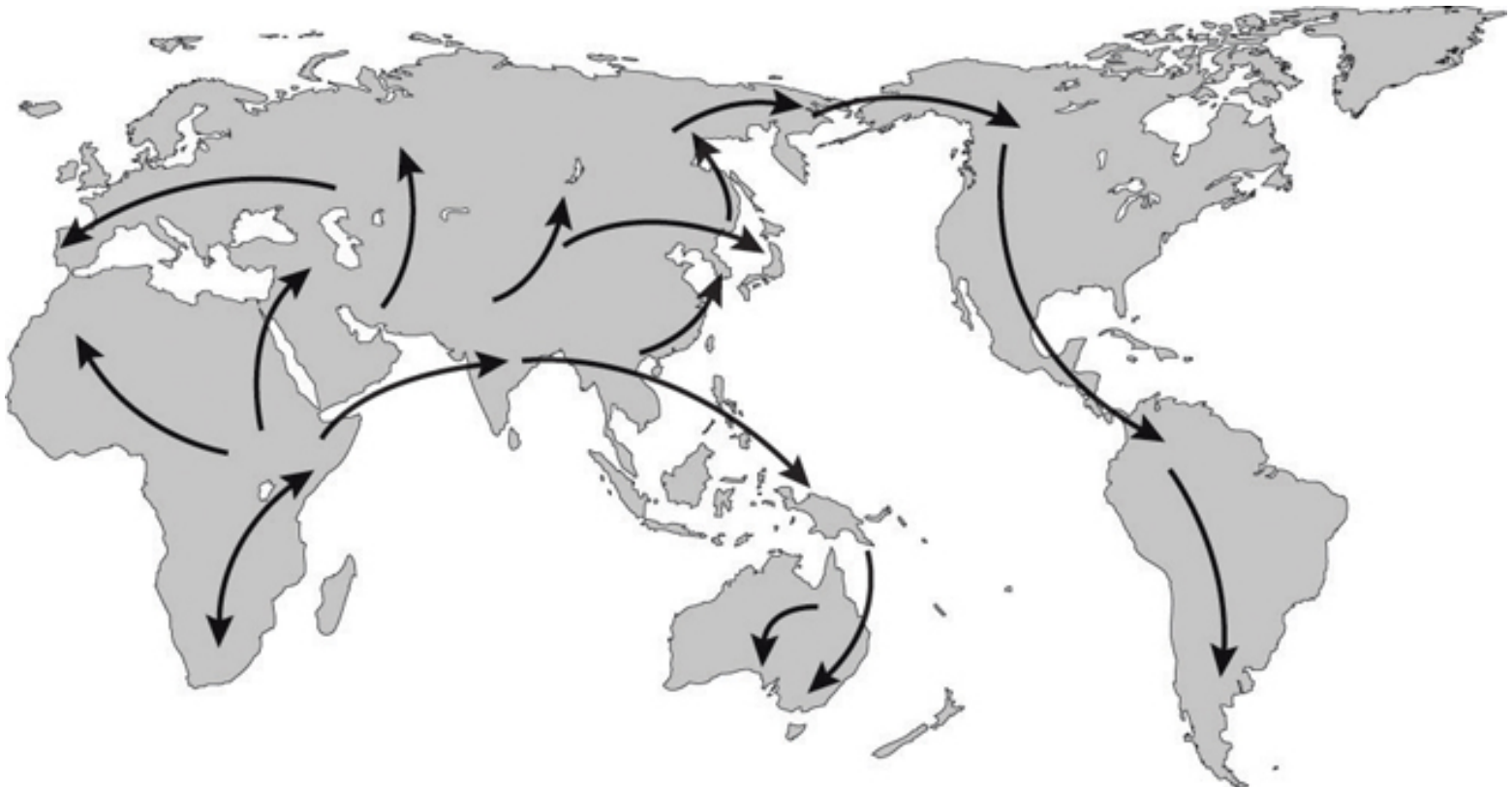
Lecture 5: Population Structure Inference

Timothy Thornton and Michael Wu

Summer Institute in Statistical Genetics 2015

Background: Population Structure

- Humans originally spread across the world many thousand years ago.
- Migration and genetic drift led to genetic diversity between isolated groups.



