Title:

NanoDel: Identification of large-scale mitochondrial DNA deletions using long-read sequencing

Fearn, C. 1, Poulton, J. 2, Fratter C. 3, Oliva, C. 4, Griguer, C. 4, Baldock, R. A. 1, Robson, S.C. 1,5, and McGeehan, R.E. *1.

1. School of Medicine, Pharmacy and Biomedical Sciences, Institute of Life Sciences and Healthcare, University of Portsmouth, UK; 2. Nuffield Dept of women's reproductive health, University of Oxford, UK; 3. Oxford Genetics Laboratories, Oxford University Hospitals NHS Foundation Trust, UK; 4. Free Radical and Radiation Biology Program, University of Iowa, USA, 5. Centre for Enzyme Innovation, Institute of Life Sciences and Healthcare, University of Portsmouth, UK. *Corresponding author.

Abstract:

Motivation: Traditional methods for detecting large-scale mitochondrial DNA (mtDNA) deletions (LSMDs) in cells present challenges, i.e. *a priori* information, high DNA inputs, poor sensitivity and are not always quantitative. Mitigation can be achieved through high throughput DNA sequencing using e.g. Illumina and Oxford Nanopore Technologies (ONT), in combination with LSMD breakpoint identification and quantification using bioinformatic tools. Splice-aware RNA alignment tools increase the sensitivity for detecting LSMD breakpoints compared with DNA aligners. Long-read sequencing (LRS) also offers potential advantages over short read sequencing, e.g. greater read lengths and capturing variants on

single reads. No existing pipelines capture the benefits of both a splice-aware alignment tool and LRS.

Results: We developed "NanoDel", a LRS pipeline, to sensitively and accurately detect cellular LSMDs. Using artificial datasets, "NanoDel" was more sensitive and accurate than other pipelines. In samples diagnosed with mitochondrial disease, it identified both known and previously uncharacterised (including mixtures) of LSMDs, without *a priori* information. LSMD breakpoints were found in *mt-co1, mt-cyb, mt-nd6* and *mt-nd5* genes. Analysis of selected LSMDs revealed proximity to repeat and putative G-quadruplex motifs, and occurrence in a range of healthy and pathological tissues, indicating potential for a shared vulnerability landscape in mtDNA, shaped by sequence motifs and structural constraints. "NanoDel" combined with one-amplicon, not two-amplicon, LR-PCR offers a robust strategy with clinical application for detecting LSMDs across a variety of cell/tissue samples, and it's application across a broader range of samples, will yield new mechanistic insights into LSMD formation, and further our understanding of mtDNA instability.

Availability and implementation: NanoDel is available at https://github.com/uopbioinformatics/NanoDel and raw read data are available through the NCBI Sequence Read Archive (SRA) under BioProject accession code PRJNA1369153 (https://www.ncbi.nlm.nih.gov/bioproject/1369153).

Contact: Rhiannon.mcgeehan@port.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

Keywords:

MtDNA, mitochondria, deletions, long-read, sequencing, disease

46

47 **Introduction:**

48 Mitochondria are cellular organelles that have their own double-stranded, supercoiled genome;
49 mitochondrial DNA (mtDNA). MtDNA has 38 loci, including 13 protein coding genes, 2
50 rRNAs, 22 tRNAs and the control region (D-loop). The D-loop is integral for producing the
51 multi-protein mitochondrial respiratory chain (MRC) complexes (I-IV), which along with
52 complex V forms part of the oxidative phosphorylation (OXPHOS) system. This system, which
53 also includes approximately 70 nuclear-encoded subunits, is central to ATP generation and
54 cellular metabolism (reviewed in (Vafai and Mootha 2012)). Direct and/or indirect alterations
55 in the mitochondrial genome can impact the function of the OXPHOS system, and therefore
56 cellular energy generation, contributing to disease (for examples, see Table S1). Large-scale
57 mtDNA deletions (LSMDs), which are direct alterations to mtDNA, result in genes crucial for
58 cellular energy production via OXPHOS being removed and, therefore, primary mitochondrial
59 disease.

60 Several traditional methods, such as Southern blotting (Moraes et al. 1989)(Santorelli et al.
61 1996), PCR (Schon et al. 1989) and Sanger sequencing (Johns et al. 1989)(Poulton, Deadman,
62 and Gardiner 1989) (Johns et al. 1989) (Mita et al. 1990), have been used historically to detect
63 LSMDs. However, the accurate breakpoint and heteroplasmy determinations using these
64 techniques is challenging, requiring *a priori* breakpoint information, and high DNA inputs.
65 Often such methods also lack sensitivity, are not completely quantitative, and can be labour
66 intensive. One way to identify the breakpoints where multiple unknown LSMDs are present
67 would be to plasmid clone and select all molecules prior to Sanger sequencing, however this is
68 even more labour intensive (Schon et al. 1989).

A useful and sensitive alternative for identifying unknown LSMDs is by LR-PCR to amplify whole mtDNAs, followed by high-throughput DNA sequencing using platforms from e.g., Illumina (Hjelm et al. 2019) or Oxford Nanopore Technologies (ONT), and subsequent breakpoint identification and quantification using bioinformatic tools. Indeed, in 2019, Hjelm *et al.* published a pipeline that combines the LR-PCR of whole mtDNAs, as single amplicons, followed by Illumina sequencing, and a splice-aware RNA alignment tool for breakpoint identification and quantification. This demonstrated increased the sensitivity in detecting true positives compared to traditional DNA alignment tools (e.g. Burrows-Wheeler aligner) by around 60% (Hjelm et al. 2019). While examples of LMSDs detection by ONT sequencing exist, to date these have not used LR-PCR or a splice-aware aligner (Vandiver et al. 2023)(Frascarelli et al. 2023). However, employing LR-PCR and ONT sequencing for LSMD detection could offer additional advantages over short read (e.g., Illumina) sequencing, including less amplification of nuclear mitochondrial DNA segments (NUMTs) and longer individual read lengths (>1kb vs. <300bp) that span breakpoints, meaning a single read can represent an entire variant. This allows for the possibility of analysing variants, including SNVs, in tandem on the same read.

With improvements to sequencing chemistry and base-calling models now achieving quality scores above Q20, ONT is an attractive alternative to Illumina for the accurate breakpoint identification and quantification of LSMDs. However, the lack of an ONT pipeline comparable to that of (Hjelm et al. 2019), prompted us to develop “NanoDel”. This pipeline endeavours to amplify whole mtDNAs as either one or two amplicons by LR-PCR, followed by ONT sequencing, and a bespoke combination of bioinformatic tools for accurate breakpoint identification and quantification in cells/tissues.

In this study we first compared the ability of four pipelines to determine LSMDs breakpoints at different heteroplasmy levels using two artificial data sets. The four pipelines included two

that use long-read sequences either in combination with a splice-aware RNA aligner ("NanoDel", Pipeline 1, this study) or a DNA aligner ("Wood ONT", Pipeline 3, (Wood et al. 2019)), and two that use short read Illumina sequences either in combination with a splice-aware RNA aligner ("Splice-Break", Pipeline 2, (Hjelm et al. 2019)) or a DNA aligner ("Wood Illumina", Pipeline 4, (Wood et al. 2019)). Second, we directly compared the two different ONT-pipelines ("NanoDel" and "Wood ONT") to understand the impact of the selected alignment strategy, and the two LR-PCR approaches (one and two-amplicon) to determine known LSMDs in Kearns-Sayre syndrome (KSS) tissues. Third, we compared the LSMDs information generated by "NanoDel" and one-amplicon PCR against LSMDs information obtained diagnostically (using traditional methods) for additional patients with mitochondrial disease. This provides a comprehensive benchmarking of the power of NanoDel for exploring unknown LSMDs from clinical samples, allowing for fine-tuned understanding of the impact of mtDNA instability on disease.

Methods and Materials:

Patient-derived tissues, cell cultures and DNA extraction

Three total DNA samples extracted from the kidney, cerebellum, and muscle of a KSS patient of known LSMD composition determined by Southern blotting (Table S2) were provided by Professor Joanna Poulton at the Nuffield Department of Women's and Reproductive health (Patient 1 in (Poulton et al. 1993)(Poulton, Deadman, and Gardiner 1989)(Poulton et al. 1995)), University of Oxford. Additionally, DNA samples from muscle (muscle_RNASEH1, muscle_TWNK and muscle_TYMP) and cultured cells (cells_nd and fibroblasts_nd) were provided by Professor Joanna Poulton and by Carl Fratter at the Oxford Genetics Laboratories (OGL). These DNA samples were prepared and de-identified prior to transfer to our laboratory.

