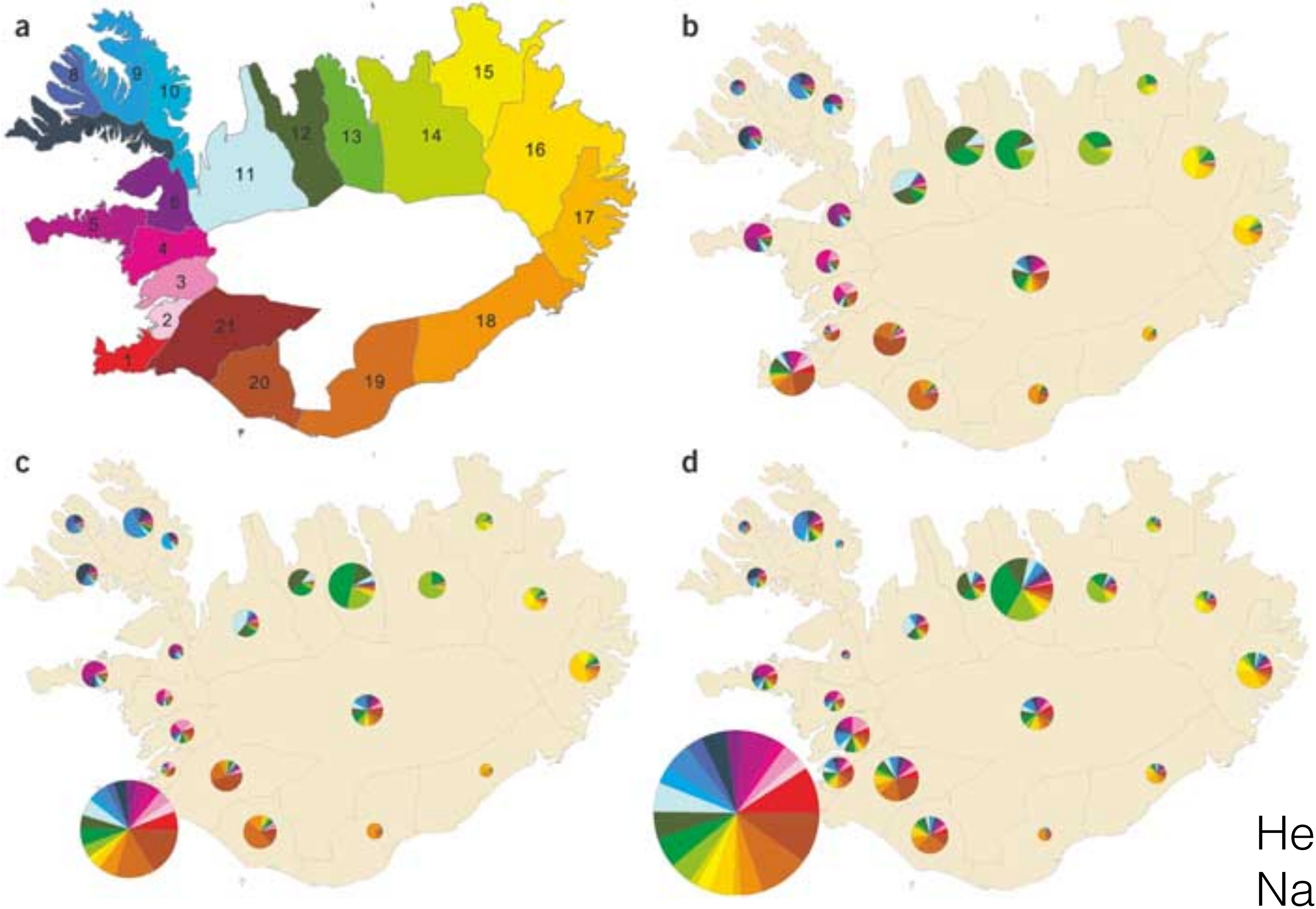
Cryptic Relatedness and fine scale population structure

Learning objectives

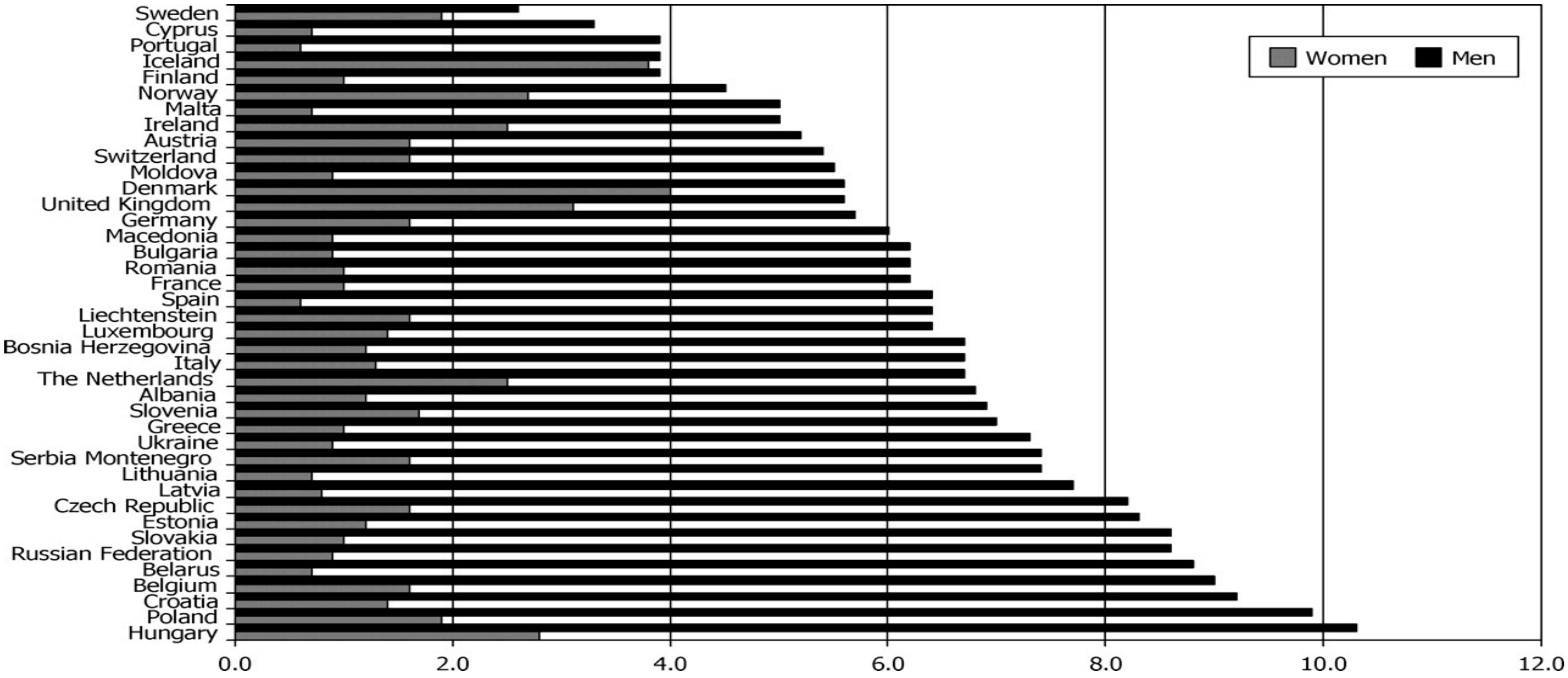
- Define fine scale population structure and cryptic relatedness
- How is it identified
    - Identity-by-descent
    - Rare variation
- Why it can be important for association analyses, especially of rare variants.

Cryptic Population Structure



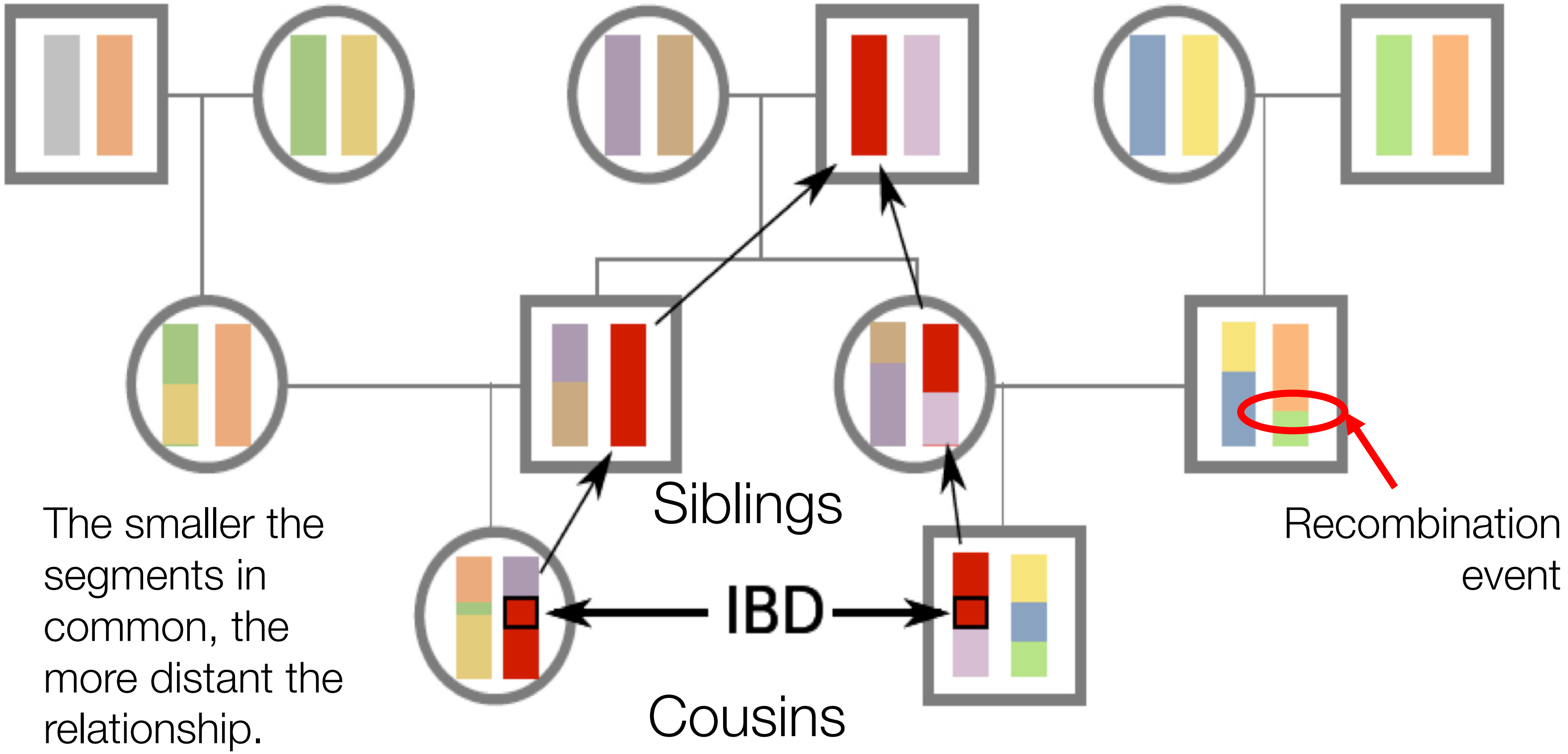
Helgason et al. (2004)
Nature Genet.

Lung Cancer Prevalence in Europe



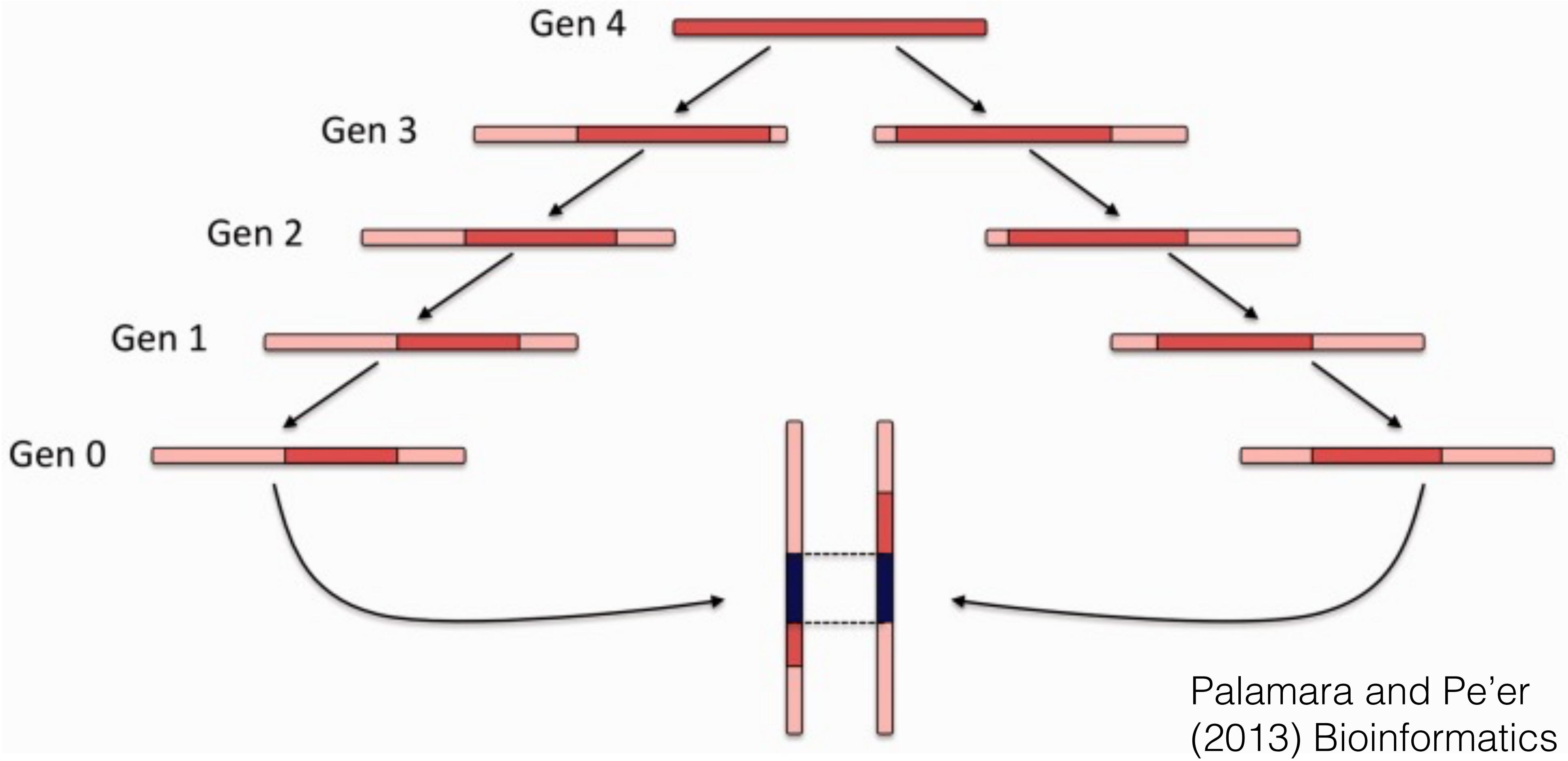
Boyle and Ferlay (2005) Annals of Oncology

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



The smaller the segments in common, the more distant the relationship.

IBD length is correlated with historical relationships.

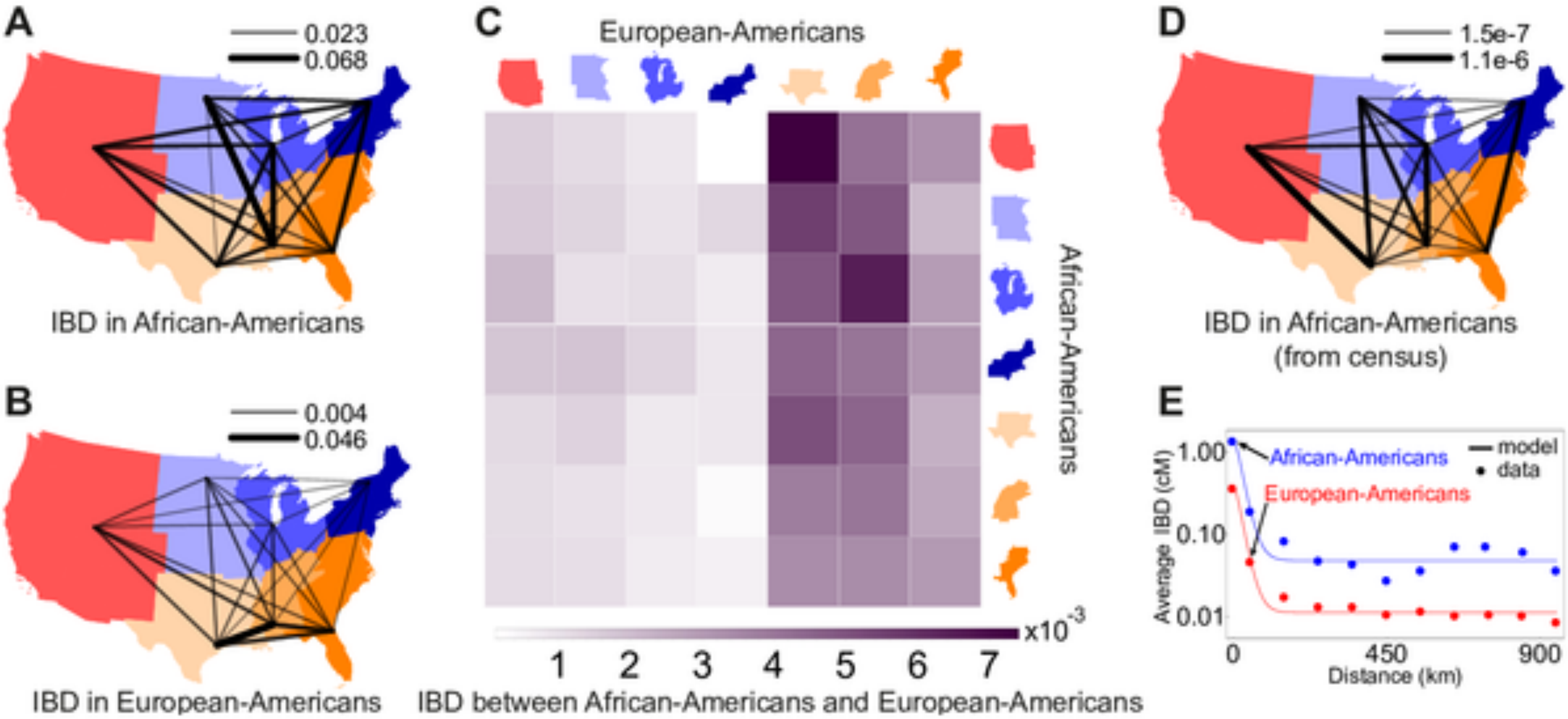


Palamara and Pe’er (2013) Bioinformatics

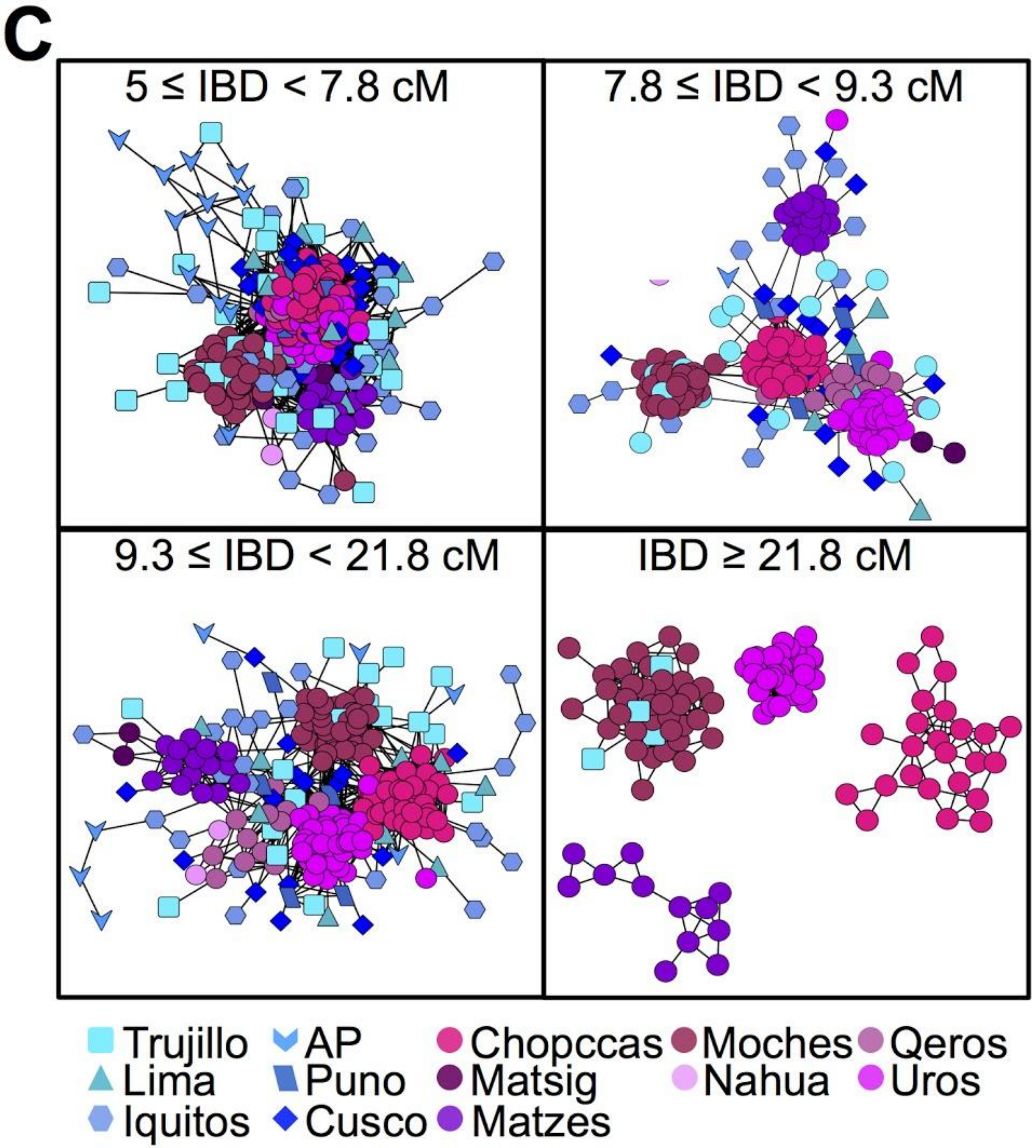
$$E[g|l] \cong \frac{3}{2 * l}$$

Baharian et al. (2016) PLoS Genet.

