

# Maximizing the acquisition of unique reads in noninvasive capture sequencing experiments

Claudia Fontserè<sup>1</sup>  | Marina Alvarez-Estape<sup>1</sup>  | Jack Lester<sup>2</sup> | Mimi Arandjelovic<sup>2</sup> | Martin Kuhlilm<sup>1</sup> | Paula Dieguez<sup>2</sup> | Anthony Agbor<sup>2</sup> | Samuel Angedakin<sup>2</sup> | Emmanuel Ayuk Ayimisin<sup>2</sup> | Mattia Bessone<sup>2</sup>  | Gregory Brazzola<sup>2</sup> | Tobias Deschner<sup>2</sup> | Manasseh Eno-Nku<sup>3</sup> | Anne-Céline Granjon<sup>2</sup> | Josephine Head<sup>2</sup> | Parag Kadam<sup>4</sup> | Ammie K. Kalan<sup>2</sup> | Mohamed Kambi<sup>2</sup> | Kevin Langergraber<sup>5,6</sup> | Juan Lapuente<sup>2,7</sup> | Giovanna Maretta<sup>2</sup> | Lucy Jayne Ormsby<sup>2</sup> | Alex Piel<sup>8</sup> | Martha M. Robbins<sup>2</sup> | Fiona Stewart<sup>4,8</sup> | Virginie Vergnes<sup>9</sup> | Roman M. Wittig<sup>2,10</sup> | Hjalmar S. Kühl<sup>2,11</sup> | Tomas Marques-Bonet<sup>1,12,13,14</sup> | David A. Hughes<sup>15,16</sup> | Esther Lizano<sup>1,14</sup> 

<sup>1</sup>Institut de Biologia Evolutiva, CSIC-Universitat Pompeu Fabra, PRBB, Barcelona, Catalonia, Spain

<sup>2</sup>Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

<sup>3</sup>WWF Cameroon Country Programme Office, Yaoundé, Cameroon

<sup>4</sup>School of Biological and Environmental Sciences, Liverpool John Moores University, James Parsons Building, Liverpool, UK

<sup>5</sup>School of Human Evolution and Social Change, Arizona State University, Tempe, AZ, USA

<sup>6</sup>Institute of Human Origins, Arizona State University, Tempe, AZ, USA

<sup>7</sup>Comoé Chimpanzee Conservation Project, Kakpin, Comoé National Park, Ivory Coast, Côte d'Ivoire

<sup>8</sup>Department of Anthropology, University College London, London, UK

<sup>9</sup>Wild Chimpanzee Foundation (WCF), Abidjan, Côte d'Ivoire

<sup>10</sup>Tai Chimpanzee Project, Centre Suisse de Recherches Scientifiques, Abidjan, Côte d'Ivoire

<sup>11</sup>German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

<sup>12</sup>CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Catalonia, Spain

<sup>13</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

<sup>14</sup>Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain

<sup>15</sup>MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

<sup>16</sup>Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

## Correspondence

Esther Lizano, Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), PRBB, Doctor Aiguader 88, Barcelona 08003, Spain.  
Email: esther.lizano@upf.edu

Tomas Marques-Bonet, Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), PRBB, Doctor Aiguader 88, Barcelona 08003, Spain.  
Email: tomas.marques@upf.edu

David A. Hughes, MRC Integrative Epidemiology Unit at University of Bristol,

## Abstract

Noninvasive samples as a source of DNA are gaining interest in genomic studies of endangered species. However, their complex nature and low endogenous DNA content hamper the recovery of good quality data. Target capture has become a productive method to enrich the endogenous fraction of noninvasive samples, such as faeces, but its sensitivity has not yet been extensively studied. Coping with faecal samples with an endogenous DNA content below 1% is a common problem when prior selection of samples from a large collection is not possible. However, samples classified as unfavourable for target capture sequencing might be the only representatives of unique

Bristol, UK.

Email: d.a.hughes@bristol.ac.uk

**Funding information**

H2020 European Research Council, Grant/Award Number: 864203; FPI (Formación de Personal Investigador) Ministerio de Ciencia, Universidades e Investigación, Grant/Award Number: PRE2018-083966; CERCA Programme del Departament d'Economia i Coneixement de la Generalitat de Catalunya, Grant/Award Number: GRC 2017 SGR 880; Agencia Estatal de Investigación (AEI), Grant/Award Number: "Unidad María de Maeztu"; Secretaria d'Universitats i Recerca; Howard Hughes International Early Career; "La Caixa" doctoral fellowship program; MINECO/FEDER, UE, Grant/Award Number: BFU2017-86471-P and CGL2017-82654-P; Heinz L. Krekeler Foundation; Max Planck Society; Obra Social "La Caixa"; "La Caixa" Foundation, Grant/Award Number: 100010434; Max Planck Society Innovation Fund

specific geographical locations, or to answer the question of interest. To explore how library complexity may be increased without repeating DNA extractions and generating new libraries, in this study we captured the exome of 60 chimpanzees (*Pan troglodytes*) using faecal samples with very low proportions of endogenous content (<1%). Our results indicate that by performing additional hybridizations of the same libraries, the molecular complexity can be maintained to achieve higher coverage. Also, whenever possible, the starting DNA material for capture should be increased. Finally, we specifically calculated the sequencing effort needed to avoid exhausting the library complexity of enriched faecal samples with low endogenous DNA content. This study provides guidelines, schemes and tools for laboratories facing the challenges of working with noninvasive samples containing extremely low amounts of endogenous DNA.

**KEYWORDS**

chimpanzees, conservation genomics, faecal samples, molecular complexity, noninvasive samples, target capture

## 1 | INTRODUCTION

Studies of wild animal populations that are not amenable to invasive sampling (e.g., trapping or darting) often rely on the usage of low quality and/or quantity DNA samples (Schwartz et al., 2007; Vigilant & Guschanski, 2009), traditionally restricting the analysis of neutral markers or genetic loci such as microsatellites (Arandjelovic et al., 2011; Inoue et al., 2013; Mengülluğlu et al., 2019; Orkin et al., 2016), autosomal regions (Fischer et al., 2004) and the mitochondrial genome (Fickel et al., 2007; Thalmann et al., 2004). Depending on the researcher's question, these neutral genetic markers may continue to be the most economical and efficient method (Shafer et al., 2015). However, for other questions such as cataloguing genetic diversity, assessing kinship, making fine inferences of demographic history, or evaluating disease susceptibility, it is increasingly relevant to acquire a more representative view of the genome (Ouborg et al., 2010; Primmer, 2009; Shafer et al., 2015; Städele & Vigilant, 2016; Steiner et al., 2013).

Conservation genomics of ecologically-crucial, non-model organisms, and especially threatened species such as great apes, have largely benefitted from the current advances in next-generation sequencing (NGS) technologies (Gordon et al., 2016; Locke et al., 2011; Mikkelsen et al., 2005; Scally et al., 2012). The ability to simultaneously interrogate hundreds of thousands of genetic markers across an entire genome allows greater resolution on inferences of demographic parameters, genetic variation, gene flow, inbreeding, natural selection, local adaptation and the evolutionary history of the studied species (De Manuel et al., 2016; Prado-Martinez et al., 2013; Xue et al., 2015).

The major impediment to the study of wild, threatened, natural populations continues to be the difficulties in acquiring samples of known location from a large number of individuals. To avoid

disturbing and negatively influencing endangered species (alteration of social group dynamics, infections and stress) (Morin et al., 1993; Taberlet et al., 1999), but also to track cryptic or monitor reintroduced species (De Barba et al., 2010; Ferreira et al., 2018; Reiners et al., 2011; Stenglein et al., 2010), sampling often relies on noninvasive (NI) sources of DNA such as faeces and hair, rather than invasive samples such as blood or other tissues, which yield better DNA quality and quantity.

NI samples have a complex nature: they are typically composed of low proportions of host or endogenous DNA (eDNA), are highly degraded (Perry et al., 2010; Taberlet et al., 1999), and contain genetic material from the microbiota of the host and from species living in the environment where the sample was collected (i.e., exogenous DNA) (Hicks et al., 2018). The proportion of endogenous vs. exogenous DNA can be highly variable (Hernandez-Rodriguez et al., 2018) and as previous literature has proposed, may depend on the environmental conditions, with humidity and ambient temperature having the highest influence (Goossens et al., 2000; Harestad & Bunnell, 1987; King et al., 2018; Nsubuga et al., 2004). Because of this, the employment of techniques that generate sequences of the whole genomic content of the samples, such as NGS, has not been economically feasible until recently. Target enrichment technologies, also known as capture, have become a common and successful methodology in ancient DNA studies (Burbano et al., 2010; Carpenter et al., 2013; Maricic et al., 2010) and have allowed for a more cost-effective use of NGS on NI samples, as the endogenous to exogenous DNA ratio greatly improves, thus reducing the sequencing effort (Perry et al., 2010; Snyder-Mackler et al., 2016; van der Valk et al., 2017). Capture methods reduce the relative cost of sequencing and improve the quality of the data by building DNA libraries that are hybridized to complementary baits for selected target regions (partial genomic regions, a chromosome, the exome, or

the whole genome) increasing the proportion of the targeted eDNA to be sequenced.

Despite the existence of technical studies describing the use of NI samples for the genomic study of wild chimpanzees (*Pan troglodytes*) (Hernandez-Rodriguez et al., 2018; White et al., 2019) many aspects remain to be investigated. For instance, in Hernandez-Rodriguez et al. (2018) samples were selected to cover the entire range of observed average fragmentation lengths and percentage of eDNA, in order to be as representative as possible. As a result, they observed a sequencing bias due to the different percentage of endogenous content in captured samples. To avoid that outcome, they proposed performing equi-endogenous pools instead of the standard pooling of libraries according to molarity. White et al. (2019) followed this recommendation and yielded a more balanced representation across samples. However, their experiments were limited to only those samples with a proportion of eDNA above 2% (White et al., 2019). As shown by Hernandez-Rodriguez et al. (2018) there is a positive association between endogenous content and the amount of data acquired from a sample, such that when possible, one should use those samples with higher endogenous content. However, the proportion of chimpanzee faecal samples with eDNA above 2% is often very low (<20%) (White et al., 2019).

Here, we look to expand on the methods presented in Hernandez-Rodriguez et al. (2018) and White et al. (2019) by focusing on very low endogenous content samples. These previous studies have illustrated the value and quality of genotype data derived from target capture enrichment protocols using complex noninvasive samples. Here, we focus on methods to improve the acquisition of unique, endogenous or host DNA reads - the variable most important in increasing the amount and quality of genotype data.

The NI chimpanzee samples used in this study were collected from 15 different geographic sites across the whole species' ecological habitat in Africa and included all four subspecies, thus representing a wide variety of sampling and environmental conditions. With this screening approach we were able to examine how the proportion of eDNA content varies between each site, revealing that the majority of collected samples in some sites have low proportions of eDNA (<1%). Therefore, when prior selection of samples from a large collection is not possible, the only ones representing a specific location, or that are relevant to the scientific question, might be those with extremely low proportions of endogenous content. Because of that, in this study we have focused our efforts on developing approaches to retrieve the maximum data possible from challenging samples.

