Module 17: Computational Pipeline for WGS Data

Population Structure Inference from Genetic Data

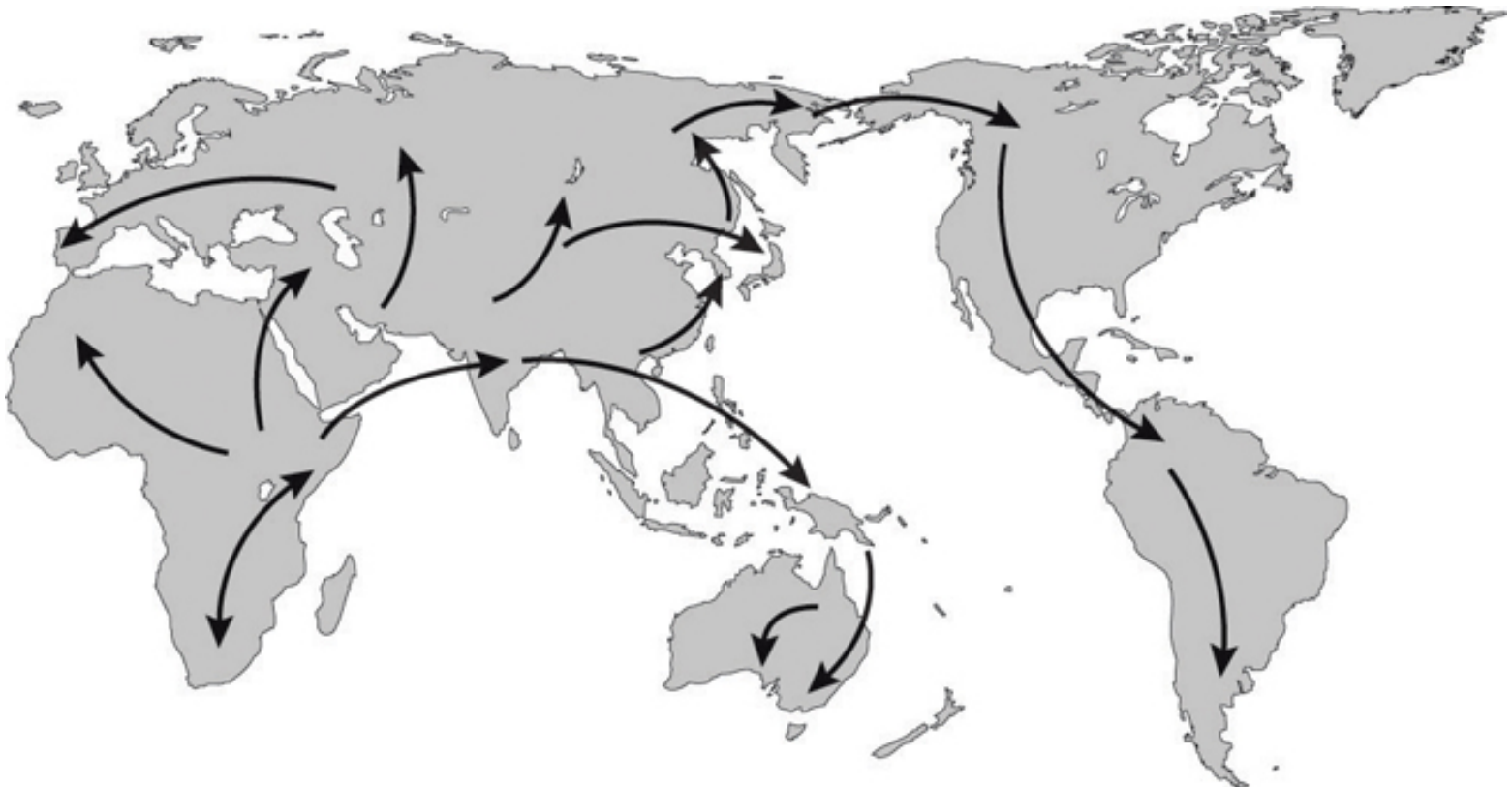
Timothy A. Thornton

University of Washington, TOPMed Data Coordinating Center

Summer Institute in Statistical Genetics
July 2019

Background: Population Structure

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



Population Structure Inference

- 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 are expected to have "similar" ancestry for that axes.

Standard Principal Components Analysis (sPCA)

- sPCA is an unsupervised learning tool for dimension reduction in multivariate analysis.
- Widely used in genetics community to infer population structure from genetic data.
  - Belief that top principal components (PCs) will reflect population structure in the sample.
- Orthogonal linear transformation to a new coordinate system
  - sequentially identifies linear combinations of genetic markers that explain the greatest proportion of variability in the data
  - these define the axes (PCs) of the new coordinate system
  - each individual has a value along each PC
- EIGENSOFT (Price et al., 2006) is a popular implementation of PCA.

Data Structure

- Sample of N individuals, indexed by $i = 1, 2, \ldots, N$.
- Genome screen data on M genetic autosomal markers, indexed by $m = 1, 2, \ldots, M$.
- At each marker, for each individual, we have a genotype value, g_{im}.
- Here we consider bi-allelic markers, so g_{im} takes values 0, 1, or 2, corresponding to the number of reference alleles.
- We center and standardize these genotype values:

$$z_{im} = \frac{g_{im} - 2\hat{p}_m}{\sqrt{2\hat{p}_m(1-\hat{p}_m)}}$$

where $\hat{p}_m$ is an estimate of the reference allele frequency for marker m.

Genetic Correlation Estimation

- Create an N x M matrix, $\mathbf{Z}$, of centered and standardized genotype values, and with $\mathbf{Z}$ we can obtain an N x N genetic relatedness matrix (GRM) for all possible pairs in the sample:

$$\widehat{\mathbf{\Psi}} = \frac{1}{M}\mathbf{Z}\mathbf{Z}^T$$

- The (i, j)th element of this matrix is

$$\widehat{\mathbf{\Psi}}_{ij} = \frac{1}{M}\sum_{m=1}^{M}\frac{(g_{im} - 2\hat{p}_m)(g_{jm} - 2\hat{p}_m)}{2\hat{p}_m(1-\hat{p}_m)},$$

  where $\widehat{\mathbf{\Psi}}_{ij}$ can be viewed as an estimate of the genome-wide average genetic correlation between individuals i and j.
- Individuals from the same ancestral population are expected to have genotypic values that are more correlated than individuals from different ancestral populations.

Standard Principal Components Analysis (sPCA)

- PCA is performed by obtaining the eigen-decomposition of $\widehat{\mathbf{\Psi}}$; that is, we find **eigenvectors** and **eigenvalues** such that $\widehat{\mathbf{\Psi}} = \mathbf{V}^T\mathbf{L}\mathbf{V}$ where
  - $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_N]$ is a N x N matrix consisting of N eigenvectors, each of length N
  - $\mathbf{L}$ is a diagonal matrix of N eigenvalues, $(\lambda_1 > \lambda_2 > \ldots > \lambda_N)$, that are in decreasing order, i.e.,

$$\mathbf{L} = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \ldots & 0 & \lambda_N \end{bmatrix}$$