118 The original tissues were surplus to the quantity required for diagnosis, with the KSS patient
119 tissues donated for research on the 1 January 1990 with permission from the patient's family
120 and surgeon, and the remaining patient tissue donated to research between 19 May 2005 and
121 13 December 2024 with written informed consent from all participants and with the approval
122 of the United Kingdom National Research Ethics Service (South Central-Berkshire, #REC
123 20/04, IRAS 162181). Previous observations, i.e., LSMD composition determined by a range
124 of different techniques following diagnosis at the OGL, were made available for these samples
125 following sequencing analysis to prevent bias (Table S2). Details of cell lines used as mtDNA
126 positive and negative amplification controls are provided in the Supplementary materials and
127 methods. Total DNA was extracted from all cell culture samples using the QIAmp DNA mini
128 kit as per the manufacturer's instructions (Qiagen, 51304).

129 Whole mitochondrial genome sequence generation

130 Whole mtDNAs were amplified from patient DNA in either two overlapping amplicons (Lloyd
131 et al. 2012)(Lloyd et al. 2015)(Keatley et al. 2019) or as a single amplicon (Hjelm et al. 2019)
132 by LR-PCR using human mtDNA-specific primers from total DNA (Table S3). Briefly,
133 Nanopore libraries were prepared from either two or single LR-PCR amplicons using the
134 Native Barcoding Kit 96 V12 library kit (ONT, NBD112.96), barcoded using the NBD112.96
135 kit and the NEB Blunt T/A Ligase MasterMix (NEB, M0367S), multiplexed and then cleaned
136 using AMPure XP beads (Beckman Coulter, A63880). Adapter sequences were ligated to the
137 multiplexed DNA sample using the NEBNext Quick Ligation module (New England
138 Bioscience, E6056S). The prepared, washed and quantified DNA library was then loaded to an
139 R10.4 flow cell and sequenced for 72 hours on an Oxford Nanopore Technologies GridION
140 X5. Full details of the one- and two-amplicon LR-PCR, library preparation and sequencing are
141 provided in the Supplementary materials and methods.

142 Bioinformatics, artificial data generation, alignment and LSMD analyses

143 First and briefly, two styles of artificial FASTQ format raw sequence data files were generated
144 using VISOR (v1.1, (Bolognini et al. 2020) in: 1) an ONT style with a longer average read
145 length of ~8kb and a "Nanopore 2020" style error model, and 2) an Illumina style with 250bp
146 length reads and an "Illumina" style error model. Each style contained either controlled
147 quantities of reads designed to emulate a single deletion (mtDNA 4977) at different levels; i.e.,
148 75%, 50%, 25%, 5%, 1% and 0.25% of reads (artificial data 1) or multiple deletions (m.8471-
149 13449del, m.6620_8440del, m.3451_5500del and m.12538_14572del) at a fixed quantity
150 (10%) of reads (artificial data 2). These files were then used to evaluate and compare the
151 efficacy of the four separate alignment and deletion calling strategies outlined above (Pipelines
152 1-4). Second, raw reads from the patient mtDNA LR-PCR libraries (one- and/or two-amplicon)
153 were base-called and demultiplexed (including barcode trimming) using Guppy (v6.2.7 and
154 v6.3.9) and transferred to a local high-performance computing cluster running on Ubuntu
155 (v20.04.5). Base-called reads were initially filtered to remove reads of <20bp in length or a Q-
156 score <10 using NanoFilt (v2.8.0, (De Coster et al. 2018) prior to processing through the two
157 best performing pipelines: "NanoDel" (Pipeline 1) and "Wood ONT" (Pipeline 3). In this study
158 all reads (artificial and patient) were aligned to the rCRS reference sequence (NC_012920.1,
159 (Bolognini et al. 2020)(Andrews et al. 1999)). Full details of the artificial data generation,
160 alignment and deletion calling strategies, including all read quality control, software,
161 benchmarking and filtering are given in the Supplementary materials and methods.

Fig 1. Flow diagram of large-scale mtDNA deletions using Pipeline 1 ("NanoDel") developed for this study. A) Raw read processing. Raw reads were base-called following barcode trimming and demultiplexing at run-time in MinKnow using Guppy (v6.2.7 or v6.3.9) and transferred from the GridION to a local high-performance computing cluster (HPC) running on Ubuntu (v20.4.5). Where the version 14 chemistry was used, samples were aligned to the bacteriophage lambda reference genome (NC_001416.1) using minimap2 (v2.21) to remove reads mapping to the spike-in DNA control. All mapped reads were then subsequently removed from their constituent FASTQ files. Individual FASTQ files were then filtered to remove all reads <20bp and/or with a Q-score <10 using NanoFilt (v2.8) post-barcode trimming, but before being processed through "NanoDel". B) "NanoDel" was constructed and run on the HPC cluster with Ubuntu (v20.4.5). First, filtered reads were aligned to the revised Cambridge reference sequence (rCRS, NC_012920.1, Andrews et al., 1999), using Minimap2 (v2.21), with splice

175 *alignment settings enabled (-ax splice), but preferences for canonical splice junctions disabled*
176 *(-u no). These outputs are piped directly to SAMtools, where the resulting data were converted*
177 *to BAM format using SAMtools view (-S and -b options for SAM format input and BAM output)*
178 *and sorted using the SAMtools sort function. Splice junction locations were then extracted from*
179 *the BAM files using the JWR_Checker.py script from NanoSplicer (v1.0), with output CSV*
180 *format options enabled (–output_csv). Simultaneously, the per-base depth was calculated by*
181 *the SAMtools depth function for all positions regardless of coverage (option -aa) and later used*
182 *for heteroplasmy calculation. Custom R scripts were then available as options within the*
183 *pipeline (filt or no_filt) that either reformatted the data and removed any deletion junction calls*
184 *with 5'- or 3'-breakpoints within 500bp of a primer binding site, or not. This is because (Hjelm*
185 *et al. 2019) demonstrated an increase in false positive breakpoint calls within this distance of*
186 *primer boundaries. The remaining breakpoint junctions were then clustered using a custom*
187 *python script that employed the sci-kit learn package (v1.2). A density-based clustering*
188 *approach (DBScan) was used to group all LSMD calls with both 5'- and 3'-breakpoints that*
189 *were within 10bp of other 5'- and 3'-breakpoints. Average 5'- and 3'-locations were then taken*
190 *from each cluster and reported with the number of reads generating that cluster for each*
191 *sample and output in CSV format, with all points not contributing to a cluster being subset as*
192 *'noise points'. C) Finally, a mean coverage benchmark was calculated from the depth output*
193 *between positions m.500_1000 (according to the rCRS reference sequence) and used in the*
194 *calculation of the relative read % (a proxy for heteroplasmy) for each LSMD using a custom*
195 *R-script. As such the number of reads containing each specific LSMD was expressed as a*
196 *percentage of the coverage benchmark for each respective sample.*

197 Repeat and G4 sequence, previous tissue/ disease association and statistical analyses

198 Sequences flanking the 5'- and 3'-breakpoints of the mitochondrial disease patient LSMDs
199 occurring at ≥50% of reads detected in the filtered "NanoDel" data (Pipeline 1) were analysed

for repeat sequences and their potential to form G-quadruplexes (G4s) using QGRS mapper (Kikin, D'Antonio, and Bagga 2006)) and AlphaFold 3 (AF3)(Abramson et al. 2024)).

Previous tissue and/or disease associations of the LSMDs, both "NanoDel" (Pipeline 1) and adjusted positions (Table S6), and 10bp flanking either side were investigated using the Mitobreak database, which contains 1369 human mtDNA deletions previously described in the literature (Damas et al. 2014).

Statistical analyses of post-alignment read metrics were conducted using GraphPad Prism (v9.4.1). The normality of each metric for each pipeline was assessed using the Shapiro-Wilk test (Shapiro and Wilk 1965). Where assumptions of normality were met, a Brown-Forsythe and Welch ANOVA was employed to model differences between the approaches. If the assumption of normality was violated, the non-parametric Kruskal-Wallis test (Kruskal and Wallis 1952) was used instead. Further post-hoc pairwise comparisons were made using either Welch's t-test or the non-parametric Mann-Whitney test, respectively. Full details on the repeat and G4 analyses are given in the Supplementary materials and methods.

Results:

The pipeline developed for this study, "NanoDel" (Pipeline 1), was tested against three other pipelines using two artificial datasets to compare the performance in detecting the percentage of reads within which each LSMD occurs, and its location. NanoDel offered the best balance between accuracy, sensitivity and specificity when identifying the heteroplasmy and breakpoint location, both as a single deletion (mtDNA4977) in artificial data set 1, and as mixed deletions in artificial data set 2 (Fig 2-3, and Supplementary Results).

