

学位論文（要約）

COMPREHENSIVE HUMAN MITOCHONDRIAL GENOME
ANALYSIS FOR POORLY-PRESERVED FOSSIL REMAINS

(劣化古人骨試料の包括的ミトコンドリアゲノム分析)

平成 29 年 12 月博士（理学）申請

東京大学大学院理学系研究科

生物科学専攻

石谷 孔司

Abstract

This paper summarizes the results of research aimed at realizing the accurate and efficient human mitochondrial genome analysis for poorly-preserved fossil remains. This thesis comprises four chapters, which are explained below.

Chapter 1 is an introduction describing the history of ancient DNA (aDNA) research and various challenges in ancient human mitochondrial genome analysis. The history of aDNA research can be divided into three periods. At the dawn of its study in the mid-1980s, scientists studied short, partial DNA sequences using only soft tissue of archeological samples. From the 1990s to the early 2000s, it was possible to analyze ultra-low amounts of DNA from hard tissues such as bone fragments by polymerase chain reaction techniques, and partial analysis of the hypervariable region (HVR) of the mitochondrial genome was often used for aDNA analysis. From 2006 to the present, next-generation sequencing (NGS) technologies have caused a dramatic paradigm shift in aDNA research. NGS allows ancient “genome-scale” research using complete mitochondrial or nuclear genome sequences. In contrast, there are still several challenges in analyzing contaminated or extremely degraded ancient samples.

第二章については、5年以内に雑誌等で刊行予定のため非公開

第三章については、5年以内に雑誌等で刊行予定のため非公開

In Chapter 4, I discuss the development of a comprehensive human mitochondrial analysis tool for NGS data and its convenience. Issues related to the determination of mitochondrial haplogroups, the detection of contamination, DNA damage, variant annotation, and the assembly of mitochondrial genome sequences are particularly important in analyzing degraded and aged human bone samples. This tool can estimate mitochondrial haplogroups with high accuracy—0.969, 0.991, 0.994, 0.994, and 0.997 at 5, 10, 50, 150, and 200-fold depth of coverage, respectively—and can be operated using a mouse pointer (graphical user interface operation), and analysis can be completed in approximately 1 min even for deep sequencing data with >1,000-fold coverage. Because the analysis results are visualized, the genome position where the damage or aDNA contamination is detected can be easily identified. This tool allows for a highly reliable ancient human mitochondrial genome analysis based on multiple items.

Table of Contents

Chapter 1. General Introduction	6
1-1 History of ancient DNA (aDNA) studies.....	6
1-2 Human mitochondrial genome	6
1-3 Challenges in ancient genome analysis using poorly-preserved fossil remains.....	7
1-4 Significance of this study	8

第二章については、 5年以内に雑誌等で刊行予定のため非公開

第三章については、 5年以内に雑誌等で刊行予定のため非公開

第三章については， 5 年以内に雑誌等で刊行予定のため非公開

Chapter 4. Development of a comprehensive human mitochondrial genome analysis tool for high-throughput sequencing data.....	48
4-1 Introduction	48
4-2 Material and methods	50
4-2-1 Format for input data.....	50
4-2-2 Test datasets.....	50
4-2-3 File processing.....	51
4-2-4 Detection of heteroplasmic sites	52
4-2-5 Optional functions	52
4-2-6 Software availability	54
4-3 Results and discussion	54
Figures, Figure Legends, and Tables	60
Conclusion Remarks	87
Literature Cited	89
Acknowledgements	102

Chapter 1. General Introduction

1-1 History of ancient DNA (aDNA) studies

The first aDNA study was the analysis of partial DNA sequences of the mitochondrial genome from soft tissues of Quagga (*Equus quagga*), which is an extinct member of the horse family (Higuchi *et al.*, 1984). In the following year, ancient human DNA analysis in 2,400-year-old Egyptian mummies was reported (Pääbo, 1985). From the 1990s until the early 2000s, the polymerase chain reaction (PCR) method allowed extremely small amounts of DNA fragments in archeological samples to be amplified, and the partial regions of the hyper variable region (HVR) of the mitochondrial genome have been analyzed (e.g., Handt *et al.*, 1996; Oota *et al.*, 1999; Wang *et al.*, 2000). However, the contamination by exogenous (typically modern) DNA in ancient samples becomes a serious problem in aDNA studies, and several strict guidelines for experimental processes have been proposed so as to prevent exogenous DNA contamination (Cooper & Poinar, 2000). From 2006 to the present, NGS technologies have expanded ancient “DNA” research to ancient “genome” research. NGS technologies allow massively parallel approaches to read very large numbers of DNA fragments, which enables genome-scale analysis of mitochondrial or nuclear genomes from archeological samples (Green *et al.*, 2006; 2008; Meyer *et al.*, 2012; Prüfer *et al.*, 2014).

1-2 Human mitochondrial genome

The human mitochondrial genome is a circular genome of approximately 17 kb resident in the mitochondrial organelle. It is present in hundreds to thousands of copies per cell compared to the diploid nuclear genome; therefore, the possibility of preservation is greatly increased, even in archeological samples. NGS technologies have made it possible to access the nuclear genome in fossil

remains, but the mitochondrial genome is often used to evaluate the extent of endogenous or exogenous DNA contamination in ancient samples (Fu *et al.*, 2013). The human mitochondrial genome has been used to study maternal phylogenetic relationships (Mishmar *et al.*, 2003; Macaulay *et al.*, 2005; Behar *et al.*, 2008) as well as disease-related mutations (Taylor & Turnbull, 2005) and for DNA identification (Holland & Parsons, 1999), and it is an important genomic region not only in molecular anthropology but also in medical and forensic fields.

1-3 Challenges in ancient genome analysis using poorly-preserved fossil remains

DNA molecules are degraded after the biological death of an organism. However, the degree and rate of DNA degradation enormously varies depending on environmental conditions, such as temperature or humidity, which greatly affect aDNA preservation (Hofreiter *et al.*, 2015). Major ancient genomics studies have reported little DNA contamination and have targeted well-preserved archeological samples with >1% endogenous aDNA. Well-preserved archeological samples with >1% endogenous aDNA can be analyzed on the genomic scale even with older samples of tens of thousands of years ago (e.g., Meyer *et al.*, 2012; Orlando *et al.*, 2013). However, most samples excavated in warm, humid, or dry areas are often poorly-preserved samples with rather <0.1%–1% endogenous DNA. In these poorly-preserved samples, only partial and incomplete sequences, even for the mitochondrial genome, can be obtained (e.g., Mohandesan *et al.*, 2016; Mizuno *et al.*, 2017). In addition, and unfortunately, the contamination by exogenous DNA remains a potential challenge, no matter how much effort is made to prevent it. For an accurate aDNA analysis, quality control must be thoroughly performed using the mitochondrial genome as an indicator of contamination. Comprehensive quality control is required to develop a new computational framework or software to

quickly and accurately confirm contaminated resources involving the human mitochondrial genome analysis.

1-4 Significance of this study

Until now, analytical methods widely applicable to contaminated and extremely degraded old human remains have not been established. In this paper, I employ NGS data of contaminated and degraded samples and discuss analytical methods and tools to resolve potential challenges involving poorly-preserved fossil remains. I believe that these approaches regarding contaminated or degraded samples will help promote the use of poor data that are otherwise discarded because of their low-quality and incompleteness.

本章については、5年以内に雑誌等で刊行予定のため非公開

本章については、5年以内に雑誌等で刊行予定のため非公開

Chapter 4. Development of a comprehensive human mitochondrial genome analysis tool for high-throughput sequencing data

4-1 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 sequences than does traditional capillary sequencing based on Sanger's method (Sanger, Nicklen & Coulson, 1977). NGS technologies also allow whole genome sequencing, exon sequencing, and gene expression profiling at high speeds and low cost-performance (Metzker, 2010; Hawkins, Hon & Ren, 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, short reads obtained by NGS can help estimate the probability of exogenous DNA sources in forensic samples (Just, Irwin & Parson, 2015). NGS technologies also promotes ancient human mitochondrial genome analysis in degraded archeological samples (Coia *et al.*, 2016; Llamas *et al.*, 2016; Schuenemann *et al.*, 2017).

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 human mitochondrial genome. MitoSeek (Guo *et al.*, 2013)

is one such character user interface (CUI) 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. These CUI-based tools are complicated in operation, and they are not designed as a general-purpose tool applicable to various samples. Because NGS data can contain various errors derived from sequencing platforms, experimental or sample conditions, it is necessary to investigate more items for reliable human mitochondrial genome analysis. There are also several web-based tools supporting NGS 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 these issues, I developed MitoSuite, a general-purpose and 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. MitoSuite 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. MitoSuite is the

first standalone, GUI software for comprehensive profiling of the mitochondrial genome, using high-throughput sequencing data with intuitive operability.

4-2 Material and methods

4-2-1 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.

4-2-2 Test datasets

I used seven sets of empirical sequencing data (NA11920, HG01112, NA18941, HG00096, HG00273, NA18548, NA18510) of 1,000 genomes project data (The 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, I generated simulated NGS reads using 324 worldwide mitochondrial genome sequences (Table 4-1) with ART, a simulation tool to generate synthetic NGS reads (Huang *et al.*, 2012). These sequence sets were selected from PhyloTree (van Oven, 2015) (<http://www.phylotree.org>), and included all known macro-haplogroups in nearly equal proportions. I 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 my in-house Python scripts. Based on the Illumina sequencer model in ART, I assumed 1% sequencing error, single-end 100-base reads, and average depth of 1–1000× in the simulated data. I aligned these simulated reads against rCRS using BWA, and then used these BAM files as simulation datasets.

4-2-3 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.

4-2-4 Detection of heteroplasmic sites

MitoSuite can also detect heteroplasmic sites that may be attributed to exogenous contamination, sequencing errors, amplification errors, or heteroplasmy. My 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:

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

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. 4-2A). This percentage ($P_{support}$) is computed as follows:

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

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. 4-2).

4-2-5 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, 1996) (<http://www.mitomap.org/MITOMAP>) (Fig. 4-3). 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). MitoSuite 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:

$$P_{mismatch} = \sum_{i=1}^k \left\{ \frac{(N_{mismatch}/N_{match})}{k} \right\} \times 100$$

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.

