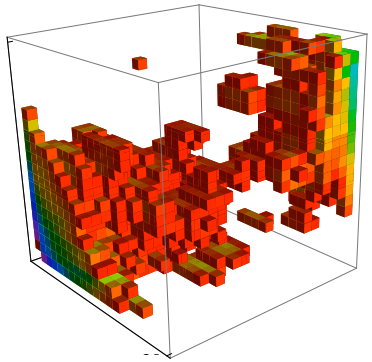
Population genetic inference in the personal genome era



Carlos D. Bustamante

Department of Genetics

Stanford University School of Medicine

Motivation

- Technological developments have dramatically driven down the cost of sequencing and genotyping
- Large-scale projects underway to document genome-wide variation in many species (e.g., human, dogs, rice, cattle) across individuals
  - Quantify genetic differences within and among populations
  - Map genes for traits of interest (e.g., disease-susceptibility, morphology)
  - Reconstruct demographic history and detect targets of recent natural selection
- Personal genomic sequences opens new opportunities and challenges for population genetics.

Motivation and Objectives

- Genome-wide association mapping has been quite successful in identifying common variation (e.g., MAF > 5%) influencing disease risk in Northern European populations
- However, for many traits, only a small proportion of the expected heritability is explained by current GWAS hits (e.g., height -> 150 hits explain 2-4%) and many groups are under-represented in medical genomics research
- Understanding the contribution of rare and common genetic genetic variants will likely require multi-ethnic and trans-ethnic genome-wide studies that compare completely sequenced genomes of many individuals with and without a particular disease
- It will be critical to account for the role of population stratification at fine scales both in terms of genomic and geographic location in these studies

Population Reference Sample (POPRES)

- Assemble large repository of genetically diverse DNA samples (~6,000 samples)
- Generate dense genotype data using key marker panels (Affy 500K)
- Establish resource for studying human population genetics, recent demography, and admixture
- Demographic and genotype data publicly available

Matt Nelson (GSK)

John Novembre (UCLA)

Bustamante Lab GSK





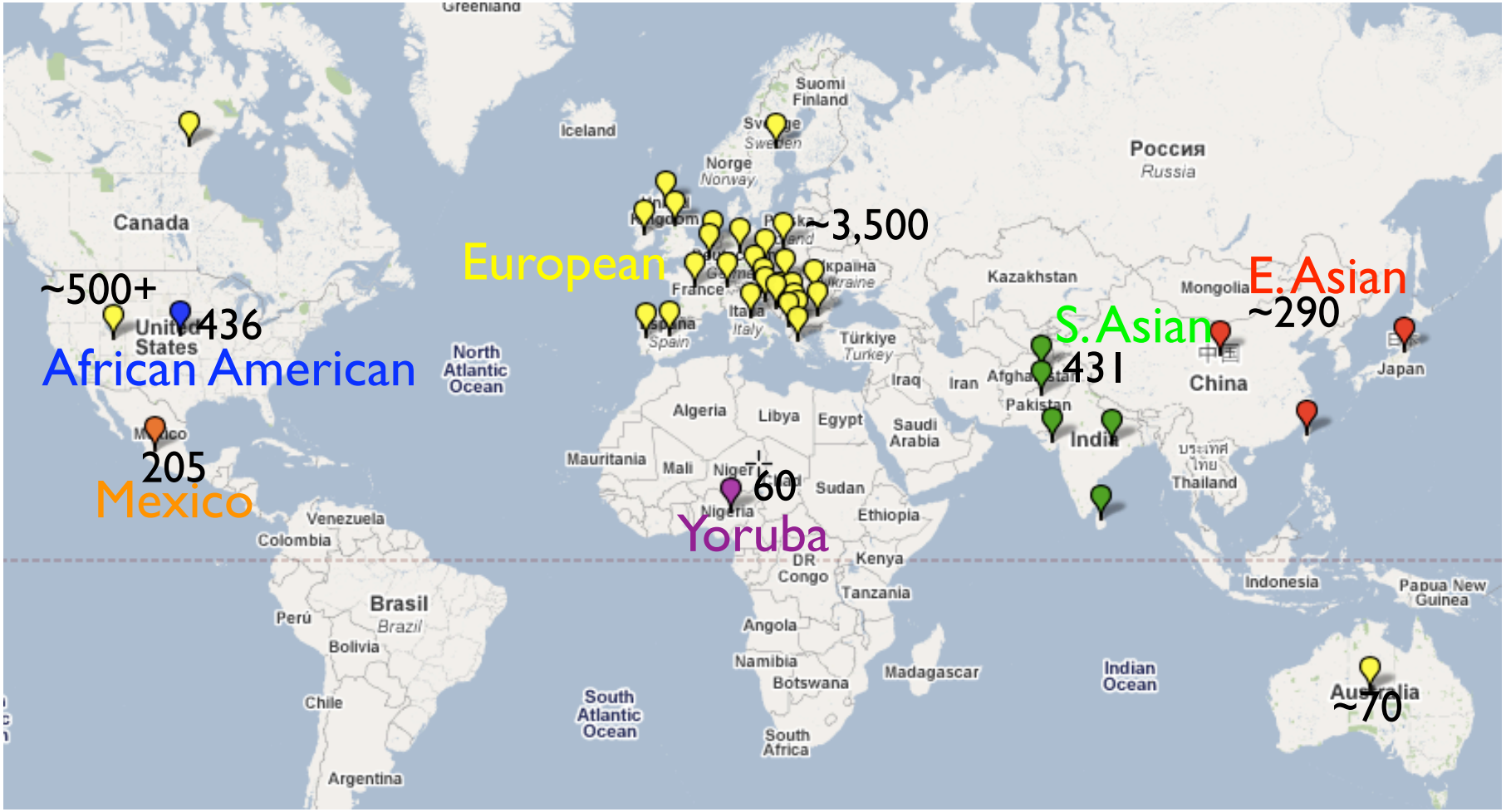

Nelson, *et al.*, AJHG (Sept. 2008)

Questions

- To what degree are modern human populations accessible for medical genetic studies sub-structured?
- What are accurate models for describing human genetic diversity?
- Can we use these kinds of data to reconstruct recent "personal" genetic history?

Nelson, *et al.*, AJHG (Sept. 2008)

PopRes + HapMap Phase II



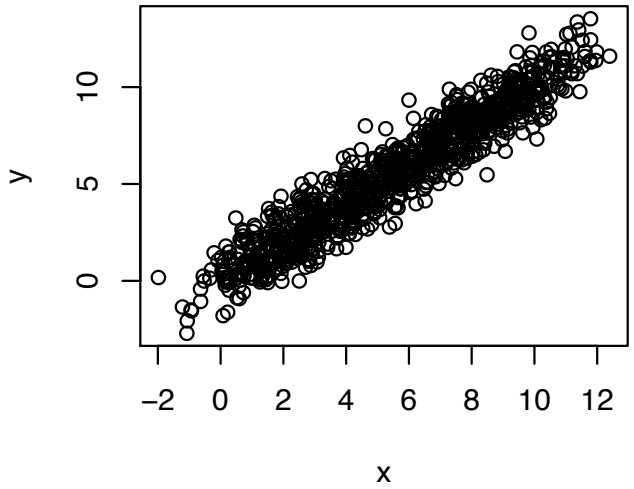
Nelson, *et al.*, AJHG (Sept. 2008)

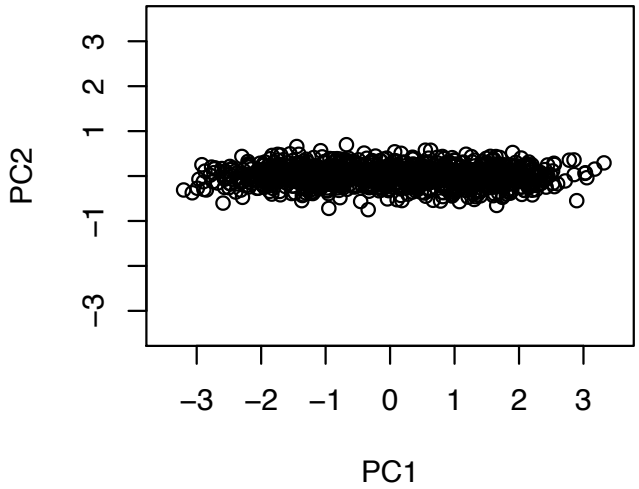
Data Management and Visualization is a Challenge!

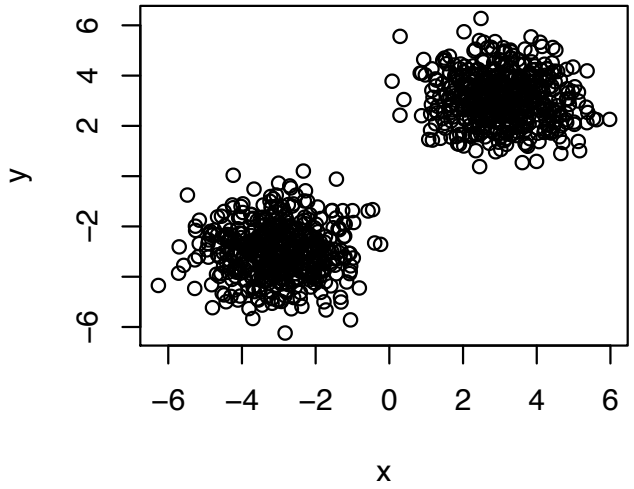
	SNP 1	SNP 2	SNP 3	...	SNP 500,000
Indiv 1	0	2	2	...	2
Indiv 2	1	1	1	...	1
- - -	- - -	- - -	- - -	...	
Indiv *n*	1	0	2	...	0

