Genome Wide Association Quality Control

James O'Reilly
Student Number: r0773125

1 Introduction to PLINK

Question One

How many SNPs are there included on chromosome 16? I first used input filtering to access only the data for SNPs from chromosome 16, and output this to a file. I then view the chr16.log to view the number of SNPs on chromosome 16.

```
./plink --bfile gwa_raw --chr 16 --snps-only --freq --out chr16
cat chr16.log
```

There are 6435 SNPs on chromosome 16.

Question Two

What is the minor allele frequency of rs2066845 in controls? And in cases? I first created a file with the frequencies for cases and controls for the given allele. I then checked this file for the case and control MAF.

```
./plink.exe --bfile plink --freq case-control --snp rs2066845
```

The minor allele frequency of this allele in controls and in cases was 0.02483 and 0.05562, respectively.

2 Quality Control

Question Three

Give the full PLINK command used to check for discordant sex information. I first create a `plink.sexcheck` file and then use `grep` and wc to get the number of individuals with discordant sex information. Their FID and IID was then listed in a new file `fail-sexcheck-qc.txt`.

```
./plink --bfile gwa_raw --check-sex
grep PROBLEM plink.sexcheck | wc -l
grep PROBLEM plink.sexcheck | awk -v OFS='\t'  '{print $1,$2}'  > fail-sexcheck-qc.txt
```

This gives 264 individuals with discordant sex information.

Question Four

How many individuals have a genotyping failure rate > 0.05*? Use PLINK to answer this question, and also specify the full PLINK command you used.* First I create a map file. and then use the`--mind` command to filter out all individuals with missing call rates exceeding the provided value.

```
./plink --bfile gwa_raw --recode
./plink --bfile gwa_raw --mind 0.05 --make-bed --out gwa_raw_filtered
```

All individuals were removed as they have a genotyping failure rate > 0.05.

Could you also have found this number not using a specific PLINK command? If so, how? Yes. You can first create a `.imiss` file and then check the number of of individuals with an `FMISS` value > 0.5.

```
./plink --bfile gwa_raw --missing --out gwa_raw
sed '1d' gwa_raw.imiss
awk '$6>0.05' gwa_raw.imiss
```

This gives 2543 individuals with a genotyping failure rate greater than 0.05.

Question Five

Add the scatter plot of sample call rate against heterozygosity to this document, and provide a legend to the figure, explaining what can be seen, and how you interpret the plot.

The thresholds for heterozygosity rate and missing genotype proportion were chosen using Figures 4 and 5, which are presented in the appendix.

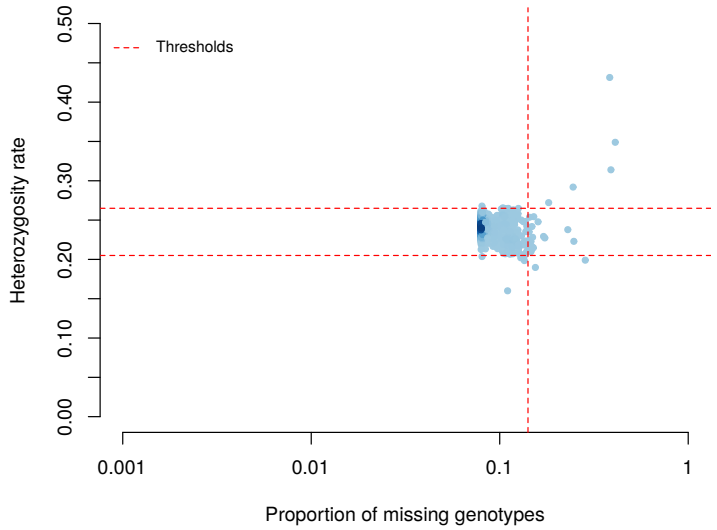

Figure 1: Scatter plot of missingness vs heterozygosity. The upper and lower thresholds for heterozygosity are at 0.265 and 0.205, respectively. The missingness threshold is at 0.15.