4-2-6 Software availability

MitoSuite is freely available from <https://mitosuite.com>. This tool mainly supports UNIX-like operating systems (OS) such as Mac OSX and Linux. MitoSuite for Mac OSX also provides the graphical installer package (Fig. 4-4). 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 on <https://mitosuite.com>.

4-3 Results and discussion

MitoSuite is designed for better usability, especially for non-bioinformaticians unfamiliar with typing complicated commands (Fig. 4-5). This tool provides a drag-and-drop functionality for loading a BAM file and automatically displays an output destination directory. MitoSuite 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 perform 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 results are finally outputted as a single html file (Fig. 4-6).

MitoSuite outputs analytical results for the quality of alignment data and genetic profiles that are haplogroup and polymorphisms on the mitochondrial genome. 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 in the visualized outputs, including categories on data quality and genetic profiles; this item shows an overview of the NGS data (Fig. 4-7). 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. 4-8). (2) Figure 4-9 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. Here, I 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 MitoSuite was easily able to analyze these ultra-deep data. Analysis of a sample dataset with MitoSuite is shown in Table 4-2. 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 web server independent environment. MitoSuite also successfully detected the fragmentation and deamination pattern of ancient DNA-like on empirical reads from Olalde *et al.*, 2014 (Fig. 4-9C, E). In addition, the most

likely haplogroup estimated by MitoSuite is “U5b2c1” that is consistent with reported one in Olalde *et al.*, 2014 (Table 4-2).

I 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. 4-10).

To examine whether MitoSuite can detect heteroplasmic sites previously reported in empirical high-throughput sequencing data, I used MPS raw read data from Avital *et al.* (2012) (Avital *et al.*, 2012). I 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 4-3). 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.*, 2016b) also supports the variant call format (VCF) storing DNA polymorphism data, as well as the two above-mentioned formats. mtDNA-Server

(Weissensteiner *et al.*, 2016a), 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 heteroplasmy from a single individual or tissue provides useful information in clinical or forensic cases. MToolBox, MitoSeek, MitoBamAnnotator, mtDNA-Server, mit-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 analysis environments must be very restrictive in terms of security systems. MitoSuite provides a server-independent system that brings advantages especially to medical and forensic researchers in terms of security.

MitoSuite can be a user-friendly all-in-one solution for many investigators unfamiliar with advanced data-processing techniques to deal with poorly-preserved samples. MitoSuite 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, post mortem damage (PMD) detection, and interactive dynamic graphics across the complete mitochondrial genome in NGS data (Table 4-4). Especially, haplogroup assignment, detection of contamination, PMD detection, variant annotation, and the assembly of mitochondrial genome sequences are important functions to realize an accurate human mitochondrial genome analysis for highly-degraded or contaminated samples. I expect MitoSuite to promote comprehensive human mitochondrial genome studies in the fields of anthropological science as well as forensic casework, and medicine.

Figures, Figure Legends, and Tables

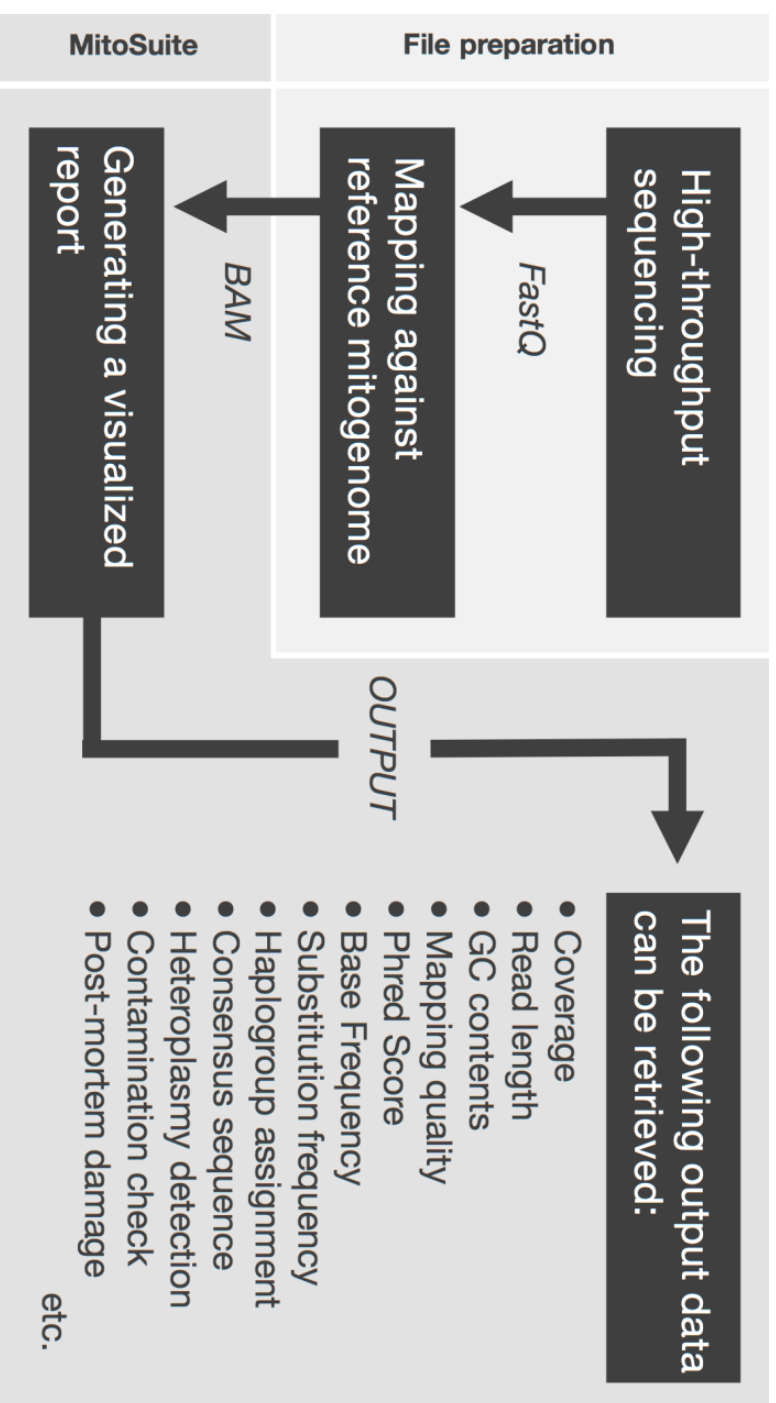


Figure 4-1. File preparation and processing flow for MitoSuite.

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.

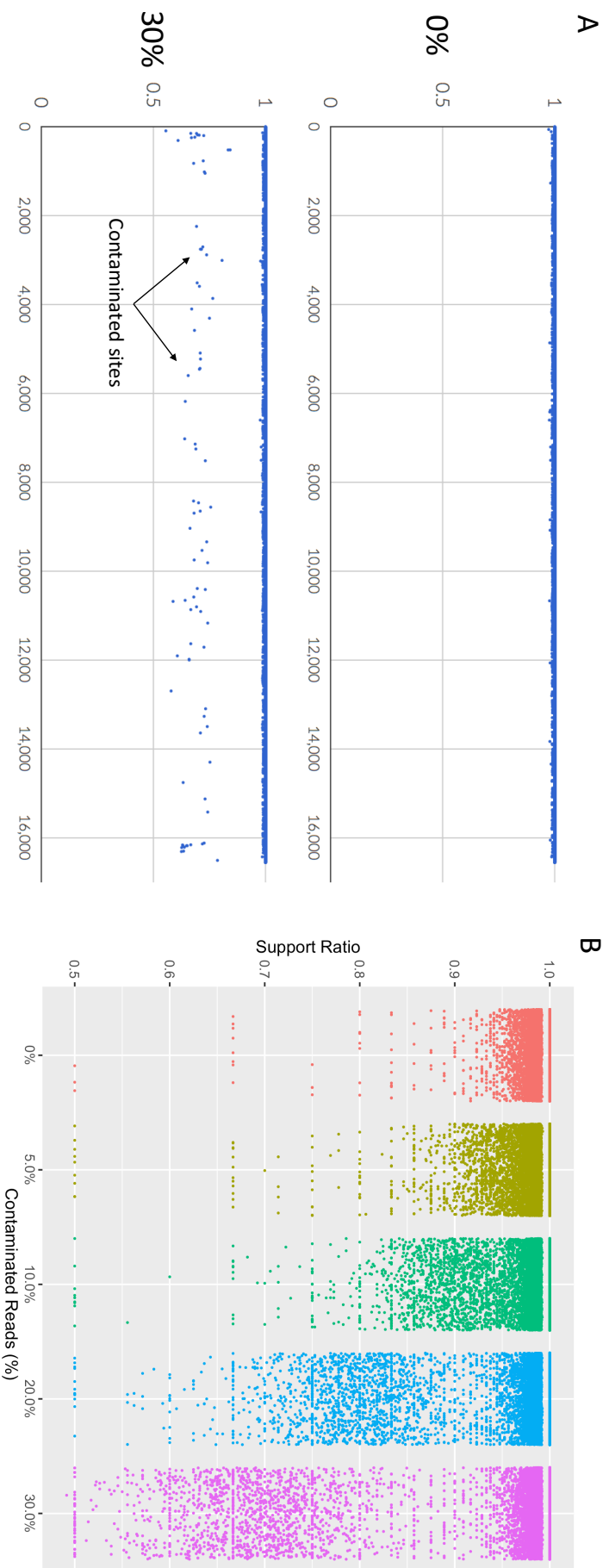


Figure 4-2. Support ratio in simulated NGS reads.

I used simulated reads of two distinct haplogroups (H and L0). These simulated reads are assumed 1% base-call error, average depth of 1–100 \times , and 0 – 30 % contamination rate that is the percentage of the number of contaminated reads in the total reads. (A) Two figures are scatter plots output from MitoSuite in the presence (30%) and the absence (0%) of exogenous contaminants, respectively. (B) This figure shows the support ratio for all simulated data set (1-100 \times ; 0-30 %).

Annotation of disease-related variants (B)
Reports of Disease-Associations in MITOMAP(Jan. 04, 2017 version).

