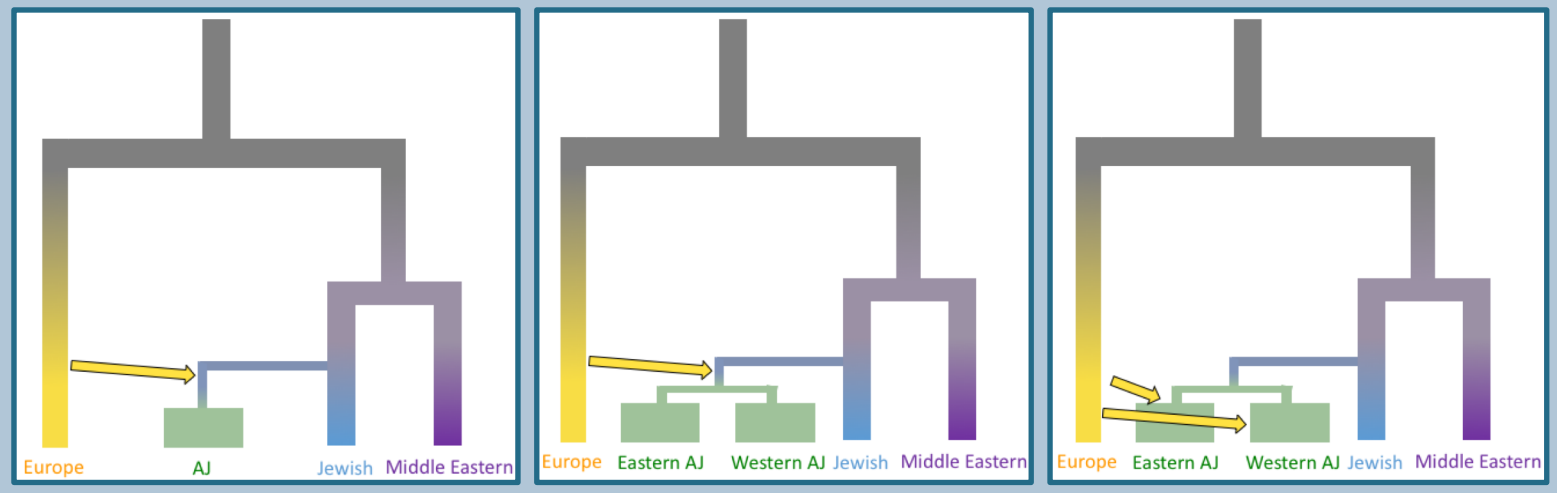
INFERENCE OF RECENT DEMOGRAPHIC HISTORY OF A POPULATION ISOLATE USING SNP ARRAY AND WHOLE GENOME DATA

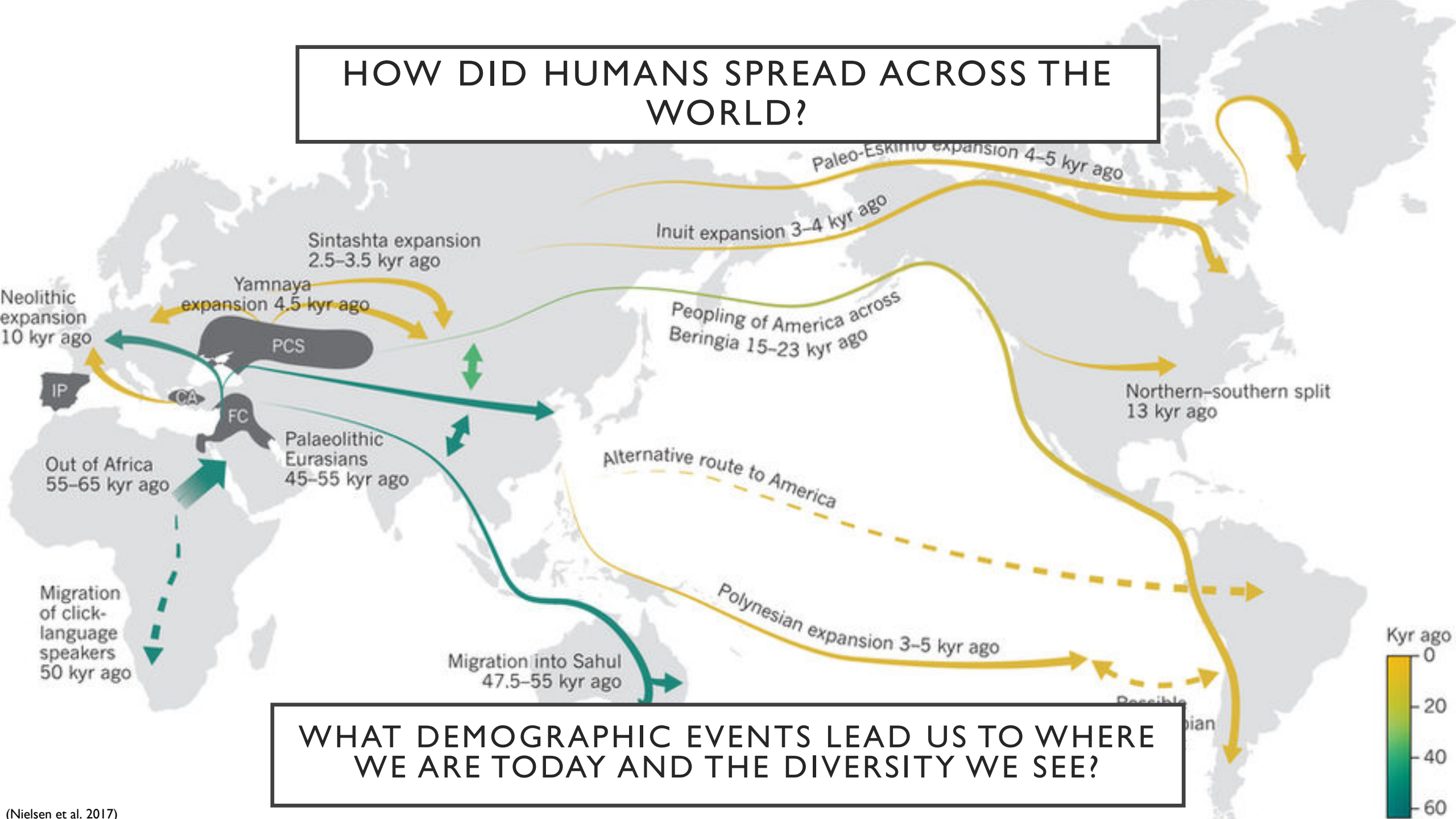


Ariella Gladstein

Ecology and Evolutionary Biology

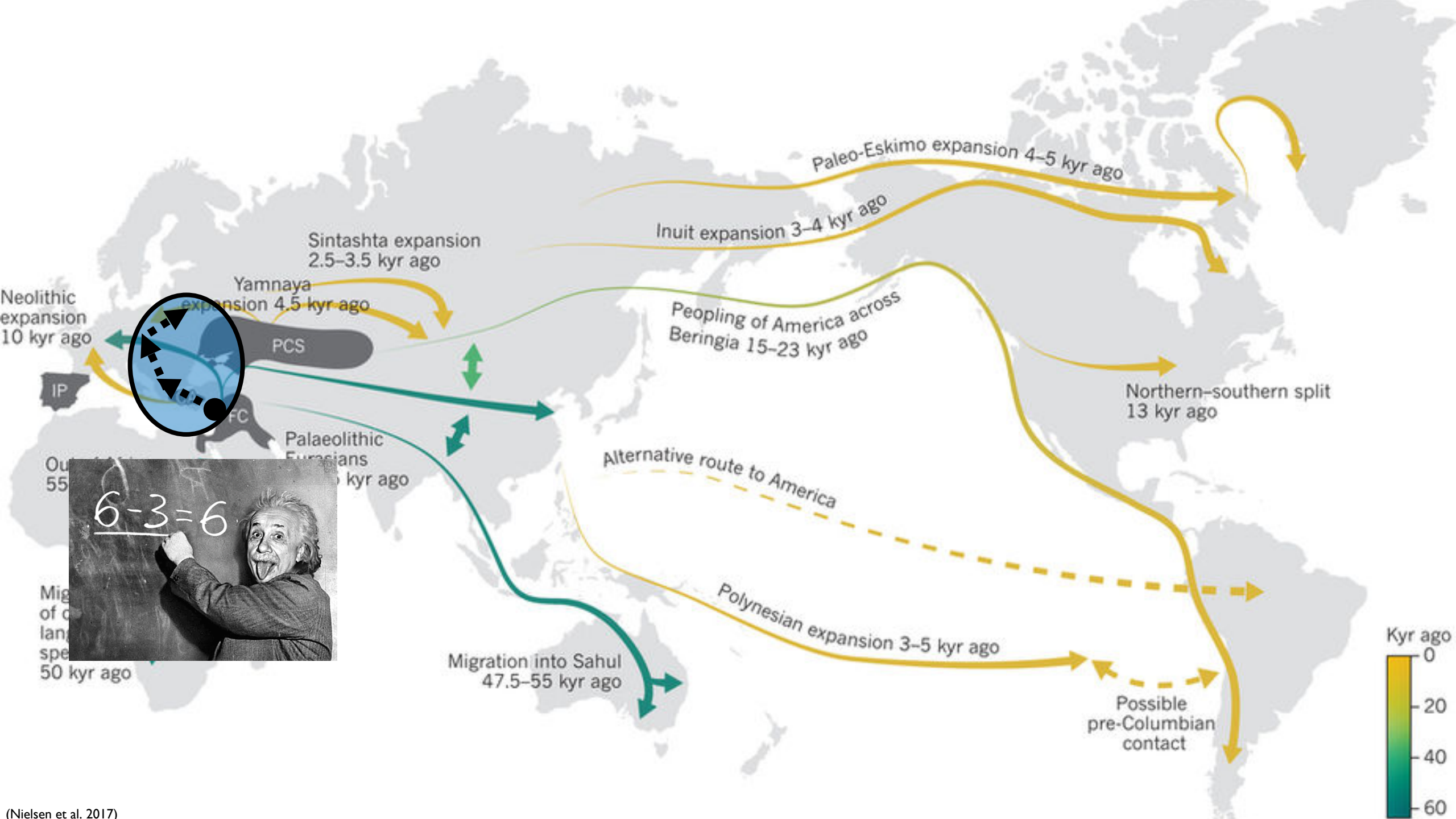
The University of Arizona

HOW DID HUMANS SPREAD ACROSS THE WORLD?

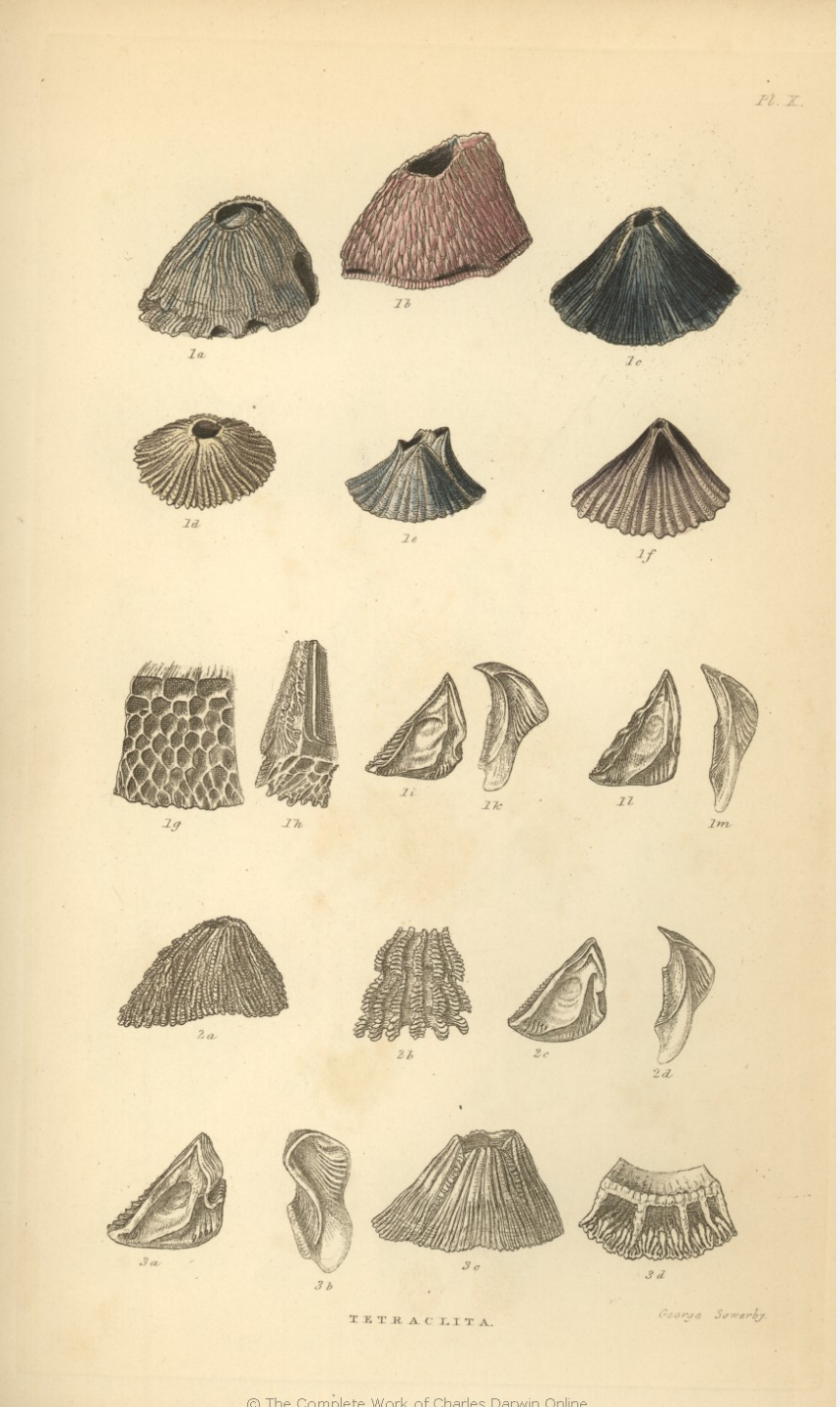


WHAT DEMOGRAPHIC EVENTS LEAD US TO WHERE WE ARE TODAY AND THE DIVERSITY WE SEE?

(Nielsen et al. 2017)

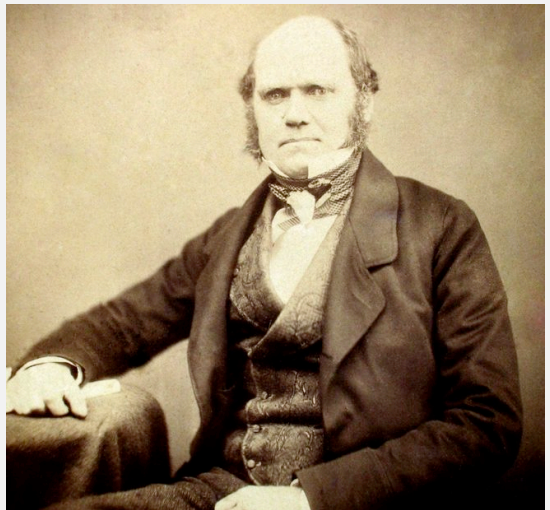


(Nielsen et al. 2017)



BASIC RESEARCH IS IMPORTANT!

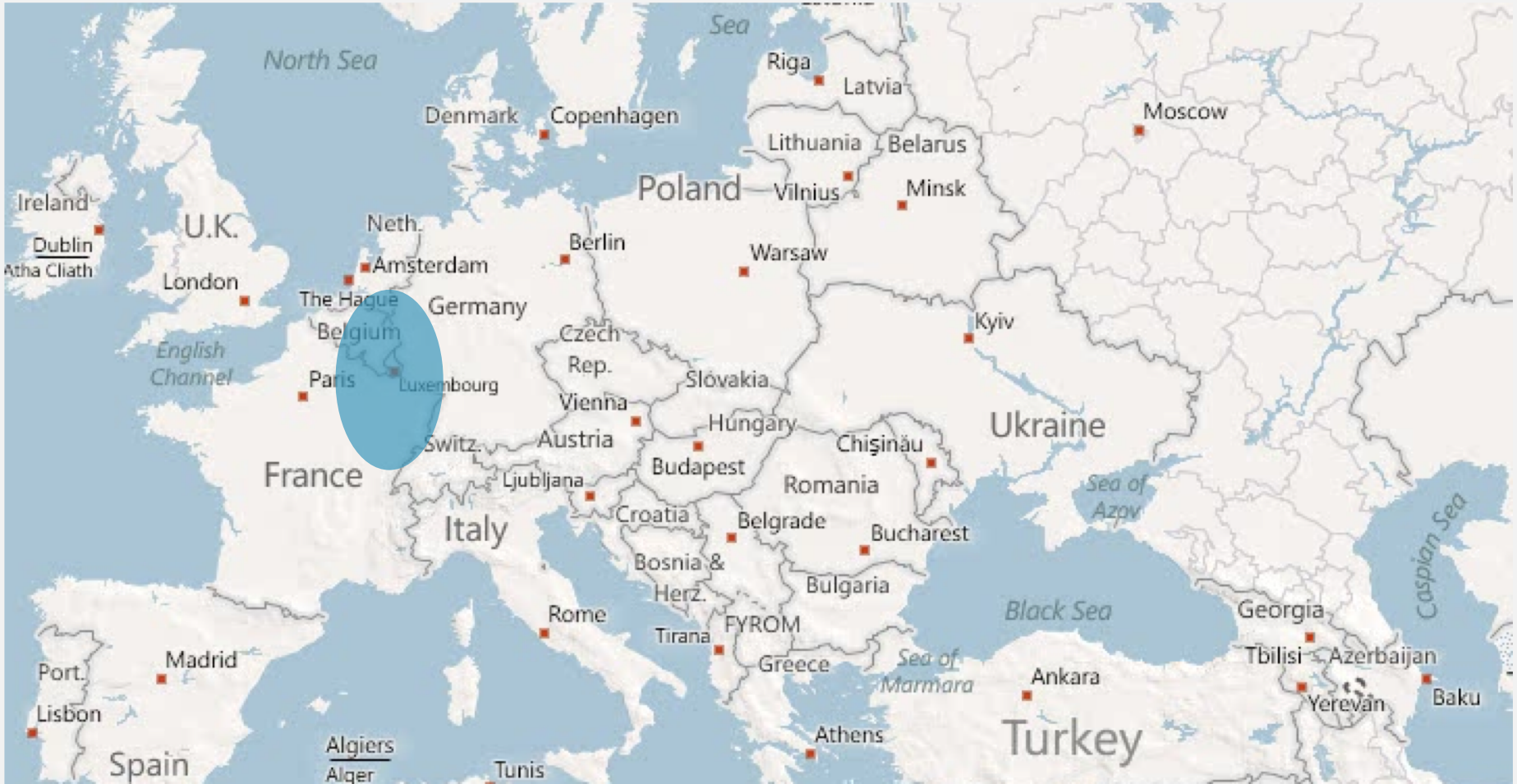
- My research focuses on a small part of basic evolutionary biology questions.
- Huge computational resources and modern techniques to contribute to basic evolution questions



THEMES OF DISSERTATION

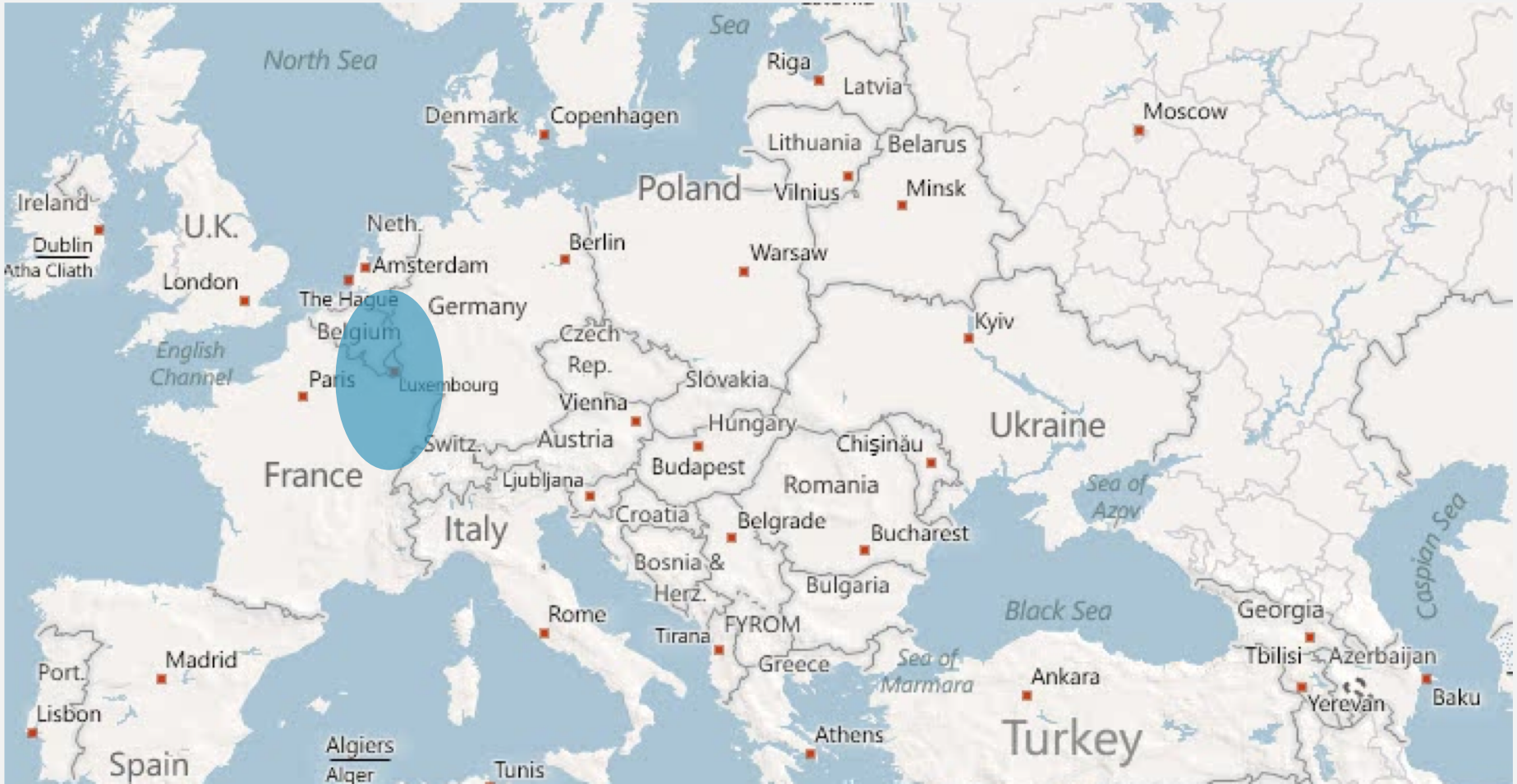
- Detection of runs of homozygosity from SNP arrays
  - Improving identification of runs of homozygosity (Ch. 2)
  - Correcting ascertainment bias in runs of homozygosity (App. C)
- Scaling up Approximate Bayesian Computation for whole chromosomes
  - Create efficient pipeline to simulate demographic models and calculate summary statistics (App. A)
  - Create generalized high throughput workflow (Ch. 4)
- Infer history of the Ashkenazi Jews
  - Substructure in AJ? (Ch. 5)
  - Khazarian origin? (App. B)

WHO ARE THE ASHKENAZI JEWS?



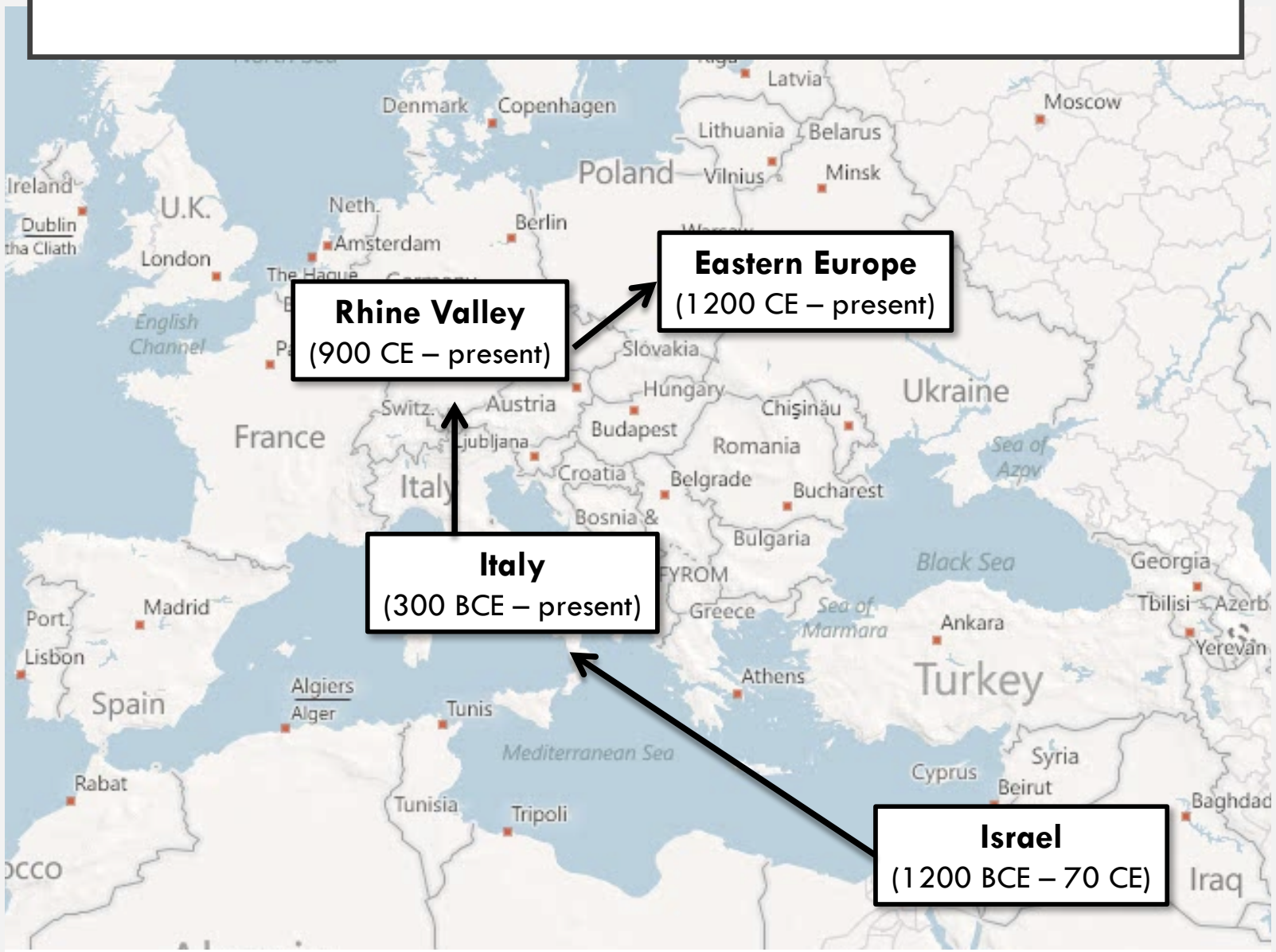
Culturally, religiously, and linguistically identify as Jews whose ancestors came from the Rhine Valley.

ASHKENAZI JEWS: AN INTERESTING STUDY POPULATION

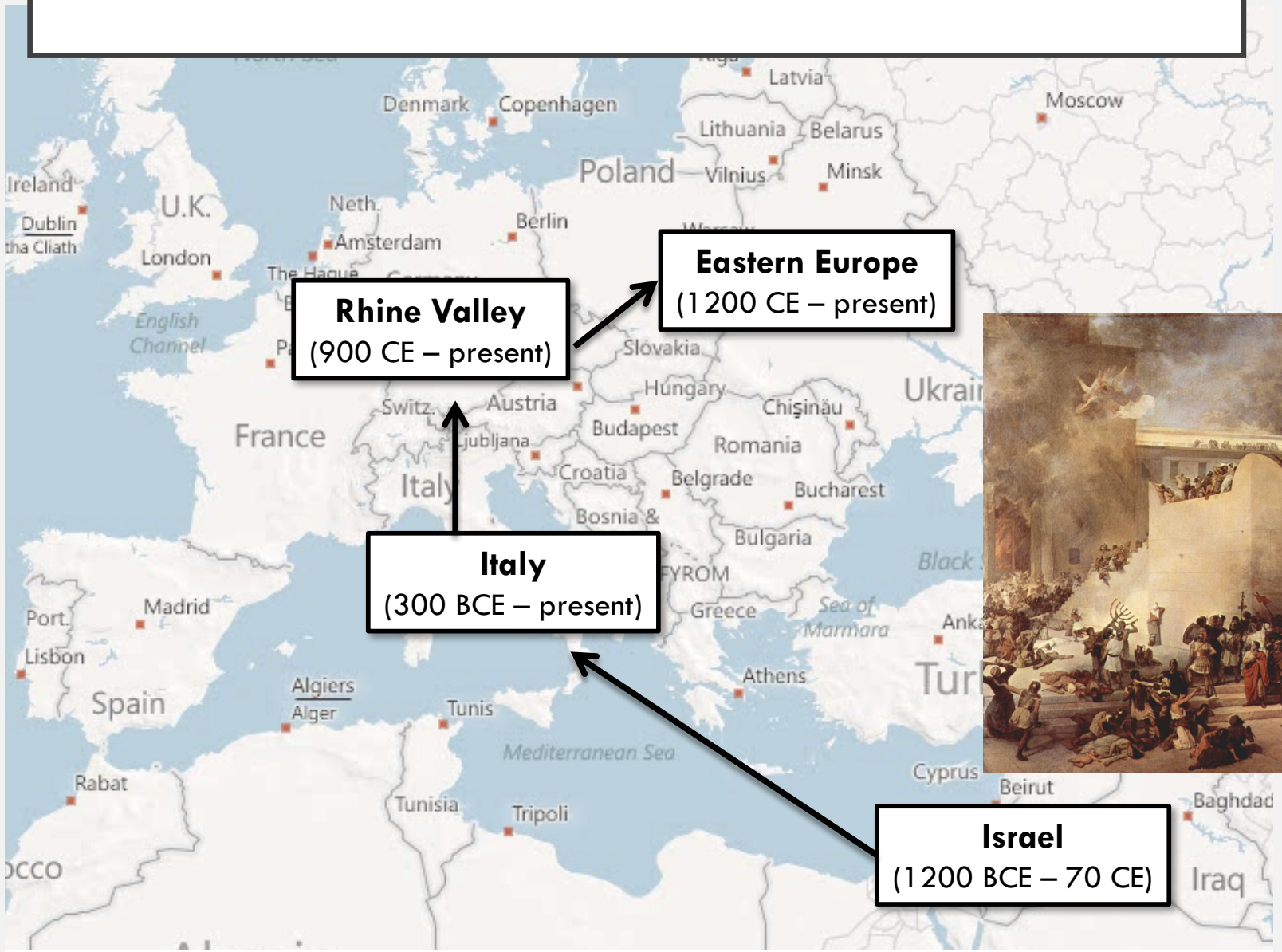


- High frequency of genetic disorders
- Population isolate
- Complex demographic history
- Well documented historical record

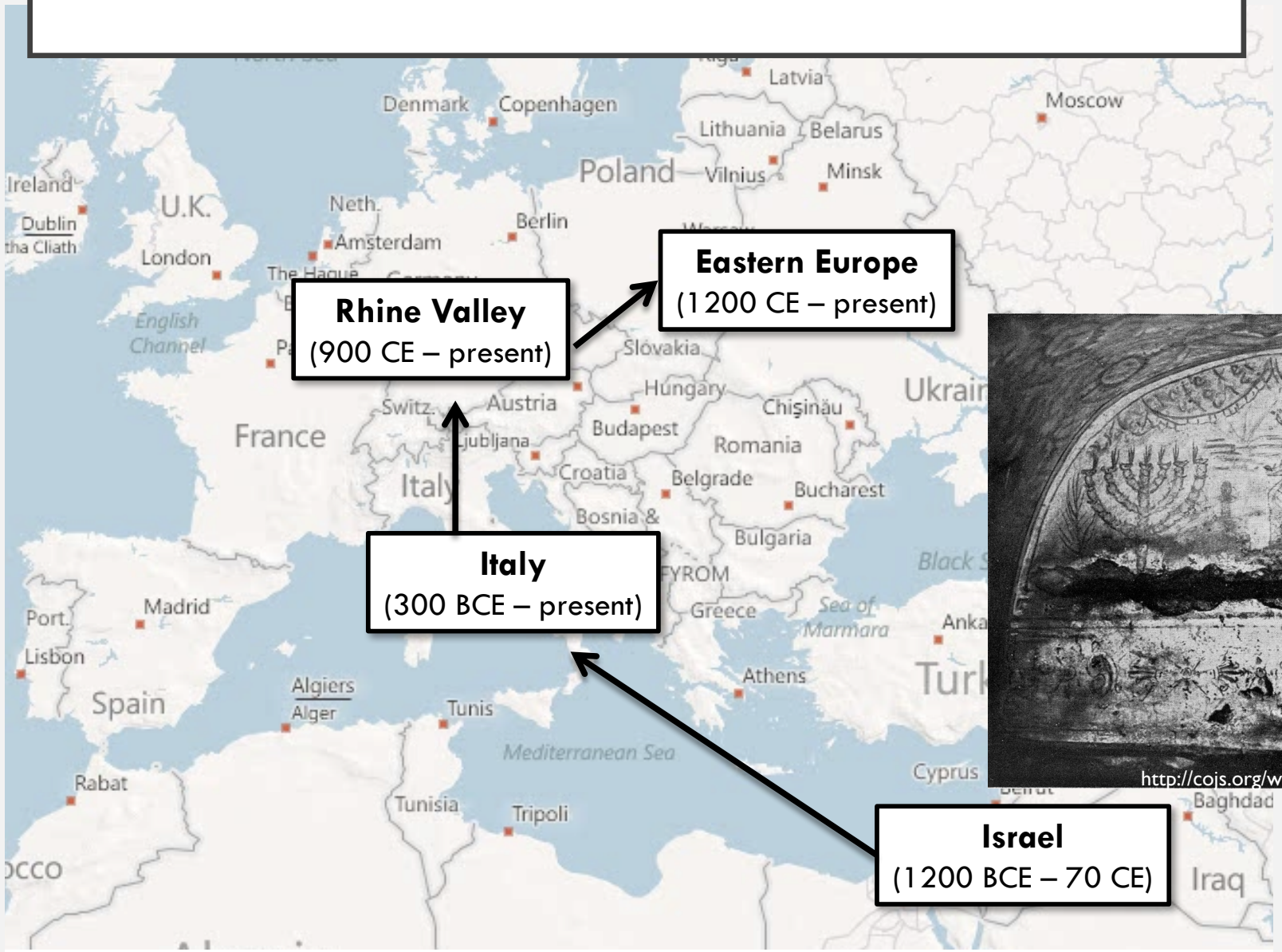
HYPOTHESIS OF ASHKENAZI ORIGINS



HYPOTHESIS OF ASHKENAZI ORIGINS



HYPOTHESIS OF ASHKENAZI ORIGINS



http://cojs.org/wp-content/uploads/2015/11/Torlonia.jpg

HYPOTHESIS OF ASHKENAZI ORIGINS

https://fotoeins.com/2016/08/24/shum-cities-schum-staedte-speyer-worms-mainz/

Rhine Valley
(900 CE – present)

Eastern Europe
(1200 CE – present)

Italy
(300 BCE – present)

Israel
(1200 BCE – 70 CE)

HYPOTHESIS OF ASHKENAZI ORIGINS

https://www.historyextra.com/period/medieval/the-jews-of-medieval-england/

Rhine Valley
(900 CE – present)

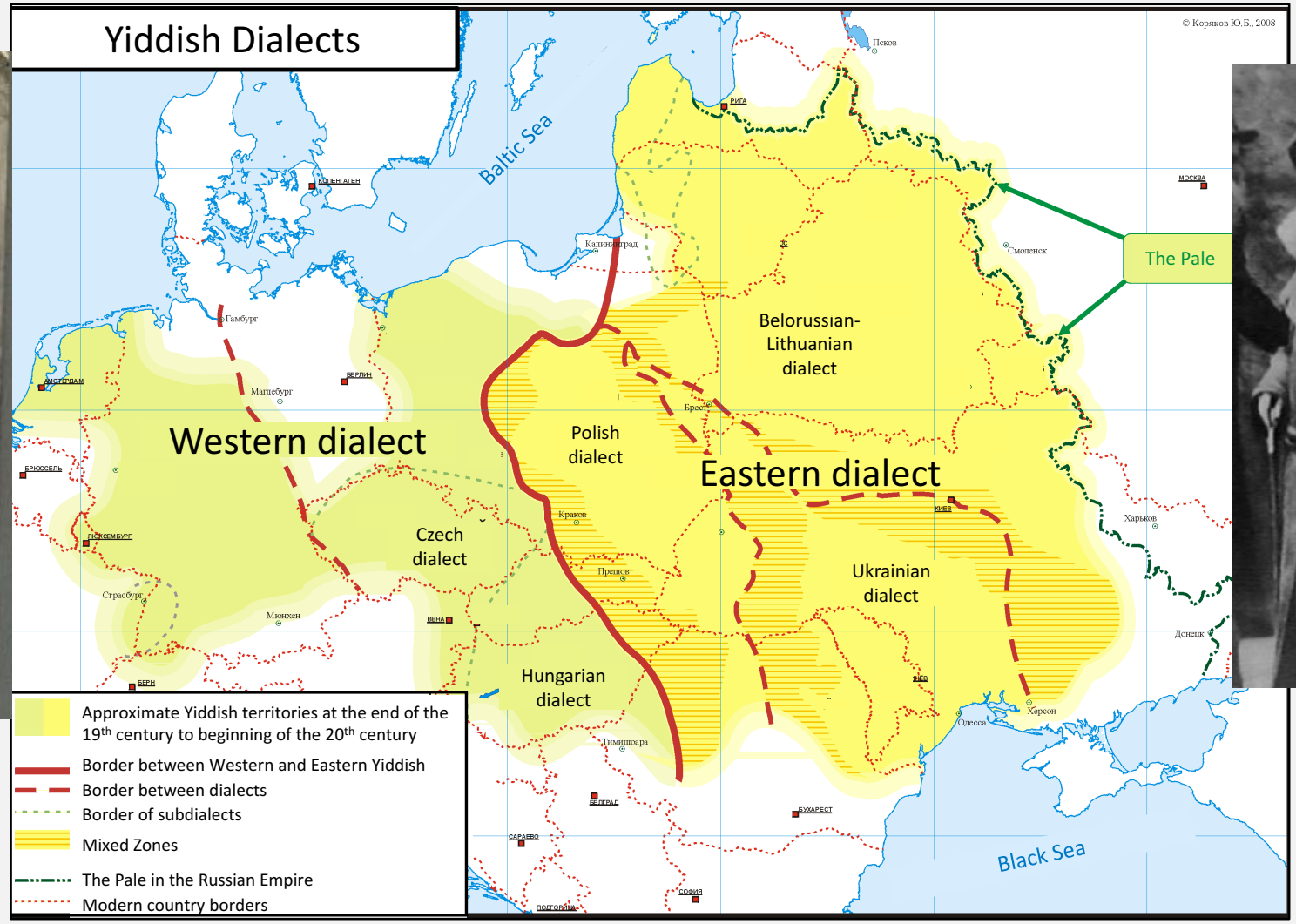
Eastern Europe
(1200 CE – present)

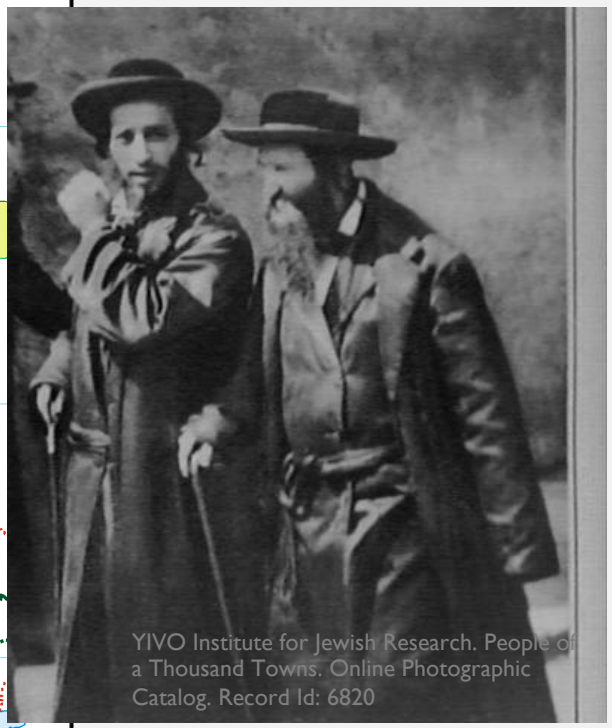
Italy
(300 BCE – present)

Israel
(1200 BCE – 70 CE)

