Full Title:

Urine as a high-quality source of host genomic DNA from wild populations

Short title:

Noninvasive genomic methods

Key words:

Genomic methods, endangered populations, primates, population genetics – empirical

Manuscript subject:

Molecular and statistical advances

Authors:

Andrew T. Ozga*[1,2], Timothy H. Webster*[3,4], Ian C. Gilby[,5,6], Melissa A. Wilson[2,3], Rebecca S. Nockerts[7], Michael L. Wilson[7,8], Anne E. Pusey[9], Yingying Li[10], Beatrice H. Hahn[10], and Anne C. Stone[2,5,6]

Affiliations:

*Contributed Equally

[1]Halmos College of Natural Science and Oceanography, Nova Southeastern University
[2]Center for Evolution and Medicine, Arizona State University
[3]School of Life Sciences, Arizona State University
[4]Department of Anthropology, University of Utah
[5]School of Human Evolution and Social Change, Arizona State University
[6]Institute of Human Origins, Arizona State University
[7]Department of Anthropology, University of Minnesota
[8]Department of Ecology, Evolution and Behavior, University of Minnesota
[9]Evolutionary Anthropology, Duke University
[10]Departments of Medicine and Microbiology, Perelman School of Medicine, University of Pennsylvania

Data citation:

Ozga AT, Webster TH, Gilby IC, Wilson MA, Nockerts RS, Wilson ML, Pusey AE, Li Y, Hahn BH, Stone AC. *Pan troglodytes schweinfurthii* raw sequence reads, NCBI SRA, PRJNA508503.

To whom correspondence should be addressed:

Andrew T. Ozga
aozga@nova.edu
Timothy H. Webster
timothy.h.webster@utah.edu

Abstract

The ability to generate genomic data from wild animal populations has the potential to give unprecedented insight into the population history and dynamics of species in their natural habitats. However, in the case of many species, it is impossible legally, ethically, or logistically to obtain tissues samples of high-quality necessary for genomic analyses. In this study we evaluate the success of multiple sources of genetic material (feces, urine, dentin, and dental calculus) and several capture methods (shotgun, whole-genome, exome) in generating genome-scale data in wild eastern chimpanzees (*Pan troglodytes schweinfurthii*) from Gombe National Park, Tanzania. We found that urine harbors significantly more host DNA than other sources, leading to broader and deeper coverage across the genome. Urine also exhibited a lower rate of allelic dropout. We found exome sequencing to be far more successful than both shotgun sequencing and whole-genome capture at generating usable data from low-quality samples such as feces and dental calculus. These results highlight urine as a promising and untapped source of DNA that can be noninvasively collected from wild populations of many species.

Introduction

The development of methods to generate genetic data from noninvasively collected samples revolutionized the study of wild animal populations, allowing for DNA research without the capture or even observation of species of interest (Kohn & Wayne, 1997; Waits & Paetkau, 2005). While studies of individual DNA markers improved our understanding of behavior, ecology, and evolution, recent advances in massively parallel sequencing strategies make it possible to incorporate information from across the entire genome, giving unprecedented insight into the evolution and population history of non-model species (Ellegren, 2014). However, for many species, it is impossible legally, ethically, or logistically to obtain high-quality tissue samples required for large-scale genomic analyses. It is therefore critically important to develop and evaluate methods for sampling and capturing genome-scale data from noninvasive and alternative sources.

While a variety of noninvasively collected biological materials have been used in DNA analyses, feces have been the primary target of recent attempts to generate genomic data. Rich in gut epithelial cells and often the most abundant, easiest to collect source of DNA in the environment, feces have long played a role in noninvasive genetic analyses (Constable, Ashley, Goodall, & Pusey, 2001; Hoss, Kohn, Paabo, Knauer, & Schroder, 1992; Kohn & Wayne, 1997). However, the retrieval of DNA from feces presents a number of difficulties. Challenges, including low DNA yields, DNA fragmentation and degradation (Deagle, Eveson, & Jarman, 2006) and the presence of PCR inhibitors, can lead to genotyping errors (Taberlet, Waits, & Luikart, 1999). Moreover, DNA recovered from fecal material is dominated by microbes (>95% exogenous DNA), which further complicates genotyping (Perry, Marioni, Melsted, & Gilad, 2010). For genetic analyses involving small number of markers, these challenges are well understood and can be overcome. However, these problems are exacerbated in massively parallel sequencing, which typically requires higher quantities and qualities of input DNA, and generates almost entirely microbial data due to the very low levels of host DNA in samples.

The main strategy that has been employed to combat these problems is enrichment of host DNA. In this vein, there have been three major methodological developments. Perry and colleagues (2010) first enriched DNA from feces on a genomic scale by using custom chimpanzee baits designed to capture approximately 1.5 Mb of sequence across six western chimpanzees (*Pan troglodytes verus*). While successful, this method required a reference genome to design baits and was cost prohibitive for producing genome-scale datasets. To address these challenges, Snyder-Mackler and colleagues developed a protocol to create RNA baits from high-quality host DNA and improve post-capture enrichment (2016). However, for this method, bait requirements—notably high-quality host DNA—and low sequencing coverage of host DNA remain barriers for some study systems and questions. Recently, Chiou and Bergey introduced a method that exploits differences in CpG-methylation densities between vertebrate and bacterial genomes to capture host DNA, alleviating the need for high-quality host material or reference genome to design baits (2018). However, CpG content varies substantially across the genomes of primates and other mammals (Han, Su, Li, & Zhao, 2008), thus targeting these regions specifically may bias the regions captured.

Despite these improvements to both capture and enrichment, DNA capture from feces is still far less efficient than from high-quality tissues. This leads to a tradeoff: attempting to capture large genomic regions leads to very low sequence coverage; however, targeting a subset of the genome can lead to biases. A compromise would be to target a small subset of the genome that is biologically important. One potential option is exome sequencing, a capture-based method that targets the entire coding region of the genome, comprising approximately 1.5% of the total length of the genome. Coding regions are among the best understood in the genome and are of great evolutionary and conservation interest (Bataillon et al., 2015; George et al., 2011; Hvilsom et al., 2012). Because exome sequencing is so widely used in human genomics, many commercial kits are available and much cheaper than custom alternatives. Human exome baits have been successfully used in a number of nonhuman primate studies (George et al., 2011; Jin et al., 2012; Vallender, 2011) and have been shown to work in primate species as distantly related from humans as Strepsirrhines (Webster, Guevara, Lawler, & Bradley, 2018). Moreover, recent work has shown that exome capture successfully enriches host DNA in chimpanzee fecal samples (Hernandez-Rodriguez et al., 2018; White et al., 2019). However, some of this work involves first screening for endogenous content using quantitative PCR (qPCR), which although successful, can be a limiting factor for smaller labs at the scales for population genomics. For example, after screening 1,780 fecal samples, White and colleagues estimated 101 samples contained enough endogenous DNA for sequencing (>1%) (White et al., 2019).

In addition to methodological development, turning to other sources of biological material might improve sequencing success in wild populations. Efforts up to this point have focused almost exclusively on feces, and many other noninvasive alternatives remain to be explored. Urine, in particular, is abundant for many large-bodied species, and has been used, albeit infrequently, as a source of DNA collected noninvasively from the environment (Hedmark et al., 2004; Sastre et al., 2009; Valiere & Taberlet, 2000). Although difficult to obtain in certain field conditions, urine contains far fewer microbes than feces, does not contain traces of dietary DNA, and lacks many inhibiting compounds commonly found in feces that impact PCR success (Hausknecht, Gula, Pirga, & Kuehn, 2007; Inoue, Inoue-Murayama, Takenaka, & Nishida, 2007;

Thomas-White, Brady, Wolfe, & Mueller, 2016). Another source of interest is skeletal material, which is often found at field sites and in museum collections. Dentin and dental calculus, in particular, are both capable of yielding host nuclear DNA (Ziesemer et al., 2018). Combining data from historic populations with those from contemporary populations has the potential to provide genomic insight into wild populations on a scale not yet fully realized.

In this study, we evaluate the success of several sources of host DNA and capture methods in generating genome-scale data in a population of wild, endangered animals. Specifically, we extracted and captured endogenous DNA from feces, dental calculus, dentin, and urine recovered from wild chimpanzees (*P. t. schweinfurthii*) from Gombe National Park, Tanzania. From these data we compared the success of whole-genome capture and targeted exome capture. We demonstrate that urine harbors the highest concentration of endogenous DNA of the materials sampled in this study. For other sources, whole-genome sequencing appears possible, but not cost-effective. Employing a targeted approach, such as exome capture, reduces the amount of sequence obtained in the genome, but it may result in increased sequencing efficiency. Finally, we show that genotypes generated from fecal and urine samples exhibit high levels of concordance and argue that genotypes from urine are less subject to contamination. Together, our results demonstrate that, while further methodological advances might improve host DNA extraction in feces, dentin, and dental calculus, urine is a promising source of noninvasive DNA from which genome-scale data can be easily generated. We anticipate the ability to generate genomic data from urine to be broadly useful across study systems, including many protected species.

