Agency: National Institute of Justice

Award number: 2014-DN-BX-K022

Project Title: Measuring Rates of mtDNA Heteroplasmy and Assessing Transmission of Variants

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: 12/31/2016

DUNS: [redacted]

EIN: [redacted]

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

Award Period: 01/01/2015 to 03/31/2017 (no cost extension)

Reporting Period End Date: 12/31/2016

Signature of Submitting Official:

29 Dec 2016

Mitchell M. Holland

PURPOSE

The ability to resolve, report, and leverage the discrimination potential of heteroplasmy will significantly enhance the value of using mitochondrial (mt) DNA analysis in forensic casework [1]. A massively parallel sequencing (MPS) approach will allow the community to achieve this goal. The first part of our study focused on establishing rates of heteroplasmy for the control region (CR) of the mtGenome through MPS analysis of 550 individuals of European ancestry. The study was comprised of male and female participants in three age groups (18-29, 30-49, and ≥50 yoa) to evaluate potential gender and age effects on rate estimates. Rates were assessed on a population and nucleotide position (np) basis. The second part of the study evaluated the transmission of heteroplasmic sequence variants in three tissue types (blood and buccal cells, hair shafts) collected from multiple maternal lineages, and in one case, from an individual lineage across multiple generations. Pipelines and software tools were developed to conduct a thorough and complete analysis of the data, which will be available to the forensic community before or at the end of the grant period; the code for error assessment of MPS data is provided in the Appendix as an example. Collective data were used to evaluate the impact of error rates on reporting low-level heteroplasmy, and to calculate heteroplasmy frequency estimates. Scholarly articles have been published and submitted, or are in preparation, including recommendations which can be used to develop best practices when conducting mtDNA MPS analysis in forensic laboratories.

EXPERIMENTAL DESIGN & METHODS

Genomic DNA was collected and isolated from cheek swabs using the Gentra Buccal Cell Kit (QIAGEN). Each sample was obtained using an individually wrapped buccal collection brush and promptly stored in the supplied cell lysis buffer. Samples were stable at room temperature for up to two years, although the actual time samples were stored in the lysis buffer varied, with no sample stored for more than a period of one month prior to DNA extraction. Samples were extracted throughout collection, following the manufacturer protocol, as soon as batches of 24 samples became available. Appendix Table 1 lists metadata for the samples. Of the 550 target samples, 494 samples were collected by our laboratory, and 56 samples were provided by Professor Mark Shriver's laboratory at Penn State. Genomic DNA was isolated from liquid saliva samples collected by the Shriver laboratory using an organic method.

For the rate study, enrichment of the mtDNA CR was accomplished through amplification of a 1 kilobase (kb) target spanning nps 15997-16569 and 1-926 with transposase adapter primers [2] or a 2 kb target spanning nps 15600-16569 and 1-960 [3]. 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) using the MiSeq Reporter integrated computer software platform (MSR; v2.1.43 and 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 (sequence and quality scores) was performed using NextGENe® (v.2.3.3) and GeneMarker® HTS software (v.06162016). Additional secondary analysis was achieved using pipelines developed by our laboratory using a combination of UNIX line commands and the R environment.

