Maximizing the acquisition of unique reads in non-invasive capture sequencing experiments

Running title: Increasing coverage of captured fecal DNA

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

[1] Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), PRBB, Doctor Aiguader 88, Barcelona, 08003, Spain.

[2] Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

[3] WWF Cameroon Country Programme Office, BP6776; Yaoundé, Cameroon.

[4] School of Biological and Environmental Sciences, Liverpool John Moores University, James Parsons Building, Byrom street, Liverpool, L3 3AF, UK.

[5] School of Human Evolution and Social Change, Arizona State University, 900 Cady Mall, Tempe, AZ 85287 Arizona State University, PO Box 872402, Tempe, AZ 85287-2402 USA.

[6] Institute of Human Origins, Arizona State University, 900 Cady Mall, Tempe, AZ 85287 Arizona State University, PO Box 872402, Tempe, AZ 85287-2402 USA.

[7] Comoé Chimpanzee Conservation Project, Kakpin, Comoé National Park, Ivory Coast.

[8] Department of Anthropology, University College London, 14 Taviton St, Bloomsbury London.

[9] Wild Chimpanzee Foundation (WCF) 23BP238 Abidjan, Côte d'Ivoire 23.

[10] Taï Chimpanzee Project, Centre Suisse de Recherches Scientifiques, BP 1301, Abidjan 01, CI, Côte d'Ivoire.

[11] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig.

[12] CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain.

[13] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia 08010, Spain.

[14] Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Columnes s/n, 08193 Cerdanyola del Vallès, Spain.

[15] MRC Integrative Epidemiology Unit at University of Bristol, Bristol, BS8 2BN, UK.

[16] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, BS8 2BN, UK.

* Esther Lizano and David A. Hughes should be considered joint senior author.

[†]**Corresponding author:** Esther Lizano and Tomas Marques-Bonet

Abstract

Non-invasive 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 non-invasive samples, such as feces, but its sensitivity has not yet been extensively studied. Coping with fecal 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 unfavorable for target capture sequencing might be the only representatives of unique 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, here we have captured the exome of 60 chimpanzees (*Pan troglodytes*) using fecal 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. Lastly, we have specifically calculated the sequencing effort needed to avoid exhausting the library complexity of enriched fecal samples with low endogenous DNA content.

This study provides guidelines, schemes and tools for laboratories facing the challenges of working with non-invasive samples containing extremely low amounts of endogenous DNA.

Keywords: Non-invasive samples, fecal samples, target capture, molecular complexity, conservation genomics, chimpanzees.

Introduction

Studies of wild animal populations that are unamenable to invasive sampling (eg: trapping or darting) often rely on the usage of low quality and/or quantity DNA samples (Schwartz, Luikart, & Waples, 2007; Vigilant & Guschanski, 2009), traditionally restricting the analysis to neutral markers or genetic loci such as microsatellites (Arandjelovic et al., 2011; Inoue et al., 2013; Mengüllüoğlu, Fickel, Hofer, & Förster, 2019; Orkin, Yang, Yang, Yu, & Jiang, 2016), autosomal regions (Fischer, Wiebe, Pääbo, & Przeworski, 2004) and the mitochondrial genome (Fickel, Lieckfeldt, Ratanakorn, & Pitra, 2007; Thalmann, Hebler, Poinar, Pääbo, & Vigilant, 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 cataloging 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, Pertoldi, Loeschcke, Bijlsma, & Hedrick, 2010; Primmer, 2009; Shafer et al., 2015; Städele & Vigilant, 2016; Steiner, Putnam, Hoeck, & Ryder, 2013).

Conservation genomics of ecologically-crucial, non-model organisms, and especially threatened species such as great apes, have largely benefited 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, Wallis, Moore, Chakraborty, & Woodruff, 1993; Taberlet, Luikart, & Waits, 1999), but also to track cryptic or monitor reintroduced species (De Barba et al., 2010; Ferreira et al., 2018; Reiners, Encarnação, & Wolters, 2011; Stenglein, Waits, Ausband, Zager, & Mack, 2010), sampling often relies on non-invasive (NI) sources of DNA such as feces 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, Marioni, Melsted, & Gilad, 2010; Taberlet et al., 1999), and contain genetic material from the host's microbiota and from species living in the environment where the sample was collected (i.e., exogenous DNA) (Hicks et al., 2018). The proportion of endogenous versus 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, Chikhi, Utami, De Ruiter, & Bruford, 2000; Harestad & Bunnell, 1987; King, Schoenecker, Fike, & Oyler-McCance, 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, Whitten, & Pääbo, 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, Lona Durazo, Dalén, & Guschanski, 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., 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. 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. 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 fecal 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 non-invasive samples. Here, we will 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, 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 fecal 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 fecal 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.

Material and Methods

Samples and Library Preparation

Chimpanzee fecal samples from 15 different sites in Africa were collected as part of the PanAf (Figure 1A). Approximately 5g (“hazelnut-size”) of feces were collected from each chimpanzee fecal 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 fecal sample were extracted using a Qiagen robot with the QIAamp Fast DNA Stool Mini Kit (Qiagen) with modifications (Lester et al, in review, 2020). The extractions, including blanks, were screened using a microsatellite genotyping assay (Arandjelovic et al., 2009; Arandjelovic et al., 2011) and up to 20 samples from each PanAf field site were selected as follows: (1) those that amplified at the most loci of the 15 loci panel, (2) represented unique individuals, and (3) were ascertained to have a low probability of being first degree relatives (Csilléry et al., 2006) (302 samples) (Supporting Information 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 fecal sample. Total DNA concentration and fragmentation were measured on a Fragment Analyzer using a Genomic DNA 50Kb 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, Chambers, Boesch, & Vigilant, 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%) (Supporting Information 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 (Supporting Information Figure S2), barcoding 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 Supplementary Information.

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 (µg) to the final pool, creating an equi-endogenous pool. So, 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 µg 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 µg 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 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, Supporting Information 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 (full protocol is available in Supporting Information), and started with either 1 µg 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, Sawyer, & Meyer, 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 reminder of the article when we use the word “capture” or “hybridization”, we will always be referring to the dual-capture or two consecutive rounds of capture hybridizations that are described above.

Sequencing and Mapping

Captured libraries were pooled into 3 sequencing batches and sequenced on a total of 3.75 lanes of a HiSeq 4000 with 2x100 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).

Demultiplexed FASTQ files were trimmed with Trimmomatic (version 0.36) (Bolger, Lohse, & Usadel, 2014) to remove the first 7 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). From now on, we will refer to those reads remaining after filtering as “reliable reads”. To retrieve the reliable reads on-target we used intersectBed from 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 (Supporting Information 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.

Capture performance

Capture performance was evaluated by calculating the enrichment factor (EF), capture specificity (CSp), 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 (64Mb) to the genome size (~3Gb). CSp 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), ten (DP10), twenty (DP20) or fifty (DP50) - divided by the total number of 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 has on performance summary statistics (number of unique reads, reliable reads, EF, LC, CS and CSp). We down-sampled 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, percent 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).

Results

Sample Description

Samples were collected from 15 different PanAf sites distributed across the entire range of chimpanzees in Africa (Figure 1A and Supporting Information Table S1). The 302 screened samples had an average eDNA of 1.49%, ranging from 0 to 47.75% (Figure 1B, Supporting Information 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 bp (ranging from 72 to 17,966 bp) (Supporting Information Figure S1B and Table S1).

We observe variation on the average endogenous content among geographical sites (Figure 1B), and also variation on fragment length among geographical sites (Supporting Information Figure S1B). For instance, samples collected in a specific location such as Campo Ma'an (Cameroon) have 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) have samples with higher than average eDNA (6.95%) (Supporting Information Table S3). This might be explained by the influence of weather, humidity and temperature on DNA preservation and bacterial growth in the fecal sample before collection as well as a product of sample age and quality of sampling conditions (Brinkman, Schwartz, Person, Pilgrim, & Hundertmark, 2010; Goossens et al., 2000; Harestad & Bunnell, 1987; King et al., 2018; Nsubuga et al., 2004; Wedrowicz, Karsa, Mosse, & Hogan, 2013).

A total of 60 samples with a mean percent endogenous content of 0.58% (range from 0.41% to 0.85%), and with a median human contamination of 0.0875% (range 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 3 pools with 10, 20 and 30 samples each (Figure 2). Samples were divided into pools based on their percent endogenous content, such that those samples with higher levels of percent 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 percent endogenous DNA is highly structured among the three pools, explaining 81% of the variation in eDNA (univariate linear model using rank normal transformed % eDNA; p-value = 2.05x10^{-91}) (Supporting Information Figure S5A).

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 3 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%) (Supporting Information 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 (Supporting information 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 will elaborate on outcome and improvements below.

The ~1.40 billion reads were not equally distributed among the 3 pools (production reads explained by pools; r^2 = 0.41, p-value = 3.24x10^{-16}) or 18 hybridizations (r^2 = 0.62, p-value = 2.59x10^{-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). 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 (Supporting Information Figure S3A, S3B and Table S5). We therefore reduced subsequent sequencing efforts, as discussed in section “Optimization of required production reads”, for the remaining replicate hybridizations.

All capture performance summary statistics (Supporting Information Table S4), to the exception of capture specificity (CSp), are strongly correlated with the number of production reads acquired (median correlation coefficient = 0.422, CI = 0.03 to 0.93; Supporting information 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 down-sampled 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 (Supporting Information Figure S5B, Table S7 and S8). The effect each experimental variable has 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 are 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 varies among our hybridizations and sequencing batches. These tendencies can be observed in Figure 5A-C. We account for this 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 percent 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.

Relevance of Equi-Endogenous Pools

The observations of Hernandez-Rodriguez et al. and White et al. 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 can be observed in some hybridizations of P2 (Supporting Information 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 find 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 (Supporting Information Figure S8).

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 twelve hybridizations we used the manufacturer's suggested amount of starting material, 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 fecal 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 (Supporting Information Figure S3B). However, given that production reads also vary between these two conditions, we down-sampled 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 a 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 (Supporting Information Figure S9). Therefore, whenever there is sufficient library available, it is advisable to start with 2 μg rather than 1 μg.

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. 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 Supporting Information Figure S4 for a schematic representation).

We performed the analysis of molecular complexity in libraries belonging to P3 since more hybridization replicates were available (8 in total) for 30 libraries. We found that for the majority of the libraries, performing additional hybridizations increased the number of unique reads retrieved (Supporting Information Figure S10, example library N259-5). However, there were libraries that quickly hit exhaustion where performing additional hybridizations would add little extra information (Supporting 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 (Supporting Information Figure S11). Interestingly, no sample covered 100% of target space. Looking carefully into this, we observed that precisely the same 3,804 regions (1.54%) were never covered in any replicate hybridizations, suggesting that some regions are either difficult to capture (Kong, Lee, Liu, Hirschhorn,

& Mandl, 2018) or are too divergent between *Homo* and *Pan* to either capture or map these particular sequences (Supporting Information Figure S12).

For deeper coverage of at least 4 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 (Supporting Information Figure S11).

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 fecal 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 2 hybridizations for P1, the first 4 hybridizations for P2, and the first 6 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 (Supporting Information Table S5), indicative of high levels of PCR duplicates due to library exhaustion. To avoid over-sequencing in our next 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 2 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 (Supporting Information 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 (Supporting Information Figure S14). We note that these values exceeded what we observed in the previous hybridization experiments. An outcome we attribute to the increase in starting material (2 µg), also used in these experiments, as noted above.

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, Supporting Information 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 Supporting Information Figure S16). If we remove the effect of having a variable number of production reads across experiments by down-sampling, this observation still remains (Supporting Information 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=3.47x10^{-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 = 2.7x10^{-4}; Supporting Information 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 (Supporting Information 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 Supporting Information Figure S4 for a schematic representation), we can acquire the same number of unique reliable reads (Figure 6C, Supporting Information S17).

Discussion

Capturing host DNA from fecal samples is a challenging endeavor. Previous work has shown that the retrieval of genomic data from fecal 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 fecal samples is not always possible. Because of that, many studies have characterized the technical difficulties of capturing DNA from non-invasive 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 analyzed 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 fecal 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 dataset. 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 fecal 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-barcoded 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 fecal samples, we recommend screening your set of samples based on quantity and quality of the DNA extracted. If having related or identical individuals in the study should be avoided, microsatellite genotyping could be an option, helping as well to discard samples with 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 over-sequencing can result in rapidly depleting the economic feasibility of the experiment.

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

Acknowledgments

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 Forets, 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 le 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.

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. doi: 10.1111/j.1755-0998.2008.02387.x

Arandjelovic, M., Head, J., Rabanal, L. I., Schubert, G., Mettke, E., Boesch, C., … Vigilant, L. (2011). Non-invasive genetic monitoring of wild central chimpanzees. *PLoS ONE*, *6*(3), e14761. doi: 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*, Vol. 526, pp. 68–74. doi: 10.1038/nature15393

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi: 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. doi: 10.1007/s10592-009-9928-7

Burbano, H. A., Hodges, E., Green, R. E., Briggs, A. W., Krause, J., Meyer, M., … Pääbo, S. (2010). Targeted investigation of the neandertal genome by array-based sequence capture. *Science*, *328*(5979), 723–725. doi:

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. doi: 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. *American Journal of Human Genetics*, *93*(5), 852–864. doi: 10.1016/j.ajhg.2013.10.002

Csilléry, K., Johnson, T., Beraldi, D., Clutton-Brock, T., Coltman, D., Hansson, B., … Pemberton, J. M. (2006). Performance of marker-based relatedness estimators in natural populations of outbred vertebrates. *Genetics*, *173*(4), 2091–2101. doi: 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. doi: 10.1111/j.1365-2664.2009.01752.x

De Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., … Marques-Bonet, T. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, *354*(6311), 477–481. doi: 10.1126/science.aag2602

Ferreira, C. M., Sabino-Marques, H., Barbosa, S., Costa, P., Encarnação, C., Alpizar-Jara, R., … 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). doi: 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. doi: 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. doi: 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. doi: 10.1023/A:1026535006318

Gordon, D., Huddleston, J., Chaisson, M. J. P., Hill, C. M., Kronenberg, Z. N., Munson, K. M., … Eichler, E. E. (2016). Long-read sequence assembly of the gorilla genome. *Science*, *352*(6281), aae0344. doi: 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. doi: 10.2307/3801624

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. doi: 10.1111/1755-0998.12728

Hicks, A. L., Lee, K. J., Couto-Rodriguez, M., Patel, J., Sinha, R., Guo, C., … Williams, B. L. (2018). Gut microbiomes of wild great apes fluctuate seasonally in response to diet. *Nature Communications*, *9*(1), 1786. doi: 10.1038/s41467-018-04204-w

Inoue, E., Akomo-Okoue, E. F., Ando, C., Iwata, Y., Judai, M., Fujita, S., … Yamagiwa, J. (2013). Male genetic structure and paternity in western lowland gorillas (Gorilla gorilla gorilla). *American Journal of Physical Anthropology*, *151*(4), 583–588. doi: 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. doi: 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. doi: 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. doi: 10.1038/gim.2018.51

Kühl, H. S., Boesch, C., Kulik, L., Haas, F., Arandjelovic, M., Dieguez, P., … Kalan, A. K. (2019). Human impact erodes chimpanzee behavioral diversity. *Science (New York, N.Y.)*, *363*(6434), 1453–1455. doi: 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. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Locke, D. P., Hillier, L. W., Warren, W. C., Worley, K. C., Nazareth, L. V., Muzny, D. M., … Wilson, R. K. (2011). Comparative and demographic analysis of orang-utan genomes. *Nature*, *469*(7331), 529–533. doi: 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. doi: 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). doi: 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. doi: 10.1038/nature04072

Morin, P. A., Chambers, K. E., Boesch, C., & Vigilant, L. (2001). Quantitative PCR analysis of DNA from noninvasive samples fro accurate microsatellite genotyping of wild chimpanzees. *Molecular Ecology*, 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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., … Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, *499*(7459), 471–475. doi: 10.1038/nature12228

Primmer, C. R. (2009). From conservation genetics to conservation genomics. *Annals of the New York Academy of Sciences*, Vol. 1162, pp. 357–368. doi: 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. doi: 10.1093/bioinformatics/btq033

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. doi: 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. doi: 10.1101/gr.128124.111

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., … Durbin, R. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, *483*(7388), 169–175. doi: 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*, Vol. 22, pp. 25–33. doi: 10.1016/j.tree.2006.08.009

Shafer, A. B., Wolf, J. B., Alves, P. C., Bergströ, L., Bruford, M. W., Brä nnströ, I., … Zielin, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*, *30*(2), 78–87. doi: 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., … Tung, J. (2016). Efficient genome-wide sequencing and low-coverage pedigree analysis from noninvasively collected samples. *Genetics*, *203*(2), 699–714. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 10.1111/1755-0998.12101

White, L. C., Fontsere, C., Lizano, E., Hughes, D. A., Angedakin, S., Arandjelovic, M., … 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. doi: 10.1111/1755-0998.12993

Xue, Y., Prado-Martinez, J., Sudmant, P. H., Narasimhan, V., Ayub, Q., Szpak, M., … Scally, A. (2015). Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, *348*(6231), 242–245. doi: 10.1126/science.aaa3952

Data Accessibility

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).

