The Impact of Imputation Quality on Family Based Analysis

Mahdi Mir

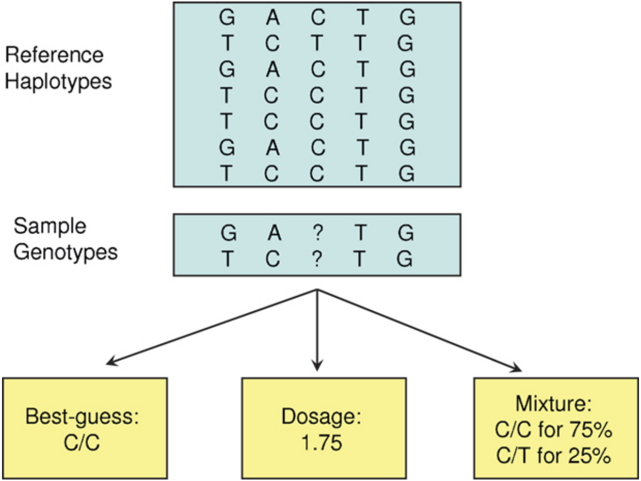
SSGAC

Oct 2024

What is imputation?

- Imputation is a process in which we predict the genotypes that are not directly observed.
- We use the information from the observed genotypes and a reference panel to predict the unobserved genotypes.

What is imputation?



Reference Haplotypes

G	A	C	T	G
T	C	T	T	G
G	A	C	T	G
T	C	C	T	G
T	C	C	T	G
G	A	C	T	G
T	C	C	T	G

Sample Genotypes

G	A	?	T	G
T	C	?	T	G

Best-guess: C/C

Dosage: 1.75

Mixture: C/C for 75% C/T for 25%

Source: Zheng et al. (2011)

Why Imputation?

- Imputation is used to increase the number of SNPs that are available for analysis.
- Imputation is used in GWAS to increase the power of the study.

Motivation

- We are concerned that low-quality imputed genotypes may not work for family-based analysis.
- We are interested in understanding the impact of imputation quality on downstream analysis.
- Family-based research designs rely on special properties of family data that might not be preserved in low-quality imputed genotypes.
- E.g.: In theory by the Mendelian laws we expect the correlation between sibling pairs and parent-offspring pairs genotypes to be 0.5.
- Current imputation methods do not take into account the relationships between the individuals.

Introduction

- We are going to show some correlation analysis between sibling pairs and parent-offspring pairs.
- We show the low-quality imputed genotypes correlation deviates from the theoretical expectations.
- E.g.: Howe et.al (2022) Sib-GWAS paper used low-quality imputed SNPs in its analysis.

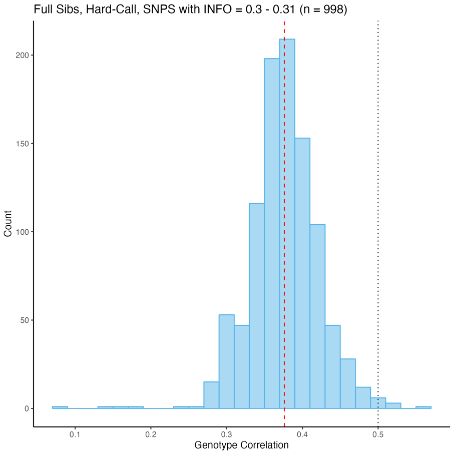
Correlation Analysis on the Imputed Data

- In theory, we should see 0.5 correlation between both full-sibling and parent-offspring pairs genotypes.
- So if the imputed genotypes are of high quality, then we should see the distribution of correlations to be concentrated around the half.
- Expectation: we would expect more deviation for low-quality SNPs from the theoretical expectation.

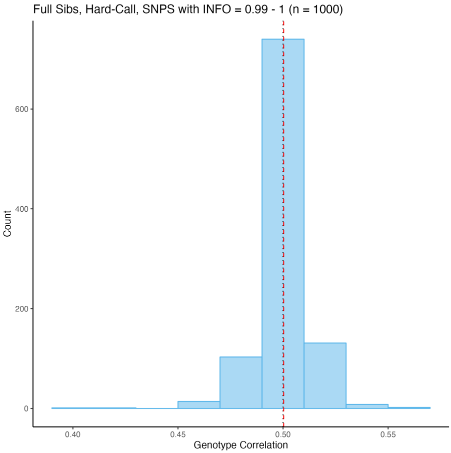
Sample

- UKB Imputed Data.
- 19K sibling pairs.
- 4K parent-offspring pairs.

