Genome-wide Association Analysis - Data Quality Control

Introduction

In this exercise, you will learn how to perform data quality control (QC) by removing markers and samples that fail QC quality control criteria. You will also examine your samples for individuals that are related to each other and/or are duplicate samples. Each sample will also be tested for excess homozygosity and heterozygosity of genotype data. Each SNP will be tested for deviations from Hardy-Weinberg Equilibrium. These exercises will be carried out using PLINK 1.9.

1. Using PLINK

PLINK can upload data in different formats please see the PLINK documentation (https://www.cog-genomics.org/plink/1.9/input) for additional details. The data for this exercise is in PLINK/LINKAGE file format. There are two files: a pedfile (`GWAS.ped`) and a map file (`GWAS.map`).Please note the commands must be given in the directory where the data reside. For this exercise the directory which contains the files for the exercise is `data` , which you can navigate on the file directory on the left sidebar.

Please examine these files and the PLINK documentation. To preview the files run the commands below.

In [ ]:
```
# Set working directory
cd ./data
```

In [ ]:
```
plink --file GWAS
```

Note, that PLINK outputs a file called plink.log that contains the same output which you see on the screen. To see all options, type `plink --help` for more information. Determine how many samples there are in your data set and fill in Oval 1 of the flowchart below.

2. Data Quality Control

a. Removing Samples and SNPs with Missing Genotypes.

You will exclude samples that are missing more than 10% of their genotype calls. These samples are likely to have been generated using low quality DNA and can also have

higher than average genotyping error rates.

In [ ]:
```
plink --file GWAS --mind 0.10 --recode --out GWAS_clean_mind
```

Examine `GWAS_clean_mind.log` to see how many samples are excluded based on this criterion and fill in Box 1.

We will now create two versions of your dataset, one with SNPs with a minor allele frequencies (MAFs) >5% and the other with SNPs with a MAFs <5%.

You will now remove SNPs with MAFs>5% that are missing >5% of their genotypes and then remove SNPs with MAFs<5% that are missing >1% of their genotypes. SNPs which are missing genotypes can have higher error rates than those SNP markers without missing data.

In [ ]:
```
plink --file GWAS_clean_mind --maf 0.05 --recode --out MAF_greater_5
plink --file GWAS_clean_mind --exclude MAF_greater_5.map --recode --out MAF_
plink --file MAF_greater_5 --geno 0.05 --recode --out MAF_greater_5_clean
```

Fill in Box 2a.

In [ ]:
```
plink --file MAF_less_5 --geno 0.01 --recode --out MAF_less_5_clean
```

Fill in Box 2b. Merge the two files.

In [ ]:
```
plink --file MAF_greater_5_clean --merge MAF_less_5_clean.ped MAF_less_5_cle
```

A more stringent criterion for missing data is used, samples missing >3% of their genotypes are removed.

In [ ]:
```
plink --file GWAS_MAF_clean --mind 0.03 --recode --out GWAS_clean2
```

Fill in Box 3.

b. Checking Sex

Error of the reported sex of an individual can occur. Information from the SNP genotypes can be used to verify the sex of individuals, by examining homozygosity (F) on the X chromosome for every individual. F is expected to be <0.2 in females and >0.8 in males. To check sex run

In [ ]:
```
plink --file GWAS_clean2 --check-sex --out GWAS_sex_checking
```

Examine the `GWAS_sex_checking.sexcheck` file at the command line and determine if there are individuals whose recorded sex is inconsistent with genetic sex.

In [ ]:
```
awk 'NR==1 || $5=="PROBLEM"' GWAS_sex_checking.sexcheck
```

NA20530 and NA20506 were coded as a female (2) and from the genotypes appear to be males (1). In addition, 3 individuals (NA20766, NA20771 and NA20757) do not have enough information to determine if they are males or females and PLINK reports sex = 0 for the genotyped sex.Fill in the table below:

Table 1: Sex check

FID	IID	PEDSEX	SNPSEX	STATUS	F
NA20506	NA20506				
NA20530	NA20530				
NA20766	NA20766				
NA20771	NA20771				
NA20757	NA20757				

Reasons for these kinds of discrepancies, include the records are incorrect, incorrect data entry, sample swap, unreported Turner or Klinefelter syndromes. Additionally, if a sufficient number of SNPs have not been genotyped on the X chromosome it can be difficult to accurately predict the sex of an individual. In this dataset, there are only 194 X chromosomal SNPs. If you cannot validate the sex of the individual they should be removed. For this exercise, we are going to assume that when the sex was checked, we found it was incorrectly recorded (i.e. these samples were male). Therefore, this error could simply be corrected.

c. Duplicate Samples

The following PLINK command can be used to check for duplicate samples

In [ ]:
```
plink --file GWAS_clean2 --genome --out duplicates
```

Examine the `duplicates.genome` file at the command line:

In [ ]:
```
head duplicates.genome
```

We are interested in the Pi-Hat (the estimated proportion IBD sharing) value. You may notice that there is more than one duplicate (Pi-Hat=~1). Also, examine the output for pairs of individuals with high Pi-Hat values which can indicate they are related. The amount of allele sharing [Z(0), Z(1) and Z(2)] across all SNPs provides information on the type of relative pair.