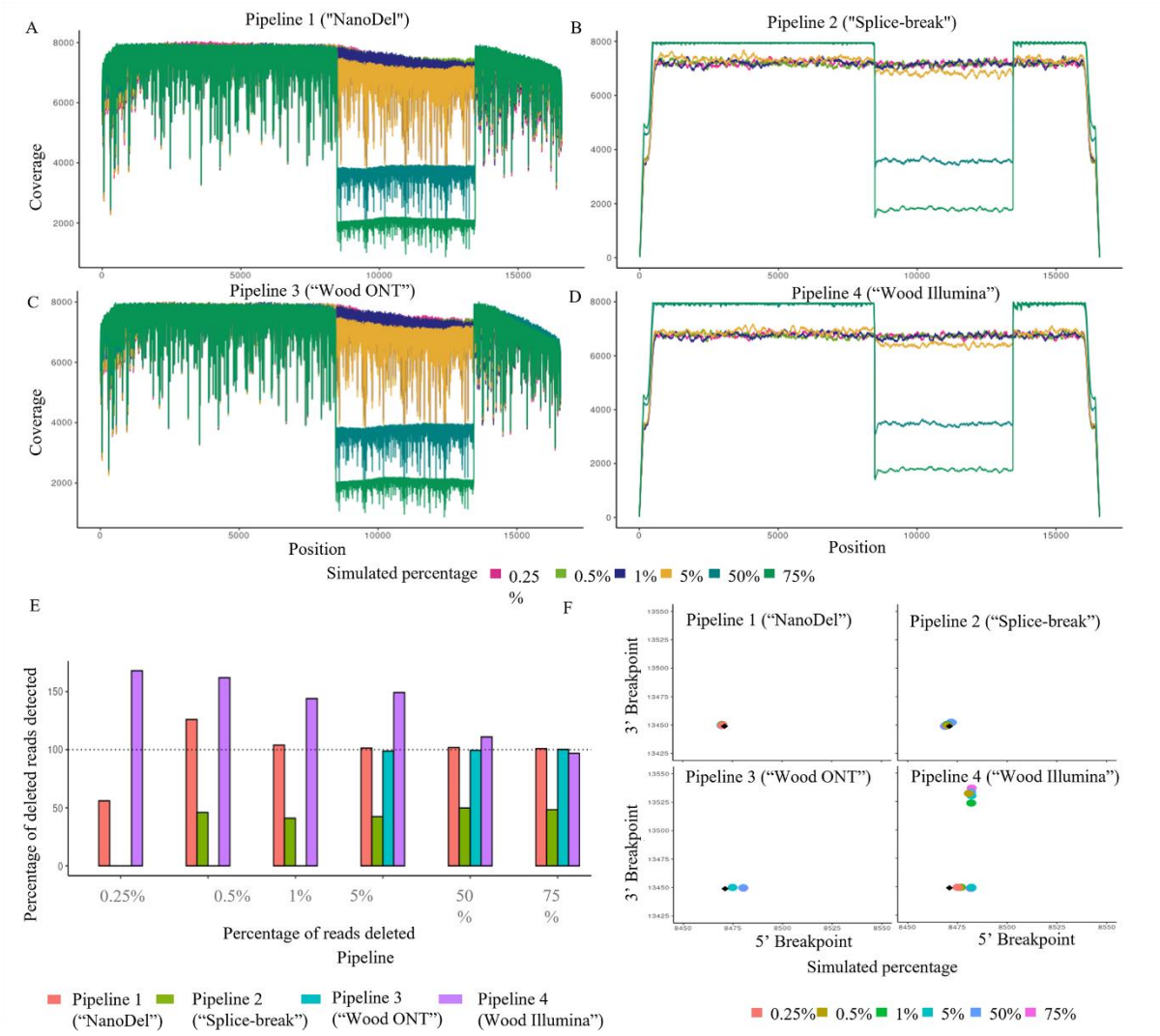
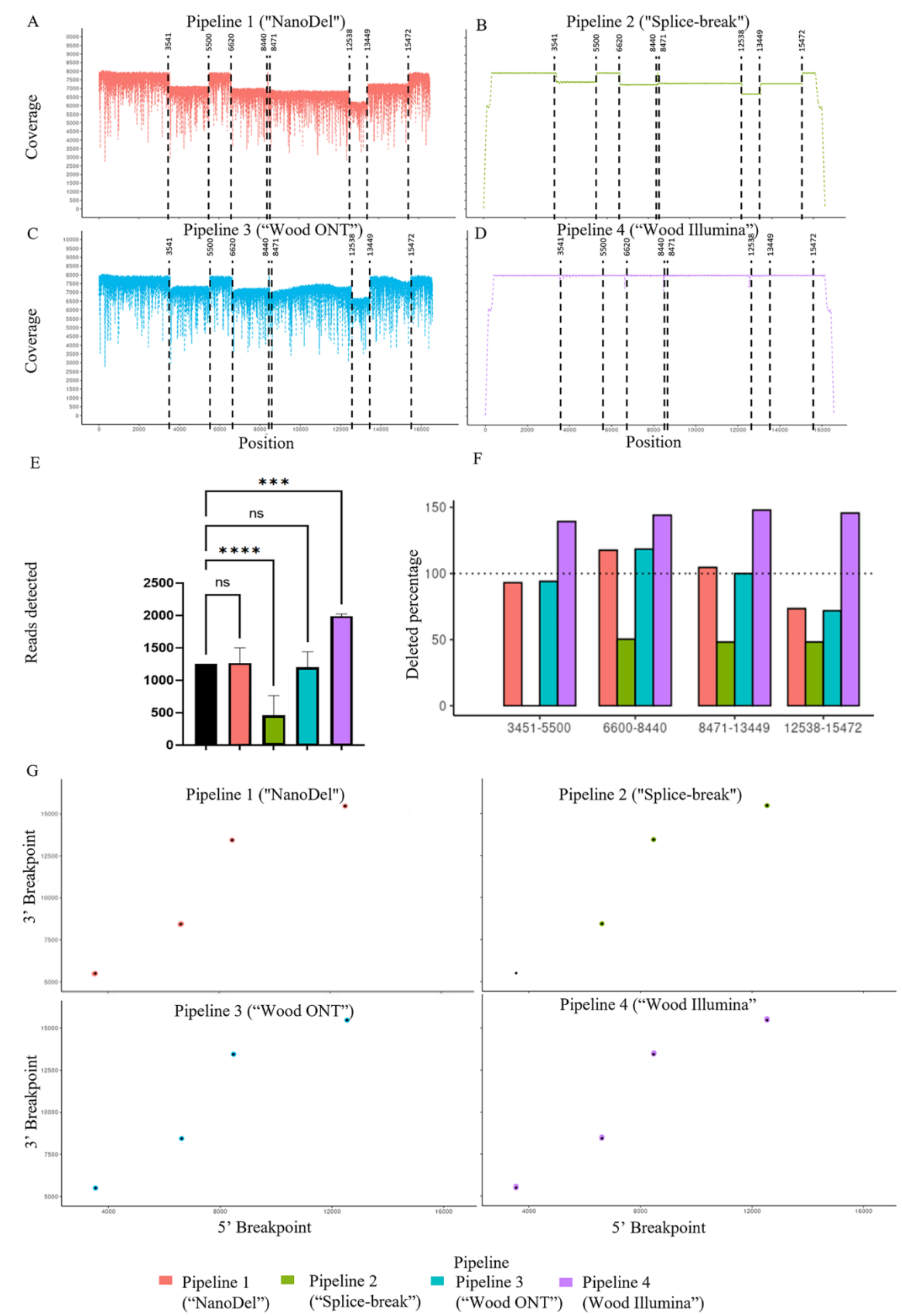


Fig 2. Analysis of artificial data set 1 that contains a mixture of the mtDNA4977 (m. 8471_13449del) and wild type mtDNA at different ratios. A-D) The per-position levels of coverage generated by Pipeline 1 ("NanoDel"), Pipeline 2 ("Splice-break"), Pipeline 3 ("Wood ONT"), and Pipeline 4 ("Wood Illumina"), respectively. E) The percentage of the expected number (7500 (75%), 5000 (50%), 500 (5%), 100 (1%), 50 (0.5%) and 25 (0.25%)) of deleted reads detected out of 10,000 identified by each pipeline. F) Breakpoint mapping locations identified at each deleted read% level (75%-0.25%) in relation to the true breakpoint locations (highlighted in black).

