The Impact of Imputation Quality on Family Based Analysis

Mahdi Mir

May 2024

Outline

Introduction

- We are concerned that low quality imputed genotypes may not work for family based analysis.
- E.g.: Howe et.al (2022) Sib-GWAS paper used low quality imputed SNPs in its analysis.
- They use hard-calls data for SNPs with info score > 0.3.
- Initially, I was focused only on the imputed data. I exteneded one of the Tammy's analysis to the whole genome.

Background

- There are these family-based research desings like what we have in the within-family project that rely on special properties of family data.
- E.g.: random assignment of genotypes within families.
- Howe et.al. (2022) used the genetic differences between sibs.

Motivation

- Based on the Mandellian laws we expect that the correlation between parent-offspring and sibling pairs to be 0.5.
- One way to detect issues in the imputed data is that if the correlation between sibling pairs and parent-offspring pairs deviates from 0.5.
- Assume i and j to be either siblings pairs or parent-offspring pairs, then theoretically we have:

$$Corr(G_i, G_j) = 0.5$$

- But the question is:

$$Corr(\hat{G}_i, \hat{G}_j) \ =? \ 0.5$$

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 is high quality imputed 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.
- We have a sample of $\backsimeq 20,000$ of sibling pairs.
- $\backsimeq 4000$ parent-offspring pairs.
- For each info score we selected 1000 SNPs randomly from the enitre genome.

What is Info Score?

- For each imputed SNP we have an info score between zero and one which gives us the quality of imputation.
- If it is closer to 1 the higher the quality of imputation is and the more we are confident that the imputation is close to reality.
- And the more it is closer to 0 we are less confident that is true.

$$\forall \text{ Imputed SNP } : 0 < \text{Info Score} < 1$$

Dosages vs. Hard-Calls Genotypes

- Imputation methods gives us for each SNP a discrete probablity distribution on 3 points.
- If we get the expectation of those probabilities we call it dosages.

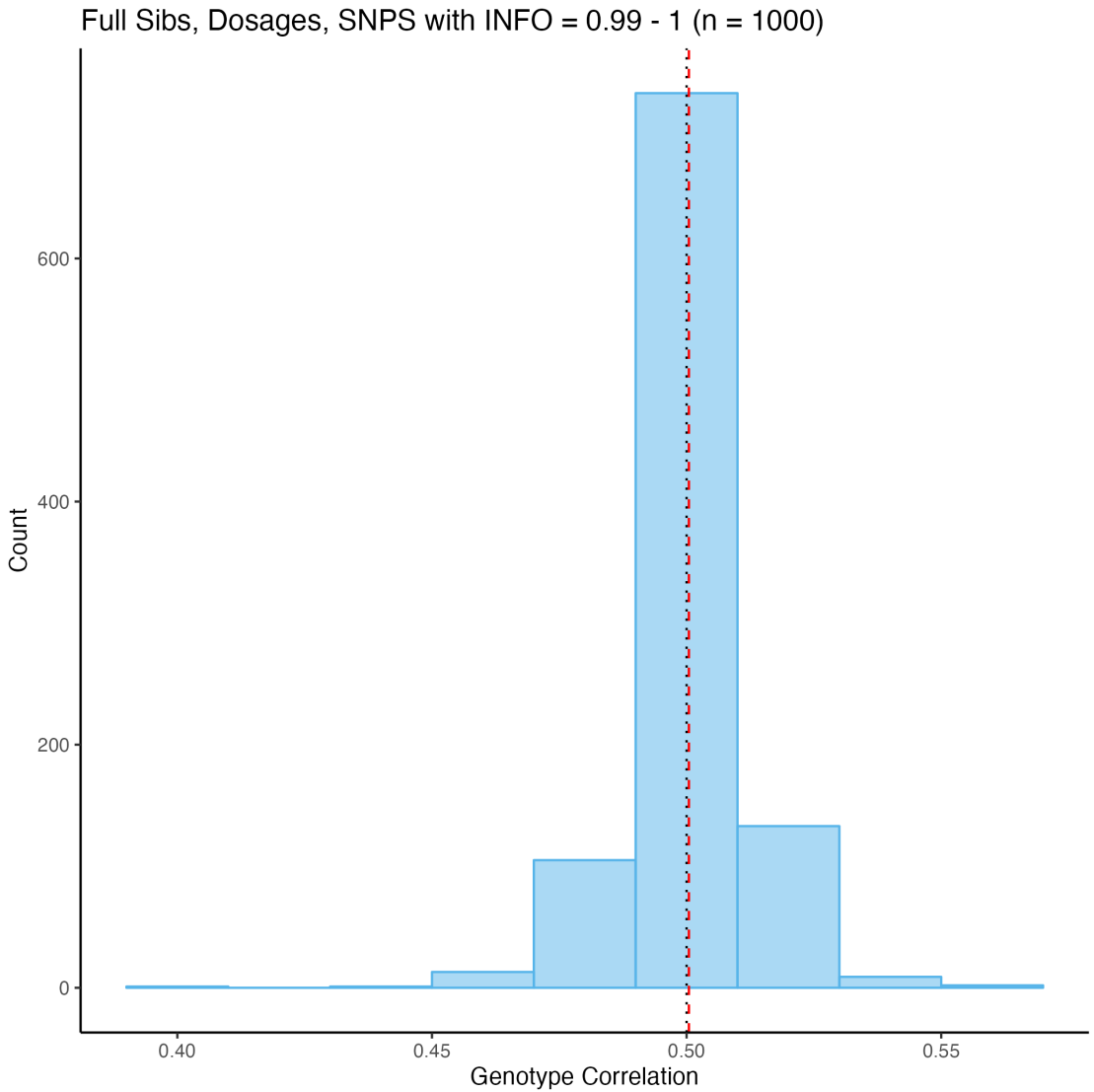
$$\text{Dosages} = \mathbb{E}(\hat{G}) = 0 \times P[G = 0] + 1 \times P[G = 1] + 2 \times P[G = 2]$$

