Mapping of mitochondrial DNA deletions and duplications using deep sequencing

Abstract : *Duplication and deletion in mitochondria genome leads to variety of rare disorders which cause and related to diseases such as cancer , diabetes type 2 and age related disorder so they set up a computational method that can accurately map and classify mtDNA deletions and duplications using high-throughput sequencing, Mitochondrial structure alterations used for accurate identification, quantification and visualization of mtDNA deletions and duplications from genomic sequencing data, they tested on human samples with single deletion and duplication and for application this methodology they use mouse models maintenance disease that show the ability to detect this events at low levels of heteroplasm.*

Introduction : *Mitochondria have their own genome, which encodes the oxidative phosphorylation system's essential subunits as well as the RNA molecules needed for mitochondrial translation. Human mtDNA is a small circular molecule of 16.6 billion base pairs with few non-coding regions. Mitochondrial gene activity is almost disrupted when mtDNA events is happened. These structural alteration may occur naturally or as a result of mutations(deletion or duplication in the gene sequence) in the nuclear-encoded mtDNA maintenance machinery. Mitochondrial disorders are often caused by deletions which related to these diseases : Cancer, diabetes, neurodegenerative diseases, and the ageing process. Duplications are less common mutation, but nearly have been described in patients with disease-causing mutations in MGME1 or mice expressing a proof-reading-deficient variant of Pol, (we must know that duplications can form as a result of mutations in mitochondrial replication factors).*

Most mtDNA alterations are heteroplasmic, which means that wild-type mtDNA coexists with mutant variants and because of the complexity of the DNA landscape, it is difficult to characterise mtDNA variants with Low-level heteroplasmic that are explicitly hard to detect .

Related work : *The detection methods Southern blotting and longrangePCR have limited resolution , even a variant present at high levels can remain undetected relying on the primers, probes, or restriction enzymes and beforehand these methods duplications have been incorrectly classified as deletions.*

To solve this and for more accurate mapping of these alterations they use high-throughput sequencing (also known as next-generation sequencing and it intend to sequence DNA and RNA in effeictive manner) , It permit for the survey of mtDNA deletions and duplications in a large body of preceding sequenced data. Identification of discordant paired-end reads or gapped alignment of individual reads to the reference genome are the fundamental bioinformatics principles for determining structural alterations from short read sequencing. However, implementation details may have a significant impact but the small size of the mitochondrial genome simplifies the problem that mitochondrial deletions often occur near repeated sequences which makes it more difficult, and mapping structural events on a circular genome adds to the difficulty.

Many methods for identifying mtDNA deletions have recently been developed from high-throughput short read sequencing, including MitoDel, Splice-Break , eKLIPse , MitoMut and a PERL script, these methods depend on gapped and split alignments to predict deletions, but fail to recognize that every event affecting the arc complementary to the deleted part but correct identification and classification of such alterations is thus an important requirement for any bioinformatics method concern to mtDNA structural changes analysis.

Nowdays the first high-throughput computational data pipeline, MitoSAlt (Mitochondrial Structural Alterations) for identifying, quantifying, and visualising mtDNA deletions and duplications was cautiously instituted using sequencing data , patient samples , and mouse models of mtDNA maintenance disease , MitoSAlt also introduces a method of visualising the results that clearly indicated duplications and deletions, as well as start and end positions , they also evince that disease-causing mutations that influence particular steps in mtDNA replication cause distinct structural alterations in mtDNA using MitoSAlt.

Results : Detection of deletions and duplications with MitoSAlt

For mapping mtDNA , mitoSAlt take a single or paired seq reading as input to predict these events which visualized in a circular plot along with tab delimited tables detailing the breakpoint positions and heteroplasmy levels .The pipeline depends on initial alignments of seq reads to the nuclear and MT genes , there's optional step which accelerate the analusis using HISAT2 *to remove nuclear reads while retaining mtDNA mapped reads and unmapped one but probably be displayed working with species having wide nuclear mtDNA region as mouse to avoid patchwise they reduce mtDNA reading , after that alignment mtDNA using* LAST *process which based on split align and classification of these events resulting in identification of duplication and dele , however in whole genome seq when no mtDNA or nuclear enrichment has been performed, MitoSAlt can compare mitochondrial and nuclear read counts to estimate relative mtDNA levels, which are indicative of mtDNA copy number.*