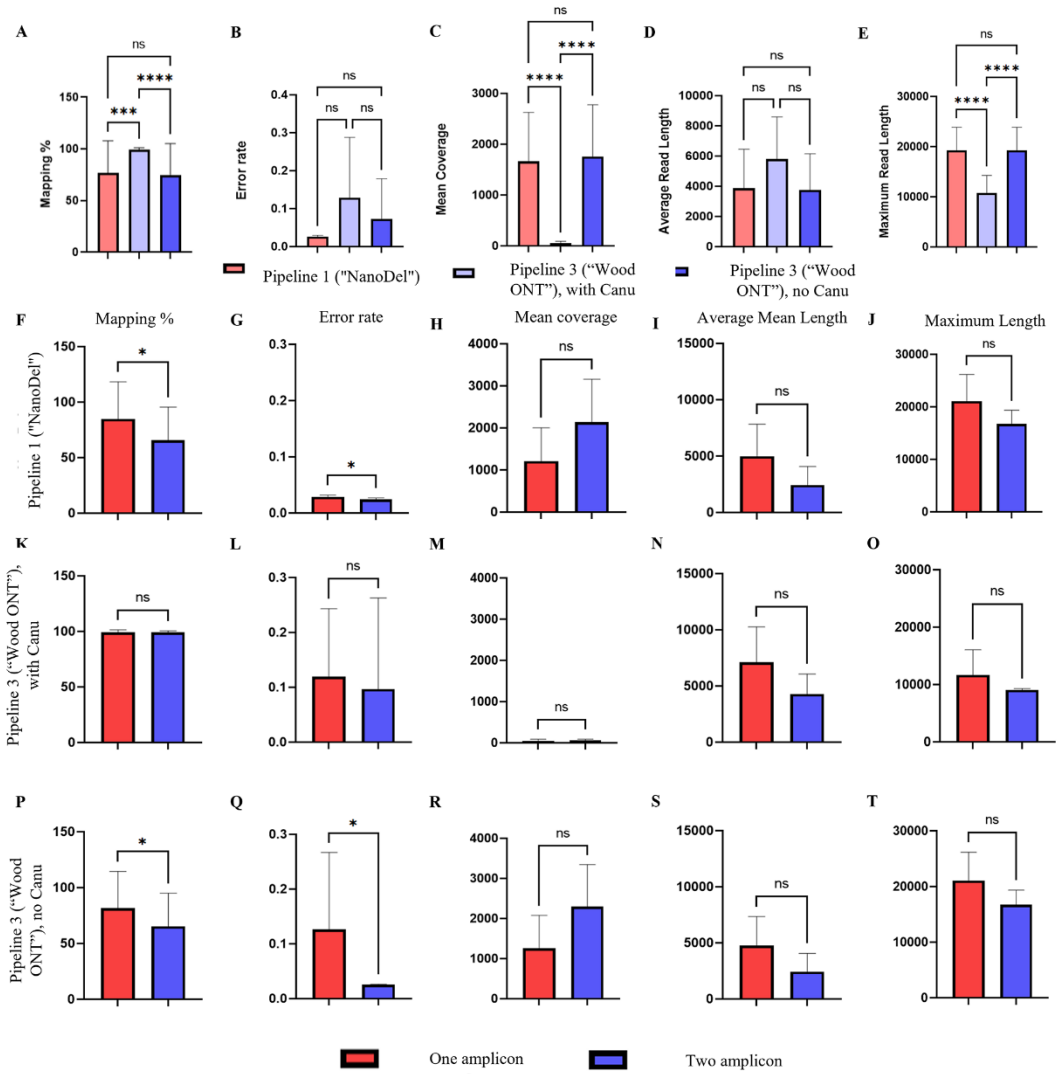

A
Pipeline 1 ("NanoDel")
B
Pipeline 2 ("Splice-break")
C
Pipeline 3 ("Wood ONT")
D
Pipeline 4 ("Wood Illumina")
Coverage
Position
3541
5500
6620
8440
8471
12538
13449
15472
E
Reads detected
***
ns
****
ns
F
Deleted percentage
3451-5500
6600-8440
8471-13449
12538-15472
G
Pipeline 1 ("NanoDel")
Pipeline 2 ("Splice-break")
Pipeline 3 ("Wood ONT")
Pipeline 4 ("Wood Illumina")
3' Breakpoint
5' Breakpoint
Pipeline
Pipeline 1 ("NanoDel")
Pipeline 2 ("Splice-break")
Pipeline 3 ("Wood ONT")
Pipeline 4 (Wood Illumina")

231

232 *Fig 3. Analysis of artificial data set 2 that contains a mixture of deletions (i.e., between*
233 *positions m. 8471_13449del, m. 6620_8440del, m. 3451_5500del and m. 2538_14572del), set*
234 *at a read level of 10% and 90% wild type mtDNA. A-D) The per position levels of coverage*
235 *generated by Pipeline 1("NanoDel"), Pipeline 2 ("Splice-break"), Pipeline 3 ("Wood ONT"),*
236 *and Pipeline 4 ("Wood Illumina"), respectively. Breakpoint locations are denoted by black*
237 *dashed lines. E) The average number of deleted reads detected by each pipeline; each one is*
238 *simulated out of 1250 reads. The percentage deleted reads detected by deletion reads (so reads*
239 *detected out of 1250*100). Breakpoint mapping locations of each deletion identified in relation*
240 *to the true breakpoint locations (highlighted in black). * = P<0.05, ** = P<0.01, *** =*
241 *P<0.001, and **** = P<0.0001.*

242 Due to the performance of "NanoDel" (Pipeline 1) in detecting LSMDs within the artificial
243 data, it was then employed to detect LSMDs and determine their breakpoints in two small
244 mitochondrial disease patient cohorts (Table S2) following both one- and two-amplicon LR-
245 PCR. MtDNA breakpoints were also determined following LR-PCR amplification in these
246 samples using "Wood ONT" (Pipeline 3) for comparison. LSMDs were observed in clinically
247 diagnosed mitochondrial disease samples on agarose gels following different LR-PCR
248 approaches, revealing single or multiple deletion types (Figure S1, and Supplementary results).

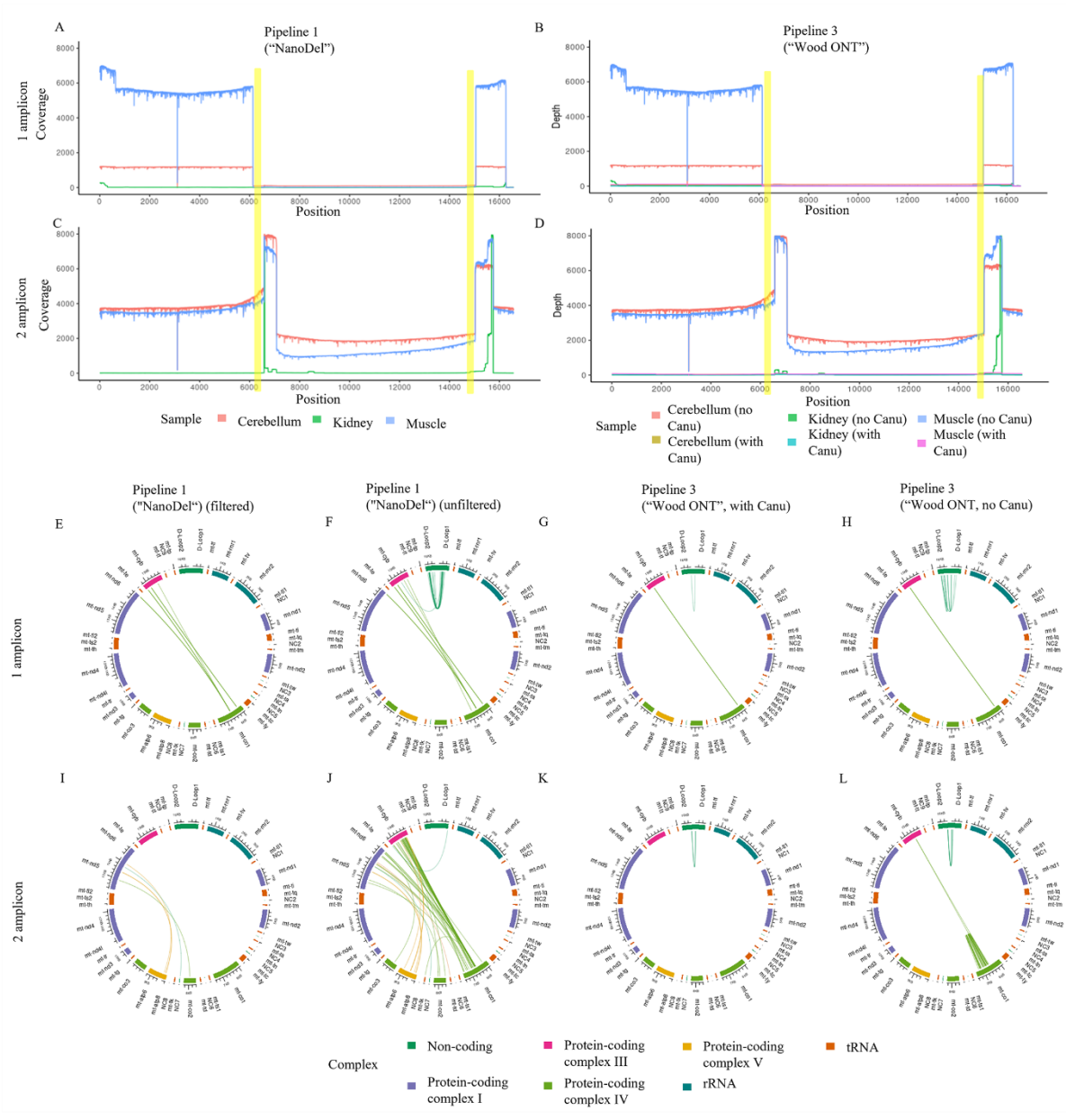
249 Post-alignment, "NanoDel" yielded lower error rates, higher mean coverage and higher
250 maximum read lengths compared to "Wood ONT" (Pipeline 3). Furthermore, "NanoDel"
251 (Pipeline 1) and "Wood ONT" (Pipeline 3) had increased mapping rate following one- rather
252 than two-amplicon LR-PCR (Fig 4, Table S4 A-B, and Supplementary results).



