INFERENCE OF EVOLUTIONARY HISTORY WITH APPROXIMATE BAYESIAN COMPUTATION

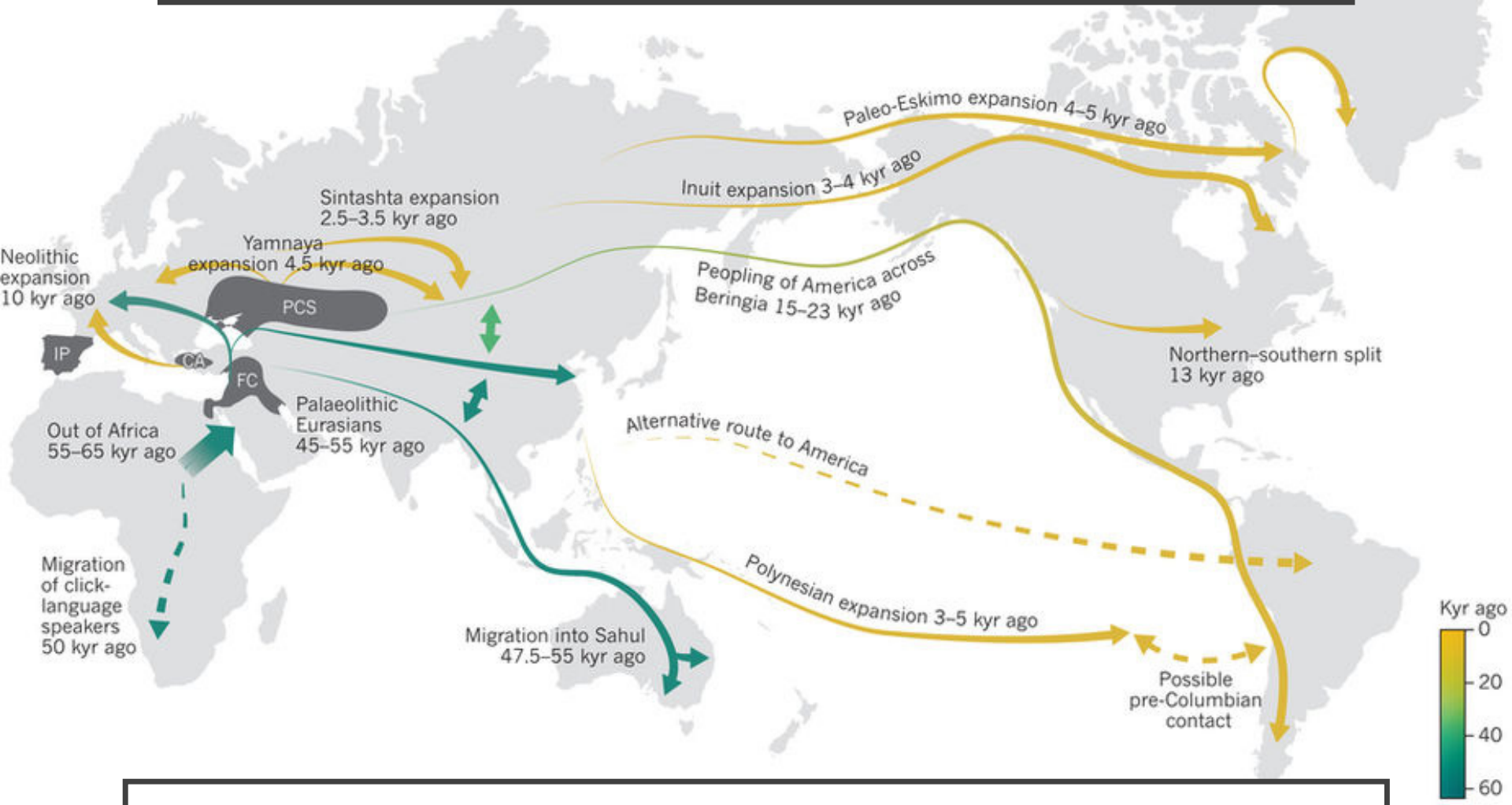
Ariella Gladstein

Ecology and Evolutionary Biology

University of Arizona

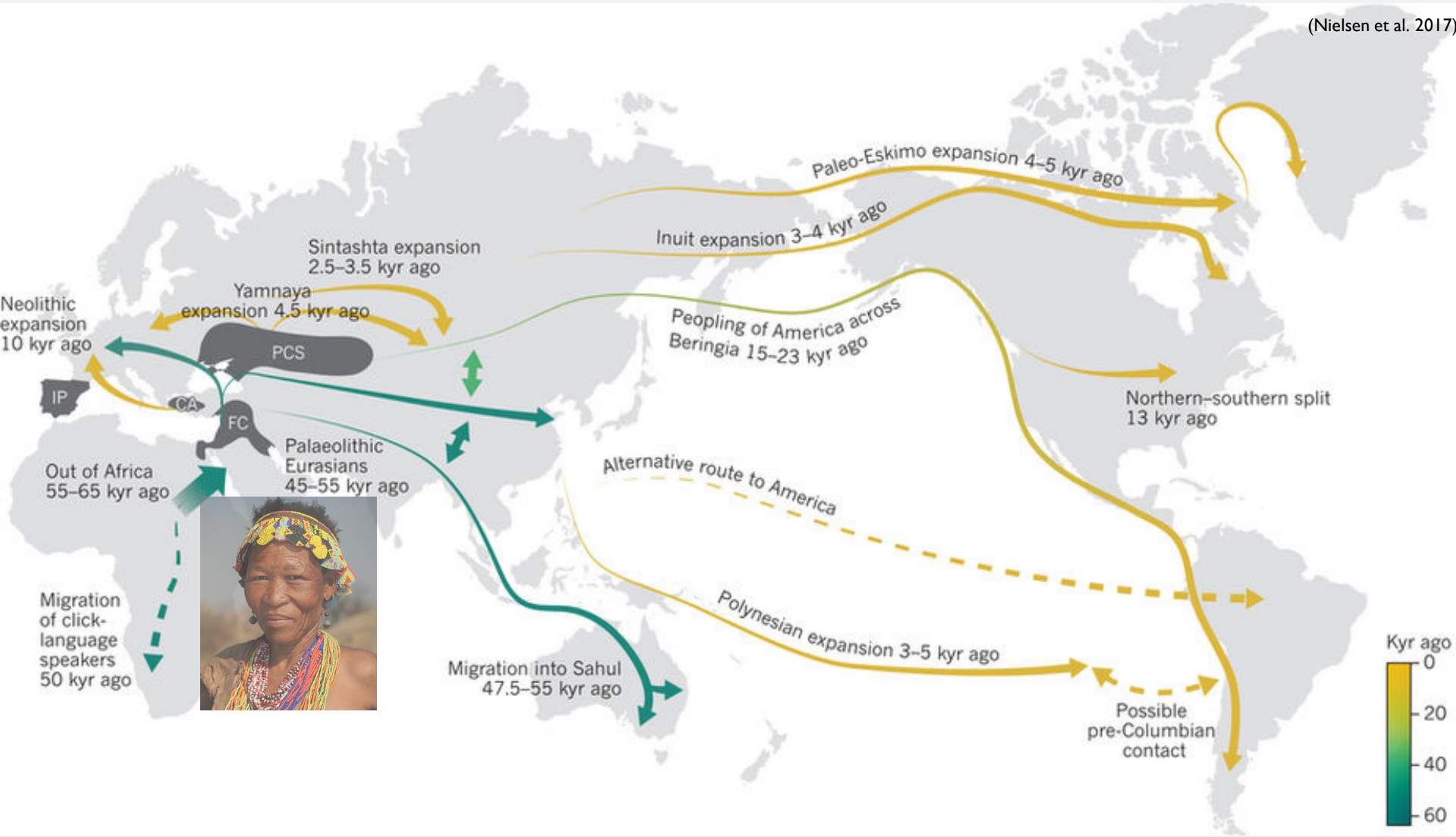
HOW DID HUMANS SPREAD ACROSS THE WORLD?

(Nielsen et al. 2017)

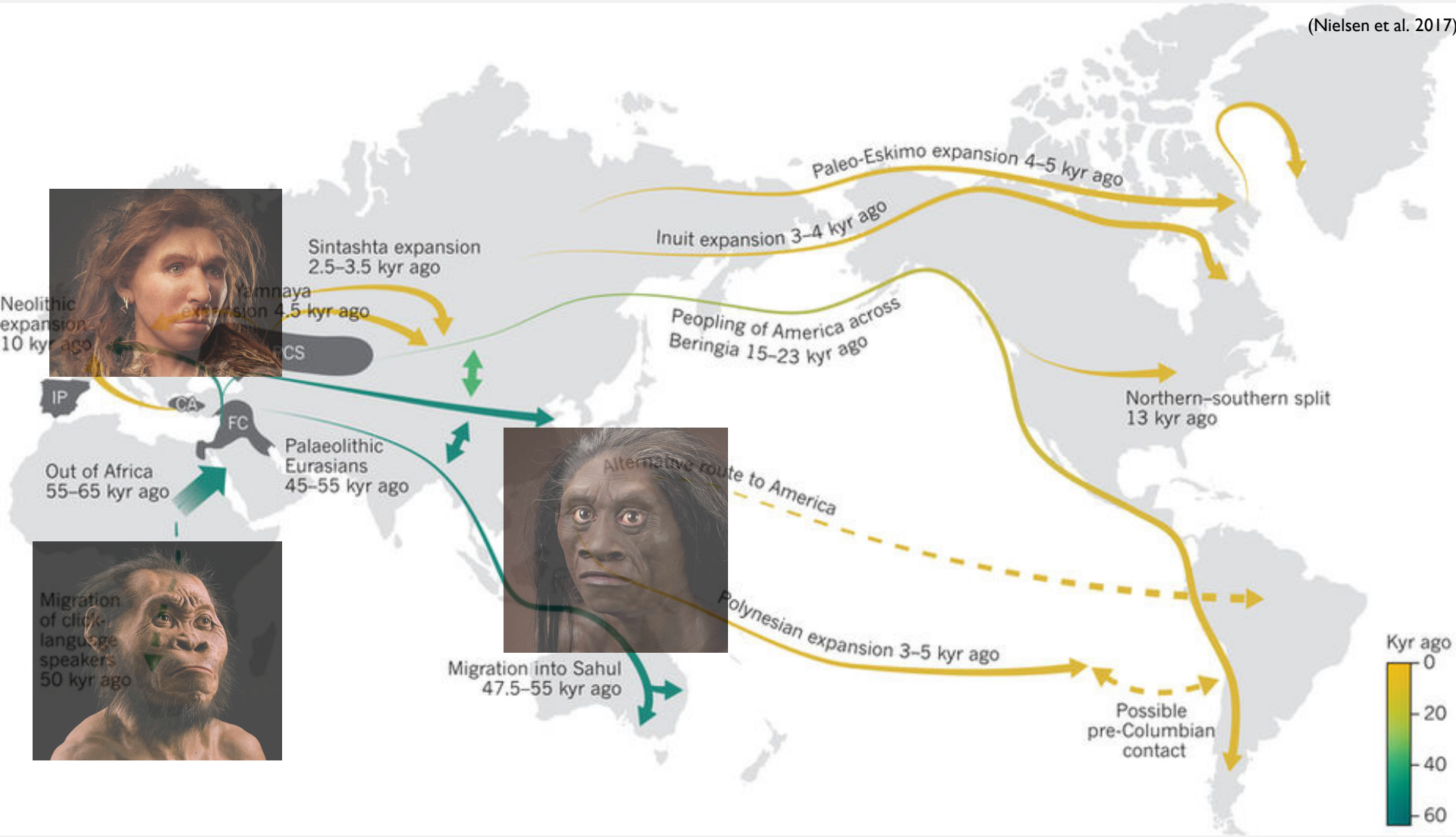


WHAT DEMOGRAPHIC EVENTS LEAD US TO WHERE WE ARE TODAY AND THE DIVERSITY WE SEE?

