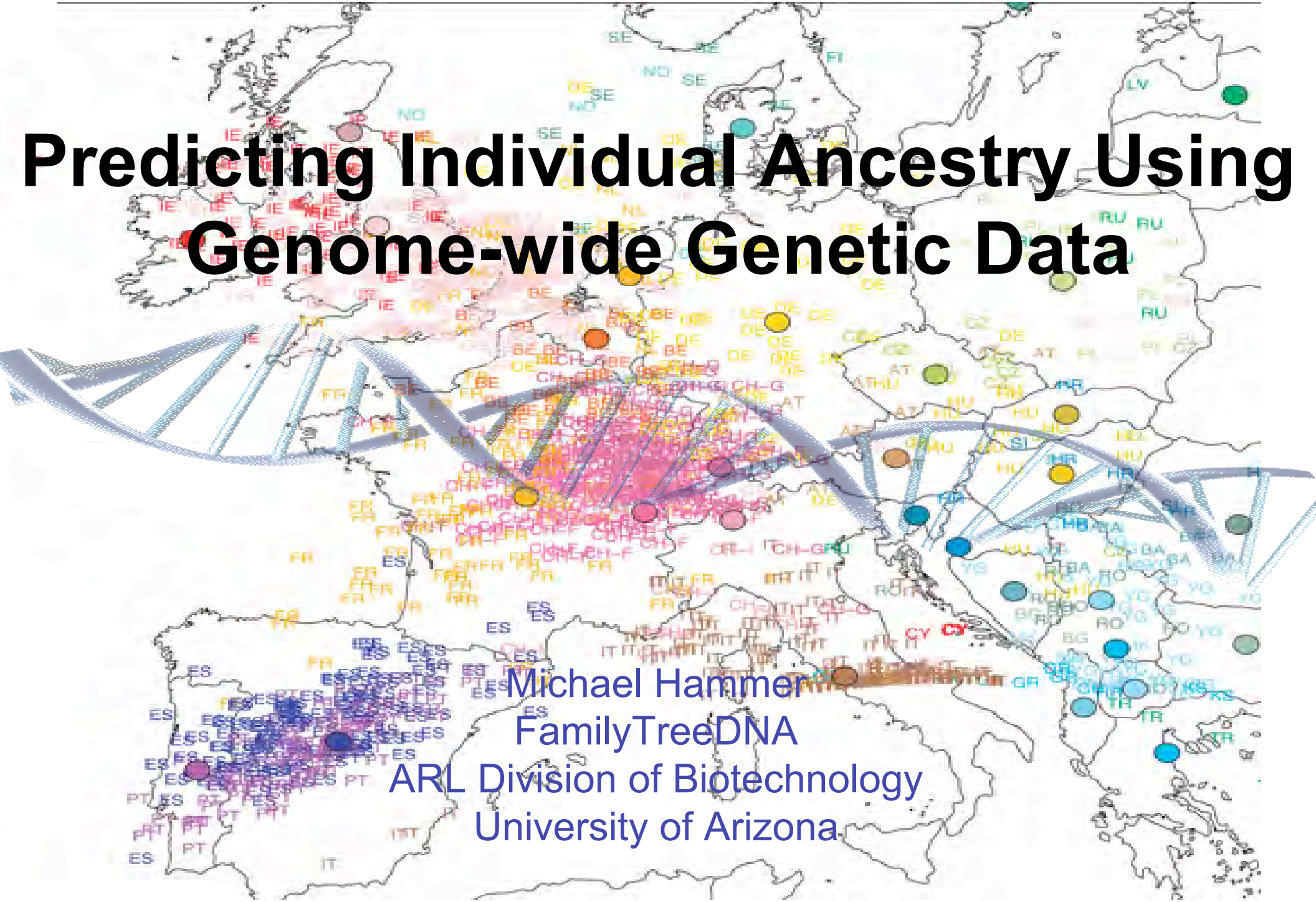
Predicting Individual Ancestry Using Genome-wide Genetic Data

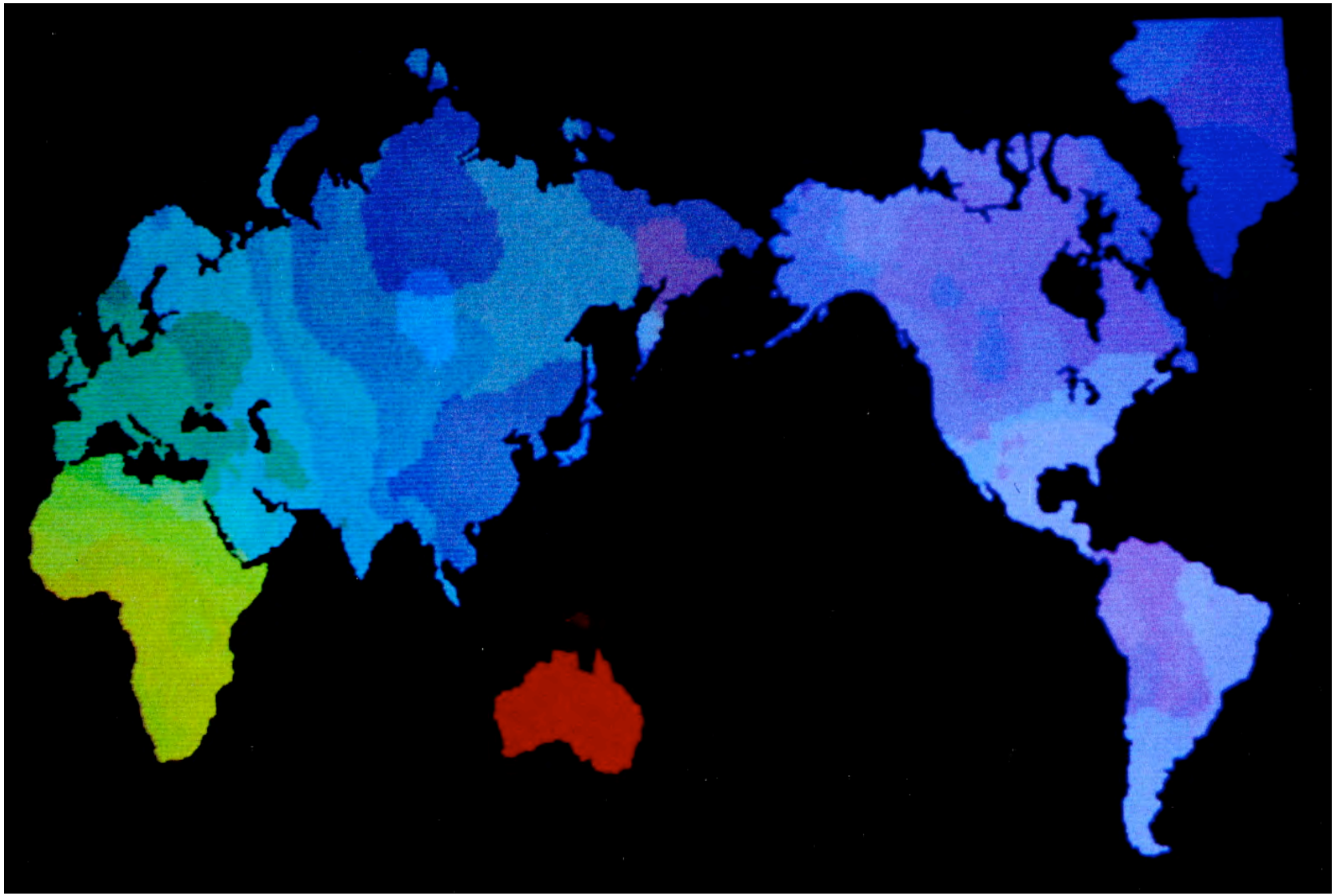


Michael Hammer
FamilyTreeDNA
ARL Division of Biotechnology
University of Arizona

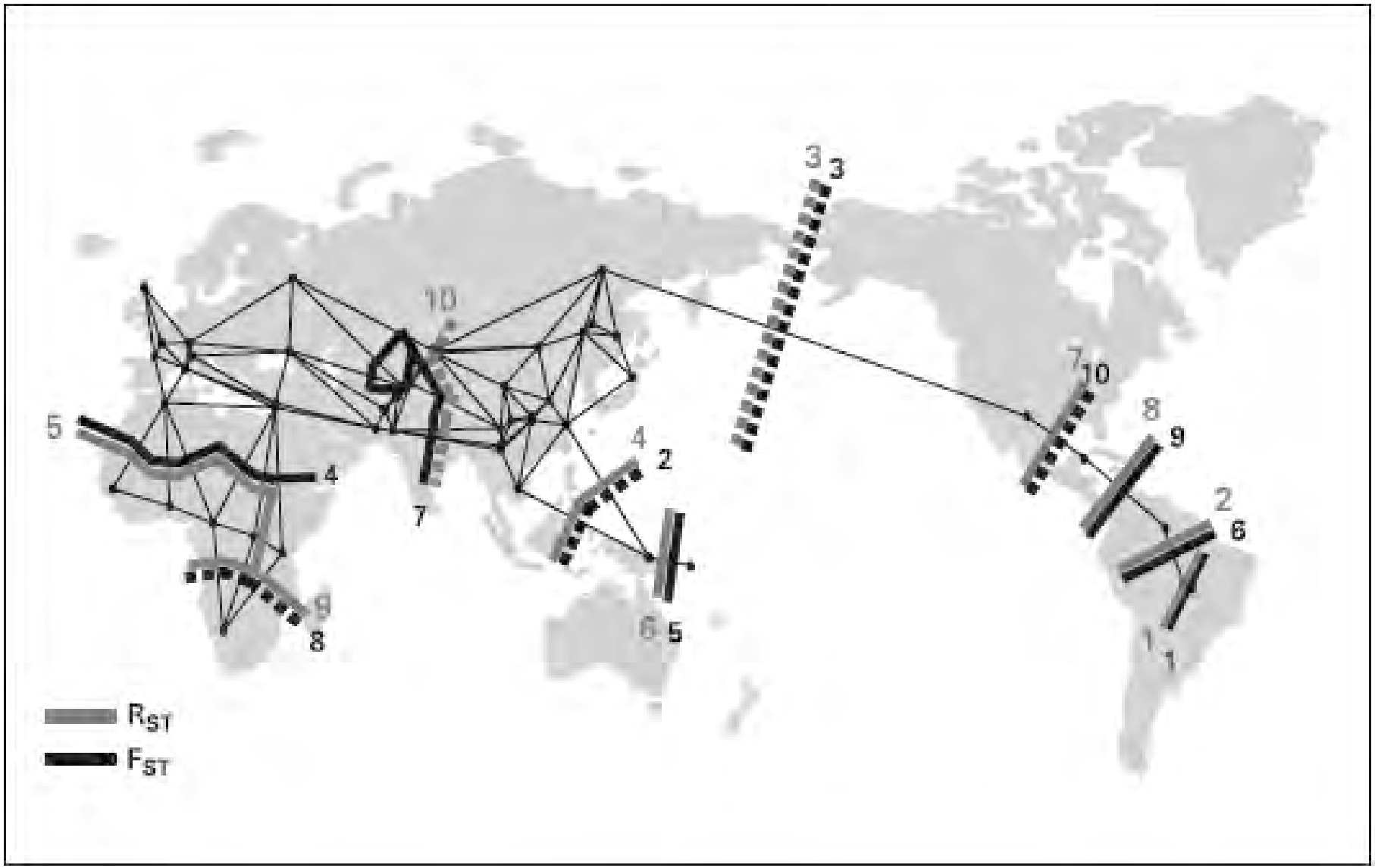
- **How is human genetic variation distributed geographically and among groups of populations?**

