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.

From benchmarking to application: a scalable mitochondrial DNA analysis and variant identification framework for whole-genome sequencing

Kelly Chen
Walter and Eliza Hall Institute of Medical Research
Michael Milton
Walter and Eliza Hall Institute of Medical Research
Victoria E. Jackson
Walter and Eliza Hall Institute of Medical Research
Haloom Rafehi
Walter and Eliza Hall Institute of Medical Research
Penny Snell
Murdoch Children's Research Institute, Royal Children's Hospital
Kayli C. Davies
Murdoch Children's Research Institute, Royal Children's Hospital
Aleena Eldo
Murdoch Children's Research Institute, Royal Children's Hospital
Paul J. Lockhart
Murdoch Children's Research Institute, Royal Children's Hospital
Melanie Bahlo
Walter and Eliza Hall Institute of Medical Research
Longfei Wang
wang.lo@wehi.edu.au

Walter and Eliza Hall Institute of Medical Research

Research Article

Keywords:

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11468975/v1

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

Additional Declarations:
No competing interests reported.

Tables 1 to 4 are available in the supplementary files section

From benchmarking to application: a scalable mitochondrial DNA analysis and variant identification framework for whole-genome sequencing

Kelly Chen[1,2,3], Michael Milton[1,2], Victoria E. Jackson[1,2], Haloom Rafehi[1,2], Penny Snell[4], Kayli C. Davies[4,5], Aleena Eldo[4], Paul J. Lockhart[4,5], Melanie Bahlo[1,2], Longfei Wang[1,2]*

[1] Walter and Eliza Hall Institute of Medical Research, 1G Royal Pde Parkville, Melbourne, VIC 3052 Australia

[2] Department of Medical Biology, University of Melbourne, Parkville, VIC 3052 Australia

[3] UWA Medical School, University of Western Australia, Perth, WA 6009 Australia

[4] Bruce Lefroy Centre, Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, VIC 3052 Australia

[5] Department of Paediatrics, University of Melbourne, Royal Children's Hospital, Parkville, VIC 3052 Australia

*Corresponding author

Longfei Wang (wang.lo@wehi.edu.au)

Abstract

Purpose: Mitochondrial DNA (mtDNA) variants are implicated in a wide range of diseases. The increasing availability of whole-genome sequencing enables mtDNA analysis in large-scale clinical and population studies. However, optimal tools for accurate variant detection from short-read data remain unclear.

Methods: We systematically benchmarked four scalable mtDNA variant calling tools: MitoHPC, mtSwirl, mity, and mtDNA-Server 2. Using 1000 Genomes Project (1KG) data and simulations across diverse haplogroups, we evaluated performance in detecting homoplasmic and heteroplasmic variants. Using the optimal tool, we built MARVel

(https://github.com/bahlolab/MARVel), an integrated workflow combining variant calling, automated quality-control reporting, and interactive variant curation, including an mtDNA principal component reference from 1KG that captures haplogroup structure and enables projection of external cohorts. We then applied MARVel to an independent ataxia cohort to assess research utility.

Results: MitoHPC and mtSwirl showed the strongest overall performance, particularly in non-European populations, with MitoHPC preferred for its ease of implementation. mity excelled in detecting low variant allele frequency single nucleotide variants (SNVs) but underperformed for insertions/deletions and showed inconsistent multi-nucleotide polymorphism normalisation. mtDNA-Server 2 maximised SNV sensitivity but had a higher false-positive rate and missed variants in the *MT-RNR2* region. Applying MARVel (MitoHPC) to the ataxia cohort recovered known pathogenic mtDNA variants and identified additional candidates for investigation.

Conclusion: We provide guidance for mtDNA tool selection in large-scale sequencing studies, along with a reproducible benchmarking framework for future evaluation as methods evolve. We also developed MARVel to facilitate disease-associated mtDNA variant discovery and reporting, supporting robust and scalable mtDNA analysis.

Keywords: mitochondrial DNA, variant calling, benchmarking, whole-genome sequencing, mitochondrial disease

Introduction

The mitochondrial genome is a circular, double-stranded DNA molecule of 16,569 base pairs in the revised Cambridge Reference Sequence (rCRS; NC_012920.1) that encodes 13 protein-coding genes, 22 transfer RNAs (tRNAs), and 2 ribosomal RNAs (rRNAs), all of which are essential for cellular adenosine triphosphate (ATP) production through oxidative phosphorylation (OXPHOS) [1]. Unlike nuclear DNA (nDNA), mitochondrial DNA (mtDNA) is inherited exclusively through the maternal lineage [2,3], and each cell contains multiple copies of mtDNA, with quantities ranging from none in erythrocytes to several hundred thousand in oocytes [4].

The polyploid nature of the mitochondrial genome results in characteristic features of mitochondrial genetic variation: homoplasmy and heteroplasmy. Homoplasmic variants are present in all mtDNA copies within a cell, whereas heteroplasmic variants are those that occur in only a subset of mtDNA molecules within any given cell [1]. The proportion of mutant mtDNA, often quantified as variant allele frequency (VAF), influences phenotypic expression and disease severity. Clinical manifestations typically emerge when a pathogenic heteroplasmy threshold is exceeded [5]. Heteroplasmy levels vary across tissues and cell types [6], in particular, heteroplasmy levels are typically higher in disease-affected tissues and lower in more accessible tissues, such as blood, reflecting selection against deleterious variants in rapidly dividing cell populations [2,7].

The mitochondrial genome is prone to high levels of variation, where the mtDNA mutation rate is estimated to be approximately 20-fold higher than that of the nuclear genome [8]. Variants in mtDNA are implicated in various human diseases and are highly heterogeneous in their clinical presentations [9,10]. Primary mitochondrial disorders (PMDs) can arise from pathogenic variants in either mtDNA or nDNA genes that impair OXPHOS, often leading to multisystemic syndromes, such as mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes (MELAS) and Leigh syndrome. New legislative frameworks in multiple countries, such as the UK and Australia, have now introduced mitochondrial donation into clinical practice, enabling mothers with PMD-associated mtDNA variants to utilise *in vitro* techniques that transfer healthy mitochondria to their children. Beyond PMDs, mitochondrial dysfunction has been implicated in common complex disorders, including Parkinson's disease, motor neuron disease, and epilepsy.

In these conditions, mtDNA variation is thought to influence disease susceptibility, progression, and clinical heterogeneity [11–13]. Accurate mtDNA variant calling is essential for identifying pathogenic variants, quantifying heteroplasmy, and supporting accurate clinical variant interpretation, which is fundamental to understanding the aetiology of mitochondrial diseases.

Short-read whole-genome sequencing (WGS) has transformed biomedical research and clinical diagnostics by enabling the comprehensive detection of genetic variations across both nuclear and mitochondrial genomes, unlike short-read whole-exome sequencing, which provides only incidental and inconsistent coverage of the mitochondrial genome. The clinical utility of WGS has expanded rapidly in recent years, driven by substantial reductions in sequencing costs and the recognition of novel genetic variation only captured by WGS. This has also led to policy developments, such as the adoption of WGS for diagnostic use within public healthcare systems, including in Australia. Despite these advances, mtDNA analysis has often been overlooked or inconsistently integrated into standard WGS pipelines. Large-scale population resources, such as the UK Biobank [14], provide a powerful opportunity to systematically characterise mtDNA variations and identify novel disease-associated mtDNA variants, but require high-fidelity variant calling for evaluation. Leveraging these datasets can improve our understanding of disease mechanisms, inform clinical stratification, and support the development of precision medicine approaches and future therapeutic trials. Beyond variant discovery, high mtDNA copy number gives far greater sequencing depth than the diploid nuclear genome, enabling the detection of low-level alternative mtDNA alleles arising from cross-sample contamination. This allows identification of low-level contamination that nDNA methods (e.g., VerifyBamID [15]) miss at standard 20-30x coverage, with contamination detectable down to ~1% [16]. As a result, mtDNA-based screening can serve as an effective quality control (QC) step for improving the reliability of true rare somatic variant detection. Collectively, the diverse roles of mtDNA highlight the need for accurate and robust variant calling methods.

Accurate detection of mtDNA variants from short-read sequencing data remains challenging because of several intrinsic features of the mitochondrial genome, including its circular structure and the presence of nuclear mitochondrial DNA segments (NUMTs) [17]. The circular genome poses challenges for read alignment. Traditional analysis pipelines developed for the linear nuclear

genome require the insertion of an artificial breakpoint usually introduced at the origin of the mitochondrial genome to convert the circular genome into a linear one. Consequently, reads spanning the origin may be misaligned or only partially mapped to the genome.

NUMTs are mtDNA-origin sequences that have been inserted into the nuclear genome. Ancient NUMTs, which predate the emergence of anatomically modern humans, are largely fixed in the human genome. Polymorphic NUMTs represent more recent insertions that vary between individuals [18,19]. The presence of NUMTs leads to read mismapping and false-positive variant calls, as short reads from NUMTs in the nuclear genome may be incorrectly aligned to the mitochondrial genome or vice versa [20,21]. These characteristics complicate the distinction between true biological variations and technical artefacts, particularly for low-level heteroplasmic variants.

Several computational tools have been developed specifically for mtDNA variant calling [22,23], including pipelines that explicitly account for circular structure and NUMTs, and are able to detect low-level heteroplasmy. Early tools were designed for mitochondrial genome (MT)-targeted sequencing and small cohorts, limiting their scalability for large WGS datasets, which are increasingly being analysed in cloud and high-performance computing (HPC) environments. Recent tools have addressed these limitations and enabled population-scale analyses, which are essential for discovering novel pathogenic variants and disease risk factors. However, substantial differences in algorithms, filtering strategies, and annotations lead to variability in the variant calls across methods, particularly for heteroplasmic variants and indels. Therefore, systematic benchmarking of scalable mtDNA variant calling tools is essential to guide tool selection and establish best practices. To date, comprehensive benchmarking studies of mtDNA variant calling remain limited. Previous benchmarking studies [24,25] have focused on heteroplasmy detection using amplicon sequencing or smaller WGS cohorts, relying on earlier pipelines not designed for scalable cloud/HPC deployment, limiting their applicability to contemporary large-scale WGS analyses.

Here, we present comprehensive benchmarking of the four most widely used and scalable mtDNA variant calling tools: MitoHPC [26], mtSwirl [27], mity [28], and mtDNA-Server 2 [29]. Using

both high-coverage WGS data from the 1000 Genomes Project (1KG) [30,31] and in silico spike-in simulations, we systematically evaluated the performance of these mitochondrial analysis tools in detecting both homoplasmic and heteroplasmic variants across a range of VAFs. Based on these results, we provide much needed guidance for tool selection in large-scale sequencing studies. Furthermore, we created a streamlined mtDNA analysis and variant discovery workflow, MARVel (https://github.com/bahlolab/MARVel). Together, these resources establish a robust and scalable framework for accurate mtDNA variant identification and discovery in large cohorts, enabling comprehensive mtDNA analysis and reliable mtDNA variant calling across diverse sequencing datasets.

Materials and methods

Tools

In this study, we focused on four recently developed mtDNA variant calling tools for single-nucleotide variants (SNVs) and insertions/deletions (indels), that are suitable for large-scale WGS data and can be deployed on cloud computing platforms. We summarize the key features of the evaluated tools in Table 1 and describe each tool in detail below.

