MitoSuite: a graphical tool for human mitochondrial genome profiling in massive parallel sequencing

Koji Ishiya and Shintaroh Ueda

Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo, Japan

ABSTRACT

Recent rapid advances in high-throughput, next-generation sequencing (NGS) technologies have promoted mitochondrial genome studies in the fields of human evolution, medical genetics, and forensic casework. However, scientists unfamiliar with computer programming often find it difficult to handle the massive volumes of data that are generated by NGS. To address this limitation, we developed MitoSuite, a user-friendly graphical tool for analysis of data from high-throughput sequencing of the human mitochondrial genome. MitoSuite generates a visual report on NGS data with simple mouse operations. Moreover, it analyzes high-coverage sequencing data but runs on a stand-alone computer, without the need for file upload. Therefore, MitoSuite offers outstanding usability for handling massive NGS data, and is ideal for evolutionary, clinical, and forensic studies on the human mitochondrial genome variations. It is freely available for download from the website https://mitosuite.com.

Subjects Genetics, Genomics, Molecular Biology

Keywords MitoSuite, Mitochondrial genome, Next-generation sequencing, Graphical tool, Massively parallel sequencing

Submitted 16 December 2016
Accepted 11 May 2017
Published 30 May 2017

Corresponding author
Koji Ishiya, mitosuite@gmail.com

Academic editor
Keith Crandall

Additional Information and Declarations can be found on page 11

DOI 10.7717/peerj.3406

INTRODUCTION

The human mitochondrial (mt) genome encodes important information that governs the development of various diseases (*Taylor & Turnbull, 2005*). It also reflects maternal lineage (*Mishmar et al., 2003*; *Macaulay et al., 2005*; *Behar et al., 2008*) and evolutionary history (*Cann, Stoneking & Wilson, 1987*; *Torroni et al., 2006*; *Underhill & Kivisild, 2007*). High-throughput, next-generation sequencing (NGS) technologies allow more rapid sequencing of a larger number of samples than does traditional capillary sequencing based on *Sanger, Nicklen & Coulson (1977)* method. NGS technologies also allow whole genome sequencing, exon sequencing, and gene expression profiling at high speeds and low costs (*Metzker, 2010*). The advent of these high-throughput technologies has led to a dramatic improvement in studies on the human mitochondrial genome. For instance, NGS has aided the discovery of variants and heteroplasmic mutation in the human mitochondrial genome (*Tang & Huang, 2010*). In addition, NGS data can help estimate the probability of exogenous DNA sources in forensic samples (*Just, Irwin & Parson, 2015*).

Consequently, the demand for advanced tools for analyzing the massive volume of data that NGS generates has also increased. Currently, there are several command-line tools available to analyze high-throughput sequencing data for the mitochondrial genome.

How to cite this article Ishiya and Ueda (2017), MitoSuite: a graphical tool for human mitochondrial genome profiling in massive parallel sequencing. **PeerJ 5:e3406; DOI 10.7717/peerj.3406**

MitoSeek (Guo et al., 2013) is one such character-based tool that provides information on mtDNA copy numbers, and alignment quality, somatic annotations, and structural variants of the mitochondrial genome. MToolBox (Calabrese et al., 2014) is another bioinformatics pipeline for analyzing mitochondrial genome data from NGS platforms, with functions similar to those of MitoSeek. There are also several web-based tools to analyze such data. For instance, MitoBamAnnotator (Zhidkov et al., 2011) assesses the functional potential of heteroplasmy. mit-o-matic (Vellarikkal et al., 2015) is another web-based pipeline for clinical annotations of mtDNA variants. However, these tools have some limitations with regard to uploading files on their servers. For example, the maximum file size that can be uploaded to mit-o-matic is restricted to less than 25 MB. To address this issue, in this study, we developed MitoSuite, a stand-alone tool which does not involve file-uploading procedures like web-based tools. The "uploading-free" process offers advantage for shortening actual run time, since it eliminates file-uploading and queue times required to begin analysis. Furthermore, the uploading-free platform is suitable for leakage prevention of personal genome data in clinical or forensic cases. MitoSuite also provides a graphical user interface (GUI), which offers user-friendly operability for researchers who are unfamiliar with the command-line interface. Our tool comprehensively supports quality check of alignment data, variant annotation, building consensus sequences, haplogroup classification, and detection of heteroplasmic sites, exogenous contamination, and base-substitution patterns for mitochondrial genome data obtained by high-throughput sequencing. The output summary is provided in the HTML format, which can be easily visualized using a web browser, without complicated programming processes. To our knowledge, MitoSuite is the first standalone, GUI software for comprehensive profiling of the mitochondrial genome, using high-throughput sequencing data with intuitive operability.

MATERIAL AND METHODS

Format for input data

MitoSuite supports the BAM format, a binary version of Sequence Alignment/Map (SAM), which is a tab-delimited text format for high-throughput sequencing alignment (Li et al., 2009). Since the genome size of mitochondria is small (approximately 17 kb), it is easy to manipulate the mitochondrial genome in simple text files such as those in FASTA format. However, because FASTA files do not contain information on either sequencing quality or alignment processes, it is difficult to detect problems with base-call errors or contamination, using sequence data in the FASTA format. In contrast, BAM files contain alignment conditions or base substitutions at each position of the mitochondrial genome, as well as reads of high-throughput sequencing. By using BAM files, MitoSuite can not only check mapping and sequencing quality, but can also detect mismatches potentially attributed to exogenous contamination, sequencing errors, or heteroplasmy. The input file is mtDNA alignment data, a BAM file mapped against a reference sequence of the human mitochondrial genome. MitoSuite supports multiple human mitochondrial reference sequences, including not only rCRS (Andrews et al., 1999), but also RSRS (Behar et al., 2012), chrM in hg19, chrMT in GRCh37, and chrMT in GRCh38.