Author Contributions

CF, TMB, DAH and EL designed the study. MA and HSK direct the Pan African Programme: The Cultured Chimpanzee. MA and HSK obtained funding for the project. MA, PD, AA, SA, EAA, MB, GB, TD, MEN, ACG, JH, PK, AKK, MK, KL, JL, GM, LJO, AP, MMR, FS, VV and RMW supervised, conducted field work and collected samples. CF, MAE, EL, JL, MA performed experiments. CF and DAH performed the analysis. MAE, MK, DAH, TMB, EL provided analytical support. CF wrote the manuscript with input from all authors.

Supporting Information

Additional supporting information with extended methods and supplementary figures and tables can be found online in the Supporting information section at the end of the article.

Conflict of Interest

Authors declare no conflict of interest.

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 1st and 3rd quartile).

FIGURE 2. Pooling strategy illustration. P1 has 10 libraries with average endogenous of 0.81%. We performed two primary pools of 2 μg 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 μg and 4 μg were distributed into six hybridization pools of 1μg and two hybridization pools of 2 μg each. Colors represent the sequencing batch.

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 colors 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 1st and 3rd quartile).

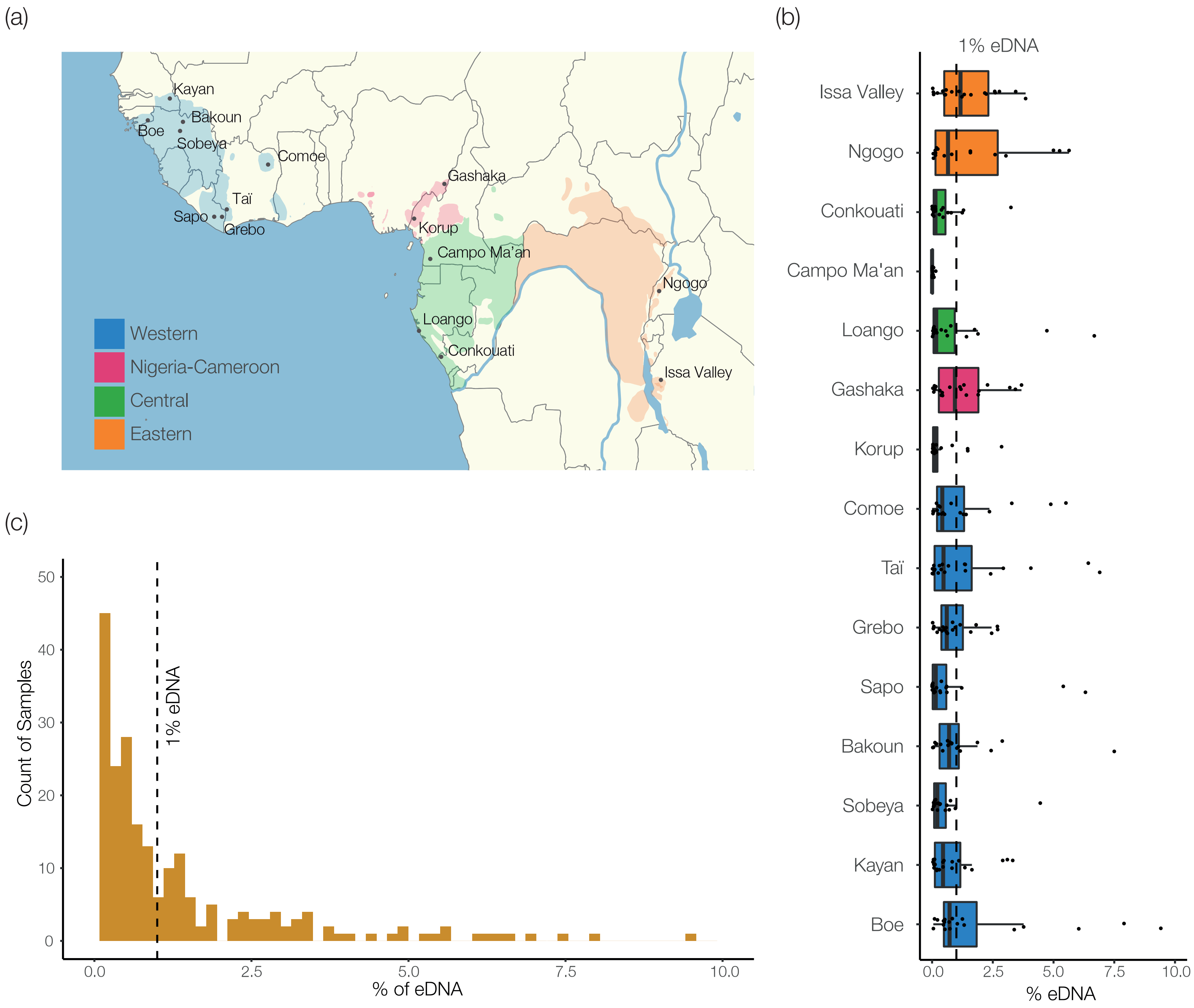
FIGURE 4. Analysis of variance. (a) Estimated variance explained from univariate linear models after rank normal transforming each performance summary statistic (columns). LC stands for library complexity and DP describes read depth at different cutoffs (1, 4, 10, 20 and 50 reads) (b) Multivariate type I ANOVA of the experimental variables affecting Capture Sensitivity (CS) at depth 1. Both models are built down-sampling libraries to 1,500,000 reads.

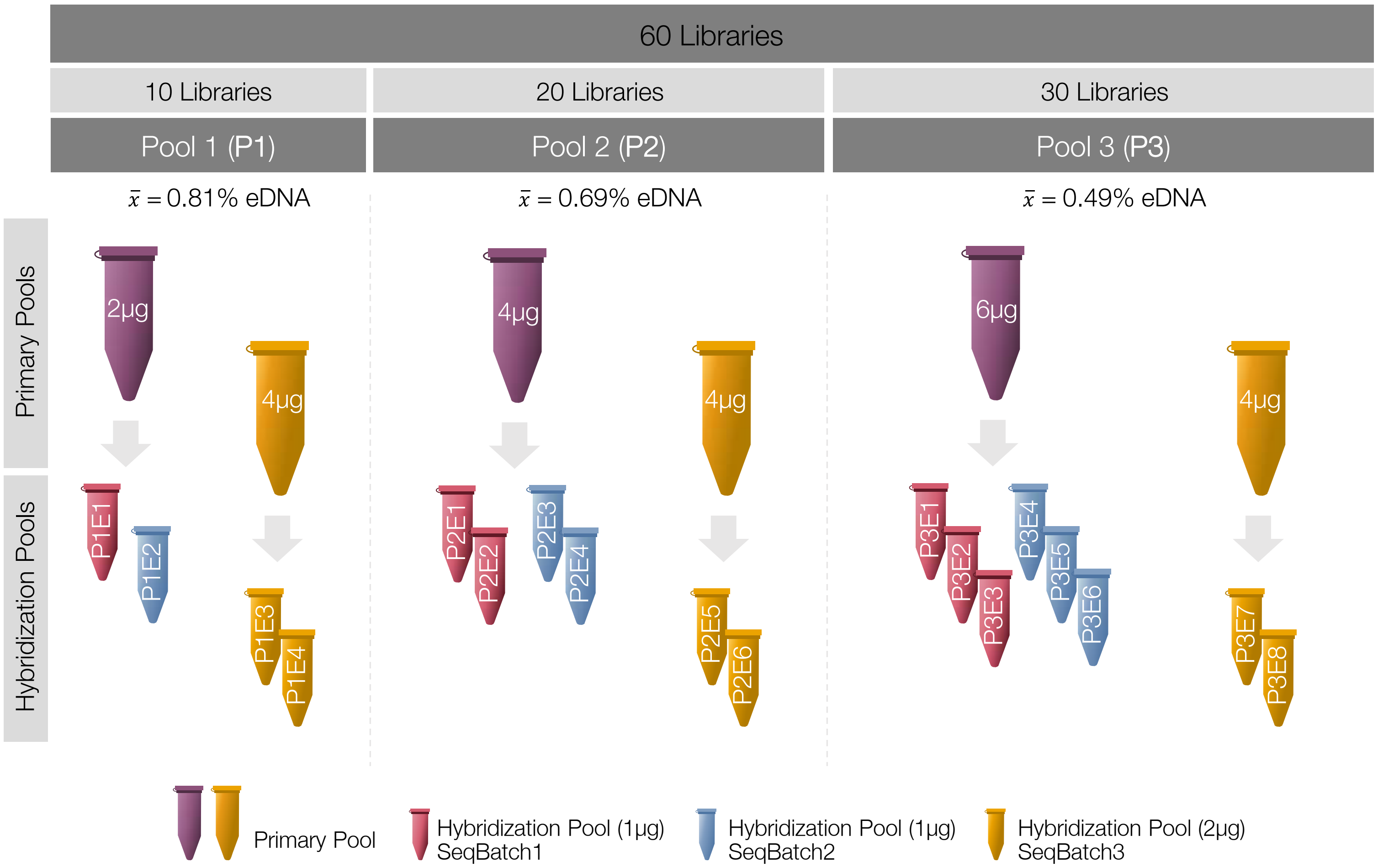
FIGURE 5. Summary stats after down-sampling to 1,500,000 reads: (a) Enrichment factor and (d) Capture Specificity (c) Capture Sensitivity at depth 1 for the 18 hybridizations in P1, P2 and P3; colors illustrate sequencing batch. (d) Library complexity contrasting the amount of starting DNA (1 μg or 2 μg) in down-sampled 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 1st and 3rd quartile).

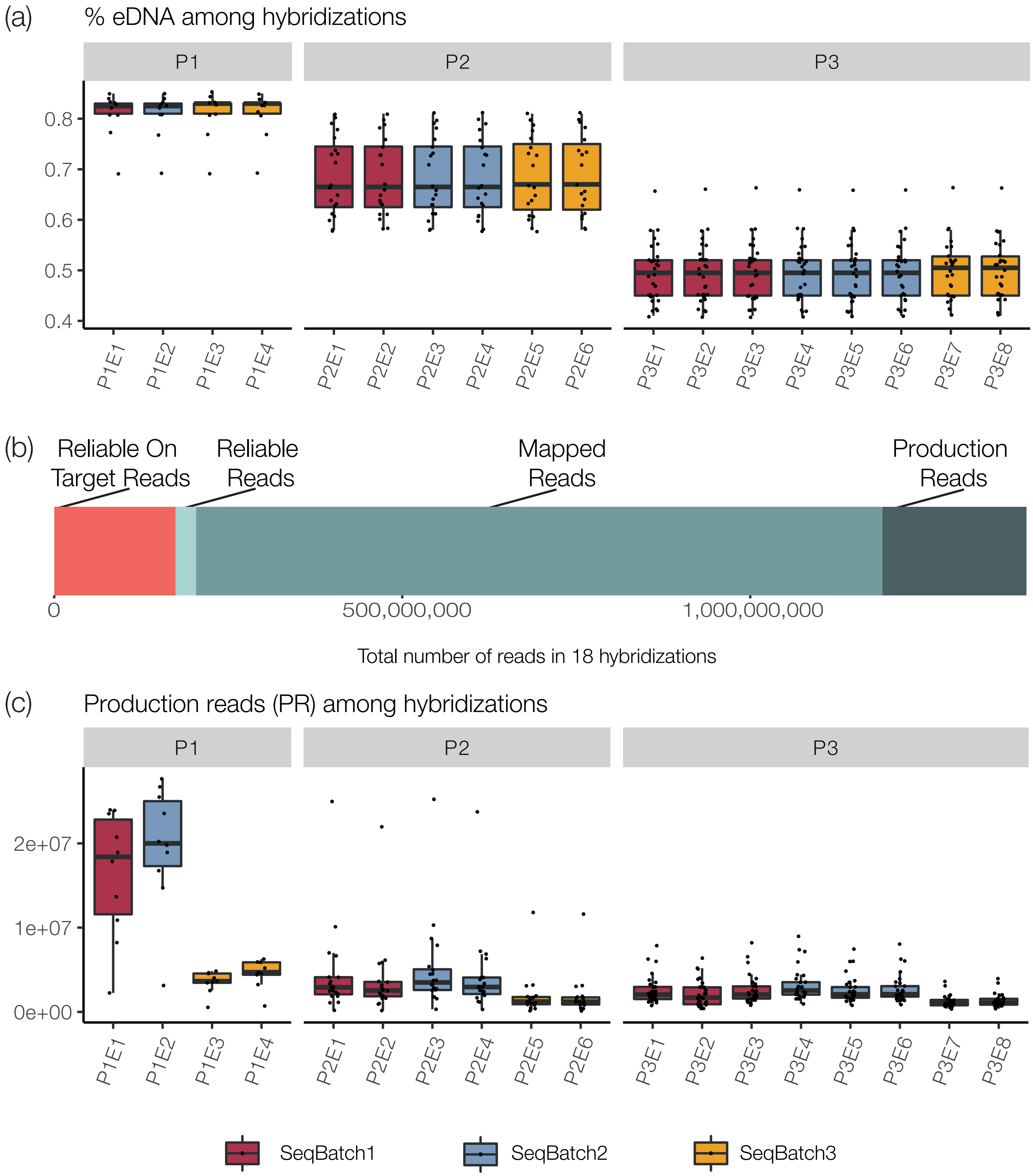
FIGURE 6. Analysis of coverage and LC with hybridizations done with 1 μg. (a) Coverage after merging data from additional hybridizations with up to 2, 4 and 6 for P1, P2 and P3. (b) Comparison of average LC curves of individual hybridizations belonging to pools with different size. Each line is the average 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.

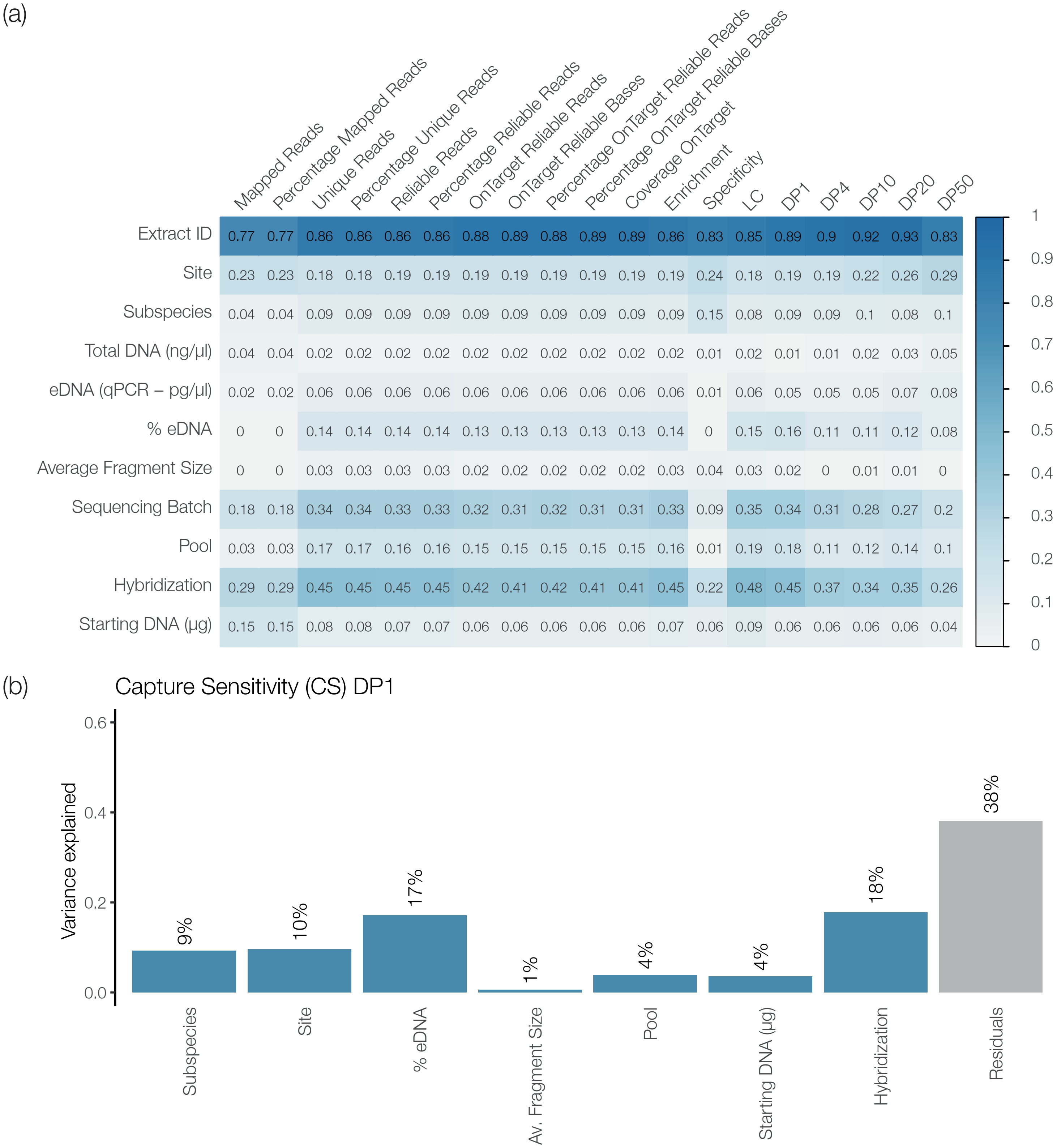
Pool	Average eDNA content (range)	Hybridization ID	Number of pooled libraries	Total DNA	Sequencing Batch
Pool 1 (**P1**)	0.81% (0.60% - 0.85%)	P1E1	10	1 µg	SeqBatch1
		P1E2	10	1 µg	SeqBatch2
		P1E3	9	2 µg	SeqBatch3
		P1E4	9	2 µg	SeqBatch3
Pool 2 (**P2**)	0.69% (0.58% - 0.80%)	P2E1	20	1 µg	SeqBatch1
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

TABLE 1. Pooling Strategy. Sixty libraries were divided into 3 pools for capture hybridization experiments in 4 replicates for P1, 6 replicates for P2 and 8 replicates for P3. Total DNA represents the starting material for each capture hybridization.

