TALLINN UNIVERSITY OF TECHNOLOGY
DOCTORAL THESIS
15/2020

Massively Parallel Sequencing of Human Mitochondrial Genome for Forensic Analysis

MONIKA STOLJAROVA

TAL TECH PRESS

TALLINN UNIVERSITY OF TECHNOLOGY
School of Science
Department of Chemistry and Biotechnology
This dissertation was accepted for the defence of the degree 19/05/2020

Supervisor: Dr. Anu Aaspõllu
Department of Chemistry and Biotechnology
Tallinn University of Technology
Tallinn, Estonia

Co-supervisor: Dr. Bruce Budowle
Center for Human Identification
University of North Texas Health Science Center
Fort Worth, TX, USA

Opponents: Prof Ángel Carracedo Álvarez
Institute of Forensic Sciences
University of Santiago de Compostela
Santiago de Compostela, Galicia, Spain

Prof Mait Metspalu
Institute of Genomics
University of Tartu
Tartu, Estonia

Defence of the thesis: 22/06/2020, Tallinn

Declaration:
Hereby I declare that this doctoral thesis, my original investigation and achievement, submitted for the doctoral degree at Tallinn University of Technology has not been submitted for doctoral or equivalent academic degree.

Monika Stoljarova

----
signature

European Union
European Regional
Development Fund

Investing
in your future

ISSN 2585-6898 (publication)
ISBN 978-9949-83-552-2 (publication)
ISSN 2585-6901 (PDF)
ISBN 978-9949-83-553-9 (PDF)

TALLINNA TEHNIKAÜLIKOOL
DOKTORITÖÖ
15/2020

Inimese mitokondriaalse genoomi massiivselt paralleelne sekveneerimine forensiliseks analüüsiks

MONIKA STOLJAROVA

TALTECH
KIRJASTUS

Contents

List of Publications

The list of author's publications, on the basis of which the thesis has been prepared:

I King, J. L., LaRue, B. L., Novroski, N. M., Stoljarova, M., Seo, S. B., Zeng, X., Warshauer, D. H., Davis, C. P., Parson, W., Sajantila, A., & Budowle, B. (2014). High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. *Forensic Sci Int Genet*, 12, 128-135. doi:10.1016/j.fsigen.2014.06.001

II Stoljarova, M., King, J. L., Takahashi, M., Aaspollu, A., & Budowle, B. (2016). Whole mitochondrial genome genetic diversity in an Estonian population sample. *Int J Legal Med, 130*(1), 67-71. doi:10.1007/s00414-015-1249-4

III Churchill, J. D., Stoljarova, M., King, J. L., & Budowle, B. (2018). Massively parallel sequencing-enabled mixture analysis of mitochondrial DNA samples. *Int J Legal Med, 132*(5), 1263-1272. doi:10.1007/s00414-018-1799-3

Author’s Contribution to the Publications

Contribution to the papers in this thesis are:

I Performance of the experimental work, analysis and interpretation of the data

II Performance of the experimental work, analysis and interpretation of the data, preparation of the manuscript

III Performance of the experimental work, analysis and interpretation of the data

Introduction

In forensic identification mitochondrial DNA (mtDNA) is used for kinship analysis and in the cases of degraded/low quality DNA samples (e.g. teeth, hair, bones, touch traces). Although mtDNA has lower discrimination power compared to nuclear DNA short tandem repeat (STR) markers, lack of recombination and strictly maternal inheritance make mtDNA a valuable marker system for kinship analysis. In addition, high number of mtDNA copies per cell – on average 500 copies versus two copies of nuclear DNA – increases the success rate of the analysis of low level DNA samples.

Based on commonly shared single nucleotide polymorphisms (SNPs) mtDNA sequences, i.e. haplotypes, are divided into genealogical groups (haplogroups) that share a common ancestor. The migration map of women 150,000 years before present from Africa onto different continents has been reconstructed using mtDNA analysis. Haplogroup assignment has been suggested to be used as a quality-control measure for mtDNA profiles. In the forensic context mtDNA typing of population samples is important for database establishment and estimating mtDNA haplotype frequencies.

Mixtures - samples originating from two or more donors and are commonly recovered from touch, sexual and/or physical assault crime scenes - are one of the most challenging samples in forensic investigations. Mixture deconvolution via autosomal STR loci and the following statistical assessment of the results can be complicated (e.g., addressing allele overlap among an unknown number of contributors and allele drop out). The allele ratios of mtDNA mixture positions can be used as an alternative or adjunct tool for mixture interpretation.

The gold standard for mtDNA analysis in forensic investigations is the Sanger-type sequencing (STS) of hypervariable regions I and II (HVI and HVII, respectively). Sequencing beyond HVI/HVII, that is approximately the other 96.3% of the mitochondrial genome (mtGenome), is rarely attempted as the technique is laborious, time consuming and expensive. However, it has been shown that discrimination power of the entire mtGenome data is much higher than that of HVI/HVII and can provide resolution of common HVI/HVII haplotypes. Moreover, it has been shown that STS has a number of limitations when mixture deconvolution is attempted.

Massively Parallel Sequencing (MPS) provides substantial increase in throughput which makes it a desirable alternative to STS. MPS tracks the incorporation of each nucleotide of each molecule (or clone) as the DNA chain is elongated during primer extension. Therefore, interpretation of point as well as length heteroplasmy can be attempted. The sequencing of the entire mtGenome via MPS as well as the quantitative data generated can provide additional information for mixture interpretation.

The aim of this thesis was, firstly, to evaluate the feasibility of an MPS system for generating entire mtGenome sequences. Secondly, to compare the discrimination power of mtGenome to HVI/HVII data in different population samples. Thirdly, the interpretation of mixture samples by using MPS generated mtDNA quantitative data, phasing information and phylogenetic assignment.

Abbreviations

AFA	African-American population
aiSNP	ancestry informative SNP
ATP	adenosine triphosphate
bp	base pair
CAU	Caucasian population
CE	capillary electrophoresis
ChrX	chromosome X
ChrY	chromosome Y
CO	cytochrome c oxidase
CODIS	Combined DNA Index System
CR	control region
CRS	Cambridge Reference Sequence
cyt	cytochrome
ddNTP	di-deoxynucleotidetriphosphates
ENFSI	European Network of Forensic Science Institutes
ESS	European Standard Set
EST	Estonian population
FBI	US Federal Bureau of Investigation
Gb	gigabases
GD	genetic diversity
HFH	high-frequency heteroplasmy
HIS	Hispanic population
HVI	hypervariable region I
HVII	hypervariable region II
IGV	Integrative Genomic Viewer
Indel	insertion/deletion
MMR	mismatch repair
MPS	massively parallel sequencing
mtDNA	mitochondrial DNA
mtGenome	mitochondrial genome
mt-MRCA	the Mitochondrial Eve
MYR	million years
NADH	reduced nicotinamide adenine dinucleotide
ND	reduced nicotinamide adenine dinucleotide dehydrogenase
nDNA	nuclear DNA
NDNAD	UK National DNA Database
NER	nucleotide excision repair
NGS	next generation sequencing
NHEJ	nonhomologous end joining
np	nucleotide position

NTC	no-template control
NUMT	mitochondrial DNA segments inserted in the nuclear genome
O_H	origin of heavy strand
O_L	origin of heavy strand
PCR	polymer chain reaction
PGM	Personal Genome Machine
PHP	point heteroplasmy
rCRS	revised Cambridge Reference Sequence
RLP	relative locus performance
RMP	random match probability
rRNA	ribosome RNA
SBS	sequencing-by-synthesis
SD	standard deviation
SNP	single nucleotide polymorphism
STR	short tandem repeat
STS	Sanger-type sequencing
SWGDAM	The Scientific Working Group on DNA Analysis Methods
tRNA	transfer RNA
YBP	years before present

1 Literature overview

1.1 DNA typing for human identification

DNA analysis is a valuable approach in forensic investigation in case of biological evidence such as body fluids (blood, saliva and semen stains), hair, teeth and bones. A number of markers used in forensic genetics have been established. The choice for DNA analysis approaches to be performed, and therefore markers used, is based on the collected sample type, assumption of the DNA quantity and quality, and the reference samples available. One the biggest challenges in human identification with this technology is low level DNA samples as well as mixture samples (Scientific Working Group on DNA Analysis Methods, 2017; ENFSI DNA Working Group, 2017).

Short tandem repeats (STRs) have become the primary markers in forensic genetics due to their high discrimination power as well as relatively short amplicon size. Among the various types of STRs, tetranucleotide (four base pair, bp) repeats are most commonly used markers as they enable design of small polymer chain reaction (PCR) product size assays that increase the success rate of the recovery of data from degraded samples; have a low stutter formation increasing the interpretation success of the data and are suitable for separation during capillary electrophoresis (Scientific Working Group on DNA Analysis Methods, 2017). In order for data exchange between local, national and global jurisdictions, core STR loci have been chosen. The UK National DNA Database (NDNAD) launched in 1995 by the United Kingdom Home Office was the world's first national DNA database and originally stored data from six STR loci. The Combined DNA Index System (CODIS) based on 13 STR loci was adopted in 1998 in the United States (Budowle, 1998). One year later, the DNA working group of the European Network of Forensic Science Institutes (ENFSI) decided on a European Standard Set (ESS), which included seven loci and adapted five additional loci in 2009 (Council of the European, 2001, 2009). Using all ESS 12 STR loci the random match probability among unrelated Caucasian individuals is approximately 9.66 × 10–16 (Butler, Hill & Coble, 2012). The US Federal Bureau of Investigation (FBI) CODIS Core Loci Working Group announced the adoption of an additional seven STR loci expanding the CODIS core loci to 20 starting from the 1st of January, 2017 (Hares, 2015). The random match probability for CODIS 20 STR loci is approximately 9.35 × 10–24 (Butler *et al*., 2012).

Single Nucleotide Polymorphisms (SNPs) is the second group of most common markers used in forensic genetics. The benefits of SNPs over STRs are: no shutter artifacts; smaller PCR product size which should increase the successful recovery of DNA from degraded samples; more loci can be multiplexed than with current STRs; and provide additional data on phenotypic traits (hair, skin, iris colour etc.) as well as ethnic origin. Data on external visible characteristics and ethnicity provide important investigation leads in case STR data yield no DNA database match or there is no apprehended suspect to compare with the evidence DNA profile. Walsh *et al.*, presented IrisPlex, the initial eye colour prediction test based on six SNPs and demonstrated its capability of generating complete results from common crime scene sample with DNA quantity as low as 31 pg (Walsh, Liu, *et al.*, 2011). The team suggested IrisPlex is ready for immediate implementation and use in any accredited forensic laboratory for aiding DNA intelligence investigations (Walsh, Lindenbergh, *et al.*, 2011). IrisPlex was followed by development and validation of the HIrisPlex assay which targeted 24 eye and hair colour predictive DNA variants including all six IrisPlex SNPs (Walsh *et al.*, 2014; Walsh *et al.*, 2013). In 2018 Chaitanya *et al.*, published data on the HIrisPlex-S system composed of two multiplex assays – previously validated HIrisPlex assay targeting 24 SNPs and a novel 17-plex assay targeting skin colour-prediction SNPs. These 41 targeted SNPs generate probabilities for three eye colour, four hair colour and five skin colour categories. The majority of the skin 17-plex SNP assay PCR amplicons' size is 170 bp or less. The assay was validated based on the Scientific Working Group on DNA Analysis Methods (SWGDAM) guidelines, and full profiles were generated from samples with as little as 63 pg of input DNA. The eye, hair, and skin colour predictions were in concordance in 28 of the 30 samples (93.3%) (Chaitanya *et al.*, 2018). A number of SNP panels for bioancetry (origin) determination have been developed for analysis via capillary electrophoresis (CE) as well as massively parallel sequencing (MPS) (de la Puente *et al.*, 2016; Jager *et al.*, 2017; Phillips *et al.*, 2014). The Kidd and Seldin union panel of 170 biogeographical ancestry informative SNPs (aiSNPs) was used to generate and analyse the genotype data of 3,933 individuals from 81 worldwide populations (Pakstis *et al.*, 2019). Jäger *et al.*, reported MiSeq FGx™ Forensic Genomics System (Illumina, San Diego, CA, USA) which simultaneously amplifies up to 231 forensic loci including 56 biogeographical aiSNPs to meet SWGDAM developmental validation guidelines (Jager *et al.*, 2017).

In missing person cases, where reference samples from family members are compared indirectly by kinship analysis autosomal STR markers also are the primary genetic markers as well as, on a case-by-case basis, STRs located on the sex chromosomes X and Y (ChrX and ChrY, respectively). ChrX STR typing adds value in specialized kinship analysis cases such as when a female child and a missing putative father are compared (Trindade-Filho, Ferreira, & Oliveira, 2013) or in some incest cases; while ChrY typing yields data on relatedness via a paternal lineage, as in the identification case of the son of Tsar Nicholas II (Coble *et al.*, 2009). Additionally, ChrY analysis is used for analysis of the male DNA profile from mixture samples in sexual assault cases, as autosomal STR typing could be limited due to high female DNA background in the sample that masks the minor amount of male DNA (Prinz & Sansone, 2001). As ChrY is inherited via paternal transmission, Y-STRs (as well as maternally inherited mitochondrial DNA SNP markers for maternal lineages) are used in studying human migration patterns and genealogical research (Lappalainen *et al.*, 2008).

Insertion/deletion (indel) polymorphism typing has been proposed by a number of groups as an adjunct or viable alternative to STR and SNP analysis for human identification. A number of multiplex panels have been developed (Bastos-Rodrigues, Pimenta, & Pena, 2006; LaRue *et al.*, 2014; Pereira *et al.*, 2009; Zidkova, Horinek, Kebrdlova, & Korabecna, 2013). The following characteristics of indel typing is considered also advantageous over STRs and/or SNPs analyses: short amplicon size, low mutation rate, deriving from a single mutational events, absence of stutter, and can be analysed by CE. (Weber *et al.*, 2002).

1.2 Genetics of mitochondrial DNA

In addition to DNA present in the nucleus (nuclear DNA, nDNA) within the eukaryotic cells, mitochondria, adenosine triphosphate (ATPs) generating organelles, separately contain a number of small, circular genomes, which encode a series of crucial proteins for cellular respiration. On average each mitochondrion contains four to five mitochondrial DNA (mtDNA) molecules, each cell contains up to 100 mitochondria, resulting in at least 500 mtDNA molecules per cell compared to two copies of 23 nDNA molecules (Butler, 2012). Mitochondrial DNA copy number per cell as high as 800 and 1720 have been reported in human lung fibroblasts and rabbit lung macrophage, respectively (Robin & Wong, 1988). Miller *et al.*, presented a PCR-based mtDNA copy number measuring assay and reported 3,650 ± 620 (mean ± standard error) and 6,970 ± 920 as an average for mtDNA copy number per cell in skeletal muscle and myocardium tissues, respectively (Miller, Rosenfeldt, Zhang, Linnane, & Nagley, 2003). While Miller *et al.*, saw no mtDNA copy number correlation with age in human skeletal and cardiac muscle (Miller *et al.*, 2003), other groups reported mtDNA copy number correlation with age dependent of the tissue type: negative correlation in peripheral blood of a healthy Chinese population sample (Xia *et al.*, 2017), negative correlation in skeletal muscle samples and positive correlation in liver samples (Wachsmuth, Hubner, Li, Madea, & Stoneking, 2016). The variance (increase and decrease) in mtDNA copy number has been linked to a number of disorders like tumorigenesis (Keseru *et al.*, 2019), neurodegenerative diseases (Alvarez-Mora *et al.*, 2019), cardiovascular diseases (Lin *et al.*, 2019) and type 2 diabetes (Skuratovskaia, Zatolokin, Vulf, Mazunin, & Litvinova, 2019).

The double-stranded circular mitochondrial genome (mtGenome) is approximately 16,569 bp long but can vary in length due to small insertions and deletions (Fig. 1). The purine-rich strand is designated as the heavy strand, and the pyrimidine-rich strand as called the light strand. The mtDNA molecule is divided into a ~15,448 bp long "coding region" comprised of 37 genes that are transcribed into 13 proteins, 2 ribosomal RNAs (rRNAs) and 22 transfer RNAs (tRNAs), and a ~1,121 bp long D-loop or the "control region" (CR) that contains the origin of replication for the heavy strand. As no gene products are encoded in the D-loop region, it is commonly referred to as the "non-coding region" (Anderson *et al.*, 1981; Butler, 2012). The highest polymorphism concentration in the mtGenome resides within the CR as there are fewer constraints for nucleotide variability due to lack of transcribed sequences. Two regions within the CR with the highest concentration of variations are commonly referred to as hypervariable region I and II (HVI and HVII, respectively) (Greenberg, Newbold, & Sugino, 1983).

Mitochondrial DNA exhibits a higher mutation rate compared to nDNA, with HVI and HVII exhibiting the highest mutation frequency in the mtGenome (Pakendorf & Stoneking, 2005). Previously, a limited mtDNA repair mechanism was considered a reason for the higher mutation rate. However a recent study suggests, based on the repair mechanism enzymes found in the mitochondria, that mtDNA uses the same repair mechanisms [nucleotide excision repair (NER), mismatch repair (MMR), homologous recombination, and classical nonhomologous end joining (NHEJ)] as does the nDNA (Zinovkina, 2018). Oxidative damage to DNA, in the most common form of urinary excretion of 8-oxodG, has been considered as one of the main drivers of mtDNA (as well as nDNA) mutagenesis (Kirkwood & Kowald, 2012). However, Kennedy *et al.*, argue that replication errors by DNA polymerase γ and/or spontaneous base hydrolysis, but not the oxidative damage, are responsible for the majority of point mutations that accumulate in mtDNA, the frequency of which increases with age (Kennedy, Salk, Schmitt, & Loeb, 2013).

As there are a number of mtDNA molecules in a cell, a mutation (i.e., a de novo variant) can affect either all of the mtDNA molecules resulting in a homoplasmy or a proportion (one or more) of mtDNA molecules leading to heteroplasmy – a notable feature of mtDNA genetics. The level of heteroplasmy can vary between cells in the same tissue, between tissues of the same individual and between individuals from the same maternal lineage. Heteroplasmy has been reported in all, several or just one tissue of an individual

(Li, Schroder, Ni, Madea, & Stoneking, 2015). Moreover, Li *et al.*, reported the existence of high-frequency heteroplasmy (HFH) sites that are allele and tissue, as well as age specific, e.g. nucleotide position (np) 72 shows high levels of heteroplasmy in the liver and kidney, moderate levels in skeletal muscle, and low levels in all other tissues, whereas np 189 shows high levels of heteroplasmy in skeletal muscle but low heteroplasmy levels in other tissues (Li *et al.*, 2015). Heteroplasmy has been associated with a wide range of metabolic diseases, degenerative diseases, cancer and aging, with the severity of the disease linked to the heteroplasmy level and the copy number (Grady *et al.*, 2018; He *et al.*, 2010; Khusnutdinova *et al.*, 2008; Lin *et al.*, 2019; Szklarczyk, Nooteboom, & Osiewacz, 2014; Zhang *et al.*, 2009).

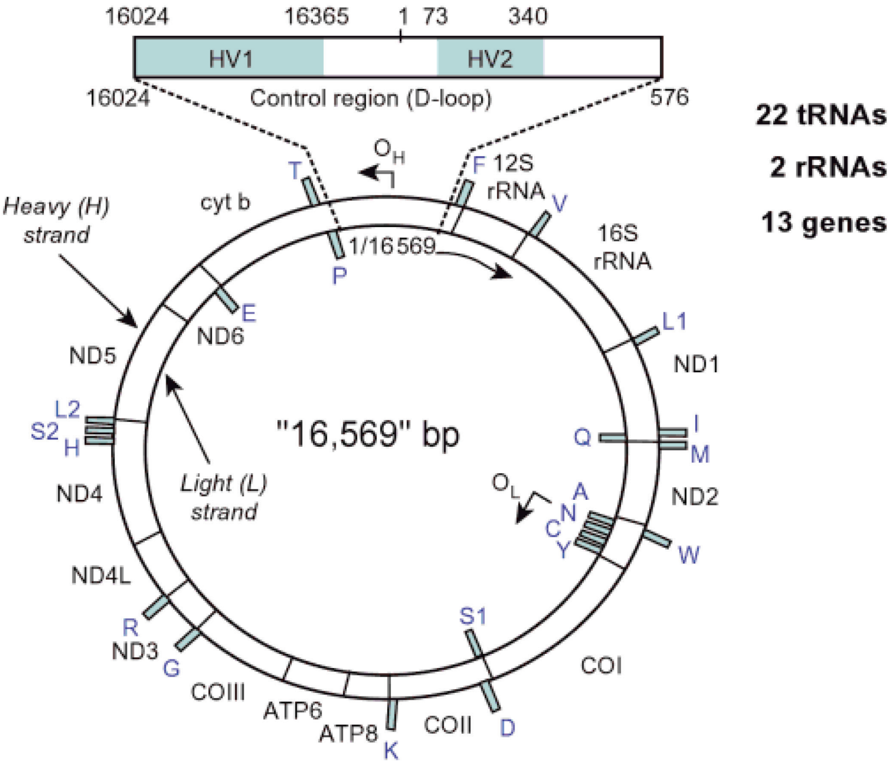


Fig. 1. The structure of mtGenome (edited (Butler, 2012)). The 2 rRNAs and 13 protein coding genes are marked next to the strand from which they are transcribed. Grey rectangles mark the 22 tRNA genes. The zoom in of the control region (D-loop) is shown at the top with HVI (np 16024-16365) and HVII (np 73-340) regions marked in grey. Abbreviations: O_H – origin of heavy strand; O_L – origin of light strand; ND – reduced nicotinamide adenine dinucleotide (NADH) dehydrogenase; CO – cytochrome c oxidase; cyt – cytochrome; ATP - adenosine triphosphate synthase membrane subunit.

Another characteristic of mtDNA genetics is the presence of mitochondrial DNA segments that have inserted in the nuclear genome (NUMTs). These NUMTs are specific regions of the mitochondrial genome that have corresponding nuclear mitochondrial pseudogenes that are distributed across multiple nDNA chromosomes and may have been inserted multiple times. Based on the data from The Human Genome Project, Mourier *et al.*, reported that all positions of the mitochondrial genome are represented in the nDNA. They reported 296 NUMTs ranging between 106 and 14,654 bp in size with four NUMTs covering the complete control region. Ongoing genetic material transfer between mtDNA and nDNA also has been observed (Mourier, Hansen, Willerslev, & Arctander, 2001). Dayama *et al.*, reported 141 polymorphic (non-fixed) NUMTs positions in 999 individuals. One sample had a 16,106 bp long NUMT integrated into a potential regulatory region in the first intron of the SDC2 gene (chromosome 8) (Dayama, Emery, Kidd, & Mills, 2014). Hazkani-Covo *et al.*, summarized a number of cases where NUMTs caused misinterpretation of amplified mtDNA data and misleadingly associated to diseases. However in five cases the research team reported involvement of NUMTs in medical disorders like plasma factor VII deficiency, Pallister-Hall syndrome, mucolipidosis IV and Usher syndrome (Hazkani-Covo, Zeller, & Martin, 2010).

The entire human mtGenome was sequenced for the first time in 1981 from placental material of an individual of European descent in the laboratory of F. Sanger by Anderson *et al.*, and is referred to as the Cambridge Reference Sequence (CRS) (Anderson *et al.*, 1981). In 1999 due to the improvement of sequencing technologies it was decided to re-sequence the original placental material used by Anderson *et al.* The reanalysis of the mtDNA sequence revealed 11 mismatches all of which positioned outside the HVI and HVII regions. The revised Cambridge Reference Sequence (rCRS) became the accepted standard for comparison (Andrews *et al.*, 1999) and is available on GenBank (accession number NC_012920.1). Notably, in order to maintain the historic numbering and thus alignment of the mtDNA sequence, an rCRS deletion (loss of a single C nucleotide) in the position 3107 was substituted with an “N” (Andrews *et al.*, 1999).

1.3 Mitochondrial DNA phylogenetics

The mitochondrial genome analysis has been used in evolutionary biology and molecular anthropology studies to determine the ethnic as well as biogeographical origin of individuals. Strictly maternal inheritance, lack of recombination and high mutation rate make mtDNA a valuable marker for determining the relationship among individuals and groups. A phylogenetic tree of human mtDNA variation has been constructed. Based on commonly shared SNPs mtDNA haplotypes are divided into genealogical groups (haplogroups) that share a common ancestor. The phylogenetic tree is currently comprised of over 5,400 nodes (haplogroups) with their defining variants (Fig. 2) (van Oven & Kayser, 2009). The migration map of women 150,000 years before present (YBP) from Africa onto different continents has been reconstructed using mtDNA analysis (Fig. 3).

Wallace *et al.* using restriction fragment length polymorphism analysis of samples collected from various populations to define haplogroups. L is designated as an African-specific haplogroup by two restriction enzyme sites: *HpaI* site at np 3592 *DdeI* site at np 10394. This lineage is subdivided into two sublineages, L1 and L2. The remaining African mtDNAs form a heterogeneous array of four lineages, designated haplogroup L3, which appear to be the progenitor of half of European, Asian and Native American mtDNA haplogroups. Moreover, four primary African populations studied (the Senegalese of West Africa, the Mbuti Pygmies, the Biaka Pygmies and the Vasikela Kung) showed that each of the populations has a distinctive set of its own specific haplotypes (Wallace, Brown, & Lott, 1999). A different study on the mtDNA CR analysis of Ghana population showed the majority (98.4%) belonged to the three Africa-specific haplogroups L1–L3. Two samples (1%) were defined belonging to U6a haplogroup (Fendt *et al.*, 2012).

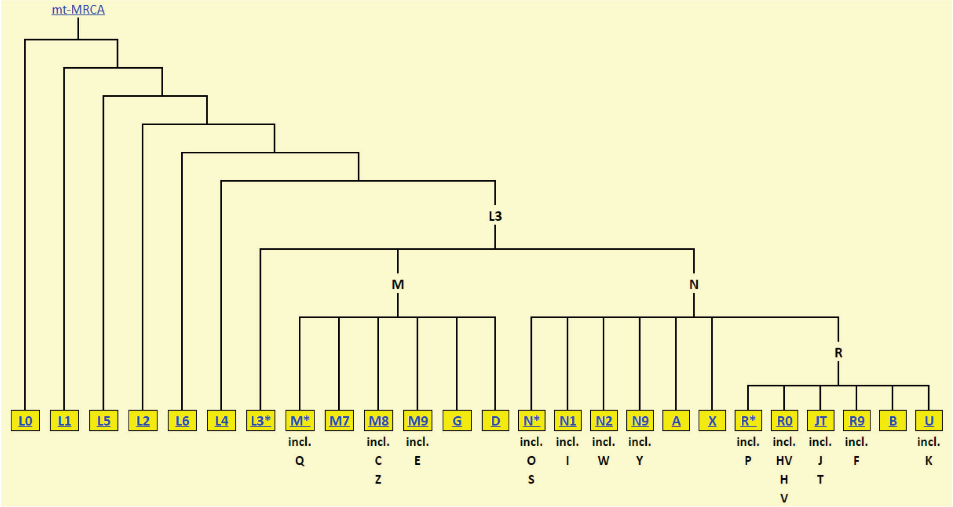


Fig. 2. Human mtDNA phylogenetic tree (Build 17) (van Oven, 2016). Haplogroups are marked with corresponding letters and numbers (e.g. L0, JT, and F). Asterisk (*) defines paragroups – lineages within a haplogroup that are not defined by any additional unique markers. Abbreviations: mt-MRCA – the Mitochondrial Eve, is the matrilineal most recent common ancestor (MRCA) of all living humans.

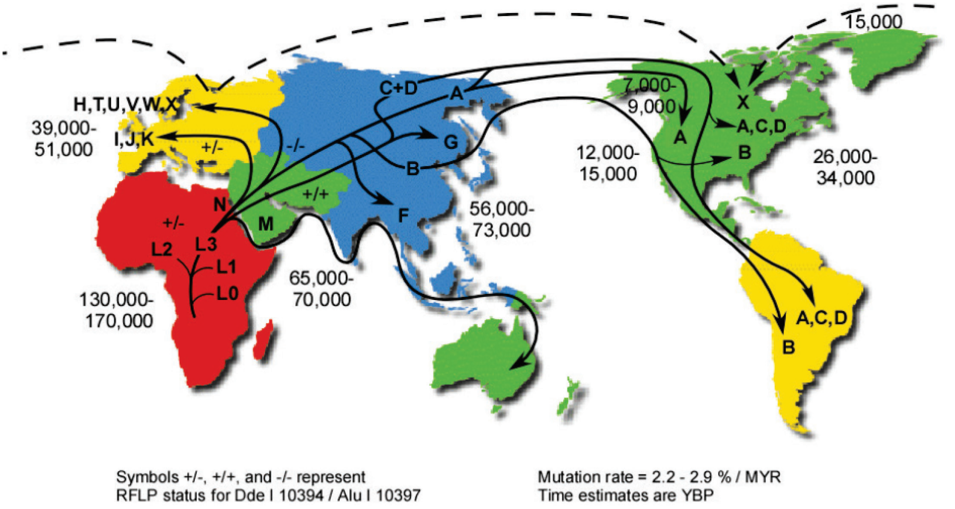


Fig. 3. Human mtDNA migrations map ("Human mtDNA Migrations", 2016). Abbreviations: MYR – million years; YBP – years before present

Of all European mtDNA types 98% fall into nine distinct haplogroups. Six of the haplogroups lack the 10394 site: H, T, U, V, W, and X. Three other major haplogroups retain the *DdeI* site at np 10394: I, J, and K. The most common Europe-specific haplogroup is H (lacking the *DdeI* np 10394 site as well as *AluI* site at np 7025 (Wallace *et al.*, 1999). Haplogroup H has been reported to have a frequency of 40-50% among Europeans and up to 60% in the region of Basque country (Northern Spain) (Loogväli *et al.*, 2004; Richards, Macaulay, Torroni, & Bandelt, 2002). While haplogroup H overall in Europe is rather uniform, haplogroup H subclades show localized frequency peaks, e.g. haplogroup H4 and H6a are centred in Iberian peninsula while haplogroup H5a shows frequency peaks centred around Finland, Balkans and Iberian peninsula (Alvarez-Iglesias *et al.*, 2009). In a Polish population study the most frequent were West-Eurasian haplogroups (H, U, J, T) forming 82.4% of all studied samples, with haplogroup H being the most frequent (43.4%). Asian and African haplogroups were also found but at very low frequencies of less than 1% (Jarczak *et al.*, 2019).

Based on mtDNA HV1 analysis of 117 Estonian population samples, the most common haplogroups belong to the H* haplogroup (17.9%) followed by T* (11.1%). Haplogroup H subclades combined (incl. haplogroup H*) have a frequency of 41.1%. Compared to other Baltic Sea region populations, Estonians were shown to have the highest number of mean pairwise differences (Lappalainen *et al.*, 2008). Loogväli *et al.*, reported a 43.9% frequency of haplogroup H in the Estonian population sample based on HVI region (Loogväli *et al.*, 2004). In a recent study, Saag *et al.*, compared mtDNA haplogroups of 41 tooth samples from Bronze, Iron and Middle Age graves/cemeteries located in Estonia and Ingria (Russia) with over 2,000 present-day Estonian whole mtDNA sequences (unpublished data). All mtDNA haplogroups observed in graves/cemeteries samples (haplogroups H, I, J, K, T, U, W and X) were present in the modern Estonian cohort (Saag *et al.*, 2019). To our knowledge no entire mtGenome data on Estonian individuals are available.

All Asia-specific mtDNA types can be subdivided into two macro-haplogroups defined by the presence or absence of the polymorphic sites *DdeI* at np 10394 and *AluI* site at np 10397. Macro-haplogroup M, comprising both restriction sites, subdivides into haplogroups C, D, G and E. Asia-specific haplogroups A, F and B are defined as (-/-) haplogroups as they lack the aforementioned restriction sites. Haplogroup F is found at high frequency in Southeast Asia, but decreases toward Northeast Asia. On the other hand four haplogroups A, B, C and D [A and B are (-/-), C and D are (+/+)] are found at a low frequencies in Southern Asia, but increase to high frequencies in Northeast Asia, and contributing predominately to the mtDNA pool in Americas (Wallace *et al.*, 1999). In a recent study based on sequenced CR of newly collected 622 mtDNA samples from six different Vietnam locations that represent seven ethnic groups, in addition to all aforementioned Asia-specific haplogroups, haplogroups N and R9’F were reported (Pischedda *et al.*, 2017).

Five haplogroups comprise all of Native American mtDNA types. Paleo-Indians of North, Central, and South America carry haplogroups A, B, C, and D which are found in Asia, and the Paleo-Indians of North Central, North America also carry haplogroup X found only in Europe. The last arrival of Native Americans to the continent is estimated to have happened 7,200–9,000 YBP (Wallace *et al.*, 1999). The study based on freshly collected blood samples from 306 unrelated individuals from Brazil (a highly admixed population) reported prevalence of African macro-haplogroups L (49.02%), followed by the typical Native American haplogroups A, B, C and D (33.33%), with A2 being the most common haplogroup. The remaining 17.65% distributed among other haplogroups including X1, H, U and R0 (Freitas, Fassio, Braganholi, & Chemale, 2019).

1.4 Mitochondrial DNA typing in forensic casework

The high copy number of mtDNA compared to nuclear DNA makes the typing of mtDNA a highly valuable method for human identification in cases where samples are of low nDNA quantity and/or highly degraded, e.g. human remain (bones, teeth) and hair samples. In addition, strictly maternal inheritance and lack of recombination provide lineage information that can be useful in cases with no reference data from first-degree relatives (Wilson, DiZinno, Polanskey, Replogle, & Budowle, 1995). A well-known case is the identification of Tsar Nicholas II's wife Tsarina Alexandra Fyodorovna's remains, as her remains were confirmed through mtDNA haplotype comparison to her distant cousin, Prince Philip. The five children of Romanov family were identified further via mtDNA comparison to Tsarina (Coble, 2011; Coble *et al.*, 2009; Gill *et al.*, 1994). A consistent heteroplasmy detected in the mtDNA of Tsar Nicholas II and his brother Georgij Romanov was used to confirm the remains of the former (Gill *et al.*, 1994; Ivanov *et al.*, 1996).

Traditionally, the gold standard for forensic mtDNA investigation has been to focus on the control region of mtDNA, mainly the HVI and HVII, due to the high variability and the feasibility of these regions to be analysed by the well-established Sanger-type sequencing (STS). Therefore, to date, DNA databases contain limited coding region data. However, it has been shown that mtGenome data provide greater discrimination power and allow resolution of common HVI/HVII haplotypes (Brandstatter, Parsons, & Parson, 2003; Coble *et al.*, 2004; Levin, Cheng, & Reeder, 1999; Parsons & Coble, 2001). Unfortunately, STS chemistry is labour intensive, costly and limited technically to be utilized in whole mtGenome sequencing by application-oriented laboratories.

In addition to human remains, mixtures are one of the most challenging forensic samples. Mixture samples originate from two or more donors and are commonly recovered from touch, sexual and/or physical assault and murder crime scenes. A "complex DNA mixture" may contain more than two donors with one or more of the donors contributing a low amount of DNA, or the contributor(s)' biological material may be somewhat degraded (i.e. missing data). The genotype assignment of STR loci and the following statistical evaluation of the results from mixture samples can be complicated by a number of factors: possible allele overlap among an unknown number of contributors, 3-allele patterns, stochastic fluctuation with low quantity or degraded template, the semi-quantitative activity of Taq polymerase, and microvariants in primer binding sites (Ladd, Lee, Yang, & Bieber, 2001). The variant ratios of mtDNA in the mixtures samples, which look like point heteroplasmies in single-source samples, can be used as an additional tool for mixture interpretation. However, STS has been reported to express peak height ratios that do not correspond with the relative quantity of the contributing donors' sequence variants (Parson *et al.*, 2014) as well as not to be sufficiently quantitative to resolve mixtures based on proportions of variants in mtDNA profiles (Holland, McQuillan, & O'Hanlon, 2011; H. Kim, Erlich, & Calloway, 2015).

Population samples are needed to estimate expected frequencies of mtDNA haplotypes especially in the forensic context. A number of mtDNA databases has been established (Congiu *et al.*, 2012; Kogelnik, Lott, Brown, Navathe, & Wallace, 1996); however issues with the quality of included mtDNA sequences have been reported with mistakes such as base shift, reference bias, phantom mutation, base mis-scoring and artificial recombination (Bandelt, Lahermo, Richards, & Macaulay, 2001; Dennis, 2003; Rohl, Brinkmann, Forster, & Forster, 2001). Haplogroup assignment has been suggested to be used as a quality-control measure for mtDNA profiles (Bandelt *et al.*, 2001; Bandelt, van Oven, & Salas, 2012). Haplogroups are based on haplotypes that have certain "in phase" variants – e.g. variants 2706A and 7028C are present in haplotypes belonging to the haplogroup H, variants 5178A and 16362C are present in haplotypes belonging to the haplogroup D (van Oven & Kayser, 2009). The web-application HaploGrep compares mtDNA profile's variants to the ones defined in Phylotree for haplogroup assignment and provides a quality score for the assigned haplogroups based on their found and missed in phase variants (Kloss-Brandstatter *et al.*, 2011). The discordance with Phylotree defined in phase variants can be an indication of error in mtDNA profile data. However, it can also be a real variant that occurred by back mutation. These data would be suggested to be analysed further. On the other hand, the concurrence with Phylotree confirms mtDNA profile variants. In addition, haplotypes might have variants, known as private mutations, which are not associated with assigned haplogroups. HaploGrep defines polymorphisms as "local private mutation," when it is not associated with the current haplogroup, but has been observed in Phylotree at least once or a "global private mutation" to a polymorphism that is neither associated with the current haplogroup nor has it ever been seen in Phylotree. These private mutations are also included in quality score calculations (Kloss-Brandstatter *et al.*, 2011; Weissensteiner *et al.*, 2016).

In 2006 Parson *et al.*, launched the mtDNA database EMPOP with quality control protocols and measures that set forensic standards as well as provide means that can be used for a quality check of mtDNA sequences. The release 1 of EMPOP comprised data from 5,173 sequences (Parson & Dur, 2007). As of March 2020, EMPOP (v4) release 13 held 48,572 mtDNA sequences from 128 different origins, with 46,963 HVI/HVII sequences, 38,361 entire control region sequences and 4,289 entire mtGenome sequences (https://empop.online/, accessed March 1st 2020). The last revision and extension of guidelines for mtDNA typing by the DNA Commission of the International Society for Forensic Genetics was published in 2014. This revision included a number of recommendations concerning heteroplasmy interpretation and reporting, naming of a haplotype with respect to differences from the rCRS, the subjection of sequences to analytical software tools that facilitate phylogenetic checks for data quality control and others (Parson *et al.*, 2014).

1.5 Massively parallel sequencing for human identification

Massively Parallel Sequencing (MPS), also termed earlier as Next Generation Sequencing (NGS), has been shown to be a feasible alternative to Sanger-type sequencing (STS), particularly by providing substantial increase in throughput (Bodner *et al.*, 2015; McElhoe *et al.*, 2014; Parson *et al.*, 2013b; Seo *et al.*, 2015). While STS DNA chain elongation is terminated by di-deoxynucleotidetriphosphates (ddNTPs) and generated fragments are separated by length (Sanger, Nicklen, & Coulson, 1977), MPS tracks the incorporation of each nucleotide as the DNA chain is elongated (sequencing-by-synthesis, SBS). The three main types of MPS sequencing chemistry are pyrosequencing, sequencing by reversible termination and sequencing by detection of hydrogen ions (pH-mediated sequencing) (Ambardar, Gupta, Trakroo, Lal, & Vakhlu, 2016).

