International Journal of Legal Medicine

Assessment of Illumina® Human mtDNA Genome assay: workflow evaluation with development of analysis and interpretation guidelines

--Manuscript Draft--

Manuscript Number:	
Full Title:	Assessment of Illumina® Human mtDNA Genome assay: workflow evaluation with development of analysis and interpretation guidelines
Article Type:	Original Article
Corresponding Author:	Marina Korolija Forensic Science Centre "Ivan Vucetic" CROATIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Forensic Science Centre "Ivan Vucetic"
Corresponding Author's Secondary Institution:	
First Author:	Viktorija Sukser
First Author Secondary Information:	
Order of Authors:	Viktorija Sukser
	Filip Rokić
	Lucija Barbarić
	Marina Korolija
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	Mitochondrial DNA (mtDNA) is a small but significant part of the human genome, whose applicability potential has gradually increased with the advent of massively parallel sequencing (MPS) technology. Knowledge of the particular workflow, equipment and reagents used, along with extensive usage of negative controls to monitor all preparation steps constitute the prerequisites for confident reporting of results. In this study, we performed assessment of Illumina® Human mtDNA Genome assay on MiSeq FGx TM instrument. Through analysis of several types of negative controls, as well as mtDNA positive controls, we established thresholds for data analysis and interpretation, consisting of several components: minimum read depth (220 reads), minimum quality score (41), percentage of minor allele sufficient for analysis (3.0%), percentage of minor allele sufficient for interpretation (6.0%), and percentage of major allele sufficient for homoplasmic variant call (97.0%). Based on these criteria, we defined internal guidelines for analysis and interpretation of mtDNA results obtained by MPS. Our study shows that the whole mtDNA assay on MiSeq FGx TM produces repeatable and reproducible results, independent of analyst, which are also concordant with Sanger-type sequencing results for mtDNA control region, as well as with MPS results produced by NextSeq®. Overall, established thresholds and interpretation guidelines were successfully applied for sequencing of complete mitochondrial genomes from high-quality samples. The underlying principles and proposed methodology on definition of internal laboratory guidelines for analysis and interpretation of MPS results may be applicable to similar MPS workflows, primarily in forensic genetics and molecular diagnostics.
Author Comments:	
Suggested Reviewers:	Mitchell Holland mmh20@psu.edu
	Walther Parson

	walther.parson@i-med.ac.at

Title

Assessment of Illumina® Human mtDNA Genome assay: workflow evaluation with development of analysis and interpretation guidelines

Viktorija Sukser[1], Filip Rokić[2], Lucija Barbarić[1], Marina Korolija[1,*]

[1] Biology and Fibres Department, Forensic Science Centre "Ivan Vučetić", Ilica 335, 10000 Zagreb, Croatia

[2] Laboratory for Advanced Genomics, Division of Molecular Medicine, "Ruđer Bošković" Institute, Bijenička cesta 54, 10000 Zagreb

* Correspondence should be addressed to the following author:
Marina Korolija, PhD
Biology and Fibres Department
Forensic Science Centre "Ivan Vučetić"
Ilica 335, 10000 Zagreb, Croatia
mkorolija@mup.hr

Abstract

Mitochondrial DNA (mtDNA) is a small but significant part of the human genome, whose applicability potential has gradually increased with the advent of massively parallel sequencing (MPS) technology. Knowledge of the particular workflow, equipment and reagents used, along with extensive usage of negative controls to monitor all preparation steps constitute the prerequisites for confident reporting of results. In this study, we performed assessment of Illumina® Human mtDNA Genome assay on MiSeq FGx™ instrument. Through analysis of several types of negative controls, as well as mtDNA positive controls, we established thresholds for data analysis and interpretation, consisting of several components: minimum read depth (220 reads), minimum quality score (41), percentage of minor allele sufficient for analysis (3.0%), percentage of minor allele sufficient for interpretation (6.0%), and percentage of major allele sufficient for homoplasmic variant call (97.0%). Based on these criteria, we defined internal guidelines for analysis and interpretation of mtDNA results obtained by MPS. Our study shows that the whole mtDNA assay on MiSeq FGx™ produces repeatable and reproducible results, independent of analyst, which are also concordant with Sanger-type sequencing results for

mtDNA control region, as well as with MPS results produced by NextSeq®. Overall, established thresholds and interpretation guidelines were successfully applied for sequencing of complete mitochondrial genomes from high-quality samples. The underlying principles and proposed methodology on definition of internal laboratory guidelines for analysis and interpretation of MPS results may be applicable to similar MPS workflows, primarily in forensic genetics and molecular diagnostics.

Keywords

MiSeq, mitochondrial DNA, Nextera XT, evaluation, analysis thresholds

Declarations

Funding: This work was supported by Ministry of the Interior of the Republic of Croatia.

Conflicts of interest: The authors declare that they have no conflict of interest.

Ethics approval: This study involved samples collected from human participants. All procedures performed in the study were in accordance with the institutional and national ethical standards.

Consent to participate: Informed consent was obtained from all individual participants included in this study.

Consent for publication: Not applicable.

Availability of data and material: The datasets generated and analysed during this study are available from the corresponding author on reasonable request.

Code availability: Not applicable.

Acknowledgements

The authors thank all participants in the study for their valuable contributions in the form of samples and detailed informed consents.

Authors are also thankful to Oliver Vugrek, PhD, Head of Laboratory for Advanced Genomics, Division of Molecular Medicine at “Ruđer Bošković” Institute, and their laboratory staff for collaboration in concordance study.

Authors are grateful to Sara Rožić and Ivana Račić, PhD, who made valuable contributions in the experimental part of this study.

1. Introduction

For such a relatively small portion of the human genome, mitochondrial DNA (mtDNA) exhibits extraordinary variability and unique features. The size of human mitochondrial genome approximates 16,569 basepairs (bp; length may slightly vary due to insertions and deletions), which is on a scale of about 1:200,000 compared to the nuclear DNA. Despite its diminutiveness, mtDNA is essential for cellular energy production and, thus, presents a vital part of our genome. It is enclosed within double-layered membranes of cell's energy factories – mitochondria. Due to its well-protected location, as well as circular nature, and the fact that there may be as many as several thousand copies of mtDNA per one cell (as opposed to nuclear DNA, present only in two copies per cell), this small genome is more resistant to environmental conditions and degradation than nuclear DNA. Therefore, it may well be the only source of genetic information recoverable in some cases, and even though it may not be used for individual identification (as all maternal relatives have the same mitochondrial genome sequence, with tolerable variations in indels and heteroplasmies), it is certainly preferable to no result at all. The aforementioned characteristics have established mtDNA as a valuable source in many fields of science, such as evolutionary biology, molecular anthropology, forensics etc. [1].

Until fairly recently, the only part of mtDNA extensively investigated was the control region (CR), approximately 1100 bp in length, encompassing the origin of replication, other regulatory elements and hypervariable regions (or segments; HVS-I, HVS-II and HVS-III). The most of mitochondrial sequence variation is concentrated in HVS, and mtDNA CR analysis by Sanger-type sequencing (STS) has become the gold standard employed in routine forensic casework, where sample material is scarce and challenging to process for various reasons (degradation, inhibitors, etc.). However, CR equals only around 7% of complete mitochondrial genome, and in cases of more common mitochondrial haplotypes, this information alone cannot provide the resolution sufficient for forensic purposes [2]. Therefore, sequencing of the entire mtDNA clearly has great value, as inter-individual variation comes to the fore by revealing all 16,569 bp length of genetic information. Besides ethical and legal issues which stem from accessing the coding region sequence, analysis of whole mitochondrial genomes was simply not feasible previously with Sanger sequencing method, as it was costly, laborious, time-consuming and nearly impossible to apply on a large scale – few studies endeavoured to employ STS to produce whole mtDNA data (e.g. [3, 4]). In addition, population samples usually contain abundance of genetic material of high quality, whereas forensic casework samples rarely come in such pristine state, meaning STS of whole mtDNA would be even more difficult in the latter case.

Over the recent years we have witnessed great technological leaps that brought about next generation of sequencing platforms and chemistries, or rather as it is more commonly called, the massively parallel sequencing (MPS). It has advanced research in many areas of biology, including forensic

science [5], where the focus of forensic genetics is gradually shifting from allele length-based identification to sequence variants, enabling even better power of discrimination. The field is being transformed into forensic genomics, since the sequencing of entire genomes (nuclear and/or mitochondrial) is not unachievable feat in routine laboratory workflow anymore. The true challenge is to assemble all steps of the sequencing protocol into a single workflow, suited for particular study, with sequencing data analysis being a singular challenge on its own [6]. Analysis and reporting for forensic purposes relies on compliance with internationally agreed and prescribed guidelines, wherefore the method needs to be evaluated through internal validation performed by each laboratory [5, 7, 8]. Current mtDNA guidelines [9, 10] have been updated to some extent to accommodate MPS methods, and will certainly undergo further refinements as more and more MPS data are generated. Various studies have already shown repeatability, reproducibility, concordance to STS data, and overall reliability of MPS assays for analysis of whole mtDNA [11-17]. However, their approach to data analysis and interpretation differed, with bioinformatics solutions encompassing commercially available software, free online software, in-house developed and tailored pipelines, along with almost as diverse threshold settings.

In this work, we evaluated Illumina® Human mtDNA Genome assay on MiSeq FGx™ benchtop sequencer, in conjunction with BaseSpace® Sequence Hub applications for mtDNA analysis (namely, mtDNA Variant Processor and mtDNA Variant Analyzer). The assay is based on Nextera® XT library preparation, which consists of target enrichment by long-range PCR (mtDNA amplified in two overlapping amplicons), fragmentation and tagging (performed by Nextera® XT transposome), dual index barcoding, and subsequent library purification and normalization. Libraries are pooled, denatured and diluted prior to loading on instrument, to undergo paired-end sequencing-by-synthesis reactions. From there, it is natural to proceed with data analysis in Illumina's bioinformatics online platform, thus streamlining the workflow and enabling faster data processing. We present here our approach to setting analysis and interpretation thresholds for the whole mtDNA analysis workflow, as well as evaluation of the entire workflow. Internal interpretation guidelines were developed herein, defined by multiple components of the thresholds (encompassing read depth, allele percentages and quality), but the underlying principles of the approach hold potential for wider application in other similar MPS workflows. Our aim was to establish a reliable system suitable for sequencing complete mitochondrial genomes from high-quality samples of the type to be used for population study (i.e. buccal swab samples and blood), which is one of the prerequisites for using mitochondrial sequence information for forensic purposes.

2. Materials and Methods

2.1. Sample collection and plan of experiments

For the purpose of this study, reference samples were collected from 11 volunteers. All participants gave detailed informed consent. From each person, two types of samples were collected: buccal swabs (collected on Whatman™ Sterile Omniswab, GE Healthcare, UK) and blood (collected on Whatman™ FTA™ Classic Cards, GE Healthcare, UK). DNA was extracted from buccal swabs using the EZ1® DNA Investigator® kit on EZ1® Advanced XL instrument (Qiagen, Hilden, Germany), following the manufacturer's instructions [18]. As for dried blood on FTA™ Cards, QIAamp® DNA Micro Kit (Qiagen, Hilden, Germany) was used for DNA extraction, also according to the manufacturer's instructions [19]. All DNA extracts were subsequently quantified on Qubit™ 3.0 Fluorometer using Qubit™ dsDNA High Sensitivity kit (Thermo Fisher Scientific, Waltham, MA, USA). Apart from the collected reference samples, Standard Reference Material® (SRM) 2392 and 2392-I from National Institute of Standards and Technology (NIST, Gaithersburg, MD, USA) [20, 21] were obtained. Of those, SRM® 2392 Component #1 CHR (abbreviated as SRM-C) and SRM® 2392-I HL-60 (abbreviated as SRM-H) were used as positive controls (i.e. probative samples). To monitor the presence of contamination and to assess the level of experimental and instrument noise, negative controls were introduced in each step of the workflow: reagent blanks in DNA extraction (NC-EX), as well as in long-range PCR (NC-PCR) and in library preparation (NC-LIB).

