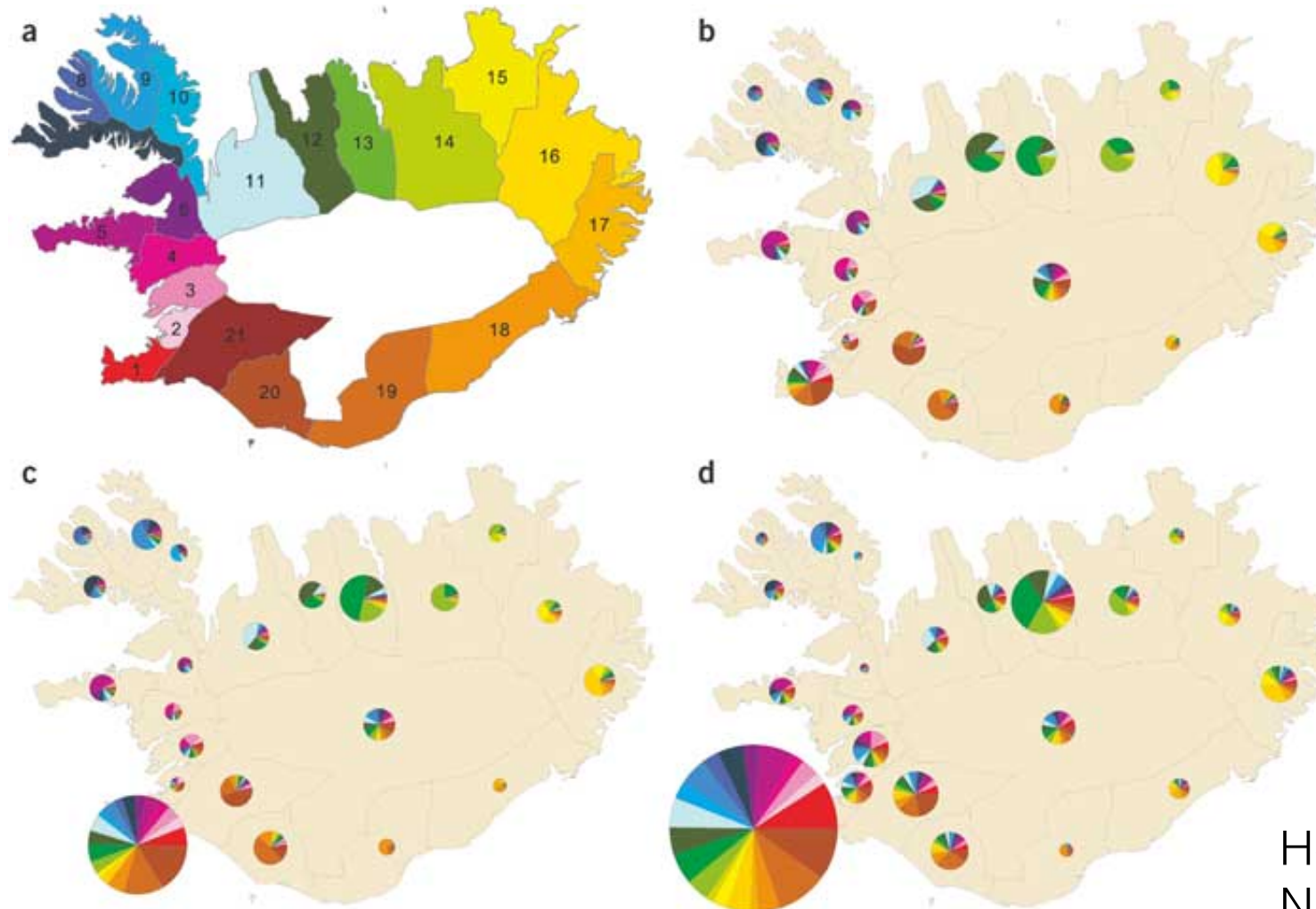


Cryptic Relatedness and fine scale population structure

Learning objectives

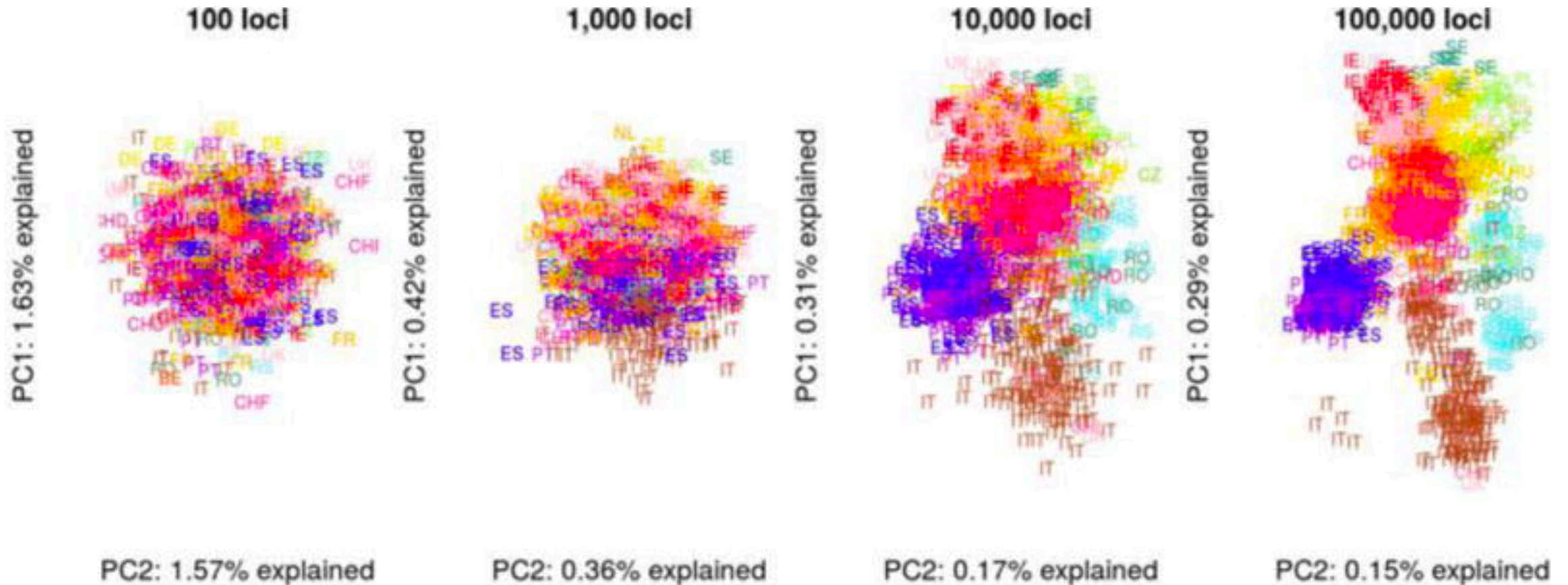
- Define fine scale population structure and cryptic relatedness
- How is it identified
 - Identity-by-descent
 - Rare variation
- Why it can be important for association analyses, especially of rare variants.

Cryptic Population Structure

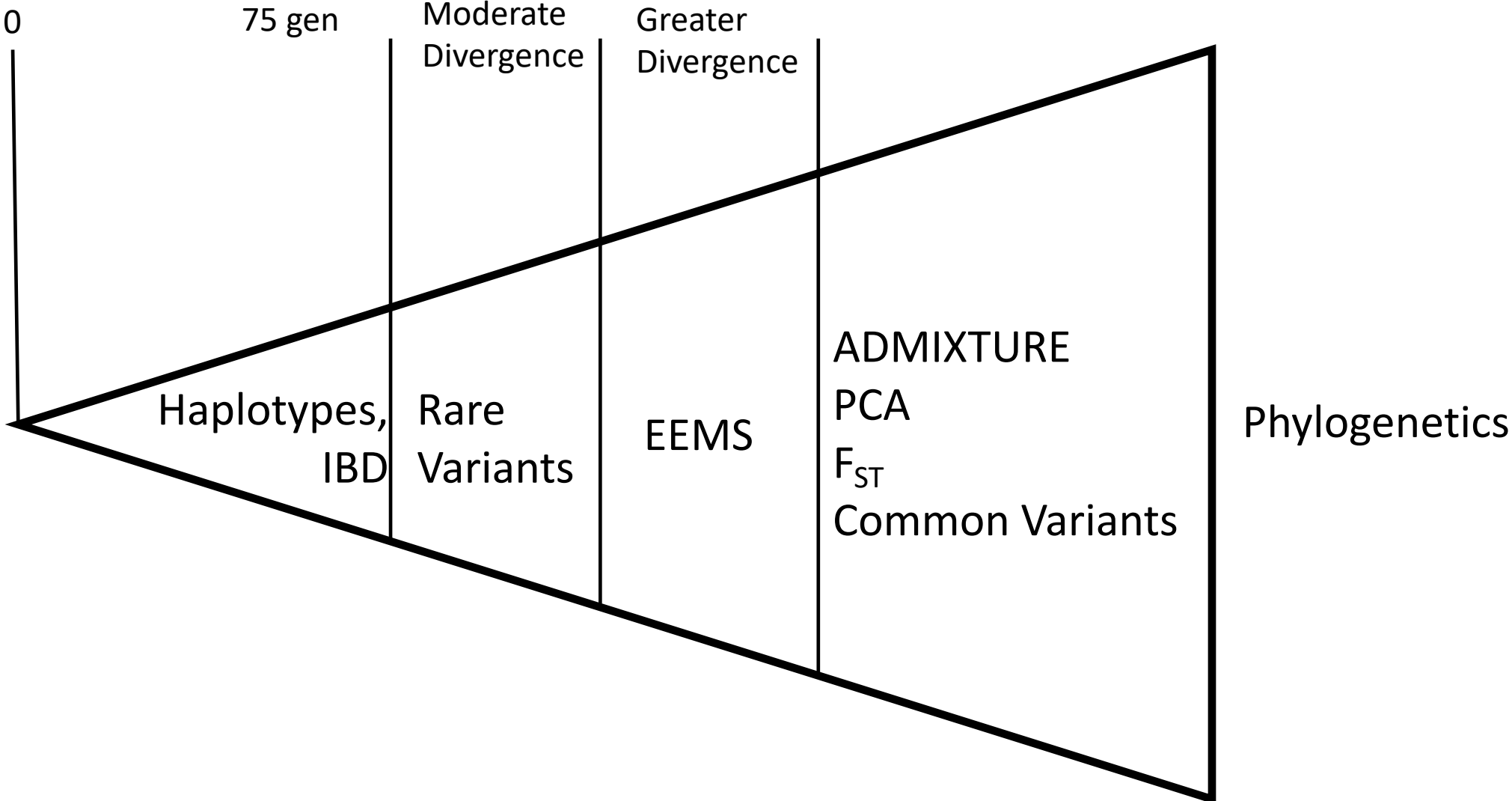


Helgason et al. (2004)
Nature Genet.

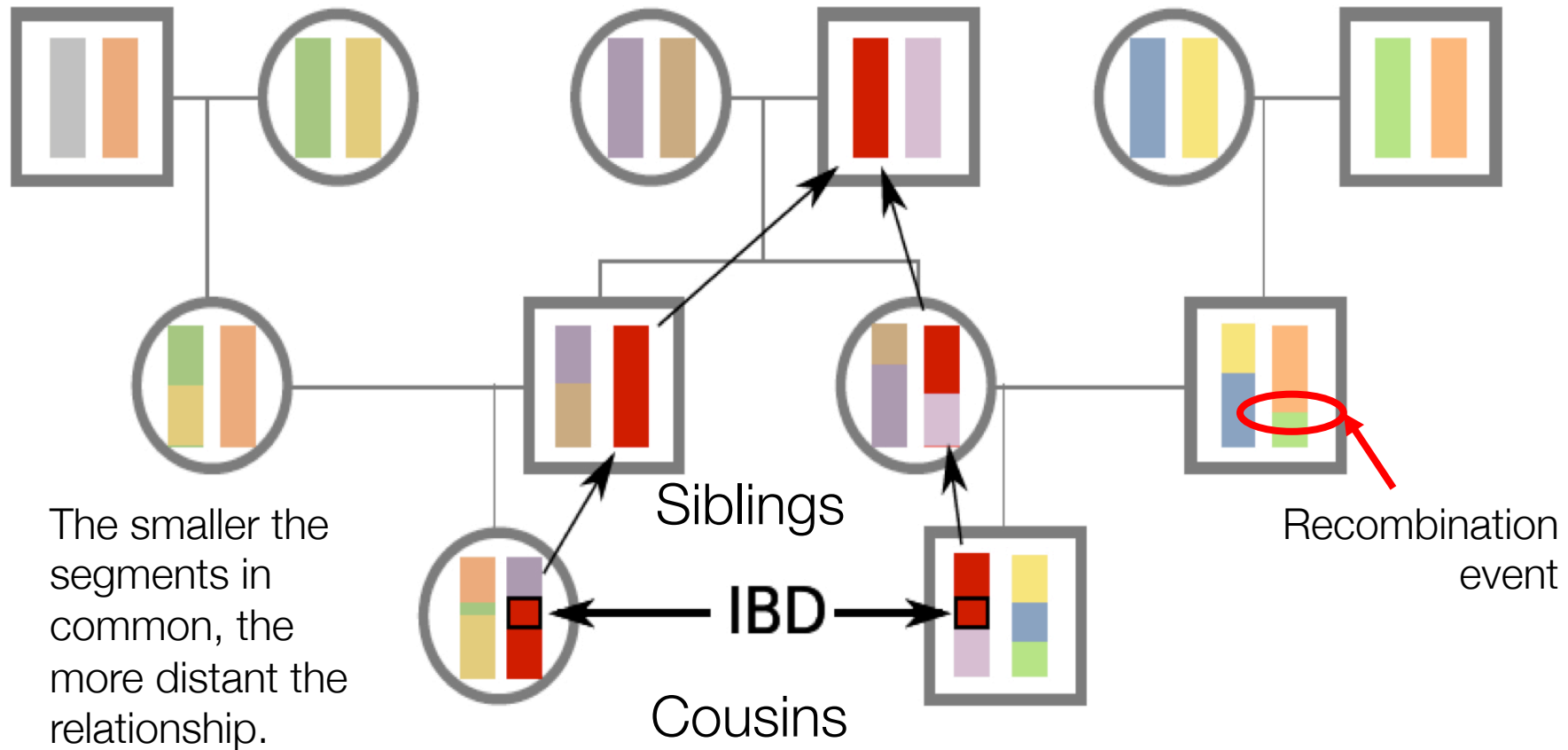
“A large number of loci is required to reveal fine-scale population structure using PCA”



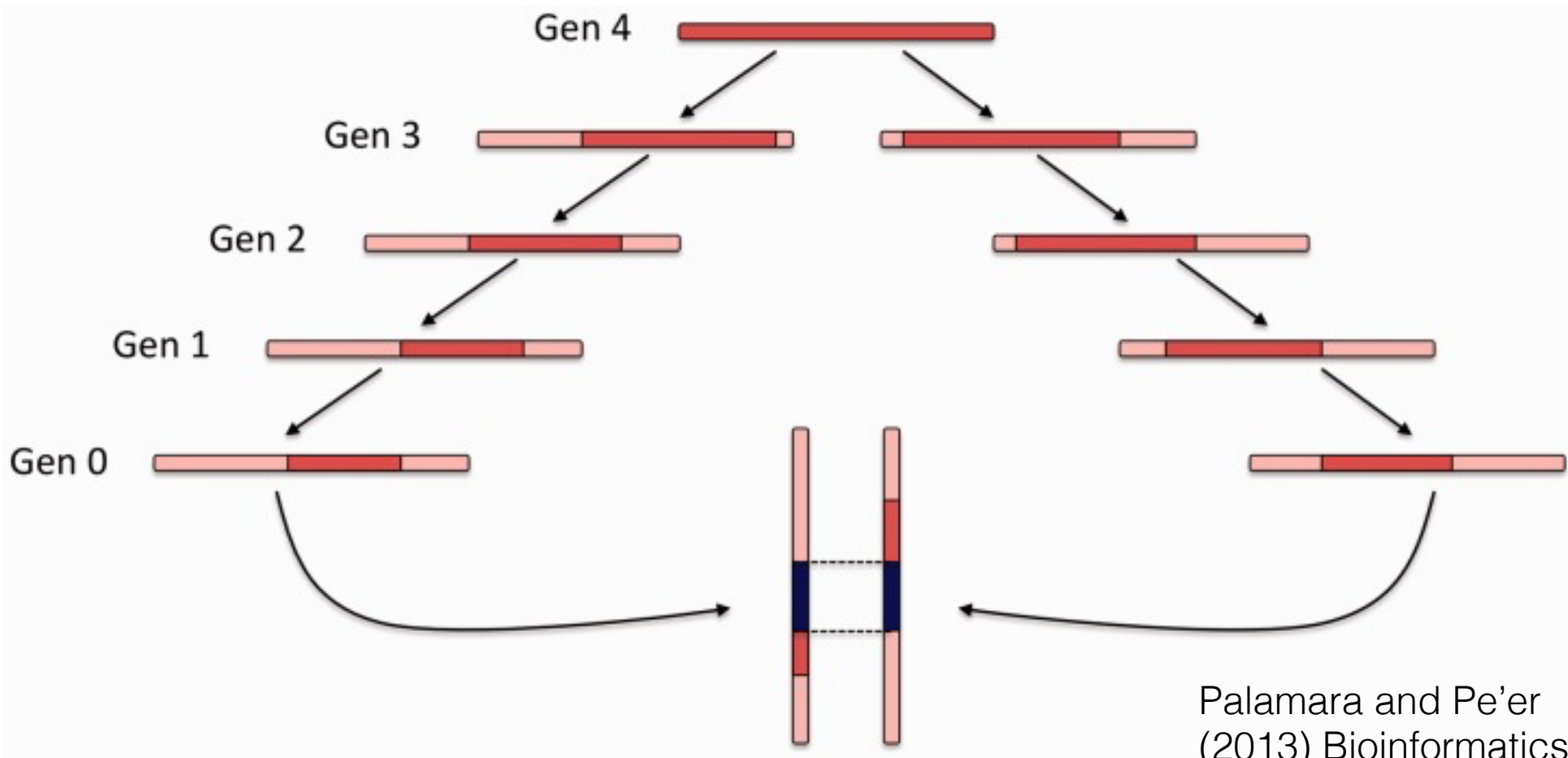
Demographic Inference Time Frames



Identity by Decent (IBD): A method to find both distant and recent relationships



IBD length is correlated with historical relationships.

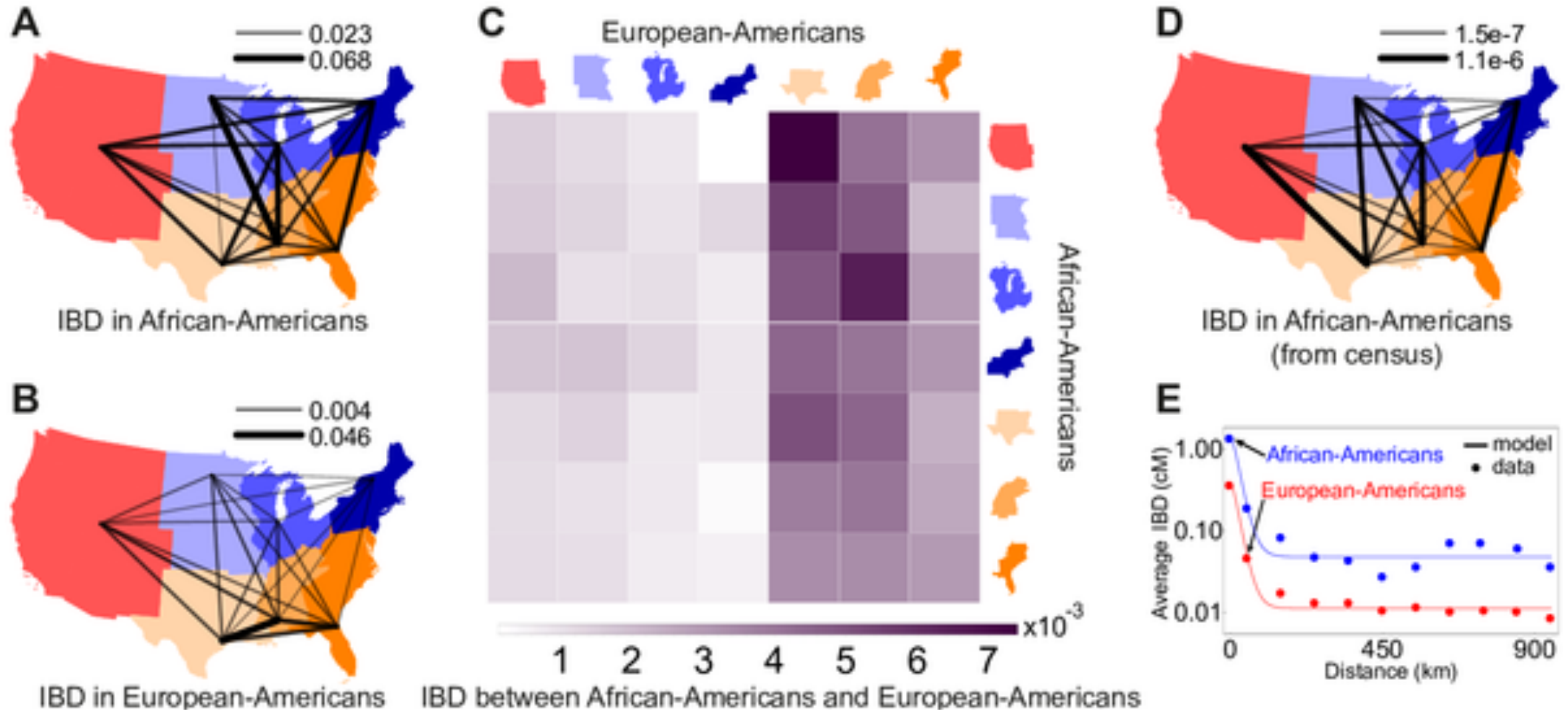


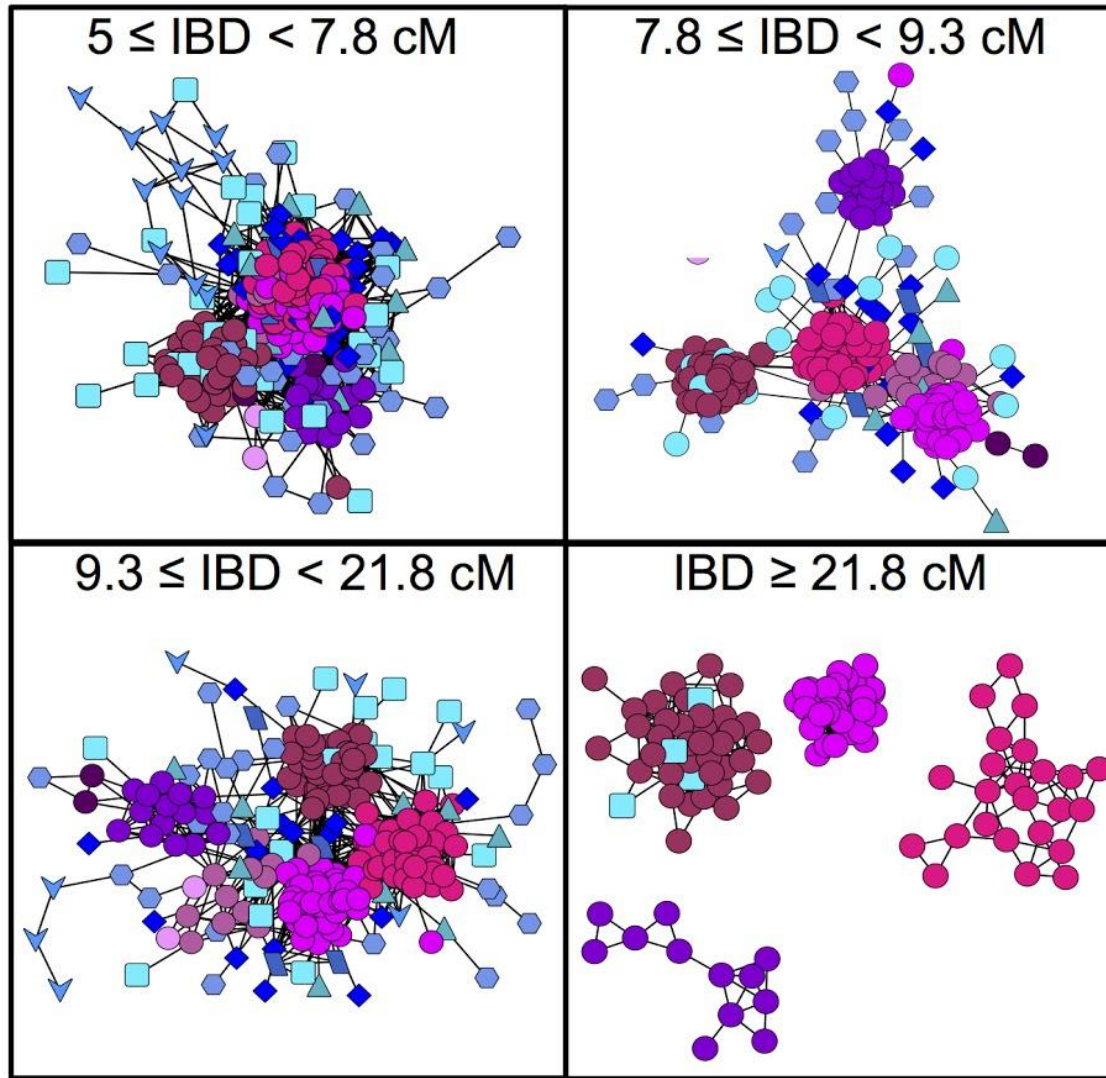
$$E[g|l] \cong \frac{3}{2 * l}$$

Baharian et al. (2016)
PLoS Genet.