The majority of individuals which failed, did so due to a high proportion of missing genotypes. There were also a large number of individuals with high heterozygosity rates. The FID and IID of individuals failing call rate or heterozygosity rate were added to a text file using the R script given below.

```
library(dplyr)
names(het)
het_failed <- het %>% filter(meanHet < 0.205 | meanHet > 0.265) %>% dplyr::select(FID,
    IID)
imiss_failed <- imiss %>% filter(F_MISS > 0.15)  %>% dplyr::select(FID, IID)

all_failed <- rbind(het_failed, imiss_failed)
all_failed <- unique(all_failed)

write.table(as.matrix(all_failed), file = 'fail-miss_het-qc.txt', sep="\t", col.names = F
    , row.names = F)
```

The number of individuals with failing rates can then be found using the `wc -l` command. Using these thresholds, 21 individuals failed.

Question Six

What is the lowest and highest pi_hat you observed between the related samples? The bash commands given below return the highest and lowest `pi_hat` values of 1 and 0, respectively.

```
cat GWA_raw.genome| awk  -v OFS='\t'  '{print $10}'| sort | head
cat GWA_raw.genome| awk  -v OFS='\t'  '{print $10}'| sort -r | head
```

What do you conclude from these values about their relatedness? The `pi_hat` value represents the proportion of identity-by-descent (IBD), for each pair of individuals. The underlying P(IBD=0/1/2) estimator sometimes yields numbers outside the range [0,1], however numbers outside this range are clipped by default. Ultimately, a `pi_hat` value of 0 means the samples are completely unrelated, while a value of 1 indicates they are very much related. However, the clipping mentioned earlier means that a value of 1 or 0 for a given pair, doesn not necessarily represent the same level of relatedness as in another pair of individuals.

Question Seven

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

```
./plink --bfile gwa_sampleqc --freq --out gwa_sampleqc
```

The distribution of minor allele frequencies, along with the cut-off threshold of 0.01 is given in Figure 2 below.

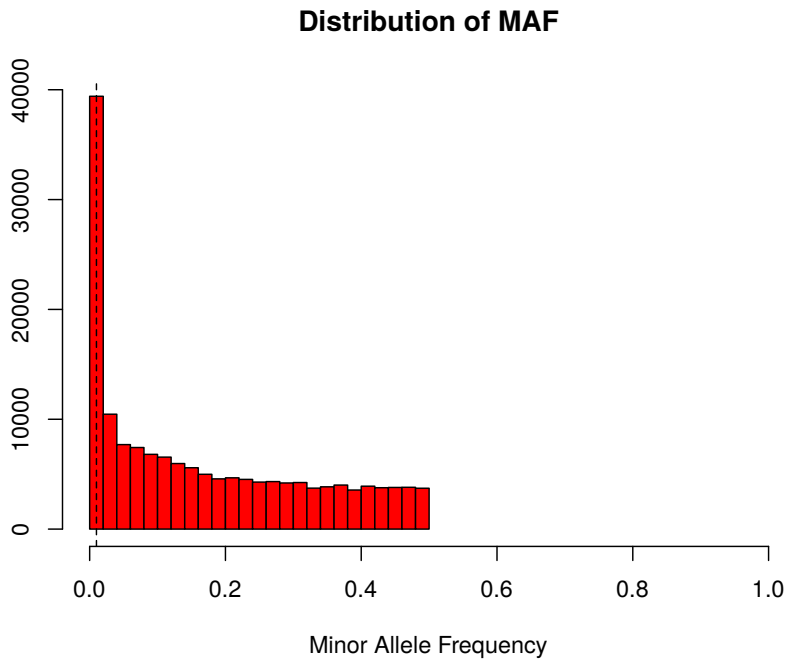

Figure 2: Histogram showing the distribution of minor allele frequencies. The cut-off threshold of 0.01 is shown as a dotted line.

Removing SNPs with MAFs below the given default threshold of 0.01:

```
./plink --bfile gwa_sampleqc --maf --make-bed --out gwa_sampleqc_filteredMAF
```

Question Eight

*Why do you think did we re-run the –missing command, and not simply used the *.lmiss output file from step ii of the sample QC part?*

In step (ii) we removed individuals with a high missingness rate. In this case, we must re-run the `missing` command as there could still be poor quality SNPs present which would not have been removed previously.

Question Nine

How many markers have a missing genotype rate greater than 0.05. Use PLINK to answer this question, and also specify the full PLINK command you used.

```
./plink --bfile gwa_sampleqc_filteredMAF --geno 0.05 --make-bed --out
    gwa_sampleqc_filteredMAF
```

15548 variants were removed due to missing genotype data. the histogram for missing genotype rate is given in Figure 6

Could you also have found this number not using a specific PLINK command? If so, how? Yes. The `.lmiss` file is a variant-based missing data report. We can return the number of variants with a missing genotype rate greater than 0.05 with the following command:

```
sed '1d' gwa_sampleqc_filteredMAF
awk '$5 >0.05' gwa_sampleqc_filteredMAF.lmiss | wc -l
```

Which again shows that 15548 variants have a missing genotype rate greater than 0.05.

