Qs:

* What is being mean normalized? The mean of genotypes are zero by subtracting the mean from each individual genotype
* Why g1 + g2 is orthogonal to g1-g2?
* How we can re write yi1 in terms of g1-g2 and g1+g2?
* How you calculate this expression?

1

Today I am going to talk about the reliability of imputed genotype data for family based analyses

2

So we are concerned that low quliaty imputed genotypes may be suitable for family based anlayese.

And we know that FGWASs is designed to use the Mendelian inheritance as a clean natural experiment to obtain unbiased estimates but with imputation especially with low quality imputation we may lose that.

3

As you might remember I did some analyses on the imputed data. I showed that contrary to the theoretical expectation the genotype correlation between siblings and parent-offspring pairs which should be equal to 0.5 for both siblings and parent-offspring pairs depends on the imputation quality and even for high quality imputed SNPs with high info score the correlation was different from 0.5. The correlation was worse when we used hard-calls instead of dosages and it is worse for the parent-offspring pairs compared to sibling pairs.

4

So now consider FGWAS models for a sibling pairs from family i:

Y and g are phenotype and genotype of the individual, the gpar is the sum of paternal and maternal genotypes in family i, delta is the direct genetic effect which is target of FGWAS and sib-GWAS analyses, and the alpha is the average non-transmitted coefficient.

5

We can rewrite the yi1 in terms of g1-g2 and g1+g2 and because g1-g2 and g1+g2 are uncorrelated with each other this estimate will give us an unbiased estimate of delta which is the goal of FGWAS and sib-GWAS

6

And so we can then ask what if we used imputed data instead of the real data?

7

And then if we regress g1-g2 from WGS data onto g1-g2 from imputed data that gives us the bias coming from sib-difference and the regression of g1+g2 from WGS onto g1-g2 from imputed data gives us the bias coming from sib sum compnents.

So the slopes from these two regressions enables us to quantify the quantify the bias and confounding in terms of the delta and alpha from the FGWAS model.

8

So we have two types of models that I have estimated for all the white British sib pairs in the UKB data. In the first type of models which I call it minus models we have on the left hand side the difference between sibling’s genotypes from WGS data and on the right-hand side we have the difference between the sibling’s genotypes from imputed data. So basically we can say in the minus models we are regressing x on x because imputed data is supposed to be the same as the actual data. So for the minus models we expect to see intercept of zero and slope of 1. For the second type of models here which I call it plus models the right hand side is the same as the minus models but the left hand side now is the sum of the sibling’s genotypes from WGS data instead of their difference. For the plus models we expect the slope or beta here to be zero that is because the sum of genotypes should be orthogonal to their difference so that means the slope should be zero.

9

So now let’s look at the results. Here I have summarized the results for minus models basically meaning the regression of g1-g2 from WGS data onto g1-g2 from imputed data. I used hard-calls imputed data here. The left column of plots shows the distribution of the intercepts and the right column shows the distribution of the slope. Red bars shows the distribution of high quality imputed SNPs and blue bars are for low quality imputed data.

The first row of plots shows both high and low quality imputed distribtions of intercept for comparsion, the second row shows the intercept and slope distribution only for high quality SNPs the third row is only for low quality SNPs. Vertical lines shows the theoretical expectations on each of the plots and the empirical means. These plots show that for the low quality SNPs the slope is very different from the theoretical expectation of 1 it is actually more close to zero and so having no correlation with real data as the empirical mean of is very different from 1 and even for the high quality SNPs this problem is still there. And also for the intercepts these plots show that for both high and low quality SNPs the distrubtion of intercepts should be concentrated around zero but it is not for both high and low quality SNPs especially for the low quality SNPs.

10

Now considering plus models which is the regression of sum of the sibling’s genotypes from WGS data onto the genotype difference from imputed data. For this type of models the intercept just shows the distribution of dependent variable and doesn’t have a meaningful interpretation so I removed the intercept from these plots here I just have the slope for these plot. The theoretical expectation is that the slope should be zero but these distributions show us that for both high and low quality SNPs the empirical mean is different from the theoretical mean. Here I used the hard calls data.

11

Here in this table I summarized some of the statistics about plus and minus models. We used 68 high quality SNPs with mean info score of 96% and 48 low quality SNPs with mean info score equal to 31%.

