1. Today, I will discuss the reliability of imputed genotype data in the context of family-based analyses.

1. 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.
2. As you may recall, and I have previously presented. I conducted analyses first on the imputed data alone. I demonstrated that, contrary to theoretical expectations, the genotype correlation between siblings and parent-offspring pairs, which should be 0.5 for both, depends on the imputation quality. Even for high-quality imputed SNPs with high info scores, the correlation differed from 0.5 for both sibling pairs and parent-offspring pairs. The correlation was worse when hard-calls were used instead of dosages, and it was worse for parent-offspring pairs compared to sibling pairs.
3. Now, consider FGWAS models for sibling pairs from family i: Y and g represent the phenotype and genotype respectively, gpar is the sum of paternal and maternal genotypes in family i, delta is the direct genetic effect and that is target of FGWAS and sib-GWAS analyses, and alpha is the average non-transmitted coefficient or NTC.
4. We can rewrite yi1 in terms of g1-g2 and g1+g2, the resulted equation is this equation on here. Since g1-g2 and g1+g2 are uncorrelated or orthogonal, the estimation of this model will provide an unbiased estimate of delta, which is the goal of FGWASs and sib-GWASs
5. We can then consider what would happen if we used imputed data instead of real data. Or how much bias comes from using imputed data. Here we take the correlation of both sides of the previous slide equation with g1-g2 from imputed data.

1. If we regress g1-g2 from WGS data onto g1-g2 from imputed data, it reveals the bias from sib-difference, and regressing g1+g2 from WGS onto g1-g2 from imputed data shows the bias from sib sum components. The slopes from these regressions allow us to quantify the bias and confounding in terms of delta and alpha from the FGWAS model. That’s a nice feature of this setup because we can quantify the confounding and bias in terms of the original FGWAS model.
2. I have estimated two types of models for all the white British sibling pairs in the UKB data. The first type, called minus models, has the difference between siblings' genotypes from WGS data on the left-hand side and the difference from imputed data on the right-hand side. Essentially, in the minus models, we are regressing x on x because imputed data should match the actual data. We expect the minus models to have an intercept of zero and a slope of 1. The second type, I call it plus models, has the same right-hand side as the minus models, but the left-hand side is now the sum of siblings' genotypes from WGS data instead of their difference. We expect the slope or b here in the plus models to be zero because the sum of genotypes should be orthogonal to their difference, that means the slope should be zero.
3. For the following results I am going to show The data I used consists of 68 high Quality Imputed SNPs with mean info score of 96% and 46% low quality imputed SNPs with mean info score of 31%.  
     
   We used the data for all of the white British sibling pairs in the UKB that is 19 thousand and 52 sibling pairs. We only considered bi allelic SNPs with minor allele frequency of greater than 1%
4. Now, let's examine the results. I have summarized the results for minus models, which involve regressing g1-g2 from WGS data onto g1-g2 from imputed data, using hard-calls imputed data. The left plot shows the distribution of intercepts of all the models fitted for all the SNPs in the sample the histogram on the right shows the distribution of slope or b coefficient in the minus models.

Red bars represent high-quality imputed SNPs, while blue bars represent low-quality imputed SNPs. Vertical lines indicate the theoretical expectations and the empirical means on each plot. The black vertical line shows the theoretical mean which is at 1 for the slope as we except slope to be 1 and it is at zero for the intercept again because we expect a or intercept in the minus models to be zero.

Red vertical line shows the mean of high quality group and the green vertical line shows the low quality group mean.

These plots reveal that for low-quality SNPs, the slope is very different from the theoretical expectation of 1, being closer to zero and showing almost no correlation with real data, as the empirical mean is far from 1. Even for high-quality SNPs, this issue persists. On top of that, for intercepts, these plots show that the distribution should be concentrated around zero for both high and low-quality SNPs, but it is not, especially for low-quality SNPs.

