Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Nanomito enables amplification-free full-length analysis of mitochondrial DNA variants and deletions

Valérie Desquiret-Dumas
Université d'Angers
Solène Duverger
Université d'Angers
Patrizia Amati-Bonneau
Université d'Angers
Xavier Dieu
Université d'Angers
Delphine Mirebeau-Prunier
Université d'Angers
Pascal Reynier
Université d'Angers
Marc Ferré
marc.ferre@univ-angers.fr

Université d'Angers

Article

Keywords: mitochondrial DNA, heteroplasmy, large-scale deletions, amplification-free sequencing, long-read sequencing, nanopore sequencing

Posted Date: July 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10113992/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: Competing interest reported. M.F. previously received non-financial support through the Oxford Nanopore Technologies MinION Access Programme in 2014, including access to an early MinION sequencer and free flow cells through 2016. The authors declare that they have no other competing interests.

Version of Record: A version of this preprint was published at Scientific Reports on October 28th, 2026. See the published version at https://doi.org/10.1038/s41598-026-74050-0.

Abstract

Accurate analysis of mitochondrial DNA (mtDNA) requires reliable quantification of heteroplasmy and robust detection of large-scale deletions. However, routine PCR-based short-read workflows can introduce amplification bias and have limited structural resolution. Here, we developed Nanomito, an amplification-free long-read workflow for full-length sequencing of native mtDNA using Oxford Nanopore Technologies, and compared it with the routine long-range PCR-based Illumina approach in 15 retrospectively selected clinical samples previously characterized in mitochondrial diagnostics. The long-read workflow detected all clinically confirmed causal point variants and identified one low-level m.3243A > G variant, present at 3.8% heteroplasmy by digital PCR, that was detected only among quality-filtered short-read calls. Relative to digital PCR, long-read heteroplasmy estimates showed a more stable quantitative relationship than the PCR-based short-read workflow. In deletion-positive samples, long-read sequencing confirmed the major deletions detected by routine testing with closely matching breakpoint coordinates and resolved a complex deletion spectrum in one sample. These findings support amplification-free full-length nanopore sequencing as a promising framework for integrated mtDNA variant and deletion analysis.

Introduction

Mitochondrial DNA (mtDNA) is a compact 16,569-base-pair double-stranded circular genome that encodes 13 respiratory chain subunits, 22 tRNAs, two rRNAs, and eight mitochondria-derived peptides (MDPs) [1, 2]. This genome has several distinctive features, including maternal inheritance, copy numbers ranging from a few to several hundred thousand per cell, depending on cell type, absence of recombination, and a higher mutation rate than the nuclear genome [3–5]. mtDNA is highly polymorphic and frequently affected by pathogenic mutations, such as point variants or large-scale rearrangements, that cause a wide range of phenotypic manifestations, from isolated symptoms to devastating multisystem disorders, whose severity often depends on the proportion of mtDNA molecules carrying the variant, that is, the heteroplasmy level [6–8]. Accurate detection of mtDNA variants and accurate quantification of heteroplasmy levels, as well as the detection and quantification of large-scale structural variations, are crucial for clinical diagnosis, genetic counseling, and therapeutic decision-making [9, 10]. Current diagnostic approaches still rely predominantly on short-read high-throughput sequencing of libraries generated from overlapping PCR amplicons spanning the mitochondrial genome [11]. Although these techniques have significantly transformed mtDNA molecular diagnosis by enabling high-throughput analysis with deep coverage at lower cost, they have inherent limitations that may restrict certain diagnostic investigations [10, 12].

Limitations of current short-read sequencing approaches

PCR amplification introduces significant biases, such as differential amplification efficiencies across amplified regions and nucleotide misincorporation caused by DNA polymerase errors. Short-read sequencing workflows include several polymerase-driven amplification steps. First, long-range PCR is used to enrich for mtDNA, followed by clonal amplification of DNA fragments to generate sufficient signal for sequencing. The specific process for this clonal stage varies by platform. It may involve flow cell-based solid-phase amplification or microbead-based emulsion PCR, followed by sequencing [13]. Polymerase misincorporations introduced early in any of these amplification steps can be clonally propagated and may inflate apparent low-level heteroplasmy [14]. These biases can distort heteroplasmy detection [15–17]. They can also cause inaccurate quantification of low-level heteroplasmic variants, which may fall below the detection threshold or be misquantified by preferential amplification of certain mtDNA copies [18].

Short-read data are inherently fragmented and require computational reconstruction of the mitochondrial genome. This process can introduce mapping artifacts and reduce accuracy in repetitive or structurally complex regions [19]. The short length of the reads (typically 150–300 base pairs) limits their ability to span large-scale structural variants, complex rearrangements, and repetitive sequences that are increasingly recognized as clinically relevant [20, 21].

Amplification-based workflows alter the native molecular structure of mtDNA, precluding direct detection of DNA methylation patterns, protein-DNA complexes, and other epigenetic modifications that could yield valuable diagnostic or prognostic information [22–25].

The presence of numerous nuclear mitochondrial sequences (NUMTs) in the nuclear genome remains a challenge for short-read mtDNA analysis. These pseudogenes, which originate from mitochondrial genomes and are scattered throughout the nuclear genome, share high sequence similarity with authentic mtDNA [26, 27]. They can lead to false-positive variant calls or inaccurate heteroplasmy quantification [12, 28, 29]. Using mtDNA-specific primers in long-range PCR protocols substantially reduces the likelihood of NUMT co-amplification. However, it may not completely eliminate the risk [30]. In particular, tissue- and age-related somatic NUMT insertions could sometimes be mistakenly amplified and generate short reads that align to the mitochondrial genome, potentially distorting apparent variant allele frequencies [31].

Potential of long-read and amplification-free nanopore sequencing approaches

Third-generation long-read sequencing technologies, particularly the Oxford Nanopore Technologies (ONT) platform, directly address these limitations of short-read approaches. Unlike traditional sequencing methods, nanopore technology can sequence native, intact DNA molecules without prior amplification or polymerization, preserving the original molecular composition and enabling direct detection of base modifications [32, 33]. The ability of nanopore sequencing to generate read lengths exceeding 100 kilobases offers compelling possibilities for exhaustive mtDNA analysis. These long reads can cover the entire mitochondrial genome in a single sequencing run, eliminating assembly artifacts and providing direct evidence of full-length structural variants, large-scale deletions, and complex rearrangements that are challenging to detect using short-read methods [34–36].

Recent ONT chemistry and pore improvements, including R10.4 flow cells and Q20 chemistry, increased modal single-read accuracy to above 99% in benchmarking experiments [37]. Although this raw-read error rate may be higher than the most accurate short-read sequencing techniques, downstream performance in SNP calling and genome-wide analyses can be comparable to short-read approaches when using the most recent ONT pipelines, while

preserving long-read advantages such as phasing and improved structural-variant resolution [38]. These advances, along with sophisticated base-calling algorithms and error-correction strategies, have made nanopore sequencing a reliable platform for clinical diagnostics [36, 39].

Amplification-free strategies represent an additional major advance in the analysis of mitochondrial genomes. Direct sequencing of native mtDNA molecules preserves the individuality of each mtDNA molecule and native heteroplasmy levels while eliminating PCR-induced biases. Additionally, it maintains the molecular integrity necessary for detecting epigenetic modifications [36, 40, 41]. However, implementing amplification-free approaches requires multi-step enrichment techniques to overcome the obstacle of insufficient biological material.

The use of Cas9-mediated targeted enrichment has become a potent tool for the selective extraction and sequencing of mtDNA [34, 36, 42]. By utilizing programmable guide RNAs, this method guides the Cas9 endonuclease to cleave specific target sites within the mitochondrial genome, followed by preferential adapter ligation and sequencing of the targeted molecules. The Cas9 cleavage sites serve a dual function, both as guides for the endonuclease and markers for the subsequent sequencing process. They allow for the specific enrichment of mtDNA, while at the same time providing molecular barcodes for the multiplexed processing of samples.

The methodology described by Keraite *et al.* represents an important advance in this field [36]. It demonstrates the ability to sequence full-length, single-molecule human mtDNA using Cas9-guided enrichment and nanopore sequencing. This method provides precise mtDNA enrichment while maintaining the native molecular structure essential for accurate heteroplasmy detection and structural variant characterization. However, the method remains technically demanding to implement and highly dependent on DNA quality and single-guide RNA design, and can lead to heterogeneous coverage. Because clinical diagnostics require robust and standardized workflows, simpler enzymatic digestion-based enrichment strategies may offer a more practical route to routine implementation.

Clinical applications

From a clinical perspective, improving the interrogation of the mitochondrial genome addresses two complementary needs. First, primary mitochondrial diseases are individually rare but, collectively, constitute one of the most common groups of inherited metabolic disorders. When both nuclear and mitochondrial genetic causes are considered, the estimated adult prevalence is about 1 in 4,300, whereas disorders specifically related to pathogenic mtDNA mutations have a minimum estimated prevalence of about 1 in 5,000 [6, 43]. Therefore, the accurate identification and quantification of these pathogenic mtDNA variants and deletions are crucial for diagnosis and genetic counseling. Second, beyond Mendelian mitochondrial disorders, mtDNA instability, including somatic mutations, large deletions and mtDNA depletion, is increasingly implicated in common conditions such as neurodegenerative disease [44] and aging [45], and can also be triggered iatrogenically by nucleoside analogue antiviral therapies through mitochondrial toxicity mechanisms that include mtDNA depletion [46].

The clinical impact of improved mtDNA analysis extends to many areas in medical genetics and precision medicine. Precise heteroplasmy quantification is crucial for prognostic biomarkers [47], as many mitochondrial disorders exhibit threshold effects where clinical manifestations correlate with the proportion of mutant mtDNA in affected tissues [7, 48]. The ability to detect low-level heteroplasmic variants is particularly important for carrier detection and genetic counseling [49]. These variants can expand in subsequent generations through genetic bottlenecks during oogenesis [50, 51].

Large-scale deletions of mtDNA can cause severe disorders such as Kearns-Sayre syndrome, Pearson syndrome, and progressive external ophthalmoplegia. Historically, it has been difficult to characterize them using short-read methods due to their size and structural complexity [35, 52]. In the last decade, bioinformatics tools were developed to detect these deletions in short-read sequencing data, including MitoSAlt [53], eKLIPse [52], and Splice-Break [54]. However, these tools rely heavily on secondary mapping of soft-clipped reads and remain prone to alignment errors in repeated regions, with heteroplasmy estimates also affected by read-quality filtering. Conversely, long-read sequencing technology provides precise mapping of deletion boundaries and accurate quantification of deletion loads, which are critical for clinical management and genetic counseling [35, 36, 55]. The analysis of multiple mtDNA deletions, whether caused by genetic defects or physiological aging, is further complicated by numerous low-level deleted molecules that may be artifactually amplified during PCR. In clinical practice, such multiple mtDNA deletions are investigated primarily in muscle tissue.

