Principal Component Analyses of Genetic Data

Loic Yengo, PhD
Institute for Molecular Bioscience
The University of Queensland
l.yengo@imb.uq.edu.au

International Statistical Genetics Workshop • Boulder, Colorado •
IBG

Outline

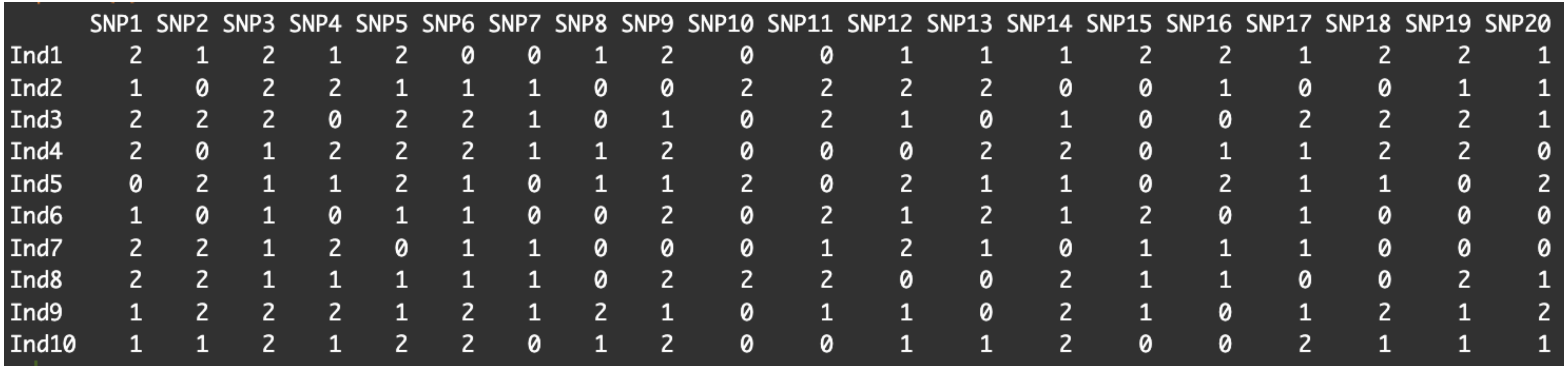
- What is PCA?
- What do we get when applying PCA to genetic data?
- Using PCA to correct confounding in association studies
- Practical considerations about PCA
- More on PCA…

What is PCA?

Origin: Karl Pearson (1901)

Harold Hotelling (1930)

Intuition 1: Visualization



```
      SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11 SNP12 SNP13 SNP14 SNP15 SNP16 SNP17 SNP18 SNP19 SNP20
Ind1     2    1    2    1    2    0    0    1    2     0     0     1     1     1     2     2     1     2     2     1
Ind2     1    0    2    2    1    1    1    0    0     2     2     2     2     0     0     1     0     0     1     1
Ind3     2    2    2    0    2    2    1    0    1     0     2     1     0     1     0     0     2     2     2     1
Ind4     2    0    1    2    2    2    1    1    2     0     0     0     2     2     0     1     1     2     2     0
Ind5     0    2    1    1    2    1    0    1    1     2     0     2     1     1     0     2     1     1     0     2
Ind6     1    0    1    0    1    1    0    0    2     0     2     1     2     1     2     0     1     0     0     0
Ind7     2    2    1    2    0    1    1    0    0     0     1     2     1     0     1     1     1     0     0     0
Ind8     2    2    1    1    1    1    1    0    2     2     2     0     0     2     1     1     0     0     2     1
Ind9     1    2    2    2    1    2    1    2    1     0     1     1     0     2     1     0     1     2     1     2
Ind10    1    1    2    1    2    2    0    1    2     0     0     1     1     2     0     0     2     1     1     1
```

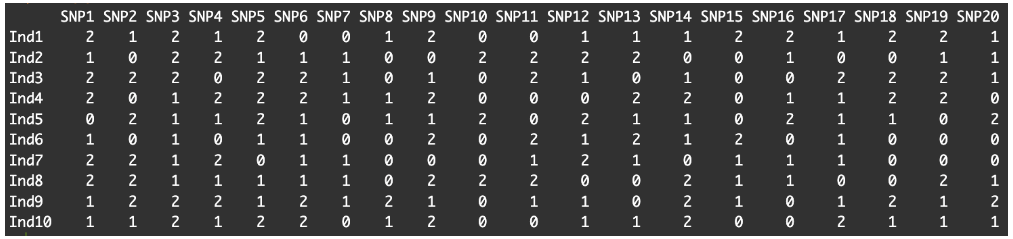
How do we visualize this data?

Yes! We need to summarize the data somehow to fit into 2D or 3D space?
So, how?

PCA solution: “Combine the data **linearly** to **maximize the separation** between **data points**.”

