Mitochondrial Genome Analysis Using Next Generation Sequencing for Forensic Applications

KIMBERLY STURK-ANDREAGGI
Dissertation presented at Uppsala University to be publicly examined in Room C8:321, Biomedicinskt centrum (BMC), Husargatan 3, Uppsala, Thursday, 19 May 2022 at 13:00 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English. Faculty examiner: Professor Antti Sajantila (Department of Forensic Medicine, University of Helsinki, Finland).

Abstract

Mitochondrial DNA (mtDNA) analysis plays a specialized role in forensic applications, overcoming certain limitations of autosomal DNA markers. The high copy number and uniparental inheritance pattern of mtDNA are advantageous in cases involving shed hairs and aged skeletal elements, especially decades-old missing persons cases. Though the discriminatory power of mtDNA is limited by common haplotypes, next generation sequencing (NGS) offers feasible access to entire mitochondrial genome (mitogenome) data that can provide increased resolution of common haplotypes to unique sequences. The primary implementation challenge of mitogenome analysis is a lack of forensic-quality reference data, which are required to determine the evidentiary weight of a match. A better understanding of NGS methods and data analysis is also necessary to ensure the generation of reliable mitogenome data. Furthermore, appropriate quality control (QC) measures must be established as analysis can be complicated by nuclear mtDNA segments (NUMTs), misalignment of homopolymer regions, sequencing errors, and other artefacts. Including such false variants in mtDNA haplotypes can lead to erroneous conclusions based on misinterpreted data.

This thesis aimed to address the implementation challenges of mitogenome analysis and facilitate the transition to NGS in forensic laboratories. **Paper I** assessed the feasibility of generating forensic-quality mitogenome data from whole genome sequencing (WGS) data, which are valuable sources of mitogenome haplotypes for population studies. Due to NUMT interference, a 10% variant frequency threshold was necessary to produce haplotypes consistent with high-quality mitogenome datasets. Since length heteroplasmy (LHP) can also complicate mtDNA data analysis, **Paper II** characterized LHP in data generated on two NGS platforms as well as with Sanger-type sequencing. Different patterns of LHP were observed across sequencing technologies, further supporting current guidelines to ignore LHP in database searches and match comparisons in forensic analyses. Phylogenetic information can provide a valuable QC check of mtDNA data, identifying errors like artificial recombination and phantom mutations. Therefore, three haplogrouping tools were examined in **Paper III**, comparing their ability to predict an accurate haplogroup based on different mitogenome target ranges. The tools performed similarly, but EMPOP’s SAM2 algorithm produced more precise haplogroup predictions than the other two tools regardless of phylogeny or interpretation range. Building upon the previous three studies, **Paper IV** characterized 934 forensic-quality Swedish mitogenomes from a population genetics perspective. The complete mitogenome data demonstrated high haplotype diversity (0.9996) with a random match probability of 0.15%. In summary, these papers combine important insights to facilitate the application of mitogenome NGS analysis in forensic laboratories.

**Keywords:** Mitochondrial DNA, next generation sequencing, massively parallel sequencing, forensic genetics, nuclear mitochondrial DNA segment (NUMT), length heteroplasmy, haplogroup, quality control, population data

Kimberly Sturk-Andreaggi, Department of Immunology, Genetics and Pathology, Medicinsk genetik och genomik, Rudbecklaboratoriet, Uppsala University, SE-751 85 Uppsala, Sweden.

© Kimberly Sturk-Andreaggi 2022

ISSN 1651-6206
URN urn:nbn:se:uu:diva-470324 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-470324)
To my sons,
Max and Luca
List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


Reprints were made with permission from the respective publishers.
Related papers


# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>11</td>
</tr>
<tr>
<td>Mitochondrial DNA</td>
<td>11</td>
</tr>
<tr>
<td>Disease association</td>
<td>13</td>
</tr>
<tr>
<td>Phylogenetics and biodiversity</td>
<td>14</td>
</tr>
<tr>
<td>Nuclear mitochondrial DNA segments</td>
<td>15</td>
</tr>
<tr>
<td>Biparental inheritance claims</td>
<td>16</td>
</tr>
<tr>
<td>Sanger-type sequencing</td>
<td>17</td>
</tr>
<tr>
<td>Forensic use of mitochondrial DNA</td>
<td>18</td>
</tr>
<tr>
<td>Match comparisons</td>
<td>19</td>
</tr>
<tr>
<td>Statistical weight</td>
<td>20</td>
</tr>
<tr>
<td>Mitochondrial genome analysis</td>
<td>22</td>
</tr>
<tr>
<td>Advantages</td>
<td>22</td>
</tr>
<tr>
<td>Limitations</td>
<td>23</td>
</tr>
<tr>
<td>Next generation sequencing</td>
<td>25</td>
</tr>
<tr>
<td>Platforms</td>
<td>25</td>
</tr>
<tr>
<td>Illumina sequencing</td>
<td>25</td>
</tr>
<tr>
<td>Ion sequencing</td>
<td>27</td>
</tr>
<tr>
<td>Long-read sequencing</td>
<td>28</td>
</tr>
<tr>
<td>Library preparation</td>
<td>29</td>
</tr>
<tr>
<td>Fragmentation</td>
<td>30</td>
</tr>
<tr>
<td>Indexes</td>
<td>31</td>
</tr>
<tr>
<td>Mitochondrial genome enrichment</td>
<td>33</td>
</tr>
<tr>
<td>Long-range amplification</td>
<td>33</td>
</tr>
<tr>
<td>Small-amplicon approaches</td>
<td>34</td>
</tr>
<tr>
<td>Single-primer extension</td>
<td>36</td>
</tr>
<tr>
<td>Hybridization capture</td>
<td>36</td>
</tr>
<tr>
<td>Data analysis</td>
<td>39</td>
</tr>
<tr>
<td>Software</td>
<td>39</td>
</tr>
<tr>
<td>Workflow parameters</td>
<td>40</td>
</tr>
<tr>
<td>Quality control</td>
<td>41</td>
</tr>
</tbody>
</table>
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>aDNA</td>
<td>Ancient DNA</td>
</tr>
<tr>
<td>AFDIL</td>
<td>Armed Forces DNA Identification Laboratory</td>
</tr>
<tr>
<td>AFMES</td>
<td>Armed Forces Medical Examiner System</td>
</tr>
<tr>
<td>AQME</td>
<td>AFDIL-QIAGEN mtDNA Expert</td>
</tr>
<tr>
<td>auDNA</td>
<td>Autosomal DNA</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence interval</td>
</tr>
<tr>
<td>codR</td>
<td>Coding region</td>
</tr>
<tr>
<td>COI</td>
<td>Cytochrome oxidase subunit I</td>
</tr>
<tr>
<td>CR</td>
<td>Control region</td>
</tr>
<tr>
<td>C-stretch</td>
<td>Polycytosine region</td>
</tr>
<tr>
<td>ddNTP</td>
<td>Dideoxynucleotide triphosphate</td>
</tr>
<tr>
<td>dNTP</td>
<td>Deoxynucleotide triphosphate</td>
</tr>
<tr>
<td>dsDNA</td>
<td>Double-stranded DNA</td>
</tr>
<tr>
<td>EDNAP</td>
<td>European DNA Profiling Group</td>
</tr>
<tr>
<td>EMPOP</td>
<td>EDNAP Mitochondrial Population database</td>
</tr>
<tr>
<td>FBI</td>
<td>Federal Bureau of Investigation</td>
</tr>
<tr>
<td>FFPE</td>
<td>Formalin-fixed paraffin embedded</td>
</tr>
<tr>
<td>HVS</td>
<td>Hypervariable segment</td>
</tr>
<tr>
<td>IGG</td>
<td>Investigative genetic genealogy</td>
</tr>
<tr>
<td>LHON</td>
<td>Leber's Hereditary Optic Neuropathy</td>
</tr>
<tr>
<td>LHP</td>
<td>Length heteroplasmy</td>
</tr>
<tr>
<td>LR</td>
<td>Long-range</td>
</tr>
<tr>
<td>Mitogenome</td>
<td>Mitochondrial genome</td>
</tr>
<tr>
<td>MM</td>
<td>Major (or dominant) molecule</td>
</tr>
<tr>
<td>MRCA</td>
<td>Most recent common ancestor</td>
</tr>
<tr>
<td>mtDNA</td>
<td>Mitochondrial DNA</td>
</tr>
<tr>
<td>nDNA</td>
<td>Nuclear DNA</td>
</tr>
<tr>
<td>NGS</td>
<td>Next generation sequencing</td>
</tr>
<tr>
<td>np</td>
<td>Nucleotide position</td>
</tr>
<tr>
<td>NUMT</td>
<td>Nuclear mtDNA segment</td>
</tr>
<tr>
<td>ONT</td>
<td>Oxford Nanopore Technologies</td>
</tr>
<tr>
<td>PacBio</td>
<td>Pacific Bioscience</td>
</tr>
<tr>
<td>PHP</td>
<td>Point heteroplasmy</td>
</tr>
<tr>
<td>QC</td>
<td>Quality control</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>rCRS</td>
<td>Revised Cambridge Reference Sequence</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal RNA</td>
</tr>
<tr>
<td>RSRS</td>
<td>Reconstructed Sapiens Reference Sequence</td>
</tr>
<tr>
<td>SBS</td>
<td>Sequencing-by-synthesis</td>
</tr>
<tr>
<td>SMRT</td>
<td>Single molecule real-time</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SPE</td>
<td>Single primer extension</td>
</tr>
<tr>
<td>ssDNA</td>
<td>Single-stranded DNA</td>
</tr>
<tr>
<td>STR</td>
<td>Short tandem repeat</td>
</tr>
<tr>
<td>STS</td>
<td>Sanger-type sequencing</td>
</tr>
<tr>
<td>UDI</td>
<td>Unique dual-index (or dual indexed)</td>
</tr>
<tr>
<td>UMI</td>
<td>Unique molecular identifier (or index)</td>
</tr>
<tr>
<td>UNTCHI</td>
<td>University of North Texas Center for Human Identification</td>
</tr>
<tr>
<td>U.S.</td>
<td>United States</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole genome sequencing</td>
</tr>
</tbody>
</table>
Introduction

Mitochondrial DNA

Mitochondrial DNA (mtDNA) is a small, circular genome located in the mitochondria, the “powerhouse” organelle of the cell (Figure 1). The mtDNA genome (mitogenome) has 16,569 base pairs (bps), compared to billions in the nuclear genome [1]. Also, in contrast to the nuclear genome, mtDNA is a haploid genome and is passed down from mother to child (i.e., maternally inherited). The mitogenome consists of a small, highly variable non-coding region that spans the genome’s origin (nucleotide positions [nps] 1-576 and 16024-16569), known as the control region (CR) or D-loop, and the coding region (codR) that encompasses nps 577-16023. The codR of the mitogenome contains 37 genes (Figure 1), including 13 protein-coding genes essential for oxidative phosphorylation and 24 genes that encode for either transfer or ribosomal RNA (rRNA) [1]. Several thousand copies of the mitogenome may be present in each cell, with multiple copies of mtDNA per mitochondrion, compared to only one or two of each chromosome [2]. There are approximately hundreds to thousands of mitogenome copies relative to each diploid version of the nuclear genome. The copy number of mtDNA varies across tissue types as those requiring more energy, like muscles, have more mitochondria and therefore more mtDNA copies [3,4]. In comparison, cells that require less power generated by the mitochondria, such as those in whole blood, have a lower mtDNA to nuclear DNA (nDNA) ratio (i.e., fewer mitogenomes to each nuclear genome) than energetically expensive tissues [4].

The mutation rate of mtDNA has been estimated to be $1.30 \times 10^{-8}$ [5] to $1.89 \times 10^{-8}$ [6] per base per year. A recent study by Cabrera et al. showed that the mtDNA mutation rate may be three-fold higher in modern populations than these previous estimates have indicated [7]. This may be due to the method employed for sequencing and/or the rate calculation. The region of the mitogenome also has an impact on this rate, with the CR mutation rate estimated to be ten-fold higher than the observed rate in the codR [8,9]. Depending on the region (CR or codR), the mtDNA mutation rate is ten to 100 times higher than the mutation rate of the nuclear genome [10]. Due to the relatively high mutation rate and the millions of mtDNA molecules in an individual, mitogenome sequences may vary within an individual, a single cell or even a single mitochondrion [11]. Heteroplasmy, a state in which mtDNA sequences are variable, is observed as sequence or length variation in the mtDNA.
molecules. Point heteroplasmy (PHP) occurs when there is a base difference between mtDNA molecules at a single position. This type of heteroplasmy is known to occur at CR hotspots, specifically nps 146, 152, 204, and 16093 [12]. By contrast, heteroplasmic hotspots have not been identified in the codR [13,14], which is thought to be due to the ten-fold lower mutation rate in the codR compared to the CR [8,9]. Length heteroplasmy (LHP) is observed when mtDNA molecules have different sequence lengths. LHP typically occurs in homopolymeric or repeat unit regions, especially the polycytosine regions (C-stretches) of the hypervariable segment (HVS) regions of the CR [11,15,16].

Figure 1. Circular mitochondrial DNA genome (mitogenome). There are 13 protein-coding genes (blue), 22 transfer RNA genes (yellow), two ribosomal RNA genes (orange), and the non-coding control region (green). The black dashed line indicates the origin of the mitogenome.