Materials and Methods

Sample Collection and Extraction

We collected fecal samples in RNAlater from four wild chimpanzees as described (Stone et al., 2010) (7069, 7150, 7365, and 7507) from Gombe National Park between August 2011 and January 2014 and shipped them to the University of Pennsylvania for storage at -80°C. Using a sterile cut pipette tip, we removed roughly 200 µL of the fecal slurry and extracted DNA using QIAamp DNA Stool Minikit (Qiagen) according to manufacturer's protocol. To obtain enough DNA, we repeated this process 8-12 times for each sample, then pooled and desiccated each sample down to 50-100 µL. We combined a total of 2 µg of DNA and molecular grade H_2O into a 50 µl tube and then sheared DNA using a Covaris Sonicator for 4min at 150 bp according to manufacturer specifications.

We retrieved dental calculus from two skeletons (individuals 7057 and 7433; less than ten mg per sample) and dentin from one skeleton (individual 7057; less than 50 mg per sample) at the University of Minnesota using a sterile dental scaler. We decontaminated calculus using exposure to UV irradiation for five min. This was followed by an initial 0.5M EDTA (Ambion) wash in a 2.0 mL tube for 15 min. We subjected samples to a two day 0.5 EDTA and proteinase K (10 mg/mL; Qiagen) digestion, at which point we combined the resulting solution with 12 mL of PB buffer and followed standard MinElute PCR Purification Kit (Qiagen) protocol. Our dentin protocol followed previously published methods (Nieves-Colón et al., 2018). We did not shear dental calculus and dentin samples prior to shotgun library builds.

We collected urine from seven wild Gombe chimpanzees—three with matched fecal samples (7150, 7365, and 7507) and four others (7072, 7323, 7535, and 7650)—in the early morning using fresh plastic bags attached to sticks suspended below chimpanzee nests. We immediately transferred between 10 mL and 30 mL of urine to a 50 mL tube and centrifuged the material for ten min at 3k rpm. We removed supernatant and covered the resulting pellet with 5 mL of RNAlater for storage in the field. In the lab, we extracted samples using the Urine DNA Isolation Kit (Abcam) according to manufacturer protocols. We sheared the resulting elution using the Covaris sonicator as previously described and desiccated the resulting solution down to 20 µL.

Shotgun Build and Amplification

We built shotgun libraries using the resulting elutions from feces, urine, dentin, and calculus extractions. For initial blunt end repair, we added a total of 20 µL (~800 ng) of DNA to 5.0 µL NEB Buffer, 0.50 µL dNTP mix (2.5mM), 4.0 µL BSA (10 mg/mL), 5.0 µL ATP (10mM), 2.0 µL T4 PNK, 0.40 µL T4 Polymerase, and 13.10 µL ddH_2O. We incubated this solution at 15°C for 15 min followed by 25°C for 15 min. We then cleaned the solution using PCR MinElute Purification Kit according to manufacturer protocol before eluting into 18 µL EB buffer. For adapter ligation, we added 18 µL of template DNA to 20 µL Quick Ligase Buffer, 1.0 µL Solexa Mix (Meyer & Kircher, 2010), and 1.0 µL Quick Ligase and incubated the solution at room temperature for 20 min. We then cleaned again using PCR MinElute Purification (Qiagen) according to manufacturer protocol and eluted the solution into 20 µL EB buffer. For the final fill in portion of the shotgun build, we added 20 µL of template DNA to 4.0 µL Thermo pol buffer, 0.50 µL dNTP mix (2.5mM), 2.0 µL Bst polymerase, and 13.50 µL ddH_2O. We incubated the solution at 37°C for 20 min followed by 80°C for 20 min. We amplified shotgun libraries using Amplitaq Gold before splitting libraries into four identical PCR reactions which contained 9.0 µL of DNA, 9.27 µL PCR Buffer II (10x), 9.27 µL $MgCl_2$ (25mM), 3.68 µL dNTP mix (10nM), 2.21 µL BSA (10 mg/mL), 2.0 µL P5 primer, 2.0 µL P7 primer, 61.09 µL of ddH_2O, and 1.48 µL of Amplitaq Gold enzyme. We used the following PCR conditions: initial denaturation at 95°C for 15 min, followed by cycling of 95°C for 30 sec, 58°C for 30 sec, and 72°C for 45 sec, with a final elongation of 72°C for ten min. We amplified each sample between 8 and 13 cycles (Table S1) using Illumina adapter primers with unique forward and reverse barcodes. We then purified samples using the Minelute PCR Purification Kit according to manufacturer protocol before eluting into 30 µL of EB buffer. We used a total of 7 µL of amplified calculus, dentin, and fecal DNA for each of the capture sets. For urine, we desiccated amplified material from 30 µL down to 7 µL before undergoing a single exome capture.

Whole-Genome and Exome Capture Kits

We used two whole-genome kits (chimpanzee and human baits) and one human exome kit to capture host DNA from the variety of samples. For the whole-genome chimpanzee kit, Arbor Biosciences produced a custom whole-genome capture MYBaits kit using *Pan troglodytes schweinfurthii* DNA. Genomic DNA extracted from the blood of a chimpanzee (Stone et al., 2010) was used as source material for baits. We pooled extractions for a total of 5 ug of DNA which Arbor Biosciences then used to produce the whole-genome capture baits. For the human whole-genome capture baits, we used a MYcroarray whole-genome human capture kit (using African/Masai male DNA). Finally, we also used the IDT xGen Exome Research Panel (v1.0), a commercially available human exome capture kit.

For feces, we used an input total of 7 uL of amplified material regardless of concentration for each of the three capture kits: the *P. t. schweinfurthii* MYBaits capture, the human MYBaits capture, and the IDT xGen Exome Research Panel. For the chimpanzee whole-genome capture MYBaits kit, we captured each sample according to MYbaits Kapa HiFi HotStart ReadyMix protocol with a hybridization time of 24 hours and a final post-capture PCR amplification of 14 cycles. We purified all samples post-capture through removal of beads, cleanup using the MinElute PCR Purification Kit, and elution into 30 µL. We re-amplified a second time using identical PCR conditions, with the number of cycles dependent upon the outcome of quantification from a Bioanalyzer DNA 1000 chip (Agilent). We purified all samples post-capture using the MinElute PCR Purification Kit according to manufacturer specifications and eluted into 30 µL.

For the MYbaits human whole-genome kit, we captured each of the four amplified fecal samples in the same manner, using the same amount of starting amplified material. However, during the final phase of the MYBaits protocol, all samples were amplified 14 cycles instead of the usual 12 cycles. As such, no samples were re-amplified post-capture after we confirmed high concentrations using a Bioanalyzer DNA 1000 chip.

For the xGen Exome Research Panel, from IDT, the unique P5 and P7 7 nt barcodes used to identify the amplified samples necessitated custom xGen Universal Blocking oligos from IDT. We used a total of 7 ul of amplified material from each sample (greater than the suggested 500 ng input of DNA) for the capture in accordance with manufacturer protocol. The exception to this was for urine, which we desiccated from a starting volume of 30 µL, due to the initial low concentrations. We amplified each capture pool to 12 cycles using KAPA HiFi Hotstart ReadyMix, purified each using Agencourt AMPure beads, and eluted into 22 µL of EB Buffer (Qiagen) as suggested by the protocol. Lastly, we quantified the samples using a Bioanalyzer High Sensitivity DNA chip and amplified each for six more cycles.

Samples were then pooled (see Table 1 for breakdown) before being sent for sequencing at the Yale Center for Genome Analysis. Samples were sequenced on four different Illumina

HiSeq2500 Rapid runs (2x100 paired-end) and an Illumina HiSeq2500 standard run (2x150 paired-end).