As many methods mitoAlt depends on identification of reads alignments (split alig or gapped one) to the linear MT genome, we also should know that on the circular genome every split read can represent either a deletion or, alternatively, a duplication of the mtDNA arc complementary to the deletion, and these two possibilities are indistinguishable when using short read sequencing but MitoSAlt handles this by primarily presuming that all events are deletions, then complementation and reclassification as a duplication in cases where the changed mtDNA molecule is deemed incapable of reproducing due to the loss of one or both origins (OriH or OriL; positions are user-definable) , the optimal approach is thus one in which both origin are left unchanged or, if this is not achievable, none are removed. Because only one interpretation will match the criteria while the other will reject them, the deletion/duplication categorization is always non-ambiguous. Furthermore, because mtDNA is circular, both deletions and duplications can result in alignments where the split segments map in reverse order to the linear reference and care has been taken for MitoSAlt to handle and interpret this correctly.

They first validated the pipeline's ability to detect and identify duplications and deletions using a small set of simulated mutations present at high heteroplasmy levels. These were created to overcome

the major classes of possible occurrences that could nevertheless maintain mtDNA replicability. To this end, deletions (2,001–3,999 and the so-called common deletion at 8,470–13,446; coordinates indicate start and end of the affected segment) and duplications (16,069–500, 2,500–3,500, 5,000–6,000, and 9,000–10,000) were introduced into the human reference mitochondrial genome (rCRS), each one at 16.7% heteroplasmy. These were paired with the nuclear genome to simulate a mitochondrial copy number of 6,000, and 10 million reads (5 million 2 126 bp) were obtained using an Illumina HiSeq model. Both alignment steps were accomplished (nuclear and mtDNA using HISAT2 followed by LAST on unmapped and Mt aligned reads). Furthermore, 98.4% of mitochondrial readings (n = 210,235) were mapped to mtDNA resulting in a mean coverage of 1600

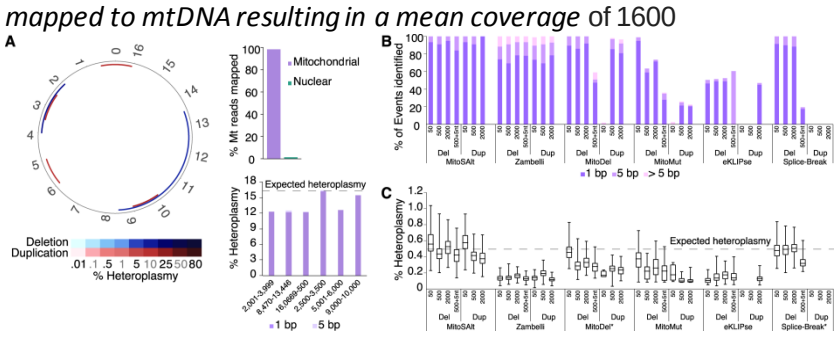


Fig 2. MitoSAlt pipeline performance on simulated data.