Plan of experiments and samples used are described in Supplementary Table S1. They were designed to encompass the following studies: repeatability (Supplementary Table S1a), reproducibility (Supplementary Table S1b), mixtures study (Supplementary Table S1c), concordance MPS to MPS, as well as concordance MPS to STS (Supplementary Table S1d). Simulated mixed samples were obtained by combining two persons' buccal swab sample DNA extracts in a particular ratio (0.5%, 1.0%, 2.5% and 5.0%; Supplementary Table S1c) prior to enrichment and library prep. In mixtures study, the sensitivity of minor contributor detection was assessed, but also repeatability, since there were three replicates for each ratio of contributors. Contamination study consisted of analysing negative controls (NCs) from all sequencing runs (including, but not limited to these studies only). The general idea was to use NCs to assess the noise level and characteristics, along with assessment of noise and errors in replicates of positive controls SRM-C and SRM-H. From this information, analysis and interpretation thresholds would be calculated, and subsequently applied to other samples included in the evaluation in order to test parameters of repeatability and reproducibility of the assay.

2.2. Target enrichment, library preparation and sequencing

Long-range PCR approach was adopted to obtain whole mitochondrial genomes in two overlapping amplicons. Primer pairs described in [22] were used (MTL-F1, MTL-R1, MTL-F2 and MTL-R2) to produce amplicons of sizes 9.1 kbp and 11.2 kbp, with the overlap covering the entire mtDNA control region. PrimeSTAR® GXL (TaKaRa, Kusatsu, Japan) was used for long-range PCR, with the following thermal cycling conditions: 25 cycles x [98°C 10 sec + 60°C 15 sec + 68°C 9 min 6 sec] for 9.1 kbp fragment, and 25 cycles x [98°C 10 sec + 68°C 10 min] for 11.2 kbp fragment. Input into target enrichment was 1 ng of genomic DNA extract in total reaction volume of 12.5 µL, or 2 ng in reaction volume of 25 µL, otherwise prepared according to manufacturer's instructions [23]. Quality of PCR products was evaluated via agarose gel electrophoresis: 1% agarose gel, with the addition of 1 µL Midori Green Advanced DNA Stain (Nippon Genetics Europe GmbH, Düren, Germany), was run for 45 minutes, 80 V, in SubCell® GT system (Bio-Rad, Hercules, CA, USA). Gels were visualized via GelDoc™ system and Image Lab™ software (Bio-Rad, Hercules, CA, USA), whereupon they were inspected for yield, as well as for expected band size and specificity. In case of any artefacts observed by gel electrophoresis, long-range PCR was repeated for the affected sample. PCR products were quantified with Qubit™ dsDNA High Sensitivity kit, and were then normalized in a two-step manner with ultra-filtered water and resuspension buffer (RSB, Illumina®) down to the final concentration of 0.2 ng/µL. Equal volumes of both mtDNA amplicons were pooled for each sample, resulting in a single tube per sample, now containing entire mtDNA in two fragments. Total amount of 1 ng of each sample was taken further for library preparation, as per protocol [22].

Libraries were prepared using Nextera® XT Library Prep Kit (Illumina, San Diego, CA, USA) according to the manufacturer's instructions [22]. Briefly, DNA was enzymatically fragmented and tagged with adapter oligonucleotides in a single reaction (tagmentation) performed by Nextera® XT transposome. Afterwards, Index 1 (i7) and Index 2 (i5) adapters were added to the tagged DNA in limited-cycle PCR. Indexed libraries underwent bead-based purification with Agencourt AMPure XP magnetic beads (Beckman Coulter, Brea, CA, USA). Afterwards, either bead-based normalization or individual normalization was applied. In the former, libraries were normalized using LNA1/LNB1 magnetic beads solution (components provided in Nextera® XT Library Prep Kit) as described in the protocol [22], while in the latter case libraries were quantified with LabChip® DNA High Sensitivity Assay on LabChip® GX Touch HT (PerkinElmer, Waltham, MA, USA) and then individually normalized to 2-3 nM using RSB. Normalized libraries were pooled in batches of 24-48 samples per run, denatured and diluted as described in Illumina® protocol [24], with a 5% spike-in of PhiX Sequencing Control v3 (Illumina, San Diego, CA, USA). Paired-end sequencing was performed on Illumina® MiSeq FGx™ instrument using MiSeq® Reagent Kit v2, standard flow-cell, 300 cycles (2 x 151 bp).

As part of concordance study, separate set of libraries (48 in total) was prepared using Nextera® XT Library Prep Kit from the same PCR amplicons that were used for repeatability, reproducibility and mixtures study. Libraries were further processed in an independent laboratory by their staff: they were quantified with Agilent High Sensitivity DNA Kit on Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) according to the manufacturer's instructions, and were subsequently normalized and pooled for sequencing on Illumina® NextSeq®500 platform following protocol as described in [25]. NextSeq®500/550 Mid Output Kit v2.5, 150 cycles, was used for paired-end sequencing (2 x 75 bp). Resulting haplotypes from both MPS platforms were compared to each other for concordance, as well as to Sanger-type sequencing (STS) results generated and described previously [26].

2.3. Data analysis

On MiSeq FGx™ instrument, software Real-Time Analysis (RTA) v.1.18.54 and MiSeq® Reporter v.2.5.1.3 (Illumina®) provided primary and secondary analysis of sequencing results, applying the "mtDNA workflow" as specified in sample sheet settings prior to each run. Quality metrics were reviewed in Illumina® Sequencing Analysis Viewer (SAV) v.1.11.1 software. FASTQ files generated by MiSeq® Reporter were extracted and uploaded to Illumina® BaseSpace® Sequence Hub online platform, where they were processed by BaseSpace® mtDNA Variant Processor v1.0.0 App [27]. The application performs adapter trimming, alignment to circular reference genome, realignment of regions with indels, removal of primer contribution from reads, variant calling, read filtering and quality scoring, and generation of output files (e.g. BAM and VCF). Of the few settings that could be user-defined in mtDNA Variant Processor, common settings that were applied to all analyses comprised: minimum basecall quality score for a call = 30, and genome used for alignment = rCRS (revised Cambridge reference sequence) [28, 29]. Values for analysis and interpretation thresholds (AT and IT, respectively) varied: the first stage of analysis encompassed negative and positive controls analysed at AT = 0.1%, IT = 0.1%, and minimum read count = 2 (Fig. 1). This way all signals, both true variants and false positives (noise signals and errors), were detected and taken into consideration for the calculation of thresholds, as well as for noise level assessment and characterization. All signals detected in negative controls were treated as noise originating from reagents (DNA extraction, long-range PCR, library preparation, sequencing) and/or instrument detection. Calculated values were expressed as number of reads (read depth, DP) and included the following: minimum (MIN), maximum (MAX), average (AV), standard deviation (SD), limit of detection (LOD) and limit of quantitation (LOQ) – applying principles similar to assessing thresholds in STR markers' analysis in capillary electrophoresis [30].

Afterwards, samples of positive control samples (SRMs) were analysed in a two-fold manner:

1) data from known variants assigned to controls' haplotypes (according to [31]) were used to calculate parameters of variant quality (known as "GQ" in genome VCF files, or "Q score" in BaseSpace mtDNA Variant Analyzer reports) and percentage of homoplasmic variant (i.e. percentage required of a base in order to classify the position as homoplasmic);
2) signals detected from all other variants not belonging to the defined haplotypes (both identical to, or differing from, rCRS) were perused similarly as in negative controls, to estimate noise level within positive controls, as well as to calculate minimum criteria for reliable variant analysis and interpretation (read depth and percentages of minor alleles), which would eventually constitute analysis and interpretation thresholds.

Overall results were used to estimate our internal analysis thresholds in terms of: minimum read depth for a reliable variant call, percentage of allele for genotype allele (i.e. calling of a homoplasmic variant at particular position), percentage of alternative allele (for point heteroplasmy calls), and genotype quality score (GQ; in Phred scale). Thus, internal analysis thresholds (INT) consisted of several components, which all variants had to comply with in order to produce a valid call.

Second stage of analysis consisted of applying the newly calculated INT to re-analyse samples of negative and positive controls to confirm the validity of thresholds. This was followed by the final stage of analysis, in which INT were applied to analyse all other evaluation samples, wherefrom repeatability, reproducibility and concordance were assessed. At all stages of analysis, samples were visually inspected via BaseSpace® mtDNA Variant Analyzer v1.0.0 App, which allowed review of coverage profiles and sequences, as well as export to Excel-format reports. All sample reports were manually reviewed, and final variant lists (i.e. mitochondrial haplotypes) were produced for sample comparison, in accordance with the current guidelines [9, 10]. When necessary, BAM files were reviewed in Integrative Genomics Viewer (IGV) tool v.2.4.16 [32, 33] to resolve ambiguous calls.

3. Results and Discussion

3.1. Quality metrics assessment

Evaluation of sequencing quality (Q) metrics is an essential step in sequencing data analysis, since it is a good indicator of what to expect regarding the quality of results. High metrics quality usually means better usage of data, therefore more abundant and reliable results. All runs in this study exhibited excellent quality, as shown in the summary of selected Q metrics parameters (Table 1). Despite the variations in cluster density (491 – 1062 K/mm^2), which were sometimes below the optimal range for MiSeq Reagent Kit v2 chemistry according to [34, 35], runs maintained high level of quality regarding

both the percentage of clusters passing filter (PF) and percentage of bases with Q score equal or higher than 30 (% Bases ≥Q30; Phred scale). Clusters PF amounted to >90% in all runs, meaning almost all of the data were always usable for downstream analysis and, judging from % Bases ≥Q30, the great majority of bases were of sufficiently high quality for downstream analysis (variant calling, eventually). Suboptimal cluster density in runs 1, 6 and 7 affected total yield and total number of reads PF, which in turn impacted average read depth per position per sample (Table 1) in the way that validation samples in these runs received lower average coverage than expected from calculated coverage values based on chemistry used and targeted region (whole mtDNA). In connection to the cluster density was also the percentage of reads aligned to PhiX sequencing control (% Aligned). As described earlier, we used 5% PhiX spike-in, therefore we expected % Aligned to approximate 5%. However, as spike-in percentage was in fact volume ratio, while % Aligned represented proportion of reads detected as PhiX reads in the total pool of reads PF, we observed that % Aligned in some runs deviated from the expected percentage (Table 1). Runs with high cluster density exhibited lower % Aligned and vice versa (runs with low cluster density contained more PhiX reads). Therefore, % Aligned parameter is directly dependent on the accuracy of library quantification and subsequent loading concentration: the former may not be as accurate using gel electrophoresis on LabChip, as opposed to qPCR [36]. The quantity of libraries may easily be over- or underestimated, thus influencing both their and PhiX's share in the total reads available (which is ultimately reflected in % Aligned value). As overclustering poses a risk to overall success of a sequencing run, we aimed for loading concentrations safely within the manufacturer's specifications (ranging 8-15 pM) in order to avoid potential loss of quality. Judging by almost all Q metrics parameters, runs 2 and 5 displayed optimal values for our data, although the other runs were only affected in the sense of quantity of results and not the quality, which was still well above the specifications.

