Document Title:	Measuring Rates of mtDNA Heteroplasmy in Different Population Groups
Author(s):	Mitchell M. Holland
Document Number:	252863
Date Received:	April 2019
Award Number:	2016-DN-BX-0171

Agency: National Institute of Justice

Award Number: 2016-DN-BX-0171

Project Title: Measuring Rates of mtDNA Heteroplasmy in Different Population Groups

PI: Mitchell M. Holland
Associate Professor BMB-Forensics
mmh20@psu.edu
814-865-5286

Submitting Official: Mitchell M. Holland
Associate Professor BMB-Forensics

Submission Date: 09/28/18

[redacted]

Recipient Organization: The Pennsylvania State University – Univ Park
Office of Sponsored Programs, 110 Technology Center Building
University Park, PA 16802-7000

Award Period: 01/01/2016 to 09/30/2018 (no cost extension)

Reporting Period End Date: 09/30/2018

Signature of Submitting Official:

Mitchell M. Holland

PURPOSE

Several objectives will need to be satisfied before massively parallel sequencing (MPS) methods will be ready for implementation in forensic laboratories, including the development and validation of protocols for forensic applications. Reliable methods of MPS analysis for the mitochondrial (mt) DNA genome for pristine samples have already been developed, including a method we developed on the MiSeq instrument from Illumina (McElhoe et al., 2014). In addition, efforts have been made to address the analysis of sample types such as hair shafts (Bintz et al., 2013; Stawski et al., 2013) and low quantity DNA samples (Just et al., 2014).

Given our understanding of the diverse nature of mtDNA haplotypes across different population groups, and that mutations resulting in low-level heteroplasmy may be associated with local sequence and their effects on the replication process, the goals of this research were to measure the rate of heteroplasmy across the mtDNA control region (CR) on an individual and per nucleotide basis for people of African, East Asian and Latino ancestry, and compare the rates across all population groups; including a study we recently completed on Europeans. Statistical evaluation of the European dataset has indicated possible correlations of heteroplasmy to haplotype. Therefore, this will be assessed for all population groups. Findings will be used to help guide the refinement of best practices regarding the reporting of mtDNA heteroplasmy; including statistical analysis. The collective findings will help forensic laboratories as they prepare to report heteroplasmy in casework, significantly enhancing the discrimination potential of mtDNA testing. This outcome may also serve as motivation for a broader range of laboratories to adopt mtDNA analysis using an MPS approach.

EXPERIMENTAL DESIGN & METHODS

A total of 282 buccal swab samples have been collected by our laboratory, and 509 saliva samples were provided by Mark Shriver's laboratory at Penn State. To date, genomic DNA has been isolated from all of the cheek swabs (n=282) collected by our laboratory using the Gentra Buccal Cell Kit (QIAGEN, Germantown, MD). Each sample was obtained using an individually wrapped buccal collection brush and promptly stored in the kit supplied cell lysis buffer, in which samples are stable at room temperature for up to two years. The amount of time the samples were stored in the lysis buffer varied, but no sample was stored for more than a period of one month. Samples were extracted following the manufacturer protocol. Genomic DNA was isolated from all (n=509) saliva samples collected by the Shriver laboratory using the Gentra Blood kit (QIAGEN), following the manufacturer protocol.

Enrichment of the mtDNA CR was accomplished through amplification of a 1 kilobase (kb) target spanning nps 15997-16569 and 1-926 with transposase adapted primers. Library preparation was conducted using the Nextera® XT approach and sequencing was performed on a MiSeq benchtop sequencer using a 300 cycle kit (v.2 chemistry) with 150 x 150 paired-end reads. Sequence data was mapped to the revised Cambridge Reference Sequence (rCRS; GenBank ID NC_012920.1; Andrews et al., 1999; Anderson et al., 1981) using the MiSeq Reporter integrated computer software platform (MSR; v2.2.29), which operates on a Burrows-Wheeler Aligner (BWA) and the Genome Analysis ToolKit (GATK) for variant calling of single nucleotide polymorphisms (SNPs) and short insertions and deletions (indels). Secondary analysis of the MSR generated FASTQ data was performed using GeneMarker® HTS software (v1.2.2; SoftGenetics, Inc., State College, PA)(Holland et al., 2017a). Additional secondary analysis was achieved using pipelines developed by our laboratory using a combination of UNIX line commands (Bash, 2007) and the R environment (Team R, 2015).