In that regard, we sought to capture the exome of 60 chimpanzee faecal samples as part of the Pan African Programme: The Cultured Chimpanzee (PanAf) (<http://panafrican.eva.mpg.de/>) (Kühl et al., 2019) with eDNA estimates below 1%. We used a commercial human exome to evaluate how the coverage of targeted genomic regions may be increased in a collection of samples that may be regarded as unfavorable for target capture sequencing. We confirmed the importance of the correct estimation of eDNA and the pooling of libraries accordingly to avoid sequencing bias across samples

(Hernandez-Rodriguez et al., 2018). We also expanded on previously explored and unexplored guidelines to ensure the maintenance of the captured molecule diversity or library complexity such as the number of libraries in a pool, the performance of additional hybridizations and increasing the total DNA starting material for capture (Hernandez-Rodriguez et al., 2018; Perry et al., 2010; Snyder-Mackler et al., 2016; White et al., 2019).

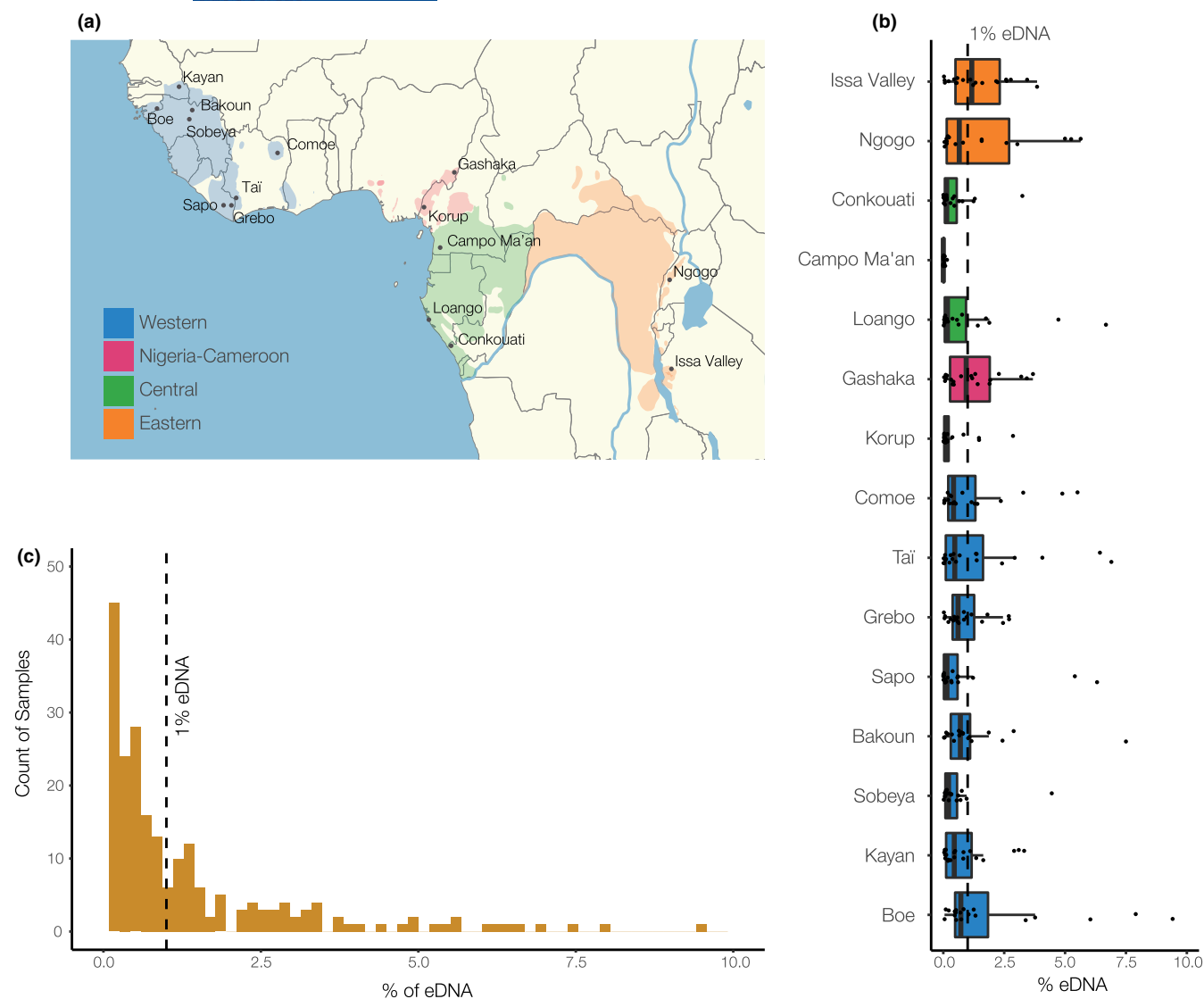
Our results provide the most comprehensive exploration to date of target enrichment efficiency in very low eDNA faecal samples, and guidelines to improve the quality of the data without re-extracting DNA and preparing new libraries. These findings could greatly benefit the conservation effort on great apes, as well as any other species with similar DNA sampling limitations.

## 2 | MATERIALS AND METHODS

### 2.1 | Samples and library preparation

Chimpanzee faecal samples from 15 different sites in Africa were collected as part of the PanAf (Figure 1a). Approximately 5 g ("hazelnut-size") of faeces were collected from each chimpanzee faecal sample and stored in the field using a two-step ethanol-silica preservation method (Nsubuga et al., 2004). Depending on the density of the sample, between 10 and 80 mg of dry faecal sample were extracted using a Qiagen robot with the QIAamp Fast DNA Stool Mini Kit (Qiagen) with modifications (Lester et al., 2020). The extractions, including blanks, were screened using a microsatellite genotyping assay (Arandjelovic et al., 2009, 2011) and up to 20 samples from each PanAf field site were selected as follows: (i) those that amplified at the most loci of the 15 loci panel; (ii) represented unique individuals; and (iii) were ascertained to have a low probability of being first degree relatives (Csilléry et al., 2006) (302 samples) (Table S1). None of the blanks amplified in the microsatellite assays. To ensure sufficient template DNA for library preparation, the 302 samples were re-extracted using the same QIAamp kit and between 100 and 200 mg of dry faecal sample. Total DNA concentration and fragmentation were measured on a Fragment Analyzer using a Genomic DNA 50 Kb Analysis kit (Advanced Analytical) and the fragmentation level was calculated with PROSize Data Analysis Software (Agilent Technologies). Endogenous DNA content (fraction of mammalian DNA, relative to gut microbial and other environmental genetic material) was estimated by qPCR (Morin et al., 2001). Finally, percentage of endogenous content for each sample was calculated by dividing the chimpanzee eDNA concentration by the total DNA concentration. We selected 60 samples with an intermediate percentage of eDNA (0.41%–0.85%, average 0.61%) from the 302 screened samples (range of endogenous distribution: 0–47.57%, average 1.49%) (Figure S1 and Table S2).

A single library was prepared for each of the 60 samples following the BEST protocol (Carøe et al., 2018) starting with 200 ng total DNA (from a sample) with minor modifications. Specifically, double in-line barcoded adapters were used (Figure S2), barcoding



**FIGURE 1** Sample description. (a) Geographical location of the 15 sites from the Pan African Programme: The Cultured Chimpanzee (PanAf). (b) Endogenous DNA (eDNA) content for all screened samples according to geographic origin. The maximum value of the x-axis has been set to 10% eDNA for visual purposes. (c) eDNA distribution for all screened samples. Samples with >10% eDNA are excluded ( $N = 5$ ). In the boxplot, lower and upper hinges correspond to first and third quartiles and the lower and upper whiskers extend to the smallest or largest value no further than 1.5 times the interquartile range (distance between the first and third quartile) [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

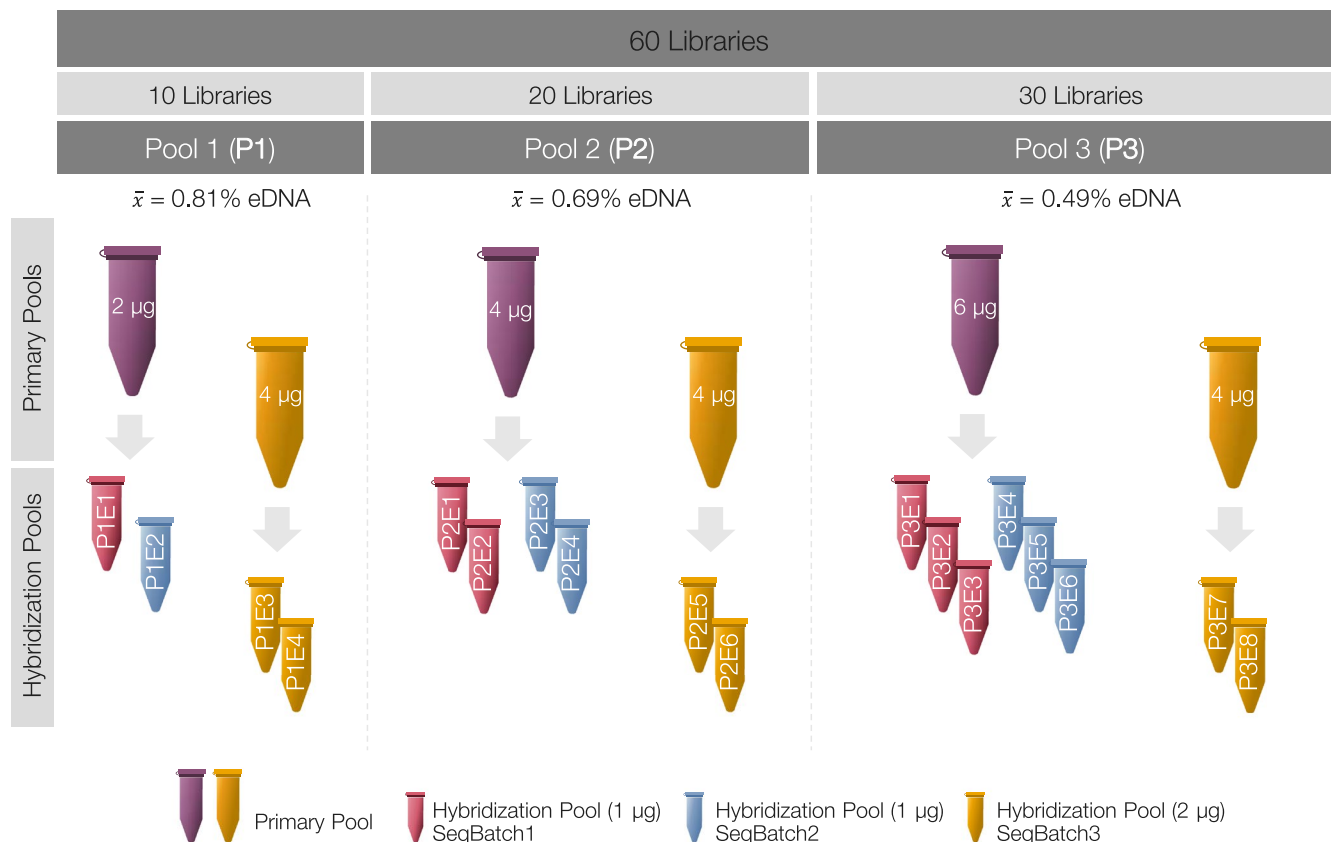
each sample at both ends of its library to allow for its unique identification within a pool (Rohland & Reich, 2012). Library concentration was calculated using Agilent 2100 BioAnalyzer and DNA7500 assay kit. A detailed protocol for library construction can be found in Supporting Information.