WESTERN VS. EASTERN ASHKENAZI JEWS

Germany, 1900's

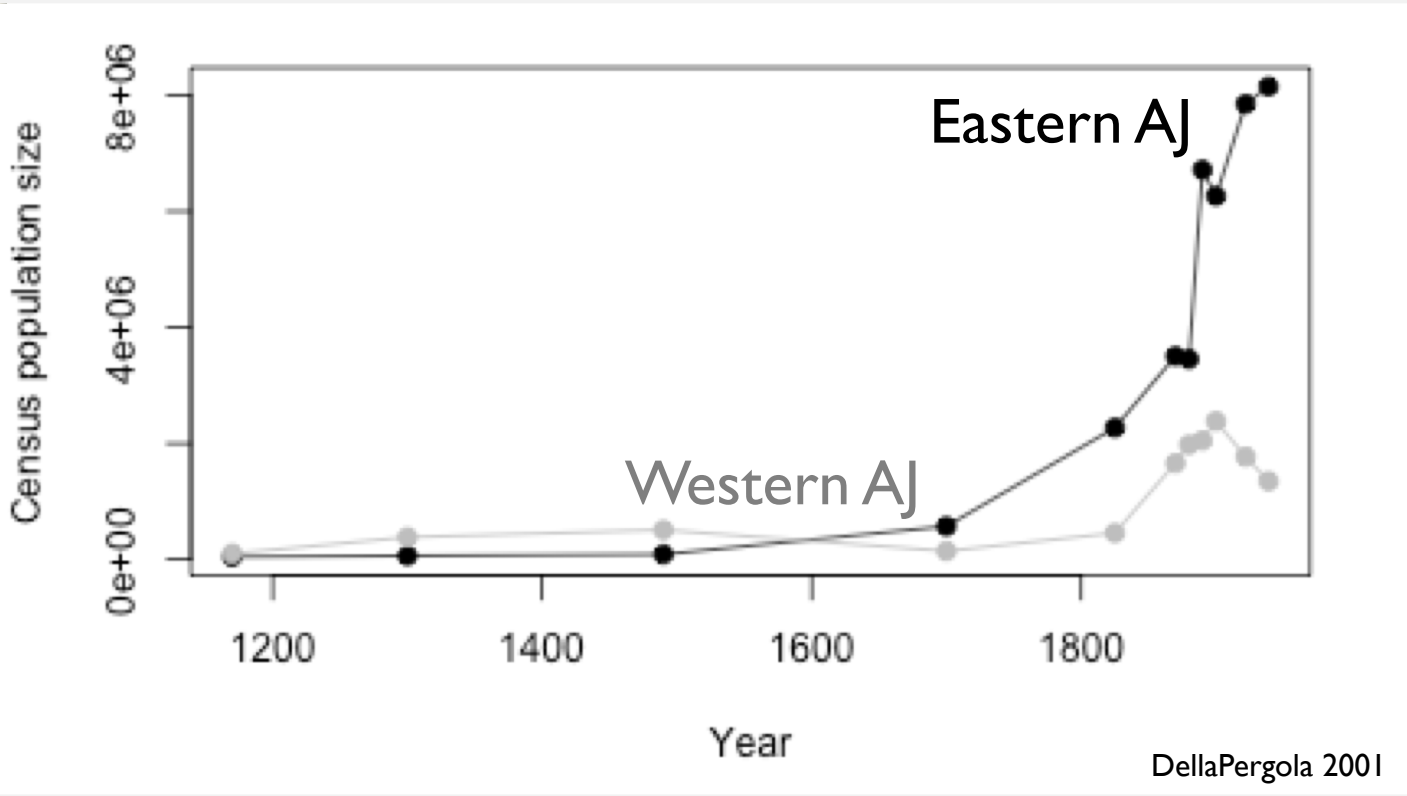




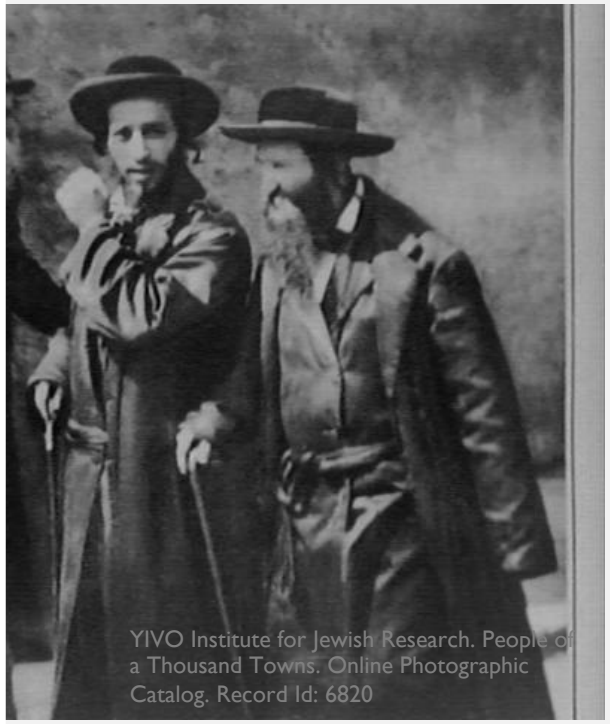
Cracow, Poland. 1932

WESTERN VS. EASTERN ASHKENAZI JEWS

Germany, 1900's



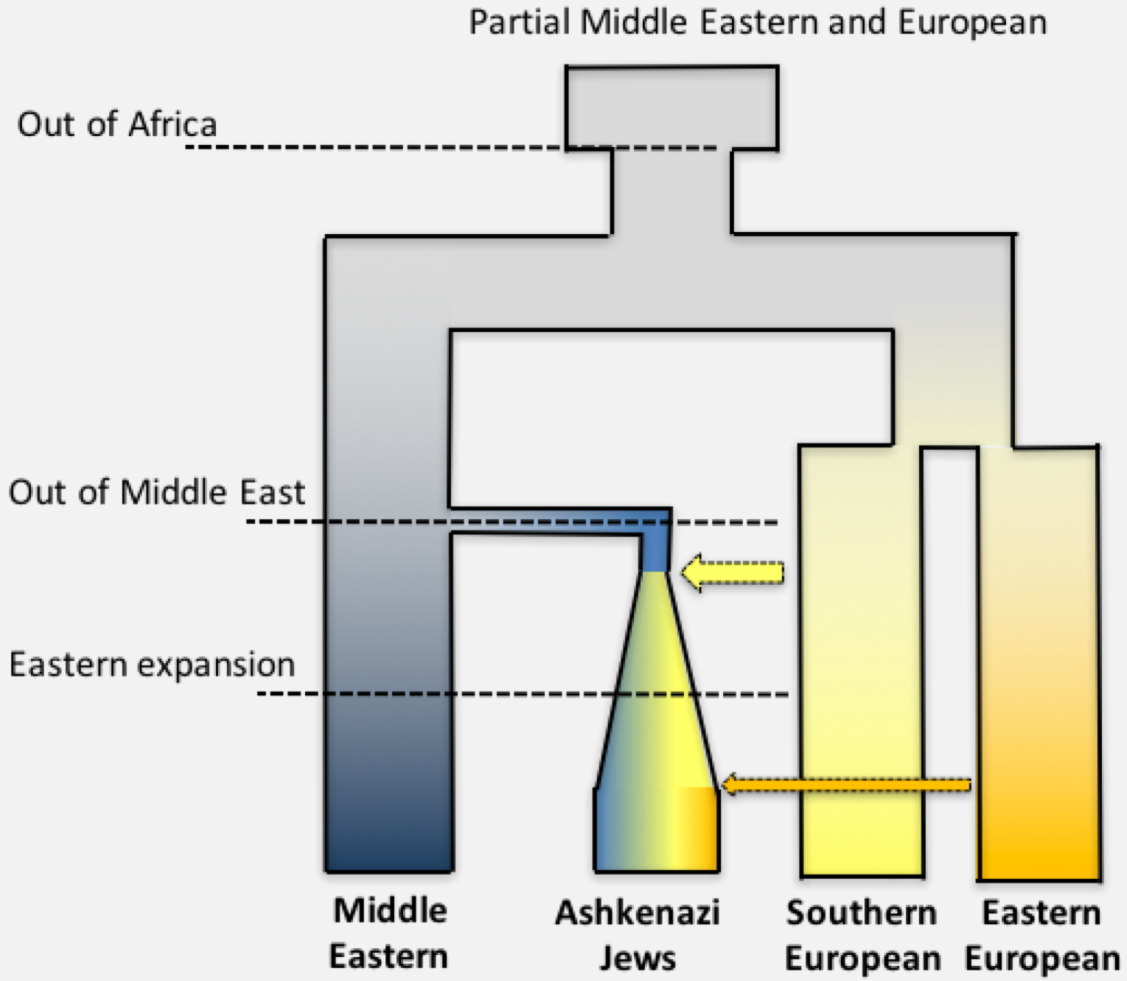
DellaPergola 2001



Cracow, Poland. 1932

PREVIOUS STUDIES

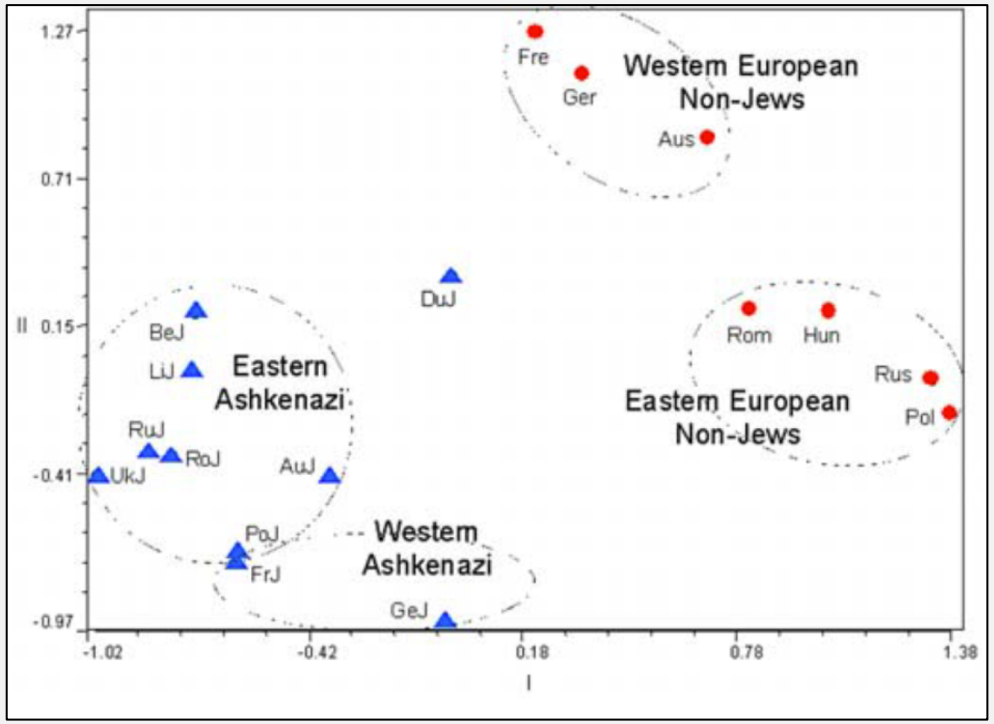
- Most closely related to other Jewish populations
- Middle Eastern origin
- Bottlenecks
  - Population reduction ~30 gen ago
  - Exponential growth ~30 gen ago
- At least 2 admixture events
  - Southern Europeans ~35-60%, ~25-50 gen ago
  - Eastern Europeans ~15-25%, ~10-20 gen ago



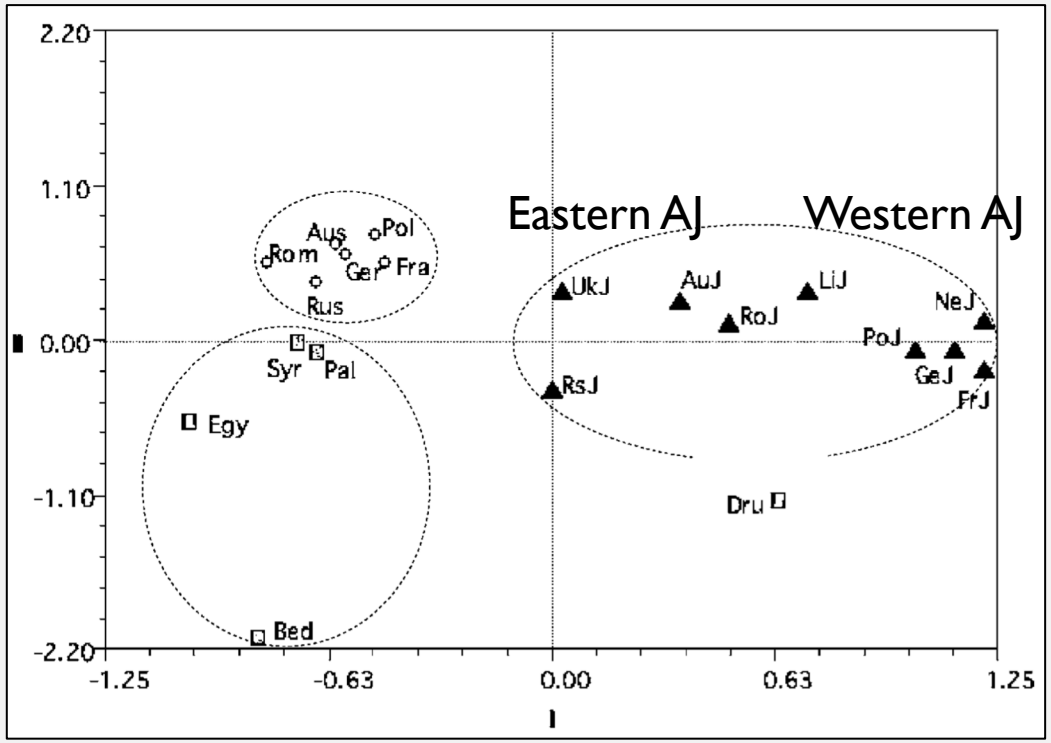
Behar et al. 2010; Palamara et al. 2012; Carmi et al. 2014; Xue et al. 2016

PREVIOUS STUDIES: AJ SUBSTRUCTURE

Y chromosome and mtDNA markers show differences among AJ from different countries



Behar et al. 2004a



Behar et al. 2004b

PREVIOUS STUDIES: AJ SUBSTRUCTURE

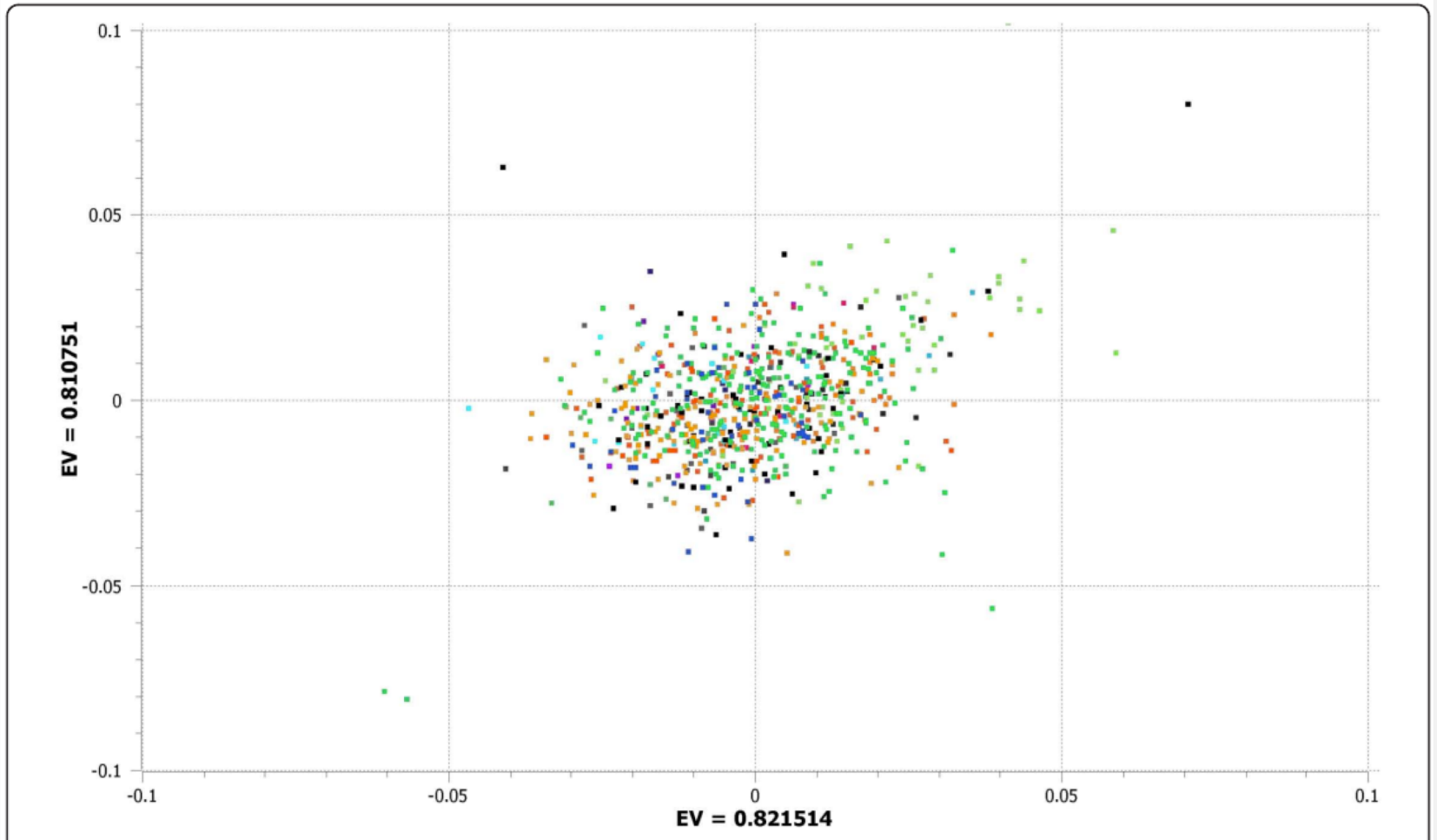
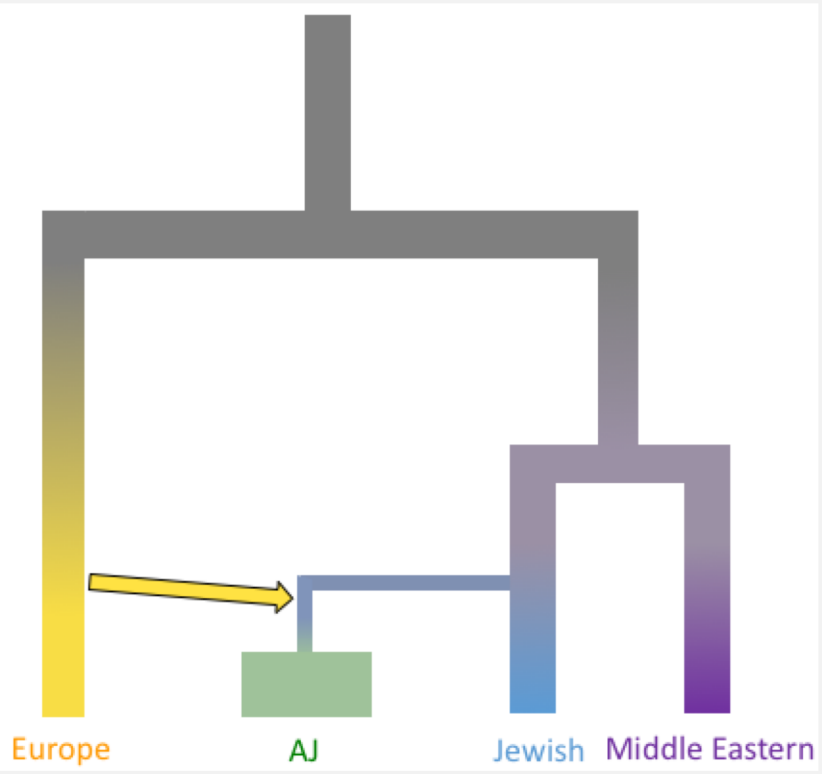
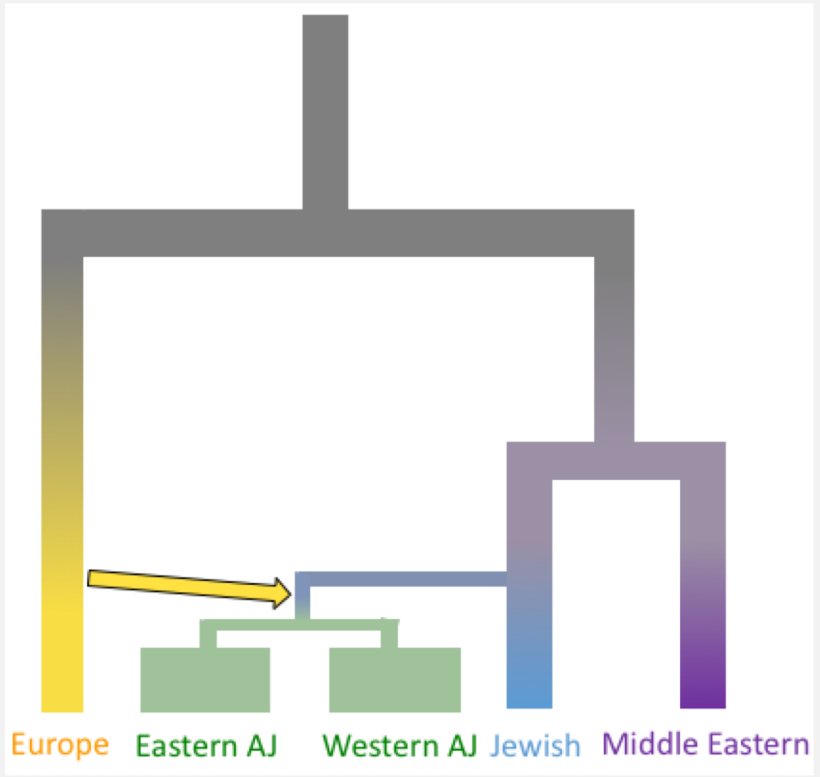


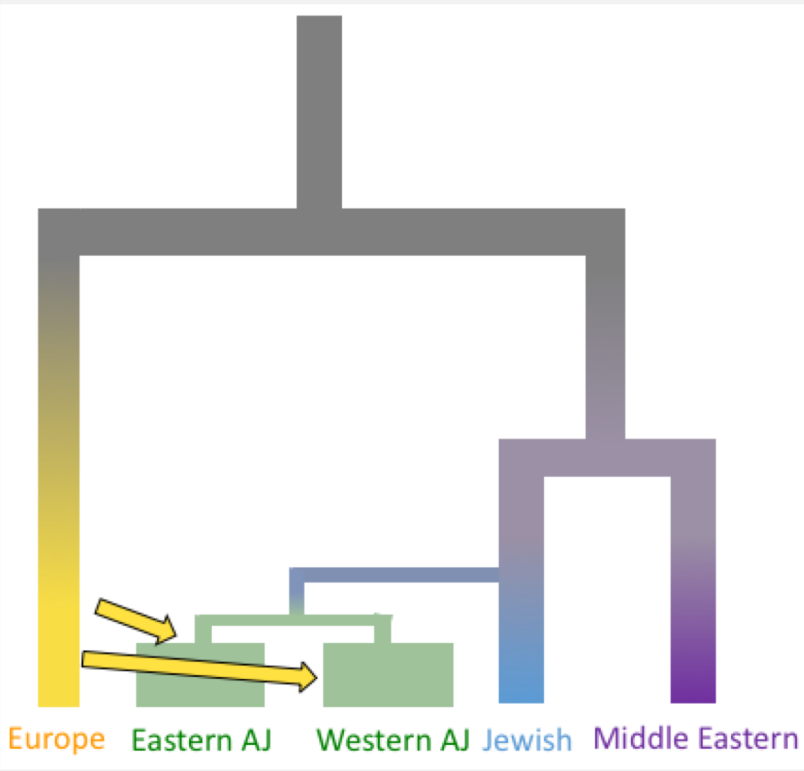
Figure 6 Intra-population principal components analysis of 1,312 Ashkenazi Jewish (AJ) individuals with cluster 3 (C3) scores > 0.6 derived from ADMIXTURE analysis. The x-axis represents the eigenvalue (EV) for principal component 1 (PC1) and the y-axis represents the eigenvalue for principal component 2 (PC2). Different colors represent different geographical origins of AJ individuals.

No intra-population structure found with SNP array data

Guha et al. 2012







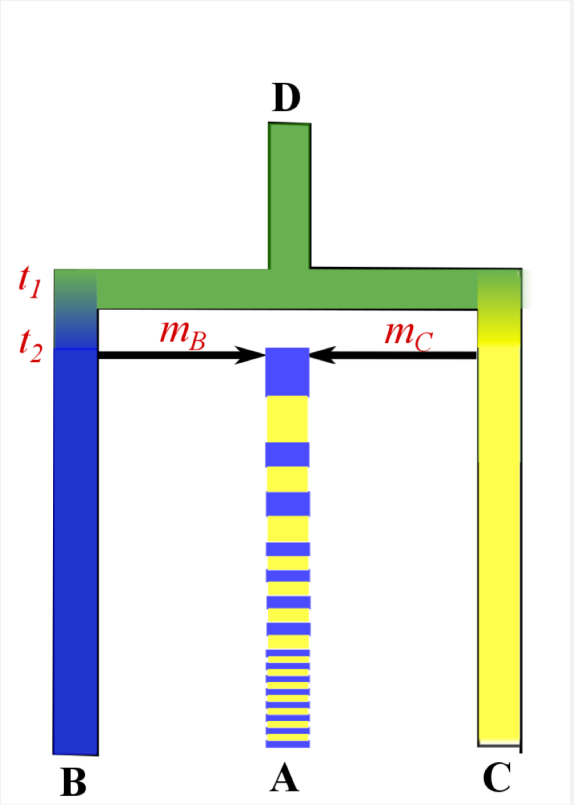
MODELS OF ASHKENAZI HISTORY

THEMES OF DISSERTATION

- Detection of runs of homozygosity from SNP arrays
  - Improving identification of runs of homozygosity (Ch. 2)
  - Correcting ascertainment bias in runs of homozygosity (App. C)
- Scaling up Approximate Bayesian Computation for whole chromosomes
  - Create efficient pipeline to simulate demographic models and calculate summary statistics (App. A)
  - Create generalized high throughput workflow (Ch. 4)
- Infer history of the Ashkenazi Jews
  - Substructure in AJ? (Ch. 5)
  - Khazarian origin? (App. B)

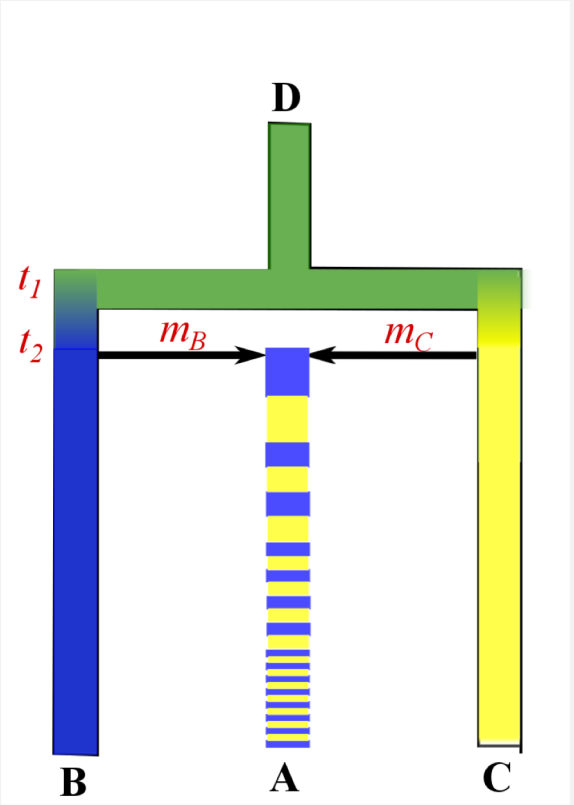
HAPLOTYPES CAN BE USED TO INFER HISTORY

- Over time, recombination breaks up segments of the genome in predictable ways
  - Segments identical by descent (IBD)
    gene flow, effective size, relatedness
  - Runs of homozygosity (ROH)
    effective size, relatedness, random mating
  - Linkage disequilibrium blocks (LD)
    gene flow, effective size
  - Ancestry blocks
    gene flow



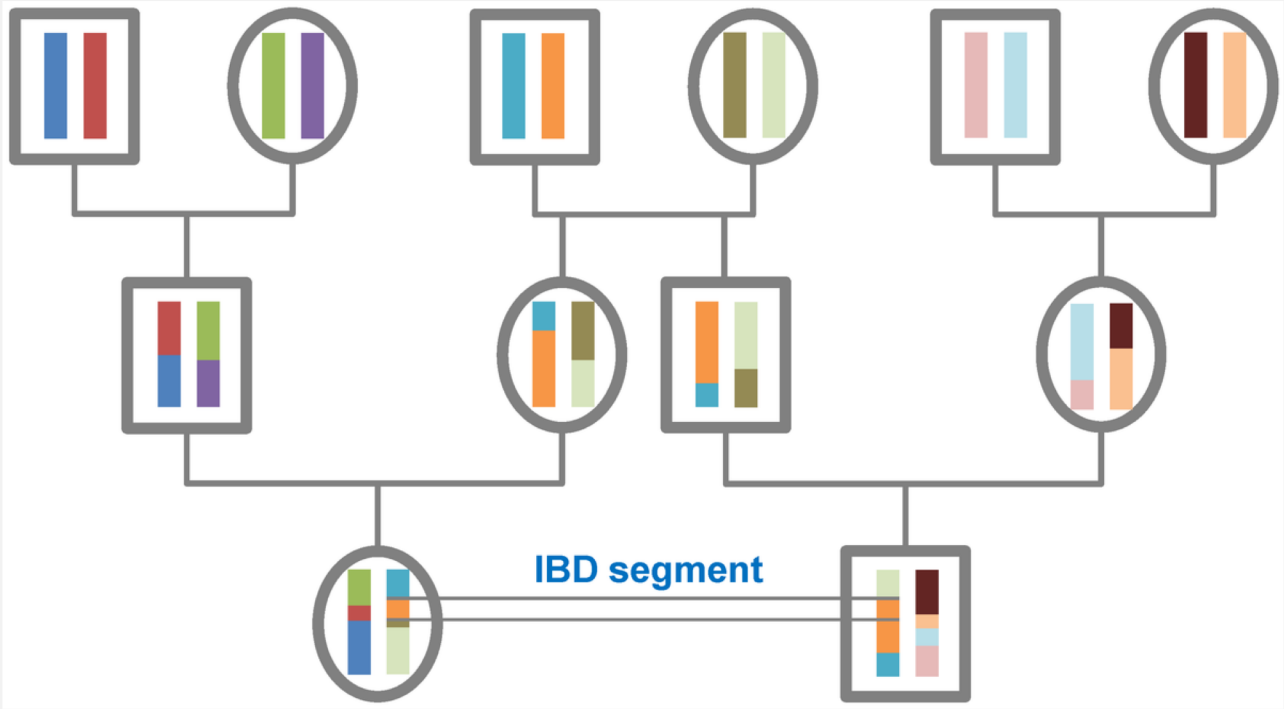
HAPLOTYPES CAN BE USED TO INFER HISTORY

- Over time, recombination breaks up segments of the genome in predictable ways
  - Segments identical by descent (IBD)
    gene flow, effective size, relatedness
  - Runs of homozygosity (ROH)
    effective size, relatedness, random mating
  - Linkage disequilibrium blocks (LD)
    gene flow, effective size
  - Ancestry blocks
    gene flow



SEGMENTS IDENTICAL BY DESCENT (IBD)

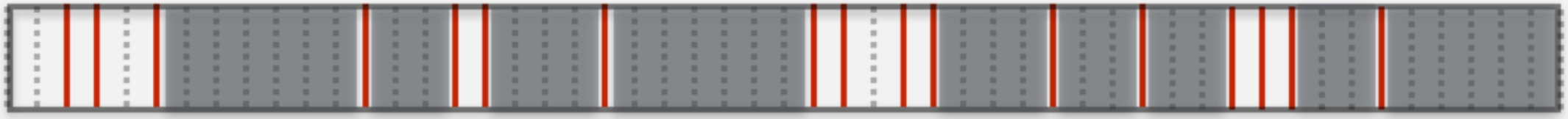
A region on two chromosomes that was inherited from a common ancestor



- Every site is (technically) IBD
- Practically, we define IBD based on a minimum length

RUNS OF HOMOZYGOSITY (ROH)

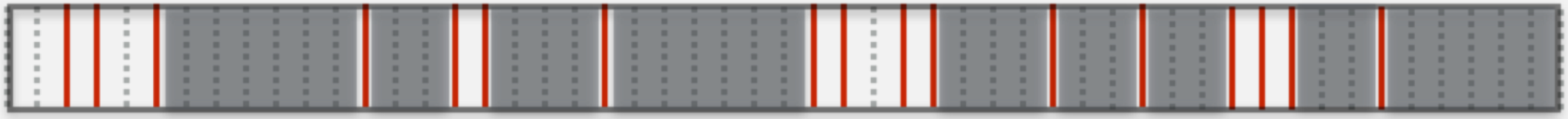
- A ROH is a genomic segment of continuous homozygous sites.
- ROH are defined based on a minimum length



ROH on a chromosome. Each dotted grey line is a homozygous site and each red line is a heterozygous site. Each shaded grey area is a ROH

RUNS OF HOMOZYGOSITY (ROH)

- A ROH is a genomic segment of continuous homozygous sites.
- ROH are defined based on a minimum length



ROH on a chromosome. Each dotted grey line is a homozygous site and each red line is a heterozygous site. Each shaded grey area is a ROH

ROH reflect relatedness of ancestors

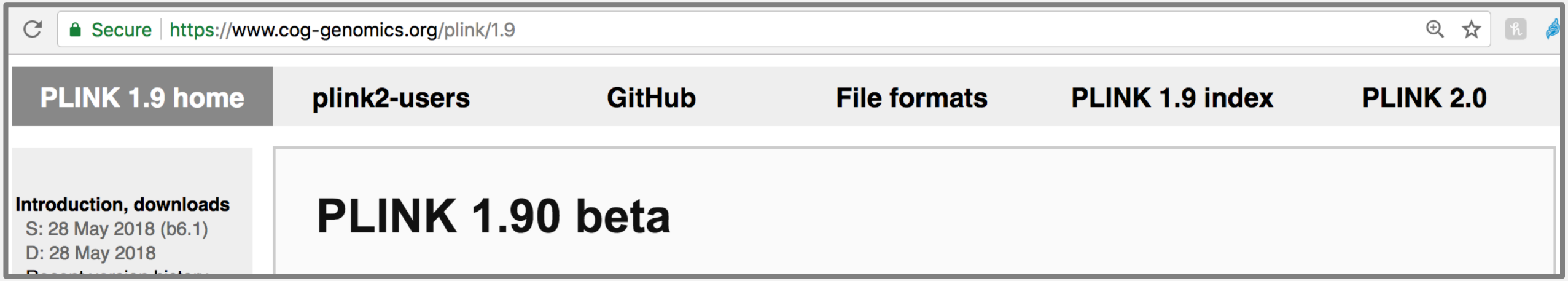
Smaller Ne increases likelihood of creating ROH

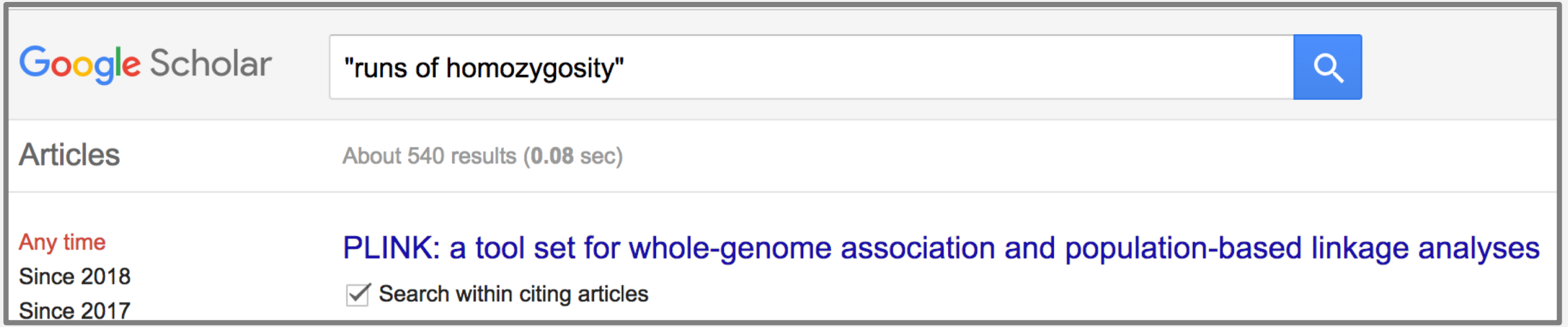
SINGLE NUCLEOTIDE POLYMORPHISM (SNP) ARRAYS

- Genome-wide and many SNPs (100 K's – millions)
- Benefits:
  - Inexpensive
  - Low genotyping error rates
  - Easy to work with
- Disadvantages:
  - Ascertainment bias – reduction of represented genetic diversity

HOW DOES BIAS FROM SNP ARRAYS AFFECT HAPLOTYPE STATISTICS?

- Extensive work on the effect of ascertainment bias on the allele frequency spectrum.
- Haplotype statistics considered to be less sensitive to ascertainment bias.


Secure
https://www.cog-genomics.org/plink/1.9
PLINK 1.9 home
plink2-users
GitHub
File formats
PLINK 1.9 index
PLINK 2.0
Introduction, downloads
S: 28 May 2018 (b6.1)
D: 28 May 2018
PLINK 1.90 beta


Google Scholar
"runs of homozygosity"
Articles
About 540 results (0.08 sec)
Any time
Since 2018
Since 2017
PLINK: a tool set for whole-genome association and population-based linkage analyses
Search within citing articles

PLINK 1.90 beta

Runs of homozygosity

```
--homozyg <group | group-verbose> <consensus-match> <extend> <subtract-1-from-lengths>
--homozyg-snp [min SNP count]
--homozyg-kb [min length]
--homozyg-density [max inverse density (kb/SNP)]
--homozyg-gap [max internal gap kb length]
--homozyg-het [max hets]
--homozyg-window-snp [scanning window size]
--homozyg-window-het [max hets in scanning window hit]
--homozyg-window-missing [max missing calls in scanning window hit]
--homozyg-window-threshold [min scanning window hit rate]
```

If any of these flags are present, a set of run-of-homozygosity reports is generated using PLINK 1.07's scanning algorithm. See the original documentation for more details.

Colorectal cancer risk is not associated with increased levels of homozygosity in Saudi Arabia

Abdul K. Siraj, PhD[1], Hanif G. Khalak, PhD[2], Mehar Sultana, MSc[1], Maha
Prashant Bavi, MD[1], Nasser Al-Sanea, MD[3], Fouad Al-Dayel, MD[4], Shah
Fowzan S. Alkuraya, MD, FACMG[5], Khawla S. Al-Kuraya, MD,

harvard.edu/purcell/plink/) was used, with default parameters
except for minimum length of ROH, minimum number of SNPs
per ROH, and maximum number of heterozygous SNPs per

Sardinians Genetic Background Explained by Runs of Homozygosity and Genomic Regions under Positive Selection

a Francesca Ortu[3], Fabio Rosa[2], Simonetta Guarrera[2],
[4,5], Cristina Barlassina[4,5], Chiara Troffa[3],
Fresu[3], Nicola Glorioso[3], Alberto Piazza[1,2],

oftware (*–homozyg* option)). The following

Inbreeding and homozygosity in breast cancer survival

Hauke Thomsen[1], Miguel Inacio da Silva Filho[1], Andrea Woltmann[1], Robert Johansson[2],
Jorunn E. Eyfjörd[3], Ute Hamann[4], Jonas Ma
Roger Henriksson[2,8], Stefan Herms[9,10], Per
Kari Hemminki[1,11], Per Lenner[2] & Asta Först

Genomic inbreeding estimation in small populations: evaluation of runs of homozygosity in three local dairy cattle breeds

S. Mastrangelo[1†], M. Tolone[1], R. Di Gerlando[1], L. Fontanesi[2], M. T. Sardina[1] and B. Portolano[1]

The Association of Genotype-Based Inbreeding Coefficient with a Range of Physical Human Traits

Karin J. H. Verweij[1,2], Abdel Abdellaoui[2], Juha Veijola[3], Sylvain Seb
Matthew C. Keller[6,7], Marjo-Riitta Järvelin[4,5,8,9,10], Brendan P. Zietsch

SNPs and, hence, to balance the number a
window. We set the remaining options to
dow, thereby ensuring >90% positive-predictive value of
eters for "homozyg-snp" option according to our heuristic

following criteria were used to define the ROH: (i) the minimum number of SNPs included in the ROH was fixed to 40; (ii) the minimum length that constituted the ROH was set to 4 Mb; (iii) two missing SNPs were allowed in the ROH; (iv) minimum density of one SNP every 100 kb; (v) maximum gap between consecutive SNPs of 1 Mb. Moreover, the number of allowed heterozygous SNPs was set to different values: from one to three. Mean F_{ROH} values obtained

In this study we defined ROHs (based on
from Howrigan et al. [26], as stretches of at leas
homozygous SNPs (not allowing any heterozygo
pruned SNP data. To minimize underestimation
runs, three (approximately 5%) missing geno
otherwise unbroken homozygous segment were

Runs of Homozygosity in European Populations

Ruth McQuillan,[1] Anne-Louise Leutenegger,[2] Rehab Abdel-Rahman,[1,7] Christopher
Marijana Pericic,[3] Lovorka Barac-Lauc,[3] Nina Smolej-Narancic,[3] Branka Janicijevic,[3]
Albert Tenesa,[5] Andrew K. MacLeod,[6] Susan M. Farrington,[5] Pavao Rudan,[3] Carolin
Veronique Vitart,[7] Igor Rudan,[1,8,9] Sarah H. Wild,[1] Malcolm G. Dunlop,[5] Alan F.
Harry Campbell,[1] and James F. Wilson[1,*]

5000 kb (minimum 50 SNPs) across the genome to detect long contiguous runs of homozygous genotypes. An occasional genotyping error or missing genotype occurring in an otherwise-unbroken homozygous segment could result in the underestimation of ROHs. To address this, the program allows one heterozygous and five missing calls per window.

