ARTICLE Genetics inMedicine

eKLIPse: a sensitive tool for the detection and quantification of mitochondrial DNA deletions from next-generation sequencing data

David Goudenège, PhD[1,2], Celine Bris, PharmD PhD[1,2], Virginie Hoffmann, MS[1], Valerie Desquiret-Dumas, PhD[1,2], Claude Jardel, MD PhD[3], Benoit Rucheton, PhD[3], Sylvie Bannwarth, PhD[4], Veronique Paquis-Flucklinger, MD PhD[4], Anne Sophie Lebre, PharmD PhD[5], Estelle Colin, MD PhD[1,2], Patrizia Amati-Bonneau, MD[1,2], Dominique Bonneau, MD PhD[1,2], Pascal Reynier, MD PhD[1,2], Guy Lenaers, PhD[1] and Vincent Procaccio, MD PhD[1,2]

Purpose: Accurate detection of mitochondrial DNA (mtDNA) alterations is essential for the diagnosis of mitochondrial diseases. The development of high-throughput sequencing technologies has enhanced the detection sensitivity of mtDNA pathogenic variants, but the detection of mtDNA rearrangements, especially multiple deletions, is still poorly processed. Here, we present eKLIPse, a sensitive and specific tool allowing the detection and quantification of large mtDNA rearrangements from single and paired-end sequencing data.

Methods: The methodology was first validated using a set of simulated data to assess the detection sensitivity and specificity, and second with a series of sequencing data from mitochondrial disease patients carrying either single or multiple deletions, related to pathogenic variants in nuclear genes involved in mtDNA maintenance.

Results: eKLIPse provides the precise breakpoint positions and the cumulated percentage of mtDNA rearrangements at a given gene location with a detection sensitivity lower than 0.5% mutant. eKLIPse software is available either as a script to be integrated in a bioinformatics pipeline, or as user-friendly graphical interface to visualize the results through a Circos representation (https://github.com/dooguypapua/eKLIPse).

Conclusion: Thus, eKLIPse represents a useful resource to study the causes and consequences of mtDNA rearrangements, for further genotype/phenotype correlations in mitochondrial disorders.

Genetics in Medicine (2019) 21:1407–1416; https://doi.org/10.1038/s41436-018-0350-8

Keywords: mitochondrial diseases; mitochondrial genome; mtDNA deletions; next-generation sequencing; soft clipping

INTRODUCTION

Mitochondria are responsible for the production of adenosine triphosphate (ATP) through oxidative phosphorylation (OXPHOS), and each cell contains approximately hundreds of mitochondria, the number depending on cellular energy requirements.[1] The 16,569-nucleotide-long mitochondrial DNA (mtDNA), present in hundreds to thousands of copies per cell, is circular, intron-less, and exclusively maternally inherited. It encodes 13 essential polypeptides required for OXPHOS plus 22 transfer RNAs (tRNAs) and 2 ribosomal RNSa (rRNAs) involved in mitochondrial translation.

Due to its proximity to the OXPHOS complexes, which are a major source of reactive oxygen species, the mtDNA accumulates variants at a much higher rate than the nuclear genome. Consequently, due to the stochastic segregation of the mtDNA in daughter cells, the percentage of mutated mtDNAs can drift over time, differently according to the cell lineages, toward either a pure mutated or wild-type condition (homoplasmy), or to a combination of both (heteroplasmy).[1] This often defines the severity of the disease related to the level of heteroplasmy, which is the ratio of mutant to normal mtDNA copies.

Previous studies have shown that about 20% of respiratory chain complex deficiencies are related to heteroplasmic mtDNA alterations.[2–4] Among them, two main categories of mtDNA rearrangements have been identified: single deletions, occurring spontaneously in affected tissues almost exclusively sporadic,[5] and multiple deletions due to inherited pathogenic

[1]MitoLab, UMR CNRS 6015-INSERM U1083, MitoVasc Institute, Angers University, Angers, France; [2]Biochemistry and Genetics Department, Angers Hospital, Angers, France; [3]Biochemistry Department and Genetics Center, APHP, GHU Pitié-Salpêtrière, Paris, France; [4]Université Côte d'Azur, CHU de Nice, INSERM, CNRS, IRCAN, Nice, France; [5]CHU Reims, Hôpital Maison Blanche, Pole de biologie, Service de génétique, Reims, France. Correspondence: Vincent Procaccio (ViProcaccio@chu-angers.fr)
Joint First Authors: David Goudenège and Celine Bris.

Submitted 23 August 2018; accepted: 17 October 2018
Published online: 5 November 2018

variants in mitochondrial nuclear-encoded genes involved in mtDNA maintenance.[6] Those rearrangements are associated with a variety of clinical symptoms.[2–4]

mtDNA deletions occur between regions flanked with short nucleotide repeat sequences, which are either perfect (class I deletions) in 60% of cases or imperfect repeats (class II deletions) in 30% of cases with mismatches between the recombined nucleotide sequences.[7,8] About one-third of all deletion events occur between two direct repeats of 13 base pairs (bp) at positions 8470–8482 and 13447–13459, removing 4977 nucleotides between the *MT-ND5* and *MT-ATP8* genes, the so-called common deletion.[2,8,9]

Of note, mtDNA rearrangements and variants accumulate during life mainly in postmitotic tissues, which are more prone to aging effects.[1] The age-related accumulation of somatic mtDNA variants in postmitotic human tissues correlates with the decline in mitochondrial OXPHOS enzymes during the aging process and is a critical parameter in age-related disorders and neurodegenerative diseases.[10]

The use of massive parallel sequencing technologies now allows the precise analysis of sequence variations in the mtDNA;[11–13] nevertheless, the detection of sporadic and multiple mtDNA deletions still remains elusive. Thus, the accurate identification and quantification of mtDNA rearrangements is challenging, especially for multiple mtDNA deletions, which have a large variety of different lengths and are usually present at very low frequency.

To date, in silico tools have been developed to analyze the copy-number variations (CNVs) from next-generation sequencing (NGS) data, either from paired-end or single-end DNA-sequencing. Their concepts are divided into four categories: read pair, split read, read depth, and de novo assembly methods.[14] Read pair methods predict the rearrangement breakpoints by comparing observed and expected insert sizes, whereas split read methods detect the mapping discrepancy between read pairs. Read depth approaches link the CNV with the depth of variation coverage, and de novo assembly methods search rearrangements by comparing the reference sequence with assembled contigs. The first two methods are the most effective but their usage is restricted to paired-end sequencing platforms. Read depth methods are adapted to all sequencing platforms, but do not predict low mutant loads and overlapping rearrangements, which are characteristics of multiple mtDNA deletions. Finally, de novo assembly approaches are limited by the presence of homopolymers in mtDNA, and therefore are not relevant for the detection of mtDNA deletions.