The general workflow of MPS platforms initially is the same as standard forensic DNA typing, i.e. DNA extraction and DNA quantitation. Subsequently MPS employs library preparation in which a DNA (or RNA) sample(s) is randomly fragmented by mechanical shearing or enzymatic treatment in order to reduce fragment size. Alternatively, fragments with suitable size for sequencing can be generated during PCR. Ligation of defined nucleotide sequences (adapters and indexes) follows. Adapters are used to attach the fragment to a surface (beads or flow-cell) for subsequent sequencing, while indexes are molecular barcodes that are used to identify fragments during post-sequencing bioinformatics data analysis. Next, because of the high throughput of MPS, multiple samples are pooled to form a "molecular library" that proceeds into the library amplification step followed by sequencing and data analysis (Shendure & Ji, 2008).

The utilization of molecular indexes in MPS workflow enables the sequencing of large scale multiplex assays that is not feasible by STS. A number of human identification multiplex panels for MPS workflow have been developed. Ganschow *et al.*, proposed a MPS-STR multiplex kit targeting simultaneously 21 forensic markers including the German DNA database STR marker SE33 (Ganschow, Silvery, & Tiemann, 2019). The ForenSeq™ DNA Signature Prep Kit (Verogen, San Diego, CA, USA), evaluated by a number of research groups (Hollard *et al.*, 2019; King *et al.*, 2018; Xavier & Parson, 2017; Xu *et al.*, 2019), enables simultaneous amplification of 27 global autosomal STRs, 24 Y-STRs, 7 X-STRs, 94 identity SNPs, 22 phenotypic SNPs and 56 biographical ancestry SNPs, with the SNP amplicon size range of 63-180 bp (Verogen, 2018). Hollard *et al.*, described and evaluated a protocol for full automation of ForenSeq™ DNA Signature Prep Kit library preparation steps (Hollard *et al.*, 2019).

In regards solely to mtDNA, Parson *et al.*, reported the amplification of whole mtGenomes from hair shaft samples with two multiplex reactions consisting of 31 non-overlapping primers producing amplicons in the size range of 300-500 bp (Parson *et al.*, 2015). The Precision ID mtDNA Whole Genome Panel (Thermo Fisher Scientific, Waltham, MA, USA) composed of 162 amplicons (amplified in two multiplexes) with a length of ~163bp was evaluated on buccal swabs as well as blood samples (Cho, Kim, Lee, & Lee, 2018; Woerner *et al.*, 2018).

Historically, 15% of signal was considered an operational threshold for detecting heteroplasmy (i.e. a minor component) via STS (Duan, Tu, & Lu, 2018). Reliable detection of heteroplasmy positions with minor component <1% has been reported with MPS platforms (Holland *et al.*, 2011; H. Kim *et al.*, 2015). In the recent study González *et al.*, proposed a 3% heteroplasmy frequency with a minimal read depth of 1,000X or 1.5% heteroplasmy frequency with a minimal read depth of 3,000X as a reliable threshold for mtDNA heteroplasmy detection with MPS (Gonzalez, Ramos, Aluja, & Santos, 2020). In recent population study conducted in order to expand the family pedigree and population data necessary for sequence comparisons and statistical analysis and produce mtGenome sequences to be deposited in the EMPOP database Marshall *et al.*, used a minimum read depth of 100X or 350X for 5% or 2% variant detection, respectively (Charla Marshall, 2019).

The ability of MPS to detect the minor component in considerably lower levels compared to STS has led to attempts to interpret mixture samples based on the minor component and phylogenetic assignment (Holland *et al.*, 2011; Li *et al.*, 2010; Lindberg *et al.*, 2016). It has been emphasized that NUMTs can interfere with detection of putative heteroplasmy, especially in case of small amplicon assays. Santibanez-Koref *et al.,* proposed that sequencing data alignment to the human nuclear genome hg19 in combination with rCRS be considered as opposed to solely rCRS. Other parameters in a number of cases regarding amplicon size and sample origin should be considered (Santibanez-Koref *et al.*, 2019). In addition a number of bioinformatics approaches that include high stringency mapping to the rCRS, NUMT-associated variants filtering, alignment to the rCRS and nuclear genome and consensus-based alignment prior to variant calling in order to minimize false variant calls have been proposed and evaluated (Roth, Parson, Strobl, Lagacé, & Short, 2019; Ring, Sturk-Andreaggi, Alyse Peck, & Marshall, 2018; Smart *et al.*, 2019).

2 Aims of the study

In forensic genetics the typing of mtDNA HVI/HVII regions with Sanger-type sequencing (STS) is considered the gold standard. Although the mtGenome has been shown to provide greater discrimination power compared to HVI/HVII, the use of STS technology is costly, labour intensive and time consuming. In addition, STS has proven to be problematic in mixture sample analysis.

The main objective of the present study was the assessment of the validity of massively parallel sequencing technology for mtGenome analysis in forensic genetics.

To this end, the following objectives were set:

- To determine the throughput level of MiSeq platform and Nextera XT DNA Preparation Kit when sequencing mtGenome and evaluate the overall feasibility of the MPS system in generating mtGenome population data by looking at read depth, strand bias, the sequencing results of the poly-cysteine tracts and variants called.

- To compare the discrimination power of mtGenome vs HVI/HVII data though haplogroup and haplotype assignment, random match probabilities and genetic diversities in three main US and an Estonian population samples.

- To compare the haplogroups observed with the Estonian population sample based on mtGenome data with previously reported Estonian mtDNA data that did not span the entire mtGenome.

- To evaluate the feasibility of two-person and three-person mixture interpretation from different and similar phylogenetic origin by using MPS generated mtDNA quantitative and phasing data together with phylogenetic assessment.

3 Materials and methods

The brief description of materials and methods used in the thesis are listed below. The detailed protocols are presented in the corresponding publications.

Sample preparation and target amplification

- Whole blood samples were collected from a total of 283 unrelated individuals from the three US populations (African American, n = 87; Caucasian, n = 83; Southwest Hispanic, n = 113). See publication for sample use approval. (Publication I)
- Buccal swabs were collected from 114 unrelated individuals from the Estonian population. See publication for sample use approval. (Publication II)
- Amplification of the mtGenome was accomplished by long-range PCR in two separate reactions. (Publications I and II)
- Six single-source samples of self-identified Asian and Caucasian individuals were used in the mixture study. See publication for sample use approval. (Publication III)
- Following mixture samples based on nuclear DNA quantity were prepared:
- Two-person mixtures with contributors of different macrohaplogroups in 1:1, 5:1, 10:1, and 20:1 ratios for both individuals.
- Two-person mixtures with contributors belonging to the same macrohaplogroup in 1:1, 5:1, and 1:5 nuclear DNA ratios.
- Three-person mixtures with contributors of different macrohaplogroups in 1:1:1 and 5:1:1 ratios. (Publication III)
- Single-source and mixed samples were amplified with the Precision ID mtDNA Whole Genome Panel (Thermo Fisher Scientific, Waltham, MA, USA). (Publication III)

Library preparation, sequencing and data generation

- Sequencing libraries were prepared using Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA, USA). (Publications I and II)
- The 12 pM pooled libraries were sequenced with MiSeq v2 (2×250 bp) (Illumina). (Publication I and II)
- Sequencing libraries were prepared using the Precision ID Library Kit (Thermo Fisher Scientific) and template preparation was completed on the Ion Chef (Thermo Fisher Scientific). (Publication III)
- Ion 318 Chip v2 (Thermo Fisher Scientific) for the Personal Genome Machine (PGM; Thermo Fisher Scientific) and an Ion 530 Chip (Thermo Fisher Scientific) for the Ion S5 (Thermo Fisher Scientific) sequencing run were used. (Publication III)
- Generated sequences were aligned to the rCRS (Publication I and II) or the rCRS+80 reference genome to account for the Precision ID mtDNA Whole Genome Panel's tiled, overlapping design (Andrews *et al.*, 1999; Thermo Fisher Scientific, 2016). (Publication III)
- Of-board software generated VCF files were converted into haplotypes using MitoSAVE (King, Sajantila, & Budowle, 2014). (Publication I, II and III)

Data analysis

- The following criteria were used for variant calling: a quality threshold of 70; a heteroplasmy threshold of 0.18; and a read depth threshold of 40X (Publication I and II)
- The following criteria were used for variant calling: minimum read depth of 10X and allele ratio of 0.10 as thresholds for generating haplotype calls. (Publication III)
- mtDNA variants were confirmed manually using BAM files and Integrative Genomic Viewer (IGV) software (Thorvaldsdottir, Robinson, & Mesirov, 2013). (Publications I, II and III)
- In order to compare the variant count results indels at poly-C tract positions 309, 315 and 16193 were included in the analysis of the US and Estonian populations. Each Indel position was counted as one variant (e.g. 573.1A, 573.2C, 573.3A, 573.4C as 4 variants) as according to SWGDAM nomenclature guidelines (Scientific Working Group on DNA Analysis Methods, 2019). The data were updated in relation to publication I and II.
- Indels at the positions at poly-C tracts and length heteroplasmy (300-315, 451-463, 515-524, 568-573, 956-965, 5895-5899, 8272-8278, 8281-8289, 12 414-12 425, 16 180-16 193) were not included in haplotype analysis. The data were updated in relation to publication I and II.
- Heteroplasmic positions were used for sample differentiation and unique haplotype count. The data were updated in relation to publication II.
- Random match probability (RMP) and genetic diversity (GD) were calculated according to the methods described by Stoneking *et al.* and Tajima, respectively (Stoneking, Hedgecock, Higuchi, Vigilant, & Erlich, 1991; Tajima, 1989). (Publication I and II)
- Mean pairwise comparison was calculated using MEGA 6 (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013). (Publication I and II)
- HaploGrep 2 (v2.2) software based on Phylotree 17 was used for haplogroup assignment (van Oven & Kayser, 2009; Weissensteiner *et al.*, 2016). (Publication I and II modification)
- For the mixed samples, the ratio of the reference allele and alternate allele compared to the total read depth for each SNP was obtained from mitoSAVE to use as a quantitative assessment of each contributor's proportion of the mixture. (Publication III)
- A phylogenetic check of the final haplotype calls was performed in HaploGrep 2 (v2.1.1) and EMPOP v3 (Parson & Dur, 2007; Weissensteiner *et al.*, 2016; Zimmermann *et al.*, 2011). (Publication III)

4 Results and discussion

4.1 The throughput level of MiSeq platform and Nextera XT DNA Preparation Kit based on read depth, strand bias and variant calls when sequencing mtGenomes (*Publication I*)

In order to evaluate the feasibility of the MPS system to generate mtGenome data for forensic use metrics such as throughput, read depth, strand bias, the sequencing results of the poly-cysteine tracts and variant calling were analysed. These metrics are the quality indicators for a generated data set. In addition, the concordance of MPS data with the data generated via routinely used STS were analysed in order to confirm the variants called by MPS.

The Nextera XT DNA Sample Preparation Kit and MiSeq v2 chemistry were used to prepare and sequence the libraries from 283 US population samples. Libraries were multiplexed as n = 24, 48, 76, 79 and 96 samples. A circus plot with coverage, strand bias data and variants called was constructed (Fig. 4).

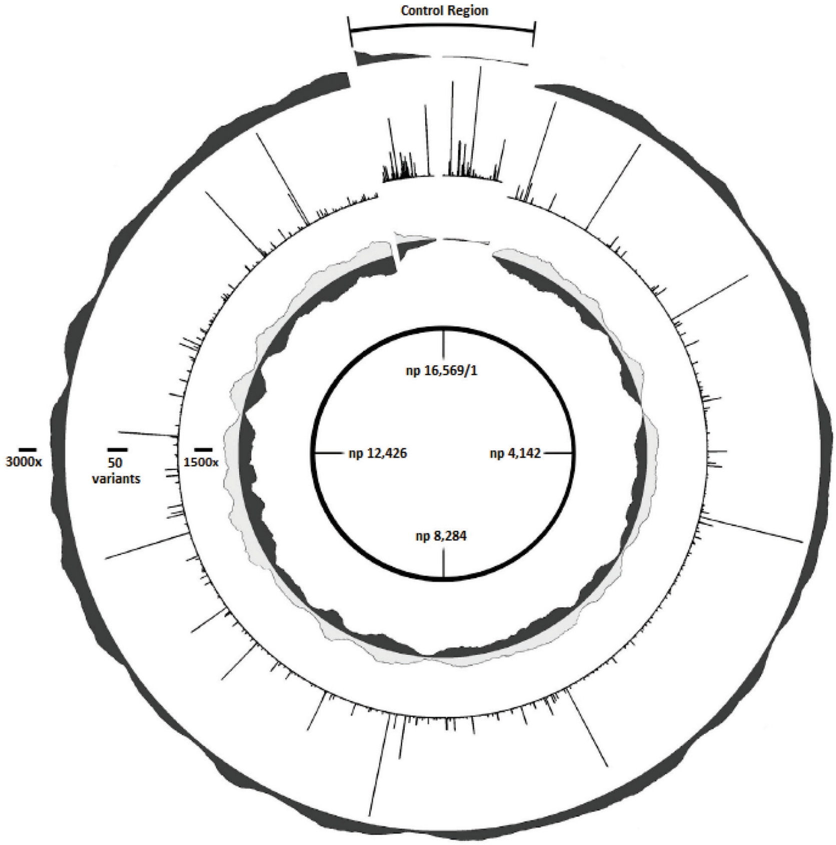


Fig. 4. The Circos plot of mtGenome that shows the mean coverage (outer circle, 24 samples), variant count (second outside circle, 24 samples) and the strand balance (third outside circle, forward and reverse strands indicated with different greys, 24 samples) per nucleotide position (King, LaRue, et al., 2014). The scale bars are shown on the left side of the circles and the nucleotide positions are indicated with the rose diagram (the innermost circle).

One MiSeq v2 chemistry run generated approximately 8.8 Gigabases (Gb) of data. Assuming equal coverage along the whole 16,569 nucleotide long mtGenome, sequencing of the maximum amount of samples (96 samples, according to the available indexes) the coverage at each nucleotide position is expected to be 530,000. However, in practice, generated data expressed a variation in the coverage along mtGenome (Fig. 5).

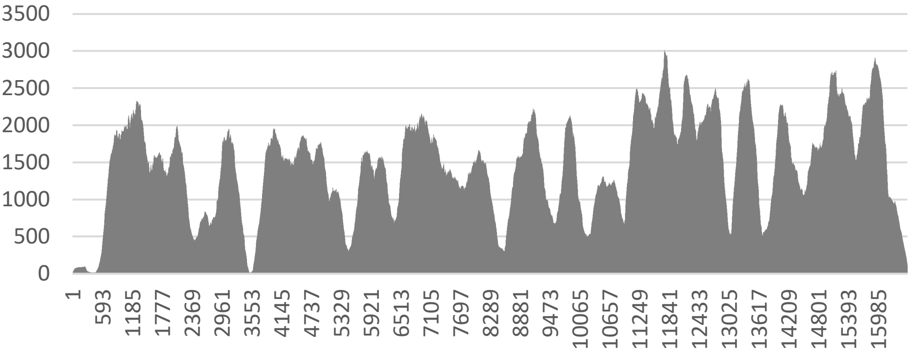


Fig. 5. The mean coverage (24 samples) across mtGenome. Nucleotide positions are on the x-axis, mean read depth on the y-axis.

Regions with the lowest coverage were poly-C tracts in HVII and <300 bp long region within NADH Dehydrogenase subunit I (NDI) gene around the np 3500 (Fig. 5). The low coverage at these positions has been reported by others, including studies with other MPS platforms (McElhoe *et al.*, 2014; Mikkelsen, Frank-Hansen, Hansen, & Morling, 2014; Parson *et al.*, 2013a; Seo *et al.*, 2015). Given that sample amplification was performed with long-range PCR, the target enrichment step can be excluded as a potential source of coverage variance. In order to test if the aligning of a circular genome to a linear reference could lead to fragments covering the positions around np 0 being filtered out, the reference genome covering HVII and 200 bp of the genome from other end was constructed and used for generated data alignment. As a result the 5-fold increase in the coverage of ~50 bases of poly-C tracts was seen, however low coverage was still present. McElhoe *et al.*, attempted a similar rearranged rCRS alignment and similarly found no sufficient coverage increase (McElhoe *et al.*, 2014). GC content correlation with sequencing coverage reported by a number of studies was not confirmed for poly-C tracts in HVII and NDI gene position by McElhoe *et al.*, (McElhoe *et al.*, 2014). Seo *et al.*, compared the effect of enzymatic versus Covaris system shearing of DNA template during library preparation for sequencing on PGM platform. The results indicated the coverage variance coming from processing steps subsequent to fragmentation (Seo *et al.*, 2015). Issue with low coverage areas should be taken into account during future sequencing chemistry, data filtering and alignment developments. Nevertheless, the coverage of the aforementioned regions in the current study was sufficient enough (≥100X) in all samples to be reliably analysed. The lowest coverage nucleotide position(s) drive(s) the throughput of samples that can be analysed. With a coverage threshold of 100X and the throughput of 8.8 Gb, 603 whole mtGenomes would in theory be sequenced assuming equal coverage along mtGenome.

From 227 samples that were indexed and sequenced as n = 24, 48, 76 and 79 multiplexes, 223 (98.2%) expressed sufficient coverage (≥100X), high quality (base quality score ≥Q30 (Phred-style scale) and were fully interpretable. The Q score gives the probability that a given base is called incorrectly by the sequencer. A Q score of 30 (Q30) assigned to a base is equivalent to the probability of an incorrect base call one in 1000 times and means that the base call accuracy (i.e., the probability of a correct base call) is 99.9% (Ewing & Green, 1998).

The four samples with areas expressing coverage <100X had sufficiently lower DNA library concentrations before pooling is a likely explanation of low sequence coverage compared to the rest of the samples in the same library pool. When the pool of 96 samples was sequenced, 26 samples (27.1%) expressed areas with coverage <100x (between 0 and 99). Full results were obtained from 17 of the 26 samples with ≥40x coverage at all variant positions. The variants were confirmed with the re-sequencing of all 26 samples at higher coverage. These data support that an operationally-selected coverage threshold of 40X is sufficient for generating reliable data. The explanation for the higher amount of low coverage samples in the 96 sample run could be the higher library concentration and therefore the reduced number of quality clusters.

A plot of average coverage at each base position from both reverse and forward strands was constructed to see if strand bias occurs (Fig. 6). In theory both strands are sequenced equally (strand balance of 50%). Of all positions, 16062 nps (96.9%) had a strand balance ≥40%. Generally, the same areas that expressed low coverage showed strand imbalance. The areas of low (≤40%) strand balance were the poly-C tract in HVII, positions near nps 16569 and 1 as well as areas around nps 3500 and 8600. The explanation for low strand balance is unknown, and it has been reported with other MPS platforms (Seo *et al.*, 2015). As in the case of read depth, the alignment of a circular genome to a linear reference might be a contributor to strand bias. Regardless, strand bias did not interfere with variant calling in the current study. Although strand bias did not affect the quality of the data, having good depth with both the reverse and forward mapped reads provides additional support for variants called. Therefore, sequencing areas with strand bias should be examined with care.

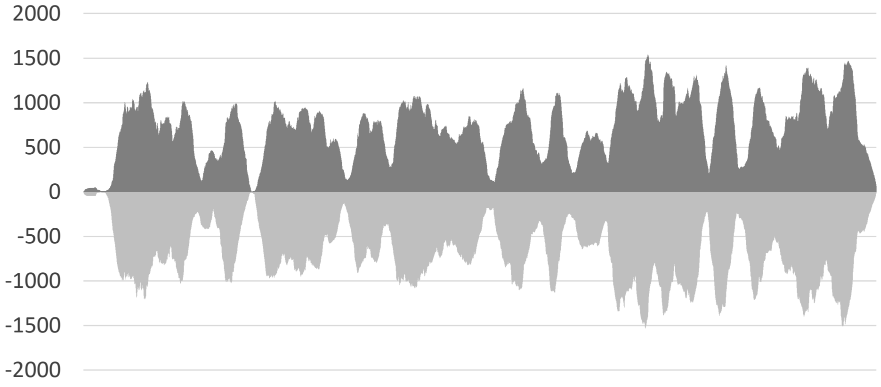


Fig. 6. Strand balance plot, showing the mean depth of coverage of forward (light grey) and reverse (dark grey) strands.

In total 11,656 variants (n = 283 samples) were called in relation to the rCRS, including length heteroplasmy. All MPS data were concordant at all positions with the HVI/HVII STS data. Due to individual sequencing of each molecule (or clonal cluster) versus all amplicons sequenced simultaneously by STS, MPS allowed for length heteroplasmy interpretation. The interpretation of length heteroplasmy in mtDNA sequences produced with MiSeq platform has been reported before (Davis, Peters, Warshauer, King, & Budowle, 2015), however shortcomings of length heteroplasmy interpretation with semiconductor MPS platforms (PGM and S5 systems) have been observed (Parson *et al.*, 2013a; Seo *et al.*, 2015; Strobl *et al.*, 2019). In the recent study by Sturk-Andreaggi *et al.*, STS and two aforementioned MPS chemistry platforms were compared regarding length heteroplasmy detection and interpretation (Sturk-Andreaggi, Parson, Allen, & Marshall, 2020). STS resulted in data with the longest poly-C track length, while the semiconductor sequencing (Ion) platform generated data with the shortest poly-C tracks. The impact on the data inconsistency was shown to come from enrichment methods, sequencing chemistries and analysis software/workflows (Sturk-Andreaggi *et al.*, 2020). As it has been shown that length of homopolymeric sites varies between tissues, currently in human identification homopolymeric sites are not relied upon by Sanger sequencing data interpretation and should not be used with MPS generated data for queries and sample comparison (Scientific Working Group on DNA Analysis Methods, 2019; Sturk-Andreaggi *et al.*, 2020). Substitution stable sights that are within the HVI and HVII C-tracts (e.g. T310C, T16189C) have been shown to be consistent across STS and two MPS platforms (MiSeq and PGM) and analytical methods and thus are reliable to interpret (Sturk-Andreaggi *et al.*, 2020).

In 283 samples, point heteroplasmy was detected in 68 samples (24.0%) at 89 positions, from which 26 (29.2%) resided in HVI/HVII region. Point heteroplasmic positions were concordant with available STS data. Strobl *et al.*, reported possible NUMTs being amplified and sequenced with large multiplex panel for the entire mtGenome. The panel is comprised of 162 amplicons in 2 reactions (Precision ID mtDNA Whole Genome Panel by Thermo Fisher Scientific) and spans the entire mtGenome (Strobl *et al.*, 2019). Amplification of NUMTs can lead to false mixture positions that are falsely called as point heteroplasmy. As long-range PCR was used for mtDNA amplification in the current study, the possibility for called heteroplasmic positions to be NUMTs artifacts is minimal – only two primer pairs are used for whole mtGenome amplification (vs high multiplex panel) what decreases the chance of a primer to anneal to a NUMT region of similar amplicon lengths. In addition to long-range PCR Marquis *et al.*, described a method in which the entire mtGenome is amplified in a single reaction via rolling circle amplification, which also reduces the possibility for NUMTs amplification (Marquis *et al.*, 2017).

Recent validation studies have been performed in order to implement MPS workflow, for partial mtDNA sequences, as well as ancestry- and phenotype-informative SNP analysis, into operational casework (Brandhagen, Just, & Irwin, 2020; Sidstedt *et al.*, 2019). Just *et al.*, reported sequencing 588 forensic quality full mtGenomes with STS (Just *et al.*, 2015), however the throughput capacity of MPS technology for generating high coverage data is considerably more feasible for entire mtGenome sequencing compared to STS even using a highly automated process. Current as well as other studies have shown the greater power of MPS in generating data over Sanger-type sequencing (Avila *et al.*, 2019; Churchill, Novroski, King, Seah, & Budowle, 2017; Parson *et al.*, 2013a; Strobl *et al.*, 2019). Barring point and length heteroplasmy, MPS data concordance with STS data has been reported by other studies including other MPS chemistry platforms (Davis *et al.*, 2015; McElhoe *et al.*, 2014; Mikkelsen *et al.*, 2014; Parson *et al.*, 2013a; Strobl *et al.*, 2019; Sturk-Andreaggi *et al.*, 2020).

After target enrichment, hands-on preparation was conducted within 2 standard working days. Sequencing on MiSeq v2 platform lasted for approximately 39 hours.

Sequencing cost per sample (mtGenome, 16,569 nucleotides, 96 samples per run) was approximately $50 or $0.003 per nucleotide. In comparison, sample preparation and sequencing of HVI/HVII regions on STS platform Genetic Analyzer 3500xl (Thermo Fisher Scientific) requires one working day and the approximate cost of $50 per sample (HVI/HVII, 610 nucleotides, 6 samples per run, good quality samples) or $0.082 per nucleotide.

According to these results, MPS outperformed STS regarding throughput, read depth and for point and length heteroplasmy interpretation. No error regarding variants called was detected. A threshold of 40X was suggested to be sufficient for generation of reliable data, although each laboratory should perform its own studies to establish a threshold. Compared to STS MPS drastically reduces the amount of laboratory work, time needed and costs for mtGenome data generation.

4.2 The comparison of discrimination power of mtGenome and HVI/HVII regions (*Publications I and II*)

In order to determine if whole mtGenome data provide a higher discrimination power and the degree of that increased power compared to routinely analysed HVI/HVII region data, haplogroup and haplotype assignment, random match probabilities and genetic diversities were analysed in an Estonian and three US populations.

Variant count

From the total of 11,656 variants detected at 1,369 nucleotide positions in US samples, 2,949 variants (25.3%) resided in HVI/HVII region. Therefore 8,710 variants that make 74.7% of the total number of detected variants were found outside of HVI/HVII region. In order to visualize the distribution of variants across the mtGenome a plot was constructed (Fig. 7A). As expected regions HVI and HVII that make up 3.7% of mtGenome expressed the highest density of variants.

From 11,656 variants called 676, 217 and 90 variants were detected in one, two and three samples, respectively. Three variants (263G, 4769G and 15326G) were observed in all US samples and are the haplogroup nodes for the rCRS haplogroup. The rest of the variants were observed between four and 282 samples from which 1,252 variants (10.7%) were observed in 20 or less samples. A total of 16 variants, five of these residing in HVI/HVII regions, were detected in half of the samples and comprise 3,793 (32.5%) of total variants called. Removing the high frequency variants (which could be an artifact of relying on the rCRS as a reference) did not change the distribution of the variants across mtGenome.

In the Estonian population sample in total 2903 variants were detected over 529 positions (Fig. 7B). Seven positions (263, 315.1, 750, 1438, 4769, 8860 and 15326) expressed a variation with respect to rCRS in 111 of 114 samples. The detection of these variants in majority of the samples is the reflection of reference used, as these positions (except for 315.1) are haplogroups H2 nodes. The remaining three samples (EST-9, EST-19 and EST-40) expressed up to three variants each in relation to rCRS from which the majority were mutation hotspots or private mutations. Seven variants 73G, 309.1C, 2706G, 7028T, 11719A, 14766T and 16519C were observed in ≥50% of the samples. Variants 73G and 11719A are the haplogroups nodes for the haplogroup R, variants 2706G and 7028T for haplogroup H, and variant 14,766T for haplogroup HV. Variants

309.1C and 16519C are considered mutation hotspots and are not included in haplogroups assignment.

From all variants called 257, 82, 53, 27 and 12 variants were detected in one, two, three, four and five samples, respectively. From the 2,903 variants detected 508 (17.5%) variants were observed in 20 or less samples. From the total amount of variants detected 2,094 (72.1%) resided outside HVI/HVII region. Although a number of differences regarding unique haplotype and haplogroup numbers as well as RMP and GD values between four studied populations will be discussed further, the amount of variants residing outside HVI/HVII region in the Estonian population is concordant with our previous study on 283 US samples (King, LaRue, *et al.*, 2014).

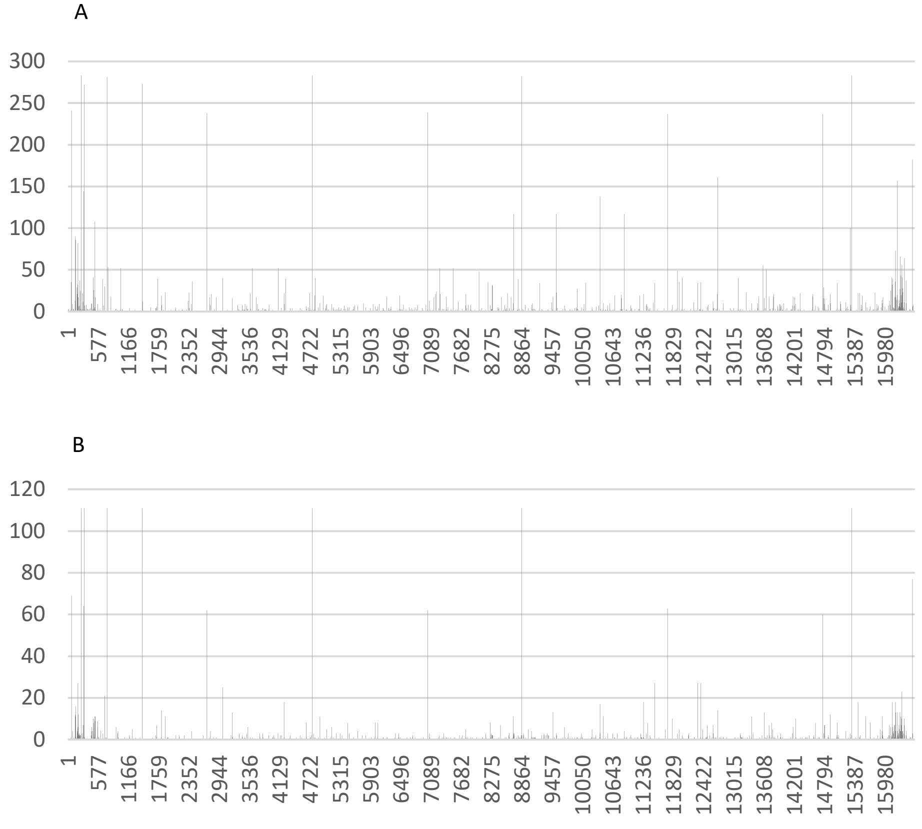


Fig. 7. The plot of variant count per nucleotide position in (A) African-American, Caucasian and Hispanic populations (B) and Estonian population sample analysed in the current study. The number of variants is indicated by the y-axis and the nucleotide position by x-axis.

Haplotypes

A total of 85 (97.7%), 83 (100%) and 111 (98.2%) unique mtGenome haplotypes were observed within African American, Caucasian and Hispanic populations, respectively (Table 1). When mtGenome of the 114 Estonian population samples was analysed, 108 (94.7%) unique haplotypes were observed. Compared to Caucasian samples, Estonian population had a slightly lower diversity (100% versus 94.7%).

Table 1. The number of unique haplotypes and haplogroups seen with HVI/HVII and mtGenome data. The count of haplogroups and haplotypes is presented for each population studied. AFA – African-American; CAU – Caucasian; HIS – Hispanic; EST – Estonian.

	HVI/HVII				mtGenome			
	AFA	CAU	HIS	EST	AFA	CAU	HIS	EST
Number of individuals	87	83	113	114	87	83	113	114
Unique haplogroups	59	72	64	79	76	80	83	87
Unique haplotypes	76	77	96	85	85	83	111	108

Haplotype diversity within and among populations was analysed via pairwise comparison (Table 2). The highest pairwise difference in the current study was determined within the African-American while the lowest was within the Caucasian population. As expected, the Estonian population sample expressed similar however slightly lower pairwise difference within the population compared to the US Caucasian population. These data are concordant with haplotypes diversity studies with HVI/HVII regions by Budowle *et al.* (Budowle *et al.*, 1999).

Table 2. Pairwise difference calculation within and between populations as according to King et al., (King, LaRue, et al., 2014) and Stoljarova et al., (Stoljarova, King, Takahashi, Aaspollu, & Budowle, 2016). Range of difference indicates how many variants between compared haplotypes were observed. AFA – African-American; CAU – Caucasian; HIS – Hispanic; EST – Estonian; SD – standard deviation.

	AFA	CAU	HIS	EST	AFA/ CAU	AFA/ HIS	CAU/ HIS	AFA/ CAU/ HIS
Number of pairwise differences (Mean ± SD)	55±22	30±11	36±15	27±11	47±20	47±19	35±12	43±18
Range of differences	0-104	1-55	0-90	0-54	0-102	0-102	0-91	0-101

In the current study HVI/HVII data showed fewer unique haplotypes compared to mtGenome data – 76 (87.4%), 77 (92.8%), 96 (85.0%) within US African American, Caucasian and Hispanic populations respectively (Table A). In the Estonian population 85 (74.6%) unique haplotypes were observed in HVI/HVII region. According to these results the increase in unique haplotypes was the highest in the Estonian population (27.1%), followed by the Hispanic population (15.6%), African American (11.8%) and Caucasian populations (7.8%).

In the Estonian population 47 individuals had one or more identical mtDNA sequence when the HVI/HVII data were solely analysed. The 27.1% increase in unique haplotypes in the Estonian sample can be attributed to 23 samples that formed identical sample groups of six (H1a group), two groups of four (H2a2a and U4 group) and three groups of three (N1a1a1a, T2b and U8a1a group) with HVI/HVII data. The rest, 24 non-unique samples formed 12 pairs. From these ≥3 identical sample groups (groups H1a, H2a2a, N1a1a1a and U8a1a) were fully differentiated with mtGenome data. The other two groups (group U4 and T2b) were each left with two identical haplotype samples.

When the whole mtGenomes of these 47 undifferentiated Estonian population samples were analysed, a total of 56 variants differentiated an additional 35 samples. Of these 56 variants five (8.9%) were observed in the control region and 51 (91.1%) in the coding region. The entire control region data increased the number of unique haplotypes from 85 to 89, a 4.7% increase (mutational hotspot 16519C for the sample EST-5 included), compared to the 27.1% with whole mtGenome data. Interestingly, the amount of coding region variants (including five heteroplasmic positions) that were observed in the 47 samples undifferentiated by HVI/HVII data were two times higher (1.4 variants per sample vs 0.7) for haplogroup H samples compared to others that included samples from haplogroups M, N, T, U and W.

The addition of coding region data benefits differentiation of samples with haplogroups that have low amount of variants within HVI/HVII region, e.g. haplogroup H. Inclusion of mtDNA CR to the analysis of Caucasian samples also was proposed by Coble *et al.* (Coble *et al.*, 2004). In addition to haplogroup H, the importance of mtDNA coding region variants for sample discrimination has been shown for Native American haplogroup B2 which lacks variants in the CR to further resolve haplogroup B4 (Wood *et al.*, 2019). The increase in discrimination power afforded by analysing the coding region has been shown for haplogroups in Han Chinese, Brazilian and US population samples (Avila *et al.*, 2019; Just *et al.*, 2015; Yao, Xu, & Wan, 2019).

In 283 US samples, heteroplasmy was detected in 68 samples (24.0%) at 89 positions. One, two, three and four heteroplasmic positions were detected in 52, 12, three and one samples, respectively. Four heteroplasmic positions (204Y, 1832R, 2650Y, 16129R) were detected in the sample USA_TX_0146 with haplogroup T1a1 and the quality score of 97.9%. The highest amount of samples with heteroplasmy was detected in the Caucasian population where 26.5% of samples expressed positions with heteroplasmy. Hispanic and African-American population had 23.9% and 21.8% of samples with heteroplasmic positions, respectively. In the Estonian population, heteroplasmy was detected in 13 (11.4%) samples at 16 positions. The maximum number of heteroplasmic positions per sample in Estonian population was 2 (samples EST-33, EST-81 and EST-106).

Of the total of 89 heteroplasmic positions in the US population samples 63 (70.8%) resided outside HVI/HVII region and 58 (65.2%) outside the CR (np 16024-576). In one case heteroplasmy in the CR enabled Hispanic population sample differentiation: samples USA_TX_0213 and USA_TX_0214 were differentiated by the 15184Y heteroplasmic position.

All heteroplasmic positions detected in the Estonian population were unique. From these 16 positions 12 (75.0%) resided outside the CR. Heteroplasmy position 9117Y enabled differentiation of samples EST-86 and EST-55.

A number of studies on heteroplasmy frequencies in mtDNA CR have been published, however the reported values vary greatly with Naue *et al.*, reporting heteroplasmy frequencies of 18.0% and 16.2% for blood and buccal swab samples, respectively (Naue *et al.*, 2015) and Irwin *et al.*, reporting 1.0%-9.5% for blood samples and 4.3%-15.5% heteroplasmy frequency for buccal swab samples of different populations (Irwin *et al.*, 2009). If solely the CR is considered, the heteroplasmy frequencies observed in the current study are more similar in value reported by the latter study. However our study shows that the majority of heteroplasmic positions are located outside the CR. In the recent study based on almost 1800 samples, Marshall *et al.*, reported that coding region heteroplasmic positions are more discriminating that those in the CR as the frequency of latter is higher (Marshall, 2019).

The high intra-individual variance of heteroplasmy frequency has been reported with common forensic samples such as hair (Barrett *et al.*, 2020; Bendall, Macaulay, & Sykes, 1997). Unlike many other tissues (e.g. blood) hairs develop from a small number of cells. By week 16–20 of gestation, the hair follicles are formed each from a small group of stem cells that have gone through a somatic bottleneck. Hair growth requires a high number of mitochondria that will be clonally produced. Developing hair follicles give rise to the hair root and shaft via rapid mitotic cell division, with each division leading to a stochastic segregation of mtDNA and thus additional genetic drift (Barrett *et al.*, 2020; Linch, Whiting, & Holland, 2001; Paneto *et al.*, 2007). Barrett *et al.*, observed that variance of heteroplasmy frequency in hairs increase with the age of individuals, that minor and major allele frequency can shift between different hairs from the same individuals and that minor allele frequency in hair differ from those in blood and buccal samples (Barrett *et al.*, 2020). Moreover, Naue *et al.*, reported heteroplasmy intra-individual variance with 90% of individuals not expressing heteroplasmy in all 8 tissues sampled (buccal cells, blood, hair, bone, skeletal muscle, heart muscle, brain, lung, and liver) (Naue *et al.*, 2015). Therefore, using heteroplasmy-based evidence in human identification should be done with caution especially when comparing data from different tissue sources.

Random match probability (RMP) and genetic diversity (GD), as common parameters used in forensic genetics, were compared between HVI/HVII and mtGenome data (Table 3). From the US populations the Hispanic population showed the largest difference between the HVI/HVII and mtGenome data results (72.2% decrease in RMP), followed by the Caucasian population (61.4%). The African American population had a RMP decreased of 47.6% when mtGenome data were used compared to HVI/HVII data. The RMP and GD for the Estonian population was 4.52% and 96.32% for HVI/HVII data and 1.15% and 99.72% for the mtGenome data, respectively, which is a 74.5% decrease in RMP, the largest decrease in RMP within the four populations studied. The differences between HVI/HVII and mtGenome RMP and GD values were statistically significant – p = 0.01659 for RMP and p = 0.01645 for GD (Student T-test, paired, two-tailed).

Table 3. Random match probability (RMP) and genetic diversity (GD) calculated for the studied populations based on HVI/HVII and mtGenome data. AFA – African-American; CAU – Caucasian; HIS – Hispanic; EST – Estonian; n – number of samples; SD – standard deviation.

		HVI/HVII		mtGenome	
Populations	n	RMP	GD	RMP	GD
AFA	87	2.50%	98.64%	1.31%	99.84%
CAU	83	3.12%	98.06%	1.20%	100.00%
HIS	113	3.52%	97.35%	0.98%	99.91%
EST	114	4.52%	96.32%	1.15%	99.72%
Mean ± SD		3.41 ± 0.85%	97.59 ± 1.00%	1.16 ± 0.14%	99.87 ± 0.12%

RMP reported by Budowle *et al.*, for HVI/HVII regions of US populations was lower than found for the three US populations in the current study (e.g. 0.9% vs 2.4% for African-American population) (Budowle *et al.*, 1999). A notably lower RMP for US Caucasians in Budowle *et al.*, study is likely due to a larger sample size as the RMP can be impacted by sample size. RMP values for HVI/HVII as well as mtGenome reported by Just *et al.*, are also lower compared to the current study (Just *et al.*, 2015). As with values from Budowle *et al.* study, larger population sizes might be the explanation for variation in the reported RMPs. However the possibility of different degrees of variation between different samplings of the same population from the current study cannot be excluded.

