Developing Methods for Access to High Quality Genome Sequences from Wild Ape Populations

Dylan Charles Koundakjian

Honors Thesis submitted in partial fulfillment of the requirements for graduation with Distinction in Biology in Trinity College of Duke University.

Under the supervision of Dr. Gregory Wray,
Biology Department, Duke University, Durham, NC

April, 2014

Abstract:
Modern evolutionary study of wild ape populations requires the collection of genomic DNA from individuals living in their natural habitat. In order to be maximally useful, these samples must be robust enough for the amplification and subsequent assembly of genomic sequences, which are driving much of modern evolutionary research. Additionally, conservation efforts require that these samples be collected with zero intervention on the study species, because all great apes are now critically endangered. Consequently, the conventional method for genomic DNA collection has been extraction from cells present in fecal samples. However, this approach presents multiple difficulties to investigators, including extensive contamination of sequences from gut microbiota and limited storage time. The purpose of this study is to explore alternative procedures of noninvasive DNA collection to overcome these challenges. Specifically, the study looks at DNA extraction from hair follicle cells and from cells present in urine. These sources of genomic information confer a number of advantages over feces, such as smaller volumes of collection, much lower levels of microbial contamination, and relative ease of storage and transport. In this study, a method for isolating genomic DNA from chimpanzee (*Pan troglodytes*) hair follicle cells is developed and tested for limit of detection using a decreasing number of hairs per extraction. Validation of the method is then established through the determination of the frequency of polymorphisms due genomic amplification error by comparing sequences obtained from three identically handled samples. After laying out the next steps of development for this method, the study also suggests a similar investigation for samples derived from urine. The overall aim of these studies is a future incorporation of these procedures into the suite of DNA collection techniques available to researchers working with natural populations of great apes and other mammals.

Introduction:

The study of wild ape populations has offered a wide variety of avenues for biological research, including investigations in ecology, development, evolution, phylogeny, and medical applications. Since the seminal work of Jane Goodall on the Kasakela chimpanzee population of the Gombe reserve in Tanzania, researchers have sought to learn all they can from natural populations of our phylogenetic relatives. Goodall's research in 1960 and beyond provided much of the basis for our current understanding of primate behavior. Her groundbreaking discoveries included several distinct behavioral patterns of affection, aggression, and tool-use, phenomena essential to the evolutionary study of both apes and humans (Goodall 1986). She also demonstrated the importance of keeping these critically endangered specimens in their natural environment (Goodall 1986). This is an important measure in preventing the development of specific domesticated traits in captivity which may eclipse the expression of natural behaviors necessary to our understanding of primate evolution (Price 1999). Furthermore, once developed, these traits can make reintroduction into natural habitats extremely difficult, which can have drastic effects on endangered populations (Price 1999). Since these first studies, primate researchers have aimed to learn all they can with the least amount of intervention on their wild subjects.

Beyond these behavioral studies, research on wild ape populations has led to many key medical and evolutionary discoveries. Most notably, the work of Beatrice Hahn and colleagues identified the origin of the Human Immunodeficiency Virus (HIV), the cause of Acquired Immune Deficiency Syndrome (AIDS) in humans, from the closely related Simian Immunodeficiency Virus (SIV) present in several ape species (Hahn 2011). Crucial parts of her epidemiological model were based on research tracking strains of HIV to west equatorial

Africa, where they began circulating in human populations (Hahn 2000). This location was considered the likely source because it contained a natural range of chimpanzees infected with strains of SIV genetically similar to the HIV strains present (Gao 1999). This similarity helped Hahn make the connection between the two types of viruses, information which has led to further discoveries of viral phylogeny and potential treatments for AIDS-infected patients (Hahn 2000). Their work also reaffirmed the commitment of ape researchers to be as noninvasive to their subjects as possible, as new methods were developed specifically for the least intrusive collection of samples (Hahn 2011). Hahn and colleagues have continued this path of research into the present day in hopes of one day finding a cure to this incredibly serious disease.

Hahn's research centers on the study of samples of DNA isolated from the wild chimpanzee populations of central Africa (Hahn 2011). The data provided by these kinds of samples are used in all fields of primate biology, from discerning pathways in primate cognitive evolution to phylogenetic analyses of ape speciation (Fisher & Marcus 2006, Telfer 2003). In our laboratory, we are using genetic data collected from a captive adult mandrill to determine its region of origin for its successful reintroduction into the wild. Moreover, today's modern advances in DNA sequencing and genomic analysis have made these samples more useful than ever before. Particular focus is set on obtaining the full genomic sequences of apes for their value in various applications, such as tracing broad patterns of molecular evolution and their use as a baseline in human population genetics (The Chimpanzee Sequencing and Analysis Consortium 2005). Yet, while the value of this information has increased, it remains important not to disturb these endangered species by engaging in invasive methods of collection (Taberlet 1999). Herein lies the general aim of

our study: to develop a highly robust method for the assembly of genomic information from natural samples with little to no intervention on the study individuals. Since we wish to keep our collection efforts away from interaction with our donors, we choose to focus on what they leave behind, namely feces, urine, and hair.