- If we just pick the most probable genotype then we have the hard-call for the SNP.

$$\text{Hard-Call} = \underset{G \in \{0,1,2\}}{\text{argmax}}\ P(G)$$

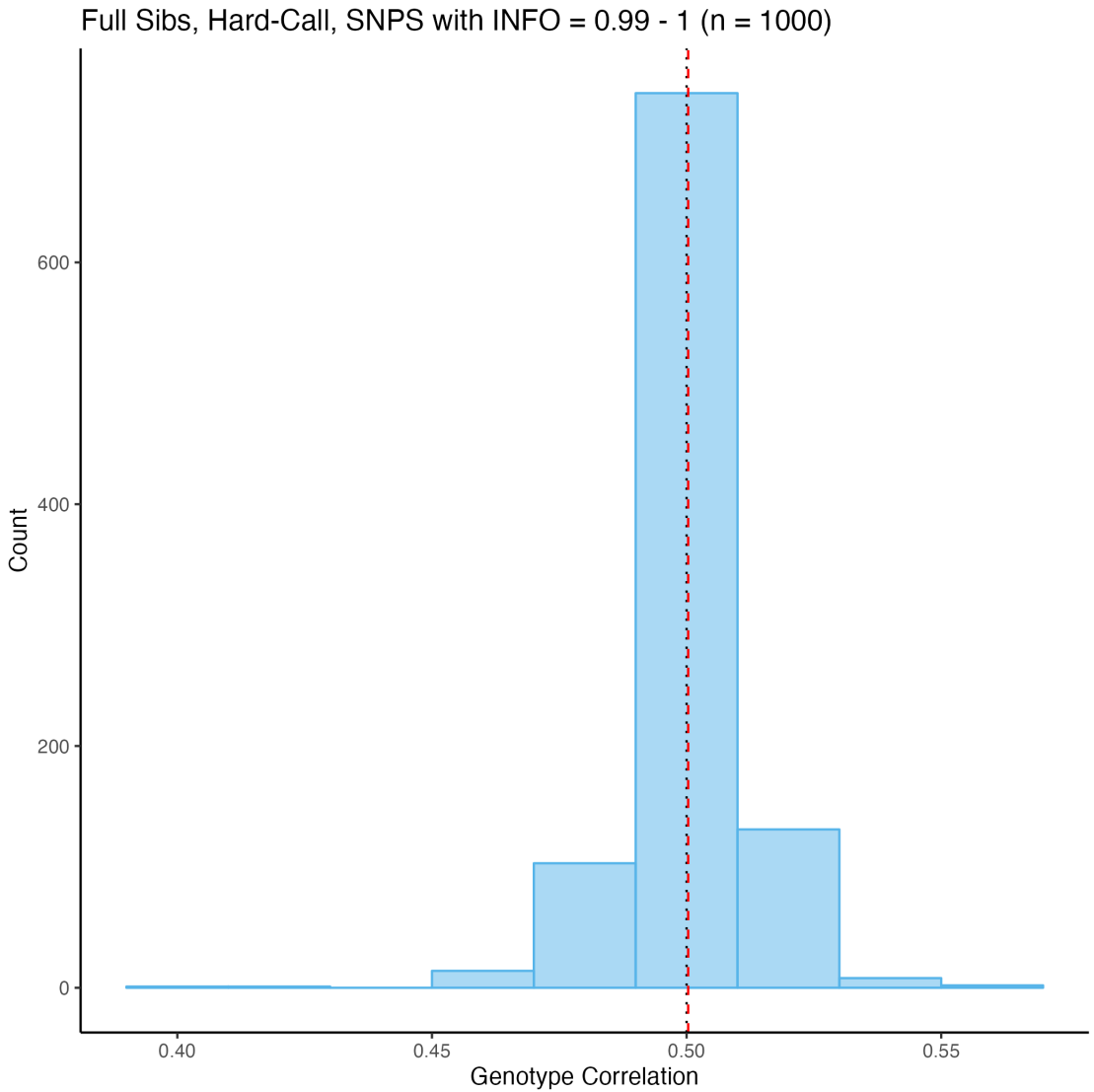
Correlations Distribution

High Quality SNPs - Full Siblings - Dosages



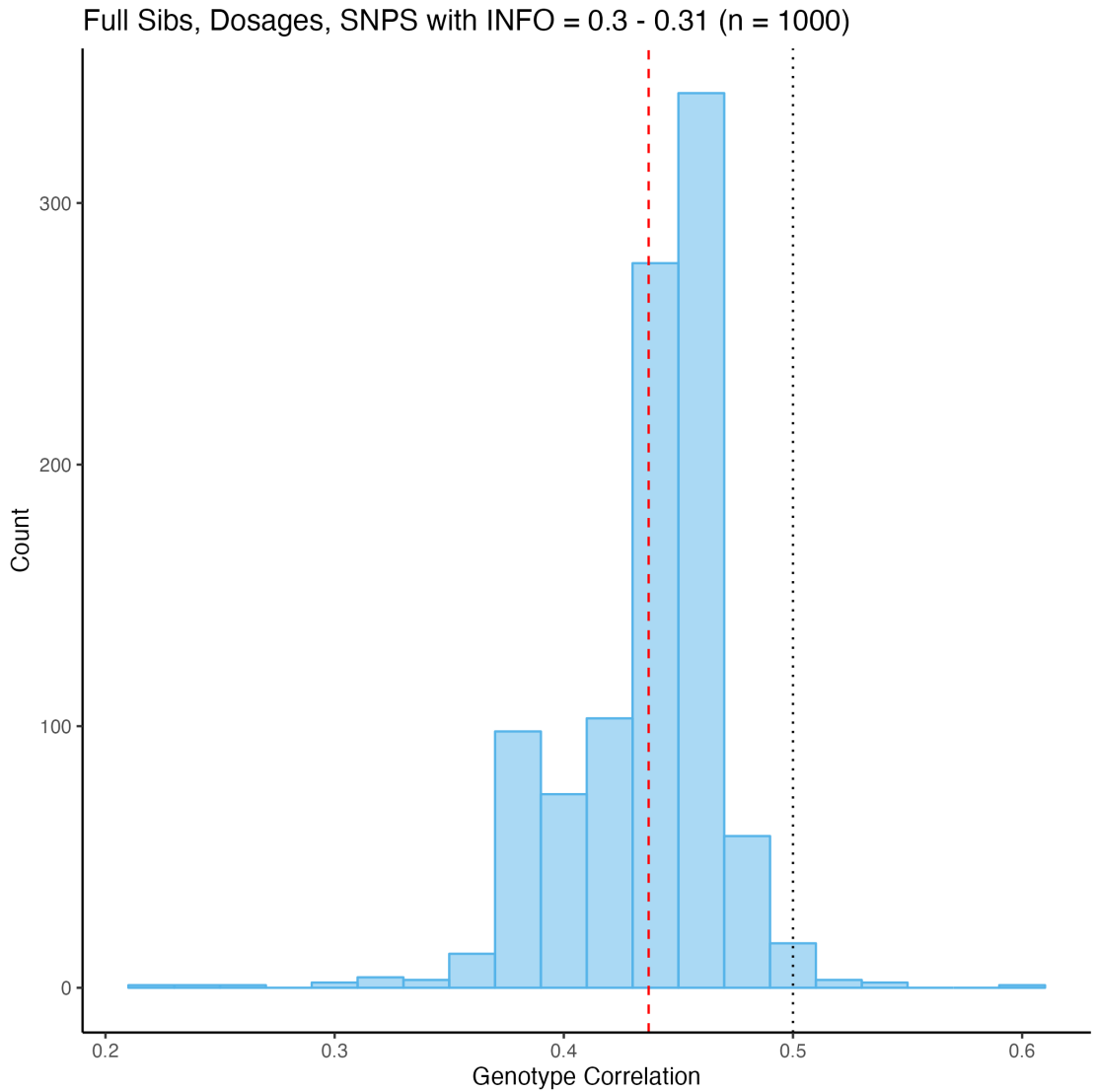
Correlations Distribution

