**Long title:**

# NanoDel: a long-read sequencing pipeline for identifying large-scale mitochondrial DNA deletions validated in patient samples clinically diagnosed with mitochondrial disease and evaluated in glioblastoma.

**Short title:**

NanoDel: long-read sequencing pipeline for detecting large-scale mitochondrial DNA deletions

Fearn, C. 1, Oliva, C. 2, Griguer, C. 2, Poulton, J. 3, Fratter C. 4, McGeehan, J.E. 5, Baldock, R. A. 1, Robson, S.C. 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 Free Radical and Radiation Biology Program, University of Iowa, USA, 3 Nuffield Dept of women's reproductive health, University of Oxford, UK, 4 Oxford Genetics Laboratories, Oxford University Hospitals NHS Foundation Trust, UK, 5 Centre for Enzyme Innovation, Institute of Life Sciences and Healthcare, University of Portsmouth, UK. *Corresponding author.

**Abstract:**

Primary glioblastoma is deadly, with limited improvements in prognosis. We developed "NanoDel", a long-read sequencing pipeline to sensitively detect large-scale mitochondrial DNA (mtDNA) deletions (LSMDs) in glioblastoma to explore their prognostic significance. Following validation using artificial datasets and samples from patients diagnosed with mitochondrial disease, "NanoDel" identified 171 low-level heterogeneous glioblastoma-LSMDs, clustering in mt-co1, mt-cyb genes, and adjacent to repeat, and G4 motifs. These "hotspots" were also seen in other tissues, revealing a shared vulnerability in mtDNA. While LSMD burden alone was not prognostic, mtDNA copy number and complex IV activity strongly stratified glioblastoma patient survival. "NanoDel" offers a clinically and cross-organismal applicable tool for LSMD detection and new insights into mtDNA markers and instability mechanisms.

**Keywords**:

Glioblastoma, mtDNA, mitochondria, deletions, sequencing, disease, prognosis

## Introduction:

Glioblastoma (glioblastoma; WHO grade IV) is the most aggressive and common primary brain tumor in adults (1). Patients diagnosed with glioblastoma who are healthy enough to

undergo the standard of care treatments have a median survival time of only 14.6 months following diagnosis (2). The challenge of effective treatment is complicated by numerous factors, including the diverse molecular biology found both between different tumors and within a single tumor. This is further hindered by the reliance on the current standard of care treatment that involves resection, radiotherapy, and DNA alkylating agents such as temozolomide (TMZ), which demonstrably increases survival by 2-3 months (3).

Over the past decade, research has identified the isocitrate dehydrogenase 1 (IDH1) R132 mutation as being associated with improved overall survival. However, it only occurs in 5-10 % of adult glioblastoma as it is inversely associated with grade in diffuse glial tumors (4). The methylation status of O6-methylguanin-DNA-methyltransferase (MGMT), typically occurring in 30-50 % of glioblastoma, is also linked to a better response to TMZ (5). Furthermore, six major subgroups have emerged based on cellular origin, anatomical location, various molecular changes, and transcriptomic profiles, each associated with different prognoses or therapy responses (6)(7) Despite these discoveries, there is still considerable need for molecular markers that meaningfully guide treatment options that improve patient quality of life and/ or survival.

Mitochondrial changes are a key characteristic of glioblastoma, distinguishing it from normal brain cells and even varying between different glioblastomas (8)(9). As a result, researchers have been exploring their role in tumor development, progression, and drug sensitivity, with the goal of discovering new predictive and prognostic markers to guide treatment decisions and support drug development.

Mitochondria are cellular organelles that have their own double-stranded, supercoiled genome, mitochondrial DNA (mtDNA). MtDNA has 38 loci, including 13 protein coding genes, 2 rRNAs, 22 tRNAs and the control region (D-loop), integral for producing the multi-protein mitochondrial respiratory chain (MRC) complexes (I-IV), which along with complex V forms part of the oxidative phosphorylation (OXPHOS) system. This system, which also includes approximately 70 nuclear-encoded subunits, is central to ATP generation and cellular metabolism (reviewed in (10)). Direct and/or indirect alterations in the mitochondrial genome can impact the function of the OXPHOS system and therefore cellular energy generation and contribute to disease (for examples, see Table S1).

Alterations in nuclear genes that encode succinate dehydrogenase (SDH, MRC complex II) contribute to the pathology of cancers by altering corresponding MRC complex II activity and substrate levels. Such alterations in turn can also alter the production of ROS, cell redox state and consequently, activate transcription factors such as HIF1-alpha and FOS-JUN, which can also alter gene expression and stimulate cancer cell proliferation (11)(12). Increased ROS levels can also lead to increases in basal levels of oxidative stress, making cancer cells vulnerable to agents that further increase ROS levels. Together with the knowledge that a functional MRC is essential for glioblastoma growth in vivo (13), it is possible that the mtDNA alterations (either the same or like those that cause mitochondrial disease) that encode MRC complexes could also affect their activities and contribute to the pathology of cancers. However, proving the mtDNA genotype-phenotype relationships directly (e.g., by introducing

SNVs into human mtDNA) has historically been challenging due to difficulties in delivering base editing tools into the mitochondria and/ or DNA repair mechanisms, preventing incorporation of desired edits. Although recent advances mean that base editing tools for mtDNA are evolving (14), indirect (association) studies remain valuable.

Changes in MRC complex I–IV activities have been described in glioblastoma compared to the normal brain, and between different glioblastoma samples (15)(16). Furthermore, changes in MRC complex IV activity have been useful in stratifying patients into subgroups with different survival outcomes, either independently (16) or in combination with MGMT methylation promotor status (17). Although the mtDNA SNVs and copy number spectrums of glioblastoma differ from those of normal brain and are highly variable between different glioblastoma samples (18)(19), whether differences in these mtDNA alterations contribute to the differential MRC activities, and patient survival outcomes, has been investigated within the last decade.

Recent evidence suggests that the inherited mtDNA SNV m.14798T>C, which causes an F18L substitution within the Qi-pocket of MT-CYB within MRC complex III, is associated with alterations in individual MRC complex activities in patient-derived biopsy cell cultures. F18L is associated with a 75% increase in MRC complex II_III activity, a 34% reduction in CoQ10 (the natural substrate for MRC complex III) levels, an 87% increase in intracellular ROS, an altered cellular distribution of mitochondrial-specific ROS, and a 64% increased sensitivity to clomipramine (a repurposed MRC complex III-targeting drug). In a small cohort of patients, F18L-carriers that received the current standard of care treatment had a poorer prognosis than non-carriers (373 days vs. 415 days, respectively), although this effect was not statistically significant (20). While accumulating evidence suggests that elevated mtDNA-CN correlates with longer glioblastoma patient survival, this correlation may be dependent on the age of the patients (21)(22).

In addition to mtDNA SNV and CN changes being observed in glioblastoma, large scale mtDNA-deletions (LSMDs) have occasionally been identified in glioblastoma, e.g., in cells that have acquired TMZ resistance, through damage to the mtDNA via chronic TMZ treatment in vitro (23). Furthermore, the 4977-bp deletion, often referred to as the “common deletion” or mtDNA4977 (specifically m.8471-13449del) was identified more frequently in high vs. low grade gliomas, in males vs. females, and in older vs. younger patients. Correlations with gender and tumor grade were significant, but not with age (24).

The relative paucity of studies documenting and exploring the importance of LSMDs, compared to SNV and CN changes, in glioblastoma could be attributed in part to studies using cell cultures rather than tissues (18)(20). LSMDs are not necessarily maintained during cell culture (25). The paucity could also be attributed to studies using the whole exome sequence analysis of mtDNAs (26) rather than methods where the whole mtDNAs are enriched as a first step, e.g., as one-amplicon using LR-PCR (27). LSMDs are detected far less frequently because traditional probes used in whole exome sequencing are designed to avoid encompassing a deletion breakpoint, thus reads that flank these breakpoints will be less likely to be amplified and observed compared to wild-type molecules.

Several traditional methods, such as Southern blotting (28)(29)PCR (30) and Sanger sequencing (31) (32) (33), have been used historically to detect LSMDs. However, the accurate breakpoint and heteroplasmy determinations using these techniques is challenging, requiring a priori breakpoint information, and high DNA inputs. Often such methods also lack sensitivity, are not completely quantitative, and can be labour intensive. One way to identify the breakpoints where multiple unknown LSMDs are present would be to plasmid clone and select all molecules prior to Sanger sequencing, however this is even more labour intensive.

A useful 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 (27) 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, demonstrating that this increased the sensitivity in detecting true positives compared to traditional DNA alignment tools (e.g. Burrows-Wheeler aligner) by around 60% (27). While examples of LMSDs detection by ONT sequencing exist, these have not used LR-PCR or a splice-aware aligner (34)(35). However, employing ONT sequencing for LSMD detection could offer additional advantages over short read (e.g., Illumina) sequencing, including 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 improving quality scores above Q20, ONT is an attractive alternative to Illumina for the accurate breakpoint identification and quantification of LSMDs.

The limited number of studies exploring the importance of LSMDs in glioblastoma tissues, and the lack of an ONT pipeline comparable to that of (27), prompted us to develop the pipeline "NanoDel". This pipeline involves LR-PCR of whole mtDNAs as single amplicons, followed by ONT sequencing, and bioinformatic tools for accurate breakpoint identification and quantification in glioblastoma tissues.

In this study we first compared the ability of four pipelines to determine LSMDs breakpoints at different heteroplasmy levels using 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, (36)), and two that use short read sequences either in combination with a splice-aware RNA aligner ("Splice-Break", Pipeline 2, (27)) or a DNA aligner ("Wood Illumina", Pipeline 4, (36)). 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 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. Fourth, we used "NanoDel" and one-amplicon PCR to generate the first catalogue of LSMDs in glioblastoma tissues. Finally, we used survival/cluster analysis to determine the prognostic significance of LSMDs to glioblastoma patients, either alone or in combination with

other parameters (e.g., mtDNA-CN, MRC CIV activity, age, gender, and MGMT methylation promotor status).

Compared to alternatives, "NanoDel" in combination with one, rather than two, amplicon long range (LR)-PCR, identified singular and mixed LSMDs breakpoints and heteroplasmy levels most accurately and sensitively in artificial and clinical datasets. We identified a total of 171 low level heterogeneous LSMDs in our glioblastoma samples. While we were able to stratify patients based on cumulative LSMD metrics into groups, with our small cohort, these did not correlate with differences in median overall survival (p=0.08). In contrast, patients stratified based on mtDNA-CN or complex IV activity levels, had a greater impact on prognosis (p=0.02 and p=0.001, respectively). Patients with low complex IV activity, rather than high complex IV activity, had better survival, higher mtDNA-CN, and a greater accumulation of LSMDs, compared to those with high complex IV activity. We further show that LSMDs in clinically diagnosed mitochondrial disease and glioblastoma patients are associated with: mt-co1 and mt-cyb, repeat and putative G4 structure-forming sequences. Furthermore, the breakpoints detected by "NanoDel" overlap with those reported in a range of healthy and pathological tissues, indicating a shared vulnerability landscape shaped by mtDNA sequence motifs and structural constraints.

Taken together, "NanoDel" combined with one-amplicon LR-PCR strategy offers a robust strategy with clinical application for detecting LSMDs across a variety of samples, reveals insights into their association with other mitochondrial and clinical parameters in glioblastoma, and yields insight into their mechanism of formation, contributing new tools to, and furthering our understanding within the mtDNA instability and glioblastoma space.

**Methods and Materials:**

Patient-derived tissues and cell cultures

Clinically diagnosed mitochondrial disease cases:

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 (37)(31)(38), University of Oxford. Additionally, DNA samples from muscle (MM, PS and PM) and cultured cells (ML and PM) 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. The original tissues were surplus to the quantity required for diagnosis, donated for research between January 1990 and June 2021 with written informed consent from all participants and with the approval of the United Kingdom National Research Ethics Service (South Central-Berkshire, #REC 20/04, IRAS 162181). Previous observations, i.e., LSMD composition determined by a range of different techniques following diagnosis at the OGL, were made available for these samples following sequencing analysis to prevent bias (Table S2).

200 Primary glioblastoma and normal brain tissue:

201 Eighteen total DNA samples extractedfrom the tumor tissue of patients diagnosed with primary
202 untreated glioblastoma (characterised for age, MGMT status, gender, CIV and citrate synthase
203 (CS) activity with matched survival data), and two normal brain tissue specimens from epilepsy
204 patients were obtained following surgical resection at the University of Alabama at
205 Birmingham (UAB) Hospital between January 2001 and November 2011 from the UAB Brain
206 Tumor Tissue Bank. These DNA samples were prepared and de-identified and provided by
207 Professors Corinne Griguer and Claudia Oliva, formerly from the UAB prior to transfer to our
208 laboratory. None of these patients received chemotherapy or radiotherapy before the surgery.
209 The original tissues were procured under the institutional review board for human use at the
210 UAB (IRB #X050415007 and X110418007). All patients provided written and informed
211 consent to the surgical procedures and gave permission for the use of restricted specimens (16).

212 Amplification control samples

213 Two human adult non-neoplastic astrocyte cell lines derived from the cerebral cortex (control)
214 were obtained commercially (ScienCell research laboratories, #1800) for use as additional
215 normal brain and mtDNA positive amplification controls. The 143B osteosarcoma and 143B
216 rho-0 (a derivative of the 143B cells that have lost their mtDNA) cell lines were obtained
217 commercially (Kerafast, ESA106), for use as mtDNA positive and mtDNA negative
218 amplification controls, respectively (Table S2).

219 Total DNA extraction and LR-PCR

220 Total DNA was extracted from all cell culture samples using the QIAmp DNA mini kit as per
221 the manufacturer's instructions (Qiagen, 51304), and whole mtDNAs were amplified in either
222 two overlapping amplicons (39)(18)(20) or as a single amplicon (27) by LR-PCR using human
223 mtDNA-specific primers (Table S3). Total DNA samples were also assayed for the number of
224 mtDNA copies per cell.

225 Two-amplicon LR-PCR:

226 Whole mtDNAs (including any LSMDs) were amplified in two overlapping amplicons of
227 8,573bp and 9,128bp in length using human mtDNA specific primers in a 50µL reaction
228 volume using the Expand Long Range dNTPack kit (Roche) (39)(18) (20). The 50µL reaction
229 contained 150nM (Amplicon 1) or 300 nM (Amplicon 2) of each primer, 1x PCR buffer,
230 500µM of each dNTP, 3% DMSO, 3.5U of polymerase and 8ng/µL of total DNA. The reaction
231 was heated for 40 cycles with the lid heated to 103°C, 92°C initial denaturation, followed by
232 10 cycles of a 10 second 92°C denaturation, 15 second 58°C annealing and 10 minutes 68°C
233 extension. This is followed by 30 cycles of 10 seconds at 92°C, 15 seconds at 58°C, 10 minutes
234 at 68°C increasing by 10 seconds per cycle and a final extension at 68°C for 7 minutes. PCR
235 products were assessed by Bioanalyser Fragment Length Analysis (1µL) using the Agilent
236 DNA 12000 kit, and the Agilent 2100 Bioanalyser instrument, as described by the
237 manufacturer's protocol. Products were then purified using the QIAquick® PCR purification
238 Kit (Qiagen).

239 One-amplicon LR-PCR:

240 Each sample was also amplified as a single 16,266bp amplicon using the Takara LA Taq DNA
241 polymerase kit (Takara Biosciences, RR002M). PCR reactions were completed in a 50µL
242 reaction volume containing 200nM of each primer, 1x LA PCR buffer, 400nM each dNTP,
243 2.5U of LA Taq and 0.4ng/µL of total DNA input. The reaction was heated 94°C for a 2-minute
244 initial denaturation. It was then heated to 94°C for 10 seconds, 65°C for 15 minutes for 10
245 cycles followed by 94°C for 10 seconds and 65°C for 15 minutes increasing by 20 seconds per
246 cycle for 30 cycles. Finally, they were heated to 72°C for 10 minutes for a final extension.
247 Agarose gel electrophoresis using a 0.8% agarose gel was used to analyse approximate
248 fragment lengths. PCR products were then purified using AMPure XP beads (Beckman
249 Coulter, A63881) and quantified using Qubit Fluorimetry (Thermofisher, Q33238).

250 MtDNA copy number analysis

251 Standards for the quantification of the number of mtDNA copies relative to nDNA copies were
252 generated by conventional PCR using a Veriti 96-well Thermal Cycler (Applied Biosystems),
253 and mt-nd4 and beta-actin primers, respectively (Table S3). Reaction mixtures with a final
254 volume of 50 µL contained 20 ng/ µL of DNA, 10x PCR buffer (Bio-Rad), 50mM MgCl2 (Bio-
255 Rad), 10 mM dNTP (Fisher), 4 µM primers and 0.5 U iTaq DNA polymerase (Bio-Rad). The
256 reaction conditions used were initial denaturation at 94 $^{o}$C for 15 min, 35 cycles of 94 $^{o}$C for
257 40 seconds, annealing at 52 $^{o}$C (mt-nd4) and 59 $^{o}$C (beta-actin) for 40 seconds, extension at 72
258 $^{o}$C for 1 min and final extension at 72 $^{o}$C for 10 minutes. For each gene, 10 µL of the PCR
259 product was resolved on a 2% agarose gel and the product size was confirmed against a 100 bp
260 ladder (New England BioLabs). PCR products were purified using QIAquick PCR Purification
261 Kit (Qiagen) as described by the manufacturer's protocol. After NanoDrop (Thermo Fisher
262 Scientific, USA) quantification, a series of 10-fold dilutions were prepared to generate a
263 standard curve for each gene.

264 Real-time PCR was performed on samples and standards using the QuantStudio™ 5 Real-Time
265 PCR System (Applied Biosystems, Thermofisher). Reaction mixtures with a final volume of
266 16µL were prepared containing 8µl of iTaq Universal SYBR Green Supermix (Bio-Rad), 6.4
267 µM of each primer, 6.96 µL of nuclease free water and 0.8 µL of standard (range: 0.2 ng/ µL –
268 2x10-4 ng/ µL) or 8 ng/ µL DNA sample solutions. Each reaction was run in duplicate for
269 standards, samples, and negative controls. For mtDNA, reaction conditions were initial
270 denaturation at 95 $^{o}$C for 10 min, followed by 40 cycles of denaturation at 95 $^{o}$C for 15 seconds
271 and annealing at 52 °C (mt-nd4) and 59°C (beta-actin) for 10s. The melt curve was analysed
272 to check the specificity of the PCR products. As the nuclear gene beta-actin is present twice in
273 all cells, the beta-actin amplification signal was divided by two to compute the number of cells
274 present in each sample. The number of mitochondrial genomes per cell was then calculated by
275 dividing amplification signal by the number of cells (40)(41).

276 Library Preparation and Sequencing

277 Nanopore Library Preparation:

Library preparation was completed using the Native Barcoding Kit 96 V12 library kit (ONT, NBD112.96) according to manufacturer's specifications with small alterations. Individual amplicons underwent end preparation using the NEB Ultra End Prep and dA tailing module (New England Bioscience, E6053S). Individual amplicons were balanced in a 96-well plate to 8ng/ul with nuclease free water, and a total DNA quantity of 100ng. To each reaction 1.75µL Ultra II End Prep Reaction Buffer and 0.75µL Ultra II End Prep Enzyme Mix was added. . The 96-well plate was sealed and incubated in a thermocycler at 20°C for 5 minutes followed by 60°C for 5 minutes.

End repaired samples were barcoded using the barcodes supplied in the NBD112.96 kit and the NEB Blunt T/A Ligase MasterMix (NEB, M0367S). Each barcoding reaction contained 5µL Blunt T/A MasterMix, 3µL nuclease free water, 1.25µL end prepped DNA and 0.75µL of each respective barcode. The barcoding reactions were then incubated again at room temperature for 20 minutes. Post-incubation 1µL of EDTA was added to each individual reaction to halt the reaction.

Samples were then multiplexed into a singular 1.5mL tube and cleaned using AMPure XP beads (Beckman Coulter, A63880). A bead volume of 0.4x pooled sample volume was used for bead washing. The sample/bead mixture was incubated on a Hula mixer for 10 minutes at room temperature. The bead pellet was washed in 700µL of freshly prepared ethanol which was subsequently removed and discarded. After being allowed to dry, the bead pellet was resuspended in 35µL of nuclease free water to elute the DNA. This was incubated at 37°C for 10 minutes, being agitated every 2 minutes. The eluate was then retained.

Adapter sequences were ligated to the multiplexed DNA sample using the NEBNext Quick Ligation module (New England Bioscience, E6056S). This reaction contained 30µL of the DNA sample, 5µL of adapter mix II , 10µL of the quick ligation reaction buffer and 5µL of the T4 DNA ligase. Excess adapter mix was then removed with bead washing using a 0.4x reaction volume of AMPure XP beads. This was incubated on a hula mixer for 10 minutes at room temperature and then spun down and pelleted on a magnetic rack. The bead pellet was washed in 125µL of short fragment buffer (ONT, NBD112.96) twice, pelleted and then DNA was eluted in 15µL of elution buffer (ONT, NBD112.96), incubated at 37°C for 10 minutes agitated every 2 minutes. DNA was then quantified using a Qubit Fluorometer (Thermo Fisher Scientific, USA). Where possible the prepared library was then mixed with 37.5µL of Sequencing Buffer (SQB) and 25.5µL of Loading Beads (LB) (ONT, NBD112.96) to a concentration of 1.4ng/µL and a final volume of 75µL.