Haplogroups

HaploGrep (Kloss-Brandstatter *et al.*, 2011; Weissensteiner *et al.*, 2016) and Phylotree (van Oven & Kayser, 2009) were used to assign haplogroups to sequenced haplotypes based on mtGenome data, coding region and HVI/HVII data. Based on mtGenome data the US samples were assigned to 14 macrohaplogroups with Hispanics being the population with the highest number of different macrohaplogroups (n = 12), followed by Caucasians (n = 10) and African Americans (n = 6) populations. In the Estonian population data 12 macrohaplogroups were observed, of which 3 macrohaplogroups X, N and R were not observed in the US populations. Macrohaplogroups A, B, C and L, frequent in the American and African populations, were not observed in the Estonian population sample (Table 4).

Table 4. Macrohaplogroups and number observed in the data set of 4 populations studied. AFA – African-American; CAU – Caucasian; HIS – Hispanic; EST – Estonian.

Macrohaplogoups	AFA	CAU	HIS	EST
A	1	1	37	-
B	1	1	18	-
C	-	1	21	-
D	-	-	5	1
H	1	32	12	54
I	-	3	1	2
J	-	11	1	7
K	-	6	1	3
L	81	-	7	-
M	2	-	-	2
N	-	-	-	3
R	-	-	-	1
T	-	7	3	11
U	1	20	6	24
V	-	1	-	-
W	-	-	1	5
X	-	-	-	1
Total samples	87	83	113	114
Total Macrohaplogroups	6	10	12	12

Seven of the 283 US population samples (2.5%) changed haplogroup clades when expanding the HVI/HVII data to those of mtGenome data (Table 5). Indeed, five of the seven samples changed macrohaplogroups. For example sample USA_TX_0057 was changed from Asia-specific mtDNA haplogroup D4j1b2 assigned based on HVI/HVII data to Africa-specific L3b1a7a haplogroup. Top rank haplogroups assigned for these samples by HaploGrep varied widely. Based on mtGenome data HaploGrep ranks haplogroups H32, H+152 and H as the top three for the sample USA_TX_0257, while based on HVI/HVII data the top three were ranked as haplogroups P5, U5b2a1a and H32 or R30a1 (latter two have the same overall quality score). Interestingly, less samples went through a haplogroup change when haplogroup assignment was performed using Phylotree build 17 (in the current study) compared to Phylotree build 15 used in publication I (King, LaRue, *et al.*, 2014).

Macrohaplogroup change in the Estonian population sample was less drastic. The biggest change was seen with the sample EST-59 that had a macrohaplogroup change from U with HVI/HVII data to H with mtGenome data. Four other samples showed a change between closely related haplogroup clades (Table 5).

The lowest pairwise difference as well as the highest increase in the number of unique haplotypes based on mtGenome and HVI/HVII data within the Estonian population show that the addition of data from regions outside of HVI/HVII benefits the most for the least diverse populations. As the Estonian population expressed the lowest increase in unique haplogroups compared to three US populations studied (Table 1), the increase in the unique haplotypes was due to the prevalence of private mutations outside the HVI/HVII regions. These variants are of importance when genetically close mtDNA samples are possibly able to be distinguished.

Bias regarding rCRS and Phylotree described previously (Bandelt & Salas, 2012; Behar *et al.*, 2012) could be seen with the Estonian sample EST-19. Sample EST-19 had a total of two variants (14598C, 16294T) in relation to rCRS and was given a quality score of 53% by HaploGrep. Therefore the true haplogroup for the EST-19 sample (defined with H5e haplogroup and the quality score of 53% by HaploGrep) is likely to be in the H2 lineage.

Table 5. Samples that changed haplogroup clade assignment between HVI/HVII data and mtGenome data.

	HVI/HVII		mtGenome	
	Assigned haplogroup	Quality	Assigned haplogroup	Quality
USA_TX_0052	M73'79	95.2%	L3b1a+@16124	93.5%
USA_TX_0057	D4j1b2	90.6%	L3b1a7a	98.8%
USA_TX_0063	N2	95.1%	L3e1f	95.0%
USA_TX_0108	HV0	94.0%	V2	95.3%
USA_TX_0132	R+16189	88.0%	H4a1a1a1a1	97.9%
USA_TX_0250	HV21	72.3%	H1ag	85.0%
USA_TX_0257	P5	96.3%	H32	97.0%
EST-111	H2a2a+(16235)	75.8%	HV	81.4%
EST-12	H1e1a4	100.0%	HV16	96.1%
EST-13	JT	88.8%	T2b4+152	97.4%
EST-36	JT	88.8%	T2b4+152	97.4%
EST-59	U5b2a1a2	86.7%	H	81.1%

According to these results over 70% of variants reside outside routinely analysed HVI and HVII regions. Entire mtGenome data show a significantly higher discrimination power compared to HVI/HVII data. A number of samples changed their haplogroup when haplogroup assignment was done based on entire mtGenome haplotypes compared to HVI/HVII. Therefore, based on these data, typing of mtGenome is preferred over HVI/HVII.

4.3 Genetic description of an Estonian population sample (*Publication II*)

As the whole mtGenome data for the Estonian population was published for the first time, the mtGenome data of 114 Estonians sequenced in the current work were compared to previously available data on Estonian mtDNA as well as descried in relation to haplogroups defined in other populations.

Twelve major haplogroup clades (van Oven & Kayser, 2009) and 87 unique haplogroups were observed in the Estonian population sample (Tables 1 and 4). The majority of samples (68.4%) belonged to clades H (47.4%) and U (21.1%). Clades T and J comprised 9.6% and 6.1% of samples, respectively. The rest of the haplogroup clades included less than 5.0% of the samples. Haplogroup clades D, R and X had one sample each with the quality score of 98.2%, 94.1% and 98.0%, respectively. Interestingly the Asia-specific M haplogroup was observed in the dataset.

In the recent study by Saag *et al.*, described mtDNA haplogroup data produced by shotgun sequencing from 41 tooth samples from Bronze, Iron and Middle Age graves/cemeteries located in Estonia and Ingria (Russia) (Saag *et al.*, 2019). All major haplogroup clades found in the Saag *et al.*, study were present in the haplogroup pool of the current Estonian population study. The major clades of the current Estonian population study haplogroups that were not observed in the Saag *et al.*, study were D4e4b, M10a2, N1a1a1a1, R1b1 and X2c1.

Haplogroup H is known to be the most frequent (44.5-48.2%) haplogroup in Europe with high frequency peaks also in the Near East, and northern Caucasus (Richards *et al.*, 2000). The frequency of haplogroup H has been reported to be 44.6% based on 17 coding region SNP data and 43.9% based on HVI data in the Baltic Sea region and Estonian populations, respectively (Achilli *et al.*, 2004; Lappalainen *et al.*, 2008; Richards *et al.*, 2000). These results are is consistent with our findings based on the entire mtGenome data. Haplogroup H divides into a number of subhaplogroups with considerable variation in subhaplogroup variations in Europe. Loogväli *et al.*, reported 57 basic branches stemming from the major haplogroup H based on the data from 267 coding region sequences (Loogväli *et al.*, 2004). In our Estonian population sample subhaplogroup H1 (15.8%, including all H1 subhaplogroups) showed a considerably higher frequency compared to the rest of the subhaplogroups that were under 6.1%. Haplogroup H1 characterized with large frequency peaks (up to 27.7%) in Western Europe is centered in Iberia and has been shown to decline toward the northeast and southeast. However frequency peaks in Austria and Estonia (16.7%), similar to our results, have been shown (Achilli *et al.*, 2004). A slightly lower H1 haplogroup frequency (12.2%) can be seen with the data from Saag *et al.*, from ancient DNA samples from grave/cemetery sites in Estonia and Ingria (Saag *et al.*, 2019).

European specific subhaplogroup U5, with a frequency in the European population up to 10.3% and proposed as the main haplogroup of Europe's first settlements (Richards *et al.*, 2000), was the second most abundant subhaplogroup in the current study with the frequency of 11.4%. Similar frequency of 12.9% for Estonian and slightly higher 15.0% for Baltic Sea region populations in general have been reported (Lappalainen *et al.*, 2008). Saag *et al.*, data show a 12.2% frequency for the U5 haplogroup in the ancient DNA samples (Saag *et al.*, 2019).

Similarly to the current study, haplogroup X has been observed by Lappalainen *et al.*, in one out of 117 Estonian as well as one out of 307 Swedish population samples (Lappalainen *et al.*, 2008). Haplogroups X2e and X2c were observed by Hedman *et al.*, in one and four Finnish population samples (n = 200), respectively (Hedman *et al.*, 2007).

Haplogroup X has a wide geographic range but low frequency (<5%) in West Eurasian and North African population. Two subhaplogroups X2b and X2c cover one-third of the European X haplogroup sequences (Reidla *et al.*, 2003).

The frequency of haplogroup D is the highest in eastern (10-43%), northern (11-34%) and central (14-20%) Asian groups, declining towards south and west (M. Derenko *et al.*, 2010). Haplogroup D4e has been reported in a number of Asian populations in EMPOP v4 release 13 database (https://empop.online/hg_tree_browser, accessed December 24th 2019). One sample in the current study was reported with a D4e4b haplogroup (quality score of 98%). Haplogroup D5a has been reported at low frequencies in Estonian samples (Loogväli *et al.*, 2004). To our knowledge, the rare haplogroup D4e4b reported in Tatars and Russians (M. Derenko *et al.*, 2010) has not been observed in the Estonian population.

The major haplogroups M and N derive from the African haplogroup L3 and gave rise to western Eurasian and Asia-specific haplogroups. While haplogroup N is a pre-haplogroup for Europe specific haplogroups (with haplogroup R as the main branch) and three Asia-specific haplogroups (A, B and F), major haplogroup M is a precursor for Asian-specific haplogroups (Wallace, 2015). Haplogroup N (more specifically N1) and R have been shown to occur in low frequencies in European populations including Finns (Hedman *et al.*, 2007; Kushniarevich *et al.*, 2015; Richards *et al.*, 2000). Pliss *et al.*, analysed mtDNA HVI region of 409 Estonian population samples and observed haplogroups N1a and N1b with the frequencies of 1.2 and 0.2, respectively. Haplogroups M* (with subhaplogroup M10) and R were not observed that population sample (Pliss *et al.*, 2006). Haplogroup M10 has been observed with a low frequency in Volot (north-western part of European Russia) population (Grzybowski *et al.*, 2007). Haplogroup M10a2 has been observed in the Kazakhstan (Westeurasian) population and the South-Korean population (Irwin *et al.*, 2010; Lee *et al.*, 2006). All three haplogroups (N, M and R) have been observed in the South Siberian populations (M. V. Derenko *et al.*, 2003).

According to these results, 10 of the 12 major haplogroup clades observed in the Estonian population sample based on entire mtGenome data have been observed in previous Estonian population studies based on partial, mostly HVI, mtDNA sequences. Two haplogroups – M and R – have been reported in geographically close populations.

4.4 Mixture sample mtDNA analysis with MPS (*Publication III*)

Mixtures are one of the most challenging samples in forensic casework especially in cases with low levels of DNA. Quantitative as well as qualitative data provided by MPS can be utilized to distinguish individuals contributing to mtDNA mixtures.

Six single-source samples, seven two-person mixture samples in the ratio of 1:1, 1:5, 1:10 and 1:20 with distinct major haplogroups (HV and F1a1a), three two-person mixture samples in the ratio of 1:1, 1:5 and 5:1 with similar haplogroups (subclades of U2e) and two three-person mixture samples in the ratio of 1:1:1 and 5:1:1 with distinct major haplogroups (HV, F1a1a and U2e2a1a) were sequenced and analysed using quantitative, phasing and phylogenetic data.

Control samples

Negative (no-template control, NTC) and positive control samples were added to each of the sequencing runs. NTC samples resulted in 0X to 179X and 0X to 51X read depth per nucleotide position for Ion S5 and PGM sequencing runs, respectively. The ratio of the average read depth of NTC samples and the average read depth of the single-source samples across mtDNA was calculated. The ratios were between 0.04 and 2.55% which is lower than established variant call threshold of 10% for this study.

The positive control NIST standard 9947A was used. Sequence results of the positive control were in concordance with NIST data (Riman, Kiesler, Borsuk, & Vallone, 2017), except for position 1393G/A with the minor allele not meeting the threshold of 10%. The minor allele was seen at 3% and 4% of the nucleotide position total read depth in PGM and Ion S5 runs, respectively. Sequencing chemistry as well as the variation in lots of the cell line might be the cause for reduced frequency of the minor 1393G/A variant. The minor allele of the heteroplasmic position 7861Y was seen at 17% and 18% in Ion S5 and PGM runs, respectively.

Single source samples

The average read depth of six single source samples and two positive controls was 270X to 18,836X and 366X to 24,224X for the PGM and Ion S5 runs, respectively. Slight differences in average read depth are expected between Ion Chips used with the sequencing platforms. The relative locus performance (RLP), calculated by dividing the read depth of the nucleotide position with the average read depth of the sample and multiplying by the length of the mtGenome, was used to normalize the two sequencing runs and visualize the relative sequencing performance across the mtGenome. Average RLP for the single-source samples was 5.95E-05 to 3.50E-04. The strand balance for the positive strand was 0.02 to 0.75 with 84% of the nucleotide positions at or above 0.40.

Noise possibly originating from PCR errors, sequencing errors or NUMTs was analysed. In general, average noise for the single-source samples ranged from 0.00% to 4.86% of the total read depth across the mtGenome with 8 noise positions above 3%. These 8 positions were associated with homopolymeric regions. Shortcomings with homopolymeric tract sequencing on Ion Torrent platforms have been reported by a number of studies (Bragg, Stone, Butler, Hugenholtz, & Tyson, 2013; Chaitanya *et al.*, 2015; Churchill, King, Chakraborty, & Budowle, 2016; Parson *et al.*, 2013a; Seo *et al.*, 2015). In a recent evaluation study of the Precision ID mtDNA Whole Genome Panel Stoble *et al.*, reported sequence artefacts present consistently across samples (Strobl *et al.*, 2019). Bioinformatics improvements could lead to better noise and off-target alignment filtering and thus allowing for a reduction of heteroplasmy and mixture thresholds.

One single-source sample (sample 011) showed a point heteroplasmic position that varied slightly depending on the sequencing platform – 14386Y, minor variant at 32% and 23% in Ion S5 and PGM runs, respectively. No heteroplasmic positions were detected in the rest of the single-source samples.

The mtDNA profiles from the single-source samples used in the current study were compared and found concordant to the same samples sequenced via long-range PCR on the MiSeq and PGM platforms in earlier studies (Churchill *et al.*, 2016; King, LaRue, *et al.*, 2014).

Mixtures

The average read depth and the RLP for the mixtures was 401X to 17,466X and 7.86E-06 to 3.47E-04, respectively. The strand balance for the positive strand was 0.02 to 0.71 with 82% of the nucleotide positions at or above 0.40. The average noise for the mixture samples ranged from 0.00% to 4.54% of the read depth across the mtGenome with seven nucleotide positions above 3%. These positions were associated with homopolymeric regions as was observed with single-source samples. The performance metrics of the mixture samples were similar to the single source samples results.

The quantitative data per variant position were obtained from mitoSAVE (King, Sajantila, *et al.*, 2014). The variants were divided into one of the three categories: a) variant present in both of the contributors' haplotypes; b) variant present in the major contributor's haplotype; c) variant present in the minor contributor's haplotype. No allele ratios were set for the categories before the variant assignment as no prior data on the ratios were available. Variants that were not able to be assigned to one of the aforementioned categories were scrutinized further regarding phasing and phylogenetic data. Phasing data were collected during manual variant verification with IGV (Thorvaldsdottir *et al.*, 2013). Phylogenetic data were collected through HaploGrep (Weissensteiner *et al.*, 2016) and EMPOP (Parson & Dur, 2007).

Two-person mixtures

Two-person mixtures with distinct haplotypes (HV and F1a1a) were analysed. The complete major contributor's mtDNA profile was detected for all two-person mixtures. Complete mtDNA profiles of both (major and minor) contributors were detected from the 1:1 and 1:5 mixtures. It is important to note that the 1:1 mixture can have a major and a minor contributor as the DNA amount in the sample in this study is based on the nuclear DNA amount. The amount on mtDNA is related to the total DNA amount however it varies between individuals and thus will differ to some degree (Robin & Wong, 1988; Shay, Pierce, & Werbin, 1990). Also manipulations during sample preparation can induce variation.

In the 1:10 mixture all variants were observed except for one variant of the minor contributor. Manual checking of the BAM files in IGV showed the missed variant 4092A at 8.0%. A partial minor contributor's profile was also detected from the 5:1 mixture with 19 positions out of 42 observed (45.2%, excluding minor contributor's point heteroplasmic positions). Therefore only major contributor's complete profile was detected from the 10:1, 1:20 and 20:1 mixtures.

The quantitative results (alternative allele to total allele count ratio) for two-person mixtures are shown in the table 6. The table includes only the variants that could be assigned to one of the three categories – both contributors, major contributor, minor contributor.

Table 6. Quantitative data results of two-person mixtures with contributors from distinct haplogroups. The average ratio of alternative allele read depth to total read depth is shown with standard deviation in parentheses. Only the major contributor profile was generated from analysis of 10:1, 1:20 and 20:1 mixtures and therefore no values for "minor" and "both contributors" are given for these samples.

	1:1 mixture	1:5 mixture	5:1 mixture	1:10 mixture	10:1 mixture	1:20 mixture	20:1 mixture
Both contributors	99.50% (0.76%)	99.38% (0.92%)	99.63% (0.74%)	98.44% (3.24%)			
Major contributor	62.38% (2.00%)	73.23% (5.14%)	89.44% (1.81%)	85.73% (1.82%)	96.94% (2.99%)	94.00% (3.24%)	98.12% (1.58%)
Minor contributor	35.96% (3.04%)	24.00% (2.00%)	11.27% (0.90%)	12.86% (1.21%)			

Phasing information was used in the 1:1 mixture where the variants 4086T and 4092A were difficult to assign quantitatively based on frequencies of 47.7% and 48.0%, respectively (Fig. 8). These variants were not in-phase and therefore were parsed (Fig. 9). A phylogenetic check was performed via HaploGrep or EMPOP to confirm blind phasing assignment (Parson & Dur, 2007; Weissensteiner *et al.*, 2016). HaploGrep assigned position 4086T to the minor contributor and therefore according to phasing result variant 4092A was assigned to the major contributor. According to HaploGrep variant 4092A is a local private mutation and therefore without phasing data it would not be possible to assign the variant to one specific contributor. According to the single-source samples variant 4092A belongs to the minor contributor of the 1:1 mixture.

The quantitative data of positions 249del, 16162G and 16172C in the 1:1 and 1:5 mixtures could not be assigned to one of three categories with confidence based on a quantitative assessment. However, the positions were assigned to the correct contributor via phylogenetic assignment (Fig. 10).

Point heteroplasmic position 14386Y from sample 011 was observed in the 1:5 (alternative allele at 18.3%), 1:10 (19.0%) and 1:20 (20.5%) mixtures, where sample 011 was the major contributor. In the 1:20 mixture point heteroplasmy (PHP) (14386Y) stands out from the quantitative data of the rest of the positions (Fig. 11). As only the major contributor's profile was obtained from the 1:20 mixture, position 14386Y can be assigned as a heteroplasmic position of the major contributor. However, in case of the 1:5 and 1:10 mixtures, the PHP amount is close to the minor contributor's average quantitative data results and could be falsely assigned to the minor contributor. Phasing nor phylogenetic assignment could help with determining the contributor of the position – no other variant was found in the nearby position for phasing analysis nor was it a definitive haplogroup variant. Two profiles per major and minor contributor were generated for the 1:5 and 1:10 mixtures. Therefore, difficulties with detecting and correct assignment of point heteroplasmic positions can be expected.

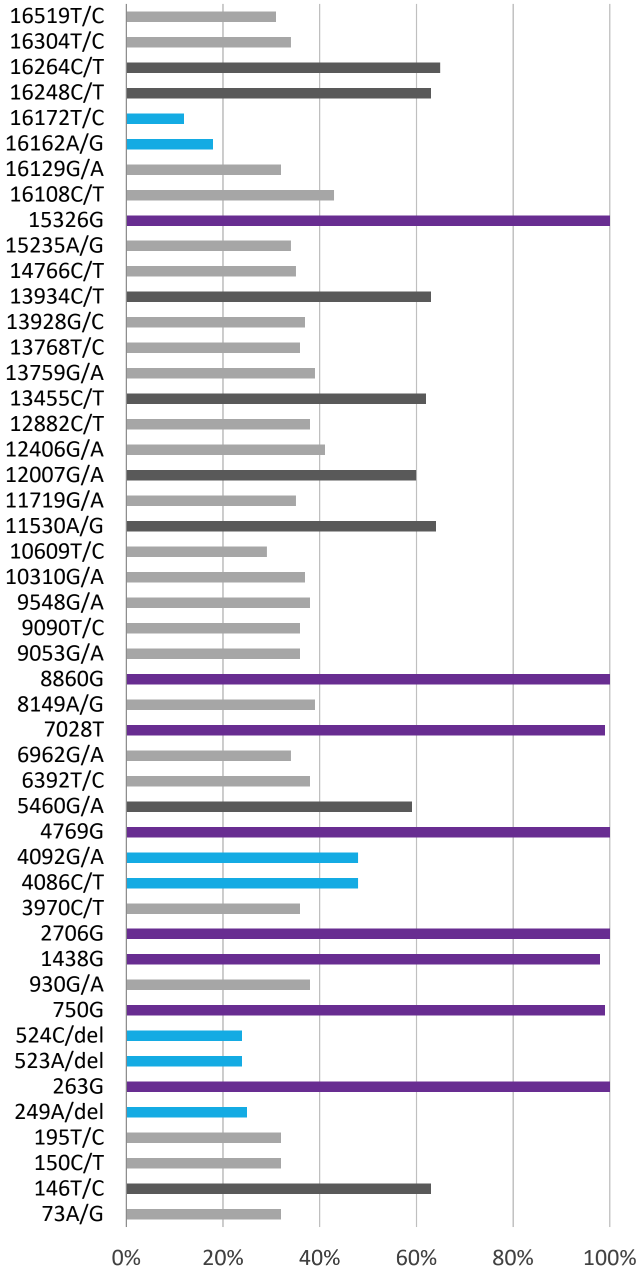


Fig. 8. Two-person 1:1 mixture with contributors from distinct haplogroups (HV and F1a1a) showing the ratio (in %) of alternative allele read depth to np total read depth per variant detected. Variants present in both contributors are highlighted with purple, variants present in the major contributor are highlighted with dark grey, variants present in the minor contributor are highlighted with light grey, variants that could not be assigned to a contributor based on the quantitative data and required additional phasing and phylogenetic analysis are highlighted with blue.

Taking into account the above presented results from two-person mixture with contributors from distinct haplogroups, analysis of two-person mixtures with contributors from similar haplogroups (U2e1a1 and U2e2a1a) with mixtures of 1:1, 1:5 and 5:1 was performed. Full major and minor contributor profiles were obtained from all mixtures. Quantitative data results for the mixtures are shown in the table 7. In the current study no additional phasing or phylogenetic assignment data were needed (Fig. 12). Nevertheless, in the case of phylogenetically close haplogroups the use of phylogenetic assignment in mixture deconvolution is less practical as fewer differences compared to distinct haplogroups are expected. From the variants detected in the 1:1 mixtures 27.3% (from the total of 44 variants) belonged to only one of the contributors in the U2e1a1 and U2e2a1a haplogroups mixture, while 79.2% (from the total of 48 variants) belonged to only one of the contributors in the HV and F1a1a haplogroups mixture.

Table 7. Quantitative data results of two-person mixtures with contributors from similar haplogroups (U2e1a1 and U2e2a1a). The average ratio of alternative allele read depth to total read depth is shown with standard deviation in parentheses.

	1:1 mixture	5:1 mixture	1:5 mixture
Both contributors	99.09% (0.93%)	99.13% (0.94%)	99.06% (1.39%)
Major contributor	59.00% (2.53%)	85.33% (1.75%)	76.50% (3.94%)
Minor contributor	38.80% (1.75%)	12.17% (1.33%)	21.33% (1.21%)

Three-person mixtures

For the three-person mixtures contributors from distinct haplogroups (U2e2a1a, HV and F1a1a) were used. The three-person mixture in the ratio of 1:1:1 resulted in no profiles for contributors as variants could not be assigned to the contributor categories based on quantitative data (Fig. 13). Only variants present in all contributors were distinguishable. All the variants from three contributors were detected. Therefore, while quantitatively variant assignment was not possible, the mixture analysis results can be used to exclude individuals from the contributors' list.

The number of contributors greater than two was indicated by the phasing information as well as a three-allelic position. Variants 146C, 150T, 152C, 195C and 217C reside in the same read area however were in-phase as follows - 150T and 195C, 152C and 217C (Fig. 14A). However, manual parsing of contributors by using phasing data would generate a high number of possible profiles and therefore would not be a practical analysis method. A three-allelic position (16129) was seen in both 1:1:1 as well as 5:1:1 mixtures (Fig. 14B and C).

A correct major contributor's profile was generated from the three-person 5:1:1 mixture. The profiles of the two minor contributors could not be parsed due to similar quantitative results as well as a number of variants from two minor contributors were not detected.

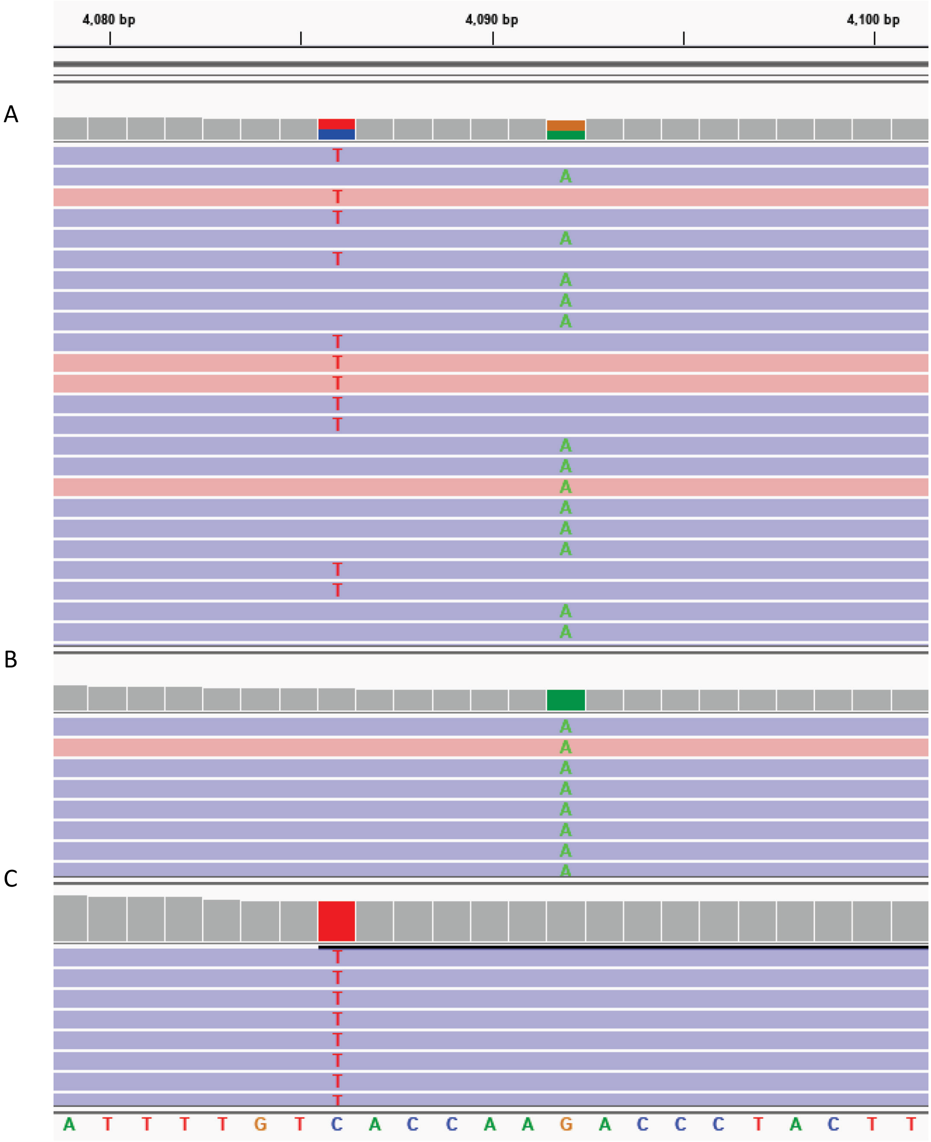
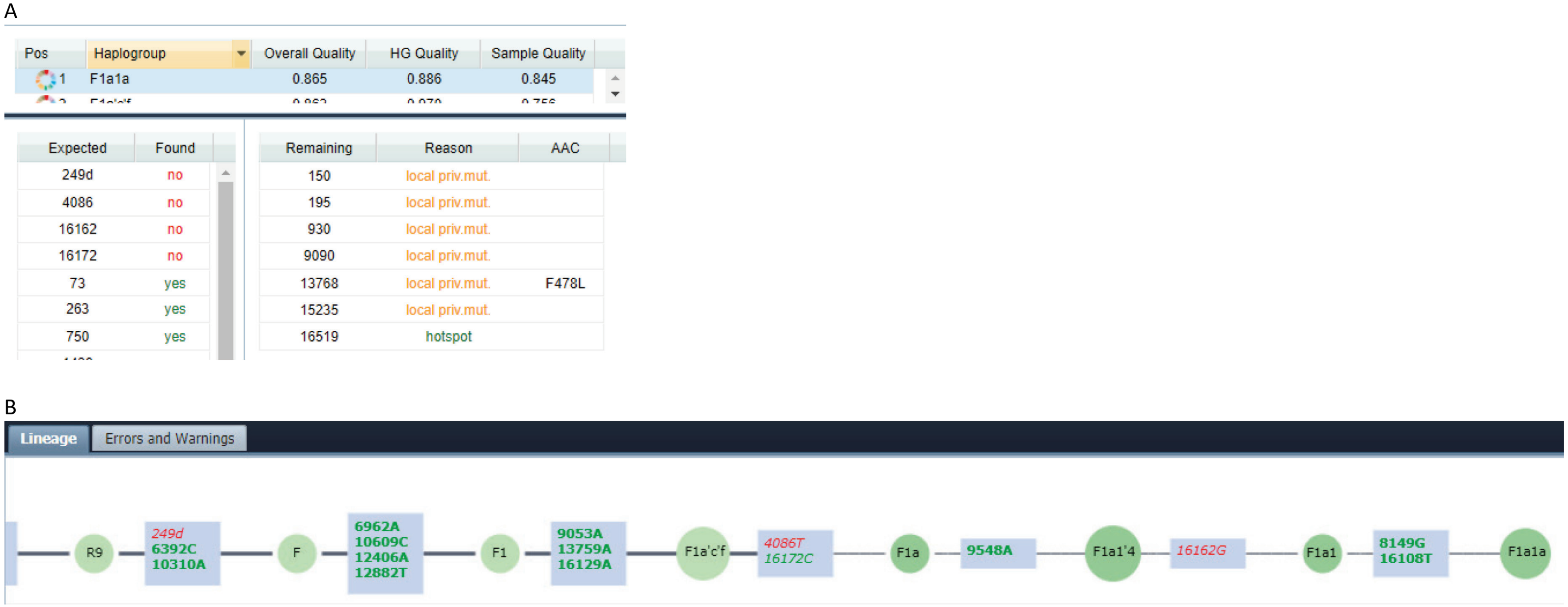


Fig. 9. Image from IGV showing variant positions out of phase. (A) The 1:1 mixture sample with positions 4086T and 4092A were not in-phase. (B and C) Single-source samples used for the aforementioned mixture expressing 4086T and 4092A variants. Reference genome (rCRS+80) sequence is shown in the bottom panel.



A

Pos	Haplogroup	Overall Quality	HG Quality	Sample Quality
1	F1a1a	0.865	0.886	0.845
2	F1a'c'f	0.863	0.970	0.756

Expected	Found
249d	no
4086	no
16162	no
16172	no
73	yes
263	yes
750	yes

Remaining	Reason	AAC
150	local priv.mut.	
195	local priv.mut.	
930	local priv.mut.	
9090	local priv.mut.	
13768	local priv.mut.	F478L
15235	local priv.mut.	
16519	hotspot	

B

Fig. 10. Image from HaploGrep showing phylogenetic assignment data. The major contributor's profile excluding positions 249del, 4086T, 16162G and 16172C was uploaded to HaploGrep. Aforementioned positions are the nodes for the F1a1a haplogroup and therefore were expected in the profile, as shown in (A) the table as well as (B) lineage chart. According to HaploGrep these positions were not expected in the minor contributor's profile. Please note: position 16172C in the lineage chart should be visualized in red font, however an occurring error with the graphical representation has been confirmed (via e-mail communication) by the HaploGrep author (Weissensteiner, H.).

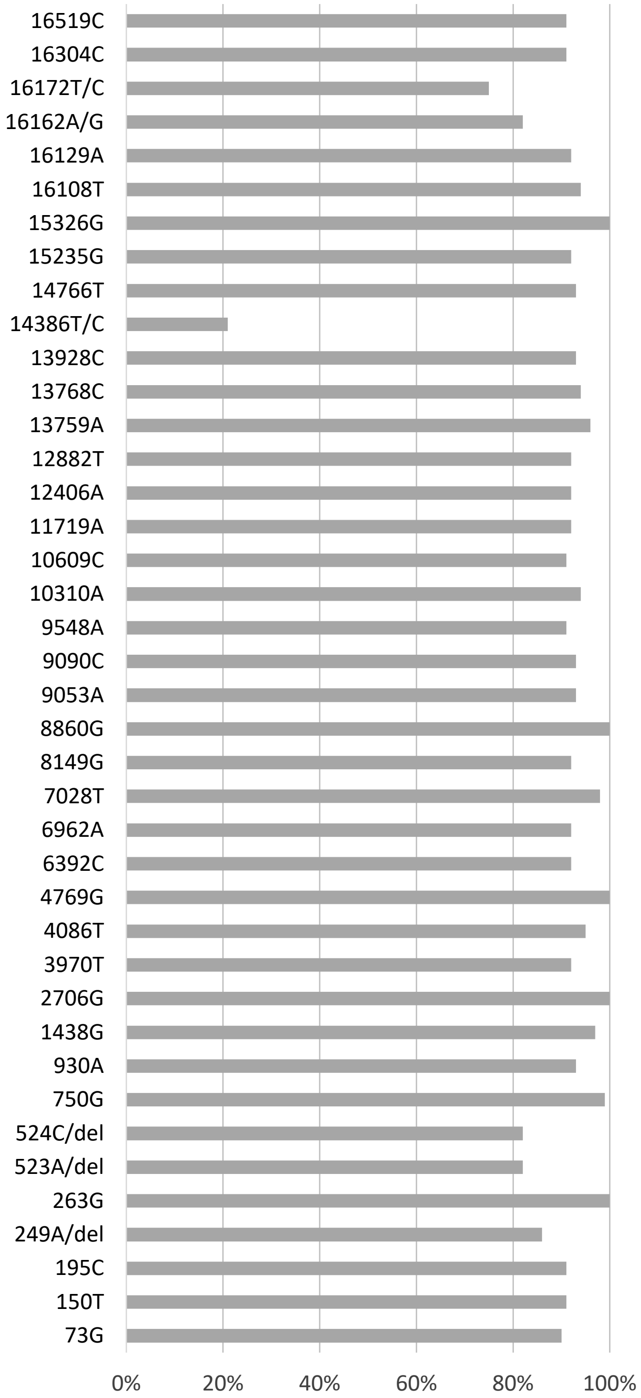


Fig. 11. Two-person 1:20 mixture showing the ratio (in %) of alternative allele read depth to np total read depth per variant detected.

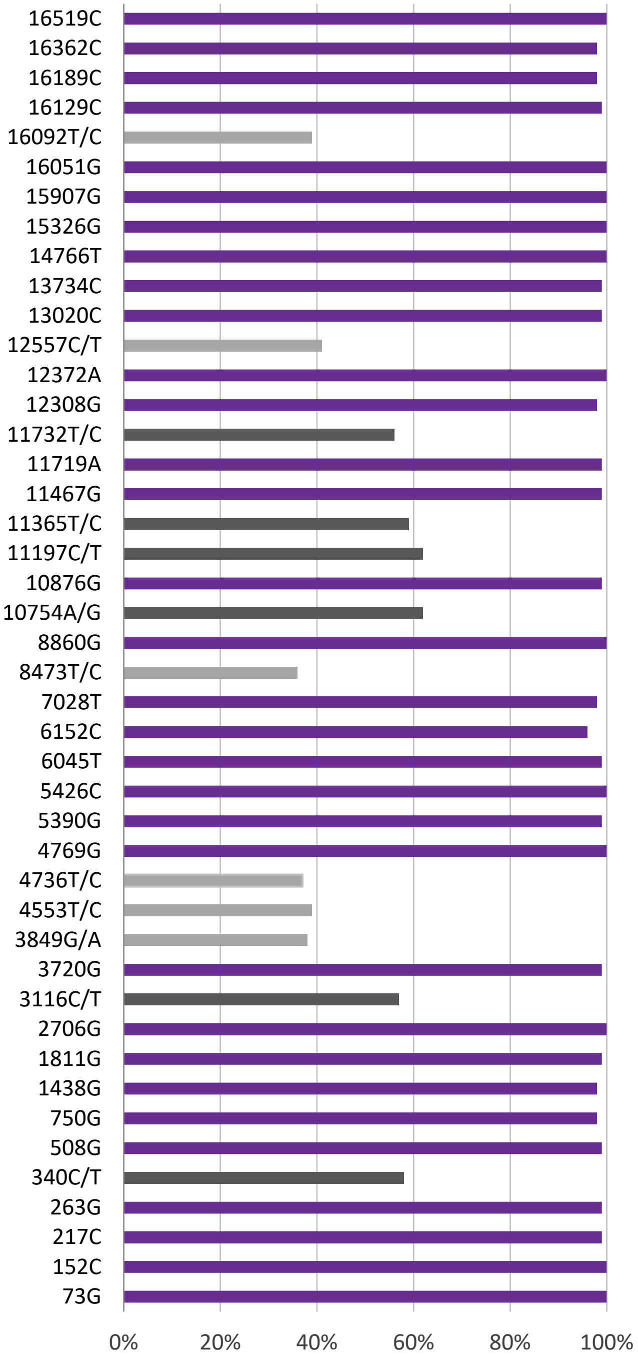


Fig. 12. Two-person 1:1 mixture with contributors from similar haplogroups (U2e1a1 and U2e2a1a) showing the ratio (in %) of alternative allele read depth to np total read depth per variant detected. Variants present in both contributors are highlighted with purple, variants present in the major contributor are highlighted with dark grey, and variants present in the minor contributor are highlighted with light grey.

