BLISS-SCHRYER, MICHAEL, M.A., MAY 2024 ANTHROPOLOGY

IDENTIFYING *ATELES GEOFFROYI* INDIVIDUALS NONINVASIVELY USING THIRD-GENERATION SEQUENCING TECHNOLOGIES (92 pp.)

Thesis Advisor: Rafaela Takeshita

Genotyping animals is necessary for various field-based applications that require precise knowledge of the sampled individuals. Though feces are considered a low-quality source of host DNA, molecular techniques are increasingly prioritizing its usage for field-based noninvasive projects. Here, we describe a reproducible workflow to genotype individuals using a whole-genome sequencing approach with the portable, high throughput MinION MK1B and the BWA-GATK variant calling pipeline. After filtering, only 4 of the original 5,394 SNPs passed the filtering criteria, leading to an unsuccessful attempt to generate an informative multiloci SNP panel to confidently and accurately differentiate animals. In the filtered SNPs, 5 samples were entirely void of genotyping data. The majority of SNPs exhibited allelic dropout and a lack of called heterozygote genotypes, leading to the presumable false genotypes of the sampled individuals. On average, approximately 97% of the genome remained unsequenced, with only about one read covering each base in the mapped regions. Despite the limitations of employing a whole-genome sequencing approach to differentiate individuals with the MinION using feces, for species lacking known variants, this strategy may be an effective way to initially identify SNPs for subsequent resequencing and genotyping. Future studies are necessary to validate the authenticity of the identified SNPs and to assess their ability to discriminate individuals effectively with enrichment and targeted sequencing techniques.

IDENTIFYING *ATELES GEOFFROYI* INDIVIDUALS NONINVASIVELY USING THIRD-GENERATION SEQUENCING TECHNOLOGIES

A thesis submitted

to Kent State University in partial fulfillment

of the requirements for the

degree of Master of Arts

by

Michael Bliss-Schryer

May 2024

Thesis written by

Michael Bliss-Schryer

B.S., Temple University, 2020

M.A., Kent State University, 2024

Approved by

_____, Advisor

Rafaela Takeshita

_____, Chair, Department of Anthropology

Mary Ann Raghanti

_____, Dean, College of Arts and Sciences

Mandy Munro-Stasiuk

LIST OF FIGURES

LIST OF TABLES

Acknowledgments

I want to say thank you, first and foremost, to the Kent State University Department of Anthropology's faculty members for their continued support and mentorship. Thank you to all my supportive friends, anthropology cohort members, and family, especially my mom, my partner, and my cat, for being with me through all the ups and downs. Thank you to my fellow lab mates and collaborators, Brandie Nelson, Salan Ghaju, Emilee Hart, and Prashant Ghimire for all of your help and support throughout my thesis. Thank you to Dr. Rafaela Takeshita and Dr. Wilson Chung for teaching me how to be more productive scientist and academic and for encouraging me to discover and realize my potential. Thank you to Dr. Sangeet Lamichhaney for working through the bioinformatics pipeline with me and for his supervision of the MinION library preparation and loading steps. Thank you Roy Heath for helping to set up our lab's Virtual Machine. Thank you Dr. Linda Spurlock for all of our fun and stimulating conversations which helped me get through the difficult days. Thank you to Barbara Davis for helping me through the logistical and financial steps of my thesis and for her unwavering encouragement and support. Thank you Dr. Anthony Di Fiore for his insights in generating the SNP panel. Thank you Mariah Donohue for introducing me to adaptive sampling and enrichment techniques to overcome the limitations of fecal sequencing. Thank you Morgan Chaney for the MinION library preparation and loading advice. Thank you Irina Milke Pavlova for their design skills and help with the aesthetics of the figures. Thank you to Dr. Rafaela Takeshita, Dr. Richard Meindl, Dr. Anthony Tosi, and Dr. Sangeet Lamichhaney for being a part of my thesis committee and providing constructive critiques and comments. Thank you to Brittany Canfield and all the primate keepers at the Nashville Zoo for collaborating on this project. Thank you Kari Bagnall,

Sara Smith, and all the primate keepers at Jungle Friends Primate Sanctuary for your collaboration on this project, and for your hospitality and mentorship of Abbigail Swiney.

Chapter I - Introduction

Sequencing Background

Over the past half-century, advancements in sequencing technologies have markedly enhanced the accessibility and versatility of genomic applications (for a comprehensive review of sequencing history, see Heather & Chain, 2016). The advent of portable, lightweight sequencing devices like the MinION by Oxford Nanopore Technologies (ONT) (Oxford, UK) has revolutionized sequencing capabilities and possibilities (Hyden, 2015). Sequencing can now be conducted in a myriad of environments and conditions, including exceedingly inaccessible and daring locations such as the international space station (Castro-Wallace et al., 2017), Antarctica (Johnson et al., 2017), and the Ecuadorian Chocó rainforest (Pomerantz et al., 2018), among others. Moreover, the MinION offers a wide array of sequencing applications, including transcriptomics (e.g., Sahlin & Medvedev, 2021), genome assembly (e.g., Pozo et al., 2024), metabarcoding (e.g., van der Reis et al., 2022), genotyping (e.g., Cornelis et al., 2017), and epigenetics (e.g., Simpson et al., 2016), versatile to fit a suite of research needs and demands.

The emergence of sequencing methodologies and technologies, marking the inception of first-generation sequencing, closely followed the advancements in the conceptual framework developed in the mid-20th century regarding DNA's structure and function. This includes methods such as the chemical cleavage technique (Maxam & Gilbert, 1977), shotgun sequencing (Anderson, 1981), and Sanger sequencing (Sanger et al., 1977), the cornerstone of first-generation DNA sequencing technology. Sanger sequencing is a chain termination method, utilizing synthetic fluorescent dideoxynucleotide triphosphate (ddNTPs) molecules to cease DNA synthesis on the template strand, resulting in fragments of different lengths which are visualized though electrophoretic techniques. This methodology was ultimately automated and

scaled (Smith et al., 1986) by Applied Biosystems and utilized to sequence the first human genome (Venter et al., 2001). The development of Polymerase Chain Reaction (PCR) described by Saiki et al. in 1988, which utilizes heat-stable Taq polymerase to synthesize DNA, and recombinant techniques such as plasmid construction through bacterial transformation demonstrated by Cohen et al. in 1973, significantly enhanced sequencing capabilities, marking key advances in genetic and genomic research.

Second-generation sequencing methodologies set the stage for the development of massively parallel, high-throughput next-generation sequencing (NGS) platforms, including the 454 system by 454 Life Sciences and the Genome Analyzer by Illumina (see Shendure & Ji, 2008; Quail et al., 2012). These approaches transformed the field of molecular biology, inciting the expansion and development of bioinformatics and computationally-capable equipment to handle the significantly increasing data generation (see Gauthier et al., 2019). Three sequencing techniques tend to define this era: the pyrosequencing method, the Solexa sequencing method, and the Ion Torrent method. The pyrosequencing method, developed by Nyrén & Lundin in 1985, determines the DNA sequence by measuring the intensity of light emitted for each nucleotide incorporated into the template strand during DNA synthesis. Light emission is proportional to the amount of pyrophosphate released, wherefore the downstream catalytic effects of the enzymes ATP sulfurylase and luciferase produce characteristic light profiles. Alternatively, the Solexa sequencing method, later acquired by Illumina, utilizes bridge amplification and fluorescent reversible-terminator dNTPs, where each synthetic nucleotide emits a characteristic light profile. These molecules are read sequentially, as the attached fluorophore, which occupies the 3′ hydroxyl position, must be cleaved before the next nucleotide is added to the template strand (Turcatti et al., 2008). This method continues to be widely used

today because of its accuracy, scalability, and cost-effectiveness. The Ion Torrent method was the first technique to move away from optical detection, as it does not utilize fluorescence or luminescence detection mechanisms to determine the sequence. Instead, it measures changes in the pH resulting from the release of hydrogen ions during DNA synthesis (Rothberg et al., 2011).

While many first and second-generation sequencing methods utilize a sequence-by-synthesis (SBS) approach (i.e., interpret DNA sequences by nucleotide additions to the template strand through direct DNA polymerase action) third-generation sequencing adopts a single molecule sequencing (SMS) approach, enabling the direct detection of individual nucleotides. The first third-generation sequencing technology (Braslavsky et al., 2003) commercialized by Helicos BioSciences was vastly important because it did not require amplification stages. This laid the groundwork for the development of the single molecule real-time (SMRT) sequencing platform by Pacific Biosciences and ONT's highly anticipated nanopore technology (Clarke et al., 2009). These platforms enable instantaneous base calling and leverage the benefits of long reads, thus overcoming the challenges posed by highly scaffolded and fragmented genomes, enabling more precise detection of variant and repeat regions.

Nanopore MinION Sequencing

Nanopore sequencing operates by passing single-stranded DNA molecules through a nanopore (Church et al., 1995; Kasianowicz et al., 1996). Nanopores fall into two general categories: biologically derived protein molecules sourced from bacteria (e.g., α-hemolysin from Staphylococcus aureus) or synthetically produced solid-state nanopores (for a comprehensive review of nanopore types, see Haque et al., 2013). There are 2,048 nanopores distributed across the 512 channels on the MinION's flow cell. As single-stranded DNA molecules pass through a

nanopore, the nucleotides induce characteristic disruptions in the nanopore's electrical current, referred to as 'squiggles'. These signals are captured in a FAST5 format file utilizing ONT's sequencing software, MinKNOW. Guppy, the base calling program integrated into MinKNOW's interface, interprets the FAST5 file by identifying squiggles that correspond to specific nucleotide signals. The result is a Fast Quality (FASTQ) file containing reads along with their associated Phred Quality Scores (PQS).

Nanopore technology, notably the MinION, made available to early-access users in May 2014 (Jain et al., 2016), holds significant potential for decentralizing sequencing applications and drastically reducing sample costs. We are close to achieving the milestone set by the National Institutes of Health, which aims to sequence and catalog mammalian genomes for under 1,000 United States Dollars (USD). The MinION starter pack, available for roughly 1,000 USD, includes the device itself, along with a single flow cell and sequencing reagents. For subsequent projects, additional consumables can be purchased separately. Flow cells typically range from 500-900 USD (depending on quantity), while the cost of reagents for library preparation and sequencing varies depending on project specifications. Multiplexing sequencing can significantly reduce per-sample costs, with certain kits capable of simultaneously sequencing 96 samples. This high-throughput technology is capable of reading fragments spanning thousands of bases, thereby enhancing both sequence and variant resolution, essential for genome assembly and genotyping-based tasks. Nanopores have the theoretical capability to read fragments of any length, contingent upon the size of the DNA molecules. Jain et al., 2018, for example, achieved reads of up to 882 kb in length while sequencing the human genome using the MinION.

Despite the significant sequencing advantages afforded by nanopore technology, several challenges still require further improvement. These include enhancing the recognition and

resolution capabilities of base calling (see Branton et al., 2008) and improving read accuracies compared to established NGS short-fragment methods. ONT has made substantial improvements to their base calling algorithms and flow cell chemistries (see Magi et al., 2018; Leggett & Clark 2017). The transition from the initial R6 model to the current R10.4 platform has led to notable increases in base calling accuracy. The earliest models demonstrated an approximately 60% accuracy rate (Laver et al., 2015; Goodwin et al., 2015), whereas more recent studies report accuracy levels exceeding 85% for the newer models (Jain et al., 2017; Tyson et al., 2017; Seki et al., 2019). Moreover, the utilization of post-sequencing correction tools can increase accuracy levels above 99% (see Rang et al., 2018). Although the MinION may not be the first choice for routine sequencing tasks, its portable and adaptable nature makes it indispensable for numerous research applications beyond the scope of the traditional laboratory setting.

Noninvasive Sequencing Approaches

For studies requiring precise individual identification (e.g., behavioral endocrinology), field-based sequencing can be a tremendously advantageous tool. Many animal populations pose significant challenges in distinguishing individuals, especially in monomorphic species where phenotypic differences are minimal. Additionally, environmental conditions (e.g., dense canopy coverage) and behavioral circumstances (e.g., tightly clustered group configurations) can hinder observation, further complicating visual identification. Addressing these challenges often entails years of studying the population to recognize individual qualities beyond distinctive physical traits, such as behavioral patterns and hierarchical relationships, a process that may still yield uncertain results. Alternatively, employing a sequencing strategy can help researchers accurately and confidently identify samples, even when direct observation is unattainable (e.g., Eriksson et

al., 2004; Waits & Paetkau, 2005; Beja-Pereira et al., 2009; Yang et al., 2012). In field-based endocrinology studies, for example, that rely on fecal samples containing hormone metabolites, researchers can split the sample; one half for hormone analysis, and the other for genotyping.

Field studies utilizing molecular techniques have increasingly prioritized the use of non-invasive sampling (NIS) or minimally invasive sampling (MIS) approaches (see Carroll et al., 2018 for a comprehensive review) following the pioneering application of MIS techniques using feces to measure genetic variation in a brown bear (*Ursus arctos*) population (Höss et al., 1992). Fecal collection has been at the forefront of this revolution because it enables researchers to isolate DNA without harming the subject animals or disrupting their natural behaviors, maintaining ethical standards and achieving project requirements. Invasive approaches, like sedative darting, present several challenges (see Cunningham et al., 2015). Not only is the equipment expensive and its usage disrupts natural behaviors, but it poses risks to both animals and humans in the vicinity. Inaccuracies may result in injury or death to the animal, especially arboreal species prone to falling as a consequence of the administered sedatives. Additionally, preanesthetic assessments of the animal's condition are often limited and inadequate (Mosley & Gunkel, 2007).

Even certain MIS techniques that utilize specialized darts equipped with collection mechanisms, such as barbs or adhesive patches (or directly adhering these collection gadgets to environmental substrates) to gather hair or tissue samples, can pose challenges, potentially causing undue stress and or open wounds. While collaring individuals can be advantageous for tracking animals and differentiating them, this method should not be solely relied upon, especially for species inhabiting highly complex environments and or social systems. In such instances, each collected sample necessitates 100% identification certainty, requiring a direct and

uninterrupted line of sight of both the animal (more specifically its recognizable collar) and the defecation, which can be challenging and occur infrequently under normal field-conditions. Moreover, collaring every individual is often impractical due to funding and time constraints. Plus, it is not uncommon for animals to remove them. In the previously described scenario, where the researcher separates the fecal sample for hormonal analysis and genotyping, after conducting a comprehensive survey and collecting from all individuals of interest (e.g., adult females), feces can be collected without a complete knowledge of the individual's identity or a direct line of site. The primary requisite to warrant collection in this scenario is a general consensus that the individual is presumably female, that is, if representative traits (e.g., sexual dimorphisms, body size) are evident. This enables researchers and field technicians to collect more samples, increasing the sample size of the targeted demographic population. In the case of certain population genetic studies, random feces (without any knowledge of the individual's identity) can be used to measure characteristics such as nucleotide diversity and the effective population size (e.g., Eriksson et al., 2004; Perry et al., 2010; Yang et al., 2012).

Obtaining blood or tissue samples from every individual in a population, especially for endangered species, presents significant challenges due to the difficulty in obtaining permits for invasive studies, the size of populations, and the behaviors (e.g., high dispersal rates) of the species. Creating a genotyping panel exclusively from high-quality sources of DNA requires extensive time and resources and may still fail to catalogue every individual. Instead, many researchers opt to sample a small subset of the population to discover informative variant sites. They then utilize feces to generate genotype panels and to profile individuals (e.g., Perry et al., 2010 for the western chimpanzee, Kraus et al., 2015 for the grey wolf, and Fitak et al., 2016 for pumas). However, obtaining blood or tissue samples is not always feasible, ethically acceptable,

or appropriate. The ability to generate genotyping panels exclusively from low-quality sources of DNA, such as feces, has the potential to significantly transform the landscape of molecular-based field research and conservation efforts. Moreover, it could help democratize genotyping, making it more accessible to a larger human population with varying degrees of experience, expertise, funding, permits, and ethical concerns. Though NIS techniques using feces have inherent limitations, they represent the future of field-based molecular approaches and applications, offering solutions to many of the drawbacks of traditionally employed invasive methods.

One of the more obvious limitations for utilizing feces is that host DNA is just a fraction compared to contributions from microorganisms (Stephen & Cummings, 1980) and ingested biota, typically with less than 1% of the total sequencing reads belonging to the host (Wanner et al., 2021; Sharma et al., 2019). Remarkably, microbial cells outnumber human cells in the body by a magnitude of ninefold (Hooper et al., 1998), showcasing incredible diversity, potentially comprising up to 1,000 distinct bacterial species within the human microbiome alone (Sekirov et al., 2010) which play essential roles in physiological functions such as carbohydrate digestion and vitamin synthesis (McFarland, 2000). Mammalian host DNA in feces is a result of the shedding of epithelial cells lining the intestinal wall (van der Flier & Clevers, 2009). Cell turnover is so rapid, the intestinal wall replaces itself every 4-5 days (van der Flier & Clevers, 2009). Though the amounts of host DNA in feces may vary by the day, individual, and species, theoretically, there should be traces within each stool sample that can be extracted. In addition to quantity concerns, DNA extracted from feces is generally of lower quality compared to other biological substrates, such as blood or tissue, typically exhibiting significant fragmentation (Taberlet et al., 1996). Moreover, fecal samples frequently contain chemicals and metabolites that can inhibit and reduce the effectiveness of certain reactions such as PCR, (Kohn & Wayne,

1997), which further contributes to the difficulties of extracting and amplifying high quality DNA from feces.

To minimize interference from microbial DNA, pre-sequence enrichment techniques like FecalSeq (Chiou & Bergey, 2018) have been developed to preferentially isolate host DNA or deplete bacterial DNA. Additionally, real-time sequencing features, such as adaptive sampling (also referred to as adaptive sequencing, selective sequencing, or 'Read Until'), have been designed to reject fragments that do not align to a specified reference genome by reversing the voltage across the nanopore (https://github.com/nanoporetech/read_until_api) (Loose et al., 2016). The sequences that are rejected are unlikely to be read again due to the change in position of their motor protein, which is necessary to initiate the sequencing process. These enrichment techniques can enhance the proportion and concentration of host DNA (or deplete host DNA in microbial studies; see Marquet et al., 2022), improving the performance of sequencing technologies with constrained capacities such as the MinION, which experiences pore degradation during sequencing. The goal of these approaches are to ultimately enhance host-specific sequencing yields with increased coverage.

Most MinION studies utilizing feces as the biological substrate have predominantly investigated the microbiome (e.g., Moss et al., 2020), with only a handful seeking to study the host organism's genome itself, whether nuclear or mitochondrial (e.g., Wanner et al., 2021). Wanner et al., 2021 sequenced and assembled the mitochondrial genome (mitogenome) of the golden lion tamarin (*Leontopithecus rosalia*) using the unfinished mitogenome of the closely related emperor tamarin (*Saguinus imperator*) for assembly with adaptive sampling. Using this approach, they were able to achieve 258x coverage, doubling the concentration of host DNA compared to non-enrichment sequencing. Adaptive sampling holds immense potential for

enhancing coverage by selectively enriching for host-specific reads. However, it can be computationally intensive, with many methods demanding substantial processing power beyond the capabilities of a typical laptop's central and graphic processing units (CPUs & GPUs).

