Day 3 Practical: Imputation

Imputation is a technique that can help in identification as well as refinement of association results. Most of the widely used imputation panels are only available via online resources. In this practical, we will conduct pre-imputation QC and prepare the data for submission into Michigan server (https://imputationserver.sph.umich.edu/index.htm). As these runs often take hours, even days to complete, we will not wait for it but rather use predownloaded files from these runs, to conduct post-imputation QC. You can use the same steps to impute this dataset using the Genome Asia panel (on Michigan server) and the TopMed panel (https://imputation.biodatacatalyst.nhlbi.nih.gov/).

Input Dataset

Please find the input dataset in the Plink format here in your VM

```
cd Day3_Imputation/Day3_Impute_input
```

1. Pre-Imputation QC

The aim here is to prepare a dataset for imputation. The QC is similar to what you do for a normal GWAS, but is often slightly more stringent.

a. Filtering out SNPs with high missingness

```
plink --bfile chr21 --geno 0.05 --make-bed --out chr21.geno
```

b. Filtering out SNPs with high deviation for HWE

```
plink --bfile chr21.geno --hwe 0.0001 --make-bed --out
chr21.geno.hwe
```

c. Filtering out low MAF SNPs

```
plink --bfile chr21.geno.hwe --maf 0.01 --make-bed --out
chr21.geno.hwe.maf
```

d. Removing individuals with high missingness

```
plink --bfile chr21.geno.hwe.maf --mind 0.05 --make-bed --out
chr21.geno.hwe.maf.mind
```

e. Identifying duplicate SNPs

```
cp chr21.bim chr21.bim.backup        # For safety!

mv chr21.bim chr21.bim.original

awk '{$2=$1":"$4; print}' chr21.bim.original > chr21.bim

awk '{print $2}' chr21.bim > chr21.pos.txt
```

```
sort chr21.pos.txt | uniq -d. > chr21.dup.txt
```

f. Everything in a single step!

In real world you often run all the above commands above in the same go -

```
plink --bfile chr21 --geno 0.05 --hwe 0.0001 --maf 0.01 --mind 0.05
--exclude chr21.dup.txt --snps-only 'just-acgt' --make-bed --out
chr21.qc
```

Note: The exact QC parameters and the order they are employed in are indicative. As this particular dataset was aligned to positive strand, we did not require to align alleles. For other genotype datasets, the participants might need additional steps to get their alleles aligned to the positive strand and match the reference alleles using the resources at http://www.well.ox.ac.uk/~wrayner/tools/.

2. Preparing input files - Converting file to vcf format

```
plink --bfile chr21.qc --recode vcf-iid --out chr21.qc.final

bgzip chr21.qc.final.vcf

tabix -p vcf chr21.qc.final.vcf.gz
```

3. Running imputation at the Michigan server - Demo in separate pdf

a. Uploading data
b. Pre-phasing
c. Imputation

4. Post imputation QC

After the imputation job has finished, you will receive a web link to download the results in a (compressed) chr_21.zip file. You will also receive a password for unzipping it in your email. After unzipping chr_21.zip, you will get 2 compressed files:

```
chr21.info.gz

chr21.dose.vcf.gz
```

For the interest of time we have provided these two files downloaded from the Michigan server here:

a. Get a feel of the imputed data

```
cd Day3_Imputation/

unzip Day3_PostImpute.zip   #unzip the imputed data

cd Day3_PostImpute

bcftools view -h chr21.dose.vcf.gz > header.txt

bcftools view -H chr21.dose.vcf.gz | head -n 20 | cut -f 1-10 >
vcf_info.txt
```

You can open these two files – 'header.txt' and 'vcf_info.txt' and have a look

b. Filtering by genotyping probability and other parameters

The first step is to remove SNPs with high missingness; low minor allele frequency and low imputation quality

```
bcftools filter chr21.dose.vcf.gz -e "F_MISSING > 0.05 || MAF < 0.01
|| R2<0.6 " | bcftools convert -Oz -o chr21.filter.vcf.gz
```

Note: this step can take 10-20 minutes depending on your computational resources

c. Preparing index files

Prepare index files to easily study the summary of the data generated

```
tabix -p vcf chr21.filter.vcf.gz

bcftools index -n chr21.filter.vcf.gz        #Number of SNPs in VCF file
```

d. Conversion

Finally depending on your choice of tool for association testing you might need to convert the file into a suitable format. Here we do a final round of QC and export the file into plink. The QC includes an additional step - removal of extreme HWE outliers:

```
plink --vcf chr21.filter.vcf.gz --hwe 1e-5 --keep-allele-order --
allow-no-sex --make-bed --out chr21.filter.qc
```

Note : You can bring back the case control information into the plink dataset for association testing. For this you use a ".phe" file that has the case-control information and use an additional step (plink --bfile chr21.filter.qc --allow-no-sex –your_file.phe --make-bed --keep-allele-order --out chr21.filter.qc.final).