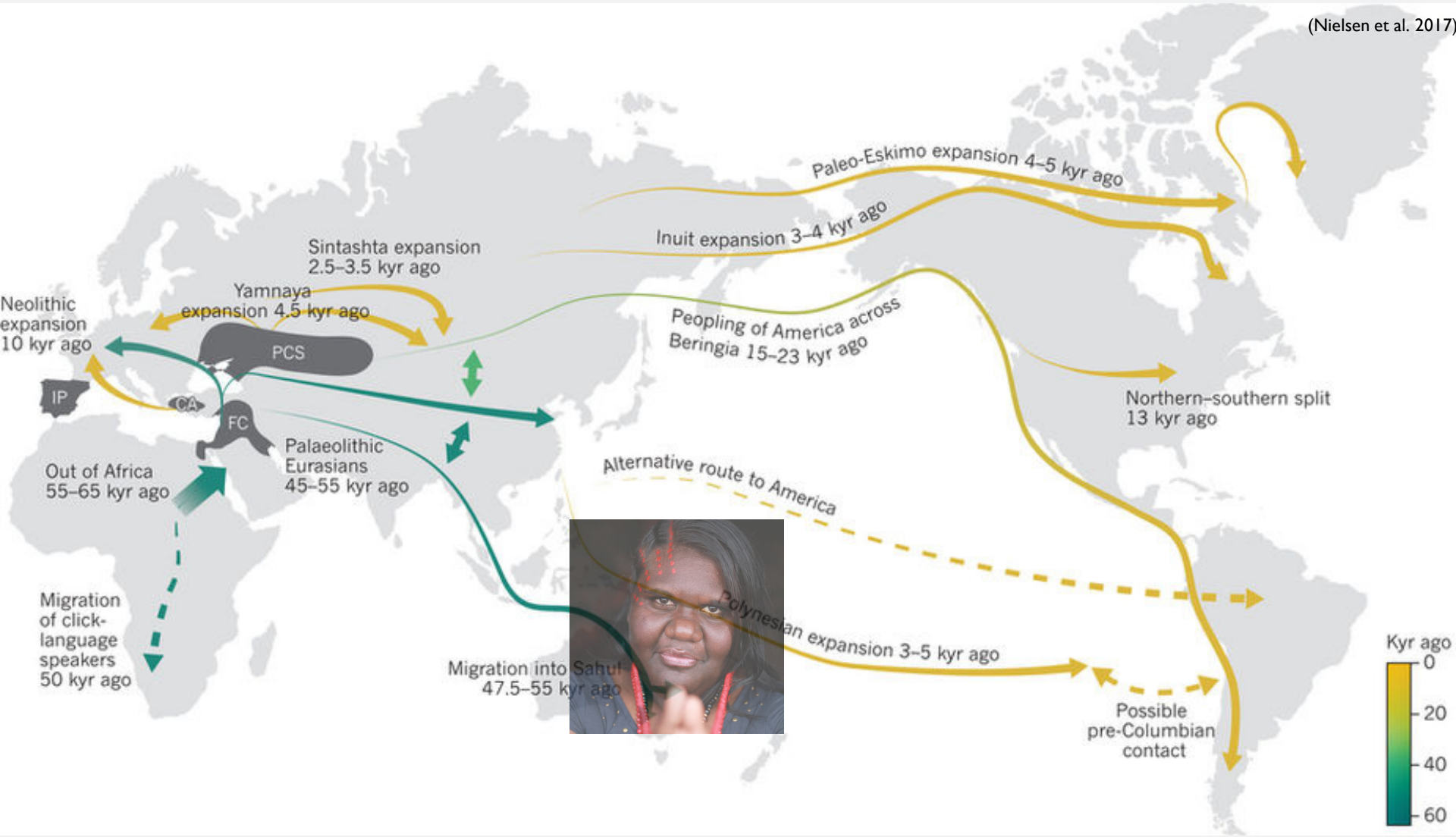
(Nielsen et al. 2017)



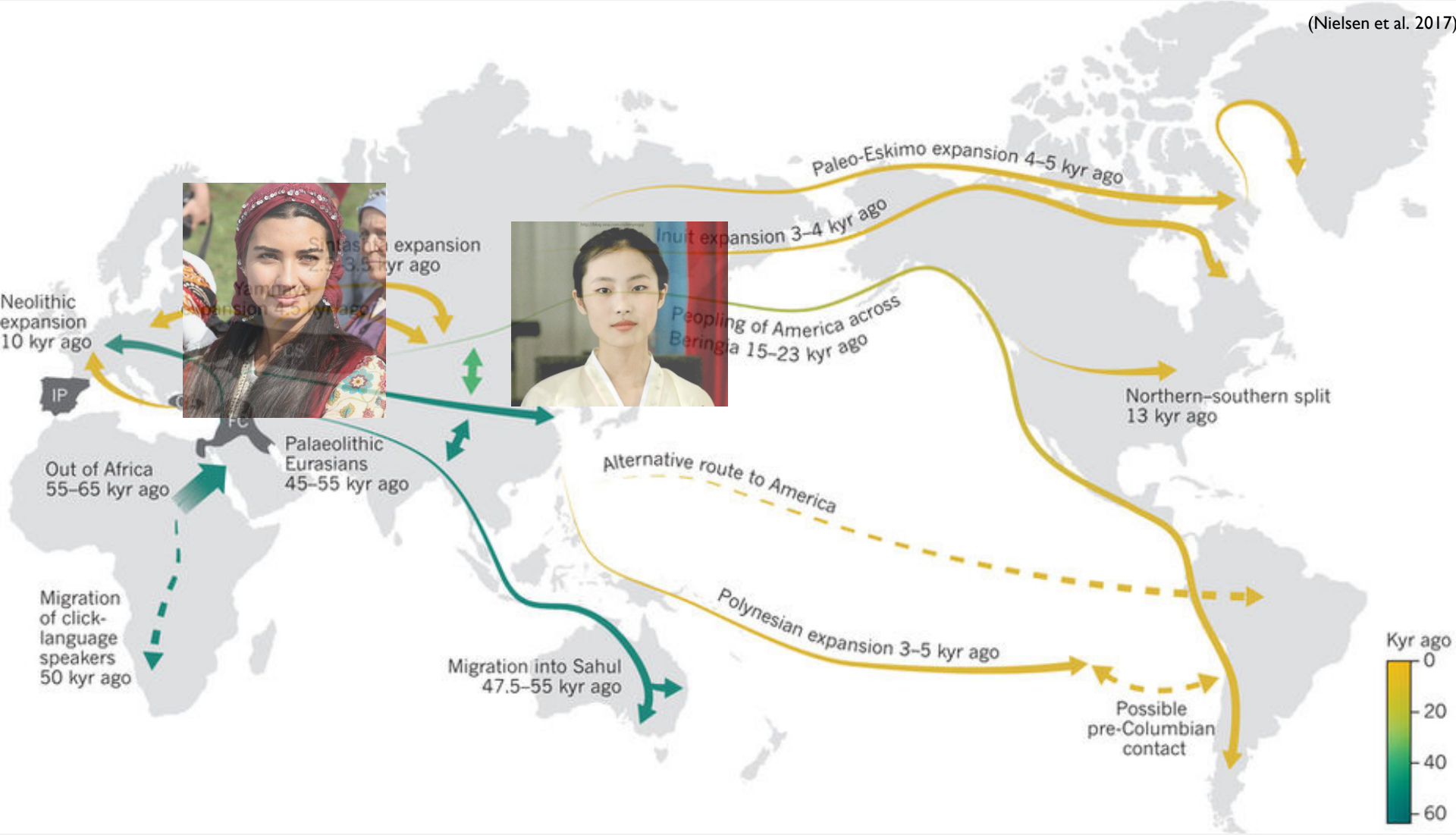
(Nielsen et al. 2017)



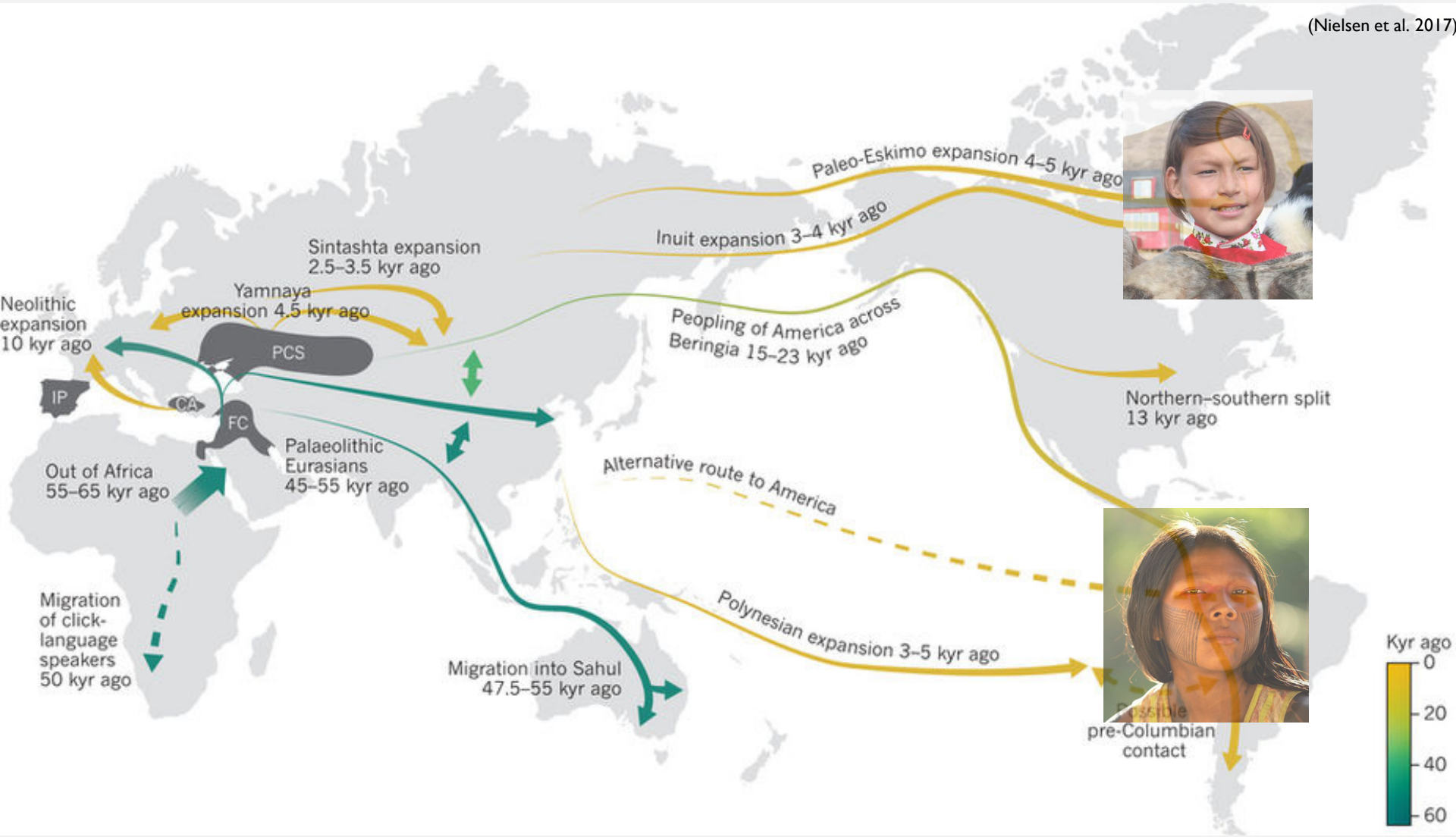
(Nielsen et al. 2017)



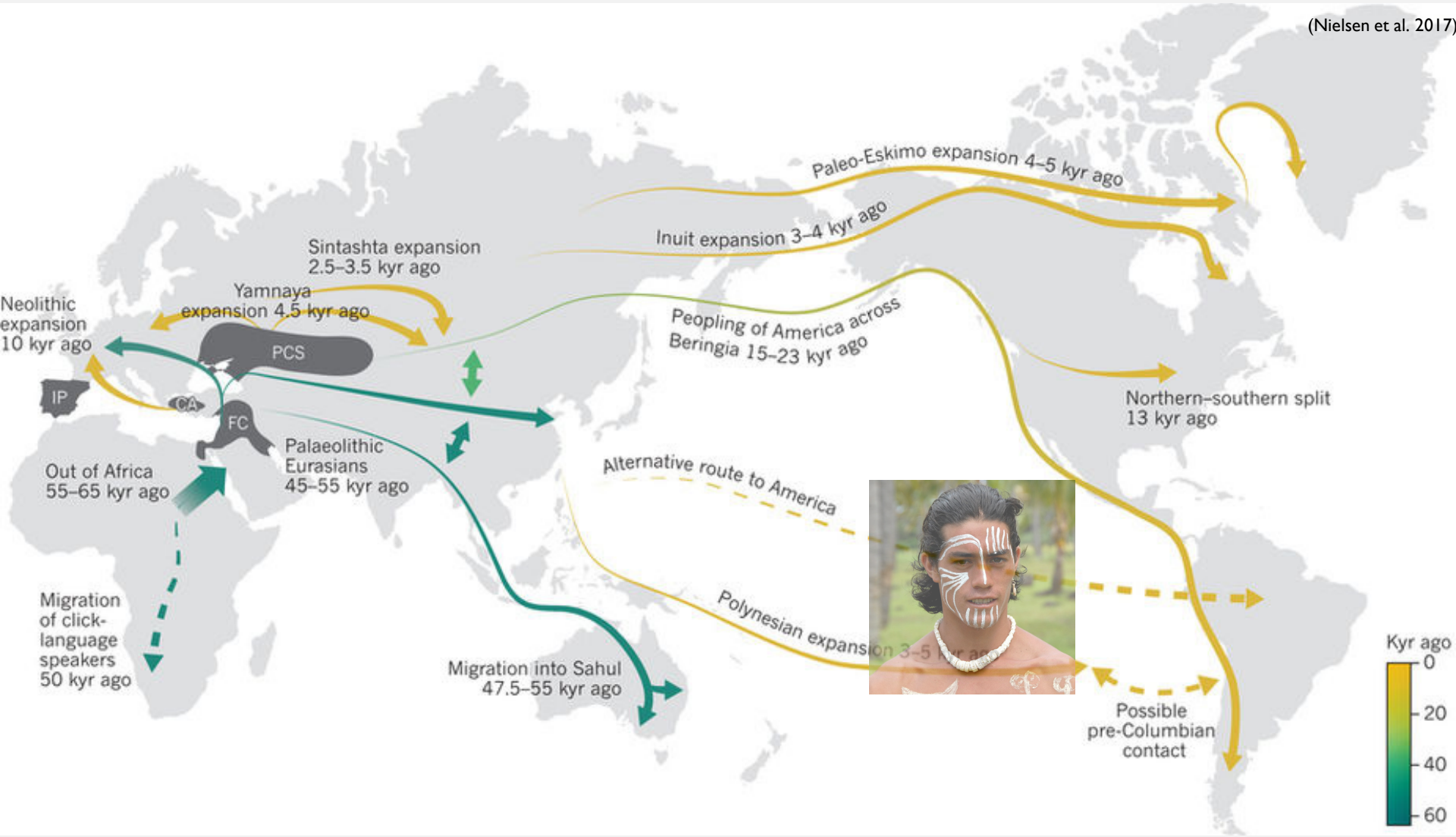
(Nielsen et al. 2017)



