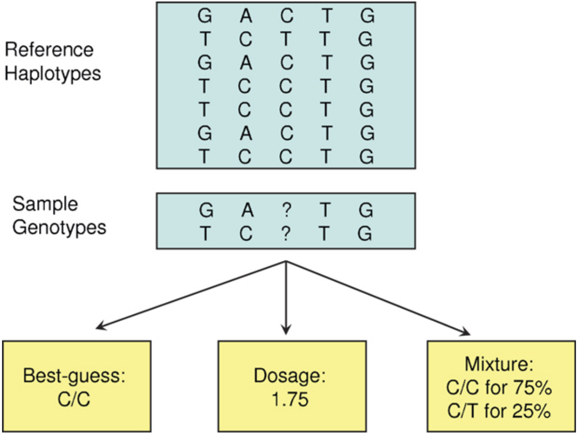
The Impact of Imputation Quality on Family-Based Analysis

Mahdi Mir (UCLA & SSGAC)
Alexander Young (UCLA & SSGAC)

Oct 2024

Imputation from a Reference Panel



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%

$$\text{Hard Call} = \text{Best Guess} \qquad \text{Dosages} = \mathbb{E}[\hat{G}]$$

$$0 \leq \text{Info Score} \leq 1 \qquad \uparrow \text{Imputation Quality} \iff \uparrow \text{Info Score}$$

Source: Zheng et al. (2011)

Reference Based Imputation VS Mandelian Imputation

- Mandellian imputation as done in SNIPAR (Young et.al, Nature Genetics 2022) is very different from reference based imputation.

- Reference based imputation is not taking into account the relationships between the individuals.

Motivation

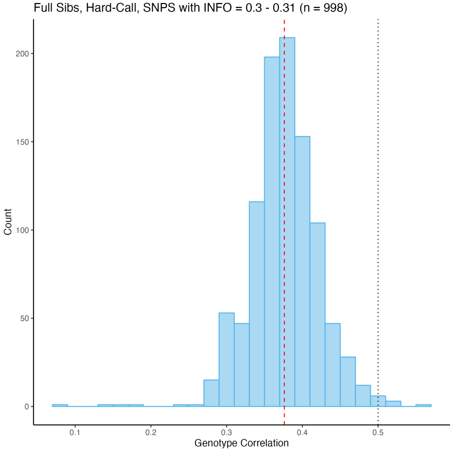
- Family-based research designs rely on special properties of the data.
- Low-quality imputed genotypes may not work for family-based analysis.
- Understaning the impact of imputation quality on downstream analysis by comparing to the WGS data

Correlation Analysis in UK Biobank

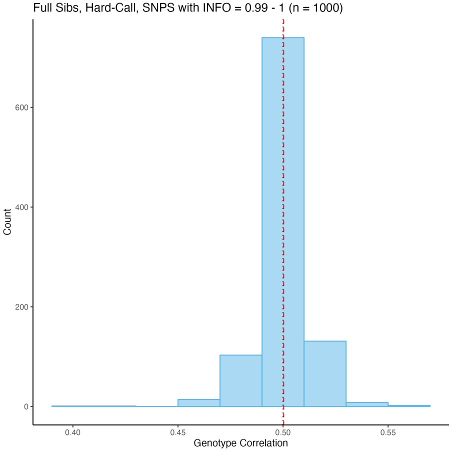
- UK Biobank Imputed Data
  - White British subsample
  - $19K$ sibling pairs
  - $4K$ parent-offspring pairs
  - SNPs with $MAF > 1\%$
- Howe et.al (2022) Sib-GWAS used low-quality (Info Score > 0.3) imputed SNPs.

