COMPUTING IN EVOLUTIONARY ANALYSIS

Evolutionary Genetics
SJSU
Spring 2018

HOW DO WE TYPICALLY ANSWER EVOLUTIONARY QUESTIONS?

- Step 0: idea -> “Proteins have important biological functions. Maybe this means that natural selection acts on proteins in some way”

HOW DO WE TYPICALLY ANSWER EVOLUTIONARY QUESTIONS?

- Step 0: idea -> "Proteins have important biological functions. Maybe this means that natural selection acts on proteins in some way"
- Step 1: Hypothesis -> "Non-synonymous sites in the human genome are likely to be under stronger selection than synonymous sites"

HOW DO WE TYPICALLY ANSWER EVOLUTIONARY QUESTIONS?

- Step 0: idea -> "Proteins have important biological functions. Maybe this means that natural selection acts on proteins in some way"
- Step 1: Hypothesis -> "Non-synonymous sites in the human genome are likely to be under stronger selection than synonymous sites"
- Step 2: Design a test for the hypothesis -> "Non-synonymous sites are likely to have lower allele frequencies than synonymous sites if they are under selection"

HOW DO WE TYPICALLY ANSWER EVOLUTIONARY QUESTIONS?

- Step 0: idea -> "Proteins have important biological functions. Maybe this means that natural selection acts on proteins in some way"
- Step 1: Hypothesis -> "Non-synonymous sites in the human genome are likely to be under stronger selection than synonymous sites"
- Step 2: Design a test for the hypothesis -> "Non-synonymous sites are likely to have lower allele frequencies than synonymous sites if they are under selection"
- Step 3: Run the experiment -> sequence some genomes, and see if the pattern we expected holds up.

ARE NON-SYNONYMOUS SITES UNDER SELECTION?

