Lecture 3: Population structure inference & Admixed populations

Instructors: Andrey Ziyatdinov and Timothy Thornton

The Unidad de Medicina Experimental, Faculty of Medicine,
National Autonomous University of Mexico

Lecture Overview

1. Population Structure & Inference with PCA
2. Accounting for Relatedness
3. PCA Best Practices
4. Admixed Populations
5. Inference of the Global Ancestry Proportions

Background: Population Structure

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

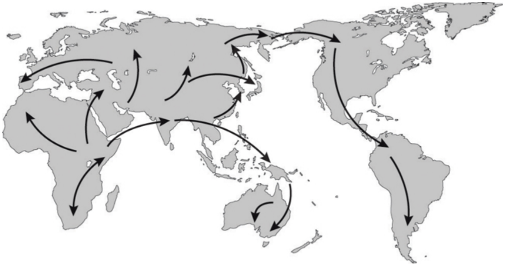


Figure: https://science.education.nih.gov

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
- 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.

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 $l = 1, 2, \ldots, m$.
- At each marker, for each individual, we have a genotype value, G_{il}.
  - Here we consider SNP data, so G_{il} takes values 0, 1, or 2, corresponding to the number of minor alleles.
- We center and standardize these genotype values:

$$z_{il} = \frac{G_{il} - 2\hat{p}_l}{\sqrt{2\hat{p}_l(1 - \hat{p}_l)}}$$

where $\hat{p}_l$ is an estimate of the minor allele frequency for marker l.

Genetic Correlation Estimation

- Create an n x m matrix, $\mathbf{Z}$, of centered and standardized genotype values, and from this, a n x n genetic correlation matrix (GRM):

$$\widehat{\boldsymbol{\Psi}} \quad = \quad \frac{1}{m}\mathbf{Z}\mathbf{Z}^T$$

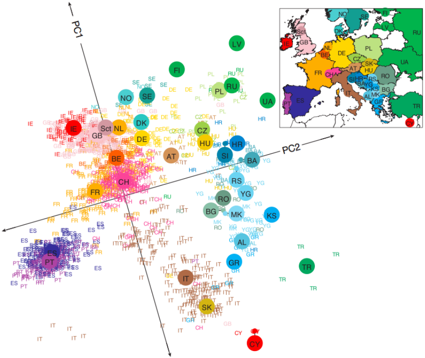
- $\widehat{\boldsymbol{\Psi}}_{ij}$ is an estimate of the genome wide average genetic correlation between individuals i and j.
- PCA is performed by obtaining the eigendecomposition of $\widehat{\boldsymbol{\Psi}}$
  - Single Value Decomposition (SVD) on $\mathbf{Z}/\sqrt{m}$ is equivalent to the eigendecomposition of $\widehat{\boldsymbol{\Psi}}$

Standard Principal Components Analysis (sPCA)

- Identify orthogonal axes of variation, i.e. linear combinations of SNPs, that best explain the genotypic variability between the *n* sample individuals.
- The result is:
  - a set of *n* length *n* eigenvectors, $(\mathbf{V}_1, \mathbf{V}_2, \ldots \mathbf{V}_n)$, where $\mathbf{V}_d$ is a column vector of coordinates of each individual along axis *d*
  - each principal component is a different linear combination of the *m* markers
  - and a corresponding set of *n* eigenvalues, $(\lambda_1 > \lambda_2 > \ldots > \lambda_n)$, in decerasing order.
  - 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 $\mathbf{V}_d$.
- These eigenvectors (PCs) are used as surrogates for population structure

PCA of Europeans

- Application of PCA in European samples (Novembre et al., *Nature* 2008)
- Among Europeans for whom all four grandparents originated in the same country, the first two PCs computed using 200k SNPs could map their country of origin quite accurately



Relatedness Confounds sPCA

- Recall that the GRM used by sPCA, $\widehat{\mathbf{\Psi}}_{ij}$, and is an estimate of the genome wide average genetic correlation between individuals i and j.
- It can be shown:

$$\mathbf{\Psi}_{ij} = 2\left[\phi_{ij} + (1-\phi_{ij})A_{ij}\right]$$

  - ϕ_{ij}: kinship coefficient - a measure of familial relatedness
  - A_{ij}: a measure of ancestral similarity
- PCA is an unsupervised method; in related samples we don't know the correlation structure each eigenvector is 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.
- Association studies have known or cryptic relatedness!

sPCA: Best practices

- Apply QC to variants & samples:
  - Restrict to common variants (e.g. MAF ≥ 0.01)
  - Remove variants with high missing genotypes rates (e.g. ≥ 0.01)
  - Remove variants which fail HWE test (e.g. p-value $\leq 10^{-10}$)
  - Remove samples with high missing genotypes rates (e.g. ≥ 0.1)
  - Keep only variants on autosomal chromosomes
- Remove related individuals (e.g. 3rd degree related or closer)
- Prune variants in linkage disequilibrium (LD) (e.g. $r^2 \geq 0.2$) include long-range LD regions (Price et al., *AJHG*, 2008)