Transmission was evaluated through whole mtGenome sequencing of a maternal lineage consisting of three generations, with multiple family members across each generation; three grandparents, seven children, and eight grandchildren, including identical twins. Two tissue types (blood or buccal cells) were collected from the eighteen family members (36 samples). Enrichment was accomplished using long-range PCR (~8.5 kb fragments), library preparation via Nextera® XT [4,5], and sequencing on a MiSeq using a 600 cycle reagent kit with 300 x 300 paired-end reads. Data for each sample was generated in duplicate for a total of 72 sequencing events (plus controls). In addition, a broad assessment of variant transmission in hair shafts was conducted for 15 different maternal lineages. The MPS profile of the mtDNA CR from blood and buccal cells was generated for each participant. Hairs were collected from five individuals with no heteroplasmy in their blood and buccal cells, five with low-level heteroplasmy (<10%), and five with high-level heteroplasmy (>10%); hairs were collected from five different regions of the scalp. The DNA from 2 cm hair shaft cuttings (five hairs from for each participant), was extracted using a method developed at Western Carolina University (WCU). The extraction method was assessed as part of a project to compare the yield of extracted DNA when using the WCU method in comparison to two methods used by operating forensic laboratories, with the WCU method outperforming the other two methods [6]. Enrichment was accomplished using the PowerSeq™ Mito System kit from Promega, a prototype, nested 10-plex approach for analysis of the CR. The amplicons were prepared for sequencing using the 10-plex library kit, and sequencing was performed on the MiSeq using the 600 cycle kit (v.3 chemistry) with 300 x 300 paired-end reads, for a total of 110 sequencing events (plus controls).

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 became apparent as the labor involved in data analysis was extraordinary during the early stages of this project. Therefore, we collaborated with SoftGenetics, Inc, to develop a software package for forensic researchers and practitioners; GeneMarker® HTS [7,8]. Existing software did not allow for the proper alignment of sequence data, producing flawed reports and requiring extensive manual analysis to identify and correct the errors. The regions of sequence that are typically difficult to align are homopolymeric stretches and patterns of SNPs and indels, both of which can produce inconsistent reporting outcomes.

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 [9]. In this study, difficulties in alignment, including the separation of major and minor allele calls, made the evaluation of heteroplasmic positions a multi-layered process requiring numerous repeat analyses. The MPS data for this study were analyzed a final time using GeneMarker® HTS (v.06162016) at a 1% analytical threshold and a 2% (n=537) or 3% reporting threshold (n=13) 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 MPS data associated with mixtures, and generated with the D-Loop Protocol from Illumina [10].

The error rate of the collective MPS process was evaluated to establish an analytical threshold based on the measured “noise”. This is a critical step to ensure that reported heteroplasmy is reliable. Major allele calls, by definition, have SNP percentages >50%; it has been clearly illustrated that MPS analysis for haplotype determination is concordant with traditional Sanger-type sequencing [11]. We chose a conservative approach to assess error rates, considering MPS data with nucleotide calls observed in <50% of the sequencing reads as assumed error. This inherently captures all positions of heteroplasmy, ensuring that the error rate is not biased by removal of this data. Coverage and base call information, for the 230 samples used for this analysis, were generated using GeneMarker® HTS and processed using a combination of UNIX line commands and R Studio for assumed error assessment. The MiSeq is known to have a low error rate, but the empirical error associated with the combined sequencing and alignment procedures has not been established.

FINDINGS

A total of 717 buccal samples were collected for this study; 130% of the proposed number. MPS data was generated for 569 samples, including 550 individuals of European descent. As expected, a percentage of the individuals who self-reported as European were of other ancestral origins (Appendix Table 1). For the purposes of this report, we focused on the 550 Europeans.

The development and evaluation of GeneMarker® HTS required numerous meetings with the team at SoftGenetics, and evaluation of multiple iterations of the developing software over a 12-month period. The fully developed software was assessed for; 1) proper alignment to a circular version of the mtGenome to span the transition point in the mtGenome numbering system, 2) consistent reporting associated with the mtGenome numbering system, 3) features for user-defined filtering and production of meaningful and accurate reports, 4) export of reports to address forensic considerations and allow for import into tertiary analysis tools such as EMPOP, 5) proper alignment of homopolymeric sequences, challenging SNP and indel motifs, and identify phylogenetically correct primary haplotypes with minimal user input, and 6) identification of heteroplasmic variants with minimal user input. We included our consultants (Walther Parson, University of Innsbruck and Ann Gross, MN BCA) in the development and evaluation process at an early stage to address their interests and solicit their feedback. The GeneMarker® HTS software has recently become commercially available, and the team at

SoftGenetics has reported to us that the FBI Laboratory has initiated the purchasing process to acquire a copy. Thus far, and to the best of our knowledge, the following laboratories in the forensic community have evaluated the software, or have expressed interest in the software: CA DOJ, MN BCA, OCME, AFDIL, Bode Technology Group, the Netherlands Forensic Institute, and the Institute of Legal Medicine Innsbruck Medical University.

