KU LEUVEN

Evolutionary and Quantitative Genetics – I0D53A
Quantitative Genetics – Assignment 1
GWAS - QC

Due date of the assignment: 18.12.2023 at noon (12PM).

Your last and first name *Plewa Daria*
Your student number r0976669

---

*Please compile your answers into this document. Try to be as **specific and concise as possible** in your answers. If you feel you would like to add additional information (screen shots, code…) related to specific answers, you can do so (no obligation!), but please then do this as **supplementary information at the end of the document**, and referring to it in the respective answer.*

Part 1: Introduction to the PLINK association analysis tool
##################################

1. How many SNPs are there included on chromosome 15? Use PLINK (!) to answer this question, and also specify the full PLINK command you used.

plink --bfile GWA_raw --chr 15 --make-bed --out test

2418

2. What is the minor allele frequency of rs13402123 in controls? And in cases? Use PLINK (!) to answer this question, and also specify the full PLINK command(s) you used.

plink --bfile GWA_raw --snp rs13402123 --filter-controls --freq --out test

Minor Allele Frequency for controls: 0.0186

CHR	SNP	A1	A2	MAF	NCHROBS
2	rs13402123	A	G	0.01862	1772

Screen of Plink result. The minor allele frequency of rs13402123 in controls

plink --bfile GWA_raw --snp rs13402123 --filter-cases --freq --out test

Minor Allele Frequency for cases: 0.01482

CHR	SNP	A1	A2	MAF	NCHROBS
2	rs13402123	A	G	0.01482	3306

Screen of Plink result. The minor allele frequency of rs13402123 in cases

---

Please upload your answer in a **pdf format** under Toledo > Assignments

KU LEUVEN

Part 2: QC for GWAS
############

3. Give the full PLINK command used to check for discordant sex information. How many individuals did you exclude?

plink --bfile GWA_raw -check-sex --out test
grep -c "PROBLEM" test.sexcheck
264

4. How many individuals have a genotyping failure rate >0.05? Only (!!!) use PLINK to answer this question, and also specify the full PLINK command you used.

```
plink --bfile GWA_raw --mind 0.05 --make-bed --out test
Error: All people removed due to missing genotype data (–mind).
All individuals have a genotyping failure rate below 0.05
```

What threshold would you choose for individual missingness? How many individuals are removed at this threshold?

```
plink --bfile GWA_raw --mind 0.1 --make-bed --out test

175135 variants and 2196 people pass filters and QC.
Among remaining phenotypes, 1330 are cases and 864 are controls.  (2
phenotypes are missing.)
```

5. What is the genotyping failure rate for the individual with IID = 679? Also specify how you obtained this value.

```
plink --bfile GWA_raw --missing --out test
grep "679" test.imiss
```
Genotyping failure equals 0.08148 for IID = 679

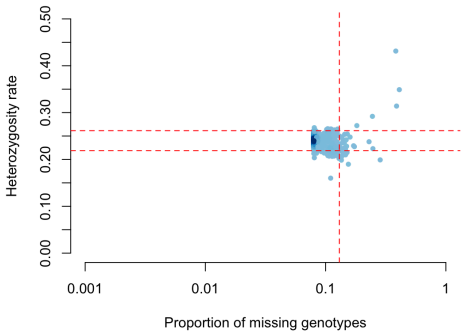
```
679   679           N    14270   175135   0.08148
```

Screen of Plink file genotyping failure rate for individual with IID = 679

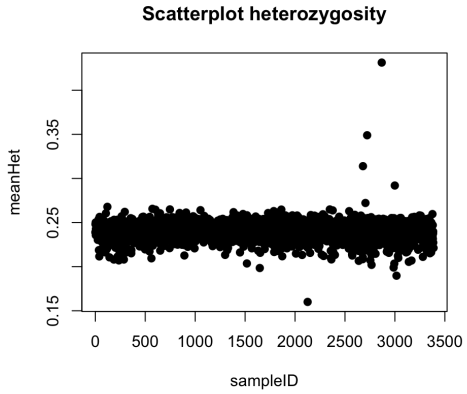
Please upload your answer in a **pdf format** under Toledo > Assignments

6. How many individuals did you remove because of outlying heterozygosity? How did you decide on the threshold used?

I removed 4 individuals whose heterozygosity exceeded 0.29, and 4 individuals whose heterozygosity hadn't reached the level of 0.2. Together I removed 8 individuals. I chose the threshold based on the repetitiveness of the results of heterozygosity. Most of the results landed below 0.29 and 0.2 thresholds with a rather constant level of heterozygosity.



