Taylor & Francis
Taylor & Francis Group

Expert Review of Molecular Diagnostics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/iero20

Enhanced mitochondrial genome analysis: bioinformatic and long-read sequencing advances and their diagnostic implications

William L. Macken, Micol Falabella, Chiara Pizzamiglio, Cathy E. Woodward, Elizabeth Scotchman, Lyn S. Chitty, James M. Polke, Enrico Bugiardini, Michael G. Hanna, Jana Vandrovcova, Natalie Chandler, Robyn Labrum & Robert D.S. Pitceathly

To cite this article: William L. Macken, Micol Falabella, Chiara Pizzamiglio, Cathy E. Woodward, Elizabeth Scotchman, Lyn S. Chitty, James M. Polke, Enrico Bugiardini, Michael G. Hanna, Jana Vandrovcova, Natalie Chandler, Robyn Labrum & Robert D.S. Pitceathly (2023) Enhanced mitochondrial genome analysis: bioinformatic and long-read sequencing advances and their diagnostic implications, Expert Review of Molecular Diagnostics, 23:9, 797-814, DOI: 10.1080/14737159.2023.2241365

To link to this article: https://doi.org/10.1080/14737159.2023.2241365

Published online: 29 Aug 2023.

EXPERT REVIEW OF MOLECULAR DIAGNOSTICS
2023, VOL. 23, NO. 9, 797–814
https://doi.org/10.1080/14737159.2023.2241365

Taylor & Francis
Taylor & Francis Group

REVIEW

Enhanced mitochondrial genome analysis: bioinformatic and long-read sequencing advances and their diagnostic implications

William L. Macken[a,b], Micol Falabella[a], Chiara Pizzamiglio[a,b], Cathy E. Woodward[b,c], Elizabeth Scotchman[c], Lyn S. Chitty[c], James M. Polke[b,c], Enrico Bugiardini[a,b], Michael G. Hanna[a,b], Jana Vandrovcova[a], Natalie Chandler[c], Robyn Labrum[b,c] and Robert D.S. Pitceathly[a,b]

[a]Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, UK; [b]NHS Highly Specialised Service for Rare Mitochondrial Disorders, Queen Square Centre for Neuromuscular Diseases, The National Hospital for Neurology and Neurosurgery, London, UK; [c]Rare and Inherited Disease Laboratory, North Thames Genomic Laboratory Hub, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK

ABSTRACT

Introduction: Primary mitochondrial diseases (PMDs) comprise a large and heterogeneous group of genetic diseases that result from pathogenic variants in either nuclear DNA (nDNA) or mitochondrial DNA (mtDNA). Widespread adoption of next-generation sequencing (NGS) has improved the efficiency and accuracy of mtDNA diagnoses; however, several challenges remain.

Areas covered: In this review, we briefly summarize the current state of the art in molecular diagnostics for mtDNA and consider the implications of improved whole genome sequencing (WGS), bioinformatic techniques, and the adoption of long-read sequencing, for PMD diagnostics.

Expert opinion: We anticipate that the application of PCR-free WGS from blood DNA will increase in diagnostic laboratories, while for adults with myopathic presentations, WGS from muscle DNA may become more widespread. Improved bioinformatic strategies will enhance WGS data interrogation, with more accurate delineation of mtDNA and NUMTs (nuclear mitochondrial DNA segments) in WGS data, superior coverage uniformity, indirect measurement of mtDNA copy number, and more accurate interpretation of heteroplasmic large-scale rearrangements (LSRs). Separately, the adoption of diagnostic long-read sequencing could offer greater resolution of complex LSRs and the opportunity to phase heteroplasmic variants.

PLAIN LANGUAGE SUMMARY

Mitochondria generate our bodies' energy, and they contain their own circular DNA molecules. Changes in this mitochondrial DNA can cause a wide range of genetic diseases. Improved computer processing of the sequence of this DNA and new techniques that can read the full DNA sequence in one experiment may enhance our ability to understand these genetic variants.

ARTICLE HISTORY
Received 24 March 2023
Accepted 24 July 2023

KEYWORDS
Bioinformatics; primary mitochondrial diseases; large-scale rearrangements; long-read sequencing; mitochondrial DNA

1. Introduction

1.1. Primary mitochondrial diseases

Primary mitochondrial diseases (PMDs) are genetic disorders caused by pathogenic variants in mitochondrial DNA (mtDNA) or nuclear DNA (nDNA) genes that encode mitochondrial proteins or RNAs. mtDNA variants can be present in all copies of mtDNA (homoplasmy) or co-exist with wild-type mtDNA (heteroplasmy). PMDs are characterized by perturbed oxidative phosphorylation or other aspects of mitochondrial functioning [1]. Mitochondria are commonly described as the 'powerhouses' or 'batteries' of cells given they are the central site of energy generation. However, as they house a myriad of other biochemical functions (e.g. calcium homeostasis and apoptosis), the biochemical manifestations of PMDs are correspondingly diverse [2–4]. In this review, we focus on molecular diagnostics pertaining to PMDs due to mtDNA variants; these include matrilineal mtDNA variants and secondary changes in mtDNA due to mtDNA maintenance defects (MDMDs).

1.1.1. Advances in the field of mitochondrial medicine

Molecular diagnoses enable personalized management and counseling in PMDs. Given the progress in reproductive options and clinical trials, the motivation to secure a diagnosis is growing ever stronger [5,6]. Prenatal options for PMDs now include preimplantation genetic diagnosis (PGD) and mitochondrial donation [7]. These *in vitro* fertilization-type approaches save couples the considerable distress of undergoing invasive diagnostic procedures and the termination of an affected pregnancy. Regarding treatments, there have been exciting developments in gene therapy (adeno-associated virus therapies) for acute Leber Hereditary Optic Neuropathy (LHON) and small molecules are entering clinical trials for mitochondrial myopathies [6,8,9]. Importantly, a molecular diagnosis is a prerequisite for recruitment to trials emphasizing the need for effective diagnostics to allow patients to access emerging therapies. Additionally, preclinical research is also investigating the

CONTACT Robert D.S. Pitceathly r.pitceathly@ucl.ac.uk Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, Queen Square, London WC1N 3BG, UK

Article highlights

- In most diagnostic laboratories mtDNA is enriched with a long-range PCR and then sequenced using NGS. This technique is very accurate and reliable for heteroplasmic single nucleotide variants. However, the long-range PCR step can result in preferential amplification of short amplicons meaning deletions in mtDNA can be exaggerated. Whole genome sequencing could overcome this problem as it does not include a PCR step. In addition, as WGS includes both mtDNA and nDNA, it could allow for a more efficient diagnostic process by analyzing both concomitantly.
- Innovations in WGS bioinformatics for mtDNA now enable improved alignment, coverage, and NUMT elimination, thereby increasing its suitability for diagnostic use.
- Structural variant tools that process and interpret multiple heteroplasmic LSRs offer realistic solutions for the application of WGS to muscle DNA.
- Accuracy of long-read technologies has improved so that they may potentially contribute toward mtDNA diagnostics, including superior LSR detection and mtDNA variant phasing.
- Innovations in Cas9 applications have enhanced targeted mtDNA long-read sequencing.
- Significant barriers to long-reads diagnostic use remain, including depth of coverage and innovative bioinformatic solutions to data analysis.

potential role of mtDNA editing and various heteroplasmy modulation techniques; however, clinical application of these methods remains remote for now [10–13].

1.1.2. The changing landscape of mtDNA diagnostics

While PMD diagnostics have traditionally been laborious, placing whole genome sequencing (WGS) at the center of diagnostics could revolutionize the sector. Concomitant sequencing of nDNA and mtDNA has been possible for several years; however, recent bioinformatic advances discussed in this article are now, for the first time, offering the prospect of all-in-one WGS-based testing for PMD diagnostics. Furthermore, long-read sequencing methods, previously thought to have insufficient sensitivity for use in diagnostics, are now achieving superior accuracy, even in homopolymeric regions [G]. The next decade is likely to see a transition away from traditional multistep testing toward next-generation sequencing (NGS) WGS, and eventually long-read WGS.

In this article, we summarize current targeted NGS approaches to mtDNA and consider how bioinformatic innovations may facilitate the wide-ranging implementation of WGS in routine practice. We review the opportunities that increasingly sophisticated bioinformatic callers for complex and multiple large-scale rearrangements (LSRs) offer for WGS, especially from muscle-extracted DNA. Finally, we consider improvements in long-read technologies and targeting approaches for mtDNA and how these might improve mtDNA diagnostics going forward.

2. Body

2.1. Mitochondrial DNA diagnostics

2.1.1. 2.1 Characteristics of mitochondrial DNA and their diagnostic implications

Mitochondrial genomes are relatively short (~16.6kb) double-stranded multicopy DNA molecules located in the mitochondrial matrix. The length and sequence of human mtDNA were resolved in 1981 by Anderson et al. [14]. Using Sanger sequencing, they established that the mtDNA sequence is 16,569 bases in length. The original sequence included an error (extra base) at position 3107 which has been retained to avoid confusion and the need for renaming of variants [15]. The mitochondrial genome displays 'extreme economy,' whereby most genes are not separated by a non-coding sequence. Rather, mtDNA contains only 37 genes, of these 13 are protein-coding genes, which are separated for the most part by one of the 22 transfer RNA (tRNA) genes (Figure 1A). The two remaining genes encode mitochondrial ribosomal RNAs (rRNAs). The two strands of mtDNA are known as the heavy strand (H-strand) and light strand (L-strand) owing to the higher proportion of heavy purine nucleotides (guanines) in the former. The 'start/end' sites of the mitochondrial genome are contained within the 'control region,' a non-coding section which includes the displacement loop (D-Loop) and transcriptional (HSP and LSP) and replication (O_H) promoters (Figure 1B). The name D-loop arises from the fact that the area often contains a short additional strand of mtDNA, hydrogen bonded to the L-strand which displaces the H-strand outwards to form a D shape [16,17]. This third strand, known as '7S DNA' is the result of early termination of H-strand replication which originates in the D-loop at the origin of heavy strand replication (O_H) [18]. Parts of the D-loop are 'hypervariable' and subject to a high rate of variation between ethnicities [19]. While the control region contains the start site of H-strand replication, the start site of L-strand replication (O_L) is located further downstream. The DNA between O_H and O_L (in the direction of replication) is known as the major arc, with the area between O_L and O_H known as the minor arc (Figure 1B).

mtDNA contains several repetitive sections and homopolymeric regions (runs of guanines), which can cause difficulties for certain sequencing technologies [20–22]. mtDNA has less effective DNA repair mechanisms and as such is subject to deletion (and other variant) formation, especially at these repetitive sections, which include perfect repeats (Class I deletions) and imperfect/interrupted repeats (Class II deletions) [23,24]. Homopolymeric regions, in particular, can give rise to non-canonical DNA structures (G-quadruplexes) [G] that contribute to deletion formation and abnormal mitochondrial functioning [25]. As a result, LSRs, especially deletions, are common and play an outsized role in mtDNA-related PMDs. Indeed, the most common cause of mitochondrial myopathy is a recurrent Class I deletion (~m.8470_13446del4977) which accounts for many cases of chronic progressive external ophthalmoplegia (CPEO) and Kearns Sayre Syndrome (KSS) [26,27].

The mtDNA molecule is circular in its native form. This can pose challenges to NGS, given that aligners are designed to deal with linear genomes and do not map sequence reads across a start/end site. Therefore, reads in this area may be removed, resulting in poor coverage of the control region. Uniform control region coverage is further compounded by its repetitive and highly variable nature, which may also contribute to suboptimal alignment. Also,

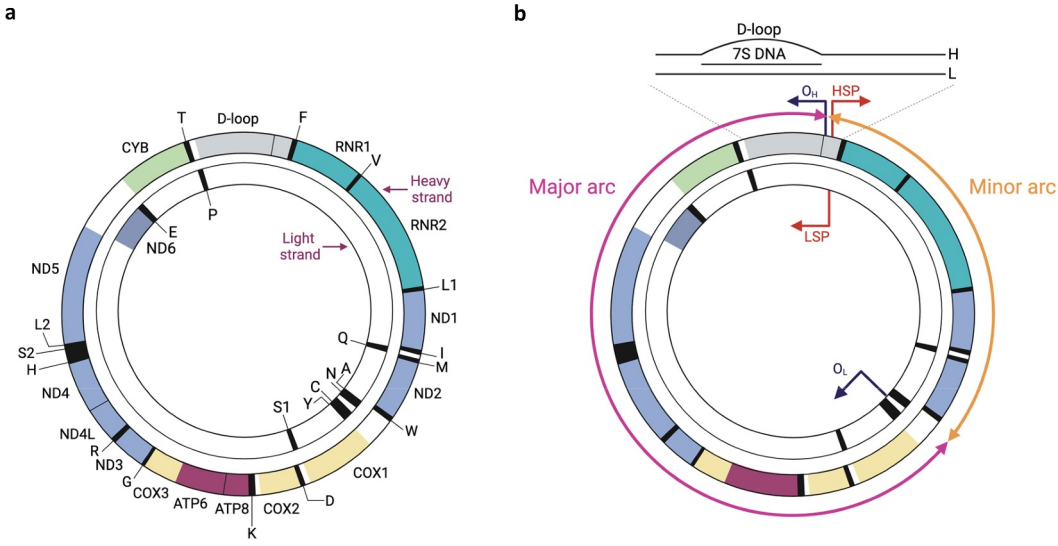

Figure 1. **Mitochondrial DNA**. (A) Mitochondrial DNA (mtDNA) structure highlighting 13 genes encoding 13 oxidative phosphorylation proteins, 22 tRNAs (black) and 2 ribosomal RNAs (green). Control region including D-loop and regulatory elements is represented in grey| (B) mtDNA demonstrating the origins of mtDNA replication (O_H and O_L), the transcriptional promoters (HSP and LSP), and major (pink) and minor (orange) arcs.

the fact that this is the typical location for primer placement for mtDNA enrichment means it is underrepresented in sequence reads from PCR-enriched samples (Figure 2A).