Many factors can impact the presence and detection of heteroplasmy in an individual. First, the transmission of heteroplasmy from mother to offspring is complicated as there is a genetic bottleneck during the production of oocytes [17]. A limited number of mtDNA molecules are transferred into each oocyte, which can alter the observed level of heteroplasmy from one generation to another [18]. Second, the frequency of a heteroplasmy (i.e., the ratio of one mtDNA molecule to another) can differ across, and even within, tissue types. Calloway et al. showed that muscle tissue had the most heteroplasmic positions [19], which is likely due to the greater number of mtDNA copies in muscle and thus more opportunities for sequence differences [4]. In the case of hair, the ratio between the mtDNA molecules that cause heteroplasmy often differs between hairs from the same individual [20,21]. Lastly, mutations can
occur throughout an individual’s lifetime and cause an increase in mtDNA heteroplasmy with age or disease [22-24]. It is possible that changes in mtDNA heteroplasmy (i.e., the number of PHPs or the ratio of mtDNA molecules) that are observed within an individual over time may also be impacted by fluctuations in mtDNA copy number, which can alter the proportions of mtDNA molecules. Like heteroplasmy, differences in mtDNA copy number are also known to occur due to tissue type, disease, age, and even sex [4,25-29].

Disease association

Within the codR, mutations in the mitogenome may result in non-synonymous amino acid changes. In many cases, these mutations occur in utero and are not conducive to life (i.e., fatal to the foetus). Studies have investigated mtDNA mutations and, as of 2022, nearly 1000 have been linked to various diseases according to MITOMAP [30,31]. Yet, only 10% are confirmed pathogenic mutations. MITOMAP is a valuable resource that reports published data on human mtDNA variation including disease-associated mtDNA variants. Since the primary function of the mitochondria is related to cellular metabolism (i.e., energy production through respiration/oxidative phosphorylation), it is not surprising that most mtDNA-associated diseases are metabolic in nature (e.g., Leigh’s Disease, Pyruvate Dehydrogenase Complex Deficiency). One example of a mitochondrial disease is Leber's Hereditary Optic Neuropathy (LHON), which is strongly associated with three codR mutations (G3460A, G11778A, T14484C) [32]. However, medical research has shown that mitochondrial disease is complex as it is impacted by the level at which disease-associated mutations are present, known as the “threshold effect” [32]. The severity of these diseases is associated with heteroplasmy and the frequency of the wild-type (non-pathogenic) molecules at specific positions in the mitogenome [33]. Furthermore, a mother and offspring(s) will likely show different ratios of mutant and wild-type base frequencies as the result of the bottleneck effect during oocyte production, and thus varying expressions of the associated mitochondrial disease [17]. In addition to variation in the sequence of the mtDNA, the number of mtDNA copies in an individual has also been linked to disease. For example, mtDNA depletion syndromes are defined by reduced mtDNA content in certain tissues due to mutations that impact mtDNA maintenance [34]. Additionally, studies have demonstrated a reduction of mtDNA copy number in individuals suffering from age-related disorders like Parkinson’s disease and Alzheimer’s disease [27,34,35]. Based on its association with numerous diseases and aging, the analysis of mtDNA plays a vital role in medical research.
Phylogenetics and biodiversity

The maternal inheritance pattern of mtDNA is an invaluable tool in population genetics and human migration studies. Because a mother and offspring share the same mtDNA haplotype that is inherited without undergoing recombination, lineages can be traced back to the most recent common ancestor (MRCA). Phylogenetic trees can be built to visualise the evolutionary relationship between haplotypes such as the PhyloTree mtDNA tree [36]. At the root of the mtDNA phylogenetic tree is a hypothetical woman commonly referred to as Mitochondrial Eve, to whom all current living humans are descendants (mtMRCA). The root (Mitochondrial Eve or mtMRCA) haplotype has been established as the Reconstructed Sapiens Reference Sequence (RSRS) based on the reassessment of human mtDNA phylogeny [37]. From the root, each branch of the mtDNA phylogenetic tree represents a group of mtDNA haplotypes that share mutations, known as a haplogroup. As the branches of the tree extend, subhaplogroups are refined by additional haplogroup-diagnostic mutations or reversions, which occur when a mutation reverts back to its ancestral base. Since these mutations are passed down the maternal lineage, the distribution of mtDNA haplogroups within contemporary populations provides information about human migration. Further, an individual’s maternal ancestry can be inferred based on insights into the geographic origins of matrilineal ancestors gained from the mtDNA haplogroup [30].

The application of mtDNA research extends beyond human studies due to the existence of mtDNA in nearly all eukaryotic cells [1]. The same characteristics of mtDNA that are useful for human evolutionary and phylogenetic studies are also essential for investigations involving other species. As in human studies, the region of the mitogenome targeted varies depending on the application. Due to its elevated mutation rate, the mtDNA CR may be sequenced and compared across individuals of the same species. However, interspecies variation studies typically target other regions of the mitogenome that are more conserved. One such example is the 16S rRNA gene, which is often used in order to characterize a broad range of species [38-40]. The sequencing of this region is pivotal in metagenomics, the study of the genetic diversity of microbial communities [40]. Metagenomics has made significant contributions in many fields including ecology, biodiversity, and medical research. Other conserved regions of the mitogenome are also employed for taxological research. The cytochrome oxidase subunit I (COI) gene is used to characterize various species as part of the Barcode of Life project, a public collection of reference sequences that is managed by the National Center for Biotechnology Information [41,42]. A “barcode”, which is typically a 650-bp region of the COI gene, is sequenced and submitted to GenBank [43] to be used for species identification purposes.
Nuclear mitochondrial DNA segments

Nuclear mtDNA segments (NUMTs), or nuclear pseudogenes, are portions of the nuclear genome that are homologous to the mitogenome [44]. NUMTs result from the incorporation of mtDNA into the nuclear genome, which has been theorised to occur through non-homologous end joining at double-strand breaks [45-47]. Due to the lower mutation rate of nDNA compared to mtDNA, NUMT sequences act as “molecular fossils” since the ancestral mitogenome haplotypes are preserved [48]. Older NUMTs are fixed in the human population due to their insertion prior to the divergence of humans and thus represented in the human reference genome [49]. In contrast, polymorphic NUMTs have not yet reached fixation. Polymorphic NUMTs were inserted over the past one million years, most in the last 100,000 years based on phylogenetic analyses [49]. To date, a majority (82.3%) of the 1090 documented NUMT sequences are less than 500-bp in length [50]. However, 41 of the NUMTs documented in [51] are 5000 bps or longer, including three that encompass the entire mitogenome (≥16,569 bps). NUMTs can complicate mtDNA analysis by co-aligning with mtDNA molecules, thus mimicking authentic heteroplasmy at positions where the NUMT and mtDNA sequences differ (Figure 2). Since NUMTs are typically quite divergent from modern mtDNA haplotypes due to their ancestral origins [49], the presence of multiple PHPs in a haplotype can indicate NUMT interference [14].

![Figure 2. Example of co-alignment of mitochondrial DNA (mtDNA) and nuclear mtDNA segment (NUMT) sequences. Mismatches between the mtDNA and NUMT reads are observed at nucleotide positions 16390, 16399, 16444, 16496, 16519, and 16527 (outlined in the black boxes). The NUMT reads are consistent with several known polymorphic NUMTs, such as HSA_NumtS_587 [49].](image-url)

The occurrence of multiple copies of the entire mitogenome tandemly inserted, termed a “mega-NUMT” [52], has been recently documented. The size and copy number of these mega-NUMTs further complicate the analysis of mtDNA data. In a study by Lutz-Bonengel et al. [53], a mixture of two modern
haplogroups (V and U4c1) was observed in both a blood and buccal swab sample from a healthy individual. Depending on the sample type, the two haplotypes were detected at different ratios. When additional tissues were tested, the U4c1 haplotype was only detected in samples that contained nDNA, whereas a single-source V haplotype was observed from tissues that contain little to no nDNA, such as hair shafts. Additional experiments supported the presence of a mega-NUMT, including the detection of only the U4c1 haplotype in mtDNA-depleted $\rho^o$ cells. Ultimately, the mega-NUMT hypothesis was confirmed using fluorescence in situ hybridization, which located approximately 50 copies of the U4c1 haplotype on chromosome 14 [53]. Due to their size, the detection of mega-NUMTs in an affected individual is unavoidable. However, mega-NUMTs are rare and do not pose a routine problem in mtDNA analysis. Regardless, it is important to consider the potential impact of mega-NUMTs, as well as NUMTs generally, when drawing conclusions from research studies involving mtDNA.

Several studies have identified specific diseases caused by NUMTs as the result of coding mitogenome portions inserted into the nuclear genome [54-58]. It is also plausible that increases in mtDNA heteroplasmy associated with aging and certain diseases could actually be due to NUMTs [59]. Lending further support to this possibility, the decrease in mtDNA copy number is also linked to age and disease [4,25-29], and reduced ratios of mtDNA to nDNA will increase the likelihood of NUMT interference. Mixed positions caused by variants associated with NUMTs can be misinterpreted as authentic mtDNA heteroplasmy and result in incorrect conclusions based on such data [14,60].

Biparental inheritance claims

It has long been accepted that mtDNA inheritance occurs without recombination with paternal mtDNA [61]. This belief is based on all mitochondria transmitted to the offspring being solely from the ovum. The reason is that only the head of the sperm, which contains exclusively nuclear genetic material (no mitochondria), is retained following fertilization. This, however, has been a matter of debate as several studies have presented evidence of biparental inheritance of mtDNA [62-64]. In the most recent study, Luo et al. examined three pedigrees, each with one individual suspected of suffering from mitochondrial disease [64]. All families showed a high level of heteroplasmy. The authors deduced that the minor haplotype was passed from the father to the offspring, though the father’s mitogenome was not analysed. This report, and the other previous attempts to demonstrate paternal co-inheritance of mtDNA, led to responses by the scientific community to dispute these assertions [65-69]. The alternative explanation for these mtDNA mixtures is that the heteroplasmic mutations were derived from NUMTs [70-72]. Since NUMTs are located on nuclear chromosomes, they can be passed down from father to offspring. During mtDNA analysis, these “paternal” NUMTs co-align with the
maternal mtDNA due to homology of the sequences, resulting in mixed haplotypes. This could be especially problematic in cases involving mega-NUMTs and modern haplogroups (opposed to ancestral haplogroups), as described by Lutz-Bonengel et al. [53]. Therefore, without additional experiments, maternal inheritance continues to be the central dogma for mtDNA [73].

Sanger-type sequencing

The analysis of mtDNA is performed most commonly through sequencing [74]. Historically, Sanger-type sequencing (STS), also known as dye-terminator sequencing, has been the preferred sequencing method for nearly 40 years. STS was developed by Frederick Sanger in the 1970s [75], for which he received the Nobel Prize in Chemistry in 1980. STS is similar to PCR, another Nobel Prize awarded method, which involves the amplification of a DNA target region using a DNA polymerase, primers and deoxynucleotide triphosphates (dNTPs) to create double-stranded DNA (dsDNA) replicates. The difference between amplification PCR and STS is that STS utilises only one primer and a small portion (10%) of fluorescently labelled dideoxynucleotide triphosphates (ddNTPs) to produce single-stranded DNA (ssDNA) copies of the template DNA. The ddNTPs are chain-terminator bases that prevent extension of the DNA target once these bases have been incorporated into the sequence. Since there is a limited amount of ddNTPs in the reaction, different lengths of fragments are produced. Detection of the labelled fragments is performed through electrophoresis in which fragments migrate at different rates depending on their size (i.e., smaller fragments migrate faster than larger fragments). As the fragments migrate through a gel or polymer, a laser excites the fluorophores bound to the ddNTP molecules and the light emitted is captured by a detector. The emission wavelength is then utilised to determine the base at a particular position in the sequence. Current instruments use automated electrophoresis through capillaries and specialized software that interprets the output to determine the DNA sequence. Although STS has been the gold standard for sequencing for decades, the method has limited throughput and sensitivity. Advances in genomics have led to a shift in DNA sequencing from STS to second-generation sequencing, also known as next generation sequencing (NGS). NGS offers high-throughput sequencing with increased sensitivity and costs that have rapidly decreased over the last 15 years [76].
Forensic use of mitochondrial DNA

The analysis of mtDNA is an important tool utilised by the forensic community, especially in cases that include specimens with little to no nDNA. Generally, autosomal short tandem repeat (STR) typing is preferred in forensic investigations due to the high power of discrimination [77,78]. Additionally, there has been recent emphasis on the use of high-density single nucleotide polymorphisms (SNPs) and their application to investigative genetic genealogy (IGG) [79-81]. However, mtDNA has numerous advantages over these more widely-used autosomal DNA (auDNA) markers. In many cases, the DNA is too degraded to yield informative STR data. Furthermore, the relatively high abundance of the mtDNA in compromised samples, such as skeletal remains and shed hairs, makes it more likely to produce data for the mtDNA locus than for auDNA markers [82-86]. Therefore, in spite of its limited discrimination power, mtDNA analysis is essential in forensic cases that involve nDNA-depleted specimens.