Separately, 30µL of Flow Tether (FLT) was added to a tube of thawed Flow buffer (FB) (ONT, EXP-FLP002). A P1000 pipette is then used to remove excess air from the R10.4 flow cell (ONT, FLO-MIN112) and 800µL of FLT/FB mix pipetted slowly into the priming port. This was left to incubate for 5 minutes before an extra 200µL of FLT/FB was pipetted into the priming port with the spot-on cover open. The 75µl library was then added to the flow cell through the spot-on port dropwise. Sequencing was conducted for 72 hours with the flow cell being refuelled every 16 hours after DNA was loaded with an additional 200µL of fresh FLT/FB mix. If excess library was available, smaller loading libraries would be prepared with

319 a DNA concentration of 0.7ng/ul and used to reload the flow cell every 24 hours for the
320 remainder of the sequencing duration. The sequencing was run using the MinKnow software
321 from ONT, with the barcode demultiplexing, trim barcodes, mid-read barcode filtering, and
322 super-accurate base calling settings enabled, with outputs in both Fastq and Fast5 formats. All
323 sequencing was conducted using an Oxford Nanopore Technologies GridION X5.

324 Bioinformatics:

325 Raw reads were base called using Guppy (v6.2.7 and v6.3.9) and transferred to a local high-
326 performance computing cluster running on Ubuntu (v20.04.5). As barcodes had already been
327 trimmed, base called reads were initially filtered to remove reads of <20bp in length or a Q-
328 score <10 using NanoFilt (v2.8.0, (42). In this study all reads were aligned to the rCRS
329 reference sequence (NC_012920.1, (43)).

330 Artificial data:

331 Two sets of artificial Fastq files were generated using VISOR (v1.1, (44). These were generated
332 in two ways: 1) an ONT style with a longer average read length of ~8kb and a "Nanopore 2020"
333 style error model, and 2) an Illumina style with 250bp length reads and an "Illumina" style
334 error model. The first set (artificial data 1) contained controlled quantities of reads designed to
335 emulate mtDNA4977, specifically m.8471-13449del previously found in glioblastoma (24).
336 These deletions were simulated in 75%, 50%, 25%, 5%, 1% and 0.25% of reads. The second
337 set of artificial Fastq files (artificial data 2) contained multiple deletions at nucleotides m.8471-
338 13449del, m.6620-8440del, m.3541-5500del, and m.12538-15472del, all simulated in
339 approximately 10% of reads. These files were used to evaluate and compare four separate
340 alignment and deletion calling strategies highlighted below.

341 Alignment and LSMD analysis strategies:

342 In this study four pipelines were compared to analyse the efficacy of using a splice-aware RNA
343 alignment (for DNA reads) vs. a traditional DNA alignment algorithm to identify deletions.
344 For "NanoDel" (Pipeline 1), compiled specifically for this study, Fig S1), reads were filtered
345 further removing those with a read length <200bp and a Q-score of 15. Reads were aligned
346 using minimap2 (v2.21, (45)) with the "splice" setting enabled but preferences for canonical
347 splicing disabled. The BAM files were indexed using SAMtools (v1.7,(46). Analysis of
348 junctions within reads was completed using the JWR_Checker python script from NanoSplicer
349 (v1.0, (47)) and used to output a .csv file containing all junctions within reads. The .csv files
350 were formatted using an R script and deletion breakpoints were clustered using sci-kit learn
351 (v1.1.1, (48)), DBScan (49) within a distance (eps) of 10 and a minimum cluster requirement
352 of 2 samples. A mean value of each breakpoint in the cluster was then taken and reported back
353 as the deletion, with a summation of the total reads for each deletion.

354 For "Splice-Break" (Pipeline 2), files were deduplicated and reformatted using BBmap
355 (V38.18, BBTools (50)). Reformatted paired end Fastq files were then aligned using RNA
356 aligner Mapsplice2 (v2.4, (51)) using the non-canonical double anchor setting. The "Splice-
357 break".sh bash script (27) was then used for junction analysis, which created a .txt file
358 containing deletion breakpoints.

Approaches for “Wood ONT” (Pipeline 3), and “Wood Illumina” (Pipelines 4) are based on those described in (36). “Wood ONT” (Pipeline 3) made use of Minimap2 (v2.21, (45)) with the DNA alignment “Map-ONT” setting. Structural variants in the resulting BAM file were then analysed using NanoSV (v1.2.4, (52)), outputting a VCF file. Structural variants were then called from the VCF file using sniffles (v1.0.12, (53)). “Wood Illumina” (Pipeline 4) aligned reads using BWA-mem (v0.7.17). Split and discordant reads were extracted from the .bam files using BWA-mem and SAMtools. Structural variant analysis was conducted using LumpyExpress (v0.2.14, (54), and the VCF file generated was genotyped using SVTyper (v0.7.1, (55)). “Wood ONT” (Pipeline 3) typically uses the ‘Canu’ software to correct and trim the sequences prior to alignment with Minimap2. However, during tests on artificial Fastq data, Canu removed all reads from the data set, due to aggressively down-sampling coverage. This read correction was likely confounded by the effects of heteroplasmy, leading to excessive reads being removed. Consequently, for the analyses of artificial data sets we also proceeded without using Canu. The “NanoDel” pipeline is freely available for use on GitHub: https://github.com/uopbioinformatics/NanoDel.

Benchmarking and filtering:

For “NanoDel” (Pipeline 1), “Wood ONT” (Pipeline 3, (36)) and "Wood Illumina" (Pipeline 4,(36)), the number of deleted reads were normalised for each sample against a coverage benchmark to provide a relative read % for each LSMD. Coverage benchmark regions were established through the visible identification of a 500bp region with an even un-interrupted coverage across samples between positions m.500-1000del. “Splice-Break” (Pipeline 2, (27)) has an in-built benchmarking approach that uses two 250bp regions between positions m.957-1206del and m.15357-15606del to normalise the deleted reads and provide a relative read % for each LSMD between samples.

Additionally, for “NanoDel” (Pipeline 1) and “Splice-break” (Pipeline 2, (27)), a separate positional filter was included for deletion calls. Any deletions with a breakpoint occurring within 500 base positions of a primer boundary were removed. According to Hjelm et al. 2019, this position filter removes false-positive detection calls that occurred due to incorrect primer binding and putative deletions that occur within the spike in coverage that can occur near the site of LR-PCR primers. For “NanoDel” (Pipeline 1), putative deletions needed to be supported by 2 reads or more to be reported, or present in >0.1% of reads to be called. “Wood ONT” (Pipeline 3) and “Wood Illumina” (Pipeline 4) made use of the assembly tool Canu (v2.2, (56)(57)), specifically the correction and trimming functions. These functions replaced noisy read sequences with consensus sequences computed from overlapping reads, to decide which regions contain “high-quality” sequences, trimming the read to retain the largest “high-quality” chunk of sequence.

Cumulative deletion metrics:

To investigate the cumulative effect of LSMDs (normalised to the benchmark coverage) within a given sample, three metrics were calculated (27): 1) Cumulative read % is the summation of the read %s, for all deletions called in that sample, 2) Number of deletions per 10k coverage is the number of unique deletions identified, divided by the benchmark coverage multiplied by

10,000, and 3) Average % burden per deletion is the cumulative read %, divided by the number of unique deletions.

### Repeat and G4 forming sequence and structure analysis

Selected sequences flanking the 5'- and 3'-breakpoints of the LSMDs detected by "NanoDel" (Pipeline 1) were analysed for repeat sequences and their potential to form G-quadruplexes (G4). G4s are four-stranded nucleic acid structures formed by guanine-rich sequences in DNA and RNA. They are non-canonical structures, meaning they deviate from the typical double-helix structure of B-form DNA. G4s are formed by Hoogsteen hydrogen bonding of four guanine bases in a planar arrangement called a G-quartet, stabilised by a cation (e.g., potassium), which then stack on top of each other to form the four-stranded structure. They can exist in different topologies depending on the orientation of the guanines within the quartet, and can be found in various genomic locations, including telomeres, promoters, untranslated regions of mRNA, and mtDNA. G4s are implicated in diverse biological processes, including DNA replication, transcription, translation, and genome stability. Samtools (46) faidx was employed to extract 50bp for the repeat sequence analysis or 100 bp for the G4 analysis on either side of the respective 5'- and 3'-breakpoints from the rCRS (which is displayed as the L-strand), respectively.

For the repeat sequence analysis, the extracted sequences flanking the 5'- and 3'-breakpoints were aligned using ClustalW (58). Regions containing the longest stretches of contiguous perfect or imperfect repeated sequence were identified visually from the alignments and reported. For the G4 analysis, the reverse complements (representing the H-strand) of these extracted sequences were generated using the online Java Sequence Manipulation Suite (59). The QGRS Mapper tool was used to predict the G4-forming sequences from both the original and reverse complement sequences, applying default settings of a maximum length of 30 bases and a minimum G-group size of 2 (60). QGRS Mapper provided either single or multiple G4-forming sequences along with a G-score for each sequence indicating its likelihood to form G4 structures. The G-score is based on the length and evenness of loop regions and the number of guanine tetrads, with higher (>30) scores suggesting a strong likelihood to form. A sequence of 30 bp could have a maximum score of 105, corresponding to the sequence 5'-GGGGGGTGGGGGGTGGGGGGTGGGGGG-3'. Where multiple putative G4-forming sequences were found, the one with the highest G-score was reported and chosen for investigation using AlphaFold 3 (AF3) in the presence and absence of a K+ cations. The formation of G4s is dependent on monovalent cations, including K+, which is often seen at high (~140mM) intracellular concentration (61). Where only a single G4-forming sequences was reported, this too was investigated using AF3. AF3 is the most recent version of the Google DeepMind's artificial intelligence modelling system is now able to generate 3D structures using nucleic acids sequences not just proteins (62).

### Previous tissue/disease association analysis

Repeat regions are often used to annotate LSMDs. For example, mtDNA4977 has two flanking 13-bp direct repeats (8470-8482 and 13447-13459). However, whether the first or last base of the repeats is used for annotation can lead to slight inconsistencies in numbering. For the 5

441 LSMDs investigated in more detail (this study), we took the first base of both the 5'- and 3'-of
442 their associated repeats and presented adjusted positions (Table S8). Previous tissue and/or
443 disease associations of the LSMDs, both "NanoDel" (Pipeline 1) and adjusted positions, and
444 10bp flanking either side were investigated using the Mitobreak database, which contains 1369
445 human mtDNA deletions previously described in the literature (63).

446 Statistical and graphical analysis

447 Statistical analyses of post-alignment read metrics were conducted using GraphPad Prism
448 (v9.4.1). The normality of each metric for each pipeline was assessed using the Shapiro-Wilk
449 test (64). Where assumptions of normality were met, a Brown-Forsythe and Welch ANOVA
450 was employed to model differences between the approaches. If the assumption of normality
451 was violated, the non-parametric Kruskal-Wallis test (65) was used instead. Further post-hoc
452 pairwise comparisons were made using either Welch's t-test or the non-parametric Mann-
453 Whitney test, respectively.

454 Due to high variability of the cumulative deletion metrics between glioblastoma samples and
455 the availability of other matched data, which varied between samples, including CIV activity
456 and mtDNA-CN, cluster analyses were performed using each metric either alone or in
457 combination with other metrics using PAST (v4.03, (66)) classical clustering function. The
458 Shapiro-Wilk test was used to evaluate the assumption of normality of datasets due to sample
459 sizes, and appropriate models were used to determine differences between clusters as above.

460 All survival analyses were conducted using GraphPad Prism (v9.4.1). Overall survival was
461 defined as the time from diagnosis (in months) to death (TTD). Time from diagnosis (in
462 months) to last follow up was used for including and censoring individuals that were recorded
463 as alive at the end of the study period. Among the cohort analysed, there was 8 males and 10
464 females, 7 unmethylated-MGMT (U-MGMT), 8 methylated-MGMT (M-MGMT), and not
465 determined-MGMT (n.d.-MGMT), and the age at diagnosis ranged from 18 to 78 years old
466 (Table S2C). All graphs were generated using GraphPad Prism 9. In all case, where multiple
467 values were combined and averaged, they were presented as mean ± SD.

468

469 **Results:**

470 NanoDel most accurately identifies the mtDNA4977 at low levels in simulated data

471 The pipeline developed for this study, "NanoDel" (Pipeline 1), was designed to detect the
472 location of LSMD breakpoints and the percentage of reads in which each LSMD occurs. This
473 was compared against three other pipelines: "Splice-break" (Pipeline 2, (27)), "Wood ONT"
474 (Pipeline 3, (36)), and "Wood Illumina" (Pipeline 4, (36)).

475 Relative read percentage (percentage of all reads in which the LSMD was found relative to
476 benchmark region coverage, read%) was used here as a proxy for heteroplasmy. This metric
477 was used to assess the sensitivity of each pipeline in its ability to detect LSMDs at different
478 read percentage levels, as well as their accuracy when detecting the correct locations of 5'- and
479 3'-breakpoint pairs for each respective LSMD. Artificial data 1 were designed to reflect

480 controlled ratios of reads containing mtDNA4977 (m.8471-13449del) found in glioblastoma
481 (24) and wild-type mtDNA, recapitulating the effects of heteroplasmy in silico. Visual
482 inspection of coverage traces revealed that the expected LSMD was visibly identifiable at read
483 % levels of ≥5% using all four pipelines (Fig 1 A-D).

484 *Fig 1. Analysis of artificial data set 1 that contains a mixture of the mtDNA4977 found in*
485 *glioblastoma (m. 8471-13449del) and wild type mtDNA at different ratios. A-D) The per*
486 *position levels of coverage generated by Pipeline 1 ("NanoDel"), Pipeline 2 ("Splice-break"),*
487 *Pipeline 3 ("Wood ONT"), and Pipeline 4 ("Wood Illumina"), respectively. E) The percentage*
488 *of the expected number (7500 (75%), 5000 (50%), 500 (5%), 100 (1%), 50 (0.5%) and 25*
489 *(0.25%)) of deleted reads detected out of 10,000 identified by each pipeline. F) Breakpoint*
490 *mapping locations identified at each deleted read% level (75%-0.25%) in relation to the true*
491 *breakpoint locations (highlighted in black).*

492 "NanoDel" (Pipeline 1) and "Wood Illumina" (Pipeline 4) were the most sensitive to low
493 heteroplasmy LSMDs, as they were the only two pipelines that identified m.8471-13449del
494 within 20 base positions of its true 5'- and 3'-breakpoints at all simulated read% levels (as low
495 as 0.25%). "NanoDel" (Pipeline 1) was able to identify 98.4% ± 22.9 (mean ± SD) of the
496 simulated reads containing a LSMD at each respective deleted read% level (Fig 1 E), with an
497 average sensitivity of 0.926 ± 0.180, whilst maintaining a specificity of 0.992 ± 0.011 across
498 all deleted read% levels. "Wood Illumina" (Pipeline 4) identified 138.5% ± 28.5 of deleted
499 reads on average, over-representing the true level of the LSMD, with a higher average
500 sensitivity of 0.995 ± 0.013, but also a lower specificity of 0.976 ± 0.043. In contrast "Splice-
501 break" (Pipeline 2) and "Wood ONT" (Pipeline 3) both failed to detect LSMDs at 0.25 read %
502 levels (Fig 1 E).

503 The ability of each pipeline to accurately detect the true location of different LSMD
504 breakpoints was also determined (Fig 1 F). "NanoDel" (Pipeline 1)"NanoDel" (Pipeline 1) was
505 the most accurate at detecting 5'- and 3'-breakpoints in the least distance from the true
506 simulated genomic position. "NanoDel" was able to identify each 5'- and 3'-breakpoint of
507 m.8471-13449del within 1 ± 0 nucleotide position, representing the LSMD as m.8470-
508 13450del at all deleted read% levels (Fig 1 F). This was due to "NanoDel" implementing 1-
509 based rather than a 0-based indexing system. "Wood Illumina" (Pipeline 4) identified two pools
510 of LSMDs with similar 5'-breakpoints but differing 3'-breakpoints, but breakpoint calls made
511 by "Wood Illumina" (Pipeline 4) were 26.5 ± 35 away from their true genomic position (Fig 1
512 F). "Splice-Break" (Pipeline 2) was the second most accurate at detecting the true breakpoint
513 locations within 1.3 ±0.19 positions, in deleted read% levels of ≥0.25% (Fig 1 F). Similarly,
514 "Wood ONT" (Pipeline 3) was able to determine breakpoints within 4.2 ± 3.9 positions, in
515 deleted percentage levels of ≥5% (Fig 1 F).

516 NanoDel most accurately identifies breakpoints and heteroplasmy levels in mixed
517 deletion populations in simulated data

518 The second set of artificial data were designed to reflect a more complex mixture of four
519 (including some overlapping) LSMDs (which were deleted between positions m.8471-
520 13449del, m.6620-8440del, m.3451-5500del and m.12538-14572del), in 10% of reads,

521 respectively (Fig 2 A-D). This was done to recapitulate more complex biological samples.
522 Drops in coverage were visible within deleted loci following the alignment stages of each
523 pipeline (Fig 2 A-C), except for "Wood Illumina" (Pipeline 4, Fig 2 D). The coverage generated
524 by "Wood Illumina" (Pipeline 4) remained static, making it difficult to deduce the presence of
525 LSMDs visually. However, using "NanoDel" (Pipeline 1), "Splice-Break" (Pipeline 2) and
526 "Wood ONT" (Pipeline 3), stepwise drops in coverage were visible due to the presence of
527 overlapping LSMDs. In general, these overlapping LSMDs obscure breakpoint boundaries
528 making it difficult to discern their true location through visual inspection alone (Fig 2 A-D).

529 *Fig 2. Analysis of artificial data set 2 that contains a mixture of deletions (i.e., between*
530 *positions m. 8471-13449del, m. 6620-8440del, m. 3451-5500del and m. 2538-14572del), set at*
531 *a read level of 10% and 90% wild type mtDNA. A-D) The per position levels of coverage*
532 *generated by Pipeline 1("NanoDel"), Pipeline 2 ("Splice-break"), Pipeline 3 ("Wood ONT"),*
533 *and Pipeline 4 ("Wood Illumina"), respectively. Breakpoint locations are denoted by black*
534 *dashed lines. E) The average number of deleted reads detected by each pipeline; each one is*
535 *simulated out of 1250 reads. The percentage deleted reads detected by deletion reads (so reads*
536 *detected out of 1250*100). Breakpoint mapping locations of each deletion identified in relation*
537 *to the true breakpoint locations (highlighted in black). * = P<0.05, ** = P<0.01, *** =*
538 *P<0.001, and **** = P<0.0001.*

539 All pipelines, except for "Splice-Break" (Pipeline 2), could identify all 4 LSMDs within the
540 mixed LSMD profile of artificial data 2. "Splice-Break" (Pipeline 2) was unable to identify the
541 LSMD m.3451-5500del but identified the remaining 3 (Fig 2 F). "NanoDel" (Pipeline
542 1)"NanoDel" (Pipeline 1) and "Wood ONT" (Pipeline 3) were able to detect 97.2% ± 18.7 and
543 96.1% ± 19.2 of deleted reads, respectively (Fig 2 F), with no significant differences between
544 the pipelines (P>0.99 and P>0.97 respectively, Fig 2 E). "NanoDel" (Pipeline 1) only displayed
545 a false positive detection rate of 0.01 and a false negative detection rate of 0.08, while "Wood
546 ONT" (Pipeline 3) displayed the same false positive rate of 0.01 but a false negative detection
547 rate of 0.09. Both Illumina centric pipelines (Pipelines 2 and 4), identified 36.7% ±24.5 and
548 144.2% ±3.6 of deleted reads on average, respectively (Fig 2 F) showing significant differences
549 from the expected (p=0.0001 and p=0.001, respectively) (Fig 2 E). "Splice-Break" (Pipeline 2)
550 displayed a much higher false negative rate of 0.63 while "Wood Illumina" (Pipeline 4)
551 displayed a higher false positive rate of 0.06.

552 "Splice-Break" (Pipeline 2) was the most accurate at determining the correct locations of 5'-
553 and 3'-breakpoints, identifying them within 1 ± 0 (mean distance from simulated breakpoint ±
554 SD), however, it was only able to identify 3 out of 4 LSMDs (m.6620-8440del, m.8471-
555 13449del, m.12538-15472del). "NanoDel" (Pipeline 1) was able to identify all LSMDs within
556 artificial data 2, within 1.25 ± 1 nucleotide positions of their respective 5'- and 3'-breakpoints
557 on average. In comparison, "Wood ONT" (Pipeline 3) and "Wood Illumina" (Pipeline 4)
558 identified them within 1.38 ± 0.7 and 3.13 ± 1.5, respectively. "Wood Illumina" (Pipeline 4)
559 also identified a small sub-population of 4 separate LSMDs with the same corresponding 5'-
560 breakpoints as those simulated but differing 3'-breakpoints 82.2 ± 2.6 nucleotide positions
561 away. This sub-population of LSMDs was not included in the previous average distance in

562 which the pipeline could detect 5'- and 3' -breakpoints (Fig 2 G) as these were deemed to be
563 different LSMDs resulting from false positive detection.

564 LSMDs revealed in clinically diagnosed mitochondrial disease samples following different
565 LR-PCR approaches

566 Due to the sensitivity, consistency, and accuracy of "NanoDel" (Pipeline 1) in detecting
567 LSMDs within the artificial data, "NanoDel" was then employed to detect LSMDs and
568 determine their breakpoints in two small human subject cohorts. MtDNA breakpoints were also
569 determined in these samples using "Wood ONT" (Pipeline 3) for comparison, both with and
570 without 'Canu'. The first cohort contained DNA from 3 tissue samples (muscle, kidney and
571 cerebellum) from a KSS patient (Table S2B), and the second cohort contained DNA from 5
572 samples (ML cultured cells, MM muscle, PS muscle, LD cultured cells, and PM muscle) from
573 different mitochondrial disease patients (Table S2B).

574 Following two-amplicon LR-PCR, full length wild type mtDNAs were visible as two-
575 amplicons at the expected sizes of 8.5kb and 9kb in the muscle and cerebellum from the KSS
576 patient, of cohort 1, but no smaller amplicons suggestive of deleted mtDNAs were visible (Fig
577 S2A-B). Full length wild type mtDNAs were also visible as two-amplicons at the correct size
578 in 4 out the 5 samples (ML cultured cells, MM muscle, PS muscle, LD cultured cells, but not
579 PM muscle due to the lack of material being available) from the patients diagnosed with
580 mitochondrial disease in cohort 2. However, unlike cohort 1, smaller amplicons suggestive of
581 deleted mtDNAs were visible in two samples (MM and ML) but not LD and PS using the first
582 primer set, and only in 3 samples MM, LD, and PS, but not ML using the second primer set
583 (Fig S2A-B).