Collection of DNA samples from feces has been the most trusted noninvasive method for genomic sequencing used by primate biologists up to this point (Kohn 2010). This method takes advantage of the intestinal wall epithelial cells present in feces as sources of nuclear DNA which can be purified and amplified on a large scale (Perry 2010). Fecal samples are readily available in the field and are permitted for transport by the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) (CITES 1994). However, there are several challenges associated with extraction from these samples. First off, fecal samples contain a large amount of contamination from the individual's gut microbiota (Marrero 2009). These microorganisms present in the digestive tracts of all mammals often come along for the ride when waste is eliminated, posing a severe contamination hazard (Marrero 2009). Secondly, fecal samples can often contain chemicals which inhibit the polymerase chain reaction (PCR), a major difficulty since PCR is used in most modern genetic analyses (Perry 2010). Lastly, fecal matter has a limited useful shelf-life and is simply not very enjoyable to work with (Wasser 1997). It is therefore desirable to find other sources of equally robust DNA samples which come from safer, less contaminated sources.

One alternative for the collection of genomic DNA samples is isolation from urine. Urine samples contain epithelial cells and white blood cells that are being expelled as waste (Smuts & Pogue 1999). Fortunately for primate genomics researchers, noninvasive methods

for collecting urine from apes have already been developed because of its use in hormone testing, a procedure carried out to ensure the health of wild populations (Knott 1997). But in contrast to feces, urine samples have less contamination from gut microbiota (Siddiqui 2011). Consequently, collection of DNA from urine for genomic sequencing seems preferable to isolation from feces, aside from one drawback. To this point, not enough testing has been done on urine samples to determine if it is an efficient enough method of DNA isolation for genomic sequencing. Initial testing shows genotyping from DNA isolated from urine is less accurate than genotyping from feces, so it remains unclear whether isolated material is of high enough quality for consistent whole genome amplification and sequencing (Inoue 2007). However, this is not an insurmountable obstacle, and the discussion of this paper proposes further study of urine-based extraction methods.

The final proposed source of genomic DNA material is isolation from hair follicle cells. A procedure for DNA extraction from hair was first proposed by Von Beroldingen *et al.* in 1987 and this procedure was actually the first noninvasive method used for collecting nuclear DNA from primates, predating extraction from feces by eight years (Constable 1995). Additionally, the first comprehensive noninvasive genetic studies used DNA samples derived from hair as their primary resource (Morin 1994, Taberlet 1999). However, recent genetic analyses have preferred extraction from feces over hair because of its higher DNA content and greater genotyping accuracy (Morin 2001). While DNA extracted from hair is generally lower in both quality and quantity compared to that obtained from fecal and urine samples, it does possess several advantages over these methods. Firstly, DNA isolated from a hair does not contain the same levels of exogenous bacterial DNA or PCR inhibiting chemicals as fecal or urine samples do (Morin 2001). Additionally, hair samples can be

maintained in the smallest volumes, have the greatest stability over long periods of time, and require the least stringent preservation conditions of these methods (Morin 2001). Despite these advantages, a method has not yet been established for use of isolated hair follicle DNA in genomic sequencing. Thus, the need arises to examine the feasibility of such a method regardless of the lower quality of hair follicle-derived DNA. The development of hair-based method would confer additional benefits not associated with fecal extraction and would also provide maximum adaptability to potential investigative constraints, including sample availability and extended study timelines.

For our experiment, we chose two specific aims focusing specifically on the methodology of DNA extraction and genomic amplification from hair follicles. The first specific aim was to determine a minimum number of hairs required for the detection of extracted nuclear DNA. Since the hairs must be collected from nests without intervention on the inhabitant, it is important to know how much collection is necessary, as samples are often limited. This first aim was tested using a DNA extraction protocol on sample sets ranging from 30 hairs down to 5 hairs, followed by a Whole Genome Amplification (WGA) and subsequent PCR analysis. The second aim of our experiment was the validation of this amplification using comparisons of sequence data to ensure that the genetic information collected is not only robust, but consistent and relatively free of errors. We tested for these errors by performing three WGAs side-by-side from the same extraction sample, amplifying and sequencing two specific fragments, and comparing the results of the three amplifications for polymorphisms both within replicates of the same amplification and between amplifications. This second aim is especially important for ensuring that results from this method are repeatable enough to be useful in actual genomic analyses. The results of these

analyses met our first specific aim, as they demonstrated the possibility of robustly amplifying hair follicle-extracted DNA using material from a minimal number of hairs. However, we cannot report on our second aim as we have not yet obtained sequences of high enough quality for accurate comparison between sequence replicates and determination of polymorphisms.

Materials and Methods:

All testing was performed using hairs collected from chimpanzees (*Pan troglodytes*) "Lance" and "Ebi" kept at the North Carolina Zoological Park in Asheboro, NC. Upon delivery to our lab, the hairs were stored away from light at room temperature.