MPS was conducted on the 550 samples and data was analyzed with GeneMarker® HTS. Consistent with previous findings [12], ~75% (412) of the 550 haplotypes were unique in the dataset, with a total of 460 different haplotypes (~84%). The slightly elevated values, when compared to previous data, reflect expanded analysis of the CR in the current study. Forty-eight haplotypes were shared by more than one individual (48/460=~10.5%); 27 by two individuals, 12 by three individuals, 4 by four individuals, 2 by five individuals, 1 by six individuals, 1 by seven individuals, and 1 by nine individuals. Using the GeneMarker® HTS software resulted in a decrease in frequency for the most common sequence profile in the data set (263G, 315.1C, 16519C); from ~1.3% to ~4% in our previous reports. The lower frequency is due to the improved alignment capabilities of the new software and better resolution of length variants. Traditionally, length variants have been largely ignored in MPS data due to alignment challenges leading to erroneous typing from miscalled indels, heteroplasmy, and complete substitutions [13]. Improved alignment with the GeneMarker® HTS significantly increased the number of times length variants were resolved in our samples set, especially related to homopolymeric sequences; nps 16182-16193 and 303-315. For example, haplotype 263G, 309.1C, 315.1C, 16519C was observed nine times and haplotype 263G, 309.1C, 309.2C, 315.1C, 16519C was observed six times. If the additional length heteroplasmy were ignored 263G, 315.1C, 16519 would be the most common haplotype in our data set with 22 observations (4.0%). The haplogroups for the 550 individuals were H (268), U (79), J (47), T (47), K (52), and I, M, N, P, R, S, V, W and X (57), all with European origins, confirming the ancestral roots of the dataset. A quality assessment of the dataset was performed for entry of the haplotypes into EMPOP; 494 of the 550 profiles were sent to EMPOP for upload, as 54 samples from the Shriver laboratory are still being assessed to determine if consent is suitable for upload to a public database, and 2 samples were presumed duplicates from previous studies.

Prior to analysis of heteroplasmy in our dataset, the average assumed error for each nucleotide (A, C, G, and T) was assessed by considering all base call information <50% of the read density. The consensus statistic report that is generated by GeneMarker® HTS was manipulated in order to combine the forward and reverse reads, the frequency of reads based on total coverage was calculated, and the frequencies were transformed back into counts to produce the assumed error. A combination of Terminal and R Studio was used to mine the data and calculate the assumed total and individual nucleotide error rates. A summary of the R Studio output is presented in Appendix Table 2. The numbers in the summary table represent a percent of the total reads, which can also be described as the number of calls made in error for every 100 nucleotides assessed. The average total error rate was 0.17±0.06 erroneous base calls for every 100 nps. The average assumed error for each nucleotide (A, C, G, and T) was 0.04±0.01, 0.05±0.02, 0.04±0.02, and 0.04±0.01 per 100 nps, respectively. Appendix Figure 1 is a boxplot of the data for each nucleotide position, along with the total combined error (a sum of the error rates for the individual nucleotides), illustrating that a significant different between the error rates for the individual nucleotides does not exist. The average total assumed error was well below our analytical threshold of 1%, and our reporting threshold of 2%, indicating that heteroplasmic positions reported at 2% are clearly above the system noise. Using this approach has proven to be robust when reporting minor sequence variants, as illustrated through precision studies conducted on mtDNA MPS data [10]. We are in the process of developing a tool, that will be made available to the forensic community (at no cost), to assess datasets for assumed error. It should be noted that development of the final tools was impacted by instability in existing laboratory methods and the repeated release of GeneMarker® HTS iterations, extending the length of the analysis process.