584 Following one-amplicon LR-PCR of the KSS samples, full length mtDNAs were visible as
585 one-amplicon at the expected size (16.3kb), as were shorter amplicons suggestive of deleted
586 mtDNAs, in muscle and cerebellum (Fig S2C). Full length mtDNAs were also visible as one-
587 amplicon in LD, PS and PM along with multiple smaller amplicons suggestive of deleted
588 mtDNAs. MM and ML appeared to contain either multiple deleted mtDNAs or a single
589 homoplasmic deleted mtDNA, respectively, as no full length mtDNAs were visible (Fig S2C).

590 NanoDel yields lower error rates, higher mean coverage and higher maximum read
591 lengths in pipeline comparisons

592 The read metrics from the two and one-amplicon LR-PCRs were initially combined after
593 alignment of the ONT reads with "NanoDel" (Pipeline 1) and "Wood ONT" (Pipeline 3), read
594 quality and length filtering (Table S4A, B). We observed overall that the average mapping rate
595 percentage (Fig 3A), using "Wood ONT" (Pipeline 3, with Canu) (99.2% ± 1.8, mean ± SD)
596 was significantly higher than that which was observed using both "NanoDel" (Pipeline 1) (76.6
597 ± 30.2, p=0.001) and "Wood ONT" (Pipeline 3, without Canu) (75.0 ± 29.2, p=0.0001). No
598 significant differences were identified between "NanoDel" (Pipeline 1) or "Wood ONT"
599 (Pipeline 3, without Canu). However, this increased mapping rate appears to have come at the
600 cost of average coverage as "Wood ONT" (Pipeline 3, with Canu) (58.6 ± 31.5) showed
601 significantly lower levels of coverage compared to "NanoDel" (Pipeline 1) (1662.7 ±

932.0,p=0.0001) and “Wood ONT” (Pipeline 3, without Canu) (1759.1 ± 982.1, p=0.0001), which is most likely due to Canu’s correct and trim functions removing the vast majority of reads. Again, no significant differences were determined between “NanoDel” (Pipeline 1) and “Wood ONT” (Pipeline 3, without Canu) (Fig 3C). “NanoDel” (Pipeline 1) was found to have the lowest average error rate of the three approaches; however, it was not significantly lower than the others (Fig 3B). No significant differences in the average read lengths aligned were identified between each of the pipelines (Fig 3D). However, significant differences were determined between the maximum read lengths aligned. “Wood ONT” (Pipeline 3, with Canu) on average aligned reads with a shorter maximum length (10775 ± 3499) compared to “NanoDel” (Pipeline 1) (19276 ± 4592, p=0.0001), and “Wood ONT” (Pipeline 3, without Canu) (19276 ± 4592, p=0.0001) (Fig 3E).

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

NanoDel and one-amplicon LR-PCR has an increased mapping rate and lower error rate than Wood ONT

Comparisons between the sequencing and alignment quality metrics following one-amplicon and two-amplicon LR-PCR were conducted to assess the impact of these approaches on mapping rate, error rate, coverage, and read length. The proportion of reads mapped to the reference genome were significantly higher following one-amplicon LR-PCR using “NanoDel” (Pipeline 1) (85% ± 31.1%) compared with two-amplicon LR-PCR (65.7% ± 27.6, p=0.02, Fig 3F). Similarly, “Wood ONT” (Pipeline 3, without Canu) exhibited a significantly higher mapping rate following one-amplicon LR-PCR (82% ± 29.6), compared to two-amplicon LR-PCR (65.3 ± 27.6, p=0.04, Fig 3P). No significant differences were identified between one-amplicon LR-PCR and two-amplicon LR-PCR when using “Wood ONT” (Pipeline 3, with Canu) as 99.14% ± 2.1 and 99.14% ± 1.4 of reads were mapped respectively (Fig 3K).

The error rate using “NanoDel” (Pipeline 1) following one-amplicon LR-PCR (0.03 ±0.003) was significantly higher compared to two-amplicon LR-PCR (0.02 ± 0.002, p=0.02, Fig 3G). Similarly, the error rate using “Wood ONT” (Pipeline 3, without Canu) following one-amplicon LR-PCR (0.13 ± 0.13) was significantly higher compared to two-amplicon LR-PCR (0.02 ± 0.001, p=0.03, Fig 3Q), but not when using “Wood ONT” (Pipeline 3, with Canu)(Fig 3L). While the error rate using “NanoDel” (Pipeline 1) following one-amplicon LR-PCR was significantly higher than the two-amplicon LR-PCR approach, the magnitude of the increase was low (Fig 3G). Furthermore, “NanoDel” (Pipeline 1) error rate was still lower (Fig 3G) than the error rates seen using “Wood ONT” (Pipeline 3, with and without Canu (Fig 3L and 3Q). All pipelines also generated a higher average coverage following two-amplicon LR-PCR.

643 However, the increases were not considered significant in comparison to the coverage
644 generated following one-amplicon LR-PCR (Fig 3H, M, R). The increase in coverage
645 following two-amplicon LR-PCR was expected due to the overlapping nature of the two-
646 amplicons causing specific regions to have 2x more coverage. Unsurprisingly, both mean and
647 maximum read lengths are higher following one-amplicon LR-PCR for each pipeline however
648 in each incidence the increase is not considered significant (Fig 3 I, N, S, and J, O, T,
649 respectively).

650 NanoDel and one-amplicon LR-PCR best validates known LSMDs in KSS patient
651 tissues

652 In addition to the differences between the quality metrics (Fig 3, Table S4A-B), visible
653 differences were also apparent upon inspection of the coverage traces generated post-alignment
654 using reads from the one or two-amplicon LR-PCR of the KSS samples. Homogenous drops
655 in coverage, indicative of a singular LSMDs, were clearly visible at ~m.6000-15000 following
656 one-amplicon LR-PCR using both "NanoDel" (Pipeline 1, Fig 4A) and "Wood ONT" (Pipeline
657 3, without Canu, Fig 4B). This was unlike those amplified using the two-amplicon LR-PCR
658 approach where increases in coverage could also be seen in the regions ~m.6500-7000del and
659 ~m.1500-15500del, using both "NanoDel" (Pipeline 1) and "Wood ONT" (Pipeline 3, without
660 Canu) (Fig 4C and D, respectively). These spikes were presumed to be the result of overlapping
661 ends of PCR products causing inflations in coverage. As such, drops in coverage appeared to
662 be between ~m.7000-15000del. In all instances no drops in coverage were evident using
663 "Wood ONT" (Pipeline 3, with Canu), due to low overall coverage (Fig B and D). Using
664 "NanoDel" (Pipeline 1), under stringent filters including removing deletions called within
665 ±500bp of a primer boundary, and focussing on a read percentage level of 5% or more, just one
666 deletion (m.8469-13446del) was identified in one KSS tissue (kidney) using two-amplicon LR-
667 PCR (Table S5A). In the absence of filtering however, an additional 33, 16 and 6 deletions
668 were identified in the muscle, kidney, and the cerebellum samples, respectively (Table S5A).
669 Using the same stringent filters, one deletion (m.6130-15055del) was identified using
670 "NanoDel" (Pipeline1) in all three tissues (muscle, kidney, and cerebellum) at differing read
671 levels relative to the benchmark (112%, 57%, and 92%, respectively) of the KSS patient
672 following one-amplicon LR-PCR (Table S5A). These were consistent with the coverage traces
673 for these samples, and the breakpoints previously determined in these samples via Southern
674 blotting (Poulton et al., 1993, Poulton et al., 1989, Poulton et al., 1995, Table S2B). On the
675 other hand, the average read % of m.6130-15055del was higher in the muscle, kidney and
676 cerebellum using "NanoDel" and not correlated with the heteroplasmy % previously
677 determined using Southern blotting (47%, 54% and 23%, respectively). In the absence of
678 stringent filtering, a further 2, 18, and 6 deletions were called using "NanoDel" in the muscle,
679 kidney, and cerebellum samples of the KSS patient (Table S5A).

680 *Fig 4. Analysis of coverage and deletion boundaries in different tissues of a KSS patient. A-D)*
681 *The levels of coverage displayed as read depth at each position for each sample amplified as*
682 *either one A, B) or two C, D) amplicons, aligned with Pipeline 1 ("NanoDel", A, C) or Pipeline*
683 *3 ("Wood ONT", B, D). E-L). Circos plots showing the 5'- and 3'-breakpoints of deletions*
684 *following one E-H) and two I-L) amplicon LR-PCR aligned with Pipeline 1 ("NanoDel") post-*

*filtering E, I) and unfiltered F, J) and Pipeline 3 ("Wood ONT") with G, K) and without H, L) Canu. The genomic links highlight the connected 5'- and 3'-breakpoints coloured by sample. Links proceed in a clockwise direction starting from the D-loop 1 region.*

Using "Wood ONT" (Pipeline 3, with Canu), and a read % of 5% or above, a set of 'breakends' spanning positions m.1-16568del were identified in each sample (muscle, kidney, and cerebellum) (see Table S5B) following two-amplicon LR-PCR. It was suspected that these sets of 'breakends' signified the ends of the reads relative to the rCRS reference, as the amplicon boundaries did not align directly with the rCRS. This increased to two deletions per sample when Canu filtering was removed (m.6835-6932del and m.6969-7068del in muscle, m.6578-8878del and m.6740-8699del in kidney, and m.6612-6714del and m.6745-6888 in cerebellum) (Table S5B). Consistent with "NanoDel" (Pipeline 1), deletion (m.6130-15055del) was also identified using "Wood ONT" (Pipeline 3 with Canu) in all three tissues (muscle, kidney, and cerebellum) of the KSS patient following one-amplicon LR-PCR, in addition to a set of 'breakends' spanning positions m.1-16568del in kidney (Table S5B). Without Canu, m.6130-15055del was also apparent in the three tissues, and two additional deletions (m.216-16189del and m.315-16184del) were also apparent in the kidney sample of the KSS patient (Table S5B).

The read% determined for m.6130-15055del in muscle kidney and cerebellum were correlated between "NanoDel" (Pipeline 1) and "Wood ONT" (Pipeline 3 with (106, 24 and 65%) and without Canu (112, 57 and 93%).

The lack of amplicons following either one or two-amplicon LR-PCR (Fig S2), consistently poor mapping rate, average read length and coverage (Table S4) for the kidney sample across all pipelines, suggests poor quality, and the importance of high-quality DNA when interpreting LSMDs. While the lack of full-length amplicons after one-amplicon LR-PCR (Fig S2) for the MM and ML samples could also be suggestive of poor quality, this is unlikely due to the amplification of full-length amplicons post two-amplicon PCR (Fig S2), and consistent mapping rates, average read lengths and coverage of these samples across pipelines (Table S4).

Fig 4 show the 5'- and 3'-prime gene involvement of the deletions for filtered and unfiltered data, respectively, for "NanoDel" and "Wood ONT" (Pipeline 3), and that two-amplicon LR-PCR (Fig 4 I-L), but not one-amplicon LR (Fig 4 E-H), prevents deletion calling across the amplicon boundaries.

As the two-amplicon LR-PCR completely failed to detect the known deletion (m.6130-15056del) in the KSS samples (and cannot detect deletions across amplicon boundaries), subsequent analyses of the other mitochondrial disease samples focussed on "NanoDel" (Pipeline 1) and "Wood ONT" (Pipeline 3) using the one-amplicon LR-PCR.

NanoDel and one-amplicon LR-PCR yield new insights into LSMDs in mitochondrial disease patients

After alignment with "NanoDel" (Pipeline 1) and "Wood ONT" (Pipeline 3) without Canu, drops in coverage were visible following one-amplicon LR-PCR for all patients. Samples ML and LD showed stepwise drops in coverage characteristic of singular LSMDs. However, samples MM, PM and PS did not show distinct homogeneous drops in coverage necessitating

725 other tools for identifying breakpoints (Fig 5A and C). The exception to this is after alignment
726 with "Wood ONT" (Pipeline 3) with Canu, where no obvious drops in coverage were visible,
727 due to the poor overall coverage (Fig 5B).

728 *Fig 5. Analysis of coverage and deletion breakpoints in five samples clinically diagnosed with*
729 *mitochondrial disease. A-C) Shows the coverage at each position for each sample generated*
730 *with Pipeline ("NanoDel", A), and Pipeline 3 ("Wood ONT") with Canu, B) and without Canu*
731 *C), respectively. D-G) Shows the 5'- and 3'-breakpoints of deletions identified by each pipeline*
732 *in each sample as a circos plot following one-amplicon LR-PCR aligned with Pipeline 1*
733 *("NanoDel") post-filtering D) and unfiltered E) and Pipeline 3 ("Wood ONT") with F) and*
734 *without G) Canu. . The genomic links highlight the connected 5' and 3' breakpoints coloured*
735 *by sample. Links proceed in a clockwise direction starting from the D-loop 1 region.*

736 "NanoDel" (Pipeline 1) identified more LSMDs within this cohort compared to the number
737 identified using "Wood ONT" (Pipeline 3), with and without Canu. However, only 6 LSMDs
738 were found to occur after stringent filtering (m.800-9530del and m.496-567del in MM,
739 m.8469-13446del in LD, and m.5787-13921del, m.555-14749del, and m.3262-14412del in
740 PM) in 3 of the mitochondrial disease patients MM, LD and PM (Table S5A). When the
741 stringent filters were removed a total of 534 putative LSMDs were identified across all 5
742 samples with an average 106.8±114.2 LSMDs per sample (Table S5A). Only 19 had a read %
743 greater than 5%. Additional LSMDs identified with filtering removed (but a read level of 5%
744 or greater) included, LSMDs in MM (m.1666-16070del, m.2818-16068del, m.10184-
745 16074del, m.457-15889, m.1766-16070del, m.1646-16033del,1664-15971del, m.3166-
746 16072del, m.2059-16069del, m.2270-16070del), no additional LSMDs in LD, additional
747 LSMDs in PM (m.3270-16089del, and m.104-16071del), and a single LSMD in ML (m.8636-
748 16072del).

749 10 LSMDs were identified in 3 out of 5 samples (MM, LD and PM), using "Wood ONT"
750 (Pipeline 3, with Canu) occurring in 12.8% ± 4.6 of reads on average with 9 out of 10 LSMDs
751 occurring in 5% or greater reads including m.806-9519del, m.1664-15972del, m.1672-
752 16071del, and m.10186-16076del in MM, m.8469-13267del and m.8482-13447del in LD, and
753 m.562-14181del, m. 5789-13923del and m.6630-13994del in PM (Table S5B). This increased
754 to 35 LSMDs, of which 17 occurred in 5% or greater of reads identified by "Wood ONT"
755 (Pipeline 3, without Canu). This included additional LSMDs in MM (m.496-568del, m.460-
756 15890del, m.1668-16071del, m.1765-16071del, m.2058-16071del, m.2273-16071del, m.2819-
757 16071del, and m.3166-16071del), no additional LSMDs in LD, additional LSMDs in PM
758 (m.105-16071del, m.3264-14414del, m.3272-16068del, and m.5788-13918del), and a single
759 LSMD in ML (m.8648-16073del) (Table S5B).

760 These deletion profiles are consistent with the "single deletion type" determined sing
761 traditional approaches for LD and ML and the "multiple deletion type" determined for MM
762 and PM previously (Table S2B, Fig S2).

763 Previously, other methods could not determine the breakpoints of the deletions in MM,
764 although the major deletion observed by Southern blotting was sized at ~8.6kb (Table S2B).
765 "NanoDel" (Pipeline 1) offers novel insights in relation to MM that: (1) the major deletion is

766 m.800-9530del (8.7kb), (2) a further deletion (m.496-567del) is present but has historically
767 been missed, and (3) the deletions have a relative abundance (based on read%) of 19 and 11%,
768 respectively. The mtDNA profile observed is consistent with impaired mtDNA replication in
769 MM, with pathogenic variants in RNASEH1, encoding ribonuclease H1, causing autosomal
770 recessive progressive external ophthalmoplegia (PEO) with mtDNA deletions (67). RNASEH1
771 is an enzyme that degrades DNA-RNA hybrid structures such as those crucial for initiating
772 mtDNA replication.

773 While multiple deletions were previously observed in PM, their breakpoints were not
774 determined (Table S2B). "NanoDel" (Pipeline 1) suggests deletions at m.5787-13921del,
775 m.555-14749del, and m.3262-14412del each with a relative abundance of 13%, 8% and 5%,
776 respectively. The mtDNA profile observed is consistent with the diagnosis of thymidine
777 phosphorylase deficiency in PM, leading to thymidine build up and damage to mtDNA, which
778 is often linked with the co-occurrence of multiple deletions, depletion, and somatic mtDNA
779 point substitution due impaired mtDNA replication. PM's diagnosis of thymidine
780 phosphorylase deficiency, presenting as autosomal recessive mitochondrial
781 neurogastrointestinal encephalomyopathy (MNGIE)(68), was previously confirmed by the
782 detection of pathogenic variants in TYMP, encoding thymidine phosphorylase.

783 A single deletion m.8483-13459del was previously described in LD (Table S2B). "NanoDel"
784 (Pipeline 1) reported the exact 5'- and 3'-positions of the deleted segment as m.8469-13446del.
785 These are both in fact mtDNA4977, as both have a deletion size of almost 4977bp, and involve
786 regions around 8470-13459. In the former case (Table S2B) the last base of the 5'- and 3' -
787 repeats have been used for annotation. "NanoDel" also suggests a relative abundance of 79%
788 (relative to benchmark region coverage) which had not previously been determined.

789 Similarly, a single deletion m.8649-16084del was previously detected at ~72-76%
790 heteroplasmy by qPCR in ML (Table S2B). The "NanoDel" (Pipeline 1) called very similar
791 breakpoints m.8636-16072del (but only without stringent filtering), at a higher relative
792 abundance of 100%. While a diagnosis was not noted for ML, according to MitoBreak (the
793 mtDNA breakpoints database (63), m.8649-16084del is associated with ageing tissues (69)
794 (70). While there is not an exact match for m.8636-16072del (7436bp), other LSMDs with
795 breakpoints located at m.8636 have been found in epidermal keratocytes irradiated with UVB
796 (71) and m.16072 has 57 hits in MitoBreak, with clinical diagnoses including mitochondrial
797 myopathy and PEO (72)(73).

798 A total of 207 LSMDs were identified in sample PS using "NanoDel" (Pipeline 1), all with a
799 relative abundance <5%. The LSMDs with the highest read percentage levels: m.5788-
800 13922del, m.5790-13922del and m.6327-13990del were found to have relative abundances of
801 4.2%, 3% and 2.3%, respectively. This was consistent with the range of LSMD sizes identified
802 via Southern blotting and a multiple deletion type (Table S2B) and gel electrophoresis
803 following one-amplicon LR-PCR (Fig S2), the impaired mtDNA replication often observed in
804 patients like PS, who was previously diagnosed with autosomal dominant twinkle (TWNK)
805 related Progressive External Ophthalmoplegia with mtDNA deletions (TWNK)(74). TWNK is
806 a mtDNA helicase, which unwinds ds-mtDNA necessary for mtDNA replication.

807 NanoDel and one-amplicon LR-PCR reveals heterogeneous low level LSMDs in
808 glioblastoma patients

809 A third cohort containing 20 samples from 18 glioblastoma and 2 normal brain control tissues
810 was analysed using LR-PCR (Table S2C). Initially, two-amplicon LR-PCR was performed on
811 these samples. Full length wild type mtDNAs were successfully amplified using the first primer
812 set in samples 1-5, 7-10, 12-24 but not 6 or 11 (Fig S3A). Using the second set of primers, full
813 length mtDNAs were successfully amplified in 1-5, 7, 9-24 but not 6 or 8. Additionally, with
814 both primer sets, most samples amplified contained multiple smaller bands (Fig S3B). The
815 presence of multiple bands (either full length or smaller) is indicative of LSMD heteroplasmy.
816 The absence of full-length PCR products in some samples is likely due to the two-amplicon
817 LR-PCR primers being unable to bind in deleted regions.

818 Following one-amplicon LR-PCR, full length mtDNAs were visible in all samples but for most
819 samples additional smaller bands were less evident compared to two-amplicon PCR (Fig S4).
820 Due to the inability of two-amplicon LR-PCR to detect known deletions in the validation tests,
821 only one-amplicon PCR amplicons were subjected to ONT sequencing and analysis with our
822 preferred "NanoDel" (Pipeline 1).

823 Glioblastoma samples yielded an average of 3119 ± 3837 reads, mapping rate of 51 ± 29%,
824 read length of 3253 ± 2465 and coverage of 331 ± 372 per sample. While normal brain samples
825 generated a similar number of reads (3083 ± 1946, p=0.13), mapping rate was higher (93 ± 5%,
826 p=0.04), read length was not statistically longer (6483 ± 3636, p=0.81) and coverage was not
827 statistically higher (878 ± 554, p=0.27) per sample than in glioblastoma samples (Table S6).

828 Using "NanoDel" (Pipeline 1), and filtering ± 500bp within the primer boundary, 171 putative
829 deletions were identified in glioblastoma samples, with an average read % of 1.60 ± 3.82%
830 (Table S7, Fig 6A). This decreased to just 10 with an average read % of 11.9% ± 11.6% when
831 looking at read% levels of 5% or more (Fig 7A). This contrasts with just 38 deletions detected
832 within the normal brain sample controls (although there are also fewer samples), at a lower
833 read % (0.73 ± 0.75%)(Table S7, Fig 6B), with no deletions detected above read% levels of 5
834 or more (Fig 7A, p=0.19).