Limit of Detection Testing

DNA Extraction from Hair Follicles

Six sets of hair follicles were collected from the two chimpanzee donors by clipping the follicles off from the hair shafts. Sets from "Lance's" hair consisted of 30, 13, and 7 follicles, and sets of "Ebi's" hair consisted of 13, 8, and 5 follicles. The follicles were then treated as previously described in a User-Developed Protocol (DY04 Aug-06) for the purification of total DNA from hair using the DNeasy® Blood & Tissue Kit (QIAGEN).

Whole Genome Amplification (WGA)

Once purified from the follicle cells, the six extracted samples were amplified using a REPLI-g Whole Genome Amplification kit (QIAGEN), incubated at 30^{o}C for 16 hours, followed by a 3 minute deactivation of the DNA polymerase. The samples were then diluted 1:20 in TE buffer (QIAGEN).

Polymerase Chain Reaction (PCR) and Gel Electrophoresis

The six amplified samples were prepared in separate 13 µl PCR reactions using two sets of primers developed by Mr. W.J. Nielsen, designated "D4" and "D18" (Sequences in Supplementary Materials). These primers had been proven to work with chimpanzee genomic DNA in a separate paternity study conducted by Mr. Nielsen, and were expected to return DNA fragments approximately 270 bp in size for "D4" and 290 bp for "D18". The PCR reaction used a Phusion DNA polymerase (New England Biolabs®) with an annealing temperature of 67^{o}C. After amplification, the resulting products were run on a 1% agarose gel to visualize the success of the reaction.

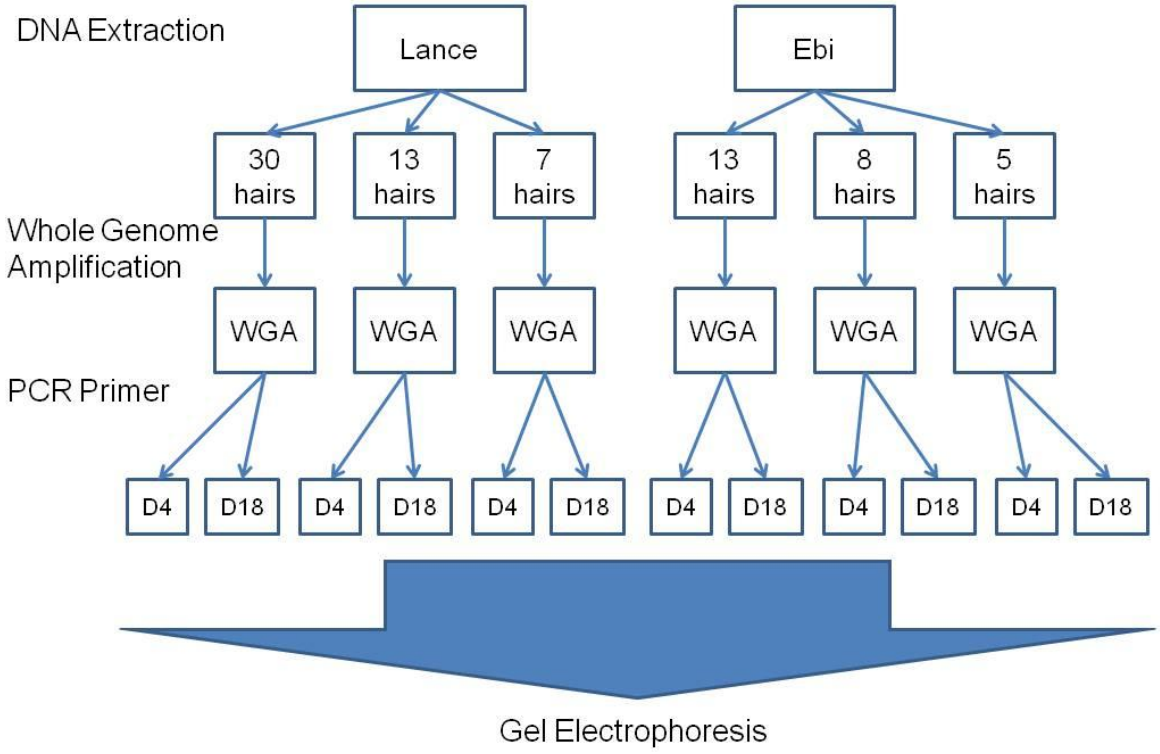

Figure 1: Flow Chart of Basic Genomic Amplification and PCR Procedure
Six DNA extractions were performed using different numbers of hairs. Lowest limit tested was extraction from 5 hairs. After extraction, samples underwent whole genome amplification, PCR using "D4" and "D18" primers, and visualization by gel electrophoresis.