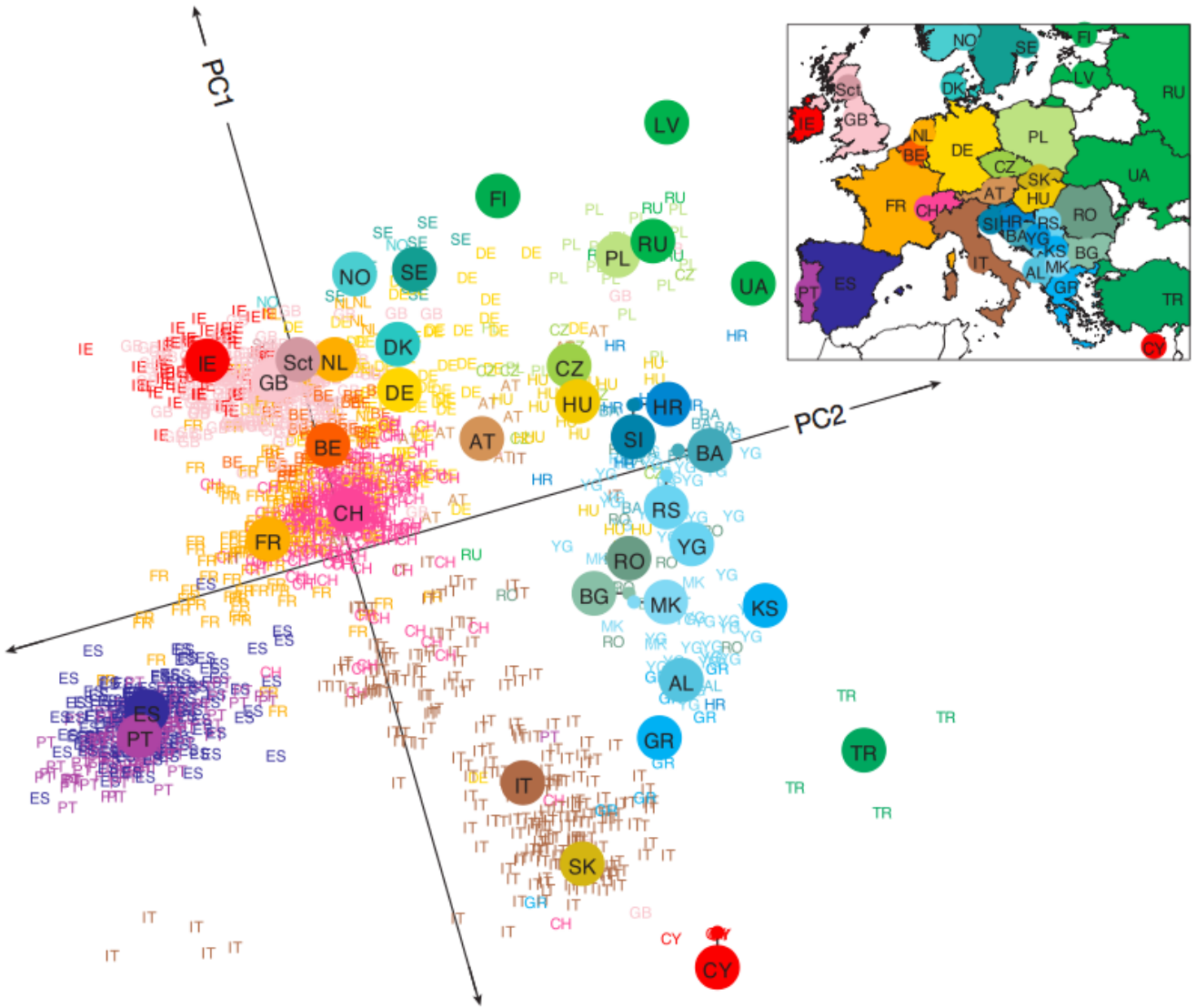
Standard Principal Components Analysis (sPCA)

- 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 $\mathbf{V}_d$.
- The top principal components are viewed as continuous axes of variation that reflect genetic variation that best explain genotypic variability amongst the N sample individuals.
- Individuals with "similar" values for a particular top principal component are expected to have "similar" ancestry for that axes.
- As a result, eigenvectors (PCs) are often used as surrogates for ancestry (or population structure).

PCA of Europeans

- In a very influential paper, an application of PCA to genetic data from European samples (Novembre et al., 2008) illustrated 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 a plane

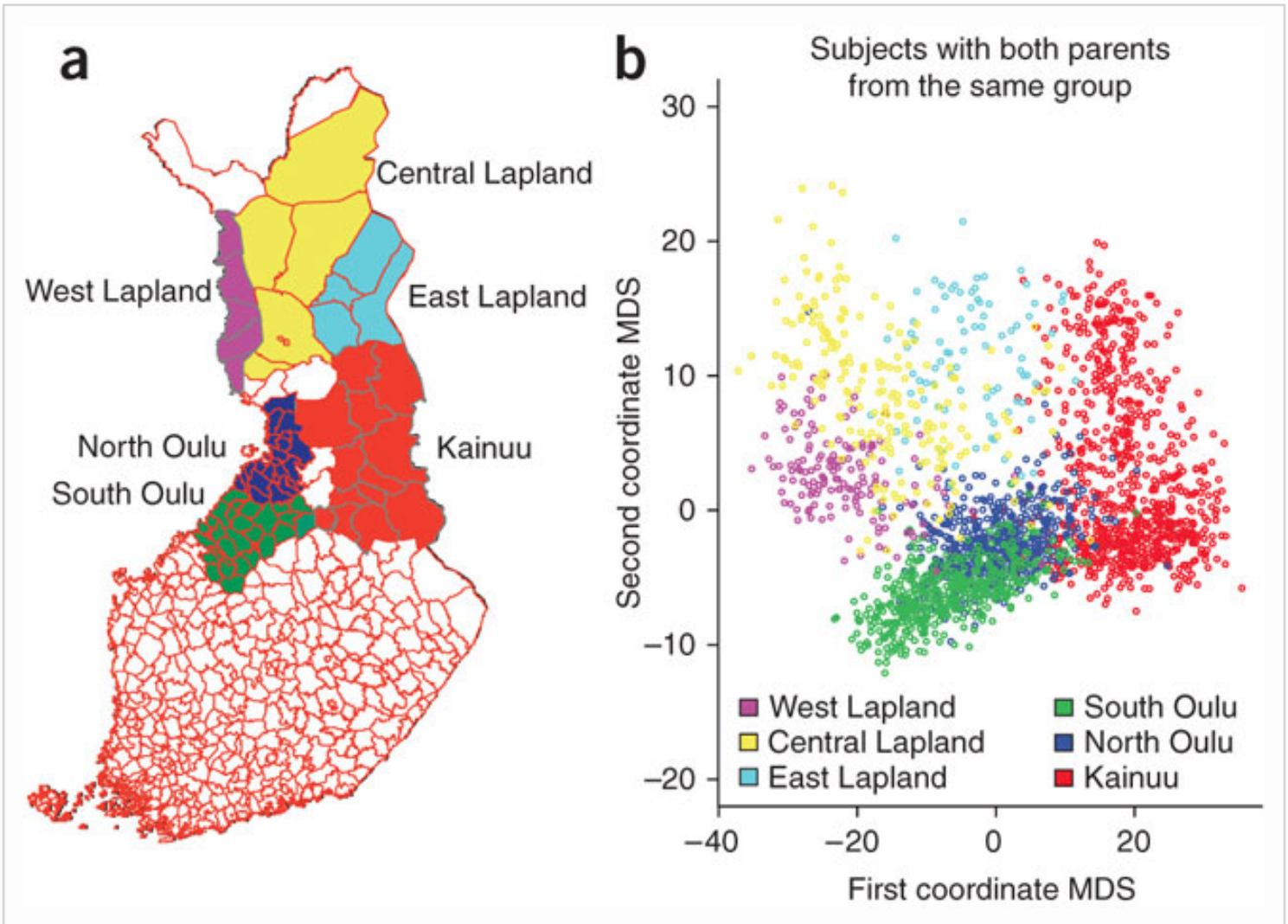
PCA of Europeans



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.

