Quality Control of the ROS/MAP Genotype Dataset

Initial Quality Control of the Genotype data

n1,686: contains all raw output and filtered SNP data following TOPmed imputation for n=1,686 ROS/MAP subjects genotyped on the Illumina HumanOmniExpress Chip

n381: contains all raw output and filtered SNP data following TOPmed imputation for n=381 ROS/MAP subjects genotyped on the Affymetrix 6.0 Chip merged: contains merged filesets combining n=1686 and n=381 subjects for consolidated analyses

SNP QC

We are using imputed data of the ROS/MAP cohort study. The genotype data was filtered with the following criteria:

- imputation R2 < 0.8
- minor allele frequency < 0.0025
- All triallelic variants (removed)

Now, we will apply the extra quality control on the dataset:

- Removing SNPs with MAF < 0.01
- Removing SNPs with low P-value from the HWE's Fisher's test of 1e-06
- Excluding SNPs with missingness of 0.01
- Excluding individuals with high rate of missing genotype at 0.01

```bash
#!/bin/bash --login
#SBATCH --time=01:00:00
#SBATCH --nodes=4
#SBATCH --ntasks=4
#SBATCH --job-name=QC
#SBATCH --output QC_rosmap.out.txt

cd /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid/

module load PLINK2/1.90b3.46

plink \
    --bfile /external/rprshnas01/netdata_kcni/dflab/data/rosmap/genotype/TOPmed_imputed/vcf/merged/merge
    --maf 0.01 \
    --hwe 1e-6 \
    --geno 0.01 \
    --mind 0.01 \
    --write-snplist \
    --make-just-fam \
    --out ROSMAP.QC
```

- A total of 9,329,439 variants were detected from the bim file
- No one was removed due to missing genotype data
- 978 variants removed due to missing genotype data
- 132 variants removed due to Hardy-Weinberg exact test
- 1,549,210 variants removed due to minor allele threshold
- 7,779,119 variants and 2067 people pass filters and QC

Extracting the SNP list from the ROS/MAP dataset:

```
#!/bin/bash --login
#SBATCH --time=01:00:00
#SBATCH --nodes=4
#SBATCH --ntasks=4
#SBATCH --job-name=QC
#SBATCH --output QC_rosmap2.out.txt

cd /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid/

module load PLINK2/1.90b3.46

plink \
    --bfile /external/rprshnas01/netdata_kcni/dflab/data/rosmap/genotype/TOPmed_imputed/vcf/merged/merge
    --keep ROSMAP.QC.fam \
    --extract ROSMAP.snplist \
    --out ROSMAP.QC
```

Heterozygosity

```
#!/bin/bash --login
#SBATCH --time=01:00:00
#SBATCH --nodes=4
#SBATCH --ntasks=4
#SBATCH --job-name=QC
#SBATCH --output QC_rosmap3.out.txt

cd /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid/

module load PLINK2/1.90b3.46

plink \
    --bfile /external/rprshnas01/netdata_kcni/dflab/data/rosmap/genotype/TOPmed_imputed/vcf/merged/merge
    --keep ROSMAP.QC.fam \
    --het \
    --out ROSMAP.QC
```

The output file ROSMAP.QC.het gives statistics for rates of SNP heterozygosity across the genome for each subject. Higher-than-expected rates indicate possible sample contamination (possibly inbreeding).

By using the .het file, we identify subjects with higher than expected heterozygosity. We identified outliers with F-values that are greater or less than 3 standard deviations from the sample mean.

```
dat <- read.table("ROSMAP.QC.het", header=T) # Read in the ROSMAP.het file, specify it has header
m <- mean(dat$F) # Calculate the mean
s <- sd(dat$F) # Calculate the SD
valid <- subset(dat, F <= m+3*s & F >= m-3*s) # Get any samples with F coefficient within 3 SD of the p
write.table(valid[,c(1,2)], "ROSMAP.valid.sample", quote=F, row.names=F) # print FID and IID for valid
```

15 people were removed from this process.