Human mtDNA displays several characteristics that add to the complexity of PMD diagnostics. These include tissue heteroplasmy, temporal heteroplasmy variation, increased mutability, nuclear-encoded maintenance, age-related mutation accumulation, and NUMTs (nuclear mitochondrial DNA segments) [G]. These are defined and explored further in Table 1.

2.1.2. Historical perspective of mitochondrial DNA diagnostics

The first molecular cause of a PMD, mitochondrial myopathy due to large-scale deletions of mtDNA, was identified at UCL Queen Square Institute of Neurology, London, by Holt, Harding, and Morgan-Hughes in 1988 [33]. The most prevalent variants accounting for mitochondrial syndromes [LHON, KSS, Mitochondrial Encephalomyelopathy with Lactic Acidosis and Stroke-like Episodes (MELAS), and Myoclonic Epilepsy with Ragged Red Fibres (MERRF)] soon followed [34–38]. In the pre-next generation sequencing era, diagnostics were slow and involved Sanger sequencing and Southern blotting, which were relatively insensitive to low-level heteroplasmy. In addition to molecular techniques, diagnostics also include histopathological and biochemical correlations with genetic defects [4]. A detailed discussion of these important methods is outside of the scope of this article.

2.1.3. Current strategies in mitochondrial DNA genetic diagnostics

The current state of the art in mtDNA diagnostics increasingly centers on targeted NGS of mtDNA. Following improvements in enrichment techniques, especially two-amplicon and single-amplicon long-range PCR (LR PCR), mtDNA is now widely enriched prior to sequencing using this method [39,40]. NGS on PCR-enriched mtDNA delivers extremely deep coverage.

Depth (the number of times a given locus is sequenced in NGS data) is important in mtDNA diagnostics. In nDNA sequencing, a depth of ~30× is typically sufficient for the reliable calling of variants (e.g. a heterozygous variant would be expected in roughly 15/30 reads). However, since mtDNA variants are heteroplasmic, they may only be present in a small minority of reads if the heteroplasmy level is low. At a depth of 30× a low-level heteroplasmic variant might only be present in a single forward or reverse read or might be entirely absent. Furthermore, as random errors occur in NGS reads, it may be difficult to distinguish an error from a true heteroplasmic variant at 30×. Diagnostic laboratories overcome this issue by using very deep sequencing of mtDNA. Deeply sequencing a locus (e.g. 1000×) increases the likelihood that a variant will be identified in multiple forward and reverse reads, differentiating it from a random error (true variants should be present in an equal proportion of forward and reverse reads). As a result, laboratories typically use a depth in excess of 500×.

Notably, other mtDNA enrichment techniques have been developed, including rolling circle amplification, the concentration of mitochondria using centrifugation, and enzymatic digestion of linear (nuclear) DNA [41–46]. However, these are not routinely applied in most diagnostic laboratories. Although NGS is the mainstay, the exact methodologies and ordering of molecular tests vary depending on the diagnostic laboratory, specific patient presentation, fitness for invasive testing, and clinical urgency. For example, recurrent SNVs (e.g. m.3243A>G, m.8993T>G/C m.8344A>G, m.11778 G>A) may be screened in blood with inexpensive non-NGS methods [(e.g. Sanger sequencing or restriction fragment length polymorphism analysis (RFLP)] in some jurisdictions. Other laboratories may progress straight to broader NGS of the mitochondrial genome for time-saving efficiency (mtDNA sequencing allows common and rare SNVs to be identified or out-ruled concurrently) [47]. A major consideration when testing is choosing DNA from the appropriate tissue. While most pathogenic mtDNA variants can be identified in children and young people (under 20 years) in blood-extracted DNA, the

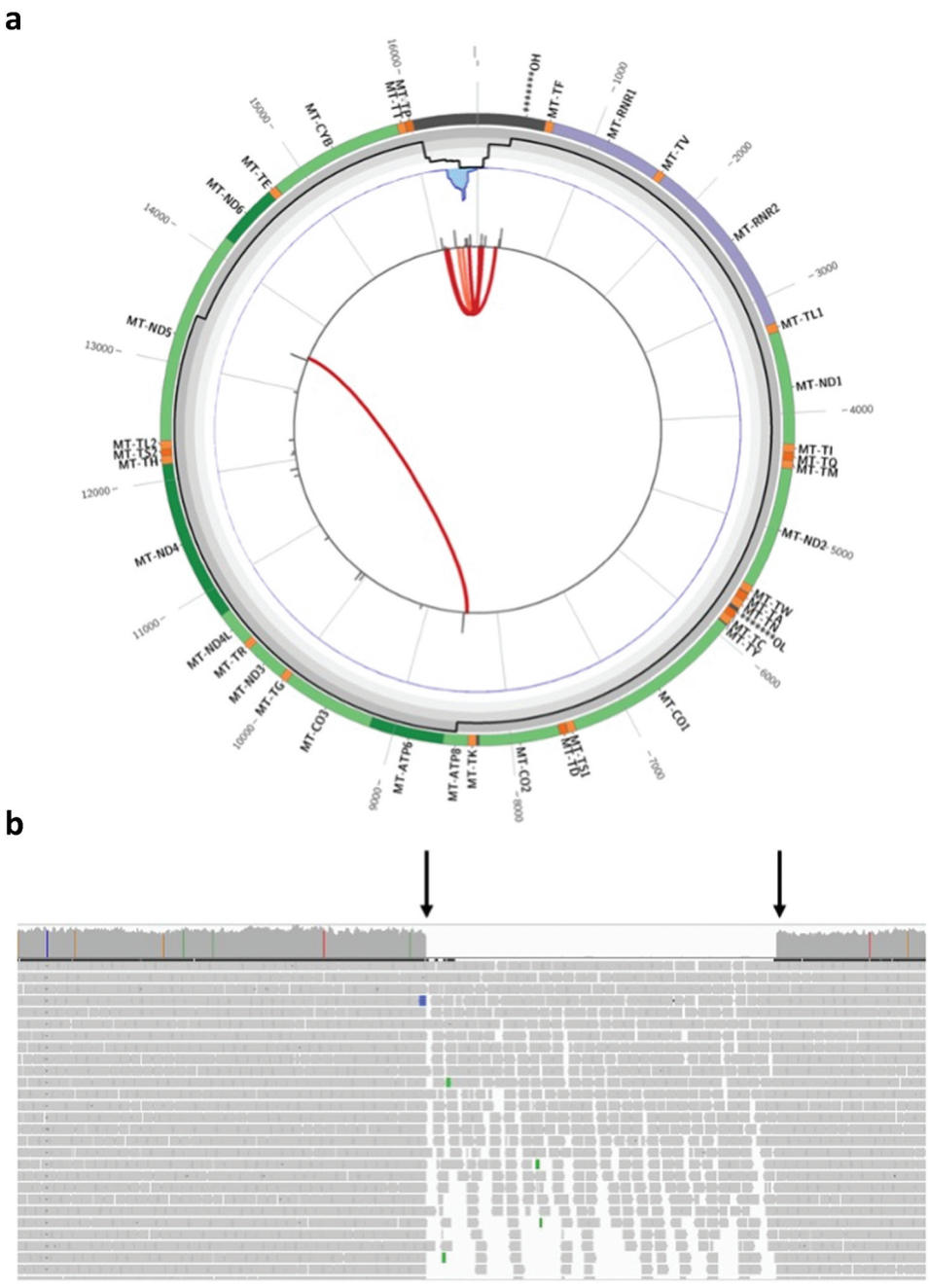

Figure 2. Single deletion in mitochondrial DNA | 2A shows the common deletion (m.8470 –13,477) as represented in a Circos plot generated using eKlipse (red arc links deletion breakpoints). Note the drop-off in coverage at the D-loop secondary to the primer placement and alignment difficulties which are incorrectly called as deletions. | 2B shows the same patient's deletion as evidenced by a drop-off in coverage visualized using IGV (Integrative Genomics Viewer).

same is not true for adults. Some variants, such as single deletions, may be lost from blood very early in childhood. The heteroplasmy levels of some SNVs fall over time, while others accumulate in post-mitotic tissues [28,48–50]. Consequently, testing muscle-extracted DNA is often mandated in adults.

Invasive testing has long played an important role in the investigation of PMDs. Tissue biopsies, such as muscle and liver biopsies, can provide valuable insights into biochemical dysfunction e.g. through respiratory chain enzyme analysis and histopathological abnormalities, such as ragged red fibers and cytochrome *c* oxidase (COX) negative fibers. Findings can be used to help classify mtDNA variants e.g. the level of a putative heteroplasmic mt-tRNA should segregate strongly with a COX-negative staining pattern if it is disease-causing. Separately, identifying a higher level of a heteroplasmic variant in muscle (a post-mitotic tissue), compared to blood (a dividing tissue) is supportive of the variant being deleterious as it suggests a negative effect on cell fitness. Additionally, mtDNA from post-mitotic tissues e.g. liver and muscle are needed for accurate measurement of certain types of mtDNA-changes, including depletion in copy number and multiple deletions.

Table 1. Mitochondrial DNA characteristics that complicate molecular diagnostics [28–32]. Abbreviations; MDMDs, Mitochondrial DNA Maintenance Defects; mtDNA, mitochondrial DNA; nDNA, nuclear DNA; NGS, next-generation sequencing; NUMT, nuclear mitochondrial DNA segments; PMD, primary mitochondrial disease; SNVs, single nucleotide variants.

mtDNA feature	Definition	Implication for Diagnostics
Heteroplasmy	The co-existence of wild-type and mutant mtDNA in an individual.	mtDNA variants may be present at a lower variant allele fraction than in medelian disease warranting dedicated variant calling strategies in NGS data. mtDNA must be sequenced to a sufficient depth to identify low-level heteroplasmy.
Temporal heteroplasmy variation	The proportion of some heteroplasmic variants can change over time e.g. deleterious SNVs or deletions may be selected against in a tissue with rapid turnover e.g. leukocytes; conversely some mutations may accumulate in postmitotic tissues e.g. muscle or brain.	Some diagnoses may require DNA samples from multiple tissues e.g. blood, urinary epithelial cells and muscle cells.
Increased mutability	mtDNA lacks the sophisticated repair mechanisms of nDNA, therefore individually rare but benign 'polymorphisms' may be encountered relatively commonly.	Clinical scientists cannot rely on rarity alone when filtering out variants and must consider additional factors including rare haplogroup signatures, conservation and (predicted) consequence. Heteroplasmy measurement is often warranted across multiple tissues to assess how deleterious a rare variant is.
Nuclear encoded mtDNA maintenance	Dominant and recessive variants in nuclear encoded mtDNA maintenance genes may in turn have a deleterious effect on mtDNA (MDMDs).	Multiple deletions, copy number depletion and accumulation of SNVs may occur secondary to an nDNA variant. These require specialist bioinformatic approaches and simultaneous testing of nDNA.
Age-related and secondary disease-related deletion accumulation	Variants, especially deletions, may accumulate in post-mitotic tissues over time unrelated to PMD. These may be due to aging or a non-PMD disease e.g. inclusion body myositis.	Unpicking which variants are due to unidentified MDMDs and which result from aging or disease can be challenging. High resolution assessment of deletion patterns, and emerging machine learning tools may help differentiate these groups in the future.
NUMT	Non-functioning nuclear pseudogenes which originate from mtDNA.	NUMT sequences can closely resemble mtDNA and cause issues with mapping. This could affect accuracy of heteroplasmy measurement.
Circular genome	mtDNA is a circular structure, resembling a plasmid.	Genomic alignment tools are built to map reads to linear genomes meaning mapping may be affected at the start/end site of mtDNA.

The subtype of the variant in mtDNA is an important consideration when choosing the testing methodology: 1) heteroplasmic SNVs are typically readily identifiable in NGS data if coverage is of sufficient depth; 2) LSRs (including single deletions, multiple deletions, and duplications) have traditionally been challenging to define and identify with NGS in clinical testing and are often tested for using LR-PCR or digital droplet PCR (ddPCR); 3) mtDNA copy number (depletion testing) is not possible with most NGS approaches, though this may change in the future (see below) and is typically testing for using ddPCR.