Correlations Distribution - Full Siblings

Low Quality Imputed

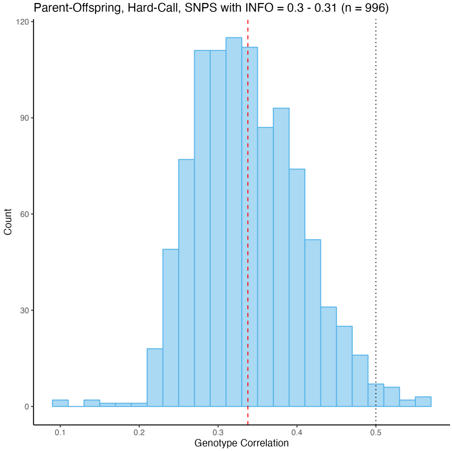


High Quality Imputed

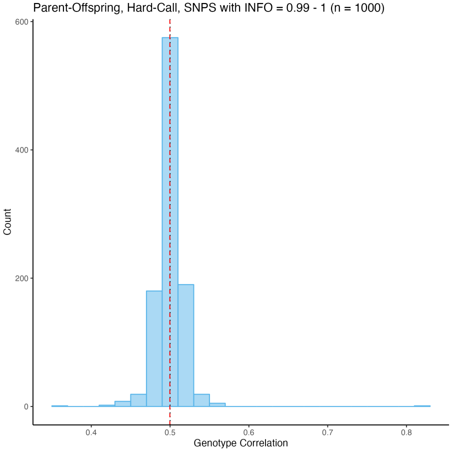


Correlations Distribution - Parent-Offspring

Low Quality Imputed

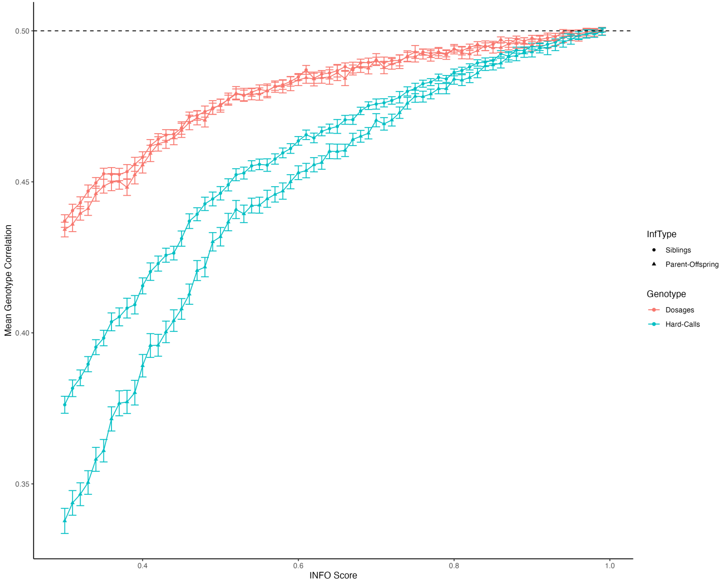


High Quality Imputed



Mean Genotype Correlation

As a Function of Info Score



Correlation Analysis Conditional on IBD states

- Quantitative genetics theory tells us the correlation between siblings' genotypes depends on their IBD state.

- IBD state records how many allels they share by descent from their parents.

- Suppose i and j are siblings. Then in theory [under random-mating] we have:

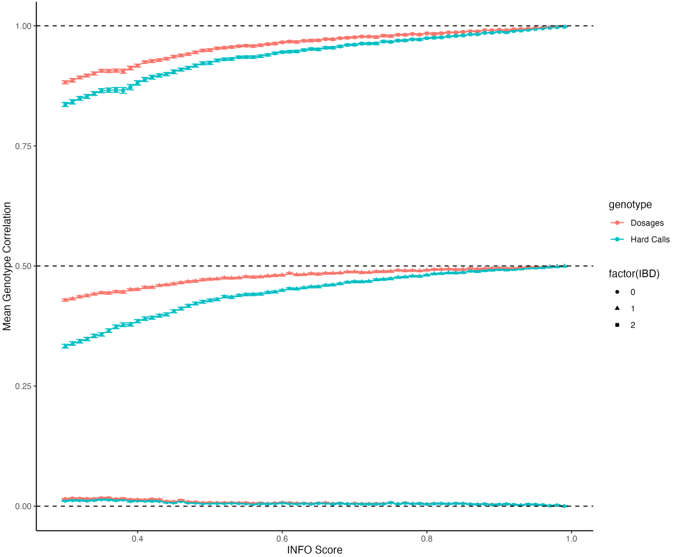
$$Corr(G_i, G_j|IBD = 0) = 0$$

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

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

Mean Genotypes Correlation Conditional on IBD States

As a Function of 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 Family-Based analysis.

- We can do that by comparing the results of Family-Based analysis using imputed genotypes and the WGS data.

Conclusion

- Genotyps imputed from a reference panel do not preserve Mandellian laws except for the very highest quality imputed variants.
- This is worse for best guess (Hard Calls) genotypes than for dosages.
- We are interested in developing reference-based imputation methods that take into account the relationships between the individuals.

Thank You!