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.

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/

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