Validation Protocol

Obtaining DNA Sequences from Simultaneous WGAs

Three separate WGAs were created simultaneously using the "Lance" 13 hair extraction prepared in the first part of the experiment, following the same REPLI-g protocol. PCR was then conducted on all three amplifications using an expanded set of primers, designated "D4," "D18", and "D9," using a Phusion DNA polymerase and an annealing temperature of 58^{o}C. The resulting products were checked on a 1% agarose gel to ensure the reaction had worked correctly. These products were purified of primer-dimer and excess nucleotides using a QIAquick® PCR Purification kit. The purified products of the "D4" and "D18" reactions were then cloned into a TOP10 competent strain of *Escherichia coli* using a PCR4® Blunt TOPO plasmid vector (not enough vector was available for cloning the "D9" samples). The *E. coli* cells were then grown overnight at 37^{o}C. Six of the resulting colonies were picked per sample plate for overnight growth in broth culture. The plasmid DNA was subsequently purified from 24 of the broth cultures, four per sample, using a Wizard® Mini-Prep kit (Promega). The 24 preps (4 preps multiplied by 2 primers multiplied by 3 WGAs) were run through a Big Dye v3.1 (Life Technologies®) reaction and submitted to Eton Biosciences Inc. for sequencing.

Sequence Analysis

Sequence information was analyzed using Molecular Evolutionary Genetics Analysis (MEGA) 5.01 software to observe the frequency of polymorphism due to amplification error (Tamura 2011). Polymorphism is defined as variation in base sequence at the same location in replicates of the amplified fragment, and was considered both within sequence replicates from the same WGA and between replicates from different WGAs. A reference wild-type

sequence was to be defined through a majority rule among replicates, wherein the base occurring in the sequences of 2/3 of the WGAs would be considered the original genotype and variation present in the sequence of the third WGA would be considered a mutation. Cases where there is no clear majority in a binary polymorphism among all replicates would indicate heterozygosity; these cases were not to be included in the calculated frequency of error-derived polymorphism. Different base calls in the same location for all three WGAs were to be considered inconclusive and most likely an incidence of amplification error.

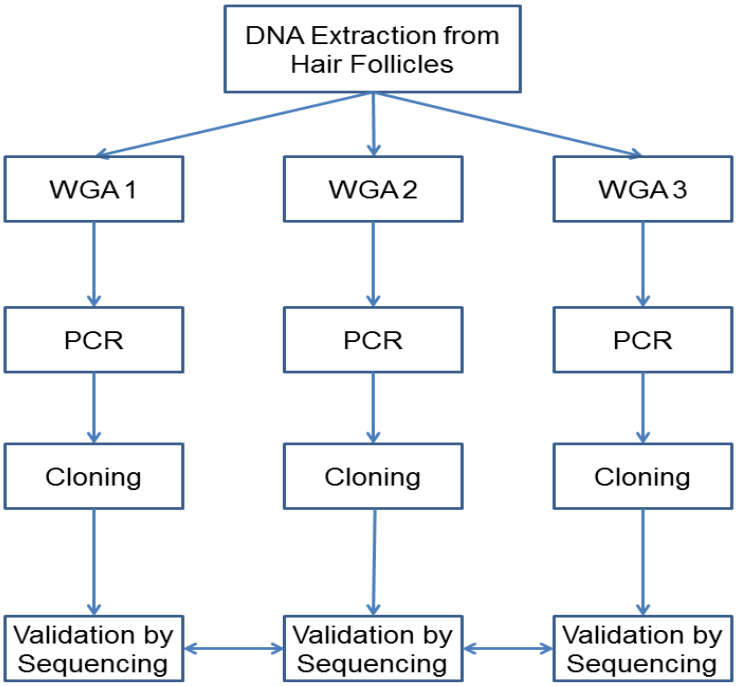

Figure 2: Validation Protocol
DNA from a single extraction is used to create three simultaneous WGAs. Material from these WGAs are amplified, cloned, and sequenced. Validation of this method was to be established from resulting sequences by calculating the frequency of polymorphisms due to amplification error.

For each base, polymorphisms occurring at a frequency of 1/6 or greater among replicates from all WGAs were to be considered the result of allelic heterozygosity. Polymorphisms occurring at a frequency of 1/6 or greater among replicates from two WGAs

but not in the third would be considered likely to be the result of heterozygosity with an allelic drop, while polymorphisms present at this frequency only in replicates from one WGA were to be considered likely due to error early in amplification and possible random sampling error. Polymorphisms occurring at a frequency below 1/6 in all three WGAs were to be considered as a result of a point mutation in a gene copy before genome amplification took place. Lastly, polymorphisms occurring at a frequency below 1/6 among replicates of only two WGAs or one WGA would be considered as errors from the amplification process. The total occurrences of polymorphism due to amplification error and inconclusive variation were to be tallied to determine the rate of error in the amplification process.

Results:

PCR Amplification and Limit of Detection

The results of the Limit of Detection PCR are displayed in **Figure 3**.