A threshold was set for the minimum length (kb) needed for a tract to qualify as homozygous. Because strong linkage disequilibrium (LD), typically extending up to about 100 kb, is common throughout the genome,[48–51] short tracts of homozygosity are very prevalent. For exclusion of these short and very common ROHs that occur in all individuals in all populations, the minimum length for an ROH was set at 500 kb. All empirical studies

Response to "Cross-Species Application of SNP Chip Suitable for Identifying Runs of Homozygosity" by S Miller, and Kardos

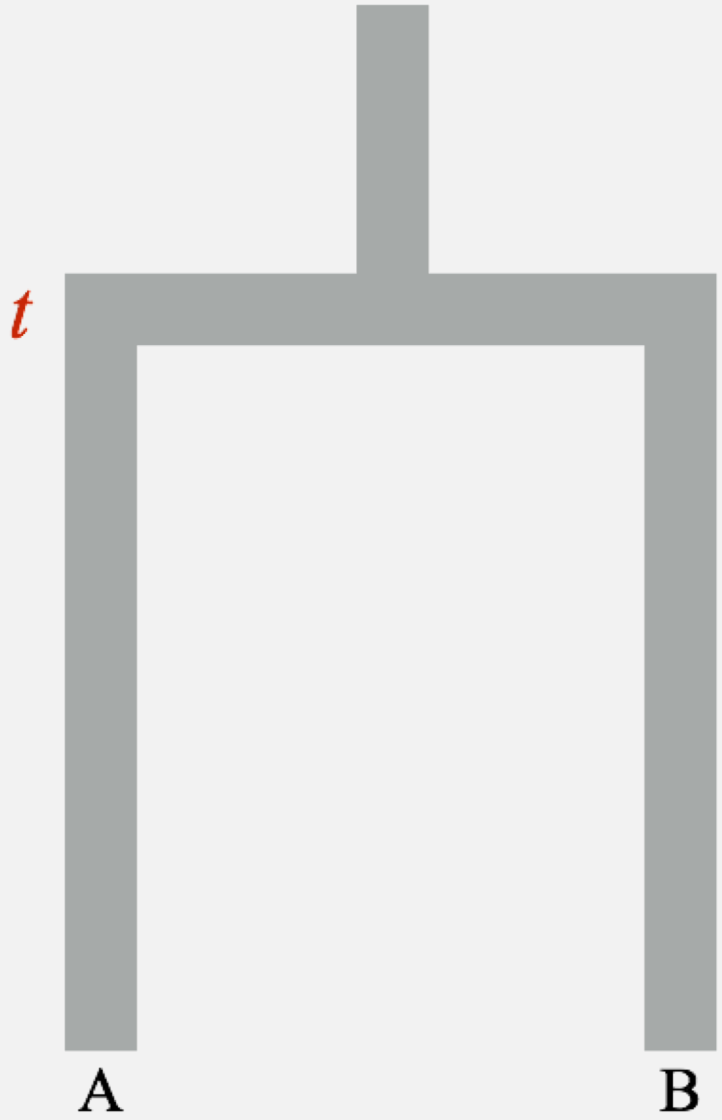
Veronika Kharzinova, Alexander A. Sermyagin, Elena A. Gladyr, Got
and Natalia A. Zinovieva

We would like to clarify that we used a sliding 100-kb window
a size of 100 SNPs to research ROH. In general, the window size
10 000 kb, not 100K SNPs as described in Shafer et al. By de
PLINK has a minimum density of 1 SNP/50 kb (Purcell et al. 2007).

mapping in a family presenting with epilepsy and hearing impairment

n Maclean[2], Muhammad Irfan[3], Farooq Naeem[4], Stephen Cass[2],
lter J Muir[1], Douglas HR Blackwood[1] and Muhammad Ayub[5]

ns of homozygosity' analytical tool set. Inspection of
mozygous tract lengths was limited to five or more
secutive SNPs, and low SNP densities in centromeric
ions were excluded. The length of homozygous regions
s taken to be from the most proximal to the most distal
mozygous SNP, and the programme allows one heterozygous SNP within this run. Marker positions are

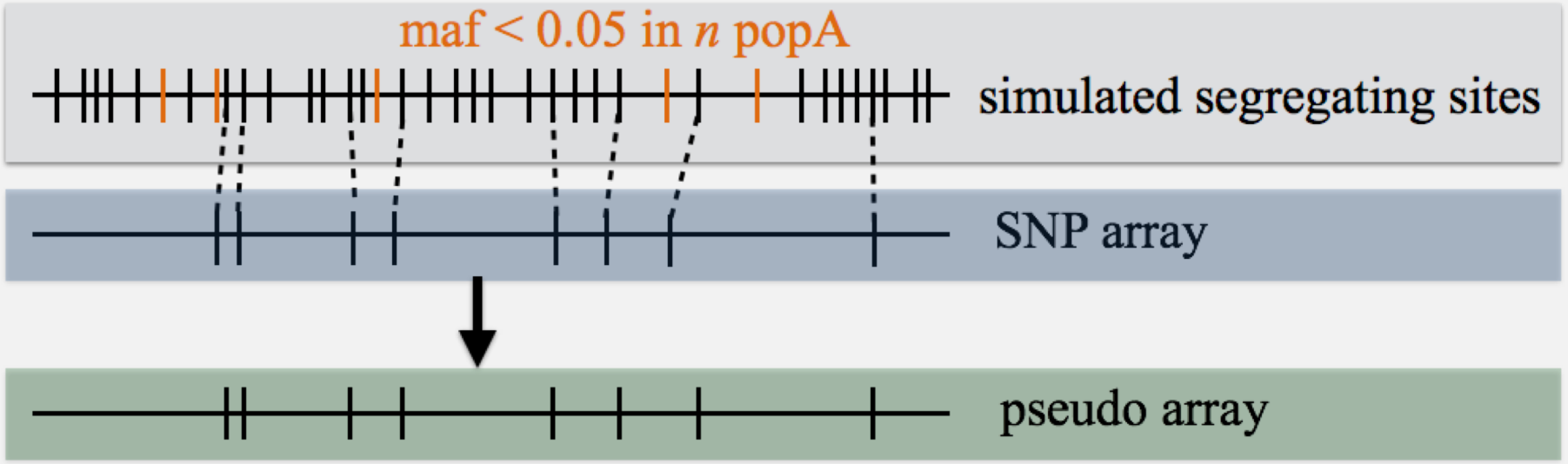


SIMULATE GENOME

- Coalescent simulation
- 100 iterations
- *Ne* = 1000
- Random *t*, such that *Fst* = [0,0.2]

CREATE PSEUDO ARRAY

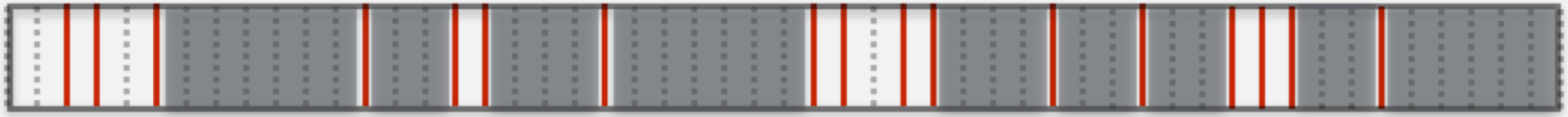
Use samples from population *A* to make pseudo array



FIND ROH

- **Genome ROH**: Script that finds pure ROH longer than *k*.

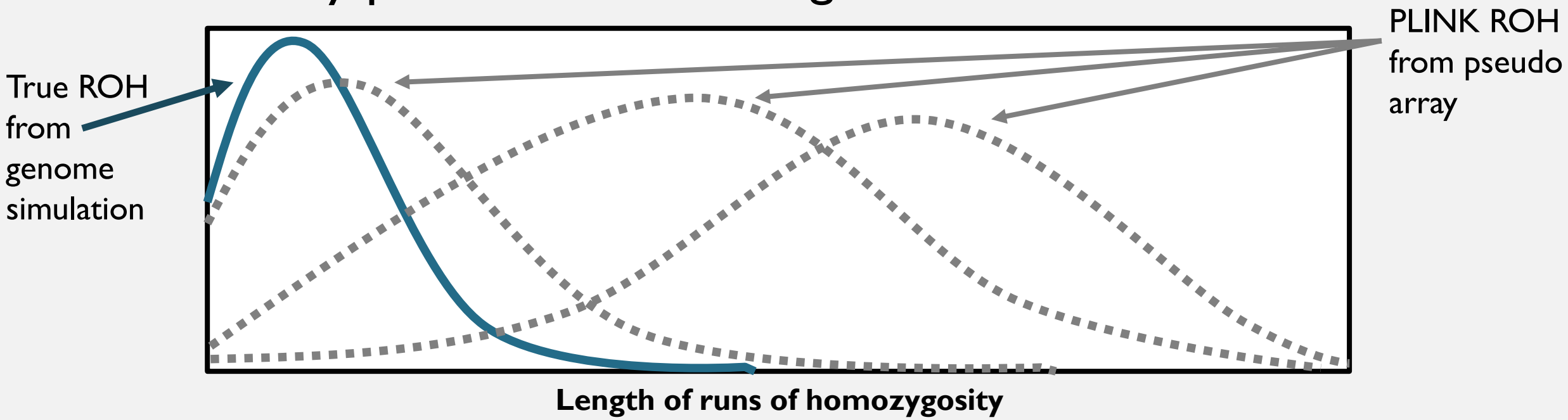
- **Pseudo array ROH:**
  - Script that finds pure ROH longer than *k*.
  - PLINK 1.09 (program used for SNP array data)



ROH on a chromosome. Each dotted grey line is a homozygous site and each red line is a heterozygous site. Each shaded grey area is a ROH

OPTIMIZE PLINK PARAMETERS

- Run PLINK on pseudo array with grid search of parameters (6,561 parameter sets)
- Identify parameter sets that give ROH closest to true ROH



IDENTIFYING CLOSEST ROH TO GENOME ROH

Mean ROH length

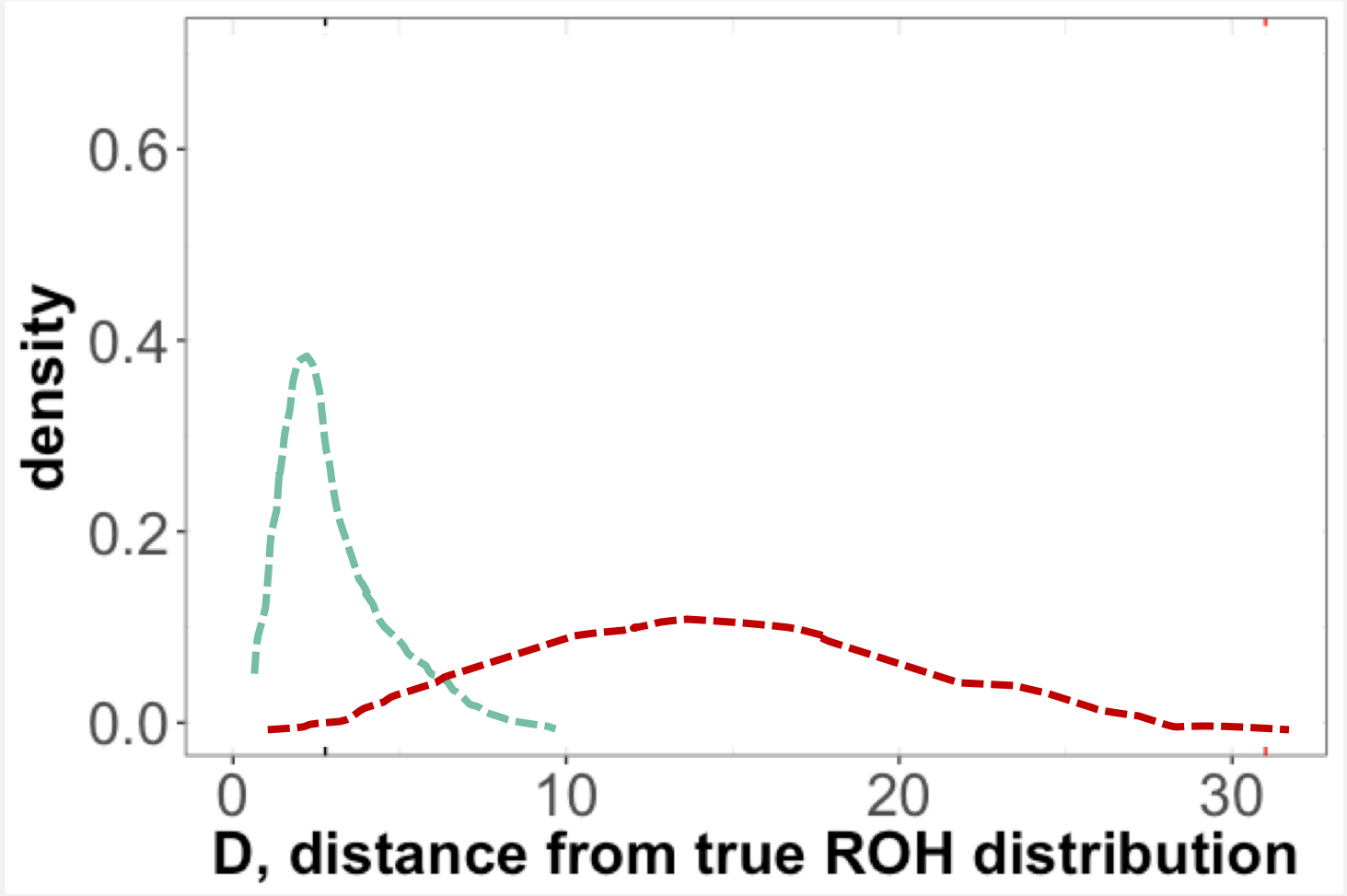
Variance of ROH length

Number of ROH

$$D = \left[\ln\left(\frac{\bar{\mathbf{x}}}{\bar{\mathbf{y}}}\right)\right]^2 + \left[\ln\left(\frac{\mathrm{var}(\mathbf{x})}{\mathrm{var}(\mathbf{y})}\right)\right]^2 + \left[\ln\left(\frac{m}{n}\right)\right]^2,$$

where $\mathbf{x}$ is the set of m ROH found by plink and $\mathbf{y}$ is the set of n real ROH. $D = 0$ is ideal.

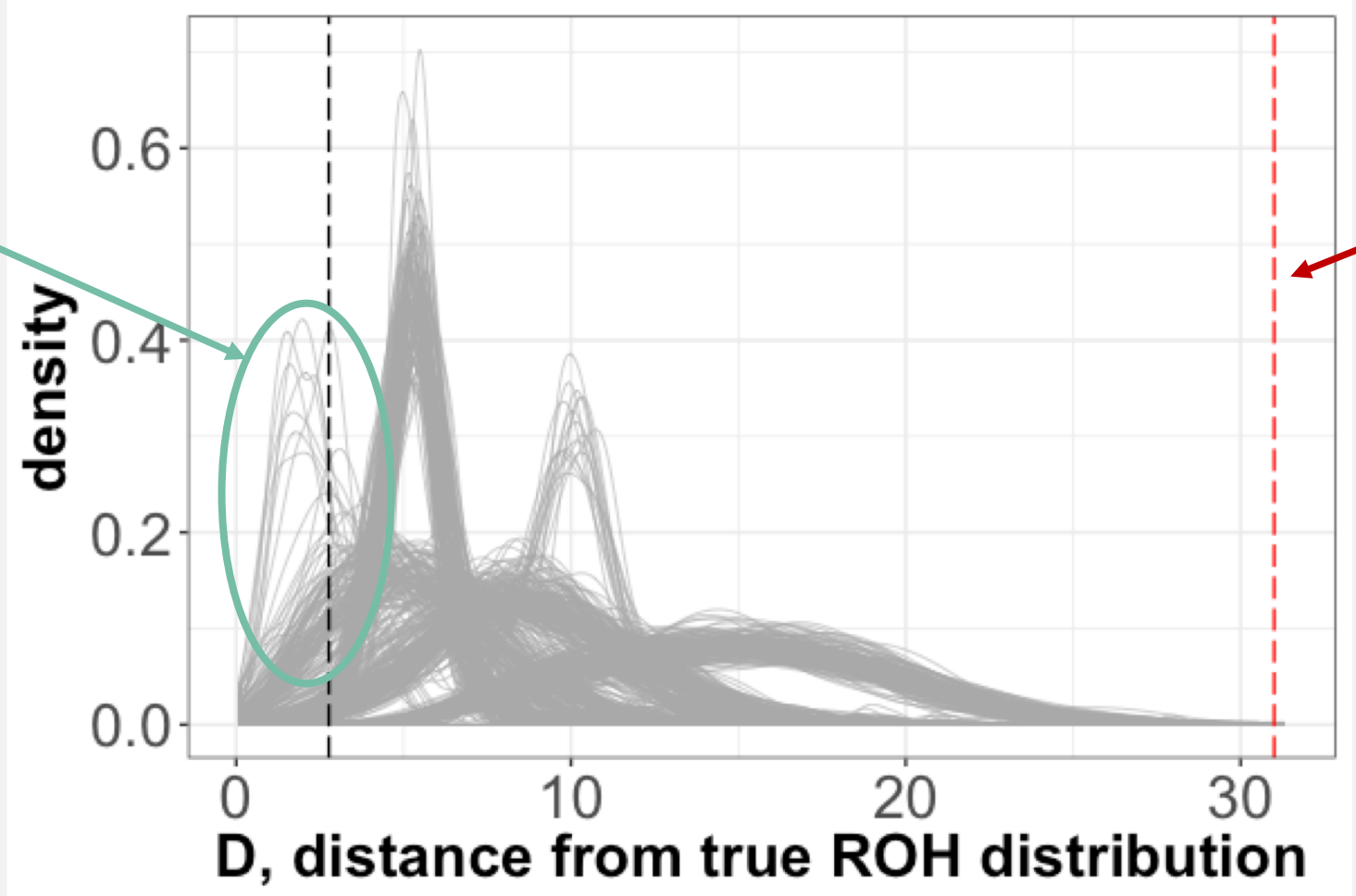
RESULTS: BEST PLINK PARAMETERS



*Not real data

RESULTS: BEST PLINK PARAMETERS

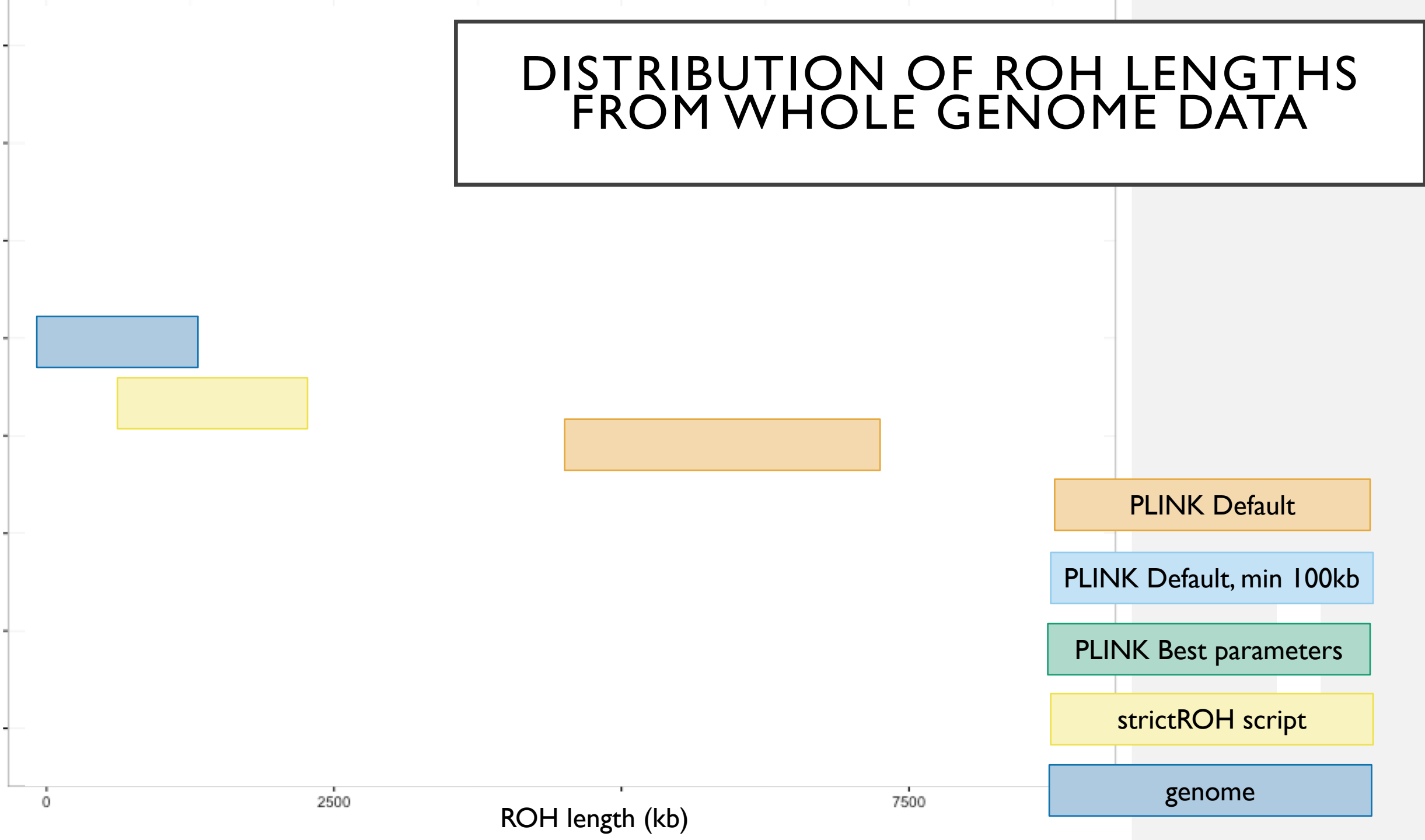
These parameter sets consistently find ROH with similar length distribution to real ROH



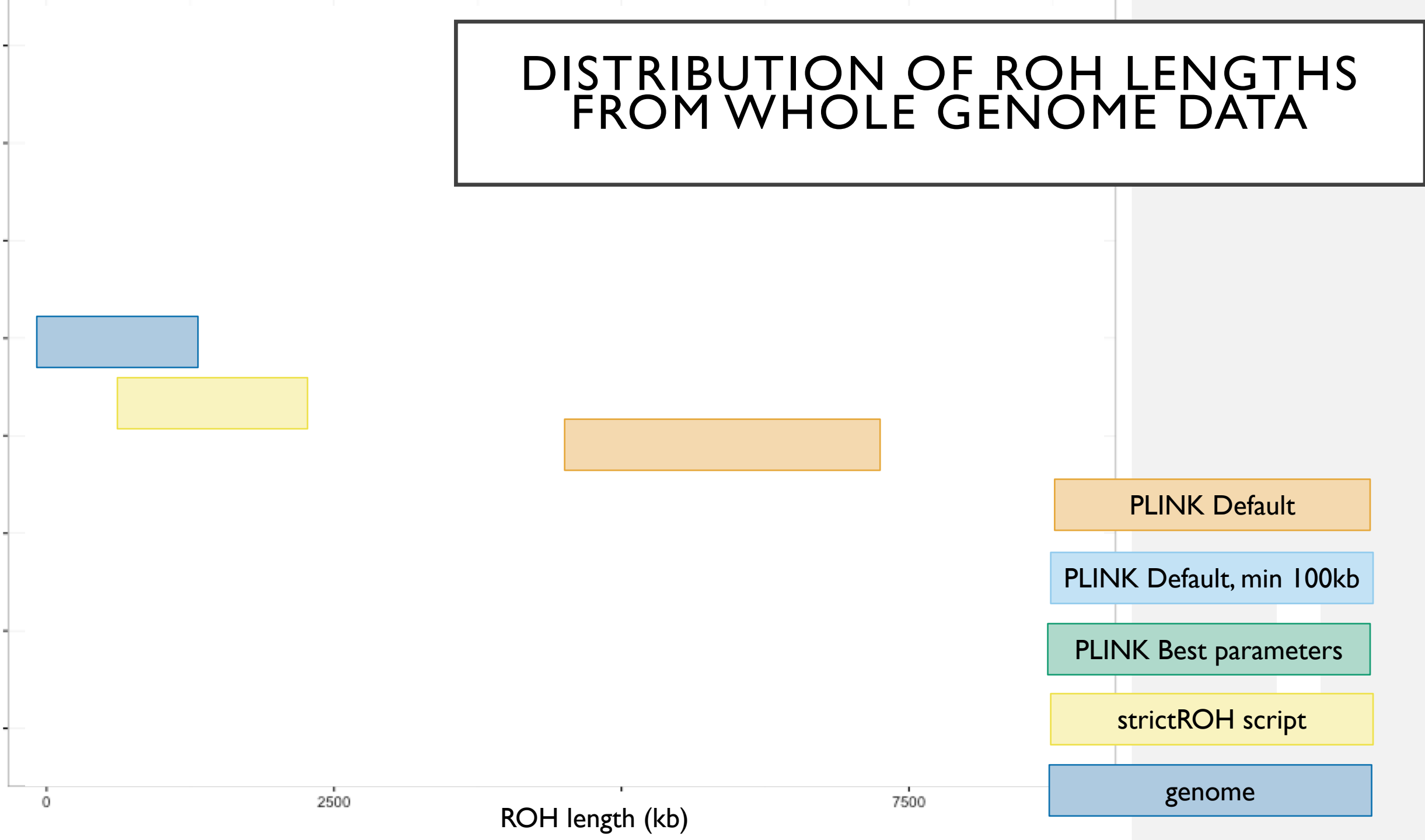
Default parameters

The distribution of *d* for each parameter set in the top half of the grid search, with 100 simulations. Each line is a different parameter set. The dashed line is the 5% cutoff from the distribution of *d* with all parameter sets.

DISTRIBUTION OF ROH LENGTHS FROM WHOLE GENOME DATA

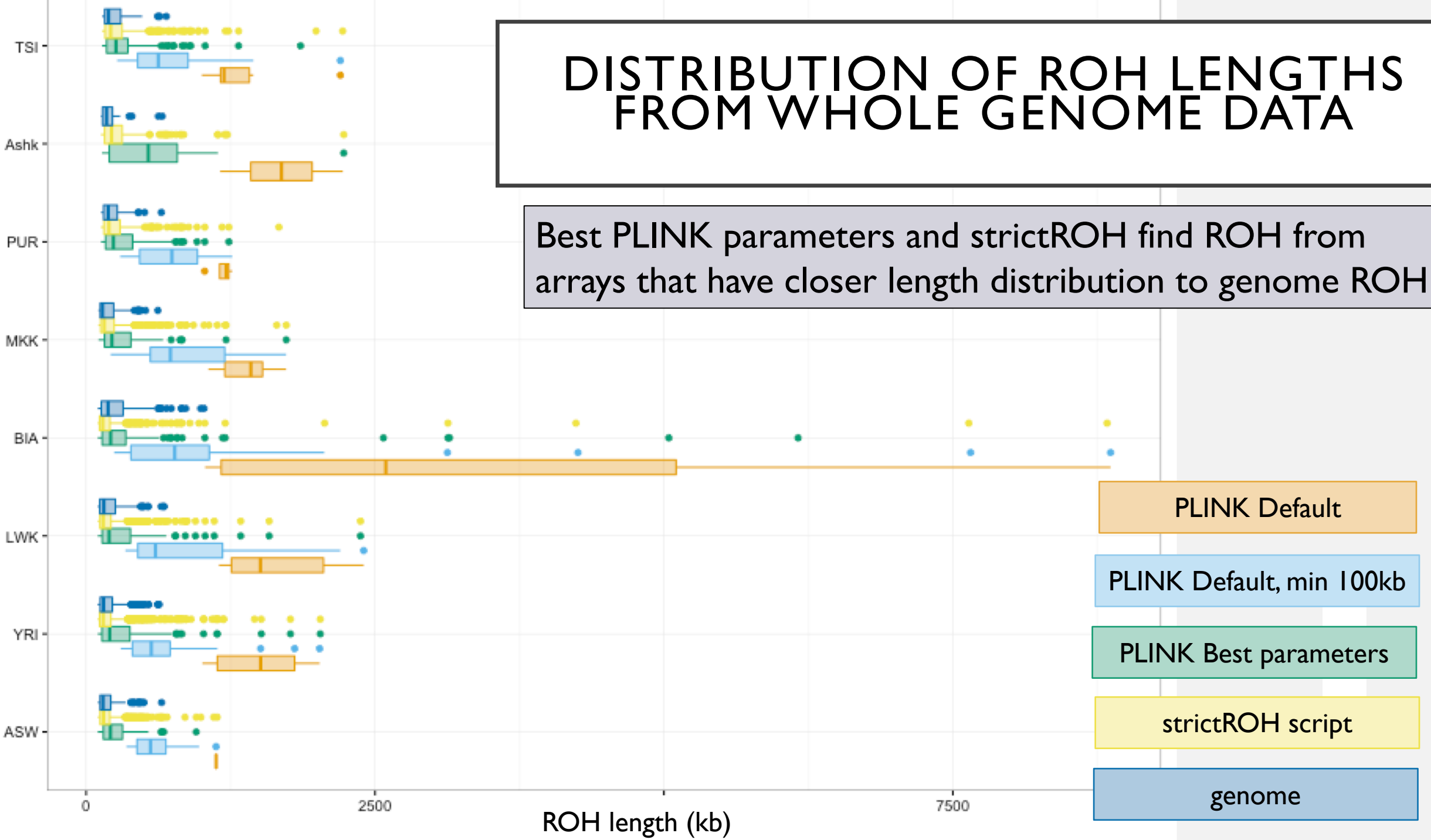


DISTRIBUTION OF ROH LENGTHS FROM WHOLE GENOME DATA



DISTRIBUTION OF ROH LENGTHS FROM WHOLE GENOME DATA

Best PLINK parameters and strictROH find ROH from arrays that have closer length distribution to genome ROH



DEVELOPED CORRECTION FOR ASCERTAINMENT BIAS WITH STRICTROH AND BEST PLINK PARAMETERS

Effect of ascertainment bias on ROH in humans not substantial

- 128 AJ whole genomes published

Can check for ascertainment bias in IBD

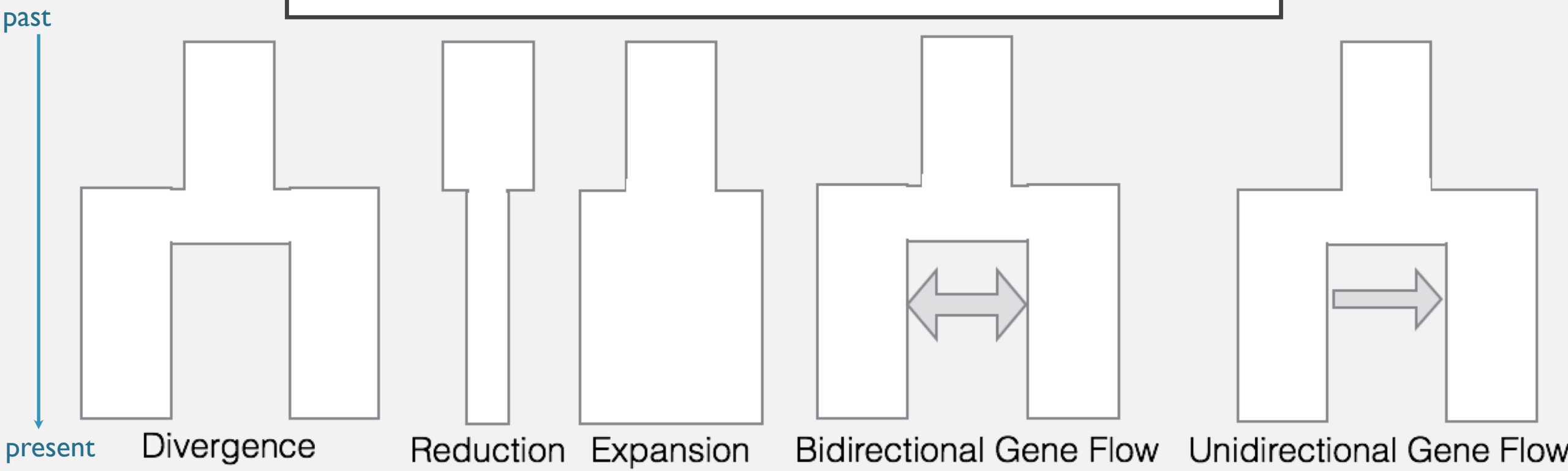
- Can incorporate SNP array ascertainment into model for ABC

Effects of ascertainment bias taken care of

THEMES OF DISSERTATION

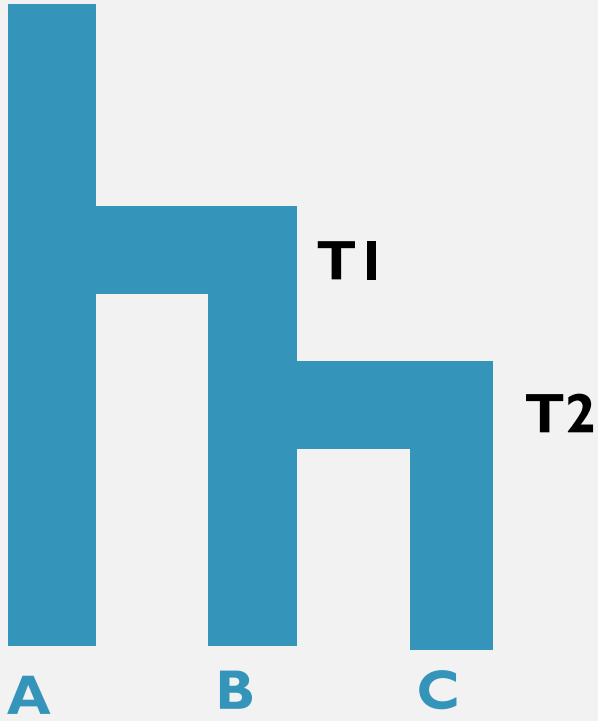
- Detection of runs of homozygosity from SNP arrays
  - Improving identification of runs of homozygosity (Ch. 2)
  - Correcting ascertainment bias in runs of homozygosity (App. C)
- Scaling up Approximate Bayesian Computation for whole chromosomes
  - Create efficient pipeline to simulate demographic models and calculate summary statistics (App. A)
  - Create generalized high throughput workflow (Ch. 4)
- Infer history of the Ashkenazi Jews
  - Substructure in AJ? (Ch. 5)
  - Khazarian origin? (App. B)

EXAMPLES OF DEMOGRAPHIC EVENTS



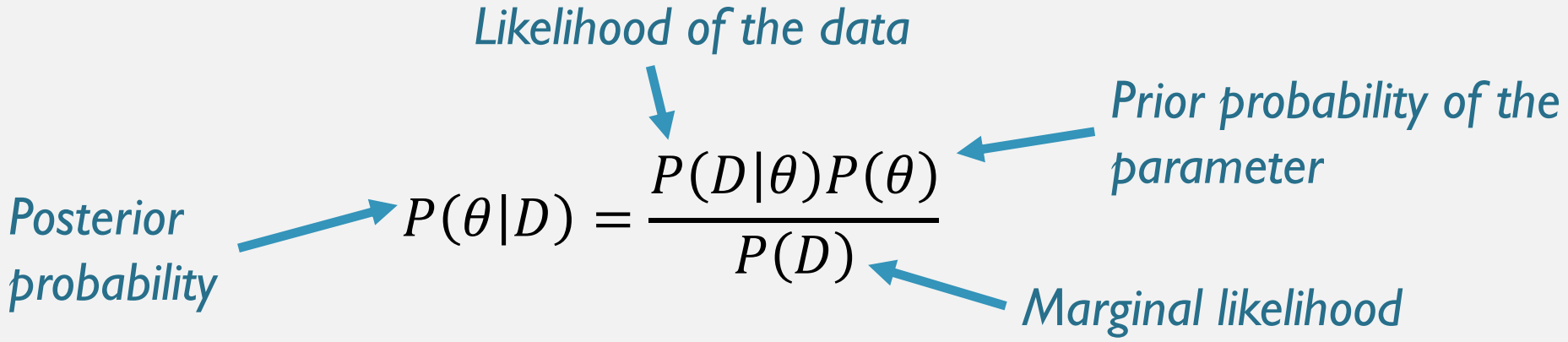
DEMOGRAPHIC PARAMETERS DEFINE POPULATIONS' HISTORIES

- A demographic model generates data, determined by a set of parameters
- Parameter examples:
  - population sizes,
  - divergence times,
  - admixture proportions, etc.



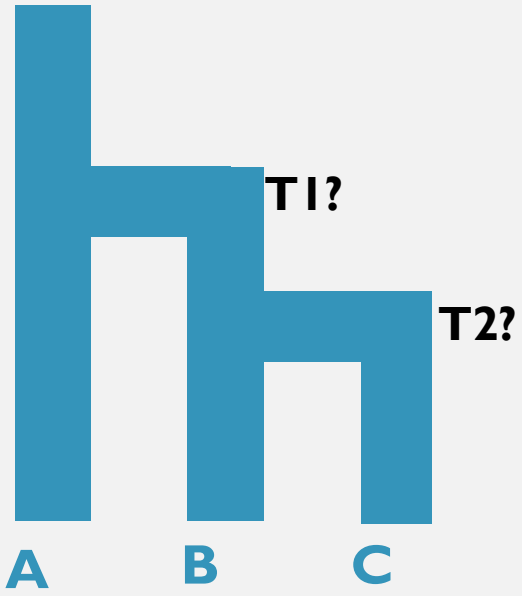
WHAT IS ABC?

- We want the posterior probability of the parameters given the data *(D)*

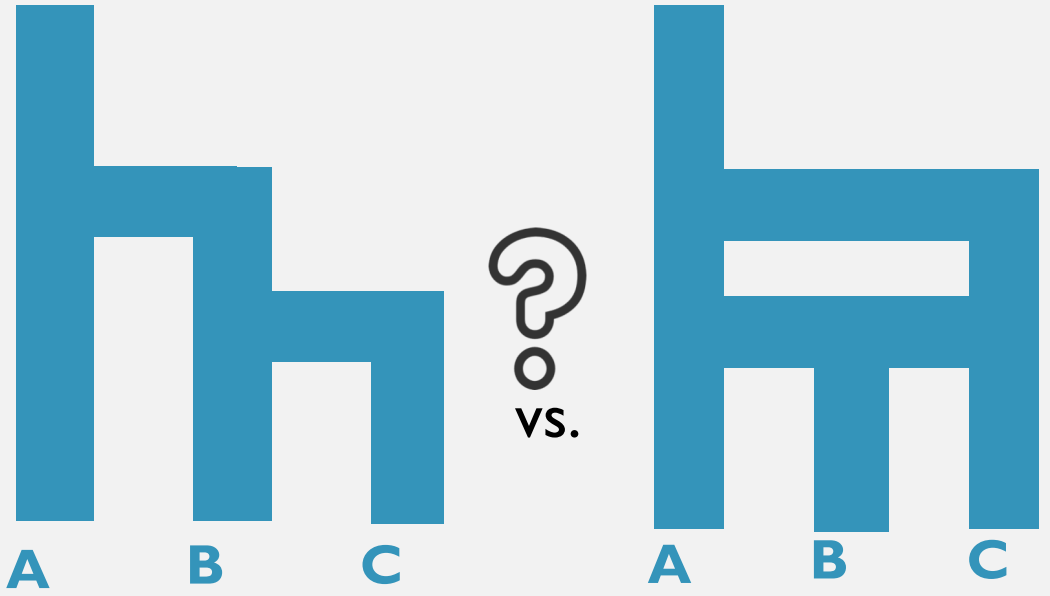


- Approximate the likelihood function by simulations that are compared to the data *(D)*

WHAT IS APPROXIMATE BAYESIAN COMPUTATION (ABC) USED FOR?



