cases, but, as above, we found no difference in results when they were excluded and thus present the full data set. Full–null model comparisons were made using null models which included our control predictor, namely the number of unique exome reads. R-code for this model implementation is also available on github (https://github.com/mtrw/white_etal_preseq_hybcap).

To explore what may be driving bias within samples we determined the GC content (using BEDTOOLS*nuc* applied to the panTro4 reference) and the proportion of mismatches between the human (hg19, with which the bait sequences were designed) and chimpanzee (panTro4) reference genomes (using the axt alignment files provided by UCSC) for each target region (again excluding the X and Y chromosomes). We also extracted the GC content per read (using unix command line tools, *sed* and *grep*), and the length distribution of mapped reads (using PICARDTOOLS) from bam files of a subset of samples (for computational efficiency and ease of visualization, 10 captured in one round, 10 captured in two rounds; Pools 1 and 2) before and after capture.

2.7 | Data and prediction validation

To confirm that PRESEQ produced reliable projections with our data set, we further sequenced 10 enriched libraries (Pool 1) using one additional lane of the Illumina HiSeq 2500 (125-bp read lengths, paired-end). After sequence processing as above, we ran PRESEQ on these data using the *c_curve* option, which uses down-sampling to yield the number of unique on-target reads for experiments using the same or less sequencing effort. After correcting the amount of required sequencing by the fraction of on-target reads as above, we

compared these results with our predictions (based on shallow sequencing) of unique on-target reads.

To further confirm the acquisition of biologically meaningful information using hybridization captured DNA data, we used the 10 deeply sequenced samples to examine chimpanzee genetic variation in a subspecies context. We combined the exome data from these 10 samples, with high coverage (16–26×) chimpanzee whole genome shotgun data from eight individuals (Prado-Martinez et al., 2013) and moderate to high coverage (4–49×) chimpanzee exome data from six individuals (Hernandez-Rodriguez et al., 2018). These additional data represent all four chimpanzee subspecies, including additional individuals from Kibale National Park. We used this combined data set to extract genotype likelihoods using the program ANGSD (Korneliussen, Albrechtsen, & Nielsen, 2014), using the default settings of the GATK model, but requiring sites to be covered in all 24 samples. These genotype likelihoods were then used to perform principal components analysis (PCA) using PCANGSD (Meisner & Albrechtsen, 2018).

3 | RESULTS

3.1 | Sample selection and estimation of subject DNA proportion

The 1,780 faecal extracts that were successfully genotyped represent 738 individuals from across Kibale National Park. Of these extracts, 316 (17.8%), representing 235 individuals, contain >2% subject DNA, and had total DNA concentrations >6 ng/μl (Supporting Information Figure S2). From this subset we chose the 110 unique individual extracts, in addition to the two tissue extracts, for exome sequencing.

We constructed libraries from each of the 112 extracts and shotgun sequenced an average of 1,223,582 reads per library (range = 751,451–2,036,321, Supporting Information Table S2) to examine how well our qPCR/Fragment Analyzer estimate reflected the true subject DNA proportion of our DNA libraries. We found only a moderate correlation between percentage subject DNA estimated from qPCR/Fragment Analyzer and from shotgun sequencing (Pearson's $r = 0.58$, $p < 0.001$, Supporting Information Figure S3). The mean per cent of subject DNA in the faecal sample extracts as estimated by shotgun sequencing was 11.6% (range = 1.7%–37.4%), while the average estimated using qPCR/Fragment Analyzer was 7.6% (range = 2%–58.6%). As we consider the estimate from shotgun sequencing to be more reliable, we used these values to calculate volumes for pooling of libraries prior to capture and hereafter refer to this estimate. No library blank control yielded any reads mapping to the chimpanzee genome after filtering (Supporting Information Table S2).

3.2 | Capture success

After performing one or two rounds of capture on our library pools, we sequenced on average 2,646,199 reads per captured library (range = 208,215–5,435,336, Supporting Information Table S2). The variation in the raw data acquired per library varied much less than was reported by Hernandez-Rodriguez et al. (2018; range 0.2–5.5 million reads here compared to 0.7–45.9). Unlike that observed in Hernandez-Rodriguez et al., we did not observe a correlation between number of raw reads and precapture percentage subject DNA (Pearson's $r = -0.03$, $p = 0.69$, Supporting Information Figure S4), indicating a reduction in sequencing bias due to equimolar pooling (as suggested by Hernandez-Rodriguez et al., 2018) within our capture pools. Below we provide a description of shotgun sequencing and capture results before describing our model results in the next section.

For the shotgun-sequenced faecal DNA libraries, the average percentage of reads mapping to the chimpanzee genome was 13.7% (range = 1.9%–45.8%, please note this differs from our calculation in the section above as it includes duplicated and unfiltered reads), the average percentage of on-target reads (i.e., those mapping to the chimpanzee exome) was 0.4% (range = 0.08%–1.4%) and the duplication rates (i.e., the average number of times a unique fragment was sequenced) were low (mean = 1, range = 1–1.02). By comparison, after enrichment the average percentage of reads mapping to the chimpanzee genome was 88.3% (one round, range = 73.1%–97%) and 97.1% (two rounds, range = 95%–98.3%), the average percentage of on-target reads was 67.1% (one round, range = 57%–74.3%) and 85.5% (two rounds range = 81.4%–87.2%), and the average duplication rates were 1.16 (one round, range = 1.06–1.35) and 1.61 (two rounds, range = 1.32–2.26). The mapping data for individual samples can be found in Supporting Information Table S2.

This inverse relationship of mapping and duplication rates in libraries captured in one or two rounds resulted in a small difference in the percentage of unique exome reads: 55% (range = 44.4%–66%) for libraries captured in one round and 51.4% (range = 33.8%–63.1%) for libraries captured in two rounds. Thus, enrichment of exome reads in faecal extracts ranged from 46- to 538-fold. Saturation curves predicted by PRESEQ analysis showed that libraries with higher precapture percentage subject DNA (as estimated by shotgun sequencing) were generally more complex (i.e., had a steeper saturation curve) after capture (Supporting Information Figure S5).

For the two tissue samples, 3% of reads were on-target in shotgun-sequenced libraries and duplicate rates were again low (range = 1–1.02). After enrichment the average percentage of chimpanzee genome reads, average percentage of on-target reads, duplication rates and percentage unique on-target reads for the two singly captured libraries were 98.2% (*SD* = 0.6), 72.3% (*SD* = 2.2%), 1.05 (*SD* = 0.00) and 66.1% (*SD* = 2.2%), respectively, resulting in an average enrichment of 24-fold (Supporting Information Table S2). Saturation curves for these two tissue samples revealed them to be more complex than all the faecal samples, as expected (Supporting Information Figure S5).