Long-read technologies can provide the comprehensive genomic coverage needed to detect even more complex structural variants that are rarely, if ever, assessed in routine practice [56, 57]. These include duplications, inversions, and translocations. Collectively, these infrequent variants may contribute substantially to the genetic basis of mitochondrial disorders and may help explain previously unresolved clinical presentations [58]. Long-read techniques are now uncovering previously unidentified pathogenic variants and enabling genetic diagnosis in situations where traditional short-read methods are insufficient [59, 60]. They have also demonstrated the feasibility of full-length, single-molecule mtDNA sequencing for sensitive detection of heteroplasmy and complex deletions [35, 36]. These early studies support the potential of amplification-free, long-read mtDNA sequencing for comprehensive mtDNA analysis.

Study objectives

Our primary goal was to develop an analytical and computational framework for amplification-free sequencing of native mtDNA using ONT. Our secondary objective was to compare this workflow with the conventional long-range PCR-based short-read approach and to assess its ability to quantify heteroplasmy and detect large deletions and complex structural variants in a retrospective comparative dataset.

Methods

Clinical samples

This retrospective comparative study evaluated an amplification-free long-read mtDNA workflow against the routine long-range PCR short-read workflow in 15 clinically characterized samples selected from our cohort of patients previously referred for mtDNA analysis in the Department of Biochemistry and Molecular Biology at Angers University Hospital. Samples were selected to represent typical categories encountered in mitochondrial diagnostics, including heteroplasmic SNVs across a range of levels and single or multiple large-scale deletions, but they were not intended to constitute a consecutive or unbiased diagnostic series. All participants provided written informed consent for genetic testing performed to establish the molecular diagnosis of their condition. This study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee (CEIC) of the Angers University Hospital (approval number 2024-031) for reuse of these samples to evaluate the long-read sequencing workflow.

For routine diagnostic purposes, total genomic DNA was extracted from peripheral blood leukocytes using the Nucleomag Blood Kit (Macherey-Nagel) on the MICROLAB STARlet automation platform (Hamilton). The DNA concentration was quantified with the dsDNA Broad Range Assay Kit (Thermo Fisher Scientific) on the Qubit fluorometer (Thermo Fisher Scientific) and stored at -20°C until used. Before inclusion in this study, all samples were analyzed in the standard clinical workflow, which consisted of long-range PCR followed by Illumina short-read next-generation sequencing. Digital PCR (dPCR) was used when necessary to accurately quantify mtDNA variant heteroplasmy levels and served as the reference method for heteroplasmy comparisons.

Digital PCR

Heteroplasmy of pathogenic mtDNA variants was accurately quantified using digital PCR. Specifically, 0.1 ng of total DNA was amplified in the 26,000 partitions of a 24-well dPCR plate (QIAcuity Nanoplate 26k 24-well, Qiagen) using primers flanking the pathogenic mtDNA variant and TaqMan probes specific for either the wild-type or mutant sequence (primer and probe sequences are available upon request). Partitions positive for either wild-type or mutant molecules were quantified using QIAcuity Software Suite 1.2 (Qiagen). The ratio of mutant to total (mutant + wild-type) copies was calculated to determine heteroplasmy levels.

Short-read sequencing and analysis

Library preparation and sequencing

For routine diagnostics, mitochondrial DNA was analyzed using a short-read sequencing workflow following standard clinical protocols. Two rounds of long-range PCR were performed to amplify mitochondrial genomes, generating two overlapping amplicons spanning the entire 16.6 kb genome, as described by Luo *et al*. [61]. Libraries were prepared using the Illumina DNA Prep Kit, following the manufacturer's instructions, and sequenced on a MiSeq instrument (Illumina) with 2 × 250 bp paired-end reads, achieving an average coverage depth of over 1000×.

Read processing and alignment

Illumina FASTQ files, a text-based sequence format that stores nucleotide calls and base-quality scores, underwent adapter trimming and quality filtering with fastp v0.23.2. Subsequently, the processed reads were aligned to the revised Cambridge Reference Sequence (rCRS, NC_012920.1) using BWA-MEM v0.7.17. elPrep v5.1.2 identified and removed duplicate reads, while SAMtools organized and indexed the Binary Alignment/Map (BAM) files.

Variant calling and annotation

In this study, to ensure compatibility of variant call format (VCF) files for comparison with nanopore data, SNVs and small indels were called from Illumina mtDNA data using the recently published mtDNA analysis tool Mitopore v1.0.1 by uploading aligned BAM files to the Mitopore web server [62]. Large-scale mtDNA deletions were then detected and quantified with eKLIPse v2.1 [52].

Long-read sequencing and analysis

Sample preparation and sequencing

Genomic DNA (gDNA) extracted from clinical specimens (see "Clinical samples") underwent PCR-free mtDNA enrichment based on the circular topology of mtDNA using an adapted protocol from Keraite *et al*. [36]. Five micrograms of total DNA were treated with exonuclease V (RecBCD), an ATP-dependent dsDNA exonuclease, to preferentially digest linear DNA, thereby enriching circular mtDNA and reducing nuclear DNA background (New England Biolabs). Samples were treated with 20 U of exonuclease V in the presence of 1 mM ATP and incubated for 30 minutes at 37°C. To prevent mtDNA degradation, the reaction was followed by a 30-minute heat treatment at 70°C. The remaining linear genomic DNA was dephosphorylated using 15 U of a recombinant calf intestine alkaline phosphatase (Quick CIP, New England Biolabs). The DNA was incubated with Quick CIP for 20 minutes at 37°C, followed by a 3-minute inactivation step at 80°C. The enriched circular mtDNA pool was linearized with 40 U of the PvuII restriction enzyme (30-minute digestion at 37°C). This enzyme has a single cleavage site in the human reference mitochondrial genome and produces a ~ 16.6 kb linear molecule suitable for full-length sequencing. After all the enzymatic steps, DNA was purified using a 1.5× bead-based DNA cleanup (AMPure DNA Clean Beads, Beckman Coulter) to eliminate enzymes, salts, and short fragments. It was then quantified fluorometrically on a Qubit instrument (Thermo Fisher Scientific) and used as input for ligation-based nanopore library construction.

The libraries were prepared using the native barcoding workflow for the Native Barcoding Kit V14 (Oxford Nanopore Technologies) following the manufacturer's instructions. They were then sequenced on MinION Mk1B or PromethION 2 Solo instruments with FLO-MIN114 or FLO-PRO114M flow cells (Oxford Nanopore Technologies). Sequencing runs were configured to support multiplexing of clinical samples per flow cell when DNA yield and expected coverage permitted.

Initial basecalling and primary processing

To reduce the computational burden and data transfer requirements, all reads with mitochondrial genome alignment were preprocessed before being uploaded to a high-performance computing (HPC) cluster.

The raw electrical signals, stored in Oxford Nanopore POD5 raw signal files, were basecalled using Dorado v1.3.0 (https://github.com/nanoporetech/dorado). To balance accuracy and computational efficiency, we used the "hac" model (high accuracy), which simultaneously performed barcode demultiplexing based on the supplied sample sheet. Basecalling was performed on a Windows-based workstation with an RTX 5000 Ada 32G graphics processing unit (GPU; NVIDIA Corporation, Santa Clara, CA, USA), with on-the-fly alignment to the complete human reference genome build 38 (GRCh38/hg38) to enable rapid identification of mitochondrial reads. The aligned BAM files were then analyzed using SAMtools v1.17 and pysam, a Python library, to identify reads aligning to the mitochondrial chromosome (chrM, NC_012920.1). A dictionary mapping read IDs to parent IDs was generated from the Dorado BAM tag pi:Z to maintain relationships between split reads generated during basecalling. POD5 files were then processed with pod5-tools v0.3.0, retaining only reads or parent reads associated with mitochondrial alignments.

This preprocessing step generated three outputs: (1) filtered POD5 files containing only mitochondrial reads, (2) a read-to-parent ID mapping file, and (3) a list of mitochondrial parent read identifiers. These files were then uploaded to the Genouest HPC cluster (Rennes, France) for downstream analysis.

High-performance basecalling

After being uploaded to the HPC cluster, the selected mitochondrial POD5 files underwent a second round of basecalling with Dorado v1.3.0 and GPU acceleration. This second basecalling step used the "sup" model (super-high accuracy) to maximize variant-calling accuracy. Basecalling was performed on compute nodes with Tesla V100-PCIE-16GB GPUs (NVIDIA Corporation, Santa Clara, CA, USA), with processing optimized for the available GPU memory.

Demultiplexing of barcoded reads was performed simultaneously with basecalling using Dorado's integrated barcode classification. It assigned each read to its corresponding sample identity based on its barcode sequence. As a result, sample-specific FASTQ files were automatically generated at the basecalling stage, labeled with their demultiplexed sample identifier.

Selection of reads corresponding to enzyme cuts

Reads corresponding to the enzymatic cleavage site introduced during sample preparation were selected in a dedicated demultiplexing step. Cut-site filtering was performed with ont_demult v0.3.4, a demultiplexing tool designed for mitochondrial DNA analysis.[36] The "both" strategy (sequences flanking the cut must be detected at both ends) was used to maximize specificity and eliminate incomplete or potential NUMT reads.

Demultiplexed FASTQ files were organized by patient identifier and aligned to a customized mitochondrial genome reference sequence. The reference consists of two tandem copies of the revised Cambridge Reference Sequence (rCRS, RefSeq: NC_012920.1) with masked regions around the expected enzymatic cut sites. This reference design follows the methodology of Keraite *et al.* [36] and improves alignment of reads spanning the expected mtDNA molecule boundaries. Alignment was performed using minimap2 v2.28-r1209 with the Oxford Nanopore-specific -ax map-ont option.

Variant calling

Variant calling was performed with Baldur v1.2.2. This tool is optimized for long-read mtDNA analysis and has been reported to enable sensitive detection of heteroplasmy below 1% and precise deletion breakpoint mapping [36]. The Baldur parameters were configured to take into account the unique characteristics of mitochondrial genomes (options: --mapq-threshold 20 --qual-threshold 10 --max-qual 30 --max-indel-qual 20 --homopolymer-limit 4 --output-deletions).

Variants were annotated with SnpSift v5.1 (SnpEff suite) [63] using specialized resources for mitochondrial DNA: the gnomAD v3.1 mitochondrial genome dataset [64] and MITOMAP [65]. Annotated VCFs were converted to tab-delimited tables using BCFtools v1.17 [66].

Mitochondrial haplogroup assignment and contamination assessment were performed with HaploCheck v1.3.3 using default parameters [67]. In some long-read samples, HaploCheck reported a "Contamination Status: YES" because minor haplogroup signals were detected. In the present context, these minor components were compatible with low-level heteroplasmy rather than true cross-sample contamination, as supported by concordant major haplogroup assignments and the overall sample-specific variant profiles. Corresponding HaploCheck raw outputs are included in the Supplementary Data files.