A fifth approach is based on the concept of soft clipping. The clipping process consists of trimming the extremities of reads that do not align with the reference sequence, to avoid aligning sequencing errors occurring mainly at the ends of the fragments (3'-clipping).[15] The process is called soft clipping if the trimmed sequence is kept, or hard clipping if the sequence is removed, and can occur to either the 5' or 3'-end, or both. Few in silico tools use this approach to predict genome breakpoints by realigning soft-clipped sequences to another genome location. The soft-clipping approach was first used by ClipCrop[15] to detect structural DNA variations, with a good discovery rate and call accuracy. Recently, a new tool dedicated to mtDNA, MitoDel, used a BLAT-based algorithm to realign imperfectly aligned reads and predict deletion breakpoints.[16] Unfortunately, both tools are restricted to paired-end sequences, and MitoDel counts the soft-clipped read realignments without providing the precise deletion load. MATCHCLIP[17] should also be cited because this program detects deletion breakpoints, using reads CIGAR strings, but was originally implemented for paired-end sequences only, and was not dedicated to mtDNA.

Thus, we have designed a novel, automated, and user-friendly software called eKLIPse, which is a sensitive and specific tool based on soft clipping, allowing the univocal detection of mtDNA rearrangements and breakpoints, and their quantification from both single and paired-end NGS data. Scripts, Wndows portable version, and documentation are available at https://github.com/dooguypapua/eKLIPse.

MATERIALS AND METHODS

eKLIPse bioinformatics pipeline for the detection of mtDNA deletions

eKLIPse script was developed using Python2 (v2.7) requiring two types of input, the BAM or SAM alignment files (with header) and the corresponding mitochondrial genome (GenBank format). In case of multiple sequences in the header (WES, WGS), eKLIPse searches the reference with the same length than the GenBank file. Input alignments are read using the package pybam (https://github.com/JohnLonginotto/pybam/blob/master/pybam.py). Soft-clipped sequences are aligned using BLASTN[18] by limiting to only one BLAST result per query and filtering by the percentage of identity and coverage, gap open, and gap extend cost, which are editable parameters. To facilitate the analysis, the downsampling step was performed using SAMtools[19] and the graphical plots were generated using Circos.[20] All eKLIPse steps can be done in parallel by specifying the required thread number (one thread per alignment). A Windows portable version is also available with a graphical user interface developed in Qt5.10.1 including all scripts, packages, and executables.

Read simulation

The simulation of reads was done using the ART simulator[21] for Illumina data (coming from HiSeq 2500 with at least 150-bp fragment lengths) and Curesim[22] for Ion Torrent data. The original and deleted FASTA files of the revised Cambridge Reference Sequence (rCRS; NC_012920.1) were combined to simulate sporadic mtDNA deletions with different sizes (from 51 to 9769 bp), and with the presence of sequencing repeats or not (from no repeat to 13 pb repeats) and heteroplasmy levels. All simulated breakpoints were selected from the MitoBreak database (http://mitobreak.portugene.com/cgi-bin/Mitobreak_home.cgi).[23] For each sequencing technology, a minimal threshold of 200,000 reads (2400× of mean coverage depth) was selected for further analysis.

Analysis of patients carrying mtDNA deletions

To test the sensitivity and specificity of eKLIPse, a series of data was selected from patients carrying sporadic (P1 and P2) or multiple mtDNA deletions related to nuclear-encoded gene pathogenic variants (P3, P4, and P5). Patient samples were collected after obtaining their written informed consent (Institutional Review Board Committee of the University Hospital of Angers; authorization number: AC-2012-1507). Their complete clinical phenotype and familial history are available as supplementary material and methods.

Next-generation sequencing

Ion torrent single-end sequencing was performed on patients P1, P3, and P4 as described elsewhere.[24] The mtDNA was amplified in two overlapping fragments, using polymerase chain reaction (PCR) primer pairs tested first on Rho zero cells devoided of mtDNA to remove the amplification of nuclear mitochondrial DNA sequences (NUMTs) (supplementary material and methods). Libraries were generated using an enzymatic fragmentation approach. Emulsion PCR and enrichment were performed using pooled libraries using an Ion Chef and sequenced on the Ion S5-XL (Thermo Fisher Scientific, Waltham, MA, USA), according to manufacturer protocol.

Illumina paired-end sequencing was performed on patients P2 and P5 as follows. The mtDNA was amplified in two overlapping fragments of 12,932 and 15,913 bp (supplementary material and methods). Libraries were generated using an enzymatic fragmentation approach using the Nextera® XT DNA kit (Illumina, Eindhoven, The Netherlands), following the manufacturer protocol. Sequencing was performed on the MiSeq sequencer (Illumina) with the MiSeq Reagent Nano Kit v2.

Statistical analysis

The performances of eKLIPse on simulated data were assessed using nonparametric tests. The correlations between theoretical deletion length or heteroplasmy and eKLIPse results were compared using a Passing–Bablok regression and calculating the Spearman rank correlation coefficient. Statistical analysis was done using MedCalc version 18.5 statistical software (Medical Software, Ostend, Belgium). Values $p < 0.05$ were considered as statistically significant.

RESULTS

Principle of eKLIPse workflow based on the soft clipping of sequencing reads

eKLIPse starts by retrieving all alignment files (BAM or SAM) and a chosen mtDNA reference (*.gbk) (Fig. **1a, b**). Before processing the data, an additional downsampling step is launched to speed up computing by reducing the read number (200,000 by default). A second step consists in the alignment analysis, which is divided into the read depth and the soft-clipping modules (Fig. **1a, c**). The read depth module counts the number of total and soft-clipped reads for each nucleotide position, which will be used to calculate the percentage of deletions, and visualize the coverage on a final Circos graphical plot. The soft-clipping module extracts each soft-clipped sequence, with the upstream mapped read sequence. As mtDNA deletions generally occur at regions flanked with perfect or imperfect sequence repeats, they can generate misalignments of the sequence for the same deletion because one repeat sequence is lost and the other retained.[2,7,8] For example, the 5-kb common deletion surrounded by two perfect sequence repeats of 13 nucleotides could be assigned to the nucleotide positions 8482–13460 or 8469–13447 depending on the alignment either on the left or the right repeat. Consequently, misalignments may generate breakpoint errors and mutant load bias. By adding the upstream mapped perfect match sequence to the soft-clipped sequence, eKLIPse can detect the sequence repeats, and then left-align all deletion breakpoints (Fig. **1d**).

All forward and reverse soft-clipped and upstream matched nucleotide sequences are written in a FASTA file before being aligned with the BLASTN tool. eKLIPse keeps only one BLAST result per query, and filters the results according to the alignment scores (identity, coverage, gap costs). For instance, the Ion Torrent platform, which uses semiconductor sequencing technology, has a higher insertion/deletion error rate within homopolymeric regions, therefore it is preferable to adjust these filters to the quality of the sequencing data and technology used.

