```
# 2024-04-05 Edited script for Gonçalo report
# to instal admixtools need a specific vs prog: RTools42
# following Rui Machado instruction download:
https://cran.r-project.org/bin/windows/Rtools/rtools42/rtools.html

install.packages("devtools")
devtools::install_github("uqrmaie1/admixtools")

library("admixtools")
# confirm your are in the same folder as the datasets needed in Allen database
(will be used later)
# read the original dataset from Olalde et al. 2019 and confirm a few
individuals = Allen dataset; these are three files GENO, IND and SNP
geno = read_packedancestrymap("./Olalde_et_al_genotypes")

#### example of data being saved as an object (read and saved). If you see this
is good news!
# Olalde_et_al_genotypes.geno has 278 samples and 1233013 SNPs.
# Reading data for 278 samples and 1233013 SNPs
# Expected size of genotype data: 2880 MB
#1233k SNPs read...
#1233013 SNPs read in total

# check format table of object geno (its a List, mixing many types of data),
element geno
head(geno$geno)
# you can see the ind and SNP using this command line, simply change the element
name after $
head(geno$snp)
head(geno$ind)

# get info for the SNP of interest, saving as an a object (vector type) - in
this case Also known as "C/T(-13910)" or just 13910T, and located in the MCM6
gene but with influence on the lactase LCT gene
mcm1 <- geno$geno["rs4988235",]
#basic plots... Just curious to confirm heterozygotes (genotype = 1) are missing
in aDNA
hist(mcm1)
plot(mcm1)

#########################################################################
# get the aadr dataset restricted to PIberia
#
https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genot
ypes-present-day-and-ancient-dna-data
# dataset: 1240k (less individuals/samples, many more SNPs)
#########################################################################
# Id samples to keep from the aadr dataset.
# selected Spain and Portugal, deselected modern from the .anno file open in
excel
# Filter by "Political Entity" (keeping only Spain and Portugal) and by "Full
Date" (excluding present). Copy the first column "Genetic ID" to a text file
called keep_samplesGROUPid_PIberia.txt
# select and saved individuals tag (i.e., names) present in the 1st column of
```

```
the .anno file open in excel into a .txt file named keep_samplesGROUPid_PIberia
# import the .txt file into R using finctionscan() and save as an object.
keep_samplesGROUPid <- scan("keep_samplesGROUPid_PIberia.txt", what =
"character") # for Gonçalo work
# you should have 583 items, i.e, ancient samples, no modern ones!

# read the data set from the Allen lab, restrict to Peninusla Iberia a DNA based
on the vector produced above using the option/parameter inds
genoIberianP = read_packedancestrymap("./v54.1.p1_1240K_public", inds =
keep_samplesGROUPid) # too large, won't work so select interest indv
#### example of data being saved as an object (read and saved). If you see this
is good news!
# v54.1.p1_1240K_public.geno has 16389 samples and 1233013 SNPs.
# Reading data for 583 samples and 1233013 SNPs
# Expected size of genotype data: 5889 MB
# 1233k SNPs read...
# 1233013 SNPs read in total

# check format table
head(genoIberianP$geno[1:100,1:10])

# get info for the SNP of interest, saving as a object
mcm12 <- genoIberianP$geno["rs4988235",]

# write file for object mcm12 which has all the info and then open it in Excel
write.table(mcm12, "aDNA_MCM1_genotypes_Iberia_AllenDataset.txt")

# basic plot, same result (same data)
hist(mcm1b)
plot(mcm1b)




## Comparing with the appearance of lactase persistance in Central Europe
(excluded medieval individuals and present, using filters in Excel)
## For this I opened in Excel the original .anno file, got rid of some columns,
replaced empty spaces with NA, replaced commas with _ and the saved it as .csv
file.
annotationfile <-read.csv("v54.1.p1_1240K_public_annotation_file.csv",
header=TRUE)



keep_samplesGROUPid2 <- scan("keep_samplesGROUPid_CentralEurope.txt", what =
"character")
genoCentralEurope = read_packedancestrymap("./v54.1.p1_1240K_public", inds =
keep_samplesGROUPid2) ## this creates a very large vectors that the computer
might not be able to handle.

# check format table
head(genoCentralEurope$geno[1:100,1:10])

# get info for the SNP of interest, saving as a object
```

```
mcm12 <- genoCentralEurope$geno["rs4988235",]

# write file for object mcm12 which has all the info and then open it in Excel
write.table(lac2, "aDNA_MCM1_genotypes_EastEurope_AllenDataset.txt")

# basic plot, same result (same data)
hist(mcm12)
plot(mcm12)

## In Excel I opened the aDNA_MCM1_genotypes_CentralEurope_AllenDataset.txt,
ordered these according to genotype ("0" on top) and copied the individuals into
a file called "CentralEuropeIndividualsWithLP.txt"


##To create a new dataframe with only the individuals of interest (those that
had a "0" in the genotype, ie have LP)
## Read the file containing individual names
individual_names <- read.table("CentralEuropeIndividualsWithLP.txt", header =
FALSE, stringsAsFactors = FALSE)

# Assuming your original dataframe is called 'annotationfile'
# Assuming the column containing individual names in your dataframe is called
'individual_name'
# Subset the original dataframe based on the names mentioned in the file into an
object called new_dataframe

new_dataframe <- annotationfile [annotationfile$Genetic.ID %in%
individual_names$V1, ]

## Save the file
write.table(new_dataframe, "DataOnCentralEuropeIndividualsWithLP.txt", sep=",")
```