Table 1. (See Table file 1: Table 1)

MitoHPC. Mitochondrial High Performance Caller (https://github.com/dpuiu/MitoHPC) is a bioinformatic pipeline designed to estimate mtDNA copy number (mtDNA-CN) and identify mitochondrial variants from WGS data [26]. MitoHPC employs a two-step variant calling strategy. In the first iteration, variants are called using rCRS as the reference. This initial call set is then used to construct a sample-specific mitochondrial reference sequence, which is applied in a second iteration of variant calling to improve the detection of heteroplasmic variants. To address the challenges arising from NUMTs, MitoHPC recovers reads that are falsely mapped to known NUMTs regions on chromosomes 1 and 17 and reduces false positive (FP) heteroplasmy calls from polymorphic NUMTs by leveraging the sample-specific consensus. To mitigate the coverage bias introduced by the circular genome, MitoHPC remaps reads to a circularized mitochondrial genome reference with a 300 bp extension beyond position m.16569. The pipeline integrates three

variant callers: GATK Mutect2 [32], mutserve [16], and FreeBayes [33], and provides an option to downsample mtDNA reads to 222,000 reads (~2000× coverage) to improve computational efficiency. Beyond the annotations generated by Mutect2, MitoHPC further augments variant annotations with the mitochondrial genomic context, including mitochondrial hypervariable regions and gene-level information.

mtSwirl. mtSwirl (https://github.com/rahulg603/mtSwirl) [27] is a scalable pipeline for calling mtDNA variants and estimating mtDNA-CN from WGS data. It extends the gnomAD mtDNA workflow [34] by constructing an individual-specific self-reference sequence and explicitly accounting for NUMTs. mtSwirl has been reported to outperform the original gnomAD pipeline [27], with improved mtDNA coverage, particularly for individuals of African ancestry, and fewer FP calls at very low heteroplasmy levels. As this improvement has already been benchmarked, we decided to not re-evaluate the original gnomAD pipeline separately in our comparison. Similarly to MitoHPC, mtSwirl is built on GATK Mutect2, using a two-step realignment strategy, and mitigates the "edge effect" at the artificial mtDNA breakpoint via a shifted reference. It further extends these approaches by using a curated set of 385 putative NUMT intervals (instead of focusing just the ancient NUMTs on chr1 and chr17) and applies a larger control-region shift, 8000bp rather than 300bp. mtSwirl was originally designed for execution on cloud platforms only. To broaden accessibility, we adapted the workflow for HPC environments by splitting it into modular tasks, addressing Hail memory limitations, optimizing task-level memory allocation, replacing HTTP-based imports with local file imports, and removing DNAnexus-specific functionality (https://github.com/bahlolab/mtSwirl_HPC).

mity. The mity (https://github.com/KCCG/mity) [28] analysis pipeline was developed to detect, filter, and annotate SNVs and indels in mtDNA from WGS or MT-only data. The pipeline consists of three primary modules: *mity-call*, *mity-report,* and *mity-merge*. *mity-call* performs variant detection using optimized FreeBayes v1.2.0 [33] settings: minimum alternate reads (C) = 4, minimum heteroplasmy (F) = 0.005, mapping quality $\geq$ 30, base quality $\geq$ 24. It also applies a custom Phred-scaled variant quality score (q) based on a binomial model and noise threshold (p). Additional filters are applied to reduce FPs and sequencing artefacts, including a strand-bias filter and exclusion of known problematic regions (m.302-319 and m.3105-3109). *mity-report* generates

easily interpretable reports with extensive annotations, including predicted functional impact (Variant Effect Predictor [35]), clinical information (MITOMAP [36], MitoTIP [37]), and population allele frequency data (Medical Genome Reference Bank [38]). *mity-merge* facilitates integration into existing WGS pipelines by combining mtDNA variants from the mity Variant Call Format (VCF) with nuclear variants from a genome-wide VCF. A major strength of mity is its high sensitivity for detecting low-level heteroplasmy of <1%. However, mity does not explicitly account for coverage bias in the control region because of its circular structure or NUMT-related FPs. Instead, it recommends reporting Tier 1 variants (VAF ≥1%) as a depth-based heuristic to minimize NUMT-related artefacts.

mtDNA-Server 2. mtDNA-Server2 (https://github.com/genepi/mtdna-server-2) [29] is a major update to the widely used mtDNA-Server web service, which showed the best performance in a previous benchmarking study [24] compared with earlier tools such as mitoCaller [39], MToolBox [22], and MitoSeek [23]. The updated version retains the core capabilities of the original pipeline, including its mtDNA-optimized SNV caller, mutserve [40], which uses a Bayesian model to account for sequencing error profiles, alongside integrated quality/contamination checks and a user-friendly web interface. A key enhancement is the "fusion" calling mode, which combines SNVs from mutserve2 with indels from GATK Mutect2 to address the limited indel sensitivity in mutserve. The workflow has also been migrated to Nextflow DSL2 [41] and containerized, enabling reproducible execution on local HPC clusters and cloud platforms. Heteroplasmy detection uses a default 2% threshold but is user-configurable, and variant calls can be further used for haplogroup estimation with Haplogrep 3 [42] and contamination detection with Haplocheck [16]. Similar to MitoHPC, users have the option to subsample the input BAMs to ~2000× coverage to reduce the runtime. However, the pipeline currently accepts only BAM input, despite CRAM being widely used.

Empirical data: 1000 Genomes Project dataset

The 1000 Genomes (1KG) database is a large public catalogue of human genetic variations from globally diverse samples [43]. It encompasses 26 populations across five continents, providing a useful empirical framework for benchmarking mtDNA variant calling across diverse mitochondrial genetic ancestries. Samples are available as lymphoblastoid cell lines (LCLs) and

LCL-derived DNA through the Coriell Cell Repository, and the dataset is consented for full public release without access or use restrictions. It has been used as a benchmarking dataset in hundreds of publications.

As empirical short-read data for benchmarking, we used the expanded 1KG high-coverage WGS dataset, comprising 3,202 individuals sequenced to a targeted depth of ~30× on the Illumina NovaSeq 6000 platform and aligned to GRCh38. We accessed the aligned reads as CRAM files and performed individual-level analyses on the Terra platform. Specifically, we developed a WDL workflow on Terra to run the MitoHPC and mtSwirl pipelines and extract mtDNA reads, as mtDNA-Server 2 cannot be executed directly on Terra. We then downloaded the MitoHPC and mtSwirl outputs, together with the extracted mtDNA CRAMs, and ran mity and mtDNA-Server on our local HPC cluster. To compare homoplasmic and heteroplasmic variant calling, we analyzed 604 mother-offspring pairs with available pedigree information. For mitochondrial PCA, we used 2,589 unrelated individuals in the analysis, verified using KING [44].

Ataxia cohort

The ataxia dataset is described in [45] and has since grown to 557 individuals recruited across Australia and New Zealand. Blood-derived DNA was subjected to WGS using Illumina sequencing technology (target mean coverage 30×). Raw reads were aligned to GRCh38, with nuclear variants called using the GATK Best Practices workflow. This cohort was used to evaluate the research utility of MARVel by assessing its ability to detect disease-associated mtDNA variants and to demonstrate the variant analysis workflow.

Evaluation of homoplasmy and heteroplasmy detection using 1KG

To compare the performance of the four tools in calling homoplasmic mtDNA variants, we leveraged 604 mother-offspring pairs available in the 1KG dataset. Because homoplasmic variants are present in essentially all mtDNA molecules in the mother, they are expected to be transmitted to the offspring and remain homoplasmic in the offspring. Therefore, we used mother-offspring concordance as an internal consistency metric: truly homoplasmic sites should be shared within a pair, and consequently a higher number of overlapping homoplasmic calls indicates more reliable

homoplasmic variant calling. In addition, we quantified cross-tool disagreement and visualized tool-specific patterns by identifying calls missed by each tool and mapping their genomic positions along the mitochondrial genome.

Heteroplasmic variants typically arise from somatic mutations and, more rarely, reflect maternal inheritance [46]. We used the same 604 mother-offspring pairs to evaluate heteroplasmy detection, but under a different assumption: within each individual, most heteroplasmic variants represent *de novo* somatic events. We defined high-confidence heteroplasmic variants as those detected by ≥3 of 4 tools. Per the UpSet plot (Additional file 1: Figure S1), 89.5% of variants (2,594/2,899) had three- or four-tool support, versus 10.5% with minority support, likely reflecting tool-specific artefacts. To address potential bias toward shared tool behaviour introduced by this consensus approach, we performed a simulation study. Tool-specific performance was further assessed by identifying variants detected by three tools but missed by the fourth and comparing both the counts and genomic distributions of these missed calls to characterise tool-specific limitations.

The evaluated tools apply inconsistent definitions and thresholds for homoplasmy and heteroplasmy. Specifically:

- **MitoHPC**: defines heteroplasmy as 3% ≤ VAF < 95%; it does not explicitly define homoplasmy but, effectively treats variants with VAF ≥95% as homoplasmic.
- **mtSwirl**: explicitly defines homoplasmy as VAF ≥95% and filters heteroplasmic calls below a 5% VAF to minimize FPs from NUMTs.
- **mity**: identifies homoplasmy via the genotype (GT = 1) without a stated VAF cutoff, and heteroplasmic variants (GT = 0) are further stratified into reporting tiers, with Tier 1 requiring a VAF ≥1%.
- **mtDNA-Server 2**: uses a Bayesian model with a default heteroplasmy detection limit of 2% VAF and reports variants by "Type" in its output; homoplasmic SNVs are classified as Type 1 when the posterior probability favours the homoplasmic genotype, whereas heteroplasmic SNVs are classified as Type 2 and indels as Type 3.

To ensure comparability across tools in subsequent analyses, we adopted unified criteria: homoplasmic variants were defined as VAF ≥95%, and heteroplasmic variants as 3% ≤ VAF < 95%.

Quality control

Mity and mtSwirl provide a blacklist to exclude artefact-prone sites, whereas the other two tools do not. Mity and mtSwirl provide different blacklists: mity excludes two regions, m.302-319 and m.3105-3109, while mtSwirl excludes six variants: m.301, m.302, m.310, m.316, m.3107(N), and m.16182. These variants mainly fall into two categories: 1) m.3107, which contains an unknown base ("N") in the reference genome, leading to misalignment issues, and 2) homopolymer tracts. A homopolymer tract is a region of DNA composed of consecutive repeats of the same nucleotide, typically defined as a run of ≥3–5 identical bases (e.g. CCCCC). Homopolymer tracts are prone to sequencing and alignment errors: short-read sequencing technologies (e.g. Illumina) are particularly susceptible to slippage in these regions, and variant callers often show elevated FP rates as a result. In practice, variant callers inconsistently detect variants within homopolymer tracts (Additional file 2: Table S1). To simplify and standardise the subsequent analyses, we generated a single unified blacklist to exclude all artefact-prone sites.