Table 1 Summary of selected quality metrics parameters from evaluation runs on MiSeq FGxTM instrument. Values % Bases ≥Q30 and Error rate are given as average for the entire run

Run	Samples per run	Cluster density (K/mm^2)	Clusters PF (%)	Yield (Gigabase)	Reads PF (million)	% Bases ≥Q30	Error rate (%)	% Aligned to PhiX	% Reads Identified [a] [Expected %]	Read depth [a, b]
1	24	491	97.8	3.0	9.4	97.1	0.5	13.1	3.9 ± 1.8 [4.2]	5885 ± 3934
2	24	1062	91.5	5.8	18.5	93.8	0.5	4.8	4.1 ± 1.6 [4.2]	11048 ± 8683
3	48	864	94.9	5.0	15.9	95.9	0.5	4.1	2.1 ± 0.7 [2.1]	5084 ± 2816
4	23	939	94.1	5.3	17.0	93.0	0.5	6.5	4.0 ± 1.3 [4.3]	9065 ± 4979
5	28	948	93.7	5.4	17.1	96.0	0.4	4.8	3.7 ± 0.7 [3.6]	10935 ± 6160
6	30	539	97.2	3.2	10.1	96.2	0.4	11.2	3.0 ± 0.5 [3.3]	4901 ± 2405
7	24	551	96.0	3.3	10.6	95.7	0.5	7.6	4.1 ± 1.7 [4.2]	7770 ± 5618
8	28	745	95.8	4.4	14.0	95.3	0.6	6.1	3.4 ± 1.2 [3.6]	8203 ± 5307

[a] Expressed as average ± standard deviation; [b] Only calculated for samples included in validation experiments

Depending on the number of samples multiplexed per sequencing run, there is an expected proportion of reads identified for each library (e.g. if there are 24 samples in a run, expected percentage of reads identified is 100/24 = 4.2% of reads assigned to each library, under condition of ideally even distribution). The values designate proportions of unique index combinations detected in total amount of reads, and their distribution within runs gives valuable information on the efficiency of the particular lab's workflow. In our runs, percentage of reads identified for validation samples closely approximated the expected values (Table 1). Greater standard deviation was usually observed in runs where bead-based normalization was applied (runs 1, 2 and 4), as opposed to standard normalization applied in the remaining runs. It has been noted previously that bead-based normalization introduces greater variation between libraries [16].

A drop in quality was generally observed in the second read of paired-end sequencing when compared to Read 1, manifesting in parameters of % Bases ≥Q30, phasing, prephasing and error rate (Supplementary Table S2). It is not an uncommon observation, particularly since it is known that in paired-end sequencing the quality drops both in the second read, as well as towards the end of both reads [15, 37-39]. Nevertheless, this did not affect the overall quality of sequencing runs, which was unquestionable.

Regarding the coverage of mtDNA, there was a reproducible pattern across all samples: reads were unevenly distributed along the entire mitochondrial genome, with extreme drops in coverage at certain positions (Supplementary Fig. S1), regardless of sample origin (type, person, etc.). This phenomenon has been reported on numerous occasions [2, 14-16, 40], all including Nextera XT library preparation. Some read-depleted regions correspond to low-complexity (homopolymer) stretches that are known as problematic for both sequencing and alignment (e.g. positions 300-600 which harbour hypervariable segments II and III). However, the cause of coverage drops in other regions (e.g. positions 3400-3700, 5400-5600, 10900-11000, 13000-13100, 13600-13800) is still unknown. Some proposed that non-uniform coverage was a by-product of alignment issues because of the circular reference genome (which was shown not to be the case, after all) [14, 15], and others that it was the result of the combination of library preparation and challenging alignment [2, 16]. Still others hypothesized that such coverage pattern resulted from Nextera XT transposome bias [16], i.e. the enzyme probably preferring certain regions of mtDNA, rather than acting randomly. We are inclined towards the latter explanation, since we observed almost identical coverage profiles in our libraries sequenced on NextSeq (data not shown) as part of concordance study, and also because it was shown that other library preparation chemistry (for example, [40]) produced different, more uniform coverage pattern. Depending on the purpose, some studies will certainly require different library preparation approach to achieve the necessary coverage uniformity – for example, uneven coverage

may be acceptable for population studies (which aim for genotype variants), but less so for minor allele detection (where sufficient read depth is of paramount importance, and non-uniformity risks the loss of true variant signal).

3.2. Contamination study and noise level assessment

Library preparation protocols consist of multiple handling steps, which increase susceptibility to introduction of exogenous contaminant DNA, facilitate cross-contamination between samples, and (by means of bead-based purification and normalization) may inflate the amount of eventual contamination – because of this, some proportion of reads is commonly found (even expected, it might be said) in NCs [2, 16, 17]. Therefore, it is recommended that NCs be introduced in various stages during library preparation, to monitor the level of background noise and the presence of contamination, so that both can be appropriately characterized and the level of tolerance established – level below which detected noise/contamination has no effect on results and can be classified as acceptable [2, 9, 10, 16]. To thoroughly assess the level of noise and its contents, as well as to estimate safe thresholds for reliable data analysis and interpretation, we analysed the total of 35 negative controls (NCs), sequenced as part of both assessment and other studies carried out on our MiSeq FGx instrument. Of these, 25 negative controls were reagent blanks introduced in the step of DNA extraction (NC-EX), six were amplification negative controls from long-range PCR (NC-PCR), and the remaining four negative controls were reagent blanks introduced in the step of limited-cycle PCR (NC-LIB).

In sequencing pools, NCs were represented with 0.0004-0.0096% of the total number of reads PF. Detailed analysis of genome VCF files (GVCF) exported from BaseSpace mtDNA Variant Processor (workflow I in Fig. 1) produced the following results. Signals were detected in total of 206,856 positions in all 35 NCs, averaging to 5,910 positions per NC covered with both forward and reverse reads. However, vast majority of these positions (142,395 in total) were detected in NC-EX, out of which 91% (i.e. 129,393 positions) had read depth of ≤10 reads, while only 47 positions exhibited elevated read count of >200 reads. NC-PCR and NC-LIB consisted of similar amount of positions with signals detected (33,699 and 30,762 respectively).

Analyses and calculations were performed both cumulatively for all NCs, for each NC-type separately, and also for each base (A, C, G, and T) to investigate potential influence of NC-type or particular dye channel (base detection) on the level and/or nature of noise signals. As shown in Table 2a, maximum depth (DP) for any NC-EX equalled 1221 reads, which is extremely high, while maximum DP for NC-PCR and NC-LIB was 57 and 21 reads, respectively. By reviewing positions with extreme read DP, we identified two regions of interest (Fig. 2): 1873-1893 (coding region, 16S rRNA) and 16128-16455

(control region, HVS-I). Region 1873-1893 showed conspicuous read depth in seven NC-EXs (>1000 reads in one, 100-1000 reads in one, 10-100 reads in five), and in one NC-PCR (40-60 reads). Start and end coordinates of this region correlated to MTL-R1 primer, used in long-range PCR for amplification of mtDNA fragment 9.1 kbp. By visualizing BAM files in IGV tool, we confirmed that indeed increased read depth originated from the primer (Supplementary Fig. S2, upper and middle panels). The purification of libraries may not have always been equally efficient, depending on the analyst and on handling the magnetic beads, thus a certain amount of primer might have persisted through to the sequencing. However, since we detected no signal from any of the other three primers in negative controls, it is possible this feature is specific to MTL-R1 alone. The discovery of primer signal was quite surprising, considering that all primer read contributions should have been removed by mtDNA Variant Processor [27]. For comparison, no primer reads were present in BAM files extracted from MiSeq Reporter software (Supplementary Fig. S2, lower panel), which indicates that BaseSpace application's pipeline may have issues with recognizing and removing this particular primer. Because of this phenomenon, variants detected in mtDNA positions 1873-1893 must be interpreted with caution, particularly in case of heteroplasmy calls, since the minor allele signal might in fact originate from primer reads, instead of true positive variant call from the sample. Most of the times such ambiguities can be successfully resolved by visual inspection in genome browsers such as IGV. The second detected region (16128-16455), unlike previous, was not connected to any of the primers used in long-range PCR, but it was found in eight NC-EX (>100 reads in one, 20-100 reads in others). The presence of these two regions of increased coverage was more or less random in NC-EX and NC-PCR (independent of normalization method, analyst, number of libraries per run, etc.), and while it is an interesting observation, it also warrants caution when interpreting variant calls occurring there.

Table 2 Calculations and estimations of read depth analysis threshold, based on negative controls data from: GVCF files (a), GVCF files excluding primer MTL-R1 reads (b), and BaseSpace® mtDNA Variant Analyzer report files (c). Negative controls originated from various stages of the workflow: DNA extraction (NC-EX), mtDNA enrichment (NC-PCR) and library preparation (NC-LIB). Values were calculated for each base separately and then cumulatively, by each NC-type and also jointly for all NCs. Limit of detection (LOD) value was calculated as: average + 3x standard deviation. Limit of quantitation (LOQ) value was calculated as: average + 10x standard deviation

a

	NC-EX					NC-PCR					NC-LIB					All NCs				
	A	**C**	**G**	**T**	**Cum.**	**A**	**C**	**G**	**T**	**Cum.**	**A**	**C**	**G**	**T**	**Cum.**	**A**	**C**	**G**	**T**	**Cum.**
MIN	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
MAX	1221	1197	1205	1196	1221	56	57	56	51	57	20	21	18	20	21	1221	1197	1205	1196	1221
AVERAGE	6	6	6	6	6	5	5	5	5	5	4	4	4	4	4	6	6	6	5	6
ST.DEV.	17	13	21	13	15	3	3	3	3	3	2	2	2	2	2	14	11	17	11	13
LOD	57	45	69	45	51	14	14	14	14	14	10	10	10	10	10	48	39	57	38	45
LOQ	176	136	216	136	156	35	35	35	35	35	24	24	24	24	24	146	116	176	115	136

Read depth threshold estimation = 220 reads

b

	NC-EX					NC-PCR					NC-LIB					All NCs				
	A	**C**	**G**	**T**	**Cum.**	**A**	**C**	**G**	**T**	**Cum.**	**A**	**C**	**G**	**T**	**Cum.**	**A**	**C**	**G**	**T**	**Cum.**
MIN	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
MAX	213	212	205	216	216	18	19	18	19	19	20	21	18	20	21	213	212	205	216	216
AVERAGE	6	6	6	6	6	5	5	5	5	5	4	4	4	4	4	5	5	5	5	5

ST.DEV.	9	9	7	7	8	3	3	3	3	3	2	2	2	2	2	7	7	6	6	7
LOD	33	33	27	27	30	14	14	14	14	14	10	10	10	10	10	26	26	23	23	26
LOQ	96	96	76	76	86	35	35	35	35	35	24	24	24	24	24	75	75	65	65	75

Read depth threshold estimation = 100 reads

c

	NC-EX					**NC-PCR**					**NC-LIB**					**All NCs**				
	A	**C**	**G**	**T**	**Cum.**	**A**	**C**	**G**	**T**	**Cum.**	**A**	**C**	**G**	**T**	**Cum.**	**A**	**C**	**G**	**T**	**Cum.**
MIN	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
MAX	172	155	55	163	182	10	12	11	11	15	11	7	7	7	13	172	155	55	163	182
AVERAGE	8	10	5	8	12	4	5	5	5	6	4	4	4	4	5	6	9	5	7	10
ST.DEV.	19	23	6	19	23	2	3	3	3	3	2	2	2	2	3	16	21	5	16	20
LOD	65	79	23	65	81	10	14	14	14	15	10	10	10	10	14	54	72	20	55	70
LOQ	198	240	65	198	242	24	35	35	35	36	24	24	24	24	35	166	219	55	167	210

Read depth threshold estimation = 240 reads

393

Comprehensive calculations based on all signals detected in negative controls according to workflow I (Fig. 1) were made: including primer MTL-R1 reads (Table 2a), and with primer MTL-R1 reads removed (Table 2b). Results are shown by NC-type, by base for each NC, as well as cumulative values. Following the more conservative approach, estimation of our internal analytical threshold of read depth (INT-DP) was based on the highest LOQ value. In the case when primer reads were excluded (Table 2b), estimated INT-DP equalled 100 reads. However, since primer reads could not be ignored in the analysis pipeline used, we decided to keep calculations from Table 2a, and estimated INT-DP accordingly: highest LOQ value was found in NC-EX for base G (216 reads), and by estimating the threshold at 220 reads, all signals in negative controls would have been eliminated except for the primer reads (Fig. 2). This actually corresponded well to calculations in Table 2b, because maximum read depth equalled 216 reads in any negative control after primer contribution was removed. Thus, INT-DP threshold of 220 reads was applicable both scenarios.