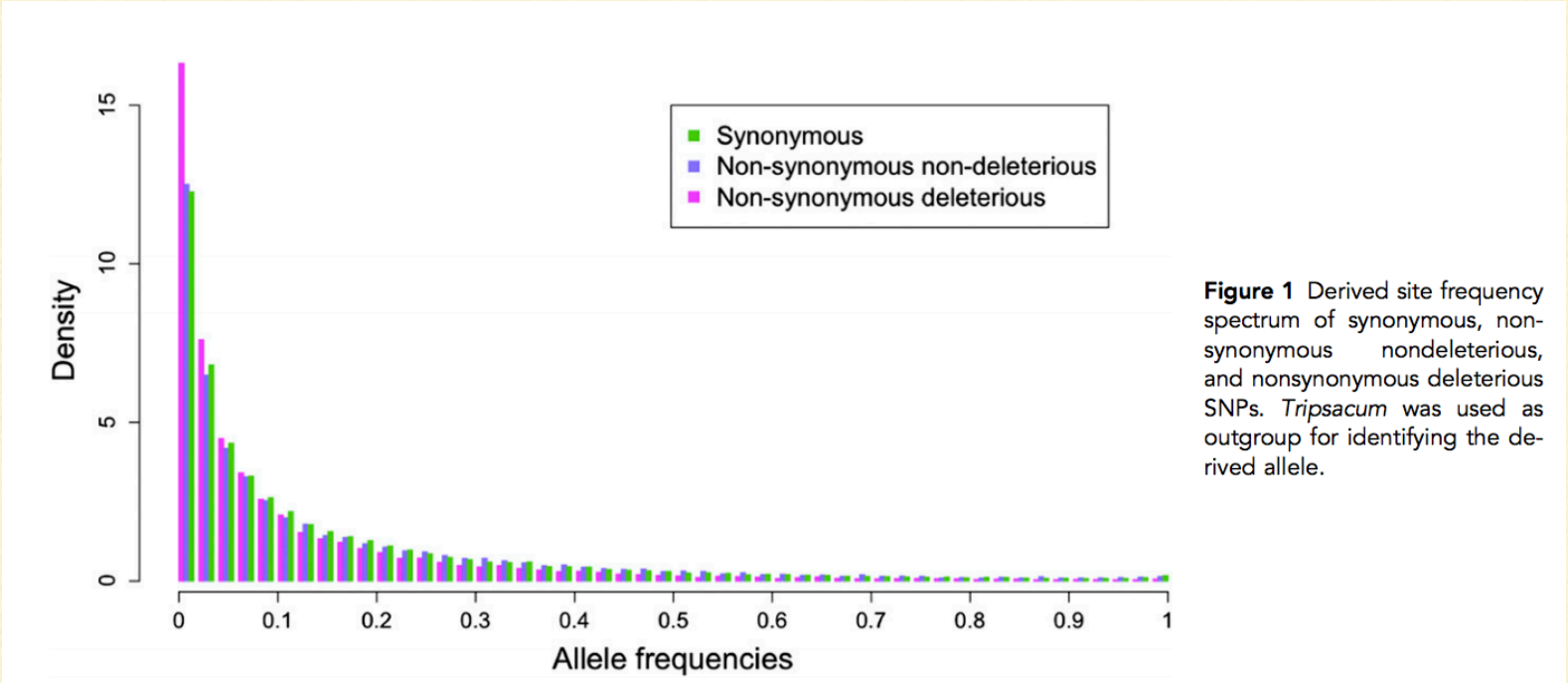


Figure 1 Derived site frequency spectrum of synonymous, non-synonymous nondeleterious, and nonsynonymous deleterious SNPs. *Tripsacum* was used as outgroup for identifying the derived allele.

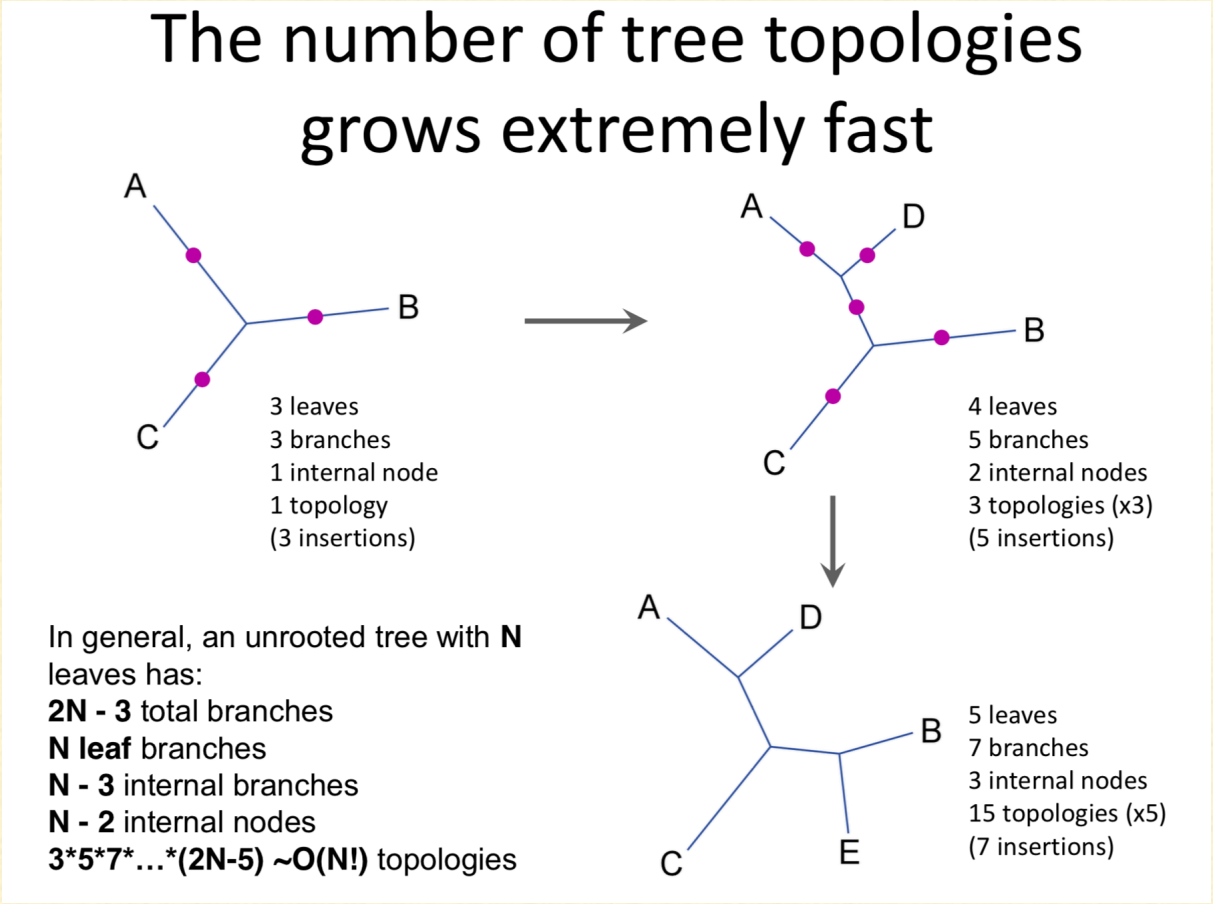
HOW DO COMPUTERS FIT IN?