Following the final analysis of our MPS dataset, the rate of observing heteroplasmy at a reporting threshold of 2% (n=537) was 41.15% (221/537), with 9.87% of individuals exhibiting more than one position of heteroplasmy (Appendix Figure 2). Heteroplasmy was reported a total of 292 times, with 78 (26.71%) of the observations having a read frequency of 10-49.93%. These 78 sites were observed in 75 individuals, for a rate of 13.64% at Sanger-type sequencing (STS) detection levels. While this is consistent with previous findings [14], it is a high frequency given that most of the frequencies were close to 10%, or the limits of STS. Interestingly, we saw no significant correlation between rates of heteroplasmy and age (Appendix Figure 3) or gender. Lastly, data can become "noisy" given the introduction of low-level foreign DNA and the sensitivity of mtDNA testing. To account for this, the threshold was raised to 3% for 13 of the 550 samples in the current project.

A total of 86 nps (~7.7%) exhibited heteroplasmy across the 1,122 sites in the CR. The most prevalent type of heteroplasmy was C/T-based with 172 observations (60.1%), followed by A/G (79, 27.6%), A/C (34, 11.9%), and G/T (1, 0.35%). The C/G and A/T transversions produced no observations. Overall, the vast majority of nps exhibited no heteroplasmy (~93.3%). 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 (~6.36%), would result in an increase in the LR to ~15,750. 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.

We assessed the transmission of mtDNA sequence variants in different tissue types focusing on hair shafts, a common source of evidence in forensic cases. This dataset is in the process of being thoroughly analyzed, but preliminary findings suggest that known variants drift rapidly between tissue types and different hairs, and new sites of presumed heteroplasmy are revealed which may be associated with DNA damage. In addition, we assessed the transmission of mtDNA sequence variants in different tissue types through a multi-generational family study. As expected, we observed reproducible tissue-specific heteroplasmy and differences in heteroplasmy between maternal relatives, consistent with previous findings [5]. However, we also observed positions of heteroplasmy which appeared to exhibit recurrent mutational events that accumulate with age, and that are selected against during germline transmission [15].

CRIMINAL JUSTICE IMPACT

In total, we presented our work on this project at least 12 times over the course of 24 months. We organized a one-day workshop in Minneapolis, MN for 32 scientists at the MN BCA and other regional laboratories; provided training to MN BCA examiners on MPS data analysis; gave oral presentations at 5 different workshops in the U.S. (NC State University/Promega 2015, CA DOJ/Promega 2016, OCME/Promega 2016, Indianapolis, IN/Promega 2016, Bode Mid-

Atlantic/Illumina 2016), reaching scientists from the CA DOJ, OCME, NC State Police, and Philadelphia PD; gave oral presentations at 2 different conferences in the U.S. (ISHI 2016, AAFS/NIJ Forensic Science R&D Symposium 2016), reaching broad forensic audiences; gave oral presentations at 3 different international conferences (ISABS 2015, ISFG 2015, Genetics in Forensics Congress 2016), reaching broad groups of international scientists; and gave multiple poster presentations.

Our work with two consultants has ensured that the project is relevant in both National and International forensic circles. Ann Gross is a member of SWGDAM and is responsible for leading the development of guidelines for mtDNA MPS analysis. We are in the process of working with the MN BCA laboratory to help them develop a plan for the implementation of an mtDNA MPS method. Walther Parson is the current President of the International Society for Forensic Genetics, and is a leading mitochondrial geneticist.

Ultimately, the most important outcome of our project will be publications. We already have one publication in press [7], have a second publication submitted [8], and are in the process of writing at least three additional manuscripts [2,6,8]. In addition, the work behind one other publication [3] and a manuscript in preparation [10] had a meaningful impact on the outcomes of this project, and include information that relates to the project findings. Therefore, we anticipate that at least five publications will result from this project, with at least two additional papers directly tied to elements of the project. Each of our papers provides recommendations that the forensic community can use to develop best practices as they implement mtDNA MPS methods in their laboratories. Therefore, the outcomes of this project should have a significant impact on forensic mtDNA casework and the criminal justice system.