Pairwise genetic relatedness across

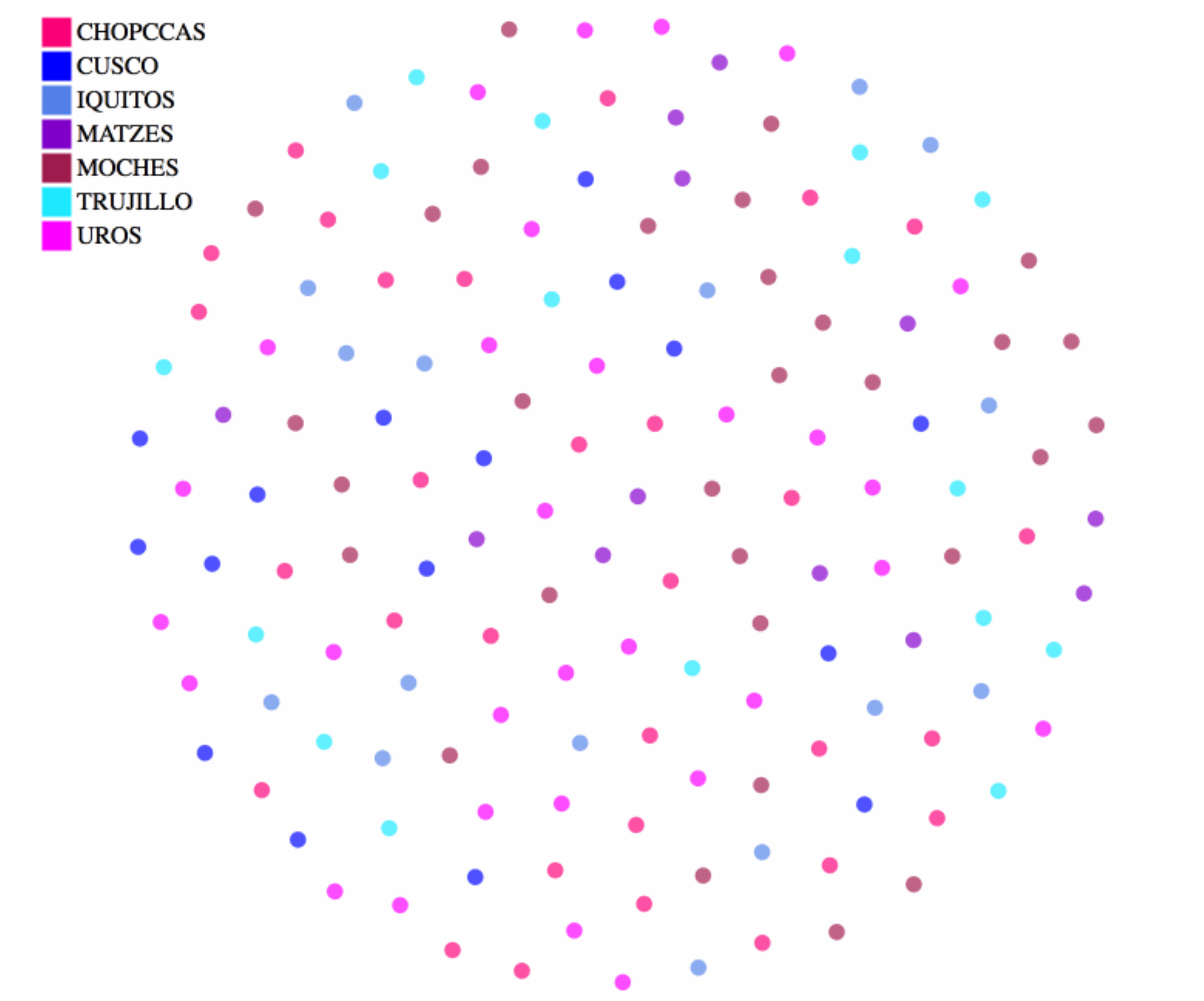


Baharian et al. (2016) PLoS Genet.



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

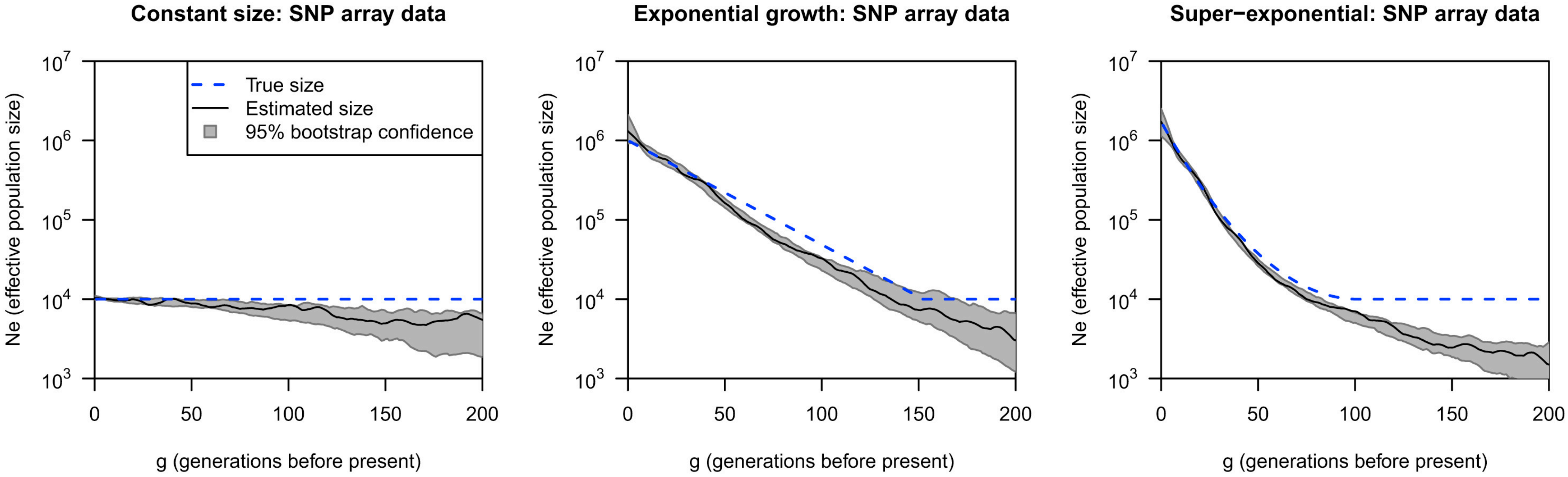
Harris et al. (2018) PNAS



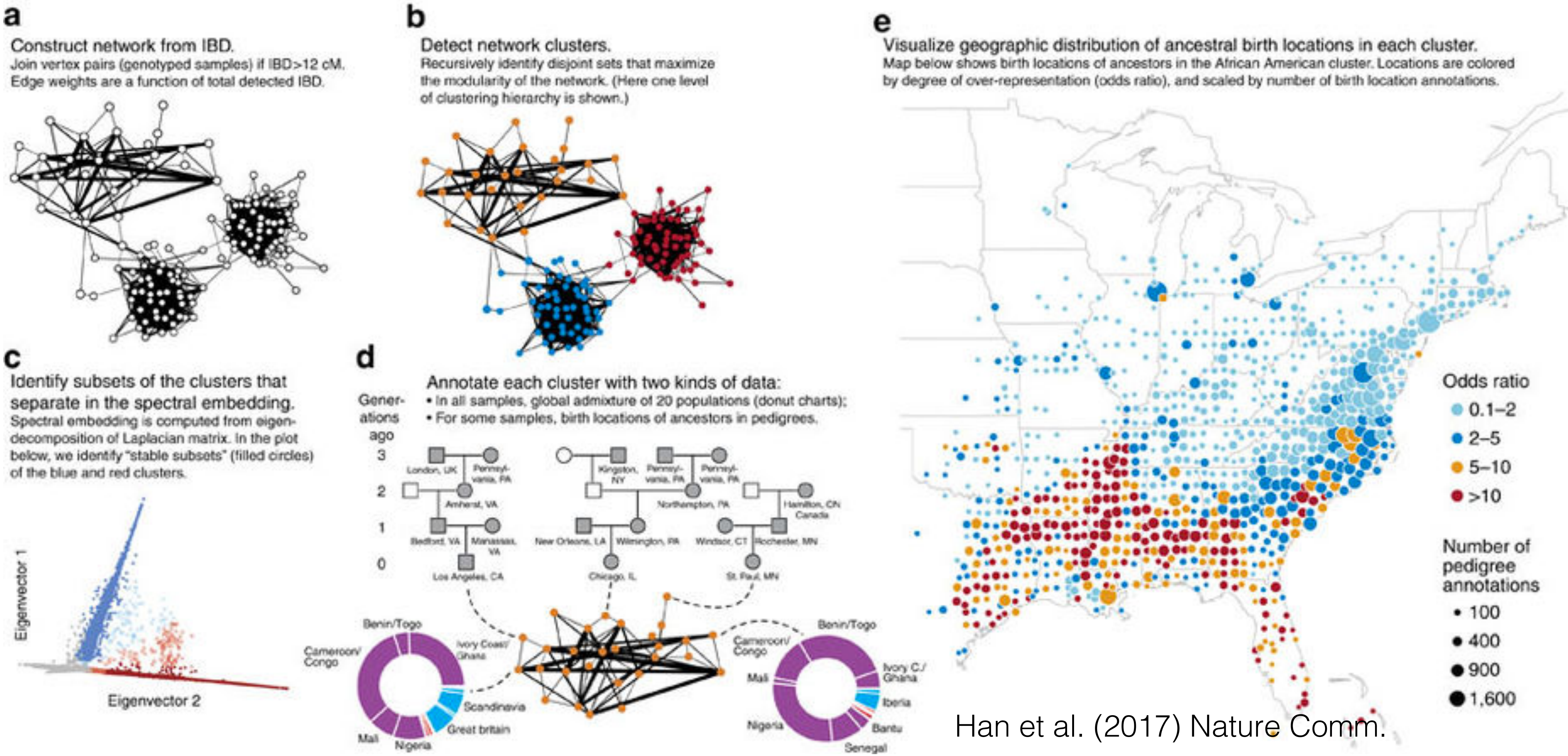
Identity-by-descent as a means to look at fine-scale structure over time

Harris et al. (2018) PNAS

IBD can estimate effective population size over time.



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.)

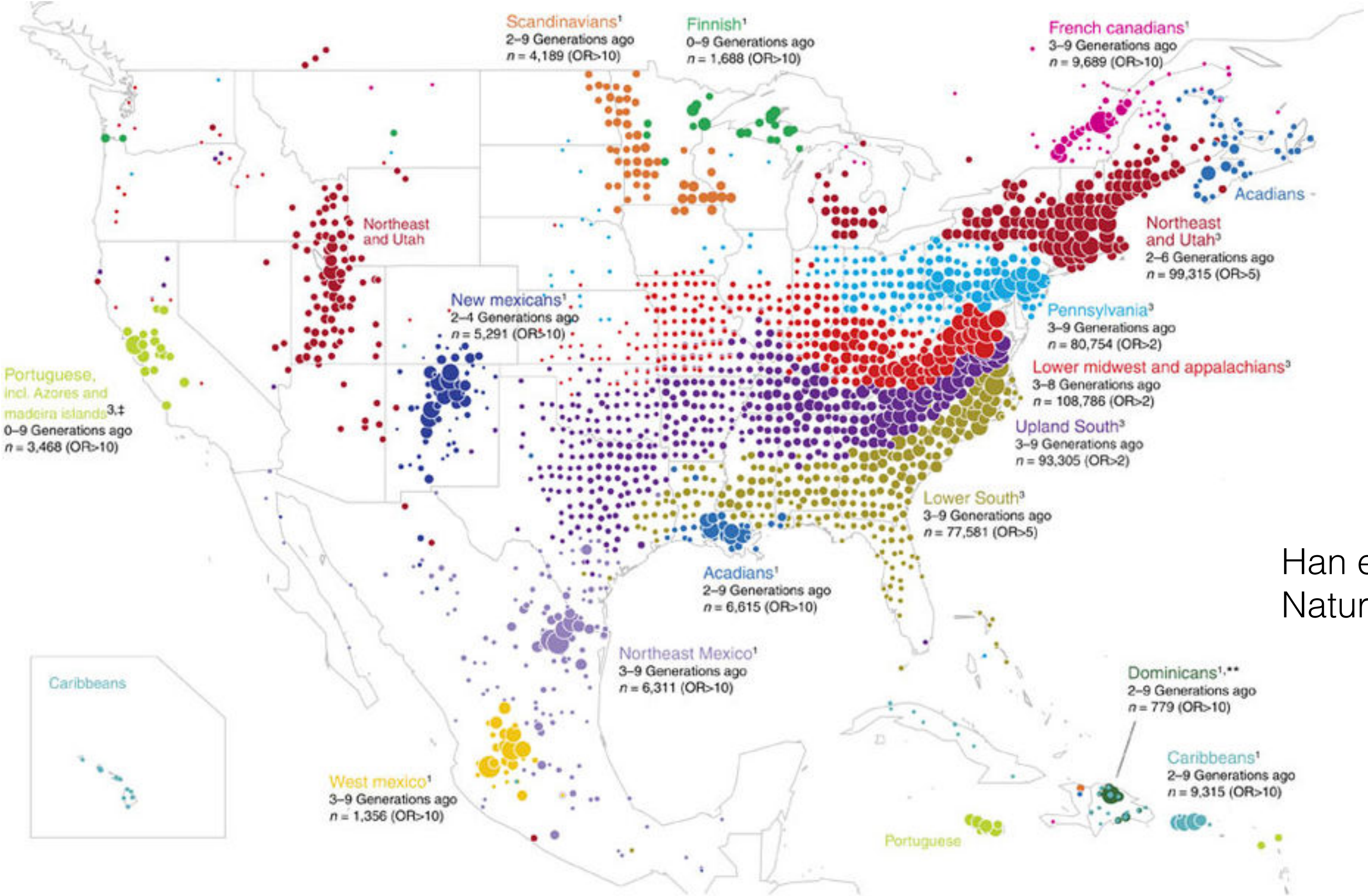
c Identify subsets of the clusters that separate in the spectral embedding.
Spectral embedding is computed from eigen-decomposition of Laplacian matrix. In the plot below, we identify "stable subsets" (filled circles) of the blue and red clusters.

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.

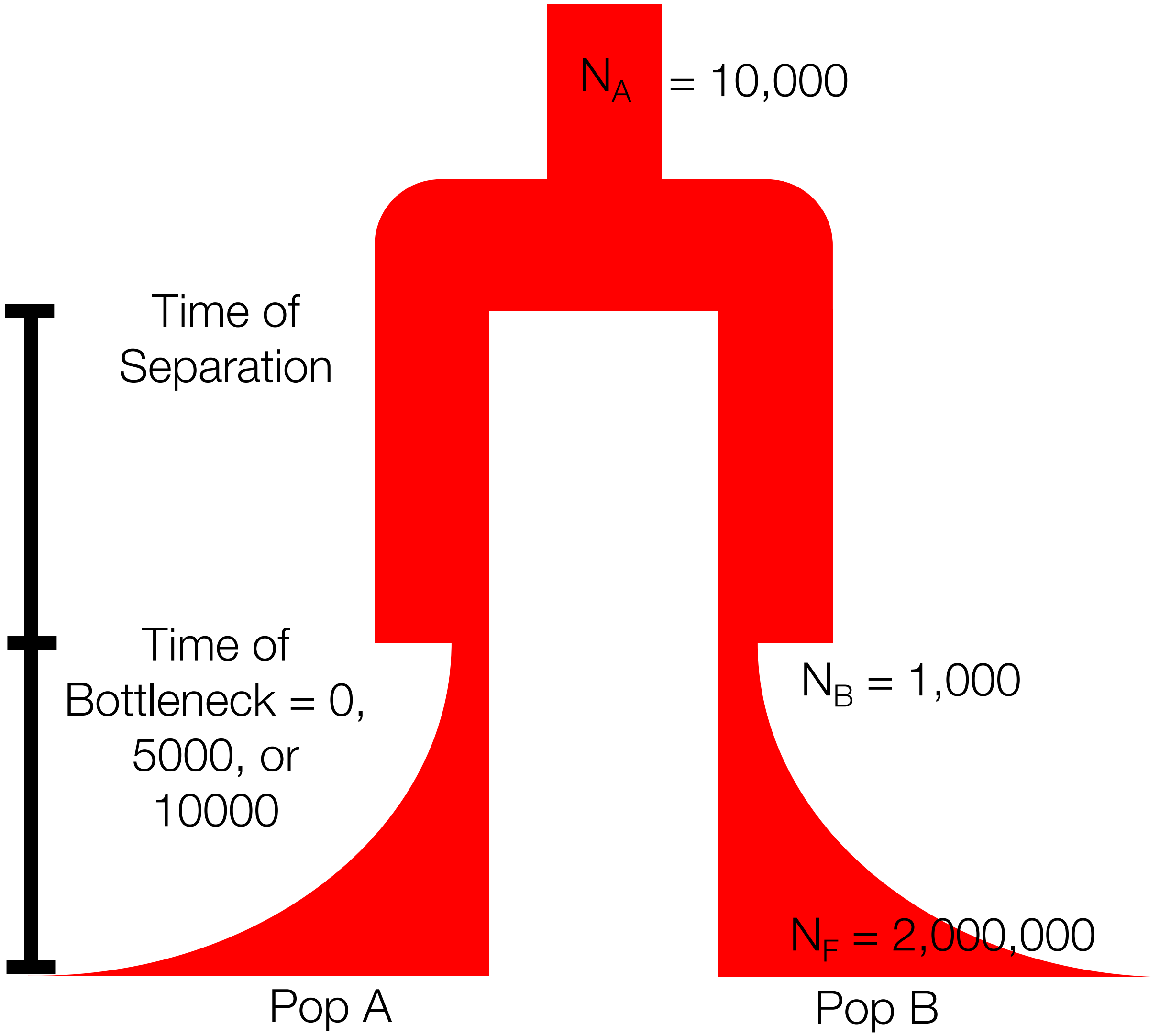
Han et al. (2017) Nature Comm.

IBD on a large scale



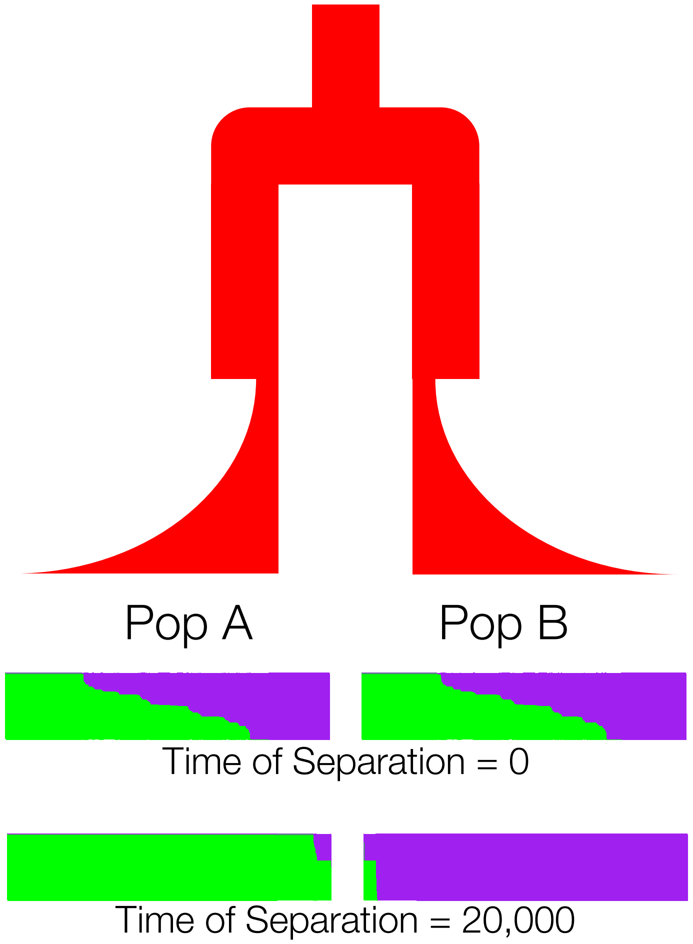
Han et al. (2017)
Nature Comm.

