Low quality genotype data is not appropriate for family-based analyses

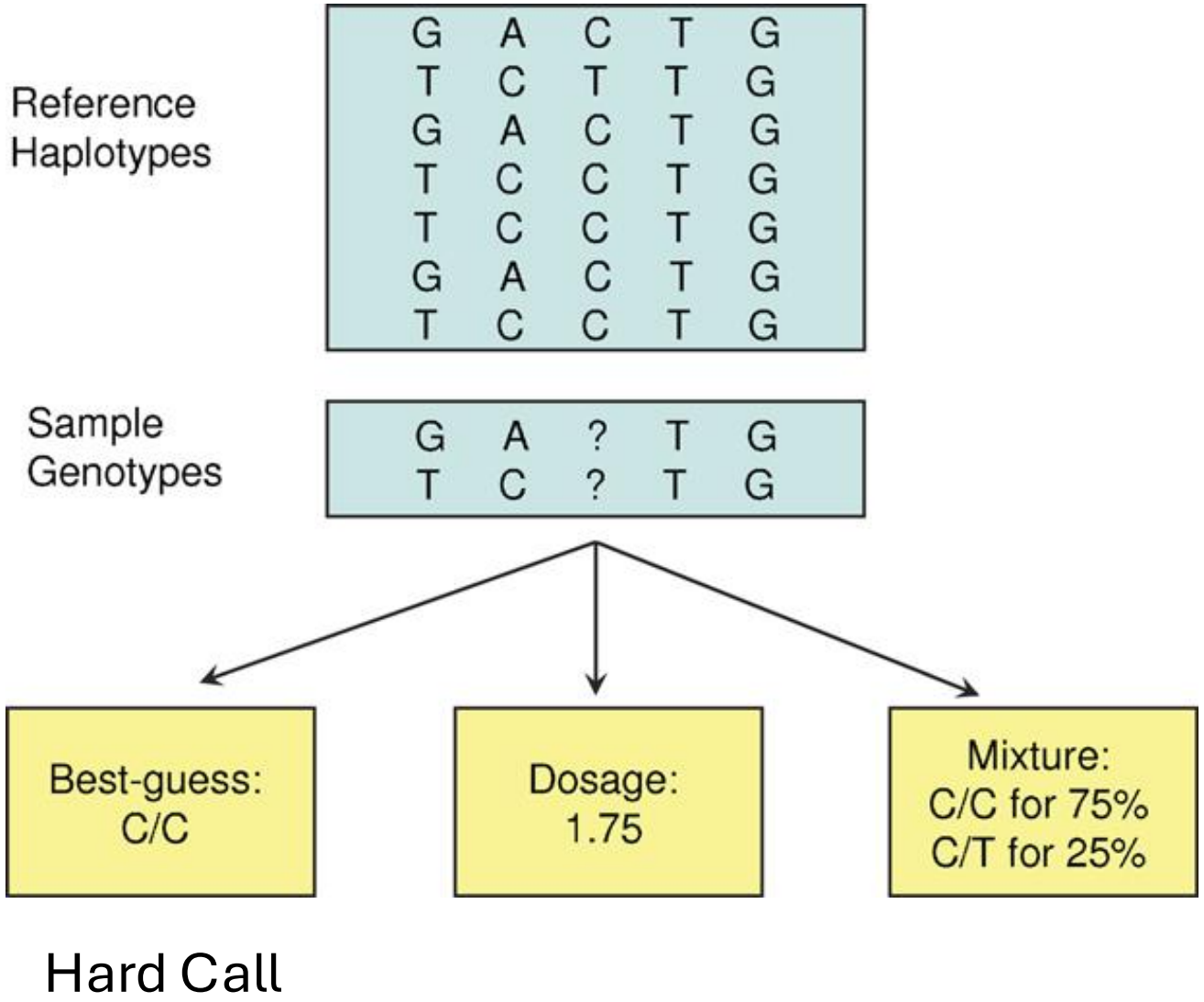
Mahdi Mir (UCLA & SSGAC)
Alexander S. Young (UCLA & SSGAC)

Oct 2024

Motivation

- We are concerned that the low-quality imputed genotypes may not be suitable for family-based analyses.
- FGWAS is designed to leverage Mendelian inheritance as a clean ***natural experiment*** to obtain unbiased estimates, but with imputation we may lose that.

Imputation from a Reference Panel



Hard Call

Source: Zheng et al. (2011)

Reference Based vs Mendelian Imputation

- Mendelian imputation, as implemented in SNIPAR (Young et al., Nature Genetics, 2022), differs entirely from reference-based imputation.

- Reference-based imputation does not consider the relationships between individuals and performs the imputation for everyone separately.