835 *Fig 6. Analysis of deletion breakpoints identified in glioblastoma and normal brain tissue*
836 *samples following one-amplicon LR-PCR aligned with Pipeline 1 ("NanoDel") prior to*
837 *filtering. A, B) Circos plots show the 5'- and 3'-breakpoints of deletions identified by each*
838 *pipeline in glioblastoma A) and normal brain B). The genomic links highlight the connected*
839 *5'- and 3'-breakpoints coloured by sample. Links proceed in a clockwise direction starting*
840 *from the D-loop 1 region. C-F) Bar graphs show the frequency of 5'- and 3'-breakpoints in C,*
841 *D) glioblastoma and in E, F) normal brain tissue samples.*

842 *Fig 7. Analysis of the A-D) deletion metrics and E-H) other matched data in glioblastoma*
843 *(n=18) and normal brain (n=2) tissue samples. A) The number of unique individual LSMDs*
844 *per sample. B) The cumulative deleted read % calculated as the total number of reads*
845 *containing LSMDs expressed as a % of benchmark coverage per sample. C) The deletions per*
846 *10K which is the number of individual LSMDs standardized for a read coverage of 10,000 per*

847 *sample. D) The average burden per deletion the cumulative read %, divided by the number of*
848 *unique deletions per sample. E) The mtDNA-CN per cell determined using qPCR in each*
849 *sample. F) The CIV activity per sample. G) The differences in the average number of LSMDs,*
850 *and the H) cumulative LSMD metrics and mtDNA-CN between glioblastoma and normal brain*
851 *tissue samples.*

852 Mt-co1 and mt-cyb harbour breakpoints most frequently in glioblastoma

853 Within the glioblastoma samples, 5'-breakpoints were identified in 18 out of the 38 mtDNA
854 regions, with the highest (in descending order) in mt-co1 (63), D-Loop (31), mt-rnr1 (20), mt-
855 nd1 (14), mt-rnr2 (7), mt-co2 (6), mt-tf (5), mt-atp6 (5), mt-tl1 (4), mt-atp8 (4), mt-nd2 (3), mt-
856 tv (2), mt-tS2 (2), mt-tq (2), mt-tc (1), mt-co3 (1), mt-nd4l (1), mt-nd4 (1). 3'-breakpoints were
857 only identified in 8 out of the 38 mtDNA regions, with the highest (in descending order) in mt-
858 cyb (94), mt-nd6 (42), mt-nd5 (29), mt-nd4 (3), mt-tm (1), mt-tl (1), mt-tl2 (1), mt-te (1) (Fig
859 6C,D). Within normal brain controls, 10 out of the 38 mtDNA regions contained 5'-breakpoints
860 and 6 out of the 38 regions contained 3'-breakpoints. The gene distribution of 5'- and 3'-
861 breakpoints were similar in glioblastoma vs. normal brain samples, with the most common (in
862 descending order) 5'-breaks being mt-co1 (10), mt-rnr1 (6), D-loop (5), mt-rnr2 (5), mt-nd1
863 (4), mt-tf (2), mt-tl1 (1), mt-nd2 (1), mt-co2 (1), mt-atp6 (1), and the most common (in
864 descending order) 3' -breaks being mt-cyb (21), mt-nd6 (9), mt-nd5 (4), mt-nd4 (2), mt-nd3
865 (1), mt-te (1) in normal brain controls (Fig 6E, F).

866 The most commonly occurring LSMDs in glioblastoma samples with a read% of 5% or more
867 had 5'-breakpoints between m.6600-6603 and 3'-breakpoints of either m.14646-14649 or
868 m.14825-14826. These LSMD "hot spots" occurred in 100% of glioblastoma samples, but also
869 in 100% of normal brain controls. The average read percentage of these LSMDs was higher in
870 glioblastoma (5.3 ± 9.3, range 0.94-41.68%) than in normal controls (2.44± 2.52, range 0.66±
871 4.22%), although this increase was not statistically significant (p=0.34) (Table S7).

872 Complex IV activity and mtDNA-CN are more prognostically relevant than cumulative
873 LSMD metrics

874 As the majority of LSMDs identified in the glioblastoma samples were present in a low
875 percentage of reads (116/171, <5%, Table S7), cumulative LSMD metrics were calculated:
876 cumulative deleted read%, deletions per 10k reads, and average burden per deletion and used
877 to explore the relationship of LSMDs with median overall survival (OS), CIV activity, mtDNA-
878 CN, age, MGMT status and gender in glioblastoma patients. The glioblastoma samples
879 exhibited inter-tumoral heterogeneity across all data types (Fig 7A-F, Table S2C).

880 An average of 10 ± 20 deletions was observed per glioblastoma sample (Fig 7A, G). This was
881 in contrast with the normal brain controls that contained 19 ± 4 deletions on average (Fig 7A,
882 G), although this difference was not statistically significant (p=0.10). The cumulative deleted
883 read% (14.9 ± 17.2 vs. 12.2 ± 17, p=0.99), and average burden per deletion (3.3 ± 5.3 vs. 0.5
884 ± 0.7, p=0.21) are both higher in glioblastoma compared to normal brain controls, albeit not
885 significantly (Fig 7H). While the deletions per 10k reads (333.6 ± 315.2 vs. 365.5 ± 399.7,

886 p=0.93) and mtDNA-CN (907 ± 760 vs. 1690 ± 1814, p=0.76) were also lower in glioblastoma
887 samples compared to normal brain controls, this was not significant (Fig 7H).

888 Due to the inter-tumoral heterogeneity across all data types, the glioblastoma samples were
889 clustered into groups using each parameter either alone or in combination with each other using
890 PAST's classical clustering. Three distinct groups were identified by cluster analysis using
891 cumulative deleted read %, distinguishable by significant differences in cumulative deleted
892 read% (p=0.001), deletions per 10k reads (p=0.001), average burden per deletion (p=0.001),
893 and mtDNA-CN (p=0.02) but not CIV activity (p=0.25) (Fig S5A-C):

894 (1) High cumulative read %: n=3 (samples 10, 18 and 21) with a cumulative deleted
895 read% of 31.5%, deletions per 10k reads of 625, average burden per deletion of 2.3, CN of
896 1462.2, Complex IV activity of 2.4, and time to death (TTD) of 22.7 months.

897 (2) Low cumulative read %: n=9 (samples 9, 20, 17, 13, 15, 16, 3, 2, 12) with a
898 cumulative deleted read% of 4.5%, deletions per 10k reads of 93.8, average burden per deletion
899 of 1.4%, CN of 1026.6, Complex IV activity of 3.8, and TTD of 13.3 months.

900 (3) Medium cumulative read %: n=5 (samples 7,11,1, 14 and 4) with a cumulative
901 deleted read% of 14.3%, deletions per 10k reads of 477.6, average burden per deletion of 4.5,
902 CN of 424.3, Complex IV activity of 4.5, and TTD of 6.5 months.

903 Integration of the above clusters with survival (TTD) data identified potential (but not
904 significant) differences (p=0.06) between median overall survival (OS) of cluster 1 vs. 2 (22.7
905 vs. 13.3 months, p=0.18), 2 vs. 3 (13.3 vs. 6.5 months, p=0.09) and 1 vs. 3 (22.7 vs. 6.5 months,
906 p=0.09).

907 Three distinct groups were identified by cluster analysis using deletions per 10k,
908 distinguishable by significant differences in cumulative deleted read percentage (p=0.001),
909 deletions per 10k reads (p=0.001), and mtDNA-CN (p=0.04) but not the average burden per
910 deletion (p=0.12) or CIV activity (p=0.49) (Fig S5D-F):

911 (1) High deletions per 10k reads: n=6 (7, 6, 21, 10, 18 and 14) with a cumulative deleted
912 read% of 31.8 %, deletions per 10k reads of 713.2, average burden per deletion of 6.32%, CN
913 of 943.4, Complex IV activity of 4.2, and TTD of 12.4 months.

914 (2) Medium deletions per 10k reads: n=3 (4, 11 and 1) with a cumulative deleted read%
915 of 15 %, deletions per 10k reads of 272.5, average burden per deletion of 4.9%, CN of 477.4,
916 Complex IV activity of 4.7and TTD of 4.8 months.

917 (3) Low deletions per 10k reads: n=9 (12, 13, 16, 2, 3, 9, 15, 17 and 20) with a
918 cumulative deleted read% of 4.3 %, deletions per 10k reads of 93.8, average burden per deletion
919 of 1.4%, CN of 1026.6, Complex IV activity of 3.8 and TTD of 13.3 months.

920 Integration of the above clusters with survival (TTD) data again identified potential (but not
921 significant) differences (p=0.08) between median OS of cluster 1 vs. 2 (12.4 vs. 4.84 months,
922 p=0.10), 2 vs. 3 (4.84 vs. 13.3 months, p=0.05) and 1 vs. 3 (12.4 vs. 13.3 months, p=0.58).

Three distinct groups were identified by cluster analysis using average burden per deletion, distinguishable by significant differences in the average burden per deletion (p=0.001), but not cumulative deleted read% (p=0.24), deletions per 10k reads (p=0.24), mtDNA-CN (p=0.09) and CIV activity (p=0.84)(Fig S5G-I):

(1) High burden: n=4 (11, 1, 14 and 18) with a cumulative deleted read% of 16.5 %, deletions per 10k reads of 391.4, average burden per deletion of 5.85%, CN of 473.4, and Complex IV activity of 3.3 and TTD of 12.4 months.

(2) Low burden: n=10 (10, 15, 12, 2, 4, 20, 16, 13, 3 and 21)) with a cumulative deleted read% of 11.7%, deletions per 10k reads of 218.1, average burden per deletion of 1%, CN of 1209.9, and Complex IV activity of 3.8 and TTD of 12 months.

(3) Medium burden: n=3 (7, 17 and 9) with a cumulative deleted read% of 7.4%, deletions per 10k reads of 453.7, average burden per deletion of 2.8%, CN of 584.7, and Complex IV activity of 4.2.

Integration of the above clusters with survival (TTD) data showed no significant difference in median overall survival time between the clusters (p=0.96).

Three distinct groups were identified by cluster analysis when all three cumulative metrics were analysed simultaneously, distinguishable by significant differences in cumulative read% (p=0.001), deletions per 10k reads (p=0.001) and mtDNA-CN (p=0.04) but not by average burden per deletion (p=0.12) or CIV activity (p=0.49 (Fig 8 A-C):

(1) High cumulative deleted read % and high deletions per 10k reads: n=6 (7, 6, 21, 10, 18 and 14)) with a cumulative deleted read% of 31.8%, deletions per 10k reads of 713.2 , average burden per deletion of 6.3%, CN of 943.4, Complex IV activity of 4.2, and TTD of 12.4 months.

(2) Medium cumulative read % and low deletions per 10k reads: n=3 (4, 11 and 1) with a cumulative deleted read% of 15%, deletions per 10k reads of 272.5, average burden per deletion of 4.9, CN of 421.7, and Complex IV activity of 4.3 and TTD of 4.8 months.

(3) Low medium cumulative read % and medium deletions per 10k reads: n=9 (12, 13, 16, 2, 3, 9, 15, 17 and 20) with a cumulative deleted read% of 4.3%, deletions per 10k reads of 93.9, average burden per deletion of 1.4%, CN of 1083, and Complex IV activity of 3.9 and TTD of 13.3 months.

*Fig 8. Analysis of clusters generated using A-C) the three cumulative LSMD metrics combined; cumulative deleted read %, deletions per 10k reads and average burden per deletion. A) Clustering generated 3 distinct groups. B) Kaplan-Meier curve showing borderline significant difference (P=0.08) between median OS of cluster 1 vs 2 (12.4 vs 4.8 months, P=0.10), 2 vs 3 (4.84 vs 13.3 months, P=0.05) and 1 vs 3 (12.4 vs 13.3 months, P=0.59) . C) Significant differences in cumulative read% (P=0.001), deletions per 10k reads (P=0.001) and mtDNA-CN (P=0.04) but not by average burden per deletion (P=0.12) or CIV activity (P=0.49) were identified between groups. D-F) MtDNA-CN. D) Clustering generated 2 distinct groups of 2 and 16 samples, respectively. E) Kaplan-Meier curve showing (P=0.02) between median OS*

962 *of cluster 1 vs 2 (25.8 vs 11 months). F) Significant differences in CN (P=0.02) and CIV activity*
963 *(P=0.03), but not cumulative deleted read% (P=0.47), deletions per 10k reads (P=0.47) or*
964 *average burden per deletion (P=0.21) were identified between groups. G-I) Complex IV*
965 *activity data. G) Clustering generated 2 distinct groups, containing 12 and 6 samples,*
966 *respectively. H) Kaplan-Meier curve showed a highly significant difference (P=0.0002)*
967 *between median OS of cluster 1 vs 2 (4.7 vs 15.6 months). I) Significant differences in CIV*
968 *activity (P=0.001) but not in cumulative deleted read% (P=0.07), deletions per 10k reads*
969 *(P=0.18), average burden per deletion (P=0.28) or mtDNA-CN (P=0.33) were observed*
970 *between groups. * = P<0.05, ** = P<0.01, *** = P<0.001, and **** = P<0.0001.*

971 Integration of the above clusters with survival (TTD) data also identified potential (but not
972 significant) differences (p=0.08) between median OS of cluster 1 vs. 2 (12.4 vs. 4.8 months,
973 p=0.10), 2 vs. 3 (4.84 vs. 13.3 months, p=0.05) and 1 vs. 3 (12.4 vs. 13.3 months, p=0.59) (Fig
974 8 A-C). Furthermore, age, MGMT status or gender did not differ significantly with these
975 combined cumulative deletions metric groupings (p=0.13, p=1.0, p=1.0), CN (p=0.3, p=0.48,
976 p=0.48), respectively) and CIV activity (p=0.14, p=1.0, p=1.0, respectively).

977 Two distinct groups were identified by cluster analysis using mtDNA-CN, distinguishable by
978 significant differences in CN (p=0.02) and CIV activity (p=0.03), but not cumulative deleted
979 read% (p=0.47), deletions per 10k reads (p=0.47) or average burden per deletion (p=0.21). (Fig
980 8D-F):

981 (1) High CN: n=2 (20 and 21) with a cumulative deleted read% of 24%, deletions per
982 10k reads of 453.7, average burden per deletion of 0.8%, CN of 2785.4, Complex IV activity
983 of 1.8, and a TTD of 25.8 months.

984 (2) Low CN: n=16 (4, 6, 12, 1, 9, 17, 18, 7, 14, 10, 11, 3, 15, 2, 13 and 16) with a
985 cumulative deleted read% of 14.1%, deletions per 10k reads of 314.6, average burden per
986 deletion of 4.0, CN of 672.6, Complex IV activity of 4.3, and a TTD of 11 months.

987 Integration of the above clusters with survival (TTD) data showed a significant difference
988 (p=0.02) between median OS of cluster 1 vs. 2 (25.8 vs. 11 months).

989 Two distinct groups were identified by cluster analysis using complex IV activity,
990 distinguishable by significant differences in CIV activity (p=0.001) but not in cumulative
991 deleted read% (p=0.07), deletions per 10k reads (p=0.18), average burden per deletion (p=0.28)
992 or mtDNA-CN (p=0.33) (Fig 8 G-I):

993 (1) High complex IV activity: n=6 (3, 7, 4, 1, 2 and 6) with a cumulative deleted read%
994 of 23.2%, deletions per 10k reads of 516.8, average burden per deletion of 6.9, CN of 638.3,
995 Complex IV activity of 7.4, and a TTD of 4.7 months.

996 (2) Low complex IV activity: n=12 (9, 11, 10, 13, 12, 14, 17, 16, 20, 21, 15 and 18)
997 with a cumulative deleted read% of 12.8%, deletions per 10k reads of 275.0, average burden
998 per deletion of 2.5, CN of 1015.9, Complex IV activity of 2.6, and a TTD of 15.6 months.

999 Integration of the above clusters with survival (TTD) data showed a highly significant
1000 difference (p=0.0002) between median OS of cluster 1 vs. 2 (4.7 vs. 15.6 months).

While older age (74.8± 2.8), unmethylated MGMT (n=7) and male gender (n=8) patients exhibited worse prognosis than younger (52.6± 6.11), methylated MGMT (n=8) and female sex (n=10) patients in our glioblastoma cohort, median OS was not significantly different between groups (age, 4.57 vs. 13.27 months, p=0.2; MGMT methylation, 13.27 vs. 14.11 months, p=0.47; sex, 13.2 vs. 14.1 months; p=0.44)(Fig S6).

LSMDs identified by NanoDel in mitochondrial disease and glioblastoma tissues lie adjacent to repeat sequences and putative G4 structure-forming sequences

Common features found to be associated with the formation of LSMDs are repeat sequences (27) and non-canonical secondary structures in DNA formed by guanine-rich sequences, known as G4-quadruplexes (G4s, (75)).

All the breakpoints identified for m.6130-15055del in the KSS samples, m.8469-13446del in LD, m.8636-16072del in ML, and m.6602-14648/14824del in glioblastoma by "NanoDel" fell within 2-18bp of a repeat sequence identified within the rCRS (Table S8). Similarly, all these breakpoints fell within 2-55bp of a putative G4-structure forming sequence identified by the QGRS Mapper with G-scores ranging from 5 to 20 identified within the rCRS (Table S8). The exception here was m.8469-13446del, where the 5'-breakpoint and the repeat sequence fell inside a putative G4-structure forming sequence. All the putative G4-structure forming sequences were found on the H-strand, except for the 3'-breakpoint associated with m.8636-16107del (Tabel S8).

Although the G-scores of the putative G4-forming sequences were considered low (<30), indicating they were unlikely to generate G4-structures under physiological conditions (60), we still sought to investigate their propensity to fold into G4 structures using AF3 (62).

AF3 predicted multiple G4 structures from sequence candidates we identified with the QGRS Mapper. Consistent with the low G-scores, the AF3 scores were also generally low, possibly due to an absence of these specific topologies in the PDB (76), Table S8). However, including ions appeared to increase the AF3 prediction scores as might be expected (77)(Table S8). One exception was the putative G4-sequence m.14806-14821, associated with the 3'-end of one of the most commonly occurring LSMDs detected in our glioblastoma samples, i.e. m.6602-14824del (Table S8). AF3 generated a convincing G4 model with 3 stacked planes of coordinated guanine tetrads (77)Fig 9A) that was predicted with high confidence (Table S8). NMR studies of homologous structures demonstrate that bases extending out (and with lower scores) with the main G4 core are likely to be relatively mobile (78)in solution While it will be important to experimentally validate this AF3-prediction, a search for similar structures revealed a very closely related topology from a G4 library study that was previously characterized by NMR (PDB: 6YY4, (79)Fig 9B).

*Fig 9. Structural prediction and comparison of G-quadraplex structures. AlphaFold 3 was used to predict the 3-dimensional structure of the putative G4-sequence m. 14806-14821, associated with the 3'-end of one of the most commonly occurring LSMDs detected in our glioblastoma samples (top row, A-D), compared with a known G-quadraplex NMR structure (PDB: 6YY4, Volek et al., 2021)(bottom row, E-H). A, E) The DNA sequence and topology diagram of each*

*structure with guanine bases depicted as blue rectangles with numbers corresponding to the sequence below (guanines colored blue), and arrows tracing the primary sequence from 5' to 3' using the Tetrado software (Adamczyk et al., 2023). B, F) The core structure of each G-quadraplex with three stacking G-quartets (colored orange, green, and pink) are shown in stick representation. K+ ions were modelled into the AF3 and NMR structures based on their observed propensity to stabilize G-quadraplexes in solution (Engelhart et al, 2004) and were found to occupy the spaces between planes (grey spheres), as seen in other quadraplex crystal structures. C, G) The full models of each sequence are shown, including flipped out bases that do not contribute to the G-quartets (lilac and yellow, respectively). D, H) Alternative side-on views (90° rotation around the x-axis) are shown to highlight similarities in the organisation of the 3 planar tetrads and intercalating K+ ions in each model. NMR studies of homologous structures demonstrate that bases extending out with the main G4 core are likely to be relatively mobile in solution (78).*

LSMDs identified by NanoDel in mitochondrial disease and glioblastoma tissues are associated with a range of healthy and pathological tissues

For the five deletions investigated in detail above, we provided adjusted "NanoDel" positions based on their associated repeats. We then used these ± 10bp for overlap analysis with the MitoBreak database. Interestingly, the deletions overlapped frequently with breakpoints in the MitoBreak database and were associated with both a range of tissues and clinical features, including (and in particular) aged, metabolically active (e.g. brain), and tumor tissues, and neurodegenerative diseases (Table S8).

**Discussion:**

Development, validation, and potential clinical application of NanoDel

In this study, we first developed and then evaluated the ability of four pipelines to accurately detect the location and quantify the read% of LSMDs using artificial data sets. These pipelines included both DNA and splice-aware RNA alignment methodologies and were designed based on either short read (Illumina) or long read (ONT) data sets. The pipelines selected for comparison against "NanoDel" (Pipeline 1, ONT splice-aware alignment, this study) were: "Splice-Break" (Pipeline 2, Illumina splice-aware alignment, (27)), "Wood ONT" (Pipeline 3, ONT DNA alignment, (36)) and "Wood Illumina" (Pipeline 4, Illumina DNA alignment, (36)). Splice-Break was chosen as a comparator as it uses Illumina short read sequencing in combination with a non-traditional splice-aware RNA alignment strategy that is not negatively affected by the presence of large-scale deletions, shown previously to improve LSMDs calling accuracy (27). While "Wood ONT" (Pipeline 3) and "Wood Illumina" (Pipeline 4) were chosen as they have previously been tested in a UK diagnostic setting and use either Illumina short- or ONT long-read sequencing, respectively, in combination with a more traditional DNA alignment strategy (36).

