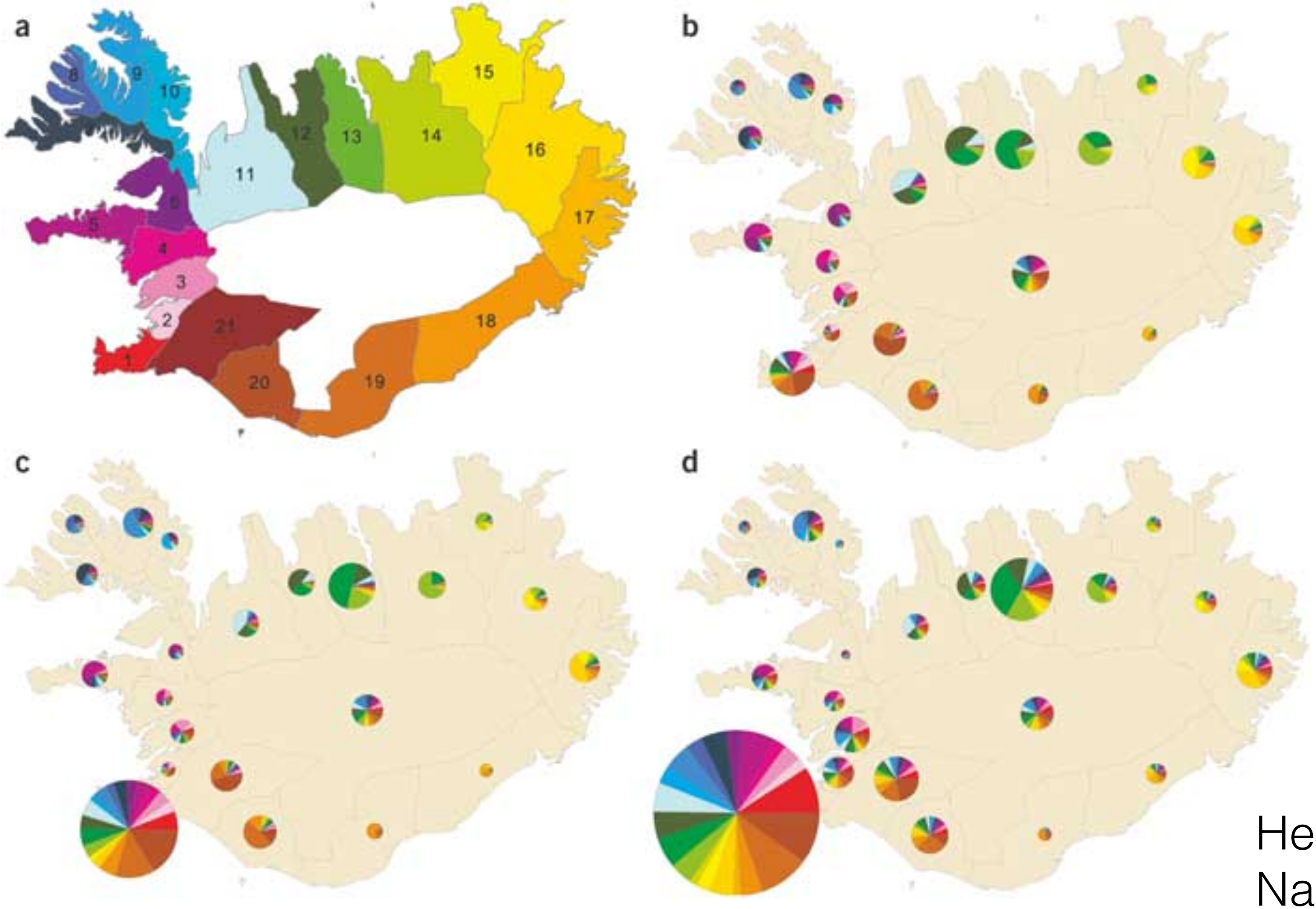
Cryptic Relatedness and fine scale population structure

Learning objectives

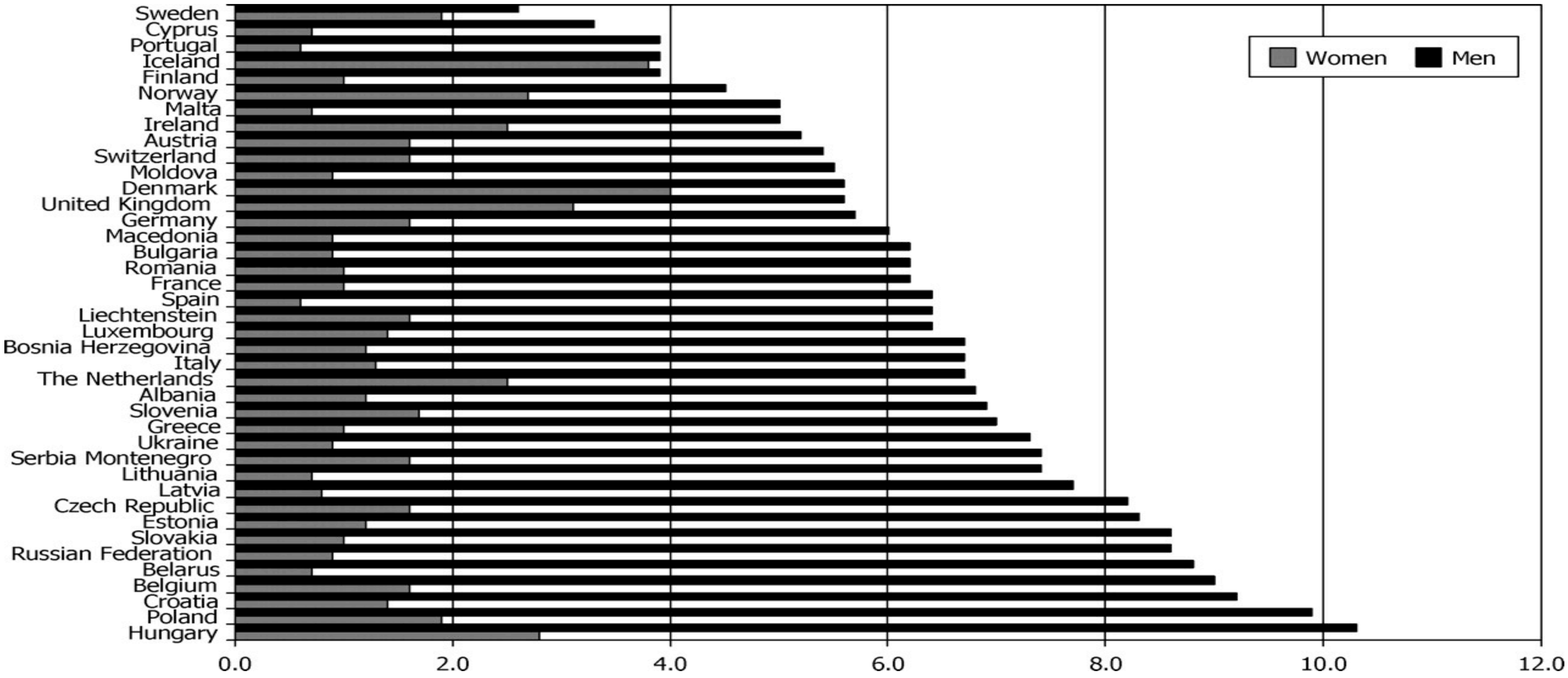
- Define fine scale population structure and cryptic relatedness
- How is it identified
  - Identity-by-descent
  - Rare variation
- Why it can be important for association analyses, especially of rare variants.

Cryptic Population Structure



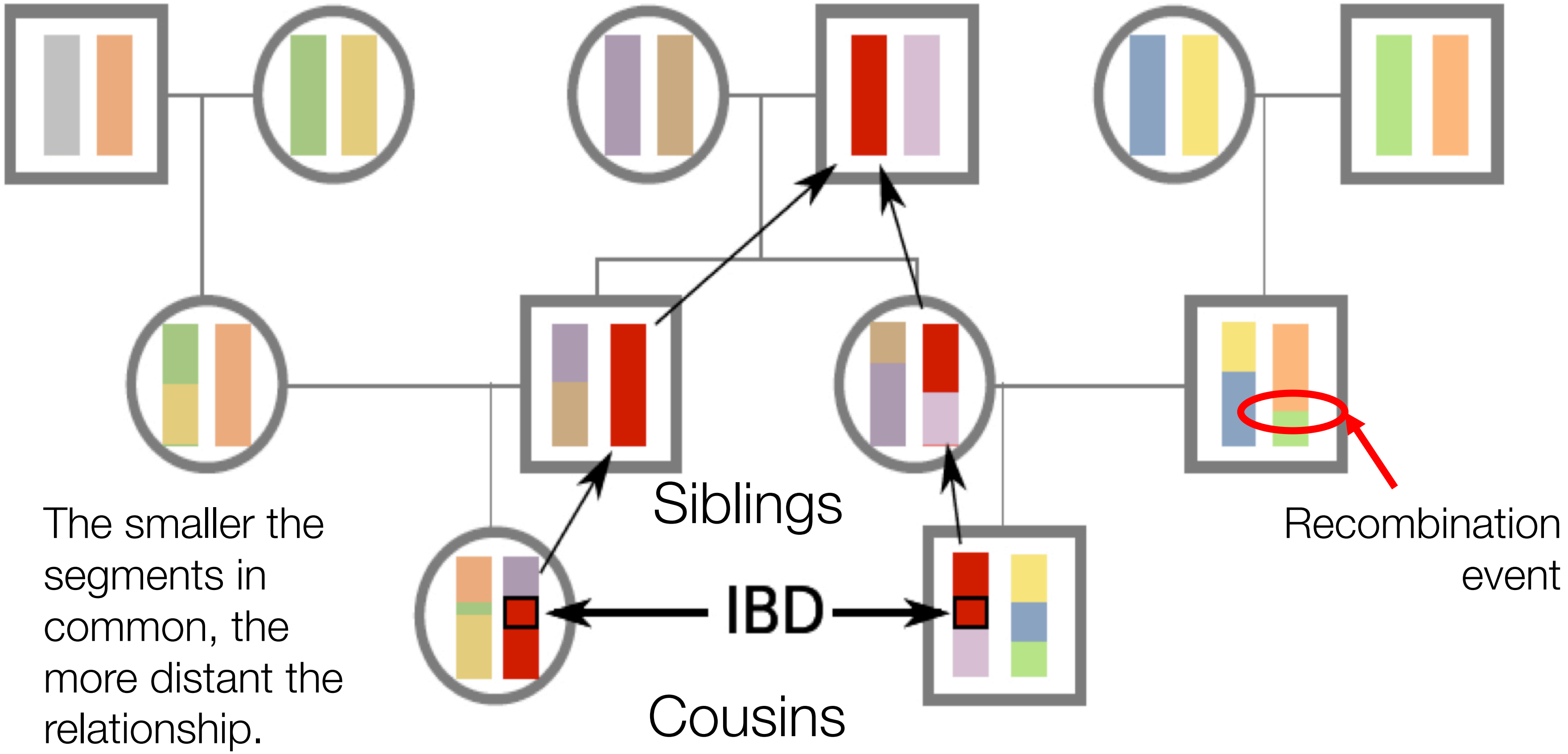
Helgason et al. (2004)
Nature Genet.

Lung Cancer Prevalence in Europe



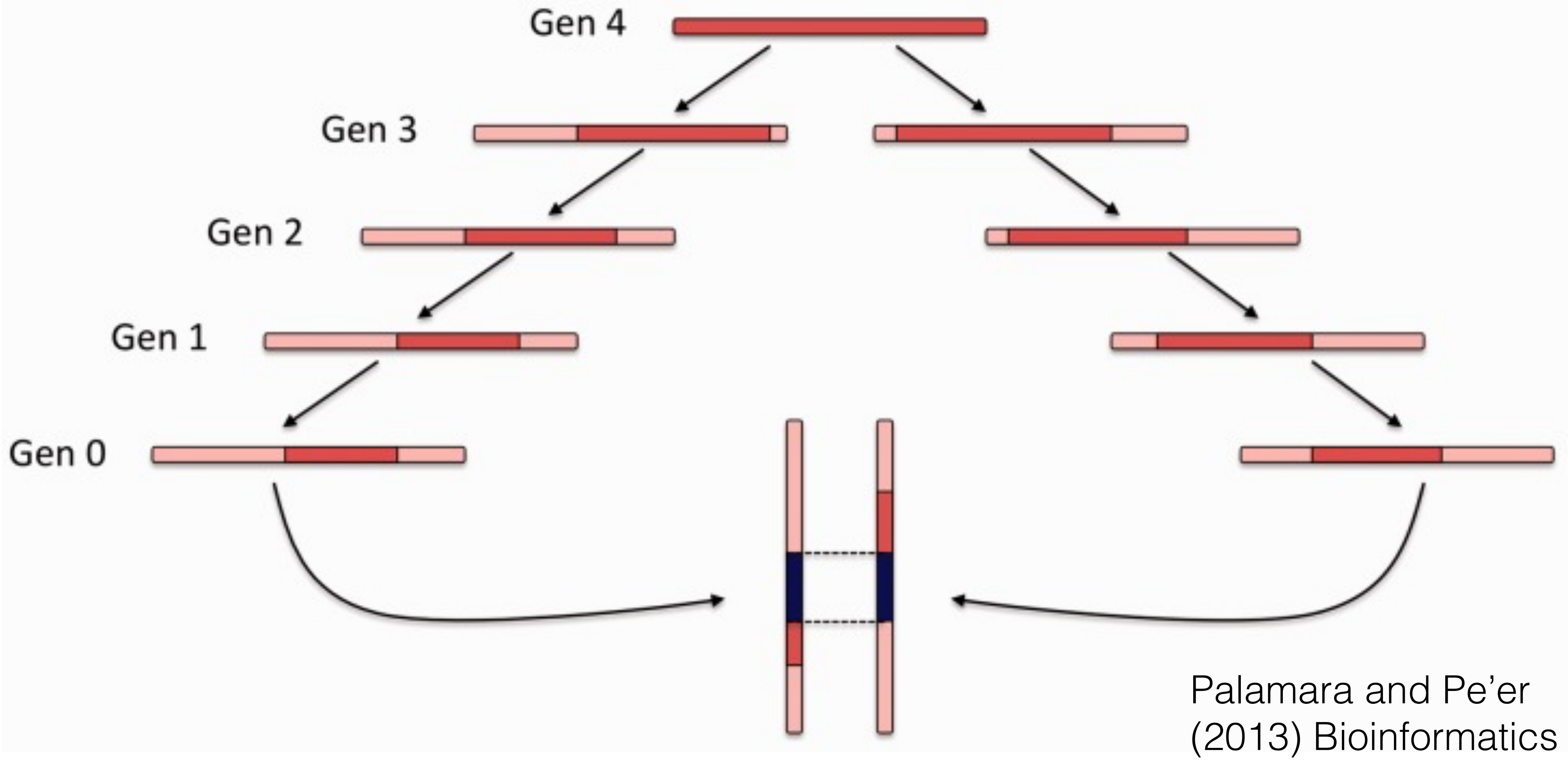
Boyle and Ferlay (2005) Annals of Oncology

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.

