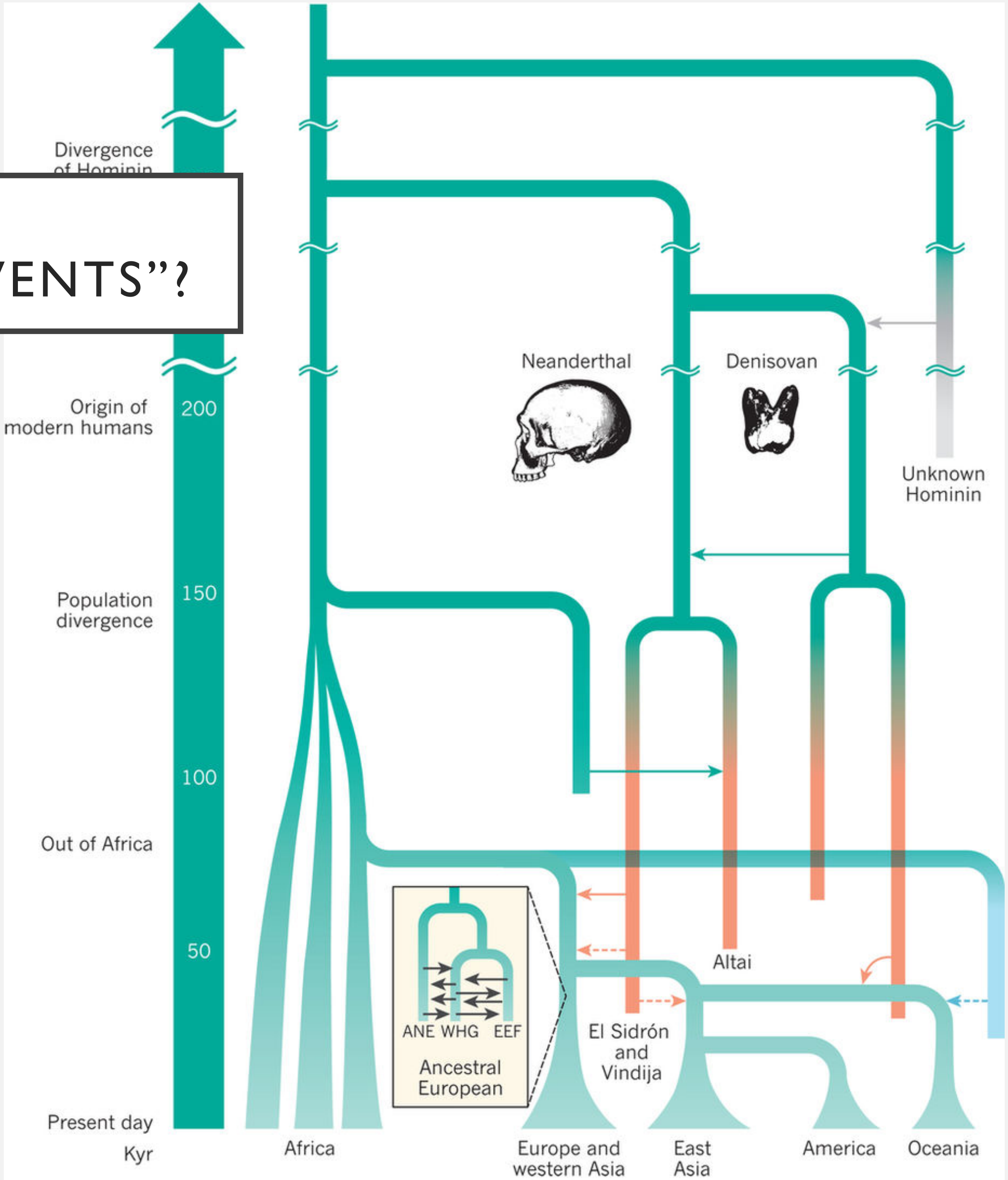
INFERENCE OF EVOLUTIONARY HISTORY WITH APPROXIMATE BAYESIAN COMPUTATION

Ariella Gladstein

Ecology and Evolutionary Biology

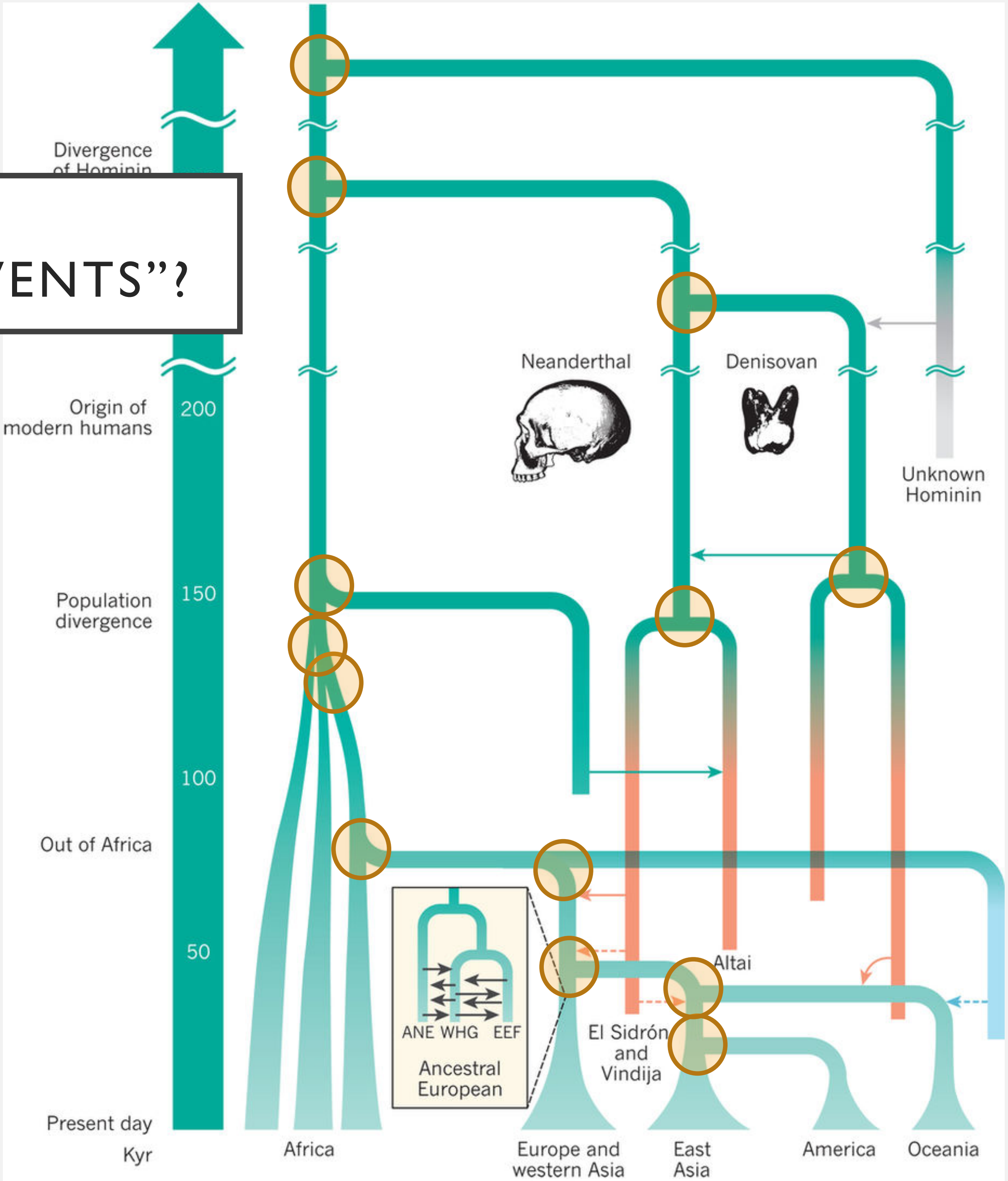
University of Arizona

WHAT ARE "DEMOGRAPHIC EVENTS"?



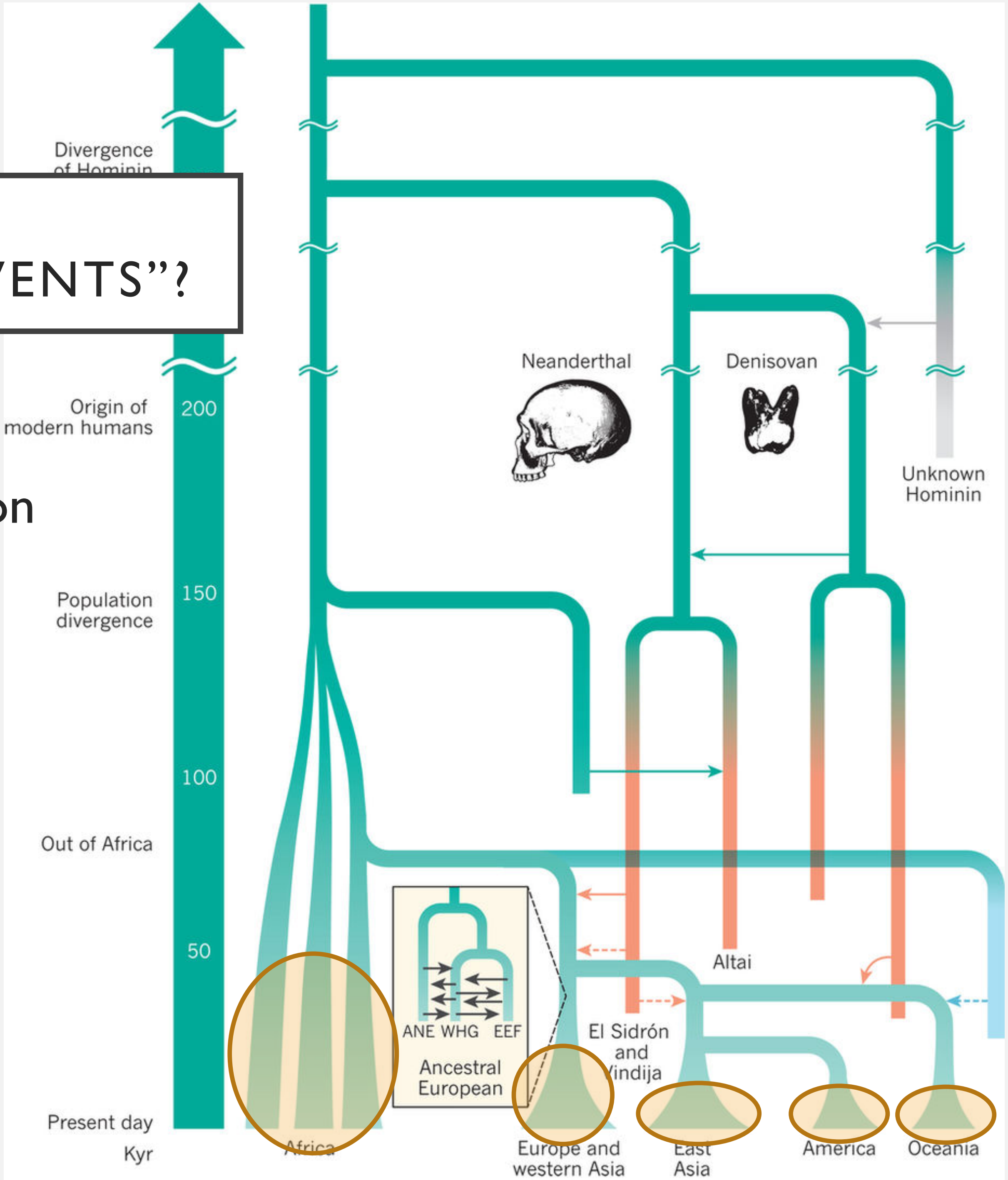
WHAT ARE "DEMOGRAPHIC EVENTS"?

