TECHNICAL SUMMARY

Project Title: **Mitogenome Sequencing of 10,000 Population Samples Using a Highly Accurate, Single-Molecule Approach**

Penn State University
Principal Investigator: Mitchell M. Holland, Professor

Co-Principal Investigator: Jennifer A. McElhoe, Associate Research Professor

National Institute of Justice
Research and Development in Forensic Science for Criminal Justice Purposes
NIJ Award: 15PNIJ-22-GG-04395-DNAX
Award Amount: $771,418
Award Start Date: January 1, 2023
Original Award End Date: December 31, 2024
No-cost Extension: December 31, 2025

Submitting Official: Mitchell M. Holland, Professor, Forensic Science Program, Biochemistry & Molecular Biology Department, Contact Information: 014 Thomas Building, , University Park, PA 16802; email address mmh20@psu.edu

Report Submission Date: April 30, 2026
Recipient Organization: Penn State University, University Park, PA 16802

Abstract:

High-quality, human mitochondrial genome (mitogenome) search databases are an essential component of forensic investigations when providing weight estimates. The European DNA Profiling group (EDNAP) Mitochondrial DNA Population database (EMPOP) is considered the gold standard for this purpose, serving as a reference database and a quality-control tool. The current study developed a sequencing pipeline that is robust and cost effective for uploading more than 10,000 mitogenome sequences to EMPOP. Batches of whole blood or buffy coat samples were extracted using the Zymo Research Quick-DNA Miniprep Plus kit. Amplification of the mitogenome was performed using two overlapping long-range amplicons of approximately 8.5 kb. Amplicons from 372 samples, plus eight DNA extraction reagent blanks and four amplification negative controls, were normalized and pooled using SequalPrep plates. A library of amplicons was prepared by ligation of SMRTbells (single molecule, real time adaptors), and prepared libraries were run on the PacBio Sequel IIe instrument using a high-fidelity (HiFi) approach. The total time for laboratory processing of 384 samples, prior to SMRTbell ligation, was up to 62 working hours. Total cost of reagents and supplies for all steps was approximately 20 U.S. dollars (USD) per sample. Including labor, the cost was approximately 30 USD. The success rate for the 10,394 samples analyzed was 98.2%, when one of the two target amplicons was considered a failure to produce suitable sequence data. On a per amplicon basis, the success rate was 99.2%. Concordance studies using two short-read sequencing methods confirmed the reliability of the long-read approach. The long-read pipeline can be easily adopted by laboratories and used in high-throughput studies involving quality biological samples to generate large mitogenome databases, including those for upload to EMPOP. A manuscript covering the pipeline and analysis of more than 10,000 mitogenome sequences has been submitted for publication. A second manuscript covering the analysis of observed heteroplasmy is in preparation.

Major Project Goals:

Massively parallel sequencing (MPS) approaches have advanced to the point where full mitogenome sequences can be generated routinely from as little as 1-5 mm of human hair shaft (2019-DU-BX-0045, Canale et al., 2022 and Korber, et al., 2023). However, large population databases of mitogenome sequences for forensic purposes are not yet available, reducing the value of an mtDNA profile match. The outcomes of the proposed work included the generation

of highly accurate mitogenome haplotypes from at least 10,000 population samples; a detailed analysis of the population data, including the rate and location of heteroplasmy through the analysis of deep-coverage data; and evaluation of concordance to short-read mitogenome MPS data on a subset of samples sequenced using the long-read approach. These outcomes addressed important interests of the criminal justice system and National Institute of Justice; specifically, providing the forensic community with better statistical power when considering mitogenome haplotype matches between evidence and reference samples in forensic cases, and providing a high throughput, cost effective method for mitogenome sequencing. Partners for the project included the Institute for Preventative Medicine, Hershey Medical Center; Penn State Neuroscience Institute, College of Medicine, Hershey, PA; the Institute of Legal Medicine, Medical University of Innsbruck; Mitotyping Technologies, LLC, a division of SoftGenetics; and the Armed Forces DNA Identification Laboratory, Armed Forces Medical Examiner System. In addition, students from the Forensic Science program at Penn State University were involved throughout the project, contributing to efforts that foster the development of a highly skilled workforce.

Objectives:

1. Develop a ***Highly Skilled Workforce*** (STEM Education & Engagement).

Students at Penn State University were involved in all facets of the project; sample acquisition, extraction, and testing, data analysis, EMPOP uploads, and pipeline dissemination. Students from Penn State included Jasmeen K. Khosa, Claudia Prieto Alcaide, Erin D. Brownfield, Jade T. Korber, Miriam G. Foster, Rachel T. Cundey, and Valeria Garcia, as well as dozens of others who were informed of the project design and findings. Of the students listed, four have joined forensic DNA laboratories. In addition, an exchange student, Floor Claessens, from the Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, worked on the project for three months as part of a doctoral program.

2. Conduct a ***Population Study*** on at least 10,000 samples.

Samples were provided by the Institute for Personalized Medicine at Hershey Medical Center in Hershey, PA, the Penn State Neuroscience Institute, College of Medicine, Hershey, PA, and the Institute of Legal Medicine, Medical University of Innsbruck. Following DNA extraction of

blood and buffy coat samples, enrichment was accomplished through long-range amplification of the mitogenome in two long-range amplicons of approximately 8.5 kilobases using indexed primers, PacBio SMRTbell template preparation, and sequencing on the PacBio Sequel II instrument. Data analysis was completed using the NextGENe® LR software from SoftGenetics, a Dotmatics company.

a. Received blood and buffy coat samples, or DNA extracts, over the course of the project.
b. The study required 28 runs of 372 samples each, plus eight reagent blank and four PCR negative controls per four combined 96-well plates.
c. The maximum capacity of a HiFi PacBio Sequel II system run is 8,000,000 SMRT cells. The expected number of viable cells is 50-60% or at least 4,000,000 cells, with at least 2,000,000 producing high-quality, single-molecule data. Given that each cell contained a read of approximately 8.5 kilobases, read depths were consistently at least 2500 per sample. Along with the reported haplotype, this allowed for analysis of heteroplasmy across the mitogenome at a minor variant frequency of 10%.