DATA ANALYSIS

A critical ingredient necessary for the adoption of an MPS approach in forensic laboratories is the availability of a suitable software package for data analysis. The lack of available software solutions led to a collaboration with SoftGenetics, Inc, to develop a software package, GeneMarker® HTS, for forensic researchers and practitioners (Holland & McElhoe 2015; Holland et al., 2017a). This software package allows for the proper alignment of sequence data in regions of sequence that are typically difficult to align, including homopolymeric stretches and complex patterns of SNPs and indels, both of which typically produced inconsistent reporting outcomes with traditional alignment software.

Careful examination of mtDNA MPS data is important, as illustrated by the publication of several high profile reports that have been deemed in error due to an inability to distinguish between heteroplasmy and other sources of mixed data, including those associated with software alignment anomalies (Just et al, 2014). The MPS data for this study were analyzed using GeneMarker® HTS (v1.2.2) at a 1% analytical threshold and a 2% reporting threshold for minor sequence variants. While the data can be evaluated at the 1% threshold, we are recommending the use of a reporting (interpretation) threshold of 2%. Application of this approach was successful when analyzing mixture MPS data generated with the D-Loop Protocol from Illumina (Holland et al., 2017b).

The error rate of the entire MPS process was evaluated to determine if the "noise" was less than our analytical threshold of 1%. The MiSeq benchtop sequencer is known to have a low error rate (Ross, et al., 2013; Schirmer et al., 2015; Li and Stoneking, 2015), but with sequencing systems constantly improving, and with our specific procedure, we needed to establish the error associated with both our sequencing and alignment procedures. To estimate the error rate, we chose to take a highly conservative approach and use all sequencing data calls observed at ≤50% in calculation of the assumed error. Major allele calls, by definition, have frequencies >50% and it has been shown that MPS data for major alleles is concordant with traditional Sanger-type sequencing (Loman, et al. 2012). Coverage and base call information was generated using GeneMarker® HTS and subsequently processed using a combination of Unix line commands and R Studio for assumed error assessment.

FINDINGS

Haplotype

Currently, 757 buccal and saliva samples have been sequenced. Of the sequenced samples, 99 samples were determined to be of European ancestry, and seven samples were contaminated, for a total of 650 samples of African, Asian, and Latino ancestry. Our evaluation of haplotype and assigning haplogroups to population groups indicates that 290 (290/250 or 116%) samples belong to African haplogroups, 247 (247/250 or 98.8%) to Asian haplogroups, and 113 (113/250 or 45.2%) to Latino haplogroups. Of the 650 haplotypes 83.8% (n=545) were unique in the dataset, with a total of 587 different haplotypes (90.3%). Of the shared haplotypes, 28 were shared by two individuals, nine were shared by 3 individuals, four shared by 4 people, and the most common haplotype was shared by six individuals (6/587 or 1.0%). Based on self-reported ethnicity, 86.7% (650/750) of the 750 samples sequenced to date, produced haplotypes that corresponded to the expected haplogroups.

The inconsistency between reported ancestry and actual haplotype/haplogroup is the reliance on self-reporting of ethnicity for sample collection. This is a challenge that has been documented by others and is a result of the complex nature of ancestry labels that are based on elements such as visible traits, including skin color, and from cultural and geographical factors (Klimentidis et al., 2009). In contrast, the assignment of mtDNA haplogroups is based on single nucleotide polymorphisms (the haplotype) that segregate individuals into lineages. To date, our research on which haplogroups belong to specific population groups indicate that African haplogroups are relatively straightforward, with the majority falling into the major haplogroup L, but also containing sub-haplogroups U6 and M1 (Secher et al., 2014), while the determination of haplogroups for Asian and Latino populations is more challenging. It is known that Latino populations generally have a mixed heritage of African, European, and Native American (Bedoya et al 2006 PNAS; Bryc et al., 2015), and several major haplogroups contain sub-haplogroups that are reflective of Asian descent along with other sub-haplogroups that indicate Latino ancestry. For example, A2a, A2b, B4, B5, C, D4, D5, M7, M8, M9, and Na9 are Asian haplogroups (Alves-Silva et al., 2000; Yao et al., 2002; Allard et al., 2004), while A2, B2, C1, C4, D1, and D2 are Latino/Native American haplogroups (Forster et al., 1996; Perego et al., 2010). Such fine-grained delineation of population groups based on sub-haplogrouping complicates our ability to use major haplogroups for sample grouping based on African, Asian, and Latino population groups, but allows us to group samples with similar control region sequences. The ability to group samples with similar sequence is critical for our statistical evaluation of possible heteroplasmy/haplotype linkage.

