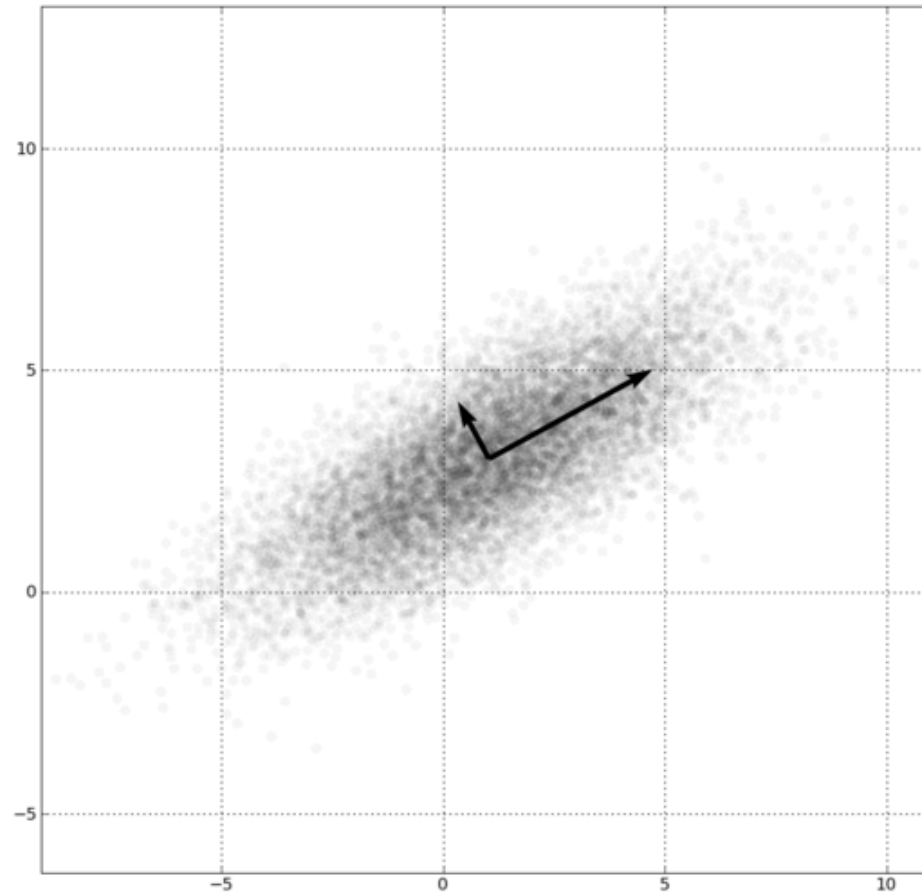


Population Structure Analysis

Learning objectives

- Methods to identify global estimates of population structure
 - Principal Component Analysis (PCA)
 - Admixture
- Local ancestry can identify segments of the genome corresponding to different ancestries.
- Local ancestry can be applied in a number of different ways
 - Demographic modeling
 - Selection
 - Refining PCA signals
 - Association analyses

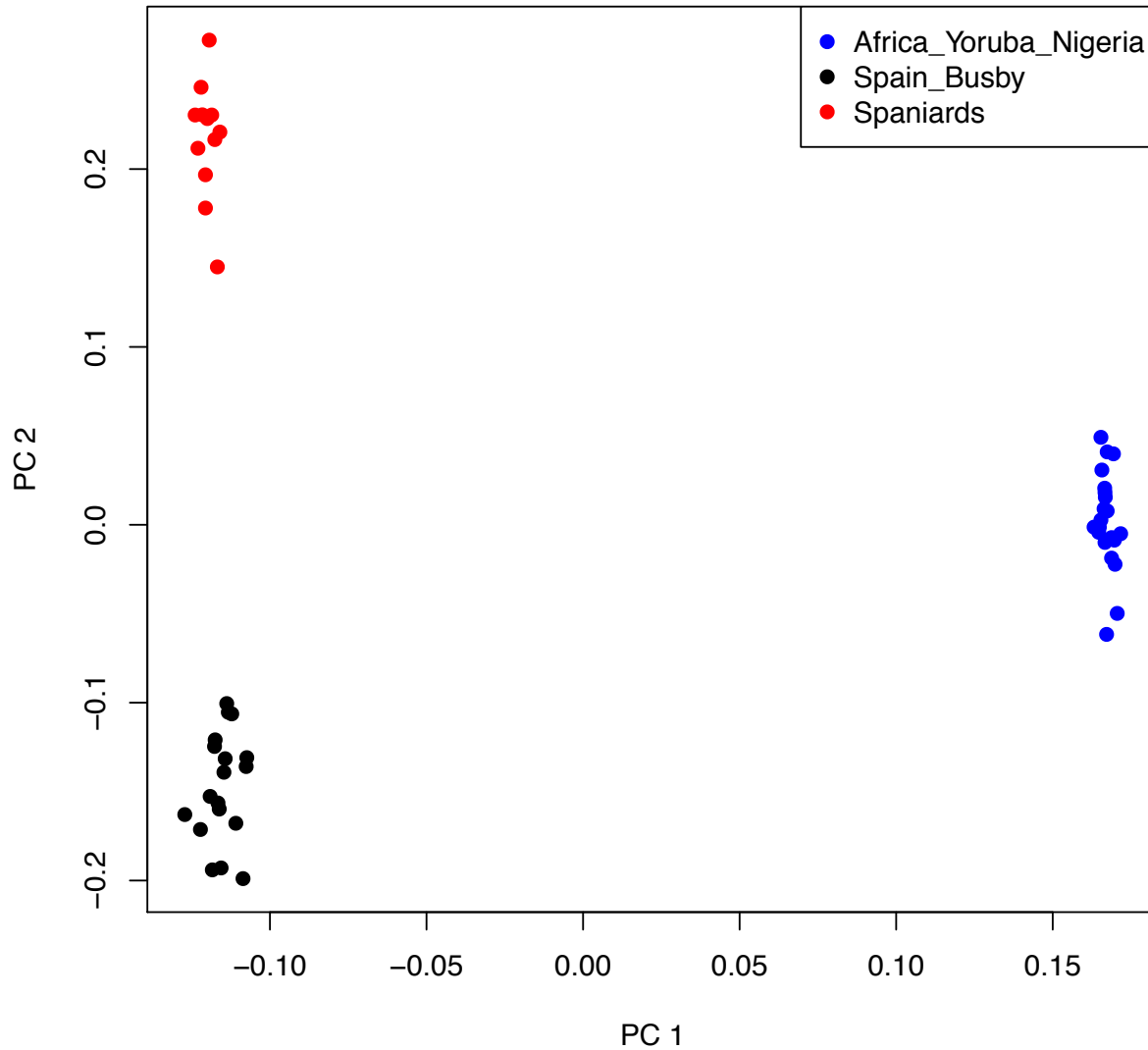
Principal Component Analysis (PCA)



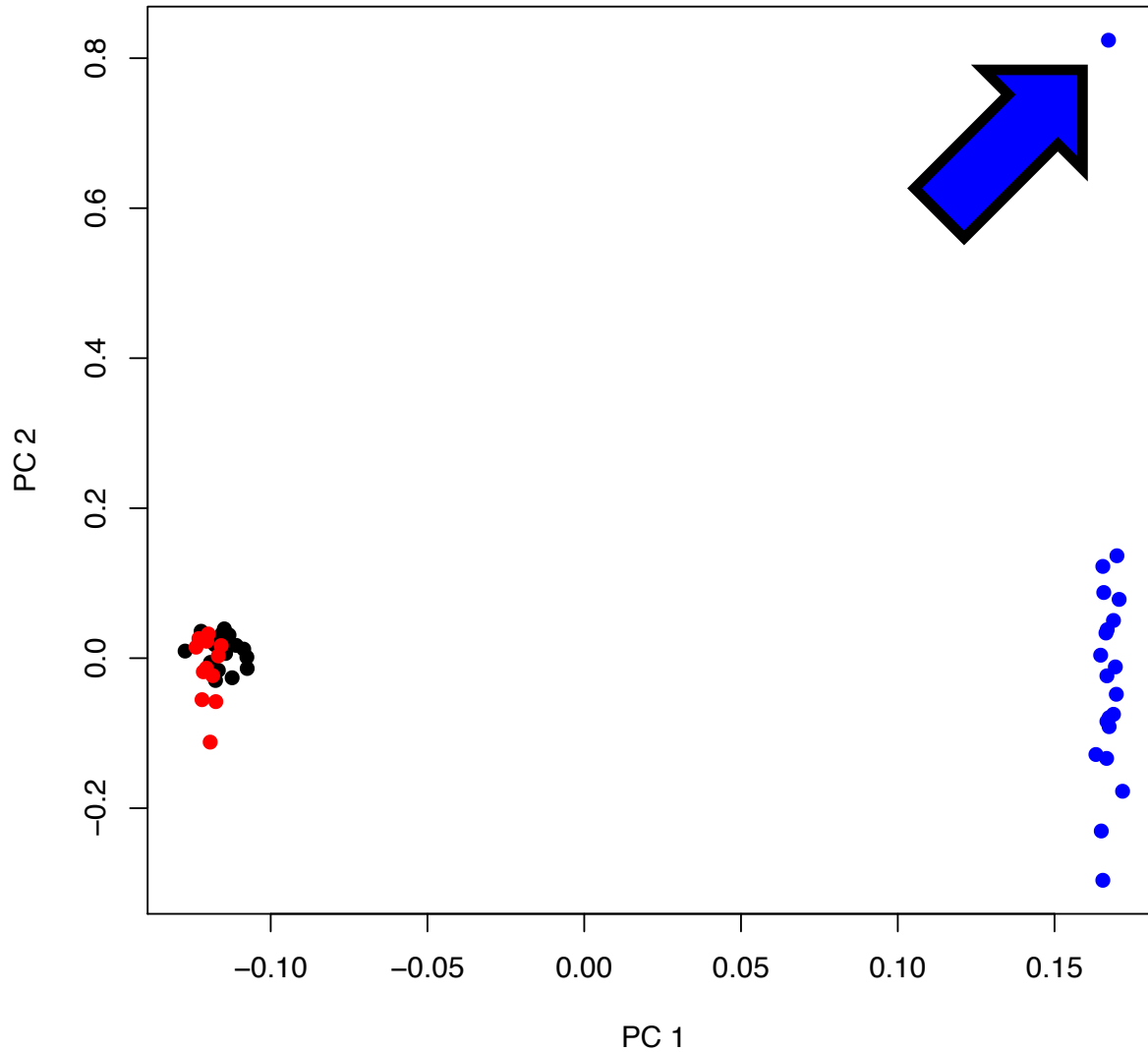
PCA

- Uses
 - Highly sensitive summary of all the data
 - Summarize population structure
 - Identify groups within data
 - Sanity check for study design
 - E.g. Diseased individuals cluster vs controls
 - Sanity check when combining data
- Pitfalls
 - Only look at the first few PCs
 - All axes are biological (once first few are)
 - Identifying significance of an axis is non-trivial
- Assumptions
 - Linear relationship between data
 - Variants are independent (LD)