Adaptive sampling requires that a decision be made whether the reads align to the reference before the read is fully sequenced. Over the years, several methodologies have been developed that interact directly with ONT's 'Read Until' application programming interface (API). Loose et al., 2016 first developed an adaptive sampling approach by using a Dynamic Time Warping (DTW) algorithm to match the electrical signals (without base calling) against a simulated reference (which they referred to as a reference squiggle). This was encouraged by the similarity between nanopores' squiggles and audio signals, with its usage to catch pronunciation errors (Miodonska et al., 2016) and compare gene sequences for phylogenetic investigations (Skutkova et al., 2015). The idea of using DTW for sequence comparison has existed since the early 1980's (Sankoff & Kruskal, 1983). Besides from being computationally expensive, this method only works for matching reads to shorter references, maxing out at about 10 kb. The method 'Utility for Nanopore Current ALignment to Large Expanses of DNA' (UNCALLED), developed by Kovaka et al., 2021 (https://github.com/skovaka/UNCALLED) was introduced several years later. This method converts squiggles into events that are matched to specific known *k*-mers (i.e., various combinations of DNA sequences). These *k*-mers are then searched for throughout the reference using the Ferragina–Manzini index (Ferragina & Manzini, 2000). In contrast, the tool 'Read Until with Basecall and Reference-Informed Criteria' (RUBRIC) developed by Edwards et al., 2019 (https://github.com/sandialabs/RUBRIC) is unlike the previous methods in that it takes advantage of real-time base calling, avoiding the need to recreate reference models to map the raw squiggles to. Similarly, Readfish (Payne et al., 2021)

(https://github.com/LooseLab/readfish) utilizes MinKNOW's base caller Guppy and the aligner Minimap2 (Li, 2018) to make read decisions.

Even the newest algorithms and methods are exceedingly computationally intensive, far surpassing the required computing capabilities of a standard laptop. Stevanovski et al., 2022 utilized 3090 GPUs (NVIDIA RTX), Payne et al. 2021 employed 1080 GPUs (NVIDIA GeForce GTX), Wanner et al., 2021 utilized 2080 GPUs (NVIDIA GeForce RTX), and Frank et al., 2023 employed either 4000 GPUs (Nvidia Quadro RTX) or 3080 GPUs (Nvidia GeForce RTX) for successful execution of adaptive sampling. These requirements are debilitating to the portability of the MinION for the use of host enrichment using non-invasively collected substrates like feces under field conditions. Currently, for large genomes (> 1 Gb), there are no 'portable' solutions to run adaptive sampling. However, software-hardware solutions intended to bring down the computational requirements are beginning to make adaptive sampling more accessible (e.g., Shih et al., 2023), though they are still limited to smaller genomes and require further refinement.

Fecal DNA Extraction Optimization

Various commercially available kits, such as the E.Z.N.A Stool DNA Kit from Omega BioTek (Norcross, GA), along with standard laboratory techniques like the phenol chloroform extraction method (PCEM), can effectively isolate DNA from feces. The extraction methods researchers utilize are contingent on the project's objectives, conditions, and funding. While commercial kits are capable of providing high quality DNA in a few hours, they tend to be more expensive than in-house techniques. Therefore, if the project necessitates a fast turn-around time and has sufficient funding, using a commercial kit might be preferrable. With many different

techniques and products available, it is important to find the right extraction method to meet the project's needs.

For advancing non-invasive approaches in molecular biology, it is crucial to continue enhancing existing protocols and developing novel ones for extracting DNA from fecal samples. In addition to utilizing the E.Z.N.A Stool DNA Kit, we aimed to optimize the PCEM to extract high quality DNA from *Ateles geoffroyi* feces. The PCEM, developed and refined in the mid-20th century, leverages two fundamental chemical properties: differential solubility and phase separation. After lysing cells with a detergent-based solution to break down cell membranes and release intra-cellular compounds, proteases are introduced to digest contaminant proteins, including nucleases that may degrade nucleic acids. Following this, phenol and chloroform are added to form two distinct layers: the less dense aqueous phase and the more dense organic phenol chloroform phase. The cellular compounds are preferentially dissolved into either phase contingent on their polarity. The aqueous phase, containing nucleic acids, is carefully extracted and further purified. The addition of salts and alcohol then precipitates the nucleic acids out of solution, rendering them insoluble, after which centrifugation pellets the DNA to be reconstituted.

The PCEM is an effective extraction method (e.g., Ghaheri et al., 2016), capable of yielding DNA suitable for amplification through PCR (Barbosa da Silva et al., 2020). When compared to the reference, these PCR products exhibited identity statistics greater than 95% (Barbosa da Silva et al., 2020). While further investigation is needed to assess the practicality of the PCEM for successful sequencing using long-read technology, Trigodet et al., 2021 demonstrated that it may not be as effective as other commercially available kits for MinION sequencing. Despite these initial results, the affordability of the PCEM remains a leading

incentive for its continued usage and optimization. Price comparisons against the popular QIAamp DNA Mini Kit by QIAGEN (Germantown, MD) demonstrate a per-sample cost reduction of approximately 3 USD (Celerino da Silva et al., 2023), with each sample costing a little more than a 1 USD (Koshy et al., 2017). Keeping extraction method costs low will help improve accessibility and facilitate its refinement.

Similar to the importance of extraction techniques, sample storage is critical for the preservation of DNA under field-based conditions. Prior to departure, it is essential to consider the available equipment, as refrigeration options may be limited. While immediate collection of fresh samples is ideal for sequencing, it is often impractical for larger population sampling studies, due to the inability to collect enough samples for multiplexing in a single field session. This necessitates that sample collection occurs over several days, weeks, or even field seasons. Thus, preservation methods that can deliver comparable DNA qualities and concentrations to fresh samples are needed. Hale et al., 2015 demonstrated that freezing samples or storing them in ethanol or RNAlater yielded similar purity scores through time, though freezing and ethanol preservation most closely resembled the quality of fresh feces. Larsen et al.'s (2015) study alternatively shows that RNAlater is the preferred storage medium, outperforming freezer preservation. Nsubuga et al., 2004 document the advantages of silica-dried feces and introduce a novel silica-desiccation method following a brief ethanol storage period. Soto-Calderón et al., 2008 did not observe demonstrable differences in DNA quality across conditions when stored for one week. However, they found that silica-dried feces performed best for nuclear microsatellite markers, while RNAlater was most successful for mitochondrial analyses. Ethanol, alternatively, did not perform well and is not recommended by the authors. These studies demonstrate that nucleic acid quality and concentrations are dependent on factors such as storage time and

condition, downstream applications, and the species under investigation. Therefore, finding the right storage preservation technique to fit the project's specifications is critical for maximizing sequencing outputs.

Genetic Markers for Genotyping

There are numerous types of variants in eukaryotic genomes. Even before the advent of sequencing technology, researchers sought to identify and utilize genetic markers for individual differentiation (for a review see Allendorf, 2017). One of the earliest methods involved allozymes, isoforms of an enzyme encoded by different alleles at the same locus (Prakash et al., 1969). However, this technique had obvious limitations, as not every enzyme is polymorphic. In the 1970s, restriction fragment length polymorphisms (RFLPs) emerged with the discovery of restriction enzymes, cleaving DNA fragments at targeted recognition sites (Avise et al., 1979; Williams, 1989). These polymorphic loci exhibit variable numbers of restriction sites among individuals. Both allozymes and RFLP techniques rely on electrophoretic visualization to detect variations in either fragment sizes or the number of fragments. As more genetic markers were discovered, the use of these methods declined. Microsatellites, also referred to as simple sequence repeats (hereafter referred to as SSRs), emerged in the 1980s and were first mentioned for their potential application and significance for population genetics investigations by Bruford & Wayne, 1993. SSRs significantly gained traction as access to PCR technology increased.

While SSRs have often been, and in some cases still are, the preferred genetic marker due to their ease of development and ability to produce highly informative information across loci (Selkoe & Toonen, 2006), single nucleotide polymorphisms (SNPs) are increasingly being favored due to their abundance and distribution across various regions of the eukaryotic genome

(e.g., coding, noncoding). SNPs occur roughly every 300-2,000 nucleotides in humans (Brookes, 1999; Lindpaintner, 1999; Nelson et al., 2004), accounting for 90% of all sequence variation (Collins et al., 1998), affording high-resolution information for intra and interspecific investigations. These point mutations involve the substitution of a single base, (e.g., adenine for cytosine), occurring within the population at a frequency of at least 1% .

Being multiallelic, SNPs can exhibit significant variation across individuals, especially at sites with representation from each nucleotide (maximum 10 possible genotypes per loci). However, their appeal and usefulness for sequencing are not contingent nor dependent on their multiallelic diversity, but rather when loci exhibit biallelic propensities. Due to their binary nature, biallelic loci are easy to model, score, and analyze in comparison with SSRs sites (Brumfield et al., 2003; Garvin et al. 2010; Carroll et al., 2018) which have the potential to exhibit high degrees of intraspecific variation. While less SSR loci are needed to identify individuals within a population compared to SNPs (Morin et al. 2004) due to their high variability as a result of SSRs high mutation rates (see Zhang & Hewitt, 2003; Ellegren, 2004), characterizing and scoring SSRs can be challenging and may contribute to incorrect genotype assignments (Hoffman & Amos, 2005). Moreover, Morin et al., 2004 suggest that SSR's high mutation rates present the challenge of homoplasy and data ambiguity, potentially reducing their biological usefulness (Hedrick, 1999). With large sample sizes, sequencing is preferred over electrophoretic visualization. SNPs, which have lower mutation rates (Nachman & Crowell, 2000; Conrad et al., 2011), allow for simpler and less complex analyses, often resulting in more successful genotype assignments compared to SSRs (Fitak et al., 2016). SNP identification is well-suited for high-throughput sequencing technologies, affording a cost-effective approach to investigate large populations across multiple loci simultaneously (for its application in NGS, see

Nielsen et al., 2011). Furthermore, numerous bioinformatic workflows and pipelines, such as the Genome Analysis Toolkit (GATK), are available to analyze and visualize the data effectively for genotyping (McKenna et al., 2010; Van der Auwera et al., 2013). When GATK's HaplotypeCaller detects variations compared with the reference, it reassembles the reads and calls potential haplotypes based on their likelihoods, allowing for simultaneous SNP calling through local de novo assembly (Li et al., 2018).

SSRs have been particularly valuable for their contributions to forensic analysis and human identification, though recent advancements in the generation of SNP panels (see Kayser & de Knijff, 2011) have resulted in the influx of investigations into this alternate method for paternity testing and genotyping. First proposed and developed by Syvanen et al., 1993, for these biallelic markers to have the same discriminatory power as 12 SSR loci, SNP panels require approximately 50 loci (with allele frequencies ranging between 20-80%; Gill, 2001). The lower mutation rates of SNPs make them more reliable to determine relationships in applications such as paternity testing, as they are more likely to represent the alleles inherited from the parents (see Børsting et al., 2012). The likelihood of paternity exclusion (probability of not being the father) across the 50 loci mentioned in Gill 2001 is greater than 99%, wherefore the discriminatory power of 4.2 SNPs with allele frequencies of 50% is equal to that of one SSR loci for paternal exclusion (Krawczak, 1999). Ayres, 2005 likewise confirms that SNPs with allele frequencies of 50% are optimal for paternity exclusion testing.

SNPs have been generated using the MinION (for a review of the forensic application potential with the MinION, see Plesivkova et al. 2019), where Cornelis et al., 2017 demonstrated its feasibility with all but one loci being correctly genotyped and Ren et al., 2021 showed over a 99.9% success rate with only one incorrect genotype out of 2,926. One of the potential pitfalls of

using the MinION in forensic applications is that it requires 400-1,000 ng of DNA (depending on the library preparation kit used), which might not be possible to obtain in circumstances where not enough genetic material is available at the scene. This could also translate as a potential issue for studies trying to isolate ancient DNA (e.g., Hui et al., 2020) or for projects under field conditions using low quality sources of DNA like that from feces.

Genotype data is imperfect and prone to exhibit false genotypes due to a number of factors, including the use of low quality samples and human errors along the way, among others (Pompanon et al., 2005). Bonin et al., 2004 discuss approaches to assess and combat genotyping errors, recommending the use of negative controls and independent repeatability tests, among other things. Allelic dropout, where one allele remains undetected, can lead to the false appearance of homozygosity. This issue is especially common and problematic in sequencing data obtained from feces or other low-quality biological substrates due to their limited ability to generate sufficient coverage (Gagneux et al., 1997; Nielsen et al., 2011). Using higher quality samples (e.g., blood or tissue) to originally generate the SNP panel is often preferred to avoid erroneous or missing genotypes, wherefore feces can be utilized subsequently (resequencing) for future genotyping applications.

However, Buerkle & Gompert, 2013 demonstrated that it is possible to obtain informative population-wide data with low coverage reads (1x), suggesting the possibility of generating a SNP panel with feces alone. In current variant calling workflows and pipelines, the preferred method for SNP calling is to utilize genotype likelihood statistics alongside population-wide allele frequencies. Likelihood statistics represent the probability of a given genotype being accurately assigned given the quality and parameters of the available sequencing data (Nielsen et al., 2011). To further filter for likely SNP loci, researchers measure deviations in genotype

distribution from Hardy-Weinberg Equilibrium (HWE). Loci showing significant deviations ($p < 0.05$) in observed versus expected heterozygosity (under HWE) are discarded (except when there is a valid explanation) as they typically indicate genotyping errors (Hosking et al., 2004; Wigginton et al., 2005). The use of genotype likelihood models, which are designed to provide unbiased and accurate genotyping data, is extremely advantageous for the generation of SNP panels using low-quality biological material.

Variant Identification & Resequencing

To initially identify variants, researchers commonly employ two main sequencing strategies: whole-genome or targeted sequencing. Whole-genome sequencing involves sequencing the entire genome, allowing for the downstream detection of variations against a high-quality reference (e.g., Hillier et al., 2008; Ng & Kirkness, 2010; Subbaiyan et al., 2012). This approach can help identify rare variants (see Cirulli et al., 2010). Genome-wide association studies (GWAS), which can be lumped into the whole-genome sequencing approach, involves the analysis of large genomic datasets of multiple individuals in parallel to characterize the associations between variants and phenotypic traits (e.g., Nagasaki et al., 2014; Littiere et al., 2020; Yengo et al., 2022). In contrast, targeted sequencing selectively isolates DNA fragments containing specific regions of interest across the genome (e.g. restriction fragments, exons). This can be accomplished with techniques like Restriction site-Associated DNA sequencing (RAD-seq) and double digest RAD-Seq (ddRAD-Seq) (see Baird et al., 2008; Andrews et al., 2016; Peterson et al., 2012) or PCR reactions using complementary oligonucleotides (see Goswami, 2016). While NGS and nanopore technologies are high-throughput and have the capacity to achieve high coverage across the entire genome, targeted sequencing is a cost-effective

alternative capable of generating sufficient amounts of SNPs for genotyping or phylogenetic analyses (Valencia et al., 2018; Martins et al., 2023). The choice between approaches depends on factors such as funding availability, sequencing methodologies, and the specific research question being addressed.

To perform genotyping via resequencing after the variants have been validated, these loci are typically amplified through PCR to ensure adequate coverage. Depending on the study's objectives and conditions (e.g., sample size, available equipment), various post-PCR methods may be employed, including sequencing or gel electrophoresis. Gel electrophoretic visualization, which is specifically used for SSR analysis, can be labor-intensive and require more consumables than the on-site utilization of third-generation sequencing technology. Additionally, the extensive infrastructure often required for visualization (e.g., thermal cycler, benchtop UV Transilluminator) can necessitate shipping samples to a primary laboratory, which can be costly and bureaucratically challenging, especially for projects where fieldwork is based at a remote location. Though compact and versatile technology such as the Bento Lab (e.g., Hirabayashi et al., 2021) (https://bento.bio/product/bento-lab/) can perform PCR amplification and gel electrophoresis, these smaller devices can only perform small sample batches, meaning more processing time and less data, often restricting the scope of the study to smaller populations.

Available Ateles Genotyping Data

To date, there are eight species within the genus *Ateles* with complete reference genomes (*Ateles geoffroyi*, *Ateles hybridus*, *Ateles paniscus*, *Ateles marginatus*, *Ateles belzebuth*, *Ateles chamek*, *Ateles fusciceps*), each with varying degrees of annotation. To the author's knowledge, there is only one publicly available genotyping record for these species. *Ateles belzebuth* has

been the subject of decades of molecular research, including dispersal (Di Fiore et al., 2009) and demographic (Link et al., 2018) studies. In 2004, Di Fiore & Fleischer identified several SSR markers for *Lagothrix lagotricha* identification, which subsequently work for *Ateles belzebuth* genotyping. To date, the majority of the genotyping studies on platyrrhine primates have employed SSRs, including species *Alouatta palliata* (Ellsworth & Hoelzer, 1998), *Alouatta belzebul* (Gonçalves et al., 2004), *Saimiri boliviensis* (Witte & Rogers, 1999), *Cebus apella* (Escobar-Páramo, 2000), *Callithrix jacchus* (Nievergelt et al., 2000), *Leontopithecus rosalia* (Grativol et al., 2001), *L. chrysopygus* (Perez-Sweeney et al., 2005) and *L. chrysomelas* (Galbusera & Gillemot, 2008). As previously mentioned, many of the developed markers work across species, including *L. caissara* (Martins & Junior, 2011), *Aotus azarai* (Babb et al., 2011), *Callicebus moloch* (Menescal et al., 2009), and broadly speaking for species from the Atelidae, Pitheciidae, and Cebidae families (Di Fiore & Fleischer, 2004).

Study Objectives

To address the limitations presented by SSR visualization, we aimed to develop a genotyping protocol for *Ateles geoffroyi* to enable on-site sequencing of SNP loci using the MinION. *Ateles geoffroyi*, native and endemic to the Central American countries of Belize, Costa Rica, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, and Panama, is considered an endangered species (Cortes-Ortíz et al., 2021). Population projections indicate a potential 50% decline by 2065 if current rates of deforestation persist (Cortes-Ortíz et al., 2021). Although ~70% of the original forest cover has been destroyed or converted for human use (Estrada et al. 2006), there are more than 400 protected areas within *Ateles geoffroyi's* distribution. Occupying diverse habitat types, including tropical rainforests, deciduous, and montane forests, spider monkeys are

primarily arboreal, venturing to the ground almost exclusively to access minerals through salt licks or rotten wood (Campbell et al., 2005; Blake et al., 2009). Similar to many other platyrrhine species, *Ateles* forms multi-male, multi-female groups. Notably, the genus *Ateles* showcases unique social dynamics, with groups that periodically fission and fuse in response to seasonal variations in food availability and distribution (Cant, 1977; van Roosmalen, 1985; Spehar et al., 2010).

Many species of *Ateles* are endangered, necessitating urgent monitoring and research to safeguard their populations. Research groups, like the Proyecto Primates Research Group (PPRG) at the Tiputini Biodiversity Station in Amazonian, Ecuador (76°08′W, 0°38′S), are actively studying spider monkey populations across various field stations. The implementation of noninvasive genotyping techniques could significantly enhance these efforts by facilitating both direct research applications and wildlife monitoring. Here, we propose the generation of genotyping panels exclusively from fecal samples and opportunistically collected high-quality DNA sources to overcome the limitations of traditional invasive sampling methods. This approach offers the potential to detect informative variant sites and profile and identify individuals while simplifying genotyping efforts and ensuring feasibility, particularly in areas and for species where regulations prohibit invasive techniques. Moreover, utilizing NIS techniques streamlines the collection process to focus primarily on obtaining fecal samples from each individual.

We conducted whole-genome sequencing of *Ateles geoffroyi* fecal samples using the MinION MK1B platform (MinKNOW v23.11.3). SNP calling was performed using the BWA-GATK germline variant pipeline, and loci were filtered with BCFtools v1.19 and PLINK v2.0 (Purcell et al., 2007; Chang et al., 2015) (https://www.cog-genomics.org/plink/2.0/). Our

objective was to establish a standardized workflow for developing custom SNP panels for genotyping by leveraging existing software and informatics suites for easy reproducibility across species. While the utilization of feces has been the subject of many previous genotyping studies, our investigation represents, to the best of our knowledge, the first exploration of MinION technology for this purpose. Despite historic challenges in calling genotypes from fecal samples, we sought to overcome this limitation by capitalizing on the high-throughput capabilities of the MinION. Additionally, we attempted to use MinKNOW's adaptive sampling feature to maximize *Ateles geoffroyi* reads, aiming to evaluate its efficacy and potential for enriching host DNA using a standard laptop without external computational resources. Lastly, we describe an optimized fecal extraction method that enables successful PCR amplification and accurate detection with Sanger sequencing.

Chapter II - Methods

Ethics Statement

This study adhered to the American Society of Primatologists (ASP) Principles for the Ethical Treatment of Non-Human Primates and complied with the Institutional Animal Care and Use Committee (IACUC) at Kent State University (T 121 RT 21-02).