- Divergence



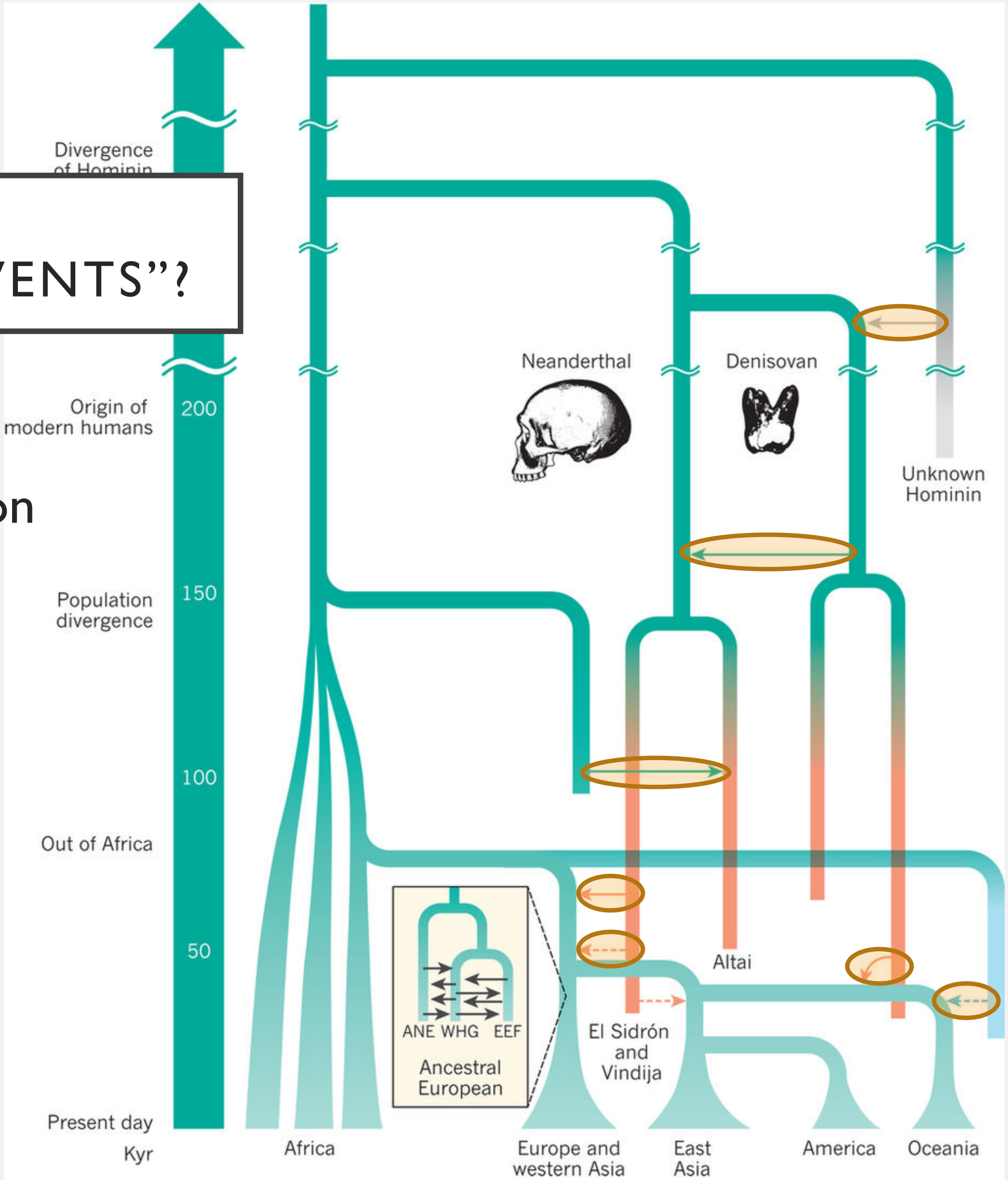
WHAT ARE "DEMOGRAPHIC EVENTS"?

- Divergence
- Expansion or reduction

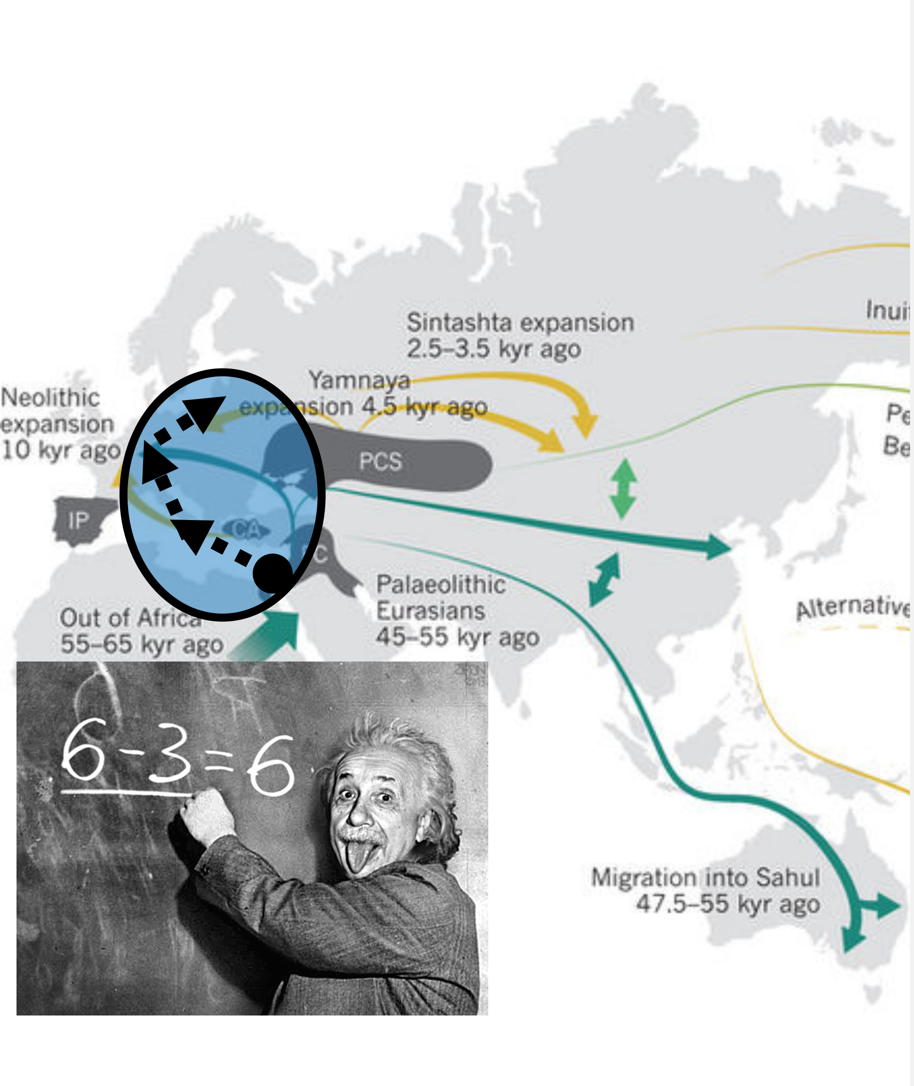


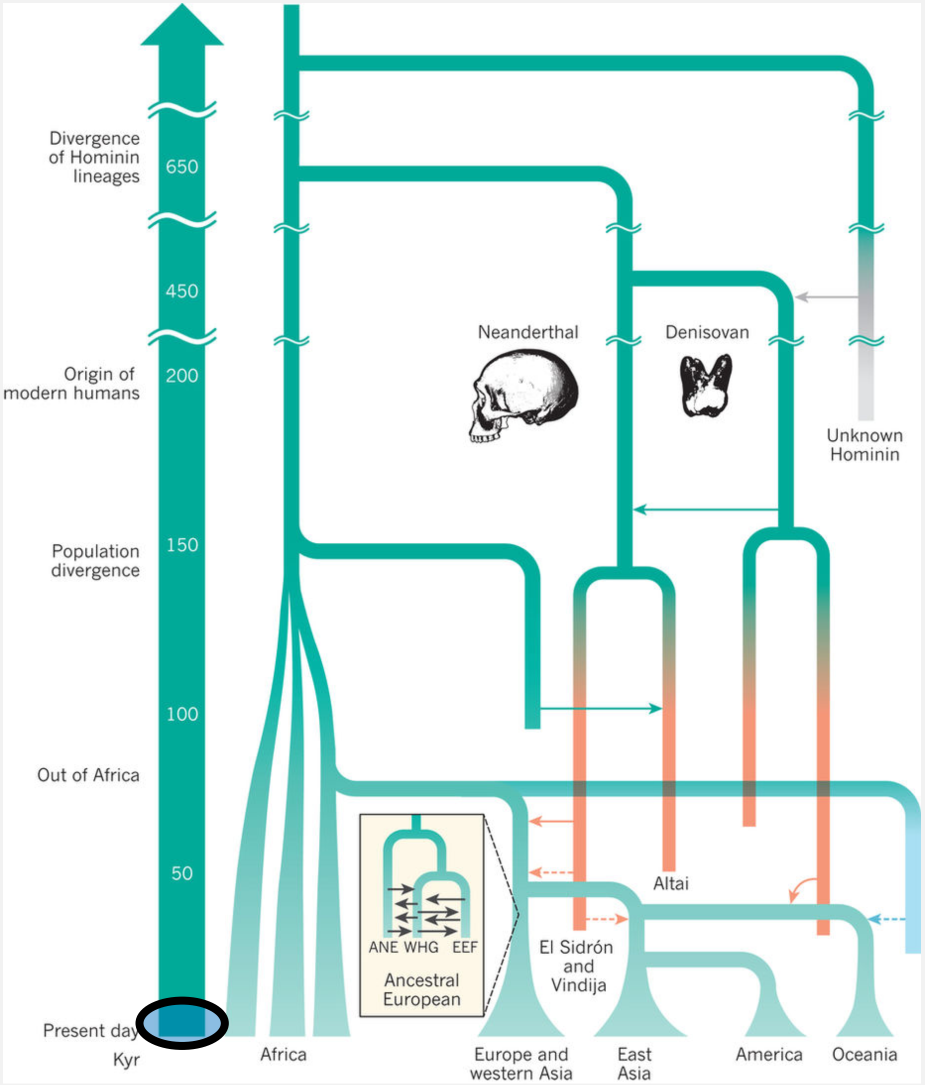
WHAT ARE “DEMOGRAPHIC EVENTS”?