The main step corresponds to the deletion filtering module, called "bidirectional BLAST," which analyzes BLASTN results to search associated forward and reverse soft-clipping positions. This association is validated if the soft-clipping position of a given read corresponds to the soft-clipped sequence alignment of another read and reciprocally (Fig. **1e**). During the filtering process, other situations are removed, such as unidirectional BLAST, or if the number of BLASTs is insufficient or unbalanced between left and right soft-clipped sequences.

Once all reciprocal breakpoints are determined, eKLIPse estimates the deletion load with the formula: $(S \times B)$ / $(N \times SC)$, where S is the total number of soft-clipped reads, B is the number of valid BLASTs, N is the total number of reads, and SC is the number of filtered soft-clipped sequences (>25 bp).

The filtering process is divided into two main steps with editable thresholds. The first step is the minimum BLAST number used to validate a breakpoint, which is related to the sequencing read depth. Secondly, eKLIPse filters out deletions leading to smaller mitochondrial genomes (by default less than 1000 bp).

The final step is the generation of the different output files. eKLIPse generates two tables (.csv files), one with all the deletion prediction results associated to the breakpoint positions, the deletion loads, the BLAST results, and the surrounding repeats (Fig. **1f**). The second file contains the cumulative frequency of deletions per gene (Fig. **1f**, supplementary Table 1).

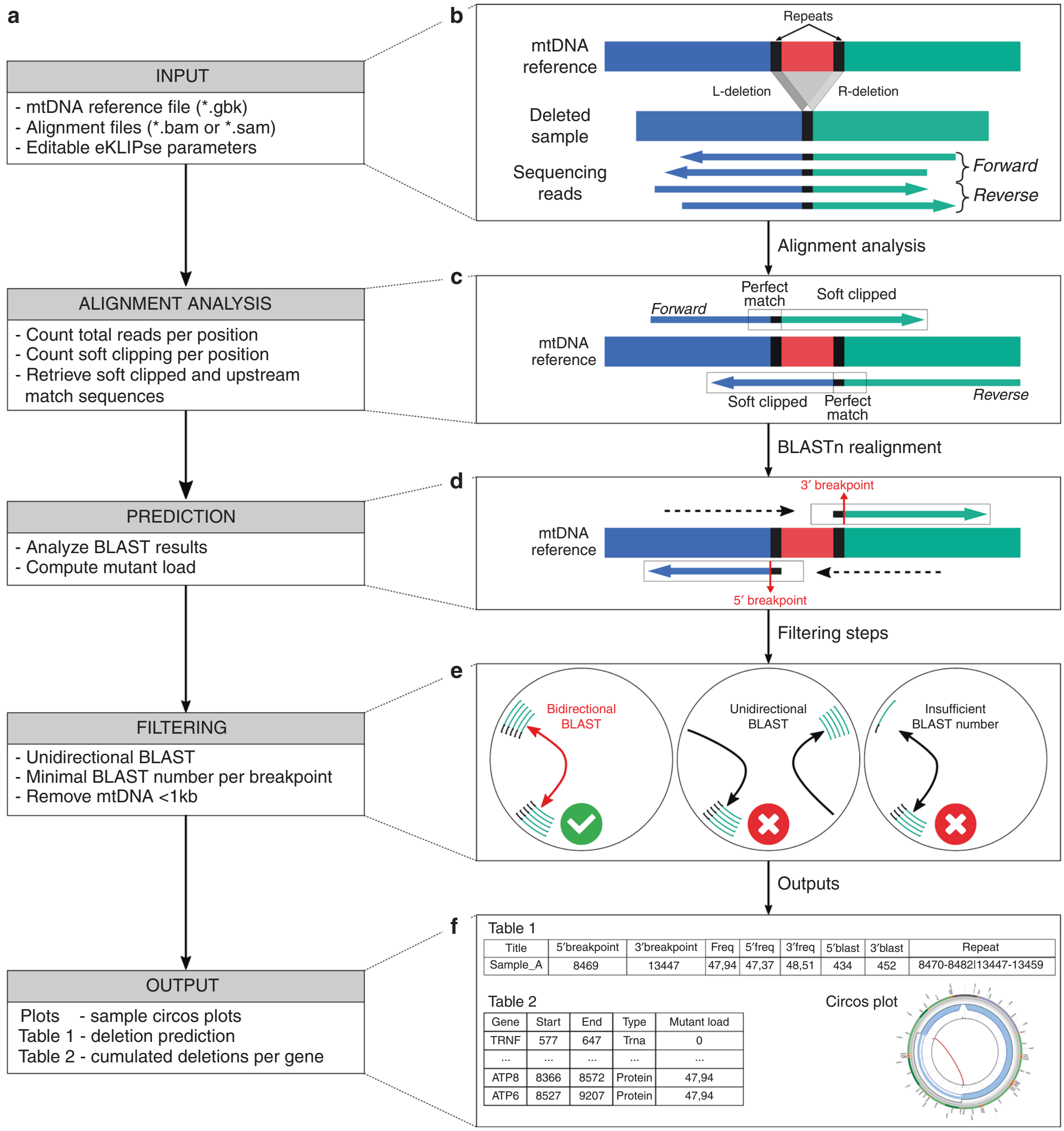


Table 1

Title	5′breakpoint	3′breakpoint	Freq	5′freq	3′freq	5′blast	3′blast	Repeat
Sample_A	8469	13447	47,94	47,37	48,51	434	452	8470-8482I13447-13459

Table 2

Gene	Start	End	Type	Mutant load
TRNF	577	647	Trna	0
...	...	...	...	...
ATP8	8366	8572	Protein	47,94
ATP6	8527	9207	Protein	47,94

Fig. 1 eKLIPse workflow. A four-step process for the detection and quantification of mitochondrial DNA (mtDNA) deletions. (**a**) eKLIPse workflow diagram. (**b**) Detailed alignment of the reference sequence (top) to deleted mtDNA region (in red) analyzed by high-throughput sequencing reads. The two repeat flanking regions are indicated in black and only one repeat is retained in the sequence of the deleted molecule. (**c**) Forward and reverse reads containing the deleted region are aligned to the mtDNA reference leading to two soft-clipping events occurring at positions called forward and reverse soft-clipping positions. (**d**) Soft-clipped and upstream perfect match sequences are retrieved and realigned using BLASTN. (**e**) Breakpoint positions are predicted by searching bidirectional BLAST between forward and reverse soft-clipping positions. Unidirectional BLAST and insufficient BLAST number are removed from the analysis. (**f**) Final outputs of eKLIPse analysis. eKLIPse provides a Circos graphical representation and two descriptive tables summarizing the main findings of each mtDNA deletion (breakpoint nucleotide locations, deletion sizes, mutant loads) and the cumulative deletion percentage related to each mtDNA gene locus.

Finally, eKLIPse creates a Circos plot summarizing most features, visualizing in silico deletion predictions in a very intuitive way (Fig. 2). Indeed, this circular plot uses the genome organization from the reference GenBank file with the different annotated mtDNA genes, to report bidirectional BLASTs represented by red arcs with an intensity proportional to the mutant load, soft-clipping distribution along the genome, cumulated percentage of deletion load, coverage depth, and alignment coverage features.

