EMBO Population Genomics Practical 1

Andrea Manica

Introduction

ADMIXTOOLS is a widely used software package for calculating admixture statistics and testing population admixture hypotheses. However, although powerful and comprehensive, it is not exactly known for being user-friendly.

A typical ADMIXTOOLS workflow generally involves a combination of sed/awk/shell scripting and manual editing to create text configuration files. These are then passed as command-line arguments to one of ADMIXTOOLS commands, and control how to run a particular analysis. The results are then redirected to another file, which has to be parsed by the user to extract values of interest, often using command-line utilities again or (worse) by manual copy-pasting. Finally, the processed results are analysed in R, Excel or another program.

This workflow is quite cumbersome, especially if one wants to explore many hypotheses involving different combinations of populations. Most importantly, however, it makes it difficult to coduct reproducible research, as it is nearly impossible to construct fully automated "pipelines" that don't require user intervention.

We will use 'admixr', an R package that makes it possible to perform all stages of ADMIXTOOLS analyses entirely from R, completely removing the need for "low level" configuration of individual ADMIXTOOLS programs.

Connect to the server using ssh -X (to be able to use the X server for graphics), and start R by simply typing "R".

Installation (not needed for this course, for reference only)

For this course, we will connect to the server and run the analyses there. If you wanted to run the analyses on your computer, you will have to install both ADMIXTOOLS and the `adxmixr`. Detailed instrusctions are provided on the github site for `admixr`: https://github.com/bodkan/admixr

Loading in the data

ADMIXTOOLS software uses a peculiar set of genetic file formats, which may seem strange if you are used to working with VCF files. However, the basic idea remains the same: we want to store and access SNP data (REF/ALT alleles) of a set of individuals at a defined set of genomic positions.

EIGENSTRAT datasets always contain three kinds of files:

```
ind file - specifies a unique name, sex (optional - can be simply "U" for "undefined") and label (such a
snp file - specifies the positions of SNPs, REF/ALT alleles etc.;
geno file - contains SNP data (one row per site, one character per sample) in a dense string-based forma
    0: individual is homozygous ALT
    1: individual is a heterozygote
    2: individual is homozygous REF
    9: missing data
```

Therefore, a VCF file is essentially a combination of all three files in a single package.

Today, we will use a dataset that combined modern human populations (the Human Origin panel) and a number of ancient genomes. This dataset is found in Eigenstrat format in the practical1 folder under day4 on

the server (files with the prefix ho_anc). Assuming you are in the practical1 directory, you can see the names of those files with:

```
dir()
```

We can load the data using the function `eigenstrat` from the `admixr` package (note that the path assumes that you are in the practical1 directory).

```
library(admixr)
ho_anc <- eigenstrat("ho_anc")
```

We can summarise the inidivduals in the dataset:

```
ind <- read_ind(ho_anc)
table(ind$label)
```

And the number of SNPs available with:

```
ho_snp <- admixr::read_snp(ho_anc)
nrow(ho_snp)
```

Outgroup f3

We want to know which, among a set of samples of interest, is most similar to a Neolithic sample from the Linearbandkeramik culture, which is called LBK in our dataset. We load a list of names for modern populations of interest and turn it into a vector (read.csv reads in data as data.frames):

```
ref_pops <- read.csv("LBK.modern.pops.csv",colClasses = "character")[,1]
ref_pops
```

Now we compute the outgroup f3, using Mbuti (a divergent African population) as an outgroup:

```
LBK_outf3 <- f3(A = ref_pops, B = "LBK", C = "Mbuti", data = ho_anc)
head(LBK_outf3)
```

Now, let's check which populations are closest to LBK (have the highest f3)

```
head(LBK_outf3[order(LBK_outf3$f3, decreasing=T),],n=10)
```

Which populations share the most drift with LBK?

You could produce a nice plot by using:

```
library(ggplot2)
ggplot(LBK_outf3, aes(A, f3)) +
  geom_point() +
  geom_errorbar(aes(ymin = f3 - 2 * stderr, ymax = f3 + 2 * stderr)) +
  labs(y = "Shared drift with LBK", x = "populations") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

But it would be nicer if our populations were ordered by the level of shared drift:

```
LBK_outf3$A<-factor(LBK_outf3$A, levels = LBK_outf3$A[order(LBK_outf3$f3)])

ggplot(LBK_outf3, aes(A, f3)) +
  geom_point() +
  geom_errorbar(aes(ymin = f3 - 2 * stderr, ymax = f3 + 2 * stderr)) +
  labs(y = "Shared drift with LBK", x = "populations") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

Admixture f3

We now want to test whether a dataset of Africa American has detectable European ancestry:

```
aa_f3 <- f3(ho_anc, "Yoruba","French","AA")
aa_f3
```

Do we have evidence for admixture?

Now, we want to test admixture in two East African target populations (Somali from Somalia, Dinka from Sudan); Source populations are Mota (~4,000 year old individual from Ethiopia) and different modern and ancient (LBK) Eurasian populations:

```
euras_sources <- c("French","Spanish","Sardinian","LBK")
```

Test first the Somali against all possible combinations of Mota and these Eurasian sources, and then test the Dinka. What can you conclude?

We now want to investigate admixture in four modern Europeans (French, Basque, Tuscan and Sardinian) from three possible sources: Loschbour (Mesolithic hunter-gatherer); LBK (Early Neolithic farmers); Yamnaya (Steppe pastoralists). We will do that in two stages, first testing the input from Loschbour and LBK, and then from LBK and Yamnaya.

Which populations show neolithic farmer / hunter-gatherer admixture? Which populations have evidence for Yamnaya admixture

f4/D statistics

We keep focussing on our four modern target populations (French, Basque, Tuscan and Sardinian) and ask whether they forms a clade with LBK to the exclusion of Yamnaya. We will use again the Mbuti as an outgroup:

```
eur_d <- d (ho_anc, "Mbuti", "Yamnaya", "LBK", modern_targets)
eur_d
```

Which configuration is consistent with a simple tree? How can we interpret the configurations that fail the test?

f4 ratio estimation

Let's go back to the question of European admixture into African Americans. We will use an f4 ratio to estimate the proportion of European admixture:

```
AA_f4ratio <- f4ratio(ho_anc, "AA", "Han","French","Yoruba","pan_troglodytes")
AA_f4ratio
```

What is the proportion of European contribution into African Americans in the dataset? Repeat the analysis using other European populations (Sardinian, Spanish and Basque). Do the results change? ##