In [ ]:
```
awk 'NR==1 || $10>0.4' duplicates.genome
```

Table 2: Duplicate and Related Individuals

FID1	IID1	FID2	IID2	Z(0)	Z(0)	Z(0)	PI_HAT

FID1 - Family ID for 1st individual; ID1 - Individual ID for 1st individual; FID2 - Family ID for 2nd individual; ID2 - Individual ID for 2nd individual; Z(0) - P(IBD=0); Z(1) - P(IBD+1); Z(2) - P(IBD=2); PI_HAT - P(IBD=2)+0.5*P(IBD=1) (proportion IBD)

Note: Pi-hat can be inflated and individuals appear to be related to each other if you have samples from different populations. This explains why we observe pairs of individuals with Pi-hat >0.05 since three distinct populations were analyzed. Additionally, this phenomenon can be observed if a subset(s) of samples have higher genotyping/sequencing error rates, which creates two or more "populations" and the individuals within these "populations" incorrectly appear to be related.

Using the command below, observe how many sample pairs have pi-hat >0.05:

In [ ]:
```
awk 'NR==1 {print $1, $2, $3, $4, $10} $10>0.05 {print $1, $2, $3, $4, $10}'
```

Create the following txt file:

```
1344 NA12057
1444 NA12739
M033 NA19774
```

name it `IBS_excluded.txt` and save it to the folder with your PLINK data. Give the command:

In [ ]:
```
cat > IBS_excluded.txt << 'EOF'
1344 NA12057
1444 NA12739
M033 NA19774
EOF
```

In [ ]:
```
plink --file GWAS_clean2 --remove IBS_excluded.txt --recode --out GWAS_clear
```

Fill in Box 4 and Oval 3.

As part of QC usually the data is examined for outliers by plotting the first and second principal or multidimensional scaling (MDS) components. Using a subset of markers that have been pruned to remove LD (e.g., r2<0.1). Principal components analysis (PCA) and MDS will be performed in the second part of the exercise to detect outliers and control for populations substructure. Outlier can be due to study subjects coming from different populations e.g. European- and African-Americans or batch effects. If it is suspected that outliers are due to study subjects having been sampled from different populations than data from HapMap can be included to elucidate population membership, e.g. for a study of European-Americans if African-American study subjects are included they would cluster between the European and African HapMap samples. If you perform this type of analysis you should remove the HapMap samples and re-estimate the MDS or PC components before adjusting for population substructure or stratification. For this

exercise data is used from HapMap Phase III which consists of CEU (Europeans from Utah), MEX (Mexicans from Los Angeles) and TSI (Tuscans from Italy). Three clusters can be observed that consist of the three data sets but no extreme outliers are observed. This data set is being used for demonstration purposes. Different populations should be analyzed separately and the results can be combined using meta-analysis. In part two of this exercise MDS and PC components will be constructed and analyzed.

d. Hardy-Weinberg Equilibrium (HWE):

To test for HWE we will test separately in each ancestry group and by case-control status. Therefore, we will need to use information on ancestry and cases-control status. Please note that this should be tested in the 3 different populations separately (CEU, MEX, TSI), but due to the small sample sizes, we tested it in the 3 populations together for example purposes. It should also be noted if the sample sizes are small it is difficult to detect a deviation from HWE.

In [ ]:
```
plink --file GWAS_clean3 --pheno pheno.txt --pheno-name Aff --hardy
```

Examine the file `plink.hwe` at the command line and look for SNPs with p-values of 10^{-7} or smaller.

In [ ]:
```
awk 'NR==1 || $9<0.0000009' plink.hwe
```

Using a criterion of p $<10^{-7}$ to reject the null hypothesis of HWE, how many SNPs fail HWE in the cases? Fill out Oval 5 and Box 4. Using the same criteria, how many SNPs fail HWE in the controls? Complete Table 3 with this information.