					Search
					MITOMAP References
Position	Mutation	Reference	Query	Disease	
150	C150T	C	T	Longevity / Cervical Carcinoma / HPV infection risk	C150T
4282	A4282G	A	G	CPRO Plus	A4282G
10398	A10398A	A	A	Invasive Breast Cancer risk factor; AD; PD; BD lithium response; Type 2 DM	A10398A
11467	A11467G	A	G	Altered brain pH / sCJD patients	A11467G
12308	A12308G	A	G	CPRO / Stroke / CM / Breast & Renal & Prostate Cancer Risk / Altered brain pH /sCJD	A12308G
12372	G12372A	G	A	Altered brain pH / sCJD patients	G12372A
13637	A13637G	A	G	Possible LHON factor	A13637G
16192	C16192T	C	T	Melanoma patients	C16192T
16270	C16270T	C	T	Melanoma patients	C16270T
16519	T16519T	T	T	Cyclic Vomiting Syndrome with Migraine /metastasis	T16519T

*Position column shows a reference sequence position (rCRS).Mutation column means that of Allele in MITOMAPPlease check more detail information at [MITOMAP](#)

Annotation of an in-house database (B)

Mutations are annotated by using in-house_annotation_data.csv.

					Search
Position	Mutation	Reference	Query	Annotation	
114		C114C	C	C	annotation1
150		C150T	C	T	annotation2
195		T195T	T	T	annotation3

*Position column shows a reference sequence position (rCRS).

Figure 4-3. Screenshot of annotation results of the visualized outputs of MitoSuite.

Upper table shows the annotation result of reported disease-related variants. Position column shows a reference sequence position (rCRS). Mutation column means that of Allele in MITOMAP (Jan. 04, 2017 version). Lower table shows the annotation result of an in-house customizable annotation database. 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), and related information (e.g., related-disease names) in each designated column. The template of the annotation file is available from MitoSuite’s support page (<https://mitosuite.com>) or can be download by the installer.

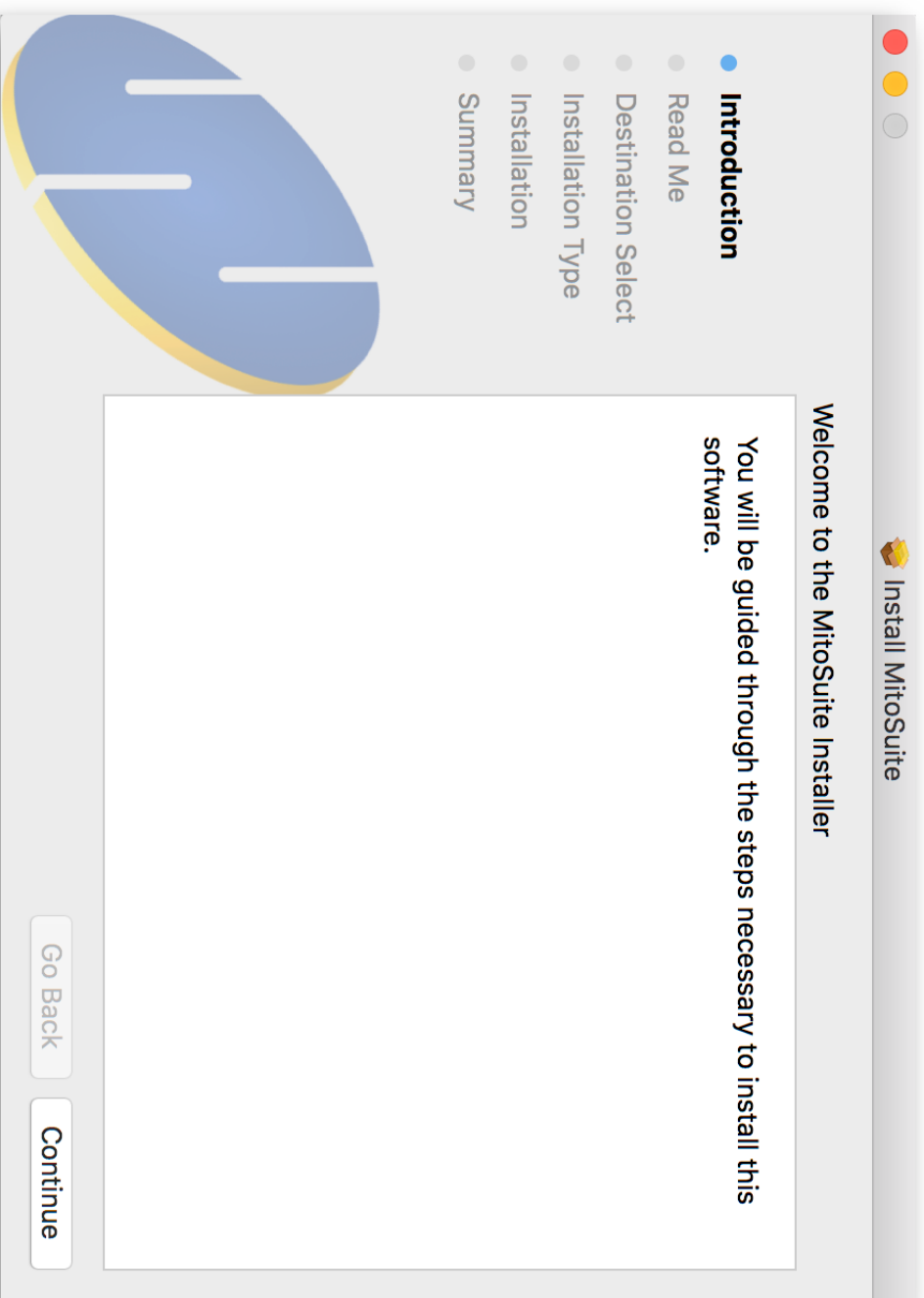


Figure 4-4. A screenshot of an automated installer package of MitoSuite.

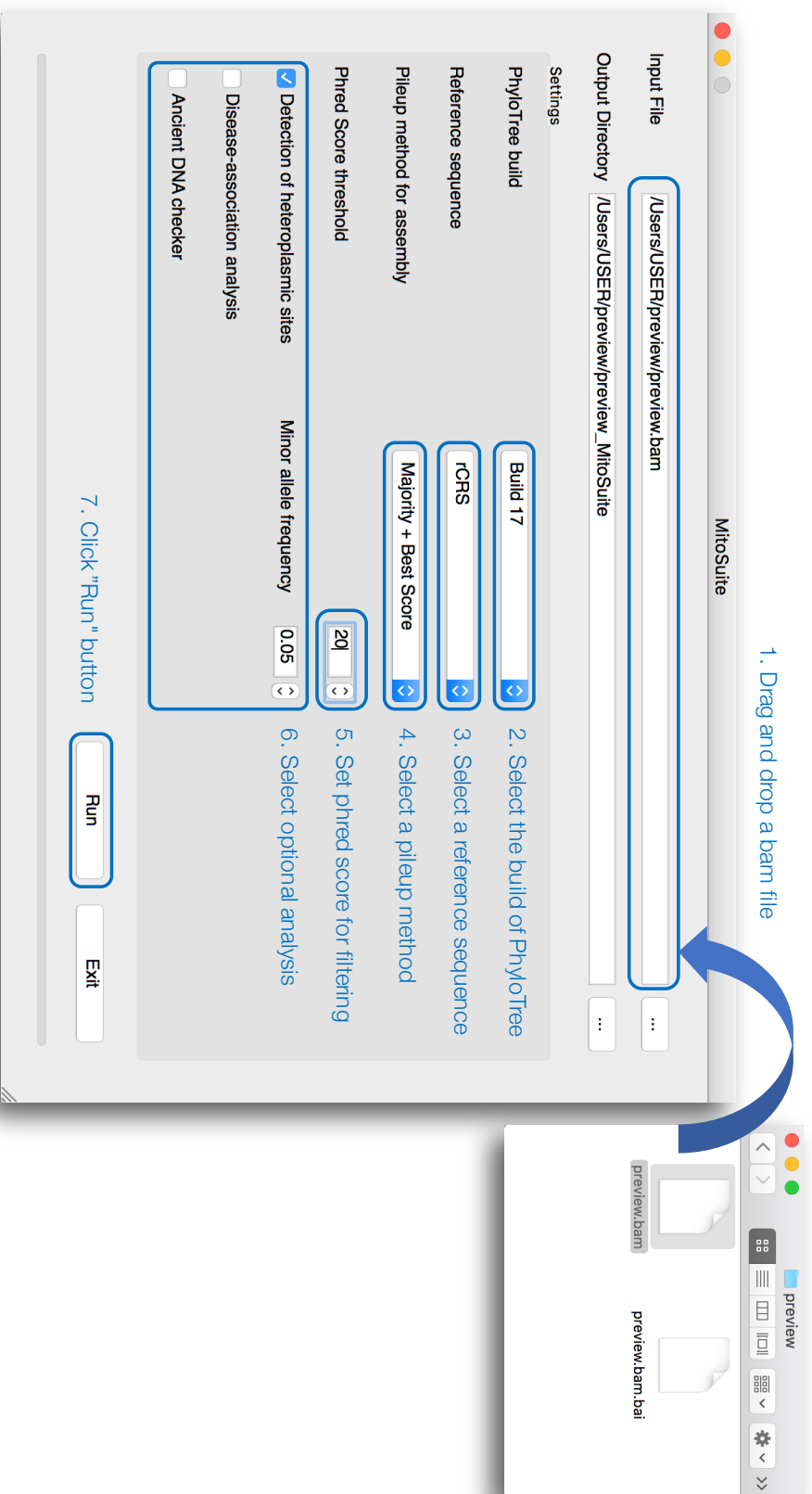


Figure 4-5. A screenshot of the graphical user interface of MitoSuite.

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.