Read Processing, Read Mapping, Variant Calling, and Depth of Coverage

Before mapping reads, we examined read quality using FastQC (v0.11.7; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC (v1.5.dev0; (Ewels, Magnusson, Lundin, & Käller, 2016)), and trimmed adapters and low-quality sequence from reads using BBDuk (v37.90; https://jgi.doe.gov/data-and-tools/bbtools/) with the following parameters: “ktrim=r k=21 mink=11 hdist=2 tbo tpe qtrim=rl trimq=10 minlength=30”. Using default parameters, we mapped reads to the chimpanzee reference genome (panTro4; (Waterson, Lander, Wilson, The Chimpanzee, & Analysis, 2005)) with BWA-MEM (v0.7.17-r1188; (Heng Li, 2013). We then used SAMtools to fix mate pairings, and sort and index BAM files (v1.7; (Heng Li & Durbin, 2009). Because we sequenced some of the samples across multiple lanes (Table S1**)**, we used Sambamba to merge BAM files from these samples using default parameters (v0.6.6; (Tarasov, Vilella, Cuppen, Nijman, & Prins, 2015). Note that we only merged BAM files within individual, biological material, and sequencing library (i.e., samples from the same individual but different source material or capture method were left unmerged and treated separately, as these were different units in our analyses). Finally, we marked duplicates using Picard (v2.18.10; http://broadinstitute.github.io/picard).

We next called variants on each processed BAM file separately using Genome Analysis Toolkit’s (GATK’s) HaplotypeCaller with default parameters (v4.0.8.1; (Van der Auwera et al., 2013)). We then filtered each VCF using BCFTools (v1.6; (H. Li et al., 2009)) . We included sites for which mapping quality >= 20, site quality (QUAL) >=30, and genotype quality >= 30.

Because some of the downstream coverage analyses are affected by differing number of raw reads across samples, we downsampled merged BAM files (without duplicate marking) to 40 million reads. To do so, we used SAMtools view (v1.7; (H. Li et al., 2009)) with the flag "-s *downsample_fraction*", where *downsample_fraction* is equal to 40 million divided by the sample's total number of raw reads. Note that for analyses requiring downsampling, we only included samples with 40 million or more reads. We next marked duplicates, as above, using Picard (v2.18.10; http://broadinstitute.github.io/picard). We used downsampled BAM files for coverage analyses, but not endogenous content estimates or variant calling.

To calculate depth of coverage from BAM files, we first used SAMTOOLS view (v1.7; (H. Li et al., 2009) with the flags '-F 1024 -q 20' to remove duplicates and only retain reads with the minimum mapping quality of 20. We then used Bedtools GenomeCov (v2.27.1; (Quinlan & Hall, 2010)) with the flag -bg to output a bedgraph file with coverage statistics. Next, again using Bedtools, we intersected bedgraph files with Ensembl coding sequences (CDS) for the panTro4 genome downloaded from the UCSC Table Browser (Karolchik et al., 2004). Finally, using a custom python script, "Compute_histogram_from_bed.py" (see Data Accessibility), we calculated histograms of CDS depth.

Analysis

We used the SAMtools stats tool to calculate basic metrics related to fraction of reads mapping, duplicates, etc. (v1.7; (H. Li et al., 2009)) across all sample types. We calculated these metrics both with and without duplicates, and for primary and downsampled BAM files. To remove duplicates, we first used SAMtools view with the '-F 1024' flag, before piping output to SAMtools stats. From these metrics, we estimated post-capture endogenous content as the fraction of reads mapping to the reference genome. To test for statistical differences in post-capture endogenous content among sample sources we used an ANCOVA test in R (R Development Core Team, 2014).

Within R, we generated "reverse cumulative" plots (Reed, Meade, & Steinhoff, 1995) of coverage across CDS for feces vs. urine and exome vs. whole-genome for feces ("plot_coverage.R"; see Data Accessibility). These plots display the proportion of total panTro4 CDS (Y-axis) covered by X or more reads (where X is a value on the X-axis).

Using exome data, we examined genotype concordance between paired urine and fecal samples for three individuals (7150, 7507, 7365), and paired calculus and dentin samples for one individual (7057). To estimate concordance, we ensured that variant calls were made at identical sites in the paired samples. We did this by first using BCFtools merge (v1.6; (H. Li et al., 2009)) with the flag "-m all" to merge the paired (urine and feces, or calculus and dentin) exome VCF files for each individual. We then conducted a second round of variant calling using GATK's HaplotypeCaller (v4.0.8.1; (Van der Auwera et al., 2013)) as described above, with the addition of the flag "-ERC BP_RESOLUTION" and the merged VCF as an interval file via the "-L" flag. These flags force HaplotypeCaller to call genotypes at the same sites—any site called in either the urine or fecal sample (or calculus or dentin sample) from a given individual—in both samples. We then, for each site, compiled genotype, depth, mapping quality, and genotype quality measures from the newly generated VCFs using the custom Python script "Compare_vcfs.py". From this compiled dataframe, we removed "random" (containing "_random") and unplaced (containing "chrUn") scaffolds. We then used the Python script "Process_dropout.py" (see Data Accessibility) to estimate genotype concordance for paired samples at four different minimum depths (4x, 6x, 8x, and 10x). The script finds all sites passing minimum quality thresholds (minimum depth >= value described previously, mapping quality >= 30, and genotype quality >= 30) in both samples, and from those sites counts the number of sites with shared genotypes, genotypes consistent with allelic dropout, and ambiguous genotypes. We considered genotypes consistent with dropout if one of the two samples was heterozygous, while the other was homozygous for one of the alleles in the first sample's genotype (e.g., "0/1" in urine and "1/1" in feces would be counted as dropout in feces). Genotypes were classified as ambiguous if they were not shared and did not fit a pattern consistent with dropout; for example, if the urine sample had a genotype of "1/1" while the fecal sample had a genotype of "0/2".

Data Accessibility

We deposited raw reads in NCBI's Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under Bioproject PRJNA508503. We implemented the full assembly and analysis pipeline in Snakemake (Köster & Rahmann, 2012), and managed software using Bioconda (Grüning et al., 2018). All code, scripts, and software environments are available on Github (https://github.com/thw17/Gombe_noninvasive_genomic_methods).

Results

We processed a total of 14 samples from ten different chimpanzees in Gombe National Park, Tanzania from urine (n=7), feces (n=4), dental calculus (n=2), and dentin (n=1) (Table S1). We then captured and sequenced samples using at least one of the following: undirected shotgun amplification (n=2), MYBaits *Pan troglodytes schweinfurthii* capture (Arbor Biosciences; n=4), MYBaits *Homo sapien*s capture (Arbor Biosciences; n=6), xGen (human) Exome Research Panel

(IDT; n=26) (Table S1). In total, we analyzed 38 different combinations of individual, source, and sequencing protocol (Table S1).

Concentrations of extracted DNA varied widely across samples (Table S1). Initially, concentrations ranged from 0.11 ng/µL to 65.6 ng/µL with a single urine sample from 7365 too low to be measured. Sequencing success was similarly variable (Table S2). After merging BAM files from the same sample across multiple runs, we generated between 6.9 and 169.5 million reads per sample and while we successfully produced data for the problematic urine sample (individual 7365), it produced the fewest reads (Table S2). We observed high duplication rates likely resulting from PCR amplification during library construction and capture in most, but not all samples (range from 0.05% to 89.4%; Table S2). In general, exome capture had higher duplication rates than whole-genome capture, which, in turn, had higher duplication rates than shotgun sequencing (Table S2; Figure S1). We also observed a linear increase in duplication rate with an increasing number of mapped reads for whole-genome capture, but not exome capture or shotgun sequencing (Figure S1). After filtration and duplicate removal, we were left with between 1.4 and 26.2 million passing reads per sample (Table S2).

Interestingly, we found that samples ranged in the amount of post-capture endogenous DNA (i.e., DNA from the host after sequence capture, as opposed to other sources) from 33.9% to 99.1% (Figure 1; Table S2). We discovered that this effect was driven by the source of the sample (Figure 1; ANCOVA: F(3,18) = 125.493; $p < 0.001$). Upon further investigation, a post hoc Tukey test revealed that urine (n =7; mean endogenous percentage = 96.4%) had significantly more endogenous DNA than dentin ($p = 0.03$; n=1; mean=75.9%), feces ($p < 0.001$; n= 12; mean= 44.9%), and calculus ($p < 0.001$; n=6 ; mean=38.3%).

