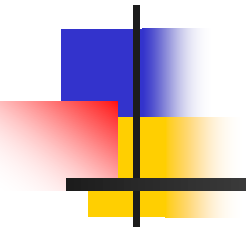


Dimensionality Reduction in the Analysis of Human Genetics Data

Petros Drineas

Purdue University
Department of Computer Science

Google drineas

PCA

~~Dimensionality Reduction~~ in the Analysis of Human ~~Genetics~~ Genotype Data

Petros Drineas

Purdue University
Department of Computer Science

Google drineas

We are quite similar, but we are different…

The average genome (~2x3 billion base pairs) contains:

- 4-5 million single nucleotide variations, compared to the reference sequence (Single Nucleotide Polymorphisms – SNPs)
- ~0.5 million small insertions or deletions 'indels' (1-100bp)
- ~5,000 larger insertions or deletions (>100bp)

Variation across all (~23,000) genes - the 'exome'

~18,000 variants

~8-9,000 functional variants

~95% of variants are common

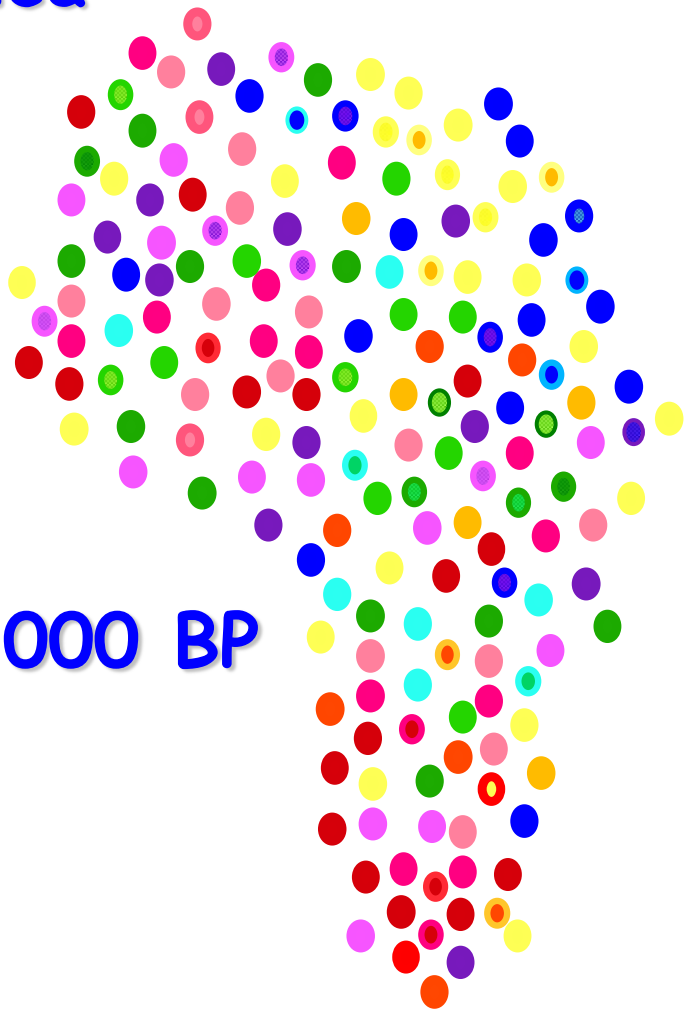
~500-1000 genes with new mutations

~100-200 knock-out mutations

Genetic variation is shaped by evolutionary forces

- Mutation
- Genetic drift
- Population structure (inbreeding, mating patterns)
- Gene flow and admixture
- Natural selection

Early *Homo sapiens sapiens* in Africa



150,000 to 100,000 BP

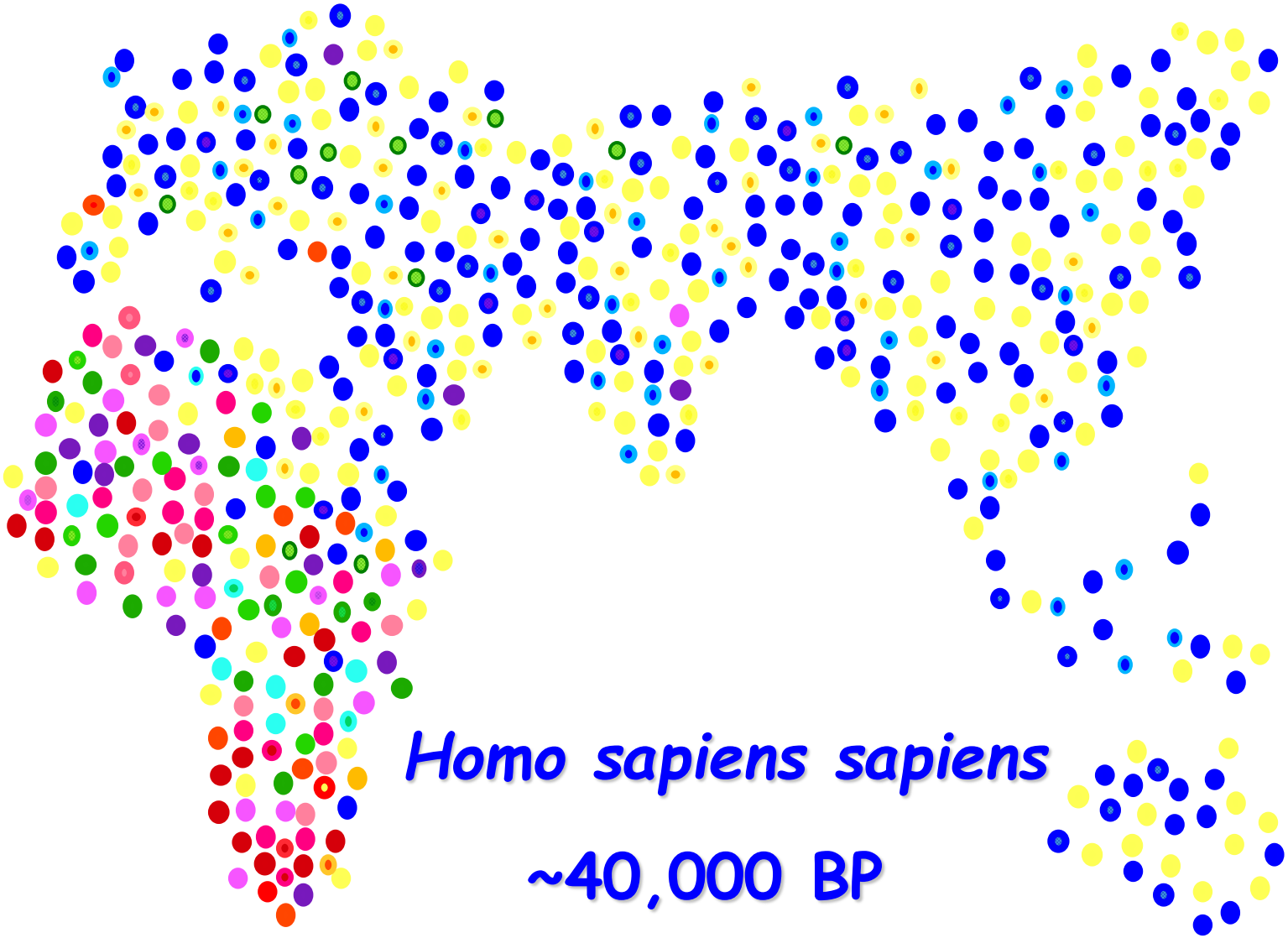
Kidd Lab, Yale University, http://info.med.yale.edu/genetics/kkidd/point.html

Homo sapiens sapiens

Colonizing south west Asia

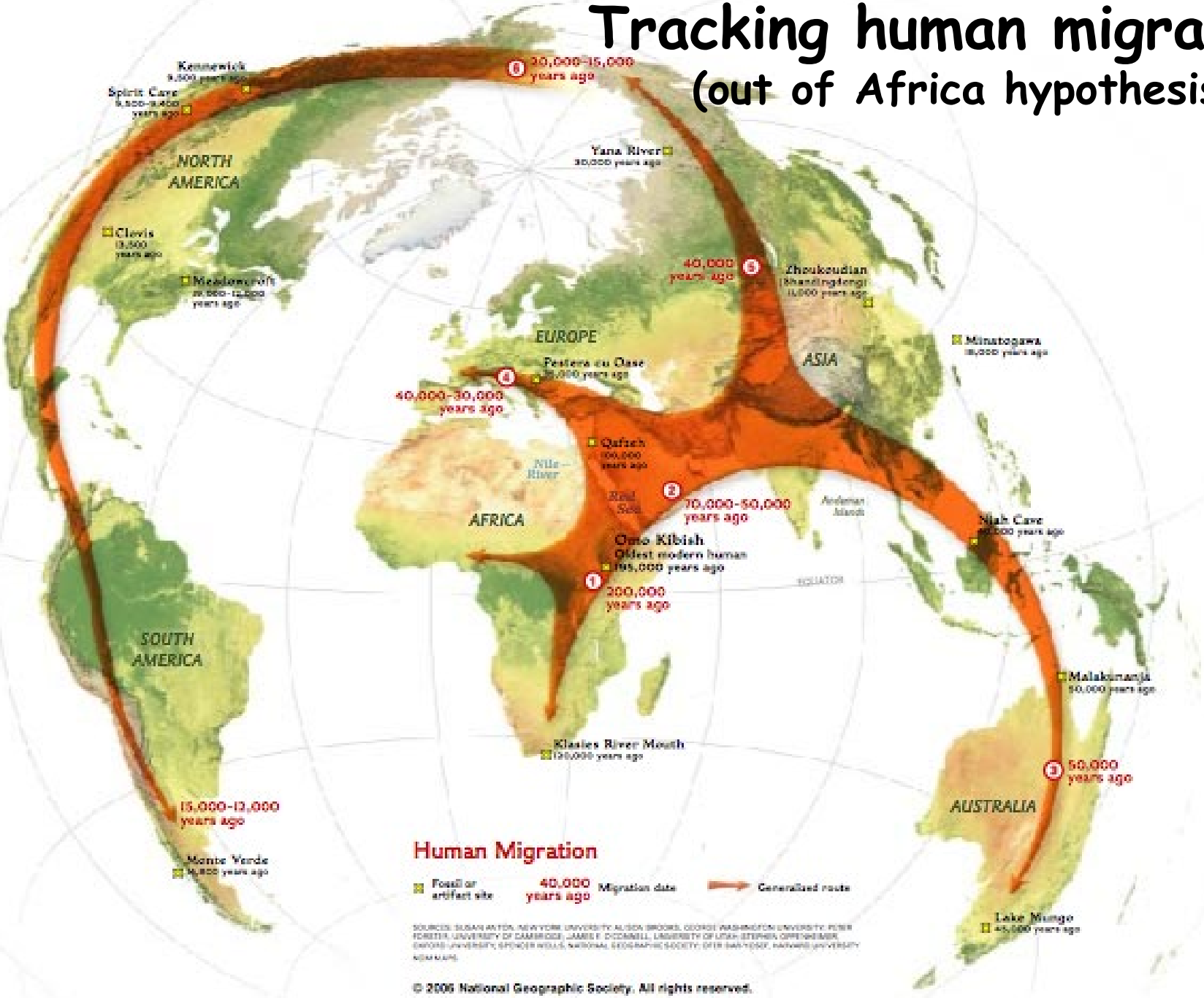
~100,000 BP

http://info.med.yale.edu/genetics/kkidd/point.html



http://info.med.yale.edu/genetics/kkidd/point.html

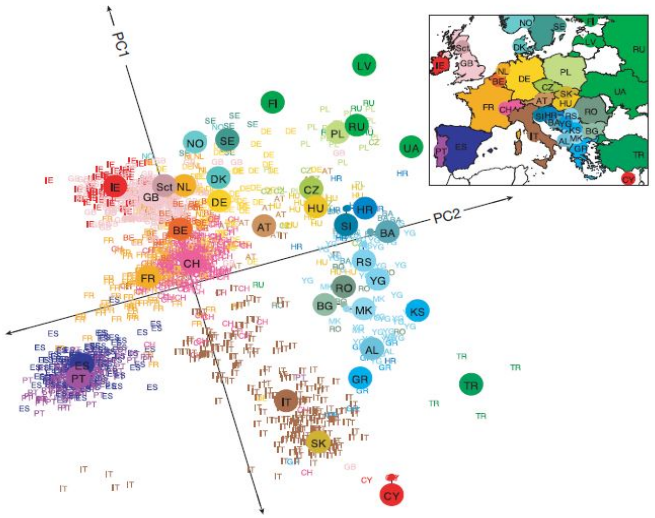
Tracking human migrations
(out of Africa hypothesis)



PCA for human genetic data analysis

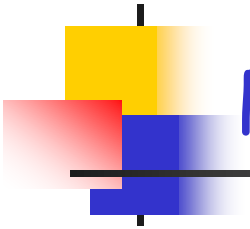
Fact:

Linear Dimensionality Reduction techniques (such as Principal Components Analysis – PCA) separate different populations and result in plots that correlate well with geography or *geo-demographics*.



$r^2 = 0.77$ for PC1 vs Latitude
$r^2 = 0.78$ for PC2 vs Longitude

Novembre et al. (Nature 2008)



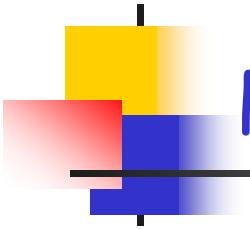
PCA for human genetic data analysis

The success of PCA in (human) genetics is remarkable!

- PCA has been around for over a century (Pearson 1901, Hotelling 1933).

- PCA in human genetics goes back to (at least) Menozzi, Piazza, & Cavalli-Sforza (Science 1978).

- Algorithms for PCA (meaning algorithms for SVD and eigendecompositions) have been a topic of intense research in numerical linear algebra and applied math for 70+ years.



PCA for human genetic data analysis

The success of PCA in (human) genetics is remarkable!

- PCA has been around for over a century (Pearson 1901, Hotelling 1933).

- PCA in human genetics goes back to (at least) Menozzi, Piazza, & Cavalli-Sforza (Science 1978).

- Algorithms for PCA (meaning algorithms for SVD and eigendecompositions) have been a topic of intense research in numerical linear algebra and applied math for 70+ years.

- PCA has been *very (?)* successful in many domains:

    - Imaging: remember Eigenfaces?
    - Document-term data: remember Latent Semantic Indexing (LSI)?
    - Web search: remember HIITS and pagerank?

    BUT the aforementioned domains have concluded that other (typically *very* non-linear) dimensionality reduction techniques are better in extracting structure in their respective **modern** datasets!

PCA for human genetic data analysis

Fact:

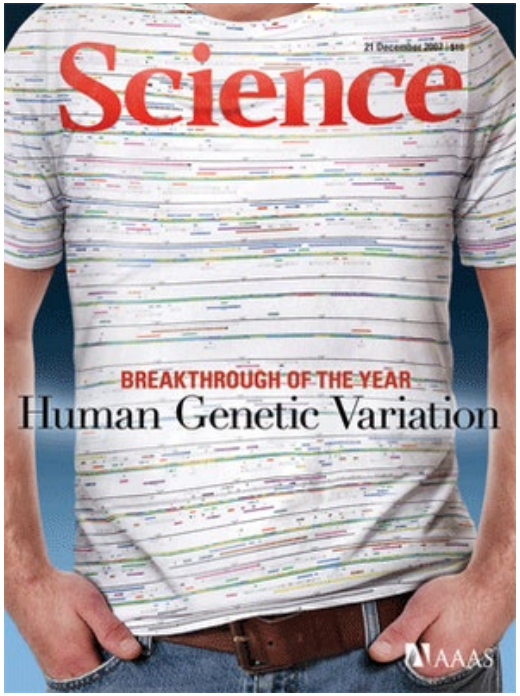
Linear Dimensionality Reduction techniques (such as Principal Components Analysis – PCA) separate different populations and result in plots that correlate well with geography or *geo-demographics*.

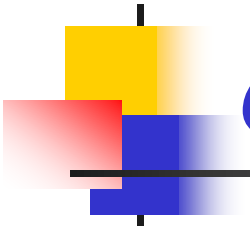
Leverage this observation:

While we invariably use many other statistical techniques and software tools to analyze human genetic data, PCA plots are ***always*** the starting point and they often "set the tone" for other analyses.

Why do we care about and population structure?

- Population genetics & histories of human populations
- Mapping causative genes for common complex disorders
  - Correcting stratification in Genome-Wide Association Studies (GWAS)
- Conservation studies
- Forensics
- Genealogy





Overview

- Scaling PCA to millions of samples/markers
- Selecting Ancestry Informative Markers (AIMs)
- PCA and Geodemographics

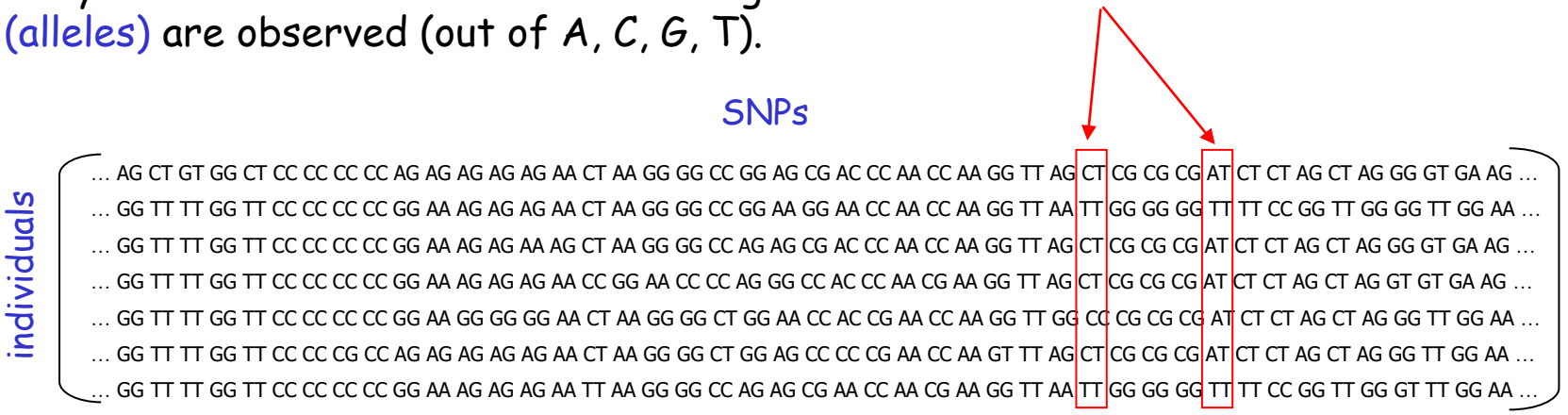
Mathematical apparatus:

- Subspace iteration vs. Krylov subspace methods to approximate principal components
- From the Singular Value Decomposition (SVD) to the CX decomposition, the Column Subset Selection Problem (CSSP), and beyond

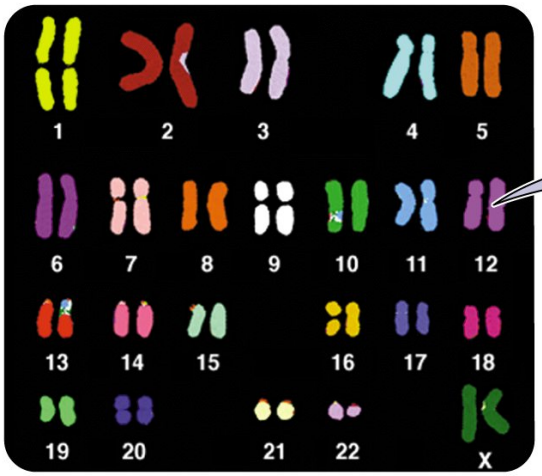
Single Nucleotide Polymorphisms (SNPs)

Single Nucleotide Polymorphisms: the most common type of genetic variation in the genome across different individuals.

