Estimates of Genetic Ancestry

Chaolong Wang

Sequence Analysis Workshop
December 2014 @ University of Michigan

Outline

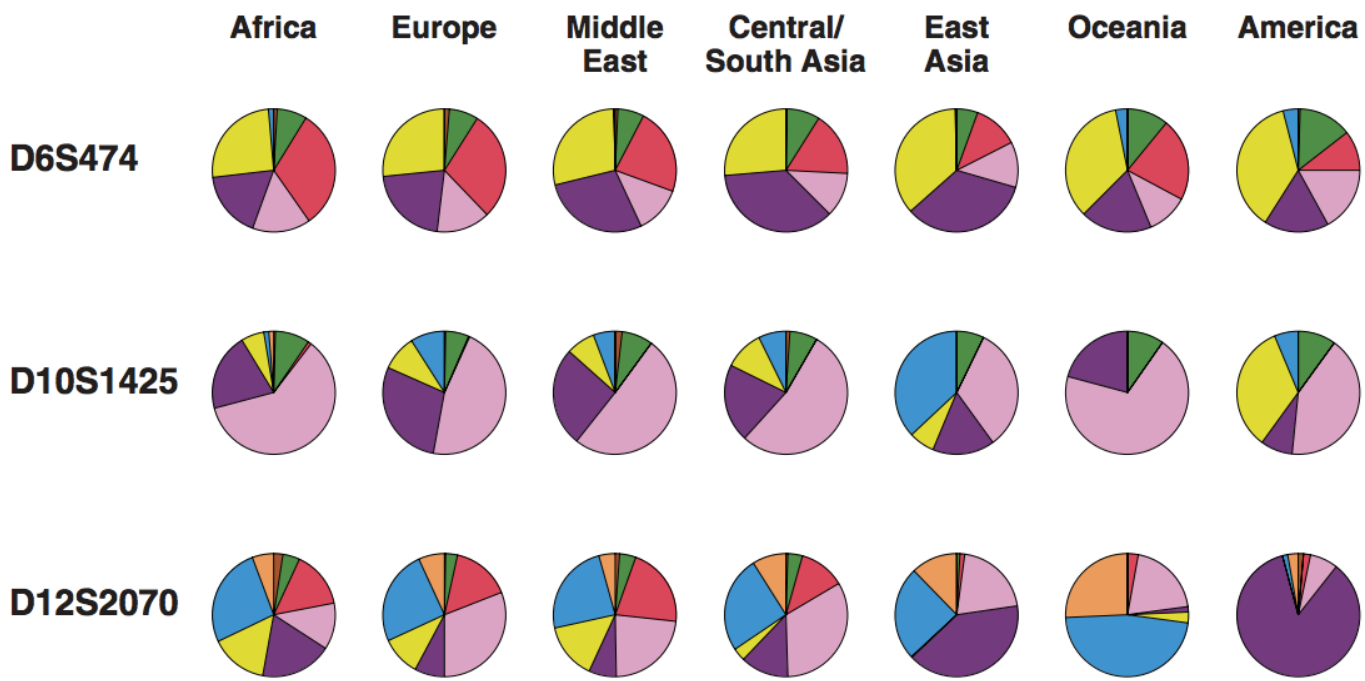
- Background
  - Population structure: causes and consequences
  - Population stratification in genetic association studies
  - Existing methods to estimate genetic ancestry

- Our approaches to estimate genetic ancestry
  - How to infer individual ancestry from small amounts of sequencing or genotyping data?
  - Simulations and empirical examples
  - Potential applications

Population structure

Population structure: Different populations differ in allele frequencies at loci across the genome.

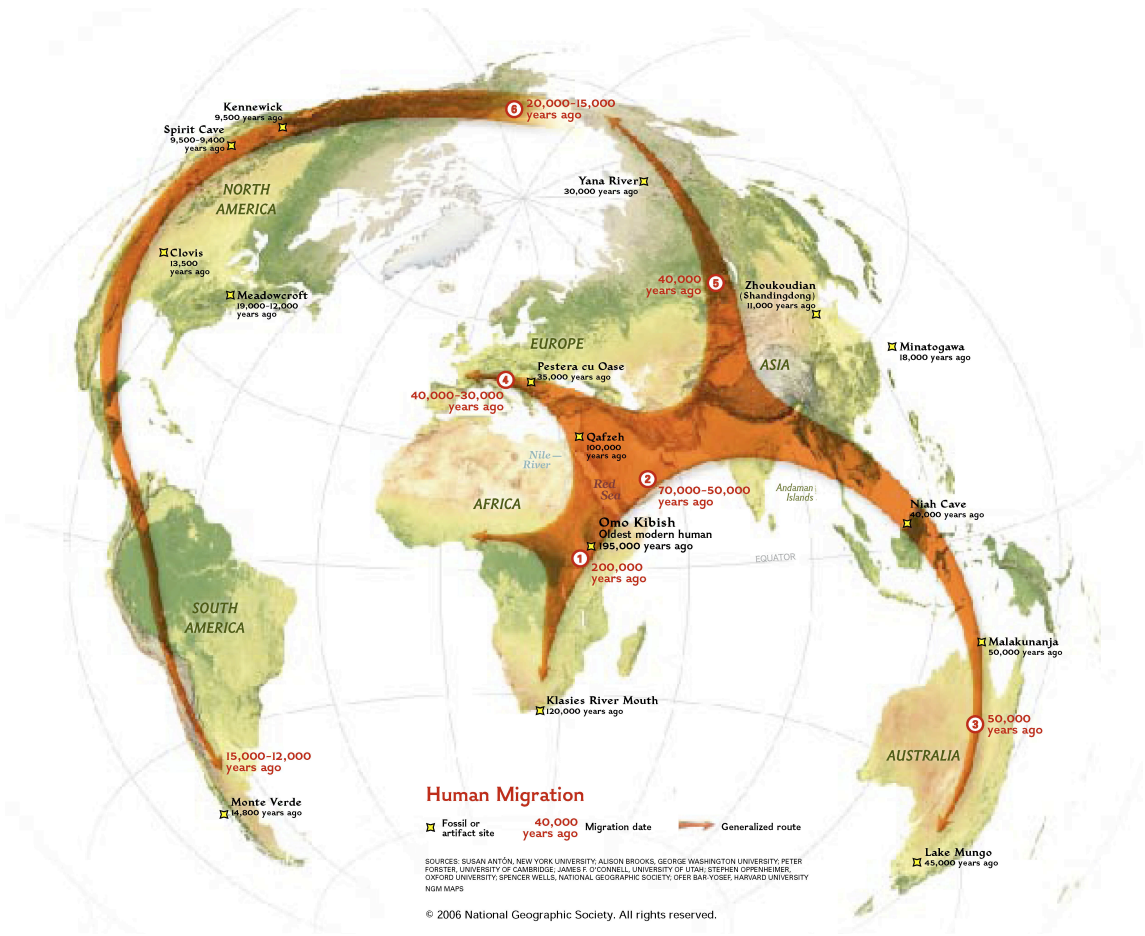
Microsatellite examples:

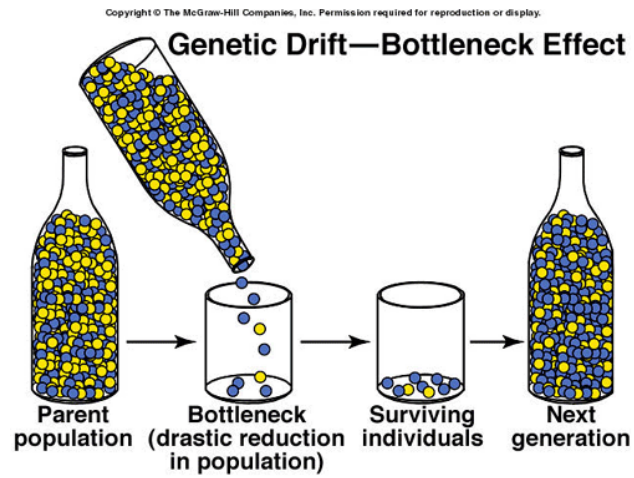


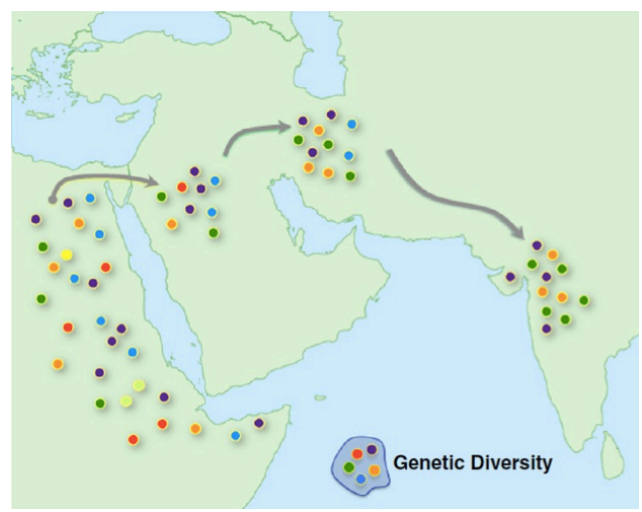
Rosenberg (2011) *Human Biology*

Causes of population structure

Human migration:





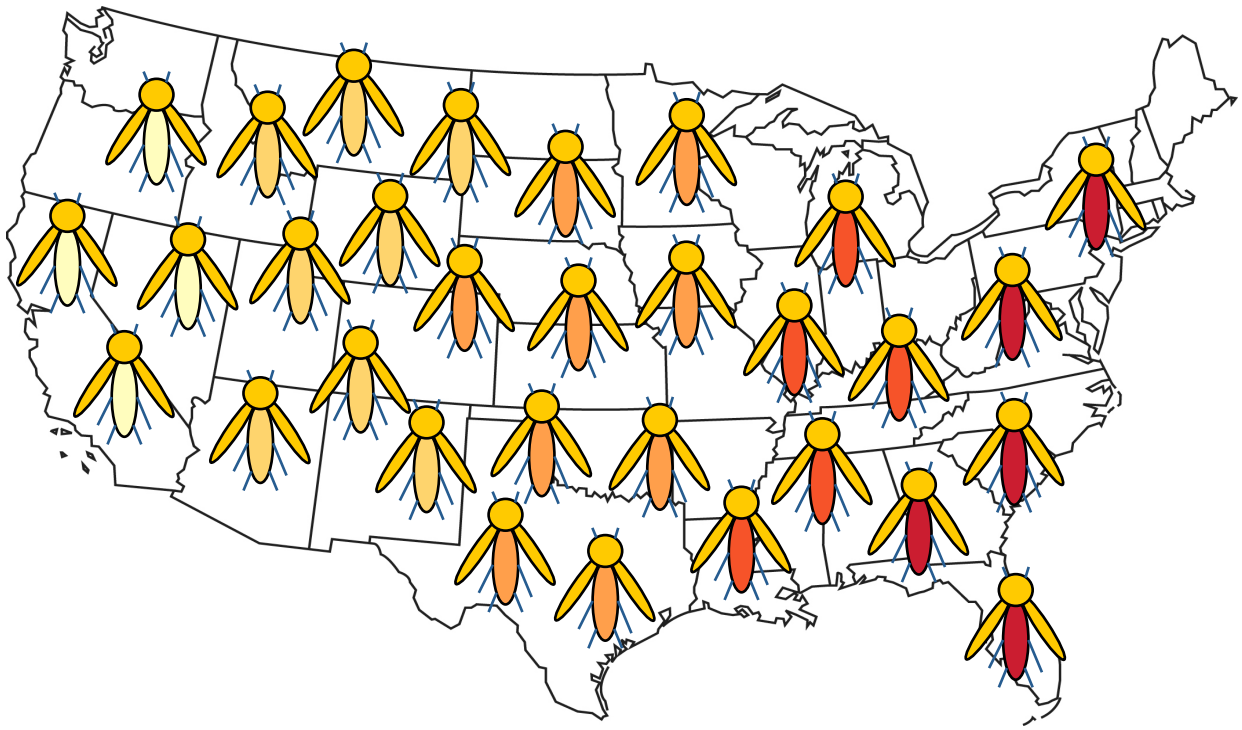


National Geographic

Henn *et al.* (2012) *PNAS*

Causes of population structure

Isolation by distance: gene flow occurs more frequently between neighboring groups.

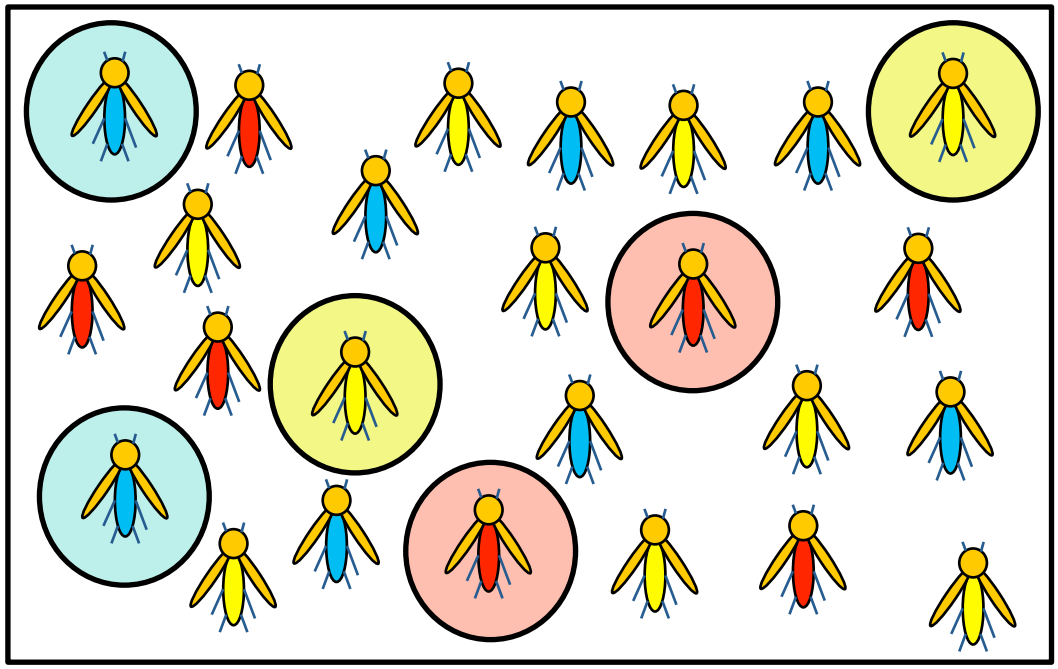


Genetic similarity decreases as geographic distance increases.

Figure: Michael DeGiorgio

Causes of population structure

Assortative mating: mating occurs more often between individuals from similar "classes" (color, education, social stratification).

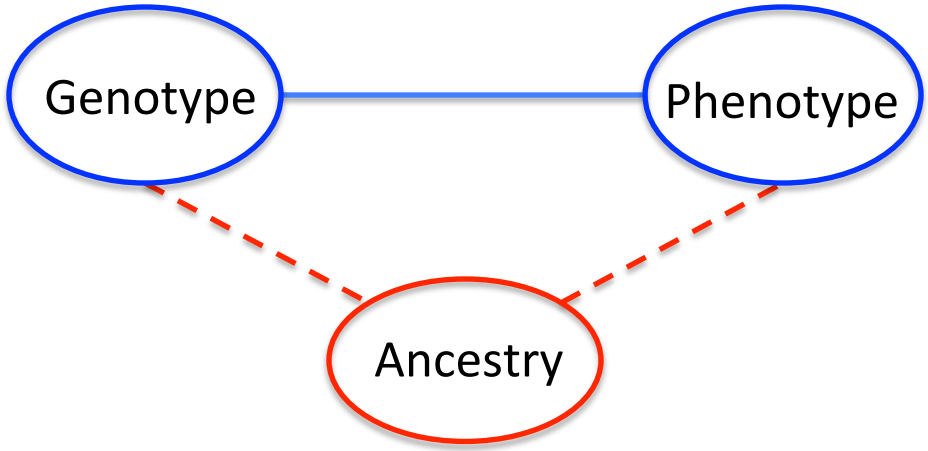


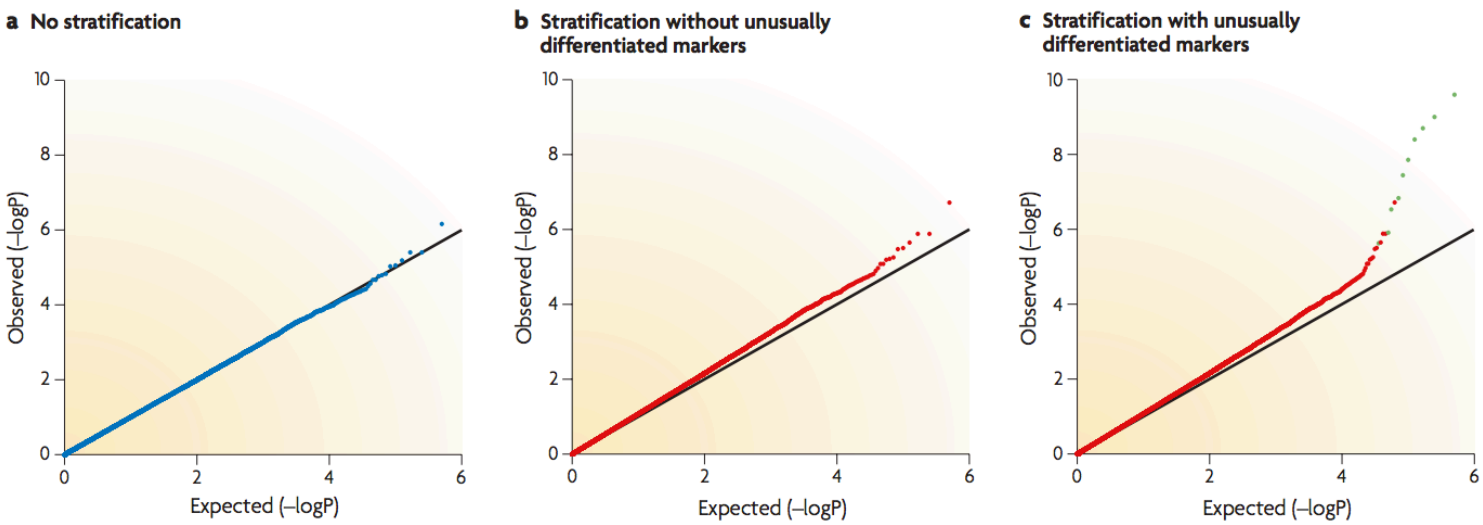
People from the same "class" are genetically more similar.
Example: upper caste and lower caste in India

Figure: Michael DeGiorgio

Population stratification in genetic association studies

Population stratification: systematic ancestry differences between subjects with different phenotypes, leading to spurious association.