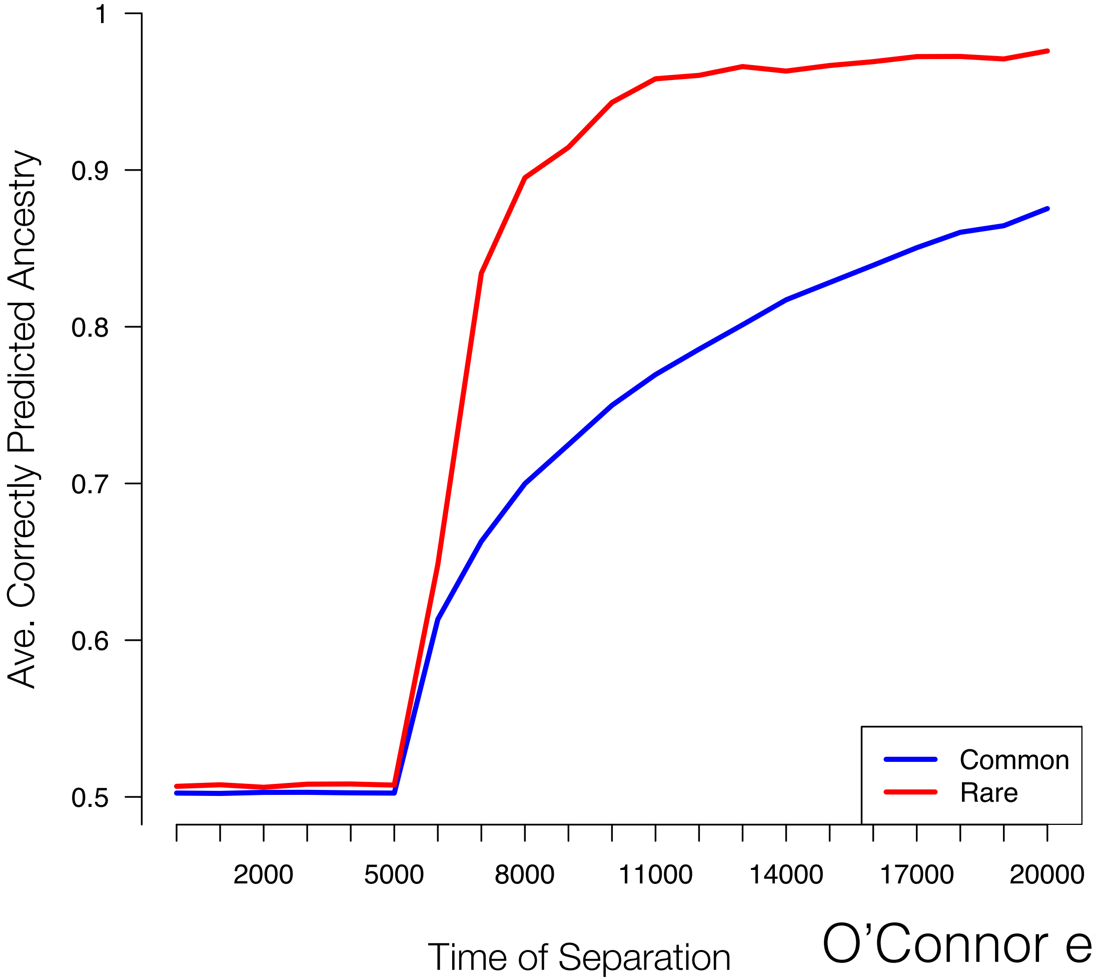
Rare VS Common: Population Structure Simulations



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?

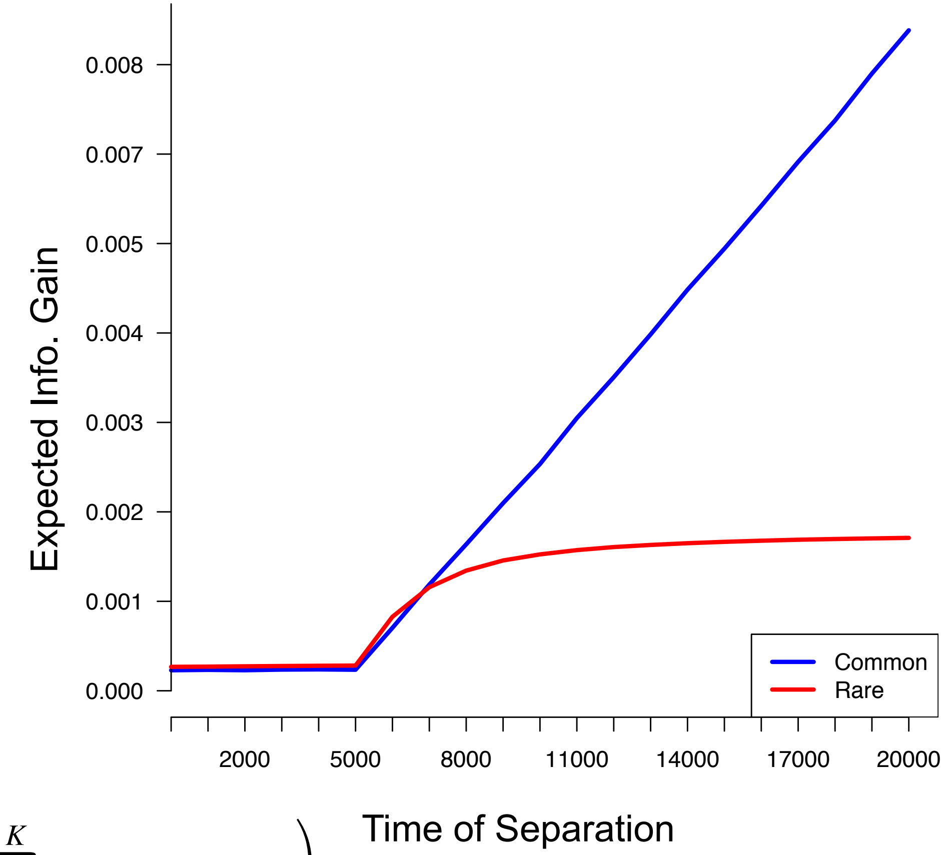
Information Gain: how well a variant can distinguish between populations. (Rosenberg et al. 2003)

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

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)

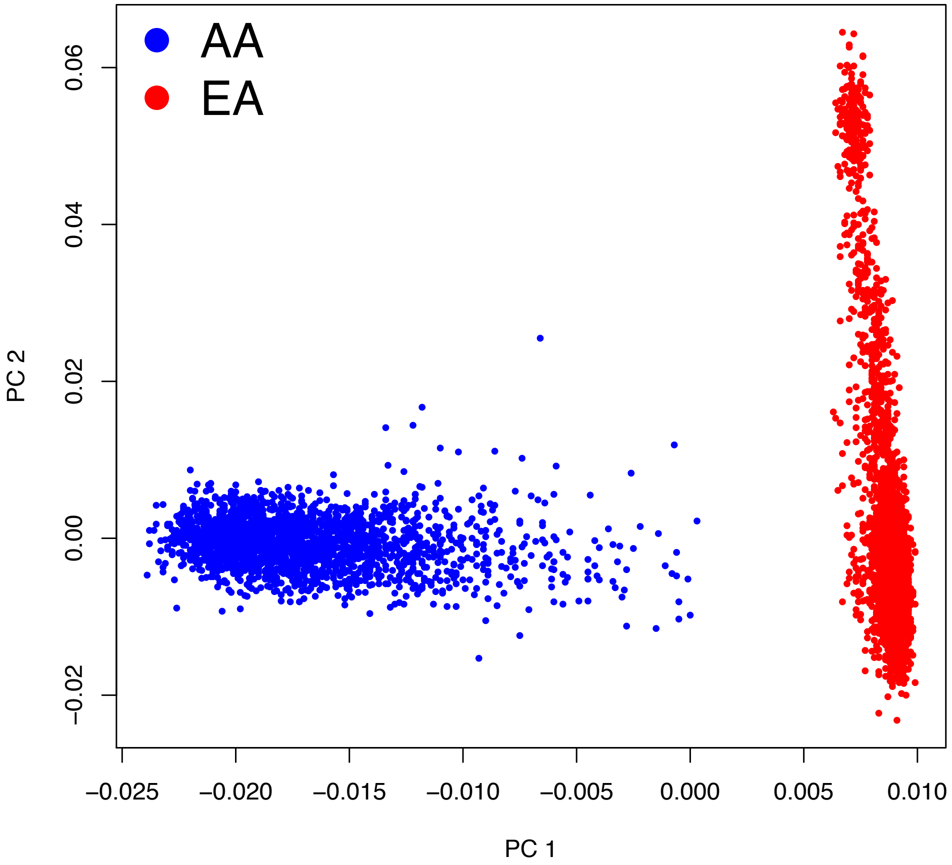
$$E(I_n \mid C, M) = \sum_{m \in M} \sum_{l=0}^{C} r_{lm} \times \sum_{j=1}^{N}\left(-p_{jlm} \ln p_{jlm} + \sum_{i=1}^{K} q_i p_{ijlm} \ln p_{ijlm}\right)$$



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

Rare Variants Identify Cryptic Populations

Common (MAF > 10%)

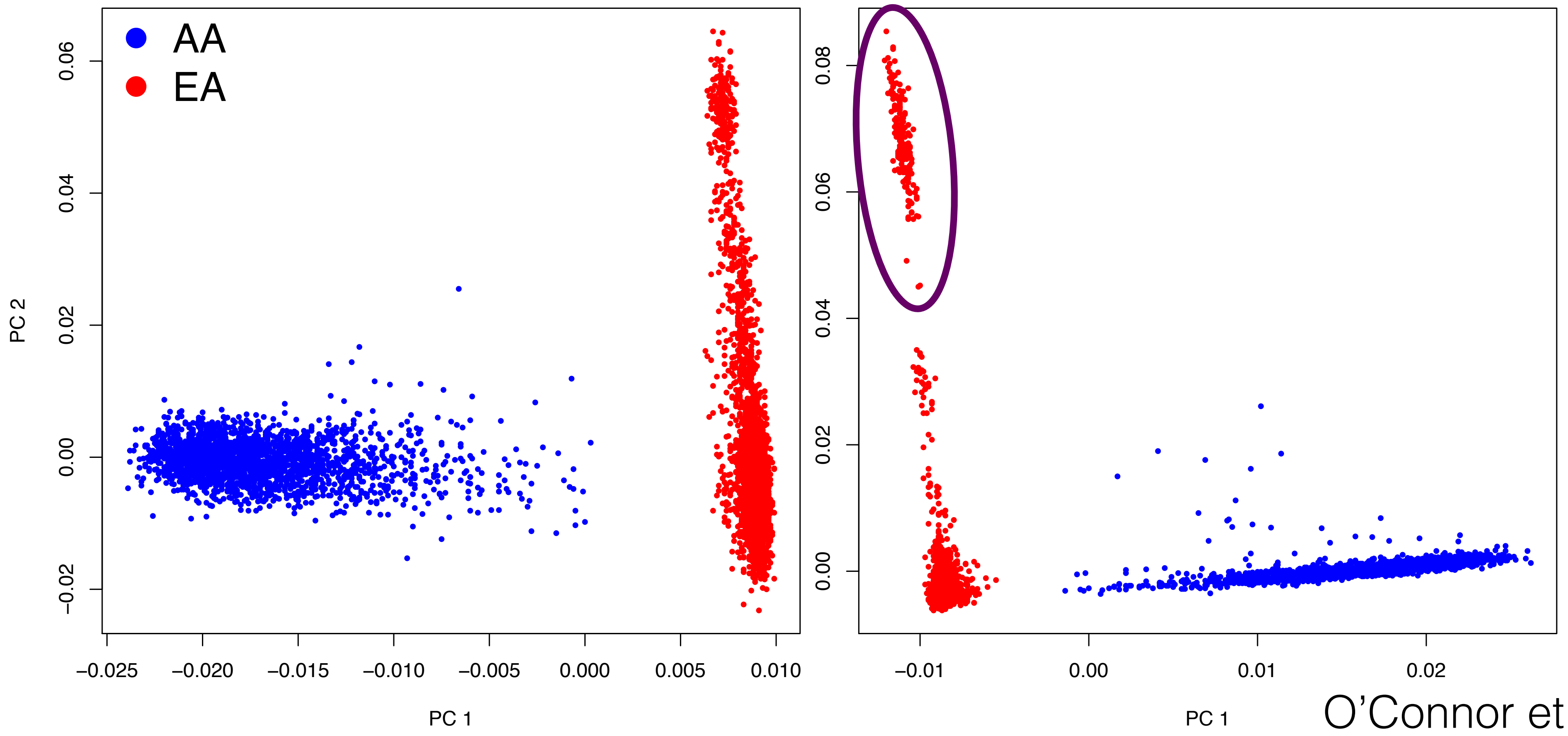


O’Connor et al. (2014)
Mol. Biol. Evol.

Rare Variants Identify Cryptic Populations

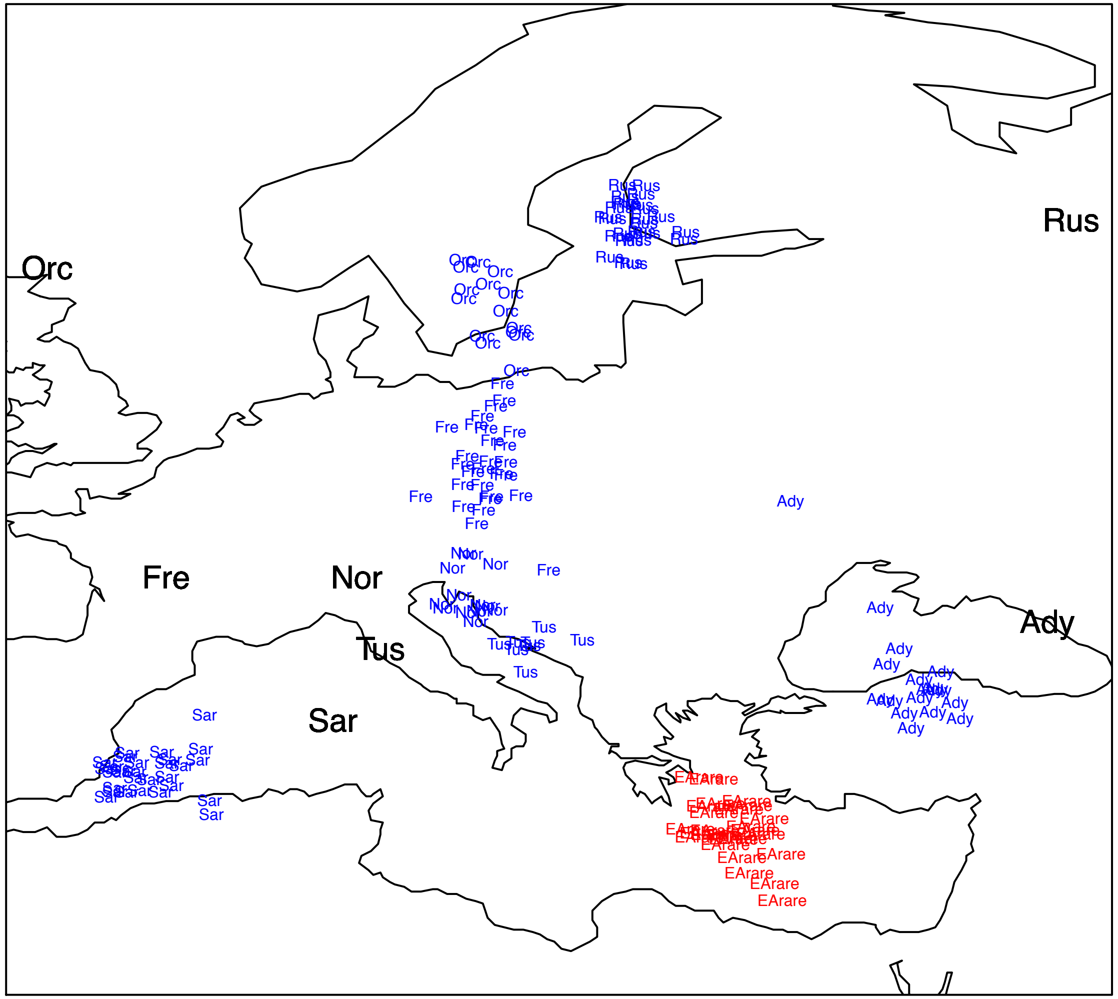
Common (MAF > 10%)

Rare (MAF < 0.5%)



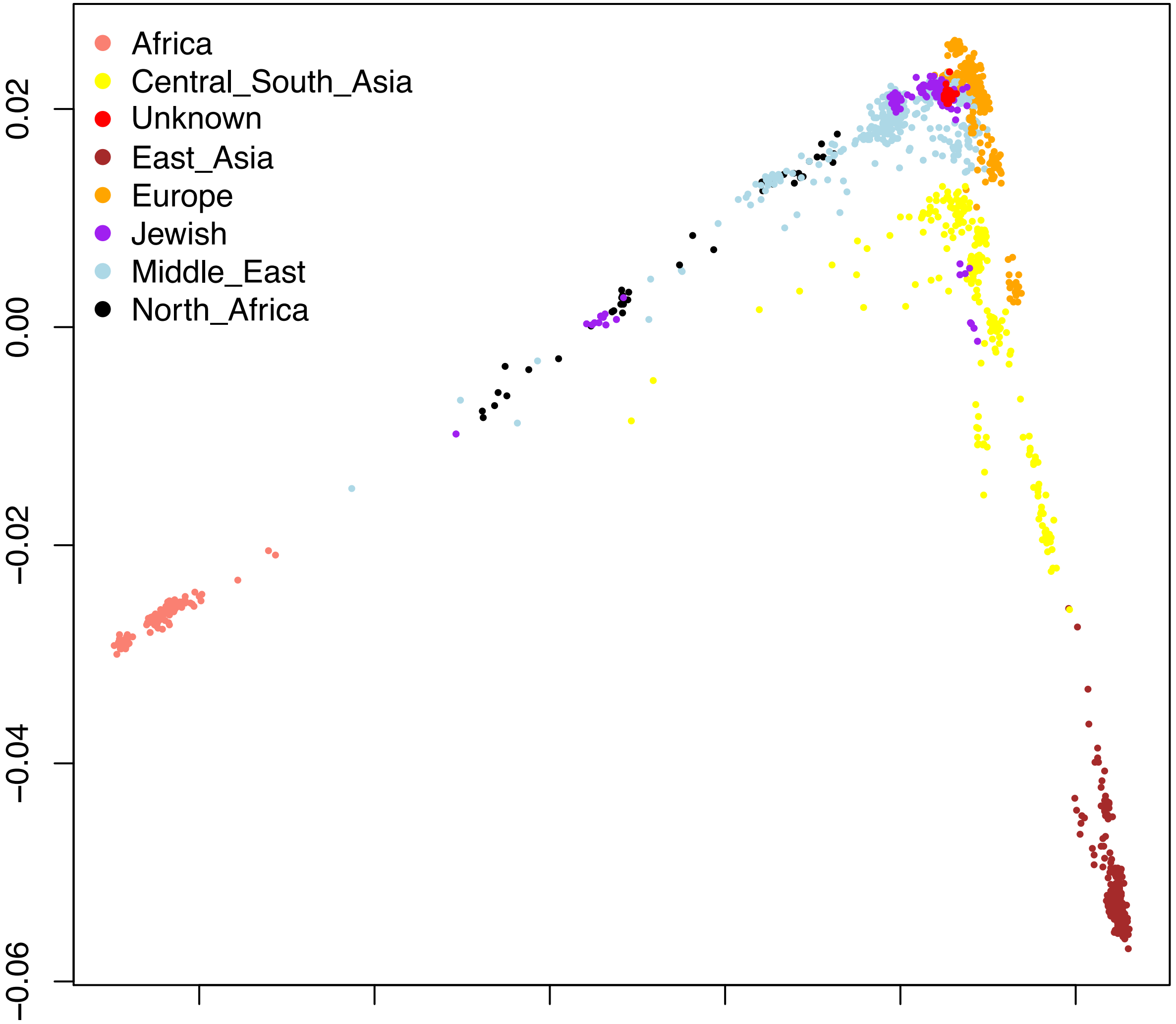
O’Connor et al. (2014)
Mol. Biol. Evol.

What is Their Geographic Ancestry?