Palamara and Pe'er
(2013) Bioinformatics

Pairwise genetic relatedness across

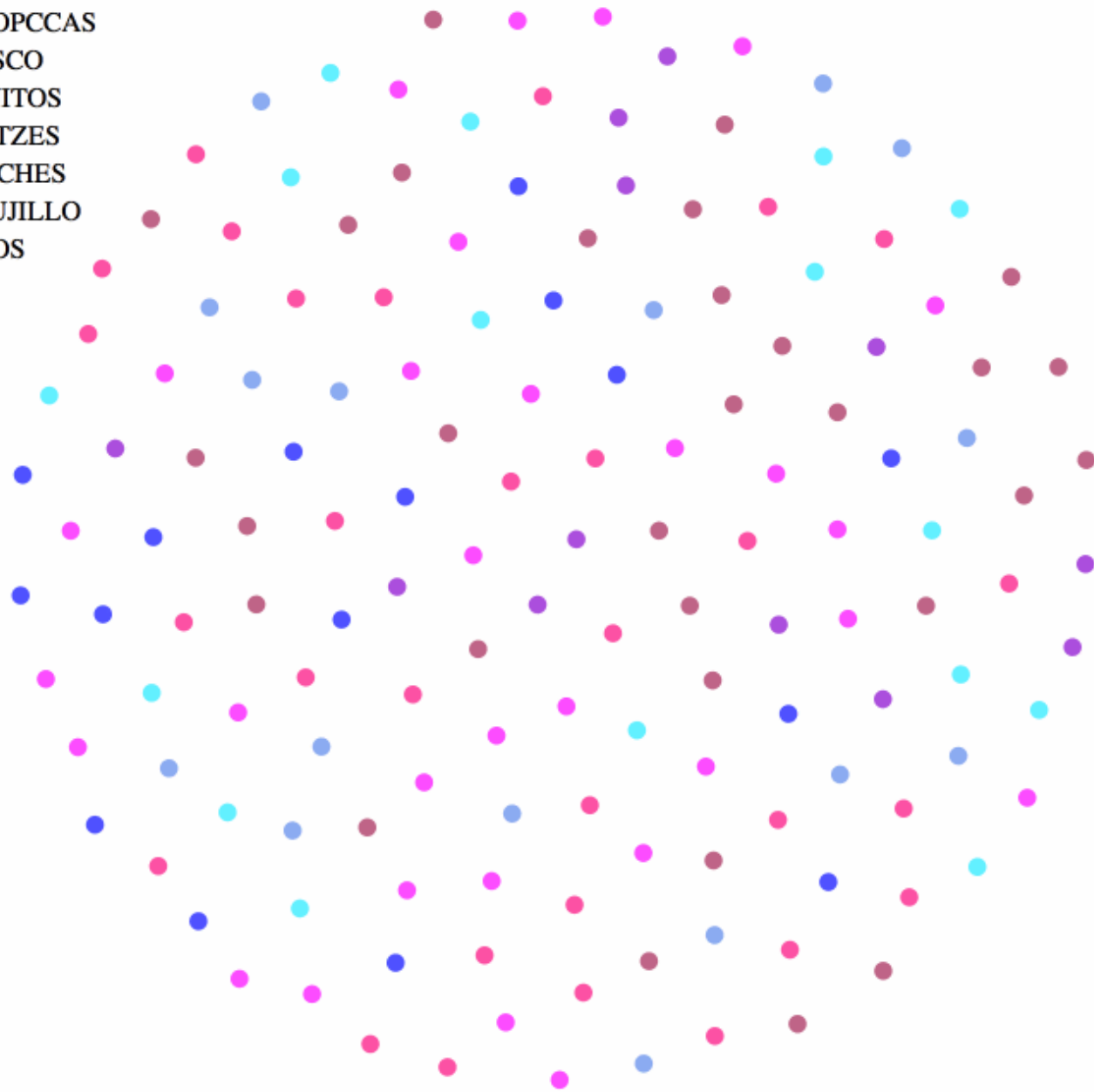


C

Identity-by-descent as a means to look at fine-scale structure over time



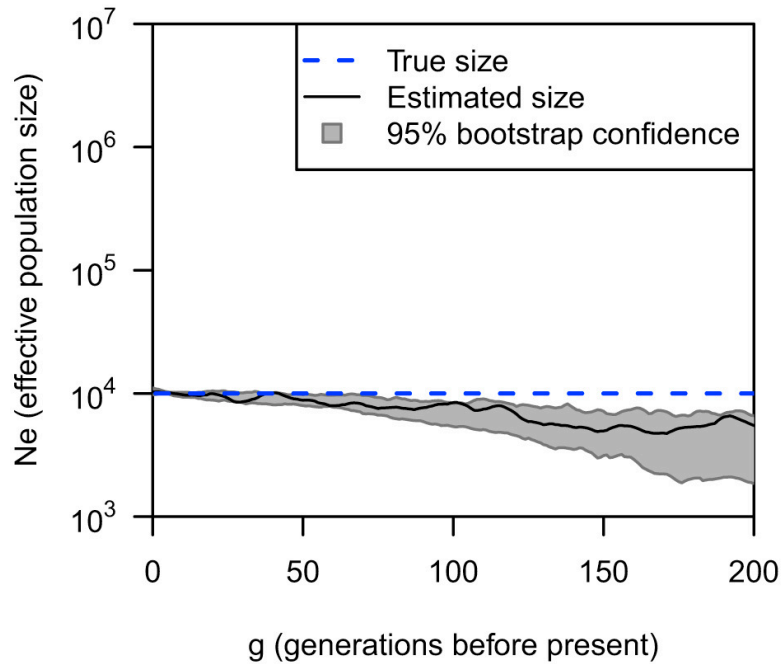
- CHOPCCAS
- CUSCO
- IQUITOS
- MATZES
- MOCHES
- TRUJILLO
- UROS



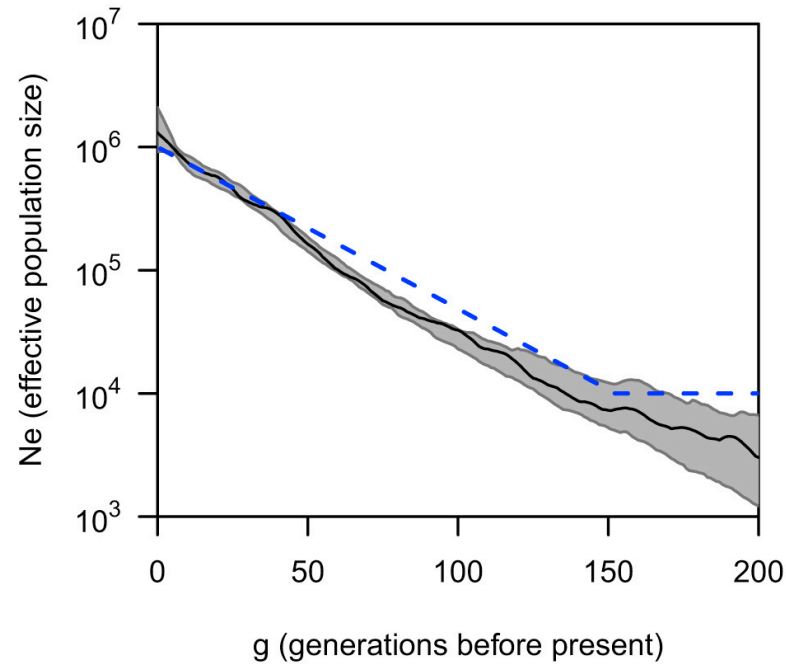
Identity-by-descent as a means to look at fine-scale structure over time

IBD can estimate effective population size over time.

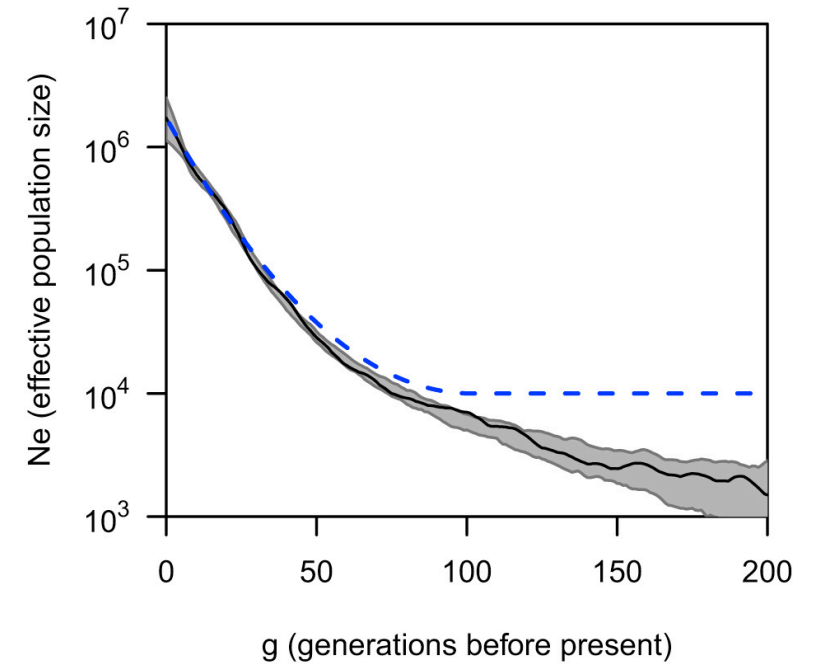
Constant size: SNP array data



Exponential growth: SNP array data

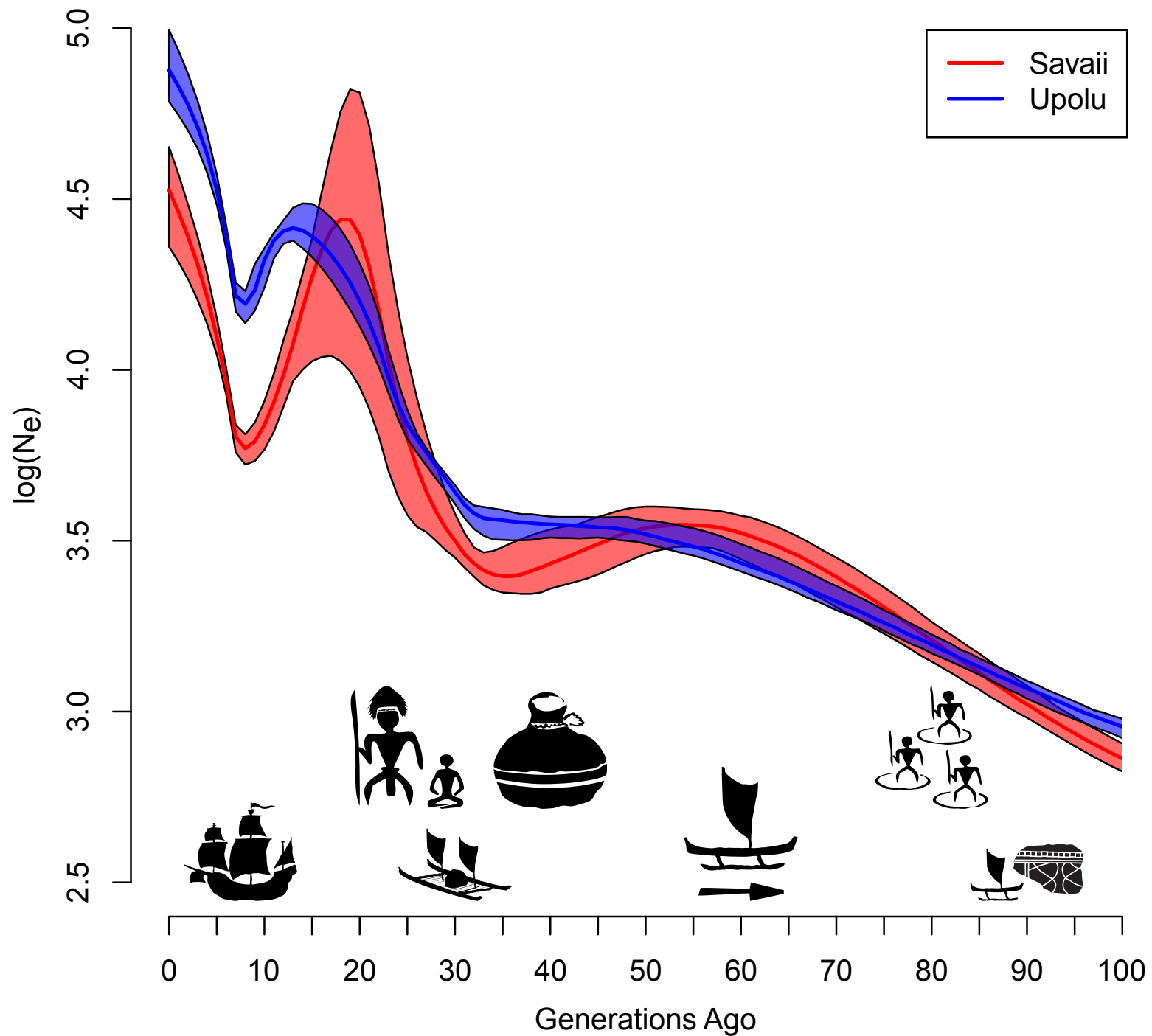


Super-exponential: SNP array data



IBDNe in Samoa!

Harris et al.
(https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3329885)

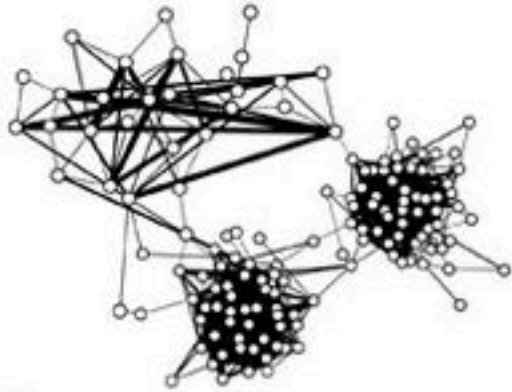


IBD on a large scale

a

Construct network from IBD.

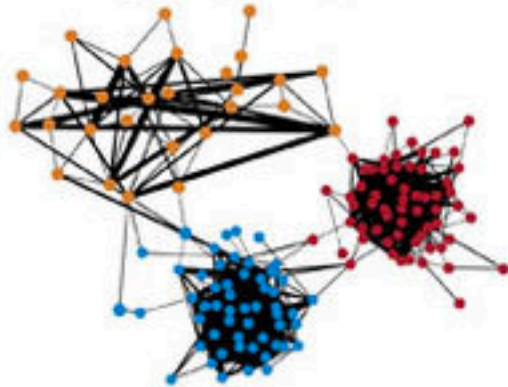
Join vertex pairs (genotyped samples) if $IBD > 12$ cM. Edge weights are a function of total detected IBD.



b

Detect network clusters.

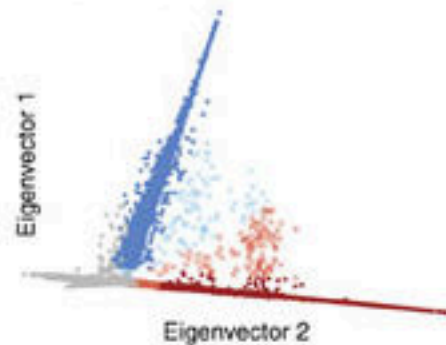
Recursively identify disjoint sets that maximize the modularity of the network. (Here one level of clustering hierarchy is shown.)



c

Identify subsets of the clusters that separate in the spectral embedding.

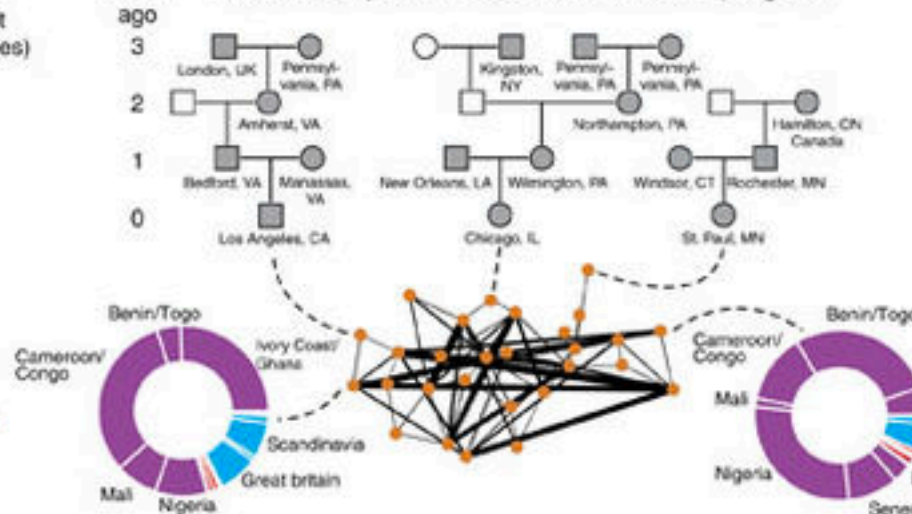
Spectral embedding is computed from eigen-decomposition of Laplacian matrix. In the plot below, we identify "stable subsets" (filled circles) of the blue and red clusters.



d

Annotate each cluster with two kinds of data:

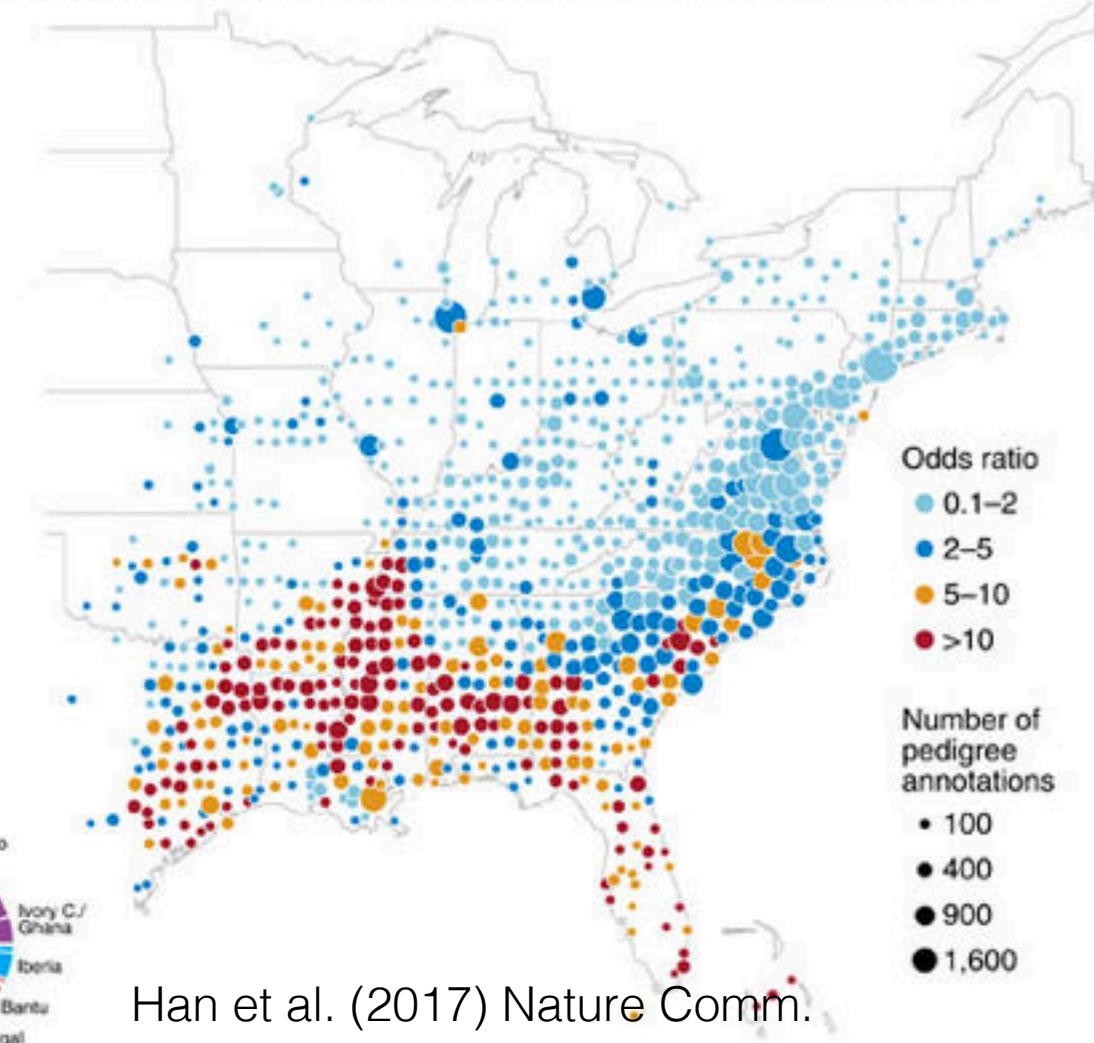
- In all samples, global admixture of 20 populations (donut charts);
- For some samples, birth locations of ancestors in pedigrees.



e

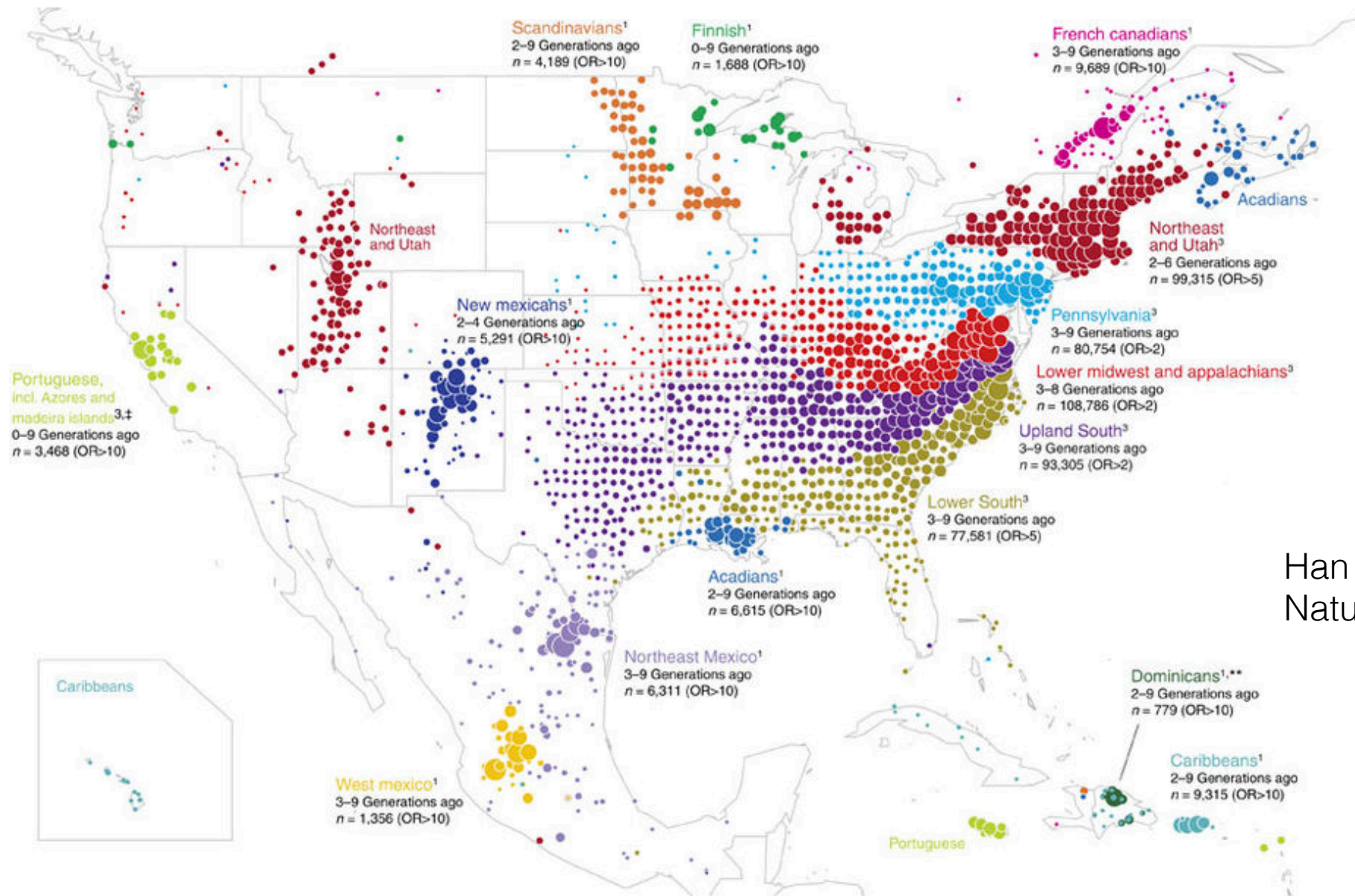
Visualize geographic distribution of ancestral birth locations in each cluster.