- Divergence
- Expansion or reduction
- Gene flow

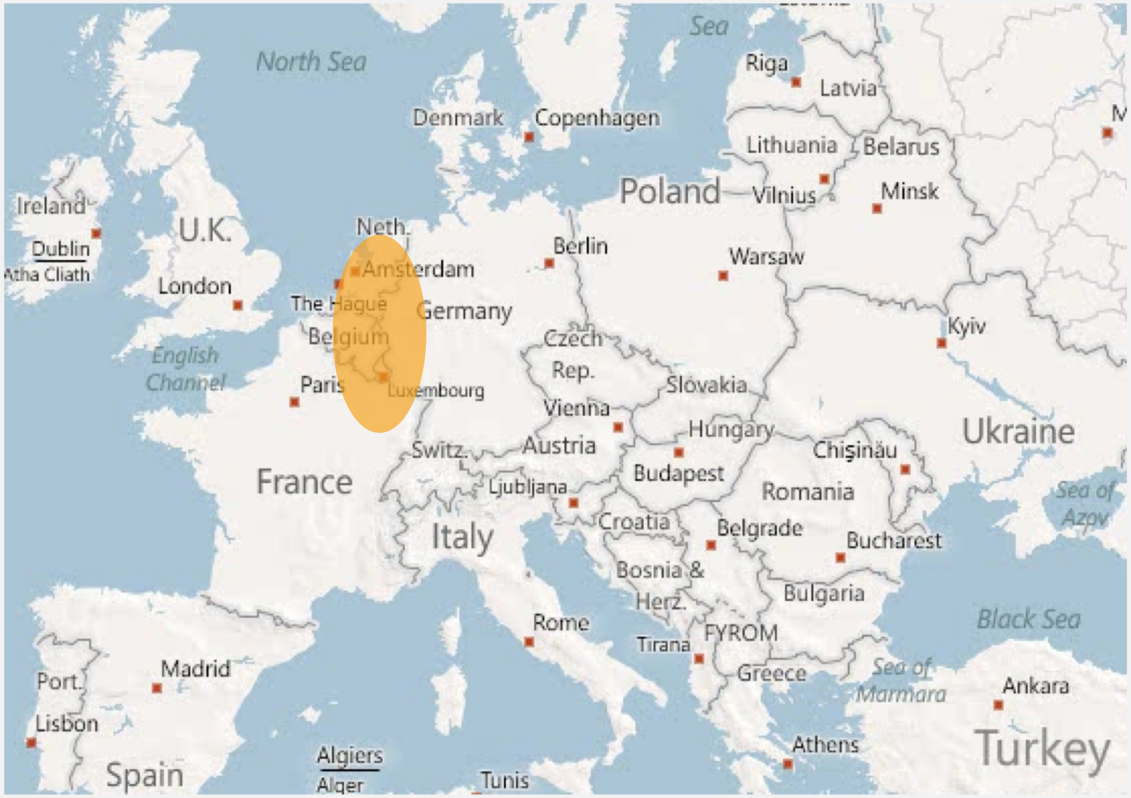


AIM: INFER THE DEMOGRAPHIC HISTORY OF THE ASHKENAZI JEWS.



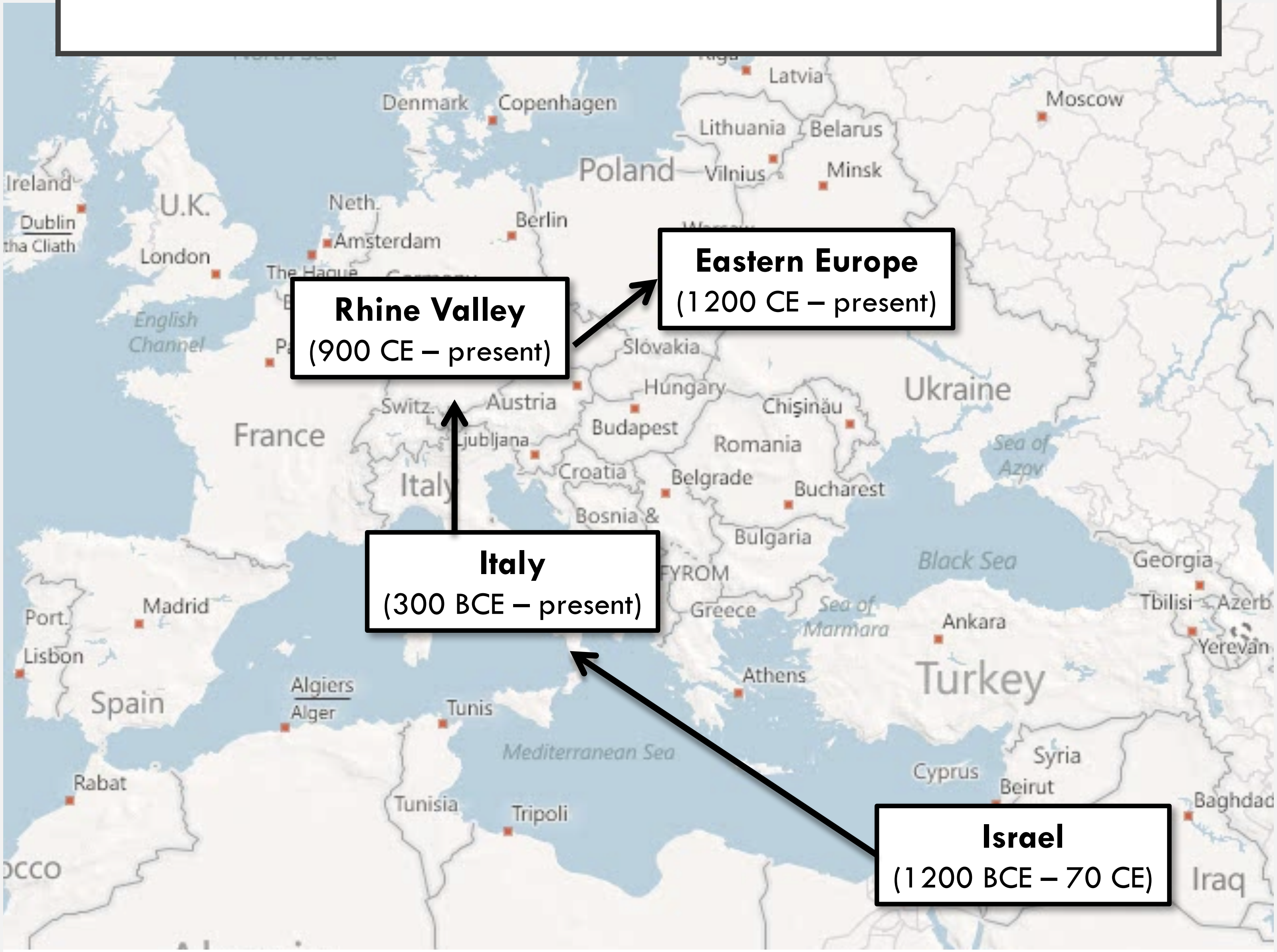


ASHKENAZI JEWS: AN INTERESTING STUDY POPULATION



The Ashkenazi Jews are a group that culturally, religiously, and linguistically identify as Jews whose ancestors came from the Rhine Valley.

HYPOTHESIS OF ASHKENAZI ORIGINS



Rhine Valley
(900 CE – present)

Eastern Europe
(1200 CE – present)

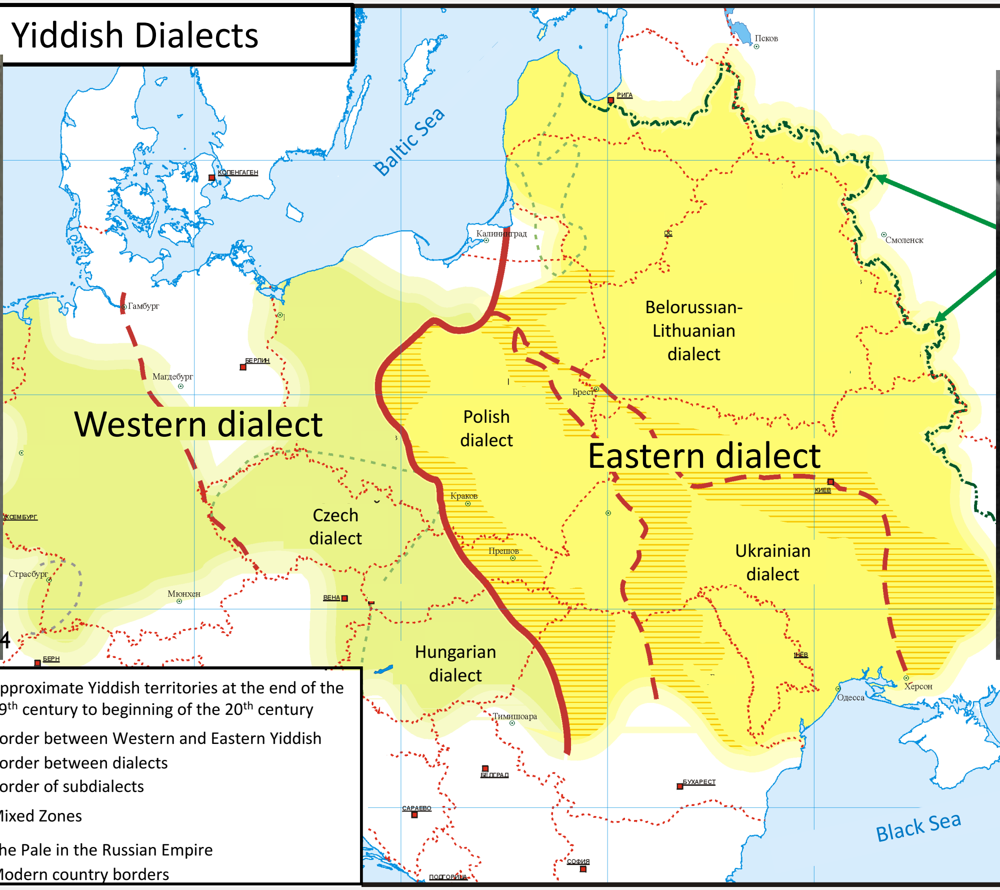
Italy
(300 BCE – present)

Israel
(1200 BCE – 70 CE)

