Day2: Haplotype phasing and Imputation

The goal of this practical is to introduce multiple key software packages commonly used for genotype phasing and imputation. Specifically, we will use the following ones:

1. SHAPEIT (Link). Fast and accurate haplotype estimation (i.e. phasing) software.
2. IMPUTE2 (Link). Imputation of un-typed genotypes onto either phased or unphased genotype data.

The raw material for this practical has been generated yesterday (chr20.FINAL.vcf.gz). It contains genotype data for multiple Spanish individuals on chromosome 20 for which we already have ensured high quality. We will now perform phasing and imputation of this data. Note that from the command lines described in the following, you should be able to build a phasing/imputation pipeline in your own research context. The code color in this practical is as follows:

1. **Grey** means command lines,
2. **Green** means questions that need answers.

0. Preparing the input data for phasing/imputation

Phasing and imputing entire chromosomes in a single run is feasible but very computationally demanding. This is why we usually require splitting the genotype data into MB sized chunks. This can be performed by specifying the coordinates of the chunk directly in the phasing/imputation command. But here, for the sake of clarity, we will first split the data prior to any other procedure.

BCFtools is very efficient to perform this task and requires as input a VCF indexed with tabix. First, copy the raw data in your working directory and make sure it is indexed using:

BASH

```
cp /home/delaneau/chr20.FINAL.vcf.gz .
tabix –fp vcf chr20.FINAL.vcf.gz
```

All along this practical, we are going to work on a 2 Mb region located between 1Mb and 3Mb. To avoid edge effects, we also add 200kb on each side of the region. The region coordinates are therefore [20:800000-3200000]. Extract the genotype data located in this region using BCFtools:

BASH

```
bcftools view -r 20:800000-3200000 -Oz -o chr20.chunk1.vcf.gz chr20.FINAL.vcf.gz
tabix -p vcf chr20.chunk1.vcf.gz
```

Q: How many genotyped variants are there in this region?

As it is often the case with genomic software, SHAPEIT and IMPUTE2 have their own file formats designed for imputed data. Fortunately, BCFtools includes conversion procedures. To convert the chunk of data into the correct file format, use:

BASH

```
bcftools convert -g chr20.chunk1 chr20.chunk1.vcf.gz
```

Look at the generated files using *less -S*. The *chr20.chunk1.samples* file contains the individual IDs while the *chr20.chunk1.gen.gz* file contains the actual genotype data in the form of genotype probabilities. Each genotype is encoded by a triplet. For instance, a genotype [1 0 0] means that the individual is homozygous reference allele while [0 1 0] means that it is heterozygous. Missing data is encoded with [0.33 0.33 0.33].

1. Phasing genotype data

To phase the genotype data, SHAPEIT requires a genetic map in which are specified the expected recombination rates between variants. This helps to build haplotypes with recombination events located at recombination hotspots. This file can be found here:

- /home/delaneau/reference/chr20.genetic_map.txt.gz

Now, let's proceed with the phasing using the following command:

BASH

```
shapeit --input-gen chr20.chunk1.gen.gz chr20.chunk1.samples
        --input-map /home/delaneau/reference/chr20.genetic_map.txt.gz
        --output-max chr20.chunk1.phased
        --output-log chr20.chunk1.phased.log
```

This should run in a couple of minutes. Let's look at the output using *less -S*:

BASH

```
less -S chr20.chunk1.phased.haps
```

Each row is a variant and each column is a haplotype (i.e. combination of 0s and 1s). Two consecutive columns give the two haplotypes for an individual following the same order than in the sample file.

2. Compute haplotype frequencies

From the phased haplotype file, we can now calculate haplotype frequencies for subsets of variants. Hereafter, the command you need to extract this information for 4 variants with IDs [*876578_T_C, 877205_G_T, 880964_A_C, 882429_A_G*]:

BASH

```
cat chr20.chunk1.phased.haps | grep '876578_T_C\|877205_G_T\|880964_A_C\|882429_A_G' |
cut -d' ' -f6- | /home/delaneau/transpose | sort | uniq –c
```

Note that */home/delaneau/transpose* is a simple *awk* script to transpose matrices. Now, compute the frequencies from the counts given by the command above. To do so, you can write the output of the command in a file and load it into ***R***.