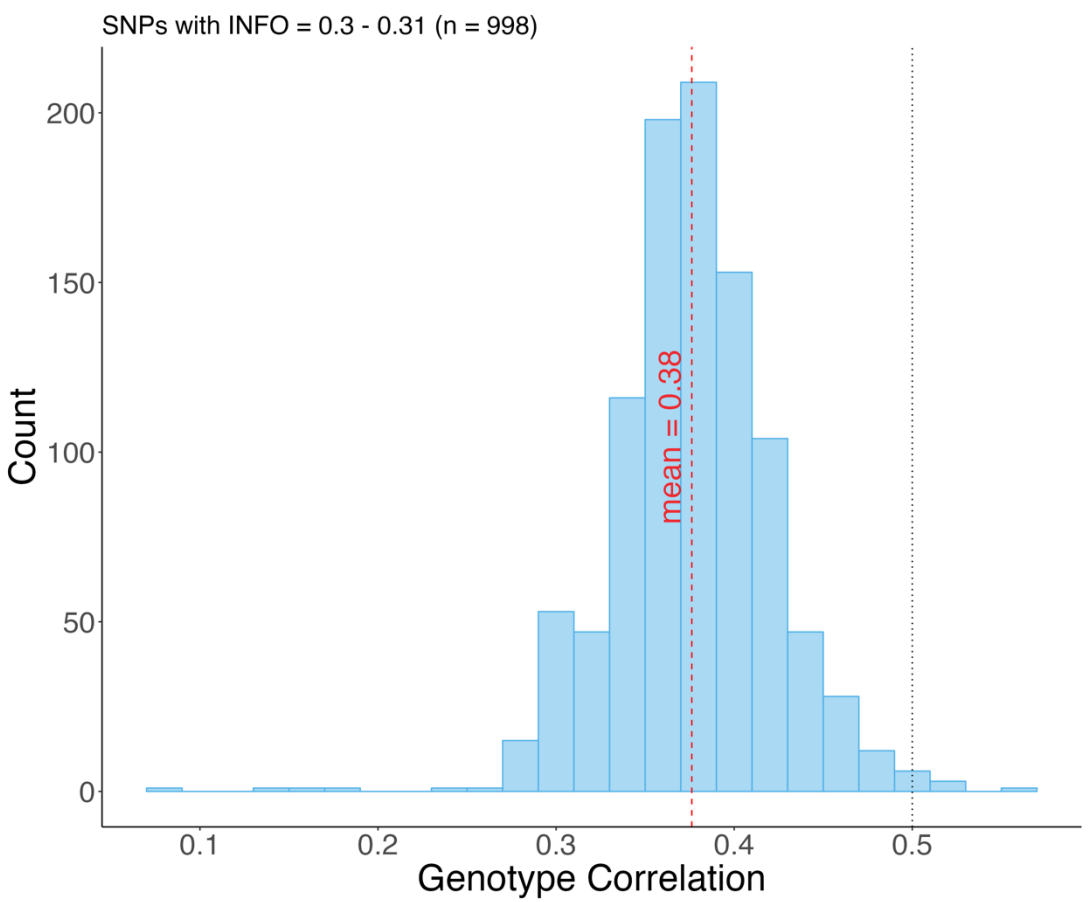
Correlation Analysis in UK Biobank

- UK BioBank Imputed Data
  - SNPs with MAF > 1%
  - Info Score (Imputation Quality) > 30%
  - 70K SNPs
  - White British Subsample
  - 19K Full Siblings Pairs
  - 4K Parent-Offspring Pairs