Map below shows birth locations of ancestors in the African American cluster. Locations are colored by degree of over-representation (odds ratio), and scaled by number of birth location annotations.



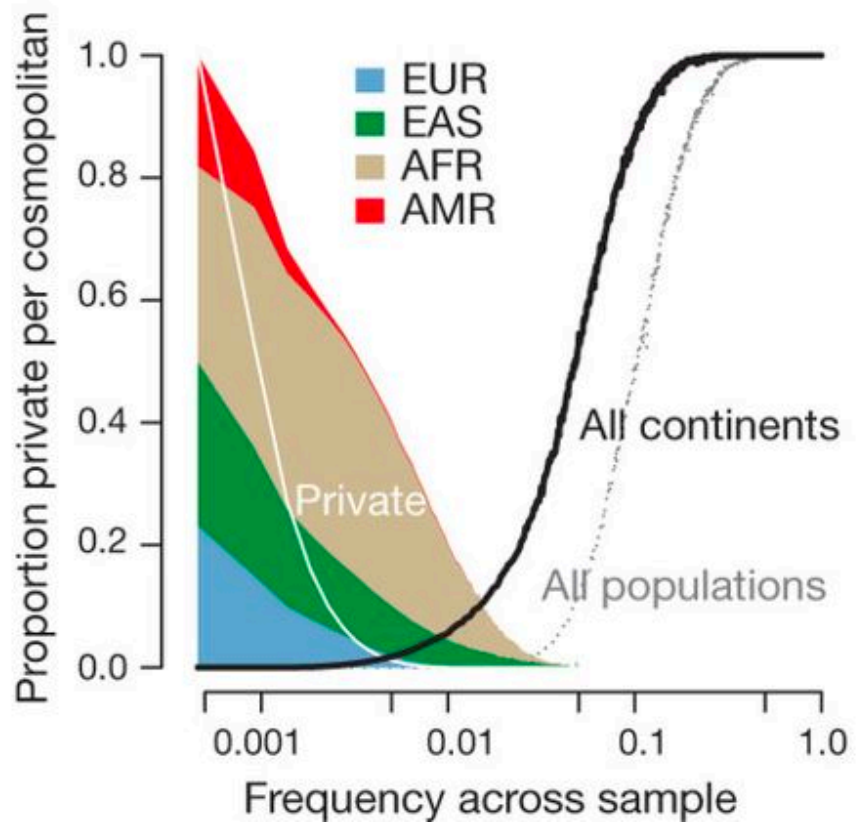
Han et al. (2017) Nature Comm.

IBD on a large scale



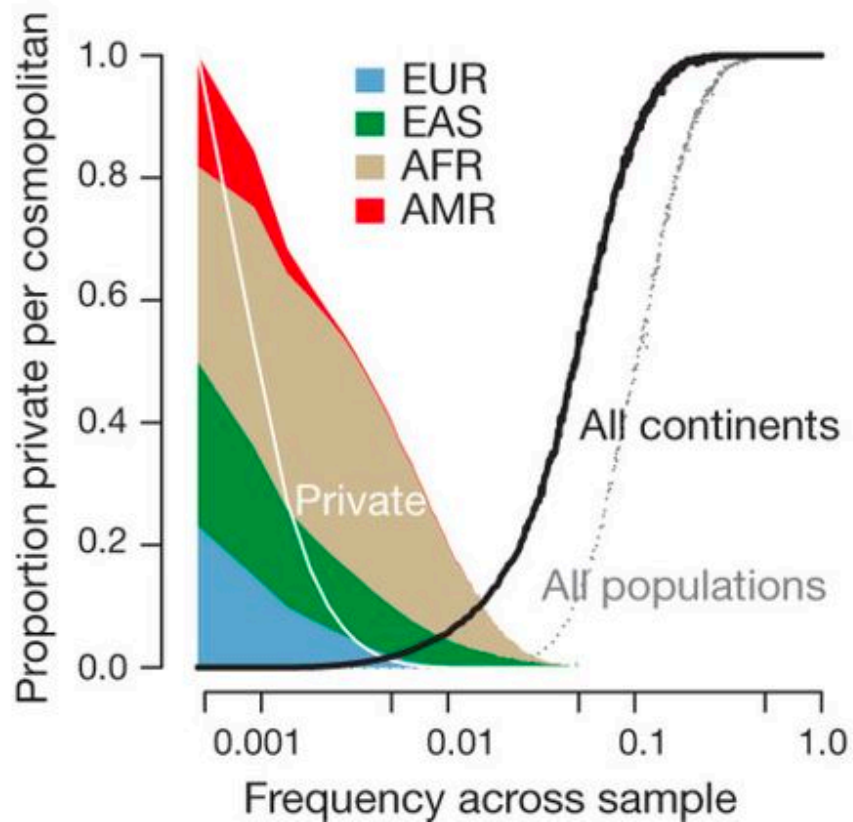
Han et al. (2017)
Nature Comm.

Do rare variants help identify recent population structure?

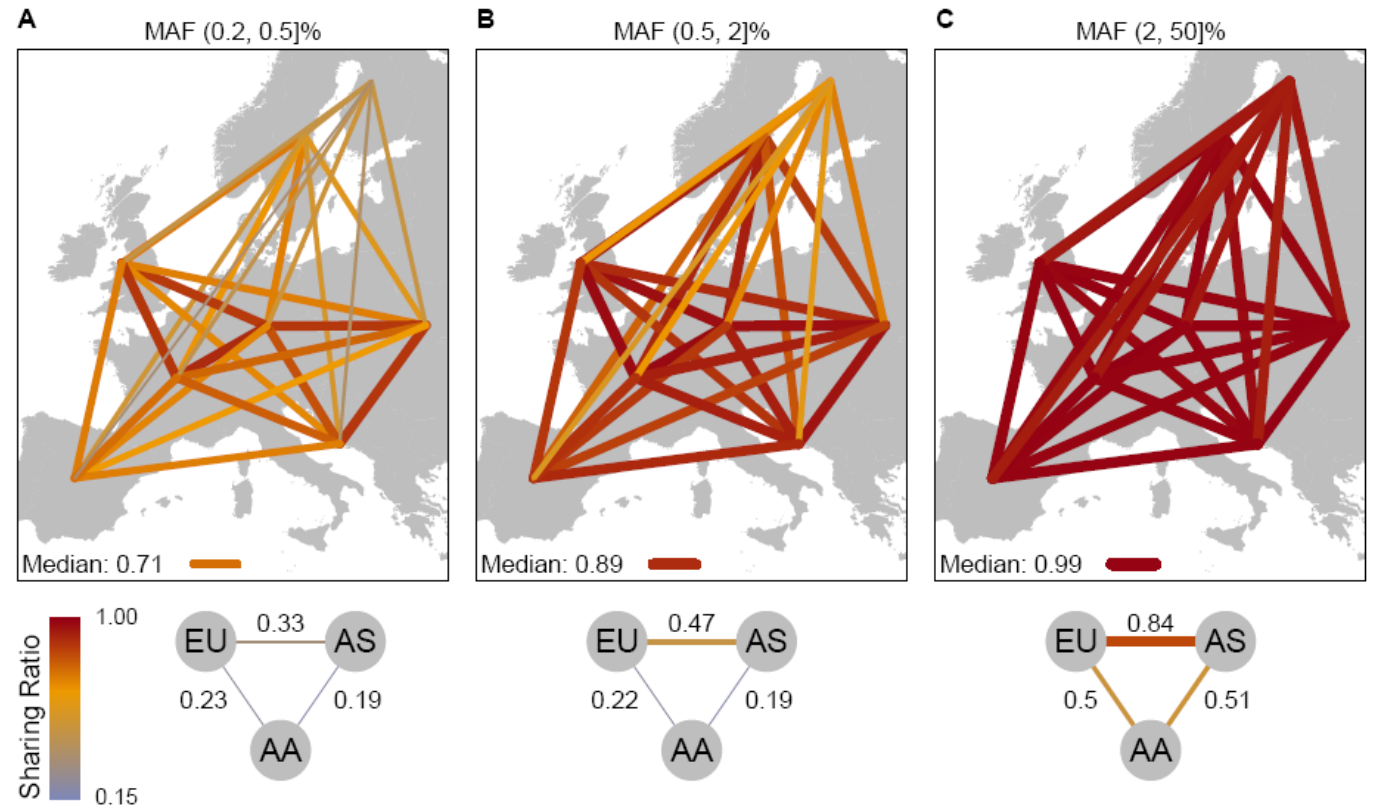


1000 Genomes Project (2012) Nature

Do rare variants help identify recent population structure?

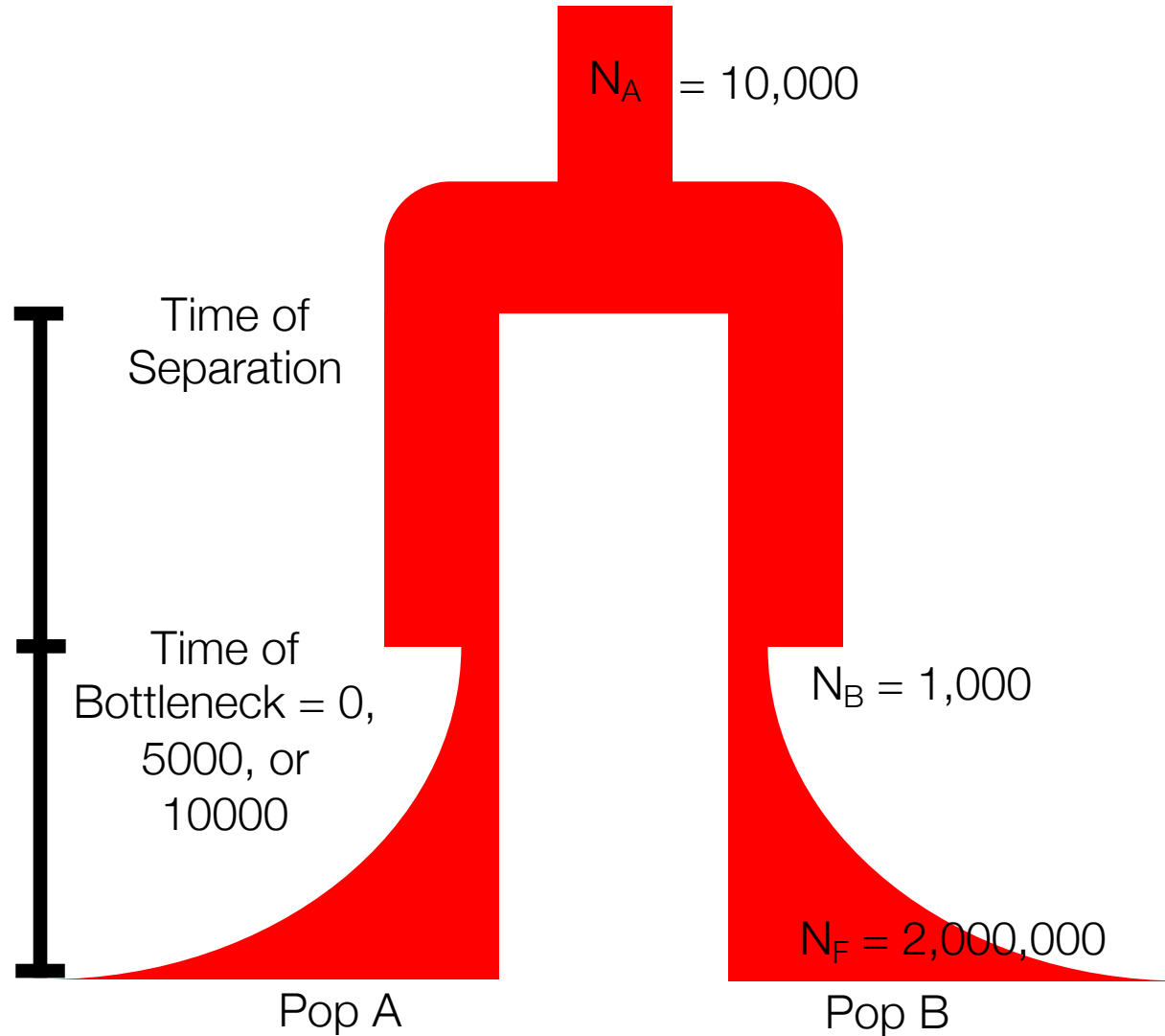


1000 Genomes Project (2012) Nature



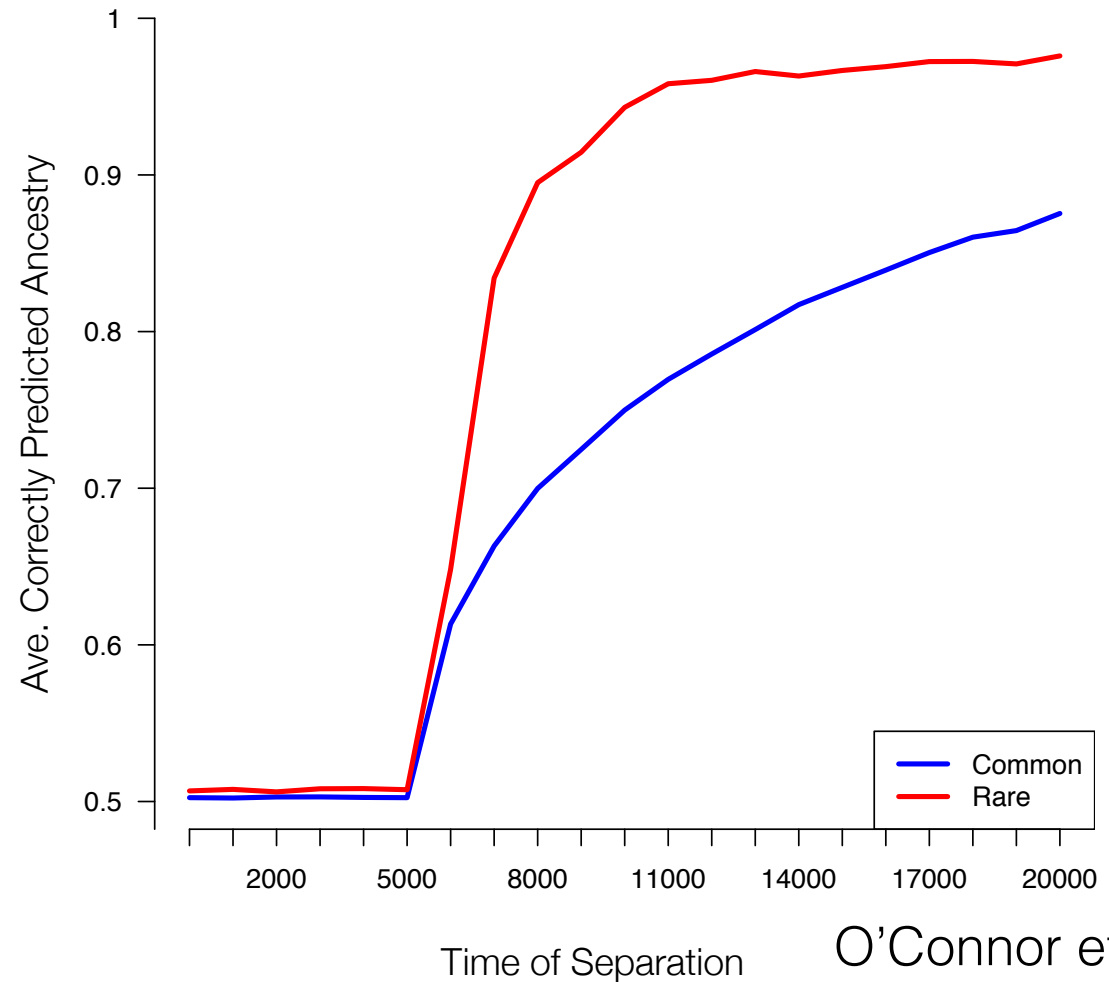
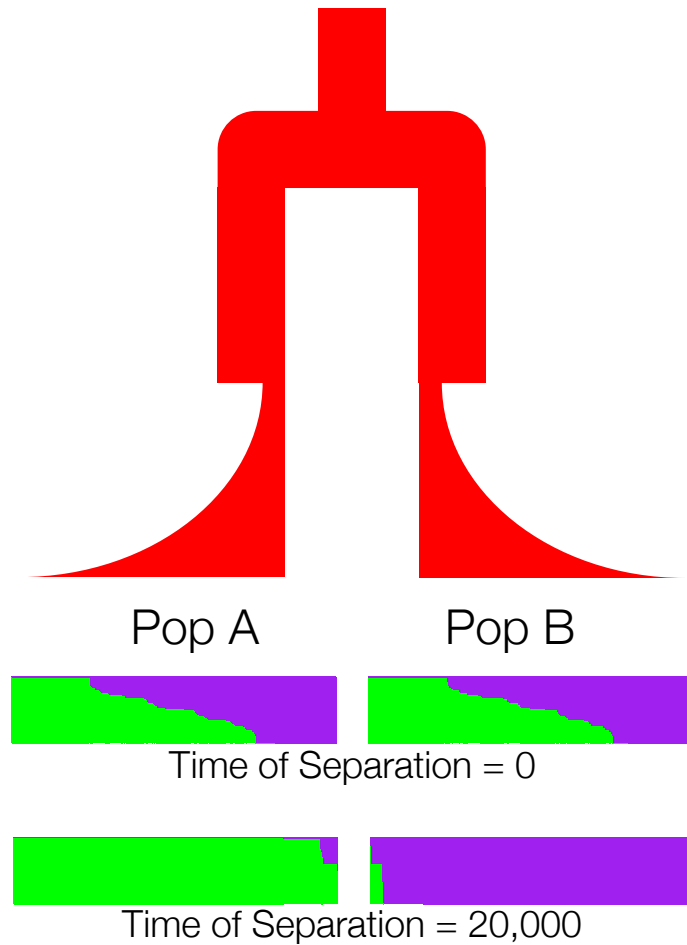
Novembre et al. (2012) Science

Rare VS Common: Population Structure Simulations



O'Connor et al. (2014)
Mol. Biol. Evol.

Rare VS Common: Assignment of Ancestry Proportions



O'Connor et al. (2014)
Mol. Biol. Evol.

Rare VS Common: Which has Greater Information? And When?

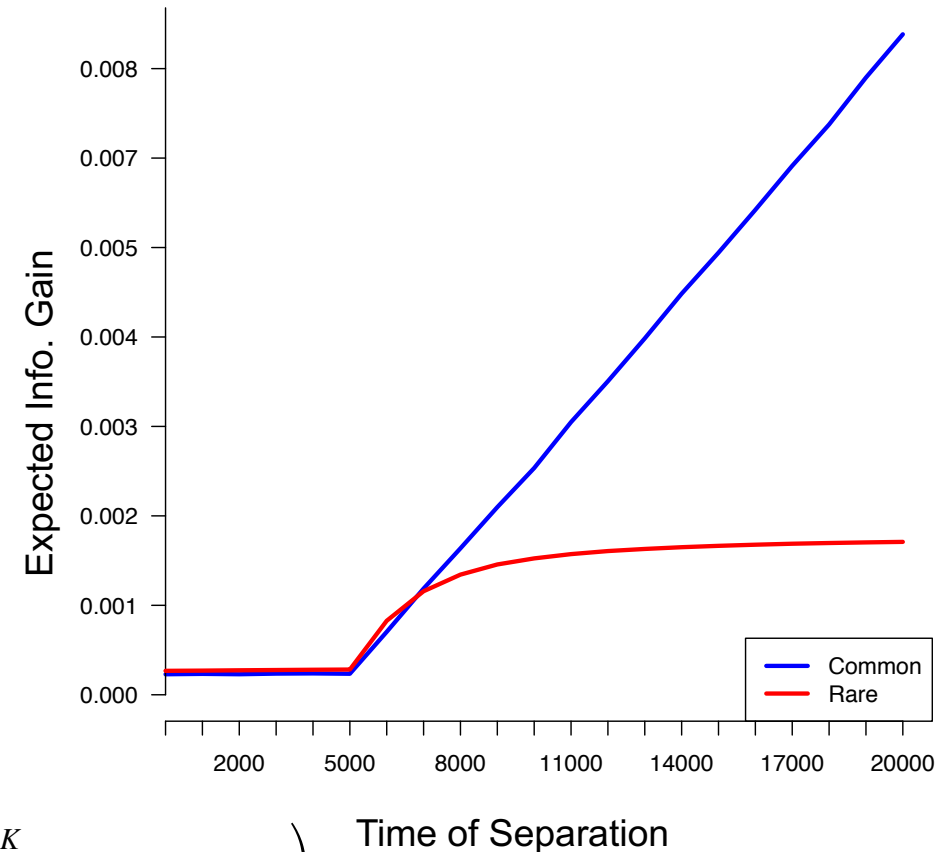
Information Gain: how well a variant can distinguish between populations.
(Rosenberg et al. 2003)

$$I_n(Q; J) = \sum_{j=1}^N \left(-p_j \ln p_j + \sum_{i=1}^K q_i p_{ij} \ln p_{ij} \right)$$

Expected Information Gain

- Calculate for a specific site count
- Correct for missing data
- Weighted average to calculate across a range of frequency (rare or common)

$$E(I_n | C, M) = \sum_{m \in M} \sum_{l=0}^C r_{lm} \times \sum_{j=1}^N \left(-p_{jlm} \ln p_{jlm} + \sum_{i=1}^K q_i p_{ijlm} \ln p_{ijlm} \right)$$

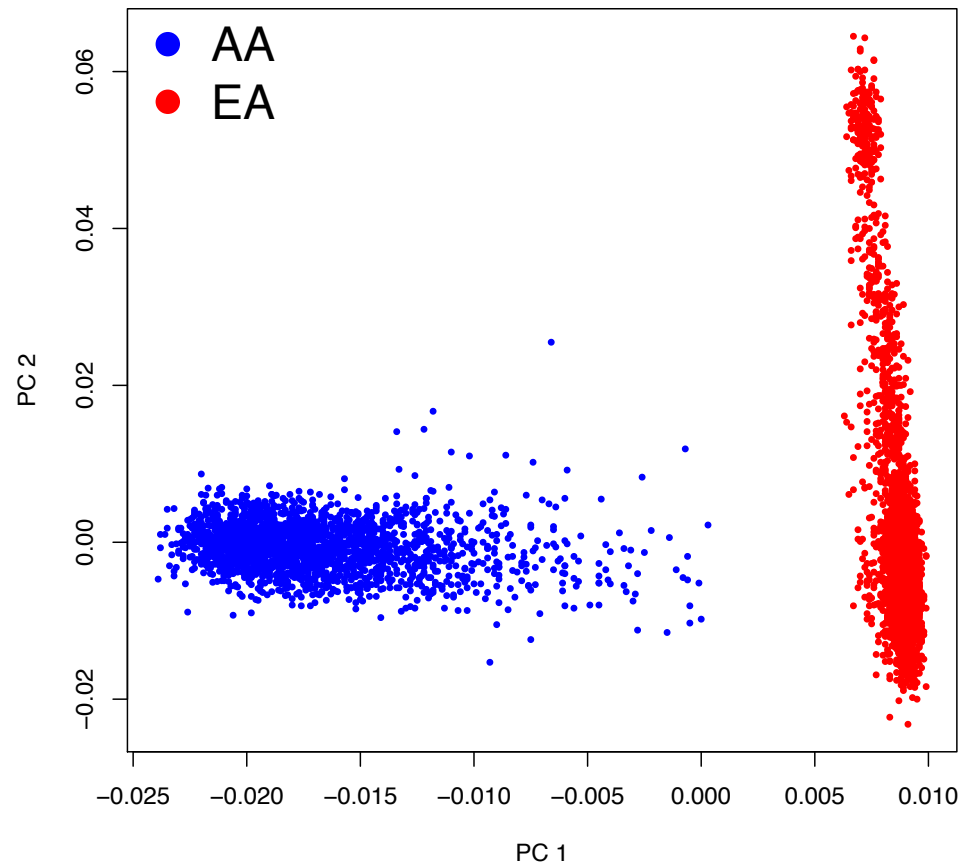


Time of Separation

O'Connor et al. (2014)
Mol. Biol. Evol.