To test the validity of estimated INT-DP threshold, calculations analogous to those described above were performed on Excel reports data exported from BaseSpace mtDNA Variant Analyzer (workflow II in Fig. 1). Reports produce lists of variants, i.e. differences from rCRS. Therefore, a large quantity of signals that are visible in GVCF files are actually not present in Excel reports, which includes primer reads (as their sequence is identical to rCRS). Nevertheless, reports may be more relevant for consideration, since a negative control (despite some portion of reads regularly expected) should not produce any variants, and no variant calls must be present in reports for NCs when validated analysis threshold is applied. Calculations resulting from negative controls' BaseSpace reports data (Table 2c) differ from the results in Table 2a and Table 2b, particularly regarding NC-EX, where greater variation among bases is evident: larger standard deviation led to higher LOD and LOQ values (highest LOQ=242 reads for cumulative NC-EX, and LOQ=240 reads for C in NC-EX), even though maximum read depth detected in any NC cumulatively equalled "only" 182 reads. Estimation of read depth threshold at 240 reads, while not considerably higher than 220 reads, would nevertheless be over-conservative, since by applying the latter threshold all signals from negative controls' reports could easily be eliminated, thus establishing the tolerable level of noise below which NCs would be regarded as truly negative.

Considering the content of noise signals, i.e. whether any of the bases (A, C, G, or T) occurred more often than the other, occurrence of each base was counted from GVCF files (Supplementary Table S3). Bases A and C were most commonly detected and in almost equal measures, followed by T, while G was the least commonly detected signal in NCs. This trend was evident in all negative controls, regardless of the type. However, since the ratio of each base count to the total number of detected bases (both by NC-type and cumulative) closely approximated its corresponding ratio in mtDNA (specifically, in rCRS: A=31%, C = 31%, G=13%, T=25%; Supplementary Table S3), we concluded that

the distribution of noise was random across the entire mtDNA, not preferring any particular base to the other.

To finalize negative controls assessment, we decided to maintain the estimated read depth threshold (INT-DP) at 220 reads, which was applicable both in GVCF files analysis and BaseSpace mtDNA Variant Analyzer reports analysis. The meaning of this thresholds was to set a limit for safe interpretation in terms of read depth: above the set threshold reliable variant calls can be made, and below is the area of background noise, contamination and possible erroneous calls. Of course, detection of false positive signals is always a possibility, but the aim is to minimize that risk with carefully set thresholds, while at the same time balancing against the loss of true positive signals in the process.

3.3. Positive controls assessment

For further threshold calculations, samples of positive controls – SRM-C and SRM-H, with known and previously well characterized sequences [20, 21, 31] – were analysed according to workflow III in Fig. 1. Indels in hypervariable regions (HVS I-III) were ignored at this time, along with point heteroplasmy variants that were reported in [31], since they cannot be considered as either errors or true variant calls until validated thresholds are applied. Therefore, indels and PHPs were excluded from calculations, as well as position 16183 in sample SRM-C. The latter was detected as ambiguous variant call: a mixture of two bases (A and C) and deletion. It is in fact homoplasmic variant A16183C which, in conjunction with T16189C (also present in SRM-C), produces uninterrupted homopolymer stretch of 11 cytosines, resulting in alignment issues, which were reported and elaborated in [31]. As mentioned earlier, ambiguities such as this may be resolved in most cases by visual inspection of read alignment in tools such as IGV.

In eight samples (four replicates of each SRM), signals were detected in 3,280 positions, in total. Single bases (100% variant from rCRS) were called for 194 positions, while in all other positions between one and three alternative alleles were detected (bases and/or deletions) in addition to the major base. Calculations were performed cumulatively for all SRMs, and also separately for each base (Table 3).

Table 3 Thresholds based on data from BaseSpace® mtDNA Variant Analyzer report files of positive controls SRM-C and SRM-H (SRM® 2392 CHR and SRM® 2392-I HL-60, respectively). Thresholds for read depth and percentage of alternative alleles were calculated and estimated from alternative (non-haplotype) alleles detected in positive controls (a). Thresholds for genotype (homoplasmic) alleles and quality scores (GQ; Phred scale) were calculated and estimated from known haplotype variants in positive controls (b). Limit of detection (LOD) value was calculated as: average + 3x standard deviation. Limit of quantitation (LOQ) value was calculated as: average + 10x standard deviation

a

	Read depth					Alternative allele				
	A	**C**	**G**	**T**	**Cum.**	**A**	**C**	**G**	**T**	**Cum.**
MIN	2	2	2	2	2	0.1%	0.1%	0.1%	0.1%	0.1%
MAX	46	449	130	155	449	1.3%	5.5%	0.6%	2.4%	5.5%
AVERAGE	3	7	4	3	5	0.1%	0.2%	0.1%	0.1%	0.2%
ST.DEV.	4	20	8	7	12	0.1%	0.5%	0.1%	0.1%	0.3%
LOD	15	67	28	24	41	0.4%	1.7%	0.3%	0.4%	0.9%
LOQ	43	207	84	73	125	0.9%	5.1%	0.6%	1.1%	2.8%
	Estimated read depth threshold = 210					**Estimated alternative allele threshold = 3%**				

b

	Genotype allele					GQ score				
	A	**C**	**G**	**T**	**Cum.**	**A**	**C**	**G**	**T**	**Cum.**
MIN	94.5%	94.0%	99.4%	96.5%	94.0%	25	31	27	27	25
MAX	100.0%	100.0%	100.0%	100.0%	100.0%	50	50	49	50	50
AVERAGE	99.8%	99.9%	99.9%	99.9%	99.9%	46	48	46	47	47

ST.DEV.	0.4%	0.2%	0.1%	0.2%	0.3%	5	3	4	4	4
LOD[a]	98.6%	99.2%	99.7%	99.1%	99.0%	41[b]	45[b]	42[b]	43[b]	43[b]
LOQ[a]	95.8%	97.6%	99.2%	97.5%	97.2%					
	Estimated genotype allele threshold = 97%					**Estimated GQ score threshold = 41**				

[a] Since max. value is 100%, LOD and LOQ calculated as [average - 3x standard deviation] and [average - 10x standard deviation], respectively; [b] Standard LOD and LOQ formulas not applicable (since GQ scores are in Phred scale), therefore calculated as: average - 1x standard deviation

Regarding read depth calculations for alternative alleles (Table 3a), results were concordant with those obtained for negative controls, wherefrom estimated coverage threshold (INT-DP) of 220 reads would be applicable to SRMs as well. Although cumulative LOQ was considerably lower (125 reads), we decided to keep the minimum read count at 220 reads, since the highest LOQ was calculated for C (207 reads; Table 3a), which is just short of the estimated negative controls' threshold (Table 2). As visible in Fig. 3a, there are two positions where maximum read depth of alternative allele exceeded the threshold: in particular, variants detected were A2487M and T16189d. However, these two would not be taken into consideration for true variant calls: the former exhibited extremely poor GQ value (26-29, Phred score) in both control samples, and the latter consisted of ambiguous calls (C and deletion, or C and T and deletion) only in SRM-C, mirroring the same problem described above for the A16183C – in this case, variant T16189C contributed to prolongation of homopolymeric C-stretch and subsequent issues in alignment.

The other parameter calculated from alternative allele signals was percentage of minor alleles, with the maximum of 5.5% (Table 3a, Fig. 3b) detected at A2487M – the same position that showed elevated read depth earlier. Estimating from the cumulative calculated LOQ, analytical threshold for minor alleles (INT-AN) would be 3%. By applying this threshold, 99% of signals would be successfully eliminated since in the total of 3,155 alternative alleles detected in all SRMs, only 39 were >1%. However, as evident from Table 3a and Fig. 3b, alternative alleles with considerably higher minor allele percentages may occur, and that prompted us to establish additional, interpretation threshold for minor alleles (INT-IT) which equalled 6%. The meaning of this dual threshold system is as follows: PHP calls with alternative (minor) alleles >6% are safe for interpretation, under condition of sufficient read depth; PHPs with minor alleles between 3% and 6% are required to undergo additional scrutiny of other quality parameters before they are reported; minor alleles <3% are in the area where it is virtually impossible to distinguish between noise signals and true positive calls (without alternative confirmation method), therefore they cannot be reported as such.

Regarding variants reported as haplotypes (i.e. genotype alleles, GT), calculations were performed analogously to the ones described for alternative alleles above (Table 3b). As a result, threshold for homoplasmic genotype alleles (INT-GT) was estimated at 97% according to cumulative calculations. Notably, minimum values detected for bases A and C were <97% (94.5% and 94.0%, respectively; Table 3b), but by additional review, we found that minimum signal for A originated from A2487M, a low-quality variant call, while the minimum for C was in fact caused by the sum of two minor alleles at the same position (namely, 2.4% T and 3.6% deletion). Overall, we decided to keep the estimated genotype variant threshold (INT-GT) at 97%, meaning that at any position a variant allele exceeding 97% would be considered homoplasmic, i.e. single base variant call – no PHP call would be allowed for

this position. This is in accordance with previously calculated minor allele analysis threshold (INT-AN) of 3%.

In addition to the threshold of percentages for genotype alleles, we performed calculations for quality values (GQ) of genotype positions (Table 3b). Since the use of standard LOD and LOQ formulas (i.e. 3x and 10x standard deviations from average, respectively) was not feasible in this case, we opted for a modified formula more appropriate for the GQ values: *average – 1 x standard deviation*. Cumulative GQ threshold (INT-GQ) equalled 43 (Table 3b), however, we decided to keep the threshold at 41 to accommodate for values of all bases (and calculations for base A produced the value of 41). Intriguingly, position 2706 exhibited GQ lower than other genotype positions in SRMs (GQ 37-41), but also in all other analysed samples (GQ values ranging from 33 to 49, of which more than 80% were <41). Because of this, and similar exceptions to the other threshold components, we must bear in mind that, for a reliable variant call, thresholds defined for all parameters must be met and considered as a whole, rather than as individual, independent requirements.

3.4. Finalized definition of analysis thresholds

Based on the calculations described in previous sections, we finalized the values proposed as our internally evaluated thresholds (INT) for whole mtDNA analysis, encompassing multiple parameters:

- INT-DP = 220 reads
- INT-GT = 97%
- INT-GQ = 41
- INT-AN = 3%
- INT-IT = 6%

Accordingly, we defined our internal guidelines for whole mtDNA analysis and interpretation as follows:

- Minimum depth of 220 reads is required for variant allele to be taken for analysis.
- Quality score (GQ) ≥41 is required for a position to be reliable for variant calling. Otherwise, the position is most likely to contain erroneous variant calls.
- All positions with major allele ≥97% are considered homoplasmic and single base variant is called.
- Alternative alleles <3% are not analysed nor interpreted, since they reside within the area of background noise.
- Alternative alleles between 3% and 6% are taken into analysis. They may be interpreted and subsequently reported, if read depth and quality score thresholds are complied with.

