Novel approach for accurate detection of contaminating human mitochondrial DNA in next-generation sequencing data

Koji Ishiya¹, and Shintaroh Ueda²,³,⁴

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

E-mail: ¹koji.ishiya@aist.go.jp; ²sueda@bs.s.u-tokyo.ac.jp

Abstract. DNA molecules are degraded after the death of an organism. However, the degree and rate of DNA degradation enormously vary depending on environmental conditions, such as temperature or humidity, which greatly affect DNA preservation. Most samples excavated in warm, humid, or dry areas are often poorly preserved samples with from <0.1%-1% endogenous DNA. In these degraded samples, the contamination by exogenous DNA remains a potential challenge, no matter how much effort is made to prevent it. For an accurate DNA sequence analysis, quality control must be thoroughly performed, using the mitochondrial DNA as an indicator of exogenous DNA contamination. Here, we propose a practical approach for detecting exogenous human mitochondrial macro haplogroups, and discuss the effectiveness of this approach using simulated data. Our approach is based on the Bayes classification, which is a supervised machine learning algorithm, and it can detect a contaminating macro haplogroup in high-throughput sequencing data. This approach can help validate the quality of high-throughput sequencing data from possibly contaminated or degraded human samples.

1. Introduction

Human mitochondrial genome research has been performed on worldwide regional and indigenous populations, and these analyses have offered new insights into the geographical distribution and evolutionary history of human maternal lineages [1, 2, 3, 4, 5, 6, 7], disease-related mutations [8], and DNA identification [9]. Therefore, the human mitochondrial genome is a meaningful target in the molecular anthropology, medical, and forensic medicine fields. The human mitochondrial genome also has a key role as a source of obtainable genetic profiles for highly degraded samples. The mitochondrial genome is easier to access than the nuclear genome, because the mitochondrial genome exists in several hundreds to thousands of copies in a cell. Therefore, the human mitochondrial genome has also been studied in archeological remains. Recent comparative analyses of archaic hominins and modern humans have also generated new findings surrounding the diversity of the hominin mitochondrial genome [10]. In comparison with the previous approach combining the conventional PCR and Sanger sequencing platforms, high-throughput sequencing technologies have made it possible to perform genome-scale analyses of archeological or degraded samples [11, 12]. However, although these samples have been carefully
handled, unexpected risks of exogenous DNA contamination remain [13, 14]. For example, exogenous DNA contamination may occur not only during excavation [15] but also by the use of contaminated reagents [16] and laboratory instruments. DNA contamination from genetically distant and not-closely related species may be computationally removed because of differences in sequence homology among the species involved; however, it would be difficult to exclude high-homology DNA sequences from other individuals or closely related species. DNA contamination derived from closely related species may cause erroneous bias in phylogenetic or population genetic analyses [17]. The contamination of DNA fragments derived from a genetically distant macro-haplogroup may also have an impact on the establishment of phylogenetic relationships and the detection of heteroplasmy or nuclear mitochondrial pseudogene (NUMT) sequences [18]. The inconsistency percentage (mismatch rate) across mitochondrial haplogroup-defining sites is often used as a measure of the degree of exogenous DNA contamination. However, this method cannot identify the contaminating mitochondrial haplogroup in the sequencing data. This percentage is only quantitative, and not a qualitative evaluation of exogenous DNA contamination.

Several studies of ancient genomes have been reported, using well-preserved archeological samples excavated in comparatively stable climate environments such as caves [19, 20, 21]. Unfortunately, widely applicable approaches have not been established for highly-degraded samples with only < 1% of the original amount of endogenous DNA [22, 23]. There are experimental [24] and computational [25] methods to selectively acquire damaged DNA sequences, and these methods can be effective for such sequences with the post-mortem deamination of cytosine to uracil. However, these methods probably ignore large quantities of endogenous DNA fragments because of the damage-selective filtering, and may not be suitable for modern contaminated samples or endogenous DNA fragments with less post-mortem damage (PMD). For samples with low endogenous DNA content or non-damaged DNA, fewer losses through filtering would be preferable.

By utilizing the existing worldwide human mitochondrial genome sequences, we applied a supervised algorithm to analyze contaminated, high-throughput sequencing data. Our approach can accurately identify a contaminating human mitochondrial haplogroup among the representative human mitochondrial macro-haplogroups. This procedure can serve as a qualitative evaluation approach to establish more accurate human mitochondrial genome analyses in the NGS data of potentially contaminated DNA samples.

2. Experimental

2.1. Data collection of learning data

We obtained 24,275 worldwide human mitochondrial genome sequences from the databases of human mitochondrial sequences considered for tree construction in PhyloTree build 17 [26] and then used 23,257 full-length sequences, which could determine the mitochondrial haplogroups. These mitochondrial genomes were oriented to the position of the human mitochondrial reference sequence rCRS [27] using MAFFT [28], and then used as the panel (learning) data.