We demonstrate that our pipeline, "NanoDel" (Pipeline 1), which calls LSMDs from long-read data using a splice-aware alignment approach, offers improvements in accuracy and sensitivity

1081 over the other publicly available tools. This is true in terms of both breakpoint location and
1082 read% quantification of mtDNA4977 found in glioblastoma (m.8471-13449del, (24)) and
1083 mixed populations of deletions, including at low levels, in artificial datasets. "NanoDel"
1084 (Pipeline 1) identified the correct breakpoint locations and number of reads for all the LSMDs.
1085 "Wood ONT" (Pipeline 3, DNA alignment) performed most similarly to "NanoDel" (Pipeline
1086 1). This may be in part due to two factors: their use of a shared alignment tool 'Minimap2'
1087 (albeit using different settings, with "NanoDel" making use of the splice preset settings); and
1088 the fact that they both analysed the same style of long-read data compared to the other pipelines
1089 analysing 'short-read style' simulated data. In comparison, both "Splice-break" (Pipeline 2,
1090 Illumina RNA alignment) and "Wood Illumina" (Pipeline 4, DNA alignment) could not
1091 identify the LSMDs at the lowest 0.25 read% levels ("Splice-break"), or identified false
1092 positives ("Wood Illumina"), and were the least able to accurately quantify the correct number
1093 of reads containing LSMDs. Both these pipelines made use of short Illumina-style reads.

1094 While these findings align with separate prior reports emphasizing the importance of using
1095 splice-aware aligners in LSMDs calling (27) and the potential utility of ONT sequencing in
1096 mitochondrial genomics (36) (34) (35), they also highlight the potential strength of combining
1097 splice-aware alignment methodologies with long-read sequencing (i.e. "NanoDel") in LSMDs
1098 calling. Consequently, we wanted to deploy "NanoDel" (Pipeline 1), and "Wood ONT"
1099 (Pipeline 3) as a comparator, to investigate LSMDs in clinical samples from patients diagnosed
1100 with mitochondrial disease and varying amounts of LSMDs and compare with the information
1101 obtained using traditional techniques such as Southern blotting.

1102 In doing so, we found that "NanoDel", in combination with one rather than two-amplicon LR-
1103 PCR, outperformed the alternative "Wood ONT" in terms of mapping and error rate when
1104 investigating LSMDs in different tissues from mitochondrial disease patients. "NanoDel" in
1105 combination with one-amplicon LR-PCR was able to identify the previously observed LSMD
1106 (m.6130-15056del) (37) (31) (38) in KSS patient samples in a tissue specific manner.
1107 Furthermore, "NanoDel" was not only able to identify previously observed LSMDs in different
1108 tissues from five other patients diagnosed with mitochondrial disease, but also revealed
1109 additional, previously unreported LSMDs information. This demonstrated "NanoDel" has
1110 power to identify LSMDs in different sample types (tissues/conditions), including more
1111 complex biological samples (including those without a known genetic cause), and potential in
1112 a clinical setting.

1113 Our findings demonstrated that "NanoDel", in combination with one-amplicon LR-PCR,
1114 outperformed our two-amplicon approach. This is consistent with an original report that
1115 detailed the development and potential benefits of the one-amplicon LRPCR approach over
1116 multi-amplicon approaches, including increased discriminating power against nuclear
1117 mitochondrial DNA segments (NUMTs), more even coverage, and increased LSMD detection
1118 resolution by limiting amplicon dropout (80). NUMTs are pieces of mtDNA that have been
1119 transferred to and integrated into the nuclear genome and can cause false positives during PCR
1120 amplification. Consistent with the original report, we also showed that NUMTs with
1121 "NanoDel" were not amplified in our rho-zero mtDNA depleted controls with either one- or
1122 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 include (but are not restricted to) colorectal cancer (81), pancreatic cancer (82) and hypertension (83) 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 (36), only now is improved chemistry generating base-calling quality scores of Q20+ routinely. This, taken together with "NanoDel"'s (in combination with one-amplicon LR-PCR) ability to yield a sensitive and accurate, and thus a comprehensive view of the LSMDs breakpoints/landscape without a priori information in individual samples, offers improvements over traditional methods, in this context. However, the limited of correlation in read% between "NanoDel" (and "Wood ONT"), in combination with one-amplicon LR-PCR, and the heteroplasmy measures obtained using traditional methods could be due to the preferential amplification of shorter fragments during PCR in our samples (84). While such preferential amplification has been shown to be minimised via the use of LR-PCR (84), its impact could be verified in our samples by determining the ratio of deleted molecules to wild type molecules using digital PCR. Thus, for now, the read percentages should not be interpreted as heteroplasmy rates directly. Alternative PCR-free methods, e.g. purifying mitochondria/mtDNA prior to sequencing or WGS in combination with adaptive sequencing (where ONT pores screen for 'expected' sequences in silico, and reject strands that do not meet the expected requirements) may be used to increase sequencing yield of an intended target like mtDNA (85) , and may more accurately quantify heteroplasmy.. However, these methods also present challenges. It can be difficult to obtain suitable yields of purified mtDNA (often due to its fragility), and WGS can generate poor coverage of mtDNA for some samples even before adaptive sampling that down samples further.

NanoDel reveals novel findings in glioblastoma

Using "NanoDel" in combination with one-amplicon LR-PCR, we present a comprehensive catalogue of previously unreported LSMDs in glioblastoma tissue samples, revealing that LSMDs profiles are complex and heterogeneous. Indeed, a total of 171 LSMDs (mainly at low level, <5%) were identified across our glioblastoma cohort, significantly without any a priori information. These findings expand significantly over previous reports investigating LSMDs in glioblastoma, where either multi-amplicon PCR or targeted PCR was used (23) (24). Indeed, former studies may have misrepresented the true nature of the LSMDs in glioblastoma, due to the increased possibility of NUMTs being amplified, LSMDs not being amplified across amplicon boundaries, and because next/third generation sequencing was not subsequently employed (or other alternatives) to identify LSMD breakpoints or quantify heteroplasmy. The paper investigating a single LSMD did not investigate others present at all (24)

Using "NanoDel", we identified increased numbers of 5'-breakpoints in the gene mt-co1 and 3'-breakpoints in the gene mt-cyb, suggestive of LSMD hotspots. This is also consistent with

1164 the previous report investigating LSMDs in non-tumorigenic brain samples that also showed
1165 5'-breakpoints were most frequent in mt-co1 and 3'-breakpoints were also frequent in mt-cyb
1166 (with mt-nd5 being the most affected). Differences in sample number and sequencing and
1167 bioinformatics strategies likely contributed to the slightly different in observations, with (27)
1168 using short-read Illumina sequencing and their pipeline, "Splice-Break". Indeed, the most
1169 common LSMDs identified in our study centred around m.6600-6603 in mt-co1 with either
1170 breakends around m.14646-9 or m.14825-6 in mt-cyb, detected in all glioblastoma and healthy
1171 control samples. Specifically, m.6602-14649del was also identified at low levels (<0.03%) in
1172 64 of 93 non-tumorigenic brain samples in Hjelm et al.'s, (2019) study. The enrichment of
1173 breakpoints in mt-cyb and mt-co1 genes in glioblastoma, and the genes deleted in between,
1174 could contribute in part to the global OXPHOS complex dysregulation observed in
1175 glioblastoma tissues (23)(20). One possibility is that sequence-specific features (e.g. repeat
1176 regions) and/or structural features (e.g. G4s) can predispose regions of the mtDNA to deletion
1177 events. Taken together, this raises the possibility that LSMDs might be contributing to
1178 mitochondrial functional and oncogenic adaptation in glioblastoma, and this might be repeat
1179 region/ G4 mediated.

1180 Consequently, we first evaluated how the glioblastoma LSMDs detected by "NanoDel" relate
1181 to other factors using matched data on TTD, mtDNA-CN, CIV activity, age, methylated
1182 MGMT promotor status and gender, and using cluster and survival analyses. Since the majority
1183 of LSMDs were present at low read percentages, we analysed their potential impact using
1184 cumulative LSMD metrics, including cumulative deleted read percentage, deletions per 10k
1185 reads, and average burden per deletion, as described by (27). Consistent with separate previous
1186 reports, we show that patients with low mtDNA-CN, compared to high mtDNA-CN (22); and
1187 high complex IV activity, compared to low complex IV activity (16), exhibited worse overall
1188 survival. We show that complex IV activity has a greater impact on overall survival than
1189 mtDNA-CN and cumulative LSMD metrics. Finally, we show that the high complex IV activity
1190 (poorest surviving group) also tends to have lower mtDNA-CN but also higher cumulative
1191 deletion read %, deletions per 10k reads and average burden per deletion levels. Although there
1192 were potential differences in median OS identified between the LSMDs clusters, they were
1193 non-significant in our small sample set, future work with larger sample cohorts may identify a
1194 significant cluster-specific effects, and to what extent each of these mtDNA parameters could
1195 interact to affect complex IV/ median OS.

1196 Consistent with the potential finding that high mtDNA-CN and LSMD levels are associated
1197 with low CIV activity levels of LSMDs, is a previous study showing cells can elicit a
1198 compensatory increase in mtDNA-CN in response to certain variants, including LSMDs and
1199 duplications (86). Furthermore, LSMDs have been linked to MRC CIV deficiency and poor
1200 outcomes in other conditions, such as KSS and presbycusis, where deletions in genes like mt-
1201 co1, mt-co2, and mt-co3 have been shown to impair CIV activity (87)(88). Paradoxically,
1202 however, CIV deficiency is typically associated with more negative outcomes in mitochondrial
1203 disease patients, rather than the better outcomes observed in glioblastoma. In glioblastoma, it
1204 is proposed that higher CIV activity indicates more efficient mitochondrial coupling, leading

1205 to reduced mtROS production. This reduction in mtROS may suppress apoptotic signalling,
1206 thereby making tumors more resistant to treatment, hence poorer outcomes (89).

1207 Also, in glioblastoma, it has been demonstrated that, in part, the regulation of mtDNA-CN is
1208 linked to the methylation of nDNA-encoded subunits for the mtDNA polymerase POLγ. One
1209 study found that methylation of CpG islands in exon 2 of the POLγ subunit A reduces its
1210 expression, leading to a decrease in mtDNA-CN (90). Additionally, mtDNA-CN is known to
1211 play a key role in cell differentiation, with stem cells exhibiting lower mtDNA-CN, via the
1212 methylation of various nuclear encoded factors involved with mtDNA transcription and
1213 replication, compared to differentiated cells (91)(92). The low mtDNA-CN/ high complex IV
1214 activity group (this study) may be indicative of a more pseudo-differentiated state with more
1215 methylation of nuclear encoded factors, stem-like characteristics, and poorer survival,
1216 compared to the high mtDNA-CN group (also this study) that exhibits better survival. Taken
1217 together, these studies raise the possibility that the low mtDNA-CN (high Complex IV, high
1218 LSMDs) group (this study) could also be related to the known differential methylation subtypes
1219 with different stem cell characteristics observed in glioblastoma (93), although this remains to
1220 be investigated.

1221 Our findings also reaffirm the idea that both mtDNA-CN and CIV activity could serve as
1222 potential therapeutic targets. As it has been demonstrated that patients with high CIV activity
1223 have shorter survival, the inhibition of CIV may serve to have therapeutic benefits. A study by
1224 (94) assessed the ability of the compound Chlorpromazine, a CIV inhibitor, to inhibit the
1225 proliferation of glioma stem cells in vitro. Mechanistically, they found that Chlorpromazine
1226 inhibited CIV activity in chemoresistant glioma cells expressing the CIV subunit 4 isoform
1227 cytochrome oxidase 4-2 (COX4-2) but not in chemosensitive glioma cells expressing COX4-
1228 1. Similarly, it is possible to modulate mtDNA-CN. It has been shown that drug compounds
1229 such as Idebenone, a synthetic analogue of CoQ10, can upregulate mtDNA-CN in specific cell
1230 types, such as neural progenitors, and ameliorate mitochondrial phenotypes in E3 ubiquitin
1231 ligase UBE3A Angelman syndrome ube3am-/p+ (maternal loss) mouse models (95)(96). This
1232 may be of therapeutic benefit for patients with high complex IV activity and/or low mtDNA-
1233 CN.

1234 Mechanistic insights and broader implications

1235 Finally, we looked at the location of selected LSMDs breakpoints relative to repeat regions and
1236 putative secondary structural motifs. These features have long been implicated in the formation
1237 of mtDNA deletions through slipped-strand mispairing, and replication stalling mechanisms
1238 (97). Here, identical, or highly similar DNA sequences are mispaired during replication by the
1239 mtDNA polymerase, which slides along the template leading to the removal of the region
1240 between the repeats. Furthermore, replication may stall at non-canonical DNA structures such
1241 as G4s, causing it to become damaged through ds-breaks. One of the main DSBR repair
1242 mechanisms in the mtDNA is microhomology-mediated end joining (MMEJ), which has been
1243 implicated in the formation of LSMDs. Since MMEJ requires micro-homologous regions of 5-
1244 22 bp in length, these imperfect repeat sequences may have determined the specific location
1245 where these LSMDs formed. Thus, a plausible mechanism for LSMD formation in

1246 glioblastoma involves genomic instability caused by G4 structures, replication fork slippage,
1247 transcription arrest, leading to the introduction of double-strand breaks (DSBs) via
1248 Endonuclease G that are then repaired via MMEJ, ultimately resulting in the formation of
1249 LSMDs (98). This might be particularly true in the context of mtDNA, which is very G-rich
1250 and has fewer ROS-protective mechanisms and is therefore prone to destabilization and the
1251 formation of G4s (99), and in the context of glioblastoma (and other cancers), which have
1252 elevated mitochondrial ROS (20).

1253 All 5 of the LSMDs investigated more closely in this study were adjacent to mainly perfect,
1254 but also imperfect, repeat, and potential G4-forming motifs, identified using QGRS mapper
1255 and AF3, although in both cases scores were low. However, AF3 suggests certain putative G4
1256 sequences identified with QGRS Mapper are more likely to form in the presence of K+ ions,
1257 this is as expected, since metal cations coordinate the G-tetrads leading to the stabilisation of
1258 G4s (100). While AF3 represents a significant advancement in macromolecule (particularly
1259 protein) structure prediction, its ability to predict structures is influenced by the availability of
1260 accurate experimental and training structures and/or evolutionary information, such
1261 homologous sequences e.g. from the Protein Databank (PDB, (76)) and other databases. While
1262 there are some 70 G4-structures in the PDB, none are from mtDNA, which may have limited
1263 the ability of AF3 to predict structures with confidence for the LSMDs m.6130-15055del in
1264 the KSS samples, m.8469-13446del in LD, m.8636-16072del in ML, and m.6602-14646-9del
1265 (but not 14825-6del) in glioblastoma.

1266 AF3 did however predict a G4-structure with the sequence m.14806-14821 with high
1267 confidence for the 3'-end break of deletion m.6602-14825-6del in glioblastoma. Curiously,
1268 despite a different sequence, the AF3 structure generated has very similar topology to an NMR-
1269 structure (6YY4, (79)). This G4 has a rare 5'-snapback topology and has high thermal stability.
1270 In such structures, the G-rich sequence folds back on itself forming a snapback loop where the
1271 distant 5'-G is brought into proximity of an incomplete 5'-G-quartet (99). Two other cases of
1272 5'-top snapback G4s have been reported one has an isolated 5'-G completes a hybrid-type G4
1273 derived from the telomeric DNA of *Tetrahymena thermophila* (101), and the second has the
1274 snapback loop forming a short double- helical segment, with the 5'-G inserted into the G-
1275 quartet at the G4–duplex interface (102). NMR-structure (6YY4, (79)) and the G4 generate
1276 from m.14806-14821 (this study) by inference are also examples of 3-plate G4s, that are known
1277 to be associated with deletion breakpoints in humans at a high frequency (98). While the AF3
1278 model of this putative G4 has some geometry issues, particularly in divergence from planarity
1279 in the guanine tetrads, the overall structure of 3 stacked tetrads contributed by 12 guanines
1280 stabilised by a central K+ ion looks to be potentially stable and worthy of further investigation.

1281 While it is possible some G4-structures are more likely to form than others, based on specific
1282 sequences, the balance between ds-DNA and G4-stability and the chemical environment at a
1283 particular point in time, other "spare" G-tracts may too form into G4s, and may serve to
1284 mitigate the effects of the damaged G4s, preserving G4-function under stress conditions (99).
1285 The other predicted G4-sequences in this study yet unresolvable by AF3, might represent such
1286 "spare" regulatory elements. As more G4, especially mtDNA-G4 structures become available
1287 the ability of AF3 to predict structures accurately is set to increase and should shed light on the

authenticity/importance of the other putative G4-sequences identified by QGRS Mapper in the future.

The m.14806-14821 region in which the G4-structure (this study) is predicted to occur is intriguing, as it also lies near m.14798T>C. We have previously shown that this SNV causes F18L substitution in MT-CYB, the central subunit of MRC complex III. We also predict F18L would change the shape of the Qi-site binding pocket in MT_CYB, affecting the on/off rate of ubiquinone, complex III activity, mitochondrial ROS levels, and the efficacy of complex III targeting drugs. We have also shown that glioblastoma cells that carry this SNV have higher complex III activity, mitochondrial ROS, and greater sensitivity to complex III targeting drugs (18)(20). In parallel, this SNV has been shown to be significantly associated with deletions and a repeat region in specific parts of the human brain (103). Given such repetitive regions are hypothesised to be used by MMEJ in DSBR, governing the site of deletions, Hjelm and colleagues suggested the SNV increasing the length of the associated-repeat sequence in mtDNA might explain the observed association between m.14798T>C and deletions. However, we suggest that perhaps G4-structure from m.14806-14821 (this study) could be responsible for the initial genomic instability in this area, which then in turn leads to DSBs, that are then repaired via MMEJ, which is mediated by repeat sequences in the area, causing deletions. Whether m.14798T>C is heightening G4 formation/stability via its effect on complex III/mitochondrial ROS, and/or affecting the efficiency of MMEJ due to increasing repeat sequence length (or other changes caused by m.14798T>C to the local DNA sequence context) remains to be determined.

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, several LSMDs investigated in more detail in glioblastoma and 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" (30) as a unifying feature of mitochondrial dysfunction across multiple pathological states.

**Author Contributions:**

Conceptualization – REM (lead), Data Curation – REM and CFe (equal), Formal Analysis – REM and CFe (equal), JEM (supporting), Funding Acquisition – REM (lead), Investigation – CFe and REM (equal), JEM (supporting), Methodology – CFe and REM (equal), JEM (supporting), Project Administration – REM (lead), Resources – REM and SCR (equal), CG, and CO, JP, CFr, JEM (supporting), Software – CFe (lead), SCR (supporting), Supervision – REM (lead), SCR and RB (supporting), Visualization – CFe (lead), REM and JEM

(supporting), Writing – Original Draft Preparation – REM (lead), CFe (supporting), Writing – Review & Editing – REM (lead), CFe, CO, CG, JP, CFr, JEM, RB, SCR (supporting).

**Acknowledgments:**

Shelest, E., and Lundgren, A.P. aided in the design of the script for clustering LSMD breakpoints in the "NanoDel" pipeline. Beckett, A. and Dent, H. helped with the library preparation and sequencing. C.Fe. was supported by a University of Portsmouth Faculty-PhD studentship and an anonymous donor (awarded to REM).Some running costs were provided by the charity Brain Tumour Research (awarded to Pilkington, G.P.).

**References:**

1. Ostrom QT, Bauchet L, Davis FG, Deltour I, Fisher JL, Langer CE, et al. The epidemiology of glioma in adults: A state of the science review. Vol. 16, Neuro-Oncology. 2014.

2. Dubrow R, Darefsky AS, Jacobs DI, Park LS, Rose MG, Laurans MSH, et al. Time trends in glioblastoma multiforme survival: The role of temozolomide. Neuro Oncol. 2013;15(12).

3. Stupp R, Mason WP, Van Den Bent MJ, Weller M, Fisher B, Taphoorn MJB, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. New England journal of medicine. 2005;352(10):987–96.

4. Labussiere M, Sanson M, Idbaih A, Delattre JY. IDH1 Gene Mutations: A New Paradigm in Glioma Prognosis and Therapy? . Oncologist. 2010;15(2).

5. Butler M, Pongor L, Su YT, Xi L, Raffeld M, Quezado M, et al. MGMT Status as a Clinical Biomarker in Glioblastoma. Vol. 6, Trends in Cancer. 2020.

6. Phillips HS, Kharbanda S, Chen R, Forrest WF, Soriano RH, Wu TD, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell. 2006;9(3):157–73.

7. Sturm D, Witt H, Hovestadt V, Khuong-Quang DA, Jones DTW, Konermann C, et al. Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. Cancer Cell. 2012;22(4).

8. Deighton RF, Le Bihan T, Martin SF, Gerth AMJ, McCulloch M, Edgar JM, et al. Interactions among mitochondrial proteins altered in glioblastoma. J Neurooncol. 2014;118(2).

9. Ordys BB, Launay S, Deighton RF, McCulloch J, Whittle IR. The role of mitochondria in glioma pathophysiology. Mol Neurobiol. 2010;42:64–75.

1363 10. Vafai SB, Mootha VK. Mitochondrial disorders as windows into an ancient organelle.
1364 Vol. 491, Nature. 2012.

1365 11. Niemann S, Muller U. Mutations in SDHC cause autosomal dominant paraganglioma,
1366 type 3. Nat Genet. 2000;26(3).

1367 12. Baysal BE, Ferrell RE, Willett-Brozick JE, Lawrence EC, Myssiorek D, Bosch A, et
1368 al. Mutations in SDHD, a mitochondrial complex II gene, in hereditary paraganglioma. Science
1369 (1979). 2000;287(5454).

1370 13. Brisudova P, Stojanovic D, Novak J, Nahacka Z, Oliveira GL, Vanatko O, et al.
1371 Functional mitochondrial respiration is essential for glioblastoma tumour growth. Oncogene.
1372 2025.

1373 14. Fan Y, Xu W, Gao BQ, Qin H, Wu X, Wei J, et al. Leveraging base excision repair for
1374 efficient adenine base editing of mitochondrial DNA. Nat Biotechnol. 2025.

1375 15. Günther Feichtinger R, Weis S, Adalbert Mayr J, Zimmermann F, Geilberger R, Sperl
1376 W, et al. Alterations of oxidative phosphorylation complexes in astrocytomas. Glia.
1377 2014;62(4).

