Sociogenomics 2022/23

Final project

José Gabriel Escarraman
Benedetta Felici
Valentina Idda

Part 1. Work with genomics data.

a. Quality control

In the quality control process, several filters were applied to ensure the reliability and integrity of the genomics data. Firstly, variants with a Minor Allele Frequency (MAF) below 0.05 were excluded from the analysis, as they may have limited impact on the population. Next, individuals with a genotyping rate below 95% were removed from the dataset to ensure high-quality genotype data. This step helps eliminate individuals with a high proportion of missing genotypes, which can introduce bias or affect downstream analyses.

Additionally, SNPs with a call rate below 95% were filtered out to exclude variants with a low rate of successful genotyping. This filter ensures that only SNPs with sufficient data quality are included in the analysis. Lastly, SNPs that deviated significantly from the Hardy-Weinberg equilibrium (HWE) with a p-value below 0.00001 were excluded. This step helps identify and remove SNPs that may be subject to genotyping errors or other issues affecting their distribution.

By applying these quality control measures, we end up with 781,538 variants and 1092 individuals.

b. GWAS results

The Additive Genome-Wide Association Study (GWAS) on education aimed to explore the genetic factors associated with educational attainment. The analysis was performed on a dataset consisting of 781,538 variants and 1,092 individuals. The individuals included 525 males and 567 females. The analysis was performed on a quantitative phenotype, specifically on the standardized values of educational attainment. The calculation of allele frequencies and genotyping rate yielded satisfactory results, with a total genotyping rate of 1, indicating good data quality.

The GWAS results revealed several variants that showed statistically significant associations with the phenotype of interest. We observed a strong association (p < 0.05) in more than 137 thousand variants, suggesting that these SNP's may represent loci that are potentially linked to educational attainment.

c. 10 variants with the smallest p-value.

CHR	SNP	BP	NMISS	BETA	SE	R2	T	P
20	rs11699751	41436729	1092	0.2655	0.04342	0.03315	6.114	1.355e-09
6	rs1341585	128900663	1092	-0.3777	0.06423	0.03074	-5.879	5.468e-09
5	rs10223241	88412388	1092	0.3472	0.05982	0.02998	5.804	8.492e-09
2	rs11674589	239638781	1092	0.3027	0.05411	0.02792	5.595	2.786e-08
14	rs11628824	22580762	1092	0.2743	0.04980	0.02709	5.509	4.495e-08
3	rs13059163	137189600	1092	0.3369	0.06120	0.02706	5.506	4.583e-08
8	rs7843555	5206751	1092	-0.2344	0.04305	0.02648	-5.445	6.408e-08
13	rs7328309	97204926	1092	-0.2444	0.04500	0.02635	-5.431	6.893e-08
1	rs6704508	34756509	1092	-0.2261	0.04164	0.02634	-5.430	6.957e-08
6	rs9502570	7258617	1092	-0.2219	0.04099	0.02618	-5.413	7.607e-08

The analysis of the dataset focused on identifying the variants with the strongest associations with educational attainment. After filtering the results, we narrowed down the dataset to the top 10 variants with the smallest p-values. These variants exhibited significant associations with educational outcomes, suggesting their potential role in influencing educational attainment. These variants, found on different chromosomes, indicate that the genetic variations influencing educational attainment are not exclusive to a specific genomic region.

The NMISS column is equal to the total amount of individuals (1092) for all variants, indicating that there is no missing data associated with the SNPs for any of the variants. This comprehensive inclusion of data allows for a robust analysis of the genetic associations with educational attainment.

The effect sizes of these 10 variants exhibit a mix of both negative and positive BETA values, suggesting the presence of both increase-impact and reduction-impact alleles with respect of phenotype. The absolute values of the BETA coefficients, ranging between 0.22 and 0.35, indicate a moderate effect size for these variants on educational attainment.