- **Can we use genetic data to predict the ethnic or geographic origin of a human sample?**

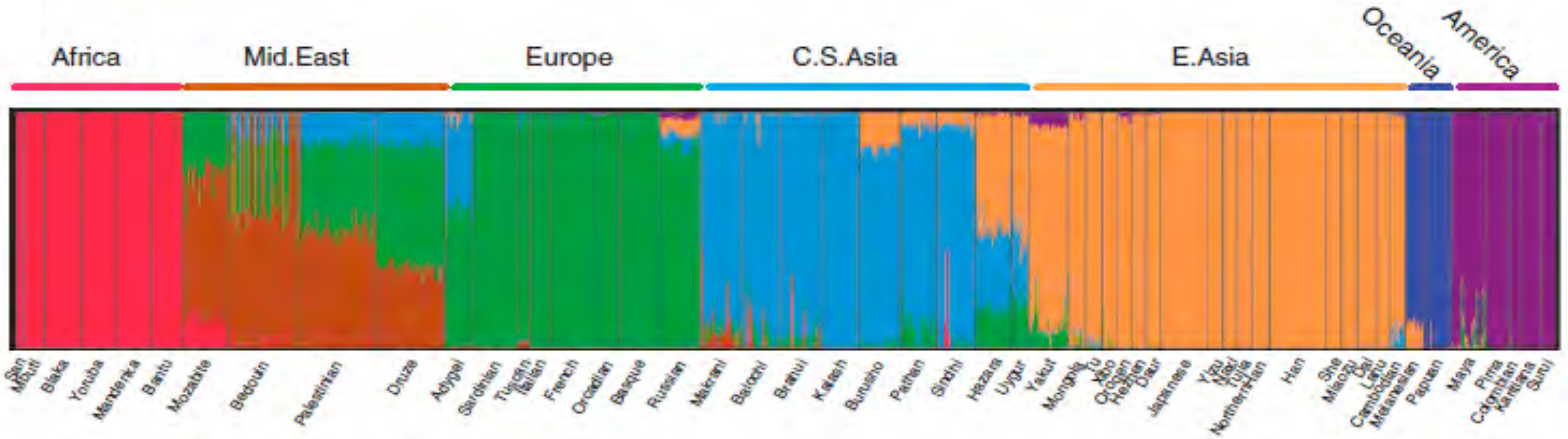
Is Human Genetic Diversity Continuous?



Is Human Genetic Diversity Discontinuous?



Genetic variation in humans is sometimes described as being **discontinuous** among continents or among groups of individuals, and by some this has been interpreted as genetic support for “races.”



Our results show that when individuals are sampled homogeneously from around the globe, the pattern seen is one of **gradients of allele frequencies** that extend over the entire world, rather than discrete clusters. Therefore, there is no reason to assume that major genetic discontinuities exist between different continents or “races.”

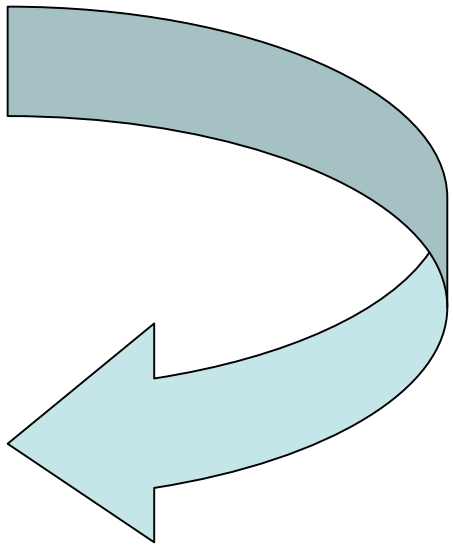
Serre & Paabo 2004

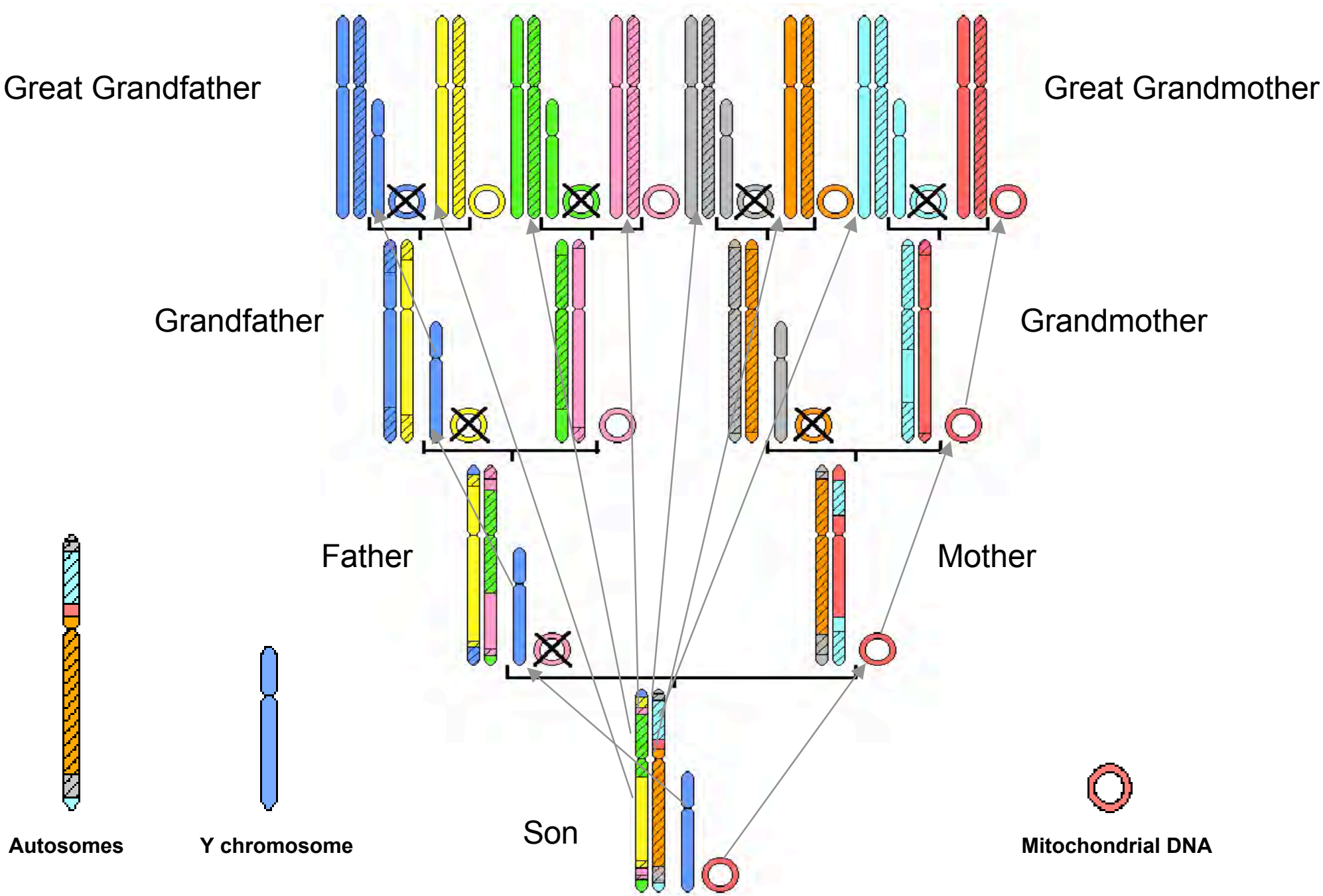
Predicting Ancestry Outline

- Y chromosome as testing ground
  - STRs predict SNP haplogroups
  - STRs/SNPs predict ancestry
- Genome-wide SNPs: Predicting ancestry at continental and regional scales