3. Mitochondrial DNA in whole exome and whole genome sequencing

Both WES and WGS include mtDNA and nDNA. WES has been adopted almost universally over the last 10–15 years, with unprecedented success in enhancing diagnostic rates and discovering novel causes of nDNA PMD [51–54]. However, its utility for mtDNA diagnostics has limitations. As WES uses baits to capture exons for library preparation, mtDNA must be specifically targeted; otherwise, off-target reads that incidentally capture mtDNA must be analyzed. Several publications have successfully reported mtDNA diagnoses in WES [55–58]. However, both approaches to mtDNA in WES (targeted and off-target) fail to achieve very deep coverage, thus identification of low-level heteroplasmic variants is limited. As such, this technique is most suited for opportunistic, but not diagnostic mtDNA analysis. In comparison, modern WGS does not involve a targeted library capture with PCR, hence it benefits from the naturally occurring high copy number of mtDNA to deliver very deep sequencing (>1000×) which is sufficient for reliable heteroplasmy calling. WGS also allows simultaneous uniform nDNA coverage at ~30×. This optimal genome-wide coverage has prompted calls for a WGS-first approach, avoiding targeted testing for recurrent mtDNA variants and separate sequencing of nDNA panels [59]. WGS is especially useful in young people under 20 years in whom levels of pathogenic variants are more likely to be detectable in blood [59,60]. Several studies from the UK, Australia, and Egypt have affirmed the utility of WGS in diagnosing both nDNA PMD and heteroplasmic SNVs in mtDNA [61–65].

The WGS-first approach introduces a number of diagnostic opportunities: 1) PMD phenocopy genes; for example, neuromuscular or neurodevelopmental conditions, can be difficult to discern clinically but can be identified by applying additional virtual panels or using a gene agnostic trio approach while screening PMD genes [64]; 2) dual pathology compound genetic phenocopies in an individual with two separate genetic diseases) can also be detected in WGS data with careful application of virtual panels; 3) digenic diseases (where variants in mtDNA and nDNA combine to give rise to a disease) can theoretically be identified, although this is a complex undertaking [66–68]; 4) mtDNA copy number, an important measure for children with mtDNA depletion syndromes, can potentially be estimated by comparing the proportions of nDNA and mtDNA reads in WGS data [69]; 5) assessment of nDNA background in which mtDNA disease variants occur, which might enable more precise counseling for risk of future symptoms. For example, the risk of stroke-like episodes and psychiatric involvement

in m.3243A>G-related disease was recently localized to a small number of regions of nuclear regions with relatively large effect sizes [70]. Finally, the nuclear background may provide insights regarding the susceptibility to falling mtDNA copy number and heteroplasmic variant accrual [71].

As an 'all in one' genetic test WGS is likely to offer improved diagnostic efficiency to patients. Whether this approach will also deliver a cost saving will depend on several factors, including the availability of WGS infrastructure (sequencers, staffing, bioinformatic pipelines, and data storage) and the scale of the testing. For example, in a healthcare setting where first-line WGS is already in use at scale (e.g. the UK) the existing infrastructure and the economy of scale could deliver an overall cost saving over multi-step testing (WGS replacing targeted NGS of mtDNA, WES for nuclear gene panels, and LR-PCR for rearrangements). By comparison, in jurisdictions where WES is the mainstay of testing for rare diseases, a first-line WGS-based approach for PMD may prove very expensive, especially in cases where basic testing such as RFLP or Sanger sequencing could deliver a diagnosis inexpensively (e.g. 3243A>G or common LHON variants). Ultimately, health economic studies will be needed to inform the most cost-efficient methodology.

Given the opportunities it presents, it is likely that the investigation of suspected PMDs, as with other genetic disorders, will utilize WGS of blood DNA as the mainstream first-line molecular diagnostic tool. Adopting WGS as the primary tool to diagnose PMDs will also minimize the invasive investigations often required for metabolic and pathological tests (e.g. muscle, skin, or liver biopsies), particularly in children. Previously, we have published on the barriers facing the introduction of first-line WGS from blood DNA; namely, the unavoidable need for muscle biopsies in adults with muscle-specific mtDNA variants, and for muscle tissue for functional and histological studies, and the lack of accurate and sensitive bioinformatic tools for structural variants including multiple deletions and duplications [60]. One alternative strategy could be to accept the disadvantages of an early muscle biopsy and undertake WGS directly on muscle-extracted DNA, particularly in complex, presumed genetic myopathies in adults. This will deliver reliable mtDNA sequence data, specifically for SNVs and single mtDNA deletions, which can be detected using standard NGS CNV (copy number variant) callers [62].

4. Bioinformatic innovations for mitochondrial DNA in next-generation sequencing data

4.1. Improvements in mitochondrial DNA alignment and variant calling

Most bioinformatic variant callers are designed to identify heterozygous and homozygous SNVs present at a standard Variant Allele Fraction (VAF) (i.e. ~50% or ~100% of reads) and therefore some will disregard low VAFs as sequencing errors. However, since heteroplasmic mtDNA variants can be present in any VAF, the pipeline utilized for variant calling in NGS must call sequence variants at any level. This issue is common to somatic cancer mutation calling, and the Mutect2 somatic cancer variant caller has been utilized successfully for mtDNA analysis [58,72]. A variety of tools specifically designed for mtDNA SNV or insertions/deletions (INDELs) calling and/or mtDNA mapping and mtDNA copy number estimation is now available, including Mity, MToobox, mt-DNA Server (Mitoverse), MitoHPC (Mito High Performance Caller), and MitoScape [69,73–76] (See Table 2). Mity has been employed in two large studies investigating the diagnostic utility of WGS in PMD diagnostics [62,63]. It has been shown to be ultrasensitive to low-level heteroplasmy, identifying variants that are well below 1% VAF and demonstrating an impressive correlation with gold standard clinical techniques for heteroplasmy measurement. These tools take several approaches to overcome mtDNA/NUMT mapping problems with variable shortcomings, primarily due to reliance on unique sequence alignment to mtDNA or hard filtering of possible NUMTs [76]. For example, MToolbox only includes reads that singularly map to mtDNA (not to mtDNA and nDNA), which though useful, could potentially remove genuine mtDNA reads leading to inaccurate heteroplasmic VAFs [74]. Similarly, Mity only includes mtDNA reads with extremely high mapping quality scores, thus potentially excluding true mtDNA reads containing unusual variants [73] though this concern has not been born out in clinical studies. Other tools, such as 'mt-DNA Server,' remove NUMTs by identifying variants located close to known NUMT-flanking sequences, which are not included in analysis [75]. Again, this may remove consequential variants and the method relies on NUMT sequence knowledge, which is poorly described and constantly evolving.

The MitoHPC tool from 2022 uses a novel assembly method [69]. It performs an initial mapping of all reads to the whole genome reference to identify reads that map to both mtDNA and NUMTs regions. NUMT regions are determined agnostically by identifying areas of read pile-up when known mtDNA reads are mapped to nDNA. Unmapped reads, where only one mate in a pair maps to the mtDNA sequence, were identified and included in further analysis of NUMT regions. The use of a circularized mapping delivered uniform coverage, including across the control region/D-Loop. Variants are called on this remapped circular alignment using any heteroplasmy calling tool (e.g. Mutect2). To further boost coverage and improve variant calling MitoHPC then uses the homoplasmic and high-level heteroplasmic variants called during this process to build a reference sequence specific to that individual. When reads are re-mapped to this new bespoke reference, MitoHPC removes false-positive calls, especially in homopolymeric regions.

MitoScape, primarily designed for use in common diseases, uses yet another novel approach [76]. A machine learning tool, MitoScape distinguishes true mtDNA from NUMTs. It is informed by several NGS datasets, including data from enriched mtDNA sequencing and WGS from wild-type cells and mtDNA depleted (rho zero) cells. MitoScape is trained on numerous features from these datasets. For example, it uses the fact that variants on paired-end reads are co-inherited due to linkage disequilibrium (haplogroups) as a feature that distinguishes true mtDNA reads from NUMTs.

Table 2. Selected bioinformatic tools for calling mitochondrial DNA variants. Abbreviations: SNV, single nucleotide variant; INDEL, insertion or deletion; GUI, graphical user interface; NUMT, nuclear mitochondrial DNA.

Tool	Notes	Website or github entry
Mutect2 (mitochondria-mode)	Widely-used variant caller suitable for heteroplasmic SNVs and INDELs. Initially designed for somatic (tumor) variants.	https://gatk.broadinstitute.org/hc/en-us/articles/360042477952-Mutect2
Mity	Ultrasensitive heteroplasmic SNV and INDEL caller for WGS data, validated in clinical studies. Input is aligned data (BAM) file. Filters out homopolymeric regions (m.302–319, and m.3105–3109). Performs an extensive annotation of variants.	https://github.com/KCCG/mity
MToolbox	Heteroplasmic SNV/INDEL caller suitable for use on WGS and WES data including off-target reads from WES. (Input can be aligned or unaligned data). Includes annotation of variants. Available as a GUI for small datasets.	https://github.com/mitoNGS/MToolBox
Mt-DNA Server (Mitoverse)	A user-friendly cloud-based application for heteroplasmic SNV and INDEL calling.Tags low complexity regions and known NUMTs.Input can be WES, WGS or targeted mtDNA data. (Input can be aligned or unaligned data. Mt-DNA Server is part of the Mitoverse suite).	https://mitoverse.i-med.ac.at/index.html#!
MitoScape	A pipeline to call heteroplasmic SNVs and INDELs in WGS primarily designed for complex disease. Uses a novel machine learning approach for extremely accurate calling and removal of NUMTs. Performed well when compared to MToolbox and Mt-DNA Server. Can be used to estimate mtDNA copy number.	https://github.com/larryns/MitoScape
MitoHPC	A pipeline to measure mtDNA copy number (as a ratio of mtDNA:nDNA coverage) in WGS data. Also calls and annotates heteroplasmic SNVs and INDELs. Performs an additional circularized alignment and generates an individual-specific mtDNA 'reference' sequence which reduces false positive variant calls. Flags homopolymeric, hypervariable and NUMT regions.	https://github.com/dpuiu/MitoHPC
MitoDel	A multiple deletion caller. Extracts unaligned reads and reads with CIGAR strings suggesting the presence of a deletion. Performs a second alignment to call multiple deletions based on split reads.	http://mendel.gene.cwru.edu/laframboiselab/
eKLIPse	Tool designed to identify multiple deletions breakpoints (in soft clipped reads) and generate Circos plots.	https://github.com/dooguypapua/eKLIPse
MitoSAlt	Allows for quantification of deletions and duplications based on split reads. Includes a second alignment step to identify split reads in mapped and unmapped data. Suitable use for on WGS data.	https://sourceforge.net/projects/mitosalt/

4.2. Copy number measurement

Several recessive MDMDs are associated with mtDNA depletion, causing severe disease in childhood. However, measuring and interpreting the depletion of mtDNA copy number is a complex process [47]. In most clinical laboratories, this is undertaken via a quantitative PCR method (either qPCR, RT qPCR, or ddPCR) on liver or muscle tissue [47]. While a very low copy number is a relatively reliable indication of a maintenance defect, intermediate levels are a challenge, and defining age and tissue-specific normal parameters is difficult. Furthermore, the quality of source tissue is important as abnormal tissue (e.g. muscle tissue with fatty infiltration) may harbor low mtDNA populations unrelated to any defect in mtDNA maintenance. A further challenge is the placement of primers; simplex experiments may include primers in a locus that has been deleted due to an mtDNA maintenance disorder, altering the number of copies detected. Several new bioinformatic methods now include copy number variant estimation via NGS data. For example, the aforementioned MitoHPC and MitoScape tools can estimate copy number as a ratio of mtDNA coverage to whole genome coverage [69,76]. While these methods are rudimentary and have not been robustly assessed head-to-head against current diagnostic tests, one can envisage that when applied to large comparative/control datasets they may become more accurate. As mtDNA copy number in muscle is also an important prognostic marker for disease trajectory, this technology may also be of use in individualized counseling [50]. Recent evidence suggests that certain nuclear haplotypes may alter the copy number/heteroplasmy level by imparting a replicative advantage to certain mtDNA molecules. Going forward, our understanding of the role of mtDNA copy number in PMD manifestation is likely to improve. Delineating the nuclear background which influences it may assist personalized disease prognostication.

4.3. Large-scale rearrangements in next-generation sequencing data

In previous years, the standard method for identification of large-scale rearrangements, including multiple deletions, involved Southern blotting of muscle-extracted DNA. While this process was highly manual and required relatively large DNA inputs, the results were reliable. Traditional Southern blotting did not require a PCR step; therefore, results did not include very low-level heteroplasmic deletions that could have a spurious relationship to the presenting symptoms. In comparison, LR PCR can potentially introduce a bias toward short amplicons, exaggerating low-level deletant molecules that may be part of normal aging or a non-PMD myopathic process [77]. While other methods, such as multiplexed ddPCR, are highly sensitive, primer placement may introduce an additional bias, and ddPCR cannot resolve breakpoints. Currently, clinical labs may use NGS to identify multiple mtDNA deletions. By visualizing alignments, one can utilize coverage drop-offs as a proxy for deletions (Figures 2B, 3B). This process is relatively straightforward for a single mtDNA deletion, and regular copy number variant callers can be employed.