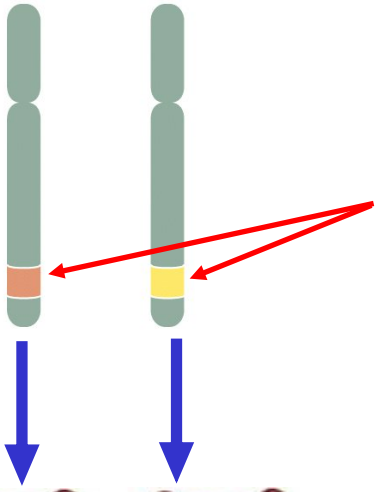
They are known locations at the human genome where two alternate nucleotide bases (alleles) are observed (out of A, C, G, T).



There are millions of SNPs in the human genome, so this matrix could have millions of columns.

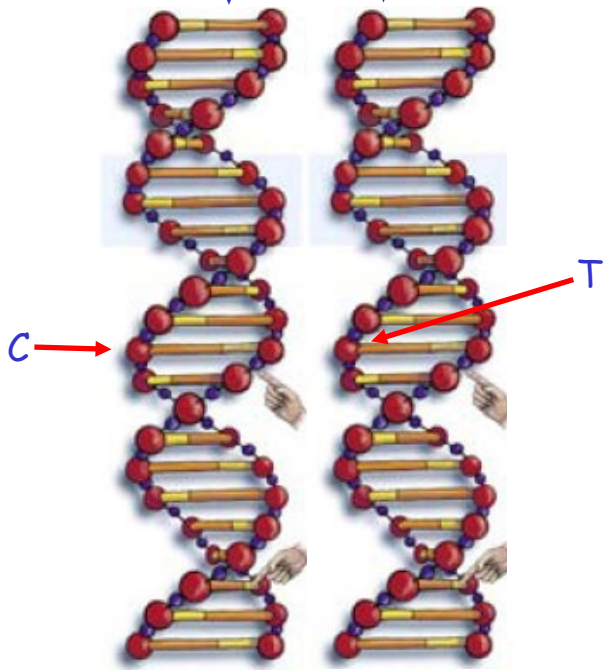


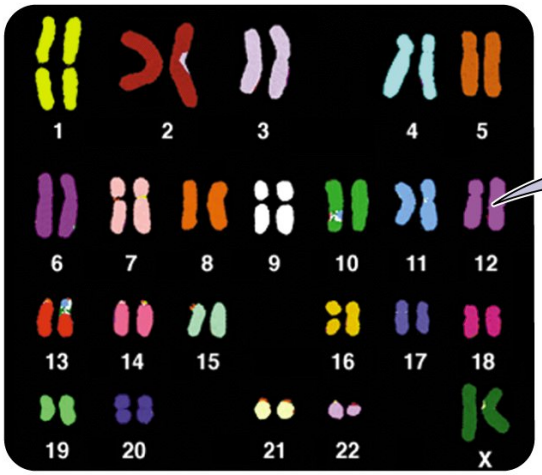
Two copies of a chromosome
(father, mother)

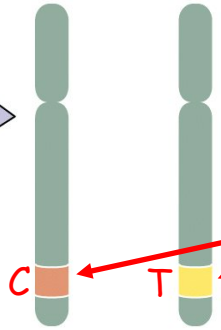


Focus at a specific locus and assay the observed nucleotide bases (alleles).

SNP: exactly two alternate alleles appear.







Two copies of a chromosome
(father, mother)

Focus at a specific locus and assay the observed alleles.

SNP: exactly two alternate alleles appear.

An individual could be:

- Heterozygotic (in our study, CT = TC)

SNPs

individuals

… AG CT GT GG CT CC CC CC CC AG AG AG AG AG AA CT AA GG GG CC GG AG CG AC CC AA CC AA GG TT AG CT CG CG CG AT CT CT AG CT AG GG GT GA AG …

… GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA CT AA GG GG CC GG AA GG AA CC AA CC AA GG TT AA TT GG GG GG TT TT CC GG TT GG GG TT GG AA …

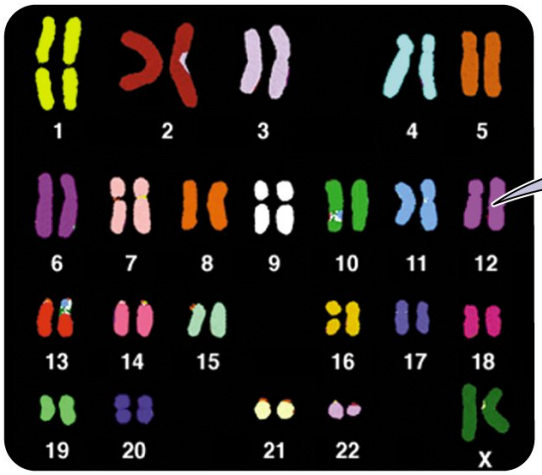
… GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA AG CT AA GG GG CC AG AG CG AC CC AA CC AA GG TT A**CT** G CG CG AT CT CT AG CT AG GG GT GA AG …

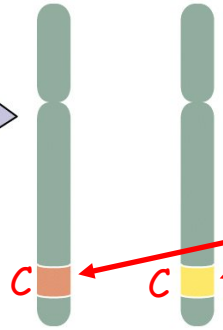
… GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA CC GG AA CC CC AG GG CC AC CC AA CG AA GG TT AG CT CG CG CG AT CT CT AG CT AG GT GT GA AG …

… GG TT TT GG TT CC CC CC CC GG AA GG GG GG AA CT AA GG GG CT GG AA CC AC CG AA CC AA GG TT GG CC CG CG CG AT CT CT AG CT AG GG TT GG AA …

… GG TT TT GG TT CC CC CG CC AG AG AG AG AG AA CT AA GG GG CT GG AG CC CC CG AA CC AA GT TT AG CT CG CG CG AT CT CT AG CT AG GG TT GG AA …

… GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA TT AA GG GG CC AG AG CG AA CC AA CG AA GG TT AA TT GG GG GG TT TT CC GG TT GG GT TT GG AA …





Two copies of a chromosome
(father, mother)

Focus at a specific locus and assay the observed alleles.

SNP: exactly two alternate alleles appear.

An individual could be:

- Heterozygotic (in our studies, CT = TC)
- Homozygotic at the first allele, e.g., C

SNPs

individuals

… AG CT GT GG CT CC CC CC CC AG AG AG AG AG AA CT AA GG GG CC GG AG CG AC CC AA CC AA GG TT AG CT CG CG CG AT CT CT AG CT AG GG GT GA AG …

… GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA CT AA GG GG CC GG AA GG AA CC AA CC AA GG TT AA TT GG GG GG TT TT CC GG TT GG GG TT GG AA …

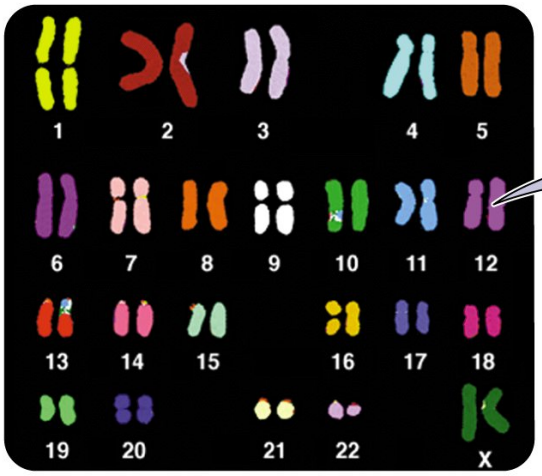
… GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA AG CT AA GG GG CC AG AG CG AC CC AA CC AA GG TT AG CT CG CG CG AT CT CT AG CT AG GG GT GA AG …

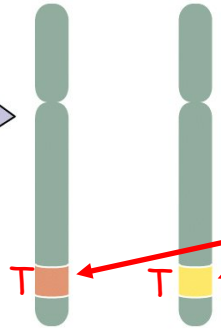
… GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA CC GG AA CC CC AG GG CC AC CC AA CG AA GG TT AG CT CG CG CG AT CT CT AG CT AG GT GT GA AG …

… GG TT TT GG TT CC CC CC CC GG AA GG GG GG AA CT AA GG GG CT GG AA CC AC CG AA CC AA GG TT G**CC**G CG CG AT CT CT AG CT AG GG TT GG AA …

… GG TT TT GG TT CC CC CG CC AG AG AG AG AG AA CT AA GG GG CT GG AG CC CC CG AA CC AA GT TT AG CT CG CG CG AT CT CT AG CT AG GG TT GG AA …

… GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA TT AA GG GG CC AG AG CG AA CC AA CG AA GG TT AA TT GG GG GG TT TT CC GG TT GG GT TT GG AA …





Two copies of a chromosome (father, mother)

Focus at a specific locus and assay the observed alleles.

SNP: exactly two alternate alleles appear.

An individual could be:

- Heterozygotic (in our studies, CT = TC) → Encode as 1
- Homozygotic at the first allele, e.g., C → Encode as 0
- Homozygotic at the second allele, e.g., T → Encode as 2

SNPs

individuals

… AG CT GT GG CT CC CC CC CC AG AG AG AG AA CT AA GG GG CC GG AG CG AC CC AA CC AA GG TT AG CT CG CG CG AT CT CT AG CT AG GG GT GA AG …

… GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA CT AA GG GG CC GG AA GG AA CC AA CC AA GG TT AA TT GG GG GG TT TT CC GG TT GG GG TT GG AA …

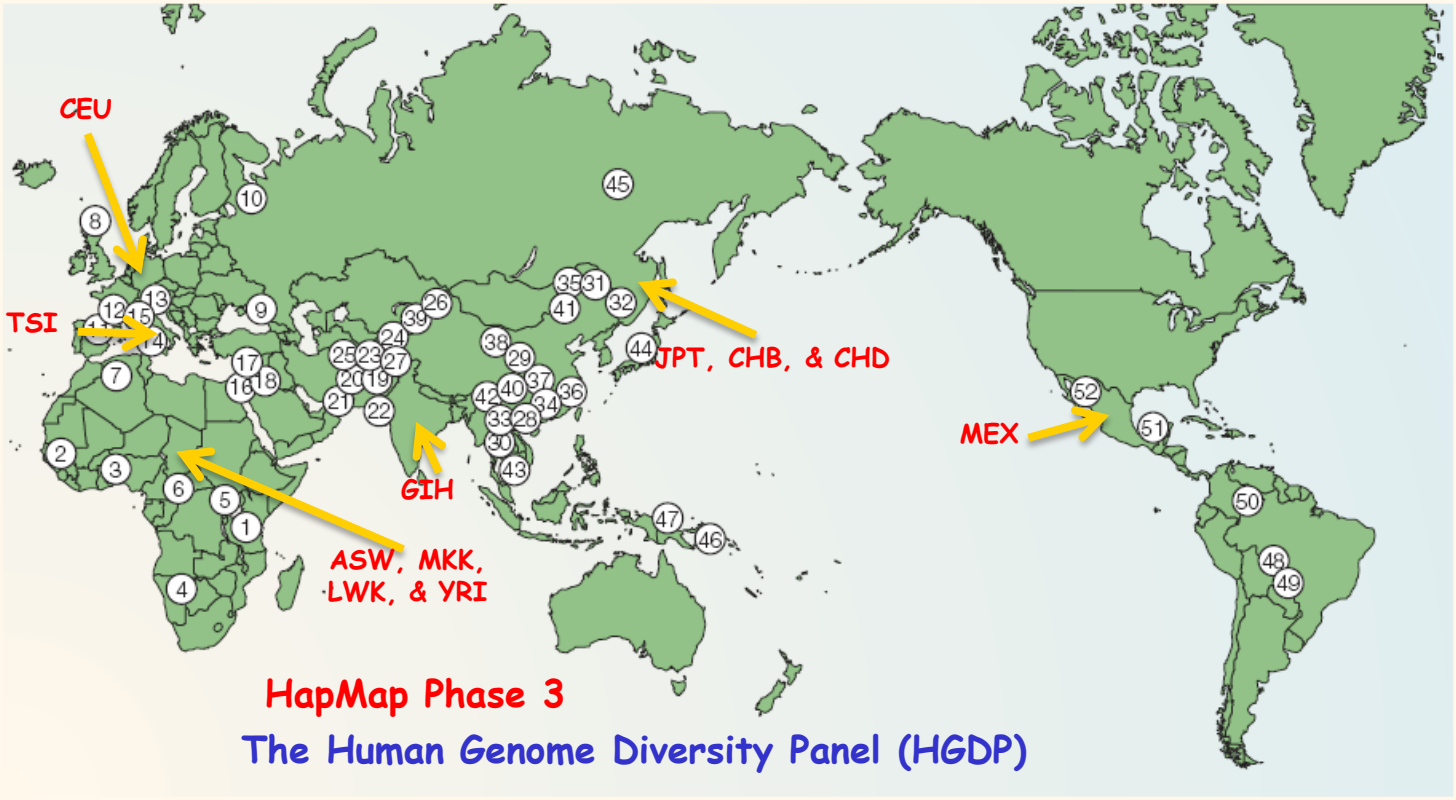
… GG TT TT GG TT CC CC CC CC GG AA AG AG AA AG CT AA GG GG CC AG AG CG AC CC AA CC AA GG TT AG CT CG CG CG AT CT CT AG CT AG GG GT GA AG …

… GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA CC GG AA CC CC AG GG CC AC CC AA CG AA GG TT AG CT CG CG CG AT CT CT AG CT AG GT GT GA AG …

… GG TT TT GG TT CC CC CC CC GG AA GG GG GG AA CT AA GG GG CT GG AA CC AC CG AA CC AA GG TT GG CC CG CG CG AT CT CT AG CT AG GG TT GG AA …

… GG TT TT GG TT CC CC CG CC AG AG AG AG AG AA CT AA GG GG CT GG AG CC CC CG AA CC AA GT TT AG CT CG CG CG AT CT CT AG CT AG GG TT GG AA …

… GG TT TT GG TT CC CC CC CC GG AA AG AG AG AA TT AA GG GG CC AG AG CG AA CC AA CG AA GG TT A**TT** G GG GG TT TT CC GG TT GG GT TT GG AA …



Africans
1 Bantu
2 Mandenka
3 Yoruba
4 San
5 Mbuti pygmy
6 Biaka
7 Mozabite

Europeans
8 Orcadian
9 Adygei
10 Russian
11 Basque
12 French
13 North Italian
14 Sardinian
15 Tuscan

Western Asians
16 Bedouin
17 Druze
18 Palestinian

Central and Southern Asians
19 Balochi
20 Brahui
21 Makrani
22 Sindhi
23 Pathan
24 Burusho
25 Hazara
26 Uygur
27 Kalash

Eastern Asians
28 Han (S. China)
29 Han (N. China)
30 Dai
31 Daur
32 Hezhen
33 Lahu
34 Miao
35 Oroqen
36 She
37 Tujia
38 Tu
39 Xibo
40 Yi
41 Mongola
42 Naxi
43 Cambodian
44 Japanese
45 Yakut

Oceanians
46 Melanesian
47 Papuan

Native Americans
48 Karitiana
49 Surui
50 Colombian
51 Maya
52 Pima

Cavalli-Sforza (2005) *Nat Genet Rev*

Rosenberg et al. (2002) *Science*

Li et al. (2008) *Science*

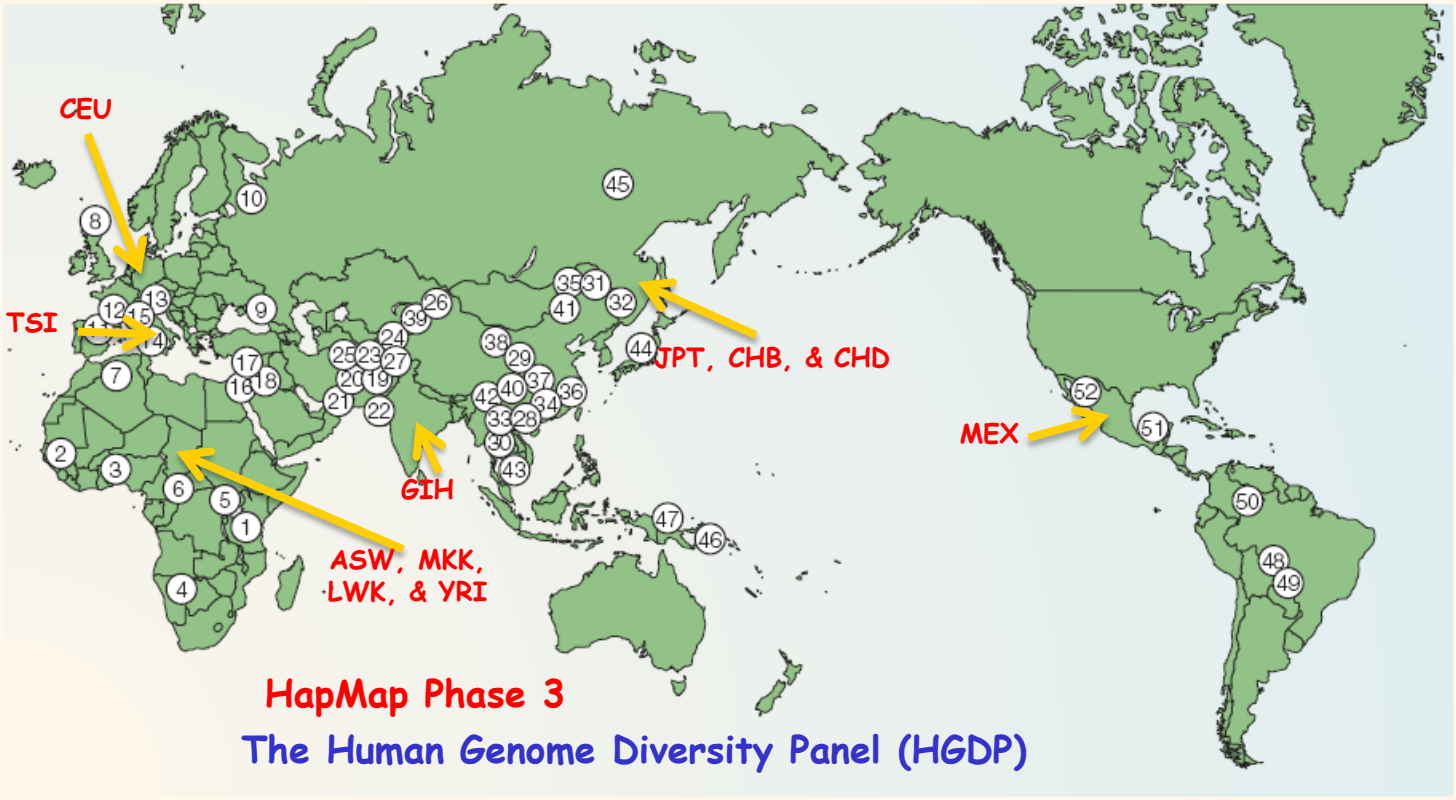
The International HapMap Consortium (2003, 2005, 2007) *Nature*

HGDP data

- 1,033 samples
- 7 geographic regions
- 52 populations

HapMap Phase 3 data

- 1,207 samples
- 11 populations



Africans
1 Bantu
2 Mandenka
3 Yoruba
4 San
5 Mbuti pygmy
6 Biaka
7 Mozabite

Europeans
8 Orcadian
9 Adygei
10 Russian
11 Basque
12 French
13 North Italian
14 Sardinian
15 Tuscan

Western Asians
16 Bedouin
17 Druze
18 Palestinian

Central and Southern Asians
19 Balochi
20 Brahui
21 Makrani
22 Sindhi
23 Pathan
24 Burusho
25 Hazara
26 Uygur
27 Kalash

Eastern Asians
28 Han (S. China)
29 Han (N. China)
30 Dai
31 Daur
32 Hezhen
33 Lahu
34 Miao
35 Oroqen
36 She
37 Tujia
38 Tu
39 Xibo
40 Yi
41 Mongola
42 Naxi
43 Cambodian
44 Japanese
45 Yakut

Oceanians
46 Melanesian
47 Papuan

Native Americans
48 Karitiana
49 Surui
50 Colombian
51 Maya
52 Pima

Cavalli-Sforza (2005) *Nat Genet Rev*

Rosenberg et al. (2002) *Science*

Li et al. (2008) *Science*

The International HapMap Consortium (2003, 2005, 2007) *Nature*

HGDP data

- 1,033 samples
- 7 geographic regions
- 52 populations

HapMap Phase 3 data

- 1,207 samples
- 11 populations

We will apply SVD/PCA on the (joint) HGDP and HapMap Phase 3 data.

