MitoIMP: A Computational Framework for Imputation of Missing Data in Low-Coverage Human Mitochondrial Genome

Bioinformatics and Biology Insights
Volume 13: 1–9
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1177932219873884
SAGE

Koji Ishiya[1], Fuzuki Mizuno[2], Li Wang[3] and Shintaroh Ueda[2,3,4]

[1]Computational Bio Big Data Open Innovation Lab (CBBD-OIL), National Institute of Advanced Industrial Science and Technology (AIST)—Waseda University, Tokyo, Japan. [2]Department of Legal Medicine, School of Medicine, Toho University, Tokyo, Japan. [3]School of Medicine, Hangzhou Normal University, Zhejiang, China. [4]Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo, Japan.

ABSTRACT: The incompleteness of partial human mitochondrial genome sequences makes it difficult to perform relevant comparisons among multiple resources. To deal with this issue, we propose a computational framework for deducing missing nucleotides in the human mitochondrial genome. We applied it to worldwide mitochondrial haplogroup lineages and assessed its performance. Our approach can deduce the missing nucleotides with a precision of 0.99 or higher in most human mitochondrial DNA lineages. Furthermore, although low-coverage mitochondrial genome sequences often lead to a blurred relationship in the multidimensional scaling analysis, our approach can correct this positional arrangement according to the corresponding mitochondrial DNA lineages. Therefore, our framework will provide a practical solution to compensate for the lack of genome coverage in partial and fragmented human mitochondrial genome sequences. In this study, we developed an open-source computer program, MitoIMP, implementing our imputation procedure. MitoIMP is freely available from https://github.com/omics-tools/mitoimp.

KEYWORDS: Missing data, low-coverage, high-throughput sequencing, mitochondrial DNA, ancient DNA

RECEIVED: July 27, 2019. **ACCEPTED:** August 13, 2019.

TYPE: Technical Advances

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by a Grant-in-Aid for JSPS Fellows 15J08490 (to K.I.), JSPS KAKENHI Grant Number 19K16246 (to K.I.), JSPS KAKENHI Grant Number 17K07588 (to F.M.), Zhejiang Provincial Natural Science Foundation of China Grant Number LZ13H020001 (to L.W.), and JSPS KAKENHI Grant Number 25291104 (to S.U.).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Koji Ishiya, Computational Bio Big Data Open Innovation Lab (CBBD-OIL), National Institute of Advanced Industrial Science and Technology (AIST)—Waseda University, Room 5-20, Bldg. 63, Nishi-Waseda Campus, 3-4-1 Ookubo, Shinjuku-ku, Tokyo 169-8555, Japan. Email: koji.ishiya@aist.go.jp

Introduction

The human mitochondrial genome has been used to study maternal phylogenetic relationships[1-6] and disease-related mutations,[7] as well as for DNA identification,[8] and it is an important genomic region not only in molecular anthropology but also in the medical and forensic fields. High-throughput genotyping and parallel sequencing technologies have made it possible to perform human mitochondrial genome population analyses on a massive scale.[9-11] These high-throughput technologies also allow mitochondrial genome analyses of ancient humans or archaic hominins, as well as modern humans.[12-17] However, the quantity of endogenous DNA in archeological remains is extremely low, which makes it difficult to obtain complete mitochondrial genome sequences. As an experimental method to deal with this problem, several target enrichment approaches have been employed,[18-21] but these problems are still not completely resolved in the cases of extremely degraded samples. The stability of endogenous DNA in postmortem resources is greatly affected by the environmental conditions, such as temperature, humidity, and pH.[22] A simulation study based on a statistical model suggested that the endogenous DNA in archeological samples is rapidly degraded in warm climates.[23] Actually, it is often difficult to obtain complete mitochondrial genome sequences from degraded samples excavated in areas with warm or hot climates, such as East or Southeast Asia.[24,25]

There are several practical approaches for analyzing incomplete data with missing values. One of these approaches is the listwise deletion approach,[26] which deletes categories, including missing values. However, this approach could limit comparisons with samples and discard some of the valuable data. Therefore, multiple imputation approaches have been proposed as methods to solve these problems.[27,28] These approaches fill the missing values with approximate or estimated values, which allows samples to be analyzed without deleting categories including missing values. The imputation against the human mitochondrial genome facilitated the performance of mitochondrial haplogroup prediction,[29] improved the statistical power in a genome-wide association study,[10] and also minimized the impact of missing nucleotides in a population genetics analysis.[30] However, these previous studies did not sufficiently assess the effect of the imputation on the worldwide mitochondrial DNA lineages, and as far as we know, there is no available imputation tool designed for the human mitochondrial genome. In this article, we propose a computational approach for deducing the missing nucleotides in partial human mitochondrial genome sequences and also assessing the effects of the imputation on the worldwide human mitochondria DNA lineages. Our computational approach will provide a practical solution to compensate for the missing data in the low-coverage human mitochondrial genomes.

Materials and Methods

Simulated data set

To assess the impact of our approach on the global mitochondrial DNA lineages, we generated artificial low-coverage human mitochondrial genome sequences for representative mitochondrial macro-haplogroups (A, B, C, D, E, F, G, H, HV, I, J, K, L0, L1, L2, L3, L4, L5, L6, M, N, O, P, Q, R, R0, S, T, U, V, W, X, Y, and Z) in PhyloTree (Build 17; http://www.phylotree.org).[31] Based on the mitochondrial genome sequences of these haplogroups, we simulated artificial next-generation sequencing (NGS) reads with ART (v. MountRainier-2016-06-05),[32] which is a simulation tool to generate synthetic NGS reads. In this study, we assumed paired-end reads based on the Illumina sequencer model in ART and mitochondrial genome coverage of ×10 to ×90. Next, these simulated short reads were aligned against the human mitochondrial reference sequence rCRS (revised Cambridge Reference Sequence),[33] using BWA (Burrows-Wheeler Aligner; v. 0.7.15).[34] We then obtained artificial low-coverage mitochondrial genome sequences from these aligned reads with MitoSuite (v. 1.0.9).[35] Finally, we applied our approach 500 times for each mitochondrial DNA lineage and investigated its precision. The precision of the imputation is computed as follows: TP/(TP + FP). True positive (TP) is the number of nucleotides imputed and validated. False positive (FP) is the number of nucleotides imputed but failed in validation.