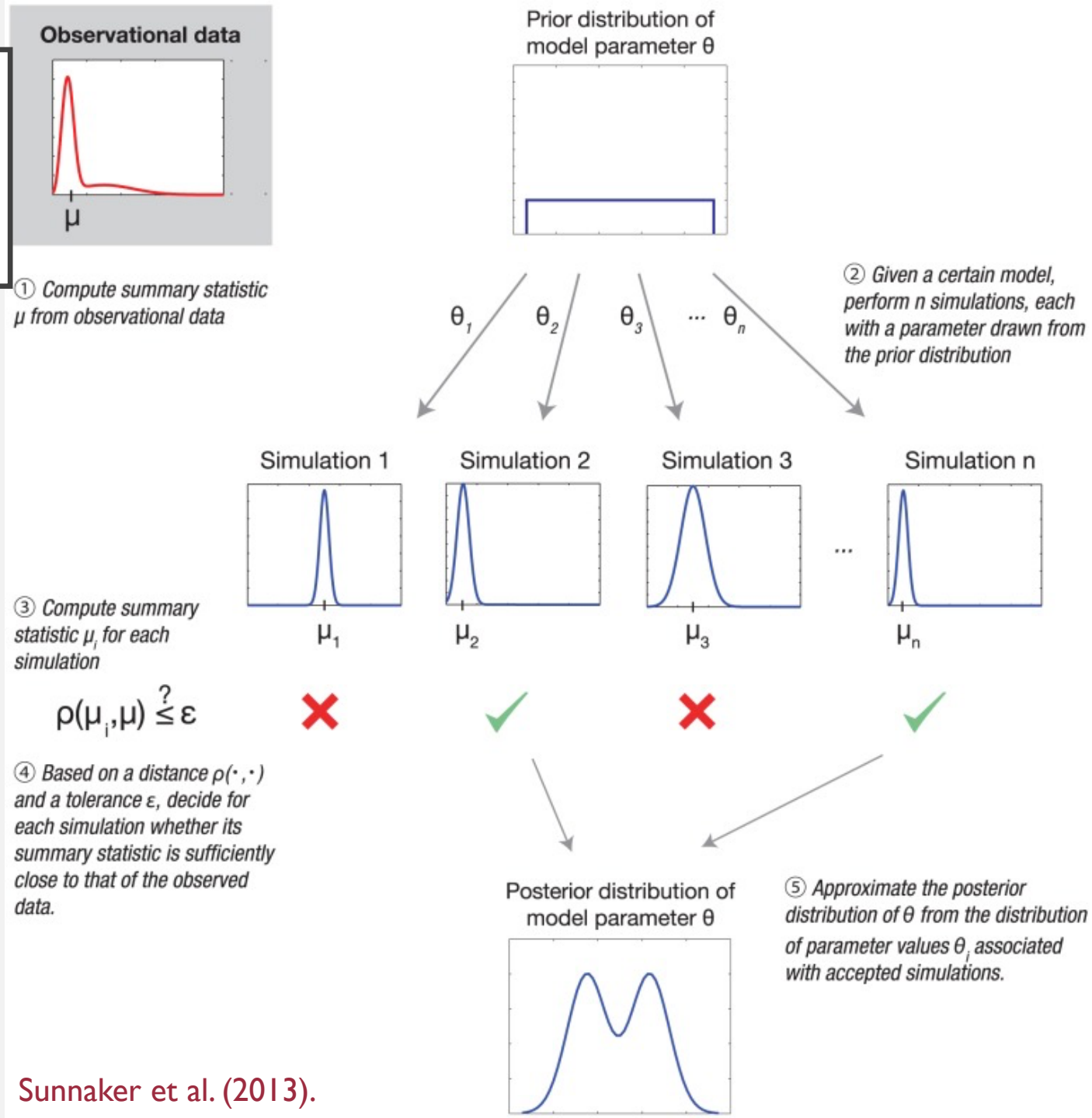
Infer parameter values



Choose among models

OVERVIEW OF ABC STEPS

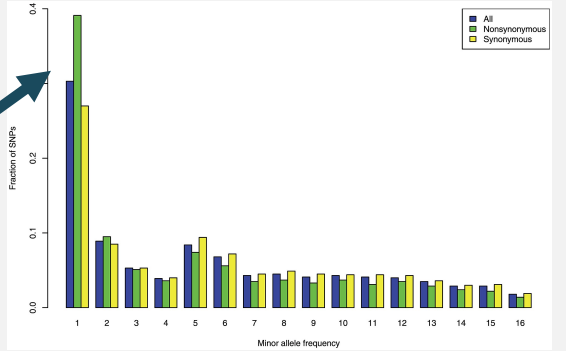
1. Define priors of parameters of model
2. Simulate data many times
3. Summarize genetic data with population genetics statistics
4. Choose model and estimate parameters based on simulations closest to observed data



Sunnaker et al. (2013).

IMPLEMENTATION OF ABC WITH ABCTOOLBOX:
0. COLLECT DATA AND CALCULATE SUMMARY STATISTICS

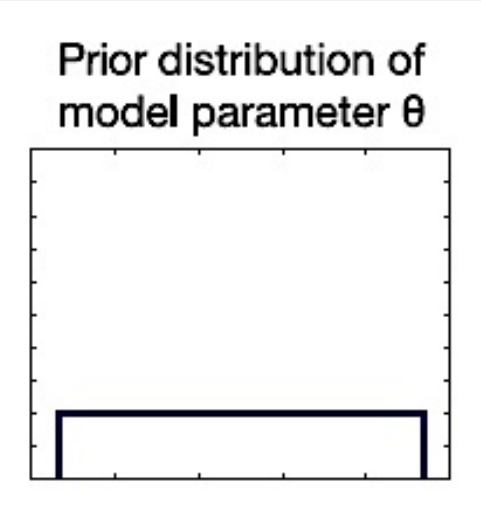
```
1 rs3094315 0 742429 0 1 0 1 0 0 0 0 1 1 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0
1 rs12562034 0 758311 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1
1 rs3934834 0 995669 1 1 0 0 1 1 0 0 1 1 1 1 1 1 0 1 0 1 0 1 1 1 1 1 1 1
1 rs9442372 0 1008567 0 1 0 1 0 1 0 0 0 1 1 1 1 1 0 0 0 0 1 1 1 1 1 1 1 1
1 rs3737728 0 1011278 0 0 0 1 0 1 1 1 0 1 1 1 1 1 0 1 1 1 0 1 1 1 0 1 1 1
1 rs11260588 0 1011521 1 1 1 1 0 1 1 1 0 1 0 1 0 1 1 1 1 1 1 1 1 1 0 1 0 1
1 rs9442398 0 1011558 0 0 0 1 0 1 0 0 0 1 1 1 1 1 0 0 0 1 0 1 1 1 0 1 1 1
1 rs6687776 0 1020428 0 1 0 1 0 1 1 1 0 1 0 0 0 1 0 1 0 0 0 1 1 1 0 1 0 1
1 rs9651273 0 1021403 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 rs4970405 0 1038818 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 rs12726255 0 1039813 0 1 0 1 0 1 1 1 0 1 0 1 0 0 0 0 1 1 0 1 1 1 0 0 0 1
1 rs7540009 0 1050098 1 1 1 1 0 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1
1 rs11807848 0 1051029 1 1 0 0 0 0 1 1 0 0 1 1 0 0 1 1 1 1 0 1 1 1 0 0 1 1
1 rs9442373 0 1052501 0 1 0 0 0 0 1 1 0 0 1 1 0 0 0 0 1 1 0 1 1 1 0 0 1 1
1 rs2298217 0 1054842 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 rs12145826 0 1055892 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 rs4970357 0 1066927 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 rs9442380 0 1077546 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 rs7553429 0 1080420 0 1 1 1 0 1 1 1 1 1 1 1 1 0 0 0 1 1 1 0 1 0 0 1 1 1
1 rs4970362 0 1084601 1 1 0 1 0 1 0 0 1 1 1 1 1 1 0 1 0 0 0 1 0 1 0 0 0 1
```

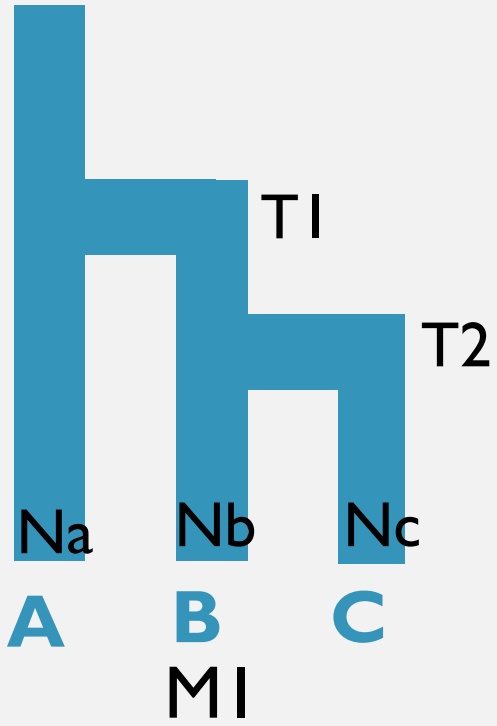


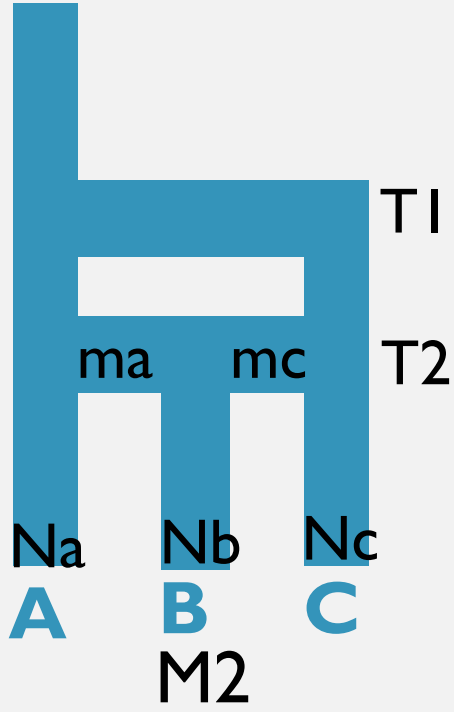
- Number of segregating sites
- nucleotide diversity
- Fst
- Tajima's D
- IBD stats, etc.

IMPLEMENTATION OF ABC WITH ABCTOOLBOX: 1. PICK MODELS AND PRIORS

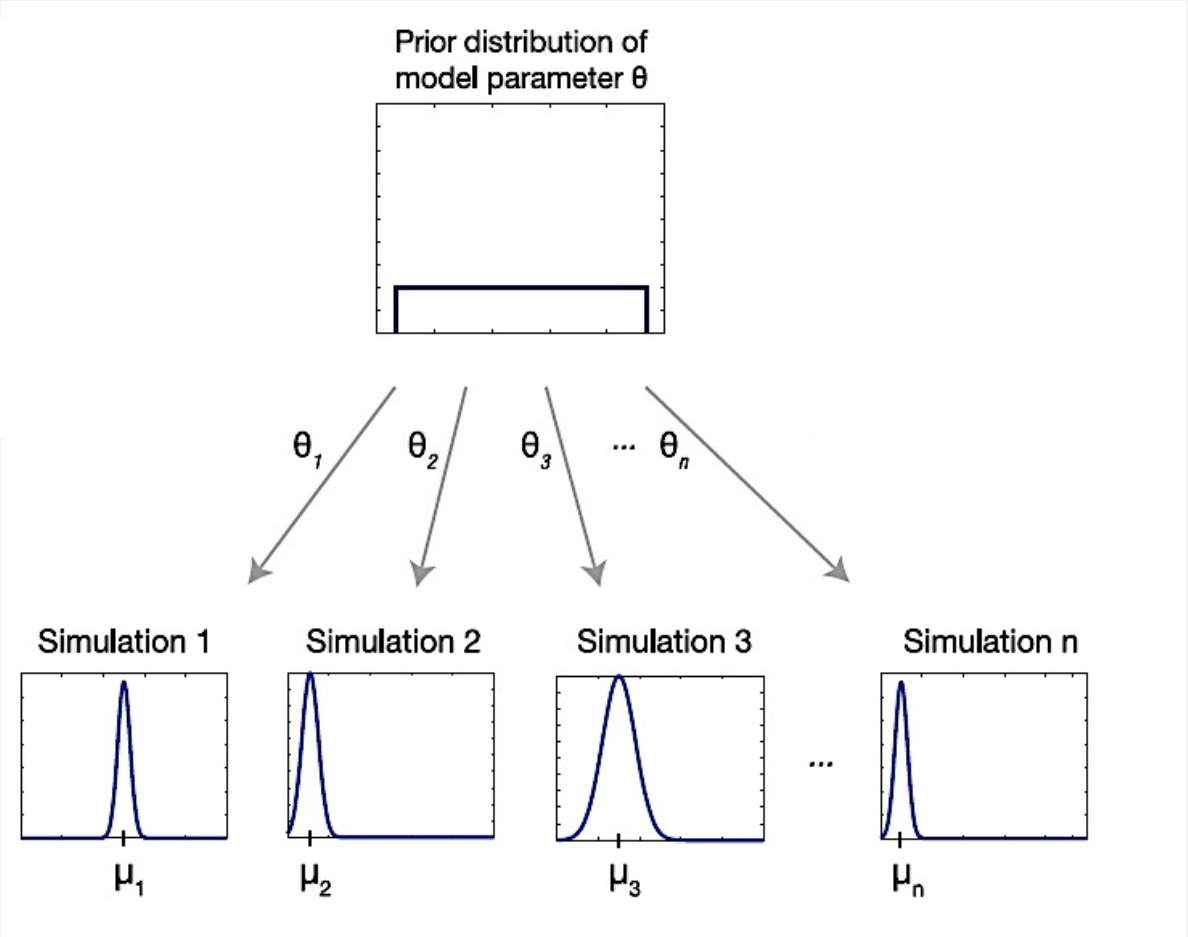
- Parameters (θ):
  - Divergence times (Ti)
  - Population sizes (Nj)
  - Proportion of gene flow (mj)
  - etc…







IMPLEMENTATION OF ABC WITH ABCTOOLBOX:
2. SIMULATE MODELS ACCORDING TO THE PRIORS AND CALCULATE SUMMARY STATISTICS



Sunnaker et al. (2013).

IMPLEMENTATION OF ABC WITH ABCTOOLBOX: 3. ADDRESS CORRELATIONS AMONG STATISTICS

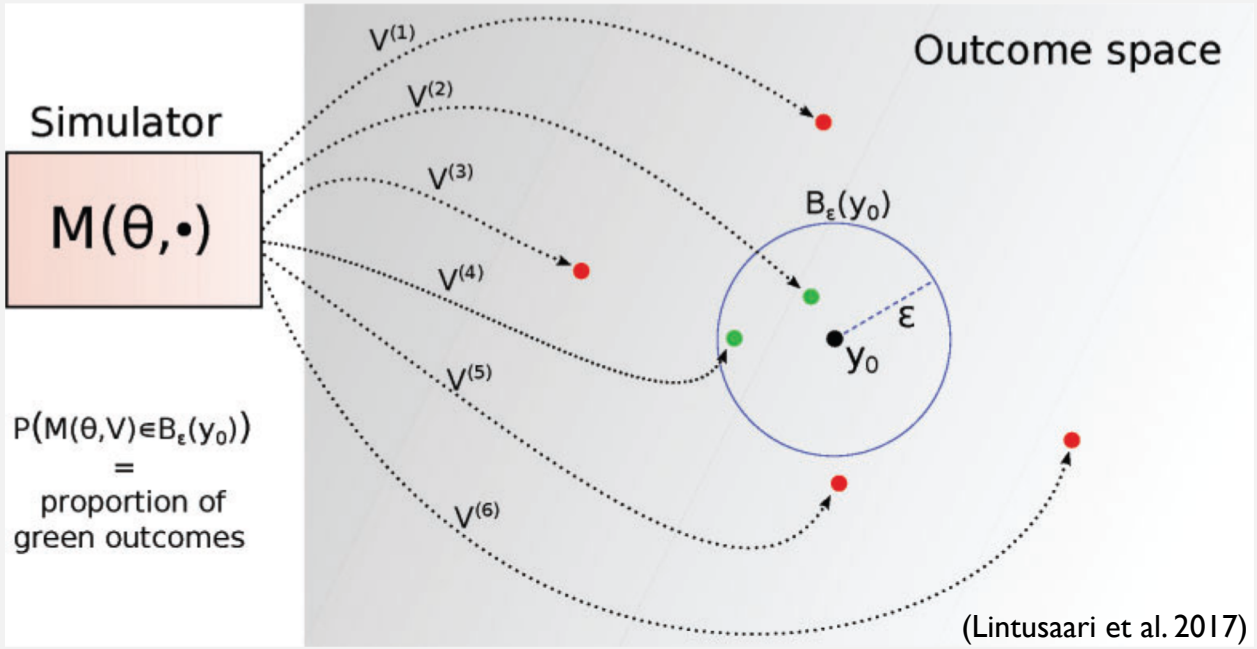
- Prune statistics for high pairwise correlation

or

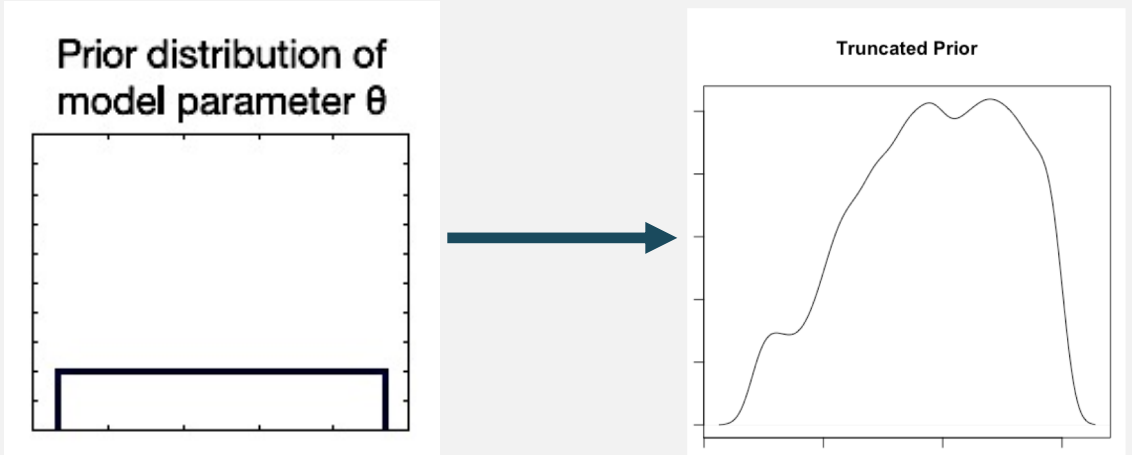
- Transform statistics with Partial Least Squares (PLS)

IMPLEMENTATION OF ABC WITH ABCTOOLBOX: 4. RETAIN N CLOSEST SIMULATIONS TO OBSERVED DATA

- Creates truncated prior by accepting some proportion of parameters and summary stats pairs closest to observed data
- Closest is defined by Euclidean distance between the simulated and observed summary statistics

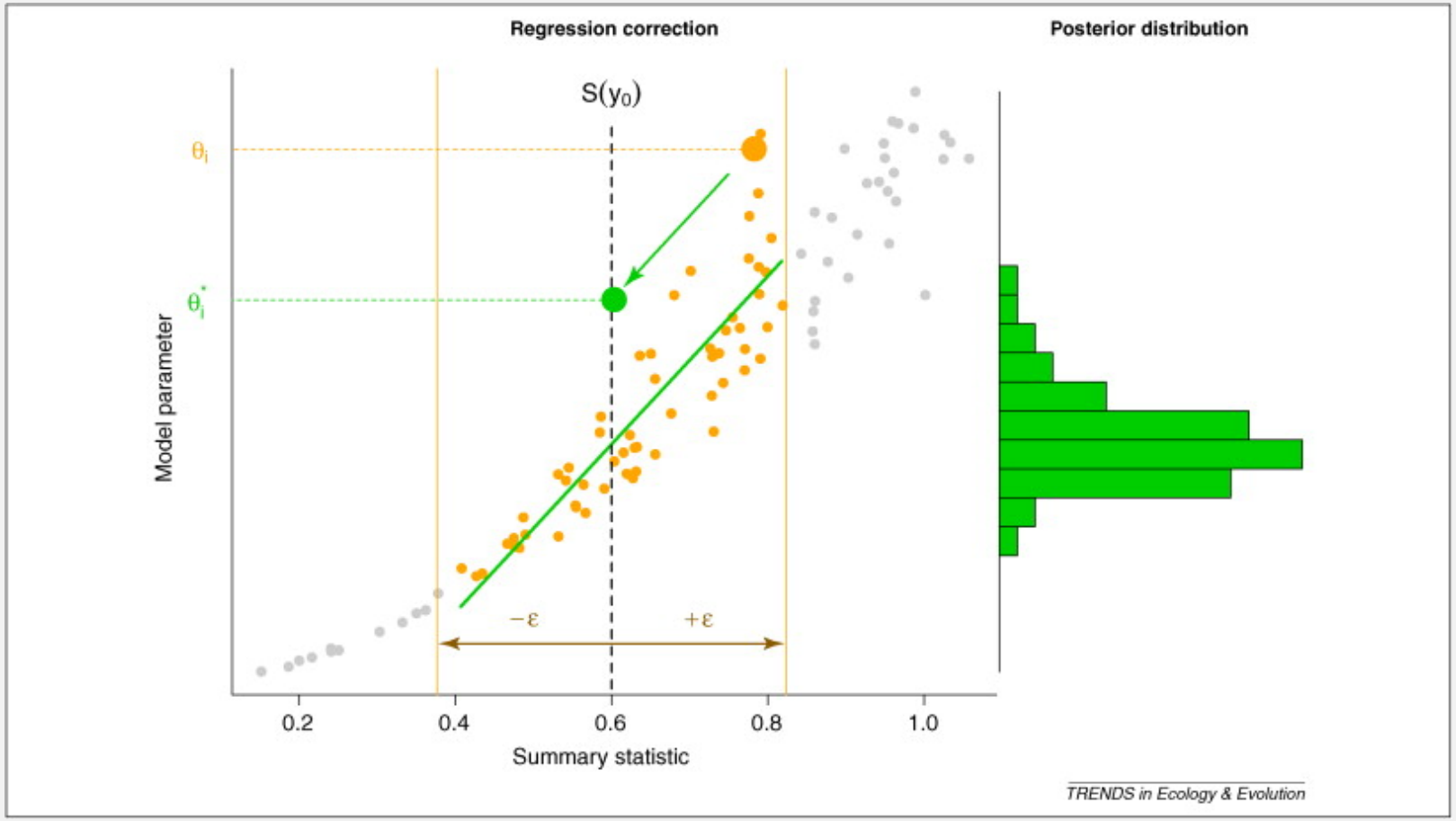


(Lintusaari et al. 2017)



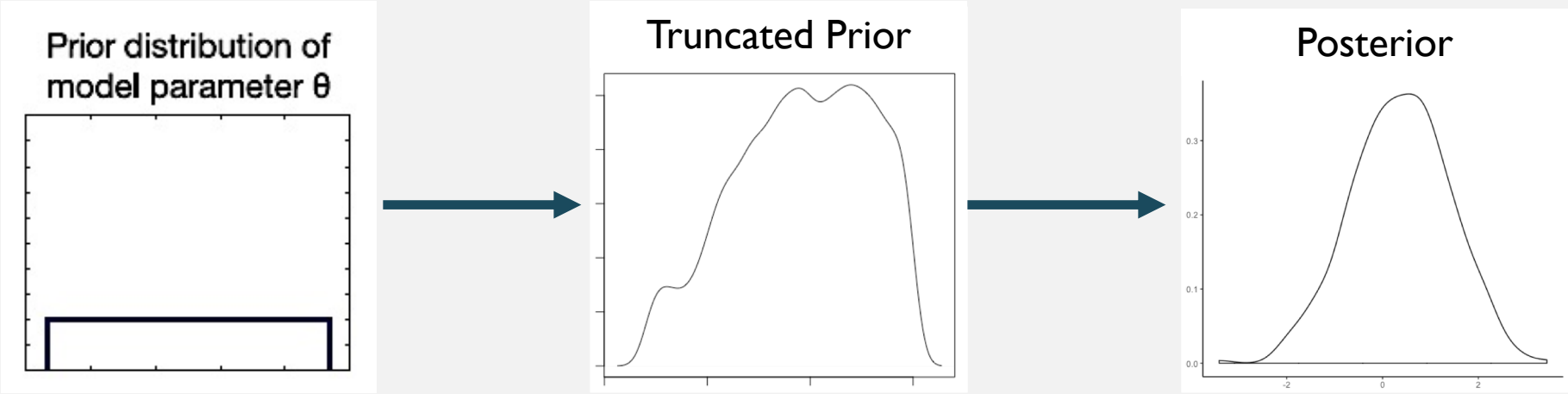
IMPLEMENTATION OF ABC WITH ABCTOOLBOX: 5. LINEAR REGRESSION ON THE SUMMARY STATISTICS AND TRUNCATED PRIOR

- Retained parameter values adjusted according to a linear transformation
- New parameter values form a sample from the posterior

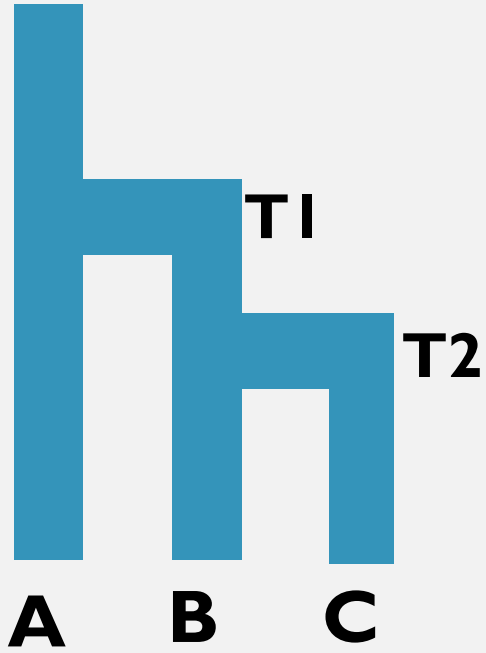


(Csillery et al. 2010)

IMPLEMENTATION OF ABC WITH ABCTOOLBOX: 7. BUILD POSTERIOR DISTRIBUTION OF PARAMETERS

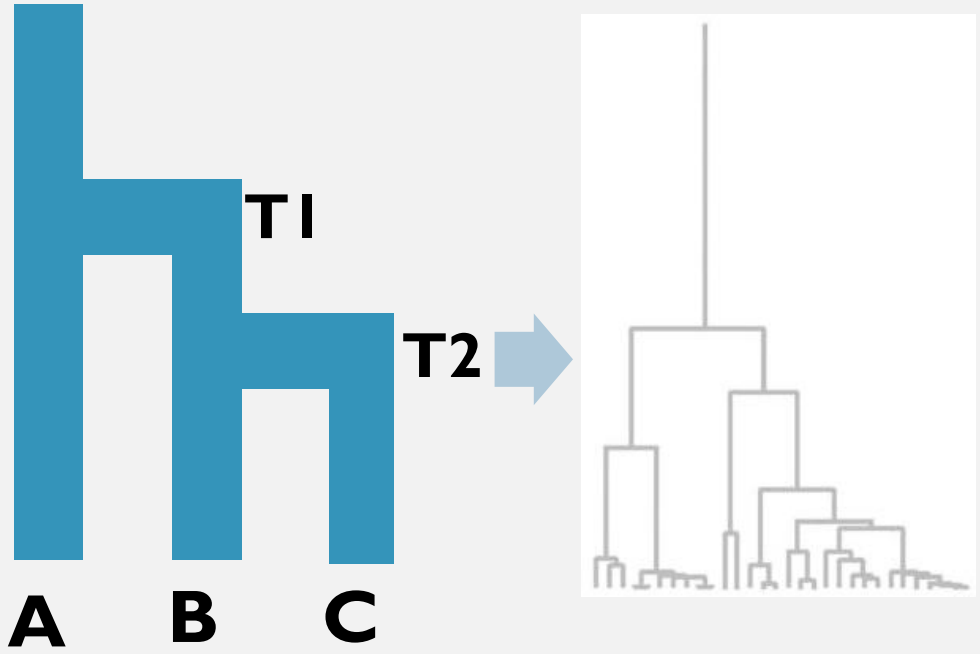


HOW DO WE PERFORM SIMULATIONS AND CALCULATE SUMMARY STATISTICS?



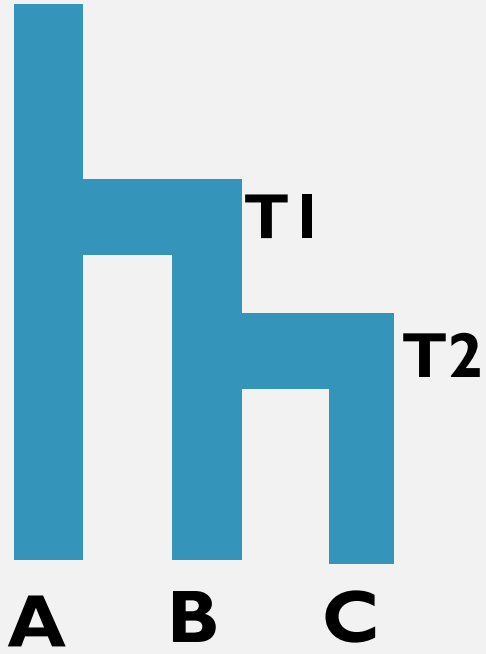
Define model

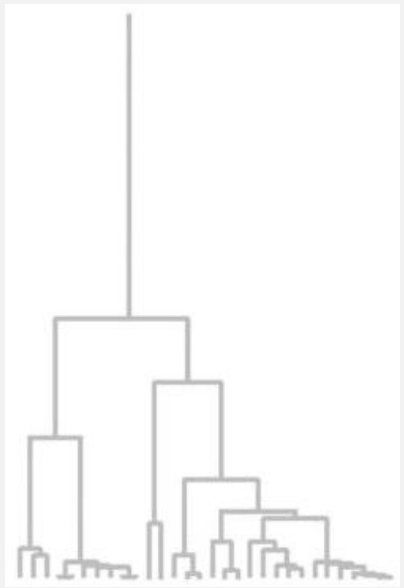
HOW DO WE PERFORM SIMULATIONS AND CALCULATE SUMMARY STATISTICS?



Define model → Coalescent simulation

HOW DO WE PERFORM SIMULATIONS AND CALCULATE SUMMARY STATISTICS?

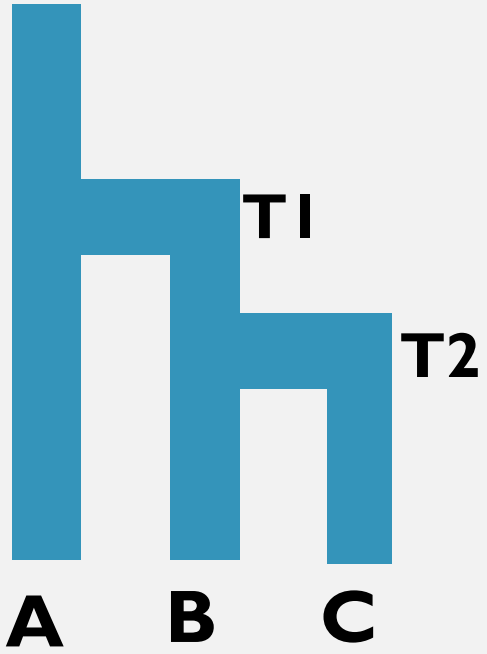


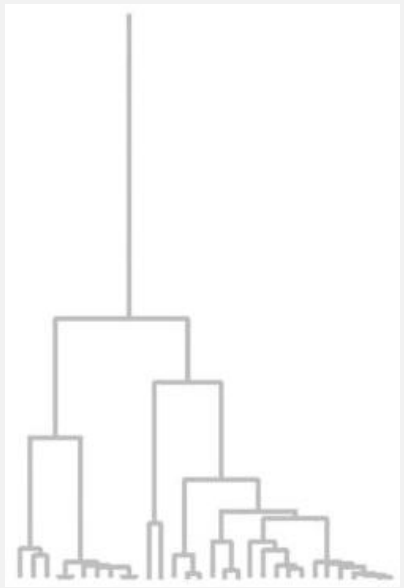


```
110000111101111011
110011110000100001
001111000011110111
101100001111000010
000100110000001001
000000000000111100
001000010000001111
000010000100001100
001111011110111111
000011110111101111
110000111101111011
```

Define model → Coalescent simulation → Store genotype data

HOW DO WE PERFORM SIMULATIONS AND CALCULATE SUMMARY STATISTICS?



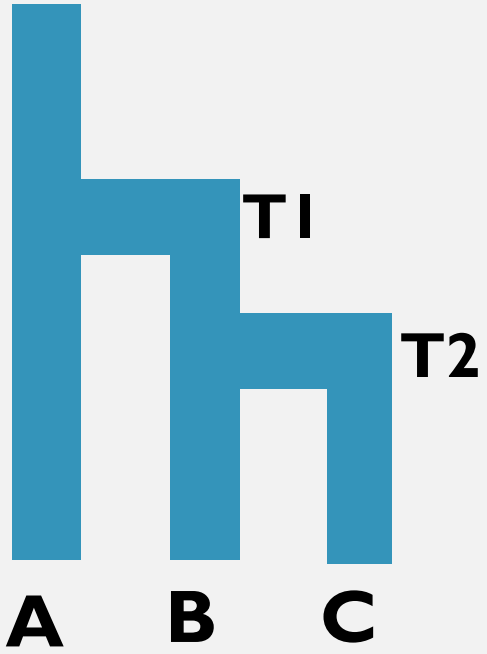


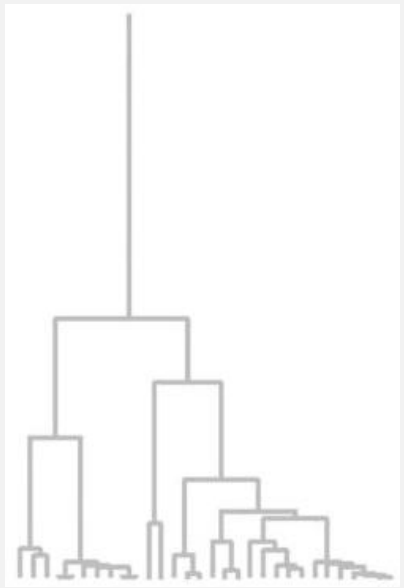
```
110000111101111011
110011110000100001
001111000011110111
101100001111000010
000100110000001001
000000000000111100
001000010000001111
000010000100001100
001111011110111111
000011110111101111
110000111101111011
```

```
def
FST2(seq1,pi1,nseq1,seq2,pi2,nseq2):
   k3=0
      pw=(pi1+pi2)/2
      for i in xrange(len(seq1)):
            si = seq1[i]
            for j in xrange(len(seq2)):
         k3=k3+hd(seq1[i],seq2[j])
      pb=k3/(float(nseq1)*float(nseq2))
      if (pb==0):
            return '0'
      else:
            fst=float(1-(pw/pb))
            return fst
```

Define model → Coalescent simulation → Store genotype data → Calculate statistics

HOW DO WE PERFORM SIMULATIONS AND CALCULATE SUMMARY STATISTICS?



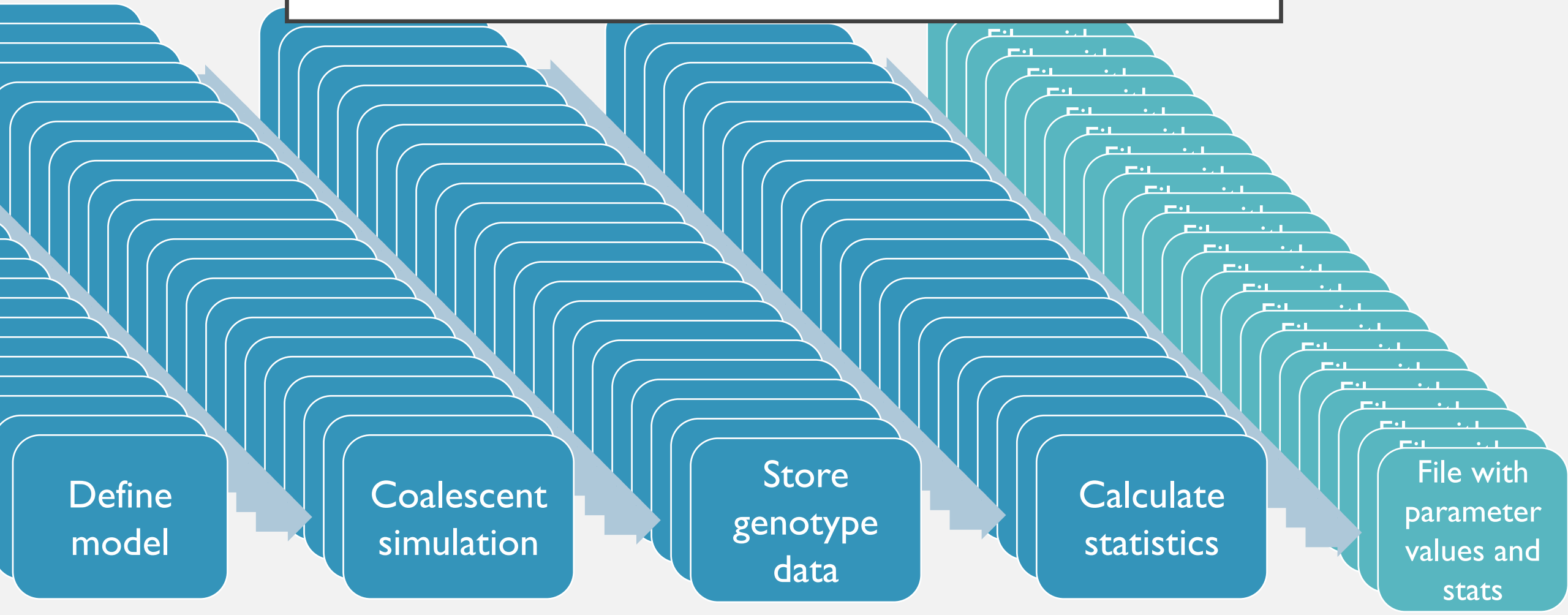


```
110000111101111011
110011110000100001
001111000011110111
101100001111000010
000100110000001001
000000000001111100
001000010000001111
000010000100001100
001111011110111111
000011110111101111
110000111101111011
```

```
def
FST2(seq1,pi1,nseq1,seq2,pi2,nseq2):
    k3=0
        pw=(pi1+pi2)/2
        for i in xrange(len(seq1)):
            si = seq1[i]
            for j in xrange(len(seq2)):
            k3=k3+hd(seq1[i],seq2[j])
        pb=k3/(float(nseq1)*float(nseq2))
        if (pb==0):
            return '0'
        else:
            fst=float(1-(pw/pb))
            return fst
```

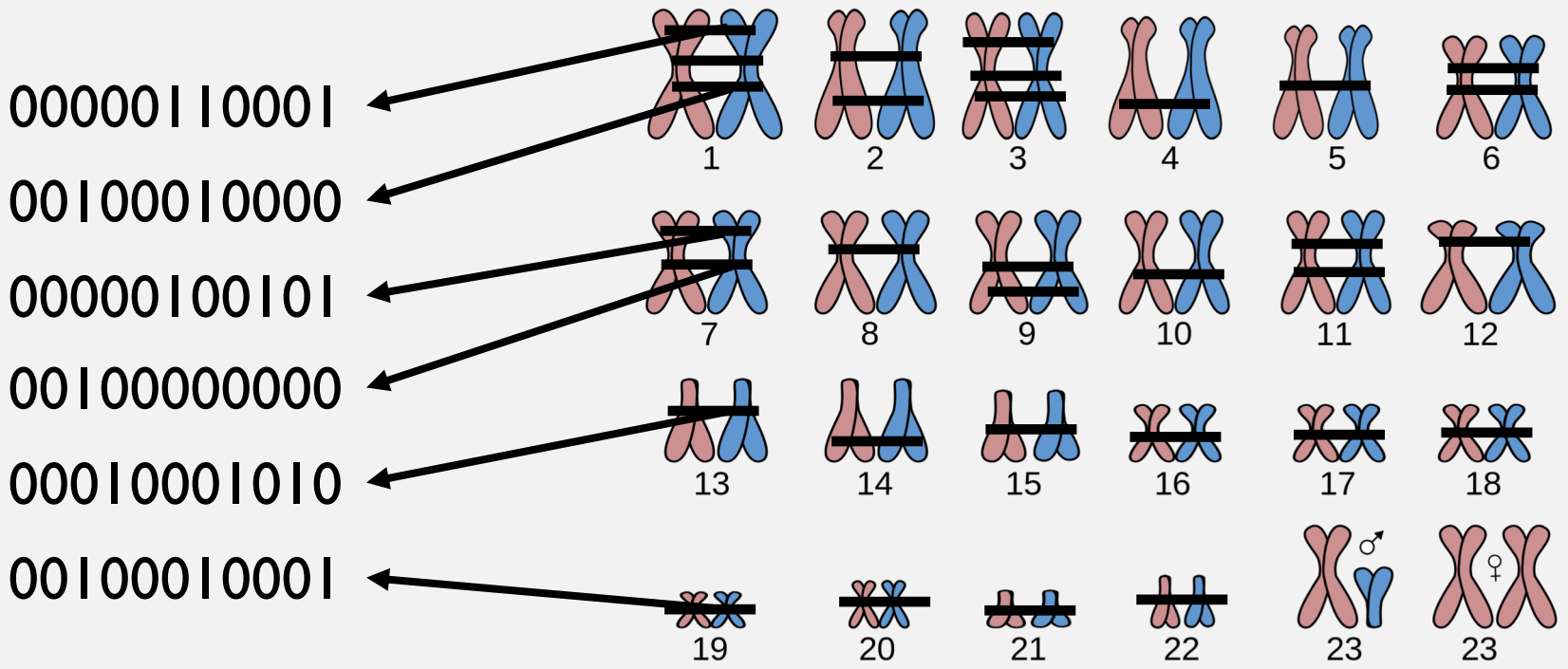
Define model → Coalescent simulation → Store genotype data → Calculate statistics → File with parameter values and stats

PLEASANTLY PARALLEL!