We evaluated capture success using reverse cumulative plots to assess the proportion of CDS in PanTro4 (i.e., the fraction of sequence in PanTro4 annotated as coding sequence) sequenced at different depths. For all samples, we started with a fixed 40 million reads before duplicate removal. We first used fecal samples to compare exome capture with whole-genome capture, and found that exome capture, despite its higher duplication rate, led to broader and deeper coverage across CDS than whole-genome capture (Figure 2). In addition, when comparing urine and fecal samples (exome capture), urine outperformed feces (Figure 3). Across all urine samples, more than 90% of CDS was captured, while only two fecal samples generated data covering more than 50% of CDS (Figure 3). This pattern became even more pronounced as depth increased; for example, at a minimum depth of 8x, more than 75% of CDS was captured in all urine samples, while all fecal samples fell below 10% CDS covered (Figure 3). Finally, when comparing calculus and dentin, we found more than 85% of CDS was captured for the single dentin sample, with 20% of CDS captured at a minimum depth of 8x (Figure 4). However, less than 25% of CDS was captured in both analyzed calculus samples, which decreased to less than 1% at a minimum depth of 8x (Figure 4).

We measured genotype concordance in the three individuals for which we sequenced at least 40 million reads each for paired fecal and urine samples (Table 1; 7150, 7365, 7507) and a single additional individual for paired dentin and calculus (7057). Likely due to the differences in endogenous DNA content and coverage described above, we obtained very few variant sites (i.e., sites with one or both alleles differing from the reference genome) passing quality and depth filters in feces compared to urine (Table 1). For example, at a minimum depth of 10x, we obtained 227, 368, and 2014 sites from the fecal samples from the three individuals, while we obtained 4952, 93,244, and 115,955 sites from the same individuals from urine samples. In total, we were able to compare between 59 and 1,309 sites depending on the individual and depth threshold used (Table 1). Overall, genotypes were overwhelmingly concordant, with less than 11% of sites discordant across all comparisons. Most discordant sites were consistent with a pattern of allelic dropout—that is, one sample was heterozygous, while the other was homozygous for one of the two alleles present in the first sample. Among these dropout sites, at a minimum depth of 10x, feces exhibited higher rates of dropout than urine in two of our three comparisons (fecal dropout = 2-8% of all sites; urine dropout = 0.8-4% of all sites). We also observed "ambiguous" sites—discordant sites inconsistent with the dropout pattern described above—at 1-3% of all sites (Table 1). For calculus and dentin, we compared between 27 and 291 shared sites and observed calculus as having the highest dropout rates of any source of DNA at depth thresholds of 8x and 10x (17.86% and 18.42%, respectively). Although we observed less dropout in dentin, these rates are comparable to our highest observed dropout rates for feces (7.89% dropout at a depth of 10x in dentin, 7.86% dropout in feces at a depth of 10x for individual 7507).

Discussion

The development of noninvasive genomic methods is critically important for studying wild populations, particularly those that cannot otherwise be legally or ethically sampled. In this study, we evaluated four biological sources of DNA that can be sampled from wild populations of many taxa: feces, urine, dentin, and dental calculus. Feces and urine may be noninvasively sampled from contemporary living populations, while dentin and dental calculus can often be sampled from skeletal collections of wild populations present in collections at museums and field sites. We assessed the quality of these sources in three different ways. First, we determined post-capture endogenous content, the amount of captured DNA is derived from the host. Next, we evaluated the breadth and depth of sequencing coverage across genomic targets. Finally, we measured the concordance of genotypes between pairs of samples captures from different sources from the same individual.

In regard to post-capture endogenous content, of the four sources, we found that urine samples contained the highest proportion of host DNA. While post-capture endogenous content was similar in calculus and feces (ranging from approximately 30-50%), all urine samples contained more than 95% host DNA. Previous studies have demonstrated both that host DNA is present in urine and can be successfully extracted and amplified (Hausknecht et al., 2007; Hayakawa & Takenaka, 1999; Hedmark et al., 2004; Nota & Takenaka, 1999; Valiere & Taberlet, 2000; Waits & Paetkau, 2005); however, our results show for the first time that urine in fact has an high fraction of host DNA compared to other sources of DNA, like feces, that are far more commonly used in genetic studies of wild animals, and thus is well-suited for genomic analysis. While we measured post-capture endogenous content in the same way across the sources of DNA that we tested, we are unable to determine for certain from this study whether the difference in endogenous content directly reflect raw differences in the fraction of host DNA across sources. We cannot easily envision a process that would cause sources of DNA to differ in endogenous content after capture but not before, but future work could aim to measure pre-capture differences to confirm our results.

We found that these differences in endogenous content meaningfully impact downstream sequencing success, as exome capture and sequencing of urine samples led both broader and deeper coverage across the coding sequence of the chimpanzee reference genome than any of the other sources of DNA. With the exception of a single problematic sample, all of the urine samples captured more than 90% of coding sequence at a depth of 4x or greater (after duplicate removal), despite extremely high duplication rates. This means that, without optimization or any other methodological considerations, our urine samples produced sufficient data for most evolutionary and population genetic analyses. In contrast, not a single fecal, calculus, or dentin sample in our study produced enough data for downstream analyses (Figures 3 and 4). Rather than suggest that any of these sources of DNA are more or less useful for genomic analyses, we instead argue that these results indicate that urine might work well “out of the box” similar to other high-quality sources like blood and other tissues, while the other sources that we tested require additional methodological considerations for use, like the many developments for feces (Chiou & Bergey, 2018; Perry et al., 2010; Snyder-Mackler et al., 2016).

Our analyses revealed that genotypes generated from feces and urine from the same individual were broadly concordant, especially when a minimum depth threshold of 10x was used. Urine fared better generally, with fewer sites ambiguously discordant or consistent with dropout. However, we only had paired fecal and urine samples for three individuals, so these results must be taken as preliminary. Regardless of whether genotypes from urine are comparable or better than those of feces, the low rates of allelic dropout underscore the quality of urine as a source of DNA for genomic analyses. In addition, while we are unable to test it at this time, we hypothesize that urine might be less susceptible to problematic contamination than feces. As discussed above, estimates of the proportion of exogenous DNA in urine before capture are unknown; however, it is well known that feces contain overwhelmingly exogenous DNA (Chiou & Bergey, 2018; Perry et al., 2010; Qin et al., 2010). In addition to the microbiota that dominate feces, fecal samples also contain dietary DNA from food items consumed by the host (Bradley et al., 2007; Clayton et al., 2016). In the case of chimpanzees, food items include a wide array of plant and animal items, including nonhuman primate prey (Gilby, 2006; Hobaiter, Samuni, Mullins, Akankwasa, & Zuberbühler, 2017; Mitani, Watts, & Muller, 2002; Pruetz et al.; Uehara, 1997). Because of the extremely high proportion of microbiota in feces, some sort of DNA capture is required to target endogenous DNA (Chiou & Bergey, 2018; Hernandez-Rodriguez et al., 2018; Perry et al., 2010; Snyder-Mackler et al., 2016; White et al., 2019). However, baits can successfully capture sequence across more than 65 million years of divergence (i.e., across the entire primate order) and much of this captured sequence will map to a reference genome equivalently divergent (Webster et al., 2018). This means that the same baits designed to capture host DNA in the feces will also likely successfully capture DNA from primate prey species and that these contaminant sequences will successfully map to the host reference genome, introducing artifacts into genotyping. This possibility needs to be studied further, but if present in our samples, it would artificially increase our observed rates of allelic dropout for urine (calculated as heterozygous sites in feces that are homozygous for one of the alleles in urine). We thus consider our estimates of allelic dropout in urine to be conservative overall.

Our analyses of genotype concordance in dentin and calculus were limited, as we only had a single individual with data from both sources and we recovered very little usable data in the calculus sample. However, in that comparison, we observed a rate of allelic dropout in calculus more than double that of any other tissue. Our estimates for dentin were similar to feces at our most rigorous depth threshold. These results are consistent with previous research showing that yields and quality of host genetic material are lower in calculus compared to dentin (Mann et al., 2018). Yet, calculus has been used to recover full mitochondrial and nuclear genomes from human calculus samples (Ozga et al., 2016; Ziesemer et al., 2018). We therefore suggest that more work is needed to explore and optimize DNA capture from calculus in wild populations.

Taken together, we suggest using urine as a primary source of noninvasive genomic DNA. However, urine is not universally available in sufficient quantities for collection and extraction, and its availability and collectable volume will vary by organism body size, study habitat, and level of habituation. When using other noninvasive biological materials, our results build on previous research (Chiou & Bergey, 2018; Hernandez-Rodriguez et al., 2018; White et al., 2019) showing that targeting a smaller subset of the genome leads to an increase in usable data. In particular, we argue that exome capture is an ideal option, as it targets a small subset of the genome commonly used in evolutionary analyses and there are commercially available human kits that can be used across the entire primate order (Webster et al., 2018). However, like other methods of DNA capture, exome capture requires additional considerations when working with noninvansive samples. First, a multitude of factors impact the quality of host genomic material in a natural environment, including time elapsed since excretion (DeMay et al., 2013), field/laboratory storage conditions (Nsubuga et al., 2004; Panek et al., 2018), and enzymatic activity (Deagle et al., 2006). Second, depending on sample quality, it may be necessary to undergo repeated extractions for the same sample, along with multiple double stranded DNA library builds and multiple indexing amplifications. Third, a single capture of the indexed DNA library may lead to a higher duplication rate, which has been cited in several studies as being a barrier to inexpensive and accurate host genome capture (Bansal & Pinney, 2017; Ebbert et al., 2016; García-García et al., 2016).