Plot of missingnes vs heterozygosity



Manhattan plot of heterozygosity

7. What is the lowest (excluding the very low values around 0) and highest pi_hat you observed between the related samples? What do you conclude from these values about the relatedness between the individuals? How many individuals did you exclude because of relatedness?

The lowest obtained pi_hat value equals 0.2135, while the highest is 1. I excluded 204 specimens because of the threshold 0.2. The higher values represented the higher relatedness between the individuals. Values exceeding 0.2 typically represent potential duplicate samples or relatedness between individuals.

8. Give the full PLINK command you used to calculate the MAF for all included markers.

```
plink --bfile GWA_sampleqc --recode --freq --out GWA_sampleqc
```

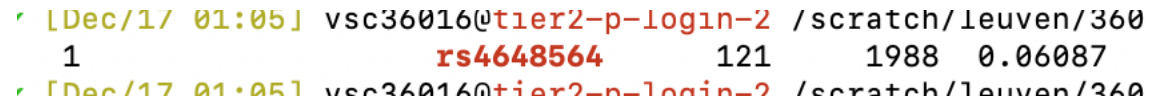
9. How many markers have a missing genotype rate >0.05? Use PLINK to answer this question, and also specify the full PLINK command you used. What threshold would you choose for SNP missingness? How many variants are removed at this threshold?

plink --bfile GWA_sampleqc --geno 0.05 --make-bed --out test
128143 - 0.01
17056 - 0.05
14244 - 0.1
The threshold values often obtain values between 0.01 and 0.1. I would choose the 0.05 threshold for that particular result because 0.01 can exclude too many individuals and 0.1 not enough.

10. What is the missingness rate for SNP with ID = rs4648564? Only (!!!) use PLINK to answer this question, and also specify the full PLINK command you used.

--bfile GWA_sampleqc --missing --snp rs4648564 --out GWA_sampleqc

For rs4648564 the missingness rate equals 0.06087

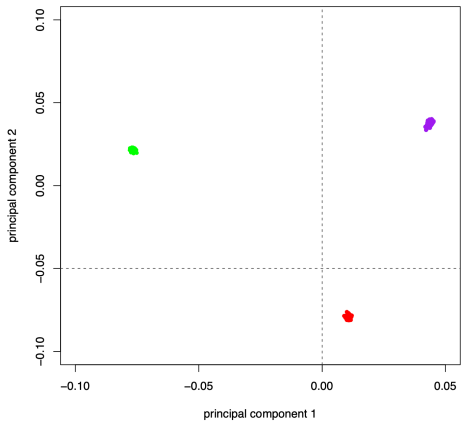


Screen of plik result: missingness rate for SNP with ID = rs4648564

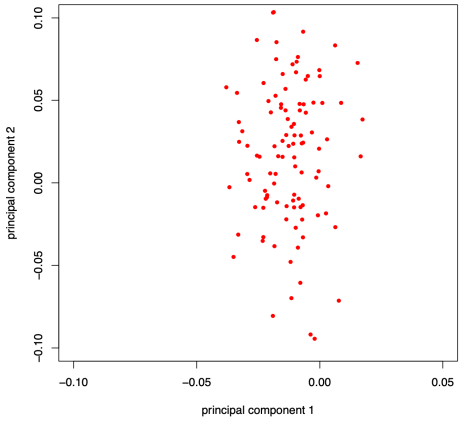
11. Give the full PLINK command you used to remove all SNPs failing QC, indicating your thresholds chosen for MAF, SNP call rate, and HWE, and also removing SNPs with differential missingness between cases and controls. How many SNPs eventually pass QC?

plink --bfile GWA_sampleqc_1 --maf 0.01 --hwe 0.001 --geno 0.05 —exclude fail-diffmiss-qc.txt —make_bed --out test_final
127111 SNPS finally passed QC

12. What cut-offs did you choose for PC1 and PC2 to retain only individuals of European ancestry? How many individuals did you remove because of non-european ethnicity? Also copy in your PCA plot (scatter plot) plotting the first two PCs of your dataset merged with the HapMap dataset; as well as the PCA plot of your final dataset (without HapMap, and with the ethnic outliers excluded).



PCA of full dataset with HapMap



PCA of data set with European ancestry without HapMap

For the full dataset, I excluded individuals exceeding 0 (on x-axis, PC1) and -0.05 values (on y-axis, PC2) The red dots represent the Europan dataset, purple - Han Chinese and Japanese and green individuals of Yoruba.

For PCA of data with European ancestry, I excluded 283 individuals.

Please upload your answer in a **pdf format** under Toledo > Assignments