Practical Course: Natural Selection

Owner	Patricio Pezo
Tags	

Directory

```
mkdir EMBO_course_2025
   mkdir input
   mkdir data_process
```

Goal

> Our goal is to explore approaches and methods, which seek to identify regions of the genome with signatures of natural selection. We will use real genomic data and two classes of tests: one based on population differentiation and another based on extended haplotype homozygosity.

Dataset

> Whole genome sequencing data by NGS (WG-NGS) from the 1000 Genomes Project phase III can be accessed through the link: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

> **The files to download are at:** https://github.com/HunemeierLab/EMBO_Practical_Course_2024

Data pre-processing

> To optimize our time, we will analyze a pre-processed dataset for chromosome 2 corresponding to individuals sampled from the African (504 individuals), European (503 individuals), and East Asian (504 individuals) populations of the 1000 Genomes).

For now, repeating these filters is unnecessary, but here are the commands used:

```
input=/Users/patriciopezo/Desktop/EMBO_course_2025/input
out=/Users/patriciopezo/Desktop/EMBO_course_2025/data_process

#I. Removing INDELS and Singletons
time vcftools --gzvcf $input/ALL.chr2.phase3_shapeit2_mvncall_integrated_v5b.20130502.genotypes.vcf.gz --remove-ind
els --min-alleles 2 --max-alleles 2 --maf 0.001 --max-maf 0.999 --recode --out $out/SNPs_Chr2_filter

#II. Selecting samples of individuals from the AFR, EAS and EUR populations (~ 30min) and filter to maf 0.05
vcftools --vcf SNPs_Chr2_filter.recode.vcf --keep pop_AFR_EAS_EUR_1000g.txt --min-alleles 2 --max-alleles 2 --maf 0.05
--max-maf 0.95 --recode --out SNPs_Chr2_AFR_EUR_EAS_maf

#III. Select individual samples for each population (for pairwise Fst comparison)
vcftools --vcf SNPs_Chr2_AFR_EUR_EAS_maf.recode.vcf --keep pop_AFR_1000g.txt --recode --out SNPs_Chr2_AFR_maf
&

vcftools --vcf SNPs_Chr2_AFR_EUR_EAS_maf.recode.vcf --keep pop_EAS_1000g.txt --recode --out SNPs_Chr2_EAS_maf
&

vcftools --vcf SNPs_Chr2_AFR_EUR_EAS_maf.recode.vcf --keep pop_EUR_1000g.txt --recode --outSNPs_Chr2_EUR_maf &

#IV. Estimating the Fst index between pairs of populations (~20 min each)
/vcftools --vcf ./dados/SNPs_Chr2_AFR_EUR_EAS_maf.recode.vcf --out AFR_EAS_maf --chr 2 --weir-fst-pop ./dados/pop
_AFR_1000g.txt --weir-fst-pop ./dados/pop_EAS_1000g.txt &

vcftools --vcf ./dados/SNPs_Chr2_AFR_EUR_EAS_maf.recode.vcf --out AFR_EUR_maf --chr 2 --weir-fst-pop ./dados/pop_
AFR_1000g.txt --weir-fst-pop ./dados/pop_EUR_1000g.txt &

vcftools --vcf ./dados/SNPs_Chr2_AFR_EUR_EAS_maf.recode.vcf --out EAS_EUR_maf --chr 2 --weir-fst-pop ./dados/pop_
EAS_1000g.txt --weir-fst-pop ./dados/pop_EUR_1000g.txt &
```

Part I | Genetic Differentiation as Evidence of Selection (FST-based Methods)

Through the exercises, discuss and answer the following questions:

1. The estimate of Fst by the Weir and Cockerham metric can sometimes generate negative values and “NA.” What does that mean? How can this interfere with the results?
2. The Fst values observed between pairs of populations for the SNP rs3827760 (position 109513601) fall within which distribution quantiles of Fst values for the studied chromosome? Can they be considered outliers?
3. From the observed Fst values between population pairs and the significance estimates, what can we say about the rs3827760 SNP differentiation between populations?
4. Discuss how these results justify performing another type of analysis based on PBS (population branch statistics).

What does the PBS analysis reveal? What is the difference between PBS and FST analysis?

Practice: Fst

```
#I. Read the files with the Fst estimates (AFR_EUR.weir.fst, AFR_EAS.weir.fst and EAS_EUR.weir.fst)

#II. Remove duplicated positions

#III. Take a look at the weir.fst file
```

#IV. Exclude NAs position in Fst estimations

#V. Overlaping SNPs

#VI. Convert negative values to zero

#VII. Check if the SNP rs3827760 (pos 109513601) is a candidate for natural selection
#1. Check if the SNP rs3827760, located at position 109513601, is an outlier in the FST distribution for any of the populatio n pairs

#2. Check which quartile percentile the rs3827760 distribution fall in each analyzed population pair?

#3. Ploting FST values in a 10,000 base pair region adjacent to the SNP at position 109513601. Highlight the SNPs that are outliers in the 95th percentile in each population pair.