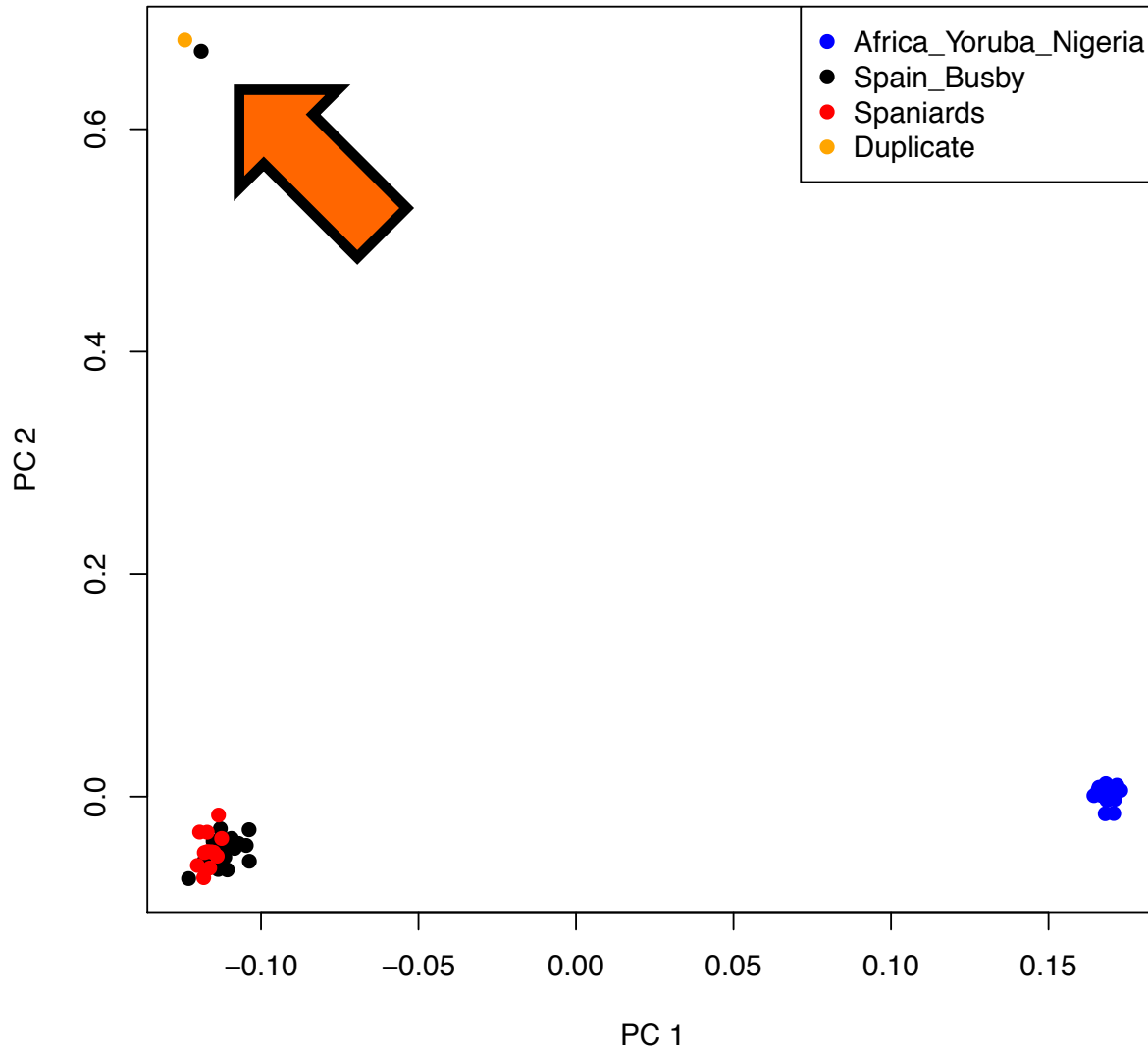
PCA Example: Strandedness



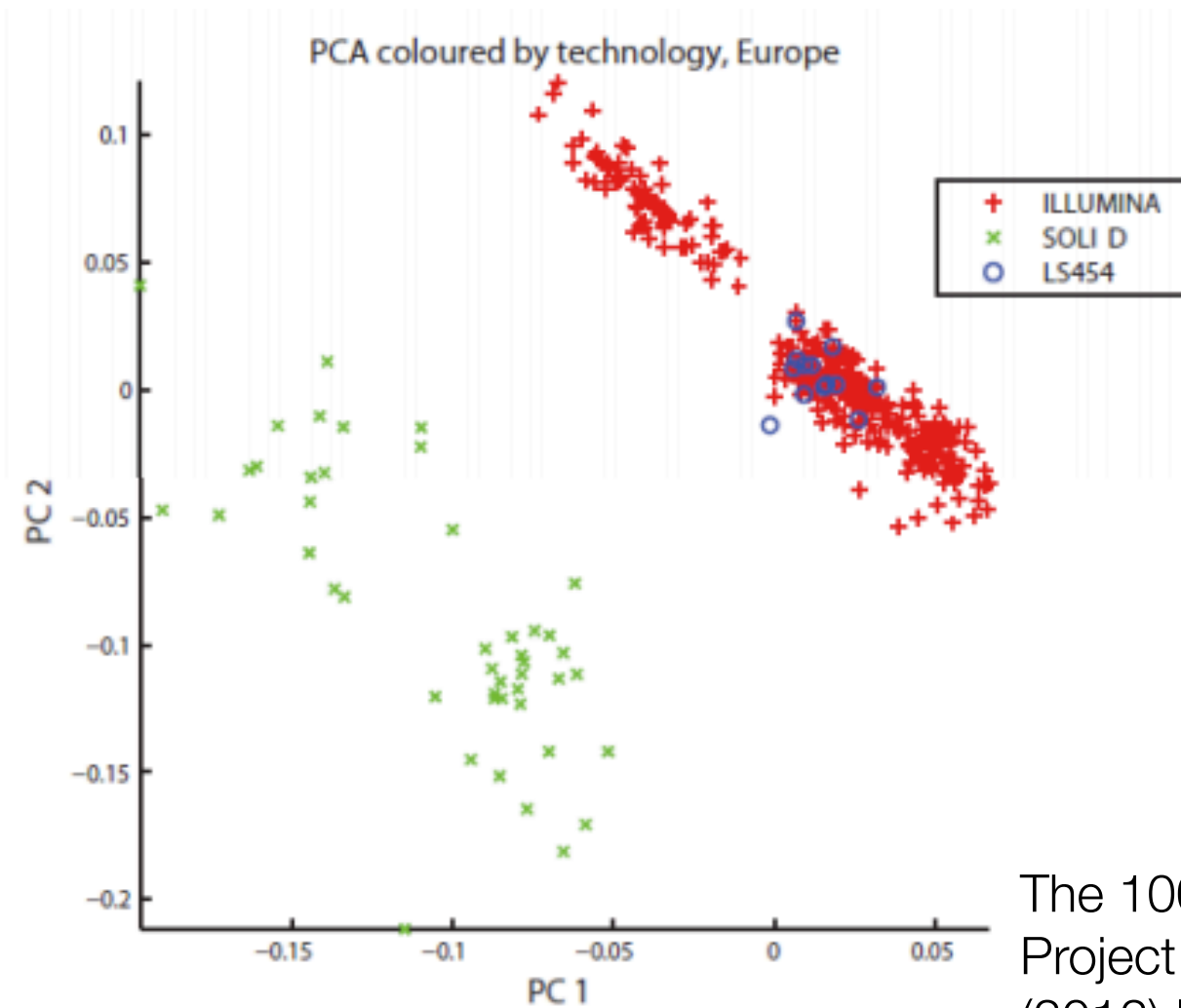
PCA Example: Insignificance



PCA Example: Relatedness

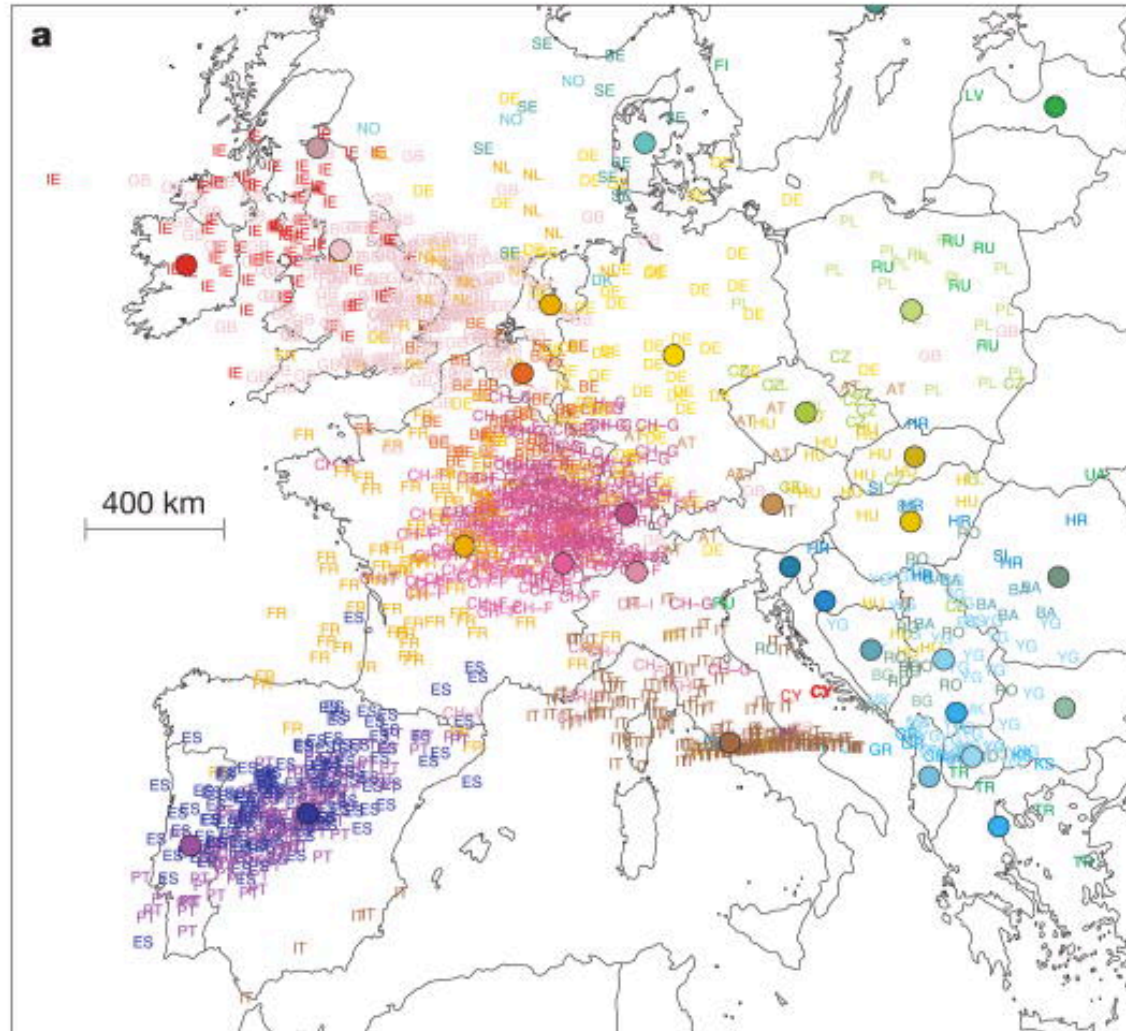


PCA Example: Technical Issues



The 1000 Genomes
Project Consortium
(2012) Nature

“Genes mirror geography within Europe”