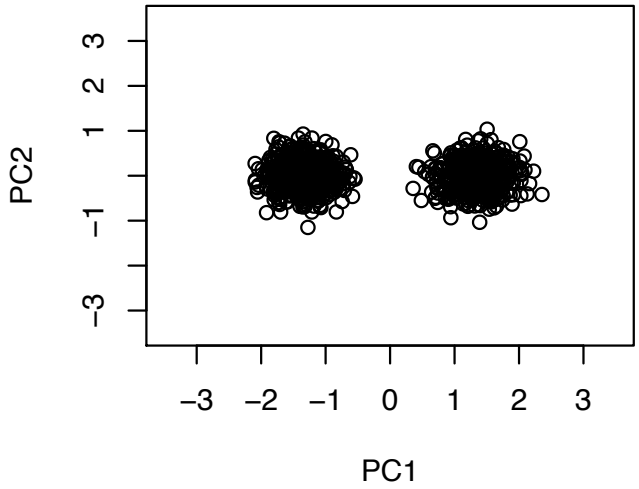
aa	Aa	AA
0	1	2

Principle Components = major axes of variation

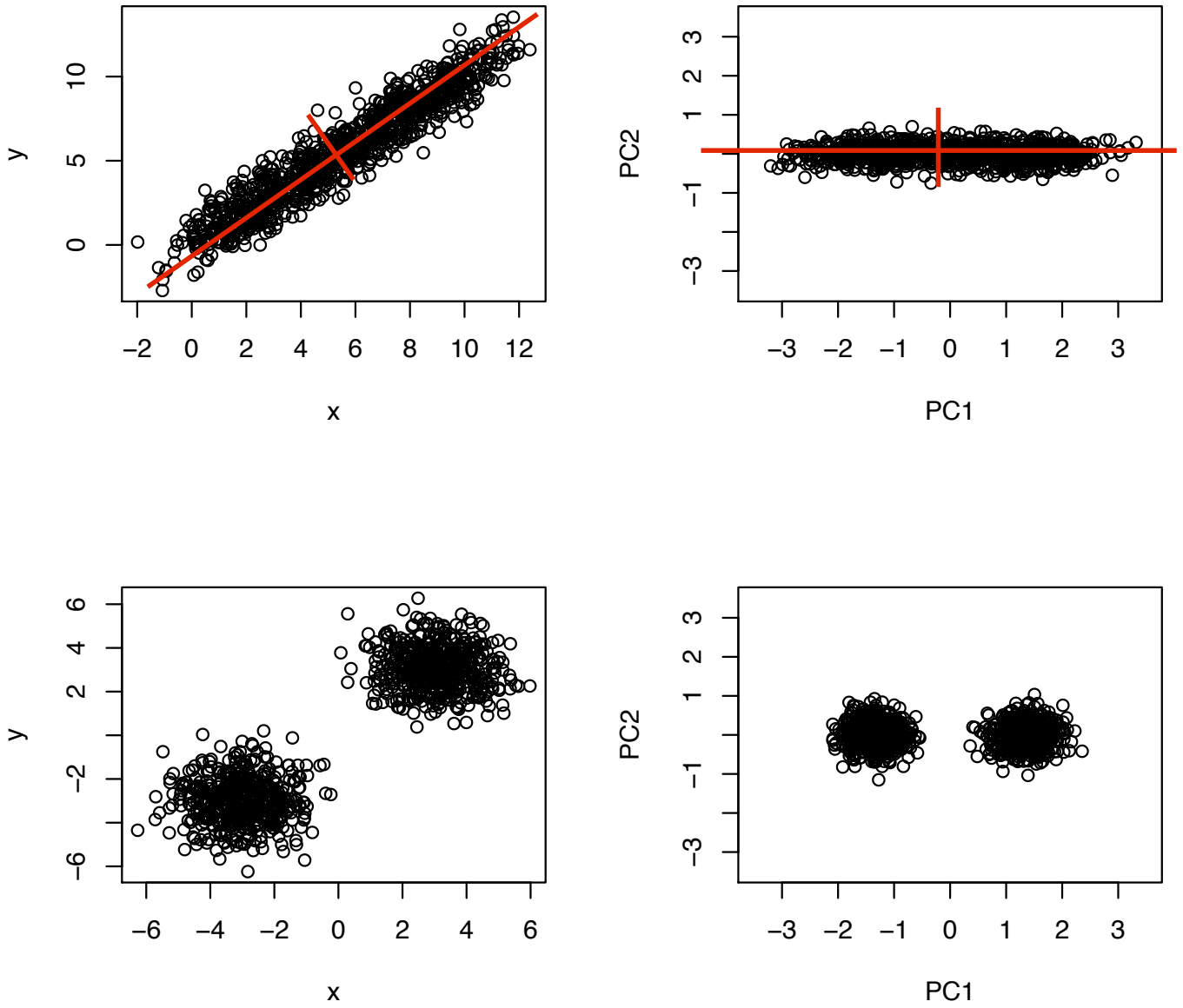


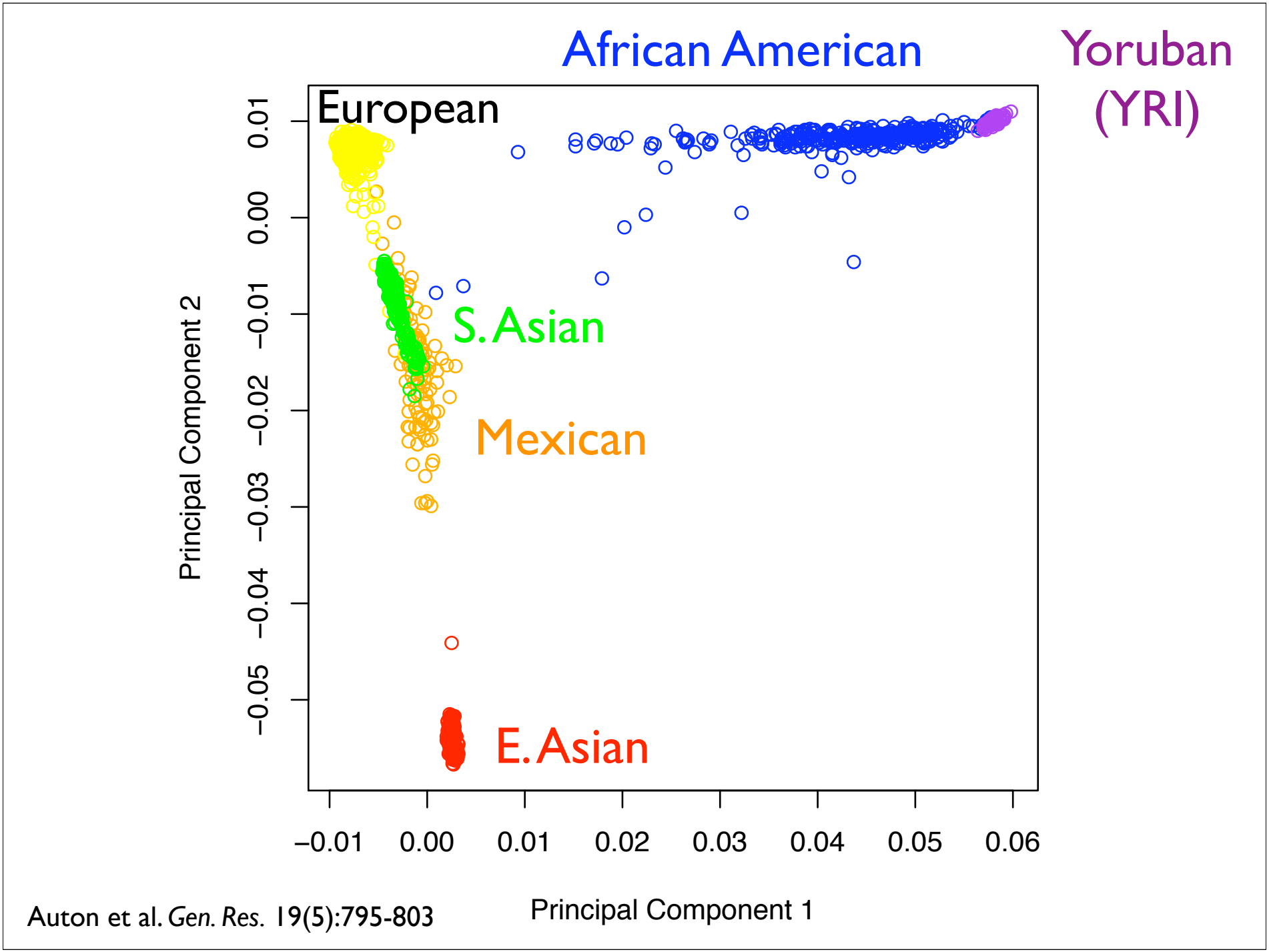




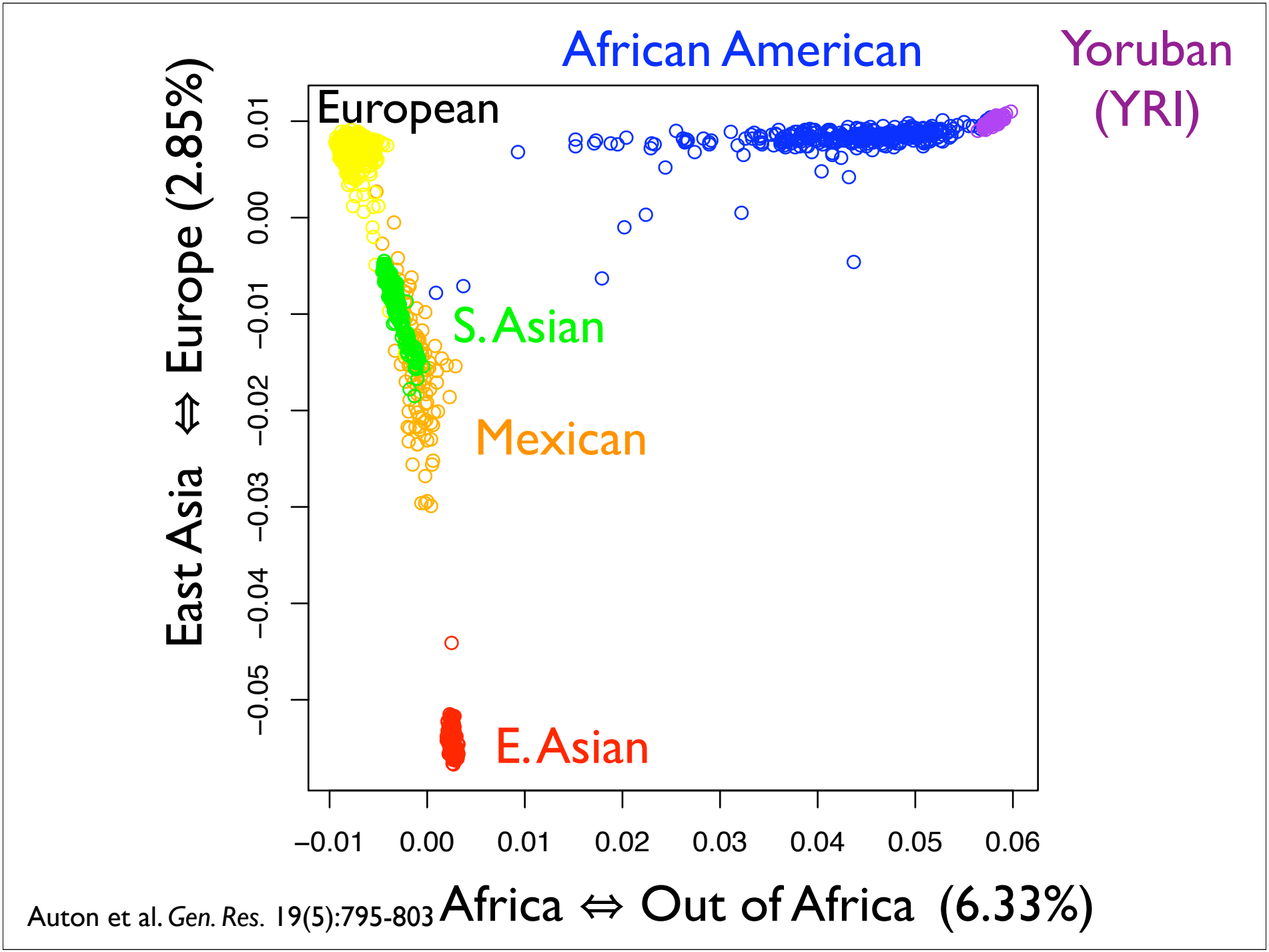


Principle Components = major axes of variation