Study Site & Subjects

Ateles geoffroyi fecal samples were collected from Nashville Zoo (Nashville, TN) and Jungle Friends Primate Sanctuary (JFPS) (Gainesville, FL). We extracted and sequenced the DNA of 5 individuals from Nashville Zoo and 6 individuals from JFPS (N=11; 8 female, 3 male), all of which are included in our genotyping dataset. Individuals' ages range from 1.5-36 years (mean of 22.4 ± 10.46).

Sample Collection

To collect fecal samples from *Ateles geoffroyi* individuals, we utilized food markers (food-grade dyes and glitter) to differentiate amongst samples. All dyes and glitters used are FDA-approved and considered safe for human consumption. Keepers collected approximately 5 g of feces from each individual between 8-10 am, which were then stored in Ziplock plastic bags labeled with the individual ID, date, and time. The samples were immediately frozen at -20°C. All samples were shipped on dry ice to Kent State University, where they were stored at -20°C.

Primer Design

We used NCBI (National Center for Biotechnology Information) reference sequence accession NC_064194.1 (complete mitochondrial genome for *Ateles geoffroyi*, 16,563 bp) to design our primers. Using NCBI Primer-BLAST (Basic Local Alignment Search Tool) with default parameters, we specified a PCR product of approximately 500 bp. We blasted potential primer pairs and PCR product sequences through the NCBI Nucleotide Blast (BLASTN) database, selecting primers with the least amount of coverage and percentage identity with other organisms (e.g., bacteria, insects, fruits) which might confound our results. Since only a fraction of fecal DNA is from the host, we selected 5 primers that were the most unique to *Ateles geoffroyi* (NFE12.mgr01, ATP6, COX2, ND1, and NFE12.mgt22). We purchased all primers from Integrated DNA Technologies (IDT) (Coralville, IA). Genes APT6, COX2, and ND1 code for protein products, while NFE12.mgr01 and NFE12.mgt22 encode for RNA products (ribosomal and transfer RNA, respectively). All primers have an annealing temperature of approximately 57°C. Of the five primers, primers 5 (only worked for E.Z.N.A), 8, and 24 worked effectively for Sanger sequence validation (see Tables 3 and 4).

Table 1. Characterization of five *Ateles geoffroyi* primers for PCR amplification

Primer #	Gene(s)	Oligonucleotide Primer Sequence	Length	Type
5	NFE12.mgr01	F* 5'-GACCTATCCGTGAAGAGGCG-3'	494	rRNA
		R 3'-TTGATCTGTGAGGGCGCTTT-5'		
8	ATP6, COX2	F* 5'-GACCAGGCCTGTTTTACGGA-3'	488	Protein
		R 3'-AGGGAGGAGAGACGATTGCT-5'		
24	ND1	F* 5'-TCCGGATGAGCATCCAACTC-3'	497	Protein
		R 3'-TGGTCATAGCGGAATCGAGG-5'		
25	NFE12.mgt22	F* 5'-AACCGTACATAGCAACGCCA-3'	487	tRNA
		R 3'-CGCGATGACAGCATAAAGCC-5'		
35	ATP6, COX2	F* 5'-ATGCGACCAGGCCTGTTTTA-3'	490	Protein
		R 3'-GGAGGAGAGACGATTGCTGG-5'		

We selected five unique mitochondrial regions of *Ateles geoffroyi* for amplification. The presence of *Ateles geoffroyi* DNA was validated using Sanger sequencing of primers 5, 8, and 24. F* indicates the forward strand, while R indicates the reverse strand. Length refers to the size of the PCR product.

DNA Extraction with the E.Z.N.A DNA Stool Kit

We extracted DNA from fecal samples 1-11 using the E.Z.N.A Stool DNA Kit (hereafter referred to as E.Z.N.A) following the human extraction protocol as per the manufacturer's instructions. To maximize DNA yields, we performed two rounds of elution with 50 uL each, instead of the suggested single elution with 100 uL. After extraction, we used a spectrophotometer (BioTek, Synergy H1 Hybrid Reader) to measure DNA quality (260, 280, 320, and 260/280 ratio (nm units)) and yields (ng/uL).

DNA Extraction with the optimized phenol chloroform extraction method

We utilized an optimized PCEM protocol to extract DNA from fecal sample 12 (see results and Figure 4). We experimented with various extraction conditions to tailor the protocol for extraction with *Ateles geoffroyi* feces. Starting with a baseline PCEM protocol (200 mg feces, 2 uL Proteinase K, 24 hour incubation, 1 phenol chloroform purification step), we experimented

with four primary conditions to improve DNA quality and concentrations suitable for sequencing with the MinION: feces and Proteinase K amounts, incubation time, and number of purification steps (see Figure 1). Samples 4 (extracted using E.Z.N.A) and 12 are derived the same fecal sample collected from individual AG0004. We eluted sample 12 in 150 uL of DEPC.

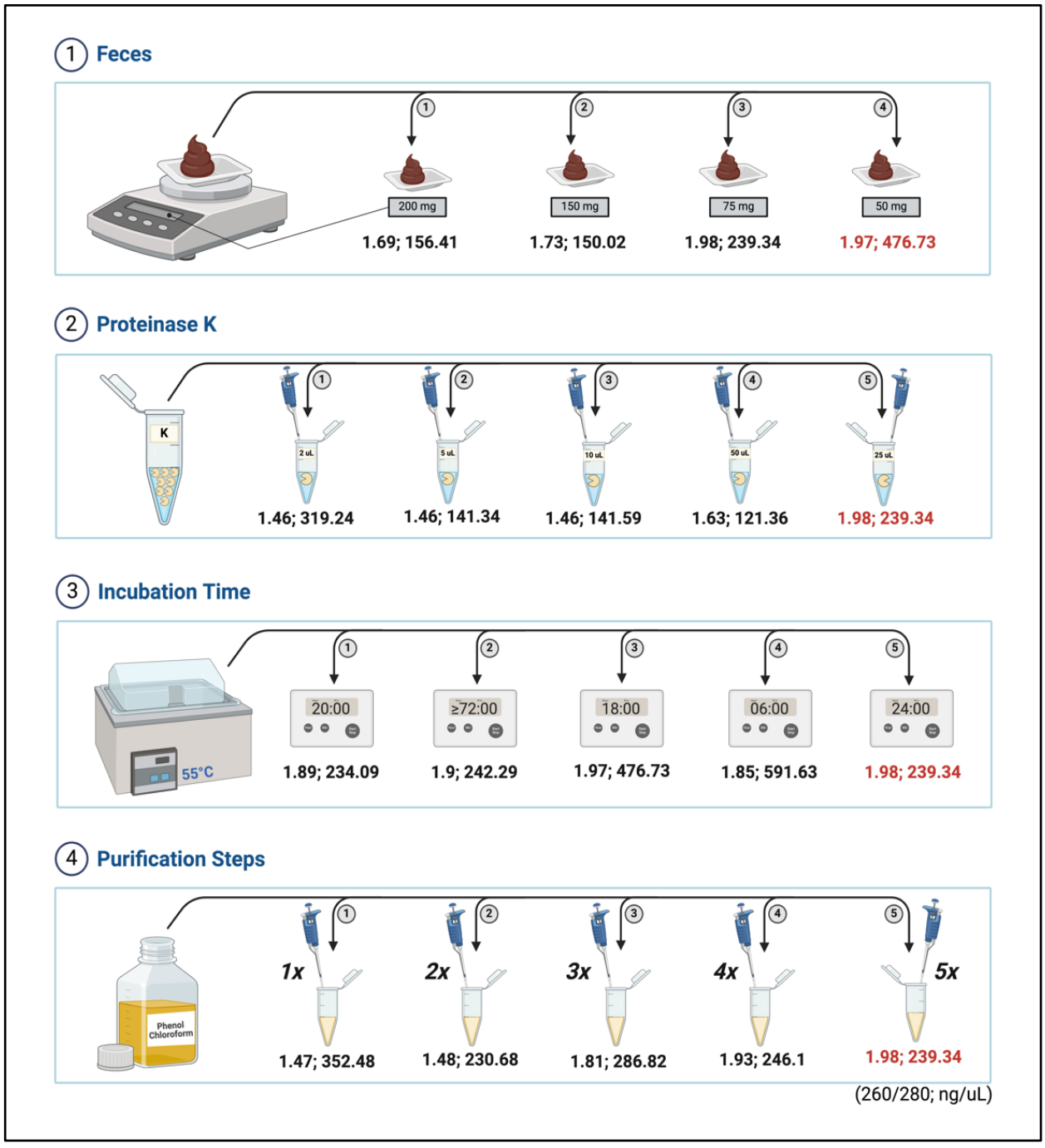

Figure 1. Optimization of the phenol chloroform extraction method (Adapted from BioRender, 2021)

Table 2. DNA yields for *Ateles geoffroyi* feces and *Macaca mulatta* tissue

Sample #	ID	Method	260/280 (nm)	ng/uL
Feces1	AG0002	EZNA	1.97	181.519
Feces2	AG0003	EZNA	1.946	94.294
Feces3	AG0005	EZNA	1.959	92.02
Feces4	AG0004	EZNA	2.015	282.533
Feces5	AG0001	EZNA	1.892	195.782
Feces6	AG0006	EZNA	2.011	195.846
Feces7	AG0010	EZNA	1.959	47.944
Feces8	AG0009	EZNA	2.038	116.717
Feces9	AG0012	EZNA	1.933	127.984
Feces10	AG0011	EZNA	1.928	72.372
Feces11	AG0008	EZNA	1.969	139.187
Feces12	AG0004	PCEM	1.92	346.652
Tissue1	MM0001	MGPK	2.063	78.536
Tissue2	MM0002	MGPK	2.099	51.151

We extracted DNA from fecal samples 1-11 using the E.Z.N.A Stool DNA Kit, fecal sample 12 with an optimized PCEM, and tissue samples 1-2 with the Monarch Genomic DNA Purification Kit (MGPK).

PCR

We conducted two rounds of PCR using each of the five primer sets to amplify DNA extracted from both E.Z.N.A and PCEM conditions. We diluted each primer to a concentration of 10 pmol and then mixed the complementary 5' (forward) and 3' (reverse) primers together with 80 uL of DEPC. We then prepared a master mix, incorporating the following components per sample: 0.5 uL primer (IDT), 0.5 uL dNTP (New England BioLabs), 0.125 uL BioReady rTaq (BIOER; Hangzhou, China), 2.5 uL 10x reaction buffer (BIOER), and 19.375 uL DEPC (Sigma-Aldrich; Burlington, MA). After adding 23 uL of the master mix solution into labeled 0.2 mL PCR tubes, we added 1 uL of DNA to the respective tubes (except negative control), and centrifuged (Denville Mini Mouse C1301, acquired by Thomas Scientific; Swedesboro, NJ) them briefly. We then placed the PCR tubes into a thermal cycler (Bio Rad, T100; Ann Arbor,

MI) and programmed the following cycle: (i) 5 minutes at 94°C, (ii) repeat cycle 35 times at 94°C for 30 seconds, 60°C for 45 seconds, and 72°C for 1 minute, (iii) 72°C for 7 minutes, and (iv) 8°C for ∞. After completing the initial cycle, we performed a second round of PCR (hereafter referred to as PCR2) as a quality control measure. For PCR2, we added 1 uL of the PCR product instead of the fecal DNA to its respective PCR tube. We stored the remaining PCR products from the initial batch (hereafter referred to as PCR) in the refrigerator at 4°C until PCR2 was completed.

Gel Electrophoresis

To prepare the agarose gel, we mixed 0.3 g of agarose powder (MidSci, General Purpose Bulls Eye Agarose GP2; Fenton, MO) with 30 mL of 1x TAE Buffer. We then microwaved the solution for 30 seconds and added 1 uL of 95% ethidium bromide (Thermo Fisher Scientific; Waltham, MA). After the gel hardened, we placed the gel onto a gel electrophoresis cast (Thermo Fisher Scientific, EasyCast B1 Mini) and added 2.5 uL ladder (GoldBio, 100 bp PLUS DNA Ladder; Saint Louis, MO). We then added a mixture of 1 uL of blue loading dye (NEB) and 10 uL of the PCR product to its corresponding well. We set the electrophoresis device (ENDURO, E0303 300V Power Supply; Oakland, CA) to 110 V for 21 minutes, which separated the fragments by size. After gel electrophoresis was complete, we removed the gel from the cast and placed it under UV light (UVP, Benchtop Variable 115 V Transilluminator with PhotoDoc-It 60 Imaging System; Upland, CA) for imaging. We also conducted electrophoresis without PCR by directly applying DNA samples to the gel to visualize their relative molecular weights (see figure 5).

Gel Extraction & Purification

After inspecting the bands relative to the ladder to confirm the approximate lengths of the PCR products compared to the anticipated PCR product size, we excised and purified the bands using a different UV machine (Spectroline Bi-O-Vision, TD-1000R UV/White Light Fixed-Intensity Transilluminator; Melville, NY) using the Monarch DNA Gel Extraction Kit (NEB) following the manufacturer's instructions. We performed two elution rounds using 10 uL of DEPC instead of using the provided Elution Buffer, which was recommended by the sequencing companies Eurofins Genomics (Louisville, KY) and Ohio State University's (OSU) Comprehensive Cancer Center (Columbus, OH). We prepared 10 uL of the purified PCR product (5 uL PCR product diluted to ~25 ng and 5 uL of 10 pmol primer mix) for Eurofins Genomics and 12 uL of the purified PCR product (6 uL PCR product diluted to ~25 ng and 6 uL of 10 pmol primer mix) for OSU to perform Sanger sequencing.

Sanger Sequencing

We sent the purified PCR products to Eurofins Genomics and OSU for Sanger sequencing. The data generated included .seq files containing the nucleotide results in FASTA format, along with .ab1 files containing the electropherogram data. We visualized the sequences using Geneious Prime (https://www.geneious.com/) and analyzed them using NCBI's BLAST database (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Specifically, we selected the BLASTN option and utilized the 'Align Two or More Sequences' feature to compare the sequences against the reference genome. We also compared the PCR and PCR2 conditions for each gene to assess the degree of difference as a quality control measure, as the sequences are expected to be identical.

MinION Sequencing

We performed whole-genome sequencing using the MinION MK1B (see Figure 2). To sequence multiple samples simultaneously, we utilized the Native Barcoding Kit 24 V14 (SQK-NBD114.24). Sequencing was conducted on a R10.4 flow cell (FLO-MIN114), which upon delivery had 1,471 active pores. We set parameters in MinKNOW to include fragments that were > 200 bp with a minimum PQS threshold of 8. We assigned each sample an identifiable barcode.

To prepare the library, we utilized ~400 ng of DNA per sample, following the manufacturer's protocol for sequencing ≥ 4 samples simultaneously. First, we performed the DNA Repair & End-Prep stage, enzymatically correcting reads lacking a blunt end. Next, we performed the Native Barcode Ligation stage, attaching unique barcode sequences to the fragments to facilitate sample identification during multiplex sequencing. We then attached adapter sequences to the repaired and barcoded reads in the Adapter Ligation & Clean-Up stage, and purified and enriched the library using Short Fragment Buffer (SFB) to remove contaminant molecules, reagents, and unligated adapter sequences. Lastly, we primed and loaded the library (with a final pooled DNA concentration of 418.5 ng) into the MinION for sequencing.

We attempted to use MinKNOW's adaptive sampling feature, selecting the most complete nuclear genome (accession GCA_023783555.1), which is assembled at the contig level (2,723 contigs, 2,683,028,796 bp, N50 29.2 mb). Our initial sequencing attempt was terminated after just 12 minutes. We performed a second MinION run under identical filtering conditions without the adaptive sampling feature enabled. We let the MinION run for 72 hours. Following sequencing completion, we merged the data from both the adaptive and non-adaptive sequencing runs to conduct the analysis.

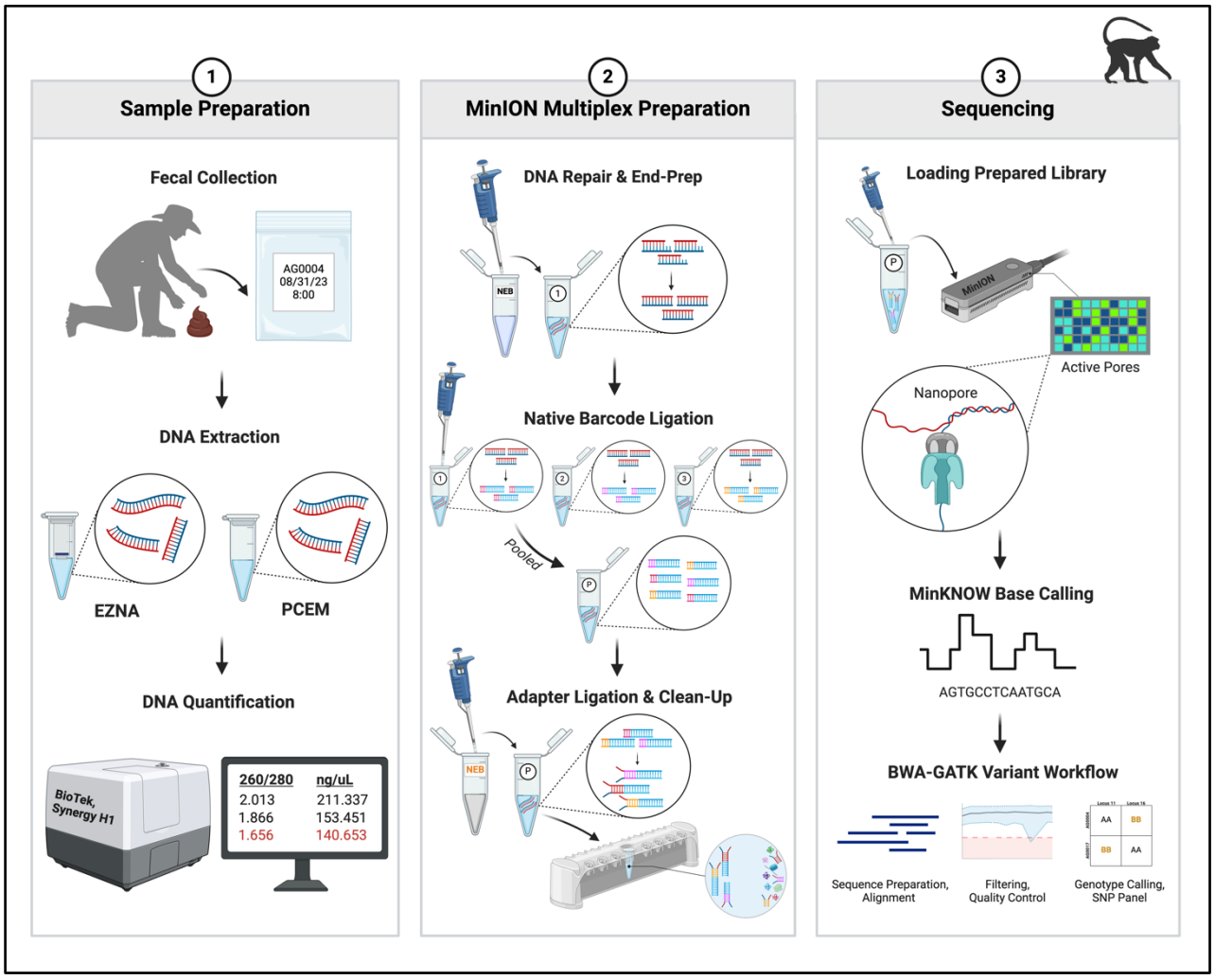

Figure 2. Sample preparation, loading, and sequencing with the MinION (Adapted from BioRender, 2020)

Bioinformatics

All sequencing and bioinformatics steps were performed on a MacBook Pro laptop (Apple M1 Chip; 8 CPUs, 8 GB Memory). The MinKNOW-generated FASTQ sequence files were output onto an external hard drive (LaCie 2 TB). They were output into the directory MinKNOW_Output, and each barcode was separated into a unique folder (e.g., barcode01, barcode02). We transferred these files from the external hard drive to the Ohio Supercomputer Center (Columbus, Ohio) terminal where all of the bioinformatics analytical steps were

completed in a conda environment (v23.7.4). We employed the BWA-GATK germline variant calling pipeline to call SNPs (see figure 3), specifically using BWA-MEM (Maximal Exact Match) capable of aligning longer reads (Li, 2013).