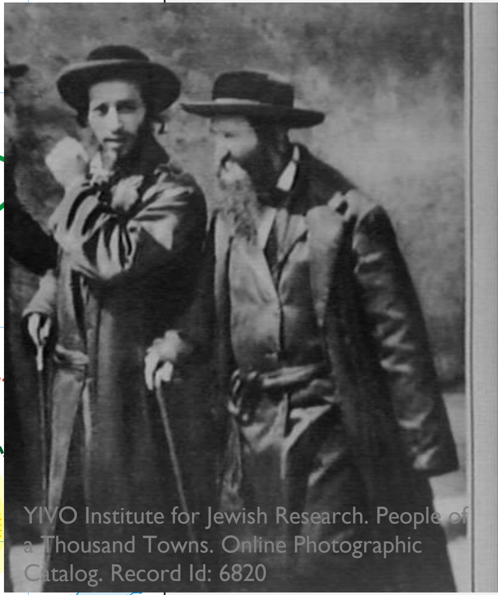
WESTERN VS. EASTERN ASHKENAZI JEWS

JDC Archives. Reference Code: NY_02044

Germany, 1900's



- 🟨 Approximate Yiddish territories at the end of the 19th century to beginning of the 20th century
- ━ Border between Western and Eastern Yiddish
- ╍ Border between dialects
- ┈ Border of subdialects
- ▤ Mixed Zones
- ─·─ The Pale in the Russian Empire
- ┈ Modern country borders

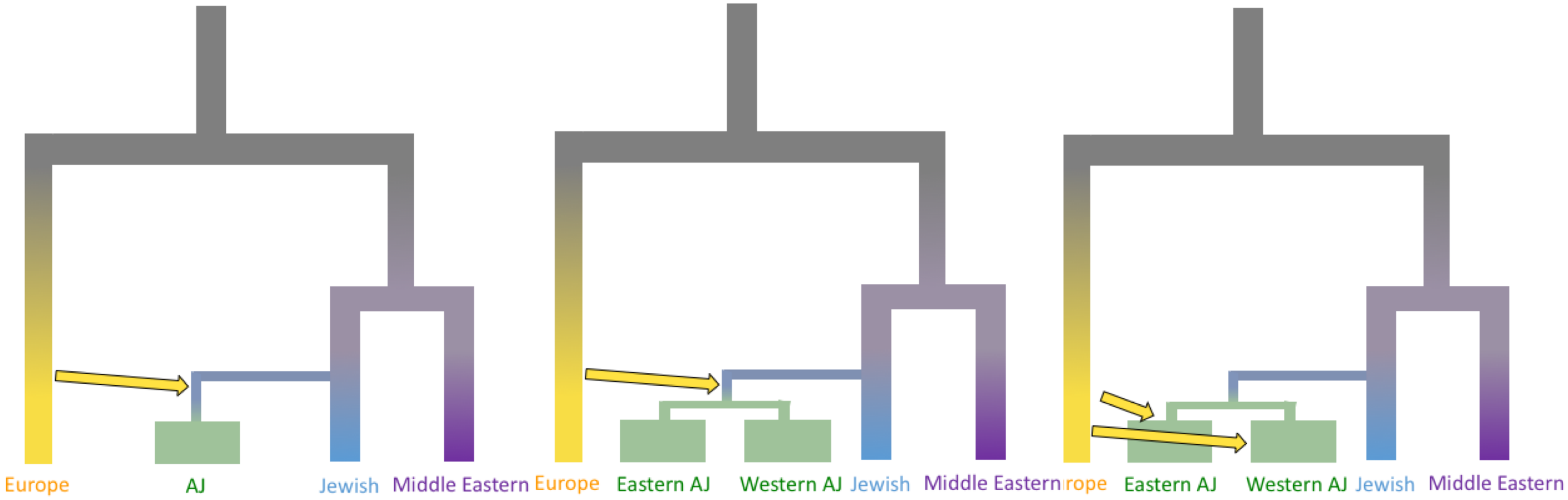


YIVO Institute for Jewish Research. People of a Thousand Towns. Online Photographic Catalog. Record Id: 6820

Cracow, Poland. 1932

MOTIVATION

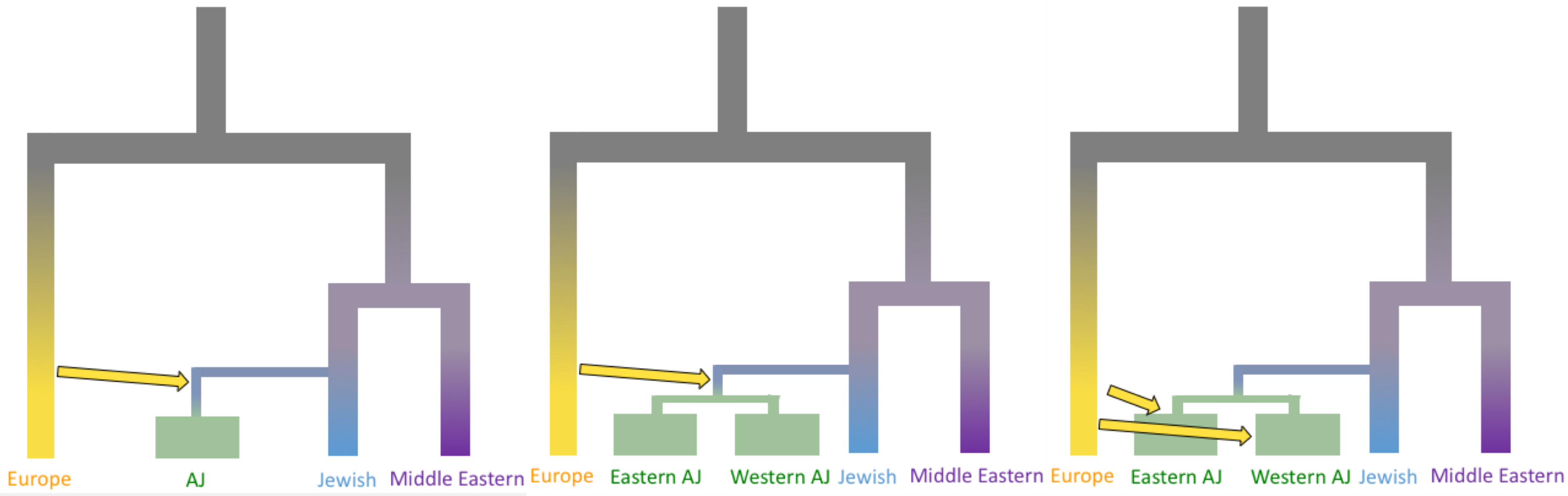
- Numerous genetic studies on the Ashkenazi Jews.
  - All genome-wide studies treat Ashkenazi Jews as one population.
- Preliminary work consistent with genetic differentiation.
  - Not informative of cause of differentiation.



MODELS OF ASHKENAZI HISTORY

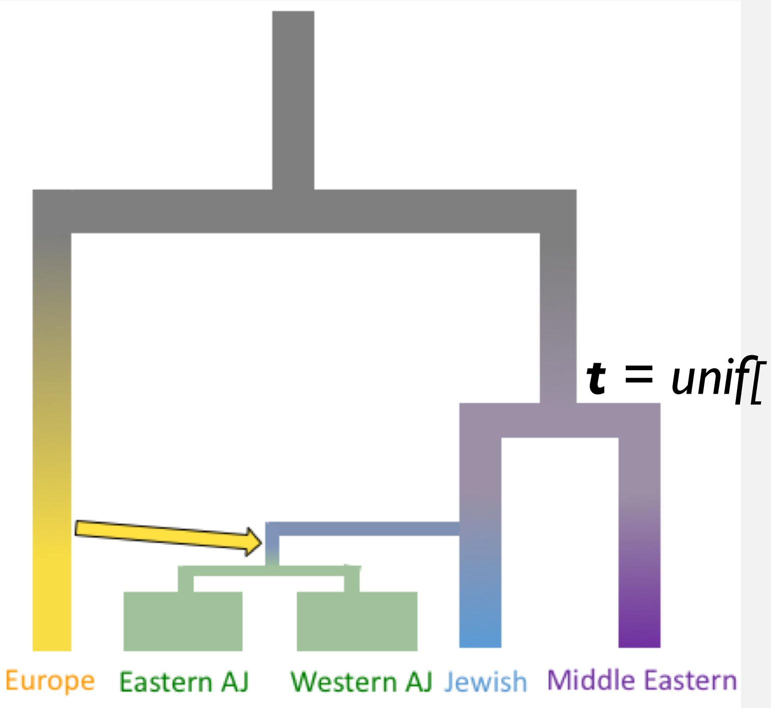
APPROXIMATE BAYESIAN COMPUTATION

- Infer parameter values
- Choose among models



APPROXIMATE BAYESIAN COMPUTATION

1. Define priors of parameters of model



t = *unif[10:1000]*

t = time (generations) of divergence between Jewish and Middle Eastern populations

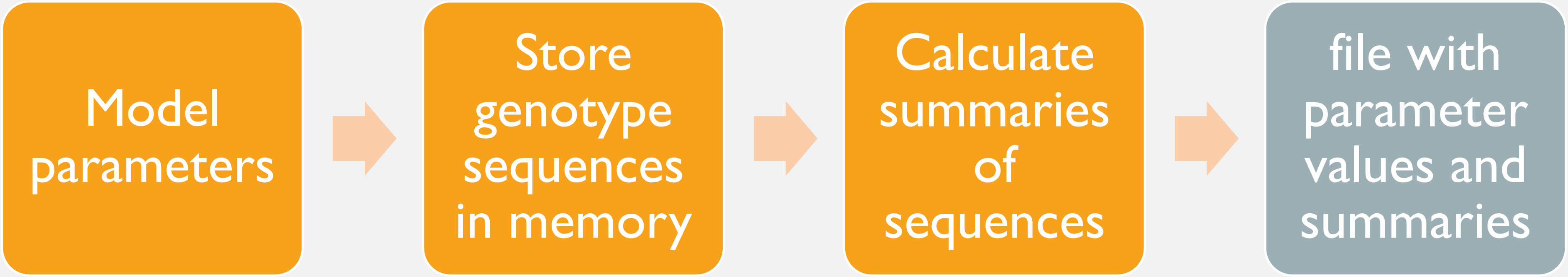
APPROXIMATE BAYESIAN COMPUTATION

1. Define priors of parameters of model
2. Simulate data many times

APPROXIMATE BAYESIAN COMPUTATION

1. Define priors of parameters of model
2. Simulate data many times
3. Choose model and estimate parameters based on simulations closest to real data