Matrix dimensions:

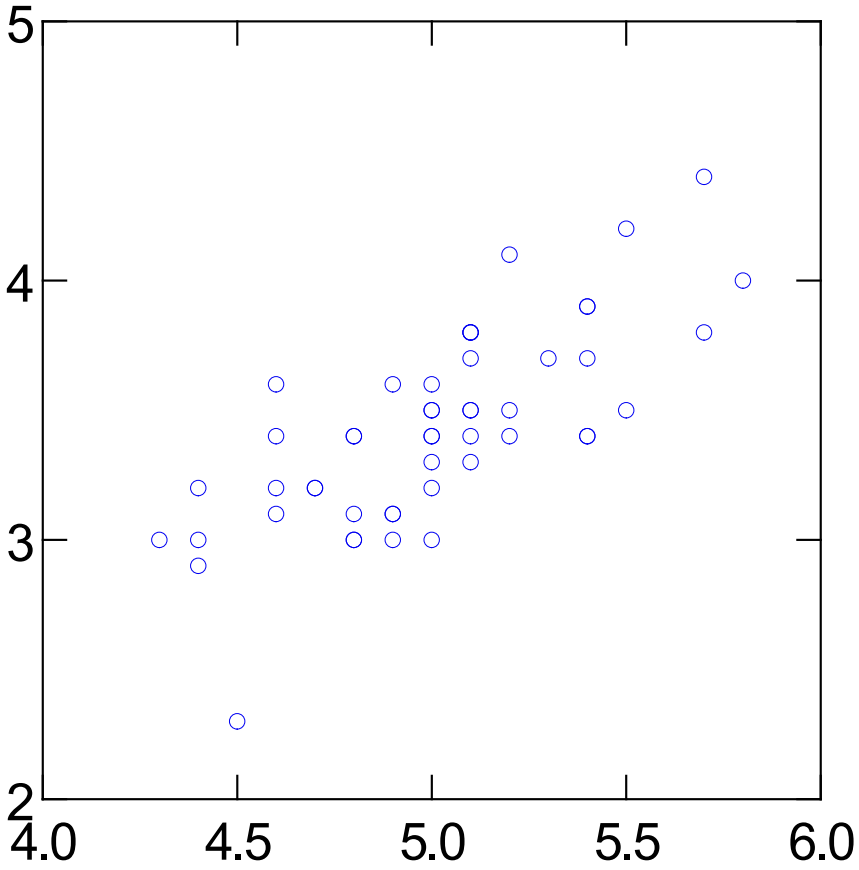
2,240 subjects (rows)

447,143 SNPs (columns)

Dense matrix:

over one billion entries

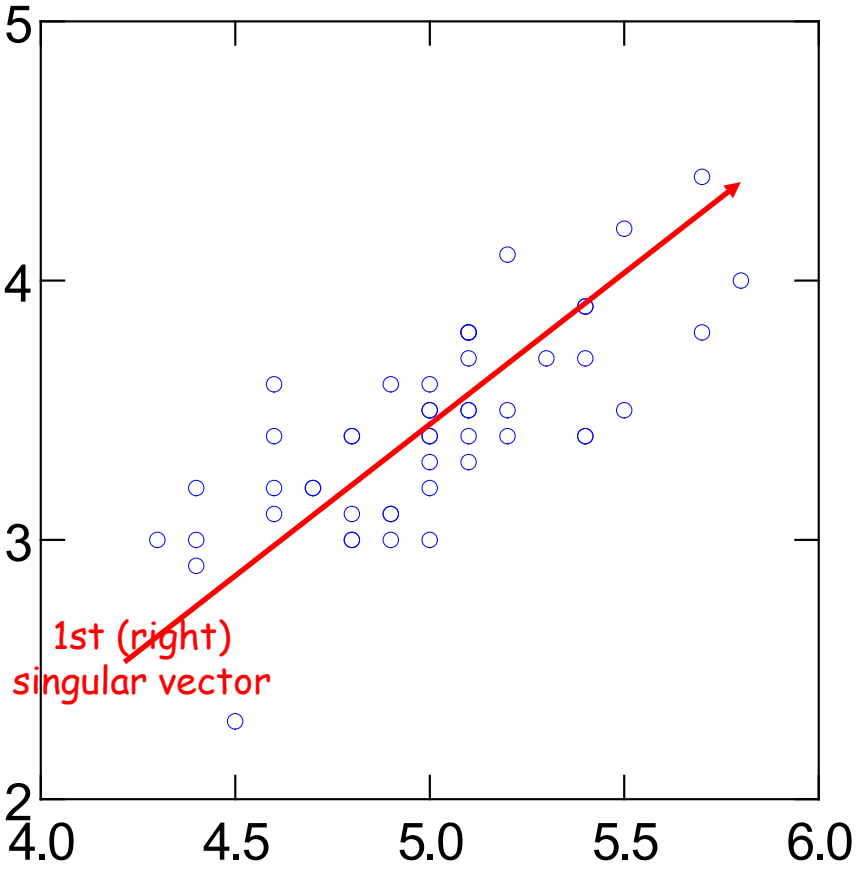
The Singular Value Decomposition (SVD)



Let the blue circles represent *m* data points in a 2-D Euclidean space.

Then, the SVD of the *m-by-2* matrix of the data will return …

The Singular Value Decomposition (SVD)



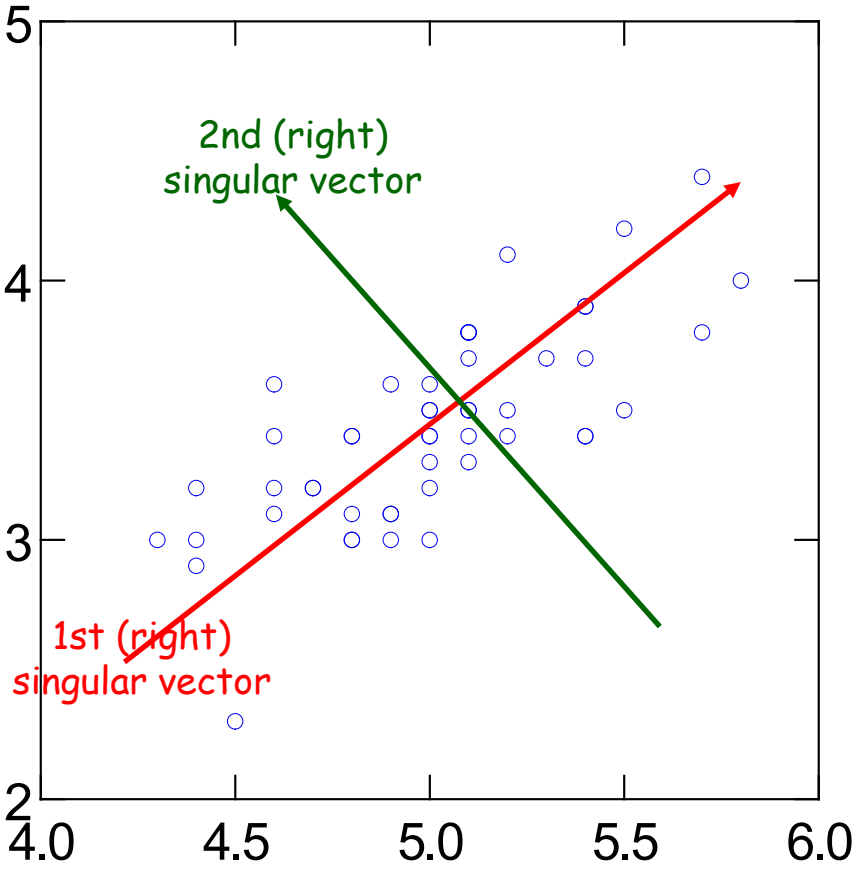
Let the blue circles represent *m* data points in a 2-D Euclidean space.

Then, the SVD of the *m-by-2* matrix of the data will return …

1st (right) singular vector:

direction of maximal variance,

The Singular Value Decomposition (SVD)



Let the blue circles represent *m* data points in a 2-D Euclidean space.

Then, the SVD of the *m*-by-2 matrix of the data will return …

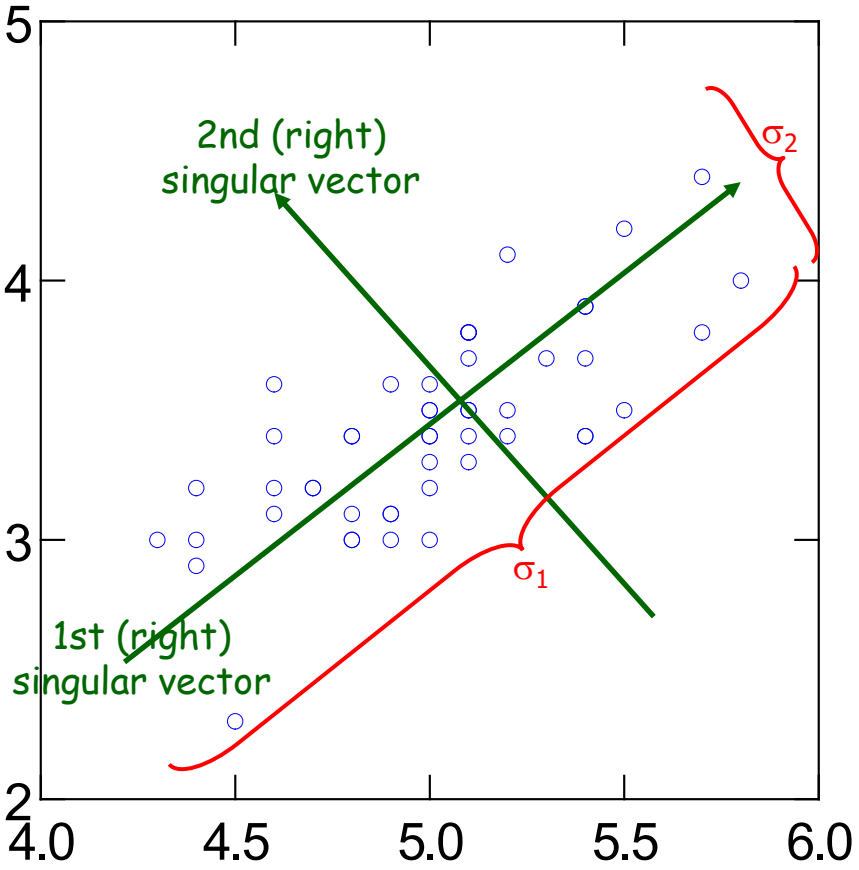
1st (right) singular vector:

direction of maximal variance,

2nd (right) singular vector:

direction of maximal variance, after removing the projection of the data along the first singular vector.

Singular values



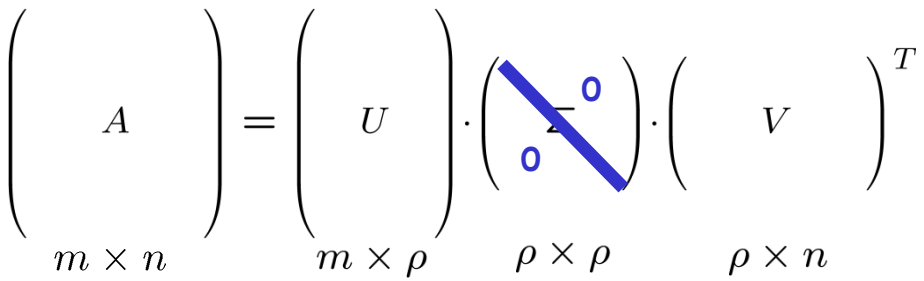
σ_1: measures how much of the data variance is explained by the first singular vector.

σ_2: measures how much of the data variance is explained by the second singular vector.

Principal Components Analysis (PCA) is done via the computation of the Singular Value Decomposition (SVD) of a (mean-centered) covariance matrix.

Typically, a small constant number (say k) of the top singular vectors and values are kept.

SVD: formal definition



ρ: rank of *A*

U (V): orthogonal matrix containing the left (right) singular vectors of *A*.

Σ: diagonal matrix containing the singular values of *A*.

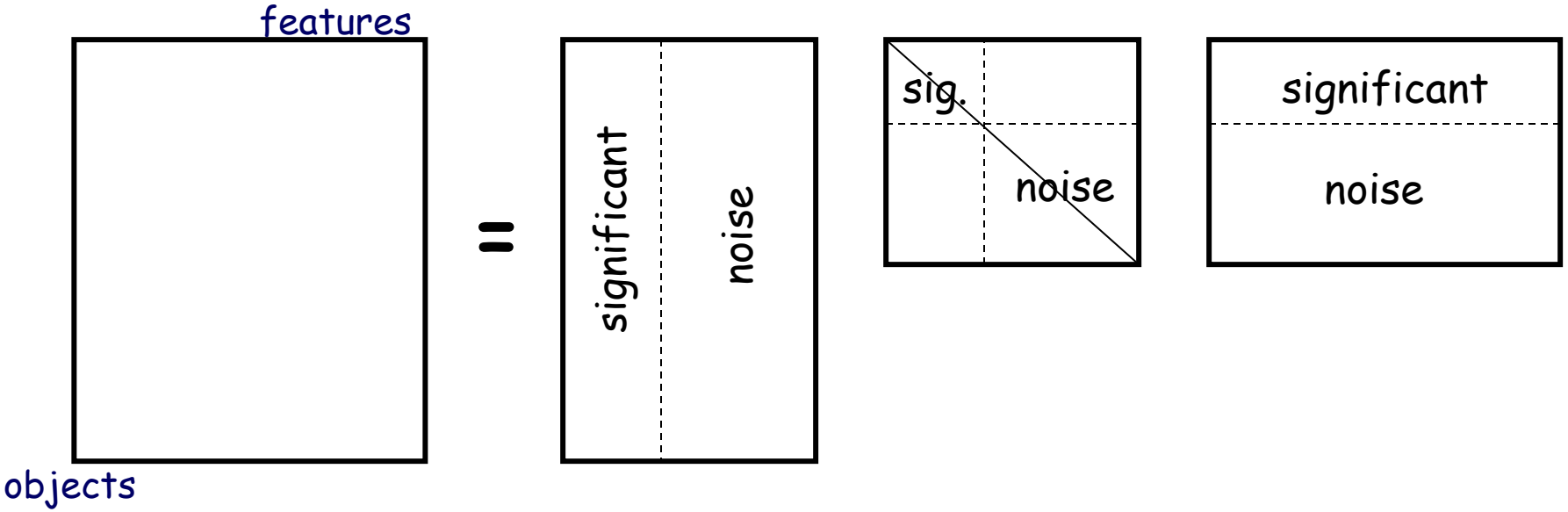
Let σ_1 , σ_2 , … , σ_ρ be the entries of Σ.

Exact computation of the SVD takes $O(\min\{mn^2 , m^2n\})$ time.

The top *k* left/right singular vectors/values can be computed faster using iterative methods.

Rank-k approximations via the SVD

$$A = U \Sigma V^T$$

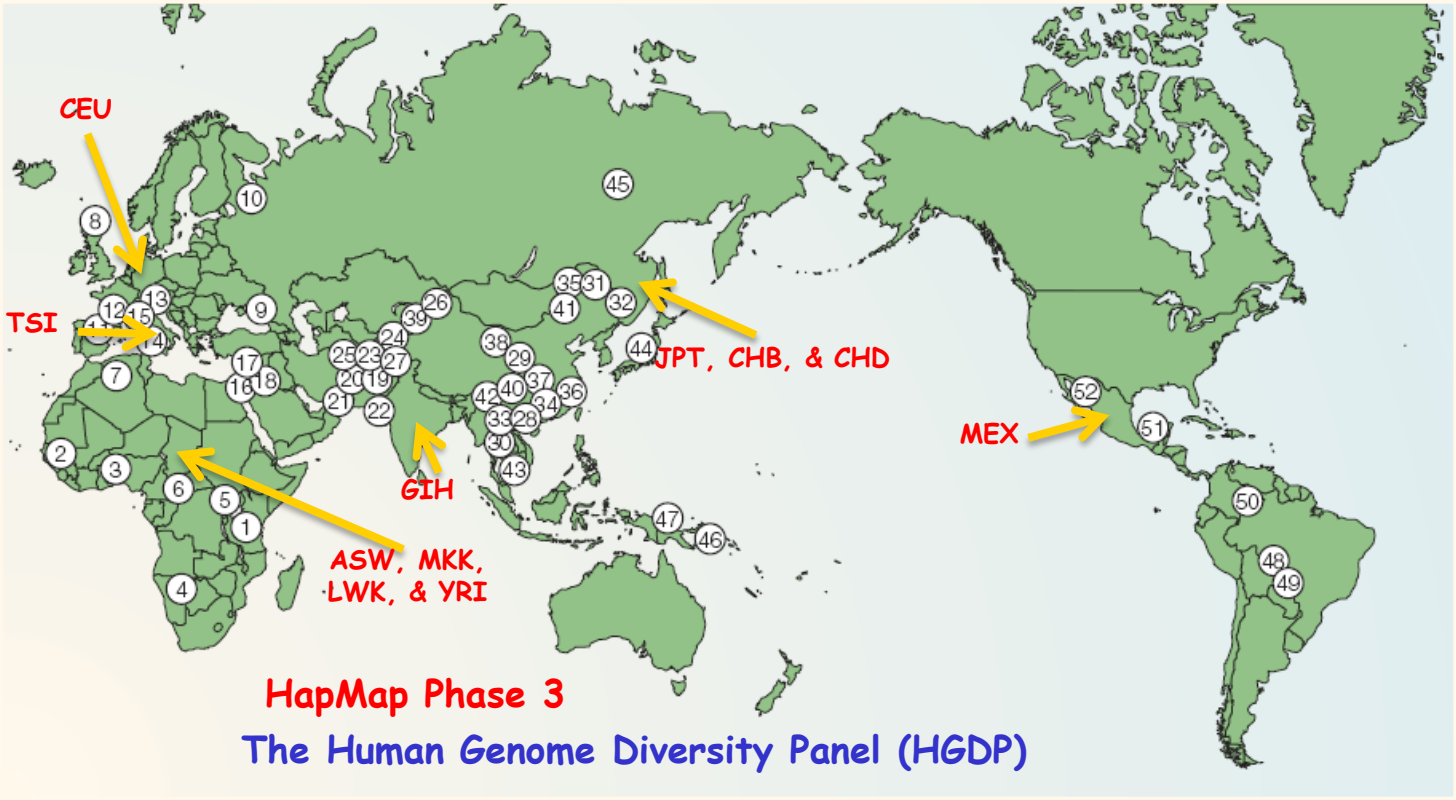


Rank-k approximations (A_k)

$$\underset{m \times n}{A_k} = \underset{m \times k}{U_k} \cdot \underset{k \times k}{\Sigma_k} \cdot \underset{k \times n}{V_k^T}$$

U_k (V_k): orthogonal matrix containing the top k left (right) singular vectors of A.