APPENDIX

BIBLIOGRAPHY

1. Ivanov PL, Wadhams MJ, Roby RK, Holland MM, Weedn VW, Parsons TJ. Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. Nat Genetics. 1996;12:417-20.
2. McElhoe JA, Holland MM. Measuring rates of low-level mtDNA heteroplasmy in a European population, manuscript in preparation.
3. Rathbun MM, McElhoe JA, Parson W, Holland MM**.** Considering DNA damage when interpreting mtDNA heteroplasmy in deep sequencing data. Forensic Sci Int: Genetics. 2017(online 28 September 2016);26:1-11.
4. 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.
5. Rebolledo-Jaramillo B, Su MS-W, Stoler N, McElhoe JA, Dickins B, Blankenberg D, Korneliussen T, Chiaromonte F, Nielsen R, Holland MM, et al. Maternal Age Effect and Severe Germline Bottleneck in the Inheritance of Human Mitochondrial DNA. Proc Natl Acad Sci. 2014;111:15474-9.
6. Gallimore J, Holland CA, McElhoe JA, Holland MM. Transmission of mtDNA heteroplasmy in human hairs using an MPS approach, manuscript in preparation. A summary of the findings can be found at the following link; http://www.ishinews.com/the-comparison-of-dna-extractions-from-hair-shafts-and-evaluation-of-the-10-plex-powerseq-kit-by-promega-with-low-copy-number-samples/.

7. 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.

8. Holland MM, E Pack, JA McElhoe. Evaluation of GeneMarker® HTS for improved alignment of mtDNA MPS data, haplotype determination, and heteroplasmy assessment. Forensic Sci Int: Genetics, submitted for consideration in October 2016.

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

10. Holland MM, Wilson LA, Copeland S, Dimick G, Holland CA, Bever, B, McElhoe JA. MPS analysis of the mtDNA hypervariable regions on the MiSeq with improved enrichment. Int J Legal Med, submitted for consideration in July 2016, resubmitted in December 2016.

11. Peck MA, Brandhagan MD, Marshall C, Diegoli TM, Irwin JA, Sturk-Andreaggi K. Concordance and reproducibility of a next generation mtGenome sequencing method for high-quality samples using the Illumina MiSeq. Forensic Sci Int: Genetics. 2016;24:103-11.

12. Holland MM, Parsons T. Mitochondrial DNA sequence analysis – validation and use for forensic casework. Forensic Sci Rev. 1999;11:21-50.

13. Just RS, Irwin JA, Parson W, Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. Forensic Sci Int: Genetics. 2015;18:131-139.

14. Irwin JA, Saunier JL, Niederstatter H, Strouss KM, Sturk KA, Diegoli TM, Brandstatter A, Parson W, Parsons TJ. Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. J. Mol. Evol. 2009;68:516–27.

15. Holland MM, McElhoe J, Recurrent Mutations at Positions 185 and 189 of the Mitochondrial DNA Control Region Show Evidence of Selection and Age-Related Accumulation, manuscript in preparation.

CODE FOR ERROR ASSESSMENT IN MPS DATA

```
#error estimation

#terminal to prepare the data for R analyses

#create folder with all the consensus statistic files

"#create another, empty folder called ""no_header"""

#remove the header information from all the consensus statistic files

"$for file in *.txt; do tail -n +2 $file> ""$(basename $file)_new.txt""; done"


"#move all the no header files into ""no_header"""

"#concatonate all the *.txt files in ""no_header"""

$cat *.txt>PR1All.txt


#open the file in excel or the like and re-add the header information

#save the file with the header information and proceed with R script
```