Noninvasive samples have been used across a variety of disciplines for addressing many evolutionary and ecological questions (Beja-Pereira, Oliveira, Alves, Schwartz, & Luikart, 2009) including investigations into dietary niches, social structures, and diversity of endangered animals (Carroll et al., 2018). Chimpanzees, currently listed as endangered on the IUCN red list, are considered to be flagship species and indicators of environmental stressors in the surrounding area (Wrangham, 2008). Thus, noninvasive genomic methods are critical for monitoring the health of wild populations as well as aspects of local adaptation and population history important for conservation management. This is especially important for small, isolated populations such as that of Gombe National Park, for which there is an effort to maintain genetic diversity (Pusey, Pintea, Wilson, Kamenya, & Goodall, 2007). The results of our study highlight urine as a promising and untapped source of DNA for this and other genomic work in not only chimpanzees, but wild populations of other protected species as well.

Acknowledgements

We gratefully acknowledge Nasibu Zuberi and Dismas Mwacha for their assistance in sample collection; the Gombe Stream Research Center and the Jane Goodall Institute; TAWIRI, COSTECH, and Tanzania National Parks for permission to conduct research; the Leakey Foundation, Arizona State University Institute for Human Origins, and the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH; grant R35GM124827 to MAW) for generously providing funding; and Research Computing at Arizona State University for providing computing resources.

Conflict of Interest Statement

The authors declare no competing interests.

References

Bansal, A., & Pinney, S. E. (2017). DNA methylation and its role in the pathogenesis of diabetes. *Pediatric diabetes, 18*(3), 167-177. doi:10.1111/pedi.12521

Bataillon, T., Duan, J., Hvilsom, C., Jin, X., Li, Y., Skov, L., . . . Schierup, M. H. (2015). Inference of Purifying and Positive Selection in Three Subspecies of Chimpanzees (Pan troglodytes) from Exome Sequencing. *Genome Biology and Evolution, 7*(4), 1122-1132. doi:10.1093/gbe/evv058

Beja-Pereira, A., Oliveira, R., Alves, P. C., Schwartz, M. K., & Luikart, G. (2009). Advancing ecological understandings through technological transformations in noninvasive genetics. *Molecular Ecology Resources, 9*(5), 1279-1301. doi:10.1111/j.1755-0998.2009.02699.x

Bradley, B. J., Stiller, M., Doran-Sheehy, D. M., Harris, T., Chapman, C. A., Vigilant, L., & Poinar, H. (2007). Plant DNA sequences from feces: potential means for assessing diets of wild primates. *American Journal of Primatology, 69*(6), 699-705. doi:10.1002/ajp.20384

Carroll, E. L., Bruford, M. W., DeWoody, J. A., Leroy, G., Strand, A., Waits, L., & Wang, J. (2018). Genetic and genomic monitoring with minimally invasive sampling methods. *Evolutionary Applications, 11*(7), 1094-1119. doi:10.1111/eva.12600

Chiou, K. L., & Bergey, C. M. (2018). Methylation-based enrichment facilitates low-cost, noninvasive genomic scale sequencing of populations from feces. *Scientific Reports, 8*(1), 1975. doi:10.1038/s41598-018-20427-9

Clayton, J. B., Vangay, P., Huang, H., Ward, T., Hillmann, B. M., Al-Ghalith, G. A., . . . Knights, D. (2016). Captivity humanizes the primate microbiome. *Proceedings of the National Academy of Sciences, 113*(37), 10376-10381. doi:10.1073/pnas.1521835113

Constable, J. L., Ashley, M. V., Goodall, J., & Pusey, A. E. (2001). Noninvasive paternity assignment in Gombe chimpanzees. *Molecular ecology, 10*(5), 1279-1300. doi:10.1046/j.1365-294X.2001.01262.x

Deagle, B. E., Eveson, J. P., & Jarman, S. N. (2006). Quantification of damage in DNA recovered from highly degraded samples--a case study on DNA in faeces. *Frontiers in zoology, 3*, 11-11. doi:10.1186/1742-9994-3-11

DeMay, S. M., Becker, P. A., Eidson, C. A., Rachlow, J. L., Johnson, T. R., & Waits, L. P. (2013). Evaluating DNA degradation rates in faecal pellets of the endangered pygmy rabbit. *Molecular Ecology Resources, 13*(4), 654-662. doi:10.1111/1755-0998.12104

Ebbert, M. T. W., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J., . . . Ridge, P. G. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics, 17 Suppl 7*(Suppl 7), 239-239. doi:10.1186/s12859-016-1097-3

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends Ecol Evol, 29*(1), 51-63. doi:10.1016/j.tree.2013.09.008

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics, 32*(19), 3047-3048. doi:10.1093/bioinformatics/btw354

García-García, G., Baux, D., Faugère, V., Moclyn, M., Koenig, M., Claustres, M., & Roux, A.-F. (2016). Assessment of the latest NGS enrichment capture methods in clinical context. *Scientific Reports, 6*(1), 20948. doi:10.1038/srep20948

George, R. D., McVicker, G., Diederich, R., Ng, S. B., MacKenzie, A. P., Swanson, W. J., . . . Thomas, J. H. (2011). Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. *Genome Research, 21*(10), 1686-1694. doi:10.1101/gr.121327.111

Gilby, I. C. (2006). Meat sharing among the Gombe chimpanzees: harassment and reciprocal exchange. *Animal Behaviour, 71*(4), 953-963.

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., . . . The Bioconda, T. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods, 15*(7), 475-476. doi:10.1038/s41592-018-0046-7

Han, L., Su, B., Li, W.-H., & Zhao, Z. (2008). CpG island density and its correlations with genomic features in mammalian genomes. *Genome Biol, 9*(5), R79. doi:10.1186/gb-2008-9-5-r79

Hausknecht, R., Gula, R., Pirga, B., & Kuehn, R. (2007). Urine — a source for noninvasive genetic monitoring in wildlife. *Molecular Ecology Notes, 7*(2), 208-212. doi:10.1111/j.1471-8286.2006.01622.x

Hayakawa, S., & Takenaka, O. (1999). Urine as another potential source for template DNA in polymerase chain reaction (PCR). *American Journal of Primatology, 48*(4), 299-304. doi:10.1002/(sici)1098-2345(1999)48:4<299::Aid-ajp5>3.0.Co;2-g

Hedmark, E., Flagstad, Ø., Segerström, P., Persson, J., Landa, A., & Ellegren, H. (2004). DNA-Based Individual and Sex Identification from Wolverine (Gulo Gulo) Faeces and Urine. *Conservation Genetics, 5*(3), 405-410. doi:10.1023/B:COGE.0000031224.88778.f5

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

Hobaiter, C., Samuni, L., Mullins, C., Akankwasa, W. J., & Zuberbühler, K. (2017). Variation in hunting behaviour in neighbouring chimpanzee communities in the Budongo forest, Uganda. *PLOS ONE, 12*(6), e0178065. doi:10.1371/journal.pone.0178065

Hoss, M., Kohn, M., Paabo, S., Knauer, F., & Schroder, W. (1992). Excrement analysis by PCR. *Nature, 359*(6392), 199-199.

Hvilsom, C., Qian, Y., Bataillon, T., Li, Y., Mailund, T., Sallé, B., . . . Schierup, M. H. (2012). Extensive X-linked adaptive evolution in central chimpanzees. *Proceedings of the National Academy of Sciences, 109*(6), 2054-2059. doi:10.1073/pnas.1106877109

Inoue, E., Inoue-Murayama, M., Takenaka, O., & Nishida, T. (2007). Wild chimpanzee infant urine and saliva sampled noninvasively usable for DNA analyses. *Primates, 48*(2), 156-159. doi:10.1007/s10329-006-0017-y

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

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research, 32*(Database issue), D493-D496. doi:10.1093/nar/gkh103

Kohn, M. H., & Wayne, R. K. (1997). Facts from feces revisited. *Trends Ecol Evol, 12*(6), 223-227.

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics, 28*(19), 2520-2522. doi:10.1093/bioinformatics/bts480

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303.3997.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics, 25*(14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., & Homer, N. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.