253

254 *Fig 4. A comparison of post-alignment quality metrics: A, F, K and P) The percentage of reads*
255 *mapped to the reference genome, B, G, L and Q) Error rate of mapped reads, C, H, M and R)*
256 *The average mean coverage, D, I, N and S) Average read length, and E, J, O and T) Maximum*
257 *read length following the alignment of one and two-amplicon LR-PCR reads, analysed in*
258 *combination (A-E) or separately (F-T), by Pipeline 1 ("NanoDel"), Pipeline 3 ("Wood ONT",*
259 *with Canu) and Pipeline 3 (without Canu) generated from patients with clinically diagnosed*
260 *mitochondrial diseases. * = P<0.05, ** = P<0.01, *** = P<0.001, and **** = P<0.0001.*

In addition to the differences between the post-alignment quality metrics, "NanoDel" (Pipeline 1) and "Wood ONT" (Pipeline 3) in combination with one-amplicon, rather than two-amplicon, LR-PCR best validated a known LSMD (m.6130_15055del) in KSS patient tissues (Fig 5, Table S5A, and Supplementary results). In the two-amplicon LR-PCR data, deletion calling was impacted across amplicon boundaries, resulting in an impact to the LSMDs called from both the quality of the DNA and filtering applied to the variant calling.



268 *Fig 5. Analysis of coverage and deletion boundaries in different tissues of a KSS patient. A-D)*
269 *The levels of coverage displayed as read depth at each position for each sample amplified as*
270 *either one A, B) or two C, D) amplicons, aligned with Pipeline 1 ("NanoDel", A, C) or Pipeline*
271 *3 ("Wood ONT", B, D). E-L). Circos plots showing the 5'- and 3'-breakpoints of deletions*
272 *following one E-H) and two I-L) amplicon LR-PCR aligned with Pipeline 1 ("NanoDel") post-*
273 *filtering E, I) and unfiltered F, J) and Pipeline 3 ("Wood ONT") with G, K) and without H, L)*
274 *Canu. The genomic links highlight the connected 5'- and 3'-breakpoints coloured by sample of*
275 *detection, and key complexes across the mitochondrial genome are highlighted. Links proceed*
276 *in a clockwise direction starting from the D-loop 1 region. Putative G4-structure forming*
277 *regions are highlighted in yellow.*

278 As the two-amplicon LR-PCR completely failed to detect the known deletion
279 (m.6130_15056del) in the KSS samples (and could not detect deletions across amplicon
280 boundaries), subsequent analyses of the other mitochondrial disease samples focussed on
281 "NanoDel" (Pipeline 1) and "Wood ONT" (Pipeline 3) using just the one-amplicon LR-PCR.
282 "NanoDel" (Pipeline 1) and one-amplicon LR-PCR validated known information obtained
283 using traditional methods but also yielded new insights into LSMDs in the other mitochondrial
284 disease patients, consistent with, or even in the absence of, primary diagnosis (Fig 6, Table
285 S5B, and Supplementary results).

286 Specifically, "NanoDel" (Pipeline 1) in combination with one-amplicon LR-PCR determined
287 that *mt-co1, mt-nd5, mt-nd6* and *mt-cyb* harbour breakpoints most frequently in mitochondrial
288 disease samples, and that breakpoints identified for m.6130_15055del in the KSS samples and
289 m.8469_13446del in fibroblast_nd, which occurred at ≥50% of reads in filtered "NanoDel"
290 (Pipeline 1) data lie adjacent to or within repeat sequences and putative G4 structure-forming
291 sequences, and are associated with a range of healthy and pathological tissues (Fig 6, Table S6,
292 and Supplementary results).

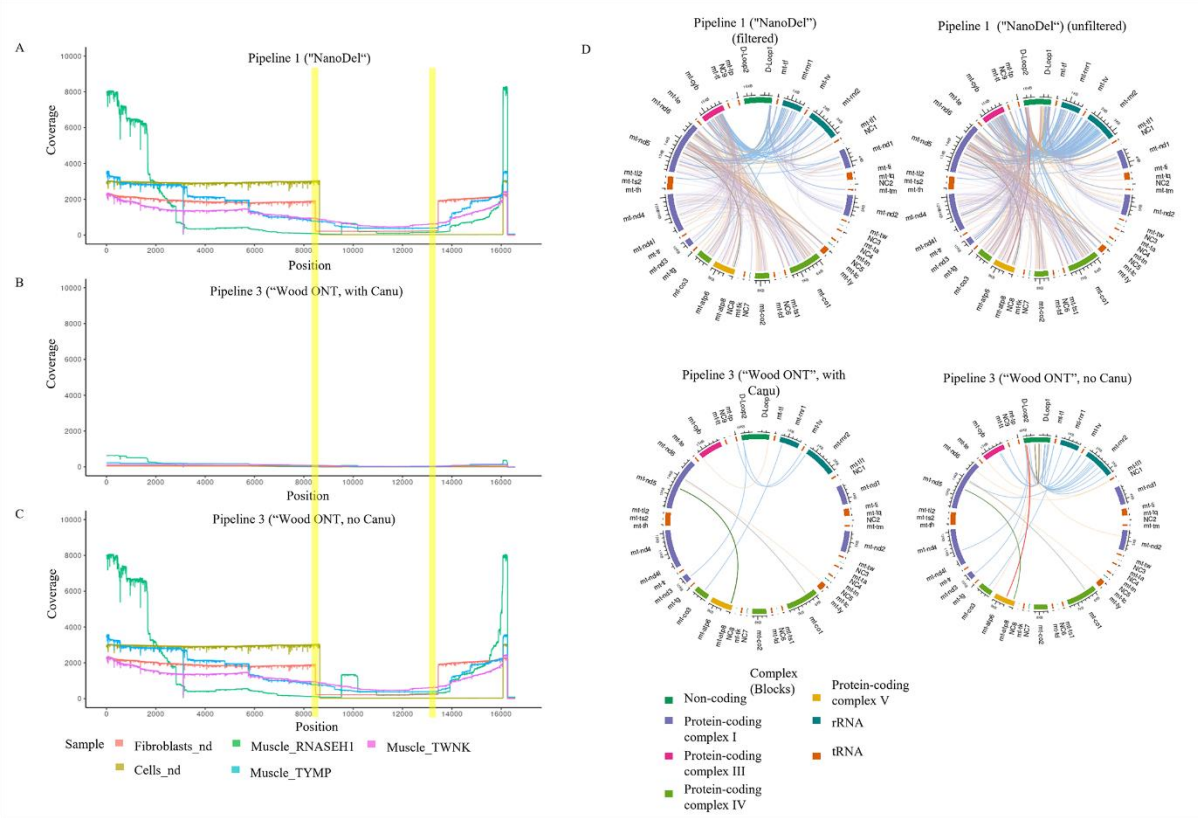


Fig 6. Analysis of coverage and deletion breakpoints in five samples clinically diagnosed with mitochondrial disease. A-C) Shows the coverage at each position for each sample generated with Pipeline ("NanoDel", A), and Pipeline 3 ("Wood ONT") with Canu, B) and without Canu C), respectively. D-G) Shows the 5'- and 3'-breakpoints of deletions identified by each pipeline in each sample as a circos plot following one-amplicon LR-PCR aligned with Pipeline 1 ("NanoDel") post-filtering D) and unfiltered E) and Pipeline 3 ("Wood ONT") with F) and without G) Canu. The genomic links highlight the connected 5' and 3' breakpoints coloured by sample of detection, and key complexes across the mitochondrial genome are highlighted. Links proceed in a clockwise direction starting from the D-loop 1 region. Putative G4-structure forming regions are highlighted in yellow.

Discussion:

Development, validation, and potential clinical application of NanoDel

307 In this study, we first developed and then evaluated the ability of four pipelines to accurately
308 detect the location and quantify the read% of LSMDs using artificial data sets. These pipelines
309 included both DNA and splice-aware RNA alignment methodologies, and were designed based
310 on either short read (Illumina) or long-read (ONT) data sets. The pipelines selected for
311 comparison against "NanoDel" (Pipeline 1, ONT splice-aware alignment, this study) were:
312 "Splice-Break" (Pipeline 2, Illumina splice-aware alignment, (Hjelm et al. 2019)), "Wood
313 ONT" (Pipeline 3, ONT DNA alignment, (Wood et al. 2019)) and "Wood Illumina" (Pipeline
314 4, Illumina DNA alignment, (Wood et al. 2019)). Splice-Break was chosen as a comparator as
315 it uses Illumina short read sequencing in combination with a non-traditional splice-aware RNA
316 alignment strategy that is not negatively affected by the presence of large-scale deletions,
317 shown previously to improve LSMDs calling accuracy (Hjelm et al. 2019). "Wood ONT"
318 (Pipeline 3) and "Wood Illumina" (Pipeline 4) were chosen as they have previously been tested
319 in a UK diagnostic setting and use either Illumina short- or ONT long-read sequencing,
320 respectively, in combination with a more traditional DNA alignment strategy (Wood et al.
321 2019).