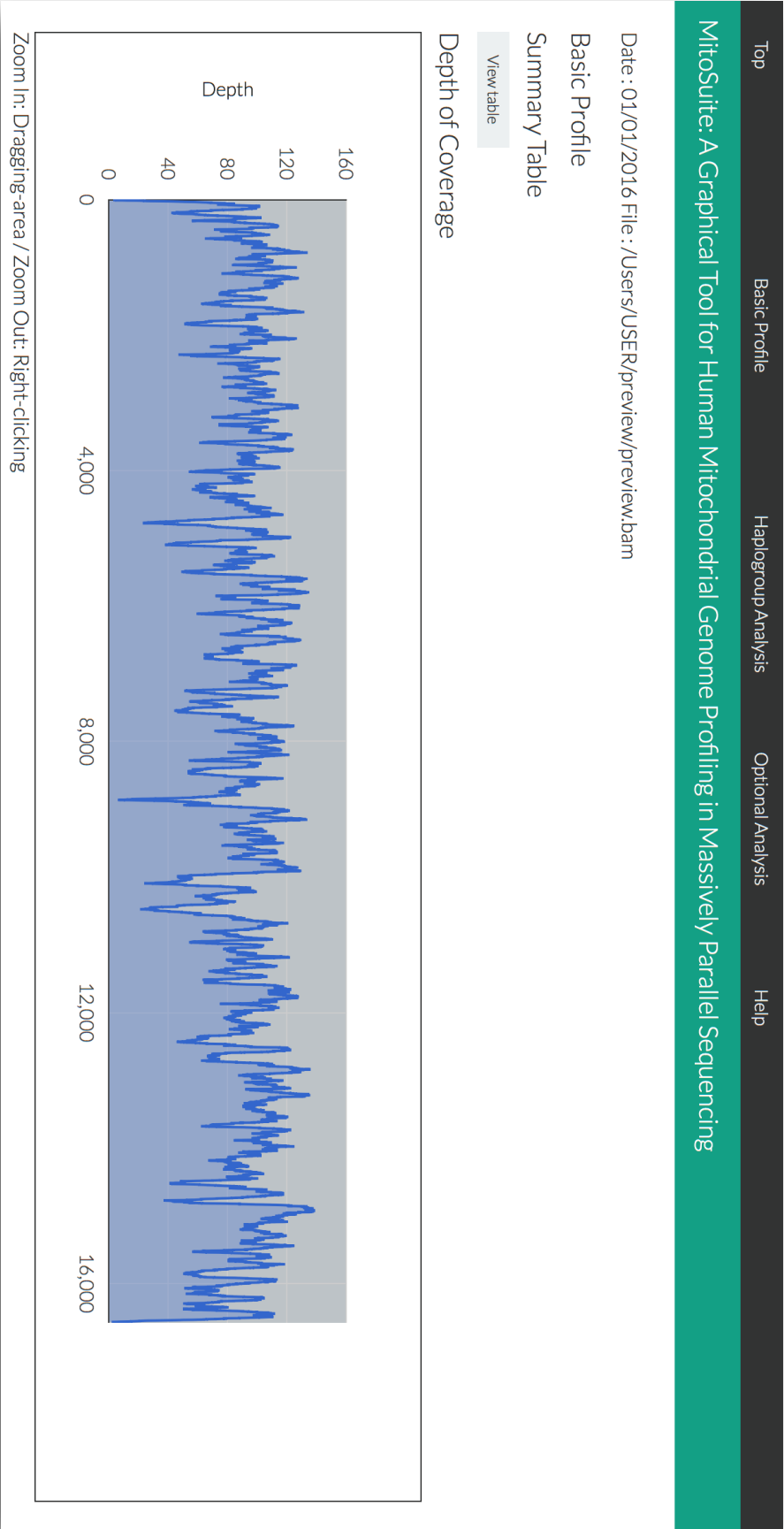


Figure 4-6. A screenshot of the visualized outputs of MitoSuite.

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.

Summary Table

[View table](#)

Click "View table" button to open the table below


Item	Value	Note
Total reads	21818	
Mapped reads	21818	Percentage of mapped : 100.0%
Unmapped reads	0	Percentage of unmapped : 0.0%
Duplicate reads	0	Percentage of duplicates : 0.0%
Reference name	gI251831106 ref NC_012920.1	
Reference length	16569	
Phred encoding	Phred+33	
Read length (avg.)	70.333	
GC% (avg.)	44.575%	
MapQ (avg.)	36.821	
Depth of coverage (avg.)	92.619	
Phred score threshold	20	
Mitogenome coverage (%)	100.0%	
Mapped sites (phred > 20)	16569	
Unmapped sites (phred > 20)	0	
Assigned haplogroup (phylotree build 16)	U5b2c1	the percentage of concordance of diagnostic sites : 100.0 %
Pileup method	Majority + Best score	
Assembled consensus sequence		
Observed base substitutions	6287	These are substitutions against an assembled consensus sequence.
The percentage of concordance of consensus bases (avg.)	99.3%	The percentage presents fraction of bases consistent with bases of an assembled consensus sequence.

Figure 4-7. A screenshot of summary statistics table in the outputs of MitoSuite.

This table shows 20 items on analysis results at a time and helps to grasp the data summary quickly.



Figure 4-8. A screenshot of area chart for depth of coverage across the human mitochondrial genome in the outputs of MitoSuite. Std; Standard deviation, Q1; 25th percentile, Q3; 75th percentile, Max; Maximum depth, Min; Minimum depth, >10-50 ; genome coverage rate over the depth value (10-50×).

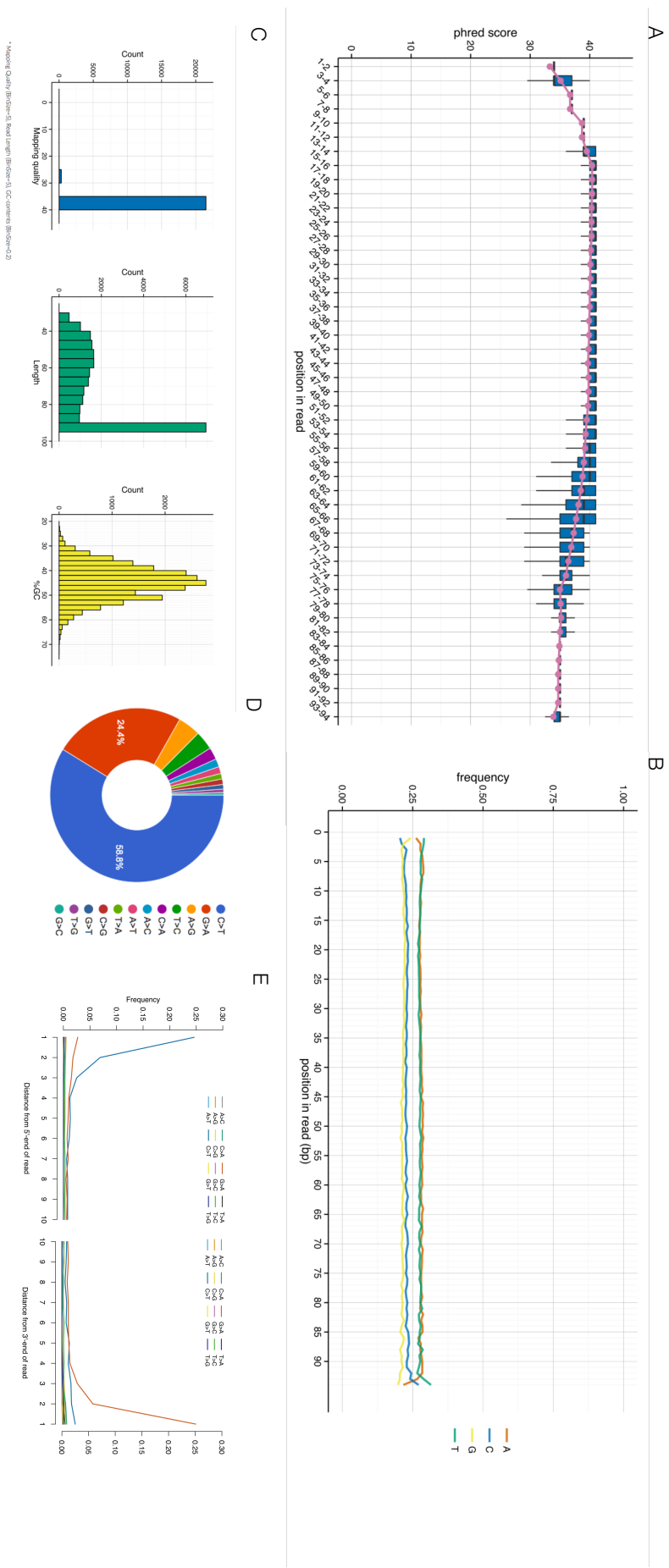


Figure 4-9. Different types of outputs given by MitoSuite.

(A) Box plot of phred scores (base quality). (B) Line plot of base frequency at each position of reads. (C) Histograms of mapping quality (blue), read length (green), and GC-contents (yellow). (D) Pie chart of proportion of base substitutions. (E) Line plot of base substitutions at each position of reads. This plot shows the deamination of cytosine to uracil, which is a post-mortem hydrolytic change representative of ancient sequences.