Σ_k: diagonal matrix containing the top k singular values of A.



HGDP data

- 1,033 samples
- 7 geographic regions
- 52 populations

HapMap Phase 3 data

- 1,207 samples
- 11 populations

Matrix dimensions:

2,240 subjects (rows)

447,143 SNPs (columns)

SVD/PCA returns…

Africans	Europeans	Western Asians	Eastern Asians	Oceanians
1 Bantu	8 Orcadian	16 Bedouin	28 Han (S. China)	46 Melanesian
2 Mandenka	9 Adygei	17 Druze	29 Han (N. China)	47 Papuan
3 Yoruba	10 Russian	18 Palestinian	30 Dai	
4 San	11 Basque		31 Daur	
5 Mbuti pygmy	12 French	Central and Southern Asians	32 Hezhen	Native Americans
6 Biaka	13 North Italian	19 Balochi	33 Lahu	48 Karitiana
7 Mozabite	14 Sardinian	20 Brahui	34 Miao	49 Surui
	15 Tuscan	21 Makrani	35 Oroqen	50 Colombian
		22 Sindhi	36 She	51 Maya
		23 Pathan	37 Tujia	52 Pima
		24 Burusho	38 Tu	
		25 Hazara	39 Xibo	
		26 Uygur	40 Yi	
		27 Kalash	41 Mongola	
			42 Naxi	
			43 Cambodian	
			44 Japanese	
			45 Yakut	

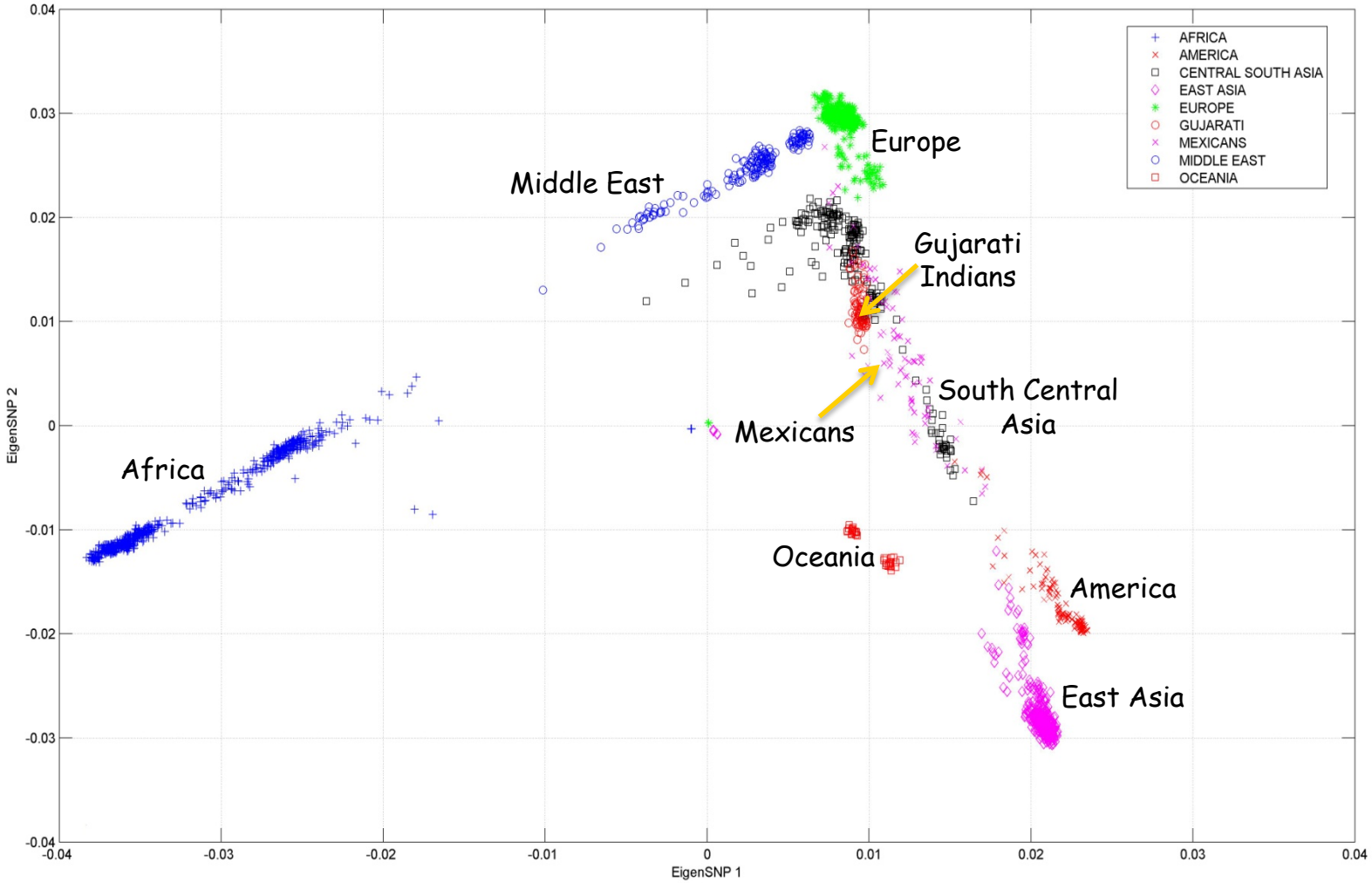
Cavalli-Sforza (2005) *Nat Genet Rev*

Rosenberg et al. (2002) *Science*

Li et al. (2008) *Science*

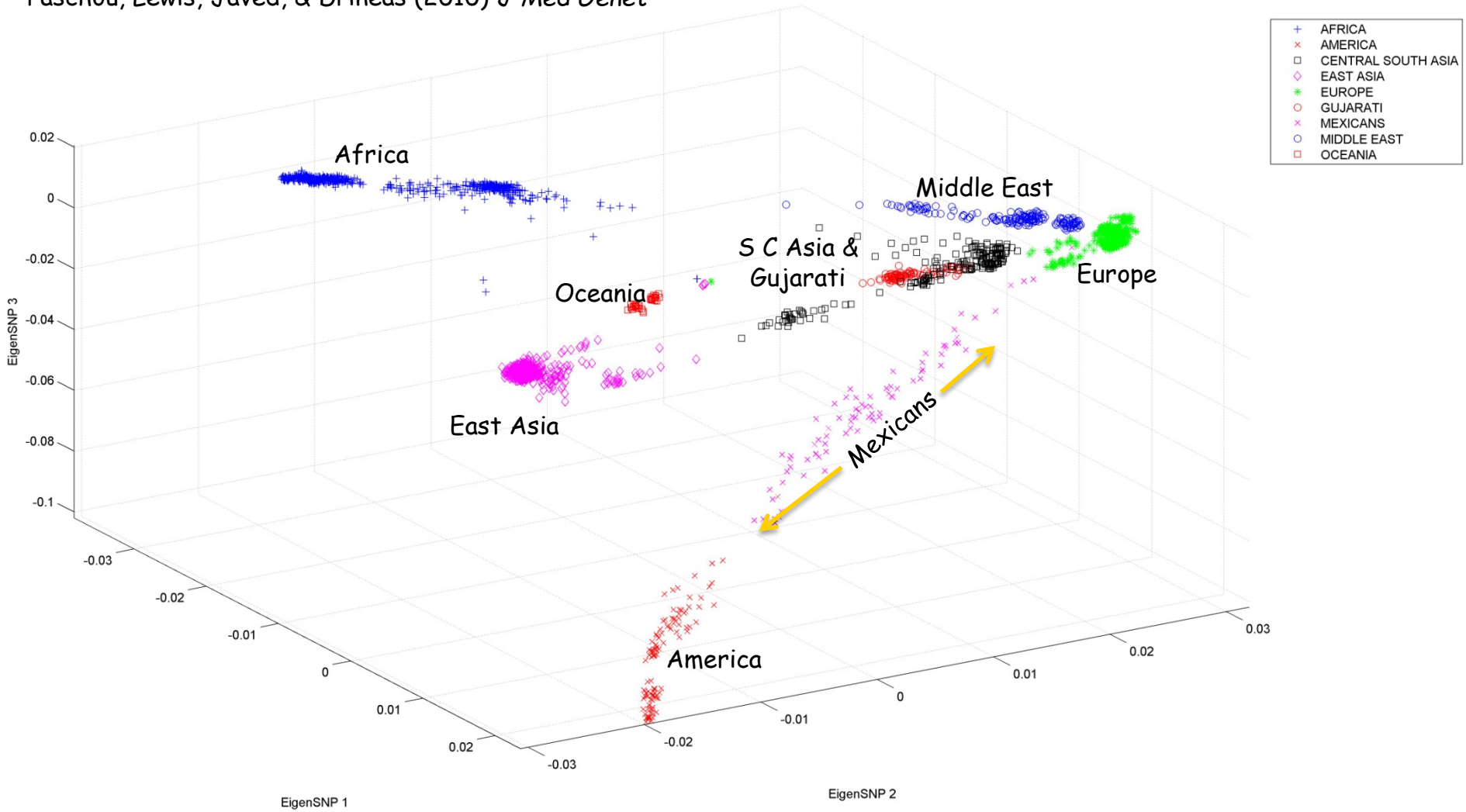
The International HapMap Consortium (2003, 2005, 2007), *Nature*

Paschou, Lewis, Javed, & Drineas (2010) *J Med Genet*



- Top two Principal Components (eigenSNPs)
- Mexican population seems out of place: we move to the top three PCs.

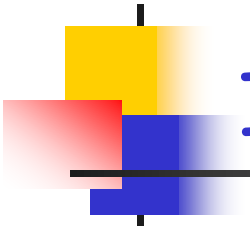
Paschou, Lewis, Javed, & Drineas (2010) *J Med Genet*



Not altogether satisfactory: the principal components are linear combinations of all SNPs, and – of course – can not be assayed!

Can we find **actual SNPs** that capture the information in the singular vectors?

Formally: spanning the same subspace.



Issues: computational time

Computing large SVDs: computational time

- In commodity hardware (e.g., a 32GB RAM, i7 laptop), using MatLab R2021, the computation of the SVD of the dense 2,240-by-447,143 matrix A takes about 4 minutes.
- Computing this SVD is not a one-liner, since we (I?) could not load the whole matrix in RAM (runs out-of-memory in MatLab R2021); we compute the eigendecomposition of AA^T.
- **Current needs:** we need to compute SVDs on biobank scale data (0.5M-1M samples genotyped on millions of SNPs).

Issues: computational time

Computing large SVDs: computational time

- In commodity hardware (e.g., a 32GB RAM, i7 laptop), using MatLab R2021, the computation of the SVD of the dense 2,240-by-447,143 matrix A takes about 5 minutes.
- Computing this SVD is not a one-liner, since we (I?) could not load the whole matrix in RAM (runs out-of-memory in MatLab R2021); we compute the eigendecomposition of AA^T.
- **Current needs:** we need to compute SVDs on biobank scale data (0.5M-1M samples genotyped on millions of SNPs).

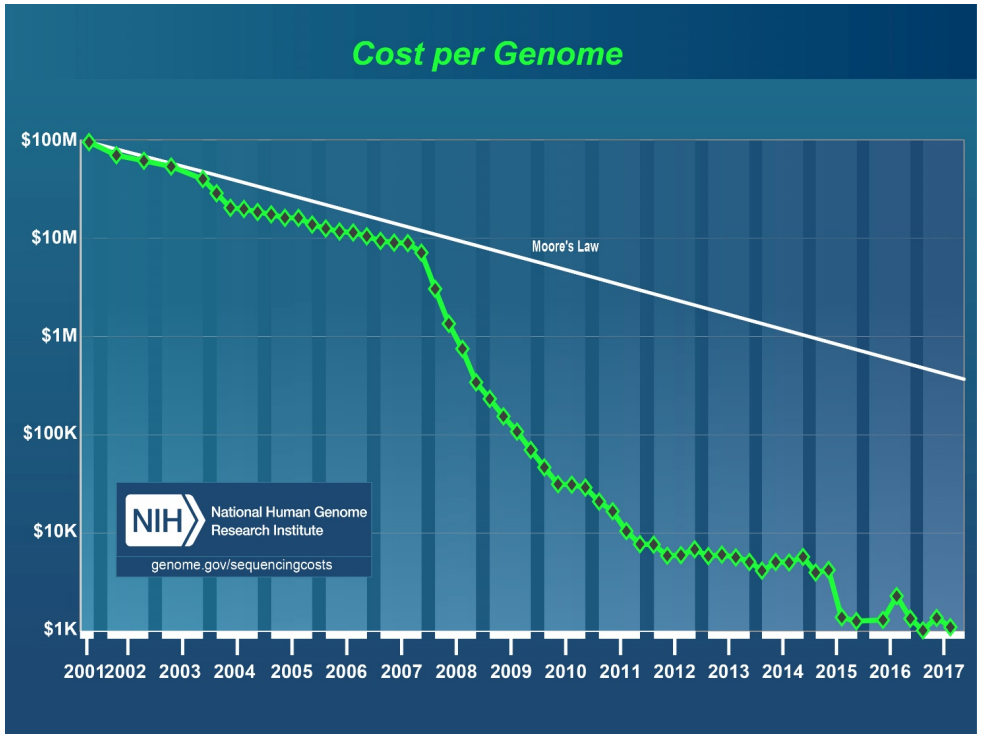
Running time will always be a concern, but: we only need the top few principal components; machine-precision accuracy is not necessary!

- Data are noisy.
- Approximate singular vectors suffice.

Iterative methods with random starting points are well-explored in numerical linear algebra.

- Subspace iteration, Krylov subspace methods, etc.
- Careful implementations that scale are important.

Growing scale of Sequencing



- Cost of sequencing and genotyping has gone down exponentially in recent years. Number of individuals sequenced has thus resulted in an exponential growth.
- From the start of Human Genome project, to Human Genome Diversity Panel (1043 individuals, 660K SNPs) to now, UK Biobank having 500K individuals and ~95 million SNPs.
- Biotech companies such as 23andMe, AncestryDNA, etc. have successfully sequenced around 2 million individuals and about 20-30 million (M) SNPs.

Bose et al. "TeraPCA: a fast and scalable software package to study genetic variation in tera-scale genotypes," Bioinformatics, 2019

Subspace Iteration: Methods

- Subspace Iteration method is essentially a generalization of power method to approximate a k-dimensional $(k > 1)$ invariant subspace, rather than one eigenvector at a time.

- For a square matrix, $B \in \mathbb{R}^{n \times n}$, a positive integer p and a basis matrix $X_0 \in \mathbb{R}^{n \times s}$ of an initial subspace, the subspace iteration computes the matrix: $X = B^p X_0$.

- X_0 is our initial guess matrix, for which we choose it to be random Gaussian vectors i.i.d from an $N(0,1)$ distribution.

- Given $A \in \mathbb{R}^{m \times n}, X_0$ and p , Subspace Iteration computes $X = (AA^T)^p X_0$.

Subspace Iteration: Algorithm

Input: $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{p} > 0$, initial guess $\boldsymbol{X_0} \in \mathbb{R}^{m \times s}$ and $\boldsymbol{k} \leq \min\{m, n\}$.

Output: Q $\in \mathbb{R}^{m \times \mathrm{k}}$

1: $C = A(A^T X_0)$

2: **repeat**

3: **For** $i = 2$ *to* p do

4: $Q = orth(C)$

5: $C = A(A^T Q)$

6: **End For**

7: $Q = orth(C)$

8: $C = AA^T Q$

8: $M = Q^T C$

9: Compute eigenvalue decomposition $M = \mathrm{PDP}^{\mathrm{T}}$

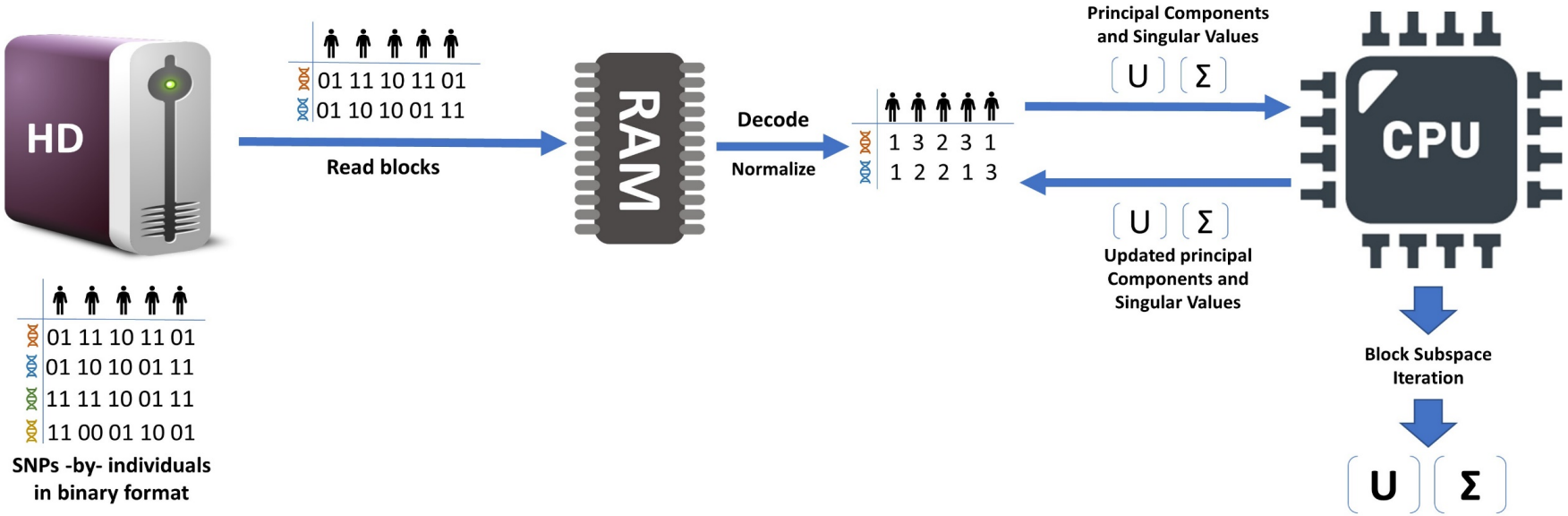
10: $C = QP$

10: **until** convergence

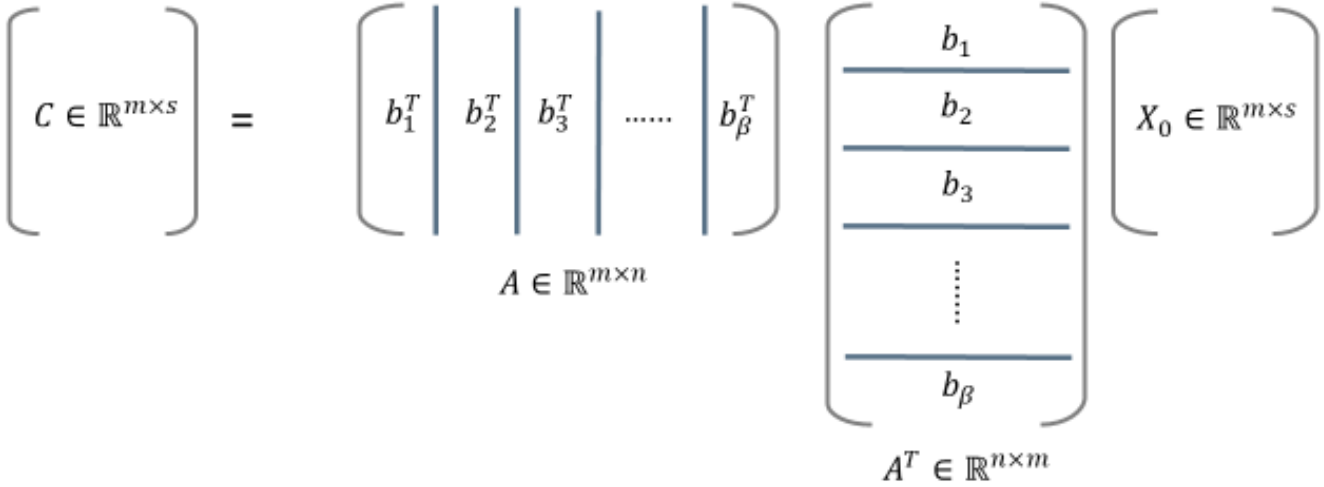
11: return Q

Computational Hotspots

The problem is RAM, not running time…



Computing $A(A^T X_0)$



We compute C as follows: $C = b_1^T b_1 X_{k-1} + b_2^T b_2 X_{k-1} + b_3^T b_3 X_{k-1} + \ ... + b_\beta^T b_\beta X_{k-1}$, where β is the number of blocks.