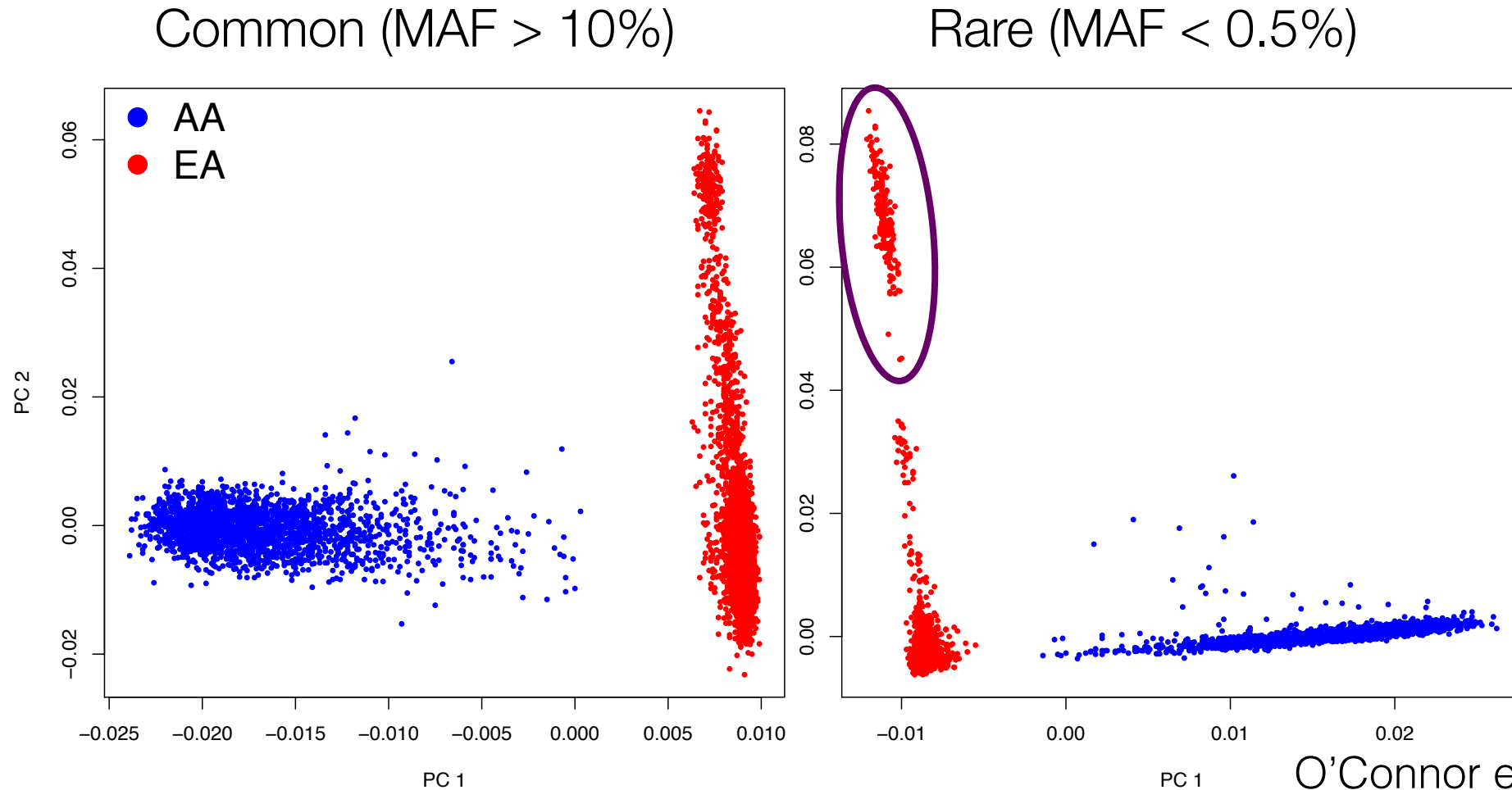
Rare Variants Identify Cryptic Populations

Common (MAF > 10%)



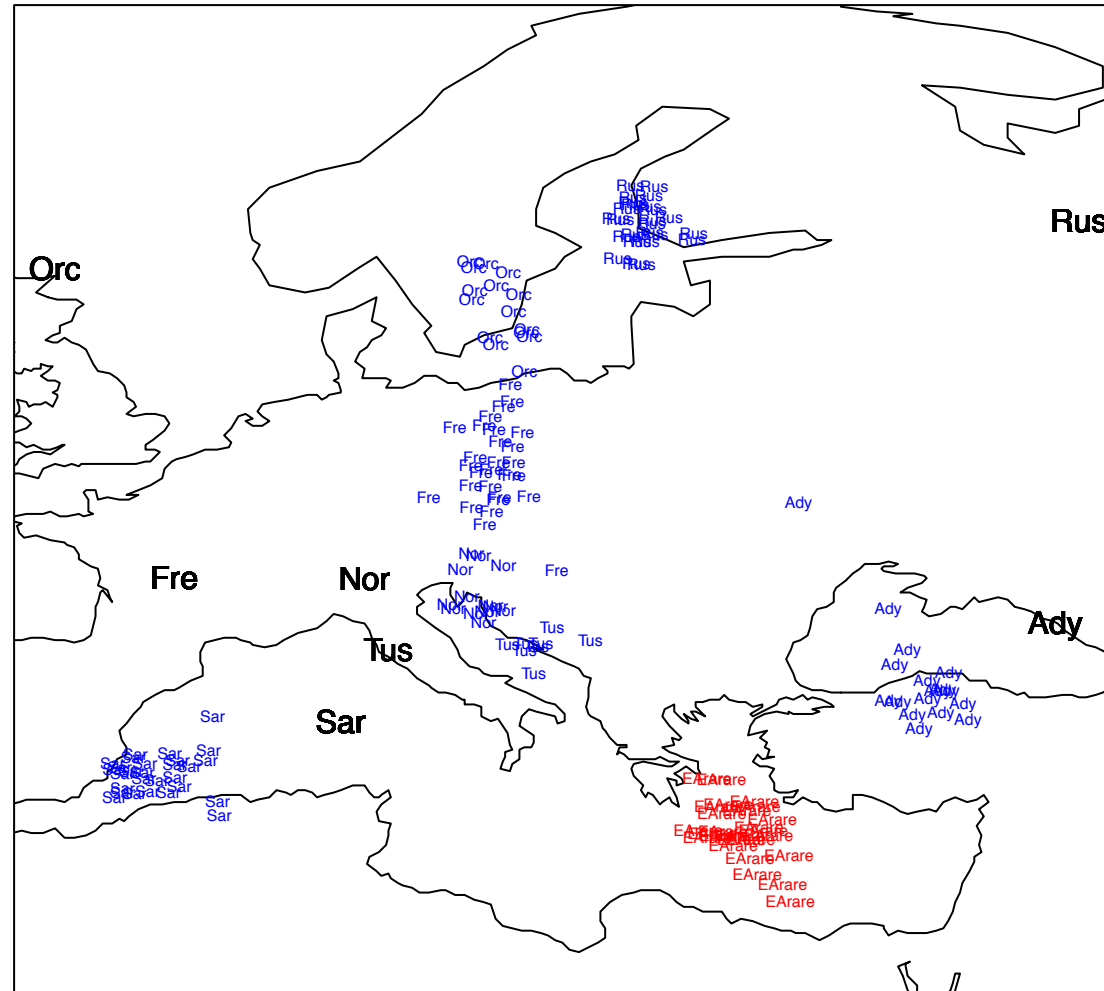
O'Connor et al. (2014)
Mol. Biol. Evol.

Rare Variants Identify Cryptic Populations



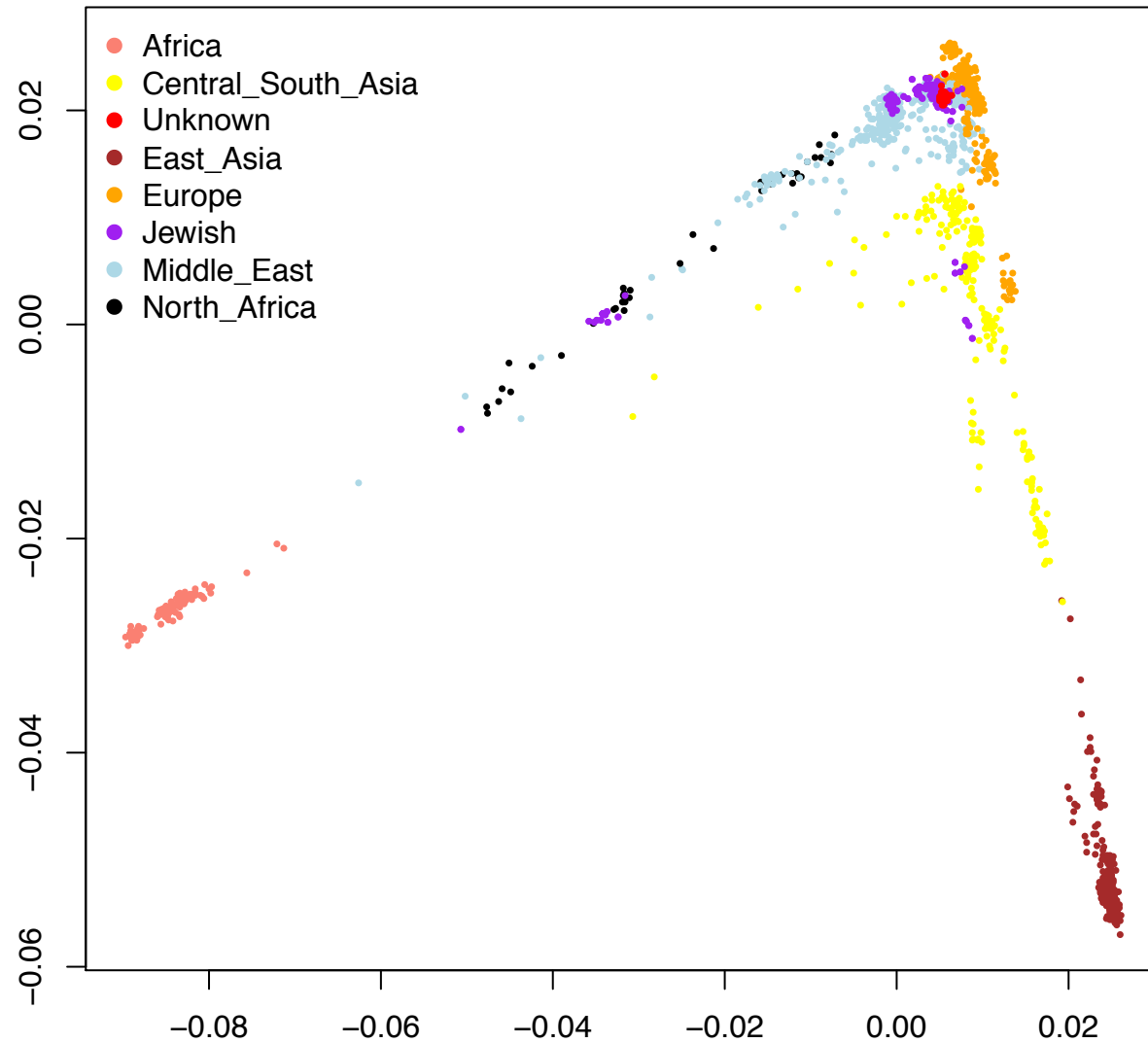
O'Connor et al. (2014)
Mol. Biol. Evol.

What is Their Geographic Ancestry?



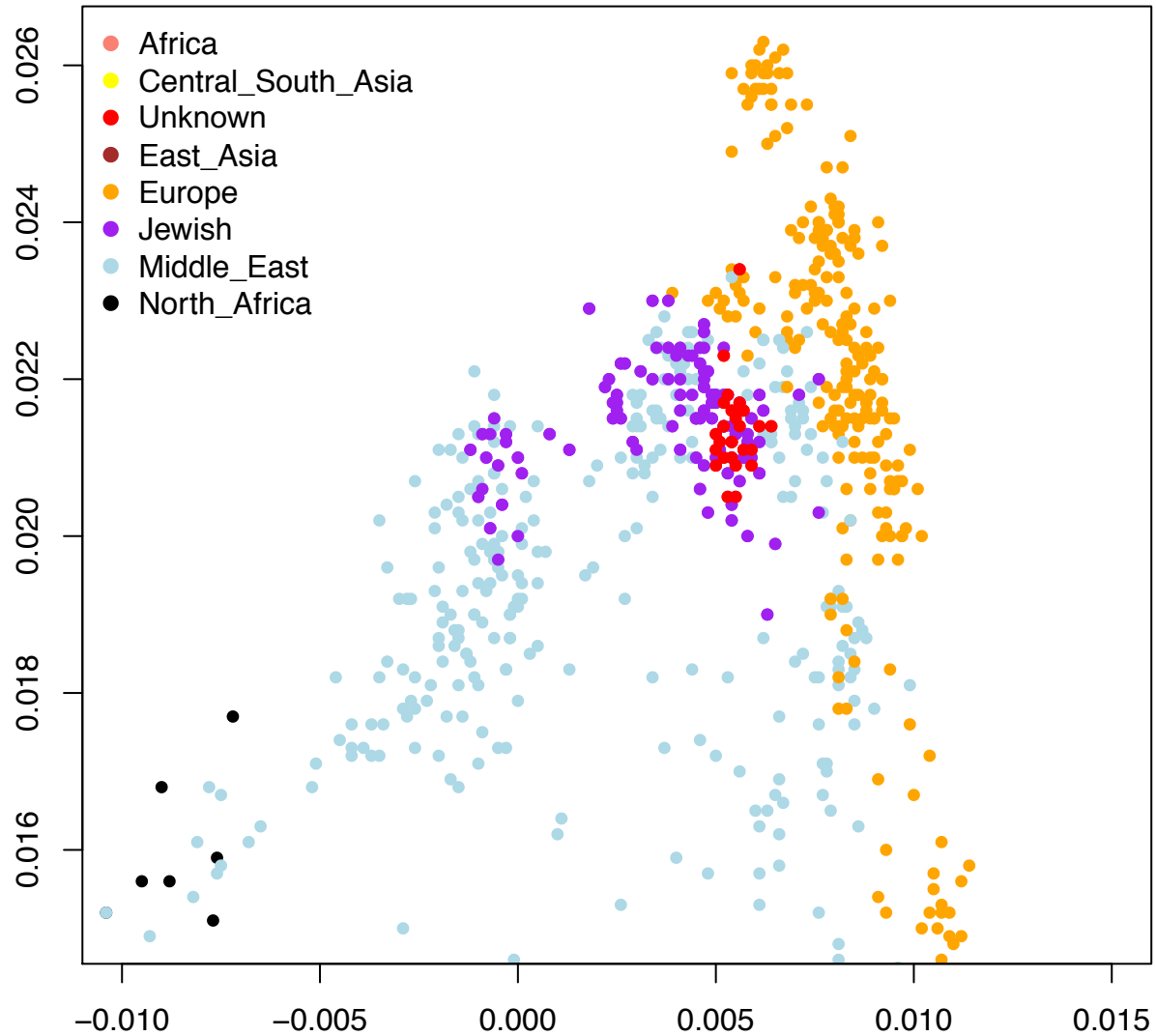
O'Connor et al. (2014)
Mol. Biol. Evol.

PCA of Global Diversity Including Cryptic Population



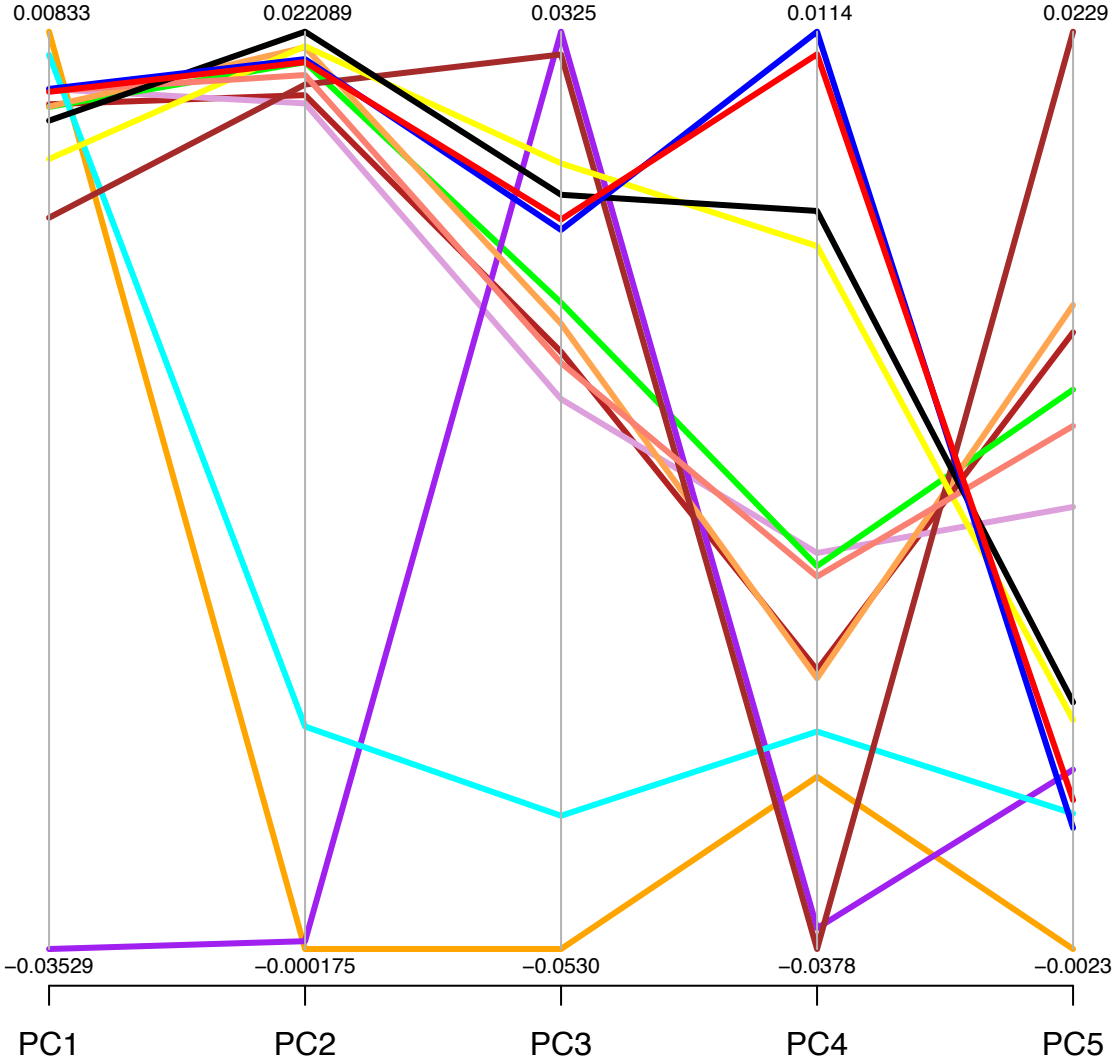
O'Connor et al. (2014)
Mol. Biol. Evol.

PCA of Global Diversity Including Cryptic Population



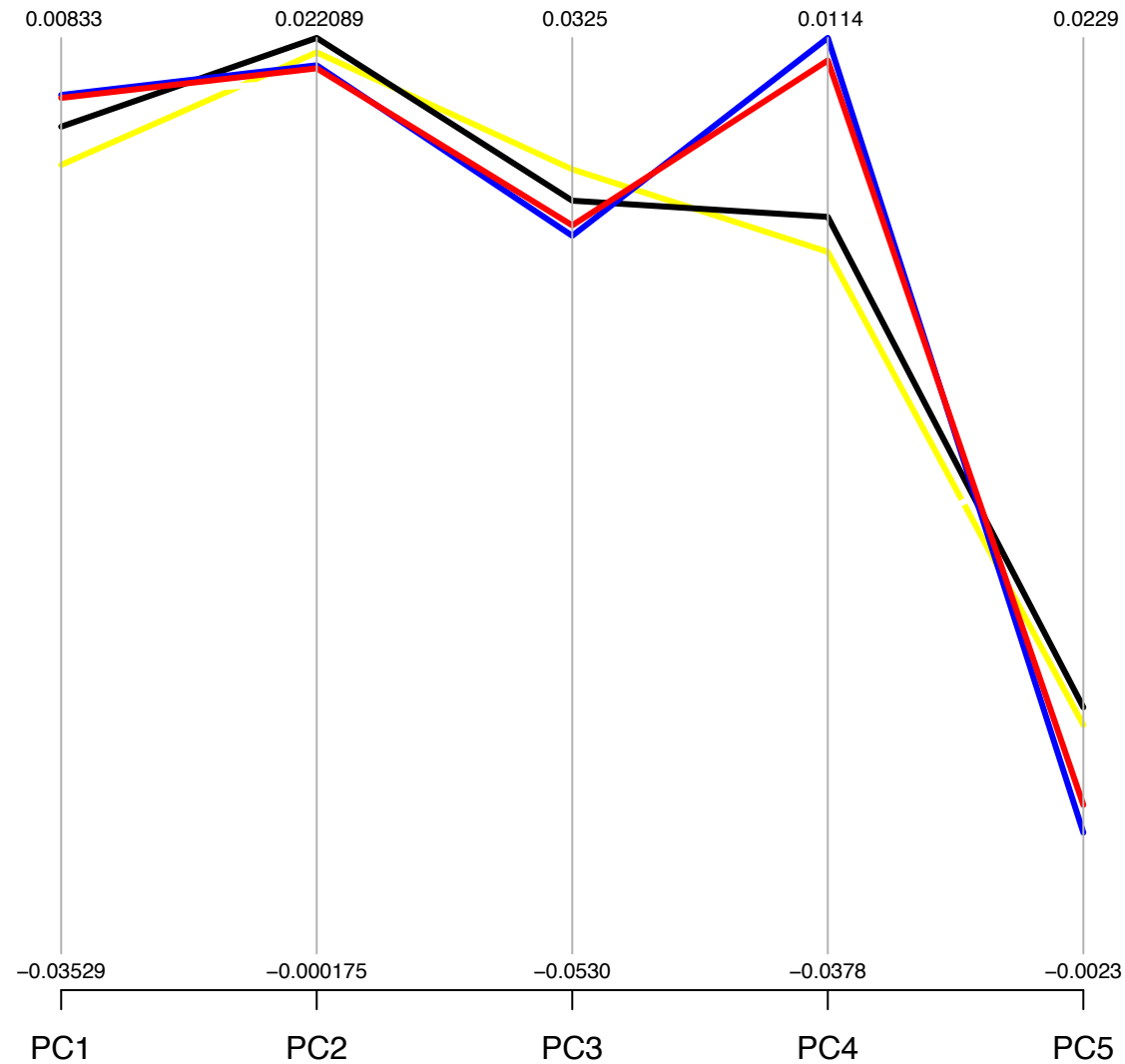
O'Connor et al. (2014)
Mol. Biol. Evol.

Population Average PCA with More Axes



O'Connor et al. (2014)
Mol. Biol. Evol.

