Forensic Science International: Genetics 28 (2017) 90–98

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Research paper

Evaluation of GeneMarker® HTS for improved alignment of mtDNA MPS data, haplotype determination, and heteroplasmy assessment

Mitchell M. Holland*, Erica D. Pack, Jennifer A. McElhoe

Forensic Science Program, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, 014 Thomas Building, University Park, PA 16802, United States

ARTICLE INFO

Article history:
Received 30 October 2016
Received in revised form 4 January 2017
Accepted 25 January 2017
Available online 6 February 2017

Keywords:
Bioinformatics tools
Mitochondrial DNA
Forensic science
Clinical
Software

ABSTRACT

Existing software has not allowed for effective alignment of mitochondrial (mt) DNA sequence data generated using a massively parallel sequencing (MPS) approach, combined with the ability to perform a detailed assessment of the data. The regions of sequence that are typically difficult to align are homopolymeric stretches, isolated patterns of SNPs (single nucleotide polymorphisms), and INDELs (insertions/deletions). A custom software solution, GeneMarker® HTS, was developed and evaluated to address these limitations, and to provide a user-friendly interface for forensic practitioners and others interested in mtDNA analysis of MPS data. GeneMarker® HTS generates an exportable consensus mtDNA sequence that produces phylogenetically correct SNP and INDEL calls using a customizable motif-based alignment algorithm. Sequence data from 500 individuals, with various alignment asymmetries and levels of heteroplasmy, were used to assess the software. Accuracy in producing mtDNA haplotypes, the ability to correctly identify low-level heteroplasmic sequence variants, and the user-based features of the software were evaluated. Analyzed sequences yielded correct mtDNA haplotypes, and heteroplasmic variants were properly identified with minimal manual interpretation. The software offers numerous user-defined parameters for filtering the data that address the interests of researchers and practitioners, and provides multiple options for viewing and navigating through the data.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

The analysis of mitochondrial (mt) DNA sequence has been employed to exonerate individuals convicted of a crime, associate perpetrators to a crime, and support the identification of mass fatality victims, military personnel and historical figures [1–5]. With the onset of massively parallel sequencing (MPS) approaches for generating mtDNA sequence data, along with a growing interest in sequence analysis of SNP (single nucleotide polymorphisms) and STR (short tandem repeat) marker systems, the forensic community will have an opportunity to expand the use of mtDNA analysis in forensic casework. Given the increased resolution power of MPS approaches, heteroplasmic variants of mtDNA sequence can be readily detected and resolved [6–8]. Therefore, the forensic community can routinely report heteroplasmic sequences in casework, enhancing the discrimination potential of the typing system [5]. In order to reach this goal, software solutions and pipelines will be needed to effectively analyze the resulting MPS data.

A number of software tools have been developed for the analysis of mtDNA MPS data. Some involve the reconstruction of mtGenomes from total genomic DNA [9], which is currently beyond the scope of forensic applications. Others have toolboxes from which pipelines can be created to assist in the process of data management [10–14], but lack the proper nomenclature conversion, analysis of the pileup of sequencing reads, or can be challenging to use. One software package, mitoSAVE, uses a simple Excel-spreadsheet-based solution [15] that converts the data into haplotypes using phylogenetically-derived nomenclature, and allows for consistent assignment of homopolymeric stretches and sequence motifs [16,17], but does not allow for analysis of the individual sequencing reads in the pileup.

The commercially available software package, NextGENe® (SoftGenetics, Inc., State College, PA), has been used by a number of researchers and clinicians to analyze MPS data for the diagnosis of cancer, detection of cardiomyopathies, and evaluation of hereditary hearing loss [18–20]. The software includes numerous user-defined parameters, and allows for detailed analysis of the pileup. NextGENe® has been used successfully to analyze mtDNA

* Corresponding author.
E-mail address: mmh20@psu.edu (M.M. Holland).

http://dx.doi.org/10.1016/j.fsigen.2017.01.016
1872-4973/© 2017 Elsevier B.V. All rights reserved.

sequence data for forensic applications [7,10], but has lacked an integrated pipeline and desired features. We previously reported on the early development of a customized version of the NextGENe® software [21] to address the following considerations; 1) alignment to a circular version of the mtgenome so that data properly spans the transition point in the mtgenome numbering system [22], 2) alignment and nucleotide numbering consistent with the revised mtgenome sequence [23], 3) recognition of SNP-associated motifs and INDELs (insertions/deletions) consistent with phylogenetic and forensic considerations [16,17], 4) identification of heteroplasmic sequences, and 5) export of reports that address forensic considerations and allow for import into tertiary analysis tools such as EMPOP; www.empop.org, v3/R11. The alignment strategies drew from previous attempts to accomplish these goals [15]. The fully developed software, including a new user interface and reporting tools, has been renamed GeneMarker® HTS (GM-HTS)(SoftGenetics, Inc.) and is commercially available. We demonstrate here the performance of GM-HTS for forensic applications through the evaluation of 500 mtDNA sequences with respect to; 1) user-based features, including the production of useful and accurate reports, 2) the ability to properly align homopolymeric sequences, SNPs and INDELs, and identify phylogenetically correct primary haplotypes with minimal user input, and 3) the identification of heteroplasmic variants with minimal user input.

2. Materials & methods

Data sets containing more than 700 MPS mtDNA sequences from random individuals in the general population have been analyzed in our laboratory with the NextGENe® software. A sampling of 500 sequences from these data sets was selected to experimentally demonstrate the utility of GM-HTS. The MPS data used in this evaluation were generated on an Illumina MiSeq (San Diego, CA). The sample source was buccal samples taken primarily from individuals of European ancestry; 21 non-European and 479 European. All laboratory work for this study was conducted under the Penn State University IRB approved project number, STUDY00000970.