Test datasets

We used seven sets of empirical sequencing data (NA11920, HG01112, NA18941, HG00096, HG00273, NA18548, NA18510) of 1000 genomes project data (*1000 Genomes Project Consortium, 2012*) to evaluate the performance of MitoSuite for high-coverage sequencing data, as well as empirical ancient sequencing data of an ancient hunter-gatherer (*Olalde et al., 2014*) to examine whether this tool can detect ancient DNA profiles. These empirical data (BAM file) were converted to FastQ files by the SamToFastQ command in Picard tools (http://broadinstitute.github.io/picard), and then realigned to the human mitochondrial reference sequence rCRS, using the Burrows-Wheeler Aligner (BWA) (*Li & Durbin, 2009*). After the realignments, duplicated reads were removed from the BAM files by the MarkDuplicates command in Picard tools. Sequence reads for the ancient hunter-gatherer were aligned against rCRS, and duplicated reads were removed in the same way. Next, to check the accuracy of mitochondrial haplogroup assignment, we generated simulated NGS reads using 324 worldwide mitochondrial genome sequences (Table S1) with ART, a simulation tool to generate synthetic NGS reads (*Huang et al., 2012*). These sequence sets were selected from PhyloTree (*Van Oven & Kayser, 2009*) (http://www.phylotree.org), and included all known macro-haplogroups in nearly equal proportions. We obtained GenBank accession numbers (https://www.ncbi.nlm.nih.gov/genbank/) from the sub tree pages on PhyloTree's site (e.g., http://www.phylotree.org/tree/L0.htm), and then downloaded FASTA files from GenBank, based on the accession numbers obtained with our in-house Python scripts. Based on the Illumina sequencer model in ART (ver.03.19.2015), we assumed 1% sequencing error, single-end 100-base reads, and average depth of 1–1000× in the simulated data. We aligned these simulated reads against rCRS using BWA, and then used these BAM files as simulation datasets.

File processing

First, MitoSuite parses a BAM file and extracts reads together with the alignment condition involved with file headers, read groups, and reference sequences. Next, this tool automatically calculates summary statistics, including the depth of coverage, GC-content, base-call quality, mapping quality, and read length. These are important statistics for the quality control of NGS data. Moreover, this tool directly estimates mitochondrial haplogroups from a BAM file. MitoSuite does not require file format conversion (e.g., BAM > FASTA, BAM > VCF) and can directly assign mitochondrial haplogroups, based on the haplogroup-defining sites of PhyloTree. Figure 1 shows the schema for the file processing that can seamlessly work as an all-in-one tool.

Detection of heteroplasmic sites

MitoSuite can also detect heteroplasmic sites that may be attributed to exogenous contamination, sequencing errors, amplification errors, or heteroplasmy. Our tool outputs a list of heteroplasmic positions with frequency greater than the minor allele frequency (MAF), which represents the frequency of inconsistent bases with the consensus sequence. MAF, a threshold for the detection of heteroplasmic sites, is given as follows:

$$\mathrm{MAF} = N_{\mathrm{diff}}/N_{\mathrm{con}}$$

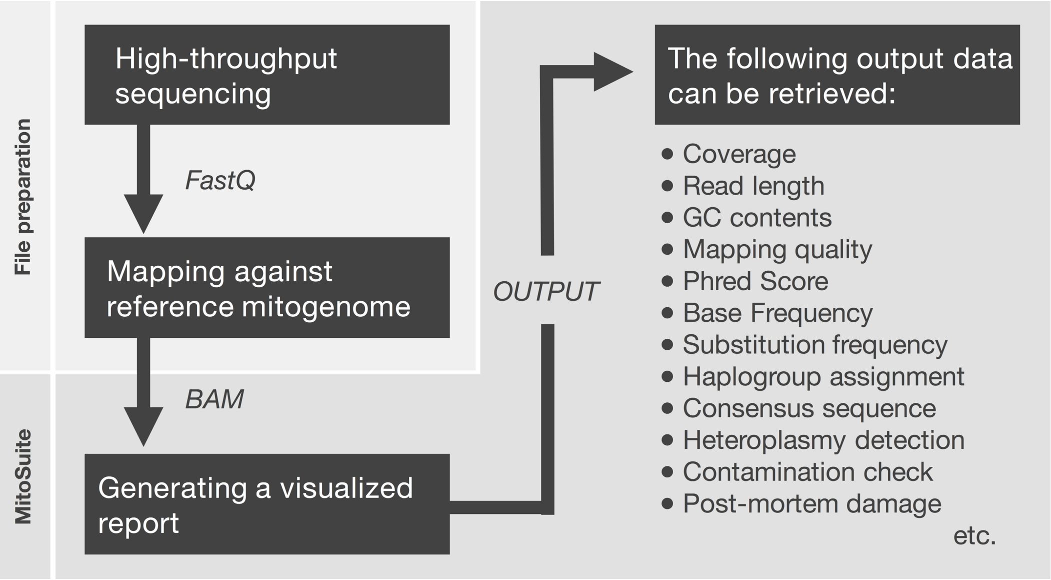


Figure 1 **File preparation and processing flow for MitoSuite.** BAM, binary version of SequenceAlignment/Map (SAM) format; FastQ, the format storing sequences and base call qualities. FastQ files are mapped against the mitochondrial genome by a mapper tool (e.g., Burrows-Wheeler Aligner). After file preparation, the BAM file is used as an input file for MitoSuite. Post-mortem damage means a series of DNA damage that occurs after biological death. These damages are observed as base substitution of C to T (G to A) and fragmentation of the DNA sequence (*Briggs et al., 2007*). Phred score is a quality value when a sequencer calls bases. This value is defined as phred score = $-10 \log 10\ P_{\text{error}}$, where P_{error} is the probability of a sequencing error.

where N_{diff} and N_{con} are the number of bases different from and identical to the consensus sequence, respectively. This means that MAF can be used as a threshold for the detection of heteroplasmic sites. MitoSuite also verifies the mitochondrial genome assembly by calculating the percentage of supporting bases of the consensus sequences in a BAM file (Fig. S1A). This percentage (P_{support}) is computed as follows:

$$P_{\text{support}} = (N_{\text{agree}}/N_{\text{depth}}) \times 100$$

where N_{agree} is the number of bases concordant with the assembled consensus sequence at each site, and N_{depth} is the depth of coverage at each site. This percentage provides clues to find unexpected contaminated sites (Fig. S1B).

Optional functions

MitoSuite also provides a few optional functions to meet user needs for other data profiles. These functions can be accessed by selecting the relevant option menus. The optional 'Annotation of disease-related variants' function provides an annotation list of disease-associated mutations, based on the list of reported mitochondrial DNA base substitution diseases at MITOMAP (*Kogelnik et al., 1996*) (http://www.mitomap.org/MITOMAP). To accommodate private and local genetic data in medical and forensic cases, MitoSuite also supports customizable polymorphic databases. Customizable annotation information is required to correspond to the position of rCRS. The annotation file is a common comma-delimited

CSV format containing two items: a mutation allele with a genomic position corresponding to that of rCRS (e.g., C150T, A4282G), and related information (e.g., related-disease name) in each designated column. The template of the annotation file is available from MitoSuite's support page or can be downloaded by the installer. In this option, MUSCLE (*Edgar, 2004*) program is used to realign a consensus sequence against a reference sequence because MitoSuite finally takes the positional consistency of the obtained consensus sequence against the reference sequence (rCRS). Our tool can also calculate the percentage of each base substitution relative to the reference sequence in the total mapped reads, and then provide a pie chart showing the proportion of each base substitution. This chart will help users find locally biased substitution patterns in total mapped reads. Biased substitution patterns are often caused by the sample or experimental conditions, rather than by the natural process of mutations. For instance, deamination of cytosine to uracil, a postmortem hydrolytic change, often occurs in ancient DNA (*Briggs et al., 2007*). With the optional 'Ancient DNA checker' function, MitoSuite can detect postmortem damages and calculate the percentage of bases inconsistent with the haplogroup-defining variants to estimate exogenous contamination. This percentage (P_{mismatch}) is computed as follows:

$$P_{\mathrm{mismatch}} = \left[\sum (i = 1 \to k)\{(N_{\mathrm{mismatch}}/N_{\mathrm{match}})\}/k\right] \times 100$$

where k is the total number of haplogroup-defining sites, N_{mismatch} is the number of bases inconsistent with the defining variant, and N_{match} is the number of bases consistent with the defining variant.

Software availability

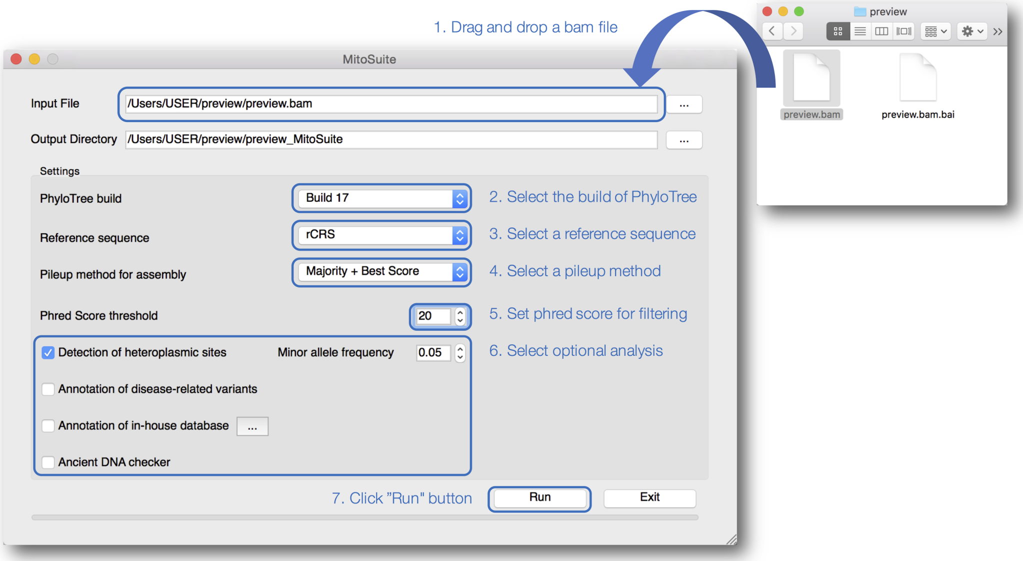
MitoSuite is freely available from https://mitosuite.com. Our tool mainly supports UNIX-like operating systems (OS) such as Mac OSX and Linux. The tool can also run on Linux-like environments (e.g., Cygwin) for Windows OS. MitoSuite for Mac OSX also provides the graphical installer package. This package can perform automatic installation without any command-line operations by the user. Installation instructions, tutorial movies, and additional technical support for MitoSuite are provided at https://mitosuite.com.