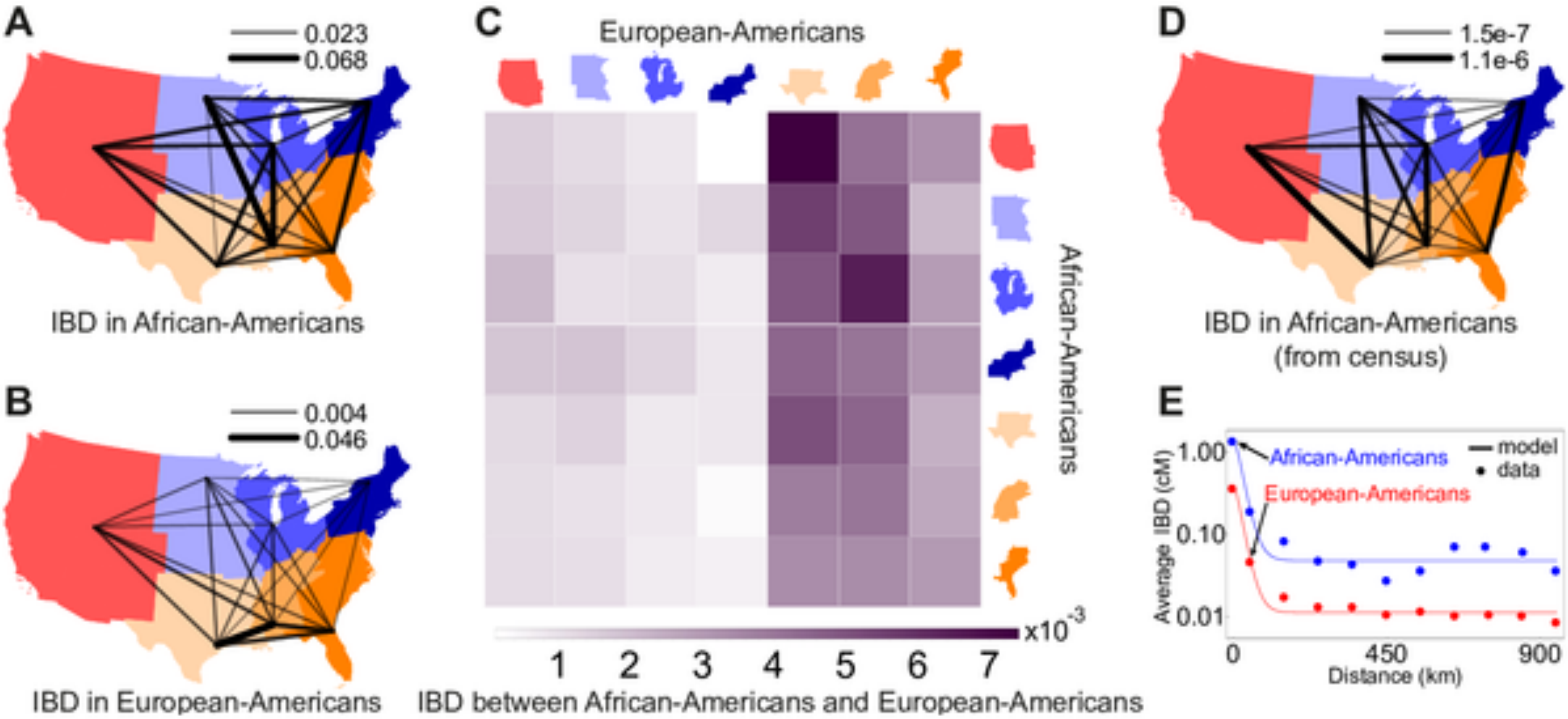


Palamara and Pe’er (2013) Bioinformatics

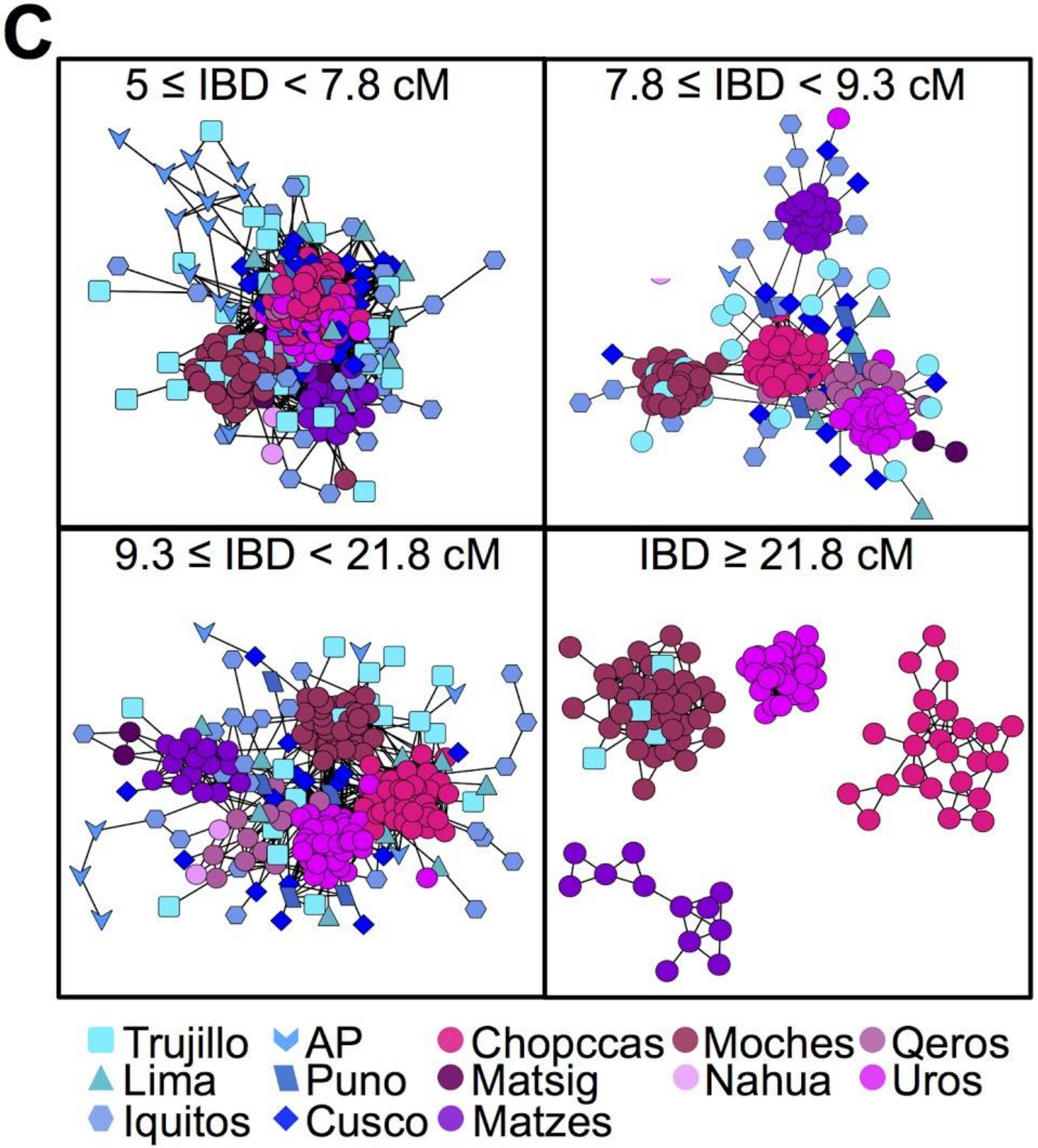
$$E[g|l] \cong \frac{3}{2 \ * l}$$

Baharian et al. (2016) PLoS Genet.

Pairwise genetic relatedness across



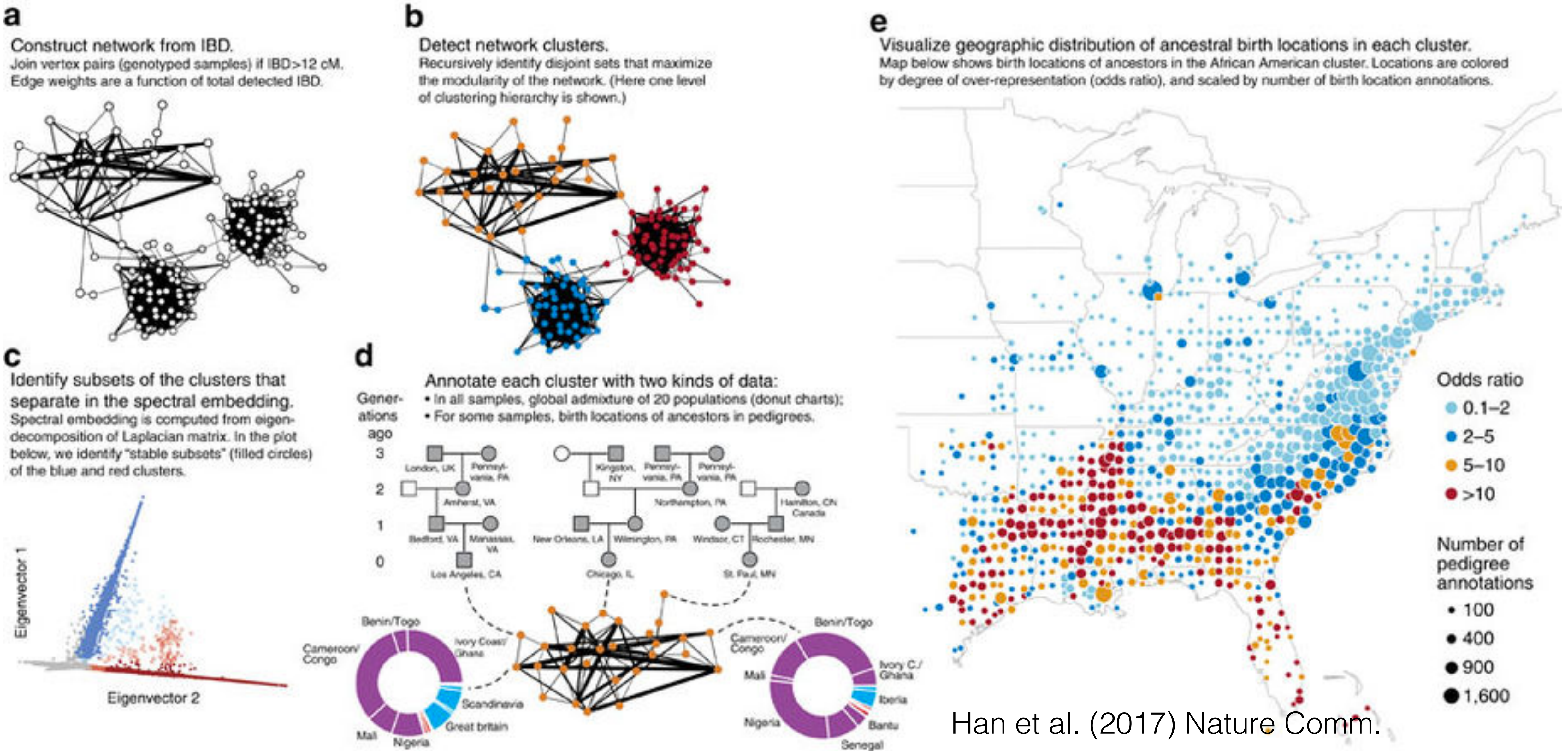
Baharian et al. (2016) PLoS Genet.



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

Harris et al. (submitted)

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.