1378 16. Griguer CE, Cantor AB, Fathallah-Shaykh HM, Gillespie GY, Gordon AS, Markert
1379 JM, et al. Prognostic relevance of cytochrome C oxidase in primary glioblastoma multiforme.
1380 PLoS One. 2013;8(4):e61035.

1381 17. Griguer CE, Oliva CR, Coffey CS, Cudkowicz ME, Conwit RA, Gudjonsdottir AL, et
1382 al. Prospective biomarker study in newly diagnosed glioblastoma: Cyto-C clinical trial.
1383 Neurooncol Adv. 2022;4(1).

1384 18. Lloyd RE, Keatley K, Littlewood DTJ, Meunier B, Holt W V, An Q, et al. Identification
1385 and functional prediction of mitochondrial complex III and IV mutations associated with
1386 glioblastoma. Neuro Oncol. 2015;17(7):942–52.

1387 19. Yeung KY, Dickinson A, Donoghue JF, Polekhina G, White SJ, Grammatopoulos DK,
1388 et al. The identification of mitochondrial DNA variants in glioblastoma multiforme. Acta
1389 Neuropathol Commun. 2014;2(1).

1390 20. Keatley K, Stromei-Cleroux S, Wiltshire T, Rajala N, Burton G, Holt W V, et al.
1391 Integrated approach reveals role of mitochondrial germ-line mutation F18L in respiratory
1392 chain, oxidative alterations, drug sensitivity, and patient prognosis in glioblastoma. Int J Mol
1393 Sci. 2019;20(13):3364.

1394 21. Radzak SMA, Khair SZNM, Idris Z, Ahmad WMAW, Patar A, Yusoff AAM.
1395 Mitochondrial Deoxyribonucleic Acid Copy Number Elevation As a Predictor for Extended
1396 Survival and Favorable Outcomes in High-Grade Brain Tumor Patients: A Malaysian Study.
1397 Eurasian Journal of Medicine. 2024;56(1).

1398 22. Sravya P, Nimbalkar VP, Kanuri NN, Sugur H, Verma BK, Kundu P, et al. Low
1399 mitochondrial DNA copy number is associated with poor prognosis and treatment resistance
1400 in glioblastoma. Mitochondrion. 2020;55:154–63.

1401 23. Oliva CR, Nozell SE, Diers A, McClugage SG, Sarkaria JN, Markert JM, et al.
1402 Acquisition of temozolomide chemoresistance in gliomas leads to remodeling of mitochondrial
1403 electron transport chain. Journal of Biological Chemistry. 2010;285(51):39759–67.

1404 24. Yusoff AAM, Khair SZNM, Abd Radzak SM, Idris Z, Lee HC. Prevalence of
1405 mitochondrial DNA common deletion in patients with gliomas and meningiomas: a first report
1406 from a Malaysian study group. Journal of the Chinese Medical Association. 2020;83(9):838–
1407 44.

1408 25. Tang Y, Manfredi G, Hirano M, Schon EA. Maintenance, of human rearranged
1409 mitochondrial DNAs in long-term cultured transmitochondrial cell lines. Mol Biol Cell.
1410 2000;11(7).

1411 26. Larman TC, DePalma SR, Hadjipanayis AG, Protopopov A, Zhang J, Gabriel SB, et al.
1412 Spectrum of somatic mitochondrial mutations in five cancers. Proc Natl Acad Sci U S A.
1413 2012;109(35).

1414 27. Hjelm BE, Rollins B, Morgan L, Sequeira A, Mamdani F, Pereira F, et al. Splice-Break:
1415 exploiting an RNA-seq splice junction algorithm to discover mitochondrial DNA deletion
1416 breakpoints and analyses of psychiatric disorders. Nucleic Acids Res. 2019;47(10):e59–e59.

1417 28. Moraes CT, DiMauro S, Zeviani M, Lombes A, Shanske S, Miranda AF, et al.
1418 Mitochondrial DNA deletions in progressive external ophthalmoplegia and Kearns-Sayre
1419 syndrome. New England Journal of Medicine. 1989;320(20):1293–9.

1420 29. Santorelli FM, Sciacco M, Tanji K, Shanske S, Vu TH, Golzi V, et al. Multiple
1421 mitochondrial DNA deletions in sporadic inclusion body myositis: a study of 56 patients. Vol.
1422 39, Annals of neurology. Wiley Online Library; 1996. p. 789–95.

1423 30. Schon EA, Rizzuto R, Moraes CT, Nakase H, Zeviani M, DiMauro S. A direct repeat
1424 is a hotspot for large-scale deletion of human mitochondrial DNA. Science (1979).
1425 1989;244(4902):346–9.

1426 31. Poulton J, Deadman ME, Gardiner RM. Tandem direct duplications of mitochondrial
1427 DNA in mitochondrial myopathy: analysis of nucleotide sequence and tissue distribution.
1428 Nucleic Acids Res. 1989;17(24):10223–9.

1429 32. Johns DR, Rutledge SL, Stine OC, Hurko O. Directly repeated sequences associated
1430 with pathogenic mitochondrial DNA deletions. Proceedings of the National Academy of
1431 Sciences. 1989;86(20):8059–62.

1432 33. Mita S, Rizzuto R, Moraes CT, Shanske S, Arnaudo E, Fabrizi GM, et al.
1433 Recombination via flanking direct repeats is a major cause of large-scale deletions of human
1434 mitochondrial DNA. Nucleic Acids Res. 1990;18(3):561–7.

1435 34. Vandiver AR, Hoang AN, Herbst A, Lee CC, Aiken JM, McKenzie D, et al. Nanopore
1436 sequencing identifies a higher frequency and expanded spectrum of mitochondrial DNA
1437 deletion mutations in human aging. Aging Cell. 2023;22(6).

1438 35. Frascarelli C, Zanetti N, Nasca A, Izzo R, Lamperti C, Lamantea E, et al. Nanopore
1439 long-read next-generation sequencing for detection of mitochondrial DNA large-scale
1440 deletions. Front Genet. 2023;14.

1441 36. Wood E, Parker MD, Dunning MJ, Hesketh S, Wang D, Pink R, et al. Clinical long-
1442 read sequencing of the human mitochondrial genome for mitochondrial disease diagnostics.
1443 bioRxiv. 2019;597187.

1444 37. Poulton J, Deadman ME, Bindoff L, Morten K, Land J, Brown G. Families of mtDNA
1445 re-arrangements can be detected in patients with mtDNA deletions: duplications may be a
1446 transient intermediate form. Hum Mol Genet. 1993;2(1):23–30.

1447 38. Poulton J, O'Rahilly S, Morten KJ, Clark A. Mitochondrial DNA, diabetes and
1448 pancreatic pathology in Kearns-Sayre syndrome. Diabetologia. 1995;38:868–71.

1449 39. Lloyd RE, Foster PG, Guille M, Littlewood DTJ. Next generation sequencing and
1450 comparative analyses of Xenopus mitogenomes. BMC Genomics. 2012;13(1):1–16.

1451 40. Bowles EJ, Lee JH, Alberio R, Lloyd REI, Stekel D, Campbell KHS, et al. Contrasting
1452 effects of in vitro fertilization and nuclear transfer on the expression of mtDNA replication
1453 factors. Genetics. 2007;176(3).

1454 41. Ponchel F, Toomes C, Bransfield K, Leong FT, Douglas SH, Field SL, et al. Real-time
1455 PCR based on SYBR-Green I fluorescence: An alternative to the TaqMan assay for a relative
1456 quantification of gene rearrangements, gene amplifications and micro gene deletions. BMC
1457 Biotechnol. 2003;3.

1458 42. De Coster W, D'hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack:
1459 visualizing and processing long-read sequencing data. Bioinformatics. 2018;34(15):2666–9.

1460 43. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N.
1461 Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA.
1462 Nat Genet. 1999;23(2):147.

1463 44. Bolognini D, Sanders A, Korbel JO, Magi A, Benes V, Rausch T. VISOR: a versatile
1464 haplotype-aware structural variant simulator for short-and long-read sequencing.
1465 Bioinformatics. 2020;36(4):1267–9.

1466 45. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.
1467 2018;34(18):3094–100.

1468 46. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence
1469 alignment/map format and SAMtools. bioinformatics. 2009;25(16):2078–9.

1470 47. You Y, Clark MB, Shim H. NanoSplicer: Accurate identification of splice junctions
1471 using Oxford Nanopore sequencing. Bioinformatics. 2022;38(15):3741–8.

48. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–30.

49. Ester M, Kriegel HP, Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings - 2nd International Conference on Knowledge Discovery and Data Mining, KDD 1996. 1996.

50. Bushnell B. BBMap. sourceforge. net/projects/bbmap. Accessed; 2020.

51. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010;38(18):e178–e178.

52. Cretu Stancu M, Van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, De Ligt J, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun. 2017;8(1):1326.

53. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019;20:1–14.

54. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15(6):1–19.

55. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods. 2015;12(10):966–8.

56. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.

57. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36(4):338–45.

58. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7.

59. Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques. 2000;28(6):1102–4.

60. Kikin O, D'Antonio L, Bagga PS. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic Acids Res. 2006;34(suppl_2):W676–W682.

61. Sen D, Gilbert W. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. Nature. 1988;334(6180).

62. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;1–3.

1507 63. Damas J, Carneiro J, Amorim A, Pereira F. MitoBreak: the mitochondrial DNA
1508 breakpoints database. Nucleic Acids Res. 2014;42(D1):D1261–D1268.

1509 64. Shapiro SS, Wilk MB. An Analysis of Variance Test for Normality (Complete
1510 Samples). Biometrika. 1965;52(3/4).

1511 65. Kruskal WH, Wallis WA. Use of Ranks in One-Criterion Variance Analysis. J Am Stat
1512 Assoc. 1952;47(260).

1513 66. Hammer Ø, Harper DAT, Ryan PD. Past: Paleontological statistics software package
1514 for education and data analysis. Palaeontologia Electronica. 2001;4(1).

1515 67. Reyes A, Melchionda L, Nasca A, Carrara F, Lamantea E, Zanolini A, et al. RNASEH1
1516 Mutations Impair mtDNA Replication and Cause Adult-Onset Mitochondrial
1517 Encephalomyopathy. Am J Hum Genet. 2015;97(1).

1518 68. Marti R, Verschuuren JJGM, Buchman A, Hirano I, Tadesse S, Van Kuilenburg ABP,
1519 et al. Late-onset MNGIE due to partial loss of thymidine phosphorylase activity. Ann Neurol.
1520 2005;58(4).

1521 69. Wei YH. Mitochondrial DNA alterations as ageing-associated molecular events.
1522 Mutation Research/DNAging. 1992;275(3–6):145–55.

1523 70. Volmering E, Niehusmann P, Peeva V, Grote A, Zsurka G, Altmüller J, et al.
1524 Neuropathological signs of inflammation correlate with mitochondrial DNA deletions in
1525 mesial temporal lobe epilepsy. Acta Neuropathol. 2016;132:277–88.

1526 71. Hwang BJ, Kuttamperoor F, Wu J, Steinberg ML. Spectrum of mitochondrial DNA
1527 deletions within the common deletion region induced by low levels of UVB irradiation of
1528 human keratinocytes in vitro. Gene. 2009;440(1–2):23–7.

1529 72. Regan C. SEQUENCE CHARACTERIZATION OF BREAKPOINT REGIONS OF
1530 HUMAN MlTOCHONDRlAL DNA DELETIONS AND POSSIBLE MECHANISMS OF
1531 FORMATION. 1999.

1532 73. Wanrooij S, Luoma P, van Goethem G, van Broeckhoven C, Suomalainen A, Spelbrink
1533 JN. Twinkle and POLG defects enhance age-dependent accumulation of mutations in the
1534 control region of mtDNA. Nucleic Acids Res. 2004;32(10).

1535 74. Spelbrink JN, Li FY, Tiranti V, Nikali K, Yuan QP, Tariq M, et al. Human
1536 mitochondrial DNA deletions associated with mutations in the gene encoding Twinkle, a phage
1537 T7 gene 4-like protein localized in mitochondria. Nat Genet. 2001;28(3).

1538 75. Dong DW, Pereira F, Barrett SP, Kolesar JE, Cao K, Damas J, et al. Association of G-
1539 quadruplex forming sequences with human mtDNA deletion breakpoints. BMC Genomics.
1540 2014;15:1–15.

1541 76. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein
1542 Data Bank. Vol. 28, Nucleic Acids Research. 2000.

1543 77. Aoki K, Murayama K. Nucleic acid-metal ion interactions in the solid state. Met Ions
1544 Life Sci. 2012;10.

1545 78. Sengar A, Heddi B, Phan AT. Formation of G-quadruplexes in poly-G sequences:
1546 Structure of a propeller-type parallel-stranded G-quadruplex formed by a G15 stretch.
1547 Biochemistry. 2014;53(49).

1548 79. Volek M, Kolesnikova S, Svehlova K, Srb P, Sgallová R, Streckerová T, et al.
1549 Overlapping but distinct: A new model for G-quadruplex biochemical specificity. Nucleic
1550 Acids Res. 2021;49(4).

1551 80. Zhang W, Cui H, Wong LJC. Comprehensive one-step molecular analyses of
1552 mitochondrial genome by massively parallel sequencing. Clin Chem. 2012;58(9):1322–31.

1553 81. Aikhionbare FO, Khan M, Carey D, Okoli J, Go R. Is cumulative frequency of
1554 mitochondrial DNA variants a biomarker for colorectal tumor progression? Mol Cancer.
1555 2004;3:1–4.

1556 82. Lam ET, Bracci PM, Holly EA, Chu C, Poon A, Wan E, et al. Mitochondrial DNA
1557 sequence variation and risk of pancreatic cancer. Cancer Res. 2012;72(3):686–95.

1558 83. Elango S, Govindaraj P, Vishwanadha VP, Reddy AG, Tamang R, Muthusami U, et al.
1559 Analysis of mitochondrial genome revealed a rare 50 bp deletion and substitutions in a family
1560 with hypertension. Mitochondrion. 2011;11(6):878–85.

1561 84. Kopsidas G, Kovalenko SA, Islam MM, Gingold EB, Linnane AW. Preferential
1562 amplification is minimised in long-PCR systems. Mutation Research - Fundamental and
1563 Molecular Mechanisms of Mutagenesis. 2000;456(1–2).

1564 85. Martin S, Heavens D, Lan Y, Horsfield S, Clark MD, Leggett RM. Nanopore adaptive
1565 sampling: a tool for enrichment of low abundance species in metagenomic samples. Genome
1566 Biol. 2022;23(1).

1567 86. Wong LJC, Perng CL, Hsu CH, Bai RK, Schelley S, Vladutiu GD, et al. Compensatory
1568 amplification of mtDNA in a patient with a novel deletion/duplication and high mutant load. J
1569 Med Genet. 2003;40(11):e125–e125.

1570 87. Mita S, Schmidt B, Schon EA, DiMauro S, Bonilla E. Detection of" deleted"
1571 mitochondrial genomes in cytochrome-c oxidase-deficient muscle fibers of a patient with
1572 Kearns-Sayre syndrome. Proceedings of the National Academy of Sciences.
1573 1989;86(23):9509–13.

1574 88. Markaryan A, Nelson EG, Hinojosa R. Major arc mitochondrial DNA deletions in
1575 cytochrome c oxidase-deficient human cochlear spiral ganglion cells. Acta Otolaryngol.
1576 2010;130(7):780–7.

1577 89. Oliva CR, Moellering DR, Gillespie GY, Griguer CE. Acquisition of chemoresistance
1578 in gliomas is associated with increased mitochondrial coupling and decreased ROS production.
1579 PLoS One. 2011;6(9):e24665.

1580 90. Kelly RDW, Mahmud A, McKenzie M, Trounce IA, St John JC. Mitochondrial DNA
1581 copy number is regulated in a tissue specific manner by DNA methylation of the nuclear-
1582 encoded DNA polymerase gamma A. Nucleic Acids Res. 2012;40(20):10124–38.

1583 91. John JCS. Mitochondrial DNA copy number and replication in reprogramming and
1584 differentiation. In: Seminars in cell & developmental biology. 2016. p. 93–101.

1585 92. Sun X, St. John JC. The role of the mtDNA set point in differentiation, development
1586 and tumorigenesis. Biochemical Journal. 2016;473(19):2955–71.

1587 93. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al.
1588 Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma
1589 characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell.
1590 2010;17(1):98–110.

1591 94. Oliva CR, Zhang W, Langford C, Suto MJ, Griguer CE. Repositioning chlorpromazine
1592 for treating chemoresistant glioma through the inhibition of cytochrome c oxidase bearing the
1593 COX4-1 regulatory subunit. Oncotarget. 2017;8(23).

1594 95. Llewellyn KJ, Nalbandian A, Gomez A, Wei D, Walker N, Kimonis VE.
1595 Administration of CoQ10 analogue ameliorates dysfunction of the mitochondrial respiratory
1596 chain in a mouse model of Angelman syndrome. Neurobiol Dis. 2015;76:77–86.

1597 96. Augustyniak J, Lenart J, Zychowicz M, Stepien PP, Buzanska L. Mitochondrial
1598 biogenesis and neural differentiation of human iPSC is modulated by idebenone in a
1599 developmental stage-dependent manner. Biogerontology. 2017;18:665–77.

1600 97. Phillips AF, Millet AR, Tigano M, Dubois SM, Crimmins H, Babin L, et al. Single-
1601 Molecule Analysis of mtDNA Replication Uncovers the Basis of the Common Deletion. Mol
1602 Cell. 2017;65(3).

1603 98. Dahal S, Siddiqua H, Katapadi VK, Iyer D, Raghavan SC. Characterization of G4 DNA
1604 formation in mitochondrial DNA and their potential role in mitochondrial genome instability.
1605 FEBS J. 2022;289(1):163–82.

1606 99. Aleksič S, Podbevšek P, Plavec J. Oxidative events in a double helix system promote
1607 the formation of kinetically trapped G-quadruplexes. Nucleic Acids Res. 2025 Apr 11;53(6).

1608 100. Phan AT. Human telomeric G-quadruplex: structures of DNA and RNA sequences.
1609 FEBS J. 2010;277(5):1107–17.

1610 101. Beseiso D, Chen E V., McCarthy SE, Martin KN, Gallagher EP, Miao J, et al. The first
1611 crystal structures of hybrid and parallel four-Tetrad intramolecular G-quadruplexes. Nucleic
1612 Acids Res. 2022;50(5).

1613 102. Lim KW, Phan AT. Structural basis of DNA quadruplex-duplex junction formation.
1614 Angewandte Chemie - International Edition. 2013;52(33).

103. Hjelm BE, Ramiro C, Rollins BL, Omidsalar AA, Gerke DS, Das SC, et al. Large Common Mitochondrial DNA Deletions Are Associated with a Mitochondrial SNP T14798C Near the 3' Breakpoints. Complex Psychiatry. 2022;8(3–4):90–8.