3. Conduct a ***Detailed Analysis of Population Data***.

a. Document the haplotype for each sample.
b. Evaluate the analytical noise in each dataset, according to McElhoe 2020.
c. Evaluated the location and rate of heteroplasmy across the mitogenome, observed above the analytical noise. The approach used for generating mitogenome sequences allowed for the phasing of multiple heteroplasmic sites and provided a far better assessment of potential NUMTs (nuclear inserts of mitochondrial DNA sequences) that can be co-amplified with functional copies of the mitogenome.
d. Performed a quality assessment of mitogenome sequences associated with each sample for inclusion into the EMPOP forensic database, including the assessment of heteroplasmy.
e. Submitted mitogenome sequences to EMPOP for upload.
f. Reported the findings through presentations and a submitted publication.

4. Conduct a ***Concordance Study*** by running previously sequenced samples using a short-read approach and MiSeq MPS.

a. Compared haplotypes for concordance.
b. A percentage of the samples previously tested presented heteroplasmic sites that were reported as a part of the haplotype. These sites were compared to assess whether single-molecule results confirm the presence of heteroplasmy or reveal sites that were not reported with the short-read approach.
c. Reported the heteroplasmy findings through presentations and a publication in preparation.

Summary of Project Design & Methods:

The primary project objective was the development of a robust pipeline for mitogenome sequencing that produces high-quality, low-cost haplotypes and allows for the detection, resolution and reporting of heteroplasmy at a threshold above analytical noise. With this first objective accomplished, the pipeline was used to sequence more than 10,000 mitogenomes and submit them for upload to EMPOP (www.empop.online). In meeting these objectives, the discrimination power of mtDNA analysis in forensic casework when sequencing the mitogenome will be significantly improved. Enhanced haplotype frequency estimates, along with the possible inclusion of heteroplasmy in the statistical analysis (McElhoe et al., 2022), will benefit the forensic community and criminal justice system.

The PacBio technology has seen recent advancements that allow for highly accurate, single-molecule sequencing of large DNA fragments of 1000's to tens of 1000's of bases in length (Wenger et al., 2019; https://www.pacb.com/technology/hifi-sequencing/how-it-works/). The basic principles of the technology are outlined in Figure 1; from Eid et al., 2009 and reproduced in Rhoads & Au 2015. Single-molecule, real-time (SMRT) sequencing involves a copy of DNA polymerase, attached to a single DNA fragment, bound to the bottom of a zero-mode waveguide (ZMW) which is a nanophotonic structure that reduces the volume of the detection zone to approximately 20 zeptoliters (10^{-21} liters). As a result, the instrument is able to detect fluorescence associated with the incorporation of a single dNTP molecule. Individual nucleotides enter the detection zone and the emission wavelength of the fluorophore is detected during the incorporation process. As the phospholinked dNTP (i.e., a fluorophore attached to the terminal phosphate group) associates with the template, the fluorescence emission increases in intensity to a level that can be detected using a high-multiplex confocal fluorescence detection system (Lundquist et al., 2008). As the phosphodiester bond forms, pyrophosphate is released,

carrying the fluorophore away from the active site of the polymerase, and thus, the detection zone. The polymerase moves to the next position on the template and awaits the incorporation of the next base. Therefore, there is no need for an intermediate step that removes the fluorophore and/or reconstitutes the 3'-end of the strand being synthesized. Therefore, the PacBio SMRT method allows for highly accurate, uninterrupted, sequential long-read sequencing.

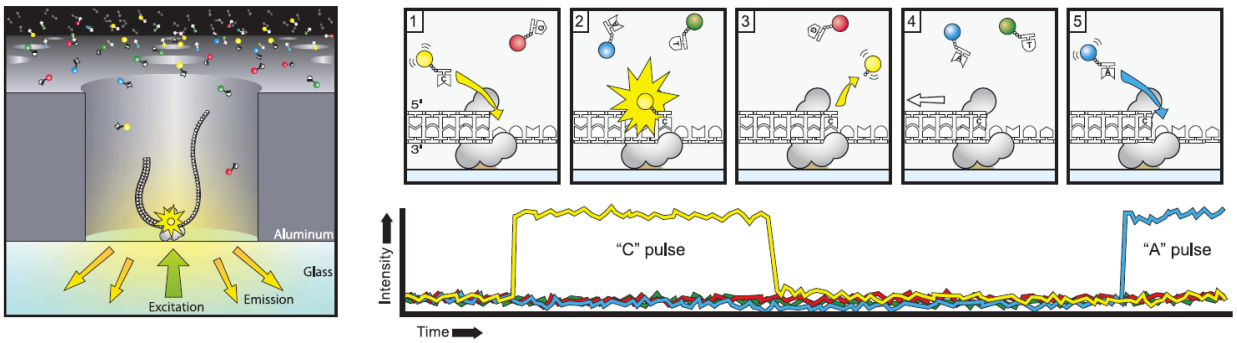

Figure 1: The basic principles of single-molecule, real-time (SMRT) sequencing.