Workflow automation and reproducibility

The bioinformatics pipeline, named Nanomito, was implemented as a modular workflow management system using the Simple Linux Utility for Resource Management (SLURM) job scheduler on the Genouest HPC cluster. The workflow comprised the following components:

1. Basecalling and demultiplexing (wf-bchg.sh v2.4.5): This step features GPU-accelerated basecalling and parallel sample demultiplexing.

2. Patient-level demultiplexing (wf-demultmt.sh v2.4.5): Molecular barcode-based sample separation that only retains reads that match the enzymatic cut sites exactly. Only reads starting and ending at the expected cut boundaries were retained. The resulting data undergo SNV/indel calling, MITOMAP, and gnomAD-based annotations for medical and population-level interpretation.

3. Comparative analysis (compare_vcf.sh v2.4.5): Evaluates platform agreement.

Extensive logging and error handling ensured reliable processing of large sample batches. Dependent tasks were automatically scheduled to execute sequentially, while independent ones were parallelized, optimizing efficiency. Generally, the complete dry-lab workflow, including the most time-consuming step of pre-processing, takes around 30 minutes per MinION flow cell and up to three hours for a PromethION flow cell.

Comparative analysis

To assess concordance between the two workflows, BCFtools isec v1.23 was used [66]. For each sample, VCF files were initially filtered to keep only the PASS variants, then compressed and indexed. This analysis generated four output files: 0000.vcf (variants unique to amplification-free long-read data analyzed with Baldur), 0001.vcf (variants unique to short-read data analyzed with Mitopore), 0002.vcf (shared variants in the Baldur format), and 0003.vcf (shared variants in the Mitopore format). An extensive set of quality metrics was compiled, including the total number of variants per platform, the number of shared variants, platform-specific variant counts, haplogroup assignments and concordance, heteroplasmy levels for shared variants, and annotations of disease-associated variants.

A custom script (compare_vcf.sh v2.4.5) was used to automate sample-by-sample comparisons. It generates a comprehensive HTML report that seamlessly integrates: a summary of statistics; haplogroup comparison tables; detailed variant tables with MITOMAP and gnomAD annotations; and highlighting of clinical significance (pathogenic, likely pathogenic, benign). Variants were classified according to MITOMAP annotations as pathogenic (P): confirmed disease-associated variants; likely pathogenic (LP): variants with strong evidence for pathogenicity; benign (B): common polymorphisms with no known clinical significance. Large-scale deletions were annotated with endpoints and lengths for precise characterization.

The processed comparative outputs, including anonymized VCF files, BCFtools isec results, HaploCheck raw results, and HTML summary reports, are provided as Supplementary Data files.

Data anonymization

Variant call files (VCF) were anonymized to protect patient privacy while preserving genomic data integrity. This was done using a custom script (anonymize_vcf.sh) that replaced all personal identifiers with unique, anonymous sample IDs. All variant information, quality metrics, and annotation fields were preserved.

Statistical analysis

All statistical analyses were conducted in R version 4.5.2, using the mcr and BlandAltmanLeh packages. Three datasets of heteroplasmy rates were compared for pathogenic point variants in mtDNA: dPCR, long-read sequencing without amplification using Baldur, and short-read sequencing amplified by long-range PCR using Mitopore. Since each variant was quantified once per method and the sample size was small (n = 10), we assessed agreement with the dPCR reference using paired nonparametric tests and method-comparison analyses rather than correlation coefficients.

To maintain quantitative uniformity, heteroplasmy levels determined by long-read sequencing with Baldur or short-read sequencing with Mitopore were compared with dPCR values using paired Wilcoxon signed-rank tests; paired t-tests were also reported as descriptive parametric summaries. Bland-Altman plots were created to visually depict the average bias and 95% limits of agreement between each sequencing technique relative to dPCR. Deming and Passing-Bablok regressions were fitted to quantify systematic bias. dPCR served as the reference (x) and sequencing heteroplasmy as the test method (y). We calculated 95% confidence intervals for the slope and intercept, comparing them to the identity line (slope = 1, intercept = 0). This allowed assessment of both constant and proportional bias.

One causal variant detected as a PASS call by the Baldur long-read workflow but present only among quality-filtered calls in the Mitopore short-read output was excluded from quantitative testing. It was classified as a false negative in PASS-based qualitative detection metrics. A two-sided p-value below 0.05 was considered statistically significant. Since the method comparison design was exploratory and the sample size was limited, no formal correction for multiple testing was applied.

Results

To evaluate the analytical performance of our amplification-free, single-molecule mtDNA sequencing workflow, we analyzed 15 retrospectively selected clinical samples that had previously been characterized in routine mitochondrial diagnostics at Angers University Hospital. The cohort was designed to represent two major categories of clinically relevant mtDNA abnormalities: heteroplasmic single-nucleotide variants (SNVs) and large-scale deletions, across a broad range of heteroplasmy levels. Ten DNA samples taken from human blood carried well-established pathogenic mtDNA variants (m.3243A > G, m.8344A > G, and m.11778G > A). Another five samples had either a single large-scale deletion or a complex deletion spectrum. All samples had previously undergone analysis using the routine diagnostic workflow, which consisted of long-range PCR followed by Illumina short-read sequencing and digital PCR (dPCR). This design enabled direct cross-platform comparison of detection and quantification, but it was not intended to estimate formal diagnostic sensitivity or specificity.

We sequenced native mtDNA molecules end to end using our amplification-free long-read workflow, which enriches circular mtDNA from total genomic DNA (Fig. 1). The mtDNA was linearized with PvuII, which cuts the mitochondrial genome once at position m.2653, creating uniform full-length templates for sequencing. Using a tandem rCRS reference, complete mtDNA molecules were aligned from the PvuII cut site through the end of the first reference copy and into the second (Fig. 2). This design allows full-length circular molecules to be represented in a linear coordinate system. The mean effective depth was 284.5 complete mtDNA molecules per sample, with a range of 85 to 832, supporting single-molecule heteroplasmy estimation and structural variant screening.

Primary pathogenic SNVs

Across the 10 blood samples carrying established pathogenic mtDNA point variants, both the Baldur variant caller-based amplification-free long-read workflow and the Mitopore variant caller-based long-range PCR short-read workflow identified the expected causal variant in most cases, indicating strong qualitative concordance for the primary finding (Table 1). However, one low-level m.3243A > G variant, present at 3.8% heteroplasmy by dPCR, was detected

directly by the Baldur long-read workflow ("PASS" status in the VCF) but was not called as a PASS variant by the Mitopore short-read pipeline; in the short-read output, it was present only among quality-filtered variant calls.

Table 1

Cross-platform qualitative and quantitative assessment of pathogenic mtDNA single-nucleotide variants (SNVs). For each individual and tissue, results fror PCR, dPCR) are compared with amplification-free long-read sequencing (Baldur) and long-range PCR short-read sequencing (Mitopore). The table reports pla variant counts, the causal variant, heteroplasmy level estimates (%), effective depth (complete mtDNA molecules for long reads; read depth for short re assignment.

Individual ID	Tissue	Variants identified by variant calling analysis			Diagnosis report with heteroplasmy determined by dPCR		Amplification-free long-read sequencing (Baldur)			Short-read sec (Mitopore)
		Baldur-only	Mitopore-only	Common to both methods	Causal variant	Heteroplasmy	Heteroplasmy	Depth (molecules)	Haplogroup	Heteroplasmy
M3243-1	Blood	18	8	38	m.3243A > G	26.1%	34.8%	446	U2e1a1a	64.5%
M3243-2	Blood	1	1	36	m.3243A > G	5.3%	6.4%	142	R / T2a1a	5.4%
M3243-3	Blood	2	3	35	m.3243A > G	52.3%	64.9%	255	K1d	61.0%
M3243-4	Blood	3	1	34	m.3243A > G	43.5%	49.3%	161	U4a3a	48.3%
M3243-5	Blood	4	1	35	m.3243A > G	3.8%	3.4%	90	R / U4b1a1a1	Quality-filtered
M3243-6	Blood	0	0	42	m.3243A > G	68.1%	72.0%	85	K1b1a	69.9%
M8344-1	Blood	9	5	29	m.8344A > G	93.4%	94.7%	221	H2 / U5a1c1a	92.8%
M8344-2	Blood	11	4	28	m.8344A > G	48.2%	54.9%	157	R / U5a1c1a	56.4%
M8344-3	Blood	3	2	12	m.8344A > G	8.5%	8.6%	189	H2 / H7h	12.8%
M11778-1	Blood	5	5	44	m.11778G > A	83.3%	87.5%	248	C4a1a + 195	84.9%

For M3243-5, "Quality-filtered" indicates that the m.3243A > G variant was present in the short-read output but did not meet PASS criteria and was therefore based quantitative comparisons.

Despite this overall qualitative concordance, heteroplasmy level estimates differed quantitatively across methods (Fig. 3). Using dPCR as the orthogonal reference, amplification-free long-read sequencing showed a modest but consistent positive bias across 10 paired samples, with a Wilcoxon pseudo-median difference of 4.03 percentage points (95% CI, 1.18–7.23; $p = 0.00586$) and a mean difference of 4.38 percentage points (95% CI, 1.40–7.35; $p = 0.00881$). Method-comparison analyses supported close proportional agreement with dPCR, with a Deming slope of 1.040 (95% CI, 0.938–1.143) and intercept of 2.63 percentage points (95% CI, -2.80 to 8.06), and a Passing-Bablok slope of 1.045 (95% CI, 0.898–1.204) and intercept of 0.81 percentage points (95% CI, -1.33 to 11.11).

By contrast, the long-range PCR short-read workflow showed greater dispersion relative to dPCR across 9 paired samples. The Wilcoxon pseudo-median difference was 4.40 percentage points (95% CI, 0.85–20.10; $p = 0.01172$), whereas the mean difference was 7.48 percentage points (95% CI, -1.79 to 16.74; $p = 0.09971$). Deming regression yielded a slope of 0.952 (95% CI, 0.593–1.312) and intercept of 9.74 percentage points (95% CI, -10.33 to 29.81), while Passing-Bablok regression yielded a slope of 0.950 (95% CI, 0.563–1.183) and intercept of 5.75 percentage points (95% CI, -0.62 to 31.54), indicating a less stable quantitative relationship with dPCR than that observed for amplification-free long-read sequencing.

Bland-Altman analysis showed that the amplification-free long-read workflow and dPCR had closer agreement at high heteroplasmy levels. The PCR-amplified short-read workflow had larger discrepancies at low heteroplasmy levels, where interpretation is often most challenging (Fig. 3).

