Evaluating Genome-Wide Imputation Accuracy Across Southern European Populations Using 1000 Genomes Data As Reference

Vasilarou M[1]., Topaloudi A[1]., Tsetsos F[1]., Drineas P.[2], Gounari E.[3], Metallinou C.[1], Galanis A.[1], Yannaki E.[3], Stamatoyannopoulos G[4], Paschou P.[1]

[1] Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupolis, Greece
[2]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, United States of America
[3]Department of Hematology, George Papanicolaou Hospital, Thessaloniki, Greece
[4]Departments of Medicine and Genome Sciences, University of Washington, Seattle, WA, United States

Introduction

The imputation of genomic data is the process of predicting genotypes, based on a reference panel, that are not directly assayed in a sample of individuals. This technique boosts the number of Single Nucleotide Polymorphisms (SNPs) contained in datasets and allows to accurately evaluate the evidence for association in Genome-wide association studies (GWAS) at genetic markers that are not directly genotyped. Genotype imputation requires phasing which is the process of statistical estimation of haplotypes from genotype data.

The purpose of our study was the estimation of imputation's quality for Greek and South European populations with Europeans from 1000 Genomes as reference panel.

Materials and Methods

For the imputation we used a dataset of 561 South European individuals genotyped on Illumina HumanOmni 2.5 with a number of 2.300.000 markers.

The pre-imputation filtering of the dataset performed with the following procedure.

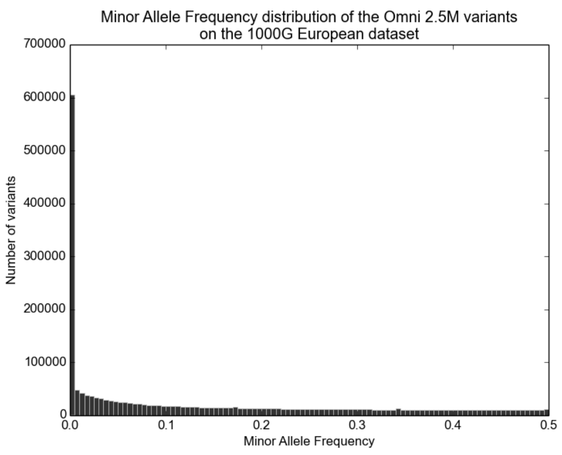
Pre-imputation data filtering

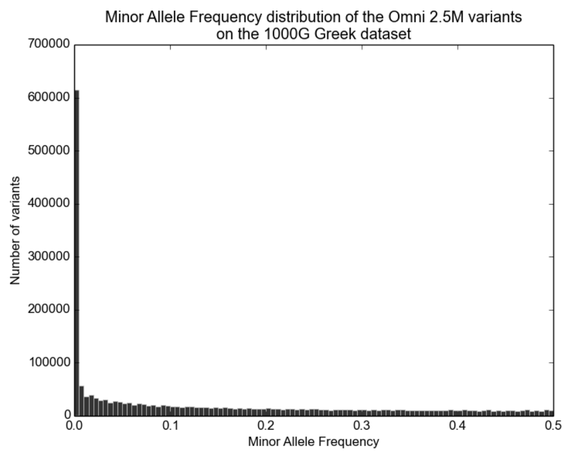
Procedure	Method
Update rs numbers according to dbSNP 142	Python script
Exclude double rs numbers and ambiguous SNPs	Python scripts
Keep only SNPs with call rate per sample 99.8%	PLINK
Set reference allele as provided from 1kG	PLINK
Extract common SNPs between dataset and Illumina Human660W-Quad	Python scripts and PLINK
Extract common SNPs between dataset and Illumina HumanOmniExpress-24	Python scripts and PLINK
Flip SNPs panel-reference populations	Python scripts and PLINK