Inferring Population Structure

- Inference on genetic ancestry differences among individuals from different populations, or **population structure**, has been motivated by a variety of applications:
  - population genetics
  - genetic association studies
  - personalized medicine
  - forensics
- Advancements in array-based genotyping technologies have largely facilitated the investigation of genetic diversity at remarkably high levels of detail
- A variety of methods have been proposed for the identification of genetic ancestry differences among individuals in a sample using high-density genome-screen data.

Inferring Population Structure with PCA

- Principal Components Analysis (PCA) is the most widely used approach for identifying and adjusting for ancestry difference among sample individuals
- PCA applied to genotype data can be used to calculate **principal components** (PCs) that explain differences among the sample individuals in the genetic data
- The top PCs are viewed as continuous axes of variation that reflect genetic variation due to ancestry in the sample.
- Individuals with similar values for a particular top principal component will have similar ancestry for that axes.

Identifying Genetic Ancestry Differences: Unsupervised Learning with PCA

- Suppose a genetic association study consists of a sample of N individuals
- Assume that genotype data is available at S SNPs in a genome-screen, where S can be very large (e.g. hundreds of thousands).
- For SNP s define $\mathbf{G_s} = (G_1^s, \ldots G_n^s)^T$ is $n \times 1$ vector of the genotypes, where $G_i^s = 0, \frac{1}{2}$, or 1, according to whether individual i has, respectively, 0, 1 or 2 copies of the reference allele at SNP s.
- We define $\mathbf{Z}$ to be an $N \times S$ standardized matrix with (i, s)-th entry

$$\mathbf{Z}_{is} = \frac{G_i^s - \hat{p}_s}{\sqrt{\hat{p}_s(1 - \hat{p}_s)}}$$

and $\hat{p}_s$ will typically be an allele frequency estimate for SNP s

Identifying Genetic Ancestry Differences: Unsupervised Learning with PCA

- Can obtain a genetic relationship matrix (GRM) $\hat{\boldsymbol{\Psi}}$ where

$$\hat{\boldsymbol{\Psi}} = \frac{1}{S}\mathbf{Z}\mathbf{Z}^T$$

The (i, j)-th entry of $\hat{\boldsymbol{\Psi}}$ is a measure of the average genetic similarity for individuals i and j in the sample.

$$\hat{\boldsymbol{\Psi}}_{ij} = \frac{1}{S}\sum_{s=1}^{S}\frac{(G_i^s - \hat{p}_s)(G_j^s - \hat{p}_s)}{\hat{p}_s(1-\hat{p}_s)}$$

- Principal Components Analysis (PCA) is a dimension reduction that can be applied to GSM to identify ancestry differences among sample individuals
- PCA is performed by obtaining the eigendecomposition of the GRM $\hat{\boldsymbol{\Psi}}$.