## 2.2 | Pooling and capture

Endogenous DNA content is a key factor in target-capture experiments directly influencing the yield of on-target reads and molecule diversity (Hernandez-Rodriguez et al., 2018). Our equi-endogenous sample pooling strategy follows two criteria. First, samples belonging to a pool have similar eDNA proportions according to a 1:2 ratio

rule: the sample with highest proportion of eDNA cannot double the sample with the lowest. Second, each sample within a pool contributes the same total amount of eDNA ( $\mu\text{g}$ ) to the final pool, creating an equi-endogenous pool. Therefore, the sample with the lowest percentage of eDNA will contribute more total DNA to the final pool compared to the sample with the highest, but the amount of eDNA per sample will be equivalent.

According to the estimates of eDNA, we pooled the 60 libraries into three primary pools (see graphical representation in Figure 2). The first pool (P1) with 2  $\mu\text{g}$  total DNA (in the pool) consisted of 10 samples with an average endogenous content of 0.81% (range 0.69%–0.85%). The second pool (P2) had 4  $\mu\text{g}$  total DNA and consisted of 20 samples and an average endogenous content of 0.69% (range 0.58%–0.80%). The 30 remaining libraries were pooled into



**FIGURE 2** Pooling strategy illustration. P1 has 10 libraries with average endogenous of 0.81%. We performed two primary pools of 2 and 4 µg each that were further divided into four hybridization pools, two at 1 µg and two at 2 µg. P2 has 20 libraries with average endogenous of 0.69%. Two primary pools of 4 µg were divided into four hybridization pools of 1 µg each and two hybridizations pools of 2 µg. P3 has 30 libraries and an average endogenous of 0.49%. Two primary pools of 6 and 4 µg were distributed into six hybridization pools of 1 µg and two hybridization pools of 2 µg each. Colours represent the sequencing batch [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13300)]

the third pool (P3) of 6 µg total DNA with an average endogenous content of 0.49% (range 0.41%–0.66%) (Table 1 and Figure 3a, Table S2). Subsequently, each initial primary pool was subdivided into two (P1E1, P1E2), four (P2E1, P2E2, P2E3, P2E4) and six (P3E1, P3E2, P3E3, P3E4, P3E5, P3E6) exome capture (E) replicates each consisting of 1 µg of total DNA.

Independently, we repeated the construction of the primary pools (P1, P2 and P3), but with each having 4 µg total DNA. Each of these new primary pools was then divided into two replicates of 2 µg each (P1E3, P1E4, P2E5, P2E6, P3E7, P3E8). As a consequence of generating replicate primary pools, six of the 60 libraries were exhausted and are not present in these replicate primary pools. As a result, across all 60 samples and 18 hybridizations there are a total of 388 individual hybridization experiments (Figure 2). All details are provided in Table 1.

Each exome capture experiment consisted of two consecutive hybridizations, or dual-capture reactions as previously recommended (Hernandez-Rodriguez et al., 2018) using the SureSelect Human All Exon V6 RNA library baits from Agilent Technologies and was performed following the manufacturer's protocol with some modifications (the full protocol is available in Supporting Information), and started with either 1 or 2 µg total DNA (Table 1 and Figure 2). After

the first hybridization reaction and the subsequent PCR enrichment, we performed the second hybridization reaction with all available material. The final captured pool was amplified with indexed primers (Kircher et al., 2012), double-indexing each library within a pool, thereby tagging each library to a specific hybridization experiment. Double inline barcoded (sample specific) and double indexed (pool specific) libraries allow for multiplexing many libraries into a single pool and sequencing many pools into a single sequencing lane, even when the same sample library is present in multiple hybridization reactions. This permits the tracking of unique experiments.

For the remainder of the article when we use the word “capture” or “hybridization”, we always refer to the dual-capture or two consecutive rounds of capture hybridizations, as described above.

## 2.3 | Sequencing and mapping

Captured libraries were pooled into three sequencing batches and sequenced on a total of 3.75 lanes of a HiSeq 4000 with 2 x 100 paired-end reads: SeqBatch1 (P1E1, P2E1, P2E2, P3E1, P3E2, P3E3), SeqBatch2 (P1E2, P2E3, P2E4, P3E4, P3E5, P3E6) and SeqBatch3 (P1E3, P1E4, P2E5, P2E6, P3E6, P3E7, P3E8) (Table 1).



TABLE 1 Pooling strategy

Pool	Average eDNA content (range)	Hybridization ID	Number of pooled libraries	Total DNA (µg)	Sequencing batch
Pool 1 (P1)	0.81% (0.60%–0.85%)	P1E1	10	1	SeqBatch1
		P1E2	10	1	SeqBatch2
		P1E3	9	2	SeqBatch3
		P1E4	9	2	SeqBatch3
Pool 2 (P2)	0.69% (0.58%–0.80%)	P2E1	20	1	SeqBatch1
		P2E2	20	1	SeqBatch1
		P2E3	20	1	SeqBatch2
		P2E4	20	1	SeqBatch2
		P2E5	19	2	SeqBatch3
		P2E6	19	2	SeqBatch3
Pool 3 (P3)	0.49% (0.41%–0.66%)	P3E1	30	1	SeqBatch1
		P3E2	30	1	SeqBatch1
		P3E3	30	1	SeqBatch1
		P3E4	30	1	SeqBatch2
		P3E5	30	1	SeqBatch2
		P3E6	30	1	SeqBatch2
		P3E7	26	2	SeqBatch3
		P3E8	26	2	SeqBatch3

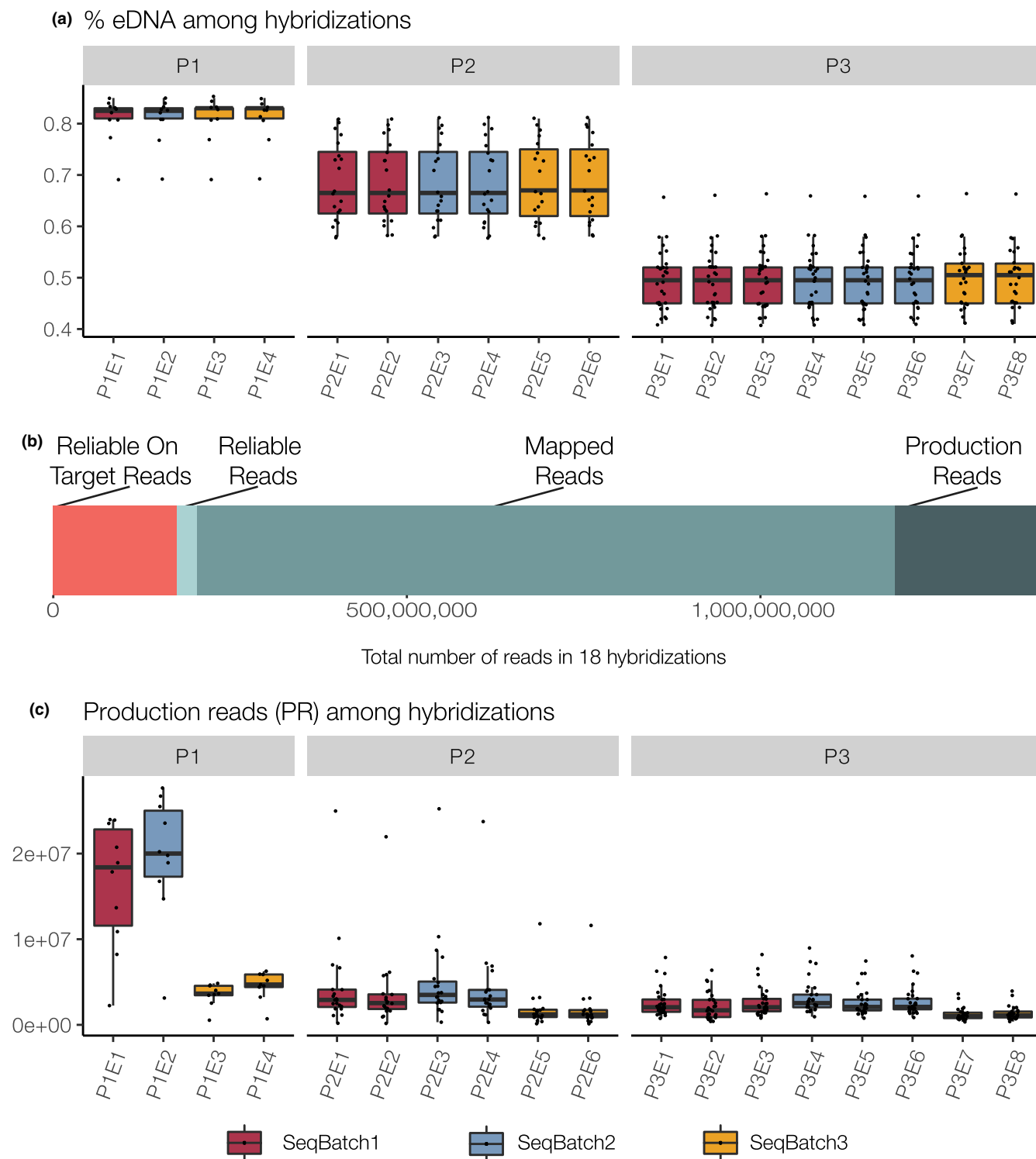
Sixty libraries were divided into three pools for capture hybridization experiments in four replicates for P1, six replicates for P2 and eight replicates for P3. Total DNA represents the starting material for each capture hybridization.

Demultiplexed FASTQ files were trimmed with Trimmomatic (version 0.36) (Bolger et al., 2014) to remove the first seven nucleotides corresponding to the in-line barcode (HEADCROP: 7), the Illumina adapters (ILLUMINACLIP:2:30:10), and bases with an average quality less than 20 (SLIDINGWINDOW:5:20). Paired-end reads were aligned to human genome Hg19 (GRCh37, Feb.2009 [GCA\_000001405.1]) using BWA (version 0.7.12) (Li & Durbin, 2009). Duplicates were removed using PicardTools (version 1.95) (<http://broadinstitute.github.io/picard/>) with MarkDuplicates option. Further filtering of the reads was carried out to discard secondary alignments and reads with mapping quality lower than 30 using samtools (version 1.5) (Li et al., 2009). The reads remaining after filtering are subsequently referred to as “reliable reads”. To retrieve the reliable reads on-target we used intersectBed from the BEDTOOLS package (version 2.22.1) (Quinlan & Hall, 2010) using exome target regions provided by Agilent. In cases where we combined sequencing data, we merged filtered bam files from different hybridizations using MergeSamFiles option from PicardTools (version 1.95) (<http://broadinstitute.github.io/picard/>). Since the merged bam files can still contain duplicates generated during library preparation, we removed duplicates and then retrieved the reliable reads on-target using the same methodology as above (Figure S3). For all previous steps, the total number of reads were counted using PicardTools (version 1.95) (<http://broadinstitute.github.io/picard/>) with CollectAlignmentSummaryMetrics option. The percentage of human contamination was estimated by using positions where modern humans and chimpanzees consistently differ. We

used previously published diversity data on high-coverage genomes from the *Pan* species (chimpanzee and bonobos) (De Manuel et al., 2016) and human diversity data from the 1000 Genomes Project (Auton et al., 2015), selecting positions where the human allele is observed at more than 98% frequency, and a different allele is observed in almost all *Pan* individuals (136 out of 138 chromosomes). Genome-wide, 5,646,707 chimpanzee-specific positions were identified. Using “samtools mpileup” (Li et al., 2009), we retrieved the number of observations of human-like alleles at these positions in the mapped reads, and estimated the human contamination as the fraction of observations for the human-like allele across all positions.