However, confidently calling and interpreting multiple deletions in NGS data can be difficult for several reasons: 1) deletions are often flanked by repetitive sequences, meaning breakpoints can be difficult to resolve; 2) NGS data retains the short amplicon bias of prior PCR enrichment; 3) coverage depth drop-offs are nonspecific (e.g. they may relate to polymerase slippage in a polycytosine tract) and could be mistaken for a deletion; 4) the circular nature of mtDNA complicates the alignment of structural variants as alignment tools are built to align linear genomes.

Thankfully, progress in bioinformatic strategies has made the calling and identification of multiple mtDNA deletions more reliable, including facilitating visualization, resolving their breakpoints, and quantifying heteroplasmy levels [46,78–81,]. EKLIPse, a bioinformatic tool published in 2019, uses retrieved soft-clipped reads [G] (reads that contain 5′ or 3′ sequence that does not align with the reference because, for example, they include a deletion breakpoint) and BLAST (a tool that searches alignment sequences) to locate breakpoints that exist in both forward and reverse reads [80]. EKLIPse can quantify deletions as low as 0.5% heteroplasmy (Figure 3A). The computational method MitoSAlt, published in 2020, allows for an accurate quantification of both deletions and duplications [81]. Duplications are likely to co-exist with multiple mtDNA deletions but are hard to identify and are understudied. Moreover, some families harbor a pathogenic single large-scale duplication in the absence of other LSRs. These are important to delineate as they are highly heritable and understanding whether their breakpoints disrupt a gene is important for their interpretation. MitoSAlt removes nDNA reads from the WGS identified during preliminary alignment and performs a second alignment to map mtDNA reads, including those that failed to map during the preliminary step, to identify split mtDNA reads [G]. Since each split can represent a deletion or duplication, the method assumes that the split is due to a deletion in the first instance, but then assesses whether such a deletion would be possible. Putative deletions which would remove the origins of replication (O_H/O_L) would not be maintained during mtDNA replication, therefore these species are assumed to be caused by duplication. An additional bioinformatic tool published in 2020, ROTLA, also uses split reads and a circular alignment and is extremely sensitive to heteroplasmic deletions [46].

4.4. Insights from multiple mitochondrial DNA deletions in next-generation sequencing data

A 2021 study using eKLIPse trained a neural network to identify significant differences in deletion patterns, or 'deletion signatures' in patients with MDMDs versus age-matched controls, with a view to supporting variant interpretation [82]. In particular, they determined there was as follows: 1) a higher variability of deletion locations in MDMDs than in control DNA; 2) a greater number of deletions in MDMDs than in controls; 3) a tendency for deletion breakpoints to be close to perfect (uninterrupted) repeat sequences in patients with MDMDs; and 4) MDMD deletions were less likely to remove the D-loop (non-coding area which is highly conserved). Lujan *et al*. used an unsupervised machine learning tool to study the outputs from their ROTLA caller [46]. They compared deletion signatures in patients with *POLG*-related MDMD to controls and demonstrated, among other findings, that 1) a greater number of deletions were seen in patients than controls; 2) repeat sequences were more characteristic of LSR breakpoints in controls than patients. Moving forwards, improved calling of both complex/multiple LSRs and machine learning interpretation may help with understanding their consequences and origins (e.g. whether they relate to an underlying MDMD variant). An additional insight by Lujan and colleagues was the importance of mtDNA 'ablation' level, a measure of total mtDNA lost due to deletion, as opposed to full copy number depletion. While copy number was relatively stable, ablation increased with age and was more characteristic of *POLG*-related MDMDs than age-related controls. An early study of multiple deletions in aging using long-read sequencing has also delivered interesting insights (see 5.3.1 Targeted Long-read sequencing of mitochondrial DNA).

4.5. Bioinformatic developments aiding mitochondrial DNA interpretation

nDNA variant interpretation has benefitted from consistent improvements in pathogenicity prediction scores, which increasingly incorporate unsupervised machine learning, rich conservation data, physicochemical considerations, and transcriptomic datasets [83–85]. Comparatively, mtDNA lacks strength *in silico* pathogenicity prediction methods. Protein-coding mtDNA genes can be interpreted in a similar fashion to nDNA, and some useful mtDNA-specific prediction models have been developed; for example, APOGEE, a 2017 machine learning tool trained on other pathogenicity prediction tools and known classified variants which has recently been optimized further (APOGEE2), and Mitoclass1, another machine learning tool trained with a curated list of missense variants from MitoMap [86–88]. However, variants in RNA genes, which dominate mtDNA-related diseases, remain challenging to interpret and are underserved by bioinformatic innovations. PON-mt-tRNA is an early machine learning tool trained on conservation, structural implications, and sequence context data [89]. It delivers more accurate pathogenicity predictions if paired with clinical data (segregation, biochemical, and histochemical). MitoTIP predicts the pathogenicity of tRNA point substitutions or one base deletions, considering their position within a tRNA molecule. Briefly, the MitoTIP model amalgamates the position of known benign and pathogenic variants in a generic tRNA molecule and takes into account the steric effect of the variants, to deliver a pathogenicity score [90]. While useful, these tools have shortcomings. To maximize accuracy, PON-mt-tRNA requires being informed of clinical data that is difficult to obtain (e.g. muscle biopsy), making it less useful for fast screening of variants. MitoTIP uses existing classification data, which can contain significant errors or oversights. In 2022, the informatic tool MitoVisualize was published. This software facilitates the visualization of mtDNA variants within the secondary structure of the relevant tRNA or rRNA, crucial for variant

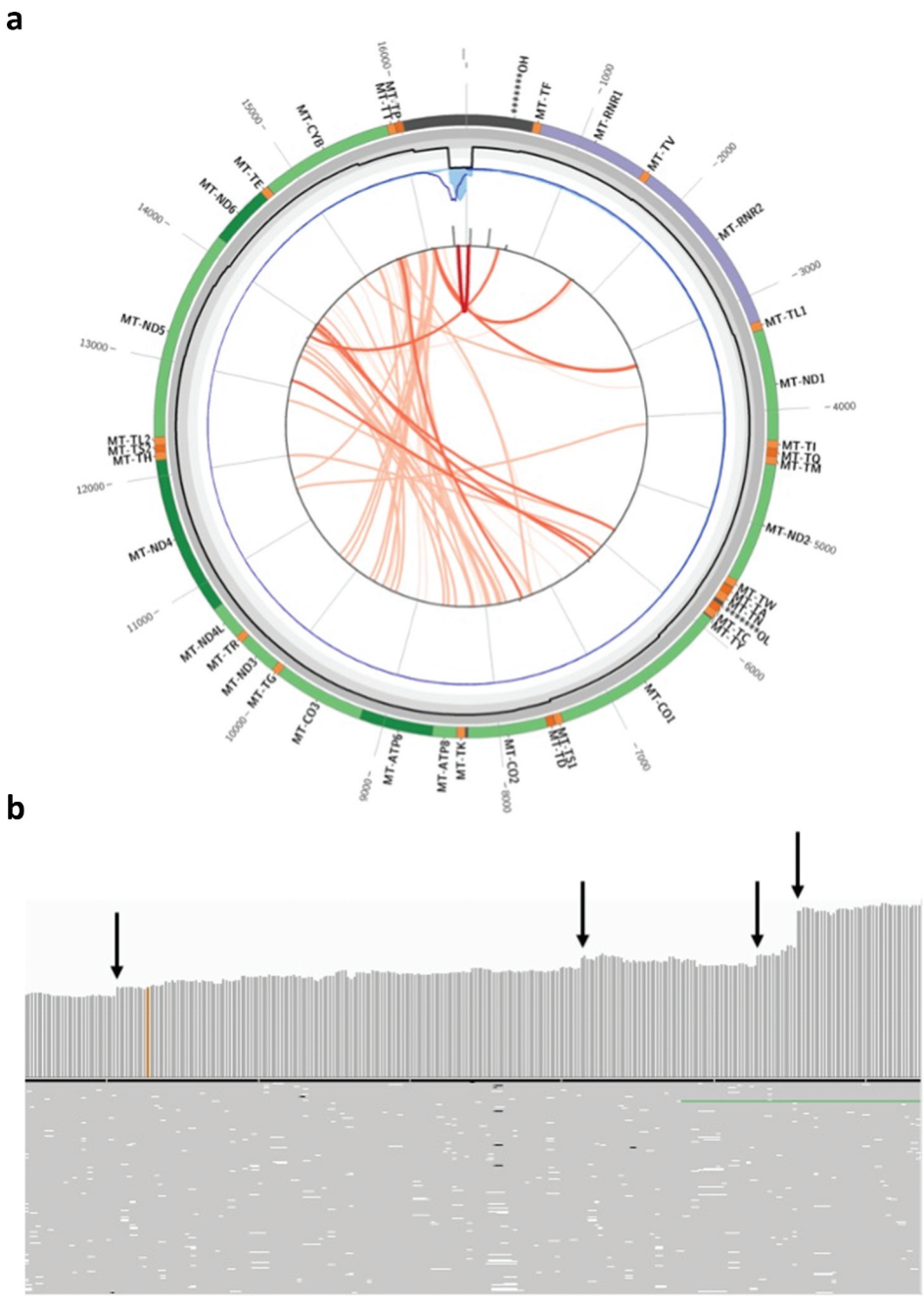

Figure 3. Multiple deletions in NGS | 3A is a Circos plot generated with eKlipse with deletion arcs illustrating the position of deletion breakpoints.| Panel B shows a coverage panel from IGV for the same patient. Black arrows delineate sharp drop-offs in coverage that are assumed to be due to deletions, however, manual inspection is neither fully sensitive nor specific.

interpretation. MitoVisualize also comprehensively annotates mtDNA variants with 1) variant frequency data and maximum heteroplasmy observed from multiple high-quality sources; 2) *in silico* predictors MitoTIP PON-mtRNA and HmtVar3 disease associations; and 3) haplogroup and conservation data.

Detailed discussions concerning the development of pathogenicity prediction tools and databases are beyond the scope of this article and are reviewed elsewhere [91].

4.6. Improvements in mitochondrial DNA control data

Specific guidance for mtDNA variant interpretation using the American College of Medical Genetics and Genomics (ACMG) criteria was published in 2020, including detailed descriptions of the nuances of mtDNA interpretation and discussion of mtDNA online databases [92]. Interpretation of mtDNA variants has been greatly assisted by improved control datasets and understanding of haplogroups. For example, the widely utilized web interface MitoMap incorporates several control

data sets, including GenBank sequences, Helix (which includes almost 200,000 sequences) and, more recently, GnomAD data [93]. In 2022, the Broad Institute published a pipeline that was utilized to call mtDNA variants over 10% heteroplasmy in the WGS arm of the GnomAD database [94]. GnomAD is undoubtedly the preeminent control dataset for modern variant interpretation. Importantly, it continues to grow and diversify the population datasets included. Consequently, the inclusion of mtDNA frequency data represents a significant practical advance for the mitochondrial medical community. A major benefit of large genomic databases is their utility in informing constraint [G] models [95–99]. However, until recently such an approach has not been possible for mtDNA. By leveraging GnomAD mtDNA sequences, researchers have created constraint metrics for mtDNA protein-coding genes, RNA genes, and the control region. Furthermore, they delivered high-resolution constraint metrics including base substitutions at every position mitochondrial local constraint (MLC) score. The MLC score generally tallied well with known pathogenic variants, however it is not completely sensitive to pathogenicity, as the score is influenced by the constraint of surrounding regions not only the locus under scrutiny.

5. Long-read sequencing of mtDNA

5.1. Long-read sequencing in mitochondrial DNA diagnostics

Long-read technologies can deliver sequences of several thousand base pairs in length, compared with 100-200bp seen in NGS. Consequently, the entire mitochondrial genome of ~16.6kb can be sequenced within a single read. Increased read lengths offer several advantages over NGS. In the context of PMDs, long-read sequencing may allow 1) improved analysis of heteroplasmic deletions contained in a single read; 2) identification of complex, very large rearrangements that are difficult to identify with short reads (e.g. duplications); 3) superior localization of NUMTs to the nuclear genome is facilitated by the ability of long-reads to include both the NUMT and the adjoining nuclear sequence; 4) phasing of variants to specific mtDNA molecules. The phasing of mtDNA variants confers several advantages. In the rare occurrence of two pathogenic mtDNA variants being identified in a single individual, understanding whether both are '*in cis*' influences counseling regarding the heritability of one or two conditions in a family. Furthermore, long-reads of the entire mtDNA which fully phase variants could facilitate *de novo* assembly (i.e. long-reads would allow sequencing of an individual's mtDNA without the need to map it to a reference for assembly and, hence, avoid errors that can occur during the mapping process). Phasing of variants to long-reads may mean that heteroplasmic variants are called more reliably and identified at lower coverage than the >500× required for NGS.