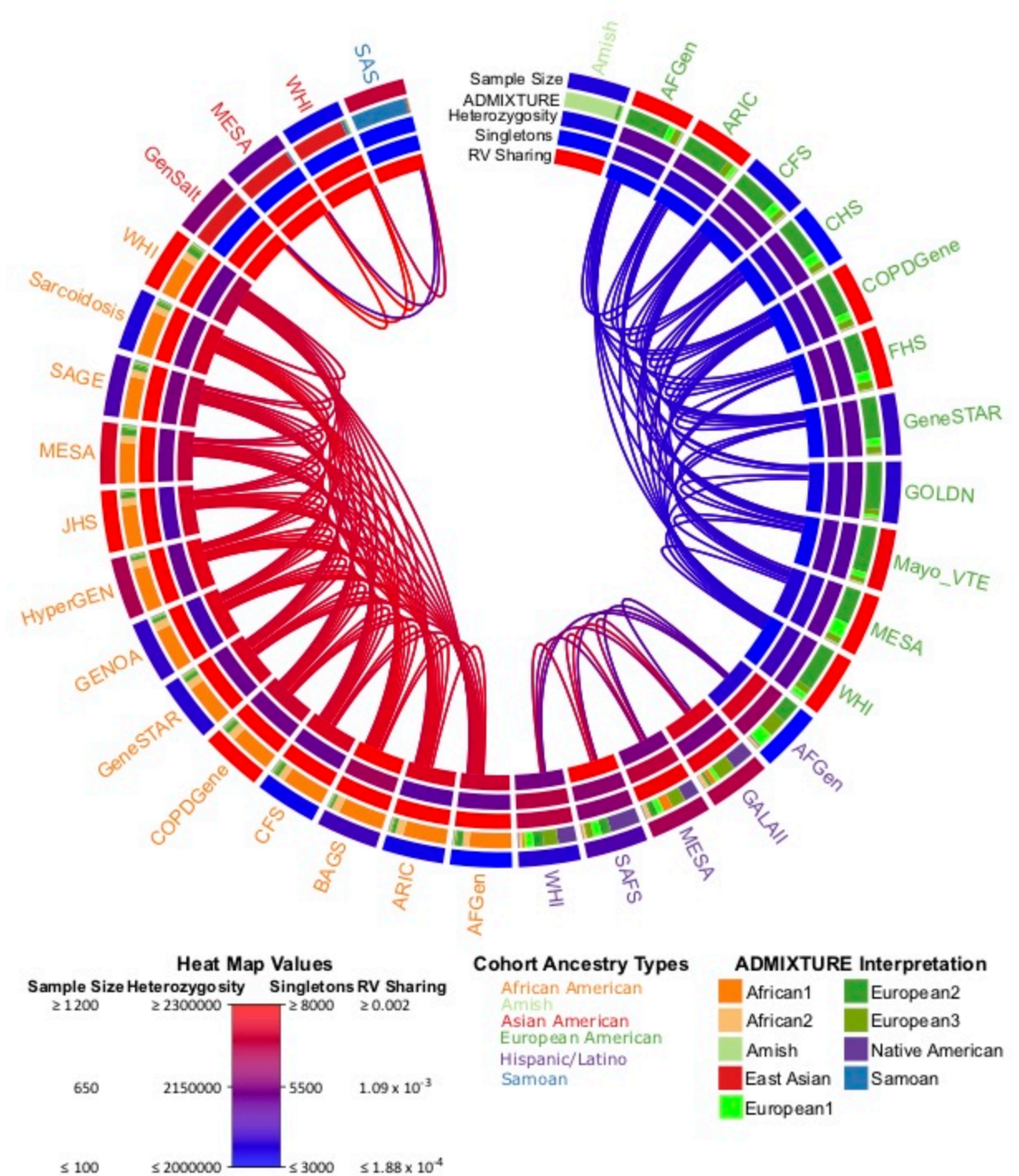
Population Average PCA with More Axes



O'Connor et al. (2014)
Mol. Biol. Evol.

Trans-Omics for Precision Medicine (TOPMed) Cohorts

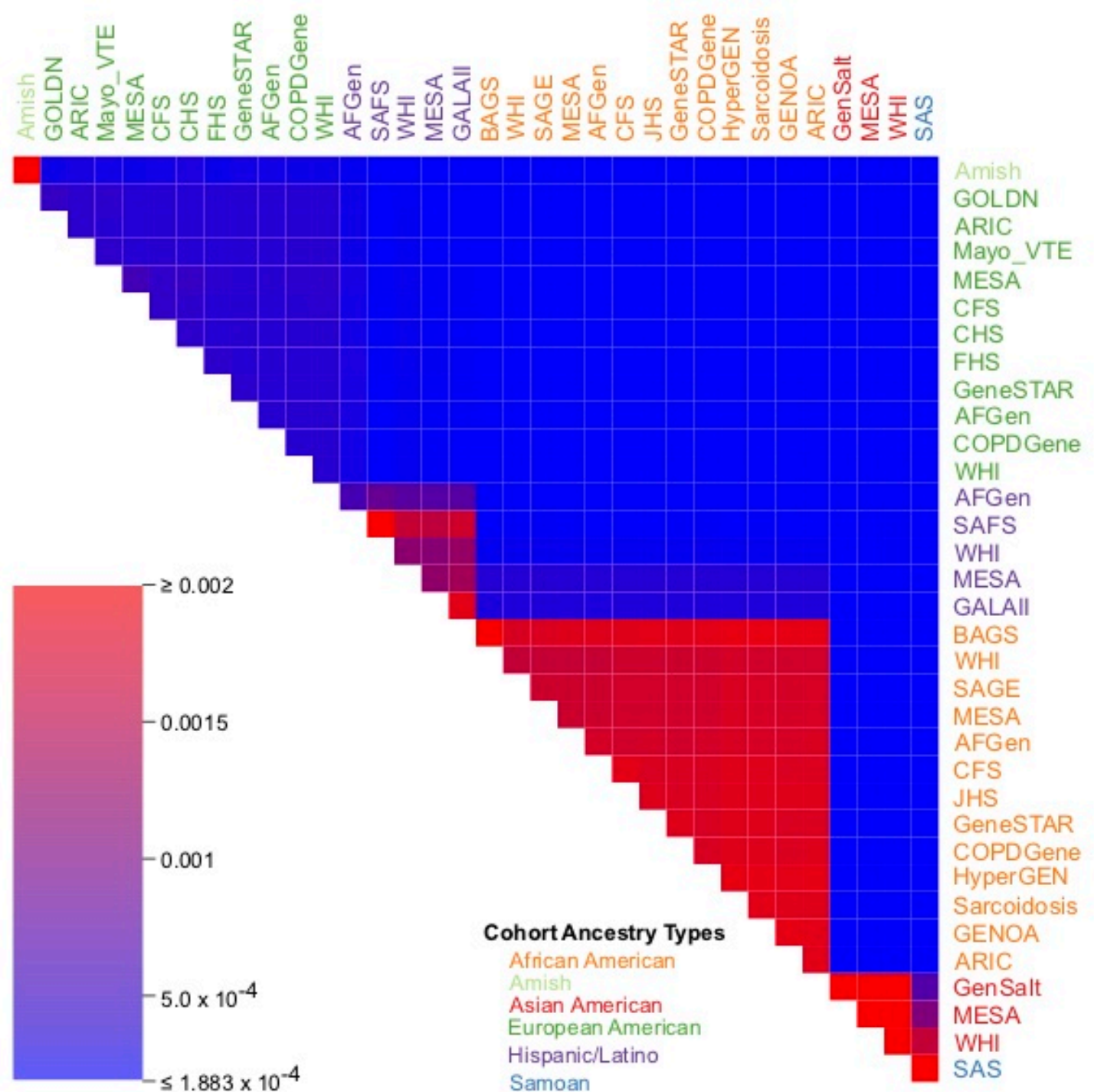
- $N \cong 55K$
- Predominantly African, Latino, and European American
 - Samoa
 - Amish
- All are well characterized for heart, lung, blood, and sleep phenotypes



Rare variant sharing across cohorts

- Allele Count 2 to 100
- Corrected for:
 - sample size
 - Genome-wide heterozygosity

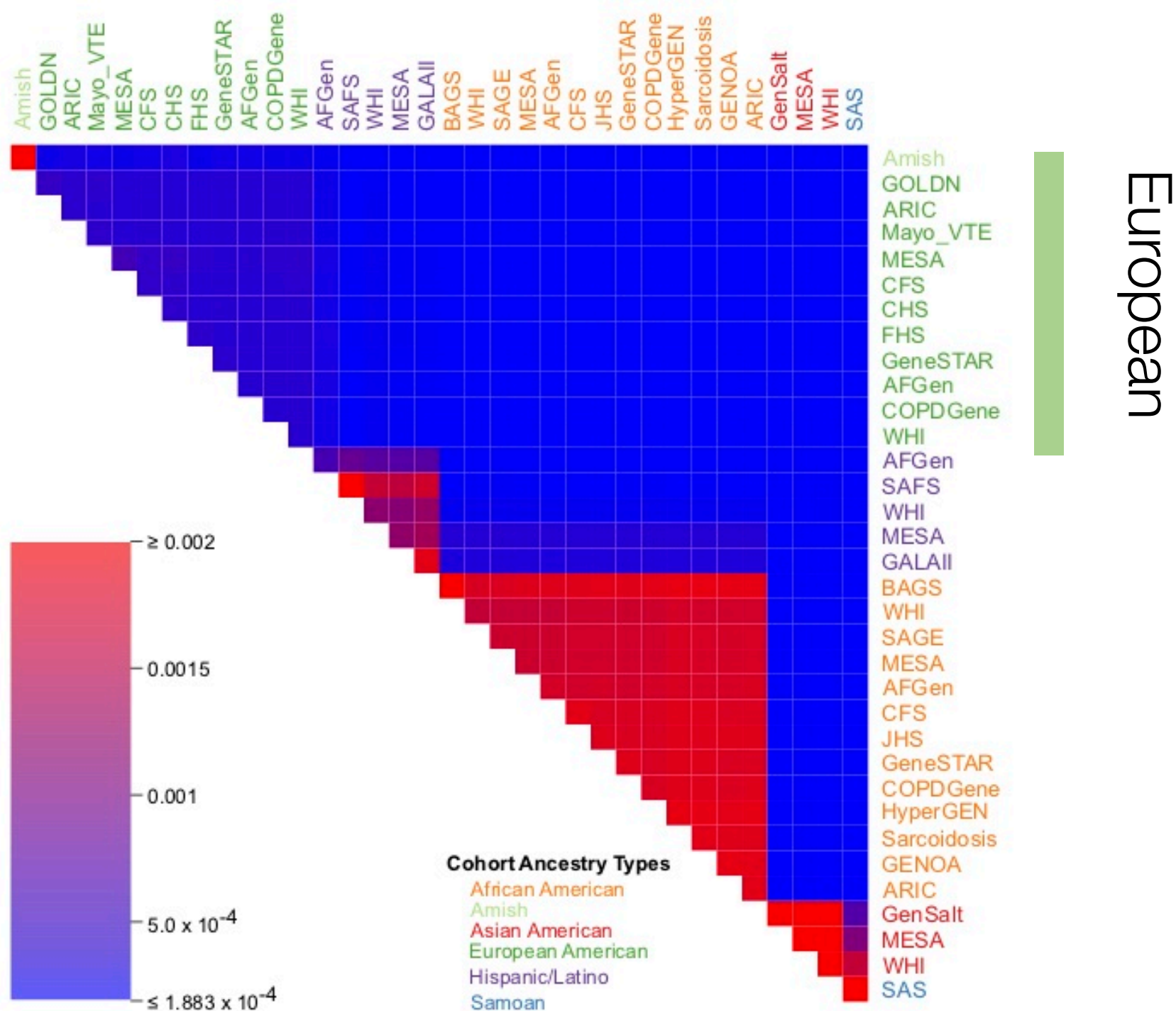
Taliun et al. (2019) Bioarxiv



Rare variant sharing across cohorts

- Allele Count 2 to 100
- Corrected for:
 - sample size
 - Genome-wide heterozygosity

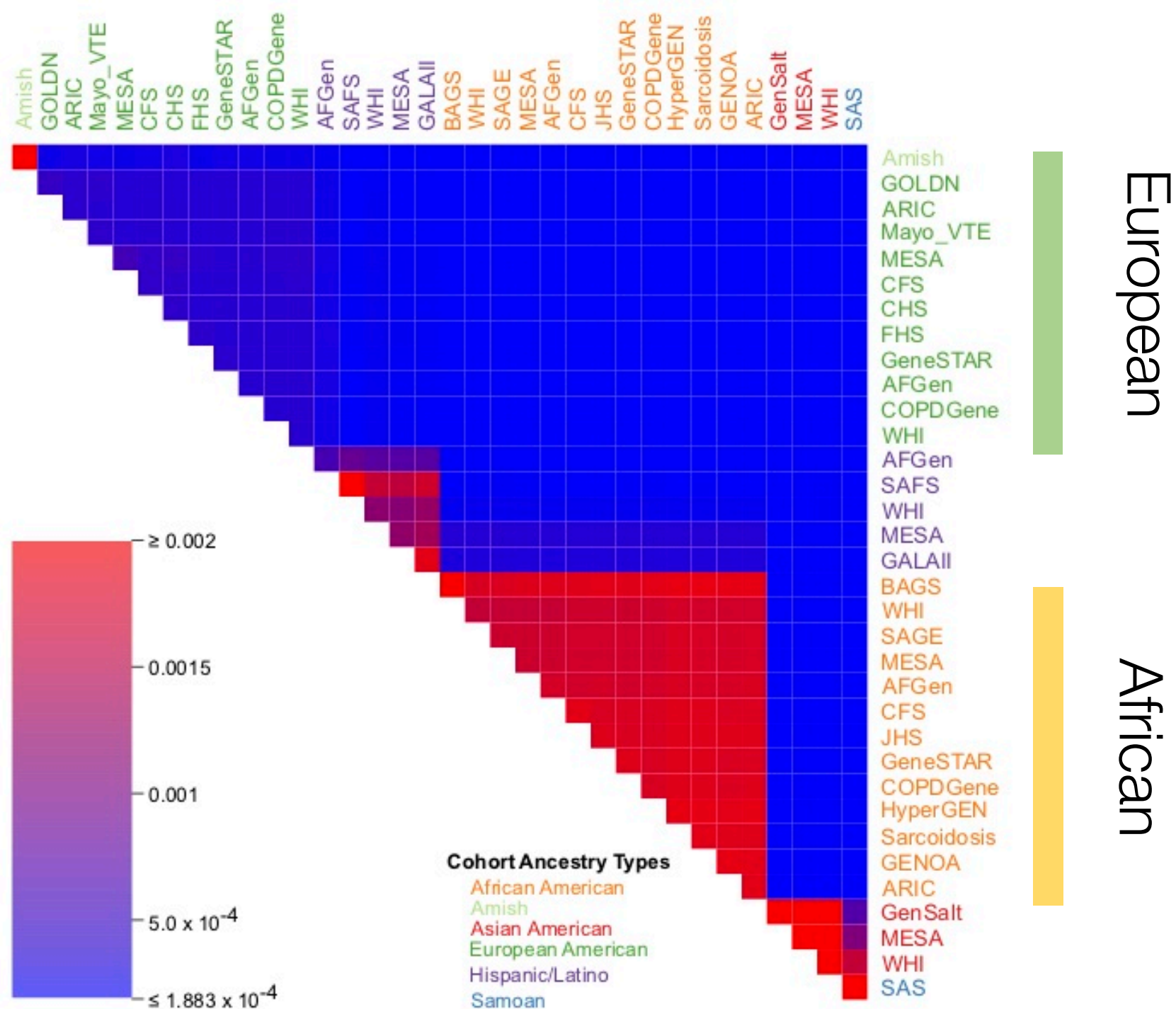
Taliun et al. (2019) Bioarxiv



Rare variant sharing across cohorts

- Allele Count 2 to 100
- Corrected for:
 - sample size
 - Genome-wide heterozygosity

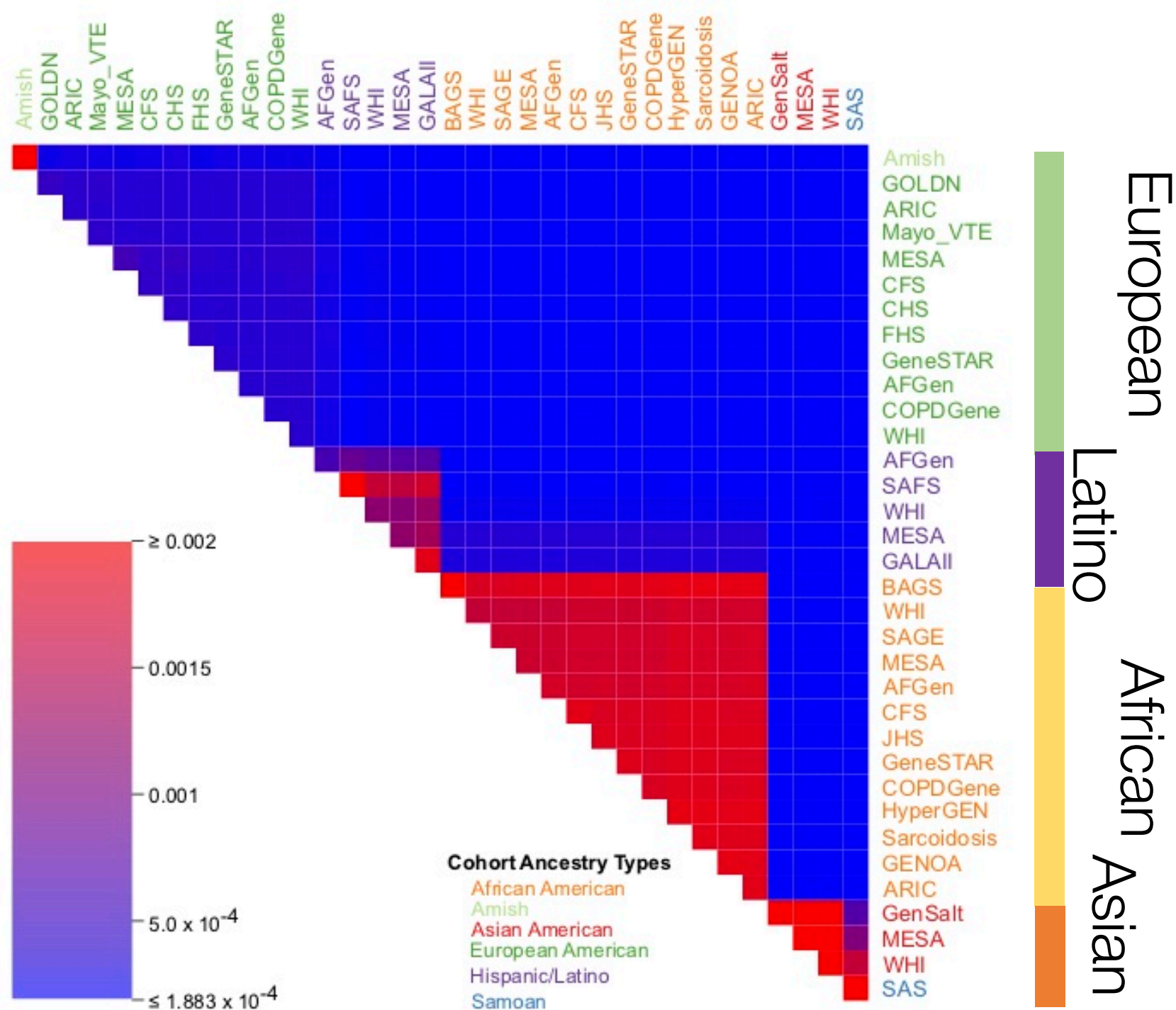
Taliun et al. (2019) Bioarxiv



Rare variant sharing across cohorts

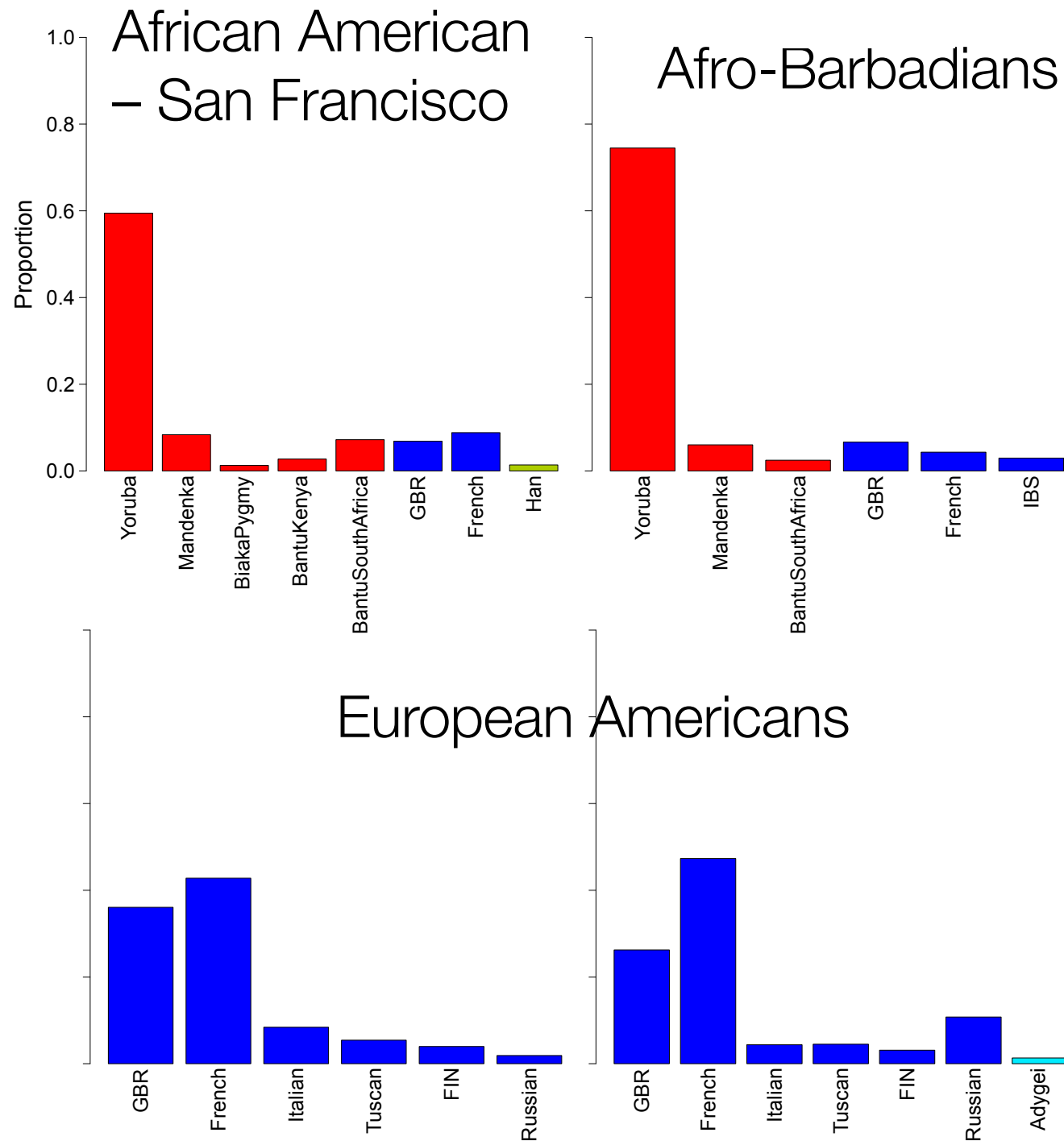
- Allele Count 2 to 100
- Corrected for:
 - sample size
 - Genome-wide heterozygosity