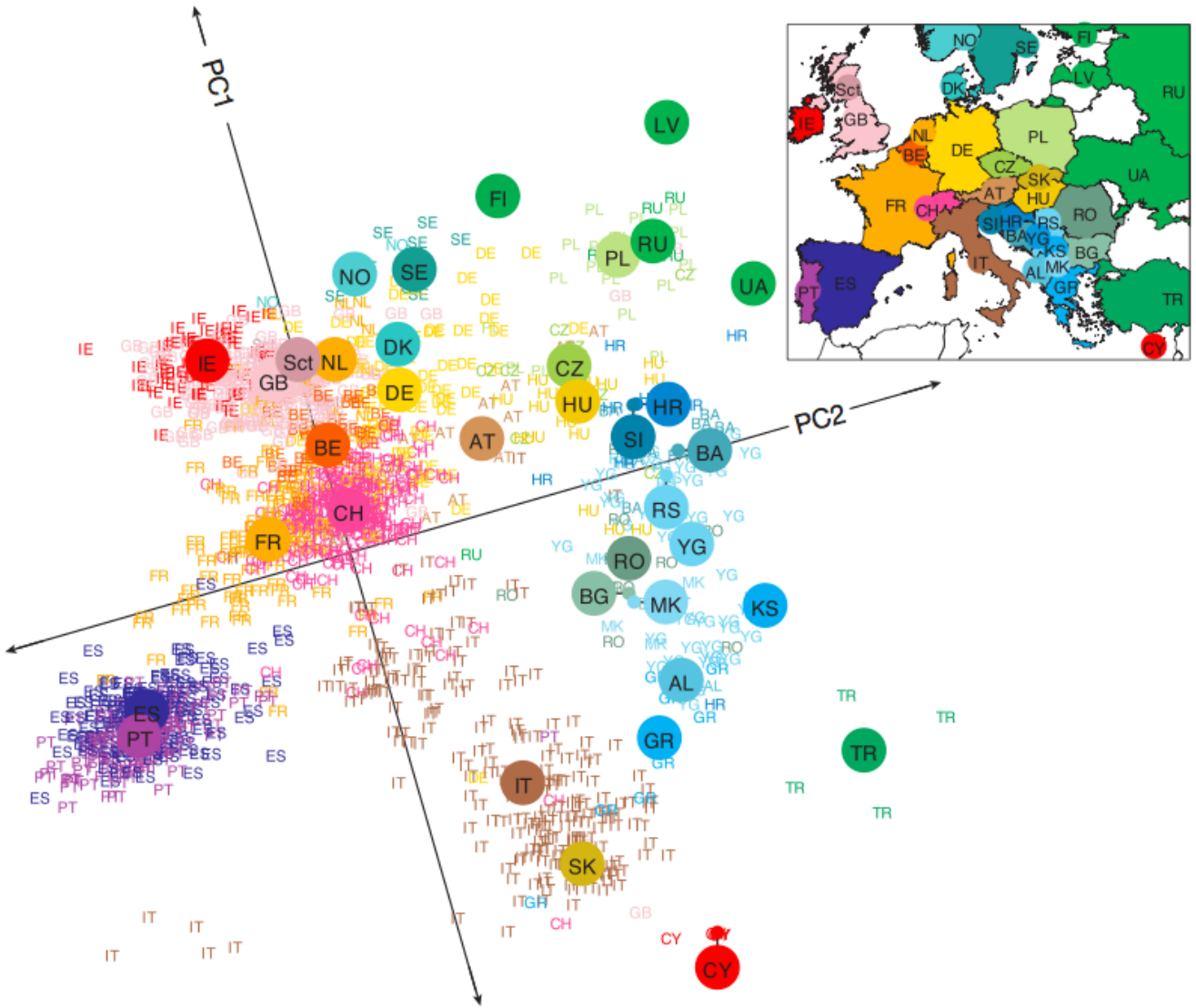
Identifying Genetic Ancestry Differences: Unsupervised Learning with PCA

- Orthogonal axes of variation, i.e. linear combinations of SNPs, that best explain the genotypic variability amongst the n sample individuals are identified.
- For the eigendecomposition we have $\hat{\mathbf{\Psi}} = \mathbf{V}\mathbf{D}\mathbf{V}^T$, where $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n]$ is an $n \times n$ matrix with orthogonal column vectors, and $\mathbf{D}$ corresponding to a diagonal matrix of the length n eigenvalue vector $\mathbf{\Lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)^T$
- The eigenvalues are in decreasing order, $\lambda_1 > \lambda_2 > \ldots > \lambda_n$. The d^{th} principal component (eigenvector) corresponds to eigenvalue λ_d, where λ_d is proportional to the percentage of variability in the genome-screen data that is explained by $\mathbf{V}_d$.