Another characteristic of mtDNA that is particularly useful for certain forensic applications is its inheritance pattern. Since individuals of the same maternal lineage possess identical mtDNA haplotypes (barring mutation), even a distant maternal relative can serve as a suitable reference for mtDNA comparison in missing persons cases. This is extremely valuable when an individual has been missing for decades and appropriate genetic references are unavailable for standard STR testing (i.e., self-reference or immediate family members). Similarly, mtDNA can extend the reach of genetic genealogy beyond that of autosomal SNPs. Forensic laboratories that are focused on missing persons cases, such as the Armed Forces Medical Examiner System’s Armed Forces DNA Identification Laboratory (AFMES-AFDIL) and the University of North Texas Center for Human Identification (UNTCHI), attempt to utilise all DNA modalities to assist in identification efforts. However, these cases often require mtDNA sequencing due to the poor quality of the samples as well as the generational gap between living family members (i.e., DNA reference donors) and missing loved ones. For example, the AFMES-AFDIL provides the DNA testing needed to support the identification of human remains for the United States (U.S.) military as well as other federal agencies. The primary mission of the laboratory is to assist in the fullest possible accounting of missing personnel from past and present military conflicts. Most cases submitted to the AFMES-AFDIL for DNA testing are at least 50 years old and
involve human remains that have been subjected to long-term exposure to the environment or undergone post-mortem chemical treatment. These insults render the DNA severely degraded and damaged with high proportions of exogenous DNA [87-90]. Therefore, mtDNA is the leading modality for most human identification efforts at the AFMES-AFDIL, with nDNA markers utilised when applicable. This approach to the DNA testing of human remains is employed by most forensic laboratories that process decades-old missing persons cases.

Match comparisons

Sequence data for a sample are aligned to the human mitochondrial reference genome, the revised Cambridge Reference Sequence (rCRS) [1,91]. The rCRS is utilised rather than the RSRS in order to maintain compatibility with mtDNA haplotypes generated for more than three decades with the rCRS notation [92,93]. The haplotype of a sample is reported as a list of differences between the rCRS and sample consensus sequence at specific positions in the mitogenome (e.g., 16519C 263G 315.1C, the most common Westeurasian CR haplotype; also notated as T16519C A263G 315.1C). Haplotypes generated for known (e.g., family reference, suspect) and unknown (e.g., unidentified human remains, hair from a crime scene) samples are compared to determine if the samples originated from the same source or maternal lineage. Analysis of mtDNA data cannot individualise samples as can be achieved with multi-locus autosomal STR or SNP markers. Due to the maternal inheritance pattern of mtDNA, the source of a sample may be from any individual within the same maternal lineage since they share a mtDNA haplotype.

According to the Federal Bureau of Investigation (FBI) Scientific Working Group on DNA Analysis Methods (SWGDAM), forensic mtDNA comparisons of CR haplotypes can result in three conclusions: 1) exclusion, 2) inconclusive, or 3) cannot exclude [94]. Not considering LHP, an exclusion is achieved if the sample haplotypes differ by two or more polymorphisms. The comparison is inconclusive if the haplotypes differ at a single position. If the haplotypes are identical, this is considered a match and the samples cannot be excluded as coming from the same source or from individuals within the same maternal lineage. Laboratories are encouraged to develop specific guidelines for the treatment of heteroplasmy. Typically, PHP and length variation (i.e., LHP) are ignored during comparisons due to heteroplasmic variation within tissues of the same individual [19-21] and individuals from the same family [8,95]. Therefore, any differences resulting from heteroplasmy are not considered. However, shared heteroplasmy may be a valuable characteristic in a case by strengthening the conclusion, as in the identification of the Romanov family and Sister Marija Krucifiksa Kozulić [96,97]. Additional modifications to these conclusion criteria can be made based on the case context in order to
account for possible increases in mutation, which may necessitate allowances for haplotype differences. Family studies, upon which these criteria are based, most often involve first degree relatives (i.e., mother-child, siblings) with only one or two meiotic events separating individuals in the pedigree [8]. However, in decades-old missing persons cases, comparisons may be performed between very distant maternal relatives, perhaps six or seven meiosis apart. The additional meiotic events and generational time will increase the possibility of mutations in the mtDNA. The consistency of haplotypes across multiple meiosis remains to be examined. Furthermore, it should also be noted that these exclusionary interpretation guidelines were established from HVS1/HVS2 and CR data [8]. Therefore, if regions of the mtDNA codR are utilised for analysis, modifications to the exclusion criteria may be necessary until the substitution rate of the extended target region has been characterized.

Statistical weight

Once an mtDNA match (i.e., cannot exclude) has been identified in a forensic case, the haplotype is queried in a population database in order to assess the statistical significance of the mtDNA evidence. Though several mtDNA population databases exist, the European DNA Profiling (EDNAP) Group's Mitochondrial Population database (EMPOP; [98]) is considered the most comprehensive source of forensic-quality population mtDNA data and therefore is recommended by the DNA Commission of the International Society for Forensic Genetics (ISFG) [99]. In Release 13 of EMPOP (28 November 2019), there are more than 48,000 mtDNA haplotypes from global populations and 38,361 of those are CR haplotypes. The haplotypes included in EMPOP undergo rigorous quality control (QC) checks in order to meet forensic standards. One particular QC tool utilised to evaluate submitted data is quasi-median network analysis [98], which is helpful in detecting potential artefacts [100]. To ensure accurate haplotype frequency estimates, EMPOP employs a string-based algorithm (SAM) to allow searches to be performed with no alignment bias. The use of SAM [101], and now SAM2 [102], converts the queried mtDNA haplotype into a string sequence and compares it to the sequences of the haplotypes stored in EMPOP. The string-based approach eliminates the possibility of missing identical sequences due to inconsistent indel alignment and rCRS-coded nomenclature. Following a database query, SAM2 will provide the phylogenetic alignment of the queried haplotype and present the nomenclature based on the predicted haplogroup, which is recommended by forensic guidelines [94,99]. The haplogroup estimation provided by EMPOP offers another tool for the QC of mtDNA data [102,103]. The phylogenetic analysis enabled with haplogroup prediction is useful for the identification of artificial recombination and other potential errors in mtDNA data [104-106]. Altogether, EMPOP provides forensic-quality reference data necessary for
haplotype frequency estimates as well as QC tools that assist in the production of only the highest-quality mtDNA haplotypes.

The result of a population database search is used to determine the match probability based on the frequency estimate of the haplotype. The SWGDAM guidelines recommend the use of a conservative statistical approach described by Clopper and Pearson [94,107]. This method reports the upper bound of a 95% confidence interval (CI) to avoid underestimates of the haplotype frequency in the population that would artificially inflate the rarity of the haplotype. However, the Clopper-Pearson calculation is heavily dependent on database size, especially for very rare or unique haplotypes (also known as “singletons”). For example, if the haplotype is not found in a database of \( n \) haplotypes, the reported haplotype frequency based on the 95% CI closely resembles \( 3/n \) [94]. Lineage markers, like mtDNA and Y-chromosomal markers, may benefit from alternative methods for calculating the match probability and likelihood ratio [108]. One such method is the \( \kappa \) (kappa) estimate that takes into account the fraction of singletons in the database as well as the database size [109]. For example, the internal mtDNA population database at the AFMES-AFDIL contains 24,223 CR haplotypes, of which 48% (11,572) are singletons. The likelihood ratio for a haplotype not observed in this database would be 50,465 using the kappa estimate whereas the Clopper-Pearson method would produce a likelihood ratio that is six times less (8074), corresponding to less support (very strong versus moderate, respectively [110]). Furthermore, databases of smaller size are typically used for haplotype frequency estimates as the result of searching only the appropriate geographic location and/or metapopulation based on the case context, consequently having a greater impact on the reported match probability. It is therefore important to employ a statistical approach that takes into account the composition of the database in order to appropriately determine the weight of the mtDNA evidence. The Discrete Laplace method is a model-based approach to estimate the haplotype frequency in a reference database, which is recommended for the analysis of Y-chromosomal markers [111]. This method clusters haplotypes based on the probability of relatedness according to a Discrete Laplace distribution [112]. Future statistical approaches may take into account mutation rates and database sampling that impact the evidential weight of mtDNA (and Y) haplotypes. The careful consideration of the applied statistical methods will allow forensic laboratories to provide a fair assessment of lineage marker data [108].
Mitochondrial genome analysis

Advantages

To date, mtDNA analysis in forensic casework has primarily focused on the two HVS regions of the CR [74,85,113,114]. Traditionally this has been accomplished using STS, which can be quite costly and labour-intensive. Therefore, targeting the HVS regions for analysis has been advantageous because it requires only 600 bps of the 16,569-bp mitogenome to be sequenced, yet still allows for a reasonable degree of inter-individual discrimination [115,116]. However, mtDNA analysis using only this small portion of the mitogenome often fails to provide the necessary discrimination due to common haplotypes. In fact, the most common Westeurasian (Caucasian) mtDNA HVS haplotype is reported to occur in roughly 7% of the U.S. population [117]. Therefore, mtDNA sequence data from individuals with this common haplotype cannot be segregated by the HVS sequence alone. As a result, the re-association of remains and human identification efforts can be hindered when common haplotypes are encountered in forensic casework. By interrogating the entire mitogenome, shared HVS haplotypes are nearly 100% resolved into unique mitogenome haplotypes within a population [118,119]. In forensic cases involving commingled remains, for example, mitogenome analyses may provide the necessary genetic information to resolve common HVS haplotypes and provide significant statistical weight in support of identification. Variation in the codR of the mitogenome has already been utilised to assist with resolving common mtDNA HVS haplotypes in specific forensic case scenarios, such as the identification of the Titanic’s unknown child [120].

Interrogation of only the HVS, or even the CR, also limits the ability to precisely determine the mtDNA haplogroup. Conversely, the use of the entire mitogenome enables a precise estimation of the mtDNA haplogroup. Accurate haplogrouping helps to inform the correct phylogenetic indel alignment and may be crucial to inferring maternal ancestry. In cases with no maternal references, maternal ancestry may provide investigative leads and help focus efforts to obtain appropriate references for particular missing individuals. For example, haplogroup assignments are regularly employed at the AFMES-AFDIL to determine if specimens with no match in the family reference database are from American casualties versus individuals native to the region from which the remains were recovered (e.g., Korea) or foreign combatants from the particular military conflict [121]. In many cases, these questions can be addressed with HVS data alone; however, the specific biogeographic origin based on the predicted mtDNA haplogroup often requires the use of entire mitogenome data to obtain a precise haplogroup. This issue is exemplified when attempting to discern between mtDNA haplogroup B4, an East Asian lineage, and its Native American subhaplogroup B2 [30,36,122]. A less precise haplogroup prediction of B4 could associate a set of remains with a
foreign national rather than an American missing servicemember of Native American ancestry (e.g., a Navajo code talker [121]), which could be discerned from the refined B2 haplogroup. The ability to determine the precise haplogroup with mitogenome data in these cases ensures that the most accurate and useful maternal ancestry information is provided to investigators.

Limitations

Ethical concerns related to the private medical information that might be revealed by analysis of the codR, particularly non-synonymous mutations, have posed a barrier to the implementation of mitogenome analysis in forensic casework [117,123]. Less than 40 mutations in the CR have reported disease associations compared to over 900 mutations observed in the codR [30]. Furthermore, none of the 96 confirmed pathogenic mtDNA mutations are located in the CR. Medical research has shown that most disease-associated mutations are heteroplasmic and only a few homoplasmic variants code for disease (such as LHON mutations) [32]. Thus, codR heteroplasmy appears to be the key to understanding mitochondrial disease, which may have profound implications for forensics if mitogenome data are to be used. A recent study evaluated the impact of filtering confirmed pathogenic mtDNA mutations on database searches using a dataset consisting of nearly 3000 forensic-quality haplotypes [124]. The authors obtained identical match statistics regardless of whether variants associated with disease were included in the sample haplotypes. Based on these findings, forensic laboratories can minimize the release of potentially sensitive data by filtering confirmed pathogenic mutations while also taking advantage of the increased discrimination power of the entire mtDNA locus.

The primary technical challenge to the implementation of mitogenomes in forensics laboratories is associated with match statistics. There is simply not enough forensic-quality mitogenome data to produce accurate haplotype frequency estimations that are needed to evaluate the weight of the DNA evidence. In comparison to the nearly 40,000 CR haplotypes in EMPOP (Release 13), less than 4500 forensic-quality mitogenomes are available. The use of STS has historically restricted the generation of mtDNA population databases to only the CR. STS of the mitogenome entails amplification of numerous small, overlapping regions due to the poor quality of DNA in degraded casework samples, thus requiring large amounts of precious evidentiary material. Alternative strategies such as custom-designed single-nucleotide base extension assays (i.e., Applied Biosystems’ SNaPshot), linear arrays, chip arrays, and mass spectrometry have all showed promise in assessing the codR variants [125-133]. However, each of these methods has been difficult to implement in a practicing forensic laboratory. Methods must be both extremely robust and sensitive in order to overcome the myriad of issues observed in mtDNA casework samples (e.g., low quantities of DNA, degradation, inhibition), and
as a result, the routine use of many of these methods is impractical. Additionally, many of these methods are homebrew assays, which introduce complicated QC factors (e.g., assay design and development). As such, the practical utility of these codR assays have prevented broad application of the mitogenome in forensic laboratories. However, the high-throughput capabilities of NGS technologies offer a feasible approach to mitogenome sequencing. NGS methods enable the efficient generation of large quantities of mitogenomes, addressing the need for forensic-quality reference data.
Next generation sequencing

Recent advances in DNA sequencing technologies have led to the use of NGS to generate large quantities of reliable sequence data in various research and clinical applications. Also known as massively parallel sequencing (MPS), NGS technologies can aid in the analysis of challenging forensic samples through the utilisation of short read lengths that are amenable to highly degraded and/or damaged samples and an increase in sensitivity with hundreds to thousands of reads covering the target region. Since mtDNA is the only locus routinely sequenced in forensic laboratories, it is not unexpected that NGS was first implemented in a forensic laboratory for mtDNA analysis [90]. Furthermore, studies have demonstrated that NGS offers a cost-effective and high-throughput method to quickly generate full mitogenome sequence data from high-quality samples [119,134-138]. Most forensic laboratories with in-house NGS capabilities have benchtop sequencers like the MiSeq FGx Sequencing System (Verogen, Inc.) or Ion GeneStudio S5 System (Thermo Fisher Scientific). Studies have demonstrated the reliability of NGS for mitogenome sequencing on both platforms as well as the concordance between NGS and STS data [137,139,140]. However, there are limitations to each NGS technology that must be accounted for to ensure accurate haplotype generation.