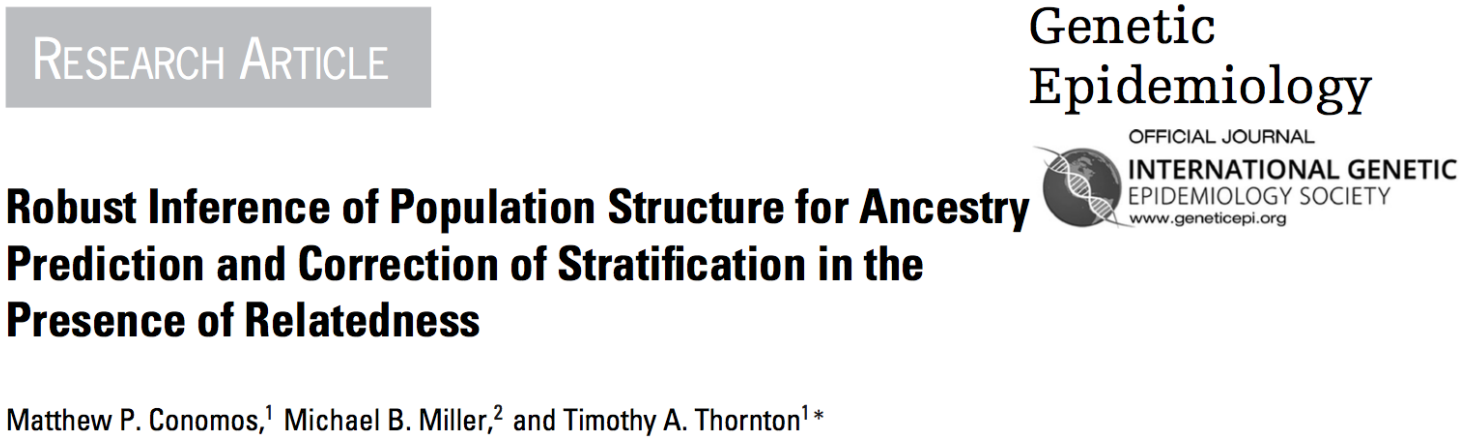
PCA in Finland



Caution: Relatedness Confounds sPCA

- Recall that the elements in the GRM used by sPCA, $\widehat{\mathbf{\Psi}}_{ij}$, are an estimate of the genome-wide average genetic correlation between individuals i and j.
- Conomos et al. (2015) showed that $\mathbf{\Psi}_{ij} = 2\left[\phi_{ij} + (1 - \phi_{ij})A_{ij}\right]$
  - ϕ_{ij}: kinship coefficient - a measure of familial relatedness (more about this later!)
  - A_{ij}: a measure of ancestral similarity
- sPCA is an unsupervised method; in related samples we don't know the correlation structure each eigenvector is actually reflecting
  - If the only genetic correlation structure among individuals is due to ancestry, $\mathbf{\Psi}$ and the top PCs will capture this.
  - If there is relatedness in the sample, the top PCs may reflect this or some combination of ancestry and relatedness.

PCA for Related Samples


RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness

Matthew P. Conomos,[1] Michael B. Miller,[2] and Timothy A. Thornton[1]*

The PC-AiR method was developed for performing a **P**rincipal **C**omponents **A**nalysis **i**n **R**elated samples. The algorithm has the following steps:

PC-AiR: PCA for Related Samples

- The PC-AiR Algorithm
  1. Use known or estimated relatedness of all pairs of individuals in the sample.
  2. Partition the sample into an unrelated set and a related set of individuals.
  3. Perform standard PCA on the set of unrelated individuals.
  4. Predict PC values for the set of related individuals based on genetic similarities with the unrelated set via projection

(1) Use either known or estimated relatedness for all pairs of individuals (More on this in the next lecture!)

- **Kinship coefficients** are common measures of relatedness for pairs of individuals; they are a measure of genetic correlation.

Relationship	ϕ
Parent-Offspring	1/4
Full Siblings	1/4
Half Siblings	1/8
Uncle-Nephew	1/8
First Cousins	1/16
First-Cousins Once Removed	1/32
Unrelated	0

Table: Expected Kinship Coefficients for Common Relationships.

(1) Infer relatedness for all pairs of individuals.

- If pedigree information is known, we can use obtain pedigree-based kinship
- When pedigree information is not available, estimation of relatedness from genotype data is challenging in samples with population structure.
- How can we accurately identify relatives in the structured populations with unknown ancestry?
- Will discuss this in more detail in the next lecture. For now, let's assume that we have reliable inference in pedigree relationships.

(2) Partition the sample into an unrelated and related set.

- We want to partition our sample into two parts: a set of unrelated individuals, and a set of related individuals.
  - The **unrelated set** will consist of individuals who are all unrelated *to each other*.
  - The **related set** will consist of individuals who are related to someone in the unrelated set.
- The optimal unrelated set would:
  - Contain as many individuals as possible.
  - Contain individuals that are representative of all ancestries in the sample.
  - Contain individuals who are highly correlated with their set of relatives in the related set.

KING: Inferring relatedness and pairwise ancestry differences

- Manichaikul A et al. (2010) propose an estimator, KING (**K**inship-based **IN**ference for **G**was), which estimates kinship coefficients for individuals in the presence of population structure.
- Their method does not require allele frequency estimates at the markers; kinship coefficient estimates are based on allele sharing counts as a measure of genetic distance between individuals.

$$\hat{\kappa}_{ij} = \frac{1}{2}\left[1 - \frac{\sum_{m=1}^{M}\left(g_{im} - g_{jm}\right)^2}{N_{Aa}^{i} + N_{Aa}^{j}}\right]$$

where N_{Aa}^{i} and N_{Aa}^{j} are the number of heterozygous genotypes for individuals i and j, respectively.

KING: Inferring differences in ancestry among samples

- The KING method gives unbiased estimates of kinship for individuals of the same ancestry.
- Gives negative estimates for unrelated individuals of different ancestry.
- Provides useful inference on ancestry difference for unrelated individuals.
- Unrelated pairs with the most divergent ancestry will have more extreme negative $\hat{\kappa}_{ij}$ values

(2) Partition the sample into an unrelated and related subsets.

- Any pair of individuals with a kinship coefficient (pedigree-based or estimated) value of $\hat{\phi}_{ij} > \delta$ are declared **related**.
  - The KING paper identified $\delta = 0.022$ as an experimental upper bound on estimates for unrelated individuals of the same ancestry.
- Any pair of individuals with KING estimate $\hat{\kappa}_{ij} < -\delta$ are declared to have **divergent** ancestry.
  - Unrelated individuals of different ancestry have negative KING estimates, and we showed that these estimates are more negative the more 'different' their ancestries are.
  - An individual with unique ancestry will appear divergent from almost everyone else in the sample.

(2) Partition the sample into an unrelated and related set.

Algorithm to obtain an optimal set of unrelated individuals:

1. Compute the number of divergent pairs for each individual: $d_i = \sum_{j \neq i} \mathbb{1}_{[\hat{\kappa}_{ij} < -\delta]}$
2. Compute the 'total kinship' for each individual with all inferred relatives: $t_i = \sum_{j \neq i} \hat{\phi}_{ij} \mathbb{1}_{[\hat{\phi}_{ij} > \delta]}$
3. Among the remaining individuals, count the number of inferred relatives for each; remove the individual with the most
   (a) If there is a tie, the one with $\min(d_i)$ is removed.
   (b) If there is another tie, the one with $\min(t_i)$ is removed.
   (c) If there is a further tie, randomize.
4. The individual removed is placed in the related set.
5. Repeat steps 3 and 4 until all remaining individuals are inferred to be unrelated ($\hat{\phi}_{ij} < \delta$); the remaining individuals are the unrelated set.

(3) Standard PCA on the unrelated set.

- Our sample has N total individuals; let n_u denote the number in the unrelated set and n_r denote the number in the related set.
- We can partition $\mathbf{Z}$ into two matrices: $\mathbf{Z}_u$ and $\mathbf{Z}_r$, containing the genotype values of the unrelated and related sets respectively.
- For the unrelated set, we can create a genetic correlation matrix

$$\widehat{\mathbf{\Psi}}_u = \frac{1}{M}\mathbf{Z}_u\mathbf{Z}_u^T$$

- We perform standard PCA on $\widehat{\mathbf{\Psi}}_u$.
- Since we have removed all relatives, the only correlation structure left in $\widehat{\mathbf{\Psi}}_u$ is that of the population structure; the PCs obtained accurately capture it.

(4) Project PC values for the related set.

- From the sPCA on the unrelated set, we obtain n_u eigenvectors $(\mathbf{V}_1^u, \mathbf{V}_2^u, \ldots, \mathbf{V}_{n_u}^u)$ and eigenvalues $(\lambda_1^u > \lambda_2^u > \ldots > \lambda_{n_u}^u)$.
- Write these in matrix form:

$$\mathbf{V}_u = \left[\mathbf{V}_1^u, \mathbf{V}_2^u, \ldots, \mathbf{V}_{n_u}^u\right] \qquad \mathbf{L} = \begin{bmatrix} \lambda_1^u & 0 & \ldots & 0 \\ 0 & \lambda_2^u & & \vdots \\ \vdots & & \ddots & \\ 0 & \ldots & 0 & \lambda_{n_u}^u \end{bmatrix}$$

- We can calculate an M x n_u SNP weight matrix that gives the relative influence of each SNP on each of the n_u eigenvectors:

$$\mathbf{W} = \mathbf{Z}_u^T \mathbf{V}_u$$

(4) Project PC values for the related set.

