Imputation and kNN-based Haplotype Refinement of Simulated Ancient Mitochondrial Genomes

Plummer N.[1], Cozzarelli A.[1], Fang Y.[1], Shamoon-Pour M.[1,2], Spake L.[1] Emery M.V.[1,3,4]

1 Department of Anthropology, Binghamton University, 4400 Vestal Pkwy E, Binghamton, NY, 13902
2 First-year Research Immersion, Binghamton University,   4400 Vestal Pkwy E, Binghamton, NY, 13902
3 School of Human Evolution and Social Change, Arizona State University, 900 Cady Mall, Tempe, AZ, 85281
4 Center for Evolution and Medicine, Arizona State University, Life Sciences C, 427 East Tyler Mall, Tempe, AZ 85281

INTRODUCTION

Ancient DNA (aDNA) is characterized by low DNA yields and post-mortem degradation, resulting in ultra-low coverage genomes[1,2]. Imputation methods, even when adapted for haploid data (i.e., mitochondrial DNA, mtDNA), reflect underlying assumptions derived from diploid genome imputation methods. The distinct evolutionary patterns of mtDNA, including lack of recombination and haplogroup-specific variation[3], create unique challenges.

We constructed an imputation workflow (**MAVEN**) that leverages a standard **Hidden Markov Model[4,5,6]** (HMM, Minimac4[7]), with a **k-Nearest Neighbor[8]** (kNN) supervised machine learning algorithm, and imputed Gargammel[9]-simulated ancient mtDNA across a range of genomic coverages to address these challenging cases (Fig. 1).

Fig 1: A) HMM (Analytics Vidhya, 2023) model and B) kNN clustering (Uddin et al., 2022)

RESEARCH QUESTIONS

To what extent can statistical imputation combined with machine learning approaches improve haplogroup assignment consistency, and what is the minimum sequencing coverage required to achieve reliable maternal lineage identification when using a dual imputation approach?

METHODS

MAVEN
Mitochondrial Analysis with Variant Enhancement Network

RESULTS

Fig. 2: Haplogrep3[10] haplotype average quality scores, grouped by minimum base call across coverages 1-10X.

Fig. 3: Number of correct haplotype & macrohaplogroup assignments by Haplogrep3[10], across imputation methods and coverages 0.25-3x when compared to the control.

Fig. 4: Number of correct haplotype & macrohaplogroup assignments by Haplogrep3[10], across imputation methods and coverages 0.25-3x when compared to the control, filtered for quality scores ≥ 0.9.

DISCUSSION

While traditional ancient DNA analysis typically excludes samples with <5X coverage, our results demonstrate that mtDNA quality substantially improves at =5X depth (Fig. 2). Our results show no significant increase in correct haplotype identification with each subsequent coverage increase at ≥5X (p-value < 0.05).
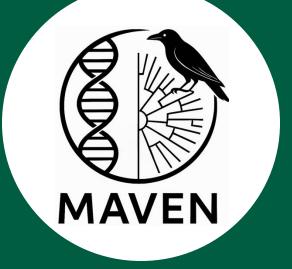
Our MAVEN imputation analysis reveals a more nuanced approach may be optimal: **For ultra-low coverage samples (0.25X-1X), MAVEN imputation dramatically improves the odds of correct haplogroup assignment (p-value=9.41×10^{-12}), while higher coverage samples (>1.5X) show diminishing returns or even decreased accuracy with imputation (Fig. 3).**

Post-imputation quality scores can serve as effective reliability indicators. Samples with scores >0.9 consistently demonstrate higher assignment accuracy across all coverage levels, with this filtering approach being particularly valuable for low-coverage samples (<1X) (Fig. 4; Table 1). **As a result, MAVEN has significantly higher counts of correct haplotype assignments across all coverages (p-value=3.187×10^{-9}).**

Coverage	High-Q Total / Correct (%)
0.25X	54 / 23 (42.6%)
0.5X	66 / 39 (59.1%)
0.75X	73 / 49 (67.1%)
1X	71 / 46 (64.8%)
1.25X	71 / 51 (71.8%)
1.5X	76 / 54 (71.1%)

Table 1: Summary of correctly identified haplogroups by MAVEN (Quality score ≥0.9).

FUTURE DIRECTIONS

Dual HMM Approach
Combined IMPUTE2[11] + Beagle[12]

Implementing a dual Hidden Markov Model system that combines strengths of IMPUTE2 and Beagle for improved variant capture.

Expected Outcome: Improved imputation accuracy across both common and rare variants

CNN Implementation
Deep Learning for mtDNA

Developing a Convolutional Neural Network approach to mtDNA imputation to better capture complex structural patterns in the mitochondrial genome.

Expected Outcome: Enhanced performance at higher coverage levels

Dynamic Algorithm
Adaptive Imputation

Developing an algorithm that dynamically determines whether imputation should be applied based on pre-imputation quality and coverage metrics.

Expected Outcome: Optimized application across diverse datasets

ACKNOWLEDGEMENTS AND REFERENCES