The NextGENe® software performs alignment using a BLAST-Like Alignment Tool (BLAT) method, which employs a Smith-Waterman [24] approach with a proprietary INDEL alignment algorithm, and provides customizable post-processing reports. The filter settings chosen followed previously established guidelines established in our laboratory [7]. FASTQ files generated by MiSeq Reporter (MSR v2.4.60; Illumina, Inc., San Diego, CA) were converted to FASTA files in NextGENe®, removing reads that did not meet the previously described quality thresholds. The successfully converted reads were then mapped to the revised Cambridge Reference Sequence (rCRS; GenBank ID NC_012920.1) [23]. Reads passing the quality filters and aligning to the reference were considered matched reads and were used to generate a mutation report based on the following settings: SNP mutations retained when the mutation percentage ≥2%; required SNP allele read count ≥40; required total read coverage ≥200; and SNPs with a read frequency less than 90% run through a read balance filter requiring a balance ratio of ≥0.2. The INDELs were retained using the same parameters except for the balance ratio; INDELs with a read frequency less than 60% were run through the filter and required a balance ratio of ≥0.1. In summary, our interpretation threshold for detecting variants was 2%, so for a minor allele to be considered reportable, any given nucleotide position required a minimum of 2000 reads. NextGENe® generated mutation reports were manually interpreted to determine the final sequencing calls used in the software comparison. For all sequencing data, regardless of software platform, variant positions were included in the major haplotype when variants were detected at a frequency greater than 50% and located within the control region (CR); nucleotide positions (nps) 16,024–16,569 and 1–576.

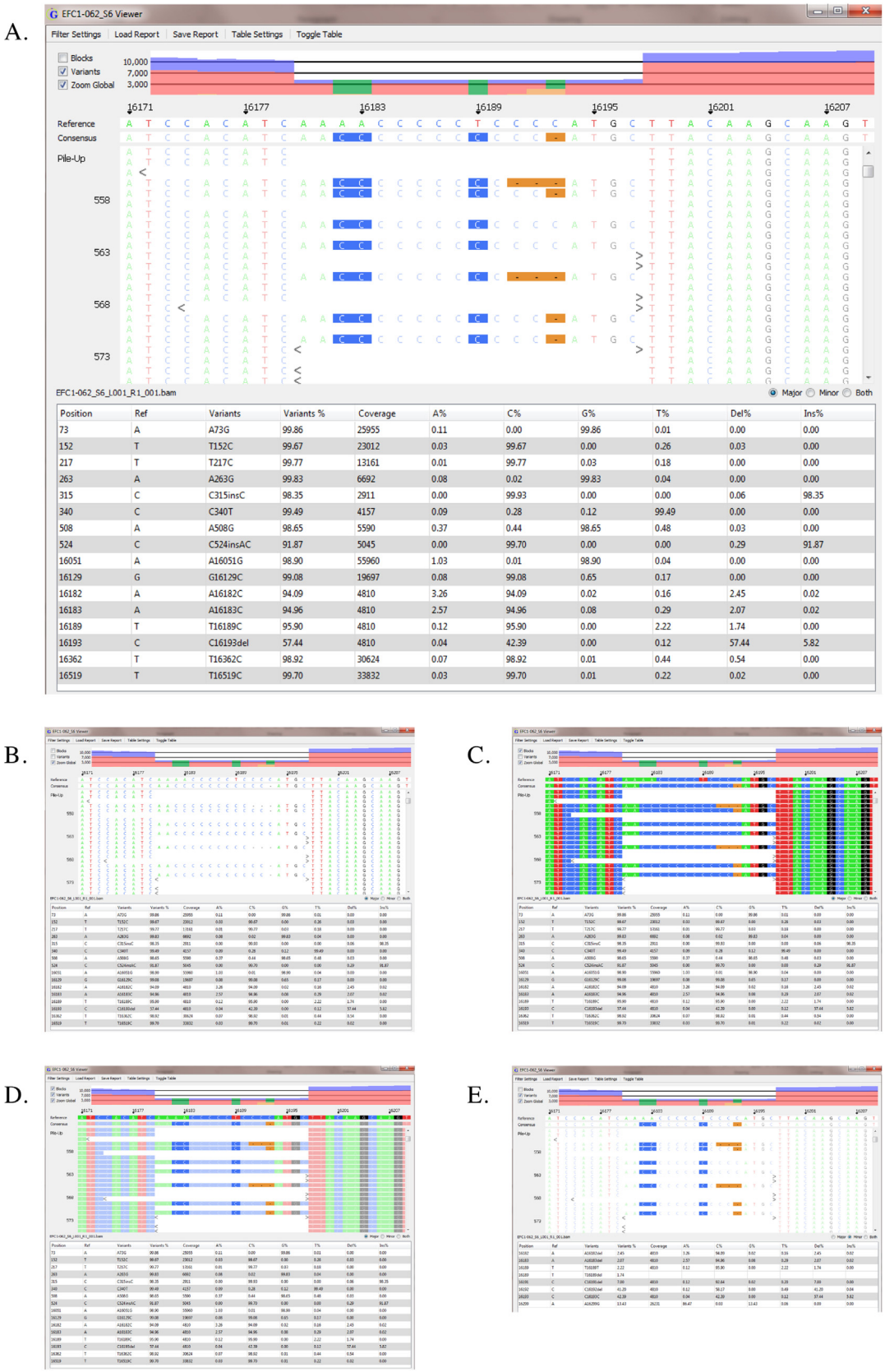
The alignment algorithm in GM-HTS performs a Burrows-Wheeler [25] hash alignment based on spaced seeds (13 bases, ignore 1 base, and 13 more bases) and fills in gaps with dynamic programming. After alignment, a motif file (built-in or user-customized) can be applied to the reads. The motif file consists of a list of variant calls that are translated into an expected sequence. Each motif region is defined by a start and end nucleotide position and is inclusive, meaning that reads that do not span the entire region are trimmed. Alignment of reads spanning a defined motif region is adjusted to match the expected alignment pattern. Output files include BAM/BAI alignment, alignment statistic, consensus sequence, consensus statistic, primary report, minor (heteroplasmy) report, project, and project settings. For this exercise, FASTQ files were mapped to the rCRS using the following alignment options: customized motif file, 85% identity, and soft clipping at locations with three consecutive bps with a quality score ≤29. Table report settings were as follows: input region nucleotide position (np) 16,024 through the origin to position 576, variant percentage ≥1% as the analytical threshold, variant allele coverage ≥40, total coverage ≥200, allele balance ratio ≤2.5, and allele score balance ≤10. A reporting threshold of 2% was used for heteroplasmic positions. The motif file, a simple text file containing phylogenetically correct sequence motifs that instructs the software which alignments are preferred by the user, contained 118 motifs that included those collected from the literature [16,17], as well as user-defined motifs based on new sequence patterns observed in the data set. Sixty (~51%) of the motifs were located in hypervariable region I (HVI; np 16,024–16,365) of the CR, with ~87% of those motifs (52/60) spanning the C-stretch surrounding position 16,189. The C-stretch surrounding position 310 had ~37% of the 43 motifs located in HVII (np 73–340), and the remaining 15 motifs were located in the CR outside of the hypervariable regions.