While these 10 variants demonstrate a significant association with educational attainment, it is important to note that the variance explained by them is relatively small. The R2 values, ranging from 0.02618 to 0.03315, suggest that these variants collectively explain only a small proportion of the total variance in educational attainment. This implies that other genetic and environmental factors beyond these specific variants play a role in determining educational attainment. In summary, these variants exhibit moderate effect sizes, representing a mix of negative and positive impact alleles. However, the cumulative impact of these variants on the observed variance in educational attainment is relatively modest.

Part 2. Polygenic risk score.

A polygenic risk score informs about the impacts of many SNPs on the desired phenotype, in our case education. Specifically, it is the sum of alleles' counts for each SNPs multiplied by their weight, that is given by GWAS summary statistics. To compute our polygenic risk index, we used PRSice-2 software that automatically implemented both clumping and thresholding.
Clumping is aimed to select independent SNPs, for this procedure we used the following parameters: 250kb and R^2=0.1. The first parameter represents the linkage disequilibrium window outside of which we believe that the variants are statistically independent. Specifically, in the selected window of SNPs the software will compute the linkage disequilibrium for every pair of SNPs and then if the R^2> 0.1, it will be selected only one SNPs, the one with the most significant p-value.
On the other hand, with thresholding it is possible to select the SNPs that respect specific p-value thresholds, in our case we use the following threshold: 5e-03, 5e-02, 5e-01 and 1. After, clumping and thresholding the software will compute for each individual the polygenic risk score.
Then, we modified the code a little bit to obtain graphical representation.

The barplot in Figure 1 represents the values of R^2 for different p-value thresholds. From this graph we can understand that a p-value threshold equal to 0.0656 gives us the highest R^2, equal to 0,0875. This means that for this p value the correspondent PGS can explain the 8.75% of variability observed in the education phenotype. The value that we find above each bar plot is the Wald test P-value for polygenic score, so we can notice that for the p-value of 0.0656 the Wald test p-value is equal to 2.4•10^{-23}, meaning that the PGS effect is highly significant.

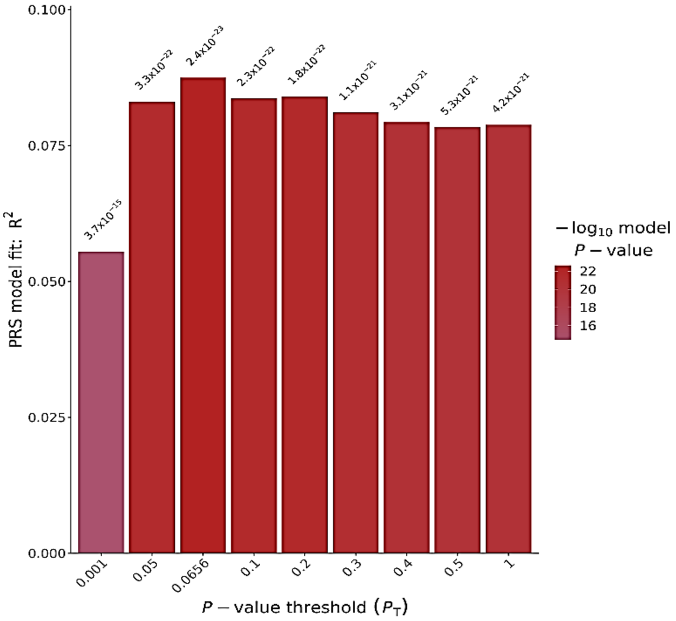

Figure 1. *Barplot for the p-values according to different values of R^2.*

Thanks to the quantile plot in Figure 2 it is possible to understand the probability of having a certain education attainment given a certain position in PGS's quartiles. It is possible to notice that the graph follows an overall increasing trend, meaning that an increase in PGS generally corresponds to an increase in educational attainment.