Question Ten

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.

First, we must make a text file with all SNPs with a significantly different missing data rate between cases and controls. This is done using the command below:

```
awk '$5 < 0.00001' gwa_sampleqc_filteredMAF.missing | awk -v OFS='\t' '{print $2}' > fail
    -diffmiss-qc.txt
```

Then we can remove all SNPs according the the chosen thresholds which we discussed earlier. The final command is given by:

```
./plink --bfile gwa_sampleqc --maf --geno 0.05  --hwe 0.00001 --makebed --exclude fail-
    diffmiss --out gwa_sampleqc_final
```

This leaves 1326 cases and 687 controls. The plot showing the distribution and cut-off threshold for SNP call rate is given in the appendix with other figures.

Suggestions for improving the QC pipeline

In search of suggestions for improving the pipeline, I read the recommendations for quality control procedures in genome wide association studies given by Turner et. al.[3] They provided a flow chart representing the key elements of the QC analysis (see Figure 3). A similar set of key steps is outlined in Marees et al.[1].

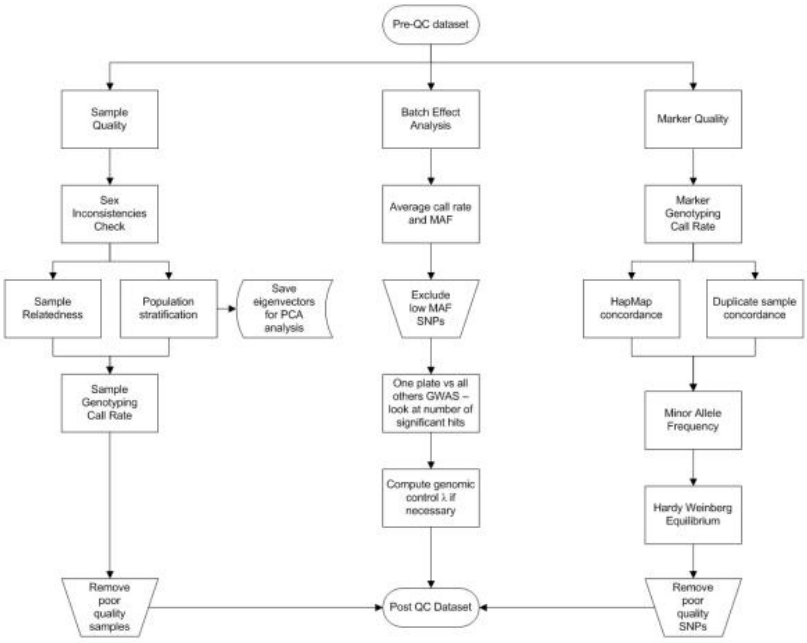

Figure 3: Key elements of quality control for genome wide association studies.

Looking at this figure, there are clearly no major components of the QC analysis which are missing here, and so any suggested improvements are likely micro adjustments to individual QC procedures. With this in mind, one element of the analysis which felt questionable to me was the method by which I chose cut-off thresholds for MAF, missingness, and heterozygosity. Simply looking at the distributions and then choosing a threshold doesn't feel like a rigorous way of removing potentially problematic variants or samples. Perhaps it would instead be better to use some rigorous statistical measures such as standard deviation, variance, or z-scores to determine where the cut-off threshold lies.

Marees et al. also suggests correcting for population stratification as part of the QC analysis (if it is suitable for the given dataset):

> *An important source of systematic bias in GWAS is population stratification. It has been shown that even subtle degrees of population stratification within a single ethnic population can exist. Therefore, testing and controlling for the presence of population stratification is an essential QC step.*

To correct for population stratification, one can use a multidimensional scaling (MDS) approach. Price et al. authored the canonical paper which outlines different approaches to correct for population stratification.[2]

3 Research Paper

Imagine you are writing a research paper about the GWA study you are doing. Write down the methods part explaining your dataset, and the QC steps you have done. Make sure it includes the necessary information to understand the dataset. For example – but not necessarily limited to – how many samples are there in total, and broken down by category; how many samples were excluded and why; how many variants were genotyped; how many variants were excluded and why... . You can include any table/figure that you feel is relevant. When doing so, make sure to provide a legend with interpretation of the figure/results.