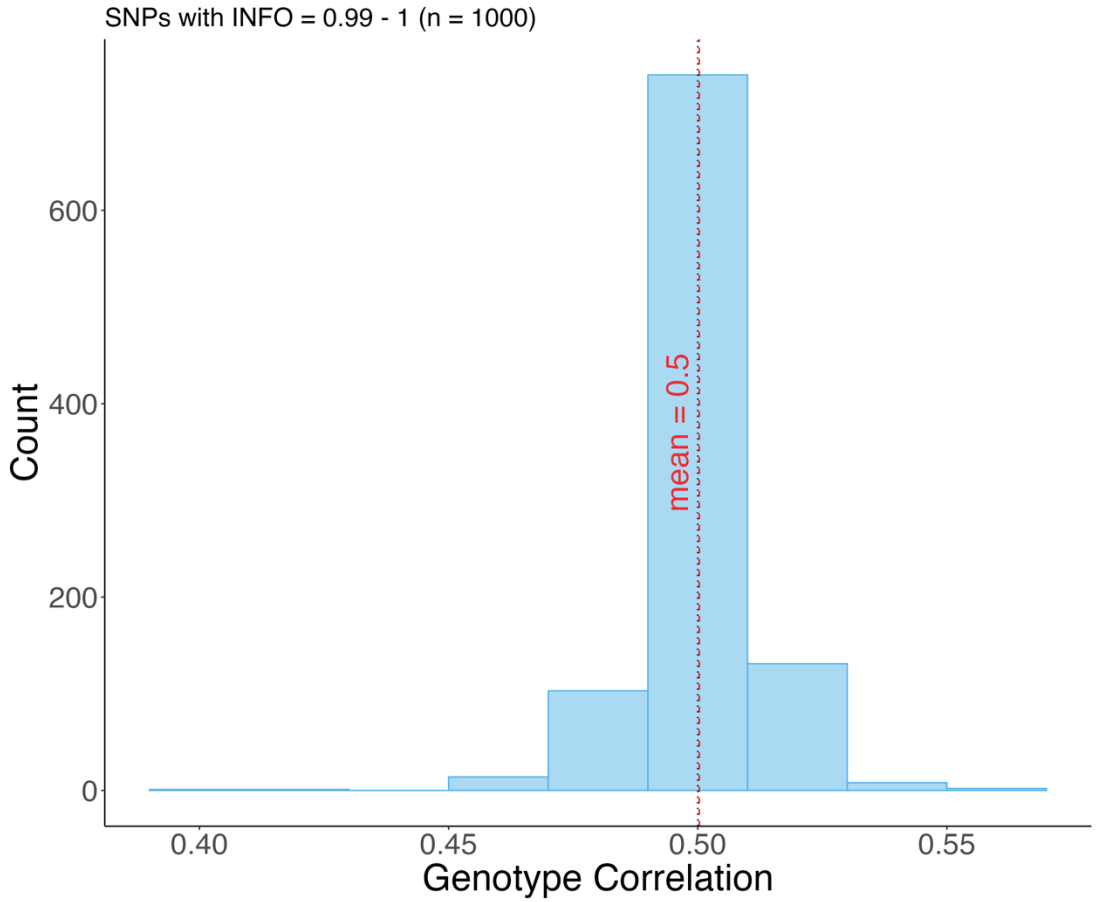
- Howe et.al (2022) Sib-GWAS used low-quality imputed SNPs (Info Score > 30%, Hard Calls).