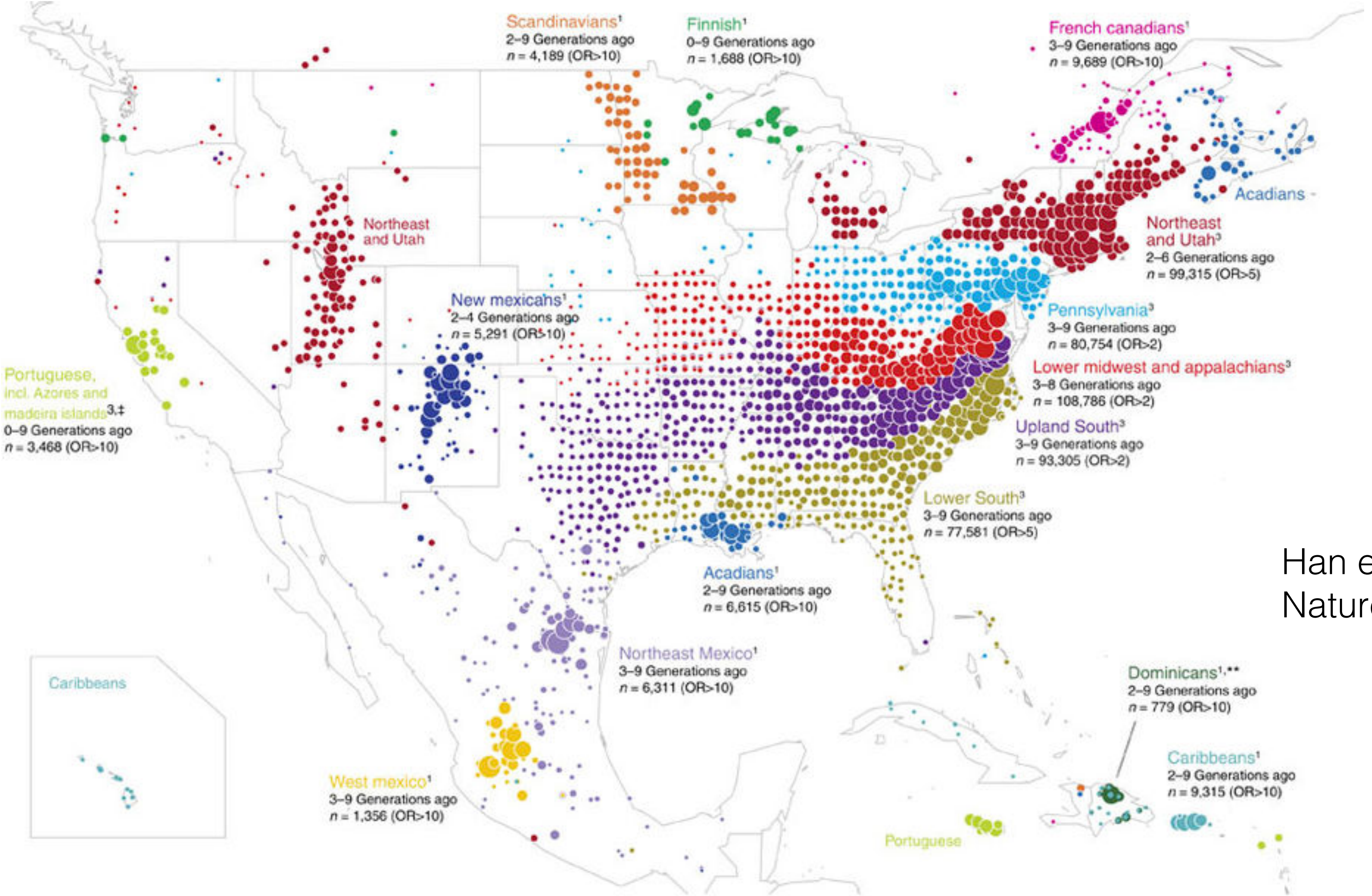
d Annotate each cluster with two kinds of data:

- 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.

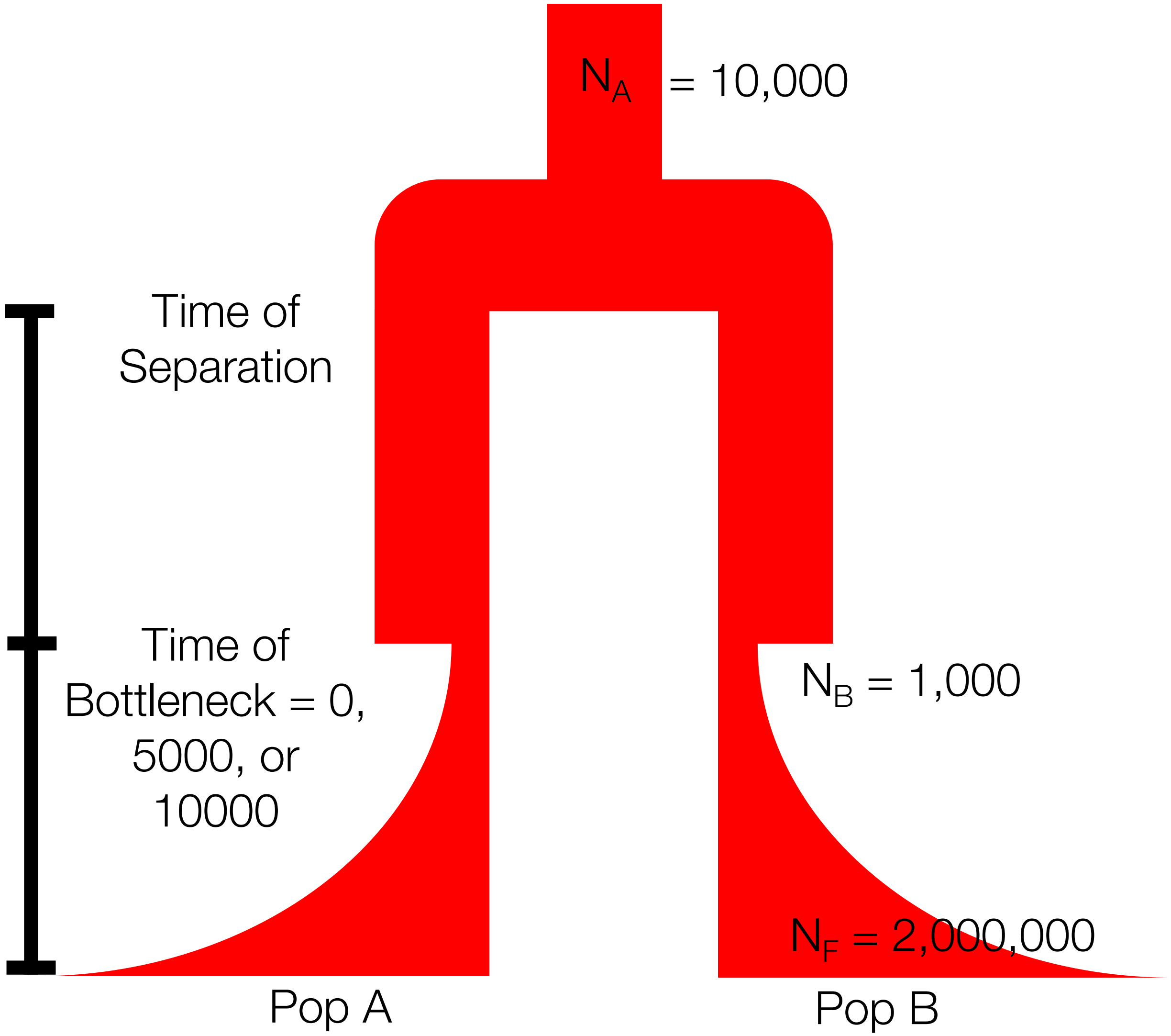
Han et al. (2017) Nature Comm.

IBD on a large scale



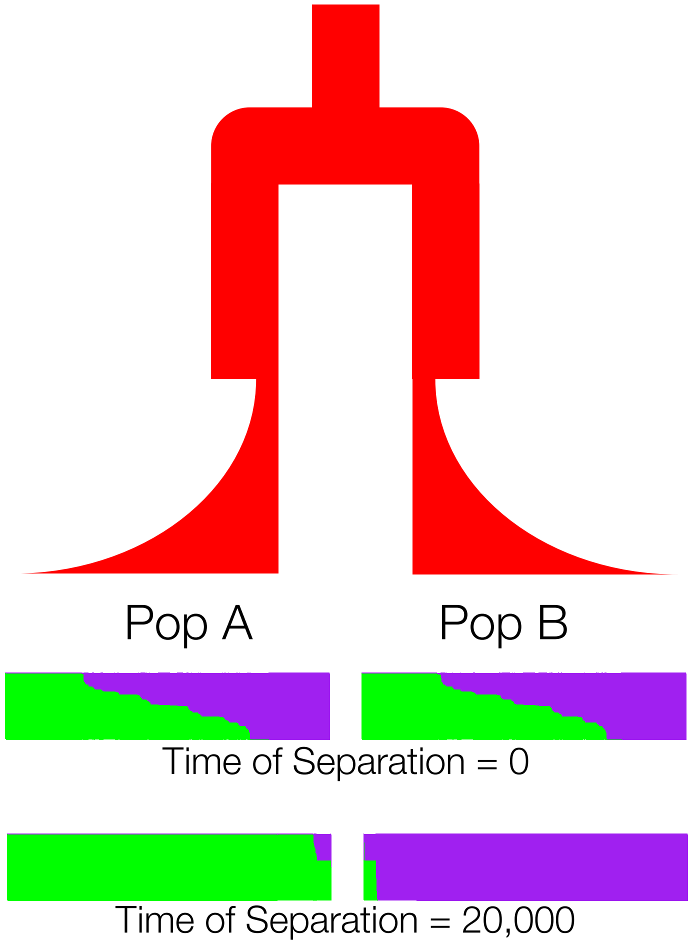
Han et al. (2017)
Nature Comm.

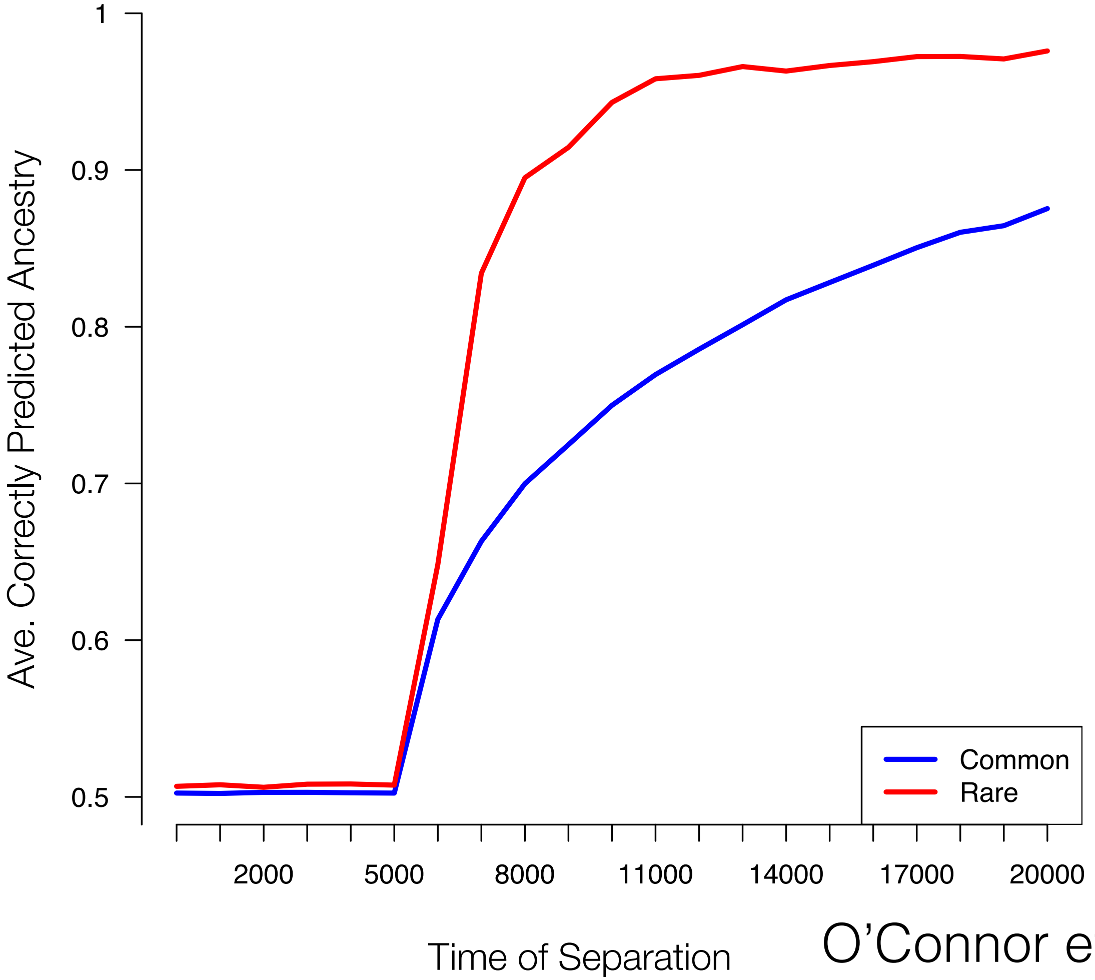
Rare VS Common: Population Structure Simulations



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?

