**UKB Sanity Checks Document:**

6/17/2022

There are several ways we can confirm that the genetic and phenotypic data was processed correctly, including a first stage and effect allele frequencies consistent with estimates produced by outside researchers, exploring the imputation quality of different genetic variants, and comparing our observational and MR results to those we found in the Health and Retirement Study.

The first stage regressions, where we regress body mass index on genetic variants, can be compared to results from the genome-wide association study (GWAS) by Locke et al. and a UK Biobank-focused GWAS replication effort by the Neale Lab. These effect estimates (represented as Betas) and standard errors should be extremely close in value.

We selected the variants based on Locke et al.’s GWAS that explored genetic variants with significant effects on body mass index in a European sample, so while our effect sizes should not be identical, they should at least be directionally consistent and highly correlated. The Neale Lab performed the same GWAS for BMI using the UK Biobank dataset directly. While we have a different sample based on our inclusion criteria, our effect estimates for different genetic variants on body mass index should align very closely with Neale et al.’s results.

Next, I explored the effect allele frequencies in our dataset compared to those found in Neale and Locke. An effect allele frequency is the average proportion of alleles in the population that are effect alleles for each genetic variant. If our effect allele frequencies are similar to those found in the other two studies, it increases both the probability that we selected and processed the genetic variants correctly and that we did not accidentally flip an effect and non-effect allele for any variants.

Because we used imputed genotype data, the imputation quality could have an effect on our results. One way of measuring imputation quality is the INFO Score, which measures what the effective sample size is for each imputed genetic variant. Generally, the strictest threshold in the literature is that each genetic variant has an imputation score at least as high as 0.8.

Lastly, we can explore the effect size differences between our observational and MR models in the Health and Retirement Study and UK Biobank. While the small sample size of the HRS (and comparatively strong first stage) might suggest the estimates will diverge somewhat, an enormous effect size divergence would come as a surprise.

**Important Limitation:**

Unfortunately, because an MR survival analysis has never been applied to explore the effect of BMI on risk of CVD, we lack any precedent to explore the plausibility of the second stage regressions (which for the IVW and weighted median models is incident CVD regressed on the genetic variants).

The key point for this second stage is that if the first stage passes muster, as long as our incident CVD case definition is being processed and we have follow up time coded correctly, then there’s no obvious reason the second stage should be inaccurate.

***TABLE OF CONTENTS (Link to Rest of Document)***

***Sanity Checklist:***

1. *[Beta and Standard Error Comparison with Locke et al. and Neale Lab](#BETA)*
2. [*Effect Allele Comparison with Locke et al. and Neale Lab*](#EAF)
3. [*Genetic Variant Imputation Quality Summary*](#IMPUTE)
4. *Beta and Standard Error Comparison with Locke et al. and Neale Lab*

Relevant Documents:

“Replicating Results from Neale Lab BMI GWAS.docx”

“BetaComparisonsFileNealeLockeUs.csv”

“LockenNealeBetas.csv”

The raw data comparing the effect and standard error estimates between our dataset, Locke et al., and the Neale Lab is in “BetaComparisonsFileNealeLockeUs.csv.” The regression itself is relatively simple, as it included only biological sex and the first 10 principal components to match the process detailed by the Neale Lab at http://www.nealelab.is/blog/2017/9/11/details-and-considerations-of-the-uk-biobank-gwas. I summarize the results of these comparisons briefly below.

Using the R File “Extracting Neale Lab Estimates UKB GWAS.R,” I found that our results correlated extremely well with those of both the Neale Lab and Locke et al., with a correlation of 0.969 between the beta coefficients estimated in our dataset and those estimated by the Neale Lab and a correlation of 0.999 between our estimated standard errors and those estimated by the Neale Lab. Likewise, our results correlate strongly with Locke, as the beta coefficients and standard errors have a 0.873 and 0.927 correlation, respectively. The fact that these estimates align closely – and that the Neale Lab results are more highly correlated with ours than Locke et al. (which did not make use of the UK Biobank) increases my confidence in our results.

Below I’ve created a 4-genetic variant comparison of betas and standard errors between studies to show how closely these variants’ estimates match those produced by the Neale Lab and Locke et al.