High Quality SNPs - Full Siblings - Hard Calls



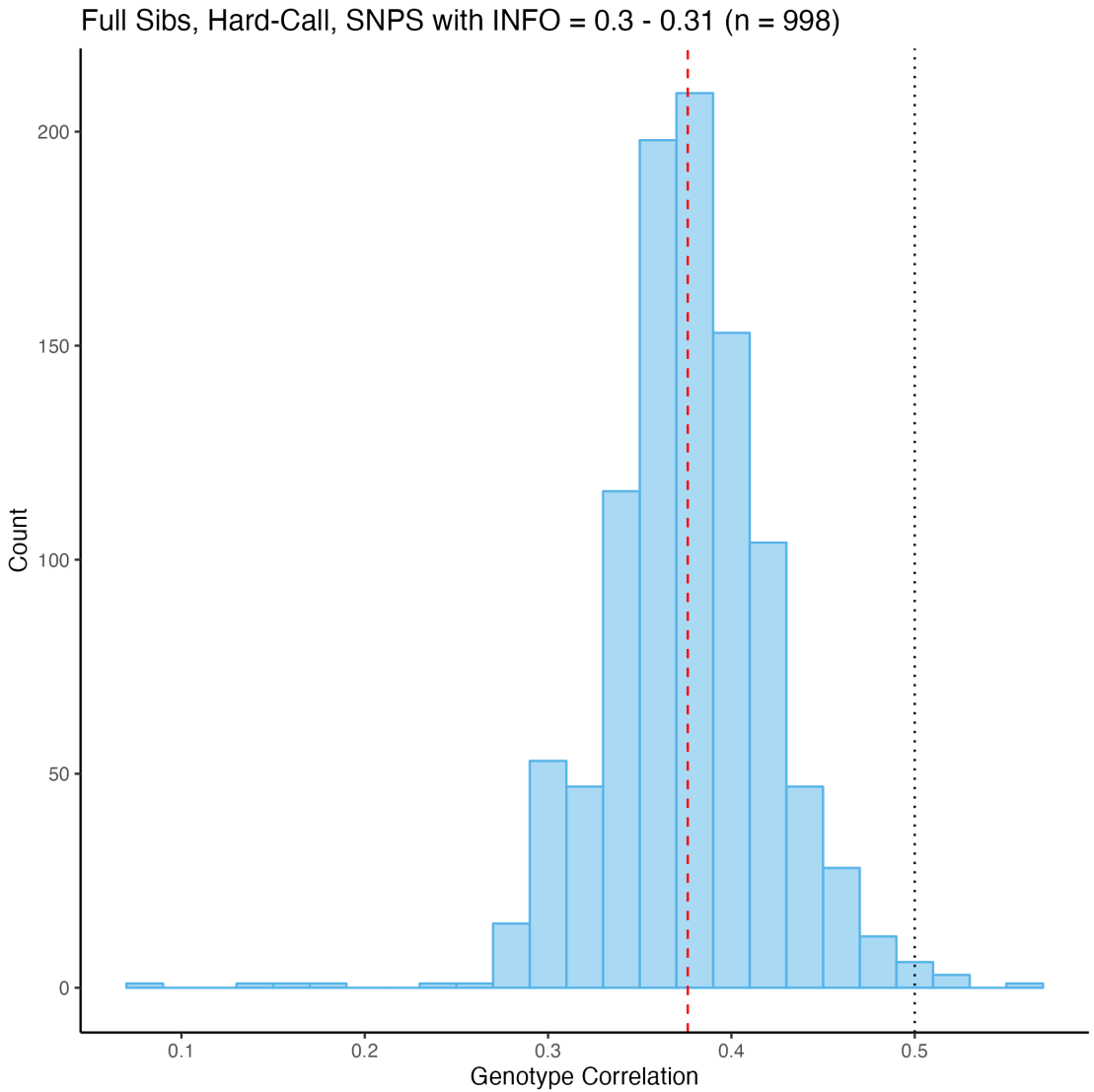
Correlations Distribution

Low Quality SNPs - Full Siblings - Dosages



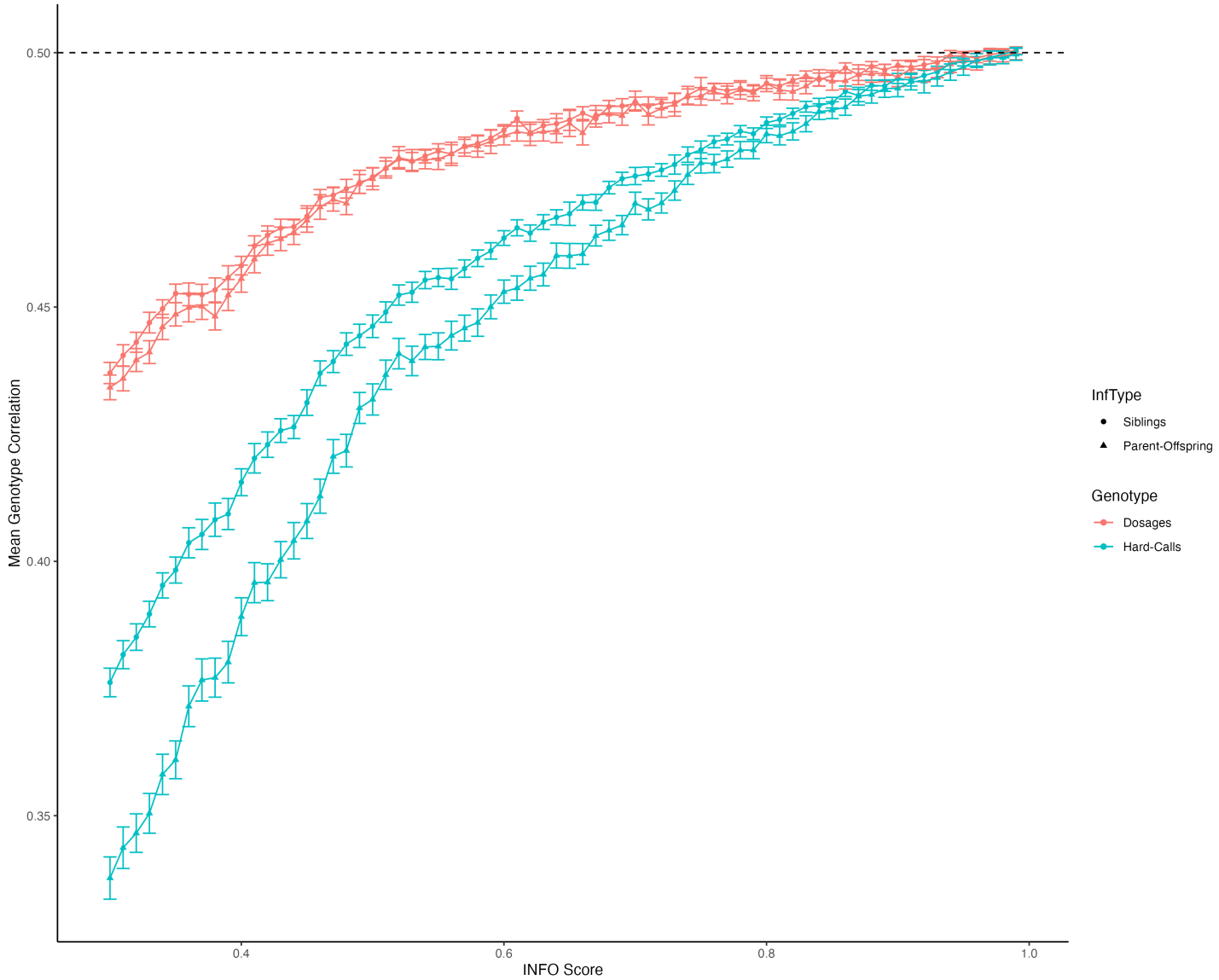
Correlations Distribution

Low Quality SNPs - Full Siblings - Hard Calls



Mean Genotype Correlation

As a function of Info Score



Correlation Analysis Conditional on IBD states

- What if we calculate all these correlations conditional on IBD states?
- Conditioning on IBD states for parent-offspring pairs doesn't mean anything because they are always IBD $= 1$ so we would get the same results as before when we didn't condition on IBD states.
- Suppose i and j are siblings. Then in theory we have