Analysis of simulated reads

A minimum number of 200,000 sequencing reads were simulated, based on deletion size or heteroplasmy levels, coming from a set of identified mtDNA deletions from MitoBreak either on single reads using CureSim[22] (Fig. 3) or paired reads, with Art.[21]

To evaluate the accuracy of eKLIPse in the detection of breakpoints and deletion lengths, a range of mtDNA deletion sizes combined with flanked repeat sequences of different

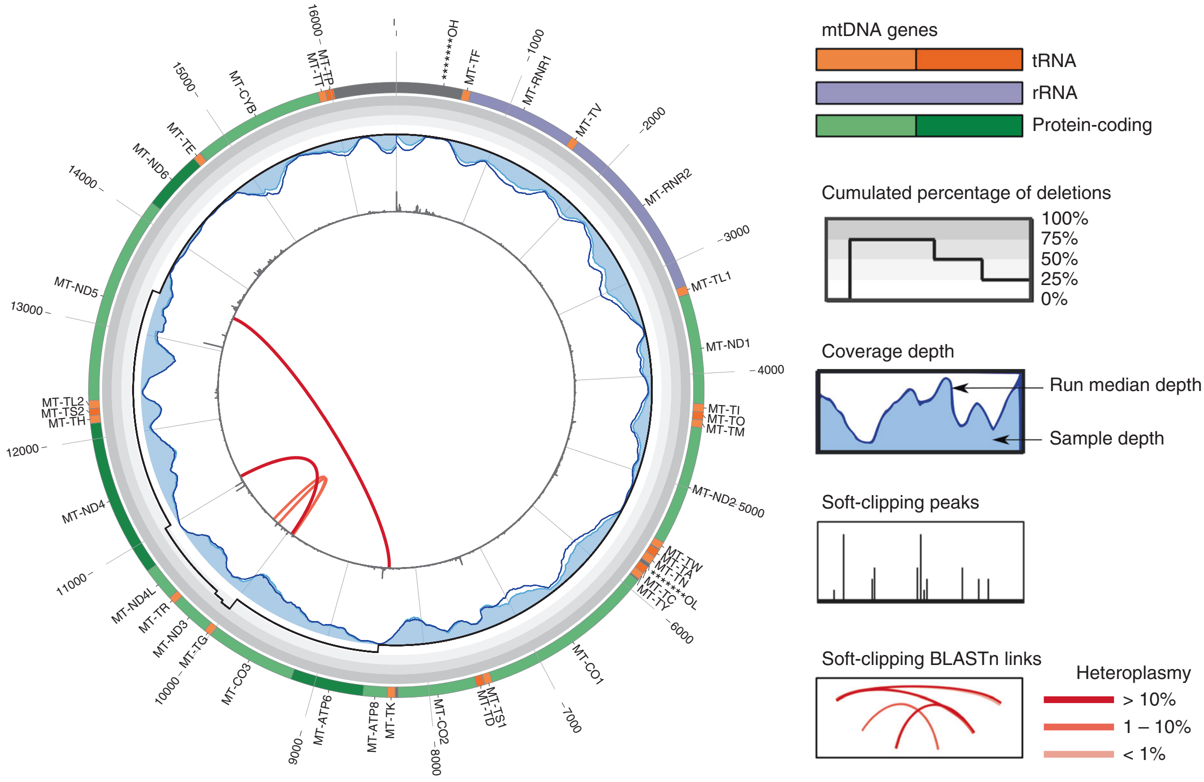


Fig. 2 Graphical representation of eKLIPse. The outer circle depicts the mitochondrial genome with annotated transfer RNAs (tRNAs) (orange), ribosomal RNAs (rRNAs) (purple), protein-coding genes (green), and noncoding regions (black). The gray gradient allows estimating the cumulative percentage of deletions, which are represented by the black line for each mitochondrial DNA (mtDNA) position. For each sample, the depth of coverage at each base is shown as a blue area, whereas the median coverage of all samples is represented as the royal blue line. Soft-clipping peaks are represented within the inner circle as black bars, whose size is proportional to the number of soft-clipped sequences. Reciprocal bidirectional BLASTs are represented by red arcs with intensities (from light to dark red) proportional to the level of mtDNA deletions. The plot was produced using Circos software, version 0.6 (http://circos.ca).

sizes was tested, including m.306-357del (51-bp deletion, 9-bp repeat sequence), m.13761-14252del (491-bp deletion, 8-bp repeat sequence), m.7874-9156del (1282-bp deletion, 3-bp repeat sequence), and m.6124-15893del (9769-bp deletion, no repeat), with a fixed 10% mutant load, on single reads (Fig. **3a**). With the default parameters, eKLIPse allowed the detection of all simulated deletion sizes with a significant correlation between their theoretical length and eKLIPse results, as supported by the Passing–Bablok regression ($p <$ 0.0001) (Fig. **3b**).

The sensitivity of eKLIPse, i.e., the detection of the mutant load, was also investigated for both technologies, using a set of different mutant loads (from 0.1% to 100%) for the common m.8470_13447del mtDNA deletion. This deletion was detectable using single reads with the accurate breakpoint identification at the lowest mutant load of 0.5%, among a total of 200.10^3 reads (Fig. **3c**). eKLIPse was also efficient for heteroplasmy quantification, as shown by the Passing–Bablok regression for single-end reads ($p < 0.0001$) (Fig. **3d**). Similar performances were obtained for a set of simulated sequences from paired-end sequencing (data not shown).

Analysis of patient sequencing data

eKLIPse was then implemented in our bioinformatics pipeline to assess the detection of mtDNA rearrangements in an automated analytical process of NGS data. A set of BAM sequences generated through single-end or paired-end sequencing was analyzed with eKLIPse for patients carrying either sporadic or multiple mtDNA deletions due to pathogenic variants in nuclear-encoded genes involved in the replisome. The results disclosed abnormal Circos profiles (Fig. **4**), confirming the mtDNA profile obtained by gel electrophoresis.

Indeed, for patient 1 (P1) a single deletion of about 5 kb was identified in muscle and uroepithelial samples, although with a stronger band corresponding to a deleted mtDNA molecule in the gel electrophoresis for the muscle sample compared with the uroepithelial sample (Fig. **4a**). In parallel, eKLIPse provided the deletion breakpoints and the heteroplasmy level quantification, revealing the m.8470_13447del. According to the software, the mutant load was estimated to reach 56% in skeletal muscle, but only 3.5% in uroepithelial cells (Fig. **4a**).

