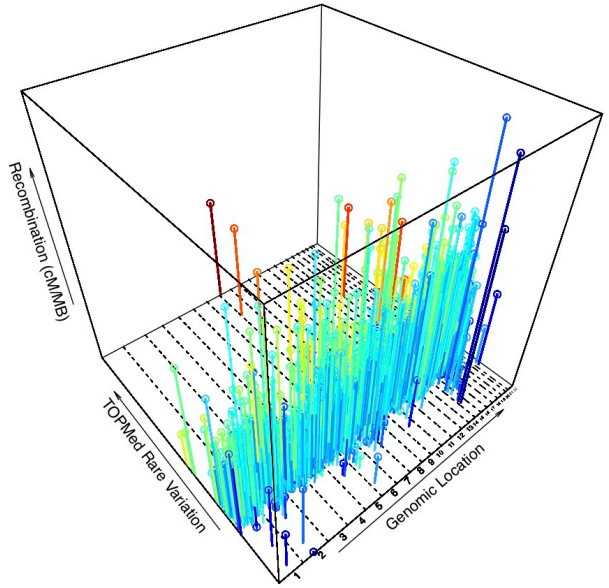
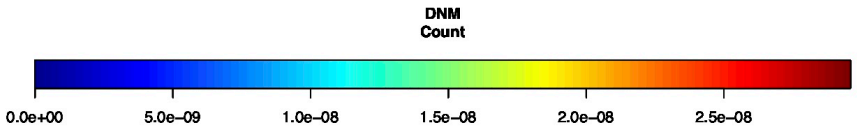
Population Structure Analysis

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

Have you ever tried visualizing 10,000 variables (dimensions)?



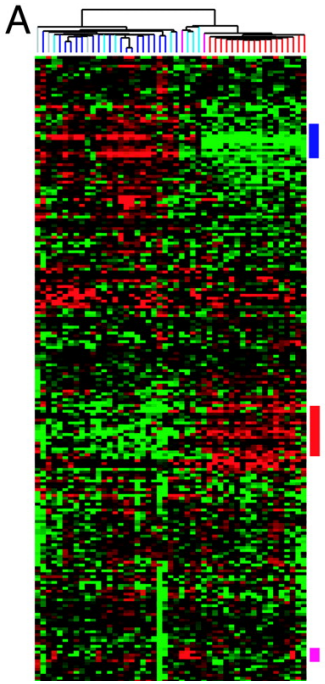


This is four variables and already pretty complicated.

Kessler et al. (2019) bioRxiv

Now we see more than 10,000 variables in many different areas

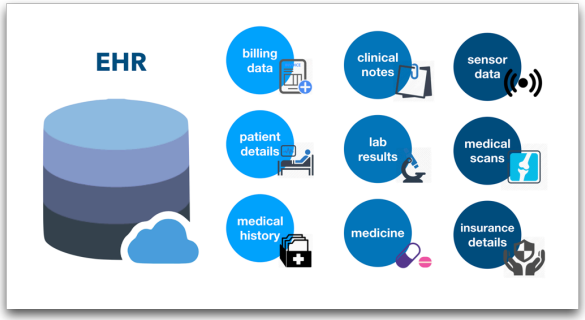
Expression



Sørlie et al. (2003) PNAS

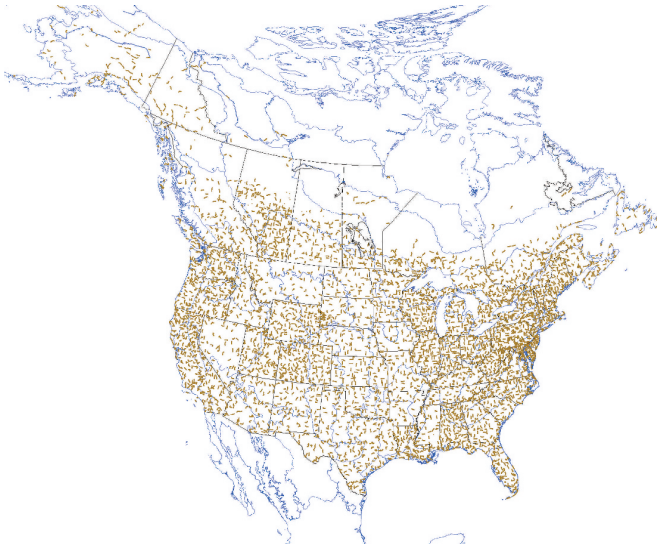
GWAS results

https://www.ebi.ac.uk/gwas/diagram



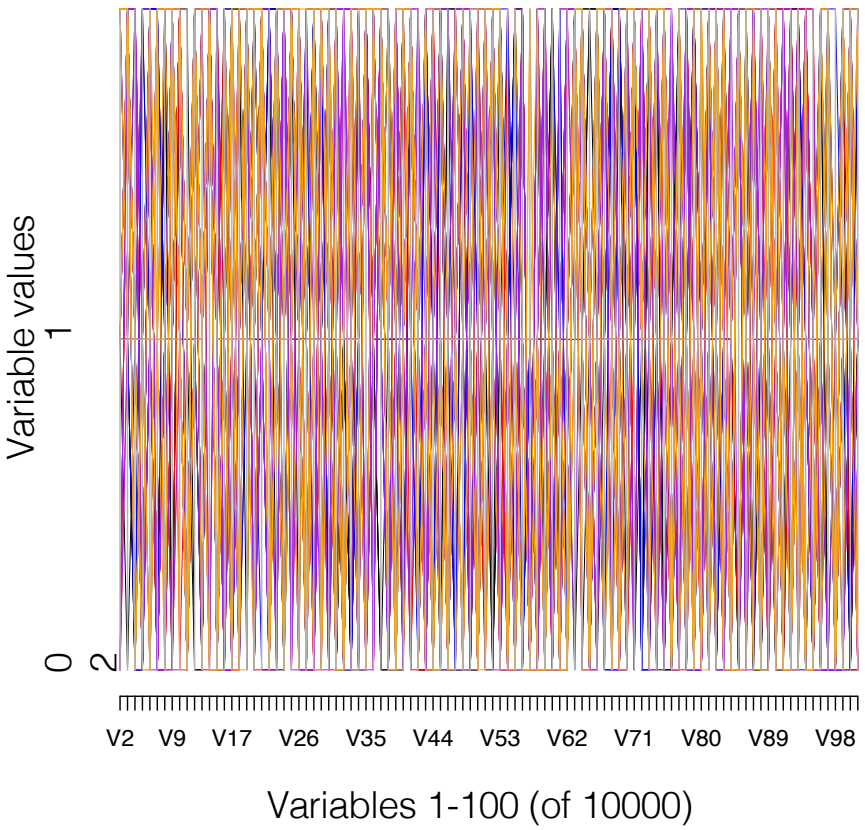
https://goku.me/blog/EHR

Ecology



Sauer et al. (2013) North American Fauna

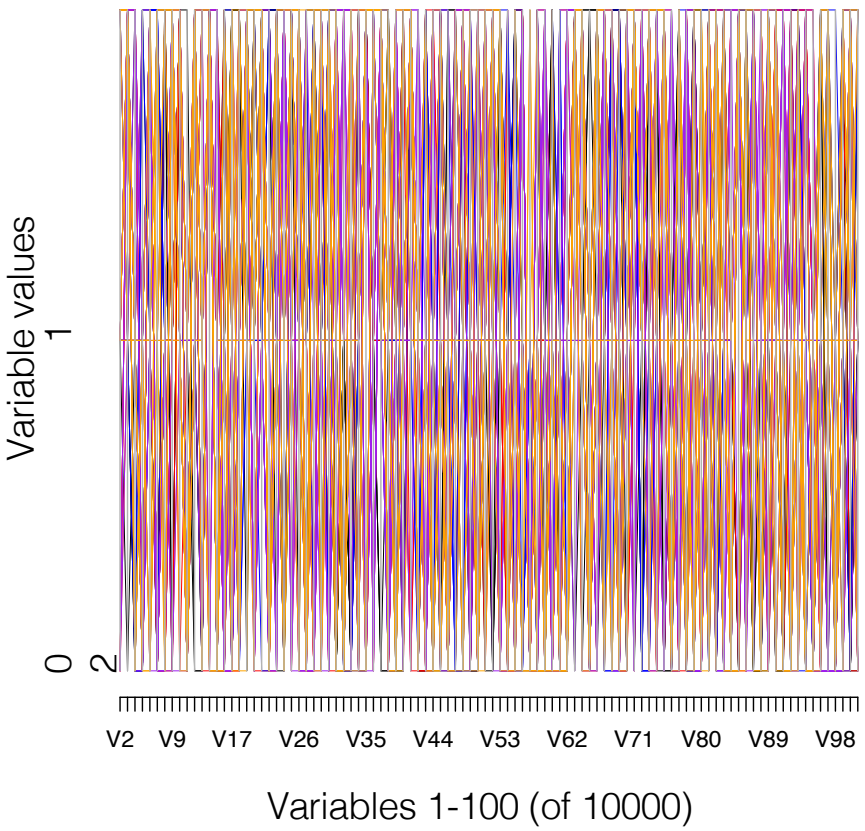
PCA: How does it work?

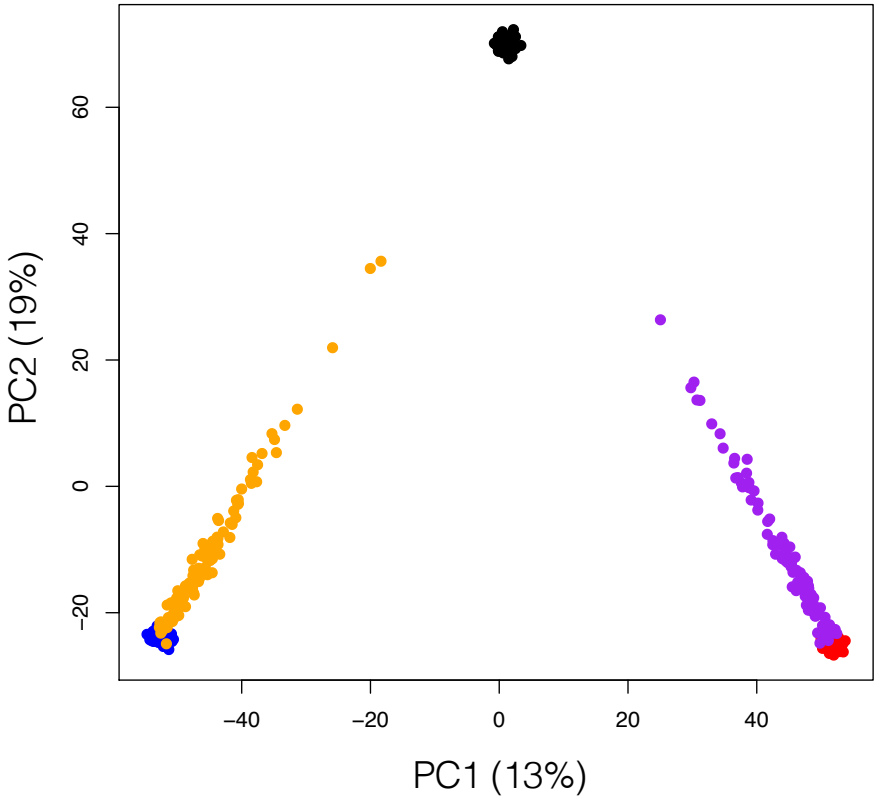



```
R code:
pca = prcomp(data,
    center=TRUE,
    scale.=TRUE)
```

1. Normalize the data
2. Find the covariance matrix between all variables
3. Calculate the eigenvectors and eigenvalues of the matrix
4. Transform or project the original data onto this new set of coordinates (PC-space)
5. PC1 is orthogonal (uncorrelated) with PC2, and so on.

