Dimensionality Reduction in the Analysis of Human Genetics Data

Petros Drineas
Purdue University
Department of Computer Science

Google drineas
Principal Components Analysis (PCA) Dimensionality Reduction in the Analysis of Human Genetics Data

Petros Drineas

Purdue University
Department of Computer Science
PCA & Human Genetics: Η Καταγωγή των Ελλήνων

Petros Drineas

Purdue University
Department of Computer Science

Google drineas
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 & Natural Selection
- Genetic drift
- Population structure (inbreeding, mating patterns)
- Gene flow and admixture
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
Homo sapiens sapiens
~40,000 BP

http://info.med.yale.edu/genetics/kkidd/point.html
Out of Africa Hypothesis

Human Migration

- Fossil or artifact site
- 40,000 years ago
- Migration date
- Generalized route

Sources: Susan Antle, New York University; Alison Brooks, George Washington University; Peter Forester, University of Cambridge; James O'Connell, University of Utah; Stephen Oppenheimer, Oxford University; Spencer Wells, National Geographic Society; Didier Goasdoué, Harvard University; National Maps.

© 2006 National Geographic Society. All rights reserved.
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**.

\[ r^2 = 0.77 \text{ for PC1 vs Latitude} \]
\[ r^2 = 0.78 \text{ for PC2 vs Longitude} \]

Novembre et al. (Nature 2008)
The success of PCA in (human) genetics is remarkable!

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

- **Genetics:** Mapping causative genes for common complex disorders
  - Correcting stratification in Genome-Wide Association Studies (GWAS)

- **Genetic history of human populations**

- **Forensics**

- **Genealogy**
Overview

- Genotype data and PCA: definitions and applications
- Scaling PCA to millions of samples/markers
- ΡΧΑ και η καταγωγή των Ελλήνων
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.
Focus at a specific locus and assay the observed nucleotide bases (alleles).

SNP: exactly two alternate alleles appear.
Focus at a specific locus and assay the observed alleles.

SNP: exactly two alternate alleles appear.

Two copies of a chromosome (father, mother)

An individual could be:
- Heterozygotic (in our study, CT = TC)
Focus at a specific locus and assay the observed alleles.

SNP: exactly two alternate alleles appear.

Two copies of a chromosome (father, mother)

An individual could be:
- Heterozygotic (in our studies, CT = TC)
- Homozygotic at the first allele, e.g., C
Focus at a specific locus and assay the observed alleles.

SNP: exactly two alternate alleles appear.

Two copies of a chromosome (father, mother)

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

... AG CT GT GG CT CC CC CC AG AG AG AA CT AA GG GG CC GG AG AG AC CC AA CC AA GG TT AG CT CG CG AT CT CT AG CT AG CT AG CT AG GT GA AG ...
... GG TT TT GG TT CC CC CC CC GG AA AG AG AA CT AA GG GG CC GG AA AA CC AA CC AA GG TT AA TT GG GG GG TT TT TT CC GG TT GG GG TT GG AA ...
... GG TT TT GG TT CC CC CC CC GG AA AG AG AA CT AA GG GG CC AG AG CG AC CC AA CC AA GG TT AG CT CG CG AT CT CT AG CT AG CT AG GT GA AG ...
... GG TT TT GG TT CC CC CC CC GG AA AG AG AA CC AA CC CC AA CG AA GG TT AG CT CG CG AT CT CT AG CT AG CT AG GT GA AG ...
... GG TT TT GG TT CC CC CC CC GG AA AG GG GA GT TT GG AA ...
... GG TT TT GG TT CC CC CC CC 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 ...
HGDP data
- 1,033 samples
- 7 geographic regions
- 52 populations

HapMap Phase 3 data
- 1,207 samples
- 11 populations

Rosenberg et al. (2002) Science
Li et al. (2008) Science
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
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 ...
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,
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

$\sigma_1$: measures how much of the data variance is explained by the first singular vector.

$\sigma_2$: 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**

\[
\begin{pmatrix}
    A \\
    m \times n
\end{pmatrix}
= 
\begin{pmatrix}
    U \\
    m \times \rho
\end{pmatrix}
\cdot
\begin{pmatrix}
    \Sigma \\
    \rho \times \rho
\end{pmatrix}
\cdot
\begin{pmatrix}
    V \\
    \rho \times n
\end{pmatrix}^T
\]

\( \rho \): rank of \( A \)

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

\( \Sigma \): diagonal matrix containing the singular values of \( A \).