Patient 2 (P2) harbored a heteroplasmic single mtDNA deletion revealed by long-range PCR in skeletal muscle with an estimated size of about 4 kb (Fig. **4b**). The retrospective

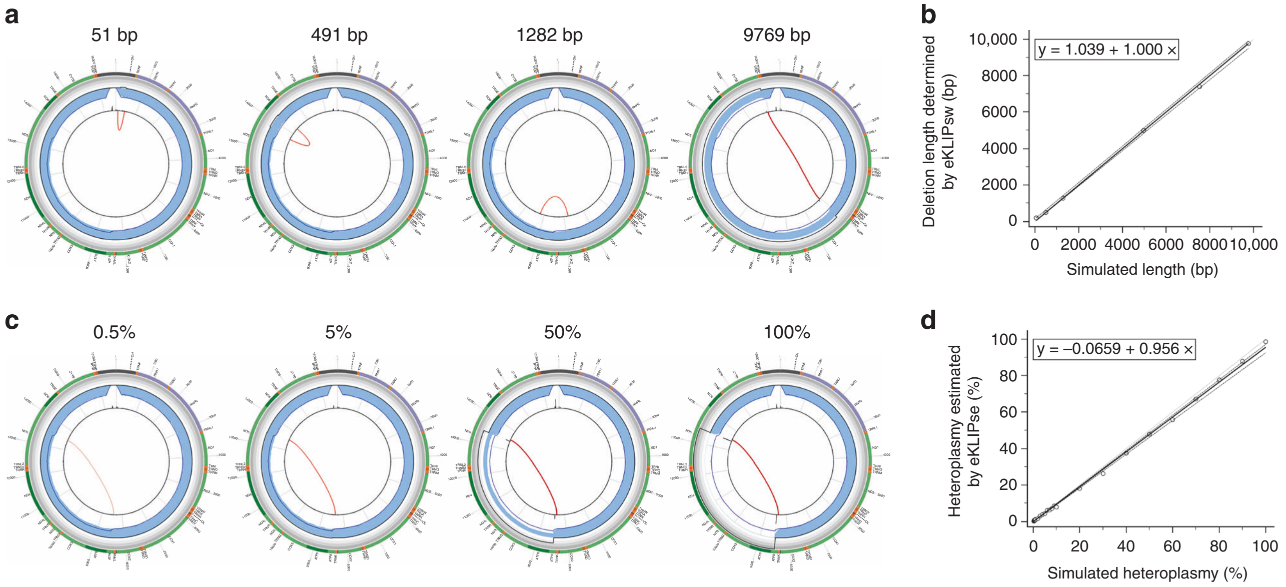


Fig. 3 eKLIPse performances of simulated data. (**a**) Determination of breakpoints and deletion lengths of single-end sequences. Graphical representations of eKLIPse for four deletions of 51, 491, 1282 and 9769 bp deletions with 10% mutant load. (**b**). Correlations between the theoretical deletion length and eKLIPse results were analyzed for a set of seven deletion sizes (51, 491, 879, 1282, 4977, 7522, 9769 bp) with 10% and 20% deletion load by a Passing–Bablok regression ($y = 1.038606 + 0.999791\times$; Spearman correlation coefficient: 1.000; $p < 0.0001$). (**c**) Quantification of heteroplasmy level of single-end sequences. Graphical representations of eKLIPse for the common deletion (m.8470-13447del) at 0.5%, 5%, 50% and 100% of deletion load. (**d**) Correlations between the theoretical heteroplasmy level and eKLIPse results were analyzed for a wide range of heteroplasmy (0.1–1% per 0.1% step, 1–10% per 1% step, 10–100% per 10% step) by a Passing–Bablok regression ($y = -0.0658805 + 0.956424\times$; Spearman correlation coefficient: 0.996; $p <$ 0001).

computing of the FASTQ files of paired-end mtDNA sequencing by eKLIPse disclosed the m.8559_12385del (3826 bp) with a 77% mutant load.

Patient 3 (P3) carried two compound heterozygous variants in *DGUOK* (NM_080916.1), c.462 T>A (p.Asn154Lys) and c.677 A>G (p.His226Arg), classified as likely pathogenic according to American College of Medical Genetics and Genomics (ACMG) guidelines.[25] eKLIPse confirmed the presence of several mtDNA fragments revealed by gel analysis (Fig. **4c**) and identified 241 deletions ranging from 798 bp to 9527 bp (mean = 4959 bp; SD = 1974 bp), each deletion with a heteroplasmy load between 0.01% and 4.03% (mean = 0.38%; SD = 0.66%), as illustrated in the Circos graph (Fig. **4c**) and listed in the Supplementary Table 1, as an example of outputs provided by the software. Importantly, the cumulated percentage of deletions reached 75% between *MT-COIII* and *MT-ND5* genes.

Patient 4 (P4) carried two compound heterozygous variants of *POLG* (NM_001126131.1), c.925 C>T (p.Arg309Cys) and c.1789C>T (p.Arg597Trp), classified as pathogenic and likely pathogenic respectively. Gel electrophoresis analysis of long-range PCR products revealed several bands included in a smear beneath the wild-type mtDNA, revealing a multiple deletion process (Fig. **4c**). eKLIPse analysis of NGS data identified a total of 338 deletions from 1652 bp to 7578 bp (mean = 5015 bp; SD = 1264 bp) with a mutant load from 0.03% to 3.4% (mean = 0.27%; SD = 0.43%) (Supplementary Table 1). The cumulated percentage of deletions reached 64% of mitochondrial genome lacking the *MT-ND4* gene (Supplementary table 1).

Patient 5 (P5) carried two compound heterozygous pathogenic variants in *SLC25A4* (NM_001151.3), the pathogenic variant c.368 C>A (p.Ala123Asp) and a likely pathogenic variant c.390del (p.Phe130Lefs*41). Retrospective computing of the FASTQ files of the paired-end mtDNA sequencing identified 102 deletions (Circos graph in Fig. **4d**) from 1425 bp to 12808 bp (mean = 7622 bp; SD = 1830 bp) with the mutant loads ranging from 0.04% to 12.9% (mean = 0.39%; SD = 1.30%), confirming the analysis of long-range PCR (Fig. **4d**). A maximal cumulative percentage of 38.9% of deletions was reached between *MT-ATP6* and *MT-ND4* (Supplementary table 1).