- Step 0: idea -> "Proteins have important biological functions. Maybe this means that natural selection acts on proteins in some way"
- Step 1: Hypothesis -> "Non-synonymous sites in the human genome are likely to be under stronger selection than synonymous sites"
- Step 2: Design a test for the hypothesis -> "Non-synonymous sites are likely to have lower allele frequencies than synonymous sites if they are under selection"
- Step 3: Run the experiment -> sequence some genomes, and see if the pattern we expected holds up.

HOW DO COMPUTERS FIT IN?

- Step 0: idea -> "Proteins have important biological functions. Maybe this means that natural selection acts on proteins in some way"
- Step 1: Hypothesis -> "Non-synonymous sites in the human genome are likely to be under stronger selection than synonymous sites"
- **Step 2: Design a test for the hypothesis -> "Non-synonymous sites are likely to have lower allele frequencies than synonymous sites if they are under selection"**
- **Step 3: Run the experiment -> sequence some genomes, and see if the pattern we expected holds up.**

COMPUTERS IN HYPOTHESIS COMPARISON

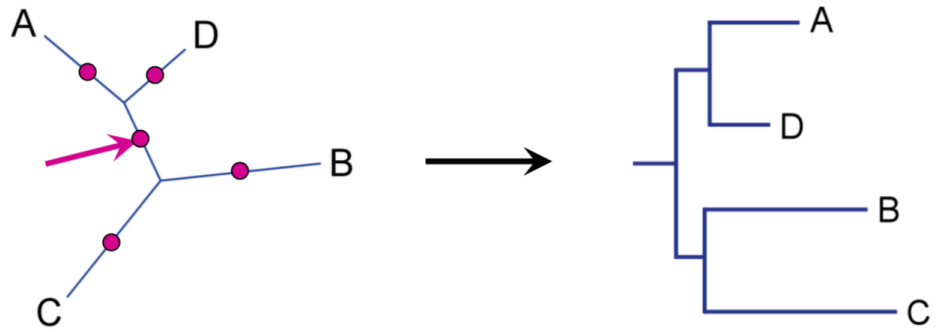


http://elbo.gs.washington.edu/courses/GS_373_12_sp/slides/W9C2-DistTrees.pdf

COMPUTERS IN HYPOTHESIS COMPARISON

There are many rooted trees for each unrooted tree

For each unrooted tree, there are 2N - 3 times as many rooted trees, where N is the number of leaves (# internal branches = 2N – 3).