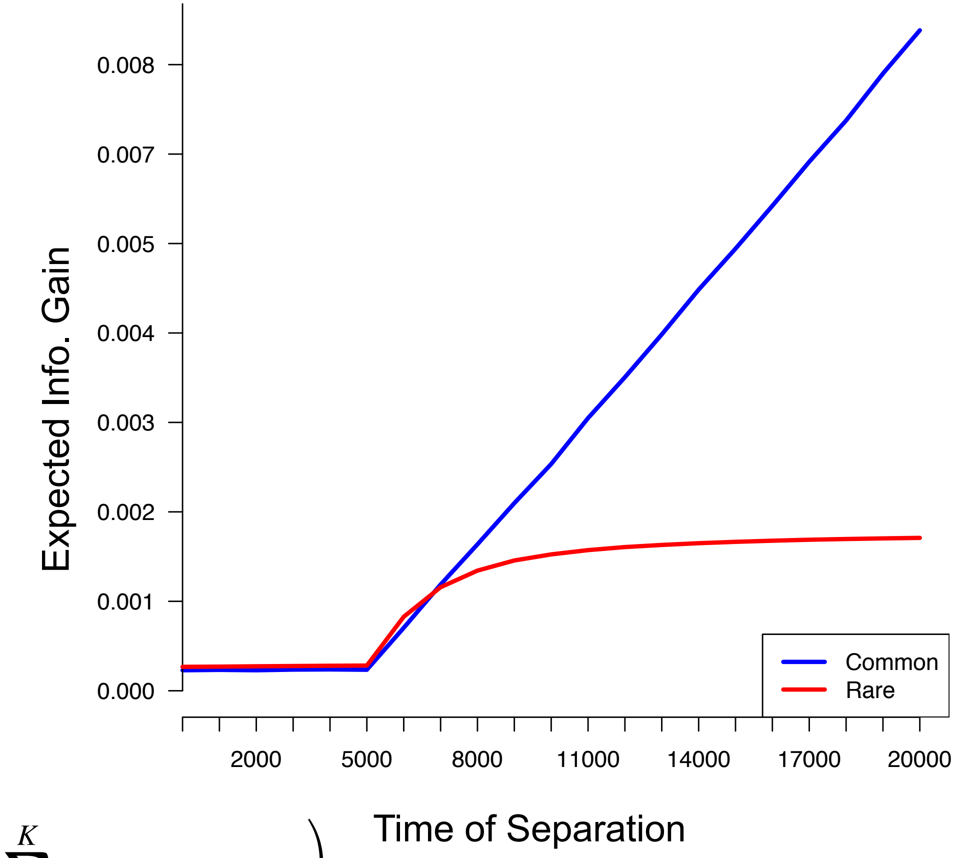
Information Gain: how well a variant can distinguish between populations. (Rosenberg et al. 2003)

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

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)

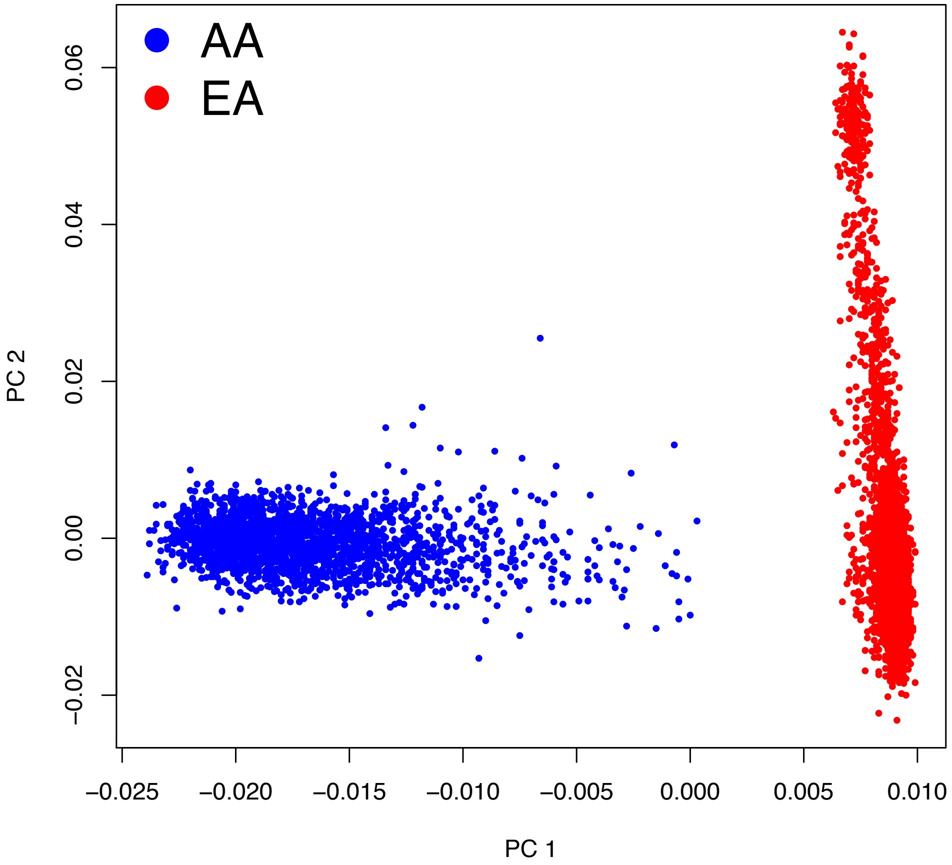
$$E(I_n \mid C,M) = \sum_{m\in M}\sum_{l=0}^{C} r_{lm} \times \sum_{j=1}^{N}\left(-p_{jlm} \ln p_{jlm} + \sum_{i=1}^{K} q_i p_{ijlm} \ln p_{ijlm}\right)$$



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

Rare Variants Identify Cryptic Populations

Common (MAF > 10%)

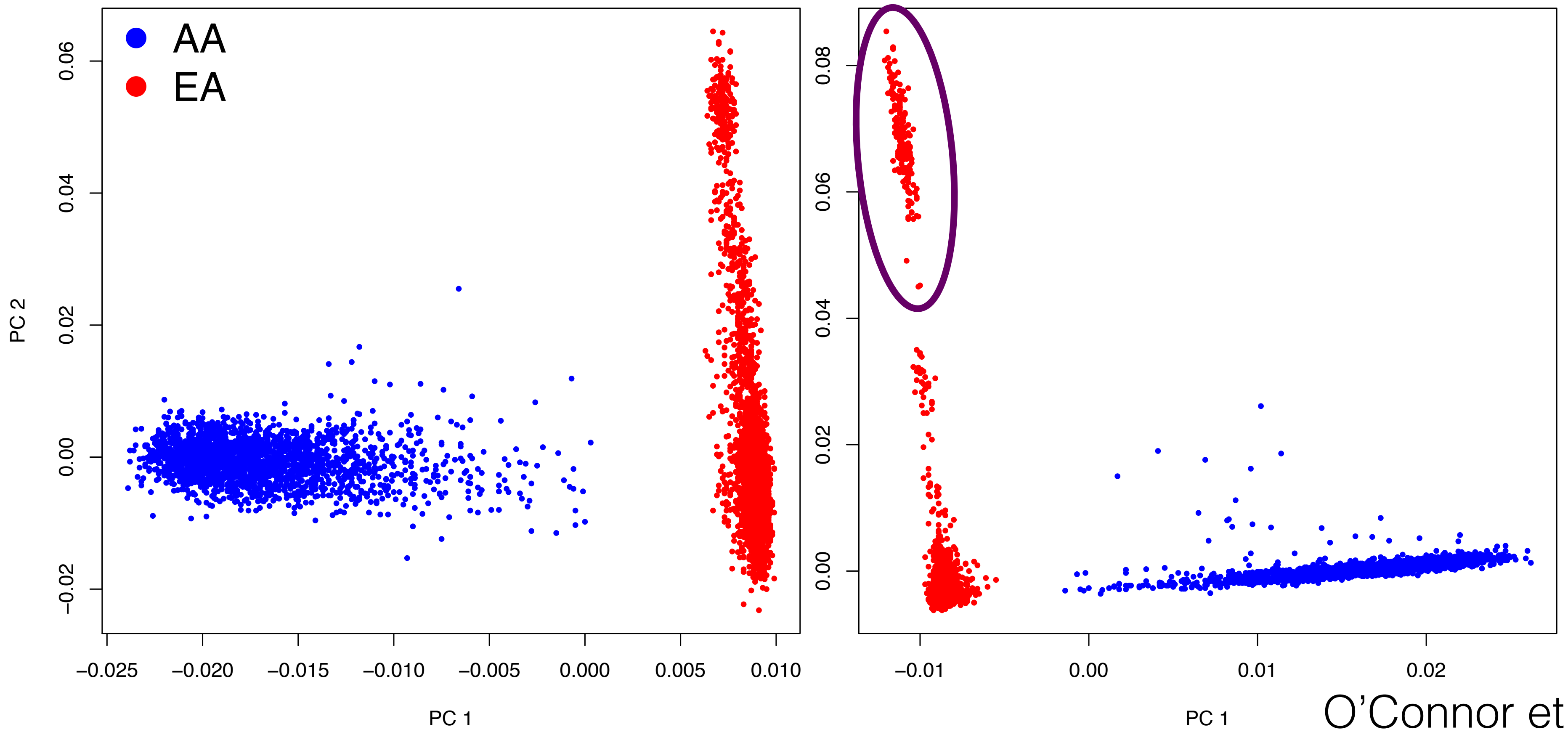


O’Connor et al. (2014)
Mol. Biol. Evol.

Rare Variants Identify Cryptic Populations

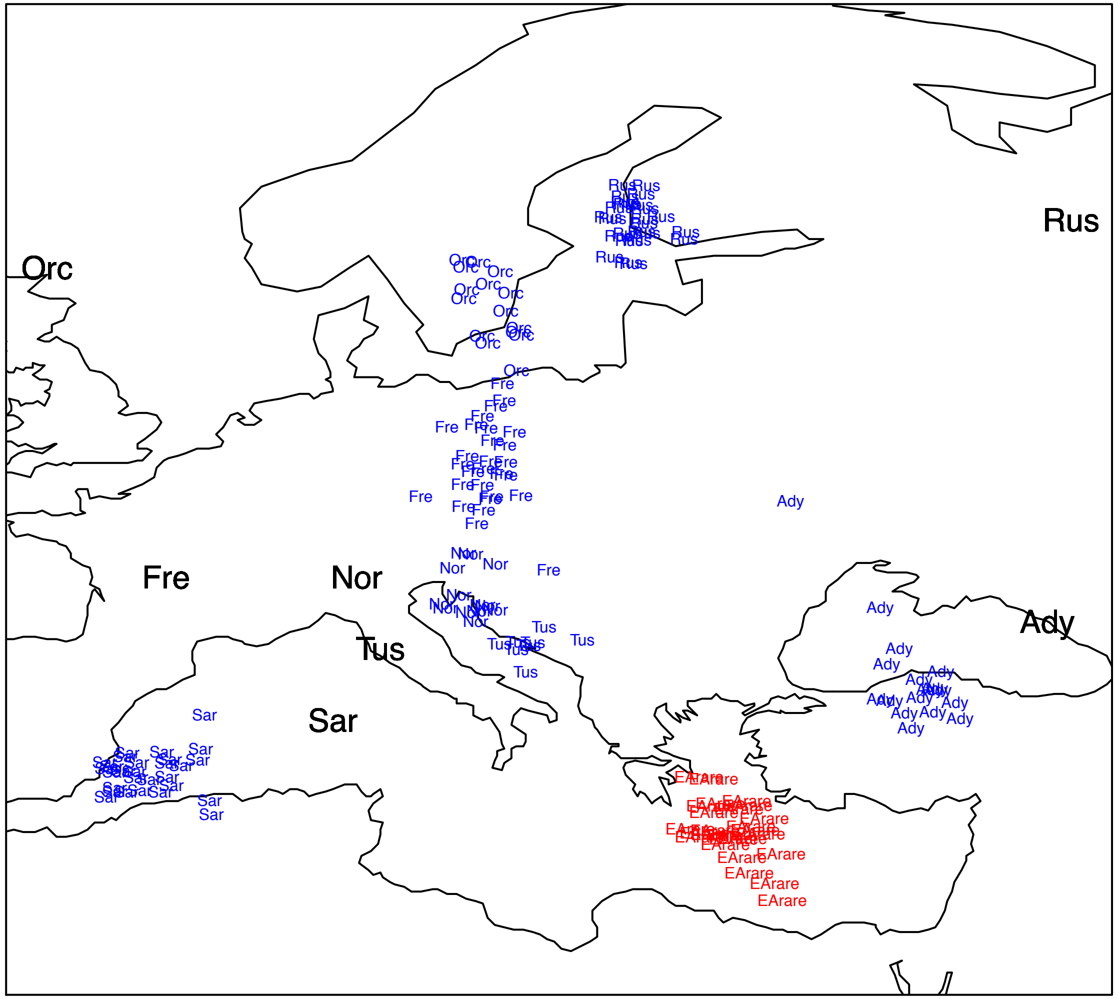
Common (MAF > 10%)

Rare (MAF < 0.5%)