## 2.4 | Capture performance

Capture performance was evaluated by calculating the enrichment factor (EF), capture specificity (CS<sub>p</sub>), library complexity (LC), and capture sensitivity (CS) as described in Hernandez-Rodriguez et al. (2018). EF is calculated as the ratio of the number of reliable reads on-target to the total reads sequenced divided by the fraction of the target space (64 Mb) to the genome size (~3 Gb). CS<sub>p</sub> is defined as the ratio of reliable on-target reads to the total number of reliable reads. LC is defined as the number of reliable reads divided by the total number of mapped reads (containing duplicated reads). Capture sensitivity (CS) is defined as the number of target regions with an average coverage of at least one (DP1) - but also four (DP4), 10 (DP10), 20 (DP20) or 50 (DP50) - divided by the total number of



**FIGURE 3** Capture performance and sequencing. (a) Percentage of eDNA among hybridizations, structured by pools (P1, P2 and P3). (b) Sequencing stats across all samples for the 18 hybridizations in 3.75 HiSeq 4000 lanes. (c) Distribution of production reads across 18 hybridizations. The colours red, blue and yellow found in the box plots for figure (a) and (c) denote the sequencing batch to which each hybridization was assigned. In the boxplots, lower and upper hinges correspond to first and third quartiles and the lower and upper whiskers extend to the smallest or largest value no further than 1.5 times the interquartile range (distance between the first and third quartile) [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

target regions provided by the manufacturer ( $n = 243,190$ ). To calculate the average coverage of the target regions we used “samtools” (version 1.5) with the option “bedcov” (Li et al., 2009).

To generate molecular complexity or library complexity curves (MC), we used the subsampling without replacement strategy implemented in Preseq software (version 2.0.7) with *c\_curve* option

(<http://smithlabresearch.org/software/preseq/>) from the bam files without removing duplicates. MCs were sequentially estimated by adding the production reads, i.e., raw reads produced by sequencing, from additional hybridizations, one at a time until all hybridizations from the same library were merged (schematic representation in Figure S4).

Correlation coefficients among all pairs of study variables were estimated. Spearman's rho (`cor.test()`, `method = "sp"`) from R stats package) was estimated when comparing two numeric variables. Among two categorical variables, we estimated Cramér's V, derived from a chi-squared test (`chisq.test()` from R stats package). When comparing a numeric and categorical variable we took the square root of the R-squared statistic derived from a univariate linear model (`lm()` from R stats package) with a rank normal transformation (`rntransform()` modified from the GenABEL package to randomly split tied values) on the dependent, numerical values. In addition, univariate and multivariate type I hierarchical analysis of variances (ANOVA; `anova()` from R stats package) were performed to estimate the variance explained (or eta-squared) each experimental variable had on performance summary statistics (number of unique reads, reliable reads, EF, LC, CS and CSp). We downsampled libraries to 1,500,000 reads ( $n = 274$ ) to remove production reads as a confounding factor. Each performance statistic was rank normal transformed with ties being randomly split to ensure normality of the dependent variable. Univariate analysis focused on the effect that subspecies, geographic sampling site, total DNA concentration, endogenous DNA concentration, percentage of endogenous DNA, average fragment length, pool, amount of DNA in a hybridization, hybridization and sequencing batch had on each performance statistic. A multivariate model was built to conform with experimental (hierarchical) order, such that each dependent variable (performance summary statistic, CS at DP1) was explained by ~subspecies + site + % eDNA + average fragment size + pool + amount of DNA + hybridization + sequencing batch + error. Again, the variance explained by each independent variable was summarized by computing the eta-square statistic derived from the sums of squares for each variable using a type I hierarchical ANOVA. All statistical analyses were performed in R (version 3.5.2) (R Core Team, 2018).

### 3 | RESULTS

#### 3.1 | Sample description

Samples were collected from 15 different PanAf sites distributed across the entire range of chimpanzees in Africa (Figure 1a and Table S1). The 302 screened samples had an average eDNA of 1.49%, ranging from 0% to 47.75% (Figure 1b, Figure S1A and Table S1) with 70.2% of the samples below 1% eDNA, according to qPCR estimates (Figure 1c). The average fragment length for screened samples was 3,479.94 base pairs (bp), ranging from 72 to 17,966 bp (Figure S1B and Table S1).

We observed variation on the average endogenous content among geographical sites (Figure 1b), and also variation on fragment length among geographical sites (Figure S1B). For example, samples collected in a specific location such as Campo Ma'an (Cameroon) had an average eDNA of 0.02%, an extremely low value compared to the average of all sites of 1.49%. On the other hand, some sites such as Ngogo (Uganda) had samples with higher than average eDNA (6.95%) (Table S3). This might be explained by the influence of weather, humidity and temperature on DNA preservation and bacterial growth in the faecal sample before collection as well as a product of sample age and quality of sampling conditions (Brinkman et al., 2010; Goossens et al., 2000; Harestad & Bunnell, 1987; King et al., 2018; Nsubuga et al., 2004; Wedrowicz et al., 2013).

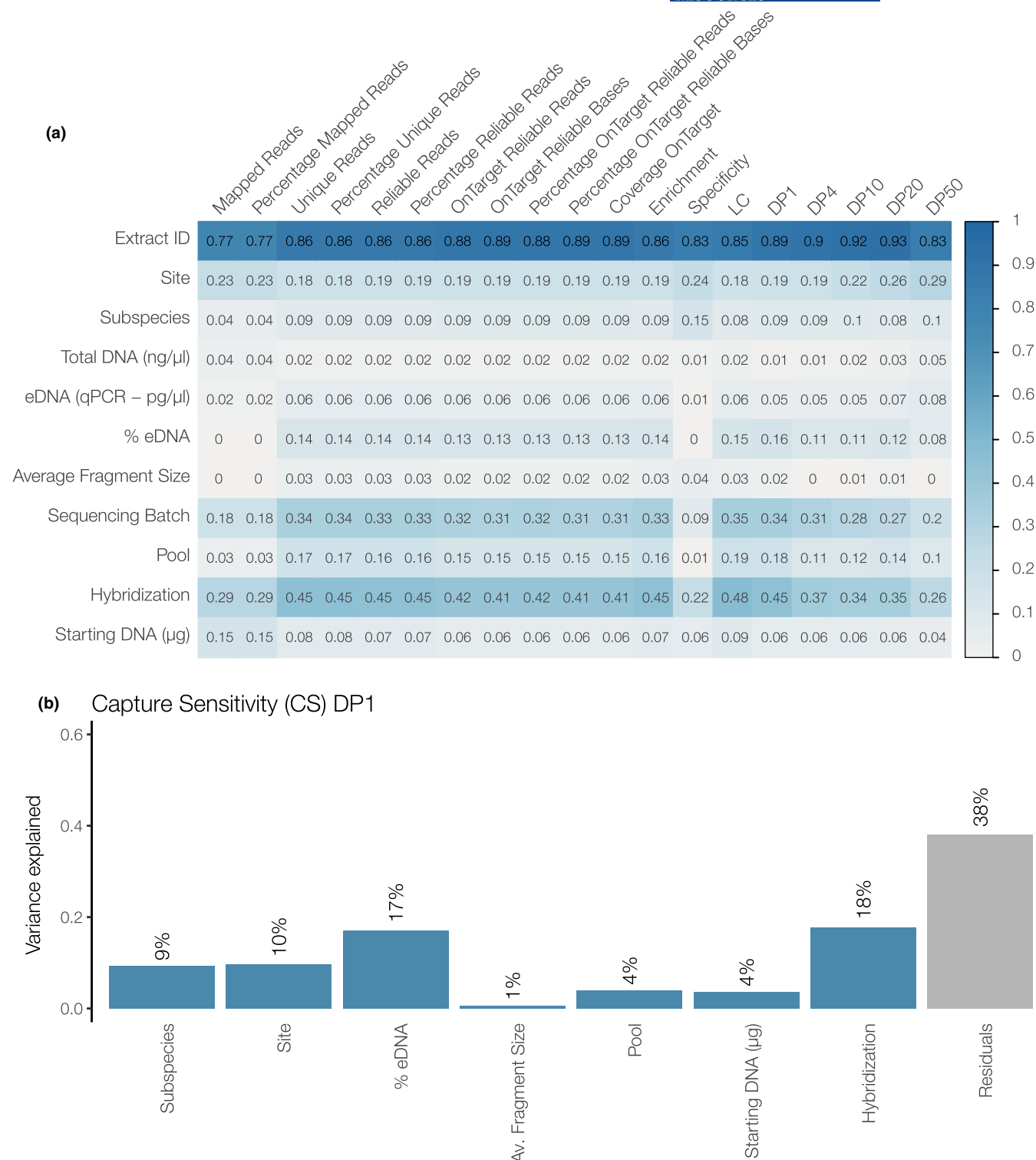
A total of 60 samples with a mean percentage of endogenous content of 0.58% (ranging from 0.41% to 0.85%), and with a median human contamination of 0.0875% (ranging from 0.04% to 7.50%) from all four chimpanzee subspecies and 14 geographic sites were carried forward into target capture enrichment experiments (Table S2). After double-inline-barcoded library production, the 60 samples were placed into three pools with 10, 20 and 30 samples each (Figure 2). Samples were divided into pools based on their percentage of endogenous content, such that those samples with higher levels of percentage of endogenous content were in P1 with 10 samples (mean = 0.81) and those with the smallest were in P3 with 30 samples (mean = 0.49; P2 mean = 0.69) (Figure 3a). As such, the percentage of endogenous DNA was highly structured among the three pools, explaining 81% of the variation in eDNA (univariate linear model using rank normal transformed % eDNA;  $p$ -value =  $2.05 \times 10^{-91}$ ) (Figure S5A).

#### 3.2 | Read summary statistics and capture performance

As illustrated in Figure 3b across a total of 18 hybridization experiments sequenced we obtained ~1.40 billion reads distributed among three pools. Of those, ~1.19 billion were mapped reads (85.19%), with ~203 million reads being considered duplicate-free, reliable reads (14.6%). After removing off-target reads, we obtained a total of ~174 million on-target-reliable reads (12.48%) (Table S4, Figure S3A). However, on average each hybridization experiment yielded an average of 17.35% on-target-reliable reads, with a range of 4.15% in our earliest experiments to 34.85% in our later experiments (Table S5). The observed high levels of duplicates are a consequence of the low endogenous content of the samples and the exhaustion of library complexity during sequencing; we elaborate on outcome and improvements below.

The ~1.40 billion reads were not equally distributed among the three pools (production reads explained by pools;  $r^2 = 0.41$ ,  $p$ -value =  $3.24 \times 10^{-16}$ ) or 18 hybridizations ( $r^2 = 0.62$ ,  $p$ -value =  $2.59 \times 10^{-30}$ ). In fact, two hybridizations of P1 (P1E1, P1E2) were sequenced to an average depth of 18 million reads, while all other hybridizations had an average depth of 3 million reads (Figure 3c).





**FIGURE 4** Analysis of variance. (a) Estimated variance explained from univariate linear models after rank normal transforming each performance summary statistic (columns). LC stands for library complexity and DP describes read depth at different cutoffs (1, 4, 10, 20 and 50 reads) (b) Multivariate type I ANOVA of the experimental variables affecting capture sensitivity (CS) at depth 1. Both models are built downsampling libraries to 1,500,000 reads [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

This very deep sequencing in P1E1 and P1E2 led to a point where the library complexity was exhausted, leading to the sequencing of a high number of PCR duplicates (Figure S3A, S3B and Table S5). We therefore reduced subsequent sequencing efforts, as discussed in

“Optimization of required production reads”, for the remaining replicate hybridizations.