RESULTS AND DISCUSSION

Our tool is designed for better usability, especially for non-bioinformaticians unfamiliar with typing complicated commands (Fig. 2). MitoSuite provides a drag-and-drop functionality for loading a BAM file and automatically displays an output destination directory. Our tool supports the latest build 17 and the previous build 16 of PhyloTree for the haplogroup assignment, and five available human mitochondrial reference sequences (rCRS, RSRS, hg19, GRCh37, and 38). MitoSuite also provides three options (Majority, Best Score, and Majority + Best Score) for calling a consensus base at each site. The "Majority" option decides the base by the majority in counting based rules. Thus, under this option, the most-read base at the site is adopted as a consensus base. For example, when counting only bases with a phred score higher than 30 defined as a base-call quality value at a site (when the base call threshold is set to 30), where the read depth of base "A" is 8 and that of base "T" is 2, the base "A" is adopted as a consensus base at a site. To avoid calling uncertain bases as much as possible, MitoSuite adopts "N" as the consensus base when multiple bases have the same read depth (e.g., A = 8, T = 8). The "Best Score" option decides the base with the highest basecall quality (phred score) at each site. The phred score is defined as the quality value when a sequencer calls bases. This option determines a consensus base at a site, using only the value of the Phred Score, regardless of the read depth. For example, when considering only bases with a phred score higher than 30 at a site, where the highest phred score of base "A" is 31, that of base "T" is 33, that of base "G" is 30, and that of "C" is 30, then the base "T" is adopted as the consensus base at the site even if base "T" is read less frequently than the other bases. Since this option does not take the read depth into consideration, it can also be applied to sites where bases cannot be determined by the majority option. However, this option adopts "N" as a consensus base at a site when multiple bases have the same maximum phred score. The "Majority + Best Score" option incorporates the "Best Score" with the "Majority" option at each site. This option firstly decides a base at each site by placing priority on majority rule, and then remaining sites that are not decided under majority rule are called by the "Best Score" option. For example, even if base "A" and base "T" have the same read depth at a site, base "A" will be adopted as a consensus base if it has the highest phred score. It is also necessary to set a threshold value for the phred score because MitoSuite adopts only bases with a phred score greater than the threshold values set by users to performs mtGenome assembly as well as haplogroup assignment, and detection of heteroplasmic sites. A consensus sequence is built on based on these basecall conditions and outputted as a FASTA file. MitoSuite also provides optional functions for detection of heteroplasmic sites, disease association, and ancient DNA according to their own purposes. The results of these analyses are finally outputted as a single html file (Fig. 3).

MitoSuite outputs analytical results for the quality of alignment data and genetic profiles that are haplogroup and polymorphisms on the mitochondrial genome (Fig. 3, Figs. S2–S4). As the results are provided in HTML format, they can be easily viewed on a web browser without depending on specific computer environments. The main output items are as follows: (1) A summary statistics table, including categories on data quality and genetic profiles (Fig. S2); this table shows an overview of the NGS data. Interactive dynamic charts for the mitochondrial genome operate with zoom and pan functions, which are useful for users to view the depth of the NGS data across mitochondrial genome (Fig. S3). (2) Figure S4 shows the other output data. All the data are saved in their respective output folders, and it is possible to individually access them. The distribution of read length, GC-content, and mapping quality are provided as histograms. Further, "retrieval" and "sort" functions in the data tables allow access to each item. These tables can be used for a quick check of quality as well as mutations at a desired destination site. MitoSuite can also automatically build a consensus sequence in the FASTA format from a BAM file, from which the phylogenetics or population genetics of the sequence can then be easily analyzed.

MitoSuite can be run even for deep sequencing data in a stand-alone environment. Here, we used seven high-throughput sequencing data from the 1000 Genomes Project as test data sets. Some of the datasets surpassed 1000× depth of coverage, and our tool was easily able to analyze these ultra-deep data. Analysis of a sample dataset with MitoSuite is shown