R Studio code

Error assessment

"The next 11 chunks were run multiple times, each time changing the working directory (i.e. Run_1_error_no_header) and changing the text file to read into a table (i.e. PR1All.txt)"

Load the concatonated data file with header information

~~~
```{r load_file}
~~~

~~~
wdir=getwd()

"setwd(""~/hetero/Run_4_error_no_header"")"

"df=read.table(""PR4All.txt"",header=TRUE)"

```



remove positions outside of the control region

```{r CR_only}

"df=rbind(subset(df, chr_pos<577), subset(df, chr_pos>16023))"

```



designate positions as categorical instead of integers

"```{r categorical, include=FALSE}"

as.factor(df$chr_pos)

```



combine forward and reverse reads

```{r total_reads}

df$A<- df$AF + df$AR

df$C<- df$CF + df$CR

df$G<- df$GF + df$GR

df$T<- df$TF + df$TR

df$del<- df$delF + df$delR

df$ins<- df$insF + df$insR
~~~

~~~
```

remove the columns containing Forward and Reverse read information
```{r rm_FandR_cols}
"df <- subset( df, select = -c(AF,AR,CF,CR,GF,GR,TF,TR,delF,delR,insF,insR) )"
```

create a column with calculation of the percent of total coverage at each chr_pos
```{r percent_cov}
df$Aper<-df$A/df$coverage*100
df$Cper<-df$C/df$coverage*100
df$Gper<-df$G/df$coverage*100
df$Tper<-df$T/df$coverage*100
df$del<-df$del/df$coverage*100
df$ins<-df$ins/df$coverage*100
```

"for a conservative estimate of error, we are going to assume any value greater than 50% is called in error.  we will set the threshold at 50% and evaluate all the values as TRUE or FALSE based on this threshold"
```{r threshold_eval}
thr=50
df$Athr<-df$Aper<=thr
~~~

~~~~
df$Cthr<-df$Cper<=thr
df$Gthr<-df$Gper<=thr
df$Tthr<-df$Tper<=thr
```
"Don't really need indel error therefore, evaluating ACGT moving forward "


"remove all values >50% based on the TRUE FALSE evaluation. since TRUE=1 and FALSE=0, any value evaluated as FALSE will not be carried through"
```{r rm_>50}
df$AthrV<-df$Aper*df$Athr
df$CthrV<-df$Cper*df$Cthr
df$GthrV<-df$Gper*df$Gthr
df$TthrV<-df$Tper*df$Tthr
```


"to calculate the total assumed error, sum the error associated with ACG&T "
```{r totAE_1}
df$TotAEper<-df$AthrV + df$CthrV + df$GthrV + df$TthrV
```


need to covert the calculated percentage of total coverage back into number of calls for each nucleotide
```{r nucleotide_calls}
~~~~

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

~~~
df$TotAEcalls<-df$TotAEper/100*df$coverage

df$AthrVcalls<-df$AthrV/100*df$coverage

df$CthrVcalls<-df$CthrV/100*df$coverage

df$GthrVcalls<-df$GthrV/100*df$coverage

df$TthrVcalls<-df$TthrV/100*df$coverage

```
~~~

to calculate the values for total assumed error and error associated with each nucleotide (ACGT)

~~~
```{r final_error_calcs}

TotAssumedError<-(sum(df$TotAEcalls))/sum(as.numeric(df$coverage))*100

Aerror<-(sum(df$AthrVcalls))/sum(as.numeric(df$coverage))*100

Cerror<-(sum(df$CthrVcalls))/sum(as.numeric(df$coverage))*100

Gerror<-(sum(df$GthrVcalls))/sum(as.numeric(df$coverage))*100

Terror<-(sum(df$TthrVcalls))/sum(as.numeric(df$coverage))*100

TotAssumedError

Aerror

Cerror

Gerror

Terror