Each sample contained multiple FASTQ sequence files (4,000 read maximum per file) which we concatenated together into a single FASTQ.gz file. We then indexed the *Ateles geoffroyi* reference genome with Burrows-Wheeler Aligner (BWA) v0.7.17-r1188 (https://github.com/lh3/bwa) and created a sequence dictionary file using Picard v2.26.0 (https://github.com/broadinstitute/picard). We then aligned and tagged (barcode specific identification marker for downstream sample merging) the sequences to the *Ateles geoffroyi* nuclear reference genome using BWA which output a Sequent Alignment Map (SAM) file. We then used samtools v1.19 (https://github.com/samtools/samtools), a program that interacts with DNA sequence alignment data in SAM format, to generate a Binary Alignment and Map (BAM) file. We then sorted the BAM file with samtools and marked duplicate reads using Picard. We indexed the sorted and marked BAM file with samtools, readying it for use through the Genome Analysis Toolkit (GATK) v4.3.0.0 (https://github.com/broadinstitute/gatk/releases/tag/4.3.0.0) variant calling pipeline. We used GATK HaplotypeCaller to covert the BAM file into a Genomic Variant Call Format (GVCF) file, a file type containing records for all sites throughout the aligned sequence, regardless if there is a variant present or not. After producing 12 unique GVCF files for all our samples, we used the Gatk CombineGVCFs function to merge them into a single GVCF file. We then ran GATK GenotypeGVCFs which produced a Variant Calling Format (VCF) file, a file type containing only the sites where variants are present. We used GATK SelectVariants to filter out variant that were not biallelic SNPs, and GATK VariantFilatration

(default settings, except we decreased the mapping quality score from quality 40 to quality 30) to filter out reads that did not meet the required quality control parameters.

To further refine the quality of our *Ateles geoffroyi* SNP loci, we employed BCFtools' filtering feature (https://samtools.github.io/bcftools/bcftools.html) to retain only loci with a genotyping call rate of ≥ 0.25 (meaning that 25% or more of the samples had called genotypes), a minor allele frequency (MAF) ≥ 0.2 (20%), and an across sample coverage of ≥ 10x. We used GATK Variant2Table with the options GT (genotype), PL (phred-scaled genotype likelihoods), and AD (allele depth) enabled to manually mark and ignore (not remove) genotypes where the PQS likelihood score for the called genotype was equal to the PQS likelihood of the second most probable genotype (e.g., 0,0,9), indicating an ambiguous genotype call where GATK was unable to differentiate between the homozygous and heterozygous condition. We then used Plink to calculate the observed and expected heterozygosity (under HWE) scores across loci and manually performed chi-squared tests [(observed – expected)^2/expected] with one degree of freedom to determine whether loci significantly deviated at $p < 0.05$. To account for multiple testing, we used the method described by Benjamini-Hochberg with a false discovery rate of 5% to correct the p-values (Benjamini & Hochberg, 1995; Benjamini et al., 2009). Additionally, we used Plink (with the option r2-phased enabled) to calculate the linkage disequilibrium (LD) between loci to investigate whether different loci on the same chromosome exhibit random or non-random associations by calculating the r^2 statistic for every variant pair. We used the LD statistic not as a filtering criteria, but rather to provide additional information about the characteristics of the SNPs. We used VCFtools genotype012 feature v0.1.16 (https://vcftools.sourceforge.net/man_latest.html) to visualize the called genotypes.

Macaca mulatta Tissue

Serving as a high-quality comparative reference to the fecal condition to demonstrate the MinION's output capacities and the capability for the BWA-GATK workflow to effectively call variants, we independently sequenced DNA from *Macaca mulatta* brain tissue following a nearly identical post-extraction workflow. We obtained the tissue samples from the Wisconsin Primate Research Center (Madison, WI). All research was conducted in compliance with the IACUC at Kent State University (T 117 RT 20-02).

We sectioned 0.1 g of frozen tissue from two *Macaca mulatta* individuals using a cryostat instrument (Leica CM1950; Wetzlar, Germany), and extracted the DNA using the Monarch Genomic DNA Purification Kit (NEB) following the manufacturer's instructions. The DNA was eluted in 10 uL of elution buffer. Library preparation and sequencing with the MinION followed an identical protocol to the *Ateles geoffroyi* fecal samples, though we did not utilize the adaptive sampling feature. The flow cell had 1,411 active pores upon delivery. Following the DNA Repair & End-Prep, Native Barcode Ligation, and Adapter Ligation & Clean-Up steps, we loaded the final pooled DNA concentration of 420 ng into the MinION for sequencing. After sequencing completion, we followed the identical BWA-GATK workflow described previously, though with more stringent filtering parameters. We utilized the most complete *Macaca mulatta* nuclear reference genome (accension GCA_003339765.3; 22 chromosomes, 2,971,314,966 bp, scaffold N50 82.3 Mb), which is assembled at the chromosome level, for alignment. Following the quality filtering of variants, we applied a more stringent set of filtering criteria to establish the genotyping panel.

We used BCFtools' filtering feature to retain only loci with a genotyping call rate of 1 (100%), a MAF ≥ 0.2 (20%), and an across sample coverage of ≥ 10x. We used GATK Variant2Table to manually remove loci where a single genotype had a 1x coverage, the second most likely genotype likelihood score was < 10 PQS units away from the called genotype, and genotype calls were the same for both samples. We used Plink to calculate the observed and expected heterozygosity (under HWE) scores across loci and manually performed chi-squared tests with one degree of freedom at $p < 0.05$. We corrected p-values using the Benjamini-Hochberg technique with a false discovery rate of 5%. We used Plink again to calculate the r^2 statistic for variant pairs, though LD statistics were not used as a filtering criteria. Finally, we utilized VCFtools genotype012 feature to visualize the called genotypes.

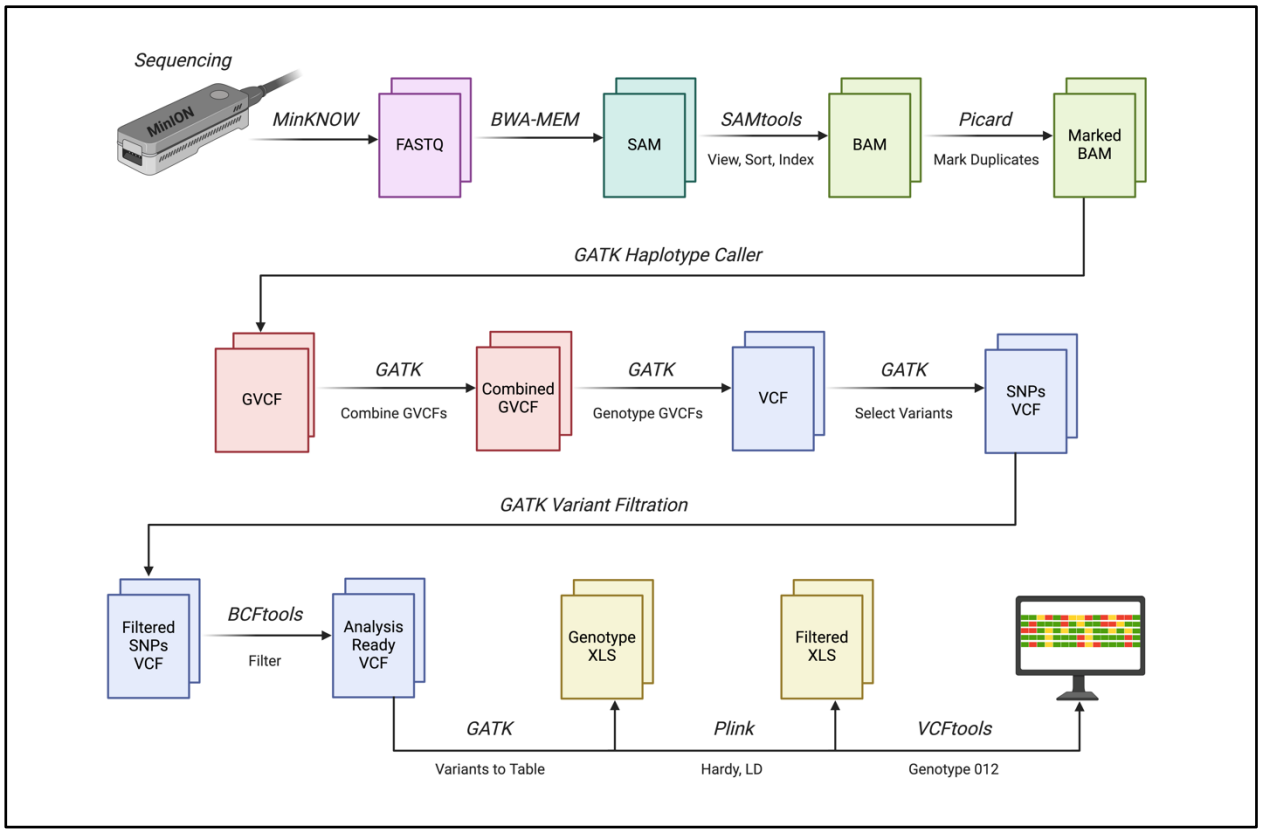

Figure 3. Genotyping with the BWA-GATK variant calling workflow (Adapted from BioRender, 2020)

Additional Tools

We used ChatGPT (v3.5) to help resolve coding errors encountered during bioinformatics steps. Additionally, we used it to identify spelling and grammatical errors in our written text, and for suggestions in improving syntax within paragraphs.

Chapter III - Results

Phenol Chloroform Extraction Method

We outlined an optimized extraction method for DNA isolation from *Ateles geoffroyi* feces, documenting improvements in DNA quality and yields by testing various experimental conditions, including feces and Proteinase K amounts, incubation time, and purification steps (see Figure 1 and Appendix A).

We examined the quantity of feces used to assess its direct impact on total yields. Surprisingly, we found that the amount of feces played a more significant role in improving overall quality. We tested four fecal weights ranging from 50-200 mg. We observed an improvement in quality scores as we decreased the amount of feces. After a 92-hour incubation, the 75 mg fecal condition yielded a much higher 260/280 quality score (1.9 nm) compared to that at 150 mg (1.63 nm). Similarly, significant improvements in 260/280 quality scores were observed during the 20-hour and 72-hour incubations, with results of 1.89 nm (75 mg feces) versus 1.73 nm (150 mg feces) and 1.85 nm (75 mg feces) compared to 1.59 nm (150 mg feces), respectively. The same is true when comparing the 50 mg and 75 mg conditions. The 50 mg condition yielded 476.73 ng/uL with a 260/280 ratio of 1.97 nm, while the 75 mg sample yielded 399.4 ng/uL with a 260/260 ratio of 1.93 nm. Though the results are similar, the 50 mg sample's yields and quality were marginally higher. Furthermore, these improvements tended to result in higher overall yields, although not for all fecal samples. For example, for the 92-hour incubation, the 75 mg feces condition yielded 242.29 ng/uL compared to 140.65 ng/uL at 150 mg.

We experimented with various amounts of Proteinase K (ranging from 2-50 uL) and different incubation times (6-92 hours) to understand the effects of these conditions specifically on quality. Since Proteinase K is responsible for breaking down protein molecules, and different

incubation times allow for differential amounts of enzymatic action and potential, we investigated whether increasing the amount of Proteinase K and incubation time would improve quality scores as a result of increased protein digestion. Even at the long incubation periods of 72 and 92 hours, we observed high quality 260/280 ratio scores (1.85 nm and 1.9 nm, respectively). When we reduced the incubation time to 6-12 hours (and one time at 18 hours), the 320 ratios increased, indicative of sample contamination. We increased the amount of Proteinase K to 25 uL, a similar quantity to some available commercial kits (e.g., E.Z.N.A uses 20 uL), and compared this condition against 50 uL. We did not observe noticeable improvements by increasing the amount to 50 uL. In fact, we noticed a reduction in both quality (1.63 nm at 50 uL compared to 1.69 nm at 25 uL) and yields (121.364 ng/uL at 50 uL compared to 156.406 ng/uL at 25 uL).

We also tested numerous purification combinations with phenol/chloroform/isoamyl (PCI) and chloroform. Beginning with one PCI stage, we ultimately increased the total number of purification steps to five, consisting of 4 PCI rounds and 1 chloroform round (though 3 PCI and 2 chloroform stages achieves similar results). Starting at baseline with 1 PCI stage which yielded a 1.46 nm 260/280 ratio, the quality improved with additional purification steps. We obtained a 1.97 nm 260/280 ratio when experimenting with 4 PCI and 1 chloroform stages. We achieved the highest 260/280 quality with 3 PCI and 2 chloroform stages (1.98 nm), though since the results were similar, we opted to continue with 4 PCI and 1 chloroform for ease of replication. When conducting experiments with chloroform as the sole purification solution, we observed a reduction in overall quality. For example, with exclusively 5 rounds of chloroform, we obtained a 260/280 ratio of 1.76 nm.

We tested multiple ethanol wash stages, yet no noticeable improvements in quality were observed with the addition of wash steps. Sometimes we observed decreases in either or both quality and yields. For example, for one ethanol wash stage (with 75 mg feces, 3 PCI and 1 chloroform stages, and a 24 hour incubation time) we achieved a 260/280 ratio of 1.98 nm with yields of 239.34 ng/uL, while two stages yielded a 1.86 nm 260/280 ratio and 255.46 ng/uL. Though the yields increased slightly with additional wash steps, the quality significantly decreased. For another extraction (200 mg feces, 2 PCI stages, and a 24 hour incubation time), one ethanol wash stage resulted in a 260/280 ratio of 1.48 nm with yields of 230.68 ng/uL, with two stages exhibiting a 260/280 ratio of 1.43 nm with yields of 119.62 ng/uL.

Optimized Phenol Chloroform Extraction Protocol

We weighed out 50 mg frozen feces and added 700 uL Lysis Buffer (0.1 M Tris-HCl pH 8.5, 0.005 M Ethylenediaminetetraacetic acid (EDTA), 0.2% Sodium dodecyl sulfate (SDS), and 0.2 M Sodium chloride (NaCl)) and 25 uL of Proteinase K (New England Biolabs; Ipswich, MA) to a 2 mL tube. We then manually homogenized the sample and incubated at 55° for 18-24 hours. Following incubation, we added 700 uL Phenol/Chloroform/Isoamyl (Acros Organics; Geel, Belgium) (25:24:1, pH 8.0) to the Lysis Buffer at a 1:1 ratio, and manually homogenized the sample before 5 minutes of centrifugation (Eppendorf 5424 R; Hamburg, Germany) at 13,000 rpm (4°C). We transferred the aqueous solution to a new 2 mL tube and repeated the previous step for three additional Phenol/Chloroform/Isoamyl stages. We then added a 1:1 ratio of Chloroform to the aqueous solution from the previous Phenol/Chloroform/Isoamyl stage. We homogenized the sample before another 5 minute centrifugation at 13,000 rpm (4°C). We then transferred the aqueous solution to a new 2 mL tube and added 0.1 volume of 3 M Sodium

Acetate (pH 5.2) and 2.5 volumes of -20°C 100% Ethanol, shaking vigorously to ensure DNA precipitation. We stored the sample at -80°C for 60 minutes. After the brief incubation period, we centrifuged the sample for 20 minutes at 13,000 rpm (4°C) and poured off the supernatant, keeping the DNA pellet intact. We added 500 uL of cold 75% Ethanol, followed by 5 minutes of centrifugation at 13,000 rpm (4°C). We then poured off the supernatant again, letting the DNA pellet air dry for 10 minutes at room temperature. We dissolved the DNA pellet in 150 uL DEPC. After extraction, we used a spectrophotometer to measure DNA quality and yields. We stored samples at -20°C.

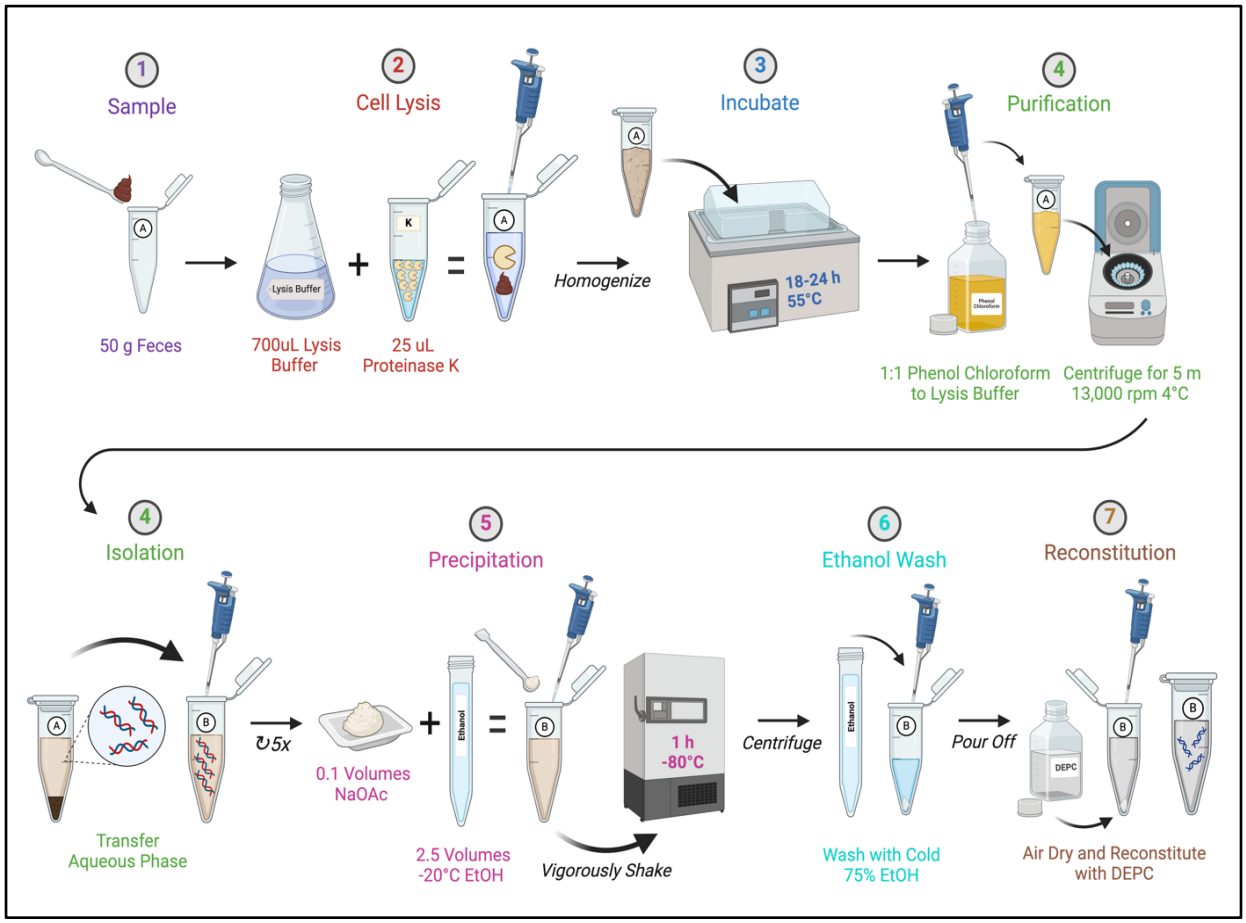

Figure 4. Optimized phenol chloroform extraction method protocol (Created with BioRender.com)

Sanger Sequencing

We analyzed the data generated from NCBI's BLASTN feature to calculate coverage (i.e., percentage of the query region that aligns to the reference) and identity (i.e., percentage of identical bases within the aligned query region compared to the reference genome) statistics. For the E.Z.N.A sample (for both PCR and PCR2 conditions), the degrees of coverage ranged from 34.0% to 94.5% (with a mean of 85.3% ± 21.16), with identity statistics ranging from 82.4% to 98.4% (with a mean of 94.2% ± 4.69). Comparing PCR to PCR2, the coverage averages ranged from 66.0% to 95.4% (mean of 82.7% ± 11.79), with identity percentages ranging from 88.1% to 99.2% (mean of 92.2% ± 4.75). We observed the lowest coverage statistic (mean of 48.7%) for the NFE12.mgr01 gene, with the OSU PCR sample failing to provide any overlap data against the reference. NFE12.mgr01's PCR2 condition for both OSU and Eurofins Genomics produced longer reads than the expected ~500 bp (1,125 bp and 710 bp, respectively).