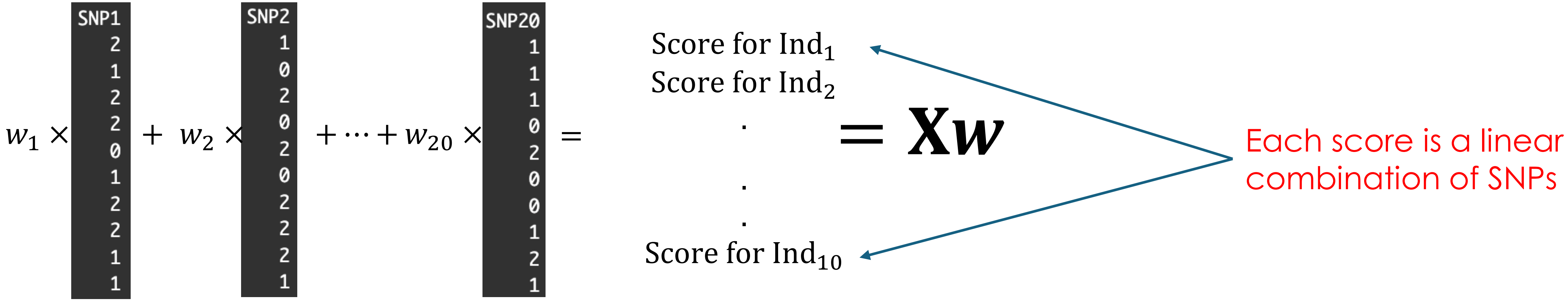
Intuition 1: Visualization

What does **linearly** mean?



```
      SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11 SNP12 SNP13 SNP14 SNP15 SNP16 SNP17 SNP18 SNP19 SNP20
Ind1     2    1    2    1    2    0    0    1    2     0     0     1     1     1     2     2     1     2     2     1
Ind2     1    0    2    2    1    1    1    0    0     2     2     2     2     0     0     1     0     0     1     1
Ind3     2    2    2    0    2    2    1    0    1     0     2     1     0     1     0     0     2     2     2     1
Ind4     2    0    1    2    2    2    1    1    2     0     0     0     2     2     0     1     1     2     2     0
Ind5     0    2    1    1    2    1    0    1    1     2     0     2     1     1     0     2     1     1     0     2
Ind6     1    0    1    0    1    1    0    0    2     0     2     1     2     1     2     0     1     0     0     0
Ind7     2    2    1    2    0    1    1    0    0     0     1     2     1     0     1     1     1     0     0     0
Ind8     2    2    1    1    1    1    1    0    2     2     2     0     0     2     1     1     0     0     2     1
Ind9     1    2    2    2    1    2    1    2    1     0     1     1     0     2     1     0     1     2     1     2
Ind10    1    1    2    1    2    2    0    1    2     0     0     1     1     2     0     0     2     1     1     1
```

If your data point is a "row" (here an individual)

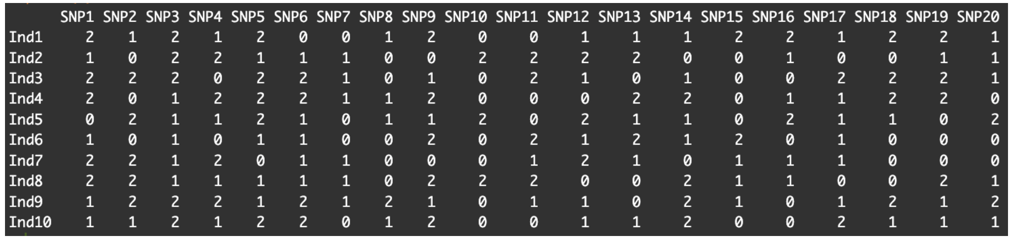


$$w_1 \times \begin{array}{r} \text{SNP1} \\ 2 \\ 1 \\ 2 \\ 2 \\ 0 \\ 1 \\ 2 \\ 2 \\ 1 \\ 1 \end{array} + \; w_2 \times \begin{array}{r} \text{SNP2} \\ 1 \\ 0 \\ 2 \\ 0 \\ 2 \\ 0 \\ 2 \\ 2 \\ 2 \\ 1 \end{array} + \cdots + w_{20} \times \begin{array}{r} \text{SNP20} \\ 1 \\ 1 \\ 1 \\ 0 \\ 2 \\ 0 \\ 0 \\ 1 \\ 2 \\ 1 \end{array} = \begin{array}{c} \text{Score for Ind}_1 \\ \text{Score for Ind}_2 \\ \cdot \\ \cdot \\ \cdot \\ \text{Score for Ind}_{10} \end{array} = \mathbf{X}\boldsymbol{w}$$

Each score is a linear combination of SNPs

Intuition 1: Visualization

What does “**maximize the separation**” mean?



```
      SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11 SNP12 SNP13 SNP14 SNP15 SNP16 SNP17 SNP18 SNP19 SNP20
Ind1     2    1    2    1    2    0    0    1    2     0     0     1     1     1     2     2     1     2     2     1
Ind2     1    0    2    2    1    1    1    0    0     2     2     2     2     0     0     1     0     0     1     1
Ind3     2    2    2    0    2    2    1    0    1     0     2     1     0     1     0     0     2     2     2     1
Ind4     2    0    1    2    2    2    1    1    2     0     0     0     2     2     0     1     1     2     2     0
Ind5     0    2    1    1    2    1    0    1    1     2     0     2     1     1     0     2     1     1     0     2
Ind6     1    0    1    0    1    1    0    0    2     0     2     1     2     1     2     0     1     0     0     0
Ind7     2    2    1    2    0    1    1    0    0     0     1     2     1     0     1     1     1     0     0     0
Ind8     2    2    1    1    1    1    1    0    2     2     2     0     0     2     1     1     0     0     2     1
Ind9     1    2    2    2    1    2    1    2    1     0     1     1     0     2     1     0     1     2     1     2
Ind10    1    1    2    1    2    2    0    1    2     0     0     1     1     2     0     0     2     1     1     1
```

$$w_1 \times \begin{array}{c} \text{SNP1} \\ 2 \\ 1 \\ 2 \\ 2 \\ 0 \\ 1 \\ 2 \\ 2 \\ 1 \\ 1 \end{array} + \; w_2 \times \begin{array}{c} \text{SNP2} \\ 1 \\ 0 \\ 2 \\ 0 \\ 2 \\ 0 \\ 2 \\ 2 \\ 2 \\ 1 \end{array} + \cdots + w_{20} \times \begin{array}{c} \text{SNP20} \\ 1 \\ 1 \\ 1 \\ 0 \\ 2 \\ 0 \\ 0 \\ 1 \\ 2 \\ 1 \end{array} = \begin{array}{c} \text{Score for Ind}_1 \\ \text{Score for Ind}_2 \\ \cdot \\ \cdot \\ \cdot \\ \text{Score for Ind}_{10} \end{array}$$

Choose the weights (w_1, w_2,…, w_{20}) such that the variance of the “**score**” is maximal.