Global SNV profiles

When the full set of mtDNA SNVs was considered, amplification-free long-read sequencing with Baldur and short-read sequencing with Mitopore yielded markedly different variant spectra (see Supplementary Data files). The long-read workflow identified more private variants overall, with up to 18 long-read-only variants in a single sample, whereas the short-read workflow identified fewer private variants, with a maximum of 8. Many long-read-only variants were low- to moderate-level polymorphisms distributed across the mitochondrial genome that passed Baldur strand-balance filters but fell below Mitopore calling thresholds or into regions flagged as artifact-prone in the short-read workflow. This pattern is consistent with the detection of low-level heteroplasmic variants that fall near or below the typical reporting threshold of PCR-based short-read mtDNA sequencing, although orthogonal validation was not systematically performed for all low-level private variants in the present dataset. By contrast, short-read-only variants were less frequent and were enriched at positions known to be problematic in short-read mtDNA analysis, particularly near homopolymeric contexts. Many were flagged as artifact-prone sites or

panel-of-normals positions in the short-read reports and were not supported by the long-read data. Some apparent discrepancies also reflected differences in how equivalent variants were normalized across VCF files, which can generate false positives in BCFtools isec comparisons.

Full-length sequencing yielded more detailed haplotype data. Each complete mtDNA molecule allowed direct phasing of the causal variant with background polymorphisms. Haplogroup inference was consistent across all 10 blood samples (Table 1). In several cases, the long-read output additionally reported the corresponding upstream macroclade together with the refined lineage (*e.g.* "R/T2a1a" or "H2/U5a1c1a"), whereas the short-read output only reported the refined haplogroup (*e.g.* "T2a1a" or "U5a1c1a"), indicating hierarchical labeling rather than a true discrepancy. This concordance supports sample identity consistency, while underscoring the added value of full-length nanopore reads for haplotype-resolved interpretation.

Large-scale deletions

In the clinical samples with large-scale mtDNA deletions (Table 2), amplification-free long-read sequencing with the Baldur variant caller identified the four single deletions previously detected by routine diagnostics in DEL-1 to DEL-4. Additionally, it revealed a more complex spectrum of low-level multiple deletions in DEL-5.

Table 2

Large-scale mtDNA deletions in deletion-benchmarking samples (DEL-1 to DEL-5) detected by long-range PCR short-read sequencing and analyzed with eKLIPse, versus amplification-free long-read sequencing and analyzed with Baldur. For each sample, the table reports deletion coordinates, deletion length (bp), heteroplasmy level (%), and coverage for each technology.

Individual ID	Tissue	Short-read sequencing (eKLIPse)				Amplification-free long-read sequencing (Baldur)			
		Variant	Length	Heteroplasmy	Coverage (reads)	Variant	Length	Heteroplasmy	Coverage (molecules)
DEL-1	Blood	m.6331_13517del	7,185 bp	72.3%	3,747	m.6332_13516del	7,185 bp	90.1%	570
DEL-2	Blood	m.8472_13447del	4,974 bp	67.5%	4,612	m.8470_13446del	4,977 bp	92.2%	233
DEL-3	Blood	m.8249_16033del	7,783 bp	84.7%	5,519	m.8250_16032del	7,783 bp	79.9%	832
DEL-4	Blood	m.6919_13554del	6,634 bp	69.8%	3,708	m.6920_13553del	6,634 bp	49.5%	100
DEL-5	Muscle	m.5787_13923del	8,135 bp	22.3%	1,348	m.5773_12770del	6,997 bp	0.7%	538
		m.6331_13994del	7,662 bp	18.9%		m.5788_13922del	8,134 bp	0.6%	
						m.6067_11883del	5,816 bp	0.6%	
						m.7499_14424del	6,925 bp	0.6%	
						m.7502_14427del	6,925 bp	0.6%	
						m.9922_16070del	6,148 bp	0.6%	
						m.5787_13922del	8,135 bp	0.4%	
						m.5916_12130del	6,214 bp	0.4%	
						m.6430_13821del	7,391 bp	0.4%	
						m.7095_14814del	7,719 bp	0.4%	
						m.7098_13922del	6,824 bp	0.4%	
						m.7862_13766del	5,904 bp	0.4%	
						m.8279_15759del	7,480 bp	0.4%	
						m.8281_14751del	6,470 bp	0.4%	
						m.8581_15730del	7,149 bp	0.4%	
						m.9506_16070del	6,564 bp	0.4%	
						m.11762_14837del	3,075 bp	0.4%	
						m.13189_15693del	2,504 bp	0.4%	
						48 additional large deletions	64–8,407 bp	0.2%	

Coverage is reported as average read depth for short-read sequencing and as the number of complete mtDNA molecules for long-read sequencing.

For DEL-5, the table lists the most frequent deletions and summarizes the remaining low-frequency deletions.

Short-read sequencing, coupled with our in-house eKLIPse workflow [52], identified one major large-scale deletion in each of DEL-1 to DEL-4, with heteroplasmy estimates of 72.3%, 67.5%, 84.7%, and 69.8%, respectively. Amplification-free long-read sequencing with the Baldur workflow [36] confirmed the corresponding major deletions in DEL-1 to DEL-4, with heteroplasmy estimates of 90.1%, 92.2%, 79.9%, and 49.5%, respectively. Representative Integrative Genomics Viewer (IGV) examples are shown for DEL-3 and DEL-4 (Figs. 4 and 5), whereas corresponding IGV views for DEL-1 and DEL-2 are provided in Supplementary Fig. S1 and Supplementary Fig. S2. Comparison of deletion coordinates showed very close agreement between short-read and amplification-free long-read sequencing. Breakpoint differences ranged from 1 to 2 bp at the 5' breakpoint and were uniformly 1 bp at the 3' breakpoint, with mean absolute differences of 1.25 bp and 1.0 bp, respectively. These small discrepancies are consistent with minor annotation differences, rather than true biological discordance.

No consistent bias direction was detected when comparing deletion load estimates (fraction of mtDNA molecules carrying deletions) obtained by amplification-free long-read and long-range PCR short-read sequencing. Instead, the inter-method differences were found to fluctuate among samples, with a mean absolute difference of 16.9 percentage points and signed differences ranging from − 20.3 to + 24.7 percentage points.

In DEL-5, eKLIPse analysis of short-read data suggested an unusual pattern dominated by two nearby deletions, m.5787_13923del at 22.3% and m.6331_13994del at 18.9% heteroplasmy. In contrast, the Baldur long-read workflow identified 66 distinct deletions in DEL-5, each present at 0.2% to 0.7% heteroplasmy, which translates to one to four out of 538 complete molecules (Fig. 6). DEL-5 was a recently obtained muscle sample with a high mtDNA yield, and 91 of 538 complete molecules (16.9%) carried deletions ranging from 64 to 8,407 bp.

The short-read VCF files lacked large-scale deletions because their detection relied on our in-house eKLIPse workflow rather than on the primary Mitopore variant-calling pipeline. In contrast, the Baldur caller integrated the most reliable long-read deletion calls directly into its analysis workflow. In our dataset, the heterogeneous deletions observed in DEL-5 were not included in the primary VCF output, although they were clearly supported by the raw nanopore reads and documented in the Baldur deletion report.

Discussion

Our results support amplification-free, full-length nanopore sequencing as a promising approach for comprehensive mtDNA analysis because it consolidates SNV detection and heteroplasmy assessment together with large-deletion characterization into a single assay performed on native mtDNA molecules. This approach eliminates the need for long-range PCR and its associated failure modes and quantification biases, which can negatively impact conventional short-read workflows. In our cohort, the amplification-free full-length workflow detected all clinically confirmed causal variants, including a low-level m.3243A > G variant that was missed by the routine short-read long-range PCR pipeline but confirmed by dPCR, highlighting a practical limitation of the short-read mtDNA testing strategy at low heteroplasmy. In deletion-positive samples, full-length non-amplified molecules also revealed subclonal and highly heterogeneous deletion spectra, exemplified by DEL-5, that were not accurately resolved by the eKLIPse-based short-read workflow. Taken together, these findings support the value of analyzing complete native molecules for both quantitative heteroplasmy assessment and characterization of structurally complex mtDNA alterations.

The SNV results are notable not only for qualitative detection, but also for how each method performed quantitatively relative to an orthogonal reference. In this comparative setting, this distinction matters because low-level heteroplasmy is often where technical bias has the greatest impact on interpretation. Despite the overall qualitative concordance for the primary finding, heteroplasmy estimates differed quantitatively across methods. Using dPCR as the orthogonal reference, amplification-free long-read sequencing showed a more stable quantitative relationship than the long-range PCR short-read workflow, which showed greater dispersion relative to dPCR and larger discrepancies at low heteroplasmy levels, where interpretation is often most challenging. This pattern is illustrated by the low-level m.3243A > G case, which was detected and quantified by amplification-free long-read sequencing but was not called as a PASS variant by the Mitopore short-read pipeline; in the short-read output, it was present only among quality-filtered variant calls. PCR-based short-read mtDNA sequencing is often interpreted with an approximate heteroplasmy detection threshold around 5%, although the exact analytical limit depends on assay design, read depth, filtering strategy, and sequence context [10, 68, 69]. While this single observation is consistent with a potential advantage of the long-read workflow near the lower end of the detectable range, it does not by itself establish a formal limit of detection. Overall, the SNV results therefore suggest that Baldur analysis of complete, amplification-free long-read molecules may provide more stable quantitative agreement with dPCR than PCR-based short-read sequencing, particularly in the low-heteroplasmy range where amplification and filtering effects are most likely to affect interpretation.

The deletion data raise a different set of interpretive issues. Although the two approaches showed close agreement for the major deletions identified in DEL-1 to DEL-4, the comparison also highlighted the difficulty of assigning biologically meaningful deletion loads when enrichment strategy, molecule length, and calling rules differ between platforms. Comparison of deletion coordinates showed very close agreement between short-read and amplification-free long-read sequencing, and the small breakpoint differences observed were consistent with minor annotation differences rather than true biological discordance. No consistent bias direction was detected when comparing deletion load estimates obtained by amplification-free long-read and long-range PCR short-read sequencing. Instead, the inter-method differences fluctuated among samples, indicating that deletion load estimates remain method-dependent and are likely influenced by both molecular and analytical factors. In contrast to point variants, for which dPCR can serve as an orthogonal quantitative reference, no broadly accepted gold-standard assay is currently available for heteroplasmy quantification of large-scale mtDNA deletions.

Nevertheless, previous work suggests that long-read nanopore sequencing can provide a more comprehensive representation of mtDNA deletions and their heteroplasmy than PCR-dependent approaches [34, 36]. This includes the Cas9-guided full-length mtDNA approach described by Keraite *et al.*, which showed that PCR-based enrichment methods preferentially amplify shorter molecules carrying large mtDNA deletions, thereby distorting deletion load estimates relative to amplification-free approaches [36]. We did not observe such overestimation in the present comparison between eKLIPse analysis of short-read

data and Baldur analysis of long-read data. This suggests that various factors, such as primer design, the Taq polymerase, and the PCR program used, may affect the phenomenon of preferential amplification.