Table 3. E.Z.N.A Sanger sequencing: sample vs. reference (top), PCR vs. PCR2 (bottom)

Primer #	Gene(s)	Condition	Company	S_Length	Q_Length	Coverage	Identity	Mismatches
5	NFE12.mgr01	PCR	Eurofins	465	354	76.1	93.2	21
5	NFE12.mgr01	PCR2	Eurofins	710	243	34.0	91.0	20
5	NFE12.mgr01	PCR	OSU	452	NA	NA	NA	NA
5	NFE12.mgr01	PCR2	OSU	1125	421	36.0	82.4	62
8	ATP6, COX2	PCR	Eurofins	656	375	57.0	91.8	28
8	ATP6, COX2	PCR2	Eurofins	461	374	81.3	97.6	7
8	ATP6, COX2	PCR	OSU	451	375	82.9	95.7	14
8	ATP6, COX2	PCR2	OSU	456	433	94.5	94.7	20
24	ND1	PCR	Eurofins	474	368	77.8	98.4	6
24	ND1	PCR2	Eurofins	483	425	88.0	96.3	14
24	ND1	PCR	OSU	457	362	79.4	98.1	6
24	ND1	PCR2	OSU	463	432	93.1	97.7	8

Primer #	Gene(s)	Company	Length	Coverage (%)	Identity (%)	Mismatches
5	NFE12.mgr01	Eurofins	311	66.0	88.1	37
5	NFE12.mgr01	OSU	337	73.0	86.4	34
8	ATP6, COX2	Eurofins	433	93.9	90.7	34
8	ATP6, COX2	OSU	403	88.0	95.3	15
24	ND1	Eurofins	465	95.4	93.4	24
24	ND1	OSU	366	80.0	99.2	2

S_Length (sequence length), the total PCR product size in bp; Q_Length (query length), the number of bases that align to the reference; Coverage, the percentage of Q_Length that covers S_Length (Q_Length/S_Length); Identity, the number of bases in Q_Length that match to the reference; Mismatches, represents the number of bases in Q_Length that do not match the reference.

For the PCEM-extracted sample (sample 12), we excluded the NFE12.mgr01 gene from our data set because the negative control resulted in a positive reading at 500 bp relative to the ladder. The degrees of coverage ranged from 34.8% to 92.7% (mean of 72.63% ± 20.33) with identity averages ranging from 84.42% to 99.28% (mean of 93.54% ± 5.24). Comparing PCR to PCR2, the coverage statistics ranged from 37.4% to 89.2% (mean of 63.73% ± 21.19) with identity percentages ranging from 84.07% to 100% (mean of 91.27% ± 7.79).

Table 4. PCEM Sanger sequencing: sample vs. reference (top), PCR vs. PCR2 (bottom)

Primer #	Gene(s)	Condition	Company	S_Length	Q_Length	Coverage	Identity	Mismatches
8	ATP6, COX2	PCR	Eurofins	658	229	34.8	86.46	28
8	ATP6, COX2	PCR2	Eurofins	520	335	64.4	94.33	12
8	ATP6, COX2	PCR	OSU	450	276	61.3	84.42	42
8	ATP6, COX2	PCR2	OSU	453	384	84.8	96.35	12
24	ND1	PCR	Eurofins	445	279	62.7	99.28	2
24	ND1	PCR2	Eurofins	469	429	91.5	96.74	11
24	ND1	PCR	OSU	463	411	88.8	94.89	18
24	ND1	PCR2	OSU	466	432	92.7	95.83	17

Primer #	Gene(s)	Company	Length	Coverage (%)	Identity (%)	Mismatches
8	ATP6, COX2	Eurofins	246	37.4	85.37	33
8	ATP6, COX2	OSU	295	65.6	84.07	46
24	ND1	Eurofins	279	62.7	100	0
24	ND1	OSU	413	89.2	95.64	16

S_Length (sequence length), the total PCR product size in bp; Q_Length (query length), the number of bases that align to the reference; Coverage, the percentage of Q_Length that covers S_Length (Q_Length/S_Length); Identity, the number of bases in Q_Length that match to the reference; Mismatches, represents the number of bases in Q_Length that do not match the reference.

Ateles geoffroyi MinION Sequencing & Bioinformatics

The merged *Ateles geoffroyi* MinION runs yielded 14.85 million reads, totaling 12.01 Gb, with 9.94 Gb meeting the filtering criteria. Among these, 88 thousand reads and 64 million passed bases were attributed to the adaptive sequencing run. The passed reads exhibited an N50 value of 864 bp. Regarding sequencing read lengths at a PQS of ≥ 20, the longest for the E.Z.N.A samples ranged from 30,011 bp to 86,938 bp (mean length of 44,486 ± 16,354 bp), while PCEM's measured 32,804 bp. Additionally, the average read lengths for the E.Z.N.A samples varied from 405 bp to 614 bp (mean of 486 ± 58 bp), while PCEM's was 513 bp.

The number of reads across the samples varied from 202,821 to 2,485,018 reads (mean of 983,339 ± 628,978). Their base counts ranged from 179,311,064 to 1,382,179,277 bp (mean of 726,946,647 ± 392,382,550). On average, approximately 24.8% or 199,366 ± 174,256 reads

(with a maximum read count of 602,249 and a minimum read count of 6,994) mapped to the *Ateles geoffroyi* reference genome at a PQS of ≥ 20. 17.67% of the total bases (with a maximum percentage of 37.64 and a minimum percentage of 2.37) mapped to the reference.

Table 5. BAM primary reads statistics

Sample	$Reads_t$	$Reads_m$	$Reads_m$ (%)	$Bases_t$	$Bases_m$	$Bases_m$ (%)	Positions (bp)	Coverage (x)
Feces1	202,821	24,690	12.17	179,311,064	12,488,542	6.96	7,204,399	0.987
Feces2	935,784	254,592	27.21	707,576,178	144,164,506	20.37	98,872,291	1.01
Feces3	626,885	291040	46.43	421,226,978	130,496,805	30.98	82,424,806	1.007
Feces4	770,841	43,881	5.69	694,594,694	24,605,191	3.54	12,627,959	0.99
Feces5	1,550,874	169,970	10.96	1,168,665,100	94,185,211	8.06	57,671,153	1.001
Feces6	629,782	251,122	39.87	498,023,397	161,882,776	32.51	117,817,974	1.02
Feces7	1,470,798	663,389	45.1	1,135,146,370	330,749,562	29.14	221,941,286	1.036
Feces8	302,772	8,720	2.88	294,779,700	6,977,455	2.37	2,242,104	0.994
Feces9	1,235,440	124,294	10.06	1,151,438,450	57,040,552	4.95	29,639,897	0.994
Feces10	671,461	311,620	46.41	439,574,892	165,447,386	37.64	113,551,468	1.013
Feces11	917,587	437,443	47.67	650,843,666	209,782,131	32.23	136,341,961	1.019
Feces12	2,485,018	82,897	3.34	1,382,179,277	45,367,152	3.28	19,816,671	0.992
Tissue1	158,498	158,389	99.93	1,026,809,226	1,026,741,003	99.99	895,093,041	1.277
Tissue2	233,381	232,423	99.59	1,643,395,949	1,642,714,264	99.96	1,433,436,363	1.42

$Reads_t$, number of total reads; $Reads_m$, number of mapped reads; $Reads_m$ (%), the percentage of reads that mapped to the reference ($Reads_m$/$Reads_t$); $Bases_t$, number of total bases in $Reads_t$; $Bases_m$, number of total bases in $Reads_m$ (number of mapped bases); $Bases_m$ (%), the percentage of bases that mapped to the reference ($Bases_m$/$Bases_t$); Positions (bp), number of unique genomic positions sequenced; coverage (x), number of bases covering each position that has sequencing data. Ateles geoffroyi reference genome size: 2,683,028,796 bp; *Macaca mulatta*'s 2,971,314,966 bp.

The percentage of the genome covered by the sequenced reads (calculated by dividing the total number of mapped nucleotide positions by the total number of bases in the reference genome) varied across samples from 0.08% to 8.27% (mean of 2.8% ± 2.47). This statistic indicates that, on average, each sample covered about 3% of the reference genome, leaving approximately 97% of the genome without coverage. The coverage (or depth) statistic for the sequenced regions exhibited relatively consistent values across samples, ranging from 0.987x to 1.019x (mean of 1.01x ± 0.015x). This suggests that, on average, each nucleotide position within the mapped regions is covered by approximately one sequencing read.

After executing the GATK bioinformatics pipeline with the specified filtering criteria (see methods), we retained 5,934 SNP sites. After filtering loci with a ≥ 0.25 genotype call rate, ≥ 0.2 MAF, and a 10x across sample coverage, only 29 loci remained. After manually removing loci without any heterozygous genotypes, 4 loci remained. No loci deviated significantly from the expected heterozygosity estimated under HWE. After filtering, samples 1, 4, 8, 9 (ignored the genotype due to likelihood ambiguity), and 12 have no remaining called genotypes. Evaluating the missingness statistics (see Appendix B), which are indicative of the number and percentage of loci where genotype calls are absent for a given sample, we observe a range of 7 to 28 missing loci (mean of 18.4 ± 7.2) for the initially filtered 29 loci, and a range of 0 to 4 missing loci (mean of 2.25 ± 1.5) for the four remaining filtered sites. This suggests that, on average, each sample lacks genotype data for approximately 18 out of the 29 loci analyzed. Among the remaining loci, on average samples lack data for approximately 2.25 of the 4 loci.

Table 6. *Ateles geoffroyi* filtering steps & loci retention rates (top), SNP viability (bottom)

Filtering Steps	SNPs (#)
Total SNPs	5,934
Genotype Call Rate ≥ 0.25	234
Minor Allele Frequency ≥ 0.2	149
Across sample coverage ≥ 10	29
No heterozygous individuals	4
Hardy-Weinberg Equilibrium p-value > 0.05	4

Chromosome	Position	N_g	N_a	H_o	H_e	P_{adj}	Significance
JAKFHY010001050.1	33683	4	2	0.2	0.42	0.462	ns
JAKFHY010001227.1	18991	3	2	0.25	0.47	0.872	ns
JAKFHY010001335.1	56547	7	2	0.75	0.47	0.078	ns
JAKFHY010001909.1	267	4	2	0.25	0.47	0.581	ns

N_g, number of individuals with called genotypes; N_a, number of alleles; H_o, observed heterozygosity; H_e, expected heterozygosity under Hardy-Weinberg Equilibrium; P_{adj}, corrected p-value; ns, non-significant (H_o does not significantly deviate from H_e at $p < 0.05$). There is no evidence for linkage between loci.

Macaca mulatta MinION Sequencing and Bioinformatics

The *Macaca mulatta* MinION run resulted in 516 thousand reads, totaling 3.21 Gb, with 2.67 Gb meeting the filtering criteria. The passed reads exhibited an N50 value of 20.49 kb. The longest read length for each sample was 191,146 bp and 170,705 bp, respectively (mean length of 180,926 ± 14,453 bp). Additionally, the average read lengths were 5,696 bp and 6,320 bp, respectively (mean of 6,008 ± 441 bp). The per sample number of reads were 158,498 and 233,381 (mean of 195,940 ± 52,950), with base counts of 1,026,809,226 and 1,643,395,949 bp, respectively (mean of 726,946,647 ± 392,382,550). Approximately 195,406 ± 52,349 reads between the two samples (with a maximum read count of 232,423 and a minimum read count of 158,389) mapped to the *Macaca mulatta* reference genome at a PQS of ≥ 20. 99.76% of the reads (with a maximum percentage of 99.93 and a minimum percentage of 99.59) mapped to the reference with 99.98% of the total bases (with a maximum percentage of 99.99 and a minimum percentage of 99.96) mapping.

The percentage of the genome covered by reads was 30.13% and 48.24% (mean of 39.19% ± 12.81), with the number of positions covered being 895,093,041 bp and 1,433,436,363 bp, respectively. This statistic indicates that, on average, the samples covered about 39% of the reference genome, leaving approximately 61% of the genome without coverage. The coverage statistic for the sequenced regions ranged from 1.28x to 1.42x (mean of 1.35x ± 0.101x). This suggests that, on average, each nucleotide position within the mapped regions is covered by approximately 1.35 sequencing reads.

After executing the GATK bioinformatics pipeline with the specified filtering criteria (outlined in the methods), we retained 103,260 SNP loci. After filtering loci with a genotype call

rate of 1, a MAF of 0.2, and a 10x coverage using BCFtools, 473 loci remained. We then manually removed loci that contained one genotype with an allele depth of 1x, which left 221 loci. We filtered out genotypes where the likelihood scores for an alternate call were < 10 PQS from the called genotype, leaving 57 loci. Finally, we manually removed loci where the called genotypes were identical (not informative for comparison), leaving 35 loci. None of these loci's observed heterozygosity scores significantly deviated from the expected under HWE.

Table 7. *Macaca mulatta* filtering steps & loci retention rates (top), SNP viability (bottom)

Filtering Steps	SNPs (#)
Total SNPs	103,260
Genotype Call Rate = 1	14,021
Minor Allele Frequency ≥ 0.2	13,306
Across sample coverage ≥ 10	473
Single genotype with < 2x coverage	221
Alternate genotype likelihood score < 10 PQS from the called genotype	57
Identical genotypes + no heterozygous individuals	35
Hardy-Weinberg Equilibrium p-value > 0.05	35

PQS, Phred Quality Score.

Chromosome	Position	N_g	N_a	H_o	H_e	P_{adj}	Significance
[1]CM014336.1	168399427	2	2	0.5	0.375	1	ns
[1]CM014336.1	168399437	2	2	0.5	0.375	1	ns
CM014342.1	242363	2	2	0.5	0.375	1	ns
[2]CM014343.1	123196	2	2	0.5	0.375	1	ns
[2]CM014343.1	133310	2	2	0.5	0.375	1	ns
[2]CM014343.1	133327	2	2	0.5	0.375	1	ns
[2]CM014343.1	133337	2	2	0.5	0.375	1	ns
CM014343.1	88218380	2	2	0.5	0.375	1	ns
CM014350.1	6460970	2	2	0.5	0.375	1	ns
CM014350.1	6470615	2	2	0.5	0.375	1	ns
CM014355.1	29845231	2	2	0.5	0.375	1	ns
[3]CM014355.1	29845680	2	2	0.5	0.375	1	ns
[3]CM014355.1	29845683	2	2	0.5	0.375	1	ns
[4]CM014356.1	61530920	2	2	0.5	0.375	1	ns
[4]CM014356.1	61530922	2	2	0.5	0.375	1	ns
[5]QNVO02000865.1	10290	2	2	0.5	0.375	1	ns
[5]QNVO02000865.1	10300	2	2	0.5	0.375	1	ns
QNVO02000909.1	18628	2	2	0.5	0.375	1	ns
QNVO02000978.1	14806	2	2	0.5	0.375	1	ns
QNVO02001305.1	9456	2	2	0.5	0.375	1	ns
[6]QNVO02001640.1	8297	2	2	0.5	0.375	1	ns
QNVO02001640.1	30758	2	2	0.5	0.375	1	ns
[6]QNVO02001640.1	30826	2	2	0.5	0.375	1	ns
QNVO02001686.1	8346	2	2	0.5	0.375	1	ns
QNVO02001734.1	9823	2	2	0.5	0.375	1	ns
ML143123.1	422209	2	2	0.5	0.375	1	ns
[7]QNVO02001792.1	33970	2	2	0.5	0.375	1	ns
[7]QNVO02001792.1	46102	2	2	0.5	0.375	1	ns
QNVO02001792.1	50423	2	2	0.5	0.375	1	ns
QNVO02002100.1	69824	2	2	0.5	0.375	0.978	ns
QNVO02002237.1	70289	2	2	0.5	0.375	0.946	ns
QNVO02002242.1	100240	2	2	0.5	0.375	0.917	ns
QNVO02002246.1	102708	2	2	0.5	0.375	0.889	ns
QNVO02002246.1	113220	2	2	0.5	0.375	0.863	ns
QNVO02002909.1	7565	2	2	0.5	0.375	0.838	ns

N_g, number of individuals with called genotypes; N_a, number of alleles; H_o, observed heterozygosity; H_e, expected heterozygosity under Hardy-Weinberg Equilibrium; P_{adj}, corrected p-value; ns, non-significant (H_o does not significantly deviate from H_e at $p < 0.05$); ^numbers; indicate linkage between loci (r^2 value = 1).

Chapter IV - Discussion

Phenol Chloroform Extraction Method

The PCEM is a cost-effective extraction method suitable for adoption in many research laboratories due to the availability of its commonly used reagents. Additionally, its straightforward protocol allows researchers of all experience levels to use it effectively. However, the phenol chloroform purification steps demand diligence to avoid pipetting off the interphase separating the aqueous and organic phases, which can introduce contaminants. This is especially crucial for the fifth and final purification round immediately preceding precipitation. To address this, we opted to use chloroform exclusively for the final purification stage, rather than phenol chloroform in combination, because of its transparent quality, making it easier to pipette off the aqueous phase which is darker in appearance. By leveraging chloroform's color distinction and ability to be readily differentiated, we were able to improve overall DNA quality.

While several conditions led to higher DNA quality and yields, they were not incorporated into the protocol due to considerations of practicality and repeatability. For example, although the results show only a small percentage of improvement between the usage of 75 mg versus 50 mg of feces, we prioritized sample conservation. This decision was driven by the need to design a protocol capable of extracting as much DNA as possible from minimal biological material. Conserving sample is crucial, particularly for field-based and forensic applications where sample availability is limited. Additionally, though the 72 and 92 hour incubations demonstrated potential to be incorporated into the protocol because of their similarly high quality outputs (1.85 and 1.9, respectively) compared to 18-24 hours (1.97 and 1.98, respectively), we did not continue their utilization because of their long incubation periods of upwards of 3-4 days. This decision was based on extraction efficiency, as incubation times of 72

and 92 hours are highly time-consuming and inefficient for project replication. Moreover, Proteinase K can begin to lose its enzymatic effectiveness over time, leading to degradation and self-digestion, which can potentially contaminate the sample. In contrast, incubation periods that are too short (6-12 hours) result in higher 320 scores, which may not allow for sufficient cell lysis and protein digestion.

The optimized PCEM protocol for DNA extraction offers several advantages. This protocol (i) yields DNA with highly quality (260/280) scores which is critical for downstream amplification and sequencing applications, (ii) is relatively simple and straightforward, requiring only basic laboratory training to perform effectively, making extraction more accessible and suitable across a wide range of research settings, and (iii) many of the reagents and materials used are commonly found in molecular laboratories and are relatively inexpensive compared to other extraction methods, such as DNA filtration columns. While there are many advantages to using this extraction method, there are also some disadvantages that should be noted, including (i) the use of harmful and hazardous chemicals, such as phenol which can be absorbed through the skin and can cause severe burns, and chloroform, a skin and eye irritant and potential carcinogen, (ii) is relatively time consuming (18-24 hour 55°C water bath incubation, 60 minute -80°C incubation, multiple centrifugation stages), (iii) generates a lot of waste, primarily tubes and pipette tips, especially if the sample size is large, and (iv) it produces highly fragmented reads, especially compared with samples extracted using the E.Z.N.A Stool DNA Kit. Moreover, even after purification and precipitation stages, peptide fragments may persist because they are similarly (to DNA) non-polar, potentially contaminating the DNA extract. While purification cartridges can provide further refinement, this step is usually unnecessary unless the goal is maximum purification.