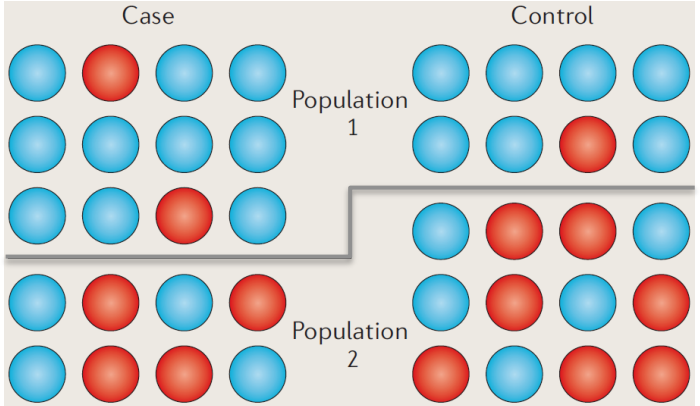


Price *et al.* 2010, *Nat Rev Genet*

Consequence of population structure in association studies

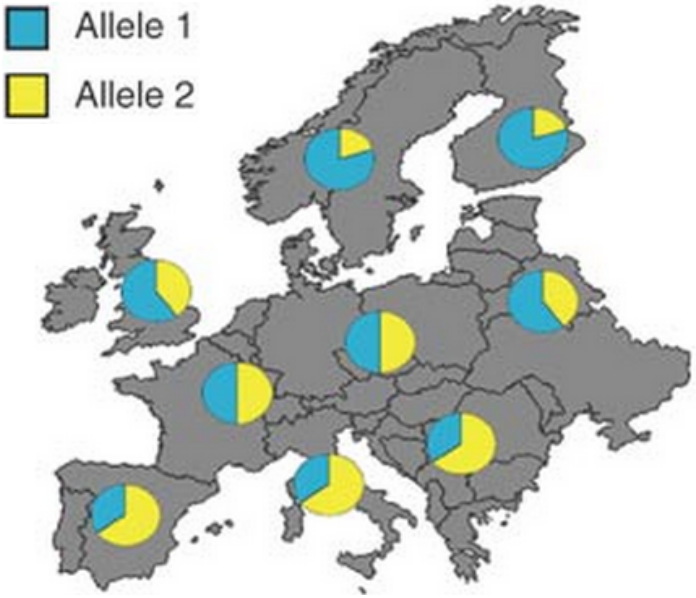
Population stratification: systematic ancestry differences between subjects with different phenotypes, leading to spurious association.

Case-control:



Balding (2006) *Nat. Rev. Genet.*

Quantitative trait: (e.g. height)



Campbell *et al.* (2005) *Nat. Genet.*

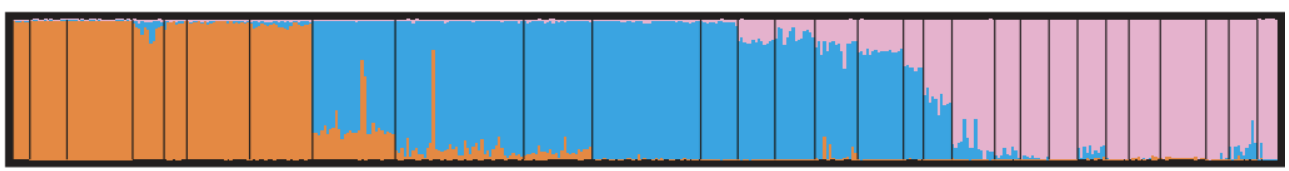
Methods to estimate population structure

- **High-dimensional genotype data (>1K samples, >100K loci)**

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	…	SNP L
Sample 1	0	2	1	0	2	…	2
Sample 2	1	0	1	-	2	…	1
…	…	…	…	…	…	…	…
Sample N	2	1	0	0	2	…	2

- **Model-based clustering methods**
  - STRUCTURE/ADMIXTURE/FRAPPE
  - Model allele frequencies of (pre-specified) *K* discrete clusters
  - Computationally challenging for large datasets
  - Not suitable for continuous population structure

Example for 29 worldwide populations (Jakobsson *et al.* 2008, *Nature*)



K=3

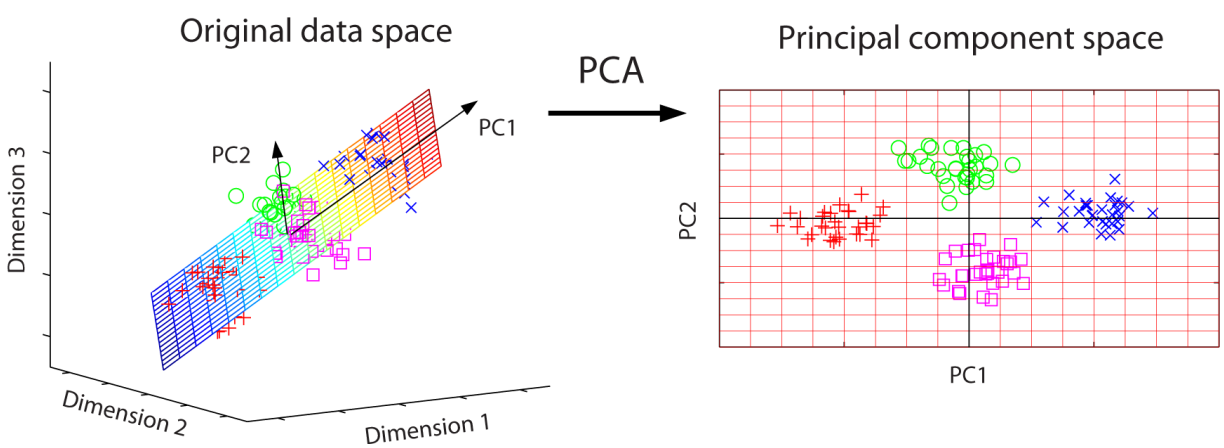
Africa Middle East/Europe Asia/Oceania/Americas

Methods to estimate population structure

- **High-dimensional genotype data (>1K samples, >100K loci)**

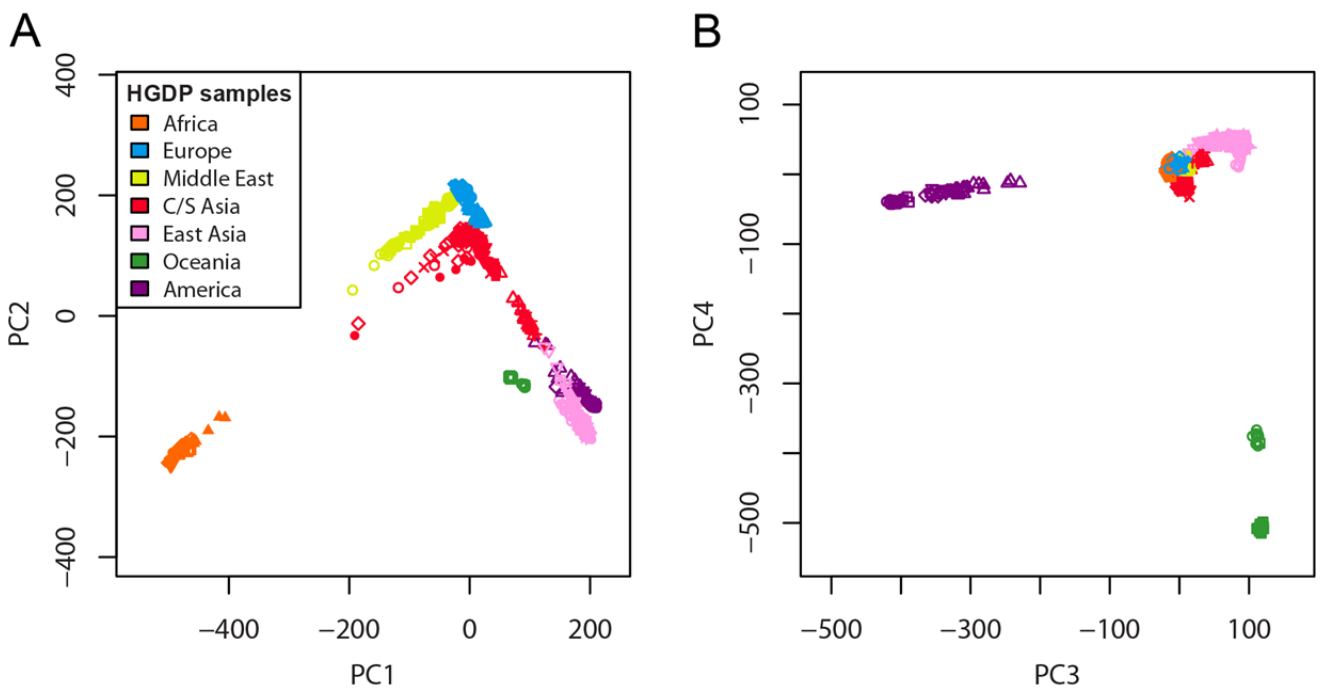
	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	…	SNP L
Sample 1	0	2	1	0	2	…	2
Sample 2	1	0	1	-	2	…	1
…	…	…	…	…	…	…	…
Sample N	2	1	0	0	2	…	2

- **Multivariate dimension reduction methods**
  - Principal components analysis (PCA)
  - Multidimensional scaling (MDS)



PCA in decomposing population structure

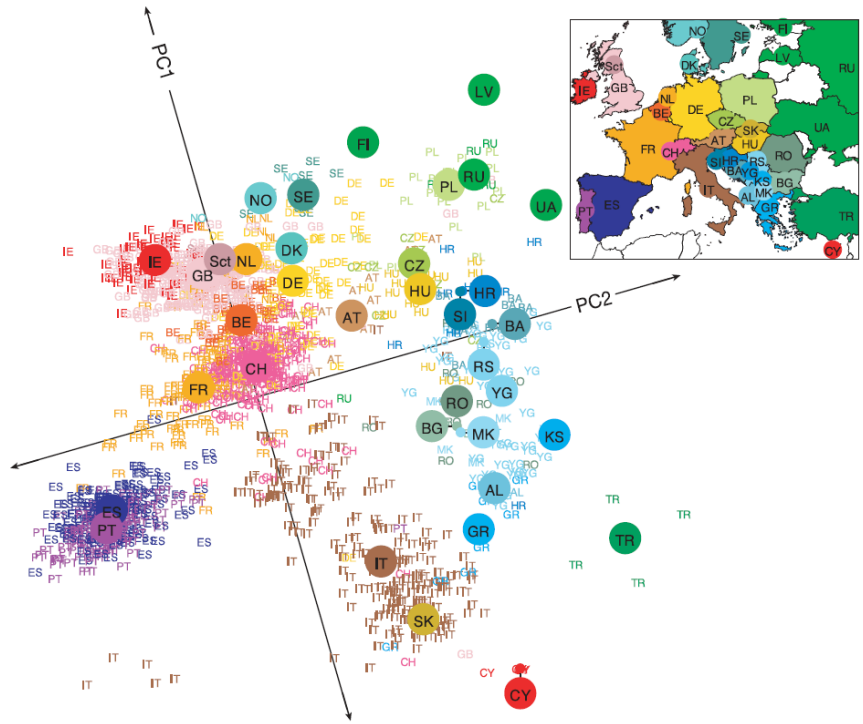
- **Worldwide population structure**
  - Human Genome Diversity Panel (HGDP, 53 worldwide populations)



Li *et al.* (2008) *Science*; Jakobsson *et al.* (2008) *Nature*

PCA in decomposing population structure

- **European population structure**
  - Population Reference Panel (POPRES, 37 European populations)



Title of the paper: Genes mirror geography in Europe

Novembre *et al.* (2008) *Nature*

Control of population stratification in association studies

- **Control for stratification using estimated ancestry:**
  - Stratified analysis of subgroups followed by meta-analysis
  - Regression on ancestry principal components
  - Genetic matching of study subjects

- **Other approaches without explicitly estimating ancestry:**
  - Linear mixed models
  - Genomic control

- **These approaches require high-quality genotype data across the genome.**
  - GWAS array genotyping data
  - Whole genome sequencing (when genotypes can be accurately estimated)

Targeted sequencing experiments

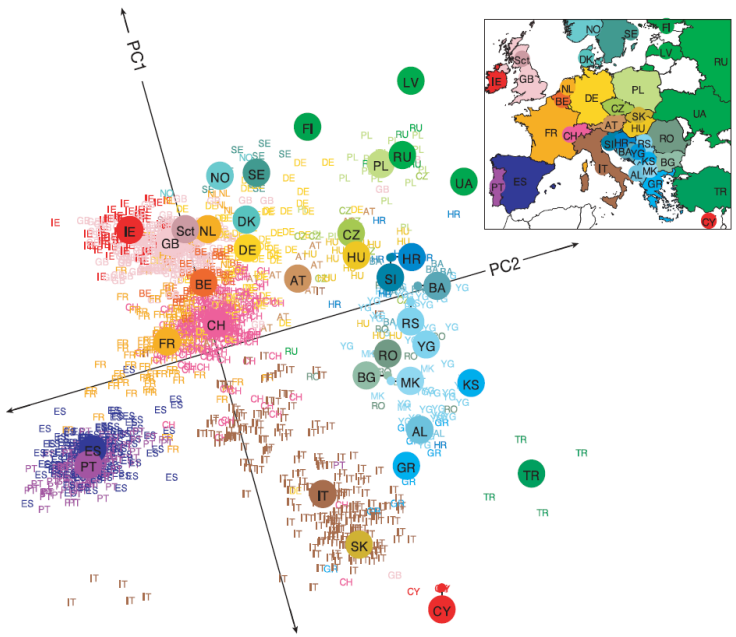
- **Targeted sequencing focuses on specific regions of interests.**
  - Our AMD study: 10 candidate regions (2MB in total)
  - Goal: search for additional high-risk (rare) variants that provide functional information about the disease

- **Large sample size is required to provide statistical power to detect association signal for rare variants.**
  - Many studies now include >10,000 individuals
  - Likely to include samples of different ancestry background

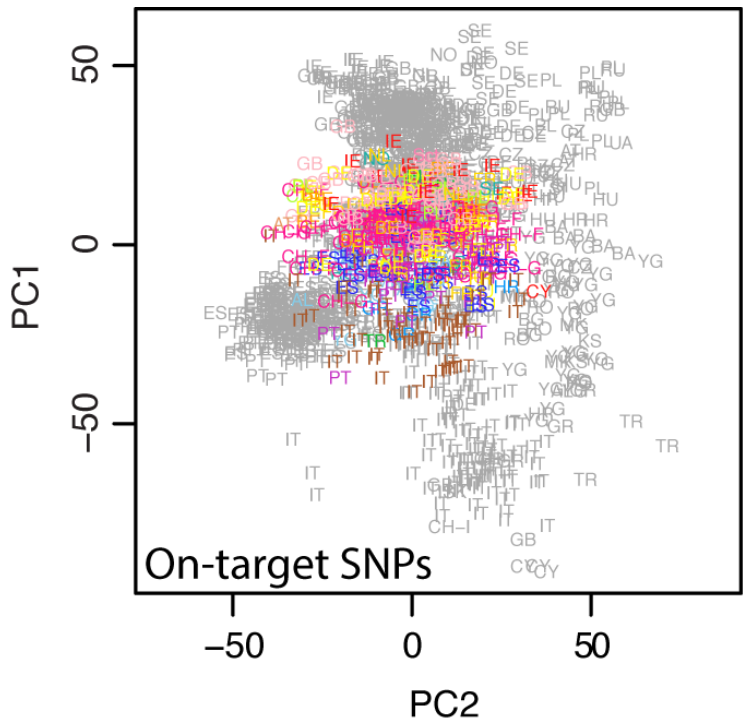
- **Correcting for population stratification is difficult for targeted sequencing experiments.**
  - Too few variant loci within targeted regions

Estimating ancestry for targeted sequencing data

What happens when we apply PCA to targeted sequencing data?

