Title:

A comparative analysis on ancient genomic data: The impact of variant discovery approach on population genetics tests

Author affiliations

İdil Taç[1], Ulaş Işıldak[2], Hande Çubukcu[1], Damla Karadavut[1], Kıvılcım Başak Vural[2], Ezgi Altınışık[3], Yılmaz Selim Erdal[3], Mehmet Somel[2], Füsun Özer[3,4], İdil Yet[1] and Gülşah Merve Kılınç[*,1]

1)Department of Bioinformatics, Graduate School of Health Sciences, Hacettepe University, 06100, Ankara, Turkey

2)Department of Biological Sciences, Middle East Technical University (METU), Ankara, Turkey

3)Department of Anthropology, Hacettepe University, Ankara, Turkey

4)Human G Lab, Department of Anthropology, Hacettepe University, Ankara, Turkey

Ancient DNA is a field of study with genomic material obtained from biological samples that are highly damaged and not stored under special conditions. However, with advances in both sequencing techniques and bioinformatics methods such as imputation, which improve the quality of the data obtained, we have the opportunity to make maximum use of the material we have. When working with ancient genomes, which are mostly low-coverage ancient genomes, we cannot follow the pipeline we apply to modern genome datasets. While modern genomes can reliably perform diploid variant calling due to their high coverage, ancient genomes with <1 coverage often undergo a pseudo-haploidization procedure. In this study, we used different approaches using pseudo-haploidization or read depths to calculate and compare the allele frequencies of neutral and disease related SNPs in Neolithic Anatolian individuals. We calculated the minor allele frequency for each SNP using (i) pseudo-haploidization with random base selection using the ANGSD tool followed by pseudo-haploidization with the Plink program and (ii) diploid variant calling with samtools -mpileup followed by an algorithm that uses read depths and finds the values that maximize the log-likelihood calculated for each population using the binomial probability distribution. With the second method we use, our aim is to maximize the use of low-coverage ancient genome sequence data by using all the information we obtain about the locus of interest from variant calling without pseudo-haploidization in ancient DNA. One-way ANOVA test was used to test whether there was a significant difference between the methods (p value <2e-16 for neutral SNPs, <2e-16 for type 2 diabetes-related SNPs), Tukey test was performed as a post-hoc test; p values: method 1 vs method 2 (0.00) for neutral SNPs, (3.64e-11) for phenotype-related SNPs.