1. Drag and drop a bam file

MitoSuite

Input File: /Users/USER/preview/preview.bam

Output Directory: /Users/USER/preview/preview_MitoSuite

Settings

PhyloTree build: Build 17

Reference sequence: rCRS

Pileup method for assembly: Majority + Best Score

Phred Score threshold: 20

☑ Detection of heteroplasmic sites — Minor allele frequency: 0.05

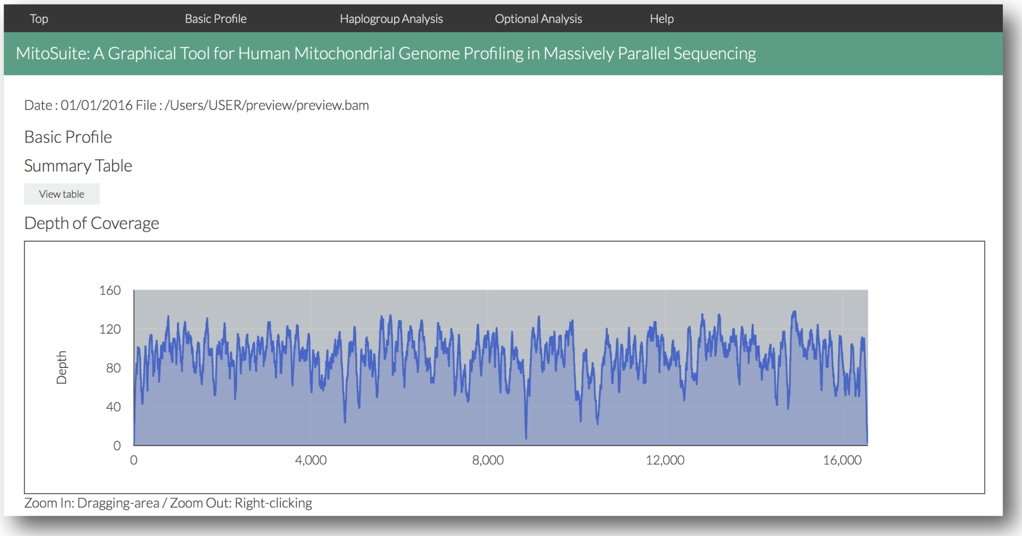
□ Annotation of disease-related variants

□ Annotation of in-house database ...

□ Ancient DNA checker

2. Select the build of PhyloTree
3. Select a reference sequence
4. Select a pileup method
5. Set phred score for filtering
6. Select optional analysis
7. Click "Run" button

Figure 2 **A screenshot of the graphical user interface of MitoSuite.** At the beginning of analysis, users can drop and drag the input file (.bam) and click “Run” after setting the optional parameters (“Detection of heteroplasmic sites” shown selected). Bullet points 1–7 point out the protocol step-by-step.



MitoSuite: A Graphical Tool for Human Mitochondrial Genome Profiling in Massively Parallel Sequencing

Date : 01/01/2016 File : /Users/USER/preview/preview.bam

Basic Profile

Summary Table

Depth of Coverage

Zoom In: Dragging-area / Zoom Out: Right-clicking

Figure 3 **A screenshot of the visualized outputs of MitoSuite.** The top page in an output file (results.html) is shown. After completion of analysis, users can quickly access the detailed information by clicking the link menu in the output page.

in Table 1. It takes about 1 min to analyze 1000× ultra-deep data at the default settings, on a desktop computer equipped with a 3.5-GHz processor and 16-GB RAM. The time required is mainly for read operation, since the tool works in a stand-alone environment. MitoSuite also successfully detected the fragmentation and deamination pattern of ancient DNA-like on empirical reads from *Olalde et al. (2014)* (Figs. S4C, S4E). In addition, the

Table 1 **Sample dataset analyzed with MitoSuite.**

Sample	Age	Haplogroup	Depth (avg)	Run (min)	Reference
NA11920	Modern	H1a1a1	1,516	1.5	1000 Genomes Project
HG01112	Modern	A2ac1	1,504	1.4	1000 Genomes Project
NA18941	Modern	N9b1a	1,297	1.2	1000 Genomes Project
HG00096	Modern	H16a1	1,212	1.1	1000 Genomes Project
HG00273	Modern	U5b1b2a	1,117	1.1	1000 Genomes Project
NA18548	Modern	C4a1b	1,021	1.1	1000 Genomes Project
NA18510	Modern	L0a1a3	746	0.8	1000 Genomes Project
La Brãna1	7,000 BP	U5b2c1	93	0.2	*Olalde et al. (2014)*

Notes.
avg, average.

most likely haplogroup estimated by our tool is “U5b2c1” that is consistent with reported one in *Olalde et al. (2014)* (Table 1).

We also validated the accuracy of mitochondrial haplogroup assignment, using simulated NGS reads including data from all macro-haplogroups. The accuracy of haplogroup assignment is computed as follows: TP/(TP + FP). True positive (TP) is the number of haplogroups predicted and validated. False positive (FP) is the number of haplogroups predicted but failed in validation. The haplogroup assignment accuracy for 5×, 10×, 50×, 150×, and 200× fold coverage sets were 0.969, 0.991, 0.994, 0.994 and 0.997, respectively, assuming 1% base-call error (Fig. 4).

To examine whether our tool can detect heteroplasmic sites previously reported in empirical high-throughput sequencing data, we used MPS raw read data from *Avital et al. (2012)*. We set MAF > 10% as the detection threshold after performing quality control analyses, including trimming of duplicates and low-quality bases (Phred Score < 20). Consequently, MitoSuite detected 14 out of 15 heteroplasmic sites with MAF > 10% in *Avital et al. (2012)* (Table S2). We think that the differences in quality control procedures and mapping tools between *Avital et al. (2012)* and this study may have changed heteroplasmic fraction in the alignment data.