So, we can write $C = \sum_{i=1}^{\beta} b_i^T (b_i X)$

Experiments

We compared the performance of TeraPCA with current industry standard, flashPCA2, as it performs the best out of the available packages. We used both real and simulated data sets to show that TeraPCA performs better than FlashPCA2 with or without invoking multithreading. We ran the following experiments:

Data sets	Size	Dimensions
HGDP	6 GB	1,043 individuals, 107,468 markers
1000 Genomes	38 GB	2,504 individuals, 808,647 markers
5K -by- 1M	19 GB	5,000 individuals, 1,000,000 markers
10K -by- 1M	38 GB	10,000 individuals, 1,000,000 markers
100K -by- 1M	373 GB	100,000 individuals, 1,000,000 markers
500K -by- 1M	1.9 TB	500,000 individuals, 1,000,000 markers
1M -by- 1M	3.7 TB	1,000,000 individuals, 1,000,000 markers

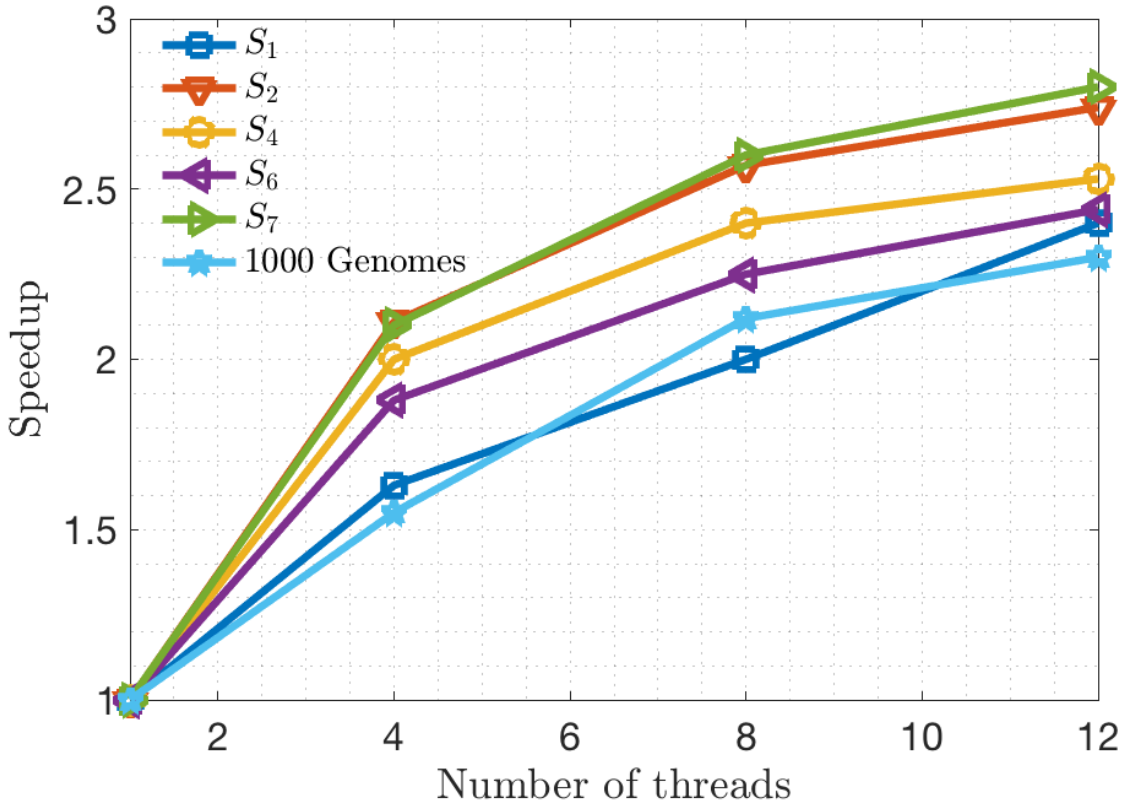
All computations were done in a single core Intel Xeon-Gold processor with 96 GB max RAM

TeraPCA: Performance Comparisons

Dataset	TeraPCA	FlashPCA2	Speedup
5K -by- 1M	26.2mins	33.3mins	1.3
10K -by- 1M	39.3mins	87.5mins	2.2
100K -by- 1M	6.99hrs	35.64hrs	4.5
500K -by- 1M	7.3hrs	n/a*	∞
1M -by- 1M	13.2hrs	n/a*	∞
100K -by- 100K	39.46mins	141.1mins	3.6
HGDP	6.45secs	7.7secs	1.2
1000Genomes	4.2 mins	3.9 mins	0.9

* n/a: not applicable as FlashPCA2 did not terminate in 75 hrs

Bose et al. "TeraPCA: a fast and scalable software package to study genetic variation in tera-scale genotypes," Bioinformatics, 2019

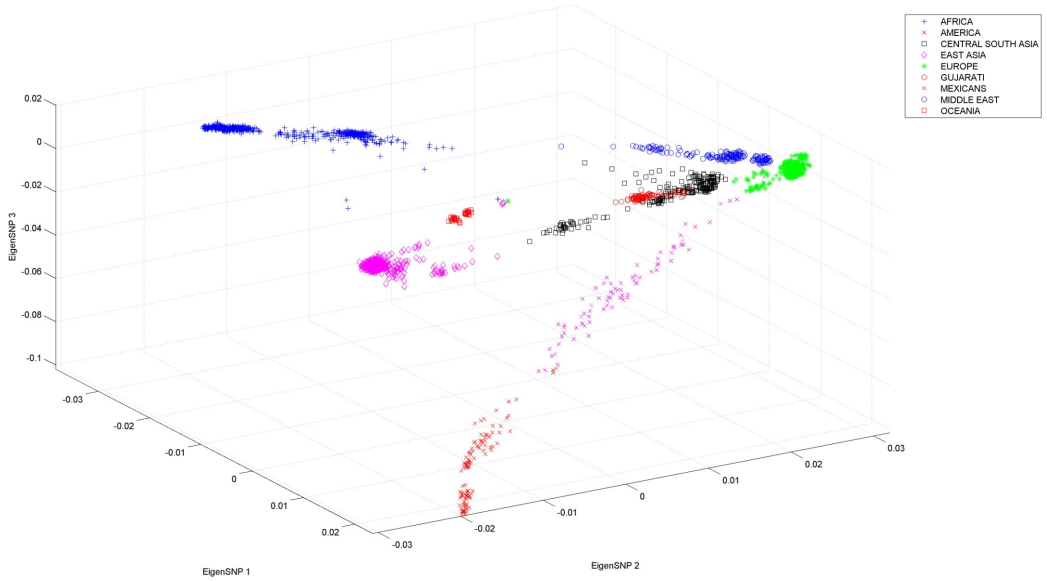


TeraPCA scales "decently" with increasing number of threads.

(C++, MPI and multithreaded implementations using Intel's OpenMP library)

Back to interpretability…

- **Selecting good columns (SNPs) that "capture the structure" of the top PCs**
  - Combinatorial optimization problem; hard even for small matrices.
  - Often called the Column Subset Selection Problem (CSSP).
  - Not clear that such columns even exist.



SVD decomposes a matrix as…

$$\underset{m \times n}{\begin{pmatrix} \\ A \\ \\ \end{pmatrix}} \approx \underset{m \times k}{\begin{pmatrix} \\ U_k \\ \\ \end{pmatrix}} \underset{k \times n}{\begin{pmatrix} \\ X \\ \\ \end{pmatrix}}$$

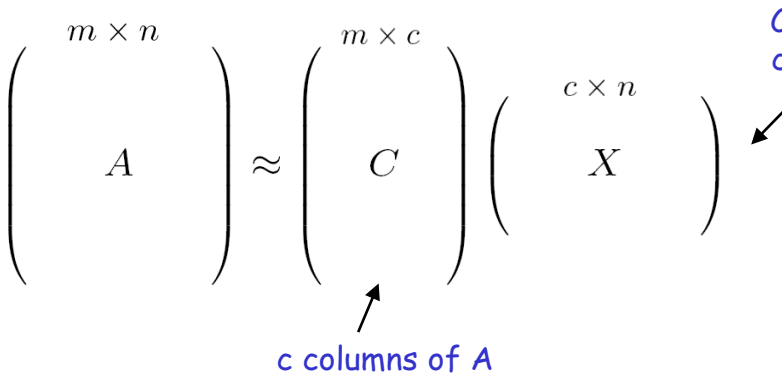
Top k left singular vectors

The SVD has strong optimality properties.

- It is easy to see that $X = \Sigma_k V_k^T = U_k^T A$.
- SVD has strong optimality properties.
- The columns of U_k are linear combinations of up to all columns of A.

The CX decomposition

Drineas, Mahoney, & Muthukrishnan (2008) SIAM J Mat Anal Appl
Mahoney & Drineas (2009) PNAS



Goal: make (some norm) of A-CX small.

Why?

If A is a subject-SNP matrix, then selecting representative columns is equivalent to selecting representative SNPs to capture the same structure as the top eigenSNPs.

We want c as small as possible!

CX decomposition

$$\underset{m \times n}{\begin{pmatrix} \\ A \\ \\ \end{pmatrix}} \approx \underset{m \times c}{\begin{pmatrix} \\ C \\ \\ \end{pmatrix}} \underset{c \times n}{\begin{pmatrix} \\ X \\ \\ \end{pmatrix}}$$

c columns of A

Easy to prove that optimal X = C^+A. (C^+ is the Moore-Penrose pseudoinverse of C.)

Thus, the challenging part is to find good columns (SNPs) of A to include in C.

From a mathematical perspective, this is a hard combinatorial problem, closely related to the so-called Column Subset Selection Problem (CSSP).

The CSSP has been heavily studied in Numerical Linear Algebra.

Relative-error Frobenius norm bounds

Drineas, Mahoney, & Muthukrishnan (2008) SIAM J Mat Anal Appl

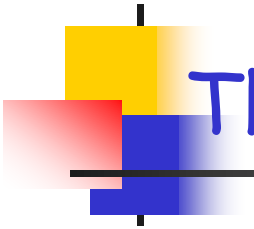
Given an m-by-n matrix A, there exists an $O(mn^2)$ algorithm that picks

at most $O((k/\varepsilon^2) \log (k/\varepsilon))$ columns of A

such that with probability at least .9

$$\|A - CX\|_F = \left\|A - CC^{\dagger}A\right\|_F \leq (1+\varepsilon)\ \|A - A_k\|_F$$

Notation: $\|X\|_F^2 = \sum_{i,j} X_{ij}^2$



The algorithm

Input: m-by-n matrix A,

$0 < \varepsilon < .5$, the desired accuracy

Output: C, the matrix consisting of the selected columns

Sampling algorithm

- Compute probabilities p_j summing to 1.
- Let $c = O((k/\varepsilon^2) \log (k/\varepsilon))$.
- In c i.i.d. trials pick columns of A, where in each trial the j-th column of A is picked with probability p_j.
- Let C be the matrix consisting of the chosen columns.

Subspace sampling (Frobenius norm)

$$\underset{m \times n}{\begin{pmatrix} \\ A_k \\ \\ \end{pmatrix}} = \underset{m \times k}{\begin{pmatrix} \\ U_k \\ \\ \end{pmatrix}} \cdot \underset{k \times k}{\begin{pmatrix} \\ \Sigma_k \\ \\ \end{pmatrix}} \cdot \underset{k \times n}{\begin{pmatrix} \\ V_k^T \\ \\ \end{pmatrix}}$$

V_k: orthogonal matrix containing the top k right singular vectors of A.

Σ_k: diagonal matrix containing the top k singular values of A.

Remark: The rows of V_k^T are orthonormal vectors, but its columns $(V_k^T)^{(j)}$ are not.

Subspace sampling (Frobenius norm)

$$\underset{m \times n}{\begin{pmatrix} \\ A_k \\ \\ \end{pmatrix}} = \underset{m \times k}{\begin{pmatrix} \\ U_k \\ \\ \end{pmatrix}} \cdot \underset{k \times k}{\begin{pmatrix} \\ \Sigma_k \\ \\ \end{pmatrix}} \cdot \underset{k \times n}{\begin{pmatrix} \\ V_k^T \\ \\ \end{pmatrix}}$$

V_k: orthogonal matrix containing the top k right singular vectors of A.

Σ_k: diagonal matrix containing the top k singular values of A.

Remark: The rows of V_k^T are orthonormal vectors, but its columns $(V_k^T)^{(j)}$ are not.

Subspace sampling in $O(mn^2)$ time

$$p_j = \frac{\left\| \left(V_k^T\right)^{(j)} \right\|_2^2}{k}$$

Normalization s.t. the p_j sum up to 1

Subspace sampling (Frobenius norm)

$$\underset{m \times n}{\begin{pmatrix} \\ A_k \\ \\ \end{pmatrix}} = \underset{m \times k}{\begin{pmatrix} \\ U_k \\ \\ \end{pmatrix}} \cdot \underset{k \times k}{\begin{pmatrix} \\ \Sigma_k \\ \\ \end{pmatrix}} \cdot \underset{k \times n}{\begin{pmatrix} \\ V_k^T \\ \\ \end{pmatrix}}$$

V_k: orthogonal matrix containing the top k right singular vectors of A.

Σ_k: diagonal matrix containing the top k singular values of A.

Remark: The rows of V_k^T are orthonormal vectors, but its columns $(V_k^T)^{(j)}$ are not.

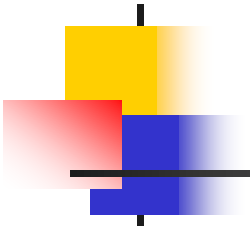
Subspace sampling in $O(mn^2)$ time

Leverage scores

(useful in statistics for outlier detection)

$$p_j = \frac{\left\| (V_k^T)^{(j)} \right\|_2^2}{k}$$

Normalization s.t. the p_j sum up to 1



Deterministic variant of CX

Paschou et al (2007) *PLoS Genetics*

Mahoney and Drineas (2009) *PNAS*

Input: m-by-n matrix A,
integer k, and
c (number of SNPs to pick)

Output: the selected SNPs

CX algorithm

- Compute the scores p_j
- Pick the columns (SNPs) corresponding to the top ***c*** scores

Deterministic variant of CX (cont'd)

Paschou et al (2007) *PLoS Genetics*

Mahoney and Drineas (2009) *PNAS*

Input: m-by-n matrix A,
integer k, and
c (number of SNPs to pick)

Output: the selected SNPs: Ancestry Informative Markers

CX algorithm

- Compute the scores p_j
- Pick the columns (SNPs) corresponding to the top ***c*** scores

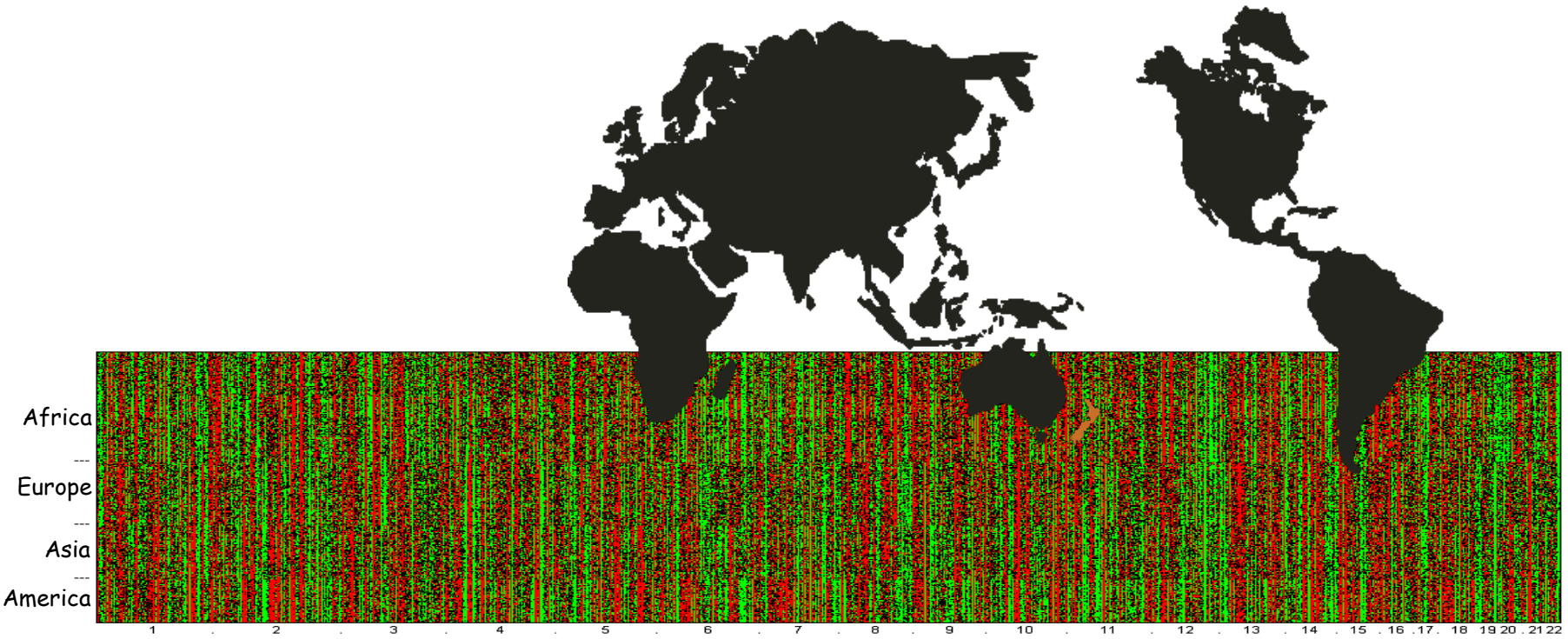
In order to estimate *k* for SNP data, we developed a permutation-based test to determine whether a certain principal component is significant or not.