Empirical Data Set

To assess the impact of our approach on the empirical sequencing data, we used several low-coverage human mitochondrial genome sequences from high-throughput sequencing data. These sequences were from Southeast Asian individuals dating from the Neolithic period through the Iron Age (4100-1700 years ago).[25] We downloaded these alignment reads according to the accession number PRJEB24939. Due to the poor DNA preservation in tropical conditions, most of the mitochondrial genomes derived from these ancient remains had incomplete partial sequences. Lipson et al[25] evaluated the ancient DNA authenticity and the exogenous contamination for each NGS library. To reduce the influence of postmortem misincorporations,[36] we clipped 2 bases from each end of the alignment reads, using BamUtil (v. 1.0.14).[37] Finally, we obtained the mitochondrial genome sequences from the clipped reads, using MitoSuite. In this study, the haplogroup assignments were performed according to the results of Lipson et al.[25]

Computational Deducing Approach

Our deducing approach couples the allele-sharing distance used for low-coverage sequencing data[38] with the k-nearest neighbor (kNN), which is a simple and effective supervised classification algorithm.[39] First, our procedure computes the allele-sharing distance to determine the pairwise distance among individual mitochondrial genome sequences, which is given by the following expression:

$$\delta_{i,j} = \frac{1}{S_{ij}} \sum_{n=1}^{S} d_{ij}^{n} \begin{cases} d_{ij}^{n} = 1\left(V_i^n \neq V_j^n\right) \\ d_{ij}^{n} = 0\left(V_i^n = V_j^n\right) \end{cases} \qquad (1)$$

where δ_{ij} is the allele-sharing distance between individuals i and j, S_{ij} is the number of polymorphic sites compared between individuals i and j, and V_i^n and V_j^n are the allele types at site n of individuals i and j, respectively. Finally, we obtain the following distance matrix D:

$$D = \begin{bmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,M} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{M,1} & \delta_{M,2} & \cdots & \delta_{M,M} \end{bmatrix} \qquad (2)$$

where M is the number of pairs in the individual mitochondrial genome sequences. Based on the distance matrix, the KNN sequences are selected, and then the missing positions are assigned to the most common alleles in the selected neighbors. To ensure the robustness of the imputed alleles, our framework also introduced the parameter f, which is the threshold frequency to determine the major alleles. In this study, we followed the parameter condition ($f = 0.7$, $k = 5$) used for kNN-based imputation procedures in Mizuno et al.[40] The condition is one of the most accurate combinations of parameter values in the previous research. Our approach also uses the reference population panel based on the complete mitochondrial genome sequences. Therefore, we obtained the worldwide human mitochondrial DNA sequences from PhyloTree Build 17 (http://www.phylotree.org).[31] These sequences were aligned using MAFFT (v. 7.407)[41] and oriented to the position of rCRS. Finally, we used 23 257 human mitochondrial DNA sequences as the default panel data set (ALL panel), including all present macro-haplogroups. The 1000 Genomes Project panel (phase 3)[42] is often used for imputation of genome-wide single-nucleotide polymorphisms (SNPs) in human population genetics, but the geographical region of the population and the mitochondrial genome lineage are limited. In this study, we designed the global panel data (ALL panel) with more than 4500 haplogroup lineages, covering known all macro-haplogroups. In addition, in our framework, there is no need to assign mitochondrial haplogroups of sequences in advance. Our approach does not force the design of panel data to be composed of the same population or maternal lineages and will be able to prevent false estimates from panels with biased population structure. To investigate whether the panel design can rescue lineage-specific mutations, we also used the worldwide macro-haplogroup panels consisting of the same macro-haplogroup lineages as the input mitochondrial genome.

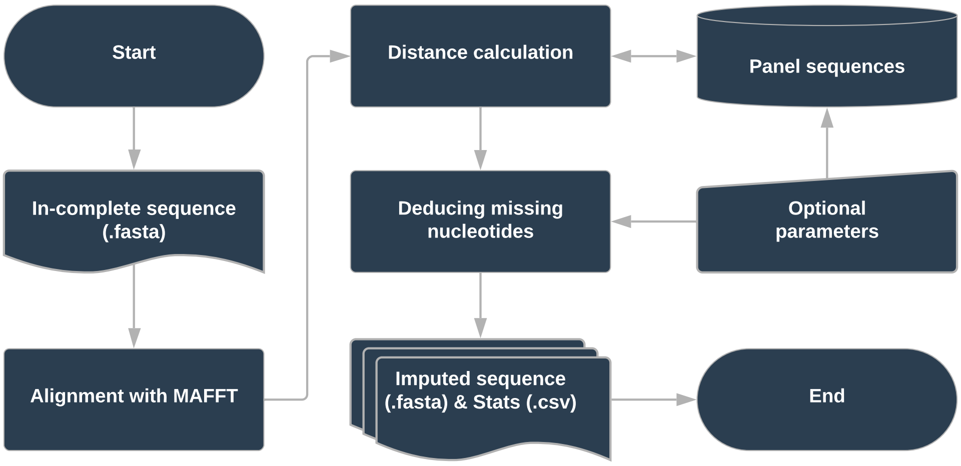
Start
In-complete sequence (.fasta)
Alignment with MAFFT
Distance calculation
Deducing missing nucleotides
Imputed sequence (.fasta) & Stats (.csv)
Panel sequences
Optional parameters
End

Figure 1. Flowchart of the imputation procedure.
This flowchart shows the imputation process in the MitoIMP program, implementing our approach. The rounded rectangles indicate the beginning and end of the procedure. The rectangles with a wavy base indicate the input and output files. The rectangular boxes represent processing or data manipulation. The MAFFT program is used to perform multiple alignments for the input sequence.