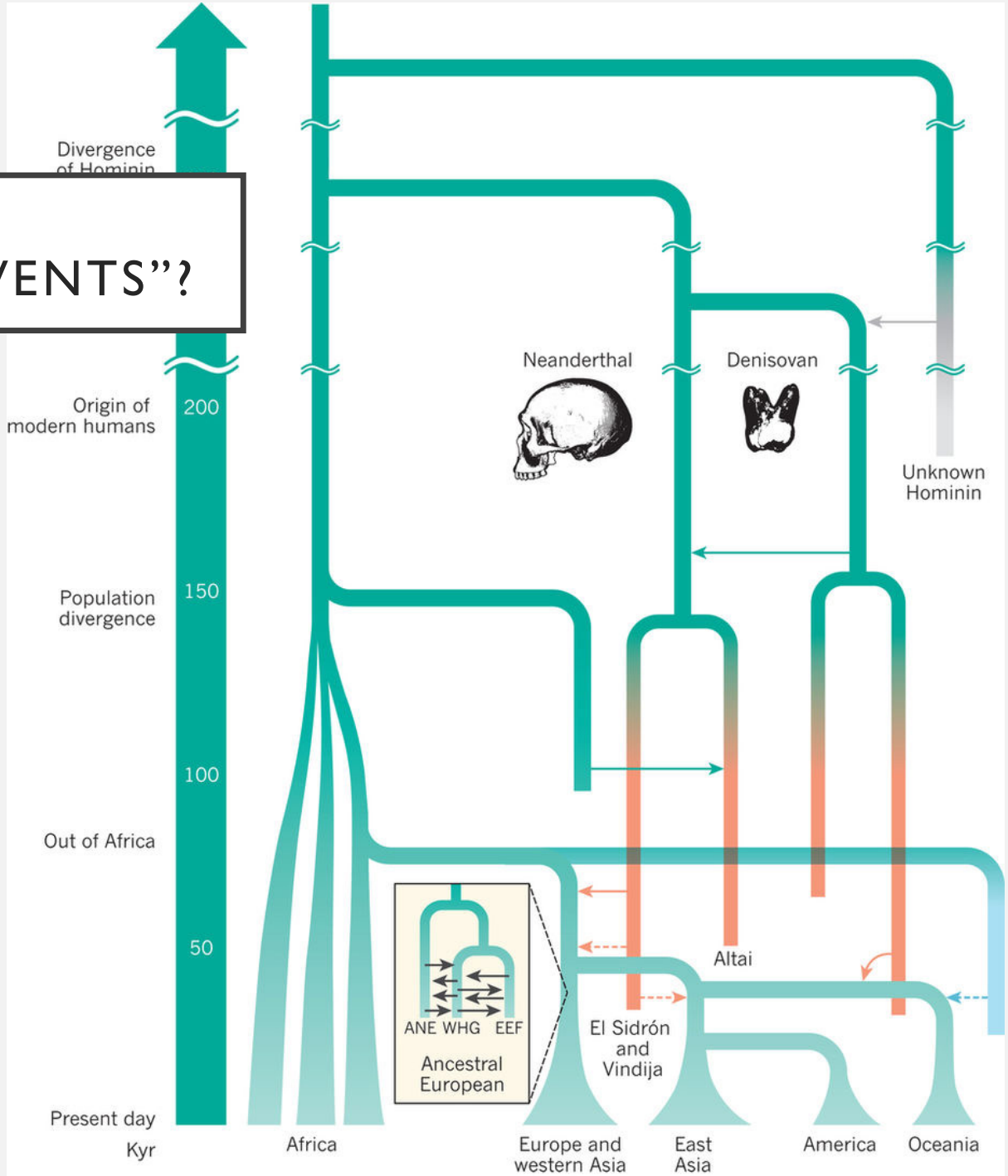
(Nielsen et al. 2017)



(Nielsen et al. 2017)

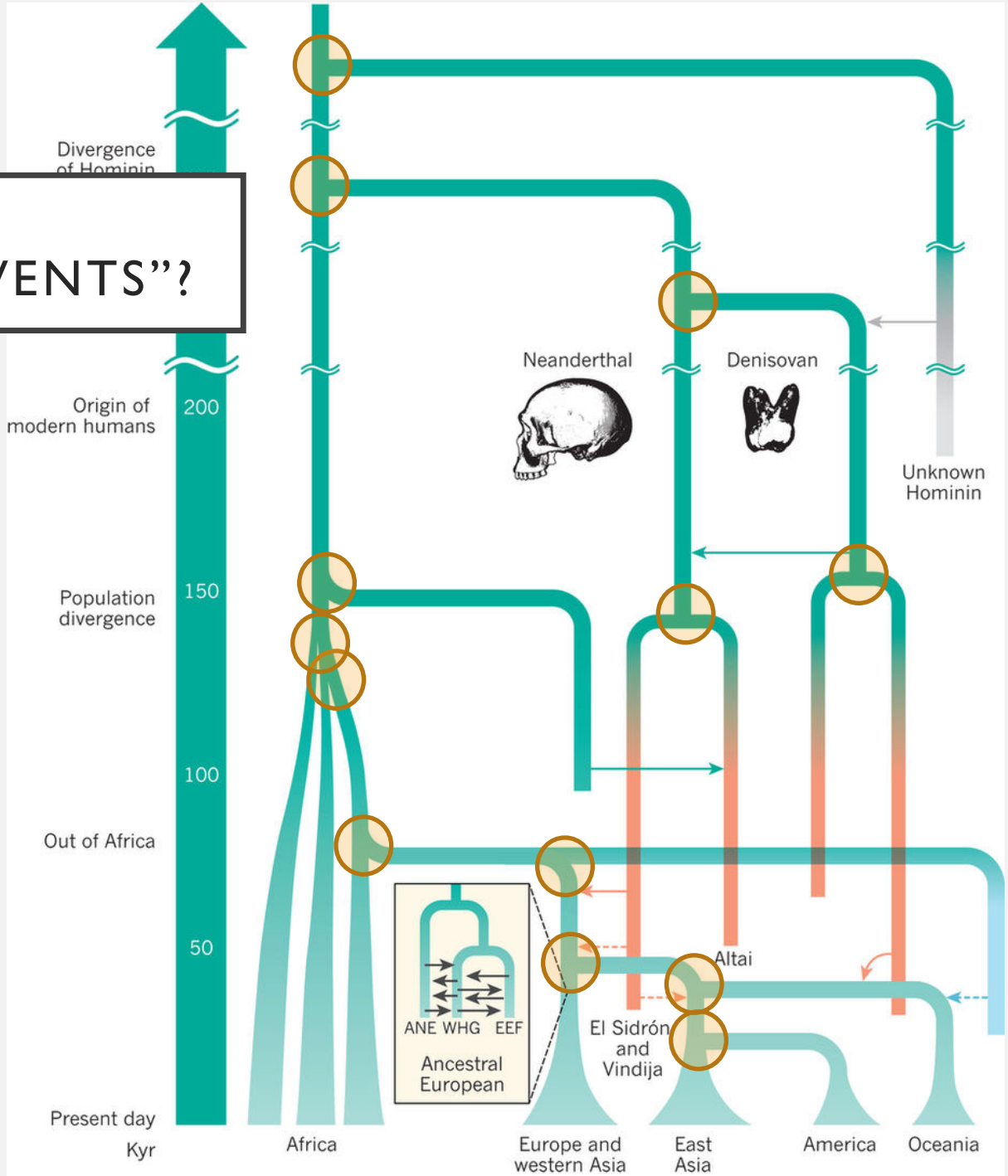


WHAT ARE "DEMOGRAPHIC EVENTS"?



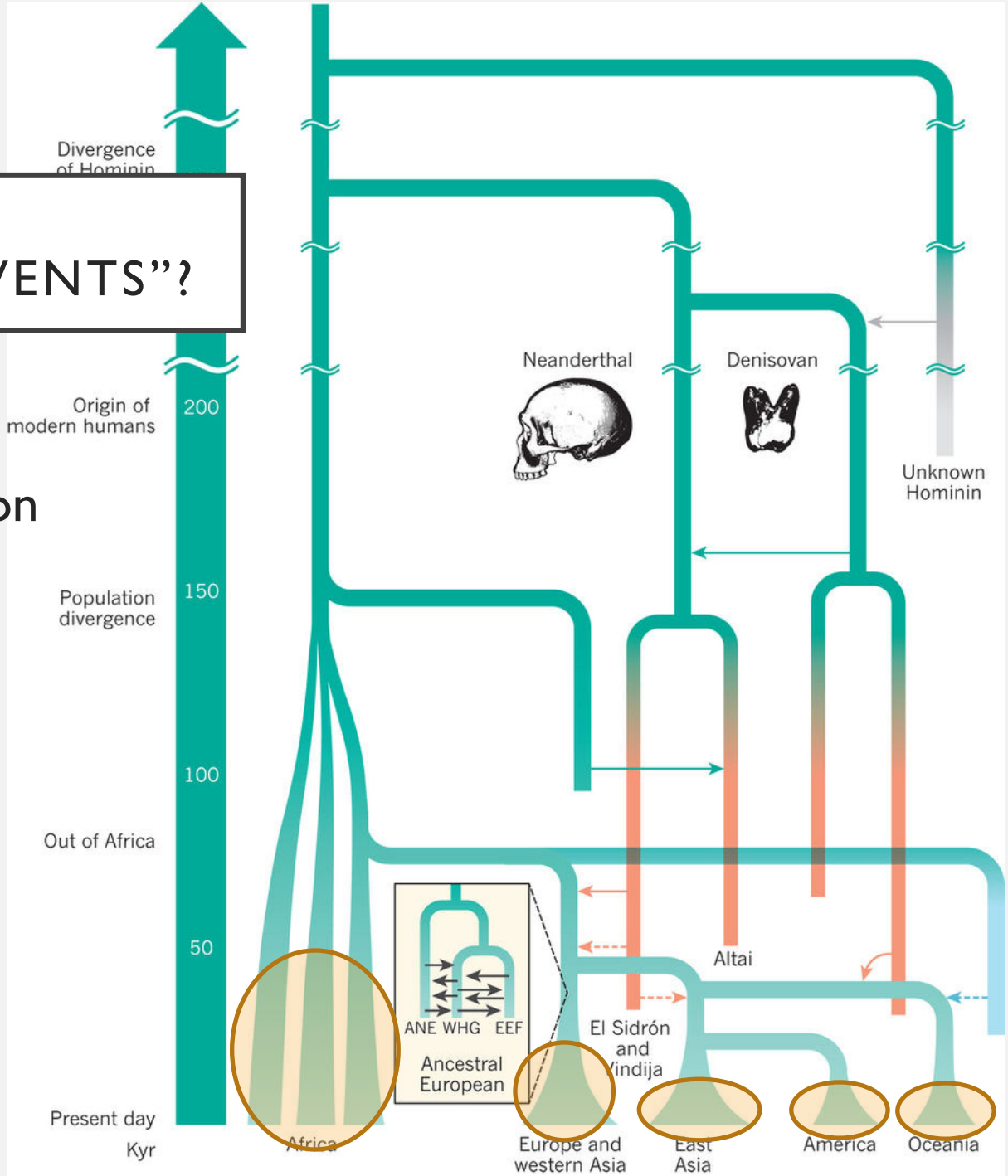
WHAT ARE "DEMOGRAPHIC EVENTS"?