As new versions of GM-HTS were developed, data sets were run through the software to ensure that the outcomes were concordant with previous analyses, and alterations were made to address shortcomings. Review of the software was a step-wise process with assessments of the early versions establishing functionality of the user interface, viewing options, processing and rendering time requirements, and report generation. Evaluation of later versions focused on properly aligning homopolymeric sequences/SNPs/INDELs and production of useful and accurate reports, including identification of phylogenetically correct primary haplotypes and heteroplasmic variants while requiring minimal user input. The most recent version, v0.20160616, was used to analyze the data for the current study. The software is available by contacting SoftGenetics, Inc. (www.softgenetics.com).

EMPOP tools, EMPcheck [26] and Network [27], were used for secondary analysis of the data as a final means of quality control. EMPcheck validates the format and the content of a text file containing haplotype information, and the Network tool highlights problematic data, possible ambiguities, and errors through visualization of the genetic structure of the lineages in the data set. Network helps to detect peculiarities in datasets and can highlight data interpretation issues by removing, or filtering, highly recurring mutations. Due to the large number of samples in the data set, a super fine filter, EMPOPall_R11, that contains all mutations observed in EMPOP, was selected to calculate and draw a quasi-median network.

3. Results & discussions

The following items were addressed when evaluating GeneMarker® HTS (GM-HTS) for forensic applications; 1) user-based



Position	Ref	Variants	Variants %	Coverage	A%	C%	G%	T%	Del%	Ins%
73	A	A73G	99.86	25955	0.11	0.00	99.86	0.01	0.00	0.00
152	T	T152C	99.67	23012	0.03	99.67	0.00	0.26	0.03	0.00
217	T	T217C	99.77	13161	0.01	99.77	0.03	0.18	0.00	0.00
263	A	A263G	99.83	6692	0.08	0.02	99.83	0.04	0.00	0.00
315	C	C315insC	98.35	2911	0.00	99.93	0.00	0.00	0.06	98.35
340	C	C340T	99.49	4157	0.09	0.28	0.12	99.49	0.00	0.00
508	A	A508G	98.65	5590	0.37	0.44	98.65	0.48	0.03	0.00
524	C	C524insAC	91.87	5045	0.00	99.70	0.00	0.00	0.29	91.87
16051	A	A16051G	98.90	55960	1.03	0.01	98.90	0.04	0.00	0.00
16129	G	G16129C	99.08	19697	0.08	99.08	0.65	0.17	0.00	0.00
16182	A	A16182C	94.09	4810	3.26	94.09	0.02	0.16	2.45	0.02
16183	A	A16183C	94.96	4810	2.57	94.96	0.08	0.29	2.07	0.02
16189	T	T16189C	95.90	4810	0.12	95.90	0.00	2.22	1.74	0.00
16193	C	C16193del	57.44	4810	0.04	42.39	0.00	0.12	57.44	5.82
16362	T	T16362C	98.92	30624	0.07	98.92	0.01	0.44	0.54	0.00
16519	T	T16519C	99.70	33832	0.03	99.70	0.01	0.22	0.02	0.00

Fig. 1. A. Sample EFC1-062 in the **Variant** nucleotide viewing mode; European (E) female (F) cheek (C) swab sample from an individual 18–30 years of age (1). This is the preferred view, as nucleotide positions with sequence variants compared to the **Reference** are highlighted. The **Reference** is the revised Cambridge reference sequence [23]. The **Consensus** reflects continuous reads of sequence in the **Pileup** that span nps 16,180 through 16,197 of the C-stretch region in HV1; nps 16,171–16,207 are visible in the window and can be shifted by right-clicking on the mouse and dragging the data in either direction. The **Pileup** is individual sequences generated on the MiSeq which have been aligned to the **Reference**. The read-based density map along the top of the window is in **Zoom Global** mode and illustrates that at least 10,000 reads per nucleotide position have been aligned from the data outside of the motif window, with greater than 3,000 reads within the window. The teal green bars in the read density map refer to

features, including the production of useful and accurate reports 2) the ability to properly align homopolymeric sequences, SNPs and INDELs, and identify phylogenetically correct primary haplotypes with minimal user input, and 3) the identification of heteroplasmic variants with minimal user input.

3.1. User-based features & reporting

Program settings within the GM-HTS software allow for user management, which requires user names and passwords for login, and limits permissions. Designation as a windows administrator is required to enable or disable user management. Persons designated as "super users" are able to create new users and designate permissions for almost every feature in the software. GM-HTS allows for multiple samples to be batch processed, with up to 72 samples being queued for alignment in a single batch for this study; requiring less than 15 min of processing time. After processing, a sample profile can be opened in a viewer window, allowing visualization of the sequencing reads through a pileup, a global view displaying the depth of coverage, and a report table listing variant calls (Fig. 1A). The pileup can be viewed as colored blocks (background becomes colored and the font is white) or letters (text is colored and displayed on a white background) designating the nucleotide sequenced at a given position, and the user can choose to highlight variants in either setting (Fig. 1B–D). The software allows for mapping to a circular genome and displays the region of interest (ROI) in the center of the viewing window. This eliminates the need to navigate across the empty space created when using a linear reference (np 1–16,569) with an ROI limited to only a partial region of the genome that spans the origin (i.e., CR sequence including np 16,024–16,569 & 1–576). The global zoom feature shows the distribution of coverage across the ROI, allows the user to visualize the position within the genome of coverage being viewed in the pileup, and highlights the location of variants. Forward reads are represented in blue, reverse reads are designated in red, and shaded areas indicate regions below the total coverage user-determined value in Table Filter Settings; Total Coverage $\geq$200. The location of variants within the global zoom appear as teal green and orange lines representing the presence of a polymorphism (SNP or INDEL) and minor sequence variants, respectively.

As seen in Fig. 1A, the read density for the motif region is lower than the flanking read coverage; ~3,000 reads to ~10,000 reads, respectively. Only continuous reads that stretch entirely across the defined motif region (in this case, np 16,180–16,197) are included during the alignment process. This approach eliminates poor alignment outcomes observed when using previous versions of software that resulted in considerable manual interpretation. However, the reduced read density in these regions may impact the interpretation of low-level heteroplasmic variants depending on the total read coverage. Using our interpretation approach, the read density for the example in Fig. 1A is suitable for reporting out minor variants at or above 2%. Fig. 1E illustrates the viewing mode with the minor table selected.