Let \( \sigma_1, \sigma_2, \ldots, \sigma_\rho \) be the entries of \( \Sigma \).

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 via the SVD

\[ A = U \Sigma V^T \]
HGDP data
- 1,033 samples
- 7 geographic regions
- 52 populations

Rosenberg et al. (2002) Science
Li et al. (2008) Science

HapMap Phase 3 data
- 1,207 samples
- 11 populations

Matrix dimensions:
2,240 subjects (rows)
447,143 SNPs (columns)

SVD/PCA returns...
Top two Principal Components (eigenSNPs)

Mexican population seems out of place: we move to the top three PCs.
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.**
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?

Boils down to the so-called Column Subset Selection Problem (CSSP).

- See work by CEID graduate Christos Boutsidis, who is now "The exceptional European VP with a special task at Goldman Sachs in NY."
- His successor in my group is another CEID graduate, Christos Boutsikas.
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 $A A^T$.
- 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 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 $A^T A$.

- **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 SNPs.


*CEID graduates
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 $X = (AA^T)^p X_0$.

- A lot of theory for these methods, as well as the closely related Krylov subspace methods.
Subspace Iteration: Methods

- **Given** $A \in \mathbb{R}^{m \times n}, X_0 \in \mathbb{R}^{m \times k}$, and integer $p$, **Subspace Iteration** computes $X_p = (AA^T)^p X_0 \in \mathbb{R}^{m \times k}$ and uses it to approximate the top singular vectors of $A$.

- **Krylov subspace methods**: Keep all intermediate iterates $(AA^T)^t X_0$ for $t = 1 \ldots p$ and use the set of vectors $[(AA^T)X_0, (AA^T)^2 X_0, \ldots, (AA^T)^p X_0]$, a total of $k \cdot p$ vectors.

- Both use **matvecs**, but Krylov subspace methods end up with a “larger” subspace. What is the difference?

- Let $\sigma_k \geq (1 + \gamma) \sigma_{k+1}$ for some positive constant $\gamma$, which could be quite close to zero. This is often called the “spectral gap” between the $k$-th and the $(k + 1)$-st singular values.
Subspace Iteration: Methods

- Both use matvecs, but Krylov subspace methods end up with a “larger” subspace. What is the difference?

- Let $\sigma_k \geq (1 + \gamma) \sigma_{k+1}$ for some positive constant $\gamma$, which could be quite close to zero. This is often called the “spectral gap” between the $k$-th and the $(k+1)$-st singular values.

- Then, to achieve roughly comparable approximation accuracy,
  - Subspace Iteration needs $p = O\left(\frac{1}{\gamma}\right)$ iterations.
  - Krylov Subspace Methods need $p = O\left(\frac{1}{\sqrt{\gamma}}\right)$ iterations.

- This is reminiscent of the improvement of, say, Conjugate Gradient linear equation solvers over naïve solvers (like Richardson).

- There is a deep connection between Conjugate Gradient linear equation solvers and Krylov Subspace Methods.
Subspace Iteration: Methods

- Then, to achieve roughly comparable approximation accuracy,
  - Subspace Iteration needs $p = O \left( \frac{1}{\sqrt{\nu}} \right)$ iterations.
  - Krylov Subspace Methods need $p = O \left( \frac{1}{\sqrt{\nu}} \right)$ iterations.

- This is reminiscent of the improvement of, say, Conjugate Gradient linear equation solvers over naïve solvers (like Richardson).
- There is a deep connection between Conjugate Gradient linear equation solvers and Krylov Subspace Methods.

References:
Drineas, Ipsen, Kontopoulou, & Magdon-Ismail SIMAX 2019
Drineas & Ipsen SIMAX 2020
Musco & Musco NeurIPS 2015
Subspace Iteration: Algorithm

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

1: $C = A(A^T X_0)$
2: repeat
3: For $i = 2$ to $p$ do
4: $Q = \text{orth}(C)$
5: $C = A(A^T Q)$
6: End For
7: $Q = \text{orth}(C)$
8: $C = AA^T Q$
9: $M = Q^T C$
10: Compute eigenvalue decomposition $M = PDP^T$
11: $C = QP$
12: until convergence
13: return $Q$
Comments:

1. For massive matrices, the problem is lack of RAM
2. Software engineering problem and not really a Numerical Linear Algebra problem
Computing $A(A^T X_0)$

We compute $C$ as follows: 

$$C = b_1^T b_1 X_{k-1} + b_2^T b_2 X_{k-1} + b_3^T b_3 X_{k-1} + \ldots + b_\beta^T b_\beta X_{k-1},$$

where $\beta$ 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:

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Size</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGDP</td>
<td>6 GB</td>
<td>1,043 individuals, 107,468 markers</td>
</tr>
<tr>
<td>1000 Genomes</td>
<td>38 GB</td>
<td>2,504 individuals, 808,647 markers</td>
</tr>
<tr>
<td>5K -by- 1M</td>
<td>19 GB</td>
<td>5,000 individuals, 1,000,000 markers</td>
</tr>
<tr>
<td>10K -by- 1M</td>
<td>38 GB</td>
<td>10,000 individuals, 1,000,000 markers</td>
</tr>
<tr>
<td>100K -by- 1M</td>
<td>373 GB</td>
<td>100,000 individuals, 1,000,000 markers</td>
</tr>
<tr>
<td>500K -by- 1M</td>
<td>1.9 TB</td>
<td>500,000 individuals, 1,000,000 markers</td>
</tr>
<tr>
<td>1M -by- 1M</td>
<td>3.7 TB</td>
<td>1,000,000 individuals, 1,000,000 markers</td>
</tr>
</tbody>
</table>

All computations were done in a single core Intel Xeon-Gold processor with 96 GB max RAM.
## TeraPCA: Performance Comparisons

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

* n/a: not applicable as FlashPCA2 did not terminate in 75 hrs
TeraPCA scales “decently” with increasing number of threads.

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

Recent progress by Li et al.: “PCAone: fast and accurate out-of-core PCA framework for large-scale biobank data” (https://www.biorxiv.org/content/10.1101/2022.05.25.493261v1)
Part 2: PCA plots & Greek DNA

Genetic analyses elucidate population relationships and provide answers to historical questions of relevance to archeology and paleoanthropology.

Again, PCA plots are quite telling.

Examples from our own work, focusing on the Greek population:

• 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.
  

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).
The George Stamatoyannopoulos Hellenic DNA collection
Documenting the Hellenic genetic heritage

A historic record of DNA and genomic data from Greece and the Greek diaspora

• Modern DNA
  Participants’ age >70

Known origin of all four grandparents from specific geographic region

Mainland Greece, Greek islands, Crete, Cyprus, Sarakatsanoi, Vlachs, Pontos, Cappadocia, Minor Asia

• Ancient DNA

Mycenaean era, Minoan era
Events that shaped European genomic variation:
Early migrations into Europe
Neolithic migrations from the Fertile Crescent

Modern Europeans are a result of admixture:

• Paleolithic inhabitants (35,000-40,000 years before present)
• Neolithic inhabitants (9,000 years before present)
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*
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 Fernández¹,²*, Alejandro Pérez-Pérez³, Cristina Gamba², Eva Prats⁴, Pedro Cuesta⁵, Josep Anfruns⁶, Miquel Molist⁶, Eduardo Arroyo-Pardo², Daniel Turbón³

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

• Neolithic colonization of Crete - 9,000 years before present

• The first European civilization established in Crete during the Early Bronze Age
The Minoans
The first advanced European civilization

- Three scenarios regarding the origin of the Minoans:
  - Refugees from Northern Egypt (Sir Arthur Evans)
  - Migration from the Cycladic islands or the Middle East
  - Established by existing inhabitants of the island
A European population in Minoan Bronze Age Crete

Jeffery R. Hughey¹, Peristera Paschou², Petros Drineas³, Donald Mastropaolo⁴, Dimitra M. Lotakis⁴, Patrick A. Navas⁴, Manolis Michalodimitrakis⁵, John A. Stamatoyannopoulos⁶ & George Stamatoyannopoulos⁶
The Minoan samples

- Two Minoan populations
  - 39 individuals from tombs near Odigitria Monastery, dating from early Minoan period I (~4900 ybp) to Middle Minoan period IB (~3800 ybp)
  - 69 individuals from a cave near Lasithi dating to Middle Minoan IIB (~3700 ybp)
- Analysis was possible for 37 samples from the Lassithi cave

*Hughey, Paschou, Drineas et al. Nature Communications 2013*
Percentage of haplotypes shared between Minoans and 71 extant populations

- 21 Minoan haplotypes - 6 unique to the Minoans
- 15 shared with modern and other ancient populations that we studied
- No African haplotypes observed in the Minoans
- Minoan haplotypes most similar to European haplotypes

*Hughey, Paschou, Drineas et al. Nature Communications 2013*
Africans are the most distant neighbors to the Minoans

Hughey, Paschou, Drineas et al. Nature Communications 2013
Modern day Lassithi inhabitants are the nearest neighbors to the Minoans!