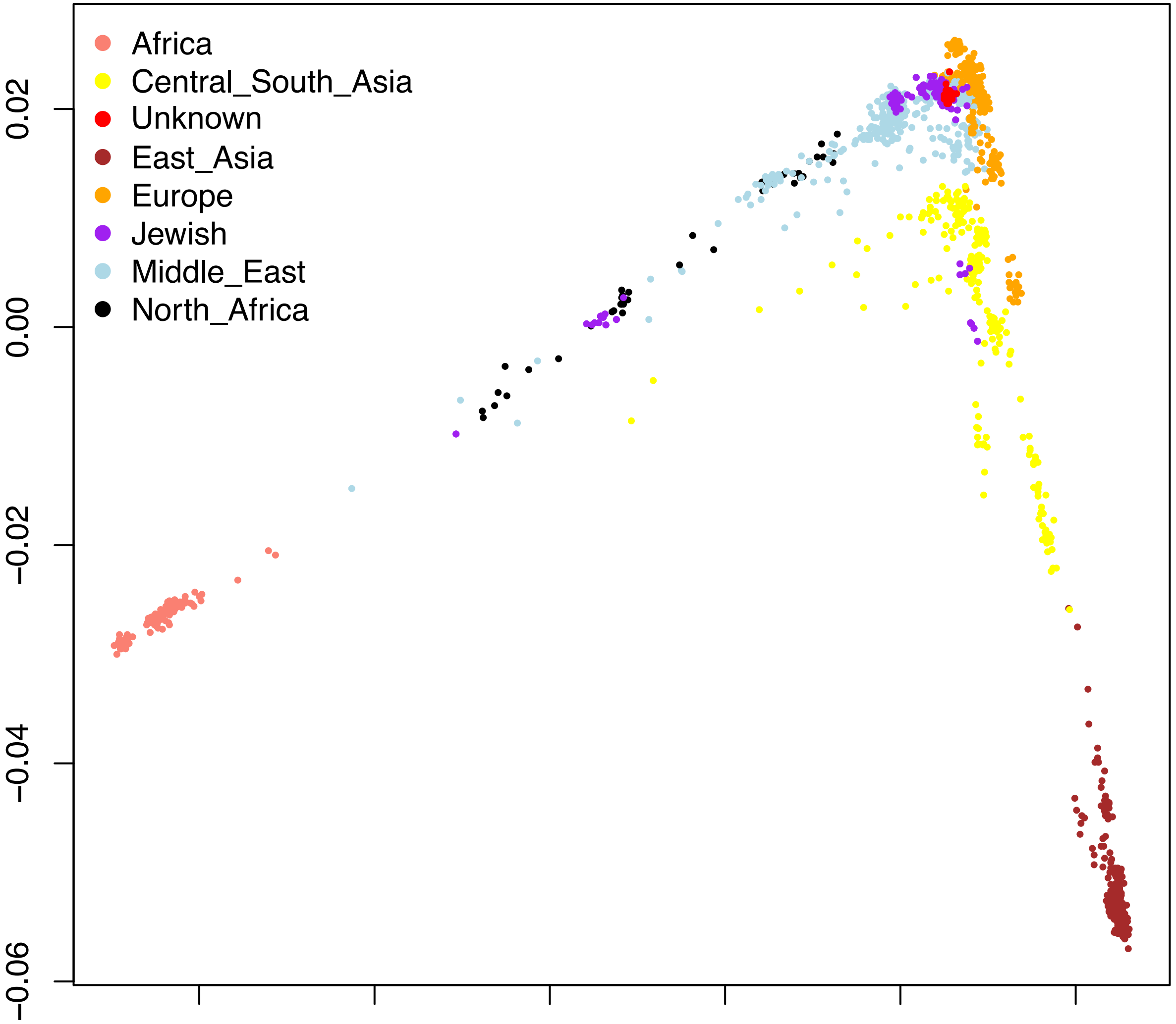
O’Connor et al. (2014)
Mol. Biol. Evol.

What is Their Geographic Ancestry?



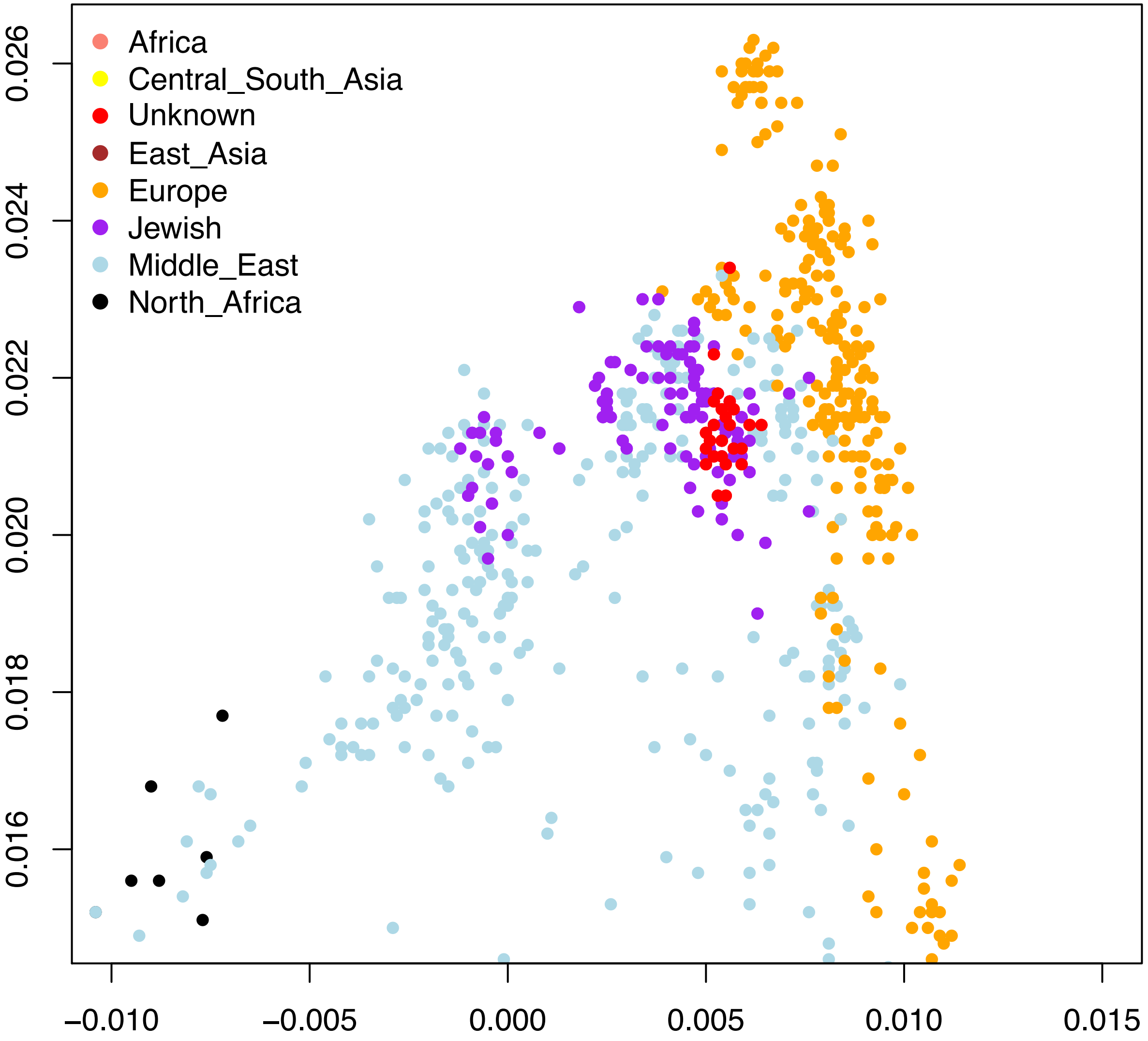
O’Connor et al. (2014)
Mol. Biol. Evol.

PCA of Global Diversity Including Cryptic Population



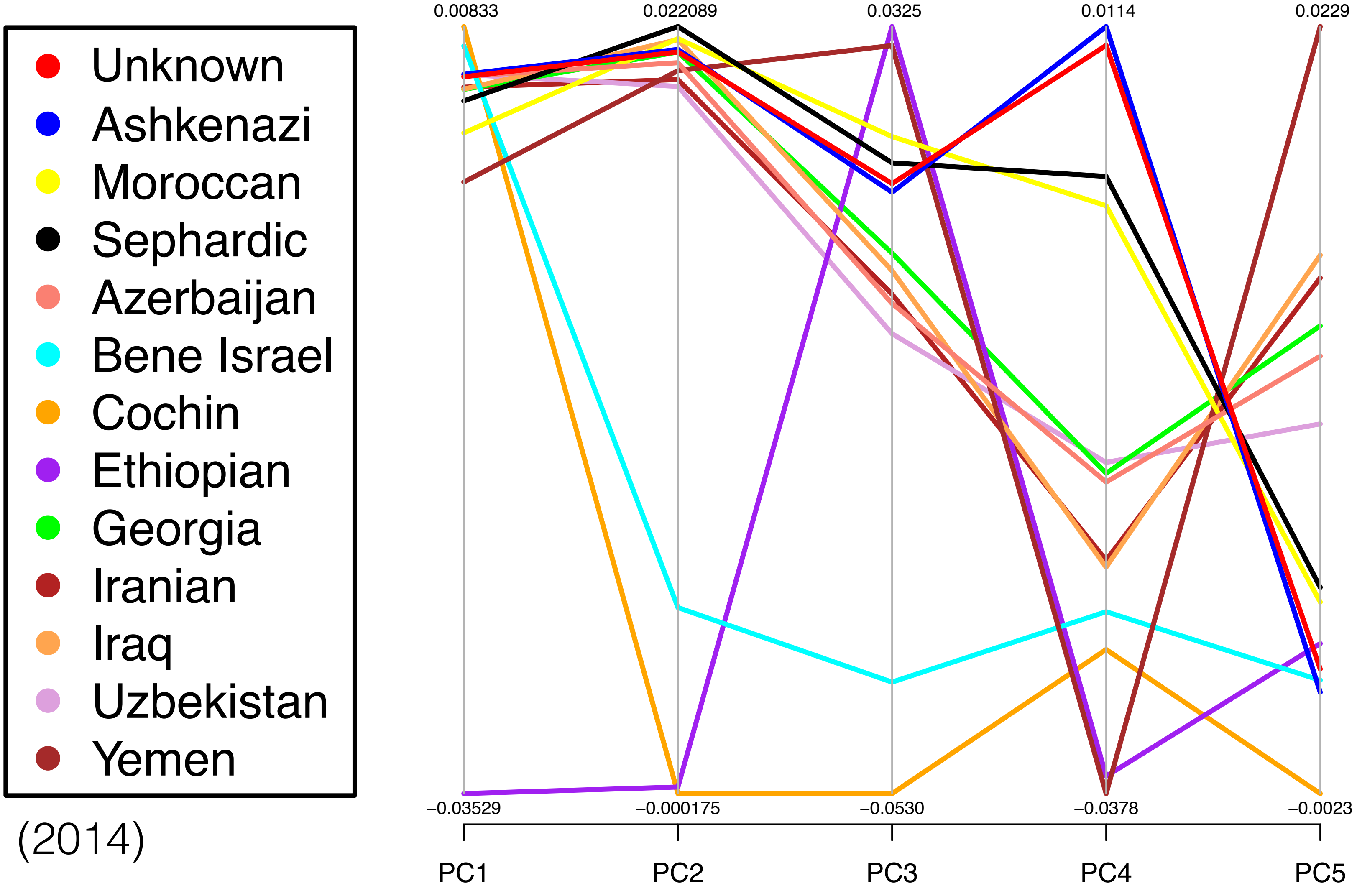
O’Connor et al. (2014)
Mol. Biol. Evol.

PCA of Global Diversity Including Cryptic Population



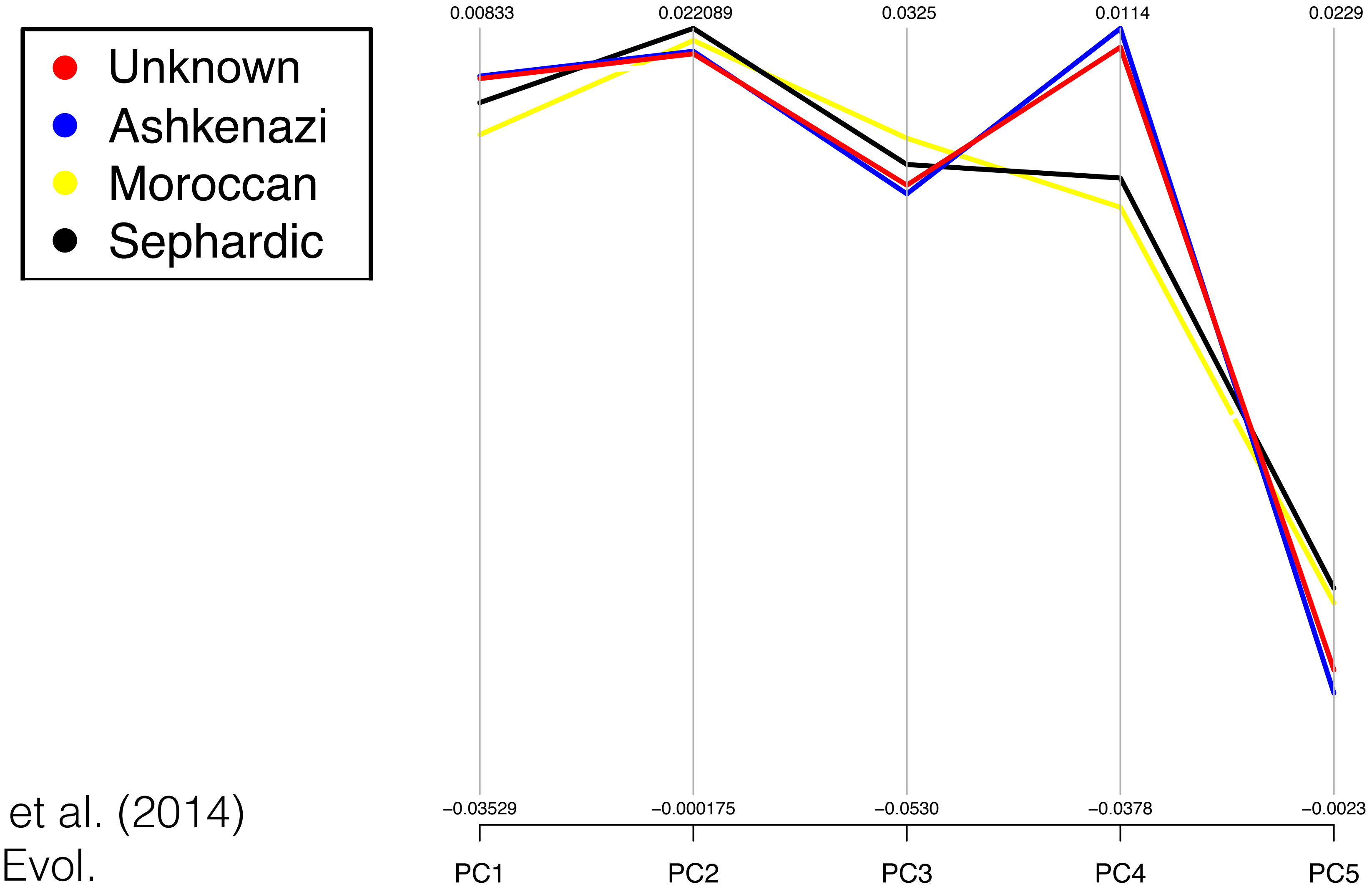
O’Connor et al. (2014)
Mol. Biol. Evol.