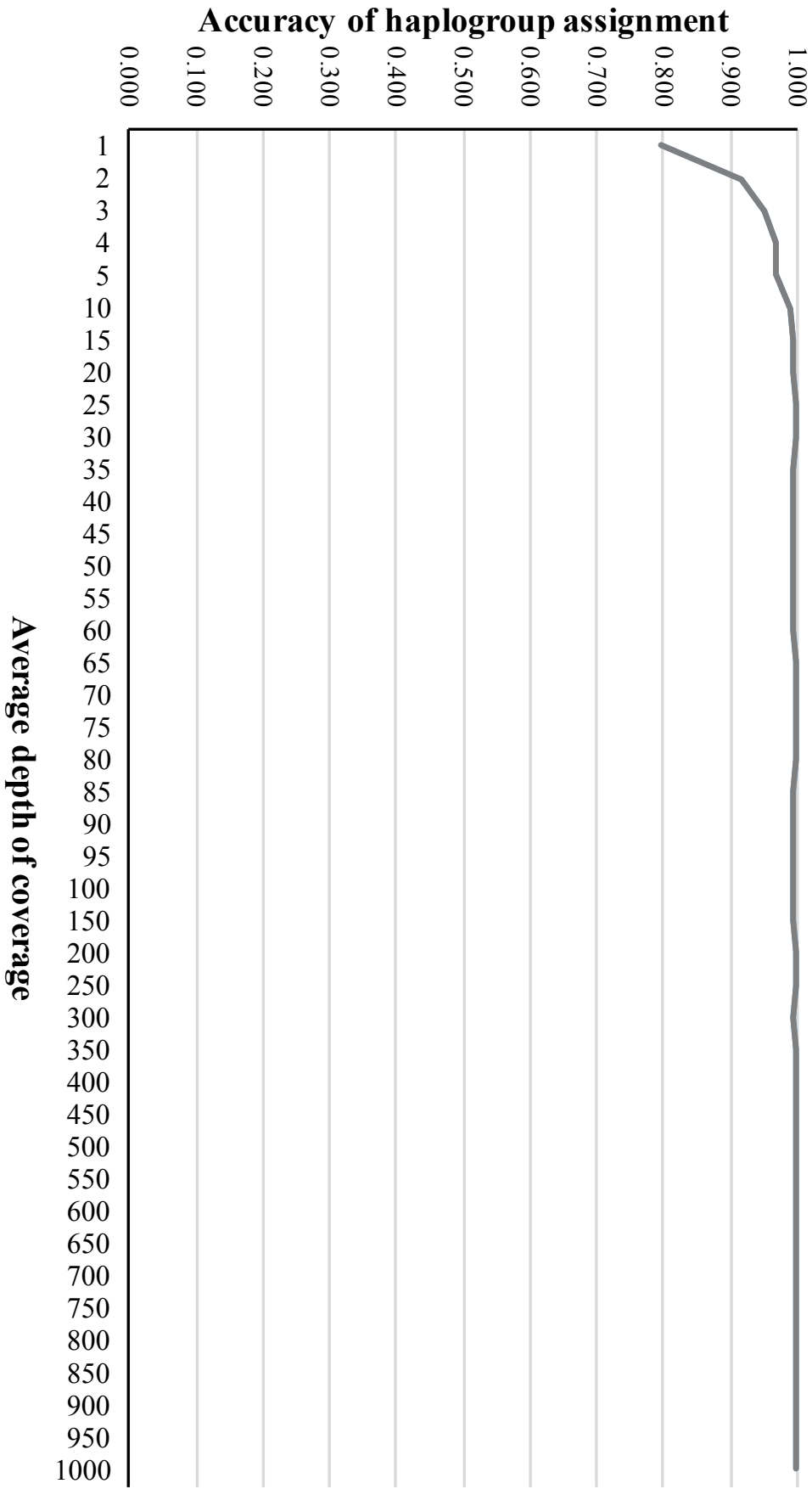


Figure 4-10. Haplogroup assignment accuracy of MitoSuite for simulated NGS reads generated from 324 worldwide mitochondrial genome sequences.

Table 4-1. Mitochondrial genome sequences to test haplogroup assignment in MitoSuite.

Accession	Haplogroup (PhyloTree Build 17)
EU092714	L0a1a1
JX303911	L0a1a2
JQ044943	L0a1a3
EU092906	L0a4
HM771161	L0a2a1
EU092936	L0b
EU092870	L0f1
KC345899	L0k1a1
KC345791	L0d1a
KC345891	L0d2d
DQ112737	L1b1
JQ705275	L1c1a
AF346992	L1c2a1a
HM771222	L1c2b1a
EU273502	L1c4a
EU273489	L1c6
JX303797	L1c5
EU092703	L1c3a
EU273488	L1c3b1a
EU273493	L1c3c
DQ341060	L5a1a
EF556173	L5a1a
EU092888	L5a1b
EU092943	L5a1b
HM771198	L5a1c
EU092699	L5a2
DQ341061	L5b1a
EU092774	L5b1b
DQ304928	L2a1a
EU092761	L2a1b
EU092747	L2b1
EU092661	L2b3c

Accession	Haplogroup (PhyloTree Build 17)
JQ044941	L2c
AY195785	L2c5
JQ045050	L2d
EU092794	L2d1a
EU092724	L2e
FJ460523	L2e1a
EU092773	L6a
EU092802	L6a
DQ341063	L6b
EU092686	L6b
FJ460531	L4a1
DQ341064	L4a1a
EU092935	L4a2
JQ044811	L4b1a
EU092942	L4b2a1
EF184627	L4b2a2
EU092938	L4b2a2a
JQ702504	L4b2b1
JN655813	L3a1a
EU092726	L3b1a
EU092891	L3f1a
EU092660	L3c
DQ112884	L3d
EU092827	L3e1
DQ341069	L3i1b
EU092822	L3k
EF556171	L3x1a1
JN655830	L3h1a1
JQ702955	M1a
AP012349	M20
KC505097	M51a
DQ408676	M3
DQ408679	M4

Accession	Haplogroup (PhyloTree Build 17)
AY922291	M65a1
EF556195	M5a
HM156679	M18a
JX289116	M38
HM346892	M9a1
AY519496	C1a
FJ951462	C4a1
FJ951515	C5a1
FJ951594	C7
FJ383641	C7b
FJ951440	C5d1
FJ951614	C5c
FJ951472	C5b1
FJ951611	C4e
GQ895155	C4d
AY519493	Z1
FJ147318	Z1a
AP008426	Z2
AP008841	Z3
EU597518	Z3a1
FJ383644	Z3b
GU392051	Z4
AP013181	Z4a1a1
AP008553	Z5
FJ383629	Z7
GQ119027	E1a
EF093539	E1a1
EF185804	E1a1a1
HQ700849	E1a2
HQ700847	E1a2a
FJ428236	E1b
EF061150	E1b1
HQ700856	E2a

Accession	Haplogroup (PhyloTree Build 17)
EF185816	E2b1
EF093542	E2b2
AP008845	G1a1a4
EF153773	G1a1
AY195762	G1b
AY255137	G1c
FJ198217	G2a1
FJ015040	G2b1a
JF824842	G2c
GU014566	G3a1
FJ748726	G3b1
AP008911	G4
DQ372885	Q1
AY289085	Q1a
AF347003	Q1b
EU597543	Q1c
HQ113226	Q2a
GQ214525	Q2a1
EF495218	Q2b
AY289079	Q3a
AY289089	Q3a1
EF061146	Q3b
JQ704974	D4
JN253391	D1
AP011004	D4b2
AP008251	D4c1b
AP008628	D4d
AP008789	D4e1a
EU660536	D2a
JF824877	D2b1
AP008519	D5c1a
AP011006	D6a1
JQ245777	N1a1a

Accession	Haplogroup (PhyloTree Build 17)
AY714031	N5
EU787451	N2a
KC867130	N3
KC887495	N7a1
JX289118	N8
AP008726	N9a1
EF495214	N13
DQ112753	N14
JF739542	N22
AP010699	A1
DQ282395	A2
AY963575	A6a
EF153771	A12
JF824996	A15a
EF153780	A16
AP013217	A3
AP008265	A5a
EU482363	A8
HM569228	A10
JQ245791	I
HM454265	I1a
FJ968796	I1b
JQ705932	I1c
EU570217	I2
JQ702041	I3a
KJ021059	I4
JQ245724	I5
JQ705382	I6a
KF146253	I7
AY289059	O
DQ404447	O1
AY289056	O1a
AF346963	S1

Accession	Haplogroup (PhyloTree Build 17)
JN226144	S1a
AY289051	S2
AY289066	S3
AY289062	S4
EF495220	S5
EU600318	X1a
EF177437	X3
GQ200588	X2
DQ523631	X2c
EU600325	X2h
JQ245730	X2i
FJ008043	X2m1
EF660942	X2n
JQ245731	X2o
HQ456226	X4
EU439939	X2a1
FJ825753	X2a2
EU935450	X2j
FJ457949	X2b
JQ705795	X2d
KF056262	W
JQ702793	W1
JQ705313	W3a
KF146279	W3b
KF146281	W4
KF146285	W5
JQ245723	W6
HM352797	W7
JQ245778	W8
JQ245759	W9
EF153813	Y1
HM776715	Y1a
GU123044	Y1b

Accession	Haplogroup (PhyloTree Build 17)
AY255138	Y1b1
JF824992	Y1b1a
AP008723	Y2
GQ119016	Y2a
GQ119013	Y2a1
KC994134	Y2a1a
AP008764	Y2b
KC985149	R1a
JQ705561	R1a1a
KC985159	R1a1
HM030522	R1a1b
KC985148	R1a1c
KC985165	R1b
JQ703633	R1b1
JX155266	R2a
EF556167	R2b
JX155271	R2c
AY255136	B6a
JX900371	B4a1a1
JQ704728	B2a
AY519494	B4b1a
AY255135	B4d1
KC733253	B4e
JN857017	B4j
KC521454	B4c1a
EU597566	B5a1a
JF824930	B5b1a
HM357817	F1a1a
AP010738	F1c
AP012346	F1f
KF849909	F2
AY255167	F3a
AP013180	F3b

Accession	Haplogroup (PhyloTree Build 17)
AP008744	F4a1a
AP013200	F4a1b
JF824892	F4a2
AY289095	F4b
JQ704041	J2a1
AY339579	J2a1a1a1
FJ348157	J2a1a1a2
JQ703568	J2a1a1a3
JQ705390	J2a1a1b
JQ797924	J2a2
EF660967	J2a2a
JQ797920	J2a2a1
JQ702563	J2b1
FJ213765	J2b2
AF347004	P1
DQ112897	P2
KC993994	P10
AY289052	P3a
EF061158	P4a
AY289063	P5
AY289053	P6
AY289054	P7
DQ404446	P8
KC993944	P9
KC405582	T1a
JQ798058	T2a1
AY495273	T2b
JQ798090	T2c
AY714037	T2d1a
EF177410	T2e
JQ703997	T2f
EU935442	T2g1
JN202724	T2h

Accession	Haplogroup (PhyloTree Build 17)
JX462725	T3
JQ702678	R0a1
HM185216	R0a1a
HM185266	R0a2
HM185261	R0a2b
HM185215	R0a2c
JF717359	R0a2d
HM185238	R0a2h
HM185265	R0a3
JQ702940	R0a4
JF717361	R0b
JQ705599	HV0
JF320654	HV1
AY713986	HV2
EF417833	HV4
EF419890	HV5
HQ658738	HV6
EU545443	HV7
EU545457	HV8
HM852849	HV12a
JF700125	HV13
HQ384174	H1
EU597492	H2a
AY738987	H3
EU935460	H4
AY495174	H5
HM765475	H6a
AY495120	H7
AY738957	H8a
EF660914	H31
HQ659700	H100
JQ702026	V1a
JQ705658	V2

Accession	Haplogroup (PhyloTree Build 17)
JQ703666	V3
AY495312	V4
AY339451	V5
FJ348207	V6
JQ705798	V7
JQ702025	V8
EF177445	V9
JQ704988	V23
HM852790	U1a1
KC521458	U5
EF064317	U6a
JQ702004	U2e2a4
FN600416	U2
HM852891	U3a
JQ703947	U4a1
AY882389	U9a
AY714004	U7
JQ702759	U8a
JQ706038	K1a
DQ301795	K1a1b1a
AY714044	K1a1b2a
JQ705454	K1a1c
JQ702168	K1b1a1a
AY495241	K2a
JX273249	K2b
JQ704064	K2b1b
DQ301796	K2c
HM852886	K3

Table 4-2. Sample dataset analyzed with MitoSuite.