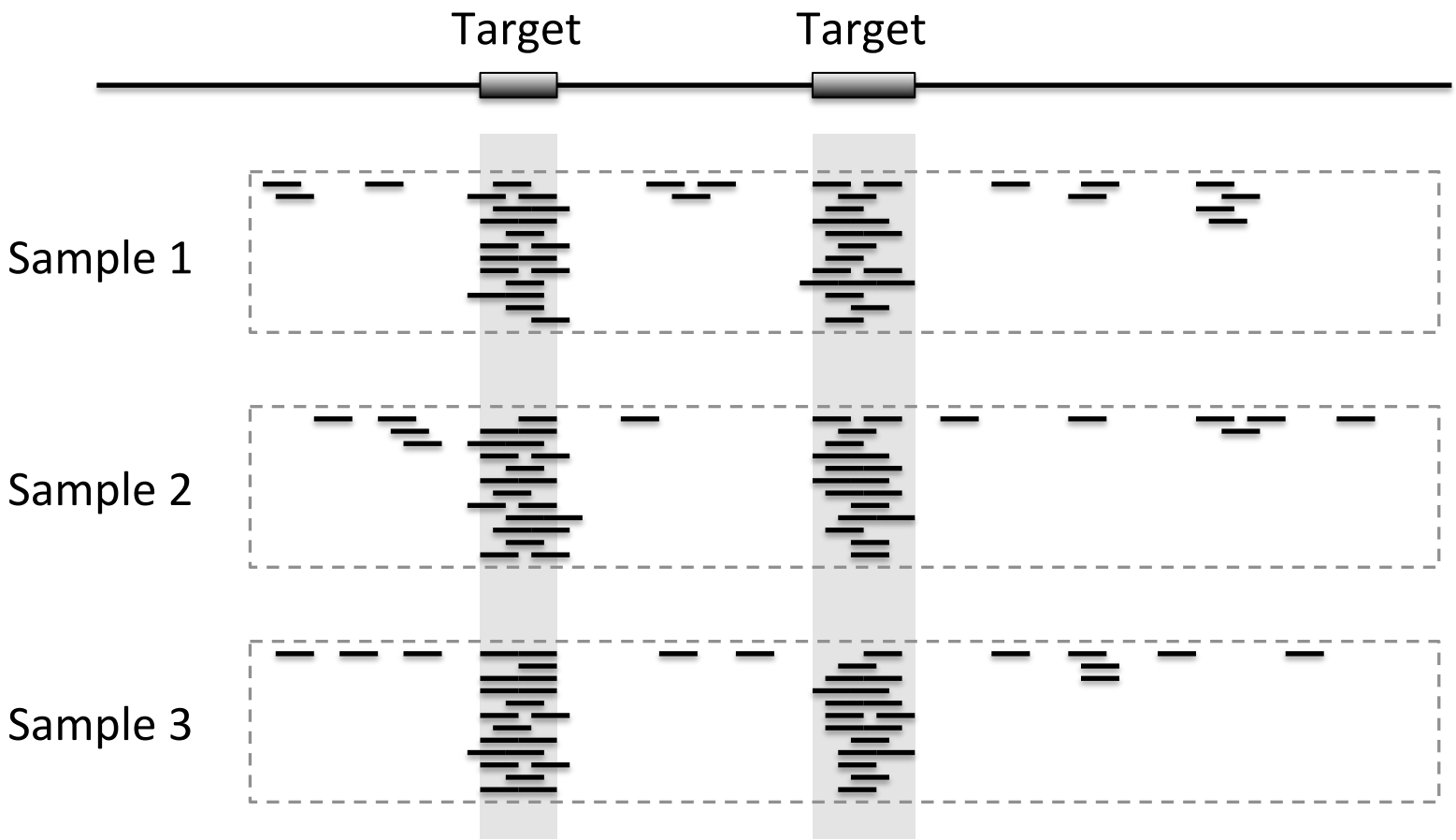


Novembre *et al.* (2008) *Nature*



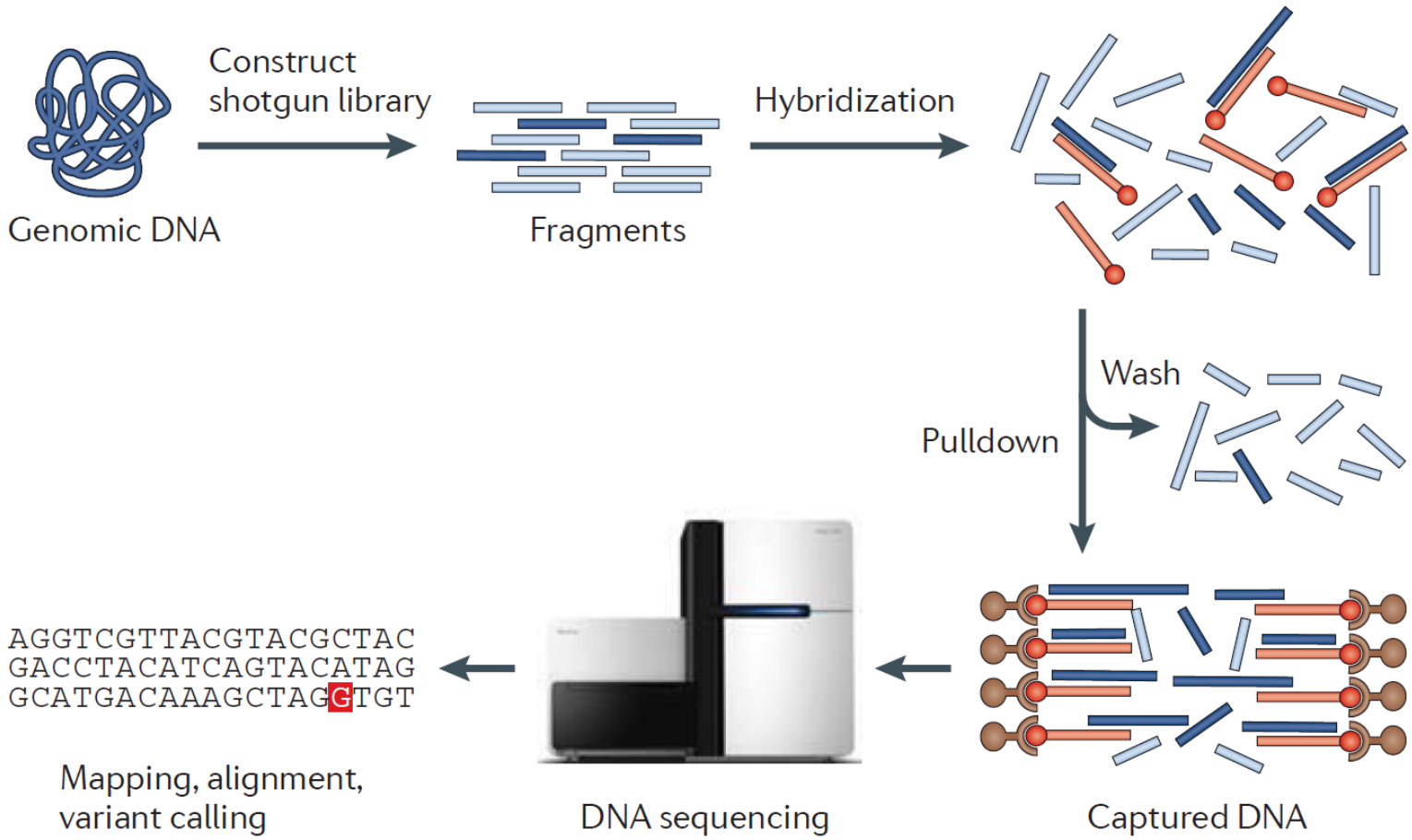
2,547 SNPs in POPRES data overlapped with whole exome sequencing

Targeted sequencing data



A lot of sequence reads distribute *randomly* and *sparsely* across the off-target genome!

Workflow of target/exome sequencing



Bamshad *et al.* (2011) *Nat. Rev. Genet.*

LASER: **L**ocating **A**ncestry from **SE**quence **R**eads

- Traditional methods such as PCA cannot be directly applied on off-target sequencing data.
  - Genotype uncertainty
  - Large amount of missing data

The LASER method:

- Use off-target sequence reads to place sequenced samples one by one into a reference PCA map of ancestry
  - Directly analyze sequence reads without calling genotypes
  - Analyze each sample with a set of reference individuals

Ancestry estimation and control of population stratification for sequence-based association studies

Chaolong Wang[1,2,10], Xiaowei Zhan[2,10], Jennifer Bragg-Gresham[2], Hyun Min Kang[2], Dwight Stambolian[3], Emily Y Chew[4], Kari E Branham[5], John Heckenlively[5], The FUSION Study[6], Robert Fulton[7], Richard K Wilson[7], Elaine R Mardis[7], Xihong Lin[1], Anand Swaroop[8], Sebastian Zöllner[2,9] & Gonçalo R Abecasis[2]

NATURE GENETICS VOLUME 46 | NUMBER 4 | APRIL 2014

Data used in LASER

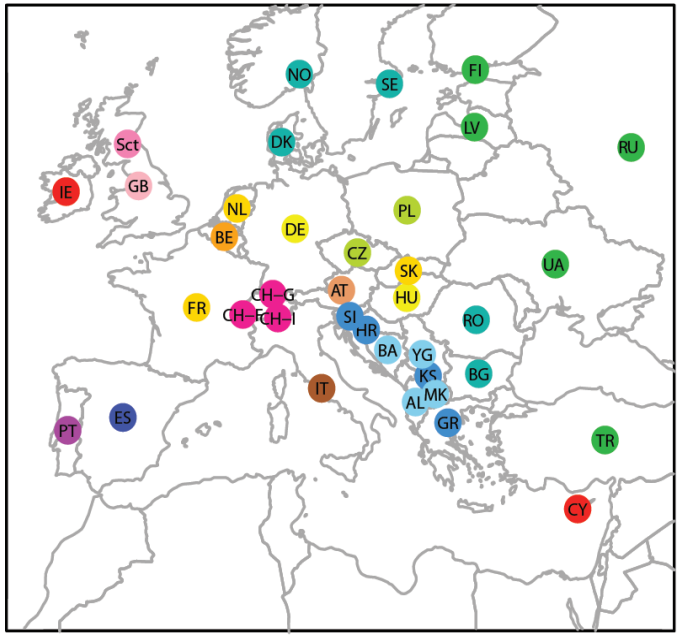
- Study samples: low-coverage sequencing reads sparsely distributed across off-target regions.

- Reference samples with known ancestry: high-quality genome-wide SNP data.
  - Human Genome Diversity Panel (HGDP)
    - 938 individuals from 53 worldwide populations
    - 632,958 autosomal SNPs after QC
    - Li *et al.* (2008) *Science*
  - Population Reference Sample (POPRES)
    - 1,385 individuals from 37 European populations
    - 318,682 autosomal SNPs after QC
    - Novembre *et al.* (2008) *Nature*

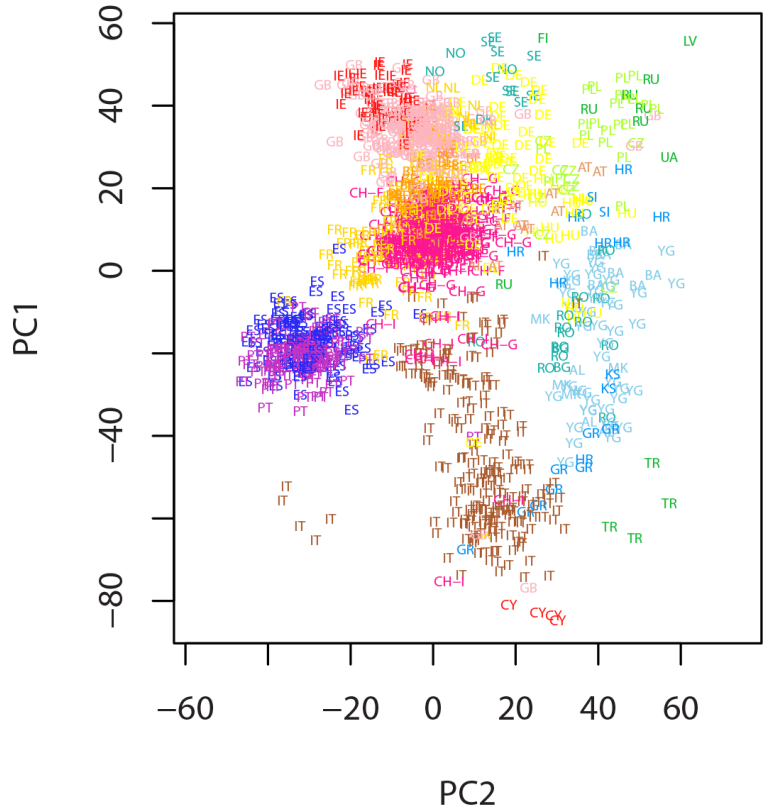
Step 1: create a reference map

- Generate a reference map by applying PCA on SNP data of N reference individuals. (Map 0)

Geographic map



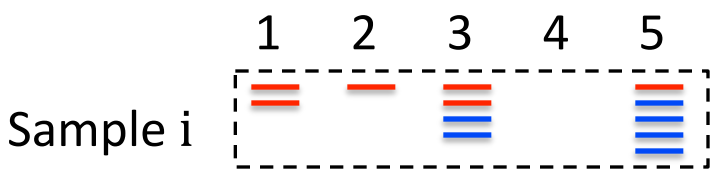
Map 0



37 populations, 1,385 individuals
318,682 autosomal SNPs
Novembre *et al.* (2008) *Nature*

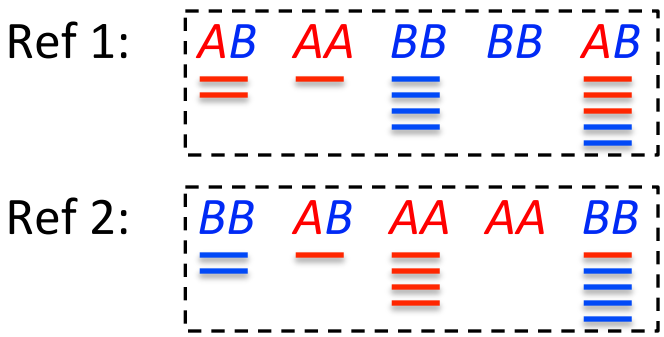
Step 2: adjust reference to each sample

Given a sample i that was sequenced with coverage C_{ij} at locus j, for j=1,2,…,L.

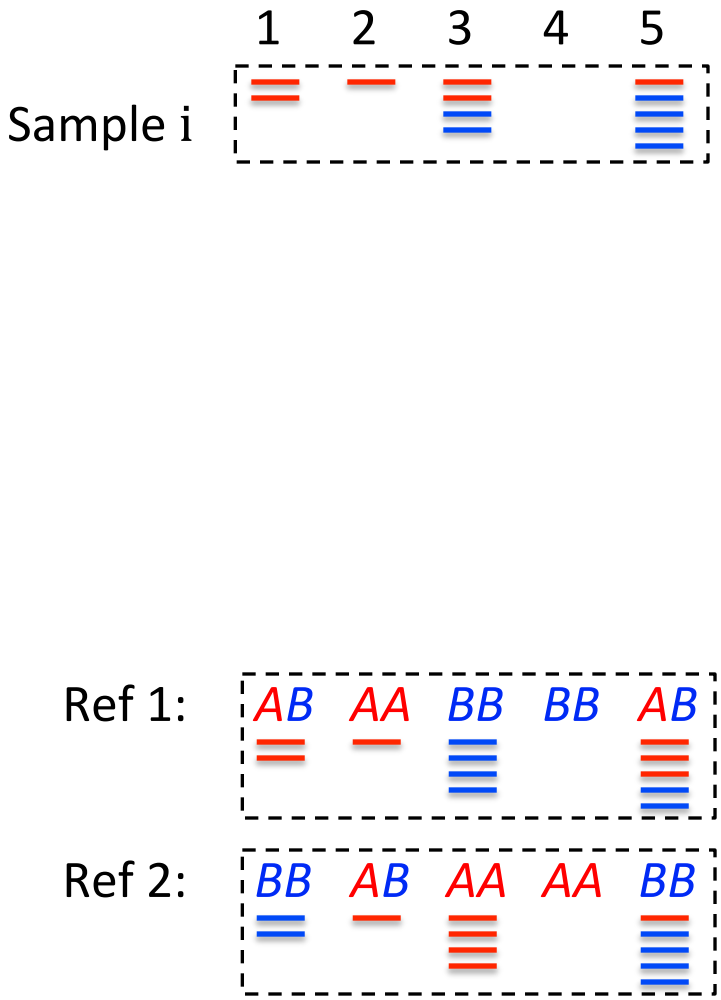


- Simulate sequence data for all reference individuals with coverage at each locus j equal to C_{ij}.

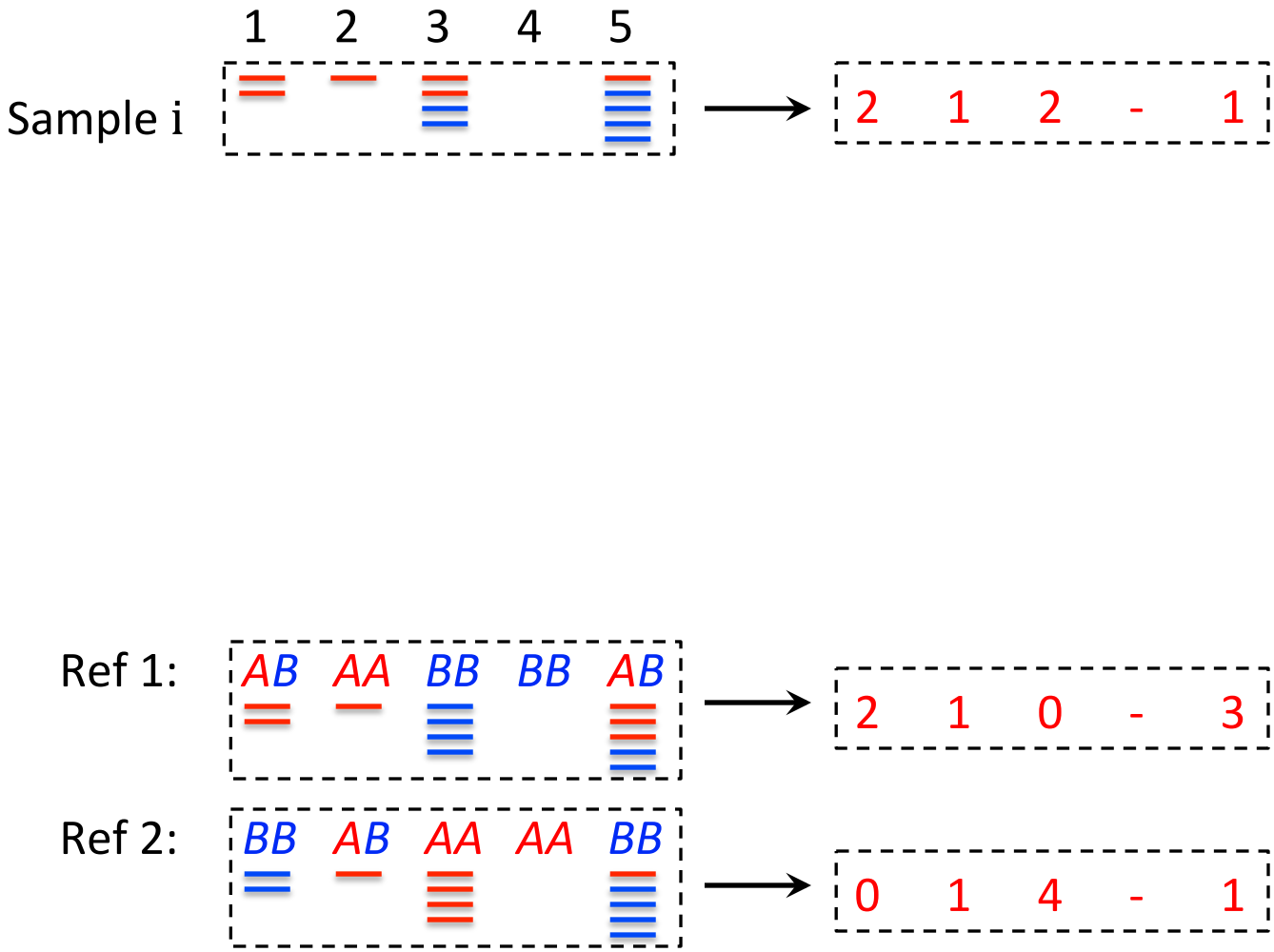
$$P(\text{drawing a read } A) = \begin{cases} 1-e & \text{if } g_{ij} = AA \\ 0.5 & \text{if } g_{ij} = AB \\ e & \text{if } g_{ij} = BB \end{cases}$$