R package bigsnpr

- Apply QC to variants & samples (relies on PLINK2)

  ```
  snp_plinkQC(plink.path, prefix.in,
  file.type="--bfile", maf = 0.01, geno = 0.1,
  mind = 0.1, hwe = 1e-10, autosome.only = TRUE )
  ```

- Remove related individuals (e.g. 3rd degree related or closer)

  ```
   extra.options = "--king-cutoff 0.0442"
  ```

- Compute PCs
  - Prune variants in linkage disequilibrium (LD) (e.g. $r^2 \geq 0.2$)
  - Removes long-range LD regions

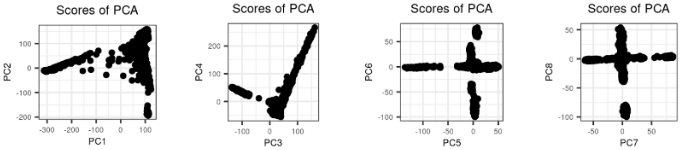
  ```
  pca <- bed_autoSVD(obj.bed, thr.r2 = 0.2, k = 20)
  predict(pca)
  ```

- Project related samples (excluded from training model)

  ```
  bed_projectSelfPCA(object.svd, obj.bed, ind.row)
  ```

R package bigsnpr

```
plot(obj.svd2, type = "scores", scores = 1:20, coeff = 0.4)
```



```
plot(obj.svd2, type = "loadings", loadings = 1:20, coeff = 0.4)
```

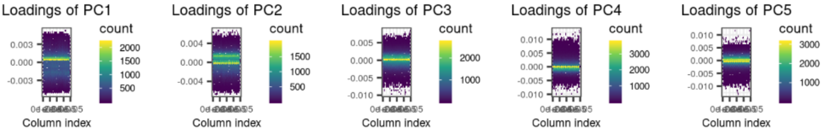
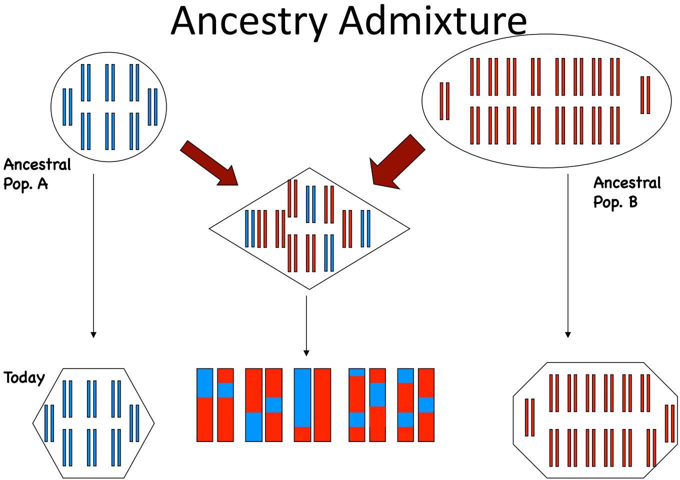


Figure: https://privefl.github.io/bigsnpr/articles/bedpca.html

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



- 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 methods for estimating admixture proportions using genetic data are available

Supervised Learning for Ancestry Admixture

- Methods, such as ADMXITURE and FRAPPE, 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:

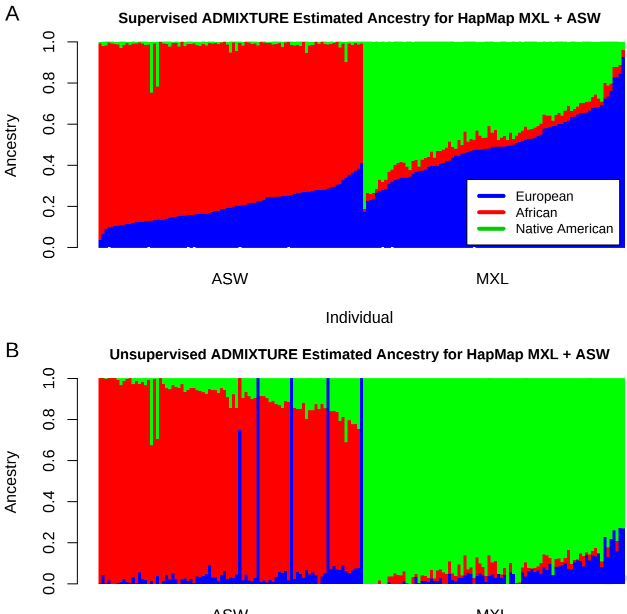
  ...TGATCCTGAACCTAGATTACAGATTACAGATTACAGATTACAATGCTTCGATGGAC...

  ...AGATCCTGAACCTAGATTACAGATTACAGATTACAGATACCAATGCTTCGATGGAC...

  ...CGATCCTGAACCTAGATTACAGATTACAGATTTGCGTATACAATGCTTCGATGGAC...

HapMap 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.