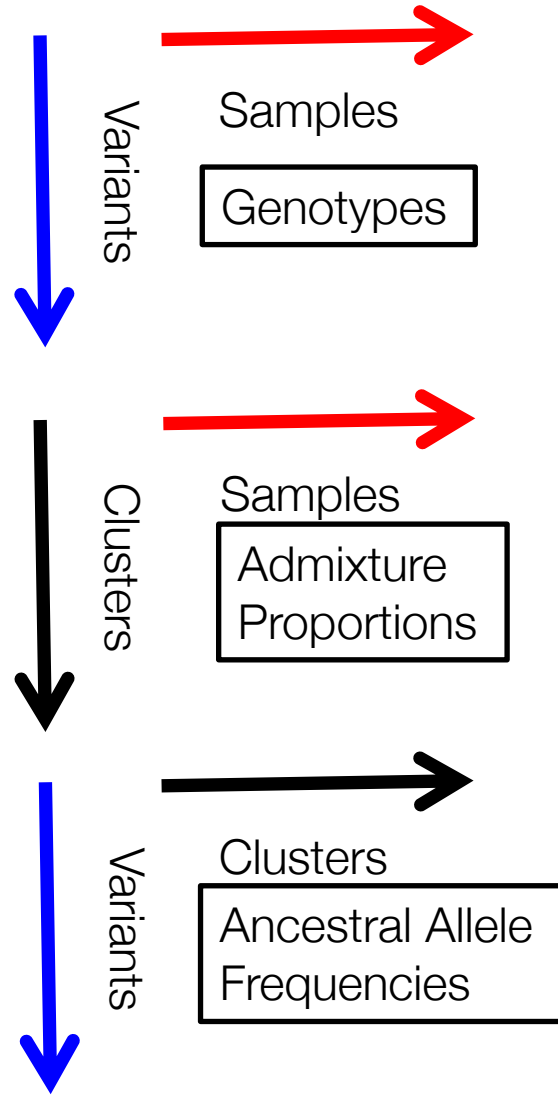
Novembre et al.
(2008) Nature

ADMIXTURE (Alexander et al. 2009)

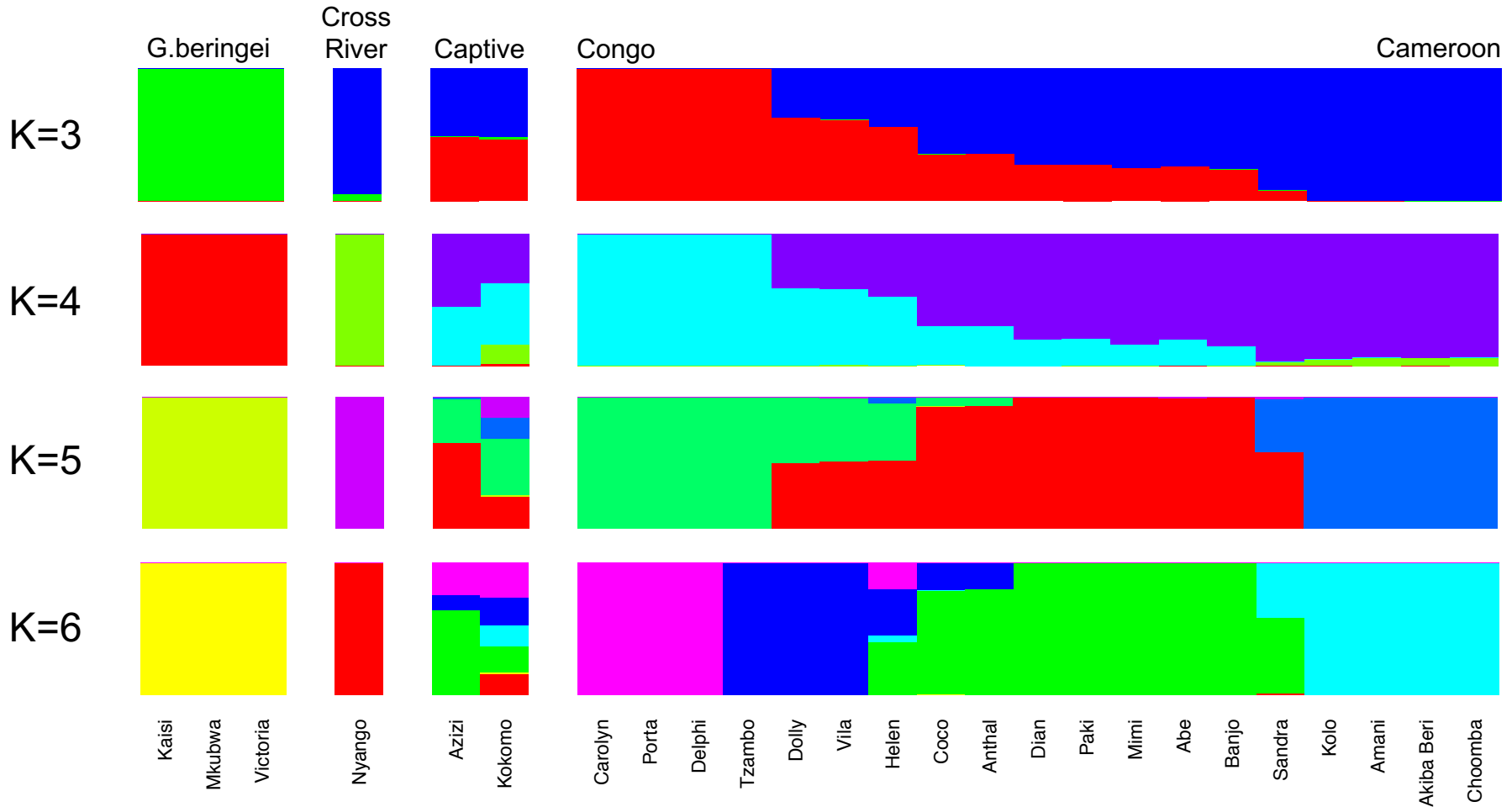
$$G = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1N} \\ g_{21} & g_{22} & \cdots & g_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ g_{M1} & g_{M2} & \cdots & g_{MN} \end{bmatrix}$$

$$Q = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1N} \\ q_{21} & q_{22} & \cdots & q_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ q_{K1} & q_{K2} & \cdots & q_{KN} \end{bmatrix}$$

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M1} & p_{M2} & \cdots & p_{MK} \end{bmatrix}$$



Admixture analyses



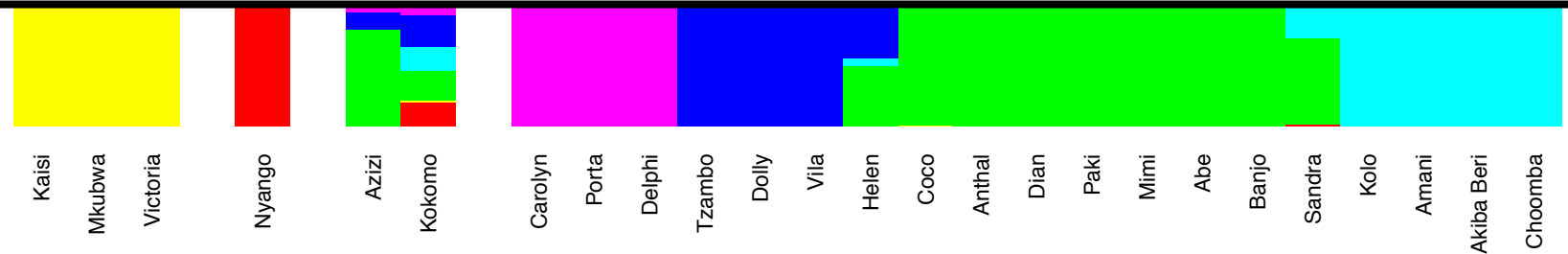
Admixture analyses: when is the K correct?



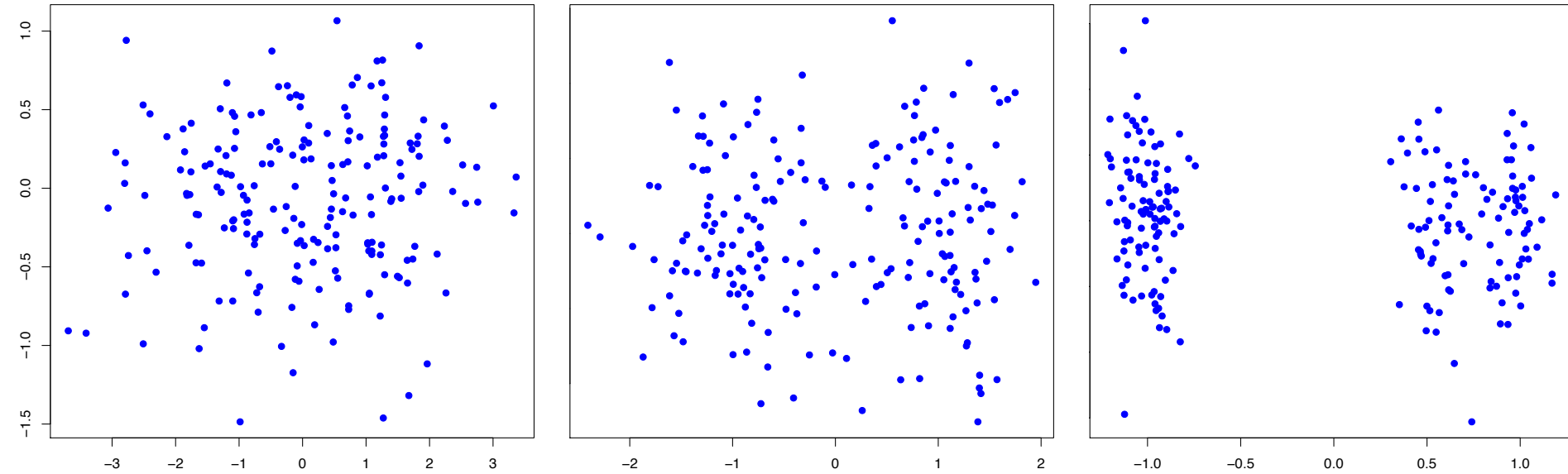
“In practice, people often try different K, and choose the K that makes most biological sense.”