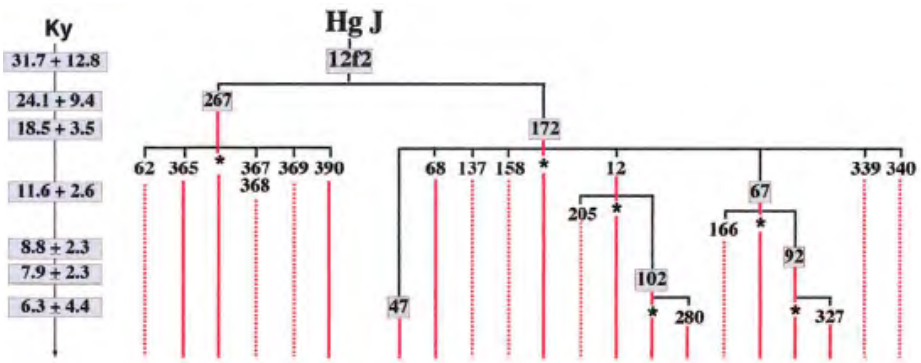
Methods

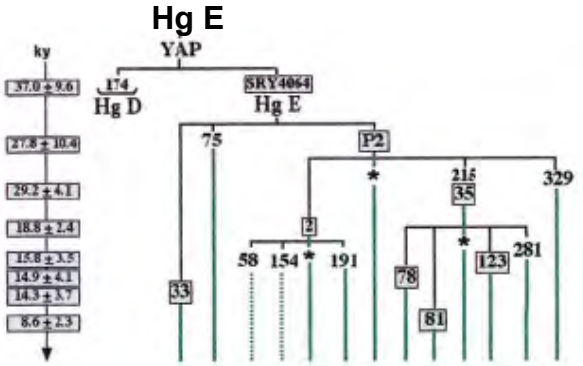
- Qualitative
- Quantitative

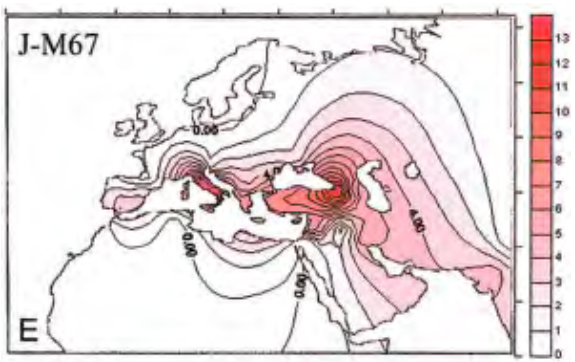


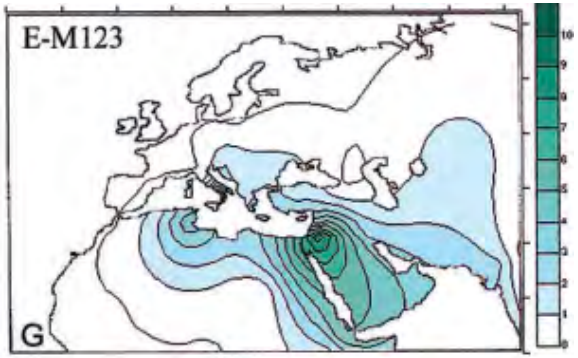

Great Grandfather
Great Grandmother
Grandfather
Grandmother
Father
Mother
Son
Autosomes
Y chromosome
Mitochondrial DNA

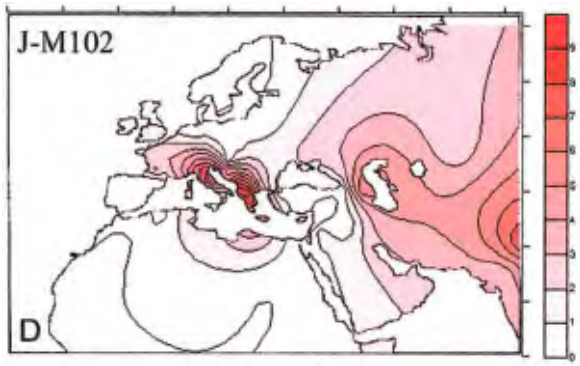
Many Y Chromosome Lineages are Geographically Localized

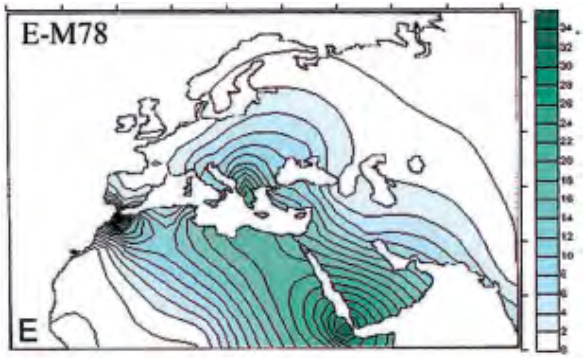






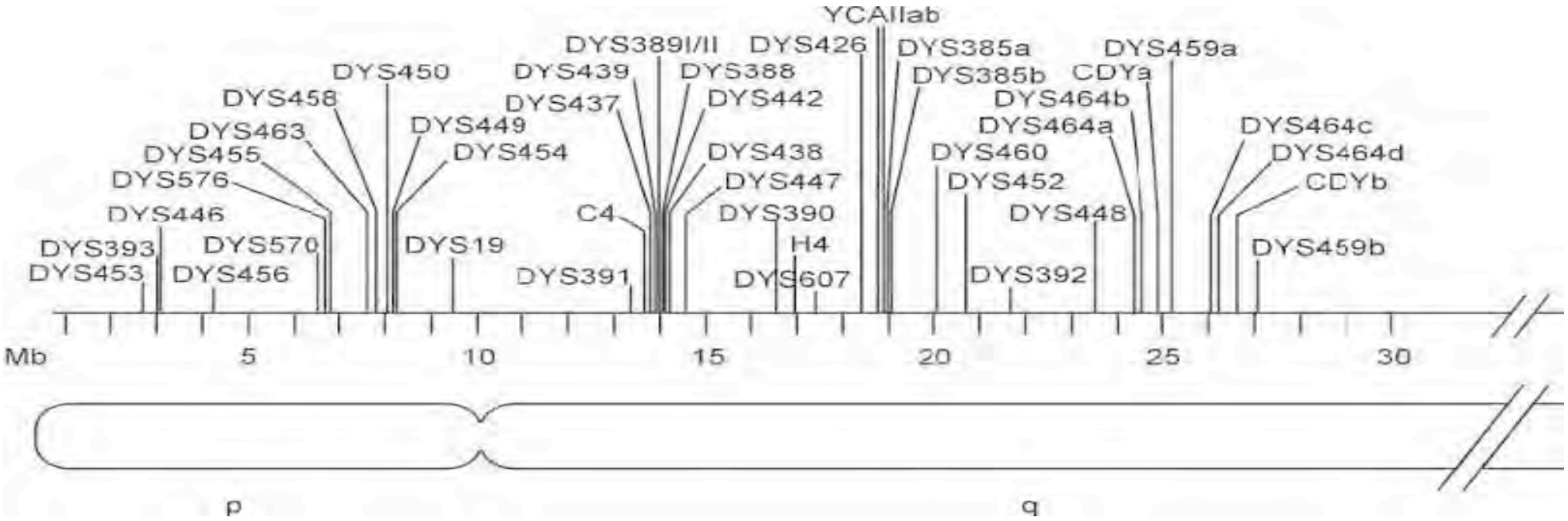






Linkage of Y-STR and Y-SNP Alleles

- All variation linked on non-recombining Y chromosome



Can We Classify Samples Using Y-STR Information?

- Predict SNP haplogroup?

- Predict other features of Y chromosome: geographic origin or ethnicity?

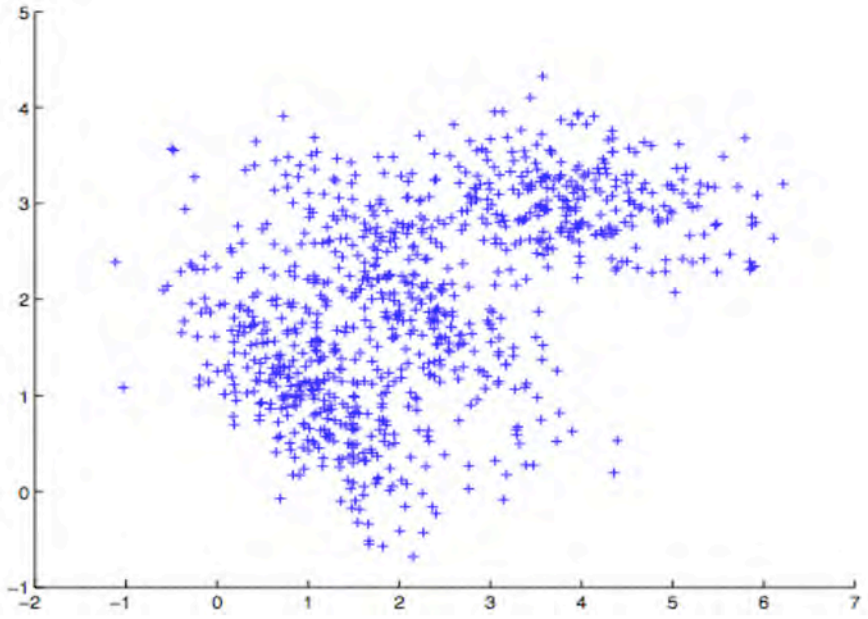
Machine Learning

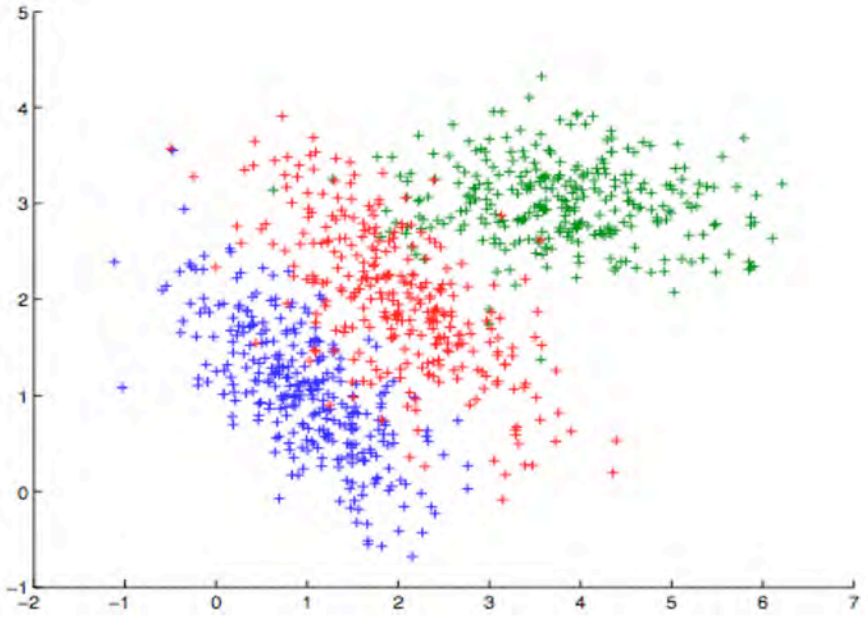
- Learn functions to map, or classify, set of Y-STR scores to a haplogroup

- Similar methods to assign haplotypes of unknown origin to populations and to predict geographic origins of unknown samples

Example: Multivariate Gaussian (MV-GMM)

- Want to define potential boundaries within cloud of data points
- A single gaussian distribution can’t capture the data, but using thee distributions allows us to capture the data better
- Use parametric method to sum across multiple peaks





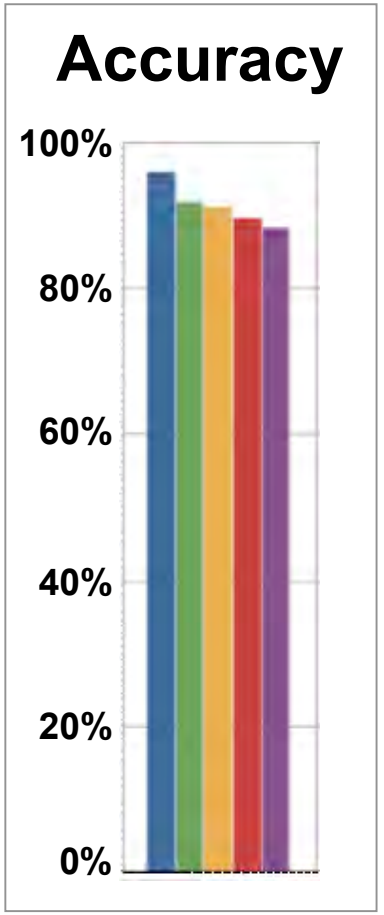
Predicting Y Haplogroup from Y-STRs

Train on truth set (majority of data)

Haplogroup
J-M267
R-M269
O-M95
A-M13
C-M217
...