DISCUSSION

Multiple mtDNA rearrangements contribute to a significant part of the pathophysiology of mitochondrial diseases, in particular those related to aging and neurodegenerative disorders.[1] Thus efficient sensitive and specific bioinformatics software characterizing this process from routine NGS mtDNA sequencing is mandatory for the diagnosis of mitochondrial diseases and to better explain the contribution of genome instability in pathological mechanisms. The eKLIPse software, based on soft clipping, was first developed for the detection and the quantification of mtDNA deletions from single-end reads generated through Ion Torrent technology, and then adapted to paired-end sequencing data generated by the Illumina technology. The soft-clipping

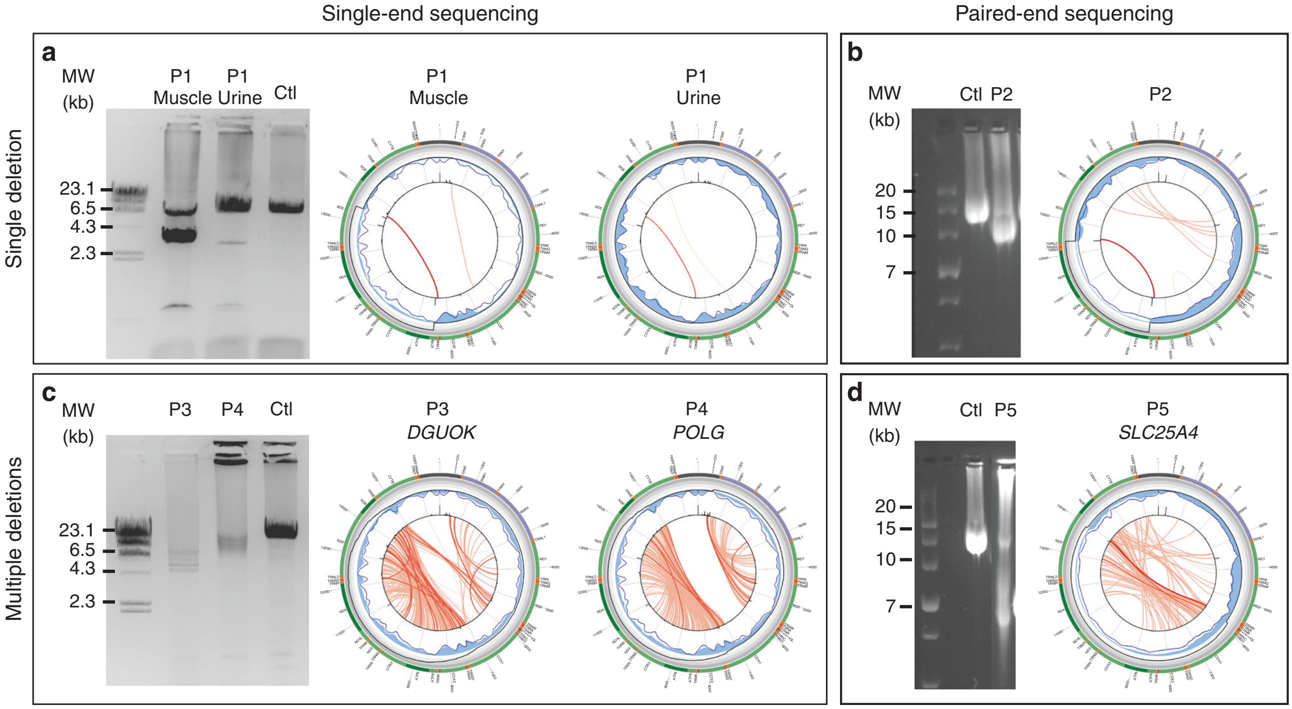


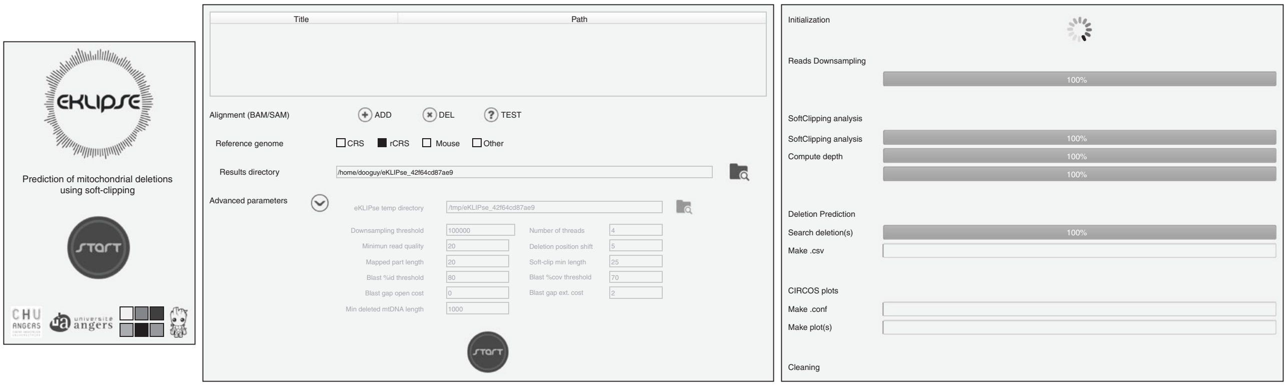
Fig. 4 Analysis of patients carrying mitochondrial DNA (mtDNA) deletions. (**a**) Detection of single deletion from single-end sequencing. Gel electrophoresis with Patient P1 harboring a heteroplasmic single deletion of approximately 5 kb detected by long-range polymerase chain reaction (PCR) analysis in muscle and uroepithelial cells. Circos graphical representations with the common deletion (m.8470_13447del) in muscle (left) and urine (right) samples with respectively 56% and 3.5% of mutant loads. (**b**) Detection of single deletion from paired-end sequencing. Gel electrophoresis with patient P2 carrying a heteroplasmic single deletion of about 4 kb in muscle. Circos graphical representation showing the presence of a 3826-bp deletion (m.8559_12385del) at 77% mutant load. (**c**) Detection of multiple deletions from single-end sequencing. Gel electrophoresis with patient P3 harboring two pathogenic variants in *DGUOK* and patient P4 carrying two likely pathogenic variants in *POLG*. Circos graphical representations of patients P3 (left) and P4 (right) with the number of deletions (red arcs) identified by eKLIPse. (**d**) Detection of multiple deletions from paired-end sequencing. mtDNA long-range PCR gel electrophoresis with patient P5 carrying two pathogenic variants in *SLC25A4* associated with multiple deletions. Circos graphical representation of patient P5 showing multiple deletions (red arcs).

process consists of trimming the ends of reads that do not align with the reference sequence, to avoid sequencing errors that occur more frequently at the ends of the fragments sequenced, and was initially used to investigate nuclear genome rearrangements.[15] After an initial validation step using simulated data to evaluate the sensitivity and specificity of the detection of rearrangements, the tool was tested on NGS data from patients carrying either sporadic or multiple mtDNA deletions, previously identified only by long-range PCR gel electrophoresis.