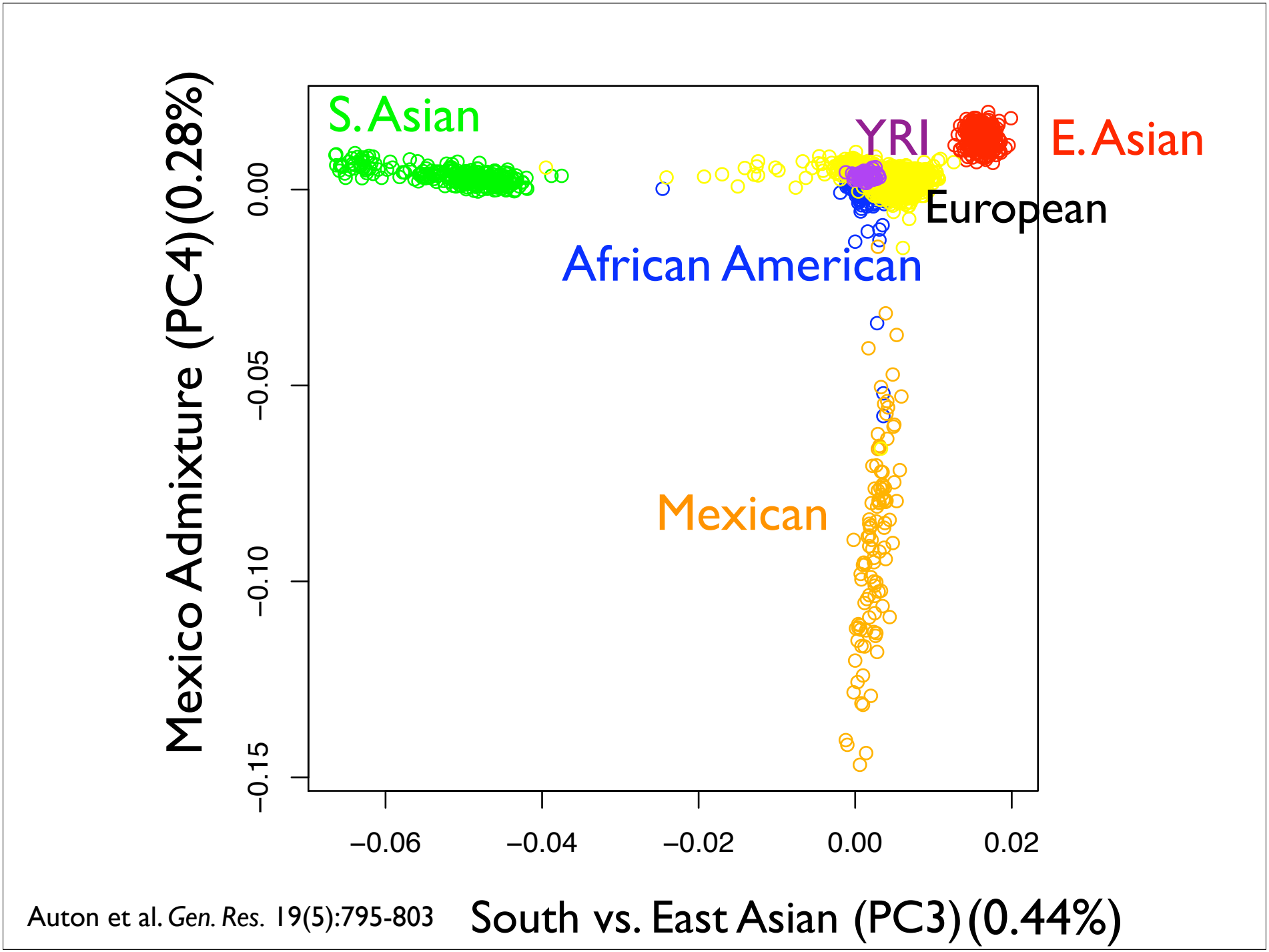




Auton et al. *Gen. Res.* 19(5):795-803

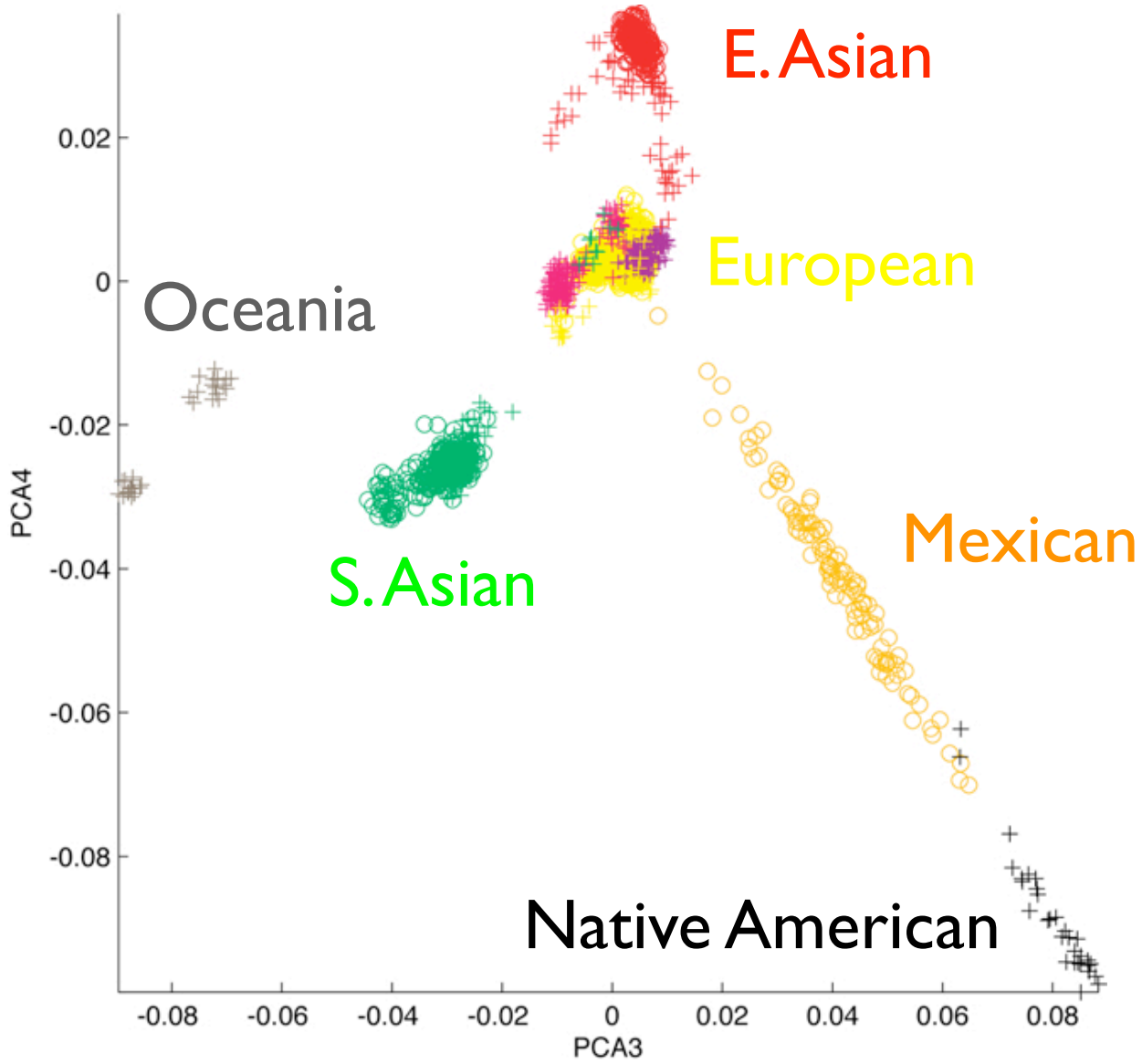


Auton et al. *Gen. Res.* 19(5):795-803



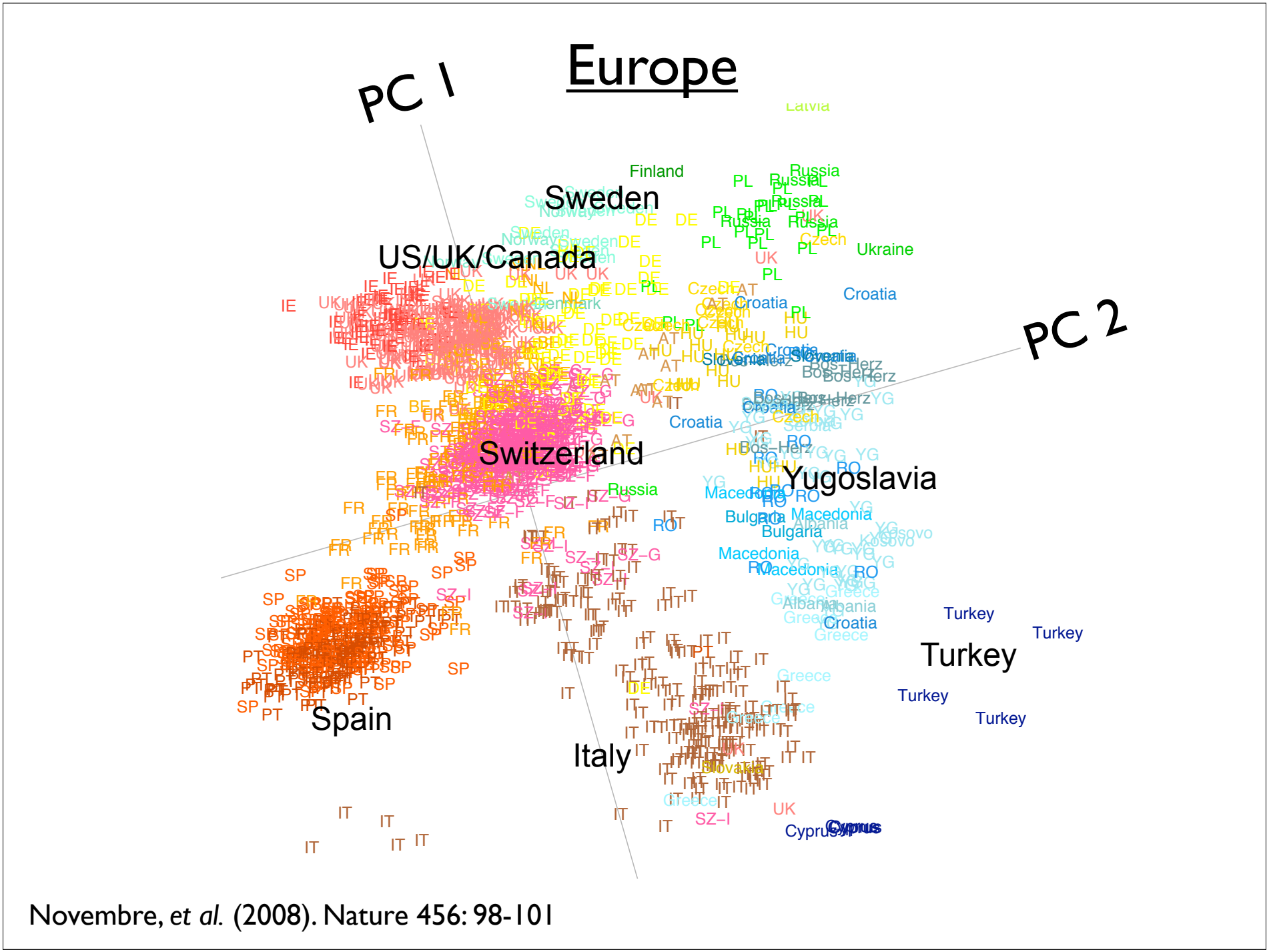
Auton et al. *Gen. Res.* 19(5):795-803