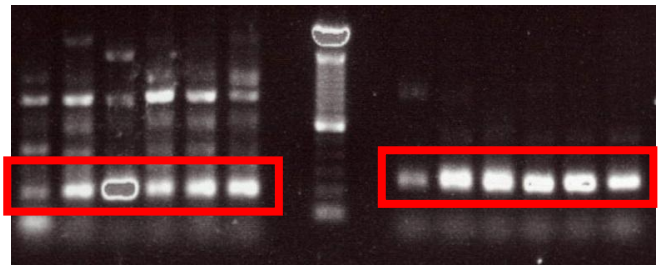

Figure 3: 1% Agarose Gel of Original PCR Amplification Products
The left side of the image contains the samples amplified using the "D4" primer, arranged in the following pattern: "Lance" 30 hair sample, 13 hair sample, 7 hair sample, "Ebi" 13 hair sample, 8 hair sample, 5 hair sample. The right side of the image contains the samples arranged in the same pattern amplified using the "D18" primer. The middle lane contains a 100bp ladder.

In all samples for both primers, the desired fragment was successfully amplified, evidenced by the succession of bright bands at approximately the 270 bp mark for primer

"D4" and 290 bp for primer "D18." Thus, the limit of PCR detection determined for DNA extraction from hair is 5 hairs. It is important to note a large amount of nonspecific amplification in the "D4" samples, evidenced by multiple bands occurring in each lane. This indicates the amplification of several other fragments besides the intended target of primer "D4."

Validation Protocol

The results of the Validation Protocol PCR are displayed in **Figure 4**.

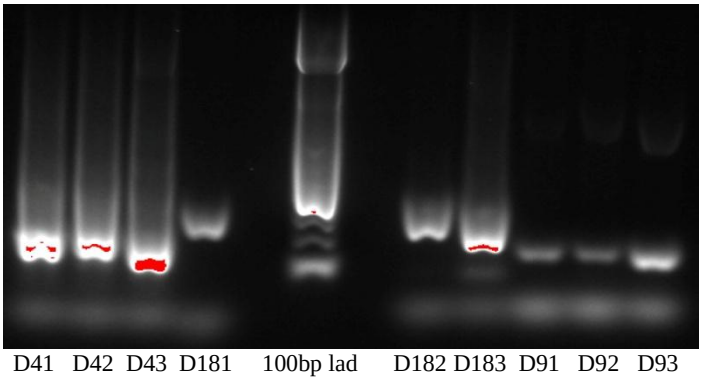

Figure 4: 1% Agarose Gel of Validation Protocol PCR Products
Samples "D41," "D42," and "D43" are amplifications using primer "D4" from WGAs 1, 2, and 3 respectively. Samples "D181," "D182," and "D183" are amplifications using primer "D18" from WGAs 1, 2, and 3 respectively. Samples "D91," "D92," and "D93" are amplifications using primer "D9" from WGAs 1, 2, and 3 respectively. A 100bp ladder is used for size estimation.

PCR amplification was observed for all nine sample types for use in cloning and sequencing. The streaking present in the six "D4" and "D18" lanes was attributed to an overload of genetic material. All three sample types were meant to be cloned and sequenced, but only enough TOPO vector was available for the "D4" and "D18" samples.

Sequences obtained from the validation protocol were not of high enough quality to align the replicates, both within and between WGAs. Chromatogram data was too ambiguous to make accurate base calls, and there was a large amount of unaccounted variation between

sequences from the same sample. The sequences of the three highest quality replicates are shown in **Figure 5.**

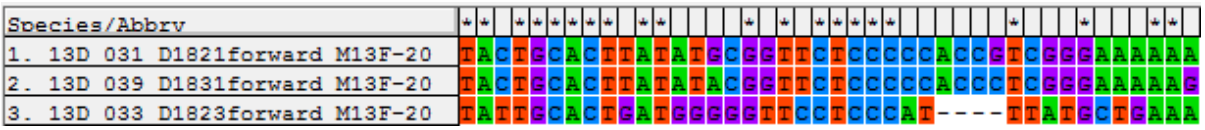

Figure 5: Alignment of the three highest quality sequences obtained
Sequences match at only 21/42 (50%) bases.

These replicates only matched at 50% of their bases. Without a proper alignment, sequences could not be compared for polymorphism frequency.

Discussion:

The results of the original DNA extractions, whole genome amplifications, and PCR visualization met the goals of our first aim by providing evidence of robust amplification of samples testing the lower limit of detection. Based on these results, we can conclude that a DNA extraction using as few as 5 hairs is sufficient for PCR amplification. There was a large amount of nonspecific amplification seen in **Figure 3** among the "D4" samples, but these are not due to faults in the extraction process and can easily be prevented in future runs by altering the PCR reaction conditions. This could include increasing the annealing temperature to prevent excess binding to undesirable fragments, changing the Magnesium concentration to affect the DNA polymerase activity, or reducing the extension time to favor the amplification of shorter fragments. Nevertheless, this low limit of detection has important implications on the application of this method in the field, as high quality hair samples are often hard to come by in the wild (Jeffery 2007). Extractions requiring a large amount of hairs would therefore be a major constraint to feasible use of this technique, so the low limit of detection obtained by this study supports the efficacy of this method.