Population Average PCA with More Axes



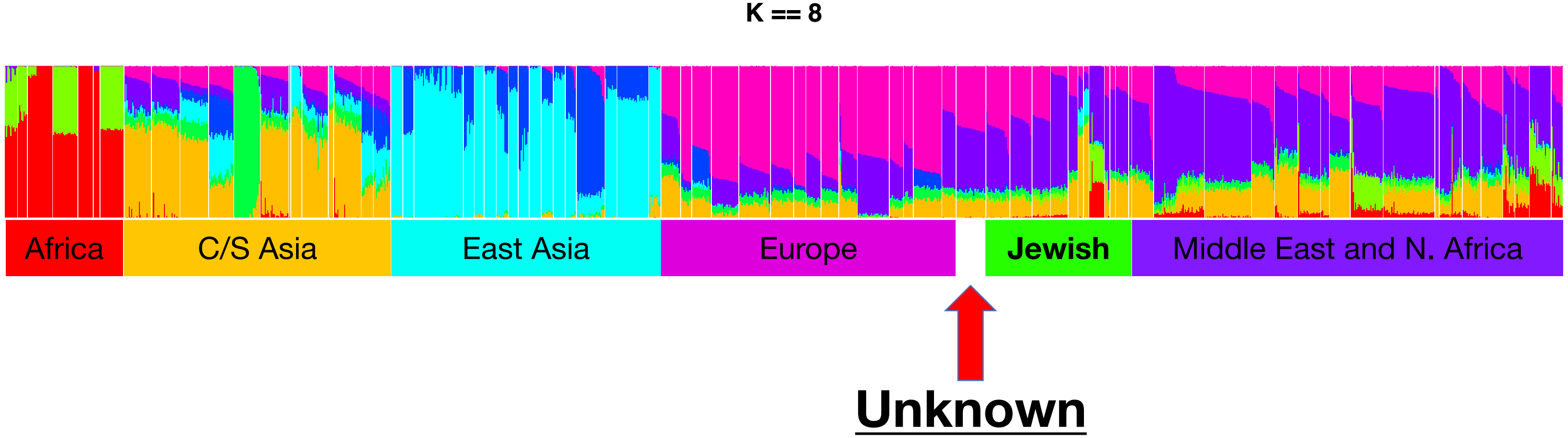
O'Connor et al. (2014)
Mol. Biol. Evol.

Population Average PCA with More Axes



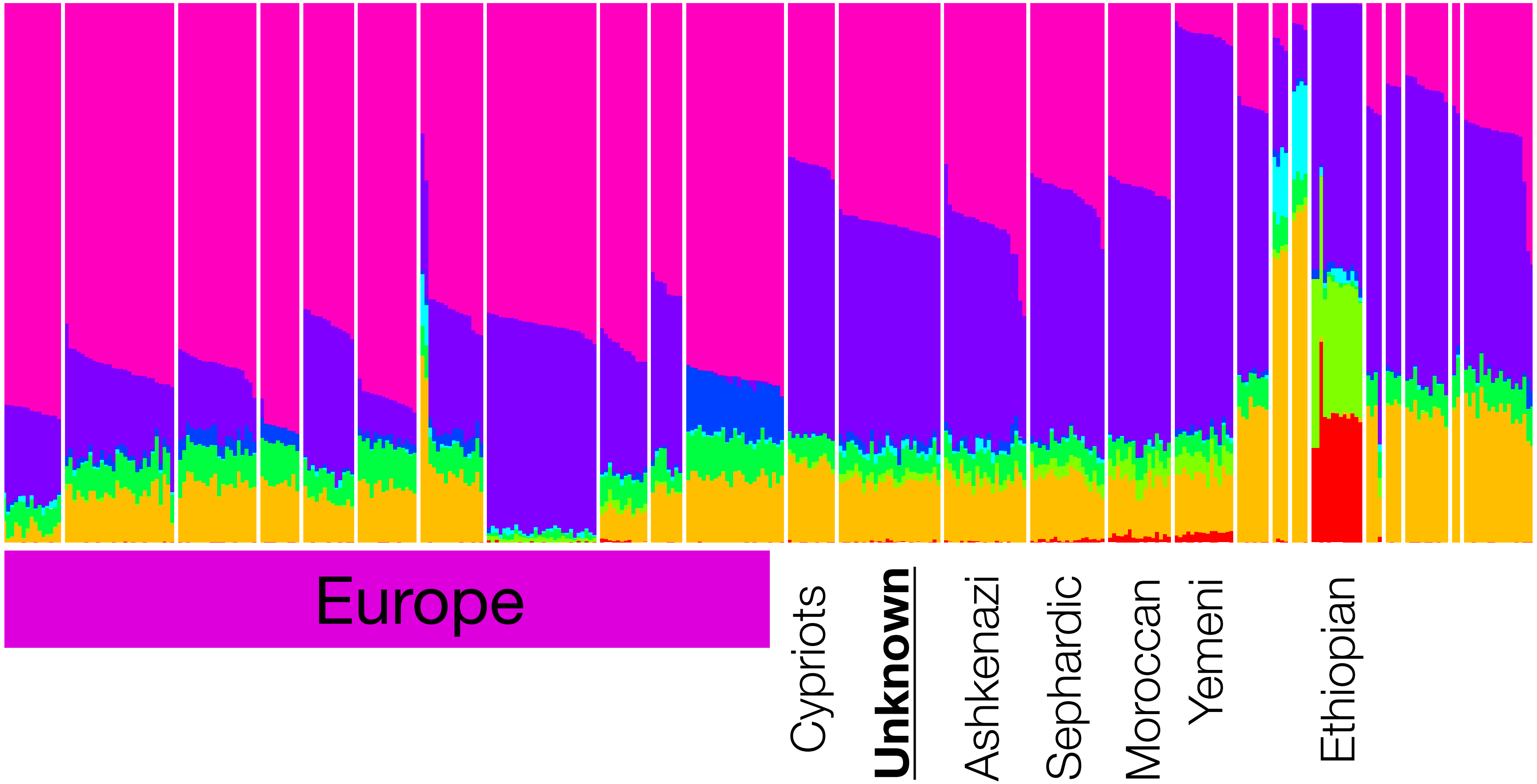
O’Connor et al. (2014)
Mol. Biol. Evol.

Cryptic Group has Similar Admixture Proportions to Jewish Groups.

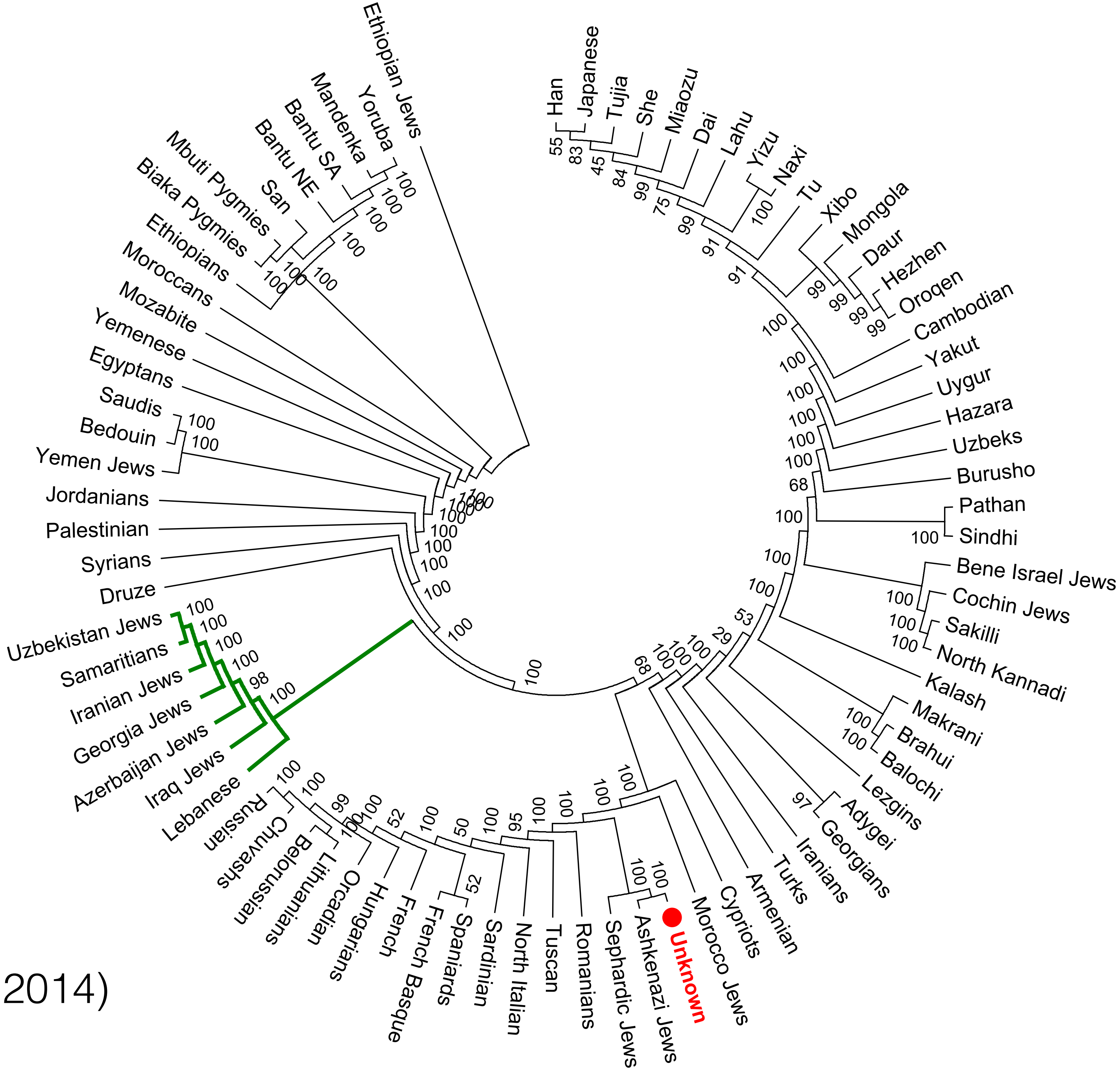


O'Connor et al. (2014)
Mol. Biol. Evol.

Cryptic Group has Similar Admixture Proportions to Jewish Groups.



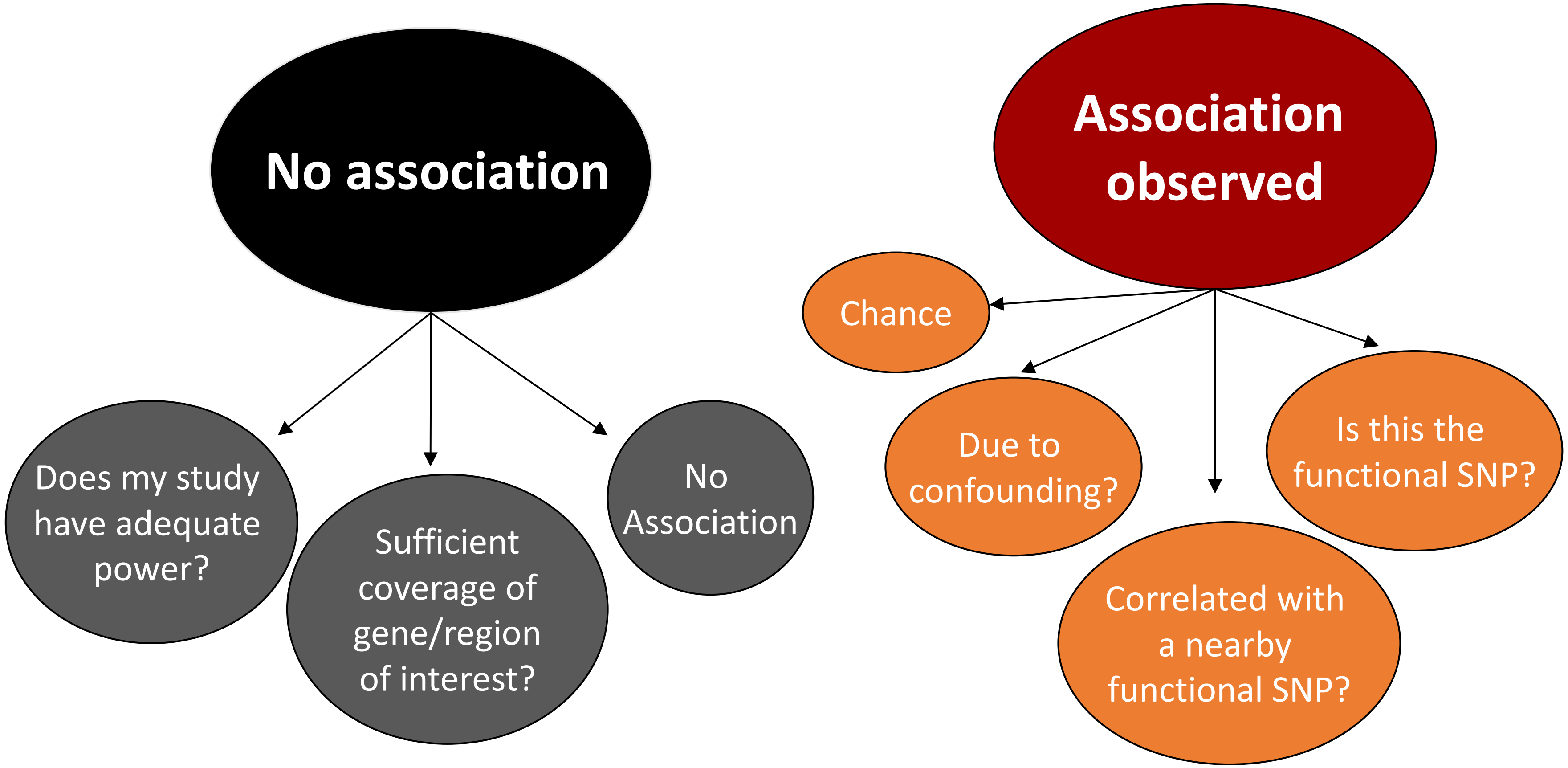
O’Connor et al. (2014)
Mol. Biol. Evol.



