A BETA SPLITTING MODEL FOR COALESCENT TREES

Detecting Natural Selection from the whole genome sequences

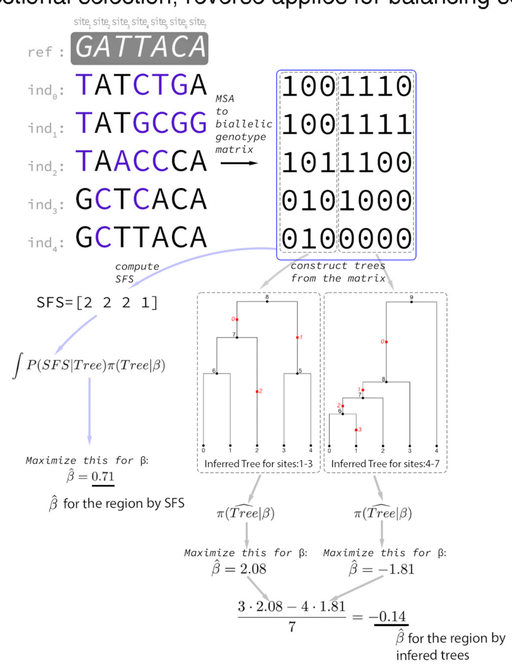
Enes Dilber, Jonathan Terhorst
Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA

Introduction

- We study a new technique for detecting signals of **natural selection** in large samples of genetic data.
- We used **coalescent process** to construct our models. It is a tool to describe evolutionary history of gene samples independently drawn from a population.
- For neutrally evolving populations the tree shape is quite easy to formulate but selection of a gene **distrots the tree topology**.
- We used a coalescent based likelihood method to detect the selection by modeling the distortion in the tree topology introduced by the selection. Our estimated parameter indicates whether it is a **balancing** or a **directional** selection signal.

Methods

- We measured the tree imbalance with our likelihood based method, **β-imbalance**. If the splits on a tree subtends uneven number of children, the estimated paremeter become negative which might imply a directional selection, reverse applies for balancing selection.

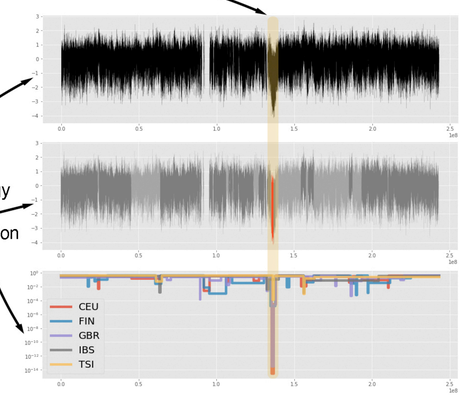


Results for 1000 Genomes Project

- Whole-genome sequencing of 2504 individuals from 26 populations (Auton et al. (2015)). Each corresponds to one of five super-populations: African (AFR), Ad Mixed American (AMR), East Asian (EAS), European (EUR), South Asian (SAS) ([1,2]).

- **LCT** is highlighted
  Lactose tolerance mutation occurred in Europe, under very strong positive selection. Here is the Genome-scan for chromosome 2.

- Algorithm:
  1) Estimate $\hat{\beta}$ for each geneology
  2) Apply change point detection
  3) Calculate p-vals for each region
  4) Do 1-3 for all populations
  5) Report for the region



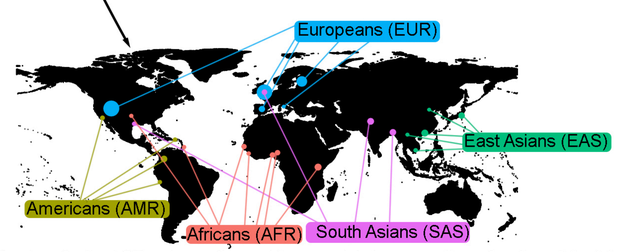
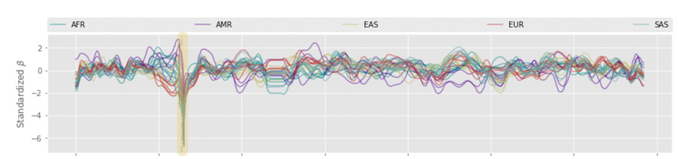


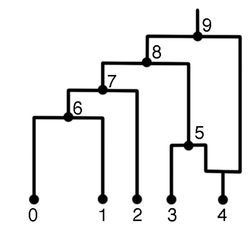
Figure: P-values for the LCT region on the world map, point sizes are a linear function of -log(p)

- **Results on MHC**
  Another region that we tried our model is Major Histocompatibility Complex (MHC) on chromosome 6. It contains a set of closely linked polymorphic genes that code for cell surface proteins essential for the adaptive immune system. All populations experience selection in this region. Here is the smoothed β-imbalance scans. MHC region is highlighted.



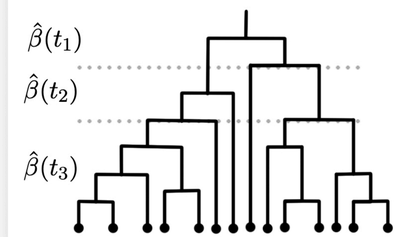
Some Extensions

- Imbalance on Ancestral Recombination Graphs (ARG)
  The ARG is a DAG whose nodes correspond to genotypes and edges correspond to either recombination or coalescence events. Here is the ARG representation of the trees in the Methods section:



Advantages:
Compact representation of the region
Not overcounting the nodes
Disadvantage:
Computation is quite challanging

- Time Dependent Model
  We can fit not a single β parameter but a function of β through the time (β(t)).



Advantage:
Might indicate period of natural selection
Disadvantages:
Computation time increases
Usually estimates are nosier
Requires larger sample sizes

Summary

- We developed a tree imbalance measure to detect the genes that experience natural selection. We provided a data application to show our model works in real life data. We provided how to extend the current model (I) to make it more robust, accurate and (II) to give more information about the natural selection.

Literature Cited

[1] Auton, A. et al. (2015) A global reference for human genetic variation. Nature, 526(7571), 68–74. Number: 7571 Publisher: Nature Publishing Group.

[2] Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., & McVean, G. (2019). Inferring whole-genome histories in large population datasets. Nature Genetics, 51(9), 1330–1338.

LSA STATISTICS
UNIVERSITY OF MICHIGAN