GSK + Stanford HGDP across 70,000 SNPs



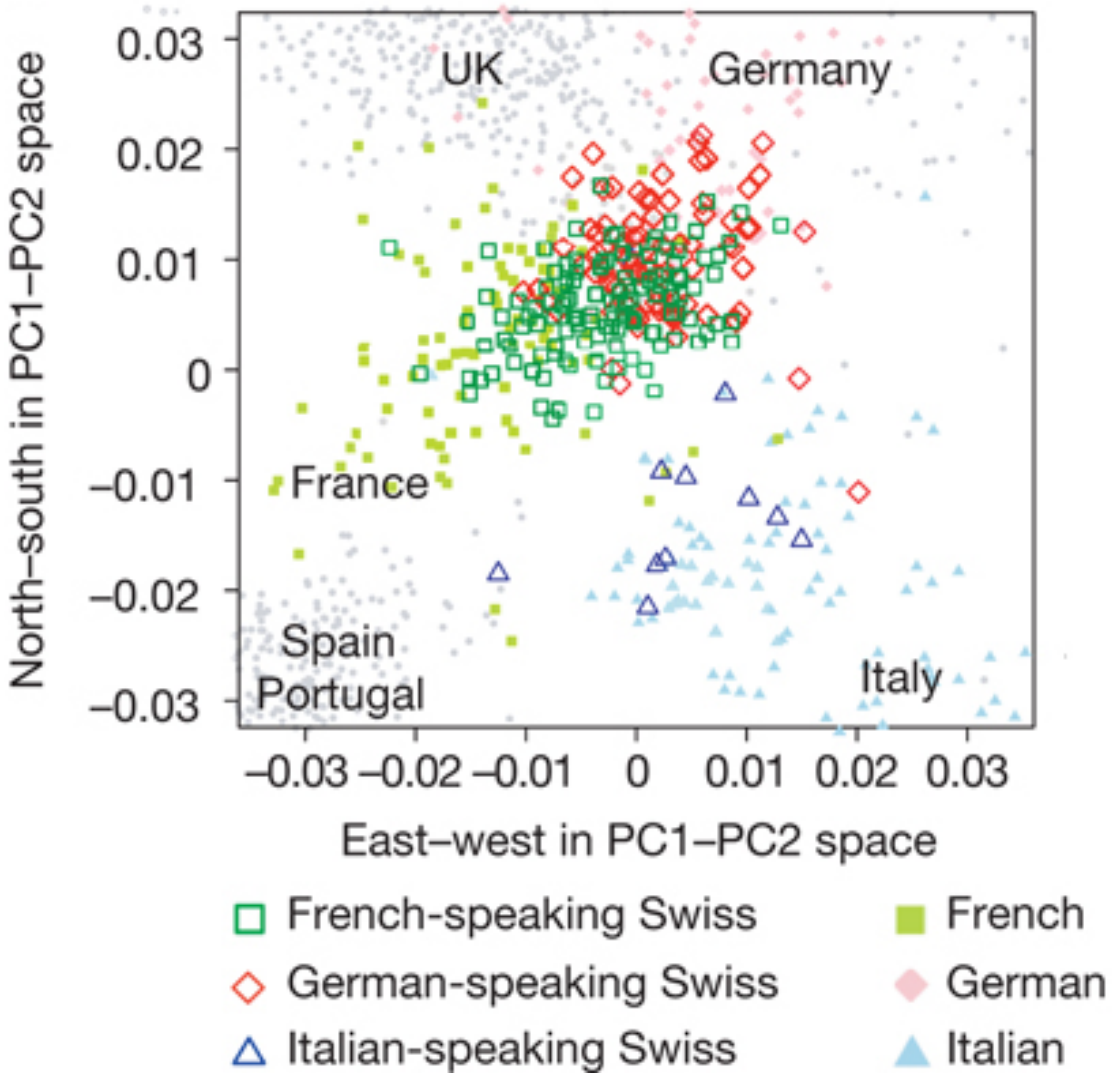
Auton et al. *Gen. Res.* 19(5):795-803

Li *et al.*, (2008) *Science* (319): 1100 - 1104



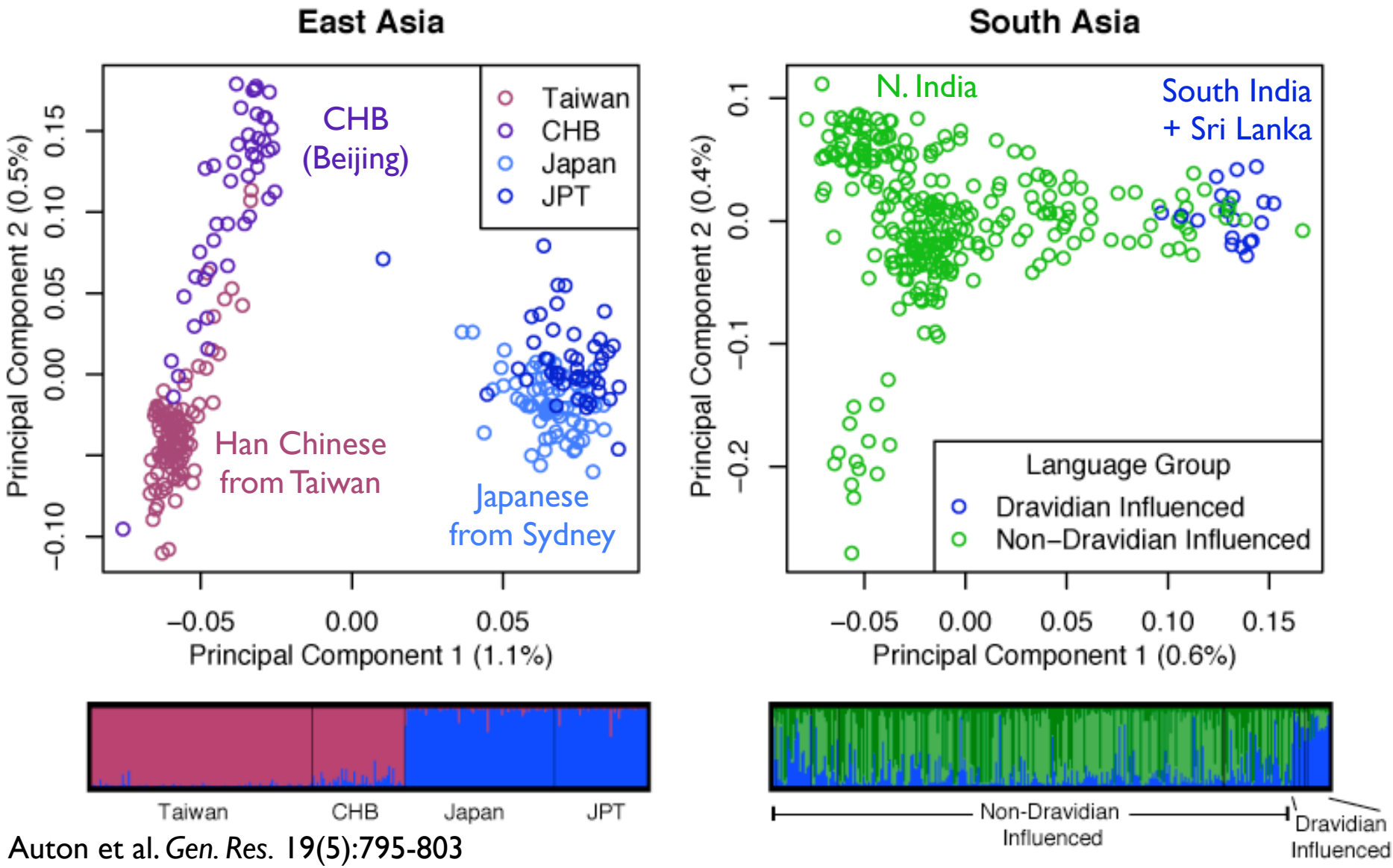
Novembre, *et al.* (2008). Nature 456: 98-101