To generate this blacklist, we assumed that true heteroplasmic variants are typically rare and unlikely to be shared across many samples. We first identified all homopolymer tracts consisting of ≥5 identical nucleotides and quantified the heteroplasmic variants occurring within these regions. Variants located within homopolymer tracts and observed in >30 samples were flagged as likely artefacts, which yielded a final unified blacklist of 13 positions: m.301, m.302, m.310, m.316, m.3107, m.5894, m.16179, m.16181-16183, m.16188, m.16189, m.16192. All subsequent analyses excluded the variants at these positions. 12 of the 13 positions fall within non-coding D-loop or intergenic regions, consistent with known homopolymer artefacts, and m.3107 corresponds to a reference-sequencing gap rather than a true variable site. This blacklist is provided as a standalone resource (Additional file 2: Table S2; also available in the MARVel GitHub repository) for use as an input filter in other mtDNA variant-calling pipelines.

In addition to artefact-prone sites, we excluded variants that failed the tool-specific filtering criteria, mainly relevant to strand bias, as outlined below:

- MitoHPC: exclude variants flagged as “strand bias” in the FILTER column.
- mtSwirl: exclude variants flagged as “strand_bias”, “base_qual”, “position”, “FAIL” or “map_qual”.
- mity: retained variants with PASS in the FILTER column, as well as variants with FAIL only when the failure was due to POS_FILTER (POS_FILTER = 0), as blacklist positions were filtered separately.
- mtDNA-Server 2: exclude variants flagged for “strand bias” in the Filter column.

Simulated data generation

Assembly-based somatic callers (e.g. Mutect2) tend to down-weight the minor haplotype, leading to biased heteroplasmy level estimates, whereas pileup-based mtDNA callers recover the expected allele fractions. To avoid this source of method-dependent bias and enable a fairer comparison across tools, we used in silico spike-in simulations. We generated read-level mtDNA simulations by spiking known SNVs and indels into empirical WGS alignments using BamSurgeon [47], which directly injects synthetic mutations at defined VAFs into existing sequencing reads. This contrasts with approaches that simulate heteroplasmy by mixing polymorphic sites from two samples [34], which instead produce a mosaic of divergent haplotypes rather than a true within-read heteroplasmic mixture.

We selected 27 background samples from 1KG, one per macro-haplogroup, and used the corresponding WGS CRAM files for analysis. To define the spike-in target sites, we first generated a set of “safe” candidate positions on mtDNA by excluding loci with pre-existing variant calls in any of the four callers in the 1KG data (on average, accounting for 0.48% of sites). Because BamSurgeon drops low-quality reads during remapping and discards a simulated variant if >10% of reads are dropped, we further excluded regions prone to this behaviour: homopolymer tracts, artificial breakpoint regions near the start/end of mtDNA (m.1-150 and m.16420-16569) and known artefact-prone hotspots (m.301-319, m.16179-16194, and m.3107). From the remaining sites, we sampled 40 SNVs across mtDNA using binned sampling with a minimum spacing constraint (≥200 bp). We also sampled ten indel targets under the same spacing constraint: five

deletions of 1/3/5/7/9 bp and five insertions with sequences C/ATG/ATGGC/ATGGCAT/ATGGCATGC.

MT reads were extracted from the background CRAM and downsampled to ~2000× depth based on the measured mean mtDNA coverage. Spike-in simulations were performed with BamSurgeon v1.4.1 (addindel.py and addsnv.py) on the downsampled mtDNA BAM, injecting 10 indels and 40 SNVs at prescribed heteroplasmy levels (VAF = 1%, 5%, 10% and 50%). Finally, the simulated mtDNA BAM was merged with the nuclear-only aligned reads from the same background sample to generate a final simulated WGS dataset for each VAF setting.

Performance evaluation on spike-in simulation

We applied the evaluated mtDNA variant calling tools to each simulated dataset. Tool outputs were harmonised into per-target tables for SNVs and indels, including pass/fail QC flags, after applying tool-specific and shared filters (e.g., removing artefact-prone sites and strand bias flagged calls). Targets dropped during remapping were also excluded. To quantify FPs, we compared calls from the simulated run against calls from the original background sample for each tool and retained calls present only in the simulated output, excluding any positions overlapping injected SNVs or indel windows.

Performance was summarised for each tool and VAF setting using (i) true positive (TP) counts for SNVs (TP_{SNV}) and indels (TP_{INDEL}), (ii) FP counts, (iii) sensitivity = TP/(TP + FN), (iv) precision = TP/(TP + FP), and (v) F1 = 2 × (precision × sensitivity)/(precision + sensitivity). In addition, we provide FP, precision, and the F1 score after applying each tool's default VAF threshold. This evaluation process was repeated across 27 simulated datasets to assess the robustness of tool performance across diverse mtDNA haplogroup backgrounds.

Principal component analysis

Principal component analysis (PCA) was performed on mtDNA variants called by MitoHPC from 1KG to establish a reference space for the projection of external cohorts. Per-sample mtDNA variant call files (VCFs) were collected for all unrelated individuals (n = 2,589). Variants were

restricted to biallelic homoplasmic single nucleotide polymorphisms (SNPs) by applying a VAF threshold of ≥ 95%. Individual VCFs were merged into a multi-sample dataset with multi-allelic sites excluded. The control region is highly variable and enriched with homopolymer tracts, resulting in elevated mutation rates and increased susceptibility to sequencing and alignment artefacts. Therefore, we excluded the control region (m.16024–16569 and m.1–576) to minimise technical and biological confounding. Variants were further filtered to those with a common allele frequency (1% ≤ AF < 99%) and low missingness (≤ 1%). After QC, we generated a filtered VCF file comprising 317 mtDNA SNPs across 2,589 individuals.

We then converted the filtered VCF into a genotype matrix, with homoplasmic alternate calls assigned a value of 1 and all other genotypes assigned 0. The genotype matrix was centred using the mean allele frequency of each variant. Principal components (PCs) were computed by singular value decomposition using the "svd()" function in R. Variant identifiers, variant loadings, and eigenvalues for the first 20 PCs are available on GitHub (https://github.com/bahlolab/MARVel/tree/main/mtDNA_report/1000G) to define a reference coordinate system and enable reproducibility for other mtDNA studies.

External cohorts with WGS can be projected into this reference space through a standardised harmonisation workflow comprising five steps: i) restriction to the same set of mtDNA SNPs (*mtPCA.variant_ids.txt*); ii) harmonisation of genotype encodings; iii) imputation of missing genotypes as reference alleles, to ensure compatibility with the reference matrix; iv) centering of genotype vectors using the reference mean (*mtPCA.mu.txt*); and v) projection of samples into the reference PCA space by multiplying the centred genotype vector by the pre-computed loadings (*mtPCA.loadings.PC1-20.txt*). This approach ensures reproducible embedding across independent datasets.

mtDNA variant curation

mtDNA Curation is an R Shiny application for interactive review of per-sample mtDNA variant calls from MitoHPC. It parses each sample's Mutect2 VCF to extract VAF, depth, genotype, FILTER status, gene, and embedded in silico scores: APOGEE [48] for protein-coding variants and mitochondrial local constraint [49] for all variants. Variants are flagged for the artefact

blacklist, hypervariable regions, homopolymer context, and indel status, then classified by VAF as homoplasmic, VAF ≥0.95, heteroplasmic, 0.03 ≤ VAF < 0.95, or low-heteroplasmic, VAF <0.03.

Variants are further annotated against gnomAD mtDNA v3.1 [34], providing overall and haplogroup-specific homoplasmic/heteroplasmic allele frequencies and counts matched to each sample's own haplogroup; MITOMAP [36], providing disease association, curation status, and homoplasmic/heteroplasmic report status; and MitoTIP [37], providing a pathogenicity score and quartile for mt-tRNA variants. We additionally annotated variants with an in-house 1KG allele frequency reference (homoplasmic and heteroplasmic allele count/frequency) computed from the 2,589 unrelated individuals used for PCA, provided as a standalone resource (https://github.com/bahlolab/MARVel/blob/main/mtDNA_curation/database/allele_freq_1000g.csv)

Adjustable filters include minimum VAF, PASS-only, strand bias, blacklist, hypervariable regions, synonymous consequence, region, gene, sample ID, and maximum in-cohort allele count. A genotype-matched gnomAD MAF threshold is also applied, under which homoplasmic variants are filtered against gnomAD homoplasmic frequency and heteroplasmic variants against gnomAD heteroplasmic frequency, each checked against both overall and haplogroup-specific populations. The resulting variant table itself supports further interactive filtering, sorting, and per-column search, allowing users to refine results beyond these predefined criteria.

Given a PLINK FAM file, the app supports maternal inheritance analysis. For a proband, it identifies maternal-line comparators (mother and/or siblings sharing the same maternal ID) and filters variants as all, shared (present in the proband and ≥1 comparator), or de novo (private to the proband), reflecting mtDNA's strictly maternal, non-recombining inheritance. Results are exportable to CSV, Excel, or PDF.

Statistical methods

To assess differences across tools, including the number of overlapping homoplasmic variants and heteroplasmy level estimates, we applied the Friedman test using the “friedman.test()” function

from the rstatix R package (version 0.7.3). For pairwise comparisons, we used the Wilcoxon signed-rank tests ("pairwise_wilcox_test()" and "wilcox_effsize()" rstatix R package, version 0.7.3) with Benjamini-Hochberg correction for multiple testing. All statistical analyses were performed using R version 4.5.2.

Results

Parameter optimization and operational performance

We first used a randomly selected 1KG sample (HG00405) to optimize the parameter settings for each mtDNA variant calling tool (MitoHPC, mtSwirl, mity, and mtDNA-Server 2) and to assess runtime, memory usage, and ease of use across both HPC and cloud computing platforms (Table 2).

Table 2. (See Table file 1: Table 2)

We evaluated the available options for each tool and selected the optimal configurations for comparison. Both MitoHPC and mtDNA-Server 2 provide an option to subsample reads to a fixed target depth of 2,000x before variant calling, to reduce the computational burden of mtDNA's naturally high depth. Although subsampling (SS) substantially reduced runtime (Table 2), it also affected VAF accuracy. For MitoHPC Mutect2, 6/51 overlapping variants (11.8%) showed discordant VAF between SS and non-subsampling (NSS) runs (Additional file 1: Figure S2A-B; differences -0.03 to 0.033), enough in some cases to shift the homoplasmy/heteroplasmy classification. This was more pronounced for MitoHPC FreeBayes, where 94/143 variants (65.7%) showed VAF differences (Additional file 1: Figure S2C-D; range -0.040 to 0.032), concentrated in the 3–10% VAF range. Therefore, we used NSS settings for both tools to prioritise accuracy (MitoHPC: HP_L=; mtDNA-Server 2: subsampling = "off"). MitoHPC provides three variant caller options: Mutect2, mutserve, and FreeBayes. The mutserve option could not be evaluated because it consistently failed to run, returning the error "Input file must be bam file, not sam file" despite providing a BAM-formatted input file. FreeBayes recovered most variants identified by Mutect2 (49/51) but additionally reported 105 heteroplasmic calls, most of which had VAF <10% (Additional file 1: Figure S3), and did not provide filtering annotations, limiting our ability to

reliably distinguish true variants from FPs. Therefore, we selected Mutect2 for MitoHPC (HP_M=mutect2). We also identified an edge-case bug in MitoHPC's alignment script, triggered when a structural deletion spans the artificial breakpoint in the linearised mitochondrial reference: reads downstream of the breakpoint were truncated from the BAM file, affecting not only depth calculation but all downstream analyses for the affected sample. We reported this upstream (issues #30) and corrected it in our fork (https://github.com/bahlolab/MitoHPC). mtDNA-Server 2 offers three calling configurations: Mutect2-only, mutserve2-only, and fusion mode (mode = "fusion"). The fusion mode integrates SNVs from mutserve2 with indels from Mutect2 to leverage complementary strengths and has been reported by the developers to provide the best overall accuracy and performance [29]. Therefore, we selected the NSS fusion mode for downstream analyses. Accordingly, comparative analyses were performed using NSS Mutect2 for MitoHPC and the NSS fusion mode for mtDNA-Server 2.