PCA: How does it work?

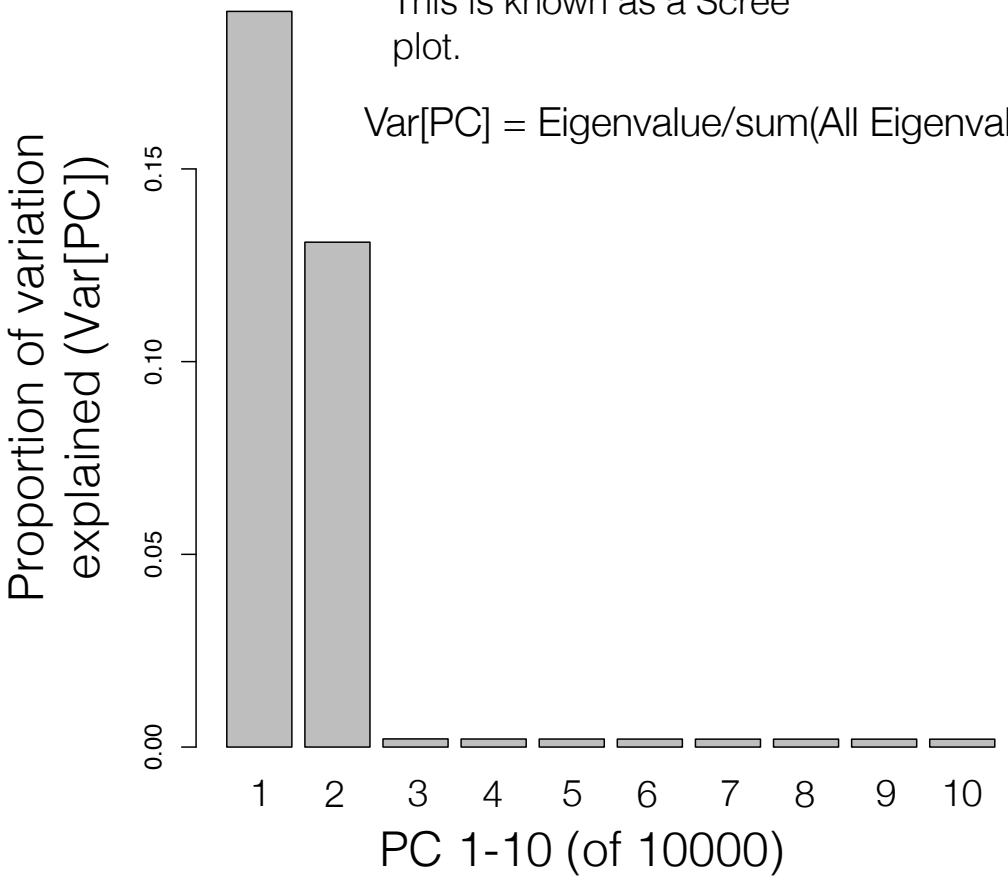




PCA: How does it work?

This is known as a Scree plot.

Var[PC] = Eigenvalue/sum(All Eigenvalues)



This is a Scree geological formation.

Mount Yamnuska, Alberta, Canada. https://commons.wikimedia.org/wiki/File:Yamnuska_bottom_cliff.jpg

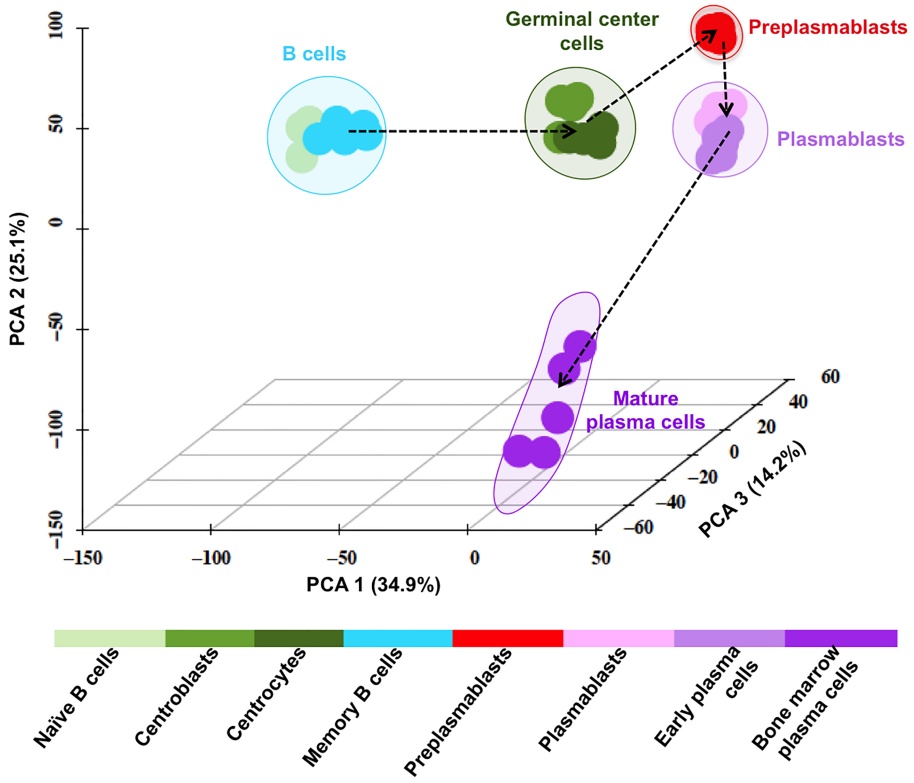
PCA: Uses

- Highly sensitive summary of all the data
- Summarize structure within the data
- Identify groups, when paired with K-means cluster
- Sanity check for study design
  - E.g. Diseased individuals cluster vs controls, first batch and second batch don't cluster together
- Sanity check when combining data

PCA: Assumptions and Pitfalls

- Assumptions
  - Linear relationship between data
  - Variables are independent
- Pitfalls
  - Only look at the first few PCs
  - All axes are biological/non-technical (once first few are)
  - Identifying significance of an axis is non-trivial

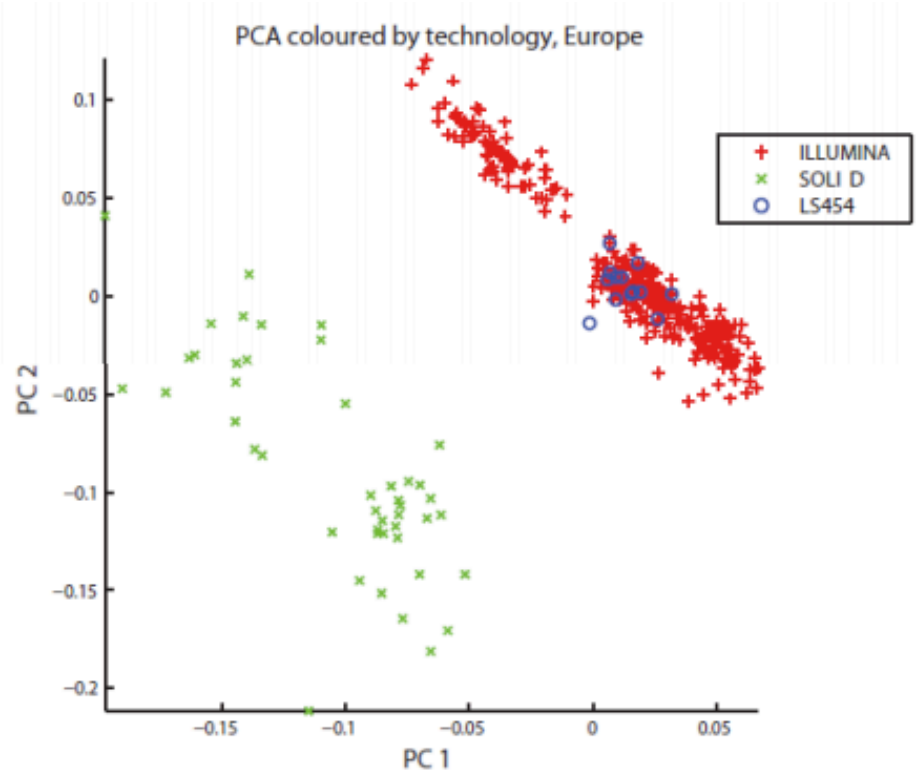
PCA Example 1: Gene expression by cell population



Expression of 9,303 genes reduced to THREE main axes of variation, explaining 75% of the original data.

Kassambara et al. (2015)
PLoS Comp. Biol.

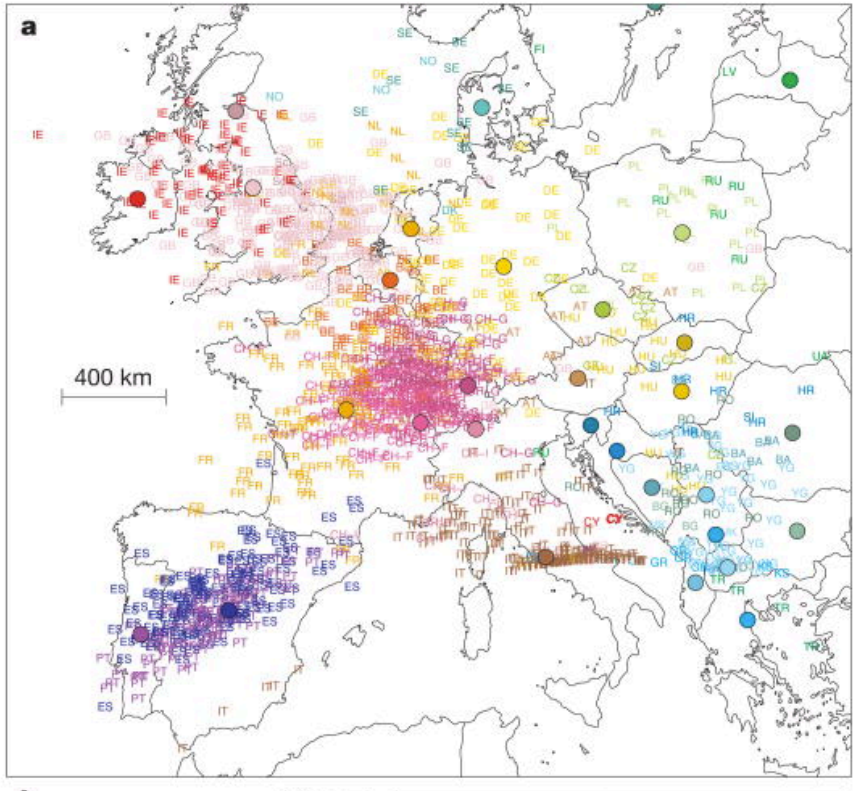
PCA Example 2: Technical Issues



Whole genome reduced to TWO main axes of variation that tightly correlate with sequence technology and geographic origin.