All capture performance summary statistics (Table S4), with the exception of capture specificity (CS<sub>p</sub>), were strongly correlated

with the number of production reads acquired (median correlation coefficient = 0.422, CI = 0.03–0.93; Figure S5A, Table S6). Given this, and also because of the distinct difference in the number of production reads between P1E1 and P1E2 and all other hybridizations we downsampled all experiments to 1.5 million production reads, retaining only those 274 sample/hybridization experiments with 1.5 million production reads, and re-estimated all capture performance summary statistics (Figure S5B, Table S7 and S8). The effect each experimental variable had on performance was estimated in a univariate linear model after rank normal transforming each summary statistic (Figure 4a). We observed a near uniformity in the variance explained by each experimental variable across each performance statistics. In short, the average, ranked order of variance explained by each explanatory variable were sample (86.50%), hybridization (38.72%), sequencing batch (28.78%), site (20.5%), pool (13%), % endogenous DNA (11%), subspecies (8.85%), starting DNA amount (7.35%), endogenous DNA concentration (5.14%), average fragmentation size (2.12%), and total DNA concentration (2.07%). Given these observations we may conclude that variation in hybridization and sequencing are crucial to performance. However, sample quality and starting material varied among our hybridizations and sequencing batches. These tendencies can be observed in Figure 5a–c. This is accounted for in a multivariate linear model followed by a decomposition of the variance in a type I hierarchical analysis of variance (ANOVA). To do so we fit a linear model ordered by experimental choices, as described in Materials and methods, to explain capture sensitivity (CS) at DP1 which is being used here as an example of capture performance. This model indicates that hybridization explains, on average, an attenuated 17.80% of the variation in performance, followed by percentage of endogenous content (17.11%), site (9.62%), subspecies (9.26%), pool (3.92%) and then the amount of DNA in the hybridization (3.58%) (Figure 4b). Results for all other performance summary statistics mirror those for CS at DP1 and can be seen in Figure S6.

### 3.3 | Relevance of equi-endogenous pools

The observations of Hernandez-Rodriguez et al. (2018) and White et al. (2019) suggest that pooling libraries by eDNA concentration (in equi-endogenous pools) prior to hybridization capture should reduce or remove the effect of variation in eDNA across samples on targeted capture sequencing performance. Indeed, eDNA did not have a major influence on production reads or on-target reads, although a slightly positive trend was observed in some hybridizations of P2 (Figure S7). Without equi-endogenous pooling, it is

expected that samples with higher eDNA would accumulate more on-target reads than other samples with lower eDNA as observed by Hernandez-Rodriguez et al. (2018). The reason why in P2 we found some outliers might be traced to both pipetting variations and inaccurate endogenous measurements from qPCR values due to the presence of inhibitors (Morin et al., 2001). Avoiding outliers is extremely important in limiting variability within a pool. For example, sample N183-5 accumulated 29.4% of total raw reads in P2, when a value 5% (1/20 of 100%) was expected (Figure S8).

### 3.4 | Impact of amount of starting DNA for capture on library complexity

One major decision when performing capture experiments is the amount of starting DNA in the pool. In 12 hybridizations we used the manufacturer's suggested amount of starting material which was 1 µg for each pool. For the last two hybridizations of each pool (a total of six hybridizations), we doubled the starting material, up to 2 µg of pooled libraries (Table 1). With this approach we aimed to test the effect on the final LC when doubling the amount of DNA and to determine how much DNA should be used for faecal capture experiments. We observed an average increase of 2.8-fold in LC for experiments using 2 µg of total DNA in the hybridization relative to those using 1 µg (Figure S3B). However, given that production reads also vary between these two conditions, we downsampled the data to 1,500,000 reads per library. After this correction we still observed 2-fold higher LC when starting the experiments with 2 µg of total DNA in all pools (Figure 5d).

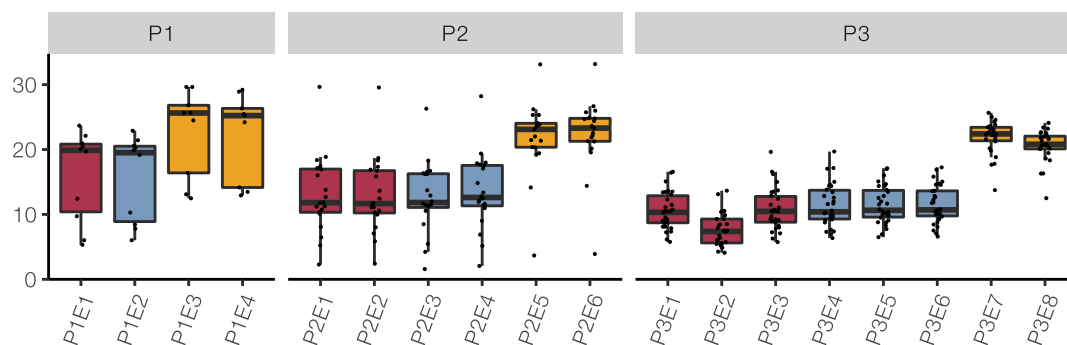
Molecular complexity, as influenced by the amount total DNA in a hybridization, was further investigated by evaluating the relationship between MC and production reads in an MC curve analysis. The MC curve for each hybridization was obtained by subsampling without replacement their reads. The results supported the conclusion above: increasing the amount of total DNA in the hybridization increased the MC (Figure S9). Therefore, whenever there is sufficient library available, it is advisable to start with 2 µg rather than 1 µg.

### 3.5 | Molecular complexity and capture sensitivity

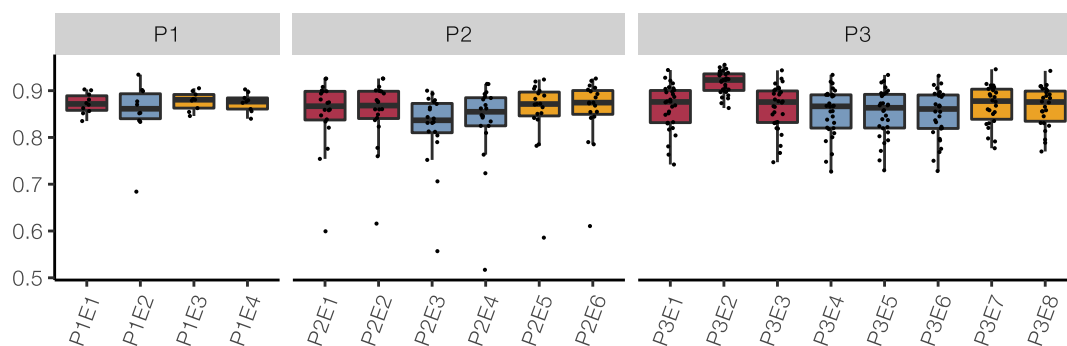
One of the critical aspects to increase coverage is to acquire as many unique on-target reads as possible without exhausting the library's molecular complexity. We applied a subsampling without replacement method to assess how many mapped reads are unique after incrementally adding production reads from replicate hybridizations.

**FIGURE 5** Summary stats after downsampling to 1,500,000 reads. (a) Enrichment factor and (d) Capture Specificity (c) Capture Sensitivity at depth 1 for the 18 hybridizations in P1, P2 and P3; colours illustrate sequencing batch. (d) Library complexity contrasting the amount of starting DNA (1 or 2 µg) in downsampled data and structured by pools (P1, Pool1; P2, Pool2; P3, Pool3). See Figure 2 for more details on pools. In the boxplots, lower and upper hinges correspond to first and third quartiles and the lower and upper whiskers extend to the smallest or largest value no further than 1.5 times the interquartile range (distance between the first and third quartile) [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13300)]

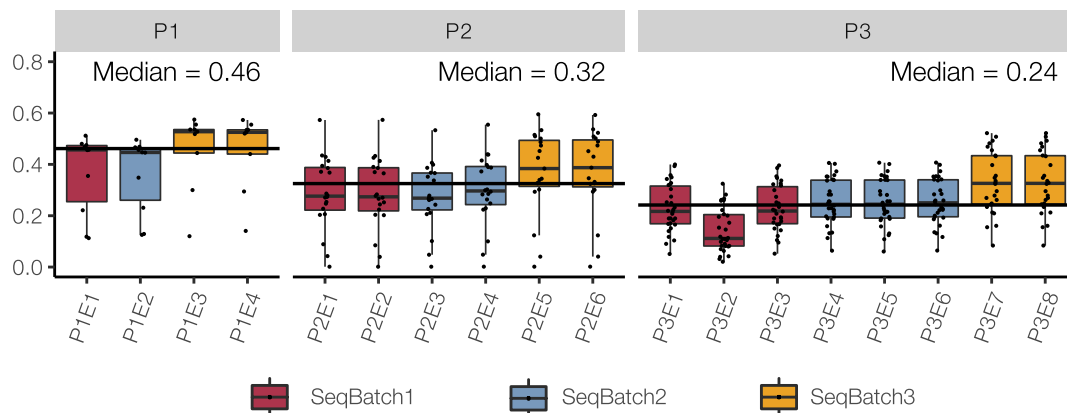
(a) Enrichment Factor (ER)



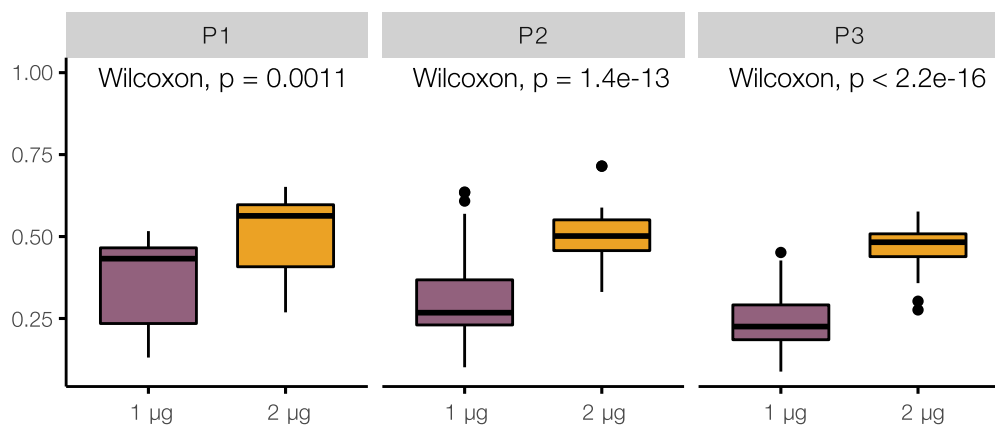
(b) Capture Specificity (CSp)



(c) Sensitivity (CS) at Depth 1



(d) Library Complexity (LC)



In principle, molecular complexity curves that plateau quickly are derived from low complexity libraries, and conversely high complexity libraries may not reach plateau. Thereby the plateau indicates when there are no new unique reads to be sampled or sequenced (see Figure S4 for a schematic representation).

We performed the analysis of molecular complexity in libraries belonging to P3 since more hybridization replicates were available (eight in total) for 30 libraries. We found that for the majority of the libraries, performing additional hybridizations increased the number of unique reads retrieved (Figure S10, example library N259-5). However, there were libraries that quickly hit exhaustion where performing additional hybridizations would add little extra information (Figure S10, example library Kay2-32). Overall, by performing additional hybridizations, it was possible to retrieve new unique reads and thus increase the final coverage (Figure 6a), because libraries themselves were not exhausted but merely their hybridization-captured molecules reached exhaustion.