- Alternative alleles ≥6% are considered safe to interpret and report, since presumably all other thresholds' criteria have already been fulfilled.

At first glance, the read depth threshold of minimum 220 reads may seem overly conservative, but its greatest advantage is that it was derived from our own experimental data, rather than set arbitrarily or taken at set value from other studies (e.g. [15-17, 40]). Detection of minor allele present at 3% would hereby require depth of 7,333 reads, while detection of minor allele at 6% would require depth of 3,667 reads. Despite large read counts, these requirements are easily met, since multiplexing of 24 samples per run gives theoretical coverage of 9,375 reads per position per sample. Even multiplexing as many as 48 samples per sequencing run gives theoretical coverage of 4,688 reads per position per sample, which is ample enough for detection of minor alleles with frequencies of 4.7% and higher. The only obstacle to detection of minor alleles is uneven coverage across the mitochondrial genome, which displays some chemistry- and sequence-dependent profile, as described earlier. Therefore, detection and interpretation of minor allele signals in presumably heteroplasmic positions should be mindful of shortcomings specific to the method used.

In addition to our internal guidelines elaborated above, interpretation and calling of indels should not be based solely on percentages obtained from BaseSpace mtDNA Variant Analyzer reports. Read alignments for any indel call are to be manually inspected by visualization in genome browsers such as IGV, prior to determining the dominant molecule [9], which would be reported as the final variant call.

Here we presented our approach to the calculation of analysis thresholds, which used multiple-parameter system to define internal guidelines for analysis and interpretation of whole mtDNA MPS results (something similar has been done in [41] for interpretation of negative controls). As the studies were performed in a forensic laboratory, the aim was to maintain similarity to the method traditionally used to derive thresholds in forensic STR markers' analysis via capillary electrophoresis. As prescribed by [10], each laboratory should develop and implement their individual interpretation guidelines based on validation and evaluation studies, which is what we aimed to do here for our own data. This approach is applicable for other laboratories performing similar studies, but it is possible that the actual threshold values would slightly vary, since each laboratory presents a unique system with its staff, equipment, consumables and environment.

3.5. Repeatability

Definition of repeatability in general terms, according to [7, 8], is the variation in measurements of results obtained by the same person (analyst) multiple times on the same instrument. This can be applied two-fold to the sequencing library preparation workflow, since replicates of a sample may

consist of PCR replicates (same sample amplified in multiple PCR reactions and from each a separate library prepared) and library replicates (i.e. technical replicates, meaning multiple libraries prepared from the same PCR reaction of a sample). Having that in mind, we tested repeatability by comparing final variant calls (final haplotypes) of PCR replicates and library replicates for the samples of buccal swabs and blood on FTA™ Cards ("B" and "F", respectively) of persons MW-0002 and MW-0020 (schedule in Supplementary Table S1a). Final haplotypes from library replicates of positive controls SRM-C and SRM-H were compared for repeatability as well. In all instances, indel and heteroplasmy calls underwent additional review and visual confirmation of read alignment in IGV. Repeatability was assessed for two analysts separately, to evaluate the variation of library preparation between different persons handling the protocol.

Library replicates of sample MW-0002-B showed 100% repeatability, regarding final variant calls, for both Analyst 1 and Analyst 2. PCR replicates of MW-0002-B showed complete repeatability as well, regardless of analyst. Both library and PCR replicates of sample MW-0002-F exhibited 100% repeatability, including point heteroplasmy T16311Y, which was consistently called across all replicates (Supplementary Table S4). In most replicates of sample MW-0020-B, there were two PHPs consistently detected: T152Y and T9325Y (Supplementary Table S4). The few exceptions occurred in instances where read depth of the minor allele did not exceed the required threshold of 220 reads, and thus required manual review below the validated thresholds. In these cases (8 in total; Supplementary Table S4a and S4b), were it not for multiple replicates for comparison, these calls would pass as homoplasmic variants. However, for the purpose of this study, presence of minor allele was considered confirmed, even for those with fewer reads than necessary. For the sample MW-0020-F, only library replicates were made, and they exhibited complete repeatability. One PHP was detected, T9325Y, which was consistently called in all replicates (Supplementary Table S4b).

Regarding technical replicates of positive controls, SRM-C exhibited 100% repeatability, including one PHP position (C64Y), which was consistently detected in all three replicates, and is concordant with [31]. Haplotypes of SRM-H replicates were repeatable as well, altogether with three heteroplasmy calls: T2445Y, C5149Y and T12071Y. Percentages of minor alleles detected were in accordance with [31] for all three PHPs. However, only T12071Y was completely repeatable (most likely due to larger proportion of minor allele), whereas for both other PHPs one or more deficiencies were observed. Read depth requirement was not met in one of three replicates for both T2445Y and C5149Y, and manual review was necessary to confirm the presence of minor allele. Besides read depth, heteroplasmy T2445Y proved more complex to interpret after application of our validated thresholds, since in all replicates GQ fell below 41 (Supplementary Table S4a) for this position. Upon inspection of this variant's environment, we determined that it resides within a region where drop in GQ is

prominent, in all replicates, and encompasses positions 2412-2487. Thus, we recommend that any variants be interpreted with caution, as this region is obviously prone to quality issues in general (the same phenomenon was observed across all samples and sample types). Regarding T2445Y in SRM-H, since it was detected in all replicates and was described previously [31] – even though the question of quality was not discussed there – this heteroplasmy was reported and included in repeatability assessment in this study. Were it not for multiple replicates and literature confirmation, the T2445Y variant would likely be omitted from final haplotype due to not meeting all threshold criteria.

Overall, 783 variant calls (differences from rCRS) were reviewed in the course of repeatability test, across 43 replicates in total. For Analyst 1, 564 variant calls were assessed in total, out of which six calls were discrepant (1.1%). Similarly, in case of Analyst 2, out of 219 variant calls that were assessed in total, two of them showed discrepancy (0.9%). Thus, repeatability equalled 98.9% and 99.1% for Analysts 1 and 2, respectively. Since discrepant calls exclusively concerned point heteroplasmies, whereby manual review confirmed the presence of minor alleles, the whole assay was appraised as completely repeatable.

3.6. Reproducibility

Reproducibility study encompassed comparison of haplotypes for two sample types of 11 persons, along with positive controls SRM-C and SRM-H. Analyst 1 and Analyst 2 independently prepared batches of libraries, which were sequenced in separate runs. As previously described for repeatability study, final variant calls (haplotypes) of samples were compared, while indels and heteroplasmy calls required additional confirmation in IGV tool to be considered for comparison.

Out of 26 pairs of haplotypes that were compared in total, six exhibited some form of discordance and were manually reviewed to determine the cause. In all cases, the main reason for observed discrepancies were inconsistently called PHPs in one sample of the pair (Supplementary Table S5). Samples MW-0078-B, MW-0020-B2, MW-0065-F, MW-0067-F and SRM-H all exhibited heteroplasmies detected in the results of Analyst 2, while apparently no corresponding heteroplasmy call was found in results of Analyst 1. The presence of minor alleles, as described in previous section, was established by manual review below the validated thresholds (220 reads), and in all instances heteroplasmy calls were confirmed. For the purpose of this study, such results were considered reproducible.

While the same effect was observed in sample MW-0087-B (variant T8955Y was detected only in one of the pair, and seemingly no minor allele signal, i.e. 0%, was detected in the other), the cause was different. To resolve this, we lowered the analysis threshold below 3%, and found minor allele C at 2.9%, despite excellent read depth (396 reads; Supplementary Table S5). Thus, heteroplasmy call was

considered confirmed for the purpose of reproducibility, even though normally it would not be detected as PHP since it does not comply with all components of our validated thresholds.

Additionally, to serve as our own internal control sample, MW-0020-B was sequenced in all our runs, 18 times in total (not limited to evaluation runs only). These results were included as part of reproducibility study, since they encompassed five different analysts who prepared libraries, and multiple runs. Haplotypes were fully reproducible, regardless of analyst and run, including two PHP calls, T152Y and T9325Y (Supplementary Table S6). Percentages of minor alleles were consistent with results from Supplementary Table S4 and Supplementary Table S5. Along with quality (GQ) and read depth (DP) parameters, they confirm the validity of our "dual" threshold system for analysis and interpretation, since all PHPs between 3-6% of minor allele conform to other INT components (GQ and read DP; Supplementary Table S6), and are therefore safe to interpret and report (after analyst review) according to our validated thresholds.

Overall, the assay produced reproducible results between analysts and different runs. The exceptions were few cases of inconsistent heteroplasmy calls: of 724 pairs of variants compared for reproducibility in total, seven pairs required manual analyst review as one of the pair did not meet a component of thresholds' criteria. Nonetheless, heteroplasmy calls were eventually confirmed, and thus considered reproducible as well in this study.

3.7. Concordance

Concordance study consisted of two parts: firstly, MPS-generated mtDNA haplotypes were compared to STS results (published previously as part of Croatian population study [26]); and secondly, MiSeq-generated results were compared to NextSeq-generated results, obtained by the same library preparation reagents, but sequenced in an independent laboratory on a different instrument.

3.7.1. MPS to STS

We compared haplotypes of 10 persons' buccal swabs used in this study to their corresponding haplotypes generated by STS. The latter encompassed only mtDNA control region, while in this study we sequenced whole mtDNA. In general, results were concordant (Supplementary Table S7), with few exceptions concerning PHP calls, as well as insertions. For example, insertions at position 573 were regularly detected in ranges of 3-10% (as reported in Excel reports from BaseSpace mtDNA Variant Analyzer application), which is far below the 50% required to call the dominant molecule. However, these percentages may not reflect the actual state: they may have been artificially produced (or, rather, reduced) by alignment artefacts. Therefore, by viewing read alignments via IGV tool, we were able to resolve apparent discrepancies between STS and MPS: insertions 573.1C-573.3C were

confirmed in MW-0012, insertions 573.1C-573.4C confirmed in MW-0026 and MW-0067, insertion 16193.1C confirmed in MW-0065, and insertions 16193.1C-16193.2C confirmed in MW-0078. The presence of insertions was sufficient to appraise results as concordant, since length variation cannot be counted as exclusion [9, 10], or discordance in this case.

Apart from indel calls, which were manually reviewed and confirmed, point heteroplasmies were the main source of discrepancies, as expected, since MPS readily detects minor alleles below 10%, which is the nominal sensitivity of detection for STS method. Thus, samples MW-0020, MW-0067, MW-0087 and MW-0088 exhibited PHPs that were not seen previously in STS results: T152Y, C16301Y, A374R and C16256Y, respectively (Supplementary Table S7). These observations were not unexpected, since in all four PHPs minor allele proportions were <10% (Supplementary Table S5), and thus passed undetected by STS. Furthermore, samples MW-0026, MW-0065 and MW-0078 exhibited homoplasmic variants in STS results (T16093C, T16093C and A200G), whereas MPS revealed these positions as actually heteroplasmic (Supplementary Table S5 and Supplementary Table S7). Minor allele T might have been detected by STS in sample MW-0026, since proportions from STS results exceeded 11%, however, the observation was probably not sufficiently confident for the PHP call.