We also developed an open-source computer program, MitoIMP, implementing our imputation procedure (Figure 1). MitoIMP is freely available from https://github.com/omics-tools/mitoimp. Our computer program supports the standard FASTA as the input and output format. In our computational framework, the genomic position of a mitochondrial genome sequence is oriented to that of the human mitochondrial genome reference sequence rCRS with MAFFT.[41] Finally, MitoIMP outputs an imputed mitochondrial genome sequence in the FASTA format and provides a summary table for the imputation procedure.

Multidimensional Scaling Analysis

To assess the impact of our approach on an empirical low-coverage mitochondrial genome, we performed a multidimensional scaling (MDS) analysis using the mitochondrial genome sequences from 18 Southeast Asian ancient remains. The MDS analysis revealed a 2-dimensional relationship among the samples, based on the genetic distance of the mitochondrial genome sequences. The distances were determined based on the allele-sharing distance.[38] The MDS analysis was performed before and after our deducing approach. Pairwise allele-sharing distances between mitochondrial genomes were calculated using MitoIMP, and the eigenvector of the distance matrix was obtained by the "cmdscale" command, which is implemented in R.[43] The scatter plots were drawn using the ggplot2 package[44] in R.

Results

Application to simulated low-coverage sequencing data

To investigate the effectiveness of our approach to worldwide mitochondrial DNA lineages, we assessed the precision of the imputation for each maternal lineage. Our simulation results showed that missing bases can be estimated with a precision of 0.99 or higher in most of the macro-haplogroup lineages (Figure 2). We also found that the imputation precision for several macro-haplogroup L lineages was more affected by the loss of genome coverage than the other macro-haplogroups (Figure 2). Our approach showed decreased precision with the loss of genome coverage, but the FPs were only a few bases even where the mitochondrial genome lacked half of the genome coverage (Supplemental Figure S1). We also examined the region-by-region effects of our approach on the mitochondrial genome.

Although the imputation is uniformly performed on the entire mitochondrial genome, the effect of this procedure has not been evaluated so far for each region. Therefore, we also examined the precision of the imputation for each region of the mitochondrial genome and found that the effect of this procedure varied by region on the human mitochondrial genome (Figure 3). Although several nucleotide positions of the D-loop showed lower precision, our approach imputed the missing nucleotides with >0.9 high precision throughout most regions of the human mitochondrial genome.

Application to Empirical Low-Coverage Sequencing Data

We used partial mitochondrial genome sequences from 18 ancient South Asian individuals for the empirical deducing approaches. These partial mitochondrial genome sequences showed different genome coverages with 5831 to 16 558 nucleotides, which are equivalent to 35.2% to 99.9% of the covered region of the human mitochondrial reference sequence rCRS.[33] Especially, VN39 was not found in 10 731 positions, corresponding to about a 65% loss of the whole mitochondrial genome. To exclude intentional filling by geographical region–specific haplogroups, we used the worldwide haplogroup sequences (ALL panel) for deducing the missing positions (parameters: $k = 5, f = 0.7$). As a result, more than 99% of the missing positions were filled after the process (Table 1).