The PCEM-extracted sample produced more fragmented and shorter reads compared to those obtained from the E.Z.N.A commercial kits (see Figure 5). Gel electrophoretic analysis of samples extracted using the E.Z.N.A Stool DNA Kit revealed a cluster of reads that far surpassed the 3,000 bp ladder, which is potentially indicative of the presence of High Molecular Weight (HMW) DNA (though there are smaller fragments present as well). The PCEM-extracted sample exhibited fragment lengths above the 3,000 bp marker, though the majority of the reads are highly fragmented and clustered around the 200 bp marker. This difference in read length (molecular weight) and fragmentation is further supported by the sequencing statistics. Sample 4's (extracted with E.Z.N.A) longest read was 53,201 bases, approximately 20,000 bases larger than the same sample extracted using the PCEM method, which had a read length of 32,804 bases. Despite these differences, when comparing the average read size and the average of the longest read, PCEM-extracted samples yielded similar results to E.Z.N.A. However, extraction with the E.Z.N.A Stool DNA Kit is preferred to fully leverage the advantages of the MinION's ability to sequence long reads. Future studies should adapt the PCEM to yield HMW DNA, essential for third-generation sequencing technologies capable of sequencing long reads of thousands of nucleotide bases.

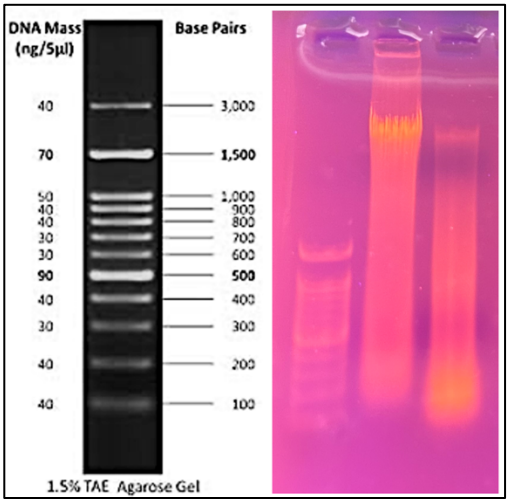

Figure 5. E.Z.N.A-extracted sample 4 (left) compared with PCEM-extracted (right) sample 12

It is also noteworthy that the PCEM-sample (sample 12) had the most total reads, 934,144 reads more than the highest E.Z.N.A sample (sample 5). However, sample 12 had 87,073 less mapped reads than sample 5. This yielded the second lowest percentage of mapped

reads among all samples (3.34%) which is almost as low as sample 8 (2.88%) which had 8.2x less total reads than sample 12. Therefore, it is important to consider that the PCEM may retain higher proportions of non-host reads, as indicated by the low percentage of mapped reads for sample 12.

Sanger Sequence Validation

Our validation statistics suggest the presence of *Ateles geoffroyi* DNA, despite the relatively high standard deviations. Some discrepancies between the reference genome and the Sanger sequence results, which resulted in < 100% identity and coverage, are in part likely due to natural individual variation and possible errors during PCR amplification and Sanger sequencing. Analysis of the electropherogram data in Geneious Prime revealed 'noisy' sequence results, with many positions experiencing the superposition of nucleotide data, which may explain the low coverage statistics observed.

MinION & SNP Panel

The MinION is capable of sequencing reads spanning thousands of bases. Therefore, it is recommended for future studies to prioritize and utilize DNA extraction methods that produce longer reads, ultimately enhancing N50 values. Long reads have the potential to improve the completeness of target and complex regions (see Huddleston et al., 2014; Pollard et al., 2018), not only leading to enhanced overall coverage, but also a more precise detection of genetic variants (see Cretu Stancu et al., 2017; Merker et al., 2018; Vollger et al., 2020) There is potential to enhance the production of longer reads during the Adapter Ligation & Clean-Up stage of MinION's library preparation. This stage offers the choice between two types of

reagents: SFB and Long Fragment Buffer (LFB). SFB preserves all DNA fragments regardless of their size, while LFB selectively enriches for fragments that are 3 kb or longer, thereby prioritizing long reads exclusively. For this experiment, given the inherent fragmentation of fecal DNA and limited knowledge of the specific fragment sizes of the samples (with only a rough estimation from gel electrophoresis), we opted to preserve all reads regardless of size.

A single MinION run has the capacity to generate millions of reads, demonstrating its potential for field-based genotyping applications. However, a significant challenge arises in obtaining sufficient quantities of host-specific mapped reads from feces. In the absence of pre-sequencing enrichment techniques, the majority of genetic material present in fecal samples is microbial. As a result, the ng/uL spectrophotometer values obtained during pre-sequencing extraction do not accurately reflect the quantities of host DNA present (see Table 2). Although adaptive sampling has been shown to improve host DNA coverage (e.g., Wanner et al., 2021), this feature should not replace, nor diminish, the importance of pre-sequencing enrichment methods. Firstly, pre-sequencing enrichment techniques provide a more accurate representation of host DNA amounts for sequencing. This enables researchers to even out host DNA concentration distributions across samples through dilution, a crucial step for effective multiplexing with the MinION. If host DNA quantities are not approximately known, host DNA is likely to be distributed unevenly across samples, as evidenced by our data (see Table 5). This leads to the overrepresentation and underrepresentation of certain samples. Secondly, it's important to note that adaptive sampling requires a reference genome. Consequently, this feature is not possible for many species. One potential alternative in such cases is to use adaptive sampling with reference genomes of closely related species. For instance, adaptive sampling has shown promising results for *Leontopithecus rosalia* (Wanner et al., 2021) utilizing a closely

related species' reference (*Saguinus imperator*). However, it is important to recognize that adaptive sampling is not truly portable nor feasible for use on devices lacking sufficient computational power, as demonstrated by our terminated sequencing run. For this reason, we sequenced all reads, which resulted in the overall reduction of host-specific reads sequenced.

Since adaptive sampling is computationally expensive, it is neither feasible nor recommended for field use with standard laptop devices which frequently have insufficient CPUs, GPUs, and storage capacities. Most of the current studies using adaptive sampling have been performed using microbial (e.g., Marquet et al., 2022) or mitochondrial (e.g., Wanner et al., 2021; Frank et al., 2023) reference genomes, or targeted gene regions (e.g, Payne et al. 2021; Stevanovski et al., 2022). The *Saguinus imperator* reference mitogenome Wanner et al., 2021 utilized to assemble the *L. rosalia* mitogenome, which was approximately 16 kb, was performed using 2080 GPUs. Adaptive sampling is capable of handling these smaller reference genomes and target regions. However, future studies are needed to address the feasibility of performing whole-genome sequencing using the adaptive sampling feature to enrich for reads within large nuclear reference genomes > 1 Gb. With various adaptive sampling techniques available (e.g., Squiggle Filter Dunn et al., 2021; Readfish, Payne et al., 2021), more studies are needed to demonstrate their ability to perform various genomic tasks (e.g., whole-genome assembly) and to effectively meet the specifications of different projects. Until further research is conducted and or substantial computational improvements are made to MinKNOW's Run Until API, it is cautioned against performing alignments using larger genomes, particularly for field usage, even with highly computationally capable laptops equipped with NVIDIA GeForce RTX (e.g., 4080) graphics cards. While pre-sequencing enrichment methods require additional processing time and materials, they provide a more reliable and field-friendly alternative to adaptive sampling.

Enhancing the coverage of individual SNPs or highly populated variant regions can be achieved through targeted amplification and sequencing. Several protocols have been developed for generating thousands of SNP sites across the genome using ddRAD-Seq (for platyrrhine primate species see Valencia et al., 2018; Martins et al., 2023). By employing this strategy to initially mine for SNPs, researchers can specifically target, amplify, and prioritize these highly populated variant regions, thereby improving the confidence of the reads and genotype calls by not burdening the sequencing system with representation of reads from the entire genome. This method is not only cost-effective (see Peterson et al., 2012), but the assessment of potential restriction sites can be readily visualized and analyzed (and the restriction enzymes purchased) for a particular sequence or species with available tools such as NEBcutter v3.0.17 (https://nc3.neb.com/NEBcutter/) (see Vincze et al., 2003 for v1.0). Additionally, REBASE, a database with catalogued information of the known restriction enzymes and their specifications (e.g., cleavage sites, commercial availability) can assist in the selection of the most beneficial restriction enzymes to meet the needs of the project (Roberts et al., 2009). In species lacking known SNPs, discovering them is possible through either a targeted sequencing approach or by conducting whole-genome sequencing to discover variant distributions and densities across the genome. Following whole-genome sequencing, another round of sequencing can be performed by selectively targeting either the regions that are highly populated with SNPs, or the individual SNPs themselves. This can be achieved through multiplex PCR, for example, or as previously discussed with the adaptive sampling feature which can be utilized to enrich for specific regions within the genome. Given the MinION's limited sequencing capacity and the potential for active pore degradation, enriching for specific regions (whether for initial SNP discovery or for

subsequent resequencing for validation and genotyping) with high variant coverage is expected to improve genotype call rates and accuracy.

We employed whole-genome sequencing to generate the SNP panel, leveraging the increased throughput of the MinION platform. However, this method presents clear drawbacks when applied to lower quality biological substrates such as feces. One of the drawbacks is low coverage, resulting in decreased genotype call rates, allelic dropout, and false-positive variant calls. Moreover, the algorithms responsible for calling variants through likelihood models are susceptible to false variant calls due to the constraints of limited read data. Even with stringent filtering criteria in place, our results are not immune from errors. Therefore, without enrichment techniques, the use of targeting sequencing to mine for SNPs might be the preferred method to maximize coverage and increase the number of informative loci.

Though we were able to generate 5,934 SNPs for *Ateles geoffroyi*, only four of these met the filtering criteria required for validation. This is primarily due to the low coverage and lack of genomic regions represented by the sequencing data. The average coverage percentage to the reference genome of 2.8% is notably low, and even in regions where reads did map, the average depth was only ~1x. Consequently, the majority of *Ateles geoffroyi* variations were not represented in our dataset because approximately 97% of the genome was not covered. Additionally, even if more loci had passed the filtering criteria (> 4), their identification and discrimination power would still have been relatively low due to the lack of genotypes called across loci (see Appendix B). Thus, to confidently and successfully perform genotyping, our SNP panel would require more loci than if we had a lesser number of highly genotyped and informative loci.

While studies using high quality sources of DNA set more stringent filtering criteria (for example, Yousefi et al., 2018 employed a genotype call rate of 0.9 to filter SNPs), we reduced the threshold to 0.25 to maintain enough loci for additional downstream filtering. With this one filtering criteria, our SNPs decreased from 5,934 to 234 loci (increasing the genotype call rate threshold per loci to 0.33 to include 4 individuals retained only 123 loci and increasing the proportion to 0.41 to include 5 individuals retained only 68 loci). Additionally, by filtering loci by a MAF of 0.2 (213 remaining after this filtering criteria, and 149 remaining with a MAF threshold to 0.1) and a minimum across sample coverage of 10x, we further decreased the number of loci to 29, which is less than the typical amounts employed for sufficient genotyping confidence and accuracy (for example, although Kidd et al., 2006 were able to genotype individuals relatively successfully using 19 SNPs, they estimated that increasing this number to 45-50 would increase genotype success to nearly identical levels achieved using the Combined DNA Index System (CODIS) markers). Out of these 29 loci, only 4 had called heterozygous genotypes, none of which significantly deviated from the expected heterozygosity under HWE.

Given the trivial percentage of called heterozygous genotypes, our data is indicative of high degrees of allelic dropout and consequently excess homozygosity. This is not surprising considering the coverage statistics across loci and genotypes. We observed an average coverage of 2.01x across loci for the 5,934 unfiltered SNPs. At the 4 validated loci, the average coverage across samples was 8.83x, with a coverage of 3.22x exclusively for the called genotypes. These 4 loci were the highest quality achieved out of the unfiltered 5,934 SNPs. The majority of SNPs were filtered out by initially filtering for a $\geq$ 10x coverage (5,934 to 181). As a result of low coverage across the dataset, individuals experience severe underrepresentation of one of their alleles. Adjusting the filtering order has been shown to enhance the representation of genotypes

across samples and loci. For example, in the red snapper SNP dataset, which is characterized by substantial missing data, initial filtering by read depth (> 5), quality score (> 20), minor allele count (> 3), and across-sample coverage per locus (> 15), then followed by filtering to remove loci and samples based on missing data, notably reduced the total missing genotypes from 75% to 35% (O'Leary et al., 2018), therefore allowing for more sample and loci representation.

The $Reads_m$ (%) data in Table 5 is particularly surprising. Many of the fecal samples show higher percentages of mapped reads than expected for feces, with four samples exceeding 30%, and two of these nearing 50%. On average, approximately 24.8% of sequenced reads mapped across all samples, despite the anticipated low percentage of host DNA in feces. In Wanner et al., 2021, for example, only 0.016% of the total reads from *L. rosalia* feces mapped to the mitogenome (0.032% mapped during the adaptive sampling run). Additionally, Sharma et al., 2019 observed that a range of 0.44% to 1.41% of the total mRNA reads from gorilla (*Gorilla gorilla gorilla)* and human (*Homo sapiens*) feces mapped to the reference genome after undergoing mRNA enrichment and rRNA depletion techniques, and 0.04% to 0.27% when exclusively depleting rRNA. These studies indicate extremely low proportions of host-specific DNA, even after selective enrichment. To understand this discrepancy, we analyzed the reads based on the number of bases they contain ($Bases_t$ and $Bases_m$ data in Table 5), for perhaps the host-specific reads were smaller and more fragmented, therefore appearing to contribute disproportionately to the total amount of genetic information present. Though the average percentage decreased from ~24.8% (mapped read percentage) to ~17% for the mapped bases percentage, this data is not adequate alone to explain this incongruity. Next, we aligned and analyzed all of the failed MinION reads (< 8 PSQ) to assess whether these disproportionally contained more failed non-host DNA (not aligning to the *Ateles geoffroyi* reference genome), for

potentially the *Ateles geoffroyi* reads were of higher quality and thus prioritized during sequencing. However, the failed data was consistent with the passed reads percentages (17.36%). Furthermore, the discrepancy cannot be accounted for by the adaptive sampling run, as the 12 minutes of selective enrichment only generated 88 thousand reads (0.59% of the total reads) and 64 million bases (0.64% of the total bases).

As indicative by the large standard deviation among the mapped read percentages (20.5%, with a minimum and maximum value of 3.34% and 46.67%, respectively), human dilution and pipetting errors may have some influence on the wide array of distribution, concentration, and representation of host DNA across samples. In addition, concentrations of DNA can vary in feces across animals and at different times of the day. Other potential causes for the high proportions of host-specific DNA may have resulted from uneven retention during library preparation, and the unsequenced reads that were still leftover in the MinION to be sequenced (which would have been sequenced given the active pores had not degraded, and if we performed sequencing > 72 hours). Given that approximately 3% of the genome was mapped, on average, there are likely a lot more DNA fragments still left in the MinION. Since we did not perform targeted sequencing, it is likely that DNA fragments for the remainder of the 97% of the genome still remains. With the limited and insufficient sequencing output, as indicative by the coverage statistics and the percentage of the genome mapped, it is difficult to determine the actual percentages of contributions from the host, given every nucleotide were to be sequenced. Though the observed high proportions of exhibited *Ateles geoffroyi* reads in our data are likely reflective of the total DNA proportions, it is a potential source for the expected discrepancy.

Using *Macaca mulatta* tissue for comparison against the fecal condition exposed the obvious limitations of using feces for whole-genome sequencing. From the data, we conclude

that the these limitations are a primary consequence of the size of the DNA fragments (N50 value of 864 bp for the *Ateles geoffroyi* fecal condition and 20.49 kb for the *Macaca mulatta* tissue condition), the proportion of host-specific reads (~25% for the *Ateles geoffroyi* fecal condition and > 99% for the *Macaca mulatta* tissue condition), and the percentage of the genome covered by reads (2.8% for the *Ateles geoffroyi* fecal condition and 39.19% for the *Macaca mulatta* tissue condition). As expected, sequencing DNA from tissue yielded more SNPs (103,260 unfiltered) compared to the fecal condition (5,934 unfiltered). This difference is even more remarkable when considering that only two individuals were represented in the *Macaca mulatta* dataset, whereas 12 *Ateles geoffroyi* individuals were sequenced for the fecal condition. Moreover, we were able to utilize more stringent filtering thresholds for the tissue-generated SNPs, thus increasing the confidence and accuracies of the called genotypes. Based on the data generated, if high-quality samples can be collected opportunistically (e.g., tissue sloshed-off from wounds), these should be used for the initial discovery of SNPs, and for the generation of the SNP panel.

Chapter V - Conclusion

While this project was unable to successfully generate an informative and comprehensive SNP panel from feces using a whole-genome sequencing approach for genotyping purposes, it contributes to the ongoing development of noninvasive genotyping methodologies. Serving as a foundational study, we establish a baseline for future noninvasive genotyping applications utilizing the portable MinION device. Although leveraging the MinION's high-throughput capabilities alone was not sufficient to generate the SNP panel, consequent primarily of the observed low coverage and lack of consensus genotype calls, our findings suggest this approach may be useful for initial SNP discovery. Further refinement of the protocol using enrichment and targeted sequencing methods holds great promise and potential to enhance genotype accuracies, consensus, and confidence.

Chapter VI - Future Directions

In future studies we will utilize enrichment and selective targeting strategies to increase host-specific yields and increase the coverage of SNP loci. We will first select the 50 most informative *Ateles geoffroyi* SNPs (including the four that were validated in this study), which will be based on a combination of quality, coverage, and genotype call rate factors (not necessarily based on MAF or heterozygosity scores, since the generated data is not representative of the population due to high rates of allelic dropout and homozygosity) to validate whether these are true SNPs and not merely artifacts which resulted from sequencing or bioinformatics errors (especially considering the relatively high sequencing errors of the MinION compared to other sequencing platforms). To accomplish this, we will perform multiplex PCRs (first described by Chamberlain et al., 1988) using the 6 samples with the most called genotypes across the 50 SNPs (with preferably half mapping to the reference genome and the other half having an alternate allele). We will use Primer3Plus (Untergasser et al., 2007) (https://www.primer3plus.com/index.html) to select primers based on their specificity (ability to target and amplify a unique sequence in the genome), binding efficiency, and melting temperature (see Hung and Weng, 2016), by extracting sequence data both 1,000 bases upstream and downstream to the SNP. We will then (i) perform multiplex PCRs to amplify the target regions, (ii) purify the PCR product, and (iii) send to the purified PCR products to OSU to perform NGS. We will compare the resultant FASTA files to the reference genome to confirm the validity of the SNPs. If we are able to validate all 50 loci, we will proceed to the enrichment and targeted sequencing preparation stages. If not, we will repeat this step by using the next most informative SNPs for validation.

Following SNP validation, we will utilize three different targeting techniques: multiplex PCR, FecalSeq enrichment, and adaptive sampling. To control for sample variability, we will use the same eight samples for each of the three conditions. Following a similar protocol to the previous step, we will perform multiplex PCRs to amplify the validated 50 loci from 8 of the individuals used in this study which had the highest percentage of mapped reads for the E.Z.N.A condition. Chen et al., 2016 observed a 90.5% total genotype accuracy (99.5% genotype accuracy with smaller sample size) with 90.4% of the 757 of the samples with uniform coverage by multiplexing 37 SNP loci simultaneously, while Podder et al., 2008 ultimately achieved a 100% genotype call rate with a > 99.9% accuracy while multiplexing 50 loci simultaneously across 49 samples. After purification, the PCR products will be ready for MinION library preparation.