INHERITED SCRIPT INTENDED FOR SMALL SEQUENCE

- Intended for millions of relatively small simulations
  - 1,389 10kb regions
  - 65 individuals
- Took a few minutes to run one simulation
- Ran parallel on U of A HPC
  - 1 million runs would take approximately 1 month.

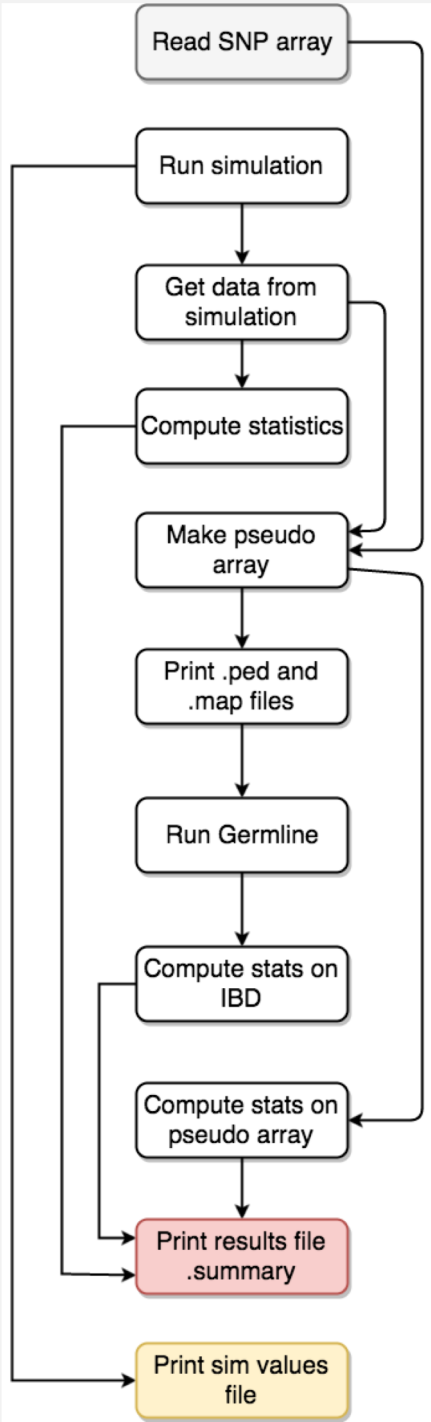


SIMULATE WHOLE CHROMOSOME

~250 million sites on human chromosome 1

SIMULATE WHOLE CHROMOSOME

- Modified Python script to
  - Simulate whole chromosome
  - Find IBD segments and calculate IBD stats



PROBLEM!

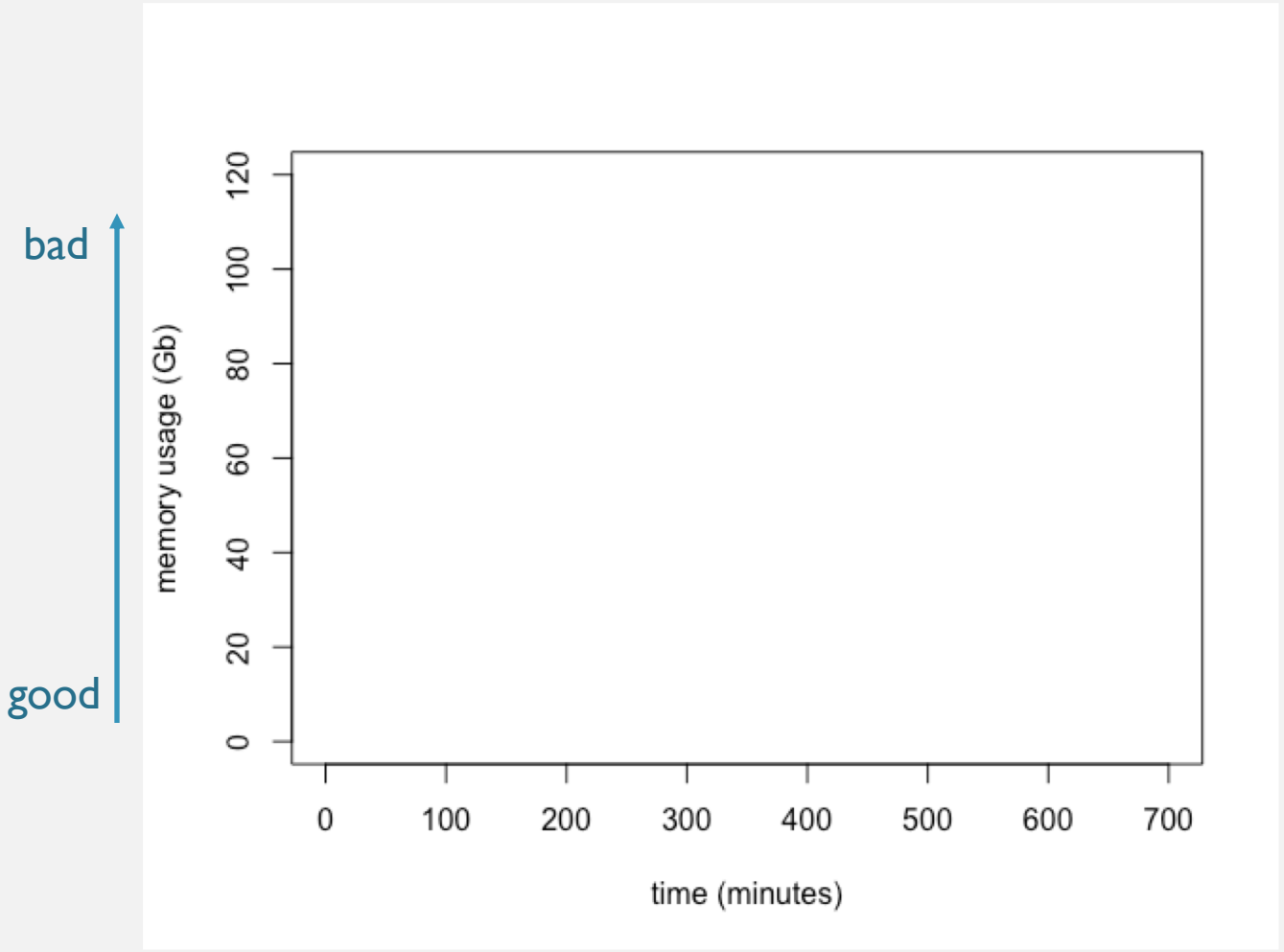
Parameters	Average Walltime	Average Memory
Minimum	00:21:00	2.7 Gb
Random	00:55:11	20 Gb
Maximum	08:02:11	117 Gb

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

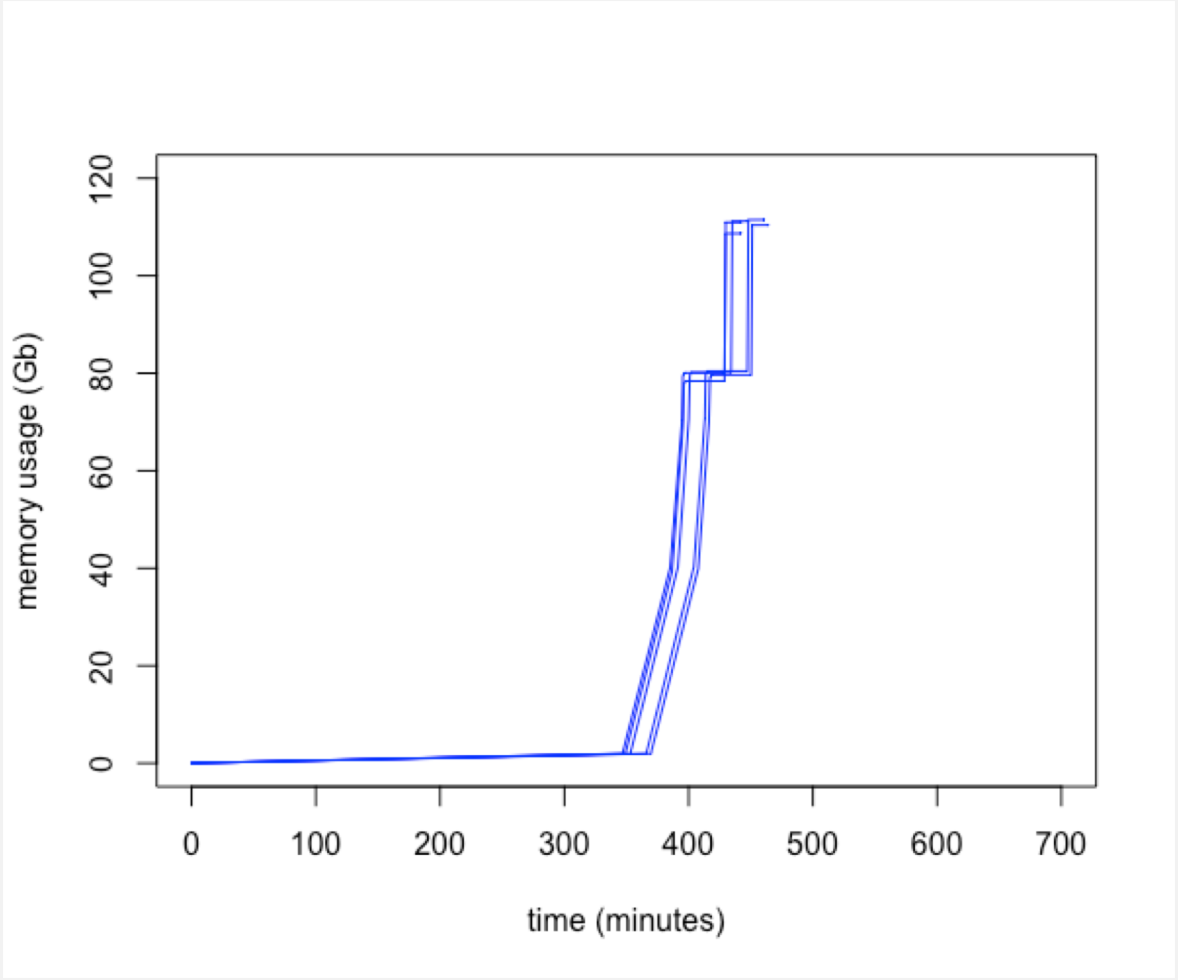
PROFILE OF PYTHON SCRIPT



PROFILE OF PYTHON SCRIPT

bad

good



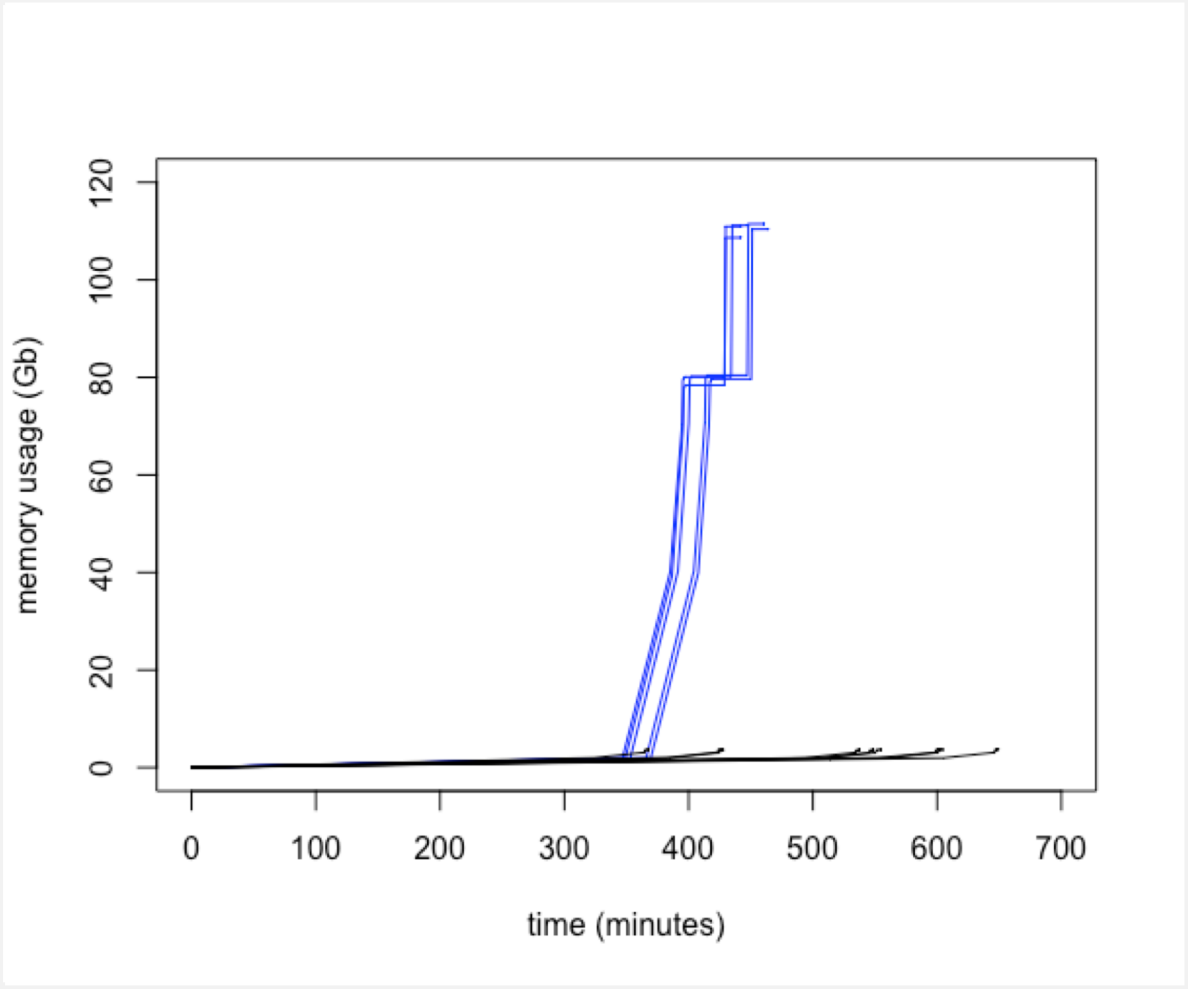
good

bad

PROFILE OF PYTHON SCRIPT

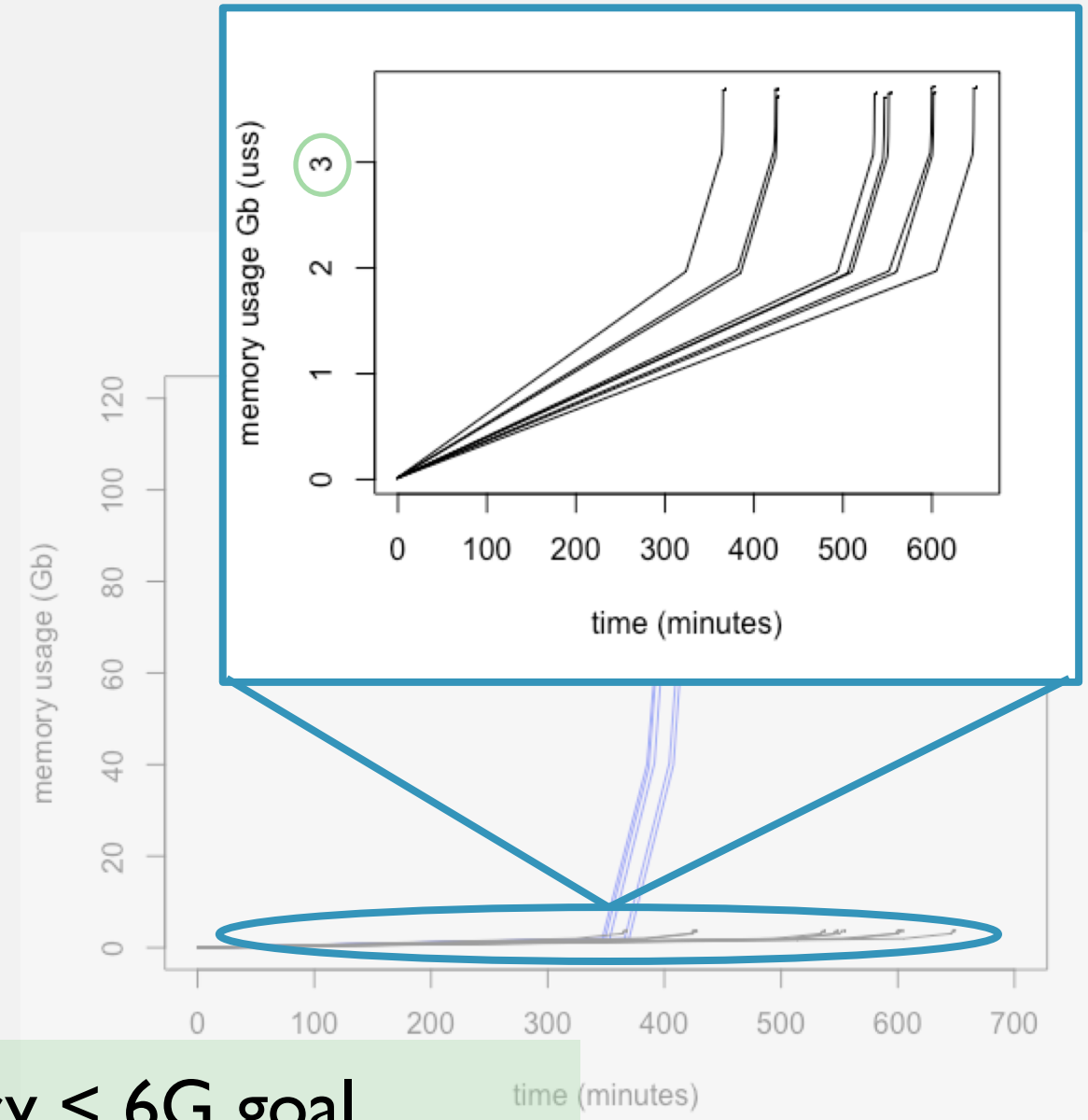
bad

good



good

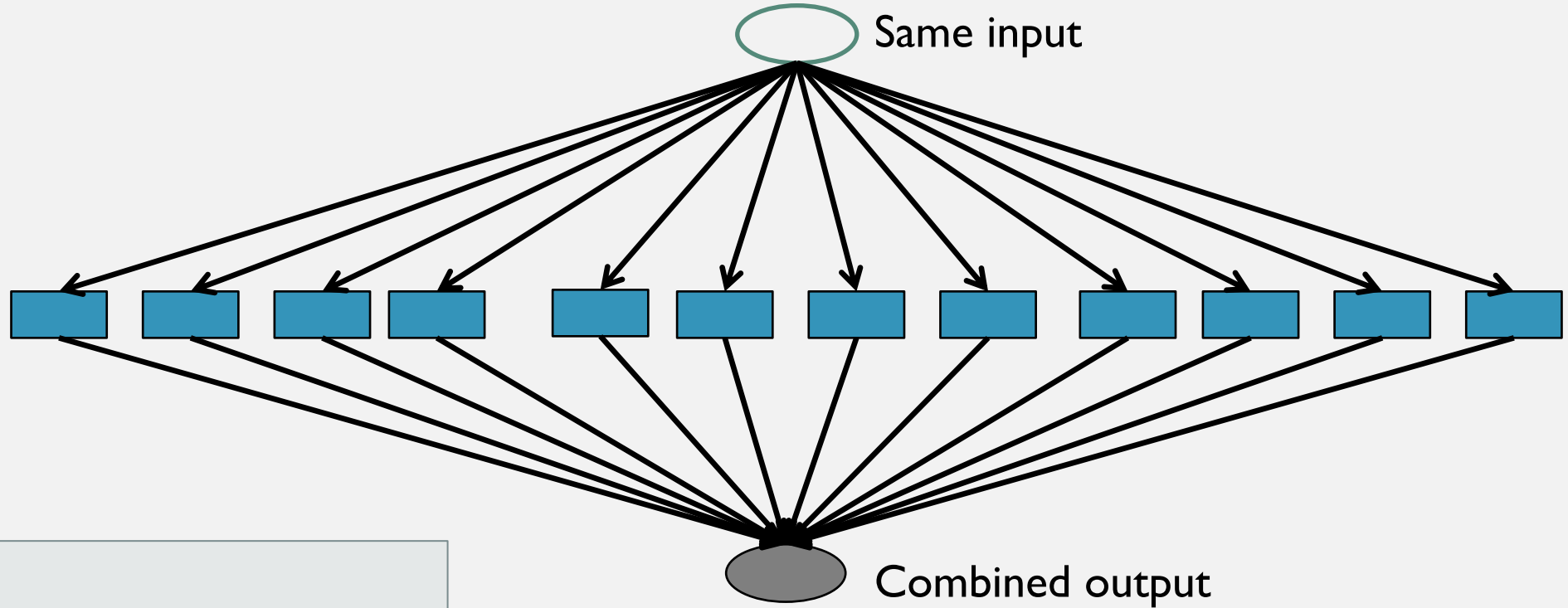
bad



Max memory < 6G goal

Can now run efficiently in parallel

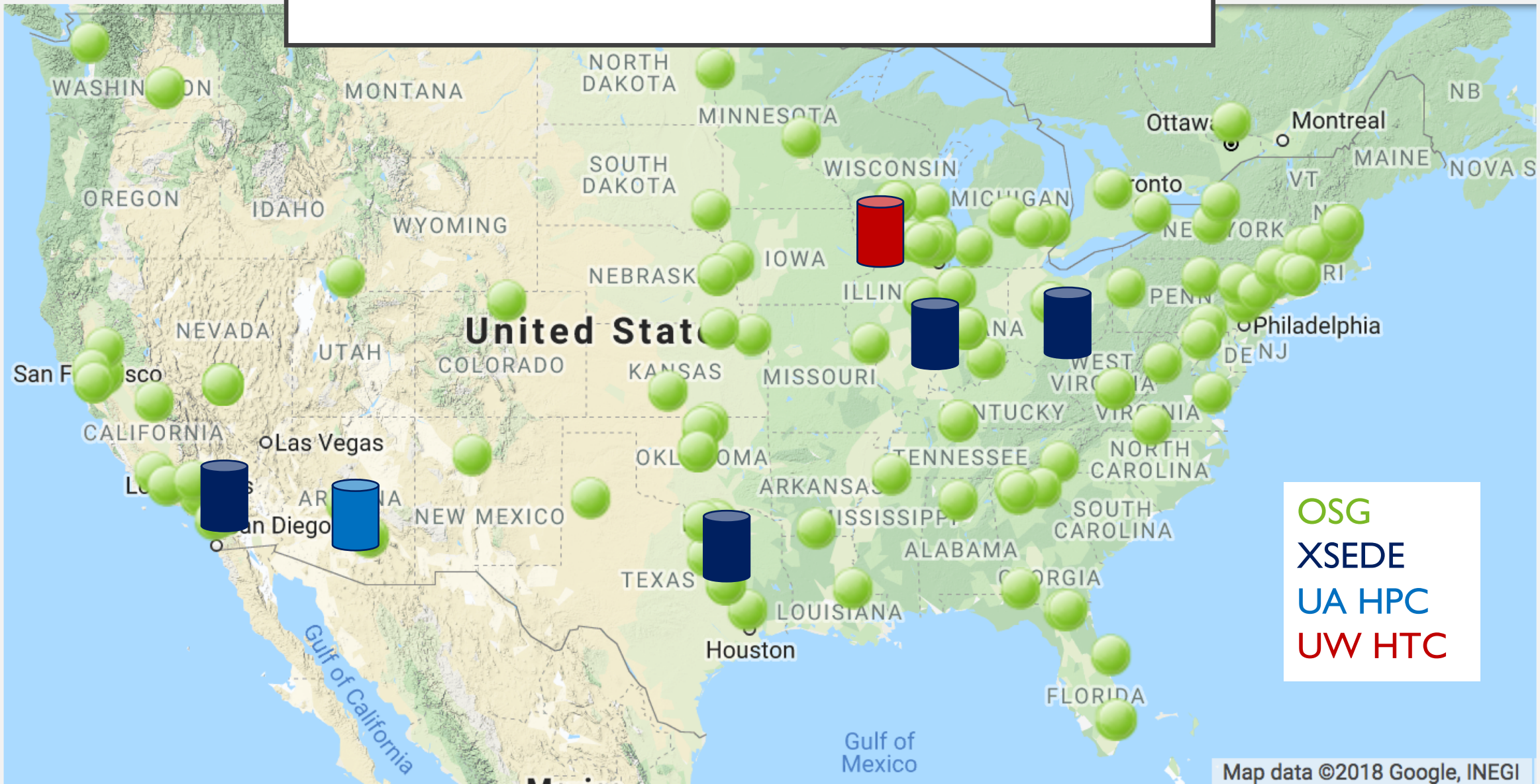
PLEASANTLY PARALLEL & RESOURCE LIGHT!



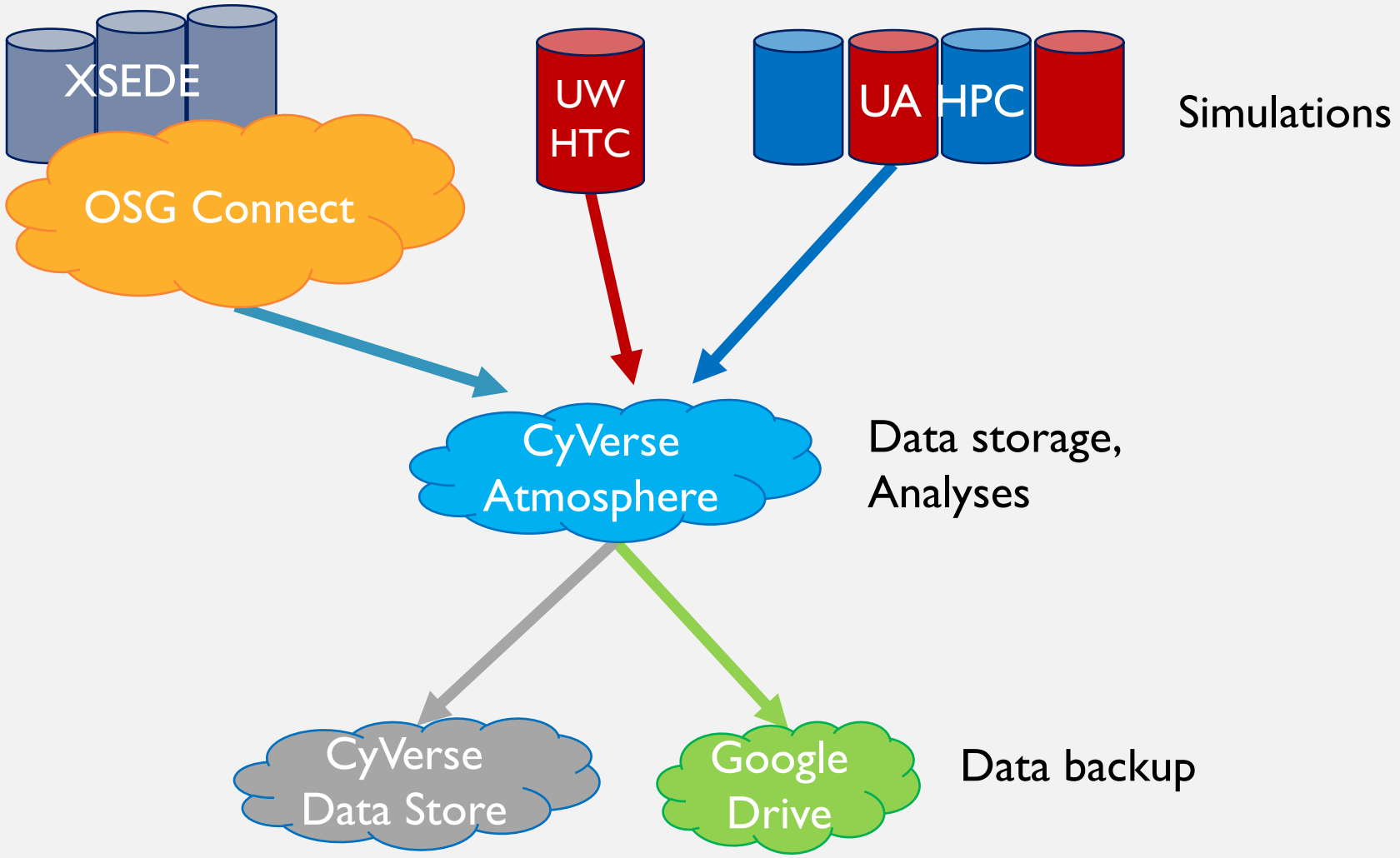
- Each job
  - runs ~40 min, and max 50 hrs
  - Uses ~1G, and max 5G memory
  - Uses ~2M in storage

Pegasus

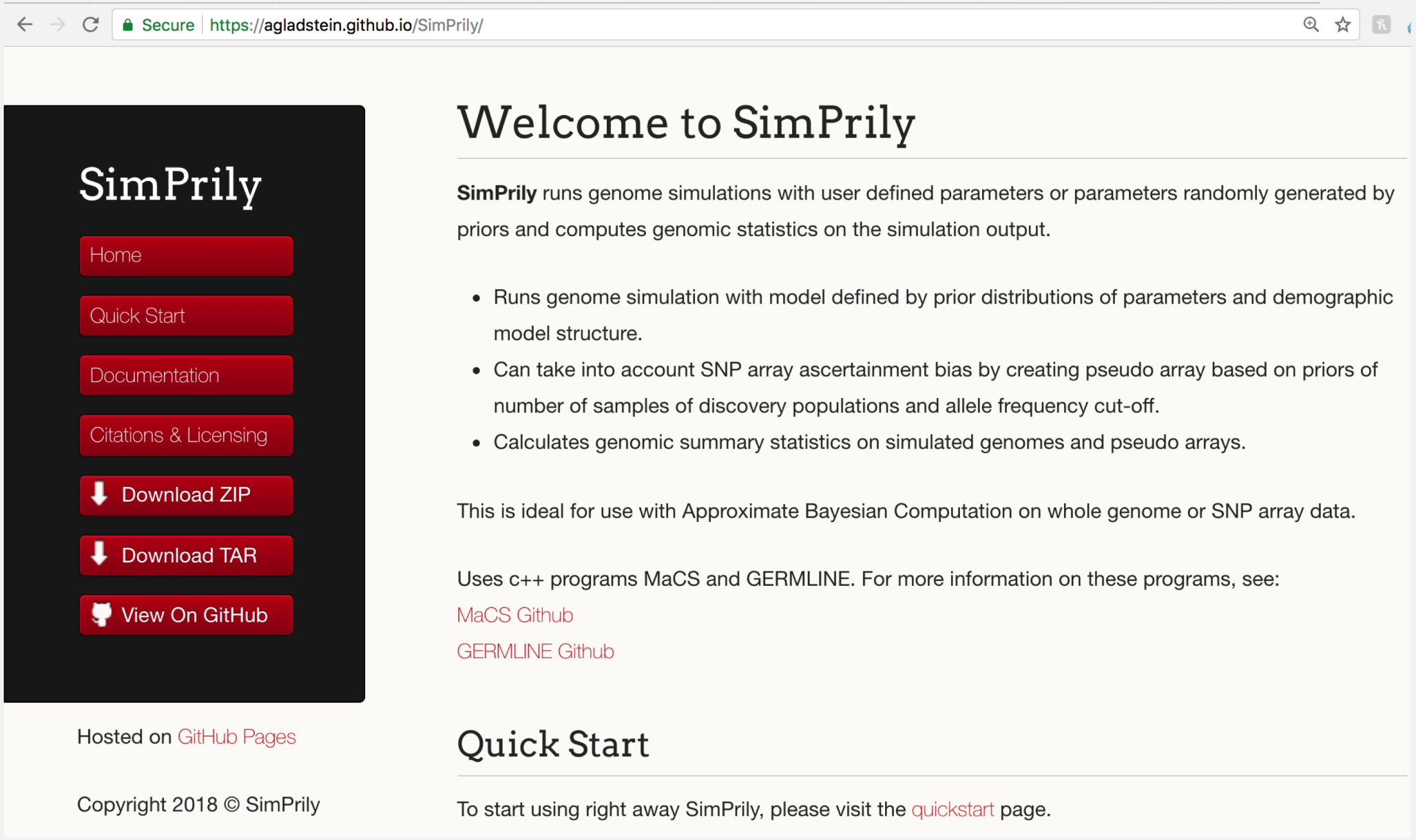
HIGH THROUGHPUT COMPUTING



SIMULATIONS ON HTC CLUSTERS, ANALYSES ON VM



GENERALIZATION OF CODE AND WORKFLOW



SIMPRILY HAS UNIQUE FEATURES

Program	Large loci	Priors	Statistics	SNP ascertainment	HTC
SimPrily (2018)	✓	✓	✓	✓	✓
Fastsimcoal2 (2013)	✓	✓			
Msprime (2016)	✓		✓		
BaySICS (2014)		✓	✓		✓
Coala (2016)		✓	✓		
SKELESIM (2017)			✓		

Comparison of SimPrily features with other simulators and wrappers.

POTENTIAL APPLICATIONS OF SIMPRILY

- Simulate genome sequence or SNP array data to
  - Test software
  - Infer demographic history with Approximate Bayesian Computation
  - Use as null model when inferring regions under selection
  - Create training and test dataset for machine learning

THEMES OF DISSERTATION

- Detection of runs of homozygosity from SNP arrays
  - Improving identification of runs of homozygosity (Ch. 2)
  - Correcting ascertainment bias in runs of homozygosity (App. C)
- Scaling up Approximate Bayesian Computation for whole chromosomes
  - Create efficient pipeline to simulate demographic models and calculate summary statistics (App. A)
  - Create generalized high throughput workflow (Ch. 4)
- Infer history of the Ashkenazi Jews
  - Substructure in AJ? (Ch. 5)
  - Khazarian origin? (App. B)

DATASET

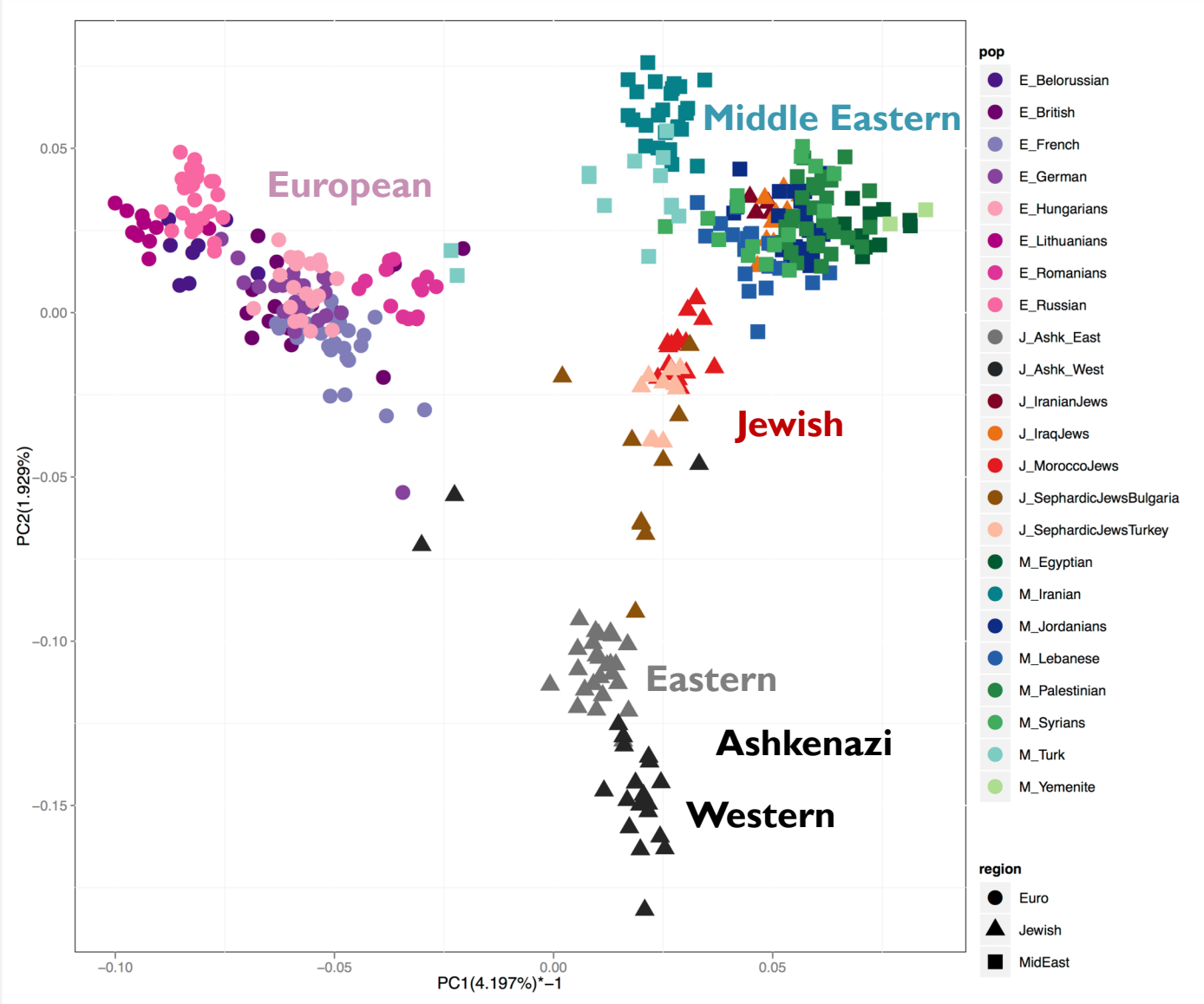
SNP array data	Sample Size	Source
Eastern Ashkenazi	239	Family Tree DNA, Behar et al. 2010
Western Ashkenazi	19	Family Tree DNA, Behar et al. 2010
Jewish (9 pops)	79	Behar et al. 2010
Middle Eastern (11 pops)	211	Behar et al. 2010, Hammer, HGDP
European (8 pops)	139	Behar et al. 2010, Hammer, HGDP

Whole genome data	Sample Size	Source
Ashkenazi	230	Carmi et al. 2014, Hammer Lab
European, African, Asian, American		CGI, 1000 Genomes

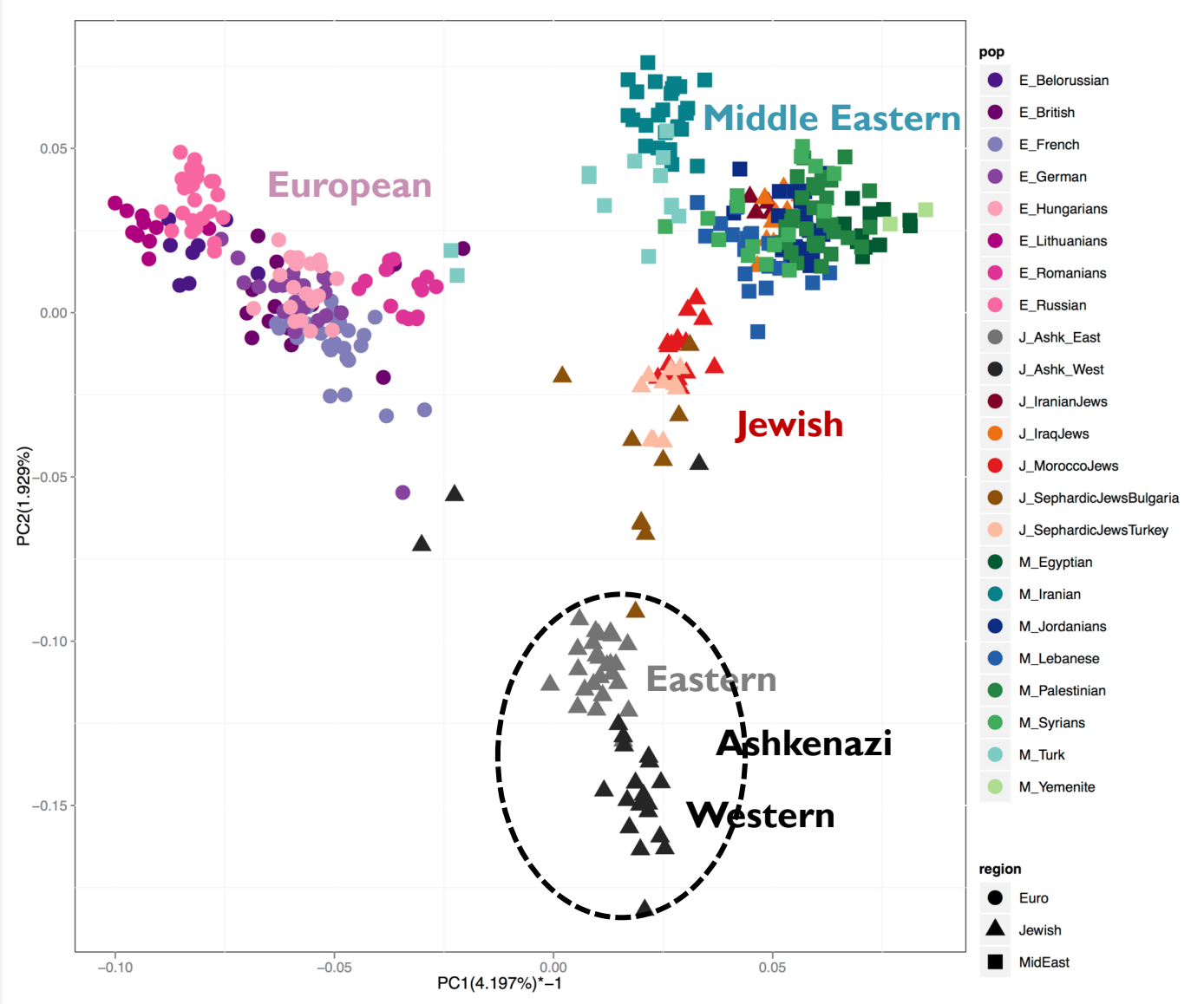
AJ GENETIC RELATIONSHIP TO MIDDLE EASTERN AND EUROPEAN POPULATIONS

- **Principal Component Analysis (PCA)** – a visualization of population genetic structure
- **ADMIXTURE** – visualization of population genetic structure