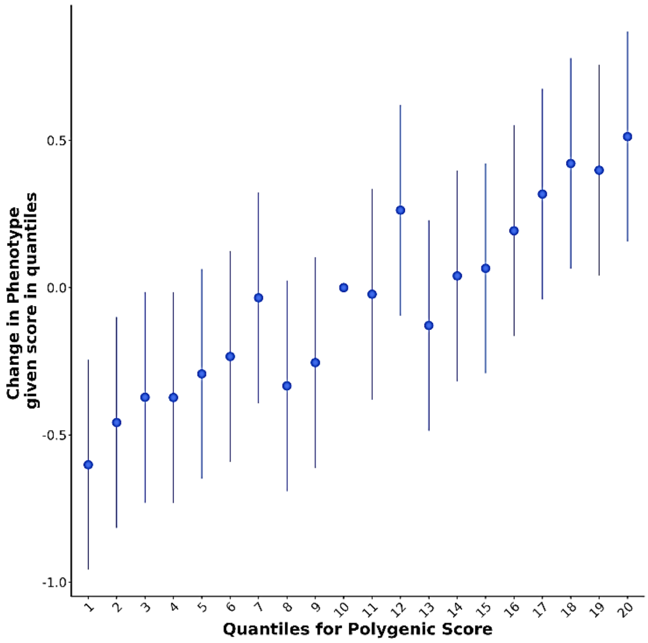

Figure 2. *Quantile plot for the polygenic score, given the change in phenotype.*

In Figure 3 it is possible to notice an example of high-resolution plot, where a green line shows the model fit at the different P-value thresholds considered in the bar plot. This graph confirms the results of Figure 1, in fact we can see that the maximum value for the -log10 transformed p-values, so the most significant PGS, corresponds to the p-value threshold equal to 0.065.

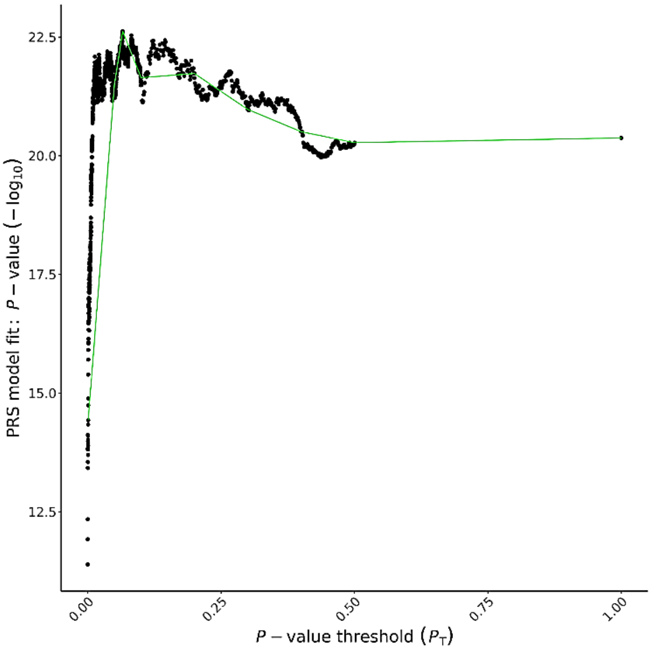

Figure 3. *High resolution plot according to the p-value threshold.*

Part 3. Work with summary statistics.

The Manhattan plot in Figure 4 represents the results of the latest genome-wide association study (GWAS Atlas, 2019) investigating the genetic basis of educational attainment. The plot displays the chromosomes and the genetic positions on the x-axis, while the y- axis represents the -log10 transformed p-values obtained from the statistical analysis for each genetic variant.