We will also perform host-specific enrichment using the method described by Chiou & Bergey, 2018 in their supplemental notes. The authors were able to achieve a 195-fold enrichment of host DNA from decade old fecal samples, increasing the percentage of mapped reads from an original proportion of 0.34% to 28.8%. After enrichment, the samples will be ready for MinION library preparation. Lastly, to ready the final eight samples to test the adaptive sampling condition exclusively, we will maintain the extracted samples as is without any pre-sequence enrichment or targeting. We will then enable the adaptive sampling feature to be conducted across all barcoded samples (n=24). We will input and specify sequences that are unique to *Ateles geoffroyi* in the 'Run Until' API to enrich for the targeted and specific SNP loci. These will be the same sequences amplified by the PCR reactions, as we will already have verified their specificity and uniqueness. After the 24 samples are prepared (8 PCR condition, 8 FecalSeq, and 8 exclusively testing the adaptive sampling feature without any moderations post-

extraction), we will prepare the library using the Native Barcoding Kit and run the MinION for 72 hours, following an identical sequencing and bioinformatics protocol used in this study. We will then compare the different conditions to each other, and to the whole genome conditions from the *Ateles geoffroyi* and *Macaca mulatta* runs, to evaluate the efficacy of these enrichment and targeted sequencing techniques to confidently and successfully generate a SNP panel exclusively from feces. Additionally, with the new panel, we will determine the probability of identity (P_{ID}), which is the probability that two individuals selected at random have the same DNA at the specified genetic markers using the method described by Waits et al., 2001. We will validate our generated *Macaca mulatta* SNPs by comparing these against known SNPs using the MACSNVdb database (Du et al., 2020), (https://big.cdu.edu.cn/macsnvdb/).

References

Allendorf F. W. (2017). Genetics and the conservation of natural populations: allozymes to genomes. *Molecular ecology*, *26*(2), 420–430.

Anderson, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic acids research*, *9*(13), 3015-3027.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81-92.

Avise, J. C., Lansman, R. A., & Shade, R. O. (1979). The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. I. Population structure and evolution in the genus Peromyscus. *Genetics*, *92*(1), 279-295.

Ayres, K. L. (2005). The expected performance of single nucleotide polymorphism loci in paternity testing. *Forensic science international*, *154*(2-3), 167-172.

Babb, P. L., McIntosh, A. M., Fernandez-Duque, E., Di Fiore, A., & Schurr, T. G. (2011). An optimized microsatellite genotyping strategy for assessing genetic identity and kinship in Azara's owl monkeys (Aotus azarai). *Folia primatologica*, *82*(2), 107-117.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one*, *3*(10), e3376.

Beja-Pereira, A. L. B. A. N. O., 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.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

Benjamini, Y., Heller, R., & Yekutieli, D. (2009). Selective inference in complex research. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1906), 4255-4271.

BioRender.com (2021). Adapted from "Blank Panels (Layout 1x4)". Retrieved from https://app.biorender.com/biorender-templates/figures/all/t-6174c37cf20a9100a9081aee-blank-panels-layout-1x4

BioRender.com (2020). Adapted from "Next Generation Sequencing Data Processing". Retrieved from https://app.biorender.com/biorender-templates/figures/all/t-5f1708b1b093b600abfff81b-next-generation-sequencing-data-processing

BioRender.com (2020). Adapted from "SARS-CoV-2 Genome Sequencing using Oxford Nanopore Technologies". Retrieved from https://app.biorender.com/biorender-templates/figures/all/t-5fc92de9cbe89f1daa8fac20-sars-cov-2-genome-sequencing-using-oxford-nanopore-technolog

Blake, J. G., Guerra, J., Mosquera, D., Torres, R., Loiselle, B. A., & Romo, D. (2010). Use of mineral licks by white-bellied spider monkeys (*Ateles belzebuth*) and red howler monkeys (*Alouatta seniculus*) in eastern Ecuador. *International Journal of Primatology*, *31*, 471-483.

Bonin, A., Bellemain, E., Bronken Eidesen, P., Pompanon, F., Brochmann, C., & Taberlet, P. (2004). How to track and assess genotyping errors in population genetics studies. *Molecular ecology*, *13*(11), 3261-3273.

Børsting, C., Tomas, C., & Morling, N. (2012). Typing of 49 autosomal SNPs by single base extension and capillary electrophoresis for forensic genetic testing. *DNA Electrophoresis Protocols for Forensic Genetics*, 87-107.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., ... & Schloss, J. A. (2008). The potential and challenges of nanopore sequencing. *Nature biotechnology*, *26*(10), 1146-1153.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., ... & Schloss, J. A. (2008). The potential and challenges of nanopore sequencing. *Nature biotechnology*, *26*(10), 1146-1153.

Braslavsky, I., Hebert, B., Kartalov, E., & Quake, S. R. (2003). Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences*, *100*(7), 3960-3964.

Brookes, A. J. (1999). The essence of SNPs. *Gene*, *234*(2), 177-186.

Bruford, M. W., & Wayne, R. K. (1993). The use of molecular genetic techniques to address conservation questions. *Molecular Techniques in Environmental Biology*, 11-28.

Brumfield, R. T., Beerli, P., Nickerson, D. A., & Edwards, S. V. (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, *18*(5), 249-256.

Buerkle, C., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: how low should we go?. *Molecular ecology*, *22*(11), 3028-3035.

Campbell, C. J., Aureli, F., Chapman, C. A., Ramos-Fernández, G., Matthews, K., Russo, S. E., ... & Vick, L. (2005). Terrestrial behavior of Ateles spp. *International Journal of Primatology*, *26*, 1039-1051.

Cant, J. G. H. (1977). Ecology, Locomotion, and Social Organization of Spider Monkeys (*Ateles Geoffroyi*). University of Calif., Davis.

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.

Castro-Wallace, S. L., Chiu, C. Y., John, K. K., Stahl, S. E., Rubins, K. H., McIntyre, A. B., ... & Burton, A. S. (2017). Nanopore DNA sequencing and genome assembly on the International Space Station. *Scientific reports*, *7*(1), 18022.

Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics, 11(6), 415-425.

Chamberlain, J. S., Gibbs, R. A., Rainer, J. E., Nguyen, P. N., & Thomas, C. (1988). Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. *Nucleic acids research*, *16*(23), 11141-11156.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), s13742-015.

Chen, K., Zhou, Y. X., Li, K., Qi, L. X., Zhang, Q. F., Wang, M. C., & Xiao, J. H. (2016). A novel three-round multiplex PCR for SNP genotyping with next generation sequencing. *Analytical and bioanalytical chemistry*, *408*, 4371-4377.

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.

Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S., & Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature nanotechnology*, *4*(4), 265-270.

Cohen, S. N., Chang, A. C., Boyer, H. W., & Helling, R. B. (1973). Construction of biologically functional bacterial plasmids in vitro. *Proceedings of the National Academy of Sciences*, *70*(11), 3240-3244.

Collins, F. S., Brooks, L. D., & Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome research*, *8*(12), 1229-1231.

Conrad, D. F., Keebler, J. E., DePristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C. L., Torroja, C., Garimella, K. V., Zilversmit, M., Cartwright, R., Rouleau, G. A., Daly, M., Stone, E. A., Hurles, M. E., Awadalla, P., & 1000 Genomes Project (2011). Variation in genome-wide mutation rates within and between human families. *Nature genetics*, 43(7), 712–714.

Cornelis, S., Gansemans, Y., Deleye, L., Deforce, D., & Van Nieuwerburgh, F. (2017). Forensic SNP genotyping using nanopore MinION sequencing. *Scientific reports*, *7*(1), 41759.

Cortes-Ortíz, L., Solano-Rojas, D., Rosales-Meda, M., Williams-Guillén, K., Méndez-Carvajal, P. G., Marsh, L. K., ... & Mittermeier, R. A. (2021). *Ateles geoffroyi* (amended version of 2020 assessment). *The IUCN red list of threatened species*, 2021-1.

Cretu Stancu, M., Van Roosmalen, M. J., Renkens, I., Nieboer, M. M., Middelkamp, S., De Ligt, J., ... & Kloosterman, W. P. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature communications*, *8*(1), 1326.

Cunningham, E. P., Unwin, S., & Setchell, J. M. (2015). Darting primates in the field: a review of reporting trends and a survey of practices and their effect on the primates involved. *International Journal of Primatology*, *36*, 911-932.

Di Fiore, A., Link, A., Schmitt, C., & Spehar, S. (2009). Dispersal patterns in sympatric woolly and spider monkeys: integrating molecular and observational data. *Behaviour*, *146*(4-5), 437-470.

Di Fiore, A., & Fleischer, R. C. (2004). Microsatellite markers for woolly monkeys (Lagothrix lagotricha) and their amplification in other New World primates (Primates: Platyrrhini). *Molecular Ecology Notes*, *4*(2), 246-249.

Du, L., Guo, T., Liu, Q., Li, J., Zhang, X., Xing, J., ... & Fan, Z. (2020). MACSNVdb: a high-quality SNV database for interspecies genetic divergence investigation among macaques. *Database*, *2020*, baaa027.

Dunn, T., Sadasivan, H., Wadden, J., Goliya, K., Chen, K. Y., Blaauw, D., ... & Narayanasamy, S. (2021, October). Squigglefilter: An accelerator for portable virus detection. In *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture* (pp. 535-549).

Edwards, H. S., Krishnakumar, R., Sinha, A., Bird, S. W., Patel, K. D., & Bartsch, M. S. (2019). Real-time selective sequencing with RUBRIC: read until with basecall and reference-informed criteria. *Scientific reports*, *9*(1), 11475.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nature reviews genetics*, *5*(6), 435-445.

Ellsworth, J. A., & Hoelzer, G. A. (1998). Characterization of microsatellite loci in a New World primate, the mantled howler monkey (*Alouatta palliata*). *Molecular Ecology*, *7*(5), 657-658.

Eriksson, J., Hohmann, G., Boesch, C., & Vigilant, L. (2004). Rivers influence the population genetic structure of bonobos (Pan paniscus). *Molecular Ecology*, *13*(11), 3425-3435.

Escobar-Páramo, P. (2000). Microsatellite primers for the wild brown capuchin monkey *Cebus apella*. *Molecular Ecology*, *9*(1), 107-108.

Estrada, A. (2006). Human and non-human primate co-existence in the Neotropics: a preliminary view of some agricultural practices as a complement for primate conservation. *Ecological and Environmental Anthropology (University of Georgia)*, 3.

Ferragina, P., & Manzini, G. (2000, November). Opportunistic data structures with applications. In *Proceedings 41st annual symposium on foundations of computer science* (pp. 390-398). IEEE.

Fitak, R. R., Naidu, A., Thompson, R. W., & Culver, M. (2016). A new panel of SNP markers for the individual identification of North American pumas. *Journal of Fish and Wildlife Management*, *7*(1), 13-27.

Frank, L. E., Lindsey, L. L., Kipp, E. J., Faulk, C., Stone, S., Roerick, T. M., ... & Larsen, P. A. (2023). Rapid molecular species identification of mammalian scat samples using nanopore adaptive sampling. *bioRxiv*, 2023-06.

Gagneux, P., Boesch, C., & Woodruff, D. S. (1997). Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. *Molecular ecology*, *6*(9), 861-868.

Galbusera, P. H., & Gillemot, S. (2008). Polymorphic microsatellite markers for the endangered golden-headed lion tamarin, *Leontopithecus chrysomelas* (Callitrichidae). *Conservation Genetics*, *9*, 731-733.

Garvin, M. R., Saitoh, K., & Gharrett, A. J. (2010). Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources*, *10*(6), 915-934.

Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Briefings in bioinformatics*, *20*(6), 1981-1996.

Ghaheri, M., Kahrizi, D., Yari, K., Babaie, A., Suthar, R. S., & Kazemi, E. (2016). A comparative evaluation of four DNA extraction protocols from whole blood sample. *Cellular and Molecular Biology*, *62*(3), 120-124.

Gill, P. (2001). An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *International journal of legal medicine*, *114*(4), 204-210.

Gonçalves, E. C., Silva, A., Barbosa, M. S. R., & Schneider, M. P. C. (2004). Isolation and characterization of microsatellite loci in Amazonian red-handed howlers *Alouatta belzebul* (Primates, Plathyrrini). *Molecular Ecology Notes*, *4*(3), 406-408.

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M. C., & McCombie, W. R. (2015). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome research*, *25*(11), 1750-1756.

Goswami, R. S. (2016). PCR techniques in next-generation sequencing. *Clinical Applications of PCR*, 143-151.

Grativol, A. D., Ballou, J. D., & Fleischer, R. C. (2001). Microsatellite variation within and among recently fragmented populations of the golden lion tamarin (*Leontopithecus rosalia*). *Conservation Genetics*, *2*, 1-9.

Hale, V. L., Tan, C. L., Knight, R., & Amato, K. R. (2015). Effect of preservation method on spider monkey (*Ateles geoffroyi*) fecal microbiota over 8 weeks. *Journal of microbiological methods*, *113*, 16-26.

Haque, F., Li, J., Wu, H. C., Liang, X. J., & Guo, P. (2013). Solid-state and biological nanopore for real-time sensing of single chemical and sequencing of DNA. *Nano today*, *8*(1), 56-74.

Hayden, E. C. (2015). Pint-sized DNA sequencer impresses first users. *Nature*, *521*(7550), 15-16.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1-8.

Hedrick, P. W. (1999). Perspective: highly variable loci and their interpretation in evolution and conservation. *Evolution*, *53*(2), 313-318.

Hillier, L. W., Marth, G. T., Quinlan, A. R., Dooling, D., Fewell, G., Barnett, D., ... & Mardis, E. R. (2008). Whole-genome sequencing and variant discovery in C. elegans. *Nature methods*, 5(2), 183-188.

Hirabayashi, A., Yanagisawa, H., Yahara, K., & Suzuki, M. (2021). On-site genomic epidemiological analysis of antimicrobial-resistant bacteria in Cambodia with portable laboratory equipment. *Frontiers in Microbiology*, *12*, 675463.

Hoffman, J. I., & Amos, W. (2005). Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular ecology*, *14*(2), 599-612.

Hooper, L. V., Bry, L., Falk, P. G., & Gordon, J. I. (1998). Host–microbial symbiosis in the mammalian intestine: exploring an internal ecosystem. *Bioessays*, *20*(4), 336-343.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., ... & Xu, C. F. (2004). Detection of genotyping errors by Hardy–Weinberg equilibrium testing. *European Journal of Human Genetics*, *12*(5), 395-399.

Höss, M., Kohn, M., Pääbo, S., Knauer, F., & Schröder, W. (1992). Excrement analysis by PCR. *Nature*, 359(6392), 199.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., ... & Eichler, E. E. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome research*, *24*(4), 688-696.

Hui, R., D'Atanasio, E., Cassidy, L. M., Scheib, C. L., & Kivisild, T. (2020). Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Scientific Reports*, *10*(1), 18542.

Hung, J. H., & Weng, Z. (2016). Designing polymerase chain reaction primers using Primer3Plus. *Cold Spring Harbor Protocols*, *2016*(9), pdb-prot093096.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., ... & Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, *36*(4), 338-345.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology*, *17*, 1-11.

Jain, M., Tyson, J. R., Loose, M., Ip, C. L., Eccles, D. A., O'Grady, J., ... & Reference Consortium. (2017). MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9. 0 chemistry. *F1000Research*, *6*.

Johnson, S. S., Zaikova, E., Goerlitz, D. S., Bai, Y., & Tighe, S. W. (2017). Real-time DNA sequencing in the Antarctic dry valleys using the Oxford Nanopore sequencer. *Journal of biomolecular techniques: JBT*, *28*(1), 2.

Kasianowicz, J. J., Brandin, E., Branton, D., & Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, *93*(24), 13770-13773.

Kayser, M., & De Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics*, *12*(3), 179-192.

Kidd, K. K., Pakstis, A. J., Speed, W. C., Grigorenko, E. L., Kajuna, S. L., Karoma, N. J., ... & Kidd, J. R. (2006). Developing a SNP panel for forensic identification of individuals. *Forensic science international*, *164*(1), 20-32.

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

Koshy, L., Anju, A. L., Harikrishnan, S., Kutty, V. R., Jissa, V. T., Kurikesu, I., ... & Sudhakaran, P. R. (2017). Evaluating genomic DNA extraction methods from human whole blood using endpoint and real-time PCR assays. *Molecular biology reports*, *44*, 97-108.

Kovaka, S., Fan, Y., Ni, B., Timp, W., & Schatz, M. C. (2021). Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nature biotechnology*, *39*(4), 431-441.

Kraus, R. H., Vonholdt, B., Cocchiararo, B., Harms, V., Bayerl, H., Kühn, R., ... & Nowak, C. (2015). A single-nucleotide polymorphism-based approach for rapid and cost-effective genetic wolf monitoring in E urope based on noninvasively collected samples. *Molecular ecology resources*, *15*(2), 295-305.

Krawczak, M. (1999). Informativity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis: An International Journal*, *20*(8), 1676-1681.

Larsen, A. M., Mohammed, H. H., & Arias, C. R. (2015). Comparison of DNA extraction protocols for the analysis of gut microbiota in fishes. *FEMS microbiology letters*, *362*(5), fnu031.

Laver, T., Harrison, J., O'neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the oxford nanopore technologies minion. *Biomolecular detection and quantification*, *3*, 1-8.

Leggett, R. M., & Clark, M. D. (2017). A world of opportunities with nanopore sequencing. *Journal of experimental botany*, *68*(20), 5419-5429.

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

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094-3100.

Li, Z., Wang, Y., & Wang, F. (2018). A study on fast calling variants from next-generation sequencing data using decision tree. *BMC bioinformatics*, *19*, 1-14.

Lindpaintner, K. (1999). Genetics in drug discovery and development: challenge and promise of individualizing treatment in common complex diseases. *British medical bulletin*, *55*(2), 471-491.

Link, A., Milich, K., & Di Fiore, A. (2018). Demography and life history of a group of white-bellied spider monkeys (*Ateles belzebuth*) in western Amazonia. *American journal of primatology*, *80*(8), e22899.

Littiere, T. O., Castro, G. H., Rodriguez, M. D. P. R., Bonafé, C. M., Magalhães, A. F., Faleiros, R. R., ... & Verardo, L. L. (2020). Identification and functional annotation of genes related to horses' performance: from GWAS to post-GWAS. *Animals*, *10*(7), 1173.

Loose, M., Malla, S., & Stout, M. (2016). Real-time selective sequencing using nanopore technology. *Nature methods*, *13*(9), 751-754.

Magi, A., Semeraro, R., Mingrino, A., Giusti, B., & D'aurizio, R. (2018). Nanopore sequencing data analysis: state of the art, applications and challenges. *Briefings in bioinformatics*, *19*(6), 1256-1272.

Marquet, M., Zöllkau, J., Pastuschek, J., Viehweger, A., Schleußner, E., Makarewicz, O., ... & Brandt, C. (2022). Evaluation of microbiome enrichment and host DNA depletion in human vaginal samples using Oxford Nanopore's adaptive sequencing. *Scientific reports*, *12*(1), 4000.

Martins, A. B., Valença-Montenegro, M. M., Lima, M. G. M., Lynch, J. W., Svoboda, W. K., Silva-Júnior, J. D. S. E., ... & Fiore, A. D. (2023). A new assessment of robust capuchin monkey (Sapajus) evolutionary history using genome-wide SNP marker data and a Bayesian approach to species delimitation. *Genes*, *14*(5), 970.

Martins, M. M., & Galetti Junior, P. M. (2011). Informative microsatellites for genetic population studies of black-faced lion tamarins (*Leontopithecus caissara*). *Genetics and Molecular Biology*, *34*, 173-175.

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, *74*(2), 560-564.

McFarland, L. V. (2000). Normal flora: diversity and functions. *Microbial ecology in health and disease*, *12*(4), 193-207.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297-1303.

Menescal, L. A., Gonçalves, E. C., Silva, A., Ferrari, S. F., & Schneider, M. P. C. (2009). Genetic diversity of red-bellied titis (*Callicebus moloch*) from eastern Amazonia based on microsatellite markers. *Biochemical genetics*, *47*, 235-240.

Merker, J. D., Wenger, A. M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., ... & Ashley, E. A. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genetics in Medicine*, *20*(1), 159-163.

Miodonska, Z., Bugdol, M. D., & Krecichwost, M. (2016). Dynamic time warping in phoneme modeling for fast pronunciation error detection. *Computers in Biology and Medicine*, *69*, 277-285.