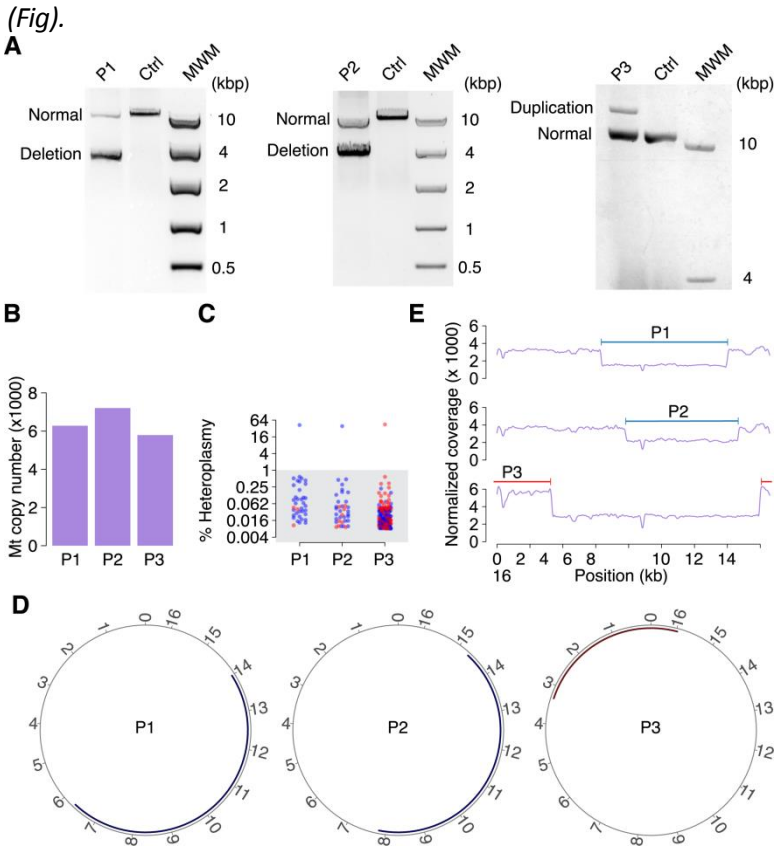
Based on generated sequencing data with two synthetic deletions and four duplications, each at 16.7% heteroplasmy (2 126 bp, 10,000,000 reads, resulting in 2,000 mtDNA coverage). The circular plot depicts segments that have been deleted (blue) or duplicated (red). The upper bar graph shows the percent of Mt reads mapped to the mitochondrial genome, while the lower indicate the heteroplasmy levels estimated by MitoSAlt for each event, Sensitivity evaluation on synthetic sequencing datasets having a large number of low heteroplasmy deletions and duplications of different sizes. Each data set includes 200 minor or major arc events with a heteroplasmy of 0.5 %(2 126 bp, 50,000,000 reads, resulting in 6,000 mtDNA coverage). The term "500+5nt" refers to 500 bp deletions followed by 5 bp non-template random insertions at the breakpoint.

MitoSAlt correctly identified all events at single bp resolution and classified them as deletions or duplications, with heteroplasmy estimates ranging from 12.5 to 16.7% (Fig). A total of 148–213 reads were accurately aligned through each breakpoint (theoretical expectation 266 without any dropouts), with a smaller number of alignments (0–4 reads) supporting breakpoints within 5 bp of the real position(FIG). Despite the presence of nuclear chromosomes in the simulation results, no further events were found. These results validated MitoSAlt's ability to effectively identify and classify deletions and duplications without the need for further false positive detections.

On the same simulated dataset, we compared MitoSAlt's performance with five other existing pipelines (S1 Table). Two of the tools, eKLIPse and the PERL script, are shown in the (Zambelli et al., 2017) All events were identified, however the duplications were characterized as complimentary arc deletions , eKLIPse and Zambelli et al reported nearly 94 and 186 breakpoint-spanning reads, respectively, indicating that eKLIPse has lower sensitivity than MitoSAlt. When breakpoints were bordered by repetitions(the 8,470–13,446 *common deletion), Zambelli et al were less accurate. Splice-Break particularly failed to detect duplications in inverse-order split alignments, implying that the technique is not designed to handle this issue. MitoMut discovered four additional minor deletions in addition to four of the six genuine changes. When simulated reads were constructed using an error model built from empirical data, similar findings were achieved.*

Next, we generated simulated datasets containing large numbers of low heteroplasmy level (0.5%) deletions or duplications of various sizes (50, 500 and 2000 bp). Each dataset contained 200 events of a single type distributed across the major and minor arcs. Additionally, a dataset with 500 bp deletions with 5 bp random insertions was generated, to test the ability to handle non-template insertions at breakpoints. Mitochondrial number was set to 6,000, and 50 million reads (5 million 2 × 126 bp) were generated for each dataset, resulting in a mean coverage of ~5,900× on chrM. All events were detected by MitoSAlt and Zambelli et al, though the latter had lower accuracy with respect to exact determination of breakpoint coordinates (Fig 2B and S1 Table). Remaining tools all showed reduced or no sensitivity with respect to small duplications or duplications in general, as well as deletions with non-template insertions. Heteroplasmy estimates reported by MitoSAlt ranged from 0.38% to 0.57% on average in each dataset (Fig 2C). No events were detected by MitoSAlt or the other tools in a simulated wild type dataset of similar size. MitoSAlt thus compared favorably to other tools in terms of sensitivity and breakpoint coordinate accuracy, in addition to being the only method capable of differentiating between duplications and deletions.