Formally!

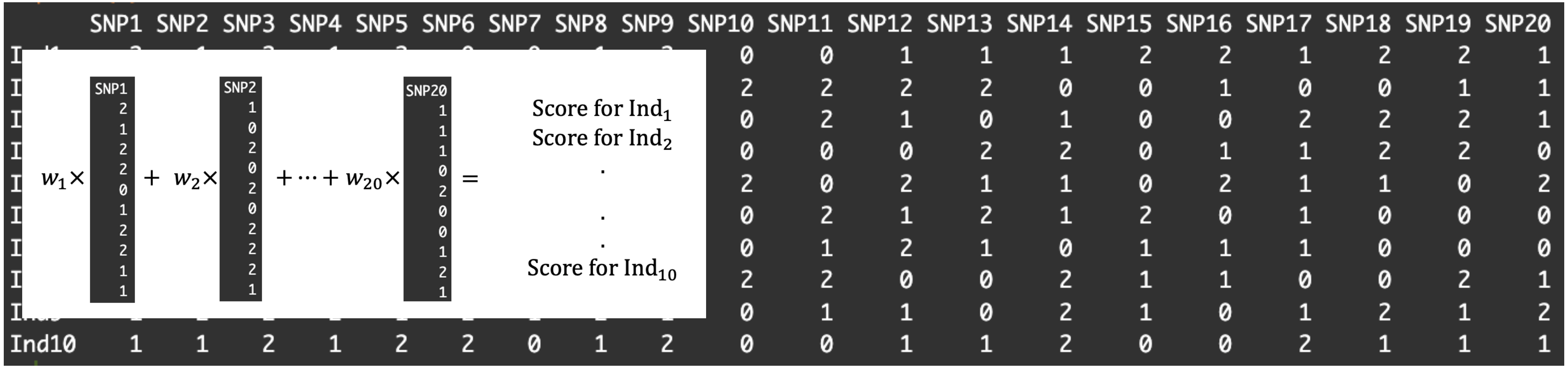
PCA solution: "Combine the data **linearly** to **maximize the separation** between **data points**."

$$\boldsymbol{w}_1 \equiv \underset{\|\boldsymbol{w}_1\|=1}{\operatorname{argmax}}\{\boldsymbol{w}_1'\mathbf{X}'\mathbf{X}\boldsymbol{w}_1\} = \operatorname{argmax}\left\{\frac{\boldsymbol{w}_1'\mathbf{X}'\mathbf{X}\boldsymbol{w}_1}{\boldsymbol{w}_1'\boldsymbol{w}_1}\right\}$$

Definition: The weights $\boldsymbol{w}$ are called principal components (PC) loadings.

Note: PC1 is "uniquely" defined except for the sign.

Intuition 2: Dimension Reduction



```
     SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11 SNP12 SNP13 SNP14 SNP15 SNP16 SNP17 SNP18 SNP19 SNP20
I                                                 0     0     1     1     1     2     2     1     2     2     1
I                                                 2     2     2     2     0     0     1     0     0     1     1
I                                                 0     2     1     0     1     0     0     2     2     2     1
I                                                 0     0     0     2     2     0     1     1     2     2     0
I                                                 2     0     2     1     1     0     2     1     1     0     2
I                                                 0     2     1     2     1     2     0     1     0     0     0
I                                                 0     1     2     1     0     1     1     1     0     0     0
I                                                 2     2     0     0     2     1     1     0     0     2     1
I                                                 0     1     1     0     2     1     0     1     2     1     2
Ind10   1    1    2    1    2    2    0    1    2    0     0     1     1     2     0     0     2     1     1     1
```

$$w_1\times\begin{array}{c}\text{SNP1}\\2\\1\\2\\2\\0\\1\\2\\2\\1\\1\end{array} + \; w_2\times\begin{array}{c}\text{SNP2}\\1\\0\\2\\0\\2\\0\\2\\2\\2\\1\end{array} + \cdots + w_{20}\times\begin{array}{c}\text{SNP20}\\1\\1\\1\\0\\2\\0\\0\\1\\2\\1\end{array} = \begin{array}{c}\text{Score for Ind}_1\\\text{Score for Ind}_2\\\cdot\\\cdot\\\cdot\\\text{Score for Ind}_{10}\end{array}$$

Is that enough to represent (summarize) the data?

PCA solution: “**Repeat** the process and look for **orthogonal** scores.”

Formally!

PCA solution: “**Repeat** the process and look for **orthogonal** scores.”

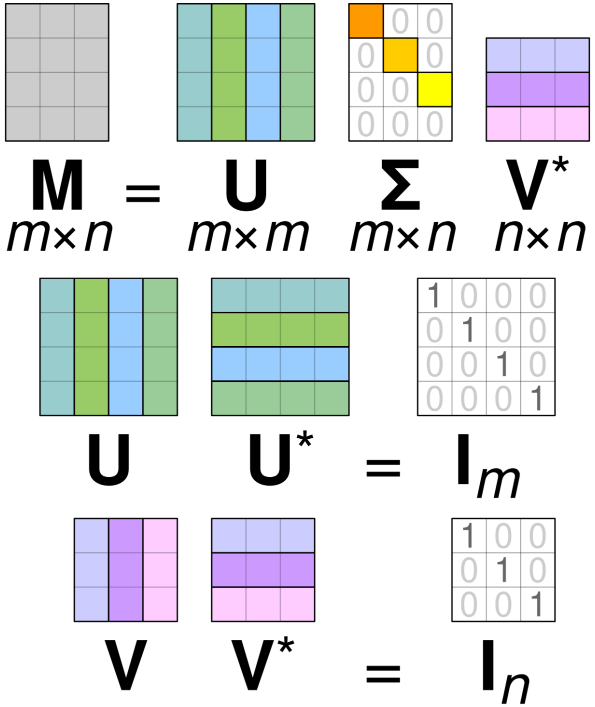
$$\boldsymbol{w}_2 \equiv \underset{\substack{\|\boldsymbol{w}_2\|=1 \\ \boldsymbol{w}_1'\mathbf{X}'\mathbf{X}\boldsymbol{w}_2=0}}{\operatorname{argmax}} \{\boldsymbol{w}_2'\mathbf{X}'\mathbf{X}\boldsymbol{w}_2\}$$