A total of 10,394 blood samples and DNA extracts were received from multiple sources under individual internal review board (IRB) protocols. These include the biorepository at the Penn State Institute for Personalized Medicine located in the Hershey College of Medicine (3,366 whole blood and 4,916 buffy coat) collected under the Penn State University IRB approved protocol #40532, 612 buffy coat samples from the Penn State Neuroscience Institute collected under PSU IRB approved protocol #STUDY00002914, and 1,500 blood extracts from the Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria collected under ethical committee of Innsbruck Medical University UN2598, session number 241/4 and underwritten informed consent. DNA extraction, if required, amplification, and sequencing were performed under IRB approved protocol #STUDY00023018. Whole blood (50 µL) and buffy coat samples (25 µL) were extracted using the Quick-DNA Miniprep Plus Kit (Zymo Research, cat. D4069) according to the manufacturer's recommended protocol. The final volume of DNA extract was 50 µL in all cases. Extracts were stored at -20°C for up to six months in screw top cryo-storage tubes (VWR cat. 89004-286) prior to being subjected to long-read sequence analysis. At approximately two hours per set of 23-24 samples (when considering extraction reagent blanks, RBs), the extraction process took up to 40 hours to complete for each run of 372 whole blood or buffy coat samples, and eight RBs.

Quantification was performed on the Qubit 4 Fluorometer using the Qubit 1X dsDNA HS Assay Kit according to the manufacturer's recommended protocol (Thermo Fisher Scientific Invitrogen, cat. Q33231). A relatively small subset of sample extracts was quantified, as it was determined that a broad range of DNA concentrations [1-100 ng/µL of total (tot) DNA template, assuming DNA content from drawn blood contains exclusively human DNA] could be amplified with consistent, quality outcomes. Therefore, quantification was not a required component of the pipeline, but instead, was used early in the development of the pipeline to determine the general range of totDNA template recovered per extract, and to spot-check the extraction process.

The following amplification protocol is suitable for any quality DNA extract that allows for the input of approximately 4-400 ngs of totDNA template. One set of 93 sample extracts, two RBs, and one no-template control (NTC) was run over a two-day period. Four sets of 96 samples and controls were prepared for each sequencing run on the PacBio (384 total samples and controls). Each set of amplification reactions required up to 3 hours of setup time in the laboratory, plus approximately 7 hours of time in the thermal cycler, so up to 12 hours of laboratory time and 28 hours in the thermal cycler per run. Each set of 96 samples resulted in two long-range amplification plates, with the A and B overlapping primer pairs covering two, approximately 8.5 kb amplicons across the mitogenome; the A primer pair consisted of F2480 (5'–AAATCTTACCCCGCCTGTTT-3') and R10858 (5'-AATTAGGCTGTGGGTGGTTG-3'), while the B primer pair consisted of F10653 (5'-GCCATACTAGTCTTTGCCGC-3') and R2688 (5'–GGCAGGTCAATTTCACTGGT-3')(Fendt et al, 2009). Amplification master mix was comprised of 42.4 µL of TaKaRa LA Taq® Hot Start polymerase (TaKaRa, Bio, cat. RR042A, 5 U/µL, final of 0.04 U/L), 530 µL of 10X TaKaRa LA PCR Buffer with 25 nM Mg^{2+} (final of 1X & 2.5 mM Mg^{2+}), 424 µL of TaKaRa dNTPs (2.5 mM each, final of 0.2 mM), 530 µL of BSA (20 mg/mL, final of 0.002 mg/µL), and 2501.6 µL of deionized MolBio Grade water, for a total of 4,028 µL. A 19 µL aliquot of master mix was added to each well of two 96-well plates. Plates of A and B custom primer pairs with different barcode combinations were acquired from Integrated DNA Technologies (5 µM each primer), with 2 µL of each primer pair added to the respective plates (final concentration of 0.4 µM for each primer pair). Lastly, a total of 4 µL of each DNA extract (typically 4-400 ngs of totDNA template), reagent blank or deionized water blank was added to their respective wells for a final volume of 25 µL. Reactions were run on an Applied Biosystems GeneAmp® PCR System 9700 using the following parameters: 93°C for 3 mins; 14 cycles of 93°C for 15 sec, 60°C for 30 sec, 68°C for 5 mins; 20 cycles of 93°C for 15

sec, 55°C for 30 sec, and 68°C for 9 mins with an increase of 10 seconds for each cycle; store at 4°C.

Post amplification, plates A and B were combined and normalized using a 96-well SequelPrep Normalization Plate (Invitrogen, cat. A1051001) according to the manufacturer's recommended protocol; 25 µL of the combined 50 µL of amplification products was added to the normalization plate, with the balance stored at -20°C. The normalized samples were combined (20 µL each, or approximately 1.92 mL of total volume) and concentrated with an Amicon Ultra 0.5 mL microcon Ultracel filter with a 100,000 MW cutoff (Merck Millipore, cat. UFC510096) to a volume of approximately 120 µL. Each of the four sets of 96 samples were combined (approximately 480 µL total volume), concentrated to a volume of approximately 120 µL. Products for each individual pool (approximately 5 µL), along with the combined pool, were assessed via a 4150 TapeStation instrument according to the manufacturers protocol. The four sets of normalizations required up to 10 hours of time in the laboratory, including incubation steps.

Twenty-eight libraries of 384 PCR-indexed samples each were prepared using the PacBio SMRTbell prep kit, version 3.0, according to the manufacturer's protocol: including extended repair and A-tailing of 1 hour incubation versus 30 minutes, adapter ligation and cleanup, and nuclease treatment and cleanup. An additional cleanup was performed, using the whole genome and metagenome library preparation guide to remove DNA fragments less than 5 kb. Sequencing on the PacBio Sequel IIe was performed using the SMRT Link user guide, version 13.0, the Sequel IIe Systems, and the SMRT sequencers operations guide, according to the manufacturer's protocols. SMRT cells 8M were used for each run. Below is a figure that illustrates the SMRTbell structure.