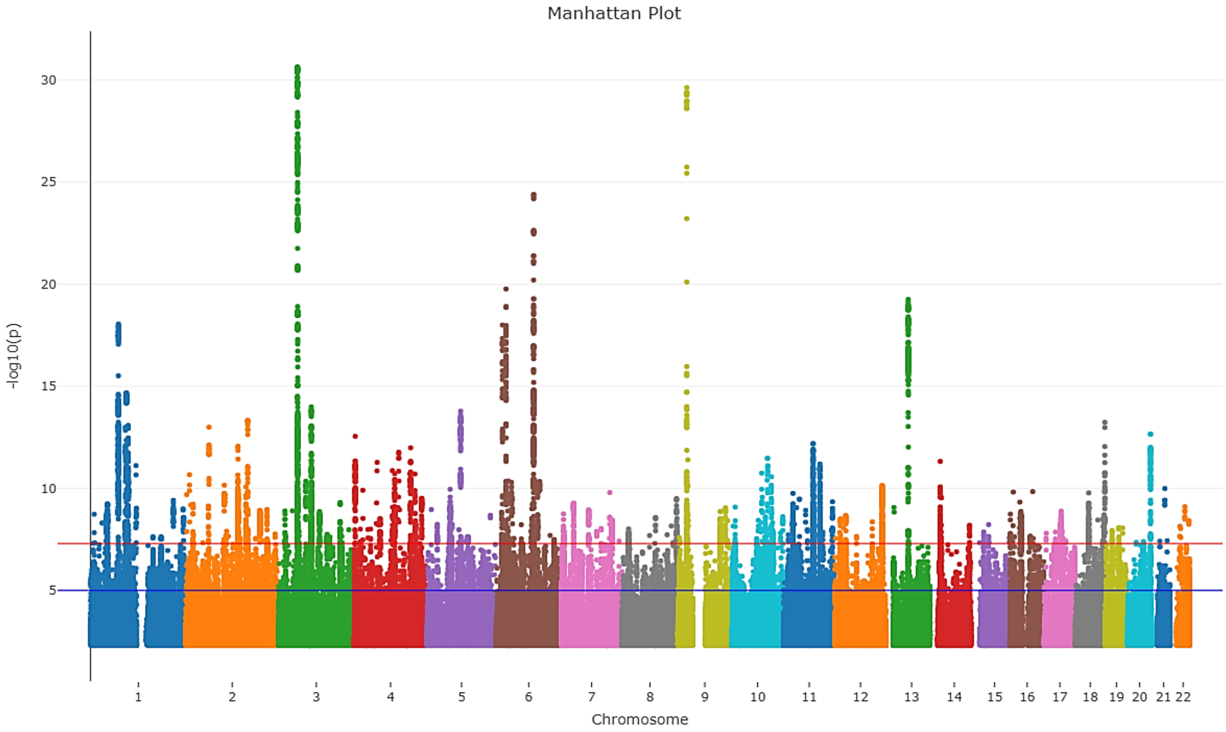

Figure 4. *Manhattan plot using the summary statistics from the latest GWAS on Educational Attainment, 2019.*

The points on the plot are coloured to distinguish different chromosomal regions. We notice a significant peak on chromosome 3, indicating a highly significant association between the genetic variant in that region and the examined trait. This result suggests that a mutation or genetic variant in that specific region could significantly produce an increase in terms of educational attainment. It's also interesting to note some other significant peaks on various chromosomes, such as 1, 6, 9, and 13 even if they are smaller compared to the peak on chromosome 3. These may represent weaker but still significant genetic associations with the disease. By examining the plot, we can distinguish if certain chromosomes show a higher density of significant associations compared to others. In this case, chromosomes 1,2,3 and 4 show a relatively higher number of associations, indicating their potential importance in influencing educational attainment.