→ Machine Learning $h = f(x_1, \ldots, x_n)$ ←

DYS19	DYS385a	DYS385b	DYS388	DYS389-I
14	13	19	17	13
14	11	14	12	13
15	15	20	12	13
15	12	12	11	13
17	12	12	12	13
...	...	...	...	...



Determine accuracy (subset of data)

SVM	MV-GMM	NB-Freq	PART	J48	Tandem
96.2	92.1	91.5	89.8	88.6	99.0

Predict unknown haplogroup

J-M67 ← $f(\cdot)$ ←

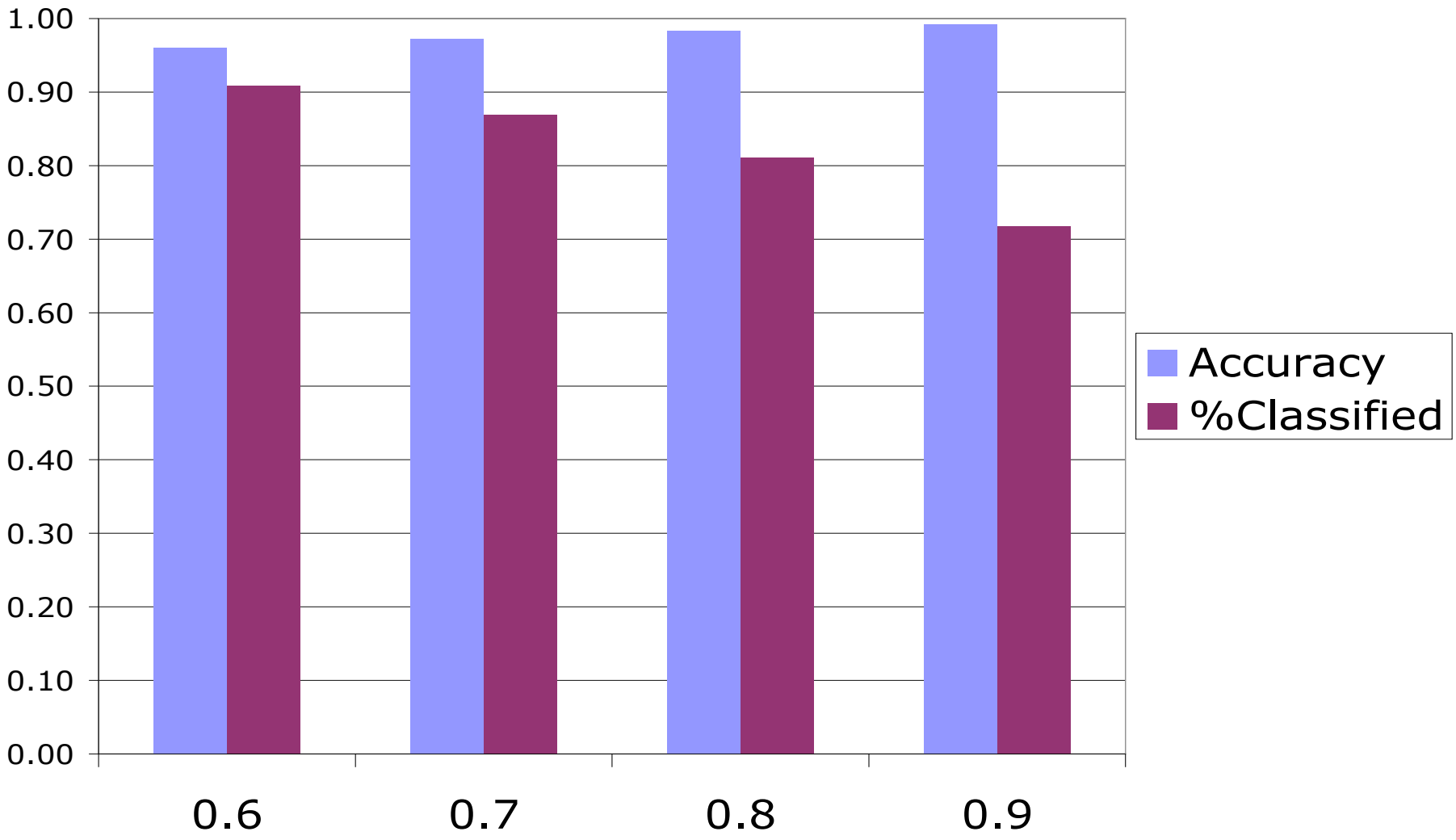
14	14	17	16	14	30	23	10

Machine Learning

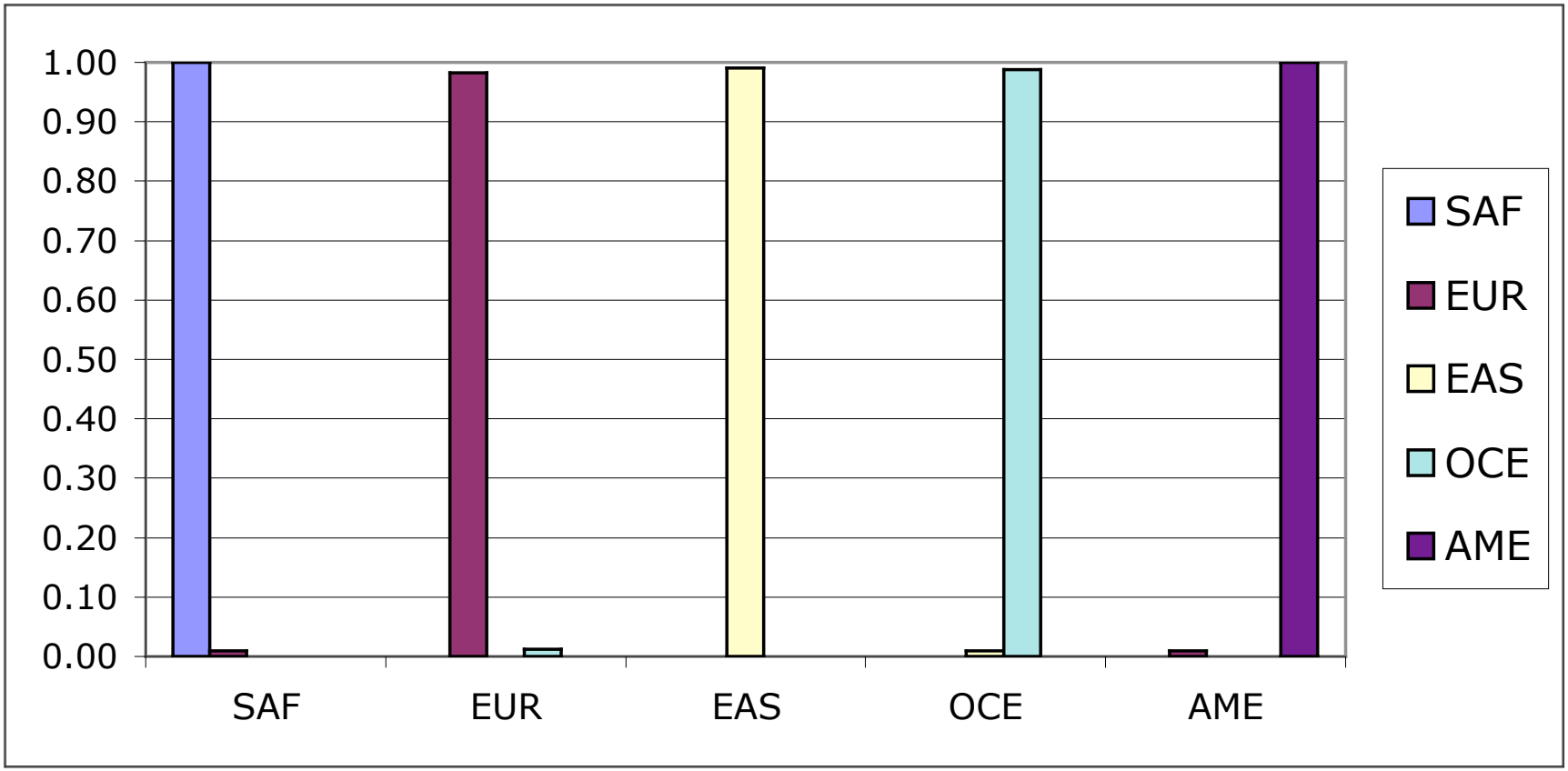
- Learn functions to map, or classify, set of Y-STR scores to a haplogroup

- Similar methods to assign haplotypes of unknown origin to populations and to predict geographic origins of unknown samples

Accuracy Y Chromosome Classification: Continental Leve



Predicting Continent of Origin from Y-STRs



91% samples classified @ 0.6 or higher (accuracy = 96.0%) *Classifier = K**

72% samples classified @ 0.9 or higher (accuracy = 99.3%)

Predicting the Extremes

- Continental groups represent periphery

- How about the interiors: populations within continents?