$$\boldsymbol{w}_k \equiv \underset{\substack{\|\boldsymbol{w}_k\|=1 \\ \boldsymbol{w}_k'\mathbf{X}'\mathbf{X}\boldsymbol{w}_1=0 \\ \cdots \\ \boldsymbol{w}_k'\mathbf{X}'\mathbf{X}\boldsymbol{w}_{k-1}=0}}{\operatorname{argmax}} \{\boldsymbol{w}_k'\mathbf{X}'\mathbf{X}\boldsymbol{w}_k\}$$

PCA and Singular Vector Decomposition

The SVD decomposition of a matrix $\mathbf{M}$ is given by a factorization of $\mathbf{M}$ as $\mathbf{M} = \mathbf{U\Sigma V}'$

Such that $\mathbf{U}'\mathbf{U} = \mathbf{I}$ and $\mathbf{V}'\mathbf{V} = \mathbf{I}$ and $\boldsymbol{\Sigma}$ is a (rectangular) diagonal matrix.



$\mathbf{M}$ = $\mathbf{U}$ $\boldsymbol{\Sigma}$ $\mathbf{V}^*$
$m\times n$ $m\times m$ $m\times n$ $n\times n$
$\mathbf{U}$ $\mathbf{U}^*$ = $\mathbf{I}_m$
$\mathbf{V}$ $\mathbf{V}^*$ = $\mathbf{I}_n$

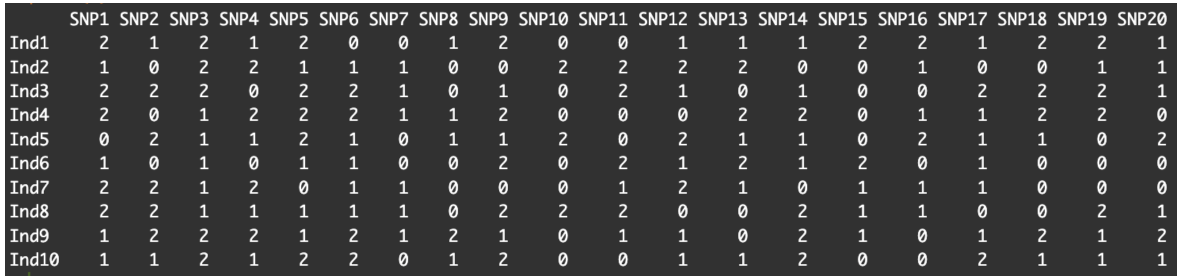
Image from Wikipedia

The columns of $\mathbf{U}$ are called eigenvectors and are **proportional** to principal components of $\mathbf{M}$.

PCA and Genomic Relatedness

Principal Components (PC) can be obtained from the eigenvectors of a genomic relationship matrix.

If $\mathbf{X}$ is a (scaled) genotyped matrix with n individuals and m SNPs



```
      SNP1 SNP2 SNP3 SNP4 SNP5 SNP6 SNP7 SNP8 SNP9 SNP10 SNP11 SNP12 SNP13 SNP14 SNP15 SNP16 SNP17 SNP18 SNP19 SNP20
Ind1     2    1    2    1    2    0    0    1    2     0     0     1     1     1     2     2     1     2     2     1
Ind2     1    0    2    2    1    1    1    0    0     2     2     2     2     0     0     1     0     0     1     1
Ind3     2    2    2    0    2    2    1    0    1     0     2     1     0     1     0     0     2     2     2     1
Ind4     2    0    1    2    2    2    1    1    2     0     0     0     2     2     0     1     1     2     2     0
Ind5     0    2    1    1    2    1    0    1    1     2     0     2     1     1     0     2     1     1     0     2
Ind6     1    0    1    0    1    1    0    0    2     0     2     1     2     1     2     0     1     0     0     0
Ind7     2    2    1    2    0    1    1    0    0     0     1     2     1     0     1     1     1     0     0     0
Ind8     2    2    1    1    1    1    1    0    2     2     2     0     0     2     1     1     0     0     2     1
Ind9     1    2    2    2    1    2    1    2    1     0     1     1     0     2     1     0     1     2     1     2
Ind10    1    1    2    1    2    2    0    1    2     0     0     1     1     2     0     0     2     1     1     1
```

Then $\mathbf{X} = \mathbf{U\Sigma V}'$ implies that $\mathbf{G} = m^{-1}\mathbf{XX}' = m^{-1}\mathbf{U\Sigma V}'\mathbf{V\Sigma}'\mathbf{U}' = \mathbf{U}(m^{-1}\mathbf{\Sigma\Sigma}')\mathbf{U}'$