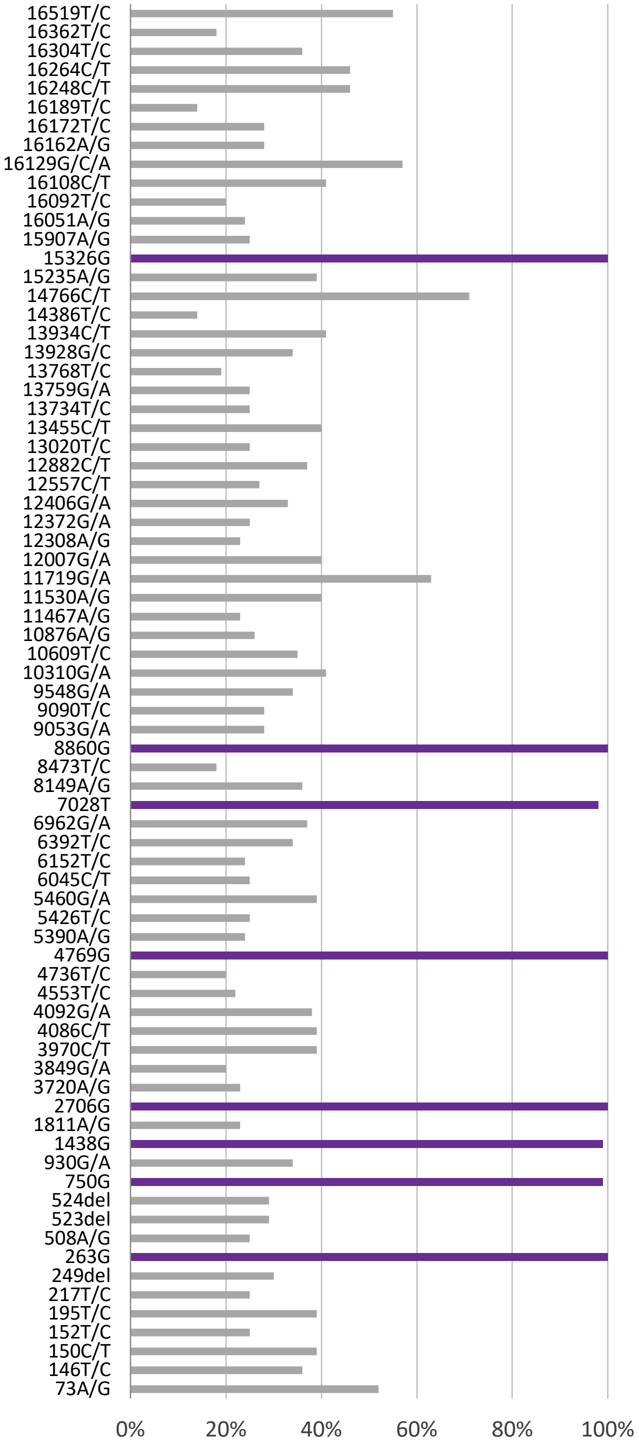
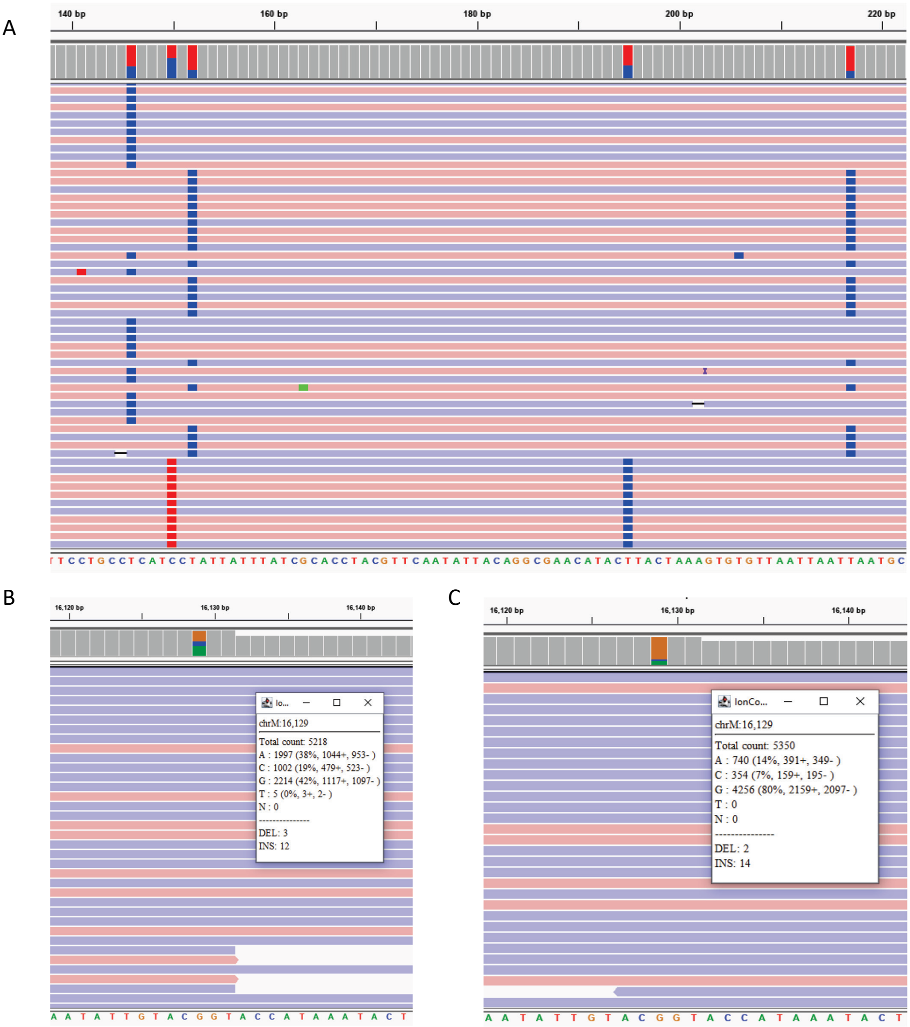


Fig. 13. Three-person 1:1:1 mixture showing the ratio (in %) of alternative allele read depth to np total read depth per variant detected. Variants present in all three contributors are highlighted with purple.


A
140 bp
160 bp
180 bp
200 bp
220 bp
TCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGC
B
16,120 bp
16,130 bp
16,140 bp
AATATTGTACGGTACCATAAATACT
C
16,120 bp
16,130 bp
16,140 bp
AATATTGTACGGTACCATAAATACT
Io...
IonCo...

chrM:16,129

Total count: 5218
A : 1997 (38%, 1044+, 953-)
C : 1002 (19%, 479+, 523-)
G : 2214 (42%, 1117+, 1097-)
T : 5 (0%, 3+, 2-)
N : 0

DEL: 3
INS: 12

chrM:16,129

Total count: 5350
A : 740 (14%, 391+, 349-)
C : 354 (7%, 159+, 195-)
G : 4256 (80%, 2159+, 2097-)
T : 0
N : 0

DEL: 2
INS: 14

Fig. 14. Images from IGV showing phasing and three-allelic data from three-person mixture samples. (A) In the 1:1:1 mixture variants 146C, 150T, 152C, 195C and 217C are situated in the same read area. Variants 150T and 195C as well as 152C and 217C are in-phase. Variant 146C is not in-phase with the other 4 variants. These results indicate a minimum of 3 contributors. (B) The three-allelic position 16129 in the 1:1:1 mixture. (C) The three-allelic position 16129 in the 5:1:1 mixture. Reference genome (rCRS+80) sequence is shown in the bottom panel. The allele count is shown in the white square. Allele count to np total read depth is shown in %. "+" and "-" as well as blue and pink lines indicate the number of forward and reverse strands.

Number of contributors

Single-source and mixture samples performance metrics (read depth, RLP, strand balance, noise) as well as variant counts were compared in order to find a potential indicator between the sample sequencing data and the number of contributors.

Read depth as well as RLP for the mixture sample expressed high and low amplicon areas across the mtGenome in the similar manner as single-source samples. Strand balance as well as detected noise of mixture samples were comparable to the stand balance of single-source samples. Therefore none of these mtDNA performance metrics analysed in the current study could be used to indicate the number of contributors in the sample.

The number of two-allelic and three-allelic positions per sample was examined next (Fig. 15). In a single-source sample these positions would be indicative of a point heteroplasmy. As expected, a number of variants expressing a two- or three-allelic position was the highest in three-person mixtures. In the 5:1:1 mixture approximately half of the two-allelic positions were detected under but close to the established threshold of 10%. Therefore, lowering the threshold might be important for detecting minor alleles of low DNA amount contributors.

Two-person mixture samples from district haplogroups with mixture ratios of 1:1, 1:5 and 1:10 exhibited a higher amount of two-allelic positions compared to other two-person mixtures (from distinct and well as similar haplogroups). The major as well as the minor contributor's mtDNA profiles were detected in the former mixtures, while only the major contributor's profile was distinguished from the distinct haplogroup contributors' mixture of 5:1, 10:1, 1:20 and 20:1. Therefore, the latter are expected to be similar to that of two-allelic position count to single-source samples. Two-person mixtures with contributors from similar haplogroups showed a higher two-allelic position count compared to single-source samples however lower compared to mixture samples with distinct haplogroups. This observation indicates that building a probabilistic model for contributor number estimation based on number of differences between populations, amount of multi-allelic positions and type of multi-allelic positions (e.g. two-allelic and three-allelic positions detected in the current study) might be possible. Population studies as well as mixture studies with high number of samples are needed.

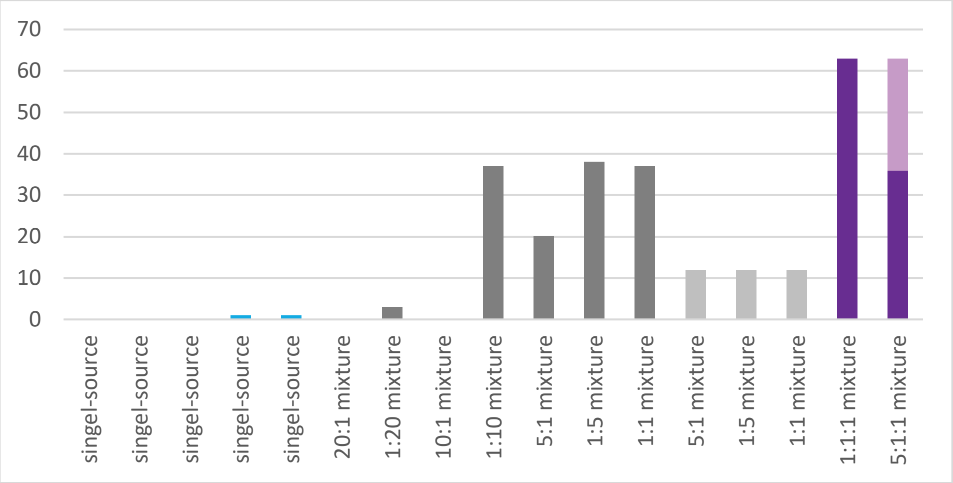


Fig. 15. Number of positions indicating a mixture. Sample compositions are shown on the x-axis, while the number of positions (per sample) indicating a mixture are shown on the y-axis. Number of positions indicating a mixture are shown in blue for single-source samples, dark grey for two-person mixtures from distinct haplogroups, light grey for two-person mixtures from similar haplogroups and purple for three-person mixtures. In the three-person mixtures dark purple indicates the positions from a mixture with the minor variant(s) over the 10% threshold. Light purple indicates the positions indicating a mixture with the minor variant(s) just below 10% threshold (6%-9%).

In publications I and II of the current study, heteroplasmy (with the threshold of 18%) was detected in 11.4% of the Estonian population sample (DNA extracted from buccal swabs) and in 26.5%, 23.9% and 21.8% of the Caucasian, Hispanic and African-American population samples (DNA extracted from whole blood), respectively (King, LaRue, *et al.*, 2014; Stoljarova *et al.*, 2016). The number of heteroplasmic positions in samples ranged from one to four, with four positions detected in only one sample which is 1.2% from samples with heteroplasmy detected and 0.3% from the total number of 397 samples. However different tissue types express different amounts of PHP per person, with higher frequency of heteroplasmic variants in hair shaft compared to blood and buccal swab samples, as reported by Kim *et al.*, (B. M. Kim, Hong, Chun, Kim, & Shin, 2019). This same study also reported the detection of 4 and 5 PHPs in hair shaft samples, however in the majority of the PHP positions expressed the minor allele in a proportion of < 10% (B. M. Kim *et al.*, 2019). Therefore, when determining the possible amounts of contributors with the number of heteroplasmic positions the tissue type should be taken into account. Moreover, a heteroplasmy threshold should be established for each tissue type for the forensic investigators to follow. In has also been shown that different software might produce different results regarding detected positions and major-minor allele ratios and thus the software used can impact detection above and below thresholds (M. Y. Kim, Cho, Lee, Seo, & Lee, 2018).

Vohr *et al.*, described the software tool mixemt for mtDNA mixture deconvolution based on phylogenetic assignment (Vohr *et al.*, 2017). The tool uses BAM files as input and estimates the probability of the read originating from each candidate haplotype via PhyloTree (van Oven & Kayser, 2009). However, taking into account the large memory requirements for computation as well as analysis time, the tool was unwieldy. Additionally the manual analysis performed in the current study outperformed the aforementioned software tool with the same mixture samples.

A number of other methods have been proposed for mtDNA mixture interpretation, e.g. haplotype-specific extraction, locked nucleic acid (LNA) mediated PCR clamping and single-cell sequencing (Asari *et al.*, 2019; Louhelainen & Miller, 2020; Morris *et al.*, 2017; Zander, Otremba, & Nagy, 2018). In two former methods specific probes are required to attach to the amplicon of interest and either separate it from the rest via beads or block its PCR amplification (Asari *et al.*, 2019; Zander *et al.*, 2018). Although promising results have been reported with these methods, these methods require high quality samples and in addition to sample sequencing, extra steps in the wet-lab sample preparation. Single-cell sequencing that allows for the analysis of DNA from a single cell or even a single mitochondrion has been demonstrated which could be another promising method for mixture interpretation (Louhelainen & Miller, 2020; Morris *et al.*, 2017). However single-cell sequencing is laborious and currently not feasible to be used routinely on forensic samples. More studies with larger sample sets are required to determine the reliability of single-cell sequencing technology for forensic applications.

With successful sequencing of 283 US and 114 Estonian population samples we have shown that high throughput makes MPS a viable alternative to routinely used STS. Moreover, generation of whole mtGenome data by MPS, which is impracticable with STS, have a significantly higher discrimination power compared to routinely analysed HVI/HVII region.

Accordingly, compared to STS, MPS enables typing low frequency (under 15%) mtDNA heteroplasmic positions along the entire mtGenome and provides quantitative data (i.e. the ratio of an alternative allele to the total read depth). Both can be used for the interpretation of one of the most challenging forensic samples – mixtures. In addition, bioinformatics tools used for MPS data analysis such as IGV (Thorvaldsdottir *et al.*, 2013), visualize sequenced fragments (or clones) and gives phasing information (i.e. multiple variants residing in one read) regarding variants detected that can also be used for mixture deconvolution. Therefore, MPS highly outperforms STS in generation of data that can be used for mixture sample analysis.

By using quantitative data, phasing information as well as phylogenetic assignment, we were able to determine the mtDNA profile of a major contributor in two-person distinct as well as similar haplogroup mixture samples with a ratio of 1:1, 5:1, 10:1 and 20:1. Regarding minor contributor, the correct mtDNA profiles were determine in two-person distinct as well as haplogroup mixture samples with a ratio of 1:1, 5:1 and 10:1. The correct mtDNA profile for the major contributor from a three-person distinct haplogroup mixture in a ratio of 5:1:1 was also achieved. Difficulties were encountered with personal point heteroplasmic positions as major contributor's heteroplasmy positions were falling into the quantitative range of minor contributor's variants and therefore could be falsely assigned as well as minor contributor's heteroplasmic positions fell below used threshold and were not detected. In addition, quantitative analysis as well as phasing information did not provide enough data to parse contributors' profiles in the three-person 1:1:1 mixture. Due to the amount of variants shared, the number of possible combinations of contributors' mtDNA profiles would be too large to parse manually. Therefore more studies with larger sample sets are needed to address these issues.

We also showed that the amount of heteroplasmic positions detected in the mixture sample could indicate the amount of contributors. Studies with higher mixture sample sets are needed to test the possibility of determining the number of contributors based on the mixture positions. MPS generated whole mtGenome population studies, as in case of 283 US and 114 Estonian population samples, provide data on the number of differences (pairwise comparison) between similar as well as distinct haplogroups. In addition, population studies can be used to determine the amount of heteroplasmy positions in single-source samples in different tissues as well as populations (B. M. Kim *et al.*, 2019; King, LaRue, *et al.*, 2014; Stoljarova *et al.*, 2016). These data lay the foundation for building a probabilistic model for contributors' determination based on number of variants detected.

With furtherer development mixture interpretation via mtDNA can become routinely used by forensic case work institutions.

Conclusion

The main conclusions are presented as follows:

- The throughput of MPS enabled sequencing of the entire mtGenome of 283 US population samples. Moreover, due to the sequencing of individual molecules (or clones) point as well as length heteroplasmic positions were typed. Although low read depth and strand balance at a number of positions were observed, the results were not affected at these regions. MPS notably outperformed the routinely used STS.
- In addition to 283 US population samples, mtGenomes were sequenced from 114 Estonian population samples. To our knowledge, these are the first and the only Estonian entire mtGenome sequences published.
- Over 70% of mtDNA variants reside outside of routinely analysed HVI/HVII regions. The discrimination power of mtGenome data is significantly higher compared to HVI/HVII data – $p = 0.01659$ and $p = 0.01645$ for RMP and GD, respectively.
- The Estonian samples showed the largest increase in unique haplotypes from HVI/HVII region data going to mtGenome data – 27.1% increase for the Estonian population, 15.6% for the Hispanic population, 11.8% for the African American population and 7.8% for the Caucasian population. Based on the studied population, we observed that the addition of coding region data benefits differentiation of samples with haplogroups that have a low amount of variants within the HVI/HVII region, e.g. haplogroup H.
- Phylogenetic assignment based on solely HVI/HVII data might lead to a falsely assigned haplogroup - five samples from 283 US population samples changed macrohaplogroups. One sample was changed from Asia-specific mtDNA haplogroup D4j1b2 assigned based on HVI/HVII data to Africa-specific L3b1a7a haplogroup.
- Twelve major haplogroup clades were observed in the Estonian population sample based on entire mtGenome data. The major haplogroups were H (with the frequency of 47.4%), U (21.1%), T (9.6%) and J (6.1%). The rest of the haplogroup clades – D, I, K, M, N, R, W and X – included less than 5.0% of the samples. All major haplogroup clades observed in the current study were reported in the Estonian and/or geographically nearby populations by previous studies based on partial mtDNA sequences.
- By using MPS generated quantitative, phasing and phylogenetic data the interpretation of one of the most challenging forensic samples – mixture samples – was achieved. The correct mtDNA profile for the major contributor was obtained from two-person distinct as well as similar haplogroup mixtures in a ratio of 1:1, 5:1, 10:1 and 20:1. The correct mtDNA profiles for the minor contributor were determined in the 1:1, 5:1 and 10:1 samples. Personal heteroplasmy positions were difficult to interpret.
- The interpretation of three-person distinct haplogroup mixtures was more difficult compared to two-person mixtures. The correct mtDNA profile for the major contributor was determined from the 5:1:1 mixture but not the 1:1:1 mixture. The use of phasing information in the case of the three-person mixtures was less helpful due to a high number of possible haplotype combinations.

More studies with a high number of contributors (with additional three person mixtures and over three person mixtures) are needed.

- The number of variants detected in the mixture sample can be an indication for the number of contributors. The number of variants detected in the three-person mixture was higher compared to the number of variants detected in the two-person mixtures. As expected, two-person mixtures from distinct haplogroups had a higher amount of variants compared to two-person mixtures from similar haplogroups. Mixture studies with higher sample sets and different populations are needed to evaluate if establishment of a probabilistic model for mixture contributor number is feasible.

References

Scientific Working Group on DNA Analysis Methods (SWGDAM) (2017). Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories – APPROVED 01/12/2017. Retrieved from https://1ecb9588-ea6f-4feb-971a-73265dbf079c.filesusr.com/ugd/4344b0_50e2749756a242528e6285a5bb478f4c.pdf

Scientific Working Group on DNA Analysis Methods (SWGDAM) (2019). Interpretation Guidelines for Mitochondrial DNA Analysis by Forensic DNA Testing Laboratories – Approved 04/23/2019. Retrieved from https://1ecb9588-ea6f-4feb-971a-73265dbf079c.filesusr.com/ugd/4344b0_f61de6abf3b94c52b28139bff600ae98.pdf

Achilli, A., Rengo, C., Magri, C., Battaglia, V., Olivieri, A., Scozzari, R., . . . Torroni, A. (2004). The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet, 75(5), 910-918. doi:10.1086/425590

Alvarez-Iglesias, V., Mosquera-Miguel, A., Cerezo, M., Quintans, B., Zarrabeitia, M. T., Cusco, I., . . . Salas, A. (2009). New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. PLoS One, 4(4), e5112. doi:10.1371/journal.pone.0005112

Alvarez-Mora, M. I., Podlesniy, P., Gelpi, E., Hukema, R., Madrigal, I., Pagonabarraga, J., . . . Rodriguez-Revenga, L. (2019). FXTAS: regional decrease of mitochondrial DNA copy number relates to clinical manifestations. Genes Brain Behav, e12565. doi:10.1111/gbb.12565

Ambardar, S., Gupta, R., Trakroo, D., Lal, R., & Vakhlu, J. (2016). High Throughput Sequencing: An Overview of Sequencing Chemistry. Indian J Microbiol, 56(4), 394-404. doi:10.1007/s12088-016-0606-4

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., . . . Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. Nature, 290(5806), 457-465.

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., & Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet, 23(2), 147. doi:10.1038/13779

Asari, M., Isozaki, S., Hoshina, C., Okuda, K., Tanaka, H., Horioka, K., . . . Shimizu, K. (2019). Discrimination of haplotype in mitochondrial DNA mixtures using LNA-mediated PCR clamping. Forensic Sci Int Genet, 41, 58-63. doi:10.1016/j.fsigen.2019.03.018

Avila, E., Graebin, P., Chemale, G., Freitas, J., Kahmann, A., & Alho, C. S. (2019). Full mtDNA genome sequencing of Brazilian admixed populations: A forensic-focused evaluation of a MPS application as an alternative to Sanger sequencing methods. Forensic Sci Int Genet, 42, 154-164. doi:10.1016/j.fsigen.2019.07.004

Bandelt, H. J., Lahermo, P., Richards, M., & Macaulay, V. (2001). Detecting errors in mtDNA data by phylogenetic analysis. Int J Legal Med, 115(2), 64-69.

Bandelt, H. J., & Salas, A. (2012). Current next generation sequencing technology may not meet forensic standards. Forensic Sci Int Genet, 6(1), 143-145. doi:10.1016/j.fsigen.2011.04.004

Bandelt, H. J., van Oven, M., & Salas, A. (2012). Haplogrouping mitochondrial DNA sequences in Legal Medicine/Forensic Genetics. Int J Legal Med, 126(6), 901-916. doi:10.1007/s00414-012-0762-y

Barrett, A., Arbeithuber, B., Zaidi, A., Wilton, P., Paul, I. M., Nielsen, R., & Makova, K. D. (2020). Pronounced somatic bottleneck in mitochondrial DNA of human hair. Philos Trans R Soc Lond B Biol Sci, 375(1790), 20190175. doi:10.1098/rstb.2019.0175

Bastos-Rodrigues, L., Pimenta, J. R., & Pena, S. D. (2006). The genetic structure of human populations studied through short insertion-deletion polymorphisms. Ann Hum Genet, 70(Pt 5), 658-665. doi:10.1111/j.1469-1809.2006.00287.x

Behar, D. M., van Oven, M., Rosset, S., Metspalu, M., Loogväli, E. L., Silva, N. M., . . . Villems, R. (2012). A "Copernican" reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet, 90(4), 675-684. doi:10.1016/j.ajhg.2012.03.002

Bendall, K. E., Macaulay, V. A., & Sykes, B. C. (1997). Variable levels of a heteroplasmic point mutation in individual hair roots. Am J Hum Genet, 61(6), 1303-1308. doi:10.1086/301636

Bodner, M., Iuvaro, A., Strobl, C., Nagl, S., Huber, G., Pelotti, S., . . . Parson, W. (2015). Helena, the hidden beauty: Resolving the most common West Eurasian mtDNA control region haplotype by massively parallel sequencing an Italian population sample. Forensic Sci Int Genet, 15, 21-26. doi:10.1016/j.fsigen.2014.09.012

Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P., & Tyson, G. W. (2013). Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput Biol, 9(4), e1003031. doi:10.1371/journal.pcbi.1003031

Brandhagen, M. D., Just, R. S., & Irwin, J. A. (2020). Validation of NGS for mitochondrial DNA casework at the FBI Laboratory. Forensic Sci Int Genet, 44, 102151. doi:10.1016/j.fsigen.2019.102151

Brandstatter, A., Parsons, T. J., & Parson, W. (2003). Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups. Int J Legal Med, 117(5), 291-298. doi:10.1007/s00414-003-0395-2

Budowle, B., Wilson, M. R., DiZinno, J. A., Stauffer, C., Fasano, M. A., Holland, M. M., & Monson, K. L. (1999). Mitochondrial DNA regions HVI and HVII population data. Forensic Sci Int, 103(1), 23-35. doi:10.1016/s0379-0738(99)00042-0

Budowle B., Moretti T.R., Niezgoda S. J., & Brown B.L. (1998). CODIS and PCR-based short tandem repeat loci: Law enforcement tools. Paper presented at the Second European Symposium on Human Identification. Retrieved from https://www.promega.com/~/media/files/resources/conference%20proceedings/ishi%2002/oral%20presentations/17.pdf

Butler J. M., Hill C. R., & Coble M. D. (2012). Variability of New STR Loci and Kits in US Population Groups. Profiles in DNA. Retrieved from https://www.promega.ee/resources/profiles-in-dna/2012/variability-of-new-str-loci-and-kits-in-us-population-groups/

Butler, J. M. (2012). Advanced Topics in Forensic DNA Typing: Methodology: Elsevier Inc.

Chaitanya, L., Breslin, K., Zuniga, S., Wirken, L., Pospiech, E., Kukla-Bartoszek, M., . . . Walsh, S. (2018). The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. Forensic Sci Int Genet, 35, 123-135. doi:10.1016/j.fsigen.2018.04.004

Chaitanya, L., Ralf, A., van Oven, M., Kupiec, T., Chang, J., Lagace, R., & Kayser, M. (2015). Simultaneous Whole Mitochondrial Genome Sequencing with Short Overlapping Amplicons Suitable for Degraded DNA Using the Ion Torrent Personal Genome Machine. Hum Mutat, 36(12), 1236-1247. doi:10.1002/humu.22905

Cho, S., Kim, M. Y., Lee, J. H., & Lee, S. D. (2018). Assessment of mitochondrial DNA heteroplasmy detected on commercial panel using MPS system with artificial mixture samples. Int J Legal Med, 132(4), 1049-1056. doi:10.1007/s00414-017-1755-7

Churchill, J. D., King, J. L., Chakraborty, R., & Budowle, B. (2016). Effects of the Ion PGM Hi-Q sequencing chemistry on sequence data quality. Int J Legal Med, 130(5), 1169-1180. doi:10.1007/s00414-016-1355-y

Churchill, J. D., Novroski, N. M. M., King, J. L., Seah, L. H., & Budowle, B. (2017). Population and performance analyses of four major populations with Illumina's FGx Forensic Genomics System. Forensic Sci Int Genet, 30, 81-92. doi:10.1016/j.fsigen.2017.06.004

Coble, M. D. (2011). The identification of the Romanovs: Can we (finally) put the controversies to rest? Investig Genet, 2(1), 20. doi:10.1186/2041-2223-2-20

Coble, M. D., Just, R. S., O’Callaghan, J. E., Letmanyi, I. H., Peterson, C. T., Irwin, J. A., & Parsons, T. J. (2004). Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. Int J Legal Med, 118(3), 137-146. doi:10.1007/s00414-004-0427-6

Coble, M. D., Loreille, O. M., Wadhams, M. J., Edson, S. M., Maynard, K., Meyer, C. E., . . . Finelli, L. N. (2009). Mystery solved: the identification of the two missing Romanov children using DNA analysis. PLoS One, 4(3), e4838. doi:10.1371/journal.pone.0004838

Congiu, A., Anagnostou, P., Milia, N., Capocasa, M., Montinaro, F., & Destro Bisol, G. (2012). Online databases for mtDNA and Y chromosome polymorphisms in human populations. J Anthropol Sci, 90, 201-215. doi:10.4436/jass.90020

Davis, C., Peters, D., Warshauer, D., King, J., & Budowle, B. (2015). Sequencing the hypervariable regions of human mitochondrial DNA using massively parallel sequencing: Enhanced data acquisition for DNA samples encountered in forensic testing. Leg Med (Tokyo), 17(2), 123-127. doi:10.1016/j.legalmed.2014.10.004

Dayama, G., Emery, S. B., Kidd, J. M., & Mills, R. E. (2014). The genomic landscape of polymorphic human nuclear mitochondrial insertions. Nucleic Acids Res, 42(20), 12640-12649. doi:10.1093/nar/gku1038

de la Puente, M., Santos, C., Fondevila, M., Manzo, L., Consortium, E. U.-N., Carracedo, A., . . . Phillips, C. (2016). The Global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs. Forensic Sci Int Genet, 22, 81-88. doi:10.1016/j.fsigen.2016.01.015

Dennis, C. (2003). Error reports threaten to unravel databases of mitochondrial DNA. Nature, 421(6925), 773-774. doi:10.1038/421773a

Derenko, M., Malyarchuk, B., Grzybowski, T., Denisova, G., Rogalla, U., Perkova, M., . . . Zakharov, I. (2010). Origin and post-glacial dispersal of mitochondrial DNA haplogroups C and D in northern Asia. PLoS One, 5(12), e15214. doi:10.1371/journal.pone.0015214

Derenko, M. V., Grzybowski, T., Malyarchuk, B. A., Dambueva, I. K., Denisova, G. A., Czarny, J., . . . Zakharov, I. A. (2003). Diversity of mitochondrial DNA lineages in South Siberia. Ann Hum Genet, 67(Pt 5), 391-411. doi:10.1046/j.1469-1809.2003.00035.x

Duan, M., Tu, J., & Lu, Z. (2018). Recent Advances in Detecting Mitochondrial DNA Heteroplasmic Variations. Molecules, 23(2). doi:10.3390/molecules23020323

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res, 8(3), 186-194. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9521922

Fendt, L., Rock, A., Zimmermann, B., Bodner, M., Thye, T., Tschentscher, F., . . . Parson, W. (2012). MtDNA diversity of Ghana: a forensic and phylogeographic view. Forensic Sci Int Genet, 6(2), 244-249. doi:10.1016/j.fsigen.2011.05.011

Freitas, J. M., Fassio, L. H., Braganholi, D. F., & Chemale, G. (2019). Mitochondrial DNA control region haplotypes and haplogroup diversity in a sample from Brasilia, Federal District, Brazil. Forensic Sci Int Genet, 40, e228-e230. doi:10.1016/j.fsigen.2019.02.006

Ganschow, S., Silvery, J., & Tiemann, C. (2019). Development of a multiplex forensic identity panel for massively parallel sequencing and its systematic optimization using design of experiments. Forensic Sci Int Genet, 39, 32-43. doi:10.1016/j.fsigen.2018.11.023

Gill, P., Ivanov, P. L., Kimpton, C., Piercy, R., Benson, N., Tully, G., . . . Sullivan, K. (1994). Identification of the remains of the Romanov family by DNA analysis. Nat Genet, 6(2), 130-135. doi:10.1038/ng0294-130

Gonzalez, M. D. M., Ramos, A., Aluja, M. P., & Santos, C. (2020). Sensitivity of mitochondrial DNA heteroplasmy detection using Next Generation Sequencing. Mitochondrion, 50, 88-93. doi:10.1016/j.mito.2019.10.006

Grady, J. P., Pickett, S. J., Ng, Y. S., Alston, C. L., Blakely, E. L., Hardy, S. A., . . . McFarland, R. (2018). mtDNA heteroplasmy level and copy number indicate disease burden in m.3243A>G mitochondrial disease. EMBO Mol Med, 10(6). doi:10.15252/emmm.201708262

Greenberg, B. D., Newbold, J. E., & Sugino, A. (1983). Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. Gene, 21(1-2), 33-49. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/6301949

ENFSI DNA Working Group (2017). DNA database management review and recommendations. Retrieved from http://enfsi.eu/wp-content/uploads/2017/09/DNA-databasemanagement-review-and-recommendatations-april-2017.pdf

Grzybowski, T., Malyarchuk, B. A., Derenko, M. V., Perkova, M. A., Bednarek, J., & Wozniak, M. (2007). Complex interactions of the Eastern and Western Slavic populations with other European groups as revealed by mitochondrial DNA analysis. Forensic Sci Int Genet, 1(2), 141-147. doi:10.1016/j.fsigen.2007.01.010

Hares, D. R. (2015). Selection and implementation of expanded CODIS core loci in the United States. Forensic Sci Int Genet, 17, 33-34. doi:10.1016/j.fsigen.2015.03.006

Hazkani-Covo, E., Zeller, R. M., & Martin, W. (2010). Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet, 6(2), e1000834. doi:10.1371/journal.pgen.1000834

He, Y., Wu, J., Dressman, D. C., Iacobuzio-Donahue, C., Markowitz, S. D., Velculescu, V. E., . . . Papadopoulos, N. (2010). Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature, 464(7288), 610-614. doi:10.1038/nature08802

Hedman, M., Brandstatter, A., Pimenoff, V., Sistonen, P., Palo, J. U., Parson, W., & Sajantila, A. (2007). Finnish mitochondrial DNA HVS-I and HVS-II population data. Forensic Sci Int, 172(2-3), 171-178. doi:10.1016/j.forsciint.2006.09.012

Holland, M. M., McQuillan, M. R., & O'Hanlon, K. A. (2011). Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy. Croat Med J, 52(3), 299-313. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/21674826

Hollard, C., Ausset, L., Chantrel, Y., Jullien, S., Clot, M., Faivre, M., . . . Laurent, F. X. (2019). Automation and developmental validation of the ForenSeq() DNA Signature Preparation kit for high-throughput analysis in forensic laboratories. Forensic Sci Int Genet, 40, 37-45. doi:10.1016/j.fsigen.2019.01.010

Human mtDNA Migrations. (2016). Retrieved 11.05.2019, from https://www.mitomap.org/foswiki/pub/MITOMAP/MitomapFigures/WorldMigrations2012.pdf

Irwin, J. A., Ikramov, A., Saunier, J., Bodner, M., Amory, S., Rock, A., . . . Parsons, T. J. (2010). The mtDNA composition of Uzbekistan: a microcosm of Central Asian patterns. Int J Legal Med, 124(3), 195-204. doi:10.1007/s00414-009-0406-z

Irwin, J. A., Saunier, J. L., Niederstatter, H., Strouss, K. M., Sturk, K. A., Diegoli, T. M., . . . Parsons, T. J. (2009). Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. J Mol Evol, 68(5), 516-527. doi:10.1007/s00239-009-9227-4

Ivanov, P. L., Wadhams, M. J., Roby, R. K., Holland, M. M., Weedn, V. W., & Parsons, T. J. (1996). Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. Nat Genet, 12(4), 417-420. doi:10.1038/ng0496-417

Jager, A. C., Alvarez, M. L., Davis, C. P., Guzman, E., Han, Y., Way, L., . . . Holt, C. L. (2017). Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories. Forensic Sci Int Genet, 28, 52-70. doi:10.1016/j.fsigen.2017.01.011

Jarczak, J., Grochowalski, L., Marciniak, B., Lach, J., Slomka, M., Sobalska-Kwapis, M., . . . Strapagiel, D. (2019). Mitochondrial DNA variability of the Polish population. Eur J Hum Genet. doi:10.1038/s41431-019-0381-x

Just, R. S., Scheible, M. K., Fast, S. A., Sturk-Andreaggi, K., Rock, A. W., Bush, J. M., . . . Irwin, J. A. (2015). Full mtGenome reference data: development and characterization of 588 forensic-quality haplotypes representing three U.S. populations. Forensic Sci Int Genet, 14, 141-155. doi:10.1016/j.fsigen.2014.09.021

Kennedy, S. R., Salk, J. J., Schmitt, M. W., & Loeb, L. A. (2013). Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet, 9(9), e1003794. doi:10.1371/journal.pgen.1003794

Keseru, J. S., Soltesz, B., Lukacs, J., Marton, E., Szilagyi-Bonizs, M., Penyige, A., . . . Nagy, B. (2019). Detection of cell-free, exosomal and whole blood mitochondrial DNA copy number in plasma or whole blood of patients with serous epithelial ovarian cancer. J Biotechnol. doi:10.1016/j.jbiotec.2019.04.015

Khusnutdinova, E., Gilyazova, I., Ruiz-Pesini, E., Derbeneva, O., Khusainova, R., Khidiyatova, I., . . . Wallace, D. C. (2008). A mitochondrial etiology of neurodegenerative diseases: evidence from Parkinson's disease. Ann N Y Acad Sci, 1147, 1-20. doi:10.1196/annals.1427.001

Kim, B. M., Hong, S. R., Chun, H., Kim, S., & Shin, K. J. (2019). Comparison of whole mitochondrial genome variants between hair shafts and reference samples using massively parallel sequencing. Int J Legal Med. doi:10.1007/s00414-019-02205-y

Kim, H., Erlich, H. A., & Calloway, C. D. (2015). Analysis of mixtures using next generation sequencing of mitochondrial DNA hypervariable regions. Croat Med J, 56(3), 208-217. doi: 10.3325/cmj.2015.56.208

Kim, M. Y., Cho, S., Lee, J. H., Seo, H. J., & Lee, S. D. (2018). Detection of Innate and Artificial Mitochondrial DNA Heteroplasmy by Massively Parallel Sequencing: Considerations for Analysis. J Korean Med Sci, 33(52), e337. doi:10.3346/jkms.2018.33.e337

King, J. L., Churchill, J. D., Novroski, N. M. M., Zeng, X., Warshauer, D. H., Seah, L. H., & Budowle, B. (2018). Increasing the discrimination power of ancestry- and identity-informative SNP loci within the ForenSeq DNA Signature Prep Kit. Forensic Sci Int Genet, 36, 60-76. doi:10.1016/j.fsigen.2018.06.005

King, J. L., LaRue, B. L., Novroski, N. M., Stoljarova, M., Seo, S. B., Zeng, X., . . . Budowle, B. (2014). High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int Genet, 12, 128-135. doi:10.1016/j.fsigen.2014.06.001

King, J. L., Sajantila, A., & Budowle, B. (2014). mitoSAVE: mitochondrial sequence analysis of variants in Excel. Forensic Sci Int Genet, 12, 122-125. doi:10.1016/j.fsigen.2014.05.013

Kirkwood, T. B., & Kowald, A. (2012). The free-radical theory of ageing - older, wiser and still alive: modelling positional effects of the primary targets of ROS reveals new support. Bioessays, 34(8), 692-700. doi:10.1002/bies.201200014

Kloss-Brandstatter, A., Pacher, D., Schonherr, S., Weissensteiner, H., Binna, R., Specht, G., & Kronenberg, F. (2011). HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat, 32(1), 25-32. doi:10.1002/humu.21382

Kogelnik, A. M., Lott, M. T., Brown, M. D., Navathe, S. B., & Wallace, D. C. (1996). MITOMAP: a human mitochondrial genome database. Nucleic Acids Res, 24(1), 177-179. doi:10.1093/nar/24.1.177

Kushniarevich, A., Utevska, O., Chuhryaeva, M., Agdzhoyan, A., Dibirova, K., Uktveryte, I., . . . Balanovsky, O. (2015). Genetic Heritage of the Balto-Slavic Speaking Populations: A Synthesis of Autosomal, Mitochondrial and Y-Chromosomal Data. PLoS One, 10(9), e0135820. doi:10.1371/journal.pone.0135820

Ladd, C., Lee, H. C., Yang, N., & Bieber, F. R. (2001). Interpretation of complex forensic DNA mixtures. Croat Med J, 42(3), 244-246. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11387631

Lappalainen, T., Laitinen, V., Salmela, E., Andersen, P., Huoponen, K., Savontaus, M. L., & Lahermo, P. (2008). Migration waves to the Baltic Sea region. Ann Hum Genet, 72(Pt 3), 337-348. doi:10.1111/j.1469-1809.2007.00429.x

LaRue, B. L., Lagace, R., Chang, C. W., Holt, A., Hennessy, L., Ge, J., . . . Budowle, B. (2014). Characterization of 114 insertion/deletion (INDEL) polymorphisms, and selection for a global INDEL panel for human identification. Leg Med (Tokyo), 16(1), 26-32. doi:10.1016/j.legalmed.2013.10.006

Lee, H. Y., Yoo, J. E., Park, M. J., Chung, U., Kim, C. Y., & Shin, K. J. (2006). East Asian mtDNA haplogroup determination in Koreans: haplogroup-level coding region SNP analysis and subhaplogroup-level control region sequence analysis. Electrophoresis, 27(22), 4408-4418. doi:10.1002/elps.200600151

Levin, B. C., Cheng, H., & Reeder, D. J. (1999). A human mitochondrial DNA standard reference material for quality control in forensic identification, medical diagnosis, and mutation detection. Genomics, 55(2), 135-146. doi:10.1006/geno.1998.5513

Li, M., Schonberg, A., Schaefer, M., Schroeder, R., Nasidze, I., & Stoneking, M. (2010). Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. Am J Hum Genet, 87(2), 237-249. doi:10.1016/j.ajhg.2010.07.014

Li, M., Schroder, R., Ni, S., Madea, B., & Stoneking, M. (2015). Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations. Proc Natl Acad Sci U S A, 112(8), 2491-2496. doi:10.1073/pnas.1419651112

Lin, L., Cui, P., Qiu, Z., Wang, M., Yu, Y., Wang, J., . . . Zhao, H. (2019). The mitochondrial tRNA(Ala) 5587T>C and tRNA(Leu(CUN)) 12280A>G mutations may be associated with hypertension in a Chinese family. Exp Ther Med, 17(3), 1855-1862. doi:10.3892/etm.2018.7143

Linch, C. A., Whiting, D. A., & Holland, M. M. (2001). Human hair histogenesis for the mitochondrial DNA forensic scientist. J Forensic Sci, 46(4), 844-853. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11451065

Lindberg, M. R., Schmedes, S. E., Hewitt, F. C., Haas, J. L., Ternus, K. L., Kadavy, D. R., & Budowle, B. (2016). A Comparison and Integration of MiSeq and MinION Platforms for Sequencing Single Source and Mixed Mitochondrial Genomes. PLoS One, 11(12), e0167600. doi:10.1371/journal.pone.0167600

Loogväli, E. L., Roostalu, U., Malyarchuk, B. A., Derenko, M. V., Kivisild, T., Metspalu, E., . . . Villems, R. (2004). Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. Mol Biol Evol, 21(11), 2012-2021. doi:10.1093/molbev/msh209

Louhelainen, J., & Miller, D. (2020). Forensic Investigation of a Shawl Linked to the “Jack the Ripper” Murders. J Forensic Sci, 65(1), 295-303. doi:10.1111/1556-4029.14038

van Oven, M. (2016). PhyloTree.org - mtDNA tree Build 17 (18 Feb 2016) Retrieved from https://www.phylotree.org/tree/index.htm

Marquis, J., Lefebvre, G., Kourmpetis, Y. A. I., Kassam, M., Ronga, F., De Marchi, U., . . . Descombes, P. (2017). MitoRS, a method for high throughput, sensitive, and accurate detection of mitochondrial DNA heteroplasmy. BMC Genomics, 18(1), 326. doi:10.1186/s12864-017-3695-5

Marshall, C., Kimberly, S.-A., Ring, J. D., Taylor, C. R., Suzanne Barritt-Ross, S., Parson, W., & McMahon, T. P. (2019). Advancing mitochondrial genome data interpretation in missing persons casework. Forensic Science International: Genetics Supplement Series, Series 7, 721-723.