DEL-5 was particularly informative because it moved the comparison beyond single dominant deletions and into the setting of structurally heterogeneous mtDNA populations. In that context, the advantage of full-length single-molecule analysis was less about confirming one canonical event than about revealing the underlying complexity of the deletion landscape. In DEL-5, eKLIPse analysis of short-read data suggested an unusual pattern dominated by two nearby deletions, whereas the Baldur long-read workflow identified 66 distinct deletions, each present at 0.2% to 0.7% heteroplasmy, corresponding to one to four out of 538 complete molecules. This pattern is consistent with a highly heterogeneous mixture of low-frequency rearrangements that short-read sequencing is poorly suited to resolve. However, the biological and clinical significance of these low-frequency deletion species remains uncertain at this stage. This extreme structural diversity explains why none of the individual deletions reached the stringent “both forward and reverse” support threshold required to be reported as a discrete variant in the VCF, even though the cumulative deletion burden was clearly documented in the Baldur deletion report.

The short-read VCF files lacked large-scale deletions because their detection relied on our in-house eKLIPse workflow rather than on the primary Mitopore variant-calling pipeline. In contrast, the Baldur caller integrated the most reliable long-read deletion calls directly into its analysis workflow. In our dataset, the heterogeneous deletions observed in DEL-5 were not included in the primary VCF output, although they were clearly supported by the raw nanopore reads and documented in the Baldur deletion report. The Baldur caller requires each structural variant to be supported by at least two independent molecules with breakpoints observed in both orientations before inclusion in the VCF, a criterion comparable to the forward/reverse read-support filters used in other short-read sequencing workflows to minimize strand-specific artifacts. This highlights the importance of the raw deletion list generated with the -D/--output-deletions option for interpreting mtDNA structural variation. For this reason, the raw deletion list was included in the final HTML report of the Nanomito pipeline, together with the VCF exported as a table, haplogroup assignment, and additional metrics generated by the wf-finalize.sh workflow.

From an implementation standpoint, the PvuII single-cut strategy appears more practical for high-throughput laboratory workflows than Cas9 guide-based enrichment. It provides a fixed and reproducible linearization point while avoiding the guide design, ribonucleoprotein assembly, and Cas9-dependent multiplexing constraints described for the Cas9 approach [10, 36, 70, 71]. In practice, Cas9-based enrichment is more difficult to standardize, more dependent on DNA quality, and more prone to coverage heterogeneity. In contrast, an enzymatic digestion workflow can be integrated as a simple and robust step in routine laboratory processing. To ensure consistent recovery of mtDNA molecules, we chose a single linearization strategy based on PvuII cleavage at position m.2653. Based on our diagnostic experience, we have not observed mtDNA deletions affecting this position, which reduces the risk that a pathogenic deletion could remove the cut site and escape capture in a single-cut design. However, this choice does not limit the analytical potential. The downstream workflow can handle multiple independent cleavage sites in the same sample, similar to the cut-site-specific demultiplexing and remapping strategy described by Keraite *et al.* [36]. In this case, distinct cleavage points would generate separate FASTQ files that can be processed jointly and merged for variant calling using the complete set of molecules.

Analytical performance is also strongly influenced by coverage. In this study, we achieved an average effective sequencing depth of 284.5 fully sequenced mtDNA molecules per sample, with a range of 85 to 832, from retrospective specimens. In this cohort, this level of coverage was sufficient to support both heteroplasmy quantification and structural variant screening. Importantly, our amplification-free workflow effectively measures the number of complete, independent full-length mtDNA molecules rather than short reads derived from clonal amplification. As a result, a few hundred complete native molecules may provide substantial analytical value because they retain molecule-level heterogeneity and structural information that are not directly accessible from conventional short-read depth alone. The lower depths in some samples were therefore interpreted in the context of independent molecular sampling and likely reflected intrinsic tissue mtDNA content together with suboptimal mtDNA recovery from retrospective DNA extracts. Collectively, these features indicate that a PvuII-linearized, amplification-free long-read workflow can achieve analytical performance at least comparable to short-read testing in this retrospective cohort while offering practical advantages over guide-based Cas9 enrichment.

In addition to detecting sequence variants, amplification-free nanopore sequencing preserves native DNA molecules and raw electrical signal information that can be used for modified-base calling [72, 73]. This allows the direct integration of mitochondrial epigenetic analyses, such as mtDNA methylation profiling, into the same sequencing dataset [24, 25]. Although methylation analysis was not part of the current diagnostic study, it has the potential to expand the analytical scope of the workflow without the need for additional wet-lab assays.

This study has several limitations. The present evaluation was retrospective and based on a limited number of clinically characterized samples that were selected to illustrate specific categories of mtDNA abnormalities rather than to represent a consecutive diagnostic series. We did not perform a formal evaluation of analytical sensitivity, specificity, repeatability, or reproducibility. In addition, deletion load estimates remained method-dependent, and no broadly accepted gold-standard assay is currently available for quantitative validation of large-scale mtDNA deletions. Prospective studies using freshly extracted samples and larger cohorts will be needed to define performance limits more precisely and to further evaluate robustness across routine laboratory conditions.

Conclusions

This study demonstrates the feasibility of amplification-free, full-length nanopore sequencing of native mtDNA for comprehensive mtDNA analysis. In our clinically characterized retrospective cohort, a single assay performed on the same set of complete native mtDNA molecules detected all primary pathogenic point variants, identified a low-level m.3243A > G variant missed as a PASS call by the short-read long-range PCR pipeline, and characterized large-scale deletions, including complex low-frequency deletion spectra not resolved by short reads. Together, these findings support amplification-free nanopore sequencing as a promising methodological framework for concurrent SNV heteroplasmy assessment and structural variant analysis from complete native

molecules. Further validation in larger and prospective cohorts will be required before establishing formal performance metrics for routine clinical implementation.

Declarations

Acknowledgements

We acknowledge the GenOuest bioinformatics core facility (https://www.genouest.org/) for providing the computing infrastructure. We thank M. Rahard and C. Fey-Billon for their technical assistance, and L. Legoff and A. M'Bodj for bioinformatics analysis of the short-read sequencing data at the Plateau de Biologie et Médecine moléculaires (PBMM), Angers.

Author contributions

M.F. initiated the project. V.D.D. and M.F. designed and supervised the project. P.R. contributed to the study design and supervision. V.D.D., P.A.B., D.M.P., X.D., and P.R. contributed to patient recruitment and clinical sample collection. V.D.D. and S.D. performed the biological analyses. M.F. performed the bioinformatics analyses. V.D.D., P.R., and M.F. drafted the manuscript, and all authors critically revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was funded by the Agence de la Biomédecine, Paris, France, and the Fondation de l'Avenir, Paris, France (grant AP-RM-25-016).

Data availability

The sequencing datasets generated during the current study are available in the European Nucleotide Archive (ENA) under accession number PRJEB115744. All processed data generated or analyzed during this study are included in this published article and its supplementary information files. Analysis scripts, configuration files, and instructions are openly available in the Nanomito GitHub repository (https://github.com/marc-ferre/nanomito).

Competing interests

M.F. previously received non-financial support through the Oxford Nanopore Technologies MinION Access Programme in 2014, including access to an early MinION sequencer and free flow cells through 2016. The authors declare that they have no other competing interests.

References

1. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–65 (1981).
2. Miller, B., Kim, S.-J., Kumagai, H., Yen, K. & Cohen, P. Mitochondria-derived peptides in aging and healthspan. *J. Clin. Invest.* **132**, e158449 (2022).
3. Sallevelt, S. C. E. H. *et al.* De novo mtDNA point mutations are common and have a low recurrence risk. *J. Med. Genet.* **54**, 73–83 (2017).
4. Pereira, C. V. & Moraes, C. T. Current strategies towards therapeutic manipulation of mtDNA heteroplasmy. *Front. Biosci.* **22**, 991–1010 (2017).
5. Giles, R. E., Blanc, H., Cann, H. M. & Wallace, D. C. Maternal inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci.* **77**, 6715–6719 (1980).
6. Gorman, G. S. *et al.* Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. *Ann. Neurol.* **77**, 753–759 (2015).
7. Rossignol, R. *et al.* Mitochondrial threshold effects. *Biochem. J.* **370**, 751–762 (2003).
8. Craven, L., Alston, C. L., Taylor, R. W. & Turnbull, D. M. Recent advances in mitochondrial disease. *Annu. Rev. Genomics Hum. Genet.* **18**, 257–275 (2017).
9. Cui, H. *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.* **15**, 388–394 (2013).
10. Slapnik, B. *et al.* The quality and detection limits of mitochondrial heteroplasmy by long read nanopore sequencing. *Sci. Rep.* **14**, 26778 (2024).
11. Palculict, M. E., Zhang, V. W., Wong, L.-J. & Wang, J. Comprehensive mitochondrial genome analysis by massively parallel sequencing. in *Mitochondrial DNA: methods and protocols* (ed. McKenzie, M.) 3–17 (Springer, New York, NY, 2016).
12. Maude, H. *et al.* NUMT confounding biases mitochondrial heteroplasmy calls in favor of the reference allele. *Front. Cell Dev. Biol.* **7**, 201 (2019).
13. Buermans, H. P. J. & den Dunnen, J. T. Next generation sequencing technology: advances and applications. *Biochim. Biophys. Acta (BBA) - Mol. Basis Dis.* **1842**, 1932–1941 (2014).
14. Potapov, V. & Ong, J. L. Examining sources of error in PCR by single-molecule sequencing. *PLOS One* **12**, e0169774 (2017).
15. Santibanez-Koref, M. *et al.* Assessing mitochondrial heteroplasmy using next generation sequencing: a note of caution. *Mitochondrion* **46**, 302–306 (2019).
16. Duan, M., Tu, J. & Lu, Z. Recent advances in detecting mitochondrial DNA heteroplasmic variations. *Molecules* **23**, 323–341 (2018).
17. He, Y. *et al.* Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature* **464**, 610–614 (2010).
18. Zhou, K. *et al.* A novel next-generation sequencing–based approach for concurrent detection of mitochondrial DNA copy number and mutation. *J. Mol. Diagn.* **22**, 1408–1418 (2020).
19. Bosworth, C. M., Grandhi, S., Gould, M. P. & LaFramboise, T. Detection and quantification of mitochondrial DNA deletions from next-generation sequence data. *BMC Bioinf.* **18**, 407–415 (2017).
20. Alkanaq, A. N. *et al.* Comparison of mitochondrial DNA variants detection using short- and long-read sequencing. *J. Hum. Genet.* **64**, 1107–1116 (2019).