Step 2: adjust reference to each sample

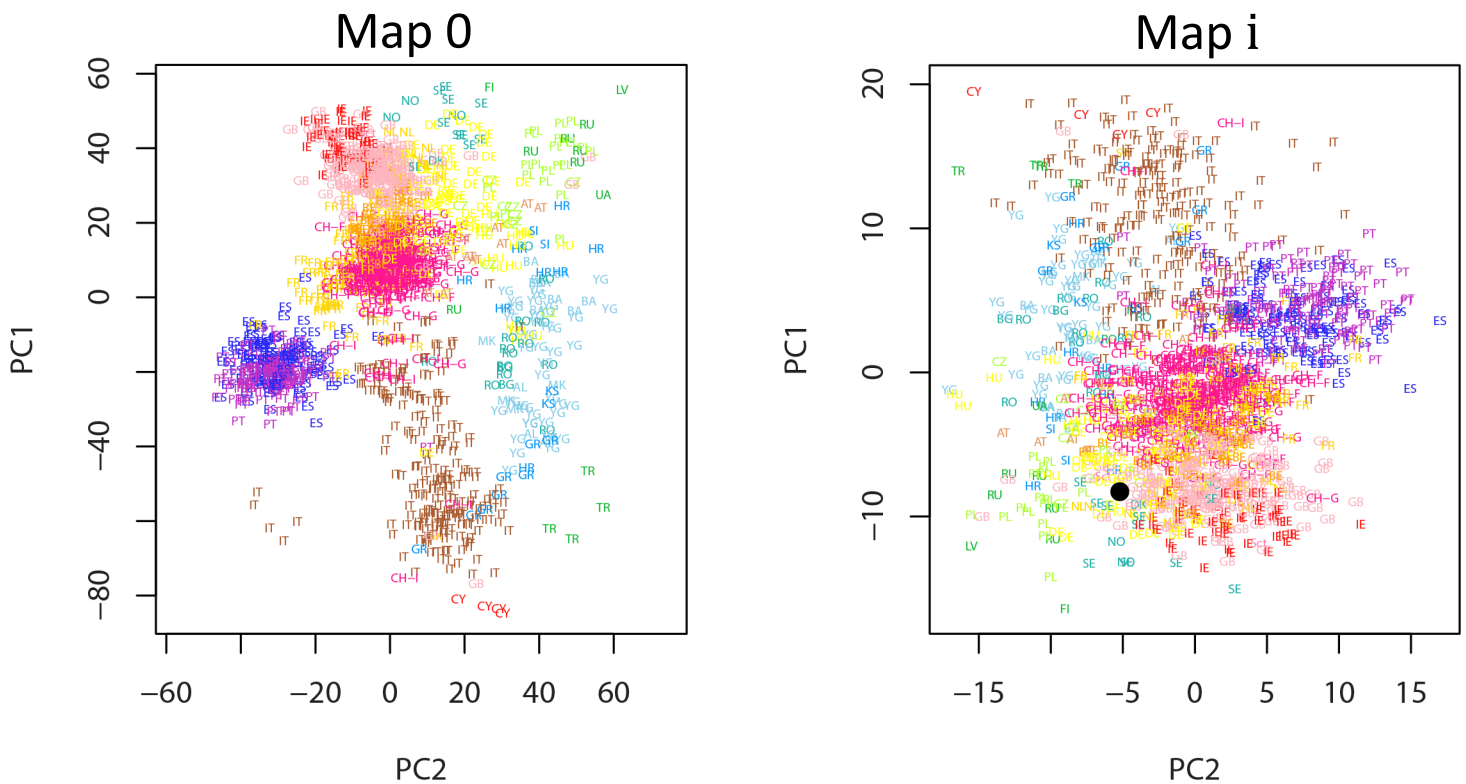


Step 3: count the variant bases at each locus



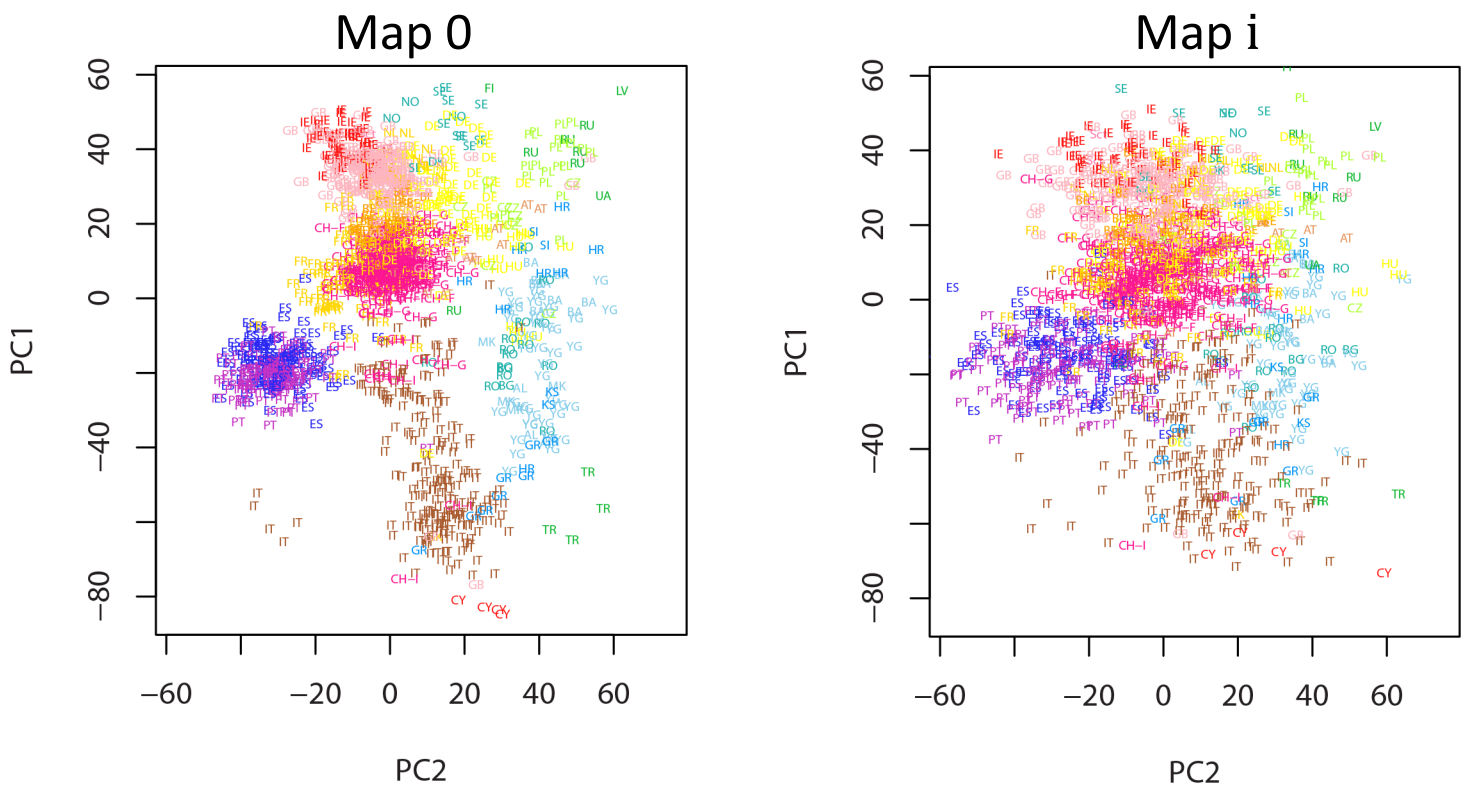
Step 4: construct a sample-specific map

- Perform PCA on combined sequencing data of sample i and N reference individuals. (Map i)



Step 5: find optimal transformation

- **Procrustes analysis**: transform Map i to optimize the similarity to Map 0 based on N reference samples.



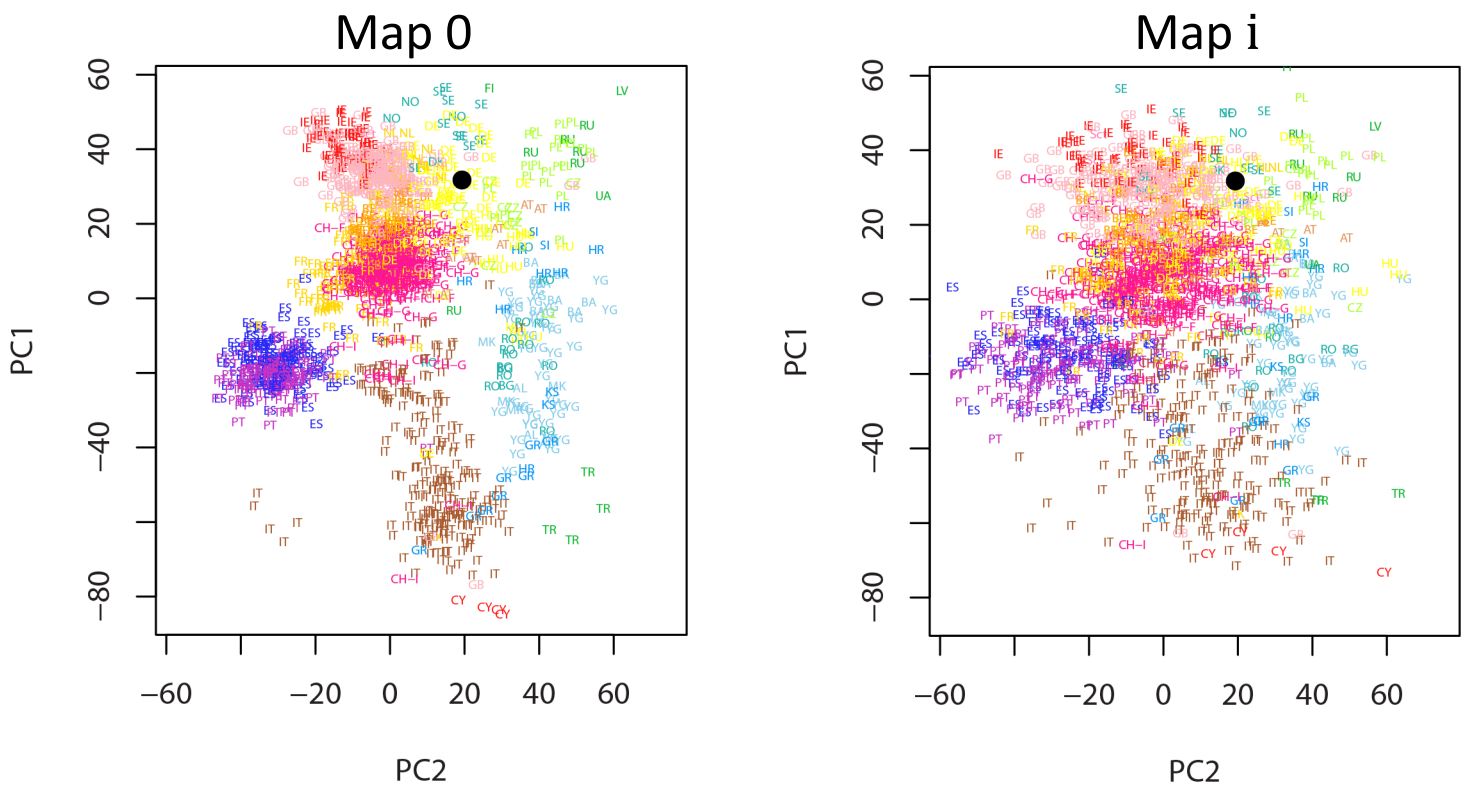
Transformations: $f(\boldsymbol{Y}) = \rho \boldsymbol{Y}\boldsymbol{A} + \boldsymbol{B}$

Minimize objective: $d(\boldsymbol{X}, f(\boldsymbol{Y})) = \sum_{i=1}^{N} (\boldsymbol{x}_i - f(\boldsymbol{y}_i))^T (\boldsymbol{x}_i - f(\boldsymbol{y}_i))$

Wang *et al.* (2010) *Stat. Appl. Genet. Mol. Biol.*

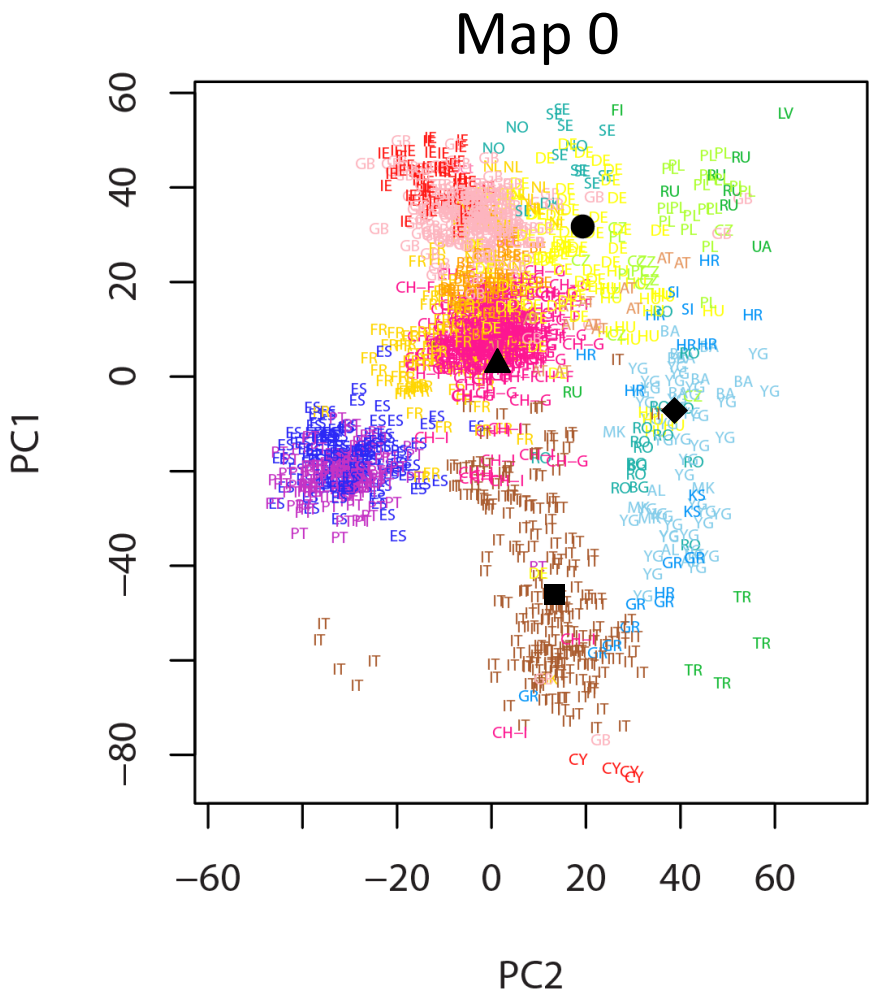
Step 6: apply transformation

- Apply the transformation on coordinates of sample i to place it into the reference PCA map.

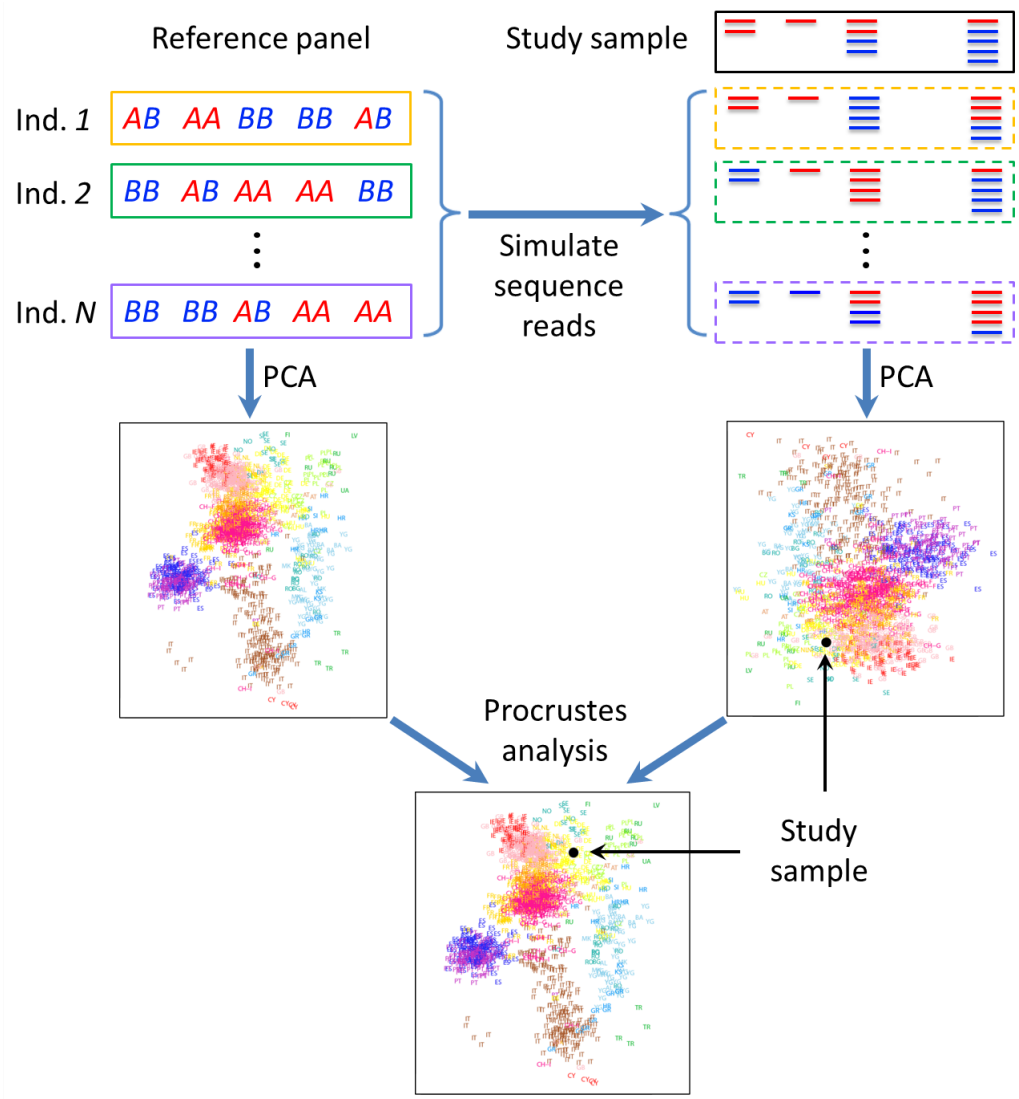


Step 7: repeat!