3.1 The Data

The SNP data are which are input for quality control are in three separate files:

- A `.bed` which contains the individual and family IDs, as well as the genotypes
- A `.bim` containing information on genetic markers
- A `fam` with information on individuals

Before quality control, the data contains information about 175,135 SNPs across 2543 individuals. Of the 2543 individuals, 886 have the phenotype. There are 1118 males, 1422 females, and 3 individuals with unknown sex.

3.2 Quality Control

Both the samples and phenotypes are filtered according to a number of criteria, given below:

- Sample quality control
  - Call rate
  - Heterozygosity
  - Discordant sex
  - Identification of duplicated or related individuals
- SNP quality control
  - Call rate
  - Minor allele frequency (MAF)
  - Hardy-Weinberg equilibrium (HWE)

Individuals with discordant sex information were first removed, excluding 264 samples from the dataset. The thresholds for heterozygosity rate and missing genotype proportion were chosen using Figures 4 and 5, excluding another 21 samples from the dataset (see Figure 1 for the distribution and thresholds). Lastly, the pairwise IBD was calculated for each pair of individuals using a pruned dataset. In each pair of individuals with an IBD less than 0.2, the individual with the lowest call rate was removed, leaving a total of 2013 samples which passed quality control.

Next, SNPs of poor quality were removed. First, all X-chromosome SNPs were removed, as they are not used in the association study that follows. SNPs with a minor allele frequency less than 0.01 were removed, along with SNPs with a HWE less than 10^{-5}. SNPs with a call rate less than 0.05 were also removed from the data. This threshold was chosen using Figure 6. Lastly, all SNPs with significantly differential missingness between cases and controls were removed. In total, 1595 SNPs were removed from the dataset, leaving 173,540 SNPs.

Appendix

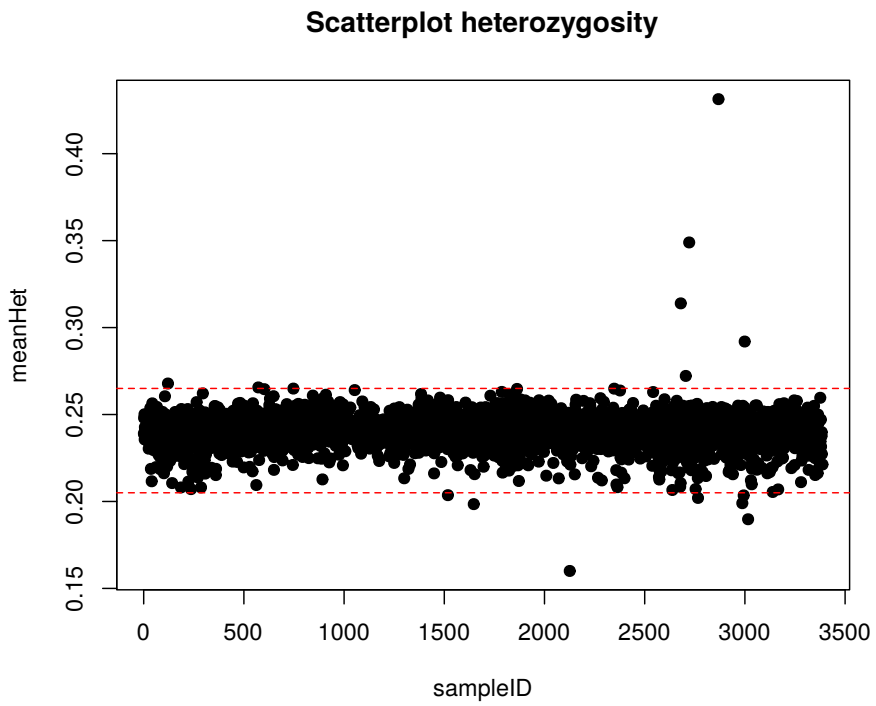

Figure 4: Scatter plot for heterogeneity of individuals, along with the given cut-off thresholds.