Taliun et al. (2019) Bioarxiv



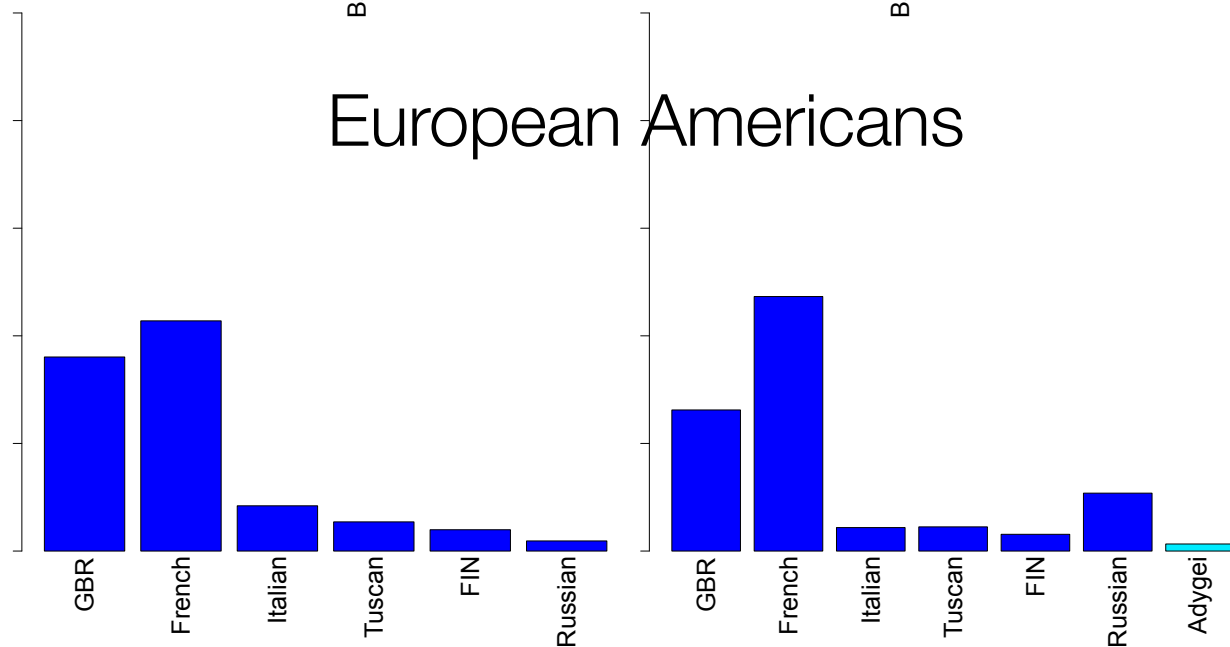
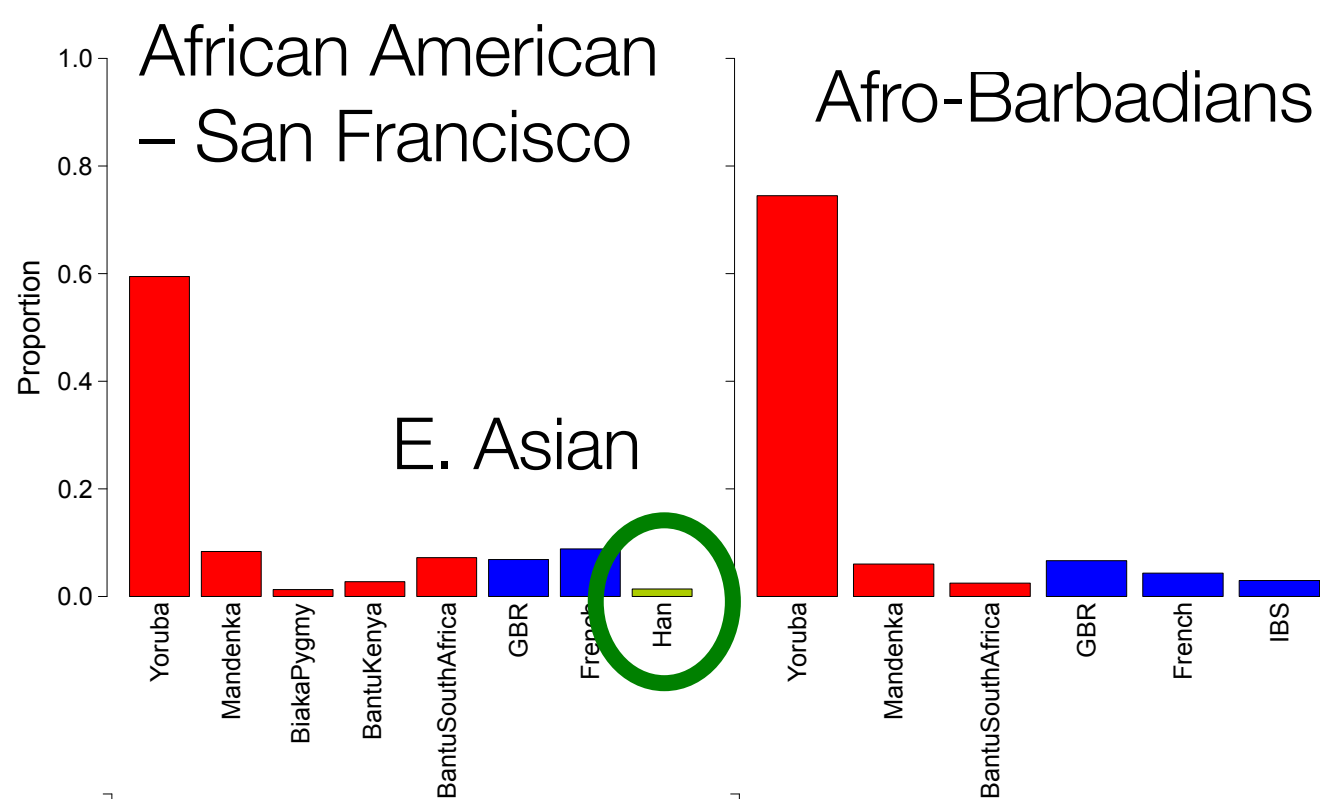
fineStructure analysis of genome-wide ancestry

- African
- Caucasiasia
- East Asian
- European



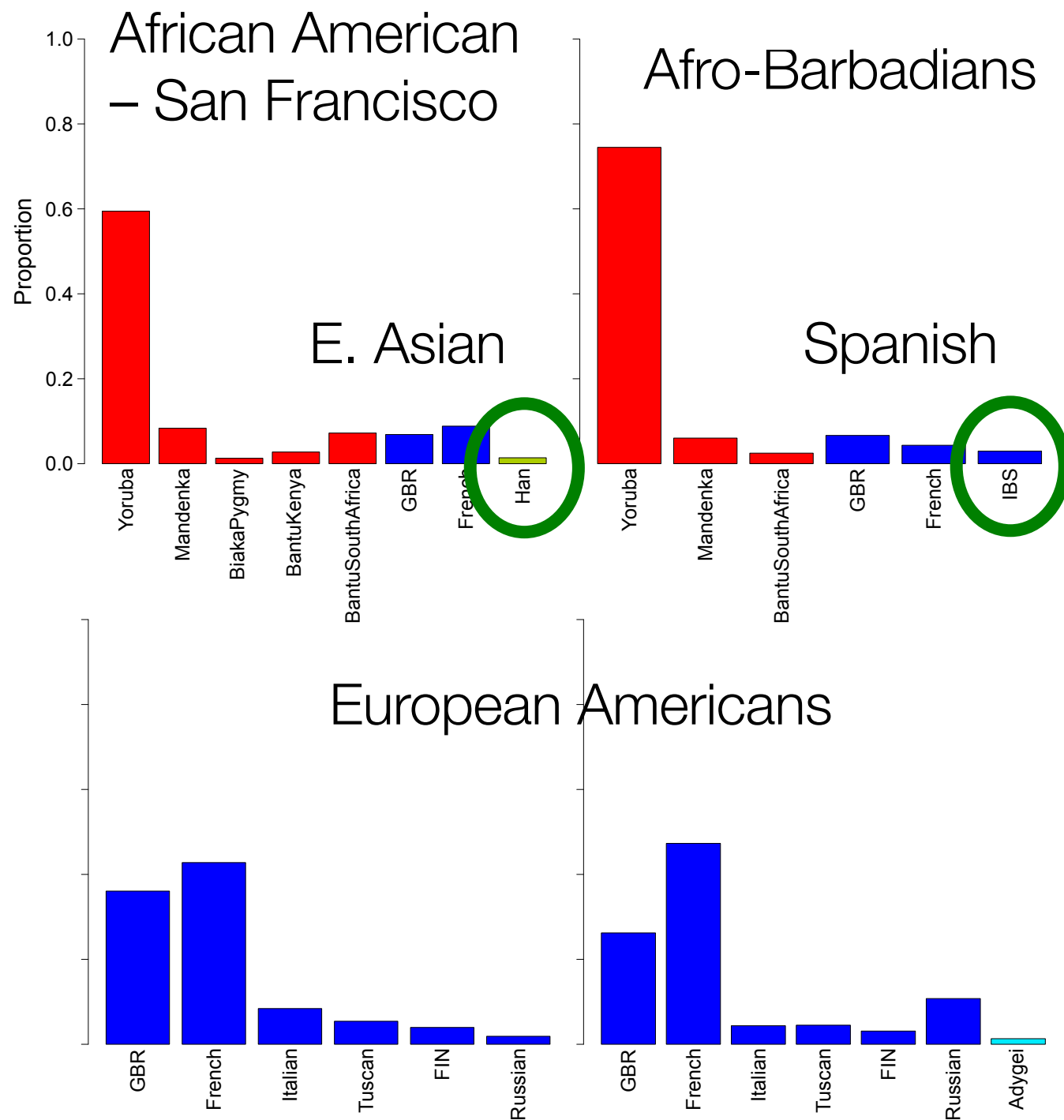
fineStructure analysis of genome-wide ancestry

- African
- Caucasiasia
- East Asian
- European



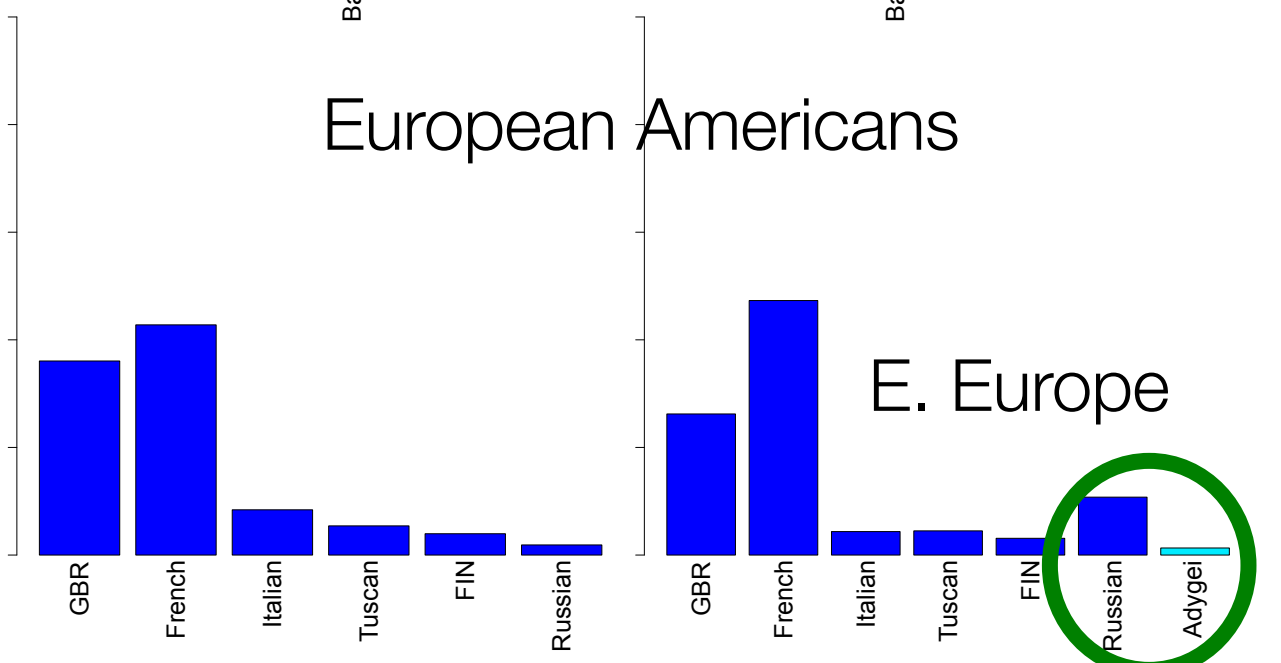
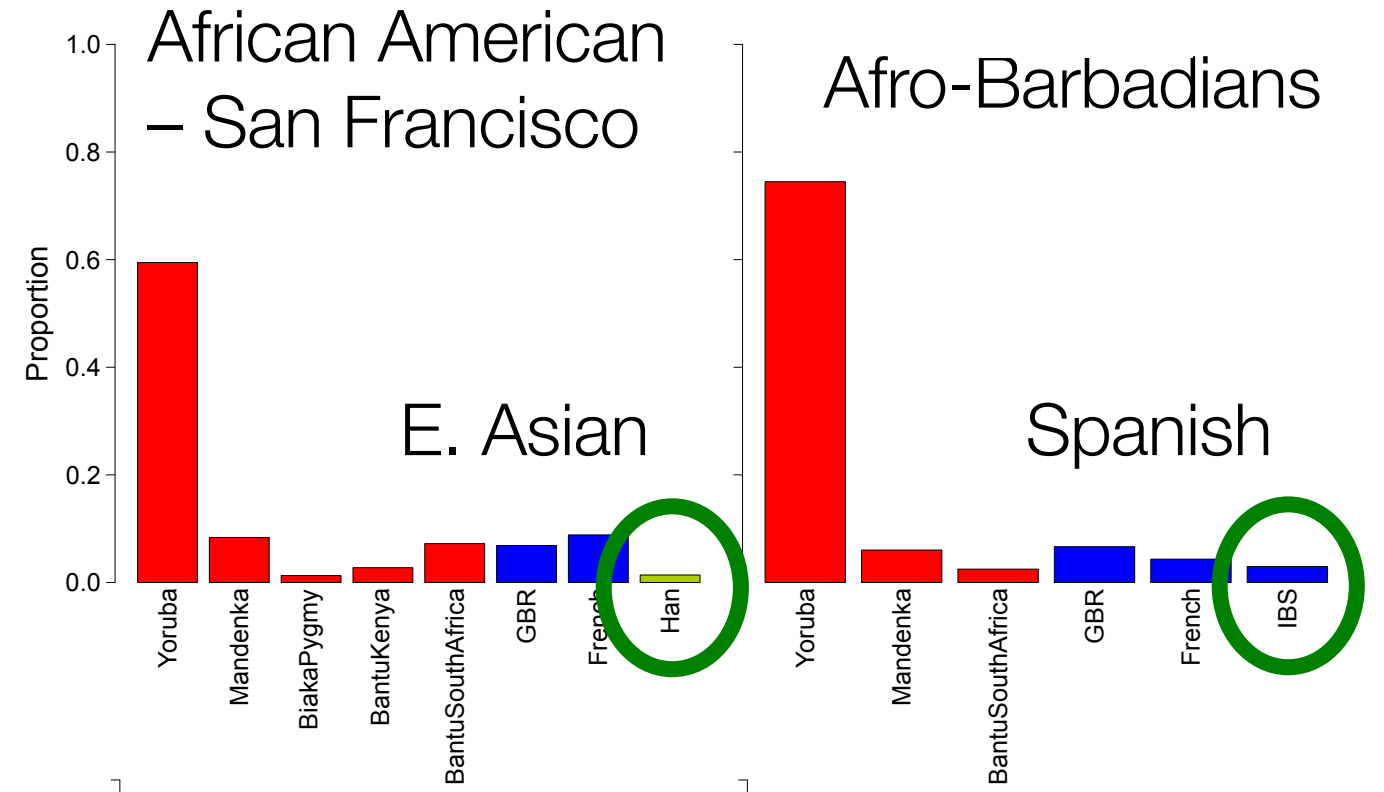
fineStructure analysis of genome-wide ancestry

- African
- Caucasiasia
- East Asian
- European



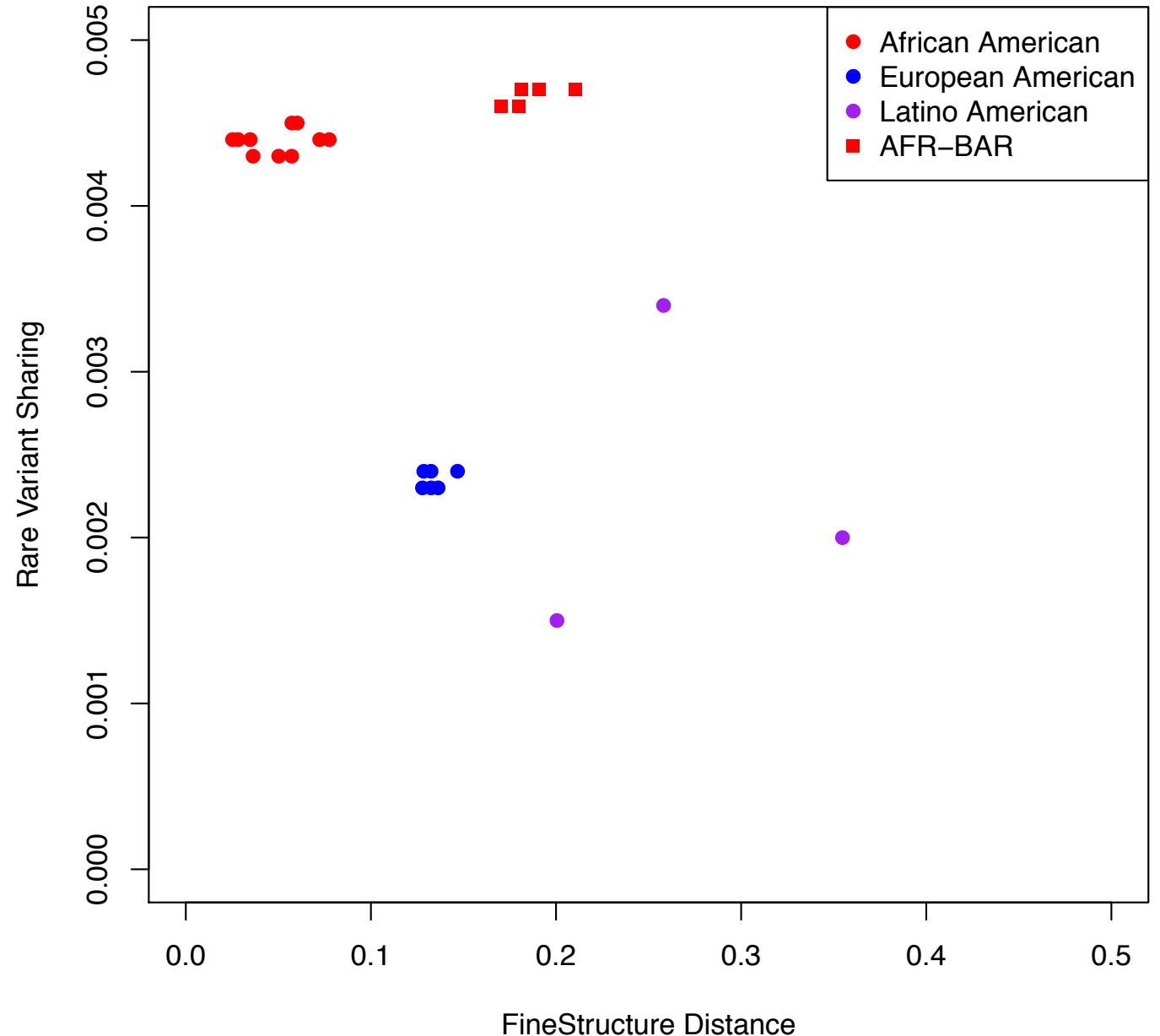
fineStructure analysis of genome-wide ancestry

- African
- Caucasiasia
- East Asian
- European



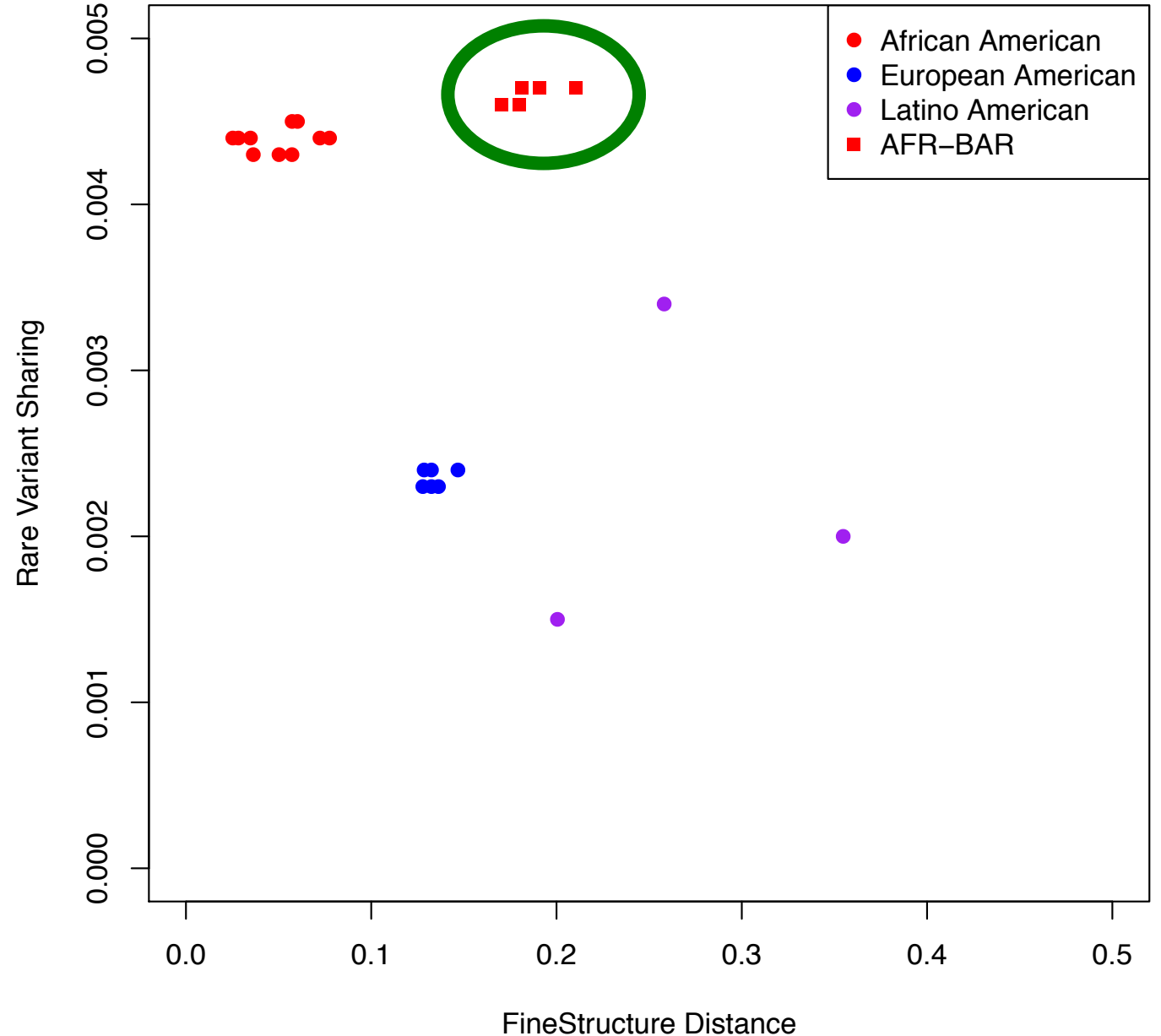
African American's have more homogeneous ancestral proportions

- Calculated Euclidian distance between fineSTRUCTURE proportions
- African American cohorts have the shortest distance and the greatest rare variant sharing