Within Switzerland



Novembre, *et al.* (2008). Nature 456: 98-101

Continental PCA and Structure Analysis



Auton et al. *Gen. Res.* 19(5):795-803

Reconstructing Personal Genetic History?

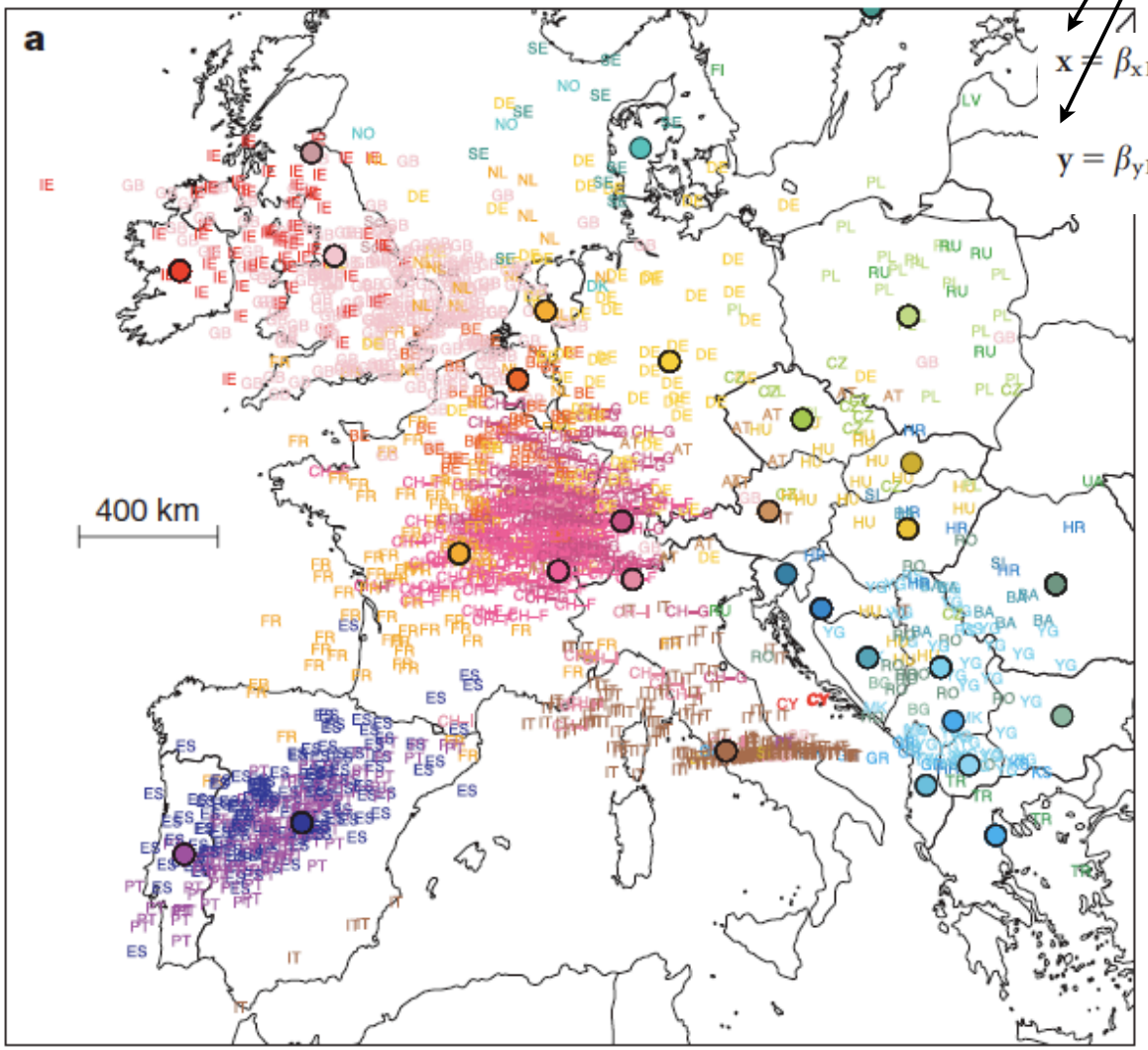
Spatial Predictions of Ancestry

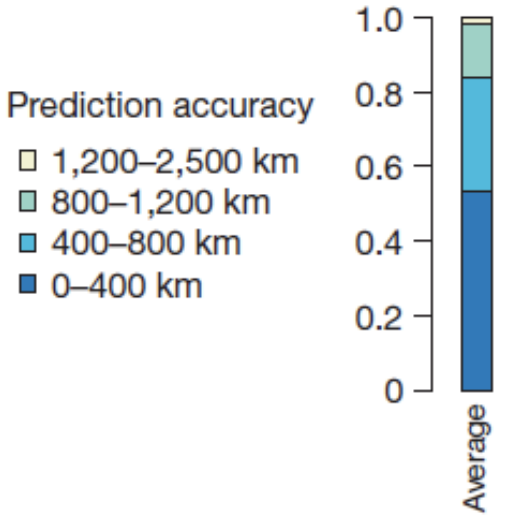
(longitude, latitude)

$$x = \beta_{x1}u_1 + \beta_{x2}u_2 + \beta_{x11}u^2{}_1 + \beta_{x22}u^2{}_2 + \beta_{x12}u_1u_2 + \varepsilon$$

$$y = \beta_{y1}u_1 + \beta_{y2}u_2 + \beta_{y11}u^2{}_1 + \beta_{y22}u^2{}_2 + \beta_{y12}u_1u_2 + \varepsilon$$

(PC1, PC2)





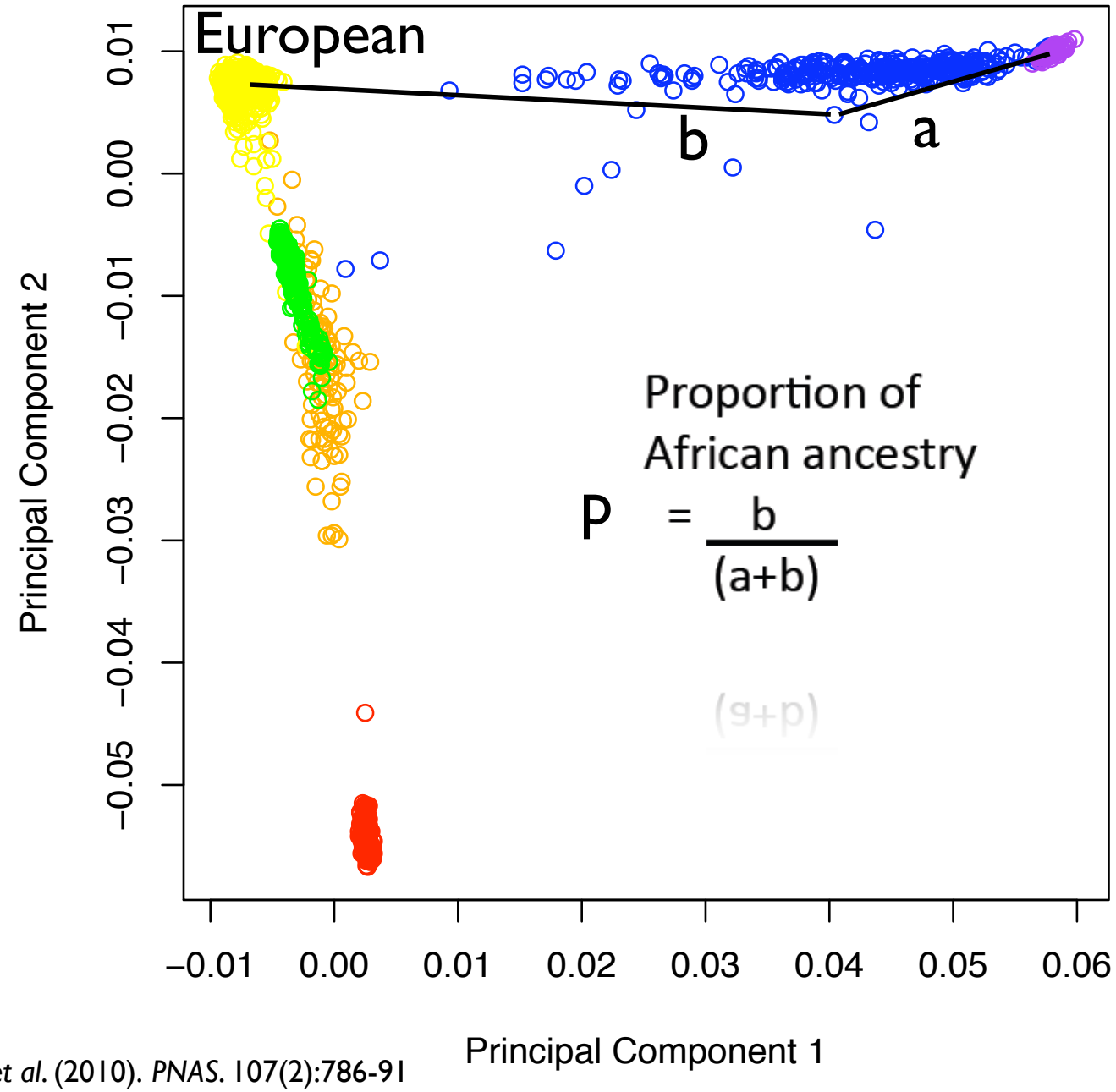
John Novembre

Matthew Stephens
(U. Chicago)