In general, MPS-generated results were concordant with STS-generated results, with few exceptions like indels and PHP calls, the first due to MPS method limitations (bioinformatic solutions still struggle with homopolymeric nucleotide stretches and other low complexity regions, thus creating artificial image of indels), and the latter due to STS method limitations (sensitivity of minor allele detection). Besides comparison of control region haplotypes, MPS of whole mtDNA evidently generates much more information and greatly complements STS data. It is particularly elucidating to see the number of variants arising in the coding region, as well as the appearance of more heteroplasmic positions. This gain of discriminatory information is particularly relevant for forensic purposes.

3.7.2. MPS to MPS (MiSeq to NextSeq)

To validate our whole mtDNA MPS results, 36 pairs of haplotypes were compared for concordance assessment between two MPS platforms: MiSeq FGx in our laboratory and NextSeq in an independent laboratory (Supplementary Table S1d). MiSeq data were analysed at the established INT thresholds, with indels and heteroplasmy calls subsequently reviewed via IGV tool as described previously. The exact same analysis thresholds, however, could not be applied to data from NextSeq instrument – different instrument, different operators, and different laboratory environment – at least not without conducting a separate evaluation to establish thresholds specific to that instrument's conditions, which was beyond the scope of this study. Therefore, all variants detected on MiSeq and reported in

final haplotypes of samples only sought confirmation in the NextSeq data, and not complete compliance with the calculated INT thresholds.

The majority of samples showed absolute concordance between results from the two sequencing platforms. Some minor discrepancies were noted, arising from heteroplasmy calls (Supplementary Table S8). For samples MW-0020-B and SRM-H, which had two and three PHPs detected, respectively, one of the three library replicates of each sample exhibited low coverage of minor alleles in MiSeq results (read depth <220 reads; Supplementary Table S8). Normally, if that one replicate were uniquely sequenced sample either for MW-0020-B or SRM-H, MiSeq calls would not have been defined as heteroplasmies, but as single variants. However, since these particular variants were detected in all other replicates of MW-0020-B and SRM-H, multiple times during repeatability and reproducibility studies (Supplementary Tables S4-S6), here they were acknowledged as PHPs as well. The presence of minor alleles for all PHPs in those two samples were unambiguously confirmed in NextSeq results, which offered much better coverage, and subsequently easier interpretation.

Further, in all three replicates of sample MW-0020-F, variant T9325Y was underrepresented in the NextSeq data, regarding both minor allele percentage and read depth (<3% and <220 reads, respectively). It is worth noting that these replicates received less than average share of reads: 0.07-0.57% reads identified, while approximately 1% would be expected since 96 samples were multiplexed for the NextSeq run. Consequently, read depth was lower in these samples, and some variants were very poorly covered (e.g. only 22 reads for minor allele C in replicate MW-0020-F2). Regardless of that, the presence of minor allele was established in all replicates and was sufficient for the confirmation of concordance. By the same analogy, heteroplasmy C16301Y in sample MW-0067-F showed minor allele at 2.9% in the NextSeq dataset, and though it may be below the established thresholds on MiSeq, it was not considered as a discordance since the confirmation was all that we needed from NextSeq.

In contrast to reproducibility study (Supplementary Table S5), sample MW-0080-B showed additional heteroplasmy call (T16093Y). Probably it passed undetected earlier because of poor read depth and/or minor allele <3%. However, it was now detected on MiSeq, and also confirmed in its corresponding pair mate in NextSeq results (Supplementary Table S8). Adversely, samples MW-0087-B and MW-0065-F experienced a loss of heteroplasmy call (T8955Y and T16093Y, respectively), in comparison to reproducibility study results (Supplementary Table S5), as their respective minor alleles probably lacked either read depth or percentage to be detected. These observations were not surprising for neither of these samples, since all three heteroplasmies exhibited minor allele proportions on the borderline of the defined analysis thresholds for MiSeq data (very close to 3%), and thus may or may not be detected, which strongly depends on sequencing run metrics in each particular case.

Overall, comparison of sequencing results comprised the total of 955 pairs of variants (differences from rCRS) between two MPS platforms. In several instances, manual review was required before confirmation of results, but they were all successfully resolved. Both datasets unequivocally showed complete concordance, as expected, since both instruments originate from the same manufacturer, and are based on the same sequencing-by-synthesis technology.

3.8. Mixtures study

As part of the repeatability study, but also to test the reliability of minor allele detection in heteroplasmy calls, as well as to discriminate between true PHPs and mixture, we prepared simulated forensic mixed samples (Supplementary Table S1c). Buccal swab samples of two female persons MW-0002 and MW-0020 were selected, since they were previously used for repeatability studies, thus sequenced multiple times, and their sequence was by now well known. They were combined in the ratios 1:199 (MIX-1 = 0.5%), 1:99 (MIX-2 = 1.0%), 1:39 (MIX-3 = 2.5%) and 1:19 (MIX-4 = 5.0%). Mixed samples underwent long-range PCR (three replicates each) and library preparation protocol as previously described for all other validation samples. The two haplotypes differed in exactly 12 positions (4 in control region, 8 in coding region; Supplementary Table S7), which we targeted for analysis with the lowered thresholds. Other positions were not eligible for analysis and interpretation, since mixture ratios were mostly below the thresholds established by this evaluation.

Read depth for the targeted positions varied (minimum 1,461 reads; maximum 30,102 reads), but in all instances it was sufficient for the detection of minor contributor at the expected ratios. Minor contributor was successfully detected in all mixtures at the expected mtDNA positions. However, percentages of minor contributor alleles differed from the theoretical values: on average, in all four mixtures minor contributor was detected in excess to the expected ratio (Table 4). It was interesting to note that at positions 2259, 4745 and 14872 minor contributor alleles were detected with as much as twice the expected ratio (e.g. 1% instead of 0.5%, 10% instead of 5%, etc.). This particular position-specific phenomenon remains inexplicable, since these mtDNA positions do not reside within error-prone regions, neither does the major contributor exhibit additional PHPs at these coordinates which would tilt the ratios to such extent. Contributing to this unusual phenomenon is the fact that NextSeq results (as mixtures were sequenced alongside other samples in concordance study) showed identical trend, and almost identical values, among minor contributor ratios, for exactly the same three positions (data not shown).

Table 4 Proportion of minor contributor alleles per each of 12 mtDNA positions differing between samples MW-0002 and MW-0020 used in the mixtures study (as minor and major contributor, respectively). Three replicates of each mix ratio was prepared and sequenced. SD = standard deviation

rCRS position	72	2259	2706	4745	5897	7028	13680	14872	15904	16231	16298	16359				
rCRS base	T	C	A	A	C	C	C	C	C	T	T	T				
Major allele	T	T	A	G	C	C	T	T	C	T	T	T				
Minor allele	C	C	G	A	T	T	C	C	T	C	C	C	Observed average ± SD	Expected	Observed ratio	Expected ratio
MIX-1-1	0.7%	0.9%	0.6%	1.1%	0.5%	0.6%	0.3%	0.7%	0.5%	0.6%	0.5%	0.6%				
MIX-1-2	1.1%	1.0%	0.7%	1.2%	0.7%	0.9%	0.7%	0.9%	0.7%	0.9%	0.7%	0.8%	0.7 ± 0.2 %	0.5%	1:135	1:199
MIX-1-3	0.7%	0.9%	0.6%	1.2%	0.6%	0.5%	0.6%	0.9%	0.6%	0.9%	0.6%	0.6%				
MIX-2-1	1.4%	2.1%	1.2%	2.8%	1.4%	1.5%	0.7%	1.5%	1.2%	1.4%	0.9%	1.4%				
MIX-2-2	1.3%	1.3%	1.1%	2.0%	1.0%	1.3%	1.1%	1.6%	1.3%	1.5%	1.1%	1.2%	1.3 ± 0.4 %	1.0%	1:76	1:99
MIX-2-3	1.1%	1.6%	1.2%	1.7%	1.1%	1.3%	0.9%	1.2%	1.0%	1.3%	0.9%	1.0%				
MIX-3-1	3.2%	4.8%	2.9%	5.3%	3.2%	3.5%	2.0%	3.1%	3.1%	3.2%	2.2%	2.6%				
MIX-3-2	3.3%	4.4%	3.3%	4.5%	3.2%	3.6%	2.4%	3.8%	3.1%	3.5%	2.5%	2.9%	3.3 ± 0.7 %	2.5%	1:31	1:39
MIX-3-3	3.2%	4.0%	2.6%	4.5%	2.9%	3.1%	2.5%	3.7%	2.9%	3.2%	2.2%	2.8%				
MIX-4-1	6.6%	9.0%	5.4%	9.4%	6.1%	6.7%	4.2%	8.2%	6.3%	6.7%	5.0%	6.0%				
MIX-4-2	6.4%	8.4%	5.1%	9.8%	6.1%	6.5%	4.1%	7.2%	5.8%	6.6%	4.7%	5.6%	6.5 ± 1.6 %	5.0%	1:15	1:19
MIX-4-3	6.2%	8.9%	5.5%	10.6%	6.0%	6.7%	4.4%	7.1%	5.5%	6.4%	4.5%	5.6%				

One possible explanation for the difference between average observed minor contributor ratios and expected values is that it might have been caused by bias during long-range PCR: one contributor's mtDNA might have been amplified more efficiently than the other's. This would introduce slight change to the ratio of contributors from the start and eventually it would manifest itself in the results. Alternatively, as indicated in [17], the skewed observed ratios may more likely be the product of differences in mtDNA vs. nDNA quantity between samples: in that case, expected mixture ratios calculated from genomic DNA concentrations would not exactly correspond to the final results where mtDNA to mtDNA ratios were observed. Notwithstanding, whole mtDNA workflow in general consists of multiple steps wherein ratios of contributors may be affected. Thus, even though sequencing is reproducible and relatively precise, this method is not suitable for accurate detection of the proportion of minor contributor in mixed samples, as multiple preparation steps, in combination with the content of mtDNA within the sample, may introduce bias to the ratio of contributors.

Besides detection of minor contributor, we monitored the presence of two PHPs characteristic to the buccal swab sample of MW-0020, as described in previous sections (Supplementary Tables S4-S6). Both heteroplasmies (T152Y and T9325Y) were consistently called in all mixtures (Supplementary Table S9), regardless of the proportion of minor contributor, and their respective values correspond well to the minor allele percentages reported in previous experiments of this study.

4. Conclusion

Based on multi-component criteria of data analysis thresholds (in terms of read depth, percentage of alleles and quality scores), which were established in this study, we defined internal guidelines for analysis and interpretation of mtDNA results obtained by MPS. The proposed methodology proved robust and confident for variant calling and reporting when applied to analysis of controls and samples alike. Our study also shows that the whole mtDNA assay on MiSeq FGx™ produces repeatable and reproducible results (both between runs for the same analyst, and between different analysts) for all samples, equally for buccal swabs and blood samples, as well as for cell-culture-derived positive control samples (SRMs 2392 and 2392-I). Moreover, results were completely concordant with STS results [26], and were also concordant with results obtained on another MPS platform. Few minor discrepancies were observed, originating from heteroplasmy calls that did not comply with at least one component of defined analysis thresholds, but all calls were eventually confirmed in both datasets after analyst review; thus, no major discordance was noted. We conclude that this assay – including enrichment strategy, library preparation reagents, sequencing reagents, sequencing instrument and accompanying analysis software – is suitable for further use in our forensic laboratory. It will be further

used for Croatian population study on whole mitochondrial genomes, in order to establish a national database for the purpose of haplotype and haplogroup frequencies.