12

So what if I 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 doesn’t have meaningful and significant difference.

13

Here is the distribution of slopes using dosages data in comparison to hard calls for the plus models. This shows pretty much the same distribution using dosages data instead of hard-calls for both high and low quality SNPs for the plus models.

14

We also did a simple regression of genotype from WGS data onto genotype from imputed data for one of the siblings, here these plots show the distribution of intercept and slope for this regression for high and low quality imputed SNPs, in this regression we expect a zero intercept and slope of 1 but results show that especially for the low quality imputed data that’s not the case.

15

Now we consider the R2 we obtained from the simple regression of WGS genotype onto imputed genotype. Info score based on its defitnition should be equal to the R2 we obtained so 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 degress or y=x line but what we see is instead is this plot that shows even for the high quality SNPs there are points that are significantly far from the y=x line and for the lower quality SNPs in our sample there is absolutely no correlation. So that tells us info score is an unreliable metric of imputation quality.

16

We can also test the quality of WGS data by regressing the sum of genotypes onto the difference of genotypes both from WGS data, in theory the sum and difference are uncorrelated so we would expect a zero slope, this plot shows the slope that is concentrated around the zero which shows us the WGS data has a good quality.

1

Today I am going to talk about the reliability of imputed genotype data for family based analyses

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

2

So we are concerned that low quliaty imputed genotypes may be suitable for family based anlayese. And we know that FGWASs is designed to use the Mendelian inheritance as a clean natural experiment to obtain unbiased estimates but with imputation especially with low quality imputation we may lose that.

We are concerned that low-quality imputed genotypes may not be suitable for family-based analyses. 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

As you might remember I did some analyses on the imputed data. I showed that contrary to the theoretical expectation the genotype correlation between siblings and parent-offspring pairs which should be equal to 0.5 for both siblings and parent-offspring pairs depends on the imputation quality and even for high quality imputed SNPs with high info score the correlation was different from 0.5. The correlation was worse when we used hard-calls instead of dosages and it is worse for the parent-offspring pairs compared to sibling pairs.

As you may recall, I conducted analyses on the imputed data. 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. The correlation was worse when hard-calls were used instead of dosages, and it was worse for parent-offspring pairs compared to sibling pairs.

4

So now consider FGWAS models for a sibling pairs from family i: Y and g are phenotype and genotype of the individual, the gpar is the sum of paternal and maternal genotypes in family i, delta is the direct genetic effect which is target of FGWAS and sib-GWAS analyses, and the alpha is the average non-transmitted coefficient.

Now, consider FGWAS models for sibling pairs from family i: Y and g represent the phenotype and genotype of the individual, 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.

5

We can rewrite the yi1 in terms of g1-g2 and g1+g2 and because g1-g2 and g1+g2 are uncorrelated with each other this estimate will give us an unbiased estimate of delta which is the goal of FGWAS and sib-GWAS

We can rewrite yi1 in terms of g1-g2 and g1+g2, and since g1-g2 and g1+g2 are uncorrelated, this estimate will provide an unbiased estimate of delta, which is the goal of FGWAS and sib-GWAS.

6

And so we can then ask what if we used imputed data instead of the real data?

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.

7

And then if we regress g1-g2 from WGS data onto g1-g2 from imputed data that gives us the bias coming from sib-difference and the regression of g1+g2 from WGS onto g1-g2 from imputed data gives us the bias coming from sib sum compnents. So the slopes from these two regressions enables us to quantify the quantify the bias and confounding in terms of the delta and alpha from the FGWAS model.

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.

8

So we have two types of models that I have estimated for all the white British sib pairs in the UKB data. In the first type of models which I call it minus models we have on the left hand side the difference between sibling’s genotypes from WGS data and on the right-hand side we have the difference between the sibling’s genotypes from imputed data. So basically we can say in the minus models we are regressing x on x because imputed data is supposed to be the same as the actual data. So for the minus models we expect to see intercept of zero and slope of 1. For the second type of models here which I call it plus models the right hand side is the same as the minus models but the left hand side now is the sum of the sibling’s genotypes from WGS data instead of their difference. For the plus models we expect the slope or beta here to be zero that is because the sum of genotypes should be orthogonal to their difference so 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, 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 beta in the plus models to be zero because the sum of genotypes should be orthogonal to their difference, meaning the slope should be zero.