```
~~~

boxplot of values generated in error assessment

~~~
```{r error_boxplot}

getwd()

"setwd(""~/hetero/"")"

"errorValues<-read.csv(""error_summary.csv"", header = TRUE)"

"boxplot(errorValues[,2:6])"

```
~~~

TABLES & FIGURES

Table 1: Metadata for the 550 European data set, including gender, age, and threshold applied. We collected 717 buccal samples; 130% of expected. The following is a list of reasons for omission of 167 collected samples from our analysis: 24 samples failed the DNA extraction step, presumably due to poor collection; 95 samples were of non-European ancestry, as reported by the donor; 17 samples were of non-European ancestry, as uncovered through our laboratory analysis; 15 samples were reported as relatives of a donor; and 16 samples were contaminated.

	Female				Male			
	18-29	30-49	50+	Tot. Female	18-29	30-49	50+	Total Male
No. samples 2% threshold	145	102	52	299	74	99	65	238
No. samples 3% threshold	2	2	3	7	1	5	0	6

Table 2: Output summary of the average assumed error rates calculated using R Studio for data taken from four MiSeq runs that represented 230 samples. Error rates presented as an estimation of the number of calls made in error per 100 nucleotides. TotAE is the total assumed error, and A, C, G, and Terror represent the error for each nucleotide.

	Tot. Assumed Error	A Error	C Error	G Error	T Error
Run1	0.13	0.03	0.04	0.04	0.03
Run2	0.22	0.04	0.07	0.06	0.05
Run3	0.23	0.05	0.07	0.06	0.05
Run4	0.12	0.03	0.04	0.02	0.03
Averge	0.18	0.04	0.05	0.05	0.04
Standard Deviation	0.06	0.01	0.02	0.02	0.01

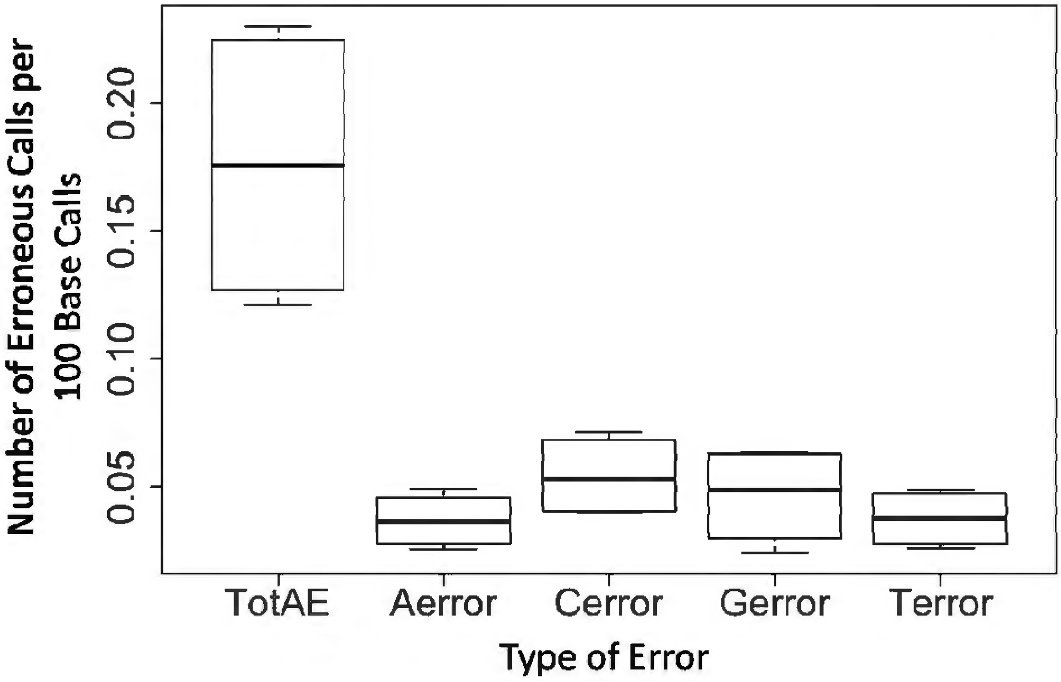

Fig 1: Boxplot of assumed error generated using R Studio. Total assumed error (TotAE) and nucleotide error (A, C, G, and Terror) is on the x-axis with the number of base calls in error per 100 nucleotides on the y-axis.

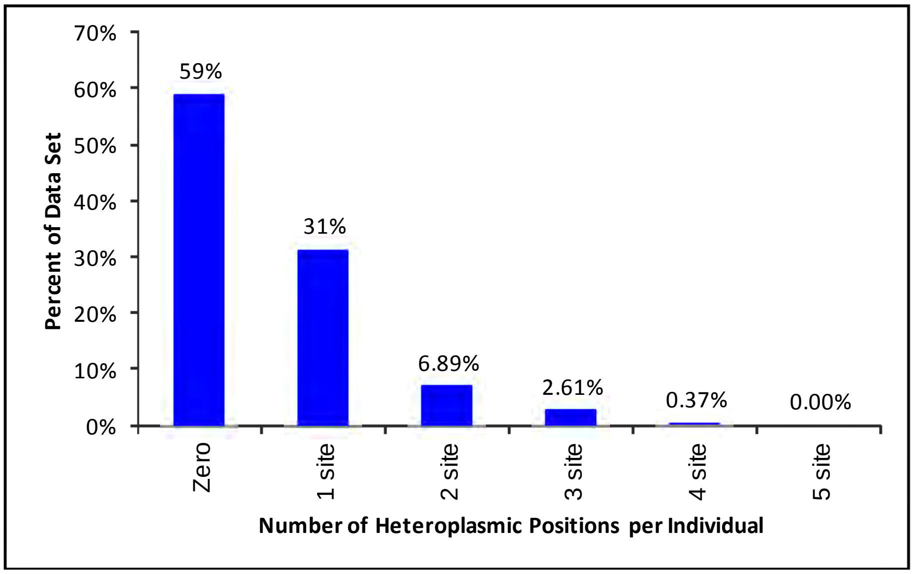

Figure 2: Rates of heteroplasmy on a per individual basis (n=537). The vast majority of individuals (90.12%) have either no heteroplasmy (58.84%) or one site of heteroplasmy (31.28%).

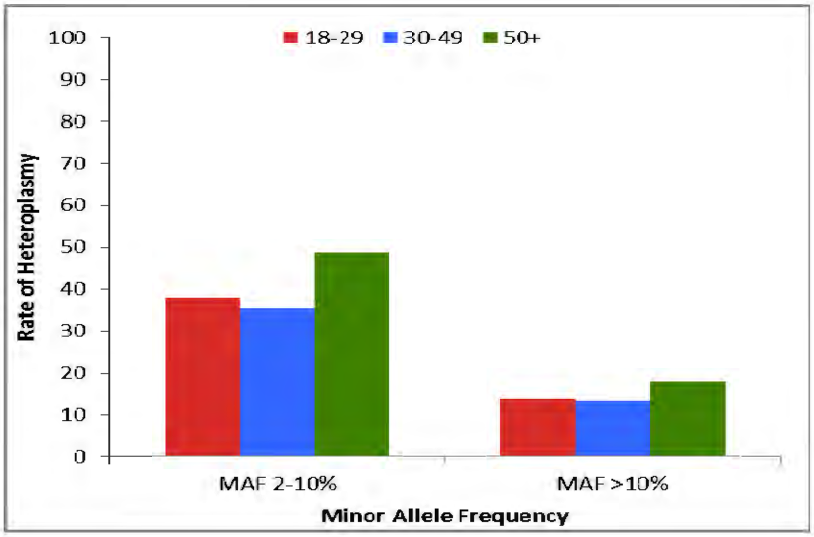

Figure 3: Rates of heteroplasmy on a per individual basis (n=537), and when considering the age of the individual.