However, traditionally long-read sequencing technologies have not been without shortcomings; most notably, identification of SNVs and small INDEL is less accurate than, especially in repetitive and homopolymeric regions of NGS (even for homoplasmic variants). While improvements in small sequence variant accuracy have been achieved and are discussed below, translational PMD research supporting widespread diagnostic implementation of long-read sequencing is limited at the present time.

5.2. Improvements in Oxford Nanopore Technologies and Pacific Biosciences sequencing

The main platforms currently in use for long-read sequencing are Oxford Nanopore Technologies (ONT) sequencing and Pacific Biosciences (PacBio) [60,100,101]. A detailed description of these methods is beyond the scope of this review. In brief, ONT sequencing involves long strands of DNA driven through a membrane pore, thereby perturbing a current. The current perturbation is specific to the nucleotide passing through the pore, enabling sequence detection. PacBio uses a single molecule real-time (SMRT) approach; a long, double-stranded DNA fragment is circularized using a special adaptor, and a complementary strand is synthesized using fluorescently labeled nucleotides. The addition of each complementary nucleotide is recorded in real-time by a specialized camera and individual fluorescence signatures for each nucleotide allow the sequence to be registered.

The greatest shortcoming of long-read technologies is their relatively high level of inaccuracy, including INDELs and substitutions [102]. ONT and, to a lesser extent, PacBio reads can struggle with homopolymeric sequences. However, both technologies consistently release new chemistry, sequencers, and informatics approaches, and both companies now claim read accuracy in excess of 99% [103,104].

PacBio accuracy has traditionally been the most accurate long-read technique and has improved significantly with the introduction of circular consensus sequencing [G], where multiple rounds of sequencing of the same sample DNA molecule generate a high fidelity or 'HiFi' read [105,106]. Each pass of the polymerase creates a 'subread' which may contain random errors; however, as subreads with differing errors are then combined into a consensus read, the number of random errors is reduced. The latest PacBio hardware, the so-called 'Revio' system, combines HiFi reads with DeepConsensus from Google Health, a machine learning tool which uses an encoder to further improve accuracy [107]. Separately, significant improvements in informatics, hardware, and chemistry for ONT deliver more accurate sequencing. For example, PEPPER-Margin-DeepVariant is a variant calling pipeline developed with ONT data, which is 'haplotype aware', i.e. it labels aligned reads with their haplotype and phase and uses this information to boost accuracy when calling potential variants [108]. This technology can also be applied to PacBio. ONT has seen consistent improvements in chemical technology. For example, its 'Q20+' chemistry improves accuracy by consecutively sequencing both strands of double-stranded DNA molecule, effectively sequencing each oligonucleotide twice [104]. Furthermore, a 2020 study demonstrated that modified nanopores (combined with a second protein to produce two constrictions in the pore) can alter electrical signal modulation during DNA transit, resulting in improved accuracy in homopolymeric sections. Such 'double constrictions' are now utilized in ONT's R10 nanopores [109].

5.3. Mitochondrial DNA sequencing using long-read technologies

As with conventional NGS, mtDNA can be sequenced as part of a whole genome approach or a targeted approach, with mtDNA enrichment prior to sequencing.

5.3.1. Targeted long-read sequencing of mitochondrial DNA

To date, a wide range of enrichment techniques have been employed for targeted long-read mtDNA sequencing. The first clinical ONT study for mtDNA was released as a pre-print in 2019. It utilized LR PCR to enrich mtDNA prior to ONT sequencing [22]. This technique allowed for a reasonable input size (1 µg of LR PCR-enriched mtDNA); however, the PCR resulted in bias amplification of deletant molecules [G] over longer wild-type mtDNA [22]. In another study, Luth et al. undertook targeted ONT on LR PCR-enriched mtDNA. Researchers assessed the performance of various aligners and variant callers by creating samples with different proportions of haplogroups as a proxy for heteroplasmic variants. The mean read length achieved was 11kb. Data showed relatively high levels of false-positive calls in low complexity/homopolymeric regions [110]. These false-positive rates were affected by the aligner used, with Minimap2 performing best. In particular, the authors found that the identification of low-level variants is very dependent on the variant caller utilized, with Mutserve2 being the most sensitive [75]. Notably, these studies were undertaken prior to the release of Q20+ chemistry and R10 flow cells, which may have influenced their capacity to call break points in repetitive regions, and false-positive calls in homopolymeric stretches. Unpublished work from PacBio available online used mitochondrial isolation before DNA extraction and SMRT sequencing [111]. However, coverage depth was a significant issue with this early method, and it did not include circular consensus sequencing.

RNA-guided Cas9 is increasingly employed to enrich DNA for targeted long-read sequencing [112–114]. Two studies from 2022 built on previous work using specific Cas9 cleavage sites as a target for adaptor ligation on target DNA sequencing by ONT, removing the need to amplify mtDNA [115,116].

First, Vandiver *et al.* developed Nanopore Cas9-Targeted Sequencing (nCATS) which used a single Cas9 guide RNA to linearize mtDNA in a large sample of genomic DNA. Genomic DNA was dephosphorylated in advance so free ends of nDNA could not serve as adaptor targets. ONT adaptors were subsequently ligated to the mtDNA cut sites [116]. While the coverage depth varied widely, several thousand reads were successfully sequenced for all samples, suggesting depth would be sufficient for clinical usage. In a follow-up research paper, the authors optimized their bioinformatic strategies to enable more reliable calling of large deletions by improving deletion detection at chimeric (split) reads [117]. The authors sequenced mtDNA from control human muscle across a range of ages, revealing a wide spectrum of deletions. Their work suggested an increased abundance of deletions with age, with a correlation between major arc deletions, and deletions spanning major and minor arcs. Second, Keraite *et al.* developed a protocol with a novel four-step enrichment [115]: 1) the sample was pre-treated with Exonuclease V to digest any linear nDNA, to enrich circular mtDNA and decrease the risk of NUMT contamination; 2) five prime ends of linear DNA (nDNA) were then dephosphorylated to prevent inadvertent ligation of ONT adaptors; 3) DNA was divided into two or more aliquots, and each aliquot was treated with a dual-RNA guided Cas9. A different guide RNA targeted a specific site in each aliquot, with the cut site serving as a barcode unique to that aliquot, allowing a multiplexed approach. Using this approach, the number of cut sites could be increased to optimize the capture of deletant molecules. Finally, the Cas9 was degraded with proteinase K and ONT adaptors were ligated to the cut sites. When the authors compared their Cas9 approach to LR PCR enriched NGS and ONT data, the authors demonstrated a significant short read bias from the PCR in a multiple deletion sample, which overestimated the prevalence of large deletions. Developments in ONT Cas9 sequencing kits are likely to further improve this technology in the future.

5.3.2. Long-read whole genome sequencing with mitochondrial DNA analysis

A 2019 study used single-pass PacBio sequencing on three mtDNA samples and compared this with NGS approaches. The depth of coverage was suboptimal and limited the ability of the long-reads to decipher heteroplasmic variants, though future work using HiFi reads is likely to improve this [21]. A recent study compared two different ONT whole genome sequencing kits: 1) the 'Rapid Sequencing Kit' – this method randomly fragments DNA (using a transposase), does not repair fragmented ends, and hence decreases adaptor ligation, thus delivers less biased, but lower yield, sequencing; and 2) the 'Ligation Sequencing Kit' – this performs dA tailing [G] to repair DNA ends, optimizing them for adaptor ligation [118]. An endonuclease digestion step was also included with the ligation kit to linearize the mtDNA. The authors demonstrated that the endonuclease cleavage site needed to be chosen carefully. One cleavage site (using BAMHI) in the major arc was often located within large deletions and resulted in coverage aberrations and affected heteroplasmy quantification. The ligation sequencing method delivered longer total reads and mtDNA reads and deeper mtDNA coverage, though the use of an endonuclease introduced coverage biases even when falling outside of a deletion. This may limit its utility for studying multiple deletions and complex rearrangements (e.g. duplications). A comparison of the two methods utilized on the same samples showed that the rapid method had a 4.4kb average length and an average depth of coverage 545× for mtDNA (MinIon R9.4.1 Flow Cell). In comparison, the ligation method delivered reads with 6.2kb in average length and a depth of coverage 921×. This study also confirmed the difficulties with accurately identifying SNVs and INDELs with ONT, though the authors acknowledged this is likely to substantially improve with more recently released chemistry/flow cells.

5.4. Large-scale rearrangements and long-read sequencing

In theory, identification, definition, and quantification of LSRs will be superior to long-read technologies for several reasons:

(1) *Coverage uniformity*. Long-reads allow superior mapping of raw sequence reads for the circular mitochondrial genome without coverage drop-off at the start and end of the genome (provided PCR primers in the control region are not utilized and an appropriate alignment is performed). Long-reads cross the start point of the mtDNA sequence (control region) with substantial anchoring sequences on either side. A split long-read can be easily aligned to the start and end of the reference sequence. This more uniform coverage is beneficial when using depth of coverage as a proxy for deletion/duplication identification. Indeed, several studies have shown excellent uniformity of mtDNA coverage [115,116,118].
(2) *Breakpoint resolution*. Long-reads are more likely to encompass a deletion within a particular read and should be more easily called than NGS data, which struggles with SV and CNV calling. Theoretically, long-reads will deliver better breakpoint resolution. While early studies struggled to define breakpoints at homopolymeric sequences [22], this technical challenge is being overcome through improved quality of both the library preparation kits and the sequencing flow cells [109].
(3) *Multiple mtDNA deletion calling*. Identification of multiple mtDNA deletions with long-read data is theoretically straightforward. Individual deletions will be contained within a single read, and hence easily defined. Age-related deletion accumulation measured with ONT has successfully identified a range of deletions across different ages, with a pronounced accumulation after the age of 60 years [116]. However, ensuring an accurate representation of each deletion subpopulation when using nonrandom fragmentation of DNA is complex (see cut site bias discussion below) [115,116,118]. Additionally, unlike NGS data, bioinformatics research to develop multiple deletion callers in long-reads is at an early stage (see bioinformatics bottleneck section).
(4) *Avoiding PCR bias*. As with traditional WGS, long-read technologies are amenable to a PCR-free approach compared with WES and targeted NGS of mtDNA. Long-read sequencing with PCR enrichment has been demonstrated to exhibit significant distortion in the proportion of large deletion subpopulations in a multiple deletion sample, which is overcome with a PCR-free approach to long-reads [115].
(5) *Identification of complex rearrangements*. The identification and delineation of rearrangements involving duplications is a challenge with traditional diagnostic techniques, given their large size and complicated configuration. However, their identification is highly relevant clinically due to their strong heritability. Long-reads have sufficient length to incorporate large mitochondrial genomes that include a duplicated segment, therefore resolving complex rearrangements potentially overlooked by other technologies. Positional information on these duplications is highly valuable in their interpretation e.g. do they interrupt a coding sequence and e.g. potentially lead to nonsense-mediated decay?

6. Barriers to use of long-reads in clinical practice

6.1. Cut site bias

Several emerging techniques use either an RNA-guided Cas9 or other endonucleases to linearize mtDNA and maximize adaptor ligation, read numbers, and length [115,116,118]. If the site chosen for cleavage is co-located within a variant (e.g. deletion, polymorphism, or SNV), the mtDNA population including that variant will be selected against during adaptor ligation. Consider the case of the recurrent 'common deletion' (m.8470–13477), frequently encountered in PMDs. A cut site contained within this deletion might miss this diagnostic mutation or obscure a breakpoint by affection coverage. Keraite *et al.* overcame this issue by dividing the sample into aliquots, each of which were Cas9 cleaved at different targeted cut sites. Cleaving the mtDNA at a site in *MT RNR2* identified up two deletant populations that were missed with more distal cut targets. An alternative is to use a more random fragmentation method, which can identify deletions in a less biased manner; however, in doing so the read length and coverage depth may be adversely affected [118]. Interestingly, as sequencing methods become more accurate, reliability may be possible at increasingly lower depths. Future research may help optimize cut site choice by avoiding known deletion hot spots or the use of multiple aliquots with varying cleavages. The issue of mtDNA subpopulation bias is not unique to long-read sequencing; for example, primer location in clinical digital PCR causes similar issues [119].

6.2. Bioinformatic bottlenecks

Despite its shortcomings, the performance of NGS is being continuously optimized through bioinformatic innovations. Specifically, mtDNA NGS analysis has benefitted from improved alignment tools and heteroplasmy/LSR callers, as discussed above. Unfortunately, there exists a comparative dearth of bioinformatic tools for mtDNA assessment in long-read data. In some cases, more generic tools and manual interpretation can be utilized for mtDNA analysis. For example, the minimap2 long-read aligner can be used successfully for mtDNA, the Sniffles long-read CNV caller can help identify single and multiple deletions, and specifically querying CIGAR strings [G]/alignment viewers can identify individual mtDNA deletions in a multiple mtDNA deletion sample [110,116,118,120].