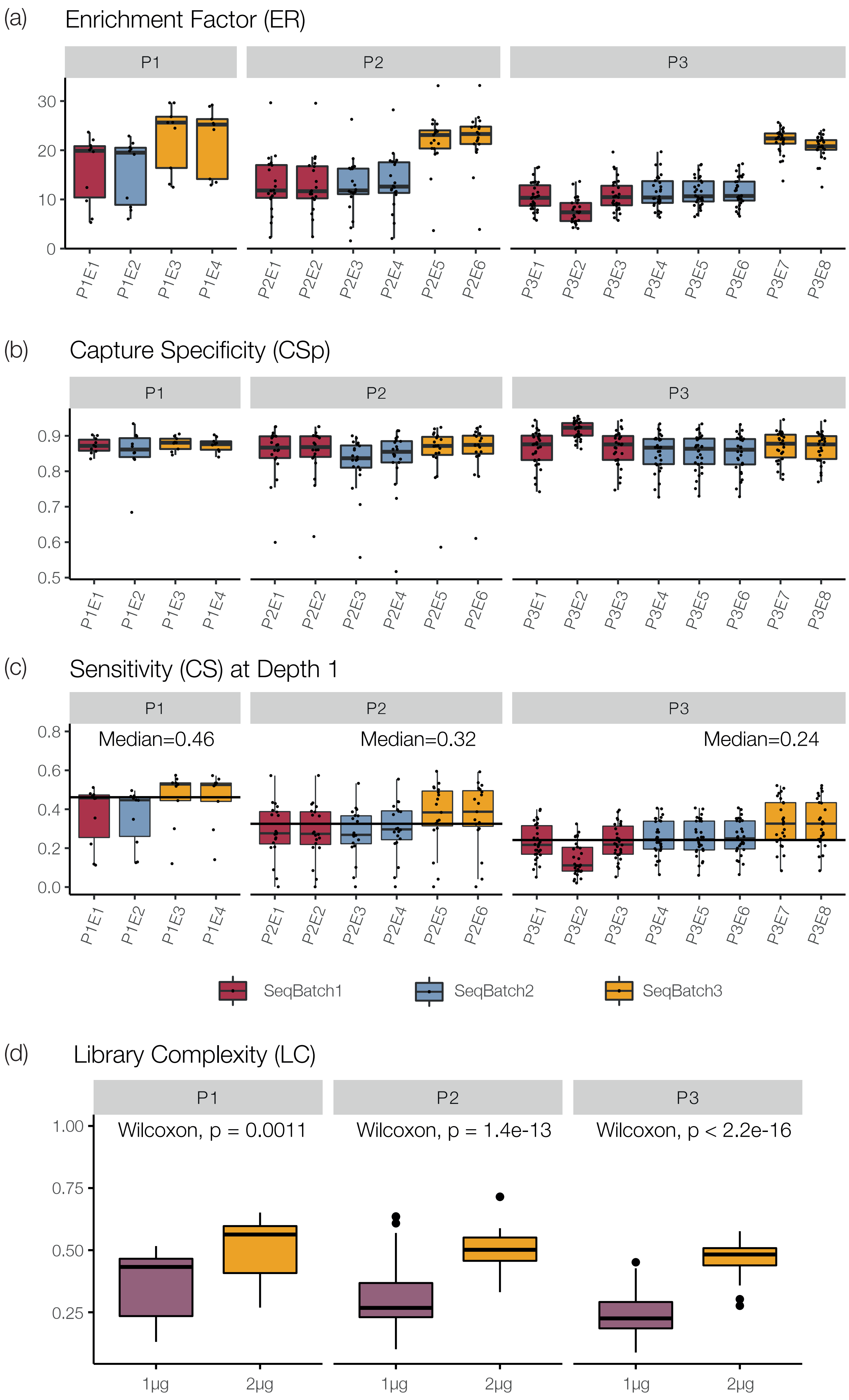

(a)
Kayan
Bakoun
Boe
Sobeya
Comoe
Gashaka
Taï
Sapo
Grebo
Korup
Campo Ma'an
Ngogo
Loango
Conkouati
Issa Valley
Western
Nigeria-Cameroon
Central
Eastern
(b)
1% eDNA
Issa Valley
Ngogo
Conkouati
Campo Ma'an
Loango
Gashaka
Korup
Comoe
Taï
Grebo
Sapo
Bakoun
Sobeya
Kayan
Boe
0.0
2.5
5.0
7.5
10.0
% eDNA
(c)
Count of Samples
0
10
20
30
40
50
1% eDNA
0.0
2.5
5.0
7.5
10.0
% of eDNA

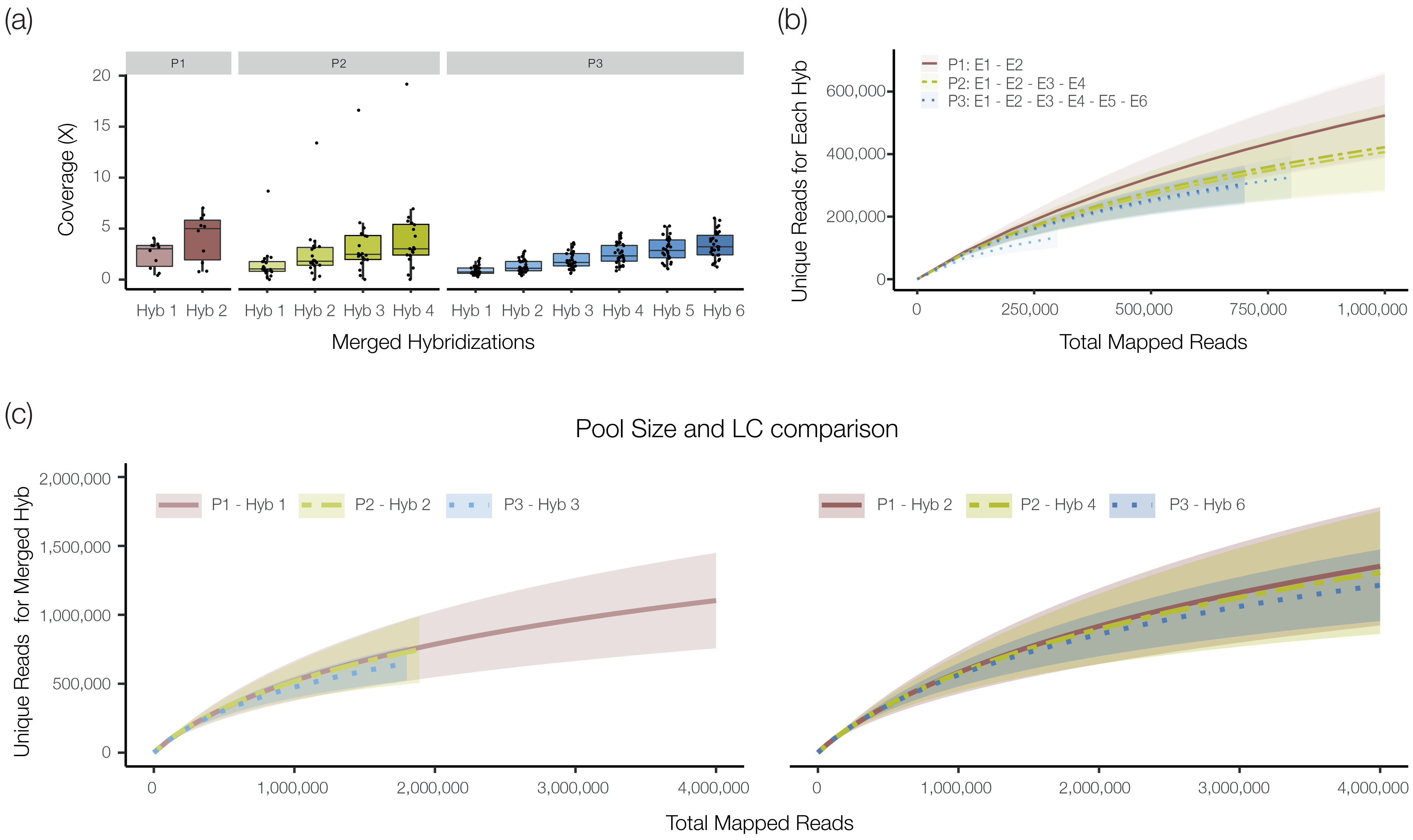

60 Libraries
10 Libraries
20 Libraries
30 Libraries
Pool 1 (P1)
Pool 2 (P2)
Pool 3 (P3)
$\bar{x}$ = 0.81% eDNA
$\bar{x}$ = 0.69% eDNA
$\bar{x}$ = 0.49% eDNA
Primary Pools
Hybridization Pools
2µg
4µg
4µg
4µg
6µg
4µg
P1E1
P1E2
P1E3
P1E4
P2E1
P2E2
P2E3
P2E4
P2E5
P2E6
P3E1
P3E2
P3E3
P3E4
P3E5
P3E6
P3E7
P3E8
Primary Pool
Hybridization Pool (1µg)
SeqBatch1
Hybridization Pool (1µg)
SeqBatch2
Hybridization Pool (2µg)
SeqBatch3


(a)
% eDNA among hybridizations
P1
P2
P3
0.8
0.7
0.6
0.5
0.4
P1E1
P1E2
P1E3
P1E4
P2E1
P2E2
P2E3
P2E4
P2E5
P2E6
P3E1
P3E2
P3E3
P3E4
P3E5
P3E6
P3E7
P3E8
(b)
Reliable On Target Reads
Reliable Reads
Mapped Reads
Production Reads
0
500,000,000
1,000,000,000
Total number of reads in 18 hybridizations
(c)
Production reads (PR) among hybridizations
P1
P2
P3
2e+07
1e+07
0e+00
P1E1
P1E2
P1E3
P1E4
P2E1
P2E2
P2E3
P2E4
P2E5
P2E6
P3E1
P3E2
P3E3
P3E4
P3E5
P3E6
P3E7
P3E8
SeqBatch1
SeqBatch2
SeqBatch3



	Mapped Reads	Percentage Mapped Reads	Unique Reads	Percentage Unique Reads	Reliable Reads	Percentage Reliable Reads	OnTarget Reliable Reads	OnTarget Reliable Bases	Percentage OnTarget Reliable Reads	Percentage OnTarget Reliable Bases	Coverage OnTarget	Enrichment	Specificity	LC	DP1	DP4	DP10	DP20	DP50
Extract ID	0.77	0.77	0.86	0.86	0.86	0.86	0.88	0.89	0.88	0.89	0.89	0.86	0.83	0.85	0.89	0.9	0.92	0.93	0.83
Site	0.23	0.23	0.18	0.18	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.24	0.18	0.19	0.19	0.22	0.26	0.29
Subspecies	0.04	0.04	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.15	0.08	0.09	0.09	0.1	0.08	0.1
Total DNA (ng/µl)	0.04	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.01	0.01	0.02	0.03	0.05
eDNA (qPCR – pg/µl)	0.02	0.02	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.01	0.06	0.05	0.05	0.05	0.07	0.08
% eDNA	0	0	0.14	0.14	0.14	0.14	0.13	0.13	0.13	0.13	0.13	0.14	0	0.15	0.16	0.11	0.11	0.12	0.08
Average Fragment Size	0	0	0.03	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.03	0.04	0.03	0.02	0	0.01	0.01	0
Sequencing Batch	0.18	0.18	0.34	0.34	0.33	0.33	0.32	0.31	0.32	0.31	0.31	0.33	0.09	0.35	0.34	0.31	0.28	0.27	0.2
Pool	0.03	0.03	0.17	0.17	0.16	0.16	0.15	0.15	0.15	0.15	0.15	0.16	0.01	0.19	0.18	0.11	0.12	0.14	0.1
Hybridization	0.29	0.29	0.45	0.45	0.45	0.45	0.42	0.41	0.42	0.41	0.41	0.45	0.22	0.48	0.45	0.37	0.34	0.35	0.26
Starting DNA (µg)	0.15	0.15	0.08	0.08	0.07	0.07	0.06	0.06	0.06	0.06	0.06	0.07	0.06	0.09	0.06	0.06	0.06	0.06	0.04


(a)
Enrichment Factor (ER)
P1
P2
P3
30
20
10
0
P1E1
P1E2
P1E3
P1E4
P2E1
P2E2
P2E3
P2E4
P2E5
P2E6
P3E1
P3E2
P3E3
P3E4
P3E5
P3E6
P3E7
P3E8
(b)
Capture Specificity (CSp)
P1
P2
P3
0.9
0.8
0.7
0.6
0.5
P1E1
P1E2
P1E3
P1E4
P2E1
P2E2
P2E3
P2E4
P2E5
P2E6
P3E1
P3E2
P3E3
P3E4
P3E5
P3E6
P3E7
P3E8
(c)
Sensitivity (CS) at Depth 1
P1
P2
P3
Median=0.46
Median=0.32
Median=0.24
0.8
0.6
0.4
0.2
0.0
P1E1
P1E2
P1E3
P1E4
P2E1
P2E2
P2E3
P2E4
P2E5
P2E6
P3E1
P3E2
P3E3
P3E4
P3E5
P3E6
P3E7
P3E8
SeqBatch1
SeqBatch2
SeqBatch3
(d)
Library Complexity (LC)
P1
P2
P3
Wilcoxon, p = 0.0011
Wilcoxon, p = 1.4e-13
Wilcoxon, p < 2.2e-16
1.00
0.75
0.50
0.25
1µg
2µg
1µg
2µg
1µg
2µg


(a)
P1
P2
P3
Coverage (X)
20
15
10
5
0
Hyb 1
Hyb 2
Hyb 1
Hyb 2
Hyb 3
Hyb 4
Hyb 1
Hyb 2
Hyb 3
Hyb 4
Hyb 5
Hyb 6
Merged Hybridizations
(b)
Unique Reads for Each Hyb
600,000
400,000
200,000
0
P1: E1 - E2
P2: E1 - E2 - E3 - E4
P3: E1 - E2 - E3 - E4 - E5 - E6
0
250,000
500,000
750,000
1,000,000
Total Mapped Reads
(c)
Pool Size and LC comparison
Unique Reads for Merged Hyb
2,000,000
1,500,000
1,000,000
500,000
0
P1 - Hyb 1
P2 - Hyb 2
P3 - Hyb 3
P1 - Hyb 2
P2 - Hyb 4
P3 - Hyb 6
0
1,000,000
2,000,000
3,000,000
4,000,000
0
1,000,000
2,000,000
3,000,000
4,000,000
Total Mapped Reads

MOLECULAR ECOLOGY
RESOURCES

Supplementary Information for:

Maximizing the acquisition of unique reads in non-invasive capture sequencing experiments

Claudia Fontsere, Marina Alvarez-Estape, Jack Lester, Mimi Arandjelovic, Martin Kuhlwilm, Paula Dieguez, Anthony Agbor, Samuel Angedakin, Emmanuel Ayuk Ayimisin, Mattia Bessone, Gregory Brazzola, Tobias Deschner, Manasseh Eno-Nku, Anne-Céline Granjon, Josephine Head, Parag Kadam, Ammie K. Kalan, Mohamed Kambi, Kevin Langergraber, Juan Lapuente, Giovanna Maretti, Lucy Jayne Ormsby, Alex Piel, Martha M. Robbins, Fiona Stewart, Virginie Vergnes, Roman M. Wittig, Richard McElreath, Hjalmar S. Kühl, Tomas Marques-Bonet, David A. Hughes[†] and Esther Lizano[†]

Table of Contents

Extended methods

Library Preparation

A single library was prepared for each sample following the BEST protocol published by Caroe *et al*. with minor modifications. A total of 200 ng of DNA in 35 μl of lowTE was sheared using a Covaris S2 ultrasonicator with the following settings to obtain 200 bp fragments: duty cycle: 10%, intensity: 5, cycles per burst: 200, time: 120 s.