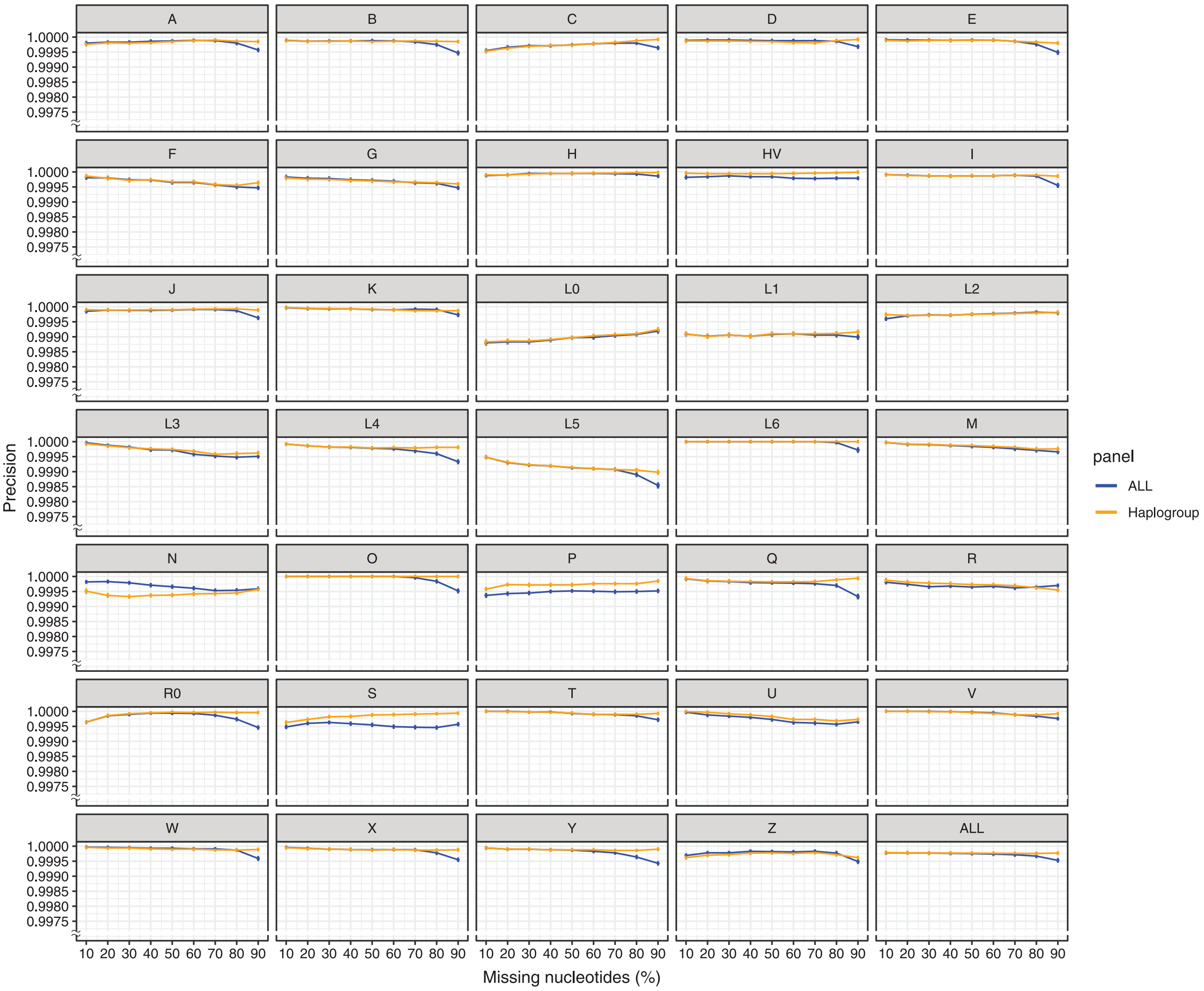

Figure 2. The assessment of imputation procedures for mitochondrial haplogroup lineages.
The vertical axis shows the precision in the assessment of the simulated imputation procedures. The horizontal axis indicates the percentage of missing nucleotides (10%-90%) in the partial mitochondrial genome sequences. Error bars indicate the standard error of the mean (SEM). The results in the case of the "ALL" panel including all macro-haplogroup lineages are indicated by the blue line, and those of the "Haplogroup" panel consisting of the same macro-haplogroup lineages are indicated by the orange line.

To investigate the relative relationships among individuals before and after the imputation, we performed an MDS analysis. The results of the MDS analysis without/with missing data showed different positional arrangements (Figure 4A). Among the partial mitochondrial genome sequences with missing data, there is no correspondence between the arrangement and their maternal lineages. The calculated distance matrix among the partial mitochondrial genome sequences did not correspond to the assigned haplogroups (Figure 4B). Our results also revealed that such partial mitochondrial genome sequences have a significant correlation between the calculated distance and the loss of coverage (Pearson correlation r^2 = .548, P = 2.384e−13; Supplemental Figure S2). On the other hand, the mitochondrial genome sequences imputed by our approach presented at least 2 clusters, belonging to macro-haplogroups B and M7 (Figure 4A and B). The individuals belonging to these 2 clusters showed different positional relationships before the filling procedure, but the imputed mitochondrial genome sequences seem to be plausibly arranged among the assigned macro-haplogroups. This result indicates that our approach can correct the unclear 2-dimensional arrangements among partial mitochondrial genome sequences.

Discussion

The effect of the loss of genome coverage

The missing data of mitochondrial genome sequences may be observed in empirical low-coverage sequencing data or genome-wide SNP array data. For example, ancient samples and museum samples have only trace amounts of endogenous DNA, and it is difficult to obtain complete mitochondrial genome sequences from such degraded samples. Although the incompleteness of the mitochondrial genome sequence may influence the calculation of the allele-sharing distances among individual sequences, our approach appears to be able to complement the deletion with the correct base in most of