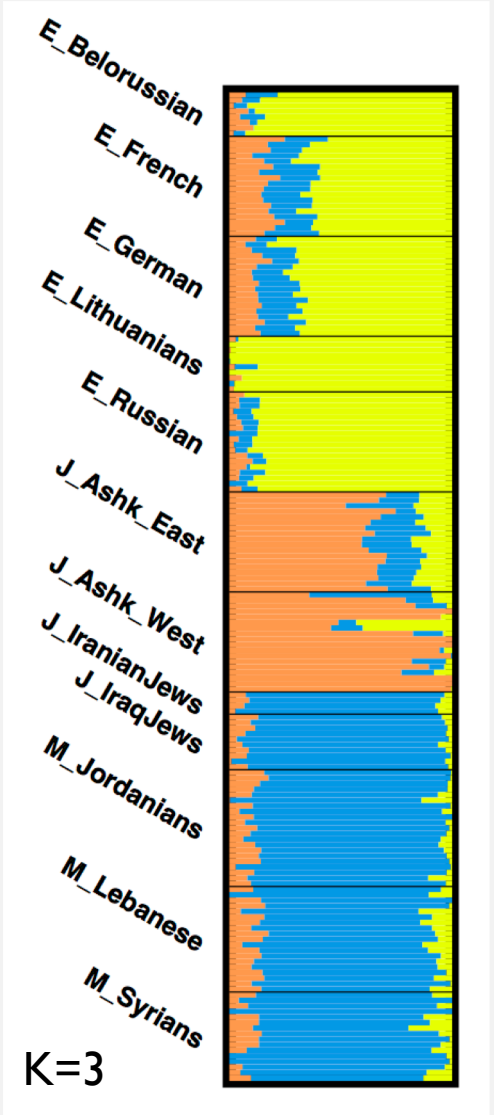
AJ GENETIC RELATIONSHIP TO MIDDLE EASTERN AND EUROPEAN POPULATIONS

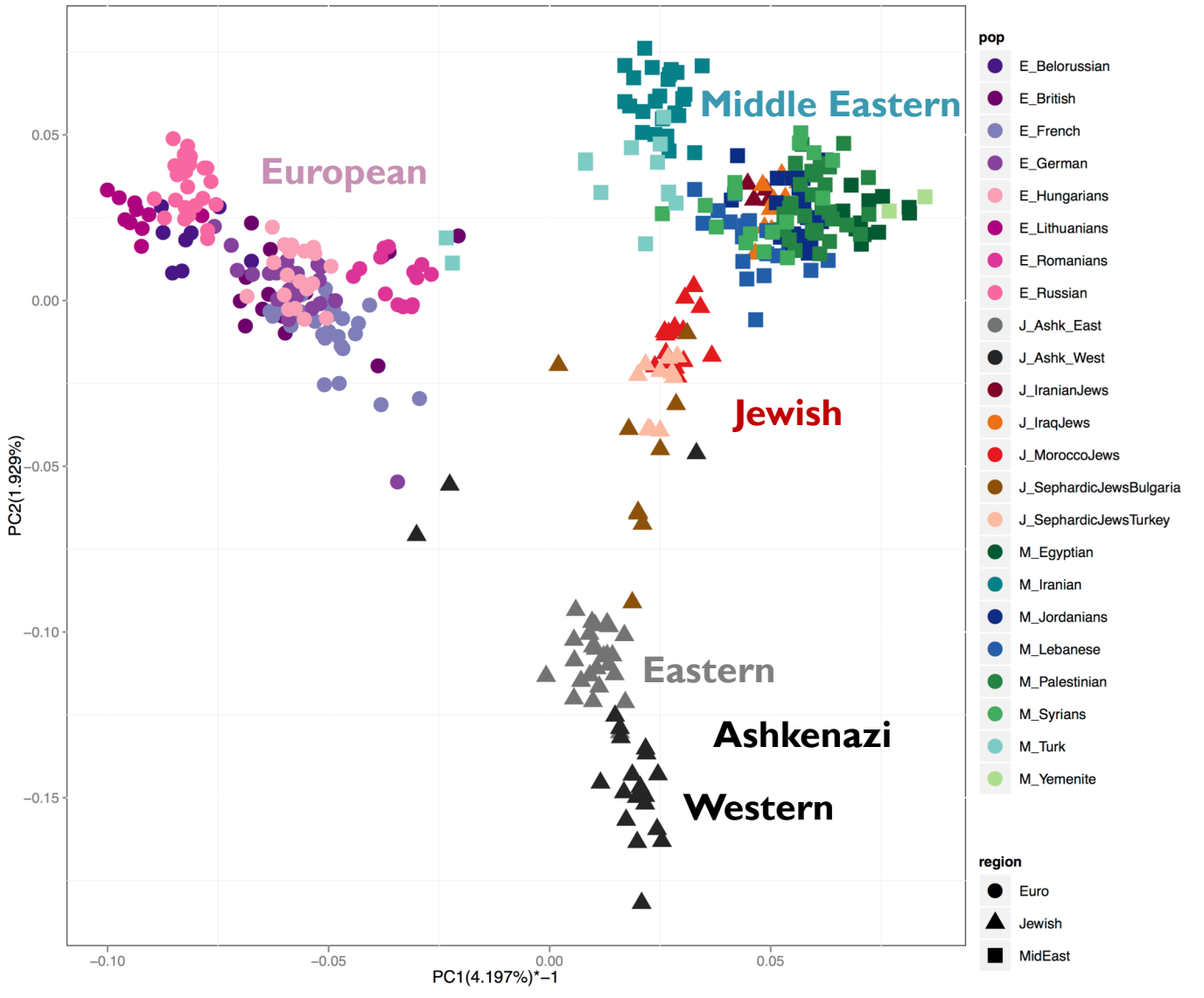


AJ GENETIC RELATIONSHIP TO MIDDLE EASTERN AND EUROPEAN POPULATIONS



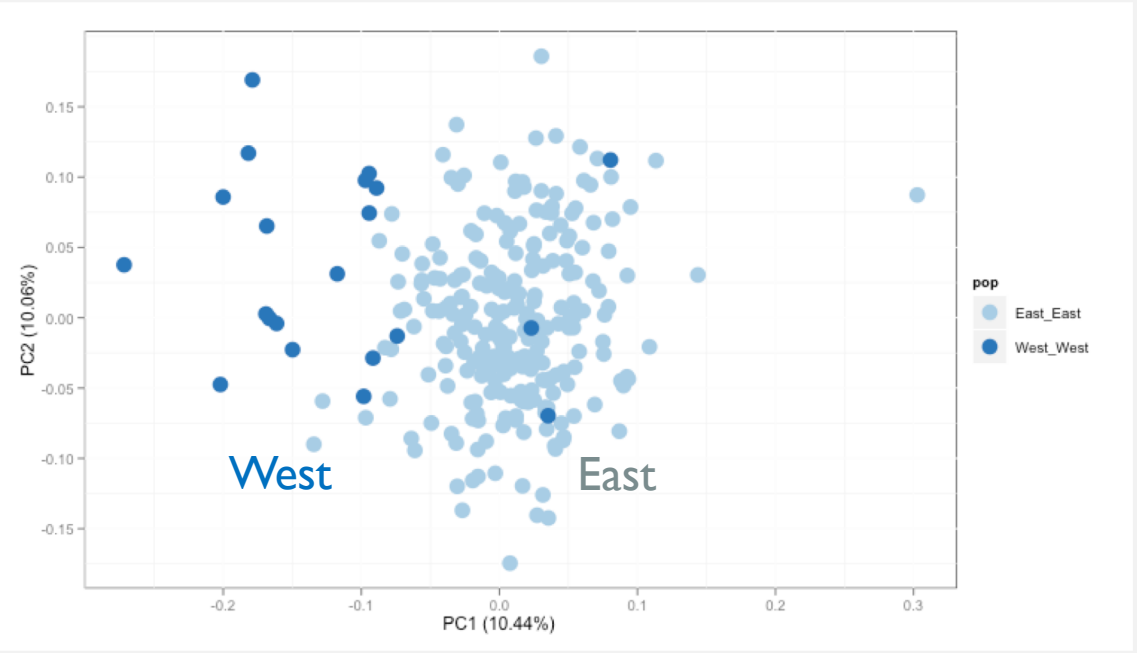
AJ GENETIC RELATIONSHIP TO MIDDLE EASTERN AND EUROPEAN POPULATIONS

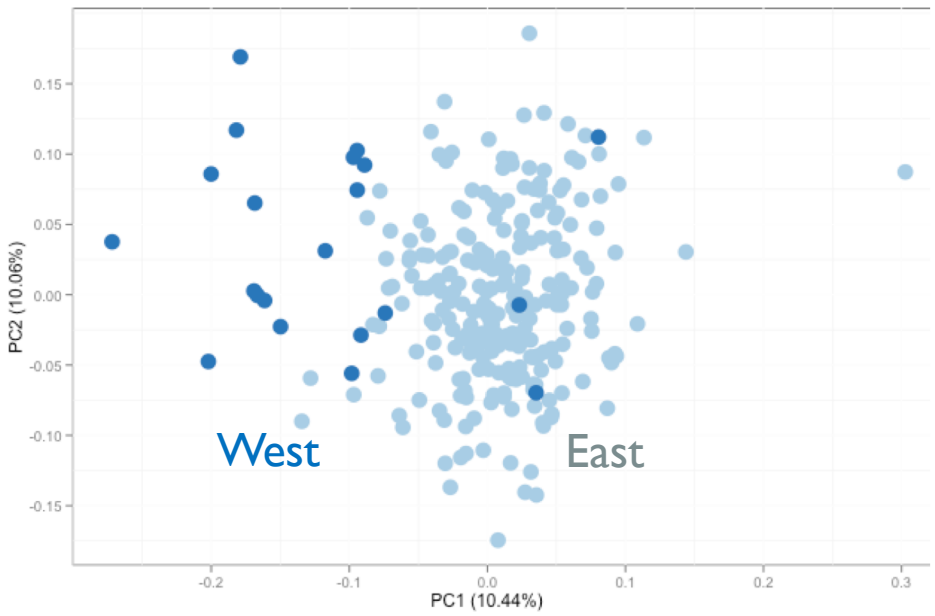


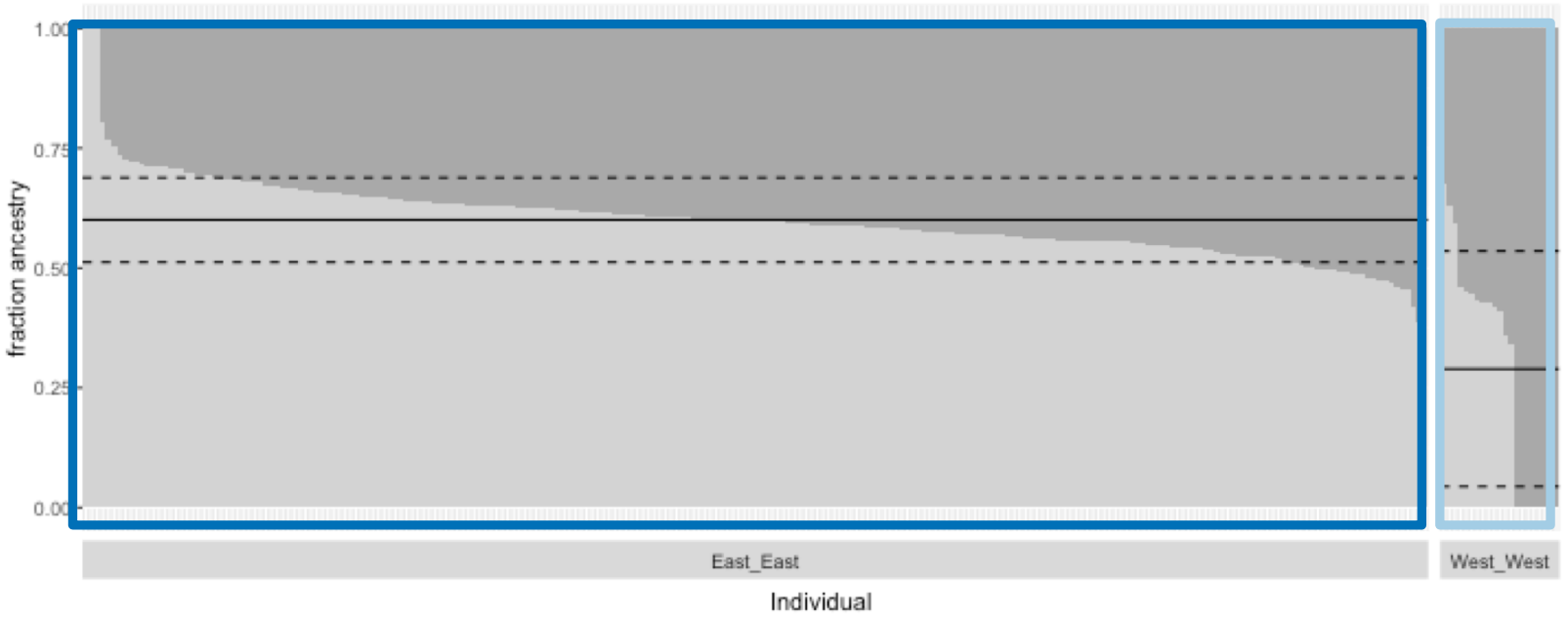


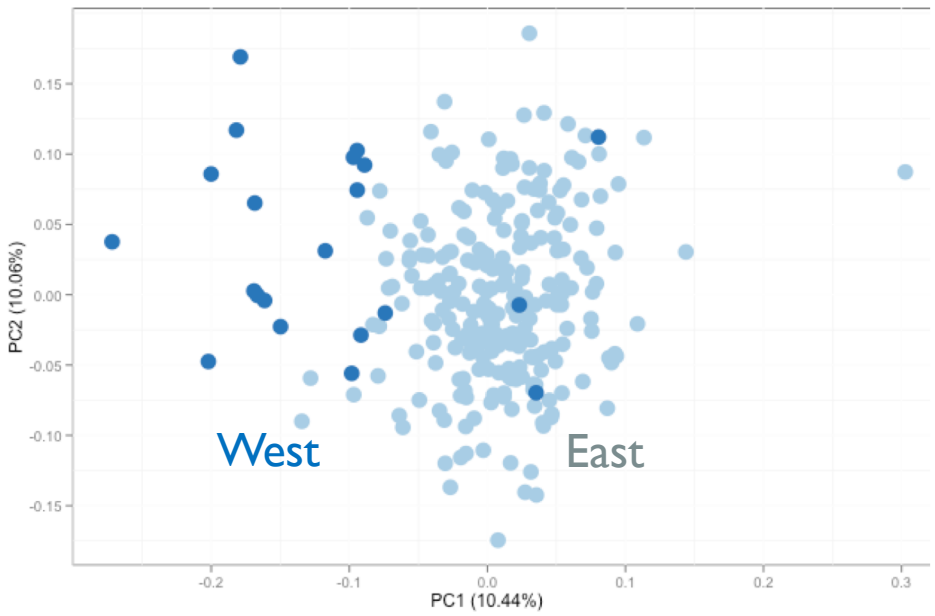
DIFFERENCE BETWEEN EASTERN AND WESTERN?

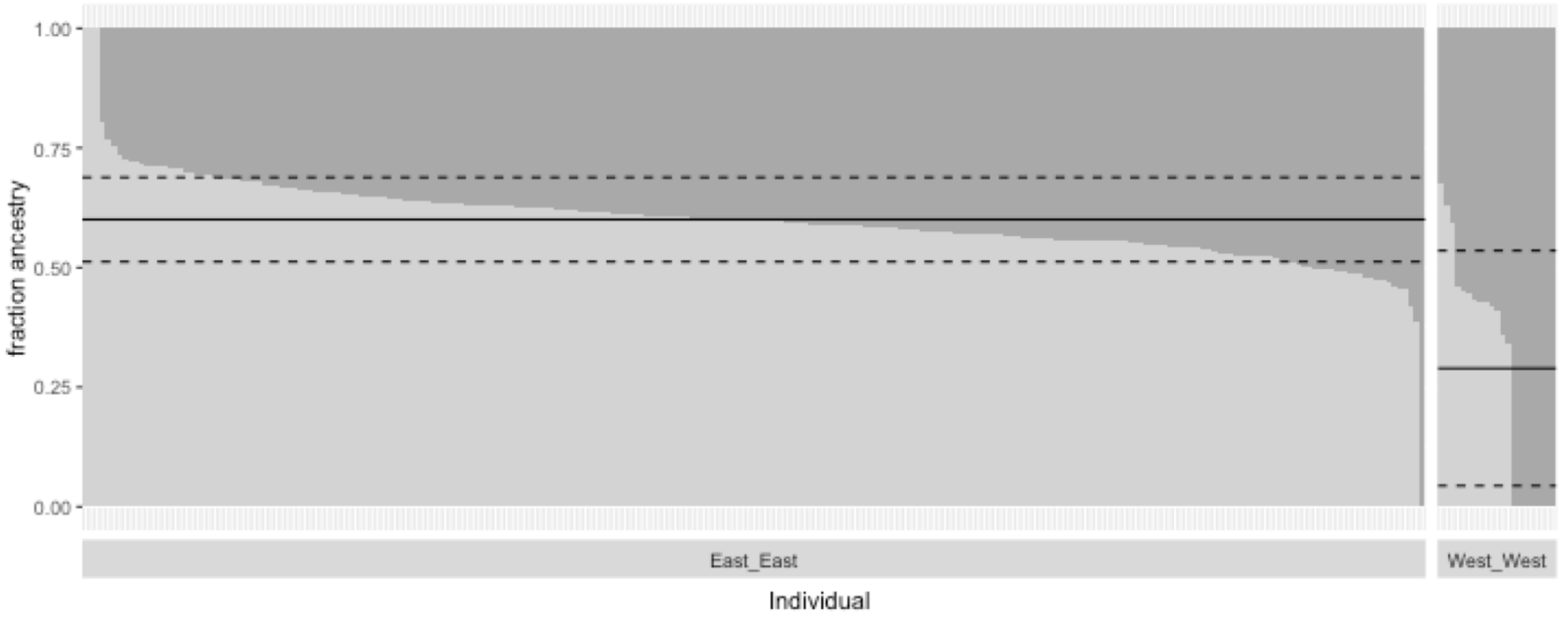
- **Principal Component Analysis (PCA)** – a visualization of population genetic structure
- **ADMIXTURE** – visualization of population genetic structure
- **Runs of homozygosity** – indicates levels of inbreeding or small effective population size
- **Identity by Descent (IBD)** – Indicates shared ancestry between individuals

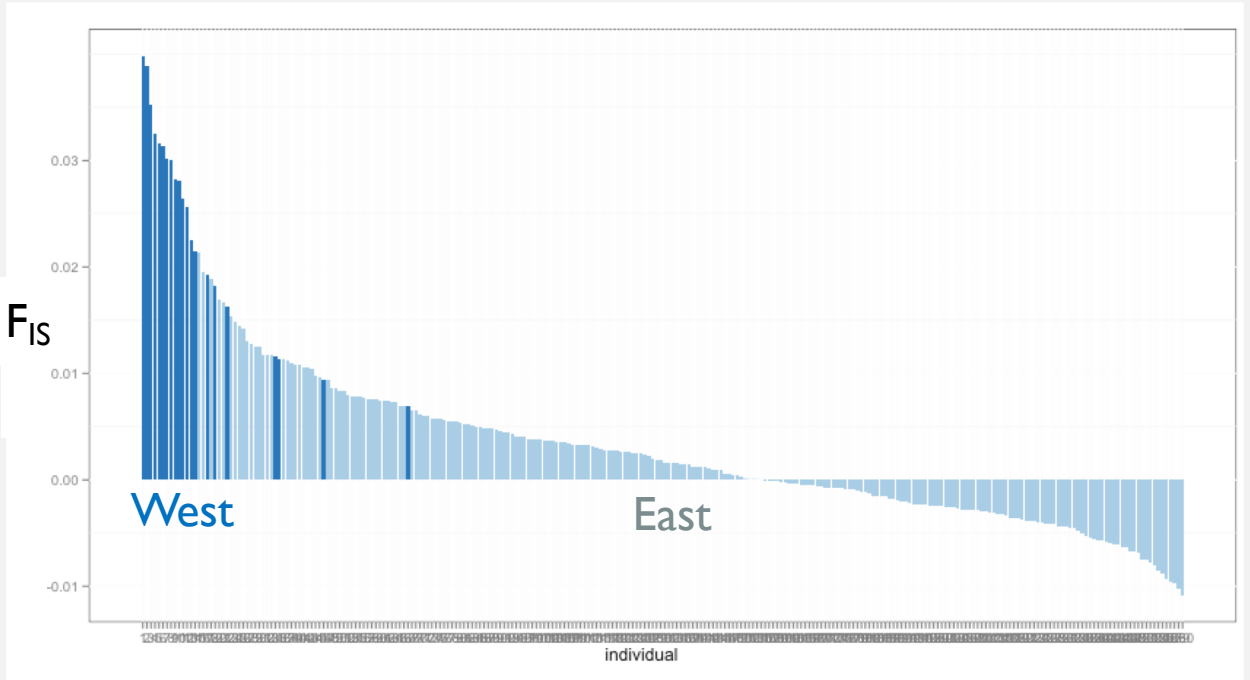

0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
PC2 (10.06%)
-0.2
-0.1
0.0
0.1
0.2
0.3
PC1 (10.44%)
West
East
pop
East_East
West_West

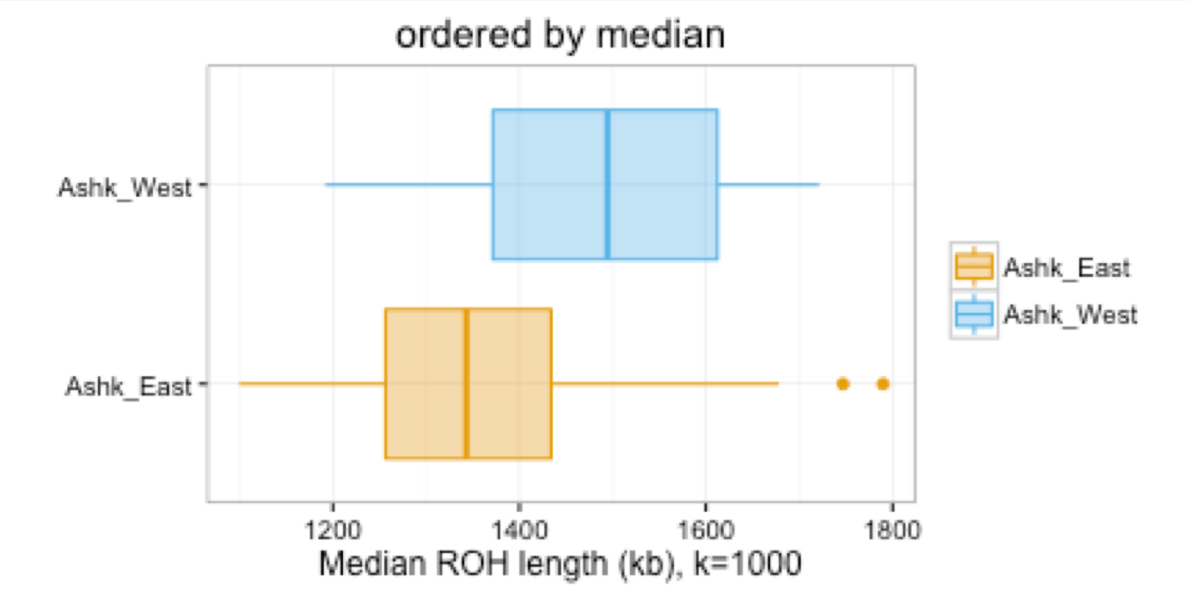

PC2 (10.06%)
PC1 (10.44%)
West
East

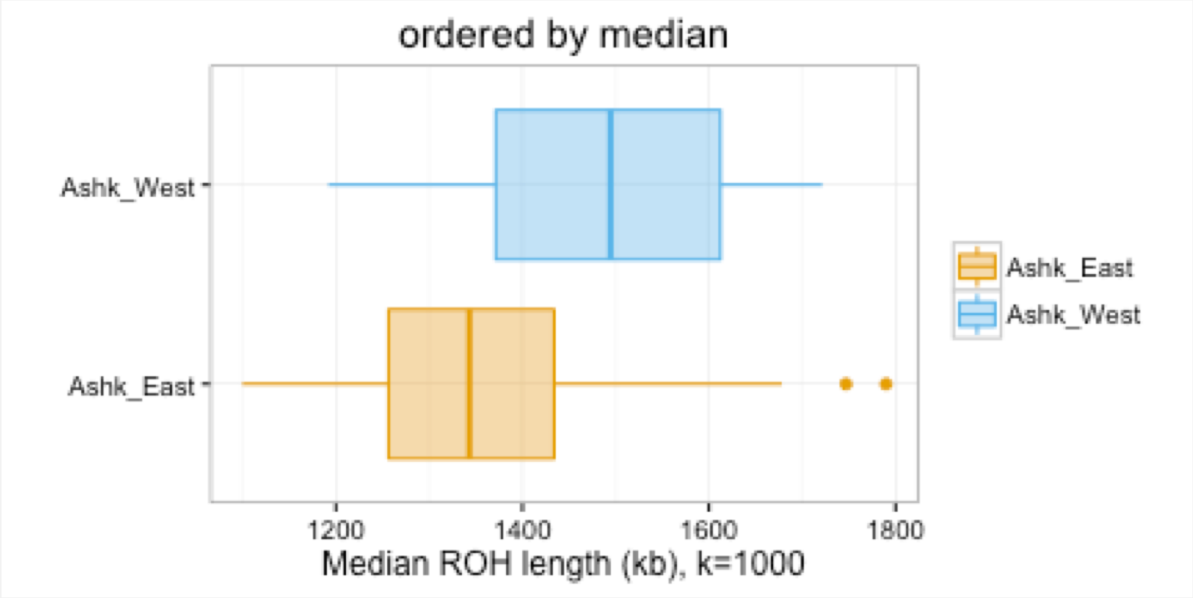

fraction ancestry
Individual
East_East
West_West

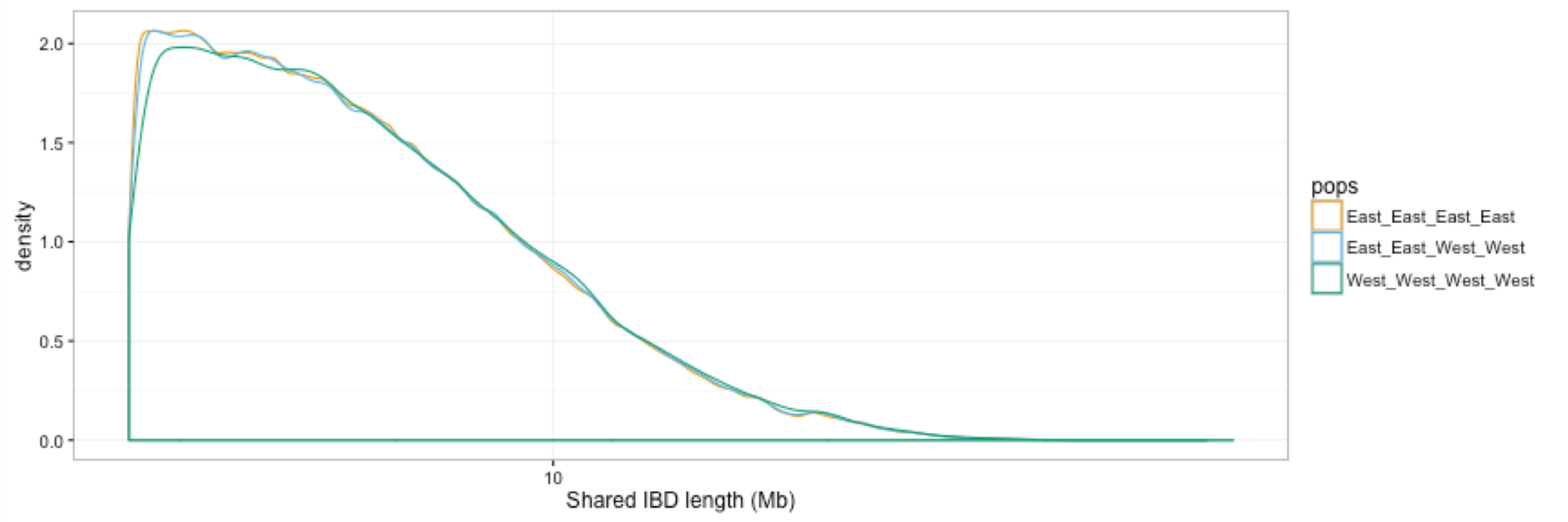

PC2 (10.06%)
PC1 (10.44%)
West
East

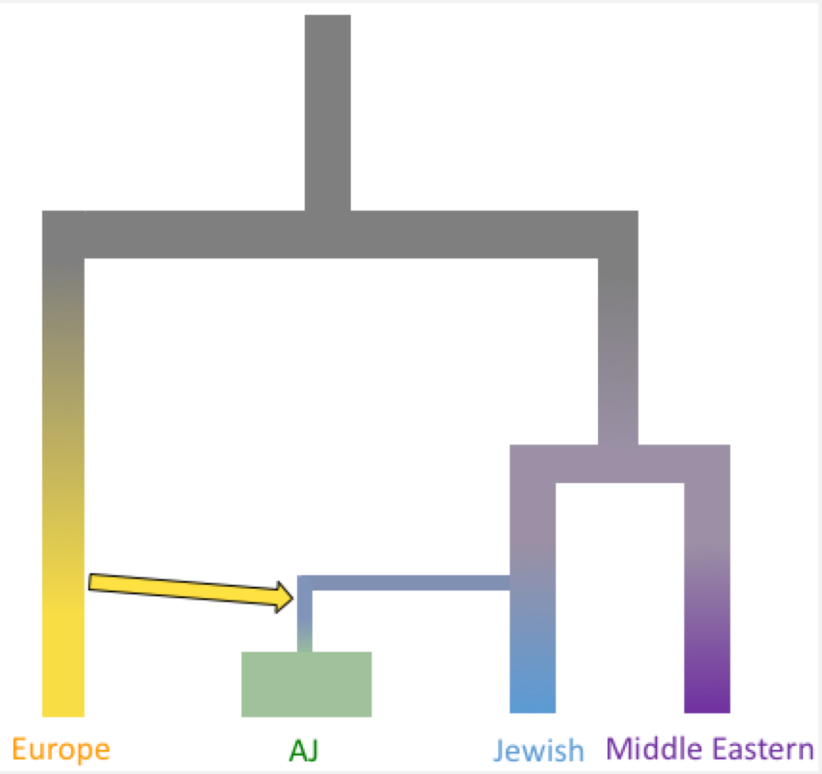

fraction ancestry
East_East
West_West
Individual

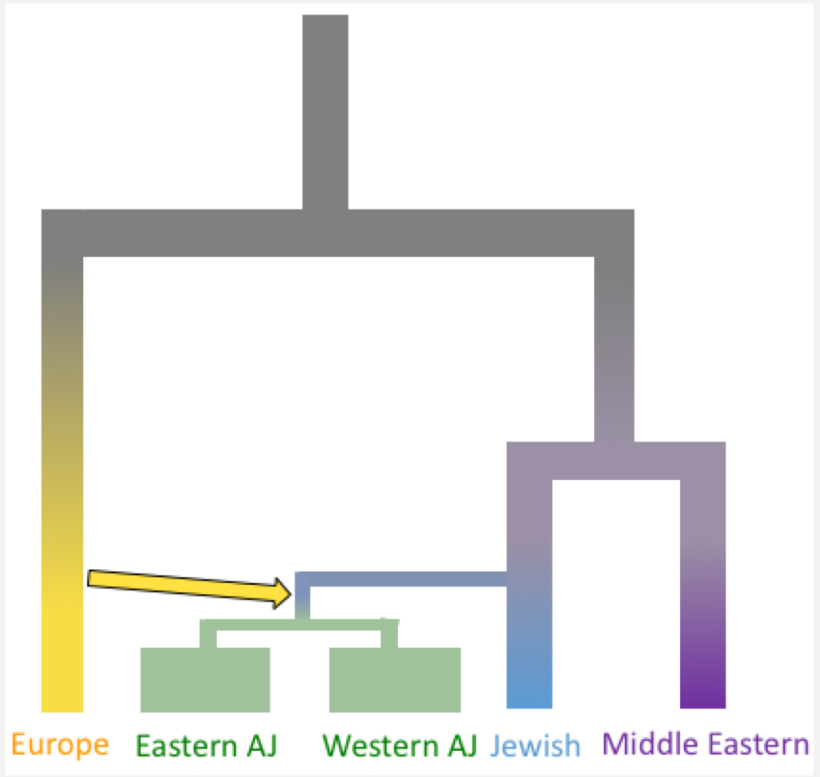

F_{IS}
West
East
individual

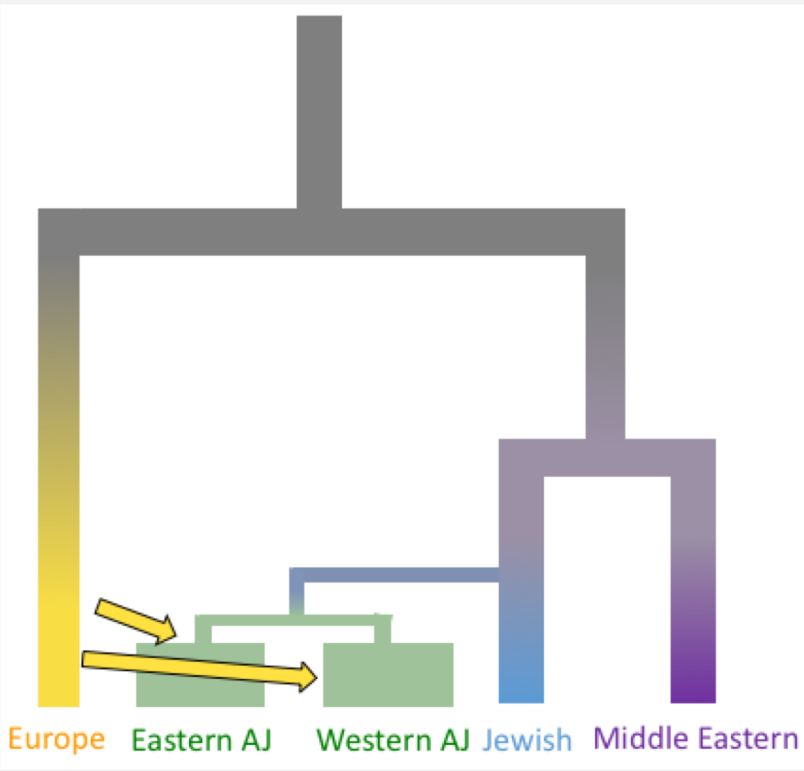

ordered by median
Ashk_West
Ashk_East
1200
1400
1600
1800
Median ROH length (kb), k=1000
Ashk_East
Ashk_West


ordered by median
Ashk_West
Ashk_East
1200
1400
1600
1800
Median ROH length (kb), k=1000
Ashk_East
Ashk_West


2.0
1.5
1.0
0.5
0.0
density
10
Shared IBD length (Mb)
pops
East_East_East_East
East_East_West_West
West_West_West_West

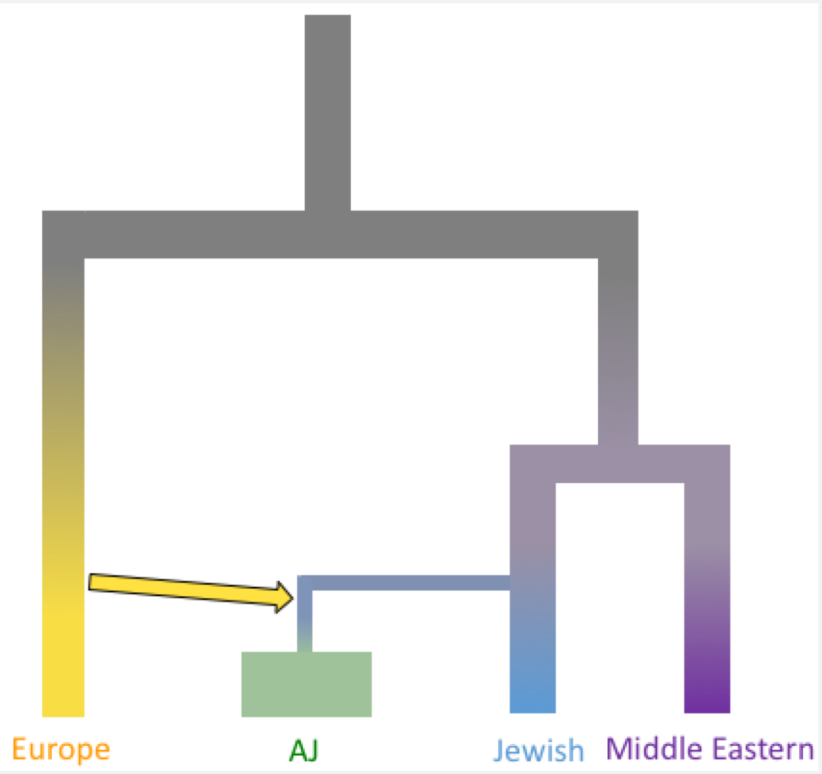


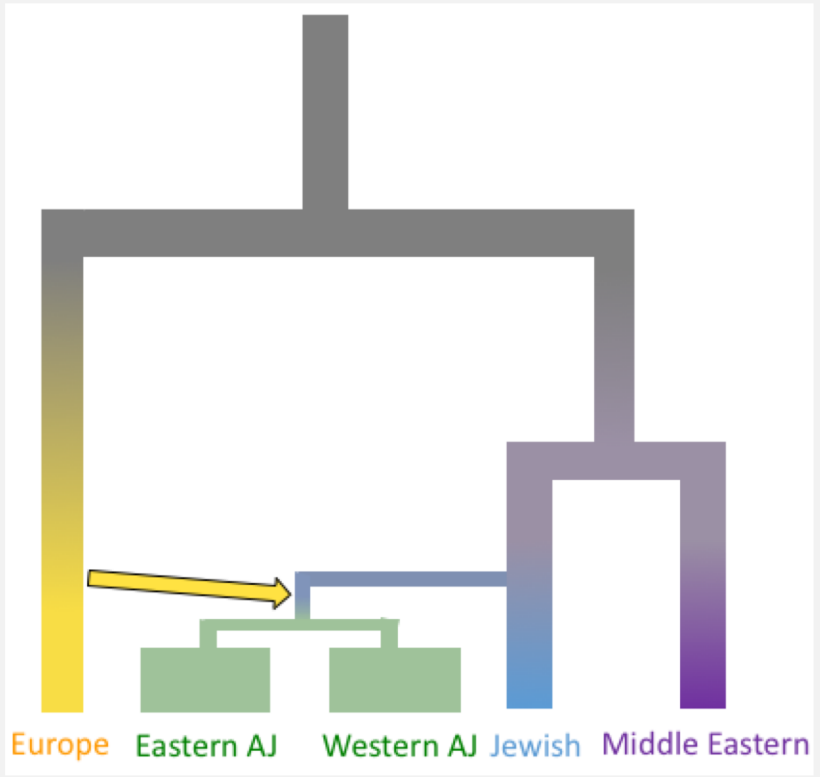


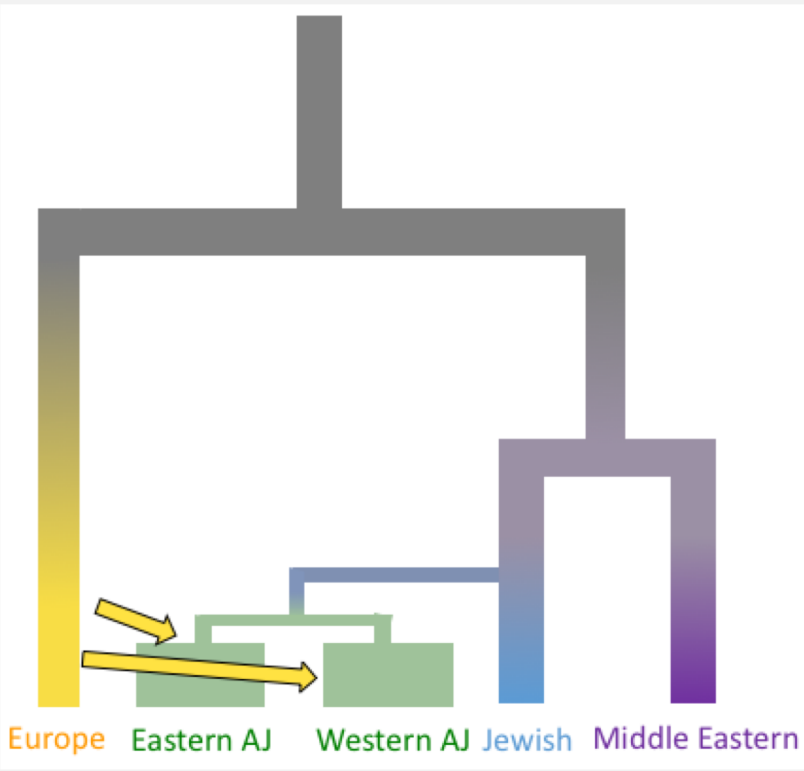


MODELS OF ASHKENAZI HISTORY

APPROXIMATE BAYESIAN COMPUTATION (ABC)