2.2. Artificially contaminated test data

In the human mitochondrial genome, contamination by sequences belonging to the exact same macro-haplogroup might not cause large biases in the phylogenetic relationship analysis. Distinguishing between distinctly different maternal lineages, such as modern human and Neanderthal, is important to determine phylogenetic relationships, as reported previously Skoglund:2014ee. Therefore, assuming data contamination with different mitochondrial DNA lineages, we used simulated NGS reads for eight representative worldwide macro-haplogroup sequences (I 0, I 3, M, D, N, A, R, and B). The percentages of contaminated reads derived from other haplogroup lineages were set between 5% and 30%. NGS reads from archeological samples
often contain post-mortem damages (PMD), such as the deamination of cytosine to uracil. To investigate the effects of such damage of short read sequences on the detection accuracy, we assumed two simulated conditions; namely, the post-mortem damage (PMD) model and the no-damage (ND) model. In the PMD model, we simulated post-mortem damaged NGS reads, using an ancient DNA (aDNA) simulator, gargammel [29]. The PMD model assumes the case of contamination by modern (non-damaged) DNA samples. In the ND model, there is no DNA damage in the simulated NGS reads. We also assumed an average coverage depth of 10-50× in all simulated data. The detection accuracy of a contaminating macro-haplogroup was computed as follows: TP / (TP + FP). True positive (TP) is the number of macro-haplogroups detected and validated, and false positive (FP) is the number of macro-haplogroups that were detected but failed validation.

2.3. Inference procedure
In this study, we propose a detection approach for contaminating mitochondrial macro-haplogroups based on the Bayes classification algorithm [30]. Our algorithm was designed with the aim of detecting DNA fragments derived from totally different maternal lineages (macro-haplogroups), which may greatly affect the genetic profile of individuals. As shown in Figure 1, the flowchart of the detection procedure is as follows: (1) The most likely macro-haplogroup included in the alignment data is estimated on the basis of the haplogroup-defining sites, as a prior procedure. This procedure may also be conducted by using existing haplogroup assignment tools [31, 32]. In this study, we refer to the most likely macro-haplogroup as the “Major Haplogroup”. (2) All expected variants differing from the “Major Haplogroup” lineages are obtained from the panel data. (3) Observed variants (polymorphic sites) against the “Major Haplogroup” sequences are extracted from the alignment data. To reduce false positive detection derived from amplification and sequencing errors as much as possible, variants with a minor allele frequency (MAF) of >1% and a PhredScore of >20 are used. In most of the ancient DNA samples without DNA damage repair procedures, mutations derived from post-mortem deamination accumulate at each read end [33]. Assuming the use of such DNA samples, we also performed an optional approach, PMD-clipped, to exclude the C to T, and G to A transition-type substitutions for 10 bases from both ends of the NGS reads.

3. Theory and calculation
3.1. Bayesian classification
Based on the obtained variant information, posterior probabilities are calculated for all haplogroups differing from the “Major Haplogroup”, using equations (1) to (3), and the haplogroup with the highest a posteriori probability is estimated as a possible contaminating haplogroup (“Minor Haplogroup”). We further tested the accuracy for eight representative worldwide macro-haplogroups (L0, L3, M, D, N, A, R, and B).

\[
P(y|X) = \frac{P(X|y)P(y)}{P(X)} \propto P(X|y)P(y) \quad (1)
\]

where \(X\) is a pattern of observed polymorphisms and \(y\) is a type of haplogroup. \(P(y|X)\) shows the posterior probability of haplogroup \(y\) when a pattern of polymorphisms \(X\) is observed. As the observed data \(X\) are the same for all haplogroups, \(P(y|X)\) is defined as being proportional to \(P(X|y) \propto P(y)\). \(P(y)\) is the frequency of the haplogroup in the panel data, and \(P(X|y)\) is the likelihood of having a combination of observed polymorphisms in the haplogroup \(y\). \(P(X|y)\) can be obtained by the product of the likelihood of each observed polymorphic site.

\[
P(X|y) = \frac{T(y, X_i)}{\sum_{X' \in V} T(y, X')} \approx \frac{T(y, X_i) + 1}{\sum_{X' \in V} T(y, X') + |V|} \quad (2)
\]
where \( X_i \) indicates a substitution \( X \) at a polymorphic site \( i \), and \( V \) is the total number of polymorphic sites. \( T(y, X_i) \) is the number of occurrences of \( X_i \) in a haplogroup \( y \). When an actually-observed polymorphism in the alignment data was unexpected in a haplogroup \( y \), \( P(X_i|y) \) becomes zero. Therefore, additive smoothing (Laplace smoothing) [34] is performed to deal with the zero-frequency problem, as shown in (2).

\[
\arg \max \left( \log(P(y) \prod_{i=1}^{n} P(X_i|y)) \right) = \arg \max \left( \log P(y) + \sum_{i=1}^{n} \log P(X_i|y) \right) \tag{3}
\]

The posterior probability of each haplogroup is calculated by using the prior probability of each haplogroup and the product of the likelihood of each observed polymorphism. To prevent underflow calculation errors, the posterior probability is logarithmically transformed, as shown in (3). The posterior probabilities for all haplogroups in the panel data are calculated, and the haplogroup with the highest a posteriori probability is estimated as a possible contaminating haplogroup.