Platforms

Illumina sequencing

Illumina platforms utilise a technology called sequencing-by-synthesis (SBS) with reversible dye terminators [141,142]. Sequencing is performed on a flow cell, which is a glass slide with one or more microfluidic channels (or lanes). Attached to both sides of the flow cell are two types oligos (P5 and P7), one of which is complementary to the adapter added to DNA fragments during library preparation. This enables the hybridization of DNA molecules to the flow cell for sequencing. Cluster generation is performed on the sequencing platform and begins with the creation of complementary strands of the hybridized DNA molecules. Using bridge amplification, the DNA strands attached to flow cell are clonally amplified, creating millions of clusters across the flow cell with each cluster representing a single DNA molecule. The sequencing
primer anneals to the adapter region of the DNA templates and sequences extend with the addition of fluorescently labelled nucleotides. Due to blocking groups (i.e., reversible terminators) on all four dNTPs, only one base can be incorporated into the complementary strand, which is determined by the sequence of the DNA template. Next, the clusters are excited by a laser and the fluorescence is emitted from each newly added base. Images at various wavelengths are collected, which are used to determine the base call of the cluster for each cycle. The generation of high-quality sequence data is dependent on the optimal cluster density (i.e., the number of clusters per square millimetre on the flow cell) as appropriate spacing between clusters is needed to detect and differentiate the emitted fluorescent signal from each cluster [141]. Once the images have been taken, the incorporated nucleotides are deblocked and the fluorophores are cleaved, which is possible since the attached terminators are reversible [142]. This allows for the next base to be added to each complementary strand, continuing the extension of the DNA templates. This process is repeated for a designated number of cycles and determines the length of the read. For example, if 150 cycles of sequencing are performed, the reads generated will be 150 bps in length (as long as the DNA insert is 150 bps or longer). Paired-end sequencing is enabled with a second sequencing read of the complementary strand. Additional reads sequence the index(es), which are added during library preparation and allow for multiple sample libraries to be sequenced in a single run. After sequencing is complete, primary analysis processes the images and converts these data into base calls for each cluster. The base call files are then demultiplexed during secondary analysis and reads are sorted based on their unique index combination into FASTQ files [142]. The number of reads generated for each sample is dependent on the Illumina instrument, reagent kit, and multiplexing approach.

The Illumina MiSeq (including the forensic-specific FGx model from Verogen) leverages Illumina’s four-channel SBS chemistry, which means that the four nucleotides are labelled with different fluorophores and four images are collected during each sequencing cycle. Depending on the reagent kit, 12 to 25 million reads (24 to 50 million paired reads) with read lengths up to 600 bps can be generated on the MiSeq. The Illumina HiSeq X instrument is a production-scale sequencer with the ability to generate up to 6 billion 300-bp reads (12 billion paired reads) in a single run. This system has been discontinued and was replaced by the NovaSeq 6000 (Illumina), which has a higher throughput capacity (10 billion reads) and longer read lengths (500 bps maximum) compared to the HiSeq X. Additionally, the NextSeq 550 System (Illumina) is a benchtop sequencer, but offers an output of 400 million 300-bp reads, and now up to 1.2 billion reads with the newer NextSeq 2000 model. These platforms, as well as the Illumina MiniSeq System (output of 22 to 25 million reads), utilise two-channel SBS chemistry that allows for faster data generation by collecting only two images per cycle instead of four-channel SBS used by the MiSeq and HiSeq systems. The smallest and most affordable
of the Illumina platforms is the iSeq 100 System, which performs the Illumina SBS chemistry in the nanowells of a metal-oxide semiconductor chip. Though the throughput is less than the other Illumina platforms (four million reads), the cost and size of the system may be suitable for forensic laboratories. Ultimately, the ideal NGS platform will be determined by the specific application.

According to Loman et al. [143], base misincorporations are the most common error in Illumina SBS data, at a rate of 0.1 substitutions per 100 bases. Slightly higher MiSeq error rates were observed in a study that characterized human mtDNA data [144], ranging from 0.18 to 0.49 per 100 nucleotides with base misincorporations most often observed at cytosine positions in the rCRS. Not yet fully characterized, two-channel Illumina SBS data are expected to have a slightly different error model than four-channel MiSeq data due to the differences in the detection of incorporated bases. This is also true for the iSeq, as little to no systematic error rate studies have been published for these newer Illumina platforms. Though one recent study compared the iSeq and MiSeq for metabarcoding purposes, and the results demonstrated similar data quality between two Illumina instruments [145]. Prior to the use of a particular platform, the level and type of sequencing error should be assessed to ensure the data generated is suitable for the targeted region or genome.

Ion sequencing

Ion Torrent sequencing is also considered an SBS technology, but the detection of the base incorporated into the DNA template molecule during synthesis differs from the Illumina reverse terminator SBS approach. Ion Torrent technology uses a semiconductor chip, which has millions of microwells. Like Illumina sequencing, Ion sequencing begins with cluster generation. A single DNA fragment is attached to a bead using an Ion-specific adapter (P1) and then amplified using emulsion PCR until the bead is covered by millions of copies of the original template [141,142]. Millions of beads, each with a different DNA molecule represented, are washed across the chip and deposited into the wells, ideally with one bead per well. The wells are then filled with one of the four nucleotides. If the nucleotide, which is an unmodified dNTP, is complementary to the next base (or bases) in the template sequence, the nucleotide (or multiple nucleotides) will be incorporated into the synthesized sequence [142]. A hydrogen is released for each nucleotide added, changing the pH in the well. A sensor detects this pH change and converts it to voltage, which indicates that the particular nucleotide was incorporated and the base is called. The number of nucleotides incorporated is determined by the change in voltage. For example, if the voltage doubles, then two nucleotides were incorporated into the sequence and two identical bases are called. If no pH change is detected, then the nucleotide is not complementary to the next base in the DNA template strand and no base is called. A different nucleotide is washed over the chip every 15 seconds, and the process is repeated up to 1400
times depending on the DNA insert size and reagent kit. The length of Ion reads correlates to the number of flows, but it is not a one-to-one relationship as it is for Illumina sequencing. Ion read lengths are approximately two-thirds of the number of nucleotide flows performed (e.g., 150 flows will produce ~100-bp reads) [146]. One main drawback to this sequencing method is that paired-end data cannot be generated. However, the advantages of Ion sequencing are the relatively low cost and short run times, which are particularly appealing to forensic laboratories.

The Ion S5 is a benchtop platform like the MiSeq and the Ion 530 chip has a maximum output of 20 million 400- to 600-bp reads. Though the MiSeq allows for longer reads and higher output, up to 80 million reads can be produced with the recently released Ion 540 chip and the new Ion 550 chip generates shorter reads but with even higher throughput (as many as 130 million 200-bp reads). The shorter reads of Ion S5 chips limit the sequencing of STRs, but 200- to 400-bp reads are sufficient for sequencing the mitogenome. The Genexus Integrated Sequencer uses the GX5 chip (both Thermo Fisher Scientific), which has four lanes that each output 12-15 million reads (up to 60 million reads total) and can be run individually or altogether depending on the throughput required. Unlike the Ion S5, which requires off-instrument template preparation and chip loading, the Genexus allows for automated processing from library preparation through analysis to streamline the Ion sequencing workflow and reduce manual steps.

In contrast to the base misincorporation errors observed in MiSeq data, spurious indels are the predominant error type observed in Ion semiconductor sequence data, occurring primarily in homopolymeric regions [147]. Sequencing on an Ion Torrent Personal Genome Machine (PGM; Thermo Fisher Scientific), the discontinued precursor to the Ion S5, results in 1.5 indel errors for every 100 bases with a substantial reduction in accuracy (as low as 60%), especially stretches with six or more of the same base [143]. The elevated indel error rate is not unexpected given the semiconductor-based technology. When the same nucleotide is incorporated multiple times during a single flow, the magnitude of the voltage becomes inaccurate and the precise number of bases to call is difficult to determine. As a result of the indel errors associated with Ion sequencing, LHP is likely the most impacted aspect of mtDNA analysis due to its occurrence in homopolymeric regions [148,149]. Therefore, characteristics of the sequencing approach must be considered during analysis of NGS mtDNA data.

Long-read sequencing

Sequencing for forensic applications has begun to transition from STS to NGS over recent years, with short-read sequencing on Illumina and Ion platforms being the most widely-used NGS technologies. This is due to their ability to generate large amounts of highly-accurate sequence data at relatively low
costs. Beyond forensics, single molecule long-read sequencing (sometimes known as third-generation sequencing) has particular benefits for biomedical and clinical applications [150]. There are two major long-read technologies, single molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio) and nanopore sequencing from Oxford Nanopore Technologies (ONT) [151,152]. Sequencing reads ranging from 10 kbps to >1 Mbps can be generated directly from native DNA using these technologies [150]; however, the utility in forensics has been limited due to cost and accuracy. Long-read sequencing has historically been used for de novo genome assembly to scaffold regions of a genome that are unknown and therefore are not amenable to targeted short-read sequencing. SMRT and nanopore sequencing have also been utilised for the detection of complex variants associated with disease and the discovery of structural variants that provide a greater understanding of human genetic diversity [150]. In terms of forensic applications, PacBio SMRT sequencing was shown to be an effective tool for differentiating between monozygotic twins based on low-level variants in their mitogenomes [153]. There has also been forensic interest in the portable MinION devices from ONT that offer real-time analysis of sequence data [154]. These devices are field deployable since they are roughly the size of a USB flash drive (MinION) or smartphone (MinION Mk1B), which has potential for security and defence purposes. Several studies have focused on the assessment of the MinION for forensic applications, including nDNA (STRs and SNPs) [155-157] and mtDNA [158,159]. Although the results of these studies have been promising, the technology and analysis of long-read sequencing have not yet met the forensic standards required for human identification. Yet, the use of the MinION for non-human forensic applications has already been validated [160,161]. As long-read technologies continue to advance, the forensic community may further evaluate these sequencing approaches, especially given the portability of the MinION devices [154].

Library preparation
To enable sequencing on any of the NGS platforms, the DNA must be prepared in a particular manner. Library preparation entails the addition of adapter sequences to the end of DNA fragments that are required for sequencing. These adapters contain platform-specific sequences (i.e., the P5 and P7 sequences for Illumina or P1 sequence for Ion) and may include barcodes (or indexes) that allow for multiple samples and/or targets to be sequenced in a single run. Library preparation begins with the repair of the DNA ends to eliminate single-stranded overhangs and the addition of a phosphate group to the 3’ end of the dsDNA fragment. The next step, which in some kits is coupled with end repair, is the addition of an adenine to the 5’ end (A-tailing). Following end repair and A-tailing, a double-stranded adapter with a thymine tail is
ligated to each end of dsDNA fragments. The shape of this adapter can vary between kits, with Y-shaped adapters allowing for paired-end sequencing on Illumina platforms. At this point a library has been created, and the DNA molecules are ready for sequencing on the NGS platform. PCR-free library preparation minimizes DNA sequence errors and bias potentially introduced by an amplification step. However, library PCR can be performed depending on the DNA input and desired library concentration. Kits such as TruSeq DNA Library Prep Kit (Illumina) and KAPA HyperPrep Library Preparation Kit (Roche Sequencing) offer PCR-free library preparation approaches. Certain library preparation methods require the PCR step to attach the full adapter sequences. In these approaches, the adapters ligated to the DNA fragments contain short universal sequences. The library PCR is then performed with oligos comprised of the sequence complementary to the universal adapter followed by the platform-specific sequences (e.g., the P5 and P7 sequences for Illumina library preparation) and sample index. The NEBNext Ultra II DNA Library Prep kit (New England Biolabs) is an example of a kit that utilises this approach for library preparation. These methods allow for the conversion of dsDNA into NGS libraries regardless of the sequencing platform.

Several dsDNA library preparation kits are available for both Illumina and Ion platforms. Each kit varies slightly in the target input, conversion efficiency, adapter incorporation method, and the number of steps including bead-based purifications. The bead ratio is important during these purification steps to ensure that the DNA fragments of interest are not lost due to their size, while also effectively removing unincorporated adapters and primers. Automated platforms designed specifically for library preparation, such as the Ion Chef (Thermo Fisher Scientific) for Ion sequencing, are available to streamline laboratory processing and increase efficiency.