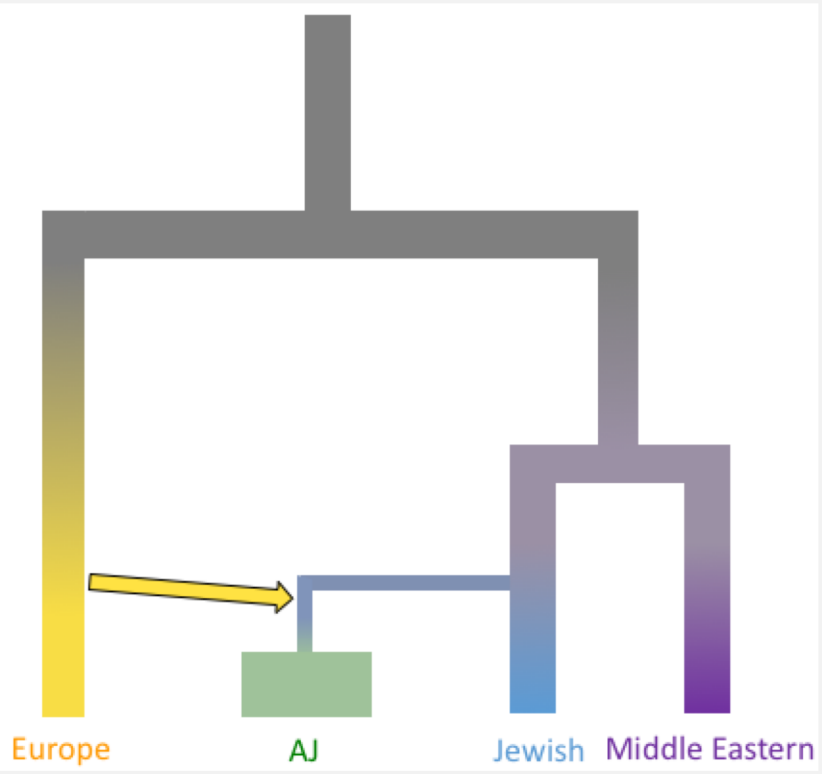


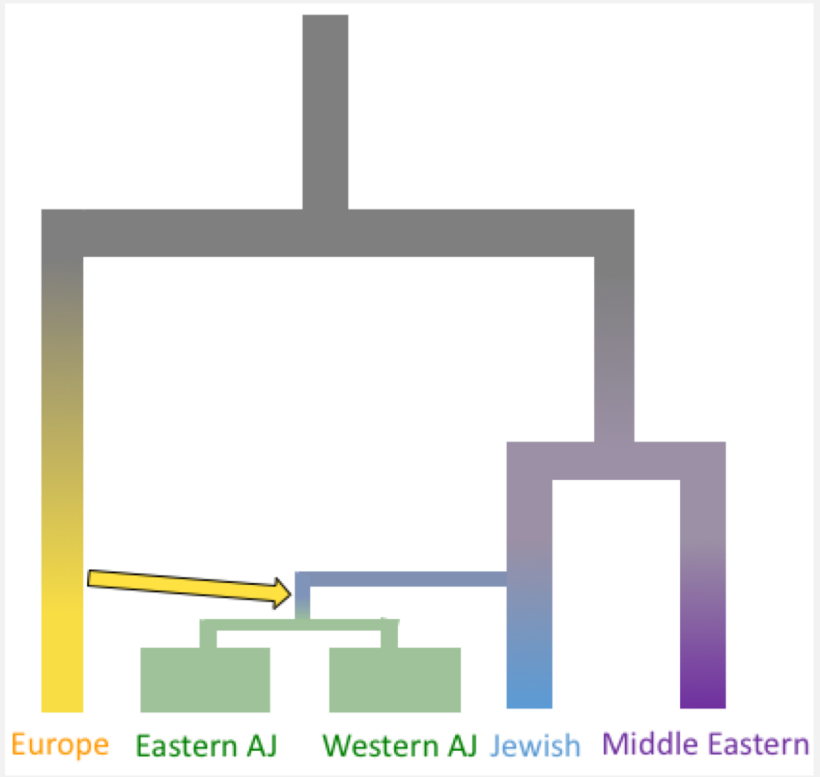


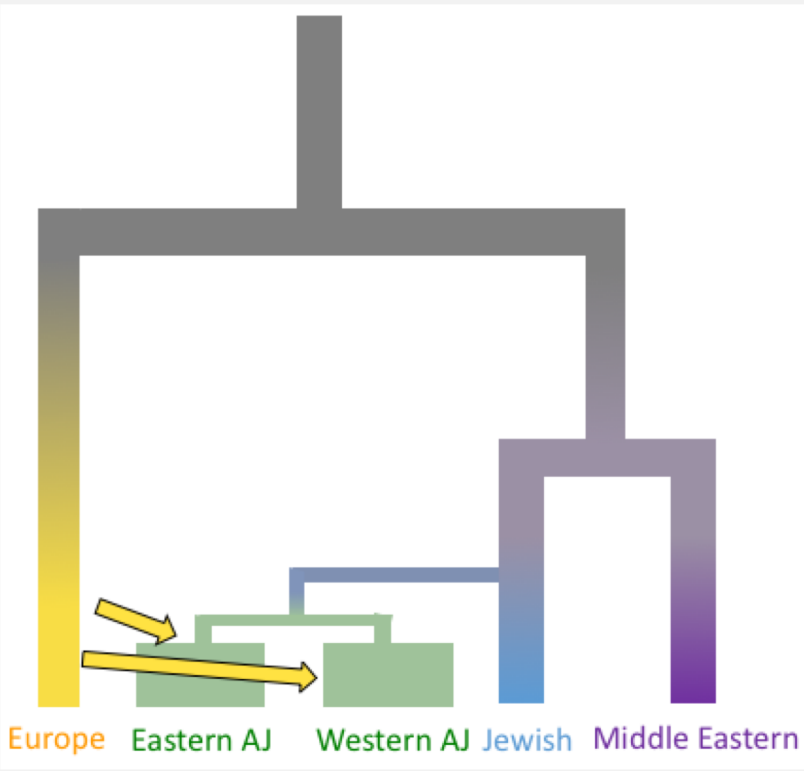


MODELS OF ASHKENAZI HISTORY

APPROXIMATE BAYESIAN COMPUTATION (ABC)







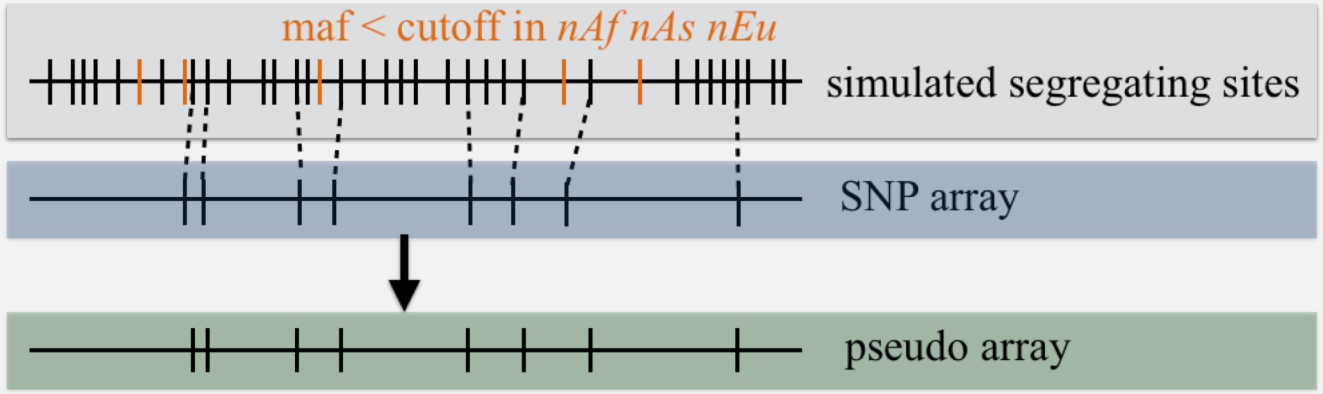
MODELS OF ASHKENAZI HISTORY

APPROXIMATE BAYESIAN COMPUTATION (ABC)

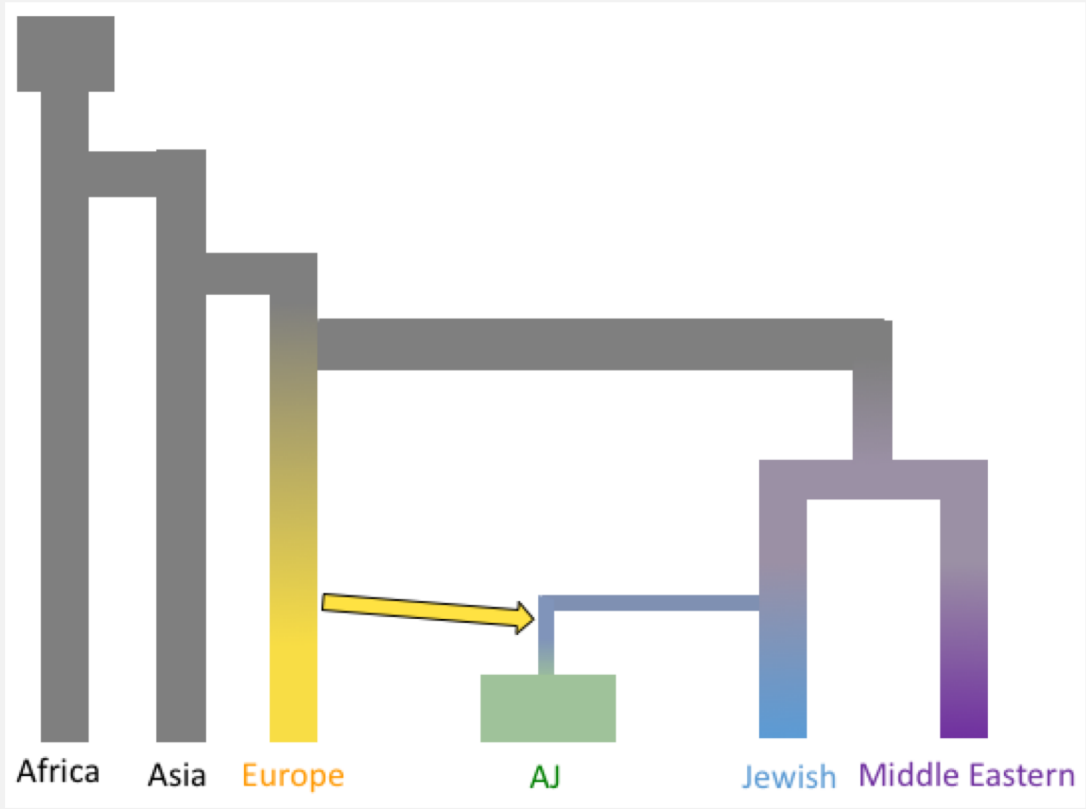
MODEL ASCERTAINMENT BIAS

Ascertainment parameters:

- Sample sizes of discovery populations
- Minor allele frequency cutoff



Discovery populations



ABC WORKFLOW

Model Choice

Parameter Estimation

ABC WORKFLOW

Model Choice

Simulate chr1 ~1×10^6 times for each model

Parameter Estimation

ABC WORKFLOW

Model Choice

Simulate chr1 ~1×10^6 times for each model

↓

Find the best stats for model choice

Parameter Estimation

ABC WORKFLOW

Model Choice

Simulate chr1 ~$1x10^6$ times for each model

Find the best stats for model choice

ABCtoolbox Greedy search algorithm:

1. For all pairs of stats, evaluate the power to distinguish the models, and retain best 10 pairs,
2. Repeat with triplets,
3. And so forth until the set of best combinations does not change anymore.

100,000 simulations, 1000 retained, 100 cross validations.

Keep 77 combinations of stats with power > 0.5, total of 20 stats

ABC WORKFLOW

Model Choice

Simulate chr1 ~$1x10^6$ times for each model

↓

Find the best stats for model choice

↓

Choose best model with ABC

Parameter Estimation

ABC WORKFLOW

Model Choice

Simulate chr1 ~$1x10^6$ times for each model

↓

Find the best stats for model choice

↓

Choose best model with ABC

Parameter Estimation

- 9, 11, 13 parameters
- ~$1x10^6$ simulations
- 1000 retained
- 1000 cross validation

ABC WORKFLOW

Model Choice

Simulate chr1 ~$1x10^6$ times for each model

↓

Find the best stats for model choice

↓

Choose best model with ABC

Parameter Estimation

Simulate chr1 ~$1x10^6$ times

Simulate genome ~$1x10^5$ times

Update chrs ~$1x10^5$ times

ABC WORKFLOW

Model Choice

Simulate chr1 ~$1x10^6$ times for each model

Find the best stats for model choice

Choose best model with ABC

Parameter Estimation

Simulate chr1 ~$1x10^6$ times

Simulate genome ~$1x10^5$ times

Update chrs ~$1x10^5$ times

Transform stats with PLS

ABC WORKFLOW

Model Choice

Simulate chr1 ~$1x10^6$ times for each model

↓

Find the best stats for model choice

↓

Choose best model with ABC

Parameter Estimation

Simulate chr1 ~$1x10^6$ times

Simulate genome ~$1x10^5$ times

Update chrs ~$1x10^5$ times

↓

Transform stats with PLS

↓

Estimate parameters

ABC WORKFLOW

Model Choice

Simulate chr1 ~$1x10^6$ times for each model

Find the best stats for model choice

Choose best model wit ABC

- 11 parameters
- ~$1x10^6$ or ~$1x10^5$ simulations
- 10 PLS components
- 1000 retained

Parameter Estimation

Simulate chr1 ~$1x10^6$ times

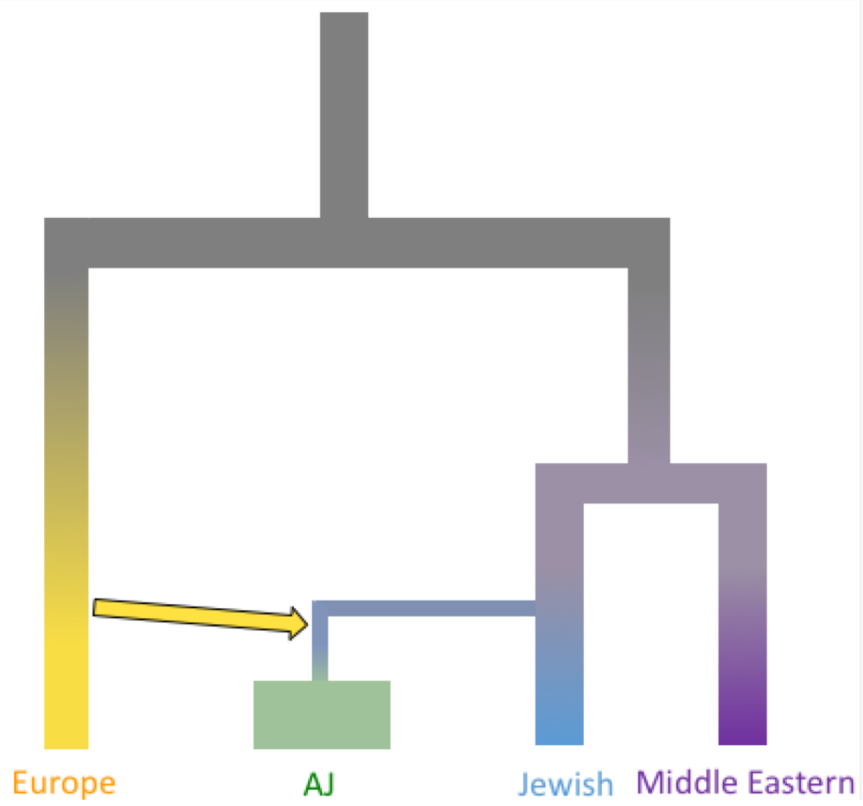
Simulate genome ~$1x10^5$ times

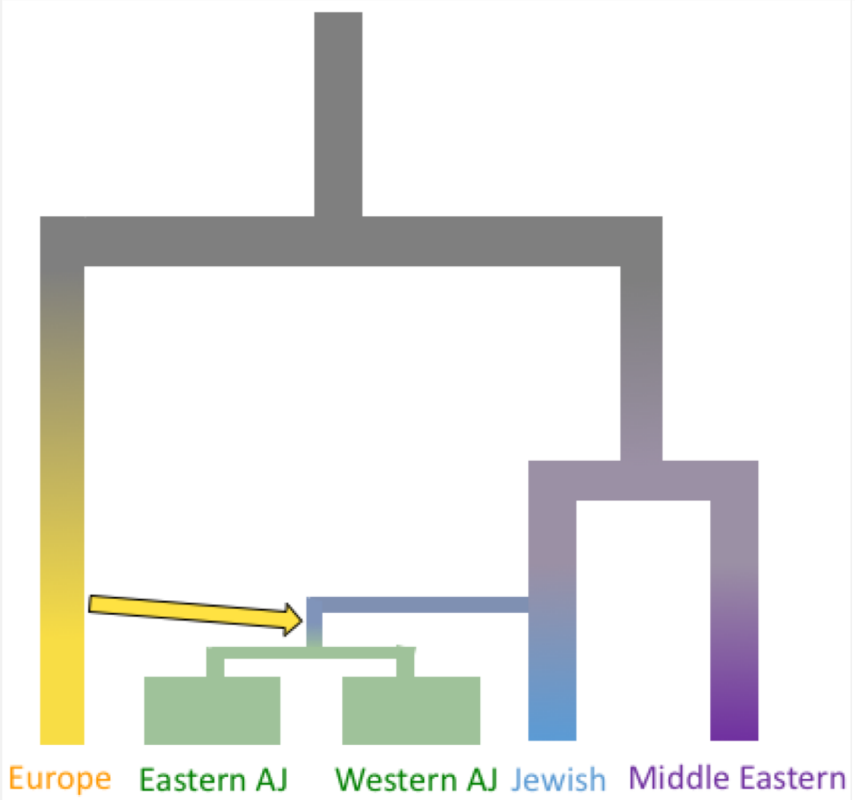
Update chrs ~$1x10^5$ times

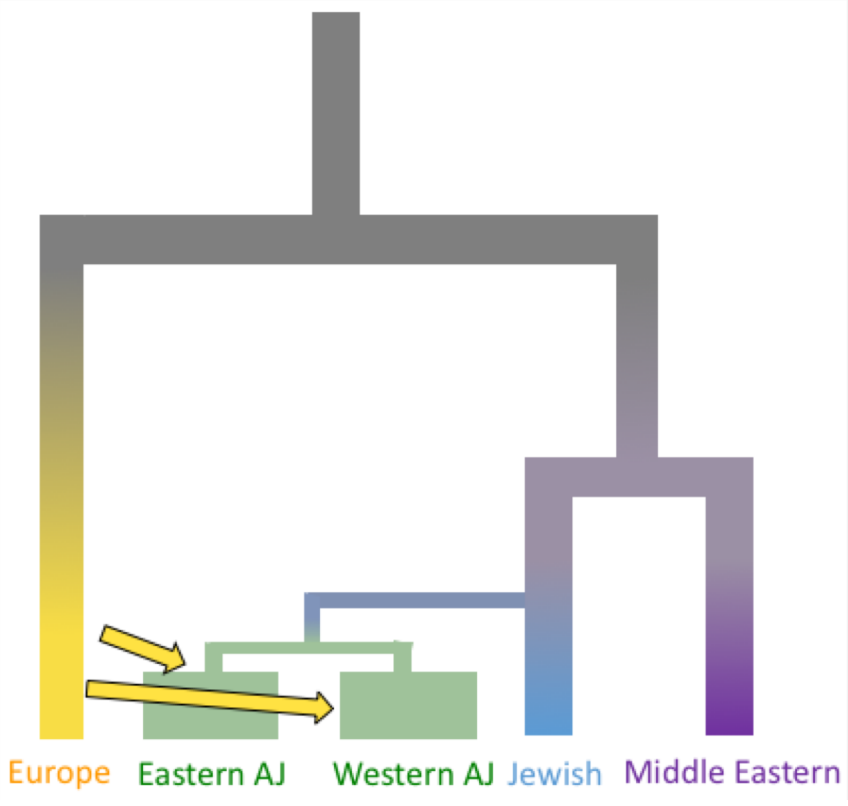
Transform stats with PLS

Estimate parameters

MODEL CHOICE

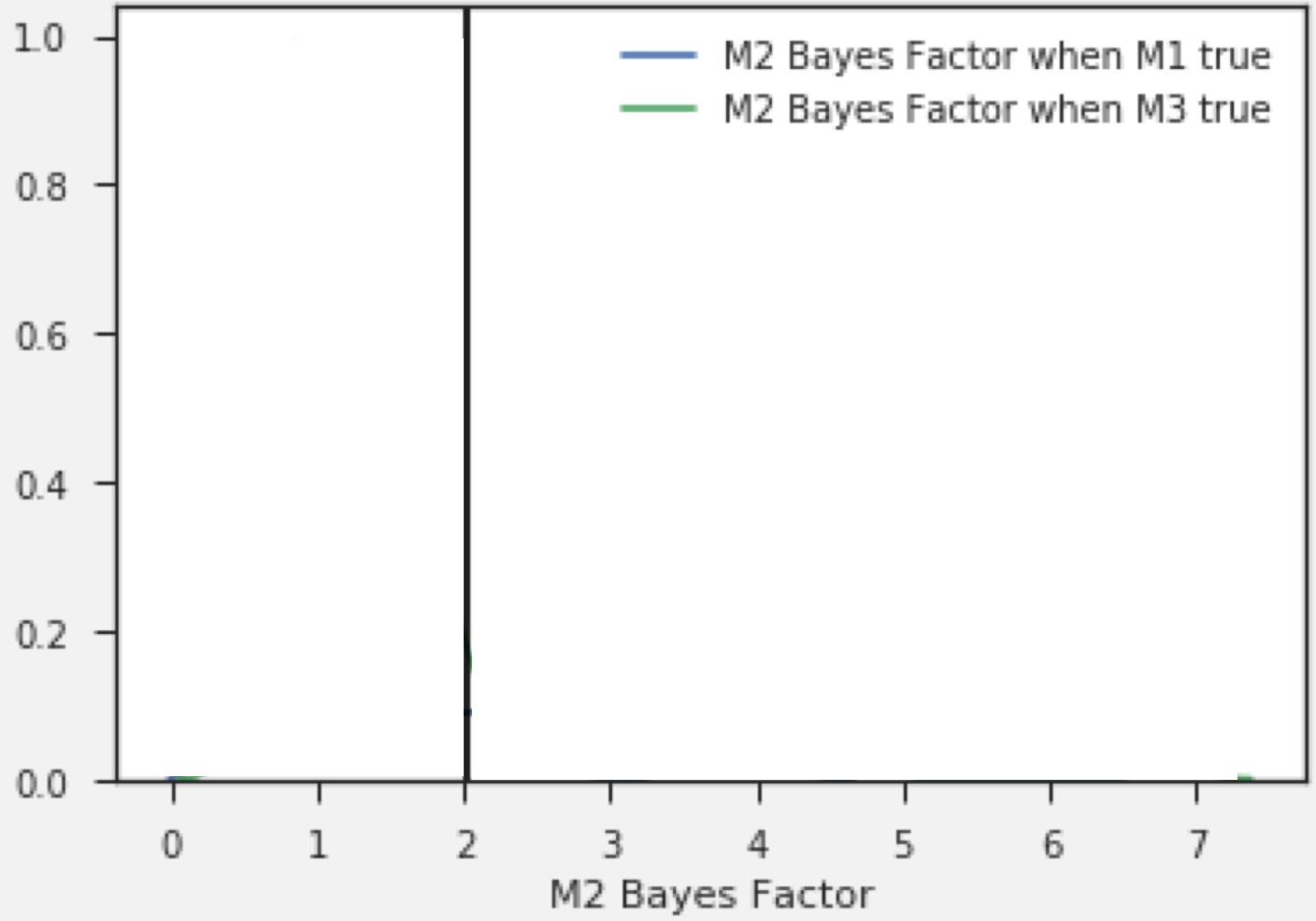




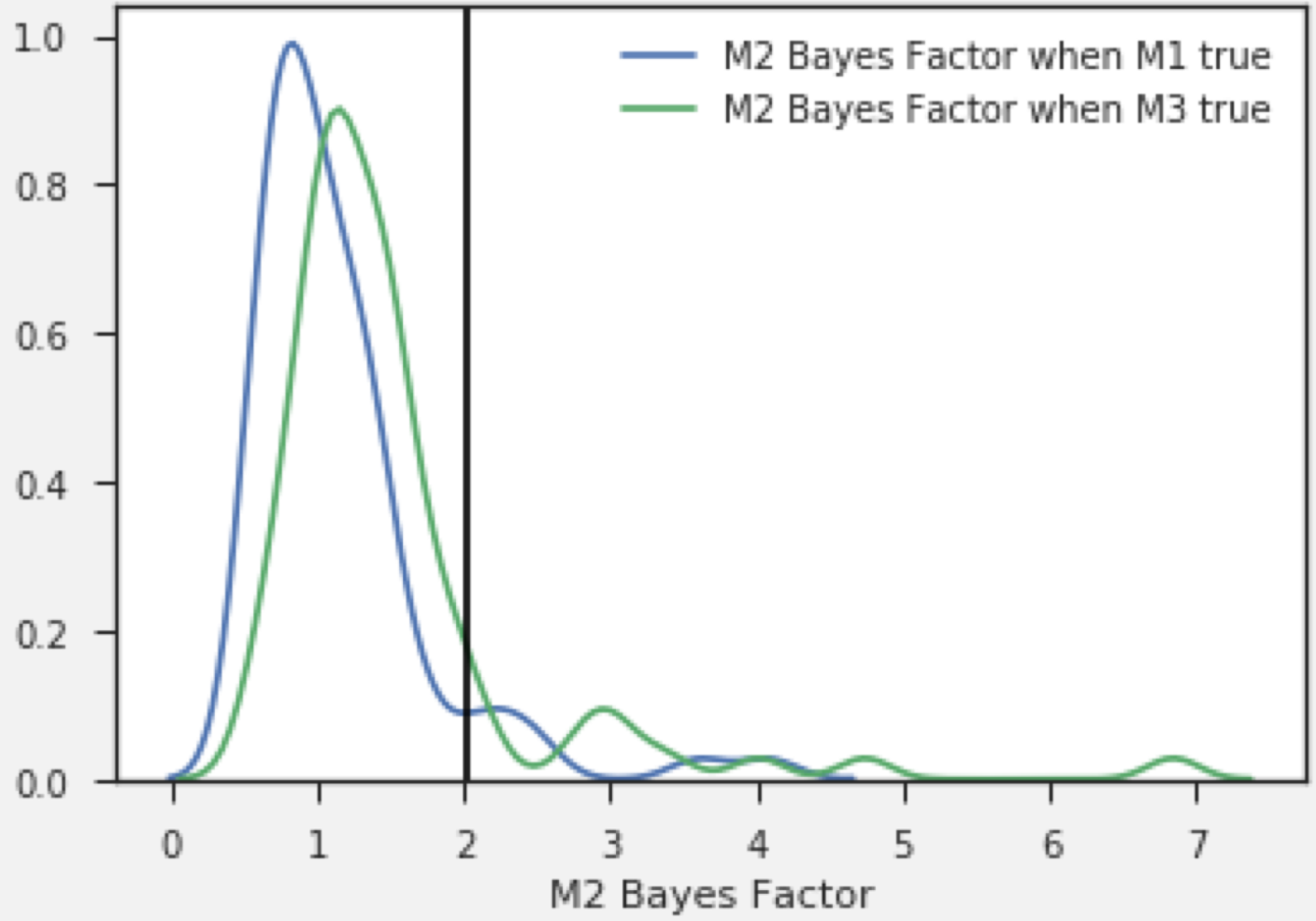


Posterior probability:	0.0005	0.67	0.33
Bayes Factor:	0.0005	2.01	0.50

CROSS VALIDATION OF MODEL CHOICE



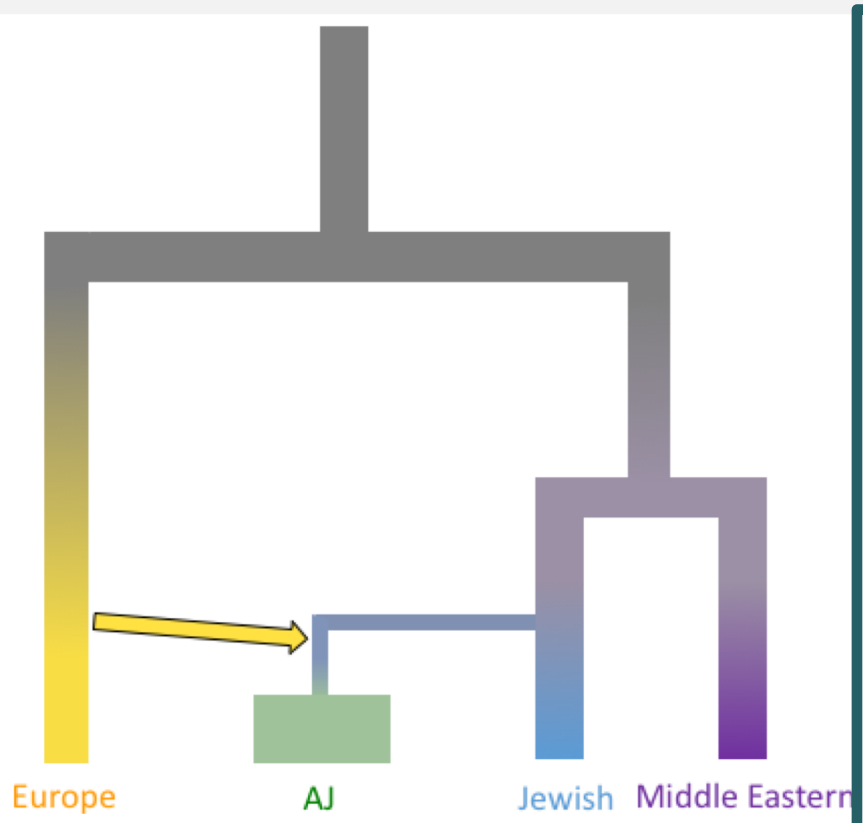
CROSS VALIDATION OF MODEL CHOICE

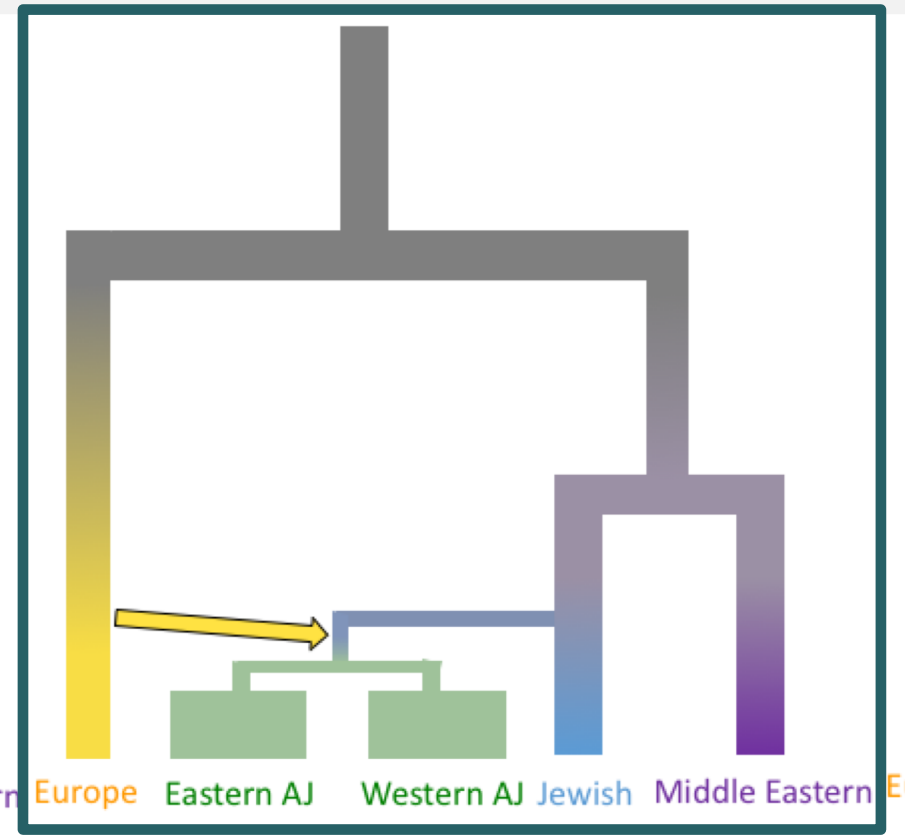


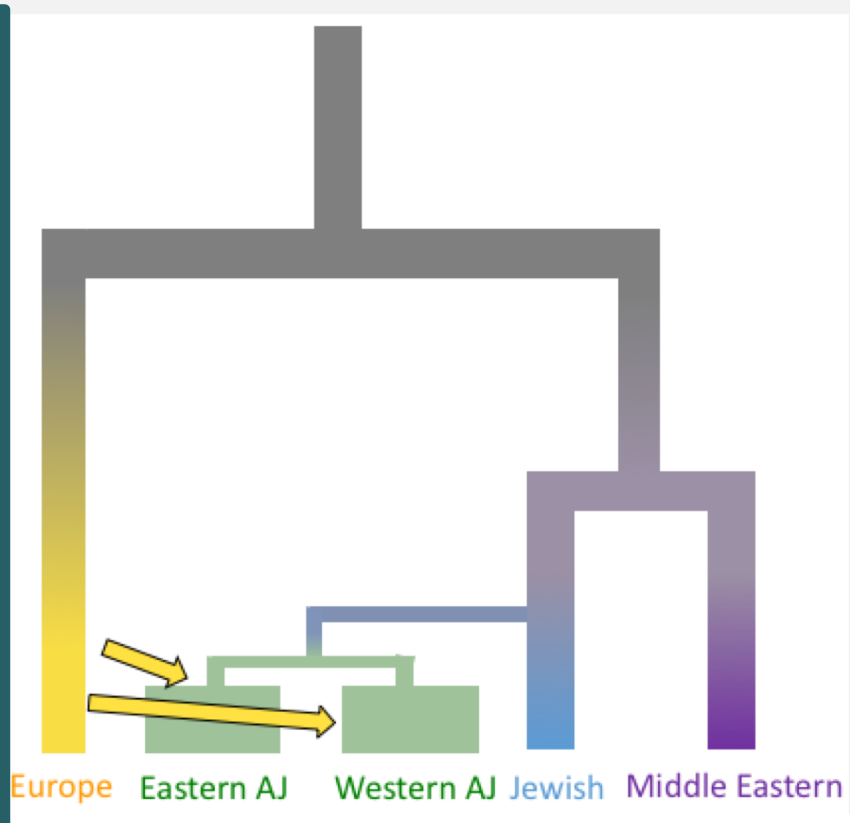
0.92 Model 2 Bayes factors greater than Model 1 Bayes factors when Model 1 is true model

0.86 Model 2 Bayes factors greater than Model 3 Bayes factors when Model 3 is true model

MODEL CHOICE



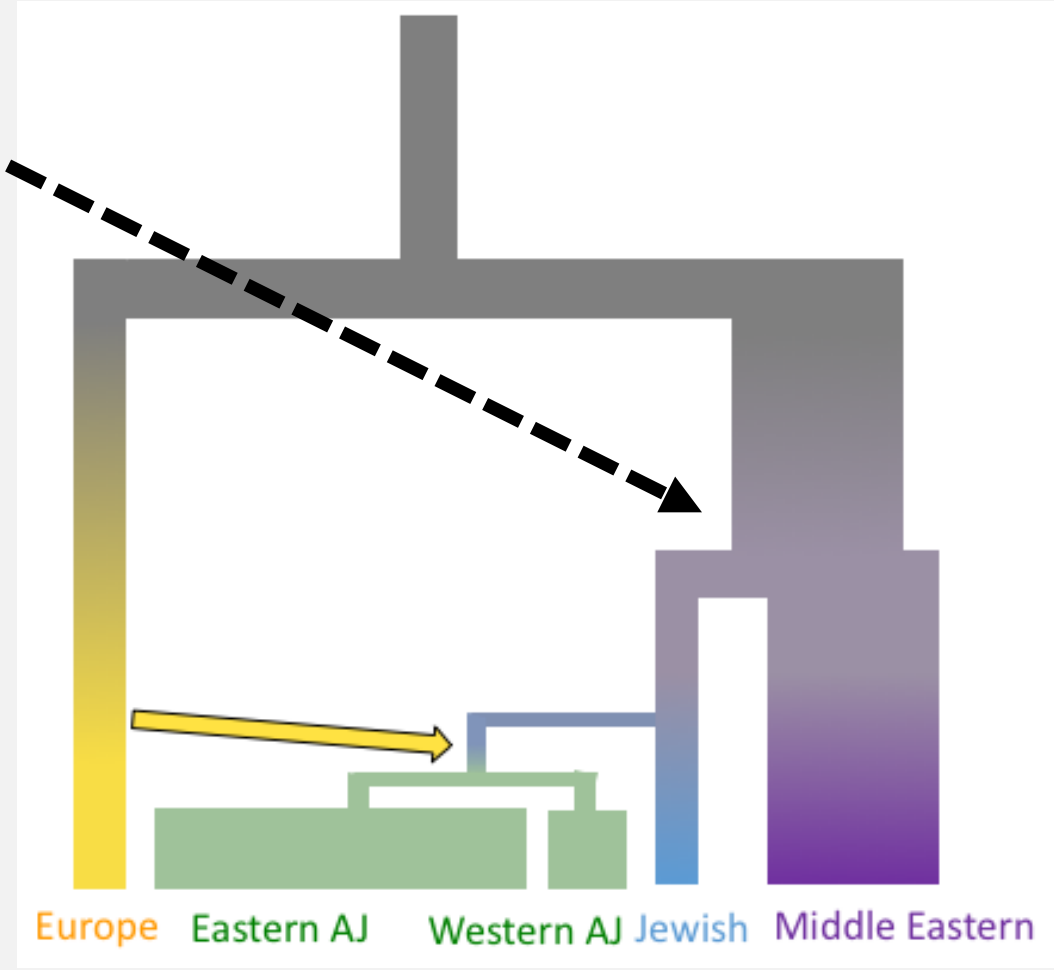




Posterior probability:	0.0005	0.67	0.33
Bayes Factor:	0.0005	2.01	0.50
Prob. false negative:	0.08		0.14

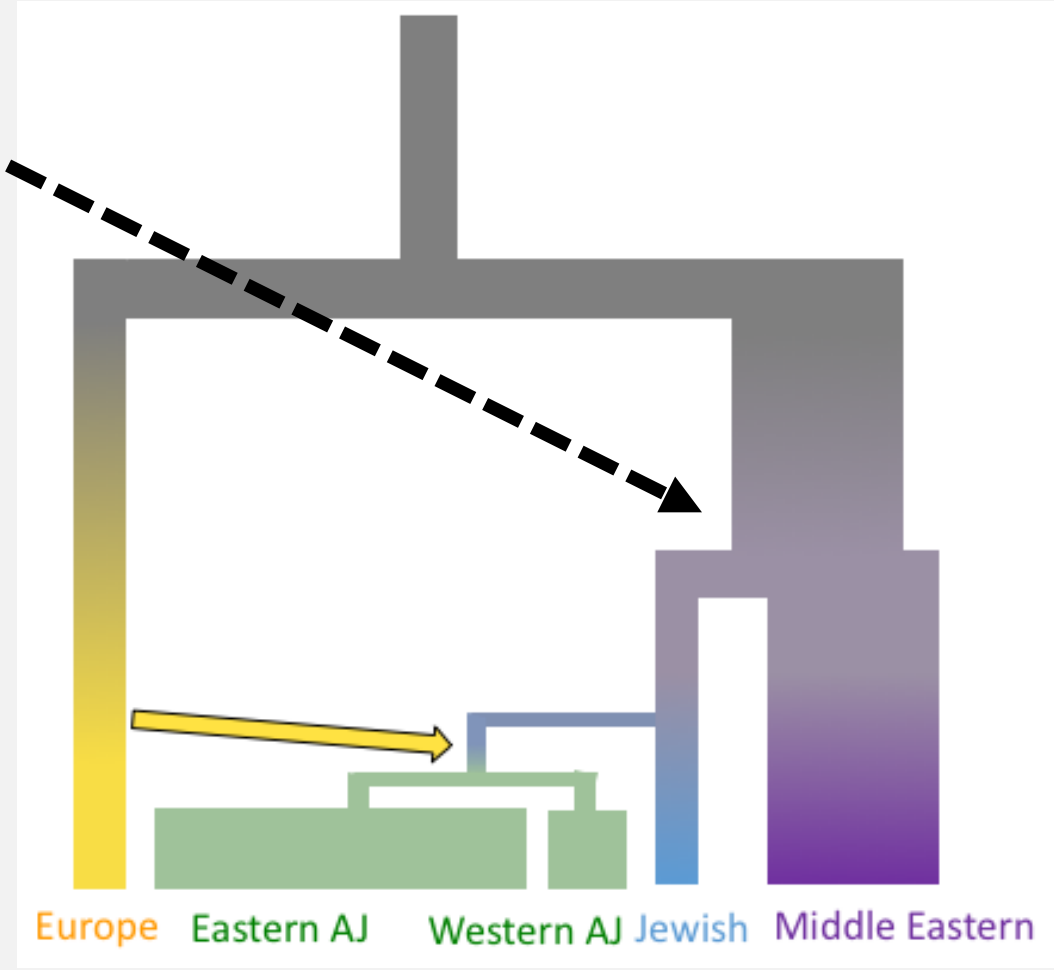
BEST MODEL

- ~ 3000 BCE ancestors of Jewish populations diverged from other Middle Eastern populations
  - Experienced extreme population size reduction