McElhoe, J. A., Holland, M. M., Makova, K. D., Su, M. S., Paul, I. M., Baker, C. H., . . . Young, B. (2014). Development and assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq. Forensic Sci Int Genet, 13, 20-29. doi:10.1016/j.fsigen.2014.05.007

Mikkelsen, M., Frank-Hansen, R., Hansen, A. J., & Morling, N. (2014). Massively parallel pyrosequencing of the mitochondrial genome with the 454 methodology in forensic genetics. Forensic Sci Int Genet, 12, 30-37. doi:10.1016/j.fsigen.2014.03.014

Miller, F. J., Rosenfeldt, F. L., Zhang, C., Linnane, A. W., & Nagley, P. (2003). Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: lack of change of copy number with age. Nucleic Acids Res, 31(11), e61. doi:10.1093/nar/gng060

Morris, J., Na, Y. J., Zhu, H., Lee, J. H., Giang, H., Ulyanova, A. V., . . . Eberwine, J. (2017). Pervasive within-Mitochondrion Single-Nucleotide Variant Heteroplasmy as Revealed by Single-Mitochondrion Sequencing. Cell Rep, 21(10), 2706-2713. doi:10.1016/j.celrep.2017.11.031

Mourier, T., Hansen, A. J., Willerslev, E., & Arctander, P. (2001). The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus. Mol Biol Evol, 18(9), 1833-1837. doi:10.1093/oxfordjournals.molbev.a003971

Naue, J., Horer, S., Sanger, T., Strobl, C., Hatzer-Grubwieser, P., Parson, W., & Lutz-Bonengel, S. (2015). Evidence for frequent and tissue-specific sequence heteroplasmy in human mitochondrial DNA. Mitochondrion, 20, 82-94. doi:10.1016/j.mito.2014.12.002

Pakendorf, B., & Stoneking, M. (2005). Mitochondrial DNA and human evolution. Annu Rev Genomics Hum Genet, 6, 165-183. doi:10.1146/annurev.genom.6.080604.162249

Pakstis, A. J., Speed, W. C., Soundararajan, U., Rajeevan, H., Kidd, J. R., Li, H., & Kidd, K. K. (2019). Population relationships based on 170 ancestry SNPs from the combined Kidd and Seldin panels. Sci Rep, 9(1), 18874. doi:10.1038/s41598-019-55175-x

Paneto, G. G., Martins, J. A., Longo, L. V., Pereira, G. A., Freschi, A., Alvarenga, V. L., . . . Cicarelli, R. M. (2007). Heteroplasmy in hair: differences among hair and blood from the same individuals are still a matter of debate. Forensic Sci Int, 173(2-3), 117-121. doi:10.1016/j.forsciint.2007.02.011

Parson, W., & Dur, A. (2007). EMPOP--a forensic mtDNA database. Forensic Sci Int Genet, 1(2), 88-92. doi:10.1016/j.fsigen.2007.01.018

Parson, W., Gusmao, L., Hares, D. R., Irwin, J. A., Mayr, W. R., Morling, N., . . . Genetics, D. N. A. C. o. t. I. S. f. F. (2014). DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. Forensic Sci Int Genet, 13, 134-142. doi:10.1016/j.fsigen.2014.07.010

Parson, W., Huber, G., Moreno, L., Madel, M. B., Brandhagen, M. D., Nagl, S., . . . Irwin, J. A. (2015). Massively parallel sequencing of complete mitochondrial genomes from hair shaft samples. Forensic Sci Int Genet, 15, 8-15. doi:10.1016/j.fsigen.2014.11.009

Parson, W., Strobl, C., Huber, G., Zimmermann, B., Gomes, S. M., Souto, L., . . . Irwin, J. (2013a). Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). Forensic Sci Int Genet, 7(5), 543-549. doi:10.1016/j.fsigen.2013.06.003

Parson, W., Strobl, C., Huber, G., Zimmermann, B., Gomes, S. M., Souto, L., . . . Irwin, J. (2013b). Reprint of: Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). Forensic Sci Int Genet, 7(6), 632-639. doi:10.1016/j.fsigen.2013.09.007

Parsons, T. J., & Coble, M. D. (2001). Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome. Croat Med J, 42(3), 304-309. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11387644

Pereira, R., Phillips, C., Alves, C., Amorim, A., Carracedo, A., & Gusmao, L. (2009). A new multiplex for human identification using insertion/deletion polymorphisms. Electrophoresis, 30(21), 3682-3690. doi:10.1002/elps.200900274

Phillips, C., Parson, W., Lundsberg, B., Santos, C., Freire-Aradas, A., Torres, M., . . . Lareu, M. V. (2014). Building a forensic ancestry panel from the ground up: The EUROFORGEN Global AIM-SNP set. Forensic Sci Int Genet, 11, 13-25. doi:10.1016/j.fsigen.2014.02.012

Pischedda, S., Barral-Arca, R., Gomez-Carballa, A., Pardo-Seco, J., Catelli, M. L., Alvarez-Iglesias, V., . . . Salas, A. (2017). Phylogeographic and genome-wide investigations of Vietnam ethnic groups reveal signatures of complex historical demographic movements. Sci Rep, 7(1), 12630. doi:10.1038/s41598-017-12813-6

Pliss, L., Tambets, K., Loogväli, E. L., Pronina, N., Lazdins, M., Krumina, A., . . . Villems, R. (2006). Mitochondrial DNA portrait of Latvians: towards the understanding of the genetic structure of Baltic-speaking populations. Ann Hum Genet, 70(Pt 4), 439-458. doi:10.1111/j.1469-1809.2005.00238.x

Prinz, M., & Sansone, M. (2001). Y chromosome-specific short tandem repeats in forensic casework. Croat Med J, 42(3), 288-291. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11387641

Reidla, M., Kivisild, T., Metspalu, E., Kaldma, K., Tambets, K., Tolk, H. V., . . . Villems, R. (2003). Origin and diffusion of mtDNA haplogroup X. Am J Hum Genet, 73(5), 1178-1190. doi:10.1086/379380

Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., . . . Bandelt, H. J. (2000). Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet, 67(5), 1251-1276. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11032788

Richards, M., Macaulay, V., Torroni, A., & Bandelt, H. J. (2002). In search of geographical patterns in European mitochondrial DNA. Am J Hum Genet, 71(5), 1168-1174. doi:10.1086/342930

Riman, S., Kiesler, K. M., Borsuk, L. A., & Vallone, P. M. (2017). Characterization of NIST human mitochondrial DNA SRM-2392 and SRM-2392-I standard reference materials by next generation sequencing. Forensic Sci Int Genet, 29, 181-192. doi:10.1016/j.fsigen.2017.04.005

Ring, J. D., Sturk-Andreaggi, K., Alyse Peck, M., & Marshall, C. (2018). Bioinformatic removal of NUMT-associated variants in mitotiling next-generation sequencing data from whole blood samples. Electrophoresis, 39(21), 2785-2797. doi:10.1002/elps.201800135

Robin, E. D., & Wong, R. (1988). Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. J Cell Physiol, 136(3), 507-513. doi:10.1002/jcp.1041360316

Rohl, A., Brinkmann, B., Forster, L., & Forster, P. (2001). An annotated mtDNA database. Int J Legal Med, 115(1), 29-39. doi:10.1007/s004140100217

Roth, C., Parson, W., Strobl, C., Lagacé, R., & Short, M. (2019). MVC: an integrated mitochondrial variant caller for forensics. Australian Journal of Forensic Sciences (51:sup1), S52-S55.

Saag, L., Laneman, M., Varul, L., Malve, M., Valk, H., Razzak, M. A., . . . Tambets, K. (2019). The Arrival of Siberian Ancestry Connecting the Eastern Baltic to Uralic Speakers further East. Curr Biol, 29(10), 1701-1711 e1716. doi:10.1016/j.cub.2019.04.026

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A, 74(12), 5463-5467. doi:10.1073/pnas.74.12.5463

Santibanez-Koref, M., Griffin, H., Turnbull, D. M., Chinnery, P. F., Herbert, M., & Hudson, G. (2019). Assessing mitochondrial heteroplasmy using next generation sequencing: A note of caution. Mitochondrion, 46, 302-306. doi:10.1016/j.mito.2018.08.003

Thermo Fisher Scientific. (2016). Precision ID panels with the Ion PGM™ System Application Guide. Revision A. Thermo Fisher Scientific, Waltham (MAN0015830). Retrieved from http://tools.thermofisher.com/content/sfs/manuals/MAN0015830_PrecisionID_Panels_IonPGM_UG.pdf

Seo, S. B., Zeng, X., King, J. L., Larue, B. L., Assidi, M., Al-Qahtani, M. H., . . . Budowle, B. (2015). Underlying Data for Sequencing the Mitochondrial Genome with the Massively Parallel Sequencing Platform Ion Torrent PGM. BMC Genomics, 16 Suppl 1, S4. doi:10.1186/1471-2164-16-S1-S4

Shay, J. W., Pierce, D. J., & Werbin, H. (1990). Mitochondrial DNA copy number is proportional to total cell DNA under a variety of growth conditions. J Biol Chem, 265(25), 14802-14807. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/2394698

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. Nat Biotechnol, 26(10), 1135-1145. doi:10.1038/nbt1486

Sidstedt, M., Junker, K., Forsberg, C., Boiso, L., Rådström, P., Ansell, R., & Hedman, J. (2019). In-house validation of MPS-based methods in a forensic laboratory. Forensic Science International: Genetics Supplement Series. 7(1), 635-636. Retrieved from https://www.sciencedirect.com/science/article/pii/S1875176819302355?via%3Dihub

Skuratovskaia, D., Zatolokin, P., Vulf, M., Mazunin, I., & Litvinova, L. (2019). Interrelation of chemerin and TNF-alpha with mtDNA copy number in adipose tissues and blood cells in obese patients with and without type 2 diabetes. BMC Med Genomics, 12(Suppl 2), 40. doi:10.1186/s12920-019-0485-8

Smart, U., Budowle, B., Ambers, A., Soares Moura-Neto, R., Silva, R., & Woerner, A. E. (2019). A novel phylogenetic approach for de novo discovery of putative nuclear mitochondrial (pNumt) haplotypes. Forensic Sci Int Genet, 43, 102146. doi:10.1016/j.fsigen.2019.102146

Stoljarova, M., King, J. L., Takahashi, M., Aaspollu, A., & Budowle, B. (2016). Whole mitochondrial genome genetic diversity in an Estonian population sample. Int J Legal Med, 130(1), 67-71. doi:10.1007/s00414-015-1249-4

Stoneking, M., Hedgecock, D., Higuchi, R. G., Vigilant, L., & Erlich, H. A. (1991). Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. Am J Hum Genet, 48(2), 370-382. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/1990843

Strobl, C., Churchill Cihlar, J., Lagace, R., Wootton, S., Roth, C., Huber, N., . . . Parson, W. (2019). Evaluation of mitogenome sequence concordance, heteroplasmy detection, and haplogrouping in a worldwide lineage study using the Precision ID mtDNA Whole Genome Panel. Forensic Sci Int Genet, 42, 244-251. doi:10.1016/j.fsigen.2019.07.013

Sturk-Andreaggi, K., Parson, W., Allen, M., & Marshall, C. (2020). Impact of the sequencing method on the detection and interpretation of mitochondrial DNA length heteroplasmy. Forensic Sci Int Genet, 44, 102205. doi:10.1016/j.fsigen.2019.102205

Szklarczyk, R., Nooteboom, M., & Osiewacz, H. D. (2014). Control of mitochondrial integrity in ageing and disease. Philos Trans R Soc Lond B Biol Sci, 369(1646), 20130439. doi:10.1098/rstb.2013.0439

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 123(3), 585-595. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/2513255

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol, 30(12), 2725-2729. doi:10.1093/molbev/mst197

Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform, 14(2), 178-192. doi:10.1093/bib/bbs017

Trindade-Filho, A., Ferreira, S., & Oliveira, S. F. (2013). Impact of a chromosome X STR Decaplex in deficiency paternity cases. Genet Mol Biol, 36(4), 507-510. doi:10.1590/S1415-47572013000400008

Council of the European Union (2001). Council Resolution of 25 June 2001 on the exchange of DNA analysis results. Official Journal of the European Communities. Retrieved from https://op.europa.eu/en/publication-detail/-/publication/07d47552-09e7-44c1-9055-5c1cde83c29c

Council of the European Union (2009). Council Resolution of 30 November 2009 on the exchange of DNA analysis results. Official Journal of the European Union. Retrieved from https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri= OJ:C:2009:296:0001:0003:EN:PDF

van Oven, M., & Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat, 30(2), E386-394. doi:10.1002/humu.20921

Verogen. (2018). ForenSeq™ DNA Signature Prep Kit. Achieve high resolution and exceptional accuracy, even with complex mixtures, or degraded DNA. Data Sheet. Data Sheet: ForenSeq DNA Signature Prep Kit (VD2018002). Retrieved from https://verogen.com/wp-content/uploads/2018/07/ForenSeq-prep-kit-data-sheet-VD2018002.pdf

Vohr, S. H., Gordon, R., Eizenga, J. M., Erlich, H. A., Calloway, C. D., & Green, R. E. (2017). A phylogenetic approach for haplotype analysis of sequence data from complex mitochondrial mixtures. Forensic Sci Int Genet, 30, 93-105. doi:10.1016/j.fsigen.2017.05.007

Wachsmuth, M., Hubner, A., Li, M., Madea, B., & Stoneking, M. (2016). Age-Related and Heteroplasmy-Related Variation in Human mtDNA Copy Number. PLoS Genet, 12(3), e1005939. doi:10.1371/journal.pgen.1005939

Wallace, D. C. (2015). Mitochondrial DNA variation in human radiation and disease. Cell, 163(1), 33-38. doi:10.1016/j.cell.2015.08.067

Wallace, D. C., Brown, M. D., & Lott, M. T. (1999). Mitochondrial DNA variation in human evolution and disease. Gene, 238(1), 211-230. doi:10.1016/s0378-1119(99)00295-4

Walsh, S., Chaitanya, L., Clarisse, L., Wirken, L., Draus-Barini, J., Kovatsi, L., . . . Kayser, M. (2014). Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage. Forensic Sci Int Genet, 9, 150-161. doi:10.1016/j.fsigen.2013.12.006

Walsh, S., Lindenbergh, A., Zuniga, S. B., Sijen, T., de Knijff, P., Kayser, M., & Ballantyne, K. N. (2011). Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence. Forensic Sci Int Genet, 5(5), 464-471. doi:10.1016/j.fsigen.2010.09.008

Walsh, S., Liu, F., Ballantyne, K. N., van Oven, M., Lao, O., & Kayser, M. (2011). IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Sci Int Genet, 5(3), 170-180. doi:10.1016/j.fsigen.2010.02.004

Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., . . . Kayser, M. (2013). The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Sci Int Genet, 7(1), 98-115. doi:10.1016/j.fsigen.2012.07.005

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C., & Marth, G. (2002). Human diallelic insertion/deletion polymorphisms. Am J Hum Genet, 71(4), 854-862. doi:10.1086/342727

Weissensteiner, H., Pacher, D., Kloss-Brandstatter, A., Forer, L., Specht, G., Bandelt, H. J., . . . Schonherr, S. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res, 44(W1), W58-63. doi:10.1093/nar/gkw233

Wilson, M. R., DiZinno, J. A., Polanskey, D., Replogle, J., & Budowle, B. (1995). Validation of mitochondrial DNA sequencing for forensic casework analysis. Int J Legal Med, 108(2), 68-74. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8547161

Woerner, A. E., Ambers, A., Wendt, F. R., King, J. L., Moura-Neto, R. S., Silva, R., & Budowle, B. (2018). Evaluation of the precision ID mtDNA whole genome panel on two massively parallel sequencing systems. Forensic Sci Int Genet, 36, 213-224. doi:10.1016/j.fsigen.2018.07.015

Wood, M. R., Sturk-Andreaggi, K., Ring, J. D., Huber, N., Bodner, M., Crawford, M. H., . . . Marshall, C. (2019). Resolving mitochondrial haplogroups B2 and B4 with next-generation mitogenome sequencing to distinguish Native American from Asian haplotypes. Forensic Sci Int Genet, 43, 102143. doi:10.1016/j.fsigen.2019.102143

Xavier, C., & Parson, W. (2017). Evaluation of the Illumina ForenSeq DNA Signature Prep Kit - MPS forensic application for the MiSeq FGx benchtop sequencer. Forensic Sci Int Genet, 28, 188-194. doi:10.1016/j.fsigen.2017.02.018

Xia, C. Y., Liu, Y., Yang, H. R., Yang, H. Y., Liu, J. X., Ma, Y. N., & Qi, Y. (2017). Reference Intervals of Mitochondrial DNA Copy Number in Peripheral Blood for Chinese Minors and Adults. Chin Med J (Engl), 130(20), 2435-2440. doi:10.4103/0366-6999.216395

Xu, M., Du, Q., Ma, G., Chen, Z., Liu, Q., Fu, L., . . . Li, S. (2019). Utility of ForenSeq DNA Signature Prep Kit in the research of pairwise 2nd-degree kinship identification. Int J Legal Med. doi:10.1007/s00414-019-02003-6

Yao, L., Xu, Z., & Wan, L. (2019). Whole Mitochondrial DNA Sequencing Analysis in 47 Han Populations in Southwest China. Med Sci Monit, 25, 6482-6490. doi:10.12659/MSM.916275

Zander, J., Otremba, P., & Nagy, M. (2018). Validation of haplotype-specific extraction for separating a mitochondrial DNA model mixture and application to simulated casework. Forensic Sci Int Genet, 35, 57-64. doi:10.1016/j.fsigen.2018.04.005

Zhang, S., Tong, A. L., Zhang, Y., Nie, M., Li, Y. X., & Wang, H. (2009). Heteroplasmy level of the mitochondrial tRNaLeu(UUR) A3243G mutation in a Chinese family is positively associated with earlier age-of-onset and increasing severity of diabetes. Chin Med Sci J, 24(1), 20-25. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19382419

Zidkova, A., Horinek, A., Kebrdlova, V., & Korabecna, M. (2013). Application of the new insertion-deletion polymorphism kit for forensic identification and parentage testing on the Czech population. Int J Legal Med, 127(1), 7-10. doi:10.1007/s00414-011-0649-3

Zimmermann, B., Rock, A., Huber, G., Kramer, T., Schneider, P. M., & Parson, W. (2011). Application of a west Eurasian-specific filter for quasi-median network analysis: Sharpening the blade for mtDNA error detection. Forensic Sci Int Genet, 5(2), 133-137. doi:10.1016/j.fsigen.2010.10.003

Zinovkina, L. A. (2018). Mechanisms of Mitochondrial DNA Repair in Mammals. Biochemistry (Mosc), 83(3), 233-249. doi:10.1134/S0006297918030045

Acknowledgements

I would like to thank Dr. Anu Aaspõllu for introducing me to the fascinating world of forensic genetics and supervising my scientific work throughout my BSc, MSc and PhD studies.

I would like to thank Dr. Bruce Budowle for accepting me for a year-long professional internship at University of North Texas Health Science Center (UNTHSC) and making me a part of his extraordinary team. During time in UNTHSC I sequenced my first full mtGenome, presented a poster at my first scientific conference and started writing my first first-author paper. I was also very close to attending (as an observer invited by Dr. Budowle) my first court case involving genetic evidence. Sadly the case was dismissed.

I would like to thank Dr. Jonathan King and Dr. Jennifer Churchill Cihlar for starting as my colleagues at UNTHSC and ending up as friends. I would like to thank them for all the mentoring they have given me on mtDNA and MPS. This defence would not be possible without them. However, I also should thank them for the Texas culinary and cultural (Texas Rodeo) experiences they have shared with me.

In addition, I would like to say thank you to the following people I worked with at UNTHSC: Dr. Pam Marshall for teaching me how to saw bones; Dr. Bobby La Rue for teaching me MiSeq; and Dr. Antti Sajantila, Dr. Seung Bum Seo, Dr. Xiangpei Zeng, Dr. Caleb Guedes and others for simply being amazing colleagues and friends.

I would like to thank Dr. Erkki Truve who was a father-figure to me and many other students at the Institute of Gene Technology in Tallinn. He fuelled our passion for science and encouraged us to continuously pursue our studies to the end. I am grateful for studying under his watch during my BSc and MSc programs. Dr. Truve passed away a couple of months before the defence of this thesis. I am sad knowing that he will not be attending the defence.

In addition I would like to thank the Baltic-American Freedom Foundation for sponsoring my professional internship at UNTHSC. The scientific work conducted during that year resulted in the successful defence of my Master's thesis, a number of scientific publications and this PhD dissertation.

This work has also been partially supported by "TUT Institutional Development Program for 2016-2022 Graduate School in Biomedicine and Biotechnology" receiving funding from the European Regional Development Fund under program ASTRA 2014-2020.4.01.16-0032 in Estonia. I am grateful to the fund for giving me the possibility to attend international courses and scientific conferences which enriched my ongoing studies.

Lastly, I would like to thank my family and friends who made this journey possible. I would like to specially thank my mother for always enabling me to put my studies first.

Abstract
Massively parallel sequencing of human mitochondrial genome for forensic analysis

Mitochondrial DNA (mtDNA) analysis is performed in forensic identification in case of kinship analysis as well as samples with low level of DNA. The gold standard for forensic mtDNA analysis is the Sanger-type sequencing (STS) of hypervariable region I and II (HVI and HVII, respectively). Sequencing beyond these regions is rarely attempted as the methodology is laborious, time consuming and expensive. However it has been shown that entire mitochondrial genome (mtGenome) has a higher discrimination power compared to HVI/HVII regions.

Due to high sequencing throughput massively parallel sequencing (MPS) is considered an alternative to STS. The ability to sequence entire mtGenome, detect heteroplasmic positions as well as give additional data such as quantitative and phasing information, MPS can be used for interpretation of mixture samples, that are one of the most challenging samples in forensic investigation to analyse.

Therefore the aim of this study was firstly, to evaluate the feasibility of generating forensic quality mtGenome data with MPS technology. Secondly, to compare the discrimination power of full mtGenome to HVI/HVII data in different population samples. Thirdly, to analyse mixture samples using MPS generated quantitative data, phasing information and phylogenetic assignment.

The throughput of MPS enabled to sequence entire mtGenome of 283 US population samples as well as 114 Estonian population samples. Point and length heteroplasmy positions were typed. Over 70% of mtDNA variants were observed outside of routinely analysed HVI/HVII regions. The discrimination power of full mtGenome data was significantly higher compared to HVI/HVII data – 27.1% increase for the Estonian population, 15.6% for the US Hispanic population, 11.8% for the US African American population and 7.8% for the US Caucasian population sample. A higher increase in unique haplotypes when mtGenome data was compared to HVI/HVII was observed in samples belonging to haplogroups that have low amount of variants in HVI/HVII regions (e.g. haplogroup H). It was shown that phylogenetic assignment based solely on HVI/HVII data might lead to a falsely assigned haplogroup.

To our knowledge, the 114 Estonian full mtGenomes generated in the current work are the first and the only Estonian entire mtGenomes published. Twelve major haplogroup clades were observed in the Estonian population mtGenome data. The most frequency haplogroups were H (with the frequency of 47.4%), U (21.1%), T (9.6%) and J (6.1%). Other haplogroups observed, haplogroups D, I, K, M, N, R, W and X, included less than 5.0% of the samples. All major haplogroup clades observed in the current study were reported in the Estonian and/or geographically nearby populations by previous studies based on partial mtDNA sequences.

The interpretation of two- and three-person mixture samples from distinct (HV and F1a1a) as well as similar (U2e1a1 and U2e2a1a) haplogroups by using MPS generated quantitative data, phasing information and phylogenetic assignment was performed. The DNA profile of a major contributor was successfully typed in the mixture samples with the ratio of 1:1, 5:1, 10:1 and 20:1, while the DNA profile of a minor contributor was successfully typed in the mixture samples with the ratio of 1:1, 5:1 and 10:1. The interpretation of three-person mixture was more complicated compared to two-person mixture analysis, as phasing information gave less value as it resulted in a

high number of possible DNA profile combinations. Nevertheless the correct mtDNA profile for the major contributor was generated from the 5:1:1 mixture. In addition, it was shown that the number of variants detected in the mixture sample could indicate the number of contributors.

These results show that MPS outperformed STS in generating data that can be used for mtDNA mixture sample interpretation.

Lühikokkuvõte
Inimese mitokondriaalse genoomi massiivselt paralleelne sekveneerimine forensiliseks analüüsiks

DNA analüüs on lahutamatu osa tänapäevasest kohtuekspertiisist ehk isikute tuvastamisest. Tavapäraselt kasutatakse selleks lühikesi kordusjärjestusi (STR, short tandem repeat) eelkõige tingituna nende lühikesest pikkusest, mis suurendab geneetilise analüüsi tulemuslikkust ja proovide eristusjõudu. Lisaks on muuhulgas loodud STR markerite-põhised andmebaasid riikidevaheliseks andmevahetuseks kuritegude lahendamiseks. Euroopas on kokku lepitud 12 STR lookusel põhineva ESS (European Standard Set) süsteemi rakendamine ja Ameerika Ühendriikides 20 STR lookusel põhinev CODIS (Combined DNA Index System) süsteem.

Täiendavalt STR markerite analüüsile viiakse läbi mitokondriaalse DNA (mtDNA) analüüsid, kui on vaja määrata sugulust ja/või uurida madala DNA sisalduse või degradeerunud proove, nagu juuresibulata karvad ja vanad säilmed. Rekombinatsiooni puudumine ja pärandumine rangelt emaliini pidi teevad mtDNA sobilikuks suguluse määramise markeriks. Madala DNA sisalduse ja/või kvaliteediga proovide puhul kasutatakse mtDNA analüüsi eelkõige selle tõttu, et mtDNA koopiaarv rakus on keskmiselt 500, võrreldes kahe tuumse DNA koopiaga. Vastavalt sellele on mtDNA analüüs nende proovide korral tulemuslikum võrreldes tuumsete STR markeritega.

Mitokondriaalse DNA analüüsi alusel jagatakse indiviidid ühe nukleotiidi polümorfismide (SNP, single nucleotide polymorphism) alusel genealoogilisteks gruppideks (haplogruppideks), kellel on ühine esivanem. Populatsiooniuuringute põhjal on toodud välja geograafilistele piirkondadele omased konkreetsed haplogrupid – nt Euroopa kõige sagedasem haplogrupp on H, samas kui aafriklasi iseloomustab haplogrupp L. mtDNA populatsiooniuuringud on olulised andmebaaside loomisel ja haplotüüpide (mtDNA profiilide) sageduse määramisel.

Segaproovid, kus bioloogiline materjal pärineb rohkem kui ühelt isikult (doonorilt), on ühed raskemini interpreteeritavad proovid. Segaproovide analüüsimine kasutades STR markereid on problemaatiline eelkõige alleelide kattuvuse ning tingituna PCR protsessile omasest produktide lisandumisest (*drop-in*) ja väljalangevusest (*drop-out*), mis raskendab doonorite arvu ja profiilide omistamist konkreetsele doonorile. Segaproovide mtDNA järjestuste määramist on pakutud üheks alternatiiviks STR analüüsile.

Tüüpiliselt määratakse kohtugeneetikas mtDNA profiilid kahe hüpervarieeruva piirkonna (vastavalt HVI ja HVII) Sanger-tüüpi sekveneerimisega (STS). HVI/HVII piirkondi, mis moodustavad ligikaudu 3.7% mitokondriaalsest genoomist (mtGenoomist), kasutatakse nendesse koondunud suure arvu SNP positsioonide tõttu. mtGenoomi täissekveneerimist teostatakse harva, kuna see on töömahukas ja kulukas. Sellegipoolest on publitseeritud andmeid, mis näitavad, et mtDNA täisgenoomi analüüsil on HVI/HVII piirkondadega võrreldes suurem eristusjõud.

Massiivselt paralleelne sekveneerimine (MPS) on alternatiiv STS meetodile, võimaldades analüüsida suuremahuliselt samaaegselt hulgaliselt DNA proove. Lisaks võimaldab MPS tehnoloogia tuvastada mtDNA järjestuse kui ka pikkuse polümorfisme. Seega, sekveneerides proovi mtGenoomi MPS tehnoloogiaga võib segaproovide analüüsi tõhusust tõsta.

Töö peamisteks eesmärkideks oli hinnata MPS rakendatavust forensiliste proovide mtGenoomide sekveneerimiseks. Teiseks, võrrelda erinevate populatsioonide täielike mtGenoomide ja HVI/HVII piirkondade eristusjõudu. Kolmandaks, analüüsida kahe ja kolme isiku kunstlikult tekitatud segaproove, lähtudes MPS kvantitatiivsetest andmetest, faasiandmetest ja fülogeneetilisest hindamisest.

Töö käigus sekveneeriti MPS tehnoloogiat kasutades 283 proovi kolmest Ameerika Ühendriikide populatsiooni valimist ning 114 proovi Eesti populatsiooni valimist. Näidati, et üle 70% varieeruvatest positsioonidest asuvad väljaspool HVI/HVII piirkondi ning mtGenoomi eristusjõud on statistiliselt oluliselt kõrgem HVI/HVII piirkondade omast – juhusliku kokkulangevuse tõenäosuse (RMP, random match probability) ja geneetiline mitmekesisuse (GD, genetic diversity) paarikaupa Studenti t-testi p-väärtused on vastavalt p = 0,01659 ja p = 0,01645.

Eesti populatsiooni valimis avaldus kõige suurem tõus unikaalsete haplotüüpide arvus mtGenoomi andmete võrdlemisel HVI/HVII andmetega - 27,1% tõus Eesti populatsiooni valimis, 15,6% tõus Ameerika Ühendriikide hispaaniakeelses (*Hispanic*) populatsioonis, 11,8% afroameeriklaste hulgas ja 7,8% tõus valgete inimeste (*Caucasian*) populatsiooni valimis. Selle alusel leiti, et mtDNA kodeeriva regiooni uurimine suurendab eelkõige vähese HVI/HVII varieeruvusega indiviidide, nagu isikud, kes kuuluvad H haplogruppi, eristamist. Lisaks näidati, et fülogeneetiline omistamine võib ainult HVI/HVII piirkondade andmete kasutamisel osutuda puudulikuks. Viie proovi korral 283 Ameerika Ühendriikide populatsioonide proovidest vahetus makro- ehk ülemhaplogrupp HVI/HVII andmete võrdlemisel mtGenoomi andmetega. Ühe proovi puhul vahetus HVI/HVII põhjal määratud Aasia-omane D4j1b2 haplogrupp Aafrika-omaseks L3b1a7a haplogrupiks.

Töö tulemusena publitseeriti esmakordselt eestlaste mtDNA täisgenoomid, mis jagunesid 12 ülemhaplogruppi, kusjuures kõrgema sagedusega kuulusid isikud haplogruppidesse H (sagedusega 47,4%), U (21,1%), T (9,6%) ja J (6,1%). Ülejäänud 8 ülemhaplogruppi (D, I, K, M, N, R, W ja X) tuvastati alla 5,0% proovidest.

Kasutades MPS tehnoloogia kaudu genereeritud kvantitatiivseid andmeid (ehk alternatiivsete alleelide ja positsiooni kõikide alleelide suhet), faasiandmeid (ehk SNP positsioonide paiknemist samal lugemil) ja fülogeneetilist teavet, hinnati kahe ja kolme isiku DNA segaproove. Õiged mtDNA profiilid omistati doonorile, kelle DNA kogus oli segaproovis suurem, segaproovides suhtega 1:1, 5:1, 10:1 ja 20:1. Õiged mtDNA profiilid omistati doonorile, kelle DNA kogus oli proovis väiksem, segaproovides suhtega 1:1, 5:1 ja 10:1. Kunstlikud segaproovid koosnesid kaugete (HV ja F1a1a) ja lähedaste (U2e1a1 ja U2e2a1a) haplogruppidega indiviidide proovidest.

Eeldatavalt osutus kolme indiviidi segaproovide tulemuste interpreteerimine, kahe indiviidi segaprooviga võrreldes, keerulisemaks. Indiviidide, kelle DNA kogus oli segaproovis suurim, õnnestus õige mtDNA profiil määrata segaproovis suhtel 5:1:1, ent suhtel 1:1:1 oli usaldusväärne eristamine problemaatiline.

Lisaks näidati, et segaproovis tuvastatud variatsioonide arv võimaldab hinnata segaproovi doonorite arvu.

Appendix

Publication I

King, J. L., LaRue, B. L., Novroski, N. M., **Stoljarova, M**., Seo, S. B., Zeng, X., Warshauer, D. H., Davis, C. P., Parson, W., Sajantila, A., & Budowle, B.

High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq.

Forensic Sci Int Genet, 2014 Sept;12:128-135.

Forensic Science International: Genetics 12 (2014) 128–135

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq

Jonathan L. King [a,1,*], Bobby L. LaRue [a,1], Nicole M. Novroski [a], Monika Stoljarova [a], Seung Bum Seo [a], Xiangpei Zeng [a], David H. Warshauer [a], Carey P. Davis [a], Walther Parson [b,c], Antti Sajantila [a,d], Bruce Budowle [a,e]

[a] Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA
[b] Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria
[c] Penn State Eberly College of Science, University Park, PA, USA
[d] Department of Forensic Medicine, Hjelt Institute, P.O. Box 40, 00014 University of Helsinki, Helsinki, Finland
[e] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 1 April 2014
Received in revised form 29 May 2014
Accepted 1 June 2014

Keywords:
Mitochondrial genome (mtGenome)
Haplogroup assignment
Mitochondrial haplotype
Massively parallel sequencing (MPS)
Illumina MiSeq
Discrimination power

ABSTRACT

Mitochondrial DNA typing in forensic genetics has been performed traditionally using Sanger-type sequencing. Consequently sequencing of a relatively-large target such as the mitochondrial genome (mtGenome) is laborious and time consuming. Thus, sequencing typically focuses on the control region due to its high concentration of variation. Massively parallel sequencing (MPS) has become more accessible in recent years allowing for high-throughput processing of large target areas. In this study, Nextera® XT DNA Sample Preparation Kit and the Illumina MiSeq™ were utilized to generate quality whole genome mitochondrial haplotypes from 283 individuals in a both cost-effective and rapid manner. Results showed that haplotypes can be generated at a high depth of coverage with limited strand bias. The distribution of variants across the mitochondrial genome was described and demonstrated greater variation within the coding region than the non-coding region. Haplotype and haplogroup diversity were described with respect to whole mtGenome and HVI/HVII. An overall increase in haplotype or genetic diversity and random match probability, as well as better haplogroup assignment demonstrates that MPS of the mtGenome using the Illumina MiSeq system is a viable and reliable methodology.

1. Introduction

Mitochondrial DNA (mtDNA) typing is used by various disciplines, such as medical genetics [1–3], genealogy and evolutionary anthropology [4–7]. The higher copy number of mitochondrial genomes (mtGenomes) per cell compared with the nuclear genome makes typing of mtDNA particularly useful for forensic human and species-identity testing, and ancient DNA analyses, where samples typically are of low quality and contain minute or undetectable amounts of nuclear DNA. Traditionally, mtDNA typing in forensic genetics has been performed with Sanger-type sequencing (STS) [8–11]. However, STS chemistry, while effective, is laborious, costly, and limited technically. The ~16,569 base mtGenome is not feasibly sequenced in a practical manner in an application-oriented laboratory. Thus, most forensic laboratories focus only on the control region (CR) of the mtGenome and, more specifically, hypervariable regions I and II (HVI and HVII) for database construction, database queries, and direct and indirect casework comparisons. mtDNA databases allow for haplotype searching [12–16] as well as variant-specific queries [16–18]. To date, forensic databases contain limited, if any, coding region data; however, some medical and population genetics-focused datasets contain whole mtGenome data. mtGenome data provide greater discriminatory power and allow resolution of common HVI/HVII haplotypes [19–22]. As technology progresses, the ability to type numerous samples from various populations and to interrogate more than a limited number of haplogroup-defining sites increases. An ancillary benefit to sequencing the entire mtGenome is these data can be used to generate more accurate haplogroup assignments in the form of phylogenetic trees (e.g., Phylotree) [6].