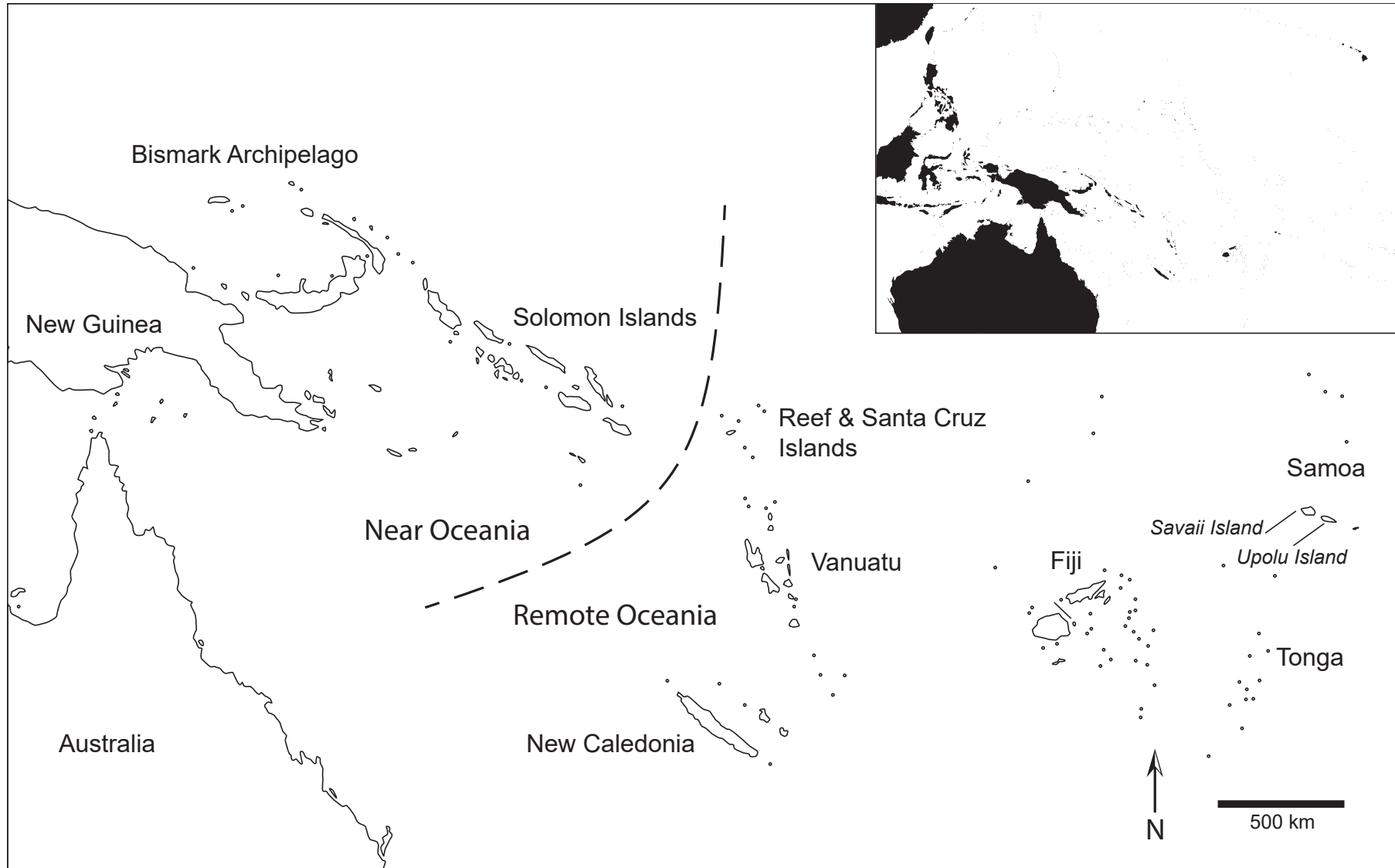


African American's have more homogeneous ancestral proportions

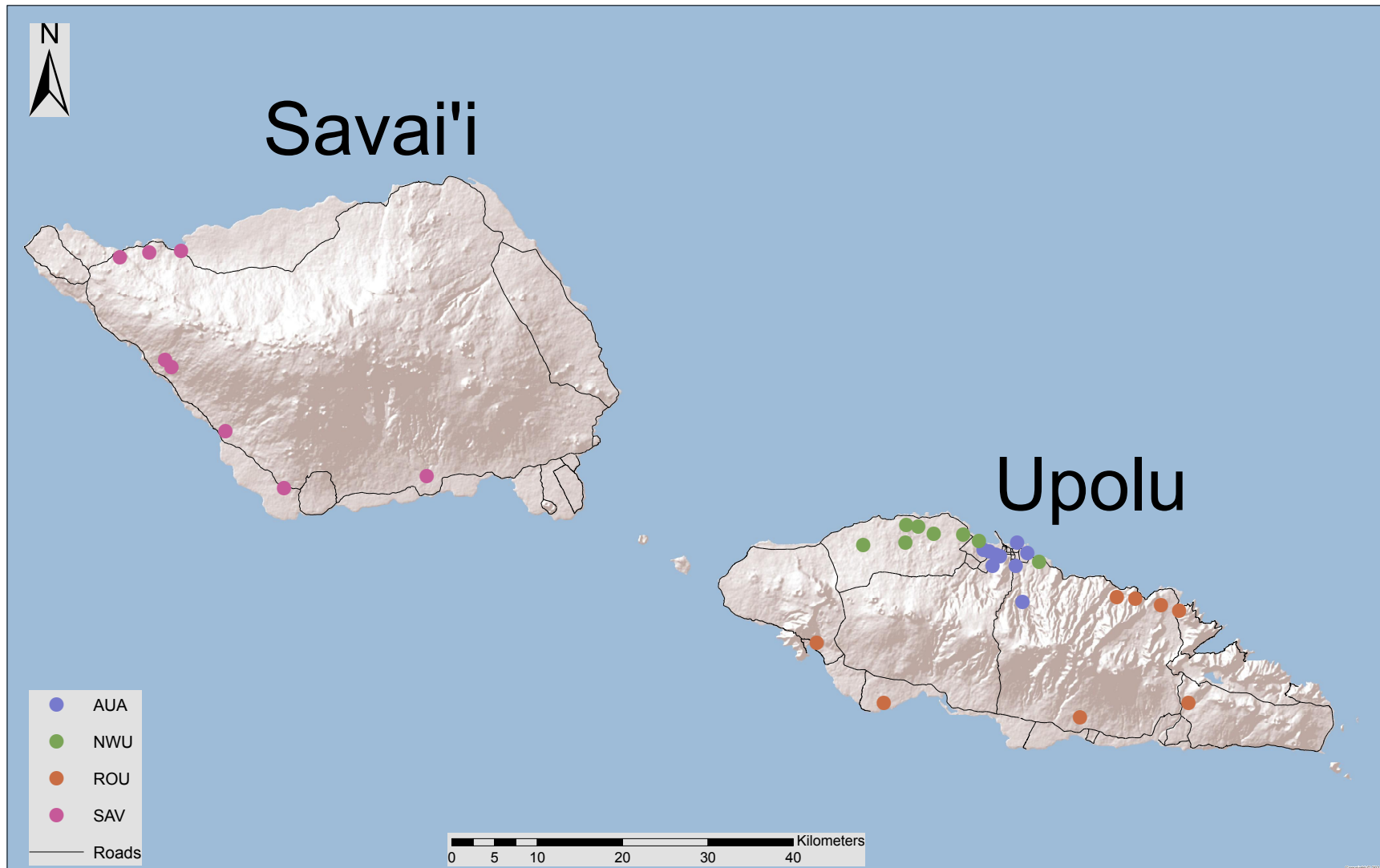
- Calculated Euclidian distance between fineSTRUCTURE proportions
- African American cohorts have the shortest distance and the greatest rare variant sharing



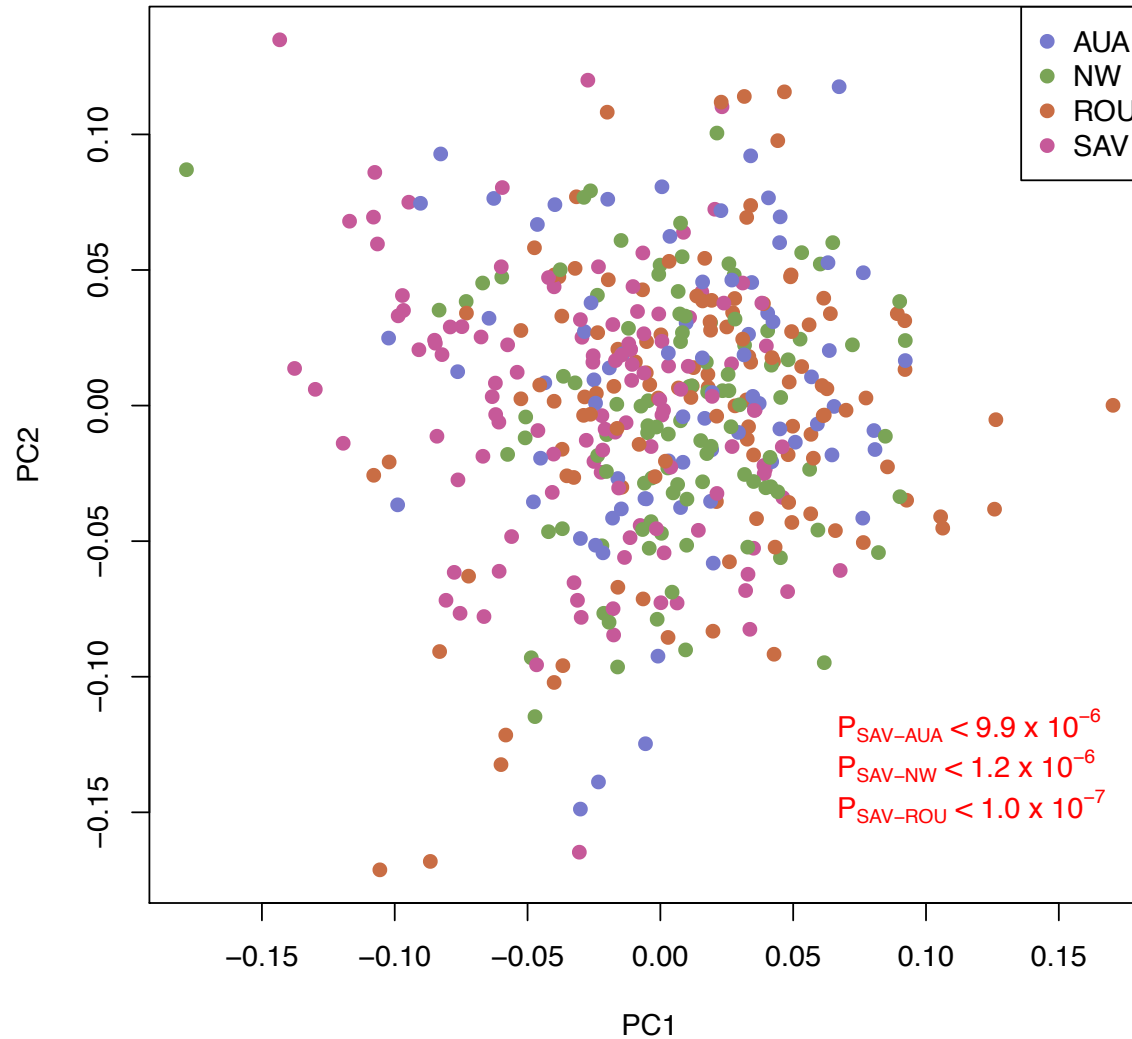
Quick background on Samoa



Quick background on Samoa

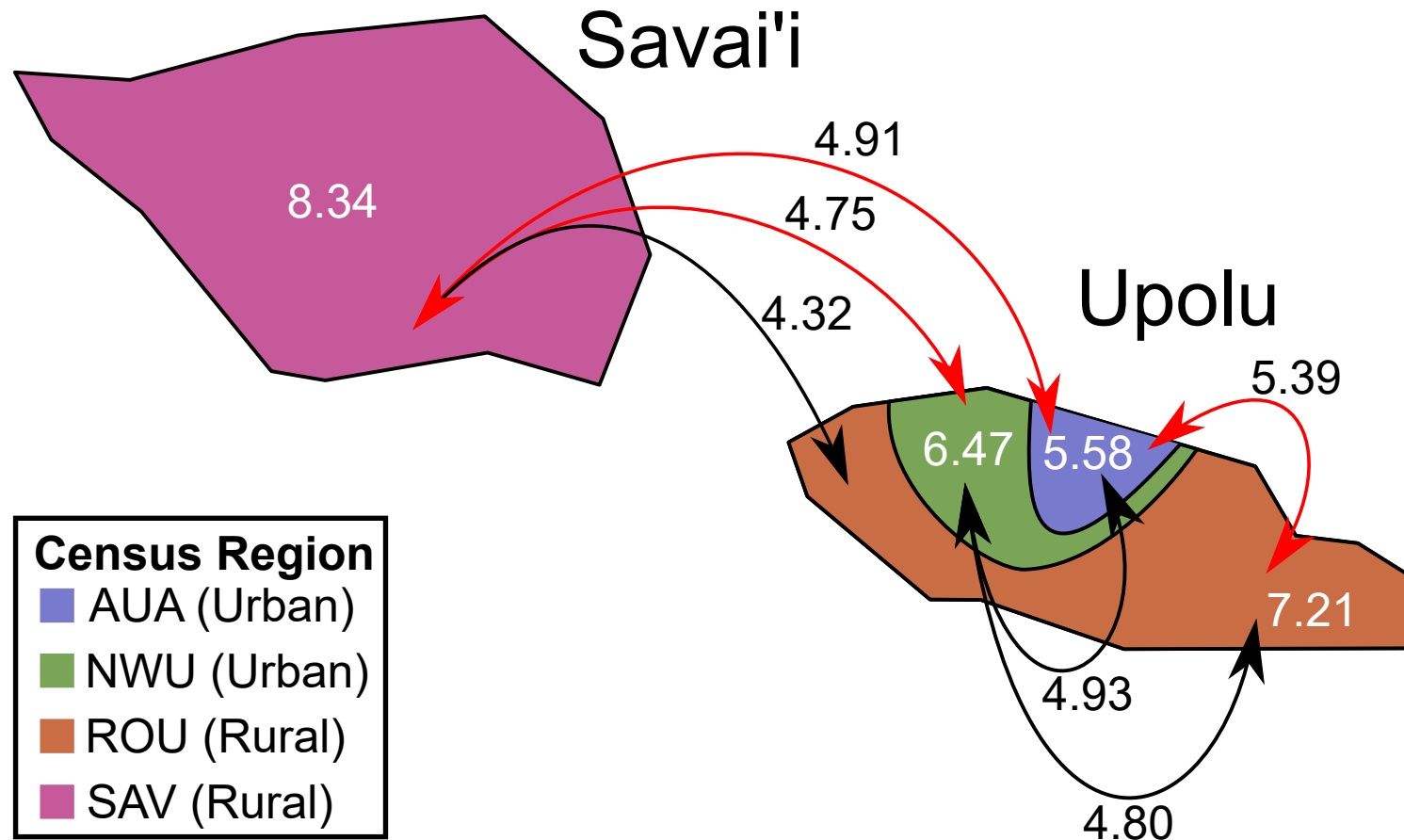


PCA with all variants can't distinguish the two islands well.



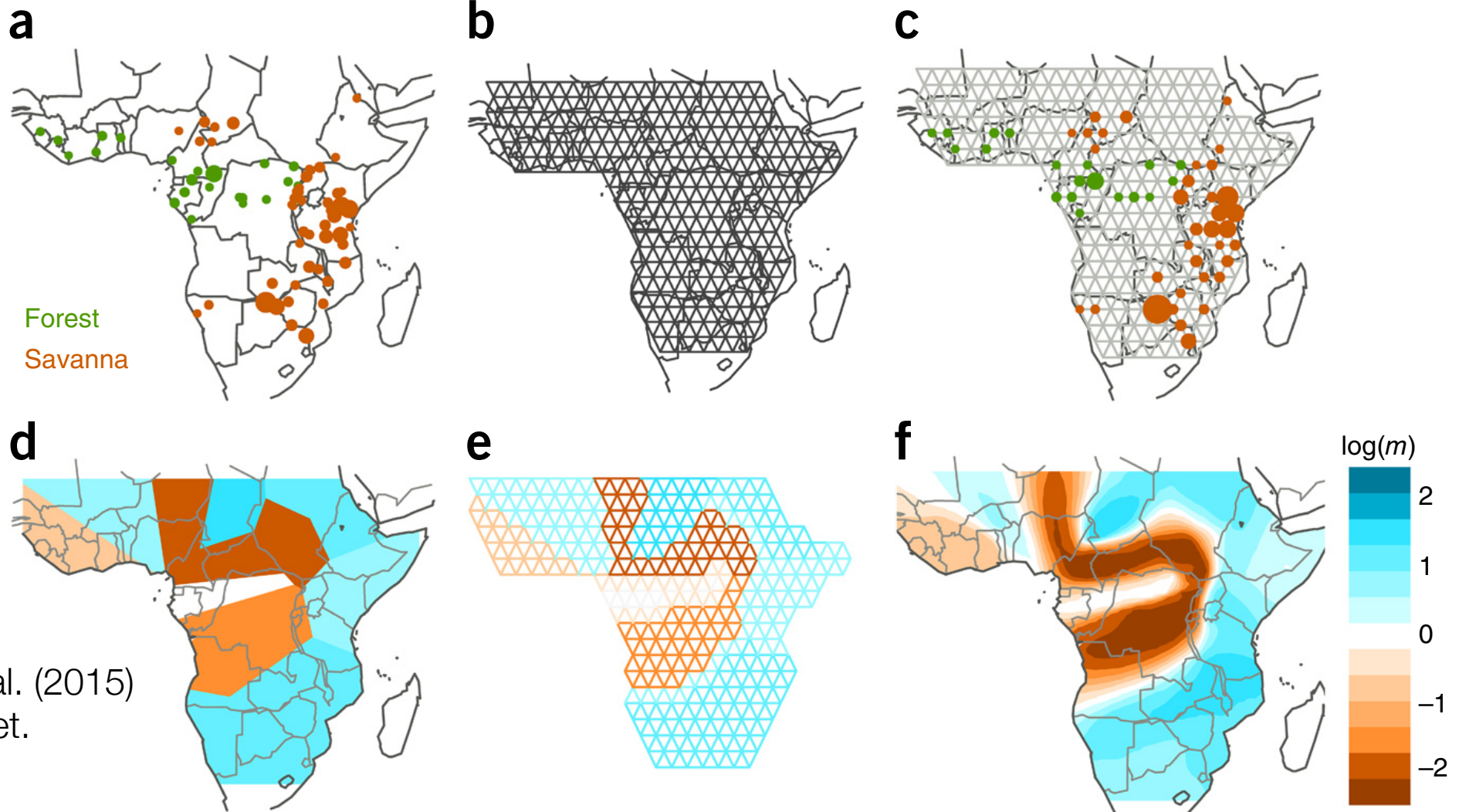
Harris et al.
(https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3329885)

Rare variant Sharing in Samoa



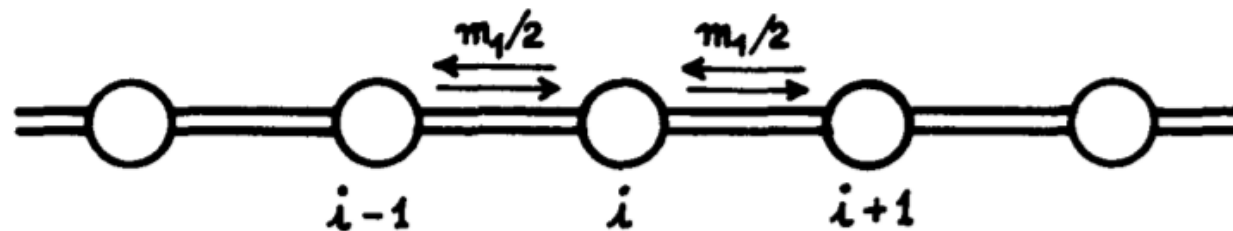
Harris et al.
(https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3329885)

Estimated Effective Migration Surfaces (EEMS)



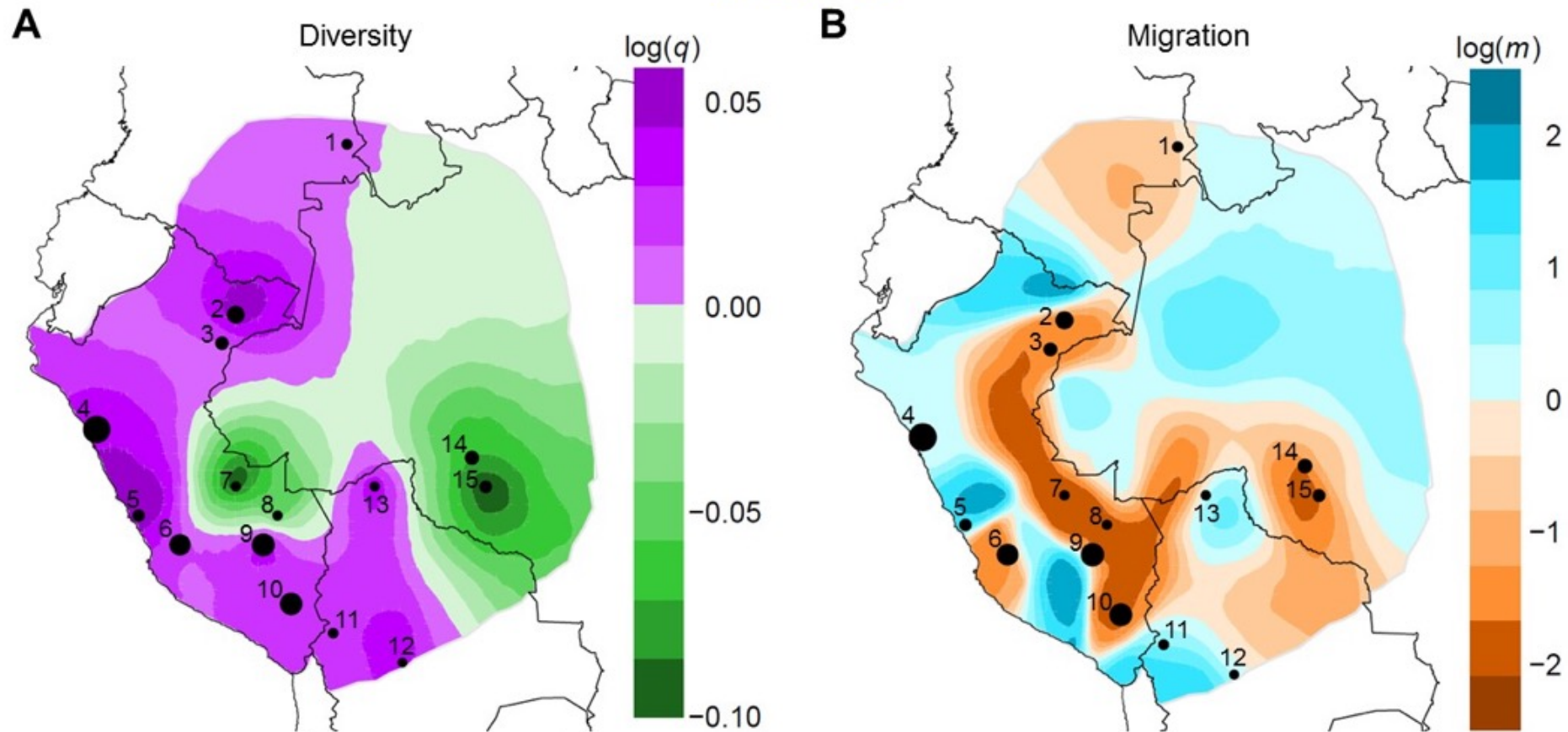
Assumptions: Stepping Stone Model

- Migration can only occur between adjacent demes
- Migration rate between each deme is assumed to be equal