Despite this initial success, it is important to note some limitations of these results. First off, our testing used only hairs with whole follicles attached. In the wild, most of hair samples collected have been shed from the animal and have often been detached from the follicle (Jeffery 2007). These samples typically contain significantly lower DNA content than "plucked" hairs including the follicle due to a lack of cellular material (Jeffery 2007). Additional testing needs to be done on samples containing only hair shafts without roots, and it is predicted that the limit of detection would be much higher for these kinds of samples. Additionally, while basic PCR amplification was robust, we have not yet accurately quantified the DNA yield of the extraction to determine if it is sufficient for whole genome sequencing. Modern sequencing technologies require relatively little template DNA to operate efficiently, the best devices needing only a few pg of starting material, but obtaining this quantity is not necessarily guaranteed, as previous quantification studies have reported the average DNA content of hair samples to be 4.4 pg/μl (Head 2014, Morin 2001). However, our inclusion of a whole genome amplification step in our protocol aims to provide enough material necessary for using these techniques, and samples amplified with a REPLI-g kit have proved effective for use in whole genome sequencing (Young 2012). Therefore, DNA obtained from our extraction method should be quantified using quantitative PCR (qPCR), using a technique developed for hair samples by Morin and colleagues (2001), to ensure the required amount of DNA is present in our samples.

Including a whole genome amplification step in our protocol is certainly useful in obtaining enough DNA for a variety of genomic applications. Yet its use can be hazardous if a moderate rate of amplification error occurs, preventing access to accurate sequences. This issue was the focus of our second aim: validation of our extraction method by comparing

sequences obtained from three identically handled samples for the frequency of amplification artifacts. Unfortunately, the sequences returned by our study were not of high enough quality to validate the consistency of this method, as an inordinate amount of variation and presence of non-allelic sequences prevented proper alignment of the sequences and the establishment of a reference genotype. This was not likely due to excessive error in the whole genome amplification, but due suboptimal PCR conditions causing nonspecific amplification. We base this conclusion on the occurrence of many conflicting chromatographic peaks at the several locations in both sample types, which are most likely the result of possible amplification of non-allelic sequences.

Errors from whole genome amplification would not cause this faulty sequencing, because even if mutations are introduced into the sequences during amplification, only one kind of fragment should be transformed into the host *E. coli* if the DNA template is relatively pure. Consequently, while it's very possible to observe base variation between replicates as a result of WGA error, this should never cause multiple fragment types to be in the same replicate. This seems to have occurred in samples, as many of the replicates returned sequences that conflict at every single base, which indicates a likely presence of two distinct fragments. This would likely stem from the amplification of nonspecific fragments in the original PCR which could then also be transformed into the *E. coli* and subsequently render the resulting sequences unreadable.

Further interpretation of our results has determined that the smearing present in **Figure 4** was not due to overload of genetic material but to an insufficient annealing temperature for samples "D4" and "D18" causing this nonspecific amplification. An annealing temperature of 58^{o}C for this reaction was used to include the "D9" sample in the

sequencing protocol, to showcase the effectiveness of the extraction on a variety of loci, but this sample set was never sequenced due to a lack of enough TOPO vector. In hindsight, using this lower annealing temperature for all of the samples was misguided, as the "D4" and "D18" samples have been optimized using a temperature of 67^{o}C. Therefore, we project that optimizing the PCR conditions for the "D4" and "D18" samples and carefully redoing the cloning and sequencing portion of our protocol should easily return high-quality sequences for both primer sets.

Given that high-quality sequences should be readily obtainable, it is useful to explain how such results will be interpreted. Since the DNA used for sequencing in each replicate group was extracted from the same source and amplified using the same primers, the theoretical sources of polymorphism in these sequences can either be normal allelic variation or errors introduced by amplification. Allelic polymorphism should hypothetically be present in a 1:1 ratio for heterozygous individuals, given that the two alleles were amplified equally, but the maximum theoretical probability of a single polymorphism due to amplification error is 1/6. This is due to the diploid nature of mammal DNA; an amplification error is a single event in only one of the three WGAs, and within that WGA, it only happens on one of two alleles. Thus, an error is expected in on only one of the six alleles present in the very first round of genome amplification across the three simultaneous WGAs. Keeping this probability in mind, we define our interpretations of different polymorphisms in **Table 1** based on the frequencies at which they occur in all sequence replicates and the number of WGA sample sets they occur in. It is important to note that these are ideal predictions, and that we expect deviation from these exact frequencies due to random sampling from a finite sample pool.

Table 1: Interpretations of Different Categories of Polymorphism

Total Frequency	**Number of WGAs**	**Interpretation**
≥1/6	3	Heterozygosity
≥1/6	2	Heterozygous with allelic dropout
≥1/6	1	Early amplification error, random sampling
<1/6	3	Point mutation in original DNA, 3rd sequence type
<1/6	2 or 1	Late amplification error

Polymorphisms occurring at an overall frequency ≥1/6 in replicate sets of all 3 WGAs must be due to heterozygosity unless 3 separate amplification errors all occurred in same spot, a very unlikely event. Polymorphisms occurring at ≥1/6 frequency in replicates of 2 WGAs but not the third is most likely also heterozygous with an allelic dropout. Allelic dropout happens when an allele is missed during amplification of small amounts of DNA, resulting in false reports of homozygosity (Taberlet 1999). Probabilistically speaking, one allele getting left behind is more likely than two separate amplification errors in same spot. Polymorphisms occurring at ≥1/6 frequency in only 1 WGA are most likely due to error early in amplification, as this is the maximum frequency at which amplification error can be observed per locus. Probability also tells us that getting a frequency of 1/6 or more in replicates of only one WGA is more likely an artifact of random sampling error than the other option of heterozygosity with two allelic dropouts.