PCA of Europeans

- The top principal components are viewed as continuous axes of variation that reflect genetic variation due to ancestry in the sample.
- Individuals with similar values for a particular top principal component will have similar ancestry for that axes.
- A application of principal components to genetic data from European samples (Novembre et al., Nature 2008) showed that among Europeans for whom all four grandparents originated in the same country, the first two principal components computed using 200,000 SNPs could map their country of origin quite accurately in the plane

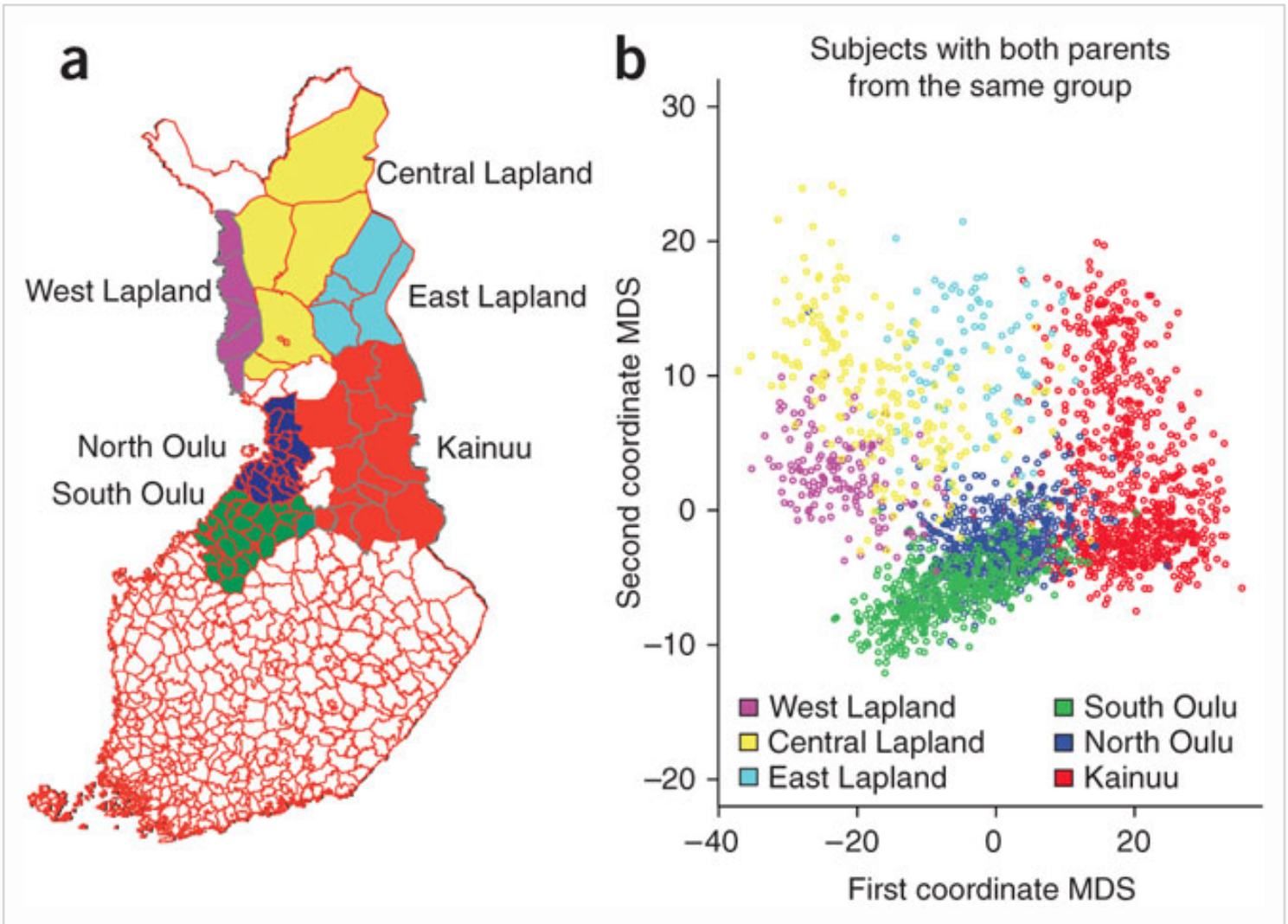
PCA of Europeans



PCA in Finland

- There can be population structure in all populations, even those that appear to be relatively "homogenous"
- An application of principal components to genetic data from Finland samples (Sabatti et al., 2009) identified population structure that corresponded very well to geographic regions in this country.