322 We demonstrate that our pipeline, "NanoDel" (Pipeline 1), which calls LSMDs from long-read
323 data using a splice-aware alignment approach, offers improvements in accuracy and sensitivity
324 over the other publicly available tools. This is true in terms of both breakpoint location and
325 read% quantification of mtDNA4977 (m.8471_13449del, (Yusoff et al. 2020)) and mixed
326 populations of deletions, including at low levels, in artificial datasets. "NanoDel" (Pipeline 1)
327 identified the correct breakpoint locations and number of reads for all the LSMDs. "Wood
328 ONT" (Pipeline 3, DNA alignment) performed most similarly to "NanoDel" (Pipeline 1). This
329 may be in part due to two factors: their use of a shared alignment tool 'Minimap2' (albeit using
330 different settings, with "NanoDel" making use of the splice preset settings); and the fact that
331 they both analysed the same style of long-read data compared to the other pipelines analysing

332 ‘short-read style’ simulated data. In comparison, both “Splice-break” (Pipeline 2, Illumina
333 RNA alignment) and “Wood Illumina” (Pipeline 4, DNA alignment) could not identify the
334 LSMDs at the lowest 0.25 read% levels (“Splice-break”), or identified false positives (”Wood
335 Illumina”), and were the least able to accurately quantify the correct number of reads containing
336 LSMDs. Both these pipelines made use of short Illumina-style reads.

337 While these findings align with separate prior reports emphasizing the importance of using
338 splice-aware aligners in LSMDs calling (Hjelm et al. 2019) and the potential utility of ONT
339 sequencing in mitochondrial genomics (Wood et al. 2019) (Frascarelli et al. 2023)(Vandiver et
340 al. 2023) (Frascarelli et al. 2023), they also highlight the potential strength of combining splice-
341 aware alignment methodologies with long-read sequencing (i.e. “NanoDel”) in LSMDs calling.
342 Consequently, we wanted to deploy “NanoDel” (Pipeline 1), and “Wood ONT” (Pipeline 3) as
343 a comparator, to investigate LSMDs in clinical samples from patients diagnosed with
344 mitochondrial disease and varying amounts of LSMDs and compare with the information
345 obtained using traditional techniques such as Southern blotting.

346 In doing so, we found that “NanoDel”, in combination with one rather than two-amplicon LR-
347 PCR, outperformed the alternative “Wood ONT” in terms of mapping and error rate when
348 investigating LSMDs in different tissues from mitochondrial disease patients. “NanoDel” in
349 combination with one-amplicon LR-PCR was able to identify the previously observed LSMD
350 (m.6130_15056del) (Poulton et al. 1995)(Poulton et al. 1993) (Poulton, Deadman, and
351 Gardiner 1989) in KSS patient samples in a tissue specific manner. Furthermore, “NanoDel”
352 was not only able to identify previously observed LSMDs in different tissues from five other
353 patients diagnosed with mitochondrial disease, but also revealed additional, previously
354 unreported LSMDs information. This demonstrated that “NanoDel” has power to identify
355 LSMDs in different sample types (tissues/diseases), including more complex biological

samples (such as those without a known genetic cause), with significant potential in a clinical setting.

Our findings demonstrated that "NanoDel", in combination with one-amplicon LR-PCR, outperformed our two-amplicon approach. This is consistent with an original report that detailed the development and potential benefits of the one-amplicon LRPCR approach over multi-amplicon approaches, including increased discriminating power against nuclear mitochondrial DNA segments (NUMTs), more even coverage, and increased LSMD detection resolution by limiting amplicon dropout (Zhang, Cui, and Wong 2012). NUMTs are pieces of mtDNA that have been transferred to and integrated into the nuclear genome and can cause false positives during PCR amplification. Consistent with the original report, we also showed that NUMTs with "NanoDel" were not amplified in our rho-zero mtDNA depleted controls with either one- or two-amplicon LR-PCR. Further, we showed the coverage was more even in our samples, with one rather than two-amplicon LR-PCR. We found that LSMDs were only identified within individual amplicon boundaries with two- rather than one-amplicon PCR. Furthermore, the inflated number of reads within overlapping regions relative to benchmark coverage affected the read% of LSMDs. These findings suggest prior reports detailing mtDNA variants following the use of multi-amplicon PCR approaches, of which there are numerous examples including (but not restricted to) colorectal cancer (Aikhionbare et al. 2004), pancreatic cancer (Lam et al. 2012) and hypertension (Elango et al. 2011), could be incomplete and contain inaccuracies and should be interpreted with caution.

While the clinical application of ONT sequencing for mtDNA variants (including LSMD detection) has been proposed previously (Wood et al. 2019), recent improvements to the ONT chemistry now routinely generate base-calling quality scores of Q20+. This, taken together with "NanoDel"'s ability to yield a sensitive and accurate, and thus a comprehensive view of the LSMDs breakpoints/landscape without *a priori* information in individual samples (in

381 combination with one-amplicon LR-PCR), offers improvements over traditional methods in
382 this context. However, the limited of correlation in read% between "NanoDel" (and "Wood
383 ONT"), in combination with one-amplicon LR-PCR, and the heteroplasmy measures obtained
384 using traditional methods, could be due to the preferential amplification of shorter fragments
385 during PCR in our samples (Kopsidas et al. 2000). While such preferential amplification has
386 been shown to be minimised via the use of LR-PCR (Kopsidas et al. 2000), its impact could be
387 verified in our samples by determining the ratio of deleted molecules to wild type molecules
388 using digital PCR. Thus, for now, the read percentages should not be interpreted as
389 heteroplasmy rates directly. Alternative PCR-free methods, e.g., purifying
390 mitochondria/mtDNA prior to sequencing or WGS in combination with adaptive sequencing
391 (where ONT pores screen for 'expected' sequences *in silico*, and reject strands that do not meet
392 the expected requirements) may be used to increase sequencing yield of an intended target like
393 mtDNA (Martin et al. 2022), and may more accurately quantify heteroplasmy. However, these
394 methods also present challenges (e.g. Herbert et al. 2025). It can be difficult to obtain suitable
395 yields of purified mtDNA (often due to its fragility), and WGS can generate poor coverage as
396 the probes are not explicitly designed for mtDNA.

397 Mechanistic insights and broader implications

398 Using "NanoDel", we identified increased numbers of 5'-breakpoints in the gene *mt-co1* and
399 3'-breakpoints in the genes *mt-cyb, mt-nd5* and *mt-nd6* in relation to other loci, suggestive of
400 LSMD hotspots. This is consistent with the previous report investigating LSMDs in non-
401 tumorigenic brain samples that also showed 5'-breakpoints were most frequent in *mt-co1* and
402 3'-breakpoints were frequent in *mt-cyb* (with *mt-nd5* being the most affected, Hjelm et al.
403 2019). Large differences in sample number/types and sequencing and bioinformatics strategies
404 likely contributed to the differences in absolute numbers of breaks, with (Hjelm et al. 2019)
405 using short-read Illumina sequencing and their pipeline, "Splice-Break". It is possible that

406 sequence-specific features (e.g., repeat regions) and/or structural features (e.g., G4s) can
407 predispose these (and other) regions of the mtDNA to deletion events. G4s are four-stranded
408 nucleic acid structures formed by guanine-rich sequences in DNA and RNA. They are non-
409 canonical structures, meaning they deviate from the typical double-helix structure of B-form
410 DNA. G4s are formed by Hoogsteen hydrogen bonding of four guanine bases in a planar
411 arrangement called a G-quartet, stabilised by a cation (e.g., potassium), which then stack on
412 top of each other to form the four-stranded structure. They can exist in different topologies
413 depending on the orientation of the guanines within the quartet, and can be found in various
414 genomic locations, including telomeres, promoters, untranslated regions of mRNA, and
415 mtDNA.

416 Consequently, we looked at the location of two LSMDs breakpoints relative to repeat regions
417 and putative secondary structural motifs. These features have long been implicated in the
418 formation of mtDNA deletions through slipped-strand mispairing, and replication stalling
419 mechanisms (Phillips et al. 2017). Here, identical, or highly similar DNA sequences are
420 mispaired during replication by the mtDNA polymerase, which slides along the template
421 leading to the removal of the region between the repeats. Furthermore, replication may stall at
422 non-canonical DNA structures such as G4s, causing it to become damaged through ds-breaks.
423 This could be mediated by microhomology-mediated end joining (MMEJ), one of the main
424 DSBR repair mechanisms in the mtDNA, which has been implicated in the formation of
425 LSMDs. Since MMEJ requires micro-homologous regions of 5-22 bp in length, these imperfect
426 repeat sequences may have determined the specific location where these LSMDs formed. Thus,
427 a plausible mechanism for LSMD formation in cells/tissues involves genomic instability
428 caused by G4 structures, replication fork slippage, transcription arrest, leading to the
429 introduction of double-strand breaks (DSBs) via Endonuclease G that are then repaired via
430 MMEJ, ultimately resulting in the formation of LSMDs (Dahal et al. 2022). This might be

431 particularly true in the context of mtDNA, which is very G-rich and has fewer ROS-protective
432 mechanisms and is therefore prone to destabilization and the formation of G4s (Aleksič,
433 Podbevšek, and Plavec 2025).