-Frappe Manual

K=6



The K Problem



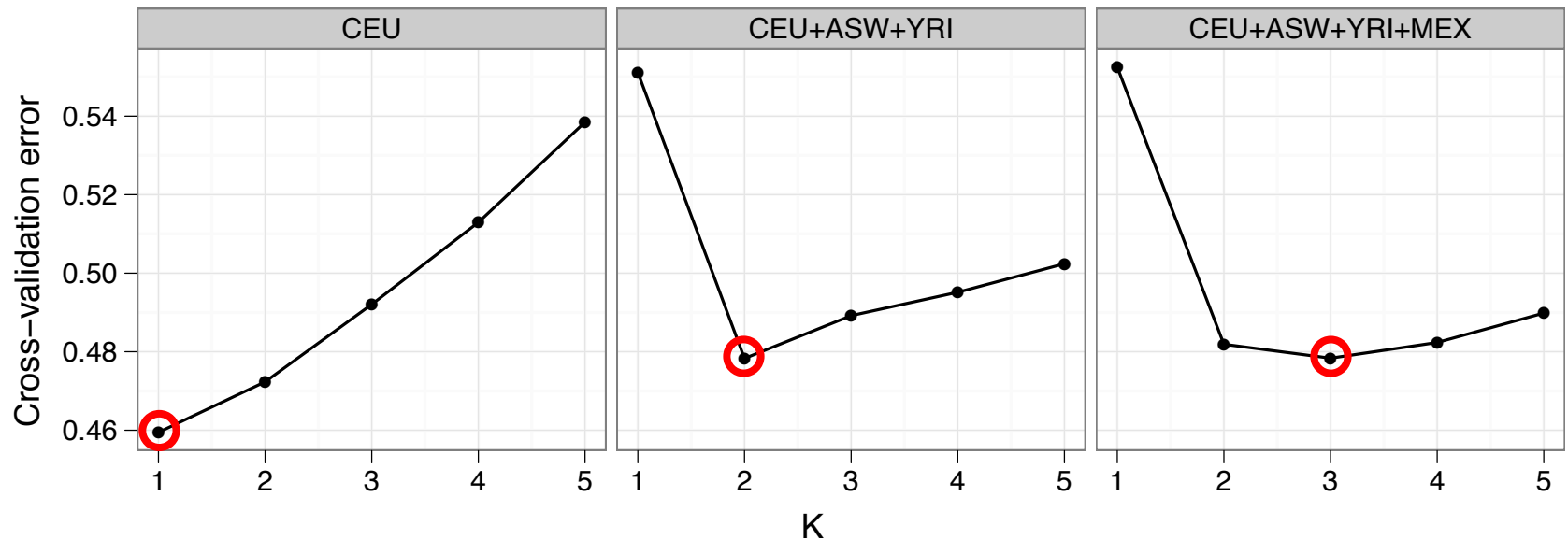
How many different means are there?

ADMIXTURE: using cross validation to identify the best K

$$G = \begin{bmatrix} g_{11} & \cancel{g_{12}} & \cdots & g_{1N} \\ \cancel{g_{21}} & g_{22} & \cdots & g_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ g_{M1} & g_{M2} & \cdots & \cancel{g_{MN}} \end{bmatrix}$$

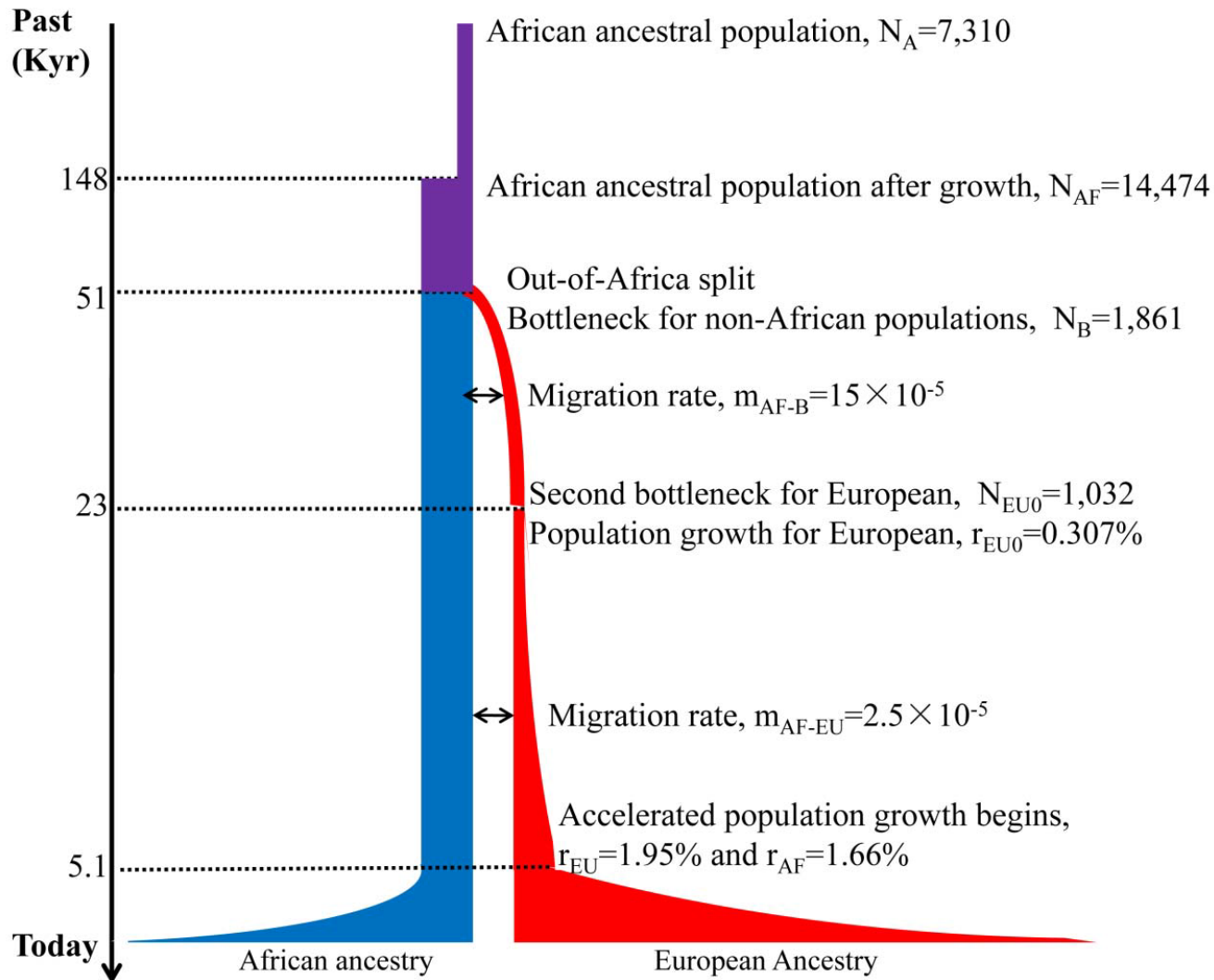
$$\hat{g}_{li} = 2 \sum_{k=1}^K p_{lk} \times q_{ki}$$