(A similar test was presented in Patterson et al (2006) *PLoS Genetics*)

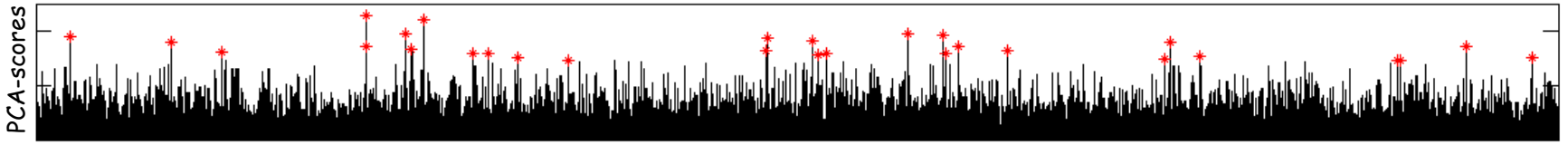
Worldwide data

274 individuals, 12 populations, ~10,000 SNPs using the Affymetrix array

Selecting PCA Informative SNPs for individual assignment to four continents (Africa, Europe, Asia, America)



* top 30 PCA-correlated SNPs

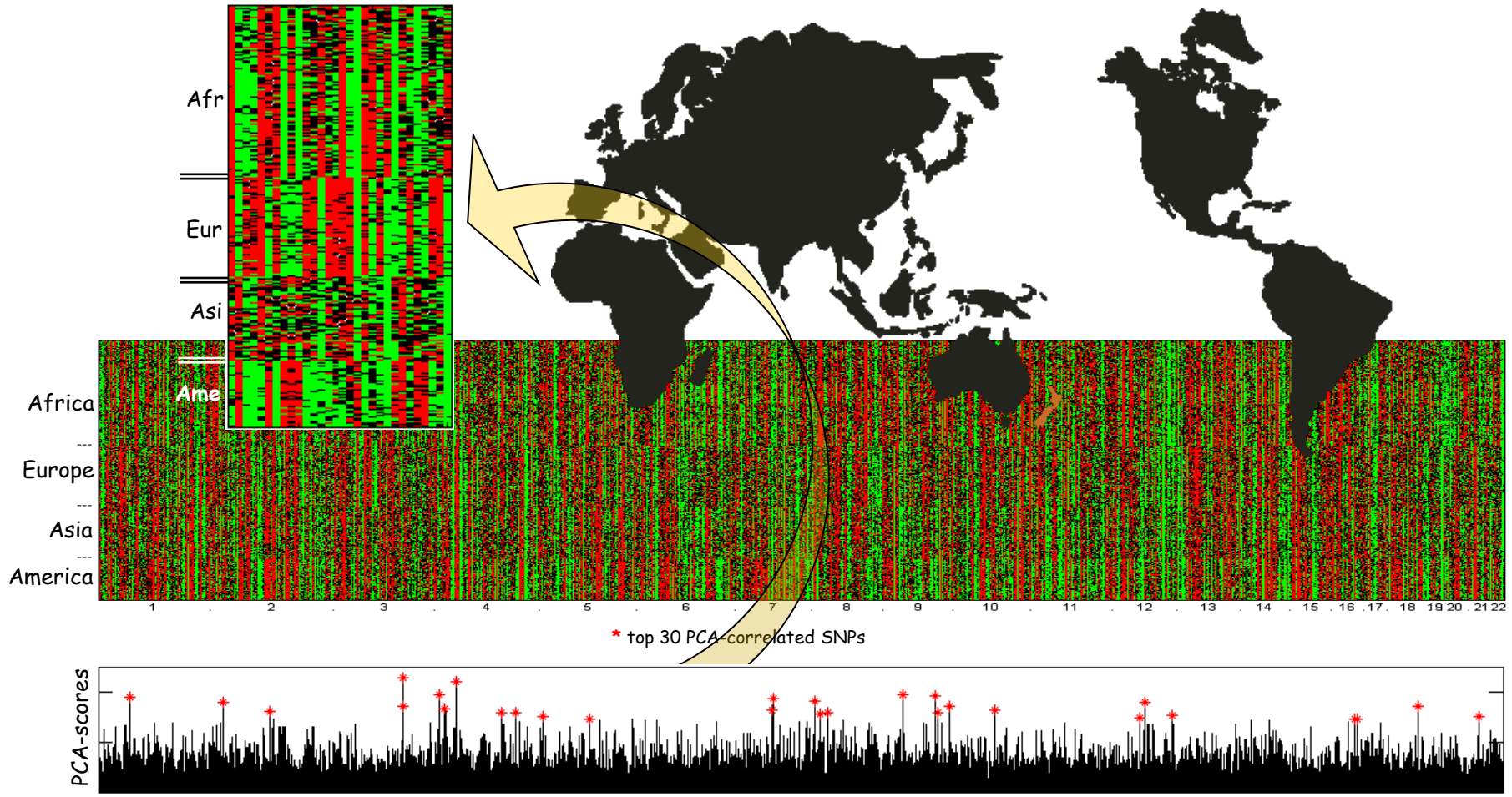


SNPs by chromosomal order

Paschou et al (2007; 2008) PLoS Genetics; Paschou et al (2010) J Med Genet; Drineas et al (2010) PLoS One

Hughey, Paschou, Drineas, et al. (2013) Nat Comm; Paschou, Drineas, et al. PNAS 2014;

Selecting PCA Informative SNPs for individual assignment to four continents (Africa, Europe, Asia, America)

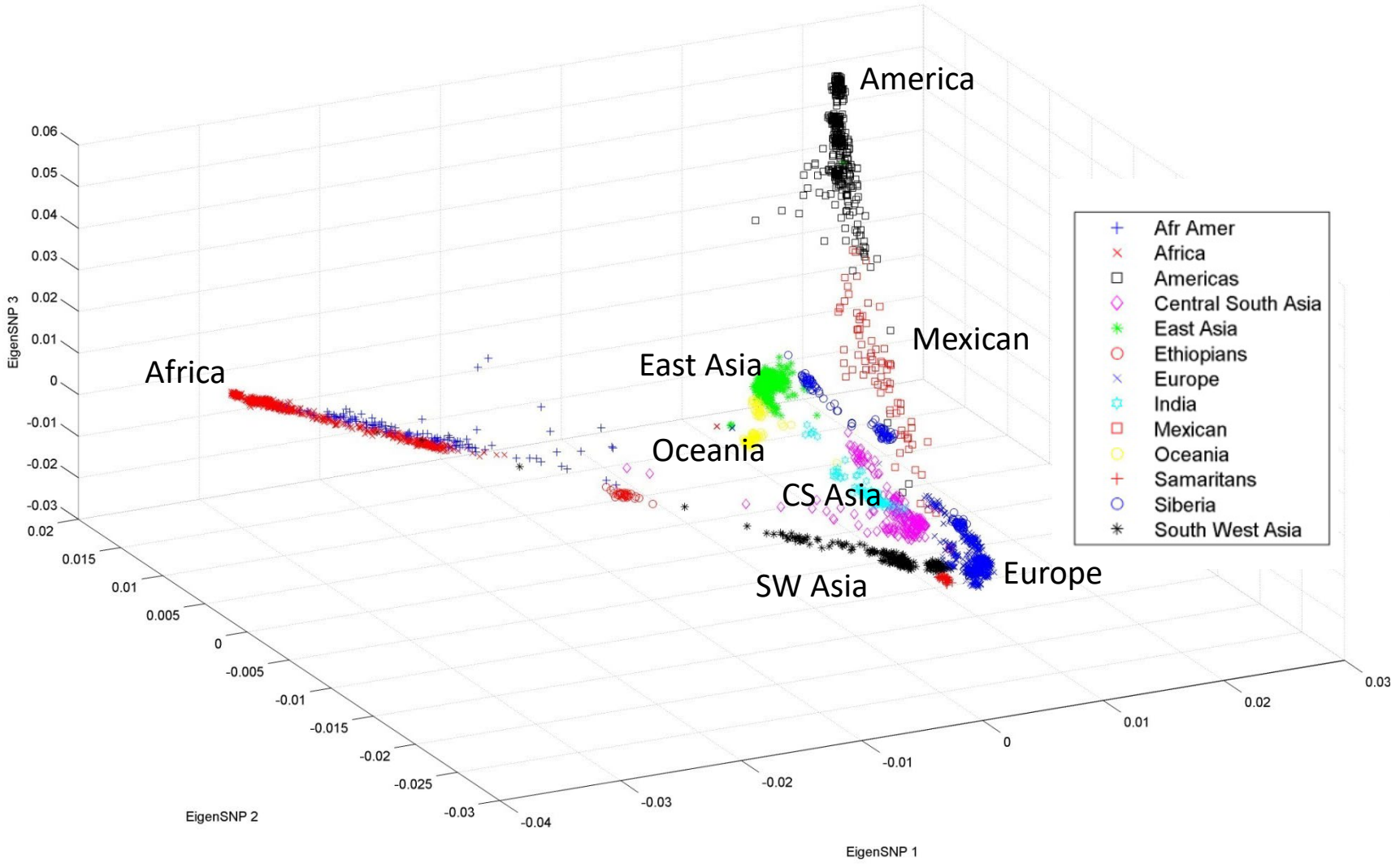


Paschou et al (2007; 2008) PLoS Genetics; Paschou et al (2010) J Med Genet; Drineas et al (2010) PLoS One

Hughey, Paschou, Drineas, et al. (2013) Nat Comm; Paschou, Drineas, et al. PNAS 2014;

A large world-wide sample: ALFRED data (K.K. Kidd's lab @ Yale)

A total of **3,567** samples from **92** populations and **442,516** common SNPs



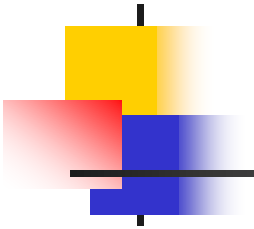
Highest scoring "genes"

Gene	Function (RefSeq)
EDAR*	Ectodermal development, hair follicle formation.
PTK6	Intracellular signal transducer in epithelial tissues. Sensitization of cells to epidermal growth factor.
SPATA20*	Associated with spermatogenesis.
MCHR1	Plasma membrane protein which binds melanin-concentrating hormone. Probably involved in the neuronal regulation of food consumption.
FOXP1*	Forkhead box transcription factors play important roles in the regulation of tissue- and cell type-specific gene transcription during both development and adulthood.
PSCD3*	Involved in the control of Golgi structure and function.
OCA2*	Skin/Hair/Eye pigmentation.
EGFR*	This protein is a receptor for members of the epidermal growth factor family. Associated with the melanin pathway.

*Barreiro et al (2008) Nat Genet

*Sabeti et al (2007) Nature

*The International HapMap Consortium (2007) Nature



A problem with the CX decomposition

Input: m-by-n matrix A, integer k, and c (number of SNPs to pick)

Output: the selected PCA Informative Markers or PCAIMs

CX algorithm

- Compute the scores p_j
- Pick the columns (SNPs) corresponding to the top c scores.

Problem:

Highly correlated SNPs (a.k.a., SNPs that are in LD) get similar – high – scores, and thus the deterministic variant would select redundant SNPs.

How do we remove this redundancy?

Rank-Revealing QR factorization

Paschou et al (2008) *PLoS Genetics*
Boutsidis, Mahoney, & Drineas (2009) *SODA*

We use a standard greedy approach (the Rank-Revealing QR factorization).

The algorithm performs k iterations:

In the first iteration, the top PCAIM is picked;

In the second iteration, a PCAIM is picked that is as uncorrelated to with the previously selected PCAIM as possible;

In the third iteration the chosen PCAIM has to be as uncorrelated as possible with the first two previously selected PCAIMs;

And so on…

Efficient implementations are available, and run in a couple of minutes for typical values of m, c, and k.

Selecting fewer columns

Problem

How many columns do we need to include in the matrix C in order to get relative-error approximations ?

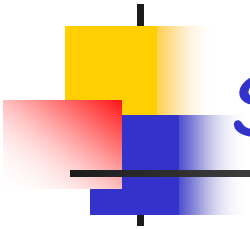
Recall: with $O((k/\varepsilon^2) \log (k/\varepsilon))$ columns, we get (subject to a failure probability)

$$\left\| A - CC^{\dagger}A \right\|_F \leq (1+\epsilon) \left\| A - A_k \right\|_F$$

Deshpande & Rademacher (FOCS '10): with exactly k columns, we get

$$\left\| A - CC^{\dagger}A \right\|_F \leq \sqrt{k} \left\| A - A_k \right\|_F$$

What about the range between k and O(k log(k))?



Selecting fewer columns (cont'd)

(Boutsidis, Drineas, & Magdon-Ismail, FOCS 2011 and SICOMP 2014)

Question:

What about the range between k and O(k log(k))?

Answer:

A relative-error bound is possible by selecting **$s=2k/\varepsilon$** columns!

Technical breakthrough;

A combination of sampling strategies with a novel approach on column selection, inspired by the work of Batson, Spielman, & Srivastava (STOC '09) on graph sparsifiers.

- **The running time is $O((mnk+nk^3)\varepsilon^{-1})$.**
- **Simplicity is gone…**

CSSP: Lower bounds & other approaches

Guruswami & Sinop, SODA 2012

Alternative approaches, based on volume sampling, guarantee

$(r+1)/(r+1-k)$ relative error bounds.

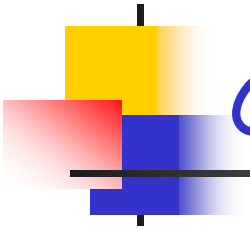
This bound is asymptotically optimal (up to lower order terms).

The proposed deterministic algorithm runs in $O(rnm^3 \log m)$ time, while the randomized algorithm runs in $O(rnm^2)$ time and achieves the bound in expectation.

Guruswami & Sinop, FOCS 2011

Applications of column-based reconstruction in Quadratic Integer Programming.

Very large body of followup work in the Theoretical Computer Science



CSSP

Massive body of follow-up work on the CSSP, including the NeurIPS 2020 best paper award for:

Michal Derezinski, Rajiv Khanna, Michael W. Mahoney, "*Improved guarantees and a multiple-descent curve for the Column Subset Selection Problem and the Nyström method*", NeurIPS 2020.

(See discussion and references in the above paper for a summary of theoretical and applied work on the CSSP.)

We also use **genetics analyses to elucidate population relationships** and provide answers to historical questions of relevance to archeology and paleoanthropology.

Again, PCA plots are quite telling.

Multiple examples from our own work:

- A maritime path for the colonization of Europe.

(Paschou et al. PNAS 2014)

- The origins of the Minoan civilization.

(Hughey et al. Nat Comms 2013)

- Disproving Fallmerayer's hypothesis (~1830s) that Byzantine and medieval Greeks (esp. Peloponneseans) were extinguished by Slavic invaders and replaced by Slavic settlers during the 6th century CE.

(Stamatoyannopoulos et al. Eur J Hum Gen 2017; Drineas et al. Hum Gen 2019)

We started collecting data to investigate these hypotheses since 2011; **joint work with P. Paschou (Purdue), J. Stamatoyannopoulos (U Washington), and G. Stamatoyannopoulos (U Washington)**.

Greece at the crossroads of Neolithic migrations into Europe

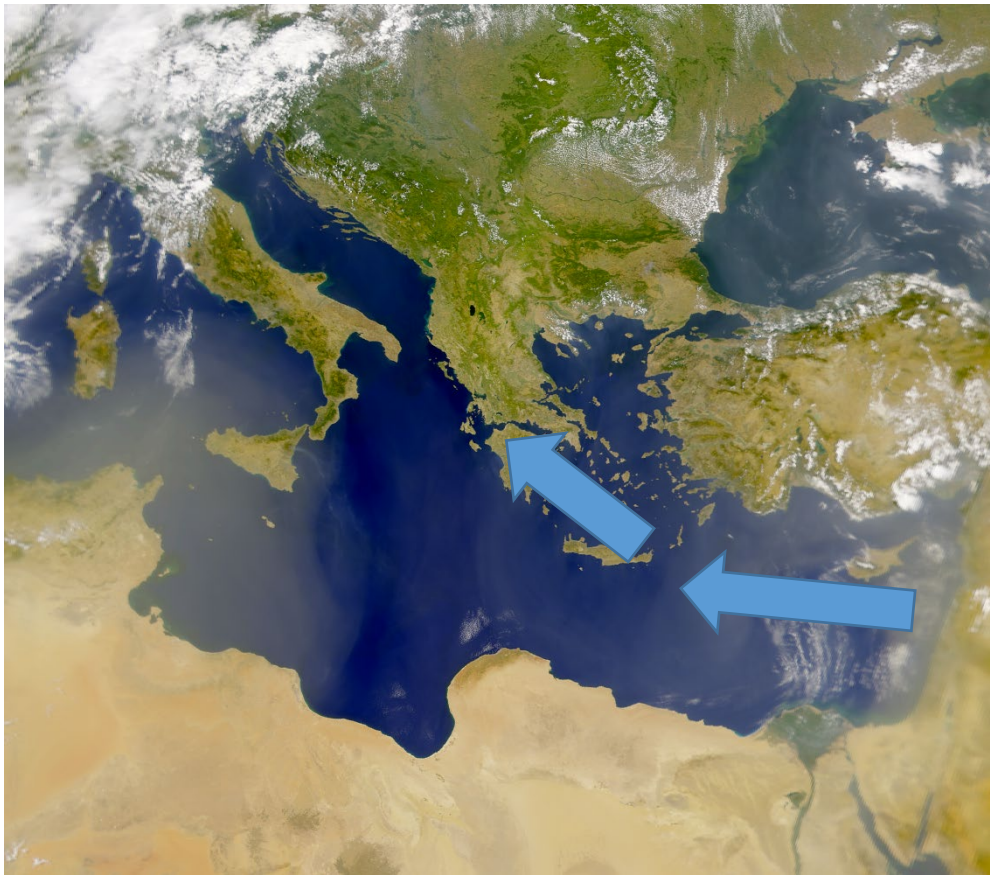
- Possible routes of migration:
  - Anatolia to Bosporus to Thrace

Greece at the crossroads of Neolithic migrations into Europe

- Possible routes of migration:
  - Anatolia to Bosporus to Thrace
  - Maritime route from the coast of Anatolia to the Aegean islands to Southeast Europe

Greece at the crossroads of Neolithic migrations into Europe

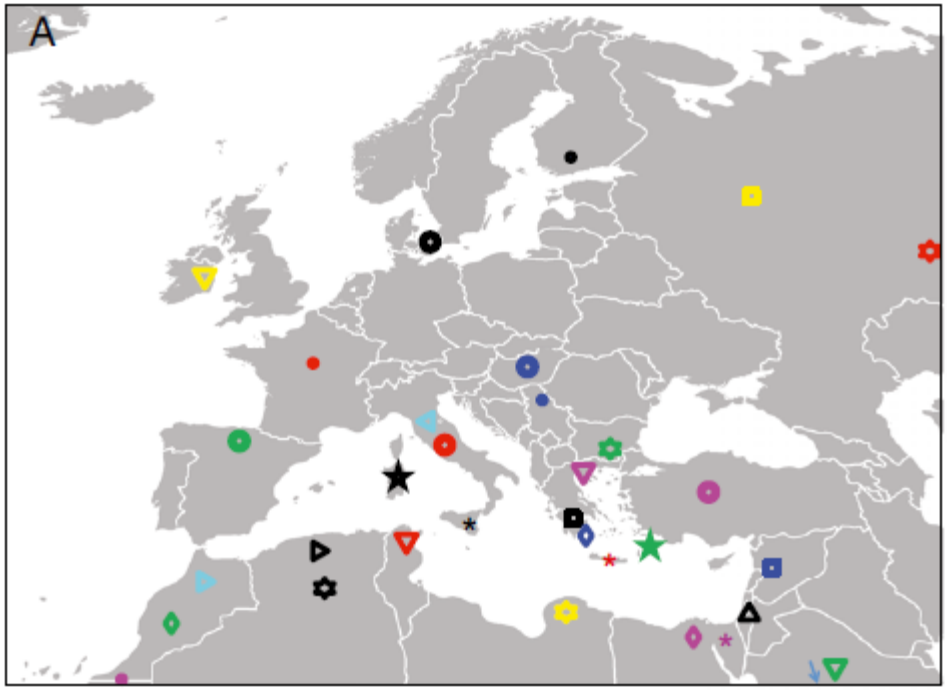
- Possible routes of migration:
  - Anatolia to Bosporus to Thrace
  - Maritime route from the coast of Anatolia to the Aegean islands to Southeast Europe
  - Middle East to the Aegean to Europe