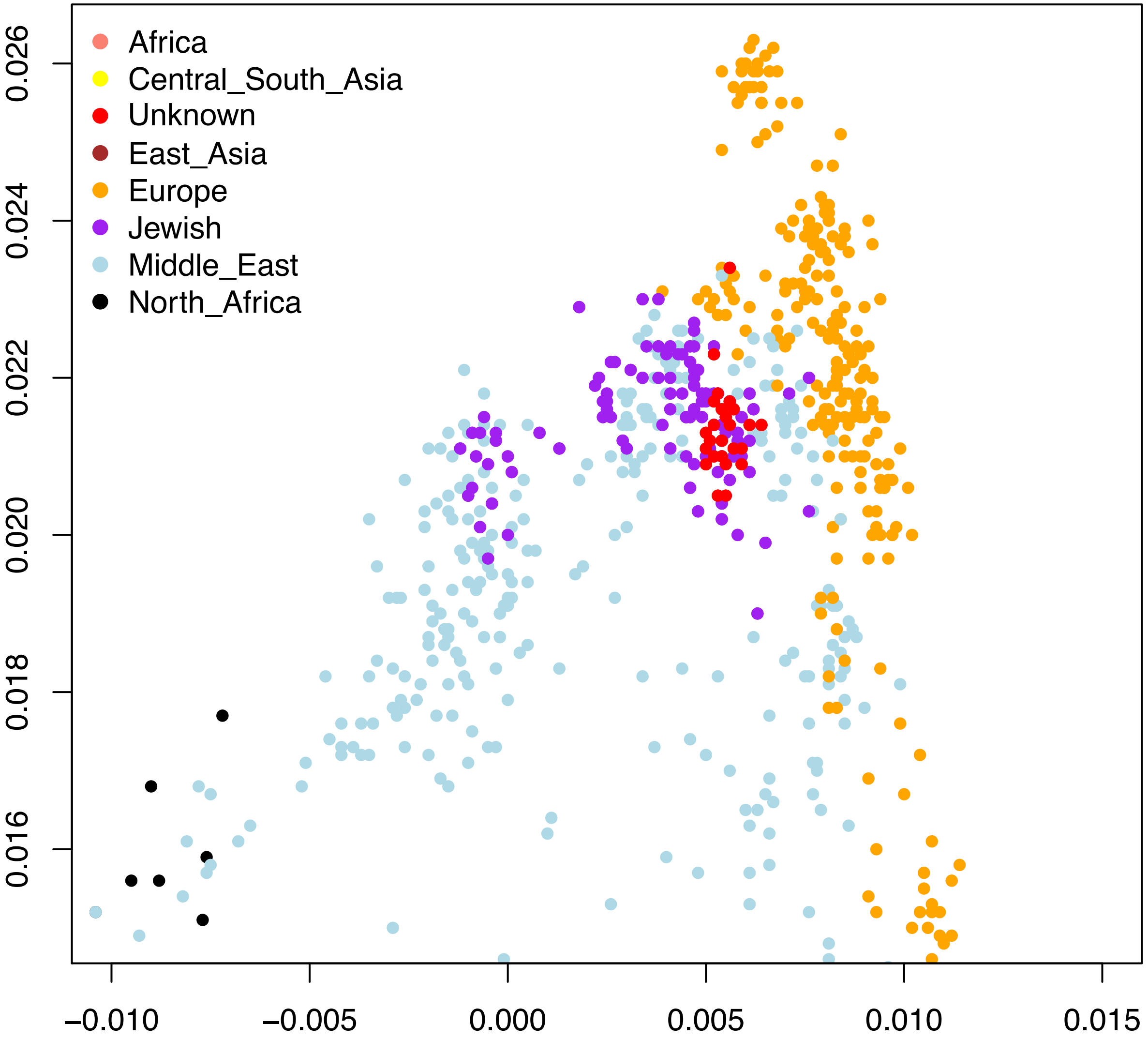
O’Connor et al. (2014)
Mol. Biol. Evol.

PCA of Global Diversity Including Cryptic Population



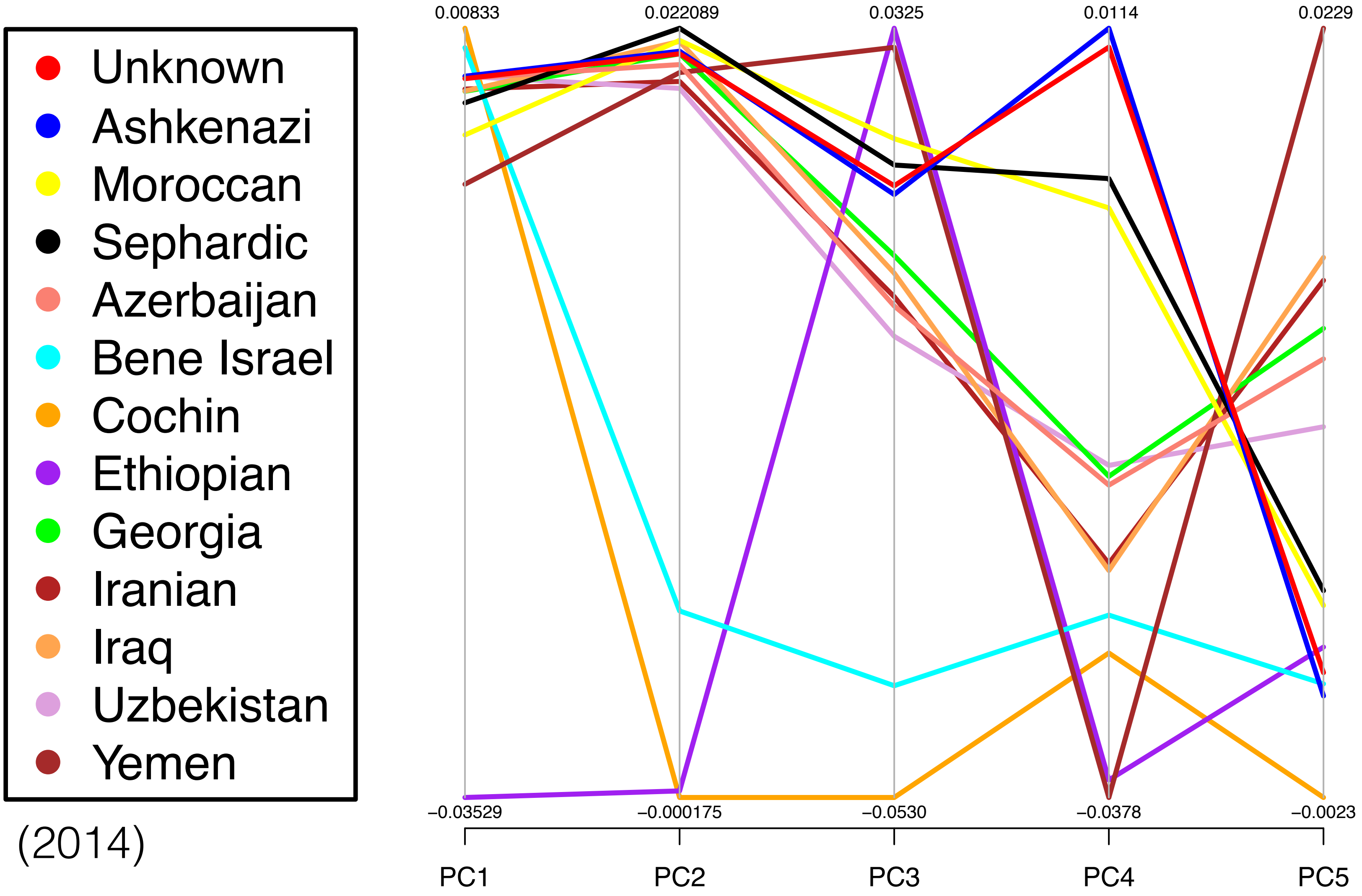
O'Connor et al. (2014)
Mol. Biol. Evol.

PCA of Global Diversity Including Cryptic Population



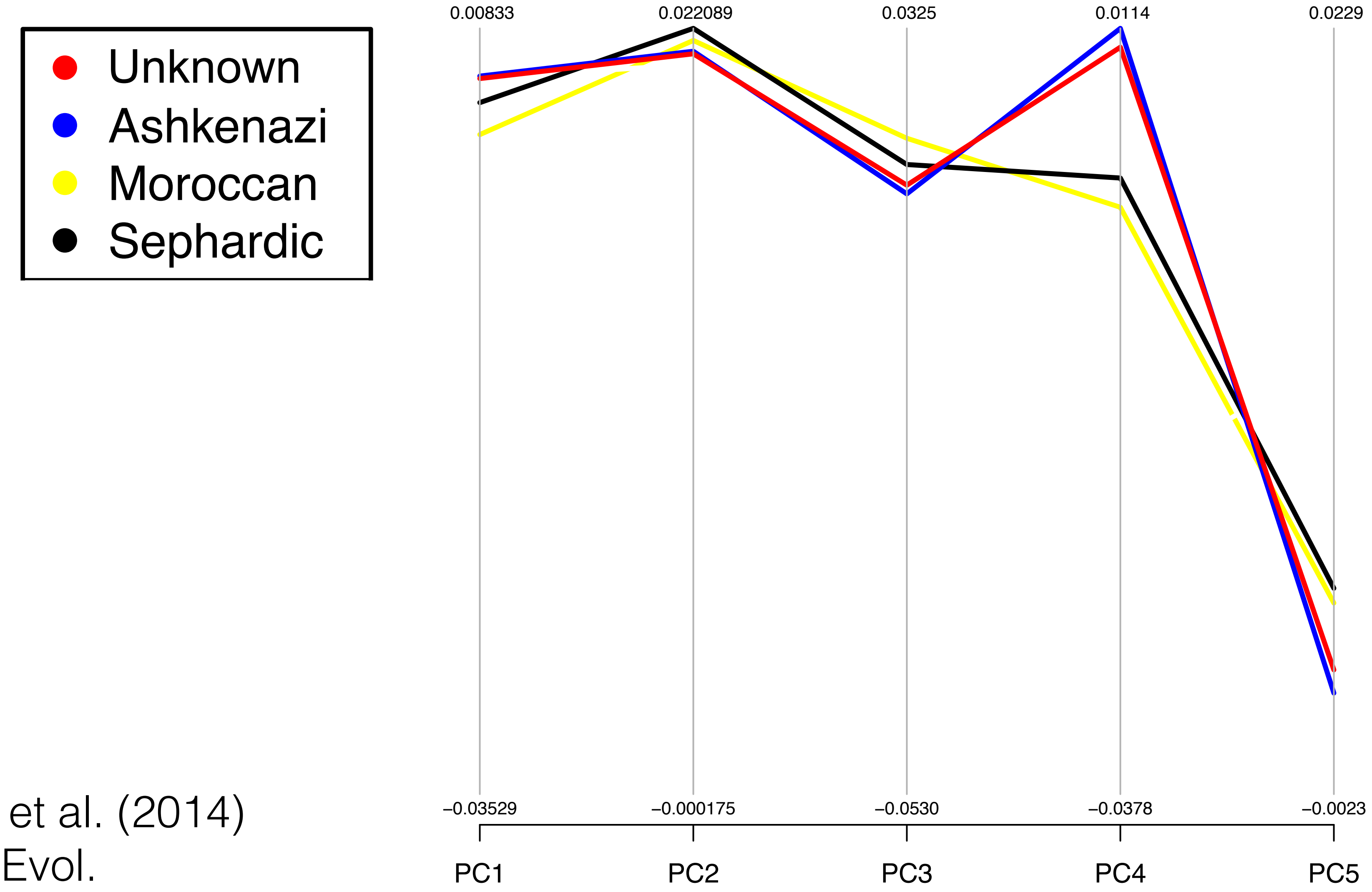
O’Connor et al. (2014)
Mol. Biol. Evol.

Population Average PCA with More Axes



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

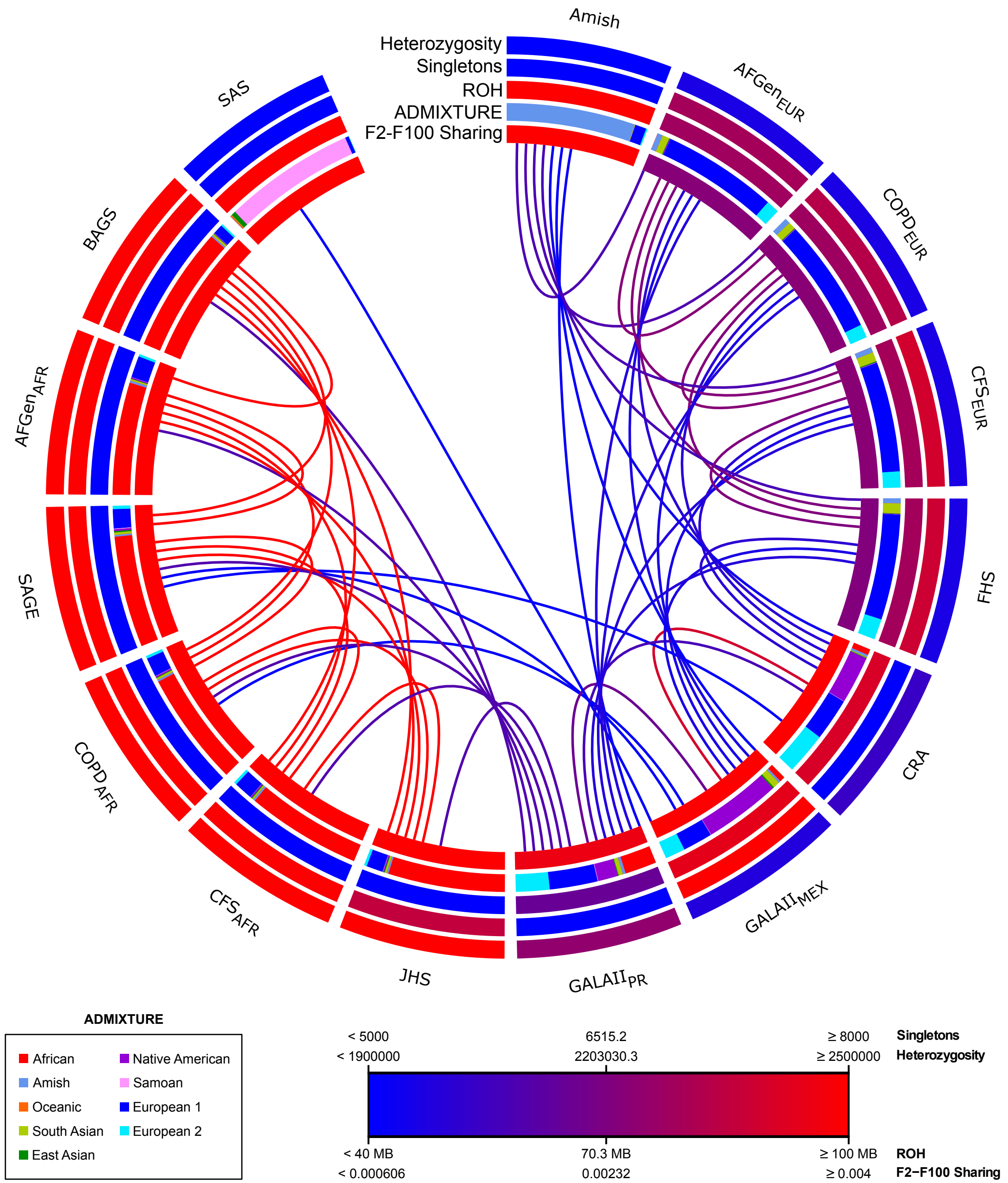
Population Average PCA with More Axes



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

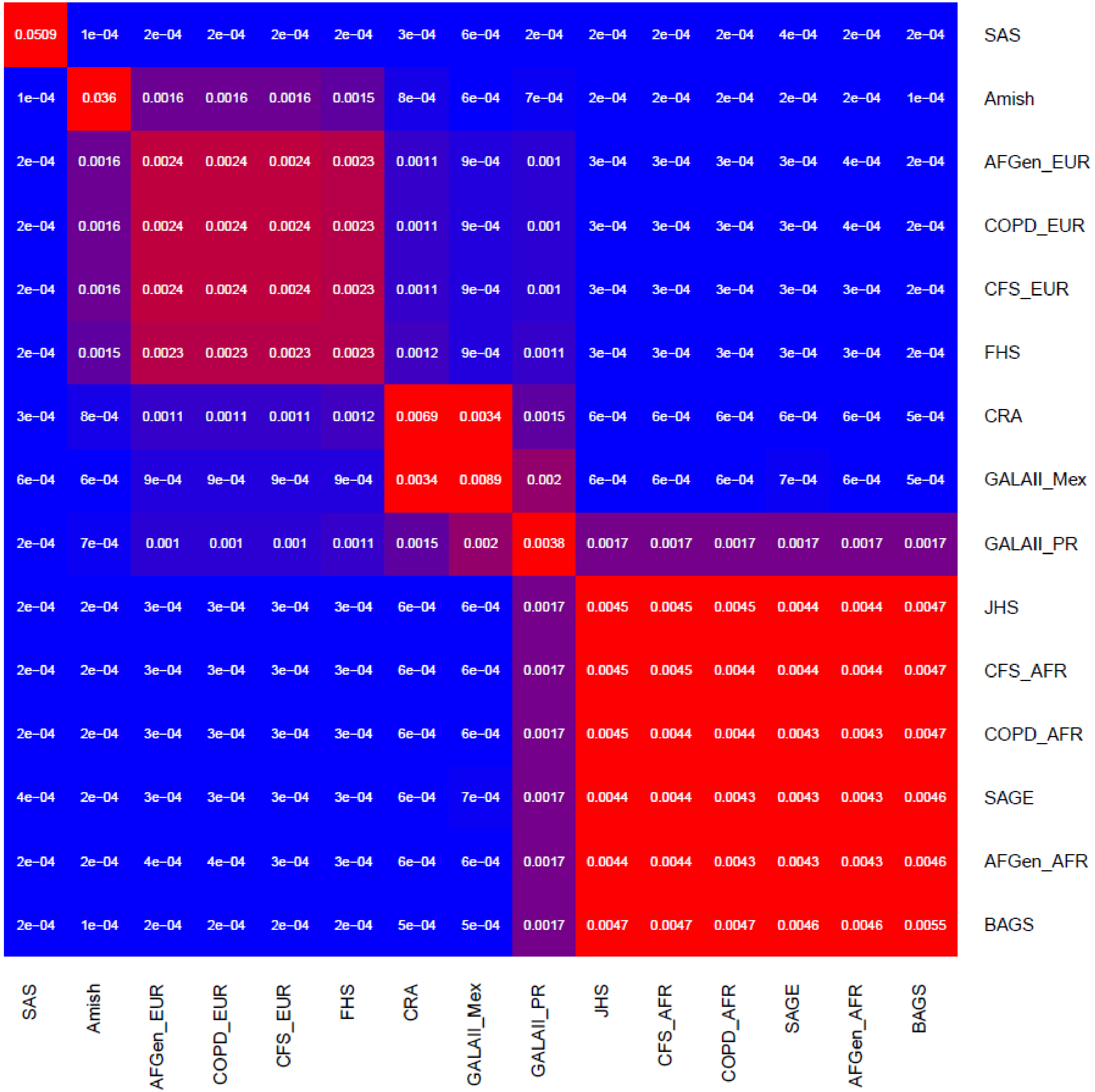
Trans-Omics for Precision Medicine (TOPMed) Cohorts

- N ≅ 18K
- This data freeze has 15 cohorts, each with 100s of samples
- Predominantly African, Latino, and European American
  - Samoa
  - Amish
- All are well characterized for heart, lung, blood, and sleep phenotypes



Rare variant sharing across cohorts

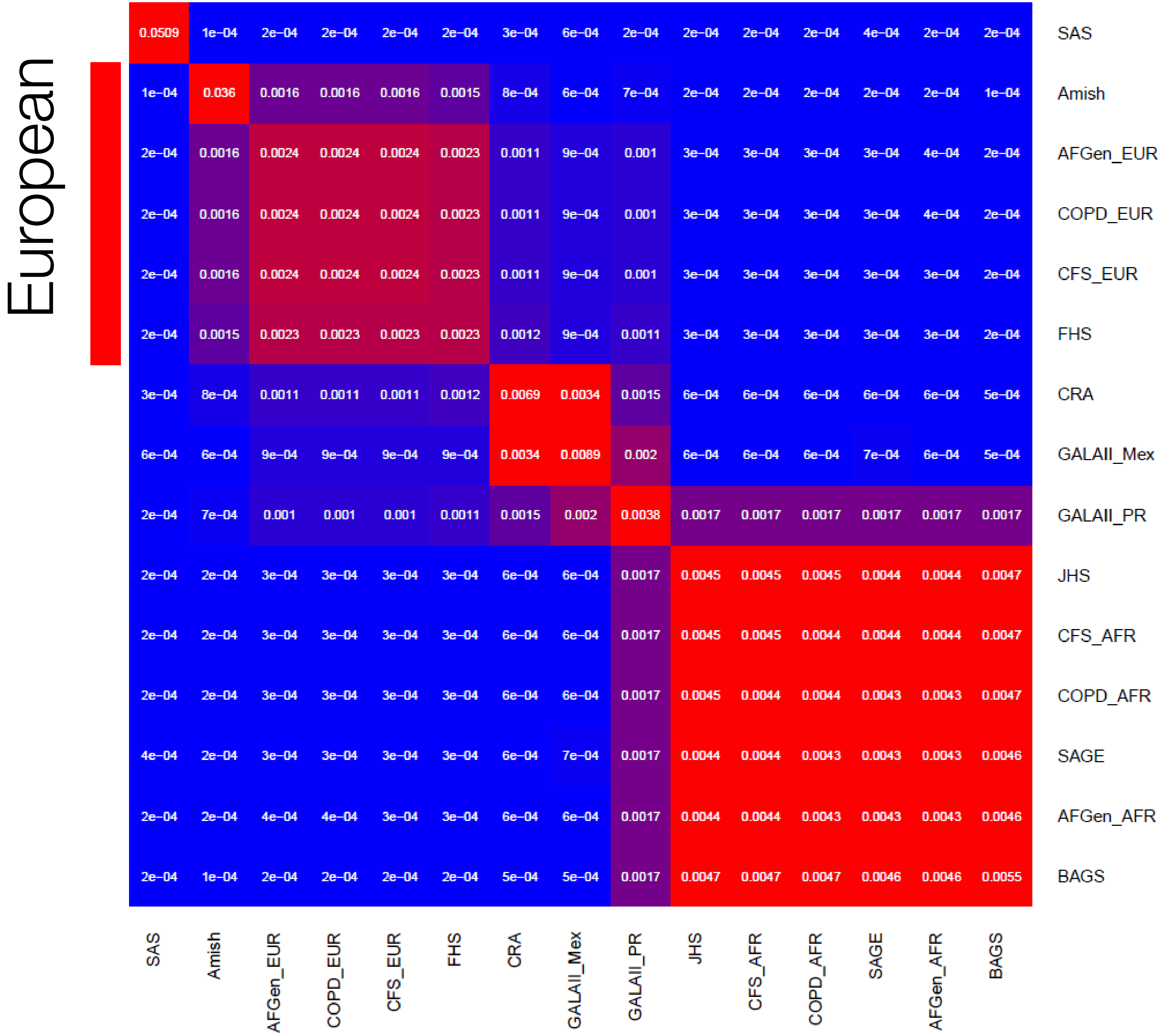
- Allele Count 2 to 100
- Corrected for:
  - sample size
  - Genome-wide heterozygosity