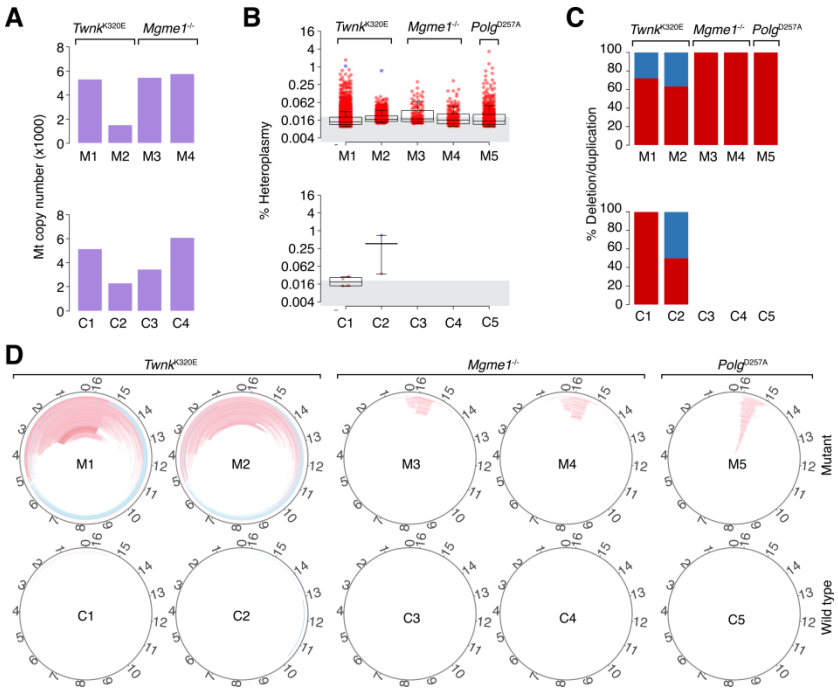
We next tested the MitoSAlt pipeline on muscle biopsy DNA from mitochondrial disease patients with single high-heteroplasmy mtDNA deletions or duplications present at high levels as detected by long-range PCR (LX-PCR). Two patients carried a deletion while the third patient had a duplication (Fig). WGS resulted in a coverage between 83,737× and 121,703× on chrM, and the estimated mtDNA levels, which can be used to predict mtDNA copy number, varied between 5,789 and 7,204 for all samples (Fig). MitoSAlt detected a single high-level heteroplasmy (>50%) deletion or duplication in each patient as expected (Fig). Additional low-level heteroplasmy (<1%) events often had breakpoints close to the main alterations, which may represent inaccurate alignments caused by sequencing errors (Fig). The major breakpoints predicted by MitoSAlt (deletions at 6,330–13,993, 7,826–14,673 and a duplication spanning the D-loop at 15,973–3,326) were compatible with the LX-PCR results and corresponded closely to breakpoints estimated from chrM read depth (Fig).



Moreover, we tested the MitoSAlt pipeline on whole genome sequencing data from three human cancers (liver, pancreatic, and skin), where read depth-based analysis already implied the presence of significant mtDNA events, and observed that these events were verified by our method. These findings provide additional evidence that MitoSAlt can appropriately detect breakpoints and classify events as deletions or duplications based on retention or loss of replication origins.

MitoSAlt detects large numbers of duplications in mouse models of mtDNA disease *: They proceeded to extend the research to more complicated DNA samples after validating the MitoSAlt pipeline on individuals with single large-scale mtDNA duplications or deletions. To that purpose, we acquired DNA from mice that had previously*

been found to have various mtDNA structural abnormalities due to mutations in the Twinkle helicase gene (TwnkK320E; two distinct animals, M1 and M2), which is a knockout of the mtDNA maintenance exonuclease Mgme1 (Mgme1-/-; two different mice, M3 and M4), or mutations in the exonuclease domain of DNA polymerase gamma Poly (PolgD257A; one mouse, M5). All three genes are important for mtDNA maintenance in mice and in humans[19,24,38–40]. The mutant mouse samples (M1-M5) and wild-type controls (C1-C5) were subjected to WGS (TwnkK320E and Mgme1-/-) or sequencing following an mtDNA enrichment protocol (PolgD257A), resulting in a coverage on chrM ranging from 35,913× to 150,182× (Fig). MtDNA level estimates for the TwnkK320E and Mgme1-/- mutants were comparable to wild type control samples , while the use of mtDNA enrichment precluded mtDNA level estimation in the PolgD257A mutant sample. A large number of events were detected in all mutant samples (ranging from 95 to 4841), mostly duplications present at low heteroplasmy levels ,maximum 3.47% and with the average per sample ranging In contrast, the negative control samples were essentially void of structural events .