Moreover, polymorphisms occurring at <1/6 total frequency and present in all three WGA replicate sets is most likely due to amplification of a mutated allele in the original DNA extraction. This third sequence may occur before amplification takes place, and can be present at a low frequency throughout all the WGA sequence sets. This is considered more likely than the occurrence of 3 identical but separate amplification errors. Finally, polymorphisms occurring at <1/6 total frequency and present in either 2 WGAs or 1 WGA

are interpreted as errors occurring late in amplification, as they are only in a small proportion of replicates. We can be fairly certain that a low-occurring polymorphism in replicates of only one WGA is an amplification error due to its scarcity. However, for low-occurring polymorphisms present in replicates from two WGAs, the interpretation is less conclusive, as it could also be due to a point mutation before amplification with a subsequent allele dropout. Regardless, while it is helpful to set out an interpretative framework for this validation, we cannot know how it will actually be applied until we obtain sequences of high enough quality for alignment and scoring. This framework is meant to be flexible, as the binomial sampling of the different cloned sequences is expected to be subject to random deviation. We plan on implementing this framework in our analysis once quality sequences are obtained.

Once we are able to validate the consistency of this extraction method, additional steps will be taken to investigate whether the genetic material obtained is sufficient for genomic sequencing and other applications. After initial quantification by qPCR, the next stage of development will be to run the extraction samples through a bioanalyzer, commercially available from Agilent Technologies, to determine the size of DNA reads produced by our method. Sufficient DNA read size is pivotal in modern evolutionary and population genetics studies using Genotyping by Sequencing (GBS) to determine the distribution of genotypes over a wide sample size (Elshire 2011). Enabling investigators to use noninvasive hair samples for this technique will have a profound effect on the efficacy and implementation of these studies on wild populations. Furthermore, once we've determined that our extraction protocol is consistent on a DNA-fragment level, it will be time to expand its scale and attempt using samples for genomic library sequencing. This could be done using a modified approach to the DNA capture and shotgun sequencing method

developed by Perry *et al.* for fecal samples (2010). Sequencing an individual's genome from hair collected in the field would revolutionize the way modern genetic and genomic analyses are conducted, providing a safe, efficient, and noninvasive manner to access the invaluable data provided by these sequences.

While this study has chosen to focus on DNA extraction from hair, it is necessary to consider other noninvasive sources of genetic material which may also be useful to genomic study. Urine has recently been used as another potential source of genetic material collected from wild subjects (Inoue 2007). It confers many of the same advantages over fecal collection as hair, such as reduced contamination from microbial sequences, and can also be collected relatively easily without intervention on the study animal (Knott 1997). Additionally, the concentration of DNA available in collected urine has been shown to be comparable to that from fecal samples, which may make it even more potentially useful than hair (Inoue 2007). We propose a study similar to the one here be conducted using collected urine samples, beginning with DNA isolation using a commercial kit available from Norgen Biotek Corp. or a comparable product, followed by similar whole genome amplification and validation sequence analysis. It is possible that the greater quantity of DNA available in urine might make it more suitable for whole genome amplification, and would thus deliver more robust results in downstream genomic applications.

Overall, the extraction of DNA from noninvasive samples for use in genomic sequencing will be crucial to advancement in modern studies on natural populations. Recent developments have made the collection of these samples safer and more efficient, and the data provided by these studies applies to all fields of animal research, including genetic, evolutionary, behavioral, phylogenetic, and population studies (Amendola-Pimenta 2009,

Bjork 2011, Fisher & Marcus 2006, Telfer 2003, Morin 1994). The overall implication of our research is not the development of a dominant method for sample collection, but rather an addition to the tools available to researchers to access high quality genomic data. We predict that DNA extraction from hair for use in genomic sequencing will be an effective method of gathering this data noninvasively, but this is not to say that samples derived from feces or from urine could be potentially as valuable. Our overarching goal is to provide the maximum amount of adaptability to researchers collecting samples in the field, so that they may access genomic data from whatever sources are readily available. The successful development of extraction protocols from these noninvasive samples will further reduce the need for intervention on study animals for data collection, affirming the efforts of the scientific community towards the conservation and support of the endangered ape species. Lastly, it is our hope that such extraction methods will be applied beyond the study of apes to all mammal groups, promoting a culture of safe and effective research on these vital members of our biosphere.

References:

Amendola-Pimenta, M., *et al.* (2009). Noninvasive collection of fresh hairs from free-ranging howler monkeys for DNA extraction. *American Journal of Primatology* 71(4) : 359-363.