3.3 | Preferred number of rounds of capture

Overall, our predictors clearly influenced the number of unique on-target reads estimated by PRESEQ (full-null model comparison:

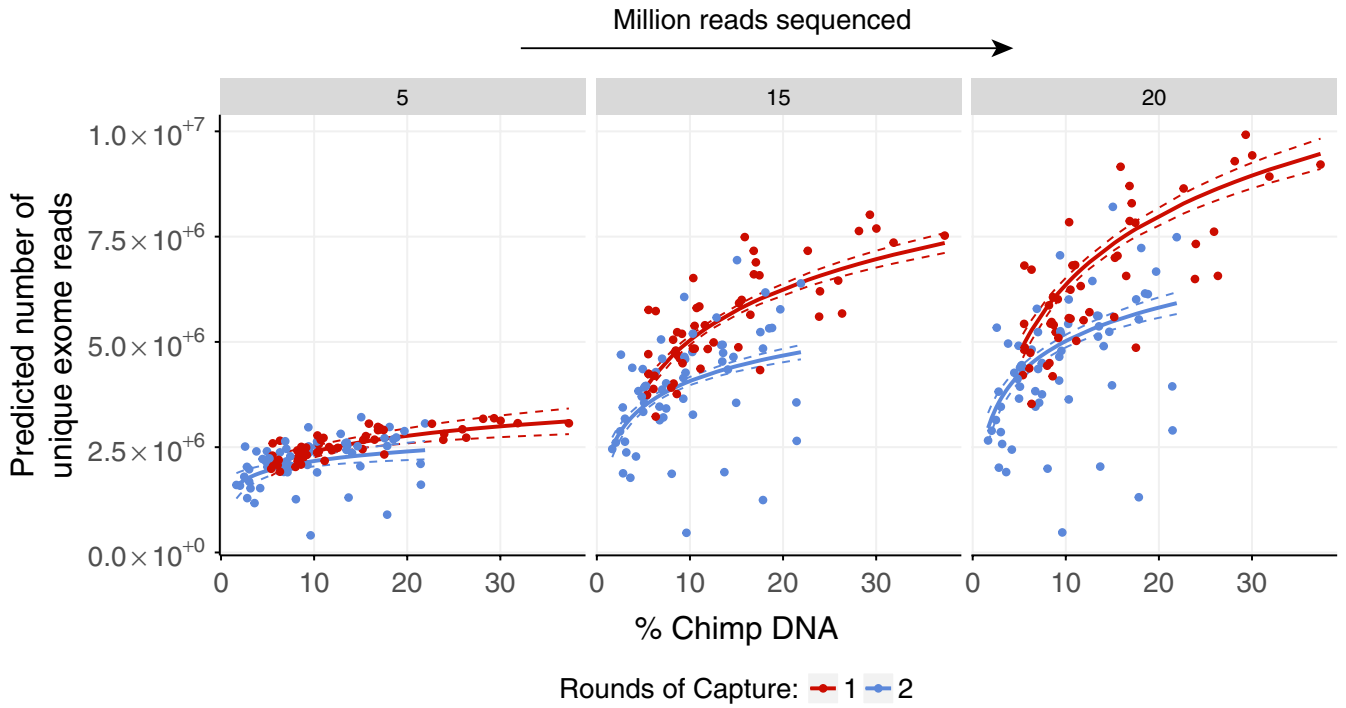

FIGURE 1 The effect of rounds of capture, precapture percentage subject DNA and sequencing effort on the number of unique exome reads (as predicted by PRESEQ). The plots show the predicted data and model at increasing sequencing effort (at 5, 15 and 20 million reads sequenced from left to right). Each point represents a captured library, the fitted model is shown as solid lines and 95% model confidence intervals are shown as broken lines [Colour figure can be viewed at wileyonlinelibrary.com]

$F_{7,872} = 413.2$, $p < 0.001$). More specifically, one round of capture led to a greater number of unique on-target reads than two rounds of capture. While this general pattern held regardless of sequencing effort or starting percentage of subject DNA (across the ranges covered by our data), the increase in unique on-target reads conferred by one round of capture compared to two was greater with increasing sequencing effort and (log-scaled) starting percentage subject DNA (three-way interaction, $p = 0.00019$, Supporting Information Table S3, Figure 1).

Our model predicts that, for samples with lower precapture percentage subject DNA (<2%–3%), the observed pattern would reverse and two rounds of capture would lead to a greater number of unique on-target reads than one round (Supporting Information Figure S6). However, our comparative data do not extend to these low values of precapture percentage subject DNA, and therefore such predictions should be taken with caution.

3.4 | Capture bias

Rounds of capture influenced both the number of drop-out regions and the number of highly sequenced regions (full–null model comparisons: drop-out: $F_{-1,110} = 77.78$, $p < 0.001$ and highly sequenced: $F_{-1,110} = 193.09$, $p < 0.001$). Performing two rounds of capture increased both the number of dropout regions and the number of highly sequenced regions (Figure 2, Supporting Information Tables S4 and S5). Visual inspection of depth of coverage across the target regions (Supporting Information Figure S7) confirmed that the bias towards or away from certain regions was largely consistent across captured samples from both faecal and tissue samples, and was not observed in whole genome shotgun sequencing data (Prado-Martinez et al., 2013). We observed an increase in average fragment length from 155.8 bp (*SD* = 63 bp) across shotgun-sequenced samples to 194 bp (*SD* = 77.2 bp) after one round of capture and 202.5 bp (*SD* = 74.3 bp) after two rounds of capture (Figure 3a). We similarly observed a change in the distribution of read GC content, with average percentage GC increasing from 46% (*SD* = 10.4%) in shotgun sequenced samples, to 52.9% (*SD* = 10.4%) after one round of capture and 56.4% (*SD* = 9.1%) after two rounds (Figure 3b). This pattern of GC bias was also observed across target regions, with average depth of coverage highest in regions with 55%–65% GC (Supporting Information Figure S8). Finally, we found no pattern of decreasing average depth of coverage with increasing divergence between the chimpanzee and human reference regardless of round of capture (Supporting Information Figure S9).