Fragmentation

Efficient library preparation and successful sequencing requires DNA of an appropriate size for the sequencing method. It is possible for samples with degraded DNA such as low-quality, aged skeletal elements or formalin-fixed paraffin embedded (FFPE) specimens to go directly into library preparation. However, most samples require fragmentation, either mechanical or enzymatic, to shear intact DNA. Mechanical shearing is commonly performed using sonication with instruments such as the Covaris platforms. The DNA sample is subjected to ultrasonic acoustic energy, which breaks the DNA into fragments of a desired size. Parameters such as frequency and time can be adjusted to achieve the desired fragment size down to as low as 100 bp. Once fragmented, the DNA samples can be prepared for sequencing using a standard library preparation method.

Library preparation kits such as Nextera XT DNA Library Prep Kit (Illumina) and KAPA HyperPlus Library Preparation Kit combine enzymatic
fragmentation and library preparation into a single workflow. The fragmentation step uses an enzyme to cut large DNA fragments to the optimal size for the sequencing platform (e.g., typically 300-500 bp for Illumina sequencing). The Nextera XT kit includes a “tagmentation” step in which a transposase cuts and adds (or “tags”) a universal sequence to the end of the DNA fragment. Indexes and platform-specific sequences are then added to the libraries with a PCR step using the transposase sequence. Although a simple process, several studies have demonstrated drastically lower coverage in specific regions of the mitogenome when library preparation is performed with Nextera XT [119,137,162]. It is believed that these reproducible drops in coverage are the result of transposase bias [163-166], amplification of GC-rich regions during library preparation [167-169], or potentially both. In a study by Ring et al. [170], the KAPA HyperPlus kit was shown to generate balanced coverage across the mitogenome and therefore greater multiplexing capabilities. The HyperPlus library preparation procedure includes an enzymatic fragmentation step like Nextera XT, but fragmentation is followed by blunt-end repair and A-tailing that allows ligation of platform-specific adapters and indexes to the DNA fragments. As a result, the HyperPlus procedure does not require a library amplification step, thus eliminating one source of bias in the NGS data. There are numerous other kit options for both Ion and Illumina library preparation that include enzymatic fragmentation, such as the NEBNext Fast DNA Fragmentation and Library Prep Set for Ion Torrent (New England Biolabs) and the QIAseq FX Library Preparation Kit available from QIAGEN for Illumina sequencing.

Indexes

The index sequences that are incorporated into the DNA fragments with the adapters during library preparation serve a variety of purposes. Index sequences range in length (e.g., six to 12 bps) and can be included on one (single-index) or both (dual-index) adapters. Depending on the indexing strategy, hundreds and even thousands of samples can be sequenced in a single NGS run. The reads associated with each index combination are bioinformatically identified and sorted into individual sequence data (FASTQ) files for analyses. Multiplexing can be maximized with the use of unique dual-index (UDI) adapters. Opposed to combinational dual-index adapters that share barcode sequences to make “unique” combinations of the two indexes, UDI adapters incorporate two unique index sequences. In addition to enabling efficient production-scale sequencing, cross-talk and index misassignment are minimized with UDIs, resulting in the confident demultiplexing required for high-sensitivity studies [171]. In addition to UDIs, unique molecular identifiers (or indexes; UMIs) can be incorporated during library preparation to improve the confidence in NGS data. A unique sequence (i.e., an UMI) is added to each individual DNA molecule prior to amplification. Sequencing reads generated
for each UMI are then bioinformatically condensed into one consensus sequence (Figure 3), thereby eliminating the presence of PCR and sequencing errors that can lead to the detection of false variants [172,173]. By utilising adapter sets that incorporate both UDIs and UMIs, the benefits of both index types can be leveraged to enable sensitive NGS data analysis [171].

Figure 3. Bioinformatic treatment of unique molecular indexes (UMIs). Prior to condensing the UMIs (above the dashed line), there 13 reads (grey) associated with six UMIs (A, B, C, D, E, and F) in the alignment. The red indicates a different from the reference sequence. After UMI reads are condensed to consensus reads (below the dashed line), only six sequences remain in the alignment, each representing an original DNA template molecule. As a result, only differences observed in all reads associated with a particular UMI will be represented in the consensus read.
Mitochondrial genome enrichment

Following library preparation, samples can be sequenced on an NGS platform to produce data from properly-adapted DNA fragments in the library. Whole genome sequencing (WGS), or shotgun sequencing, has been introduced into the forensic community for the purpose of IGG [174]. Large amounts of sequence data are required to generate the necessary breadth and depth of coverage for reliable genotyping. This coverage is often obtained with the use of higher-throughput sequencing instruments (e.g., the NovaSeq 6000), minimal multiplexing and/or multiple sequencing events. The production of WGS data is therefore cost prohibitive for routine casework. Additionally, forensic specimens that would most benefit from NGS pose a challenge to WGS due to numerous factors, such as limited sample, low DNA concentration, degraded DNA, the presence of inhibitors, and high proportions of exogenous DNA from the environment [175]. As a result, WGS is not a practical approach for mitogenome sequencing for most forensic applications. To more efficiently generate mtDNA data, enrichment of the mitogenome is performed using a variety of approaches.

Long-range amplification

One of the most widely-used mitogenome enrichment methods is long-range (LR) PCR as it minimizes the number of reactions necessary (and therefore the cost) as well as the co-amplification of NUMTs [176]. The entire mitogenome can be amplified in a single amplification using a pair of overlapping primers (e.g., F16426 and R16425) [177-179]. This approach can be utilised in medical research since the samples of interest contain high quantities of intact DNA, which is required for this mitogenome enrichment method. Two-amplicon strategies are more commonly applied due to the DNA quality of forensic samples [119,136,137,180]. There are several sets of LR primer pairs that have been published (e.g., [95,181,182]). However, an evaluation of three primer sets by Peck et al. found that the two overlapping amplicons described in [95] (Figure 4a) were the most compatible with a wide range of mtDNA haplotypes [183]. The LR primer pairs recommended by Illumina are also beneficial in that the two amplicons both include the CR [181], which increases the chances of covering the CR region and allowing comparison to
any previously generated STS data. LR PCR requires the use of enzymes and buffers that are specially designed to synthesize PCR products up to 20 kbps such as the TaKaRa LA Taq Hot Start polymerase and Advantage GC Genomic LA Polymerase (both TaKaRa). Additionally, it requires long extension times (8-20 minutes) to allow the entire amplicon to be replicated. Following LR PCR, fragmentation of the amplification products is required for library preparation and sequencing due to the large size of the amplicons. The LR enrichment approach is most often coupled with a library preparation method that includes enzymatic fragmentation to simplify processing. The main limitation of LR PCR is the DNA must be intact in order for efficient amplification to occur. If the input DNA is not of the requisite quality, the amplification will either fail or yield inefficient product, resulting in stochastic artefacts.

Figure 4. Depiction of a) the two long-range (LR) mitochondrial DNA (mtDNA) genome amplicons, b) the two Precision ID mtDNA Whole Genome Panel multiplexes (81 amplicons each), and c) the QIAseq Targeted DNA Mitochondrial Panel primers (222 total).

Small-amplicon approaches

To address samples with DNA fragments too small for LR PCR, approaches using multiplexed primer pools of overlapping small amplicons have been developed. One method includes mini- and midi-amplicons (144-237 bps and 282-444 bps, respectively) to amplify the mtDNA CR in two overlapping multiplexes, which was originally utilised for STS [184,185]. Two commercial NGS small-amplicon kits designed for Ion platforms are now available from Thermo Fisher Scientific, the Precision ID mtDNA Whole Genome and the Precision ID mtDNA Control Region Panels. The Precision ID mtDNA Whole Genome kit enriches for the entire mitogenome using two multiplexes of 81 amplicons each (Figure 4b) [134,149,186-188]. The Precision ID mtDNA Control Region kit targets only the mtDNA CR with two seven-amplicon multiplexes (14 amplicons total). Each Precision ID kit contains primer pairs that generate amplicons averaging 160 bps. The kits require as little as 100 pgs of
DNA, and Precision ID enrichment and library preparation can be automated on the Ion Chef. Although designed specifically for Ion platforms, alternative library preparation can be applied after targeted mtDNA PCR to allow for Illumina sequencing of the Precision ID enrichment products [149]. The two ForenSeq mtDNA Kits from Verogen offer a small-amplicon approach specifically designed for sequencing on the MiSeq FGx [189]. The ForenSeq mtDNA Whole Genome Kit targets the mitogenome using two multiplexes that contain a total of 245 amplicons that average 131-bp in length. The 1200-bp CR is targeted by 18 amplicons in two multiplexes in the ForenSeq mtDNA Control Region Kit, with amplicon sizes averaging 118-bp. Another small-amplicon commercial kit, the PowerSeq CRM Nested System from Promega Corporation, enriches for the mtDNA CR in a single multiplex of ten amplicon [190]. The amplicon size for this kit ranges from 144 bps to 237 bps and libraries are simultaneously prepared for sequencing on the MiSeq in a unique library preparation workflow. Although this presents a streamlined approach with one amplification reaction, coverage across the CR is imbalanced due to preferential amplification of the short overlap regions [190].  

Small-amplicon approaches are amenable to a broad range of forensic samples, from challenging casework specimens to high-quality references. To date, the Precision ID mtDNA Panels (Whole Genome and Control Region) and the PowerSeq CRM Nested System have been validated for use in several forensic laboratories in the U.S. (e.g., UNTCHI [140], California Department of Justice [191], and the FBI [192]). Mitogenome NGS analysis with these kits has also been utilised in historical cases, such as a study that examined the mtDNA of Viking Age bones using the Precision ID mtDNA Whole Genome Panel [193]. Although applicable to most forensic sample types, the small-amplicon enrichment approaches have inherent challenges. Due to the two-primer amplicon design, binding site mutations can be problematic. This is especially true in the HVS regions where numerous polymorphisms may be observed, particularly in specific haplogroups and populations. As a result, the panels include degenerate primers to ensure efficient amplification regardless of haplotype, but these can complicate downstream data analysis if the primer sequences are not completely removed [194]. The greatest challenge of mtDNA enrichment with small amplicons is the co-amplification of NUMTs [176]. This is not a substantial problem for low-quality samples with minute nDNA quantities; however, NUMT interference can be particularly problematic when analysing mitogenome data from high-quality samples [176,194,195]. The recommended input of 100 pgs of genomic DNA, or roughly 2000 mtDNA copies, for the Precision ID mtDNA panels is likely an effort to minimize NUMTs. Despite these challenges, commercial NGS mtDNA kits allow mitogenome analysis to be more accessible to forensic laboratories.
Single-primer extension

A different enrichment approach is utilised in another mitogenome commercial kit, the QIAsseq Targeted DNA Mitochondrial Panel from QIAGEN. The QIAsseq Targeted DNA panels utilise a library preparation that incorporates UMIs combined with a unique target enrichment approach. The QIAsseq procedure begins with a fragmentation step, regardless of sample type or DNA quality, to ensure that all fragments are an optimal size for library preparation and sequencing. This step also prepares the DNA fragments for adapter ligation. In addition to the sample-specific UDI, 12-bp UMIs are ligated to each individual DNA molecule prior to target enrichment. Furthermore, target enrichment (for V3 panels) is performed using single primer extension (SPE). SPE utilises one target-specific primer to enrich for the target region and then one universal primer that is complementary to the adapter that was ligated to the DNA fragment in the previous step. This approach does not have the same restrictions as traditional two-primer PCR design, which is an advantage for degraded DNA as it eliminates amplicon size constraints and minimizes primer binding conflicts. The QIAsseq Mitochondrial panel includes more than 200 primers tiled across the mitogenome to ensure uniform coverage (Figure 4c). Disadvantages of the QIAsseq Targeted DNA kit are the minimum input of 10 ng, the method is fairly labour-intensive with multiple bead purifications, and highly viscous reagents make automation difficult. Nevertheless, the combination of UMIs and SPE enrichment provide substantial advantages for NGS applications, such as uniform coverage and low-frequency variant detection. QIAsseq Targeted DNA panels have been used in cancer research and other medical studies involving cell-free DNA or FFPE samples [196-198]. Recently, the QIAsseq approach has been utilised in forensics to target mtDNA using the commercial panel [176,199] as well as custom panels targeting forensically-relevant nDNA targets [200,201].

Hybridization capture

Although current commercially available mitogenome NGS kits offer streamlined processing workflows, these methods may not be an option for extremely degraded samples. PCR-based enrichment approaches have been shown to be successful for forensic samples of varying DNA quantity/quality [202]. However, methods that employ small-amplicon enrichment will fail if a majority of the DNA fragments are less than 100 bps. The SPE enrichment of the QIAsseq Targeted DNA kit may allow for the assay to be more amenable to smaller fragment sizes, but this method has yet to be sufficiently tested on forensic samples. In decades-old human identification efforts, the remains have often been subjected to post-mortem insults that damage the DNA. Conditions such as time, temperature, soil acidity, fire, and decomposition can
break intact DNA into smaller fragments as well as induce changes to the DNA sequence (e.g., cytosine deamination) [203-205]. The quality of the DNA in severely compromised forensic samples is similar to FFPE tissues and even ancient skeletal remains [206]. Thus, the mitogenome of even the most challenging forensic samples can be successfully sequenced with methods designed specifically for the poorest-quality samples.