Correlations Distribution - Full Siblings

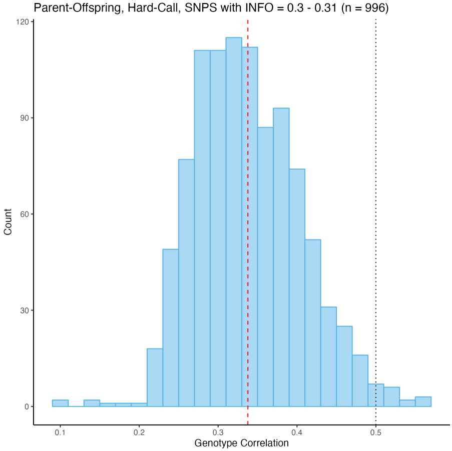


Low Quality Imputed SNPs

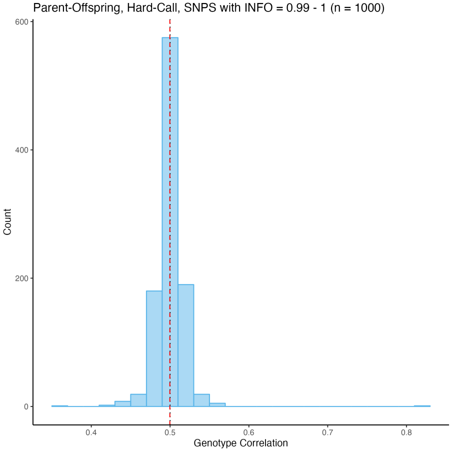


High Quality Imputed SNPs

Correlations Distribution - Parent-Offspring

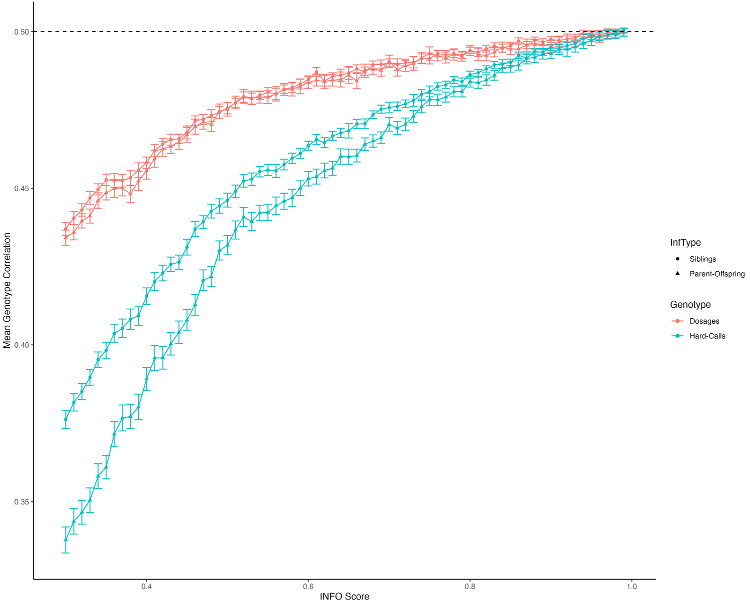


Low Quality Imputed SNPs



High Quality Imputed SNPs

Mean Genotype Correlation



Correlation Analysis Conditional on IBD states

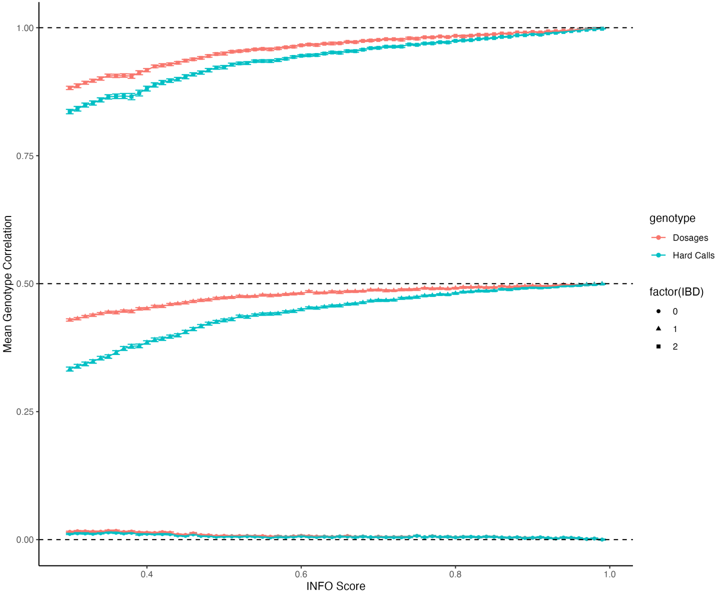
- Suppose i and j are siblings. Then in theory we have

$$Corr(G_i, G_j | IBD = 0) = 0$$

$$Corr(G_i, G_j | IBD = 1) = 0.5$$

$$Corr(G_i, G_j | IBD = 2) = 1$$

Mean Genotypes (info score)



Next Steps

- Using recently released WGS data from UKB. We are interesed to see what is the the downstream effect of using low-quality imputed genotypes in GWAS analysis.
- We can do that by comparing the results of GWAS analysis using imputed genotypes and the WGS data.

Next Steps

- We still need imputation methods because it sill increase sample size. Like parental imputation of those in UKB data.
- Imputation methods don't take into account the relationships so for a pair of siblings, the imputaion is done independently. But in reality there are some information that can be used to improve the imputation quality from the other sibling in the pair. In other words the better imputation method should keep the correlation between the genotypes of the siblings.
- We are indterested in developing imputation methods that take into account the relationships between the individuals.

Thank You!