3.5 | Data validation

The average number of reads produced for the 10 libraries (Pool 1) that were subject to deep sequencing was 14,755,628 (range = 6,341,251–18,622,980), resulting in an average depth of coverage of our target region of 17.3× per library (range = 7.6–24.5×). Comparison of the predicted and observed saturation curves produced by PRESEQ shows a good match, although

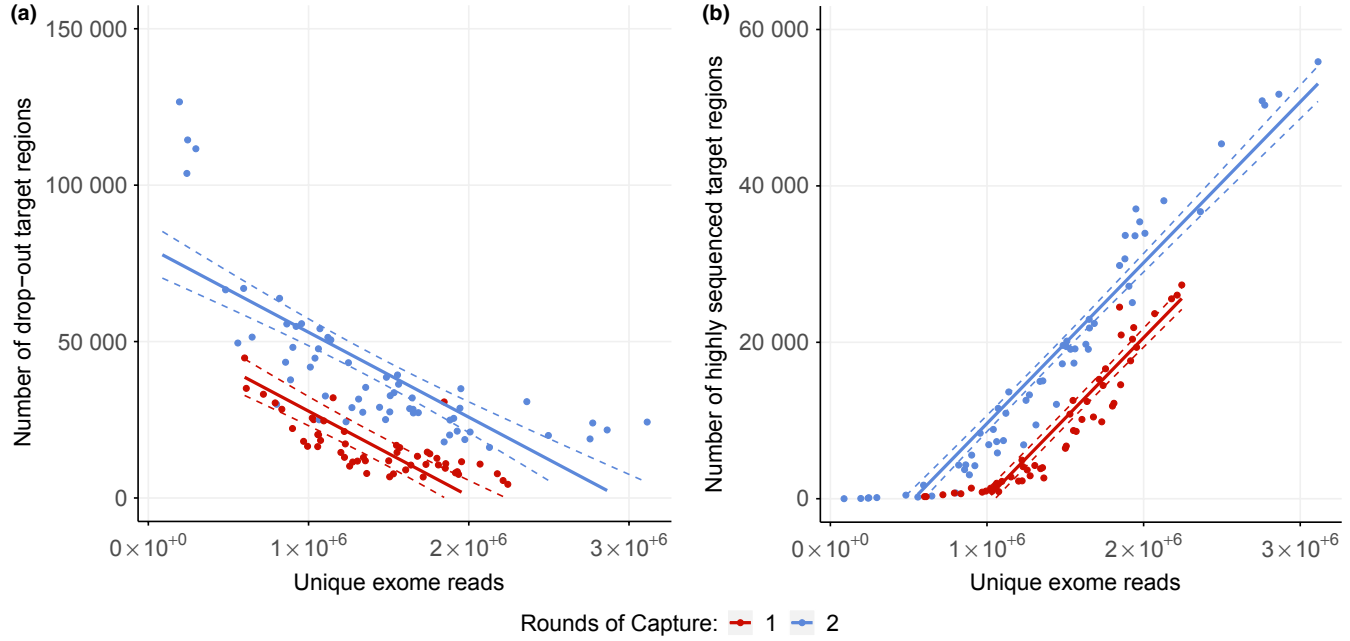

FIGURE 2 The effect of rounds of capture on the number of (a) drop-out target regions (depth = 0) and (b) highly sequenced regions (depth > 10). Each point represents a captured library, the fitted model is shown as solid lines and 95% model confidence intervals are shown as broken lines [Colour figure can be viewed at wileyonlinelibrary.com]

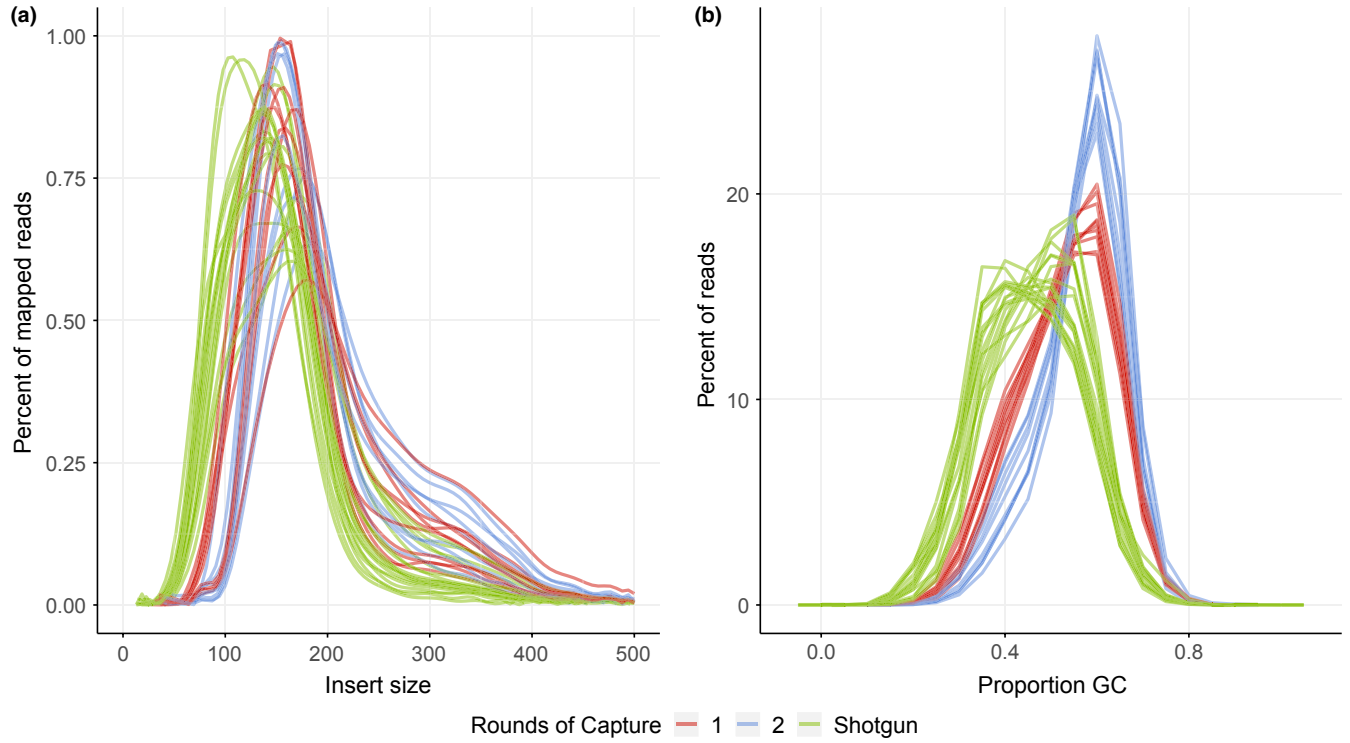

FIGURE 3 Exploration of capture bias in samples from Pools 1 and 2. Density plots of (a) percentage of mapped reads of a given insert size and (b) percentage of all reads with given GC proportion. Lines are coloured by the rounds of capture: none/shotgun in green, one round in red and two rounds in blue [Colour figure can be viewed at wileyonlinelibrary.com]

three libraries (N17311, N21602, N21608) had 1%–2% fewer unique reads on-target compared to that predicted (Supporting Information Figure S10). Finally, our PCA shows that, as expected, the 10 samples from this study cluster with other eastern chimpanzee samples – specifically, with samples originating from Kibale National Park (Figure 4).