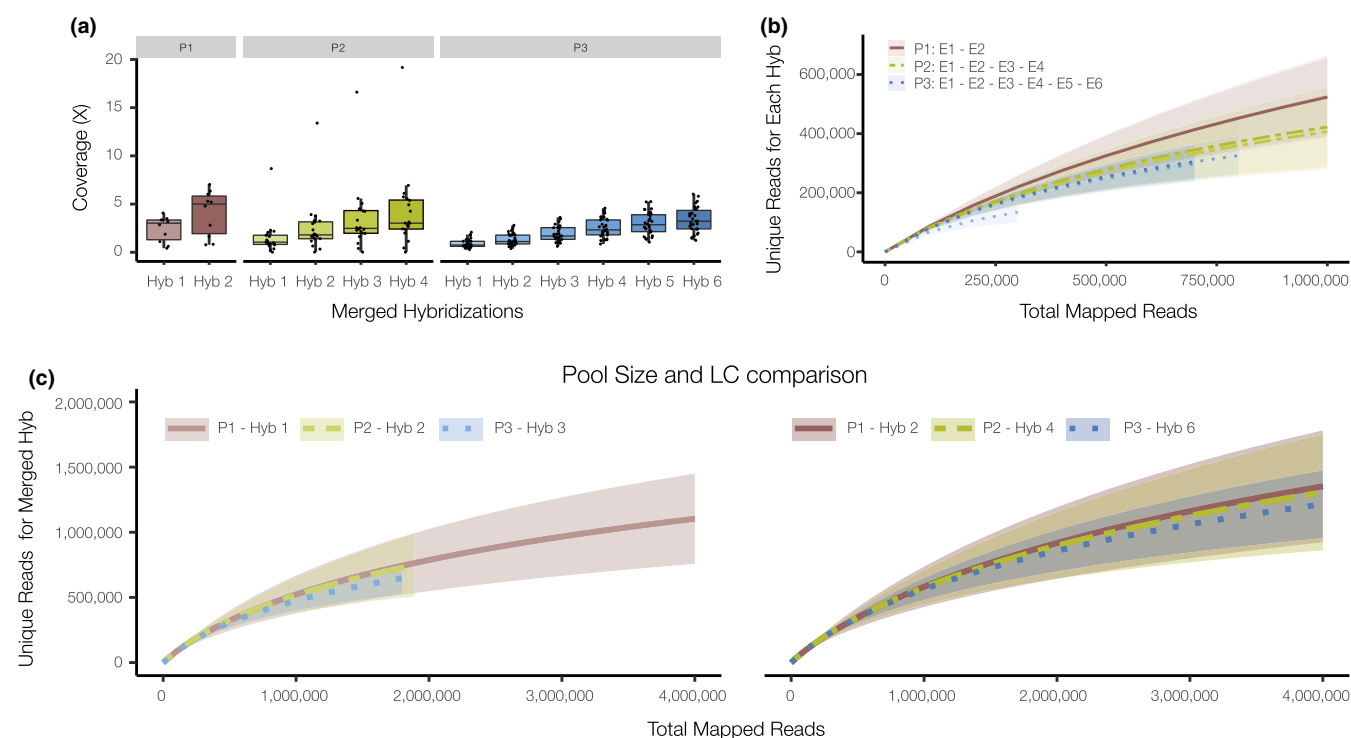
Following the same strategy, we calculated the sensitivity in P1, P2 and P3 (4, 6 and 8 replicates respectively). After cumulatively adding data from replicate hybridizations we covered 85.57% in P1 (95% CI: 74.78–96.36%), 76.23% in P2 (95% CI: 64.55–87.91%) and 79.83% in P3 (95% CI: 74.44–85.22%) on average of the target space, with at least 1 read (Figure S11). Interestingly, no sample covered 100% of target space. Looking carefully into this, we observed

that precisely the same 3804 regions (1.54%) were never covered in any replicate hybridizations, suggesting that some regions are either difficult to capture (Kong et al., 2018) or are too divergent between *Homo* and *Pan* to either capture, or map, these particular sequences (Figure S12).

For deeper coverage of at least four or 10 reads, we still observed a positive progression, with each additional hybridization increasing coverage, indicating that additional hybridizations would result in an increase of the proportion of the genome covered at these depths as well (Figure S11).

### 3.6 | Optimization of required production reads

Assessing the amount of sequencing needed is one of the major decisions when planning an experiment. As a result of the low eDNA content of most faecal samples, derived libraries can easily reach saturation (i.e., high levels of duplicated reads). Therefore, sequencing depth should be carefully calculated. Without previous knowledge, we sequenced the first two hybridizations for P1, the first four hybridizations for P2, and the first six hybridizations for P3 in three lanes of a HiSeq 4000. For P1 only ~6% and for P2 and P3 only ~13% of production reads were unique reads (Table S5), indicative of high levels of PCR duplicates due to library exhaustion. To



**FIGURE 6** Analysis of coverage and LC with hybridizations done with 1  $\mu$ g. (a) Coverage after merging data from additional hybridizations with up to 2, 4 and 6 for P1, P2 and P3. (b) Comparison of average LC curves of individual hybridizations belonging to pools with different size. Each line is the average of libraries within each hybridization and the surrounding area is the standard deviation. (c) Two examples comparing the effect of pool size on the average LC curves from merged hybridization: P1 (10 samples) - 1 hybridization, P2 (20 samples) - 2 hybridizations and P3 (30 samples) - 3 hybridizations; and P1 (10 samples) - 2 hybridizations, P2 (20 samples) - 4 hybridizations and P3 (30 samples) - 6 hybridizations. Sample Lib1-6D in P2 was removed from the analysis due to low coverage [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

avoid oversequencing in subsequent experiments, we set an arbitrary threshold to recover approximately 20% of the “informative” data (unique reads) available in a hybridization experiment. This 20% threshold was chosen to maximize the output cost ratio given the diminishing returns on further sequencing (Figure S13). Using the data from SeqBatch 1 and 2, we estimated that on average, for samples with less than 1% eDNA, we would sequence at most two million mapped reads per library (Figure S13). Given that 80% of reads mapped to the genome in these experiments, we estimated that we would need to sequence at most 2.5 million production reads per library (Table S5).

To test these estimates, we sequenced the remaining hybridizations (P1E3, P1E4, P2E5, P2E6, P3E7, P3E8) in three-fourths of a HiSeq 4000 lane. The number of average production reads obtained were 3.5, 2.0 and 1.5 million for libraries in hybridizations from P1, P2, and P3, respectively. On average ~38% (range: 8.09%–50.81%) of reads were unique reads in all pools (Figure S14). We note that these values exceeded what we observed in the previous hybridization experiments. This outcome we attribute to the increase in starting material (2 µg), also used in these experiments, as noted above.

### 3.7 | Pooling strategy

Choosing how many samples to pool is a difficult decision, since little is known on how the pool size will affect the final molecular complexity. Taking advantage of our pooling strategy (Figure 2), we assessed the effect of size on the average library complexity for all samples within each hybridization with a subsampling without replacement strategy.

When only a single hybridization was performed, a single library within a pool of 10, 20 or 30 would, on average, result in a similar number of unique molecules (Figure 6b, Figure S15). However, there is a tendency for samples in smaller pools (P1) to perform better than those in larger pools. This could be explained by our experimental design, where samples with higher eDNA content are in smaller pools. However, let us address this possibility here. Using CS as an example summary statistic, we observed that CS is higher for pools with smaller numbers of samples in them (Figure 5c). Given median estimates, a pool of 10 libraries (median CS = 0.46) had 1.44-fold higher CS than a pool of 20 libraries (median CS = 0.32), and 1.92-fold higher than a pool of 30 libraries (median CS = 0.24). Between a pool of 20 and a pool of 30, the ratio was 1.33-fold (Figure 5c and Figure S16). If we remove the effect of having a variable number of production reads across experiments by downsampling, this observation still remains (Figure S17). That is, smaller pools do have higher CS estimates, and pools linearly account for 18% of the variation in CS (univariate ANOVA,  $p$ -value =  $3.47 \times 10^{-12}$  [Figure 4a]). Finally, if we correct for all experimental variables with a multivariate analysis, as done above, we show that “Pool” only accounts for 4% of the variation in CS (Figure 4b), but the effect of pool size remains significant (multivariate ANOVA,  $p$ -value =  $2.7 \times 10^{-4}$ ; Figure S17). However, this effect on CS attenuates with additional hybridizations

(4, 6 and 8, for P1, P2 and P3, respectively) for the same pool (Figure S18). Moreover, a similar outcome can be observed when comparing the effect of pool size on LC. After sequentially adding data from replicate hybridizations in each pool (see Figure S4 for a schematic representation), we can acquire the same number of unique reliable reads (Figure 6c, Supporting Information S17).

## 4 | DISCUSSION

Capturing host DNA from faecal samples is a challenging endeavour. Previous work has shown that the retrieval of genomic data from faecal samples by target enrichment methodologies is a feasible and powerful tool for conservation and evolutionary studies (Perry, 2014; Snyder-Mackler et al., 2016). However, obtaining good quality and quantity DNA from faecal samples is not always possible. Because of that, many studies have characterized the technical difficulties of capturing DNA from noninvasive samples and proposed different strategies (Hernandez-Rodriguez et al., 2018; van der Valk et al., 2017; White et al., 2019). Van der Valk et al. (2017) captured the whole mitochondrial genome but no autosomal regions, and describe the biases introduced during capture such as DNA fragment size, jumping PCR and divergence between bait and target species. The study performed by Hernandez-Rodriguez et al. (2018) systematically analysed the capture performance and library complexity. While they described that pooling different libraries into the same hybridization is feasible, they did not discuss how many of them should be pooled. Also, they concluded that performing multiple libraries from the same extract or even from different extracts from the same sample can increase the final complexity. Finally, they recommended performing two capture rounds for the same library. On the other hand, White et al. (2019) suggested to do only one capture round, at least when eDNA is higher than 2%–3%, stressing the importance of pooling libraries as well as taking into consideration the eDNA content, as first proposed by Hernandez-Rodriguez et al.

The present study addresses these gaps left unexplored by the previous studies. We focused our analysis on a representative set of samples with very low proportions of endogenous content (<1%) as are often found in the field. After screening 302 samples, we found that up to 70% of samples are below this threshold, similar to what was already described (White et al., 2019). Hence, if time and economic reasons hinder the ability to collect and select the best samples, the only available one(s) might have low eDNA. This may be a common situation when using historical samples, aiming for a large sample size, or if an interesting sampling location is particularly challenging in terms of low eDNA (such as Campo Ma'an, Figure 1b).

For these reasons, it is of utmost importance to characterize ways to maximize the amount of data to be recovered from these types of samples. In this regard, we have extensively evaluated how to increase library complexity without doing more extractions or library preparations from the same sample, how many libraries to pool together, and how much starting amount of DNA should be used in a capture, as well as the impact of endogenous content for pooling.

Consistent with previous findings (Hernandez-Rodriguez et al., 2018; White et al., 2019), we determined that assessing the endogenous content of faecal samples and pooling them equi-endogenously is a practical way to equally distribute raw reads between samples. Importantly, the correct estimation of the proportion of eDNA is key for the success of this method. Thus, we recommend the usage of shotgun sequencing (Hernandez-Rodriguez et al., 2018) rather than qPCR estimates, since the later can easily fluctuate due to the presence of inhibitors (Morin et al., 2001).