#4. PLOT

#VIII. Can the candidate SNP be considered an outlier in all populations? What is the interpretation of this result?
#1. Estimate the p-value for the candidate SNP from the distribution of FST values for each population pair
#AfrEas

#AfrEur

#EurEas

Practice: Population Branch Statistic (PBS)

$$PBS = \frac{((-log(1 - FST\ AB)\ + (-log(1 - FST\ AC))\ -\ (-log(1 - FST\ BC))}{2}$$

Selection signals using the PBS

#1. Perform PBS test, using EAS as candidate population for selection
#Build the PBS Topology and why is it important to measure the distance

#2. Convert negative PBS values to O

#3. Add to the data.table with FST values, a new column with PBS values

#4. Check the PBS value for the candidate SNP.

#5. In which quartile of the distribution does the PBS value for the SNP rs3827760 fall?

#6. Plot the PBS values in a 10,000 base pair region adjacent to the SNP at position 109513601.
#Highlight the SNPs that are outliers in the 95th percentile.
#Select 10000bp adjacent to candidate SNP

#7. Subset the candidate SNP region

#8. Plot PBS values

Part II: Extended Haplotype Homozigosity (EHH)

Different approaches are able to detect genomic signatures of selection at different timescales. More recent selection signals can be detected from the extended haplotype homozygosity approach.

Practice: EHH

#I. Install the rehh R package

#II. Load rehh R package

#III. Use the following files
#Chr2_EDAR_LWK_500K.recode.vcf #(African population)
#Chr2_EDAR_CHS_500K.recode.vcf # (East Asian population)

#IV. What is the profile of ancestral and derived haplotypes of the rs3827760 SNP in AFR and EAS?

#1. Convert the data to haplohh format
#Use the 'data2haplohh' function

#2. Calculate the EHH for the candidate SNP (rs3827760) in AFR
#Use the 'calc_ehh' function

#3. Calculate the EHH for the candidate SNP (rs3827760) in EAS

#4. Plot EHH around “rs3827760” in AFR and EAS

#6. Calculate furcation trees around a candidate SNP in AFR

#7. Calculate furcation trees around a candidate SNP in EAS

Practice 2: integrated Haplotype score (iHS)

iHS is a measure of the amount of extended haplotype homozygosity at a given SNP along the ancestral allele relative to the derived allele. This measure is typically standardized empirically to the distribution of observed iHS scores over a range of SNPs with similar derived allele frequencies.

#1. Calculate the EHH for all SNPs in the file for AFR
#Use the 'scan_hh' function

#2. Calculate the EHH for all SNPs in the file for EAS

#3. Check eHH statistics for candidate SNP for AFR

```
#4.  Check eHH statistics for candidate SNP for EAS


#5.  Estimate the iHS in AFR (use min_maf = 0.02, freqbin = 0.01)
#Use the 'ihh2ihs' function

#6.  Estimate the iHS in EAS (use min_maf = 0.02, freqbin = 0.01)


#7.  Check the iHS score for the candidate SNP in AFR


#8.  Check the iHS score for the candidate SNP in EAS


#9.  Plot the iHS score in EAS
```

Make windows of SNPs

```
#1.  Create a function to estimate the mean in sliding windows.
slideFunct ← function(data, window, step){
 total ← length(data)
 spots ← seq(from = 1, to = (total - window + 1), by = step)
 result ← vector(length = length(spots))
 for(i in 1:length(spots)){
  result[i] ← mean(abs(data[spots[i]:(spots[i] + window - 1)]),na.rm=TRUE)
 }
 return(result)
}

#2.  Estimate the mean over a window of 50 SNPs with steps of 40 SNPs in EAS.


#3.  Identify the starting position of each window
slidePos ← function(data, window, step){
 total ← length(data)
 spots ← seq(from = 1, to = (total - window + 1), by = step)
 result ← vector(length = length(spots))
 for(i in 1:length(spots)){
  result[i] ← data[spots[i]]
 }
 return(result)
}

#4.  Put the position information and average iHS in a table


#5.  Identify the window which contains the candidate SNP


#6.  Plot the mean iHS per window
```

```
#6.  Check the distribution of iHS window in quantiles and check if the candidate SNP is an outlier.
```

#7. Add the cut line for the quartile to the graph

Practice 3: XP-EHH analysis in the candidate SNP?

Cross-population extended haplotype homozygosity (xp-EHH) method was developed to detect selective sweeps in which the selected allele has approached or achieved fixation in one population but remains polymorphic in the other.
Our candidate SNP is not polymorphic in Africans, but for the purposes of the exercise, let's perform windowed xp-EHH analysis on SNPs adjacent to rs3827760.

#1. Calculate the xp-EHH between EAS e AFR
#Use the function 'ies2xpehh'

#2. Calculate the average xp-EHH per 50 SNP window with 40 SNP steps

#3. Identify the starting position of each window

#4. Put the position information and average xpEHH in a table

#5. Identify the window which contains the candidate SNP

#6. Plot the mean xpEHH per window

#7. Check the distribution of xpEHH window in quantiles and check if the candidate SNP is an outlier.

#8. Add the cut line for the quartile to the graph and outline the candidate gene region