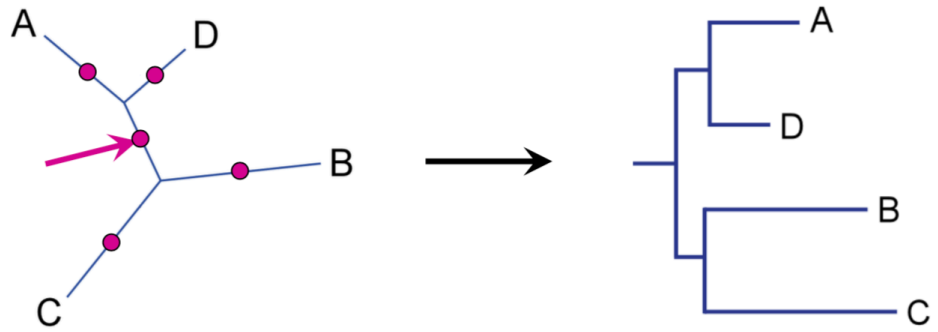
20 leaves - 564,480,989,588,730,591,336,960,000,000 topologies

http://elbo.gs.washington.edu/courses/GS_373_12_sp/slides/W9C2-DistTrees.pdf

COMPUTERS IN HYPOTHESIS COMPARISON

There are many rooted trees for each unrooted tree

For each unrooted tree, there are 2N - 3 times as many rooted trees, where N is the number of leaves (# internal branches = 2N – 3).



20 leaves - 564,480,989,588,730,591,336,960,000,000 topologies

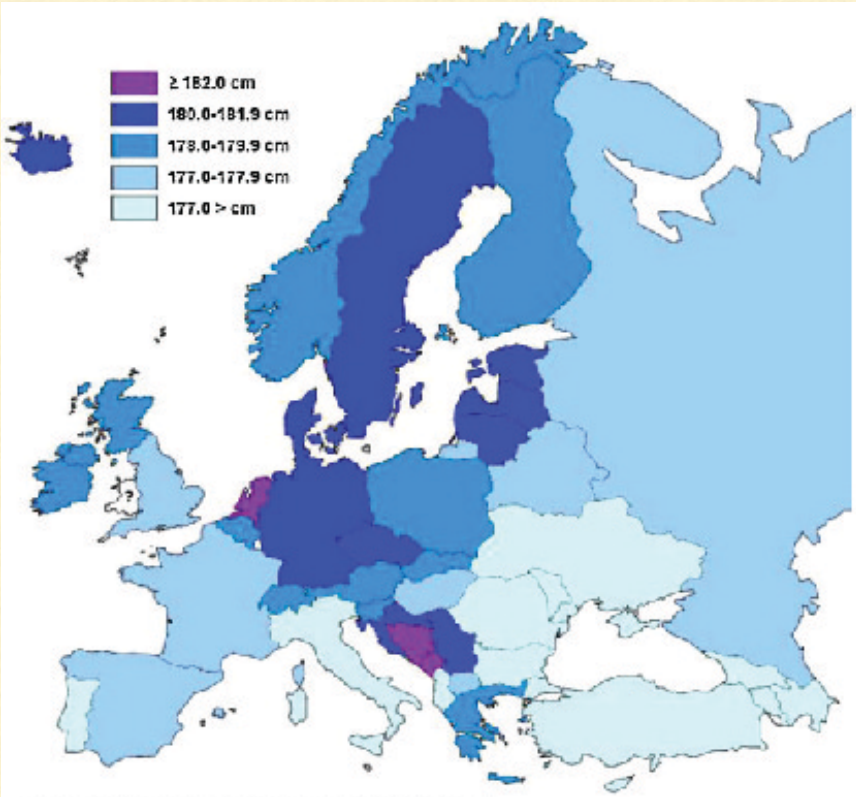
This implies that we need computers and clever algorithms to compare hypotheses when we compare many species

http://elbo.gs.washington.edu/courses/GS_373_12_sp/slides/W9C2-DistTrees.pdf

COMPUTERS IN DATA GENERATION AND STORING

- Human genome has 3,000,000,000 base pairs
- Storing a compressed human genome takes O(GB) of data
- Obviously, we can't sort through all this data by hand
- Even accessing it efficiently on a computer requires clever storage