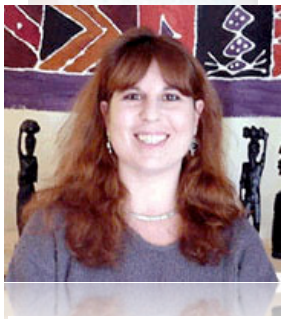
Novembre, *et al.* (2008). *Nature* 456: 98-101

African American

Yoruban (YRI)



Kasia Bryc



Dr. Sarah Tishkoff

Bryc, *et al.* (2010). *PNAS.* 107(2):786-91

Approach

- Run PCA on African Americans and diverse potential ancestral populations
- For each individual *i* at each 15-SNP window *k*, calculate

$$score_{ik} = M'_{ik} \times e_k$$

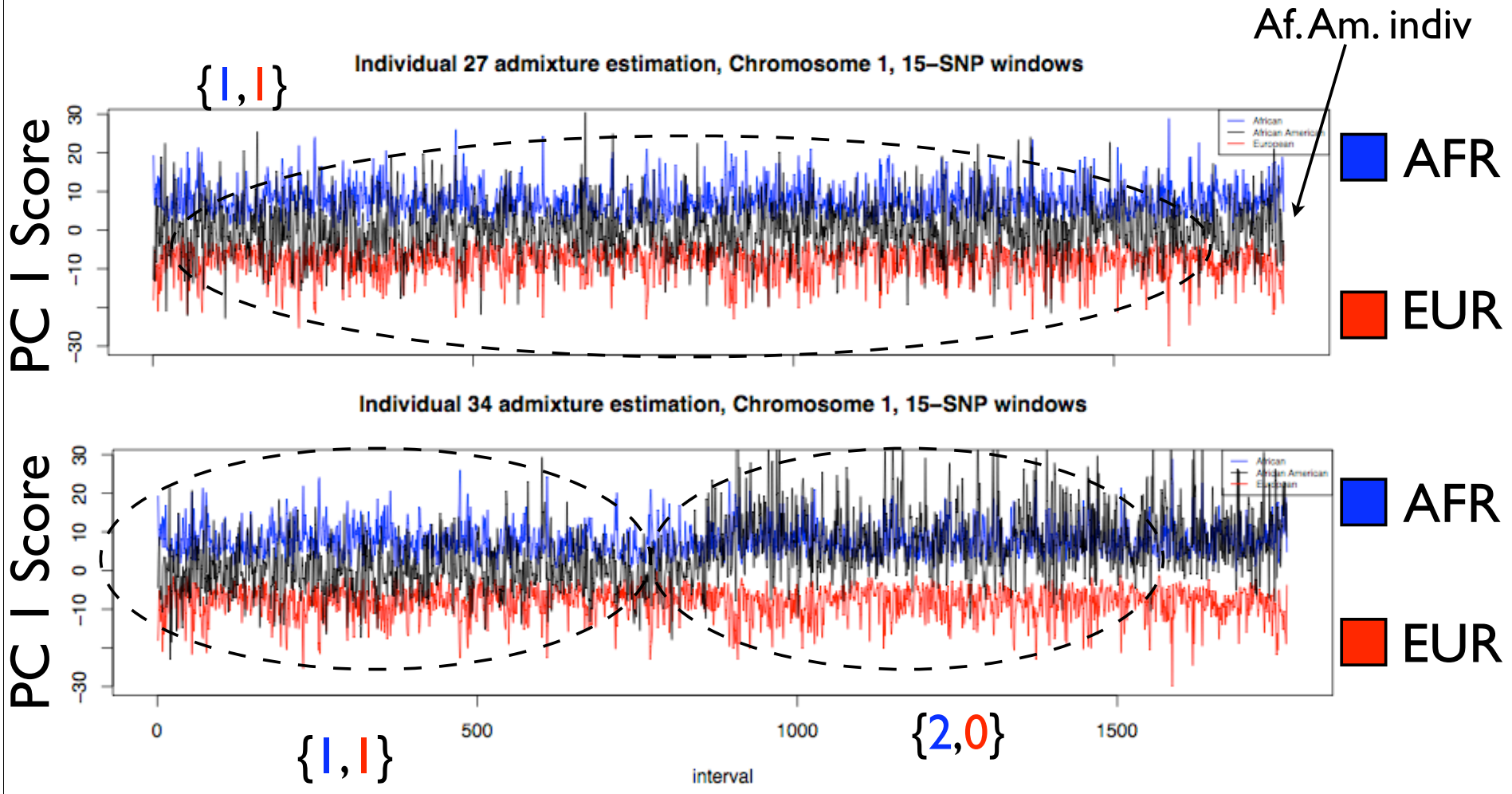
M'_{ik} are the normalized and scaled genotypes of the markers in window *k* for individual *i*

e_k is the vector of loadings corresponding to the markers in window *k*

- Overlay a Hidden Markov Model to make ancestry assignment calls

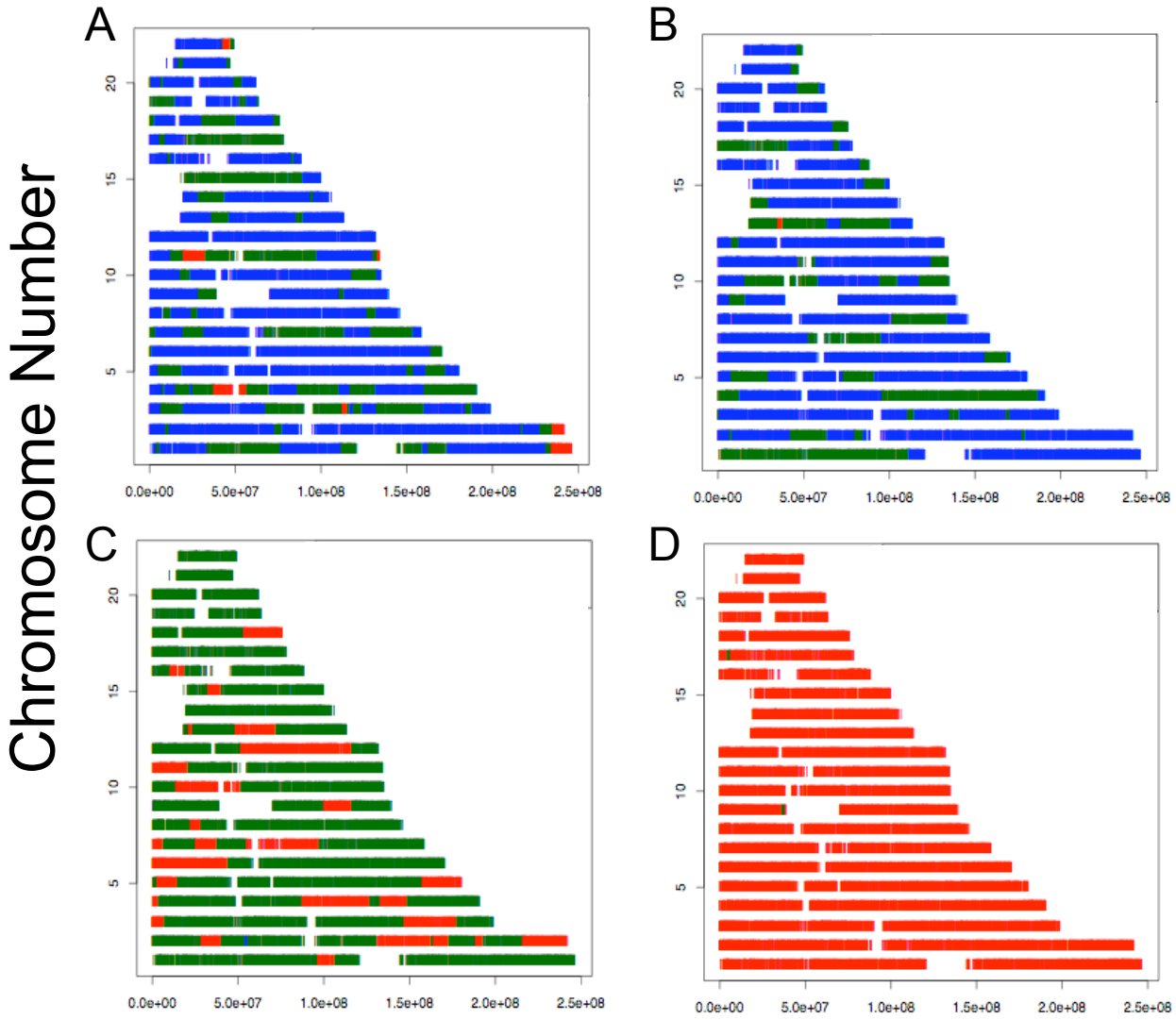
Bryc, *et al.* (2010). *PNAS*. 107(2):786-91