Performances of eKLIPse

Until recently, mtDNA deletions were identified either from NGS data due to the reduction of mitochondrial sequence coverage,[26] real-time PCR,[27] or by Southern blotting,[28] with limitations in the detection of the mutant load. To gain sensitivity and insights in the pathological process, we designed eKLIPse and assessed its performances on data generated through the two most frequent NGS technologies used to date, i.e., Ion Torrent and Illumina platforms.

eKLIPSe sensitivity was highly correlated to the total number of sequencing reads. For the in silico validation, a read aligned on the mtDNA reference sequence was required, allowing the detection of low mutant loads with a limit of detection reaching a ground level of 0.5%. This sensitivity was approximately five times higher than those reached by digital PCR technology, with the assessment of minimal mtDNA deletion load of 2.8% (ref. [29]). But, because the performances of eKLIPse are conditioned by the sequencing quality and the coverage depth, the limit of detection may vary according to the sequencing platform and the number of aligned reads. For patients analyzed in our clinical diagnostic department, the median coverage was $1.74.10^{6}$ aligned reads, enabling a detection level of mtDNA deletions at mutant load below 0.5%. This increased sensitivity improved the detection of multiple mtDNA deletions related to aging and mitochondrial diseases, consistently reducing false negative rates. Whatever the variant load, eKLIPse accurately determined the deletion breakpoints and length, even in the case of overlapping multiple deletions. As previously described, for the three patients carrying multiple mtDNA deletions, 85% of deletions were located within the major arc[30] and most deletion breakpoints occurred near direct or indirect sequence repeats.[7] The gain of sensitivity did not affect the tool specificity, false-positive deletion detection being minimized by the minimal requirements of the reciprocal bidirectional



Alignment (BAM/SAM)

Reference genome ☐ CRS ■ rCRS ☐ Mouse ☐ Other

Results directory: /home/dooguy/eKLIPse_42f64cd87ae9

Advanced parameters

eKLIPse temp directory: /tmp/eKLIPse_42f64cd87ae9

Downsampling threshold: 100000

Number of threads: 4

Minimun read quality: 20

Deletion position shift: 5

Mapped part length: 20

Soft-clip min length: 25

Blast %id threshold: 80

Blast %cov threshold: 70

Blast gap open cost: 0

Blast gap ext. cost: 2

Min deleted mtDNA length: 1000

Reads Downsampling: 100%

SoftClipping analysis: 100%

Compute depth: 100%

100%

Search deletion(s): 100%

Make .csv

Make .conf

Make plot(s)

Fig. 5 Graphical user interface of eKLIPse. Screenshots of eKLIPse from left to right: eKLIPse homepage, window for launching eKLIPse analysis with all tunable parameters, and window to display and monitor the analysis in real time.

BLAST of both perfect upstream match extension and soft-clipping sequences. Similar to other bioinformatics tools, the quality of eKLIPse results will be highly dependent on the sequencing quality of NGS data, especially removing NUMTs during the amplification process. Therefore, eKLIPse avoided the identification of false deletions due to degraded DNA samples, unlike Southern blot,[31] or leading to bad sequencing quality associated with numerous soft-clipped sequences or a reduction of the coverage depth due to PCR amplification problems. In addition to the detection of rearrangements, eKLIPse provided an accurate estimation of the deletion load, similar to other digital PCR methods[29] or real-time PCR,[27,32] but with greater detection sensitivity.

eKLIPse is a user-friendly bioinformatics tool for the detection of mtDNA deletions

eKLIPse was initially developed for the diagnostic clinical setting with a Windows interface, for an easy launch of analyses with a graphical user interface (Fig. 5). A user guide is also available to facilitate its implementation (https://github.com/dooguypapua/eKLIPse). In addition, the script is also available, allowing the tool to be easily integrated into a local pipeline by bioinformaticians. eKLIPse provides a wide set of results: a user-friendly and fast identification of mtDNA deletions with a graphical representation, and two tables for a complete description of these deletions (e.g., breakpoints, mutant loads) (Fig. **1f**). The consequences of mtDNA rearrangements with the cumulative percentage of deletions according to the mtDNA gene localization are also provided.

Digital PCR was also recently described as a method to quantify mtDNA deletions,[29] but it requires the development of a specific assay for each gene, whereas eKLIPse provides the information for all the different mtDNA loci at the same time. Another advantage of eKLIPse is its modularity. Initially developed for human mtDNA, eKLIPse can be used to analyze mtDNA from other species, with the possibility of loading a new specific reference sequence, conversely to other softwares.[16,33] In addition, all the default settings can be changed to match specific mtDNA NGS strategies or research hypotheses.

eKLIPse allows a better understanding of the consequences of mtDNA deletions

Clinical features of mitochondrial diseases associated with mtDNA sporadic or multiple deletions are highly heterogeneous.[6,30] The pathophysiology of mtDNA deletion syndrome results in protein synthesis defects, concordant with the fact that all deletions so far identified remove at least one structural tRNA or rRNA gene, which are essential for mitochondrial translation.[23] However, no genotype–phenotype correlation has been made related to the clinical severity, regarding the molecular characteristics of deletions in patients carrying single[7,34–36] or multiple deletions associated with mtDNA maintenance defects.[37] It was suggested that high mutant loads, imperfect sequence repeats surrounding the deletion breakpoints, and a reduction of the percentage of the "common" deletion are associated with higher severity and earlier onset of the clinical presentation.[7] eKLIPse has the great advantage to provide a complete list of deleted mtDNA genes, especially those encoding tRNAs, with a cumulated heteroplasmy load for a given mtDNA gene, which can be then correlated to the patient phenotype. Examples provided in this study showed that the accumulation of mtDNA deletions in patients carrying multiple deletions could reach more than 75% of one or many mitochondrial genes. Hence, it appears important to identify and quantify the genes that have been deleted, to evaluate the consequences of mtDNA rearrangements and establish genotype–phenotype correlations.

Another interesting field with the potential to improve patient management is the distinction between multiple mtDNA deletions associated with aging and those due to mtDNA maintenance defects. Interestingly enough, during the aging process, some mtDNA deletions occur preferentially in healthy tissues in the D-loop or at the origin of the L-strand replication (OL), with a preferential breakpoint between nucleotides 16001 and 16100, whereas deletions related to mtDNA maintenance defects are not associated with a loss of the origin of replication.[2] Hypothetically, the analysis with eKLIPse of larger cohorts of elderly patients compared with patients harboring pathogenic nuclear variants in genes involved in mtDNA maintenance will highlight other molecular features that can distinguish both situations. The latter analysis may potentially lead to the identification of mtDNA multiple deletion profiles specific for the different alterations of mtDNA maintenance defects, establishing correlations between the identity of the mutated gene and the mtDNA deletion profile. Future studies of large patient cohorts harboring mtDNA rearrangements will be required to explore the molecular consequences of multiple mtDNA deletions with respect to clinical presentations. In addition, eKLIPse will be useful for other applications such as evaluating the acquisition of somatic deletions during tumorigenesis by comparing healthy tissues and the corresponding tumor.

In conclusion, eKLIPse is a sensitive, specific, and accurate new tool for the detection and the quantification of mtDNA deletions from NGS, regardless of the mtDNA library construction or the NGS technology used. eKLIPse can be easily implemented in an in-house bioinformatics pipeline, allowing a complete one-step molecular analysis of mtDNA through NGS, to provide an user-friendly way to launch the analysis and visualize the results in a standard presentation that can be associated with a clinical diagnostic. In this respect, better detection and quantification of the consequences of mtDNA multiple deletion syndrome in mitochondrial disease patients will provide key information, allowing important genotype–phenotype correlations and unraveling precisely the pathophysiological mechanisms related to mitochondrial genome instability.