How well X-validation performs

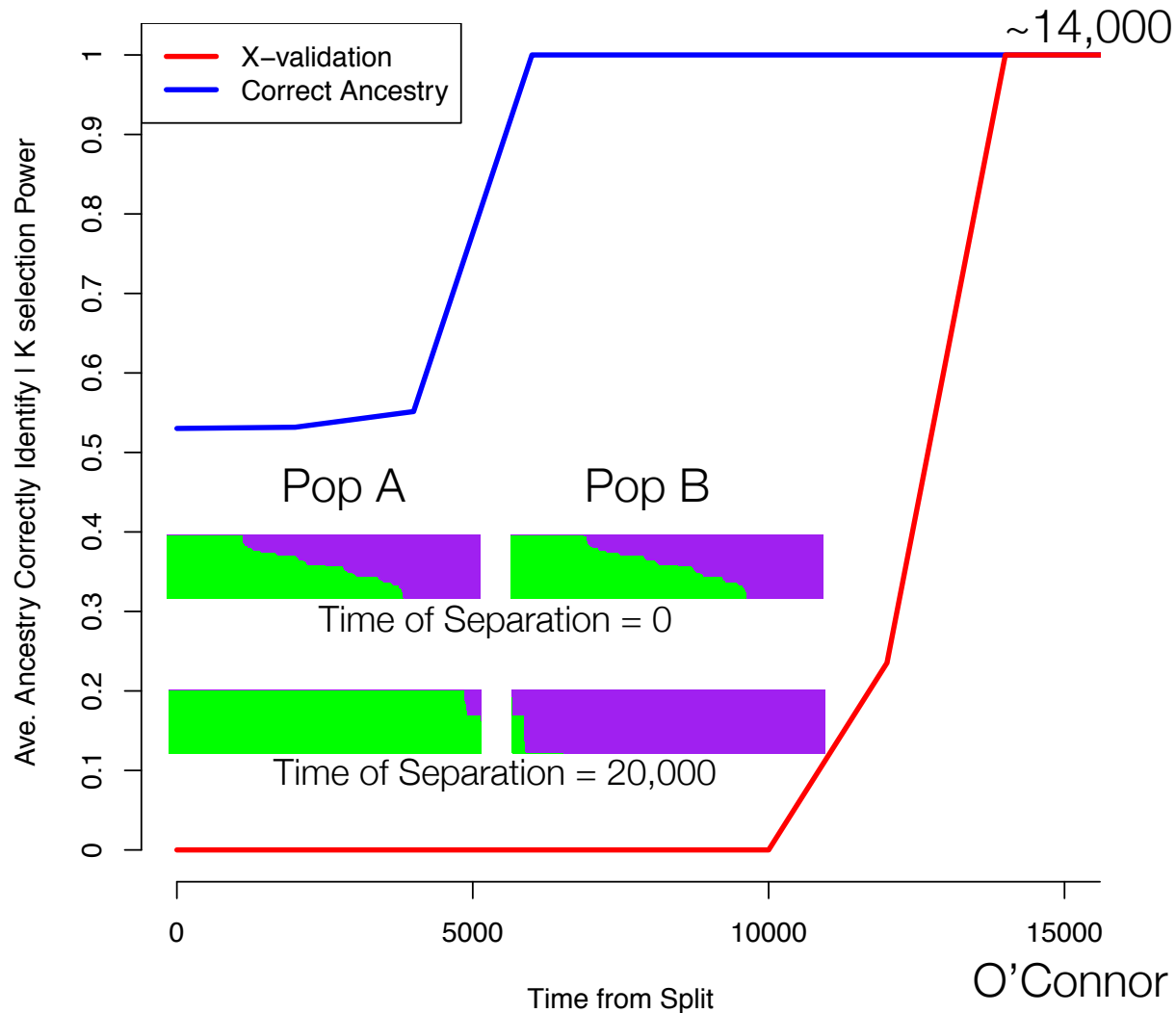


Alexander and Lange
(2011) BMC Bioinformatics

Test it with ESP inspired simulations



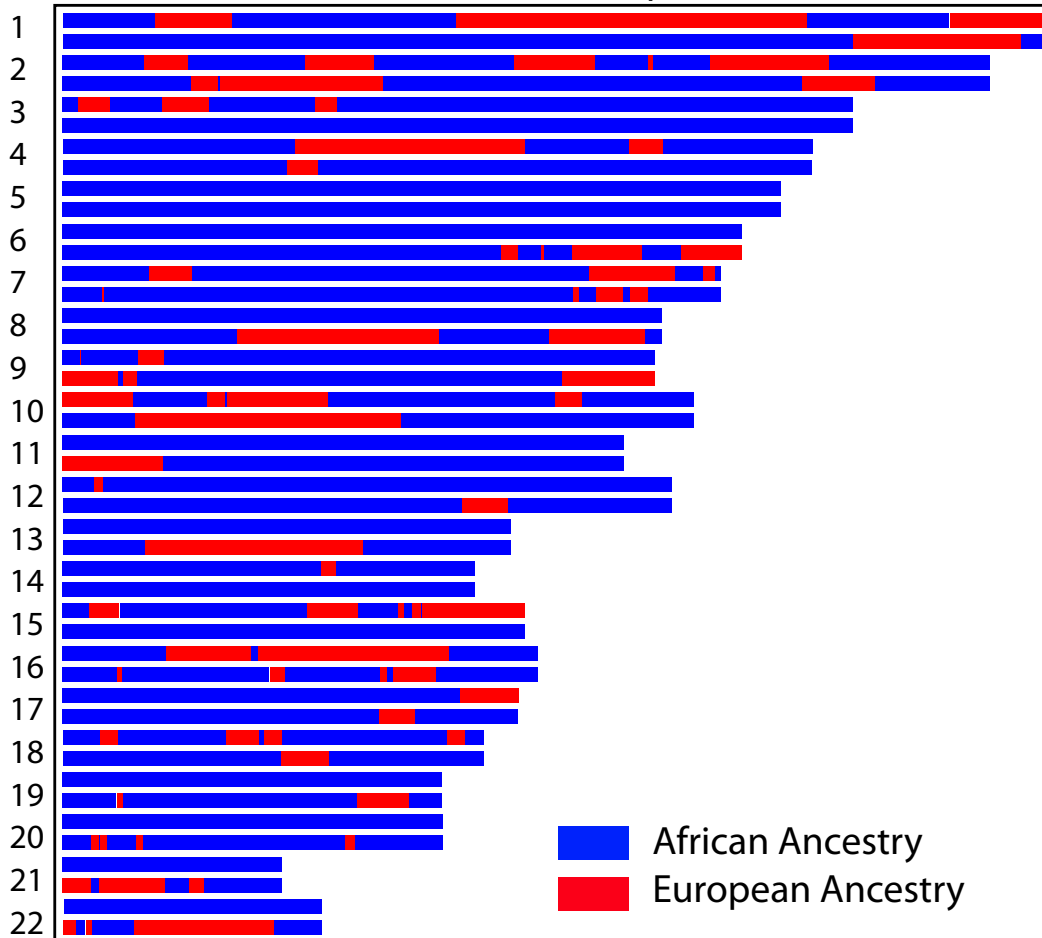
X-validation's performance as a function of split time



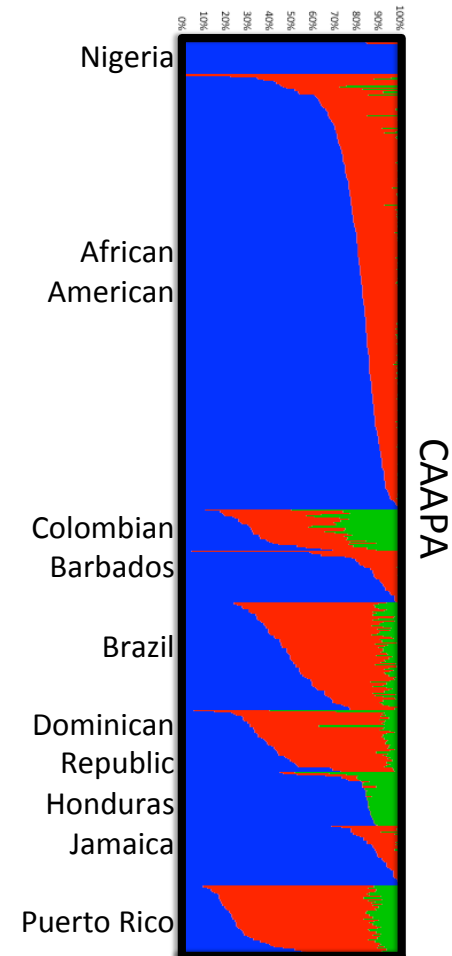
Tricks to effectively use ADMIXTURE

- This is a Maximum Likelihood framework with many parameters
 - Run multiple times (I usually use >10) for each K taking the best log-likelihood (an output parameter).
 - This deals with local minimum problems.
- Sometimes the lowest K that has X-validation identifies is less than what we thought. Though this is possible (see previous power figure), it doesn't mean we have objective evidence other than the K it found.
- Sometimes we get greater K than we expect or can explain. In such situations it might be better to move to a supervised learning version (also available in ADMIXTURE).

Local vs Global Ancestry

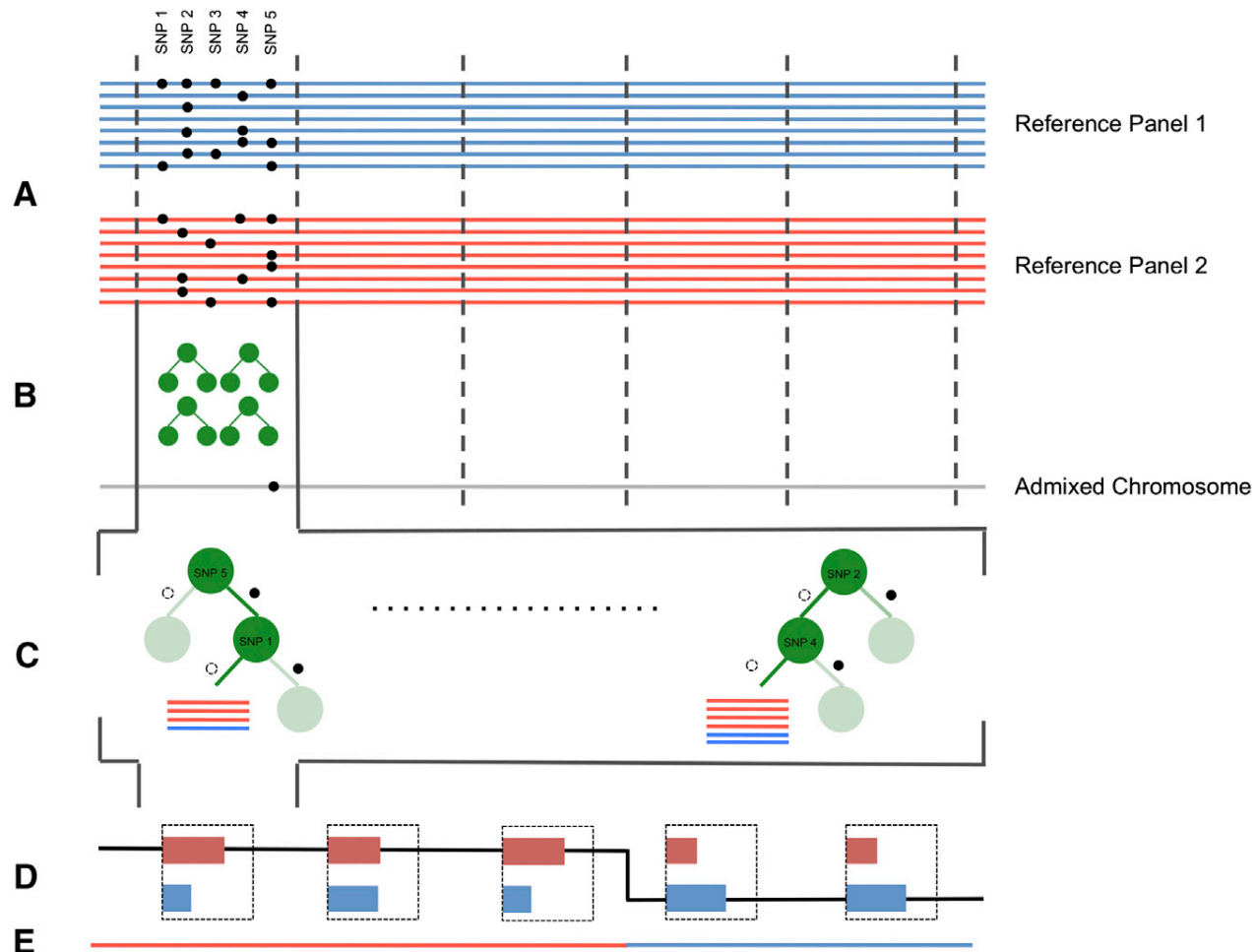


Gravel et al. (2013) Genetics

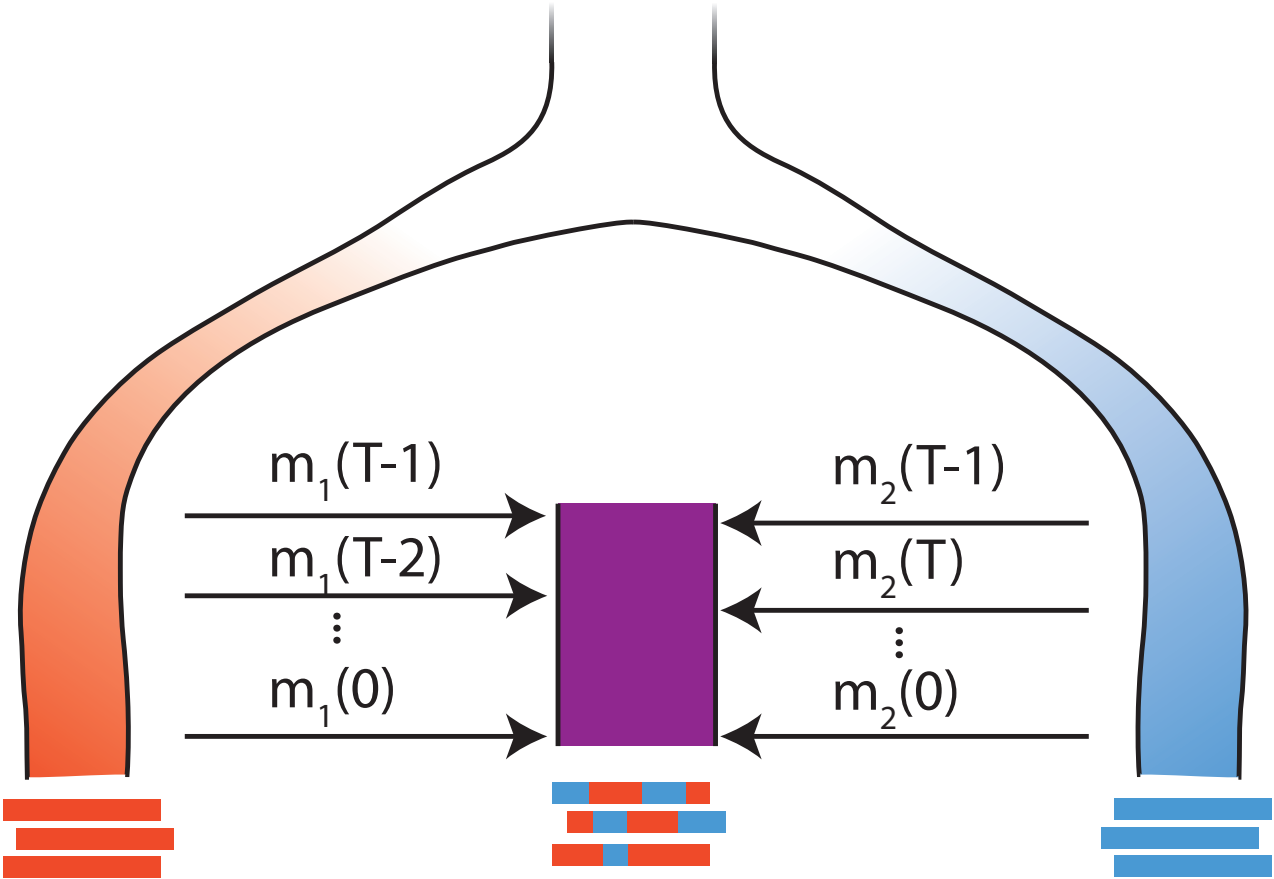


Mathias et al. (2016) Nat. Comm.