In regard to the performance of target capture sequencing experiments, gaining new unique reads is crucial to reach higher sensitivity, which is a good predictor of capture success. Here, we have established an approach to obtain new unique reads using the same prepared libraries. Since it is mainly during capture experiments when the molecular diversity is reduced, we propose to perform additional hybridizations from the same library so the final coverage can reach higher values. If the library complexity is already very low, the only solution is to re-extract DNA or prepare a new library from the same sample (Hernandez-Rodriguez et al., 2018).

We observed a better performance (MC and CS) in small pools, when evaluating initial results derived from the entire data set. However, after correcting for other variables that differ among pools, the effect is attenuated and can only explain ~4% of the variance, an effect that may be largely negligible for most studies. Moreover, performing additional hybridizations can also compensate for this effect. Therefore, we do not conclude, based on this data, that pool size is a major contributor to performance. However, in cases where libraries have small proportions of eDNA, we would advocate for the reduction of the number of samples per pool so that pipetting volumes may remain larger, and as a consequence variability due to pipetting error may be reduced. Otherwise when the eDNA proportion is not a limiting factor, pooling more libraries together and performing additional hybridizations can be a good strategy.

It is worth noting that without taking into consideration individual sample quality and the amount of starting material used, one of the most influential variables on the performance of target capture enrichment experiments is the hybridization experiment itself. After accounting for all other variables, it still explains 18% of the variation. This is due to the technical complexity and variability inherent to these experiments. Careful equipment optimization, material selection, preparation and experience will aid in minimizing this variation, although it is likely to remain a sensitive experiment that requires diligence.

Finally, we have illustrated that a sequencing effort of exome-captured faecal samples with low eDNA (< 1%) should be set at ~3 million reads per library in a pool to avoid exhausting the molecular complexity. We have benefited from the usage of double-bar-coded and double-indexed libraries to multiplex many samples in a single sequencing lane. This becomes a great advantage because we can utilize high throughput sequencing technologies at a lower price per read.

To summarize, when starting a project involving faecal samples, we recommend screening the set of samples based on quantity

and quality of DNA extracted. If having related or identical individuals in the study is to be avoided, microsatellite genotyping could be an option, also helping to discard samples with a high amount of PCR inhibitors. Further selection of samples should be based on the proportion of eDNA; we recommend using shotgun sequencing from the prepared libraries. Performing re-extractions of the most valuable samples and preparing replicate libraries from each extract can help increase the final molecular complexity. As we have shown here, another approach to achieve higher molecular complexity is based on conducting additional hybridizations of the captured libraries, always pooling libraries in an equi-endogenous manner, and starting with more library material than the standard protocol suggests. Finally, we suggest not sequencing the captured libraries very deeply, since their molecular complexity is already very low and oversequencing can result in rapidly depleting the economic feasibility of the experiment.

In the study presented here, we thoroughly explored approaches to increase the molecular diversity and capture sensitivity and hence the final coverage of exome captured faecal samples with extremely low endogenous content in an attempt to help laboratories facing the challenges of working with noninvasive samples.

## ACKNOWLEDGEMENTS

We would like to thank Linda Vigilant, Christophe Boesch and Marco Telford for helpful discussion and Roland Schroeder, Alan Riedel and Katharina Madl, for guidance and assistance in the laboratory. We thank Emmanuel Dilambaka, Devla Dowd, Annemarie Goedmakers, Vincent Lapeyre, Vera Leinert, Mizuki Murai, Emmanuelle Normand, Robinson Orume, Alexander Tickle, Els Ton, Joost van Schijndel, Sergio Marrocoli, Amelia Meier, Volker Sommer, Martijn Ter Heegde, Nadege Wangue Njomen, Joshua M Linder, Hilde Vanleeuwe, Jean Claude Dengui, Paul Telfer and Yasmin Moebius for assistance in field site coordination and sample collection. C.F. is supported by the "La Caixa" doctoral fellowship program. M.A.E. is supported by an FPI (Formación de Personal Investigador) PRE2018-083966 from Ministerio de Ciencia, Universidades e Investigación. The Pan African Programme: The Cultured Chimpanzee (PanAf) is generously funded by the Max Planck Society, the Max Planck Society Innovation Fund and the Heinz L. Krekeler Foundation. E.L. is supported by CGL2017-82654-P (MINECO/FEDER,UE). M.K. is supported by "la Caixa" Foundation (ID 100010434), fellowship code LCF/BQ/PR19/11700002. T.M.-B. is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 864203), BFU2017-86471-P (MINECO/FEDER, UE), "Unidad de Excelencia María de Maeztu", funded by the AEI (CEX2018-000792-M), 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). We thank the following government agencies for their support in conducting field research in their countries: Ministère de la Recherche



Scientifique et de l'Innovation, Cameroon, Ministère des Forêts et de la Faune, Cameroon, Ministère des Eaux et Forêts, Cote d'Ivoire, Ministère de l'Enseignement Supérieur et de la Recherche Scientifique, Cote d'Ivoire, Agence Nationale des Parcs Nationaux, Gabon, Centre National de la Recherche Scientifique (CENAREST), Gabon, Société Equatoriale d'Exploitation Forestière (SEEF), Gabon, Ministère de l'Agriculture de l'Elevage et des Eaux et Forêts, Guinea, Instituto da Biodiversidade e das Áreas Protegidas (IBAP), Guinea Bissau, Ministro da Agricultura e Desenvolvimento Rural, Guinea-Bissau, Forestry Development Authority, Liberia, National Park Service, Nigeria, Ministère de l'Economie Forestière, R-Congo, Ministère de la Recherche Scientifique et Technologique, R-Congo, Direction des Eaux, Forêts et Chasses, Senegal, Tanzania Commission for Science and Technology, Tanzania, Tanzania Wildlife Research Institute, Tanzania, Uganda National Council for Science and Technology (UNCST), Uganda, Uganda Wildlife Authority, Uganda, National Forestry Authority, Uganda.

## AUTHOR CONTRIBUTIONS

C.F., T.M.B., D.A.H., and E.L. designed the study. M.A., and H.S.K. direct the Pan African Programme: The Cultured Chimpanzee. M.A., and H.S.K. obtained funding for the project. M.A., P.D., A.A., S.A., E.A.A., M.B., G.B., T.D., M.E.N., A.C.G., J.H., P.K., A.K.K., M.K., K.L., J.L., G.M., L.J.O., A.P., M.M.R., F.S., V.V., and R.M.W. supervised, conducted fieldwork and collected samples. C.F., M.A.E., E.L., J.L., and M.A. performed experiments. C.F., and D.A.H. performed the analysis. M.A.E., M.K., D.A.H., T.M.B., and E.L. provided analytical support. C.F. wrote the manuscript with input from all authors.

## DATA AVAILABILITY STATEMENT

All raw sequencing data have been deposited at ENA and are available under the accession code PRJEB37173 (<http://www.ebi.ac.uk/ena/data/view/PRJEB37173>).

## ORCID

Claudia Fontseré  <https://orcid.org/0000-0003-2233-6026>

Marina Alvarez-Estape  <https://orcid.org/0000-0001-8242-1092>

Mattia Bessone  <https://orcid.org/0000-0002-8066-6413>

Esther Lizano  <https://orcid.org/0000-0003-3304-9807>

## 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., Rabanal, L. I., Schubert, G., Mettke, E., Boesch, C., Robbins, M. M., & Vigilant, L. (2011). Non-invasive genetic monitoring of wild central chimpanzees. *PLoS One*, 6(3), e14761. <https://doi.org/10.1371/journal.pone.0014761>
- Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., & Schloss, J. A. (2015). A global reference for human genetic variation. *Nature*, 526, 68–74. <https://doi.org/10.1038/nature15393>
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Brinkman, T. J., Schwartz, M. K., Person, D. K., Pilgrim, K. L., & Hundertmark, K. J. (2010). Effects of time and rainfall on PCR success using DNA extracted from deer fecal pellets. *Conservation Genetics*, 11(4), 1547–1552. <https://doi.org/10.1007/s10592-009-9928-7>
- Burbano, H. A., Hodges, E., Green, R. E., Briggs, A. W., Krause, J., Meyer, M., Good, J. M., Maricic, T., Johnson, P. L. F., Xuan, Z., Rooks, M., Bhattacharjee, A., Brizuela, L., Albert, F. W., de la Rasilha, M., Fortea, J., Rosas, A., Lachmann, M., Hannon, G. J., & Paabo, S. (2010). Targeted investigation of the neandertal genome by array-based sequence capture. *Science*, 328(5979), 723–725. <https://doi.org/10.1126/science.1188046>
- 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., Buenostro, J. D., Valdiosera, C., Schroeder, H., Allentoft, M. E., Sikora, M., Rasmussen, M., Gravel, S., Guillén, S., Nekhrizov, G., Leshtakov, K., Dimitrova, D., Theodossiev, N., Pettener, D., Luiselli, D., Sandoval, K., Moreno-Estrada, A., Li, Y., Wang, J., ... Bustamante, C. D. (2013). Pulling out the 1%: Whole-Genome capture for the targeted enrichment of ancient dna sequencing libraries. *American Journal of Human Genetics*, 93(5), 852–864. <https://doi.org/10.1016/j.ajhg.2013.10.002>
- Csilléry, K., Johnson, T., Beraldi, D., Clutton-Brock, T., Coltman, D., Hansson, B., Spong, G., & Pemberton, J. M. (2006). Performance of marker-based relatedness estimators in natural populations of outbred vertebrates. *Genetics*, 173(4), 2091–2101. <https://doi.org/10.1534/genetics.106.057331>
- De Barba, M., Waits, L. P., Genovesi, P., Randi, E., Chirichella, R., & Cetto, E. (2010). Comparing opportunistic and systematic sampling methods for non-invasive genetic monitoring of a small translocated brown bear population. *Journal of Applied Ecology*, 47(1), 172–181. <https://doi.org/10.1111/j.1365-2664.2009.01752.x>
- de Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., Hernandez-Rodriguez, J., Dupanloup, I., Lao, O., Hallast, P., Schmidt, J. M., Heredia-Genestar, J. M., Benazzo, A., Barbujani, G., Peter, B. M., Kuderna, L. F. K., Casals, F., Angedakin, S., Arandjelovic, M., ... Marques-Bonet, T. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, 354(6311), 477–481. <https://doi.org/10.1126/science.aag2602>
- Ferreira, C. M., Sabino-Marques, H., Barbosa, S., Costa, P., Encarnação, C., Alpizar-Jara, R., Pita, R., Beja, P., Mira, A., Searle, J. B., Paupério, J., & Alves, P. C. (2018). Genetic non-invasive sampling (gNIS) as a cost-effective tool for monitoring elusive small mammals. *European Journal of Wildlife Research*, 64(4). <https://doi.org/10.1007/s10344-018-1188-8>
- Fickel, J., Lieckfeldt, D., Ratanakorn, P., & Pitra, C. (2007). Distribution of haplotypes and microsatellite alleles among Asian elephants (*Elephas maximus*) in Thailand. *European Journal of Wildlife Research*, 53(4), 298–303. <https://doi.org/10.1007/s10344-007-0099-x>
- Fischer, A., Wiebe, V., Pääbo, S., & Przeworski, M. (2004). Evidence for a complex demographic history of chimpanzees. *Molecular Biology and Evolution*, 21(5), 799–808. <https://doi.org/10.1093/molbev/msh083>
- Goossens, B., Chikhi, L., Utami, S. S., De Ruiter, J., & Bruford, M. W. (2000). A multi-samples, multi-extracts approach for microsatellite analysis of faecal samples in an arboreal ape. *Conservation Genetics*, 1(2), 157–162. <https://doi.org/10.1023/A:1026535006318>