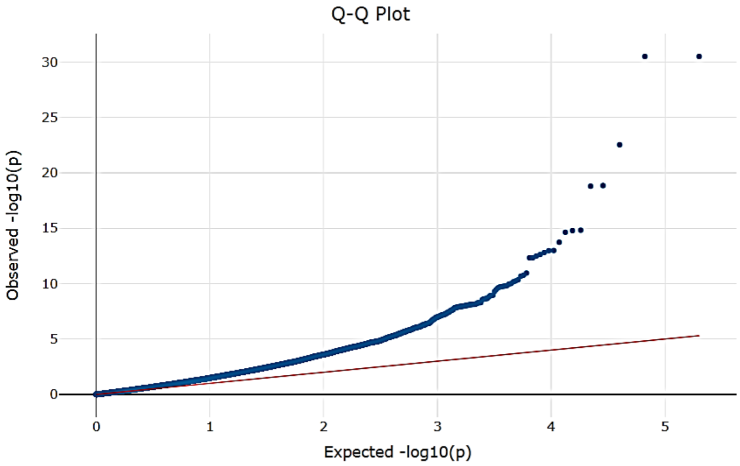

Figure 5. *QQ-plot of the summary statistics.*

With the QQ-plot in Figure 5 we evaluate the distribution of summary statistics, such as the p-value associated with each tested genetic variant. If the p-values were randomly distributed, we would expect the QQ plot to follow a 45-degree line, indicating a good match between the empirical and expected theoretical distributions. However, this graph shows that the empirical distribution of p-values significantly deviates from the theoretical distribution, and this suggests that there may be non-random associations or systematic biases in the data. For example, deviations from the expected distribution could be indicative of population stratification, which occurs when the study sample consists of subpopulations with different allele frequencies; this can lead to false positive or false negative associations. Another factor to consider is the impact of multiple testing: in GWAS many genetic variants are typically tested simultaneously, leading to multiple hypothesis testing. This can increase the chance of false positives, so adjusting for multiple testing, such as using Bonferroni or Holm correction, can help control for this problem. Nevertheless, it is important to take into account that significant deviations from the expected distribution can also suggest the presence of true genetic associations with the trait under study. If the QQ plot shows an higher number of low p-values than what is expected by chance, it may indicate the presence of genetic variants that are actually associated with the phenotype of interest.

Now we examine the genetic correlation between our trait, educational attainment, and other 10 traits that could be of interest:

- Childhood IQ
- Former/Current Smoker
- Autism Spectrum
- Bipolar
- Infant Head Circumference

- Anorexia
- LDL (low-density lipoprotein) cholesterol
- Alzheimer's
- ADHD
- Obesity Class 3

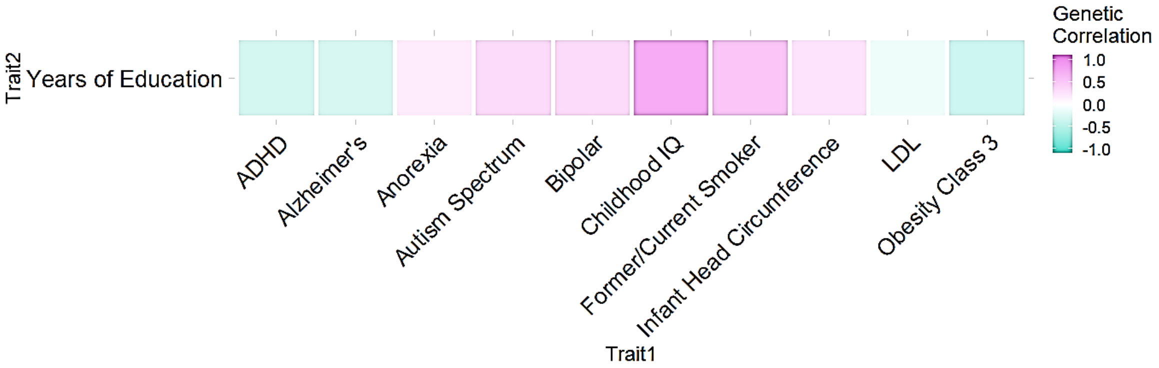

Figure 6. *Heatmap describing genetic correlations.*

Based on the heatmap analysis shown in Figure 6, we observed that childhood IQ shows the strongest positive correlation with our trait, which is coherent with our study. Additionally, we identified other positive correlations with variables such as current/former smoking, bipolar disorder, and autism spectrum. We also noticed weaker positive correlations with anorexia and infant head circumference. On the other hand, we found well-defined negative correlations with severe obesity, Alzheimer's disease, and ADHD, while the negative correlation with LDL (low-density lipoprotein) appears to be weaker compared to the others.