"….there is a great tendency in the literature to use a few populations from the extremes of continental landmasses to make worldwide inferences about substructures in the human gene pool. In fact, because human genetic diversity tends to be distributed clinally, it is especially problematic to sample the extremes of continents because this will create the impression of sharp discontinuities in the distribution of genetic variants."

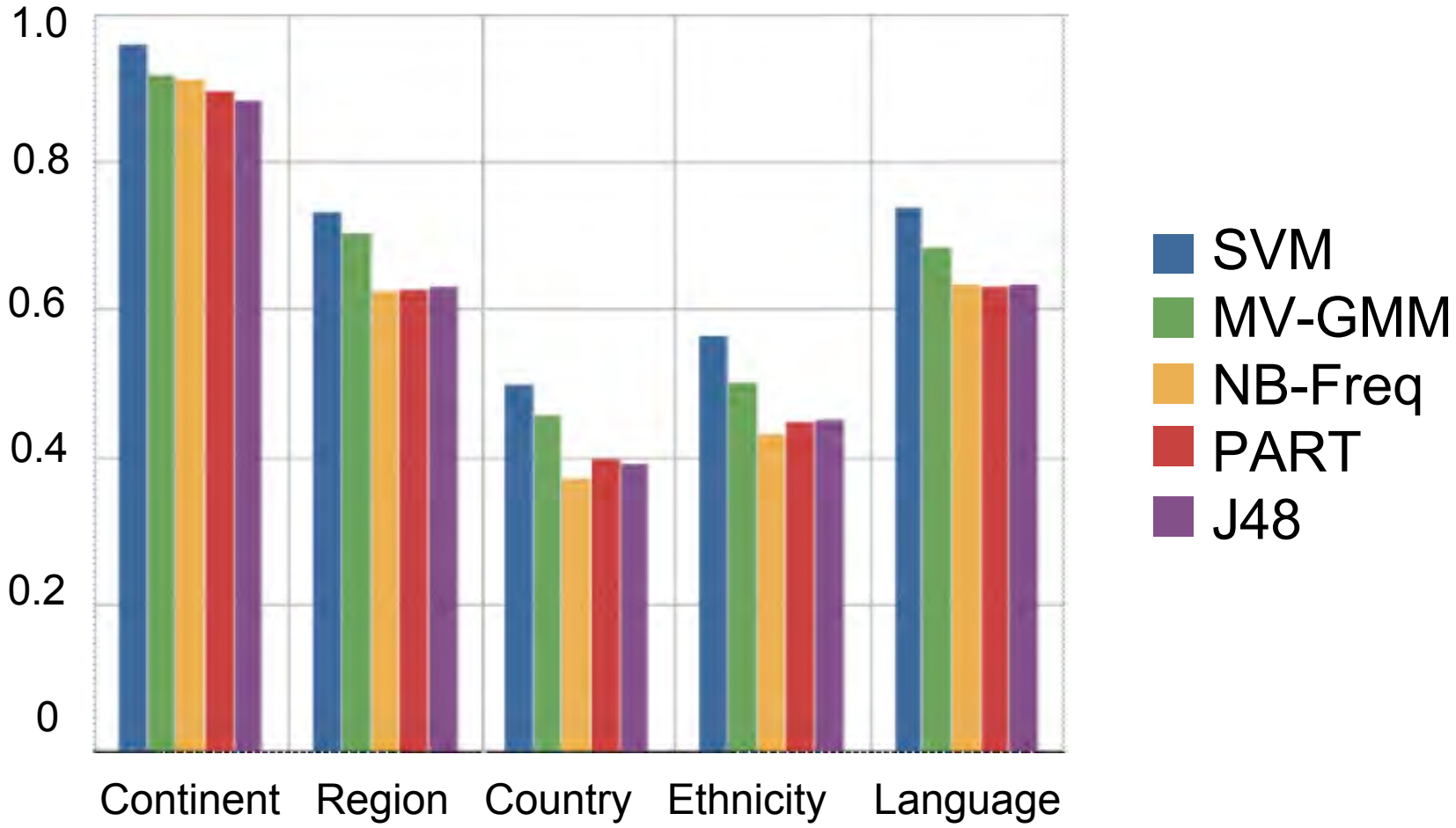
Serre and Paabo 2004

Regional Population Database

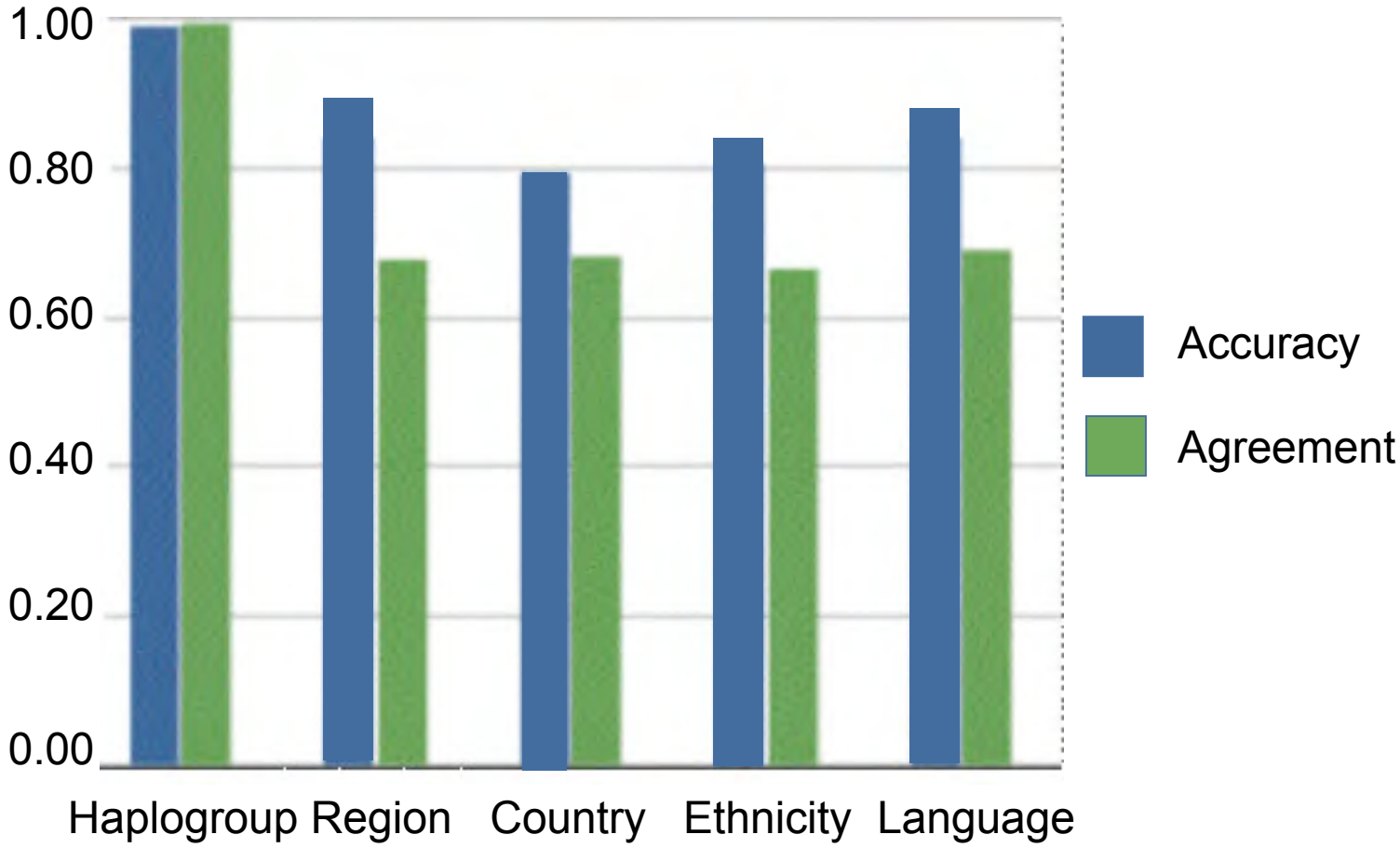
Region	N Samples	N Countries	N Ethn	N Lang. Fam.
North-East Africa (NAF)	342	5	5	5
Middle East (MEA)	670	9	14	3
Caucasus (CAU)	69	1	8	2
Central Asia (CAS)	280	7	10	4
South Asia (SAS)	496	3	12	3
Southeast Asia (SEAS)	238	2	4	1

Individual Classifier Performance

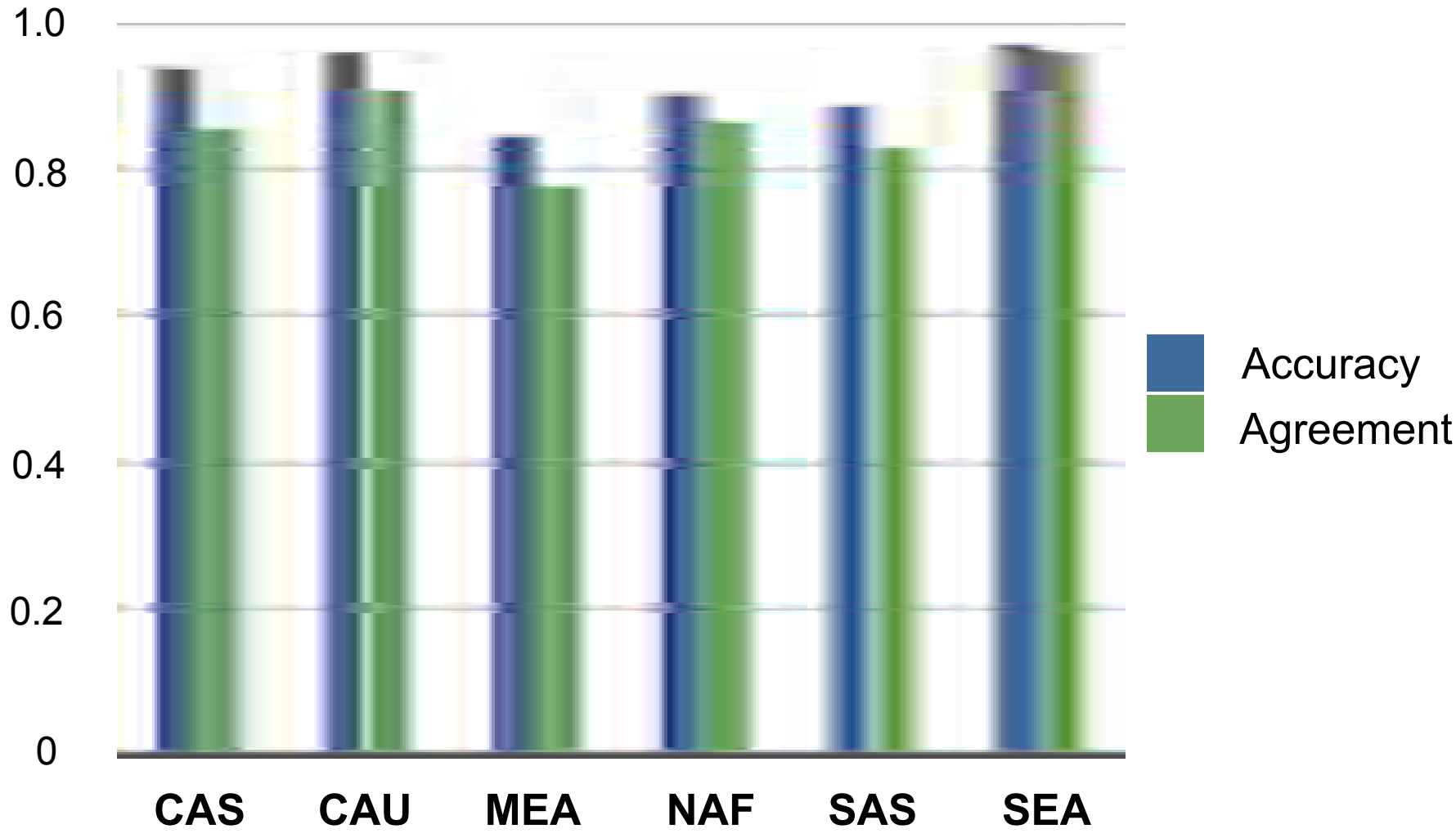
- Each algorithm has a different method of learning and accuracy