21. Macken, W. L., Vandrovcova, J., Hanna, M. G. & Pitceathly, R. D. S. Applying genomic and transcriptomic advances to mitochondrial medicine. *Nat. Rev. Neurol.* **17**, 215–230 (2021).
22. Stoccoro, A. & Coppedè, F. Mitochondrial DNA methylation and human diseases. *Int. J. Mol. Sci.* **22**, 4594 (2021).
23. Nair, A. P. *et al.* Epigenetic regulation of the nuclear genome associated with mitochondrial dysfunction in leber's hereditary optic neuropathy (LHON). *Hum. Genome Var.* **11**, 6 (2024).
24. Hao, Z. *et al.* N6-deoxyadenosine methylation in mammalian mitochondrial DNA. *Mol. Cell* **78**, 382-395.e8 (2020).
25. Koh, C. W. Q. *et al.* Single-nucleotide-resolution sequencing of human *N* 6-methyldeoxyadenosine reveals strand-asymmetric clusters associated with SSBP1 on the mitochondrial genome. *Nucleic Acids Res.* **46**, 11659–11670 (2018).
26. Woischnik, M. & Moraes, C. T. Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome Res.* **12**, 885–893 (2002).
27. Tourmen, Y. *et al.* Structure and chromosomal distribution of human mitochondrial pseudogenes. *Genomics* **80**, 71–77 (2002).
28. Li, M., Schroeder, R., Ko, A. & Stoneking, M. Fidelity of capture-enrichment for mtDNA genome sequencing: influence of NUMTs. *Nucleic Acids Res.* **40**, e137–e137 (2012).
29. Albayrak, L. *et al.* The ability of human nuclear DNA to cause false positive low-abundance heteroplasmy calls varies across the mitochondrial genome. *BMC Genomics* **17**, 1017 (2016).
30. Wei, W. *et al.* Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. *Nature* **611**, 105–114 (2022).
31. Zhou, W. *et al.* Somatic nuclear mitochondrial DNA insertions are prevalent in the human brain and accumulate over time in fibroblasts. *PLOS Biol.* **22**, e3002723 (2024).
32. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The oxford nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239–250 (2016).
33. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nat. Biotechnol.* **34**, 518–524 (2016).
34. Vandiver, A. R. *et al.* Long read mitochondrial genome sequencing using Cas9-guided adaptor ligation. *Mitochondrion* **65**, 176–183 (2022).
35. Vandiver, A. R. *et al.* Nanopore sequencing identifies a higher frequency and expanded spectrum of mitochondrial DNA deletion mutations in human aging. *Aging Cell* **22**, e13842 (2023).
36. Keraite, I. *et al.* A method for multiplexed full-length single-molecule sequencing of the human mitochondrial genome. *Nat. Commun.* **13**, 5902 (2022).
37. Ni, Y., Liu, X., Simeneh, Z. M., Yang, M. & Li, R. Benchmarking of nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. *Comput. Struct. Biotechnol. J.* **21**, 2352–2364 (2023).
38. Kolmogorov, M. *et al.* Scalable nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nat. Methods* **20**, 1483–1492 (2023).
39. Smits, D. J. *et al.* Nanopore long-read sequencing for the critically ill facilitates ultrarapid diagnostics and urgent clinical decision making. *Eur. J. Hum. Genet.* **34**, 108–118 (2026).
40. Walsh, D. J. *et al.* Mito-SiPE is a sequence-independent and PCR-free mtDNA enrichment method for accurate ultra-deep mitochondrial sequencing. *Commun. Biol.* **5**, 1269 (2022).
41. Xu, L. & Seki, M. Recent advances in the detection of base modifications using the nanopore sequencer. *J. Hum. Genet.* **65**, 25–33 (2020).
42. Gilpatrick, T. *et al.* Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.* **38**, 433–438 (2020).
43. Ng, Y. S. & Turnbull, D. M. Mitochondrial disease: genetics and management. *J. Neurol.* **263**, 179–191 (2016).
44. Pinto, M. & Moraes, C. T. Mitochondrial genome changes and neurodegenerative diseases. *Biochim. Biophys. Acta (BBA) - Mol. Basis Dis.* **1842**, 1198–1207 (2014).
45. Bartman, S., Coppotelli, G. & Ross, J. M. Mitochondrial dysfunction: a key player in brain aging and diseases. *Curr. Issues Mol. Biol.* **46**, 1987–2026 (2024).
46. Lewis, W., Copeland, W. C. & Day, B. J. Mitochondrial DNA depletion, oxidative stress, and mutation: mechanisms of dysfunction from nucleoside reverse transcriptase inhibitors. *Lab. Invest.* **81**, 777–790 (2001).
47. Grady, J. P. *et al.* mtDNA heteroplasmy level and copy number indicate disease burden in m.3243A>G mitochondrial disease. *EMBO Mol. Med.* **10**, EMMM201708262 (2018).
48. Smith, K. K. *et al.* A systematic review on the biochemical threshold of mitochondrial genetic variants. *Genome Res.* **34**, 341–365 (2024).
49. Fazzini, F. *et al.* Analyzing low-level mtDNA heteroplasmy—pitfalls and challenges from bench to benchmarking. *Int. J. Mol. Sci.* **22**, 935 (2021).
50. Cree, L. M., Samuels, D. C. & Chinnery, P. F. The inheritance of pathogenic mitochondrial DNA mutations. *Biochim. Biophys. Acta (BBA) - Mol. Basis Dis.* **1792**, 1097–1102 (2009).
51. Zaidi, A. A. *et al.* Bottleneck and selection in the germline and maternal age influence transmission of mitochondrial DNA in human pedigrees. *Proc. Natl. Acad. Sci.* **116**, 25172–25178 (2019).
52. Goudenège, D. *et al.* eKLIPse: a sensitive tool for the detection and quantification of mitochondrial DNA deletions from next-generation sequencing data. *Genet. Med.* **21**, 1407–1416 (2019).
53. Basu, S. *et al.* Accurate mapping of mitochondrial DNA deletions and duplications using deep sequencing. *PLOS Genet.* **16**, e1009242 (2020).
54. Hjelm, B. E. *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.* **47**, e59 (2019).

55. Frascarelli, C. *et al.* Nanopore long-read next-generation sequencing for detection of mitochondrial DNA large-scale deletions. *Front. Genet.* **14**, 1089956 (2023).
56. Sinha, S. *et al.* Long read sequencing enhances pathogenic and novel variation discovery in patients with rare diseases. *Nat. Commun.* **16**, 2500 (2025).
57. Moustakli, E. *et al.* Long-read sequencing and structural variant detection: unlocking the hidden genome in rare genetic disorders. *Diagnostics* **15**, 1803 (2025).
58. Miller, D. E. *et al.* Targeted long-read sequencing identifies missing disease-causing variation. *Am. J. Hum. Genet.* **108**, 1436–1449 (2021).
59. Hiatt, S. M. *et al.* Long-read genome sequencing and variant reanalysis increase diagnostic yield in neurodevelopmental disorders. *Genome Res.* **34**, 1747–1762 (2024).
60. Del Gobbo, G. F. & Boycott, K. M. The additional diagnostic yield of long-read sequencing in undiagnosed rare diseases. *Genome Res.* **35**, 559–571 (2025).
61. Luo, S. *et al.* Biparental inheritance of mitochondrial DNA in humans. *Proc. Natl. Acad. Sci.* **115**, 13039–13044 (2018).
62. Dobner, J. *et al.* mtDNA analysis using mitopore. *Mol. Ther.–Methods Clin. Dev.* **32**, 101231 (2024).
63. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w^{1118}; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
64. Laricchia, K. M. *et al.* Mitochondrial DNA variation across 56,434 individuals in gnomAD. *Genome Res.* **32**, 569–582 (2022).
65. Ruiz-Pesini, E. *et al.* An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res* **35**, D823-8 (2007).
66. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
67. Weissensteiner, H. *et al.* Contamination detection in sequencing studies using the mitochondrial phylogeny. *Genome Res.* **31**, 309–316 (2021).
68. González, M. del M., Ramos, A., Aluja, M. P. & Santos, C. Sensitivity of mitochondrial DNA heteroplasmy detection using next generation sequencing. *Mitochondrion* **50**, 88–93 (2020).
69. Liu, Y. *et al.* Optimized PCR-based enrichment improves coverage uniformity and mutation detection in mitochondrial DNA next-generation sequencing. *J. Mol. Diagn.* **22**, 503–512 (2020).
70. Yeow, D. *et al.* Long-read sequencing for diagnosis of genetic myopathies. *BMJ Neurol. Open* **7**, e000990 (2025).
71. Song, H. *et al.* A long-amplicon nanopore sequencing and analysis method for human whole mitochondrial genome. *Forensic Sci. Int.* **377**, 112668 (2025).
72. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
73. Liu, Y. *et al.* DNA methylation-calling tools for oxford nanopore sequenc

Figures

Figure 1

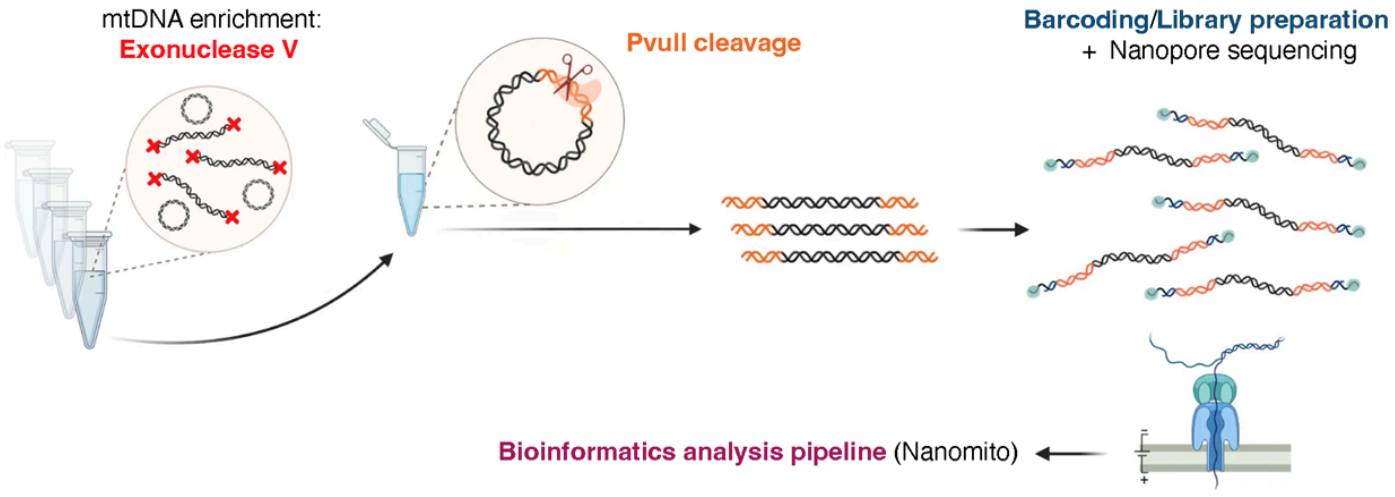


Figure 1

Overall workflow for amplification-free full-length sequencing of human mtDNA using Oxford Nanopore Technologies (ONT). Total genomic DNA is first enriched for circular mtDNA by Exonuclease V (RecBCD) digestion to remove linear nuclear DNA. Intact mtDNA molecules are then linearized at the single

PvuII cut site (m.2653, indicated in orange) to generate uniform ~16.6 kb templates for end-to-end sequencing. Sequencing libraries are prepared using ONT native barcoding and ligation chemistry. Libraries are then sequenced on MinION or PromethION instruments to generate full-length, single-molecule mtDNA reads for downstream analysis. Workflow schematic inspired by Keraite *et al.* [36].

