Evaluation Homework Evolutionary Genomics 2020

Vitor Sousa

vitor.sousa@iee.unibe.ch

Two Options for Evaluation:

- Exercise 1. R script to code coalescent simulations

OR

- Exercise 2. Fastsimcoal2 inference of demographic history based on SFS.

Exercise 1
Coalescent simulations

Enard et al. (2002) looked at polymorphism patterns in the exons of FOXP2 gene in 20 humans sampled across the globe. By sequencing 14,063 bp they found 47 segregating sites (S_{obs}=47). Assuming a mutation rate of μ=1.4 x 10^{-8} /site/generation, we would like to know if this S_{obs} value is better explained by a small (N_e=10,000) or a large (N_e=1,000,000) effective size?

The aim is to find what is the probability of observing S_{obs}=47 segregating sites (out of 14,063 bp) in sample of 20 diploid individuals from a population with a given constant size N_e. To answer this question we ask you to use coalescent simulations to obtain the distribution of the number of segregating sites S for different values of N_e. This will be done assuming that the number of mutations occurring in a gene tree follows a Poisson distribution with rate $T_L\,\mu\,L$, where T_L is the total branch length, μ is the mutation rate per site per generation and L is the length of the sequence in bp.

NOTE: FOXP2 is a gene that seems to be involved in human speech, and that has been invoked to be under recent positive selection in the human lineage. However, this has been highly debated.

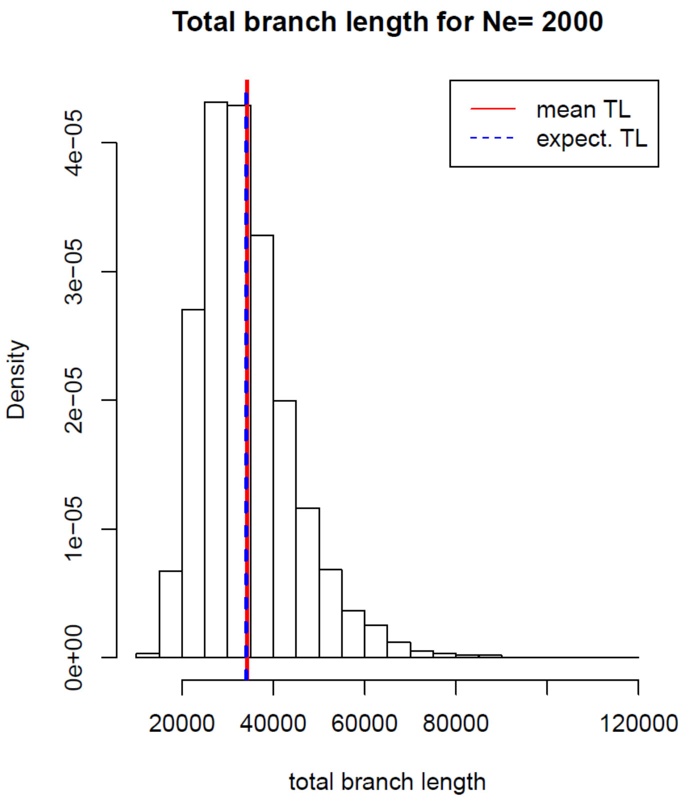
Enard et al. (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418, 869-872

Exercise 1
Distribution of Total branch length

Write an R script to perform simulations following these steps:

Write a function to simulate coalescent times that gets as input the effective size of the population (*Ne*) and the sample size *n*, and outputs the total branch length. The coalescent times (in generations) are geometrically distributed, and hence you can get random samples with the *rgeom()* function. Note that you only need to simulate the time intervals between coalescent events and record the total branch lengths. **The tree topology does not need to be simulated**. The effective size and sample size are given as number of diploid individuals.

1. For an N_e value of 10,000 perform 10,000 coalescent simulations to obtain the total branch length distribution. Consider using the *replicate* function to call a function several times.
    i. Plot the distribution of the total branch length T_L using the function *hist()*. Compare the mean of the distribution with the expected value computed in class.



Example of plots with *Ne*=2000

Exercise 1

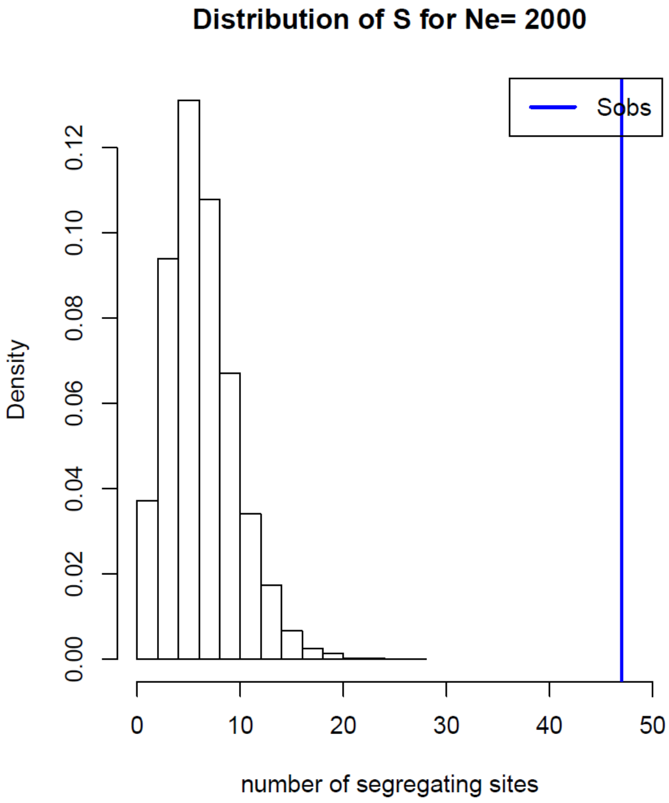
Distribution of the number of segregating sites

ii. Plot the distribution of the number of segregating sites using the function *hist()* and compare it with the observed S_{obs}=47.

iii. Approximate the likelihood Pr(S_{obs}|n,Ne,μ,L) as the proportion of simulations with S_{obs} segregating sites

Repeat the above steps with an N_e=1,000,000

Does the data support an effective size for humans of *Ne*=10,000 or *Ne*=1,000,000?



Example of plots with *Ne*=2000

Exercise 1 - Evaluation

- Send to vitor.sousa@iee.unibe.ch the R code, with the PDF of the distributions of the total branch length and number of segregating sites. The files should be named according to the following template:
  - R script: "EvolGenomics_[YOURNAME]_script.r";
  - Figures: "EvolGenomics_[YOURNAME]_figures.pdf"

- The subject of the e-mail should be "Evol Genomics Homework"

Deadline: 30th March 2020

Exercise 2 - Evaluation

- Perform exercise 3 from "Day2/PracticalFastsimcoal/PracticalFastsimcoal2.html"

- Write a report in English (max 5 pages) describing and discussing the results, answering the questions in the "To Discuss" section.

Send a PDF with the report to vitor.sousa@iee.unibe.ch. The file should be named according to the following template:

"EvolGenomics_[YOURNAME]_fastsimcoal2.pdf";

The subject of the e-mail should be "Evol Genomics Homework"

Deadline: 30th March 2020