The circular Mt genome showed two distinct patterns: Mgme1-/- and PolgD257A had a common signature involving multiple, shorter duplications in the non-coding region (NCR), whereas TwnkK320E was characterised by abundant longer duplications extending from a hotspot in the NCR to another hotspot in the middle of the minor arc (FIG), These alteration signatures may reflect similarities and differences in the underlying molecular processes leading to breakpoint formation.

Finally, MitoSAlt is a well-validated tool for accurate mapping of mtDNA structural alterations, especially for detecting and distinguishing between deletions and duplications. MitoSAlt will facilitate further dissection of the mechanistic basis underlying the formation of these types of events, and will enable detailed analysis of samples from patients with mitochondrial diseases.

Methods

MitoSAlt

The MitoSAlt pipeline is comprised of three modules combined into a single pipeline: (1) alignment of sequencing reads (using PERL wrapper third party softwares), (2) parsing aligned reads to identify Mt breakpoints (PERL and R), and (3) plotting the results on the circular Mt genome and analysis of breakpoint repeats (R programming environment).

Alignment of sequencing reads

The raw sequencing reads are aligned to the source genome (Nuclear + Mitochondrial) using HISAT2[46]. HISAT2 is run with default parameters for RNA sequencing and specific parameters are used to customize it for DNA sequencing (—no-temp-splicesite—no-spliced-alignment—max-intronlen 5000). Following the first round of alignment the reads which remain unmapped or are mapped to the mitochondrial genome are extracted and converted to a concatenated FASTQ using Samtools. The FASTQ is realigned to the mitochondrial genome using the lastal (-Q1 -e80), processed using last-split and converted from MAF to BAM and TAB format using maf-convert, where all the binaries are part of the LAST software package. The results in TAB format are parsed in PERL and R to classify the potential deletions and duplications. If the input sequencing data is enriched for mitochondrial DNA/RNA, then the pipeline skips the initial HISAT2 mapping and concatenates the FASTQ files using reformat.sh from BBMap software suite and maps the concatenated reads on the mitochondrial genome using LAST, where the downstream processing remains the same.

Parsing aligned reads to identify Mt breakpoints

The TAB formatted output is parsed in PERL to remove duplicated reads (both wildtype and mutant) and generate three output files a) BED format file with the list of split reads which may support a deletion or a duplication b) BREAKPOINT file with the list of breakpoints identified c) CLUSTER file, which groups the breakpoints at a given distance threshold and estimates the heteroplasmy at a given pair of clustered breakpoints as the ratio of reads supporting the breakpoints by the number of wildtype reads overlapping the breakpoints.

Final report and circular plots

The CLUSTER, and BREAKPOINT files are further used by an R script to generate a final table, classifying each cluster as a duplication or a deletion using the logic described in S1 Fig. This report also contains information about direct repeat sequences overlapping with or flanking the breakpoints. It should be noted that genomic coordinates in the final table refer to start and end positions of the deleted or duplicated segments, rather than junction coordinates. Finally, the breakpoint positions (at the cluster level) are plotted on a circular plot (size of the input mitochondrial genome) as arcs using the R plotrix package, where the individual arcs are colored to indicate whether they represent deletions or duplications, and where the estimated heteroplasmy is indicated by the intensity of the color.

Generation of simulated sequencing data

For the initial evaluation, involving a limited number of high heteroplasmy level events, six mutant mitochondrial reference genomes were generated, each containing a large deletion or duplication as detailed in Results. These were concatenated such that each would be present at a heteroplasmy of 16.7% and included in multiple copies together with the nuclear human chromsomes (hg19 assembly) to emulate an mtDNA copy number of 6000. Next we generated simulated reads using InSilicoSeq, a Python software package[36]. Two different error models were used: the default Illumina Hiseq model (10,000,000 2×126 bp paired-end reads) and an empirical error model base on NextSeq 500 generated whole genome sequencing data (6,000,000 2×76 bp paired-end reads). To evaluate the performance on a larger number of low heteroplasmy events, 6 separate datasets were generated, each containing 200 events as described in Results. These datasets were generated by concatenating mitochondrial genomes containing different deletions or duplications such that each would have a heteroplasmy level of 0.5%. These were combined with a nuclear human genome to emulate mtDNA copy number of 6,000. Simulated reads were generated using the Illumina HiSeq Model (50,000,000 2×126 bp paired-end reads).