Correlations Distribution – Full Sibs

Low Quality Imputed

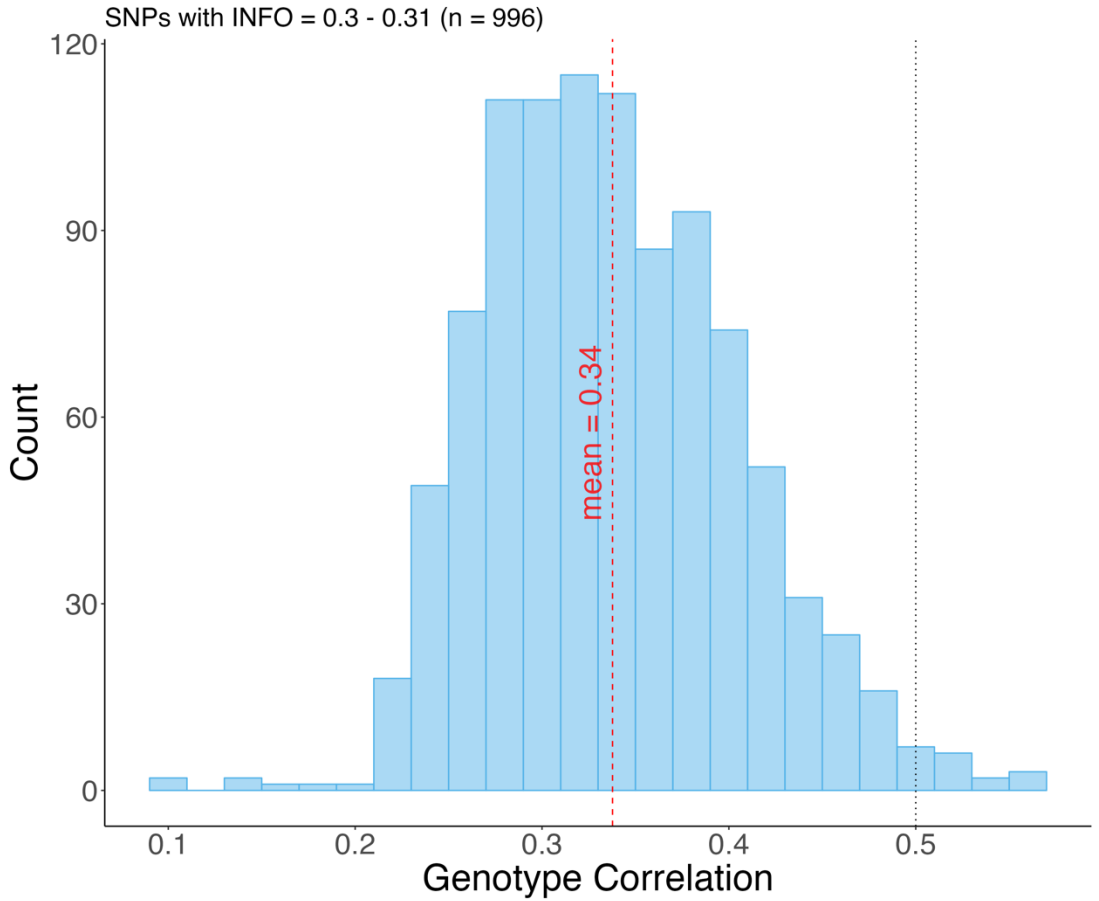


High Quality Imputed

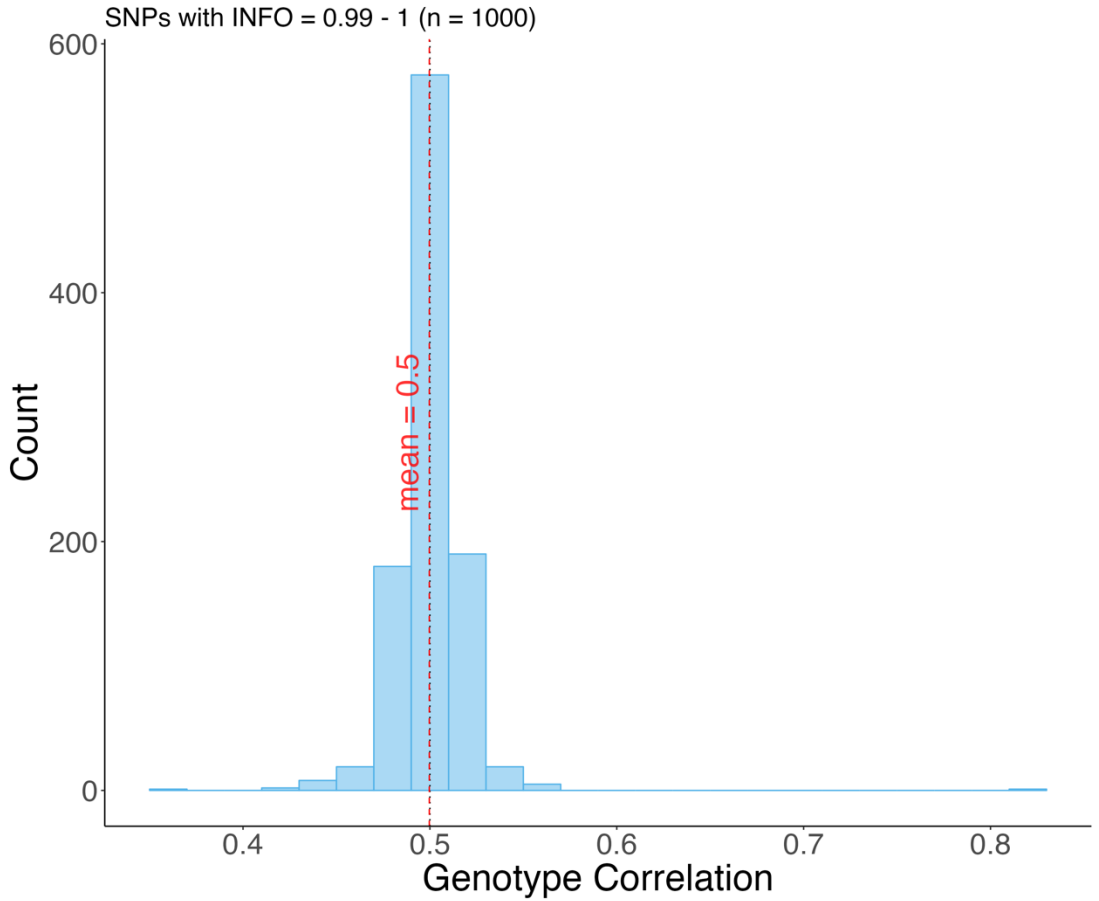


Correlations Distribution – Parent-Offspring

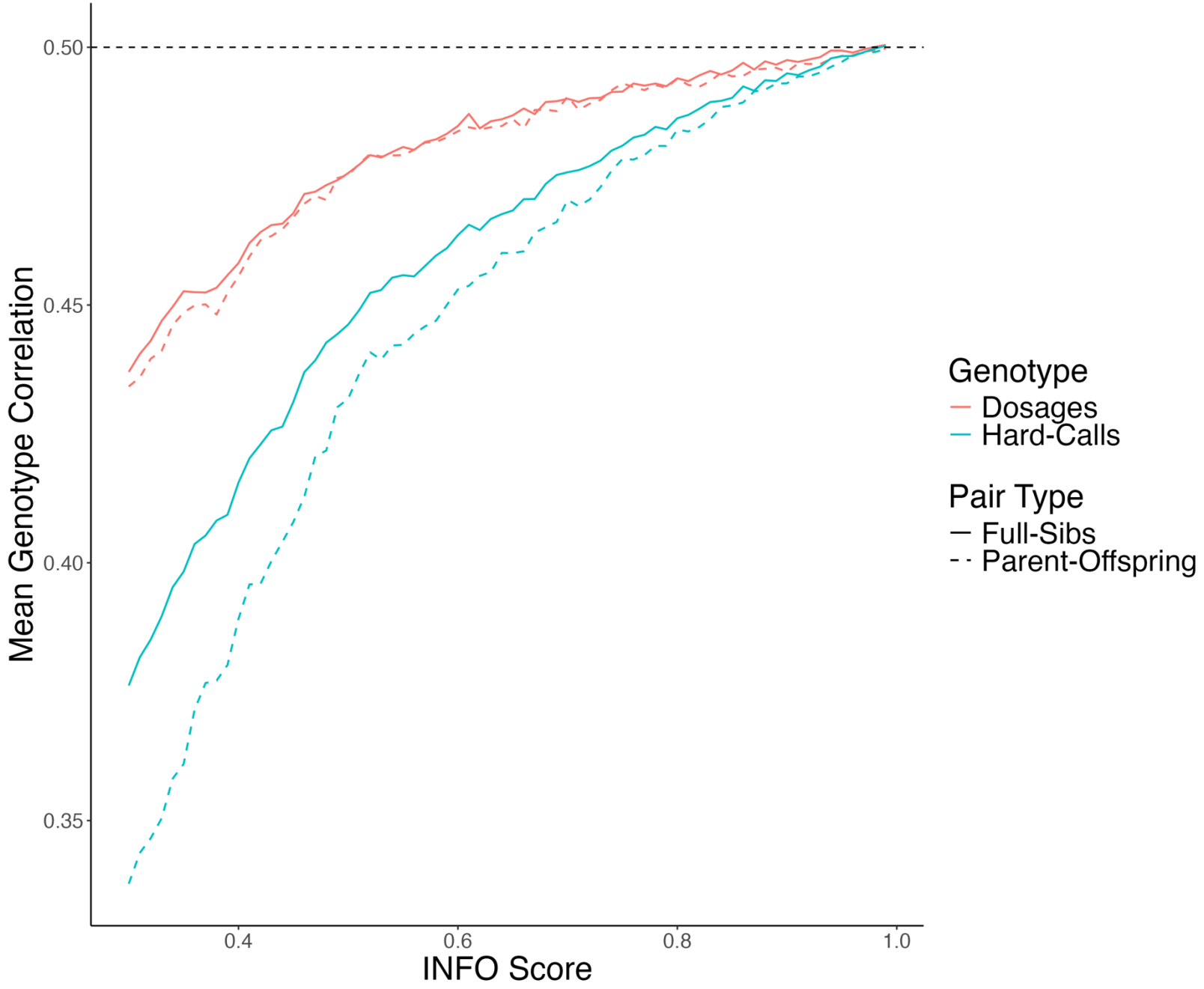