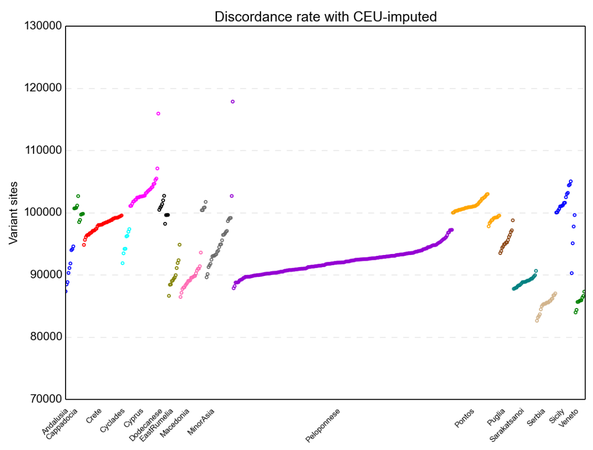
For our analysis we used Beagle 4.0 which first phases the data and then performs the imputation. Following we used BCFtools software for the evaluation of the results. We examined for each individual the discordance rate between the genotyped and the imputed markers.

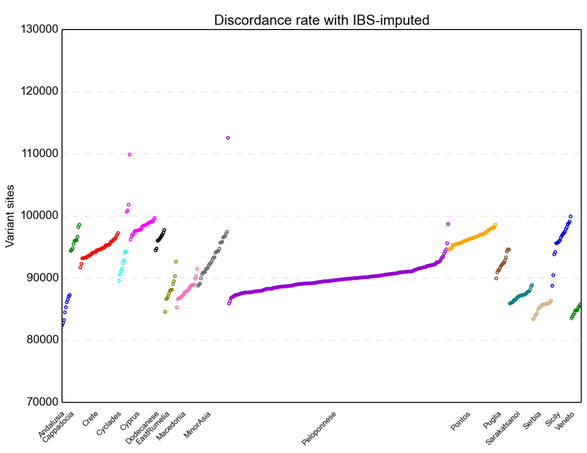
Results

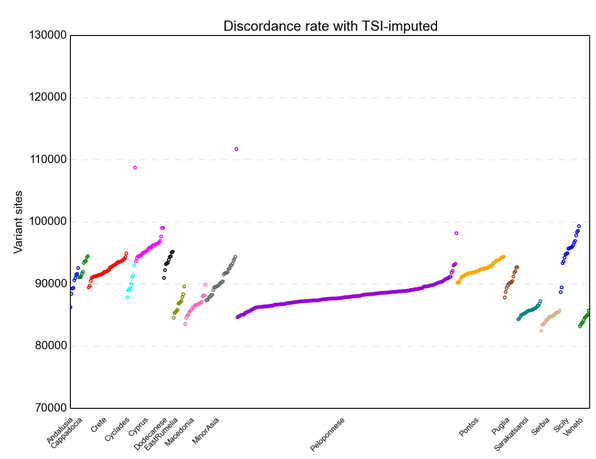
Figures: Minor Allele Frequencies on the 1000Genomes European and the Greek dataset.











In the figures: The genotype discordance between the original dataset genotyped at 2.3M markers and the reduced one that underwent imputation using the CEU, IBS and TSI panels respectively. The results are presented by population, showing the distinct differences in imputation efficiency per population subset. The imputation resulted in an average of 1.6M of total markers with a genotype probability over 90%.

Conclusion

We evaluate the suitability of the European populations of 1000Genomes as reference panels for imputation on South European samples. Our results can guide and increase the fidelity of genomic analyses based on samples of South European origin.

More specifically we found that:

1. The concordance rate of the genotyped markers differs from the imputed markers with different reference populations for the same individuals.
2. The markers imputed using as reference the TSI population show the highest concordance with the original markers. Generally, we noticed a trend in genotype discordance reducing by using imputation panels from populations that are more related to the Greek population.
3. There is a proportion of 30% of the variants that have a minor allele frequency less than 1%. For the imputation of these rare variants, a more closely related imputation panel is required.

References

1. Marchini et al,(2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*. 11, 499-511
2. Verma et al,(2014). Imputation and quality control steps for combining multiple genome-wide datasets. *Front.Genet.* 11, 5:370.
3. Browning et al, (2007). Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81:1084-97.

Supported by ARISTEIA II grant to PP (4386 - GENOMAP.GR). Operational Programme "Education and Lifelong Learning", co-funded by the European Union (European Social Fund) and national funds