In the report table, GM-HTS not only allows the user to view all calls in the report, but also conveniently allows the user to toggle between reports isolating the major and minor variant calls. Text files containing the major and minor variants are generated for each sample eliminating the need for a user-created pipeline to separate major and minor variant calls, and reducing the potential introduction of errors that naturally occurs through transcription of data. When double-clicking on a position in the report table, the pileup moves to that position in the sequence. Variants may be manually added or removed, and a comment recorded. Any variants added or removed are then shaded (green and red, respectively) in the report table window. While analyzing a sample, the filter settings window can be viewed and modified and any changes to variant calls are reflected in the report table. This application allows the user to quickly assess the effect of altering the filter parameters, an evaluation that could only be achieved by re-processing a sample with the previous version of the software.

Multiple samples from a project can be opened at the same time, and comparison tools allow for sample-to-sample and variant comparison between all samples in a project. An example is provided in Fig. 2A–C. The Sample-to-Sample comparison panels illustrate the similarity between samples based on comparing major-to-major or major-to-minor profiles using the Jaccard similarity index [28]. The Sample-to-Sample comparison matrix consists of 3×3 samples, with the rows always representing the major profile (haplotype) and the columns representing either the major (Fig. 2A) or minor (Fig. 2B) profiles; the number of rows and columns expands to reflect the number of samples in a project. The three samples used to illustrate the tool have the following haplotypes: EFC3-007 [European (E) female (F) cheek (C) swab sample from an individual greater than 50 years of age (3), 7th donor collected] is A263G, 309.1C, 315.1C, T16,209C, C16,259A, T16,519C; EMC1-050 [male (M) sample from an individual 18–30 years of age (2), 50th donor collected] is A73G, C150T, T152C, T195C, A215G, A263G, C295T, 310.1T, 315.1C, T319C, T489C, G513A, C16,069T, T16,126C, G16,145A, T16,231C, C16,261T; and EMC2-079 [sample from an individual 31–50 years of age (2), 79th donor collected] is A73G, A263G, 315.1C, T16,519C. A quick assessment of the data reveals that samples EFC3-007 and EMC1-050 share 2 of the collective 21 SNPs and INDELs represented in their respective haplotypes, or major profiles; positions A263G and 315.1C. The Variant Comparison panel (Fig. 2C) provides the variant nucleotide positions and frequencies for each sample allowing for direct comparisons, for identification of related and unrelated profiles, and can be used as an indicator of possible contamination.

3.2. Proper alignment of homopolymeric stretches, SNPs and INDELs producing phylogenetically correct haplotypes

When considering SNP data, the haplotypes for all 500 samples in the MPS data set were concordant with previous results when analyzed with GM-HTS. The improved alignment allowed for more precise reporting of length variants such as the 309.1C insertion. GM-HTS uses a Burrows-Wheeler hash alignment strategy in conjunction with a motif file alignment approach. The NextGENe® software uses an optimum alignment approach, which focuses on minimizing the number of differences between the sequenced read and the reference sequence (Fig. 3A). Since there is an established convention for mtDNA typing nomenclature, the motif file, a simple text file containing phylogenetically correct sequence motifs [16,17], instructs the software on which alignments are preferred by the user and helps the software achieve a forensic

SNPs or INDELs in the primary haplotype, and the orange bars represent minor components above the analytical threshold of 1%. The report table below the **Pileup** lists the **Major** SNPs and INDELs, or haplotype for the sample which can be exported for database searching and reporting. The list of information provided in the table is customizable. B. Sample EFC1-062 in the nucleotide viewing mode, without the variant positions highlighted. C. Sample EFC1-062 in the **Block** nucleotide viewing mode. Nucleotide positions are blocked-out and color-coded; green for A, blue for C, red for T, and black for G. D. Sample EFC1-062 in the **Block Variant** nucleotide viewing mode. Nucleotide positions are blocked-out and color-coded, with nucleotide positions consistent with the **Reference** shaded out of view. E. Sample EFC1-062 in the **Variant** nucleotide viewing mode with the **Minor** table selected. Minor variants within the viewing window include 16,189T and 16,193C, at 2.22% and 42.39%, respectively. The remaining variants in the window are INDEL-related, which is commonly observed in the HV1 C-stretch when SNPs 16,182C, 16,183C, and 16,189C are present in the haplotype. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A.

Comparison Viewer

Comparison Filter Settings | Major to Major | Major to Minor | Minor to Minor

Sample to Sample Comparison Proportion of Shared Variants: None All Save As

	EFC3-007_S67	EMC1-050_S23	EMC2-079_S22
EFC3-007_S67	6/6	2/21	3/7
EMC1-050_S23	2/21	17/17	3/18
EMC2-079_S22	3/7	3/18	4/4

Row: Both Major Minor Column: Both Major Minor

Showing (Row's Major ∩ Column's Major) / (Row's Major ∪ Column's Major) Show Percentage Use Jaccard Index

B.

Sample to Sample Comparison Proportion of Shared Variants: None All Save As

	EFC3-007_S67	EMC1-050_S23	EMC2-079_S22
EFC3-007_S67	0/6	1/9	0/7
EMC1-050_S23	0/17	0/21	1/17
EMC2-079_S22	0/4	1/7	0/5

Row: Both Major Minor Column: Both Major Minor

Showing (Row's Major ∩ Column's Minor) / (Row's Major ∪ Column's Minor) Show Percentage Use Jaccard Index

C.

Variant Comparison Major Allele Minor Allele Low Coverage Save As

	EFC3-007_S67	EMC1-050_S23	EMC2-079_S22
A73G	0%	99%	99%
C150T	0%	99%	0%
T152C	0%	99%	0%
T195C	0%	99%	17%
A215G	0%	99%	0%
A263G	99%	99%	99%
C295T	0%	99%	0%
C309insC	71%	0%	1%
T310insT	0%	90%	0%
C315insC	94%	96%	97%
T319C	0%	99%	0%
T489C	0%	98%	0%
G513A	0%	99%	0%
A523del	1%	2%	0%
C524del	1%	2%	0%
C16069T	0%	98%	0%
T16126C	0%	98%	0%
G16145A	0%	99%	0%
T16209C	99%	0%	0%
T16231C	0%	99%	0%
C16259A	98%	0%	0%
C16261C	99%	9%	99%
C16261T	0%	90%	0%
T16519C	99%	6%	99%

Fig. 2. The **Sample to Sample Comparison** shows the similarity between samples in a project based on comparing major, major to minor, or minor variants. The **Variant Comparison** shows variant positions and gives the variant frequency for each sample within a project. A. **Sample to Sample Comparison** with rows and columns representing comparison of the major profile, and applying the Jaccard index for a project consisting of non-related individuals. B. **Sample to Sample Comparison** with rows representing the major profile and columns representing the minor profile, compared by applying the Jaccard index for a project consisting of non-related individuals. C. **Variant Comparison** for the 3 × 3 sample matrix, with SNPs and INDELs listed in the first column and percentages of the variant listed in the subsequent three columns. Descriptions taken from the GeneMarker® HTS Quick Start Guide (v.28 July 2016).