- Recall that an eigenvector decomposition of a matrix can be written:

$$\begin{aligned} \mathbf{V}_u\mathbf{L}_u &= \mathbf{\Psi}_u\mathbf{V}_u = \left(\frac{1}{M}\mathbf{Z}_u\mathbf{Z}_u^T\right)\mathbf{V}_u = \frac{1}{M}\mathbf{Z}_u\mathbf{W}_u \\ \mathbf{V}_u &= \frac{1}{M}\mathbf{Z}_u\mathbf{W}_u\mathbf{L}_u^{-1} \end{aligned}$$

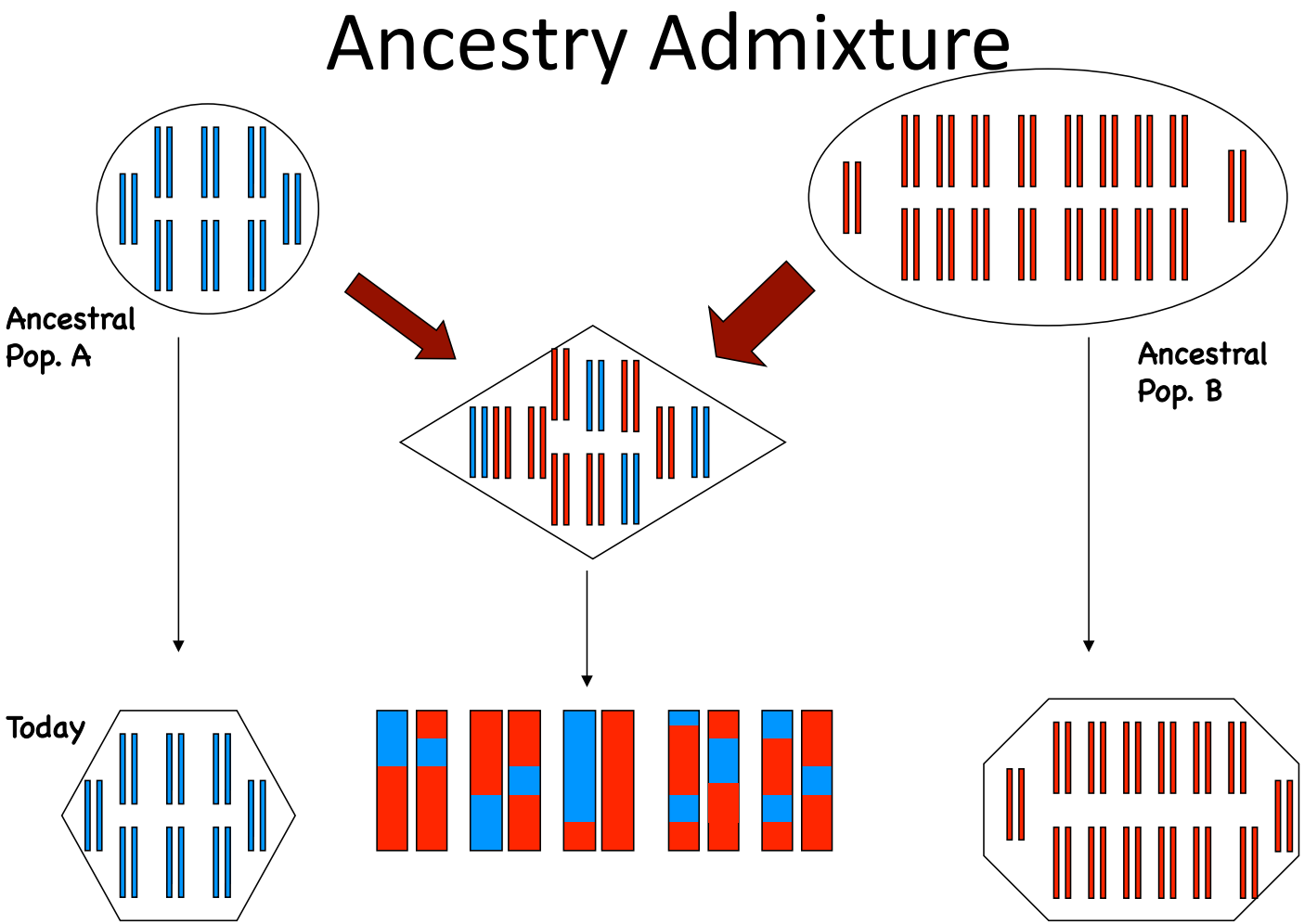
- Replacing the genotype values, $\mathbf{Z}_u$, with those from the related set, $\mathbf{Z}_r$, we can project an n_r x n_u matrix of pseudo-eigenvectors, $\mathbf{Q}_r$, for the related individuals.

$$\mathbf{Q}_r = \frac{1}{M}\mathbf{Z}_r\mathbf{W}_u\mathbf{L}_u^{-1}$$

- Combining $\mathbf{V}_u$ for the unrelated and $\mathbf{Q}_r$ for the related set, we obtain a complete set of eigenvectors for all individuals in the sample that only capture population structure, and *not* family structure.

Recently Admixed Populations

- Several recent genetic studies, including TOPMed, include sampled individuals from **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



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

Recently 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 of Ancestry Admixture

- Methods, such as ADMIXTURE (Alexander et al., 2009), 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 genome-wide or global ancestry
- Other methods, such as RFMix (Maples et al., 2013) have been implemented to infer local ancestry of admixed individuals, i.e., ancestry at specific locations on the genome.

Supervised Learning of Ancestry Admixture

- Example: We are interested in identifying the ancestry proportions for an admixed individual
- Suppose the observed sequence on a chromosome for an admixed individual is:

...TATACGTGCACCTGGATTACAGATTACAGATTACAGATTACATTGCATCGATCGAA...

- Assume that we have a suitable reference panel with diverse ancestries, and a similar sequence is observed in samples from one of the "homogenous" reference populations:

...TGATCCTGAACCTAGATTACAGATTACAGATTACAGATTACAATGCTTCGATGGAC...

...AGATCCTGAACCTAGATTACAGATTACAGATACCAATGCTTCGATGGAC...

...CGATCCTGAACCTAGATTACAGATTTGCGTATACAATGCTTCGATGGAC...

- Can infer the likelihood of the observed sequence in the admixed individuals being derived from each of the reference population samples. This can be performed across the genome.

Example: HapMap ASW and MXL Ancestry Inference

- 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.