A major component of this study is the evaluation of whether heteroplasmy at a specific chromosome location, or a specific region of the mtDNA genome is dependent on the specific haplotype of an individual. In our previous study on Europeans, a possible correlation was assessed using statistical significance testing to analyze contingency tables. The contingency tables were generated by counting different combinations of heteroplasmy observations in conjunction with haplogroup information. Next, Chi-squared and Fisher's exact testing were applied to the tables. Both tests apply a null hypothesis of independence, under which there is a hypergeometric distribution of the numbers in the counts of the table. Fisher's exact test is generally applied to 2x2 contingency tables while the Chi-squared method is best for larger tables. Analysis of independence for heteroplasmy is challenging due to the rarity of observations. Sequencing 1,122 nucleotides within the CR for 550 Europeans (617,100 positions) produced only 283 observations (0.05%) of heteroplasmy. This highlights the need for additional sequencing data to bolster the statistical analysis, and the dataset from this project more than doubled the number of observations with a current total of 345 additional observations of heteroplasmy. Even with the increase in observations, the analysis will most likely still be hampered by small sample size and/or unequal distribution among the cells of the contingency table, in which case the Fisher's exact test is more dependable, although the approximation method may still be unreliable due to a lack of data. In an attempt to minimize the effects of small sample size, multiple different groupings of heteroplasmy and major haplogroup designations were evaluated, with the final groupings only using combinations with observations greater than ten. Using the major haplogroups as a grouping method to increase the number of samples per group still only produced nine groupings for the 550 European samples.

While the statistical method appears to be valid for testing independence, given the rarity of heteroplasmy, our previous results should probably be viewed as preliminary, with the need for additional data points to increase the statistical strength of the analysis. Based on strengthening the statistical analysis and the inherent challenges with targeted sample collection using self-reported ancestry, the goal of this study will shift slightly to having an emphasis on increasing the sample size of major haplogroups rather than focusing on the characterization of African, Asian, and Latino populations based on the fine-grain detail of associating sub-haplogroups (i.e. U6 vs. the major haplogroup U) to specific ethnic populations. That being said, and to the extent possible, we will strive to maintain the originally proposed goal of an equal distribution of samples from individuals of African, Asian, and Latino descent for the final dataset. Statistical analysis to evaluate possible linkage between haplotype and heteroplasmy will be conducted, and included in the manuscript based on the final dataset, once the complete dataset is generated and analyzed.

Error

Prior to analysis of heteroplasmy in the samples sequenced to date, the average total assumed error for all nucleotides (A, C, G, and T) was assessed by assuming all base calls <50% were called in error. Briefly, the consensus statistic report that was generated by GeneMarker® HTS was manipulated in order to combine the forward and reverse reads, calculate the frequency based on total coverage, and then transform the frequencies back to counts to produce the assumed error. A combination of Terminal and R Studio was used to calculate the total assumed error and individual (ACGT) nucleotide error. The error values represent a percent of the total reads, which can also be described as the number of calls made in error for every 100 nucleotide called. All results, even the total assumed error, are well below our reporting threshold of 2%, indicating that heteroplasmic positions reported at 2% are well above the noise of the system.

The data presented here is a subset of the 650 samples, representing 628 total samples. The combined error rate was 0.169 erroneous base calls for every 100 nps. The average assumed error for each nucleotide (A, C, G, and T) was 0.035, 0.054, 0.045, and 0.035 per 100 nps, respectively. Therefore, the analytical threshold of 1% and a reporting threshold of 2% have proven to be robust when reporting minor sequence variants.