More bespoke bioinformatic approaches are in development for mtDNA. Keraite *et al.* showcased a novel sample preparation (multiplexed sample with Cas9 cut sites), which was coupled with a customized bioinformatic approach for ONT reads [115]. A basic alignment was performed and reads that were mapped to the mitochondrial genome were further filtered to ensure accuracy. In the demultiplexing process

(identification of the cut sites), a range of strategies were employed to select reads for inclusion in the analysis. For example, for maximum accuracy where possible only reads with a start and end close to the cut site (full-length reads) were included, though this had to be loosened where coverage depth was too low. Following this demultiplexing step, the selected reads were aligned to the reference, again using a custom remapping process to overcome poor coverage over read split points (cut sites), and bespoke alignment, and calling steps were developed to account for mtDNA-specific issues (heteroplasmic SNVs and large deletions, and molecule-specific phasing). For PacBio, a recently developed multi-species mtDNA alignment tool for HiFi reads, MitoHiFi, can extract, align, and annotate high-fidelity mtDNA reads allowing the identification of heteroplasmic mutation and exclusion of NUMTs [121].

Further bioinformatic innovations will be crucial to ensure effective translation of these technologies into the diagnostic sphere.

6.3. Depth of coverage

Depth has been an issue in several PCR-free studies [21,111]. As discussed previously, achieving a high depth of mtDNA coverage is crucial in distinguishing variants from sequencing errors. Low DNA quality and quantity impact the depth of sequencing coverage (e.g. DNA from uroepithelial cells, which is often sequenced in mtDNA diagnostics, may not be of high enough quality to achieve very deep coverage). However, as long-read technology improves to lower depth than seen in NGS, it may be sufficient for accurate variant calling [115]. Extremely accurate long-read technologies could mean a variant only needs to be identified in a very low number of reads to be called accurately. Additionally, information regarding the surrounding haplogroup will also help inform what variants are likely to co-occur in an mtDNA molecule, which will add to the accuracy of variant calling. Future work will be required to validate the accuracy of lower-depth long-read sequencing or to optimize DNA extraction and bioinformatic processes to boost depth further.

7. Conclusion

Opportunities exist for transformative changes in mtDNA diagnostics in the coming years. We have seen the publication of several pivotal studies supporting the adoption of a WGS approach to mtDNA diagnostics, and bioinformatic innovations will undoubtedly be required for this approach to evolve. Long-read sequencing platforms have improved significantly, and there has been a clear uplift in their application to mtDNA sequencing studies. While short-read WGS, including the mitochondrial genome, is close to introduction in clinical genetic laboratories, the outlook for long-reads, while promising, is less certain. The diagnostic community will need to actively engage with long-read innovations and research to ensure this technology successfully integrates within the clinical sector and offers tangible benefits to patients.

8. Expert opinion

The technical developments discussed in this review are likely to have a significant impact on PMD diagnostics. Bioinformatic innovations for mtDNA variant calling and NUMT region removal have established WGS as a reliable assay for investigating suspected PMDs. In the short term, we anticipate that the use of PCR-free WGS from blood DNA will increase in diagnostic laboratories, eventually replacing multistep genetic diagnostics (e.g. mtDNA sequencing, LR PCR, and nuclear gene panels). As pediatric patients are less likely to harbor mtDNA variants restricted to postmitotic tissues, they will particularly benefit from this trend, shortening the diagnostic odyssey associated with sequential genetic testing. The need for invasive techniques (i.e. muscle biopsies) will reduce in this group. In the adult setting, the utility of WGS from blood DNA is less clear. While many adults may be diagnosable via blood WGS, a significant subgroup will not. Instead, muscle WGS may become a first-line investigation in adults with a classical mitochondrial myopathy/CPEO presentation and in those who have already had a muscle biopsy for histological diagnostics. As a second-line investigation, muscle WGS would benefit patients with suspected PMD who have negative blood WGS. Bioinformatic innovations for muscle mtDNA interrogation show promise (e.g. delineating and interpreting multiple deletion patterns and inferring copy number), thus supporting the case for muscle WGS introduction in clinical laboratories.

Whether blood WGS replaces relatively inexpensive techniques entirely (e.g. RFLP and Sanger sequencing for m.3243A>G and LHON variants) will likely be guided by costs. WGS prices continue to fall, especially when undertaken in large numbers. Health economic studies comparing inexpensive techniques to large-scale, centralized WGS will need to be conducted to provide clarity in this area. It should be noted during this analysis that the power of WGS data is not limited to the identification of the diagnostic mtDNA variant itself. The nuclear background data obtained from WGS are important and may increasingly inform polygenic risk scores, enabling tailored screening and improving genetic counseling.

The capacity for WGS tools to identify very low-level heteroplasmic variants raises interesting questions. What is the significance of an ultralow-level SNV (e.g. m.3243A>G at <1% heteroplasmy)? In patients where this is a known familial variant, such a result can be presumed diagnostic. In contrast, if a variant is identified in a simplex case with nonspecific clinical features, how should it be interpreted? Data from GnomAD suggest that 1 in 250 individuals carry a pathogenic mtDNA variant at >10% heteroplasmy, and if variants at <10% are included, the carrier rate is likely to be even higher, meaning large numbers of unaffected people are likely to harbor known pathogenic variants. The introduction of diagnostic WGS will identify ultralow-level heteroplasmy incidentally in more patients, but the utility of this information is unclear, especially where blood is the only tissue available. Will these patients ever develop symptoms from such low-levels? How likely are these variants to expand through the maternal bottleneck and cause disease in offspring? Separately, how should we interpret low-level mtDNA

deletions? These deletions can occur somatically, and therefore could be an incidental finding. If we identify them at a very low level in blood, can they be considered diagnostic, or should muscle mtDNA be analyzed? These questions are complex; consequently, data accrual, including VAFs across multiple tissues, and in-depth phenotyping will be required to answer them. Technological advances have real-world implications when applied to diagnostics and should be employed judiciously until there is an improved understanding of their clinical implications.

Long-read sequencing also offers promise for mtDNA diagnostics, most notably because of its capacity to define large LSRs, which play an outsized causative role in PMDs. A major unanswered question is whether targeted mtDNA sequencing or full WGS is more suitable for clinical use. Enrichment with PCR will bias results, and Cas9-based techniques are time-consuming, manual, and could introduce bias. However, full long-range WGS creates very large data sets which are potentially difficult to store and process in diagnostic laboratories. More studies evaluating the clinical applicability of both approaches are needed. Another area requiring further study is that of heteroplasmy calling. Long-reads are likely to identify heteroplasmic variants reliably at lower coverage depths than NGS. However, optimal depths for variant calling remain undefined. Head-to-head studies comparing long-reads at varying depths to current heteroplasmy measurement techniques are necessary.

The medium-term outlook in this area is positive. In the next five years, we speculate that short-read WGS will be firmly established in diagnostic laboratories, replacing multi-step testing for many patients. More research is required to support the clinical adoption of long-reads; however, recent advances in accuracy are promising, and we anticipate that diagnostic laboratories will be piloting the use of these techniques within a similar time frame.

Glossary

(1) CIGAR string: Before being aligned to a reference, a sequence read may contain errors. A CIGAR string displays the operations associated with the alignment of the read to the reference, i.e. it compares the reference and read e.g. which bases align, and which are inserted or deleted when compared with the reference.
(2) Circular Consensus Sequencing: A method used in PacBio long-read sequencing whereby an individual DNA molecule is sequenced multiple times to improve accuracy.
(3) Constraint: A genomic region exhibits constraint when it has less variation than would be expected for its size/architecture i.e. it is intolerant of variation.
(4) dA tailing: dA tailing involves adding an additional adenosine base to the three prime end of a phosphorylated DNA fragment to facilitate adaptor ligation.
(5) Deletant molecule: An mtDNA molecule containing a large deletion, this is typically a shorter circular molecule than wild-type mtDNA.
(6) G-quadruplexes: Helical secondary DNA structures that are formed in sequences rich in the nucleotide guanine
(7) Homopolymeric Region: A repetitive run of nucleotides e.g. five C nucleotides in a row.
(8) Large-scale Rearrangement: A deletion, duplication or complex combination of both within mtDNA. These are typically >1kb in size and may clonally expand in post-mitotic tissues (e.g. muscle, brain, and fat) where the selective pressures of cell division do not exist, but mtDNA replication continues.
(9) NUMTs: Segments of mtDNA can become incorporated into mtDNA over time [122]. These pseudogenes may cause problems when mapping short reads in NGS data.
(10) Soft Clipped Reads: Reads which have part of their sequence masked during alignment as the sequence does not fit with the reference.
(11) Split Reads: Reads which do not align contiguously to a reference, e.g. because they contain the break point of a deletion, or cross the origin of mtDNA in a linear alignment model.

The suffix **[G]** indicates the term is defined in a glossary at the end of this article.

Abbreviations

ACMG	American College of Medical Genetics and Genomics
BLAST	Basic Local Alignment Search Tool
CIGAR	Compact Idiosyncratic Gapped Alignment Report
CNV	Copy Number Variant
COX	Cytochrome *c* Oxidase
D-Loop	Displacement Loop
ddPCR	Digital Droplet PCR
H-Strand	Heavy Strand
INDEL	Insertion/Deletion
IVF	*In Vitro* Fertilisation
L-Strand	Light Strand
LHON	Leber Hereditary Optic Neuropathy
LR PCR	Long Range Polymerase Chain Reaction
LSR	Large-scale Rearrangement
MDMD	Maintenance Defect of Mitochondrial DNA
MELAS	Mitochondrial Encephalomyopathy, Lactic Acidosis and Stroke-like episodes
MERRF	Myoclonic Epilepsy with Ragged-Red Fibres
mtDNA	Mitochondrial DNA
NGS	Next-Generation Sequencing
nCATS	Nanopore Cas9-Targeted Sequencing
nDNA	Nuclear DNA
NUMT	Nuclear mitochondrial DNA segments
O_H	Origin of Heavy Strand Replication
O_L	Origin of Light Strand Replication
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PGD	Preimplantation Genetic Diagnosis
PCR	Polymerase chain reaction
PMDs	Primary Mitochondrial Diseases
RFLP	Restriction Fragment Length Polymorphism
SNV	Single Nucleotide Variant
SV	Structural variant
tRNA	Transfer RNA

VAF	Variant Allele Fraction
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

Funding

This paper was not funded.

Acknowledgements

W.L.M., C.P., M.G.H., and R.D.S.P. received funding from The Lily Foundation. J.V. holds a fellowship from the Health Education England Genomics Education Programme. C.P. is supported by the Clore Duffield Foundation. W.M., R.D.S.P., and M.F. are supported by a seedcorn award from the Rosetrees Trust and Stoneygate Foundation. The University College London Hospitals/University College London Queen Square Institute of Neurology sequencing facility receives a proportion of funding from the Department of Health's National Institute for Health Research Biomedical Research Centres funding scheme. All research at Great Ormond Street Hospital for Children NHS Foundation Trust and UCL Great Ormond Street Institute of Child Health is made possible by the National Institute for Health and Care Research Great Ormond Street Hospital Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the National Institute for Health and Care Research, or the Department of Health. The clinical and diagnostic 'Rare Mitochondrial Disorders' Service in London is funded by the UK NHS Highly Specialised Commissioners. R.D.S.P. and M.F. are supported by a Medical Research Council (UK) Clinician Scientist Fellowship (MR/S002065/1). M.F., M.G.H, and R.D.S.P. are supported by the Medical Research Council (UK) award MC_PC_21046 to establish a National Mouse Genetics Network Mitochondria Cluster (MitoCluster). W.L.M., E.B., M.G.H., J.V., and R.D.S.P. are supported by the Medical Research Council (UK) strategic award MR/S005021/1 to establish an International Centre for Genomic Medicine in Neuromuscular Diseases (ICGNMD).

Declaration of interest

The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

Reviewers disclosure

Peer reviewers on this manuscript have no relevant financial relationships or otherwise to disclose.