SIMULATION



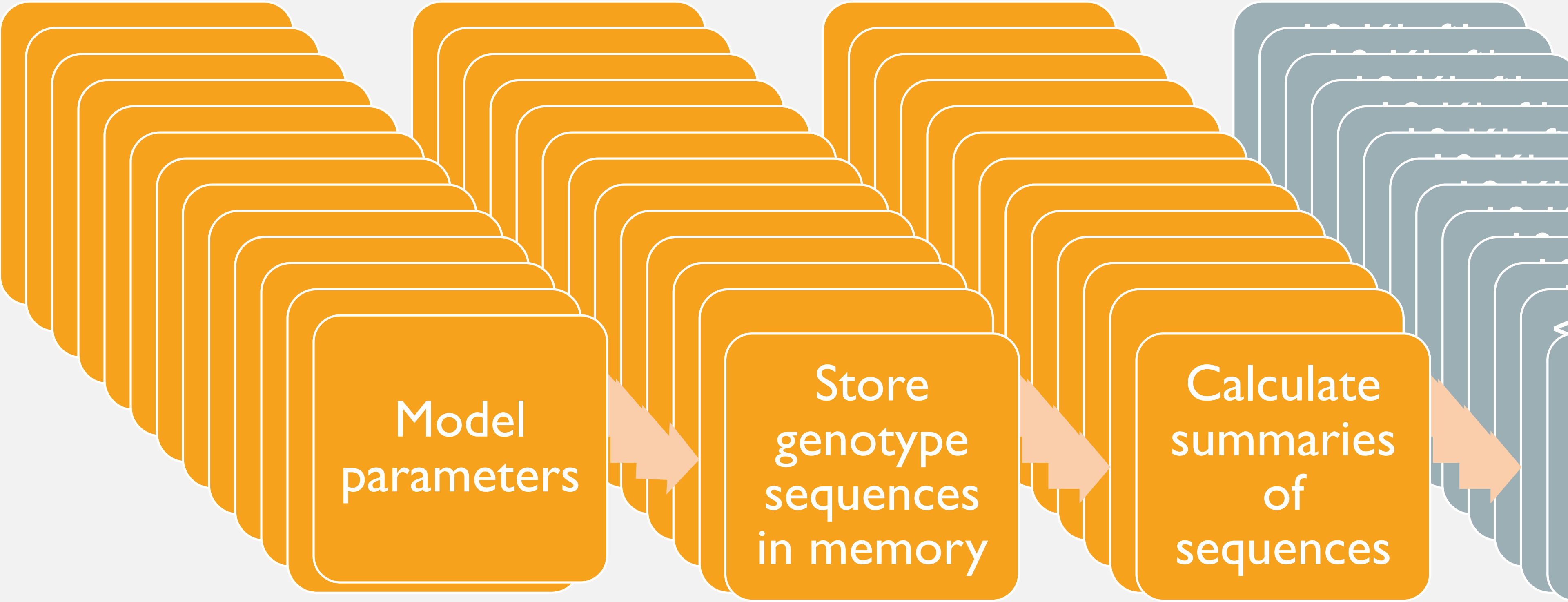
Model parameters

Store genotype sequences in memory

Calculate summaries of sequences

file with parameter values and summaries

PLEASANTLY PARALLEL!

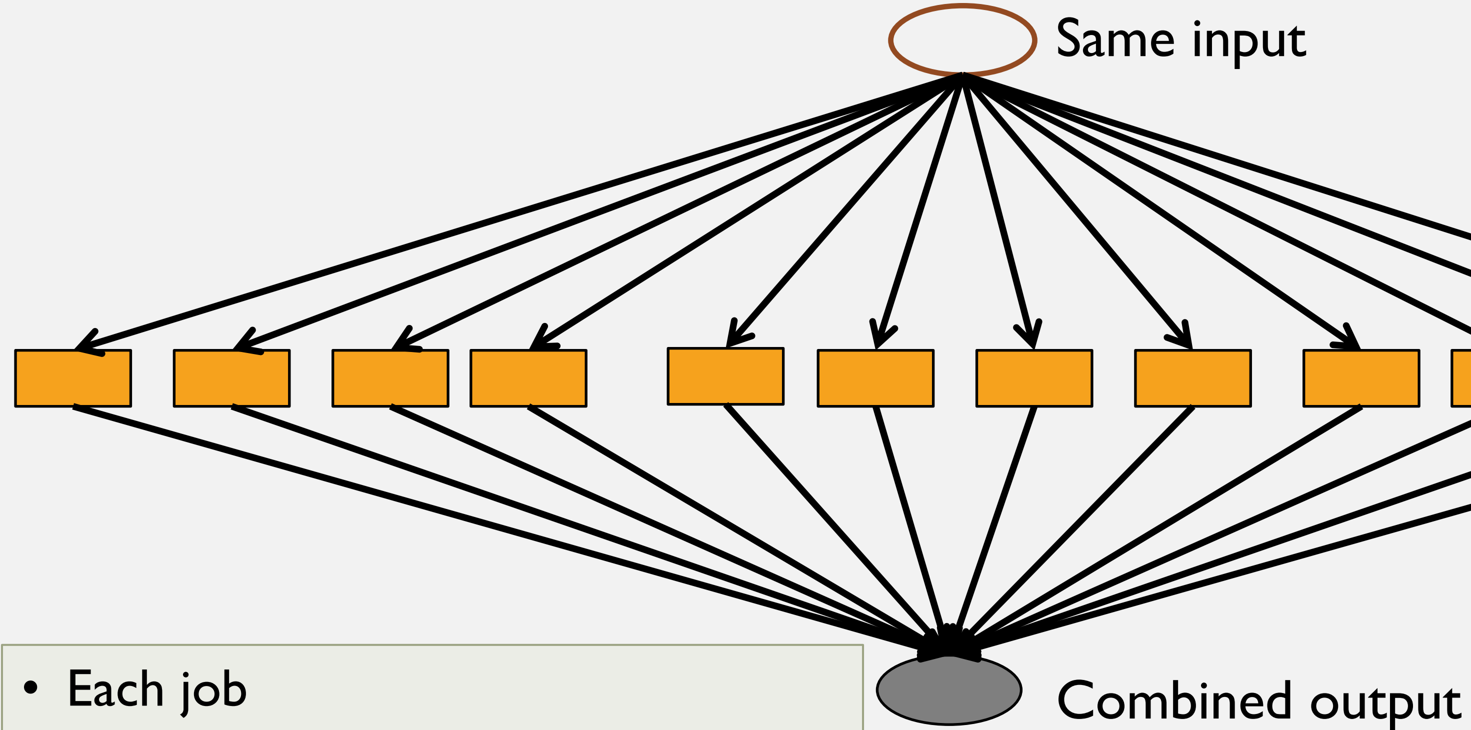


Model parameters

Store genotype sequences in memory

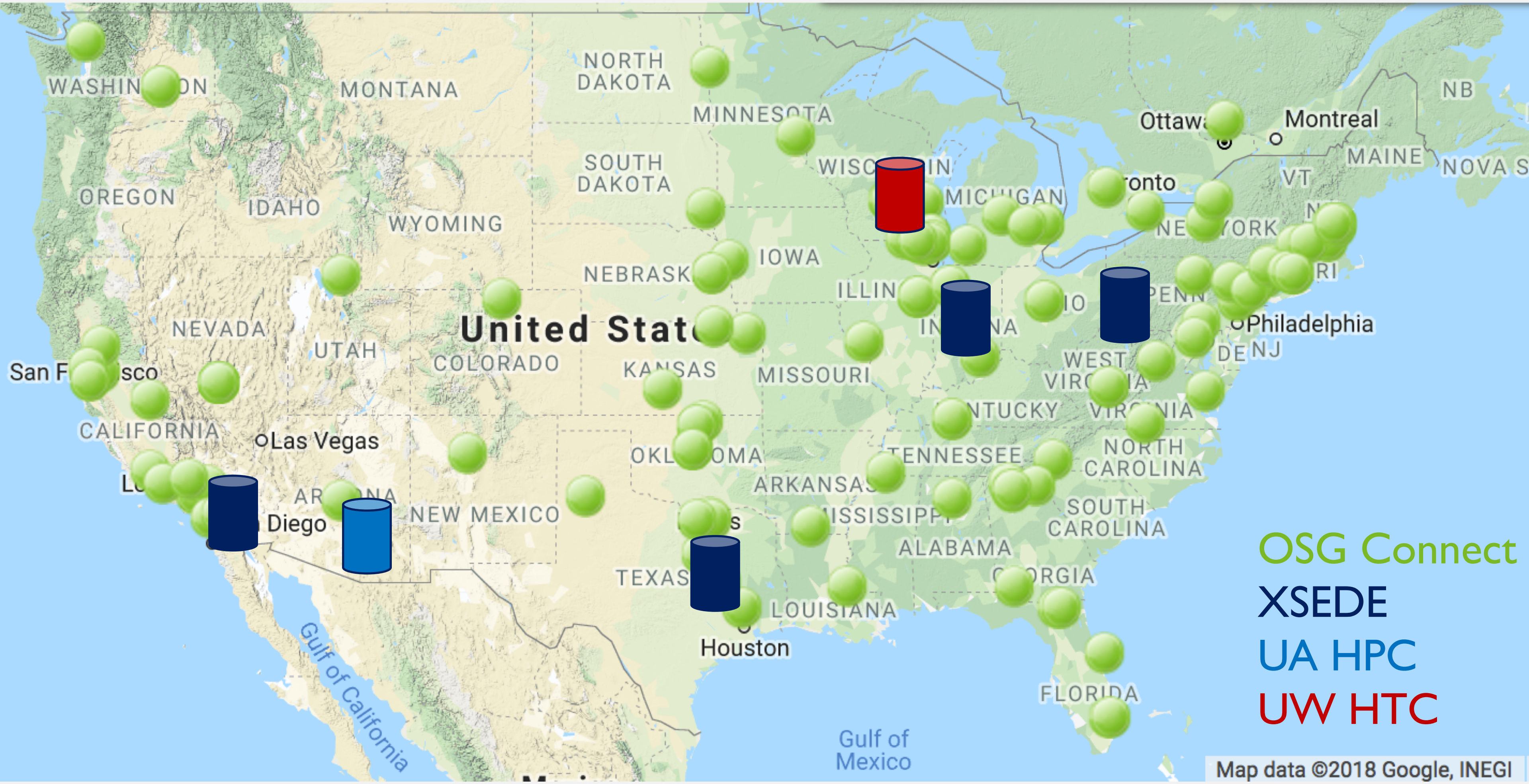
Calculate summaries of sequences

PLEASANTLY PARALLEL & RESOURCE LIGHT!