On HPC, mity and mtDNA-Server 2 were the most user-friendly to install and run and had the shortest runtimes per sample (mity: ~1 min; mtDNA-Server 2: ~7 min), as they process only the MT. In contrast, MitoHPC and mtSwirl required substantially longer runtimes (MitoHPC Mutect2: ~30 min; mtSwirl: ~1.5 h), reflecting their additional steps required to mitigate NUMT-related artefacts and construct sample-specific self-references. Memory usage was modest overall, with mity requiring the least memory (0.16 GB) and mtSwirl requiring the most (~10 GB).

On cloud computing platforms, we assessed the performance only under optimal parameter settings. mity was again the easiest and fastest to deploy (runtime 11 min, estimated cost $0.03 on Terra). MitoHPC showed moderate usability on the cloud (runtime ~1 h, cost $0.19), whereas mtSwirl required substantially longer runtime (4.5 hours) despite a comparable estimated cost ($0.16). These runtime and cost estimates are based on a single sample (HG00405, 15.16 GB CRAM) and may not generalise across datasets varying in file size or coverage. mtDNA-Server 2 is not directly compatible with Terra, as its workflow is implemented in Nextflow, whereas Terra natively supports WDL-based workflows; however, it can in principle be run on other cloud platforms that support Nextflow (e.g., Google Cloud), or via its web server for small cohorts by uploading MT-only short-read sequencing data. mtSwirl is primarily designed for cloud execution,

which may limit accessibility for users relying on local HPC infrastructure. To address this, we implemented a HPC-compatible version of mtSwirl (https://github.com/bahlolab/mtSwirl_HPC).

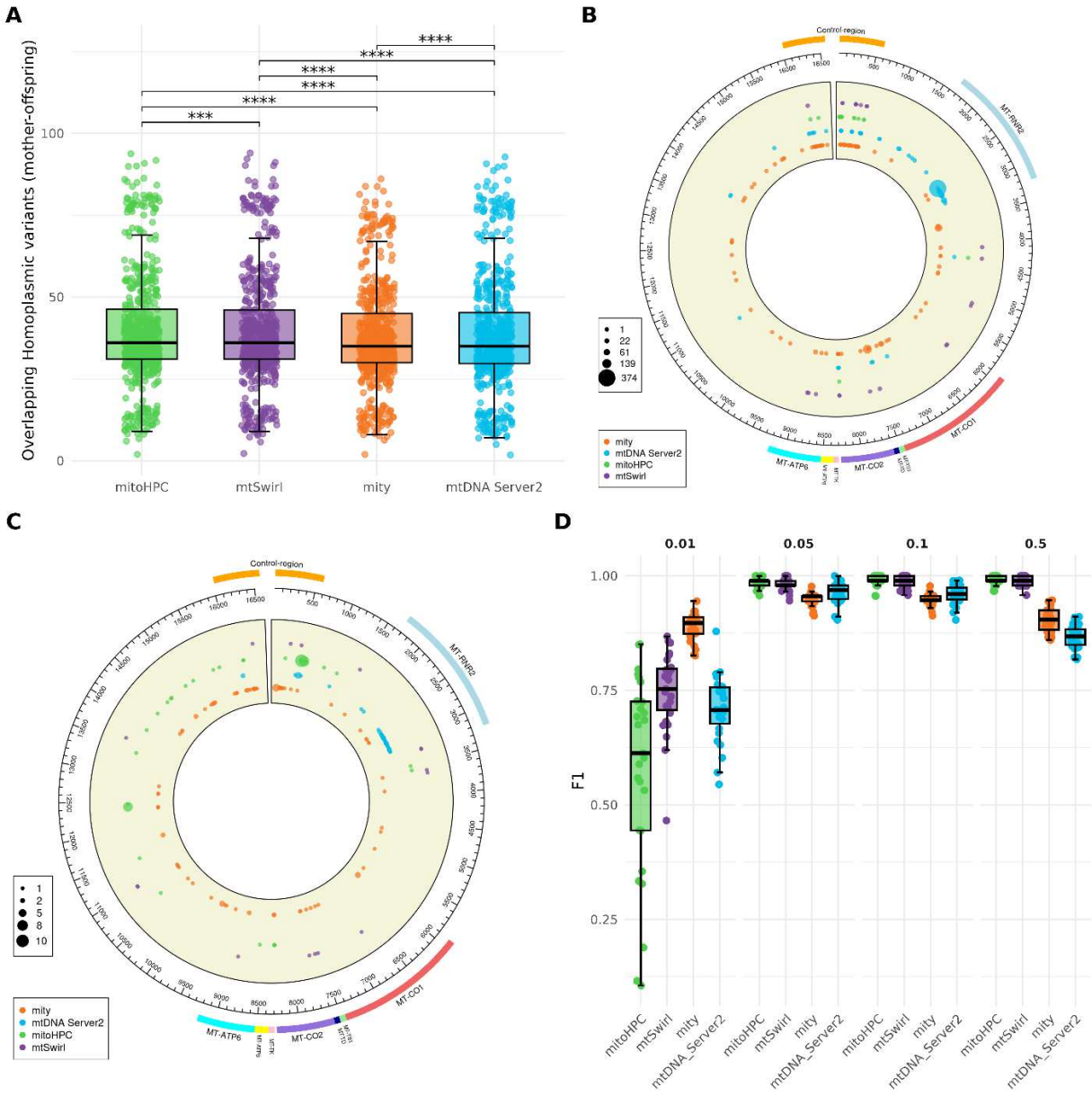

Figure 1. **Benchmarking mtDNA variant calling tools for homoplasmic and heteroplasmic variant detection. A. Concordance of overlapping homoplasmic variant calls in mother-offspring pairs across tools.** Box plots show the distribution of the total number of overlapping homoplasmic variants shared within each mother-offspring pair (n = 604) for each tool from the 1000 Genomes Project (1KG). Horizontal bars indicate statistical significance between tools,

calculated using pairwise Wilcoxon signed-rank tests (*** $p < 0.001$; **** $p < 0.0001$). **B. Genomic distribution of homoplasmic variant detection gaps across the mtDNA variant calling tools.** The plot shows 160 discordant homoplasmic variants, where at least one individual in the mother-offspring pair was missed by the labelled tool, but was identified by at least one other pipeline. Dot size represents the number of mother-offspring pairs in which a homoplasmic variant was missed at a given position. The labelled control region and *MT-RNR2* gene are particularly problematic. **C. Genomic distribution of heteroplasmic variant detection gaps across four mtDNA variant calling tools.** The plot shows the 176 heteroplasmic variants that were not called by a given tool but were detected by the other three tools. Dot size indicates the number of individuals in whom this heteroplasmic variant inconsistency was detected. **D. Detection performance of the evaluated tools across haplogroups and heteroplasmy levels in the simulations.** Boxplots showing the distribution of F1 scores for each tool across 27 haplogroups from simulated data with injected variants at various variant allele frequency (VAF) settings. F1 was calculated as 2 × (precision × sensitivity)/(precision + sensitivity), with tool-specific cut-offs applied to all VAFs, except VAF = 1%.

Homoplasmy detection evaluation

To assess the homoplasmic variant calling performance, we leveraged the 604 mother-offspring pairs from the 1KG WGS dataset. Because true homoplasmic variants are expected to be maternally inherited and therefore shared between mothers and their offspring, greater overlap was interpreted as higher calling reliability. Using a unified definition of homoplasmy as VAF ≥ 95% across all four tools, we calculated the number of overlapping homoplasmic variants within each mother-offspring pair for each tool (**Methods**). Overlap counts differed significantly across tools (Figure 1A; Friedman test: $\chi^2(3) = 1250.36$; $p < 0.0001$). Specifically, MitoHPC achieved the highest concordance of homoplasmic variants between mothers and their offspring (median = 36.00, IQR: 31.00-46.25), followed closely by mtSwirl (median = 36.00, IQR: 31.00-46.00). MitoHPC showed a small but significant increase in concordance relative to mtSwirl (pairwise Wilcoxon signed-rank test: r = 0.16, $p = 0.0003$). The overlap of mtDNA-Server 2 (median = 35.00, IQR: 29.75-45.25) was significantly lower than that of both MitoHPC (r = 0.95; $p < 0.0001$) and mtSwirl (r = 0.93; $p < 0.0001$), while mity (median = 35.00, IQR: 30.00-45.00) showed the lowest

concordance overall and was significantly lower than that of mtDNA-Server 2 (r = 0.30; p <0.0001). P-values were adjusted for multiple testing using the Benjamini-Hochberg procedure.

We then examined the cross-tool discordance in detail. Most discrepancies were attributable to mity's handling of multi-nucleotide polymorphisms (MNPs), which were not consistently normalised into constituent SNVs (e.g., m.15850TA>CG). In these cases, mity detected the same underlying variation as the other tools but failed to separate or normalise the calls consistently. After identifying all 131 MNPs (Additional file 2: Table S3) and excluding them from downstream analyses, 160 homoplasmic variants remained discordant across tools (Figure 1B; Additional file 2: Table S4). Of these, 62 (39%) were located in the control region (m.16024-16569 and m.1-576), where reduced read depth due to the artificial breakpoint in the linearised reference complicated variant detection. MitoHPC and mtSwirl recovered more variants in this region because they account for the circular genome by shifting the breakpoint by 300 bp and 8000 bp, respectively. A further 32 (20%) discordant variants mapped to *MT-RNR2* (m.1671-3229), which encodes the mitochondrial 16S rRNA, and were predominantly missed by mtDNA-Server 2, indicating a genomic region-specific limitation of this pipeline.

Among the loci discordant with mtSwirl, 10/22 (45%) fell below the 95% VAF homoplasmy threshold, indicating that mtSwirl identified these sites but classified them as heteroplasmic. Notably, 60% (6/10) of these cases occurred solely in individuals from non-European haplogroups (B, D, L, and M). This pattern is consistent with mtSwirl recovering additional NUMT-derived reads that are not retained by the other tools, which can reduce the observed VAF and shift calls from homoplasmy to heteroplasmy.

MitoHPC showed the fewest discordant loci (N = 15), with most (13/15, 87%) occurring in the homopolymer tracts, suggesting MitoHPC may be more robust to artefacts arising from sequencing errors in these challenging regions, albeit potentially at the cost of reduced sensitivity.

Other inconsistencies likely reflect tool-specific filtering and reporting behaviours. For example, several variants were removed by mity's strand bias ratio filter despite being called by other pipelines. mity also excludes reads with mapping quality <30 prior to variant calling, which

removes many NUMT-derived reads; however, this QC step can also eliminate supporting reads for true mtDNA variants in regions prone to ambiguous mapping. In addition, mtDNA-Server 2 exhibited reporting discrepancies, in which some variants appeared in the haplogroup output but were absent from the corresponding VCF files, including a subset of variants in *MT-RNR2*.