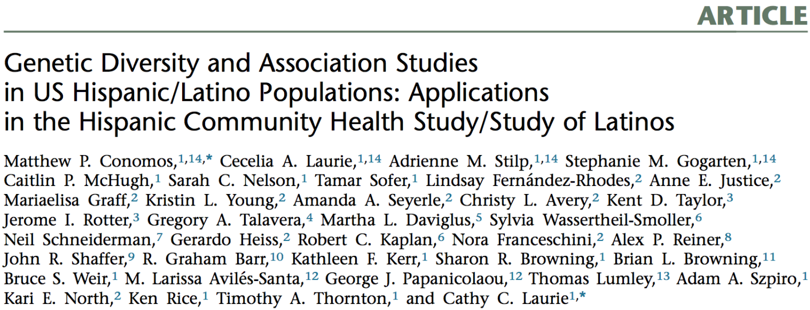


Supervised ADMIXTURE Estimated Ancestry for HapMap MXL + ASW

Unsupervised ADMIXTURE Estimated Ancestry for HapMap MXL + ASW

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

	Estimated Ancestry Proportions (SD)		
Population	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%)



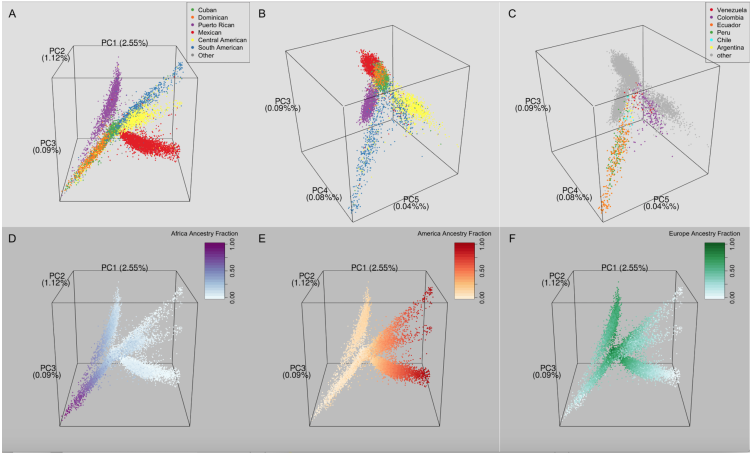
ARTICLE

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

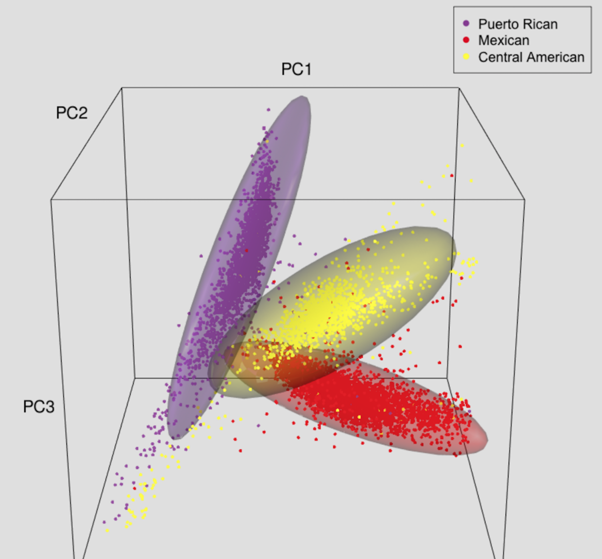
Matthew P. Conomos,[1,14,*] Cecelia A. Laurie,[1,14] Adrienne M. Stilp,[1,14] Stephanie M. Gogarten,[1,14] Caitlin P. McHugh,[1] Sarah C. Nelson,[1] Tamar Sofer,[1] Lindsay Fernández-Rhodes,[2] Anne E. Justice,[2] Mariaelisa Graff,[2] Kristin L. Young,[2] Amanda A. Seyerle,[2] Christy L. Avery,[2] Kent 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.

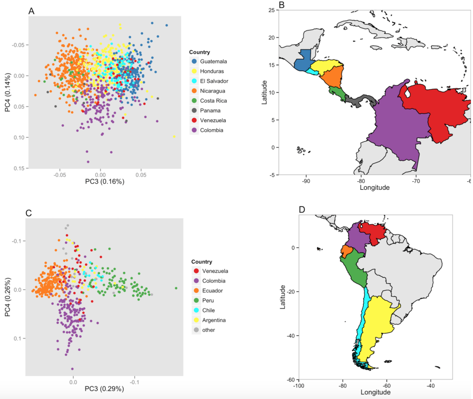
PCA-AiR: Hispanic Community Health Study



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.
- Individuals for whom all four grandparents were born in a specific country in Central or South America were used.

References

- Patterson,N., Price, A.L., Reich, D. (2006) Population structure and eigenanalysis. *PLoS Genet.* **2**, e190.
- 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.
- Alexander, D.H., Novembre, J., Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**,1655-1664.

References

- Price, Alkes L, Michael E Weale, Nick Patterson, Simon R Myers, Anna C Need, Kevin V Shianna, Dongliang Ge, et al. (2008). Long-Range LD Can Confound Genome Scans in Admixed Populations. *The American Journal of Human Genetics*, **83**(1), 132-135.
- 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
- Privé, F., Luu, K., Blum, M. G., McGrath, J. J., Vilhjálmsson, B. J. (2020). Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics*, 36(16), 4449-4457.