1. What if we used dosages instead of hard calls on the right-hand side of minus models? The right panel displays the previously shown results using hard calls, while the left panel shows the distributions using dosages data. This comparison indicates that the overall distribution is roughly the same when using dosages data, the distribution of high quality snps is more concentrated around 1 for the dosages data but still we have some problems here and for the SNPs with lower quality we have the same problem using dosages data instead of hard-calls imputed data.
2. Now, let's consider plus models, which involve regressing the sum of siblings' genotypes from WGS data onto the genotype difference from imputed data. For this model type, the intercept merely shows the distribution of the dependent variable and lacks meaningful interpretation, so I removed the intercept from these plots, showing only the slope. The theoretical expectation is that the slope should be zero, but these distributions indicate that for both high and low-quality SNPs, the empirical mean differs from the theoretical mean. Here, I used hard-calls data.
3. Again, what if we used dosages instead of hard calls in the right hand side of minus models? The right panel shows previously shown results using the hard calls and the left panel here shows the distributions using dosages data. This comparison shows that the overall distribution is the roughly the same when we use dosages data and and I think it doesn’t have meaningful and significant difference here.
4. Last two slides was only the distribution but here in this slide I show a formal statistical test to demonstrate that b or the slope distribution in the plus models have a problem and doesn’t follow the theoretical expectation. This plot show the normal QQ plot for the slope statistic using dosages data and only for high quality SNPs. We can see there are many points that doesn’t lie on the 45 degree line and a stiatistical test that shows us the distribution of b doesn’t follow theory.
5. Here is the normal QQ plot exactly like the previous slide the only difference here is that this plot is only for low quality SNPs and shows an even worse situation as the dots don’t lie on the 45-degree line.
6. We did another analyses only on the WGS data.   
   regressing the sum of genotypes onto the difference of genotypes, both from WGS data. We are testing whether g1+g2 is uncorrelated to g1 – g2 as it should be based on theory and this also can assess the quality of WGS data. so we would expect a zero slope. This plot shows the histogram of slopes and it shows that they are concentrated around zero.
7. And here is the Normal QQ-plot for the slope of regression of genotype sum on genotype difference.
8. We did a simple regression of genotype from WGS data onto genotype from imputed data for one of the siblings. To test how much imputed data resembles the real data and what would be the R2 in comparison to the info score. These plot shows the distribution of slope for this regression for high and low-quality imputed SNPs. In this regression, we expect a zero intercept and a slope of 1, but the results indicate that this is not the case, especially for low-quality imputed data the slope is very different from 1.

1. Now, let's consider the R² obtained from the simple regression of WGS genotype onto imputed genotype. The info score, based on its definition, should be equal to the R² we obtained. If we draw a scatter plot showing the R² based on the info score alongside the 45-degree line, we expect to see all the dots on the 45-degree or y=x line. However, the plot shows that even for high-quality SNPs, there are points significantly far from the y=x line, and for lower-quality SNPs in our sample, there is almost no correlation between the imputed and WGS data. This indicates that the info score is an unreliable metric of imputation quality.
2. So, we were curious about whether we similar results using the array genotyped data, that is original observed data that is used for imputation and creation of the imputed data using imputation methods. But it is a much smaller dataset compared to imputed data and the WGS data,
3. One interesting result that we saw in the gnotyped data is that when we get the
4. For next steps we are going to extend the analysis to a larger set of SNPs to increase power.

And we are also going to do parental imputation with Snipar using WGS data to make the UKB sample larger for current paper and other analyses in the future.

Today, I will discuss the reliability of imputed genotype data in the context of family-based analyses.

Today, I will talk about the reliability of imputed genotype data in family-based analyses.

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.

We are worried that imputed data, especially low-quality imputed genotypes, may not be appropriate for family-based analyses. FGWASs are intended to use Mendelian inheritance as a natural experiment to obtain unbiased estimates, but with imputation, particularly low-quality imputation, we might lose this benefit.

As you may recall, and I have previously presented. I conducted analyses first on the imputed data alone. I demonstrated that, contrary to theoretical expectations, the genotype correlation between siblings and parent-offspring pairs, which should be 0.5 for both, depends on the imputation quality. Even for high-quality imputed SNPs with high info scores, the correlation differed from 0.5 for both sibling pairs and parent-offspring pairs. The correlation was worse when hard-calls were used instead of dosages, and it was worse for parent-offspring pairs compared to sibling pairs.