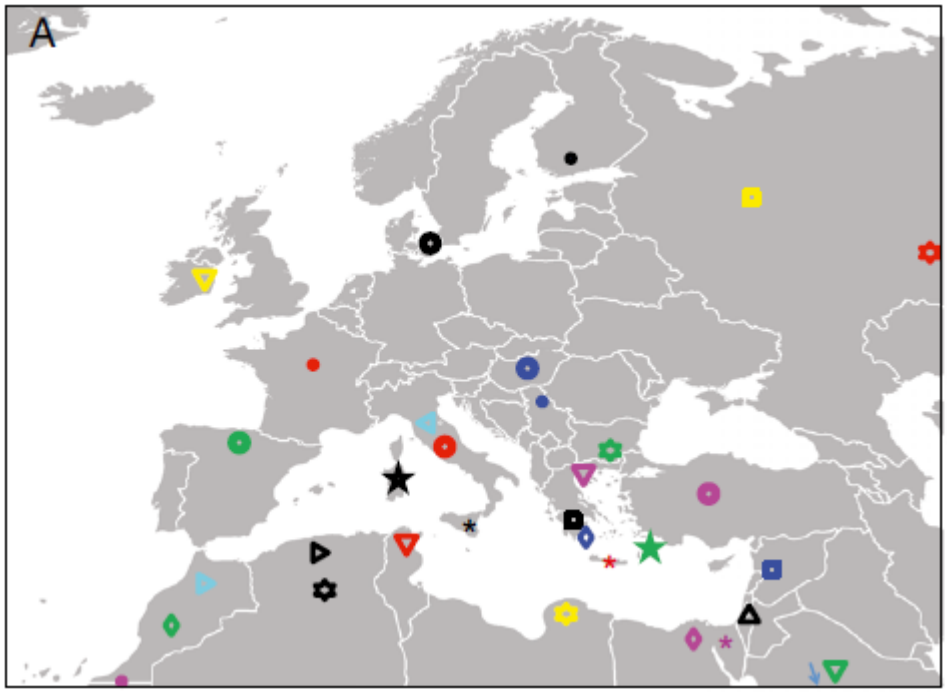
The Data

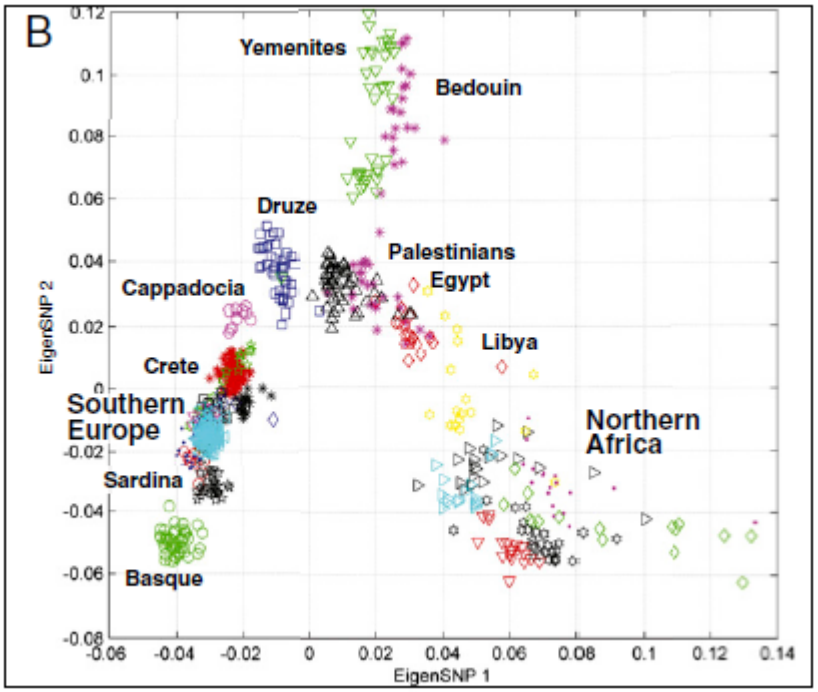
964 samples from 32 populations genotyped across 75,194 SNPs across all autosomes

- Crete, Dodecanese (Aegean islands)
- 3 populations from mainland Greece
- Cappadocia (Anatolia)
- 14 populations from Northern and Southern Europe
- 7 populations from North Africa
- 5 populations from Middle East

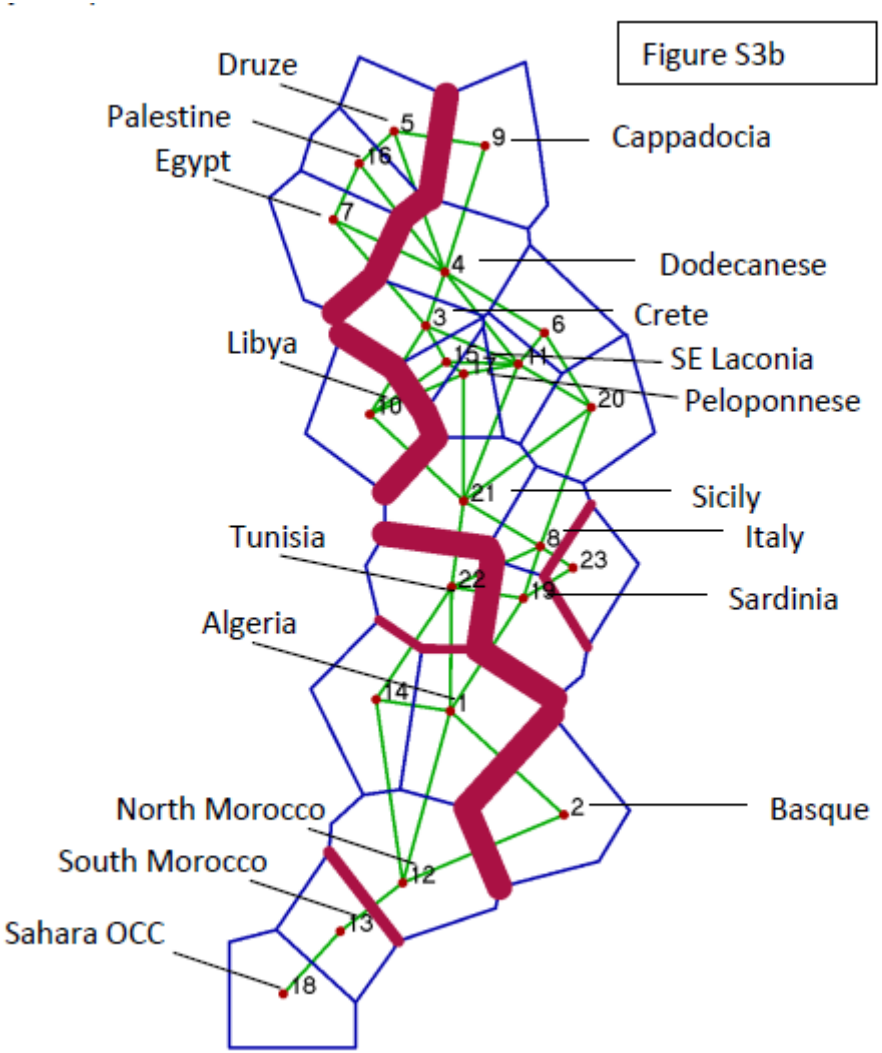


Population genetic structure around the Mediterranean





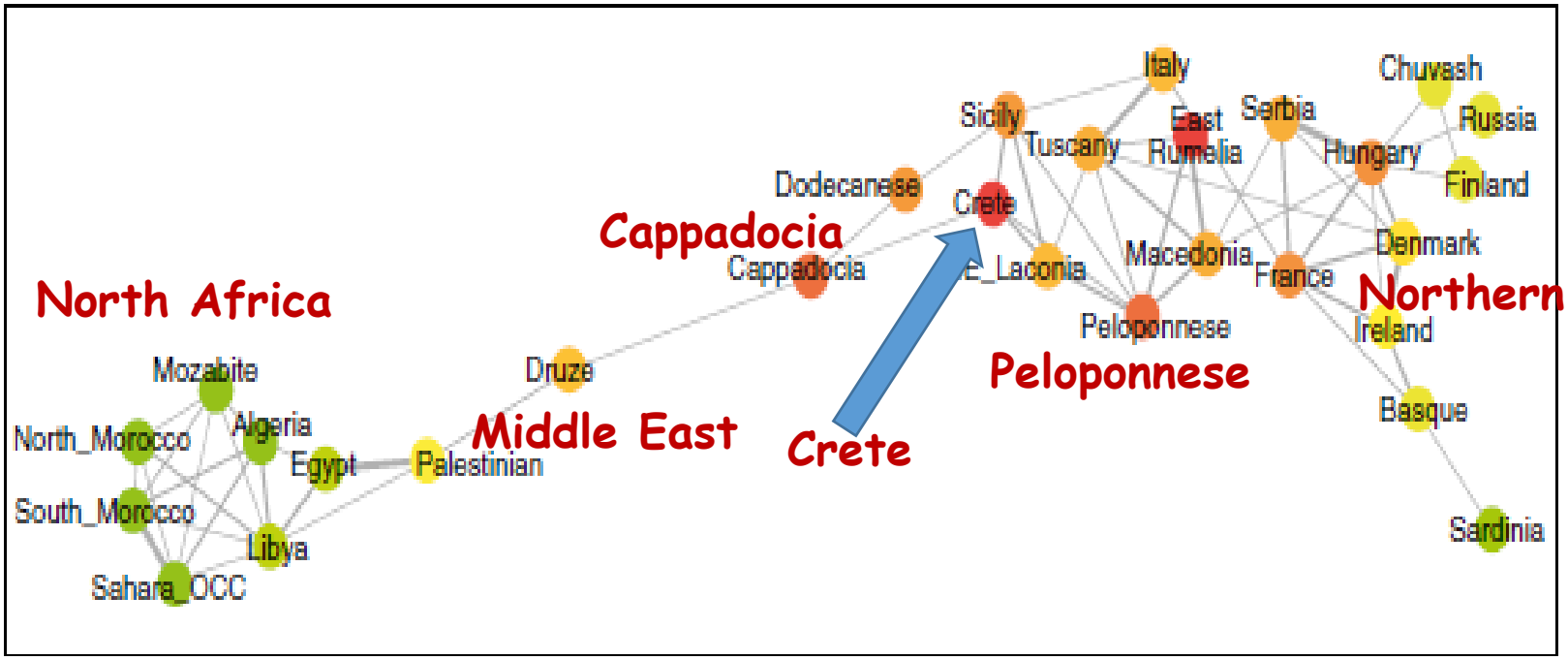
The Mediterranean as a barrier in gene flow



Analysis using BARRIER software (combination of genetic and geographic distances)

Constructing gene flow networks

The islands of Crete and the Dodecanese as a bridge connecting Anatolia to the Southern Peloponnese and the rest of Europe



Neolithic migrations to Europe via a maritime route

- The islands of the Aegean and Crete are important nodes of migration towards Europe in the Neolithic Era.

- The Mediterranean acted as a barrier for migrations to Europe from Northern Africa.

Paschou, Drineas, et al. PNAS 2014

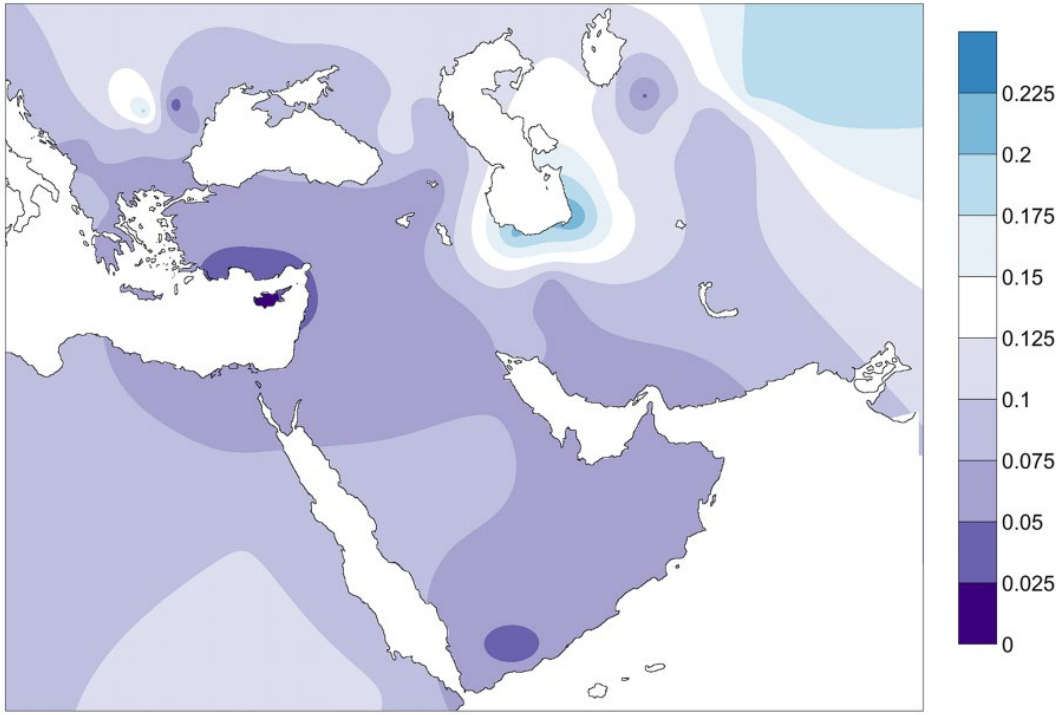
OPEN ACCESS Freely available online

PLOS | GENETICS

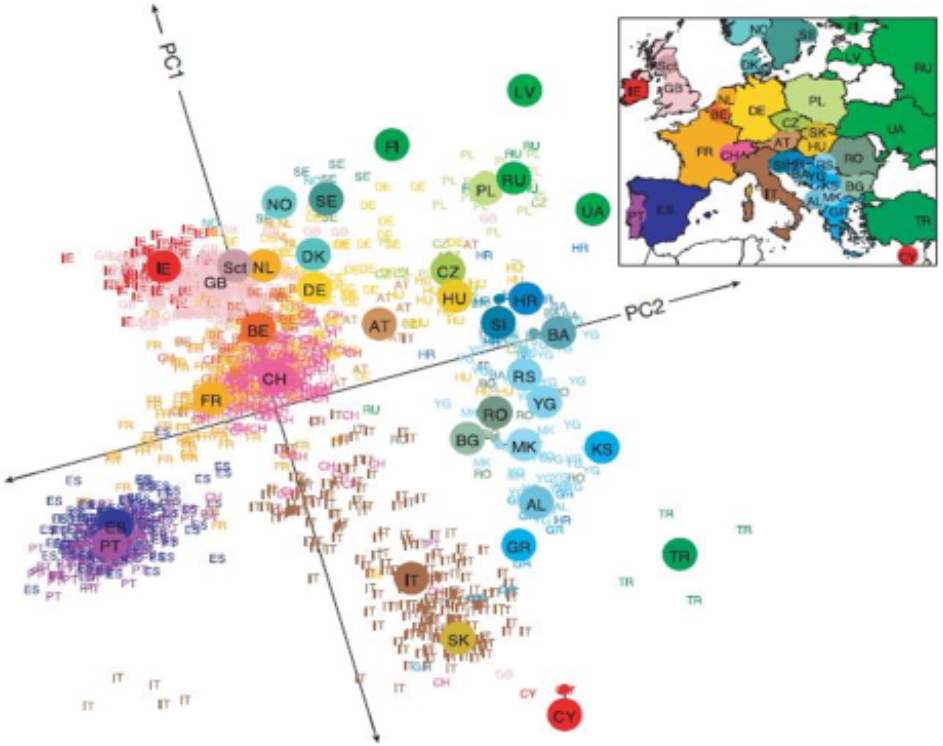
Ancient DNA Analysis of 8000 B.C. Near Eastern Farmers Supports an Early Neolithic Pioneer Maritime Colonization of Mainland Europe through Cyprus and the Aegean Islands

Eva Fernández[1,2]*, Alejandro Pérez-Pérez[3], Cristina Gamba[2], Eva Prats[4], Pedro Cuesta[5], Josep Anfruns[6], Miquel Molist[6], Eduardo Arroyo-Pardo[2], Daniel Turbón[3]

1 Research Centre in Evolutionary Anthropology and Paleoecology, Liverpool John Moores University, Liverpool, United Kingdom, **2** Laboratorio de Genética Forense y Genética de Poblaciones, Dpto. Toxicología y Legislación Sanitaria, Facultad de Medicina, Universidad Complutense de Madrid, Madrid, Spain, **3** Dpto. Biología Animal-Unidad de Antropología, Facultad de Biología, Universitat de Barcelona, Barcelona, Spain, **4** Centro de Investigación y Desarrollo, Consejo Superior de Investigaciones Científicas, Barcelona, Spain, **5** Dpto. de Apoyo a la Investigación, Servicios informáticos de la Universidad Complutense de Madrid, Madrid, Spain, **6** Dep. Prehistoria, Facultad de Filosofía y Letras, Universitat Autónoma de Barcelona, Bellaterra, Barcelona, Spain



PCA of Europeans: Genes mirroring Geography



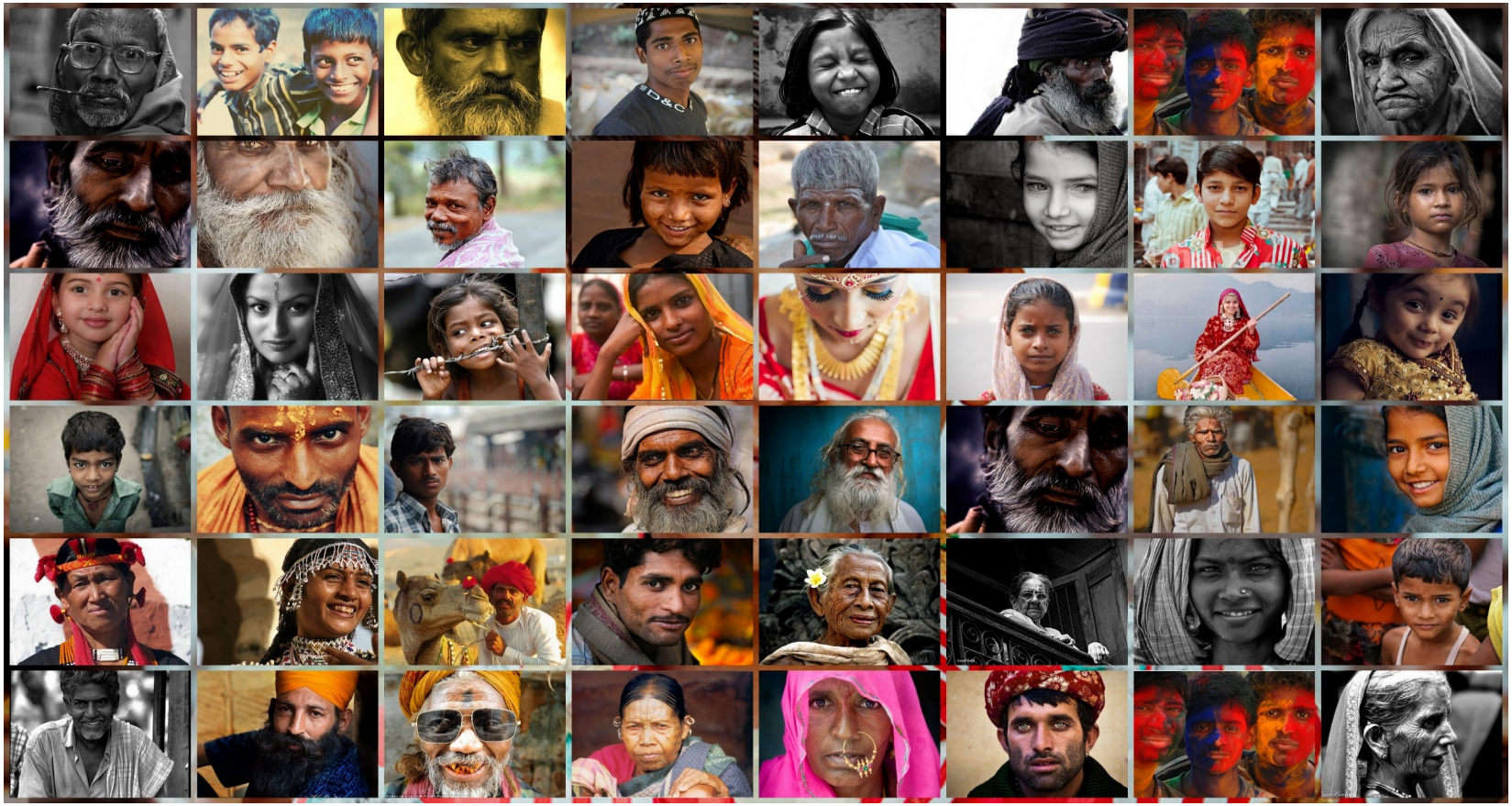
Novembre et al (Nature, 2008) showed the Pearson correlation coefficient, r^2 between the geographical coordinates and the principal components for 197,146 SNPs in 1,387 samples (POPRES project) collected across Europe to be:

0.77 for PC1 v/s Latitude
0.78 for PC2 v/s Longitude

Also in Paschou et al. (PLoS Genet,, 2010, PNAS 2013); Drineas et al. (PLoS One, 2010), Lao et al. (AJHG 2008))

Novembre et al. (Nature, 2008)

What about India?