Local ancestry calling: RFMix as an example



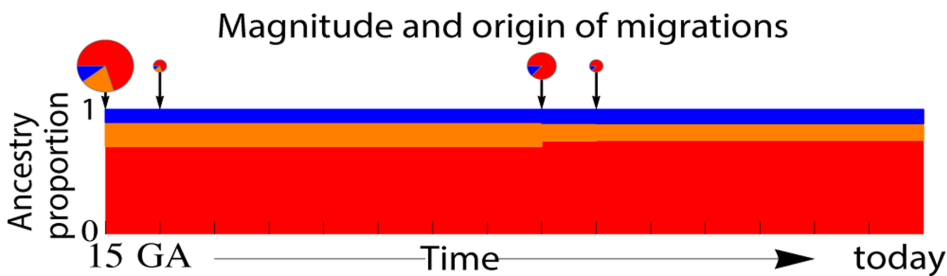
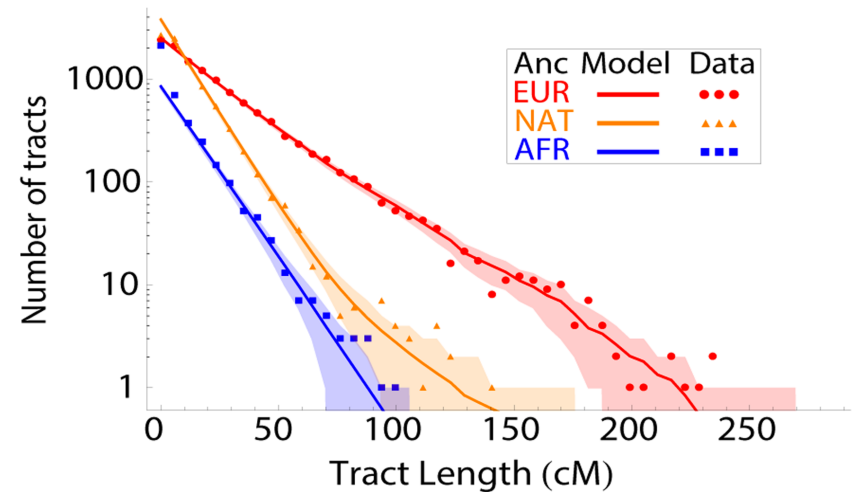
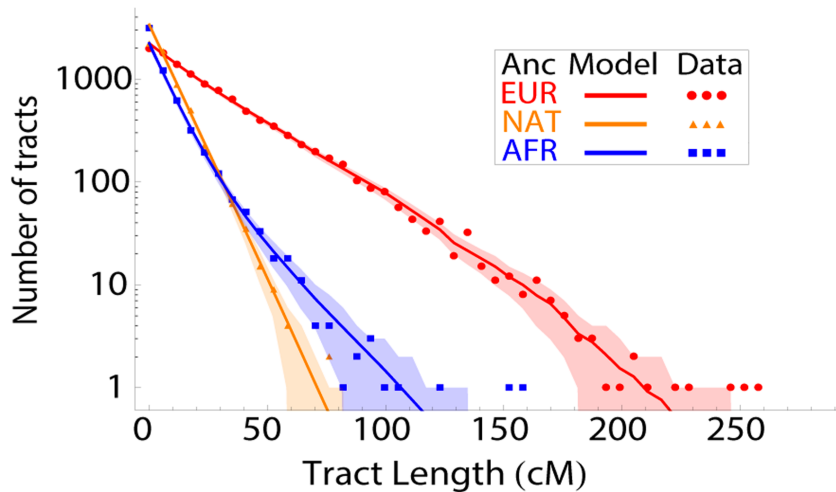
Demographic modeling with local ancestry



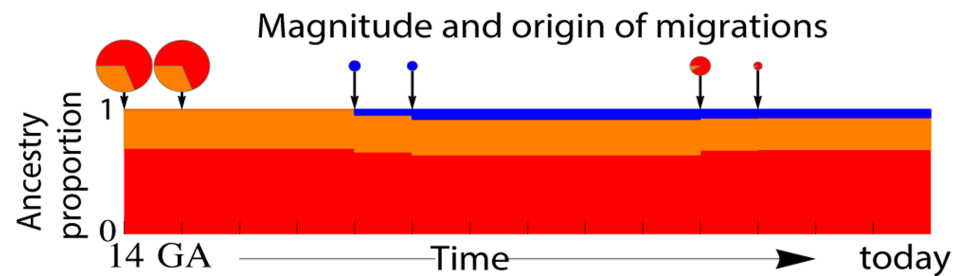
Demographic modeling with local ancestry

PUR

CLM

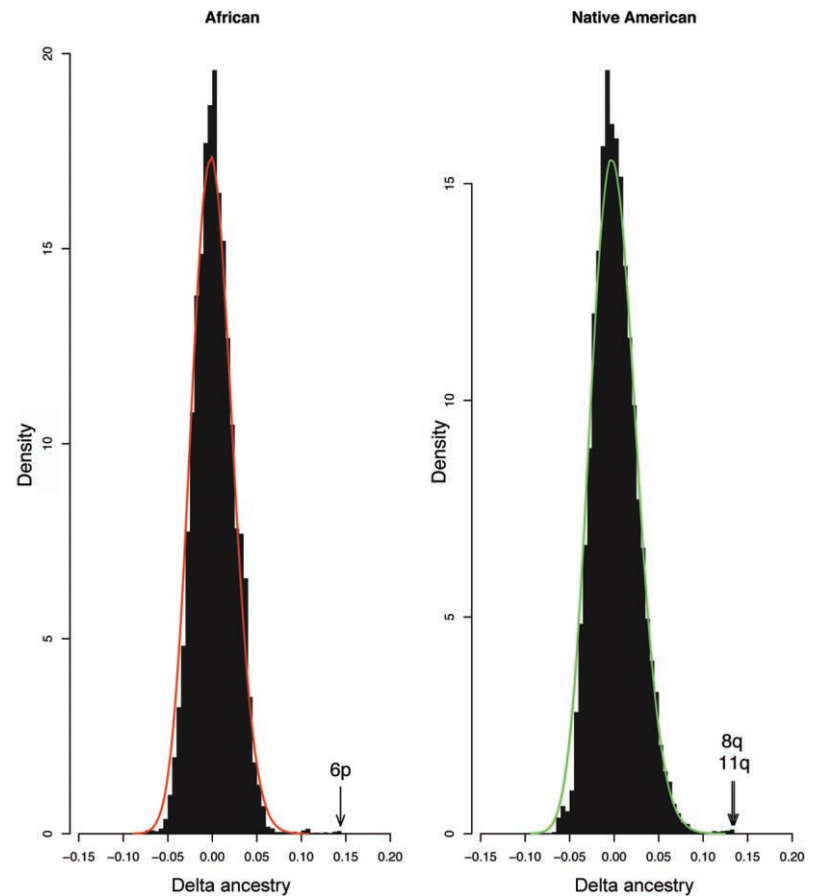
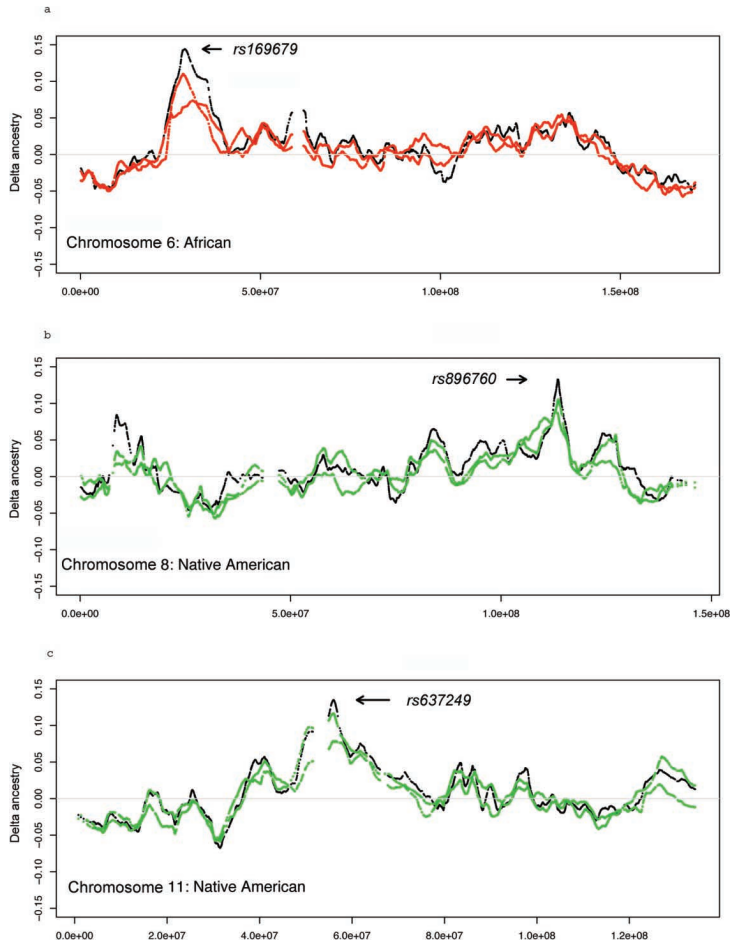


(a)



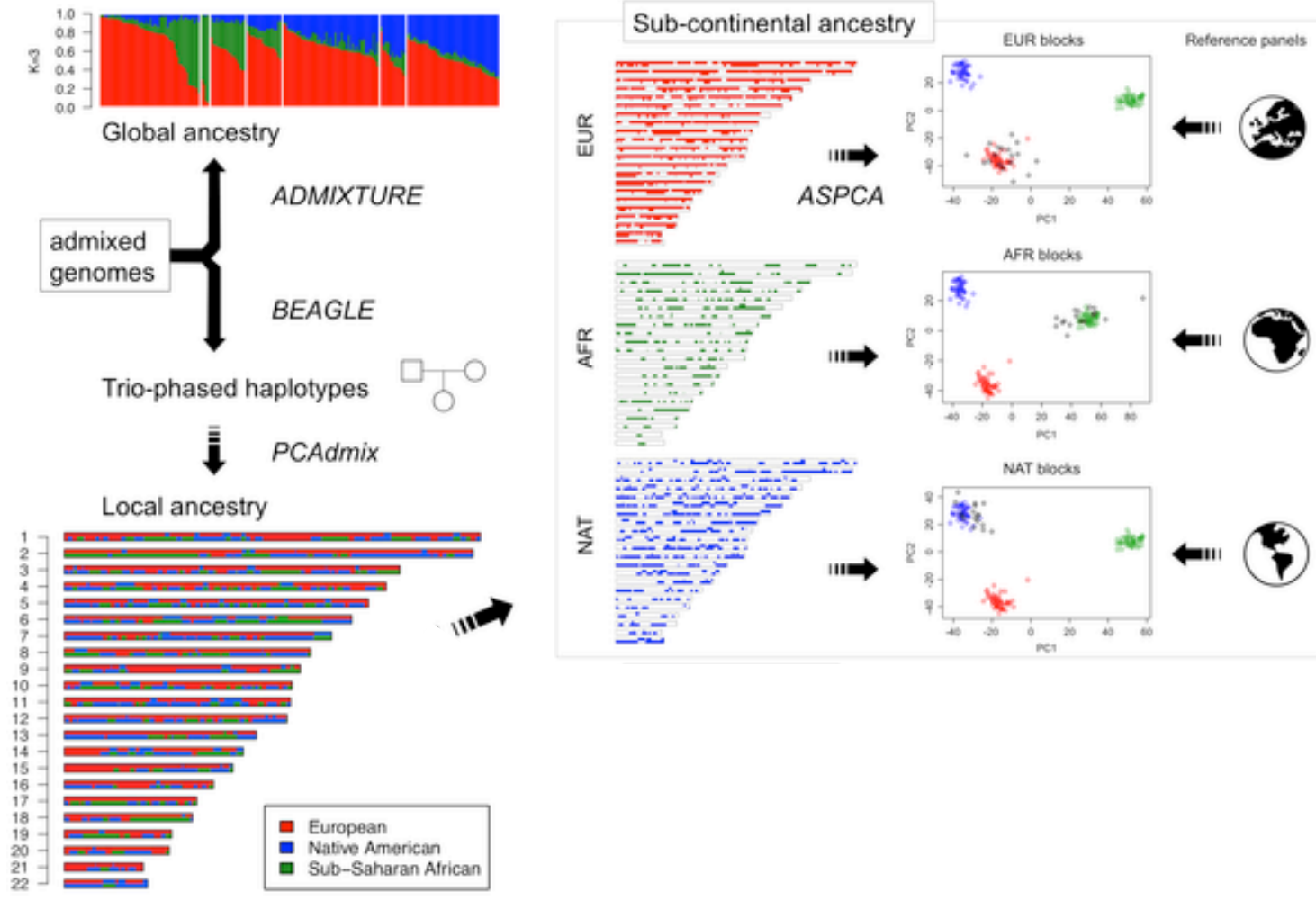
(b)