Though not routinely performed in forensic casework, haplogroup assignments allow analysts a measure of data quality control [23,24]. Haplogroup assignments can be performed

[*] Corresponding author. Tel.: +1 817 735 2940; fax: +1 817 735 5016.
E-mail address: Jonathan.King@unthsc.edu (J.L. King).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.fsigen.2014.06.001
1872-4973/© 2014 Elsevier Ireland Ltd. All rights reserved.

manually using Phylotree or with haplogroup-assignment software [13,15,25–27]. Bandelt et al. [24] recommend a semi-automated approach to haplogroup assignment that could alleviate some shortcomings of haplogroup assignment processes. Regardless, the accuracy of a haplogroup assignment still is reliant on the accuracy and amount of genetic data used (i.e., control region versus mtGenome). However, employing traditional STS for such a relatively large target as the entire mtGenome is a costly and time-consuming process which will ultimately result in low coverage data (i.e., forward and reverse; 2×).

Massively parallel sequencing (MPS), also termed Next Generation Sequencing, technologies allow for a substantial increase in throughput and depth of coverage at a relatively-affordable price. With the advent of accessible benchtop sequencers [28,29], MPS can be considered a viable technology for application-oriented and research-driven laboratories. The Nextera® XT DNA Sample Preparation Kit (Illumina®, San Diego, CA) and the MiSeq™ platform (Illumina) together enable development of a practical protocol for whole mtGenome sequencing. The aims of the study herein were: (1) to determine the throughput level when sequencing the mtGenome using the Nextera XT DNA Sample Preparation Kit and the MiSeq platform; (2) to interrogate the poly-cysteine (C) stretches for interpretation related to length homopolymers; (3) to compare the random match probabilities (RMPs) and haplotype or genetic diversities (GDs) of mtGenome and HVI/HVII data; and (4) to determine the overall feasibility of the MPS system in generating mtGenome population data.

2. Materials and methods

2.1. Sample preparation

Whole blood samples were collected by venipuncture with lavender-top Vacutainer® tubes (Becton, Dickson and Company; Franklin, NJ, USA) from a total of 283 anonymized and unrelated individuals from the three U.S. populations (African American, $n = 87$; Caucasian, $n = 83$; Southwest Hispanic, $n = 113$) according to protocols approved by the University of North Texas Health Science Center's Institutional Review Board. DNA was extracted using the QIAamp® DNA Blood Mini Kit (QIAGEN, Hilden, Germany) according to the manufacturer's recommendations. The quantity of DNA was determined using the Qubit® dsDNA BR Quantification Kit and a Qubit® 2.0 Fluorometer (Life Technologies, Foster City, CA, USA). Samples were normalized to 0.1 ng/μL and stored at either 4 °C or −20 °C until mtDNA enrichment.

2.2. Target enrichment

Amplification of the mtGenome was accomplished by long PCR in two separate reactions using the TaKaRa LA PCR Kit (TaKaRa Bio; Otsu, Shiga, Japan). The primers for each reaction were described previously by Gunnarsdóttir et al. [30]. These primer pairs generated overlapping amplicons of ~8.3 and ~8.6 kb, respectively. The total template DNA was 1.0 ng per reaction. Amplification was performed on a GeneAmp® 9700 PCR System (Life Technologies) using the following thermal-cycling parameters: an initial temperature of 94 °C for 1 min; followed by 35 cycles of 98 °C for 10 s, 60 °C for 2 min, and 68 °C for 10 min. After cycling, there was a final extension step of 72 °C for 10 min. The amplified product was maintained at 4 °C until normalization. The quantity of amplified product was determined using the Qubit dsDNA BR Quantification Kit. Next, 0.2 ng/μL normalized products were pooled and 1.0 ng of DNA was used for library preparation.

2.3. Nextera XT library preparation

Libraries were prepared using the Nextera XT DNA Sample Preparation Kit according to the manufacturer's protocol [31], unless otherwise stated. Sample libraries (multiplexed as $n = 24$, 48, 76, 79, and 96) were prepared in succession to assess system throughput. Following PCR cleanup, the libraries were quantified using the Qubit dsDNA BR kit, and evaluated for fragment size using the High Sensitivity D1K ScreenTape and Tape Station 2200 (Agilent Technologies, Santa Clara, CA, USA). Following Illumina's (Illumina®, San Diego, CA, USA) technical note for cluster optimization [32] and the resultant size and quantity data, each library was normalized for sequencing to 12 pM according to the manufacturer's protocol.

2.4. MPS sequencing and data generation

Sequencing reactions were carried out using the MiSeq v2 (2 × 250 bp and 2 × 150 bp) chemistries (Illumina). The MiSeq re-sequencing protocol for small genome sequencing was followed according to the manufacturer's recommendations. Sequencing proceeded on a MiSeq platform in an automated fashion for ~39 h. On-board software (i.e., Real-TimeAnalysis and MiSeq Reporter) converted raw data to Binary Alignment/Map (BAM) and Variant Call Format (VCF) v4.1 files using Genome Analysis Toolkit (GATK) [33]. During this process, the sequenced region of interest (ROI) was aligned to the revised Cambridge Reference Sequence (rCRS) [34]. Each nucleotide position (np) was interrogated and variations from the reference were annotated by base difference (e.g., 73G). These VCF files were analyzed subsequently using mitoSAVE [35].

2.5. Data analysis

Database querying of sequence data typically is done using a haplotype defined by differences from the rCRS rather than a "string search". Thus, variant reports (VCF v4.1 files) were converted into concise haplotypes using mitoSAVE. For the purposes of this study, the following criteria were used for variant calling: a quality threshold (GATK-assigned confidence in variant call) of 70; a heteroplasmy threshold of 0.18; and a coverage threshold of 40× (all values operationally chosen for this study). As such, positions would only be interpreted if there was a minimum of 40× coverage, and point heteroplasmy at this minimum position coverage threshold would be called as long as the alternate base displayed ≥7× coverage. Haplotypes were exported from mitoSAVE in *.hsd* or *.txt* file format for upload to HaploGrep [26], a web-based haplogroup assignment software that uses Phylotree (various builds) [6] to assign haplogroup status to haplotypes. In this study, all haplotypes were assigned haplogroup status by comparison with Phylotree build 15. Each haplogroup assignment is given an algorithm-based ranking which is displayed for the user and can be changed readily prior to export. For this study, the highest ranking haplogroup was relied upon with no assumed haplogroup status prior to assignment.

Variants not known to be associated with a haplogroup (local private mutations), not previously observed in the database (global private mutations), or variants expected, but not observed, for each haplotype were verified by manually viewing BAM files in Integrative Genomics Viewer (IGV) [36]. Concordance data were generated for HVI and HVII in a subset of samples ($n = 9$) using STS according to the method described by Wilson et al. [10]. RMP and GD were calculated according to methods described by Stoneking et al. [37] and Tajima [38], respectively. Pairwise comparisons of the mtGenomes were made using MEGA 6 [39]. Lastly, Circos plots were generated using Circos version 0.64 [40].

3. Results and discussion

The Nextera XT DNA Sample Preparation Kit was selected for library preparation because it requires only 1 ng template DNA, can be performed in a relatively-short time frame, and multiple samples can be prepared simultaneously. Sample libraries (multiplexed as n = 24, 48, 76, 79, and 96) were prepared and sequenced using the Nextera XT and MiSeq v2 chemistry, respectively. Library preparation for each multiplex could be carried out easily over the course of two standard working days after target enrichment. Sequencing occurred on the MiSeq over the course of ~39 h. The time frame for data generation varied depending on the number of samples sequenced. This method is a vast improvement in time and throughput for whole mtGenome compared with STS even with an automated process [41].

3.1. mtGenome coverage

The MiSeq generates approximately 8.8 Gigabases (Gb) of data from an optimal sequencing run using the MiSeq v2 (2 × 250 bp) chemistry. Assuming equal coverage across the mtGenome, an individual sample would be expected to have over 530,000× coverage at each base position of the mtGenome; with 96 samples (barcoded with different indices), over 5500× coverage would be expected on average for each indexed sample across the target space. However, in practice, coverage was not dispersed evenly across the mtGenome. While there was consistency of coverage among individuals, coverage varied within individual mtGenomes. Most notably, the poly-C stretch in HVII and a portion (<300 bp near np 3500) of the gene encoding the NADH Dehydrogenase subunit I (ND1) were particularly low (Fig. 1 and Supplemental Fig. 1). The combination of the poly-C stretch with the alignment of a circular genome to a linear reference could explain the apparent low coverage in HVII. To assess whether the effect was due to the reference genome alignment, an alternative reference genome was created with 200 bp from the opposite end of the original reference was appended to the HVII region. This genome allowed alignment of reads overlapping np 16,569–1. An increase of ~5 fold in coverage was observed within ~50 bases. However, this alignment shifted the positions in relation to the rCRS and thus was not feasible for high-throughput studies. The region near the poly-C stretch still presented with low coverage. Despite this observation, both regions were sufficiently covered (i.e., ≥100×) in all samples. In fact, the high depth of interrogation offered by sequencing of the poly-C stretch with MPS technology allowed elucidation of length heteroplasmies previously not afforded with STS.

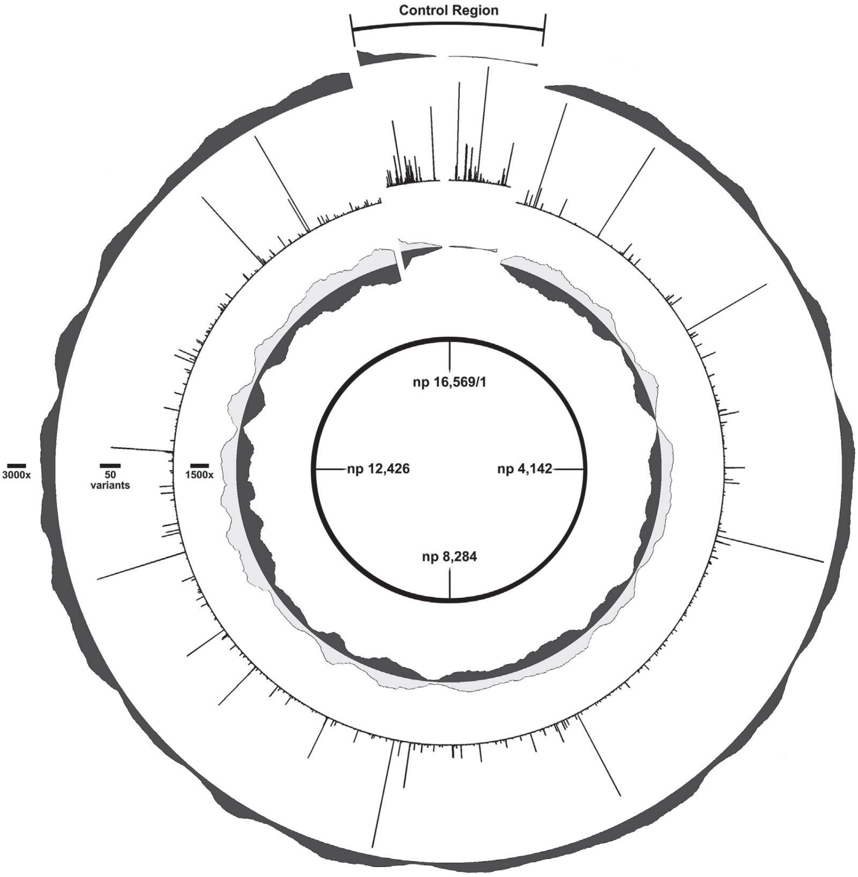


Fig. 1. A concentric Circos plot of the mtGenome representing mean coverage (outer circle; n = 24), variants observed per nucleotide position (middle circle; n = 283), and mean coverage differentiated by reverse (dark) or forward (light) strand (inner circle; n = 24). The rose diagram in the center is included for nucleotide position orientation and scale bars are included to the left of the individual plots to approximate values. The control region is offset slightly for orientation. The disproportionally-low coverage observed in HVII is likely a combination of sequencing the poly-C stretch and alignment to a linear reference.

Supplementary Fig. 1 can be found, in the online version, at doi:10.1016/j.fsigen.2014.06.001.

The first four multiplex sequencing runs contained 24, 48, 76, and 79 indexed samples and generated fully interpretable results with sufficient coverage in 223 of these 227 samples (98.2%). Initial library concentration for the four reduced coverage samples was considerably lower than samples within the same run prior to pooling (data not shown) and could explain the lack of coverage. Thus, haplotypes were considered to be accurate when variants were of sufficient coverage (i.e., ≥100×) and quality (i.e., base quality scores of ≥Q30 (Phred-style scale)) [42]. When 96 samples were indexed and sequenced, 26 samples (27.1% of the 96 indices) had coverage areas that were less than the 100× coverage for the lower number of indexed sample runs. In these samples, areas of low coverage ranged from no coverage to just below 100×. These areas also tended to have a concomitant strand bias that complicated interpretation. Of these 26, 17 provided full results (variants ≥40×); however, all 26 samples were re-sequenced at higher coverage to confirm variant calls. These reanalyzes support that the operationally-selected ≥40× minimum coverage threshold generates reproducible base calls. One possible explanation for lower overall coverage may be due to a high library concentration that could lead to a reduced number of quality clusters.

3.2. Strand bias

A plot of average coverage from both the forward and reverse strands at each base position (Fig. 1 and Supplemental Fig. 2) illustrates that few areas of the mtGenome ($n = 24$) exhibited strand bias with the protocol used herein. In fact, when observing the percentage of strand balance at all positions, there were only a small portion of positions which displayed dramatic bias (Fig. 2). In all, 16,062 nps (96.9% of all positions) had a strand balance percentage above 40%. Areas of high strand bias (≤40%) coincide with areas of low overall coverage relative to the rest of the mtGenome. In particular, positions near nps 16,569 and 1, the poly-C stretch in HVII and low areas around nps 3500 and 8600 were underperforming generally. It is not evident whether the observed strand bias is the result of library preparation, sequencing, quality filtering, and/or mapping of reads to the rCRS. The mapping of reads from a circular genome to a linear reference is inherently problematic and, most notably, resulted in a drop in coverage and strand bias in this HVII (Fig. 1). Potential strategies to overcome this bias would improve the throughput and generate higher quality haplotype data (when multiplexing) by more evenly distributing the coverage across the genome.

Supplementary Fig. 2 can be found, in the online version, at doi:10.1016/j.fsigen.2014.06.001.

3.3. Variants called

From the 283 samples sequenced, 11,607 variants, defined in relation to the rCRS, were observed (Supplemental Table 1). These variants were distributed across 1353 nucleotide positions throughout the mtGenome. Of these 1353 positions, more than one variant type was observed at 55 base positions among all samples sequenced. A total of 722, 220, and 96 of the 11,607 variants were observed in one, two and three samples, respectively, and three variants (263G, 4769G, and 15326G) were observed in all samples, which is a reflection of the reference used. The remaining variants were observed in between 4 and 282 samples with 1302 variants (approximately 92.3% of the 1411 total unique variants) being observed in 20 or less of all samples sequenced.

Supplementary Table 1 can be found, in the online version, at doi:10.1016/j.fsigen.2014.06.001.

To better illustrate the frequency of variants at each base position, a “heat map” was generated (Fig. 1 and Supplemental Fig. 3) that plots mtGenome position versus the number of variants observed at each variant position. As would be predicted, polymorphism density was clustered heavily in the HVI/HVII. Out of all observed variants, 2938 of the variants (25.3%) were observed in these two regions which comprise only 3.7% of the

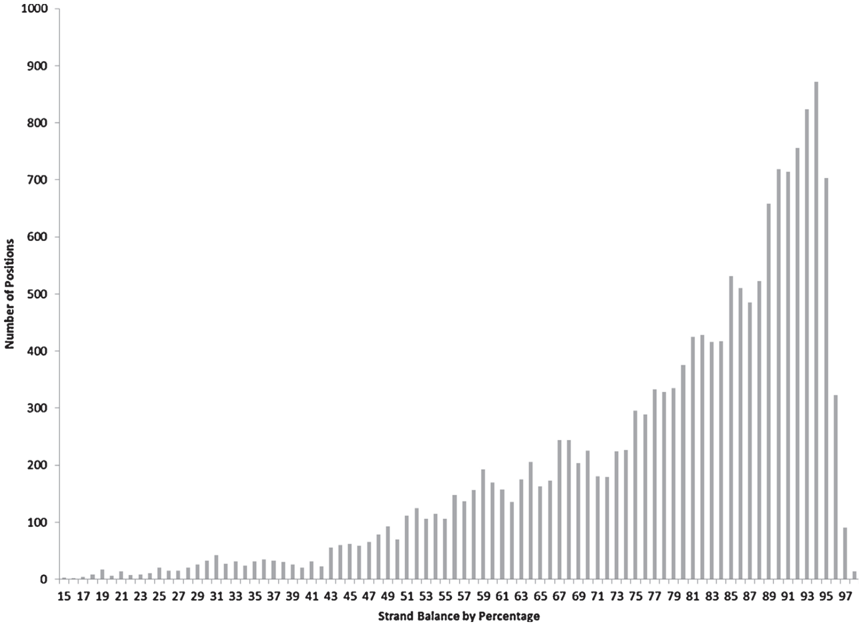


Fig. 2. Strand bias histogram displaying the distribution of strand balance across all nucleotide positions of the mtGenome for an arbitrary subset of samples ($n = 24$).

Table 1
A comparison of unique haplogroups[a] and haplotypes generated by HVI/HVII versus mtGenome versus coding region sequence data in three major US populations (n = 283).

	Breakdown by population								
	HVI/HVII			mtGenome			Coding region		
	AFA	CAU	HIS	AFA	CAU	HIS	AFA	CAU	HIS
Number of individuals	87	83	113	87	83	113	87	83	113
Unique haplogroups	55	70	56	70	79	70	70	74	73
Unique haplotypes	76	77	96	85	83	111	85	83	111

[a] As assigned by HaploGrep [6] and Phylotree [26].

mtGenome. Thus, 8669 of the variants (74.7% of all variants observed) resided outside of the HVI/HVII regions. The distribution of variants is inflated artificially, however, by high frequency variants such as those mentioned earlier. A total of 15 variants, 4 of which reside in HVI/HVII, appeared in more than half the samples and account for 3638 (31.3%) of all variants observed. These high frequency reference-alignment artifacts are unavoidable and do not change the observed distribution of variants. This distribution illustrated the untapped potential of the coding region for discriminatory power and more effective haplogroup assignment.

Supplementary Fig. 3 can be found, in the online version, at doi:10.1016/j.fsigen.2014.06.001.

3.4. Haplogroup and haplotype assessment

As part of the quality assessment of haplotype data, haplogroup assignment was performed to discern established haplogroup specific mutations from yet undescribed "private" mutations in the respective haplogroup backgrounds. The latter were subject to additional quality checks to confirm their authenticity. The online software tool HaploGrep was used to assign the haplogroup status to the observed haplotypes. From all 283 individual mtGenomes, 14 different clades were represented (Supplemental Table 2) with 208 distinct haplogroups and 279 unique haplotypes. By population, there were 70, 79, and 70 distinct haplogroups and 85, 83, and 111 distinct haplotypes for African Americans, Caucasians, and Southwest Hispanics, respectively (Table 1).

Supplementary Table 2 can be found, in the online version, at doi:10.1016/j.fsigen.2014.06.001.

The task of generating concordance mtGenome data using STS for such a large dataset is a time-consuming, arduous task which is impractical. However, previously-generated STS data for HVI and HVII were available for a subset of samples (n = 8). All MPS data were concordant at all positions with STS (data not shown). These data included point and length heteroplasmy, the latter of which previously was difficult to interpret given the nature of STS (and not considered for concordance).

To examine diversity based on individual differences in haplotypes among the population samples, pairwise comparisons of base differences in a string format were performed both within, between/among population groups. A bimodal distribution was observed in both Caucasian and Hispanic pairwise comparisons and a trimodal distribution was observed for African Americans (Supplemental Fig. 4). These modes are consistent with the gross phylogeny of the populations. The mean pairwise differences was highest in African Americans (55 ± 22) and lowest in Caucasians (30 ± 11) (Table 2). These results were consistent with previously reported HVI/HVII data [43].

Supplementary Fig. 4 can be found, in the online version, at doi:10.1016/j.fsigen.2014.06.001.

All private mutations (both global and local) denoted by HaploGrep were analyzed further in IGV and these variants were confirmed. However, 7 haplogroup-defining variants (T16189C, n = 6; G16129C, n = 1) were not included in the original VCF files. Reanalysis of the samples with the standard workflow did not resolve these "missed calls". When the original FASTQ files were analyzed with an alternative alignment workflow (a beta version of mtDNA Variant Analyzer 1.0, Illumina), the "missing" variants were assigned correctly. This result was consistent with previous observations regarding variations between aligners [44]. While mtGenome haplogroup assignments were unaffected by these missed calls, when HVI/HVII haplotypes were analyzed with HaploGrep, the haplogroup assignment of one sample (USA_TX_0102) diverged greatly from the full HVI/HVII haplogroup. When the T16189C variant was included in the haplotype, the sample was assigned to the haplogroup B4; however, using the VCF file from GATK alone, the sample was assigned to the haplogroup HV2. This shift in haplogroup assignment was observed also when assessing mtGenome and HVI/HVII haplotypes (Supplemental Table 3).

Supplementary Table 3 can be found, in the online version, at doi:10.1016/j.fsigen.2014.06.001.

Nine samples (3.2% of all samples sequenced) changed clades (e.g., G → L) between the limited HVI/HVII data and that of the mtGenome data (Table 3). In fact, six of nine samples changed macrohaplogroups (i.e., L, M, or N). Further analysis of the HVI/HVII haplogroup assignments indicated a variation in top-ranked haplogroups independent of stated quality. Quality scores for these nine samples ranged from 80.3 to 95.9 using HVI/HVII data. In fact, four of nine samples had quality scores greater than 93.0 (rank equivalent 0.930). These observations can be explained by the fact that HaploGrep's assignment is based on signature mutations indicated on the branches of Phylotree only, while other mutations present in the corresponding mtGenomes were not taken into consideration. The ranked haplogroups displayed by HaploGrep for these samples varied widely. Sample USA_TX_0257, for example, listed the following haplogroups as the top three possible haplogroups: P5 (rank-0.959), U5b2a1a (rank-0.914), H32 (rank-0.886). Conversely, the mtGenome haplotype analyzed with HaploGrep listed H2 + 152 (rank-0.925), H (rank-0.917), and H32 (rank-0.916) as the top three possible haplogroups. This observation demonstrates that haplogroup assignment can lead to even

Table 2
Haplotype diversity as measured by pairwise comparisons of individual consensus sequences[a] both within and among three major US populations.

	AFA	CAU	HIS	AFA/CAU	AFA/HIS	CAU/HIS	AFA/CAU/HIS
Number of pairwise differences (Mean ± SD)	55 ± 22	30 ± 11	36 ± 15	47 ± 20	47 ± 19	35 ± 12	43 ± 18
Range of differences	0–104	1–55	0–90	0–102	0–102	0–91	0–101

[a] As described in [43].

Table 3
Samples in which the clade[a] was reassigned based on mtGenome versus HVI/HVII sequence data.

	HVI/HVII		mtGenome	
Sample ID[b]	Haplogroup assignment	Quality (%)	Haplogroup assignment	Quality (%)
USA_TX_0028	N11a	80.3	L2a1c3	93.1
USA_TX_0052	M73'79	95.1	L3b1a + !16124	95.1
USA_TX_0057	G3	89.3	L3b1a7	97.2
USA_TX_0063	N2	95.2	L3e1f	95.1
USA_TX_0108	HV0	93.9	V2	95.4
USA_TX_0132	R0 + 16189	87.7	H4a1a1a1a1	97.8
USA_TX_0174	M33c	83.6	A2 + 64	90.3
USA_TX_0175	D4e1	82.8	A2 + 64	91.8
USA_TX_0257	P5	95.9	H32	92.5

[a] As assigned by HaploGrep [6] and Phylotree [26].
[b] As labeled in EMPOP [12].

Table 4
Comparison of population specific and mean random match probabilities (RMP)[a] and genetic diversities (GD)[b] in three major US populations with HVI/HVII versus mtGenome sequence data.

		HVI/HVII		mtGenome	
Populations	*n*	RMP	GD	RMP	GD
AFA	87	2.42%	98.72%	1.31%	99.84%
CAU	83	3.12%	98.06%	1.20%	100.00%
HIS	113	3.33%	97.53%	0.98%	99.91%
Mean ± SD		2.96 ± 0.48%	98.10 ± 0.59%	1.16[c] ± 0.17%	99.91[d] ± 0.08%

[a] As described in [37].
[b] As described in [38].
[c] Significantly different than the RMP obtained from the HV1/HVII alone ($p = 0.00358$; paired two-tailed Student's *T*-test).
[d] Significantly different than the GD obtained from the HV/HVII alone ($p = 0.00631$; paired two-tailed Student's *T*-test).

highly ranked ambiguous results when it is reduced to signature mutations as reported previously [24,25].

Bandelt et al. [45] reported shortcomings with HaploGrep particularly regarding HVI/HVII haplotypes. HaploGrep is reliant solely on the mutations listed on the branches of Phylotree for haplogroup assignment which limits its use for the assignment of those haplogroups that are defined by few or only one mutation. Its application in forensic genetics, where the control region or segments thereof are still the main target(s), can be particularly misleading in lineages that are not represented by control region polymorphisms, in Phylotree, such as a large portion of haplogroups nested in H that are only defined by coding region data. In those instances, we observed a general tendency of HaploGrep to be biased toward the rCRS or close neighbors thereof. Additionally, the fact that some polymorphisms are not considered systematically for haplogroup assignment in Phylotree (e.g., np 16,519, "AC" insertion/deletions (INDELs) between np 515 and 524, and insertions in the poly-C regions of HVI/HVII between np 16,183 and 16,194 and between np 302 and 316, respectively) explains, why the haplogroup resolution may be low in some lineages [20,46].

3.5. Forensic parameters

The haplotype and haplogroup diversity of HVI/HVII were compared with that of the mtGenome. The variant calls from HVI/HVII resulted in 38 fewer unique haplogroups (55, 70, 56 for African Americans, Caucasians, and Southwest Hispanics, respectively), and 30 fewer unique haplotypes (76, 77, 96 for African Americans, Caucasians, and Southwest Hispanics, respectively) than when whole mtGenome sequence data were assessed (Table 1). In fact, the coding region alone allowed comparable resolution of haplotypes and haplogroups when compared to mtGenome data suggesting similar discrimination power with this dataset (Table 1).

Population genetics parameters of mtDNA sample sets are reliant partially on the size of the database. Generating a total of 283 mtGenome sequences in a relatively short time was impressive compared with STS. However, it was a relatively small number for assessing mean RMP and GD. The increase in RMP can be appreciated better by comparison of HVI/HVII sequences and mtGenome sequences from the same sets of individuals. The RMPs for HVI/HVII data were 2.42%, 3.12%, and 3.33% (Table 4) in African American, Caucasian, and Southwest Hispanic, respectively. In contrast, the RMPs based on mtGenome sequences were 1.31%, 1.20%, and 0.98%, respectively. This difference was significant ($p = 0.0036$; paired, two-tailed Student's *T*-test).

A similar pattern held with GD. The GD was 0.987, 0.981, and 0.975 for HVI/HVII compared to 0.998, 1.000, and 0.999 using the mtGenome data for African Americans, Caucasians and Southwest Hispanics, respectively. The increase in GD was significant ($p = 0.0063$; paired, two-tailed Student's *T*-test). As the database increases in size, it is expected that the RMP will decrease and GD will increase. Interestingly, both the RMP and GD for the coding region alone yielded equivalent RMP and GD to the mtGenome reinforcing that there is more variation residing in the coding region compared with the control region of the mtGenome. Given the ease of generating sequence data and concomitant high quality results, MPS sequencing of the entire mtGenome should be considered as a viable approach to supplement power of discrimination, when warranted.

4. Conclusion

Previous studies have described sequencing the mtGenome by MPS on other platforms [44,47], albeit with smaller sample sizes. To the best of our knowledge, this is the first report of a relatively large number of mtGenomes that has been sequenced in a high-throughput fashion using the Illumina MiSeq system. The throughput level was high due to the use of Nextera XT DNA Sample Preparation Kit and MPS. While some strand bias was observed, it generally was limited to areas of low coverage and did not diminish the ability to assign variant calls. Also, it was possible, due to a high depth of interrogation, to type length and point heteroplasmies.

By sequencing the entire mtGenome versus only HVI/HVII, the additional variant calls significantly improved the discrimination power of haplotypes. An overall improvement in the resolution of haplogroup assignments was observed compared with only the control region to the mtGenome, while haplogroup assignment was ambiguous for HVI/HVII segments of those mtGenome sequences that were not represented by control region polymorphisms in Phylotree. In this study, we demonstrated that approximately 300 individual mtGenomes could be sequenced and analyzed by a single individual in one month at an average reagent cost of approximately 50 USD per sample. Indeed, one individual can sequence 96 samples from DNA to VCF file in approximately 4.5 days. Subsequently, haplotypes can be compiled in a matter of hours using mitoSAVE. This MPS approach will facilitate generation of whole mtGenome population data to support human evolution, forensic, and medical studies. These mtGenome haplotypes will be made available to the community via EMPOP [12] under accession numbers EMP00658-EMP00660 (http://empop.org/). While this protocol applies to analysis of relatively high quality DNA, i.e., using long PCR to generate two amplicons of ~8 kb in length, the same general sequencing strategy could be applied to more challenged samples. Research is underway to design small-sized amplicons that span the mtGenome so that degraded samples potentially may be analyzed.

Acknowledgements

The authors would like to thank Jennifer Churchill for critical review and constructive comments provided. This work was supported in part by award no. 2012-DN-BXK033 and 2011-MU-MU-K402, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the U.S. Department of Justice. AS has been supported by Finnish Foundations' Professor Pool (Paulo Foundation). WP received funding from the European Union Seventh Framework Program (FP7/2007-2013) under grant agreement no. 285487 (EuroForGen-NoE).

References

[1] S. Bannwarth, V. Procaccio, A.S. Lebre, C. Jardel, A. Chaussenot, C. Hoarau, et al., Prevalence of rare mitochondrial DNA mutations in mitochondrial disorders, J. Med. Genet. 50 (2013) 704–714.

[2] J. Nunnari, A. Suomalainen, Mitochondria: in sickness and in health, Cell 148 (2012) 1145–1159.

[3] D.C. Wallace, D. Chalkia, Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease, Cold Spring Harbor Persp. Biol. 5 (2013) a021220.

[4] T. Kivisild, M. Reidla, E. Metspalu, A. Rosa, A. Brehm, E. Pennarun, et al., Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears, Am. J. Hum. Genet. 75 (2004) 752–770.

[5] M. Richards, V. Macaulay, A. Torroni, H.-J. Bandelt, In search of geographical patterns in European mitochondrial DNA, Am. J. Hum. Genet. 71 (2002) 1168–1174.

[6] M. van Oven, M. Kayser, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation, Hum. Mutat. 30 (2009) E386–E394.

[7] A. Sajantila, A.-H. Salem, P. Savolainen, K. Bauer, C. Gierig, S. Pääbo, Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population, Proc. Natl. Acad. Sci. U.S.A. 93 (1996) 12035–12039.

[8] P. Gill, P.L. Ivanov, C. Kimpton, R. Piercy, N. Benson, G. Tully, et al., Identification of the remains of the Romanov family by DNA analysis, Nat. Genet. 6 (1994) 130–135.

[9] M.M. Holland, T.J. Parsons, Mitochondrial DNA sequence analysis-validation and use for forensic casework, Forensic Sci. Rev. 11 (1999) 21–50.

[10] M.R. Wilson, J.A. DiZinno, D. Polanskey, J. Replogle, B. Budowle, Validation of mitochondrial DNA sequencing for forensic casework analysis, Int. J. Leg. Med. 108 (1995) 68–74.

[11] J.U. Palo, M. Hedman, N. Soderholm, A. Sajantila, Repatriation and identification of Finnish World War II soldiers, Croat. Med. J. 48 (2007) 528.

[12] W. Parson, A. Dür, EMPOP—a forensic mtDNA database, Forensic Sci. Int. Genet. 1 (2007) 88–92.

[13] H.Y. Lee, I. Song, E. Ha, S.-B. Cho, W.I. Yang, K.-J. Shin, mtDNAmanager: a Web-based tool for the management and quality analysis of mitochondrial DNA control-region sequences, BMC Bioinform. 9 (2008) 483.

[14] M. Ingman, U. Gyllensten, mtDB: human mitochondrial genome database, a resource for population genetics and medical sciences, Nucleic Acids Res. 34 (2006) D749–D751.

[15] L. Fan, Y.-G. Yao, An update to MitoTool: using a new scoring system for faster mtDNA haplogroup determination, Mitochondrion 13 (2013) 360–363.

[16] M. Attimonelli, M. Accetturo, M. Santamaria, D. Lascaro, G. Scioscia, G. Pappadà, et al., HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research, BMC Bioinform. 6 (2005) S4.

[17] J. Kohl, I. Paulsen, T. Laubach, A. Radtke, A. von Haeseler, HvrBase++: a phylogenetic database for primate species, Nucleic Acids Res. 34 (2006) D700–D704.

[18] M.C. Brandon, M.T. Lott, K.C. Nguyen, S. Spolim, S.B. Navathe, P. Baldi, et al., MITOMAP: a human mitochondrial genome database—2004 update, Nucleic Acids Res. 33 (2005) D611–D613.

[19] B.C. Levin, H. Cheng, D.J. Reeder, A human mitochondrial DNA standard reference material for quality control in forensic identification, medical diagnosis, and mutation detection, Genomics 55 (1999) 135–146.

[20] M. Coble, R. Just, J. O'Callaghan, I. Letmanyi, C. Peterson, J. Irwin, et al., Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians, Int. J. Leg. Med. 118 (2004) 137–146.

[21] A. Brandstätter, T.J. Parsons, W. Parson, Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups, Int. J. Leg. Med. 117 (2003) 291–298.

[22] T.J. Parsons, M.D. Coble, Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome, Croat. Med. J. 42 (2001) 304–309.

[23] H.-J. Bandelt, P. Lahermo, M. Richards, V. Macaulay, Detecting errors in mtDNA data by phylogenetic analysis, Int. J. Leg. Med. 115 (2001) 64–69.

[24] H.-J. Bandelt, M. van Oven, A. Salas, Haplogrouping mitochondrial DNA sequences in Legal Medicine/Forensic Genetics, Int. J. Leg. Med. 126 (2012) 901–916.

[25] A.W. Röck, A. Dür, M. van Oven, W. Parson, Concept for estimating mitochondrial DNA haplogroups using a maximum likelihood approach (EMMA), Forensic Sci. Int. Genet. 7 (6) (2013) 601–609.

[26] A. Kloss-Brandstätter, D. Pacher, S. Schönherr, H. Weissensteiner, R. Binna, G. Specht, et al., HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups, Hum. Mutat. 32 (2011) 25–32.

[27] D. Vianello, F. Sevini, G. Castellani, L. Laura, M. Capri, C. Franceschi, HAPLOFIND: a new method for high-throughput mtDNA haplogroup assignment, Hum. Mutat. 34 (9) (2013) 1189–1194.

[28] M.A. Quail, M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, et al., A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, BMC Genom. 13 (2012) 341.

[29] S. Jünemann, F.J. Sedlazeck, K. Prior, A. Albersmeier, U. John, J. Kalinowski, et al., Updating benchtop sequencing performance comparison, Nat. Biotechnol. 31 (2013) 294–296.

[30] E.D. Gunnarsdóttir, M. Li, M. Bauchet, K. Finstermeier, M. Stoneking, High-throughput sequencing of complete human mtDNA genomes from the Philippines, Genome Res. 21 (2011) 1–11.

[31] Illumina Nextera® XT DNA Sample Preparation Guide, 2012.

[32] Illumina Nextera® Library Validation and Cluster Density Optimization, 2013.

[33] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297–1303.

[34] R.M. Andrews, I. Kubacka, P.F. Chinnery, R.N. Lightowlers, D.M. Turnbull, N. Howell, Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, Nat. Genet. 23 (1999) 147.

[35] J.L. King, A. Sajantila, B. Budowle, mitoSAVE: Mitochondrial sequence analysis of variants in Excel, Forensic Sci. Int. Genet. (2014), http://dx.doi.org/10.1016/j.fsigen.2014.05.013.

[36] H. Thorvaldsdóttir, J.T. Robinson, J.P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, Brief. Bioinform. 14 (2) (2012) 178–192.

[37] M. Stoneking, D. Hedgecock, R.G. Higuchi, L. Vigilant, H.A. Erlich, Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes, Am. J. Hum. Genet. 48 (1991) 370.

[38] F. Tajima, Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, Genetics 123 (1989) 585–595.

[39] K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: molecular evolutionary genetics analysis version 6.0, Mol. Biol. Evol. 30 (2013) 2725–2729.

[40] M.I. Krzywinski, J.E. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, et al., Circos: an information aesthetic for comparative genomics, Genome Res. 19 (2009) 1639–1645.

[41] R.S. Just, M.K. Scheible, S.A. Fast, K. Sturk-Andreaggi, J.L. Higginbotham, E.A. Lyons, et al., Development of forensic-quality full mtGenome haplotypes: success rates with low template specimens, Forensic Sci. Int. Genet. 10 (2014) 73–79.

[42] B. Ewing, P. Green, Base-calling of automated sequencer traces using phred. II. Error probabilities, Genome Res. 8 (1998) 186–194.

[43] B. Budowle, M.R. Wilson, J.A. DiZinno, C. Stauffer, M.A. Fasano, M.M. Holland, et al., Mitochondrial DNA regions HVI and HVII population data, Forensic Sci. Int. 103 (1999) 23–35.

[44] W. Parson, C. Strobl, G. Huber, B. Zimmermann, S.M. Gomes, L. Souto, et al., Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM), Forensic Sci. Int. Genet. 7 (2013) 543–549.

[45] H.-J. Bandelt, A. Salas, Current next generation sequencing technology may not meet forensic standards, Forensic Sci. Int. Genet. 6 (2012) 143–145.

[46] H. Oberacher, H. Niederstätter, C.G. Huber, W. Parson, Accurate determination of allelic frequencies in mitochondrial DNA mixtures by electrospray ionization time-of-flight mass spectrometry, Anal. Bioanal. Chem. 384 (2006) 1155–1163.

[47] M. Mikkelsen, R.F. Hansen, A.J. Hansen, N. Morling, Massively parallel pyrosequencing 454 methodology of the mitochondrial genome in forensic genetics, Forensic Sci. Int. Genet. 12 (2014) 30–37.

Publication II

Stoljarova, M., King, J. L., Takahashi, M., Aaspollu, A., & Budowle, B.

Whole mitochondrial genome genetic diversity in an Estonian population sample.

Int J Legal Med. 2016 Jan;130(1):67-71.

Int J Legal Med (2016) 130:67–71
DOI 10.1007/s00414-015-1249-4

ORIGINAL ARTICLE

Whole mitochondrial genome genetic diversity in an Estonian population sample

Monika Stoljarova[1,2] · **Jonathan L. King**[2] · **Maiko Takahashi**[2] · **Anu Aaspõllu**[1] · **Bruce Budowle**[2]

Received: 1 May 2015 / Accepted: 12 August 2015 / Published online: 20 August 2015

Abstract Mitochondrial DNA is a useful marker for population studies, human identification, and forensic analysis. Commonly used hypervariable regions I and II (HVI/HVII) were reported to contain as little as 25 % of mitochondrial DNA variants and therefore the majority of power of discrimination of mitochondrial DNA resides in the coding region. Massively parallel sequencing technology enables entire mitochondrial genome sequencing. In this study, buccal swabs were collected from 114 unrelated Estonians and whole mitochondrial genome sequences were generated using the Illumina MiSeq system. The results are concordant with previous mtDNA control region reports of high haplogroup HV and U frequencies (47.4 and 23.7 % in this study, respectively) in the Estonian population. One sample with the Northern Asian haplogroup D was detected. The genetic diversity of the Estonian population sample was estimated to be 99.67 and 95.85 %, for mtGenome and HVI/HVII data, respectively. The random match probability for mtGenome data was 1.20 versus 4.99 % for HVI/HVII. The nucleotide mean pairwise difference was 27±11 for mtGenome and 7±3 for HVI/HVII data. These data describe the genetic diversity of the Estonian population sample and emphasize the power of discrimination of the entire mitochondrial genome over the hypervariable regions.