Table 3: Hardy-Weinberg Equilibrium

Cases			Controls		
SNP	**Pvalue**	**Population(s)**	**SNP**	**Population(s)**	**Pvalue**

Create a text file called `HWE_out.txt` with the following SNP in it:

```
rs2968487
```

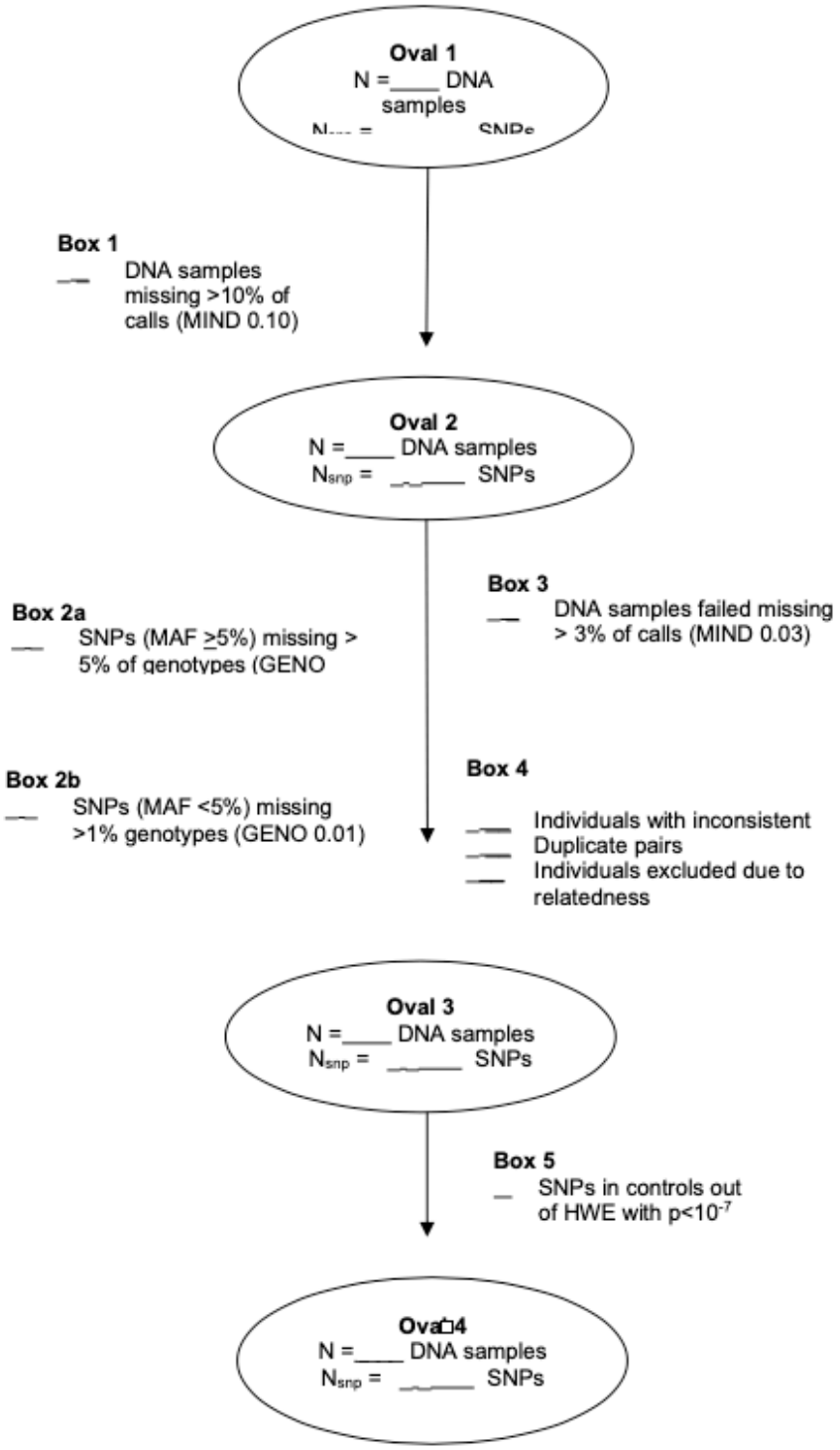
and type the following command:

In [ ]:
```
cat > HWE_out.txt << 'EOF'
rs2968487
EOF
```

In [ ]:
```
plink --file GWAS_clean3 --exclude HWE_out.txt --recode --out GWAS_clean4
```

There are a number of SNPs with HWE p-values in the range of 10^{-5} to 10^{-6} in the controls. Based on above criterion they will not be excluded however, if they reach

genome-wide significance during association testing they SNPs should be further investigated to ensure there is no genotyping error. You can now fill in Box 5 and Oval 4.



3. Questions

Question 1: Why do you expect the homozygosity rate to be higher on the X chromosome in males than females?

Question 2: How many duplicate pairs do your find (hint: Pi-Hat = ~1)? Do pairs with a Pi-Hat = ~1 have to be duplicate samples? What is another explanation? What proportion would you expect a parent/ child to share IBD? Can you find any such relationship?