Mann, A. E., Sabin, S., Ziesemer, K., Vågene, Å. J., Schroeder, H., Ozga, A. T., . . . Warinner, C. (2018). Differential preservation of endogenous human and microbial DNA in dental calculus and dentin. *Scientific Reports, 8*, 9822.

Meyer, M., & Kircher, M. (2010). Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. *Cold Spring Harbor Protocols, 2010*(6), pdb.prot5448.

Mitani, J. C., Watts, D. P., & Muller, M. N. (2002). Recent developments in the study of wild chimpanzee behavior. *Evolutionary Anthropology: Issues, News, and Reviews, 11*(1), 9-25. doi:10.1002/evan.10008

Nieves-Colón, M. A., Ozga, A. T., Pestle, W. J., Cucina, A., Tiesler, V., Stanton, T. W., & Stone, A. C. (2018). Comparison of two ancient DNA extraction protocols for skeletal remains from tropical environments. *American Journal of Physical Anthropology, 166*(4), 824-836. doi:doi:10.1002/ajpa.23472

Nota, Y., & Takenaka, O. (1999). DNA extraction from urine and sex identification of birds. *Molecular ecology, 8*(7), 1237-1238. doi:10.1046/j.1365-294X.1999.00682_2.x

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

Ozga, A. T., Nieves-Colón, M. A., Honap, T. P., Sankaranarayanan, K., Hofman, C. A., Milner, G. R., . . . Warinner, C. (2016). Successful enrichment and recovery of whole mitochondrial genomes from ancient human dental calculus. *American Journal of Physical Anthropology, 160*(2), 220-228.

Ozga, A. T., Webster, T. H., Gilby, I. C., Wilson, M. A., Nockerts, R. S., Wilson, M. L., . . . Stone, A. C. (2020). *Pan troglodytes schweinfurthii raw sequence data*.

Panek, M., Čipčić Paljetak, H., Barešić, A., Perić, M., Matijašić, M., Lojkić, I., . . . Verbanac, D. (2018). Methodology challenges in studying human gut microbiota – effects of collection, storage, DNA extraction and next generation sequencing technologies. *Scientific Reports, 8*(1), 5143. doi:10.1038/s41598-018-23296-4

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

Pruetz, J. D., Bertolani, P., Ontl, K. B., Lindshield, S., Shelley, M., & Wessling, E. G. New evidence on the tool-assisted hunting exhibited by chimpanzees (Pan troglodytes verus) in a savannah habitat at Fongoli, Sénégal. *Royal Society Open Science, 2*(4), 140507. doi:10.1098/rsos.140507

Pusey, A. E., Pintea, L., Wilson, M. L., Kamenya, S., & Goodall, J. (2007). The Contribution of Long-Term Research at Gombe National Park to Chimpanzee Conservation. *Conservation Biology, 21*(3), 623-634. doi:10.1111/j.1523-1739.2007.00704.x

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., . . . Renault, P. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature, 464*. doi:10.1038/nature08821

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

R Development Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundationg for Statistical Computing.

Reed, G. F., Meade, B. D., & Steinhoff, M. C. (1995). The Reverse Cumulative Distribution Plot: A Graphic Method for Exploratory Analysis of Antibody Data. *Pediatrics, 96*(3), 600-603.

Sastre, N., Francino, O., Lampreave, G., Bologov, V. V., López-Martín, J. M., Sánchez, A., & Ramírez, O. (2009). Sex identification of wolf (Canis lupus) using non-invasive samples. *Conservation Genetics, 10*(3), 555-558. doi:10.1007/s10592-008-9565-6

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

Stone, A. C., Battistuzzi, F. U., Kubatko, L. S., Perry, G. H., Trudeau, E., Lin, H., & Kumar, S. (2010). More reliable estimates of divergence times in Pan using complete mtDNA sequences and accounting for population structure. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1556), 3277-3288. doi:10.1098/rstb.2010.0096

Taberlet, P., Waits, L. P., & Luikart, G. (1999). Noninvasive genetic sampling: look before you leap. *Trends Ecol Evol, 14*(8), 323-327. doi:10.1016/S0169-5347(99)01637-7

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics (Oxford, England), 31*(12), 2032-2034. doi:10.1093/bioinformatics/btv098

Thomas-White, K., Brady, M., Wolfe, A. J., & Mueller, E. R. (2016). The bladder is not sterile: History and current discoveries on the urinary microbiome. *Current bladder dysfunction reports, 11*(1), 18-24. doi:10.1007/s11884-016-0345-8

Uehara, S. (1997). Predation on mammals by the chimpanzee (Pan troglodytes). *Primates*(38), 193.

Valiere, N., & Taberlet, P. (2000). Urine collected in the field as a source of DNA for species and individual identification. *Molecular ecology, 9*(12), 2150-2152. doi:10.1046/j.1365-294X.2000.11142.x

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

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., . . . DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics, 43*(1110), 11.10.11-11.10.33. doi:10.1002/0471250953.bi1110s43

Waits, L. P., & Paetkau, D. (2005). Noninvasive Genetic Sampling Tools for Wildlife Biologists: A Review of Applications and Recommendations for Accurate Data Collection. *The Journal of Wildlife Management, 69*(4), 1419-1433. doi:10.2193/0022-541X(2005)69[1419:NGSTFW]2.0.CO;2

Waterson, R. H., Lander, E. S., Wilson, R. K., The Chimpanzee, S., & Analysis, C. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature, 437*(7055), 69-87. doi:10.1038/nature04072

Webster, T. H., Guevara, E. E., Lawler, R. R., & Bradley, B. J. (2018). Successful exome capture and sequencing in lemurs using human baits. *bioRxiv*, 490839. doi:10.1101/490839

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

Wrangham, R. W. (2008). Why the link between long-term research and conservation is a case worth making. In R. W. Wrangham & E. M. Ross (Eds.), *Science and Conservation in African Forests*. Cambridge: Cambridge University Press.

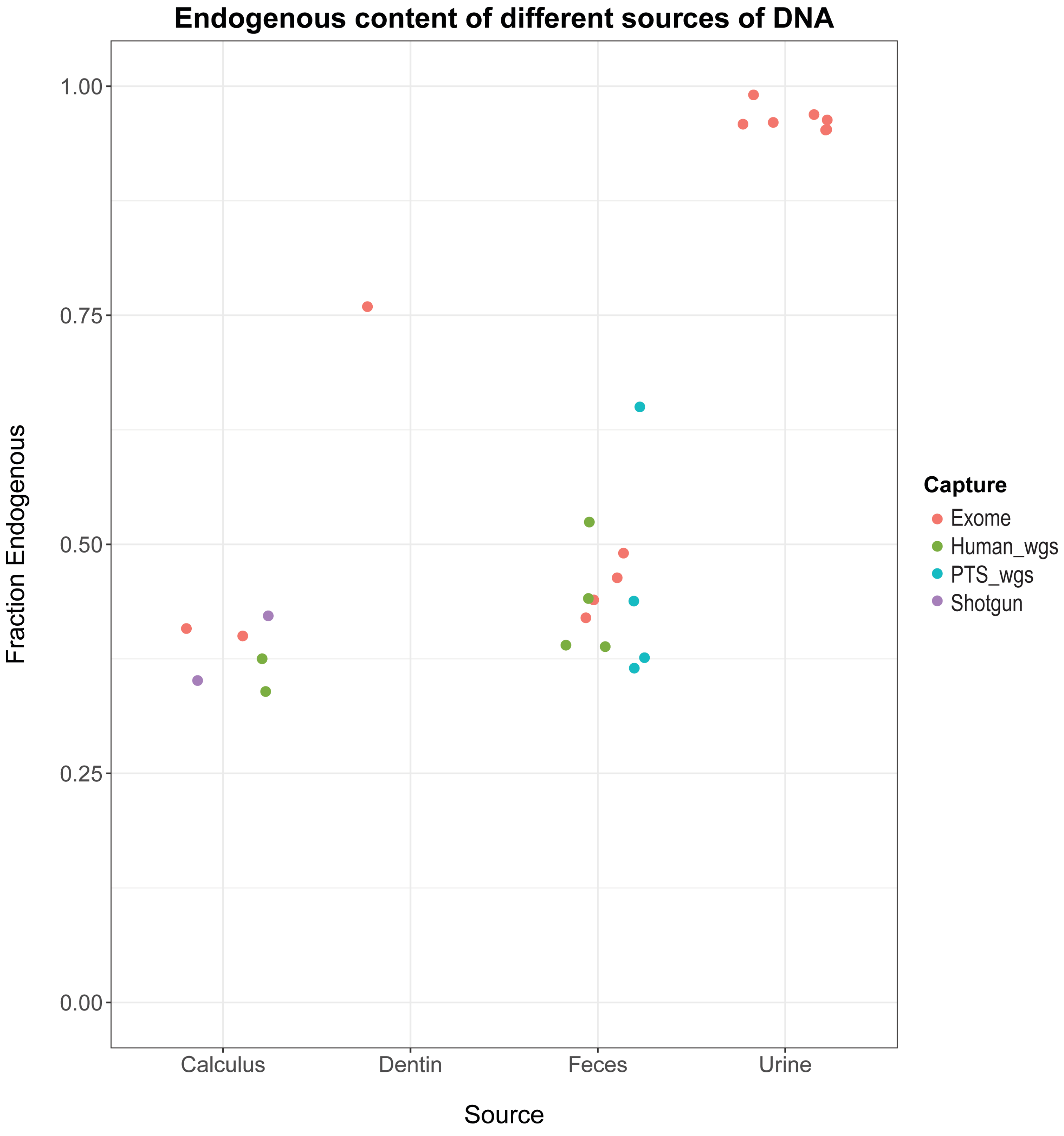
Ziesemer, K. A., Ramos-Madrigal, J., Mann, A. E., Brandt, B. W., Sankaranarayanan, K., Ozga, A. T., . . . Schroeder, H. (2018). The efficacy of whole human genome capture on ancient dental calculus and dentin. *American Journal of Physical Anthropology, Early View*. doi:doi:10.1002/ajpa.23763