The 1000 Genomes Project Consortium (2012) Nature

PCA Example 3: "Genes mirror geography within Europe"



500K single nucleotide polymorphisms reduced to TWO main axes of variation that tightly correlate with sampling origin

Novembre et al. (2008) Nature

ADMIXTURE (Alexander et al. 2009)

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

Variants

Samples

Genotypes

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

Clusters

Samples

Admixture Proportions

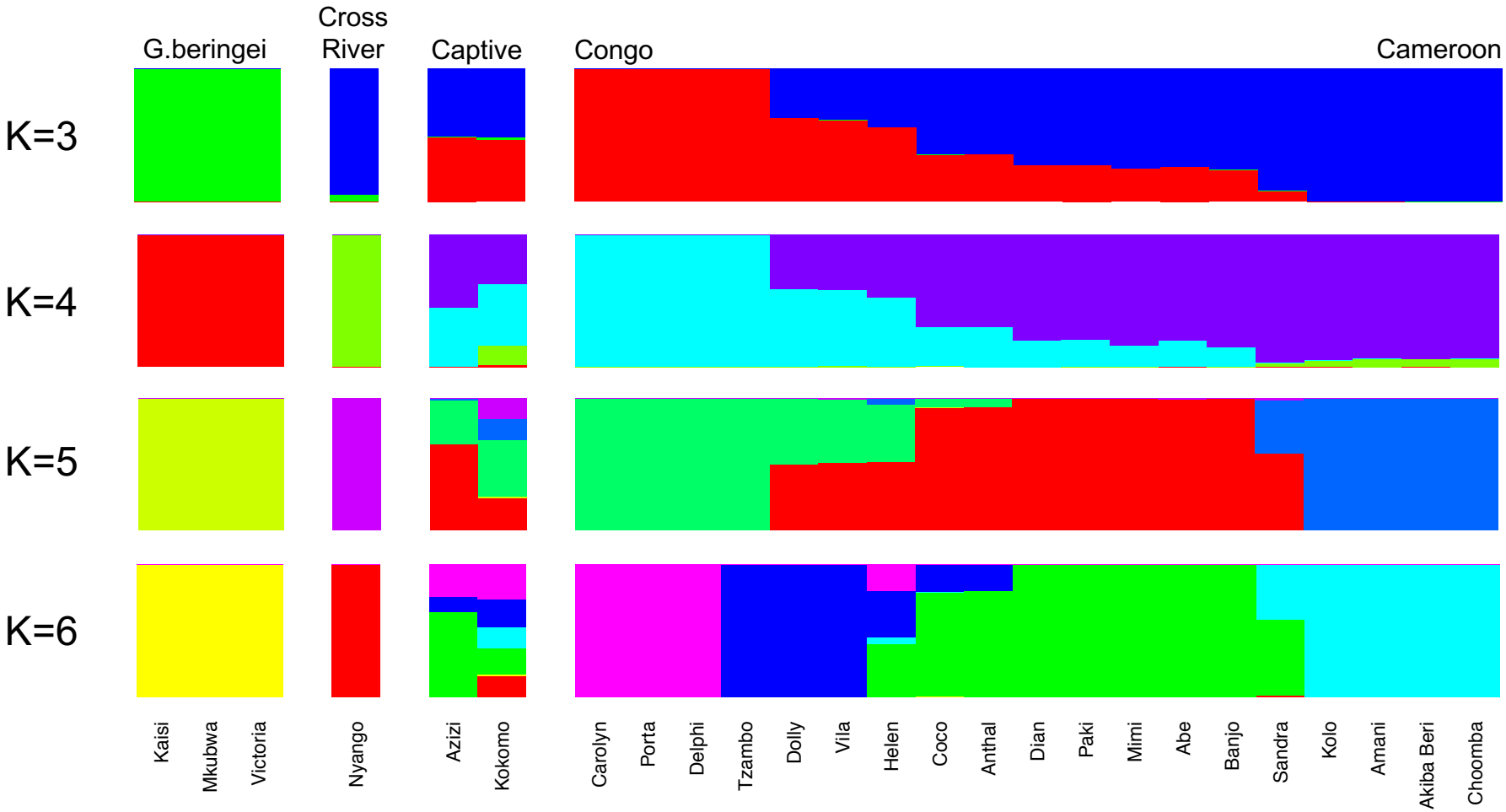
$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M1} & p_{M2} & \cdots & p_{MK} \end{bmatrix}$$

Variants

Clusters

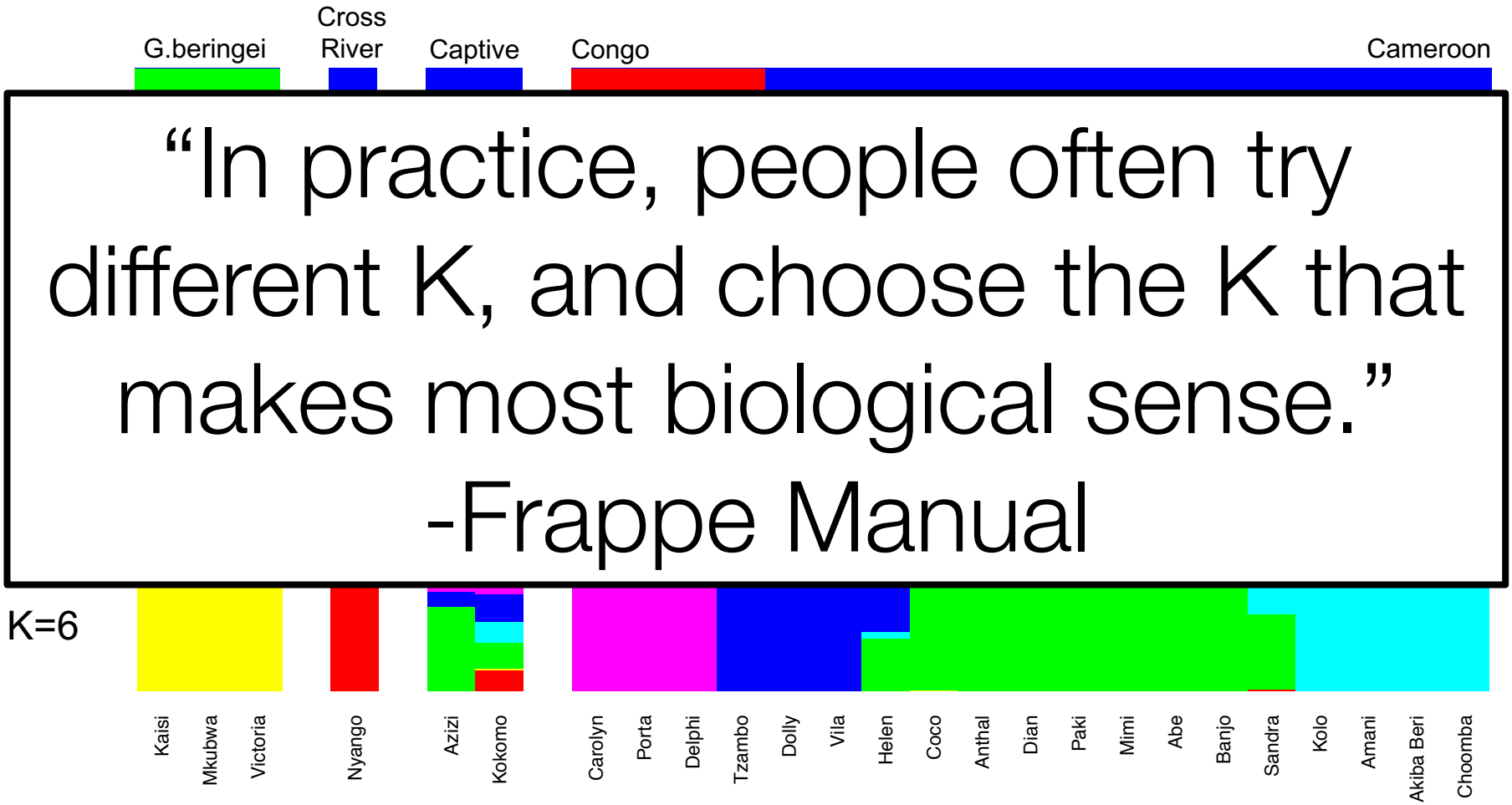
Ancestral Allele Frequencies

Admixture analyses



Prado-Martinez et al. (2013) Nature

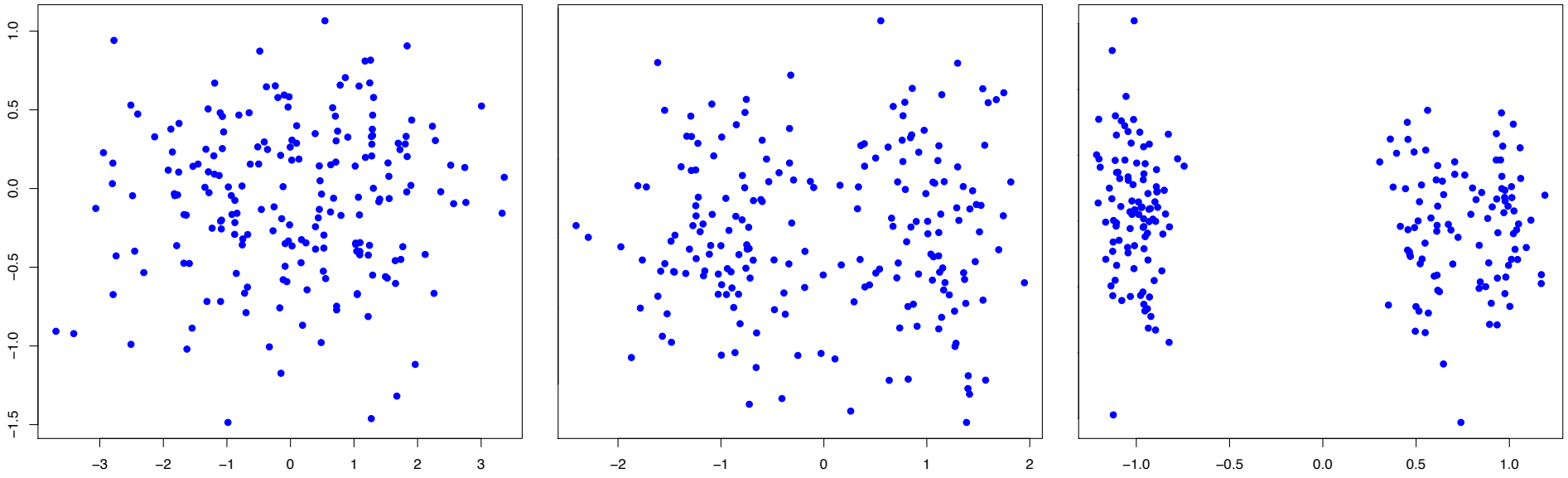
Admixture analyses: when is the K correct?



"In practice, people often try different K, and choose the K that makes most biological sense."
-Frappe Manual

Prado-Martinez et al. (2013) Nature

The K Problem



How many different means are there?