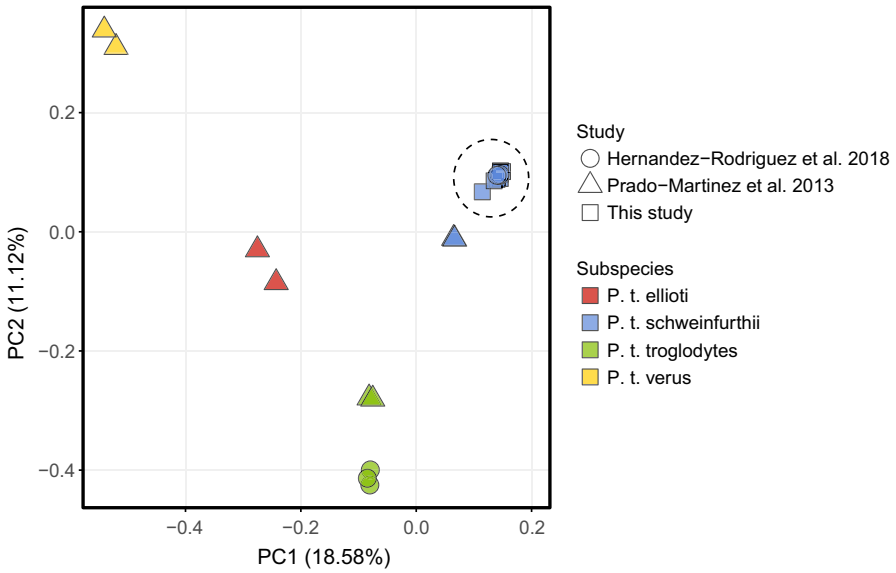

FIGURE 4 Principal components analysis showing the genetic population structure of our 10 study samples (squares), and 14 samples from other studies (circles and triangles). Samples originating from Kibale National Park, Uganda, are circled with a broken line [Colour figure can be viewed at wileyonlinelibrary.com]

3.6 | Budget details

If we aimed to produce 15 million reads per library, our model predicts that we could achieve 5,297,949 unique on-target reads (95% confidence interval: 5,196,338–5,399,560) for samples with 11.6% subject DNA (the average of our 110 chosen faecal extracts) using a single round of exome capture. Assuming the high correlation between number of unique exome reads and average depth of coverage observed in our data (r^2 = 0.97, Supporting Information Figure S11) continues when sequencing effort is higher, we can translate this predicted number of unique on-target reads to an average depth of coverage of ~13.9×. We consider this depth of coverage to be adequate for most downstream population genetic analyses and therefore provide a detailed budget (Supporting Information Table S6) based on the above conditions (i.e., a single round of capture and sequencing effort that is equal to producing 15 million reads per library). Under these conditions the cost per sample for library preparation, hybridization capture and sequencing is ~€230 at the time of writing, most of which is accounted for by sequencing costs (~€194 per library; based on multiplexing ~13 libraries per lane of an Illumina 2500, version 3 chemistry). Including a second round of capture simply doubles the cost of the capture step and brings the total cost to ~€250. The cost of qPCR/Fragment Analyzer screening was roughly €2 per extract, while the cost of sample collection, DNA extraction and microsatellite genotyping will vary between laboratories and projects. Thus, these additional costs should be considered on a per-project basis.

In Figure 5 we generalize sequencing costs to provide a guide for researchers wishing to undertake their own hybridization capture study of noninvasive samples. Figure 5 shows the predicted number of unique on-target reads for various precapture subject DNA contents at increasing sequencing effort and relates this effort to sequencing cost. Our cost estimates are based on the use of one lane of an Illumina 2500 (version 3 chemistry, which produces ~200 million reads) and relies on researchers multiplexing to the highest degree possible. For example, to produce 10 million reads per library at a cost of €125 each, 20 libraries must be multiplexed on a single lane. These predictions are based on the specific protocol described here, and we discuss below how variations in design, which other studies may wish to implement, may impact capture efficiency and thus cost.

We note that our budget is based on the cost of performing molecular analysis in Europe, and that part of the cost of sample collection involves shipping samples out of range-countries, which may lack the infrastructure to conduct such experiments. As researchers, we should work towards a situation in which such analyses can be performed *in situ*, making such extractive collection practices unnecessary. Such contributions could include support for local infrastructure, or providing mentorship to local students. For example, the Ngogo Chimpanzee Project supports local schools in Uganda and provides scholarships for Ugandan masters students to conduct research at Kibale National Park (http://ngogochimpanzeeproject.org/education/). Such efforts are important to create more equitable, sustainable, ethical and cost-effective research practices.

4 | DISCUSSION

In this study we successfully used hybridization capture to enrich 110 libraries derived from faecal DNA extracts for exome DNA. We used these data to examine the appropriate number of rounds of capture when using a range of concentrations of subject DNA relative to total DNA, explore bias introduced by capture, and provide a roadmap for future hybridization capture studies of noninvasive samples.

4.1 | Rounds of capture

Our analysis shows that one round of capture is preferable to two for samples with greater than ~2%–3% subject DNA. For lower

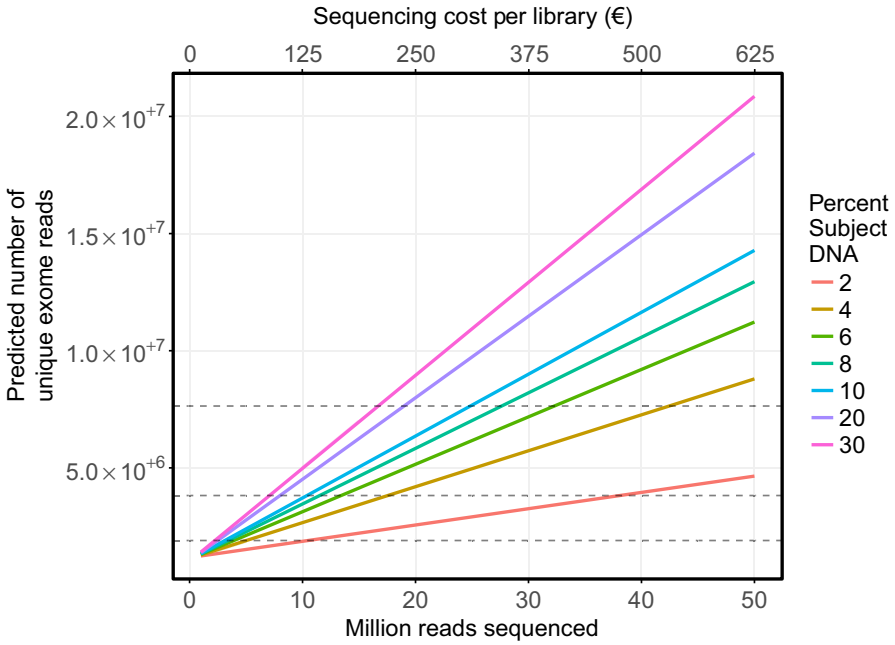

FIGURE 5 Predicted number of exome reads, for samples of various subject DNA percentages, at increasing sequencing effort/cost using a single round of capture. Broken grey lines represent the estimated number of reads needed to achieve 5×, 10× and 20× (from bottom to top) average depth of coverage of our target space (the exome). Sequencing cost per library is calculated based on the use of one flow cell of an Illumina HiSeq 2500 (version 3 chemistry) [Colour figure can be viewed at wileyonlinelibrary.com]

precapture subject DNA proportions we predict that two rounds of capture are optimal, although further experimental work is needed to confirm this. It is possible that the correlation of capture success with precapture subject DNA proportions may become nonlinear at lower values due to, for example, increased rates of misbinding between baits and library DNA, leading to deviation from our predictions. Although a number of studies focused on ancient DNA provide some evidence that the correlation of capture success with precapture subject DNA proportions holds for lower quality samples (Ávila-Arcos et al., 2015; Carpenter et al., 2013; Cruz-Dávalos et al., 2017, 2018), these studies lack the sample sizes needed for statistical tests of this hypothesis. Thus, further experimental work on samples with low proportions of subject DNA are necessary.