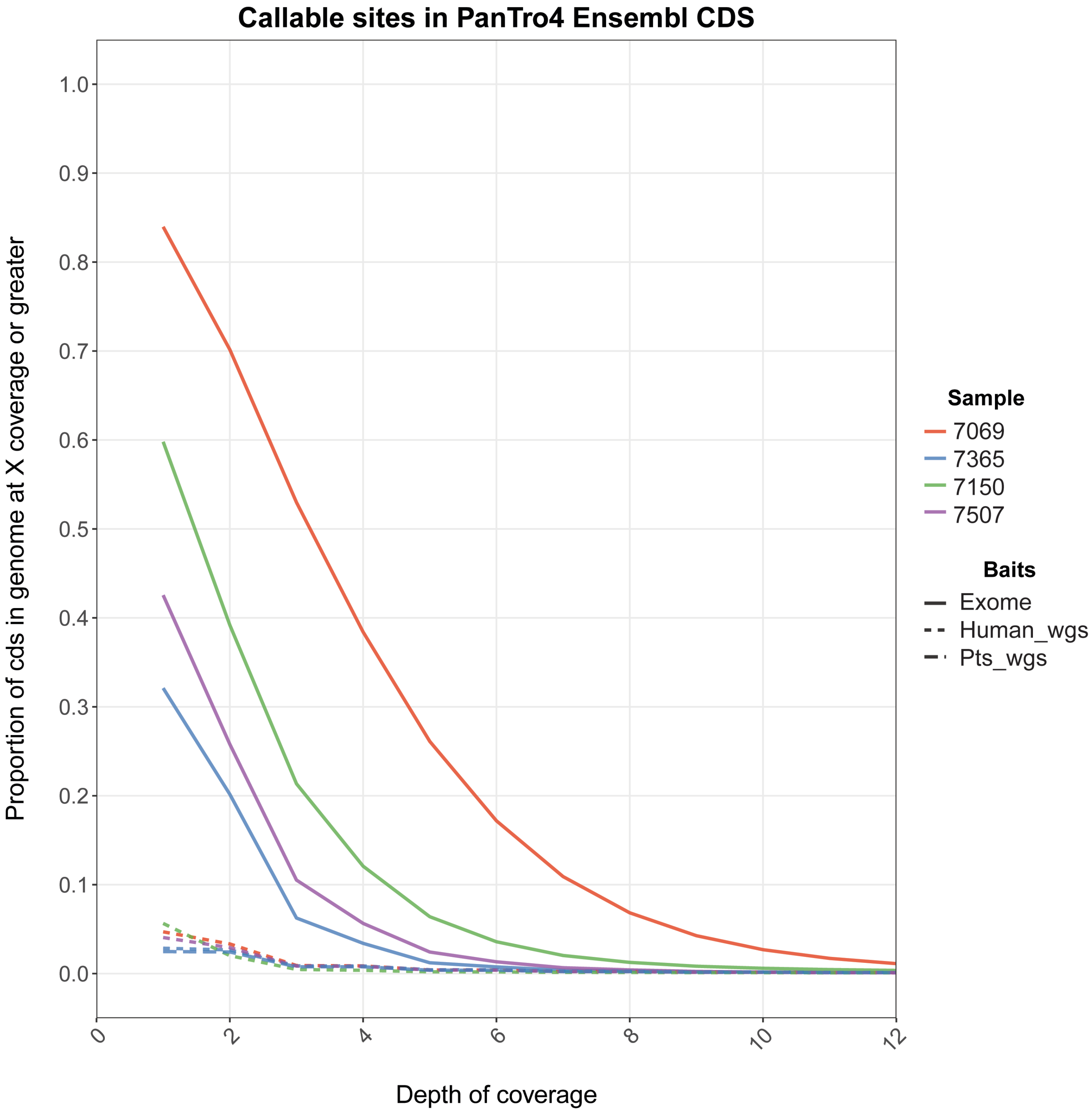
Data Accessibility

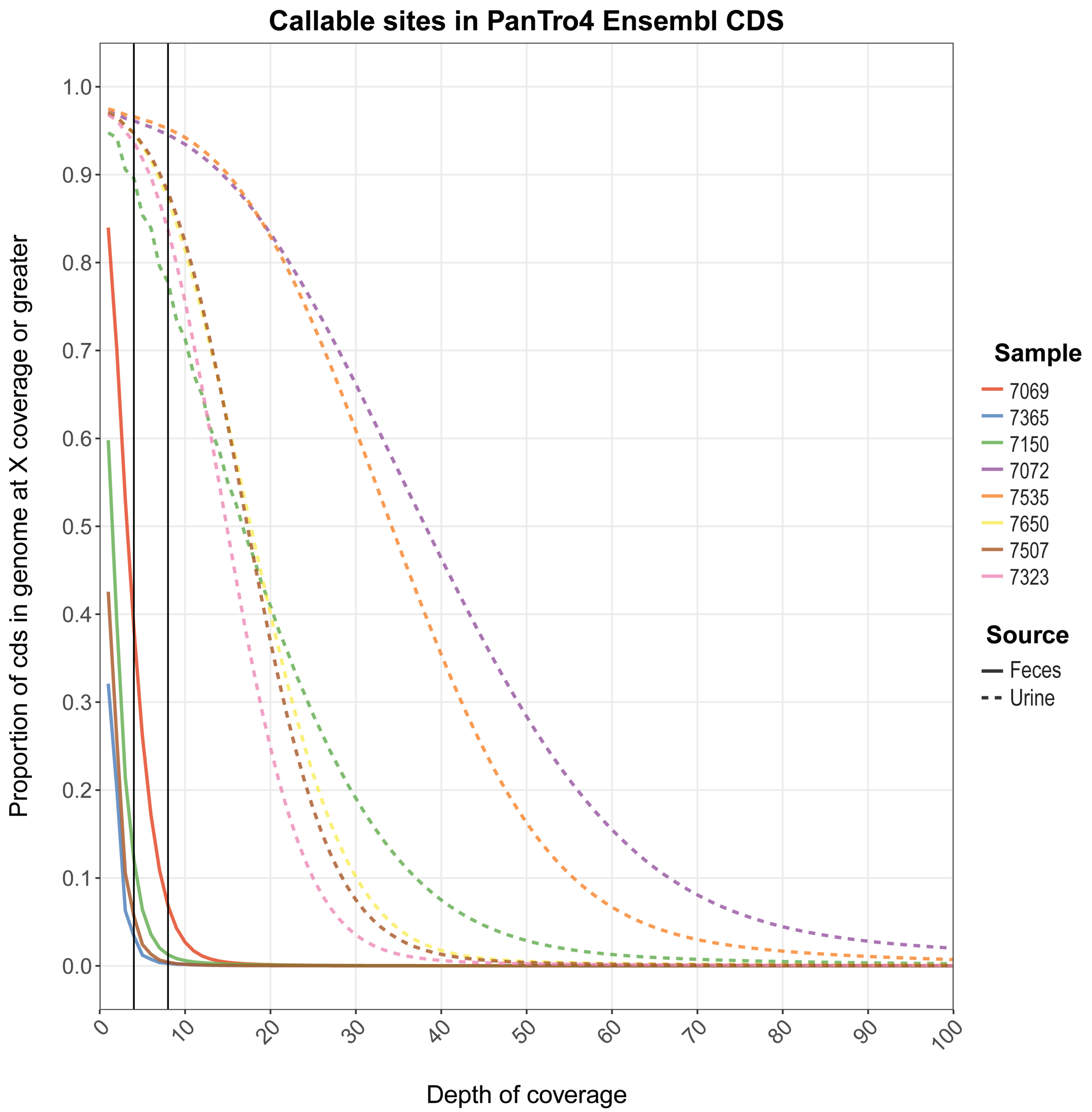
We deposited raw reads in NCBI's Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under Bioproject PRJNA508503 (Ozga et al., 2020). We implemented the full assembly and analysis pipeline in Snakemake (Köster & Rahmann, 2012), and managed software using Bioconda (Grüning et al., 2018). All code, scripts, and software environments are available on Github (https://github.com/thw17/Gombe_noninvasive_genomic_methods).