**Supplementary Information :**

*Fig S1. Flow diagram of large-scale mtDNA deletions using Pipeline 1 ("NanoDel") developed for this study. A) Raw read processing. Raw reads were base called 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), all mapped reads were then subsequently removed from their constituent FASTQ files. Individual FastQ files that were then filtered to remove all reads <20bp and 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 alignment settings enabled (-ax splice), but preferences for canonical splice junctions disabled (-u no). These outputs are piped directly to Samtools where the resulting data were converted to BAM format using Samtools view (-S and -b options for SAM format input and BAM output) and sorted using the Samtools sort function. Splice junction locations were then extracted from the BAM files using the JWR_Checker.py script from NanoSplicer (v1.0), with output CSV format options enabled (–output_csv). Simultaneously, the per base depth was calculated by the Samtools depth function for all positions regardless of coverage (option -aa) and later used for heteroplasmy calculation. Custom R scripts were then available as options within the pipeline (filt or no_filt) that either reformatted the data and remove any deletion junction calls with 5'- or 3'-breakpoints within 500bp of a primer binding site, or not. This is because (27) demonstrated an increase in false positive breakpoint calls within this distance of primer boundaries. The remaining breakpoint junctions were then clustered using a custom python script that employed the sci-kit learn package (v1.2). A density-based clustering approach (DBScan) was used to group all LSMD calls with both 5'- and 3'-breakpoints that were within 10bp of other 5'- and 3'-breakpoints. Average 5'- and 3'-locations were then taken from each cluster and reported with the number of reads generating that cluster for each sample and output in CSV format, with all points not contributing to a cluster being subset as 'noise points'. C) Finally, a mean coverage benchmark was calculated from the depth output between positions m.500-1000 (according to the rCRS reference sequence) and used in the calculation of the relative read % (a proxy for heteroplasmy) for each LSMD using a custom R-script. As*

*such the number of reads containing each specific LSMD was expressed as a percentage of the coverage benchmark for each respective sample.*

*Fig S2. LR-PCR amplification and agarose gel analysis of mitochondrial disease patient sample mtDNAs as two-amplicons overlapping primers A) CytbF and LongR/ Amplicon 1 (expected size ~8573bp), and B) LongF and Cytb/ Amplicon 2 (expected size ~9128bp); and as one-amplicon using almost back to back primers C) Alt Human LR-PCR F and Alt Human LR-PCR R (expected size 16299bp). A 1kb DNA ladder and 17kb and 15kb DNA markers were used for sizing mtDNA bands. Muscle, kidney, and cerebellum samples were from a KSS patient, and ML cultured cells, MM muscle, PS muscle, LD cultured cells, and PM muscle were from patients with either confirmed or unconfirmed genetic diagnosis (Table S1B). SC1800-1 and 2 were from two human adult non-neoplastic astrocyte cell lines derived from the cerebral cortex, represent full length mtDNA + controls. 143B osteosarcoma and Rho-0 143B (a derivative of the 143B cells that have lost their mtDNA) cell lines, represent further full length mtDNA + control and a mtDNA – control, respectively. NFW (nuclease free water), represents no DNA input (PCR) – control. Smaller than full length mtDNAs are suggestive of LSMDs being present, and depending on the LR-PCR primers used, this reveals either "single" or "multiple" LSMD types.Fig S3. LR-PCR amplification and gel analysis of glioblastoma sample mtDNAs as two-amplicons using two overlapping primers A) CytbF and LongR/ Amplicon 1, and B) LongF and Cytb/ Amplicon 2 (Agilent Bioanalyser analysis); and as one-amplicon using almost back to back primers C) Alt Human LR-PCR F and Alt Human LR-PCR R (agarose gel electrophoresis analysis). Full length mtDNAs were expected to be ~8573 and ~9128, respectively, based on the rCRS. Labels 1-21 glioblastoma tissue samples, 22-24 normal brain tissue controls.*

*Fig S4. LR-PCR amplification and gel analysis of glioblastoma sample mtDNAs as one-amplicon using almost back-to-back primers Alt Human LR-PCR F and Alt Human LR-PCR R (agarose gel electrophoresis analysis). Full length mtDNAs were expected to be ~16299bp in size, based on the rCRS. A) UAB1-21 glioblastoma tissue samples, UAB22-24 normal brain tissue controls. NFW (nuclease free water), represents no DNA input (PCR) – control. Smaller than full length mtDNAs are suggestive of LSMDs being present. A 1kb DNA ladder and 17 and 15kb DNA markers were used for sizing mtDNA bands.*

*Fig S5. Analysis of clusters generated using cumulative deleted read %, deletions per 10k of reads and average burden per deletion in glioblastoma. A-C) cumulative deleted read %, D-F) deletions per 10k of reads and G-I) average burden per deletion in glioblastoma. A) Clustering generated 3 distinct groups of samples (with one outlier), respectively. B) Kaplan-Meier curve showing a borderline significant difference (P=0.06) between median OS of cluster 1 vs 2 (22.7 vs 13.3 months, P=0.18), 2 vs 3 (13.3 vs 6.5 months, P=0.09) and 1 vs 3 (22.7 vs 6.5 months, P=0.09). C) Significant differences in cumulative deleted read% (P=0.001), deletions per 10k reads (P=0.001), average burden per deletion (P=0.001), and mtDNA-CN (P=0.02) but not CIV activity (P=0.25) were identified. D-F) Deletions per 10k of reads, D) Clustering generated 3 distinct groups of samples, respectively. E) Kaplan-Meier curve showing a borderline significant difference (P=0.08) between median OS of cluster 1 vs 2 (12.4 vs 4.84 months, P=0.10), 2 vs 3 (4.84 vs 13.3 months, P=0.05) and 1 vs 3 (12.4 vs 13.3*

*months, P=0.58). F) Significant differences in cumulative deleted read percentage (P=0.001), deletions per 10k reads (P=0.001), and mtDNA-CN (P=0.04) but not the average burden per deletion (P=0.12) or CIV activity (P=0.49) were identified. G-I) Average burden per deletion. G) Clustering generated 3 distinct groups of samples. H) Kaplan-Meier curve showing no significant difference in median overall survival time between the clusters (P=0.96). I) Significant differences in in the average burden per deletion (P=0.001), but not cumulative deleted read% (P=0.24), deletions per 10k reads (P=0.24), mtDNA-CN (P=0.09) and CIV activity (P=0.84). * = P<0.05, ** = P<0.01, *** = P<0.001, and **** = P<0.0001.*

*Fig S6. Kaplan-Meier estimates of overall survival in primary adult glioblastoma patients stratified based age, MGMT promotor status and gender. A) age (groups determined by cluster analysis, n=17; sample 15 was an outlier (18 yo) and so excluded, >62yo group (n=5) had an average age of 74.8 +/- 2.8, and <62yo group (n=12) had an average age of 52.8 +/- 6.1 years), B) MGMT promoter status, n=15; methylated (Me) group (n=8), unmethylated (U) group (n=7), samples 4, 7 and 10 were excluded as their status was not determined, C) gender, n=18, males (M) group (n=8), females (F) group (n=10). The median overall survival for all groups (in months), and p-values are also indicated.*

*Table S1. S1 MtDNA variant types and clinical features*

*Table S2A. Cell culture details*

*Table S2B. Clinically diagnosed mitochondrial disease patient details*

*Table S2C. Glioblastoma patient and normal brain sample details*

*Table S3. Primer information*

*Table S4A. Pipeline 1 (Nanodel) read statistics for mitochondrial disease patients*

*Table S4B. Pipeline 3 (Wood Illumina) read statistics for mitochondrial disease patients*

*Table S5A. Pipeline 1 (Nanodel) LSMDs for mitochondrial disease patients*

*Table S5B. Pipeline 3 (Wood Illumina) LSMDs for mitochondrial disease patients*

*Table S6. Pipeline 1 (Nanodel) read statistics for glioblastoma patients*

*Table S7. Pipeline 1 (Nanodel) LSMDs for glioblastoma patients*

*Table S8. Repeat regions and G4 sequences*

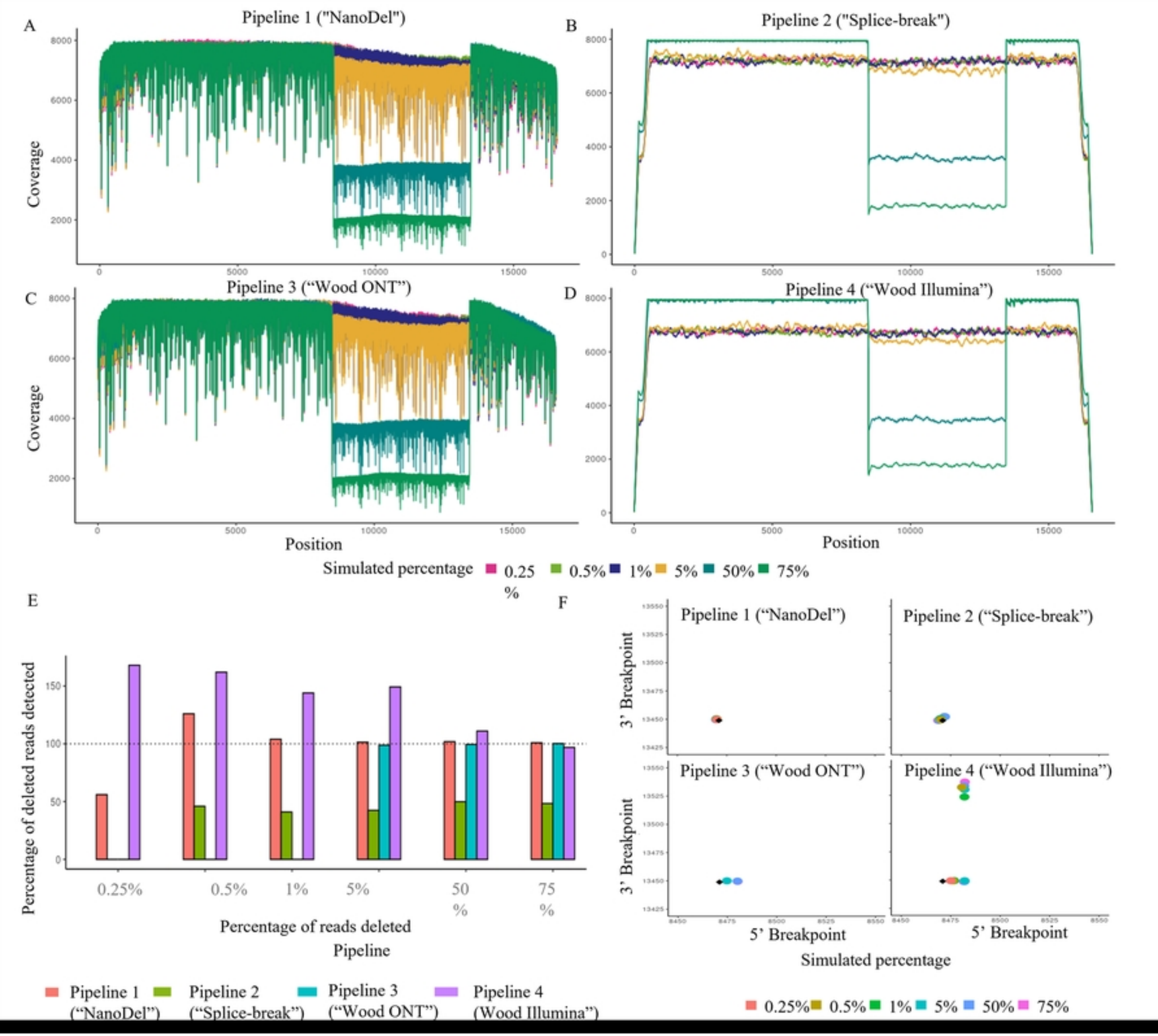
A
Pipeline 1 ("NanoDel")
B
Pipeline 2 ("Splice-break")
C
Pipeline 3 ("Wood ONT")
D
Pipeline 4 ("Wood Illumina")
Coverage
Position
Simulated percentage
0.25 %
0.5%
1%
5%
50%
75%
E
Percentage of deleted reads detected
0.25%
0.5%
1%
5%
50 %
75 %
Percentage of reads deleted
Pipeline
Pipeline 1 ("NanoDel")
Pipeline 2 ("Splice-break")
Pipeline 3 ("Wood ONT")
Pipeline 4 (Wood Illumina")
F
3' Breakpoint
5' Breakpoint
Simulated percentage
0.25%
0.5%
1%
5%
50%
75%

Figure 1

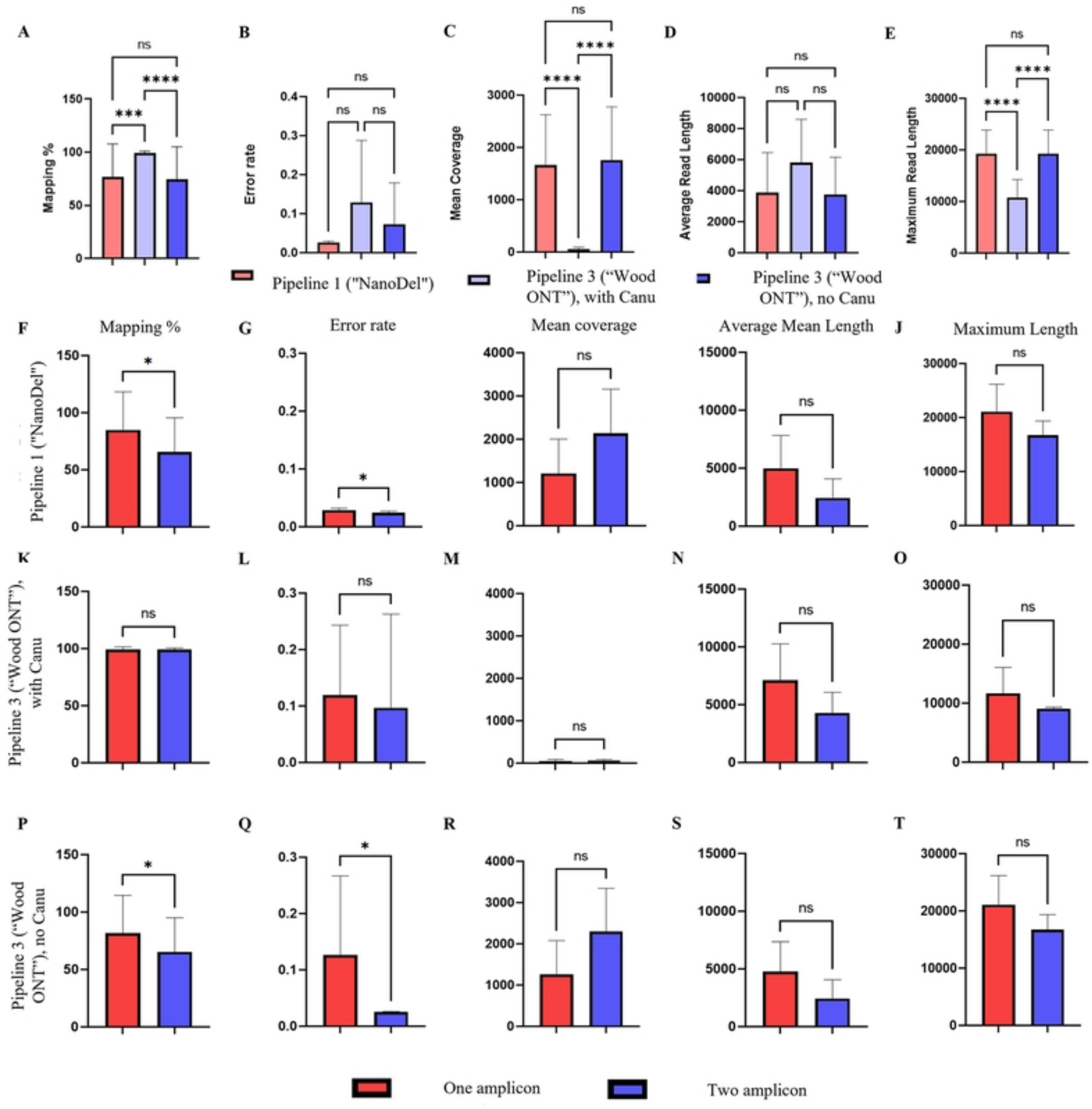
A
B
C
D
E
Mapping %
Error rate
Mean Coverage
Average Read Length
Maximum Read Length
ns
***
****
Pipeline 1 ("NanoDel")
Pipeline 3 ("Wood ONT"), with Canu
Pipeline 3 ("Wood ONT"), no Canu
F
Mapping %
G
Error rate
Mean coverage
Average Mean Length
J
Maximum Length
Pipeline 1 ("NanoDel")
K
L
M
N
O
Pipeline 3 ("Wood ONT"), with Canu
P
Q
R
S
T
Pipeline 3 ("Wood ONT"), no Canu
One amplicon
Two amplicon

Figure 3

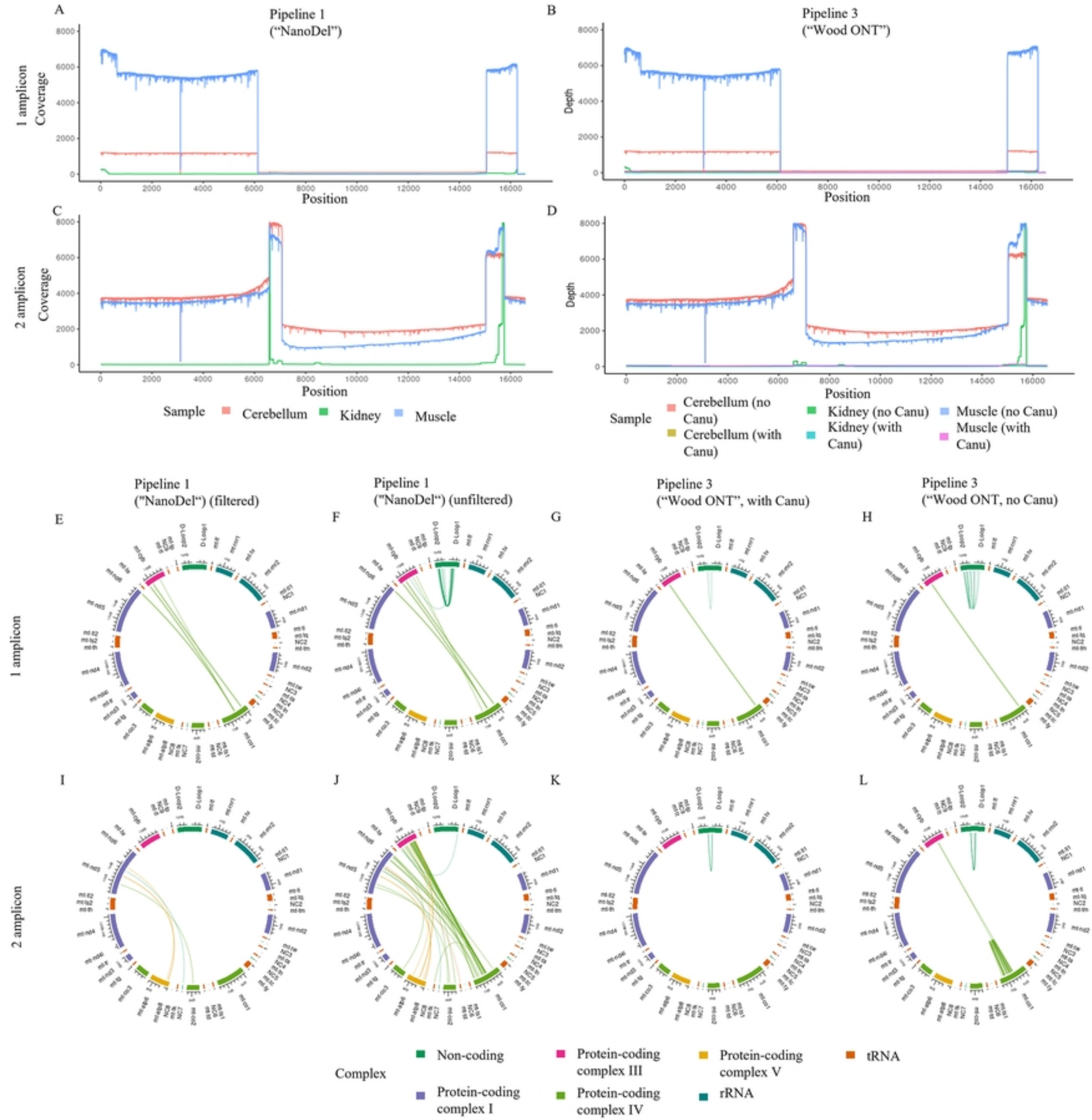
A
Pipeline 1
("NanoDel")
B
Pipeline 3
("Wood ONT")
1 amplicon
Coverage
Depth
Position
C
D
2 amplicon
Sample
Cerebellum
Kidney
Muscle
Cerebellum (no Canu)
Cerebellum (with Canu)
Kidney (no Canu)
Kidney (with Canu)
Muscle (no Canu)
Muscle (with Canu)
Pipeline 1
("NanoDel") (filtered)
Pipeline 1
("NanoDel") (unfiltered)
Pipeline 3
("Wood ONT", with Canu)
Pipeline 3
("Wood ONT, no Canu)
E
F
G
H
I
J
K
L
Complex
Non-coding
Protein-coding complex I
Protein-coding complex III
Protein-coding complex IV
Protein-coding complex V
rRNA
tRNA

Figure 4

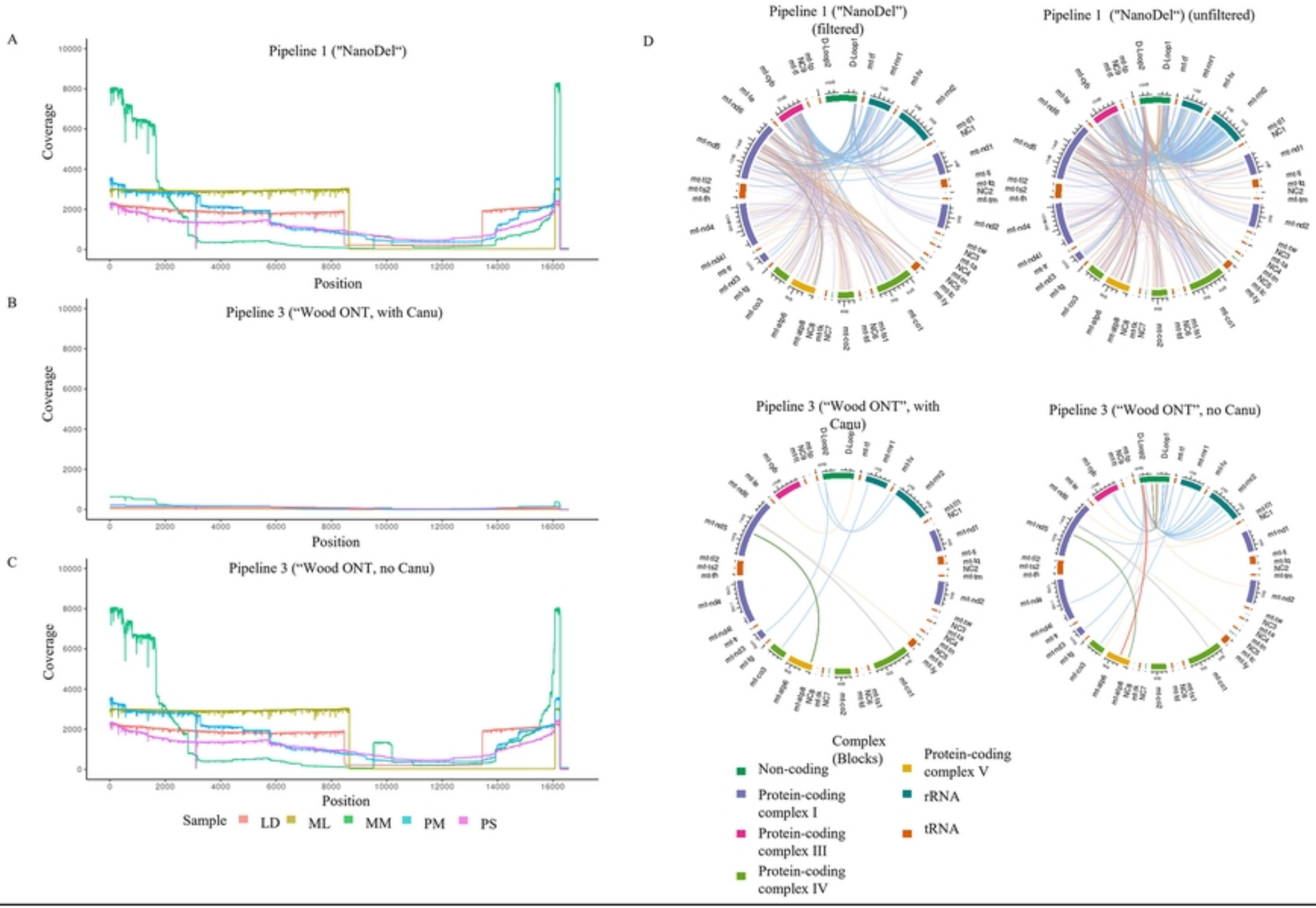
A
Pipeline 1 ("NanoDel")
Coverage
Position
B
Pipeline 3 ("Wood ONT, with Canu)
Coverage
Position
C
Pipeline 3 ("Wood ONT, no Canu)
Coverage
Position
Sample LD ML MM PM PS
D
Pipeline 1 ("NanoDel") (filtered)
Pipeline 1 ("NanoDel") (unfiltered)
Pipeline 3 ("Wood ONT", with Canu)
Pipeline 3 ("Wood ONT", no Canu)
Complex (Blocks)
Non-coding
Protein-coding complex I
Protein-coding complex III
Protein-coding complex IV
Protein-coding complex V
rRNA
tRNA