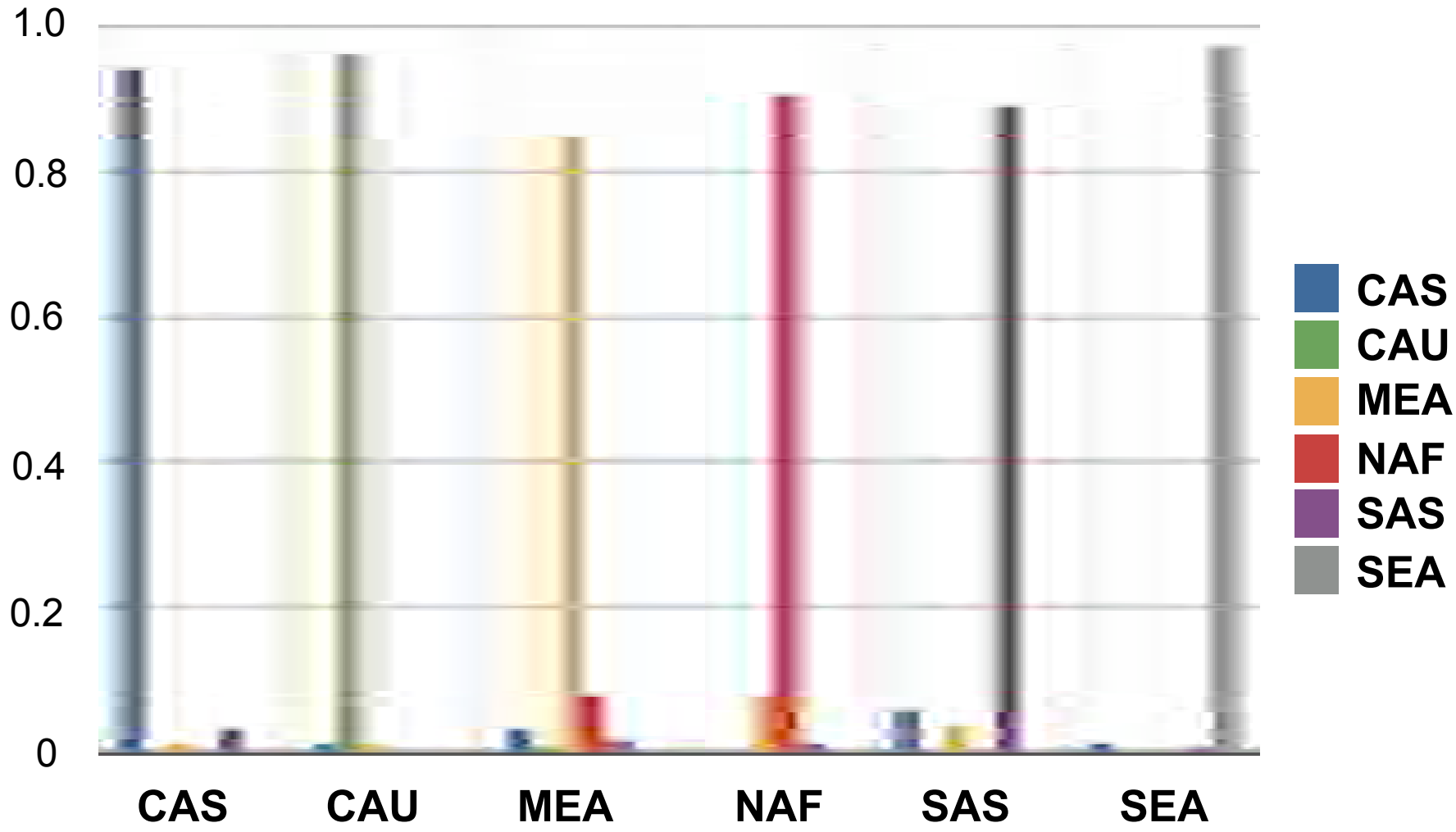
Accuracy of prediction when combining classifiers



Variation Among Regions



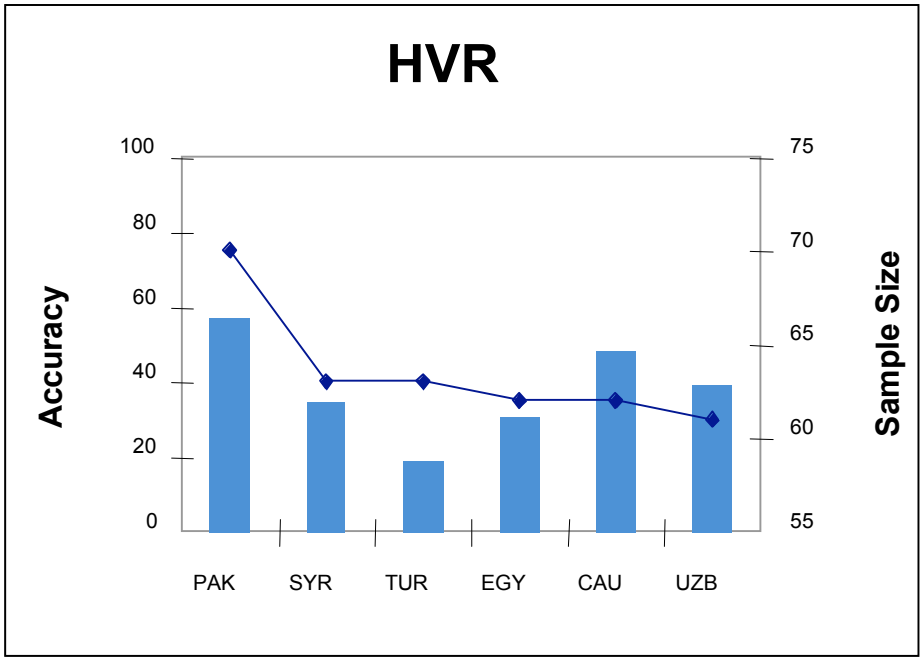
Mistakes Tend to be Geographically Localized



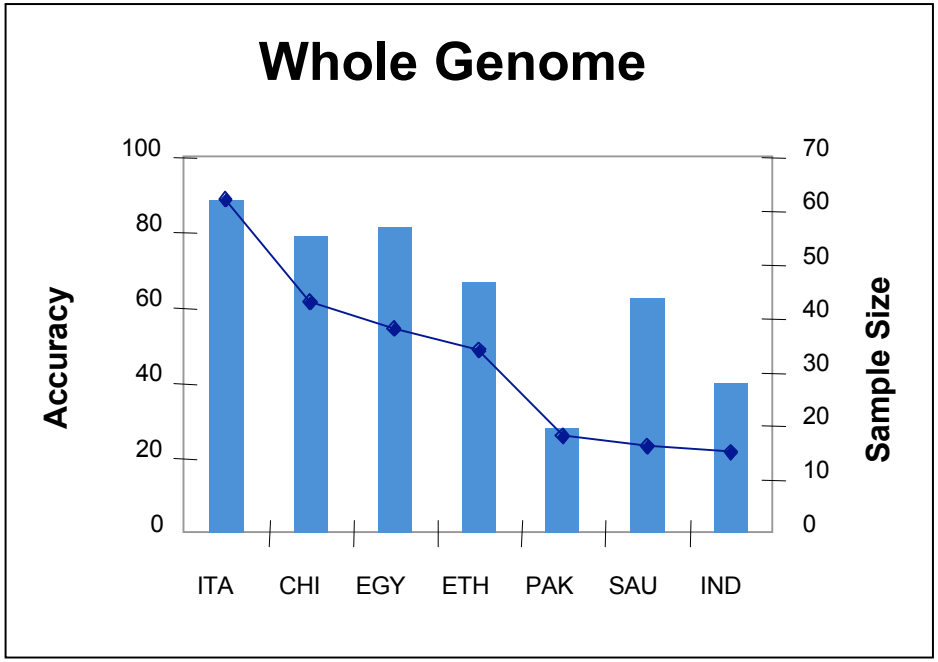
Summary: Y chromosome

- Accurate prediction of haplogroup and continent of origin
  - Geography, ethnicity and language more challenging
- Tandem classifiers improve performance
  - Good strategy: Allow one classifier to disagree
- Error is localized and non-random