- Repeat steps 2-6 for all sequenced samples one by one.

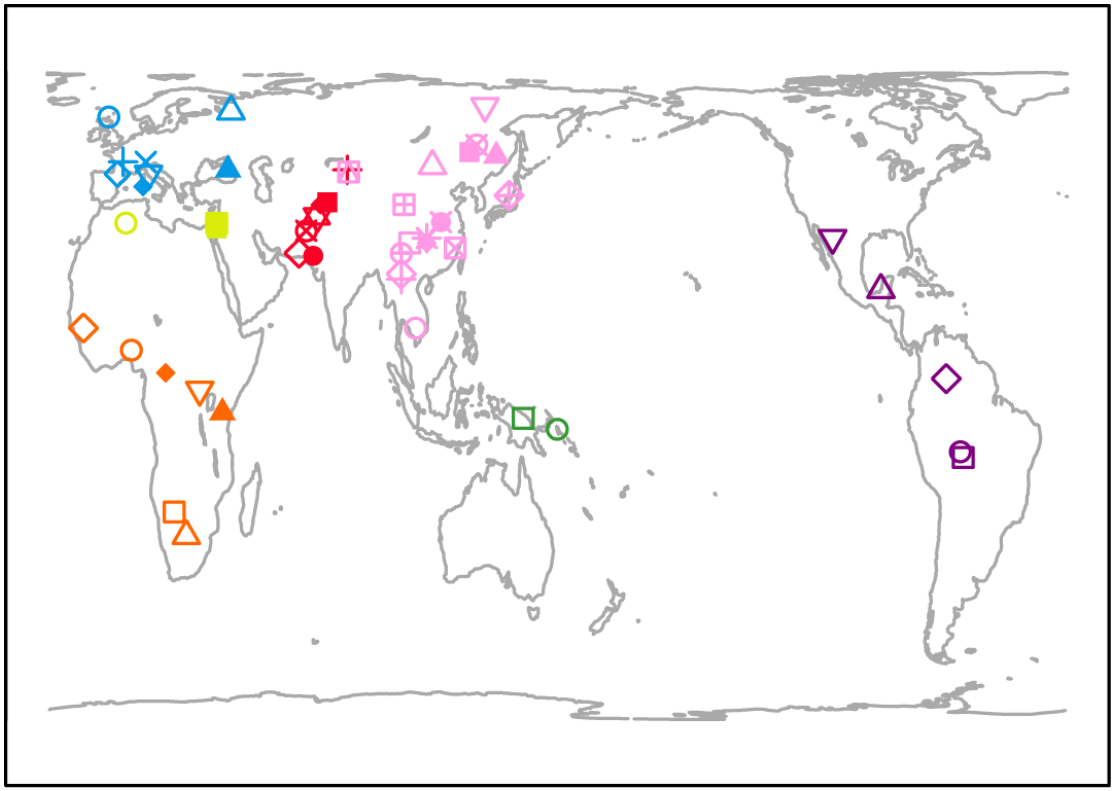


LASER: Locating Ancestry from SEquence Reads



Wang *et al.* (2014) *Nat Genet.*

Human Genome Diversity Panel (HGDP)



Africa
- Bantu (Kenya)
- Bantu (S. Africa)
- Biaka Pygmy
- Mandenka
- Mbuti Pygmy
- San
- Yoruba

Europe
- Adygei
- Basque
- French
- Italian
- Orcadian
- Russian
- Sardinian
- Tuscan

Middle East
- Bedouin
- Druze
- Mozabite
- Palestinian

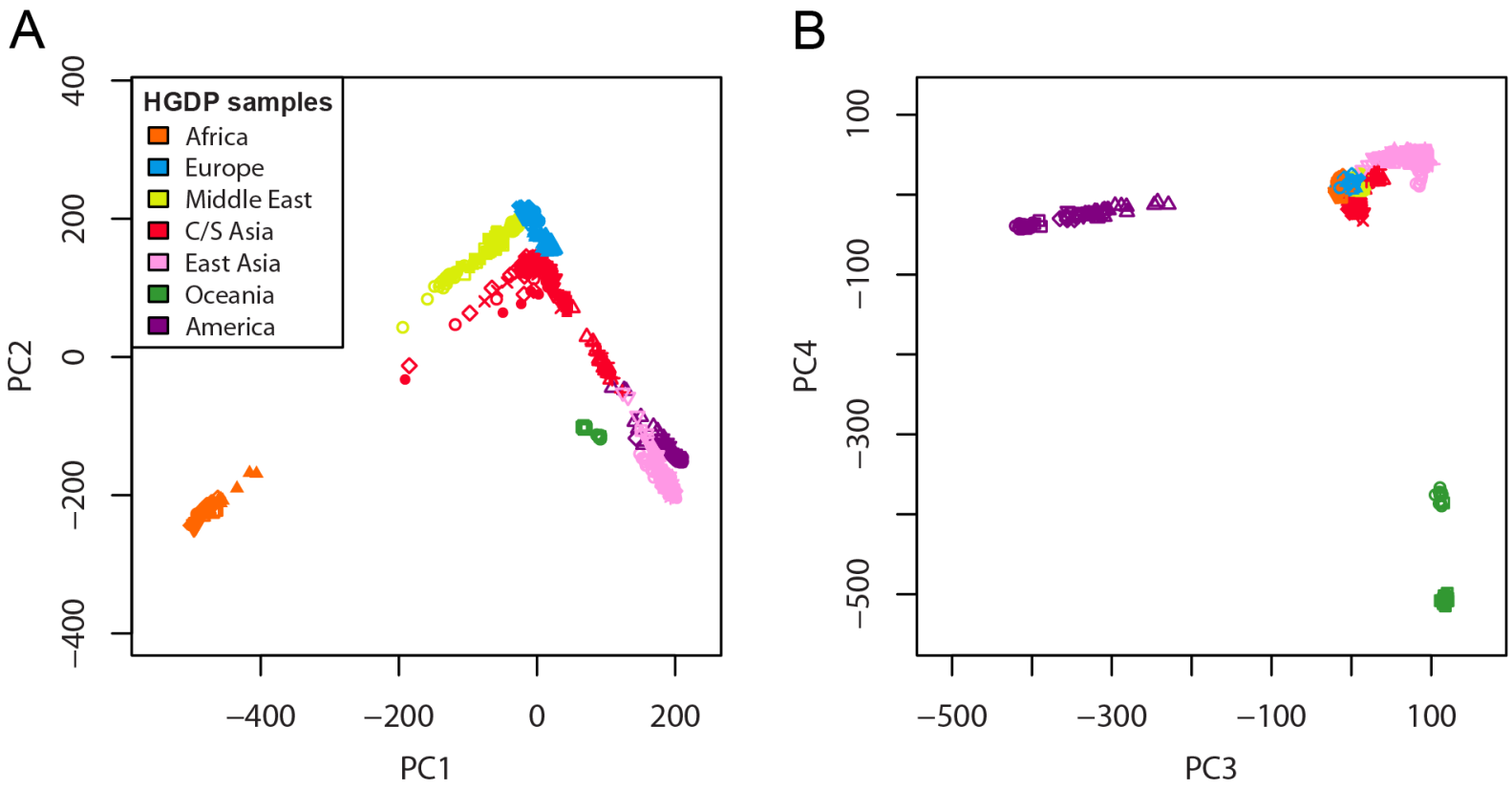
C/S Asia
- Balochi
- Brahui
- Burusho
- Hazara
- Kalash
- Makrani
- Pathan
- Sindhi
- Uygur

East Asia
- Cambodian
- Dai
- Daur
- Han
- Han (N. China)
- Hezhen
- Japanese
- Lahu
- Miao
- Mongola
- Naxi
- Oroqen
- She
- Tu
- Tujia
- Xibo
- Yakut
- Yi

Oceania
- Melanesian
- Papuan

America
- Colombian
- Karitiana
- Maya
- Pima
- Surui

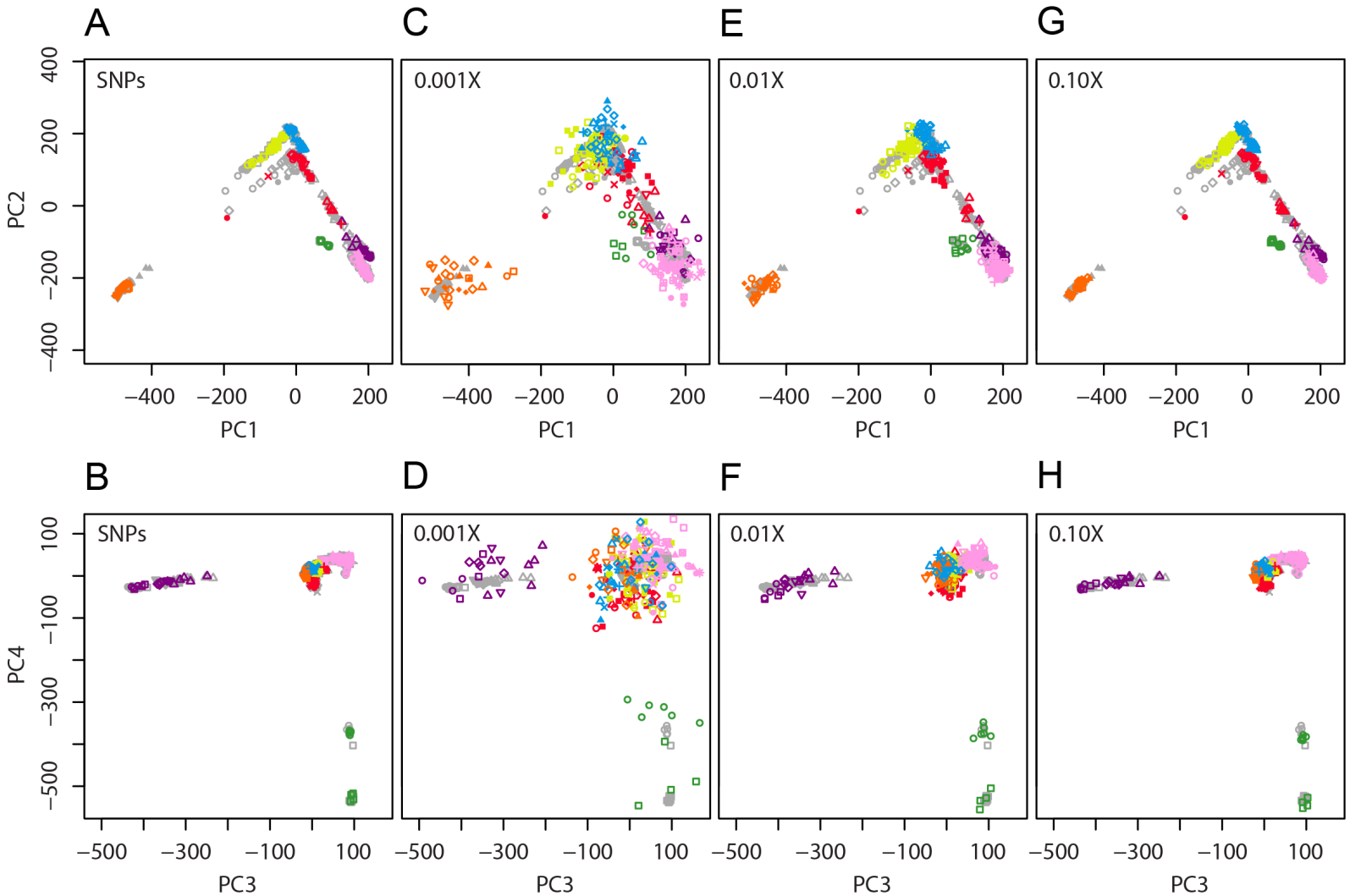
PCA on the HGDP data



Simulations based on HGDP

HGDP: 938 individuals at 632,958 autosomal SNP loci

Test set: 238 individuals; **Reference set:** 700 individuals



Simulations based on HGDP

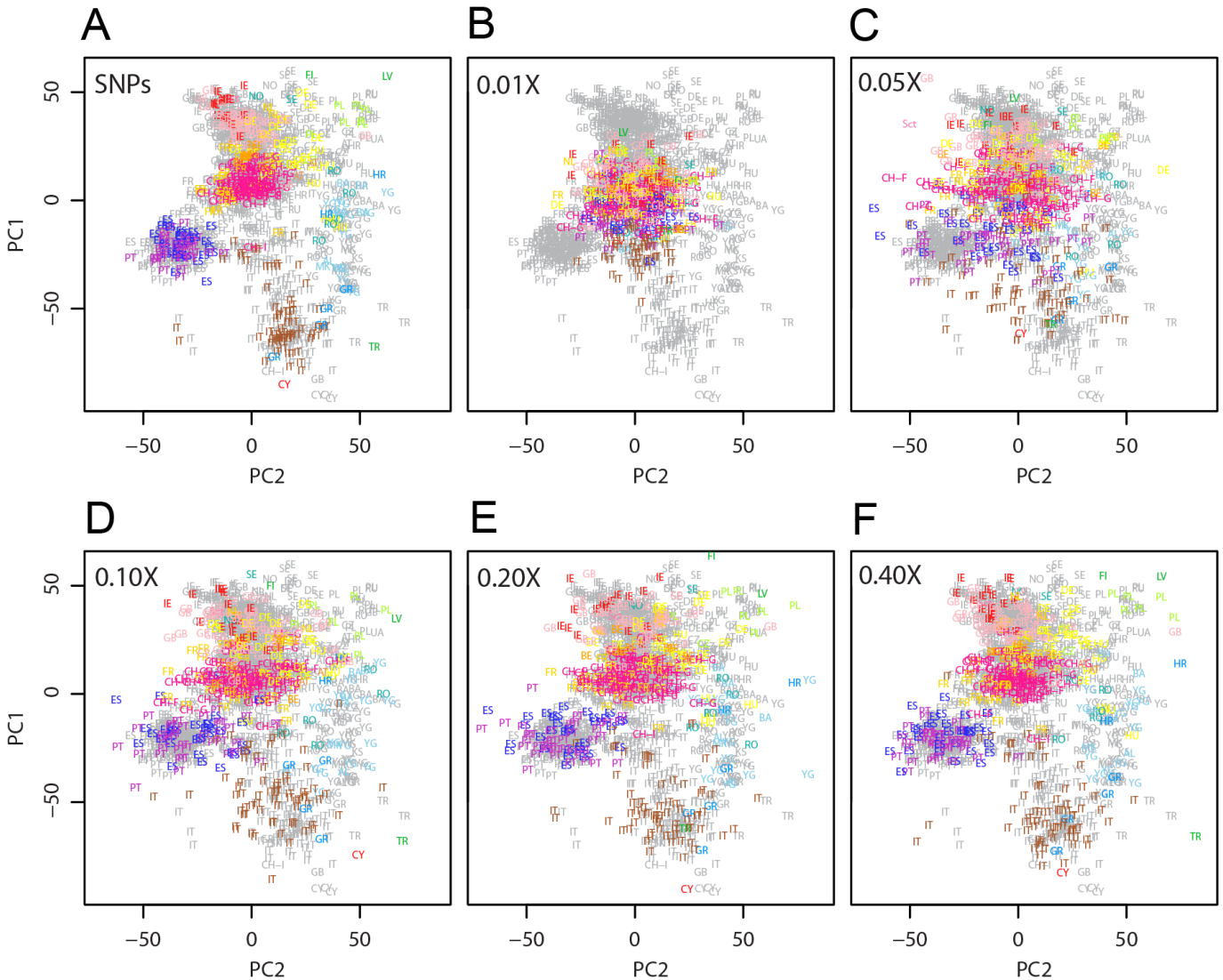
Sequence-based coordinates vs. SNP-based coordinates

Simulated mean coverage	Expected number of loci with reads	Sequence-based coordinates vs. SNP-based coordinates				
		Pearson correlation of PC1	Pearson correlation of PC2	Pearson correlation of PC3	Pearson correlation of PC4	Procrustes similarity t_0
0.25	140,010	0.9998	0.9998	0.9996	0.9994	0.9997
0.20	114,736	0.9998	0.9998	0.9996	0.9993	0.9996
0.15	88,166	0.9997	0.9998	0.9994	0.9989	0.9995
0.10	60,234	0.9996	0.9996	0.9991	0.9987	0.9993
0.05	30,870	0.9994	0.9993	0.9982	0.9973	0.9989
0.01	6,298	0.9974	0.9966	0.9909	0.9857	0.9949
0.008	5,043	0.9970	0.9960	0.9891	0.9830	0.9940
0.006	3,786	0.9948	0.9941	0.9834	0.9791	0.9911
0.004	2,527	0.9947	0.9941	0.9765	0.9668	0.9887
0.002	1,265	0.9877	0.9852	0.9468	0.9141	0.9729
0.001	633	0.9750	0.9689	0.9138	0.8600	0.9508

Simulations based on POPRES

POPRES: 1,385 individuals at 318,682 autosomal SNP loci

Test set: 385 individuals; **Reference set:** 1,000 individuals



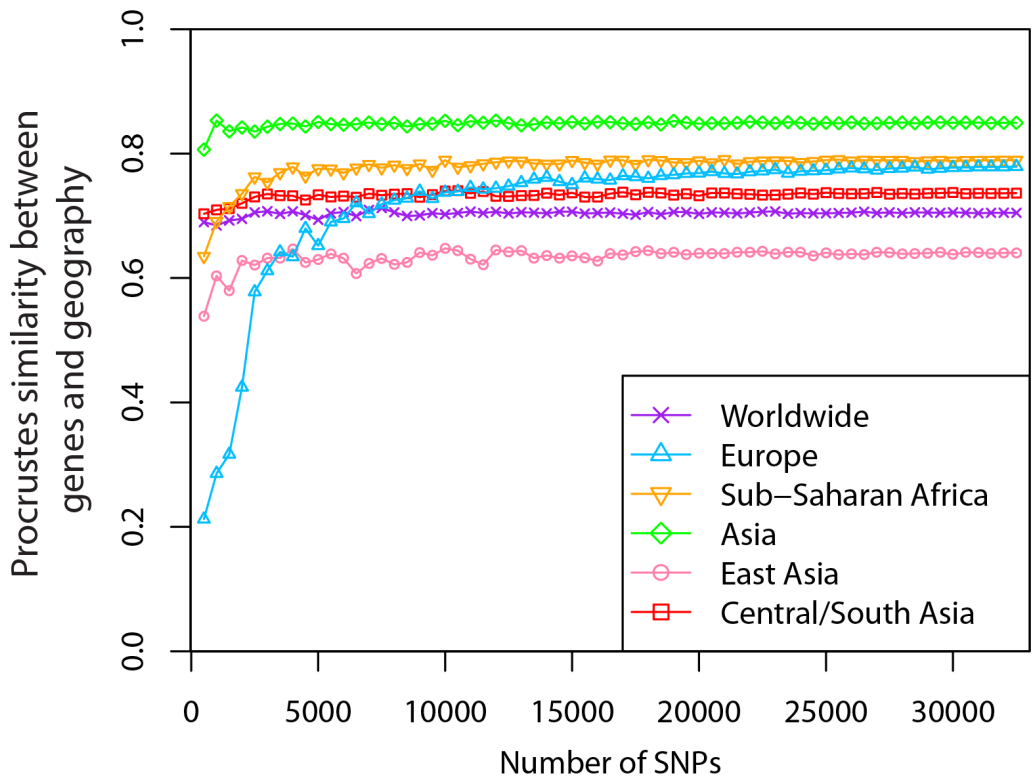
Simulations based on POPRES

Sequence-based coordinates vs. SNP-based coordinates

Simulated mean coverage	Expected number of loci with reads	Sequence-based coordinates vs. SNP-based coordinates		
		Pearson correlation of PC1	Pearson correlation of PC2	Procrustes similarity t_0
0.40	105,063	0.9927	0.9528	0.9764
0.35	94,111	0.9933	0.9458	0.9737
0.30	82,597	0.9906	0.9341	0.9671
0.25	70,492	0.9898	0.9241	0.9636
0.20	57,767	0.9868	0.8929	0.9495
0.15	44,390	0.9825	0.8811	0.9428
0.10	30,327	0.9752	0.8153	0.9126
0.05	15,542	0.9408	0.5016	0.7720
0.01	3,171	0.7541	0.1041	0.4786

Homogeneous samples need more markers

- Homogeneous samples need more markers to reveal their geographic structure of genetic variation.
- Europe is the most homogeneous continental group.