Our results appear to contrast with a recent study that also focused on exome hybridization capture using DNA from chimpanzee faecal samples. Hernandez-Rodriguez et al. (2018) found two rounds of capture to be more efficient for samples with a range of subject DNA from 0.16% to 24.6%. However, more than half of the samples in that study (10–12 of 18) were below our ~2%–3% cut-off value, and the interaction between subject DNA content and rounds of capture was not accounted for in the Hernandez-Rodriguez et al. (2018) model. Thus, their recommendation does not contradict the results presented here, but probably reflects the quality distribution of their sample set, with the majority of their samples benefiting from two rounds of capture.

4.2 | Bias introduced by capture

Our analysis of capture bias showed that samples subjected to two rounds of capture were more likely to have more highly sequenced target regions, but that this comes at the cost of drop-out at other regions. This result, and our visual inspection of depth of coverage across the target space and across samples, shows that capture is indeed biasing our sequencing results towards some target regions and away from others, and that this effect is exacerbated in libraries subject to two rounds of capture. This means that even when the number of unique reads mapping to the exome is equal across samples, samples captured in one round will have more uniform coverage across the target space compared to samples captured twice. This result has implications for the preferred number of rounds of capture. Our model of rounds of capture is based on maximizing the number of unique exome reads, regardless of the distribution of reads across the target space. Thus, although we predict that two rounds of capture will maximize the unique exome reads for samples with less than 2%–3% subject DNA, researchers dealing with such samples must also consider the increase in bias, and effective decrease in total target space, that two rounds will confer.

We explored three possible drivers of capture bias: GC content, fragment length, and divergence between bait sequence and target DNA sequence. We found that the distribution of read GC content narrowed with successive rounds of capture and that target regions with GC proportions in the range of 55%–65% had higher average depth of coverage. This indicates that GC content is a probable driver of capture bias, as has been shown previously (Ávila-Arcos et al., 2015; Cruz-Dávalos et al., 2017). Conversely, we found no decrease in average depth of coverage across target regions with increasing divergence between the chimpanzee and human genome. This is in agreement with previous work on chimpanzee exome capture using human baits (Jin et al., 2012; Vallender, 2011), and suggests that the overall divergence between the chimpanzee target DNA and human bait sequences is not large enough to interfere with capture success, with the large majority of regions having less than 2% mismatches. Also in agreement with previous studies (Ávila-Arcos et al., 2015; Carpenter et al., 2013; Enk et al., 2014), we observed an increase in average fragment length in captured libraries compared to shotgun data. Ávila-Arcos et al. (2015) hypothesized that fragment length

bias may be a result of the distribution of bait lengths, with longer baits biasing against smaller target molecules. Unfortunately, we were unable to test this hypothesis, as well as a number of other possible drivers of bias (e.g., tiling density, bait sequence content), due to our use of commercial baits, for which we have no information on bait design. Further work on capture bias should use custom bait sets to fully explore the drivers of capture bias, and such work could then be used to improve bait design and reduce bias in capture studies.

4.3 | Roadmap for hybridization capture of noninvasive samples

We were able to incorporate sequencing effort into our analysis of rounds of capture by using predictions from the program PRESEQ, which we validated by sequencing 10 libraries to higher depth to show a good match between predictions and observations. This approach allowed us to relate sample quality (as measured by subject DNA content) to sequencing costs and/or probable data yields, to assist in project planning of future research. For example, achieving an average depth of 5× for the chimpanzee exome, from samples with 2% subject DNA, requires roughly 10 million reads per sample, which translates to a cost of €125 each (plus the cost of library preparation and capture, Supporting Information Table S6).

Our results can be used as a guide to the feasibility of hybridization capture wherever researchers possess a reasonable understanding of the depth of coverage needed for the study at hand, and awareness about the quality of available samples. The necessary depth of coverage will depend on the research question and desired analytical framework of a particular study. For example, although single nucleotide variant calling is usually only recommended for samples with depths of coverage greater than 15–20× (Meynert, Ansari, FitzPatrick, & Taylor, 2014; Sims, Sudbery, Ilott, Heger, & Ponting, 2014), an increasing number of population genetic software can accommodate low to medium coverage (4–15×) data by using, for example, genotype likelihoods or haploid calling methods (Korneliussen & Moltke, 2015; Therkildsen & Palumbi, 2017; Vieira, Albrechtsen, & Nielsen, 2016; Wall et al., 2016).

With regard to sample quality, we highly recommend sampling extensively and prescreening extracts. Our results show that samples with <2% subject DNA would only yield small amounts of data using the protocols presented here, even after switching to two rounds of capture (see below for a discussion of other possible protocol modifications). The proportion of our sample set that met our criteria (2% subject DNA and 6 ng/μl ng total DNA concentration) was 17%, but this is likely to vary widely across species, collection environments and collection methods. Pilot studies that characterize the quality distribution of sample sets can inform researchers of the total number of samples that need to be collected to reach a desired number of usable samples. Generally, we expect the number of usable samples to scale proportionally with the total number of samples collected, and thus sampling as extensively as possible is recommended.

Screening extracts not only allows selection of samples most likely to yield usable data, but is also imperative for equitable pooling prior to capture to reduce sequencing bias. However, our qPCR/Fragment Analyzer estimate of subject DNA content was only moderately correlated with the estimate derived from shotgun sequencing. This could be due to differences in sensitivity to fragment length (our qPCR assay can only assess fragments larger than 81 bp, while we can map sequenced fragments as short as 35 bp), compounded by biases toward smaller fragments during library preparation (Dabney & Meyer, 2012; Enk, Rouillard, & Poinar, 2013). Unless more accurate screening assays are available or can be developed for the study species, we recommend shallow shotgun sequencing prior to pooling to re-estimate the percentage of subject DNA in each library.