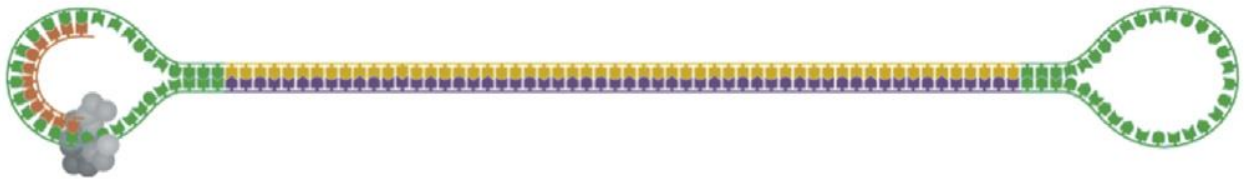

Figure 2: Hairpin adaptors are highlighted in green. The adaptors are ligated to the ends of the amplicon forming a SMRTbell. Also depicted is a copy of DNA polymerase attached to the left-most end of the molecule (Rhoads & Au 2015).

The single copy of DNA polymerase, bound to a single molecule of DNA, allows for multiple passes of the polymerase across the template of ~8.5 kb. Figure 3A (from Logsdon et al., 2020) illustrates how the polymerase continues to pass through the single-stranded template, creating multiple copies of mitogenome sequence. This is a driving factor in the accuracy of the process, as the end-result is a consensus sequence of multiple reads; termed circular consensus sequencing (CCS), versus a single pass which is termed continuous long reads (CLRs). As seen in Figure 3B, the CCS approach improves accuracy by at least one to two orders of magnitude. Most importantly, the average read accuracy routinely exceeds 99%. Numerous studies have provided evidence that long-read sequencing using the CCS approach provides the level of accuracy required to report reliable genomic sequences (for example, Wenger et al., 2019).

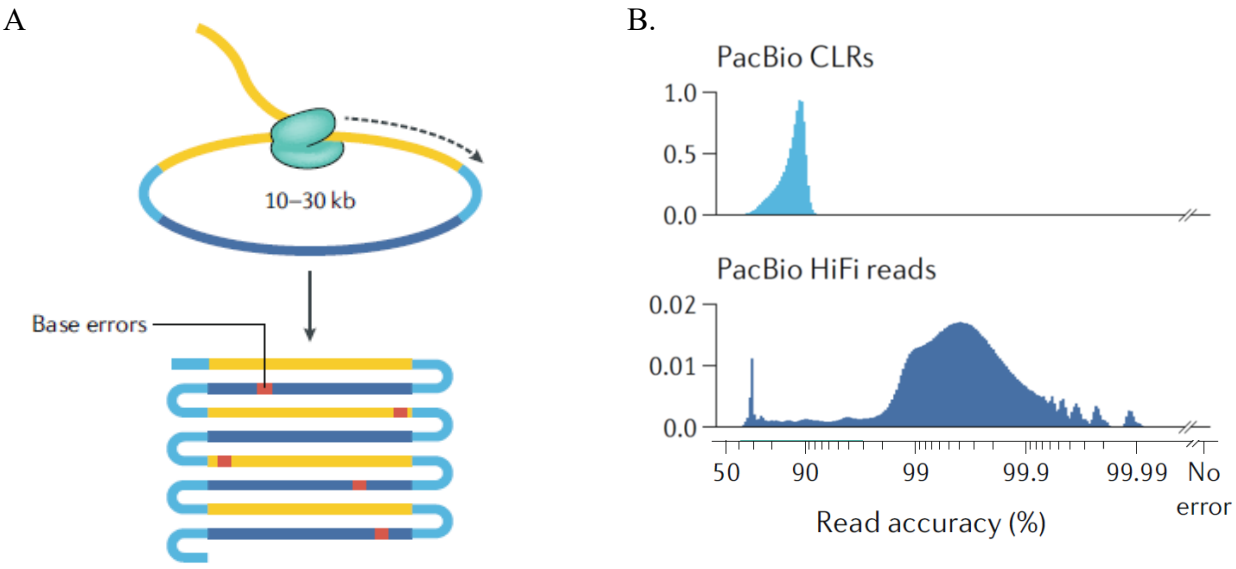

Figure 3: A) Illustrates the circular consensus sequencing (CCS) approach where the polymerase passes through the circularized amplicon multiple times allowing for a significant increase in accuracy. B) An illustration of the increased accuracy when comparing continuous long read (CLR) sequencing to CCS (PacBio HiFi reads).

Each run on the PacBio instrument resulted in unaligned BAM files which were used for secondary analysis with GeneMarker HTS (GM HTS) long-read software (GeneMarkerHTS_v2.6.0-beta2-44.16+20ab6c05c.ci; SoftGenetics, a Dotmatics Company, PA). Alignment to the revised Cambridge reference sequence (rCRS) (Anderson et al., 1981 and Andrews et al., 1999) produced haplotypes that were evaluated by the EMPOP SAM2 algorithm (Huber et al., 2018) to confirm whether they were phylogenetically plausible and to assess sites

of heteroplasmy. A bioinformatic pipeline was created that consisted of a step-by-step process that applied Linux Bourne Again SHell (Bash) v3.2.57(1), R v4.3.1, and Microsoft Office to manage and analyze raw data through EMPOP submission. To summarize:

1. **BAM file alignment - GM HTS** (v2.6.0-beta2-44.16+20ab6c05c.ci): PacBio Sequel IIe unaligned BAM files were uploaded for secondary analysis using the long-read GM HTS platform. All sequence data was mapped to the rCRS using the GM HTS alignment algorithm which employs an optimal phylogenetic alignment of data from PacBio instrument. For this study, unaligned BAM files were mapped to the rCRS using the following settings in GM HTS: variant percentage = 10; variant allele coverage = 40; total coverage = 200; allele score difference = 63; allele balance ratio = 2.5 for single nucleotide polymorphisms (SNPs) and = 5.0 for insertions and deletions (indels); and identity percent = 85. A custom motif file was used to assist with proper reporting of haplotypes associated with phylogenetically valid alignments. Default settings were applied for remaining options. SNPs present in 50% or more of the reads were considered major variants and collectively referred to as the major haplotype. Minor variants, present in less than 50% of the reads, represented sites of potential heteroplasmy. A 10% reporting threshold was used for minor variants. Given a 10% reporting threshold and the minor variant read coverage set at 40, each position required a minimum coverage of 400 reads to report variants ≥10%. Variants observed at a higher percentage could be reported with lower read depth. GM HTS analysis produced several output files including a consensus statistic file, which lists the number of reads for each nucleotide position (np), and a variant file, which includes the major haplotype, the minor haplotype, and the combined haplotype with International Union of Pure and Applied Chemistry (IUPAC) designations for mixed positions.
2. **Coverage evaluation**: **GM HTS** consensus statistic files were used to assess coverage and identify samples with nucleotide positions (nps) under the minimum coverage threshold of 200 reads. A combination of R and Bash was used to combine the coverage information for each PacBio run on a per sample basis and calculate the maximum, minimum, and average coverage of the A and B amplicons. Samples with coverage under the 200 read threshold for either A or B, or for both amplicons, were removed from further analysis. Negative controls were also assessed for coverage.

3. **Haplogrouping - Haplogrep 3:** For each PacBio run, Linux commands were used to collate the non-IUPAC haplotypes from GM HTS variant files for samples exceeding the coverage threshold of 200 reads across all nps. Haplogroups were then generated using Haplogrep 3 and following the instructions provided on the following website (https://haplogrep.i-med.ac.at/).
4. **EMPOP submission - quality control checks:** GM HTS generated IUPAC haplotypes that include heteroplasmic positions were submitted to EMPOP following the instructions provided on the following website (https://empop.online/). Before data was submitted to the EMPOP database, samples were manually evaluated, interpreted for mtDNA sequence variation, and checked for quality and potential errors. After initial submission to EMPOP, the EMPOP SAM2 software unmasked sequencing errors and assessed whether the haplotypes were phylogenetically accurate. This ensured that haplotypes were reported in accordance with standardized nomenclature. EMPOP QC also looked at other aspects of the data such as formatting (i.e. range violations, rCRS bias, nomenclature errors), plausibility (i.e. lack of expected variants), novel sequence calls (i.e. unobserved transversions, insertions, deletions), heteroplasmy, alignment of sequences, potential artefacts based on previous EMPOP sequences, and a final check of transversions given their rarity. Comments generated through EMPOP QC were manually evaluated in the aligned sequenced data using GM HTS. Once all comments were fully addressed, a response to comments was returned to EMPOP for finalization and upload to the search database.
5. **Error assessment:** An error assessment was performed using the same methodology presented in McElhoe et al., 2020, with the addition of indel error evaluation. Briefly, a conservative estimation of the substitution and indel error rates based on the number of times each nucleotide with a frequency of <50% was called in relation to the total coverage for all calls were generated using GM HTS consensus statistic reports for each base call (A, C, G, T, del, ins) and each np (1-16569). This conservative estimate of error captures all positions of true heteroplasmy as well as errors from the sequencer, damage due to cytosine deamination or depurination, PCR artifacts, and/or contamination. The error values represent a percentage of the total reads (error calls/total calls x 100), which can also be described as the number of calls made in error per 100 nucleotides. The top 25 sites of error for each base call for each PacBio run (top 25*6 types of error*28 runs =

4,200 total) were analyzed. Concordance tables and heatmaps were generated using R v4.5.1, ggplot2 v4.0.0, reshape2 v1.4.5, and viridis (lite) v0.6.5.

A subset of samples (n = 188) processed with the PacBio protocol were subjected to two different mitogenome protocols for short-read sequencing at the AFMES-AFDIL (Armed Forces Medical Examiners System, Armed Forces DNA Identification Laboratory) to assess concordance. The first protocol employed long-range amplification for enrichment, which was initially described in Taylor, et al. 2020. The same primer pairs were utilized for the amplification step as above (Fendt et al., 2009) with different conditions as described in Peck et al. 2018. A standard input of 4 µL DNA extract was added to the reactions. All post-enrichment sample preparation steps were performed on a Hamilton STARplus (Hamilton Robotics, Reno, NV) instrument. The amplicons were quantified on a 96-capillary Fragment Analyzer (Agilent Technologies, Santa Clara, CA), with amplicons A and B combined at equal concentration (ng/µL). The combined amplicon pools were purified with a 0.6X AMPure XP (Beckman Coulter, Brea, CA) purification and quantified with a Varioskan LUX instrument for a targeted input of 150 ng into KAPA HyperPlus (Roche Sequencing, Santa Clara, CA) library preparation. HyperPlus library preparation was performed in half-volume reactions and no additional library amplification step. NEXTflex-HT Barcodes (Bioo Scientific Corporation, Austin, TX), 12 nucleotide single indexed adapters for Illumina sequencing, were used in this protocol. After a 0.6X AMPure XP purification, the quality of the libraries was assessed on the Fragment Analyzer prior to library pooling. The library pools were created by combining one plate of libraries (96 samples including controls) in equal volume as normalization was performed at library input. The pools were quantified with the KAPA Library Quantification Kit for Illumina Platforms (Roche Sequencing) to ensure accurate loading concentration. Each pool was sequenced with a 600-cycle MiSeq Reagent Kit v3 (Illumina, San Diego, CA) using paired-end sequencing (301 x 301 cycles) on a MiSeq FGx Sequencing System (QIAGEN) in Research Use Only (RUO) mode.