Figure 5

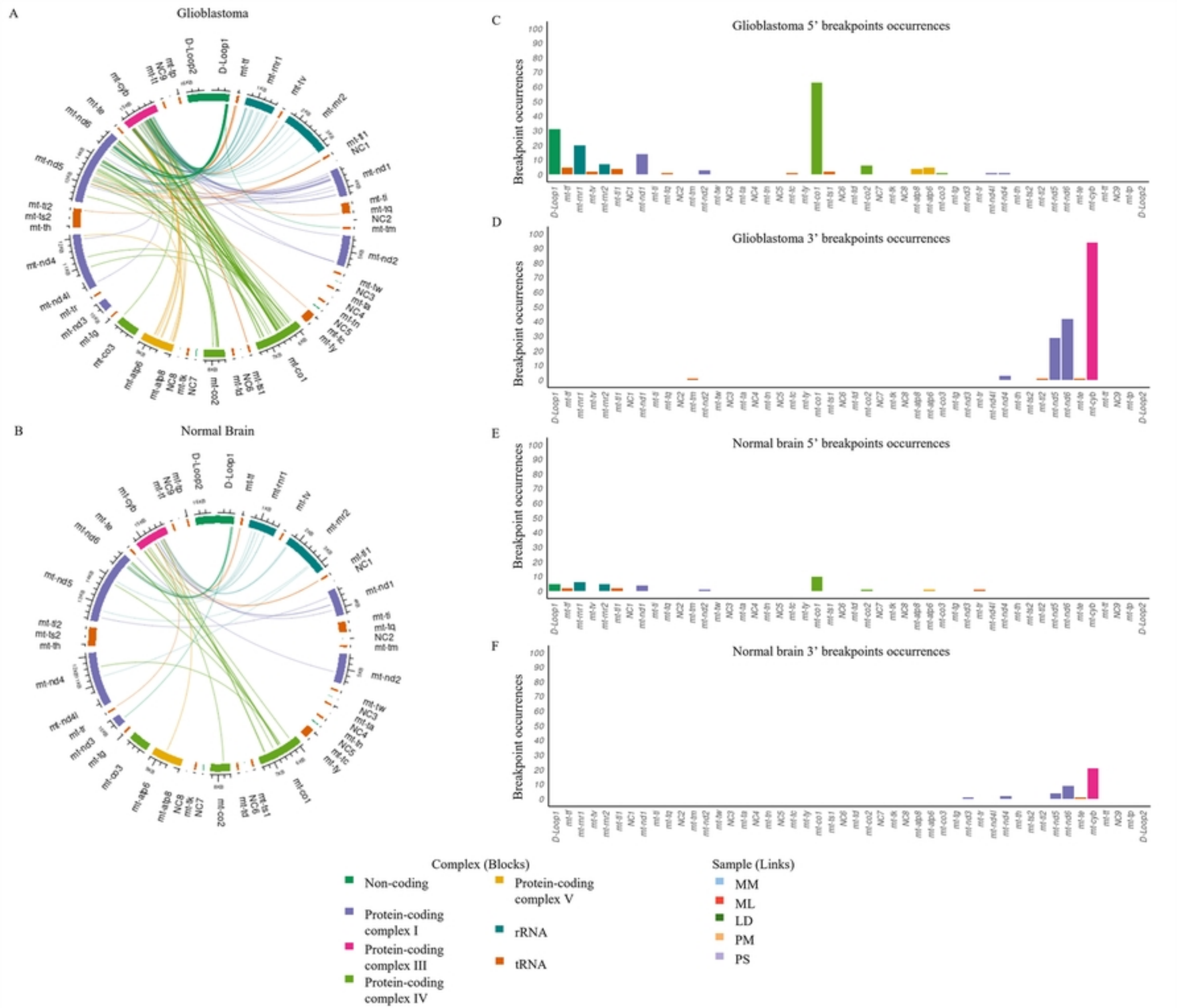
A
Glioblastoma
B
Normal Brain
C
Glioblastoma 5' breakpoints occurrences
D
Glioblastoma 3' breakpoints occurrences
E
Normal brain 5' breakpoints occurrences
F
Normal brain 3' breakpoints occurrences
Breakpoint occurrences
Complex (Blocks)
Non-coding
Protein-coding complex I
Protein-coding complex III
Protein-coding complex IV
Protein-coding complex V
rRNA
tRNA
Sample (Links)
MM
ML
LD
PM
PS

Figure 6

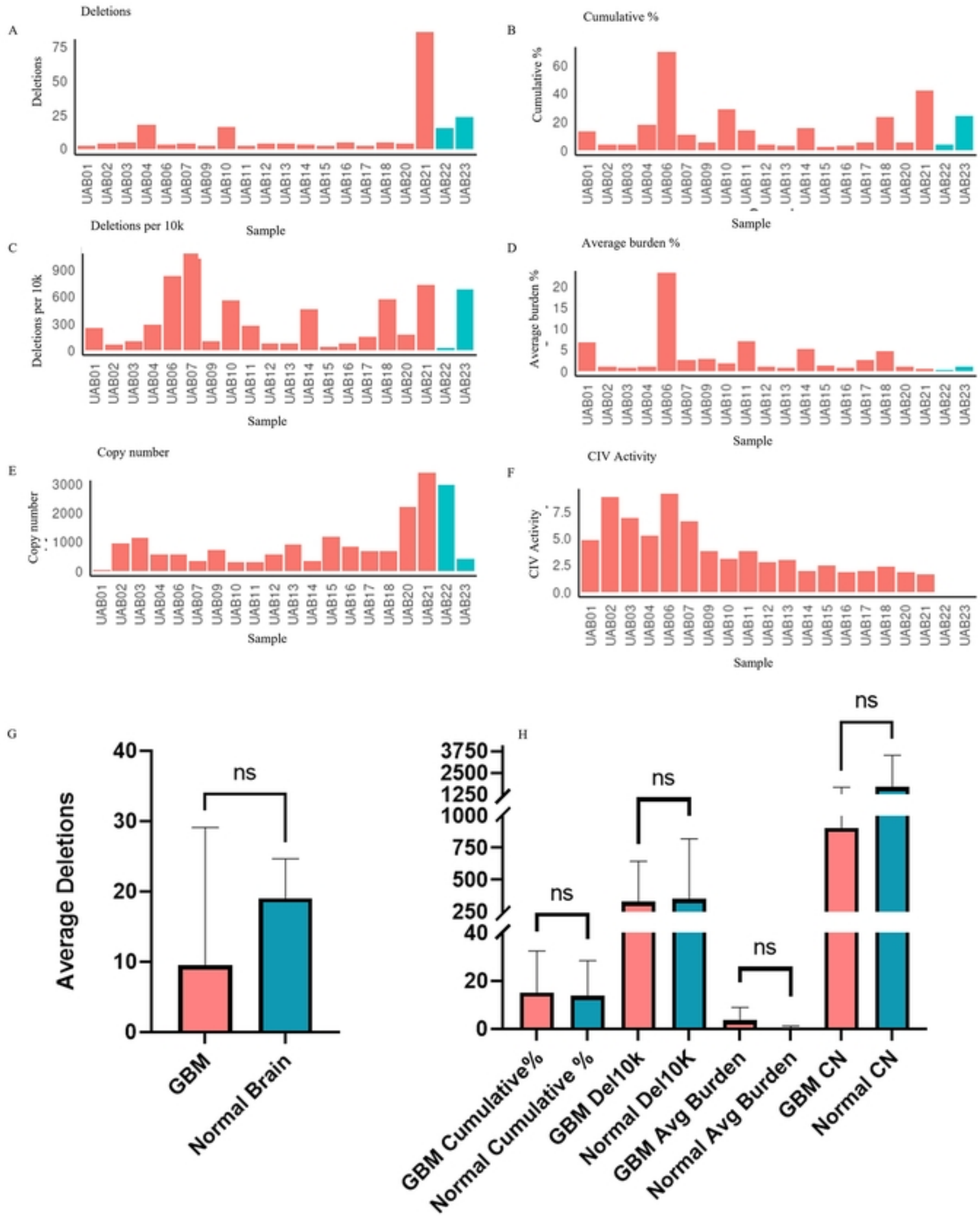
A
Deletions
Deletions
Sample
B
Cumulative %
Cumulative %
Sample
C
Deletions per 10k
Deletions per 10k
Sample
D
Average burden %
Average burden %
Sample
E
Copy number
Copy number
Sample
F
CIV Activity
CIV Activity
Sample
G
Average Deletions
ns
GBM
Normal Brain
H
ns
GBM Cumulative%
Normal Cumulative %
GBM Del10k
Normal Del10K
GBM Avg Burden
Normal Avg Burden
GBM CN
Normal CN

Figure 7

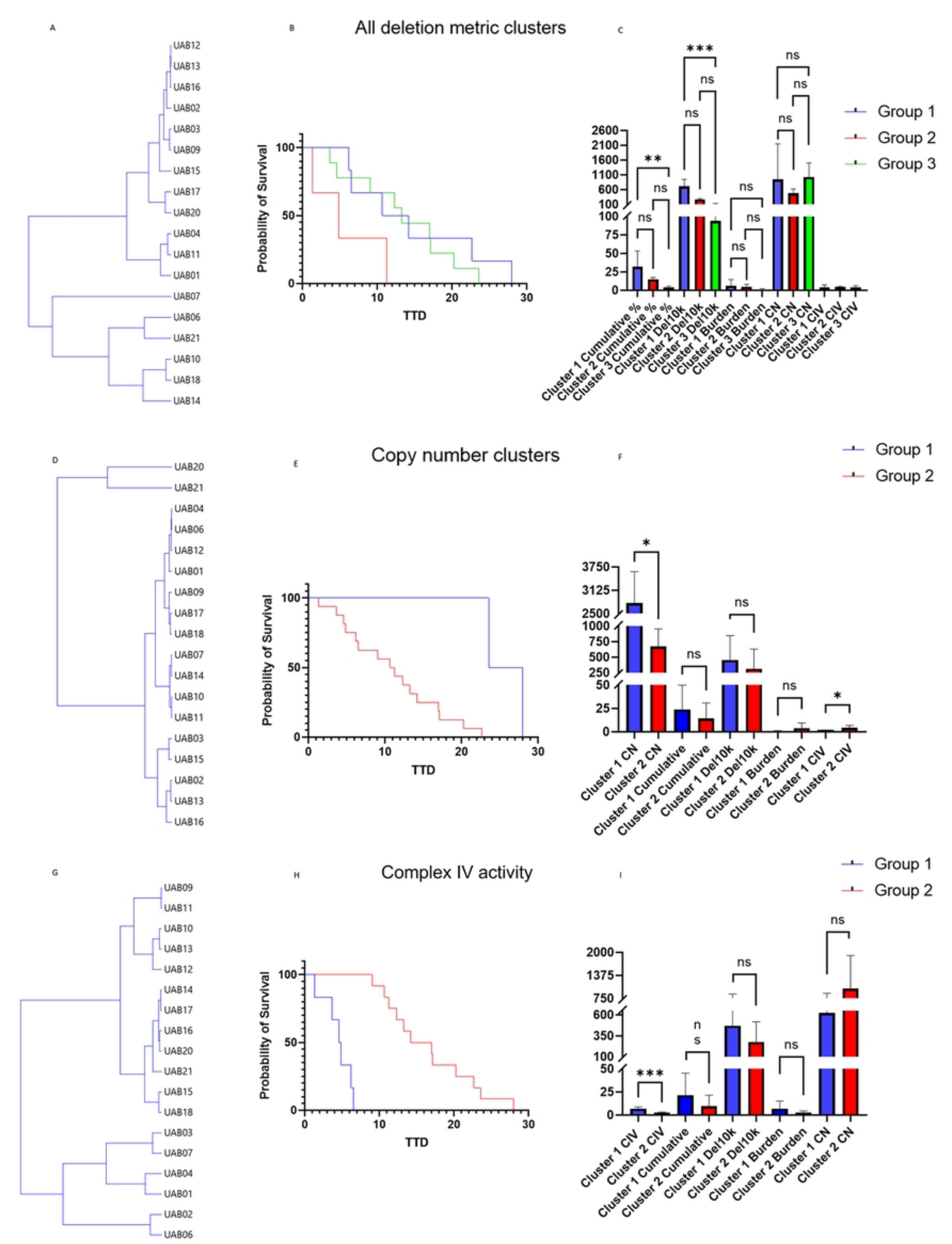
A
B
All deletion metric clusters
C
UAB12
UAB13
UAB16
UAB02
UAB03
UAB09
UAB15
UAB17
UAB20
UAB04
UAB11
UAB01
UAB07
UAB06
UAB21
UAB10
UAB18
UAB14
Probability of Survival
TTD
Group 1
Group 2
Group 3
Cluster 1 Cumulative %
Cluster 2 Cumulative %
Cluster 3 Cumulative %
Cluster 1 Del10k
Cluster 2 Del10k
Cluster 3 Del10k
Cluster 1 Burden
Cluster 2 Burden
Cluster 3 Burden
Cluster 1 CN
Cluster 2 CN
Cluster 3 CN
Cluster 1 CIV
Cluster 2 CIV
Cluster 3 CIV
D
E
Copy number clusters
F
Group 1
Group 2
G
H
Complex IV activity
I

Figure 8

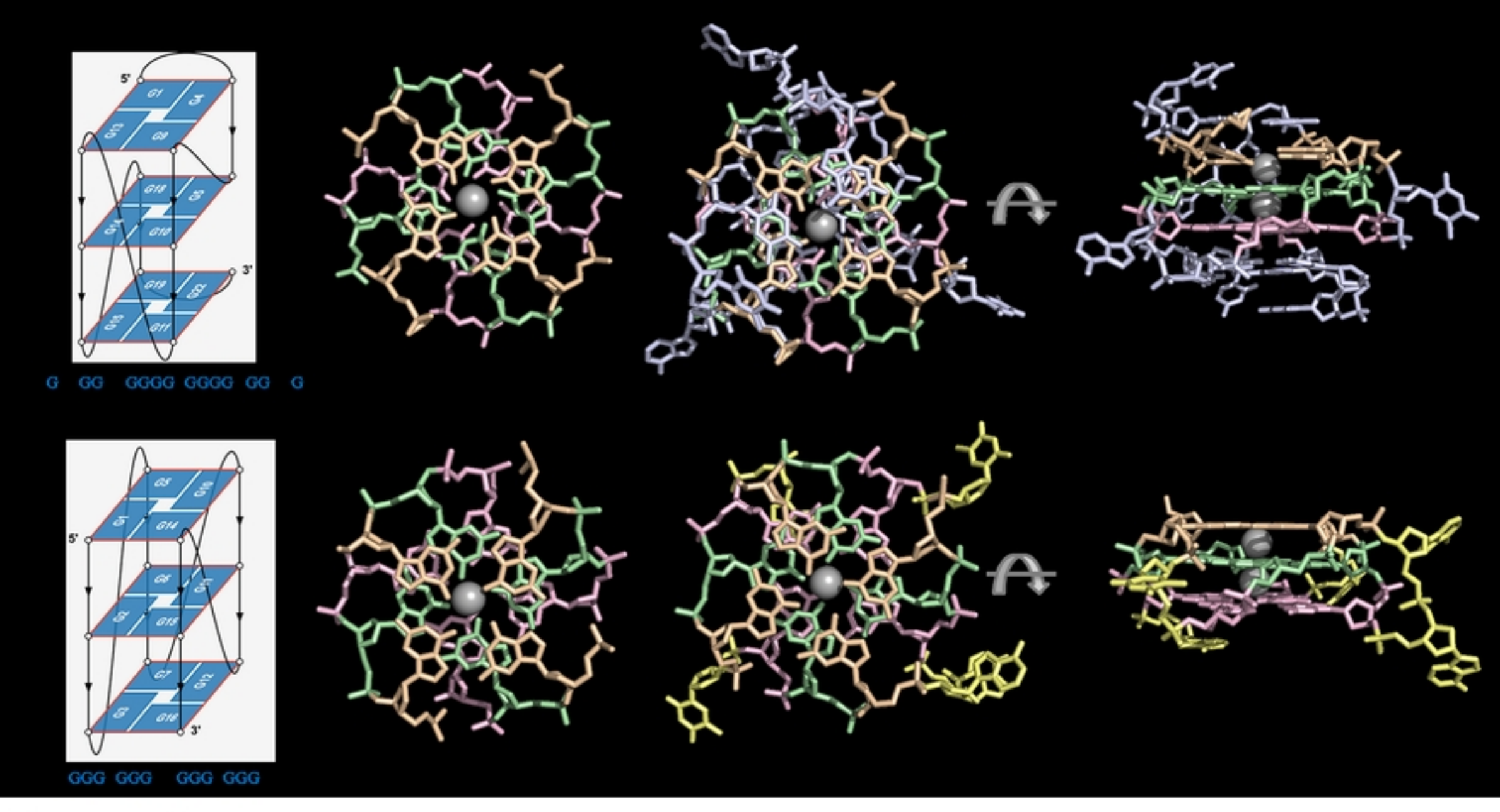


Figure 9

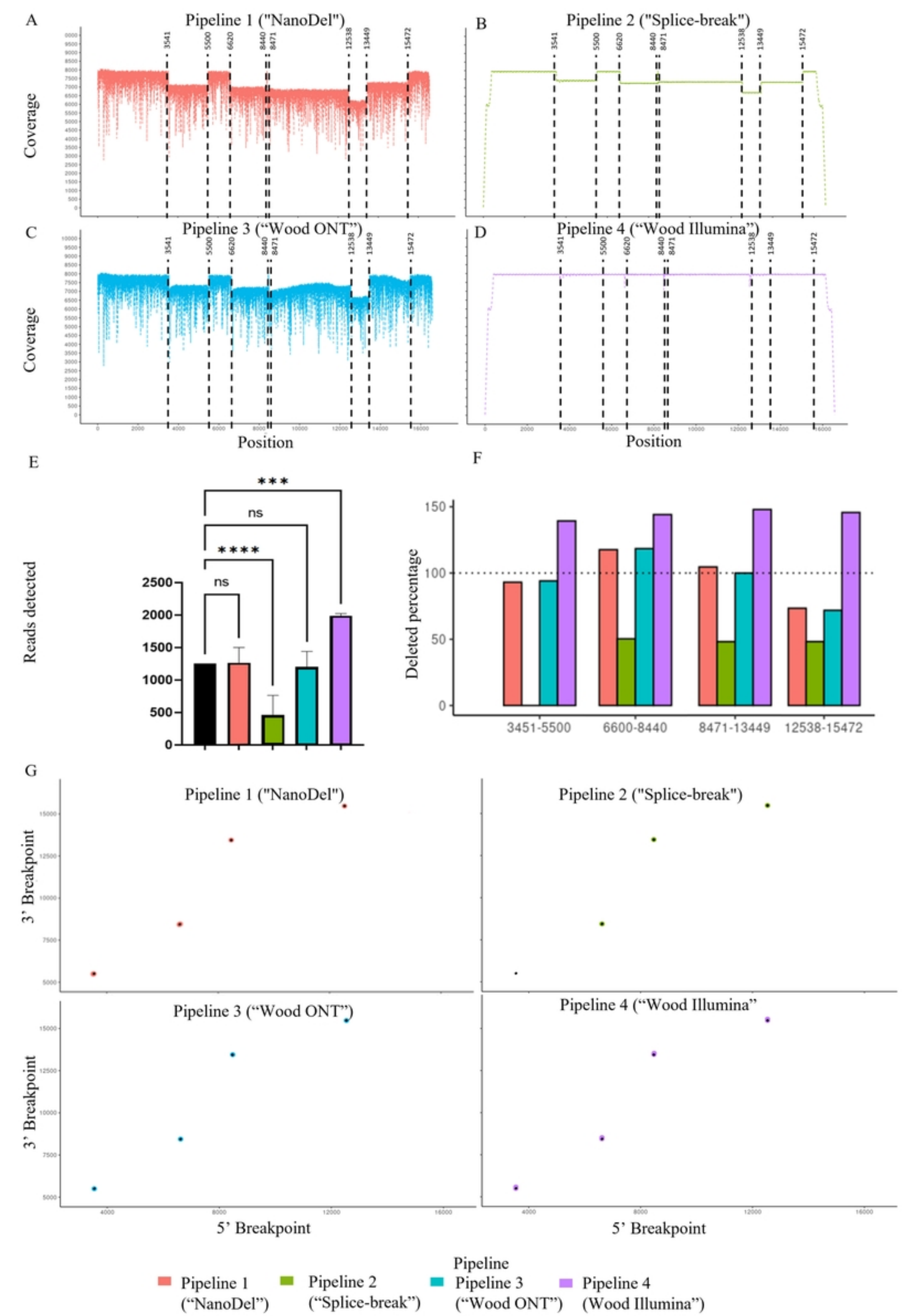
A
Pipeline 1 ("NanoDel")
B
Pipeline 2 ("Splice-break")
C
Pipeline 3 ("Wood ONT")
D
Pipeline 4 ("Wood Illumina")
Coverage
Position
3541
5500
6620
8440
8471
12538
13449
15472
E
Reads detected
***
ns
****
ns
F
Deleted percentage
3451-5500
6600-8440
8471-13449
12538-15472
G
Pipeline 1 ("NanoDel")
Pipeline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

Figure 2