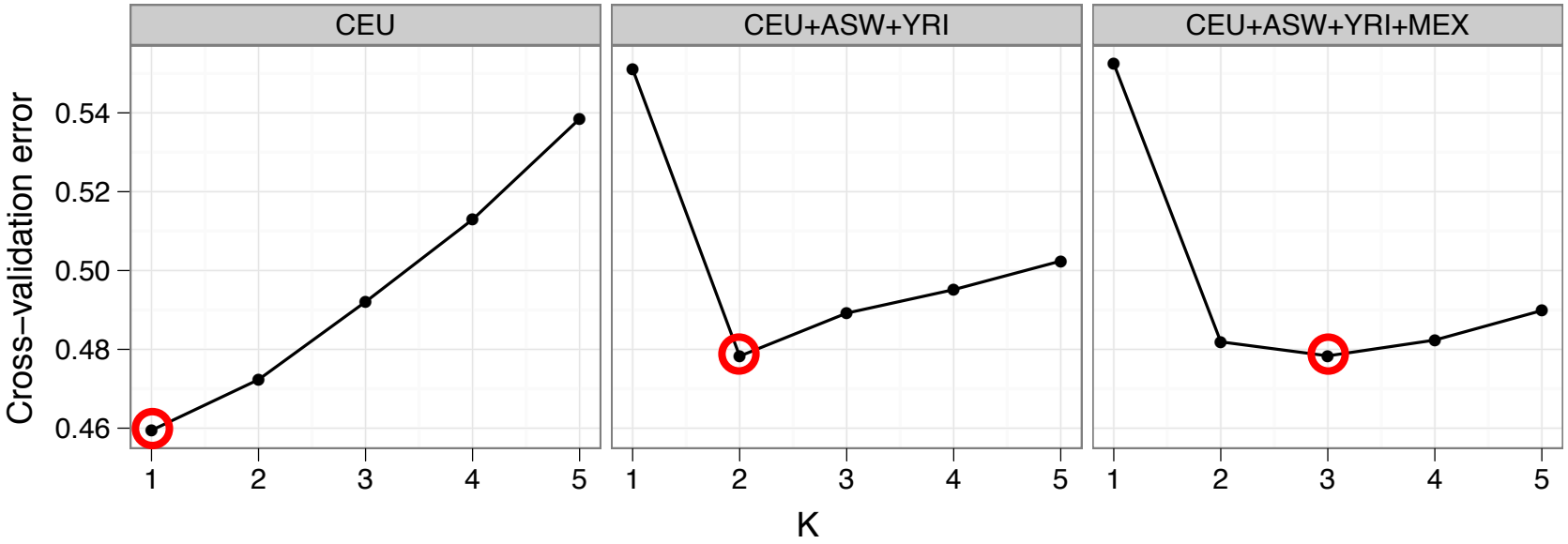
ADMIXTURE: using cross validation to identify the best K

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

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

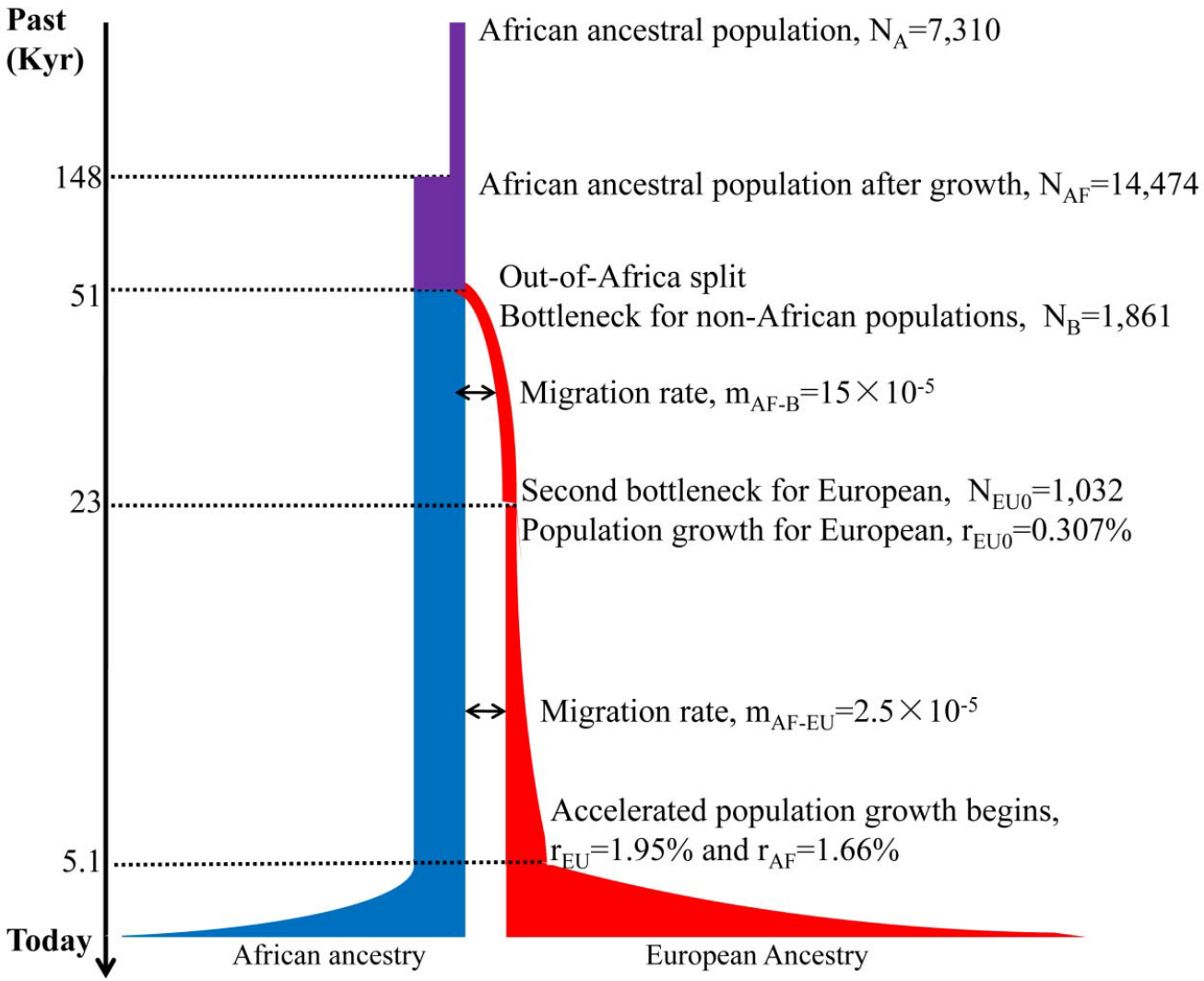
Alexander and Lange (2011) BMC Bioinformatics

How well X-validation performs



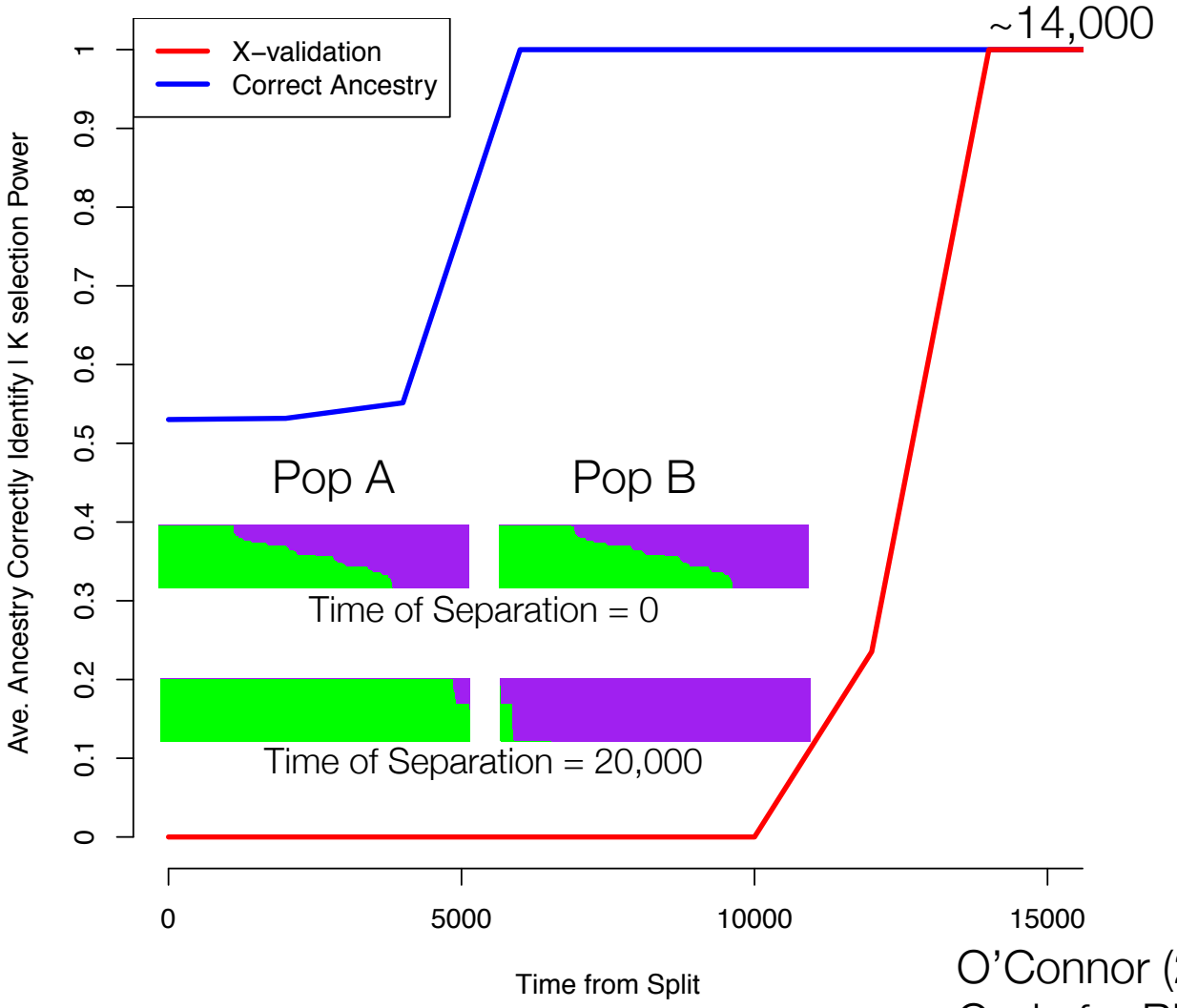
Alexander and Lange
(2011) BMC Bioinformatics