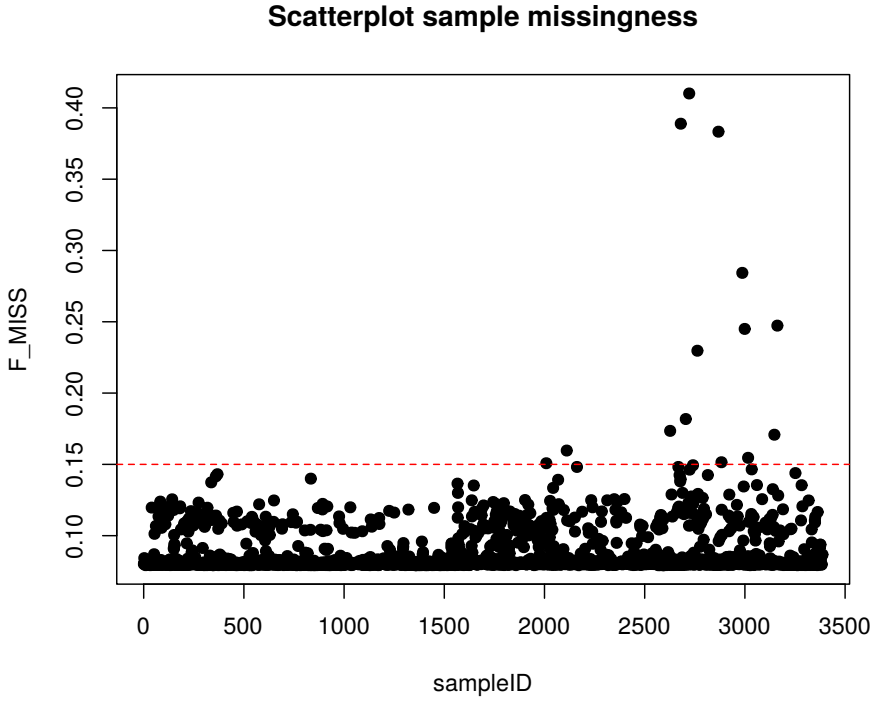

Figure 5: Scatter plot for missingness of indivudals, along with the given cut-off threshold.

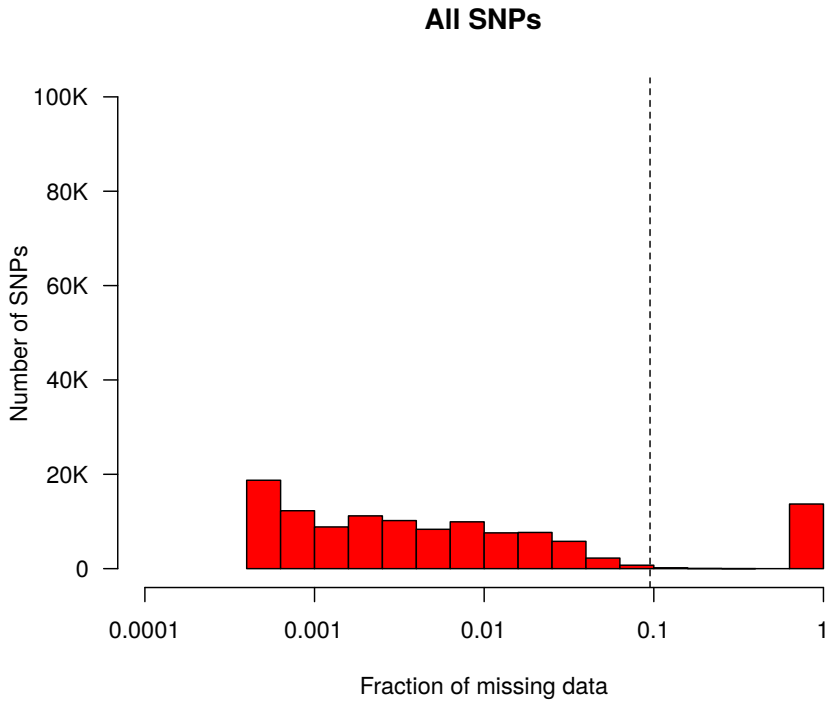

Figure 6: Histogram showing the distribution of SNP call rate, along with the given cut-off threshold.

References

[1] Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., and Derks, E. M. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research 27*, 2 (2018), e1608.

[2] Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics 11*, 7 (2010), 459–463.

[3] Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., de Andrade, M., Doheny, K. F., Haines, J. L., Hayes, G., et al. Quality control procedures for genome-wide association studies. *Current protocols in human genetics 68*, 1 (2011), 1–19.