Furthermore, to evaluate whether self-referencing improves variant calling in non-European ancestries, we quantified the cross-tool discordance separately in haplogroup H (54 pairs; 91% European) and haplogroup L (213 pairs; 94% African). Overall concordance was higher in haplogroup H, with only 11 discordant loci, whereas substantially greater discordance was observed in haplogroup L (72 discordant loci). MitoHPC and mtSwirl recovered more variants in haplogroup L (Additional file 1: Figure S4). This pattern is consistent with the rCRS reference being predominantly derived from European ancestry, which may provide a less optimal representation of more divergent non-European haplogroups. By incorporating self-referencing, MitoHPC and mtSwirl likely mitigate reference bias and improve variant identification in these samples.

Collectively, our findings demonstrate that mtDNA-Server 2 and mity show clear drawbacks in their detection of homoplasmic variants, whereas MitoHPC provides the most reliable homoplasmic calls across genomic contexts and populations, with mtSwirl showing a broadly comparable performance.

Heteroplasmy detection evaluation

Heteroplasmic variant calling performance was first assessed using 1KG mother-offspring pairs. We focused on high-confidence heteroplasmic variants only, identifying those with $3\% \leq \text{VAF} < 95\%$ detected by at least three tools. Of these high-confidence variants, a subset of 351 was detected by all four tools, allowing direct comparison of VAF estimates across pipelines for the same variant. Comparison of their VAF distributions across tools revealed a significant overall difference (Friedman test: $\chi^2(3) = 257.19, p < 0.0001$), although median VAFs were nearly identical across pipelines (VAF ≈ 8%; Additional file 2: Table S5). Pairwise comparisons indicated small but consistent differences in VAFs between the tools (Additional file 1: Figure S5). These results suggest that while all methods provide comparable central estimates of heteroplasmy, subtle tool-

specific effects influence VAF quantification across variants. Haplogroup-specific analyses of heteroplasmy detection also demonstrated that MitoHPC and mtSwirl identified more variants in non-European populations owing to their self-referencing approach (Additional file 1: Figure S6).

Next, we assessed cross-tool discordance by examining variants called by three tools but missed by the fourth. After removing MNP-related discordances in mity, 176 heteroplasmic loci remained discordant (Additional file 2: Table S6). Overall, mtSwirl missed the fewest heteroplasmic variants (N = 16), followed by MitoHPC (N = 37), mtDNA-Server 2 (N = 53), and mity (N = 74). We observed patterns similar to those seen in the homoplasmic analysis (Figure 1C). Half of the discordant loci fell within either the control region (21%) or *MT-RNR2* (30%), which were mainly missed by mtDNA-Server 2. For MitoHPC, 23 (62%) missed calls occurred in the homopolymer tracts (Additional file 2: Table S6).

Overall, mtSwirl demonstrated the highest sensitivity for heteroplasmic variant detection, as indicated by the lowest number of missed variants, whereas MitoHPC performed comparably when loci in homopolymer tracts were excluded. mtDNA-Server 2 and mity showed the same tool-specific limitations observed for homoplasmy detection.

However, this consensus-based approach may introduce bias toward shared tool behaviour. Moreover, in the absence of a known ground truth, the analysis is limited to assessing the sensitivity of high-confidence heteroplasmic variants and does not permit estimation of precision. In addition, low-level heteroplasmy (e.g., VAF <3%) cannot be reliably evaluated. To address these limitations, we generated 27 simulated datasets, one per macro-haplogroup (e.g., H, L0, and L1), with predefined VAFs using an in-silico spike-in approach (**Methods**). We then applied all four tools to the simulated data and computed performance metrics for each sample, including true positives (TP), false positives (FP), sensitivity, precision, and F1 score. Average values across the simulations are presented in Table 3.

Table 3. (See Table file 1: Table 3)

Across BamSurgeon [47] spike-in simulations, 40 SNVs and 10 indels were injected per sample at specific VAFs (1%, 5%, 10%, and 50%). However, a small number of low-quality variants were dropped during BamSurgeon remapping, leaving ~47 evaluable targets on average per sample. Tool performance diverged the most at low heteroplasmy and in controlling FPs.

At 1% VAF, performance was the most variable. mity achieved the highest F1 score (0.89; Figure 1D), driven by a strong SNV recovery (95%) and high precision (91%). However, it missed approximately half of the indels, limiting overall sensitivity to 87%. This indel-calling limitation was consistent across VAF settings, with ~40% of indels missed. mtDNA-Server 2 showed the highest sensitivity (98%), with strong recovery of both SNVs (98%) and indels (97%), but substantially lower precision due to a large FP burden (mean FP = 38.11; precision = 56%; F1 = 0.71). mtSwirl had intermediate sensitivity (68%) but retained a low FP burden (mean FP = 6.00; precision = 84%; F1 = 0.75). MitoHPC was the most conservative at 1% VAF (45% sensitivity), with the lowest FP burden of all four tools (mean FP = 2.30; precision = 90%; F1 = 0.59; Figure 1D).

mtDNA-Server 2 consistently showed the highest sensitivity for SNV detection but also accumulated the largest number of FPs (38.11→ 66.93), leading to reduced precision (56% → 41%) across all VAF settings. Applying tool-specific VAF cutoffs (mity ≥1%, mtDNA-Server 2 ≥2%, MitoHPC/mtSwirl ≥3%) substantially improved precision for tools with many low-support calls, most notably mtDNA-Server 2 (mean FP = 3.37 and precision = 93% at 5-10% VAF), indicating that many FPs were concentrated below these thresholds. At 50% VAF, FP inflation became particularly pronounced for mtDNA-Server 2 (mean FP = 66.93) and mity (mean FP = 18.59), with many FPs occurring within ±10 bp of the injected indels. For example, in the L2 haplogroup simulation, 53% of mtDNA-Server 2 FPs occurred near the injected indels, and 90% occurred for mity. In contrast, MitoHPC showed the lowest FP burden (2.30-2.63) and the highest precision (90-95%) across all VAF settings, achieving near-perfect precision after applying a VAF ≥3% cutoff.

To summarise the balance between sensitivity and precision, we compared the F1 scores across the tools after applying the VAF cutoffs (Figure 1D). MitoHPC performed the poorest at 1% VAF

but achieved the best overall performance for studies prioritising variants with VAF ≥5%, with sensitivity and precision among the highest of all tools. Compared with MitoHPC, mtSwirl provided higher sensitivity at 1% VAF (Wilcoxon test: r = 0.81, p <0.0001) and similar sensitivity at VAF ≥5% (5% VAF: p = 0.29; 10% VAF: p = 0.48; 50% VAF: p = 0.40), but required additional filtering to mitigate its persistent low-level FP signature.

In summary, in addition to the tool-specific limitations identified in the homoplasmy detection evaluation, we observed under-detection of indels in mity and a high FP rate in mtDNA-Server 2. mtSwirl and MitoHPC showed very high sensitivity and precision at moderate levels of heteroplasmy but exhibited reduced sensitivity for low-level heteroplasmic variants (VAF = 1%).

MARVel and its application to an ataxia cohort

To standardise and streamline the process from mtDNA variant calling to cohort quality control to variant identification, we developed a modular analytical workflow, MARVel (https://github.com/bahlolab/MARVel), comprising three components (Figure 2): 1) *mtDNA Calling*, which, based on our benchmarking results, provides instructions for running MitoHPC with optimised parameters, including scripts for parallelising large cohorts on HPC and a WDL workflow for cloud execution; 2) *mtDNA Report*, an interactive HTML report that aggregates coverage metrics, contamination estimates, haplogroup assignments, and variant calls and classifications across all samples for cohort-level QC; and 3) *mtDNA Curation*, an interactive R Shiny app that supports clinicians and researchers in reviewing, filtering, and prioritising mtDNA variants identified from WGS data (**Methods**). *mtDNA Curation* presents variant-level annotations, including population frequency data, disease-associated evidence, and pathogenicity information, alongside sample-specific metrics such as read depth and VAF. In addition, it supports family-based analyses to compare shared and unique variants among maternally related individuals and evaluate segregation of candidate variants. Together, these components broaden the accessibility of mtDNA analysis for all users, including clinicians, while supporting secure analysis of sensitive data within controlled computing environments.

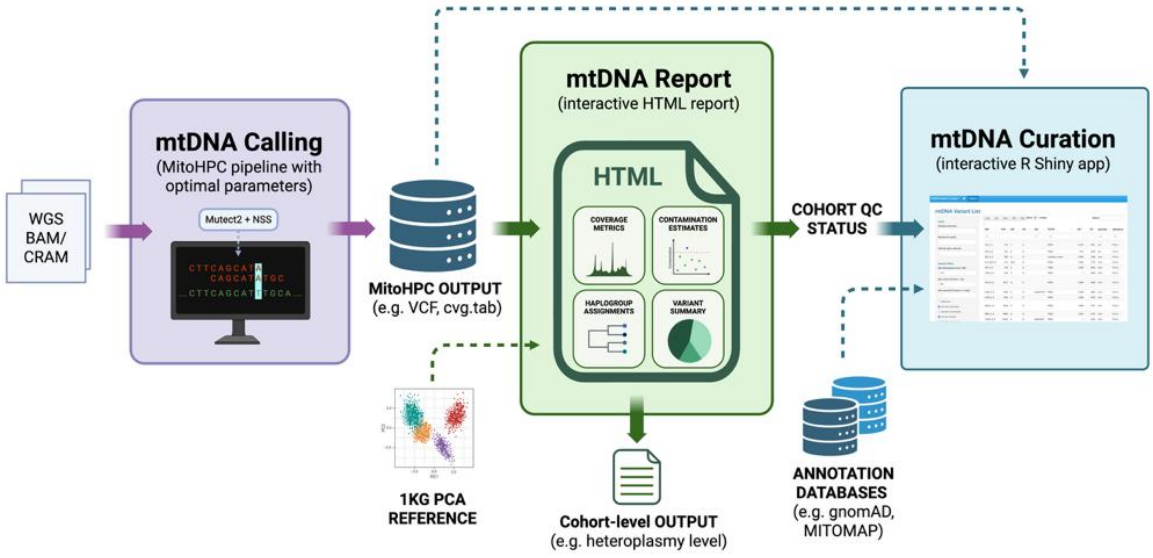

Figure 2. MARVel workflow schematic. The schematic provides an overview of MARVel, a modular workflow to facilitate mtDNA variant discovery and analysis. It comprises of 3 modules: 1) *mtDNA Calling*, which runs MitoHPC on WGS CRAM/BAM files with optimised parameters; 2) *mtDNA Report*, which combines output from MitoHPC with a 1KG PCA reference to generate an interactive HTML quality-control (QC) report and cohort-level output; and 3) *mtDNA Curation*, which integrates cohort QC status and external annotation databases into an interactive variant curation app.

We applied MARVel to an ataxia cohort of 557 participants to validate its capacity for disease-relevant variant discovery and demonstrate its research utility. Following the *mtDNA Calling* instructions, we ran all 557 samples in parallel on HPC, completing variant calling for the cohort with a net elapsed time of 6 hours, and 8 CPU cores allocated per sample.