Q: What is the most frequent haplotype? What is the less frequent? How many possible haplotypes do you expect for 4 variants? How many do you actually observe in the data?

3. Imputation of un-typed genotypes [Approach 1]

Now, let's impute un-typed genotype data onto the haplotypes we estimated in the previous section using a reference panel of haplotypes that only contains European samples. This reference panel is made of three files:

1. */home/delaneau/reference/chr20.EUR.hap.gz* contains the haplotypes,

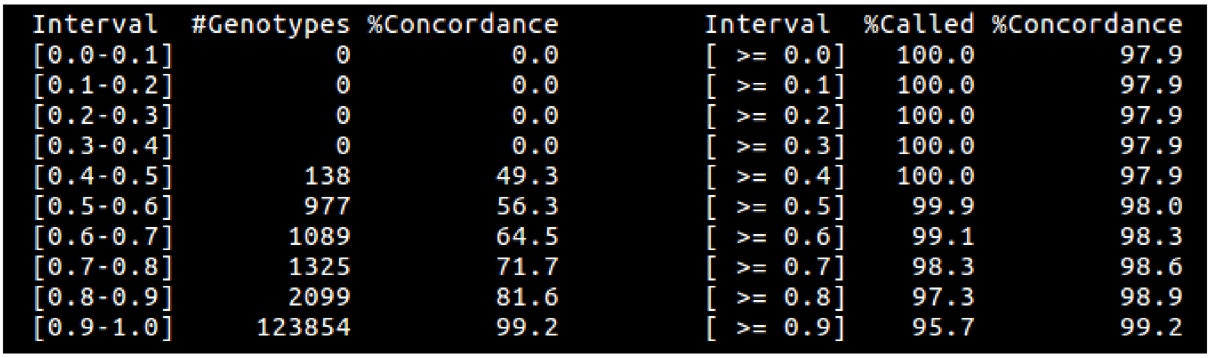
2. */home/delaneau/reference/chr20.EUR.legend.gz* contains the variant descriptions,
3. */home/delaneau/reference/chr20.EUR.samples* contains the sample IDs.

This procedure of performing first the phasing and then the imputation is usually termed as pre-phasing. To do so with IMPUTE2, use the following command below:

BASH

```
impute2 -use_prephased_g -phase                             #Pre-phasing mode
        -known_haps_g chr20.chunk1.phased.haps              #Our haplotypes
        -m /home/delaneau/reference/chr20.genetic_map.txt.gz #Genetic map
        -h /home/delaneau/reference/chr20.EUR.hap.gz        #Reference haplotypes
        -l /home/delaneau/reference/chr20.EUR.legend.gz     #Reference variants
        -int 1e6 3e6                                        #Chunk coordinates
        -buffer 200                                         #Buffer on each side in kb
        -o chr20.chunk1.imputed.approach1                   #Prefix for output files
```

Once the command completed, you should get something like this on the screen:



```
 Interval  #Genotypes %Concordance         Interval  %Called %Concordance
 [0.0-0.1]          0          0.0         [ >= 0.0]   100.0         97.9
 [0.1-0.2]          0          0.0         [ >= 0.1]   100.0         97.9
 [0.2-0.3]          0          0.0         [ >= 0.2]   100.0         97.9
 [0.3-0.4]          0          0.0         [ >= 0.3]   100.0         97.9
 [0.4-0.5]        138         49.3         [ >= 0.4]   100.0         97.9
 [0.5-0.6]        977         56.3         [ >= 0.5]    99.9         98.0
 [0.6-0.7]       1089         64.5         [ >= 0.6]    99.1         98.3
 [0.7-0.8]       1325         71.7         [ >= 0.7]    98.3         98.6
 [0.8-0.9]       2099         81.6         [ >= 0.8]    97.3         98.9
 [0.9-1.0]     123854         99.2         [ >= 0.9]    95.7         99.2
```

This constitutes a very important piece of information in order to check that imputation went well. This is obtained thanks to a cross-validation procedure implemented in IMPUTE2 in which known genotypes are masked and then imputed so that the software can then co,pare the predictions to the truth. In the first three columns is given, for multiple genotype probability bins:

1. The number of genotypes belonging to each bin,
2. The percentage of concordance between imputed and true genotypes (the higher the better).