- Each job
  - runs ~40 min, and max 50 hrs
  - Uses ~1G, and max 5G memory
  - Uses ~2M in storage

HIGH THROUGHPUT COMPUTING



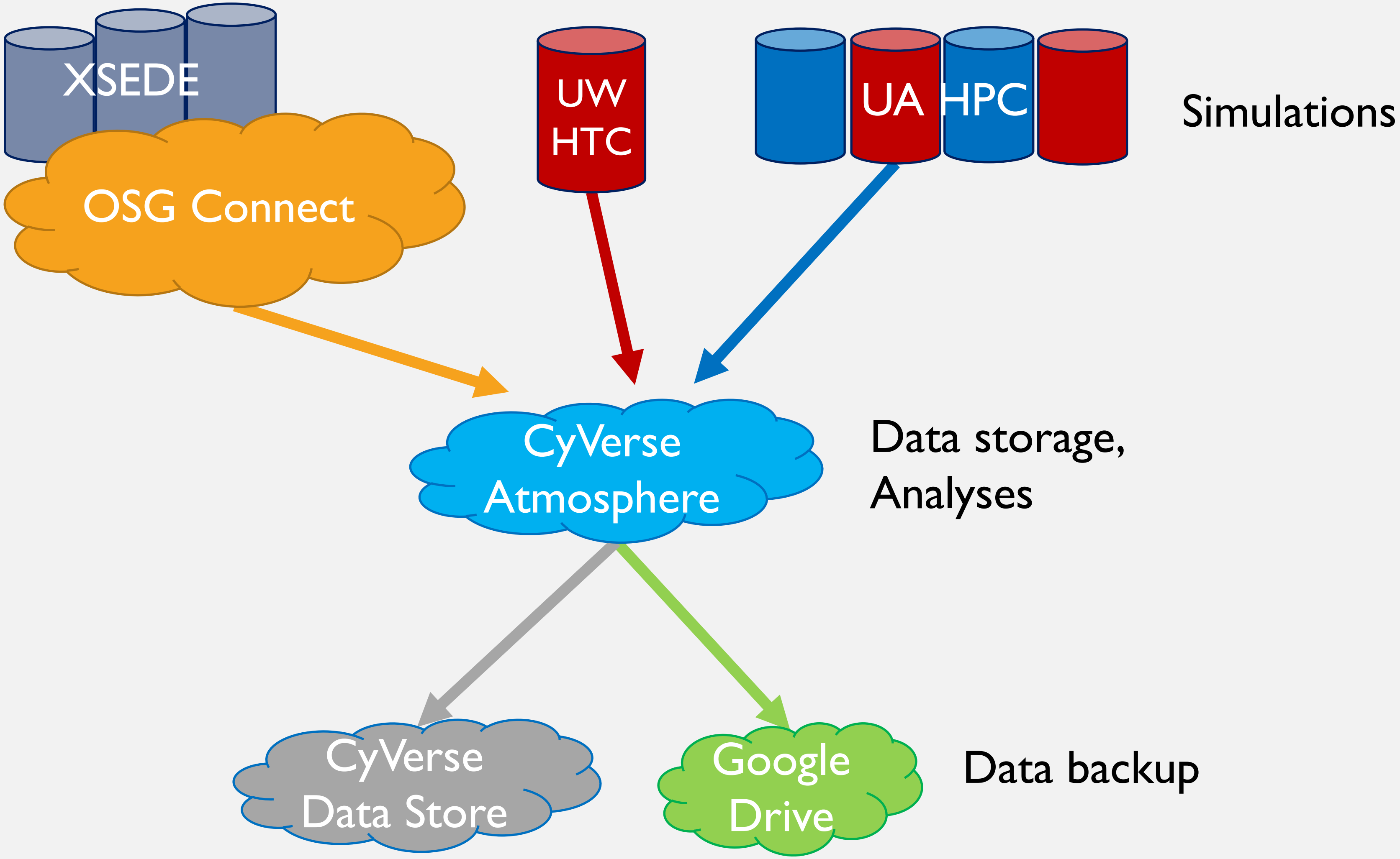
OSG Connect

XSEDE

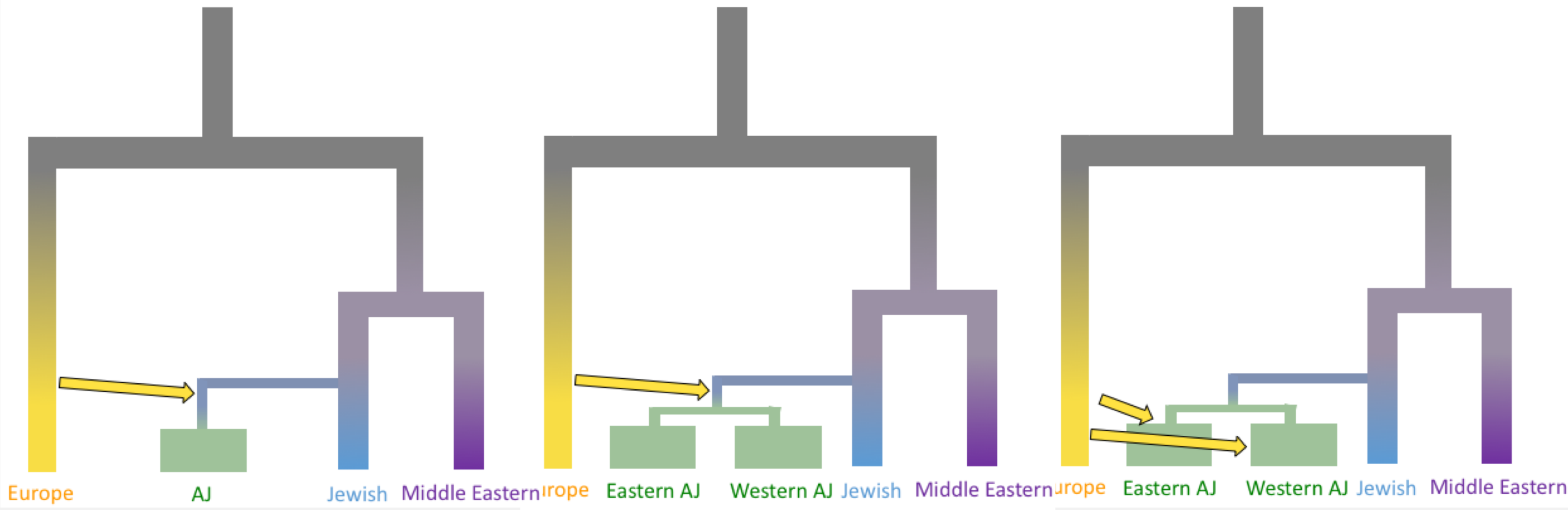
UA HPC

UW HTC

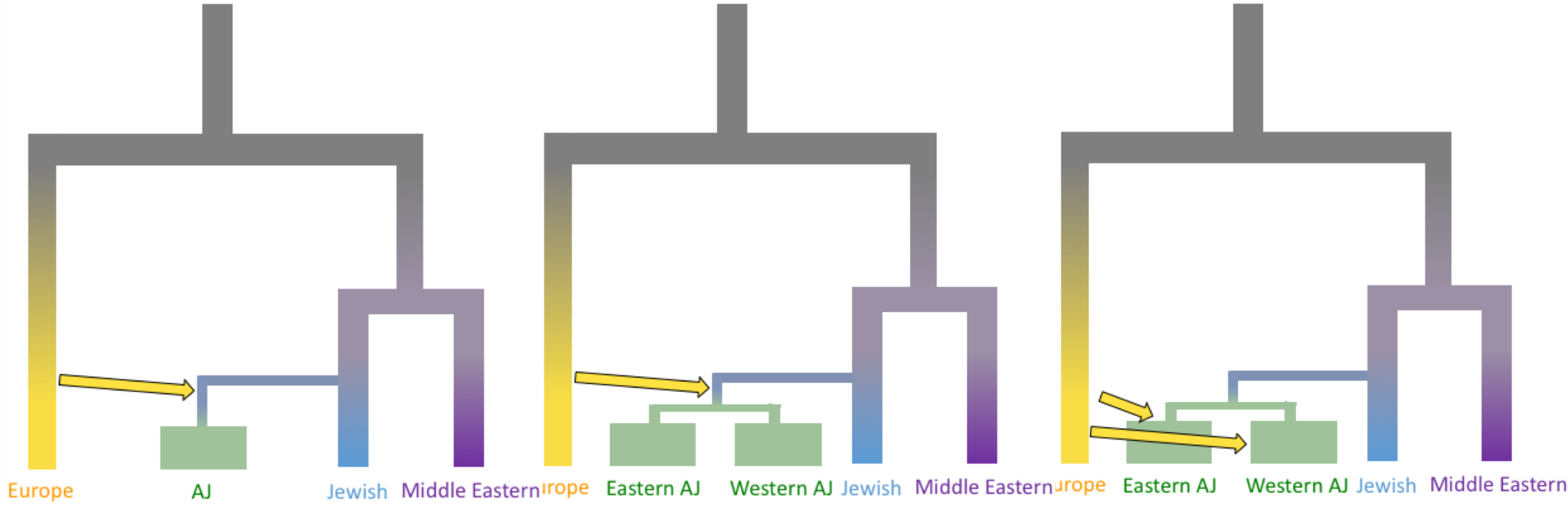
SIMULATIONS ON HTC CLUSTERS, ANALYSES ON VIRTUAL MACHINE