BEST MODEL

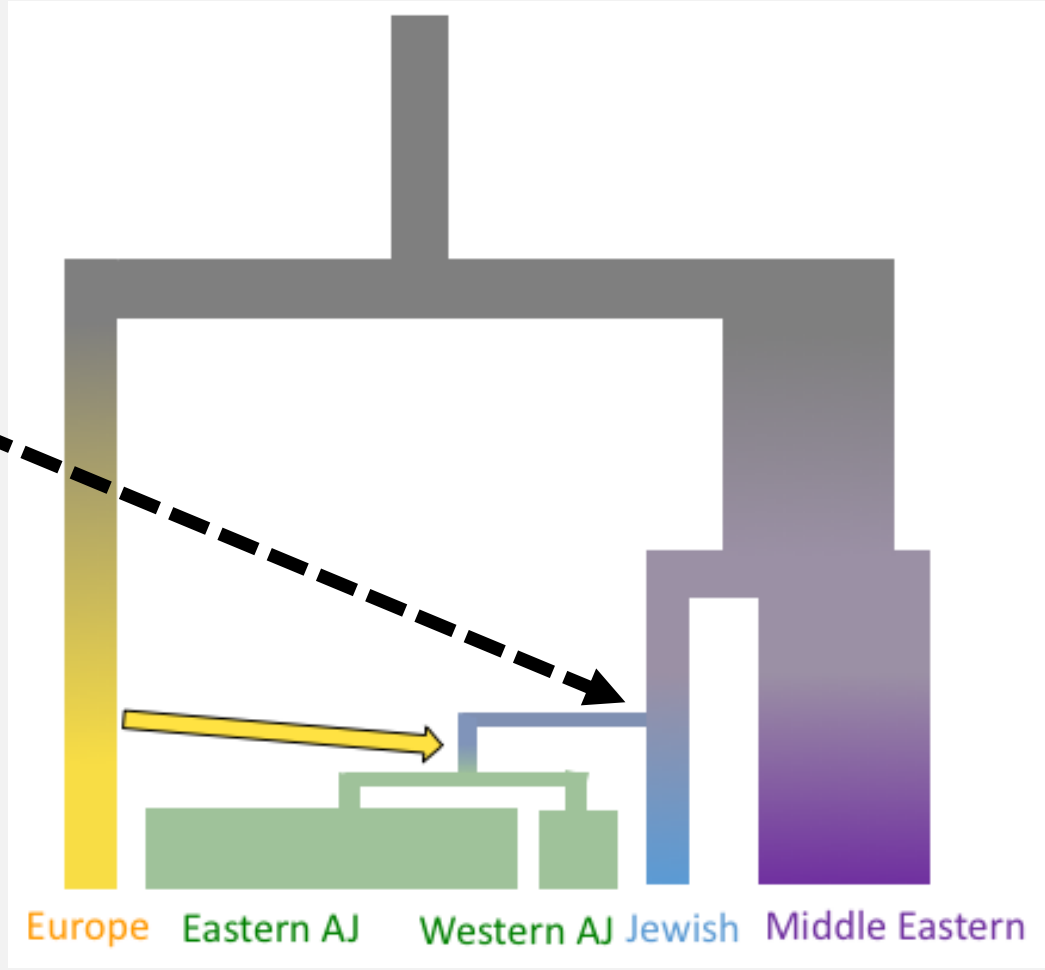
- ~ 3000 BCE ancestors of Jewish populations diverged from other Middle Eastern populations
  - Experienced extreme population size reduction



- First written accounts of "Israel" from Merneptah Stele in 1207 BCE

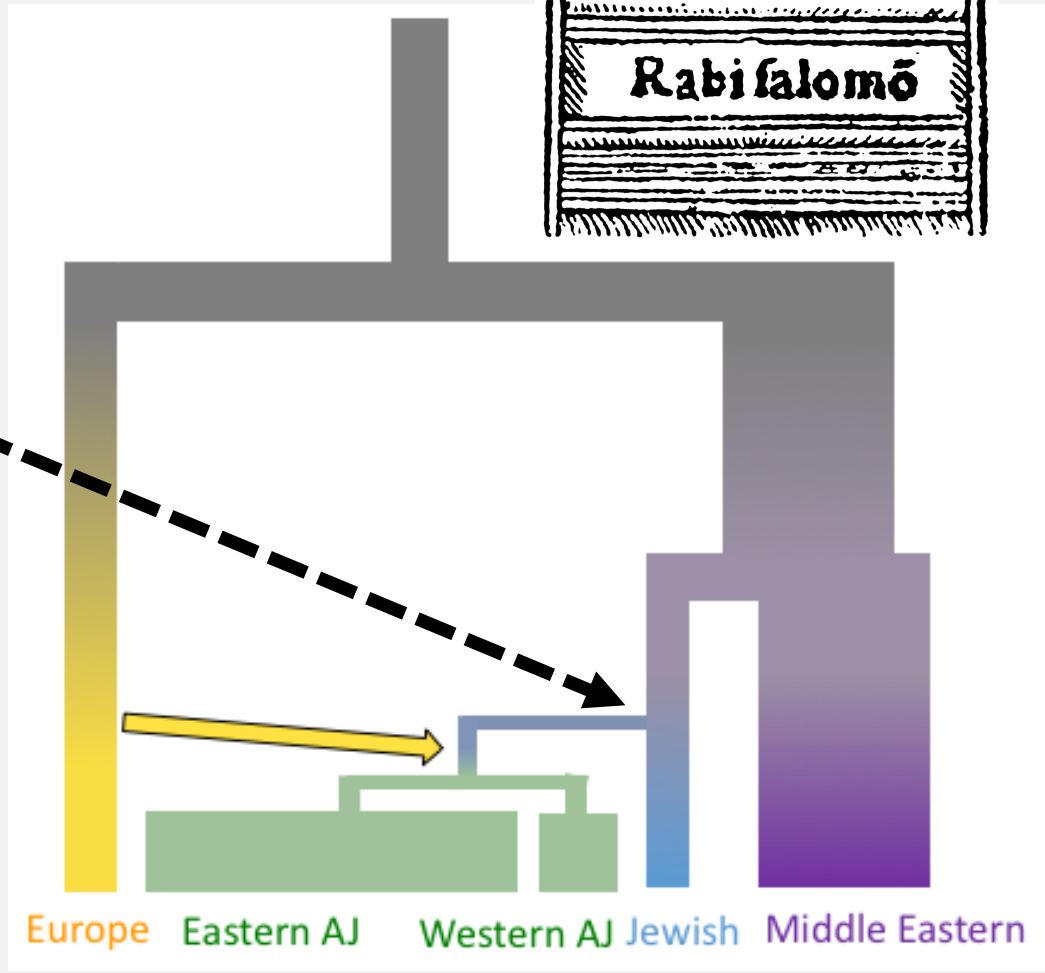
BEST MODEL

- ~ 3000 BCE ancestors of Jewish populations diverged from other Middle Eastern populations
  - Experienced extreme population size reduction
- 13th century ancestors of Ashkenazi Jews diverged from other Jewish populations
  - Experienced another population size reduction



BEST MODEL

- ~ 3000 BCE ancestors of Jewish populations diverged from other Middle Eastern populations
  - Experienced extreme population size reduction
- 13th century ancestors of Ashkenazi Jews diverged from other Jewish populations
  - Experienced another population size reduction

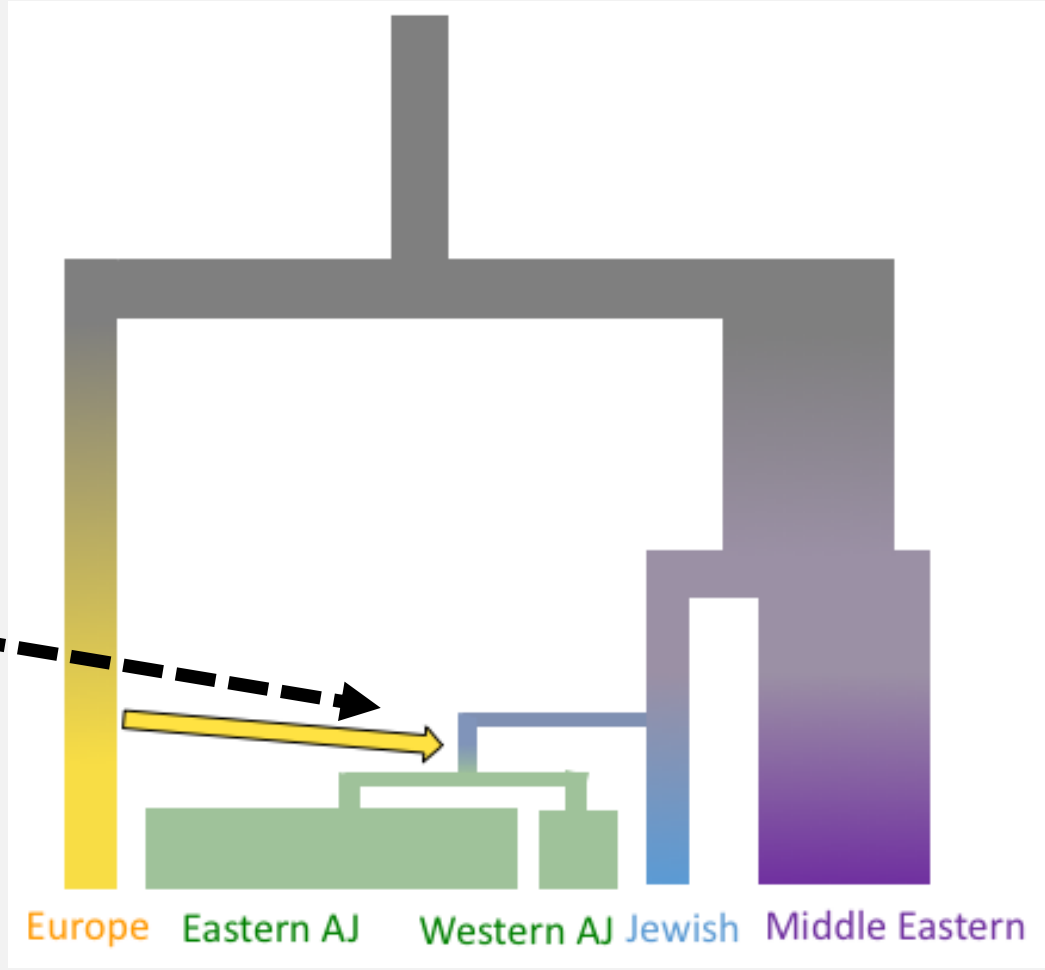


- Migrations northward from Italy led to AJ community in Rhine Valley by 10th century.
- In the late 10th, 11th, and 12th centuries charters were issued to protect Jews in towns.
- In the 11th and 12th centuries the Ashkenazi rabbinic genres formed.

BEST MODEL

- ~ 3000 BCE ancestors of Jewish populations diverged from other Middle Eastern populations
  - Experienced extreme population size reduction
- 13th century ancestors of Ashkenazi Jews diverged from other Jewish populations
  - Experienced another population size reduction
  - Experienced gene flow from Europeans

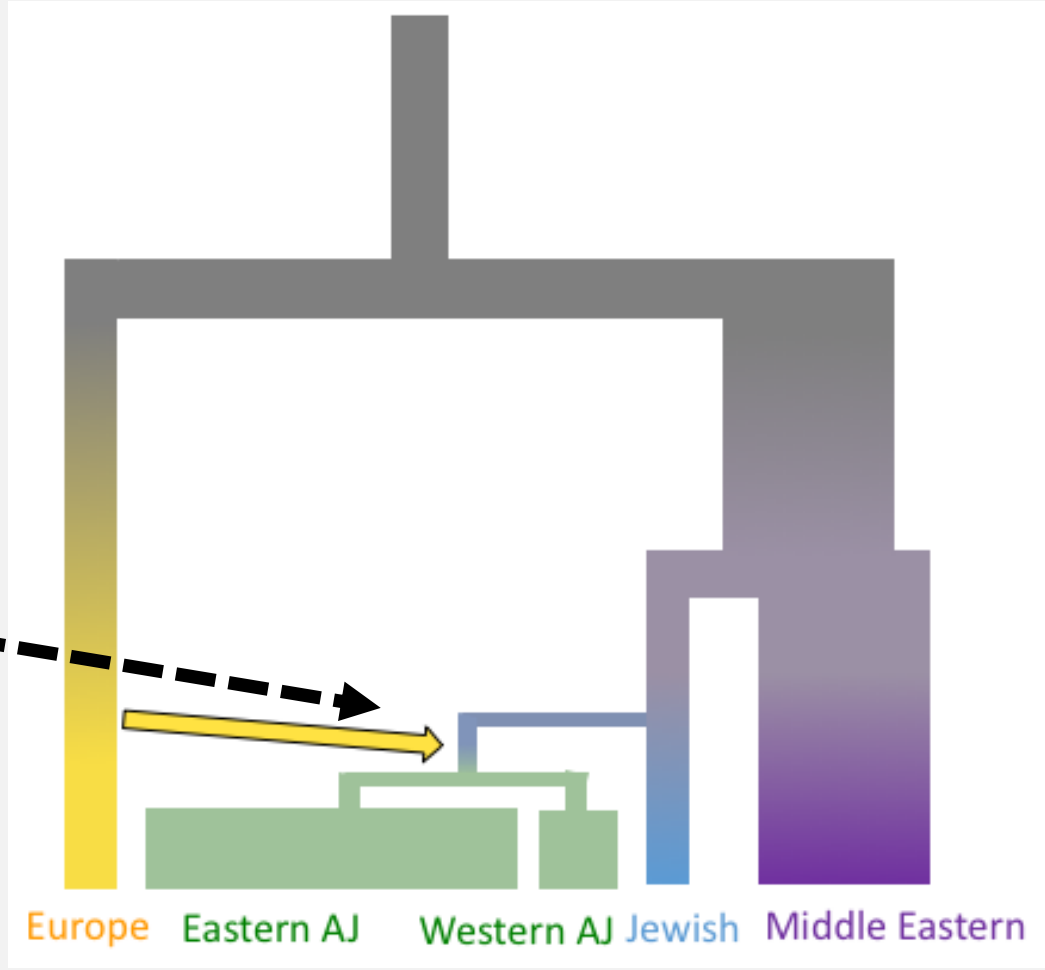
    (unresolved how much or when)



BEST MODEL

- ~ 3000 BCE ancestors of Jewish populations diverged from other Middle Eastern populations
  - Experienced extreme population size reduction
- 13th century ancestors of Ashkenazi Jews diverged from other Jewish populations
  - Experienced another population size reduction
  - Experienced gene flow from Europeans

  (unresolved how much or when)

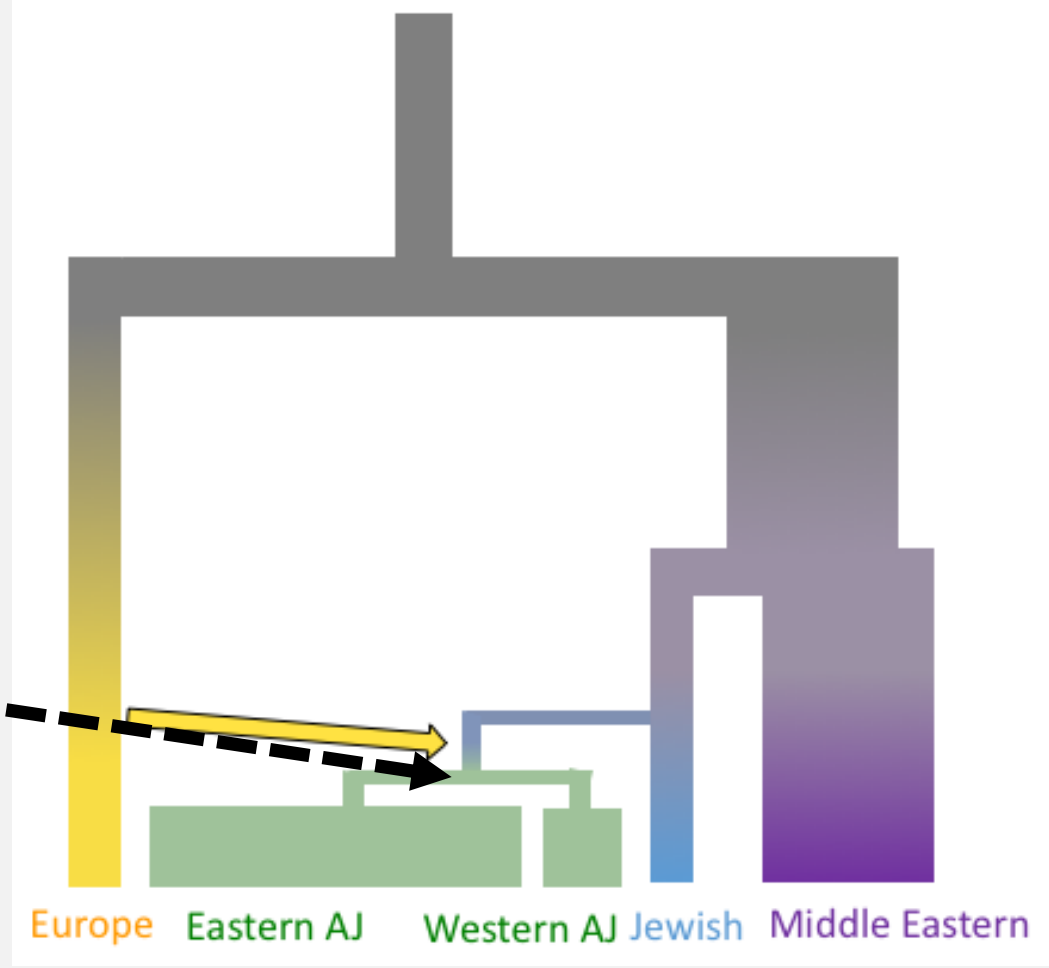


- Judaism follows matrilineal descent.
- In Central Europe Jews became increasingly integrated into gentile life.
- In Eastern Europe Jews became increasingly isolated.

BEST MODEL

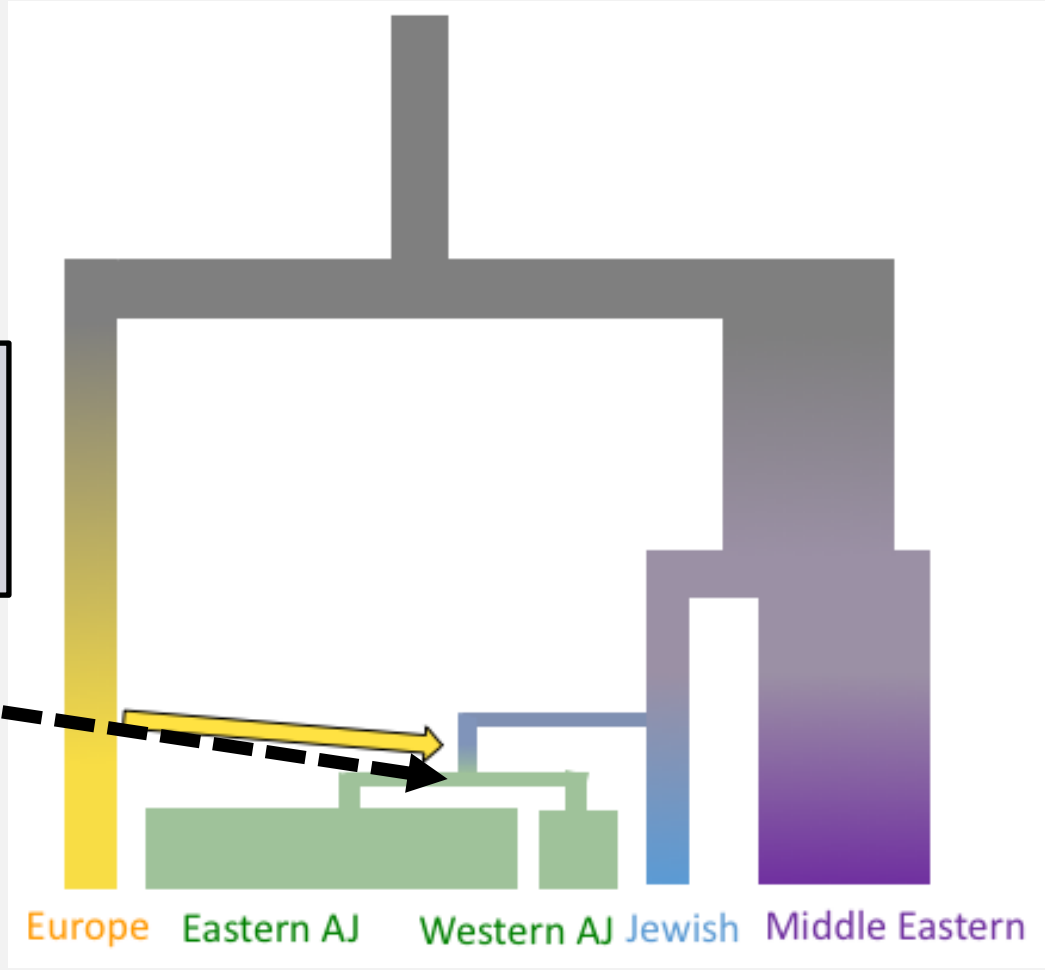
- ~ 3000 BCE ancestors of Jewish populations diverged from other Middle Eastern populations
  - Experienced extreme population size reduction
- 13th century ancestors of Ashkenazi Jews diverged from other Jewish populations
  - Experienced another population size reduction
  - Experienced gene flow from Europeans

    (unresolved how much or when)
- 16th century Eastern and Western Ashkenazi Jews diverged



BEST MODEL

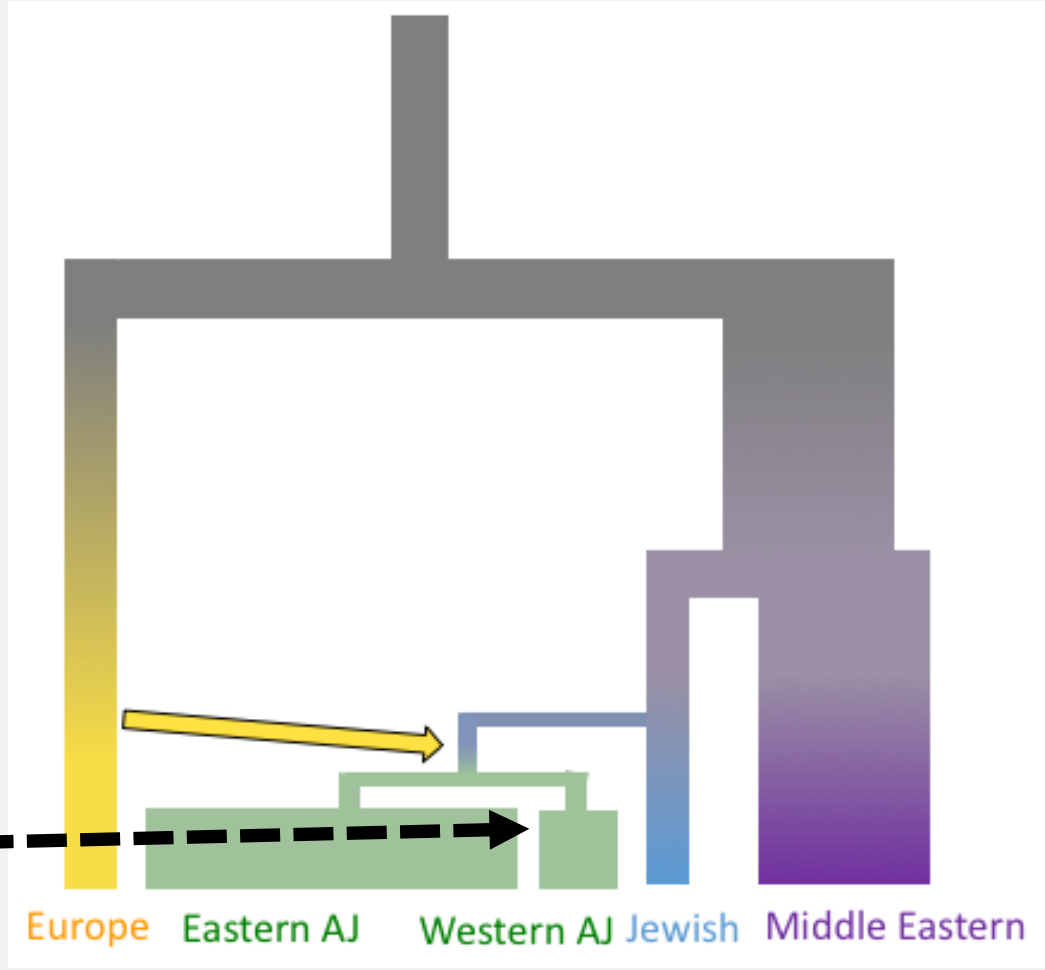
- ~ 3000 BCE ancestors of Jewish populations diverged from other Middle Eastern populations
  - Experienced extreme population size reduction
- 13th century ancestors of Ashkenazi Jews diverged from other Jewish populations
  - Migrations from Central Europe to Poland in the 14th, 15th, and 16th centuries.
  - By 16th century Polish Jewry culturally distinct.
- 16th century Eastern and Western Ashkenazi Jews diverged



BEST MODEL

- ~ 3000 BCE ancestors of Jewish populations diverged from other Middle Eastern populations
  - Experienced extreme population size reduction
- 13th century ancestors of Ashkenazi Jews diverged from other Jewish populations
  - Experienced another population size reduction
  - Experienced gene flow from Europeans

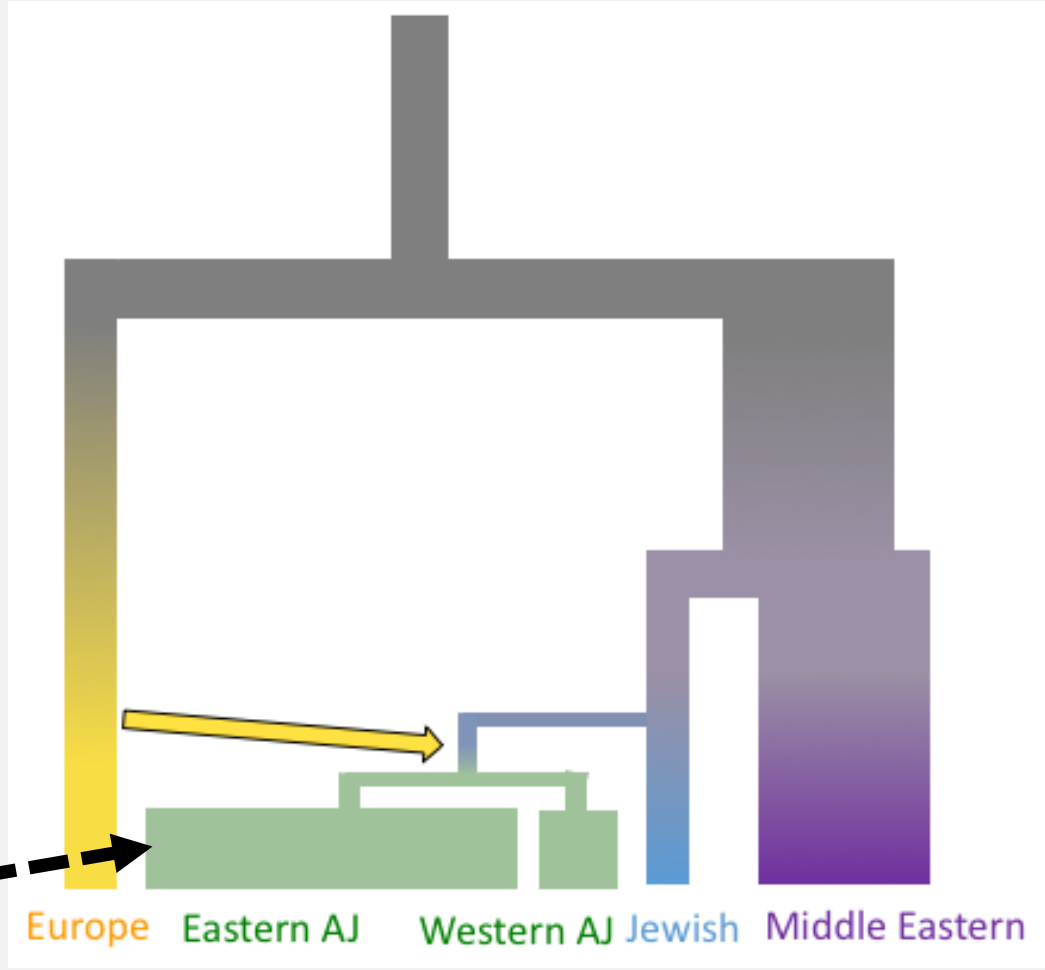
    (unresolved how much or when)
- 16th century Eastern and Western Ashkenazi Jews diverged
  - Western AJ moderately grew in size



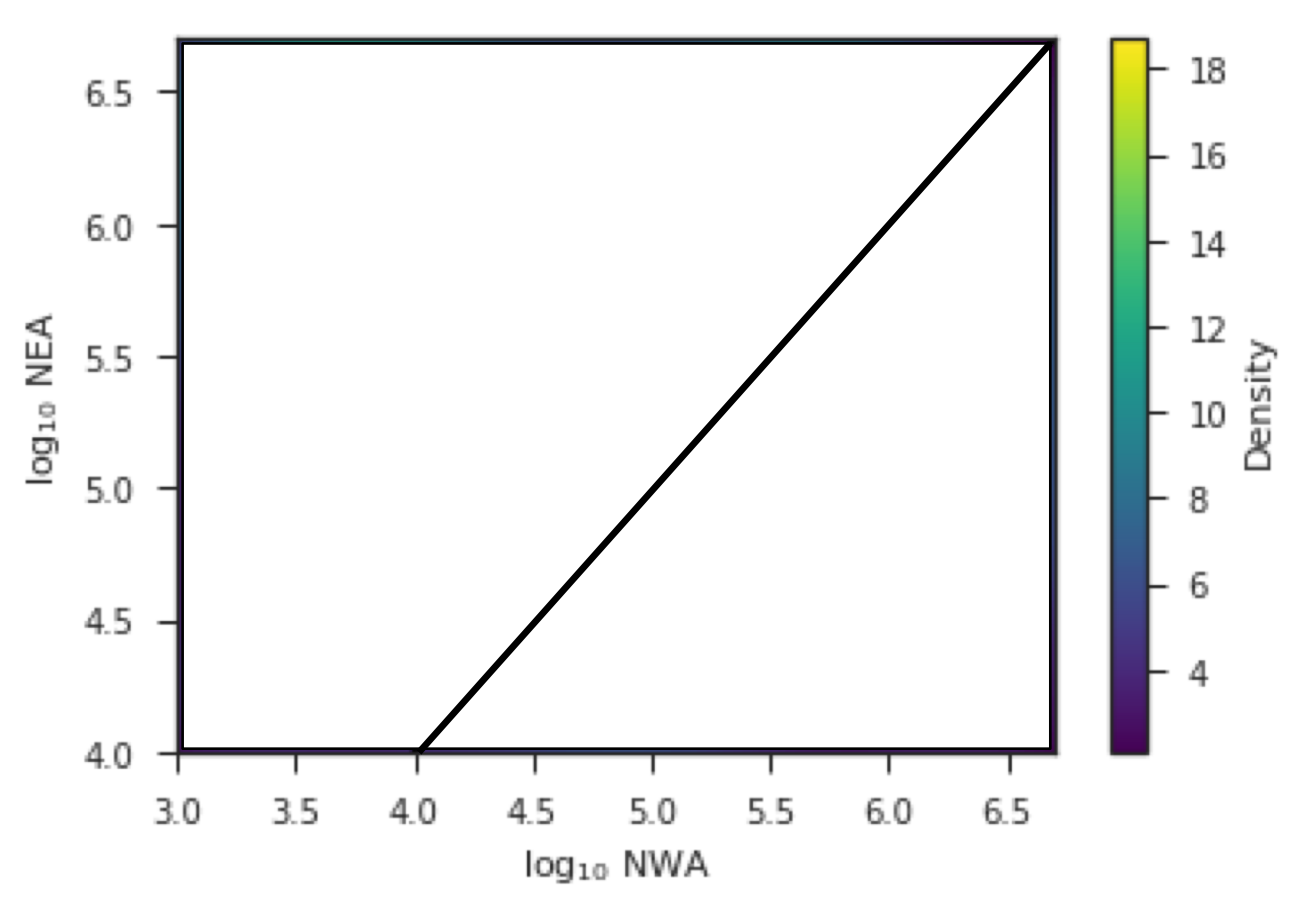
BEST MODEL

- ~ 3000 BCE ancestors of Jewish populations diverged from other Middle Eastern populations
  - Experienced extreme population size reduction
- 13th century ancestors of Ashkenazi Jews diverged from other Jewish populations
  - Experienced another population size reduction
  - Experienced gene flow from Europeans

    (unresolved how much or when)
- 16th century Eastern and Western Ashkenazi Jews diverged
  - Western AJ moderately grew in size
  - Eastern AJ massively grew in size

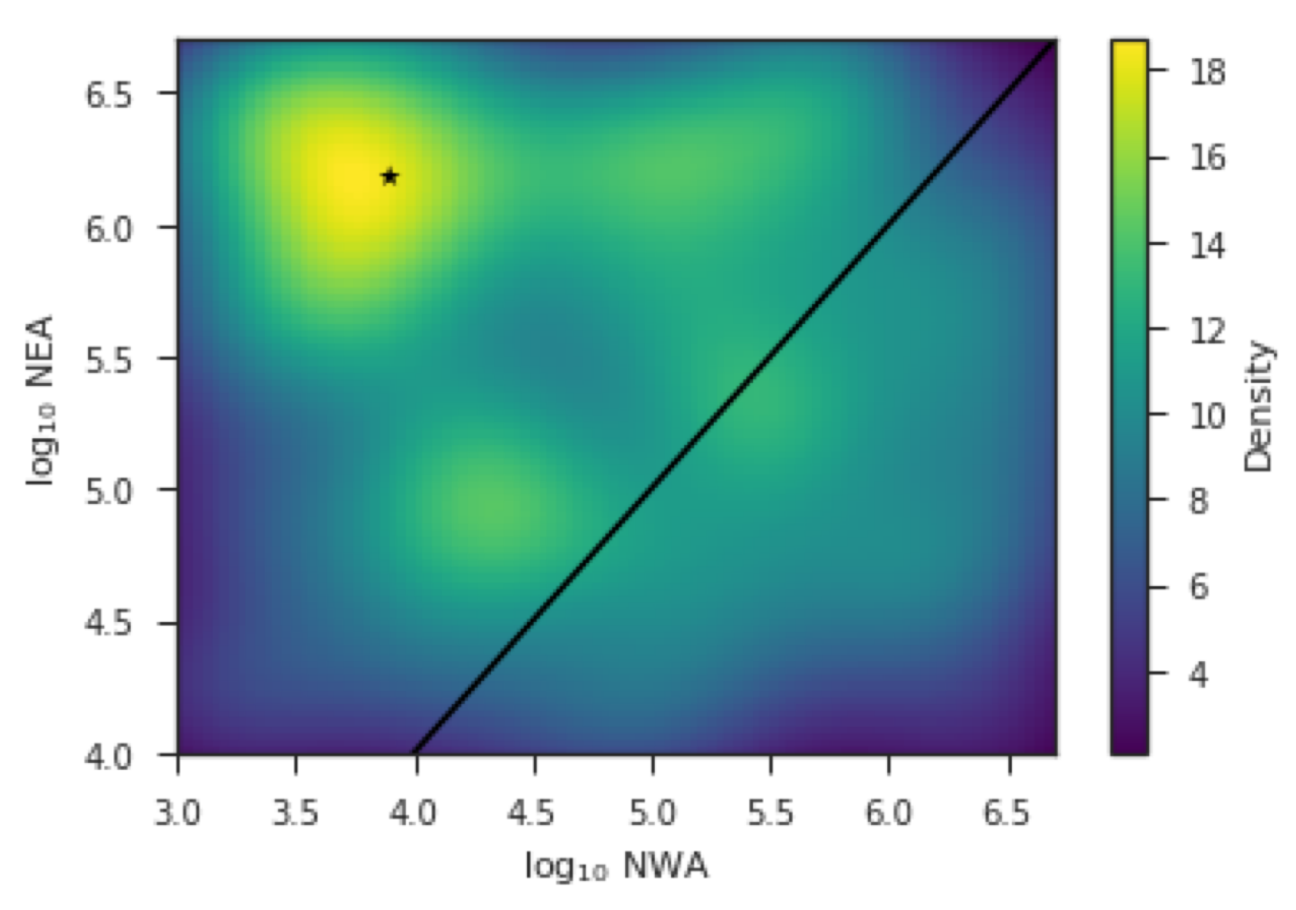


JOINT POSTERIOR OF EFFECTIVE POPULATION SIZE OF EASTERN AND WESTERN AJ



JOINT POSTERIOR OF EFFECTIVE POPULATION SIZE OF EASTERN AND WESTERN AJ

	Probability NEA > NWA
Chr1	0.69
Genome	0.62
Updated chr5	0.70



MORE GROWTH IN EASTERN AJ

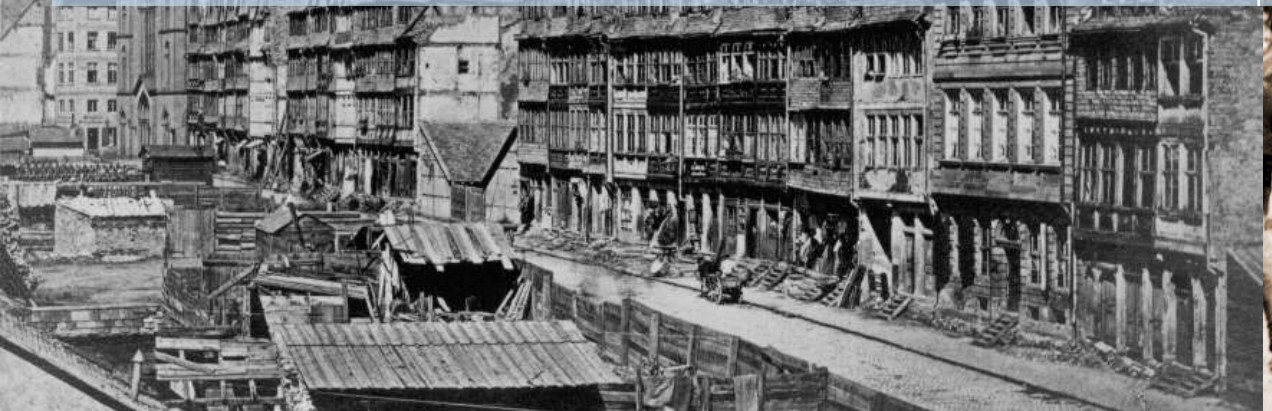
Central Europe	Eastern Europe
- Often expelled from settlements. - Strict regulations on where Jews could live and what they could do to earn a living.	- Could generally move freely. - Protected by nobles.

MORE GROWTH IN EASTERN AJ

Central Europe	Eastern Europe
- Often expelled from settlements. - Strict regulations on where Jews could live and what they could do to earn a living. - Legal limitations on the number of Jewish families. - Cramped ghettos in the 19th century .	- Could generally move freely. - Protected by nobles. - No limitations from government on number of Jewish marriages

MORE GROWTH IN EASTERN AJ

Central Europe	Eastern Europe
- Often expelled from settlements. - Strict regulations on where Jews could live and what they could do to earn a living. - Legal limitations on the number of Jewish families. - Cramped ghettos in the 19th century - Integration into non-Jewish society.	- Could generally move freely. - Protected by nobles. - No limitations from government on number of Jewish marriages - Adherence to religious and traditional norms and economic structures encouraged early marriage and high fertility.



IMPORTANCE OF WORK

Historical / Cultural	Evolution / Population genetics	Medical
Resolved controversial question of Jewish population growth in Eastern Europe.	Demonstration of inference of very recent history.	How do different growth rates in Western and Eastern AJ affect the frequency of deleterious mutations?

FUTURE DIRECTIONS

- Infer demographic history in other populations with histories of population size changes or inbreeding and admixture
- Approximate Bayesian Computation
  - Using other statistics to better infer admixture
- Machine learning
  - Without using genomic statistics

"Big Data"

THANK YOU!

HAMMER LAB (AND FORMER)

- Michael Hammer
- Consuelo Quinto-Cortes
- August Woerner
- Fernando Mendez

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

CYVERSE

- Blake Joyce
- Julian Pistorius

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