Low Quality Imputed



High Quality Imputed



Mean Genotype Correlation

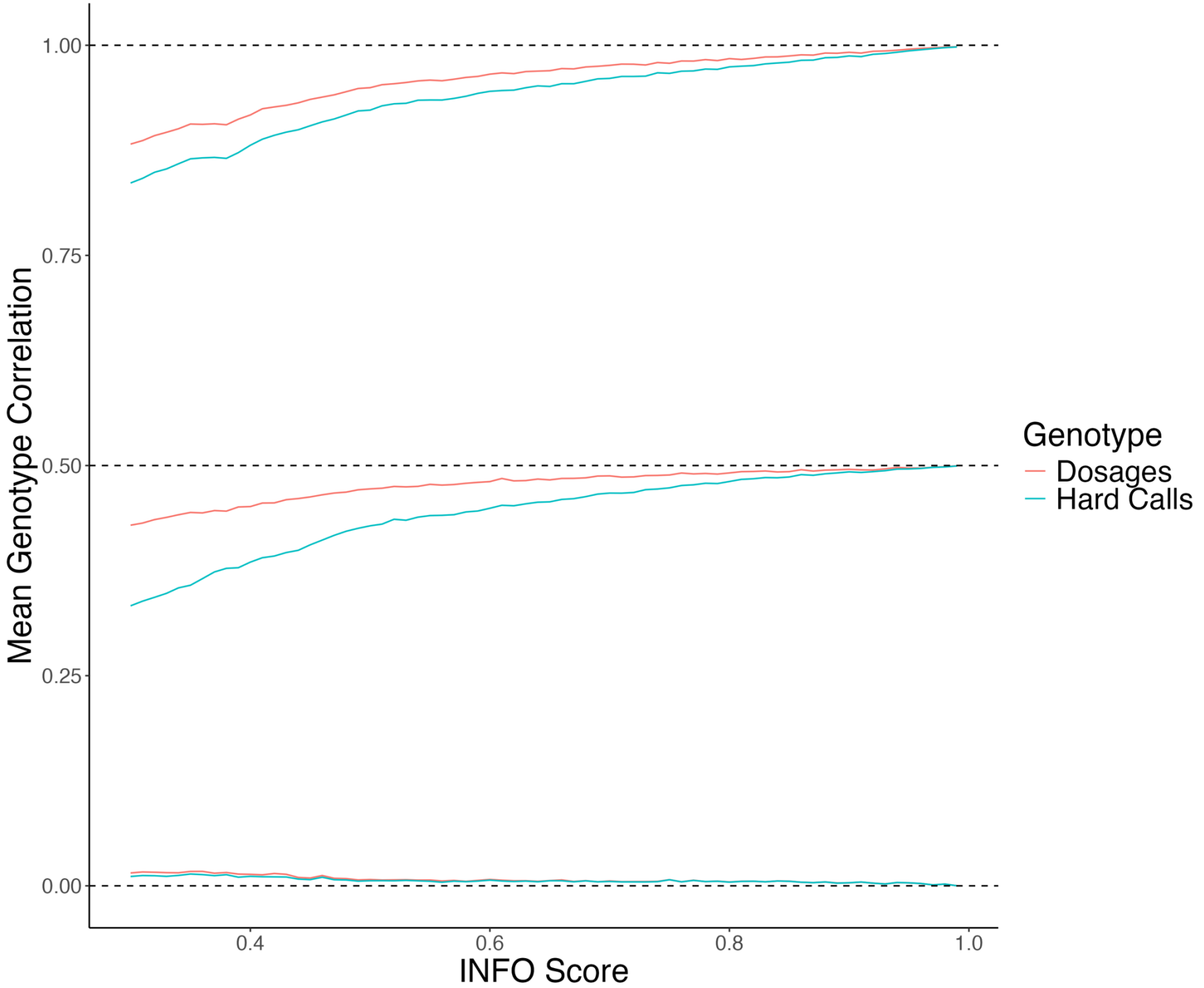


Correlation Analysis Conditional on IBD States

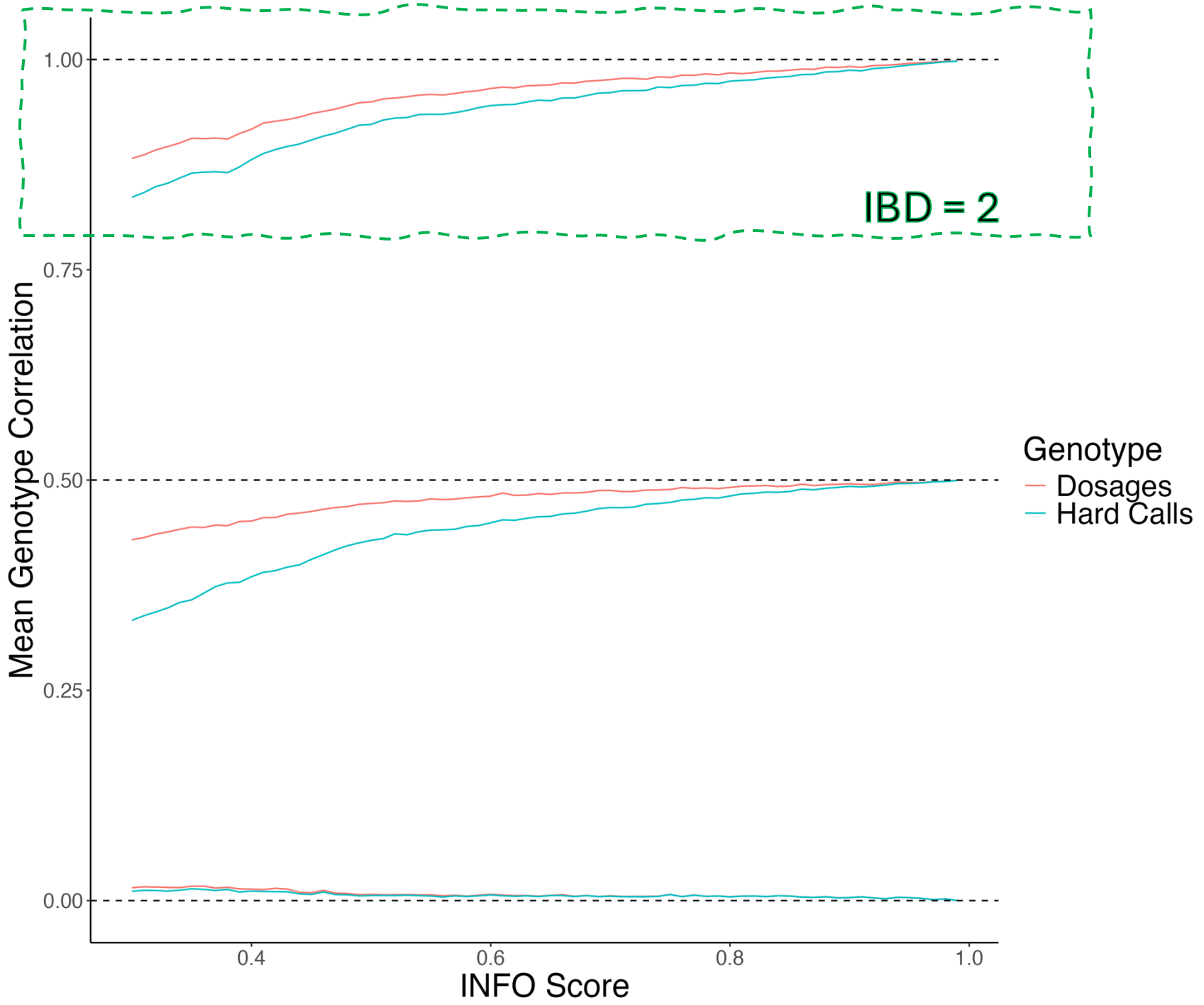
- Quantitative genetics theory tells us the correlation between siblings' genotypes depends on their IBD state.
- IBD state records how many alleles they share by descent from their parents.
- Suppose *i* and *j* are full siblings. Then in theory (under random-mating) we have:

$$\mathrm{Corr}\left(G_i, G_j \mid IBD = 0\right) = 0$$
$$\mathrm{Corr}\left(G_i, G_j \mid IBD = 1\right) = 0.5$$
$$\mathrm{Corr}\left(G_i, G_j \mid IBD = 2\right) = 1$$

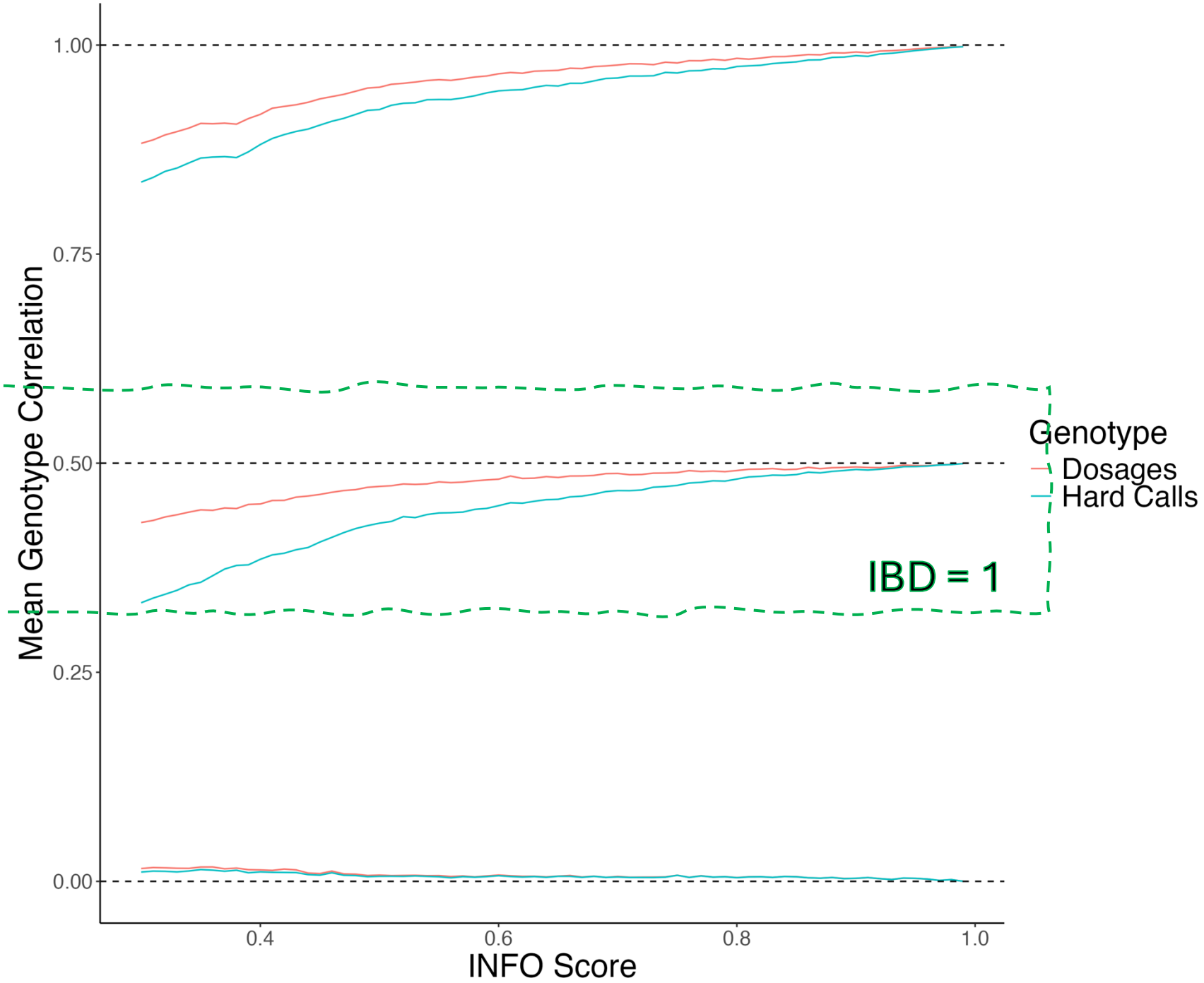
Mean Imputed Genotype Correlation Conditional on IBD State