HMM for Admixture Estimation in African Americans



Bryc, *et al.* (2010). *PNAS.* 107(2):786-91

Individual ancestry results

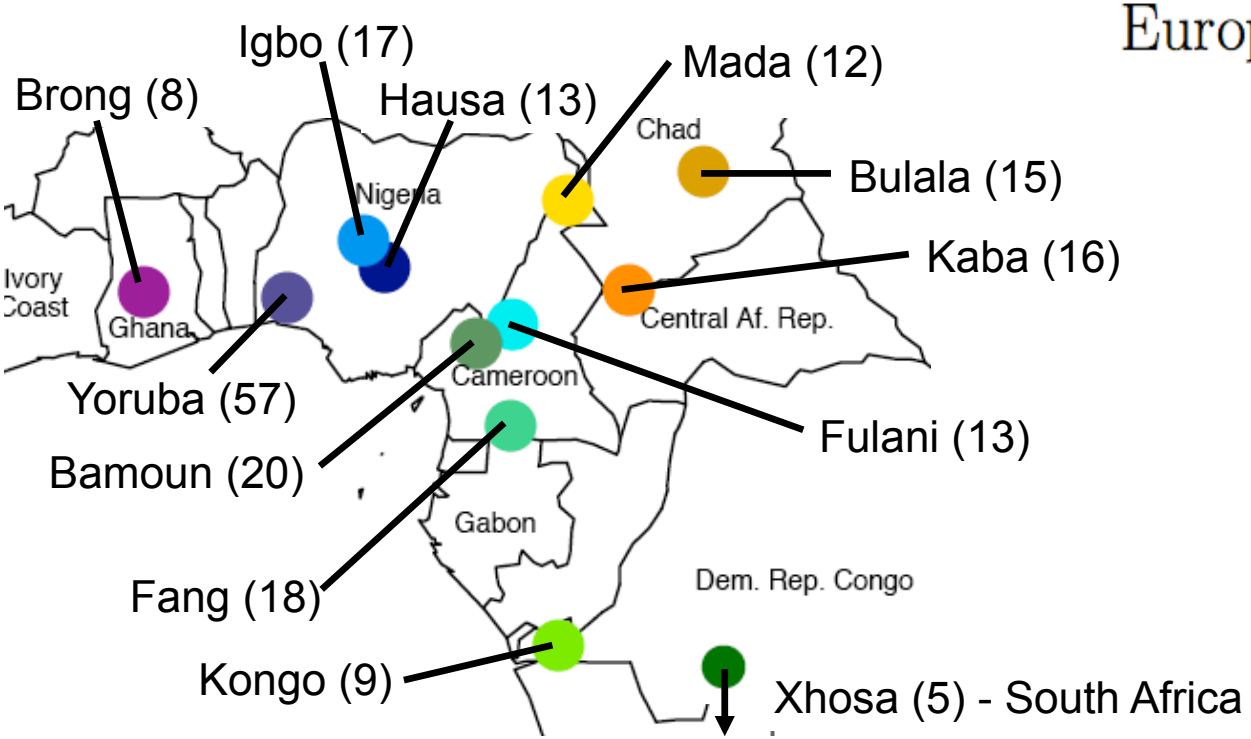


~5 minutes for all AfAm's in the data

Dataset

- Affymetrix 500K SNP arrays

Region	*n*
Africa	203
African American	365
Europe	400

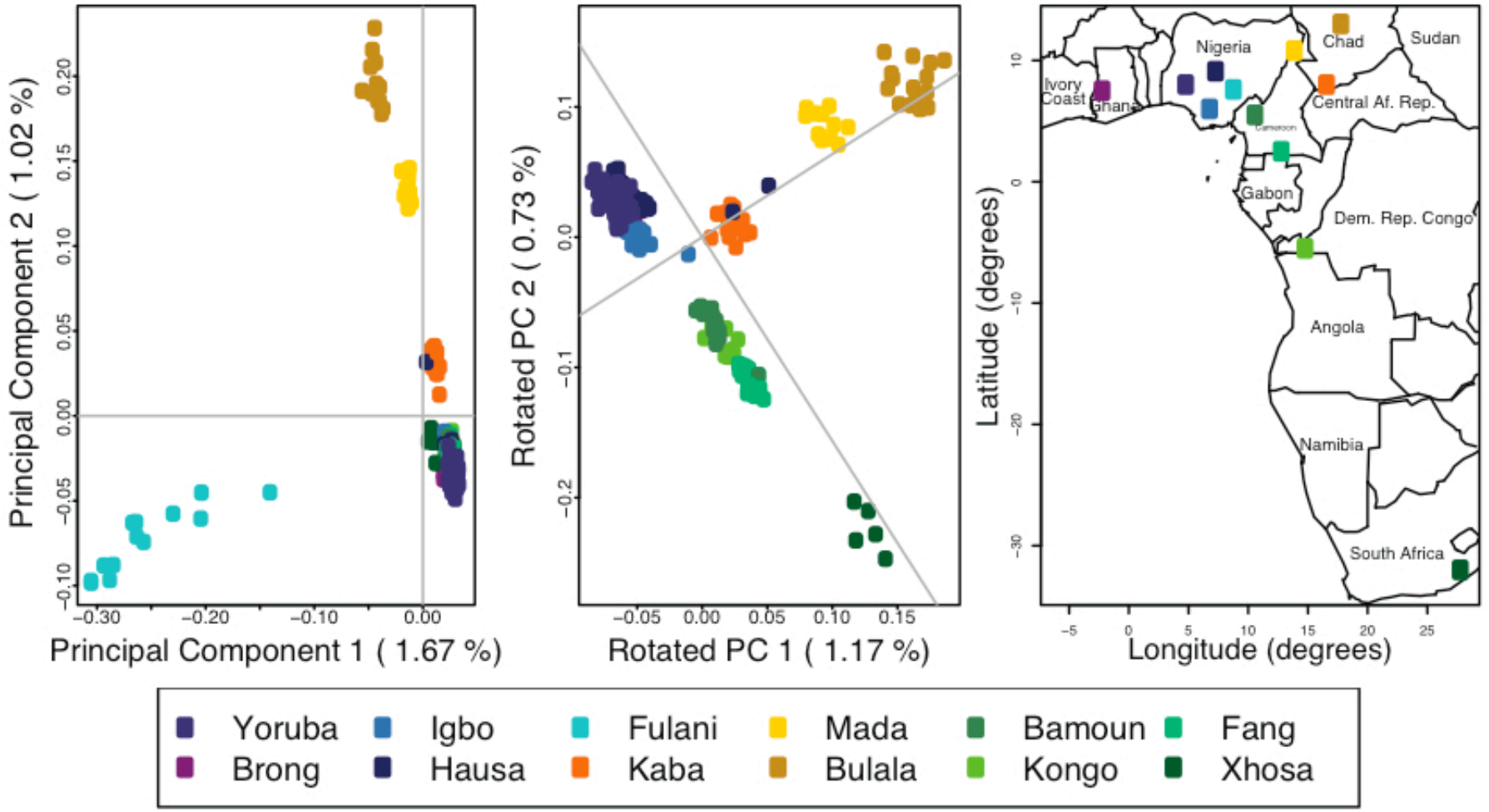


Bryc, *et al.* (2010). *PNAS.* 107(2):786-91

Matt Nelson (GSK)

Sarah Tishkoff (Penn)