Summary: mtDNA



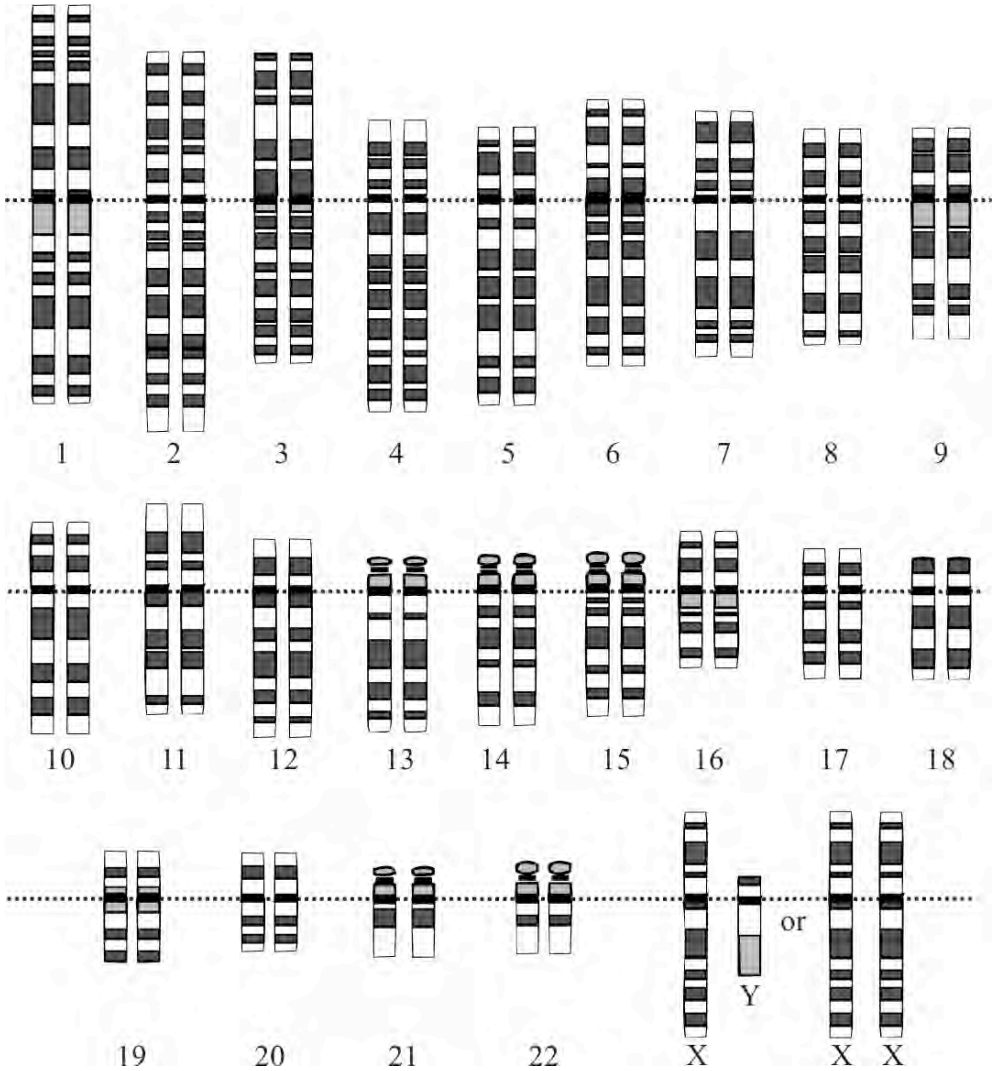
38% accuracy for Population, 44% accuracy for Region



Accuracy based on whole mtDNA sequences is much better, but partly dependent on sample size.

FGS: Accuracy rates as high as 80% can be achieved, depending on the population, as long as sample sizes are large enough

Genome-wide Data



- Genome ~3 Billion bp
- 99.9% identical
- ~3 million SNPs/person
- SNP Chips (up to 2.5M SNPs)

UA and Public SNP Data by Population

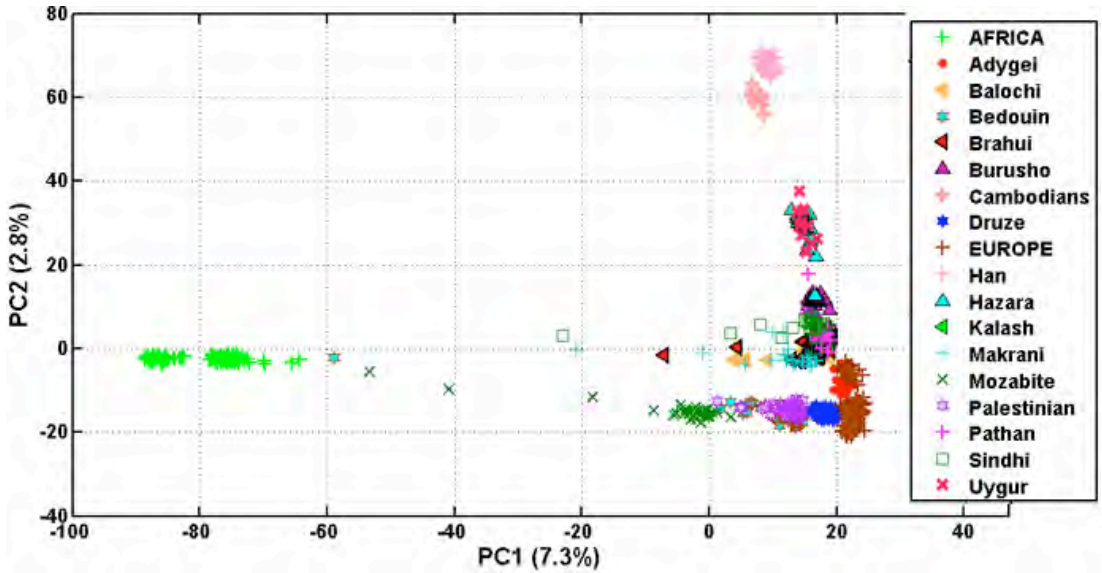
	Public data	This study	Total	
Region	N	N	N	N pop
SSA	102	80	182	7
NAF	116	75	191	10
MEA	271	219	490	18
EUR	226	90	316	18
CAU	86	75	161	10
CAS	46	128	174	12
SAS	194	50	244	12
EAS	219	0	219	16
SEA	0	100	100	4
Total	1260	817	2077	107

Classification Methods

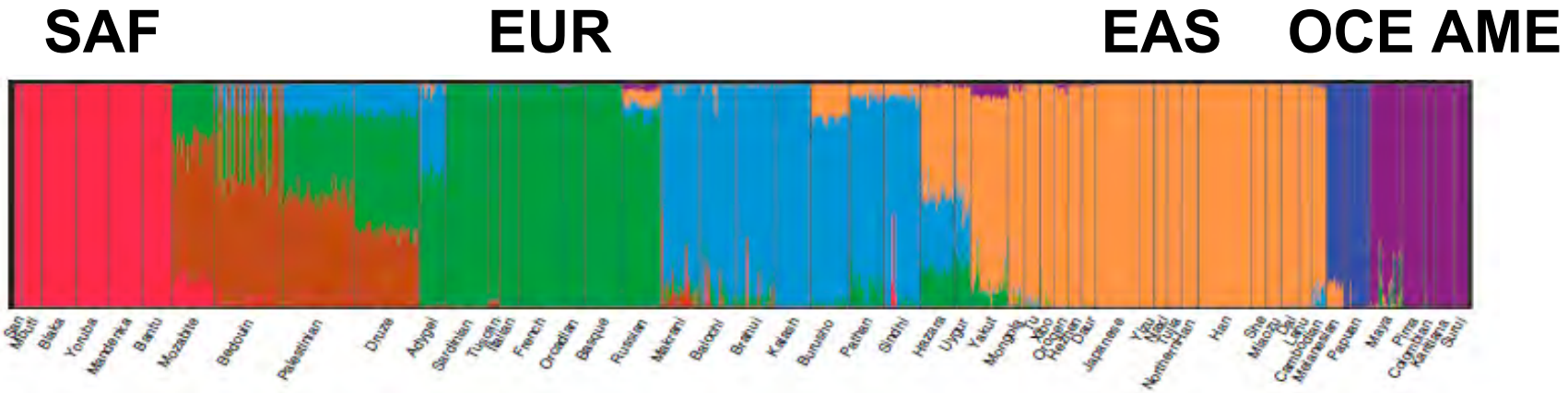
- **Naïve Bayes**
- **Neural nets (multilayer perceptron)**
- Support vector machines (SVMs)
- J48
- Random Forest
- **Logistic regression**
- **K-nearest neighbor (KNN)**
- Discriminant analysis
- Cluster analysis