Figure 2

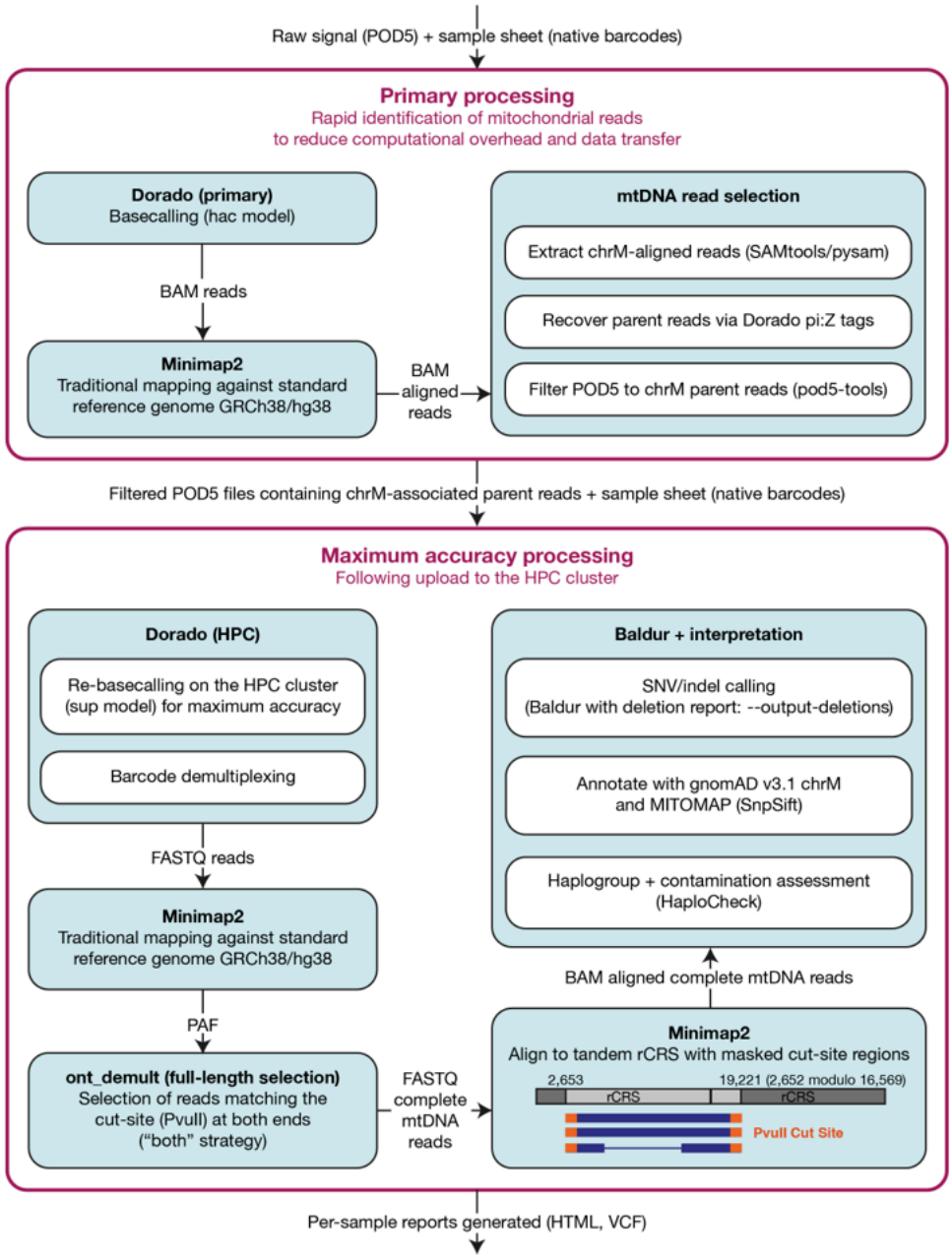


Figure 2

Read alignment strategy used in the Nanomito workflow. Full-length circular mtDNA molecules are represented in a linear coordinate system for variant and deletion detection. After cut-site-aware demultiplexing, reads are aligned to a custom mitochondrial reference comprising two tandem copies of the rCRS (tandem rCRS). This enables correct mapping of reads spanning the artificial linearization breakpoint while avoiding edge artifacts. In this coordinate system, full-length molecules align from the PvuII cut position (indicated in orange) to the end of the first rCRS copy and continue into the second copy, while mtDNA coordinates can also be interpreted as circular genome coordinates (modulo 16,569).

Figure 3

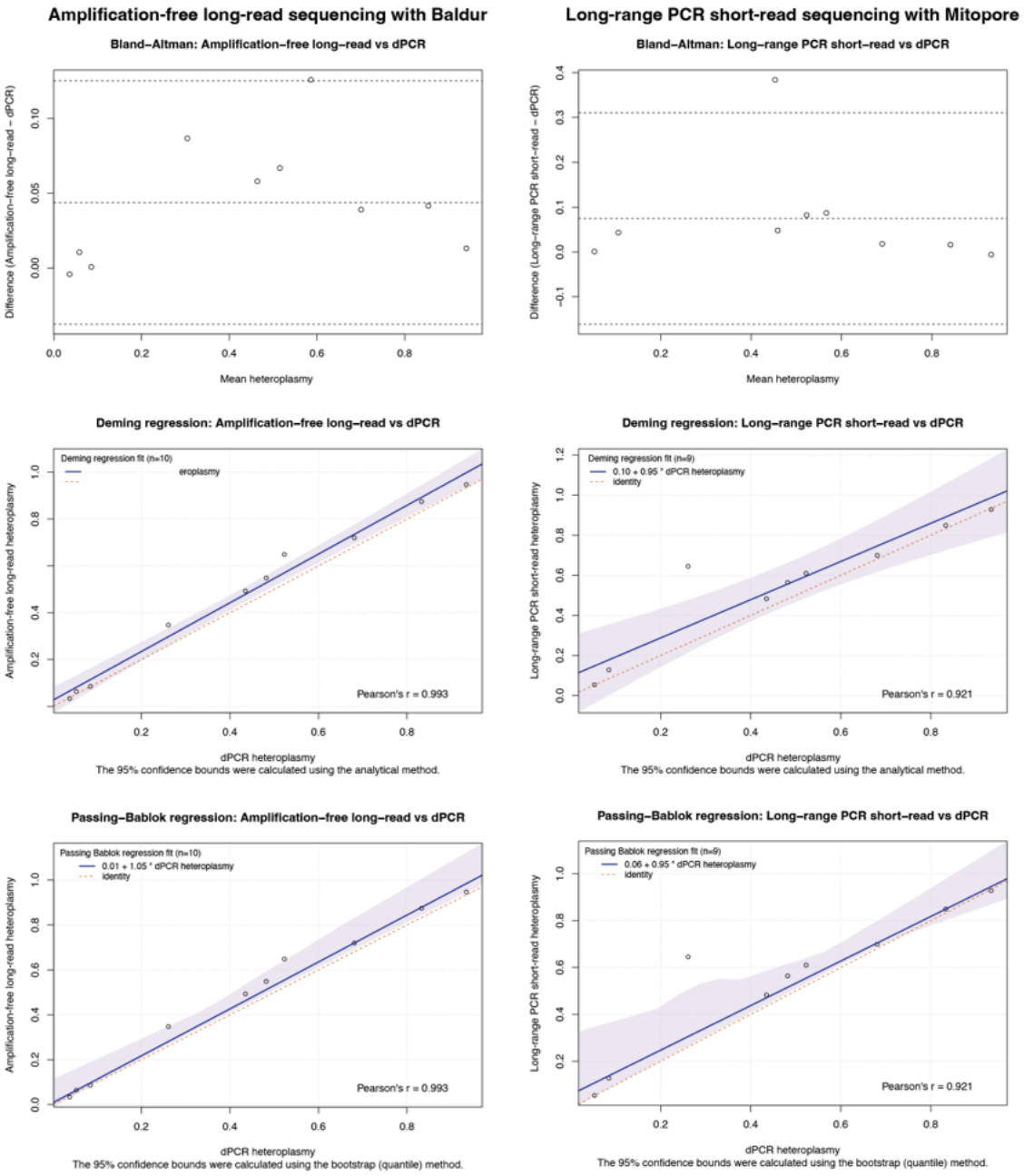


Figure 3

Comparison of sequencing-based heteroplasmy estimates with digital PCR. Bland-Altman, Deming regression, and Passing-Bablok regression analyses comparing heteroplasmy values obtained by amplification-free long-read sequencing and by long-range PCR short-read sequencing with dPCR for clinically confirmed pathogenic mtDNA point variants.