Next, DNA was end-repaired using 0.5 μl T4 polymerase (5U/μl, Thermo Scientific) 1.5 μl T4 PNK (10 U/μl, Thermo Scientific), 0.4 μl dNTPs (25mM, GE Healthcare), 10 μl T4 DNA ligase buffer (5x, Invitrogen) and 2.5 μl Reaction Enhancer (20% PEG-4000 (Thermo Scientific), 2 mg/μL BSA (New England BioLabs), 400 mM NaCl (Sigma-Aldrich). The mix was incubated 30 min at 20ºC and 30 min at 65ºC (lid at 80ºC).

For adapter ligation reaction we used 2.5 μl T4 DNA ligase buffer (5x, Invitrogen), 1.25 μl T4 DNA ligase (5 U/μl, Invitrogen) and 6.25 μl ddH_2O. At each well we added unique inline barcoded short adapters (1.25 μl each at 100uM; F_P5_7nt_XX Indexed Adapter 5'-CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNN-3'; F_P7_7nt_XX Indexed Adapter 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNN-3'; R_P5/P7_7nt_XX Indexed Common Adapter 5'-NNNNNNNAGATCGGAA-3') with the same 7 nucleotide barcode for the P5 and P7 adapters (Figure S2). Previous studies have shown a better capture efficiency when the library size is small (Rohland & Reich, 2012). Moreover, an early barcoding of the library (in the adapter ligation step rather than in the final amplification PCR) lowers the probability of indiscernible contamination from close wells. Ligation reaction was incubated 45 min at 20ºC and 10 min at 64ºC (lid at 80ºC). Fill-in reaction was done using 2 μl of Bst 2.0 WarmStart Polymerase (8 U/μl, New England BioLabs), 2.5 μl of Isothermal amp. buffer (10x, New England BioLabs)), 0.5 μl of dNTPs (25 mM, GE Healthcare) and 7.5 μl ddH_2O. Reaction was incubated for 20 min at 65ºC (lid 80ºC) and 20 min at 80ºC (lid 110ºC).

The product was purified using homemade SPRI beads (Rohland & Reich, 2012) and eluting in a final volume of 25 μl of lowTE. Finally, each library was amplified using 25 μl of Kapa HIFI HS RM (2x, Roche), and 2.5 μl of each PreHyb primers (P5: 5'-CTTTCCCTACACGACGCTCTTC-3' and P7: 5'-GTGACTGGAGTTCAGACGTGTG-3', 10 μM) and incubated 2 min at 95ºC (lid at 110ºC), followed by 8 to 12 cycles of 15 s at 98ºC, 30 s at 55ºC and 30 s at 72ºC, with a final elongation of 1 min at 72ºC.

The final library was purified using homemade SPRI beads (Rohland & Reich, 2012) and eluting in a final volume of 30 μl of ddH_2O. Libraries were quantified with an Agilent 2100 Bioanalyzer using a DNA 7500 assay kit.

Hybridization Capture

Each hybridization reaction was performed with 1 or 2 μg of pooled library (7 μl) a blocking mix containing 2.5 μg of Human cot-1 (1 μg/μl, Invitrogen), 2.5 μg of salmon sperm (10 μg/μl, Invitrogen), 2 μM of P5 and P7 blocking oligos (Rohland & Reich, 2012), heated 5 min at 95ºC (lid 105ºC) and held at 65ºC for at least 5 minutes.

Then, the prewarmed 22 µl of hybridization buffer (10x SSPE (20x, Invitrogen), 10x Denhardt's Solution (50x, Invitrogen), 10mM EDTA (0.5M, Sigma-Aldrich), 0.2% SDS (20%, Invitrogen)) was added to the previously warmed to 65 ºC for 2 min bait mix: 3 µl of SureSelect Human All Exon V6 RNA library baits (Agilent Technologies), 1 µl of SUPERase-In and 1 µl of ddH$_2$O. The capture mix was added to the pools and incubated overnight at 65ºC. After the incubation we performed several washes with homemade wash buffers (Wash Buffer #1: 1x SSC (20x, Invitrogen) and 0.1% SDS (20%, Invitrogen); Wash Buffer #2: 0.1% SSC (20x, Invitrogen) and 0.1% SDS (20x, Invitrogen)) and Streptavidin-coated beads (Dynabeads MyOne Streptavidin T1 beads, Invitrogen). Beads were washed following the manufacturer's protocol and resuspended in 200 µl of binding buffer (1M NaCl (5M, Sigma-Aldrich), 10mM Tris-HCl pH 7.5 (1M, Invitrogen), 1mM EDTA (0.5M, Sigma-Aldrich)). The captured library was transferred to the beads and incubated at room temperature on a thermomixer at 700 RPM for 30 min. Using a magnetic rack, we removed the supernatant and washed the beads with Wash Buffer #1 for 15 min at room temperature on the thermomixer at 700 RPM. Then, the beads were placed in the magnetic rack again and washed with Wash Buffer #3 three times for 10 min at 68°C and 700 RPM. Finally, the beads were resuspended in 20 µl of H$_2$O followed by an enrichment PCR with PreHyb primers (P5-F: 5'-CTTTCCCTACACGACGCTCTTC-3' and P7-R: 5'-GTGACTGGAGTTCAGACGTGTG-3'), with the same incubation protocol as in library preparation amplification but with 10-12 cycles. After cleaning the PCR product with homemade SPRI beads (Rohland & Reich, 2012) a second capture experiment was performed as recommended by Hernandez-Rodriguez et al. PCR amplification (9-12 cycles) of the final captured pool was done using the same protocol as before but with indexed primers (P5-F: 5'-AATGATACGGCGACCACCGAGATCTACACNNNNNNNACACTCTTTCCCTACACGACGCTCTT-3' and P7-R: 5'-CAAGCAGAAGACGGCATACGAGATNNNNNNNGTGACTGGAGTTCAGACGTGT-3') (Kircher, Sawyer, & Meyer, 2012) to double-index each pool of libraries with a unique pair of indices (Figure S2).

As previously described, the use of inline barcodes and P5 and P7 indexing primers allows the multiplexing of numerous libraries in a single pool. Thus, for the experiments presented here, the usage of such adapters was of high utility, since after the libraries were build, we pooled them together for capture, and subsequently pools were indexed using P5 and P7 (Rohland & Reich, 2012).

Since the captured pools were indexed, it was possible to sequence many libraries in one sequencing lane. Also, these short adapters do not interfere with hybridization experiments as complete adapters did. As suggested in Rohland et al., we increased by one nucleotide the barcode sequence in the adapters, from 6nt to 7nt, thus increasing the multiplexing power.

Supplementary Table Legends

Supplementary T1. Sample description of screened samples.

Sample description for all screened samples in this study; provided in the additional excel file.

Supplementary T2. Sample description for capture samples.

Sample description for the selected samples for capture; provided in the additional excel file.

Supplementary T3. Endogenous content by site.

Average endogenous content of samples according to site; provided in the additional excel file.

Supplementary T4. Sequencing summary statistics.

Summary of sequencing stats for each sample in each hybridization; provided in the additional excel file.

Supplementary T5. Sequencing summary statistics for independent hybridizations.

Summary of sequencing stats for independent hybridizations, each row contains the sum of all samples belonging to each hybridization; provided in the additional excel file.

Supplementary T6. Correlation matrix among all study variables.

Correlation matrix of all variables analyzed in this study. Spearman's rho was estimated when comparing two numeric variables. Cramér's V was estimated among two categorical variables. When comparing a numeric and categorical variable we took the square root of the R-squared statistic derived from a univariate linear model with no transformation on the dependent, numerical values; provided in the additional excel file.

Supplementary T7. Sequencing summary statistics for down-sampled data.

Summary of sequencing stats for each down-sampled library at 1,500,000 in each hybridization; provided in the additional excel file.

Supplementary T8. Correlation matrix among all study variables for down-sampled data.

Correlation matrix of all variables analyzed in this study after each library has been down-sampled to 1,500,000 reads. Spearman's rho was estimated when comparing two numeric variables. Cramér's V was estimated among two categorical variables. When comparing a numeric and categorical variable we took the square root of the R-squared statistic derived from a univariate linear model with no transformation on the dependent, numerical values; provided in the additional excel file.

Supplementary Figures

Figure S1. Endogenous content and fragment size across sampling sites.

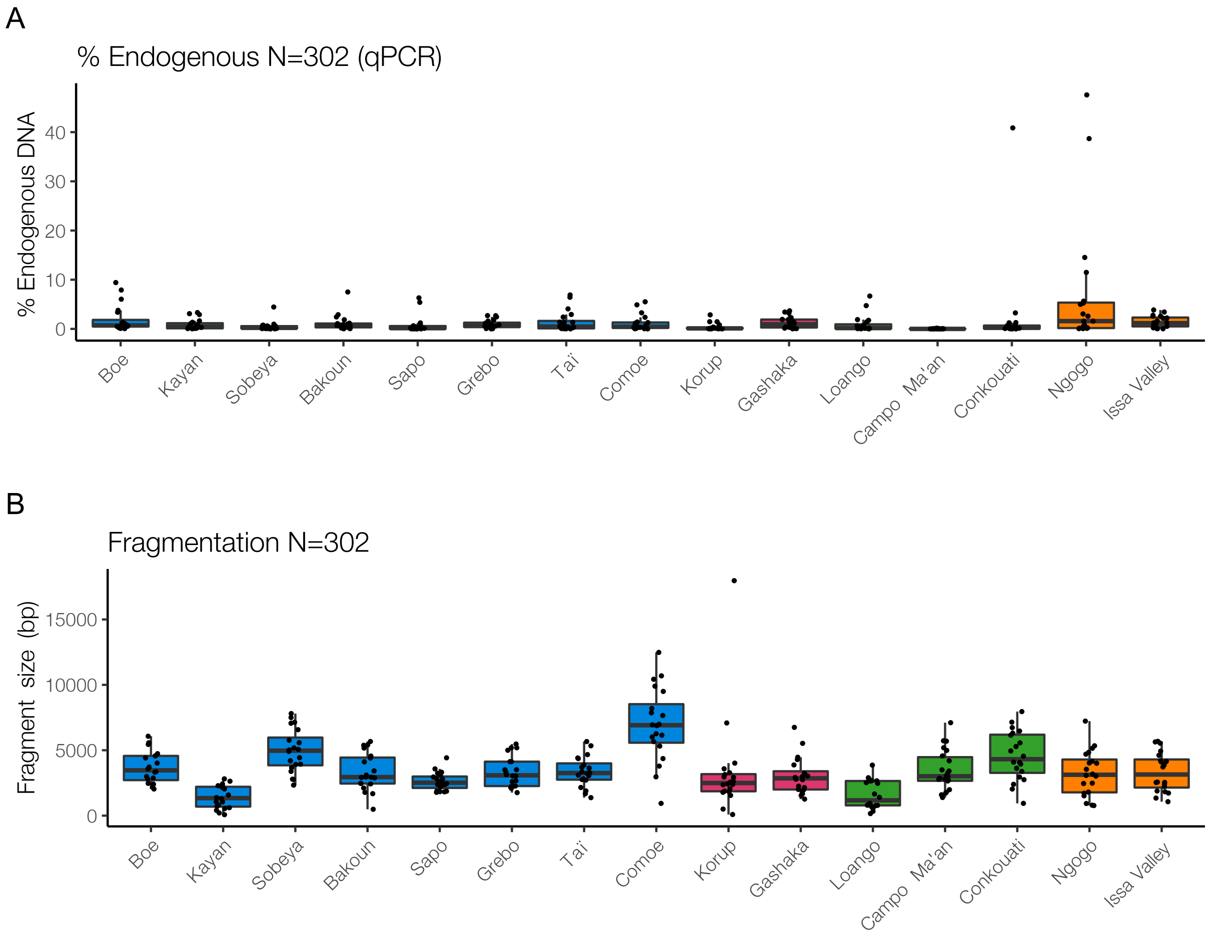
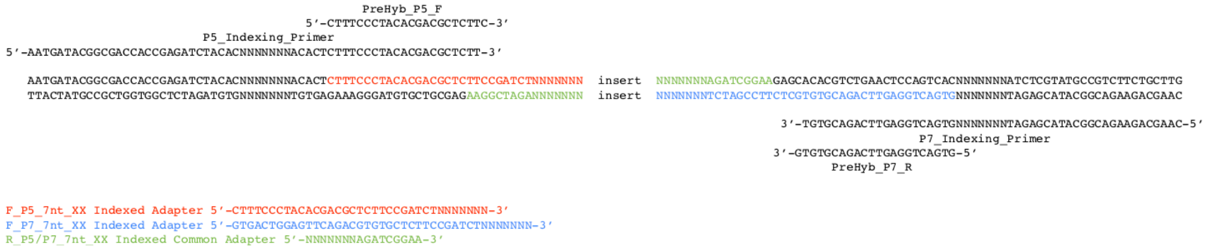


Figure S1 Legend: Distribution of (A) % endogenous content and (B) fragment size for the 302 screened samples from the 15 screened African sites in the PanAfrican programme. The boxplot colors indicate the subspecies membership as seen in Figure 1: blue (western chimpanzee), pink (Nigeria-Cameroon chimpanzee), green (central chimpanzee) and orange (eastern chimpanzee).

Figure S2. Illustration of library construction



```
                                                 PreHyb_P5_F
                                         5′-CTTTCCCTACACGACGCTCTTC-3′
                           P5_Indexing_Primer
5′-AATGATACGGCGACCACCGAGATCTACACNNNNNNNACACTCTTTCCCTACACGACGCTCTT-3′

   AATGATACGGCGACCACCGAGATCTACACNNNNNNNACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNN  insert  NNNNNNNAGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNNATCTCGTATGCCGTCTTCTGCTTG
   TTACTATGCCGCTGGTGGCTCTAGATGTGNNNNNNNTGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGANNNNNNN  insert  NNNNNNNTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGNNNNNNNTAGAGCATACGGCAGAAGACGAAC

                                                                                                          3′-TGTGCAGACTTGAGGTCAGTGNNNNNNNTAGAGCATACGGCAGAAGACGAAC-5′
                                                                                                                             P7_Indexing_Primer
                                                                                                         3′-GTGTGCAGACTTGAGGTCAGTG-5′
                                                                                                                 PreHyb_P7_R


F_P5_7nt_XX Indexed Adapter 5′-CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNN-3′
F_P7_7nt_XX Indexed Adapter 5′-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNNNN-3′
R_P5/P7_7nt_XX Indexed Common Adapter 5′-NNNNNNNAGATCGGAA-3′
```

Figure S2. Final library structure showing the sequences of the indexed adapters and primers used as well as the primers used for amplification of the partial library before and after the first round of hybridization.