Sociogenomics 2022/23
Final project

José Gabriel Escarraman
Benedetta Felici
Valentina Idda

R and Plink code

Part 1.

```
##1-Create two directories:
mkdir Final_project_data
mkdir Final_project_results
## 2- Create a link in the data directory for plink:
cd $HOME/Final_project_data ln -s $HOME/plink
## 4- Upload the files “1kg_hm3.bed, 1kg_hm3.bim, 1kg_hm3.fam” (put them inside Final_project_data directory)
## 5-Filtering of variants and individuals:
./plink --bfile 1kg_hm3 --maf 0.05 --mind 0.05 --geno 0.05 --hwe 0.00001 --make-bed --out filtered_data
## 6- Upload the file “edu_sim.phen” in the same directory.
## 7- Attach phenotype:
 ./plink --bfile filtered_data --pheno edu_sim.phen --make-bed --out final_data
## 8-Create results.
plink --bfile final_data --assoc --out $HOME/ Final_project_results /gwas_results
## 9-Selecting the 10 variants with smaller P-value:
sort -g -k9,9 gwas_results.qassoc | head -n 11 > filtered_gwasresults.txt
```

Part 2.

```
sudo apt-get install r-base r-base-dev
R
install.packages(c("ggplot2", "data.table"))

DATA="/home/benedetta01felici/Sociogenomics/Data"
RESULTS="/home/benedetta01felici/Sociogenomics/Results"
Rscript PRSice.R --dir . \
    --prsice ./PRSice_linux \
    --base ${DATA}/Edu.txt\
    --target ${DATA}/filtered_data \
    --thread 1 \
    --snp MarkerName \
    --A1 A1 \
    --A2 A2 \
    --stat Beta \
    --pvalue Pval \
    --bar-levels 5e-03,5e-02,5e-01,1 \
    --fastscore \
    --all-score \
    --no-regress \
    --binary-target F \
    --out  ${RESULTS}/Eduscore2_thresholds

Rscript PRSice.R --dir . \
    --prsice ./PRSice_linux\
    --base ${DATA}/Edu.txt \
    --target ${DATA}/filtered_data \
    --thread 1 \
    --snp MarkerName \
    --A1 A1 \
    --A2 A2 \
    --stat Beta \
    --pvalue Pval \
    --pheno-file ${DATA}/edu_sim_stan.phen \
    --interval 0.00005 \
    --lower 0.0001 \
    --quantile 20 \
    --all-score \
    --binary-target F \
    --out ${RESULTS}/EstanDUscore_graphics
```

Part 3.

```
EA = read.table(file.choose(),header=T)
EA_sub = subset(EA, P < 0.005)
library(manhattanly)
manhattanly(EA_sub, snp = "SNP" , bp="BP", p="P", chr="CHR", col=2)
qqly(EA[sample(dim(EA)[1],100000),], snp = "SNP" , bp="BP", p="P", chr="CHR", col=2)

qqly(EA[sample(dim(EA)[1],100000),], snp = "SNP" , bp="BP", p="P", chr="CHR", col=2)


RG<-read.table(file.choose(),fill =T, sep="\t", header=T, quote="")
install.packages("ggplot2")
library(ggplot2)

ggplot(data = RG, aes(Trait1, Trait2, fill = rg))+
  geom_tile(color = "white", lwd=2)+
  scale_fill_gradient2(low = "turquoise", high = "orange", mid =
                         "white",  midpoint = 0, limit =
                         c(-1.1,1.1), space = "Lab",
                       name="Genetic\nCorrelation") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 8, hjust = 1))+
  coord_fixed()
```