Keywords Mitochondrial genome (mtGenome) · Massively parallel sequencing (MPS) · Illumina MiSeq · Haplogroup · Estonia · Power of discrimination

Introduction

The analysis of mitochondrial DNA (mtDNA) has been implemented in molecular anthropology [1], evolutionary biology [2], medical genetics [3, 4], and human identity testing [5]. Strictly maternal inheritance, lack of recombination and high mutation rate make mtDNA a viable marker system for assessing genetic relationships among individuals or groups. Although having less discrimination power compared to autosomal DNA markers, the high copy number of mtDNA per cell increases the success rate of DNA typing of damaged, degraded, and low-quantity samples that fail to yield nuclear DNA profiles.

Forensic scientists have focused mainly on hypervariable regions I and II (HVI and HVII, respectively) due to high concentration of variants in those mitochondrial genome (mtGenome) regions. However, it has been reported that 75 % of the total variation within the mtGenome resides outside the control region, and therefore sequencing of the entire mtGenome increases the discrimination power and the value of generated data [6]. Sanger-type sequencing (STS) is still the main mtDNA typing technique in case work laboratories, and sequencing beyond the control region is attempted rarely as the well-established methodology is labor intensive, costly, and time consuming. In recent years, massively parallel sequencing (MPS) has been shown to be a feasible alternative to

Electronic supplementary material The online version of this article (doi:10.1007/s00414-015-1249-4) contains supplementary material, which is available to authorized users.

✉ Monika Stoljarova
monika.stoljarova@ttu.ee

[1] Department of Gene Technology, Tallinn University of Technology, Akadeemia tee 15A-604, Tallinn 12618, Estonia

[2] Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA

STS. With the utility of bench-top sequencers like Illumina MiSeq and Ion Torrent PGM (Personal Genome Machine), generation of whole mtGenome data is feasible for the application-orientated laboratory [6–10].

Population data are essential for haplotype frequency estimation. While more than 25,000 forensic mtDNA sequences and almost 35,000 mtDNA sequences in total are available currently in the EMPOP database [11, 12], these data include information only from the mtDNA control region. In addition to STS, a number of other technologies permit practical access to mtDNA coding region data through single nucleotide polymorphism assays, sequence-specific oligonucleotide probes, mass spectroscopy, and MPS technology that emphasize the importance of a whole mtGenome database [13]. Recently, 588 forensic-quality whole mtGenomes from three major US populations have been determined with Sanger sequencing and will be available for query [14]. However, with MPS, population data can be generated far more expeditiously and at a lower cost per nucleotide. Thus, more whole genome data can be generated to exploit the full power of mtDNA for forensic identity testing.

The current population size of Estonia is slightly above 1.3 million. Throughout history, the native Estonian population has been affected by migration from both east and west due to numerous conquests. Estonians have served under German, Danish, Polish, Swedish, and Russian rule. Thus, there is some expectation of genetic admixture from these populations. A European genetic map of >1500 individuals and based on ~270,000 single nucleotide polymorphism data divided the European population into four groups and placed Estonians into the Baltic region, Poland, and Western Russia group [15]. It has been reported that Estonians have a higher Y-haplotype diversity, and based on their mtDNA HVI sequence, they have higher mean pairwise differences compared with other populations in the Baltic region [16].

Despite that fact that the Estonian population data has been used for migration studies based on its mtDNA HVI region and a number of coding region single nucleotide polymorphisms [17, 18, 16], to the best of our knowledge there are no published whole mtGenome data. The objective of this study was to describe the genetic variability of mtGenome in an Estonian population sample and to compare the discrimination power of mtGenome with solely HVI/HVII data.

Materials and methods

Sample preparation and target amplification

Buccal swabs were collected from 114 unrelated Estonian volunteers according to protocols approved by the Tallinn Medical Research Ethics Committee and have been performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. In addition, all samples were collected in accordance with the University of North Texas Health Science Center Institutional Review Board. Buccal samples were collected with sterile Eurotubo® collection swabs (Deltalab, Rubí, Spain). The swabs were allowed to air dry and were stored at ambient temperature. DNA extraction was performed with the QIAamp® DNA Blood Mini Kit (QIAGEN, Hilden, Germany) according to the manufacturer's protocol. The quantity of extracted DNA was determined using the Qubit® dsDNA HS (High Sensitivity) Quantification Kit and a Qubit® 2.0 Fluorometer (Life Technologies, Foster City, CA, USA). Amplification of the mtGenome was accomplished as described by King et al. [6].

Nextera XT library preparation

For library preparation, 1.0 ng of DNA was used. Libraries were prepared using Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA, USA) according to the manufacturer's protocol [19] except for the library normalization after bead cleanup and library preparation for sequencing. Following PCR cleanup, the libraries were quantified using the Qubit dsDNA BR (Broad Range) kit (Life Technologies, Foster City, CA, USA) and evaluated for fragment size using the High Sensitivity D1000 ScreenTape and Tape Station 2200 (Agilent Technologies, Santa Clara, CA, USA). Library normalization and preparation for sequencing was performed according to an in-house protocol. Purified libraries were normalized to 2 nM and pooled. Illumina PhiX control v3 (Illumina) was diluted to 2 nM with resuspension buffer (RSB, Illumina), and 2 μl of diluted PhiX were mixed with 14 μl of pooled library (2 nM). Further, 10 μl of PhiX-library pool were mixed with 10 μl of freshly made 0.1 N NaOH and vortexed. The library-PhiX-NaOH mix was incubated for 5 min at room temperature. Pre-chilled HT1 (980 μl, hybridization buffer, Illumina) was added to the mix. Then, 600 μl of library were mixed with 400 μl of HT1 for a final 12 pM sequencing library.

Sequencing and data generation

The 12 pM pooled library was sequenced with MiSeq v2 (2× 250 bp) chemistries (Illumina). The MiSeq re-sequencing protocol for small genome sequencing was followed according to the manufacturer's recommendations. Criteria used for variant calling (quality threshold, heteroplasmy threshold, coverage threshold) were as described by King et al. [6]. On-board software (i.e., Real-TimeAnalysis and MiSeq Reporter) converted raw data to Binary Alignment/Map (BAM) [20] and Variant Call Format (VCF) v4.1 files [21] using Genome Analysis Toolkit (GATK) [22]. During this process, the sequenced regions of interest (ROIs) were aligned to the revised Cambridge Reference Sequence (rCRS) [23].

Data analysis

Generated VCF files were converted into haplotypes using MitoSAVE [24]. mtDNA variants were confirmed manually using BAM files and Integrative Genomic Viewer (IGV) software. Indels at the positions 309, 315, and 16193 were not included in the analysis. Random match probability (RMP) and genetic diversity (GD) were calculated according to the methods described by Stoneking et al. [25] and Tajima [26], respectively. Mean pairwise comparison was calculated using MEGA [27]. HaploGrep software based on Phylotree 16 [28, 29] was used for haplogroup assignment.

Results and discussion

All 114 samples resulted in whole mtGenome sequences obtained by MPS (Supplemental Table 1). Of these mtDNA profiles, 100 (87.7 %) were unique within the data set, and 12.3 % of sequenced mtGenomes were observed twice. Compared to the previous whole mtGenome population studies [6, 14], the number of shared haplotypes in the current study is relatively high. This might be a reflection of lower mtGenome diversity in Estonia or just may arise from sampling variance. The majority of the samples were collected at two educational institutions, and although an effort was made to ensure that the sampled individuals were unrelated, a long-distance kinship between these individuals cannot be excluded. In total, 2663 positions were reported as variants in relation to rCRS. These variants were distributed across 512 mtDNA positions. Variants 263G, 750G, 1438G, 4769G, 8860G, and 15326G were seen in 111 of 114 samples. The detection of these variants in majority of the samples is the reflection of reference used. The remaining three samples (EST-9, EST-19, and EST-40) exhibited few differences with respect to the rCRS (≤3), of which 1–2 were local or global private mutations as defined by HaploGrep. The low haplogroup assignment quality score for sample EST-19 was noted. The limited number of variants within the sample could indicate bias previously observed in reference to Phylotree and the rCRS [30]. Therefore, the true haplogroup for EST-19 individual (haplogroup H5e with the quality score 53 % assigned by HaploGrep) is likely to be in the H2 lineage. Six variants (73G, 2706G, 7028T, 11719A, 14766T, and 16519C) were found in ≥50 % of the samples. From these variants, 73G and 11719A were haplogroup nodes for haplogroup R, variants 2706G and 7028T for haplogroup H, and variant 14766T for haplogroup HV. Position 16519 is considered a hotspot and thus not useful for haplogroup assignment. Point heteroplasmy was detected in 14 samples (12.3 %). Observed point heteroplasmy with position coverage and heteroplasmy percentage is listed in Supplemental Table 2. Three samples (EST-33, EST-81, and EST-106) exhibited two point heteroplasmy positions each. A similar extent of heteroplasmy (16.2 %), along with heteroplasmy at multiple positions per sample, has been observed previously in buccal cell mtDNA [31].

Haplogroup assignment using HaploGrep software resulted in 11 major clades and 87 distinct haplogroups. Major clades were D, HV (including haplogroup H), I, J, M, N, R, T, U (including haplogroup K), W, and X. Two clades were dominant: 54 samples (47.4 %) belonged to the haplogroup HV (including H) and 27 samples (23.7 %) pertained to the haplogroup U (including K). Haplogroups D, R, and X were seen once. Haplogroups HV and U are the most represented haplogroups in the European population with the estimated frequency of ≥50 and ≥20 %, respectively [17]. Our results are concordant with previous mtDNA control region reports conforming of the prevalence of haplogroup H and U in the Estonian population [32, 33]. While haplogroup H was introduced to Europe from the Franco-Cantabrian region, haplogroup U5, which was observed in 13 of our samples (11.4 %), is thought to have evolved in situ [32, 17]. Haplogroup U4 has been reported in the Eastern Baltic Sea region with a frequency up to 8.8 % and associated with Volga-Ural influence [16]. While haplogroup D is the second most common haplogroup in Northern Asia, haplogroup D5 has been found with a very low frequency in several European populations including Estonians [34]. The rare subhaplogroup D4e4b, observed in one of our samples, has been reported in Tatars and Russians [35].

Out of the observed 2663 nucleotide variants, 607 (22.8 %) were identified in HVI/HVII regions; accordingly, 77.2 % of variation resided in the coding region. These results are in accordance with results reported by King et al. on 283 individuals from Caucasian, Hispanic, and African-American populations [6]. As in the case of whole mtGenome data, the proportionally smaller level of variation in HVI/HVII may be a reflection of lower mtGenome diversity in Estonia or just may arise from sampling variance. Whereas 100 unique mtGenome haplotypes and 87 distinct haplogroups were observed with whole mtGenome data, only 66 unique haplotypes and 79 distinct haplogroups were observed using HVI/HVII data. Haplogroup comparison between full mtGenome and HVI/HVII data resulted in a haplogroup clade change according to HaploGrep for 1 sample (EST-59) that changed from haplogroup U5b2a1a2 (HVI/HVII; quality score 86.8 %) to haplogroup H (mtGenome; quality score 81 %). Haplogroup assignment based on mtGenome data resulted in a quality score increase for the majority of the samples that yielded a quality score less than 100 % with HVI/HVII. A lower quality score of sample EST-59 can be explained with an abundance of local and global private mutations that were not present in HVI/HVII data. GD for mtGenome and HVI/

HVII was 99.67 and 95.85 %, respectively. RMP for mtGenome data was 1.20 versus 4.99 % for HVI/HVII. Compared to RMP values of other populations [14, 6, 36], the RMP results presented herein for HVI/HVII data are higher. This finding might be explained with 24 HVI/HVII haplotypes having ≥4 identical matches in the population sample. Mean pairwise difference within the Estonian population was 27±11 for mtGenome data, which is slightly lower than reported by King et al. [6] for Caucasians. Mean pairwise difference for HVI/HVII data within the Estonian population sample was 7±3. These results support the power of discrimination of entire mtGenome over HVI/HVII.

Conclusion

In this study, 114 mtGenome profiles from the Estonian population were generated. The results show that the use of the entire mtGenome compared to HVI/HVII data substantially improves the discrimination power of the quality of haplotypes.

Acknowledgments We would like to thank the Baltic American Freedom Foundation and Council of International Educational Exchange for providing the opportunity for the professional internship of Monika Stoljarova at the University of North Texas Health Science Center.

References

1. King TE, Fortes GG, Balaresque P, Thomas MG, Balding D, Delser PM, Neumann R, Parson W, Knapp M, Walsh S, Tonasso L, Holt J, Kayser M, Appleby J, Forster P, Ekserdjian D, Hofreiter M, Schurer K (2014) Identification of the remains of King Richard III. Nat Commun 5:5631. doi:10.1038/ncomms6631
2. Rito T, Richards MB, Fernandes V, Alshamali F, Cerny V, Pereira L, Soares P (2013) The first modern human dispersals across Africa. PLoS One 8(11), e80031. doi:10.1371/journal.pone.0080031
3. Li CT, Bai YM, Hsieh JC, Lee HC, Yang BH, Chen MH, Lin WC, Tsai CF, Tu PC, Wang SJ, Su TP (2015) Peripheral and central glucose utilizations modulated by mitochondrial DNA 10398A in bipolar disorder. Psychoneuroendocrinology 55C:72–80. doi:10.1016/j.psyneuen.2015.02.003
4. Hagen CM, Aidt FH, Hedley PL, Jensen MK, Havndrup O, Kanters JK, Moolman-Smook JC, Larsen SO, Bundgaard H, Christiansen M (2013) Mitochondrial haplogroups modify the risk of developing hypertrophic cardiomyopathy in a Danish population. PLoS One 8(8), e71904. doi:10.1371/journal.pone.0071904
5. Tokutomi T, Takada Y, Kanetake J, Mukaida M (2009) Identification using DNA from skin contact: case reports. Leg Med (Tokyo) 11(Suppl 1):S576-577. doi:10.1016/j.legalmed.2009.02.004
6. King JL, LaRue BL, Novroski NM, Stoljarova M, Seo SB, Zeng X, Warshauer DH, Davis CP, Parson W, Sajantila A, Budowle B (2014) High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int Genet 12:128–135. doi:10.1016/j.fsigen.2014.06.001
7. Seo SB, Zeng X, King JL, Larue BL, Assidi M, Al-Qahtani MH, Sajantila A, Budowle B (2015) Underlying data for sequencing the mitochondrial genome with the massively parallel sequencing platform ion torrent() PGM (). BMC Genomics 16(1):6938. doi:10.1186/1471-2164-16-S1-S4
8. Parson W, Strobl C, Huber G, Zimmermann B, Gomes SM, Souto L, Fendt L, Delport R, Langit R, Wootton S, Lagace R, Irwin J (2013) Reprint of: evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). Forensic Sci Int Genet 7(6):632–639. doi:10.1016/j.fsigen.2013.09.007
9. Bodner M, Iuvaro A, Strobl C, Nagl S, Huber G, Pelotti S, Pettener D, Luiselli D, Parson W (2015) Helena, the hidden beauty: resolving the most common West Eurasian mtDNA control region haplotype by massively parallel sequencing an Italian population sample. Forensic Sci Int-Gen 15:21–26. doi:10.1016/j.fsigen.2014.09.012
10. McElhoe JA, Holland MM, Makova KD, Su MS, Paul IM, Baker CH, Faith SA, Young B (2014) Development and assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq. Forensic Sci Int Genet 13:20–29. doi:10.1016/j.fsigen.2014.05.007
11. European DNA Profiling Group (EDNAP) (1999–2015) EMPOP. www.empop.org. Accessed 19 Mar 2015
12. Parson W, Dur A (2007) EMPOP—a forensic mtDNA database. Forensic Sci Int Genet 1(2):88–92. doi:10.1016/j.fsigen.2007.01.018
13. Just RS, Scheible MK, Fast SA, Sturk-Andreaggi K, Higginbotham JL, Lyons EA, Bush JM, Peck MA, Ring JD, Diegoli TM, Rock AW, Huber GE, Nagl S, Strobl C, Zimmermann B, Parson W, Irwin JA (2014) Development of forensic-quality full mtGenome haplotypes: success rates with low template specimens. Forensic Sci Int Genet 10:73–79. doi:10.1016/j.fsigen.2014.01.010
14. Just RS, Scheible MK, Fast SA, Sturk-Andreaggi K, Rock AW, Bush JM, Higginbotham JL, Peck MA, Ring JD, Huber GE, Xavier C, Strobl C, Lyons EA, Diegoli TM, Bodner M, Fendt L, Kralj P, Nagl S, Niederwieser D, Zimmermann B, Parson W, Irwin JA (2015) Full mtGenome reference data: development and characterization of 588 forensic-quality haplotypes representing three U.S. populations. Forensic Sci Int Genet 14:141–155. doi:10.1016/j.fsigen.2014.09.021
15. Nelis M, Esko T, Magi R, Zimprich F, Zimprich A, Toncheva D, Karachanak S, Piskackova T, Balascak I, Peltonen L, Jakkula E, Rehnstrom K, Lathrop M, Heath S, Galan P, Schreiber S, Meitinger T, Pfeufer A, Wichmann HE, Melegh B, Polgar N, Toniolo D, Gasparini P, D'Adamo P, Klovins J, Nikitina-Zake L, Kucinskas V, Kasnauskiene J, Lubinski J, Debniak T, Limborska S, Khrunin A, Estivill X, Rabionet R, Marsal S, Julia A, Antonarakis SE, Deutsch S, Borel C, Attar H, Gagnebin M, Macek M, Krawczak M, Remm M, Metspalu A (2009) Genetic structure of Europeans: a view from the North-East. PLoS One 4(5), e5472. doi:10.1371/journal.pone.0005472
16. Lappalainen T, Laitinen V, Salmela E, Andersen P, Huoponen K, Savontaus ML, Lahermo P (2008) Migration waves to the Baltic Sea region. Ann Hum Genet 72(Pt 3):337–348. doi:10.1111/j.1469-1809.2007.00429.x
17. Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Golge M, Dimitrov D, Hill E, Bradley D, Romano V, Cali F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Norby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A, Bandelt HJ (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67(5):1251–1276
18. Pliss L, Tambets K, Loogvali EL, Pronina N, Lazdins M, Krumina A, Baumanis V, Villems R (2006) Mitochondrial DNA portrait of

Latvians: towards the understanding of the genetic structure of Baltic-speaking populations. Ann Hum Genet 70(Pt 4):439–458. doi:10.1111/j.1469-1809.2005.00238.x

19. Illumina (2015) Nextera XT DNA Library preparation guide
20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079. doi:10.1093/bioinformatics/btp352
21. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Genomes Project Analysis G (2011) The variant call format and VCFtools. Bioinformatics 27(15):2156–2158. doi:10.1093/bioinformatics/btr330
22. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297–1303. doi:10.1101/gr.107524.110
23. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23(2):147. doi:10.1038/13779
24. King JL, Sajantila A, Budowle B (2014) mitoSAVE: mitochondrial sequence analysis of variants in Excel. Forensic Sci Int Genet 12: 122–125. doi:10.1016/j.fsigen.2014.05.013
25. Stoneking M, Hedgecock D, Higuchi RG, Vigilant L, Erlich HA (1991) Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. Am J Hum Genet 48(2):370–382
26. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595
27. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30(12):2725–2729. doi:10.1093/molbev/mst197
28. Kloss-Brandstatter A, Pacher D, Schonherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F (2011) HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat 32(1):25–32. doi:10.1002/humu.21382
29. van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30(2):E386–E394. doi:10.1002/humu.20921
30. Bandelt HJ, Salas A (2012) Current next generation sequencing technology may not meet forensic standards. Forensic Sci Int Genet 6(1):143–145. doi:10.1016/j.fsigen.2011.04.004
31. Naue J, Horer S, Sanger T, Strobl C, Hatzer-Grubwieser P, Parson W, Lutz-Bonengel S (2015) Evidence for frequent and tissue-specific sequence heteroplasmy in human mitochondrial DNA. Mitochondrion 20:82–94. doi:10.1016/j.mito.2014.12.002
32. Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V, Moral P, Dugoujon JM, Roostalu U, Loogvali EL, Kivisild T, Bandelt HJ, Richards M, Villems R, Santachiara-Benerecetti AS, Semino O, Torroni A (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 75(5):910–918. doi:10.1086/425590
33. Ottoni C, Ricaut FX, Vanderheyden N, Brucato N, Waelkens M, Decorte R (2011) Mitochondrial analysis of a Byzantine population reveals the differential impact of multiple historical events in South Anatolia. Eur J Hum Genet: EJHG 19(5):571–576. doi:10.1038/ejhg.2010.230
34. Tambets K, Rootsi S, Kivisild T, Help H, Serk P, Loogvali EL, Tolk HV, Reidla M, Metspalu E, Pliss L, Balanovsky O, Pshenichnov A, Balanovska E, Gubina M, Zhadanov S, Osipova L, Damba L, Voevoda M, Kutuev I, Bermisheva M, Khusnutdinova E, Gusar V, Grechanina E, Parik J, Pennarun E, Richard C, Chaventre A, Moisan JP, Barac L, Pericic M, Rudan P, Terzic R, Mikerezi I, Krumina A, Baumanis V, Koziel S, Rickards O, De Stefano GF, Anagnou N, Pappa KI, Michalodimitrakis E, Ferak V, Furedi S, Komel R, Beckman L, Villems R (2004) The western and eastern roots of the Saami—the story of genetic “outliers” told by mitochondrial DNA and Y chromosomes. Am J Hum Genet 74(4):661–682. doi:10.1086/383203
35. Derenko M, Malyarchuk B, Grzybowski T, Denisova G, Rogalla U, Perkova M, Dambueva I, Zakharov I (2010) Origin and post-glacial dispersal of mitochondrial DNA haplogroups C and D in northern Asia. PLoS One 5(12), e15214. doi:10.1371/journal.pone.0015214
36. Lembring M, van Oven M, Montelius M, Allen M (2013) Mitochondrial DNA analysis of Swedish population samples. Int J Legal Med 127(6):1097–1099. doi:10.1007/s00414-013-0908-6

Publication III

Churchill, J. D., **Stoljarova, M.**, King, J. L., & Budowle, B.

Massively parallel sequencing-enabled mixture analysis of mitochondrial DNA samples.

Int J Legal Med. 2018 Sep;132(5):1263-1272.

International Journal of Legal Medicine (2018) 132:1263–1272
https://doi.org/10.1007/s00414-018-1799-3

ORIGINAL ARTICLE

Massively parallel sequencing-enabled mixture analysis of mitochondrial DNA samples

Jennifer D. Churchill[1] · Monika Stoljarova[2] · Jonathan L. King[1] · Bruce Budowle[1,3]

Received: 11 January 2018 / Accepted: 9 February 2018 / Published online: 22 February 2018

Abstract

The mitochondrial genome has a number of characteristics that provide useful information to forensic investigations. Massively parallel sequencing (MPS) technologies offer improvements to the quantitative analysis of the mitochondrial genome, specifically the interpretation of mixed mitochondrial samples. Two-person mixtures with nuclear DNA ratios of 1:1, 5:1, 10:1, and 20:1 of individuals from different and similar phylogenetic backgrounds and three-person mixtures with nuclear DNA ratios of 1:1:1 and 5:1:1 were prepared using the Precision ID mtDNA Whole Genome Panel and Ion Chef, and sequenced on the Ion PGM or Ion S5 sequencer (Thermo Fisher Scientific, Waltham, MA, USA). These data were used to evaluate whether and to what degree MPS mixtures could be deconvolved. Analysis was effective in identifying the major contributor in each instance, while SNPs from the minor contributor's haplotype only were identified in the 1:1, 5:1, and 10:1 two-person mixtures. While the major contributor was identified from the 5:1:1 mixture, analysis of the three-person mixtures was more complex, and the mixed haplotypes could not be completely parsed. These results indicate that mixed mitochondrial DNA samples may be interpreted with the use of MPS technologies.

Keywords Mitochondrial DNA · Mixtures · Massively parallel sequencing · Ion S5 · Ion PGM

Introduction

Mitochondrial DNA has become a powerful tool for the identification of human remains and in analyses of certain types of forensic evidence from criminal cases, e.g., hair evidence. The mitochondrial genome's higher copy number per cell, compared with the nuclear genome [1], provides a high sensitivity of detection with challenged or degraded remains, where nuclear markers often provide inconclusive or negative results. In addition, maternal inheritance [2] and well-characterized phylogeny [3–15] of the mitochondrial genome offer useful lineage and bioancestry information. Currently, Sanger sequencing technologies are employed to sequence a limited portion of the mitochondrial genome, often focusing only on the hypervariable regions of the non-coding region. Because the assay is time-consuming and labor-intensive, substantial variation residing in the coding region of the mitochondrial genome is not considered. Moreover, Sanger sequencing is not sufficiently quantitative to resolve mixed mitochondrial DNA profiles [16, 17].

Massively parallel sequencing (MPS) technologies now make it feasible for forensic crime laboratories to sequence the entire mitochondrial genome. Large multiplex, small amplicon panels that amplify the entire mitochondrial genome have been designed for challenged and degraded samples [18–20]. Moreover, the technology has become reasonably robust such that the amount of time and labor needed to sequence the entire mitochondrial genome has been reduced substantially. The high throughput concomitantly provides a much larger amount of useful information. Expanding analysis to the entire mitochondrial genome enables analysis of a previously untapped resource of a large number of single

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00414-018-1799-3) contains supplementary material, which is available to authorized users.

✉ Jennifer D. Churchill
Jennifer.Churchill@unthsc.edu

[1] Center for Human Identification, University of North Texas Health Science Center, 3500 Camp Bowie Blvd, CBH-250, Fort Worth, TX 76107, USA

[2] Department of Chemistry and Biotechnology, Tallinn University of Technology, Tallinn, Estonia

[3] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

nucleotide polymorphisms (SNPs) (up to 75% of total mitochondrial DNA variation) when the mitochondrial coding region is evaluated [4, 5, 7, 21–24]. Analysis of only the mitochondrial control region may provide only limited phylogenetic information as two samples with a control region match do not necessarily belong to the same haplogroup [23]. Thus, sequence data from the entire mitochondrial genome is likely to increase phylogenetic resolution [4, 5, 7, 23, 24]. Additionally, since each molecule (or clonal cluster) is sequenced independently, heteroplasmy detection can be enhanced versus the simultaneous sequencing of each amplicon by Sanger sequencing.

Mixtures are one of the more challenging sample types encountered in forensic casework, and mitochondrial DNA mixture interpretation typically is not attempted with current sequencing technologies used in forensic crime labs. In fact, the current recommendation from the DNA Commission of the International Society of Forensic Genetics (ISFG) states that heteroplasmy evaluation depends, in part, on the limitations of the technology [25]. MPS-generated data are more quantitative than Sanger sequencing data. Studies such as Stewart et al. [26] and Davis et al. [27] visually illustrate the lack of quantitative information provided by Sanger sequencing at varying mixture ratios and heteroplasmic sites, respectively. Combining quantitative information and phylogenetic assignment may make it feasible to effect mixture deconvolution in some samples [16, 17, 20, 28–30]. In this study, two-person mixtures and three-person mixtures of individuals from differing and similar phylogenetic backgrounds were prepared in various ratios. A workflow consisting of the Precision ID mtDNA Whole Genome Panel, Ion Chef, and Ion PGM/S5 sequencer (Thermo Fisher Scientific, Waltham, MA, USA) was used to sequence the mixture samples. Finally, a bioinformatic pipeline using quantitative analysis of positions with multiple allele states, phasing information, and phylogenetics was used to parse mixed haplotypes into their individual components.

Materials and methods

Samples

The policies and procedures approved by the Institutional Review Board for the University of North Texas Health Science Center in Fort Worth, TX, were followed for the collection and use of samples. DNA used in this study was extracted using the QIAamp DNA Blood Mini Kit (Qiagen, Valencia, CA, USA) following the manufacturer's protocols [31] and quantified using the Quantifiler Trio DNA Quantification kit (Thermo Fisher Scientific) following the manufacturer's protocols [32]. Single-source reference samples ($n = 6$) of self-identified Asian and Caucasian individuals, mixed samples ($n = 12$), and positive and negative controls were included in the sequencing runs. Two-person mixtures with contributors of different macrohaplogroups were prepared in 1:1, 5:1, 10:1, and 20:1 major to minor contributor nuclear DNA ratios for both individuals. Two-person mixtures with contributors belonging to the same macrohaplogroup were prepared in 1:1, 5:1, and 1:5 nuclear DNA ratios. Three-person mixtures with contributors of different macrohaplogroups were prepared in 1:1:1 and 5:1:1 nuclear DNA ratios.

Library preparation and massively parallel sequencing

The mitochondrial genome was amplified in each single-source and mixed sample with the Precision ID mtDNA Whole Genome Panel (Thermo Fisher Scientific), a multiplexed panel which generates amplicons of 175 base pairs or less that cover the entire mitochondrial genome in a tiled, overlapping manner [33]. Each amplification was performed with one nanogram of total input nuclear DNA following the manufacturer's protocols. Sequencing libraries were prepared manually using the Precision ID Library Kit (Thermo Fisher Scientific) and the manufacturer's recommended protocols for the "2-in-1 method."

Template preparation was completed on the Ion Chef (Thermo Fisher Scientific) following the manufacturer's recommended protocols [33, 34] for both sequencing runs. Templated Ion Sphere Particles (ISPs) were loaded into an Ion 318 Chip v2 (Thermo Fisher Scientific) for the Personal Genome Machine (PGM; Thermo Fisher Scientific) sequencing run and an Ion 530 Chip (Thermo Fisher Scientific) for the Ion S5 (Thermo Fisher Scientific) sequencing run. The Ion Chips were loaded onto their respective sequencers using the Ion PGM Hi-Q Sequencing Kit for the PGM run and the Ion S5 Sequencing Kit for the S5 run and the manufacturer's respective recommended protocols [33, 34].

Concordance data

An orthogonal methodology was used to generate concordance data for the single-source reference samples included in this study. The concordance data were generated via long-PCR on the Ion PGM and MiSeq (Illumina, San Diego, CA, USA) sequencers as described in Churchill et al. [35] and King et al. [5].

Data analysis

Primary data analyses were completed with the Torrent Suite software v5.2.1. Data were aligned to an "rCRS+80" reference genome to account for the Precision ID mtDNA Whole Genome Panel's tiled, overlapping design [33, 36]. Variant calls were obtained from the variant call format (VCF) output files generated by the Variant Caller plugin v5.2.1 and were imported into mitoSAVE [37] to generate haplotype calls in standard forensic nomenclature [25, 38]. A minimum of 10 reads (X) and allele ratio of 0.10 were used as thresholds for generating haplotype calls in mitoSAVE. Length heteroplasmies were not included in the final haplotype calls. Additionally, for the mixed samples, the ratio of the reference allele and alternate allele compared to the total read depth for each SNP was obtained from mitoSAVE to use as a quantitative assessment of each contributor's proportion of the mixture. Binary alignment map (BAM) files were viewed in Integrative Genomic Viewer (IGV) for a manual verification of the haplotype calls and to identify any relevant phasing information [39, 40]. A phylogenetic check of the final haplotype calls was performed in HaploGrep v2.1.1 and EMPOP v3 [8, 41, 42]. Finally, performance metrics, including read depth, relative locus performance (RLP), strand balance, and noise, were used to evaluate the quality of the sequencing results. The read depth was used to calculate normalized RLP at each nucleotide position of the mitochondrial genome (i.e., read depth of one nucleotide position divided by the total read depth across the mitochondrial genome for that sample each multiplied by the length of the rCRS (i.e., 16,569)). Strand balance ratios were calculated by dividing the read depth of one strand by the total read depth of that nucleotide position. Noise was calculated by dividing the number of reads not attributed to nominal allele calls at a nucleotide position by the total coverage at that nucleotide position.

Results and discussion

Controls

A positive and negative control was included in each sequencing run to help evaluate the success and performance of each run. The read depth of the negative controls was compared to the read depth of the single-source reference samples by calculating the ratio of the negative controls' average read depth to the single-source samples' average read depth across the mitochondrial genome. The average read depth for both negative controls ranged from 0.04 to 2.55% of the single-source samples' average read depth across the mitochondrial genome. These results are well-below and are in-line with the use of a 0.10 point heteroplasmy threshold for making variant calls. The haplotypes generated for the positive controls in this study were concordant with the NIST standard data described in Riman et al. [43] except for the 1393G/A sequence variant call. The heteroplasmic 1393G/A sequence variant, described by Riman et al. [43], did not reach the 0.10 point heteroplasmy threshold set for this study. The "A" allele was seen at 3% (Ion PGM run) and 4% (Ion S5 run) of the total read depth at that nucleotide position. Differences may be due to sequencing chemistry or variation that can occur with different lots of a cell line.

Single-source reference samples

Haplotype calls for the single-source reference samples were compared to complete mitochondrial genome sequence data generated by long-PCR and sequenced on the MiSeq or PGM. The haplotype calls were completely concordant and then used as references for assessing mixture deconvolution.

Performance metrics of the six single-source samples and two positive controls were evaluated. Average read depth for the eight samples ranged from 368X to 22,188X across the mitochondrial genome. Samples sequenced on the Ion PGM had a slightly lower average read depth, which ranged from 270X to 18,836X, than those samples sequenced on the Ion S5, which ranged from 366X to 24,224X. This difference in average read depth is attributed to the difference in throughput capacities of the Ion Chips used on the two instruments. Thus, an average RLP was calculated to normalize the two sequencing runs and to visualize relative sequencing performance across the mitochondrial genome. Average RLP for the single-source samples ranged from 5.93E-05 to 3.50E-04 (Supplementary Fig. 1). Additionally, read depth across the mitochondrial genome was analyzed on a per strand basis to evaluate balance. Average ratios of read depth for the positive strand ranged from 0.02 to 0.75 with 84% of the nucleotide positions at or above 0.40 (Supplementary Fig. 2). Finally, the level of noise across the mitochondrial genome for the single-source samples was evaluated. Any reads not attributed to nominal allele calls were considered noise. These noise reads potentially could be the result of sequencing errors, PCR errors, alignment errors, NUMTs, or contamination. Average noise for the single-source genomes ranged from 0.0% of the total read depth to 4.86% of the total read depth across the mitochondrial genome, with only eight nucleotide positions above 3% (Supplementary Fig. 3). The nucleotide positions where noise was the highest were scrutinized further. These nucleotide positions (e.g., nucleotide position 13057) generally were associated with homopolymeric regions in the genome. Ion Torrent platforms' difficulty in sequencing through homopolymeric regions has been well-characterized [9, 18, 35, 44, 45], and Supplementary Fig. 4 illustrates the variation in reads that is generated and aligned to a homopolymer. Bioinformatic improvements (e.g., improvements to the alignment of homopolymers) may allow for better characterization of noise or off-target reads potentially allowing the thresholds for which point heteroplasmies and mixtures are called to be lowered.

Mixed samples

Performance metrics

For the mixed samples, average read depth ranged from 401X to 17,466X across the mitochondrial genome. Average RLP for the mixed samples ranged from 7.86E-06 to 3.47E-04 (Supplementary Fig. 5). When evaluating read depth on a per strand basis, average ratios of read depth for the positive strand ranged from 0.02 to 0.71 with 82% of the nucleotide positions at or above 0.40 (Supplementary Fig. 6). The level of reads not attributed to nominal allele calls (i.e., noise) ranged from 0.0 to 4.54% of the total read depth across the mitochondrial genome, with only seven nucleotide positions above 3.0% (Supplementary Fig. 7). The nucleotide positions with the highest level of noise were generally associated with homopolymeric regions in the genome (e.g., nucleotide position 13057). These performance metrics for mixtures were similar to those for single-source samples.

Mixture interpretation

The quantitative MPS data, phasing, and phylogenetics were used to deconvolve mixtures. Mixture data were analyzed quantitatively by calculating the ratio of each allele's read depth to total read depth. These ratios were used to group the sequence variants in each mixed haplotype into three groups: (1) alternate allele present in both contributors' haplotypes, (2) alternate allele present in the major contributor's haplotype, and (3) alternate allele present in minor contributor's haplotype. The "major" versus "minor" designation was decided by assigning the sequence variants with the higher ratio of alternate allele read depth to total read depth as the "major contributor." No specific ratio was selected a priori as no criteria were available to set mixture ratios. Phasing and phylogenetic information were used when sequence variants did not fall into one of the three categories. During manual verification of haplotype calls in IGV, phasing information was collected from amplicons in which two or more nucleotide positions showed evidence of a mixture. Phylogenetic information was acquired from HaploGrep and EMPOP [8, 41, 42].

Two-person mixtures

Two-person mixtures of individuals from differing phylogenetic backgrounds (i.e., haplogroups HV and F1a1a) were analyzed first to assess the bioinformatic processes' ability to parse mixtures of mitochondrial haplotypes with a relatively large amount of genetic differences between the contributors. The quantitative analysis results for each mixture (1:1, 5:1, 10:1, and 20:1) are shown in Table 1. Note that a 1:1 mixture can result as a major:minor because input DNA was based on nuclear DNA amounts in this study. The amount of mitochondrial DNA per individual is related to total DNA but does vary among individuals [1, 46]. While the major contributor's haplotype was fully and accurately identified quantitatively for each mixture, sequence variants associated with the minor contributor's haplotype only were identified above the 0.1 point heteroplasmy threshold in the 1:1, 1:5, 5:1, and 1:10 mixtures, with the 5:1 mixture exhibiting only a partial minor contributor haplotype and the personal point heteroplasmy (14386T/C) in the minor contributor of the 1:1 mixture falling below the 0.1 point heteroplasmy threshold.

Personal point heteroplasmies were difficult to assess as they were low or could be attributed to a minor contributor. In instances (i.e., the 1:5 and 1:10 mixtures) when phase and phylogenetics did not identify whether the 14386T/C was a personal point heteroplasmy from the major contributor or a mixed site with the sequence variant belonging to the minor contributor's haplotype, two possible haplotypes for both the major and minor contributor (four haplotypes in total) were generated. However, phasing and phylogenetic information did help resolve some mixtures (Supplementary Fig. 8a–g). For example, the 4086C/T and 4092G/A point mixtures in the 1:1 mixture had an alternate allele frequency of 48%. These mixture sites could not be assigned with confidence to one of the three categories of (1) alternate allele present in both contributors' haplotypes, (2) alternate allele present in the major contributor's haplotype, or (3) alternate allele present in minor contributor's haplotype (Table 1). This lack of success to parse contributors quantitatively for essentially 1:1 mixtures is expected. Manual verification of the mixture positions in IGV offered additional information as the two sites resided within one amplicon, and the two alternate alleles at these sites were not in-phase with each other (i.e., not sequenced in the same read; Fig. 1). Therefore, in this amplicon, more information about each contributor's genetic profile could be obtained, despite similar read depth for both allele states at the mixture sites. Furthermore, both EMPOP and HaploGrep were used to phylogenetically confirm (or refute) the blind phasing assignments [8, 41, 42]. These tools indicated whether or not each sequence variant would be expected to occur in this haplotype. Alignment issues with indels, reads not making it all the way through an amplicon in one direction, and differing amplification efficiencies can increase the variance for the ratio of allele read depths to the total read depth and thus affect the ability to accurately assign sequence variants to one contributor or another. While such assessments were performed manually herein, anticipated bioinformatic developments (as these applications are increasing rapidly) could facilitate interpretation and improve the ability to parse mitochondrial DNA mixtures.

