Methods for Detecting Intraspecific Natural Selection

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/

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). In this dataset we remove INDELs, Singletons and MAF less than 0.05. Next, the FST index was then estimated between pairs of populations using the vcftools program package.

For now, repeating these filters is unnecessary.

Let’s practice

Investigating a “Candidate Gene”

We refer to a “candidate gene” when we investigate whether there is evidence of natural selection in it based on previous results suggesting that it is a possible target for selection.

Detecting signatures of natural selection in the genome has the twofold meaning of (i) understanding which adaptive processes shaped genetic variation and (ii) identifying putative functional variants. In the case of humans, biological pathways enriched with selection signatures include pigmentation (Wilde et al. 2014), pathogen responses (Klunk et al. 2022; Couto-Silva, Nunes et al. 2023), and metabolic processes (Acuña-Alonzo et al. 2010).

The human Ectodysplasin A receptor gene, or *EDAR*, is part of the EDA signaling pathway which specifies prenatally the location, size, and shape of ectodermal appendages (such as hair follicles, teeth, and glands). *EDAR* is a textbook example of positive selection in East Asians (Sabeti et al. 2007) with genomic and functional experiments corroborating it. Also, genome-wide association studies found the same functional variant in EDAR associated hair morphology (Fujimoto et al. 2008) and incisor shape (Kimura et al. 2009) in East Asia populations and with several human facial traits (ear shape and chin protrusion) in Native American populations (Adhikari et al. 2016). Another plausible hypothesis stated that *EDAR* acted with *FADs* and *VDRs* genes in the Beringia Standstill (Hlusko et al. 2018), allowing these populations to survive in this extreme environment.

Natural Selection Tests

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

PRACTICE: FST

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

1. Use R to run the following commands

a. Read the files with the Fst estimates (AFR_EUR.weir.fst, AFR_EAS.weir.fst and EAS_EUR.weir.fst)
b. Eliminate duplicate positions
c. Take a look at the weir.fst file

2. The estimation of FST by Weir and Cockerham can sometimes generate negative values, and Na. What does this mean? How can it interfere with the results?

a. Exclude positions whose FST result was equal to Na
b. The datasets now have different sets of SNPs, filter the files by overlapping the SNPs between them.
c. Convert positions with estimates from FST < 0 to = 0

3. Now with the data filtered and matched, let's check if the SNP rs3827760 (pos 109513601) is a candidate for natural selection.

a. Check if the SNP rs3827760, located at position 109513601, is an outlier in the FST distribution for any of the population pairs.
b. In which quartile of the distribution does the FST value for the SNP rs3827760 fall in each analyzed population pair?
c. Please plot the 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. Can the candidate SNP be considered an outlier in all populations? What is the interpretation of this result?

a. Estimate the p-value for the candidate SNP from the distribution of FST values for each population pair.

PRACTICE: Population Branch Statistics (PBS)

Use R to:

1. Based on population differentiation methods, let's explore selection signals using the PBS (Population Branch Statistic) approach:

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

a. Perform PBS test, using EAS as candidate population for selection
b. Convert negative PBS values to 0.
c. Add to the data.table with FST values, a new column with PBS values.
d. Check the PBS value for the candidate SNP.
e. In which quartile of the distribution does the PBS value for the SNP rs3827760 fall?
f. 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
   - Subset the candidate SNP region
   - Plot PBS values

2. Based on FST and PBS tests applied so far, what can we conclude?

PART II

EXTENDED HAPLOTYPE HOMOZYGOSITY (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.

With the following exercises, we seek to answer the following questions:

1) How is the haplotype profile of genetic variants under recent positive selection?
2) What is the profile of ancestral and derived haplotypes of the rs3827760 SNP in AFR and EAS?
3) The iHS score observed for the SNP rs3827760 fall within which distribution quantiles of iHS values for the studied chromosome? Can they be considered an outlier? How can we make this analysis more robust?
4) What information does the xp-EHH analysis add about natural selection in the candidate SNP?

PRACTICE: EHH

Use R to:

Install the rehh R package

```
install.packages("rehh")
```

Load rehh R package

```
library("rehh")
```

Use the following files

```
Chr2_EDAR_LWK_500K.recode.vcf #(African population)

Chr2_EDAR_CHS_500K.recode.vcf # (East Asian population)
```

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

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

a. Convert the data to haplohh format
b. Calculate the EHH for the candidate SNP (rs3827760) in AFR
c. Calculate the EHH for the candidate SNP (rs3827760) in EAS
d. Plot EHH around "rs3827760" in AFR
e. Plot EHH around "rs3827760" in EAS
f. Calculate furcation trees around a candidate SNP in AFR
g. Calculate furcation trees around a candidate SNP in AFR

2. The integrated Haplotype Score (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.

a. Calculate the EHH for all SNPs in the file (~5min) for AFR
b. Calculate the EHH for all SNPs in the file (~5min) for AFR
c. Check eHH statistics for candidate SNP for AFR
d. Check eHH statistics for candidate SNP for AFR
e. Estimate the iHS in AFR (use min_maf = 0.02, freqbin = 0.01)
f. Estimate the iHS in EAS (use min_maf = 0.02, freqbin = 0.01)
g. Check the iHS score for the candidate SNP in AFR
h. Check the iHS score for the candidate SNP in EAS
i. Plot the iHS score in EAS

3. As we are looking at haplotypes, several individual SNPs have outlier values. One way to make the analysis more robust is to average a window of SNPs.

a. Create a function to estimate the mean in sliding windows.
b. Estimate the mean over a window of 50 SNPs with steps of 40 SNPs in EAS.
c. Identify the starting position of each window
d. Put the position information and average iHS in a table
e. Identify the window which contains the candidate SNP
f. Plot the mean iHS per window
g. Check the distribution of iHS window in quantiles and check if the candidate SNP is an outlier.
h. Add the cut line for the quartile to the graph

4. What information does the xp-EHH analysis add about natural selection 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.

a. Calculate the xp-EHH between EAS e AFR
b. Calculate the average xp-EHH per 50 SNP window with 40 SNP steps
c. Identify the starting position of each window
d. Put the position information and average xpEHH in a table
e. Identify the window which contains the candidate SNP
f. Plot the mean xpEHH per window
g. Check the distribution of xpEHH window in quantiles and check if the candidate SNP is an outlier.
h. Add the cut line for the quartile to the graph and outline the candidate gene region

PART III

Hlusko et al. (2018) using morphological data, found a strong selection signal in the *EDAR* gene in Native Americans. Although the proposed hypothesis is well supported, there was not enough data to perform genomic selection tests at that time. With that in mind, and using the additional database (1KGP Peruvian samples with over 95% Native American Ancestry), answer the following questions:

1. Is the functional allele in East Asian at high frequency in other human populations (e.g. Native Americans)?

2. Can we identify signatures of natural selection on EDAR in Native Americans using PBS?

3. Is selection targeting the same functional variant?

4. What is your conclusion based on data generated?