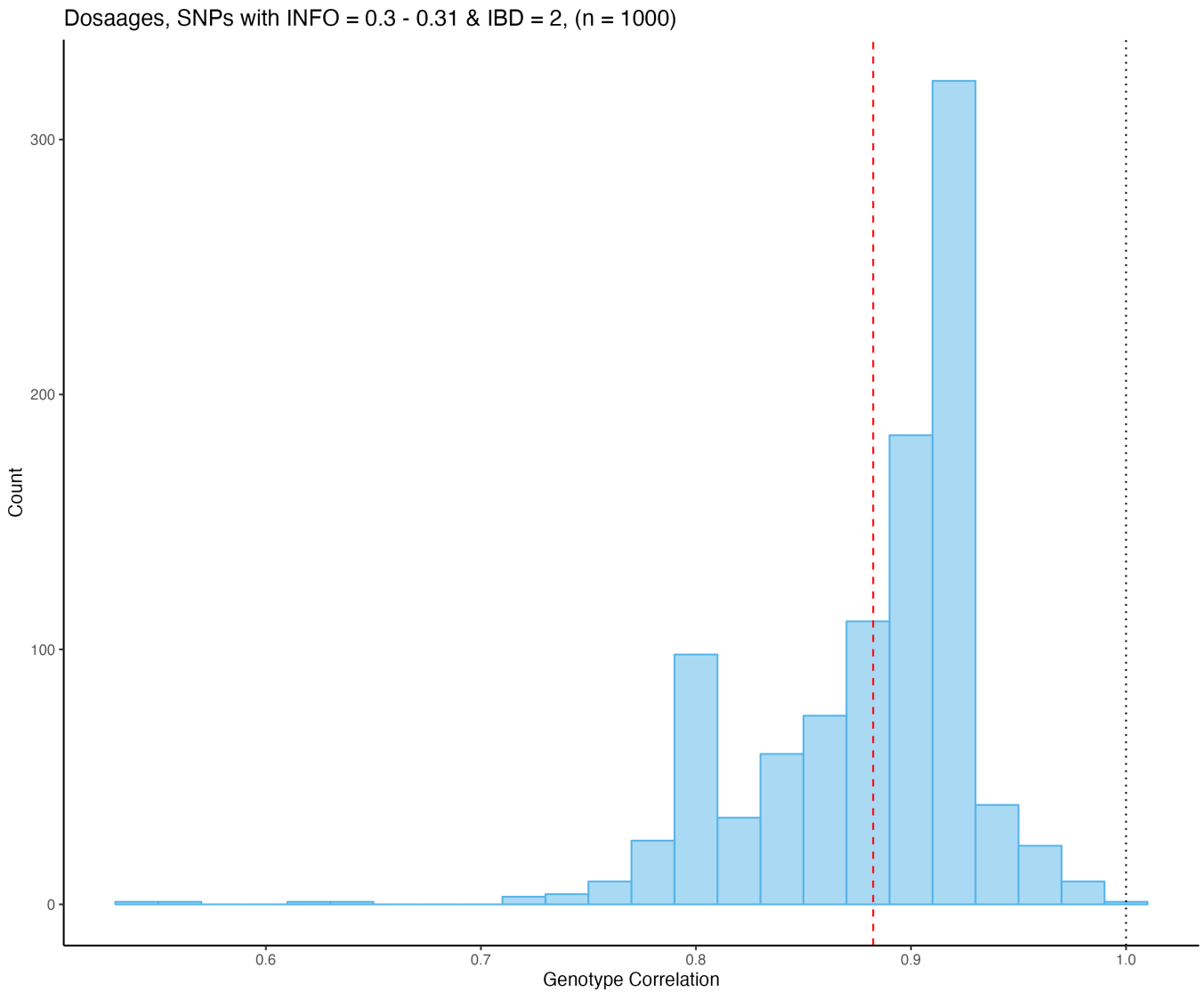
$$Corr(G_i, G_j | IBD = 0) = 0$$

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

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