	SAS	Amish	AFGen_EUR	COPD_EUR	CFS_EUR	FHS	CRA	GALAII_Mex	GALAII_PR	JHS	CFS_AFR	COPD_AFR	SAGE	AFGen_AFR	BAGS
SAS	0.0509	1e−04	2e−04	2e−04	2e−04	2e−04	3e−04	6e−04	2e−04	2e−04	2e−04	2e−04	4e−04	2e−04	2e−04
Amish	1e−04	0.036	0.0016	0.0016	0.0016	0.0015	8e−04	6e−04	7e−04	2e−04	2e−04	2e−04	2e−04	2e−04	1e−04
AFGen_EUR	2e−04	0.0016	0.0024	0.0024	0.0024	0.0023	0.0011	9e−04	0.001	3e−04	3e−04	3e−04	3e−04	4e−04	2e−04
COPD_EUR	2e−04	0.0016	0.0024	0.0024	0.0024	0.0023	0.0011	9e−04	0.001	3e−04	3e−04	3e−04	3e−04	4e−04	2e−04
CFS_EUR	2e−04	0.0016	0.0024	0.0024	0.0024	0.0023	0.0011	9e−04	0.001	3e−04	3e−04	3e−04	3e−04	3e−04	2e−04
FHS	2e−04	0.0015	0.0023	0.0023	0.0023	0.0023	0.0012	9e−04	0.0011	3e−04	3e−04	3e−04	3e−04	3e−04	2e−04
CRA	3e−04	8e−04	0.0011	0.0011	0.0011	0.0012	0.0069	0.0034	0.0015	6e−04	6e−04	6e−04	6e−04	6e−04	5e−04
GALAII_Mex	6e−04	6e−04	9e−04	9e−04	9e−04	9e−04	0.0034	0.0089	0.002	6e−04	6e−04	6e−04	7e−04	6e−04	5e−04
GALAII_PR	2e−04	7e−04	0.001	0.001	0.001	0.0011	0.0015	0.002	0.0038	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017
JHS	2e−04	2e−04	3e−04	3e−04	3e−04	3e−04	6e−04	6e−04	0.0017	0.0045	0.0045	0.0045	0.0044	0.0044	0.0047
CFS_AFR	2e−04	2e−04	3e−04	3e−04	3e−04	3e−04	6e−04	6e−04	0.0017	0.0045	0.0045	0.0044	0.0044	0.0044	0.0047
COPD_AFR	2e−04	2e−04	3e−04	3e−04	3e−04	3e−04	6e−04	6e−04	0.0017	0.0045	0.0044	0.0044	0.0043	0.0043	0.0047
SAGE	4e−04	2e−04	3e−04	3e−04	3e−04	3e−04	6e−04	7e−04	0.0017	0.0044	0.0044	0.0043	0.0043	0.0043	0.0046
AFGen_AFR	2e−04	2e−04	4e−04	4e−04	3e−04	3e−04	6e−04	6e−04	0.0017	0.0044	0.0044	0.0043	0.0043	0.0043	0.0046
BAGS	2e−04	1e−04	2e−04	2e−04	2e−04	2e−04	5e−04	5e−04	0.0017	0.0047	0.0047	0.0047	0.0046	0.0046	0.0055

Rare variant sharing across cohorts

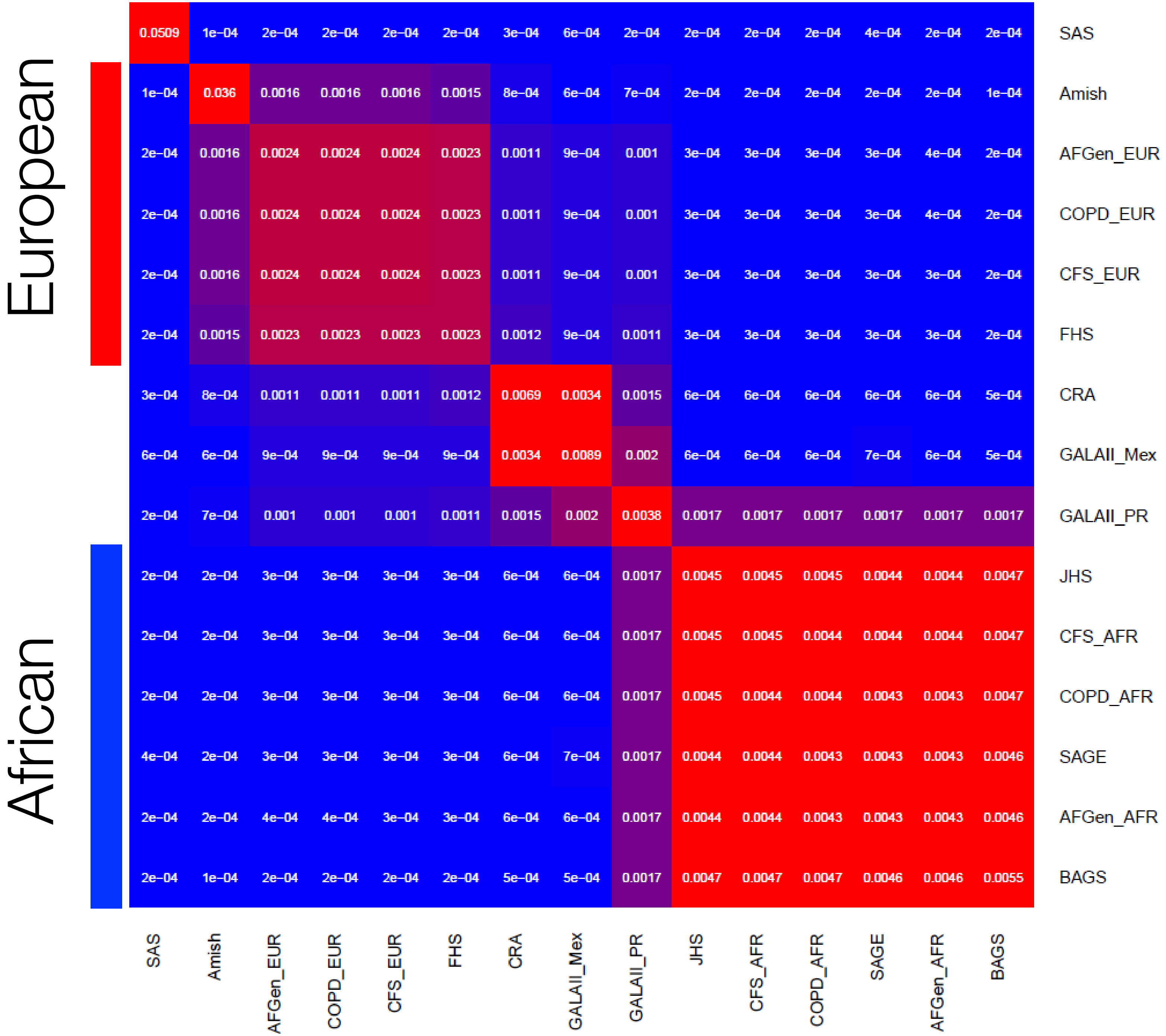
- Allele Count 2 to 100
- Corrected for:
  - sample size
  - Genome-wide heterozygosity



	SAS	Amish	AFGen_EUR	COPD_EUR	CFS_EUR	FHS	CRA	GALAII_Mex	GALAII_PR	JHS	CFS_AFR	COPD_AFR	SAGE	AFGen_AFR	BAGS
SAS	0.0509	1e−04	2e−04	2e−04	2e−04	2e−04	3e−04	6e−04	2e−04	2e−04	2e−04	2e−04	4e−04	2e−04	2e−04
Amish	1e−04	0.036	0.0016	0.0016	0.0016	0.0015	8e−04	6e−04	7e−04	2e−04	2e−04	2e−04	2e−04	2e−04	1e−04
AFGen_EUR	2e−04	0.0016	0.0024	0.0024	0.0024	0.0023	0.0011	9e−04	0.001	3e−04	3e−04	3e−04	3e−04	4e−04	2e−04
COPD_EUR	2e−04	0.0016	0.0024	0.0024	0.0024	0.0023	0.0011	9e−04	0.001	3e−04	3e−04	3e−04	3e−04	4e−04	2e−04
CFS_EUR	2e−04	0.0016	0.0024	0.0024	0.0024	0.0023	0.0011	9e−04	0.001	3e−04	3e−04	3e−04	3e−04	3e−04	2e−04
FHS	2e−04	0.0015	0.0023	0.0023	0.0023	0.0023	0.0012	9e−04	0.0011	3e−04	3e−04	3e−04	3e−04	3e−04	2e−04
CRA	3e−04	8e−04	0.0011	0.0011	0.0011	0.0012	0.0069	0.0034	0.0015	6e−04	6e−04	6e−04	6e−04	6e−04	5e−04
GALAII_Mex	6e−04	6e−04	9e−04	9e−04	9e−04	9e−04	0.0034	0.0089	0.002	6e−04	6e−04	6e−04	7e−04	6e−04	5e−04
GALAII_PR	2e−04	7e−04	0.001	0.001	0.001	0.0011	0.0015	0.002	0.0038	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017
JHS	2e−04	2e−04	3e−04	3e−04	3e−04	3e−04	6e−04	6e−04	0.0017	0.0045	0.0045	0.0045	0.0044	0.0044	0.0047
CFS_AFR	2e−04	2e−04	3e−04	3e−04	3e−04	3e−04	6e−04	6e−04	0.0017	0.0045	0.0045	0.0044	0.0044	0.0044	0.0047
COPD_AFR	2e−04	2e−04	3e−04	3e−04	3e−04	3e−04	6e−04	6e−04	0.0017	0.0045	0.0044	0.0044	0.0043	0.0043	0.0047
SAGE	4e−04	2e−04	3e−04	3e−04	3e−04	3e−04	6e−04	7e−04	0.0017	0.0044	0.0044	0.0043	0.0043	0.0043	0.0046
AFGen_AFR	2e−04	2e−04	4e−04	4e−04	3e−04	3e−04	6e−04	6e−04	0.0017	0.0044	0.0044	0.0043	0.0043	0.0043	0.0046
BAGS	2e−04	1e−04	2e−04	2e−04	2e−04	2e−04	5e−04	5e−04	0.0017	0.0047	0.0047	0.0047	0.0046	0.0046	0.0055

Rare variant sharing across cohorts

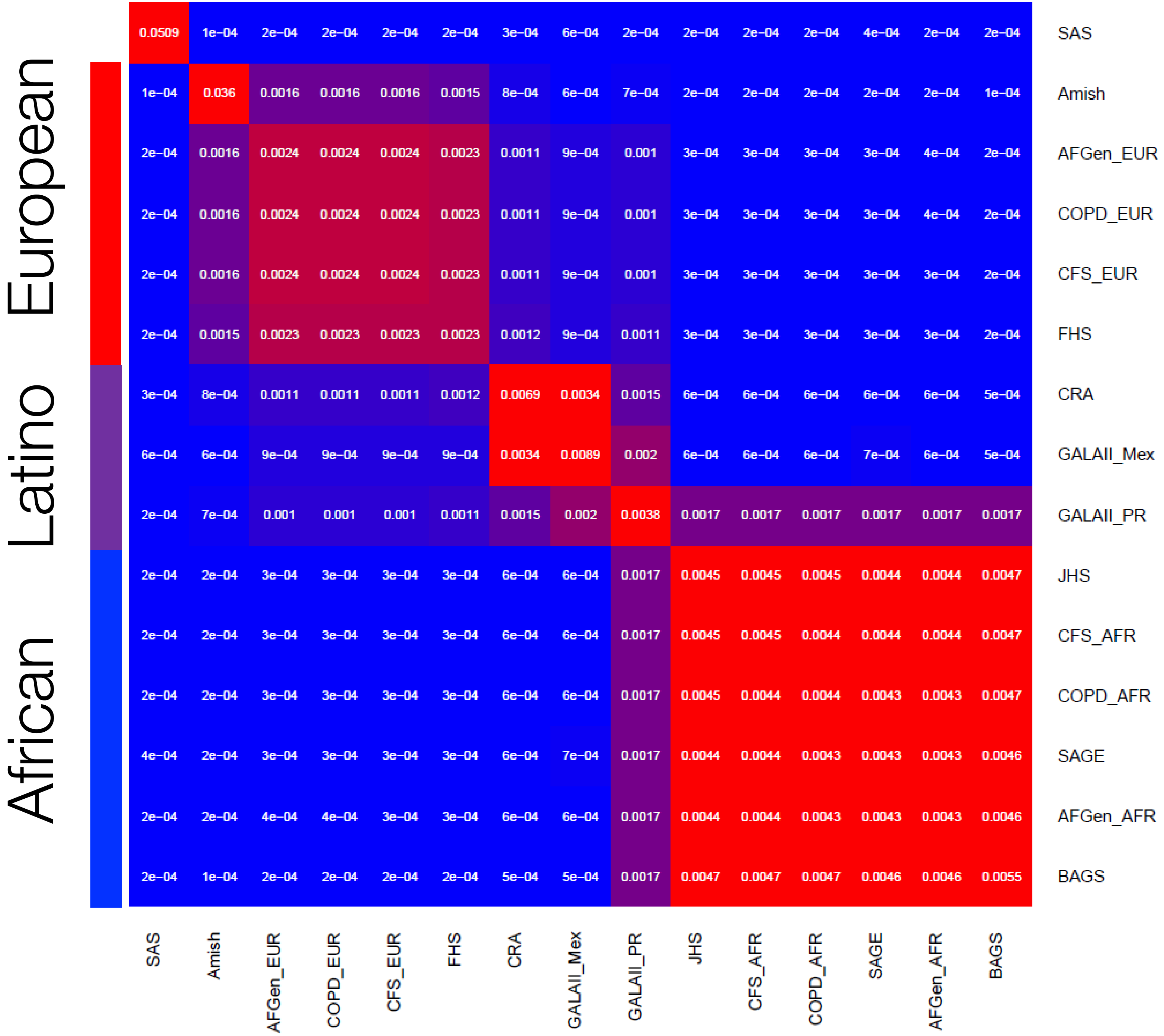
- Allele Count 2 to 100
- Corrected for:
  - sample size
  - Genome-wide heterozygosity



	SAS	Amish	AFGen_EUR	COPD_EUR	CFS_EUR	FHS	CRA	GALAII_Mex	GALAII_PR	JHS	CFS_AFR	COPD_AFR	SAGE	AFGen_AFR	BAGS
SAS	0.0509	1e−04	2e−04	2e−04	2e−04	2e−04	3e−04	6e−04	2e−04	2e−04	2e−04	2e−04	4e−04	2e−04	2e−04
Amish	1e−04	0.036	0.0016	0.0016	0.0016	0.0015	8e−04	6e−04	7e−04	2e−04	2e−04	2e−04	2e−04	2e−04	1e−04
AFGen_EUR	2e−04	0.0016	0.0024	0.0024	0.0024	0.0023	0.0011	9e−04	0.001	3e−04	3e−04	3e−04	3e−04	4e−04	2e−04
COPD_EUR	2e−04	0.0016	0.0024	0.0024	0.0024	0.0023	0.0011	9e−04	0.001	3e−04	3e−04	3e−04	3e−04	4e−04	2e−04
CFS_EUR	2e−04	0.0016	0.0024	0.0024	0.0024	0.0023	0.0011	9e−04	0.001	3e−04	3e−04	3e−04	3e−04	3e−04	2e−04
FHS	2e−04	0.0015	0.0023	0.0023	0.0023	0.0023	0.0012	9e−04	0.0011	3e−04	3e−04	3e−04	3e−04	3e−04	2e−04
CRA	3e−04	8e−04	0.0011	0.0011	0.0011	0.0012	0.0069	0.0034	0.0015	6e−04	6e−04	6e−04	6e−04	6e−04	5e−04
GALAII_Mex	6e−04	6e−04	9e−04	9e−04	9e−04	9e−04	0.0034	0.0089	0.002	6e−04	6e−04	6e−04	7e−04	6e−04	5e−04
GALAII_PR	2e−04	7e−04	0.001	0.001	0.001	0.0011	0.0015	0.002	0.0038	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017
JHS	2e−04	2e−04	3e−04	3e−04	3e−04	3e−04	6e−04	6e−04	0.0017	0.0045	0.0045	0.0045	0.0044	0.0044	0.0047
CFS_AFR	2e−04	2e−04	3e−04	3e−04	3e−04	3e−04	6e−04	6e−04	0.0017	0.0045	0.0045	0.0044	0.0044	0.0044	0.0047
COPD_AFR	2e−04	2e−04	3e−04	3e−04	3e−04	3e−04	6e−04	6e−04	0.0017	0.0045	0.0044	0.0044	0.0043	0.0043	0.0047
SAGE	4e−04	2e−04	3e−04	3e−04	3e−04	3e−04	6e−04	7e−04	0.0017	0.0044	0.0044	0.0043	0.0043	0.0043	0.0046
AFGen_AFR	2e−04	2e−04	4e−04	4e−04	3e−04	3e−04	6e−04	6e−04	0.0017	0.0044	0.0044	0.0043	0.0043	0.0043	0.0046
BAGS	2e−04	1e−04	2e−04	2e−04	2e−04	2e−04	5e−04	5e−04	0.0017	0.0047	0.0047	0.0047	0.0046	0.0046	0.0055

Rare variant sharing across cohorts

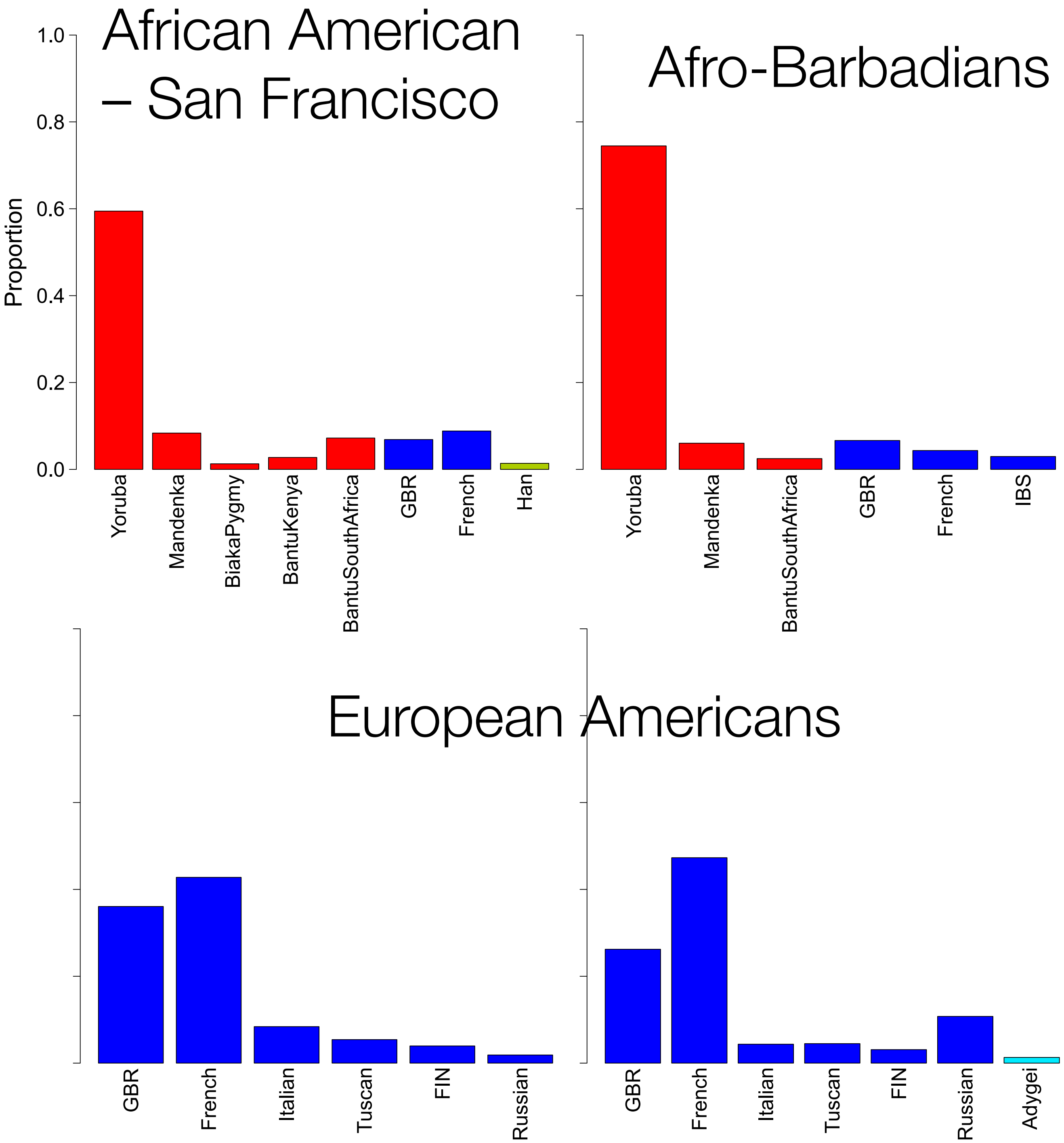
- Allele Count 2 to 100
- Corrected for:
  - sample size
  - Genome-wide heterozygosity