Morin, P. A., Luikart, G., Wayne, R. K., & SNP Workshop Group. (2004). SNPs in ecology, evolution and conservation. *Trends in ecology & evolution*, *19*(4), 208-216.

Mosley, C., & Gunkel, C. (2007). Cardiovascular and pulmonary support. *Zoo animal and wildlife immobilization and anesthesia*, *1*, 93-102.

Moss, E. L., Maghini, D. G., & Bhatt, A. S. (2020). Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nature biotechnology*, *38*(6), 701-707.

Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, *156*(1), 297-304.

Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y., ... & Yamamoto, M. (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nature communications*, *6*(1), 8018.

Nelson, M. R., Marnellos, G., Kammerer, S., Hoyal, C. R., Shi, M. M., Cantor, C. R., & Braun, A. (2004). Large-scale validation of single nucleotide polymorphisms in gene regions. *Genome research*, *14*(8), 1664-1668.

Ng, P. C., & Kirkness, E. F. (2010). Whole genome sequencing. *Methods in molecular biology (Clifton, N.J.)*, *628*, 215–226.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*(6), 443-451.

Nievergelt, C. M., Digby, L. J., Ramakrishnan, U., & Woodruff, D. S. (2000). Genetic analysis of group composition and breeding system in a wild common marmoset (Callithrix jacchus) population. *International Journal of Primatology*, *21*, 1-20.

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.

Nyrén, P., & Lundin, A. (1985). Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Analytical biochemistry*, *151*(2), 504-509.

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular ecology*, 10.1111/mec.14792.

Payne, A., Holmes, N., Clarke, T., Munro, R., Debebe, B. J., & Loose, M. (2021). Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nature biotechnology*, *39*(4), 442-450.

Perez-Sweeney, B. M., Valladares-Padua, C. L. A. U. D. I. O., Burrell, A. S., Di Fiore, A., Satkoski, J., Van Coeverden De Groot, P. J., ... & Melnick, D. J. (2005). Dinucleotide microsatellite primers designed for a critically endangered primate, the black lion tamarin (*Leontopithecus chrysopygus*). *Molecular Ecology Notes*, *5*(2), 198-201.

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.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one*, *7*(5), e37135.

Plesivkova, D., Richards, R., & Harbison, S. (2019). A review of the potential of the MinION™ single-molecule sequencing system for forensic applications. *Wiley Interdisciplinary Reviews: Forensic Science*, *1*(1), e1323.

Podder, M., Ruan, J., Tripp, B. W., Chu, Z. E., & Tebbutt, S. J. (2008). Robust SNP genotyping by multiplex PCR and arrayed primer extension. *BMC Medical Genomics*, *1*, 1-15.

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., & Sandhu, M. S. (2018). Long reads: their purpose and place. *Human molecular genetics*, 27(R2), R234-R241.

Pomerantz, A., Peñafiel, N., Arteaga, A., Bustamante, L., Pichardo, F., Coloma, L. A., ... & Prost, S. (2018). Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *GigaScience*, *7*(4), giy033.

Pompanon, F., Bonin, A., Bellemain, E., & Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, *6*(11), 847-859.

Pozo, G., Albuja-Quintana, M., Larreátegui, L., Gutiérrez, B., Fuentes, N., Alfonso-Cortés, F., & Torres, M. D. L. (2024). First whole-genome sequence and assembly of the Ecuadorian brown-headed spider monkey (*Ateles fusciceps fusciceps*), a critically endangered species, using Oxford Nanopore Technologies. *G3: Genes, Genomes, Genetics*, *14*(3), jkae014.

Prakash, S., Lewontin, R. C., & Hubby, J. L. (1969). A molecular approach to the study of genic heterozygosity in natural populations IV. Patterns of genic variation in central, marginal and isolated populations of Drosophila pseudoobscura. *Genetics*, *61*(4), 841.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), 559-575.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., ... & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, *13*, 1-13.

Rang, F. J., Kloosterman, W. P., & de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome biology*, *19*(1), 90.

Ren, Z. L., Zhang, J. R., Zhang, X. M., Liu, X., Lin, Y. F., Bai, H., ... & Yan, J. W. (2021). Forensic nanopore sequencing of STRs and SNPs using Verogen's ForenSeq DNA signature prep kit and MinION. *International journal of legal medicine*, *135*(5), 1685-1693.

Roberts, R. J., Vincze, T., Posfai, J., & Macelis, D. (2010). REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic acids research*, *38*(suppl_1), D234-D236.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., ... & Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*(7356), 348-352.

Sahlin, K., & Medvedev, P. (2021). Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nature communications*, *12*(1), 2.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., ... & Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, *239*(4839), 487-491.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, *74*(12), 5463-5467.

Sankoff, D., & Kruskal, J. B. (1983). Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. *Reading: Addison-Wesley Publication*.

Seki, M., Katsumata, E., Suzuki, A., Sereewattanawoot, S., Sakamoto, Y., Mizushima-Sugano, J., ... & Suzuki, Y. (2019). Evaluation and application of RNA-Seq by MinION. *DNA Research*, *26*(1), 55-65.

Sekirov, I., Russell, S. L., Antunes, L. C. M., & Finlay, B. B. (2010). Gut microbiota in health and disease. *Physiological reviews*.

Selkoe, K. A., & Toonen, R. J. (2006). Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology letters*, *9*(5), 615-629.

Sharma, A. K., Pafčo, B., Vlčková, K., Červená, B., Kreisinger, J., Davison, S., ... & Gomez, A. (2019). Mapping gastrointestinal gene expression patterns in wild primates and humans via fecal RNA-seq. *BMC genomics*, *20*, 1-14.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, *26*(10), 1135-1145.

Shih, P. J., Saadat, H., Parameswaran, S., & Gamaarachchi, H. (2023). Efficient real-time selective genome sequencing on resource-constrained devices. *GigaScience*, *12*, giad046.

Silva, A. N. B. D., Souza, R. D. C. M. D., Honorato, N. R. M., Martins, R. R., Câmara, A. C. J. D., Galvão, L. M. D. C., & Chiari, E. (2020). Comparison of phenol-chloroform and a commercial deoxyribonucleic acid extraction kit for identification of bloodmeal sources from triatomines (Hemiptera: Reduviidae). *Revista da Sociedade Brasileira de Medicina Tropical*, *53*, e20200189.

Silva, R. C. D., de Lima, S. C., dos Santos Reis, W. P. M., de Magalhães, J. J. F., Magalhães, R. N. D. O., Rathi, B., ... & Pena, L. (2023). Comparison of DNA extraction methods for COVID-19 host genetics studies. *Plos one*, *18*(10), e0287551.

Simpson, J. T., Workman, R., Zuzarte, P. C., David, M., Dursi, L. J., & Timp, W. (2016). Detecting DNA methylation using the oxford nanopore technologies MinION sequencer. *BioRxiv*, 047142.

Skutkova, H., Vitek, M., Sedlar, K., & Provaznik, I. (2015). Progressive alignment of genomic signals by multiple dynamic time warping. *Journal of theoretical biology*, *385*, 20-30.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., ... & Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, *321*(6071), 674-679.

Soto-Calderón, I. D., Ntie, S., Mickala, P., Maisels, F., Wickings, E. J., & Anthony, N. M. (2009). Effects of storage type and time on DNA amplification success in tropical ungulate faeces. *Molecular Ecology Resources*, *9*(2), 471-479.

Spehar, S. N., Link, A., & Di Fiore, A. (2010). Male and female range use in a group of white-bellied spider monkeys (Ateles belzebuth) in Yasuní National Park, Ecuador. *American Journal of Primatology: Official Journal of the American Society of Primatologists*, *72*(2), 129-141.

Stephen, A. M., & Cummings, J. H. (1980). The microbial contribution to human faecal mass. *Journal of Medical Microbiology*, *13*(1), 45-56.

Stevanovski, I., Chintalaphani, S. R., Gamaarachchi, H., Ferguson, J. M., Pineda, S. S., Scriba, C. K., ... & Deveson, I. W. (2022). Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing. *Science advances*, *8*(9), eabm5386.

Subbaiyan, G. K., Waters, D. L., Katiyar, S. K., Sadananda, A. R., Vaddadi, S., & Henry, R. J. (2012). Genome-wide DNA polymorphisms in elite indica rice inbreds discovered by whole-genome sequencing. *Plant Biotechnology Journal*, 10(6), 623-634.

Syvänen, A. C., Sajantila, A., & Lukka, M. (1993). Identification of individuals by analysis of biallelic DNA markers, using PCR and solid-phase minisequencing. *American journal of human genetics*, *52*(1), 46.

Taberlet, P., Griffin, S., Goossens, B., Questiau, S., Manceau, V., Escaravage, N., ... & Bouvet, J. (1996). Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic acids research*, *24*(16), 3189-3194.

Trigodet, F., Lolans, K., Fogarty, E., Shaiber, A., Morrison, H. G., Barreiro, L., ... & Eren, A. M. (2022). High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. *Molecular ecology resources*, *22*(5), 1786-1802.

Turcatti, G., Romieu, A., Fedurco, M., & Tairi, A. P. (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic acids research*, *36*(4), e25-e25.

Tyson, J. R., O'Neil, N. J., Jain, M., Olsen, H. E., Hieter, P., & Snutch, T. P. (2017). Whole genome sequencing and assembly of a Caenorhabditis elegans genome with complex genomic rearrangements using the MinION sequencing device. *BioRxiv*, 099143.

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., & Leunissen, J. A. (2007). Primer3Plus, an enhanced web interface to Primer3. *Nucleic acids research*, *35*(suppl_2), W71-W74.

Valencia, L. M., Martins, A., Ortiz, E. M., & Di Fiore, A. (2018). A RAD-sequencing approach to genome-wide marker discovery, genotyping, and phylogenetic inference in a diverse radiation of primates. *PloS one*, *13*(8), e0201254.

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*(1), 11-10.

Van Der Flier, L. G., & Clevers, H. (2009). Stem cells, self-renewal, and differentiation in the intestinal epithelium. *Annual review of physiology*, *71*, 241-260.

van der Reis, A. L., Beckley, L. E., Olivar, M. P., & Jeffs, A. G. (2023). Nanopore short-read sequencing: A quick, cost-effective and accurate method for DNA metabarcoding. *Environmental DNA*, *5*(2), 282-296.

van Roosmalen, M. G. (1985). Habitat preferences, diet, feeding strategy and social organization of the black spider monkey [*Ateles paniscus paniscus* Linnaeus 1758] in Surinam. *Acta Amazonica*, *15*, 7-238.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... & Kalush, F. (2001). The sequence of the human genome. *science*, *291*(5507), 1304-1351.

Vincze, T., Posfai, J., & Roberts, R. J. (2003). NEBcutter: a program to cleave DNA with restriction enzymes. *Nucleic acids research*, *31*(13), 3688-3691.

Vollger, M. R., Logsdon, G. A., Audano, P. A., Sulovari, A., Porubsky, D., Peluso, P., ... & Eichler, E. E. (2020). Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Annals of human genetics*, 84(2), 125-140.

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.

Waits, L. P., Luikart, G., & Taberlet, P. (2001). Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular ecology*, 10(1), 249-256.

Wanner, N., Larsen, P. A., McLain, A., & Faulk, C. (2021). The mitochondrial genome and Epigenome of the Golden lion Tamarin from fecal DNA using Nanopore adaptive sequencing. *BMC genomics*, *22*, 1-11.

Wigginton, J. E., Cutler, D. J., & Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *The American Journal of Human Genetics*, *76*(5), 887-893.

Williams, R. C. (1989). Restriction fragment length polymorphism (RFLP). *American Journal of Physical Anthropology*, *32*(S10), 159-184.

Witte, S. M., & Rogers, J. (1999). Microsatellite polymorphisms in Bolivian squirrel monkeys (Saimiri boliviensis). *American Journal of Primatology: Official Journal of the American Society of Primatologists*, *47*(1), 75-84.

Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., ... & Lee, J. Y. (2022). A saturated map of common genetic variants associated with human height. *Nature*, *610*(7933), 704-712.

Yousefi, S., Abbassi-Daloii, T., Kraaijenbrink, T., Vermaat, M., Mei, H., van 't Hof, P., ... & 't Hoen, P. A. (2018). A SNP panel for identification of DNA and RNA specimens. *BMC genomics*, *19*, 1-12.

Zhang, D. X., & Hewitt, G. M. (2003). Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular ecology*, *12*(3), 563-584.

Zhang, P., Hu, K., Yang, B., & Yang, D. (2016). Snub-nosed monkeys (*Rhinopithecus spp.*): conservation challenges in the face of environmental uncertainty. *Science Bulletin*, *61*, 345-348.

Appendices

Appendix A - Phenol chloroform extraction method spectrophotometry results

ID	Condition	Feces	K	PCI	C	EtOH	Time	260	280	320	260/280	ng/µL
AG0002	Baseline	200	2	1	0	1	24	0.32	0.22	0.21	1.46	319.24
AG0004	Proteinase K, Purification	200	4	1	0	2	24	0.35	0.24	0.34	1.47	352.48
		200	4	2	0	1	24	0.23	0.16	0.29	1.48	230.68
		200	4	2	0	2	24	0.12	0.08	0.13	1.43	119.62
AG0001	Proteinase K, Purification	200	5	1	0	1	24	0.15	0.10	0.23	1.44	150.4
		200	10	1	0	1	24	0.19	0.14	0.24	1.39	193.38
		200	5	1	0	2	24	0.14	0.1	0.22	1.46	141.34
		200	10	1	0	2	24	0.14	0.1	0.2	1.46	141.59
AG0005	Proteinase K, Purification	200	25	2	2	1	20	0.16	0.09	0.15	1.69	156.41
		200	50	2	2	1	20	0.12	0.07	0.13	1.63	121.36
AG0005	Feces, Incubation	75	25	2	2	1	20	0.23	0.12	0.1	1.89	234.09
		150	25	2	2	1	20	0.15	0.09	0.12	1.73	150.0
		75	25	2	2	1	72	0.16	0.09	0.09	1.85	156.72
		150	25	2	2	1	72	0.21	0.13	0.19	1.59	206.26
		75	25	2	2	1	92	0.24	0.13	0.11	1.90	242.29
		150	25	2	2	1	92	0.14	0.09	0.13	1.63	140.65
AG0005	Purification	75	25	3	2	1	24	0.24	0.12	0.1	1.98	239.34
		75	25	3	2	2	24	0.26	0.14	0.12	1.86	255.46
		75	25	2	3	1	24	0.26	0.15	0.13	1.75	258.02
		75	25	2	3	2	24	0.24	0.13	0.09	1.88	236.31
AG0005	Purification	75	25	5	0	1	24	0.34	0.18	0.08	1.84	335.06
		75	25	4	0	1	24	0.25	0.13	0.10	1.93	246.1
		75	25	3	0	1	24	0.29	0.16	0.09	1.81	286.82
		75	25	0	5	1	24	0.19	0.11	0.12	1.76	193.69
		75	25	0	4	1	24	0.17	0.1	0.09	1.77	173.49
		75	25	0	3	1	24	0.26	0.14	0.09	1.83	262.33
AG0005	Feces	50	25	4	1	1	18	0.48	0.24	0.10	1.97	476.73
		75	25	4	1	1	18	0.4	0.21	0.11	1.93	399.4
AG0005	Incubation	50	25	4	1	1	6	0.59	0.32	0.2	1.85	591.63
		50	25	4	1	1	12	0.87	0.48	0.28	1.81	864.59
		50	25	4	1	1	18	0.78	0.43	0.29	1.83	775.73
		50	25	4	1	1	24	0.31	0.17	0.12	1.88	312.14

K, Proteinase K amounts in uL; PCI, number of phenol/chloroform/isoamyl stages; C, number of chloroform stages; EtOH, number of ethanol wash stages; Time, incubation time in hours. All samples were diluted with 150 uL DEPC. Yellow highlight is representative of a noticeably high 320 yield indicated by the spectrophotometry machine, potentially indicating contamination.

Appendix B - Missingness statistics for 29 loci (left) & 4 loci (right)

Sample	Loci (#)	Missed (#)	Missed (%)
Feces1	29	28	0.965517
Feces2	29	15	0.517241
Feces3	29	15	0.517241
Feces4	29	25	0.862069
Feces5	29	14	0.482759
Feces6	29	10	0.344828
Feces7	29	7	0.241379
Feces8	29	28	0.965517
Feces9	29	25	0.862069
Feces10	29	14	0.482759
Feces11	29	16	0.551724
Feces12	29	24	0.827586

Sample	Loci (#)	Missed (#)	Missed (%)
Feces1	4	4	1
Feces2	4	1	0.25
Feces3	4	2	0.5
Feces4	4	4	1
Feces5	4	2	0.5
Feces6	4	0	0
Feces7	4	0	0
Feces8	4	4	1
Feces9	4	3	0.75
Feces10	4	2	0.5
Feces11	4	1	0.25
Feces12	4	4	1

Loci (#), number of total loci represented in the assessment; Missed (#), number of loci with missing genotype data; Missed (%), percentage of loci with genotype data (Missed (#)/Loci (#)).

Appendix C - BWA-GATK Variant Calling Workflow Code

Merge FASTQ Files:
```
cat * > your_file.fastq.gz
```

Burrows-Wheeler Aligner Index:
```
bwa index your_file.fasta
```

Picard SequenceDictionary:
```
java -jar picard.jar CreateSequenceDictionary \
R=your_file.fasta
O=your_file.dict
```

Burrows-Wheeler Aligner:
```
bwa mem -t 28 -R '@RG\tID:your_ID \tSM:your_sample' your_file.fasta your_file.fastq.gz >
your_file.sam
```

Samtools View:
```
samtools view -S -b your_file.sam > your_file.bam
```

Samtools Sorted:
```
samtools sort -o your_file_sorted.bam your_file.bam
```

Picard MarkDuplicates:
```
java -jar picard.jar MarkDuplicates \
-I your_file_sorted.bam -O your_file_sorted_markduplicates.bam -M
your_file_sorted_markduplicates_metrics.txt
```

Samtools Index:
```
samtools index your_file_sorted_markduplicates.bam
```

GATK HaplotypeCaller:
```
./gatk-4.3.0.0/gatk HaplotypeCaller \
   -R GCA_023783555.1/GCA_023783555.1_ASM2378355v1_genomic.fna \
   -I your_file_sorted_markduplicates.bam \
   -O your_file.g.vcf \
   -ERC GVCF
```

Gatk CombineGVCFs:
```
./gatk-4.3.0.0/gatk CombineGVCFs \
   -R your_file.fasta \
   -V your_file1.g.vcf \
   -V your_file2.g.vcf \
   -O your_file_combined.g.vcf
```

GATK GenotypeGVCFs:

```
./gatk-4.3.0.0/gatk GenotypeGVCFs \
   -R your_file.fasta \
   -V your_file_combined.g.vcf \
   -O your_file.vcf
```

GATK SelectVariants:

```
./gatk-4.3.0.0/gatk SelectVariants \
   -R your_file.fasta \
   -V your_file.vcf \
   --select-type-to-include SNP \
   -O your_file_snps.vcf
```

GATK VariantFilatration:

```
./gatk-4.3.0.0/gatk VariantFiltration -V your_file_snps.vcf \
   --filter-expression "QD < 2.0" --filter-name "QD2" \
   --filter-expression "QUAL < 30.0" --filter-name "QUAL30" \
   --filter-expression "SOR > 3.0" --filter-name "SOR3" \
   --filter-expression "FS > 60.0" --filter-name "FS60" \
   --filter-expression "MQ < 30.0" --filter-name "MQ30" \
   --filter-expression "MQRankSum < -12.5" --filter-name "MQRankSum-12.5" \
   --filter-expression "ReadPosRankSum < -8.0" --filter-name "ReadPosRankSum-8" \
   -O your_file_snps_filtered.vcf
```

GATK VariantsToTable:

```
./gatk-4.3.0.0/gatk VariantsToTable \
   -V your_file_snps_filtered.vcf \
   -F CHROM -F POS -F TYPE -F REF -F ALT -F QUAL \
   -GF GT -GF PL -GF AD \
   -O your_file_snps_filtered.table
```