- Divergence



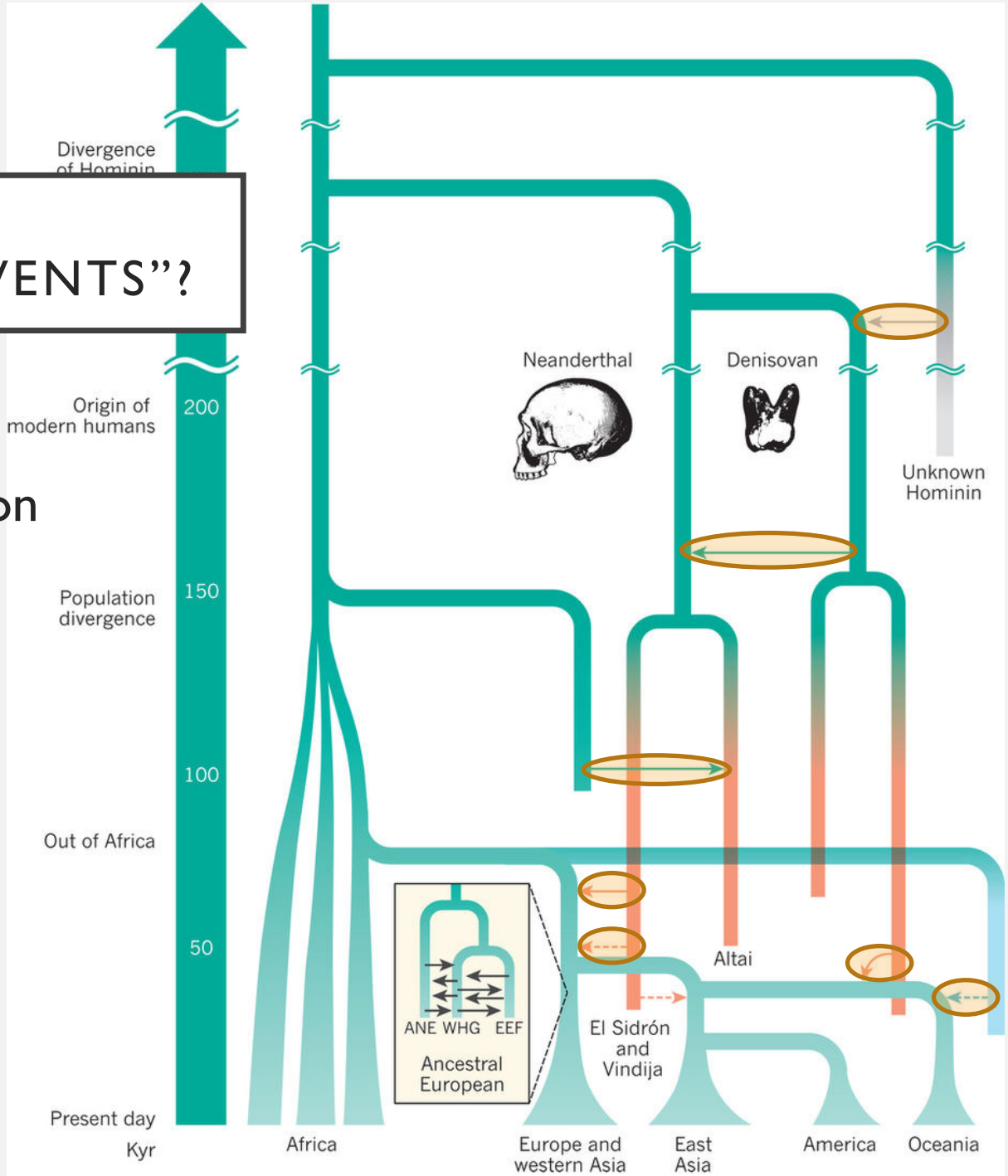
WHAT ARE "DEMOGRAPHIC EVENTS"?

- Divergence
- Expansion or reduction

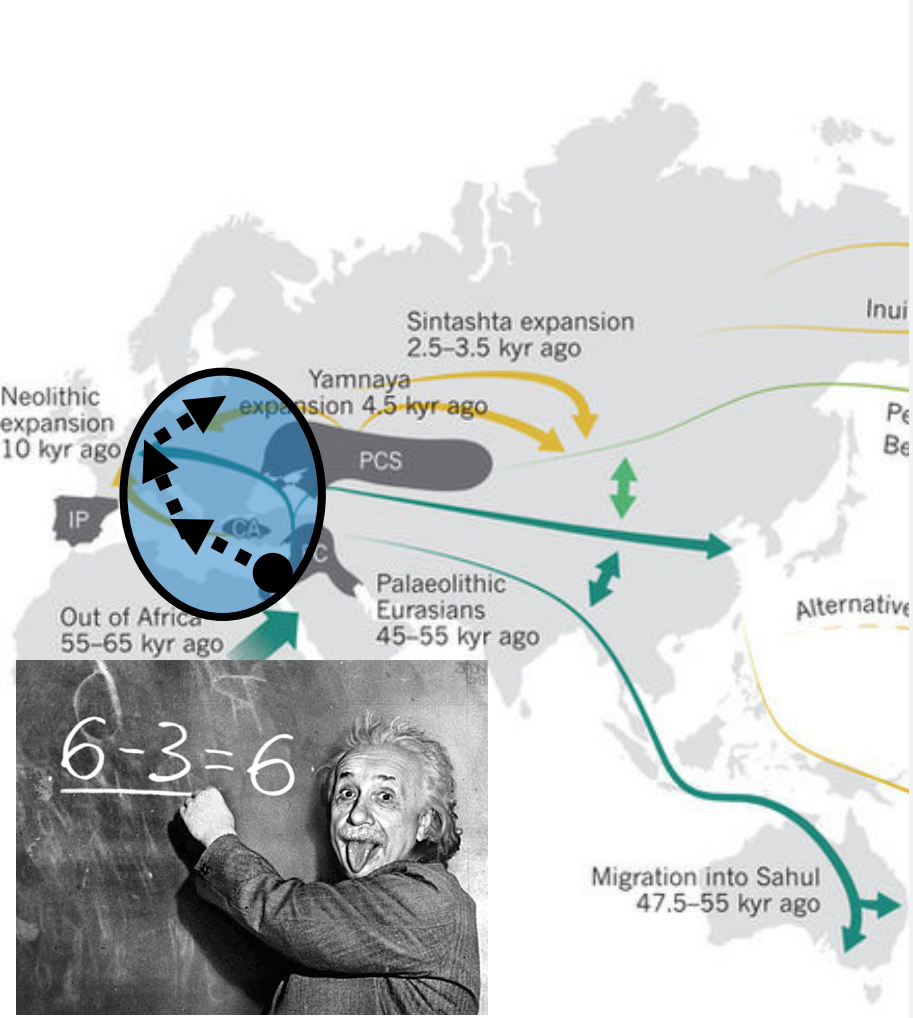


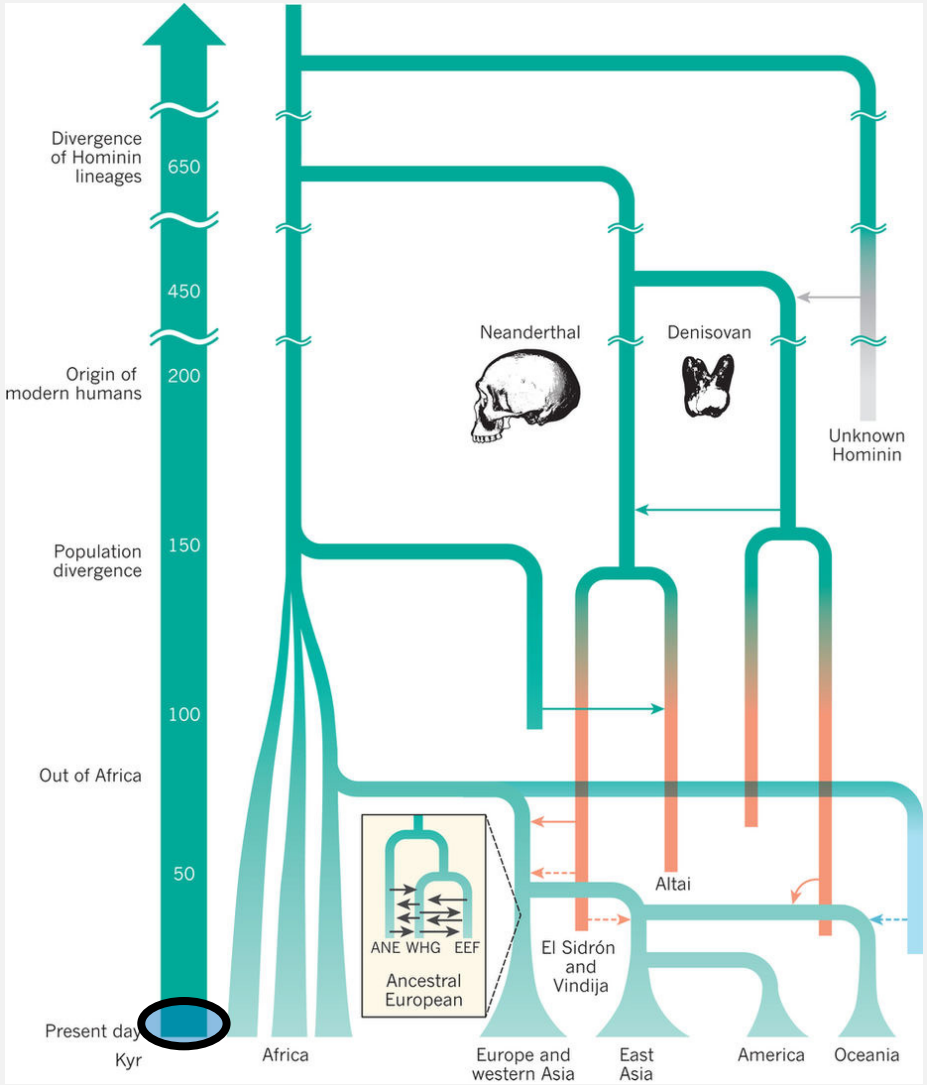
WHAT ARE "DEMOGRAPHIC EVENTS"?

- Divergence
- Expansion or reduction
- Gene flow



AIM: INFER THE DEMOGRAPHIC HISTORY OF THE ASHKENAZI JEWS.





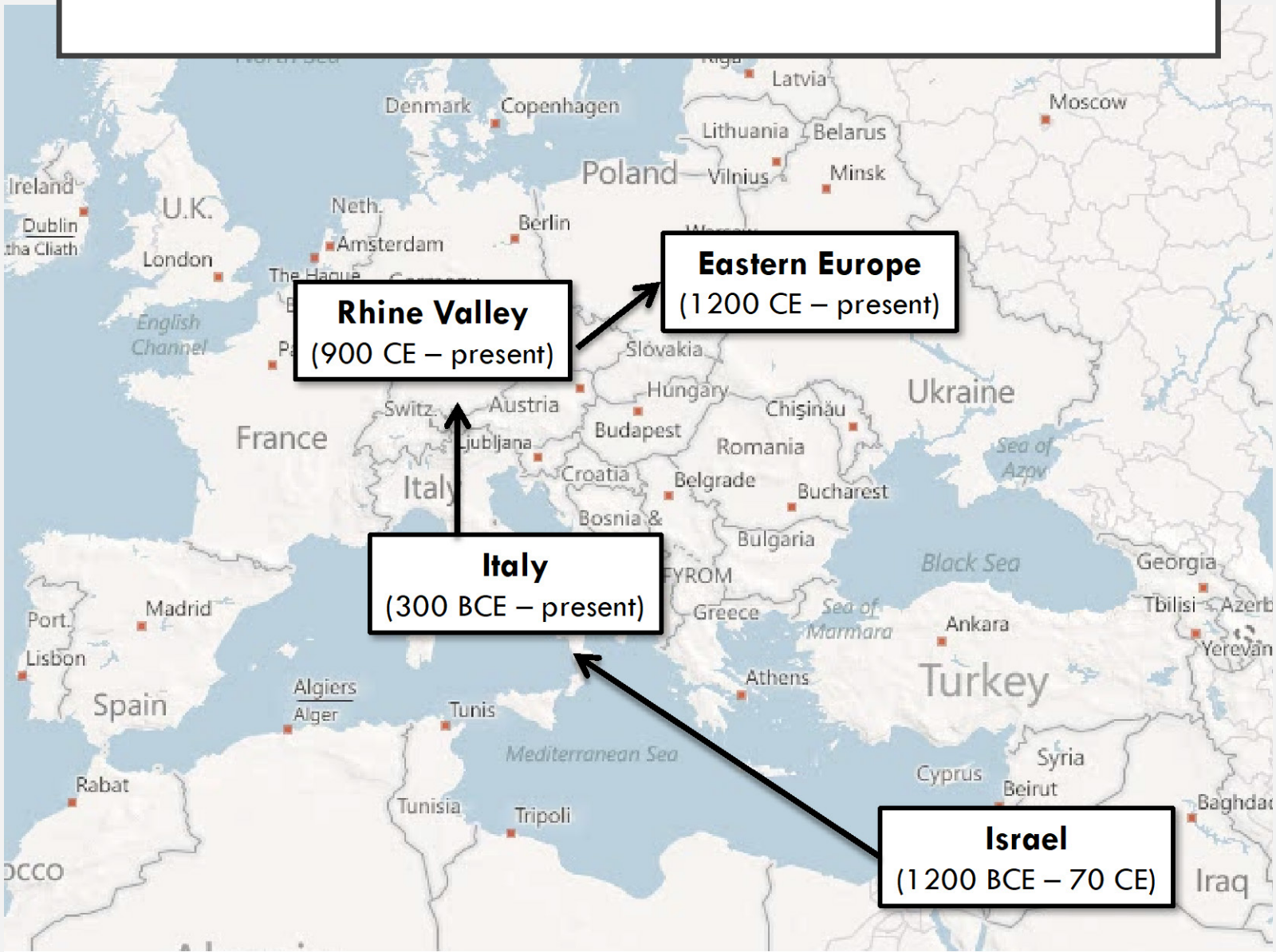
ASHKENAZI JEWS: AN INTERESTING STUDY POPULATION

- High frequency of genetic disorders
- Population isolate
- Complex demographic history
- Well documented historical record