Hybridization capture is an enrichment approach that can overcome the limitations of PCR-based enrichment methods, particularly for severely degraded DNA [207]. Hybridization capture enrichment has been successfully applied in ancient DNA (aDNA) studies [208-210], including for mitogenome analysis [211-214]. Unlike the PCR-based methods, library preparation is typically performed prior to capture enrichment to increase DNA input as well as index the samples to prevent cross-contamination. This also allows for sample libraries to be pooled for capture to reduce costs; though this modification is primarily used for higher-quality samples and requires UDIs to prevent index jumping [215]. The “capture” of target DNA is enabled by ~70-bp unique oligos, known as probes or baits, that are designed to tile the region(s) of interest. During an incubation step, which is typically 16-24 hours, the baits hybridize to target DNA in the sample library. The biotinylated baits are then pulled out of solution using streptavidin beads, thus “capturing” the targeted DNA fragments that are bound to the baits. Subsequent wash steps remove any unwanted DNA, and the captured product is then amplified to increase the concentration of the target DNA prior to sequencing. A second round of capture can be performed to further increase the on-target yield, which may be necessary for samples with high proportions of exogenous DNA [207,212]. Another modification that can streamline the capture workflow is performing a subsequent capture using an alternative bait set of the off-target supernatant from the first round of capture [216]. This “differential” capture enables the enrichment of two different target panels (e.g., autosomal SNPs and then mtDNA) without the need for additional library volume.

Though more expensive and labour-intensive in comparison to the commercial mitogenome PCR-based methods available, hybridization capture may be necessary depending on the quality of the samples in certain cases or applications. As an example, a mitogenome hybridization capture assay was implemented at the AFMES-AFDIL for a specific set of decades-old human remains that were chemically treated post-mortem, rendering the DNA severely damaged and fragmented to approximately 40-50 bps [90]. This method uses a custom myBaits Target Enrichment Kit (Arbor Biosciences), which incorporates a custom RNA bait design based on a variety of haplotypes common in U.S. populations. Through validation studies, the method was shown to be reliable and robust for casework samples submitted to the AFMES-AFDIL for mtDNA analysis [90]. In addition to the AFMES-AFDIL custom bait kit, Arbor Biosciences now offers a pre-designed mitogenome target capture kit (myBaits Expert Mito – Human Global Panel) that was
designed based upon 197 globally diverse human mitogenome sequences. Alternative capture approaches such as a custom mitogenome HaloPlex Target Enrichment System (Agilent Technologies) have also been developed. The HaloPlex method combines target enrichment and library preparation into a single kit using restriction enzymes, hybridization, purification and ligation of the fragments to produce circular amplicons containing the target sequence and Illumina-specific adapter sequences. This method has been applied in genetic investigations of ancient historical remains such as Copernicus, St. Bridget of Sweden, and several Viking mass graves. STS data was utilised for the mtDNA analysis of these samples, which was limited to the HVS regions [193,217,218]. With the use of the HaloPlex enrichment, preliminary data show successful sequencing of the entire mitogenome for these ancient remains. These hybridization capture enrichment approaches combined with NGS offer the ability to analyse the mitogenome of severely degraded samples previously resistant to traditional STS methods.
Data analysis

The adoption of NGS has forced a paradigm shift in the forensic field, particularly in terms of data analysis. Due to the overwhelming abundance of sequence data generated from NGS (thousands to millions of reads per sample), the ability to visually inspect each individual position in the mitogenome is no longer possible and data analysis must be automated. NGS software used by forensic laboratories must therefore be thoroughly validated to ensure accurate variant calling based on desired analysis and interpretation criteria. Automated strategies may enable the application of thresholds and guidelines without analyst review as well as potential enforcement of standard nomenclature to output a reliable, forensic-quality mtDNA haplotype.

Software

There are several tools utilised by the forensic community for analysis of NGS mtDNA data [219,220], which can be applied in an automated or semi-automated fashion. The AFDIL-QIAGEN mtDNA Expert, or AQME, tool is a plug-in developed for the CLC Genomics Workbench (QIAGEN) [199,221]. AQME is used in conjunction with existing CLC tools to generate an editable mtDNA haplotype that employs forensic conventions with integrated haplogroup prediction and includes the interpretation range required for mtDNA data reporting. Specific for Ion sequencing, the Applied Biosystems Converge Software Data Analysis module (Thermo Fisher Scientific) was developed specifically to analyse Precision ID mtDNA data [222,223]. In addition to generating the sample haplotype, the Converge software also flags potential haplotype artefacts based on the expected presence of a variant in the sample’s haplogroup, variant phasing (i.e., NUMTs), and strand bias. Another software, GeneMarker HTS (SoftGenetics), provides streamlined workflows for the analysis of mtDNA NGS data from both Illumina and Ion platforms [192,224,225]. Most recently, Verogen’s Universal Analysis Software (UAS) has been updated (v2.0) to allow for the analysis of data generated with ForenSeq mtDNA kits [189,226] as well as custom mtDNA assays. Finally, the Parabon Fχ Forensic Analysis Platform (Parabon NanoLabs, Inc.) is an agnostic software suitable for NGS mtDNA analysis using all sequence data types [227,228]. Future versions of the Fχ software will include an option for
pathogenic variant filtering [124] that may facilitate the admissibility of complete mitogenome haplotypes in forensic casework.

Workflow parameters

Though mtDNA sequencing has been performed for decades, data analysis is complicated by homopolymer or other complex regions [148,149], artefacts such as NUMTs [72,176,195], and forensic conventions [94,99] that must be considered. Bioinformatic pipelines apply algorithms with adjustable parameters for trimming, alignment, and variant detection of mtDNA NGS data. Ultimately, the analysis workflow can impact the resulting mtDNA haplotype just as much as the enrichment method and sequencing technology. Additionally, the workflow parameters are dependent on the sample preparation method.

The input for most bioinformatic pipelines is the FASTQ file produced by demultiplexing the base call data generated on the sequencing platform. The reads are typically trimmed to remove low quality bases and excess adapter sequences at the end of the reads. Primers from PCR enrichment may also be removed during the trimming step. Additionally, UMI reads are condensed into a single consensus sequence at this point in the workflow, effectively eliminating errors introduced during PCR and sequencing (Figure 3). The prepared reads are then mapped to the rCRS, or some variation of the human mtDNA reference genome (e.g., rCRS+80), utilising various parameters that determine the sequence similarity required for a particular read to be included in the alignment. The stringency of the mapping may have a significant effect on the haplotype, especially when off-target reads are present (e.g., NUMTs, bacterial DNA). Also, alternative mapping strategies can be implemented to ensure only endogenous mtDNA reads are included in the alignment, such as consensus or competitive mapping [176]. For WGS and capture data, PCR duplicates are marked or removed from the mapping to allow for variant detection based only on unique reads (i.e., original DNA molecules), which minimizes PCR bias [207]. A realignment step may also be performed to help with the consistent placement of indels, particularly in homopolymer regions. Once the mapping is complete, NGS data analysis can continue with the detection of variants and generation of the mtDNA haplotype.

Variant detection is performed by comparing the mapped reads to the rCRS, employing a variety of thresholds and filters. In most workflows, a minimum read depth (or coverage) threshold is established and only positions that meet or exceed that read depth are examined. This threshold is dependent on the sample preparation as well as sequencing approach. Minimum coverage thresholds are typically lower (<30X) for workflows that require the removal of PCR duplicates (e.g., WGS, capture and UMI-adapted libraries) than PCR-based enrichment methods since variant detection is based on unique reads.
Other parameters utilised for variant detection are a minimum read count for the variant and variant frequency (ratio of the variant to the reference nucleotide). These values should be determined based on an assessment of background noise observed in the sequence data for the specific sample preparation method, which may also be impacted by the read depth. For example, it has been hypothesized that heteroplasmny detection may be as low as 1%, but the deep sequencing (i.e., high coverage) required is not feasible in routine forensic processing due to cost and/or throughput [229]. Based on an NGS study employing forensic samples, a 2% frequency was shown to be a valid reporting threshold to exceed background noise [144]. Variants that meet established thresholds can then be filtered based on quality and forward/reverse balance, which are indicators of sequencing error. Further filtering of specific variants can also be employed to remove pathogenic variants [124] and known artefacts, such as variants associated with NUMTs [176]. To simplify analysis, a higher variant frequency threshold can be applied to eliminate potential low-level artefacts. For example, although a 5% threshold has been shown to be a valid threshold for LR Illumina data [13], a 10% threshold was selected for a casework NGS analysis workflow at the AFMES-AFDIL to minimize stochastic errors observed in lower-quality reference samples [183]. Similarly, a variant frequency threshold of 30% may be employed for the analysis of small-amplicon data to reduce the detection of NUMTs and other artefacts. Another aspect of mtDNA analysis is the detection and reporting of LHP, which is observed in roughly half of the human population [12,13]. LHP is commonly observed in three C-stretches within the two HVS regions, but occurs in at least seven other regions across the mitogenome [9,13,118]. In addition to difficulties in sequencing homopolymeric and repetitive regions, the alignment of reads spanning these regions can also be problematic due to variable lengths and the placement of indels within the stretch. In order to overcome this challenge, fragment analysis can be employed to identify the frequency of molecules of different lengths [221,230,231]. Based on this fragment analysis, the major (or dominant) molecule (MM) can be determined and reported in the mtDNA haplotype according to forensic guidelines [99]. Prior to the analysis of mitogenome NGS data within a forensic laboratory, appropriate analysis thresholds and interpretation criteria must be established (and subsequently validated) to ensure the output of reliable, forensic-quality mtDNA haplotypes.

Quality control

The quality of the mtDNA haplotypes produced in forensics is of the utmost importance since these data may be used to identify a missing person or lead to the conviction of an alleged criminal. It is therefore imperative that robust QC checks are performed to ensure that both casework and reference data are
reliable. First, the processing and review of the data may be replicated to ensure the reproducibility of the haplotype. Any samples with suspicious variants in the haplotype that are indicative of a mixture, stochastic amplification error, NUMTs, untrimmed adapter or primer, or sequencing errors should be flagged during review. Cytosine deamination can also complicate NGS data analysis of low-quality samples since cytosines have been converted to uracils in the original template molecule, appearing as low-level cytosine-to-thymine and guanine-to-adenine mutations in the sequence data [232]. One critical QC indicator is the number of PHPs in a haplotype, as well as the ratio of CR and codR PHPs [14]. Based on studies of high-quality sequence data, a mitogenome haplotype is not expected to have more than three PHPs [13,118]; though this may vary based on the detection threshold [14]. Reprocessing of samples, potentially with an alternate method, can resolve suspicious variants or confirm low-level PHPs to ensure that only authentic variants are included in the mtDNA haplotype. Independent reviews by different analysts can also eliminate potential human error, allowing for replicate inspections of the mapping as needed. Often, studies in other disciplines rely on the analysis workflow to identify and remove any spurious variants in the data [233,234]; however, manual review by an experienced analyst can provide an invaluable assessment of the data. In particular, inspection of the mapping can assist in the identification of NUMT interference since variants associated with NUMTs will be in-phase (i.e., on the same read) as shown in Figure 2. Second, haplogroup prediction enables a QC appraisal of the mtDNA haplotype and is recommended by forensic guidelines [99,106]. Analysis workflows that incorporate mtDNA haplogroup prediction are advantageous because the nomenclature can be adjusted based on the phylogeny and additional haplotype QC can be performed during data review. In particular, artificial recombination and other processing errors can be identified. A haplotype resulting from artificial recombination will typically contain multiple missing and/or private mutations compared to the predicted haplogroup, primarily in one amplicon or region targeted by a multiplex. Artificial recombination can be further supported by haplogrouping the haplotype of each amplicon or multiplex target region, and then evaluating the predictions (including for the mitogenome) to ensure that the haplogroups are phylogenetically plausible. Next, electronic haplotype entry and data transfer should be used. Manual entry of haplotypes is prone to transcription error, particularly for mitogenomes, which can include more than 100 polymorphisms. For example, the data issues previously found in the FBI mtDNA database were partially caused by clerical errors [235].