Some features of the analysis software may require additional attention in future upgrades, for example, dealing with leftover primer reads, treatment of indels and homopolymeric regions (a common struggle to almost every mtDNA analysis program), accommodation of forensic mitochondrial nomenclature, and also making more parameters available for user-modification in order to better tailor the analysis to specific study goals. All in all, Illumina® BaseSpace® Sequence Hub online bioinformatics platform is, at present, an acceptable solution for fast, intuitive, high-throughput data analysis which will be required for the population study.

Free online, cloud-based platforms such as BaseSpace®, with its plethora of applications, can be user-friendly, require little previous bioinformatic knowledge, and provide simple, fast, cost-effective solutions to streamline both data analysis and data storage. However, online solutions are unsuitable in a forensic setting, where data handling procedures are strictly prescribed by laws and protocols, dedicated off-line servers are used for analysis and storage of sensitive case-related information and analysis results in order to maintain their confidentiality, etc. Considering that, at some point in the future, whole mtDNA analysis by MPS will be implemented into routine forensic casework, the choice of analysis software will have to be reconsidered. Therefore, it is imperative that, in parallel to the population study, in the future, a comparison of other available analysis software be conducted, in order to decide the best bioinformatics solution for casework samples. Needless to say, they provide more challenge than reference samples used in evaluation and population studies, and would thus require a different approach not only in terms of analysis software, but in library preparation method as well.

5. References

[1] Butler JM (2012) Mitochondrial DNA Analysis. In: Butler JM (2012) Advanced Topics in Forensic DNA Typing: Methodology. Academic Press, Elsevier, USA, pp 405-456.

[2] Parson W, Huber G, Moreno L, Madel MB, Brandhagen MD, Nagl S, Xavier C, Eduardoff M, Callaghan TC, Irwin JA (2015) Massively parallel sequencing of complete mitochondrial genomes from hair shaft samples. Forensic Science International: Genetics 15:8-15.

[3] Lyons EA, Scheible MK, Sturk-Andreaggi K, Irwin JA, Just RS (2013) A high-throughput Sanger strategy for human mitochondrial genome sequencing. BMC Genomics 14:881.

[4] Just RS, Scheible MK, Fast SA, Sturk-Andreaggi K, Röck AW, Bush JM, Higginbotham JL, Peck MA, Ring JD, Huber GE, Xavier C, Strobl C, Lyons EA, Diegoli TM, Bodner M, Fendt L, Kralj P, Nagl S, Niederwieser D, Zimmermann B, Parson W, Irwin JA (2015) Full mtGenome reference data: Development and characterization of 588 forensic-quality haplotypes representing three U.S. populations. Forensic Science International: Genetics 14:141-155.

[5] Børsting C, Morling N (2015) Next generation sequencing and its applications in forensic genetics. Forensic Science International: Genetics 18:78-89.

[6] Clarke AC, Prost S, Stanton JAL, White WTJ, Kaplan ME, Matisoo-Smith EA, The Genographic Consortium (2014) From cheek swabs to consensus sequences: An A to Z protocol for high-throughput DNA sequencing of complete human mitochondrial genomes. BMC Genomics 15:68.

[7] ENFSI DNA Working Group (2010) Recommended minimum criteria for the validation of various aspects of the DNA profiling process, Issue no. 001.

[8] Scientific Working Group on DNA Analysis Methods – SWGDAM (2016) Validation guidelines for DNA analysis methods.

[9] Parson W, Gusmão L, Hares DR, Irwin JA, Mayr WR, Morling N, Pokorak E, Prinz M, Salas A, Schneider PM, Parsons TJ (2014) DNA Commission of the International Society for Forensic Genetics: Revised and extended guidelines for mitochondrial DNA typing. Forensic Science International: Genetics 13:134-142.

[10] Scientific Working Group on DNA Analysis Methods – SWGDAM (2019) Interpretation guidelines for mitochondrial DNA analysis by forensic DNA testing laboratories.

[11] Parson W, Strobl C, Huber G, Zimmermann B, Gomes SM, Souto L, Fendt L, Delport R, Langit R, Wootton S, Lagacé R, Irwin J (2013) Evaluation of next generation mtGenome sequencing using the Ion Torrent Personal Genome Machine (PGM). Forensic Science International: Genetics 7:543-549.

[12] Strobl C, Eduardoff M, Bus MM, Allen M, Parson W (2018) Evaluation of the presicion ID whole MtDNA genome panel for forensic analyses. Forensic Science International: Genetics 35:21-25.

[13] Woerner AE, Ambers A, Wendt FR, King JL, Moura-Neto RS, Silva R, Budowle B (2018) Evaluation of the precision ID mtDNA whole genome panel on two massively parallel sequencing systems. Forensic Science International: Genetics 36:213-224.

[14] King JL, LaRue BL, Novroski NM, Stoljarova M, Seo SB, Zeng X, Warshauer DH, Davis CP, Parson W, Sajantila A, Budowle B (2014) High-quality and high-throughput massively parallel sequencing of the human mitochondrial genome using the Illumina MiSeq. Forensic Science International: Genetics 12:128-135.

[15] McElhoe JA, Holland MM, Makova KD, Su MS, Paul IM, Baker CH, Faith SA, Young B (2014) Development and assessment of an optimized next-generation DNA sequencing approach for the mitogenome using the Illumina MiSeq. Forensic Science International: Genetics 13:20-29.

[16] Peck MA, Brandhagen MD, Marshall C, Diegoli TM, Irwin JA, Sturk-Andreaggi K (2016) Concordance and reproducibility of a next generation mtGenome sequencing method for high-quality samples using the Illumina MiSeq. Forensic Science International: Genetics 24:103-111.

[17] Peck MA, Sturk-Andreaggi K, Thomas JT, Oliver RS, Barritt-Ross S, Marshall C (2018) Developmental validation of a Nextera XT mitogenome Illumina MiSeq sequencing method for high-quality samples. Forensic Science International: Genetics 34:25-36.

[18] Qiagen (2014) EZ1® DNA Investigator® Handbook.

[19] Qiagen (2014) QIAamp® DNA Micro Handbook.

[20] National Institute of Standards and Technology – NIST (2018) Certificate of Analysis: Standard Reference Material® 2392 Mitochondrial DNA Sequencing (Human).

[21] National Institute of Standards and Technology – NIST (2018) Certificate of Analysis: Standard Reference Material® 2392-I Mitochondrial DNA Sequencing (Human HL-60 DNA).

[22] Illumina (2016) Protocol: Human mtDNA Genome for the Illumina Sequencing Platform, Document #15037958 v01.

[23] TaKaRa Bio Inc. (2015) PrimerSTAR® GXL DNA Polymerase Product Manual, Cat. #R050A, v201509Da.

[24] Illumina (2016) MiSeq® System Denature and Dilute Libraries Guide, Document #15039740 v01.

[25] Illumina (2019) NextSeq® System Denature and Dilute Libraries Guide, Document #15048776 v10.

[26] Barbarić L, Lipovac K, Sukser V, Rožić S, Korolija M, Zimmermann B, Parson W (2020) Maternal perspective of Croatian genetic diversity. Forensic Science International: Genetics 44:102190.

894 [27] Illumina (2016) mtDNA Variant Processor v1.0 BaseSpace App Guide, Document
895 #1000000007931 v00.
896 [28] Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP,
897 Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization
898 of the human mitochondrial genome. Nature 290(5806):457-465.
899 [29] Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis
900 and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature
901 Genetics 23(2):147.
902 [30] Gilder JR, Doom TE, Inman K, Krane DE (2007) Run-specific limits of detection and quantitation
903 for STR-based DNA testing. Journal of Forensic Sciences 52(1):97-101.
904 [31] Riman S, Kiesler KM, Borsuk LA, Vallone PM (2017) Characterization of NIST human
905 mitochondrial DNA SRM-2392 and SRM-2392-I standard reference materials by next generation
906 sequencing. Forensic Science International: Genetics 29:181-192.
907 [32] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011)
908 Integrative Genomics Viewer. Nature Biotechnology 29(1):24-26.
909 [33] Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP (2017) Variant review with the
910 Integrative Genomics Viewer (IGV). Cancer Research 77(21):31-34.
911 [34] Illumina (2015) MiSeq® System Specification Sheet.
912 [35] Illumina (2019) Cluster Optimization: Overview Guide, Document #1000000071511 v00.
913 [36] Hussing C, Kampmann ML, Smidt Mogensen H, Børsting C, Morling N (2018) Quantification of
914 massively parallel sequencing libraries – a comparative study of eight methods. Scientific
915 Reports 8:1110.
916 [37] Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ (2012)
917 Performance comparison of benchtop high-throughput sequencing platforms. Nature
918 Biotechnology 30(5):434-439.
919 [38] Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y
920 (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific
921 Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341.
922 [39] Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C (2015) Insight into biases and
923 sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids
924 Research 43(6):e37.
925 [40] Ring JD, Sturk-Andreaggi K, Peck MA, Marshall C (2017) A performance evaluation of Nextera
926 XT and KAPA HyperPlus for rapid Illumina library preparation of long-range mitogenome
927 amplicons. Forensic Science International: Genetics 29:174-180.

928 [41] Brandhagen MD, Just RS, Irwin JA (2020) Validation of NGS for mitochondrial DNA casework at
929 the FBI Laboratory. Forensic Science International: Genetics 44:102151.
930

Figure captions

Fig. 1 Schematic diagram of analysis steps performed on samples of negative and positive controls. All controls underwent analysis in BaseSpace® mtDNA Variant Processor using identical thresholds. Genome variant call format (GVCF) files were perused in detail only for negative controls. Excel reports were perused both for negative and positive controls. After performed calculations, internal analysis and interpretation thresholds (INT) were defined and estimated conservatively

Abbreviations: F = forward; R = reverse; MIN = minimum value; MAX = maximum value; AVERAGE = mean (average) value; STDEV = standard deviation; LOD = limit of detection; LOQ = limit of quantitation; NC = negative control; SRM-C = SRM® 2392 CHR; SRM-H = SRM® 2392-I HL-60; PHP = point heteroplasmy

Fig. 2 Maximum read depth per mtDNA position of all signals detected in negative controls. Two regions of interest (i.e. with conspicuously high read coverage) are marked with arrows: primer MTL-R1 coordinates (1873-1893) and part of hypervariable region HVS-I (16128-16455)

Fig. 3 Graphical representation of maximum read depth of alternative alleles per position (a) and maximum percentage (%) of alternative alleles per position (b) in positive control samples (SRM® 2392 CHR and SRM® 2392-I HL-60). Extremes detected in positions 2487 and 16189 (on both graphs) are marked with arrows