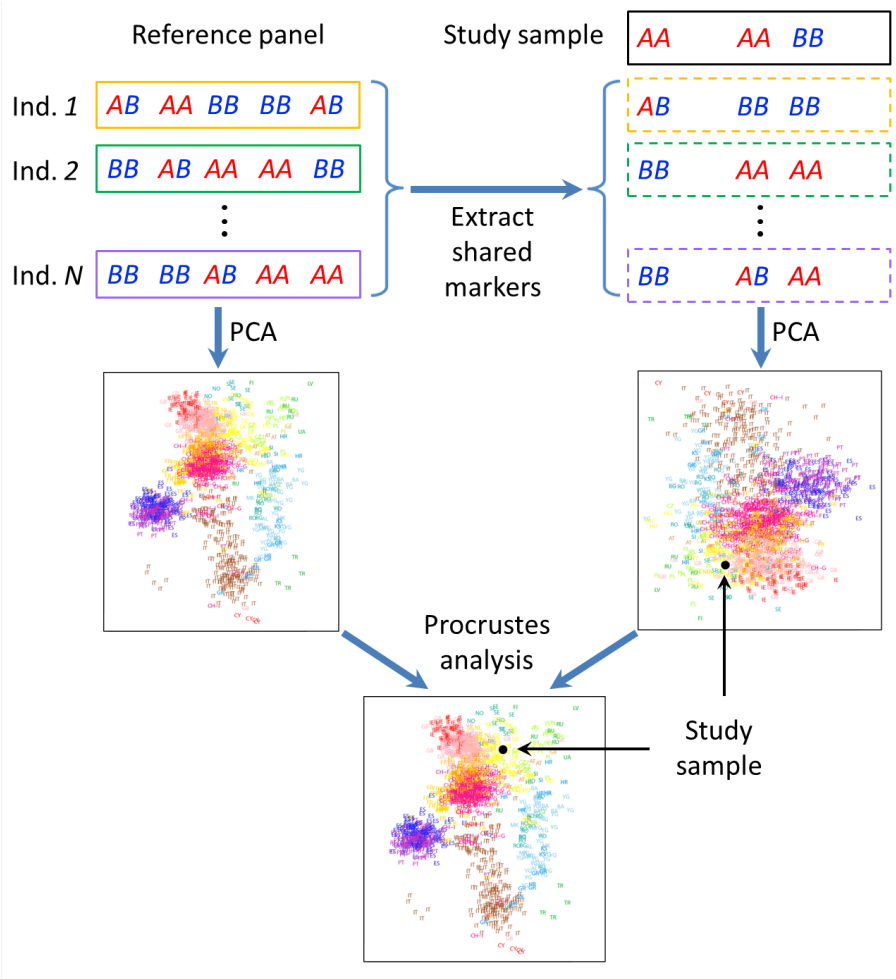


Region	F_{ST} (%)
World	9.704
Europe	0.212
Africa	1.334
Asia	4.706
E. Asia	1.874
C.S. Asia	2.140

Wang *et al.* (2012) *PLoS Genet.*

Estimate ancestry from SNP genotypes

Same framework for analyzing genotype data:

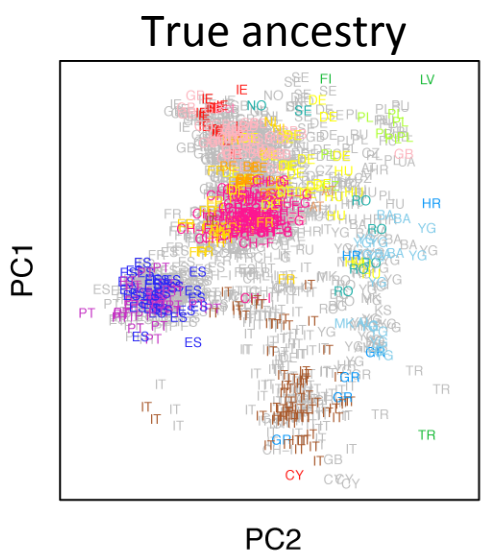


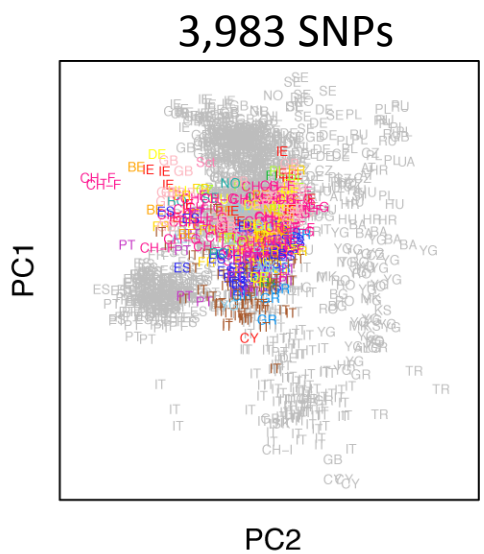
Motivations:

- Sequence reads might not be available
  - Array-genotyping data
- Joint analysis of sequencing and array-genotyping data
- Computational time scales linearly with sample size
  - PCA scales cubically
- Robust to family structure within the sample
- Can handle large amounts of missing data
  - Ancient DNA samples (Skoglund *et al.* 2012, *Science*)

Challenges

- Small number of overlapped markers
  - POPRES European reference panel: ~319K SNPs after QC
  - ExomeChip array: ~273K SNPs by design
  - Shared by POPRES and ExomeChip: 3,983 SNPs

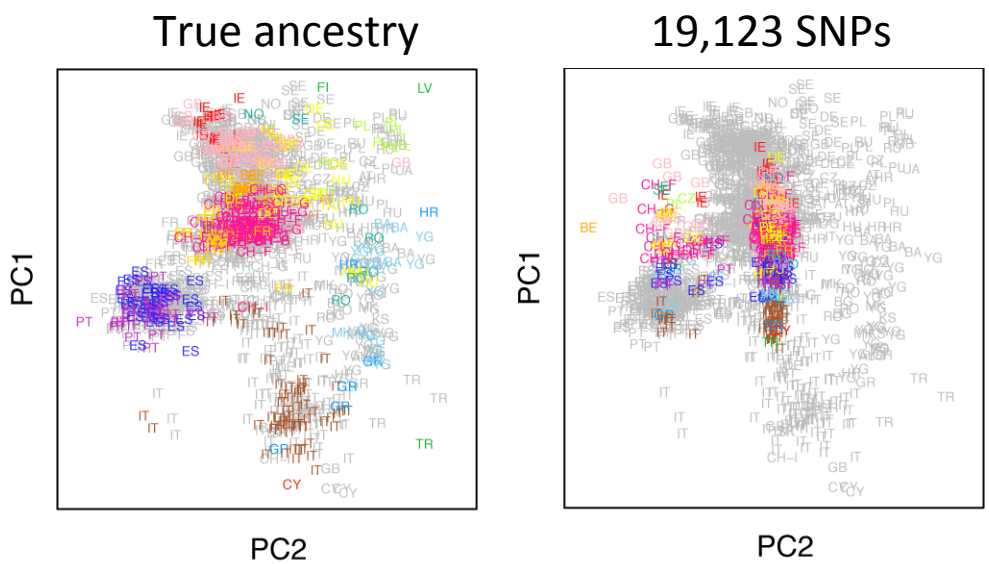




- Too expensive to whole-genome sequence a large reference sample

Impute the reference panel

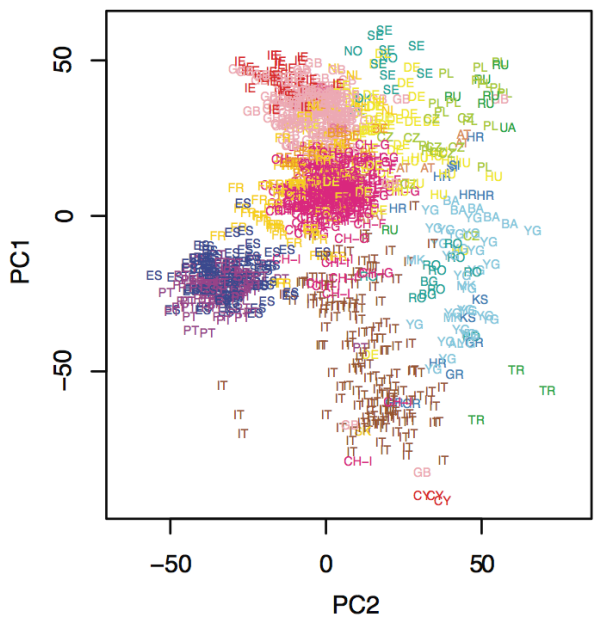
- Use 1000 Genomes data to impute POPRES
  - Imputed POPRES: 4.2 million SNPs after QC
  - Overlapped with the ExomeChip: 19,123 SNPs



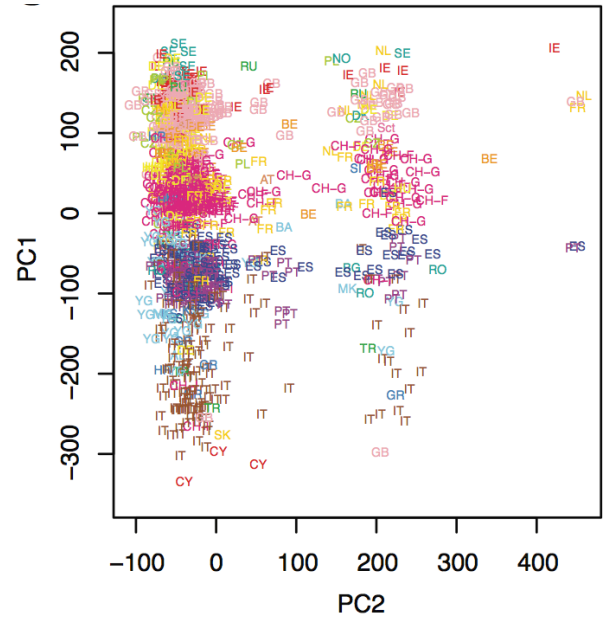
- PC1 reflects the north-south population structure.
- PC2 reflects some imputation artifacts.
- The east-west population structure is likely captured by higher order PCs.

Imputation causes artifacts in PCA

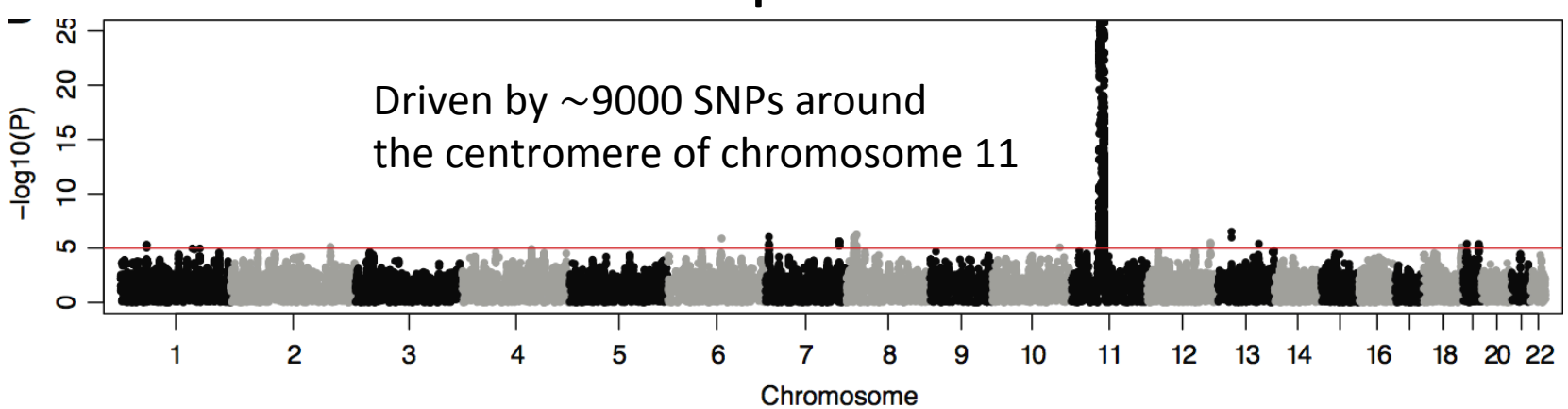
PCA on the **original** POPRES data



PCA on the **imputed** POPRES data

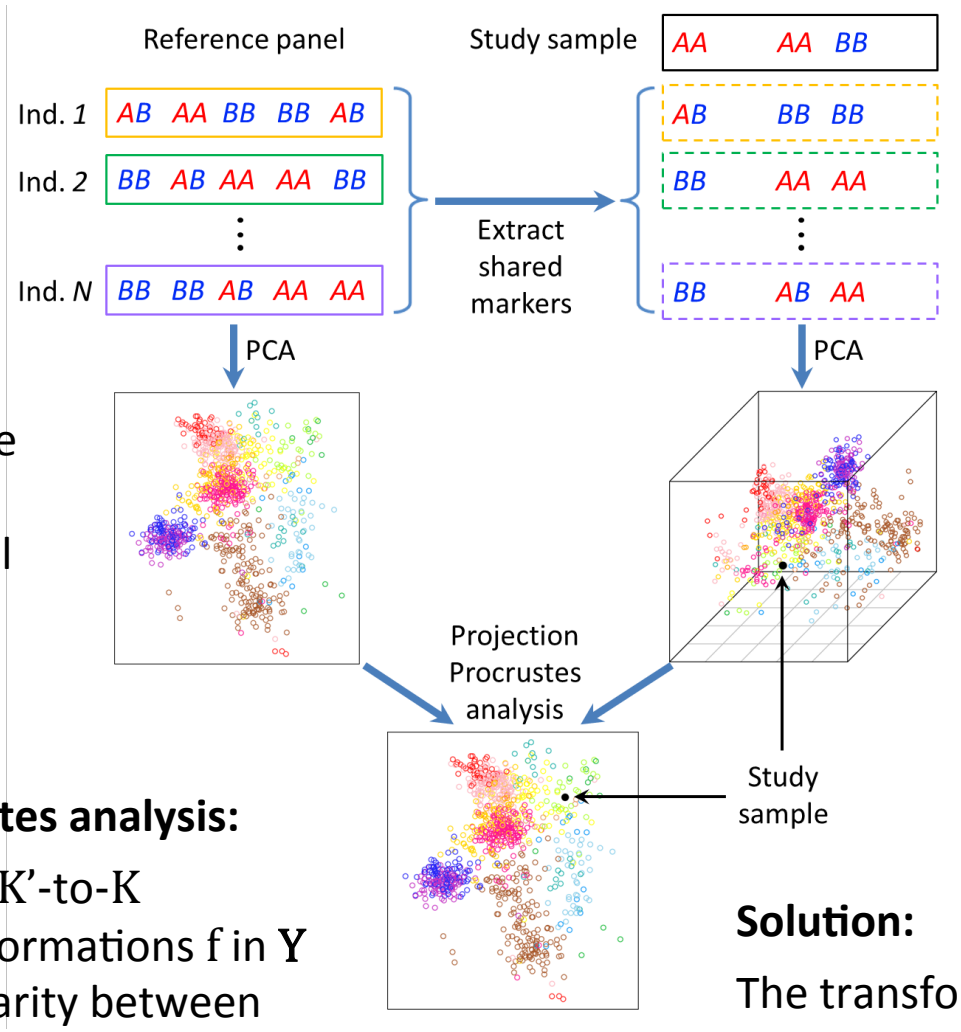


Association with PC2 of the imputed POPRES data



Project from high-dimensional PC space

K-dimensional PCA map **X** based on the genotyped SNPs of the reference panel



K'-dimensional PCA map **Y** based on the SNPs shared by the imputed reference panel and the study sample (K'≥ K)

Projection Procrustes analysis:

Search for a set of K'-to-K dimensional transformations f in **Y** such that the similarity between f(**Y**) and **X** is maximized.