A.

	16180				16185					16190					16195		
Reference	A	A	A	A	C	C	C	C	C	T	C	C	C	C	A	T	G
Optimal Alignment	A	A	-	-	C	C	C	C	C	T	C	C	C	CC	A	T	G
Forensic Alignment	A	A	C	C	C	C	C	T	C	C	C	C	C	-	A	T	G

B.

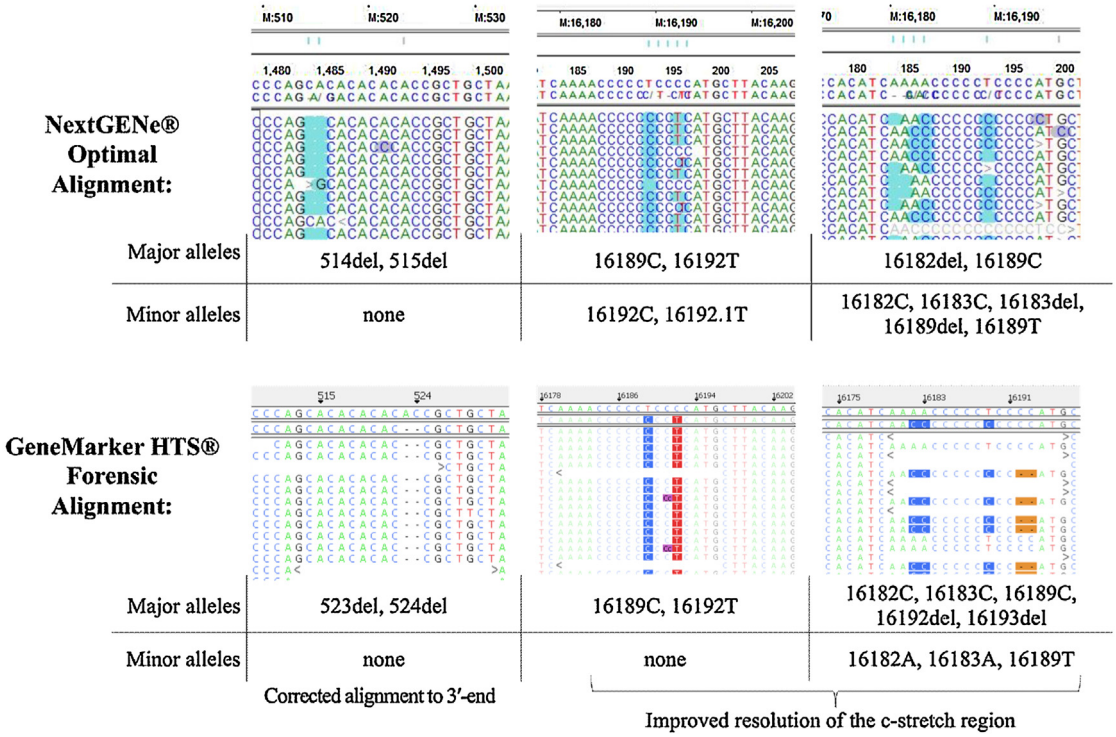


Fig. 3. Examples of optimal (NextGENe® software) versus forensic (GeneMarker® HTS software) alignment. A. Simple example of DNA sequence aligned to a reference sequence comparing an optimal alignment to a forensic alignment. B. Comparison of alignments generated using NextGENe® and GeneMarker® HTS demonstrating corrected alignment to the 3′-end of a homopolymeric stretch and improved resolution in the C-stretch region. The major and minor allele calls as determined by the respective software are given below each example. Descriptions taken from the GeneMarker® HTS Quick Start Guide, 28 July 2016.

alignment. The text file can be expanded and customized to suit the user's interests. When samples are processed, the software uses the motif file to guide alignment, producing reports containing phylogenetically correct haplotypes and heteroplasmic variants with minimal user input (Fig. 3B). A listing of all 118 motifs used for this study is provided in Supplemental Fig. 1. Examples of line items in the motif file are as follows:

#285-294
291del
290del 291del

This instructs the alignment software that between np 285–294 there are two motif driven alignments, one that results in a haplotype of 291del, and a second that results in 290del and 291del. The sequence in this segment of the CR contains a homopolymeric stretch of adenines (poly-A) that extends from np 286 to 291. Without the motif driven alignment, partial reads align incorrectly creating a pool of false heteroplasmic variants, and the software attempts to align the deletions at the 5′-end of the poly-A sequence resulting in an incorrect haplotype when considering recommended forensic and phylogenetic nomenclature. By using the motif alignment strategy, these inconsistencies are eliminated. When new sequence motifs are identified, they can be added to the text file and the sample reanalyzed to improve subsequent alignments.

Manual interpretation of haplotypes was rarely required (1.4% of the 500 sequences analyzed) when using GM-HTS. Comparing the output from GM-HTS to the manually interpreted haplotypes that were determined using NextGENe® indicated that both alignment strategies produce robust haplotypes, but the GM-HTS software eliminated the need for user-adjusted calls in homopolymeric stretches to produce an output with forensically correct 3′-end INDEL placement and SNP designations. After the haplotypes were manually compared, the EMPcheck and Network (EMPOP. online) tools were used as a final quality control measure. The EMPcheck tool validates the format and the content of a text file containing haplotype information, and the Network tool highlights problematic data, possible ambiguities, and errors through visualization of the genetic structure of the lineages in the data set. The 500 haplotypes generated using GM-HTS were extracted electronically from the individual primary reports, compiled into a text file with a valid EMPOP structure, and processed.