>1 MILLION SIMULATIONS OF EACH MODEL



MODEL CHOICE

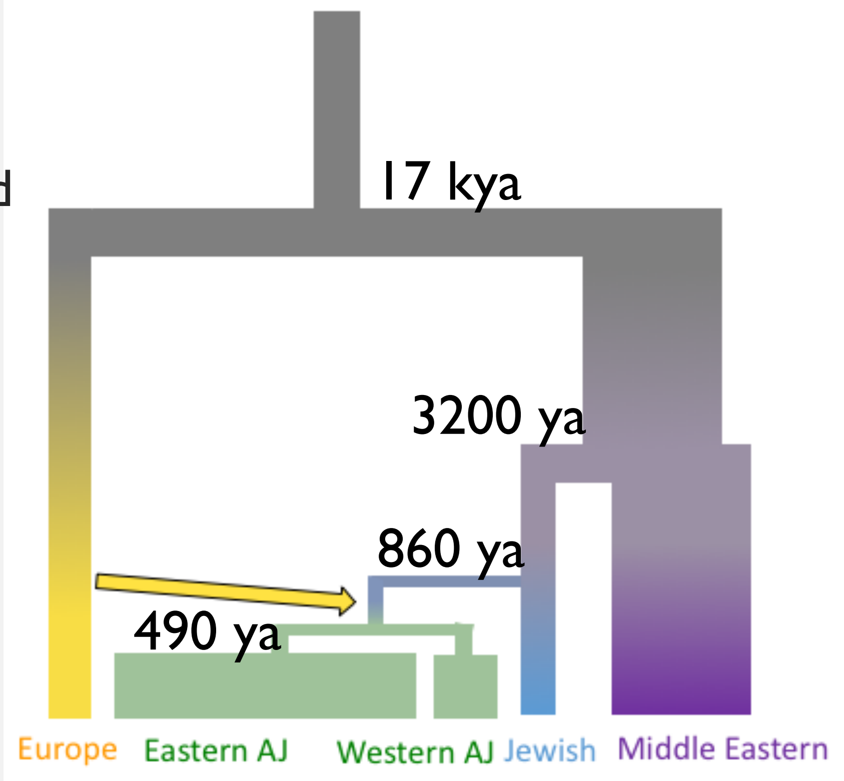


Posterior probability: 0.0065 0.85 0.14

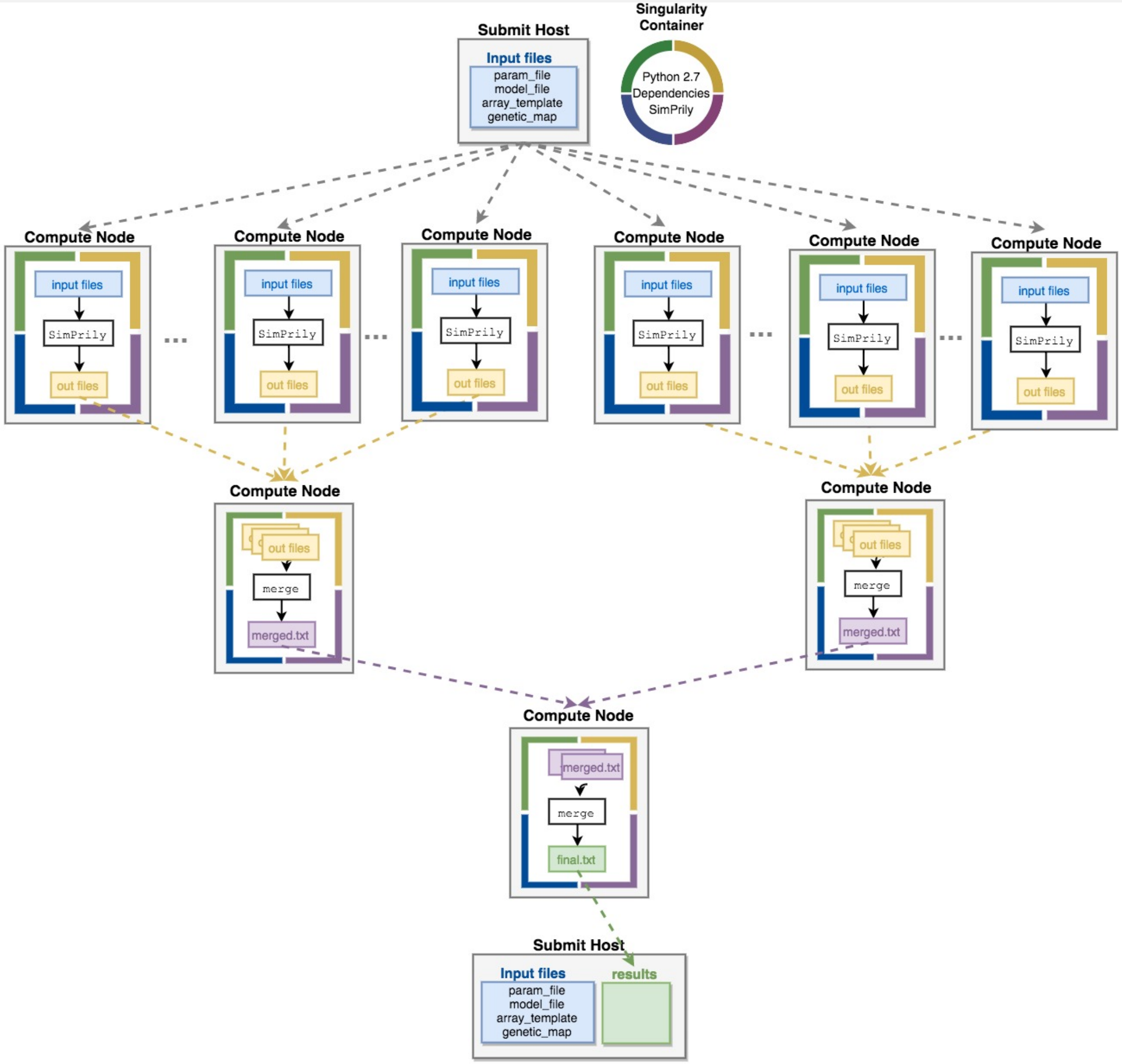
BEST MODEL

- ~1200 BCE ancestors of Jewish populations diverged from other Middle Eastern populations
  - Experienced extreme population size reduction
- ~1100 CE ancestors of Ashkenazi Jews diverged from other Jewish populations
  - Experienced another population size reduction
  - Experienced gene flow from Europeans

    (unresolved how much or when)
- ~1500 CE Eastern and Western Ashkenazi Jews diverged
  - Western AJ moderately grew in size
  - Eastern AJ massively grew in size



SIMPRILY: GENERALIZATION OF CODE AND WORKFLOW



- Developed program to simulate any demographic model
  - Memory & space efficient
- Use Singularity container
- Pegasus workflow for OSG

https://agladstein.github.io/SimPrily/

WHAT ARE THE CHALLENGES?

WHAT ARE THE CHALLENGES?

- How to be confident there are no bugs?
- How to maintain a consistent run environment?
- How to handle millions of files?
  - UA HPC has file number limit
  - If there are too many files in a directory simple things take a long time
- How to not overload UA HPC system?
- How to reliably backup data?
- Why do jobs fail?

WHAT ARE THE CHALLENGES?

- How to be confident there are no bugs? Tests!
- How to maintain a consistent run environment?
- How to handle millions of files?
  - UA HPC has file number limit
  - If there are too many files in a directory simple things take a long time
- How to not overload UA HPC system?
- How to reliably backup data?
- Why do jobs fail?

WHAT ARE THE CHALLENGES?

- How to be confident there are no bugs?

Tests!

- How to maintain a consistent run environment?

Virtual environment or Container

- How to handle millions of files?
  - UA HPC has file number limit
  - If there are too many files in a directory simple things take a long time
- How to not overload UA HPC system?
- How to reliably backup data?
- Why do jobs fail?

WHAT ARE THE CHALLENGES?

- How to be confident there are no bugs? **Tests!**
- How to maintain a consistent run environment? **Virtual environment or Container**
- How to handle millions of files? **Buckets!**
  - UA HPC has file number limit
  - If there are too many files in a directory simple things take a long time
- How to not overload UA HPC system?
- How to reliably backup data?
- Why do jobs fail?

WHAT ARE THE CHALLENGES?

- How to be confident there are no bugs? Tests!
- How to maintain a consistent run environment? Virtual environment or Container
- How to handle millions of files? Buckets!
  - UA HPC has file number limit
  - If there are too many files in a directory simple things take a long time
- How to not overload UA HPC system? "Small" batches
- How to reliably backup data?
- Why do jobs fail?

WHAT ARE THE CHALLENGES?

- How to be confident there are no bugs? Tests!
- How to maintain a consistent run environment? Virtual environment or Container
- How to handle millions of files? Buckets!
  - UA HPC has file number limit
  - If there are too many files in a directory simple things take a long time
- How to not overload UA HPC system? "Small" batches
- How to reliably backup data? Automate!
- Why do jobs fail?

WHAT ARE THE CHALLENGES?

- How to be confident there are no bugs? Tests!
- How to maintain a consistent run environment? Virtual environment or Container
- How to handle millions of files? Buckets!
  - UA HPC has file number limit
  - If there are too many files in a directory simple things take a long time
- How to not overload UA HPC system? “Small” batches
- How to reliably backup data? Automate!
- Why do jobs fail? ¯_(ツ)_/¯

LINKS

- UA HPC Dashboard
  - https://ood.hpc.arizona.edu/pun/sys/dashboard
- UA HPC Allocation + Limits
  - https://docs.hpc.arizona.edu/display/UAHPC/Allocation+and+Limits
- Demo Repository
  - https://github.com/agladstein/ECOL-346-HPC-demo

THANK YOU!

HAMMER LAB

- Michael Hammer
- Consuelo Quinto-Cortes

CYVERSE

- Blake Joyce
- Julian Pistorius

UA HPC CONSULTING

- Mike Bruck
- Dima Shyshlov

OPEN SCIENCE GRID & PEGASUS

- Mats Rynge

UW CENTER FOR HTC

- Lauren Michael
- Christina Koch

OPEN SCIENCE GRID USER SCHOOL

- Tim Cartwright
- Lauren Michael
- Christina Koch

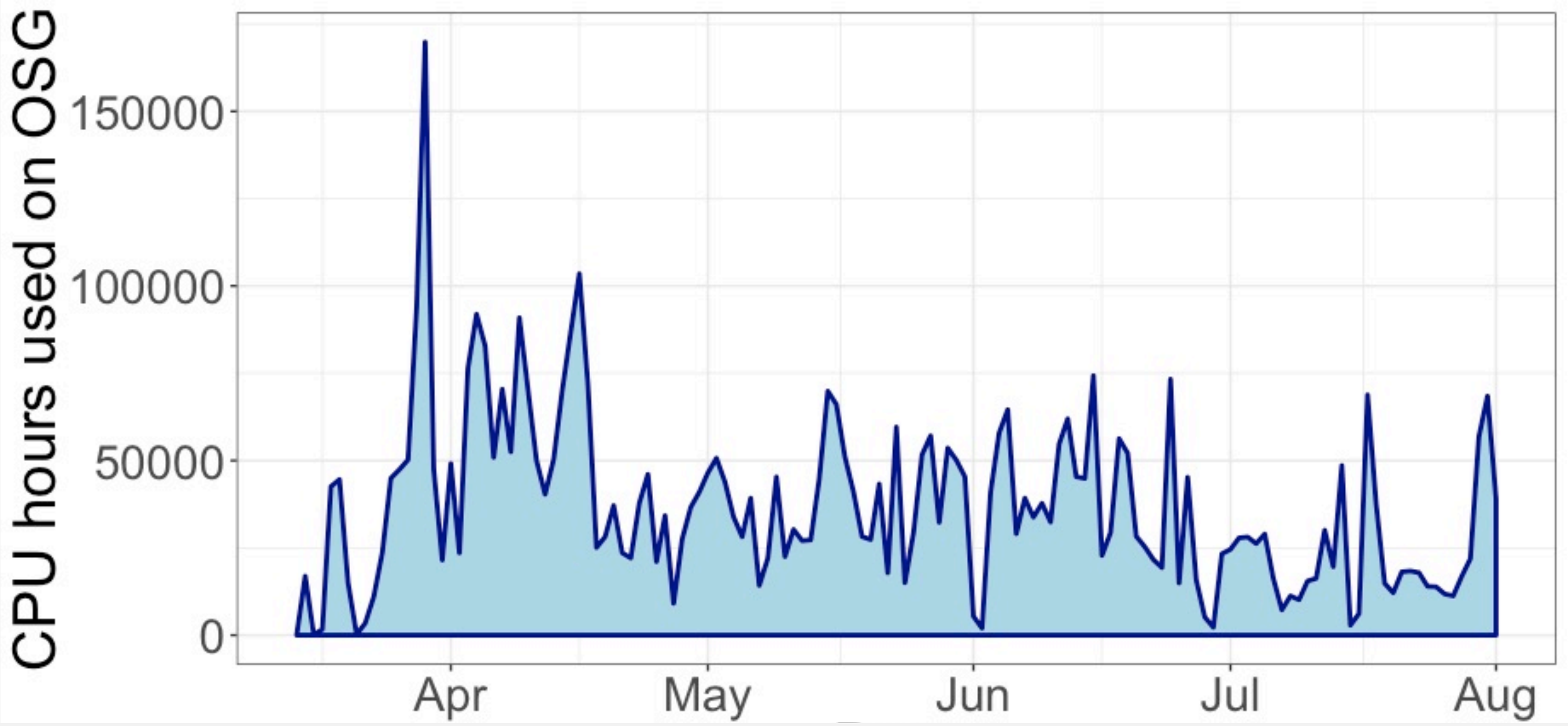
CODING MINIONS

- David Christy
- Logan Gantner
- Mack Skodiak
- Daniel Olson
- Rafael Lopez
- Kayleen Gurrola
- Katie McCready