Using *mtDNA Report,* we characterised the ataxia cohort across sequencing quality, haplogroup composition, and the mtDNA variant landscape, including heteroplasmy burden. Overall, sequencing quality was high, with all samples showing a median depth >100 (Additional file 1: Figure S7A). mtDNA copy number (mtDNA-CN) ranged from 38 to 540 across samples, consistent with blood-derived estimates [50] (Additional file 1: Figure S7B). A sample showed extreme coverage variation (CV = 0.374; Additional file 1: Figure S7C), with a coverage pattern

suggesting large structural variations or a sequencing-related technical artefact, which is also highlighted in the base coverage plot (Additional file 1: Figure S7D). No samples exceeded the 2% mtDNA contamination threshold (Additional file 1: Figure S8).

Haplogroup analysis (Additional file 1: Figure S9A) showed that macro-haplogroup H was the most common lineage (39.7%), followed by U (13.6%) and J (12.0%), consistent with a predominantly European-ancestry cohort. All assignments exceeded the 0.5 quality threshold; however, seven fell between 0.5 and 0.8, indicating moderate-confidence assignments (Additional file 1: Figure S9B). To further validate these assignments and detect sample swaps or outliers, the *mtDNA Report* projects sample haplogroups against a 1KG reference. PCA on mtDNA variants from 2,589 unrelated 1KG individuals established this projection reference (**Methods**): the first three PCs clearly separated the major founder haplogroups L0–L5, M, and N (Additional file 1: Figure S10A-B), and ataxia samples projected cleanly onto this reference (Additional file 1: Figure S10C-D). These clustering patterns reflect known human evolutionary history [2]: L3 is ancestral to the non-African founder haplogroups M and N, with M diversifying into Asian lineages (C, D, G) and N giving rise to R, which subsequently diversified into the major European haplogroups (H, J, T, U, V).

The *mtDNA Report* retains variants with VAF $\geq$ 3%, excluding those at artefact-prone loci or flagged for strand bias. Retained variants are classified by default as heteroplasmic ($0.03 \leq$ VAF < 0.95) or homoplasmic (VAF ≥ 0.95), though both thresholds are user adjustable. Across the ataxia cohort, 13,707 mtDNA variants were called, of which 96% were homoplasmic. Of the 1,822 mtDNA loci carrying at least one variant (11% of all positions), 77% were homoplasmic only, 16% heteroplasmic only, and 7% carried both. By variant type, 90% were SNV transitions, 7% SNV transversions and 3% indels (Additional file 1: Figure S11A). Synonymous variants were most abundant (793, 44%), followed by missense (418, 23%), intergenic (277, 15%), rRNA (169, 9%), and tRNA (108, 6%) (Additional file 1: Figure S11B). Synonymous and intergenic variants were predominantly homoplasmic, consistent with haplogroup-defining positions, whereas missense, tRNA and rRNA variants carried a substantial heteroplasmic fraction, consistent with purifying selection against deleterious mutations in these functionally constrained genes. The variant frequency spectrum (Additional file 1: Figure S11C) showed that singletons accounted for ~60%

of loci and were enriched for heteroplasmic-only variants, whereas homoplasmic-only variants dominated across all frequency classes.

The bimodal distribution of homoplasmic variant counts (Additional file 1: Figure S11D) reflects haplogroup structure: the first mode (peak ~11 variants) comprises predominantly N clade samples from haplogroups H and V, which are closest to the rCRS, while the second mode (peak ~31) represents N clade samples from haplogroups more divergent from the reference. The sparse L and M clade samples carry substantially more homoplasmic variants overall, largely reflecting haplogroup-defining differences from the rCRS. Heteroplasmic variants are predominantly somatic; due to negative selection in blood, most samples carry fewer than five heteroplasmic variants (Additional file 1: Figure S11E). Heteroplasmic variant count, sometimes termed heteroplasmy level (HL), accumulates with age [6] and is elevated in individuals with mitochondrial dysfunction, including those with *PINK1/PRKN* Parkinson's disease [51]. We also report the mitochondrial local constraint score sum (MSS) (Additional file 1: Figure S11F), which reflects the aggregate constraint burden of heteroplasmic variants per sample [49] and has been shown to associate with disease risk [52,53]. The report generates a per-sample QC summary table for cohort-level exploration, such as comparing heteroplasmy load between cases and controls.

Using *mtDNA Curation*, we applied a series of quality control (QC) and filtering steps to 33,054 initial mtDNA variant calls across 557 samples, following exclusion of duplicate WGS runs. First, low-quality variants flagged for strand bias, variants with a VAF <3%, and variants located within artefact-prone loci were excluded, retaining 13,784 high-confidence variants. Common variants were then removed using a genotype-matched gnomAD minor allele frequency (MAF) threshold of ≥0.1% (homoplasmic vs. heteroplasmic MAF, each checked against overall and haplogroup-specific frequencies) or an in-cohort minor allele count (MAC) >5. This filtering step reduced the dataset to 747 rare variants. After excluding synonymous, intergenic and hypervariable variants, 391 variants remained for downstream clinical review. We prioritised variants with MITOMAP-confirmed (Cfrm) or -reported (Reported) status whose associated disease relates to ataxia or the participants' own phenotype. All candidate variants are listed in Table 4.

Three variants (m.3243A>G, m.8851T>C and m.11778G>A), found across four individuals, had MITOMAP-confirmed (Cfrm) status, indicating pathogenicity independently reported by two or more laboratories [36]. However, m.11778G>A is excluded from Table 4: its associated MITOMAP disease, Leber hereditary optic neuropathy (LHON) or progressive dystonia, is not an established feature of ataxia, and the individual carrying this variant has an alternative genetic diagnosis that already explains their clinical presentation. m.3243A>G and m.8851T>C remain candidates for the cohort's ataxia phenotype: cerebellar ataxia is reported in an estimated 25–49% of MELAS cases, the phenotype predominantly caused by m.3243A>G [54], and this variant has been identified through targeted screening of previously unsolved cerebellar ataxia cohorts, at heteroplasmy levels (5.7–21%) comparable to those observed in our cohort [55]; m.8851T>C is associated with bilateral striatal necrosis and Leigh syndrome, in which cerebellar ataxia is a recognised feature, reported in up to 50% of mtDNA-associated Leigh syndrome cases [56]. In addition, after phenotype matching, we identified one further *MT-ND1* variant, m.3308T>C, whose MITOMAP-associated disease phenotype corresponds to the non-ataxia clinical feature of hypertension, that is also present in the relevant participant, alongside their ataxia. Collectively, these candidate variants (Table 4) warrant further clinical evaluation and validation in these participants.

Table 4. (See Table file 1: Table 4)

Discussion

In this study, we systematically benchmarked four recently developed, scalable mtDNA variant calling tools: MitoHPC, mtSwirl, mity, and mtDNA-Server 2; using both 1KG WGS data and in silico spike-in simulations. By evaluating the performance across homoplasmic and heteroplasmic SNVs and indels, we provide a systematic assessment of tool-specific strengths, limitations, and sources of discordance. Based on our findings, we propose operational recommendations for large-scale sequencing studies.

Because no single tool performs optimally across all scenarios, tool selection should align with study objectives and data type. For WGS data, we recommend MitoHPC (Mutect2 caller without read subsampling) as a default, given its robust overall performance, particularly in non-European ancestry samples, integrated contamination and haplogroup QC, and ease of implementation. mtSwirl offers comparable sensitivity, especially for heteroplasmic variants, but requires additional filtering to control low-level false positives. For studies prioritising low-level heteroplasmy in high-coverage WGS (e.g., ~30× from whole blood), mity offers high SNV sensitivity and precision exceeding mtDNA-Server 2, though its indel detection is limited and caution is needed at lower coverage, where NUMT-related false positives increase. When only MT-targeted sequencing is available, we recommend applying mity and mtDNA-Server 2 in parallel to offset their complementary limitations: missed *MT-RNR2* calls in mtDNA-Server 2, and MNP normalisation and indel-sensitivity issues in mity. For each tool, we also provide optimised parameter settings derived from this benchmarking, along with WDL workflows (https://github.com/bahlolab/MARVel/tree/main/mtDNA_calling/WDL) to enable straightforward deployment on cloud platforms.

Beyond variant calling, the newly developed mtDNA tool suite, MARVel, supports downstream analyses encompassing a complete mtDNA workflow. We created several resources for mitochondrial genomic community, including a unified artefact blacklist, a mitochondrial principal component analysis (mtPCA) reference for haplogroup visualisation and adjustment of population stratification in mitochondrial genome-wide association studies, as well as a 1KG-derived allele frequency panel to complement gnomAD annotations. These resources are intended to facilitate future mtDNA analyses. The *mtDNA Report* output can also be repurposed for cohort-

level analyses, such as summarising heteroplasmy levels across a study population, extending its utility beyond individual sample QC, extending to mega-cohort analyses, such as biobank data. Furthermore, MARVel's flexible deployment across secure computing environments, including local HPC infrastructure and cloud platforms, enables reproducible analysis, while also allowing sensitive genomic data to remain within the user's controlled computing environment, supporting data security and minimising unnecessary data transfer between platforms.

Our systematic benchmark of four recent, scalable mtDNA variant calling pipelines builds on Ip *et al.*'s previous benchmarking work, providing an updated assessment that is more relevant to current large-scale WGS studies and clinical workflows. Ip *et al*. [24] compared five earlier mtDNA variant calling tools – mutserve, mitoCaller, MitoSeek, MToolBox, and a custom GATK HaplotypeCaller pipeline – and reported strong agreement in homoplasmic variant detection but substantial discordance in heteroplasmic calls, with mutserve showing the highest overall accuracy. Consistent with these findings, we observed high concordance for homoplasmic variants across current scalable tools in a larger dataset. Importantly, our work further expands on this previous work by investigating the source of discordance and revealing tool-specific limitations.

Through our detailed analyses with both empirical and simulation studies, we find that each tool exhibits specific limitations. In addition to the *MT-RNR2* miscalling in mtDNA-Server 2 and MNP normalisation issue in mity, MitoHPC and mtSwirl also present tool-specific constraints. MitoHPC occasionally reported spurious REF=ALT calls at hypervariable positions (e.g., m.103G>G) and dropped genuine variants with VAF >3% during filtering. We avoided the latter by using the unfiltered `<ID>.mutect2.mutect2.00.vcf` output directly in our MARVel workflow. We also identified an edge-case bug in MitoHPC's alignment script, triggered when a structural deletion spans the artificial breakpoint in the linearised mitochondrial reference: reads downstream of the breakpoint were truncated from the BAM file, affecting all downstream analyses for the affected sample. We reported this upstream and corrected it in our fork (https://github.com/bahlolab/MitoHPC). mtSwirl, meanwhile, was originally designed for cloud platforms (Terra and DNAnexus) only, limiting its use in HPC environments. We addressed this by implementing an HPC-adapted version (https://github.com/bahlolab/mtSwirl_HPC). We reported the calling-related issues that we identified as part of our benchmarking in specific tools

to the respective developers (mity: issue #69; MitoHPC: issue #26, #27, #30; mtDNA-Server 2: issue #35). We welcome the opportunity to collaborate with tool developers in resolving identified issues.

Our evaluation was limited to the current tool versions, a common challenge in rapidly evolving bioinformatics fields, where new updates can quickly shift performance dynamics. Nevertheless, we provide a fully reproducible benchmarking framework that includes all analysis scripts and simulation data generation pipelines. This framework enables a straightforward re-evaluation as updated versions become available and facilitates extension to newly developed methods. We will maintain our framework and encourage community contributions, such as reporting improvements or the addition of new or improved mtDNA variant calling tools.