ACTIVITY: USING COMPUTATION TO UNDERSTAND HUMAN ADAPTATION



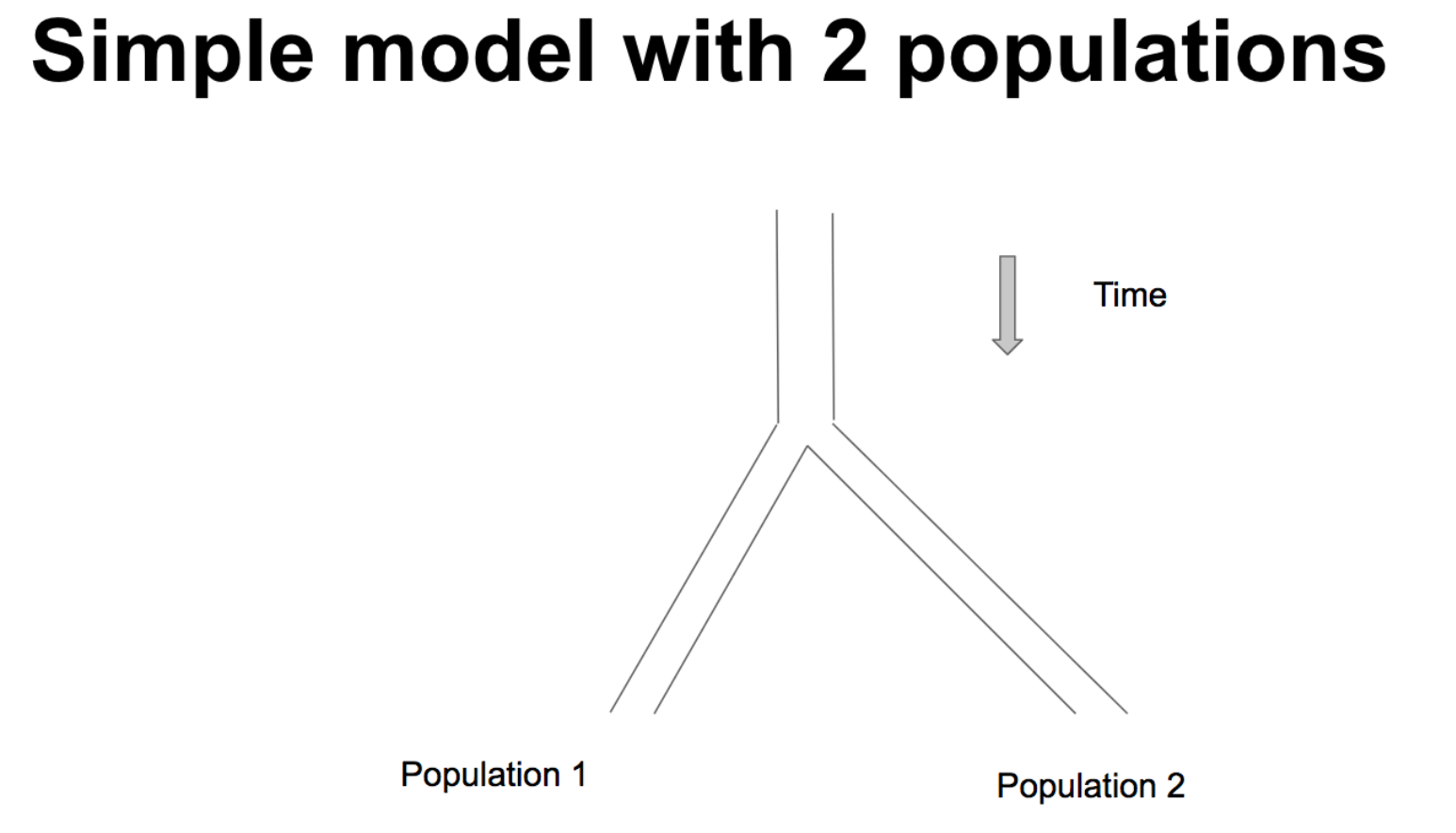
Grasgruber et al. 2014 - Mean male height