ASHKENAZI JEWS: AN INTERESTING STUDY POPULATION

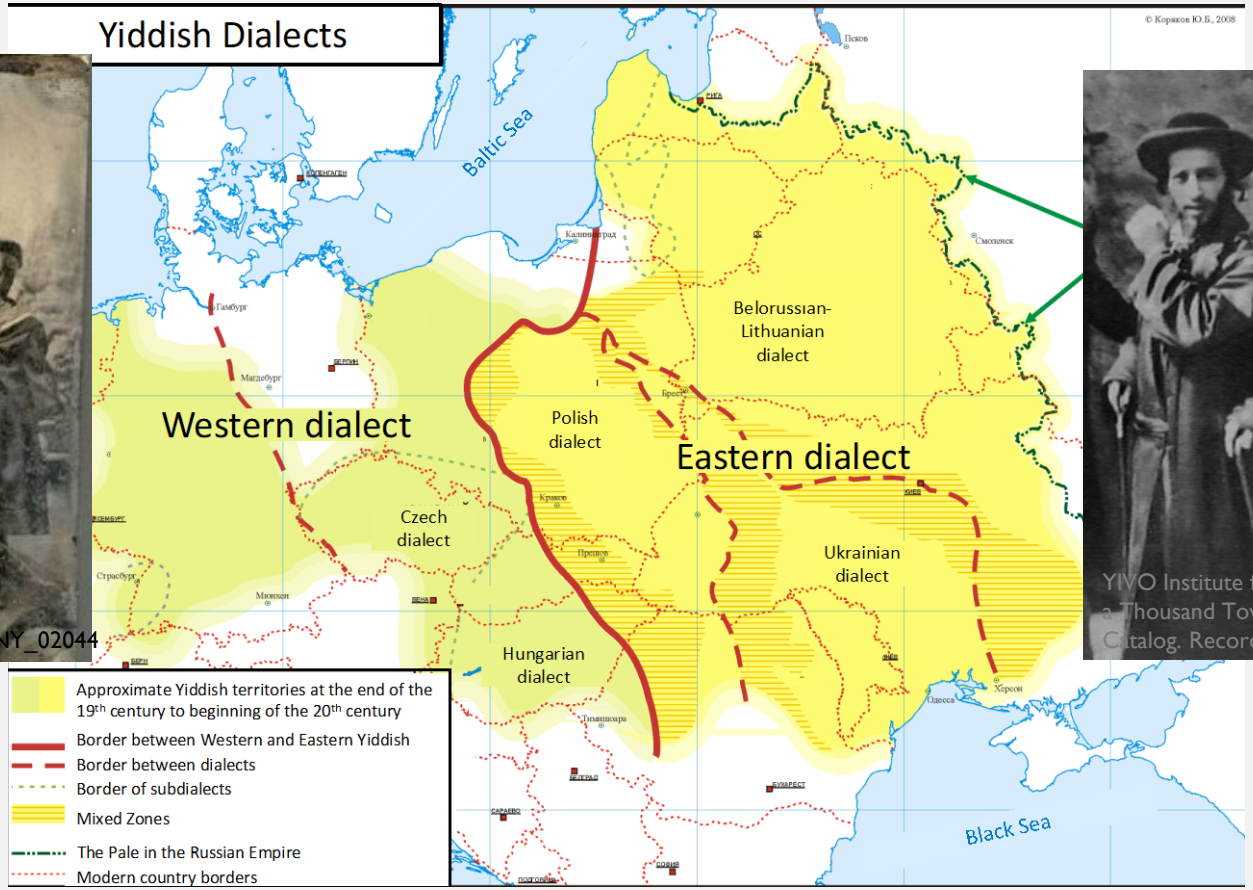
- High frequency of genetic disorders
- Population isolate
- Complex demographic history
- Well documented historical record

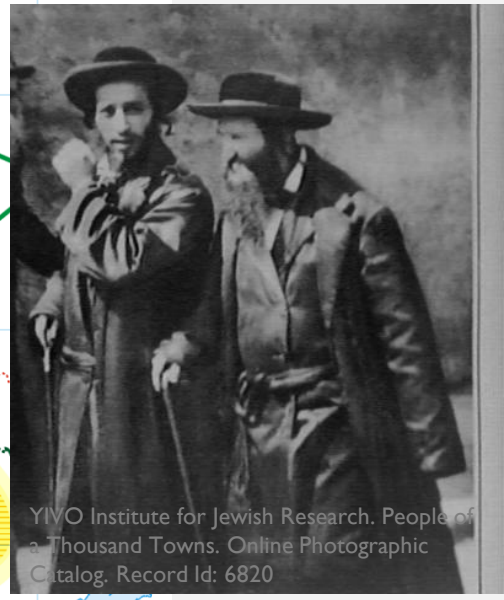
HYPOTHESIS OF ASHKENAZI ORIGINS



WESTERN VS. EASTERN ASHKENAZI JEWS

Germany, 1900's

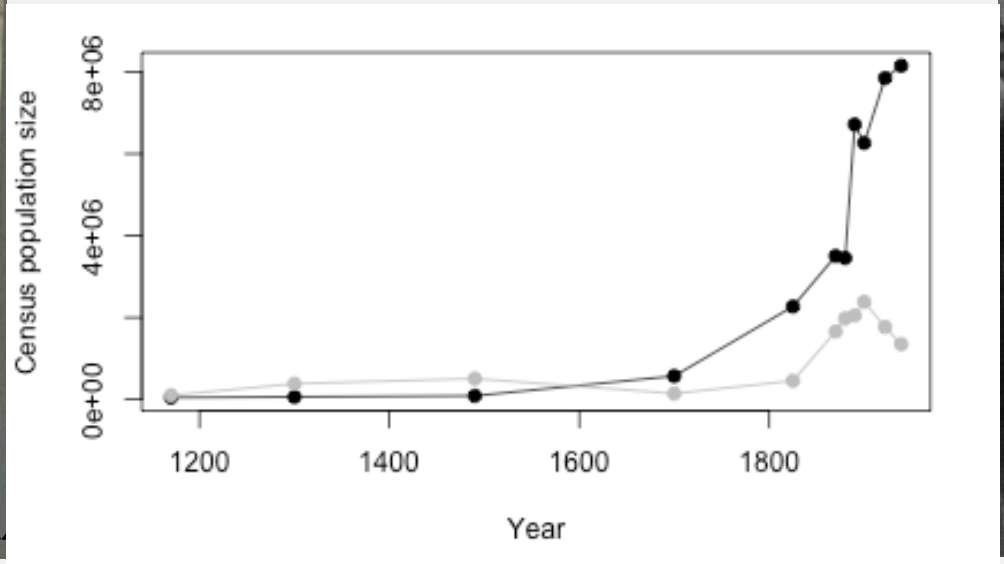




Cracow, Poland. 1932

WESTERN VS. EASTERN ASHKENAZI JEWS

Germany, 1900's

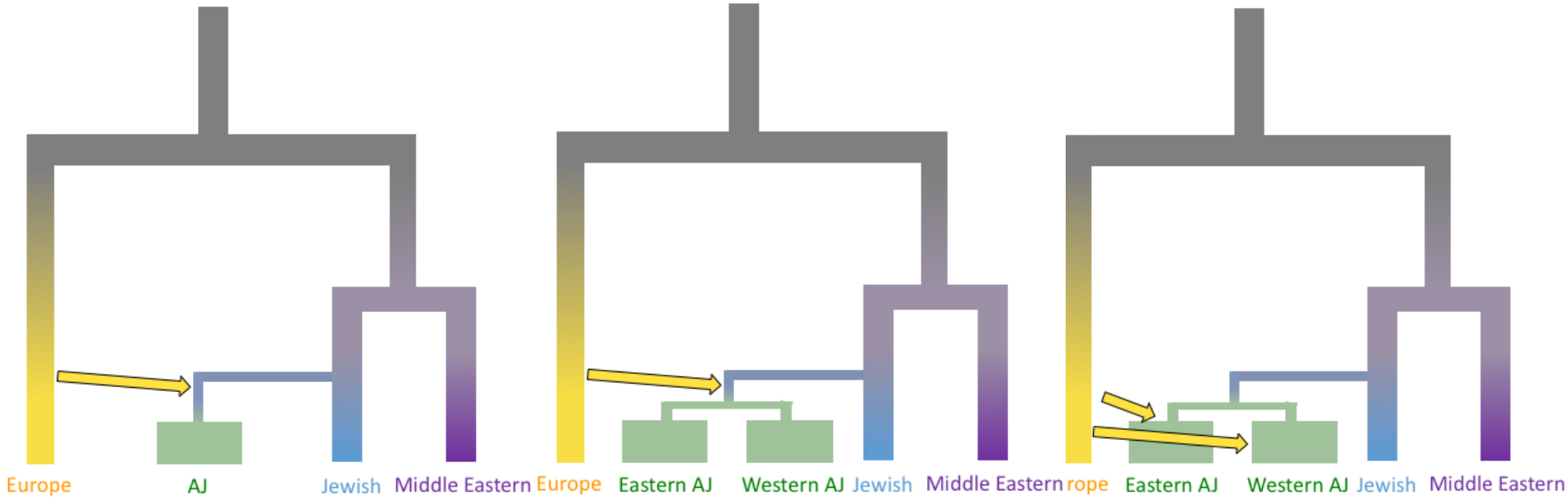


Cracow, Poland. 1932

Reference census data

MOTIVATION

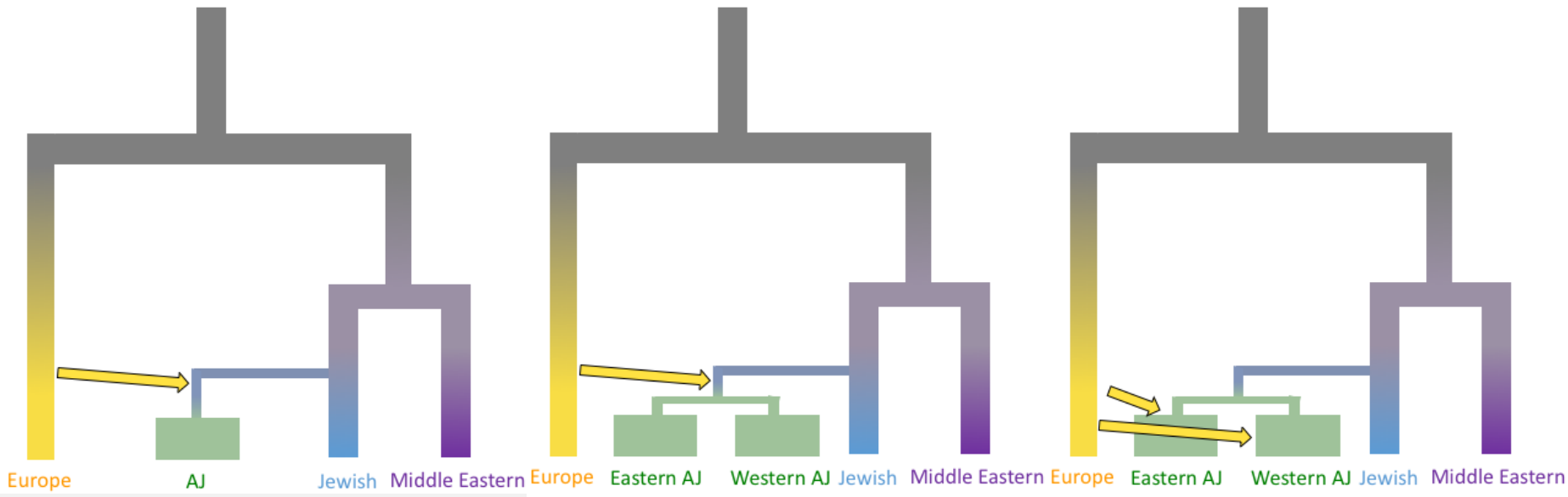
- Numerous genetic studies on the Ashkenazi Jews.
  - All genome-wide studies treat Ashkenazi Jews as one population.
- Preliminary work consistent with genetic differentiation.
  - Not informative of cause of differentiation.



MODELS OF ASHKENAZI HISTORY

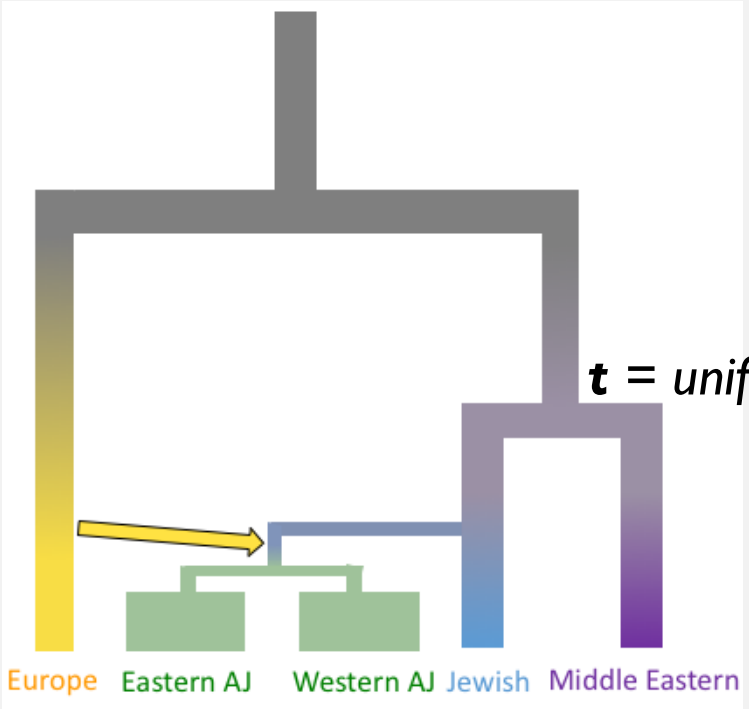
APPROXIMATE BAYESIAN COMPUTATION

- 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

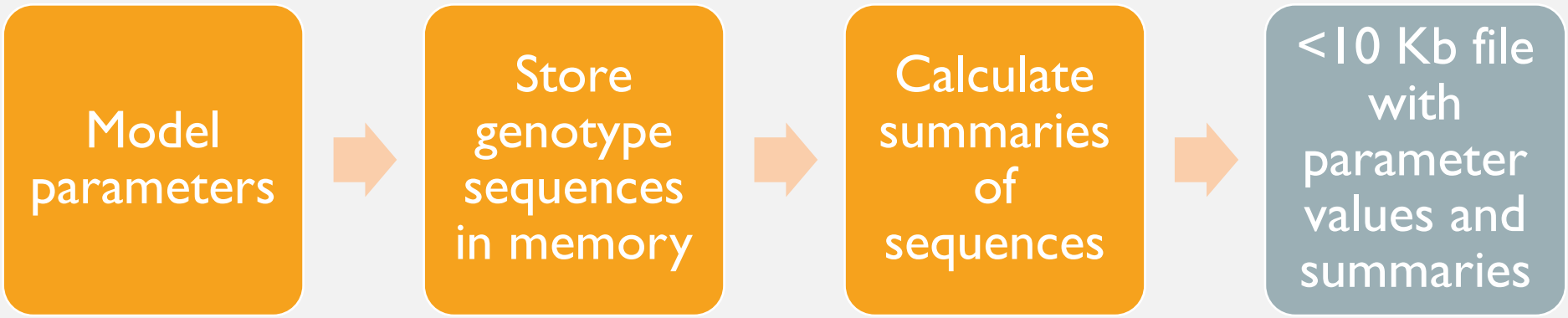
APPROXIMATE BAYESIAN COMPUTATION

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
<10 Kb file with parameter values and summaries

EMBARRASSINGLY PARALLEL!