References

1. Rahman J, Rahman S. Mitochondrial medicine in the omics era. Lancet Internet. 2018;391:2560–2574. doi: 10.1016/S0140-6736(18)30727-X
2. Bock FJ, Tait SWG. Mitochondria as multifaceted regulators of cell death. Nat Rev Mol Cell Biol Internet. 2019 [cited 2023 May 17];21(221):85–100. doi: 10.1038/s41580-019-0173-8
3. Giorgi C, Marchi S, Pinton P. The machineries, regulation and cellular functions of mitochondrial calcium. Nat Rev Mol Cell Biol Internet. 2018 [cited 2023 May 17];19(11):713–730. doi: 10.1038/s41580-018-0052-8
4. Gorman GS, Chinnery PF, DiMauro S, et al. Mitochondrial diseases. Nat Rev Dis Primers Internet. 2016;2(1):16080. doi: https://doi.org/10.1038/nrdp.2016.80
5. Falabella M, Minczuk M, Hanna MG, et al. Gene therapy for primary mitochondrial diseases: experimental advances and clinical challenges. Nat Rev Neurol. 2022;18(11):689–698. doi: 10.1038/s41582-022-00715-9
6. Pitceathly R, Keshavan N, Rahman J, et al. Moving towards clinical trials for mitochondrial diseases. J Inherit Metab Dis. 2020;44(1):22–41. doi: 10.1002/jimd.12281
7. Gorman GS, McFarland R, Stewart J, et al. Mitochondrial donation: from test tube to clinic. Lancet. 2018;392(10154):1191–1192. doi: 10.1016/S0140-6736(18)31868-3
8. Yu-Wai-Man P, Newman NJ, Carelli V, et al. Bilateral visual improvement with unilateral gene therapy injection for Leber hereditary optic neuropathy. Sci Transl Med. 2020;12(573):12. doi: 10.1126/scitranslmed.aaz7423
9. Yang S, Ma S-Q, Wan X, et al. Long-term outcomes of gene therapy for the treatment of Leber's hereditary optic neuropathy. EBioMedicine. 2016;10:258–268. doi: 10.1016/j.ebiom.2016.07.002
10. Mok BY, de Moraes MH, Zeng J, et al. A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. Nature. 2020;583(7817):631–637. doi: 10.1038/s41586-020-2477-4
11. Cho S-I, Lee S, Mok YG, et al. Targeted A-to-G base editing in human mitochondrial DNA with programmable deaminases. Cell. 2022;185(10):1764–1776.e12. doi: 10.1016/j.cell.2022.03.039
12. Zekonyte U, Bacman SR, Smith J, et al. Mitochondrial targeted meganuclease as a platform to eliminate mutant mtDNA in vivo. Nat Commun. 2021;12(1):3210. doi: 10.1038/s41467-021-23561-7
13. Pantic B, Ives D, Mennuni M, et al. 2-Deoxy-D-glucose couples mitochondrial DNA replication with mitochondrial fitness and promotes the selection of wild-type over mutant mitochondrial DNA. Nat Commun Internet. 2021;12(1):6997. doi: https://doi.org/10.1038/s41467-021-26829-0
14. Anderson S, Bankier AT, Barrell BG, et al. Sequence and organization of the human mitochondrial genome. Nature. 1981;290(5806):457–465. doi: 10.1038/290457a0
15. Andrews RM, Kubacka I, Chinnery PF, et al. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA [5]. Nat Genet. 1999;23(2):147. doi: 10.1038/13779
16. Kasamatsu H, Robberson DL, Vinograd J. A novel closed-circular mitochondrial DNA with properties of a replicating intermediate. Proc Natl Acad Sci U S A. 1971;68(9):2252–2257. doi: 10.1073/pnas.68.9.2252
17. Doda JN, Wright CT, Clayton DA. Elongation of displacement-loop strands in human and mouse mitochondrial DNA is arrested near specific template sequences. Proc Natl Acad Sci U S A. 1981;78(10):6116–6120. doi: 10.1073/pnas.78.10.6116
18. Gustafsson CM, Falkenberg M, Larsson N-G. Maintenance and expression of mammalian mitochondrial DNA. Annu Rev Biochem. 2016;85(1):133–160. doi: 10.1146/annurev-biochem-060815-014402
19. van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 2009;30(2):E386–E394. doi: 10.1002/humu.20921
20. Guiblet WM, Cremona MA, Cechova M, et al. Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. Genome Res. 2018;28(12):1767–1778. doi: 10.1101/gr.241257.118
21. Alkanaq AN, Hamanaka K, Sekiguchi F, et al. Comparison of mitochondrial DNA variants detection using short- and long-read sequencing. J Hum Genet Internet. 2019;64(11):1107–1116. doi: http://dx.doi.org/10.1038/s10038-019-0654-9
22. Wood E, Parker MD, Dunning MJ, et al. Clinical long-read sequencing of the human mitochondrial genome for mitochondrial disease diagnostics. bioRxiv. 2019;597187.
23. Nissanka N, Minczuk M, Moraes CT. Mechanisms of mitochondrial DNA deletion formation. Trends Genet. 2019;35(3):235–244. doi: 10.1016/j.tig.2019.01.001
24. Lynch M, Koskella B, Schaack S. Mutation pressure and the evolution of organelle genomic architecture. Science (1979). 2006;311(5768):1727–1730. doi: 10.1126/science.1118884
25. Falabella M, Kolesar JE, Wallace C, et al. G-quadruplex dynamics contribute to regulation of mitochondrial gene expression. Sci Rep. 2019;9(1):5605. doi: 10.1038/s41598-019-41464-y
26. Nelson I, Degoul F, Obermaier-Kusser B, et al. Mapping of heteroplasmic mitochondrial DNA deletions in Kearns-Sayre syndrome. Nucleic Acids Res. 1989;17(20):8117–8124. doi: 10.1093/nar/17.20.8117

27. Goldstein A, Falk MJ. Mitochondrial DNA Deletion Syndromes. In: A HH, RA P, editors. GeneReviews®. 2nd ed. Seattle (WA): University of Washington; 2019.
28. Rajasimha HK, Chinnery PF, Samuels DC. Selection against pathogenic mtDNA Mutations in a stem cell population leads to the loss of the 3243A→G mutation in blood. Am J Hum Genet. 2008;82 (2):333–343. doi: 10.1016/j.ajhg.2007.10.007
29. Rahman S, Poulton J, Marchington D, et al. Decrease of 3243 A→G mtDNA mutation from blood in MELAS syndrome: A longitudinal study. Am J Hum Genet. 2001;68(1):238–240. doi: 10.1086/316930
30. El-Hattab AW, Craigen WJ, Wong L-J, et al. Mitochondrial DNA maintenance defects overview. Gene Reviews. 2018;1–15.
31. Moslemi AR, Lindberg C, Oldfors A. Analysis of multiple mitochondrial DNA deletions in inclusion body myositis. Hum Mutat. 1997;10 (5):381–386. doi: 10.1002/(SICI)1098-1004(1997)10:5<381:AID-HUMU8>3.0.CO;2-I
32. Reynier P, Malthiery Y. Accumulation of deletions in MtDNA during tissue aging: analysis by long PCR. Biochem Biophys Res Commun. 1995;217(1):59–67. doi: 10.1006/bbrc.1995.2745
33. Holt IJ, Harding AE, Morgan-Hughes JA. Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies. Nature. 1988;331(6158):717–719. doi: 10.1038/331717a0
34. Shoffner JM, Lott MT, Lezza AMS, et al. Myoclonic epilepsy and ragged-red fiber disease (MERRF) is associated with a mitochondrial DNA tRnalys mutation. Cell. 1990;61(6):931–937. doi: 10.1016/0092-8674(90)90059-N
35. Goto YI, Nonaka I, Horai S. A mutation in the tRnaleu(uur) gene associated with the MELAS subgroup of mitochondrial encephalomyopathies. Nature. 1990;348(6302):651–653. doi: 10.1038/348651a0
36. DiMauro S, Garone C. Historical perspective on mitochondrial medicine. Dev Disabil Res Rev. 2010;16(2):106–113. doi: 10.1002/ddrr.102
37. Wallace DC, Singh G, Lott MT, et al. Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. Science. 1988;242(4884):1427–1430. doi: 10.1126/science.3201231
38. Zeviani M, Moraes CT, DiMauro S, et al. Deletions of mitochondrial DNA in Kearns-Sayre syndrome. Neurology. 1988;38(9):1339–1339. doi: 10.1212/WNL.38.9.1339
39. Cui H, Li F, Chen D, et al. Comprehensive next-generation sequence analyses of the entire mitochondrial genome reveal new insights into the molecular diagnosis of mitochondrial DNA disorders. Genet Med. 2013;15(5):388–394. doi: 10.1038/gim.2012.144
40. Zhang W, Cui H, Wong LJC. Comprehensive one-step molecular analyses of mitochondrial genome by massively parallel sequencing. Clin Chem. 2012;58(9):1322–1331. doi: 10.1373/clinchem.2011.181438
41. Gould MP, Bosworth CM, McMahon S, et al. PCR-free enrichment of mitochondrial DNA from human blood and cell lines for high quality next-generation DNA sequencing. PLoS One. 2015;10 (10):1–13. doi: 10.1371/journal.pone.0139253
42. Marquis J, Lefebvre G, Kourmpetis YAI, et al. MitoRS, a method for high throughput, sensitive, and accurate detection of mitochondrial DNA heteroplasmy. BMC Genomics. 2017;18(1):1–19. doi: 10.1186/s12864-017-3695-5
43. Wolff JN, Shearman DCA, Brooks RC, et al. Selective enrichment and sequencing of whole mitochondrial genomes in the presence of nuclear encoded mitochondrial pseudogenes (numts). PLoS One. 2012;7(5):1–7. doi: 10.1371/journal.pone.0037142
44. Ancora M, Orsini M, Colosimo A, et al. Mitochondrial heteroplasmy profiling in single human oocytes by next-generation sequencing. Mitochondrial DNA B Resour. 2017;2(2):542–543. doi: 10.1080/23802359.2017.1365634
45. Yao Y, Nishimura M, Murayama K, et al. A simple method for sequencing the whole human mitochondrial genome directly from samples and its application to genetic testing. Sci Rep. 2019;9(1):1–7. doi: 10.1038/s41598-019-53449-y
46. Lujan SA, Longley MJ, Humble MH, et al. Ultrasensitive deletion detection links mitochondrial DNA replication, disease, and aging. Genome Biol Genome Biology. 2020;21(1). doi: 10.1186/s13059-020-02138-5
47. Mavraki E, Labrum R, Sergeant K, et al. Genetic testing for mitochondrial disease: the United Kingdom best practice guidelines. Eur J Hum Genet. 2022;31(2):148–163. doi: 10.1038/s41431-022-01249-w
48. Chinnery PF, Samuels DC. Relaxed replication of mtDNA: A model with implications for the expression of disease. Am J Hum Genet. 1999;64(4):1158–1165. doi: 10.1086/302311
49. Durham SE, Samuels DC, Cree LM, et al. Normal levels of wild-type mitochondrial DNA Maintain cytochrome c oxidase activity for two pathogenic mitochondrial DNA mutations but not for m.3243A→G. Am J Hum Genet. 2007;81(1):189–195. doi: 10.1086/518901
50. Grady JP, Pickett SJ, Ng YS, et al. mtDNA heteroplasmy level and copy number indicate disease burden in m.3243A>G mitochondrial disease. EMBO Mol Med. 2018;10(6):10. doi: 10.15252/emmm.201708262
51. Ohtake A, Murayama K, Mori M, et al. Diagnosis and molecular basis of mitochondrial respiratory chain disorders: Exome sequencing for disease gene identification. Biochim Biophys Acta Gen Subj. 2014;1840(4):1355–1359. doi: 10.1016/j.bbagen.2014.01.025
52. Pronicka E, Piekutowska-Abramczuk D, Ciara E, et al. New perspective in diagnostics of mitochondrial disorders: Two years' experience with whole-exome sequencing at a national paediatric centre. J Transl Med. 2016;14(1):1–19. doi: 10.1186/s12967-016-0930-9
53. Taylor RW, Pyle A, Griffin H, et al. Use of whole-exome sequencing to determine the genetic basis of multiple mitochondrial respiratory chain complex deficiencies. JAMA. 2014;312(1):68–77. doi: 10.1001/jama.2014.7184
54. Wortmann SB, Koolen DA, Smeitink JA, et al. Whole exome sequencing of suspected mitochondrial patients in clinical practice. J Inherit Metab Dis. 2015;38(3):437–443. doi: 10.1007/s10545-015-9823-y
55. Diroma MA, Calabrese C, Simone D, et al. Extraction and annotation of human mitochondrial genomes from 1000 genomes whole exome sequencing data. BMC Genomics. 2014;15(S3):1–15. doi: 10.1186/1471-2164-15-S3-S2
56. Garret P, Bris C, Procaccio V, et al. Deciphering exome sequencing data: Bringing mitochondrial DNA variants to light. Hum Mutat. 2019;40(12):2430–2443. doi: 10.1002/humu.23885
57. Griffin HR, Pyle A, Blakely EL, et al. Accurate mitochondrial DNA sequencing using off-target reads provides a single test to identify pathogenic point mutations. Genet Med. 2014;16(12):962–971. doi: 10.1038/gim.2014.66
58. Poole OV, Pizzamiglio C, Murphy D, et al. Mitochondrial DNA analysis from exome sequencing data improves the diagnostic yield in neurological diseases. Ann Neurol. 2021;89(6):1240–1247.
59. Raymond FL, Horvath R, Chinnery PF. First-line genomic diagnosis of mitochondrial disorders. Nat Rev Genet. 2018;19(7):399–400. doi: 10.1038/s41576-018-0022-1
60. Macken WL, Vandrovcova J, Hanna MG, et al. Applying genomic and transcriptomic advances to mitochondrial medicine. Nat Rev Neurol. 2021;17(4):215–230. doi: 10.1038/s41582-021-00455-2
61. Schon KR, Horvath R, Wei W, et al. Use of whole genome sequencing to determine genetic basis of suspected mitochondrial disorders: cohort study. BMJ. 2021;375:e066288. doi: 10.1136/bmj-2021-066288
62. Davis RL, Kumar KR, Puttick C, et al. Use of whole-genome sequencing for mitochondrial disease diagnosis. Neurology. 2022;99(7): e730–e742. doi: 10.1212/WNL.0000000000200745
63. Riley LG, Cowley MJ, Gayevskiy V, et al. The diagnostic utility of genome sequencing in a pediatric cohort with suspected mitochondrial disease. Genet Med. 2020;22(7):1254–1261. doi: 10.1038/s41436-020-0793-6
64. Macken WL, Falabella M, McKittrick C, et al. Specialist multidisciplinary input maximises rare disease diagnoses from whole genome sequencing. Nat Commun. 2022;13(1):6324. doi: 10.1038/s41467-022-32908-7
65. ElHefnawi M, Jeon S, Bhak Y, et al. Whole genome sequencing and bioinformatics analysis of two Egyptian genomes. Gene. 2018;668:129–134. doi: 10.1016/j.gene.2018.05.048
66. Uittenbogaard M, Wang H, Zhang VW, et al. The nuclear background influences the penetrance of the near-homoplasmic