The Network tool produced a network with a root node containing 493 of the 500 haplotypes and seven nodes, each containing one haplotype, with yet unobserved mutations (Fig. 4). This quickly identified seven samples that required manual interpretation. The mutations that were identified as possible errors were manually evaluated and it was determined that five of the possible errors were correctly called while two samples had mutations that could be aligned in multiple ways, producing various haplotypes. For one of these samples, EMPcheck reported a strange CT insertion after position 56 and the Network tool identified a possible error with calls 56.1CT and 66del. This haplotype consisted of multiple SNPs and INDELs between np 55–75; 58C, 60.1T, 60.2T, 65del, 71del, and 73G. The motif table originally did not list the haplotype correctly, and reported the

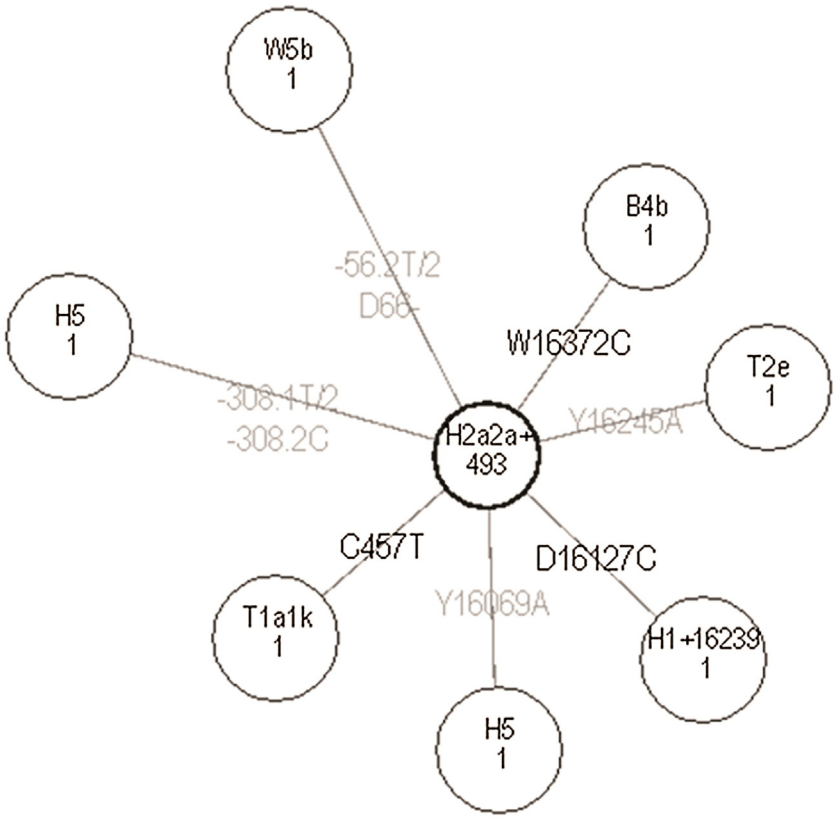


Fig. 4. Quasi-median network generated using EMPcheck and Network tools (empop.online) with 500 haplotypes, a data range of 16,024-576, and the EMPOPall_R11 filter. The filter removes all mutations observed in EMPOP and highlights only yet unobserved mutations. Each node represents filtered, reduced, and condensed haplotypes. Each circle (node) shows haplogroup information and how many samples are contained within the given node. The information between the root node (center circle) and the other nodes are mutational events, with black text representing transitions and grey text representing all other mutations (transversions, insertions, deletions).

profile incorrectly as 57CTins, 65del, 66del, and 73G. Fig. 5 provides a screen shot of the sequence the incorrect alignment. All remaining haplotype information was correct for this sample. Upon correction of the motif file (Supplemental Fig. 1), the haplotype was called correctly. Therefore, while the vast majority of sequences typically align correctly, users will need to confirm the reported information and make corrections to the motif file when encountering challenging haplotypes for the first time.

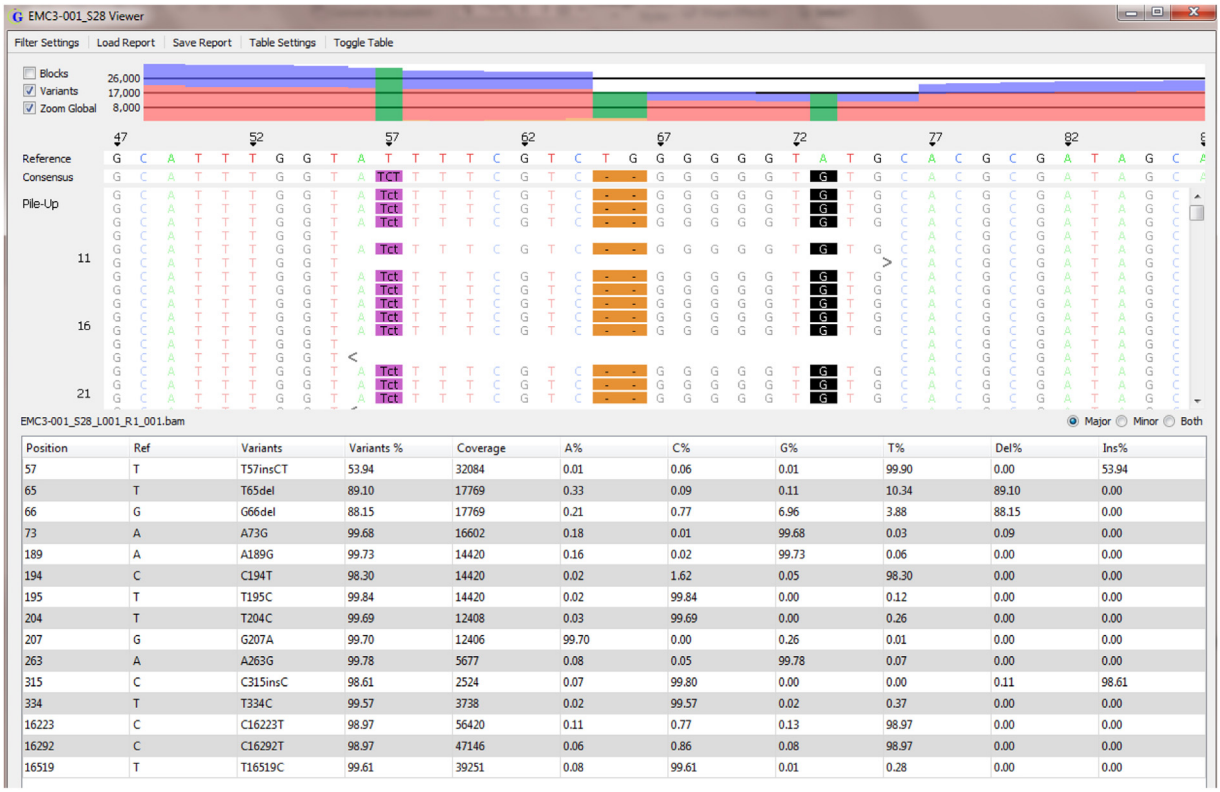
The other sample was flagged for a possible incorrect call at np 308 (TC insertion). This haplotype consisted of multiple possible SNPs and INDELs between np 303–310, with the software calling the profile as 308.1T and 308.2C. After manual inspection, the optimal alignment for this haplotype was determined to be 309T, 309.1C, and 309.2C. This motif was not listed in the motif file used for alignment, and demonstrates the utility of a modifiable motif file. Once the best alignment is determined, that motif can be added to the motif file and applied to subsequent alignments; which was done for these two samples. The haplotype table was modified to reflect the manually determined alignments for the two haplotypes and re-processed using the EMPOP tools. This produced a network with a root node containing 495 of the 500 haplotypes, and five nodes, effectively removing the two samples with multiple alignment possibilities that were previously flagged, as the haplotypes were now consistent with phylogenetic calls.