Recent selection by looking for local ancestry biases

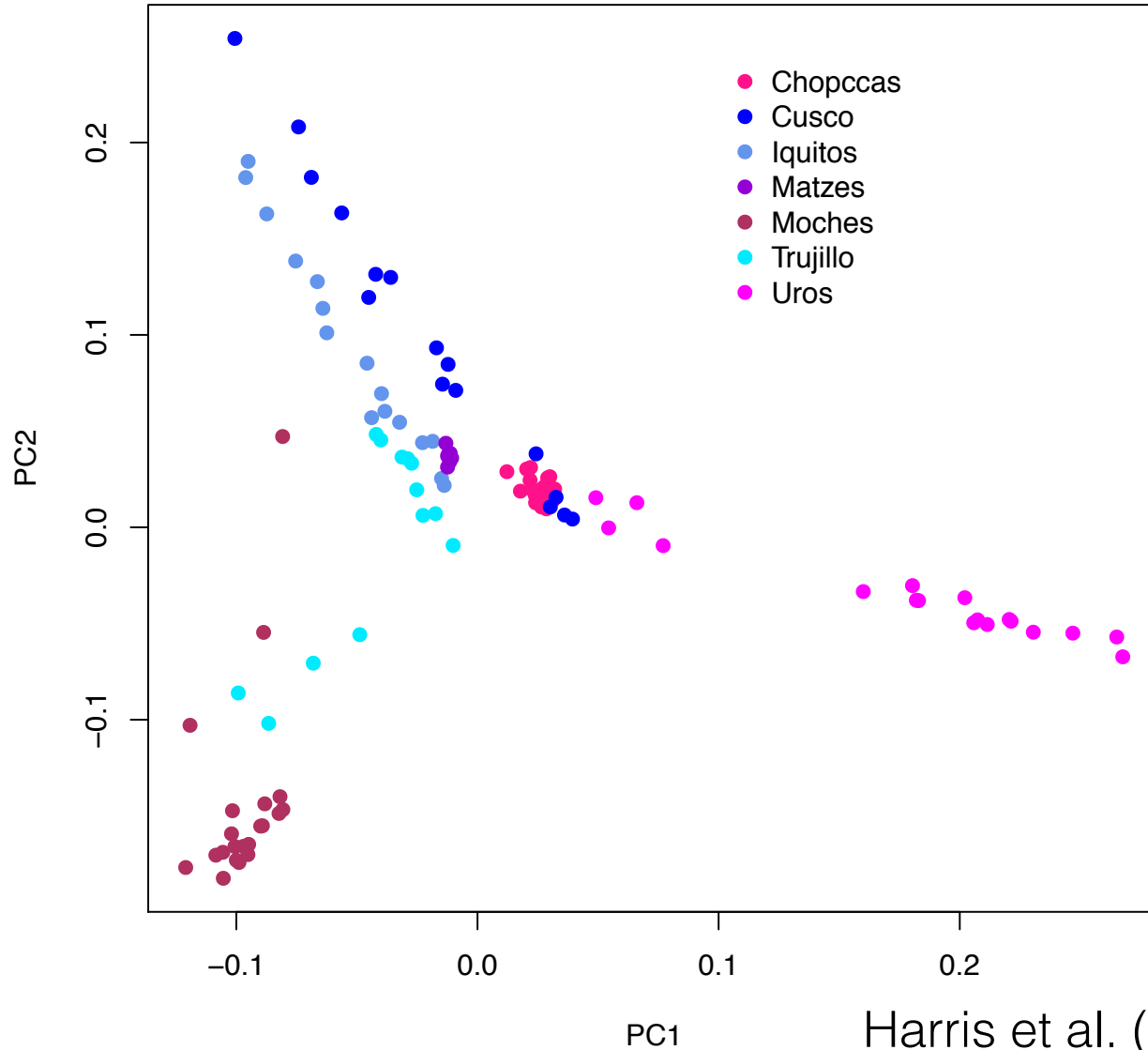


Tang et al. (2007) AJHG
Though see Bhatia et al. (2014) AJHG

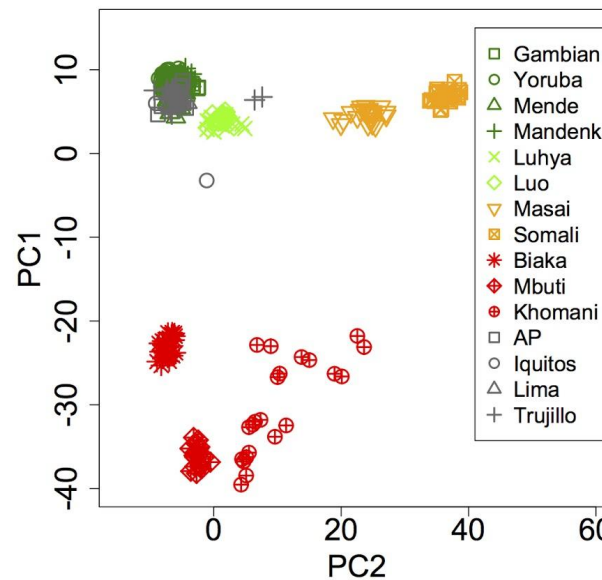
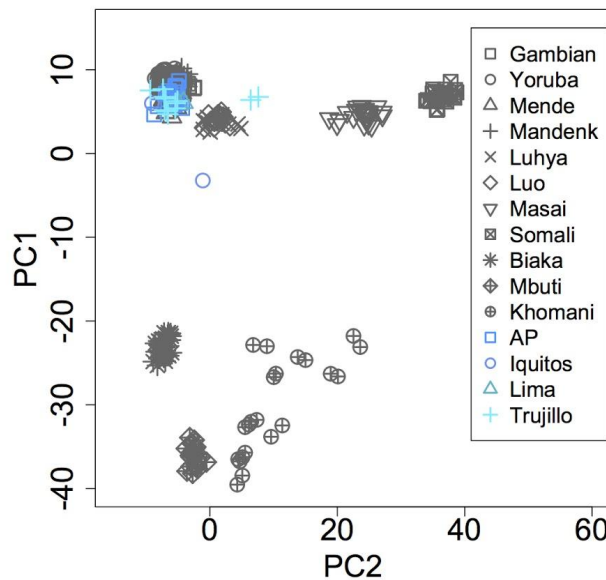
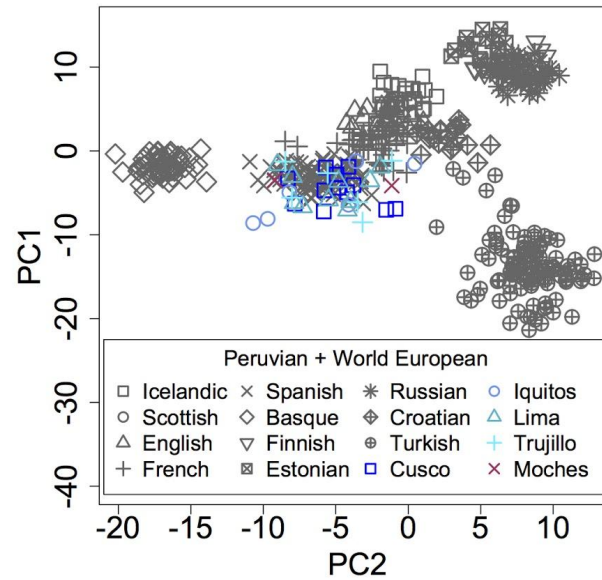
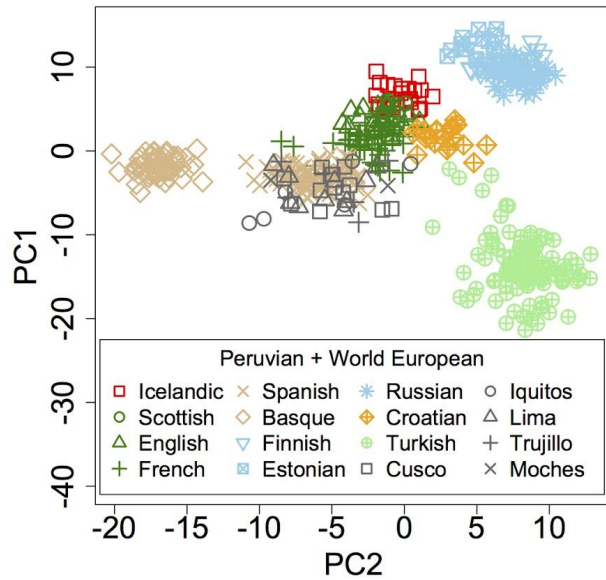
Combining Local Ancestry and PCA to give Ancestry Specific PCA (or ASPCA)