m.1630 A>G MELAS variant in a symptomatic proband and asymptomatic mother. Mol Genet Metab. 2019;126(4):429–438. doi: 10.1016/j.ymgme.2019.01.022
67. Hathazi D, Griffin H, Jennings MJ, et al. Metabolic shift underlies recovery in reversible infantile respiratory chain deficiency. Embo J. 2020;44:1–19.
68. Gustafson MA, McCormick EM, Perera L, et al. Mitochondrial single-stranded DNA binding protein novel de novo SSBP1 mutation in a child with single large-scale mtDNA deletion (SLSMD) clinically manifesting as Pearson, Kearns-Sayre, and Leigh syndromes. PLoS One. 2019;14(9):e0221829. doi: 10.1371/journal.pone.0221829
69. Battle SL, Puiu D, Verlouw J, et al. A bioinformatics pipeline for estimating mitochondrial DNA copy number and heteroplasmy levels from whole genome sequencing data. NAR Genom Bioinform. 2022;4(2):lqac034. doi: 10.1093/nargab/lqac034
70. Boggan RM, Ng YS, Franklin IG, et al. Defining the nuclear genetic architecture of a common maternally inherited mitochondrial disorder. medRxiv. 2022.
71. Gupta R, Kanai M, Durham TJ, et al. Nuclear genetic control of mtDNA copy number and heteroplasmy in humans. medRxiv. 2023.
72. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31(3):213–219. doi: 10.1038/nbt.2514
73. Puttick C, Kumar KR, Davis RL, et al. Mity: A highly sensitive mitochondrial variant analysis pipeline for whole genome sequencing data. bioRxiv. 2019;852210.
74. Calabrese C, Simone D, Diroma MA, et al. MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. Bioinformatics. 2014;30(21):3115–3117. doi: 10.1093/bioinformatics/btu483
75. Weissensteiner H, Forer L, Fuchsberger C, et al. MtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. Nucleic Acids Res. 2016;44(W1):W64–9. doi: 10.1093/nar/gkw247
76. Singh LN, Ennis B, Loneragan B, et al. MitoScape: A big-data, machine-learning platform for obtaining mitochondrial DNA from next-generation sequencing data. PLoS Comput Biol. 2021;17(11):e1009594. doi: 10.1371/journal.pcbi.1009594
77. Legati A, Zanetti N, Nasca A, et al. Current and new next-generation sequencing approaches to study mitochondrial DNA. J Mol Diagn. 2021;23(6):732–741. doi: 10.1016/j.jmoldx.2021.03.002
78. Bosworth CM, Grandhi S, Gould MP, et al. Detection and quantification of mitochondrial DNA deletions from next-generation sequence data. BMC Bioinf. 2017;18(S12):0–7. doi: 10.1186/s12859-017-1821-7
79. Hjelm BE, Rollins B, Morgan L, et al. Splice-Break: Exploiting an RNA-seq splice junction algorithm to discover mitochondrial DNA deletion breakpoints and analyses of psychiatric disorders. Nucleic Acids Res. 2019;47(10):e59–e59. doi: 10.1093/nar/gkz164
80. Goudenège D, Bris C, Hoffmann V, et al. eKlipse: a sensitive tool for the detection and quantification of mitochondrial DNA deletions from next-generation sequencing data. Genet Med. 2019;21(6):1407–1416. doi: 10.1038/s41436-018-0350-8
81. Basu S, Xie X, Uhler JP, et al. Accurate mapping of mitochondrial DNA deletions and duplications using deep sequencing. PLoS Genet. 2020;16(12):1–15. doi: 10.1371/journal.pgen.1009242
82. Bris C, Goudenège D, Desquiret-Dumas V, et al. Improved detection of mitochondrial DNA instability in mitochondrial genome maintenance disorders. Genet Med. 2021;23(9):1769–1778. doi: 10.1038/s41436-021-01206-w
83. Frazer J, Notin P, Dias M, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599(7883):91–95. doi: 10.1038/s41586-021-04043-8
84. Rentzsch P, Witten D, Cooper GM, et al. CADD: Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886–D894. doi: 10.1093/nar/gky1016
85. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting splicing from primary sequence with deep learning. Cell InternetAvailable from. 2019;176(3):535–548.e24. doi: http://dx.doi.org/10.1016/j.cell.2018.12.015
86. Castellana S, Fusilli C, Mazzoccoli G, et al. High-confidence assessment of functional impact of human mitochondrial non-synonymous genome variations by APOGEE. PLoS Comput Biol. 2017;13(6):e1005628. doi: 10.1371/journal.pcbi.1005628
87. Martín-Navarro A, Gaudioso-Simón A, Álvarez-Jarreta J, et al. Machine learning classifier for identification of damaging missense mutations exclusive to human mitochondrial DNA-encoded polypeptides. BMC Bioinf. 2017;18(1):158. doi: 10.1186/s12859-017-1562-7
88. Bianco SD, Parca L, Petrizzelli F, et al. APOGEE 2: multi-layer machine-learning model for the interpretable prediction of mitochondrial missense variants. bioRxiv. 2023 2023 Jun 18:545476.Internet: http://biorxiv.org/content/early/2023/06/20/2023.06.18.545476.abstract
89. Niroula A, Vihinen M. PON-mt-tRNA: : a multifactorial probability-based method for classification of mitochondrial tRNA variations. Nucleic Acids Res. 2016;44(5):2020–2027. doi: 10.1093/nar/gkw046
90. Sonney S, Leipzig J, Lott MT, et al. Predicting the pathogenicity of novel variants in mitochondrial tRNA with MitoTIP. PLoS Comput Biol. 2017;13(12):e1005867. doi: 10.1371/journal.pcbi.1005867
91. Bris C, Goudenege D, Desquiret-Dumas V, et al. Bioinformatics tools and databases to assess the pathogenicity of mitochondrial DNA variants in the field of next generation sequencing. Front Genet. 2018;9:632. doi: 10.3389/fgene.2018.00632
92. McCormick EM, Lott MT, Dulik MC, et al. Specifications of the ACMG/AMP standards and guidelines for mitochondrial DNA variant interpretation. Hum Mutat. 2020;41(12):2028–2057. doi: 10.1002/humu.24107
93. Ruiz-Pesini E, Lott MT, Procaccio V, et al. An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic Acids Res. 2007;35(Database):D823–8. doi: 10.1093/nar/gkl927
94. Laricchia KM, Lake NJ, Watts NA, et al. Mitochondrial DNA variation across 56,434 individuals in gnomAD. Genome Res. 2022;32(3):569–582. doi: 10.1101/gr.276013.121
95. Vitsios D, Dhindsa RS, Middleton L, et al. Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. Nat Commun. 2021;12(1):1504. doi: 10.1038/s41467-021-21790-4
96. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–443. doi: 10.1038/s41586-020-2308-7
97. Samocha KE, Robinson EB, Sanders SJ, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014;46(9):944–950. doi: 10.1038/ng.3050
98. Balick DJ, Jordan DM, Sunyaev S, et al. Overcoming constraints on the detection of recessive selection in human genes from population frequency data. Am J Hum Genet. 2022;109(1):33–49. doi: 10.1016/j.ajhg.2021.12.001
99. Havrilla JM, Pedersen BS, Layer RM, et al. A map of constrained coding regions in the human genome. Nat Genet. 2019;51(1):88–95. doi: 10.1038/s41588-018-0294-6
100. Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323(5910):133–138. doi: 10.1126/science.1162986
101. van Dijk EL, Jaszczyszyn Y, Naquin D, et al. The third revolution in sequencing technology. Trends Genet. 2018;34(9):666–681. doi: 10.1016/j.tig.2018.05.008
102. Amarasinghe SL, Su S, Dong X, et al. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21(1):30. doi: 10.1186/s13059-020-1935-5
103. PacBio. Pacific biosciences revio system [Internet]. [cited 2023 Jan 14]. Available from: https://www.pacb.com/revio/.
104. ONT. Oxford Nanopore Q20 plus chemistry [Internet]. [cited 2023 Jan 14]. Available from: https://nanoporetech.com/q20plus-chemistry.
105. Wenger AM, Peluso P, Rowell WJ, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37(10):1155–1162. doi: 10.1038/s41587-019-0217-9

106. Vollger MR, Logsdon GA, Audano PA, et al. Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. Ann Hum Genet. 2020;84(2):125–140. doi: 10.1111/ahg.12364
107. Baid G, Cook DE, Shafin K, et al. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. Nat Biotechnol. 2022. doi:10.1038/s41587-022-01435-7.
108. Shafin K, Pesout T, Chang PC, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. Nat Methods. 2021;18(11):1322–1332. doi: 10.1038/s41592-021-01299-w
109. Van der Verren SE, Van Gerven N, Jonckheere W, et al. A dual-constriction biological nanopore resolves homonucleotide sequences with high fidelity. Nat Biotechnol. 2020;38(12):1415–1420. doi: 10.1038/s41587-020-0570-8
110. Li H, Birol I. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics Internet. 2018;34(18):3094–3100. doi: 10.1093/bioinformatics/bty191
111. Chakraborty S, Rowell W, Gu J, et al. Mitochondrial DNA sequencing using PacBio SMRT technology. Adv Genom Biol Technol. 2018. InternetAvailable from: https://www.pacb.com/wp-content/uploads/Chakraborty-AGBT-2018-Mitochondrial-DNA-Sequencing-Using-the-PacBio-Sequel-System-.pdf
112. Georgieva D, Liu Q, Wang K, et al. Detection of base analogs incorporated during DNA replication by nanopore sequencing. Nucleic Acids Res. 2020;48(15):e88. doi: 10.1093/nar/gkaa517
113. Gilpatrick T, Lee I, Graham JE, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. Nat Biotechnol. 2020;38(4):433–438. doi: 10.1038/s41587-020-0407-5
114. Hafford-Tear NJ, Tsai Y-C, Sadan AN, et al. Crispr/cas9-targeted enrichment and long-read sequencing of the Fuchs endothelial corneal dystrophy–associated TCF4 triplet repeat. Genet Med. 2019;21(9):2092–2102. doi: 10.1038/s41436-019-0453-x
115. Keraite I, Becker P, Canevazzi D, et al. A method for multiplexed full-length single-molecule sequencing of the human mitochondrial genome. Nat Commun. 2022;13(1). doi: 10.1038/s41467-022-33530-3
116. Vandiver AR, Pielstick B, Gilpatrick T, et al. Long read mitochondrial genome sequencing using Cas9-guided adaptor ligation. Mitochondrion. 2022;65:176–183. doi: 10.1016/j.mito.2022.06.003
117. Vandiver AR, Hoang AN, Herbst A, et al. Nanopore sequencing identifies a higher frequency and expanded spectrum of mitochondrial DNA deletion mutations in human aging. Aging Cell Internet. 2023 [cited 2023 May 24]; (6):e13842. doi: 10.1111/acel.13842
118. Frascarelli C, Zanetti N, Nasca A, et al. Nanopore long-read next-generation sequencing for detection of mitochondrial DNA large-scale deletions. Front Genet Internet. 2023;14
119. Rygiel KA, Grady JP, Taylor RW, et al. Triplex real-time PCR–an improved method to detect a wide spectrum of mitochondrial DNA deletions in single cells. Sci Rep. 2015;5(1):1–12. doi: 10.1038/srep09906
120. Sedlazeck FJ, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15(6):461–468. doi: 10.1038/s41592-018-0001-7
121. Uliano-Silva M, Gabriel RN, Ferreira J, et al. MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. bioRxiv. 2022 2022 Dec 23;521667.
122. Wei W, Schon KR, Elgar G, et al. Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. Nature. 2022;611(7934):105–114. doi: 10.1038/s41586-022-05288-7