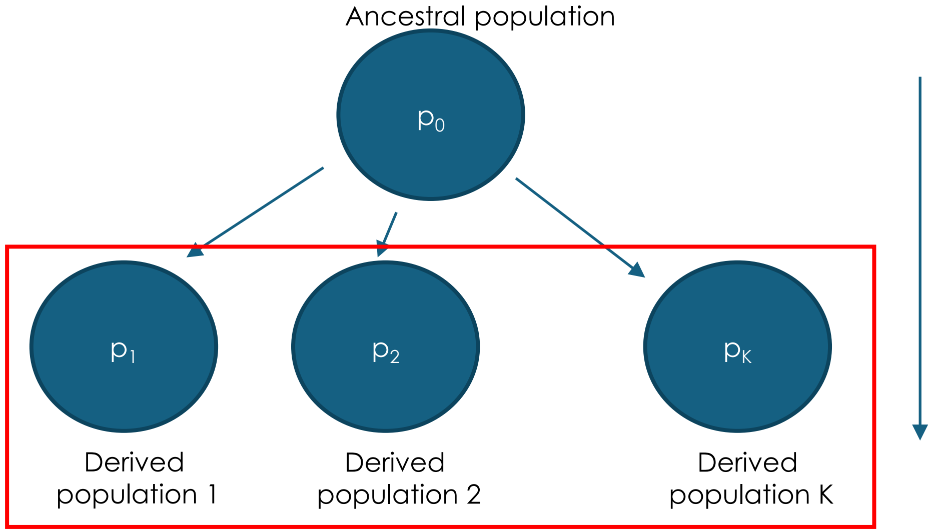
$\mathbf{G}$ is a $n \times n$ matrix also called Genomic Relationship Matrix or **GRM**.

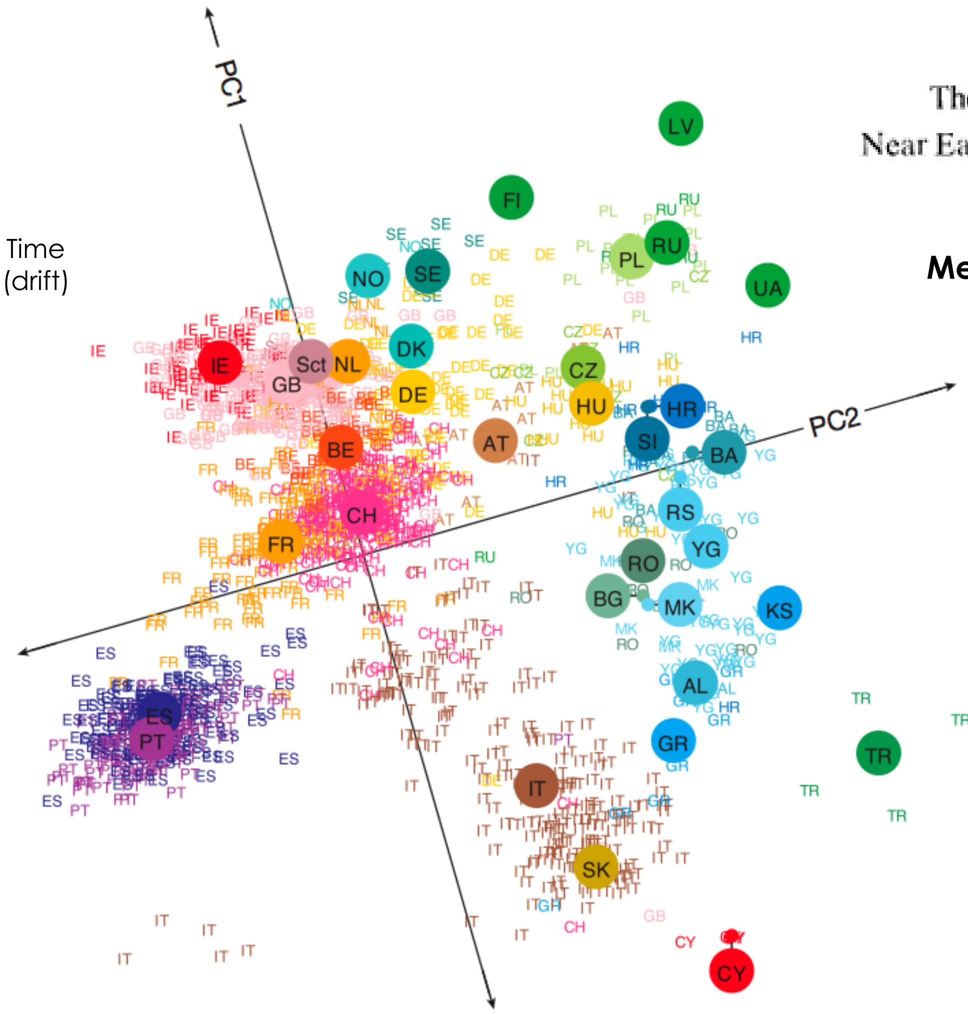
Summary

- PCA is a statistical technique to **visualize** and **reduce** the dimension of data by **summarizing the information as linear combinations of data points**.
- Those linear combinations (scores) are called **Principal Components (PCs)** and the weights **PC loadings**.
- PCA has tight links with concepts such **SVD decomposition** of **genomic relationship matrices (GRM)**.

What do we get when we apply PCA to genetic data?

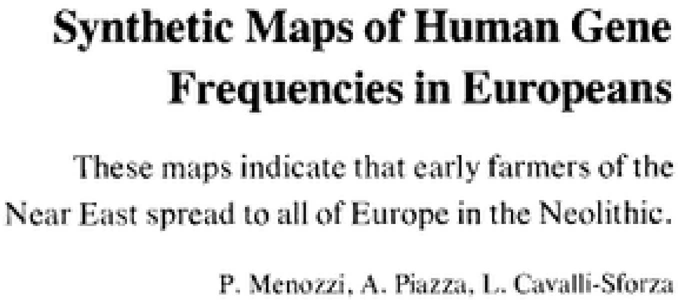
PCA detects genetic structures between (sub)populations





Novembre et al. Nature (2008)

[200K SNPs]



Synthetic Maps of Human Gene Frequencies in Europeans

These maps indicate that early farmers of the Near East spread to all of Europe in the Neolithic.