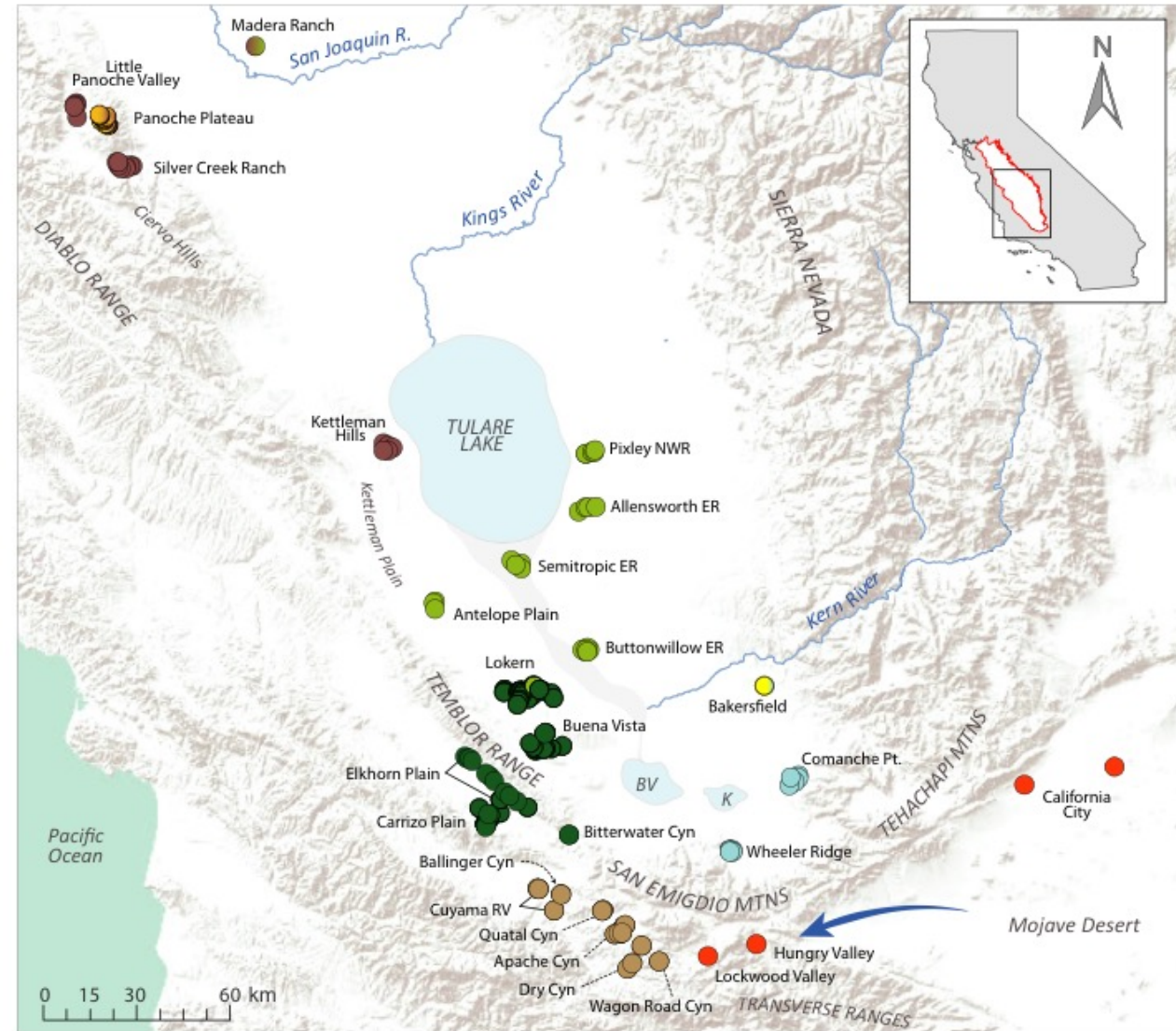
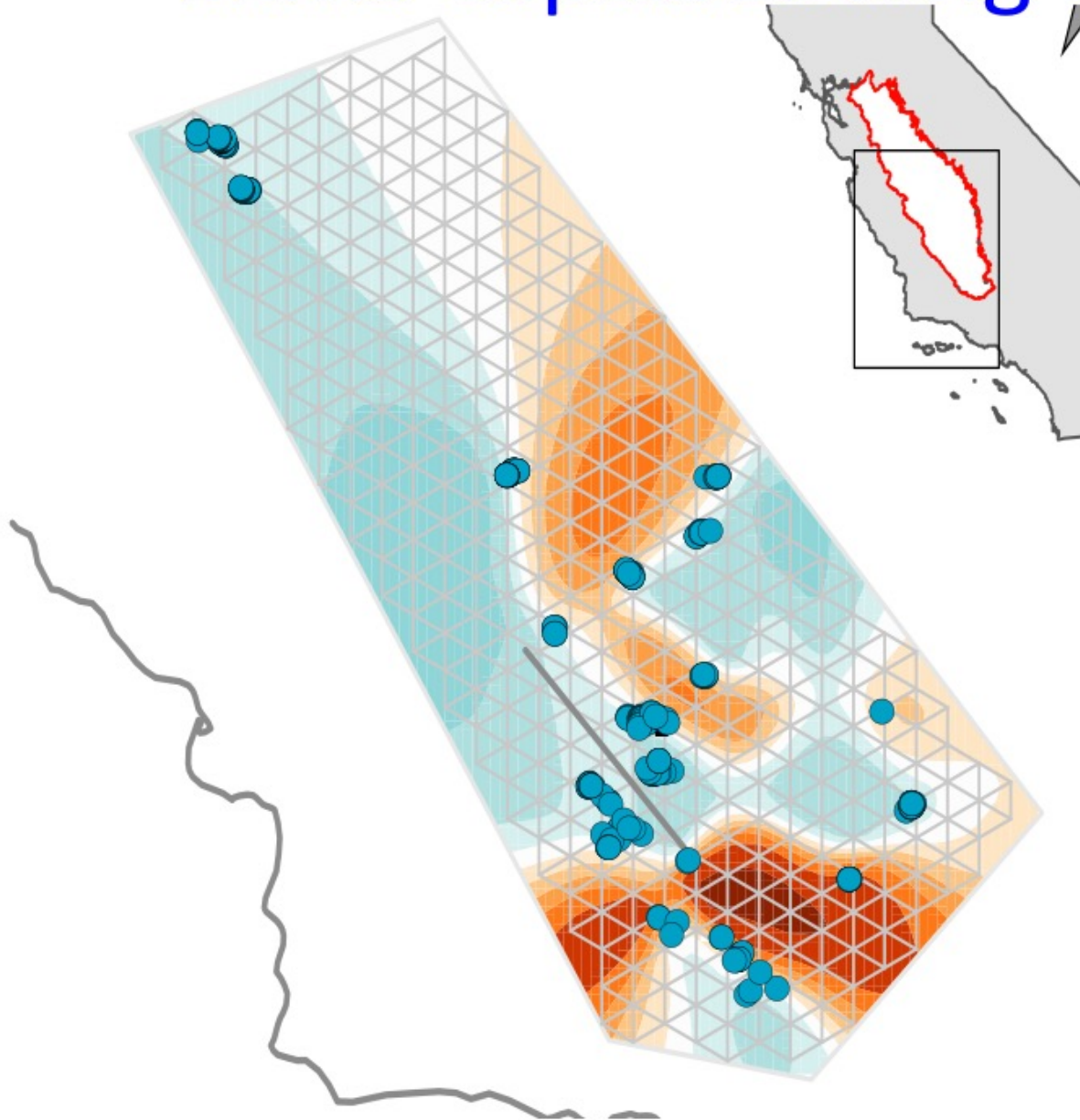
Kimura and Weiss (1964)

EEMS: Migration and diversity within Peru



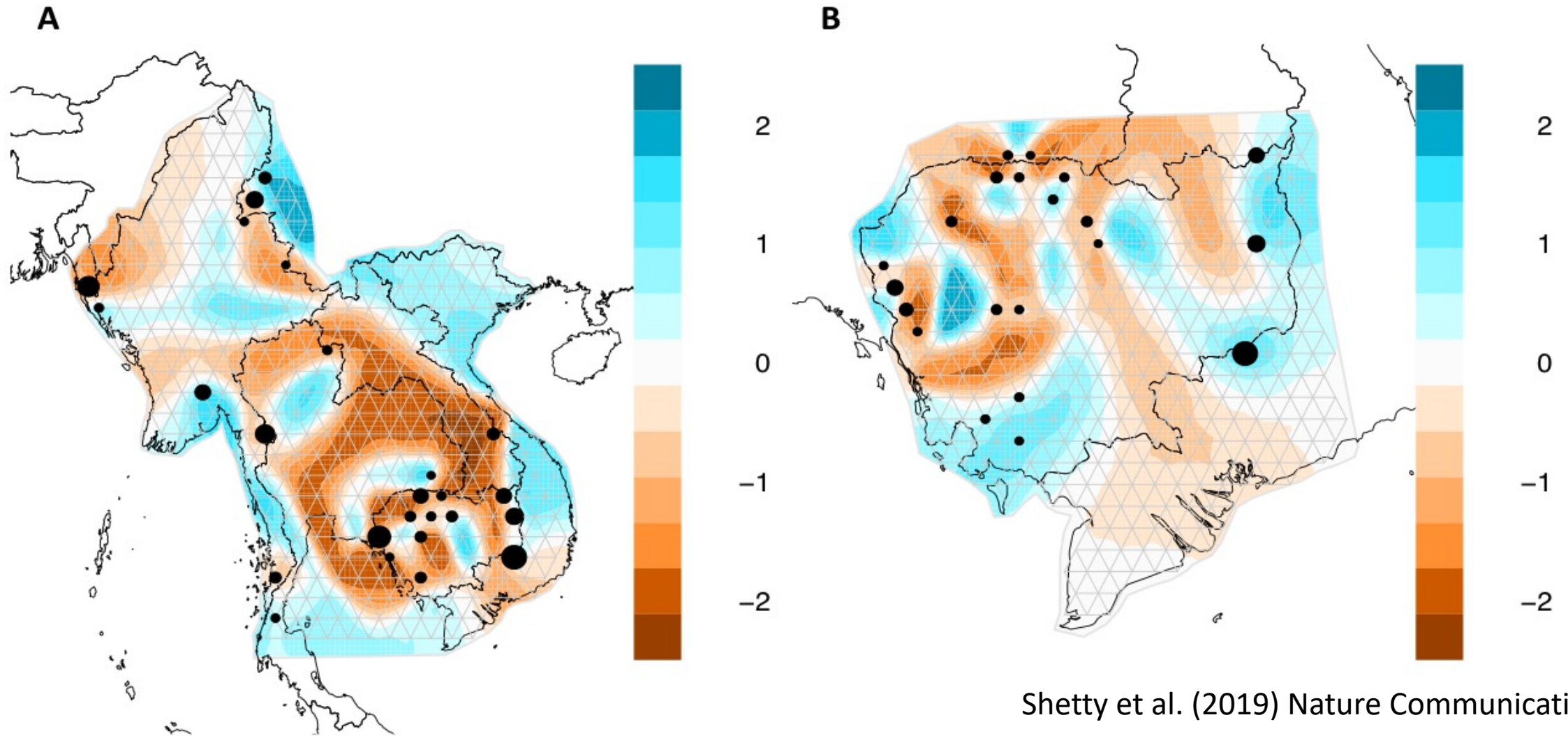
Deme	Populations	Deme	Populations	Deme	Populations	Deme	Populations
1	Piapoco	5	Lima	9	Cusco, Qeros, Quechua	15	Surui
2	Iquitos	6	Chopccas	10	Puno, Uros		
3	Matzes	7	Nahua	11,12,13	Bolivian		
4	Moches, Trujillo	8	Matsigenka	14	Karitiana		

EEMS captures long-term migration patterns



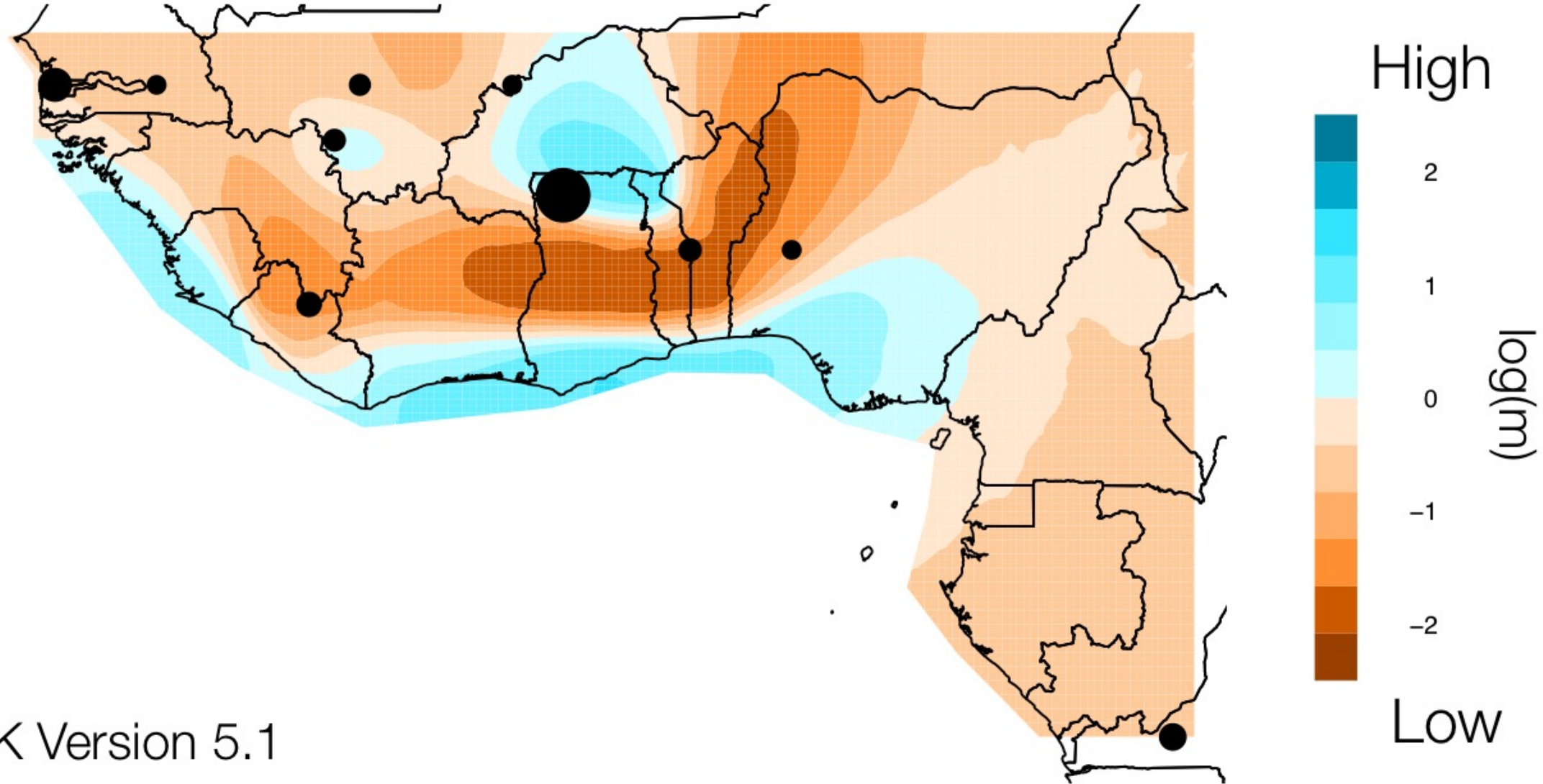
Richmond et al. (2015) Molecular Ecology

EEMS in Malaria Parasites of South East Asia



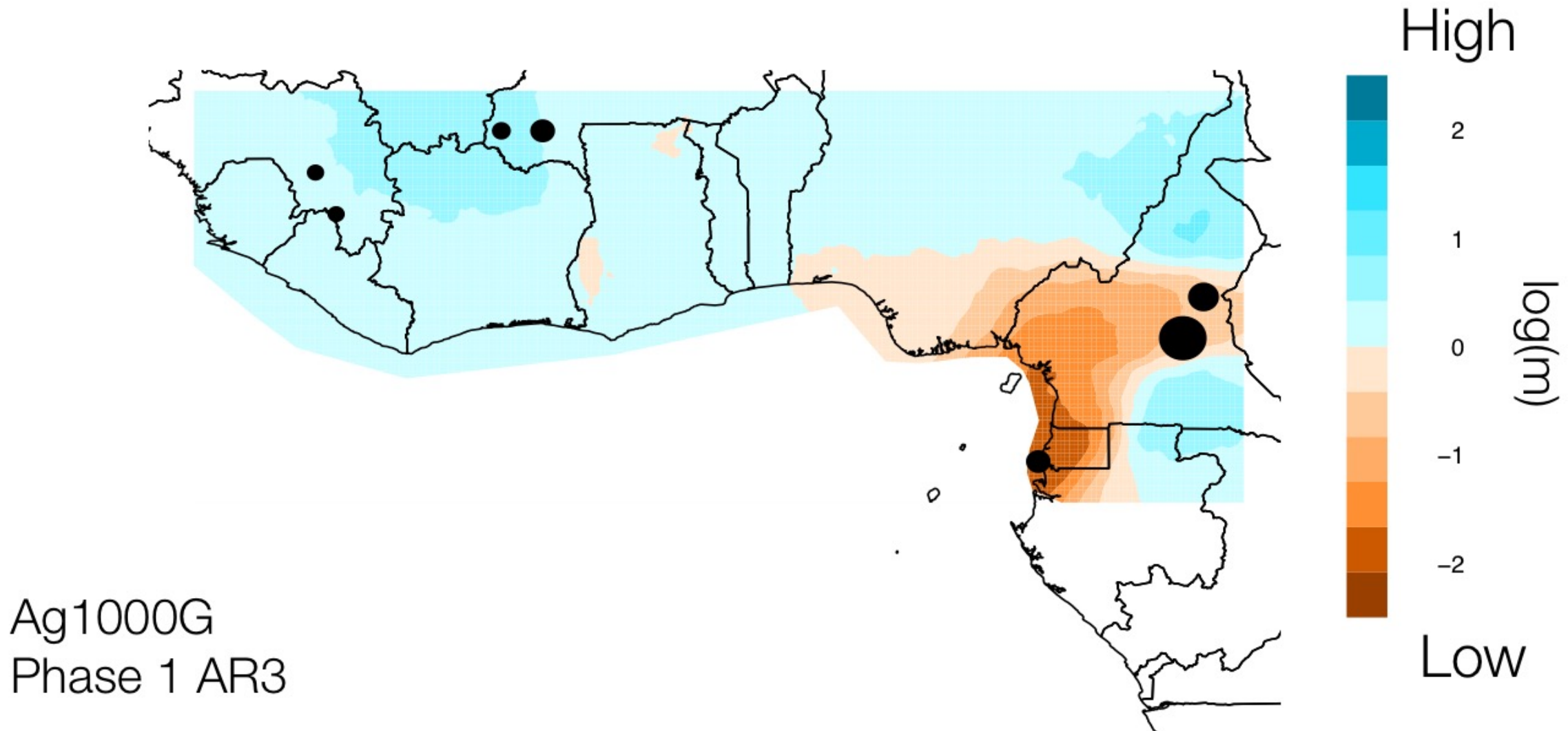
Shetty et al. (2019) Nature Communications

Application to Malaria Parasites in W. Africa

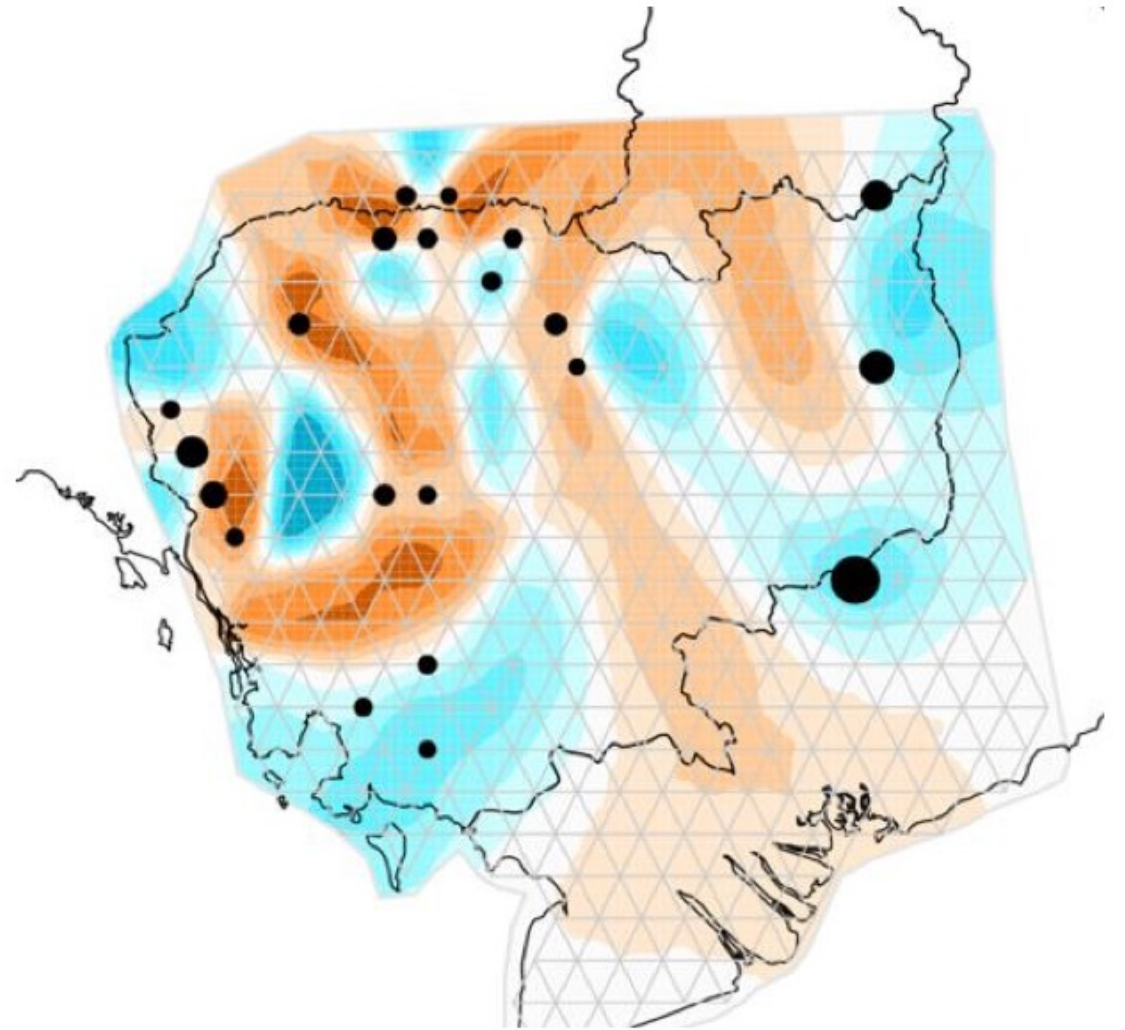
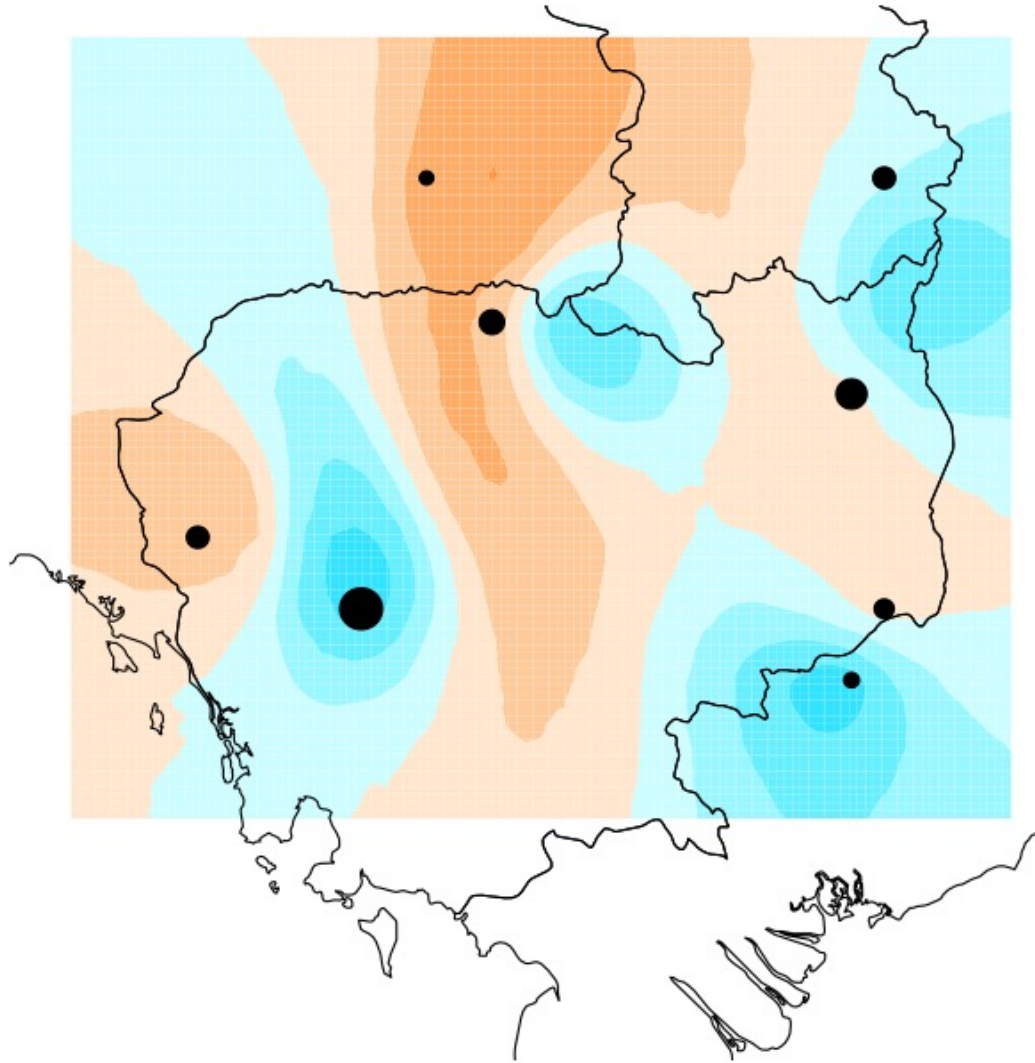


Pf3K Version 5.1

Application to Mosquito in W. Africa



Robustness of Sampling on EEMS



Concluding summary

- Fine-scale population structure is subdivisions of individuals on an ever increasingly granular scale
- Identity-by-descent and rare variant sharing are a powerful methods of identifying recent relationships and can be scaled by time.
- Cryptic population structure arises with extended relationships within a cohort, unknown to the investigators.
- EEMS can visualize migration patterns on a fine-scale illustrating cryptic structure not observed with other methods

Questions?

- CHOPCCAS
- CUSCO
- IQUITOS
- MATZES
- MOCHES
- TRUJILLO
- UROS