Figure S3. Capture performance

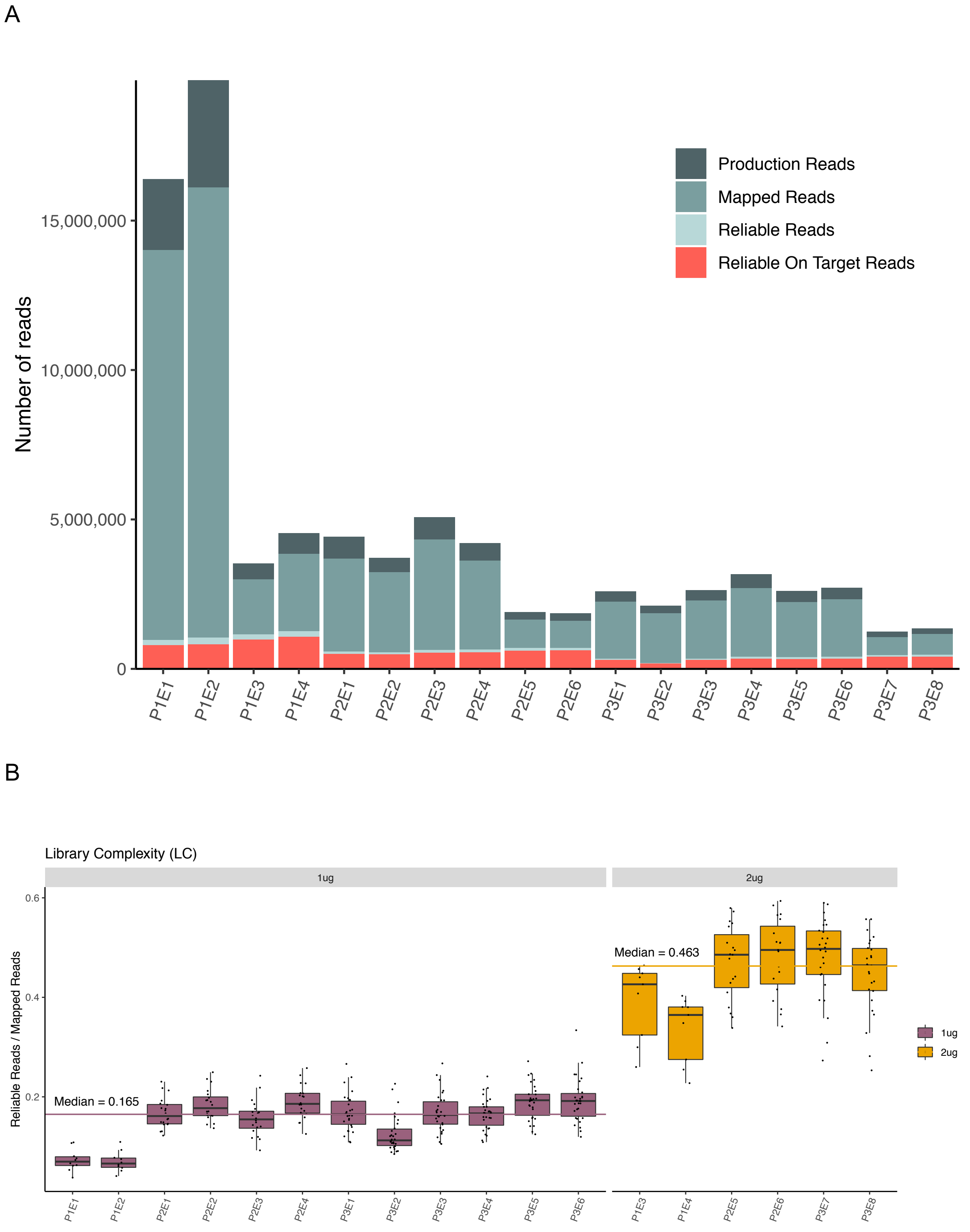


Figure S3 Legend. Capture performance analysis for each 18 capture experiments in 3,75 HiSeq 4000 lanes. (A) Sequencing stats and (B) Library complexity separated by experiments using 1 µg and 2 µg of pooled library, solid lines represent the median LC.

Figure S4. Schematic of library complexity analysis

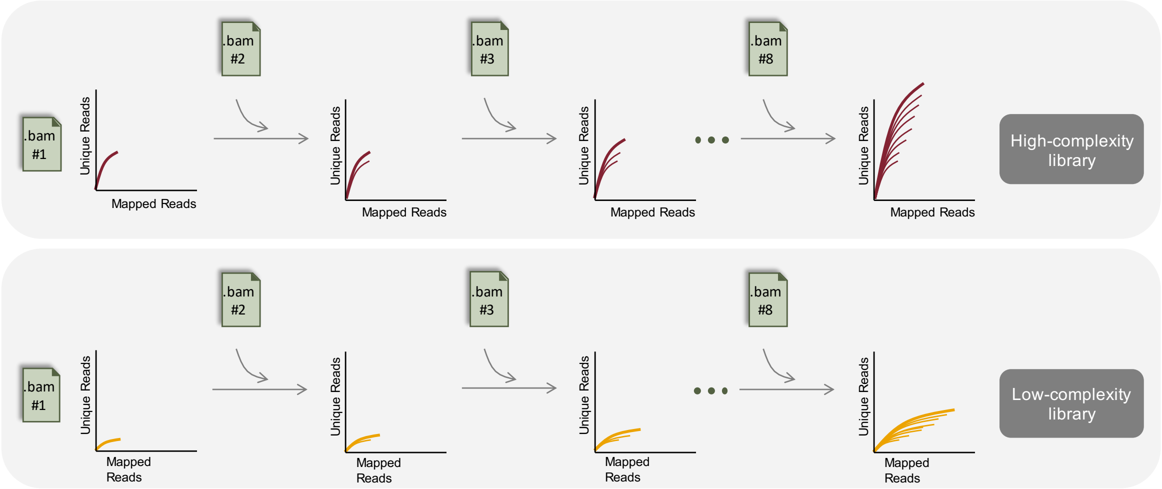


Figure S4 Legend. Schematic representation of library complexity analysis. We add data sequentially, coming from replicate hybridizations through merging BAM files. For each step we subsample without replacement each merged bam file. If the library has high molecular complexity (in red) we see a feathered distribution, where the more data we add, the more unique reads are retrieved. On the other hand, if the library has low molecular complexity, performing additional replicate hybridization does not improve the recovery of new unique reads.

Figure S5. Correlation matrixes of all variables

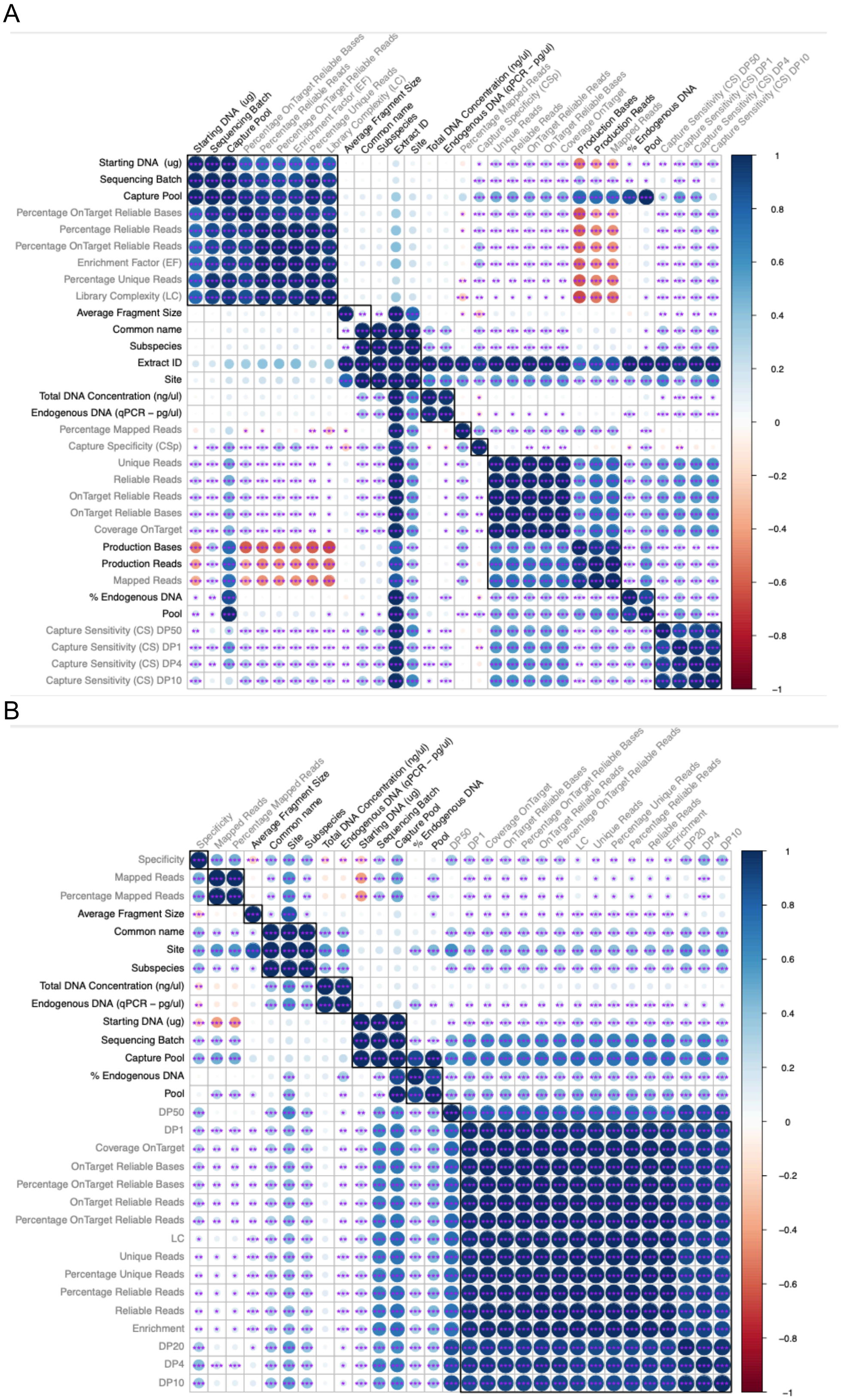
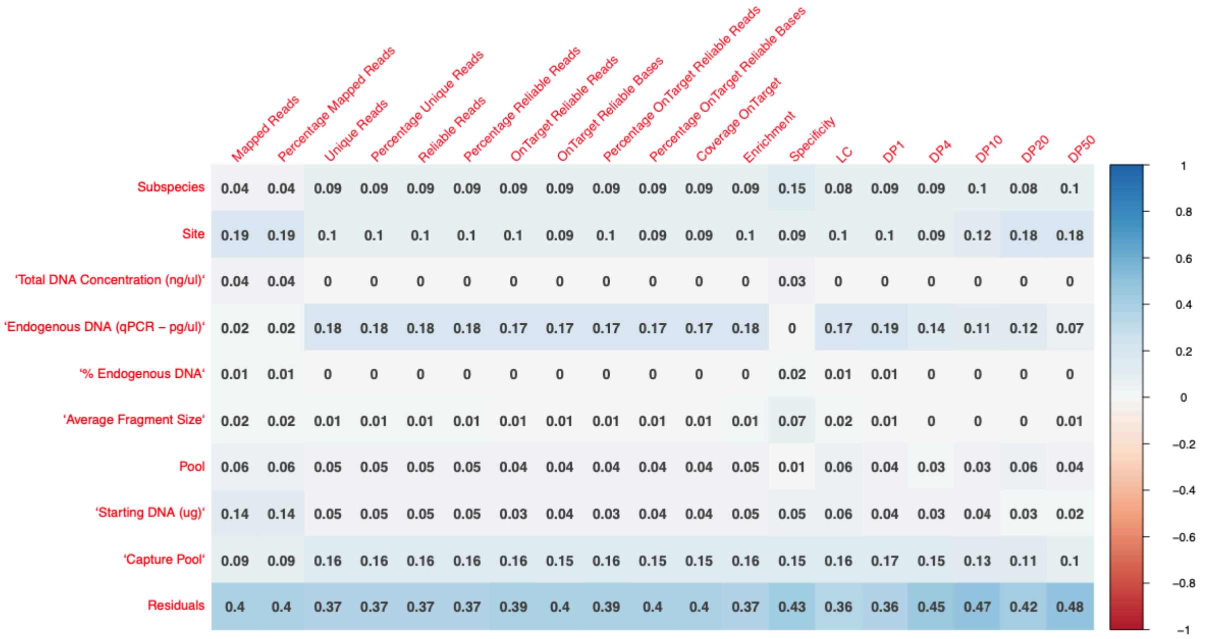


FIGURE S5 Legend. Correlation matrix of all variables included in this study in the (A) full dataset and (B) after having down-sampled each library to 1,500,000 reads. Spearman's rho was estimated when comparing two numeric variables. Cramér's V was estimated among two categorical variables. When comparing a numeric and categorical variable we took the square root of the R-squared statistic derived from a univariate linear model with no transformation on the dependent, numerical values. Experimental variables are illustrated in black text. Performance variables are illustrated in grey text. Clusters of strongly correlated variables where identified, and illustrated by the black squares, using the function cutree() on a hierarchical clustering dendrogram of the same data transformed to distances (1-abs(data)). A cut height of 0.5 was used to identify clusters where intra-cluster distances among variables are greater than or equal to 0.5, and inter-cluster correlations are smaller than 0.5.

Figure S6. Multivariate type I ANOVA

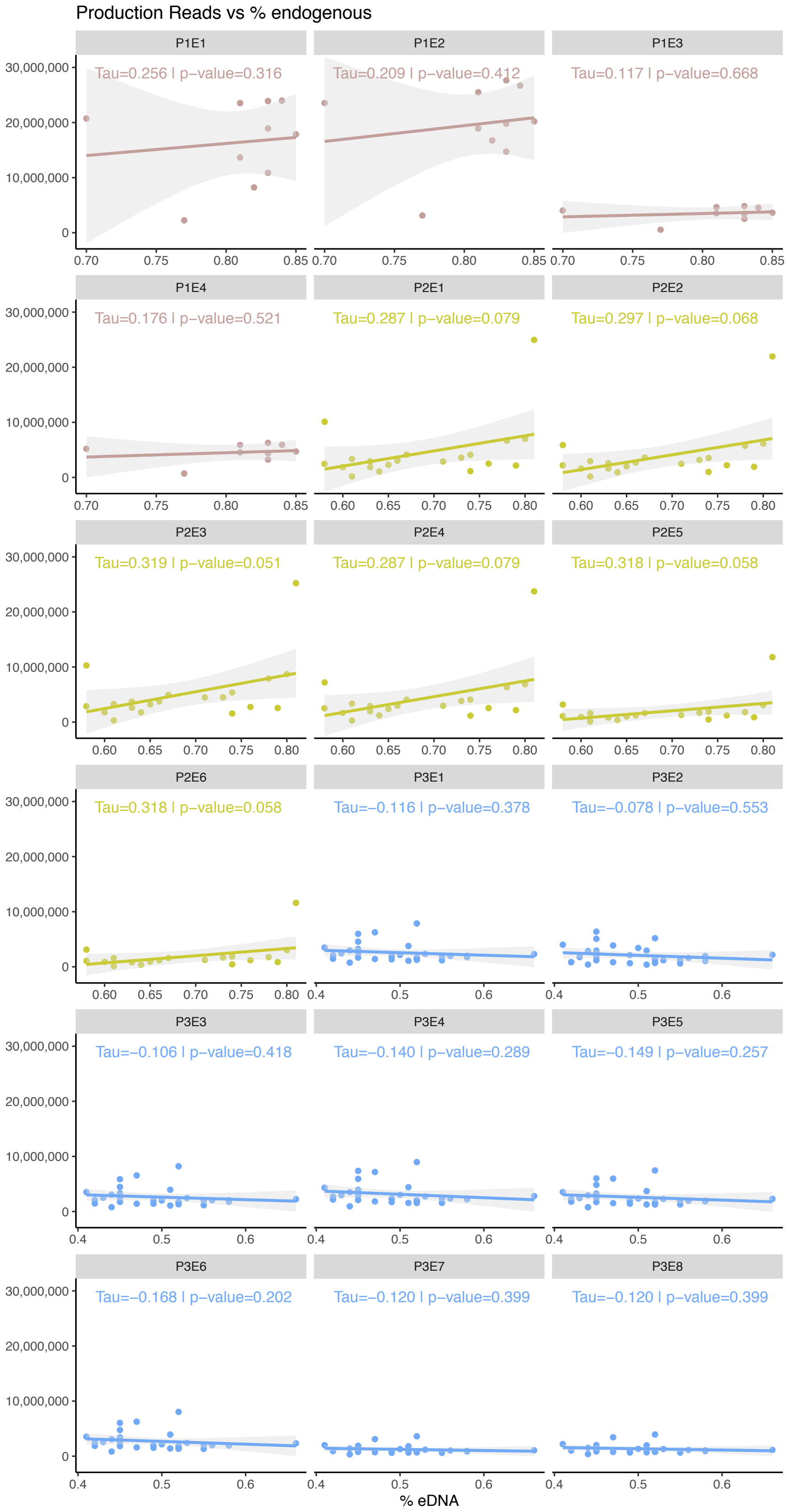


	Mapped Reads	Percentage Mapped Reads	Unique Reads	Percentage Unique Reads	Reliable Reads	Percentage Reliable Reads	OnTarget Reliable Reads	OnTarget Reliable Bases	Percentage OnTarget Reliable Reads	Percentage OnTarget Reliable Bases	Coverage OnTarget	Enrichment	Specificity	LC	DP1	DP4	DP10	DP20	DP50
Subspecies	0.04	0.04	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.15	0.08	0.09	0.09	0.1	0.08	0.1
Site	0.19	0.19	0.1	0.1	0.1	0.1	0.1	0.09	0.1	0.09	0.09	0.1	0.09	0.1	0.1	0.09	0.12	0.18	0.18
‘Total DNA Concentration (ng/ul)‘	0.04	0.04	0	0	0	0	0	0	0	0	0	0	0.03	0	0	0	0	0	0
‘Endogenous DNA (qPCR – pg/ul)‘	0.02	0.02	0.18	0.18	0.18	0.18	0.17	0.17	0.17	0.17	0.17	0.18	0	0.17	0.19	0.14	0.11	0.12	0.07
‘% Endogenous DNA‘	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0.02	0.01	0.01	0	0	0	0
‘Average Fragment Size‘	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.07	0.02	0.01	0	0	0	0.01
Pool	0.06	0.06	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.05	0.01	0.06	0.04	0.03	0.03	0.06	0.04
‘Starting DNA (ug)‘	0.14	0.14	0.05	0.05	0.05	0.05	0.03	0.04	0.03	0.04	0.04	0.05	0.05	0.06	0.04	0.03	0.04	0.03	0.02
‘Capture Pool‘	0.09	0.09	0.16	0.16	0.16	0.16	0.16	0.15	0.16	0.15	0.15	0.16	0.15	0.16	0.17	0.15	0.13	0.11	0.1
Residuals	0.4	0.4	0.37	0.37	0.37	0.37	0.39	0.4	0.39	0.4	0.4	0.37	0.43	0.36	0.36	0.45	0.47	0.42	0.48