For the simulation study, existing laboratory-based mixture datasets are limited to MT-only sequencing data [24,40] and are thus not suitable for evaluating WGS-based tools, such as MitoHPC and mtSwirl. In addition, in silico mixture approaches [34] generate mosaics of divergent haplotypes, introducing bias in heteroplasmy estimates for assembly-based somatic callers. To minimise such method-dependent bias and enable a fairer comparison across tools, we used BamSurgeon for in silico spike-in simulations. However, this approach does not adequately model NUMTs and thus precludes direct assessment of NUMT correction. Instead, we demonstrated the advantage of self-referencing indirectly using empirical data through improved variant recovery in non-European ancestry samples. Moreover, BamSurgeon removes low-quality reads during remapping and discards simulated variants when >10% of reads are dropped. We had to exclude regions prone to this behaviour, such as homopolymer tracts, from the analysis. Consequently, the performance in these regions could not be assessed in the simulation study.

It is important to note that this study provides research and operational guidance for mtDNA variant calling, not clinical or analytical validation. Our truth model, mother-offspring genotype concordance combined with in silico spike-in simulations, is well suited to benchmarking tool performance under realistic conditions, but does not constitute validation against an orthogonally confirmed truth set. We did not establish a limit of detection, nor did we assess clinical sensitivity or specificity against independently confirmed diagnoses, due to a lack of appropriate data for

these analyses. Similarly, application of MARVel to the ataxia cohort identified variant candidates, which were not orthogonally confirmed by an independent method such as digital droplet PCR or Sanger sequencing, and this work makes no claims regarding accreditation frameworks such as ISO 15189 or NATA; this demonstrates the workflow's capacity for variant discovery, not evidence of diagnostic-grade performance.

The tools evaluated in this study were restricted to those designed for short-read WGS data to facilitate mtDNA analyses in existing clinical WGS and large-scale public datasets. Emerging evidence indicates that long-read sequencing offers advantages over short-read approaches, particularly in resolving misalignment caused by NUMTs and the artificial breakpoint introduced by the circular mitochondrial genome [58], both of which can affect variant detection in clinically relevant regions. Continued development and application of long-read sequencing technologies are likely to mitigate these challenges by improving alignment accuracy and enhancing variant detection [59].

Conclusion

In conclusion, our systematic benchmarking study demonstrated that mtDNA variant calling performance varies across tools, revealing tool-specific strengths, limitations, and sources of discordance in detecting homoplasmic and heteroplasmic variants. Based on these findings, we provide evidence-based recommendations for tool selection across large-scale sequencing studies, together with a reproducible benchmarking framework to support ongoing evaluation as these tools continue to evolve. We further introduce MARVel, an integrated workflow spanning variant calling, quality-control reporting, and interactive curation, and demonstrate its research utility through application to an independent ataxia cohort. Collectively, these resources provide a robust and scalable foundation for accurate mtDNA variant identification and discovery, supporting the broader integration of mitochondrial genome analysis into large-scale genomic studies.

List of abbreviations

- 1KG = 1000 Genomes Project
- mtDNA: mitochondrial DNA
- rCRS = revised Cambridge Reference Sequence
- SNV: single nucleotide variant
- tRNA: transfer RNA
- rRNA: ribosomal RNA
- ATP: adenosine triphosphate
- OXPHOS: oxidative phosphorylation
- nDNA: nuclear DNA
- PMD: primary mitochondrial disorder
- WGS: whole-genome sequencing
- NUMT: nuclear mitochondrial DNA segment
- HPC: high-performance computing
- PCA: principal component analysis
- VAF: variant allele frequency
- mtDNA-CN: mitochondrial DNA copy number
- MT: mitochondrial genome
- NSS: non-subsampling
- SS: subsampling
- FP: false positive
- VCF: Variant Call Format
- LCL: lymphoblastoid cell line
- CA: cerebellar ataxia
- TP: true positive
- FN: false negative
- SNP: single nucleotide polymorphism
- QC: quality control
- PC: principal component
- MNP: multi-nucleotide polymorphism
- HL: heteroplasmy level
- MSS: mitochondrial local constraint score sum
- MAC: minor allele count
- mtPCA: mitochondrial principal component analysis

References

1. Taylor RW, Turnbull DM. Mitochondrial DNA mutations in human disease. Nat Rev Genet. Nature Publishing Group; 2005;6:389–402. https://doi.org/10.1038/nrg1606

2. Stewart JB, Chinnery PF. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nat Rev Genet. Nature Publishing Group; 2015;16:530–42. https://doi.org/10.1038/nrg3966

3. Wei W, Pagnamenta AT, Gleadall N, Sanchis-Juan A, Stephens J, Broxholme J, et al. Nuclear-mitochondrial DNA segments resemble paternally inherited mitochondrial DNA in humans. Nat Commun. Nature Publishing Group; 2020;11:1740. https://doi.org/10.1038/s41467-020-15336-3

4. Wai T, Ao A, Zhang X, Cyr D, Dufort D, Shoubridge EA. The Role of Mitochondrial DNA Copy Number in Mammalian Fertility1. Biol Reprod. 2010;83:52–62. https://doi.org/10.1095/biolreprod.109.080887

5. Rossignol R, Faustin B, Rocher C, Malgat M, Mazat J-P, Letellier T. Mitochondrial threshold effects. Biochem J. 2003;370:751–62. https://doi.org/10.1042/bj20021594

6. Sondheimer N, Glatz CE, Tirone JE, Deardorff MA, Krieger AM, Hakonarson H. Neutral mitochondrial heteroplasmy and the influence of aging. Hum Mol Genet. 2011;20:1653–9. https://doi.org/10.1093/hmg/ddr043

7. Li M, Schröder R, Ni S, Madea B, Stoneking M. Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations. Proc Natl Acad Sci U S A. 2015;112:2491–6. https://doi.org/10.1073/pnas.1419651112

8. Árnadóttir ER, Moore KHS, Guðmundsdóttir VB, Ebenesersdóttir SS, Guity K, Jónsson H, et al. The rate and nature of mitochondrial DNA mutations in human pedigrees. Cell. 2024;187:3904-3918.e8. https://doi.org/10.1016/j.cell.2024.05.022

9. Gorman GS, Chinnery PF, DiMauro S, Hirano M, Koga Y, McFarland R, et al. Mitochondrial diseases. Nat Rev Dis Primer. Nature Publishing Group; 2016;2:16080. https://doi.org/10.1038/nrdp.2016.80

10. Nunnari J, Suomalainen A. Mitochondria: In Sickness and in Health. Cell. Elsevier; 2012;148:1145–59. https://doi.org/10.1016/j.cell.2012.02.035

11. Niyazov DM, Kahler SG, Frye RE. Primary Mitochondrial Disease and Secondary Mitochondrial Dysfunction: Importance of Distinction for Diagnosis and Treatment. Mol Syndromol. 2016;7:122–37. https://doi.org/10.1159/000446586

12. Exner N, Lutz AK, Haass C, Winklhofer KF. Mitochondrial dysfunction in Parkinson's disease: molecular mechanisms and pathophysiological consequences. EMBO J. 2012;31:3038–62. https://doi.org/10.1038/emboj.2012.170

13. Finsterer J, Mahjoub SZ. Epilepsy in mitochondrial disorders. Seizure - Eur J Epilepsy. Elsevier; 2012;21:316–21. https://doi.org/10.1016/j.seizure.2012.03.003

14. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015;12:e1001779. https://doi.org/10.1371/journal.pmed.1001779

15. Zhang F, Flickinger M, Taliun SAG, Consortium IPG, Abecasis GR, Scott LJ, et al. Ancestry-agnostic estimation of DNA sample contamination from sequence reads. Genome Res. Cold Spring Harbor Laboratory Press; 2020;30:185–94. https://doi.org/10.1101/gr.246934.118

16. Weissensteiner H, Forer L, Fendt L, Kheirkhah A, Salas A, Kronenberg F, et al. Contamination detection in sequencing studies using the mitochondrial phylogeny. Genome Res. Cold Spring Harbor Lab; 2021;31:309–16. https://doi.org/10.1101/gr.256545.119

17. Keraite I, Becker P, Canevazzi D, Frias-López C, Dabad M, Tonda-Hernandez R, et al. A method for multiplexed full-length single-molecule sequencing of the human mitochondrial genome. Nat Commun. Nature Publishing Group; 2022;13:5902. https://doi.org/10.1038/s41467-022-33530-3

18. Dayama G, Emery SB, Kidd JM, Mills RE. The genomic landscape of polymorphic human nuclear mitochondrial insertions. Nucleic Acids Res. 2014;42:12640–9. https://doi.org/10.1093/nar/gku1038

19. Wallace DC, Stugard C, Murdock D, Schurr T, Brown MD. Ancient mtDNA sequences in the human nuclear genome: A potential source of errors in identifying pathogenic mutations. Proc Natl Acad Sci. 1997;94:14900–5. https://doi.org/10.1073/pnas.94.26.14900

20. Albayrak L, Khanipov K, Pimenova M, Golovko G, Rojas M, Pavlidis I, et al. The ability of human nuclear DNA to cause false positive low-abundance heteroplasmy calls varies across the mitochondrial genome. BMC Genomics. 2016;17:1017. https://doi.org/10.1186/s12864-016-3375-x

21. Maude H, Davidson M, Charitakis N, Diaz L, Bowers WHT, Gradovich E, et al. NUMT Confounding Biases Mitochondrial Heteroplasmy Calls in Favor of the Reference Allele. Front Cell Dev Biol. 2019;7:201. https://doi.org/10.3389/fcell.2019.00201

22. Calabrese C, Simone D, Diroma MA, Santorsola M, Guttà C, Gasparre G, et al. MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. Bioinformatics. 2014;30:3115–7. https://doi.org/10.1093/bioinformatics/btu483

23. Guo Y, Li J, Li C-I, Shyr Y, Samuels DC. MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. Bioinformatics. 2013;29:1210–1. https://doi.org/10.1093/bioinformatics/btt118

24. Ip EKK, Troup M, Xu C, Winlaw DS, Dunwoodie SL, Giannoulatou E. Benchmarking the Effectiveness and Accuracy of Multiple Mitochondrial DNA Variant Callers: Practical Implications for Clinical Application. Front Genet [Internet]. Frontiers; 2022 [cited 2025 Dec 21];13. https://doi.org/10.3389/fgene.2022.692257

25. Fazzini F, Fendt L, Schönherr S, Forer L, Schöpf B, Streiter G, et al. Analyzing Low-Level mtDNA Heteroplasmy—Pitfalls and Challenges from Bench to Benchmarking. Int J Mol Sci. Multidisciplinary Digital Publishing Institute; 2021;22:935. https://doi.org/10.3390/ijms22020935

26. Battle SL, Puiu D, TOPMed mtDNA Working Group, Verlouw J, Broer L, Boerwinkle E, et al. A bioinformatics pipeline for estimating mitochondrial DNA copy number and heteroplasmy levels from whole genome sequencing data. NAR Genomics Bioinforma. 2022;4:lqac034. https://doi.org/10.1093/nargab/lqac034

27. Gupta R, Kanai M, Durham TJ, Tsuo K, McCoy JG, Kotrys AV, et al. Nuclear genetic control of mtDNA copy number and heteroplasmy in humans. Nature. Nature Publishing Group; 2023;620:839–48. https://doi.org/10.1038/s41586-023-06426-5

28. Puttick C, Davis RL, Kumar KR, Quinn JM, Zeng T, Fares C, et al. mity: A Highly Sensitive Mitochondrial Variant Analysis Pipeline for Whole Genome Sequencing Data. J Bioinforma Syst Biol. Fortune Journals; 2024;7:5–16.

