1. Today, I’ll talk about the reliability of imputed genotype data in the context of family-based analyses.
2. We are concerned that imputed data especially low-quality imputed genotypes may not be suitable for family-based analyses.   
   In fact, Family and sib-GWAS are designed to use Mendelian inheritance as a natural experiment to obtain unbiased estimates, but with imputation, especially low-quality imputation, we may lose this advantage.
3. Before going forward I want to remind you about the imputation from a reference panel and the difference between hard-calls and dosages genotypes, imputation from a reference panel is when we estimate genotypes at unobserved locations using a reference panel. Basically, the reference panel helps us figure out the probability of different genotypes at those spots. If we go with the genotype that has the highest probability, that’s called “hard calls.” And if we use the expected value of the genotypes instead, we call that “dosages.”
4. One thing worth pointing out is that the reference-based imputation used in UKB data is completely different from the method used in Alex Young’s SNIPAR package. The UKB imputation doesn’t take relationships between individuals into account, and it performs the imputation separately for each person, that means we lose some valuable information from Mendelian inheritance that could have improved the imputation quality.
5. For the analyses I’m about to present, we used 70K randomly selected SNPs with MAF greater than 1% from UKB imputed data. These SNPs have imputation quality (INFO Score) between 0.3 and 1. We used the White British subsample of the UKB, which consists of around19,000 full sibling pairs and 4,000 parent-offspring pairs.
6. So we can assess the imputation quality by looking at siblings and parent-offspring pairs in theory we should see a correlation of half between the genotypes of both sibling and parent-offspring pairs if the data quality is good then we should see this correlation in the data but if we don’t see that we then can have doubts about the imputed data and imputation quality and a natural question would be how much bias does come from the imputation quality?
7. we plotted the empirical correlation as a function of imputation quality or INFO score, we notice that for low-quality SNPs, the correlation is well below 0.5 for both types of pairs, regardless of whether we’re looking at dosages or hard calls. Even for the highest-quality imputed SNPs, there are still deviations from the expected 0.5. This plot also highlights that the issue is more pronounced for parent-offspring pairs and for hard-calls compared to full sibling pairs and dosage genotypes.
8. Another way to test the quality of imputed data is by examining the correlation between siblings' genotypes, conditional on their IBD (Identity by Descent) states. IBD state shows how many alleles siblings share from their parents. If they share both alleles (IBD = 2), the genotype correlation at those locations should be 1. For locations with IBD = 1, the correlation should be 0.5, and for locations where no alleles are shared (IBD = 0), the correlation should be zero.
9. We plotted the mean correlation between siblings' genotypes—both for dosages and hard-calls—across different IBD states as a function of imputation quality.
10. For IBD = 2, we’d expect a perfect correlation of 1, but the results show that the correlation depends on imputation quality and deviates significantly from 1. This discrepancy also occurs for IBD = 1, which can lead to downstream issues in FGWASs.
11. For SNPs with IBD = 1, the expected genotype correlation is 0.5, but the observed values differ from this expectation. As with previous observations, the issue is more pronounced for hard calls compared to dosages.
12. For locations with IBD = 0, the correlation should ideally be zero. However, for low-quality imputed SNPs, we still observe non-zero correlations.
13. After analyzing imputed data, we started using the newly available UKB Whole Genome Sequencing (WGS) data to compare it with the imputed data and understand what might have been missed during imputation. For the analyses I’m about to show, we used 68 high-quality SNPs 46 low-quality SNPs
14. What if we perform a simple regression of WGS data onto imputed data to see how well the imputed data was representing the actual data? We did this and separated the SNPs into high- and low-quality groups, represented by red and blue bars here and we plotted the distribution of resulted slopes.

Essentially, what we are doing here is regressing x on x, so we’d expect to see a slope of 1. For the high-quality group (red), the distribution of slopes for different SNPs is mostly concentrated around 1 which is not bad. However, for the low-quality group (blue), the distribution is closer to zero, with a peak at zero, indicating that the quality imputed data is almost uncorrelated with real data and it is essentially just noise. Even for the very high-quality imputed data, where we’d expect a stronger correlation, there are still significant deviations from a slope of 1 in the distribution.

1. We then examined the R² values from this regression. The INFO score is supposed to represent the R² of this simple regression and serve as a measure of imputation quality in the imputed data. Ideally, when we plot R² against the INFO score, all the points should lie along the 45-degree line. However, what we observe is that for low-quality imputed data, the R² values are mostly zero, except for one point. For high-quality imputed data, while the R² values are closer, there are still noticeable deviations from the 45-degree line. So wan can safely say INFO score is an unreliable metric of imputation quality in UKB imputed data.
2. We conducted another analysis combining WGS and imputed data. In theory, the sibling sum genotypes should be uncorrelated with the sibling difference genotypes. So, we performed a regression with the sibling sum from the WGS data as the dependent variable and the sibling difference from the imputed data as the independent variable. Ideally, we’d expect to see a slope of zero because, theoretically, these two variables should have no correlation.
3. We plotted the qq plot for zstatistics of this regression, under the null hypothesis which is having slope of zero here the distribution of zstatistics shold be normal and the points in the qq plot should lie on the 45 degree line For low-quality imputed data, we observe that the dots don’t lie on the 45-degree line and that indicates we have data quality issues especially when we compare it to high quality imputed data. For the high quality group we have less problems but we still see assymetric deviation from the line.
4. What if we use only WGS data in this regression? We Plotted the same normal Q-Q plot for the slope, we still observe issues with data quality. We expect fewer quality issues in WGS data compared to imputed data, but some problems still persist.
5. We also used array data and did the same analyses we did on imputed and WGS data using 2000 SNPs for that and we got similar results to what we obtained from the imputed and WGS data.
6. One interesting relationship that we saw using genotyped array data is that we saw a relationship between the MAF and the level of

significance of the slope for the sib-sum sib difference regression, We see more significant slopes for lower MAF values. we don’t have an explanation for this yet.

We don’t yet have an explanation for this finding yet, it’s one of our most recent results. If anyone has insights on this, we’d appreciate hearing them.

1. Thanks for listening, we also have a postdoc position in our group working with Alex Young.