3.3. Identification of heteroplasmic variants

The improved alignment capabilities of GM-HTS reduces the amount of manual evaluation required for accurate identification of heteroplasmic sites. The identification of heteroplasmic variants was evaluated by comparing the minor allele report to the pattern of heteroplasmy as manually determined through evaluation of variant reports generated using the NextGENe® software. To be considered a match, the variant reports were required to have identical patterns of heteroplasmy. If the reports did not match, any inconsistencies were evaluated and the reason for exclusion determined. The pattern of heteroplasmy was concordant for 453 (~91%) of the 500 samples. The remaining 47 samples had inconsistencies resulting from differences in balance ratio calculations, Q-score filtering, and the increased C-stretch resolution provided by the GM-HTS software.

The inconsistencies in the pattern of heteroplasmy resulting from the balance ratio are due to differences in the way the software platforms calculate imbalance. NextGENe® evaluates the balance between variants identified in forward versus reverse reads and excludes the call if the imbalance exceeds 0.2 (1:4) in either direction. The new software also compares the balance of the minor allele identified in the forward to reverse reads, but the overall read balance at the given nucleotide position is also taken into consideration. For example, if the variant allele and the overall reads for a given nucleotide position are imbalanced (i.e., greater than 1:4), the minor allele is no longer excluded when the balance ratio is consistent with the ratio for the overall reads. This resulted in GM-HTS identifying 14 additional heteroplasmic sites [np 195, 215, 234 (x3), 237, 16,069, and 16,126, 16,183 (x6)] and excluding seven variants [np 16,183 (x3), 16,189 (x2), 16,320, and 16,355] identified using NextGENe®.

Based on Q-score, the GM-HTS software excluded two heteroplasmic positions that were reported using NextGENe®. The Q-score parameter is a new filter that evaluates the quality score of each base call and removes any variant exceeding the user-determined difference in quality (≥10 for this study). Both of these



Position	Ref	Variants	Variants %	Coverage	A%	C%	G%	T%	Del%	Ins%
57	T	T57insCT	53.94	32084	0.01	0.06	0.01	99.90	0.00	53.94
65	T	T65del	89.10	17769	0.33	0.09	0.11	10.34	89.10	0.00
66	G	G66del	88.15	17769	0.21	0.77	6.96	3.88	88.15	0.00
73	A	A73G	99.68	16602	0.18	0.01	99.68	0.03	0.09	0.00
189	A	A189G	99.73	14420	0.16	0.02	99.73	0.06	0.00	0.00
194	C	C194T	98.30	14420	0.02	1.62	0.05	98.30	0.00	0.00
195	T	T195C	99.84	14420	0.02	99.84	0.00	0.12	0.00	0.00
204	T	T204C	99.69	12408	0.03	99.69	0.00	0.26	0.00	0.00
207	G	G207A	99.70	12406	99.70	0.00	0.26	0.01	0.00	0.00
263	A	A263G	99.78	5677	0.08	0.05	99.78	0.07	0.00	0.00
315	C	C315insC	98.61	2524	0.07	99.80	0.00	0.00	0.11	98.61
334	T	T334C	99.57	3738	0.02	99.57	0.02	0.37	0.00	0.00
16223	C	C16223T	98.97	56420	0.11	0.77	0.13	98.97	0.00	0.00
16292	C	C16292T	98.97	47146	0.06	0.86	0.08	98.97	0.00	0.00
16519	T	T16519C	99.61	39251	0.08	99.61	0.01	0.28	0.00	0.00

Fig. 5. Sample EMC3-001 (European male older than 50 years of age) in the **Variant** nucleotide viewing mode with the **Major** table selected. The original motif table resulted in a haplotype between np 55–75 of 57CTins, 65del, 66del, and 73G, with the corrected file resulting in a phylogenetically correct haplotypes of 58C, 60.1T, 60.2T, 65del, 71del, and 73G.

positions were located in C-stretch regions (np 16,183 and 16,189). Finally, increased resolution in the C-stretch regions surrounding positions 310 and 16,189 resulted in GM-HTS identifying 24 additional heteroplasmic positions and excluding two positions identified using the NextGENe® software. The improved alignment removed two incorrectly identified sites in the HVI C-stretch that were reported by NextGENe® due to the misalignment of reads at positions 16,189. The correction of misaligned reads also resolved low-level heteroplasmy for 16 observations (~67%) at position 310 in HVII and eight observations (~33%) in HVI at positions 16,182, 16,183 (x2), 16,189 (x4), and 16,193.

4. Conclusions

The introduction of a massively parallel sequencing (MPS) approach for forensic mtDNA analysis will require the use of a software package that enables the examiner to easily navigate through the data, reliably report the findings, and use the outputs to effectively run database searches. GeneMarker® HTS (GM-HTS) has been developed to perform this function, and has been evaluated to ensure that it meets expectations. A data set containing 500 MPS-generated sequences yielded the correct mtDNA haplotypes, and heteroplasmic variants down to 2% were properly identified, both with minimal manual interpretation. The software addresses the primary needs of the forensic community: alignment to a circular version of the mtgenome, alignment of complex SNP and INDEL motifs, appropriate use of nomenclature, and the production of meaningful reports. The software offers numerous user-defined parameters for filtering the data that address the interests of researchers and practitioners, and provides multiple options for viewing and navigating through the data. Therefore, GM-HTS is a reliable software package for use in forensic mtDNA casework.