ELECTRONIC SUPPLEMENTARY MATERIAL

The online version of this article (https://doi.org/10.1038/s41436-018-0350-8) contains supplementary material, which is available to authorized users.

ACKNOWLEDGEMENTS

This work was supported by grants from Association contre les Maladies Mitochondriales, and the University and Hospital of Angers.

DISCLOSURE

The authors declare no conflicts of interest.

REFERENCES

1. Wallace DC, Lott MT, Procaccio V. Mitochondrial medicine: the mitochondrial biology and genetics of metabolic and degenerative diseases, cancer, and aging. In: Rimoin D, Pyeritz R, Korf B, eds. Emery and Rimoin's principles and practice of medical genetics. Oxford: Academic Press; 2013. p. 1–153.
2. Damas J, Samuels DC, Carneiro J, Amorim A, Pereira F. Mitochondrial DNA rearrangements in health and disease—a comprehensive study. Hum Mutat. 2014;35:1–14.
3. Kirby DM, McFarland R, Ohtake A, et al. Mutations of the mitochondrial ND1 gene as a cause of MELAS. J Med Genet. 2004;41:784–789.
4. Bannwarth S, Procaccio V, Lebre AS, et al. Prevalence of rare mitochondrial DNA mutations in mitochondrial disorders. J Med Genet. 2013;50:704–714.
5. Ballinger SW, Shoffner JM, Gebhart S, Koontz DA, Wallace DC. Mitochondrial diabetes revisited. Nat Genet. 1994;7:458–459.
6. El-Hattab AW, Craigen WJ, Scaglia F. Mitochondrial DNA maintenance defects. Biochim Biophys Acta. 2017;1863:1539–1555.
7. Sadikovic B, Wang J, El-Hattab A, et al. Sequence homology at the breakpoint and clinical phenotype of mitochondrial DNA deletion syndromes. PLoS One. 2010;5:e15687.
8. Holt IJ, Harding AE, Morgan-Hughes JA. Deletions of muscle mitochondrial DNA in mitochondrial myopathies: sequence analysis and possible mechanisms. Nucleic Acids Res. 1989;17:4465–4469.
9. Schon EA, Rizzuto R, Moraes CT, Nakase H, Zeviani M, DiMauro S. A direct repeat is a hotspot for large-scale deletion of human mitochondrial DNA. Science. 1989;244:346–349.
10. Keogh MJ, Chinnery PF. Mitochondrial DNA mutations in neurodegeneration. Biochim Biophys Acta. 2015;1847:1401–1411.
11. Seneca S, Vancampenhout K, Van Coster R, et al. Analysis of the whole mitochondrial genome: translation of the Ion Torrent Personal Genome Machine system to the diagnostic bench? Eur J Hum Genet. 2015;23:41–48.
12. Vancampenhout K, Caljon B, Spits C, et al. A bumpy ride on the diagnostic bench of massive parallel sequencing, the case of the mitochondrial genome. PLoS One. 2014;9:e112950.
13. Wong LJ, Boles RG. Mitochondrial DNA analysis in clinical laboratory diagnostics. Clin Chim Acta. 2005;354:1–20.
14. Pirooznia M, Goes FS, Zandi PP. Whole-genome CNV analysis: advances in computational approaches. Front Genet. 2015;6:138.
15. Suzuki S, Yasuda T, Shiraishi Y, Miyano S, Nagasaki M. ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. BMC Bioinformatics. 2011;12 Suppl 14:S7.
16. Bosworth CM, Grandhi S, Gould MP, LaFramboise T Detection and quantification of mitochondrial DNA deletions from next-generation sequence data. BMC Bioinformatics 2017;18:407.
17. Wu Y, Tian L, Pirastu M, Stambolian D, Li H. MATCHCLIP: locate precise breakpoints for copy number variation using CIGAR string by matching soft clipped reads. Front Genet. 2013;4:157.
18. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
19. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–2079.
20. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–1645.
21. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28:593–594.
22. Caboche S, Audebert C, Lemoine Y, Hot D. Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. BMC Genomics. 2014;15:264.
23. Damas J, Carneiro J, Amorim A, Pereira F. MitoBreak: the mitochondrial DNA breakpoints database. Nucleic Acids Res. 2014;42 Database issue: D1261–1268.
24. Boucret L, Bris C, Seegers V, et al. Deep sequencing shows that oocytes are not prone to accumulate mtDNA heteroplasmic mutations during ovarian ageing. Hum Reprod. 2017;32:2101–2109.
25. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17:405–424.
26. Wong LJ. Challenges of bringing next generation sequencing technologies to clinical molecular diagnostic laboratories. Neurotherapeutics. 2013;10:262–272.
27. He L, Chinnery PF, Durham SE, et al. Detection and quantification of mitochondrial DNA deletions in individual cells by real-time PCR. Nucleic Acids Res. 2002;30:e68–e68.
28. Tonska K, Piekutowska-Abramczuk D, Kaliszewska M, et al. Molecular investigations of mitochondrial deletions: evaluating the usefulness of different genetic tests. Gene. 2012;506:161–165.
29. Belmonte FR, Martin JL, Frescura K, et al. Digital PCR methods improve detection sensitivity and measurement precision of low abundance mtDNA deletions. Sci Rep. 2016;6:25186.
30. Pitceathly RD, Rahman S, Hanna MG. Single deletions in mitochondrial DNA–molecular mechanisms and disease phenotypes in clinical practice. Neuromuscul Disord. 2012;22:577–586.
31. Moraes CT, Atencio DP, Oca-Cossio J, Diaz F. Techniques and pitfalls in the detection of pathogenic mitochondrial DNA mutations. J Mol Diagn. 2003;5:197–208.

32. Grady JP, Murphy JL, Blakely EL, et al. Accurate measurement of mitochondrial DNA deletion level and copy number differences in human skeletal muscle. PLoS One. 2014;9:e114462.
33. Guo Y, Li J, Li CI, Shyr Y, Samuels DC. MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. Bioinformatics. 2013;29:1210–1211.
34. Moraes CT, Sciacco M, Ricci E, et al. Phenotype-genotype correlations in skeletal muscle of patients with mtDNA deletions. Muscle Nerve Suppl. 1995;3:S150–153.
35. Spelbrink JN, Van Oost BA, Van den Bogert C. The relationship between mitochondrial genotype and mitochondrial phenotype in lymphoblasts with a heteroplasmic mtDNA deletion. Hum Mol Genet. 1994;3:1989–1997.
36. Yamashita S, Nishino I, Nonaka I, Goto Y. Genotype and phenotype analyses in 136 patients with single large-scale mitochondrial DNA deletions. J Hum Genet. 2008;53:598–606.
37. Mancuso M, Filosto M, Oh SJ, DiMauro S. A novel polymerase γ mutation in a family with ophthalmoplegia, neuropathy, and parkinsonism. Arch Neurol. 2004;61:1777–1779.