Test it with ESP inspired simulations



Fu et al. (2012) Nature

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

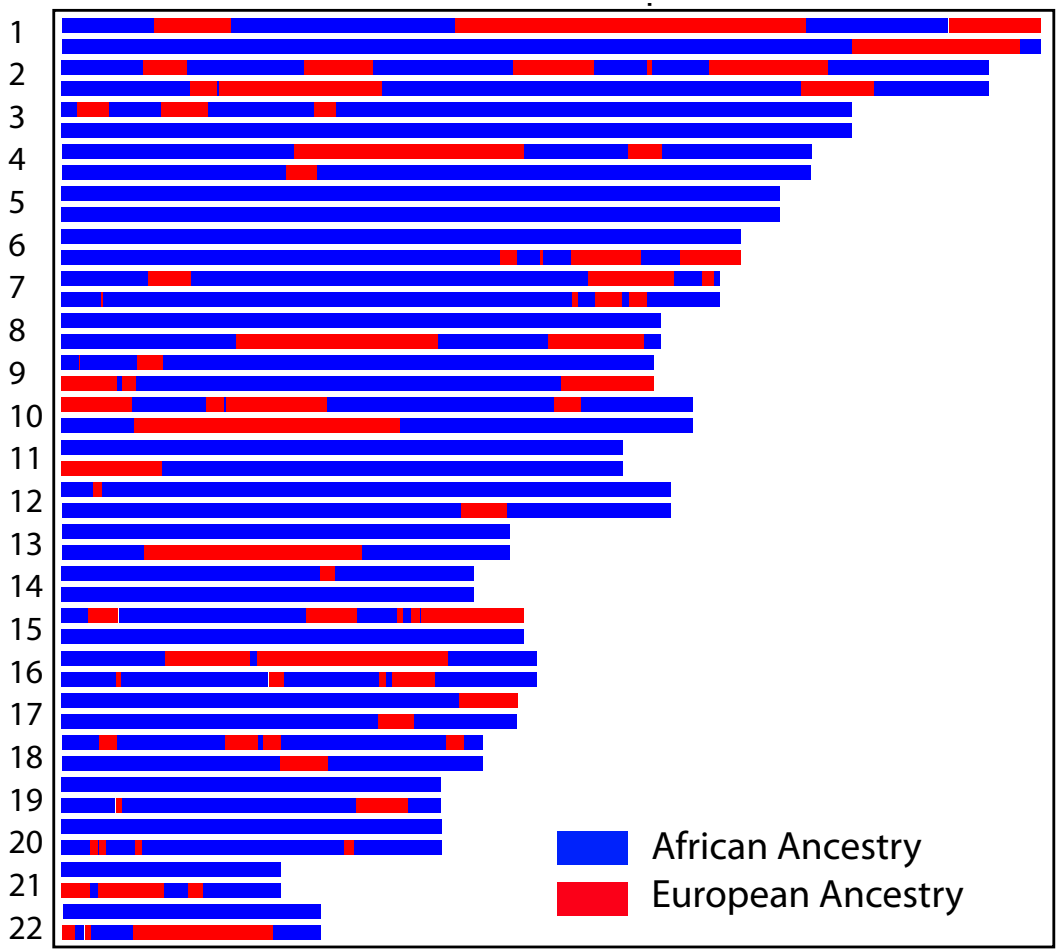


O'Connor (2015) Source Code for Biol. and Med.

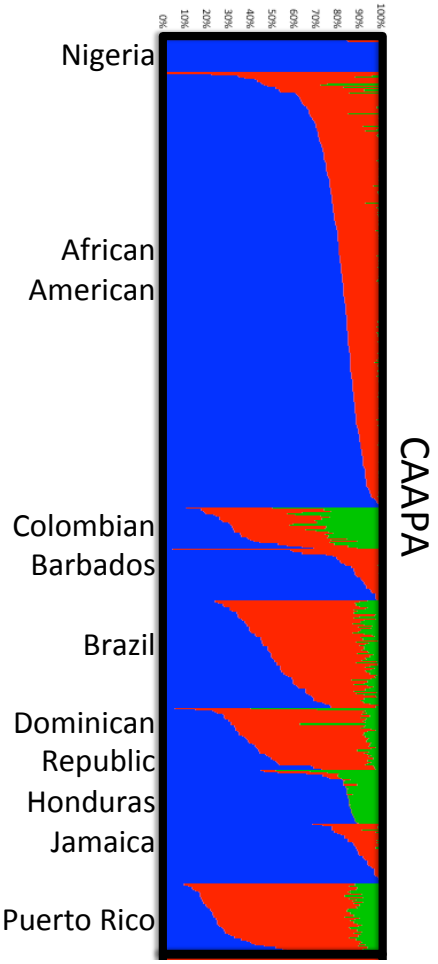
Tricks to effectively use ADMIXTURE

- 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

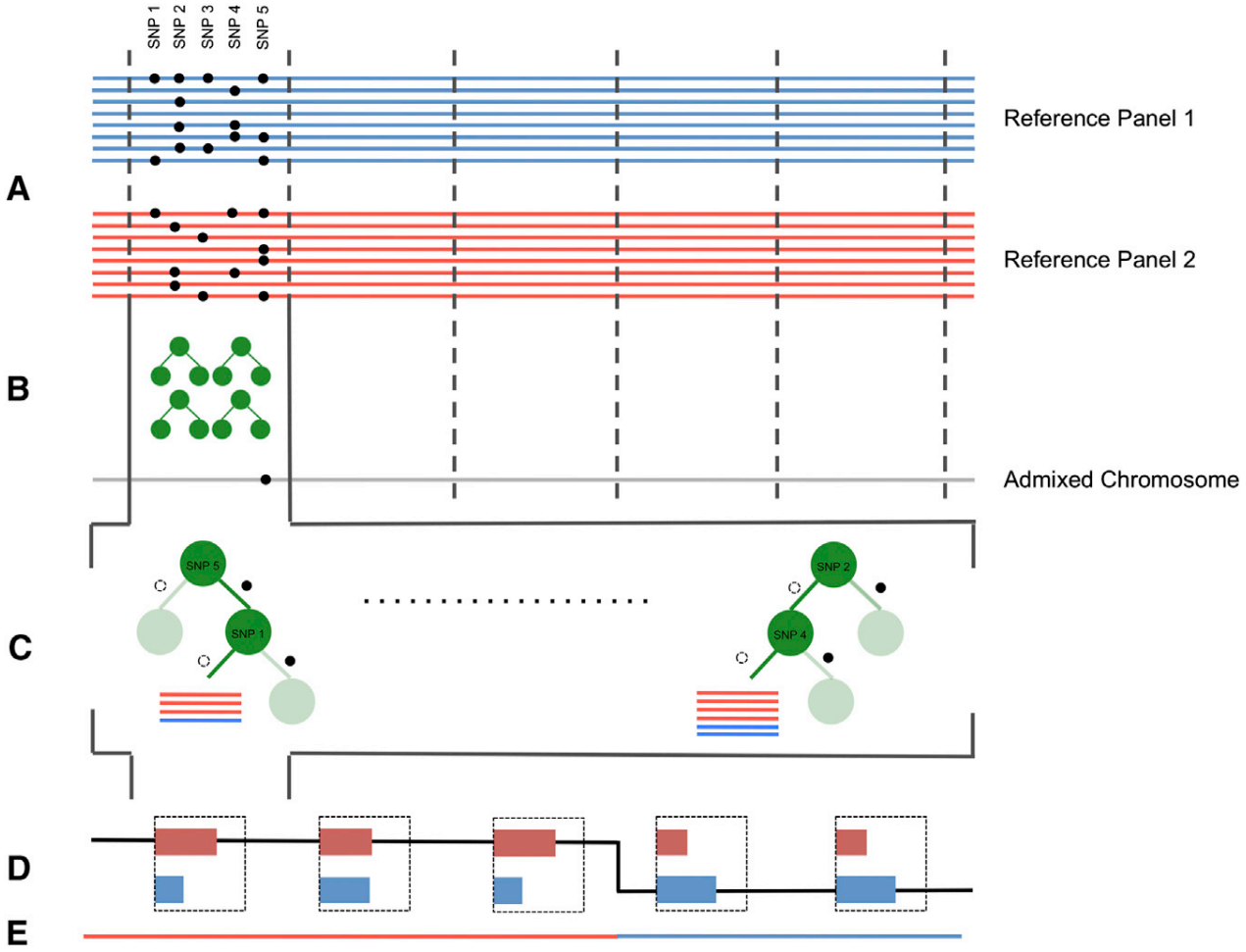


Gravel et al. (2013) Genetics



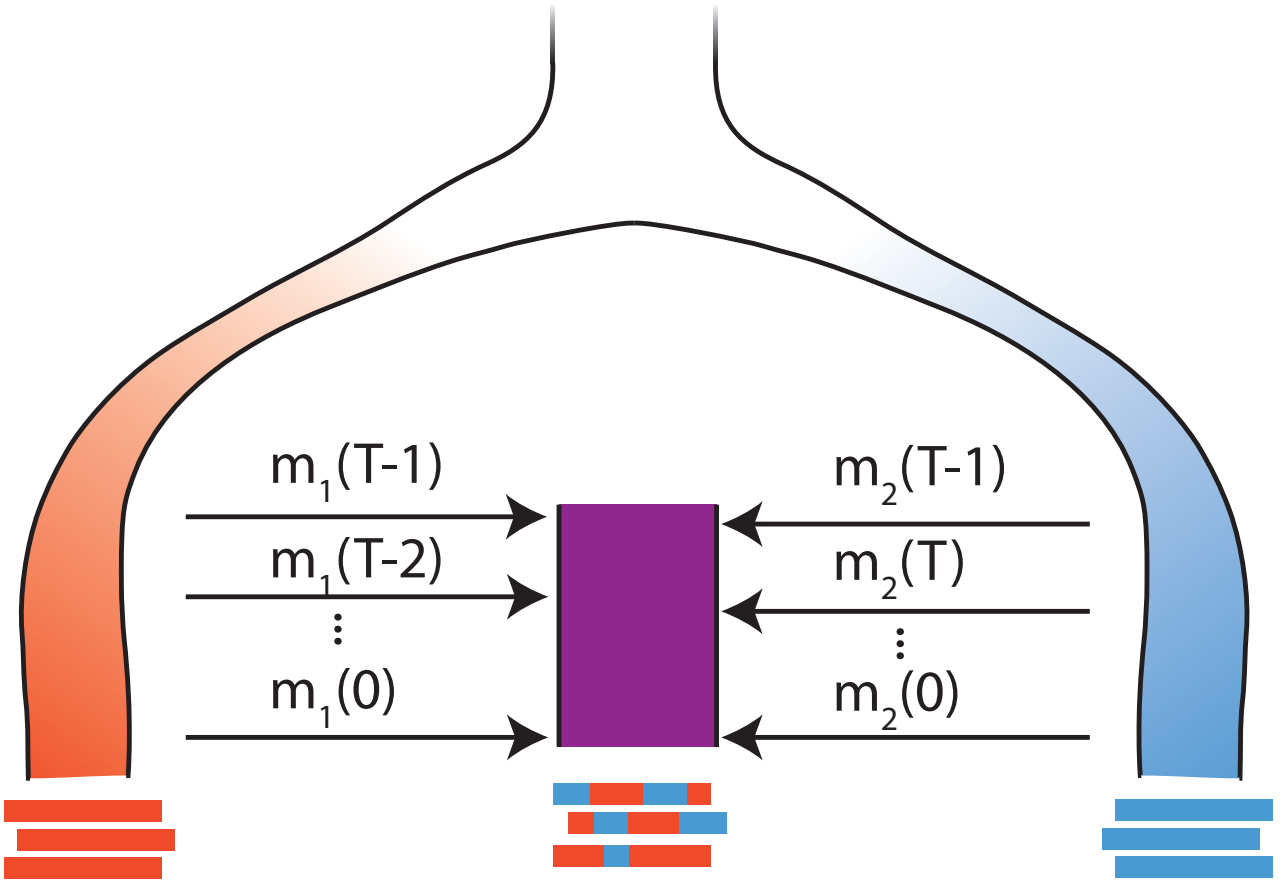
Mathias et al. (2016) Nat. Comm.