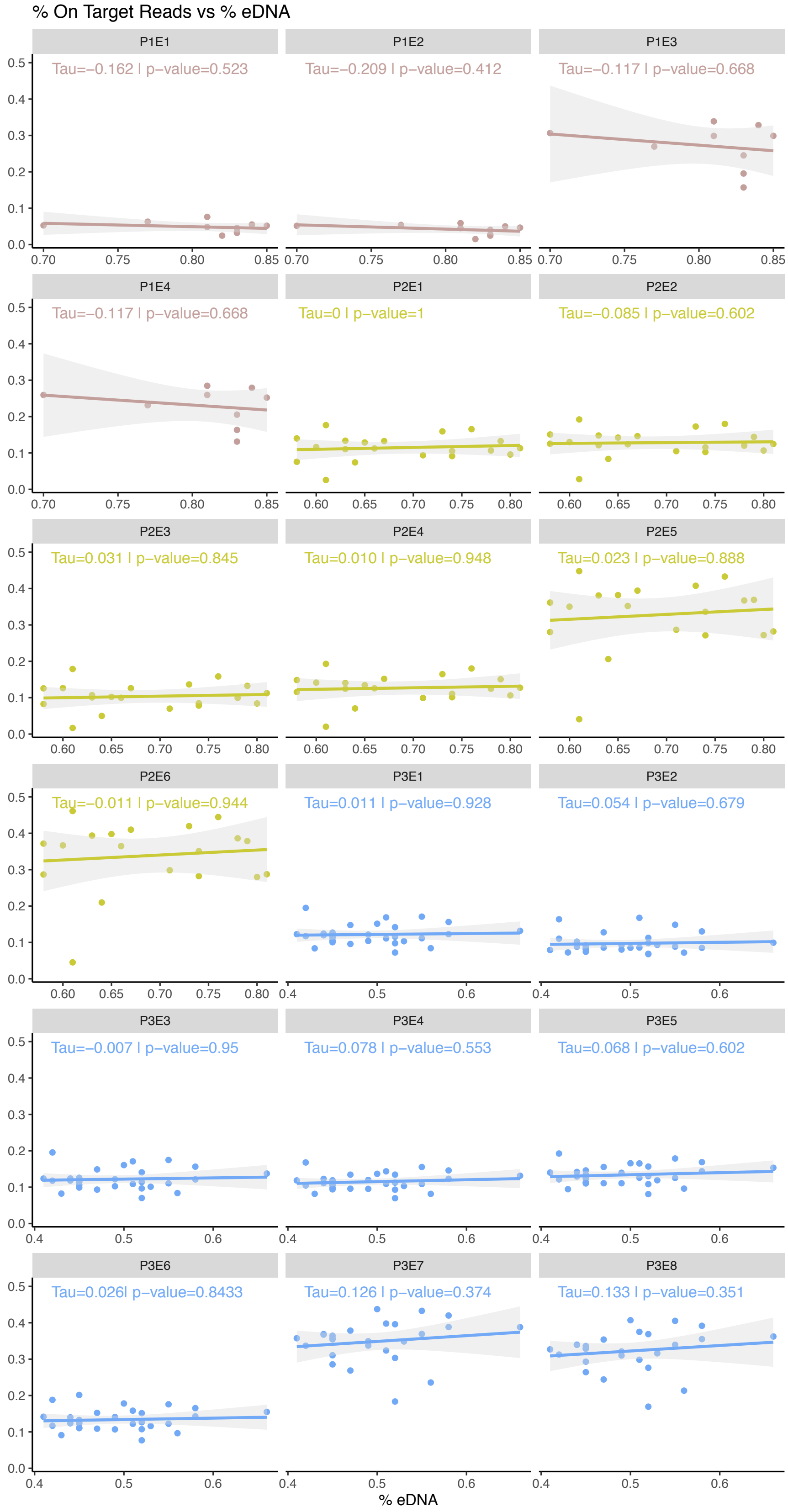
Figure S6 Legend. Multivariate type I analysis of variance. Estimated variance explained from multivariate type I ANOVA of the experimental variables affecting performance summary statistics. Figure is an extension of Figure 4. Estimates are derived from 1,500,000 read down-sampled libraries.

Figure S7. Correlation dot plots

(A)



(B)


% On Target Reads vs % eDNA
P1E1
Tau=−0.162 | p−value=0.523
P1E2
Tau=−0.209 | p−value=0.412
P1E3
Tau=−0.117 | p−value=0.668
P1E4
Tau=−0.117 | p−value=0.668
P2E1
Tau=0 | p−value=1
P2E2
Tau=−0.085 | p−value=0.602
P2E3
Tau=0.031 | p−value=0.845
P2E4
Tau=0.010 | p−value=0.948
P2E5
Tau=0.023 | p−value=0.888
P2E6
Tau=−0.011 | p−value=0.944
P3E1
Tau=0.011 | p−value=0.928
P3E2
Tau=0.054 | p−value=0.679
P3E3
Tau=−0.007 | p−value=0.95
P3E4
Tau=0.078 | p−value=0.553
P3E5
Tau=0.068 | p−value=0.602
P3E6
Tau=0.026| p−value=0.8433
P3E7
Tau=0.126 | p−value=0.374
P3E8
Tau=0.133 | p−value=0.351
0.0
0.1
0.2
0.3
0.4
0.5
0.70
0.75
0.80
0.85
0.60
0.65
0.70
0.75
0.80
0.4
0.5
0.6
% eDNA

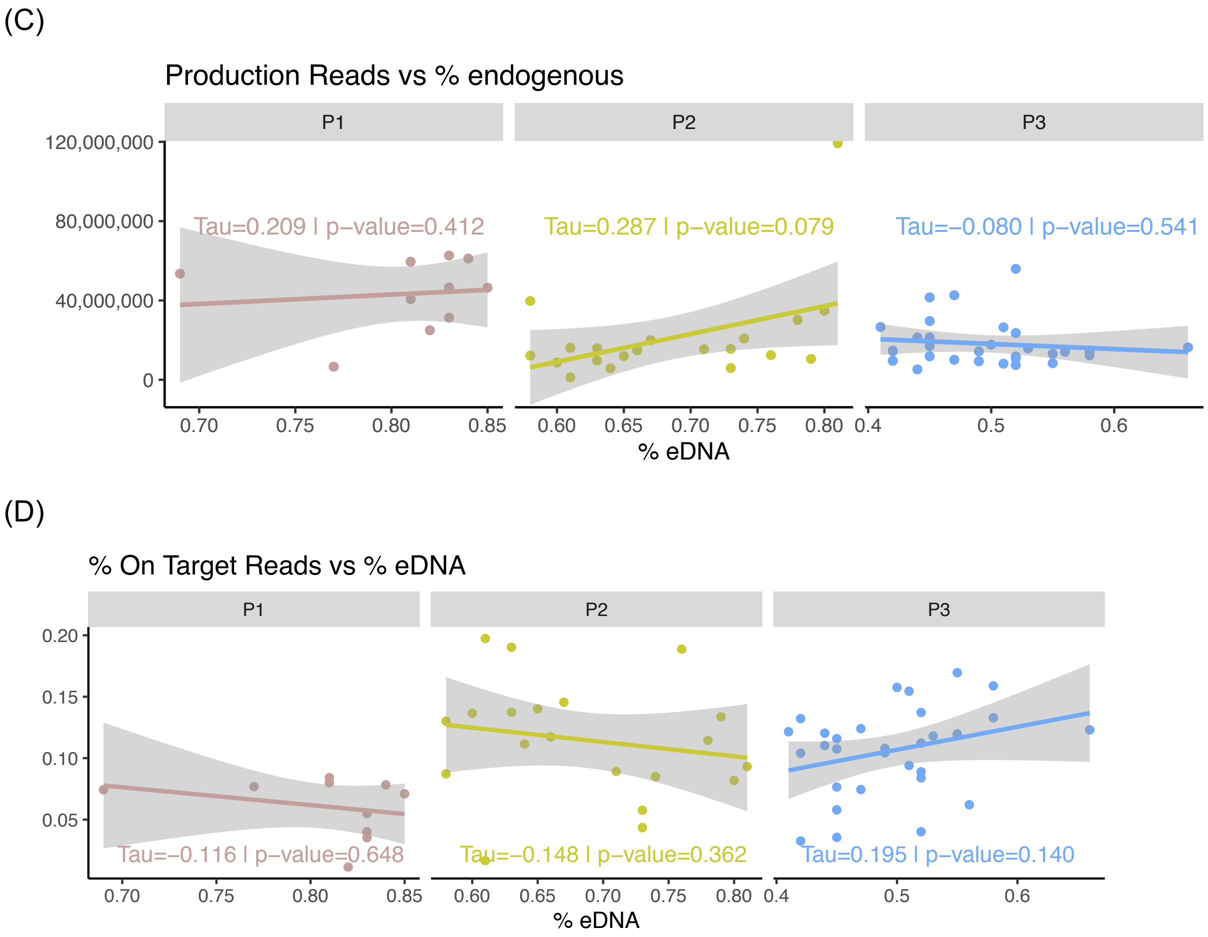


Figure S7 Legend. Kendall’s correlation between (A) Production Reads and (B) % On Target Reads versus % eDNA in each Hybridization experiment. No statistically significant correlation of eDNA content with both summary statistics although some hybridizations in P2 exhibit a slight positive correlation, possibly due to one outlier. In (C) Production Reads and (D) % On Target Reads we show the same correlation plots with % eDNA but now with data coming from merged hybridizations.

Figure S8. Distribution of raw reads across pools

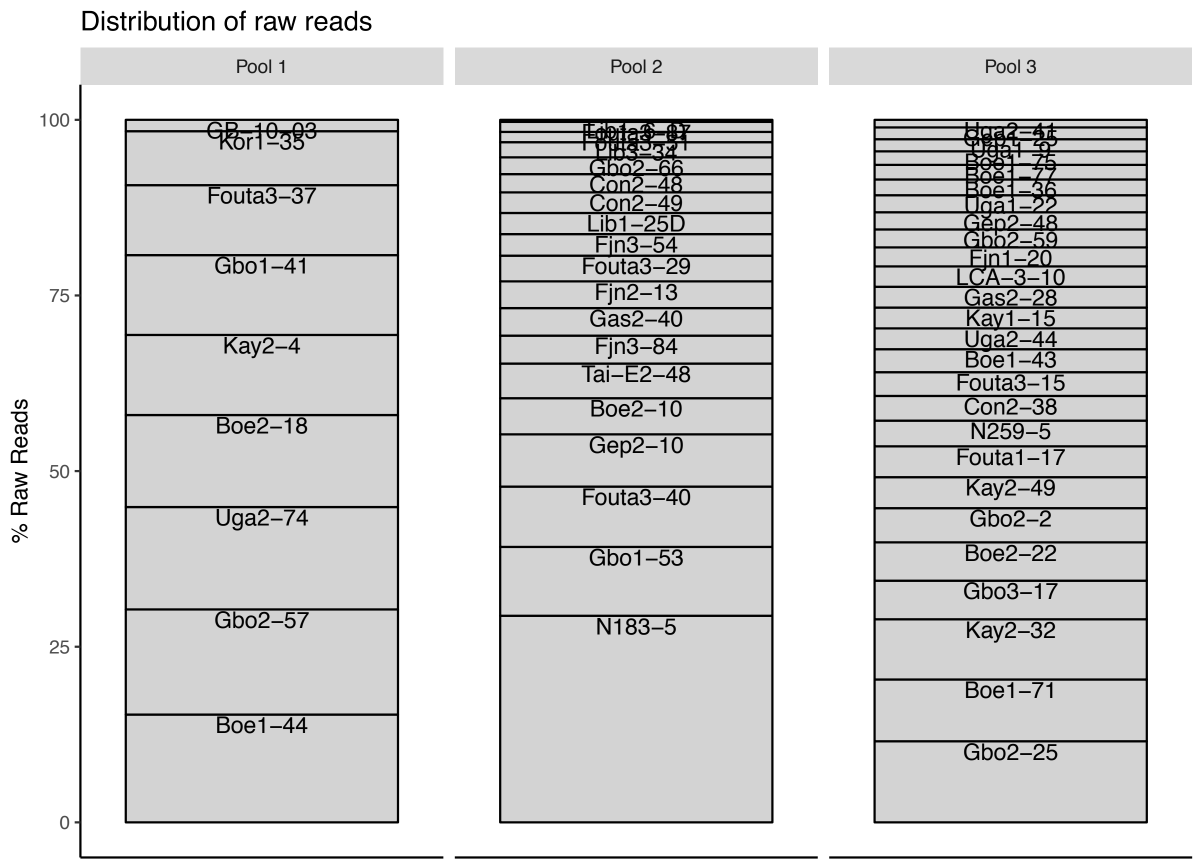


Figure S8 Legend. Percentage of raw reads (production reads) sequenced for each library in each pool to detect which samples are taking a greater proportion of the total production reads.

Figure S9. Impact of total DNA in pooled libraries on average unique read count.

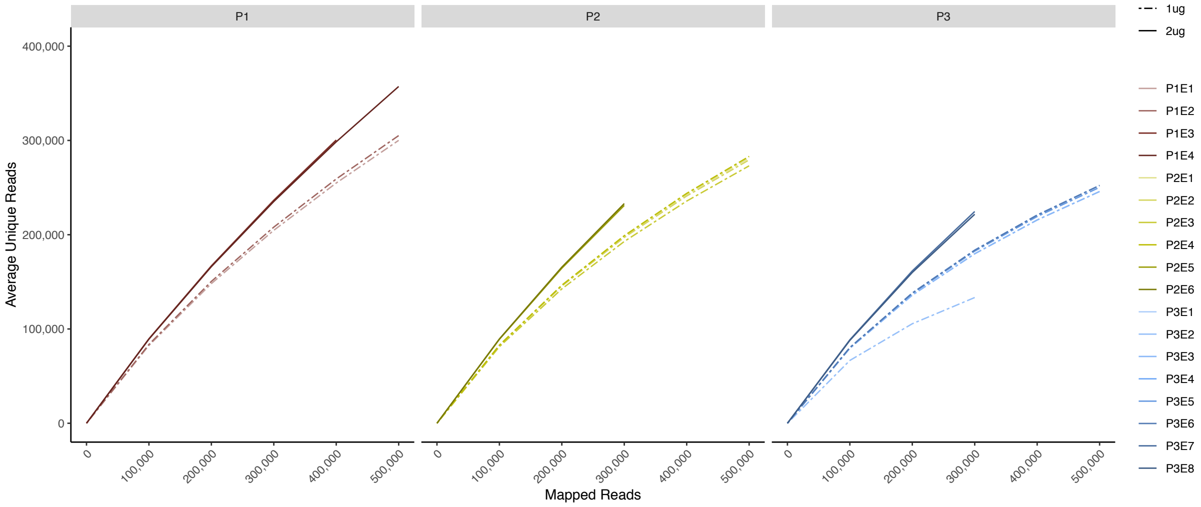


Figure S9 Legend. Comparison of pooling 1 μg or 2 μg DNA for capture. We subsampled without replacement reads in each hybridization (average of all samples within a pool) and obtained the corresponding average unique reads. The averages are done if all samples in the pool have data in any given point (for that reason sample Lib1-6D from P2 is excluded). Dashed lines indicate 1 μg of starting DNA for capture while solid lines are the hybridizations with 2 μg of starting DNA. Colors indicate each hybridization.