Sample	Age	Haplogroup	Depth (avg)	Run (min)	Reference
NA11920	Modern	H1a1a1	1517	1.5	1000 Genomes Project
HG01112	Modern	A2ac1	1505	1.4	1000 Genomes Project
NA18941	Modern	N9b1a	1297	1.2	1000 Genomes Project
HG00096	Modern	H16a1	1213	1.1	1000 Genomes Project
HG00273	Modern	U5b1b2a	1117	1.1	1000 Genomes Project
NA18548	Modern	C4a1b	1021	1.1	1000 Genomes Project
NA18510	Modern	L0a1a3	747	0.8	1000 Genomes Project
La Brăna1	7,000 BP	U5b2c1	93	0.2	Olalde <i>et al.</i> , 2014

avg: average

Table 4-3. Detected heteroplasmic sites with MAF > 10 % in Avital *et al.*, (2012).

Sample ID	mtDNA Position	Region	Cell Type	Heteroplasmic fraction (Avital <i>et al.</i>)	Heteroplasmic fraction (MitoSuite)	> 10% (Avital <i>et al.</i>)	> 10% (MitoSuite)	Run
95451	9077	ATP6	Blood	43.99	41.3	✓	✓	SRR409202
95451	9077	ATP7	Muscle	47.22	45.1	✓	✓	SRR420854
95452	9077	ATP7	Blood	38.53	48.9	✓	✓	SRR420827
95452	9077	ATP7	Muscle	41.82	28.7	✓	✓	SRR420855
70251	16213	D-L-OOP	Blood	10.64	15.2	✓	✓	SRR420842
70251	16213	D-L-OOP	Muscle	15.21	33.3	✓	✓	SRR420957
70252	16213	D-L-OOP	Blood	11.81	12.2	✓	✓	SRR420844
68842	152	D-L-OOP	Muscle	11.88	41.9	✓	✓	SRR420859
68842	11299	ND4	Muscle	11.52	11	✓	✓	SRR420859
68842	14167	ND6	Muscle	10.94	36	✓	✓	SRR420859
68842	15132	CYB	Muscle	48.71	18.9	✓	✓	SRR420859
68842	15132	CYB	Blood	48.11	39.5	✓	✓	SRR420831
68841	14167	CYB	Muscle	12.42	47.2	✓	✓	SRR420858
68841	15132	D-L-OOP	Blood	41.96	21.3	✓	✓	SRR420830
51402	385	D-L-OOP	Muscle	10.3	7.5	✓	-	SRR420863

Low-quality and duplicates reads were trimmed in QC procedures before we analyzed them using MitoSuite.

Table 4-4. The list of available bioinformatics tools for mtDNA analysis (1/2).

	MitoSuite	HaploGrep2	mtDNA-Server	mtDNAprofiler	MTtoolBOX	MitoSeek
Input file	bam	fasta, hsd, vcf	fastq, bam, vcf	fasta, variant*	fastq, bam, sam	bam
Userinterface	GUI	Web	Web	Web	CUI	CUI
Supported reference sequence	rCRS, RSRS, hg19 [†] , GRCh [‡]	rCRS, RSRS	rCRS, RSRS	rCRS	rCRS, RSRS	rCRS, hg19 [†]
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	✓	-	✓	-	✓	✓

✓ : available, - : not-avaiaible

* A text-based format for describing SNPs information against a reference seuqunce.

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

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

Table 4-4. The list of available bioinformatics tools for mtDNA analysis (2/2).

	MitoBamAnnotator	mtDNAManager	MitoTool	HAPLOFIND	mit-o-matic	Phy-Mer
Input file	bam	variant*	fasta, variant*	fasta	fastq, pileup [†]	fasta, fastq, bam
Userinterface	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	✓	-	-	-	✓	-

✓ : available, - : not-available

* A text-based format for describing SNPs information against a reference seugquence.

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

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

Conclusion Remarks

In Chapter 1, I introduced the history of previous aDNA studies and summarized the challenges faced in DNA analysis dealing with poorly-preserved fossil remains. These main challenges include exogenous human DNA contamination and missing regions in the genome data. In this thesis, I focused on the human mitochondrial genome, which is used for quality control in ancient genomics,

第二章ならびに第三章に関する内容は、5年以内に雑誌等で刊行予定のため非公開

I summarized the development of MitoSuite, which is an all-in-one tool that can comprehensively

analyze the human mitochondrial genome in high-throughput sequencing data. MitoSuite allows the comprehensive analysis of multiple essential parameters involving the haplogroup assignment, the detection of contaminated genomic position, the calculation of contamination rate, PMD detection, and the assembly of a consensus sequence, which are essential for an accurate aDNA analysis. This tool achieves much more straightforward and comprehensive ancient mitochondrial genome analysis. The computational framework and analysis tools in Chapters 2–4 could be a useful approach to contaminated or extremely degraded fossil remains that previously could not be analyzed because of their conditions.

Literature Cited

- Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, Campos PF, Samaniego JA, Gilbert MTP, Willerslev E, Zhang G, Scofield RP, Holdaway RN, Bunce M 2012. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings. Biological sciences / The Royal Society* 279:4724–4733. DOI: 10.1098/rspb.2012.1745.
- Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG 1981. Sequence and organization of the human mitochondrial genome. *Nature* 290:457–465. DOI: 10.1038/290457a0.
- 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* 23:147–147. DOI: 10.1038/13779.
- Atkinson QD, Gray RD, Drummond AJ 2009. Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa. *Proceedings. Biological sciences / The Royal Society* 276:367–373. DOI: 10.1098/rspb.2008.0785.
- 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* 21:4214–4224. DOI: 10.1093/hmg/ddc245.
- Barbieri C, Vicente M, Rocha J, Mpoloka SW, Stoneking M, Pakendorf B 2013. Ancient substructure in early mtDNA lineages of southern Africa. *American journal of human genetics* 92:285–292. DOI: 10.1016/j.ajhg.2012.12.010.
- Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli E-L, Silva NM, Kivisild T, Torroni A, Villems R 2012. A “Copernican” Reassessment of the Human Mitochondrial DNA Tree from its Root. *The American Journal of Human Genetics* 90:675–684. 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. *The American Journal of Human Genetics* 82:1130–1140. DOI: 10.1016/j.ajhg.2008.04.002.

- Biffi A, Anderson CD, Nalls MA, Rahman R, Sonni A, Cortellini L, Rost NS, Matarin M, Hernandez DG, Plourde A, de Bakker PIW, Ross OA, Greenberg SM, Furie KL, Meschia JF, Singleton AB, Saxena R, Rosand J 2010. Principal-Component Analysis for Assessment of Population Stratification in Mitochondrial Medical Genetics. *The American Journal of Human Genetics* 86:904–917. DOI: 10.1016/j.ajhg.2010.05.005.
- Bolger AM, Lohse M, Usadel B 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. DOI: 10.1093/bioinformatics/btu170.
- Bonatto SL, Salzano FM 1997. A single and early migration for the peopling of the Americas supported by mitochondrial DNA sequence data. *Proceedings of the National Academy of Sciences of the United States of America* 94:1866–1871.
- Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prufer K, Meyer M, Krause J, Ronan MT, Lachmann M, Paabo S 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences* 104:14616–14621. 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* 30:3115–3117. DOI: 10.1093/bioinformatics/btu483.
- Cann RL, Stoneking M, WILSON AC 1987. Mitochondrial DNA and human evolution. *Nature* 325:31–36. DOI: 10.1038/325031a0.
- Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft ME, Sikora M, Rasmussen M, Gravel S, Guillén S, Nekhrizov G, Leshtakov K, Dimitrova D, Theodossiev N, Pettener D, Luiselli D, Sandoval K, Moreno-Estrada A, Li Y, Wang J, Gilbert MTP, Willerslev E, Greenleaf WJ, Bustamante CD 2013. Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. *The American Journal of Human Genetics* 93:852–864. DOI: 10.1016/j.ajhg.2013.10.002.
- Coia V, Cipollini G, Anagnostou P, Maixner F, Battaggia C, Brisighelli F, Gómez-Carballa A, Bisol GD, Salas A, Zink A 2016. Whole mitochondrial DNA sequencing in Alpine populations and the genetic history of the Neolithic Tyrolean Iceman. *Scientific Reports* 6:18932. DOI: 10.1038/srep18932.

- Cooper A, Poinar HN 2000. Ancient DNA: Do It Right or Not at All. *Science* 289:1139b–1139. DOI: 10.1126/science.289.5482.1139b.
- Edgar RC 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32:1792–1797. DOI: 10.1093/nar/gkh340.
- Enk JM, Devault AM, Kuch M, Murgha YE, Rouillard JM, Poinar HN 2014. Ancient Whole Genome Enrichment Using Baits Built from Modern DNA. *Molecular biology and evolution* 31:1292–1294. DOI: 10.1093/molbev/msu074.
- Eshleman J, Smith DG 2001. Use of DNase to eliminate contamination in ancient DNA analysis. *ELECTROPHORESIS* 22:4316–4319. DOI: 10.1002/1522-2683(200112)22:20<4316::AID-ELPS4316>3.0.CO;2-V.
- Fan L, Yao Y-G 2013. An update to MitoTool: Using a new scoring system for faster mtDNA haplogroup determination. *Mitochondrion* 13:360–363. DOI: 10.1016/j.mito.2013.04.011.
- Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prüfer K, de Filippo C, Meyer M, Zwyns N, Salazar-García DC, Kuzmin YV, Keates SG, Kosintsev PA, Razhev DI, Richards MP, Peristov NV, Lachmann M, Douka K, Higham TFG, Slatkin M, Hublin J-J, Reich D, Kelso J, Viola TB, Pääbo S 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514:445–449. DOI: doi:10.1038/nature13810.
- Fu Q, Mittnik A, Johnson PLF, Bos K, Lari M, Bollongino R, Sun C, Giemsch L, Schmitz R, Burger J, Ronchitelli AM, Martini F, Cremonesi RG, Svoboda J, Bauer P, Caramelli D, Castellano S, Reich D, Pääbo S, Krause J 2013. A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. *Current Biology* 23:553–559. DOI: 10.1016/j.cub.2013.02.044.
- Gamba C, Hanghøj K, Gaunitz C, Alfarhan AH, Alquraishi SA, Al-Rasheid KAS, Bradley DG, Orlando L 2015. Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. *Molecular Ecology Resources* 16:459–469. DOI: 10.1111/1755-0998.12470.
- Gansauge M-T, Meyer M 2014. Selective enrichment of damaged DNA molecules for ancient genome sequencing. *Genome Research* 24:1543–1549. DOI: 10.1101/gr.174201.114.
- Gojobori J, Mizuno F, Wang L, Onishi K, Granados J, Gomez-Trejo C, Acuña-Alonzo V, Ueda S 2015. mtDNA diversity of the Zapotec in Mexico suggests a population decline long before the