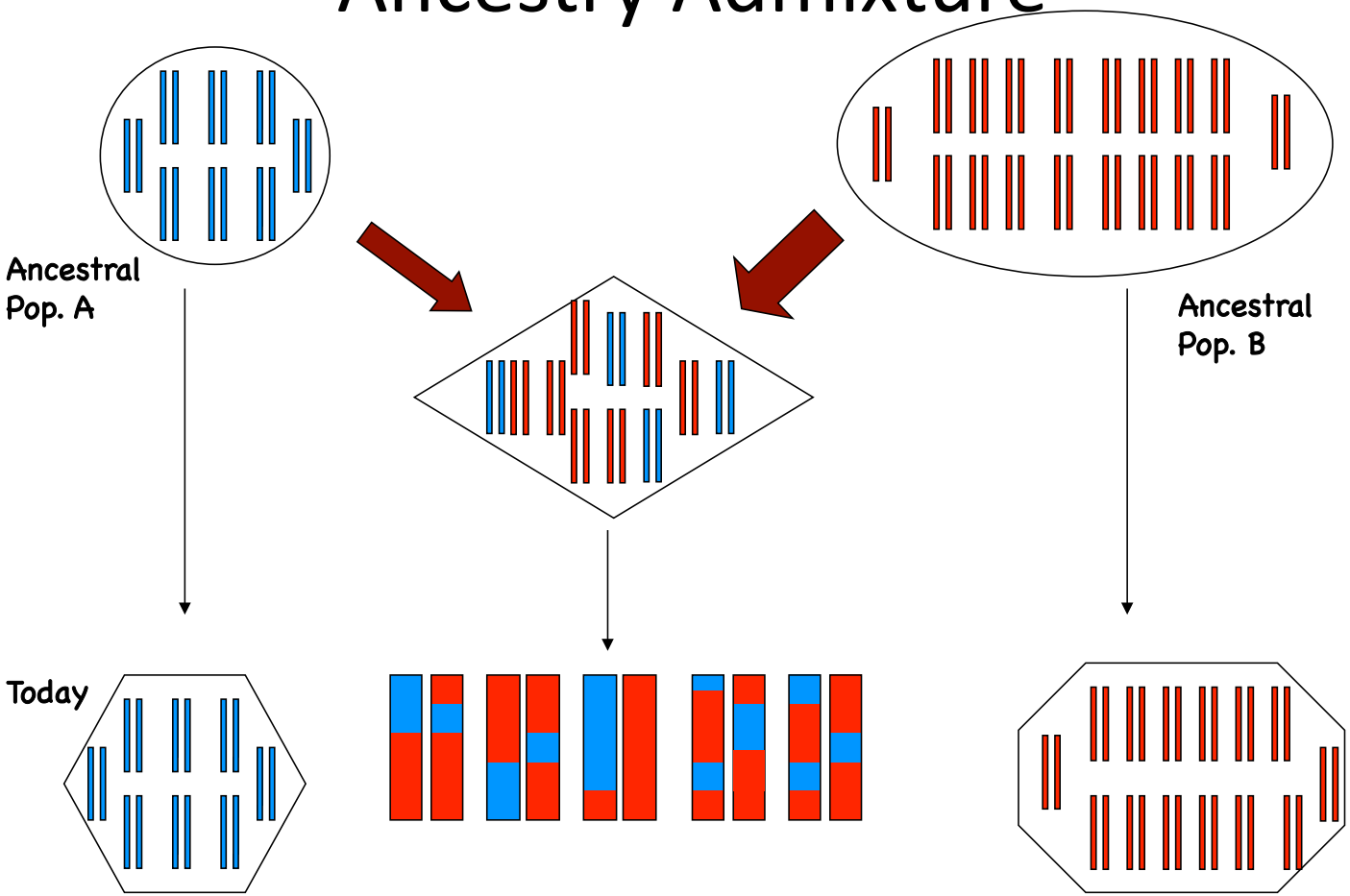
PCA in Finland



Admixed Populations

- Several recent and ongoing genetic studies have focused on **admixed populations**: populations characterized by ancestry derived from two or more ancestral populations that were reproductively isolated.
- Admixed populations have arisen in the past several hundred years as a consequence of historical events such as the transatlantic slave trade, the colonization of the Americas and other long-distance migrations.
- Examples of admixed populations include
  - African Americans and Hispanic Americans in the U.S
  - Latinos from throughout Latin America
  - Uyghur population of Central Asia
  - Cape Verdeans
  - South African "Coloured" population

Ancestry Admixture



- The chromosomes of an admixed individual represent a mosaic of chromosomal blocks from the ancestral populations.

Admixed Populations

- Can be substantial genetic heterogeneity among individuals in admixed populations
- Admixed populations are ancestrally admixed and thus have population structure.
- Statistical method for estimating admixture proportions using genetic data are available

Supervised Learning for Ancestry Admixture

- Methods have recently been developed for supervised learning of ancestry proportions for an admixed individuals using high-density SNP data.
- Most use either a hidden Markov model (HMM) or an Expectation-Maximization (EM) algorithm to infer ancestry