Relatedness

Now, we remove individuals with first or second degree relative $\hat{\pi} > 0.125$

```
#!/bin/bash --login
#SBATCH --time=01:00:00
#SBATCH --nodes=4
#SBATCH --ntasks=4
#SBATCH --job-name=QC
#SBATCH --output QC_rosmap5.out.txt

cd /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid_1

module load PLINK2/1.90b3.46

plink \
    --bfile /external/rprshnas01/netdata_kcni/dflab/data/rosmap/genotype/TOPmed_imputed/vcf/merged/merge
    --keep /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid/ROSMAP.
    --rel-cutoff 0.125 \
    --out /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid/ROSMAP.Q
```

- Nobody was removed due relatedness.

Creating the bed file

```
#!/bin/bash --login
#BATCH --time=01:00:00
#SBATCH --nodes=4
#SBATCH --ntasks=4
#SBATCH --job-name=QC
#SBATCH --output QC_rosmap.out

cd /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid

module load PLINK2/1.90b3.46

plink \
    --bfile /external/rprshnas01/netdata_kcni/dflab/data/rosmap/genotype/TOPmed_imputed/vcf/merged/merge
    --keep ROSMAP.QC.valid.sample \
    --make-bed \
    --out /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid/ROSMAP.Q
    --extract /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid/ROSMA
```

- 7,779,119 variants and 2,052 people pass filters and QC.

Removing the MHC region

Now, we remove the MHC region from the QC'd Genotype data (We used the assembly code GRCH38).

```
#!/bin/bash --login
#SBATCH --time=01:00:00
#SBATCH --nodes=4
#SBATCH --ntasks=4
#SBATCH --job-name=QC
#SBATCH --output QC_rosmapMHC.out.txt

cd /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid_MHC/

module load PLINK2/1.90b3.46

plink \
    --bfile /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid/ROSMAP
    --chr 6 \
    --from-bp 25477569\
    --to-bp 36480577\
    --write-snplist \
    --make-bed \
    --out ROSMAP.MHC.QC
```

- 55,204 out of 7,779,119 variants loaded from .bim file.

```
#!/bin/bash --login
#SBATCH --time=01:00:00
#SBATCH --nodes=4
#SBATCH --ntasks=4
#SBATCH --job-name=QC
#SBATCH --output QC_rosmapMHC1.out.txt

cd /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid_NO_MHC/

module load PLINK2/1.90b3.46

plink \
    --bfile /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid/ROSMAP
    --exclude /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid_MHC/
    --make-bed \
    --out ROSMAP.QC
```

- 7,723,915 variants and 2,067 people pass filters and QC.

Removing the APOE region

Now, we remove the APOE region from the QC'd Genotype data (We used the assembly code GRCH38).

```
#!/bin/bash --login
#SBATCH --time=01:00:00
#SBATCH --nodes=4
#SBATCH --ntasks=4
#SBATCH --job-name=QC
#SBATCH --output QC_rosmapMHCAPOE.out.txt

cd /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid_APOE_MHC/

module load PLINK2/1.90b3.46

plink \
    --bfile /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid_NO_MHC
    --chr 19 \
    --from-bp 44796743\
    --to-bp 44996742\
    --write-snplist \
    --make-bed \
    --out ROSMAP.MHCAPOE.QC
```

- 622 out of 7,723,915 variants loaded from .bim file.

```
#!/bin/bash --login
#SBATCH --time=01:00:00
#SBATCH --nodes=4
#SBATCH --ntasks=4
#SBATCH --job-name=QC
#SBATCH --output QC_rosmapMHCAPOE1.out.txt

cd /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid_NO_MHC_APOE/

module load PLINK2/1.90b3.46

plink \
    --bfile /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid_NO_MHC
    --exclude /external/rprshnas01/netdata_kcni/dflab/team/ak/Thesis/QC_Geno/ROSMAP/ROSMAP_QC_rsid_APOE
    --make-bed \
    --out ROSMAP.QC
```

- 7,723,293 variants and 2067 people pass filters and QC.

The final bed file has no MHC and APOE region.