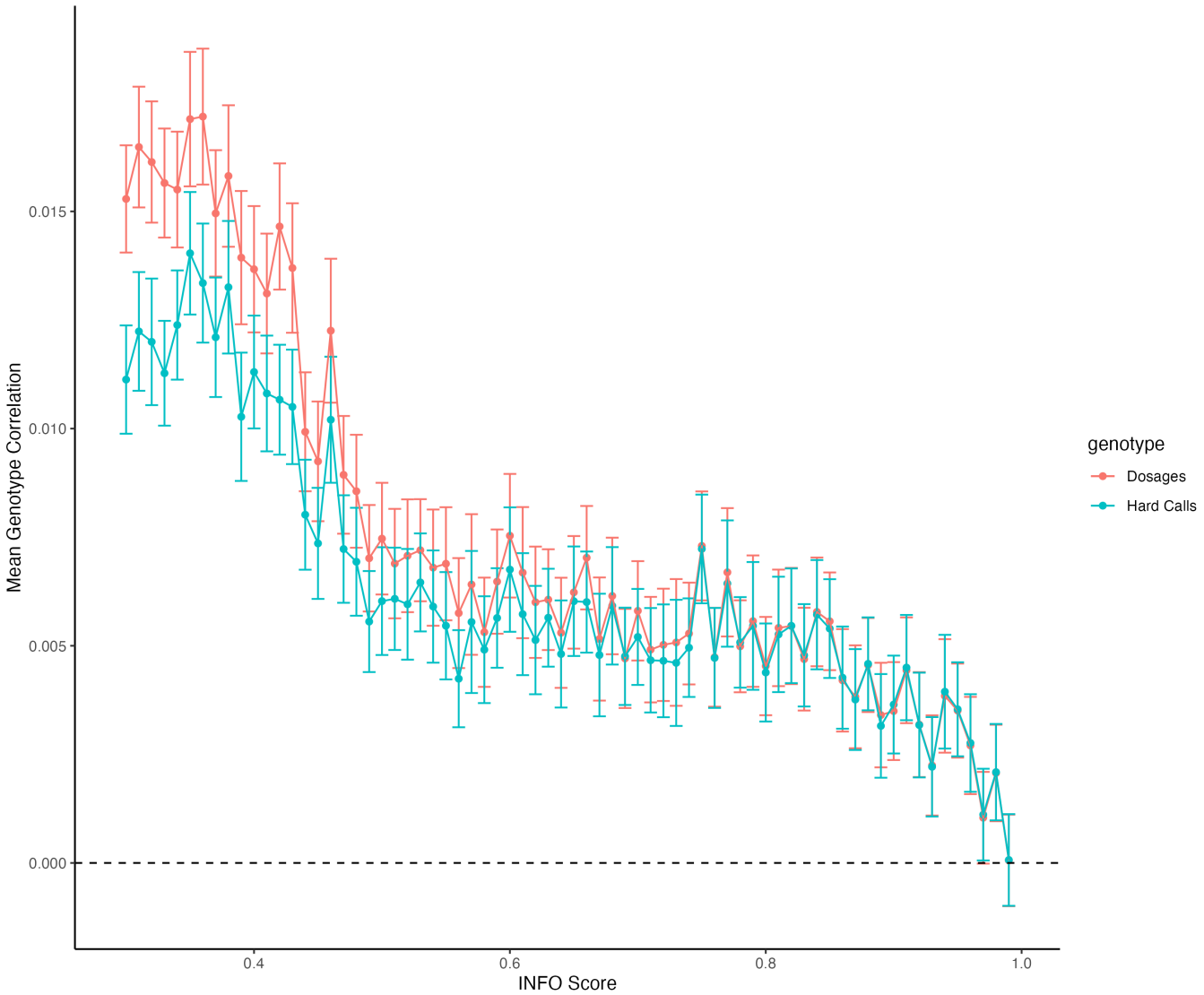
Correlations Dist. Conditional on IBD states

Low Quality SNP - Full Siblings - Dosages - IBD = 2



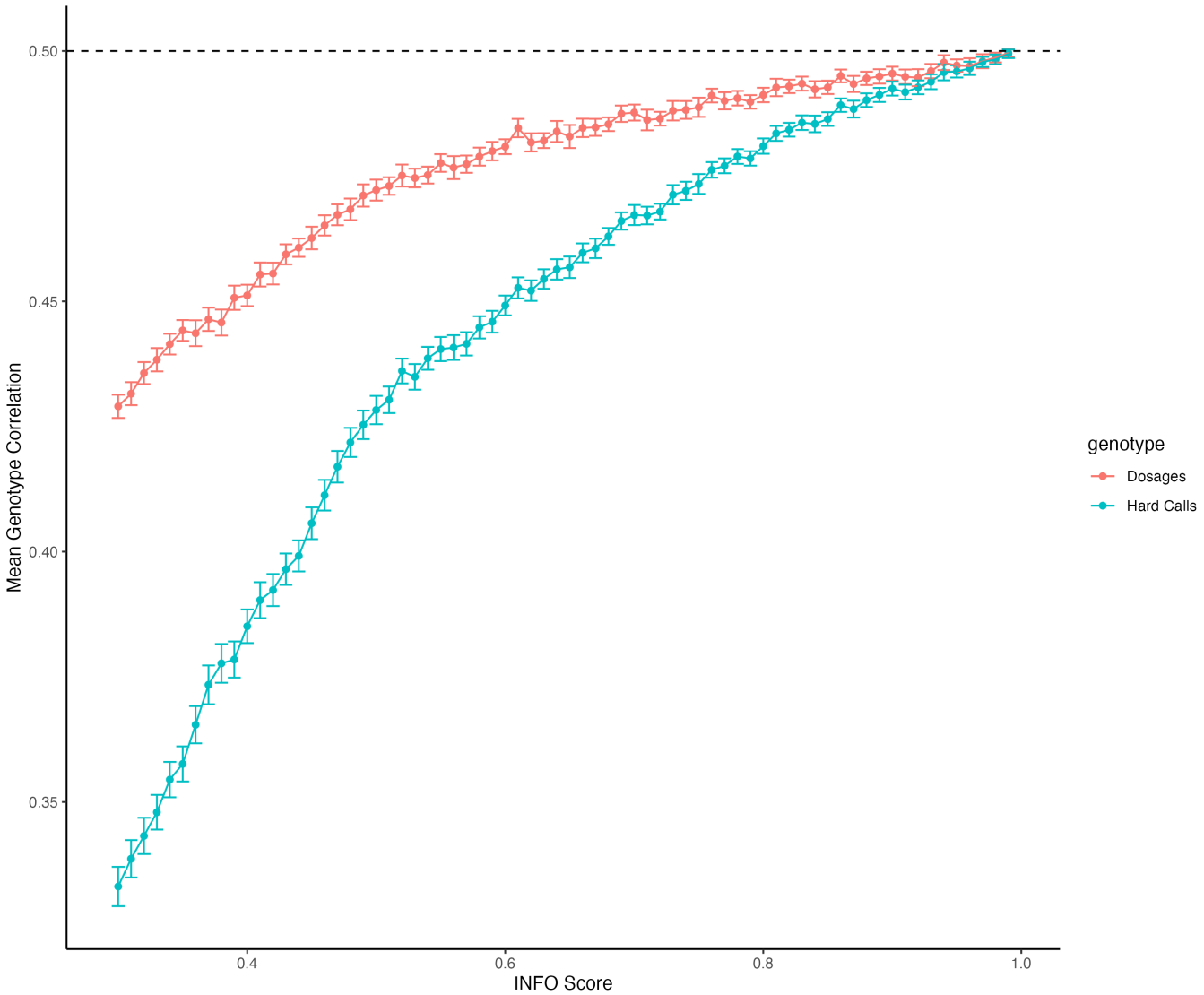
Mean Genotypes (info score)