Bjork, A. *et al.* (2011). Evolutionary History of Chimpanzees Inferred from Complete Mitochondrial Genomes. *Mol. Biol. Evol.* 28(1): 615–623.

Constable, J.J. *et al.* (1995). Nuclear DNA from primate dung. *Nature* 373: 393.

Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). Resolution Conf. 9.6 (Rev. CoP16): Trade in Readily Recognizable Parts and

Derivatives. Ninth meeting of the Conference of the Parties; 1994 07 - 18 Nov.; Fort Lauderdale (United States of America).

Elshire, R.J. *et al.* (2011). A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. *PLoS ONE* 6(5): e19379.

Fisher, S.E. and G.F. Marcus (2006). The eloquent ape: genes, brains, and the evolution of language. *Nature Reviews Genetics* 7: 9-20.

Gao, F. *et al.* (1999). Origin of HIV-1in the chimpanzee *Pan troglodytes troglodytes*. *Nature* 397: 436-441.

Goodall, J. (1986). *The Chimpanzees of Gombe: Patterns of Behavior.* Cambridge, Mass.: Belknap Press of Harvard University Press.

Hahn, B. *et al.* (2000). AIDS as a zoonosis: Scientific and public health implications. *Science* 287(5453): 607-614.

Hahn, B. and P. Sharp (2011). Origins of HIV and the AIDS Pandemic. *Cold Spring Harb Perspect Med*. 1(1): a006841.

Head, S.R. *et al.* (2014). Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques* 56(2):61–77.

Inoue, E. *et al.* (2007). Wild chimpanzee infant urine and saliva samples noninvasively usable for DNA analyses. *Primates* 48: 156-159.

Jeffery, K. J., *et al*. (2007). Biological and environmental degradation of gorilla hair and microsatellite amplification success. *Biological Journal of the Linnean Society* 91: 281–294.

Kohn, M. H. (2010). Noninvasive genome sampling in chimpanzees. *Molecular Ecology* 19: 5328–5331.

Knott, C.D. (1997) Field collection and preservation of urine in orangutans and chimpanzees. *Tropical Biodiversity* 4(1): 95-102.

Marrero, P. *et al.* (2009). Extraction of high-quality host DNA from feces and regurgitated seeds: a useful tool for vertebrate ecological studies. *Biological Research* 42(2): 147-151.

Morin, P.A. *et al.* (1994) Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science*, 265:1193–1201.

Morin, P.A. *et al.* (2001). Quantitative polymerase chain reaction analysis of DNA from noninvasive samples for accurate microsatellite genotyping of wild chimpanzees (*Pan troglodytes verus*). *Molecular Ecology* 10(7): 1835-1844.

Perry, G. H. *et al*. (2010). Genomic-scale capture and sequencing of endogenous DNA from feces. *Molecular Ecology* 19: 5332–5344.

Price, E.O. (1999) Behavioral development in animals undergoing domestication. *Applied Animal Behavior Science* 65(3): 245-271.

Siddiqui H. *et al.* (2011). Assessing diversity of the female urine microbiota by high throughput sequencing of 16S rDNA amplicons. *BMC Microbiology* 11: 244.

Smuts, A.L., and P.D. Pogue. DNA from urine as a potential source of identification. Poster session presented at: 10th International Symposium on Human Identification; 1999 Sep 29 – Oct 2; Colorado Springs, CO.

Taberlet, P., Waits, W.P., and G. Luikart (1999). Noninvasive genetic sampling: look before you leap. *Trends in Ecology and Evolution* 14(8): 323-327.

Tamura K. *et al.* (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* 28: 2731-2739.

Telfer, P.T. *et al.* (2003). Molecular evidence for deep phylogenetic divergence in *Mandrillus sphinx*. *Molecular Ecology* 12: 2019-2024.

The Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437: 69-87.

Von Beroldingen C.H. *et al*. (1987). Analysis of enzymatically amplified HLA-Dqalpha DNA from single human hairs. *American Journal of Human Genetics* 41:725.

Wasser, S.K. *et al.* (1997). Techniques for application of faecal DNA methods to field studies of Ursids. *Molecular Ecology* 6: 1091-1097.

Young, N.D. *et al*. (2012). Whole-genome sequence of *Schistosoma haematobium*. *Nature Genetics* 44: 221.

Supplementary Data

Sequence data for primers "D4" – D4s243, "D18" – D18s851, and "D9" – D9s910

Primer	**Sequence**
D4s243 Forward	5' – TCAGTCTCTCTTTCTCCTTGCA
D4s243 Reverse	5' – TAGGAGCCTGTGGTCCTGTT
D18s851 Forward	5' – CTGTCCTCTAGGCTCATTTAGC
D4s851 Reverse	5' - TTATGAAGCAGTGATGCCAA
D9s910 Forward	5' - AAGTCAGTTAGCTGAAGGTTGC
D9s910 Reverse	5' - TATATGAAGTGCTTAGAAAAAGTGC