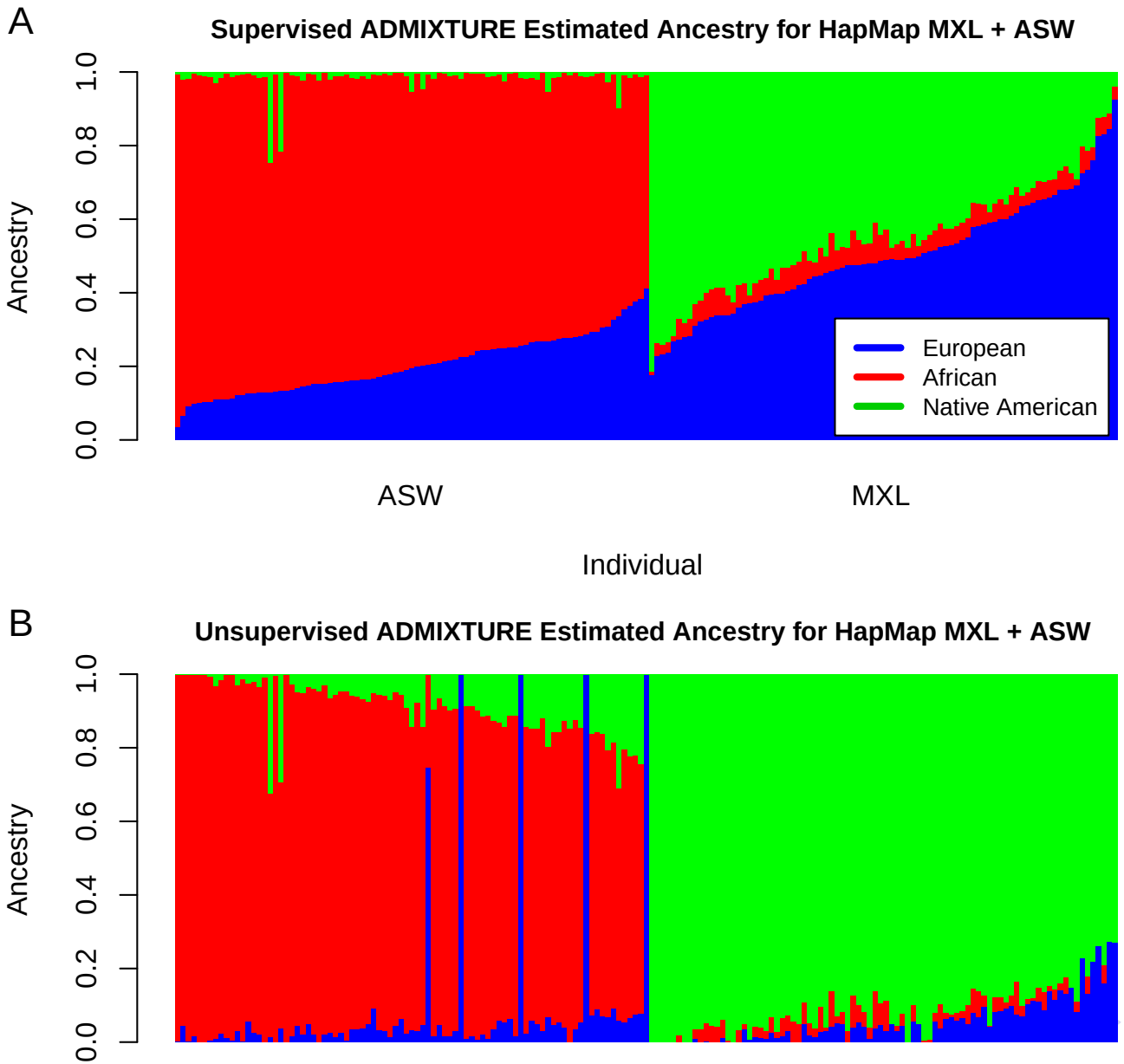
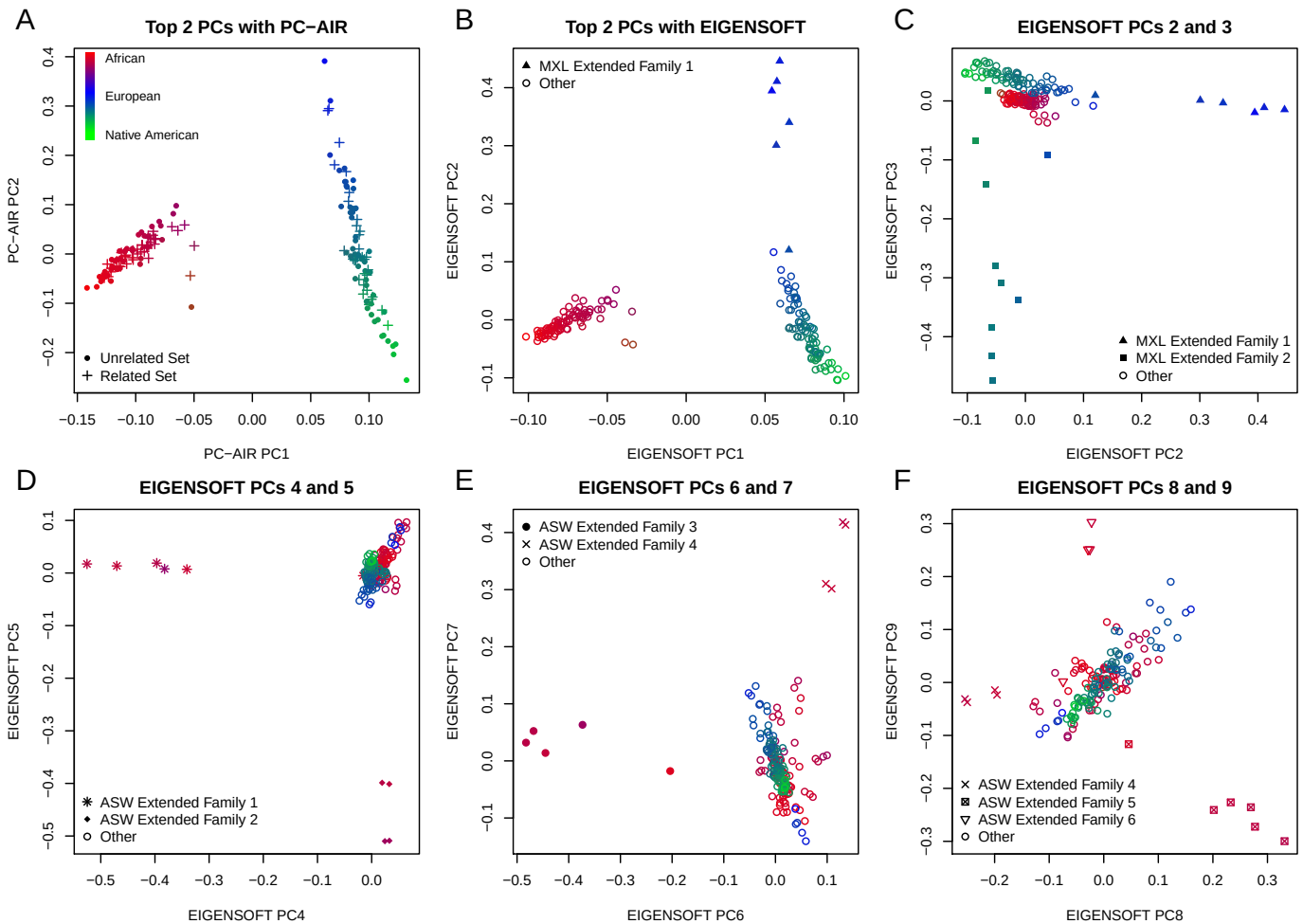

A
Supervised ADMIXTURE Estimated Ancestry for HapMap MXL + ASW
Ancestry
0.0
0.2
0.4
0.6
0.8
1.0
European
African
Native American
ASW
MXL
Individual
B
Unsupervised ADMIXTURE Estimated Ancestry for HapMap MXL + ASW
Ancestry
0.0
0.2
0.4
0.6
0.8
1.0

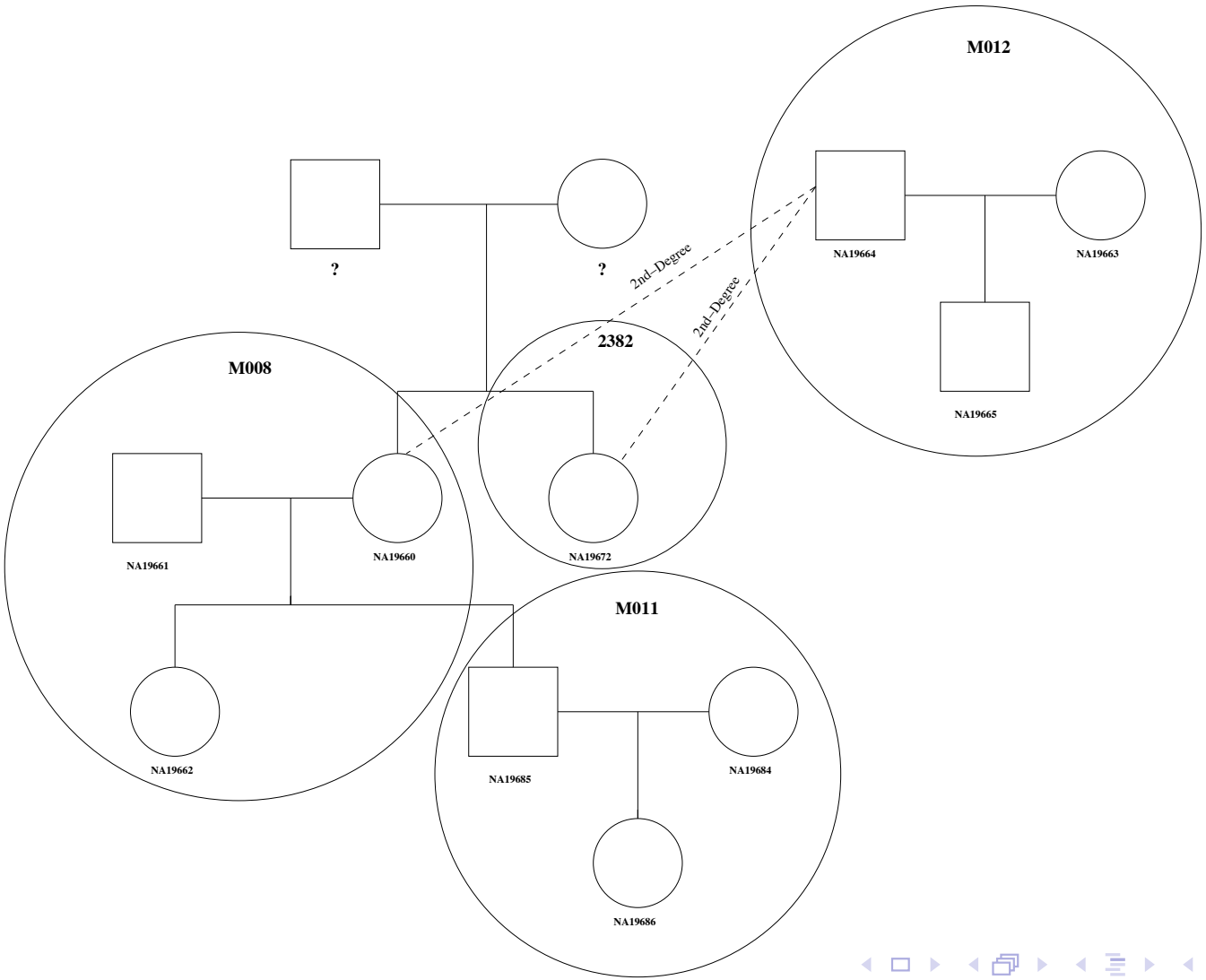
Table: Average Estimated Ancestry Proportions for HapMap African Americans and Mexican Americans