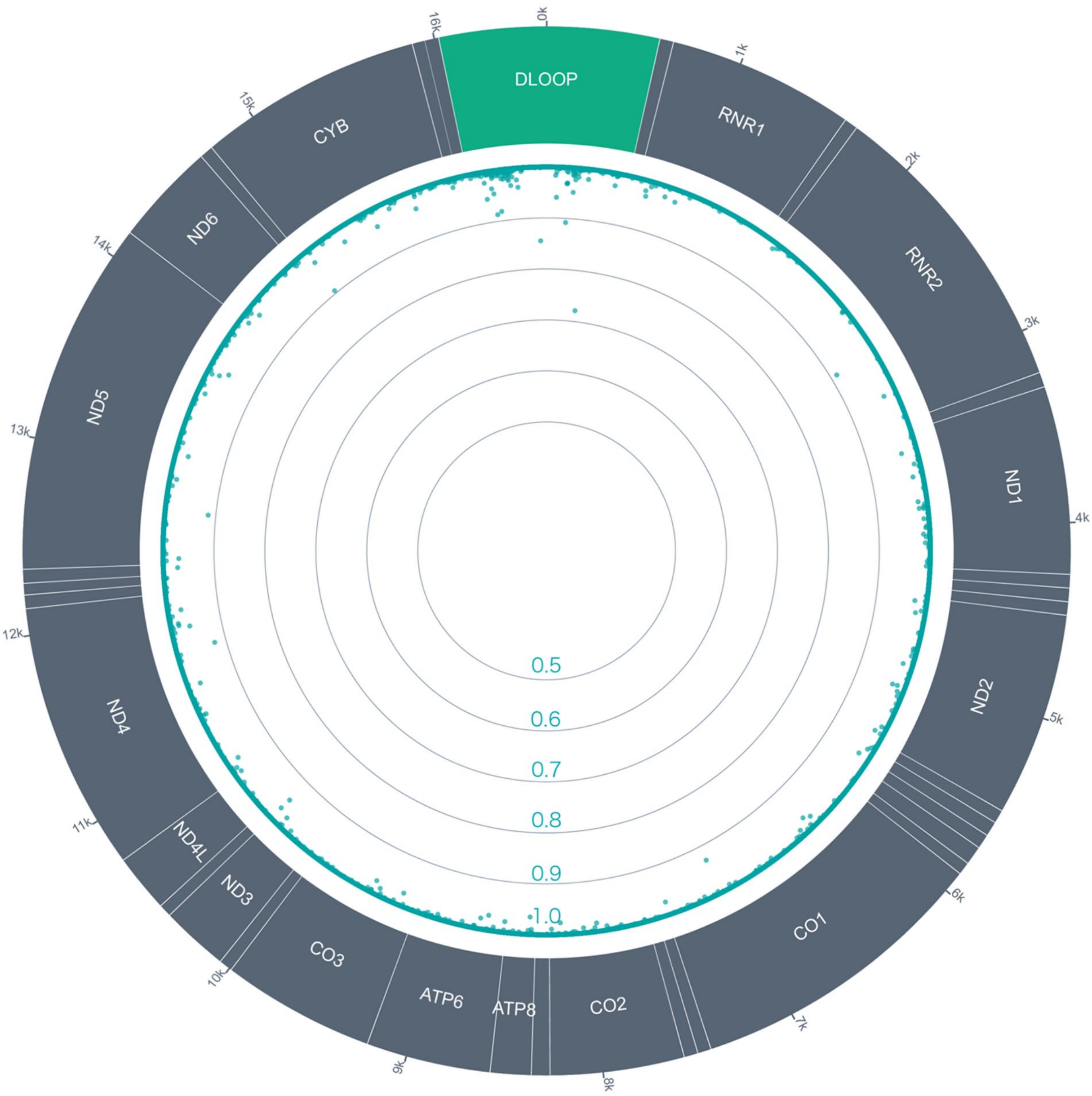

Figure 3. The assessment of imputation procedures across the human mitochondrial genome.
The scatter plot inside the circle shows the precision of the imputed nucleotides in 500 imputation trials, using worldwide haplogroup lineages. Protein- and RNA-coding regions are shown in gray and the noncoding region (D-loop) is shown in green. The abbreviations of the regions over 100 bp (base pairs) are written in white letters. The numerical value of the outer frame indicates the genomic position in the mitochondrial genome at intervals of 1000 bp. The lines inside the circle are graduated by 0.1 intervals, from 0.5 to 1.00.

the haplogroup lineages. However, the effect of the imputation approach on the missing data in the human mitochondrial genome sequence may vary among the regions. The human mitochondrial genome consists of 37 genes and a control region (D-loop). These 37 genes comprise 13 subunits of the respiratory chain complex genes, 22 transfer RNA genes, and 2 ribosomal RNA genes.[45] Therefore, we should consider the effect of our approach across the entire region of the human mitochondrial genome sequence. In this study, we found several positions (nucleotide positions: 152, 310, and 16 519) with a precision less than 0.9. These positions are located on the D-loop regions, which are known to have a higher mutation rate than that of the gene-coding regions in the human mitochondrial genome.[46] As private alleles are accumulated in the D-loop, it is not preferable to use many candidate sequences in the KNN algorithm. Therefore, we may decrease erroneous estimations in D-loop by reducing the number of the parameter *k*. Although the effect of low coverage was observed at only a few nucleotide positions, most of the missing positions were correctly filled across the human mitochondrial genome sequence. This result indicates that our approach using partially obtained sequences helps to infer a nearly complete mitochondrial genome sequence. Our method will be a useful approach to complement the missing parts of low-coverage mitochondrial genome sequences in which unidentified sites occur randomly.

Table 1. Empirical mitochondrial genome sequences used in this study.

SAMPLE	AGE (BP)	COUNTRY	HAPLOGROUP	MTGENOME COVERAGE (%)		MISSING NUCLEOTIDES		REFERENCES
				BEFORE IMP.	AFTER IMP.	BEFORE IMP.	AFTER IMP.	
AB40C	1890-1730	Cambodia	B5a1a	71.5	99.9	4726	14	Lipson et al[25]
BCES B16	2600-2400	Thailand	M72a	87.7	99.9	2031	4	Lipson et al[25]
BCES B27	3000-2800	Thailand	M74b2	68.8	99.9	5163	3	Lipson et al[25]
BCES B38	3200-3000	Thailand	B5a1a	72.8	99.9	4499	14	Lipson et al[25]
BCES B54	3200-3000	Thailand	B5a1c	70.1	99.9	4954	13	Lipson et al[25]
BCES B67	3500-3200	Thailand	F1f	44.7	99.9	9162	4	Lipson et al[25]
OAI1/S28	3200-2700	Myanmar	D4q	50.7	99.9	8163	1	Lipson et al[25]
OAI1/S29	3200-2700	Myanmar	D4h1c	44.2	99.9	9240	12	Lipson et al[25]
VN22	3835-3695	Vietnam	M13b	68.9	99.9	5161	6	Lipson et al[25]
VN29	3900-3600	Vietnam	M7b1a1	58.3	99.9	6915	3	Lipson et al[25]
VN31	3900-3600	Vietnam	No call (N.A.)	59.7	99.9	6684	6	Lipson et al[25]
VN33	3900-3600	Vietnam	B5a1a	93.7	99.9	1051	2	Lipson et al[25]
VN34	4080-3845	Vietnam	M7b1a1	99.1	100	141	0	Lipson et al[25]
VN37	3825-3635	Vietnam	M7b1a1	92.3	99.9	1276	3	Lipson et al[25]
VN39	3830-3695	Vietnam	M7b1a1	35.2	99.9	10 731	9	Lipson et al[25]
VN40	3820-3615	Vietnam	M74b	98.8	100	198	0	Lipson et al[25]
VN41	2100-1900	Vietnam	C7a	99.9	100	11	0	Lipson et al[25]
VN42	1995-1900	Vietnam	M8a2a	59.5	99.9	6705	8	Lipson et al[25]

The mtGenome coverage shows the percentages of mapped positions in the mitochondrial genome. Missing nucleotides represent the number of missing bases in the mitochondrial genome.

Understanding the Maternal Lineage for Accurate Imputation

To accurately fill the missing nucleotides in partial mitochondrial genome sequences, it is important to assess the impact of the imputation on each maternal lineage. Although the design of the population panel may lead to a bias of the filled nucleotides, a panel design with appropriate maternal lineages will improve the imputation performance. Actually, a previous study using 1500-year-old highly degraded samples showed that a closely related population panel (haplogroup A panel) can fill more missing sites, as compared with population panels including other maternal lineages.[40] However, this previous study assessed the impact of imputation on the macro-haplogroup A lineages, but that of the worldwide maternal lineages has not been verified. In this study, we used the population panels for the worldwide mitochondrial genome lineages and also tested the effect of these panels on the imputation performance. Our simulation result showed that the population panels consisting of closely related maternal lineages could improve the precision of the imputed nucleotides (Figure 2). Although the precision slightly varied due to the haplogroup lineages, our approach filled the missing sequences of the low-coverage mitochondrial genome with nucleotides with a precision of 0.99 or higher.