Bose et al. "Integrating linguistics, social structure, and geography to model genetic diversity within India," Mol Bio & Evo, 2021

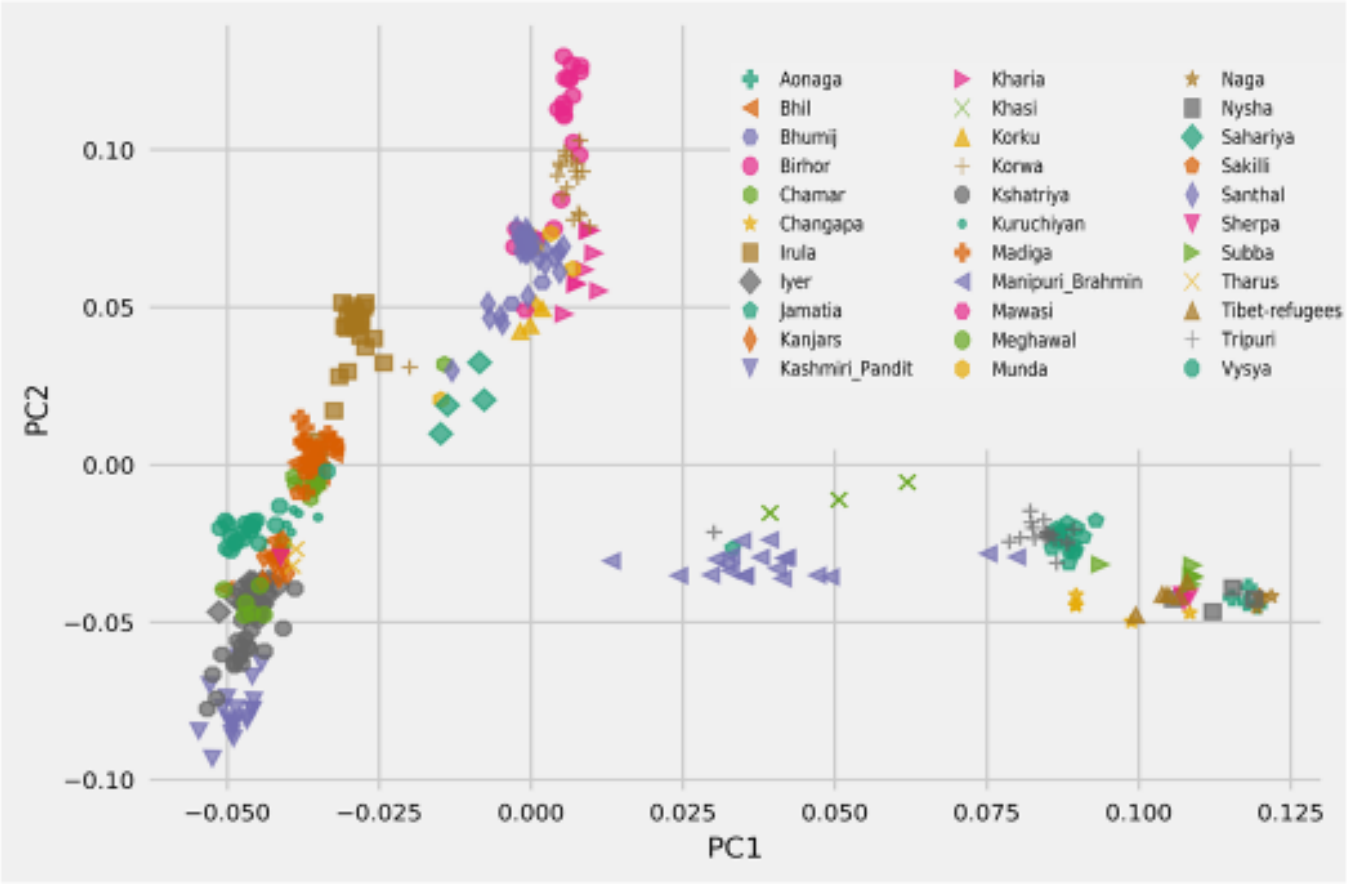
Data collection

Combining data from various sources:

Number of Samples	Number of Populations	Source
142	20	Metspalu et al. (2011)
26	10	Chaubey et al. (2010)
19	4	Behar et al. (2010)
132	10	Reich et al. (2009)
188	20	Moorjani et al. (2013)
367	20	Basu et al. (2016)
835	**84**	

Location of Samples

Map showing the locations of the 835 Indian samples (from 84 well-defined population groups) that were used as the starting point in our study. After QC, a total of 48,373 SNPs were included.

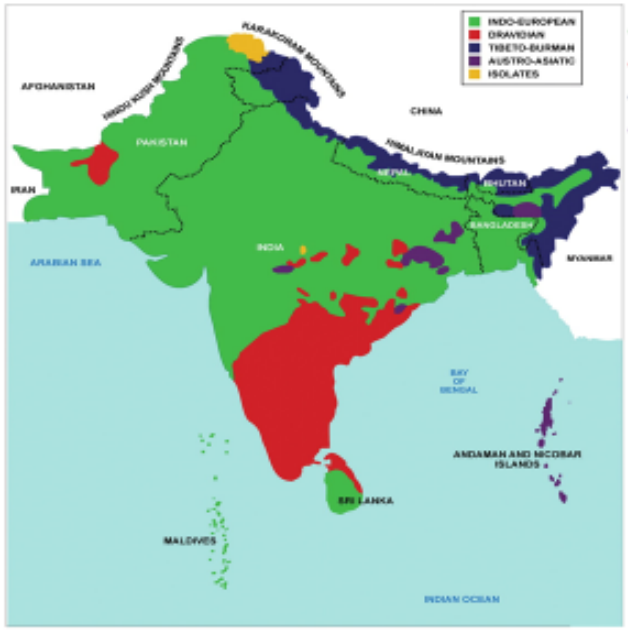


The top two PCs show poor correlation with geography in India

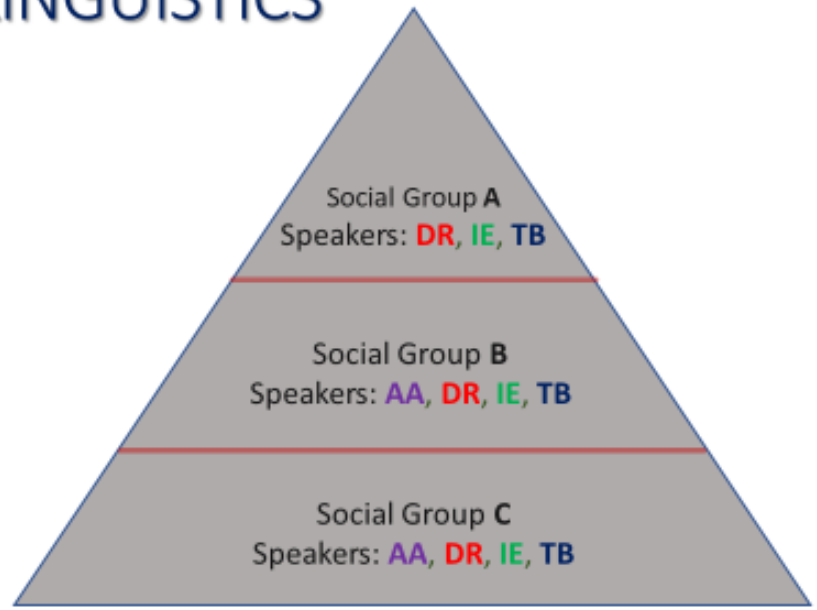
r^2 between the geographical coordinates and the principal components were:

0.6 for PC1 v/s Longitude and 0.06 for PC2 v/s Latitude.

INDIA: SOCIOLINGUISTICS



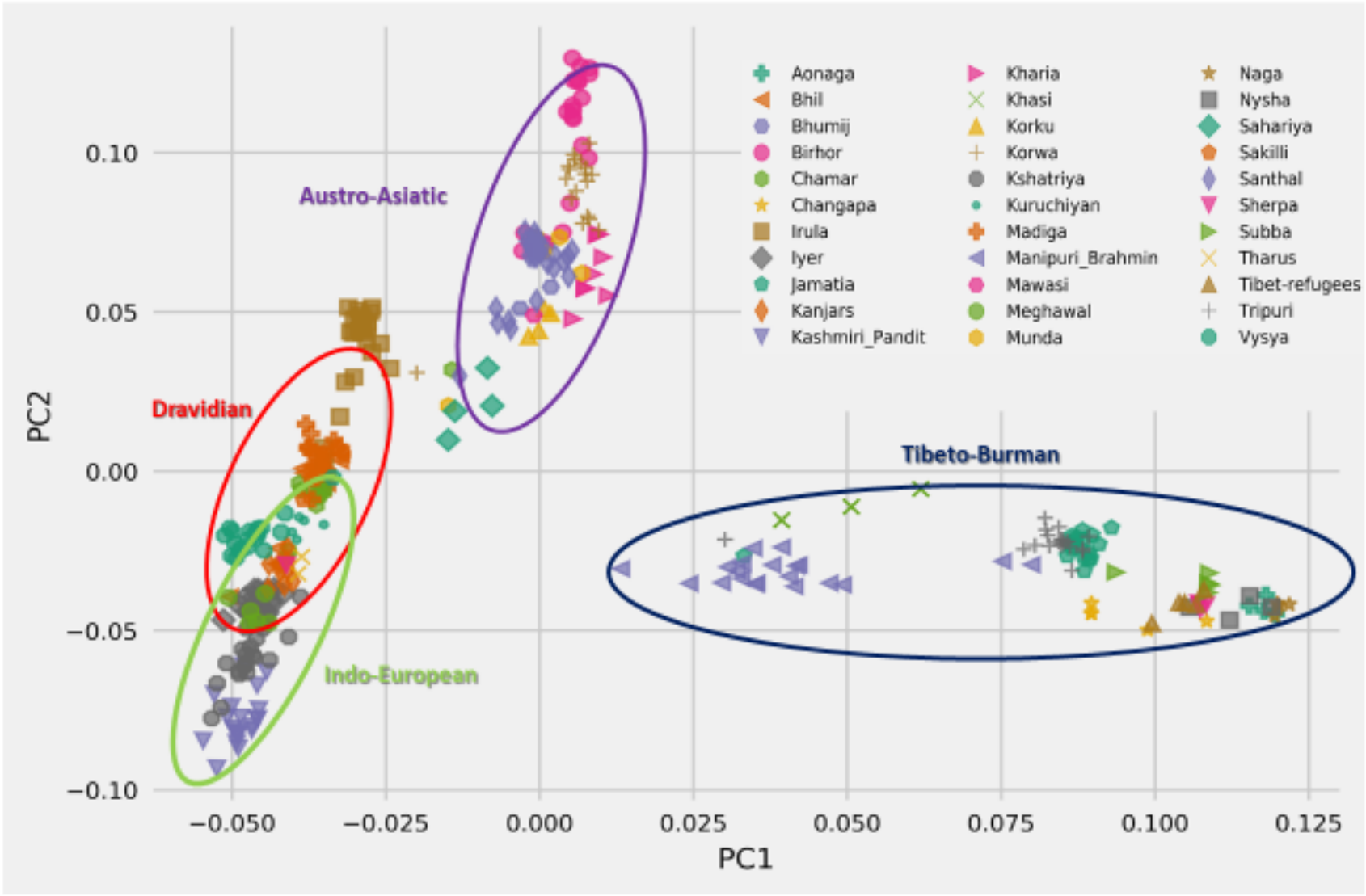
- Indo-European (IE)
- Dravidian (DR)
- Tibeto-Burmese (TB)
- Austro-Asiatic (AA)



According to 2001 census, 29 languages have more than a million native speakers, of which 22 languages are recognized as official, with a total of 1,652 mother tongues spoken across the country.

- Social stratification in terms of Caste System was documented first around 300 BC.
- There are 4,635 well-defined endogamous populations in India with 532 tribal communities constituting ~8% (2001 Census, Govt. of India) of the total population.

Ayub et al. (2009) Genetic Variation in South Asia, Fig 1


Aonaga
Bhil
Bhumij
Birhor
Chamar
Changapa
Irula
Iyer
Jamatia
Kanjars
Kashmiri_Pandit
Kharia
Khasi
Korku
Korwa
Kshatriya
Kuruchiyan
Madiga
Manipuri_Brahmin
Mawasi
Meghawal
Munda
Naga
Nysha
Sahariya
Sakilli
Santhal
Sherpa
Subba
Tharus
Tibet-refugees
Tripuri
Vysya
Austro-Asiatic
Dravidian
Indo-European
Tibeto-Burman
PC1
PC2
0.10
0.05
0.00
−0.05
−0.10
−0.050
−0.025
0.000
0.025
0.050
0.075
0.100
0.125

COGG

$$\max_{\alpha} \mathbf{Corr}\left(\underset{[n\times 1]}{\begin{bmatrix} \vdots \\ eigenSNP \\ \vdots \end{bmatrix}}, \sum_{i=1}^{9} \underset{[n\times 9]}{\left[\begin{array}{c|c|c|c|c|c|c|c|c} \alpha_1 G_1 & \alpha_2 G_2 & \alpha_3 G_3 & \alpha_4 G_4 & \alpha_5 G_5 & \alpha_6 G_6 & \alpha_7 G_7 & \alpha_8 G_8 & \alpha_9 G_9 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Latitude & Longitude & Forward\ Castes\ (SGA) & Backward\ Castes\ (SGB) & Tribals\ (SGC) & AA & DR & IE & TB \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array}\right]} \right)$$

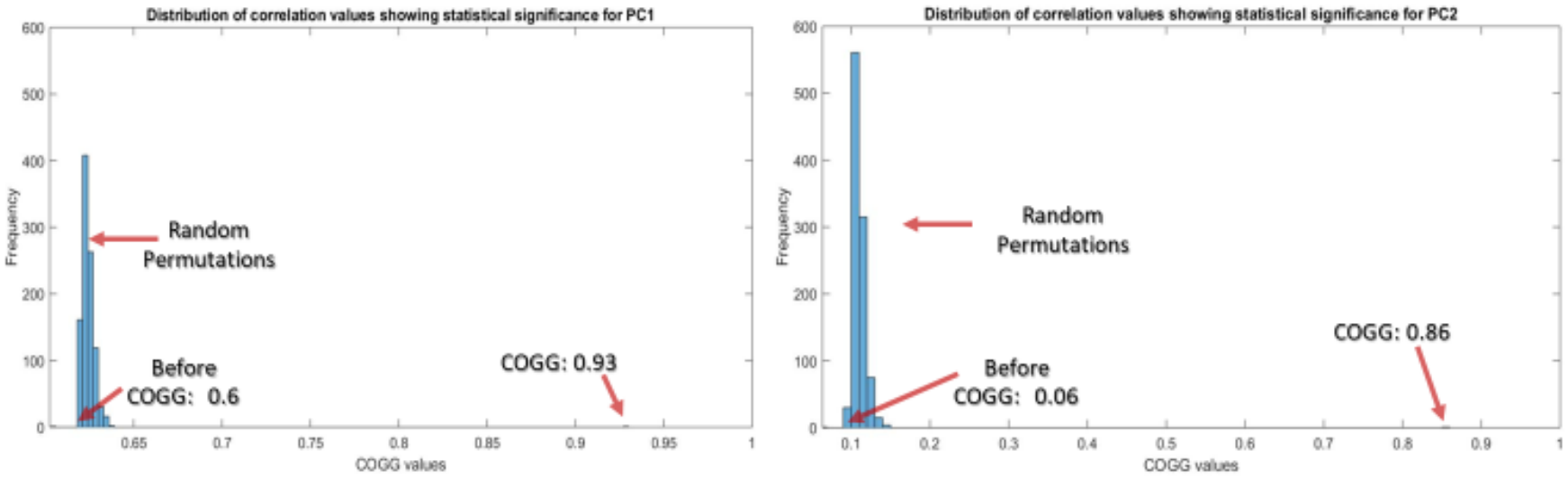
Formally:

$$\max_{\alpha} \mathbf{Corr}\left(U, \sum_{i=1}^{k} \alpha_i \cdot G_i\right)$$

where $U \in \mathbb{R}^n$, is the vector corresponding to the eigenSNPs.
$G \in \mathbb{R}^{n \times k}$, is the Geodemographic matrix.
$\alpha = (\alpha_i)$ is the unknown vector of coefficients for each feature.

A closed form solution exists for the COGG optimization problem: $\alpha = [\mathbf{Var}[U] \cdot \mathbf{Cov}[G_i, G_j]]^{-1} \cdot \mathbf{Cov}[U, G_i]$

COGG

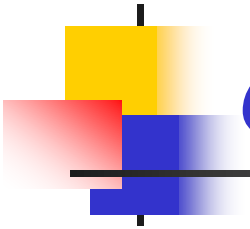


Statistical significance of the COGG output (using random permutations). Clearly, COGG is statistically significant for both the first and the second principal components.

Plugging in the value of α_{max} we get:

0.93 for eigenSNP1 v/s G
0.86 for eigenSNP2 v/s G

For more details: Bose et al. Mol Bio & Evo (2021)



Open questions

Unsupervised dimensionality reduction techniques are NOT successful in separating cases from controls in GWAS studies.

- Why? Because the disease signal is too "weak".
- Potential remedies? Supervised techniques, e.g., GLMs, SVMs, Deep Learning, etc.
- Goal? Supervised dimensionality reduction techniques that identify axes that separate cases from controls. Then, identify SNPs (and genes) that span the same subspace as those axes.
- Looks challenging, especially if the objective is to separate cases and controls (too stringent).
- Maybe relax the objective? Separating averages is too naïve; is there something more interesting?

Acknowledgements

Collaborators

P. Paschou, Purdue
E. Ziv, UCSF
K. K. Kidd, Yale University
M. W. Mahoney, UC Berkeley
J. Stamatoyannopoulos, U Washington
G. Stamatoyannopoulos, U Washington

Students

A. Javed, RPI
J. Lewis, RPI
J. Alexander, RPI
A. Bose, Purdue
F. Tsetsos, Purdue
M. Burch, Purdue
P. Jain, Purdue
Z. Yang, Purdue

Funding: NSF, NIH, DOE, EMBO, IBM, Tourette Syndrome Association, EU FP7 Programme.

Papers and preprints: google Drineas; go to Publications page.