IBD = 0



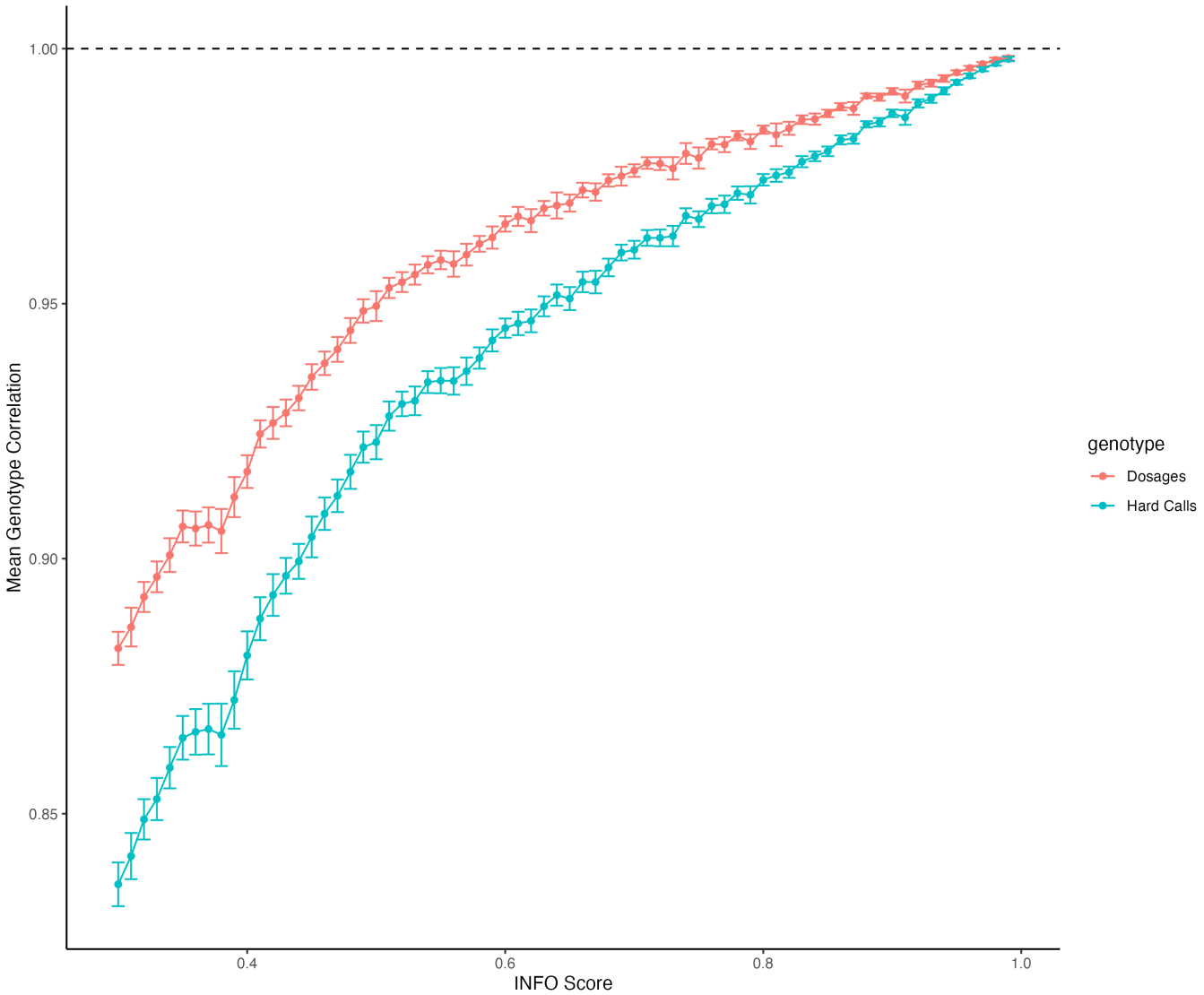
Mean Genotypes (info score)