Local ancestry calling: RFMix as an example



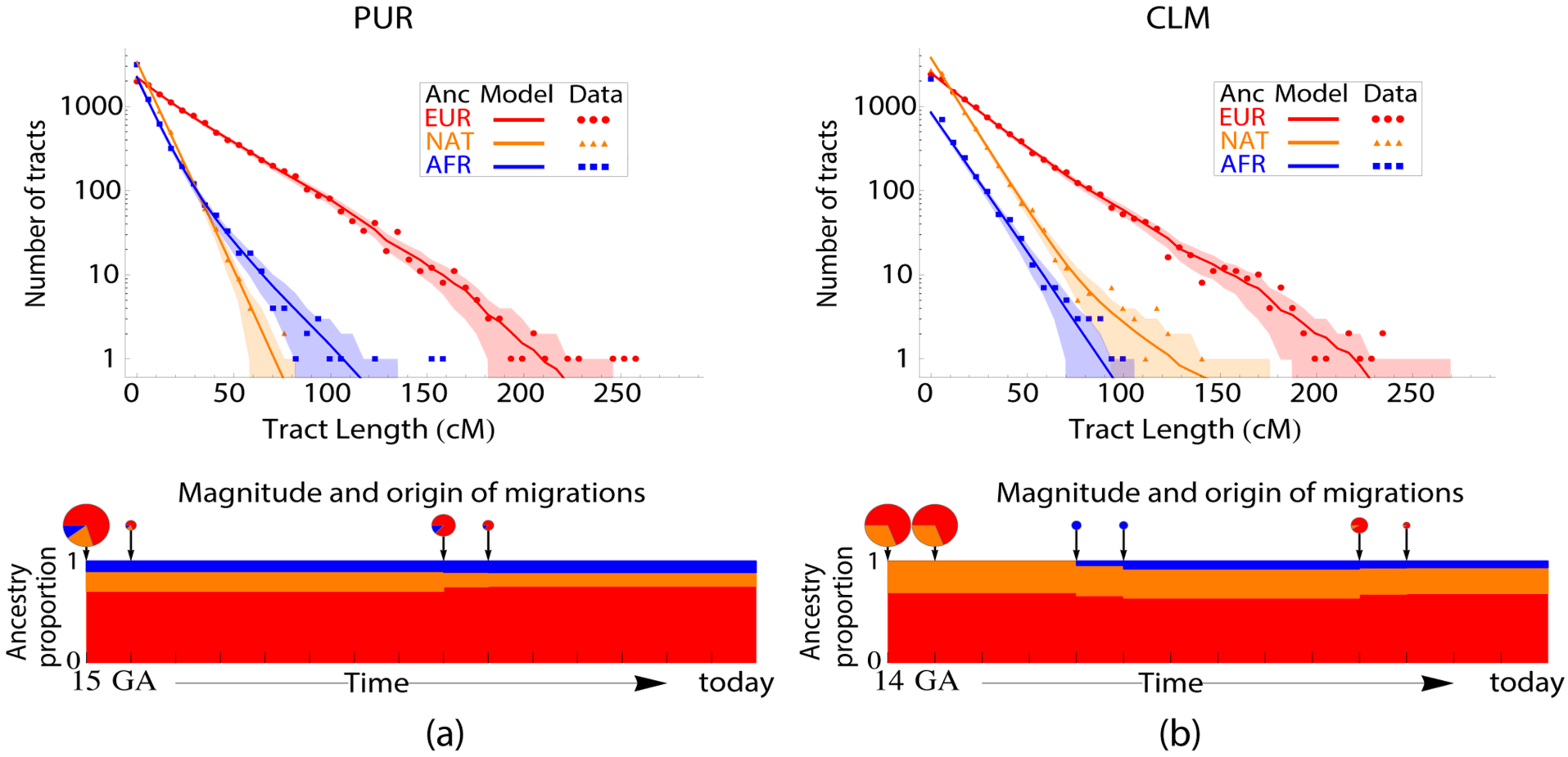
Maples et al. (2013) AJHG

Demographic modeling with local ancestry



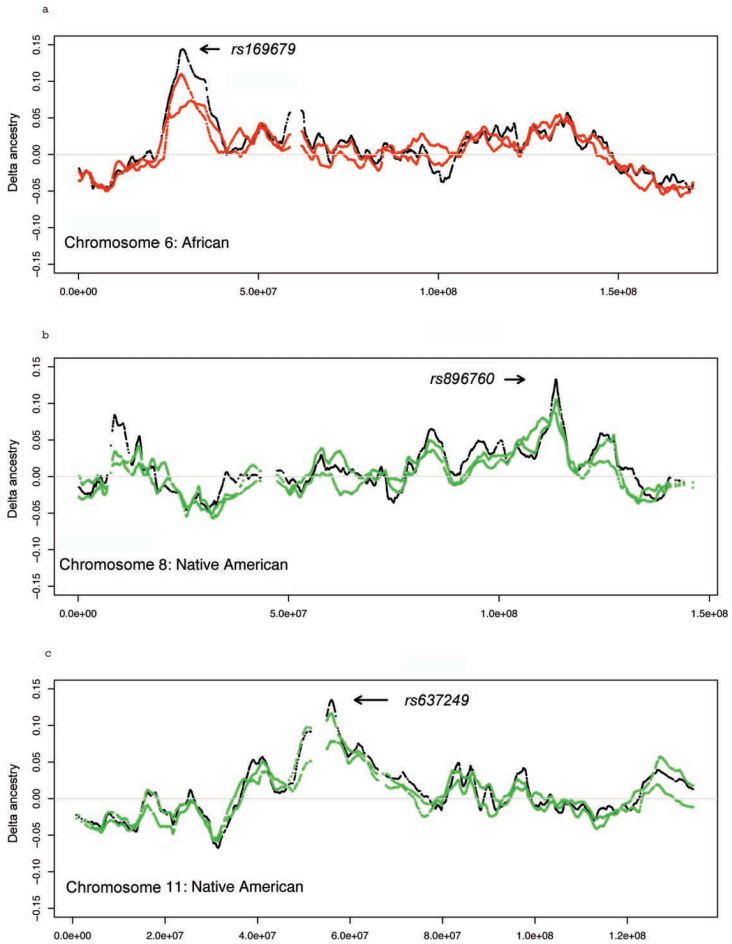
Gravel et al. (2013) Genetics

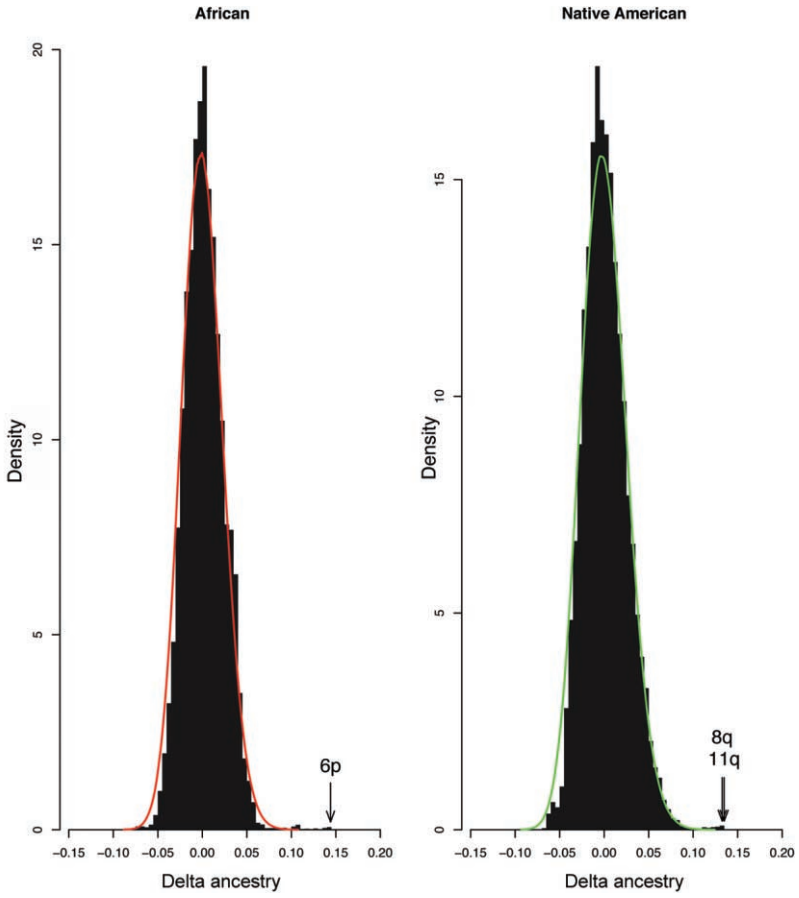
Demographic modeling with local ancestry



Gravel et al. (2013) PLoS Genet.