Competing interests

The authors have no competing interests regarding this research, including their relationship with SoftGenetics, Inc.

Fundings

This work was supported by the National Institute of Justice. Grant number 2014-DN-BX-K022, and Principal Investigator, Mitchell Holland.

Declaration of authorship

MMH and JAM contributed equally to experimental design, direction of the project, and writing of the manuscript. MMH was the architect of the motif/continuous read approach, while SoftGenetics was responsible for the software development. EP, JAM and MMH were responsible for testing iterations of the software.

Acknowledgements

The authors gratefully acknowledge the support of Jonathan Liu, John McGuigan, Michael Wiegand, and John Fosnacht from SoftGenetics for development of new versions of GeneMarker® HTS, and to Charity Holland for review of the manuscript. This work was supported in part by grant 2014-DN-BX-K022 from the National Institute of Justice (NIJ). The points of view in this

document are those of the authors and do not represent the official position or policies of the U.S. Department of Justice.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2017.01.016.

References

[1] S.A. Krieger, Why our justice system convicts innocent people, and the challenges faced by innocence projects trying to exonerate them, New Crim. Law Rev. 14 (2011) 333–402.
[2] E. Marchi, R.J. Pasacreta, Capillary electrophoresis in court: the landmark decision of the People of Tennessee versus Ware, J. Capill. Electrop. 4 (1997) 145–156.
[3] D. Hartman, O. Drummer, C. Eckhoff, et al., The contribution of DNA to the disaster victim identification (DVI) effort, Forensic Sci. Int. 205 (2011) 52–58.
[4] M.M. Holland, D.L. Fisher, L.G. Mitchell, et al., Mitochondrial DNA sequence analysis of human skeletal remains: identification of remains from the Vietnam War, J. Forensic Sci. 38 (1993) 542–553.
[5] P.L. Ivanov, M.J. Wadhams, R.K. Roby, et al., Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II, Nat. Genet. 12 (1996) 417–420.
[6] M.M. Holland, M.R. McQuillan, K.A. O'Hanlon, Second generation sequencing allows for mtDNA mixture deconvolution and high resolution detection of heteroplasmy, Croat. Med. J. 52 (2011) 299–313.
[7] J. McElhoe, M. Holland, K. Makova, et al., Development and assessment of an optimized next-generation DNA sequencing approach for the mtgenome using the Illumina MiSeq, Forensic Sci. Int. Gen. 13 (2014) 20–29.
[8] R.S. Just, J.A. Irwin, W. Parson, Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing, Forensic Sci. Int. Genet. 18 (2015) 131–139.
[9] C. Hahn, L. Bachmann, B. Chevreux, Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads —a baiting and iterative mapping approach, Nucleic Acids Res. 41 (2013) 1–9 (e129).
[10] B. Rebolledo-Jaramillo, M.S.-W. Su, N. Stoler, et al., Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 15474–15479.
[11] C. Calabrese, D. Simone, M.A. Diroma, et al., MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing, Bioinformatics 30 (2014) 3115–3117.
[12] Y. Guo, J. Li, C. Li, et al., MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis, Bioinformatics 29 (2013) 1210–1211.
[13] W. Parson, G. Huber, L. Moreno, et al., Massively parallel sequencing of complete mitochondrial genomes from hair shaft samples, Forensic Sci. Int. Gen. 15 (2015) 8–15.
[14] M.A. Peck, M.D. Brandhagen, C. Marshall, et al., Concordance and reproducibility of a next generation mtGenome sequencing method for high-quality samples using the Illumina MiSeq, Forensic Sci. Int. Gen. 24 (2016) 10–11.
[15] J.L. King, A. Sanjantila, B. Budowle, mitoSAVE: Mitochondrial sequence analysis of variants in Excel, Forensic Sci. Int. Gen. 12 (2014) 122–125.
[16] W. Parson, L. Gusmao, D.R. Hares, et al., DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing, Forensic Sci. Int. Gen. 13 (2014) 134–142.
[17] H.-J. Bandelt, W. Parson, Consistent treatment of length variants in the human mtDNA control region: a reappraisal, Int. J. Leg. Med. 122 (2008) 11–21.
[18] D. Dacheva, R. Dodova, I. Popov, et al., Validation of an NGS approach for diagnostic BRCA1/BRCA2 mutation testing, Mol. Diagn. Ther. 19 (2015) 119–130.
[19] G. Millat, V. Chanavat, R. Rousson, Evaluation of a new NGS method based on a custom AmpliSeq library and Ion Torrent PGM sequencing for the fast detection of genetic variations in cardiomyopathies, Clin. Chim. Acta 433 (2014) 266–271.
[20] T.A. Sivakumaran, A. Husami, D. Kissell, et al., Performance evaluation of the next-generation sequencing approach for molecular diagnosis of hereditary hearing loss, Otolaryngol. Head Neck Surg. 148 (2013) 1007–1016.
[21] M. Holland, J. McElhoe, A custom software solution for forensic mtDNA analysis of MiSeq data, Forensic Sci. Int. Gen. 5 (2015) e614–e616.
[22] S. Anderson, A.T. Bankier, B.G. Barrell, et al., Sequence and organization of the human mitochondrial genome, Nat. Genet. 290 (1981) 457–465.
[23] R.M. Andrews, I. Kubacka, P.F. Chinnery, et al., Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA, Nat. Genet. 23 (1999) 147.
[24] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.
[25] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (2009) 1754–1760.
[26] A.B. Brandstätter, H. Niederstätter, M. Pavlic, et al., Generating population data for the EMPOP database —an overview of the mtDNA sequencing and data evaluation processes considering 273 Austrian control region sequences as example, Forensic Sci. Int. 166 (2007) 164–175.
[27] B. Zimmermann, A.W. Röck, W. Parson, Improved visability of quasi-median networks with the EMPOP NETWORK software, Croat. Med. J. 55 (2) (2014) 15–120.
[28] P. Jaccard, The distribution of the flora in the alpine zone, New Phytol. 11 (2) (1912) 7–50.