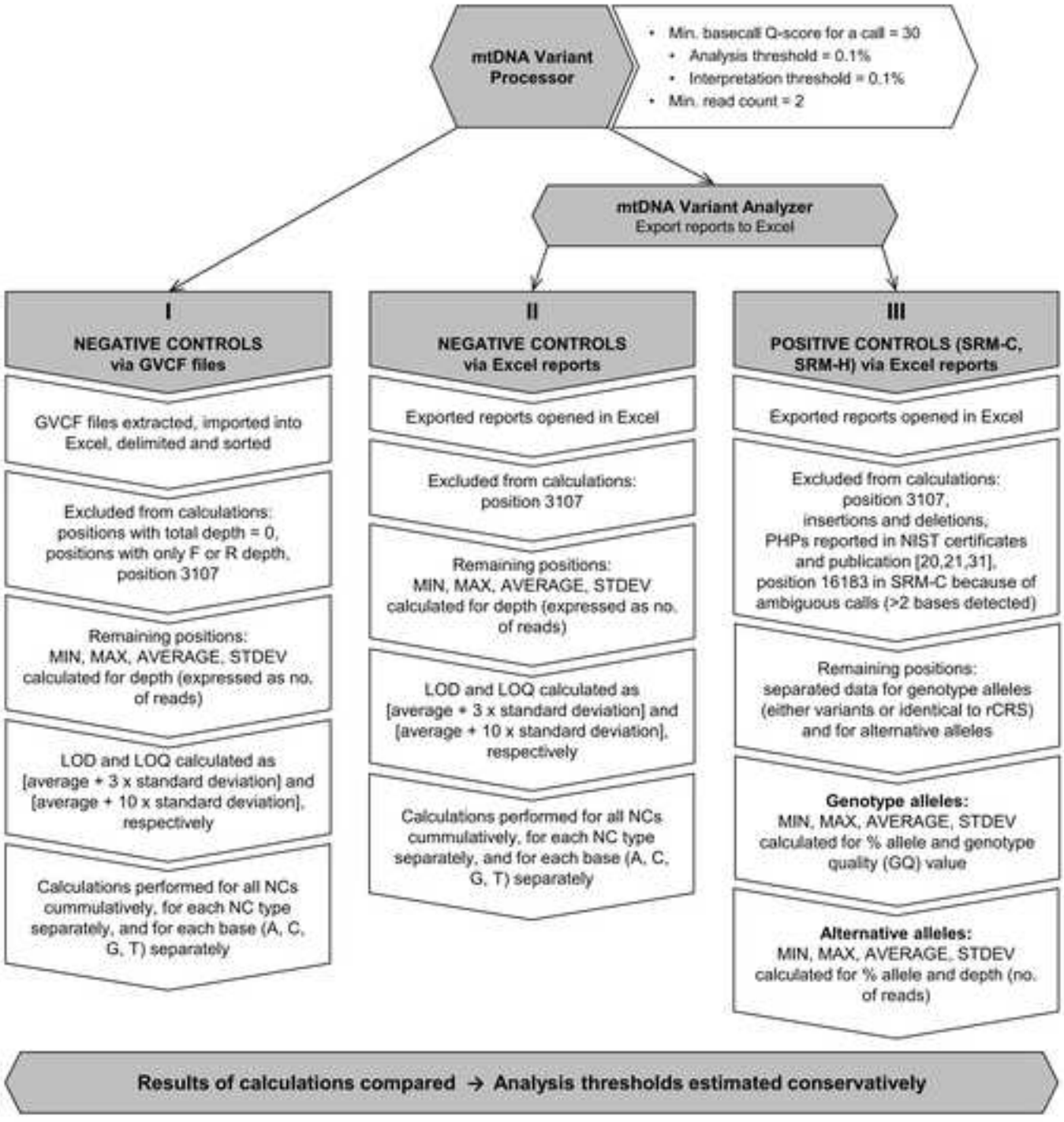
Supplementary material description

Electronic supplementary material file contains:

- **Fig. S1**: coverage profiles of selected samples (NIST SRM® controls, buccal swab and blood samples), with highlighted areas exhibiting coverage drops;
- **Fig. S2**: visualization of primer MTL-R1 reads in negative controls, via IGV tool;
- **Table S1**: outline of evaluation experiments: repeatability, reproducibility, mixtures and concordance studies;
- **Table S2**: comparison of sequencing quality metrics for Read 1 and Read 2;
- **Table S3**: base counts and their respective proportions in signals detected in negative controls;
- **Table S4**: repeatability study results for point heteroplasmy calls detected in samples included in the study;
- **Table S5**: reproducibility study results for point heteroplasmy calls detected in samples included in the study;

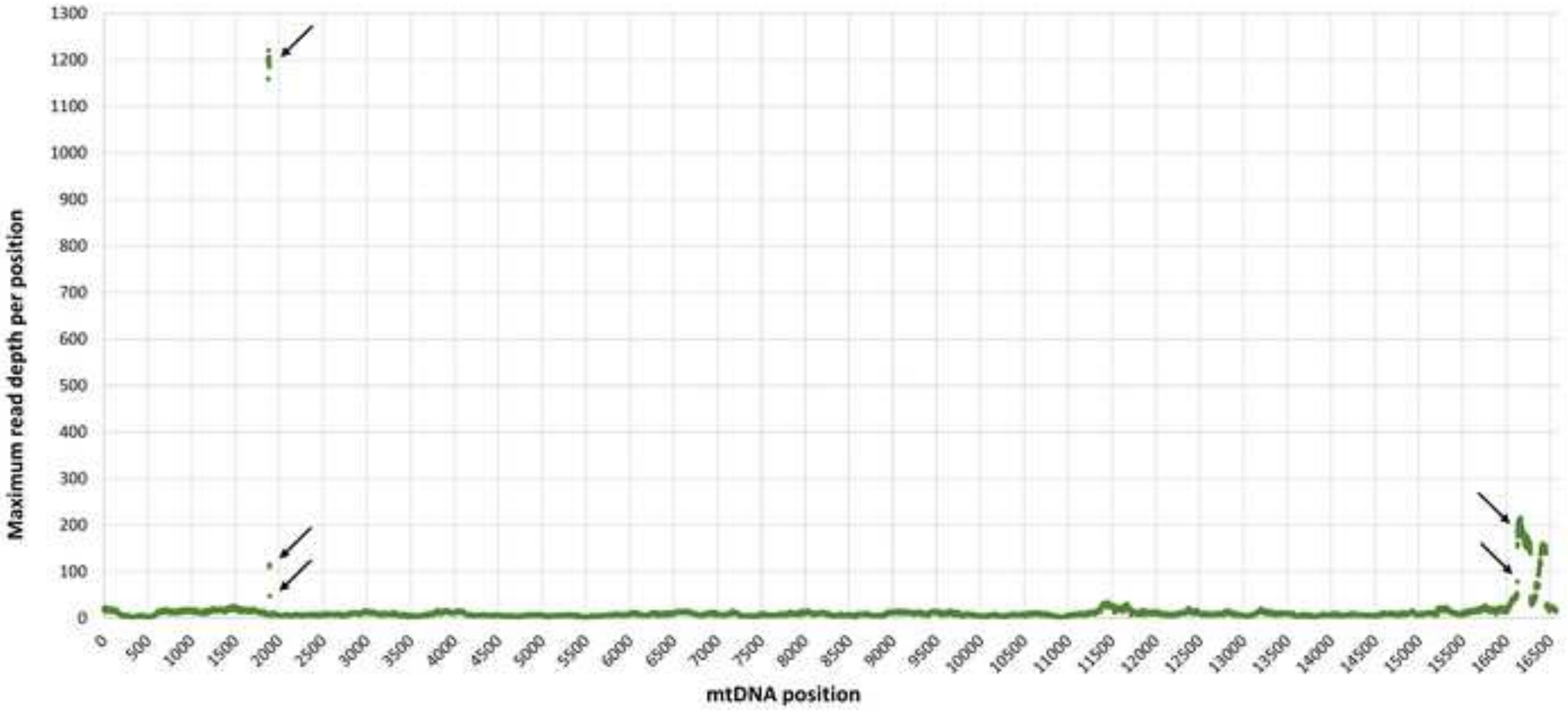
- **Table S6**: reproducibility results of point heteroplasmy calls in sample MW-0020 sequenced in all runs;
- **Table S7**: comparison of samples' haplotypes generated in this study to Sanger-type sequencing results;
- **Table S8**: concordance results of samples' point heteroplasmy calls between MiSeq FGx™ and NextSeq®500 platforms;
- **Table S9**: detected proportions of minor allele in point heteroplasmy calls of sample MW-0020 (major contributor in mixtures).

Figure



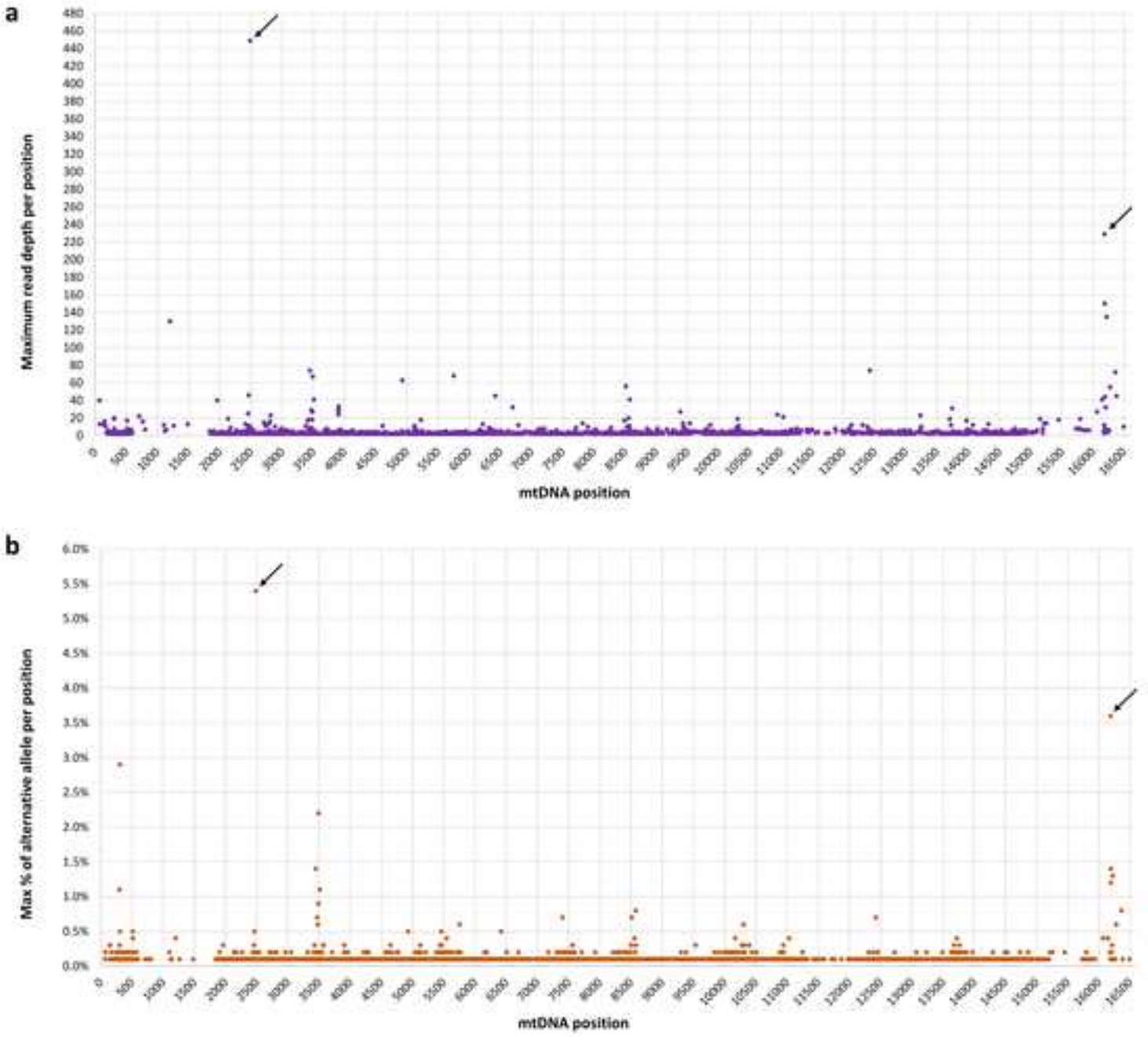
Figure

Click here to access/download;Figure;Fig2.tif



Figure

Click here to access/download;Figure;Fig3.tif



Click here to access/download

Supplementary Material

Suppl-material-IJLM.xlsx