9

So now let’s look at the results. Here I have summarized the results for minus models basically meaning the regression of g1-g2 from WGS data onto g1-g2 from imputed data. I used hard-calls imputed data here. The left column of plots shows the distribution of the intercepts and the right column shows the distribution of the slope. Red bars shows the distribution of high quality imputed SNPs and blue bars are for low quality imputed data. The first row of plots shows both high and low quality imputed distribtions of intercept for comparsion, the second row shows the intercept and slope distribution only for high quality SNPs the third row is only for low quality SNPs. Vertical lines shows the theoretical expectations on each of the plots and the empirical means. These plots show that for the low quality SNPs the slope is very different from the theoretical expectation of 1 it is actually more close to zero and so having no correlation with real data as the empirical mean of is very different from 1 and even for the high quality SNPs this problem is still there. And also for the intercepts these plots show that for both high and low quality SNPs the distrubtion of intercepts should be concentrated around zero but it is not for both high and low quality SNPs especially for the low quality SNPs.

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 column of plots shows the distribution of intercepts, and the right column shows the distribution of slopes. Red bars represent high-quality imputed SNPs, while blue bars represent low-quality imputed data. The first row of plots shows both high and low-quality imputed distributions of intercepts for comparison, the second row shows the intercept and slope distribution only for high-quality SNPs, and the third row is only for low-quality SNPs. Vertical lines indicate the theoretical expectations and the empirical means on each plot. 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 no correlation with real data, as the empirical mean is far from 1. Even for high-quality SNPs, this issue persists. Additionally, 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.

10

Now considering plus models which is the regression of sum of the sibling’s genotypes from WGS data onto the genotype difference from imputed data. For this type of models the intercept just shows the distribution of dependent variable and doesn’t have a meaningful interpretation so I removed the intercept from these plots here I just have the slope for these plot. The theoretical expectation is that the slope should be zero but these distributions show us that for both high and low quality SNPs the empirical mean is different from the theoretical mean. Here I used the 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.

11

Here in this table I summarized some of the statistics about plus and minus models. We used 68 high quality SNPs with mean info score of 96% and 48 low quality SNPs with mean info score equal to 31%.

This table summarizes some statistics

--The response length has exceeded the model's configured limit. If you wish, I can complete responses for specific sections with your next prompt.

Certainly, let's continue with the notes:

12

So what if I 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 doesn’t have meaningful and significant difference.

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, with no meaningful or significant differences.

13 Here is the distribution of slopes using dosages data in comparison to hard calls for the plus models. This shows pretty much the same distribution using dosages data instead of hard-calls for both high and low quality SNPs for the plus models.

Here is the distribution of slopes using dosages data compared to hard calls for the plus models. This shows that the distribution is quite similar when using dosages data instead of hard calls for both high and low-quality SNPs in the plus models.

14 We also did a simple regression of genotype from WGS data onto genotype from imputed data for one of the siblings, here these plots show the distribution of intercept and slope for this regression for high and low quality imputed SNPs, in this regression we expect a zero intercept and slope of 1 but results show that especially for the low quality imputed data that’s not the case.

We also conducted a simple regression of genotype from WGS data onto genotype from imputed data for one of the siblings. These plots show the distribution of intercept and 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.

15 Now we consider the R2 we obtained from the simple regression of WGS genotype onto imputed genotype. Info score based on its defitnition should be equal to the R2 we obtained so 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 degress or y=x line but what we see is instead is this plot that shows even for the high quality SNPs there are points that are significantly far from the y=x line and for the lower quality SNPs in our sample there is absolutely no correlation. So that tells us info score is an unreliable metric of imputation quality.

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 absolutely no correlation. This indicates that the info score is an unreliable metric of imputation quality.

16 We can also test the quality of WGS data by regressing the sum of genotypes onto the difference of genotypes both from WGS data, in theory the sum and difference are uncorrelated so we would expect a zero slope, this plot shows the slope that is concentrated around the zero which shows us the WGS data has a good quality.

We can also assess the quality of WGS data by regressing the sum of genotypes onto the difference of genotypes, both from WGS data. In theory, the sum and difference are uncorrelated, so we would expect a zero slope. This plot shows the slope concentrated around zero, indicating that the WGS data is of good quality.