4.4 | Caveats and other considerations

The generalizability of our results does of course have limits. A number of variables that we did not test can meaningfully impact the efficiency of hybridization-enrichment, and these factors should be considered when predicting yields and making cost estimates. First, the size of the target space is directly related to the number of reads required to reach a given depth of coverage. Our target space was ~60 Mb, and smaller or larger spaces will proportionally affect the cost of bait synthesis and sequencing. Second, studies on capture efficiency across scales of divergence indicate that increased evolutionary distance between the species for which the baits were designed and the target species will decrease capture efficiency, particularly for divergence estimates exceeding 20 million years (Bragg et al., 2016; Jin et al., 2012; Portik et al., 2016). This is especially important for species that do not yet have a high-quality reference genome. Third, because commercial baits are generally designed for high-quality DNA, these bait sets may not be as efficient for DNA from noninvasive samples, and redesign with increased bait tiling density (i.e., the average number of unique bait molecules that cover each position of the target sequence) may improve capture success (Bodi et al., 2013; Clark et al., 2011). However, such drawbacks of pre-designed, "off-the-shelf" options should be weighed against the fact that custom-designed bait sets are generally (depending on target space and tiling density) more expensive.

Other modifications to our protocol may also increase capture efficiency and accommodate lower quality samples. For example, increasing the amount of DNA per library in a capture reaction by lowering the number of pooled libraries (and keeping the total input amount the same) should increase the number of unique fragments per sample available for capture and thus increase the post-capture library complexity (McCartney-Melstad, Mount, & Shaffer, 2016). Similarly, preparing multiple libraries and/or conducting multiple captures per sample may also prove more cost-efficient than deep sequencing of a singly captured library, particularly for very low-quality samples (Hernandez-Rodriguez et al., 2018). These modifications unavoidably increase the cost of library preparation, capture and sequencing per sample, but may allow researchers to obtain useable data for samples with <2% subject DNA.

A final consideration for studies in which faecal samples are collected from unobserved animals is the possible inadvertent sampling of nontarget conspecific species (Arandjelovic et al., 2010), and complications arising from the presence of diet-related DNA (Hofreiter, Kreuz, Eriksson, Schubert, & Hohmann, 2010). Distinguishing different species' remains can be achieved through mitochondrial or microsatellite DNA analysis. Microsatellite genotyping is also necessary to identify unique individuals and thus should be considered a necessary part of the screening process. For example, our samples were first subjected to analysis at 15 microsatellite loci and categorized as individuals in the context of a long-running project (Granjon et al., 2017; Langergraber, Mitani, & Vigilant, 2007; Langergraber, Watts, Vigilant, & Mitani, 2017). We could therefore be confident that our putative chimpanzee samples were indeed from chimpanzees. A more insidious issue is the concurrent presence of diet-related DNA from closely related species. For example, chimpanzees are known to hunt and eat various species of sympatric primates (Watts & Mitani, 2002), which could result in other primate DNA being co-captured during hybridization. This has the potential to impact downstream analyses, especially at conserved genetic regions. Therefore, the development of bioinformatic tools to identify contaminated extracts from shallow shotgun data, or remove or account for contaminant reads *in silico*, as is possible for studies of ancient DNA (Racimo, Renaud, & Slatkin, 2016; Renaud, Slon, Duggan, & Kelso, 2015; Skoglund et al., 2014), should be considered a high priority.

5 | CONCLUSION

Our study adds to a growing body of literature showing that large-scale genomic studies are feasible for noninvasive samples through hybridization capture, and provides further data to help optimize such research. To our knowledge this study represents the first exploration of the appropriate rounds of capture with respect to sample quality, finding that one round of capture is more efficient for samples with more than ~2%–3% subject DNA. Additionally, we have confirmed that successive rounds of capture introduce biases to sequencing data, with fewer rounds of capture ensuring more even coverage across the target space. We have presented detailed protocols, budgets and model output to act as a guide for researchers wishing to implement hybridization capture of noninvasive samples and show that, with an understanding of sample availability, quality and necessary data yields, the feasibility of such studies can be easily assessed.

ACKNOWLEDGEMENTS

We thank Anette Nicklisch, Roland Schroeder, Amy Heilman and Sebastian Ramirez Amaya for assistance in the laboratory, Barbara Schellbach and Antje Weihmann for conducting the Illumina sequencing, Janet Kelso, Johann Visagie and Joshua Schmidt for bioinformatic help and guidance, Thomas Gilbert for sharing library preparation protocols prior to their publication, Richard McElreath and Dieter Lukas for helpful discussion, and Prof. Ávila Arcos and two anonymous reviewers for constructive comments. Collection of samples would not have been possible without the invaluable assistance of the Ngogo Chimpanzee Project Field assistants and snare team members, Lawrence Ndangizi, Godfrey Mbabazi, Alfred Tumusiime, Ambrose Twineomujuni, Brian Kamugyisha, Charles Birungi, Christopher Aliganyira, James Tibisimwa, William Sunday, Braise Mugyisha, David Sunday, Ronald Mugume, Justus Byamugisha, Joseph Tumusiime, James Kyeyune and Snofrex Turyakyira. Funding was provided by the Max Planck Society and the President's Strategic Initiative Fund of ASU. T.M.B. is supported by BFU2017-86471-P (MINECO/FEDER, UE), U01 MH106874 grant, Howard Hughes International Early Career, Obra Social "La Caixa" and Secretaria d'Universitats i Recerca and CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880) and C.F. is supported by a La Caixa PhD Fellowship.

DATA ACCESSIBILITY

Adapter-trimmed sequencing data have been deposited at NCBI's short read archive and are available under the accession code PRJNA505752. R-code and data for PRESEQ output modifications and model implementation are available on github (https://github.com/mtrw/white_etal_preseq_hybcap).

AUTHOR CONTRIBUTIONS

L.C.W. and L.V. conceived and designed the study, and drafted the manuscript. L.C.W. conducted the experiments and analysed the data. K.E.L., S.A. and C.R. supervised sample collection, and genotyping. C.R., A-C.G. and V.S. conducted sample quality assessments and genotyping. J.B.H. designed and optimized the qPCR assay. M.T.R-W. provided R-code for modifying PRESEQ output. C.F., E.L., M.A. and J.D.L. provided discussion on experimental design. C.F., D.H., E.L. and T.M-B. supplied laboratory protocols, and bioinformatics expertise. All authors gave input to the manuscript and approved the final version.