The three right most columns give the same information in a cumulative way. You can note from this that 123854 genotypes have been imputed with a probability above 0.9 and these predictions are accurate at 99.2%. This looks good!

Q: Given this report, what can you say about probability calibration?

IMPUTE2 produces multiple output files. Amongst them, the two most important ones are:

1. *chr20.chunk1.imputed.approach1_info* which contains multiple key quality metrics on a per-variant basis. It notably gives an information score for each variant on column 7.
2. *chr20.chunk1.imputed.approach1* contains the imputed genotypes in the form of triplets. Take a look at the content of this file using *less -S*. Each triplet gives the probabilities of the 3 possible genotypes at a given site for a given individual. For example the triplet [0.33 0.67 0] means that the individual is heterozygous with probability 0.67 and homozygous for the reference allele with a probability of 0.33.

To assess imputation performance, we can also look at the information scores contained in the file *chr20.chunk1.imputed.approach1_info* and plot them as a function of allele frequencies:

R

```
#Load information scores
DATA1 = read.table("chr20.chunk1.imputed.approach1_info", head=TRUE)

#Only keep imputed variants
DATA1 = DATA1[DATA1$snp_id == "---", ]

#Compute MAF
DATA1$maf = ifelse(DATA1$exp_freq_a1 < 0.5, DATA1$exp_freq_a1, 1- DATA1$exp_freq_a1)

#Bin the imputed variants by MAF
DATA1$bin = cut(DATA1$maf, breaks = seq(0,0.5,0.05), labels = 1:10, include.lowest = TRUE)

#Plot average info scores versus MAF
pdf("chr20.chunk1.imputed.approach1.pdf")
plot(by(DATA1$maf, DATA1$bin, mean), by(DATA1$info, DATA1$bin, mean), type="l", xlab="MAF",
ylab="Mean info score", xlim=c(0,0.5))
dev.off()
```

Q: How does imputation perform at rare and common variants?

4. Imputation of un-typed genotypes [Approach 2]

Now, let's impute un-typed genotype data using a slightly different approach. We simultaneously phase and impute the genotype in a single run. This has been the common approach before the introduction of pre-phasing. We therefore need to go back to the original genotype data in *chr20.chunk1.gen.gz* and go ahead with the imputation step as follows:

BASH

```
impute2 -g chr20.chunk1.gen.gz                                  #unphased input data
        -m /home/delaneau/reference/chr20.genetic_map.txt.gz    #genetic map
        -h /home/delaneau/reference/chr20.EUR.hap.gz            #reference haplotypes
        -l /home/delaneau/reference/chr20.EUR.legend.gz         #reference variants
        -int 1e6 3e6  -buffer 200                               #buffer and target region
        -o chr20.chunk1.imputed.approach2                       #prefix for output files
```

Plot the imputation performance of this approach in the context of what we previously got:

R

```
#Load information scores
DATA1 = read.table("chr20.chunk1.imputed.approach1_info", head=TRUE)
DATA2 = read.table("chr20.chunk1.imputed.approach2_info", head=TRUE)

#Only keep imputed variants
DATA1 = DATA1[DATA1$snp_id == "---", ]
DATA2 = DATA2[DATA2$snp_id == "---", ]

#Compute MAF
DATA1$maf = ifelse(DATA1$exp_freq_a1 < 0.5, DATA1$exp_freq_a1, 1- DATA1$exp_freq_a1)
DATA2$maf = ifelse(DATA2$exp_freq_a1 < 0.5, DATA2$exp_freq_a1, 1- DATA2$exp_freq_a1)
```

```
#Bin the imputed variants by MAF
DATA1$bin = cut(DATA1$maf, breaks = seq(0,0.5,0.05), labels = 1:10, include.lowest = TRUE)
DATA2$bin = cut(DATA2$maf, breaks = seq(0,0.5,0.05), labels = 1:10, include.lowest = TRUE)

#Plot average info scores versus MAF for both approaches
pdf("chr20.chunk1.imputed.approach2.pdf")
plot(by(DATA1$maf, DATA1$bin, mean), by(DATA1$info, DATA1$bin, mean), type="l", xlab="MAF",
ylab="Mean info score", xlim=c(0,0.5), ylim=c(0.6, 1.0), col="grey")
points(by(DATA2$maf, DATA2$bin, mean), by(DATA2$info, DATA2$bin, mean), type="l", col="red")
legend("bottomright", legend=c("Approach 1","Approach 2"), fill=c("grey","red"))
dev.off()
```

Q: How does this approach compare to the previous one in terms of accuracy and running times? Which one do you recommend given these results?

5. Using a bigger and more diverse reference panel [Approach 3]

In a final imputation stage, we use a different reference panel that contains the same European samples plus many more additional samples with non-european ancestries. This will tell us if using cosmopolitan reference panels can lead to better imputation performance. This new reference panel is also made of three files:

1. */home/delaneau/reference/chr20.ALL.hap.gz* contains the haplotypes,
2. */home/delaneau/reference/chr20.ALL.legend.gz* contains the variant descriptions,
3. */home/delaneau/reference/chr20.ALL.samples* contains the sample IDs.

Perform imputation using this reference panel and the approach based on pre-phasing (Approach1). Write the outcome of this run in files with prefix *chr20.chunk1.imputed.approach3* and compare its performance to the two previous approaches using R as shown before.

Q: So, cosmopolitan reference panel does help or not?

6. Processing multiple chunks

Proceed with approach 3 to impute the region from 3Mb and 5Mb in order to get the file:

- *chr20.chunk2.imputed.approach3*

Then, bring all chunks of data together by using the *cat* command as follows:

BASH

```
#concat files of imputed genotypes
cat chr20.chunk1.imputed.approach3 chr20.chunk2.imputed.approach3 >
chr20.chunkALL.imputed.approach3

#concat files with info score (removing headers)
cat chr20.chunk1.imputed.approach3_info chr20.chunk2.imputed.approach3_info | grep -v position
> chr20.chunkALL.imputed.approach3_info
```

7. Post-processing on imputed genotype data

Once all chunks of data have been merged together, one can proceed with:

1. The conversion of the files to VCF for easy data management,
2. The filtering of badly imputed genotypes.

To convert imputed genotypes into a VCF file, bcftools works great:

BASH

```
#Convert to VCF (note that you can use any of the sample file)
bcftools convert -G chr20.chunkALL.imputed.approach3,chr20.chunk1.samples | bgzip -c >
chr20.chunkALL.imputed.approach3.vcf.gz

#Index the resulting VCF
tabix -p vcf chr20.chunkALL.imputed.approach3.vcf.gz
```

Note that in the VCF, genotypes have been set as those having maximal probabilities. For instance a genotype with probs = [0.3 0.6 0.1] will be set as heterozygous. This is a big approximation. However, imputation probabilities are also encoded in the file using the GP field.

In most studies based on imputed data, people usually work on variants with an info score above 0.4. Get the list of variants matching this filtering criterion and filter the VCF accordingly:

BASH

```
cat chr20.chunkALL.imputed.approach3_info | awk '{ if ($7 >= 0.4) print "20", $3; }' >
chr20.chunkALL.imputed.approach3.filtering.txt

vcftools --gzvcf chr20.chunkALL.imputed.approach3.vcf.gz --positions
chr20.chunkALL.imputed.approach3.filtering.txt --recode --stdout | bgzip -c >
chr20.chunkALL.imputed.filtered.approach3.vcf.gz
```

Q: How many imputed and genotyped variants did you get through this procedure? Is it worth to perform imputation?