434 The two LSMDs investigated more closely in this study were adjacent to or within either
435 perfect or imperfect repeats, or potential G4-forming motifs, identified using QGRS mapper
436 and AF3. In both cases scores were low, although AF3 suggests certain putative G4 sequences
437 identified with QGRS Mapper are more likely to form in the presence of K+ ions. This is
438 perhaps as expected, since metal cations coordinate the G-tetrads leading to the stabilisation of
439 G4s (Phan 2010). While AF3 represents a significant advancement in macromolecule
440 (particularly protein) structure prediction, its ability to predict structures is influenced by the
441 availability of accurate experimental and training structures and/or evolutionary information,
442 such as homologous sequences e.g., from the Protein Databank (PDB, (Berman et al. 2000))
443 and other databases. While there are some 70 G4-structures in the PDB, none are from mtDNA,
444 which may have limited the ability of AF3 to predict structures with confidence for the LSMDs
445 m.6130_15055del in the KSS samples and m.8469_13446del in fibroblasts_nd.

446 While it is possible that some G4-structures are more likely to form than others, based on
447 specific sequences, the balance between ds-DNA and G4-stability and the chemical
448 environment at a particular point in time, other "spare" G-tracts may too form into G4s, and
449 may serve to mitigate the effects of the damaged G4s, preserving G4-function under stress
450 conditions (Aleksič, Podbevšek, and Plavec 2025). The predicted G4-sequences in this study,
451 yet unresolvable by AF3, might represent such "spare" regulatory elements. As more G4,
452 especially mtDNA-G4, structures become available the ability of AF3 to predict structures
453 accurately is set to increase and should shed light on the authenticity/importance of the other
454 putative G4-sequences identified by QGRS Mapper in the future.

Nonetheless, these findings could suggest that the mtDNA deletions, including those detected by "NanoDel", are not random, and that their formation is guided by both inherent sequence architecture and interactions with other factors. This includes ions that might be modulated within the mitochondria by the cellular microenvironment, thus contributing to our understanding of mtDNA instability in both health and disease. Interestingly, the two LSMDs investigated in more detail in our mitochondrial disease samples (this study) overlap with deletions observed in aging, metabolically active, and cancer tissues, as well as in neurodegenerative diseases, suggesting a shared vulnerability landscape shaped by mtDNA sequence motifs and structural constraints. These findings support the concept of "deletion hotspots" (Schon et al. 1989) as a unifying feature of mitochondrial dysfunction across multiple pathological states.

References:

Abramson, Josh, et al. (2024), 'Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3', *Nature* 1–3.

Aikhionbare, Felix O, et al. (2004), 'Is Cumulative Frequency of Mitochondrial DNA Variants a Biomarker for Colorectal Tumor Progression?', *Molecular Cancer*, 3, 1–4.

Aleksič, Simon, Peter Podbevšek, and Janez Plavec (2025), 'Oxidative Events in a Double Helix System Promote the Formation of Kinetically Trapped G-Quadruplexes', *Nucleic Acids Research*, 53/6, https://doi.org/10.1093/nar/gkaf260.

Andrews, Richard M, et al. (1999), 'Reanalysis and Revision of the Cambridge Reference Sequence for Human Mitochondrial DNA', *Nature Genetics*, 23/2, 147.

Aoki, Katsuyuki, and Kazutaka Murayama (2012), 'Nucleic Acid-Metal Ion Interactions in the Solid State', *Metal Ions in Life Sciences*, 10, https://doi.org/10.1007/978-94-007-2172-2_2.

Berman, Helen M., et al. (2000), 'The Protein Data Bank', in *Nucleic Acids Research*, xxviii, no. 1, preprint, https://doi.org/10.1093/nar/28.1.235.

Bolognini, Davide, et al. (2020), 'VISOR: A Versatile Haplotype-Aware Structural Variant Simulator for Short-and Long-Read Sequencing', *Bioinformatics*, 36/4, 1267–9.

Bushnell, Brian (2020), 'BBMap. Sourceforge. Net/Projects/Bbmap', preprint, Accessed.

Chiang, Colby, et al. (2015), 'SpeedSeq: Ultra-Fast Personal Genome Analysis and Interpretation', *Nature Methods*, 12/10, 966–8.

Coster, Wouter De, et al. (2018), 'NanoPack: Visualizing and Processing Long-Read Sequencing Data', *Bioinformatics*, 34/15, 2666–9.

Cretu Stancu, Mircea, et al. (2017), 'Mapping and Phasing of Structural Variation in Patient Genomes Using Nanopore Sequencing', *Nature Communications*, 8/1, 1326.

Dahal, Sumedha, et al. (2022), 'Characterization of G4 DNA Formation in Mitochondrial DNA and Their Potential Role in Mitochondrial Genome Instability', *The FEBS Journal*, 289/1, 163–82.

Damas, Joana, et al. (2014), 'MitoBreak: The Mitochondrial DNA Breakpoints Database', *Nucleic Acids Research*, 42/D1, D1261–D1268.

Dong, Dawei W, et al. (2014), 'Association of G-Quadruplex Forming Sequences with Human MtDNA Deletion Breakpoints', *BMC Genomics*, 15, 1–15.

Elango, Santhini, et al. (2011), 'Analysis of Mitochondrial Genome Revealed a Rare 50 Bp Deletion and Substitutions in a Family with Hypertension', *Mitochondrion*, 11/6, 878–85.

Ester, Martin, et al. (1996), 'A Density-Based Algorithm for Discovering Clusters A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise', *Proceedings - 2nd International Conference on Knowledge Discovery and Data Mining, KDD 1996*.

Frascarelli, Chiara, et al. (2023), 'Nanopore Long-Read next-Generation Sequencing for Detection of Mitochondrial DNA Large-Scale Deletions', *Frontiers in Genetics*, 14, https://doi.org/10.3389/fgene.2023.1089956.

Herbert, Joanna, et al. (2025), 'Impact of microbiological molecular methodologies on adaptive sampling using nanopore sequencing in metagenomic studies', *Environ Microbiome*, 20/1, 47.

Hjelm, Brooke E, et al. (2019), 'Splice-Break: Exploiting an RNA-Seq Splice Junction Algorithm to Discover Mitochondrial DNA Deletion Breakpoints and Analyses of Psychiatric Disorders', *Nucleic Acids Research*, 47/10, e59–e59.

Hwang, Bor-Jang, et al. (2009), 'Spectrum of Mitochondrial DNA Deletions within the Common Deletion Region Induced by Low Levels of UVB Irradiation of Human Keratinocytes in Vitro', *Gene*, 440/1–2, 23–7.

Jain, Miten, et al. (2018), 'Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads', *Nature Biotechnology*, 36/4, 338–45.

Johns, Donald R, et al. (1989), 'Directly Repeated Sequences Associated with Pathogenic Mitochondrial DNA Deletions.', *Proceedings of the National Academy of Sciences*, 86/20, 8059–62.

Keatley, Kathleen, et al. (2019), 'Integrated Approach Reveals Role of Mitochondrial Germ-Line Mutation F18L in Respiratory Chain, Oxidative Alterations, Drug Sensitivity, and Patient Prognosis in Glioblastoma', *International Journal of Molecular Sciences*, 20/13, 3364.

Kikin, Oleg, Lawrence D'Antonio, and Paramjeet S Bagga (2006), 'QGRS Mapper: A Web-Based Server for Predicting G-Quadruplexes in Nucleotide Sequences', *Nucleic Acids Research*, 34/suppl_2, W676–W682.

Kopsidas, George, et al. (2000), 'Preferential Amplification Is Minimised in Long-PCR Systems', *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 456/1–2, https://doi.org/10.1016/S0027-5107(00)00126-3.

Koren, Sergey, et al. (2017), 'Canu: Scalable and Accurate Long-Read Assembly via Adaptive k-Mer Weighting and Repeat Separation', *Genome Research*, 27/5, 722–36.

Kruskal, William H., and W. Allen Wallis (1952), 'Use of Ranks in One-Criterion Variance Analysis', *Journal of the American Statistical Association*, 47/260, https://doi.org/10.1080/01621459.1952.10483441.

Lam, Ernest T, et al. (2012), 'Mitochondrial DNA Sequence Variation and Risk of Pancreatic Cancer', *Cancer Research*, 72/3, 686–95.

Layer, Ryan M, et al. (2014), 'LUMPY: A Probabilistic Framework for Structural Variant Discovery', *Genome Biology*, 15/6, 1–19.