- Gordon, D., Huddleston, J., Chaisson, M. J. P., Hill, C. M., Kronenberg, Z. N., Munson, K. M., Malig, M., Raja, A., Fiddes, I., Hillier, L. W., Dunn, C., Baker, C., Armstrong, J., Diekhans, M., Paten, B., Shendure, J., Wilson, R. K., Haussler, D., Chin, C.-S., & Eichler, E. E. (2016). Long-read sequence assembly of the gorilla genome. *Science*, 352(6281), aae0344. <https://doi.org/10.1126/science.aae0344>
- Harestad, A. S., & Bunnell, F. L. (1987). Persistence of black-tailed deer fecal pellets in coastal habitats. *The Journal of Wildlife Management*, 51(1), 33. <https://doi.org/10.2307/3801624>
- Hernandez-Rodriguez, J., Arandjelovic, M., Lester, J., de Filippo, C., Weihmann, A., Meyer, M., Angedakin, S., Casals, F., Navarro, A., Vigilant, L., Kühl, H. S., Langergraber, K., Boesch, C., Hughes, D., & 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>
- Hicks, A. L., Lee, K. J., Couto-Rodriguez, M., Patel, J., Sinha, R., Guo, C., Olson, S. H., Seimon, A., Seimon, T. A., Ondzie, A. U., Karesh, W. B., Reed, P., Cameron, K. N., Lipkin, W. I., & Williams, B. L. (2018). Gut microbiomes of wild great apes fluctuate seasonally in response to diet. *Nature Communications*, 9(1), 1786. <https://doi.org/10.1038/s41467-018-04204-w>
- Inoue, E., Akomo-Okoue, E. F., Ando, C., Iwata, Y., Judai, M., Fujita, S., Hongo, S., Nze-Nkogue, C., Inoue-Murayama, M., & Yamagiwa, J. (2013). Male genetic structure and paternity in western lowland gorillas (*Gorilla gorilla gorilla*). *American Journal of Physical Anthropology*, 151(4), 583–588. <https://doi.org/10.1002/ajpa.22312>
- King, S. R. B., Schoenecker, K. A., Fike, J. A., & Oyler-McCance, S. J. (2018). Long-term persistence of horse fecal DNA in the environment makes equids particularly good candidates for noninvasive sampling. *Ecology and Evolution*, 8(8), 4053–4064. <https://doi.org/10.1002/ece3.3956>
- Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, 40(1), 1–8. <https://doi.org/10.1093/nar/gkr771>
- Kong, S. W., Lee, I. H., Liu, X., Hirschhorn, J. N., & Mandl, K. D. (2018). Measuring coverage and accuracy of whole-exome sequencing in clinical context. *Genetics in Medicine*, 20(12), 1617–1626. <https://doi.org/10.1038/gim.2018.51>
- Kühl, H. S., Boesch, C., Kulik, L., Haas, F., Arandjelovic, M., Dieguez, P., Bocksberger, G., McElreath, M. B., Agbor, A., Angedakin, S., Ayimisin, E. A., Bailey, E., Barubiyo, D., Bessone, M., Brazzola, G., Chancellor, R., Cohen, H., Coupland, C., Danquah, E., ... Kalan, A. K. (2019). Human impact erodes chimpanzee behavioral diversity. *Science (New York, N.Y.)*, 363(6434), 1453–1455. <https://doi.org/10.1126/science.aau4532>
- Lester, J. D., Vigilant, L., Gratton, P., McCarthy, M. S., Barratt, C. D., Dieguez, P., & Arandjelovic, M. (2020). Recent genetic connectivity and clinal variation in chimpanzees. In review
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Locke, D. P., Hillier, L. D. W., Warren, W. C., Worley, K. C., Nazareth, L. V., Muzny, D. M., Yang, S.-P., Wang, Z., Chinwalla, A. T., Minx, P., Mitreva, M., Cook, L., Delehaunty, K. D., Fronick, C., Schmidt, H., Fulton, L. A., Fulton, R. S., Nelson, J. O., Magrini, V., ... Wilson, R. K. (2011). Comparative and demographic analysis of orang-utan genomes. *Nature*, 469(7331), 529–533. <https://doi.org/10.1038/nature09687>
- Maricic, T., Whitten, M., & Pääbo, S. (2010). Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One*, 5(11), e14004. <https://doi.org/10.1371/journal.pone.0014004>
- Mengüllüoğlu, D., Fickel, J., Hofer, H., & Förster, D. W. (2019). Non-invasive faecal sampling reveals spatial organization and improves measures of genetic diversity for the conservation assessment of territorial species: Caucasian lynx as a case species. *PLoS One*, 14(5), <https://doi.org/10.1371/journal.pone.0216549>
- Mikkelsen, T. S., Hillier, L. W., Eichler, E. E., Zody, M. C., Jaffe, D. B., Yang, S. P., & Waterston, R. H. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055), 69–87. <https://doi.org/10.1038/nature04072>
- Morin, P. A., Chambers, K. E., Boesch, C., & Vigilant, L. (2001). Quantitative PCR analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees. *Molecular Ecology*, 10, 1835–1844.
- Morin, P. A., Wallis, J., Moore, J. J., Chakraborty, R., & Woodruff, D. S. (1993). Non-invasive sampling and DNA amplification for paternity exclusion, community structure, and phylogeography in wild chimpanzees. *Primates*, 34(3), 347–356. <https://doi.org/10.1007/BF02382630>
- 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>
- Orkin, J. D., Yang, Y., Yang, C., Yu, D. W., & Jiang, X. (2016). Cost-effective scat-detection dogs: Unleashing a powerful new tool for international mammalian conservation biology. *Scientific Reports*, 6(1), 34758. <https://doi.org/10.1038/srep34758>
- Ouborg, N. J., Pertoldi, C., Loeschcke, V., Bijlsma, R. K., & Hedrick, P. W. (2010). Conservation genetics in transition to conservation genomics. *Trends in Genetics*, 26(4), 177–187. <https://doi.org/10.1016/j.tig.2010.01.001>
- Perry, G. H. (2014). The promise and practicality of population genomics research with endangered species. *International Journal of Primatology*, 35(1), 55–70. <https://doi.org/10.1007/s10764-013-9702-z>
- 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>
- Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A. E., Malig, M., Hernandez-Rodriguez, J., ... Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, 499(7459), 471–475. <https://doi.org/10.1038/nature12228>
- Primmer, C. R. (2009). From conservation genetics to conservation genomics. *Annals of the New York Academy of Sciences*, 1162, 357–368. <https://doi.org/10.1111/j.1749-6632.2009.04444.x>
- 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>
- R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>
- Reiners, T. E., Encarnação, J. A., & Wolters, V. (2011). An optimized hair trap for non-invasive genetic studies of small cryptic mammals. *European Journal of Wildlife Research*, 57(4), 991–995. <https://doi.org/10.1007/s10344-011-0543-9>
- Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, 22(5), 939–946. <https://doi.org/10.1101/gr.128124.111>

- Sally, A., Dutheil, J. Y., Hillier, L. D. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S. H., Schwalie, P. C., Tang, Y. A., Ward, M. C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L. N., ... Durbin, R. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483(7388), 169–175. <https://doi.org/10.1038/nature10842>
- Schwartz, M. K., Luikart, G., & Waples, R. S. (2007). Genetic monitoring as a promising tool for conservation and management. *Trends in Ecology and Evolution*, 22, 25–33. <https://doi.org/10.1016/j.tree.2006.08.009>
- Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., Colling, G., Dalén, L., De Meester, L., Ekblom, R., Fawcett, K. D., Fior, S., Hajibabaei, M., Hill, J. A., Hoezel, A. R., Höglund, J., Jensen, E. L., Krause, J., Kristensen, T. N., ... Zieliński, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*, 30(2), 78–87. <https://doi.org/10.1016/j.tree.2014.11.009>
- Snyder-Mackler, N., Majoros, W. H., Yuan, M. L., Shaver, A. O., Gordon, J. B., Kopp, G. H., Schlebusch, S. A., Wall, J. D., Alberts, S. C., Mukherjee, S., Zhou, X., & 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>
- 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>
- Steiner, C. C., Putnam, A. S., Hoeck, P. E. A., & Ryder, O. A. (2013). Conservation genomics of threatened animal species. *Annual Review of Animal Biosciences*, 1(1), 261–281. <https://doi.org/10.1146/annurev-animal-031412-103636>
- Stenglein, J. L., Waits, L. P., Ausband, D. E., Zager, P., & Mack, C. M. (2010). Efficient, noninvasive genetic sampling for monitoring reintroduced wolves. *Journal of Wildlife Management*, 74(5), 1050–1058. <https://doi.org/10.2193/2009-305>
- Taberlet, P., Luikart, G., & Waits, L. P. (1999). Noninvasive genetic sampling: Look before you leap. *Trends in Ecology and Evolution*, 14(8), 323–327. [https://doi.org/10.1016/S0169-5347\(99\)01637-7](https://doi.org/10.1016/S0169-5347(99)01637-7)
- Thalmann, O., Hebler, J., Poinar, H. N., Pääbo, S., & Vigilant, L. (2004). Unreliable mtDNA data due to nuclear insertions: A cautionary tale from analysis of humans and other great apes. *Molecular Ecology*, 13(2), 321–335. <https://doi.org/10.1046/j.1365-294X.2003.02070.x>
- van der Valk, T., Lona Durazo, F., 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>
- Vigilant, L., & Guschanski, K. (2009). Using genetics to understand the dynamics of wild primate populations. *Primates*, 50(2), 105–120. <https://doi.org/10.1007/s10329-008-0124-z>
- Wedrowicz, F., Karsa, M., Mosse, J., & Hogan, F. E. (2013). Reliable genotyping of the koala (*Phascolarctos cinereus*) using DNA isolated from a single faecal pellet. *Molecular Ecology Resources*, 13(4), 634–641. <https://doi.org/10.1111/1755-0998.12101>
- White, L. C., Fontseré, C., Lizano, E., Hughes, D. A., Angedakin, S., Arandjelovic, M., Granjon, A.-C., Hans, J. B., Lester, J. D., Rabanus-Wallace, M. T., Rowney, C., Städele, V., Marques-Bonet, T., Langergraber, K. E., & Vigilant, L. (2019). A roadmap for high-throughput sequencing studies of wild animal populations using noninvasive samples and hybridization capture. *Molecular Ecology Resources*, 19(3), 609–622. <https://doi.org/10.1111/1755-0998.12993>
- Xue, Y., Prado-Martinez, J., Sudmant, P. H., Narasimhan, V., Ayub, Q., Szpak, M., Frandsen, P., Chen, Y., Yngvadottir, B., Cooper, D. N., de Manuel, M., Hernandez-Rodriguez, J., Lobon, I., Siegmund, H. R., Pagani, L., Quail, M. A., Hvilson, C., Mudakikwa, A., Eichler, E. E., ... Sally, A. (2015). Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, 348(6231), 242–245. <https://doi.org/10.1126/science.aaa3952>

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Fontseré C, Alvarez-Estape M, Lester J, et al. Maximizing the acquisition of unique reads in noninvasive capture sequencing experiments. *Mol Ecol Resour*. 2021;21:745–761. <https://doi.org/10.1111/1755-0998.13300>