IBD = 1

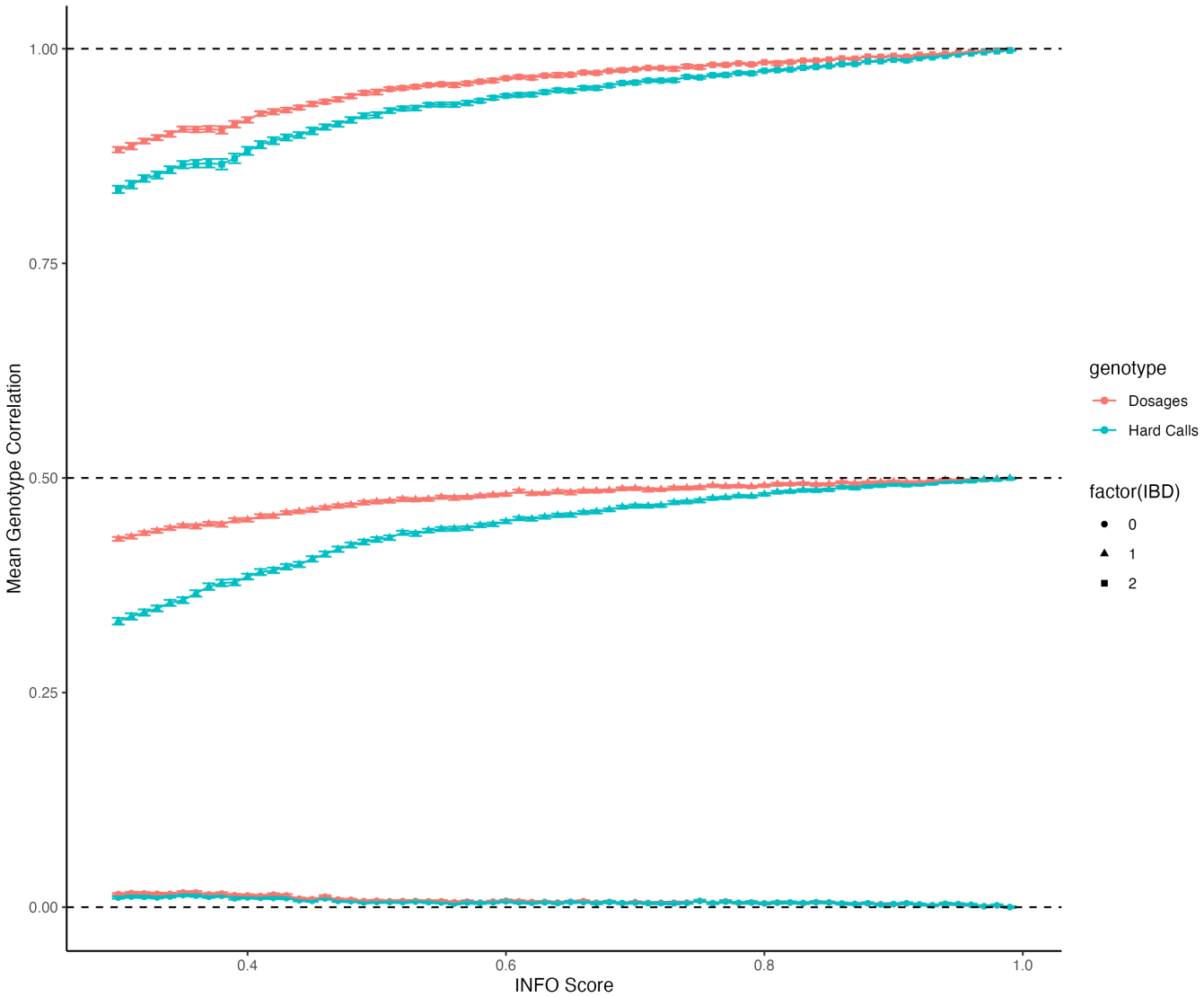


Mean Genotypes (info score)

IBD = 2



Mean Genotypes (info score)

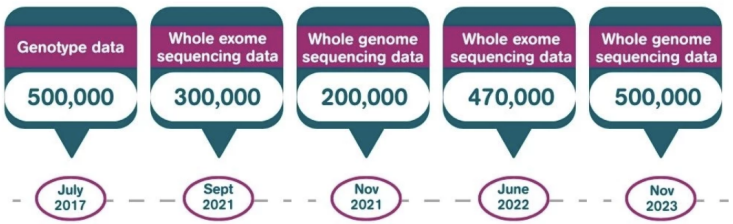


Why We Need WGS Data?

- We showed that imputed data doesn't have the expcted properties for the family based analysis. because the correlation deviates from the theoretical expectations.
- What we are concerned and don't understand is that what is the downstream consequences of this problem is for family based GWAS.
- Ideally, we would compare the imputed genotypes with true genotypes so we can get some idea what are the downstream consequences might be on Sib-GWAS.

WGS Data

- UKB has released whole genome sequencing (WGS) data on 30th November 2023 from half a million volunteers.
- First they released 200K version in 2021.



Whole genome sequencing for 500,000 participants has generated 27.5 petabytes of data!

UKB-RAP Setup

- First you create an account get approved on UKB-AMS.
- Then you should be added as a part of SSGAC on the platform.
- Then using your account you can access the RAP-Platfrom.
- On the RAP platform you have access to different tools like Jupyterlab & R studio.
- You should first create a project then you can access tools & data.
- Upon project creation you should provide your application ID to access the data.

How to access WGS Data

- Getting an application ID is a separate process in which you should apply to get access to the data and you should specify which data and for what purpose you want to use. It takes some days to get approved.
- But fortunately you don't need to do this.
- Get the application ID from Chelsea. If you are already a member of the SSGAC on the RAP you can use the application ID on your personal profile to get access to WGS data.
- You provide the application ID and you should select data bundles you need.
- Tabular data and bulk data files
- *Addtionall bulk data* files for individula level data later. It takes about 1 or 2 days to dispense data to your project.

Some Important Things

- Billing: You can used the the £40 first then you should transfer billing to SSGAC.
- WGS Individual level data is in this path: yourproject:/Bulk/DRAGEN WGS/Whole genome variant call files (VCFs) (DRAGEN) [500k release]
- Will PLINK or BGEN versions be released? PLINK 2.0 and BGEN versions for both the DRAGEN and GraphTyper joint variant calls are planned to be released in 2024.

Some Important Things (cont'd)

- Use TABIX for now with VCF Files. It is much faster than plink2 for VCF files when it is used with .tbi index.
- CAUTION: Individual IDs are getting randomized for each application ID. You might get results that are compeletly random if you don't know this.

Next Step

- We are going to estimate a sib-difference regression on the real WGS data then we can ask what we would obtain if we instead use imputed data.
- Then we can assess the bias that comes from using imputed data.

Thank You for your Attention!

Please Provide me with FEEDBACKS!