29. Weissensteiner H, Forer L, Kronenberg F, Schönherr S. mtDNA-Server 2: advancing mitochondrial DNA analysis through highly parallelized data processing and interactive analytics. Nucleic Acids Res. 2024;52:W102–7. https://doi.org/10.1093/nar/gkae296

30. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. Nature Publishing Group; 2015;526:68–74. https://doi.org/10.1038/nature15393

31. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell. Elsevier; 2022;185:3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004

32. Auwera G van der, O'Connor BD. Genomics in the cloud : using Docker, GATK, and WDL in Terra. First edition. Sebastopol, CA: O'Reilly Media; 2020.

33. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing [Internet]. arXiv; 2012 [cited 2025 Dec 14]. https://doi.org/10.48550/arXiv.1207.3907

34. Laricchia KM, Lake NJ, Watts NA, Shand M, Haessly A, Gauthier L, et al. Mitochondrial DNA variation across 56,434 individuals in gnomAD. Genome Res. Cold Spring Harbor Lab; 2022;32:569–82. https://doi.org/10.1101/gr.276013.121

35. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122. https://doi.org/10.1186/s13059-016-0974-4

36. Kogelnik AM, Lott MT, Brown MD, Navathe SB, Wallace DC. MITOMAP: a human mitochondrial genome database. Nucleic Acids Res. 1996;24:177–9. https://doi.org/10.1093/nar/24.1.177

37. Sonney S, Leipzig J, Lott MT, Zhang S, Procaccio V, Wallace DC, et al. Predicting the pathogenicity of novel variants in mitochondrial tRNA with MitoTIP. PLoS Comput Biol. 2017;13:e1005867. https://doi.org/10.1371/journal.pcbi.1005867

38. Pinese M, Lacaze P, Rath EM, Stone A, Brion M-J, Ameur A, et al. The Medical Genome Reference Bank contains whole genome and phenotype data of 2570 healthy elderly. Nat Commun. Nature Publishing Group; 2020;11:435. https://doi.org/10.1038/s41467-019-14079-0

39. Ding J, Sidore C, Butler TJ, Wing MK, Qian Y, Meirelles O, et al. Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools. PLOS Genet. Public Library of Science; 2015;11:e1005306. https://doi.org/10.1371/journal.pgen.1005306

40. Weissensteiner H, Forer L, Fuchsberger C, Schöpf B, Kloss-Brandstätter A, Specht G, et al. mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. Nucleic Acids Res. 2016;44:W64–9. https://doi.org/10.1093/nar/gkw247

41. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9. https://doi.org/10.1038/nbt.3820

42. Schönherr S, Weissensteiner H, Kronenberg F, Forer L. Haplogrep 3 - an interactive haplogroup classification and analysis platform. Nucleic Acids Res. 2023;51:W263–8. https://doi.org/10.1093/nar/gkad284

43. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res. 2020;48:D941–7. https://doi.org/10.1093/nar/gkz836

44. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26:2867–73. https://doi.org/10.1093/bioinformatics/btq559

45. Rafehi H, Fearnley LG, Read J, Snell P, Davies KC, Scott L, et al. A prospective trial comparing programmable targeted long-read sequencing and short-read genome sequencing for genetic diagnosis of cerebellar ataxia. Genome Res. 2025;35:769–85. https://doi.org/10.1101/gr.279634.124

46. Parakatselaki M-E, Ladoukakis ED. mtDNA Heteroplasmy: Origin, Detection, Significance, and Evolutionary Consequences. Life. 2021;11:633. https://doi.org/10.3390/life11070633

47. Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods. Nature Publishing Group; 2015;12:623–30. https://doi.org/10.1038/nmeth.3407

48. Castellana S, Fusilli C, Mazzoccoli G, Biagini T, Capocefalo D, Carella M, et al. High-confidence assessment of functional impact of human mitochondrial non-synonymous genome variations by APOGEE. PLOS Comput Biol. Public Library of Science; 2017;13:e1005628. https://doi.org/10.1371/journal.pcbi.1005628

49. Lake NJ, Ma K, Liu W, Battle SL, Laricchia KM, Tiao G, et al. Quantifying constraint in the human mitochondrial genome. Nature. Nature Publishing Group; 2024;635:390–7. https://doi.org/10.1038/s41586-024-08048-x

50. Wang L, Han J, Fearnley LG, Milton M, Rafehi H, Reid J, et al. Peripheral immune cell abundance differences link blood mitochondrial DNA copy number and Parkinson's disease. Npj Park Dis. Nature Publishing Group; 2024;10:219. https://doi.org/10.1038/s41531-024-00831-x

51. Trinh J, Hicks AA, König IR, Delcambre S, Lüth T, Schaake S, et al. Mitochondrial DNA heteroplasmy distinguishes disease manifestation in PINK1/PRKN-linked Parkinson's disease. Brain. 2023;146:2753–65. https://doi.org/10.1093/brain/awac464

52. Pasca S, Hong YS, Shi W, Puiu D, Lake NJ, Lek M, et al. Mitochondrial heteroplasmy is a risk factor for the development of chronic lymphocytic leukemia. Nat Commun. Nature Publishing Group; 2026;17:2898. https://doi.org/10.1038/s41467-026-69861-8

53. Hong YS, Pasca S, Shi W, Puiu D, Lake NJ, Lek M, et al. Mitochondrial heteroplasmy improves risk prediction for myeloid neoplasms. Nat Commun. Nature Publishing Group; 2024;15:10133. https://doi.org/10.1038/s41467-024-54443-3

54. El-Hattab AW, Almannai M, Scaglia F. MELAS. In: Adam MP, Bick S, Mirzaa GM, Wallace SE, Amemiya A, editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cited 2026 Sept 28]. http://www.ncbi.nlm.nih.gov/books/NBK1233/. Accessed 28 Sept 2026

55. Liao N-Y, Lai K-L, Liao Y-C, Hsiao C-T, Lee Y-C. Identification of m.3243A>G mitochondrial DNA mutation in patients with cerebellar ataxia. J Formos Med Assoc. 2023;122:1028–34. https://doi.org/10.1016/j.jfma.2023.05.031

56. Ball M, Thorburn DR, Rahman S. Mitochondrial DNA-Associated Leigh Syndrome Spectrum. In: Adam MP, Bick S, Mirzaa GM, Pagon RA, Wallace SE, Amemiya A, editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993 [cited 2026 Jan 2]. http://www.ncbi.nlm.nih.gov/books/NBK1173/. Accessed 2 Jan 2026

57. Nolte D, Kang J-S, Hofmann A, Schwaab E, Krämer HH, Müller U. Mutations in MT-ATP6 are a frequent cause of adult-onset spinocerebellar ataxia. J Neurol. 2021;268:4866–73. https://doi.org/10.1007/s00415-021-10607-5

58. Slapnik B, Šket R, Črepinšek K, Tesovnik T, Bizjan BJ, Kovač J. The quality and detection limits of mitochondrial heteroplasmy by long read nanopore sequencing. Sci Rep. Nature Publishing Group; 2024;14:26778. https://doi.org/10.1038/s41598-024-78270-0

59. Lüth T, Schaake S, Grünewald A, May P, Trinh J, Weissensteiner H. Benchmarking Low-Frequency Variant Calling With Long-Read Data on Mitochondrial DNA. Front Genet [Internet]. Frontiers; 2022 [cited 2026 Sept 28];13. https://doi.org/10.3389/fgene.2022.887644

Funding

This work was supported in part by the Australian Government National Health and Medical Research Council (NHMRC) grant (GNT2046073) to LW; NHMRC grants (GNT2028681 and MRF2024989) to PJL and MB; and a WEHI Undergraduate Vacation Scholarship to KC. MB was also supported by an NHMRC Investigator Grant (GNT1195236). This work was made possible through the Victorian State Government Operational Infrastructure Support Program and the Australian Government NHMRC IRIISS.

Data availability

Access to the 1000 Genomes Project data is available through the public '1000G-high-coverage-2019' workspace in Terra (https://app.terra.bio/#workspaces/anvil-datastorage/1000G-high-coverage-2019). Individual-level sequencing data from ataxia cohort participants who consented to controlled-access sharing are deposited in the European Genome-phenome Archive (EGA) (https://ega-archive.org/) under accession EGAS50000000573 (dataset EGAD50000000815), with access managed by the MCRI Data Access Committee. Data for the remaining participants cannot be shared publicly because participant consent does not extend to repository deposition; de-identified data are available from the authors on reasonable request, subject to approval by The Royal Children's Hospital Human Research Ethics Committee.

The MARVel pipeline, including all analysis scripts used in this study, is publicly accessible on GitHub (https://github.com/bahlolab/MARVel), with study-specific scripts available at https://github.com/bahlolab/MARVel/tree/main/Scripts. The mtSwirl HPC pipeline is publicly

accessible on GitHub (https://github.com/bahlolab/mtSwirl_HPC). Our patched fork of MitoHPC, which includes the fix described above, is available at https://github.com/bahlolab/MitoHPC.

Author contributions

Conceptualization: M.B., L.W.; Investigation: K.C., V.E.J., L.W.; Pipeline development: M.M., L.W.; Supervision: L.W., M.B.; Data and resources: H.R., P.J.L., M.B.; Writing – original draft: K.C., L.W.; Writing – review & editing: K.C., M.M., V.E.J., H.R., P.S., K.C.D, A.E., P.J.L., M.B., L.W.

Correspondence and requests for materials should be addressed to Longfei Wang.

Ethics declarations

Ethics approval and consent to participate

The ataxia cohort study was approved by the Royal Children's Hospital Human Research Ethics Committee (HREC #28097) and the Walter and Eliza Hall Institute of Medical Research (WEHI) Human Research Ethics Committee (HREC 18/06). Written informed consent for research genetic testing was obtained from all participants in the ataxia cohort or, for participants under 18 years of age, from a parent or legal guardian. The 1000 Genomes Project data are publicly available and were generated with informed consent under the original project's ethics approvals; no additional consent was required for their use in this study. All components of this study were conducted in accordance with the principles embodied within the Declaration of Helsinki.

Consent for publication

Not applicable

Competing interests

The authors declare no conflicts of interest.

Supplementary Information

Table file 1.xlsx: Manuscript tables.

XLSX file (.xlsx) containing the tables presented in the manuscript, provided as separate worksheets.

Additional file 1.docx: Supplementary figures.

DOCX file (.docx) containing the supplementary figures and corresponding figure legends.

Additional file 2.xlsx: Supplementary tables.

XLSX file (.xlsx) containing the supplementary tables, provided as separate worksheets.

Supplementary Files

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

- Tablefile1.xlsx
- Additionalfile2.xlsx
- Additionalfile1.docx