Figure S10. Library complexity by replicate hybridizations.

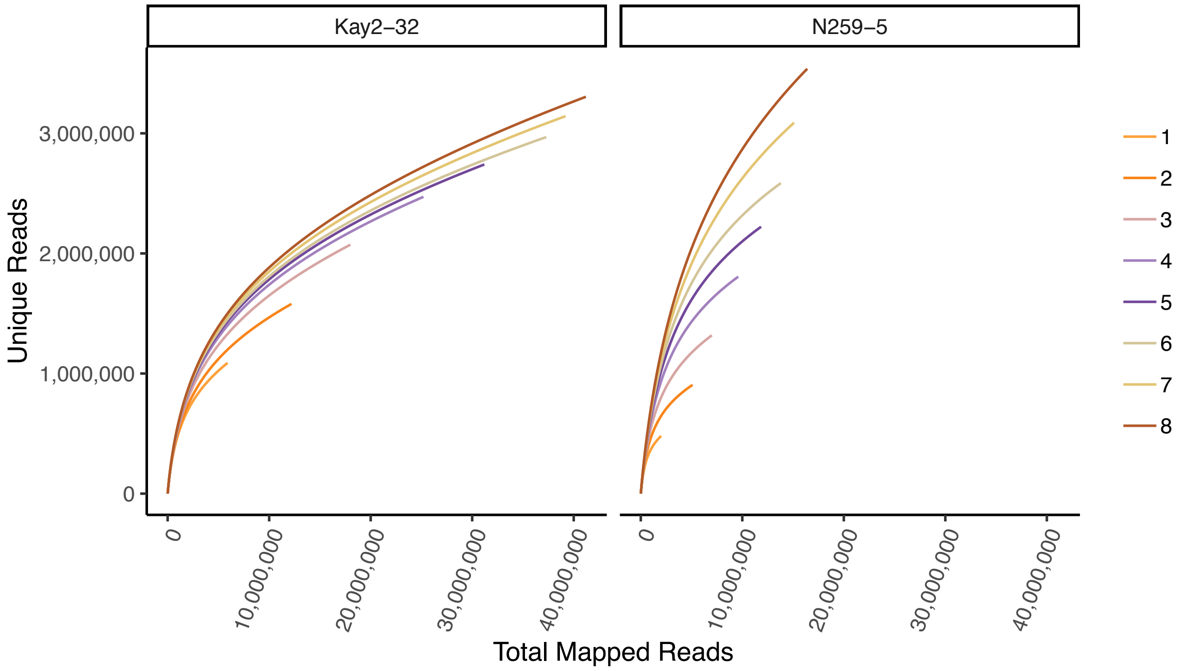


Figure S10 Legend. Library complexity plots of two samples belonging to P3. Each line represents data coming from cumulative replicate hybridizations. *Line 1* indicates data coming for only one hybridization, *line 2* indicates combined data from 2 hybridization, until *line 8* that indicates combined data from all 8 hybridization replicates. Library Kay2-32 has low library complexity and cannot be increased by additional hybridizations. However, the majority of samples behave similar to the example sample N259-5. By performing additional hybridizations, it is possible to retrieve new unique reads.

Figure S11. Capture sensitivity by depth and pool.

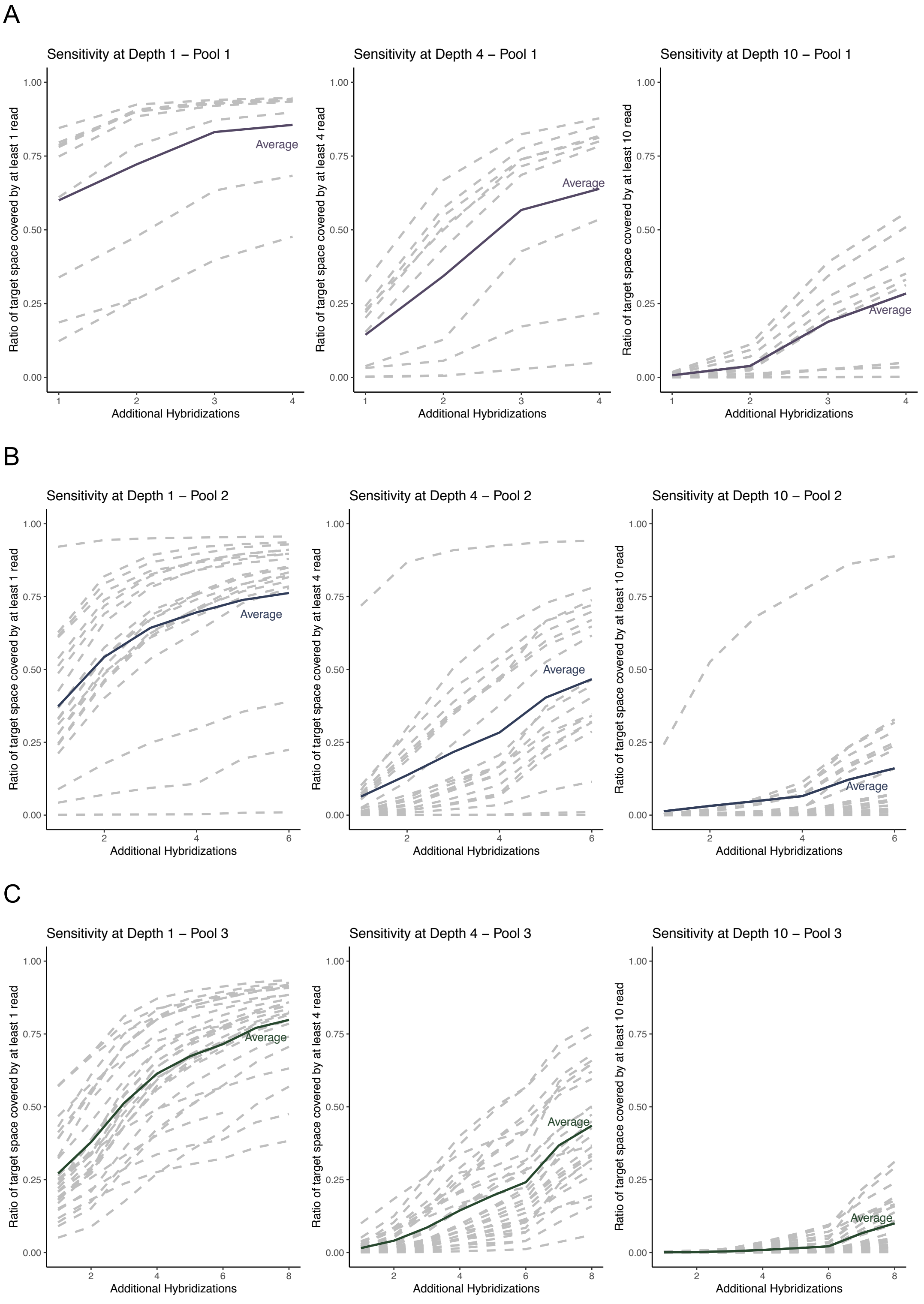


Figure S11 Legend. Sensitivity (ratio of target space covered by at least a certain number of reads) at depth 1, 4 and 10 for samples in (A) P1, (B) P2 and (C) P3. Each grey dashed line

represents a sample from each pool and the colored solid line is the average of all samples within the pool.

Figure S12. Venn diagram never covered regions.

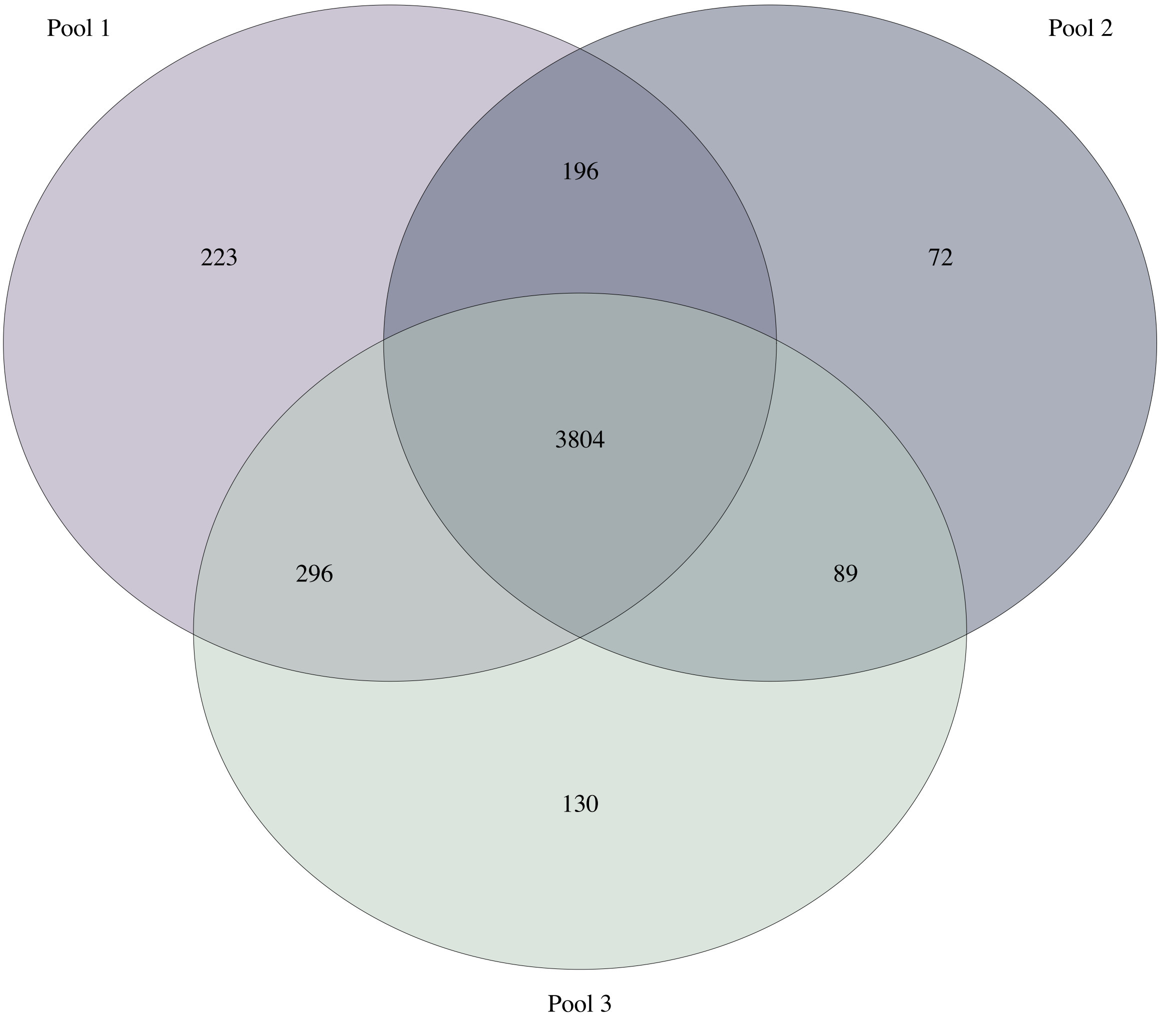


Figure S12 Legend. Intersection of regions never covered after 4, 6 and 8 additional hybridizations for Pool1, Pool2 and Pool3, respectively. In Pool1, out of the total 243,190 regions, 4,519 are never covered (1.85%); in Pool 2, it is 4161 out of 243,190 total regions (1.71%); and for Pool 3 it is 4319 out of 243,190 total regions (1.77%). From those, the same 3804 regions are never covered in all experiments (1,564%).

Figure S13. Sequencing effort data saturation.

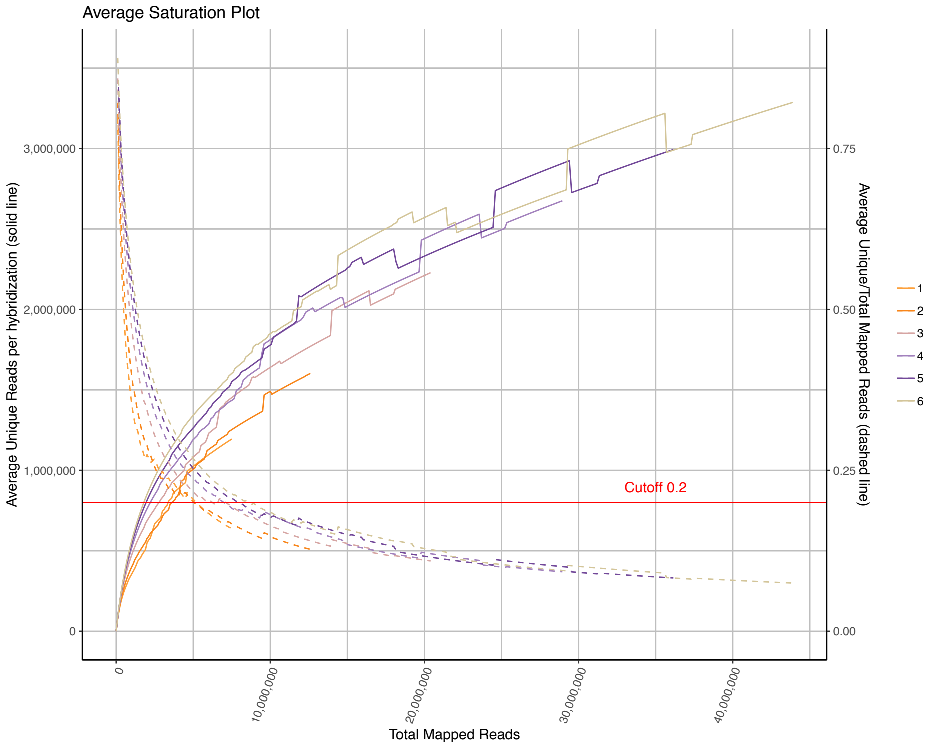


Figure S13 Legend. Sequencing Effort. Solid lines represent the sample average number of unique reads after merging data from additional hybridizations (numeric key). Dashed lines represent the average number of unique reads normalized by the number of mapped reads. The cutoff is set at 20% (right Y axis). We estimated for each additional hybridization a sample average and plotted the number of unique reads averaged across samples (left Y axis) and also the proportion of unique reads by total mapped reads averaged across samples (right Y axis), with the total mapped reads (X axis).

Figure S14. Sequencing summary statistics by SeqBatch.

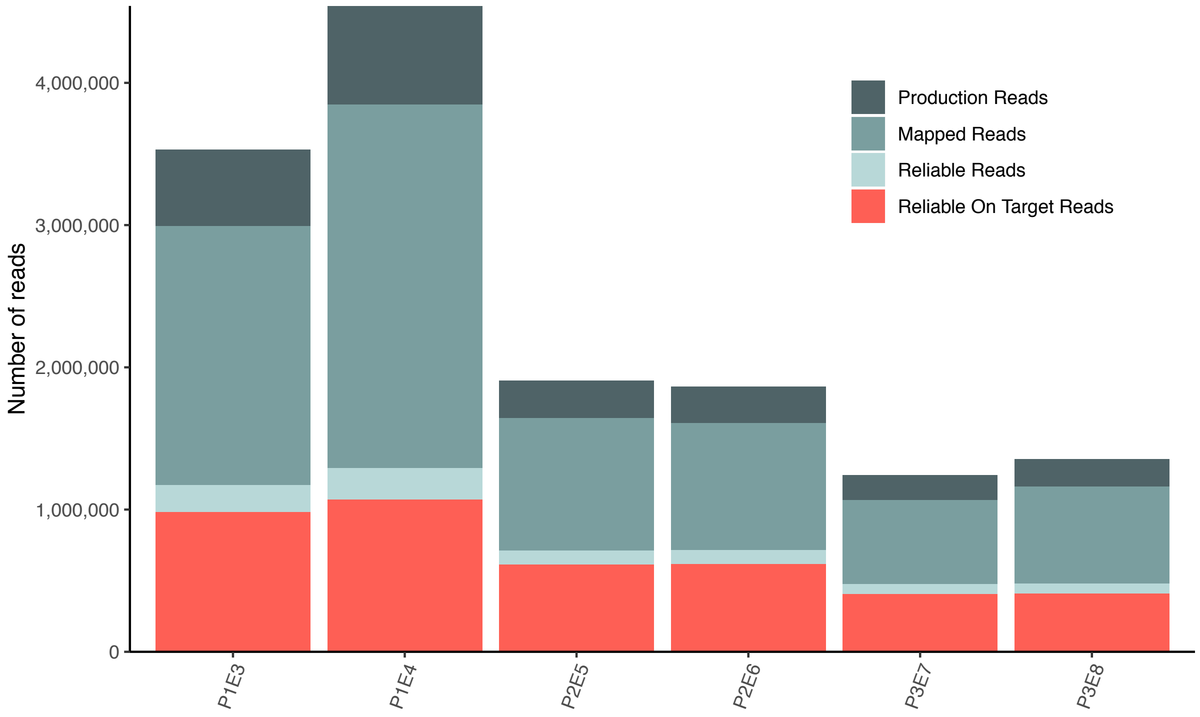


Figure S14 Legend. Sequencing stats for the SeqBatch 3 (P1E3, P1E4, P2E5, P2E6, P3E7, P3E8). Y axis represents the average number of reads per library belonging to each pool. On average we obtain 3.5 million reads per library in hybridizations from P1, around 2 million reads per library in hybridizations from P2 and around 1.5 million reads per library for hybridizations from P3. The percentage of reliable reads is 27.87% in P1E3 and 23.58% in P1E4; 32.12% in P2E5 and 33.06% in P2E6; 32.71% in P3E7 and 30.17% in P3E8.