Li, Heng, et al. (2009), 'The Sequence Alignment/Map Format and SAMtools', *Bioinformatics*, 25/16, 2078–9.

Li, Heng (2018), 'Minimap2: Pairwise Alignment for Nucleotide Sequences', *Bioinformatics*, 34/18, 3094–100.

Lloyd, Rhiannon E, et al. (2012), 'Next Generation Sequencing and Comparative Analyses of Xenopus Mitogenomes', *BMC Genomics*, 13/1, 1–16.

Lloyd, Rhiannon E, et al. (2015), 'Identification and Functional Prediction of Mitochondrial Complex III and IV Mutations Associated with Glioblastoma', *Neuro-Oncology*, 17/7, 942–52.

Mahmoud, Medhat, et al. (2019), 'Structural Variant Calling: The Long and the Short of It', *Genome Biology*, 20, 1–14.

Marti, Ramon, et al. (2005), 'Late-Onset MNGIE Due to Partial Loss of Thymidine Phosphorylase Activity', *Annals of Neurology*, 58/4, https://doi.org/10.1002/ana.20615.

Martin, Samuel, et al. (2022), ‘Nanopore Adaptive Sampling: A Tool for Enrichment of Low Abundance Species in Metagenomic Samples’, *Genome Biology*, 23/1, https://doi.org/10.1186/s13059-021-02582-x.

Mita, Shuji, et al. (1990), ‘Recombination via Flanking Direct Repeats Is a Major Cause of Large-Scale Deletions of Human Mitochondrial DNA’, *Nucleic Acids Research*, 18/3, 561–7.

Moraes, Carlos T, et al. (1989), ‘Mitochondrial DNA Deletions in Progressive External Ophthalmoplegia and Kearns-Sayre Syndrome’, *New England Journal of Medicine*, 320/20, 1293–9.

Pedregosa, F, et al. (2011), ‘Scikit-Learn: Machine Learning in Python’, *Journal of Machine Learning Research*, 12, 2825–30.

Phan, Anh Tuân (2010), ‘Human Telomeric G-Quadruplex: Structures of DNA and RNA Sequences’, *The FEBS Journal*, 277/5, 1107–17.

Phillips, Aaron F., et al. (2017), ‘Single-Molecule Analysis of MtDNA Replication Uncovers the Basis of the Common Deletion’, *Molecular Cell*, 65/3, https://doi.org/10.1016/j.molcel.2016.12.014.

Poulton, J, et al. (1995), ‘Mitochondrial DNA, Diabetes and Pancreatic Pathology in Kearns-Sayre Syndrome’, *Diabetologia*, 38, 868–71.

Poulton, J, M E Deadman, and R M Gardiner (1989), ‘Tandem Direct Duplications of Mitochondrial DNA in Mitochondrial Myopathy: Analysis of Nucleotide Sequence and Tissue Distribution’, *Nucleic Acids Research*, 17/24, 10223–9.

Poulton, Jo, et al. (1993), ‘Families of MtDNA Re-Arrangements Can Be Detected in Patients with MtDNA Deletions: Duplications May Be a Transient Intermediate Form’, *Human Molecular Genetics*, 2/1, 23–30.

Regan, Catherine (1999), *SEQUENCE CHARACTERIZATION OF BREAKPOINT REGIONS OF HUMAN MITOCHONDRIAL DNA DELETIONS AND POSSIBLE MECHANISMS OF FORMATION*.

Reyes, Aurelio, et al. (2015), ‘RNASEH1 Mutations Impair MtDNA Replication and Cause Adult-Onset Mitochondrial Encephalomyopathy’, *American Journal of Human Genetics*, 97/1, https://doi.org/10.1016/j.ajhg.2015.05.013.

Santorelli, F M, et al. (1996), ‘Multiple Mitochondrial DNA Deletions in Sporadic Inclusion Body Myositis: A Study of 56 Patients’, in *Annals of Neurology*, no. 6, preprint, Wiley Online Library, xxxix, 789–95.

Schon, Eric A, et al. (1989), ‘A Direct Repeat Is a Hotspot for Large-Scale Deletion of Human Mitochondrial DNA’, *Science*, 244/4902, 346–9.

589 Sen, Dipankar, and Walter Gilbert (1988), 'Formation of Parallel Four-Stranded Complexes
590 by Guanine-Rich Motifs in DNA and Its Implications for Meiosis', *Nature*, 334/6180,
591 https://doi.org/10.1038/334364a0.

592 Shapiro, S. S., and M. B. Wilk (1965), 'An Analysis of Variance Test for Normality
593 (Complete Samples)', *Biometrika*, 52/3/4, https://doi.org/10.2307/2333709.

594 Sievers, Fabian, et al. (2011), 'Fast, Scalable Generation of High-Quality Protein Multiple
595 Sequence Alignments Using Clustal Omega', *Molecular Systems Biology*, 7,
596 https://doi.org/10.1038/msb.2011.75.

597 Spelbrink, Johannes N., et al. (2001), 'Human Mitochondrial DNA Deletions Associated with
598 Mutations in the Gene Encoding Twinkle, a Phage T7 Gene 4-like Protein Localized in
599 Mitochondria', *Nature Genetics*, 28/3, https://doi.org/10.1038/90058.

600 Stothard, Paul (2000), 'The Sequence Manipulation Suite: JavaScript Programs for Analyzing
601 and Formatting Protein and DNA Sequences', *Biotechniques*, 28/6, 1102–4.

602 Vafai, Scott B., and Vamsi K. Mootha (2012), 'Mitochondrial Disorders as Windows into an
603 Ancient Organelle', in *Nature*, cdxci, no. 7424, preprint,
604 https://doi.org/10.1038/nature11707.

605 Vandiver, Amy R., et al. (2023), 'Nanopore Sequencing Identifies a Higher Frequency and
606 Expanded Spectrum of Mitochondrial DNA Deletion Mutations in Human Aging',
607 *Aging Cell*, 22/6, https://doi.org/10.1111/acel.13842.

608 Volmering, Elisa, et al. (2016), 'Neuropathological Signs of Inflammation Correlate with
609 Mitochondrial DNA Deletions in Mesial Temporal Lobe Epilepsy', *Acta*
610 *Neuropathologica*, 132, 277–88.

611 Wang, Kai, et al. (2010), 'MapSplice: Accurate Mapping of RNA-Seq Reads for Splice
612 Junction Discovery', *Nucleic Acids Research*, 38/18, e178–e178.

613 Wanrooij, Sjoerd, et al. (2004), 'Twinkle and POLG Defects Enhance Age-Dependent
614 Accumulation of Mutations in the Control Region of MtDNA', *Nucleic Acids Research*,
615 32/10, https://doi.org/10.1093/nar/gkh634.

616 Wei, Yau-Huei (1992), 'Mitochondrial DNA Alterations as Ageing-Associated Molecular
617 Events', *Mutation Research/DNAging*, 275/3–6, 145–55.

618 Wood, Elizabeth, et al. (2019), 'Clinical Long-Read Sequencing of the Human Mitochondrial
619 Genome for Mitochondrial Disease Diagnostics', *BioRxiv* 597187.

620 You, Yupei, Michael B Clark, and Heejung Shim (2022), 'NanoSplicer: Accurate
621 Identification of Splice Junctions Using Oxford Nanopore Sequencing', *Bioinformatics*,
622 38/15, 3741–8.

623 Yusoff, Abdul Aziz Mohamed, et al. (2020), 'Prevalence of Mitochondrial DNA Common
624 Deletion in Patients with Gliomas and Meningiomas: A First Report from a Malaysian
625 Study Group', *Journal of the Chinese Medical Association*, 83/9, 838–44.

Zhang, Wei, Hong Cui, and Lee-Jun C Wong (2012), 'Comprehensive One-Step Molecular Analyses of Mitochondrial Genome by Massively Parallel Sequencing', *Clinical Chemistry*, 58/9, 1322–31.

Author Contributions:

Conceptualization – REM (lead), Data Curation – REM, SCR and CFe (equal), Formal Analysis – REM, SCR and CFe (equal), Funding Acquisition – REM (lead), Investigation – CFe and REM (equal), Methodology – CFe, SCR and REM (equal), Project Administration – REM (lead), Resources – REM and SCR (equal), JP and CFr (supporting), Software – CFe (lead), SCR (supporting), Supervision – REM (lead), SCR and RB (supporting), Visualization – CFe (lead), REM (supporting), Writing – Original Draft Preparation – REM (lead), CFe (supporting), Writing – Review & Editing – REM (lead), CFe, CO, CG, JP, CFr, RB, SCR (supporting).

Acknowledgments:

We are thankful to Shelest, E. and Lundgren, A.P., who aided in the design of the script for clustering LSMD breakpoints in the "NanoDel" pipeline. We also gratefully acknowledge Beckett, A. and Dent, H., who helped with the library preparation and sequencing. CFe was supported by a University of Portsmouth Faculty-PhD studentship. For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC-BY) licence to any author accepted manuscript version arising from this submission.