Robinson et al. 2015 - Mean male & female height

HOW CAN WE TEST WHETHER SELECTION ON HEIGHT MAY HAVE OCCURRED?

- Hypothesis: "Natural selection acted to increase height as Europeans entered Europe"
- Alternate hypothesis: "Changes in mean height across Europe are not due to natural selection, but instead some combination of random genetic fluctuations (genetic drift) and environmental effects"

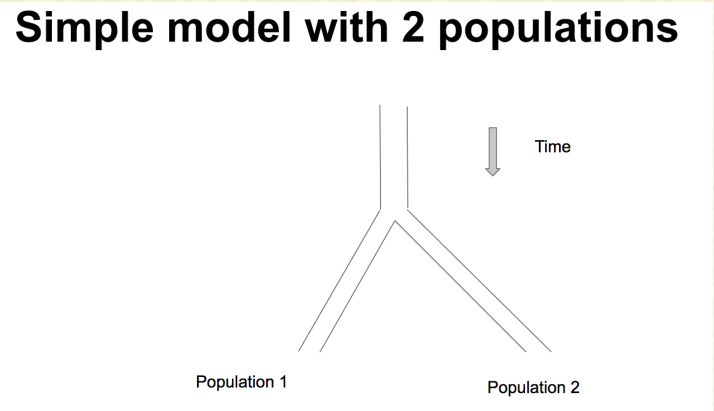
WHAT PATTERNS MIGHT SELECTION GENERATE?



THINK-PAIR-SHARE

- What factors would affect the rate of change of a trait in each population (e.g., do you expect genetic factors to be important? Environmental factors? The amount of time?)

- If selection has not acted on the trait in either population, do we expect the trait to differ between populations over time?



TESTING OUR INTUITION

- I created a simple simulator of this model for alleles and traits, which allows us to vary:
  - Time since divergence
  - Population size
  - Fraction of alleles under selection
  - Number of alleles that affect the trait

TESTING OUR INTUITION 2

- Now that you have seen the allele simulator, how do we expect the following factors to affect the rate of divergence of the trait:
  - Time since divergence
  - Population size
  - Fraction of alleles under selection
  - Number of alleles that affect the trait

TESTING THIS HYPOTHESIS IN REAL DATA

- Since we know some of the alleles that affect height, we can ask whether these alleles are systematically at higher frequencies among northern Europeans than southern Europeans

Evidence of widespread selection on standing variation in Europe at height-associated SNPs

Michael C Turchin[1–5,8], Charleston WK Chiang[1–6,8], Cameron D Palmer[1–5], Sriram Sankararaman[5,6], David Reich[5,6], Genetic Investigation of ANthropometric Traits (GIANT) Consortium[7] & Joel N Hirschhorn[1–6]

CONCLUSIONS

- Computation is increasingly important in solving evolutionary mysteries because
  - 1) even designing ways to test hypotheses requires clever algorithms and
  - 2) Generating, storing, and accessing (large) genetic datasets also requires sophisticated algorithms
- Computation/simulations can also be used to gain intuition about evolutionary processes
- We made a simple simulation to show how traits change over time with and without selection
  - When there is no selection, the trait can change randomly over time. How rapidly it changes depends on factors like the number of alleles that change the trait
  - When selection acts, these changes occur more rapidly and in a concerted way (e.g., by systematically increasing the frequencies of certain types of alleles).