Peruvian population structure with PCA

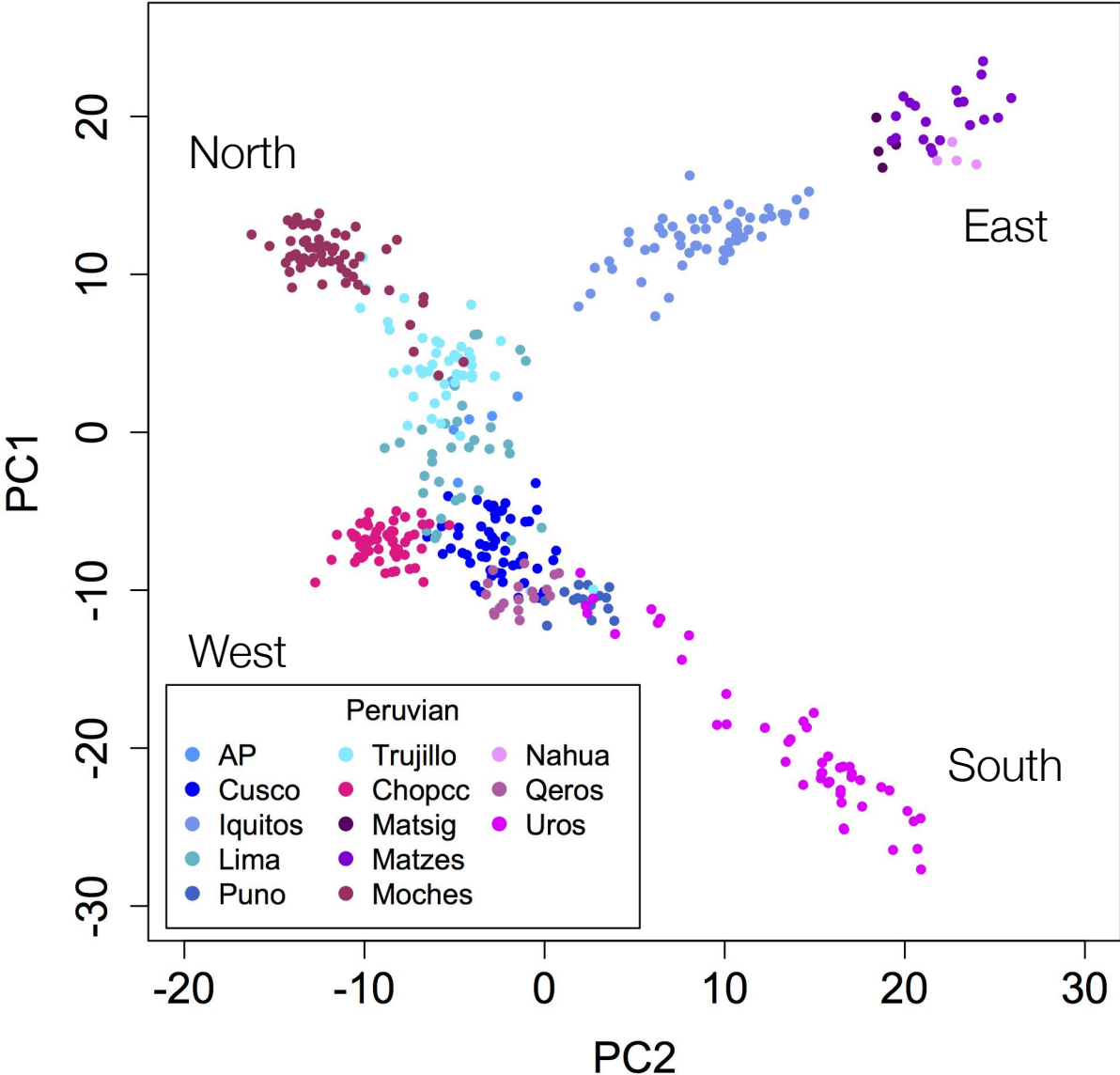


Ancestry specific PCA: Europe and Africa



Harris et al.
(submitted)

Peruvian population structure using Ancestry Specific PCA



Harris et al.
(submitted)

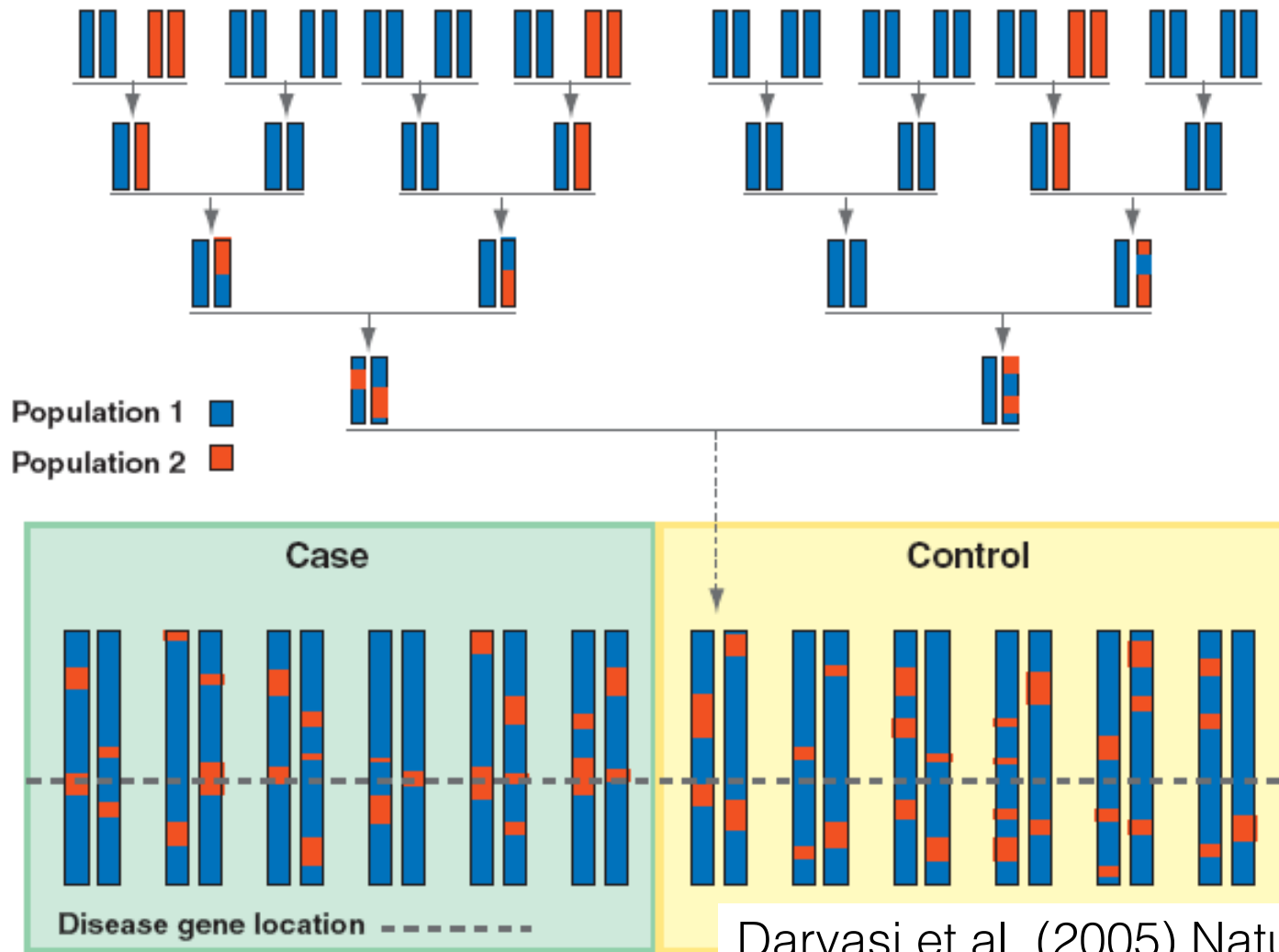
Admixture is not just a nuisance for association

- Differences in genetic architecture are not just nuisance values that need to be 'adjusted' for in association models.
 - Extension Studies
 - Admixture Mapping

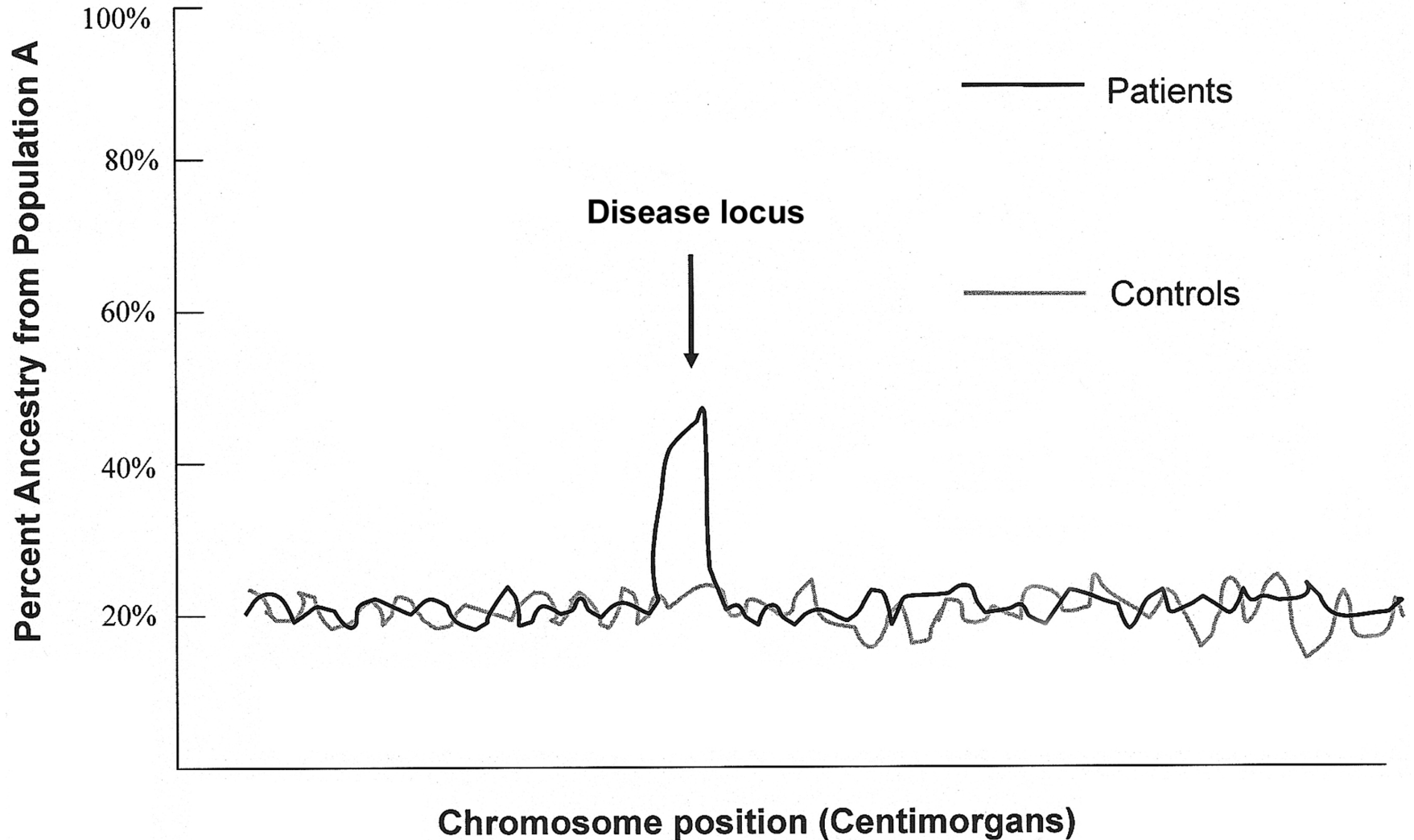
Extension Studies

- Extension of findings to other ancestries is important to:
 - Determine association's potential public health impact
 - Provide additional evidence supporting association
 - Useful in fine-mapping an association signal
 - Finding risk variation in non-homogenous populations (like African Americans)

Admixture mapping - Concept



Example of an Admixture scan



Concluding Summary

- PCA and Admixture analyses can summarize the ancestry found across the entire genome
- Local ancestry refines this inference to genomic segments with broad applications including demographic modeling and association analyses.