ORCID

Lauren C. White https://orcid.org/0000-0001-8085-9293
Claudia Fontsere https://orcid.org/0000-0003-2233-6026

REFERENCES

Arandjelovic, M., Guschanski, K., Schubert, G., Harris, T. R., Thalmann, O., Siedel, H., & Vigilant, L. (2009). Two-step multiplex polymerase chain reaction improves the speed and accuracy of genotyping using DNA from noninvasive and museum samples. *Molecular Ecology Resources*, *9*(1), 28–36. https://doi.org/10.1111/j.1755-0998.2008.02387.x

Arandjelovic, M., Head, J., Kühl, H., Boesch, C., Robbins, M. M., Maisels, F., & Vigilant, L. (2010). Effective non-invasive genetic monitoring of

multiple wild western gorilla groups. *Biological Conservation*, *143*(7), 1780–1791. https://doi.org/10.1016/j.biocon.2010.04.030

Ashley, M. V., & Dow, B. D. (1994). The use of microsatellite analysis in population biology: Background, methods and potential applications. In B. Schierwater, B. Streit, G. P. Wagner & R. DeSalle (Eds.), *Molecular Ecology and Evolution: Approaches and Applications* (pp. 185–201). Basel, Switzerland: Birkhäuser Basel. https://doi.org/10.1007/978-3-0348-7527-1_10

Ávila-Arcos, M. C., Sandoval-Velasco, M., Schroeder, H., Carpenter, M. L., Malaspinas, A.-S., Wales, N., … Gilbert, M. T. P. (2015). Comparative performance of two whole-genome capture methodologies on ancient DNA Illumina libraries. *Methods in Ecology and Evolution*, *6*(6), 725–734. https://doi.org/10.1111/2041-210X.12353

Bodi, K., Perera, A. G., Adams, P. S., Bintzler, D., Dewar, K., Grove, D. S., … Zianni, M. (2013). Comparison of commercially available target Enrichment methods for next-generation sequencing. *Journal of Biomolecular Techniques*, *24*(2), 73–86. https://doi.org/10.7171/jbt.13-2402-002

Bragg, J. G., Potter, S., Bi, K., & Moritz, C. (2016). Exon capture phylogenomics: Efficacy across scales of divergence. *Molecular Ecology Resources*, *16*(5), 1059–1068. https://doi.org/10.1111/1755-0998.12449

Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S. T., Sinding, M. H. S., Samaniego, J. A., … Gilbert, M. T. P. (2018). Single-tube library preparation for degraded DNA. *Methods in Ecology and Evolution*, *9*(2), 410–419. https://doi.org/10.1111/2041-210X.12871

Carpenter, M. L., Buenrostro, J. D., Valdiosera, C., Schroeder, H., Allentoft, M. E., Sikora, M., … Bustamante, C. D. (2013). Pulling out the 1%: Whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *The American Journal of Human Genetics*, *93*(5), 852–864. https://doi.org/10.1016/j.ajhg.2013.10.002

Clark, M. J., Chen, R., Lam, H. Y. K., Karczewski, K. J., Chen, R., Euskirchen, G., … Snyder, M. (2011). Performance comparison of exome DNA sequencing technologies. *Nature Biotechnology*, *29*(10), 908–914. https://doi.org/10.1038/nbt.1975

Cruz-Dávalos, D. I., Llamas, B., Gaunitz, C., Fages, A., Gamba, C., Soubrier, J., … Orlando, L. (2017). Experimental conditions improving in-solution target enrichment for ancient DNA. *Molecular Ecology Resources*, *17*(3), 508–522. https://doi.org/10.1111/1755-0998.12595

Cruz-Dávalos, D. I., Nieves-Colón, M. A., Sockell, A., Poznik, G. D., Schroeder, H., Stone, A. C., … Ávila-Arcos, M. C. (2018). In-solution Y-chromosome capture-enrichment on ancient DNA libraries. *BMC Genomics*, *19*(1), 608. https://doi.org/10.1186/s12864-018-4945-x

Dabney, J., & Meyer, M. (2012). Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *BioTechniques*, *52*(2), 87–94. https://doi.org/10.2144/000113809

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. *Nature Methods*, *10*(4), 325–327. https://doi.org/10.1038/nmeth.2375

Enk, J. M., Devault, A. M., Kuch, M., Murgha, Y. E., Rouillard, J.-M., & Poinar, H. N. (2014). Ancient whole genome enrichment using baits built from modern DNA. *Molecular Biology and Evolution*, *31*(5), 1292–1294. https://doi.org/10.1093/molbev/msu074

Enk, J. M., Rouillard, J.-M., & Poinar, H. (2013). Quantitative PCR as a predictor of aligned ancient DNA read counts following targeted enrichment. *BioTechniques*, *55*(6), 300–309. https://doi.org/10.2144/000114114

Forstmeier, W., & Schielzeth, H. (2011). Cryptic multiple hypotheses testing in linear models: Overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology*, *65*(1), 47–55. https://doi.org/10.1007/s00265-010-1038-5

Gienapp, P., Fior, S., Guillaume, F., Lasky, J. R., Sork, V. L., & Csilléry, K. (2017). Genomic quantitative genetics to study evolution in the wild. *Trends in Ecology & Evolution*, *32*(12), 897–908. https://doi.org/10.1016/j.tree.2017.09.004

Granjon, A.-C., Rowney, C., Vigilant, L., & Langergraber, K. E. (2017). Evaluating genetic capture-recapture using a chimpanzee population of known size. *The Journal of Wildlife Management*, *81*(2), 279–288. https://doi.org/10.1002/jwmg.21190

Hernandez-Rodriguez, J., Arandjelovic, M., Lester, J., de Filippo, C., Weihmann, A., Meyer, M., … Marques-Bonet, T. (2018). The impact of endogenous content, replicates and pooling on genome capture from faecal samples. *Molecular Ecology Resources*, *18*(2), 319–333. https://doi.org/10.1111/1755-0998.12728

Hofreiter, M., Kreuz, E., Eriksson, J., Schubert, G., & Hohmann, G. (2010). Vertebrate DNA in fecal samples from bonobos and gorillas: Evidence for meat consumption or artefact? *PLoS ONE*, *5*(2), e9419. https://doi.org/10.1371/journal.pone.0009419

Jin, X., He, M., Ferguson, B., Meng, Y., Ouyang, L., Ren, J., … Wang, X. (2012). An effort to use human-based exome capture methods to analyze chimpanzee and macaque exomes. *PLoS ONE*, *7*(7), e40637. https://doi.org/10.1371/journal.pone.0040637

Johnston, S. E., Huisman, J., Ellis, P. A., & Pemberton, J. M. (2017). A high density linkage map reveals sexual dimorphism in recombination landscapes in red deer (*Cervus elaphus*). *G3: Genes, Genomes, Genetics*, *7*(8), 2859–2870. https://doi.org/10.1534/g3.117.044198

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Korneliussen, T. S., & Moltke, I. (2015). NgsRelate: A software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics*, *31*(24), 4009–4011. https://doi.org/10.1093/bioinformatics/btv509

Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, *14*(2), 144–161. https://doi.org/10.1093/bib/bbs038

Langergraber, K. E., Mitani, J. C., & Vigilant, L. (2007). The limited impact of kinship on cooperation in wild chimpanzees. *Proceedings of the National Academy of Sciences*, *104*(19), 7786–7790. https://doi.org/10.1073/pnas.0611449104