4. Results and Discussion

4.1. Impact of sequencing depth on detection accuracy

In this study, the detection accuracy of a contaminated macro-haplogroup was calculated for simulated mitochondrial genome sequencing data with different sequencing depths (10-50×) and percentages of contaminating reads (5%-30%). When the average depth was 10×, the detection accuracy was 0.714 in the case of 5% contaminating reads in all reads, whereas when the percentage of contaminating reads was 10%, this figure changed to 0.929 (Fig. 2). The decrease in the detection accuracy at a 10× low cover might be caused by variants derived from different macro-haplogroups that were not observed, because the total number of reads obtained was small. Simulated mitochondrial sequencing data with \( \geq 20 \times \) depths showed higher detection accuracy (>0.964), even in the case of 5% contaminating reads in all reads (Fig. 2).

4.2. Impact of post-mortem damage on detection accuracy

To test the effect of the post-mortem damage on the detection accuracy, we tried our approach to damaged (PMD) and non–damaged (ND) datasets. Comparing the detection accuracy in the PMD and ND models, particularly at a low depth of coverage, a change in the detection accuracy was observed depending on the presence or absence of DNA damage. For example, under the ND model at 10× depth, the detection accuracy was 0.941, which dropped to 0.873 under the PMD model (Fig. 3, Table 1). We believe that this decrease in the accuracy was due to unexpected erroneous polymorphisms and variants derived from deamination (C>T, G>A) in short read sequences. To prevent false positive (FP) detection because of these deamination events, we also evaluated an optional procedure (PMD–clipped) to account for the deamination patterns. Post-mortem deamination is often observed near the ends of DNA fragments [33]. In this procedure, deamination-like variants (C>T or G>A), which are observed within 10 bases of both read termini, are skipped to call variants. As a result, this approach improved the detection accuracy to 0.909 at a 10 depth of coverage, even for the PMD model dataset (Fig. 3, Table 1). In addition, the decrease of the detection accuracy in the high coverage data of \( \geq 30 \times \) was slighter than that in the lower coverage sequencing data (10-20×) (Fig. 3, Table 1), which is consistent with the result of a previous simulation study, in which the increase in depth suppressed the decrease in the accuracy of variant calling under the PMD model [35].
Figure 1. Flowchart of the detection of contaminating macro–haplogroups in NGS data. This figure shows the outline for detecting a contaminating macro-haplogroup in short read data obtained by NGS. The macro-haplogroup comprising the highest percentage in the pool of reads is defined as the “Major Haplogroup”. Existing haplogroup assignment tools are used to perform this procedure. Variants differing from the “Major Haplogroup” are then extracted from the panel data and used by the classification algorithm for learning. Finally, the observed variants in the pool of reads are used to detect a contaminating macro–haplogroup by the learned classifier.
Figure 2. Detection accuracy versus the percentage of contaminating reads. The horizontal axis represents the average depth of short reads across the human mitochondrial genome, and the vertical axis represents the average value of detection accuracy. The line colors show the percentages (5%-30%) of short reads derived from a contaminating haplogroup in the total mapped reads. As for the highly contaminated data of \( \geq 15\% \), their colored lines will appear to overlap due to the almost identical detection accuracy.

Table 1. Detection accuracy of a contaminating macro-haplogroup.

| Coverage | Detection Accuracy | Conditions  |
|----------|--------------------|-------------|
| 10       | 0.941              | ND          |
| 20       | 0.997              | ND          |
| 30       | 0.997              | ND          |
| 40       | 0.994              | ND          |
| 50       | 1.000              | ND          |
| 10       | 0.873              | PMD         |
| 20       | 0.937              | PMD         |
| 30       | 0.964              | PMD         |
| 40       | 0.964              | PMD         |
| 50       | 0.972              | PMD         |
| 10       | 0.909              | PMD-clipped |
| 20       | 0.956              | PMD-clipped |
| 30       | 0.984              | PMD-clipped |
| 40       | 0.992              | PMD-clipped |
| 50       | 0.996              | PMD-clipped |
Figure 3. Impact of deamination on the detection accuracy. This figure shows the impact of the presence/absence of the deamination of cytosine to uracil, which is post-mortem damage (PMD), on the detection accuracy. The horizontal axis represents the average depth of short reads across the human mitochondrial genome, and the vertical axis represents the average value of the detection accuracy. The green line is the no-damage (ND) model without deamination, and the blue line is the PMD model assuming the deamination. The orange line indicates the detection accuracy in the PMD model, but with clipping the deamination of cytosine to uracil near the ends of DNA fragments (PMD–clipped).