The imputation performance may also reflect the differences in the phylogenetic diversity of the haplogroup lineages. The most recent common ancestor of the human mitochondrial genome is dated back to approximately 150 000 to 250 000 years ago,[2,3,47] and the phylogenetic relationships involving modern human mitochondrial genomes can be largely divided into maternal lineages in sub-Saharan Africa (African haplogroup L) and non-Africa, with the exception of the "back-to-Africa" lineages.[48] Especially, it is well known that the human mitochondrial DNA lineages among Africans are more than twice as diverse as those of non-Africans.[49] Therefore, although the precision of imputation might decrease in several haplogroup L lineages, our approach will be especially effective for most of the non-African mitochondrial DNA lineages. Depending on the maternal lineages of samples, setting the parameters may be useful to reduce biased alleles' assignments caused by population structure. For

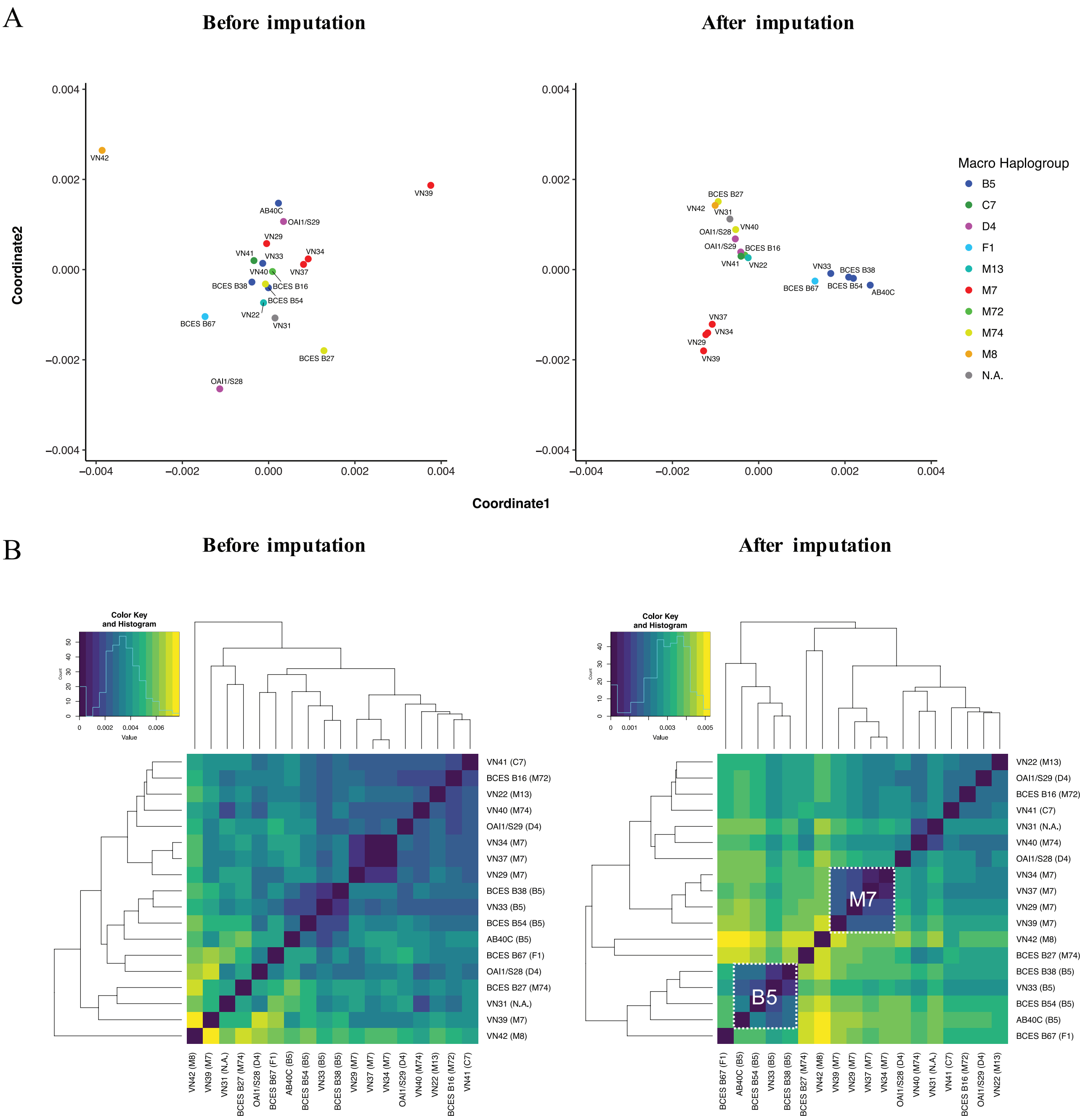

Figure 4. The relative relationships among individuals before and after the imputation. (A) The left and right figures show the results of the MDS analysis before and after the imputation procedure, respectively. The color scheme is according to macro-haplogroup lineages—B5 (blue), C7 (green), D4 (pink), F1 (light blue), M13 (cyan), M7 (red), M72 (light green), M74 (yellow), M8 (orange), N.A. (dark gray). (B) These figures are heat maps, based on the allele-sharing distance matrix among the empirical human mitochondrial genome sequences. Color keys of the distance values are shown on the upper left of each heatmap. The left figure shows the heatmap based on the allele-sharing distance before the imputation. The right figure shows the heatmap based on the allele-sharing distance after the imputation. Clusters of macro-haplogroups B5 and M7 are outlined by the white dashed lines.

example, population-specific alleles can be reduced by increasing the parameter *k* and the parameter *f*.