INHERITED SCRIPT INTENDED FOR SMALL SEQUENCE

1,389 10kb regions

00000110001

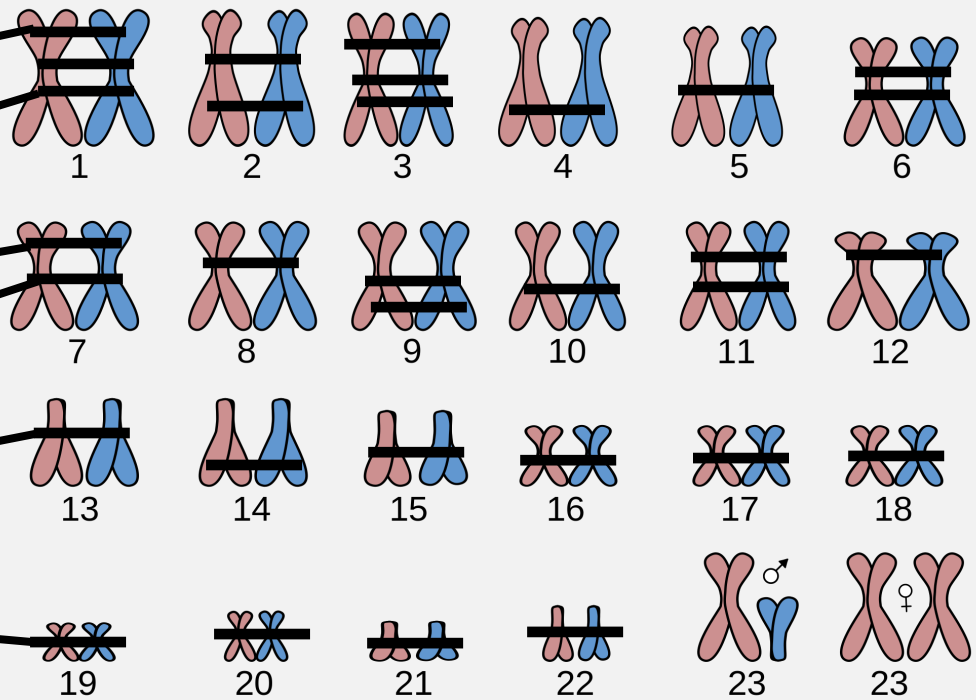
00100010000

00000100101

00100000000

00010001010

00100010001



SIMULATE WHOLE CHROMOSOME

~250 million sites on human chromosome 1

PROBLEM!

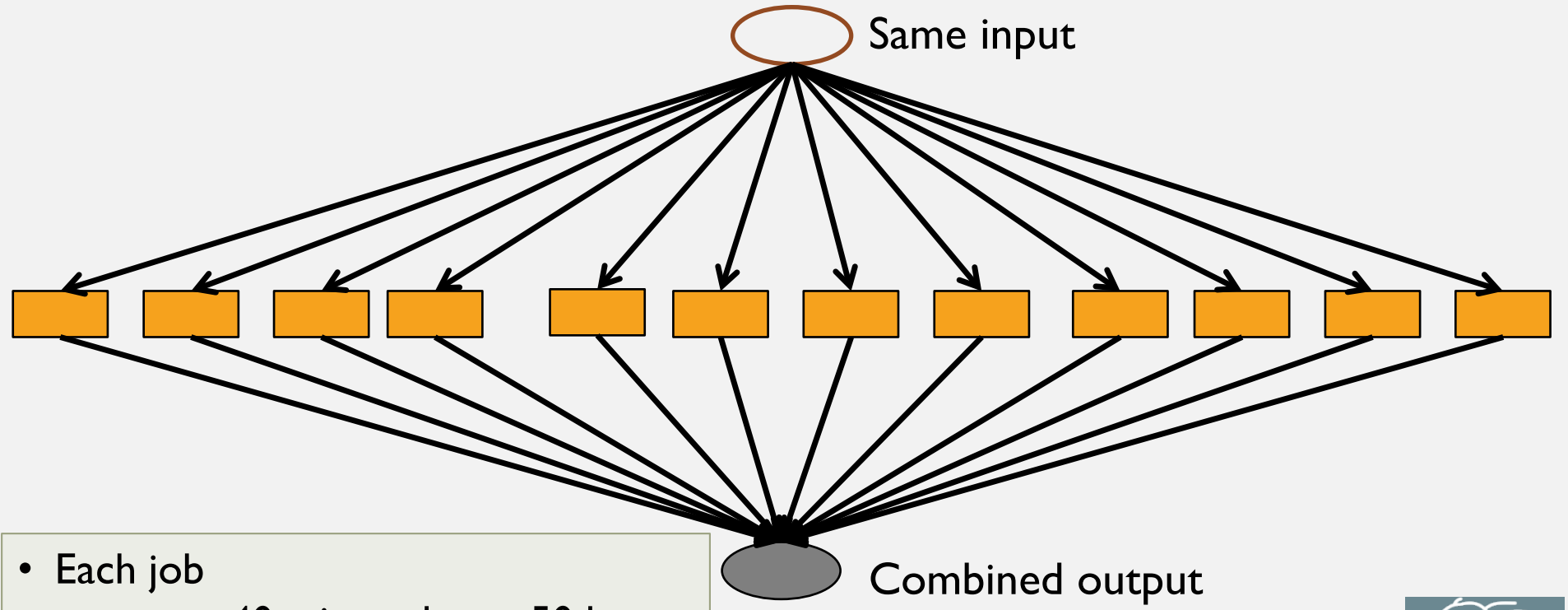
Parameters	Average Walltime	Average Memory
Minimum	00:21:00	2.7 Gb
Random	00:55:11	20 Gb
Maximum	08:02:11	117 Gb

Too much memory!

Over a decade to complete 6000 runs/month w/ UA resources

Each core on UA HPC has 6G - **Need memory < 6G** for each run

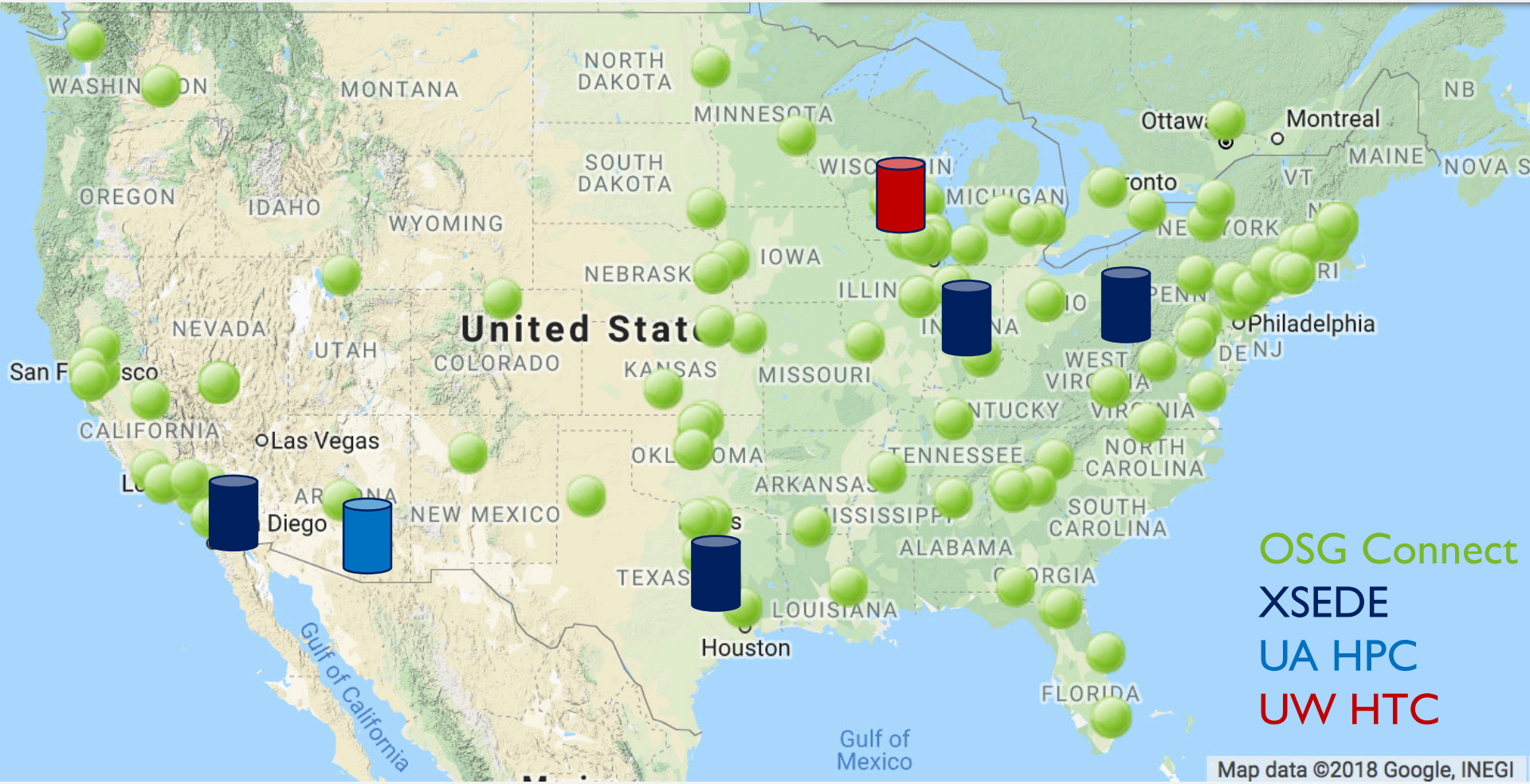
EMBARRASSINGLY PARALLEL & RESOURCE LIGHT!



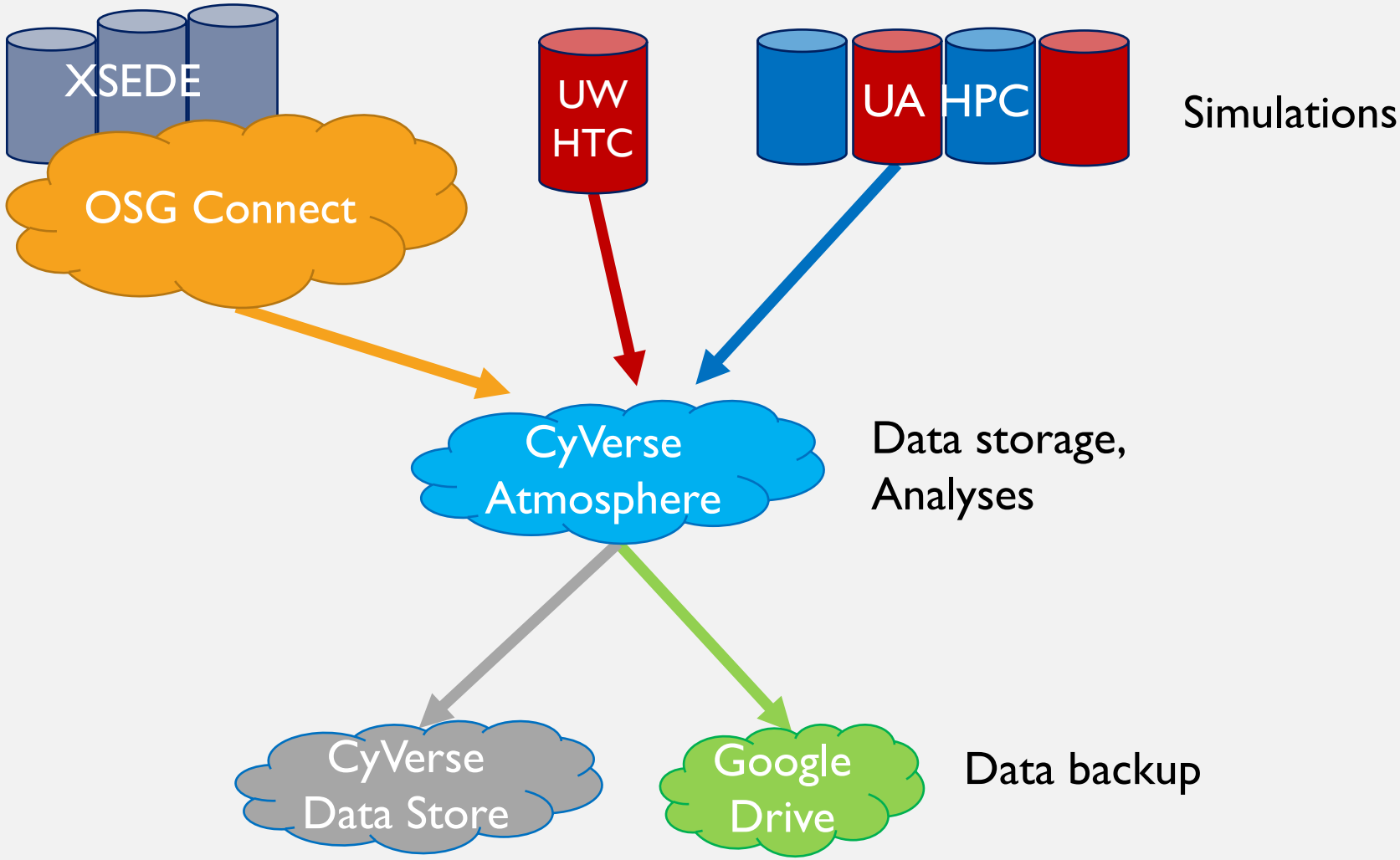
- Each job
  - runs ~40 min, and max 50 hrs
  - Uses ~1G, and max 5G memory
  - Uses ~2M in storage