A second short-read protocol utilized a small-amplicon approach with the Precision ID (PID) mtDNA Whole Genome Panel (Thermo Fisher Scientific) followed by library preparation to allow for sequencing on Illumina platforms. Amplification was performed according to a QIAGEN 2016 application note, with two PID multiplex primer pools amplified in separate reactions. Each reaction contained 12.5 µL QIAGEN Multiplex PCR Master Mix (2X), 2 µL PID mtDNA Whole Genome Panel Primer Pool (1 or 2), 5.5 µL deionized MolBio Grade water,

and 2 µL DNA extract. The following cycling conditions, which included an increase in cycle number from the recommended number for the QIAGEN Multiplex PCR Master Mix, were used on a Veriti 96-Well Thermal Cycler (Thermo Fisher Scientific): 95°C for 15 mins; 26 cycles of 94°C for 30 sec, 60°C for 90 sec, 72°C for 90 sec; 72°C for 10 min; store at 4°C. The post-enrichment sample preparation steps for the PID protocol were automated on the STARplus as described above with the following alterations. PID amplicons were quantified on the Varioskan LUX and the two multiplex products for each sample were combined in equal concentration. The PID amplicon pools were then purified with a 1.8X AMPure XP bead cleanup. To allow for Illumina sequencing, PID libraries were created using the KAPA HyperPrep Kit (Roche Sequencing) and eight-nucleotide TruSeq-compatible unique dual indexed full-length adapters (Integrated DNA Technologies, Coralville, IA). HyperPrep library preparation was used, as fragmentation was not required for short-read sequencing of PID amplicons. Unlike the LR HyperPlus protocol, the half-reaction HyperPrep PID libraries were subjected to an 8-cycle post-library amplification step following the manufacturer's protocol with the exception of half-volume reactions. The amplified PID libraries were purified with AMPure XP (1.0X) prior to pooling. All remaining steps were consistent with the LR MiSeq protocol including equal volume pooling, KAPA Library Quantification Kit used for pool quantitation, and paired-end sequencing (301 x 301 cycles) on a MiSeq FGx. MiSeq FGx fastq files were analyzed using the short-read version of GM HTS (v.2.5.5) (Holland et al., 2017) with the settings detailed in Holland et al., 2018, except for variant percentage which was set to 10% to be consistent with the variant threshold for PacBio sequenced samples. The major haplotypes were compared to long-read data from the PacBio instrument.

Summary of Results:

The success rate for the twenty-eight libraries was 98.2%, with a total of 189 failures out of 10,394 samples. Failures were considered as any sample where either the A or B amplicon sequence did not reach at least 200 reads at each np. On a per amplicon basis, the success rate was 99.2%. Successfully sequenced samples produced approximately 535 GB of unaligned BAM files and subsequently 650 GB of GM HTS data files. For the quality evaluation, metrics were compiled and considered from the PacBio sequencer and GM HTS, concordance data, and error assessment.

Valuable metrics include SMRT cell productivity, yield, read length, read quality, and sequencing control performance. Additional metrics (for example, CCS processing of double-stranded (DS) and single-stranded (SS) reads, and raw data reporting) were assessed to ensure run-to-run consistency and production of quality data. The SMRT Link onboard software generated a metrics report for 27 of the 28 runs, with data summarized for runs 1-3 and 5-28; run 4 was captured as raw data and not processed by the PacBio onboard software, therefore run metrics were not generated for this run. As expected, an average of 8,014,219 ± 765 active ZMWs were detected, illustrating the consistency between SMRT cells. ZMW productivity is defined as P0 (a ZMW that did not produce read data and is presumed to lack a copy of the polymerase), P1 (a ZMW with useful sequence data and low background noise), or P2 (a ZMW with no useful sequence data due to multiple template-polymerase complexes and/or high background noise). An optimally loaded SMRT cell has a P0 value >10% and a P1 value between 50-85%. A P0 value below 10% suggests that the cell is overloaded. The average percentage of P0 cells for the 27 runs evaluated was 22.4 ± 7.1%, with an average P1 value of 75.2 ± 6.5% and a P2 value of 2.3 ± 1.3% (Table S1) indicating quality sequencing results.

HiFi read length distribution plots were consistent with the observed physical library insert size for the two amplicons; A is 8,379 base pairs (bps) in length, B is 7,966 bps, with a difference of 413 bps (Figure 4).

In the figure, the approximate bp difference between the two tallest histogram bars align with the separation in size and further illustrates the even distribution of read density when using the normalization process associated with the laboratory pipeline. Read length density plots provide a way to quickly visualize whether the sequencing data properly reflect the expected properties of the SMRTbell library. Plots generated for this study indicated ideal performance.

A CCS processing summary was provided for each run that included an assessment of DS and SS reads. The distinction between the two was whether the instrument software determined that the forward and reverse reads were different enough to consider them as a heteroduplex. If so, the subreads were immediately split into SS reads and labeled as forward or reverse, each of which was considered a separate set of subreads resulting in two consensus reads. If the reads were not considered to be a heteroduplex, the DS read pileup became the consensus read for the particular ZMW. The average read output per run for DS reads was 3.76 million and 4.47 million for SS reads, with S.D.'s of 400 thousand and 620 thousand, respectively. A total of 14.8% of DS reads passed the initial quality filter (average of 556 thousand), while 81.0% of the SS reads passed (average of 3.62 million). Most lost DS reads (85.2% of the 3.76 million failed reads) resulted from an inability to complete a sufficient number of passes through the template. Nonetheless, these findings did not have an impact on the ability to generate high-quality, consistent read coverage for each run.

Tertiary analysis of the PacBio generated sequence (n=10,394) was completed using the long-read version of GM HTS. The software mapped a total of 76,784,402 reads, with 76,422,565 (99.5%) of those reads aligning to the rCRS. An average coverage of 3,644 ± 2,361 and 2,856 ± 1,879 was observed for the A and B fragments, respectively. The long-read approach allowed for analysis of long stretches of continuous sequence and produced highly consistent coverage across the individual fragments.