Impact of Deducing Approach for Poorly Preserved Fossil Remains

The full-length sequence of the human mitochondrial genome can facilitate the estimation of population dynamics,[47,50,51] and new findings surrounding the diversity of the human mitochondrial genome have been revealed by comparisons with the diversity of mitochondrial genomes of archaic hominins, such as Neanderthals.[52] However, extremely degraded archeological remains often contain less than 1% endogenous DNA, which makes it difficult to obtain the full-length mitochondrial genome. Environments with high temperatures and humidity rapidly fragment the endogenous DNA in postmortem samples; therefore, ancient genome research involving warm and humid environments, such as East or Southeast Asia, has been hindered. As approaches to examine such degraded ancient remains, experimental methods such as DNA extraction protocols specialized for ancient bones[53,54] and target enrichment methods[19,55,56] are often applied. However, these experimental methods are generally expensive, and when applied to highly degraded fossil remains, the complete mitochondrial genome sequence would not be successfully obtained. Therefore, short regions, such as hypervariable regions (HVRs), have often been used in previous studies on ancient DNA from warm and humid geographical regions.[57-59] However, the mitochondrial genome coverages obtained from ancient remains vary according to their DNA preservation. Actually, some empirical sequencing data from prehistoric Southeast Asian samples have large defects in the D-loop, including the HVRs. HV39 showed an especially low-coverage mitochondrial genome sequence, with more than half of the coverage missing. In addition, our simulation results suggest that the effect of the missing data on the control regions including HVRs is greater than that in other regions of the mitochondrial genome. Partial and incomplete mitochondrial DNA sequences might not be considered and neglected in detailed discussions. Therefore, we may also have limited insight into the analysis of low-coverage mitochondrial genome sequences including missing data. The filling of missing alleles for ancient genomes might be a challenging approach because ancient human genome resources are limited. However, our results show that our application can robustly impute most of the missing nucleotides in worldwide human mitochondrial DNA lineages (Figure 1 and Supplemental Figure S1). In this study, we have shown that the missing nucleotides can have a considerable impact on genetic distance calculations (Supplemental Figure S2). Our approach would be preferable to minimize the impact of missing nucleotides on the downstream analysis such as the MDS analysis.

In this study, we propose a computational deducing approach designed for such incomplete mitochondrial genome sequences. Recent improvements in NGS and target enrichment technologies have increased the opportunity to obtain randomly fragmented various partial sequences beyond specific regions of the human mitochondrial genome. Our approach imputes the missing regions using such partial sequences of the mitochondrial genome, without relying on the use of limited sequences or SNPs. Thus, this approach will work better for shotgun or target enrichment sequencing, as compared with genome-wide SNP arrays, amplicon sequencing, and polymerase chain reaction amplification. We also applied our approach to empirical low-coverage mitochondrial genome sequences. In the MDS analysis, our approach achieved the correct positional relationships by filling in the missing nucleotides. In particular, the individuals belonging to the macro-haplogroup lineages B5 and M7 appeared to be more closely clustered after the imputation. This result suggests that most of the missing regions were complemented by phylogenetically informative nucleotides. Our approach can be applied to low-coverage mitochondrial genome sequences from these empirical poorly preserved fossil remains and will prompt the use of partial mitochondrial DNA sequences that previously would have been discarded.

Acknowledgements

Discussions with our lab members greatly contributed to improving this study. Especially, we appreciate Dr Jun Gojobori and Dr Masahiko Kumagai for their expert advice.

Author Contributions

KI contributed to the design and analysis of the research, to implementation of the source code and to the writing of the manuscript. FM, LW and SU provided critical feedback and helped shape the research, analysis and manuscript.

ORCID iD

Koji Ishiya https://orcid.org/0000-0002-8715-0452

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. *Nature*. 1987;325:31-36.
2. Mishmar D, Ruiz-Pesini E, Golik P, et al. Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci U S A*. 2003;100:171-176.
3. Macaulay V, Hill C, Achilli A, et al. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science*. 2005;308:1034-1036.
4. Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H. Harvesting the fruit of the human mtDNA tree. *Trends Genet*. 2006;22:339-345.
5. Underhill PA, Kivisild T. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu Rev Genet*. 2007;41:539-564.
6. Behar DM, Villems R, Soodyall H, et al. The dawn of human matrilineal diversity. *Am J Hum Genet*. 2008;82:1130-1140.
7. Taylor RW, Turnbull DM. Mitochondrial DNA mutations in human disease. *Nat Rev Genet*. 2005;6:389-402.
8. Holland MM, Parsons TJ. Mitochondrial DNA sequence analysis—validation and use for forensic casework. *Forensic Sci Rev*. 1999;11:21-50.
9. Sequeira A, Martin M, Rollins B, et al. Mitochondrial mutations and polymorphisms in psychiatric disorders. *Front Genet*. 2012;3:103.
10. Hudson G, Gomez-Duran A, Wilson IJ, Chinnery PF. Recent mitochondrial DNA mutations increase the risk of developing common late-onset human diseases. *PLoS Genet*. 2014;10:e1004369.
11. Cai N, Li Y, Chang S, et al. Genetic control over mtDNA and its relationship to major depressive disorder. *Curr Biol*. 2015;25:3170-3177.
12. Green RE, Malaspinas A-S, Krause J, et al. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*. 2008;134:416-426.
13. Krause J, Fu Q, Good JM, et al. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*. 2010;464:894-897.
14. Meyer M, Fu Q, Aximu-Petri A, et al. A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature*. 2013;505:403-406.
15. Coia V, Cipollini G, Anagnostou P, et al. Whole mitochondrial DNA sequencing in Alpine populations and the genetic history of the Neolithic Tyrolean Iceman. *Sci Rep*. 2016;6:18932.
16. Llamas B, Fehren-Schmitz L, Valverde G, et al. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci Adv*. 2016;2:e1501385.

