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, FGWASs 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. You already know what imputation is, but just to quickly remind you—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. This type of reference-based imputation doesn’t take pedigree or relationships between individuals into account. Instead, it performs the imputation separately for each individual, which means we lose some valuable information from Mendelian inheritance that could have improved the imputation.
5. For the analyses I’m about to present, we used UKB imputed data. Specifically, we used SNPs with a minor allele frequency greater than 1% and an INFO score (or imputation quality) between 0.3 and 1. We randomly selected 70,000 SNPs within this range. We only focused on the subsample of White British individuals, which included approximately 19,000 full sibling pairs and 4,000 parent-offspring pairs.
6. We analyzed the correlation between full sibling pairs and parent-offspring pairs. In theory, we’d expect a correlation of 0.5 for both types of pairs. This serves as a test for the imputed data. When we plot the 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.
7. 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 indicates 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.
8. We plotted the mean correlation between siblings' genotypes—both for dosages and hard-calls—across different IBD states as a function of imputation quality.
9. 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.
10. 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.
11. For locations with IBD = 0, the correlation should ideally be zero. However, for low-quality imputed SNPs, we still observe non-zero correlations.
12. After analyzing only 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 (mean INFO score of 0.96) and 46 low-quality SNPs (mean INFO score of 0.31) that are also present in the WGS data. this is a small sample, it’s because we didn’t have access to pre-processed WGS data and had to process the raw VCF files ourselves. Additionally, working on the RAP platform, which is required for WGS data, have some limitations that slowes us down. We were also mindful of the costs associated with using RAP for these analyses. For this reason, we kept the sample small but plan to extend the analysis to a larger set of SNPs in the future. The subsample we used consisted of around 19,000 White British individuals, with SNPs having a minor allele frequency greater than 1%. We excluded multi-allelic SNPs and focused only on bi-allelic SNPs.
13. Now that we have WGS data, we don’t necessarily need to rely on imputed genotypes. So, we can ask: 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 analysis and separated the SNPs into high- and low-quality groups, represented by red and blue bars, respectively.

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 imputed data is almost uncorrelated and 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.
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. For this regression, if we plot a normal Q-Q plot, we’d expect the dots not to lie on the 45-degree line because, with a slope of zero, the z-statistics of the slope should follow a normal distribution. For low-quality imputed data, we observe that the dots don’t lie on the 45-degree line, which might seem surprising. However, our explanation is that this happens because the data is mostly noise, not because it’s high-quality.
4. On the other hand, when we look at the high-quality group, we see dots lying on the 45-degree line, which is contrary to what we’d expect—we’d want to see dots that deviate from the red line. This provides further evidence that the imputed data is not as reliable as we’d hope.
5. Now, let’s consider using only WGS data for this regression—using WGS data on both sides instead of including imputed data on the right-hand side. If we then draw the same normal Q-Q plot for the slope, we still observe issues with data quality.  
   While WGS is the highest-quality data available, we still expect fewer quality issues in WGS data compared to imputed data, but some problems persist.
6. We also used genotyped array data and did the same analyses we did on imputed and WGS data we used 2000 SNPs for that and we got similar results to what we obtained from the imputed and WGS data.
7. 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 of the sib-sum sib difference regression, we don’t have a explanation for this observation yet it is one of our latest results and we would be thankful if anyone has any opinion on this.
8. One interesting observation we made using genotyped array data was a relationship between the minor allele frequency (MAF) and the level of significance of the slope in the sibling-sum and sibling-difference regression. We see more significant slopes for lower MAF values.

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