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

Genetic Association Studies - Applied

You completed an association study…
NOW WHAT?



No association

Does my study have adequate power?

Sufficient coverage of gene/region of interest?

No Association

Association observed

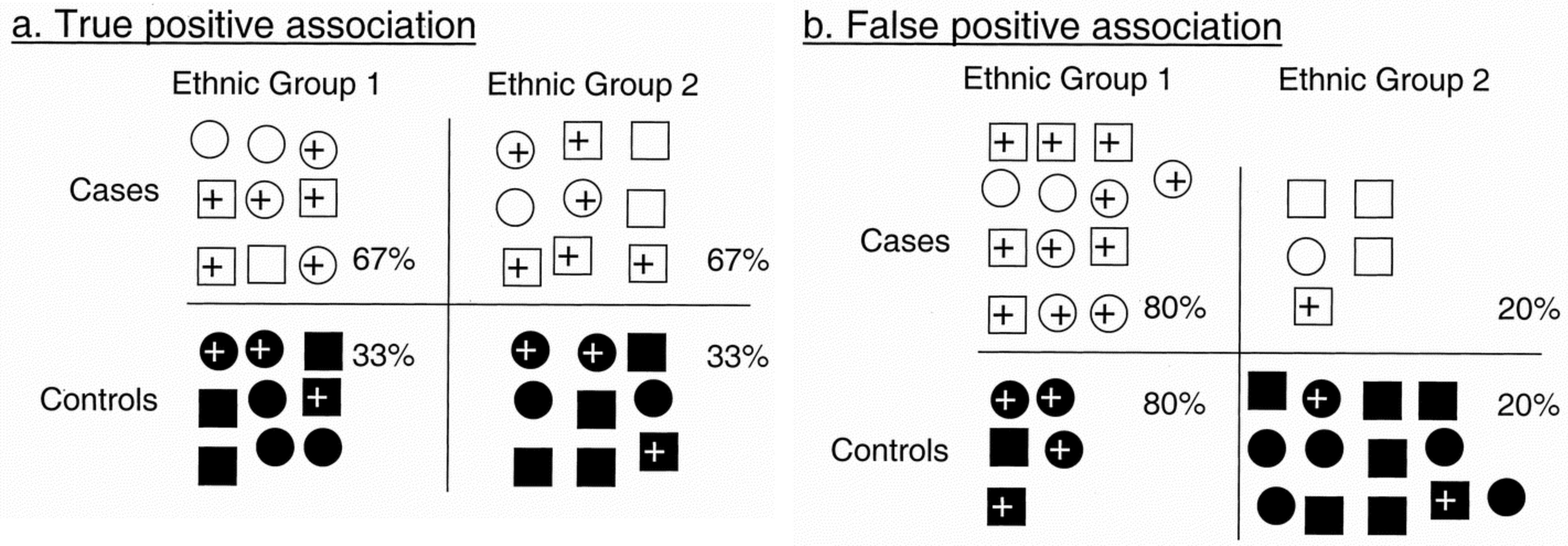
Chance

Due to confounding?

Correlated with a nearby functional SNP?

Is this the functional SNP?

Population Stratification - Concept



Hirschhorn (2002) Genetics in Medicine

Population Stratification – Example of spurious association

Population 1

% with disease: 10%

% with variant allele (A*): 20%

	Case	Cont.	
A*	2	18	20
G	8	72	80
	10	90	100

OR = (2 *72)/(8*18) = 1

Population 2

% with disease: 40%

% with variant allele (A*): 50%

	Case	Cont.	
A*	20	30	50
G	20	30	50
	40	60	100

OR = (20 *30)/(20*30) = 1

	Case	Cont.	
A*	22	48	70
G	28	102	130
	50	150	200

OR = (22 *102)/(28*48)
= 1.67

Population Stratification – thought question

The problem – poor sample matching. Cases and controls are not selected from the same source populations.

Population 1

% with variant allele (A*): 20%

Population 2

% with variant allele (A*): 50%

Is a scenario like this an issue for continuous traits vs. case-control analysis?

Determining Proportions of Case and Control

$P(i = X) =$ The proportion of the subpopulation i in the full population.

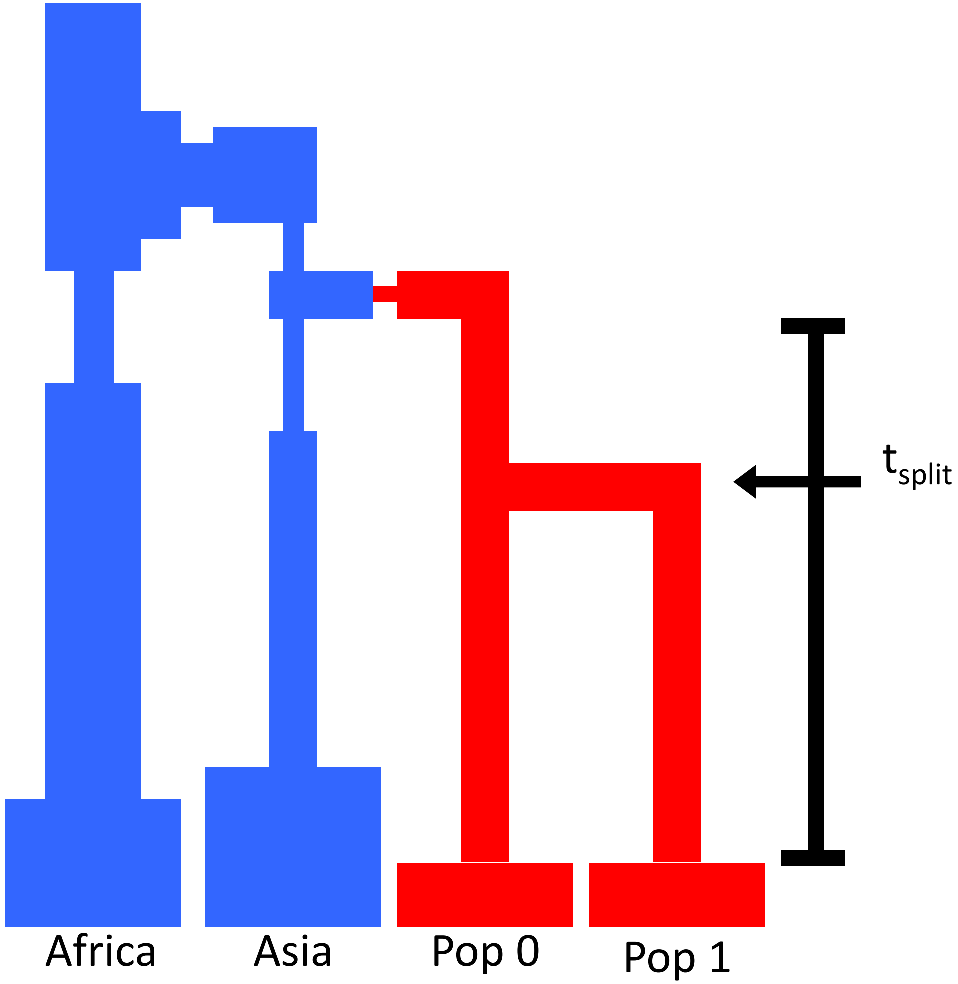
$P(d = c \mid i = X) =$ The probability of subpopulation i being a case (ie disease risk).