Langergraber, K. E., Watts, D. P., Vigilant, L., & Mitani, J. C. (2017). Group augmentation, collective action, and territorial boundary patrols by male chimpanzees. *Proceedings of the National Academy of Sciences*, *114*(28), 7337–7342. https://doi.org/10.1073/pnas.1701582114

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Subgroup, 1000 Genome Project Data Processing (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Magoč, T., & Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, *27*(21), 2957–2963. https://doi.org/doi:10.1093/bioinformatics/btr507

Mason, V. C., Li, G., Helgen, K. M., & Murphy, W. J. (2011). Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. *Genome Research*, *21*(10), 1695–1704. https://doi.org/10.1101/gr.120196.111

McCartney-Melstad, E., Mount, G. G., & Shaffer, H. B. (2016). Exon capture optimization in amphibians with large genomes. *Molecular Ecology Resources*, *16*(5), 1084–1094. https://doi.org/10.1111/1755-0998.12538

Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, *210*(2), 719–731. https://doi.org/10.1534/genetics.118.301336

Meynert, A. M., Ansari, M., FitzPatrick, D. R., & Taylor, M. S. (2014). Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics*, *15*(1), 247. https://doi.org/10.1186/1471-2105-15-247

Morin, P. A., Chambers, K. E., Boesch, C., & Vigilant, L. (2001). Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan*

troglodytes verus). *Molecular Ecology*, *10*(7), 1835–1844. https://doi.org/10.1046/j.0962-1083.2001.01308.x

Morin, P. A., Moore, J. J., Chakraborty, R., Jin, L., Goodall, J., & Woodruff, D. S. (1994). Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science*, *265*(5176), 1193–1201. https://doi.org/10.1126/science.7915048

Nsubuga, A. M., Robbins, M. M., Roeder, A. D., Morin, P. A., Boesch, C., & Vigilant, L. (2004). Factors affecting the amount of genomic DNA extracted from ape faeces and the identification of an improved sample storage method. *Molecular Ecology*, *13*(7), 2089–2094. https://doi.org/10.1111/j.1365-294X.2004.02207.x

Paetkau, D., & Strobeck, C. (1994). Microsatellite analysis of genetic-variation in black bear populations. *Molecular Ecology*, *3*(5), 489–495. https://doi.org/10.1111/j.1365-294X.1994.tb00127.x

Paijmans, J. L. A., Fickel, J., Courtiol, A., Hofreiter, M., & Förster, D. W. (2016). Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. *Molecular Ecology Resources*, *16*(1), 42–55. https://doi.org/10.1111/1755-0998.12420

Perry, G. H., Marioni, J. C., Melsted, P., & Gilad, Y. (2010). Genomic-scale capture and sequencing of endogenous DNA from feces. *Molecular Ecology*, *19*(24), 5332–5344. https://doi.org/10.1111/j.1365-294X.2010.04888.x

Portik, D. M., Smith, L. L., & Bi, K. (2016). An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (Class: Amphibia, Order: Anura). *Molecular Ecology Resources*, *16*(5), 1069–1083. https://doi.org/10.1111/1755-0998.12541

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., … Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, *499*(7459), 471–475. https://doi.org/10.1038/nature12228

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Quinn, G. P., & Keough, M. J. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511806384

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Racimo, F., Renaud, G., & Slatkin, M. (2016). Joint estimation of contamination, error and demography for nuclear DNA from ancient humans. *PLoS Genetics*, *12*(4), e1005972. https://doi.org/10.1371/journal.pgen.1005972

Renaud, G., Slon, V., Duggan, A. T., & Kelso, J. (2015). Schmutzi: Estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biology*, *16*(1), 224. https://doi.org/10.1186/s13059-015-0776-0

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, *1*(2), 103–113. https://doi.org/10.1111/j.2041-210X.2010.00012.x

Schott, R. K., Panesar, B., Card, D. C., Preston, M., Castoe, T. A., & Chang, B. S. W. (2017). Targeted capture of complete coding regions across divergent species. *Genome Biology and Evolution*, *9*(2), 398–414. https://doi.org/10.1093/gbe/evx005

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

Skoglund, P., Northoff, B. H., Shunkov, M. V., Derevianko, A. P., Pääbo, S., Krause, J., & Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences*, *111*(6), 2229–2234. https://doi.org/10.1073/pnas.1318934111

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 non-invasively collected samples. *Genetics*, *203*(2), 699–714. https://doi.org/10.1534/genetics.116.187492

Städele, V., & Vigilant, L. (2016). Strategies for determining kinship in wild populations using genetic data. *Ecology and Evolution*, *6*(17), 6107–6120. https://doi.org/10.1002/ece3.2346

Taberlet, P., & Luikart, G. (1999). Non-invasive genetic sampling and individual identification. *Biological Journal of the Linnean Society*, *68*(1–2), 41–55. https://doi.org/10.1111/j.1095-8312.1999.tb01157.x

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, *17*(2), 194–208. https://doi.org/10.1111/1755-0998.12593

van der Valk, T., Durazo, F. L., Dalén, L., & Guschanski, K. (2017). Whole mitochondrial genome capture from faecal samples and museum-preserved specimens. *Molecular Ecology Resources*, *17*(6), e111–e121. https://doi.org/10.1111/1755-0998.12699

Vallender, E. J. (2011). Expanding whole exome resequencing into non-human primates. *Genome Biology*, *12*(9), R87. https://doi.org/10.1186/gb-2011-12-9-r87

Vieira, F. G., Albrechtsen, A., & Nielsen, R. (2016). Estimating IBD tracts from low coverage NGS data. *Bioinformatics*, *32*(14), 2096–2102. https://doi.org/10.1093/bioinformatics/btw212

Wall, J. D., Schlebusch, S. A., Alberts, S. C., Cox, L. A., Snyder-Mackler, N., Nevonen, K. A., … Tung, J. (2016). Genomewide ancestry and divergence patterns from low-coverage sequencing data reveal a complex history of admixture in wild baboons. *Molecular Ecology*, *25*(14), 3469–3483. https://doi.org/10.1111/mec.13684

Watts, D. P., & Mitani, J. C. (2002). Hunting behavior of chimpanzees at Ngogo, Kibale National Park. *Uganda*. *International Journal of Primatology*, *23*(1), 1–28. https://doi.org/10.1023/A:1013270606320

Zhou, X., Guang, X., Sun, D., Xu, S., Li, M., Seim, I., … Yang, G. (2018). Population genomics of finless porpoises reveal an incipient cetacean species adapted to freshwater. *Nature Communications*, *9*(1), 1276. https://doi.org/10.1038/s41467-018-03722-x

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3–14. https://doi.org/10.1111/j.2041-210X.2009.00001.x

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: White LC, Fontsere C, Lizano E, et al. A roadmap for high-throughput sequencing studies of wild animal populations using noninvasive samples and hybridization capture. *Mol Ecol Resour.* 2019;19:609–622. https://doi.org/10.1111/1755-0998.12993