4.3. Conclusion
In fields dealing with human DNA, such as medicine, forensic medicine, and molecular anthropology, exogenous human DNA contamination reduces the reliability of the analysis results. In this study, we proposed a novel approach for the detection of contaminated human mitochondrial DNA, based on mitochondrial macro–haplogroups in public DNA databases. In our approach, mitochondrial DNA polymorphisms are used to estimate the contaminated mitochondrial macro–haplogroups by the Bayesian classification algorithm. In this study, we verified the detection accuracy of a contaminated macro–haplogroup using simulated data with various depths of coverage, contamination levels, and the presence/absence of post-mortem deamination. Our results showed that the contaminating macro–haplogroups can be identified with high accuracy. This approach may be useful for quality control in scientific fields dealing with human DNA sequences obtained from next–generation sequencing.

Acknowledgments
Discussions with our lab members greatly contributed to improving this study. 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.), and JSPS KAKENHI Grant Number 25291104 (to S.U.).

Conflict of interest
The Authors declare that there is no conflict of interest.
References

[1] Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. *Nature*. 1987;325(6099):31–36.

[2] Bonatto SL, Salzano FM. 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*. 1997;94(5):1866–1871.

[3] Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, et al. Natural selection shaped regional mtDNA variation in humans. *Proceedings of the National Academy of Sciences of the United States of America*. 2003;100(1):171–176.

[4] Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, et al. Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes. *Science*. 2005;308(5724):1034–1036.

[5] Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H. Harvesting the fruit of the human mtDNA tree. *Trends in Genetics*. 2006;22(6):339–345.

[6] Underhill PA, Kivisild T. Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations. *Annual Review of Genetics*. 2007;41(1):339–356.

[7] Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, et al. The Dawn of Human Matrilineal Diversity. *The American Journal of Human Genetics*. 2003;82(5):1130–1140.

[8] Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, et al. Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes. *Science*. 2005;308(5724):1034–1036.

[9] Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H. Harvesting the fruit of the human mtDNA tree. *Trends in Genetics*. 2006;22(6):339–345.

[10] Underhill PA, Kivisild T. Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations. *Annual Review of Genetics*. 2007;41(1):339–356.

[11] Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, et al. The Dawn of Human Matrilineal Diversity. *The American Journal of Human Genetics*. 2003;82(5):1130–1140.

[12] Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, et al. Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes. *Science*. 2005;308(5724):1034–1036.

[13] Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H. Harvesting the fruit of the human mtDNA tree. *Trends in Genetics*. 2006;22(6):339–345.

[14] Underhill PA, Kivisild T. Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations. *Annual Review of Genetics*. 2007;41(1):339–356.

[15] Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, et al. The Dawn of Human Matrilineal Diversity. *The American Journal of Human Genetics*. 2003;82(5):1130–1140.

[16] Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, et al. Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes. *Science*. 2005;308(5724):1034–1036.

[17] Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H. Harvesting the fruit of the human mtDNA tree. *Trends in Genetics*. 2006;22(6):339–345.

[18] Underhill PA, Kivisild T. Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations. *Annual Review of Genetics*. 2007;41(1):339–356.

[19] Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, et al. The Dawn of Human Matrilineal Diversity. *The American Journal of Human Genetics*. 2003;82(5):1130–1140.

[20] Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, et al. Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes. *Science*. 2005;308(5724):1034–1036.

[21] Torroni A, Achilli A, Macaulay V, Richards M, Bandelt H. Harvesting the fruit of the human mtDNA tree. *Trends in Genetics*. 2006;22(6):339–345.
[27] Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature genetics*. 1999;23(2):147–147.

[28] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*. 2013;30(4):772–780.

[29] Renaud G, Hanghøj K, Willerslev E, Orlando L. gargamnel: a sequence simulator for ancient DNA. *Bioinformatics*. 2017;33(4):577–579.

[30] Mitchell TM. Machine Learning. McGraw-Hill Series in Computer Science; 1997.

[31] Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt HJ, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Research*. 2016;44(W1):W58–W63.

[32] Ishiya K, Ueda S. MitoSuite: a graphical tool for human mitochondrial genome profiling in massive parallel sequencing. *PeerJ*. 2017;5(2):e3406.

[33] Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prufer K, et al. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences*. 2007;104(37):14616–14621.

[34] Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press; 2008.

[35] Parks M, Lambert D. Impacts of low coverage depths and post-mortem DNA damage on variant calling: a simulation study. *BMC Genomics*. 2015;16(1):19.