- Example: Suppose we are interested in identifying the ancestry proportions for an admixed individual
- Observed sequence on a chromosome for an admixed individual:

...TATACGTGCACCTGGATTACAGATTACAGATTACAGATTACATTGCATCGATCGAA...

- Observed sequence on a chromosome for samples selected from a "homogenous" reference population:

...TGATCCTGAACCTAGATTACAGATTACAGATTACAGATTACAATGCTTCGATGGAC...

...AGATCCTGAACCTAGATTACAGATTACAGATTACAGATACCAATGCTTCGATGGAC...

...CGATCCTGAACCTAGATTACAGATTACAGATTTGCGTATACAATGCTTCGATGGAC...

HapMap Admixed Populations

- The International Haplotype Map Project (HapMap) contains genotype data on admixed individuals.
- The Mexican Americans in Los Angeles, California (MXL) samples consist of 86 genotyped individuals
- The African American individuals in the southwestern USA (ASW) population samples consist of 87 genotyped individuals.

HapMap ASW and MXL Ancestry

- Genome-screen data on 150,872 autosomal SNPs was used to estimate ancestry
- Estimated genome-wide ancestry proportions of every individual using the ADMIXTURE (Alexander et al., 2009) software
- A supervised analysis was conducted using genotype data from the following reference population samples for three "ancestral" populations
  - HapMap YRI for West African ancestry
  - HapMap CEU samples for northern and western European ancestry
  - HGDP Native American samples for Native American ancestry.