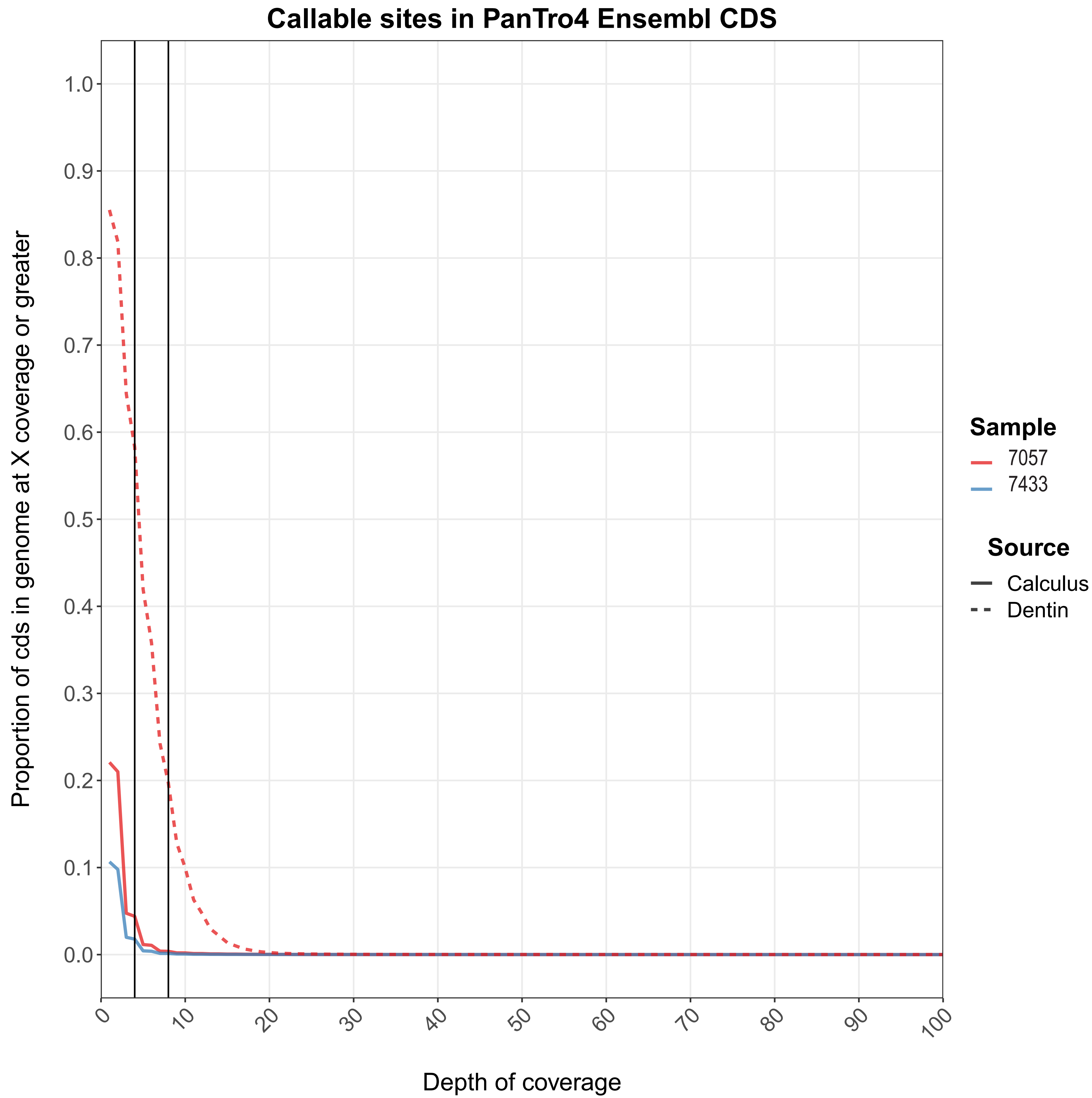
Author Contributions

ATO, THW, and ACS designed the study. RSN, ICG, AP, YL, and BH provided samples. ATO completed the lab work. THW analyzed the data. MAW and ACS provided laboratory space and funding. ATO and THW wrote the manuscript. All authors revised and approved the manuscript.


Endogenous content of different sources of DNA
Fraction Endogenous
1.00
0.75
0.50
0.25
0.00
Calculus
Dentin
Feces
Urine
Source
Capture
● Exome
● Human_wgs
● PTS_wgs
● Shotgun


Callable sites in PanTro4 Ensembl CDS
Proportion of cds in genome at X coverage or greater
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
2
4
6
8
10
12
Depth of coverage
Sample
7069
7365
7150
7507
Baits
Exome
Human_wgs
Pts_wgs


Callable sites in PanTro4 Ensembl CDS
Proportion of cds in genome at X coverage or greater
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Depth of coverage
0
10
20
30
40
50
60
70
80
90
100
Sample
7069
7365
7150
7072
7535
7650
7507
7323
Source
Feces
Urine


Callable sites in PanTro4 Ensembl CDS
Proportion of cds in genome at X coverage or greater
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
80
90
100
Depth of coverage
Sample
7057
7433
Source
Calculus
Dentin

Individual ID	Depth[a]	Source	Passing Variant Sites[b]	Passing in Both[c]	Shared Sites[d]	Dropout Sites[e]	Ambiguous Sites[f]	Dropout Rate[g]
7365	4	Feces	2,484	453	441	3	5	0.66%
7365	4	Urine	17,127	453	441	4	5	0.88%
7365	6	Feces	711	147	139	3	4	2.04%
7365	6	Urine	11,522	147	139	1	4	0.68%
7365	8	Feces	315	79	74	3	1	3.80%
7365	8	Urine	7,486	79	74	1	1	1.27%
7365	10	Feces	227	59	54	3	1	5.08%
7365	10	Urine	4,952	59	54	1	1	1.69%
7150	4	Feces	9,933	4,317	3,918	38	26	0.88%
7150	4	Urine	107,629	4,317	3,918	335	26	7.76%
7150	6	Feces	5,039	2,580	2,391	38	19	1.47%
7150	6	Urine	103,639	2,580	2,391	132	19	5.12%
7150	8	Feces	2,947	1,778	1,650	36	16	2.02%
7150	8	Urine	98,441	1,778	1,650	76	16	4.27%
7150	10	Feces	2,014	1,309	1,212	36	15	2.75%
7150	10	Urine	93,244	1,309	1,212	46	15	3.51%
7507	4	Feces	3,153	1,497	1,436	19	14	1.27%
7507	4	Urine	144,600	1,497	1,436	28	14	1.87%
7507	6	Feces	1,284	666	629	19	9	2.85%
7507	6	Urine	135,002	666	629	9	9	1.35%
7507	8	Feces	611	362	332	18	9	4.97%
7507	8	Urine	125,793	362	332	3	9	0.83%
7507	10	Feces	368	229	203	18	6	7.86%
7507	10	Urine	115,955	229	203	2	6	0.87%
7057	4	Calculus	3,555	322	291	10	8	3.11%
7057	4	Dentine	74,516	322	291	13	8	4.04%
7057	6	Calculus	1,291	132	113	10	3	7.58%
7057	6	Dentine	58,088	132	113	6	3	4.55%
7057	8	Calculus	349	56	40	10	2	17.86%
7057	8	Dentine	40,438	56	40	4	2	7.14%

7057	10	Calculus	115	38	27	7	1	18.42%
7057	10	Dentine	27,190	38	27	3	1	7.89%