DNA Phenotyping on Ancient DNA from Egyptian Mummies

Janet Cady, Mark Wilson, and Ellen Greytak*
Parabon NanoLabs, Inc.
*Correspondence to: ellen@parabon.com

Data

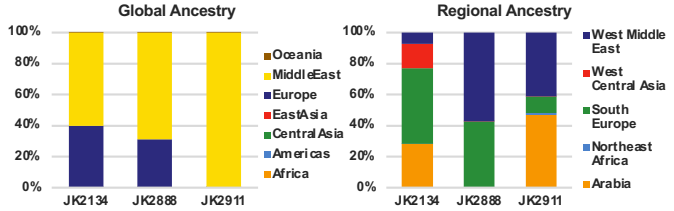
Raw sequencing reads (FASTQ files) from three ancient Egyptian mummies[1] were obtained from the European Nucleotide Archive (ENA). These samples were sequenced with a capture assay targeting 1.24 million SNPs, followed by alignment to the human genome and variant calling. Enzymatic damage repair was performed on each.

Sample ID	Date	Age (Years)	Coverage[1]
JK2134	776 – 569 BC	2,590 - 2,797	0.130X
JK2888	97 – 2 BC	2,023 - 2,118	0.229X
JK2911	769 – 560 BC	2,581 - 2,790	0.957X

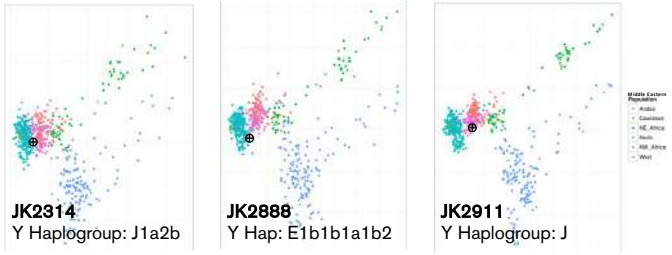
[1] Schuenemann, et al. (2017). Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. *Nature Communications*, 8, 15694.

Ancestry Prediction

The Snapshot DNA Phenotyping pipeline was applied to each subject to predict global and regional ancestry using admixture analysis.



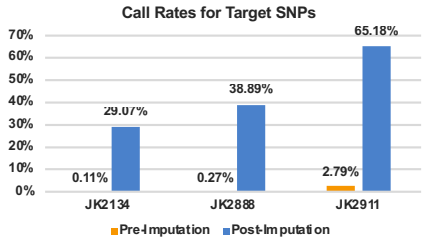
The Snapshot ancestry database of subjects with known ancestry was searched for the subjects with the most similar admixture proportions to each individual. They were found to be Jewish individuals from Yemen, Morocco, and Tunisia, respectively. Principal component analysis (PCA) within the modern Middle East and Y-haplogroup determination were also performed. Together, these analyses point to a non-sub-Saharan African origin for these ancient Egyptians, as was also found by Schuenemann et al. (2017).



Low-Coverage Imputation

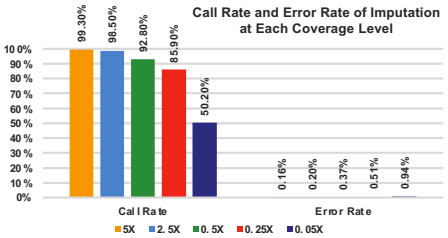
The SNPs of interest for phenotyping are those present on microarray chips. All three samples had autosomal coverage of <1X, so SNP genotypes could not be called directly from the reads. Traditional imputation approaches require calling as many SNPs as possible, then using those SNPs to infer the genotypes at uncalled sites. However, when coverage is very low, the called SNPs are too sparse for this technique to operate successfully.

Instead, a low-coverage imputation pipeline using genotype likelihoods was implemented. This technique uses hundreds or thousands of linked SNPs to statistically infer the most likely genotype for each target SNP, resulting in much higher call rates.



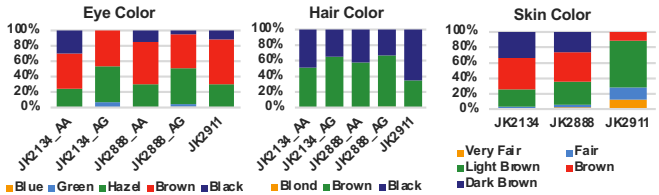
Imputation Validation

The low-coverage whole genome sequencing alignment file for subject HG00119 from the 1000 Genomes Project was downloaded, which has ~5.3X coverage. This data was randomly subsampled to 2.5X, 0.5X, 0.25X, and 0.05X. Low-coverage imputation was run on each subsample using a reference panel with subject HG00119 removed. Accuracy was determined by comparing the imputed genotypes to the genotypes in the 1000 Genomes phase 3 call set. As shown below, even at very low coverages, the vast majority of SNP genotypes can be accurately recovered by low-coverage imputation.



Phenotype Prediction

The two lower-coverage subjects were both missing the SNP rs12913832, which is the primary SNP associated with eye color, so eye color and hair color were predicted assuming AA and AG genotypes, which are the most common genotypes in the Middle Eastern population. There were also very few skin color SNPs available for these two subjects, resulting in low-confidence predictions. The graphs below show the relative consistency of each possible phenotype category.



Three-dimensional face morphology was predicted by predicting the values of face PCs, which were then transformed into 3D objects. Each subject was compared to the previous subject and heat maps were calculated to show the differences among the face predictions. These differences were then emphasized to create caricatured faces, which were combined with the pigmentation predictions to create composites of the individuals' likely appearance at age 25 by a forensic artist.

ID	Area	Width (X)	Height (Y)	Depth (Z)	Front	Profile
JK2134						
JK2888						
JK2911