- first contact with Europeans. *Journal of Human Genetics* 60:557–559. DOI: 10.1038/jhgc.2015.55.
- Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, Pääbo S 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature* 444:330–336. DOI: 10.1038/nature05336.
- Green RE, Malaspinas A-S, Krause J, Briggs AW, Johnson PLF, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, Prüfer K, Siebauer M, Burbano HA, Ronan M, Rothberg JM, Egholm M, Rudan P, Brajković D, Kućan Ž, Gušić I, Wikström M, Laakkonen L, Kelso J, Slatkin M, Pääbo S 2008. A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing. *Cell* 134:416–426. DOI: 10.1016/j.cell.2008.06.021.
- Guo Y, Li J, Li C-I, Shyr Y, Samuels DC 2013. MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. *Bioinformatics* 29:1210–1211. DOI: 10.1093/bioinformatics/btt118.
- Handt O, Krings M, Ward RH, Paabo S 1996. The retrieval of ancient human DNA sequences. *The American Journal of Human Genetics* 59:368–376.
- Hawkins MTR, Hofman CA, Callicrate T, McDonough MM, Tsuchiya MTN, Gutiérrez EE, Helgen KM, Maldonado JE 2015. In-solution hybridization for mammalian mitogenome enrichment: pros, cons and challenges associated with multiplexing degraded DNA. *Molecular Ecology Resources* 16:1173–1188. DOI: 10.1111/1755-0998.12448.
- Hawkins RD, Hon GC, Ren B 2010. Next-generation genomics: an integrative approach. *Nature Review Genetics* 11:476. DOI: 10.1038/nrg2795.
- Higuchi R, Bowman B, Freiburger M, Ryder OA, WILSON AC 1984. DNA sequences from the quagga, an extinct member of the horse family. *Nature* 312:282–284. DOI: 10.1038/312282a0.
- Hofreiter M, Paijmans JLA, Goodchild H, Speller CF, Barlow A, Fortes GG, Thomas JA, Ludwig A, Collins MJ 2015. The future of ancient DNA: Technical advances and conceptual shifts. *BioEssays* 37:284–293. DOI: 10.1002/bies.201400160.
- Holland MM, Parsons TJ 1999. Mitochondrial DNA Sequence Analysis - Validation and Use for Forensic Casework. *Forensic science review* 11:21–50.
- Huang W, Li L, Myers JR, Marth GT 2012. ART: a next-generation sequencing read simulator. *Bioinformatics* 28:593–594. DOI: 10.1093/bioinformatics/btr708.

- Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang Q-F, Li J, Bin Han 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature genetics* 42:961–967. DOI: 10.1038/ng.695.
- Hudson G, Gomez-Duran A, Wilson IJ, Chinnery PF 2014. Recent Mitochondrial DNA Mutations Increase the Risk of Developing Common Late-Onset Human Diseases. *PLoS genetics* 10:e1004369. DOI: 10.1371/journal.pgen.1004369.
- Ingman M, Kaessmann H, Pääbo S, GYLLENSTEN U 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* 408:708–713. DOI: 10.1038/35047064.
- Ishiya K, Ueda S 2017. MitoSuite: a graphical tool for human mitochondrial genome profiling in massive parallel sequencing. *PeerJ* 5:e3406. DOI: 10.7717/peerj.3406.
- 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.
- Katoh K, Standley DM 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* 30:772–780. DOI: 10.1093/molbev/mst010.
- Kihana M, Mizuno F, Sawafuji R, Wang L, Ueda S 2013. Emulsion PCR-coupled target enrichment: An effective fishing method for high-throughput sequencing of poorly preserved ancient DNA. *Gene* 528:347–351. DOI: 10.1016/j.gene.2013.07.040.
- Kim D, Song L, Breitwieser FP, Salzberg SL 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research* 26:gr.210641.116–1729. DOI: 10.1101/gr.210641.116.
- Kogelnik A 1996. MITOMAP: a human mitochondrial genome database. *Nucleic Acids Research* 24:177–179. DOI: 10.1093/nar/24.1.177.
- Krause J, Fu Q, Good JM, Viola B, Shunkov MV, Derevianko AP, Bo SPAA 2010. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* 464:894–897. DOI: doi:10.1038/nature08976.

- Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA 2009. Circos: an information aesthetic for comparative genomics. *Genome Research* 19:1639–1645. DOI: 10.1101/gr.092759.109.
- Lee HY, Song I, Ha E, Cho S-B, Yang WI, Shin K-J 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* 25:1754–1760. DOI: 10.1093/bioinformatics/btp324.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup 1GPDP 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. DOI: 10.1093/bioinformatics/btp352.
- Linderholm A 2015. Ancient DNA: the next generation – chapter and verse. *Biological Journal of the Linnean Society* 117:150–160. DOI: 10.1111/bij.12616.
- Lippold S, Xu H, Ko A, Li M, Renaud G, Butthof A, Schröder R, Stoneking M 2014. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. *Investigative Genetics* 5:13. DOI: 10.1186/2041-2223-5-13.
- Liu J, Wang L-D, Sun Y-B, Li E-M, Xu L-Y, ZHANG Y-P, Yao Y-G, Kong Q-P 2012. Deciphering the Signature of Selective Constraints on Cancerous Mitochondrial Genome. *Molecular biology and evolution* 29:1255–1261. DOI: 10.1093/molbev/msr290.
- Llamas B, Fehren-Schmitz L, Valverde G, Soubrier J, Mallick S, Rohland N, Nordenfelt S, Valdiosera C, Richards SM, Rohrlach A, Romero MIB, Espinoza IF, Cagigao ET, Jiménez LW, Makowski K, Reyna ISL, Lory JM, Torrez JAB, Rivera MA, Burger RL, Ceruti MC, Reinhard J, Wells RS, Politis G, Santoro CM, Standen VG, Smith C, Reich D, Ho SYW, Cooper A, Haak W 2016. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Science Advances* 2:e1501385–e1501385. DOI: 10.1126/sciadv.1501385.
- 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 H-J, Oppenheimer S, Torroni A, Richards M 2005. Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes. *Science* 308:1034–1036. DOI: 10.1126/science.1109792.

- Malaspinas A-S, Tange O, Moreno-Mayar JV, Rasmussen M, DeGiorgio M, Wang Y, Valdiosera CE, Politis G, Willerslev E, Nielsen R 2014. bammds: a tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS). *Bioinformatics* 30:2962–2964. DOI: 10.1093/bioinformatics/btu410.
- Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ 2010. Target-enrichment strategies for next-generation sequencing. *Nature Methods* 7:111–118. DOI: 10.1038/nmeth.1419.
- Manning CD, Raghavan P, Schütze H 2008. *Introduction to Information Retrieval*. Cambridge University Press.
- Maricic T, Whitten M, Pääbo S 2010. Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PloS one* 5:e14004. DOI: 10.1371/journal.pone.0014004.
- 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* 20:1297–1303. DOI: 10.1101/gr.107524.110.
- Meng X-L 1994. Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science* 9:538–558. DOI: 10.1214/ss/1177010269.
- Metzker ML 2010. Sequencing technologies — the next generation. *Nature Review Genetics* 11:31–46. DOI: 10.1038/nrg2626.
- Meyer M, Fu Q, Aximu-Petri A, Glocke I, Nickel B, Arsuaga J-L, Martínez I, Gracia A, de Castro JMB, Carbonell E, Pääbo S 2013. A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature* 505:403–406. DOI: 10.1038/nature12788.
- Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andrés AM, Eichler EE, Slatkin M, Reich D, Kelso J, Pääbo S 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science* 338:222–226. DOI: 10.1126/science.1224344.
- Mikami E, Fuku N, Kong Q-P, Takahashi H, Ohiwa N, Murakami H, Miyachi M, Higuchi M, Tanaka M, Pitsiladis YP, Kawahara T 2013. Comprehensive analysis of common and rare

- mitochondrial DNA variants in elite Japanese athletes: a case–control study. *Journal of Human Genetics* 58:780–787. DOI: 10.1038/jhg.2013.102.
- 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* 100:171–176. DOI: 10.1073/pnas.0136972100.
- Mitchell TM 1997. *Machine Learning*. McGraw-Hill Series in Computer Science.
- Mizuno F, Gojobori J, Wang L, Onishi K, Sugiyama S, Granados J, Gomez-Trejo C, Acuña-Alonzo V, Ueda S 2014. Complete mitogenome analysis of indigenous populations in Mexico: its relevance for the origin of Mesoamericans. *Journal of Human Genetics* 59:359–367. DOI: 10.1038/jhg.2014.35.
- Mizuno F, Kumagai M, Kurosaki K, Hayashi M, Sugiyama S, Ueda S, Wang L 2017. Imputation approach for deducing a complete mitogenome sequence from low-depth-coverage next-generation sequencing data: application to ancient remains from the Moon Pyramid, Mexico. *Journal of Human Genetics* 15:47. DOI: 10.1038/jhg.2017.14.
- Mohandesan E, Speller CF, Peters J, Uerpmann HP, Uerpmann M, De Cupere B, Hofreiter M, Burger PA 2016. Combined hybridization capture and shotgun sequencing for ancient DNA analysis of extinct wild and domestic dromedary camel. *Molecular Ecology Resources*. DOI: 10.1111/1755-0998.12551.
- Navarro-Gomez D, Leipzig J, Shen L, Lott M, Stassen APM, 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* 31:1310–1312. DOI: 10.1093/bioinformatics/btu825.
- Olalde I, Allentoft ME, Sánchez-Quinto F, Santpere G, Chiang CWK, DeGiorgio M, Prado-Martinez J, Rodríguez JA, Rasmussen S, Quilez J, Ramírez O, Marigorta UM, Fernandez-Callejo M, Prada ME, Encinas JMV, Nielsen R, Netea MG, Novembre J, Sturm RA, Sabeti P, Marques-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* 507:225–228. DOI: doi:10.1038/nature12960.
- Oota H, Saitou N, Matsushita T, Ueda S 1999. Molecular Genetic Analysis of Remains of a 2,000-Year-Old Human Population in China—and Its Relevance for the Origin of the Modern Japanese Population. *The American Journal of Human Genetics* 64:250–258. DOI: 10.1086/302197.