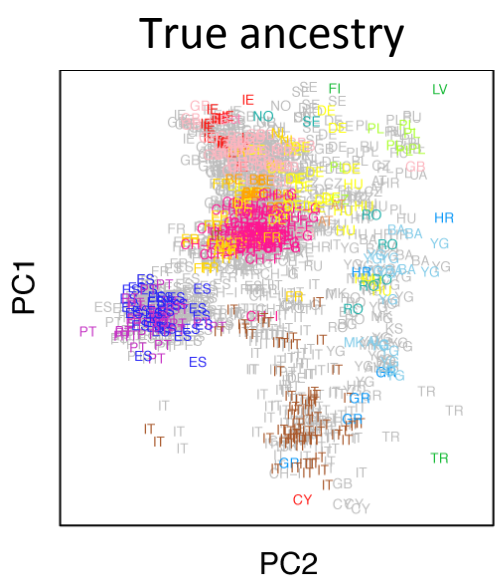
- Projection, rotation, reflection, translation, scaling

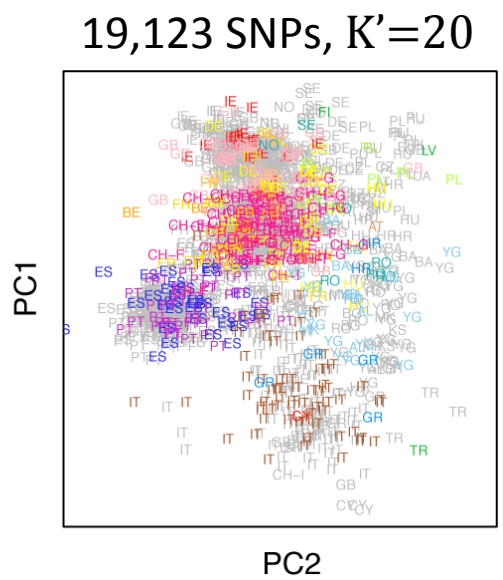
Solution:

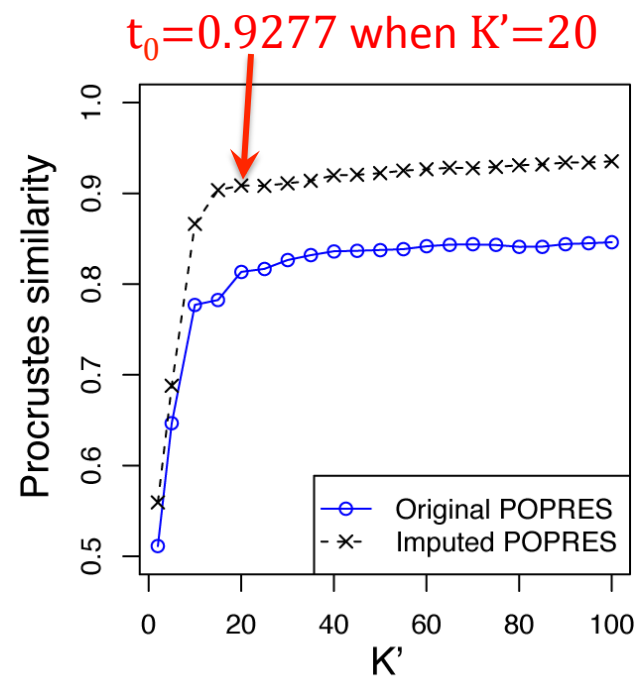
The transformation f does not have close form solution, but can be numerically solved using an iterative algorithm.

Project from high-dimensional PC space

- Combining imputation and high-dimensional projection can substantially improve the ancestry estimation!





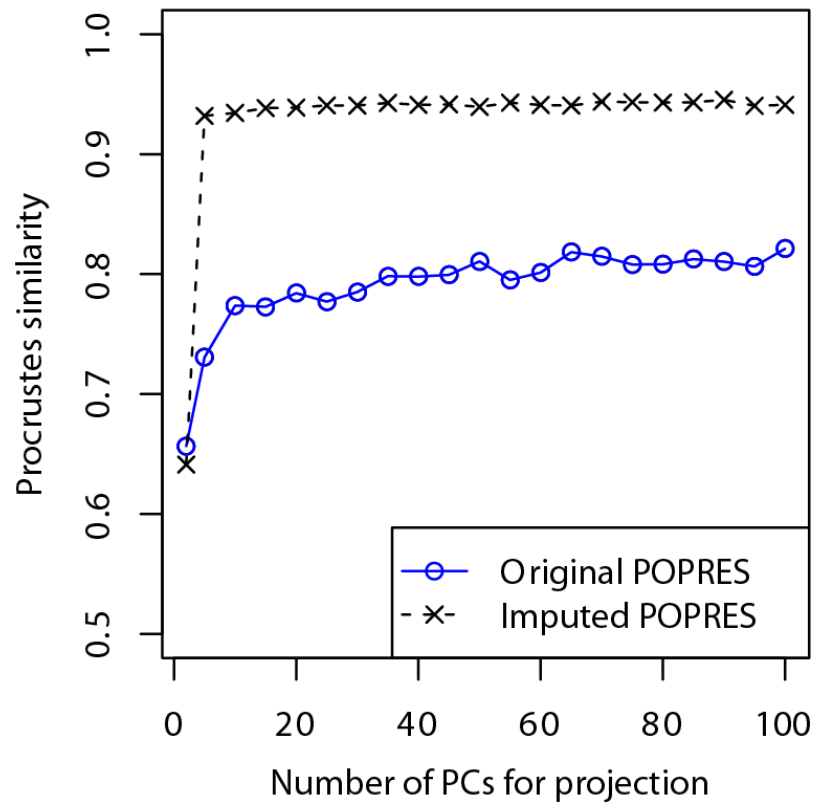


LASER 2.0

- Same strategies can be used to improve ancestry estimation from off-target sequence reads.

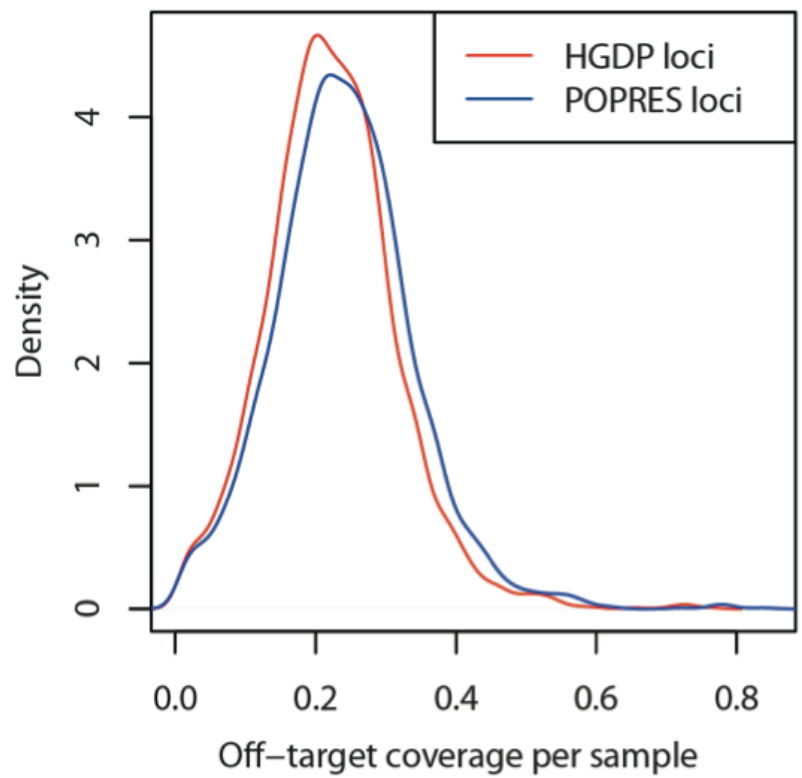
Simulation:

1. Take off-target coverage patterns from the Exome Sequencing Project, and down sample to 5% of the original coverage (~0.05X on average).
2. Simulate sequence reads based on the genotypes of 385 POPRES Europeans.

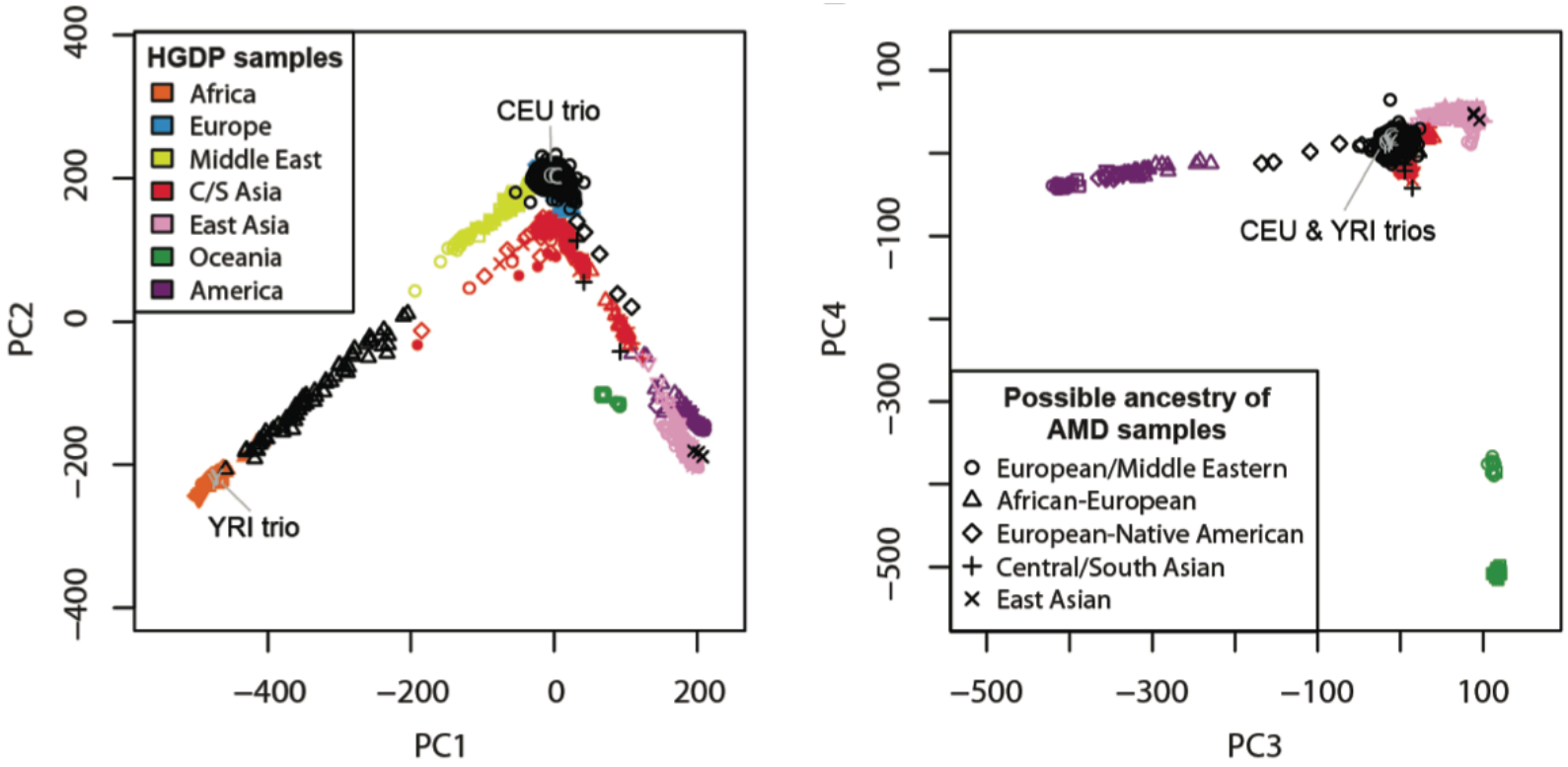


Application to the AMD study

- **Targeted sequencing of 10 AMD risk loci**
  - Sequenced at 127X across 0.97Mb targeted region
  - The off-target region is covered at ~0.2X on average
  - Sequenced 2,348 cases and 789 controls



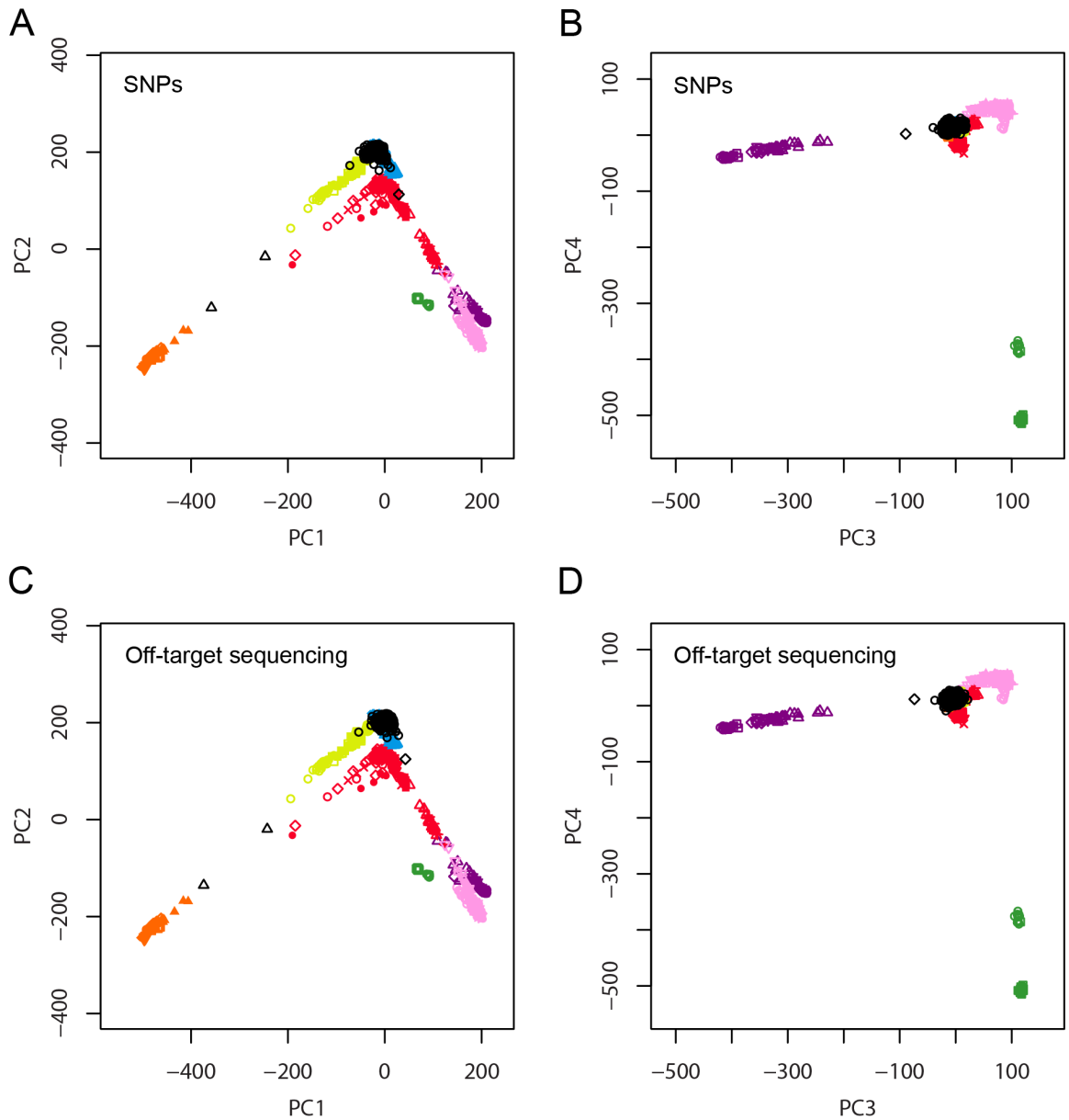
AMD samples on HGDP reference panel



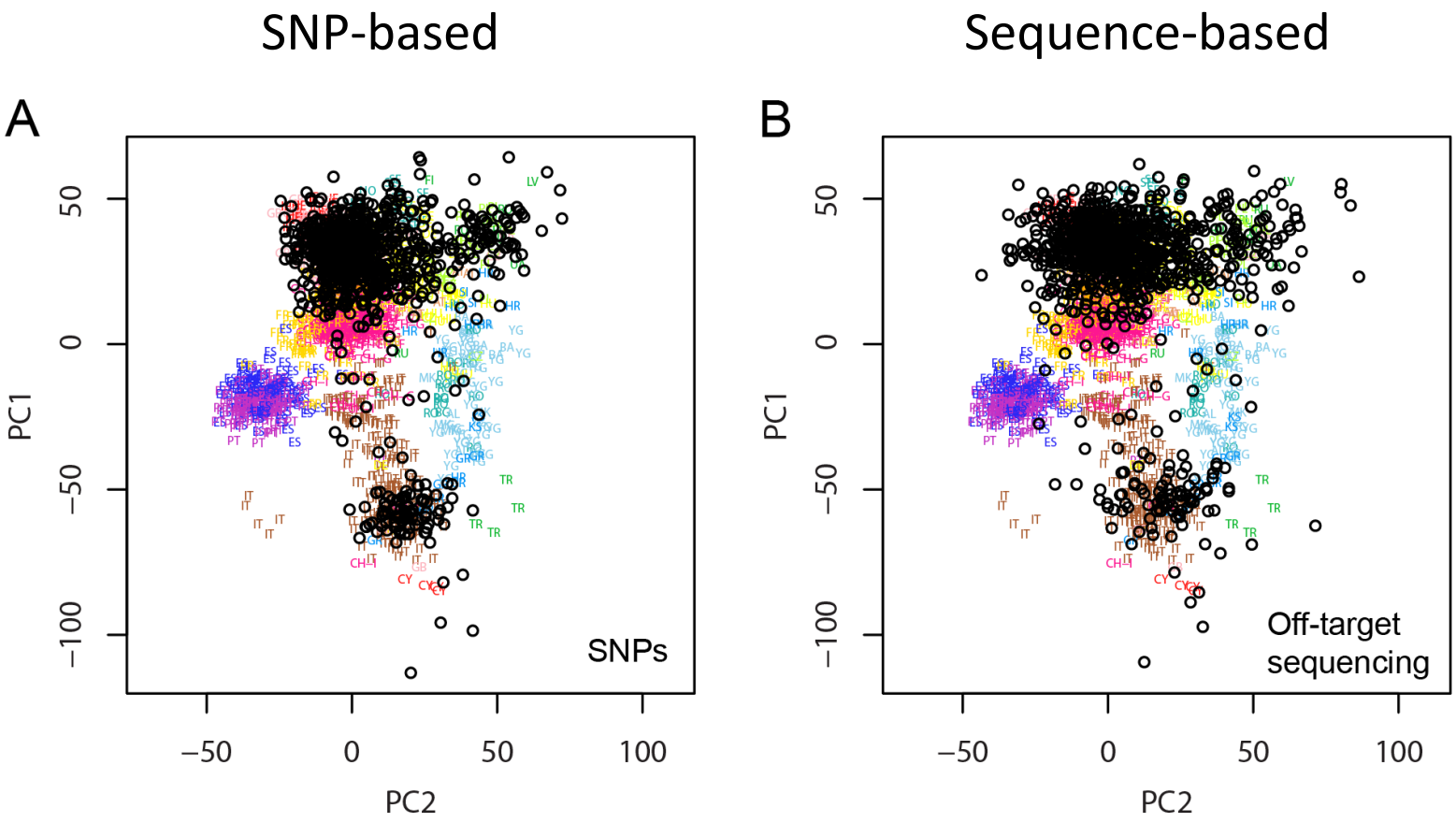
Genotype vs. sequence for 931 AMD samples