Reference datasets may require an additional QC check to validate their quality prior to inclusion in a database for forensic use. For any shared haplotypes discovered in the population data, pairwise comparisons should undergo nDNA testing to assess for close relatives (i.e., nuclear family members, and potential second-degree relatives). MtDNA databases are expected to be a
random sampling of the population and the inclusion of haplotypes from the same maternal lineage may not meet this QC requirement [236]. Generally, only one haplotype per family per population should be included in a database. However, distant relatives can now be identified with the use of large amounts of STR and/or SNP markers [80,228,237]. Distant maternal relatives may be included in the same database as a “randomly related” sampling of a population. Furthermore, the inclusion of more than one haplogroup from a maternal lineage may be appropriate for isolated or endogamous groups [236]. As part of the QC check performed during submission to EMPOP, haplotypes are compared against a curated mitogenome dataset to identify any “abnormal” variants (e.g., variants never observed before, known phantom mutations or other haplotype irregularities). Based on this dataset, and phylogenetic knowledge, certain variants are expected to occur in the presence of other variants in a haplotype. Therefore, if a particular variant is observed but associated variants are not reported (or vice versa), the mtDNA data should be reviewed to ensure the haplotype is accurate. Quasi-median networks are also useful for identifying abnormalities in mtDNA haplotypes [100]. Furthermore, haplotypes can be compared to a catalogue of variants associated with known NUMTS in order to assess possible NUMT interference in the mtDNA data [176]. Haplotypes that do not pass the QC checks and established interpretation criteria should be flagged for further review. Reanalysis of the NGS data may be required and/or replicate processing using the initial or an alternative enrichment method. Those haplotypes that cannot meet QC criteria after additional measures should be excluded from reference databases or deemed inconclusive for casework samples.
Present investigations

The objective of this thesis was to assist the forensic community in its transition from mtDNA analysis using STS to NGS. NGS allows for high-throughput, cost-effective generation of mitogenomes; however, there are limitations to the implementation of mitogenome analysis in forensic laboratories. In particular, the availability of reliable mitogenome reference data must be substantially augmented in order to make use of mitogenome data that are expected to differentiate maternal lineages. Further, laboratories previously unfamiliar with traditional mtDNA sequencing (i.e., STS) may now be inclined to implement mitogenome analysis with the commercial availability of NGS panels. While this allows broader access to mtDNA analysis for the forensic community, each of these methods and associated analysis pipelines pose a number of new challenges. To ensure the generation and usage of forensic-quality mitogenome data, QC assessment will be essential. Mitogenome NGS data can be complicated by NUMTs, misalignment of homopolymer regions, sequencing error, and other artefacts. This can then lead to erroneous conclusions based on the resulting data. The studies presented in these four papers provide a better understanding of mitogenome NGS data and address the limitations to the implementation of mitogenome analysis. In addition to offering insight into the efficiency of different approaches to generating reliable mitogenome data, over 900 high-quality Swedish mitogenome haplotypes were produced and thus expanding the potential use of mitogenome evidence in forensic casework.
Paper I: The value of whole genome sequencing for mitochondrial DNA population studies: strategies and criteria for extracting high-quality mitogenome haplotypes

Aim
Publicly available WGS data provide a valuable source of mitogenome haplotypes. However, NUMTs from chromosomal DNA may prevent the production of reliable mitogenome haplotypes from untargeted WGS reads. Reads from NUMTs co-align with mtDNA reads, mimicking authentic heteroplasmy and complicating data analysis. The inclusion of NUMTs in mtDNA haplotypes may lead to erroneous conclusions, such as claims of biparental inheritance of mtDNA [64,69], and therefore must be avoided. This study assessed two different variant detection thresholds (5% and 10%) for the development of forensic-quality mitogenome haplotypes from WGS data.

Results and discussion
This study utilised the SweGen WGS dataset, which is comprised of more than 900 genomes from unrelated Swedish individuals from the TwinGene project. The nDNA was previously analysed in Ameur et al. [238], but the mtDNA data were not examined. When a 5% variant frequency threshold was applied to the mtDNA WGS data, there were 413 variants associated with NUMTs detected in the mtDNA alignments for 91 of 917 (~8%) SweGen samples. A majority of these NUMT variants were observed in two hotspot regions (nps 12612-13105 and 16390-16527), which were consistent with previously documented NUMTs [49,239]. The detection of NUMT interference was dramatically reduced with the use of the 10% variant frequency threshold, resulting in only 13 NUMT variants detected in eight samples. Furthermore, the 10% threshold haplotypes were similar in terms of heteroplasmic variation to other high-quality mitogenome datasets [13,118], whereas the frequency of PHPs was elevated in the 5% data. As a result, a 10% frequency threshold was required to produce high-quality mitogenome haplotypes from the SweGen WGS data for forensic use. Additionally, NUMT interference correlated to average read depth, as well as the proportion of mtDNA reads to nuclear reads in the WGS data (i.e., mtDNA copy number). Therefore, it may be possible to utilise these metrics as a QC metric to assist in the classification of observed PHPs as either authentic heteroplasmy or an artefact of NUMT reads.
Paper II: The value of whole genome sequencing for mitochondrial DNA population studies: strategies and criteria for extracting high-quality mitogenome haplotypes

Aim
The present study aimed to characterize the impact of sequencing chemistry on the pattern of LHP, and to determine the validity of utilising length variation for mtDNA comparisons in forensics. NGS data generated with two well-established mitogenome methods was compared to STS data in the HVS C-stretches. Sequence data were generated for 16 high-quality samples using the following approaches: 1) STS of ~1200-bp CR amplicons [240], 2) Illumina sequencing following PCR enrichment of the entire mitogenome using LR PCR [183], and 3) Ion sequencing on the Ion S5 System with the Precision ID mtDNA Whole Genome Panel [149]. LHP analyses enabled the characterization of length molecules in the data for an in-depth investigation of the impact of sample preparation and sequencing methods.

Results and discussion
The mitogenome haplotypes generated in this study were concordant across all conditions with the exception of length and low-level variants (<30% variant frequency). In the STS data, LHP in HVS1 was observed in samples with nine or more consecutive cytosines and eight cytosines in the HVS2 region. These results were consistent with previous STS studies [12,241]. The Illumina data produced a similar pattern of LHP to that of the STS data, whereas the Ion data was noticeably different. Of particular note, in these regions of low LHP complexity the average number of length molecules in the Ion data was nearly double that of the other two methods. Since Ion sequence data are prone to homopolymer indels [143,147], length variation was also observed in portions of HVS1 and HVS2 other than the longest C-stretches (e.g., the polyadenine stretch prior to each respective C-stretch). In contrast, length variation in the STS and Illumina data occurred almost exclusively (99%) in the expected C-stretches (i.e., C-stretches of more than seven consecutive cytosines). Overall, the STS MM differed from the Illumina MM in six (20%) of the 30 regions evaluated and in 11 regions (37%) in Ion data. In general, the STS MMs were longer than the Illumina MMs, and the Ion MMs were the shortest. In addition to sample preparation and sequencing platform, supplemental analysis with alternative analysis approaches demonstrated that the LHP pattern may also be altered by the bioinformatic tool, input data (e.g., FASTQ or mapped reads), and interpretation of the output. Based on these
results, the inclusion of length variants in mtDNA comparison should be performed with caution. With broader application of NGS in forensic laboratories, varying enrichment and sequencing methods will undoubtedly be used. Thus, it is expected that minor difference in LHP patterns will be observed and it will be important to ignore C-stretch indels for mtDNA haplotype queries and comparisons.

Paper III: Tools and considerations for mitochondrial haplogroup assignment

Aim

The aim of this study was to gain greater understanding of the haplogrouping tools available and the impact of the region used for the prediction, particularly the ability to obtain an accurate and informative mtDNA haplogroup. Haplogroups were determined for 92 diverse mtDNA haplotypes by EMPOP’s SAM2 [102], the Mitochondrial Haplogrouper tool in the AQME plug-in [221], and HaploGrep2 [242]. Haplogroups were assigned based on four regions: the full mitogenome, CR, HVS1 and HVS2 (HVS1-2), and HVS-1 only. The results were compared to the manual assignment using PhyloTree Build 17 [36,243]. The results were utilised to better assess the reliability of the prediction based on the region used for haplogroup assignment and evaluate the enhanced ability to QC data with the phylogenetic information.

Results and discussion

There were only two differences (out of 92) between all three tools when using the entire mitogenome for haplogrouping, and in these instances the haplogroups were less precise by only one or two nodes. Haplogroup assignments for the CR and HVS1-2 were similar; though resulting in less precise haplogroups for the HVS1-2 haplotypes compared to the CR due to the presence of haplogroup-diagnostic mutations outside the queried region. Each haplogrouping tool has a different value that provides an indication of the prediction quality (i.e., cost for SAM2, normalized score for AQME, and overall rank for HaploGrep2). However, these haplogroup quality scores provided very little guidance for many predictions. For example, a higher cost (>10) for the SAM2 prediction of a mitogenome haplotype still resulted in the correct haplogroup. The high haplogroup prediction cost was generated due to the uniqueness of the haplogroup (i.e., the haplotype is not present in PhyloTree). The comparison of the quality scores between authentic and artificial haplotypes showed that it is possible to identify recombinant LR mitogenome haplotypes using the quality scores generated by the three tools, but the differentiation is more
difficult with HVS1-2 haplotypes due to similar haplogroup scores for artificial and authentic haplotypes.

Overall, the tools performed similarly; however, EMPOP’s SAM2 produced more precise haplogroup predictions than AQME and HaploGrep2 across all haplogroups and regions. Artificial recombination may be detected through haplogroup prediction using these tools, but the detection is dependent on the haplogroups involved and/or target regions, and analyst review is often required. Each of the tools presented advantages and disadvantages, and the optimal tool for a laboratory is ultimately dependent on the application of the haplogroup prediction. For example, both SAM2 and HaploGrep2 are freely available, but the AQME tool requires a license for both the plug-in and the CLC Genomics Workbench. However, AQME is the only tool that can be applied during analysis and analyst review of the NGS data. The most important consideration when using any haplogrouping tool, particularly for smaller regions, is that the haplogroup identified is only a prediction. Analysts must therefore evaluate the results critically before using the information.

Paper IV: Mitochondrial DNA genome variation in the Swedish population

Aim

The development of mitogenome reference data for inclusion in publicly-available population databases is currently underway, and the addition of more high-quality mitogenomes will only enhance the statistical power of this forensically-useful locus. In order to facilitate mitogenome testing in Sweden, the mtDNA data from the SweGen WGS dataset [238] were analysed. To avoid the NUMT interference observed in the SweGen WGS data, which was described in Paper I, this study utilised the 917 high-quality mitogenomes produced with the application of a 10% variant frequency threshold for the population analyses.

Results and discussion

Due to the increase in the applied variant frequency threshold from 5% to 10%, there were 17 haplotypes excluded in Paper I that were added to the SweGen mtDNA dataset, increasing the dataset to 934 total haplotypes. Despite the elevated frequency threshold, 31 NUMT variants were still observed in 13 lower coverage haplotypes. However, NUMT interference was minimal and localized to two hotspot regions, substantially reducing the analysis burden that was required for the 5% frequency threshold described in Paper I. Almost 45% of the SweGen haplotypes belonged to haplogroup H and nearly all mitogenome haplotypes (99.1%) were assigned to European haplogroups,
which was expected based on previous studies of Swedish CR [244] and HVS [245] mtDNA data. There were characteristic northern Swedish (i.e., Saami) and Finnish haplogroups present in the dataset, consistent with the nDNA analyses of the SweGen data [238]. The analysis of the complete mitogenome resulted in high haplotype diversity (0.9996) with a random match probability of 0.15%. Overall, the mitogenomes generated from the SweGen WGS data provide a mitogenome reference database for Sweden as well as contribute to the global effort to increase the availability of forensic-quality mitogenome reference data.
Concluding remarks and future perspectives

The analysis of mtDNA plays a relatively small, but significant role in forensic casework. Limitations to the use of nDNA markers can be overcome with mtDNA due to its high copy number and maternal inheritance. Though sequencing of mtDNA has limited discrimination power, especially when targeting the CR, NGS offers feasible access to mitogenome data that can be used to resolve common haplotypes. There are several challenges to the implementation of mitogenome analysis in forensic laboratories, and the largest is associated with match statistics. There is an insufficient number of forensic-quality reference mitogenomes available to accurately estimate haplotype frequencies at this time, which are needed to accurately determine the evidential weight of mtDNA data in the case of a match. NGS allows for high-throughput, cost-effective generation of mitogenomes, and this technology will therefore permit the amount of reference data available to increase rapidly. The quality of reference data, as well as data generated for a forensic sample, is of great importance. With broader access to mtDNA with commercial NGS kits, laboratories that use NGS for nDNA analysis may look to implement mtDNA as well. Therefore, the forensic community must have access to the necessary knowledge and data in order to support the correct usage of mitogenome NGS analysis.

This thesis aimed to provide a comprehensive understanding of mitogenome NGS data, while also producing high-quality reference data, in order to encourage the forensic community to utilise mtDNA and implement NGS technologies. Paper I evaluated the feasibility of mitogenome haplotype generation using previously produced WGS data. This study highlighted the complexities of analysing mtDNA data with high proportions of NUMTs. This issue is not isolated to WGS data as NUMT interference impacts small-amplicon and hybridization capture enrichment methods as well. The use of an appropriate detection threshold that effectively excludes unauthentic PHPs caused by NUMTs was shown to be imperative for the generation of high-quality mitogenome data. Similarly, characterization of sequencing error (i.e., background noise) and platform-specific artefacts is necessary to establish the limits of detection for mtDNA NGS data analysis. Each platform employs, or at least recommends, different enrichment and library preparation kits, which may impact the consistency between data generated with different sequencing technologies. As more knowledge is gained about sequencing errors, NUMTs
and other artefacts observed in mtDNA NGS data, methods can be developed to effectively identify and remove false variants that simplify analysis. However, until novel software can be developed or existing software can be improved, there is a continued need for analyst review and stringent QC measures of mitogenome NGS data.

Two studies included in this thesis further examined QC aspects of mitogenome NGS data analysis. The variation in the detection of LHP that may be observed between enrichment and sequencing approaches was described in Paper II. Based on these results, the continued need to ignore length variants in database searches and forensic match comparisons was confirmed. The study also provided suggestions for tools and parameters for the analysis of LHP regions in mtDNA NGS data. The haplogrouping tools evaluation presented in Paper III demonstrated the usefulness of haplogroup prediction for the identification of artificial recombinant haplotypes [104-106], which is applicable to NGS data generated with enrichment methods that have more than one amplicon or PCR multiplex. As the result of recent modifications to both the SAM2 [246] and HaploGrep2 algorithms [247], a reassessment is needed to evaluate the currently available haplogroup prediction tools. Overall, these three studies examined factors that can complicate mtDNA NGS data analysis and interpretation, including NUMTs, the misalignment of homopolymer regions, and other artefacts. Their findings provide guidance that enable high-quality mtDNA haplotype generation, thus avoiding erroneous conclusions based on haplotypes based on data of questionable quality.