- Orlando L, Gilbert MTP, Willerslev E 2015. Reconstructing ancient genomes and epigenomes. *Nature Reviews Genetics* 16:395–408. DOI: 10.1038/nrg3935.
- Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, Johnson PLF, Fumagalli M, Vilstrup JT, Raghavan M, Korneliussen T, Malaspinas A-S, Vogt J, Szklarczyk D, Kelstrup CD, Vinther J, Dolocan A, Stenderup J, Velazquez AMV, Cahill J, Rasmussen M, Wang X, Min J, Zazula GD, Seguin-Orlando A, Mortensen C, Magnussen K, Thompson JF, Weinstock J, Gregersen K, Røed KH, Eisenmann V, Rubin CJ, Miller DC, Antczak DF, Bertelsen MF, Brunak S, Al-Rasheid KAS, Ryder O, Andersson L, Mundy J, Krogh A, Gilbert MTP, Kjær K, Sicheritz-Pontén T, Jensen LJ, Olsen JV, Hofreiter M, Nielsen R, SHAPIRO B, Wang J, Willerslev E 2013. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* 499:74–78. DOI: doi:10.1038/nature12323.
- Parks M, Lambert D 2015. Impacts of low coverage depths and post-mortem DNA damage on variant calling: a simulation study. *BMC Genomics* 16:19. DOI: 10.1186/s12864-015-1219-8.
- Parry RM, Jones W, Stokes TH, Phan JH, Moffitt RA, Fang H, Shi L, Oberthuer A, Fischer M, Tong W, Wang MD 2010. k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *The pharmacogenomics journal* 10:292–309. DOI: 10.1038/tpj.2010.56.
- Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, Berry DL, Holland KA, Weedn VW, Gill P, Holland MM 1997. A high observed substitution rate in the human mitochondrial DNA control region. *Nature genetics* 15:363–368. DOI: 10.1038/ng0497-363.
- Pääbo S 1985. Molecular cloning of Ancient Egyptian mummy DNA. *Nature* 314:644–645. DOI: 10.1038/314644a0.
- Peng M-S, ZHANG Y-P 2011. Inferring the Population Expansions in Peopling of Japan. *PloS one* 6:e21509–6. DOI: 10.1371/journal.pone.0021509.
- Pesole G, Gissi C, De Chirico A, Saccone C 1999. Nucleotide Substitution Rate of Mammalian Mitochondrial Genomes. *Journal of Molecular Evolution* 48:427–434. DOI: 10.1007/PL00006487.

- Pilli E, Modi A, Serpico C, Achilli A, Lancioni H, Lippi B, Bertoldi F, Gelichi S, Lari M, Caramelli D 2013. Monitoring DNA Contamination in Handled vs. Directly Excavated Ancient Human Skeletal Remains. *PloS one* 8:e52524. DOI: 10.1371/journal.pone.0052524.
- Posth C, Wißing C, Kitagawa K, Pagani L, van Holstein L, Racimo F, Wehrberger K, Conard NJ, Kind CJ, Bocherens H, Krause J 2017. Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. *Nature Communications* 8:ncomms16046. DOI: 10.1038/ncomms16046.
- Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellmann I, Johnson PLF, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Pääbo S 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505:43–49. DOI: doi:10.1038/nature12886.
- Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW Jr., Orlando L, Metspalu E, Karmin M, Tambets K, Rootsi S, Mägi R, Campos PF, Balanovska E, Balanovsky O, Khusnutdinova E, Litvinov S, Osipova LP, Fedorova SA, Voevoda MI, DeGiorgio M, Sicheritz-Pontén T, Brunak S, Demeshchenko S, Kivisild T, Villems R, Nielsen R, Jakobsson M, Willerslev E 2013. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* 505:87–91. DOI: 10.1038/nature12736.
- Renaud G, Hanghøj K, Willerslev E, Orlando L 2017. gargammel: a sequence simulator for ancient DNA. *Bioinformatics* 33:577–579. DOI: 10.1093/bioinformatics/btw670.
- Rohland N, Hofreiter M 2007. Ancient DNA extraction from bones and teeth. *Nature Protocols* 2:1756–1762. DOI: 10.1038/nprot.2007.247.
- 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* 74:5463–5467.
- Schafer JL, Olsen MK 2010. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research* 33:545–571. DOI: 10.1207/s15327906mbr3304_5.

- Schuenemann VJ, Peltzer A, Welte B, van Pelt WP, Molak M, Wang C-C, Furtwängler A, Urban C, Reiter E, Nieselt K, Teßmann B, Francken M, Harvati K, Haak W, Schiffels S, Krause J 2017. Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. *Nature Communications* 8:ncomms15694. DOI: 10.1038/ncomms15694.
- Skoglund P, Jakobsson M 2011. Archaic human ancestry in East Asia. *Proceedings of the National Academy of Sciences of the United States of America* 108:18301–18306. DOI: 10.1073/pnas.1108181108.
- Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Pääbo S, Krause J, Jakobsson M 2014. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences* 111:2229–2234. DOI: 10.1073/pnas.1318934111.
- Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB 2009. Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. *The American Journal of Human Genetics* 84:740–759. DOI: 10.1016/j.ajhg.2009.05.001.
- Tanaka M, Cabrera VM, González AM, Larruga JM, Takeyasu T, Fuku N, Guo L-J, Hirose R, Fujita Y, Kurata M, Shinoda K-I, Umetsu K, Yamada Y, Oshida Y, Sato Y, Hattori N, Mizuno Y, Arai Y, Hirose N, Ohta S, Ogawa O, Tanaka Y, Kawamori R, Shamoto-Nagai M, Maruyama W, Shimokata H, Suzuki R, Shimodaira H 2004. Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Research* 14:1832–1850. DOI: 10.1101/gr.2286304.
- Tang S, Huang T 2010. Characterization of mitochondrial DNA heteroplasmy using a parallel sequencing system. *BioTechniques* 48:287–296. DOI: 10.2144/000113389.
- Taylor RW, Turnbull DM 2005. Mitochondrial DNA mutations in human disease. *Nature Review Genetics* 6:389–402. DOI: 10.1038/nrg1606.
- The 1000 Genomes Project Consortium 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65. DOI: 10.1038/nature11632.
- Torroni A, Achilli A, Macaulay V, RICHARDS M, BANDELT H 2006. Harvesting the fruit of the human mtDNA tree. *Trends in Genetics* 22:339–345. DOI: 10.1016/j.tig.2006.04.001.

- Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17:520–525. DOI: 10.1093/bioinformatics/17.6.520.
- Ueno H, Nishigaki Y, Kong Q-P, Fuku N, Kojima S, Iwata N, Ozaki N, Tanaka M 2009. Analysis of mitochondrial DNA variants in Japanese patients with schizophrenia. *Mitochondrion* 9:385–393. DOI: 10.1016/j.mito.2009.06.003.
- 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 2015. PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Science International: Genetics Supplement Series* 5:e392–e394. DOI: 10.1016/j.fsigs.2015.09.155.
- 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* 36:419–424. 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* 34:1189–1194. DOI: 10.1002/humu.22356.
- Wang L, Oota H, Saitou N, Jin F, Matsushita T, Ueda S 2000. Genetic Structure of a 2,500-Year-Old Human Population in China and Its Spatiotemporal Changes. *Molecular biology and evolution* 17:1396–1400. DOI: 10.1093/oxfordjournals.molbev.a026422.
- Weissensteiner H, Forer L, Fuchsberger C, Schöpf B, Kloss-Brandstätter A, Specht G, Kronenberg F, Schönherr S 2016a. mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. *Nucleic Acids Research*:W64–W69. DOI: 10.1093/nar/gkw247.
- Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt H-J, Kronenberg F, Salas A, Schönherr S 2016b. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Research* 44:W58–W63. DOI: 10.1093/nar/gkw233.

- Zheng H-X, Yan S, Qin Z-D, Jin L 2012. MtDNA analysis of global populations support that major population expansions began before Neolithic Time. *Scientific Reports* 2:745. DOI: 10.1038/srep00745.
- Zheng H-X, Yan S, Qin Z-D, Wang Y, Tan J-Z, Li H, Jin L 2011. Major population expansion of East Asians began before neolithic time: evidence of mtDNA genomes. *PloS one* 6:e25835. DOI: 10.1371/journal.pone.0025835.
- 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* 11:924–928. DOI: 10.1016/j.mito.2011.08.005.

Acknowledgements

I would like to express the deepest appreciation to Dr. S. Ueda, Department Biological Sciences, Graduate School of Science, University of Tokyo, for his instruction and encouragement that were indispensable to accomplish this study. I have had the support and encouragement of Dr. W. Li, School of Medicine, Hangzhou Normal University, for her excellent advice and encouragement. I would like to express my gratitude to Dr. J. Gojobori, School of Advanced Sciences, SOKENDAI (The Graduate University for Advanced Studies) and Dr. M. Kumagai, Advanced Analysis Center, The National Agriculture and Food Research Organization (NARO), for their excellent advice about population genetics and bioinformatics analysis. Dr. F. Mizuno and Ms. M. Hayashi, Department of Legal Medicine, Toho University School of Medicine, give me constructive instructions and warm encouragement for experimental procedures. I would like to offer my special thanks to all members of Laboratory for Molecular Anthropology and Molecular Evolution (S. Ueda's Laboratory), for their advice and enthusiastic discussions about this study. Special thanks to my family for everything else in my life to accomplish this work.

The result obtained in Chapter 4 has been published in *PeerJ* (Ishiya and Ueda, 2017).