RESOURCES PROVIDED BY

- University of Arizona HPC
- University of Wisconsin HTC
- CyVerse
- Open Science Grid
- XSEDE
  - Bridges
  - Comet
  - Jetstream

CPU HOURS ON THE OPEN SCIENCE GRID



DNA SEQUENCE

Indiv 1	AATCATTTCGGTTTTAATGCTTGGGCTGCATTGGGAAA AATCATATCGGTCTTAATGCTTGCGCTGCCTTGGTAAA

DNA SEQUENCE, SEGREGATING SITES

Indiv 1

AATCATTTCGGTTTTAATGCTTGGGCTGCATTGGGAAA

AATCATATCGGTCTTAATGCTTGCGCTGCCTTGGTAAA

DNA SEQUENCE, SEGREGATING SITES

Indiv 1

AATCATTTCGGTTTTAATGCTTGGGCTGCATTGGGAAA
AATCATATCGGTCTTAATGCTTGCGCTGCCTTGGTAAA

Indiv 2

AATCATTTCGGTTTTAATGCTTGGGCTGCCTTGGTAAA
AAACATTTCCGTCTTTTATGGTTGCGCTGCATTGGGGAA

DNA SEQUENCE, GENOTYPES ENCODED 0/1

Indiv 1

AATCATTTCGGTTTTAATGCTTGGGCTGCATTGGGAAA
AATCATATCGGTCTTAATGCTTGCGCTGCCTTGGTAAA

Indiv 1

00000000000000000000000100000000000001000
00000010000010000000000000000010000000000

Indiv 2

AATCATTTCGGTTTTAATGCTTGGGCTGCCTTGGTAAA
AAACATTTCCGTCTTTATGGTTGCGCTGCATTGGGGAA

Indiv 2

00000000000000000000000000000100000000
00100000010010010001000100000000001100

SEQUENCE OF GENOTYPES, ONLY SEGREGATING SITES

Indiv 1

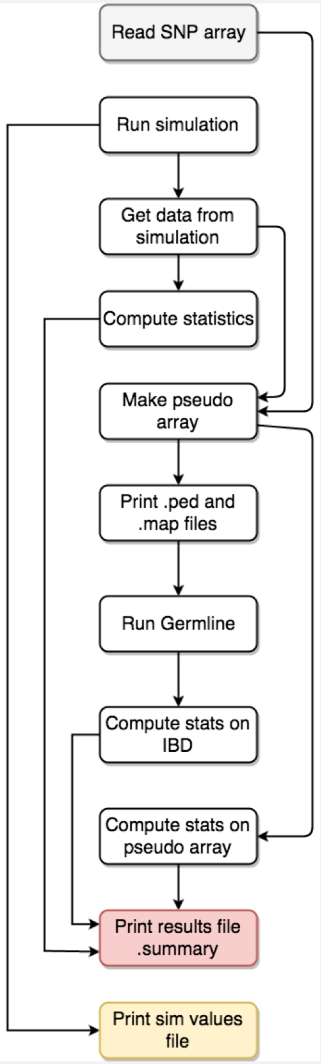
0000001010
0101000100

Indiv 2

0000000100
1011111011

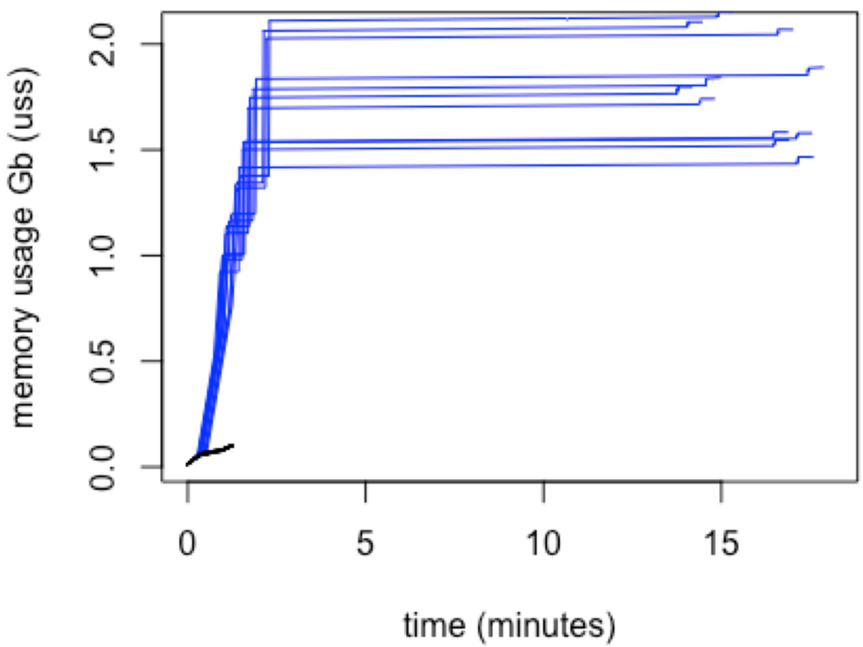
PYTHON SCRIPT: GENOME SIMULATIONS AND COMPUTE SUMMARY STATISTICS

- Inherited from lab mates
- Intended for millions of relatively small simulations
  - 1,389 10kb regions
  - 65 individuals
- Originally took a few minutes to run
- Originally ran parallel on U of A HPC
  - 1 million runs would take approximately 1 month.

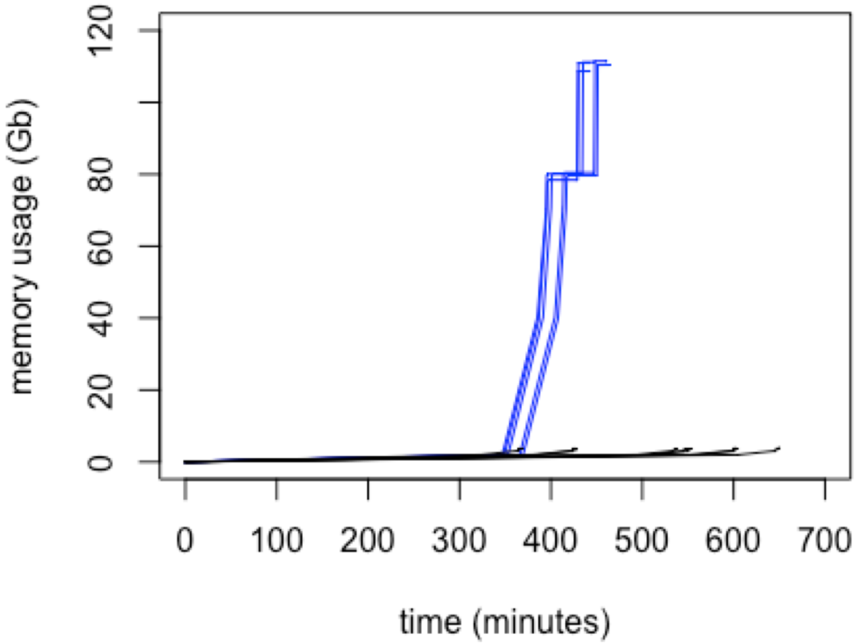


PROFILE OF PYTHON SCRIPT

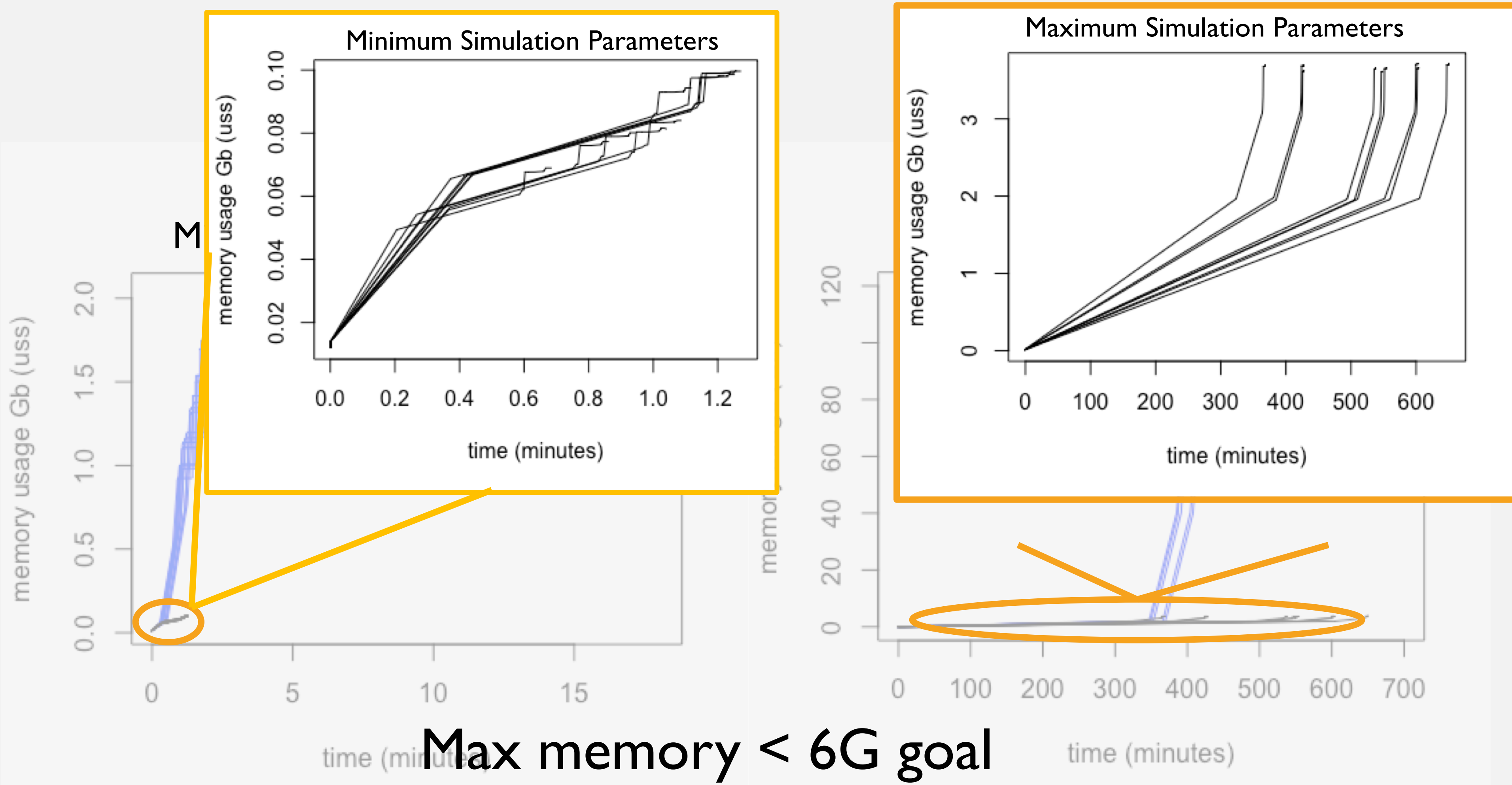
Minimum Simulation Parameters



Maximum Simulation Parameters



Note different scales



Max memory < 6G goal

Can now run efficiently in parallel

**Note different scales*