Figure 4

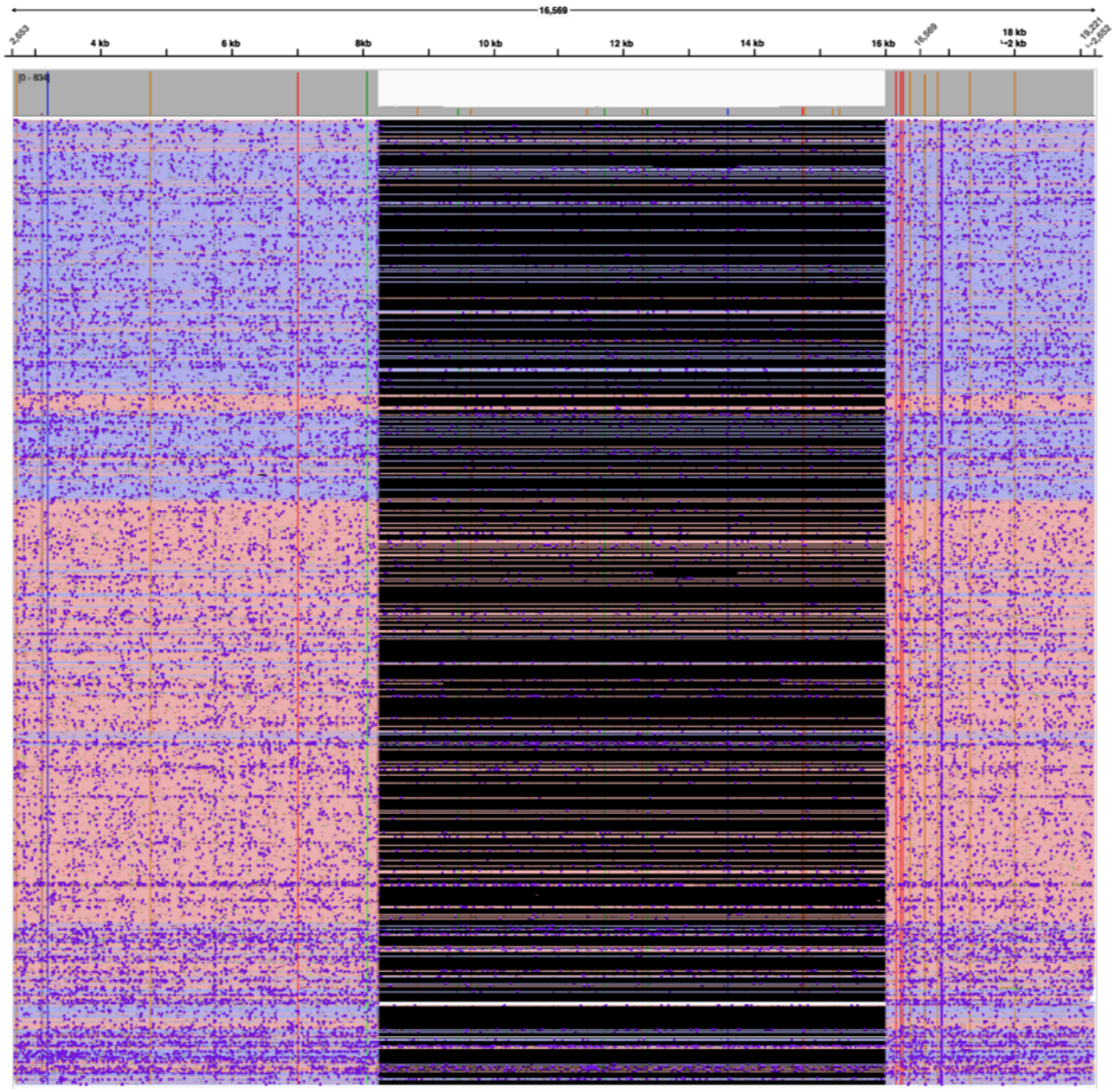


Figure 4

IGV visualization of reads from sample DEL-3 obtained by amplification-free long-read sequencing, showing a primary large-scale mtDNA deletion, m.8250_16032del, spanning 7,783 bp and detected at 79.9% heteroplasmy based on 832 complete mtDNA molecules. Alignments are displayed in squished mode with strand-specific coloring. The coverage track and coordinates, shown above the alignments, are displayed both as linear positions on the tandem reference and as circular genome coordinates (modulo 16,569). Full-length mtDNA molecules appear as continuous alignments, whereas deleted molecules show gaps corresponding to the missing segment, visible as thin black lines on a white background.

Figure 5

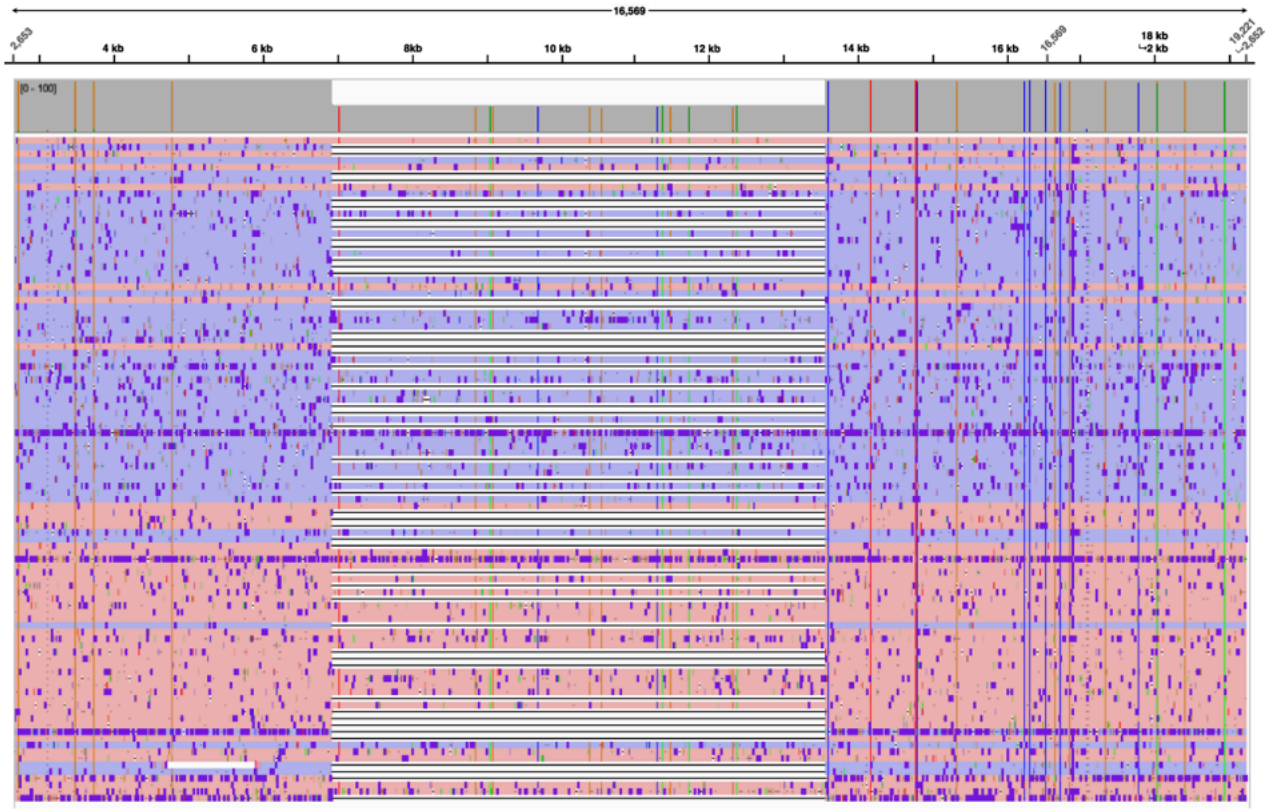


Figure 5

IGV visualization of reads from sample DEL-4 obtained by amplification-free long-read sequencing, showing a primary large-scale mtDNA deletion, m.6920_13553del, spanning 6,634 bp and detected at 49.5% heteroplasmy based on 100 complete mtDNA molecules. Alignments are displayed in squished mode with strand-specific coloring. The coverage track and coordinates, shown above the alignments, are displayed both as linear positions on the tandem reference and as circular genome coordinates (modulo 16,569). Full-length mtDNA molecules appear as continuous alignments, whereas deleted molecules show gaps corresponding to the missing segment, visible as thin black lines on a white background.

Figure 6

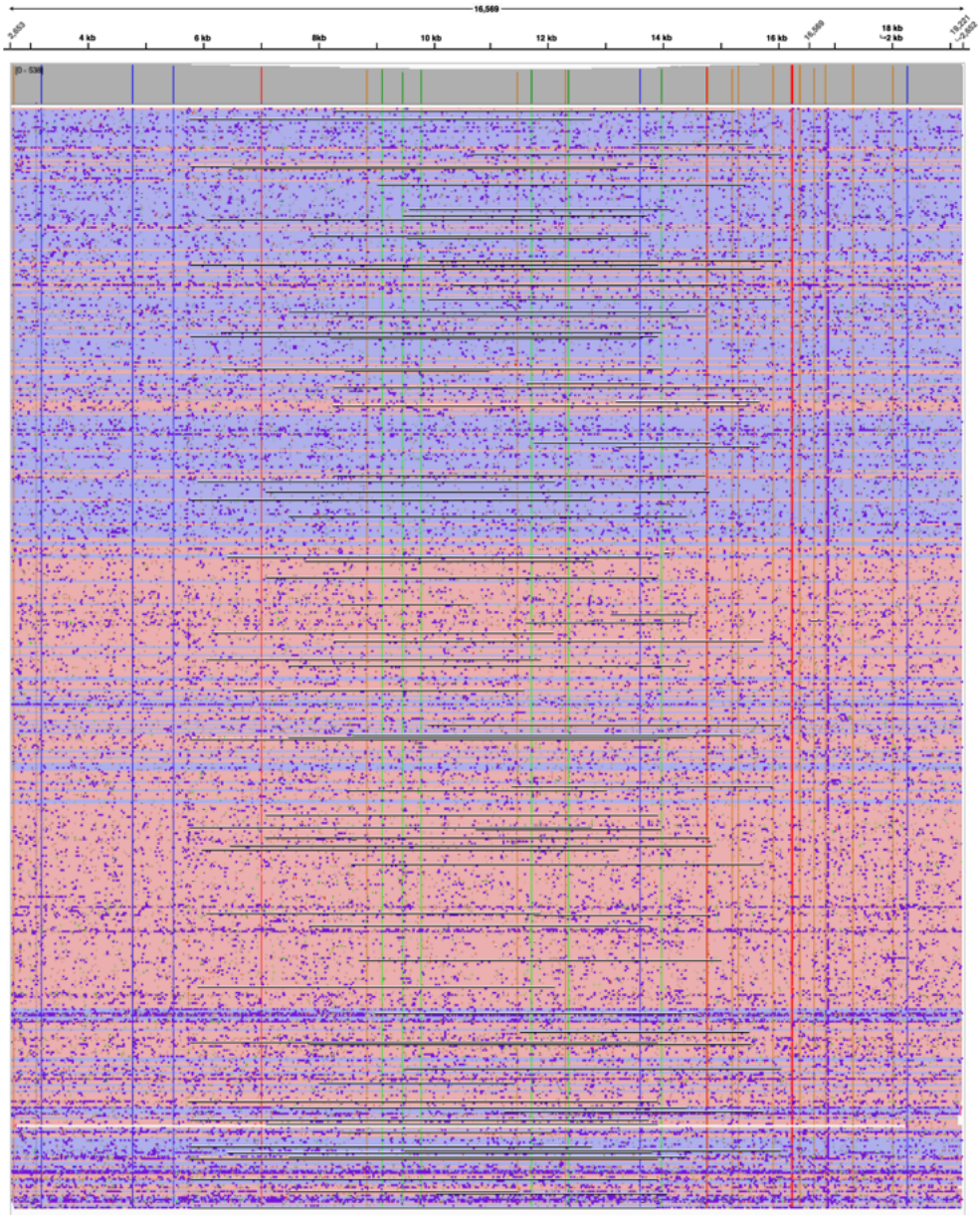


Figure 6

IGV visualization of reads from sample DEL-5 obtained by amplification-free long-read sequencing, showing a heterogeneous spectrum of 66 large-scale mtDNA deletions affecting 91 of 538 complete mtDNA molecules (16.9%). Individual deletions were detected at frequencies ranging from 0.2% to 0.7% (1–4 molecules) and spanned 64 to 8,407 bp. Alignments are displayed in squished mode with strand-specific coloring. The coverage track and coordinates, shown above the alignments, are displayed both as linear positions on the tandem reference and as circular genome coordinates (modulo 16,569). Full-length mtDNA molecules appear as continuous alignments, whereas deleted molecules show gaps corresponding to the missing segment, visible as thin black lines on a white background.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- nanomitoprocesseddata.zip
- nanomitosupplementaryfigures.pdf