As you may remember, I have previously presented my analyses on the imputed data alone. I showed that, contrary to theoretical expectations, the genotype correlation between siblings and parent-offspring pairs, which should be 0.5 for both, depends on imputation quality. Even for high-quality imputed SNPs with high info scores, the correlation was different from 0.5 for both sibling and parent-offspring pairs. The correlation was worse when hard-calls were used instead of dosages, and it was worse for parent-offspring pairs compared to sibling pairs.

Now, consider FGWAS models for sibling pairs from family i: Y and g represent the phenotype and genotype respectively, gpar is the sum of paternal and maternal genotypes in family i, delta is the direct genetic effect and that is target of FGWAS and sib-GWAS analyses, and alpha is the average non-transmitted coefficient or NTC.

Now, consider FGWAS models for sibling pairs from family i: Y and g represent the phenotype and genotype, respectively. Gpar is the sum of paternal and maternal genotypes in family i, delta is the direct genetic effect targeted by FGWAS and sib-GWAS analyses, and alpha is the average non-transmitted coefficient or NTC.

We can rewrite yi1 in terms of g1-g2 and g1+g2, the resulted equation is this equation on here. Since g1-g2 and g1+g2 are uncorrelated or orthogonal, the estimation of this model will provide an unbiased estimate of delta, which is the goal of FGWASs and sib-GWASs

We can rewrite yi1 in terms of g1-g2 and g1+g2, resulting in the equation shown here. Since g1-g2 and g1+g2 are uncorrelated or orthogonal, estimating this model will provide an unbiased estimate of delta, which is the objective of FGWASs and sib-GWASs.

We can then consider what would happen if we used imputed data instead of real data. Or how much bias comes from using imputed data. Here we take the correlation of both sides of the previous slide equation with g1-g2 from imputed data.

We can then consider the impact of using imputed data instead of real data, or how much bias arises from using imputed data. Here, we take the correlation of both sides of the equation from the previous slide with g1-g2 from imputed data.

If we regress g1-g2 from WGS data onto g1-g2 from imputed data, it reveals the bias from sib-difference, and regressing g1+g2 from WGS onto g1-g2 from imputed data shows the bias from sib sum components. The slopes from these regressions allow us to quantify the bias and confounding in terms of delta and alpha from the FGWAS model. That’s a nice feature of this setup because we can quantify the confounding and bias in terms of the original FGWAS model.

Regressing g1-g2 from WGS data onto g1-g2 from imputed data reveals the bias from sib-difference, while regressing g1+g2 from WGS onto g1-g2 from imputed data shows the bias from sib sum components. The slopes from these regressions allow us to quantify the bias and confounding in terms of delta and alpha from the FGWAS model. This is a useful feature of this setup because it enables us to quantify confounding and bias in terms of the original FGWAS model.

I have estimated two types of models for all the white British sibling pairs in the UKB data. The first type, called minus models, has the difference between siblings' genotypes from WGS data on the left-hand side and the difference from imputed data on the right-hand side. Essentially, in the minus models, we are regressing x on x because imputed data should match the actual data. We expect the minus models to have an intercept of zero and a slope of 1. The second type, I call it plus models, has the same right-hand side as the minus models, but the left-hand side is now the sum of siblings' genotypes from WGS data instead of their difference. We expect the slope or b here in the plus models to be zero because the sum of genotypes should be orthogonal to their difference, that means the slope should be zero.

I have estimated two types of models for all the white British sibling pairs in the UKB data. The first type, called minus models, has the difference between siblings' genotypes from WGS data on the left-hand side and the difference from imputed data on the right-hand side. Essentially, in the minus models, we are regressing x on x because imputed data should match the actual data. We expect the minus models to have an intercept of zero and a slope of 1. The second type, called plus models, has the same right-hand side as the minus models, but the left-hand side is now the sum of siblings' genotypes from WGS data instead of their difference. We expect the slope or b in the plus models to be zero because the sum of genotypes should be orthogonal to their difference, meaning the slope should be zero.