Useful bioinformatics tools for various mtDNA studies have been developed and are currently available to researchers. Mitochondrial haplogroup is an important genetic profile for molecular anthropological and forensic genetic investigations, and most available mtDNA tools support haplogroup assignment for various data formats. MitoTool (*Fan & Yao, 2013*), mtDNAmanager (*Lee et al., 2008*), and HAPLOFIND (*Vianello et al., 2013*) can estimate haplogroup, using the FASTA format or a text-based format containing variant information against a reference sequence. HaploGrep2 (*Weissensteiner et al., 2016a*) also supports the VCF format storing DNA polymorphism data, as well as the two above-mentioned formats. mtDNA-Server (*Weissensteiner et al., 2016b*), MToolBox, mit-o-matic, Phy-Mer (*Navarro-Gomez et al., 2015*), and MitoSuite can manipulate massively parallel sequencing (MPS) data such as the FASTQ or BAM (SAM) formats for haplogroup classification. MToolBox, MitoSeek, MitoBamAnnotator, mtDNA-Server, mit-o-matic, and MitoSuite can annotate variants in high-throughput sequencing data. The detection of

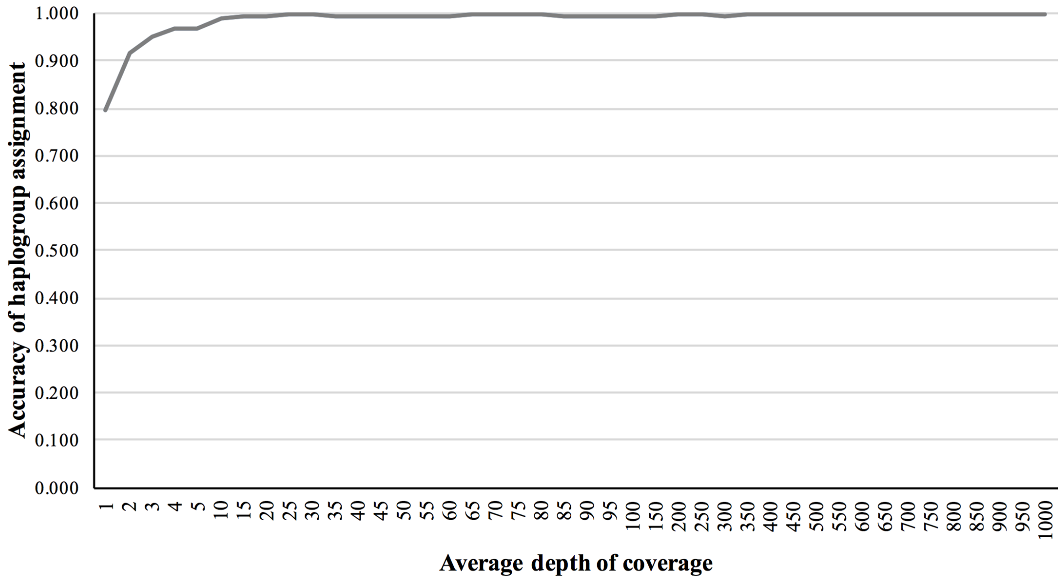


Figure 4 Haplogroup assignment accuracy of MitoSuite for simulated NGS reads generated from 324 worldwide mitochondrial genome sequences.

heteroplasmy from a single individual or tissue provides useful information in clinical or forensic cases. MToolBox, MitoSeek, MitoBamAnnotator, mtDNA-Server, mito-o-matic, and MitoSuite can also report possible point heteroplasmy (PHP), based on detection parameters (e.g., minor allele frequency; MAF) set by users.

Users can select a suitable tool that takes into consideration their application, computational environment, data size, and bioinformatics skills. Command-line tools such as MitoSeek or MitoToolBox have the advantage of flexible incorporation into customizable NGS pipelines, but their setup is still difficult for non-bioinformaticians. MitoSeek is a useful tool for the detection of structural variants or somatic mutations, but its use requires installation of Circos, which is a software package for genomic data visualization (*Krzywinski et al., 2009*). This in turn requires users to install several dependent Perl modules based on their computational environment (e.g., operating system), as well as to set the local path to executable files. Command-line tools sometimes change their command specifications when updating the version. Therefore, users need to appropriately manage the version of dependent command-line tools for proper functioning of the pipeline and set an environment path in the local host, because the pipelines contain several command-line tools, including mapper or variant callers (e.g., BWA, Picard tools, GATK *McKenna et al., 2010*). Web-based tools such as mit-o-matic or mtDNA-Server do not require complicated installation processes, and provide a system that is easy to use. However, there are still unavoidable issues that include file size limits or queue times required to start analysis on the web-server. In addition, it is often necessary to assign an email address or individual account to manage uploaded data, and few servers clarify what technology is being used in the background of the management system. Users thus need to trust the server-side management system. Indeed, the mitochondrial genome is widely used in medical and forensic fields, and thus

Table 2 The list of available bioinformatics tools for mtDNA analysis.