$$P(d = c) = \sum_{i=1}^{N} P(d = c \mid i = X) \times P(i = X)$$

$$P(i = X \mid d = c) = \frac{P(d = c \mid i = X) \times P(i = X)}{P(d = c)}$$

O'Connor et al. (2013) PLoS One

Two Population Case

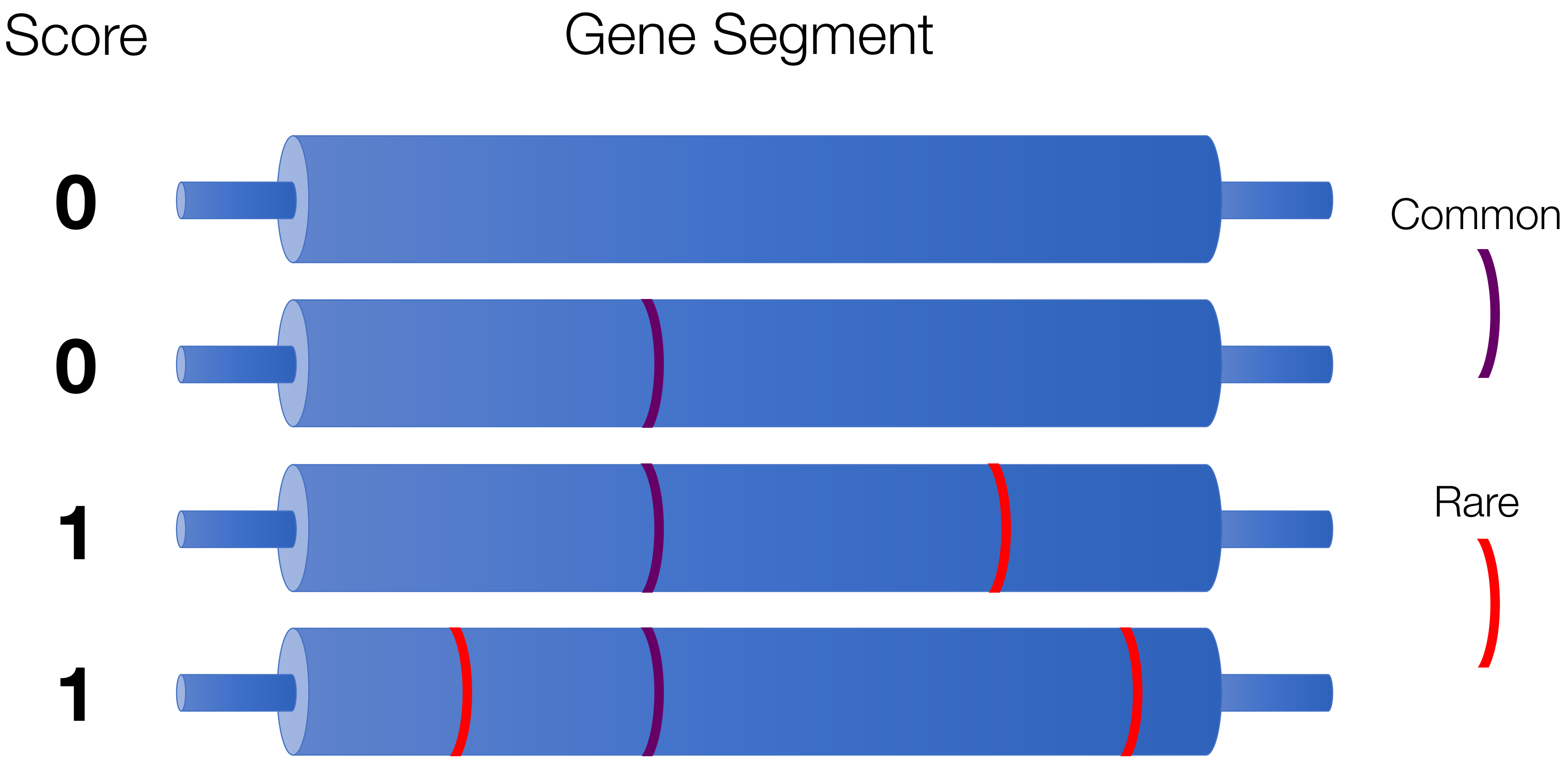


$$P(i = X) = 0.5$$

$$P(d = c \mid i = 0) = 0.02 + x$$
$$P(d = c \mid i = 1) = 0.02 - x$$

O'Connor et al. (2013) PLoS One

Collapsing Rare Variants



Score

Gene Segment

0

0

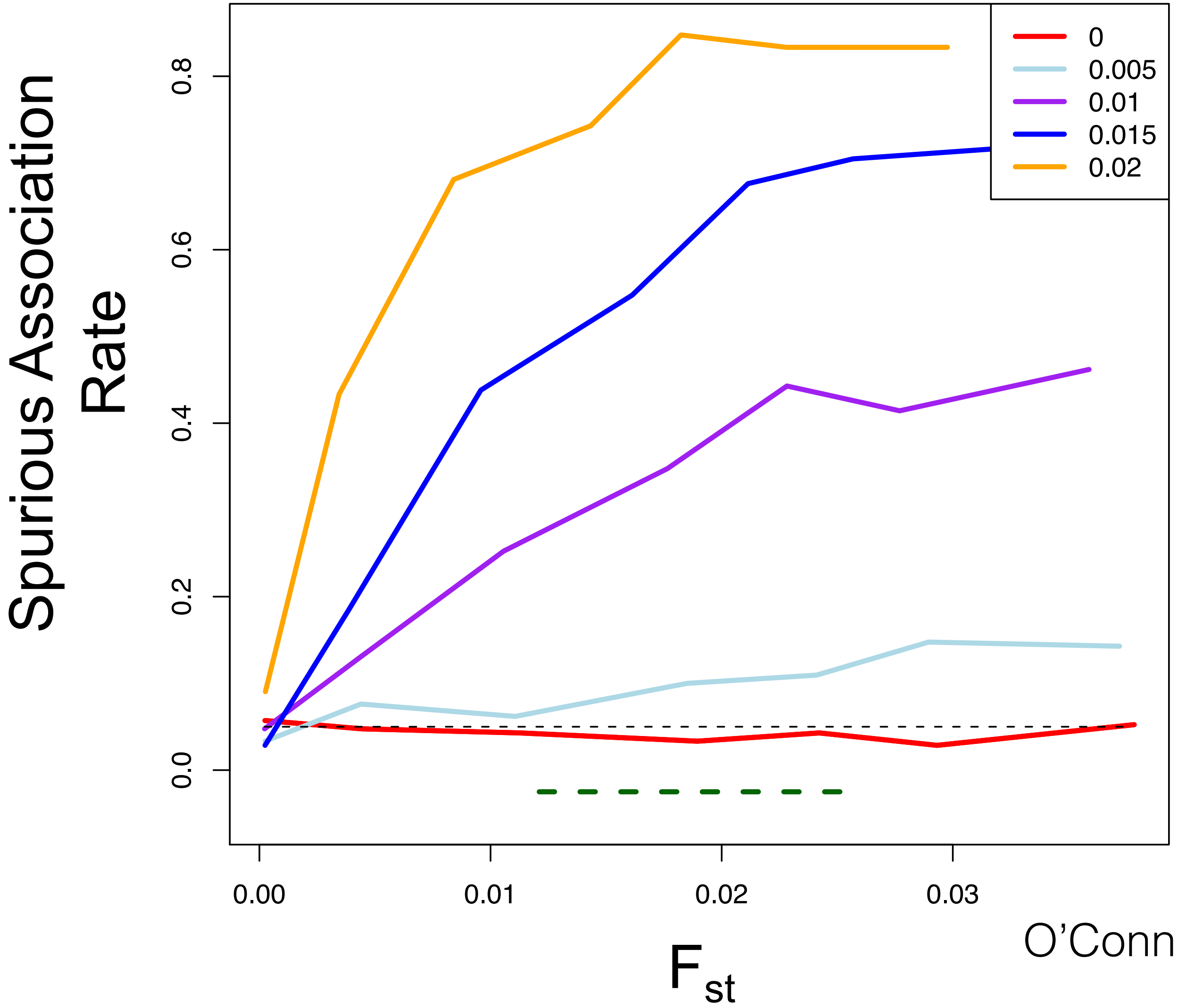
1

1

Common

Rare

Spurious associations as a function of confounding



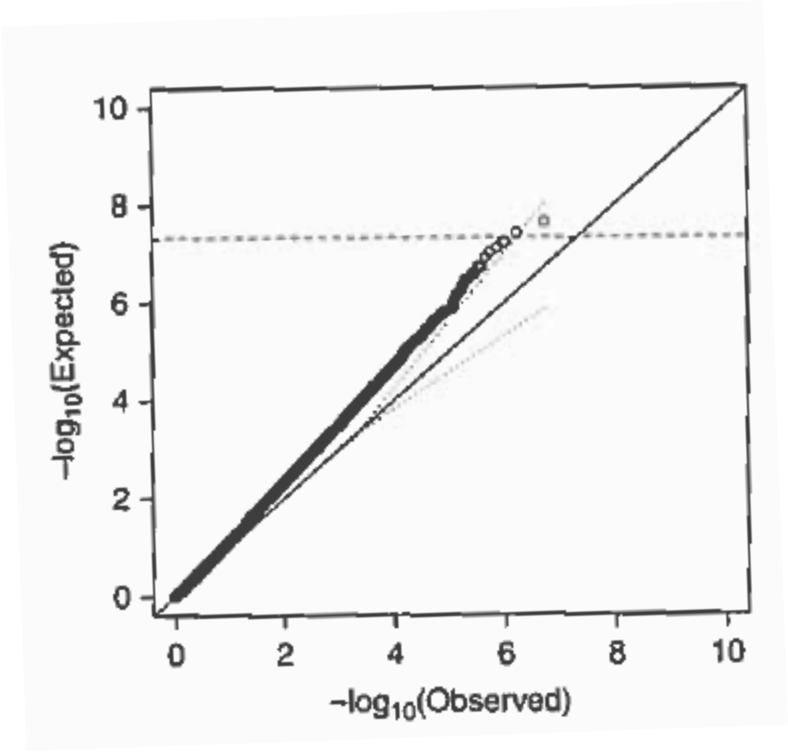
O'Connor et al. (2013) PLoS One

General Approaches to handle structure in association analyses

- Stratify by race – carefully match cases and controls.
- Control with family data
- Genomic control
- Control using genetic markers –
  - Ancestry informative markers (AIMs)
  - Principle components analysis

Genomic Control

- How inflated is this?



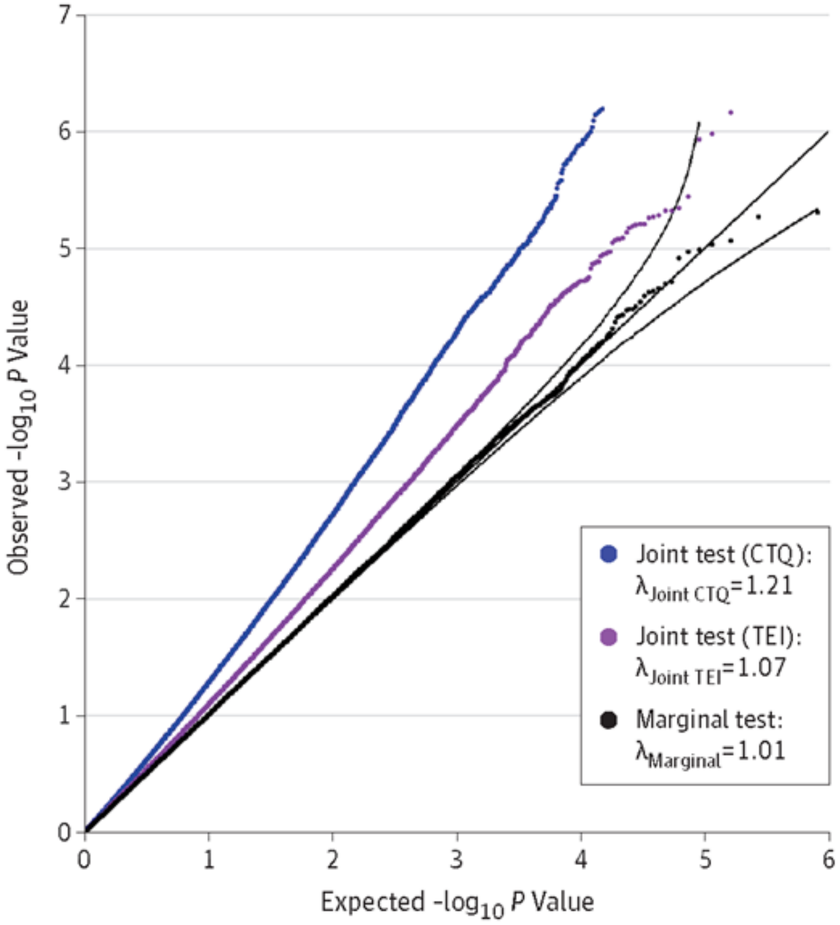
Genomic Control

- Assumption 1 - inflation affects the entire distribution of test statistics
- Assumption 2 – Underlying distribution of test statistics is χ^2 distribution
- Take median of observed test statistics
- Divide by expected =0.4549

  $\lambda = \chi^2_{\text{observed median}} / 0.4549$
- Divide all test statistics by λ

Austin. Genetic Epidemiology Methods & Applications 2013.

Genomic Control Cont.

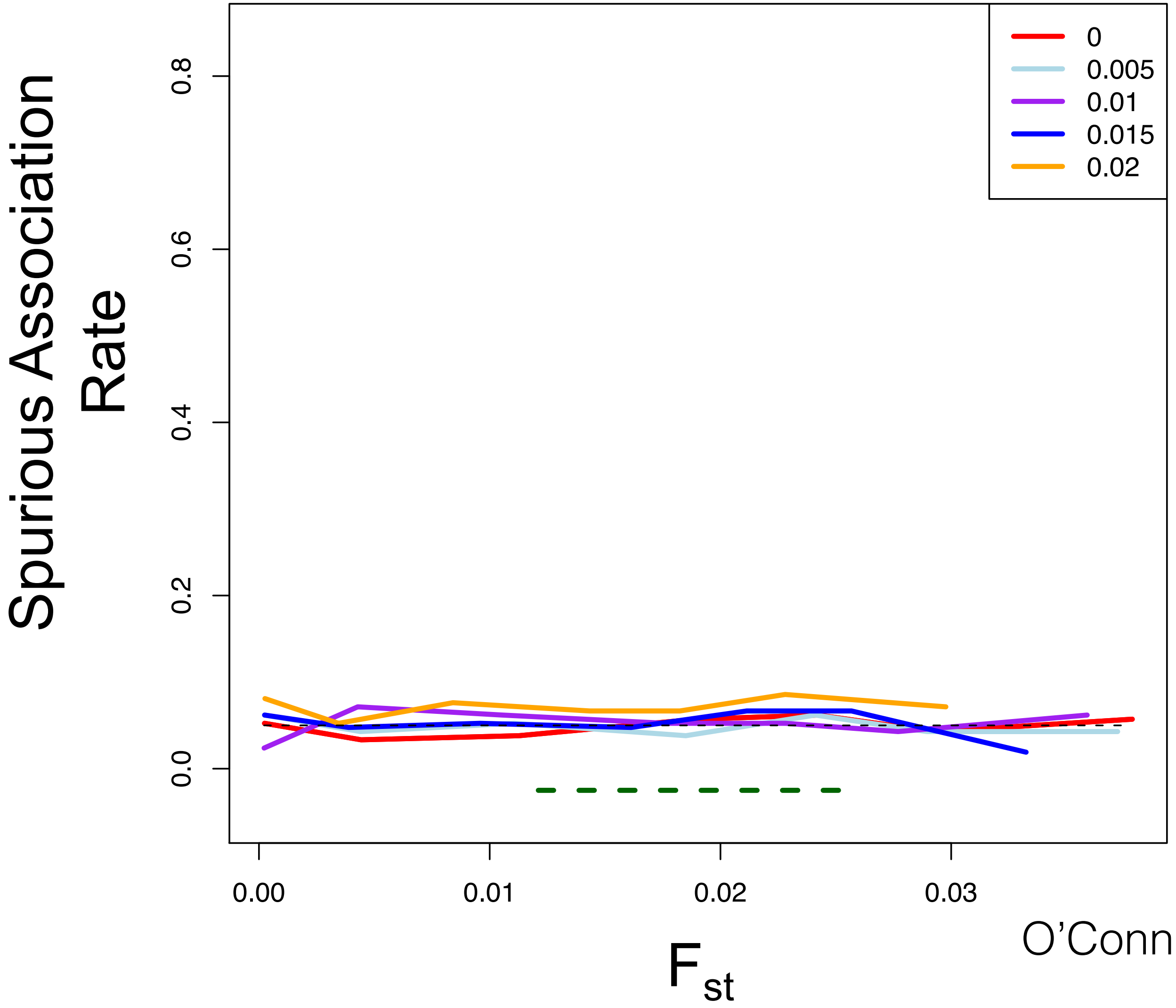


- Larger lambda → more systematic bias
- High lambda → consequence not cause
- Not only admixture (eg. cryptic relatedness)

AIMs and PCs

- General idea – Use genetic markers to measure and account for ancestry differences
- AIMs – fixed markers known to detect population substructure
- PCs – use large scale data (e.g. GWAS) to conduct principle components analysis
    - Dimension reduction
    - The first principal components summarize most of the variation
    - PCs can be used as variable in regression models

Spurious associations as a function of confounding with PC correction



O'Connor et al. (2013) PLoS One

Concluding summary

- Fine-scale population structure is subdivisions of individuals on an ever increasingly granular scale
- Cryptic population structure arises with extended relationships within a cohort, unknown to the investigators.
- Identity-by-descent and sharing of rare variants are a powerful method of identifying recent relationships and can be scaled by time.
- Cryptic relatedness can increase spurious associations for phenotype studies, but should be handled within the models.