Given the success of deconvolving two-person mixtures with different haplogroups, two-person mixtures of individuals from the same U2e subclade were analyzed to assess the bioinformatic pipeline's ability to parse mixtures of

Table 1 Quantitative results of two-person mixtures with contributors of different haplogroups (and ancestries). The average for the alternate allele read depth as a ratio to total read depth for each category is provided. The standard deviation is in parentheses

	1:1 mixture	1:5 mixture	5:1 mixture	1:10 mixture	10:1 mixture	1:20 mixture	20:1 mixture
Both contributors	99.50% (0.76%)	99.38% (0.92%)	99.63% (0.74%)	98.44% (3.24%)			
Major contributor	62.38% (2.00%)	73.23% (5.14%)	89.44% (1.81%)	85.73% (1.82%)	96.94% (2.99%)	94.00% (3.24%)	98.12% (1.58%)
Minor contributor	35.96% (3.04%)	24.00% (2.00%)	11.27% (0.90%)	12.86% (1.21%)			

mitochondrial haplotypes with less genetic differences between the contributors. The quantitative analysis results for the 1:1 and 5:1 mixtures are shown in Table 2. This quantitative assessment was able to identify full and accurate haplotypes for the major and minor contributor in each mixture, similar to the results above. Additional phasing and phylogenetic information were not needed to resolve this set of mixed samples (Supplementary Fig. 9a–c). Greater major

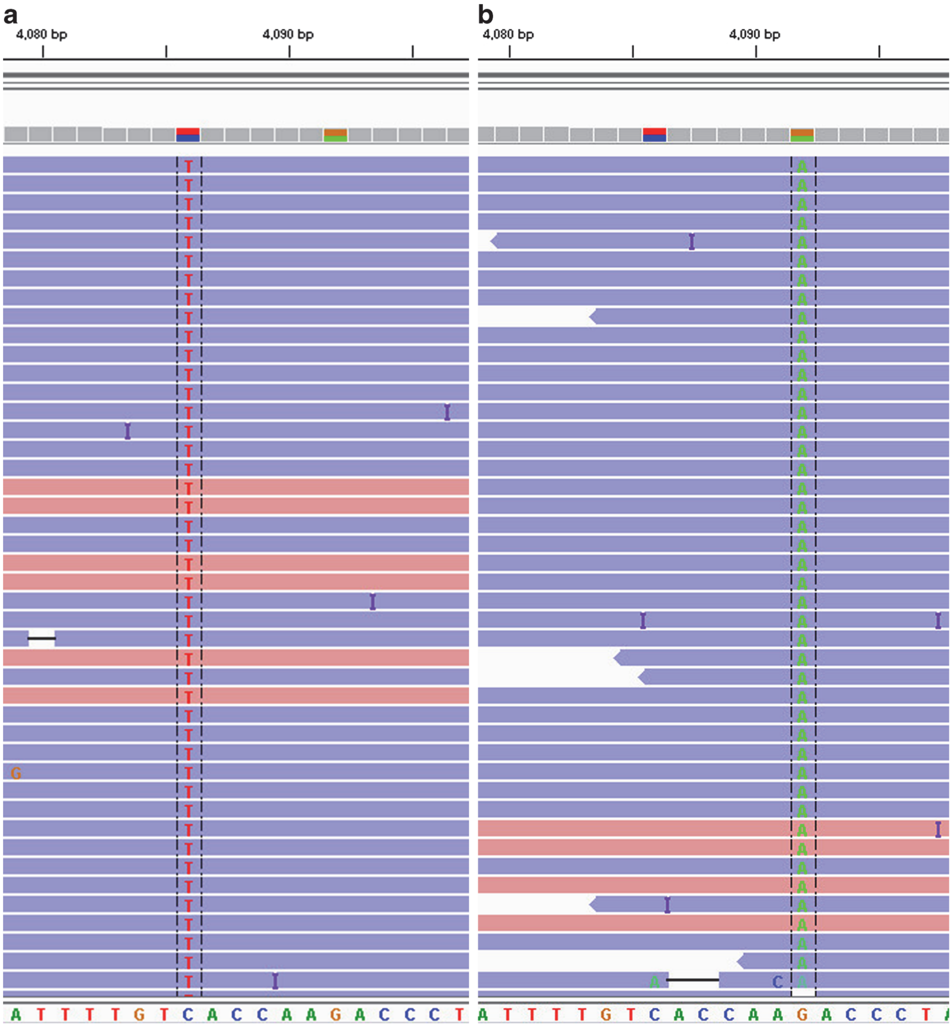

a
4,080 bp
4,090 bp
b
4,080 bp
4,090 bp
A T T T T G T C A C C A A G A C C C T
A T T T T G T C A C C A A G A C C C T A

Fig. 1 Viewing the 1:1 mixture's haplotype in IGV. Sorting alignments by base illustrated that the alternate alleles at 4086C/T and 4092G/A were not in-phase with each other (i.e., T and A, respectively were not in the same read)

Table 2 Quantitative results of two-person mixtures with contributors of the same U2e subclade. The average for the alternate allele read depth as a ratio to total read depth for each category is provided. The standard deviation is in parentheses

	1:1 mixture	5:1 mixture	1:5 mixture
Both contributors	99.09% (0.93%)	99.13% (0.94%)	99.06% (1.39%)
Major contributor	59.00% (2.53%)	85.33% (1.75%)	76.50% (3.94%)
Minor contributor	38.80% (1.75%)	12.17% (1.33%)	21.33% (1.21%)

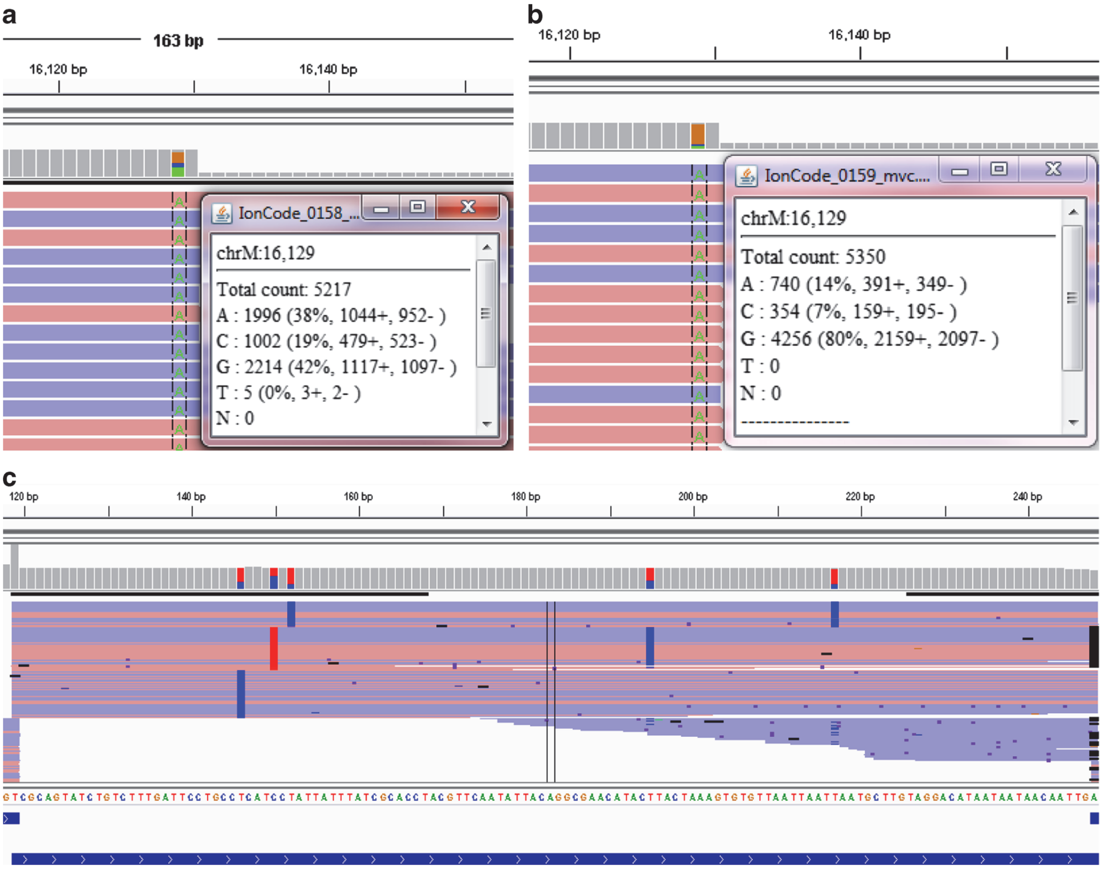
versus minor contributor ratios (i.e., 10:1 and 20:1 ratios) were not attempted based on the mixture findings discussed above.

Three-person mixtures

Mixtures of 1:1:1 and 5:1:1 ratios of individuals from different phylogenetic backgrounds (i.e., haplogroups HV, F1a1a, and U2e2a1) were generated. Phasing information and the presence of tri-allelic nucleotide positions suggested that the number of contributors was greater than two for these mixtures (Fig. 2), which could be more challenging if all three contributors were of the same haplogroup. However, quantitative assessment of the three-person mixtures did not provide a clear delineation of the alternate allele ratios into one of the three groups of (1) alternate allele present in each contributors' haplotypes, (2) alternate allele present in the major contributor's haplotype, or (3) alternate allele present in minor contributor's haplotype where complex mixtures can have more than one minor contributor (Supplementary Fig. 10a, b). While the tri-allelic nucleotide position and amplicons where all three haplotypes were observed (Fig. 2) provided an indication for assessing the quantitative contribution of each individual in the mixture, the amount of shared variants and large number of possible combinations for parsing each sequence variant in the mixture was not feasible with manual deconvolution of the 1:1:1 mixture. Phasing only provided additional information for a small number of amplicons in this mixture (Fig. 2c), but this phasing information should be considered as it could help exclude some individuals from the mixture. With the 5:1:1 mixture, quantitative assessment allowed the shared alleles present in each contributors' haplotype to be identified at a ratio of 99.38% (± 0.74%) and the alternate allele's present in the major contributor's haplotype to be identified at a ratio of 74.78% (±1.92). The range of alternate allele ratios seen for the two minor contributors was too similar to parse manually. The remaining sequence variants that were attributed to the minor contributors were uploaded to EMPOP for a phylogenetic assessment [8]. A haplogroup prediction (F1a1a) for one of the minor contributors was obtained. The sequence variants labeled as "Private Mutations" were deemed "inconclusive" prior to comparison with single-source reference samples as these variants also could have come from the second minor contributor. This phylogenetic assessment provided an accurate haplogroup prediction and an accurate, partial (77.5% complete) haplotype for one of the minor contributors. The remaining seven sequence variants could have been private mutations from minor contributor one or part of minor contributor two's haplotype, but with so many of minor contributor two's sequence variants likely falling below the 0.1 point heteroplasmy threshold, it was difficult to take advantage of any additional phylogenetic assessment. As discussed previously, alignment issues with indels, reads not making it all the way through an amplicon in one direction, and differing amplification efficiencies can increase the variance for the ratio of the alternate allele's read depth to the total read depth and affect the ability to accurately assign sequence variants to one contributor or another. However, nucleotide positions that can be phylogenetically associated to a contributor could be reported with a probability of a profile given certain genotypes. Likely, since MPS data for the most part are quite quantitative, a probabilistic genotyping approach could perform better at parsing contributors [47–50]. Vohr et al. [51] have released a software package for the analysis of mixed mitochondrial DNA samples called mixemt. This software provides a more automated approach to the quantitative and phylogenetic assessment completed manually in this study. However, use of a PCR-based amplification of the mitochondrial genome and current data capacities of Linux-based systems available for this study rendered this software package, in its current state, ineffective for analysis of previously discussed mixtures.

Evaluating number of contributors

The performance metrics, read depth, RLP, strand balance, and noise, of the single-source and mixed samples were compared to search for the presence of trends that potentially could help distinguish between single-source and mixed samples. The results for each of the performance metrics for both groups of samples were found to be comparable. The range of read depth and RLP across the mitochondrial genome was comparable for both the single-source and mixed samples. The topography of the RLP graphs (Supplementary Figs. 1 and 5) displayed similar valleys and peaks and illustrated that the higher and lower performing amplicons were the same for both the single-source and mixed samples. Strand balance and the number of reads attributed to noise also were comparable between the single-source and mixed samples. The nucleotide


a
163 bp
16,120 bp
16,140 bp
IonCode_0158_...
b
16,120 bp
16,140 bp
IonCode_0159_mvc....
c
120 bp
140 bp
160 bp
180 bp
200 bp
220 bp
240 bp
GTCGCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATAACAATTGA

chrM:16,129

Total count: 5217

A : 1996 (38%, 1044+, 952-)

C : 1002 (19%, 479+, 523-)

G : 2214 (42%, 1117+, 1097-)

T : 5 (0%, 3+, 2-)

N : 0

chrM:16,129

Total count: 5350

A : 740 (14%, 391+, 349-)

C : 354 (7%, 159+, 195-)

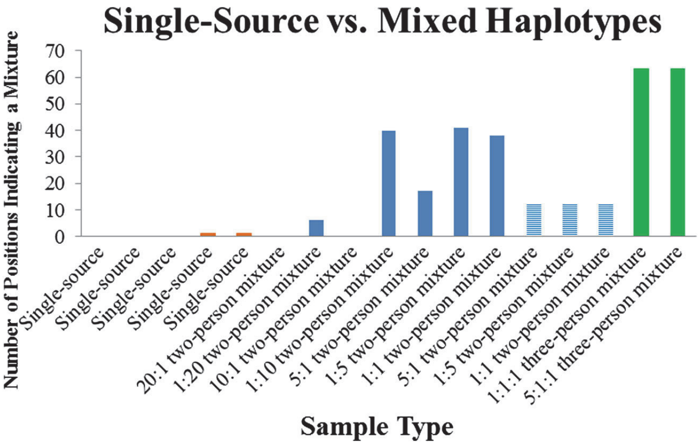
G : 4256 (80%, 2159+, 2097-)

T : 0

N : 0

Fig. 2 Images taken from IGV of nucleotide position 16129 in the 1:1:1 (**a**) and 5:1:1 mixtures (**b**) where three different alleles from the three different contributors are present. Figure 2**c** illustrates an amplicon in the 1:1:1 mixture where sorting by base in IGV allows visualization of the three haplotypes of the three-person mixture

Fig. 3 The number of positions in the sequence data indicating a mixture in each single-source and mixed sample analyzed in this study. A gradual increase in the number of mixture positions is seen from single-source to more complex mixtures. A striped fill pattern is used to indicate mixtures with contributors of similar phylogenetic backgrounds



positions with the highest level of noise across the mitochondrial genome associated with homopolymeric regions in both the single-source and mixed samples.

The final comparison evaluated the number of mixture sites (or point heteroplasmies for single-source samples). King et al. [5] provided the pairwise nucleotide differences between and among haplotypes from three major US population groups. Large population studies such as these provide necessary baseline information on the number of differences that exist between samples of different and similar phylogenetic backgrounds and concomitantly the number of positions that would indicate the presence of a mixture and the potential number of contributors of a mixed sample. Thus, plotting the number of mixture positions provides insight of the potential to predict the number of contributors in a mixed sample (Fig. 3). As expected, an increase in the number of mixture sites (point heteroplasmies for single-source samples) occurs from single-source samples to more complex mixtures. The two-person mixtures exhibited a greater range of the number of mixture positions which can be attributed to two explanations: (1) the 10:1, 1:20, and 20:1 mixtures' minor contributors were not detected above the 0.1 point heteroplasmy threshold, and thus, present more similarly to that of single-source samples and (2) two-person mixtures with contributors from the same haplogroup subclade have fewer differences (on average) between the two individuals than mixtures of individuals from different haplogroups. Despite these confounding factors, the number (i.e., in actuality the distribution of number) of positions indicating a mixture may be a good indicator of the number of contributors (at least up to three) in a mixed sample.

Conclusions

MPS offers the potential for analyzing mixtures using mtDNA sequence data. This study demonstrated, in a similar manner to that of STR typing, that a quantitative approach (i.e., the ratio of alternate alleles to total read depth) can be used to properly assign allele states to major and minor contributors. Qualitatively, phasing information (i.e., multiple SNPs residing within one amplicon) and well-characterized phylogeny of the mitochondrial genome can assist in mixture deconvolution.

Analysis was effective in identifying the major contributor in two-person mixtures with nuclear DNA ratios of 1:1, 5:1, 10:1, and 20:1. SNPs associated with the minor contributor were identified in the 1:1, 5:1, and 10:1 mixtures. For the more complex three-person mixtures, parsing was more difficult but likely can be improved substantially with additional studies and a probabilistic genotyping approach. These results indicate that MPS-based approaches that sequence mitochondrial DNA may be applicable to mixture interpretation compared to analysis with current CE technologies. With continued bioinformatic developments, mitochondrial DNA mixture analysis will become more robust and could become more routine for analysis of challenging samples.

Acknowledgements We would like to thank Thermo Fisher Scientific for providing reagents and technical support necessary to perform this study.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Robin ED, Wong R (1988) Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. J Cell Physiol 136:507–513
2. Giles RE, Blanc H, Cann HM, Wallace DC (1980) Maternal inheritance of human mitochondrial DNA. Proc Natl Acad Sci 77:6715–6719
3. Guevara EK, Palo JU, Guillen S, Sajantila A (2016) MtDNA and Y-chromosomal diversity in the Chachapoya, a population from the Northeast Peruvian Andes-Amazon divide. Am J Hum Biol 28: 857–867
4. Just RS, Scheible MK, Fast SA, Sturk-Andreaggi K, Rock AW, Bush JM, Higginbotham JL, Peck MA, Ring JD, Huber GE, Xavier C, Strobl C, Lyons EA, Diegoli TM, Bodner M, Fendt L, Kralj P, Nagl S, Niederwieser D, Zimmermann B, Parson W, Irwin JA (2015) Full mtGenome reference data: development and characterization of 588 forensic-quality haplotypes representing three U.S. populations. Forensic Sci Int Genet 14:141–155
5. King JL, LaRue BL, Novroski NM, Stoljarova M, Seo SB, Zeng X, Warshauer DH, Davis CP, Parson W, Sajantila A, Budowle B (2014) High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int Genet 12:128–135
6. Lopopolo M, Borsting C, Pereira V, Morling N (2016) A study of the peopling of Greenland using next generation sequencing of complete mitochondrial genomes. Am J Phys Anthropol 161: 698–704
7. Park S, Cho S, Seo HJ, Lee JH, Kim MY, Lee SD (2017) Entire mitochondrial DNA sequencing on massively parallel sequencing for the Korean population. J Korean Med Sci 32:587–592
8. Parson W, Dur A (2007) EMPOP-A forensic mtDNA database. Forensic Sci Int Genet 1:88–92
9. Parson W, Strobl C, Huber G, Zimmermann B, Gomes SM, Souto L, Fendt L, Delport R, Langit R, Wootton S, Lagace R, Irwin J (2013) Evaluation of next generation mtGenome sequencing using the ion torrent personal genome machine (PGM). Forensic Sci Int Genet 7:632–639
10. Irwin JA, Saunier JL, Niederstatter H, Strouss KM, Sturk KA, Diegoli TM, Brandstatter A, Parson W, Parsons TJ (2009) Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. J Mol Evol 68:516–527
11. Malyarchuk B, Litvinov A, Derenko M, Skonieczna K, Grzybowski T, Grosheva A, Shneider Y, Rychkov S, Zhukova O (2017) Mitogenomic diversity in Russians and poles. Forensic Sci Int Genet 30:51–56

12. Saunier JL, Irwin JA, Strouss KM, Ragab H, Sturk KA, Parsons TJ (2009) Mitochondrial control region sequences from an Egyptian population sample. Forensic Sci Int Genet 3:e97–e103
13. Irwin JA, Saunier JL, Beh P, Strouss KM, Paintner CD, Parsons TJ (2009) Mitochondrial DNA control region variation in a population sample from Hong Kong, China. Forensic Sci Int Genet 3:e119–e125
14. Boattini A, Castri L, Sarno S, Useli A, Cioffi M, Sazzini M, Garagnani P, De Fanti S, Pettener D, Luiselli D (2013) mtDNA variation in East Africa unravels the history of Afro-Asiatic groups. Am J Phys Anthropol 150:375–385
15. Chaitanya L, van Oven M, Brauer S, Zimmermann B, Huber G, Xavier C, Parson W, de Knijff P, Kayser M (2016) High-quality mtDNA control region sequences from 680 individuals sampled across the Netherlands to establish a national forensic mtDNA reference database. Forensic Sci Int Genet 21:158–167
16. Holland MM, McQuillan MR, O'Hanlon KA (2011) Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy. Croat Med J 52:299–313
17. Kim H, Erlich HA, Calloway CD (2015) Analysis of mixtures using next generation sequencing of mitochondrial DNA hypervariable regions. Croat Med J 56:208–217
18. Chaitanya L, Ralf A, van Oven M, Kupiec T, Chang J, Lagace R, Kayser M (2015) Simultaneous whole mitochondrial genome sequencing with short overlapping amplicons suitable for degraded DNA using the ion torrent personal genome machine. Hum Mutat 36:1236–1247
19. Parson W, Huber G, Moreno L, Madel MB, Brandhagen MD, Nagl S, Xavier C, Eduardoff M, Callaghan TC, Irwin JA (2015) Massively parallel sequencing of complete mitochondrial genomes from hair shaft samples. Forensic Sci Int Genet 15:8–15
20. Cho S, Kim MY, Lee JH, Lee SD (2017) Assessment of mitochondrial DNA heteroplasmy detected on commercial panel using MPS system with artificial mixture samples. Int J Legal Med. https://doi.org/10.1007/s00414-017-1755-7
21. Coble MD, Just RS, O'Callaghan JE, Letmanyi IH, Peterson CT, Irwin JA, Parsons TJ (2004) Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. Int J Legal Med 118:137–146
22. Parsons TJ, Coble MD (2001) Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome. Croat Med J 42:304–309
23. Bodner M, Iuvaro A, Strobl C, Nagl S, Huber G, Pelotti S, Pettener D, Luiselli D, Parson W (2015) Helena, the hidden beauty: resolving the most common West Eurasian mtDNA control region haplotype by massively parallel sequencing an Italian population sample. Forensic Sci Int Genet 15:21–26
24. Zhou Y, Guo F, Yu J, Liu F, Zhao J, Shen H, Zhao B, Jia F, Sun Z, Song H, Jiang X (2016) Strategies for complete mitochondrial genome sequencing on Ion Torrent PGM™ platform in forensic sciences. Forensic Sci Int Genet 22:11–21
25. Parson W, Gusmao L, Hares DR, Irwin JA, Mayr WR, Morling N, Pokorak E, Prinz M, Salas A, Schneider PM, Parsons TJ (2014) DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. Forensic Sci Int Genet 13:134–142
26. Stewart JEB, Aagaard PJ, Pokorak EG, Polansky D, Budowle B (2003) Evaluation of multicapillary electrophoresis instrument for mitochondrial DNA typing. J Forensic Sci 48:571–580
27. Davis C, Peters D, Warshauer D, King J, Budowle B (2015) Sequencing the hypervariable regions of human mitochondrial DNA using massively parallel sequencing: enhanced data acquisition for DNA samples encountered in forensic testing. Legal Med 17:123–127
28. Holland MM, Wilson LA, Copeland S, Dimick G, Holland CA, Bever R, McElhoe JA (2017) MPS analysis of the mtDNA hypervariable regions on the MiSeq with improved enrichment. Int J Legal Med 131:919–931
29. Lindberg MR, Schmedes SE, Hewitt FC, Haas JL, Ternus KL, Kadavy DR, Budowle B (2016) A comparison and integration of MiSeq and MinION platforms for sequencing single source and mixed mitochondrial genomes. PLoS One 11:e0167600. https://doi.org/10.1371/journal.pone.0167600
30. Li M, Schonberg A, Schaefer M, Schroeder R, Nasidze I, Stoneking M (2010) Detecting Heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. Am J Hum Genet 87:237–249
31. Qiagen (2012) QIAamp® DNA mini and blood mini handbook. Qiagen, Valencia
32. Thermo Fisher Scientific (2015) Quantifiler™ HP and Trio DNA Quantification Kits User guide. Revision E Thermo Fisher Scientific, Waltham
33. Thermo Fisher Scientific (2016) Precision ID panels with the Ion PGM™ System Application Guide. Revision A Thermo Fisher Scientific, Waltham
34. Thermo Fisher Scientific (2015) Ion PGM™ Hi-Q™ Chef Kit. Revision A Thermo Fisher Scientific, Waltham
35. Churchill JD, King JL, Chakraborty R, Budowle B (2016) Effects of the Ion PGM Hi-Q sequencing chemistry on sequence data quality. Int J Legal Med 130:1169–1180
36. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147
37. King JL, Sajantila A, Budowle B (2014) mitoSAVE: mitochondrial sequence analysis of variants in Excel. Forensic Sci Int Genet 12:122–125
38. Scientific Working Group on DNA Analysis Methods (SWGDAM). (2013) Interpretation guidelines for mitochondrial DNA analysis by forensic DNA testing laboratories http://media.wix.com/ugd/4344b0_c5e20877c02f403c9ba16770e8d41937.pdf
39. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV)high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192
40. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26
41. Kloss-Brandstatter A, Pacher D, Schonherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F (2011) HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat 32:25–32
42. Weissensteiner H, Pacher D, Kloss-Brandstatter A, Forer L, Specht G, Bandelt HJ, Kronenberg F, Antonio Salas A, Schonherr S (2016) HaploGrep2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res 44:58–63
43. Riman S, Kiesler KM, Borsuk LA, Vallone PM (2017) Characterization of NIST human mitochondrial DNA SRM-2392 and SRM-2392-I standard reference materials by next generation sequencing. Forensic Sci Int Genet 29:181–192
44. Seo SB, Zeng X, King JL, Larue BL, Assidi M, Al-Qahtani MH, Sajantila A, Budowle B (2015) Underlying data for sequencing the mitochondrial genome with the massively parallel sequencing platform Ion Torrent™ PGM™. BMC Genomics 16(Suppl1):S4
45. Bragg LM, Stone G, Margaret K, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput Biol 9:e1003031. https://doi.org/10.1371/journal.pcbi.1003031
46. Shay JW, Pierce DJ, Werbin H (1990) Mitochondrial DNA copy number is proportional to Total cell DNA under a variety of growth conditions. J Biol Chem 265:14802–14807
47. Bright J, Taylor D, McGovern C, Cooper S, Russell L, Abarno D, Buckleton J (2016) Developmental validation of STRmix™, expert

software for the interpretation of forensic DNA profiles. Forensic Sci Int Genet 23:226–239

48. Bright J, Taylor D, Curran JM, Buckleton JS (2013) Developing allelic and stutter peak height models for a continuous method of DNA interpretation. Forensic Sci Int Genet 7:296–304

49. Scientific Working Group on DNA Analysis Methods (SWGDAM). (2015) Guidelines for the validation of probabilistic genotyping systems. https://docs.wixstatic.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf

50. Moretti TR, Just RS, Kehl SC, Willis LE, Buckleton JS, Bright J, Taylor DA, Onorato AJ (2017) Internal validation of STRmix™ for the interpretation of single source and mixed DNA profiles. Forensic Sci Int Genet 29:126–144

51. Vohr SH, Gordon R, Eizenga JM, Erlich HA, Calloway CD, Green RE (2017) A phylogenetic approach for haplotype analysis of sequence data from complex mitochondrial mixtures. Forensic Sci Int Genet 30:93–105

Curriculum vitae

Personal data

Name:	Monika Stoljarova
Date of birth:	18.04.1990
Place of birth:	Estonia
Citizenship:	Estonian

Contact data

E-mail:	monika.stoljarova@gmail.com

Education

2015–…	Tallinn University of Technology, Faculty of Science, Department of Chemistry and Biotechnology, Chemistry and Gene Technology, PhD
2012–2015	Tallinn University of Technology, Faculty of Science, Department of Gene Technology, Gene Technology, MSc (*cum laude*)
2011–2011	IMC University of Applied Sciences Krems (Krems a. d. Donau, Austria), Medical and Pharmaceutical Biotechnology, exchange student
2009–2012	Tallinn University of Technology, Faculty of Science, Department of Gene Technology, Gene Technology, BSc (*cum laude*)
2006–2009	Tallinn Secondary Science School

Language competence

Estonian	Fluent
English	Fluent
Russian	Fluent

Professional employment

2013–2014	University of North Texas Health Science Center (Fort Worth, TX, USA), Visiting Scientist, 10.2013-10.2014
2013–2013	Human Research Institute (Weiz, Austria), Insternship, 06.2013-08.2013

Supervised dissertations

Eero, I. (2018). Captive bred and wild living European mink (Mustela lutreola) gut microbial community analysis. Master's thesis. Tallinn University of Technology, School of Science, Department of Chemistry and Biotechnology. Supervisors: Aaspõllu, A., Stoljarova, M.

Kuningas, K. (2016). Estimation of Genetic Diversity of Estonian and Finnish Flying Squirrel (Pteromys volans) Populations. Master's thesis. Tallinn University of Technology, Faculty of Science, Department of Gene Technology. Supervisors: Aaspõllu, A., Stoljarova, M.

Awards/Scholarships

2017	Development Fund, Tallinn University of Technology, Rickard Kruusberg "Adventure through studies" Scholarship!"
2013	Baltic-American Freedom Foundation, Professional Intership scholarship.
2011/ 2012	Rotalia Foundation (USA), "Professor, Dr Anne Jennings Smith and Gerhard Treuberg commemorative scholarship"

Publications

Churchill, J. D., Stoljarova, M., King, J. L., & Budowle, B. (2018). Massively parallel sequencing-enabled mixture analysis of mitochondrial DNA samples. Int J Legal Med, 132(5), 1263-1272. doi:10.1007/s00414-018-1799-3

Oversti, S., Onkamo, P., Stoljarova, M., Budowle, B., Sajantila, A., & Palo, J. U. (2017). Identification and analysis of mtDNA genomes attributed to Finns reveal long-stagnant demographic trends obscured in the total diversity. Sci Rep, 7(1), 6193. doi:10.1038/s41598-017-05673-7

Stoljarova, M., King, J. L., Takahashi, M., Aaspollu, A., & Budowle, B. (2016). Whole mitochondrial genome genetic diversity in an Estonian population sample. Int J Legal Med, 130(1), 67-71. doi:10.1007/s00414-015-1249-4

Ambers, A. D., Churchill, J. D., King, J. L., Stoljarova, M., Gill-King, H., Assidi, M., Abu-Elmagd M., Buhmeida A., Al-Qahtani M.,Budowle, B. (2016). More comprehensive forensic genetic marker analyses for accurate human remains identification using massively parallel DNA sequencing. BMC Genomics, 17(Suppl 9), 750. doi:10.1186/s12864-016-3087-2

Zeng, X., King, J. L., Stoljarova, M., Warshauer, D. H., LaRue, B. L., Sajantila, A., Patel J., Storts D.R., Budowle, B. (2015). High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Sci Int Genet, 16, 38-47. doi:10.1016/j.fsigen.2014.11.022

King, J. L., LaRue, B. L., Novroski, N. M., Stoljarova, M., Seo, S. B., Zeng, X., Warshauer D.H., Davis C.P., Parson W., Sajantila A., Budowle, B. (2014). High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int Genet, 12, 128-135. doi:10.1016/j.fsigen.2014.06.001

Marshall, P. L., Stoljarova, M., Larue, B. L., King, J. L., & Budowle, B. (2014). Evaluation of a novel material, Diomics X-Swab, for collection of DNA. Forensic Sci Int Genet, 12, 192-198. doi:10.1016/j.fsigen.2014.05.014

Marshall, P. L., Stoljarova, M., Schmedes, S. E., King, J. L., & Budowle, B. (2014). A high volume extraction and purification method for recovering DNA from human bone. Forensic Sci Int Genet, 12, 155-160. doi:10.1016/j.fsigen.2014.06.011

Conferences and workshops

International Society of Forensic Genetics 2019 (09.09-13.09, 2019, Prague, Czechia).

International Society of Forensic Genetics 2015 (09.09-13.09, 2019, Prague, Czechia). Preconference workshop. Kinship statistics using Familias and FamLink.

International Society of Forensic Genetics 2015 (09.09-13.09, 2019, Prague, Czechia). Preconference workshop. ISO/IEC 17025:2017.

International Society of Forensic Genetics 2015 (09.09-13.09, 2019, Prague, Czechia). Preconference workshop.
Y chromosome: YHRD, mixture interpretation, kinship, population differentiation.

ecSeq Bioinformatics 2nd Berlin Summer School (25.06-29.062018, Berlin, Germany). NGS Data Analysis.

International Society of Forensic Genetics 2017 (28.08-02.09, 2017, Seoul, South-Korea). Poster presentation.
Stoljarova, M., Jaagura, M., Nuhk, E., Aaspõllu, A. The massively parallel sequencing of bacterial 16s ribosomal rna gene from urine samples for forensic application.

International Society of Forensic Genetics 2015 (31.08-05.09, 2015, Krakow, Poland). Poster presentation.
Stoljarova, M., King, J. L., Churchill, D.J, Takahashi, M., Budowle, B. “Massively Parallel Sequencing of Multiplex Short Amplicons of mtDNA from Challenged Forensic Samples“.

International Symposium of Human Identification 2014 (29.09-02.10, 2014, Phoenix, AZ, USA). Poster presentation.
Stoljarova, M., King, J. L., Churchill, D.J, Budowle, B. “Massively Parallel Sequencing of Multiplex Short Amplicons of mtDNA from Challenged Forensic Samples”.

Elulookirjeldus

Isikuandmed

Nimi:	Monika Stoljarova
Sünniaeg:	18.04.1990
Sünnikoht:	Eesti
Kodakondsus:	Eesti

Kontaktandmed

E-post:	monika.stoljarova@gmail.com

Hariduskäik

2015–…	Tallinna Tehnikaülikool, Loodusteaduskond, Keemia ja biotehnoloogia instituut, geenitehnoloogia eriala, PhD
2012–2015	Tallinna Tehnikaülikool, Matemaatika-loodusteaduskond, Geenitehnoloogia instituut, geenitehnoloogia eriala, MSc (*cum laude*)
2011–2011	IMC University of Applied Sciences Krems (Krems a. d. Donau, Austria), Meditsiiniline ja farmatseutiline biotehnoloogia, vahetusõpilane
2009–2012	Tallinna Tehnikaülikool, Matemaatika-loodusteaduskond, Geenitehnoloogia instituut, geenitehnoloogia eriala, BSc (*cum laude*)
2006–2009	Tallinn Reaalkool

Keelteoskus

eesti keel	emakeel
inglise keel	kõrgtase
vene keel	kõrgtase

Teenistuskäik

2013–2014	Põhja-Texase ülikool, Terviseteaduse keskus (Fort Worth, TX, USA), külalisteadur, 10.2013-10.2014
2013–2013	Human Research Institute (Weiz, Austria), praktikant, 06.2013-08.2013

Juhendatud diplomitööd

Eero, I. (2018). Captive bred and wild living European mink (Mustela lutreola) gut microbial community analysis. Magistritöö. Tallinn Tehnikaülikool, Loodusteaduskond, Keemia ja biotehnoloogia instituut. Juhendajad: Aaspõllu, A., Stoljarova, M.

Kuningas, K. (2016). Estimation of Genetic Diversity of Estonian and Finnish Flying Squirrel (Pteromys volans) Populations. Magistritöö. Tallinna Tehnikaülikool, Matemaatika-loodusteaduskond, Geenitehnoloogia instituut. Juhendajad: Aaspõllu, A., Stoljarova, M.

Tunnustused/stipendiumid

2017	Arengufond, Tallinna Tehnikaülikool, Rickard Kruusbergi nimeline stipendium “Seikle õpingute läbi!”
2013	Baltic-American Freedom Foundation, Professionaalse praktika stipendium
2011/ 2012	Rotalia Foundation (USA), Professor, Dr Anne Jennings Smith’i ja Gerhard Treubergi nimeline mälestusstipendium

Publikatsioonid

Churchill, J. D., Stoljarova, M., King, J. L., & Budowle, B. (2018). Massively parallel sequencing-enabled mixture analysis of mitochondrial DNA samples. Int J Legal Med, 132(5), 1263-1272. doi:10.1007/s00414-018-1799-3

Oversti, S., Onkamo, P., Stoljarova, M., Budowle, B., Sajantila, A., & Palo, J. U. (2017). Identification and analysis of mtDNA genomes attributed to Finns reveal long-stagnant demographic trends obscured in the total diversity. Sci Rep, 7(1), 6193. doi:10.1038/s41598-017-05673-7

Stoljarova, M., King, J. L., Takahashi, M., Aaspollu, A., & Budowle, B. (2016). Whole mitochondrial genome genetic diversity in an Estonian population sample. Int J Legal Med, 130(1), 67-71. doi:10.1007/s00414-015-1249-4

Ambers, A. D., Churchill, J. D., King, J. L., Stoljarova, M., Gill-King, H., Assidi, M., Abu-Elmagd M., Buhmeida A., Al-Qahtani M.,Budowle, B. (2016). More comprehensive forensic genetic marker analyses for accurate human remains identification using massively parallel DNA sequencing. BMC Genomics, 17(Suppl 9), 750. doi:10.1186/s12864-016-3087-2

Zeng, X., King, J. L., Stoljarova, M., Warshauer, D. H., LaRue, B. L., Sajantila, A., Patel J., Storts D.R., Budowle, B. (2015). High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Sci Int Genet, 16, 38-47. doi:10.1016/j.fsigen.2014.11.022

King, J. L., LaRue, B. L., Novroski, N. M., Stoljarova, M., Seo, S. B., Zeng, X., Warshauer D.H., Davis C.P., Parson W., Sajantila A., Budowle, B. (2014). High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Sci Int Genet, 12, 128-135. doi:10.1016/j.fsigen.2014.06.001

Marshall, P. L., Stoljarova, M., Larue, B. L., King, J. L., & Budowle, B. (2014). Evaluation of a novel material, Diomics X-Swab, for collection of DNA. Forensic Sci Int Genet, 12, 192-198. doi:10.1016/j.fsigen.2014.05.014

Marshall, P. L., Stoljarova, M., Schmedes, S. E., King, J. L., & Budowle, B. (2014). A high volume extraction and purification method for recovering DNA from human bone. Forensic Sci Int Genet, 12, 155-160. doi:10.1016/j.fsigen.2014.06.011

Konverentsidel ja kursustel osalemine

International Society of Forensic Genetics 2019 (09.09-13.09, 2019, Prague, Czechia). Teaduskonverents.

International Society of Forensic Genetics 2015 (09.09-13.09, 2019, Prague, Czechia). Konverentsieelne kursus. Kinship statistics using Familias and FamLink.

International Society of Forensic Genetics 2015 (09.09-13.09, 2019, Prague, Czechia). Konverentsieelne kursus. ISO/IEC 17025:2017.

International Society of Forensic Genetics 2015 (09.09-13.09, 2019, Prague, Czechia). Konverentsieelne kursus.
Y chromosome: YHRD, mixture interpretation, kinship, population differentiation

ecSeq Bioinformatics 2nd Berlin Summer School (25.06-29.062018, Berlin, Germany). Bioinformaatika suvekool. NGS Data Analysis.

International Society of Forensic Genetics 2017 (28.08-02.09, 2017, Seoul, South-Korea). Teaduskonverents ja posterettekanne.
Stoljarova, M., Jaagura, M., Nuhk, E., Aaspõllu, A. The massively parallel sequencing of bacterial 16s ribosomal rna gene from urine samples for forensic application.

International Society of Forensic Genetics 2015 (31.08-05.09, 2015, Krakow, Poland). Teaduskonverents ja posterettekanne.
Stoljarova, M., King, J. L., Churchill, D.J, Takahashi, M., Budowle, B. "Massively Parallel Sequencing of Multiplex Short Amplicons of mtDNA from Challenged Forensic Samples".