17. Schuenemann VJ, Peltzer A, Welte B, et al. Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. *Nat Commun*. 2017;8:15694.
18. Maricic T, Whitten M, Paabo S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE*. 2010;5:e14004.
19. Carpenter ML, Buenrostro JD, Valdiosera C, et al. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *Am J Hum Genet*. 2013;93:852-864.
20. Kihana M, Mizuno F, Sawafuji R, Wang L, Ueda S. Emulsion PCR-coupled target enrichment: an effective fishing method for high-throughput sequencing of poorly preserved ancient DNA. *Gene*. 2013;528:347-351.
21. Enk JM, Devault AM, Kuch M, Murgha YE, Rouillard JM, Poinar HN. Ancient whole genome enrichment using baits built from modern DNA. *Mol Biol Evol*. 2014;31:1292-1294.
22. Hofreiter M, Paijmans JLA, Goodchild H, et al. The future of ancient DNA: technical advances and conceptual shifts. *BioEssays*. 2015;37:284-293.
23. Allentoft ME, Collins M, Harker D, et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc Biol Sci*. 2012;279:4724-4733.
24. McColl H, Racimo F, Vinner L, et al. The prehistoric peopling of Southeast Asia. *Science*. 2018;361:88-92.
25. Lipson M, Cheronet O, Mallick S, et al. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. *Science*. 2018;361:92-95.
26. Allison PD. *Missing Data*. Thousand Oaks, CA: SAGE; 2001.
27. Meng X-L. Multiple-imputation inferences with uncongenial sources of input. *Statist Sci*. 1994;9:538-558.
28. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivar Behav Res*. 2010;33:545-571.
29. Biffi A, Anderson CD, Nalls MA, et al. Principal-component analysis for assessment of population stratification in mitochondrial medical genetics. *Am J Hum Genet*. 2010;86:904-917.
30. Barbieri C, Vicente M, Rocha J, Mpoloka SW, Stoneking M, Pakendorf B. Ancient substructure in early mtDNA lineages of southern Africa. *Am J Hum Genet*. 2013;92:285-292.
31. van Oven M. PhyloTree Build 17: growing the human mitochondrial DNA tree. *Foren Sci Int*. 2015;5:e392-e394.
32. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics*. 2012;28:593-594.
33. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet*. 1999;23:147-147.
34. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25:1754-1760.
35. Ishiya K, Ueda S. MitoSuite: a graphical tool for human mitochondrial genome profiling in massive parallel sequencing. *PeerJ*. 2017;5:e3406.
36. Briggs AW, Stenzel U, Johnson PLF, et al. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci U S A*. 2007;104:14616-14621.
37. Jun G, Wing MK, Abecasis GR, Kang HM. An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. *Genome Res*. 2015;25:918-925.
38. Malaspinas A-S, Tange O, Moreno-Mayar JV, et al. bammds: a tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS). *Bioinformatics*. 2014;30:2962-2964.
39. Cover TM, Hart PE. Nearest neighbor pattern classification. *IEEE Trans Inform Theory*. 1967;13:21-27.
40. Mizuno F, Kumagai M, Kurosaki K, et al. 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. *J Hum Genet*. 2017;62:631-635.
41. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30:772-780.
42. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526:68-74.
43. R Core Team. *R: A Language and Environment for Statistical Computing*; 2018. http://softlibre.unizar.es/manuales/aplicaciones/r/fullrefman.pdf.
44. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer; 2016.
45. Anderson S, Bankier AT, Barrell BG, et al. Sequence and organization of the human mitochondrial genome. *Nature*. 1981;290:457-465.
46. Soares P, Ermini L, Thomson N, et al. Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet*. 2009;84:740-759.
47. Atkinson QD, Gray RD, Drummond AJ. Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa. *Proc Biol Sci*. 2009;276:367-373.
48. Cabrera VM, Marrero P, Abu-Amero KK, Larruga JM. Carriers of mitochondrial DNA macrohaplogroup L3 basal lineages migrated back to Africa from Asia around 70,000 years ago. *BMC Evol Biol*. 2018;18:98.
49. Ingman M, Kaessmann H, Paabo S, Gyllensten U. Mitochondrial genome variation and the origin of modern humans. *Nature*. 2000;408:708-713.
50. Zheng H-X, Yan S, Qin Z-D, et al. Major population expansion of East Asians began before neolithic time: evidence of mtDNA genomes. *PLoS ONE*. 2011;6:e25835.
51. Zheng H-X, Yan S, Qin Z-D, Jin L. MtDNA analysis of global populations support that major population expansions began before Neolithic Time. *Sci Rep*. 2012;2:745.
52. Posth C, Wissing C, Kitagawa K, et al. Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. *Nat Commun*. 2017;8:16046.
53. Rohland N, Hofreiter M. Ancient DNA extraction from bones and teeth. *Nat Protocol*. 2007;2:1756-1762.
54. Gamba C, Hanghøj K, Gaunitz C, et al. Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. *Mol Ecol Resour*. 2015;16:459-469.
55. Mamanova L, Coffey AJ, Scott CE, et al. Target-enrichment strategies for next-generation sequencing. *Nat Meth*. 2010;7:111-118.
56. Hawkins MTR, Hofman CA, Callicrate T, et al. In-solution hybridization for mammalian mitogenome enrichment: pros, cons and challenges associated with multiplexing degraded DNA. *Mol Ecol Resour*. 2015;16:1173-1188.
57. Wang L, Oota H, Saitou N, Jin F, Matsushita T, Ueda S. Genetic structure of a 2,500-year-old human population in China and its spatiotemporal changes. *Mol Biol Evol*. 2000;17:1396-1400.
58. Igawa K, Manabe Y, Oyamada J, et al. Mitochondrial DNA analysis of Yayoi period human skeletal remains from the Doigahama site. *J Hum Genet*. 2009;54:581-588.
59. Zhao Y-B, Zhang Y, Zhang Q-C, et al. Ancient DNA reveals that the genetic structure of the Northern Han Chinese was shaped prior to 3,000 years ago. *PLoS ONE*. 2015;10:e0125676.