Recent selection by looking for local ancestry biases

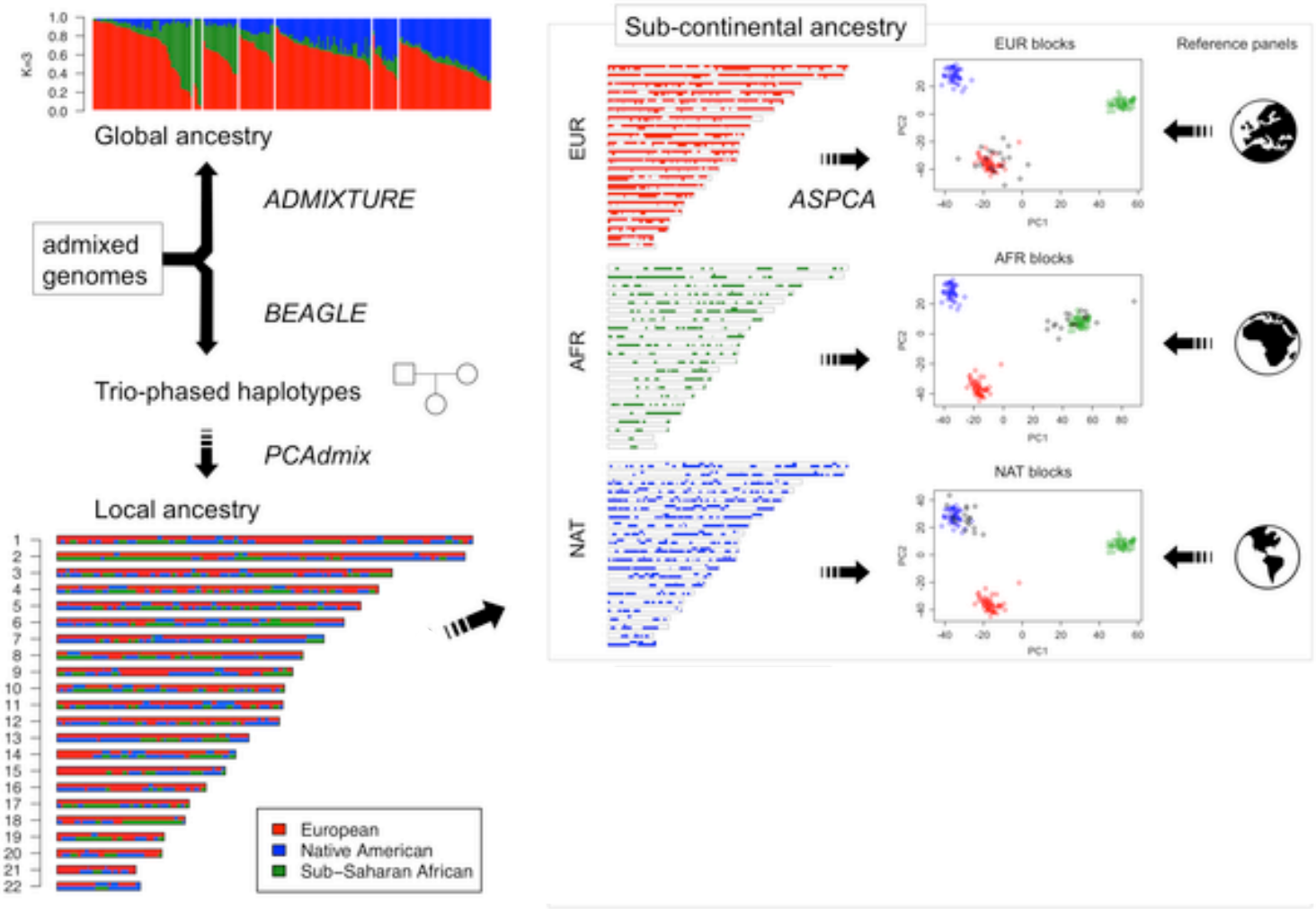




Tang et al. (2007) AJHG

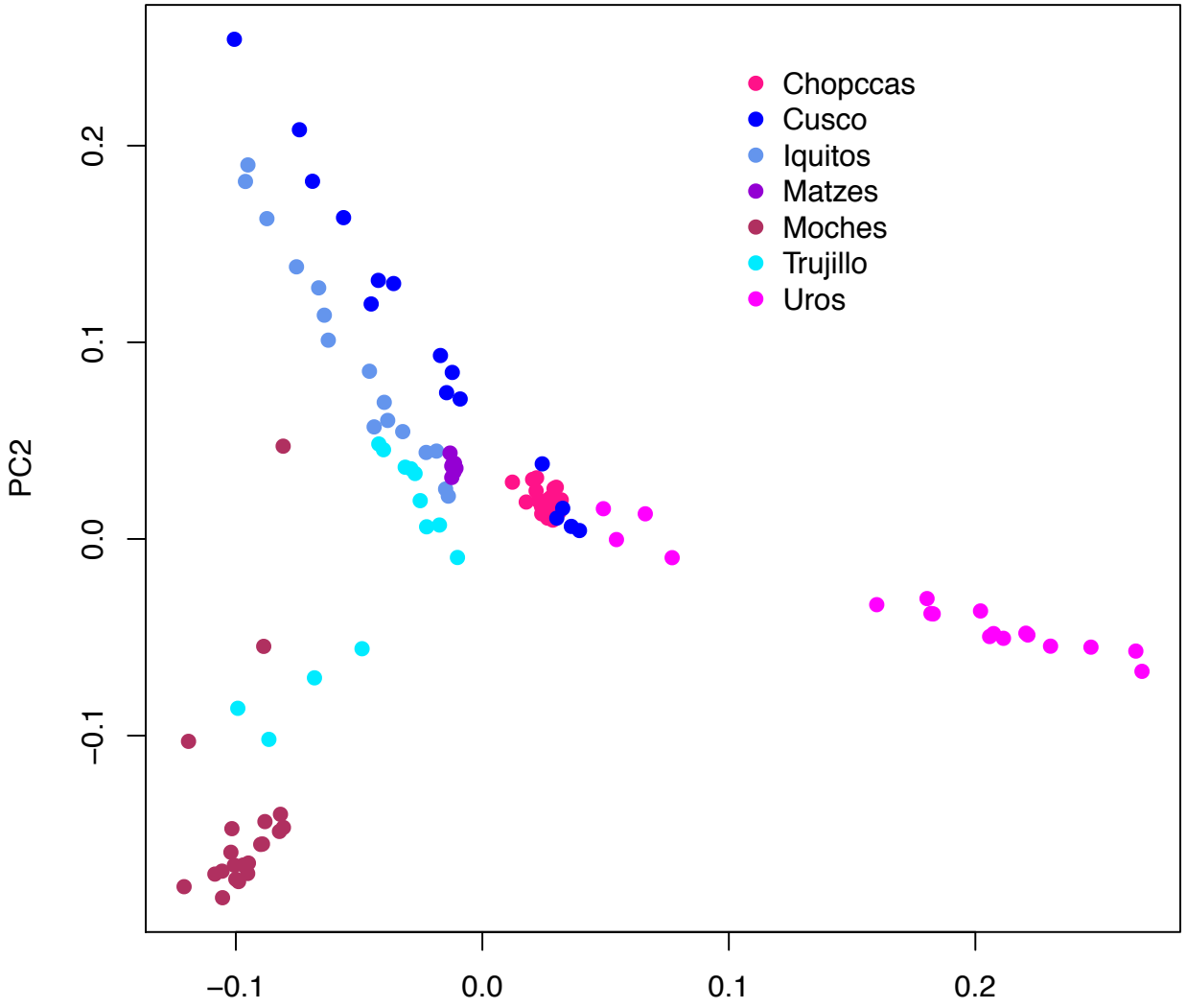
Though see Bhatia et al. (2014) AJHG

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



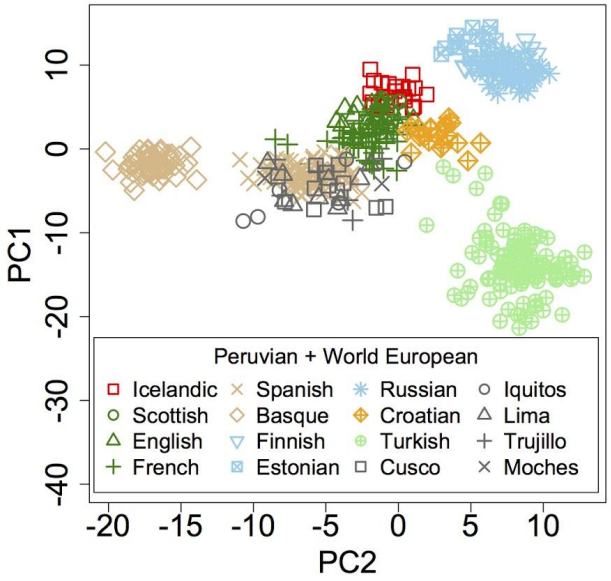
Moreno-Estrada et al. (2013) PLOS Genet.