Pegasus

HIGH THROUGHPUT COMPUTING



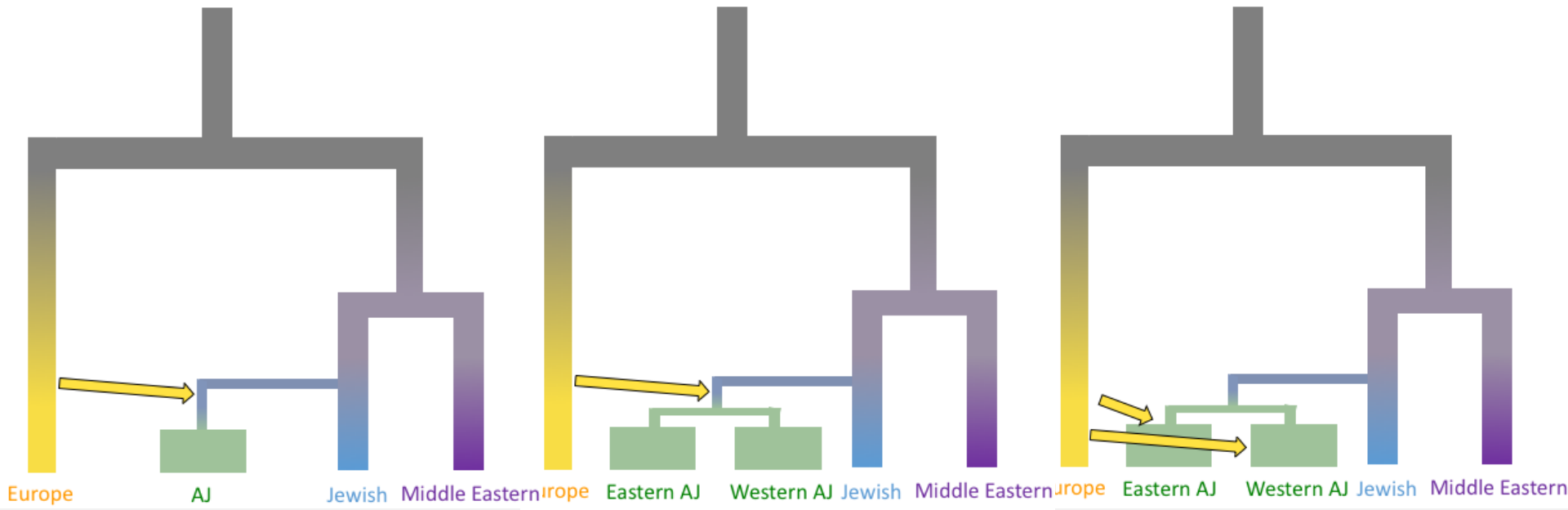
SIMULATIONS ON HTC CLUSTERS, ANALYSES ON VM



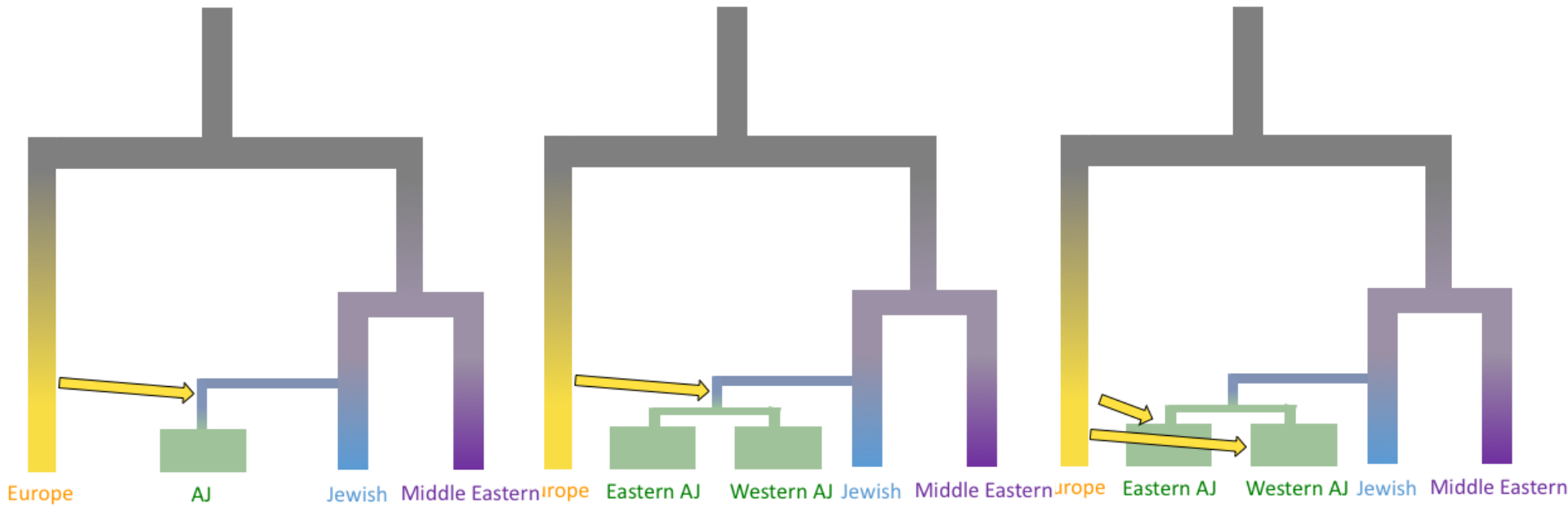
CHALLENGES: TECHNICAL

- 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?