	SAS	Amish	AFGen_EUR	COPD_EUR	CFS_EUR	FHS	CRA	GALAII_Mex	GALAII_PR	JHS	CFS_AFR	COPD_AFR	SAGE	AFGen_AFR	BAGS
SAS	0.0509	1e-04	2e-04	2e-04	2e-04	2e-04	3e-04	6e-04	2e-04	2e-04	2e-04	2e-04	4e-04	2e-04	2e-04
Amish	1e-04	0.036	0.0016	0.0016	0.0016	0.0015	8e-04	6e-04	7e-04	2e-04	2e-04	2e-04	2e-04	2e-04	1e-04
AFGen_EUR	2e-04	0.0016	0.0024	0.0024	0.0024	0.0023	0.0011	9e-04	0.001	3e-04	3e-04	3e-04	3e-04	4e-04	2e-04
COPD_EUR	2e-04	0.0016	0.0024	0.0024	0.0024	0.0023	0.0011	9e-04	0.001	3e-04	3e-04	3e-04	3e-04	4e-04	2e-04
CFS_EUR	2e-04	0.0016	0.0024	0.0024	0.0024	0.0023	0.0011	9e-04	0.001	3e-04	3e-04	3e-04	3e-04	3e-04	2e-04
FHS	2e-04	0.0015	0.0023	0.0023	0.0023	0.0023	0.0012	9e-04	0.0011	3e-04	3e-04	3e-04	3e-04	3e-04	2e-04
CRA	3e-04	8e-04	0.0011	0.0011	0.0011	0.0012	0.0069	0.0034	0.0015	6e-04	6e-04	6e-04	6e-04	6e-04	5e-04
GALAII_Mex	6e-04	6e-04	9e-04	9e-04	9e-04	9e-04	0.0034	0.0089	0.002	6e-04	6e-04	6e-04	7e-04	6e-04	5e-04
GALAII_PR	2e-04	7e-04	0.001	0.001	0.001	0.0011	0.0015	0.002	0.0038	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017
JHS	2e-04	2e-04	3e-04	3e-04	3e-04	3e-04	6e-04	6e-04	0.0017	0.0045	0.0045	0.0045	0.0044	0.0044	0.0047
CFS_AFR	2e-04	2e-04	3e-04	3e-04	3e-04	3e-04	6e-04	6e-04	0.0017	0.0045	0.0045	0.0044	0.0044	0.0044	0.0047
COPD_AFR	2e-04	2e-04	3e-04	3e-04	3e-04	3e-04	6e-04	6e-04	0.0017	0.0045	0.0044	0.0044	0.0043	0.0043	0.0047
SAGE	4e-04	2e-04	3e-04	3e-04	3e-04	3e-04	6e-04	7e-04	0.0017	0.0044	0.0044	0.0043	0.0043	0.0043	0.0046
AFGen_AFR	2e-04	2e-04	4e-04	4e-04	3e-04	3e-04	6e-04	6e-04	0.0017	0.0044	0.0044	0.0043	0.0043	0.0043	0.0046
BAGS	2e-04	1e-04	2e-04	2e-04	2e-04	2e-04	5e-04	5e-04	0.0017	0.0047	0.0047	0.0047	0.0046	0.0046	0.0055

fineStructure analysis of genome-wide ancestry

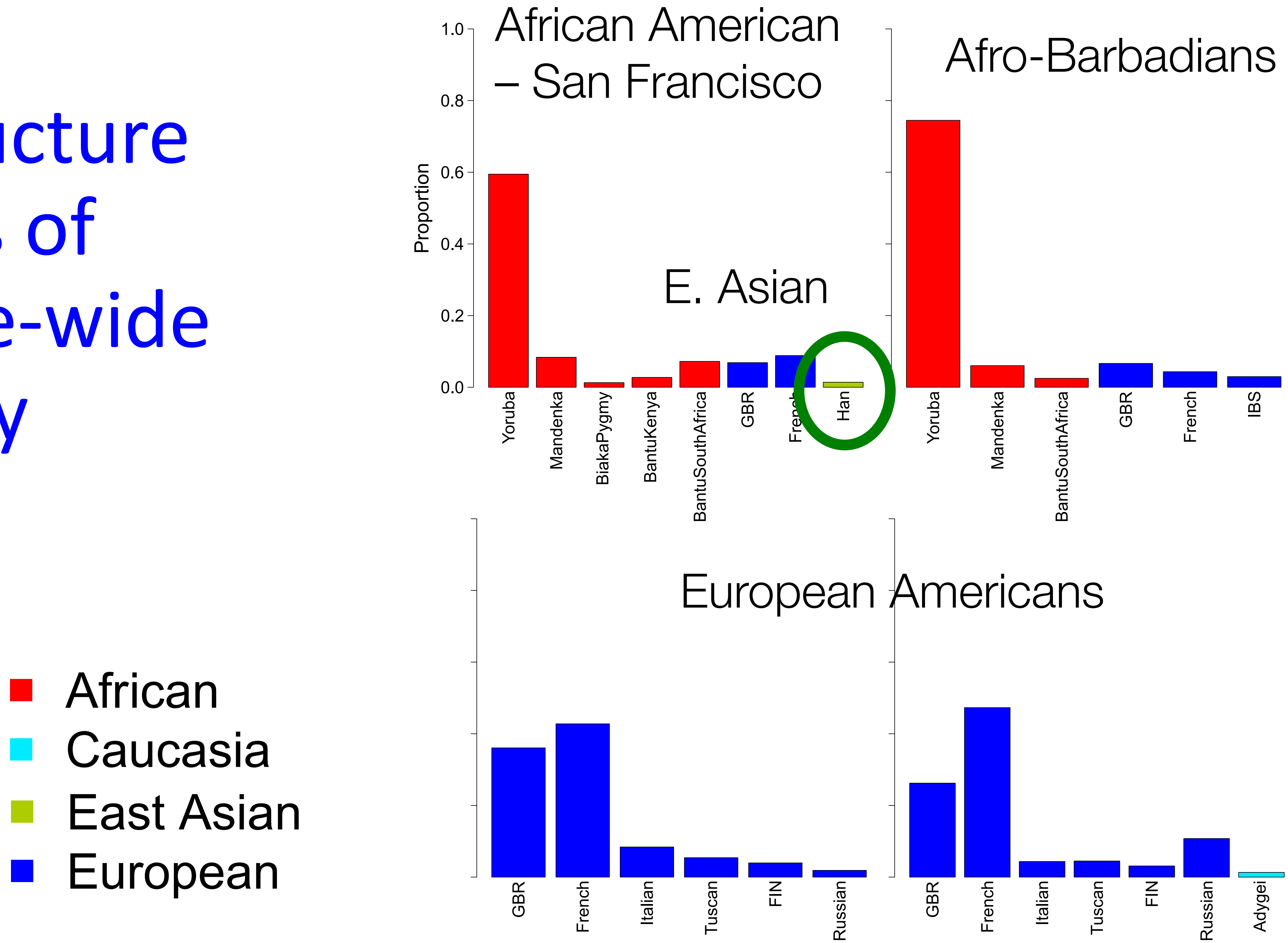
■ African
■ Caucasia
■ East Asian
■ European


African American – San Francisco

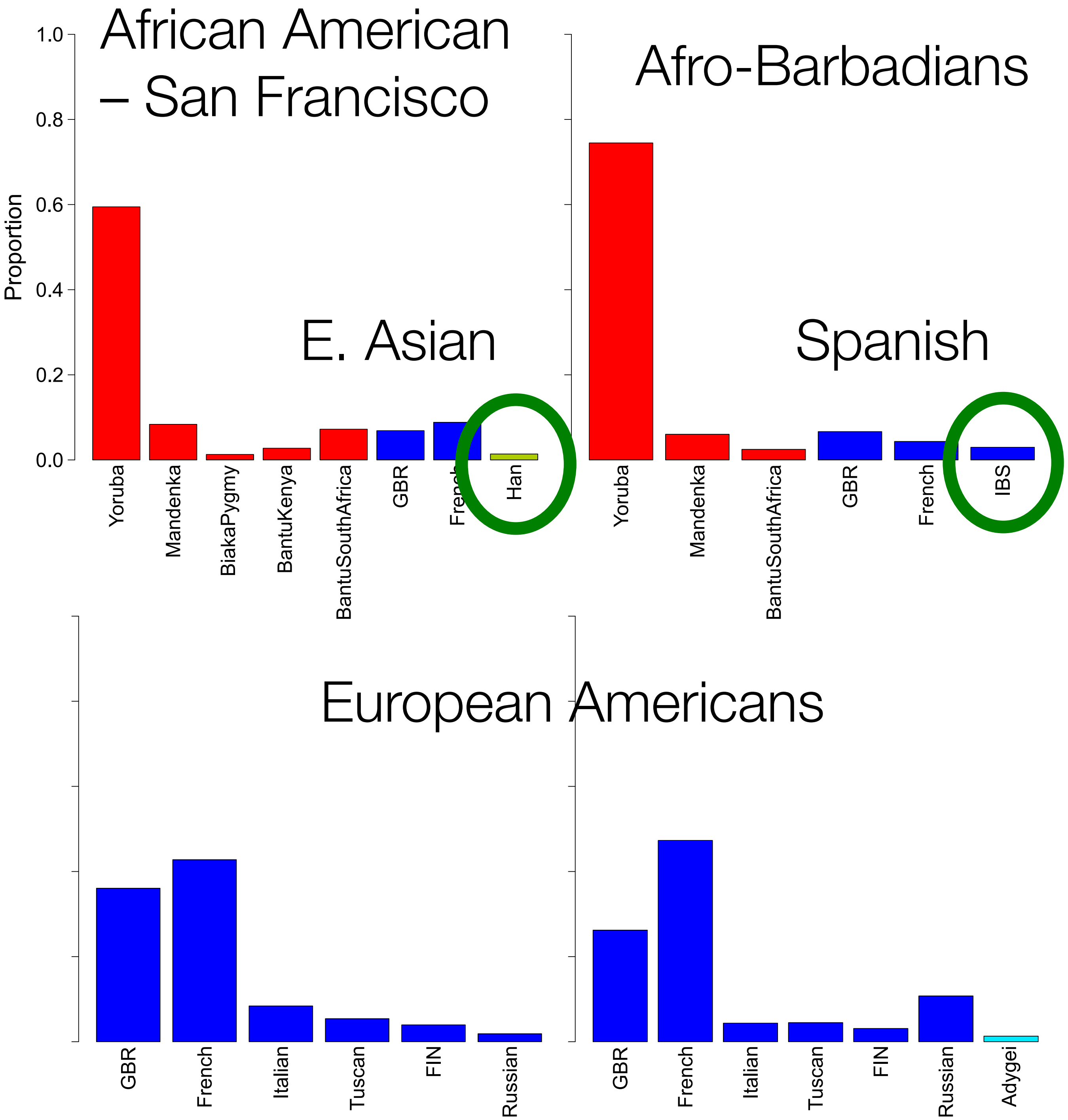
Afro-Barbadians

European Americans

fineStructure analysis of genome-wide ancestry



fineStructure analysis of genome-wide ancestry



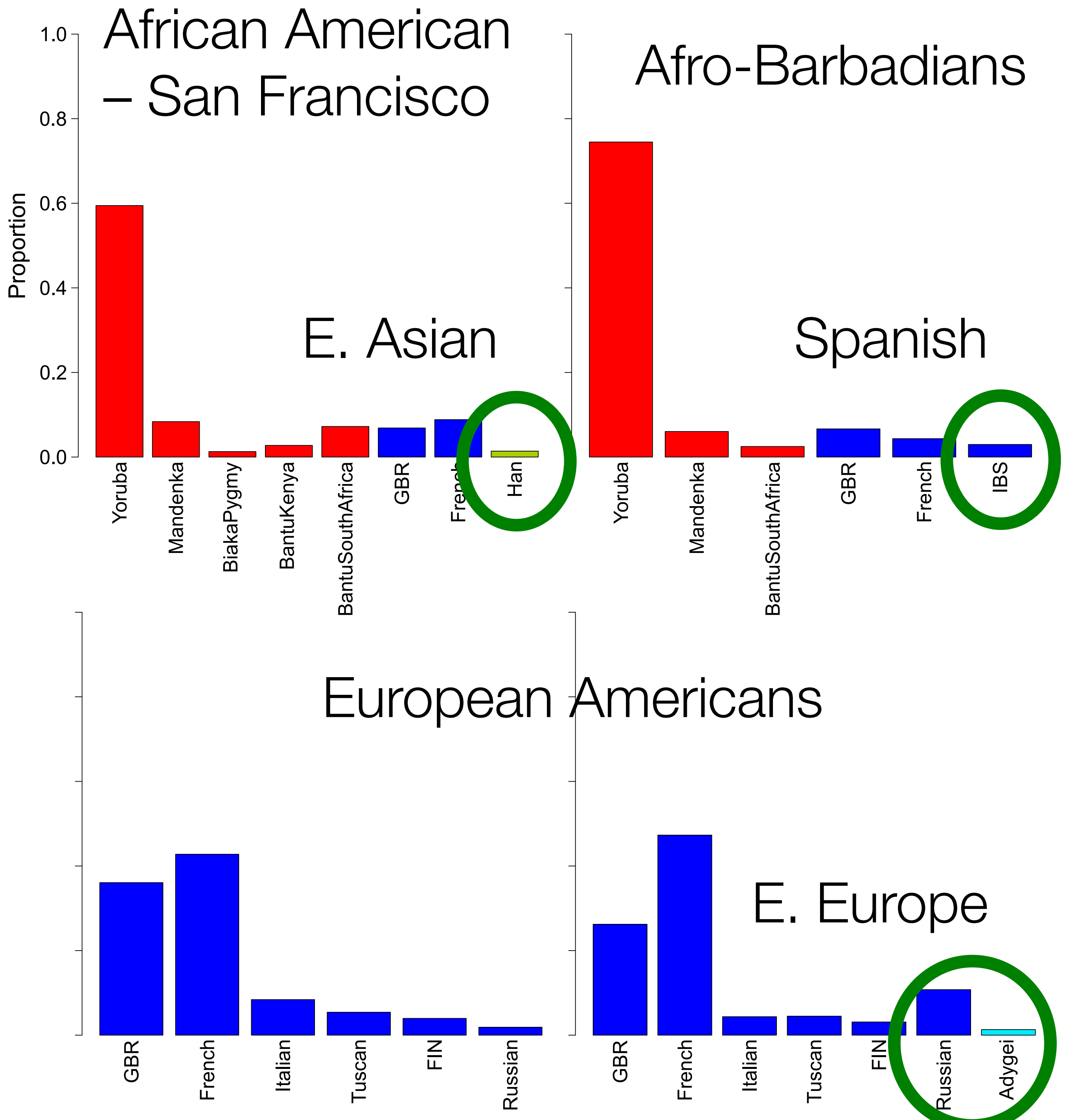
🟥 African

🟦 Caucasia

🟩 East Asian

🟦 European

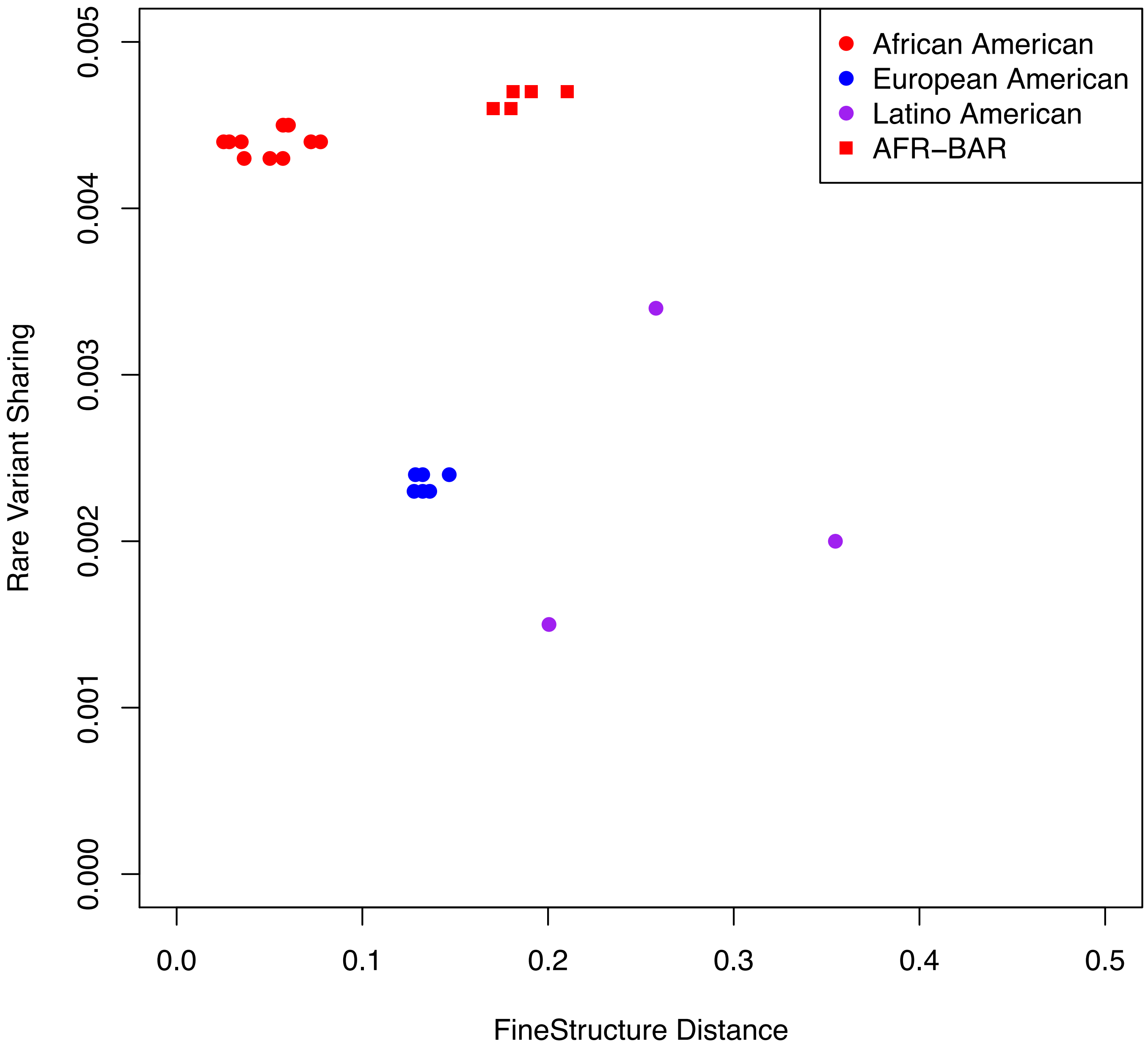
fineStructure analysis of genome-wide ancestry