SNP-based coordinates

Sequence-based coordinates



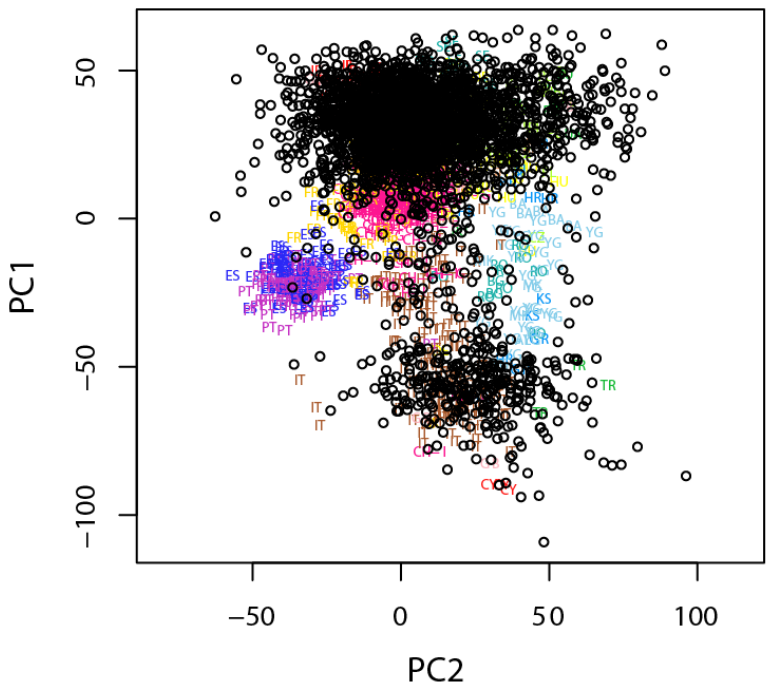
SNP vs. sequencing for 928 AMD samples



LASER 1.0 vs. LASER 2.0

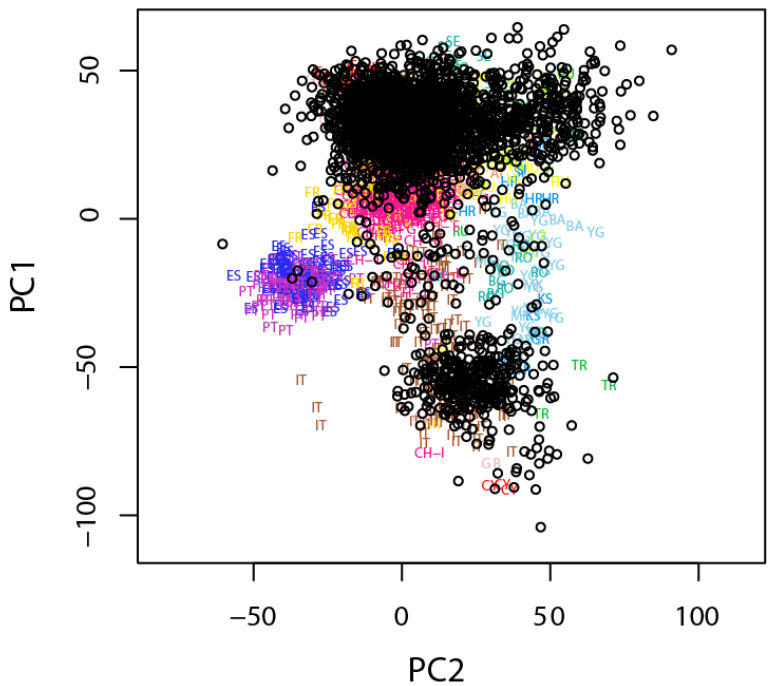
3,066 AMD sample with European ancestry

LASER 1.0
K′=2, original POPRES reference



t_0=0.9171

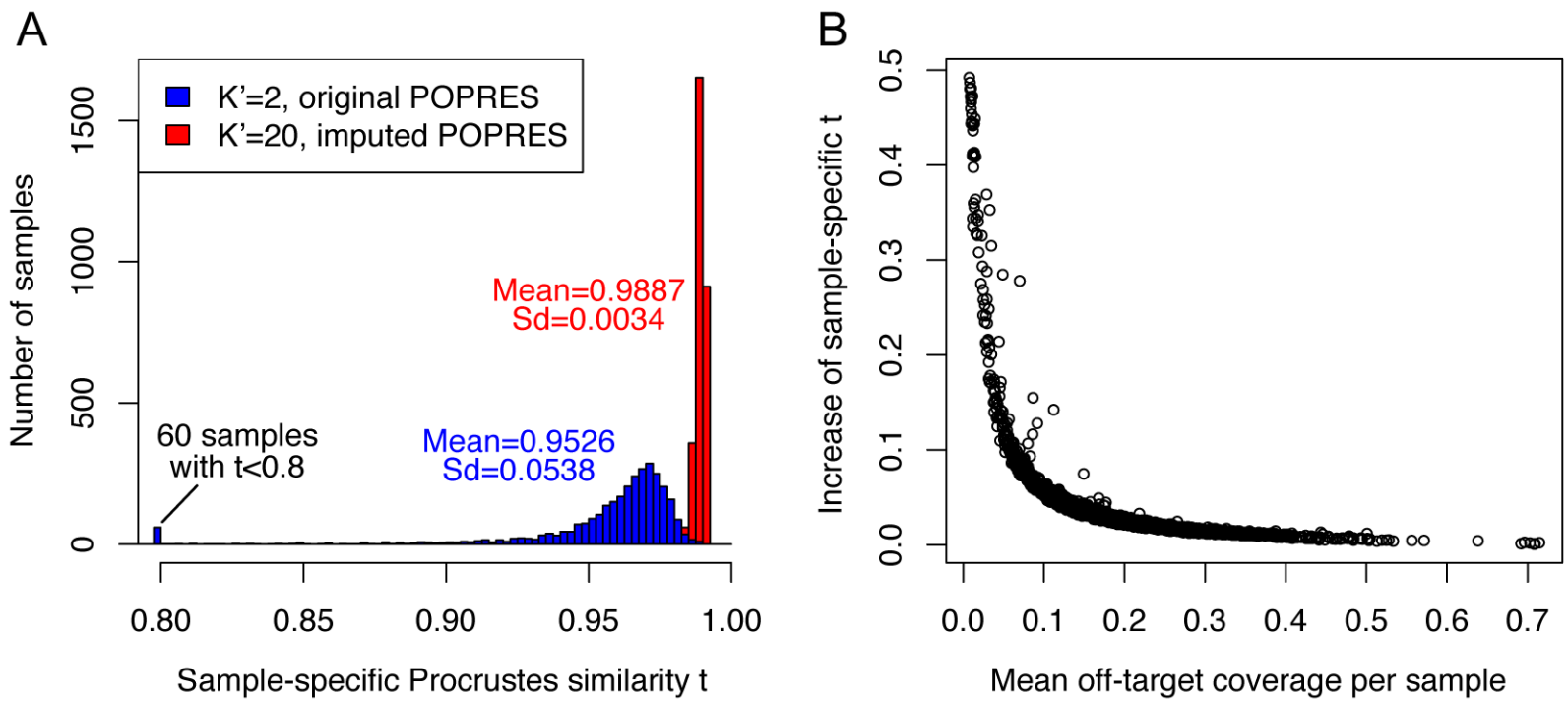
LASER 2.0
K′=20, imputed POPRES reference



t_0=0.9628

LASER 1.0 vs. LASER 2.0

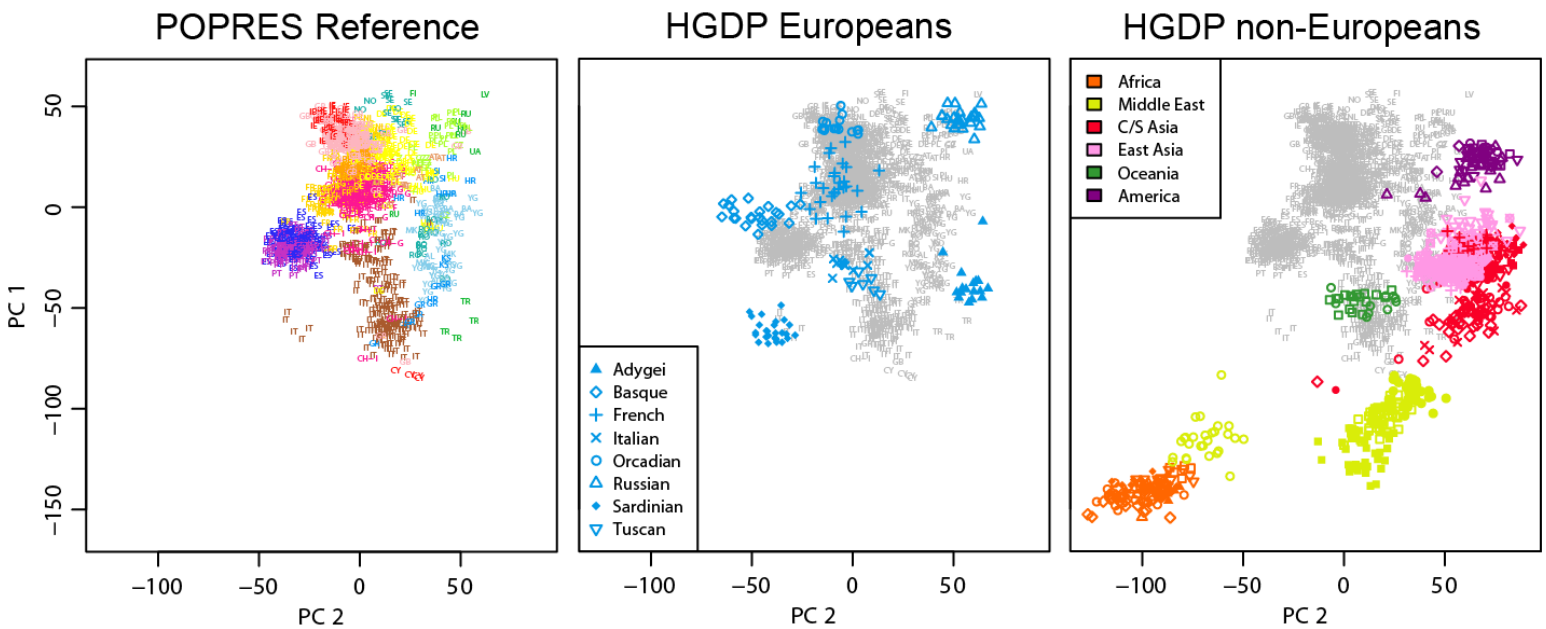
Sample-specific Procrustes similarity score (partially) reflects estimation accuracy of each sample.



Ancestry estimation improves for all samples, especially for samples that have extremely low coverage off-target data

Be cautious in choosing reference panel

- **Limitation:** results might be difficult to interpret when the reference panel does not include relevant ancestry groups.



- **Recommendation:** start with a worldwide reference panel and gradually narrow down to fine-scale regional panels.

Potential applications

- Control of population stratification in genetic association studies.
  - Regression on ancestry coordinates
  - Match ancestry background of study samples
  - Facilitate modeling of the relationship between phenotypes and ancestry coordinates

- Can apply to study sequencing data of ancient DNA samples, which often have abundant missing data.
  - Skoglund *et al.* (2012, *Science*) investigated the genetic relationship of four ancient DNA samples in Europe with modern humans using a similar approach.

Application to the AMD study

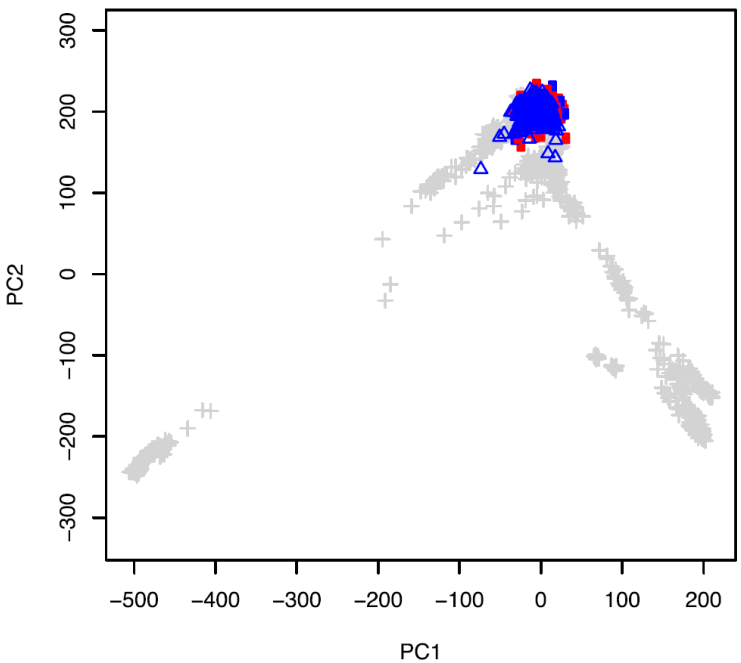
- **Targeted sequencing of 10 AMD risk loci (0.97Mb)**
  - To search for additional high-risk (rare) variants that can provide information about function
  - Sequenced 2,348 cases and 789 controls
    - Known high-risk variant R1210C in CFH gene has $P=2.6x10^{-3}$
    - Not enough sample size for studying rare variants.

- **Expanding our experiment**
  - Identify additional ancestry-matched controls from public resources to augment our sample size
  - Plan
    - Place AMD samples in the worldwide ancestry map
    - Place other sequenced samples in the same map
    - Identify matched controls for all cases

Matching results

- Search for matches from >6,800 samples in the Exome Sequencing Project
- Build matched set
  - 2,268 AMD cases
  - 2,268 matched controls
  - Focused on sites with >10X depth
  - Exclude sites near indels
  - 430 protein changing variants in both ESP and AMD experiments
- R1210C variant now has $P=2.9\times10^{-6}$ (initial $P=2.6\times10^{-3}$)
- A new rare variant K155Q in the C3 gene: $P=2.7\times10^{-4}$ (initial $P=6.3\times10^{-3}$), OR=2.68

Red: AMD cases
Blue: matched controls



Zhan *et al.* (2013) *Nature Genetics*

Validation of the K155Q variant

Table 2 Follow-up genotyping summary and meta-analysis summary

Sample set	Controls		Cases		*P* value
	N	MAF	*N*	MAF	
Discovery sample					
Sequenced samples (*N* = 4,536)	2,268	0.004	2,268	0.011	2.7×10^{-4}
Follow-up samples					
Germany: University of Regensburg (*N* = 2,976)	1,147	0.006	1,829	0.016	1.7×10^{-3}
United States: Vanderbilt/Miami (*N* = 1,819)	726	0.004	1,093	0.007	3.5×10^{-1}
Netherlands: Rotterdam Study (*N* = 1,409)	1,280	0.005	129	0.031	1.5×10^{-4}
UK: Cambridge AMD Study (*N* = 1,279)	423	0.006	856	0.015	6.2×10^{-2}
United States: University of California, Los Angeles/University of Pittsburgh (*N* = 830)	211	0.004	619	0.017	8.3×10^{-4}
deCODE study					
deCODE discovery sample (*N* = 52,578)	51,435	0.005	1,143	–[a]	1.1×10^{-7}
Meta-analysis					
All follow-up samples (*N* = 8,313)	3,787	0.005	4,526	0.013	7.7×10^{-7}
Discovery and all follow-up samples (*N* = 12,849)	6,055	0.005	6,794	0.013	1.1×10^{-9}
Discovery, all follow-up and deCODE samples (*N* = 65,427)	57,490	0.005	7,937	–[a]	1.6×10^{-15}

Zhan *et al.* (2013) *Nature Genetics*

Summary

- A statistical framework to trace individual ancestry in a reference PCA space.
  - Accurate even with small amounts of sequence/genotype data
  - Robust to family structure and sampling distribution
  - Computationally efficient, $\sim O(n)$
  - Easy for parallel computation

- High-dimensional projection and genotype imputation can substantially improve the accuracy of our ancestry estimates.

- Software package and references:
  - **LASER**: http://www.sph.umich.edu/csg/chaolong/LASER/
  - **Version 1**: Wang *et al.* (2014), Control of population stratification for sequence-based association studies. *Nature Genetics*, 46:409-415.
  - **Version 2**: Wang *et al.,* Improved ancestry estimation for both genotyping and sequencing data using projection Procrustes analysis and genotype imputation. *AJHG* (under review).

Other resources: admixed samples

- SEQMIX
  - Improve estimation of local ancestry for admixed samples using low-coverage off-target sequence data


ARTICLE

Accurate Local-Ancestry Inference in Exome-Sequenced Admixed Individuals via Off-Target Sequence Reads

Youna Hu,[1,2,4,*] Cristen Willer,[3] Xiaowei Zhan,[2] Hyun Min Kang,[2] and Gonçalo R. Abecasis[2,*]

CrossMark

The American Journal of Human Genetics 93, 891–899, November 7, 2013 891