The concordance study performed on a subset (n=188) of the samples sequenced in the PacBio study resulted in two samples that failed to produce sequence data using the LR MiSeq protocol due to failed amplifications. Comparison of major GM HTS haplotypes for the remaining 186 samples showed a concordance of 100% (186/186). Due to differences in the alignment algorithms required for mapping short- and long-read data, manual inspection of the pileups was required to resolve apparent differences in three samples where major variants (present in 50% or more of the reads) in PacBio generated sequence data were observed as minor

components (present in less than 50% of reads) in MiSeq generated sequence data. In one of the three samples, 4393T was observed at a frequency of 50.3% in the PacBio data while the same SNP was observed at 47.3% and 42.7% for the LR MiSeq and the PID MiSeq samples respectively. However, this position would be reported as a heteroplasmic variant (4393Y) in the sample's final haplotype for all three approaches and therefore would not constitute a discordant site. Two samples showed inconsistencies for a 9 bp insertion at np 8289. The insertion was reported as a major component in both samples for the PacBio technology (78.6 & 73.5%) and PID MiSeq samples (68.4 & 69.7%) while the insertion was reported as a minor component in both LR MiSeq samples (44.2 & 37.0%). Evaluation of the alignment pileup illustrated that the 9 bp repeat locus in the LR MiSeq samples is located at the end of reads and subsequent quality based soft trimming applied in the GM HTS processing and alignment is most likely causing the decrease in observations and the loss of the insertion from the major haplotype. Reanalysis of these data in an alternative MPS analysis software showed the 9 bp insertion as the major component in both LR MiSeq samples (95.4% & 80.2%), further supporting that the observed differences in the major haplotypes are an artefact of the GM HTS analysis rather than true discordance. Inconsistencies with the calling of length heteroplasmy (LH) were observed in 35 samples due to variability in the LH being reported as either the major or minor component. All samples with LH variability, fluctuating around the 50% cutoff (50 ± 20%), were ultimately determined to be concordant after manual inspection. Specifically, 309.1C and 309.2C were reported at various frequencies in 34 samples, 524.1A 524.2C in one sample, and 573.1C, 573.2C, and 573.3C in one sample.

Presumed substitution and indel error rates were calculated for each of the 28 PacBio runs. The assumed error rates were well below the established threshold of 10% indicating that a 10% minor allele frequency (MAF) threshold is robust for this dataset. The total indel error rate (ranging 0.16-0.51 errors per 100 nucleotides across the 28 PacBio runs) was higher than the total substitution error rate (ranging 0.09-0.15 errors per 100 nucleotides), with insertions having the highest rate (0.08-0.43) of misincorporation. Overall, the assumed error rate estimates indicate the PacBio system has low observed error, with indels being the predominant form of noise in the data.

To assess whether "hotspots" of error exist at specific locations, the 25 nps with the highest rate of error for each of the 28 PacBio runs were evaluated for each type of error (top 25*6 error

types*28 runs =4,200). The frequency range observed across all runs, the proportion of nps with error exceeding a frequency of 10%, and the region of error sites across the mitogenome are provided in the table below.

	Minimum Freq.	Maximum Freq.	Positions > 10% (%)	HVI Positions (%)	Enrichment Ratio*	HVII Positions (%)	Enrichment Ratio*	Positions outside of HVI/HVII (%)	Enrichment Ratio*
Substitution Error									
A	0.24	5.41	0.00	7.0	3.4	3.7	2.3	89.3	0.93
C	0.31	13.48	0.43	24.7	12.0	10.9	6.7	64.4	0.67
G	0.26	8.14	0.00	19.6	9.5	3.7	2.3	76.7	0.80
T	0.19	13.04	0.43	14.6	7.1	18.9	11.7	66.6	0.69
Indel Error									
Insertion	1.54	28.22	41.00	0.0	1.8	0.1	4.9	99.9	1.04
Deletion	2.98	24.70	33.71	3.7	0.0	8.0	0.06	88.3	0.92

*note: Enrichment Rtio= $\frac{P_{obs(A)\%}}{P_{size(A)\%}}$

Calculation of an enrichment ratio, which normalizes the percentage of observations across regions of different sizes by dividing the percentage of observations in each region by the percentage of total genome for that region, indicates that the majority of top error positions were observed in the mtDNA hypervariable I (HVI) region (nps 16024–16365) for A, C, and G substitutions, in the HVII region (nps 73-340) for T substitution and deletions, and outside the mtDNA control region (nps 577-16023) for insertion related noise in the data. Overall, the rate of indel error was substantially higher than substitution error rates. It is important to note that while certain positions had assumed error rates that surpassed the 10% threshold, these positions were overwhelmingly associated with indel errors not substitution errors, and when observed, the number of these positions were limited (generally 5 or less). Most importantly, these positions could be easily resolved through manual evaluation of the pileup.

For the first 5,004 complete mitogenomes, 3,930 unique haplotypes were identified, demonstrating substantial diversity within the dataset. The haplogroup composition was dominated by European/West Eurasian maternal lineages, particularly haplogroups H, U, T, J, and K, with additional lineages associated with African, Asian, and Latino/Native American ancestry also represented. Point heteroplasmy was detected in 20.4% of individuals using a 10% minor allele frequency (MAF) threshold. Among heteroplasmic individuals, 873 (17.4%) exhibited a single heteroplasmic site, 124 (2.5%) exhibited two sites, and fewer than 1% contained three (n = 20), four (n = 4), or five (n = 1) heteroplasmic sites. The prevalence of heteroplasmy was broadly similar across the four population groups, ranging from 15.5% in

Latino/Native American individuals to 27.3% in Asian individuals, with no statistically significant differences observed between populations. A total of 1,201 heteroplasmic observations were detected across the dataset, including 381 within the control region and 820 within the coding region. When normalized for genomic length, the control region exhibited a 6.4-fold higher rate of heteroplasmy per nucleotide than the coding region. Several recurrent heteroplasmic sites were identified, including positions 16093, 152, and 16192, which are well-recognized mutational hotspots within the mitochondrial control region. Overall, these initial findings provide additional insight into the prevalence, distribution, and recurrence of mitochondrial heteroplasmy. The broader dataset will be part of a manuscript submitted for publication.