For the following results I am going to show The data I used consists of 68 high Quality Imputed SNPs with mean info score of 96% and 46% low quality imputed SNPs with mean info score of 31%.

We used the data for all of the white British sibling pairs in the UKB that is 19 thousand and 52 sibling pairs. We only considered bi allelic SNPs with minor allele frequency of greater than 1%

For the following results, the data I used consists of 68 high-quality imputed SNPs with a mean info score of 96% and 46 low-quality imputed SNPs with a mean info score of 31%.

We used data from all the white British sibling pairs in the UKB, totaling 19,052 sibling pairs. We only considered bi-allelic SNPs with a minor allele frequency greater than 1%.

Now, let's examine the results. I have summarized the results for minus models, which involve regressing g1-g2 from WGS data onto g1-g2 from imputed data, using hard-calls imputed data. The left plot shows the distribution of intercepts of all the models fitted for all the SNPs in the sample the histogram on the right shows the distribution of slope or b coefficient in the minus models. Red bars represent high-quality imputed SNPs, while blue bars represent low-quality imputed SNPs. Vertical lines indicate the theoretical expectations and the empirical means on each plot. The black vertical line shows the theoretical mean which is at 1 for the slope as we except slope to be 1 and it is at zero for the intercept again because we expect a or intercept in the minus models to be zero. Red vertical line shows the mean of high quality group and the green vertical line shows the low quality group mean. These plots reveal that for low-quality SNPs, the slope is very different from the theoretical expectation of 1, being closer to zero and showing almost no correlation with real data, as the empirical mean is far from 1. Even for high-quality SNPs, this issue persists. On top of that, for intercepts, these plots show that the distribution should be concentrated around zero for both high and low-quality SNPs, but it is not, especially for low-quality SNPs.

Certainly, let's continue with the notes:

What if we used dosages instead of hard calls on the right-hand side of minus models? The right panel displays the previously shown results using hard calls, while the left panel shows the distributions using dosages data. This comparison indicates that the overall distribution is roughly the same when using dosages data, the distribution of high-quality SNPs is more concentrated around 1 for the dosages data, but still we have some problems here and for the SNPs with lower quality we have the same problem using dosages data instead of hard-calls imputed data.

What if we used dosages instead of hard calls on the right-hand side of minus models? The right panel displays the previously shown results using hard calls, while the left panel shows the distributions using dosages data. This comparison indicates that the overall distribution is roughly the same when using dosages data. The distribution of high-quality SNPs is more concentrated around 1 for the dosages data, but we still encounter some issues here, and for lower-quality SNPs, we have the same problem using dosages data instead of hard-calls imputed data.

Now, let's consider plus models, which involve regressing the sum of siblings' genotypes from WGS data onto the genotype difference from imputed data. For this model type, the intercept merely shows the distribution of the dependent variable and lacks meaningful interpretation, so I removed the intercept from these plots, showing only the slope. The theoretical expectation is that the slope should be zero, but these distributions indicate that for both high and low-quality SNPs, the empirical mean differs from the theoretical mean. Here, I used hard-calls data.

Now, let's consider plus models, which involve regressing the sum of siblings' genotypes from WGS data onto the genotype difference from imputed data. For this model type, the intercept merely shows the distribution of the dependent variable and lacks meaningful interpretation, so I removed the intercept from these plots, showing only the slope. The theoretical expectation is that the slope should be zero, but these distributions indicate that for both high and low-quality SNPs, the empirical mean differs from the theoretical mean. Here, I used hard-calls data.

Again, what if we used dosages instead of hard calls in the right hand side of minus models? The right panel shows previously shown results using the hard calls and the left panel here shows the distributions using dosages data. This comparison shows that the overall distribution is roughly the same when we use dosages data, and I think it doesn’t have meaningful and significant difference here.

Again, what if we used dosages instead of hard calls on the right-hand side of minus models? The right panel shows previously shown results using hard calls, while the left panel shows the distributions using dosages data. This comparison shows that the overall distribution is roughly the same when we use dosages data, and I think it doesn’t have meaningful and significant differences here.