	MitoSuite	HaploGrep2	mtDNA-Server	mtDNAprofiler	MToolBOX	MitoSeek
Input file	BAM	FASTA, hsd, VCF	FastQ, BAM, VCF	FASTA, variant[a]	FastQ, BAM, SAM	BAM
User interface	GUI	Web	Web	Web	CUI	CUI
Supported reference sequence	rCRS, RSRS, hg19[c], GRCh[c]	rCRS, RSRS	rCRS, RSRS	rCRS	rCRS, RSRS	rCRS, hg19[c]
Automatic installation	✓	–	–	–	–	–
File upload	–	✓	✓	✓	–	–
Haplogroup assignment	✓	✓	✓	–	✓	–
mtGenome assembly	✓	–	✓	✓	✓	✓
Coverage plot	✓	–	✓	–	–	–
Quality check	✓	–	✓	–	–	✓
Concordance check	✓	–	–	✓	–	–
Damage check	✓	–	–	–	–	–
Contamination check	✓	–	✓	–	–	–
Relative copy number	–	–	–	–	–	✓
Variant annotation	✓	–	✓	–	✓	✓
Structural variants detection	–	–	–	–	–	✓
Somatic mutation detection	–	–	–	–	–	✓
Heteroplasmy detection	✓	–	✓	–	✓	✓

	MitoBamAnnotator	mtDNAmanager	MitoTool	HAPLOFIND	mit-o-matic	Phy-Mer
Input file	BAM	variant[a]	FASTA, variant[a]	FASTA	FastQ, pileup[b]	FASTA, FastQ, BAM
User interface	Web	Web	Web/GUI	Web	Web/CUI	CUI
Supported reference sequence	rCRS	rCRS	rCRS, RSRS	rCRS, RSRS	rCRS	reference-independent
Automatic installation	–	–	–	–	–	–
File upload	✓	✓	✓	✓	✓	–
Haplogroup assignment	–	✓	✓	✓	✓	✓
mtGenome assembly	✓	–	–	–	✓	✓
Coverage plot	–	–	–	–	–	–
Quality check	–	–	–	–	–	–
Concordance check	–	–	–	–	–	–
Damage check	–	–	–	–	–	–
Contamination check	–	–	–	–	–	–
Relative copy number	–	–	–	–	–	–
Variant annotation	✓	–	✓	✓	✓	–
Structural variants detection	–	–	–	–	–	–
Somatic mutation detection	–	–	–	–	–	–
Heteroplasmy detection	✓	–	–	–	✓	–

Notes.

✓, available; –, not-available.

[a] A text-based format for describing SNPs information against a reference seuquence.

[b] A text-based format for describing the base-pair information of the reads against a reference sequence.

[c] Support for the mitochondrial sequence (chrM/chrMT) on hg19, GRCh37 and 38.

analysis environments must be very restrictive in terms of security systems. Our tool provides a server-independent, stand-alone system that brings advantages especially to medical and forensic researchers in terms of security. MitoSuite also provides a graphical user interface with intuitive operability, in addition to a graphical report on quality of alignment data, variant annotation, building of consensus sequences, haplogroup classification, detection of heteroplasmic sites and exogenous contamination, damage detection, and interactive dynamic graphics across the complete mitochondrial genome in NGS data (Table 2). MitoSuite for Mac OSX also provides an easy-to-use automated installer. Therefore, it provides a user-friendly solution for many investigators unfamiliar with advanced information-processing techniques. We expect our tool to promote human mitochondrial genome studies in the fields of anthropological science, forensic casework, and medicine.

ACKNOWLEDGEMENTS

The discussions with and cooperation of our lab members have greatly contributed to improving this tool.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The work was supported in part by Grant-in-Aid for Japan Society for the Promotion of Science (JSPS) Fellows, 15J08490 (to KI), and JSPS KAKENHI Grant Number 25291104 (to SU). There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Japan Society for the Promotion of Science (JSPS) Fellows: 15J08490.
JSPS KAKENHI: 25291104.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Koji Ishiya conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Shintaroh Ueda reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:

We used seven sets of empirical sequencing data (NA11920, HG01112, NA18941, HG00096, HG00273, NA18548, NA18510) of 1000 Genomes Project data (*1000 Genomes Project Consortium, 2012*).

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.3406#supplemental-information.

REFERENCES

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics* **23**(**2**):147 DOI 10.1038/13779.

Avital G, Buchshtav M, Zhidkov I, Tuval Feder J, Dadon S, Rubin E, Glass D, Spector TD, Mishmar D. 2012. Mitochondrial DNA heteroplasmy in diabetes and normal adults: role of acquired and inherited mutational patterns in twins. *Human Molecular Genetics* **21**(**19**):4214–4224 DOI 10.1093/hmg/dds245.

Behar DM, Van Oven M, Rosset S, Metspalu M, Loogväli EL, Silva NM, Kivisild T, Torroni A, Villems R. 2012. A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *American Journal of Human Genetics* **90**(**4**):675–684 DOI 10.1016/j.ajhg.2012.03.002.

Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, Makkan H, Tzur S, Comas D, Bertranpetit J, Quintana-Murci L, Tyler-Smith C, Wells RS, Rosset S. 2008. The dawn of human matrilineal diversity. *American Journal of Human Genetics* **82**(**5**):1130–1140 DOI 10.1016/j.ajhg.2008.04.002.

Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America* **104**(**37**):14616–14621 DOI 10.1073/pnas.0704665104.

Calabrese C, Simone D, Diroma MA, Santorsola M, Guttà C, Gasparre G, Picardi E, Pesole G, Attimonelli M. 2014. MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics* **30**(**21**):3115–3117 DOI 10.1093/bioinformatics/btu483.

Cann RL, Stoneking M, Wilson AC. 1987. Mitochondrial DNA and human evolution. *Nature* **325**(**6099**):31–36 DOI 10.1038/325031a0.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**(**5**):1792–1797 DOI 10.1093/nar/gkh340.