Figure S15. Average library complexity curves

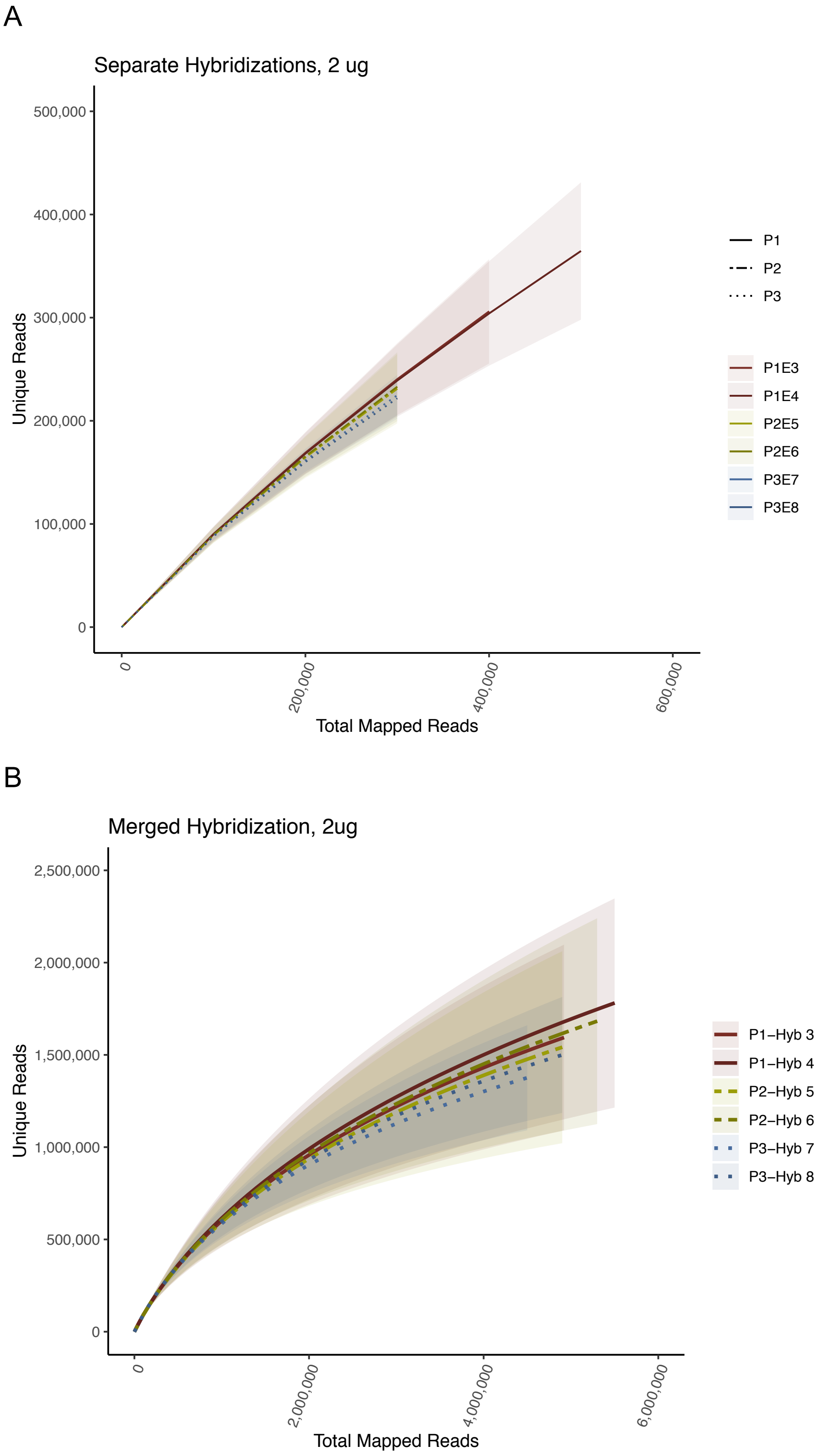


Figure S15 Legend. A) Average library complexity curve for each individual hybridization (starting with 2µg). B) Average library complexity curve for merged hybridizations (only hybridizations with starting DNA of 2 µg). Solid line is P1, two-dashed line is P2 and dotted line is P3. Sample Lib1-6D in P2 was removed from the analysis due to low coverage.

Figure S16. Sensitivity by pool at various depth.

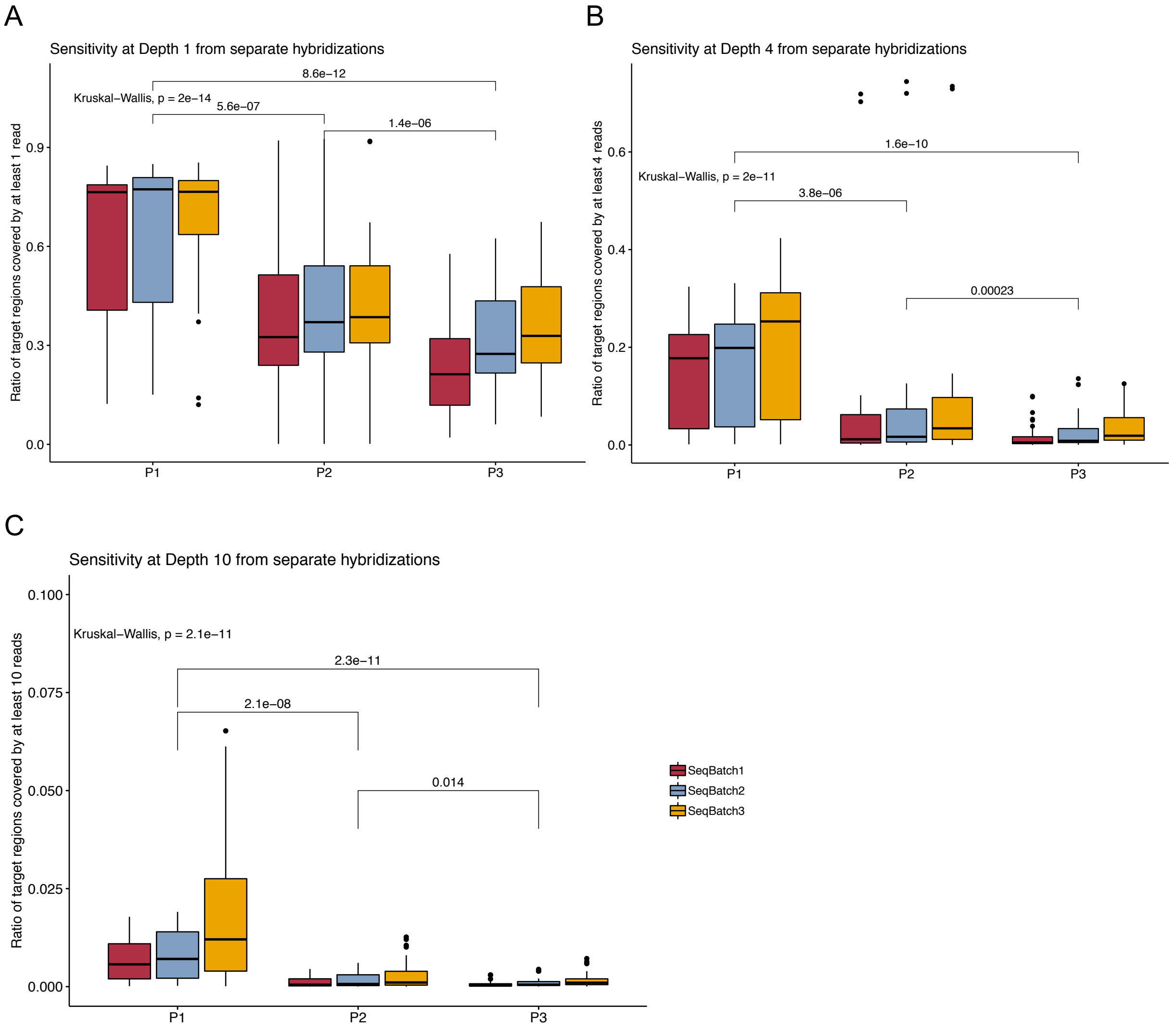


Figure S16 Legend. Capture performance analysis of sensitivity from separate hybridizations and plotting together the data coming from the same Sequencing Batch (color). Small pools have higher sensitivity than larger pools. (A) Capture sensitivity at depth 1, (B) capture sensitivity at depth 4 and (C) capture sensitivity at depth 10.

Figure S17. Variance explained by pool on capture sensitivity.

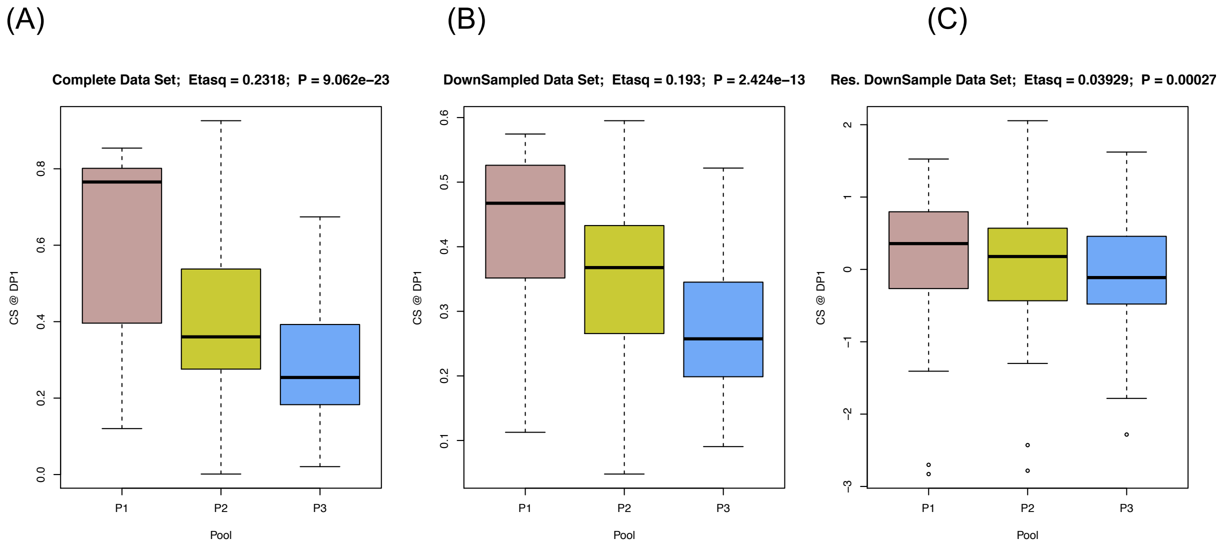


Figure S17 Legend. Multivariate Type I ANOVA of the variance explained of ‘Pool’ on capture sensitivity (CS) at Depth 1. (A) Whole data set. (B) Libraries down-sampled at 1,500,000 reads. (C) Residuals.

Figure S18. Variation in capture sensitivity across pools.

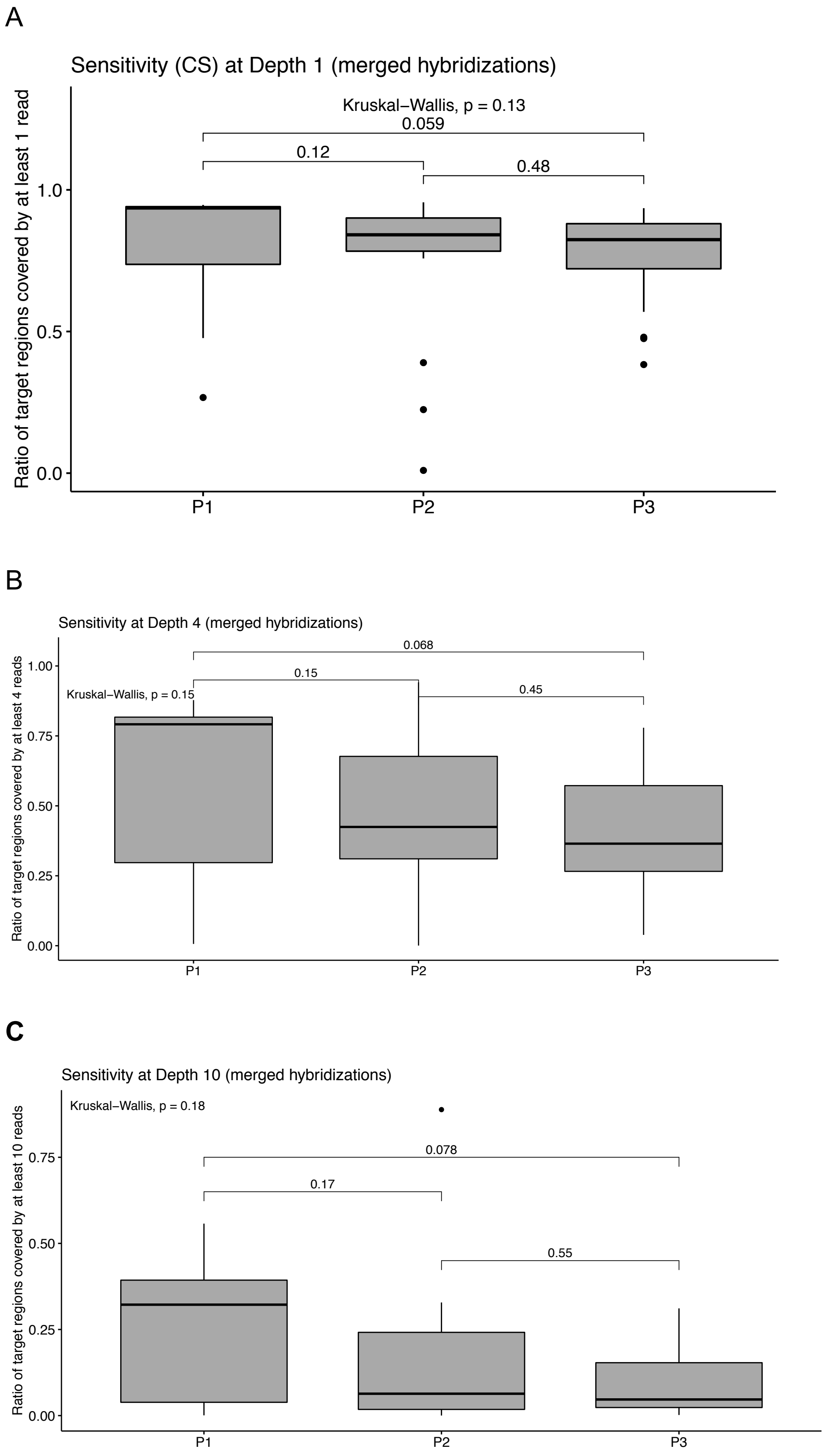


Figure S18 Legend. Capture performance analysis of sensitivity after merging data from additional hybridizations. (A) Capture sensitivity at depth 1, (B) Capture sensitivity at depth 4 and (B) capture sensitivity at depth 10.

References

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. doi: 10.1111/2041-210X.12871

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. doi: 10.1093/nar/gkr771

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, *22*(5), 939–946. doi: 10.1101/gr.128124.111