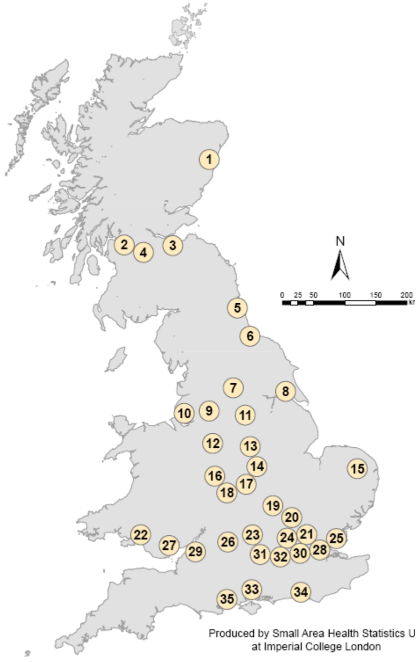
P. Menozzi, A. Piazza, L. Cavalli-Sforza

Menozzi & Cavalli-Sforza. Science (1978)

[10 markers]

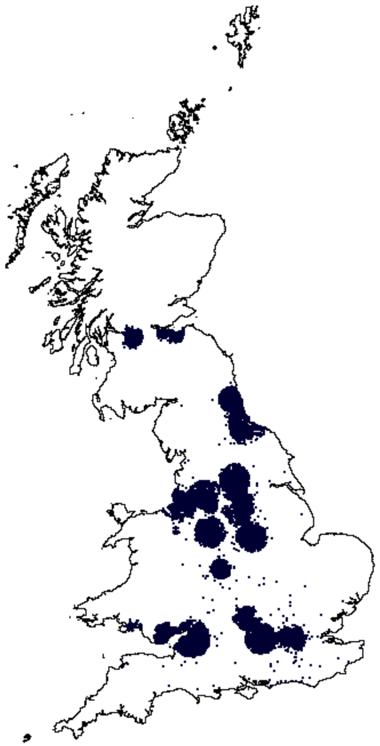
Genetic structures have different sources

biobank uk
Improving the health of future generations

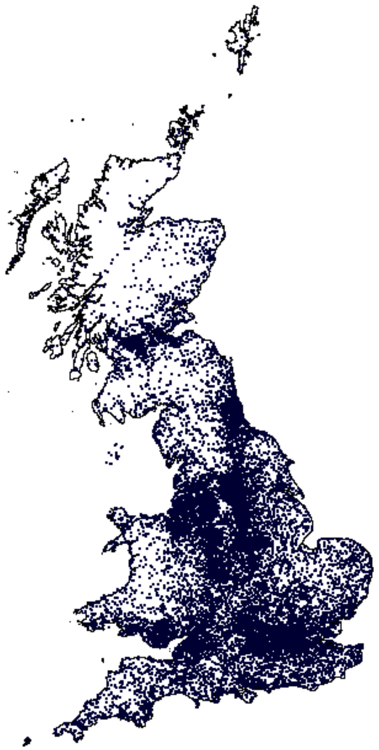


Produced by Small Area Health Statistics U
at Imperial College London

Assessment centers



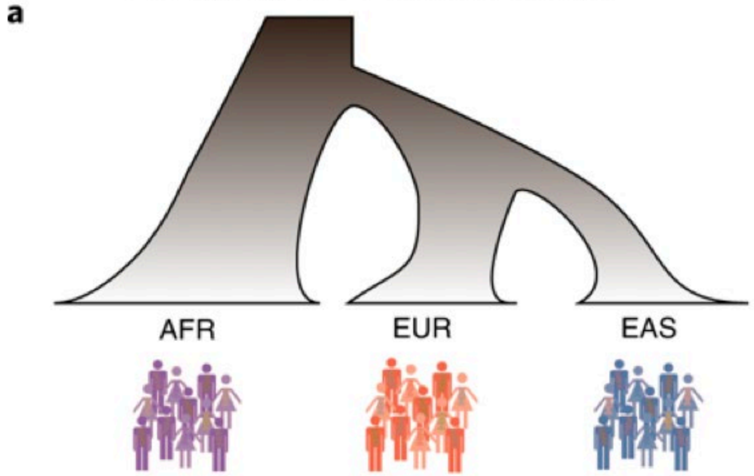
Current living address



Place of birth

Migrations

Demographic relationships



Martin et. al., *Nat. Genet* (2019)