Heteroplasmy

Point heteroplasmy, assessed at a reporting threshold of 2%, was observed at 101 nucleotide positions (9.0%) across the 1,122 total positions of the CR. Heteroplasmy was observed in 36.9% of all samples, for a total of 345 heteroplasmic observations in 240 individuals. Three quarters of the sites (77.1%) were observed at a minor allele frequency (MAF) between 2-10%. The majority of individuals had no heteroplasmy (63.1%), while 26.7% of individuals had one site of heteroplasmy (174/650), 6.1% of individuals had two sites (40/650), 2.1% of individuals had 3 (14/650), 1.7% had 4 (11/650) sites, and 1 individual had five sites of heteroplasmy (0.15%). Position 16093 was the most prevalent heteroplasmic position with 43 observations (12.5%) followed by positions 16183 (39 obs.), 16189 (25 obs.), 16192 (18 obs.), 16182 (18 obs.), 215 (14 obs.), and 204 (13 obs.). The remaining positions with observed heteroplasmy had nine or less observations per positions.

When compared to our previous study on a European population (n=550), the rate of heteroplasmy in the current study (current:36.7%; European:41%), the number of sites across the control region having observations of heteroplasmy (current:101; European:80), the most prevalent site of heteroplasmy (nucleotide position 16093; current:12.5%; European:12.4%), and the rate of heteroplasmy with a MAF between 2-10% (current:77.1%; European:73.5%) are all in good agreement. In addition, the rates of heteroplasmy per individual are quite similar for one site of heteroplasmy (current:26.7%; European:32%), two sites (current:6.1%; European:6%), and three sites (current:2.1%; European:2.4%), but differ for four sites (current:1.7%; European:0.18%), and five sites (current:0.15%; European:0%). The two data sets showed the greatest divergence when comparing the location of the observed heteroplasmic positions within the CR. Of the 345 (European:283) observed instances of heteroplasmy, 71.9% (European: 39.2%), 22.9% (European: 53.7%), and 5.2% (European: 7.1%) fell within HVI, HVII, and outside the hypervariable regions, respectively. The observed differences could simply be a reflection of differences in rates across population groups or could possibly have dependence on haplotype and SNP locations.

A total of 101 nps (9.0%) exhibited heteroplasmy across the 1,122 sites in the CR. The most prevalent type of heteroplasmy was C/T-based with 215 observations (62.3%), followed by A/G (68, 19.7%), and A/C (62, 17.9%) . The A/T, C/G, and G/T transversions produced no observations. Overall, the vast majority of nps exhibited no heteroplasmy (91.0%). Using a crude approach for determining the frequency of heteroplasmy at these nps (3/550 or ~0.55%), a likelihood ratio (LR) in a forensic case could be increased by a factor of ~180. Assuming a LR of 1000 for the haplotype, the presence of heteroplasmy at one of these positions would result in an increase in the LR to ~180,000. On the other hand, the np with the highest rate of heteroplasmy, 16093 (12.8%), would result in an increase in the LR to ~32,000. While this is not as impactful, it would still be of benefit to the trier of fact. We are in the process of assessing whether a correlation exists between haplotype and occurrence/position of heteroplasmy. Assuming a lack of correlation, reporting heteroplasmy in a case will clearly increase the discrimination potential of the testing method. The data regarding position 16093 provides a crude assessment that haplotype and heteroplasmy are not linked.

CRIMINAL JUSTICE IMPACT

We have presented our work on this project at the Northeastern Association of Forensic Scientists (NEAFS) meeting, November 6-10th, 2017, in Pocono Manor, PA. We have also shared this work during a platform presentation at the 70th annual American Academy of Forensic Sciences (AAFS) in Seattle, WA on February 19-24th, 2018. AAFS is an excellent opportunity to disseminate our work and has the potential to impact many in the forensic community, as it is one of the larger forensic-focused conferences, with a recent attendance of 4,044 (including 507 international scientists). But, the most important outcome of our project will be a scholarly publication, which we will intend to submit during the early 2019 calendar year.

BIBLIOGRAPHY

Allard, MW, Wilson, MR, Monson, KL, Budowle, B. Control region sequences for far East individuals in the Scientific Working Group on DNA Analysis Methods Forensic mtDNA Data Set. Legal Med. 2004; 6:11-24.

Alves-Silva, J, da Silva Santos, M, Guimaraes, PEM, Ferreira, ACS, Bandelt, HJ, Pena, DSJ, Prado VF. The ancestry of Brazilian mtDNA lineages. Am. J. Hum. Genet. 2000; 67:444-461.