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%)

Figure: **HapMap MXL + ASW Sample**



HapMap MXL: Known and Cryptic Relatedness



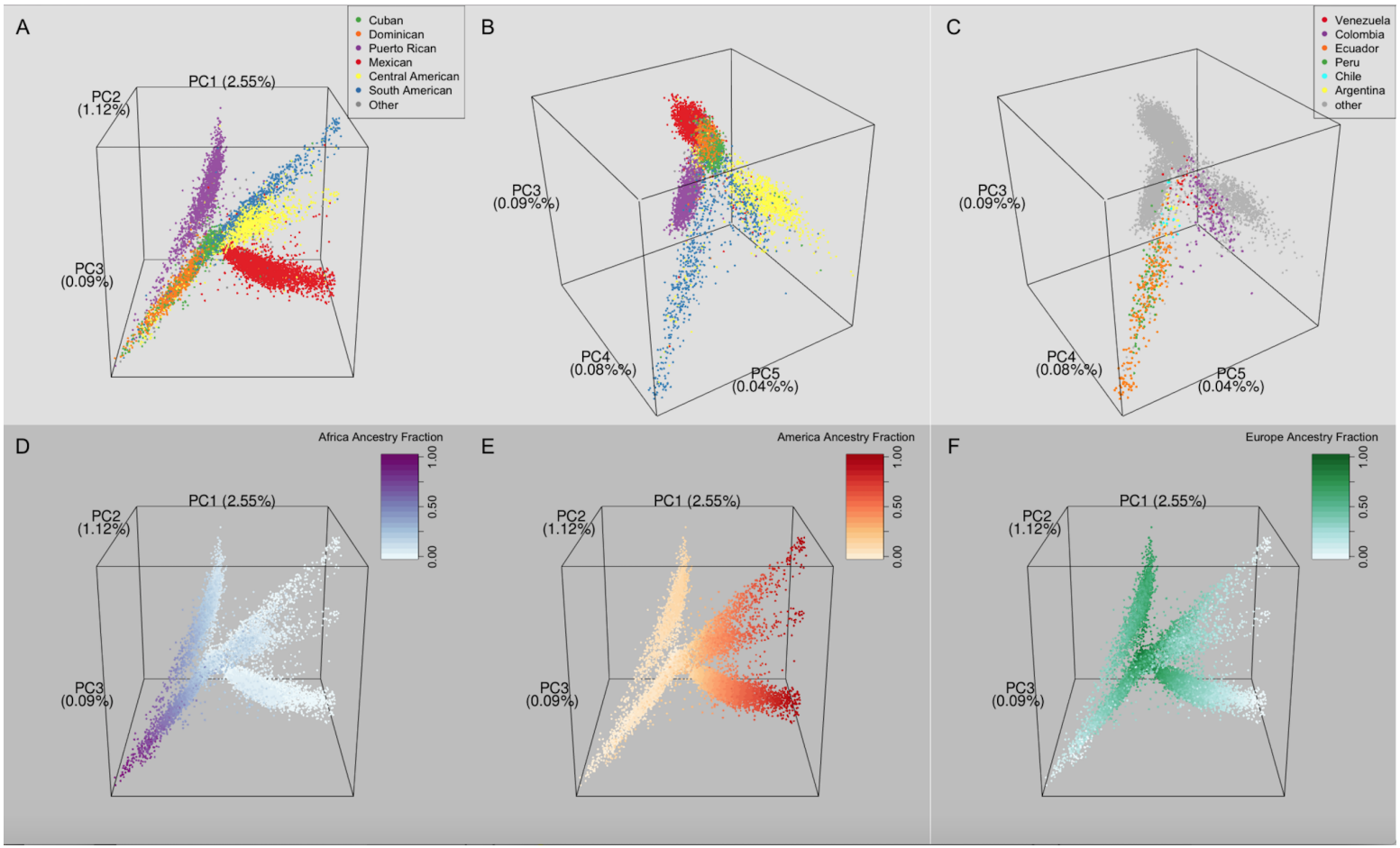
ARTICLE

Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos

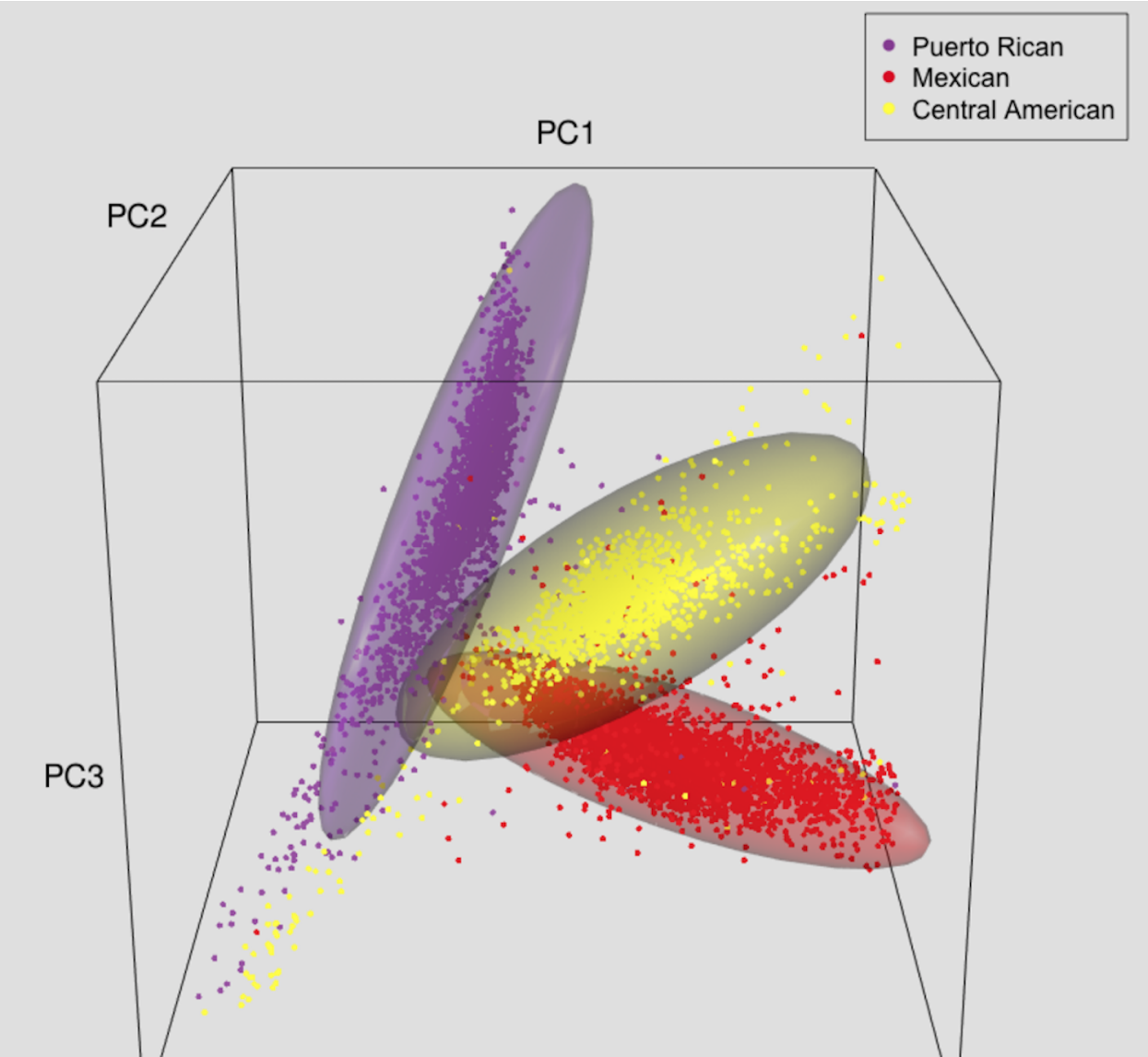
Matthew P. Conomos,[1,14,*] Cecelia A. Laurie,[1,14] Adrienne M. Stilp,[1,14] Stephanie M. Gogarten,[1,14] Caitlin P. McHugh,[1] Sarah C. Nelson,[1] Tamar Sofer,[1] Lindsay Fernández-Rhodes,[2] Anne E. Justice,[2] Mariaelisa Graff,[2] Kristin L. Young,[2] Amanda A. Seyerle,[2] Christy L. Avery,[2] Kent D. Taylor,[3] Jerome I. Rotter,[3] Gregory A. Talavera,[4] Martha L. Daviglus,[5] Sylvia Wassertheil-Smoller,[6] Neil Schneiderman,[7] Gerardo Heiss,[2] Robert C. Kaplan,[6] Nora Franceschini,[2] Alex P. Reiner,[8] John R. Shaffer,[9] R. Graham Barr,[10] Kathleen F. Kerr,[1] Sharon R. Browning,[1] Brian L. Browning,[11] Bruce S. Weir,[1] M. Larissa Avilés-Santa,[12] George J. Papanicolaou,[12] Thomas Lumley,[13] Adam A. Szpiro,[1] Kari E. North,[2] Ken Rice,[1] Timothy A. Thornton,[1] and Cathy C. Laurie[1,*]