>1 MILLION SIMULATIONS OF EACH MODEL



MODEL CHOICE

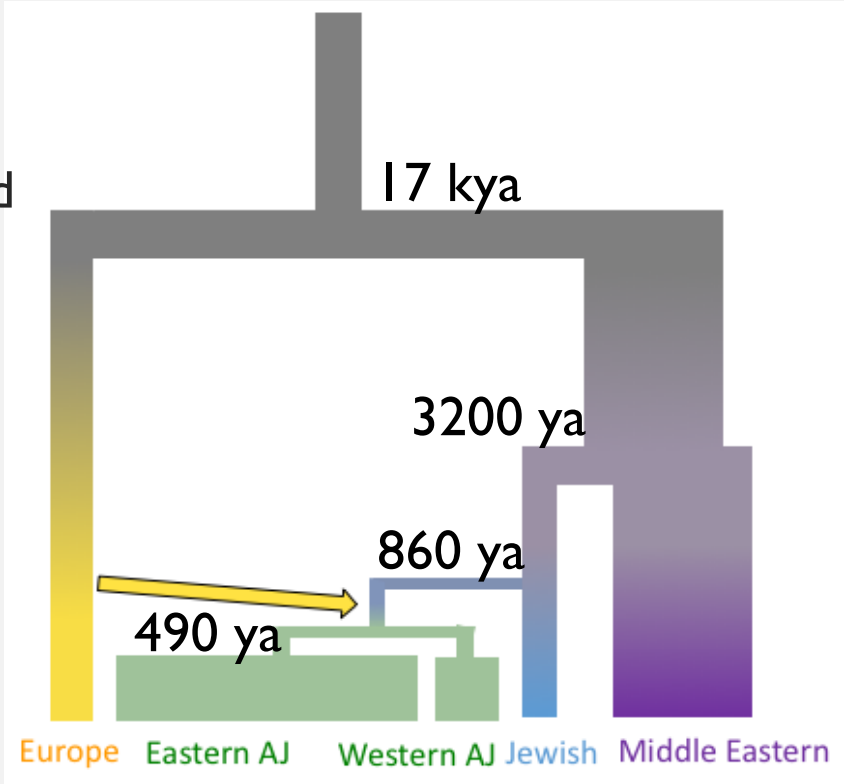


Posterior probability: 0.0065 0.85 0.14

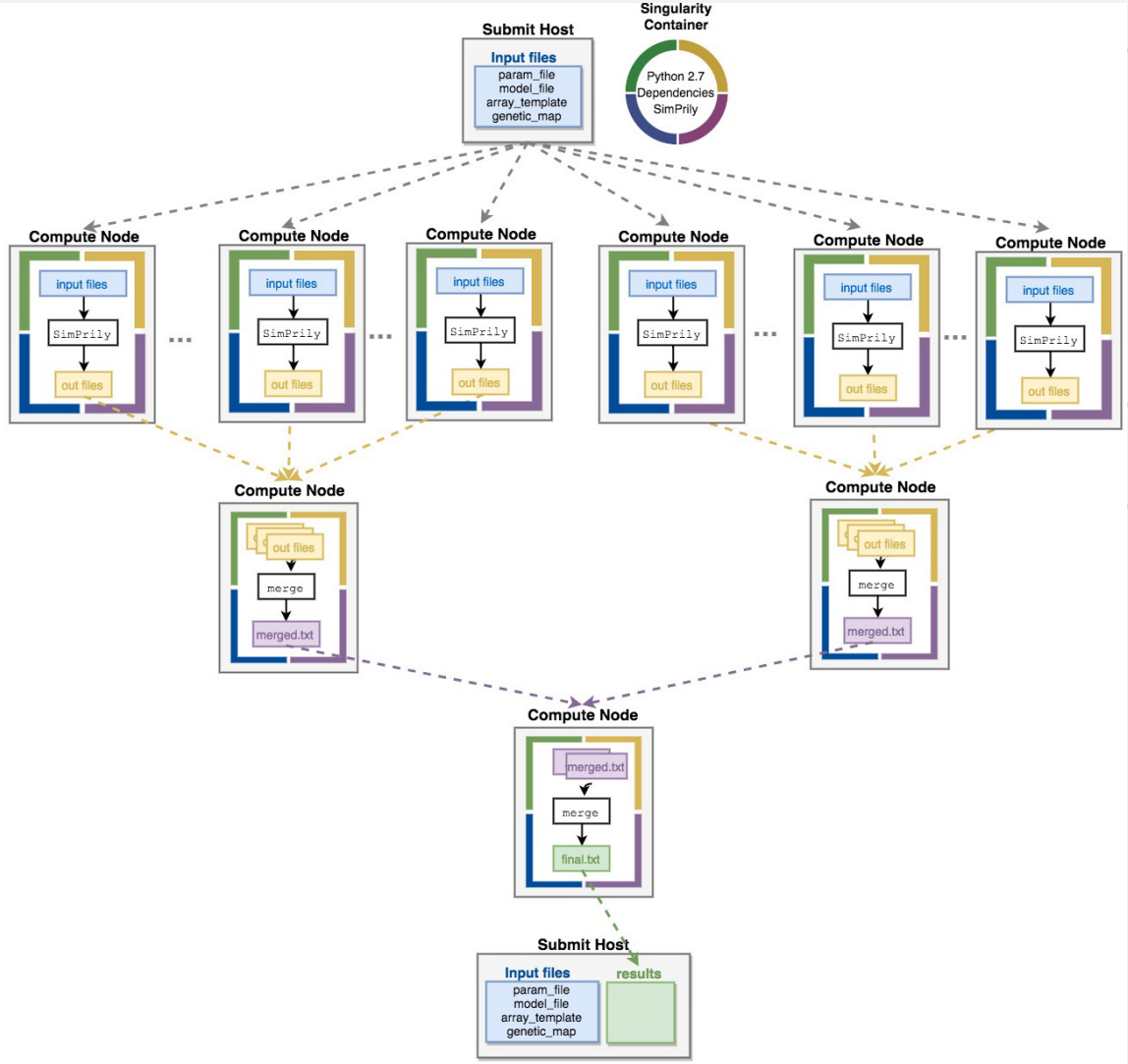
BEST MODEL

- ~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/

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

THE UNIVERSITY OF ARIZONA

WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

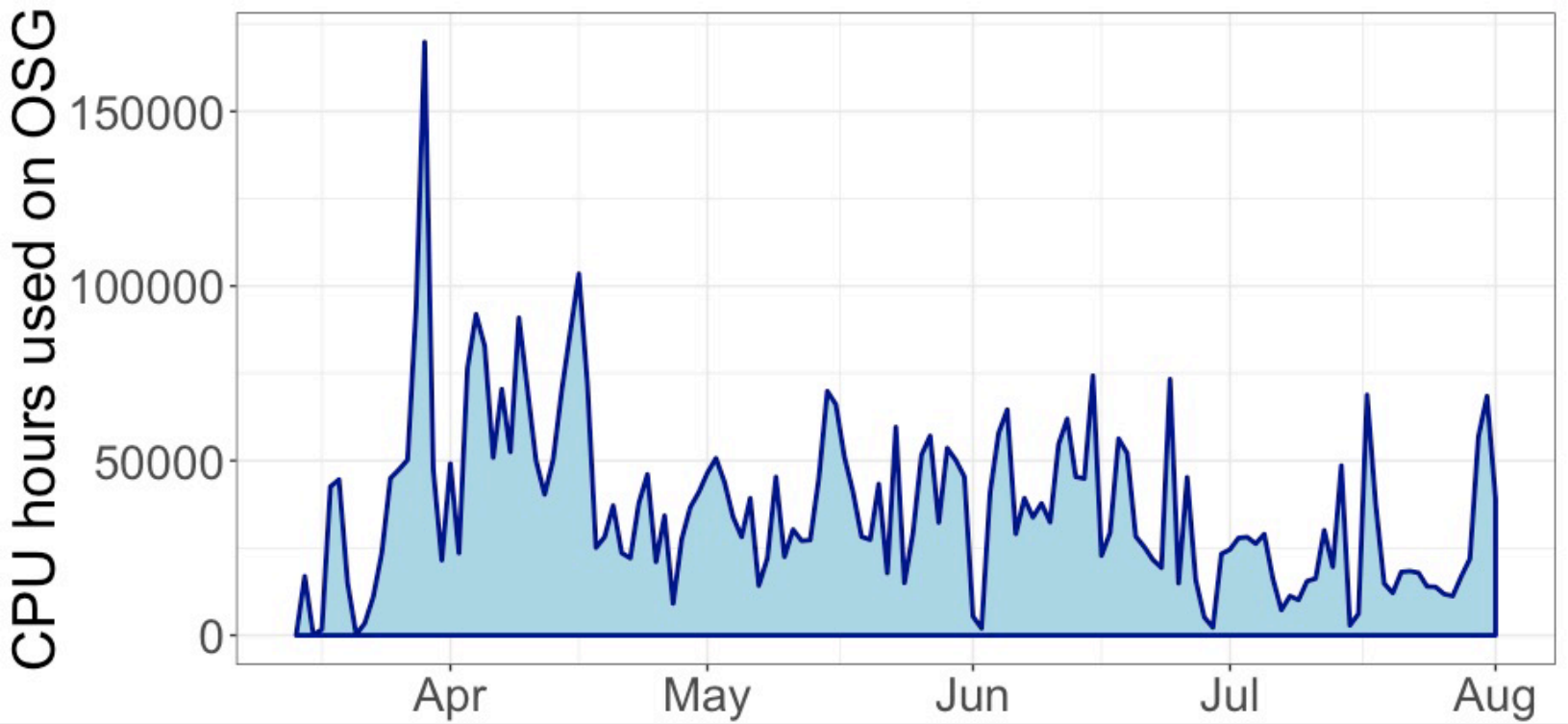
Pegasus

XSEDE Extreme Science and Engineering Discovery Environment

CYVERSE

osg connect

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
	AAACATTTCCGTCTTTATGGTTGCGCTGCATTGGGGAA

DNA SEQUENCE, GENOTYPES ENCODED 0/1

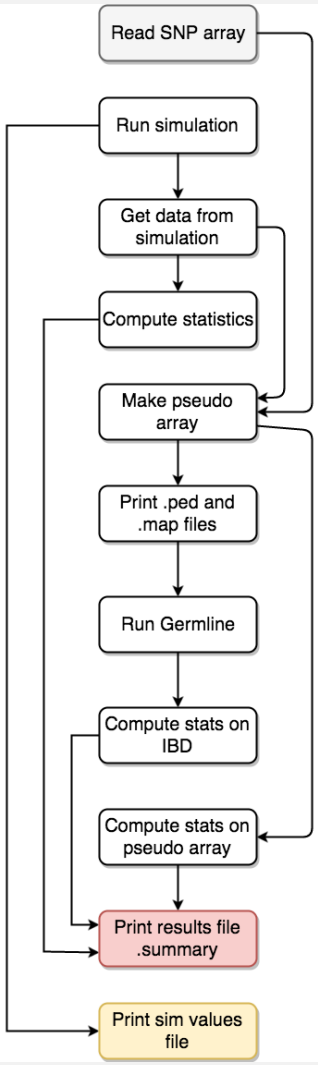
Indiv 1	AATCATTTCGGTTTTAATGCTTGGGCTGCATTGGGAAA AATCATATCGGTCTTAATGCTTGCGCTGCCTTGGTAAA
Indiv 1	000000000000000000000000100000000001000 000000100000100000000000000000100000000
Indiv 2	AATCATTTCGGTTTTAATGCTTGGGCTGCCTTGGTAAA AAACATTTCCGTCTTTATGGTTGCGCTGCATTGGGGAA
Indiv 2	00000000000000000000000000000100000000 00100000001001001000100010000000000001100

SEQUENCE OF GENOTYPES, ONLY SEGREGATING SITES

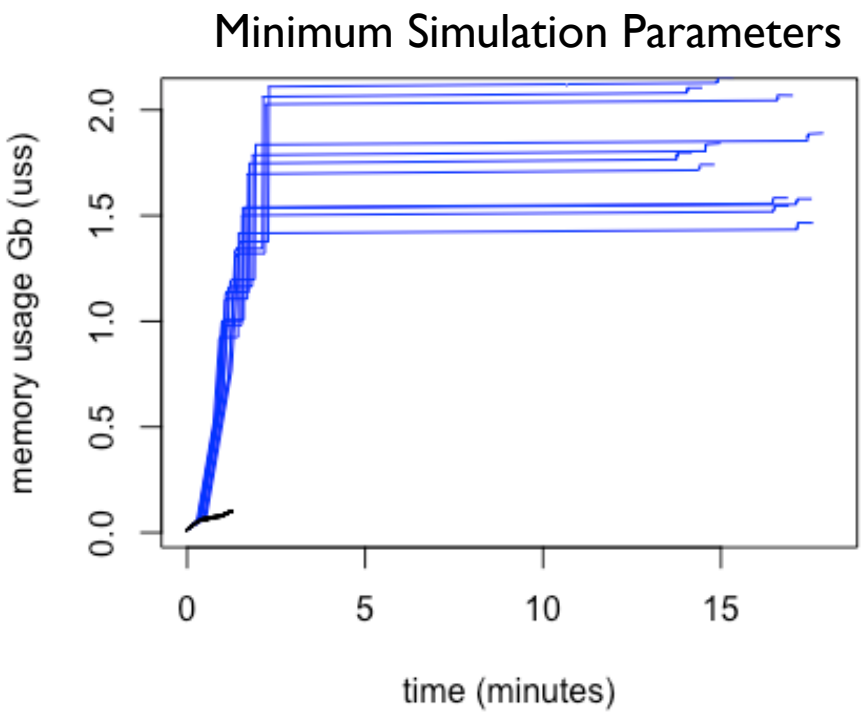
Indiv 1	0000001010 0101000100
Indiv 2	0000000100 1011111011

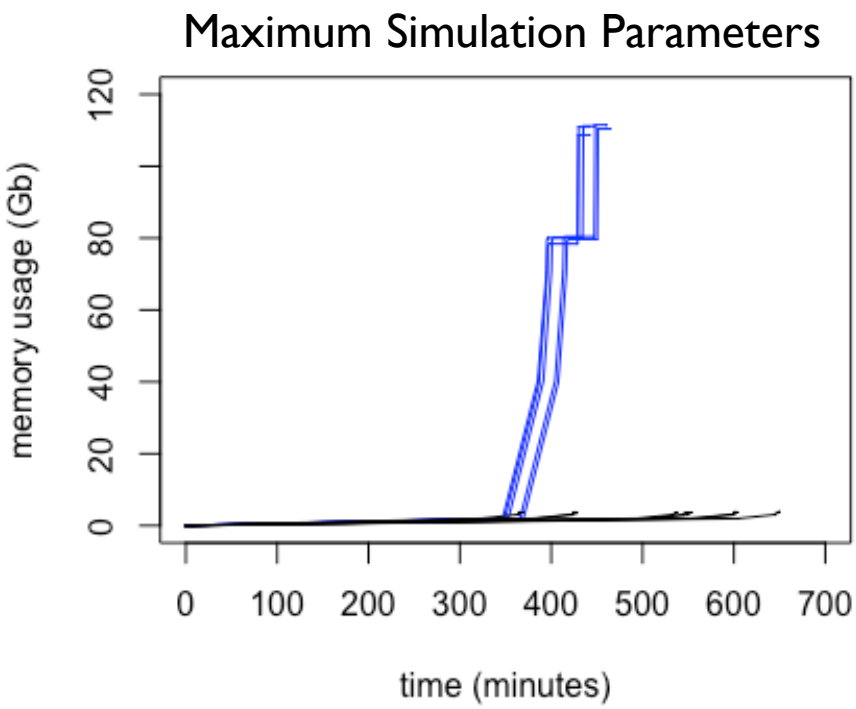
PYTHON SCRIPT: GENOME SIMULATIONS AND COMPUTE SUMMARY STATISTICS

- 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.

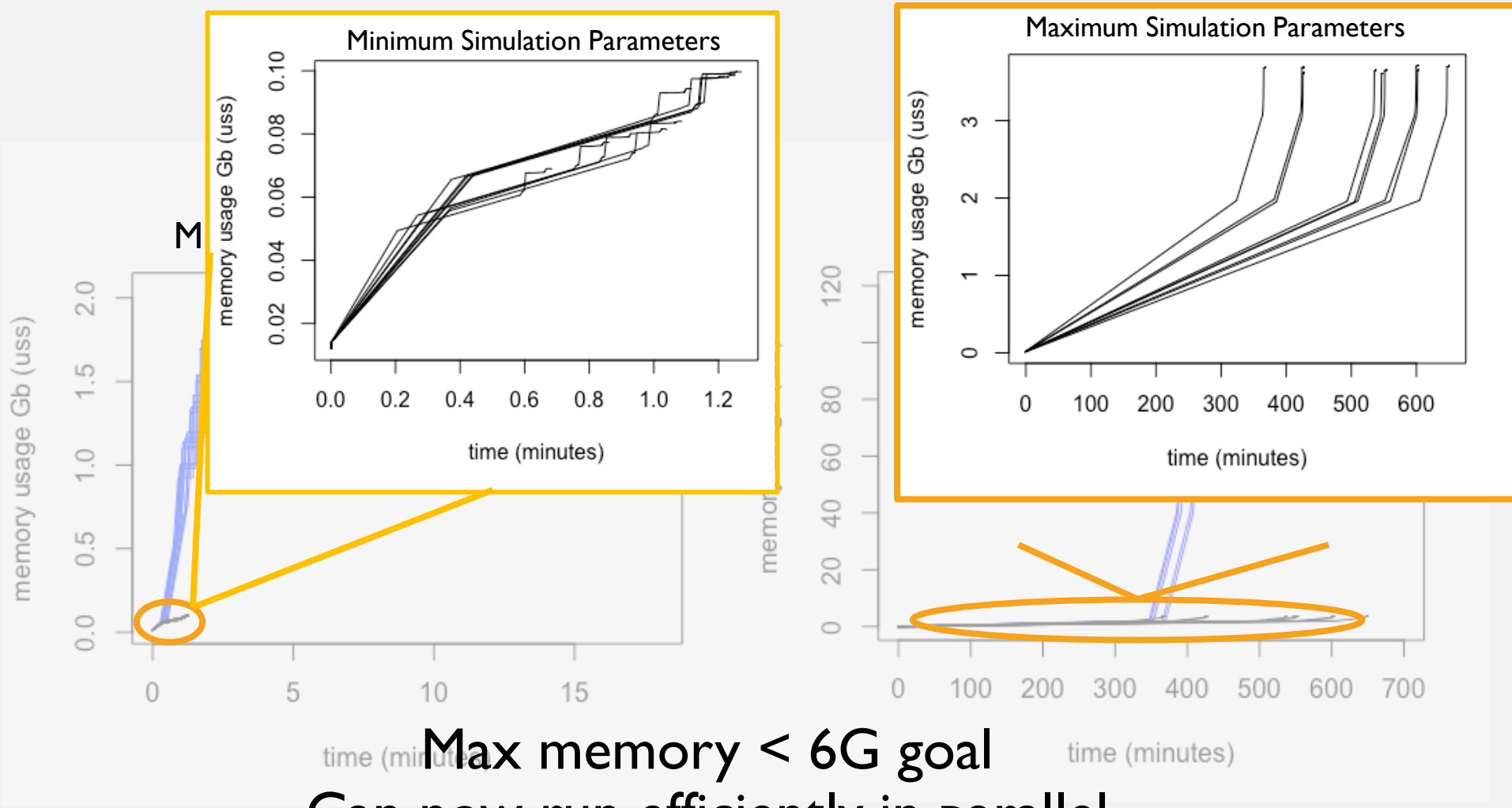


PROFILE OF PYTHON SCRIPT





*Note different scales



Max memory < 6G goal

Can now run efficiently in parallel

*Note different scales