Products:

A manuscript has been submitted for publication that reports on the 10,394 samples analyzed. A manuscript is in preparation that will report on the observed heteroplasmy in the dataset. A third manuscript is planned that will address a detailed assessment of the analytical noise in the PacBio sequencing data. Three Masters-level students (Jasmeen Khosa, Claudia Prieto Alcaide and Erin Brownfield) included work on the project as part of their final research paper. A doctoral student (Floor Claessens) included the portion of work she did on the project as part of her thesis work at the Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

The work completed on this project included the development of a laboratory-based pipeline and a bioinformatics-based process for analyzing mitogenome sequences in a high-throughput, cost effective manner. This work product, along with the assessment of its application to a sample set of more than 10,000 whole blood, buffy coat, and DNA extracts will benefit the forensic community for those who choose to pursue the development of additional datasets of mitogenome sequence. Archived data in the form of approximately 535 GBs of unaligned BAM files and resulting 650 GBs of GM HTS data files can be requested by the National Institute of Justice (NIJ). Bioinformatic tools developed during the execution of the project can be provided to the NIJ upon request; or through published material.

The following is a list of conferences and workshops where the findings of the project have been disseminated. In addition, materials (PowerPoints) are posted routinely on a research website (https://sites.psu.edu/hollandresearch/) that serves as a resource for interested parties.

1. We had multiple zoom meetings throughout the project with our consultants from the Medical University of Innsbruck and the Armed Forces DNA Identification Laboratory (AFDIL); Walther Parson and Charla Marshall, respectively. We hosted our consultants on 14-15 November 2023 for a workshop on the progress of the project. These meetings were helpful and allowed for the dissemination of our initial methods development to two interested parties who have extensive mitogenome sequencing and casework experience. Both consultants are authors on the first submitted manuscript.
2. Dr. Holland was an invited speaker at the Technical Leaders meeting at the International Symposium on Human Identification (ISHI) in Denver in September of 2023. We (Erin Brownfield, first author) also presented a poster at ISHI 2023 on the methods development aspects of the project.
3. Dr. Holland gave an oral presentation as an invited speaker at the International Society for Applied Biosciences (ISABS) conference in Split, Croatia in June of 2024. We (Claudia Prieto Alcaide, first author) also presented a poster on the methods development aspects of the project.
4. Dr. Holland gave an invited oral presentation for the laboratory staff at AFDIL in Dover, DE, in August 2025.
5. Dr. Holland will be an invited speaker at ISABS conference in Dubrovnik, Croatia in June of 2026.

References:

Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, et al. Sequence and organization of the human mitochondrial genome. Nature. 1981;290:457-65.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, et al. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature Genetics. 1999;23:147.

Canale LC, McElhoe JA, Dimick G, DeHeer KM, Beckert J, Holland MM. Routine mitogenome MPS analysis from 1 and 5 mm of rootless human hair. Genes. 2022:13;2144-2163.

Eid J, Fehr A, Gray J, Luong K, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323:133-38.

Fendt L, Zimmermann B, Daniaux M, Parson W. Sequencing strategy for the whole mitochondrial genome resulting in high quality sequences. BMC Genomics. 2009;10:1–11.

Holland MM, Makova KD, McElhoe JA. Deep-coverage MPS analysis of heteroplasmic variants within the mtgenome allows for frequent differentiation of maternal relatives. Genes. 2018;9:1-21.

Holland MM, Pack E, McElhoe JA. Evaluation of GeneMarker® HTS for improved alignment of mtDNA MPS data, haplotype determination, and heteroplasmy assessment. Forensic Science International: Genetics. 2017;28:90-8.

Huber N, Parson W, Dür. Next generation database search algorithm for forensic mitogenome analysis. Forensic Science International: Genetics. 2018;37:204-14.

Korber JT, Canale LC, Holland MM. Massively parallel sequencing of the mitogenome from human hair shafts in forensic investigations. Curr Protocols Hum Genet. 2023:3, e865. doi: 10.1002/cpz1.865

Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nature Reviews. 2020;21:597-614.

Lundquist PM, Zhong CF, Zhao P, Tomaney AB, et al. Parallel confocal detection of single molecules in real time. 2008;33:1026-28.

McElhoe JA, Holland MM. Characterization of background noise in MiSeq MPS data when sequencing human mitochondrial DNA from various sample sources and library preparation methods. Mitochondrion. 2020;52:40-55.

McElhoe JA, Wilton PR, Parson W, Holland MM. Exploring statistical weight estimates for mitochondrial DNA matches involving heteroplasmy. International Journal of Legal Medicine. 2022;136:671-85.

Peck, MA, Sturk-Andreaggi, K, Thomas, JT, Oliver, RS, Barritt-Ross, S, Marshall, C, Developmental validation of a Nextera XT mitogenome Illumina MiSeq sequencing method for high-quality samples, Forensic Sci. Int. Genet. 34 (2018) 25–36. https://doi.org/10.1016/j.fsigen.2018.01.004.

QIAGEN, A Simple One-Step Library Prep Method To Enable AmpliSeq Panel Sequencing on Illumina Platforms, Https://Www.Qiagen.Com/Us/Resources/Resourcedetail?Id=8d58fa59-7652-4ef8-9c59-8f28976c64e9&lang=en (2016) 1–4.

Rhoads A, Au KF. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics. 2015;13:278-89.

Taylor, CR, Kiesler, KM, Sturk-Andreaggi, K, Ring, JD, Parson, W, Schanfield, M, Vallone, PM, Marshall, C, Platinum-Quality Mitogenome Haplotypes from United States Populations, Genes (Basel). 11 (2020) 1290. https://doi.org/10.3390/genes11111290.

Wenger AM, Peluso P, Rowell WJ, Chang P-C, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology. 2019;37:1155-62.