- "Genetic diversity and association studies in US Hispanic/Latino populations: Applications in the Hispanic Community Health Study/Study of Latinos." (2016) *American Journal of Human Genetics* 98(1), 165-184.

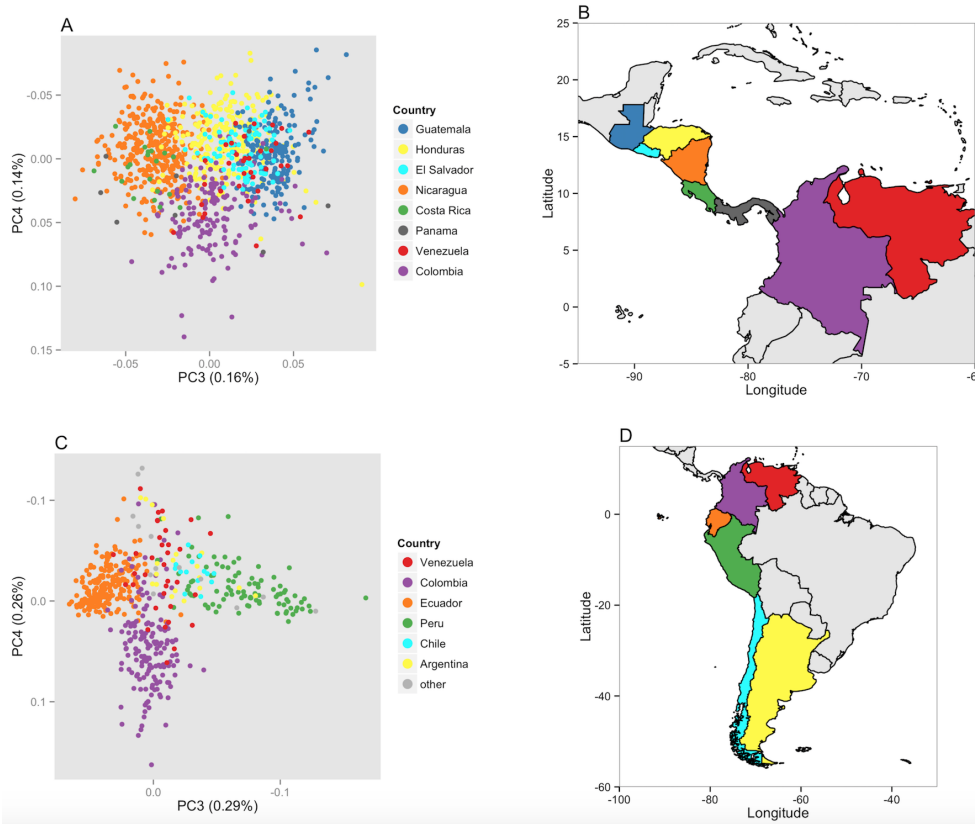
PCA-AiR: Hispanic Community Health Study



PC-AiR: Hispanic Community Health Study



PC-AiR: Hispanic Community Health Study

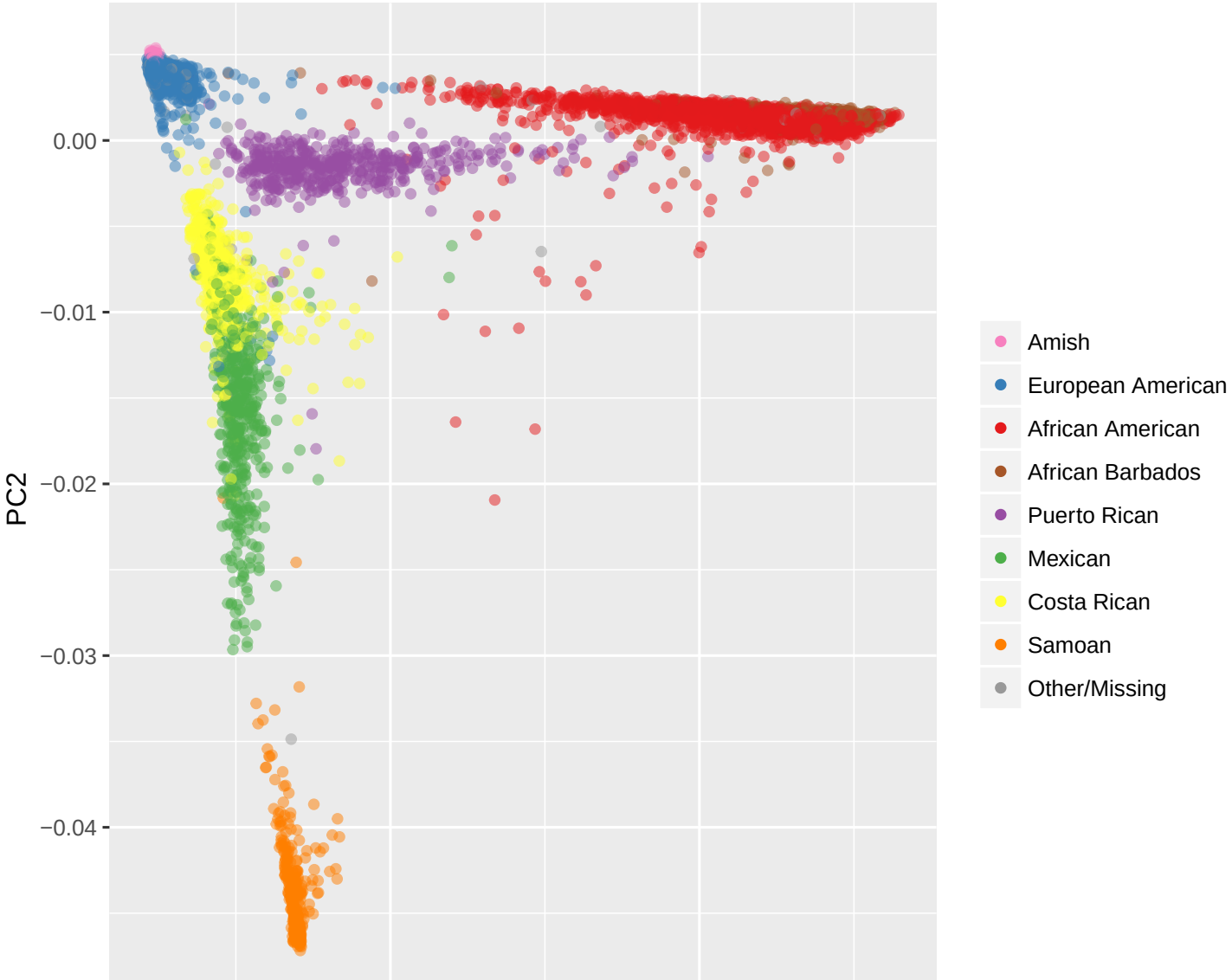


- Genetic differentiation among individuals is associated with the geography of their countries of grandparental origin.