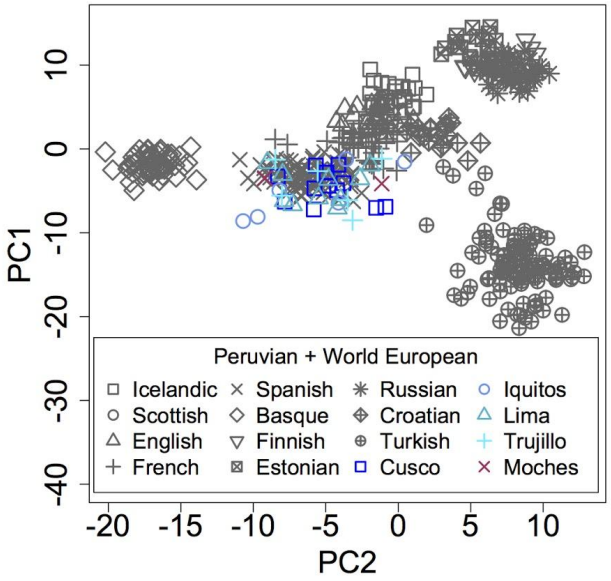
Peruvian population structure with PCA

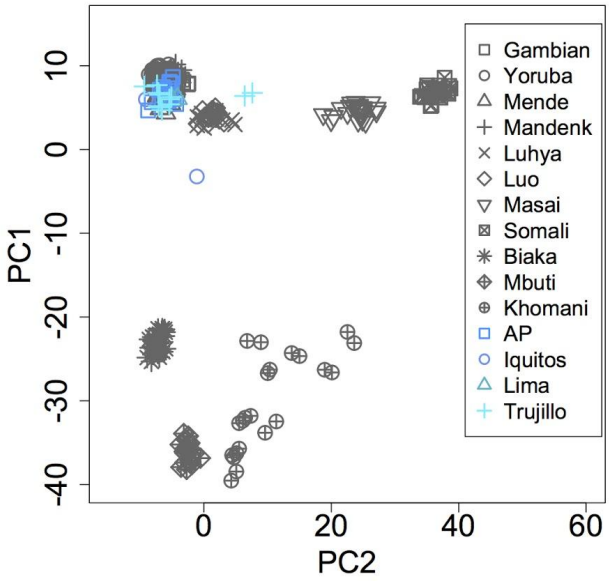


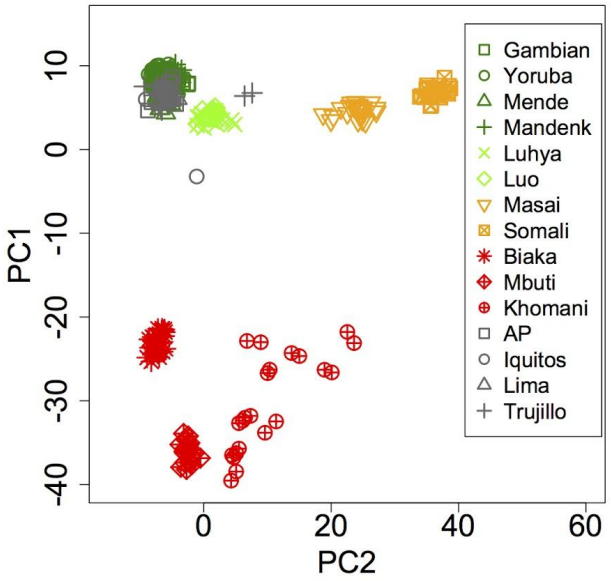
Harris et al. (2018) PNAS

Ancestry specific PCA: Europe and Africa



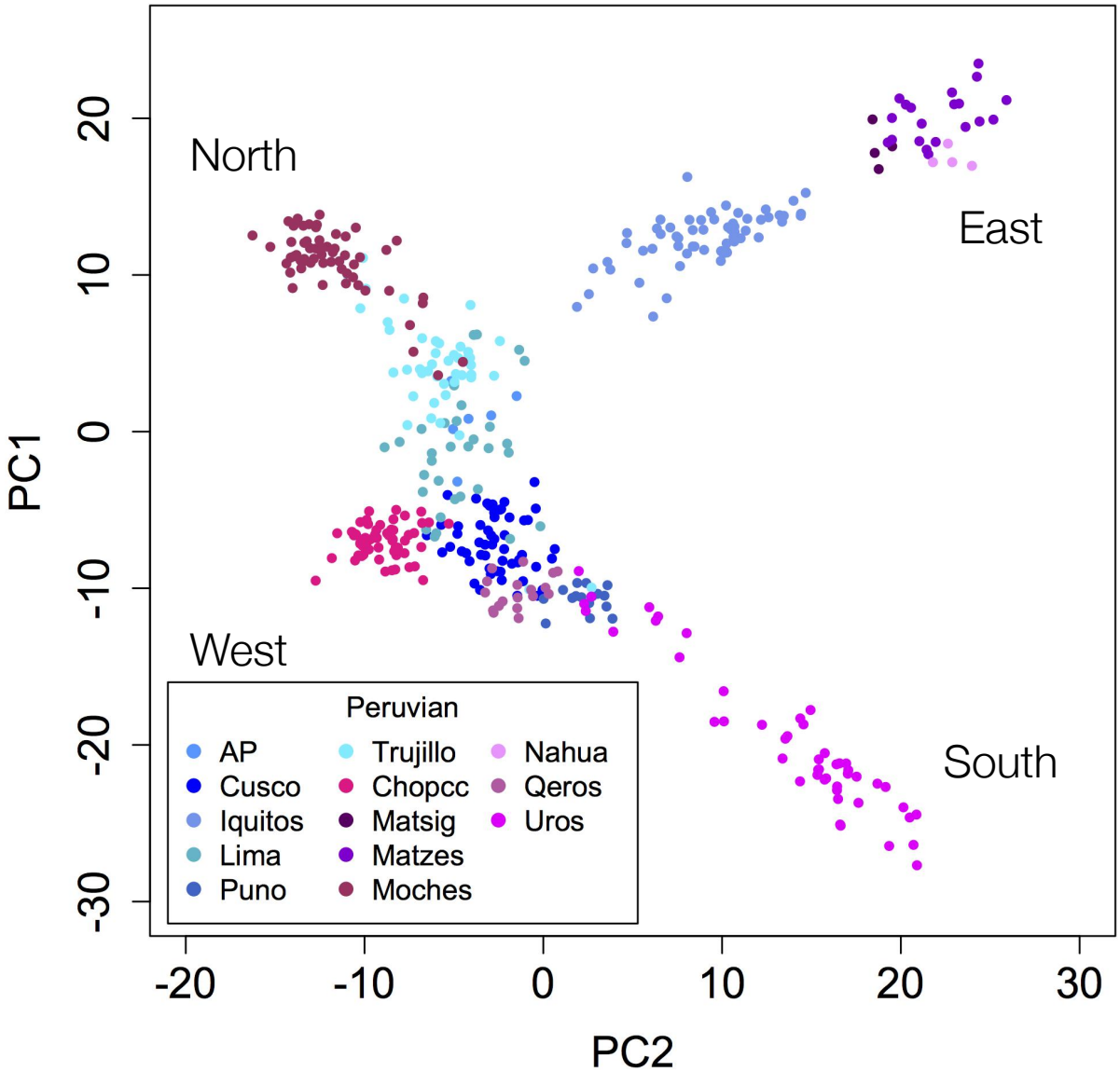






Harris et al. (2018) PNAS

Peruvian population structure using Ancestry Specific PCA



Harris et al. (2018) PNAS

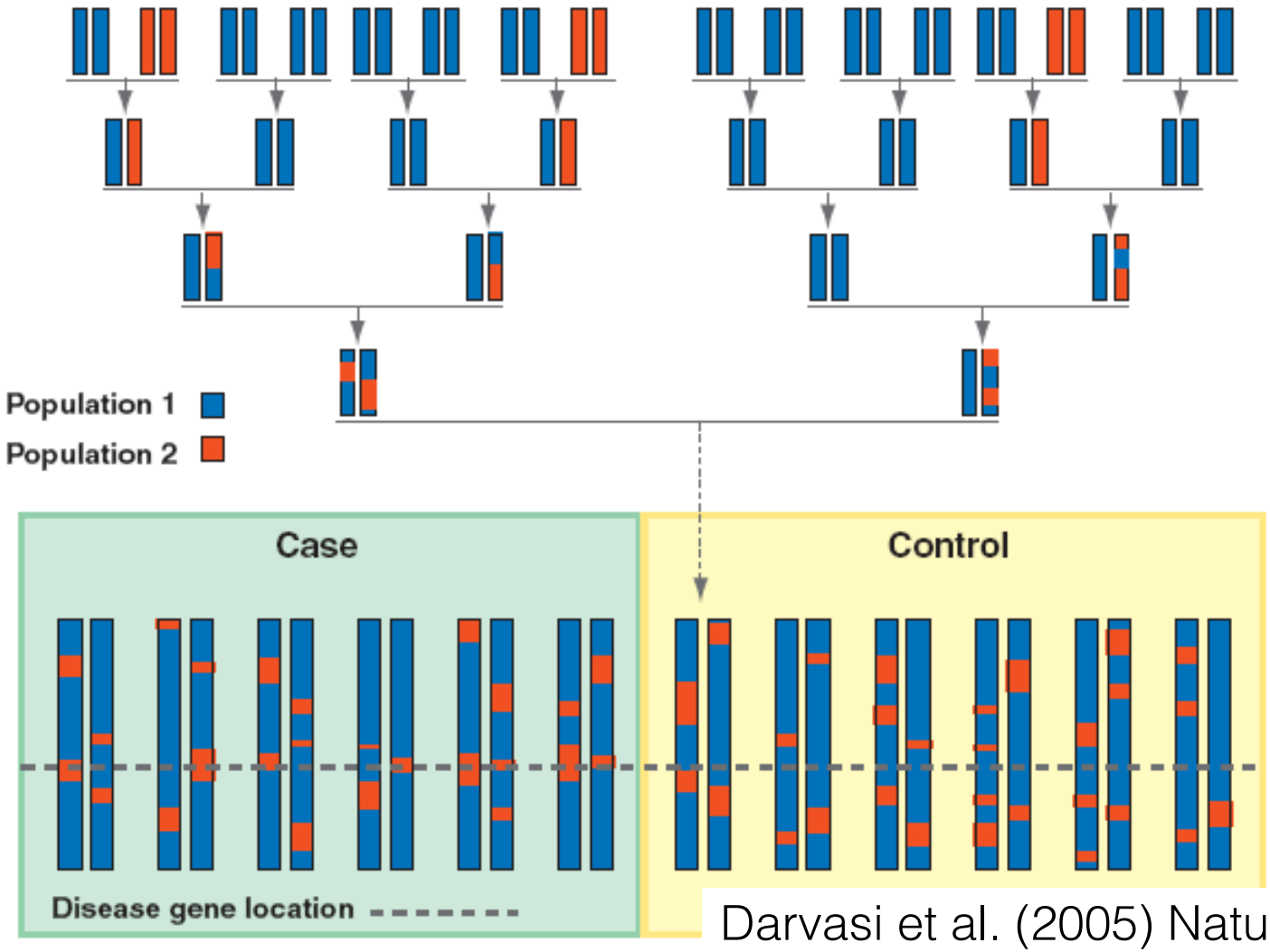
Admixture is not just a nuisance for association

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

Extension Studies

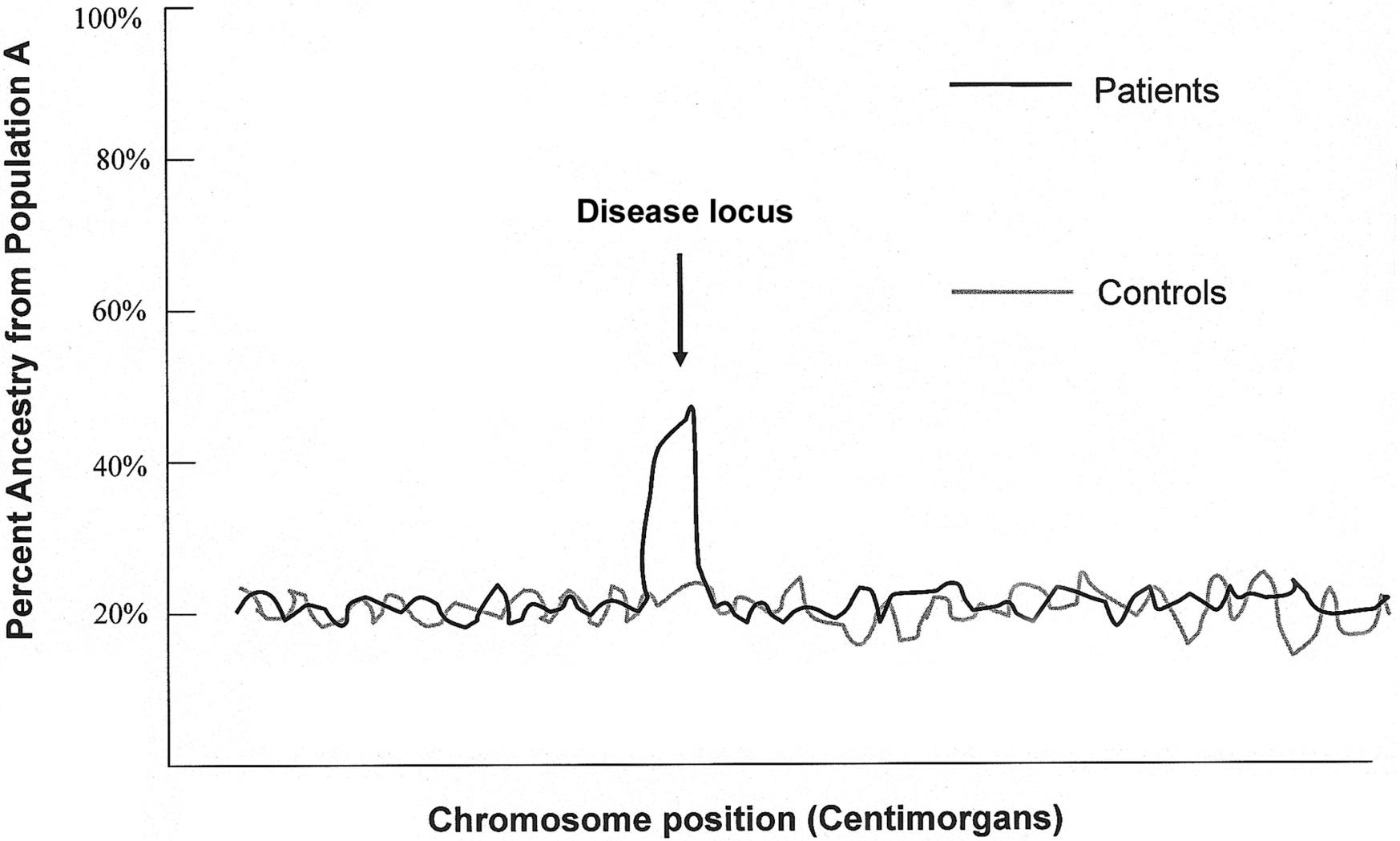
- Extension of findings to other ancestries is important to:
  - Determine association's potential public health impact
  - Provide additional evidence supporting association
  - Useful in fine-mapping an association signal
  - Finding risk variation in non-homogenous populations (like African Americans)

Admixture mapping - Concept



Darvasi et al. (2005) Nature Genet.

Example of an Admixture scan



Patterson et al. (2004) AJHG

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.