Assignment 2.1

Anders Albrechtsten

September 25, 2020

1 Genotype calling based on genotype likelihoods

1.1 The data

We will use the data from the NGSadmix exercise[1]

1.1.1 Genotype likelihoods from the 1000 genomes low depth sequencing data:

Download from here[2] or use from the ricco server

```
/data/albrecht/advBinf/ngsAdmix/input.gz
```

The input.gz file contains 3 genotype likelihoods (assuming known major and minor allele e.i. genotypes 0, 1 and 2) for each individuals and each site. The 100 individuals are numbered from 0 to 99. 50,000 SNP sites were selected (to make it fast to run). Each line is a SNP sites.

1.1.2 Population information:

Download from here[3] or use from the ricco server

```
/data/albrecht/advBinf/ngsAdmix/pop.info
```

ID and population information for each individual. Same order as in the input.gz file

1.1.3 True genotypes

We have the true genotypes for these samples which we will use to estimate accuracy. Download from here[4] or use from the ricco server

[1] http://pontus.popgen.dk/albrecht/advBinf/web/NGSadmix.html
[2] http://pontus.popgen.dk/albrecht/open/input.gz
[3] http://pontus.popgen.dk/albrecht/open/pop.info
[4] http://pontus.popgen.dk/albrecht/open/input.geno

`/data/albrecht/advBinf/ngsAdmix/input.geno`

The data is in the same order (both SNPs and individuals) as the input.gz file.

1.2 Assignment

1.2.1 Estimate allele frequencies and ancestry proportions

The individuals are admixed so we cannot use the normal EM-algorithm to obtain allele frequencies therefore we need to use NGSadmix which will estimate both admixture proportions and allele frequencies for each ancestral population. Rerun the NGSadmix analysis assuming 3 ancestral populations without filtering away SNPs and using a seed (-s) with the following command

```
NGSadmix=/data/albrecht/advBinf/prog/angsd/misc/NGSadmix
$NGSadmix -likes /data/albrecht/advBinf/ngsAdmix/input.gz -K 3 \
-P 3 -minMaf 0 -o assign3 -s 1
```

Both the ancestry (.qopt) and minor allele frequency (.fopt.gz) output will contain 3 columns - one for each ancestral populations.

- Identify the ancestral populations of each of the 3 columns (same in both files) using the inferred admixture proportions and the population information file. Which columns represents Africa, European and Asian?

1.2.2 Call genotypes using different methods

The 61th individual, named NA12750 in the population information file, have genotype likelihood column names Ind60 in the input.gz file (remember there are 3 columns for each individual).

- Write code to call the 50,000 genotypes for NA12750 assuming a uniform prior for the 3 possible genotypes and calculate the posterior probability for each called genotype. Plot a histogram of the posterior probabilities.
- For the same individual write code to call genotypes using the frequency as a prior assuming Hardy-Weinberg equilibrium. Write your choice of frequency and plot a histogram of the posterior probabilities.
- Perform haplotype imputation using BEAGLE (the input.gz is in the beagle input format) on the same data and call genotypes based on the maximum posterior probability (see exercise for example). For the same individual extract the posterior probability as well and make a histogram.

1.2.3 Evaluation

In order to evaluate the methods in a proper manner make a plot of the error rate vs. the call rate (1-fraction of missing data). This is the best way to compare the accuracy of the different methods. We want to use the posterior probability as a cutoff for calling genotypes such that only genotypes with a high enough probability gets called while the others are set to missing. Here is an R example of how to plot it

```
plotAccuracy<-function(x,p,...){
    p<-p[!is.na(x)]
    x<-x[!is.na(x)]
    ord<-order(p,decreasing=T)
    lines(1:length(x)/length(x),cumsum(!x[ord])/1:length(x),...)
}
## posterior probabilities
PP <- c(0.9,0.1,0.45,0.99,0.15,0.9999,0.75,0.60,0.88,0.98)
## indicator of correct prediction
CP <- c(T,F,T,T,T,T,F,F,T,T) #you can also use 1s and zeroes

plot(1,xlim=0:1,ylim=c(0,0.40),col="transparent",xlab="callrate",ylab="error rate")
plotAccuracy(CP,PP,lwd=3,col="hotpink")
```

Use the true genotypes (input.geno file) to compare with your genotype calls.

- Make a plot with the accuracy of the three genotyping approaches.

You should then try to make a similar analysis for the third individual (NA19663,Ind2)

- Try to use all 3 ancestral population frequencies as a prior (one at a time)
- Try to make the optimal prior using a combinations of the 3 ancestral frequencies
- Plot the results in a single plot.

Hand in your code as plain text (not pdf) alongside you answers.