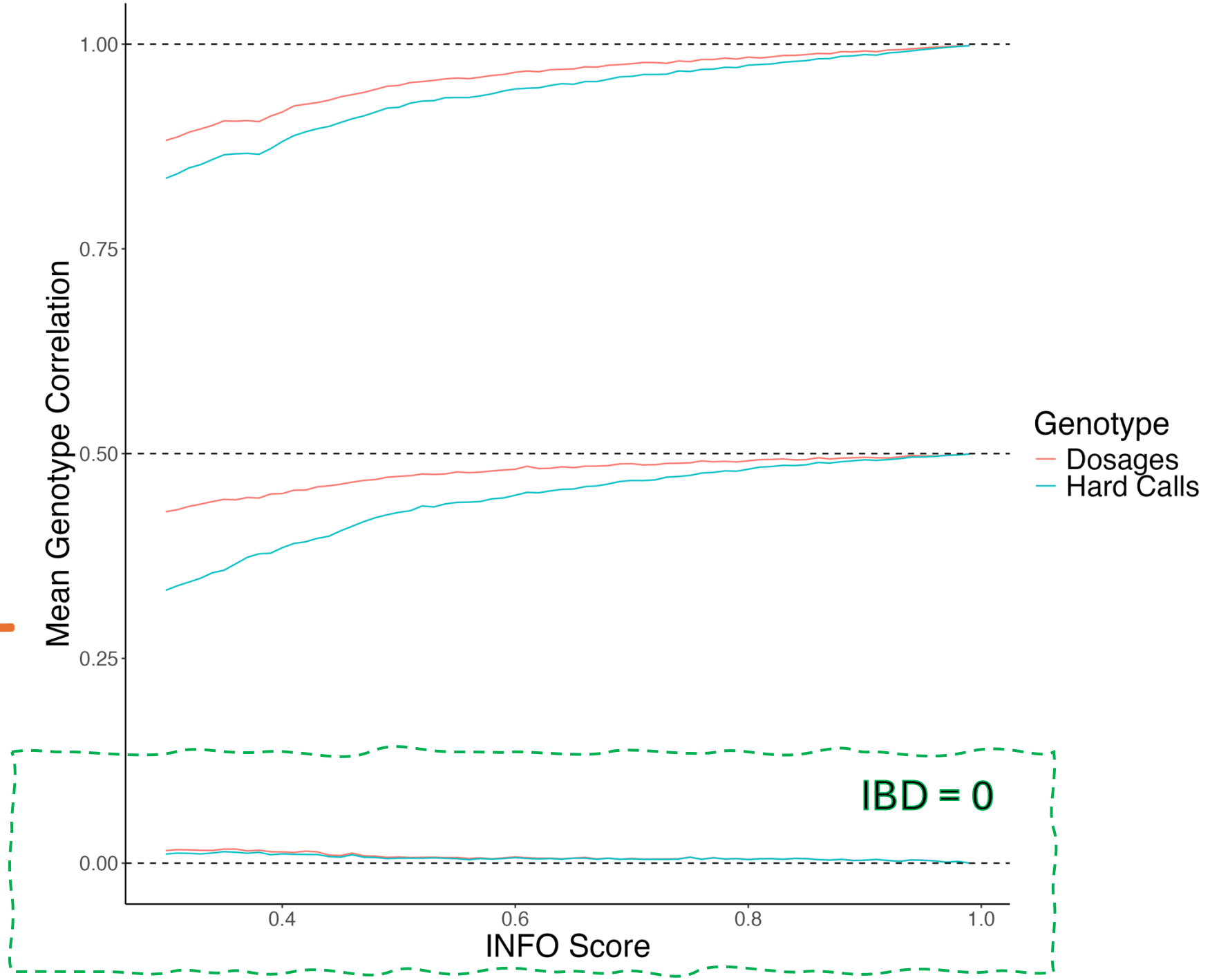
Mean Imputed Genotype Correlation Conditional on IBD State



Mean Imputed Genotype Correlation Conditional on IBD State



Mean Imputed Genotype Correlation Conditional on IBD State



Conclusion

- Genotypes imputed from a reference panel do not preserve Mendelian laws except for the very highest quality imputed variants.
- In practice with FGWAS, using imputed data, we can't ensure proper controls and unbiased estimates due to missing key data features.
- Stringent quality control is required for family-based analyses using imputed genotype data.

Next Steps

- We are working with WGS data to see the downstream effect of using low-quality imputed genotypes in FGWAS.
- Also, we are interested in developing reference-based imputation methods that consider the relationships between individuals and Mendelian laws.

THANK YOU!