Partner's Choice

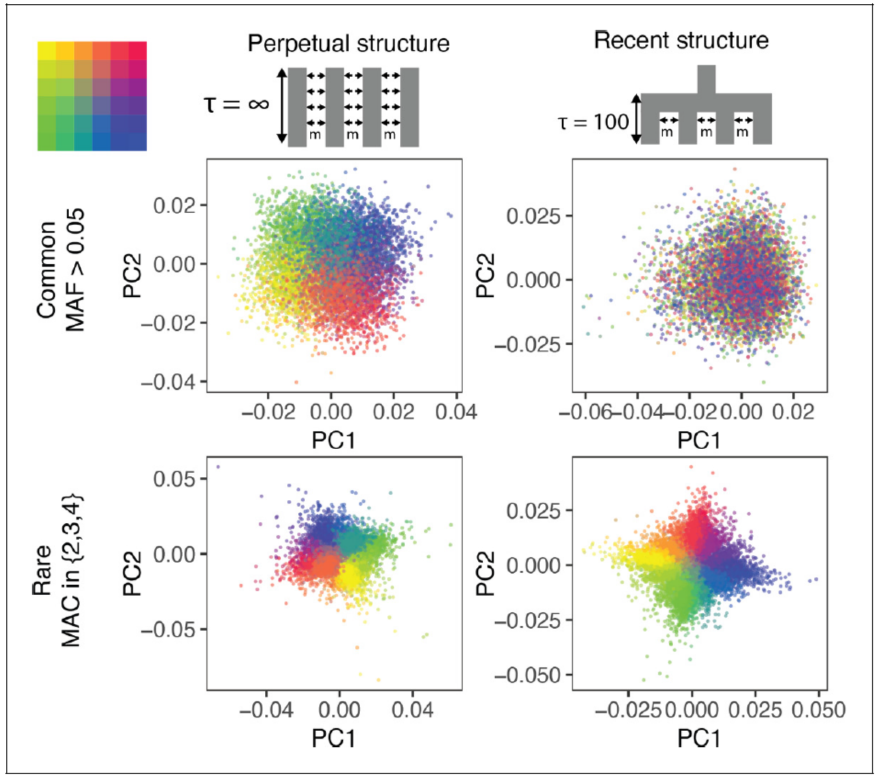
Interpretating PCs can be challenging…

- What is the “evolutionary force” causing the structure that I observe?

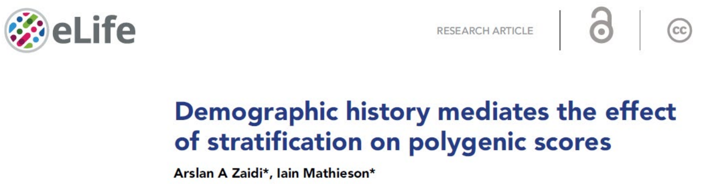
- Answering this question is **entire field of research**, which expands **beyond PCA**
  - What structures are detectable using PCA
  - How to detect structures when PCA does not work

Interpretating PCs can be challenging…

- PCA informs about population structures at different times, depending on allele frequency (rare variant => more recent history)
- Rare variant stratification (i.e., more recent history) can be missed



The type structure detected depends on the set of variants used as input!


eLife RESEARCH ARTICLE

Demographic history mediates the effect of stratification on polygenic scores

Arslan A Zaidi*, Iain Mathieson*

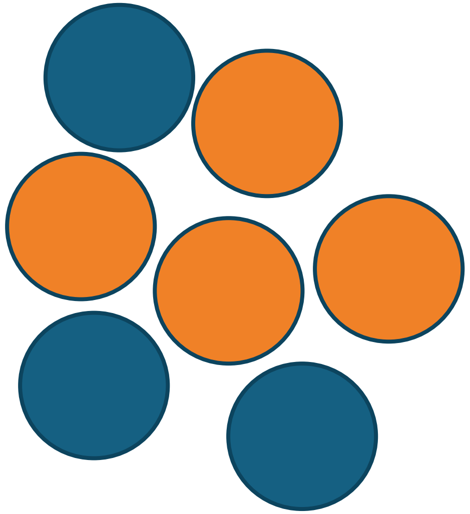
Summary

- PCA detects **genetic structures** in a sample of genomes.
- PCA is **agnostic** to the structure detected, which makes interpretation challenging.
- The type of structure depends on the set of variants used as input.

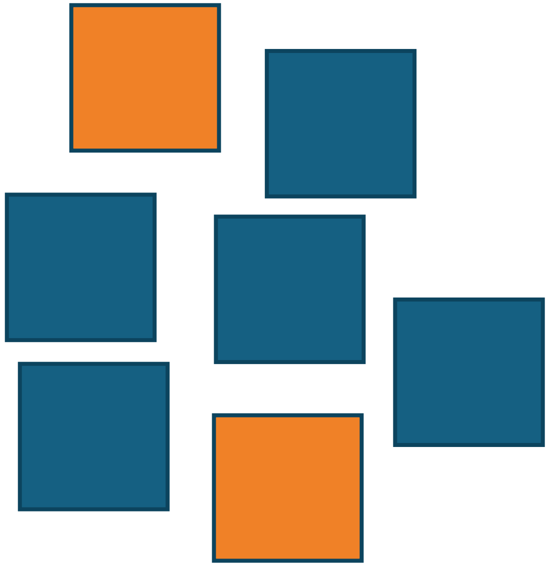
Using PCA to correct confounding in association analyses?