Last two slides was only the distribution but here in this slide I show a formal statistical test to demonstrate that b or the slope distribution in the plus models have a problem and doesn’t follow the theoretical expectation. This plot show the normal QQ plot for the slope statistic using dosages data and only for high quality SNPs. We can see there are many points that doesn’t lie on the 45 degree line and a stiatistical test that shows us the distribution of b doesn’t follow theory.

The last two slides focused only on the distribution, but here I present a formal statistical test to demonstrate that the b or slope distribution in the plus models has a problem and doesn’t follow the theoretical expectation. This plot shows the normal QQ plot for the slope statistic using dosages data and only for high-quality SNPs. We can see that there are many points that don’t lie on the 45-degree line, and a statistical test shows that the distribution of b doesn’t follow theory.

Here is the normal QQ plot exactly like the previous slide the only difference here is that this plot is only for low quality SNPs and shows an even worse situation as the dots don’t lie on the 45-degree line.

Here is the normal QQ plot, exactly like the previous slide, but the only difference is that this plot is only for low-quality SNPs and shows an even worse situation as the dots don’t lie on the 45-degree line.

We did another analyses only on the WGS data. regressing the sum of genotypes onto the difference of genotypes, both from WGS data. We are testing whether g1+g2 is uncorrelated to g1 – g2 as it should be based on theory and this also can assess the quality of WGS data. so we would expect a zero slope. This plot shows the histogram of slopes and it shows that they are concentrated around zero.

We conducted another analysis only on the WGS data, regressing the sum of genotypes onto the difference of genotypes, both from WGS data. We are testing whether g1+g2 is uncorrelated with g1-g2, as it should be based on theory, and this also assesses the quality of WGS data. We would expect a zero slope. This plot shows the histogram of slopes, and it indicates that they are concentrated around zero.

And here is the Normal QQ-plot for the slope of regression of genotype sum on genotype difference.

Here is the normal QQ plot for the slope of the regression of genotype sum on genotype difference.

We did a simple regression of genotype from WGS data onto genotype from imputed data for one of the siblings. To test how much imputed data resembles the real data and what would be the R2 in comparison to the info score. These plot shows the distribution of slope for this regression for high and low-quality imputed SNPs. In this regression, we expect a zero intercept and a slope of 1, but the results indicate that this is not the case, especially for low-quality imputed data the slope is very different from 1.

We performed a simple regression of genotype from WGS data onto genotype from imputed data for one of the siblings to test how much imputed data resembles the real data and what the R² would be in comparison to the info score. These plots show the distribution of slope for this regression for high and low-quality imputed SNPs. In this regression, we expect a zero intercept and a slope of 1, but the results indicate that this is not the case, especially for low-quality imputed data, where the slope is very different from 1.

Now, let's consider the R² obtained from the simple regression of WGS genotype onto imputed genotype. The info score, based on its definition, should be equal to the R² we obtained. If we draw a scatter plot showing the R² based on the info score alongside the 45-degree line, we expect to see all the dots on the 45-degree or y=x line. However, the plot shows that even for high-quality SNPs, there are points significantly far from the y=x line, and for lower-quality SNPs in our sample, there is almost no correlation between the imputed and WGS data. This indicates that the info score is an unreliable metric of imputation quality.

Now, let's examine the R² obtained from the simple regression of WGS genotype onto imputed genotype. The info score, based on its definition, should be equal to the R² we obtained. If we draw a scatter plot showing the R² based on the info score alongside the 45-degree line, we expect to see all the dots on the 45-degree or y=x line. However, the plot shows that even for high-quality SNPs, there are points significantly far from the y=x line, and for lower-quality SNPs in our sample, there is almost no correlation between the imputed and WGS data. This suggests that the info score is an unreliable metric of imputation quality.

For next steps we are going to extend the analysis to a larger set of SNPs to increase power. And we are also going to do parental imputation with Snipar using WGS data to make the UKB sample larger for current paper and other analyses in the future.

For the next steps, we will extend the analysis to a larger set of SNPs to increase power. We are also going to perform parental imputation with Snipar using WGS data to make the UKB sample larger for the current paper and future analyses.