Dimensionality Reduction Methods

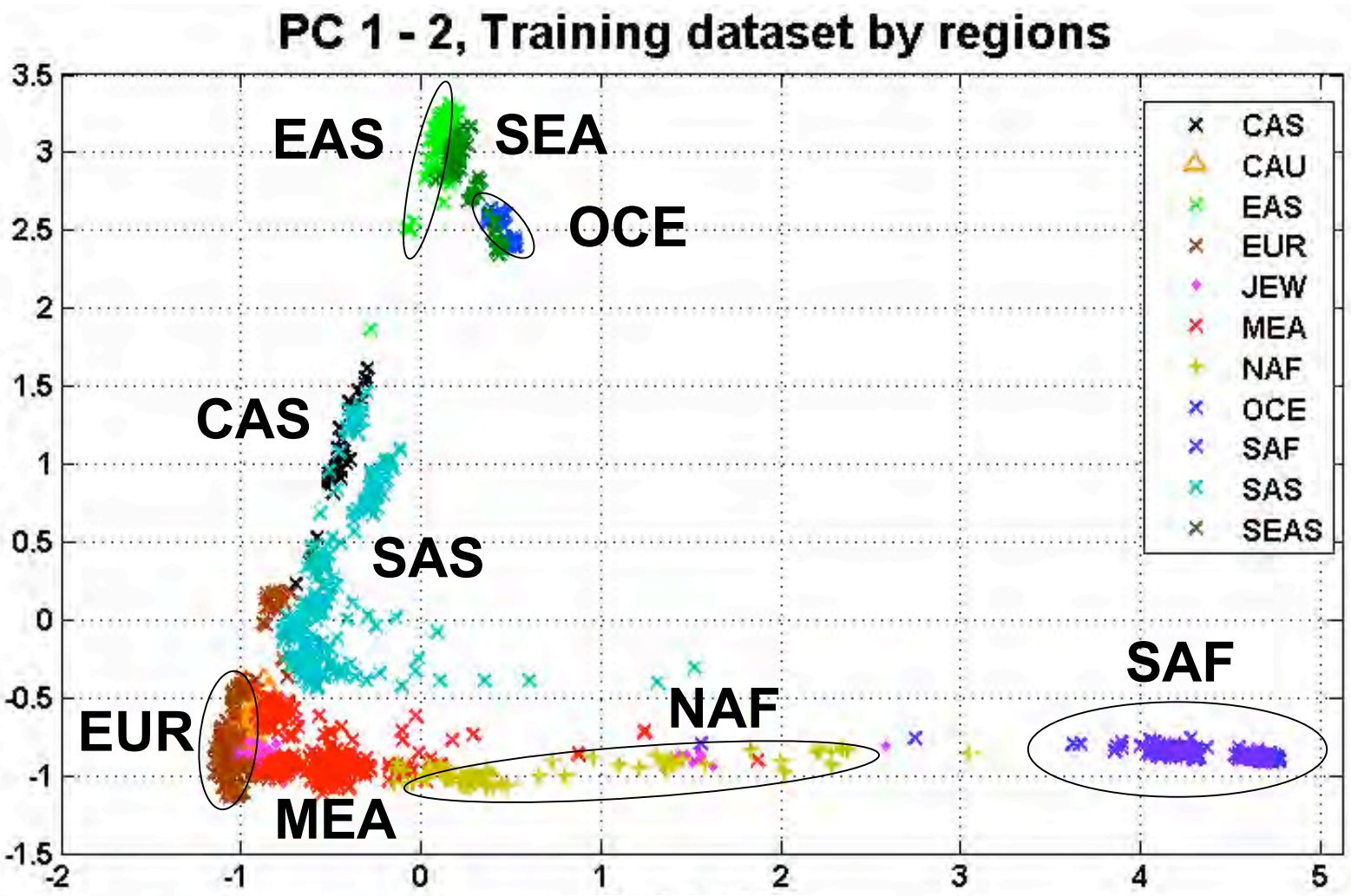
Principal Components Analysis



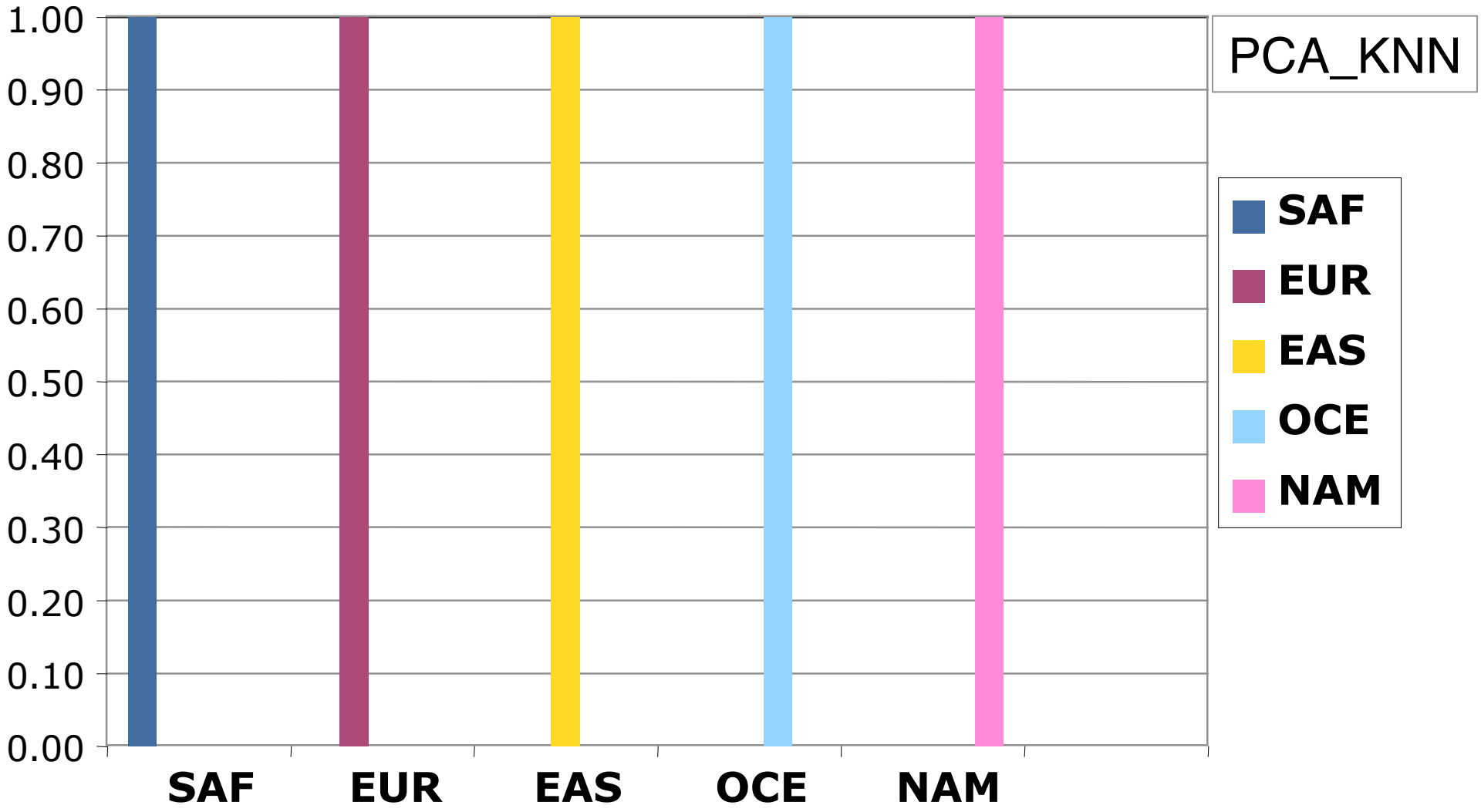
Structure-like Analysis (Admixture)



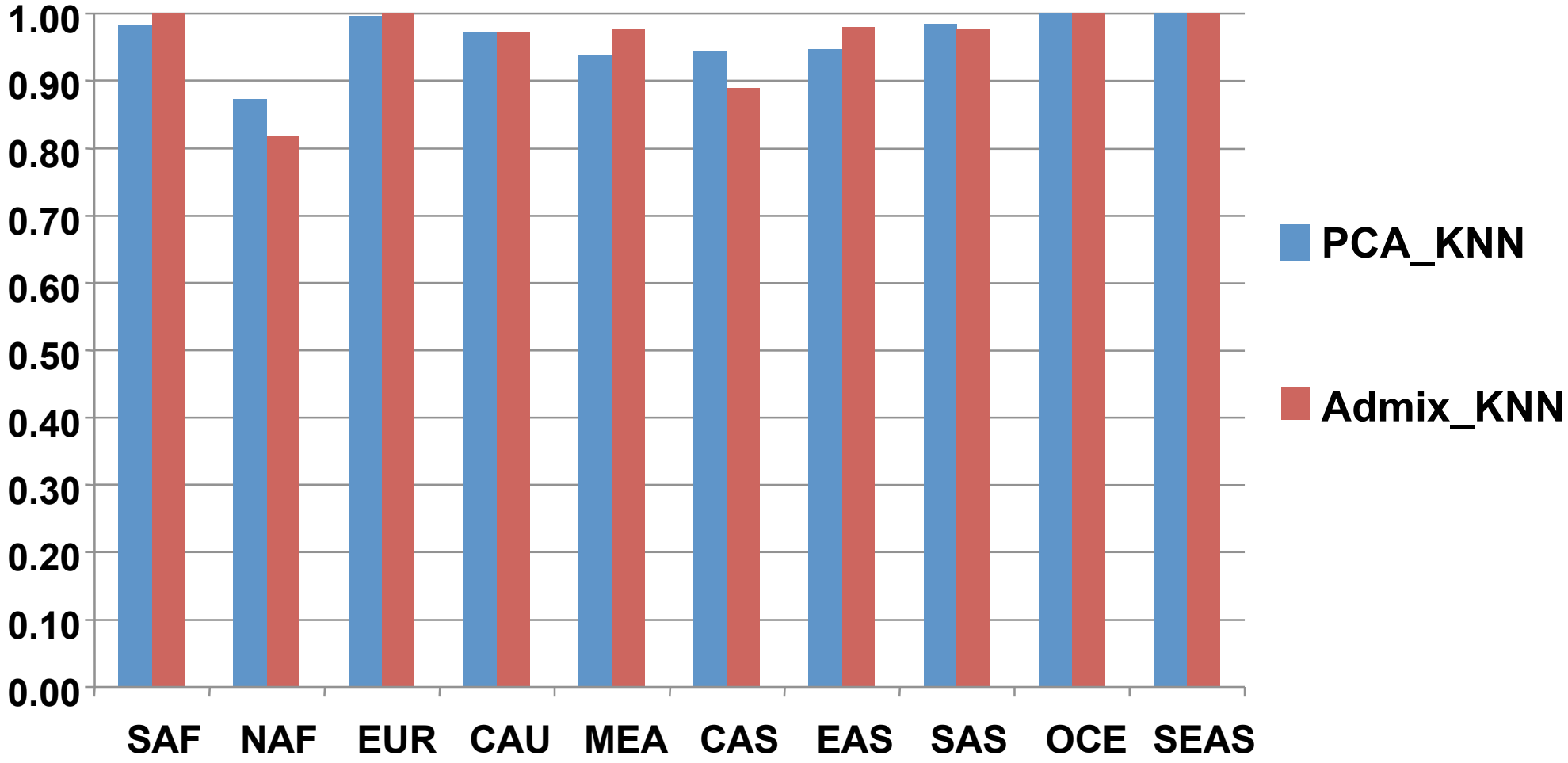
Global Principal Components Analysis



Predicting Continent of Origin from Genome-wide SNPs (100% Accuracy)



Accuracy by Region



Summary: Genome-wide SNPs

- Accurate classification of samples at continental level (near 100%)
- Regional level accuracies quite high (>90%), average accuracy at level of population ~85%
- Gene flow may obscure genetic differences; ethnic labels may be recent for some populations
- Need to Explore Combinations of Tandem Classifiers
- Explore effects of admixture