**Table 1: Comparison of Betas and SEs for Four Variants from Our Dataset and Neale Lab**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variants (Chr:Pos)** | **Beta From Our Dataset** | **SE From Our Dataset** | **Beta from Neale Lab** | **SE From Neale Lab** | **Difference in Betas** | **Difference in SEs** |
| **1:110154688** | 0.069148 | 0.008756 | 0.0871 | 0.008041 | 0.017952 | 0.000715 |
| **1:177889480** | 0.047896 | 0.003433 | 0.05877 | 0.003139 | 0.010874 | 0.000294 |
| **1:201784287** | 0.019715 | 0.002808 | 0.02419 | 0.002568 | 0.004475 | 0.00024 |
| **1:47684677** | 0.017234 | 0.002826 | 0.01386 | 0.002584 | 0.003374 | 0.000242 |

Note: Average Difference in Betas between our estimates and Neale Lab’s was 0.005 and the Average Difference in SEs was 0.0003.

**Table 2: Comparison of Betas and SEs for Four Variants from Our Dataset and Locke et al.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variants (Chr:Pos)** | **Beta From Our Dataset** | **SE From Our Dataset** | **Beta from Locke et al.** | **SE From Locke et al.** | **Difference in Betas** | **Difference in SEs** |
| **1:110154688** | 0.069148 | 0.008756 | 0.066 | 0.009 | 0.003148 | 0.000244 |
| **1:177889480** | 0.047896 | 0.003433 | 0.048 | 0.004 | 0.000104 | 0.000567 |
| **1:201784287** | 0.019715 | 0.002808 | 0.02 | 0.003 | 0.000285 | 0.000192 |
| **1:47684677** | 0.017234 | 0.002826 | 0.017 | 0.003 | 0.000234 | 0.000174 |

Note: Average Difference in Betas between our estimates and Locke et al.’s was 0.006 and the Average Difference in SEs was 0.0005.

1. *Effect Allele Comparison with Locke et al. and Neale Lab*

Relevant Documents:

“NealeUsEAFComp.csv”

The Effect Allele Frequency is a measure of the proportion of alleles (out of 2 per person) that were the effect allele (the allele that increases BMI) as compared to how many were the non-effect allele. It should tell us whether the allele distribution is similar between our dataset and the GWAS in Locke et al. and the Neale Lab. Close alignment of the effect allele frequencies would further my confidence that our genetic data was processed correctly and none of the effect alleles were flipped on accident.

Again, I present a brief summary of these results, with more complete results available in “NealeUsEAFComp.csv.” The effect allele frequencies in our dataset correlate extremely well with those of Locke et al. and the Neale Lab, with a correlation coefficient of 0.978 and 0.983, respectively. I show what this looks like for a sample of four genetic variants. I focus here on the Neale Lab results, as they should most closely align with our own.

**Table 3: Comparison of Effect Allele Frequency for Four Variants from Our Dataset and Neale Lab**

|  |  |  |
| --- | --- | --- |
| **Variants (Chr:Pos)** | **EAF from Our Dataset** | **EAF from Neale Lab** |
| **1:110154688** | 0.02589 | 0.025932 |
| **1:177889480** | 0.2076 | 0.207557 |
| **1:201784287** | 0.5672 | 0.566888 |
| **1:47684677** | 0.417 | 0.41664 |

It’s readily apparent that the effect allele frequencies are virtually identical between our estimates and those of the Neale Lab, which furthers my confidence in our results.

1. *Genetic Variant Imputation Quality Summary*

Relevant Documents:

SNPINFO.csv

The INFO Score is a measure of the imputation quality of the genetic variants. It can be thought of as the effective sample size for a given genetic variant. In practice, I’ve seen a lot of researchers propose some cutoff for the INFO Score, under which they would not make use of a variant. It is scaled between 0 and 1, with a 1 denoting perfect imputation. This is not a significant issue in our analysis, as each of the included variants has an INFO Score in excess of 0.95 (denoting extremely high genotype imputation quality). The full list of INFO Scores by variant is available in this Folder in “SNPINFO.csv”.