Hughey, Paschou, Drineas et al. Nature Communications 2013
Genetic History of the Population of Crete

17 extant Cretan populations studied with Illumina 1M or 2.5M arrays

Drineas et al. Annals of Human Genetics, 2019
High correlation with geographic coordinates (east to west axis)
ARTICLE

Genetics of the peloponnesian populations and the theory of extinction of the medieval peloponnesian Greeks

George Stamatoyannopoulos*,1, Aritra Bose2, Athanasios Teodosiadis3, Fotis Tsetsos2, Anna Plantinga4, Nikoletta Psatha5, Nikos Zogas6, Evangelia Yannaki6, Pierre Zalloua7, Kenneth K Kidd8, Brian L Browning4,9, John Stamatoyannopoulos3,10, Peristera Paschou11 and Petros Drineas2
Peloponnese
A history of migrations

- Neolithic sites established by early migrants arrived from Anatolia ca 9000 BC.
- The Mycenaeans (advanced Bronze Era), either migrated from the north around 2200 BC or were the descendants of the original Neolithic migrants.
- Invasion of Peloponnese by the Dorian Greeks (1000 BC)
- Beginning of the medieval period - migrations of the Slavs to the Balkans.
Debating the theory of extinction of the medieval Peloponnesean Greeks

Jacob Philipp Fallmerayer
Debating the theory of extinction of the medieval Peloponnesian Greeks

“The race of the Hellenes has been wiped out in Europe... Not the slightest drop of undiluted Hellenic blood flows in the veins of the Christian population of present-day Greece.” Fallmerayer, 1830
Debating the theory of extinction of the medieval Peloponnesean Greeks

Constantine Paparrigopoulous
Genetics of Peloponnesian populations
241 individuals - 2.5 million SNPs

Stamatoyannopoulos et al 2017
Peloponnese in comparison to other European populations

Stamatoyannopoulos et al 2017
Peloponnese in comparison to other European populations

Stamatoyannopoulos et al 2017
Peloponnese in comparison to other European populations

Stamatoyannopoulos et al 2017
Peloponnese in comparison to other European populations

Stamatoyannopoulos et al. 2017
Peloponnese in comparison to Slavs and other Southern Europeans

Stamatoyannopoulos et al 2017
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)
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:

<table>
<thead>
<tr>
<th>Number of Samples</th>
<th>Number of Populations</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>142</td>
<td>20</td>
<td>Metspalu et al. (2011)</td>
</tr>
<tr>
<td>26</td>
<td>10</td>
<td>Chaubey et al. (2010)</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>Behar et al. (2010)</td>
</tr>
<tr>
<td>132</td>
<td>10</td>
<td>Reich et al. (2009)</td>
</tr>
<tr>
<td>188</td>
<td>20</td>
<td>Moorjani et al. (2013)</td>
</tr>
<tr>
<td>367</td>
<td>20</td>
<td>Basu et al. (2016)</td>
</tr>
<tr>
<td>835</td>
<td>84</td>
<td></td>
</tr>
</tbody>
</table>
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*
COGG

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

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:

\[ \alpha = \left[ \text{Var} [U] \cdot \text{Cov} [G_i, G_j] \right]^{-1} \cdot \text{Cov} [U, G_i] \]
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 $\alpha_{max}$ 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)
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

**Students**

A. Javed, RPI
J. Lewis, RPI
C. Boutsidis, RPI
A. Zouzias, IBM
A. Bose, Purdue
C. Boutsikas, Purdue
M. Burch, Purdue
A. Chowdhuri, Purdue
V. Georgiou, KIT
P. Jain, Purdue
E. Kontopoulou, Purdue
F. Tsetsos, Purdue
Z. Yang, Purdue

**Funding:**

- NSF
- NIH
- DOE
- IBM
- EMBO
- EU FP7 Programme
- GSRT