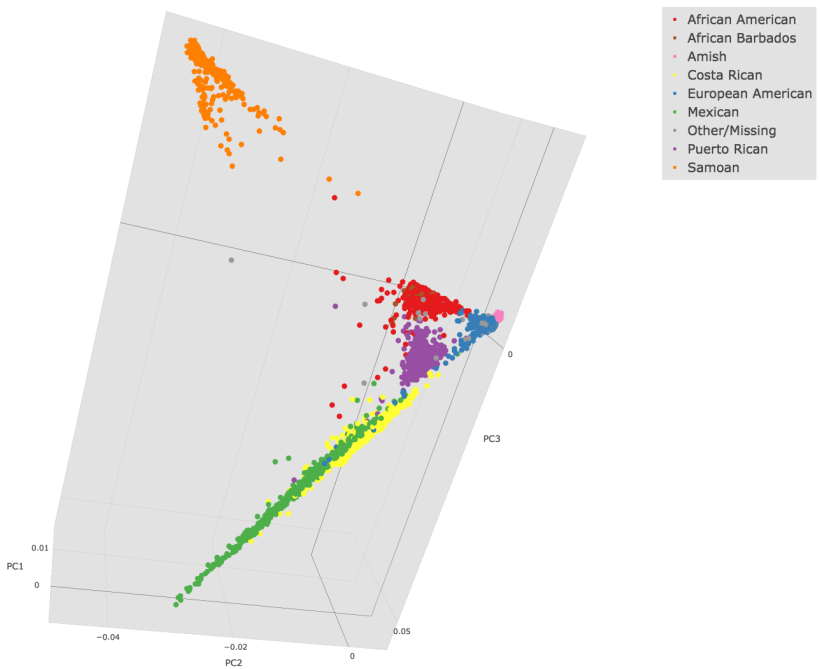
TOPMed Phase I: Population Structure Inference

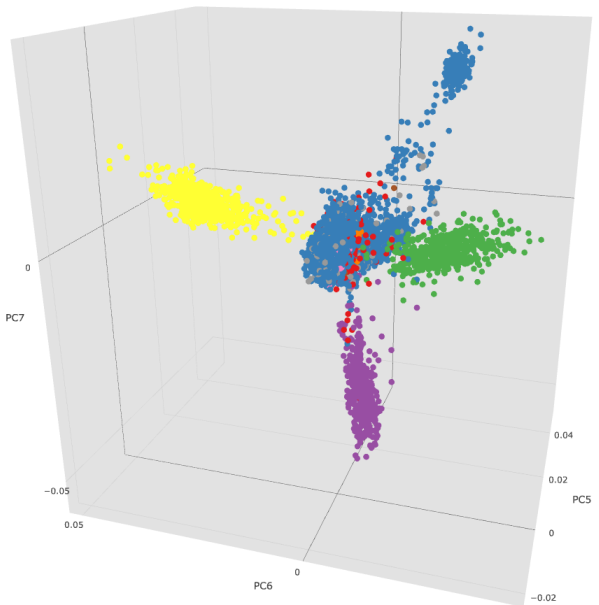
- TOPMed cohorts are multi-ethnic
- Variety of study designs: family-based, case-control, founder populations (Amish).
- PC-AiR algorithm applied to TOPMed Phase I data
- Variants with minor allele frequency $> 1\%$ (common variants) were used for population structure inference

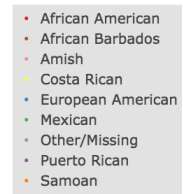
TOPMed Phase I: PC-AiR



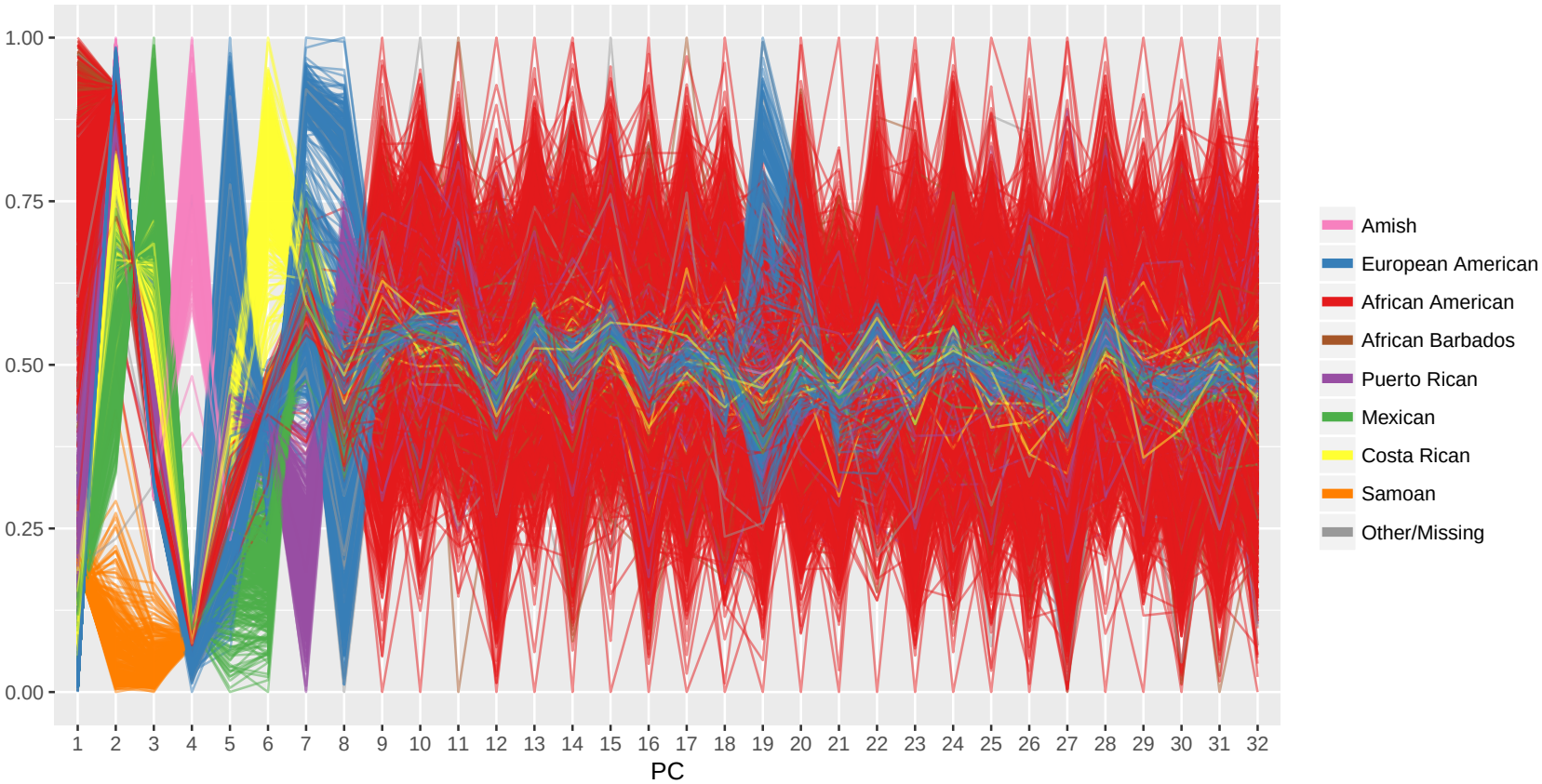
TOPMed Phase I: PC-AiR







TOPMed Phase I: PC-AiR



References

- Alexander, D.H., Novembre, J., Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**,1655-1664.
- Conomos MP, Miller M, Thornton T (2015). Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. *Genetic Epidemiology* **39**, 276-93
- Maples, B. K., Gravel, S., Kenny, E. E., Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, **93**, 278-288.
- Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**, 2867-2873.

References

- 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.