Single-locus imputation of ancient African DNA using novel regression-based approach

Ratchanon "RP" Pornmongkolsuk, Alejandro Ochoa

University Program in Genetics and Genomics

Department of Biostatistics and Bioinformatics

Duke UNIVERSITY

If I am busy, you can read below!!

Introduction

Imputation on ancient genomes is challenging due to:

- postmortem DNA degradation
- contamination

Addtionally, due to low linkage disequilibrium (LD), African genomes are particularly difficult to impute.[1]

Common imputation tools often rely on high quality reference genomes that are likely to contain haplotype information of the target genomes. Although ancient genomes may have information that is otherwise not found in modern data, their low coverage render them unlikely to be included as references.

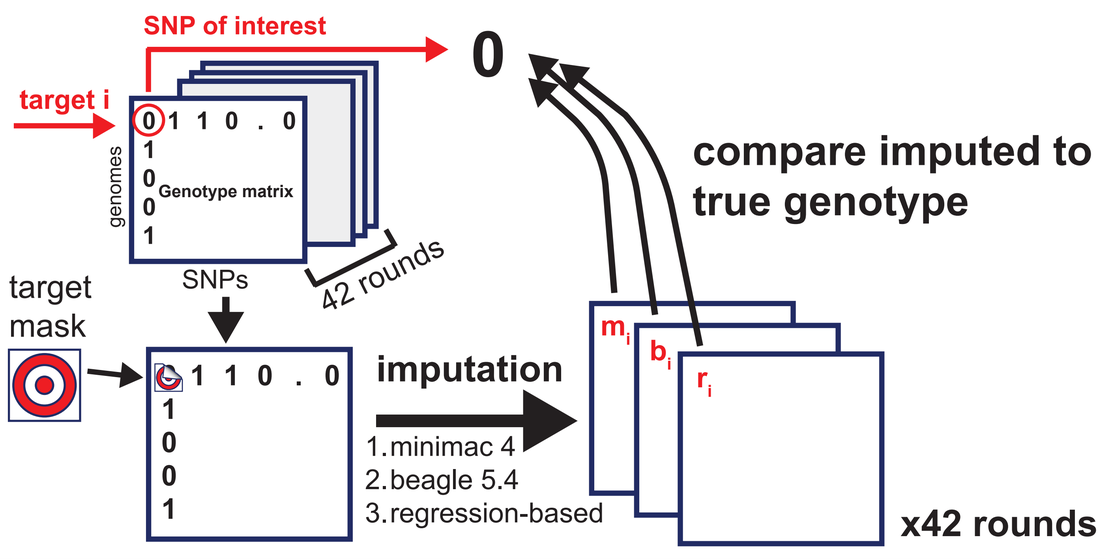
My aim is to develop a novel, regression-based method for imputing a single-nucleotide locus. The method prioritizes predictors that are present in each target genome and maximize the use of ancient data.

This project is part of my overarching dissertation in leveraging ancient DNA in studying evolution of Duffy Null variant in admixture context.

Leave-one-out cross validation

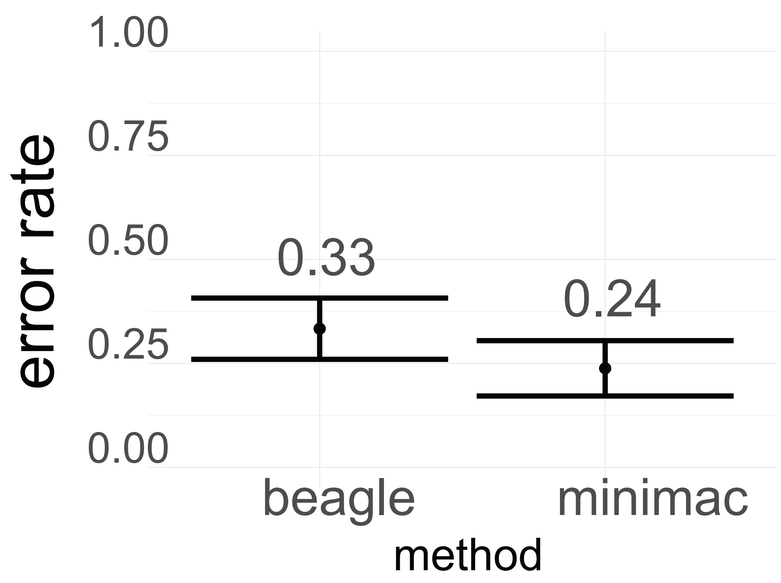
To evaluate the performance of methods in imputing a single nucleotide variant, leave-one-out cross validation is performed using Minimac 4[4], Beagle 5.4[5], and my regression-based tool (Fig 1).

Figure 1. Leave-one-out cross validation



Imputation was done on Duffy null (rs2814778) locus and on ancient African genomes that true genotypes are known (n=42). Mask the genotype, impute, and compare the results to calculate error rates (Fig. 2).

Figure 2. single-locus imputation error rates



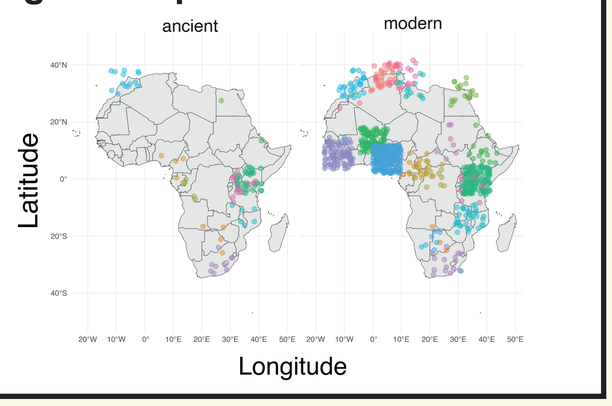
Results

Existing imputation tools perform worse when target samples are not represented in the reference panel.

Ancient DNA Data

Global modern and ancient samples were curated by the AADR[3] (v54.1.p1). The focus is on the populations in Africa.

Fig 3. Samples in Africa



Regression method

For each target, subset for loci that are present in target genome. Run independent logistic glm for each SNP, find best predictor one at a time by filtering out SNPs with correlation with top predictor > 0.05.

Potential Problem

- Missing patter correlates with age of sample and/or geography
- Missing data pattern leads to singular `glm()`
- Small sample size

Future Direction

- Simulation of ancient DNA
- Benchmark downstream analyses
- Multiple-step imputation
- Alternatives: `GLIMPSE`, `STITCH`

rp280@duke.edu
linktr.ee/arepee

Reference

1. Sousa da Mota, B. et al. Imputation of ancient human genomes. Nat Commun 14, 3660 (2023).
2. McManus, K. F. et al. Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. PLoS Genet 13, e1006560 (2017).
3. Mallick, S. et al. The Allen Ancient DNA Resource (AADR): A curated compendium of ancient human genomes. 2023.04.06.535797 Preprint at https://doi.org/10.1101/2023.04.06.535797 (2023).
4. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet 5, e1000529 (2009).
5. Browning, B. L., Zhou, Y, & Browning, S, R. A one-penny imputed genome from next generation reference panels. Am J Hum Genet 103(3):338-348. doi:10.1016/j.ajhg.2018.07.015 (2018).