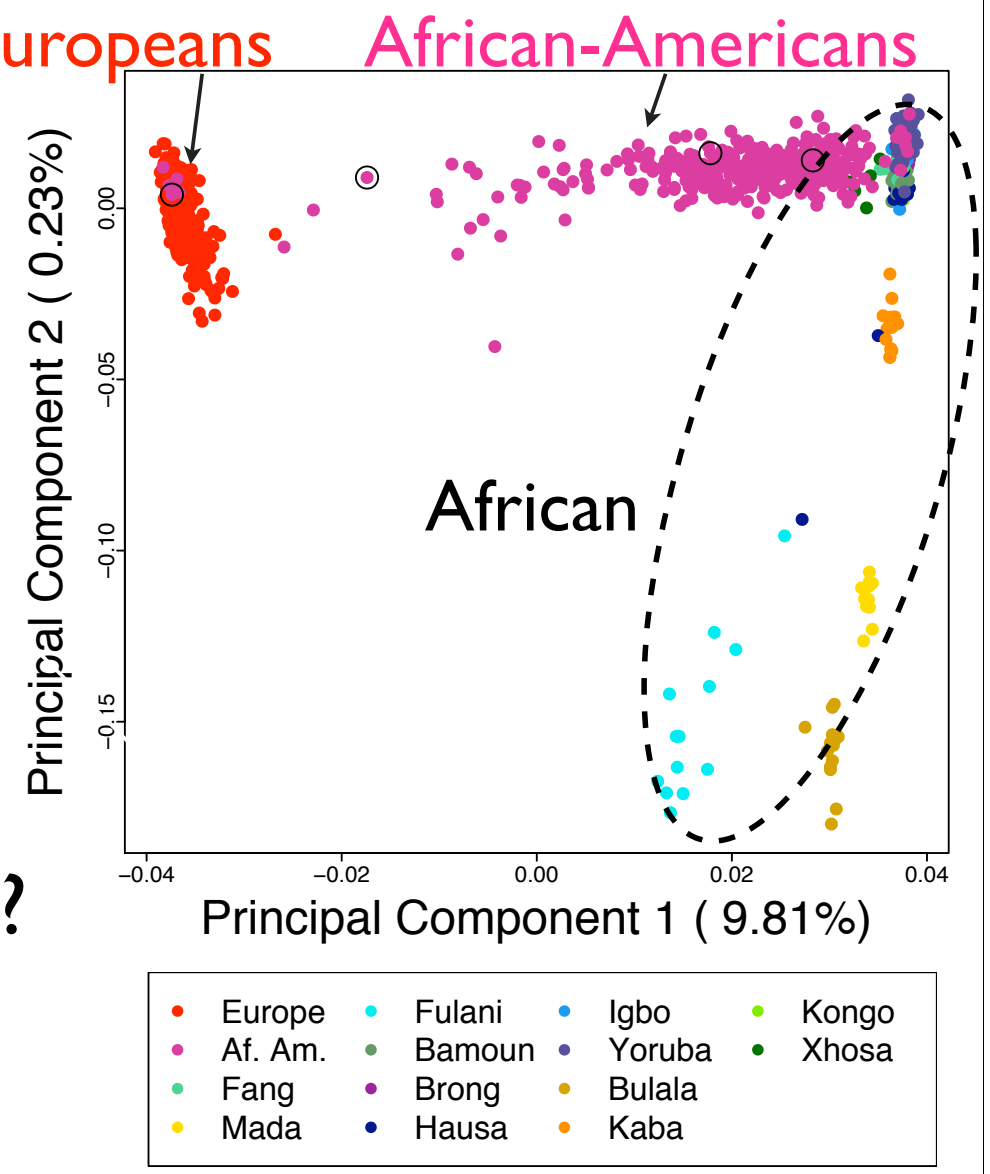
Kasia Bryc



Bryc, *et al.* (2010). *PNAS.* 107(2):786-91

Population	Language	F_{ST}
Igbo	Non-Bantu Niger-Kordofanian	0.074
Brong	Non-Bantu Niger-Kordofanian	0.077
Yoruba	Non-Bantu Niger-Kordofanian	0.089
Kongo	Bantu Niger-Kordofanian	0.112
Bamoun	Bantu Niger-Kordofanian	0.201
Xhosa	Bantu Niger-Kordofanian	0.257
Fang	Bantu Niger-Kordofanian	0.266
Hausa	Afro-Asiatic	0.325
Kaba	Nilo-Saharan	0.353
Mada	Afro-Asiatic	0.97
Bulala	Nilo-Saharan	1.581
Fulani	Non-Bantu	2.973

?



Bryc, *et al.* (2010). *PNAS.* 107(2):786-91

Placing Hispanic/Latino Genomic Diversity on the Map

Dr. Harry Ostrer

Chris Velez

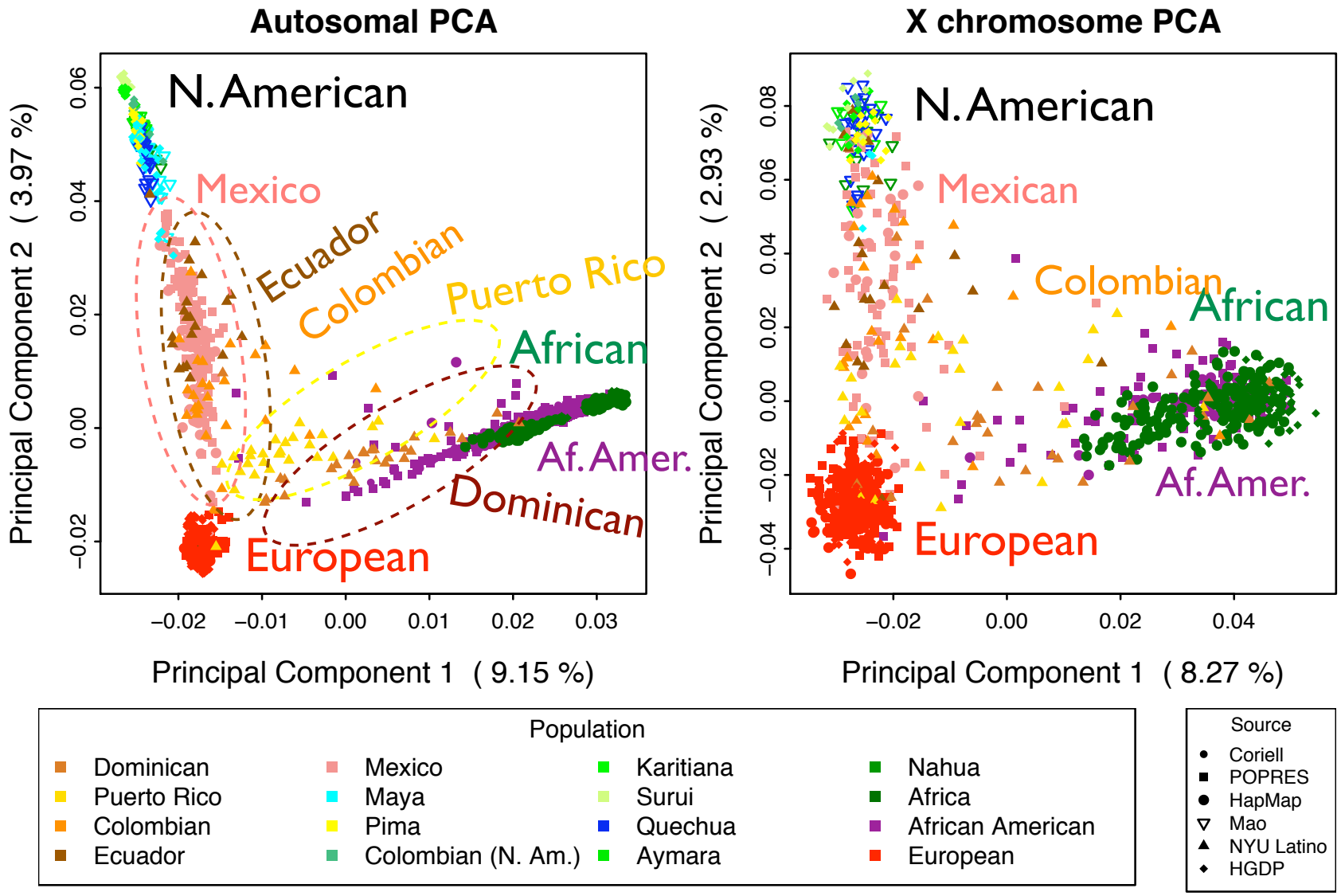
Kasia Bryc

Dr. Harry Ostrer

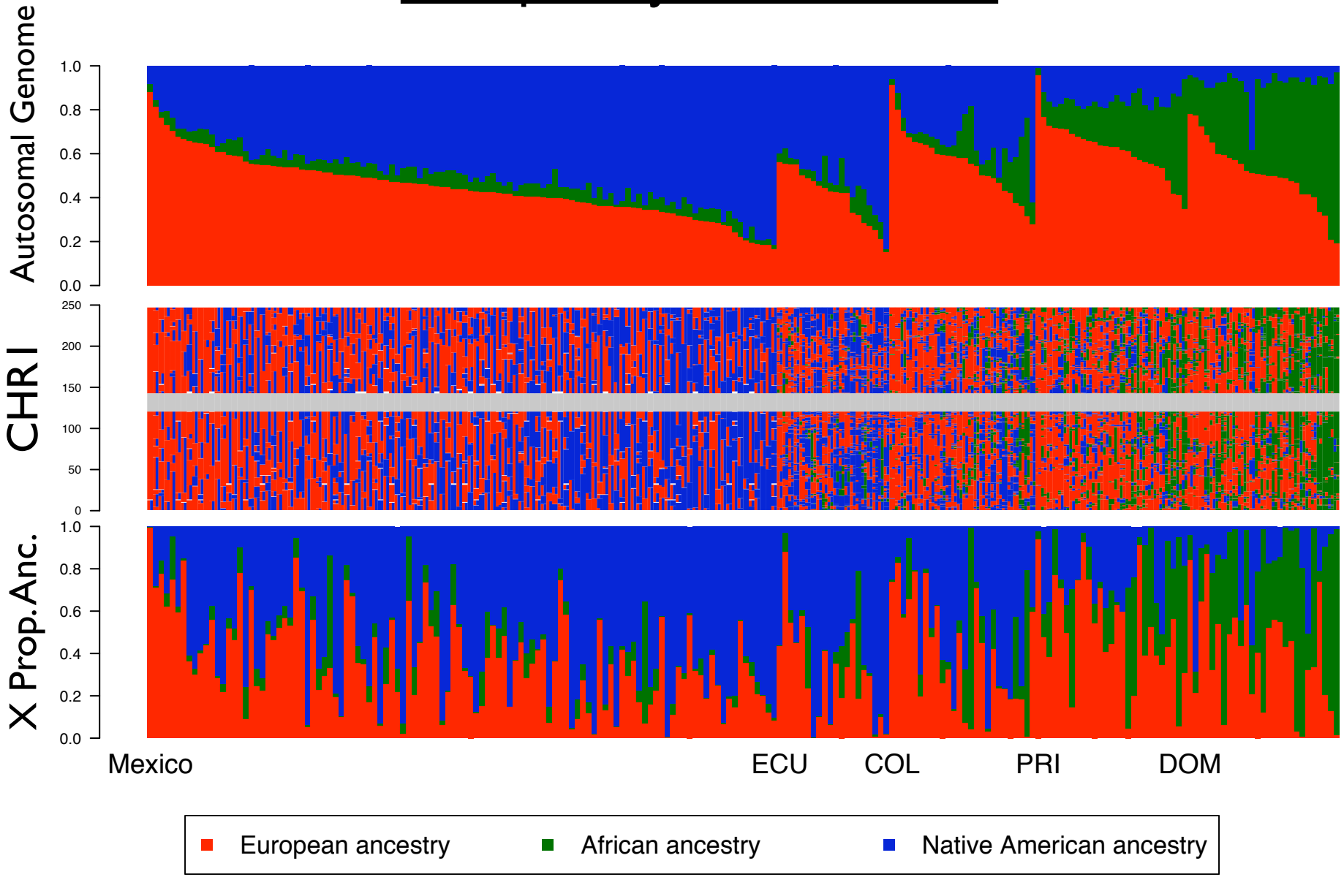
Chris Velez
Kasia Bryc

Latin American Diversity

100 HL's
Illumina 610K
(unpublished)



Complexity of Admixture



Take home message:

- Personal ancestry reconstruction (including detection of admixture tracts) is feasible on genome-wide scale
- African-Americans exhibit, on average, ~78% West African and 22% European ancestry from SNP chip data with large variation among individuals and genomic segments within individuals
- Hispanic-Latinos vary tremendously in admixture proportions from European, African, and Native American source populations
- A key question is understanding how full genome data will improve ancestry deconvolution and fine-mapping of ancestry "break points"