- 🟥 African
- ■ Caucasia
- 🟩 East Asian
- 🟦 European

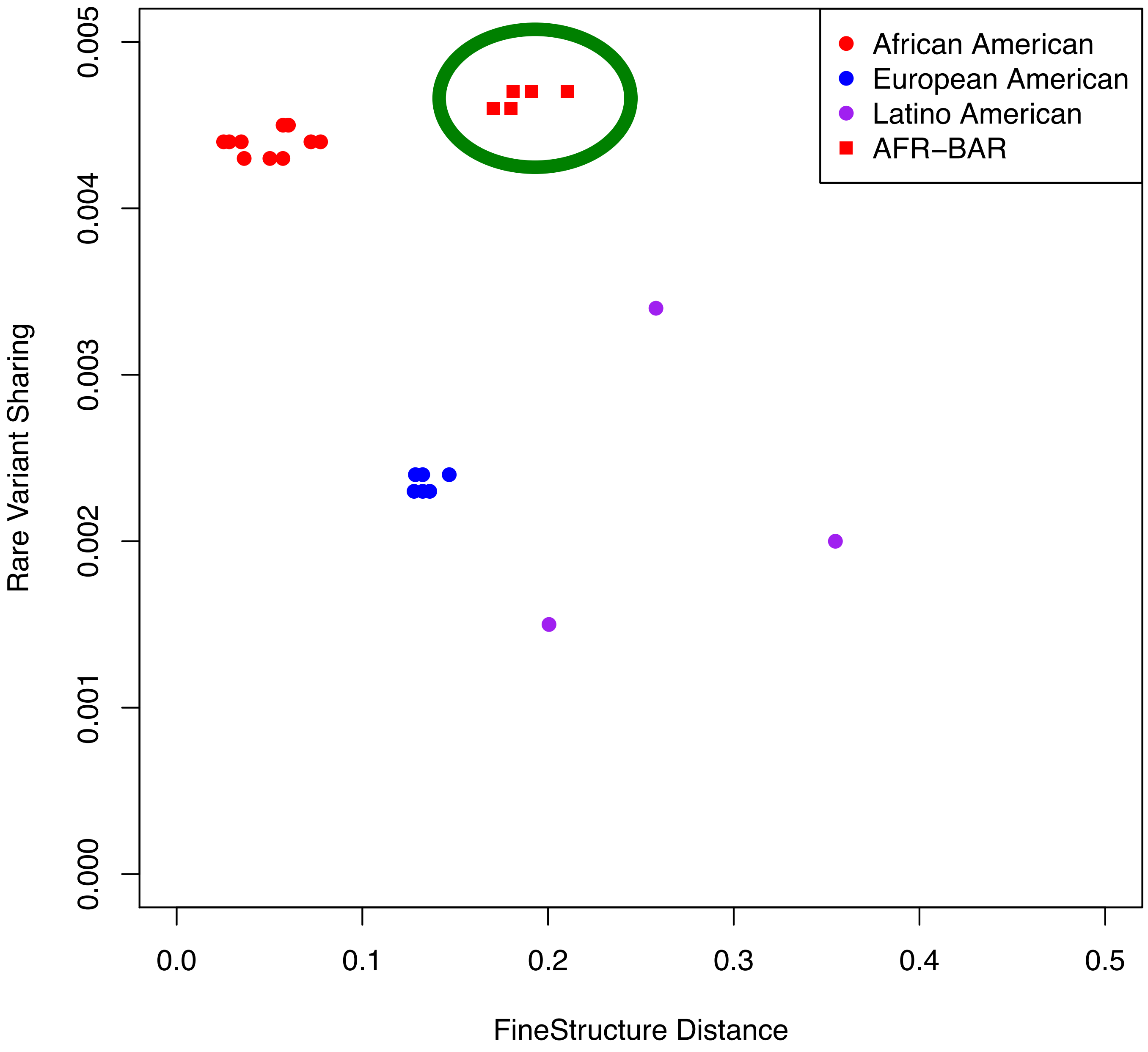
African American's have more homogeneous ancestral proportions

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

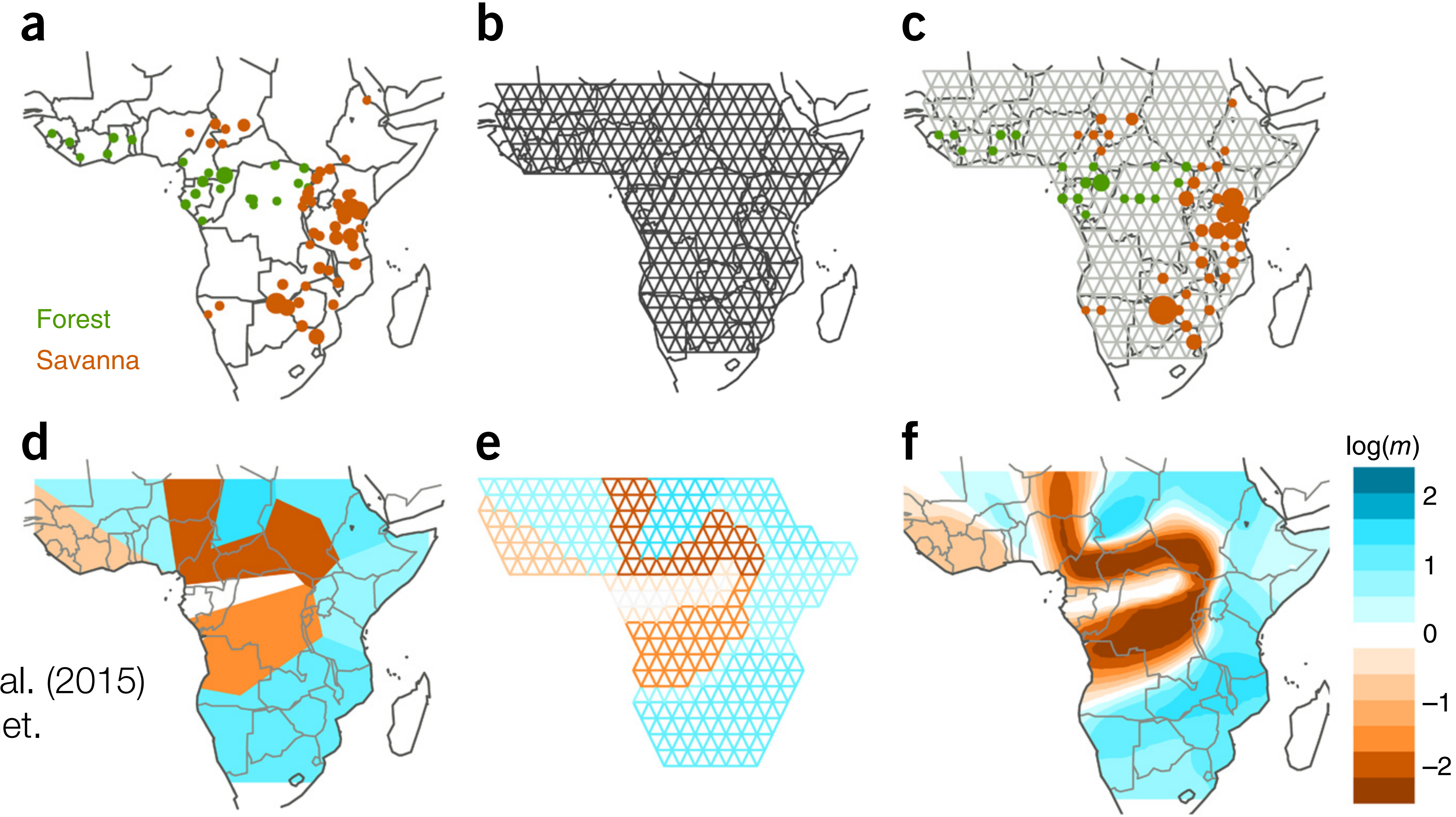


African American's have more homogeneous ancestral proportions

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



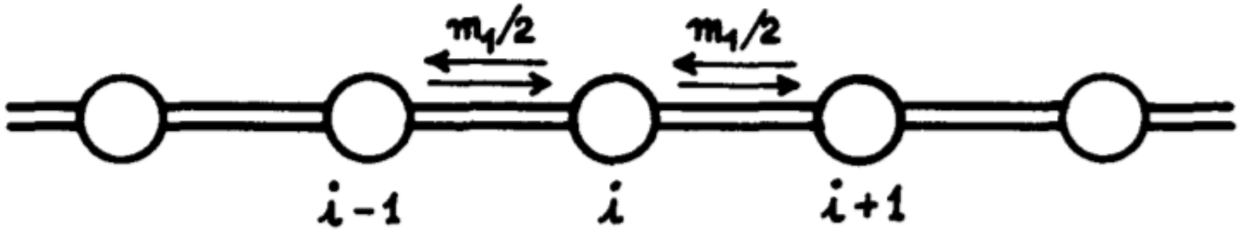
Estimated Effective Migration Surfaces (EEMS)



Petkova et al. (2015)
Nature Genet.

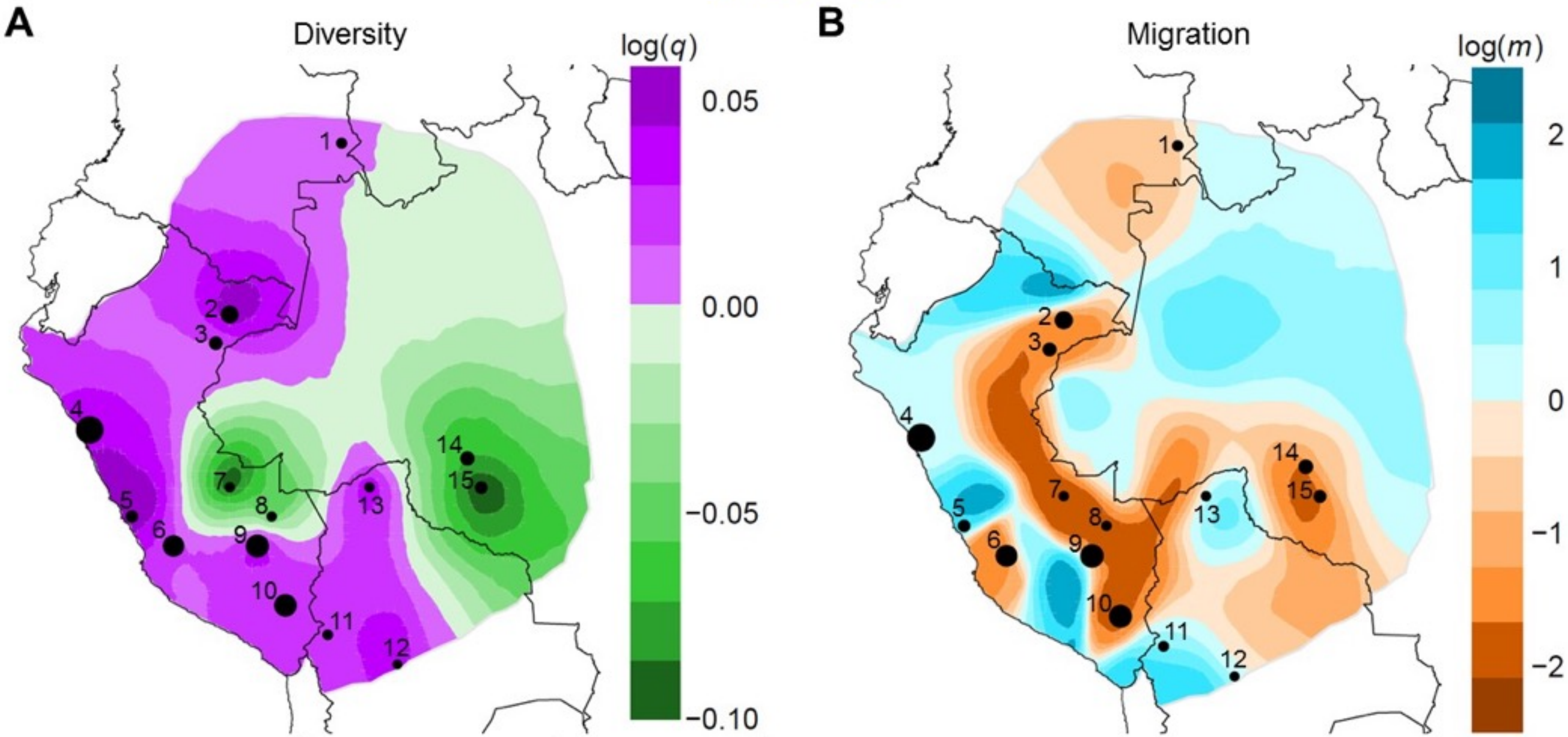
Assumptions: Stepping Stone Model

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



Kimura and Weiss (1964)

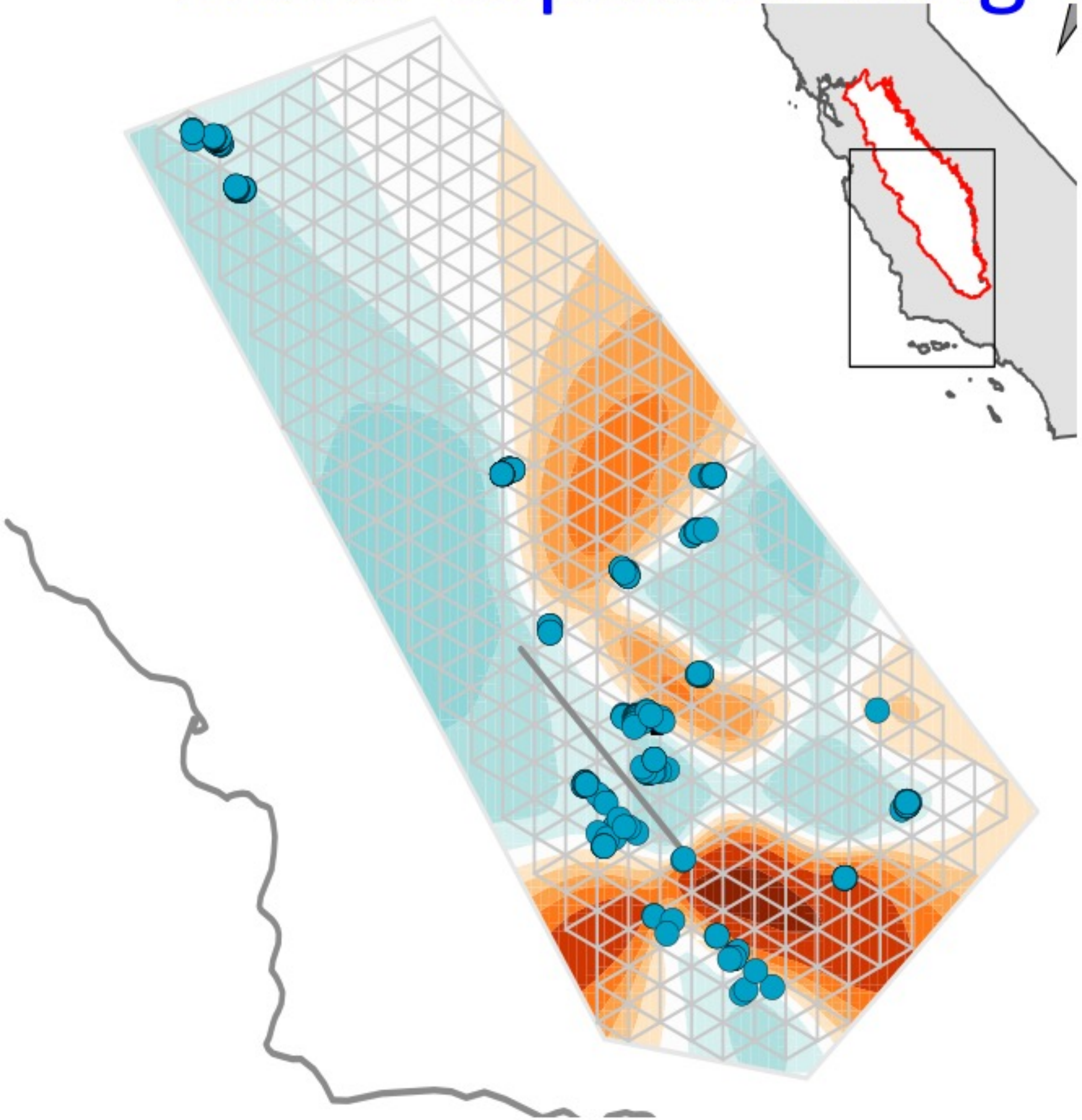
EEMS: Migration and diversity within Peru

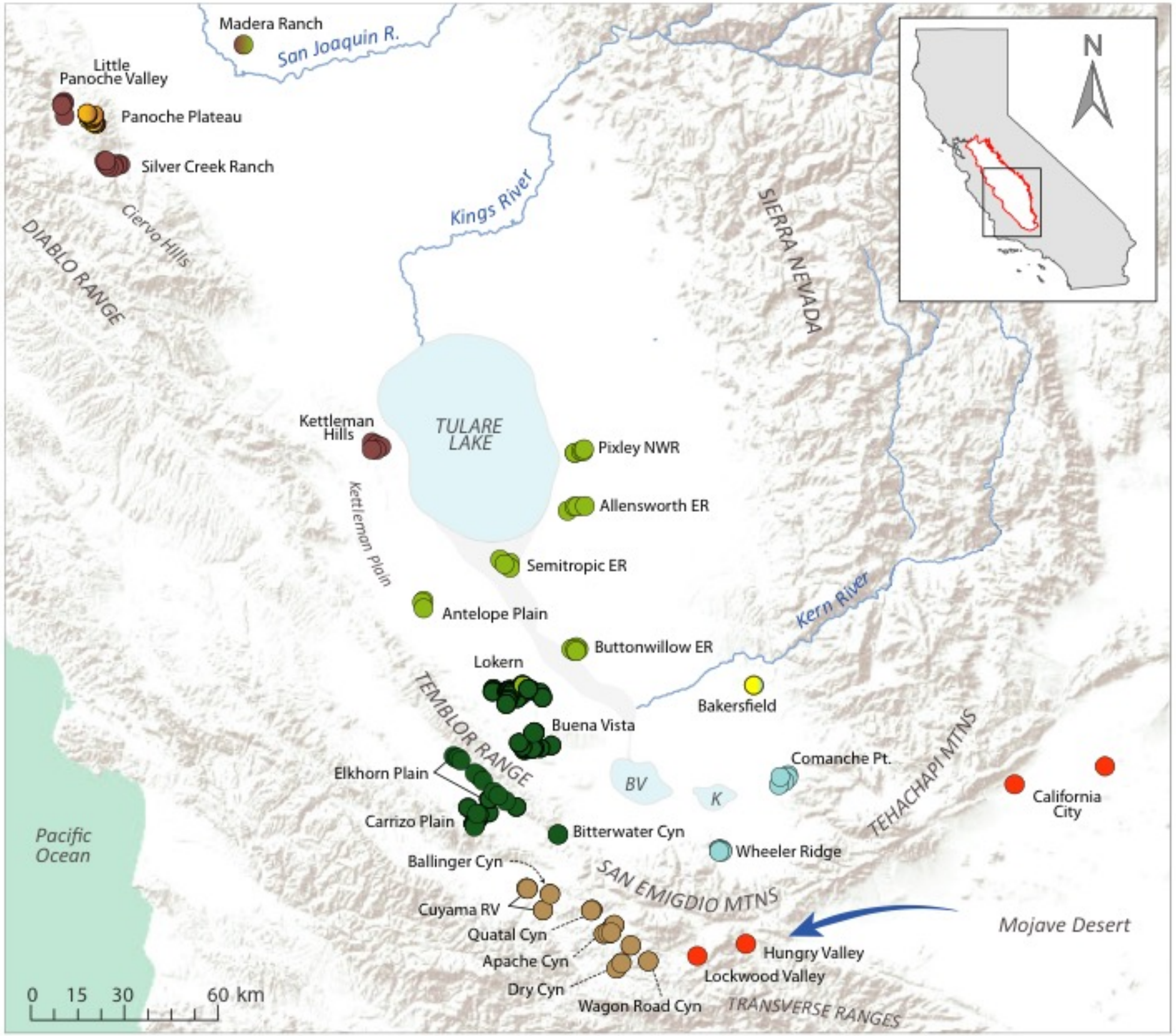


Deme	Populations	Deme	Populations	Deme	Populations	Deme	Populations
1	Piapoco	5	Lima	9	Cusco, Qeros, Quechua	15	Surui
2	Iquitos	6	Chopccas	10	Puno, Uros		
3	Matzes	7	Nahua	11,12,13	Bolivian		
4	Moches, Trujillo	8	Matsiguenka	14	Karitiana		