Supervised Ancestry Inference

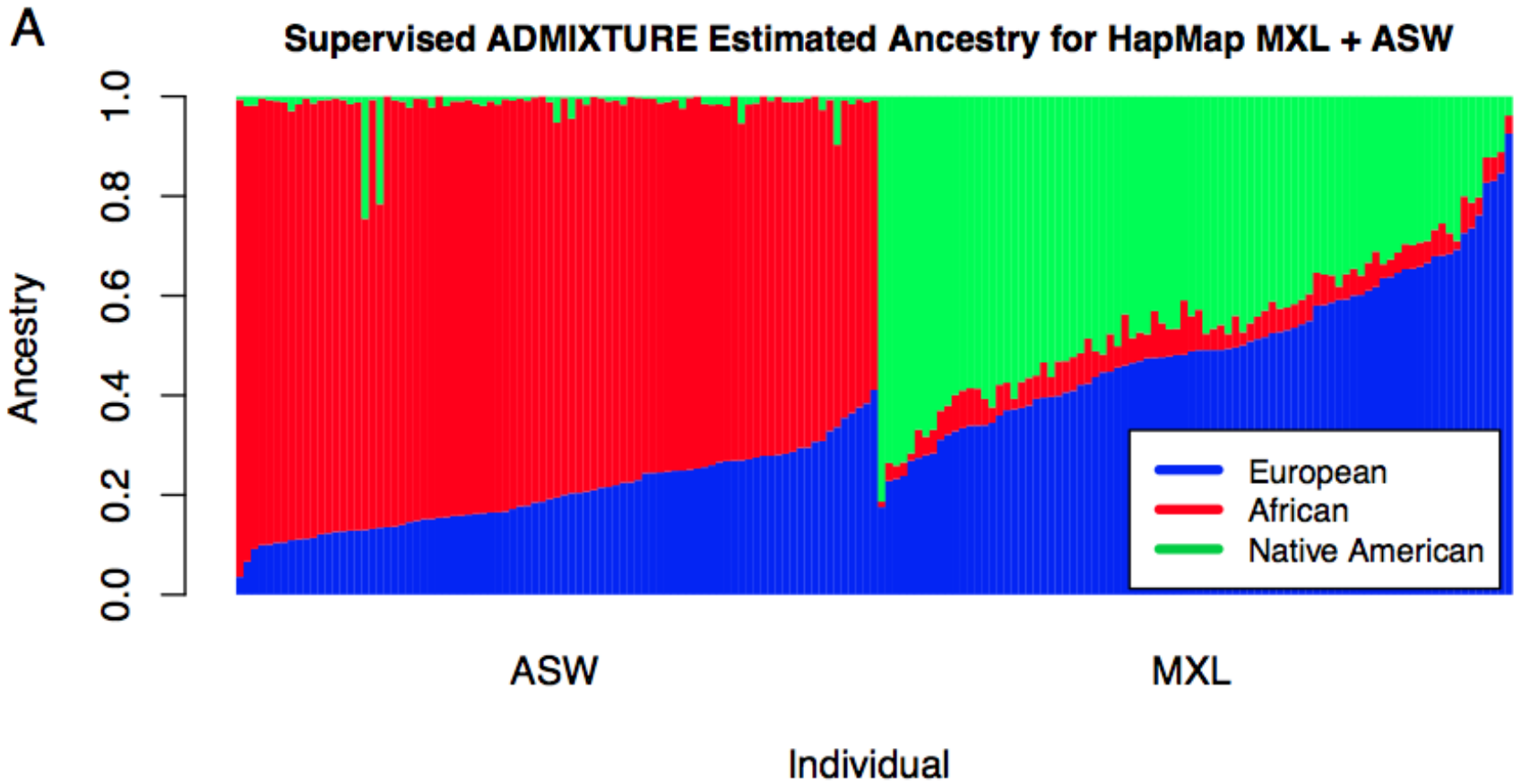


Table: Average Estimated Ancestry Proportions for HapMap African Americans and Mexican Americans

Population	Estimated Ancestry Proportions (SD)		
	European	African	Native American
MXL	49.9% (14.8%)	6%(1.8%)	44.1% (14.8%)
ASW	20.5% (7.9%)	77.5% (8.4%)	1.9% (3.5%)

References

- Alexander, D.H., Novembre, J., Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**,1655-1664.
- Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R. (2008). Genes mirror geography within Europe. *Nature* **456**, 98-101.
- Patterson,N., Price, A.L., Reich, D. (2006) Population structure and eigenanalysis. *PLoS Genet.* **2**, e190.
- Sabatti, C., Service, S. K., Hartikainen, A. L., Pouta, A., Ripatti, S., Brodsky, J., et al. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics.*, **41**, 35-46.