Fan L, Yao YG. 2013. An update to MitoTool: using a new scoring system for faster mtDNA haplogroup determination. *Mitochondrion* **13**(**4**):360–363 DOI 10.1016/j.mito.2013.04.011.

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(**7422**):56–65 DOI 10.1038/nature11632.

Guo Y, Li J, Li CI, Shyr Y, Samuels DC. 2013. MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics* **29**(**9**):1210–1211 DOI 10.1093/bioinformatics/btt118.

Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. *Bioinformatics* **28(4)**:593–594 DOI 10.1093/bioinformatics/btr708.

Just RS, Irwin JA, Parson W. 2015. Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. *Forensic Science International Genetics* **18**:131–139 DOI 10.1016/j.fsigen.2015.05.003.

Kogelnik AM, Lott MT, Brown MD, Navathe SB, Wallace DC. 1996. MITOMAP: a human mitochondrial genome database. *Nucleic Acids Research* **24(1)**:177–179 DOI 10.1093/nar/24.1.177.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Research* **19(9)**:1639–1645 DOI 10.1101/gr.092759.109.

Lee HY, Song I, Ha E, Cho SB, Yang WI, Shin KJ. 2008. mtDNAmanager: a Web-based tool for the management and quality analysis of mitochondrial DNA control-region sequences. *BMC Bioinformatics* **9**:483 DOI 10.1186/1471-2105-9-483.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25(14)**:1754–1760 DOI 10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25(16)**:2078–2079 DOI 10.1093/bioinformatics/btp352.

Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, Taha A, Shaari NK, Raja JM, Ismail P, Zainuddin Z, Goodwin W, Bulbeck D, Bandelt HJ, Oppenheimer S, Torroni A, Richards M. 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* **308(5724)**:1034–1036 DOI 10.1126/science.1109792.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20(9)**:1297–1303 DOI 10.1101/gr.107524.110.

Metzker ML. 2010. Sequencing technologies—the next generation. *Nature Review Genetics* **11(1)**:31–46 DOI 10.1038/nrg2626.

Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, Wallace DC. 2003. Natural selection shaped regional mtDNA variation in humans. *Proceedings of the National Academy of Sciences of the United States of America* **100(1)**:171–176 DOI 10.1073/pnas.0136972100.

Navarro-Gomez D, Leipzig J, Shen L, Lott M, Stassen AP, Wallace DC, Wiggs JL, Falk MJ, Van Oven M, Gai X. 2015. Phy-Mer: a novel alignment-free and reference-independent mitochondrial haplogroup classifier. *Bioinformatics* **31(8)**:1301–1302 DOI 10.1093/bioinformatics/btu825.

**Olalde I, Allentoft ME, Sánchez-Quinto F, Santpere G, Chiang CW, DeGiorgio M, Prado-Martinez J, Rodríguez JA, Rasmussen S, Quilez J, Ramírez O, Marigorta UM, Fernández-Callejo M, Prada ME, Encinas JM, Nielsen R, Netea MG, Novembre J, Sturm RA, Sabeti P, Marquès-Bonet T, Navarro A, Willerslev E, Lalueza-Fox

C. 2014. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507**(**7491**):225–228 DOI 10.1038/nature12960.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**(**12**):5463–5467 DOI 10.1073/pnas.74.12.5463.

Tang S, Huang T. 2010. Characterization of mitochondrial DNA heteroplasmy using a parallel sequencing system. *Biotechniques* **48**(**4**):287–296 DOI 10.2144/000113389.

Taylor RW, Turnbull DM. 2005. Mitochondrial DNA mutations in human disease. *Nature Review Genetics* **6**(**5**):389–402 DOI 10.1038/nrg1606.

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ. 2006. Harvesting the fruit of the human mtDNA tree. *Trends in Genetics* **22**(**6**):339–345 DOI 10.1016/j.tig.2006.04.001.

Underhill PA, Kivisild T. 2007. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annual Review of Genetics* **41**:539–564 DOI 10.1146/annurev.genet.41.110306.130407.

Van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation* **30**(**2**):E386–E394 DOI 10.1002/humu.20921.

Vellarikkal SK, Dhiman H, Joshi K, Hasija Y, Sivasubbu S, Scaria V. 2015. mit-o-matic: a comprehensive computational pipeline for clinical evaluation of mitochondrial variations from next-generation sequencing datasets. *Human Mutation* **36**(**4**):419–424 DOI 10.1002/humu.22767.

Vianello D, Sevini F, Castellani G, Lomartire L, Capri M, Franceschi C. 2013. HAPLOFIND: a new method for high-throughput mtDNA haplogroup assignment. *Human Mutation* **34**(**9**):1189–1194 DOI 10.1002/humu.22356.

Weissensteiner H, Forer L, Fuchsberger C, Schöpf B, Kloss-Brandstätter A, Specht G, Kronenberg F, Schönherr S. 2016b. mtDNA-server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. *Nucleic Acids Research* **44**(**W1**):W64–W69 DOI 10.1093/nar/gkw247.

Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt HJ, Kronenberg F, Salas A, Schönherr S. 2016a. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Research* **44**(**W1**):W58–W63 DOI 10.1093/nar/gkw233.

Zhidkov I, Nagar T, Mishmar D, Rubin E. 2011. MitoBamAnnotator: a web-based tool for detecting and annotating heteroplasmy in human mitochondrial DNA sequences. *Mitochondrion* **11**(**6**):924–928 DOI 10.1016/j.mito.2011.08.005.