Harris et al. (2018) PNAS

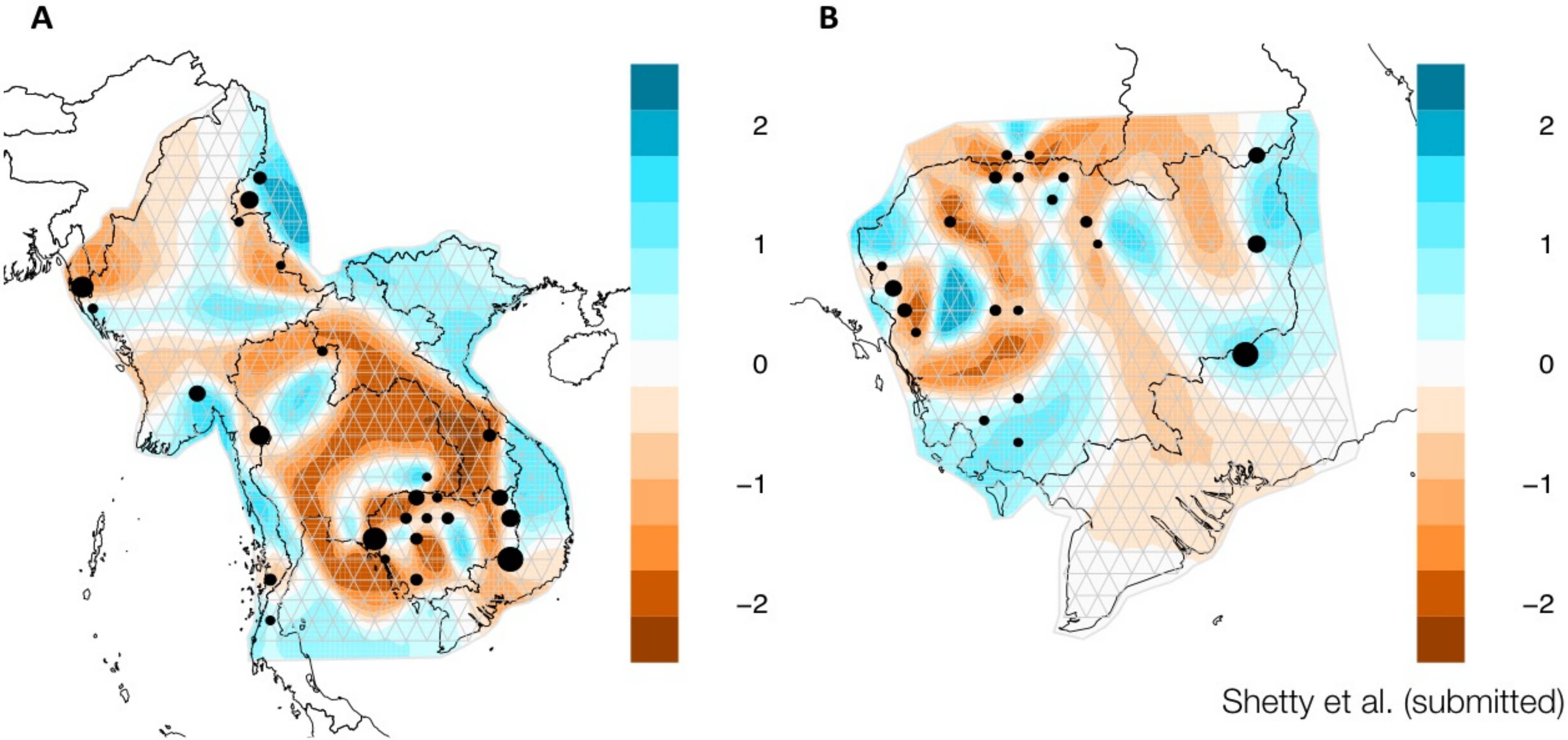
EEMS captures long-term migration patterns





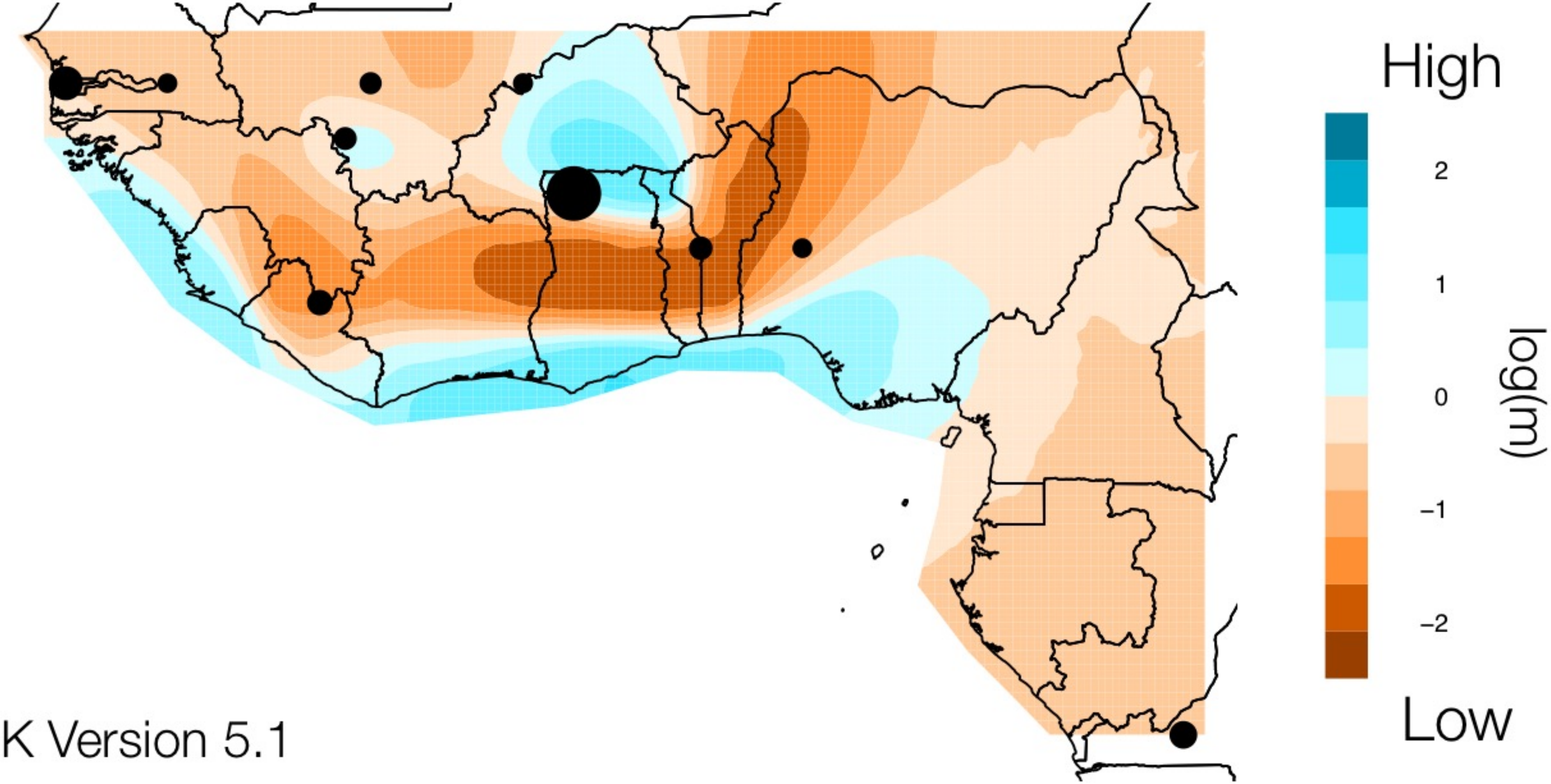
Richmond et al. (2015) Molecular Ecology

EEMS in Malaria Parasites of South East Asia



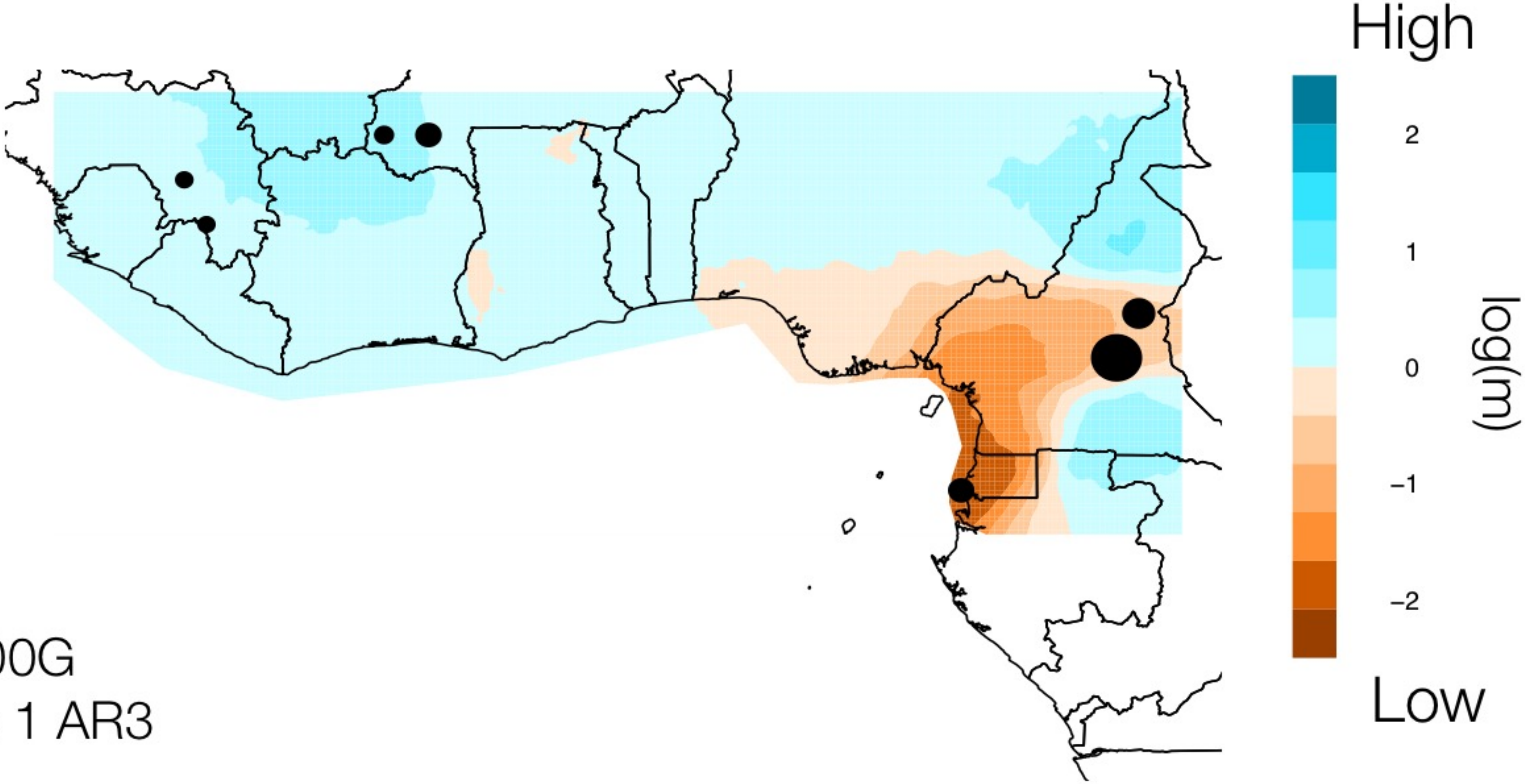
Shetty et al. (submitted)

Application to Malaria Parasites in W. Africa



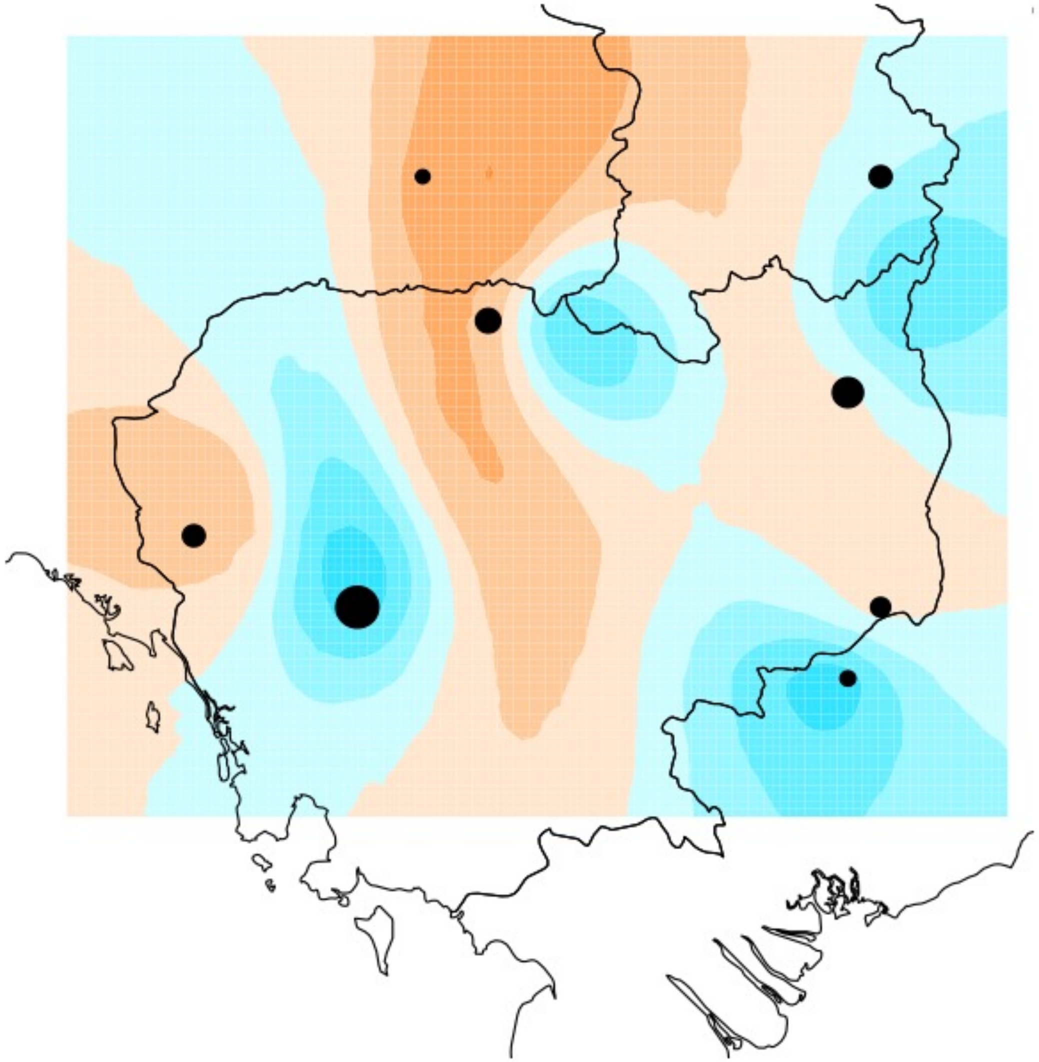
Pf3K Version 5.1

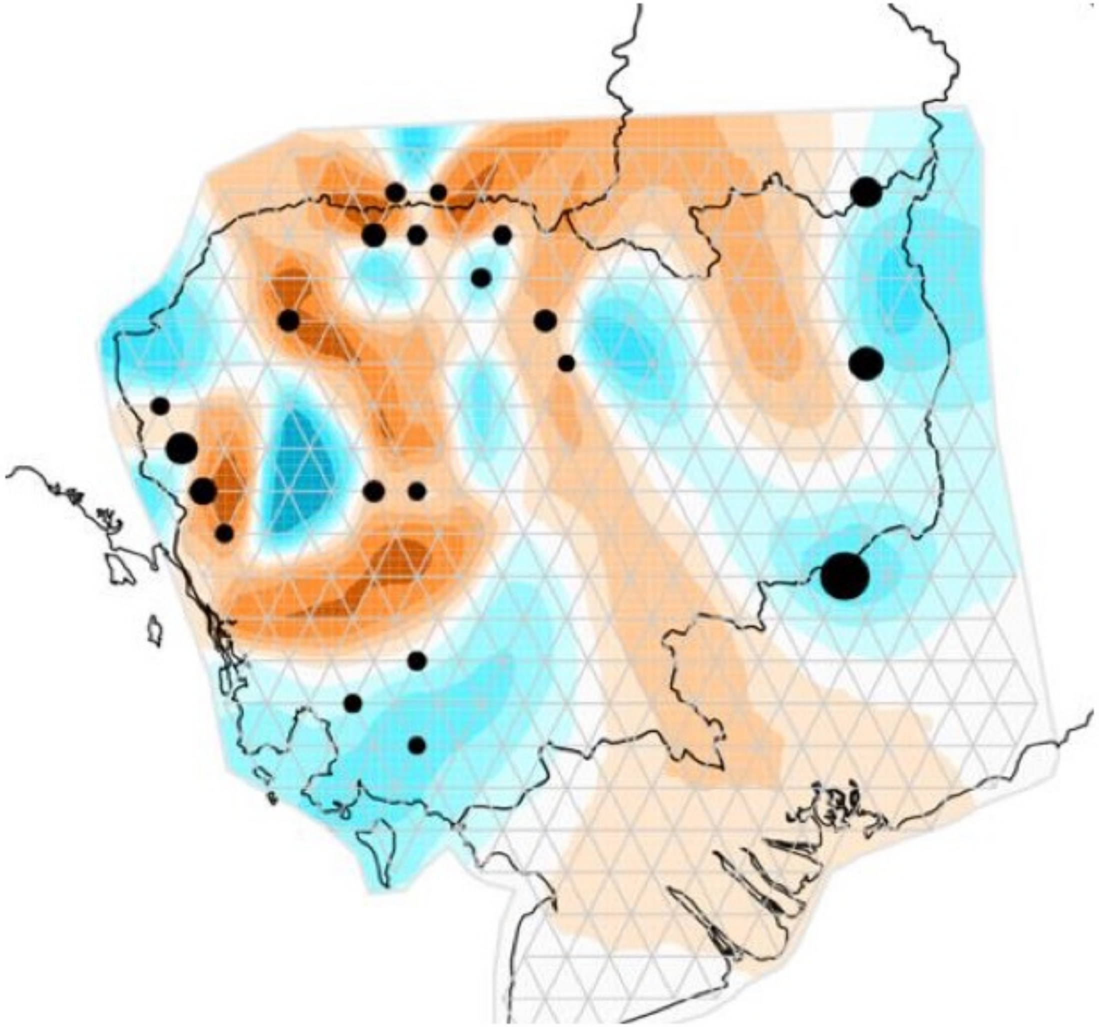
Application to Mosquito in W. Africa



Ag1000G
Phase 1 AR3

Robustness of Sampling on EEMS

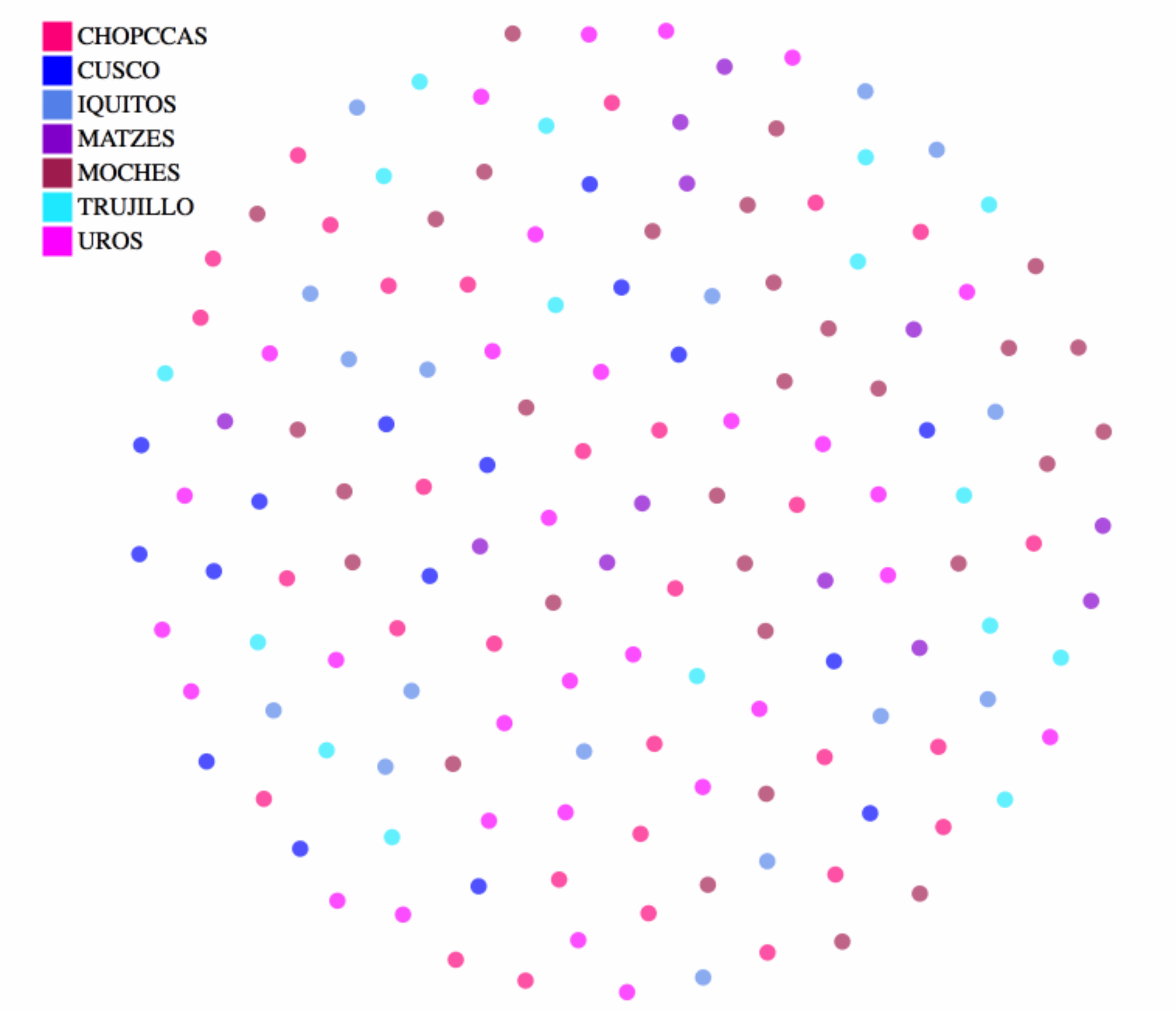




Concluding summary

- Fine-scale population structure is subdivisions of individuals on an ever increasingly granular scale
- Identity-by-descent and sharing of rare variants are a powerful method 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