Another aim of this thesis was the generation of forensic-quality mitogenome reference data. In Paper IV, mitogenome haplotypes were generated from existing production-scale WGS data from the SweGen nDNA study [238]. This allowed the generation of over 900 Swedish mitogenomes with no additional processing costs, while WGS is cost-prohibitive for most forensic applications. Established approaches to NGS sample preparation (enrichment and library generation) and sequencing have been adapted to the processing of sample types routinely encountered at forensic laboratories. Although the standard procedures are sufficient, modifications to streamline these methods would reduce costs and processing time while increasing throughput. Automation of mitogenome NGS methods could also increase throughput as well as reduce hands-on time and processing errors [248]. The Ion Chef offers automated library preparation for Precision ID panels; however, the throughput capability is limited to eight samples. The ForenSeq DNA Signature Preparation kit (Verogen) has been automated and allows for the NGS sample preparation (nDNA enrichment and library preparation) of up to 96 samples [249], which could be applied (with minimal modifications) to the ForenSeq mtDNA kits. Taylor et al. used an automated library preparation method for the sequencing of several hundred “platinum-quality” mitogenomes on a Hamilton Microlab STARplus Liquid Handling System [13]. Even lower in throughput, manual hybridization capture workflows are laborious and time-consuming.
(two weeks from extraction to data analysis) with the ability to process only a few samples at a time. As a result, the hybridization capture method is reserved for only the poorest quality samples. However, the benefits of this specific method for forensic applications have been demonstrated for mitogenome analysis as well as the targeted enrichment of 5000 to 95,000 nDNA SNPs [90,228,250]. Minor modifications to streamline and automate the procedure (e.g., capture pooling, automation) would allow for broader application of hybridization capture in routine casework [251]. To further improve the success of the most challenging samples (e.g., severely degraded, heavily damaged, low endogenous DNA proportion), single-stranded library preparation methods used in the aDNA field have shown promising results for forensic samples [206]. Specially-designed adapters used in these methods allow for the conversion of native ssDNA species as well as denatured dsDNA, enabling more endogenous DNA molecules to be sequenced [252,253]. Improvements in enrichment and library preparation combined with laboratory automation will make NGS a practical approach for routine mitogenome analysis in forensic laboratories.

This thesis evaluated different aspects of QC for mitogenome NGS analysis in Paper I (NUMT detection), Paper II (LHP) and Paper III (haplogrouping). As a result of these analyses, over 900 high-quality mitogenomes were produced from WGS data in Paper IV. These studies further characterized mtDNA NGS data and addressed the primary limitation to the analysis of mitogenome data in forensics (i.e., availability of forensic-quality mitogenome reference data used for the calculation of evidential weight). With broader implementation of NGS in forensics, the use of mtDNA analysis will likely increase over the next few years due to the availability of commercial kits and specialized software. NGS also allows for the recovery of nDNA from challenging samples when combined with enhanced sample preparation. This includes shed hairs [254,255] and historical remains [228,250], which were previously limited to mtDNA analysis. Therefore, the use of SNPs in forensic casework has the potential to expand exponentially, especially given the continued success of IGG [81,174,256,257]. Although restricted to the maternal lineage, mtDNA has the ability to identify related individuals more distant than the kinship capabilities of high-density SNPs [258]. Additionally, the use of mtDNA has been accepted in the U.S. court system for nearly three decades [259], whereas there are ethical and legal questions regarding the forensic use of SNPs and IGG that must be addressed. Therefore, mtDNA will continue to maintain a vital role in forensics as a lineage marker, especially given the throughput enabled with NGS and the enhanced discrimination power of the mitogenome. Ultimately, the expanded knowledge and data developed as part of this thesis will enable more forensic laboratories to adopt mitogenome NGS analysis to assist in human identification and criminal investigations.
The journey to write my thesis and complete my PhD has been incredibly challenging, but also one of the most rewarding experiences of my life. It would not have been possible without the amazing group of colleagues, friends and especially my family that have supported me throughout the process. I am forever grateful to you all!

First and foremost, I am sincerely thankful to my supervisor, Marie Allen for this amazing opportunity. Your invaluable advice, continuous support, and patience allowed me to finish my thesis and ultimately get my PhD. You and Anders made my visits to Uppsala so wonderful, and I truly appreciate everything you have done for me. I am very much looking forward to our future collaborations and your continued mentorship.

Thank you to Lars Forsberg for taking on the task of my co-supervisor. Additionally, I would like to thank my co-supervisor Charla Marshall, who helped me so much through this PhD journey. I appreciate all the knowledge and experience you have shared with me, especially the ability to see the most important aspect of the data. We truly make an amazing team! I am a better researcher and writer because of you. Thank you so much for your mentorship, but also your friendship!

I have been incredibly lucky to have worked with such amazing scientists throughout my forensic career. I started my professional career at AFDIL, and I have grown up there. From the very beginning, the research/ETS group has been my family. I am thankful for Jess and Toni, who were there from the start and became life-long friends. You both have been so supportive throughout this PhD adventure and I cannot wait to celebrate with you both!

To the Dover ETS crew – Joe, Jen, Cassie, Erin, Michelle, Tyler, Courtney, and Elise – I could not have done this without your help and understanding. You all have made going to work, no matter how busy or stressed I was, so enjoyable. I love seeing how you all have grown as scientists, and I am so thankful to have worked with you all!

I also need to thank the amazing mentors that I have had throughout my forensic career. Jodi Irwin, Rebecca Just, Mike Coble, and Tom Parsons – I would not be where I am today professionally without your support and mentorship early in my career. You all exposed me to so many interesting projects and connected me with incredible collaborators from around the world. Thank you to Walther Parson for your encouragement throughout my PhD. I am so
grateful that our shared love of mtDNA has brought us together, and I have
learned so much from you. Also, thank you Daniele Podini for your continued
support and friendship. I am grateful for the opportunity to have had you all
as mentors, so thank you!

There are so many others at AFDIL that have been supportive of me and
my PhD. Thank you so much to Tim McMahon, Louis Finelli and Alice Bri-
one for allowing me to take on this endeavour and achieve one of my goals.
To Tim Cambridge and Jim Ross, thank you for always dealing with all my
CLC and NGS server questions. Thank you to Suzie for your encouragement
and to Julie for your support and friendship. Kerriann, thank you for bringing
me all of your complicated and weird NGS cases. But also thank you for being
such a good friend and always being there when everything was fine!

These projects have included the expertise of so many other scientists.
Thank you to Adam Ameur for his help with the SweGen data and Hanne
for her assistance with various projects. Niki, thank you for all your help with
mtDNA haplogrouping, and hopefully we will be able to work together again
soon. To Martin and Andreas – you two are so amazing, both as scientists,
but also as people. I am so appreciative of your willingness to share your
knowledge and expertise with me. I have learned so much from you both, and
I look forward to all of our future collaborations!

Thank you to Magda and Ida for making my visits to Uppsala so much
more fun. I truly appreciate all your help navigating Uppsala and the univer-
sity. You both made this such a memorable experience and I am grateful for
your friendships.

I would not have been able to complete my PhD without the amazing sup-
port system I have around me. First, Meghan – where do I even start? After
30+ years, you are still my best friend. You have been so supportive, from our
daily morning phone calls to nightly texts about the day’s craziness. I could
not have done this without you always pushing me and encouraging me. Love
you! And to Daniela – you have encouraged me so much throughout this pro-
cess. You are such an amazing, strong woman, and I’m so glad our sons be-
came friends so that we could be friends! Thank you, Tom and Mary, for your
constant support and encouragement. Finally, to all friends and teammates,
thank you for always being there whenever I needed to laugh or cry or have a
drink. I truly appreciate it!

To my extended family – Aunt Kay, Uncle Jim, Margie, Jackie, Laura,
Tony, Laurie, and Aunt Ruth. You all have been so loving and supportive
throughout my life, and even more so during my PhD studies. The Bressler
family is a group of strong, intelligent women, and you all have inspired me
to excel. Thank you! To my grandparents, Joe and Margaret, I miss you both
every day and I know how proud you would be of this PhD. Also, thank you
to Lorenzo and Sina, I could not ask for better in-laws. And to Lisa, Jr and
Dom – thank you for your constant support. I am looking forward to spending
more time with you all now that my PhD is complete. Thank you to my entire family for their love and support over the past five years!

I am forever grateful for my parents, **Ann** and **Roland**. You truly are the most incredible, loving, supportive parents – even though you do drive me crazy from time to time. From a young age, you made sure I knew the importance of education and I have always strived to make you proud. You have been so supportive and helpful throughout this entire process, from cooking dinner to serving wine to watching the boys. I would not be where I am today without you. I am so lucky to be your daughter!

To my little men – **Max** and **Luca**! You two crazy boys bring me so much joy every day. You constantly cheered me on to finish my thesis, but you were also a good distraction when I needed a break. And when I was stressed, snuggling with the two of you helped me calm down and eventually get back to writing. I hope that I have shown you what you can do with hard work and dedication, and I hope you are as proud of me as I am of the two of you!

And above all, I have to thank my husband, **Lorenzo**. You have been my biggest supporter. You were always there to encourage me when I needed it, dealt with me when I was sleep deprived, made me laugh when I was stressed, brought me coffee and food when I was busy working, cleaned our house, and made sure our kids were still alive. You are truly my best friend, and I could not have done this without your love and support. I love you so much!
References


Newman, A.B.; Matteini, A.M.; Bandeen-Roche, K.; Boerwinkle, E.; Walston,
J.D.; Arking, D.E. Association of mitochondrial DNA levels with frailty and

Proccaccio, V.; Wallace, D.C. mtDNA Variation and Analysis Using Mitomap

31. Ruiz-Pesini, E.; Lott, M.T.; Proccaccio, V.; Poole, J.C.; Brandon, M.C.; Mishmar,
D.; Yi, C.; Kreuziger, J.; Baldi, P.; Wallace, D.C. An enhanced MITOMAP
with a global mtDNA mutational phylogeny. *Nucleic Acids Res* 2007, 35,
D823-8.

32. Schon, E.A.; DiMauro, S.; Hirano, M. Human mitochondrial DNA: roles of inher-

33. Freyer, C.; Cree, L.M.; Mourier, A.; Stewart, J.B.; Koolmeister, C.; Milenkovic,
D.; Wai, T.; Floros, V.L.; Hagstrom, E.; Chatzidakis, E.E.; Wiesner, R.J.; Sam-
uels, D.C.; Larsson, N.G.; Chinnery, P.F. Variation in germline mtDNA heter-
oplasmia is determined prenatally but modified during subsequent transmis-

34. Filograna, R.; Mennuni, M.; Alsina, D.; Larsson, N.G. Mitochondrial DNA copy
number in human disease: the more the better? *FEBS Lett* 2021, 595, 976-1002.

35. Soltys, D.T.; Pereira, C.P.M.; Rowies, F.T.; Farfel, J.M.; Grinberg, L.T.; Suemoto,
C.K.; Leite, R.E.P.; Rodriguez, R.D.; Ericson, N.G.; Bielas, J.H.; Souza-Pinto,
N.C. Lower mitochondrial DNA content but not increased mutagenesis asso-
ciates with decreased base excision repair activity in brains of AD subjects.
*Neurobiol Aging* 2019, 73, 161-170.

36. van Oven, M.; Kayser, M. Updated comprehensive phylogenetic tree of global
human mitochondrial DNA variation. *Hum Mutat* 2009, 30, E386-94.

Kivisild, T.; Torroni, A.; Vilems, R. A "Copernican" reassessment of the hu-
mam mitochondrial DNA tree from its root. *Am J Hum Genet* 2012, 90, 675-
684.

identification tests targeting the 16S ribosomal RNA coding region in mito-

39. Nelson, M.C.; Morrison, H.G.; Benjamino, J.; Grim, S.L.; Graf, J. Analysis, opti-
mization and verification of Illumina-generated 16S rRNA gene amplicon sur-

40. Salipante, S.J.; Sengupta, D.J.; Rosenthal, C.; Costa, G.; Spangler, J.; Sims, E.H.;
Jacobs, M.A.; Miller, S.I.; Hoogestraat, D.R.; Cookson, B.T.; McCoy, C.;
16S rRNA next-generation sequencing of polymicrobial clinical samples for

41. Dawnay, N.; Ogden, R.; McEwing, R.; Carvalho, G.R.; Thorpe, R.S. Validation of
the barcoding gene COI for use in forensic genetic species identification.

42. Ferri, G.; Alu, M.; Corradini, B.; Licata, M.; Beduschi, G. Species identification
through DNA "barcodes". *Genet Test Mol Biomarkers* 2009, 13, 421-426.

43. Benson, D.A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Ostell, J.; Pruitt,

44. Lopez, J.V.; Yuhki, N.; Masuda, R.; Modi, W.; O’Brien, S.J. Numt, a recent trans-
fer and tandem amplification of mitochondrial DNA to the nuclear genome of


107. Clopper, C.J.; Pearson, E.S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934, 26, 404-413.


181. Illumina Human mtDNA Genome: For the Illumina Sequencing Platform, Document #15037958 v01 ed.;2016;


A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine”.)