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

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the cambridge reference sequence for human mitochondrial DNA. Nature Genetics. 1999;23:147.

Bash [Unix shell program]. 3.2.57(1) ed: Free Software Foundation; 2007.

Bedoya, G, Montoya, P, Garcia, J, Soto, I, Bourgeois, S, Carvajal, L, Labuda, D, Alvarez, V, Ospina, J, Hedrick, PW, Ruiz-Linares, A. Admixture dynamics in Hispanics: A shift in the nuclear genetic ancestry of a South American population isolate. PNAS. 2006; 103(19): 7234–7239.

Bintz B, Burnside ES, Wilson MR. A tale of two platforms: An evaluation of the Roche GS Junior and Illumina MiSeq next-generation sequencing instruments for forensic mitochondrial DNA analysis. 24th International Symposium on Human Identification. Atlanta, GA: Promega; 2013.

Bryc, K, Durand, EY, Macpherson, JM, Reich, D, Mountain, JL. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. Am. J. Human Genetics. 2015; 96:37–53.

Forster, P, Harding, R, Torroni, A, Bandelt, HJ. Origin and evolution of native American mtDNA variation: a reappraisal. Am. J. Hum. Genet. 1996; 59:935-945.

Holland M, McElhoe J. A custom software solution for forensic mtDNA analysis of MiSeq data, Forensic Sci Int: Genetics (Suppl Series). 2015;5:e614-e16.

Holland MM, Pack ED, McElhoe JA. Evaluation of GeneMarker HTS for improved alignment of mtDNA MPS data, haplotype determination, and heteroplasmy assessment. Forensic science international Genetics. 2017a;28:90-8.

Holland M, Wilson L, Copeland S, Dimick G, Holland C, Bever R, et al. MPS analysis of the mtDNA hypervariable regions on the MiSeq with improved enrichment. International Journal of Legal Medicine. 2017b:1-13.

Just RS, Irwin JA, Parson W. Questioning the prevalence and reliability of human mitochondril DNA heteroplasmy from massively parallel sequencing data. Proc Natl Acad Sci. 2014;111:e4546-e47.

Klimentidis, YC, Miller, GF, Shriver, MD. Genetic admixture, self-reported ethnicity, self-estimated admixture, and skin pigmentation among Hispanics and Native Americans. Am. J. Phys. Anthropol. 2009; 138:375-383

Li M, Stoneking M. A new approach for detecting low-level mutations in next-generation sequence data. Genome Biology. 2012;13:R34.

Loman N, Misra R, Dallman T, Constantinidou C, Gharbia S, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. Nature Biotechnology. 2012;30:434-41.

McElhoe J, Holland M, Makova K, Su MS-W, Paul I, Baker C, Faith S, Young B. Development and Assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq. Forensic Sci Int: Genetics. 2014;13:20-29.

Perego, UA, Angerhofer, M, Pala, A, Olivieri, H, Lancioni, BH, Kashani, V, Carossa, JE, Ekins, A, Gomez-Carballa, G, Huber, B, Zimmermann, D, Corach, N et al. The initial peopling of the Americas: a growing number of founding mitochondrial genomes from Beringia. Genome Res. 2010; 20:1174-1179.

Ross M, Carsten R, Costello M, Hollinger A, Lennon N, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biology. 2013;14:R51

Schirmer M, Ijaz U, D'Amore R, Hall N, Sloan W, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Research. 2015;43:e37.

Secher B, Fregel, R, Larruga, JM, Cabrera, VM, Endicott, P, Pestano, JJ, González, AM. The history of the North African mitochondrial DNA haplogroup U6 gene flow into the African, Eurasian and American continents. BMC Evolutionary Biology. 2014; **14**:109

Stawski H, Burnside ES, Wilson MR. Evaluation of whole genome amplification kits for augmentation of mitochondrial DNA from hair shaft extracts for next generation sequencing. 24th International Symposium on Human Identification. Atlanta, GA: Promega; 2013.

Team R. RStudio: Integrated Development Environment for R. . 0.99.903 ed. Boston, MA RStudio, Inc.; 2015.

Yao, YG, Kong, QP, Bandelt, HJ, Kivisild, T, Zhang, YP. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am. J. Hum. Genet. 2002; 70:635-651.