Genome-wide association studies

Association = **allele frequency differences between cases and controls**

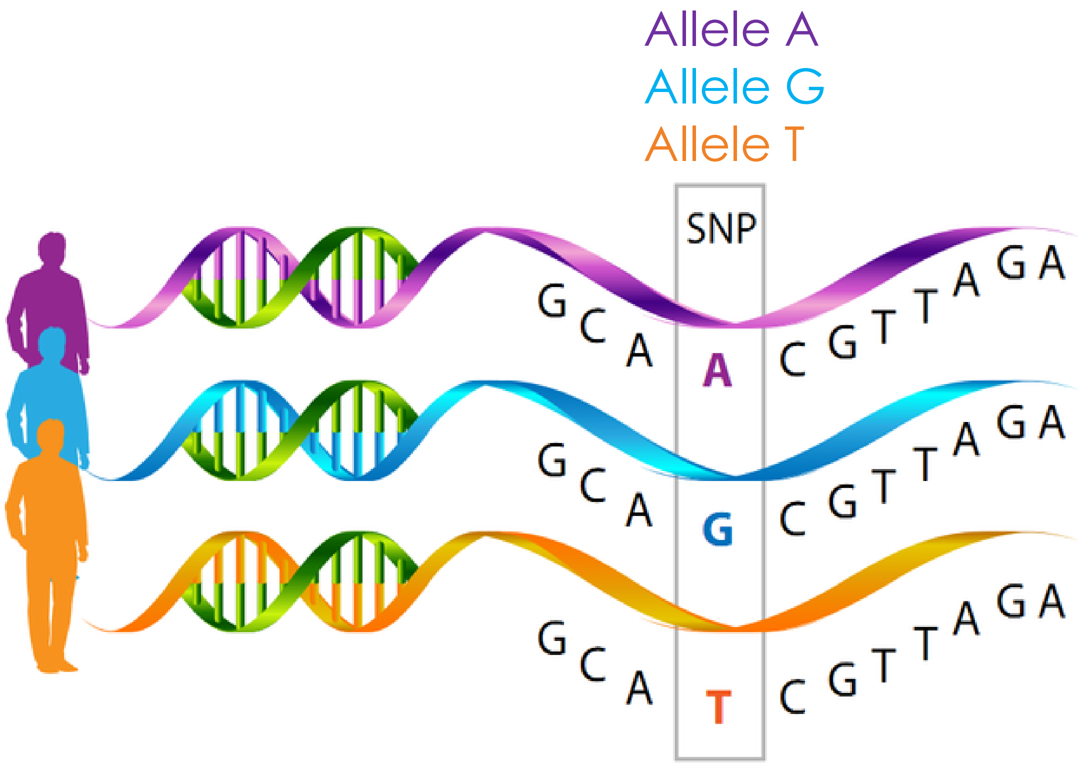


More alleles T in Cases



Less alleles T in Controls

Large sample sizes are required…



Population stratification: chopstick example

Sample 1 Americans: $\chi^2=0$, $p=1$

	Use of chopsticks		
	Yes	**No**	**Total**
Allele 1	320	320	640
Allele 2	80	80	160
Total	400	400	800

Sample 2 Chinese: $\chi^2=0$, $p=1$

	Use of chopsticks		
	Yes	**No**	**Total**
Allele 1	320	20	340
Allele 2	320	20	340
Total	640	40	680

Population stratification: chopstick example

Sample 1 Americans: χ^2=0, p=1			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	320	640
Allele 2	80	80	160
Total	400	400	800

Sample 2 Chinese: χ^2=0, p=1			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	20	340
Allele 2	320	20	340
Total	640	40	680

There is a clear difference between Americans and Chinese in proportion of "cases" and "controls"

Population stratification: chopstick example

Sample 1 Americans: χ^2=0, p=1			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	320	640
Allele 2	80	80	160
Total	400	400	800

Sample 2 Chinese: χ^2=0, p=1			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	20	340
Allele 2	320	20	340
Total	640	40	680

There is a clear allele frequency difference between Americans and Chinese

Population stratification: chopstick example

Sample 1 Americans: χ^2=0, p=1			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	320	640
Allele 2	80	80	160
Total	400	400	800

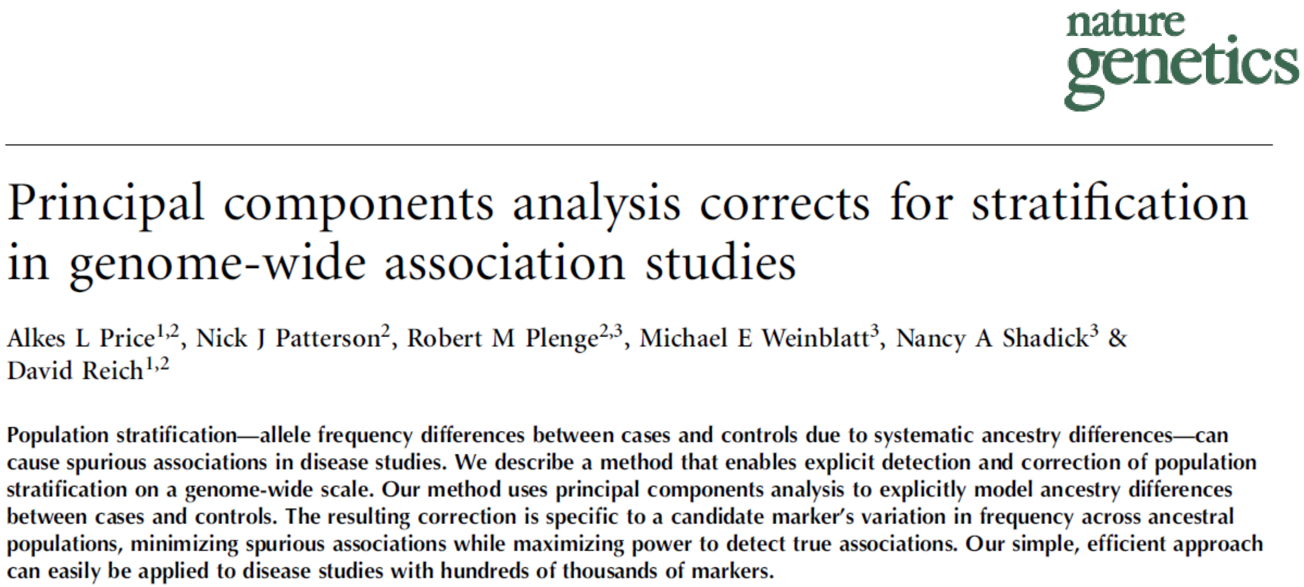
Sample 2 Chinese: χ^2=0, p=1			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	20	340
Allele 2	320	20	340
Total	640	40	680

Population stratification: chopstick example

Sample 1 Americans: $\chi^2=0$, $p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	320	640
Allele 2	80	80	160
Total	400	400	800

Sample 2 Chinese: $\chi^2=0$, $p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	20	340
Allele 2	320	20	340
Total	640	40	680

Sample 1 + 2 = Americans + Chinese: $\chi^2=34.2$, $p=4.9 \times 10^{-9}$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	640	340	980
Allele 2	400	100	500
Total	1040	440	1480



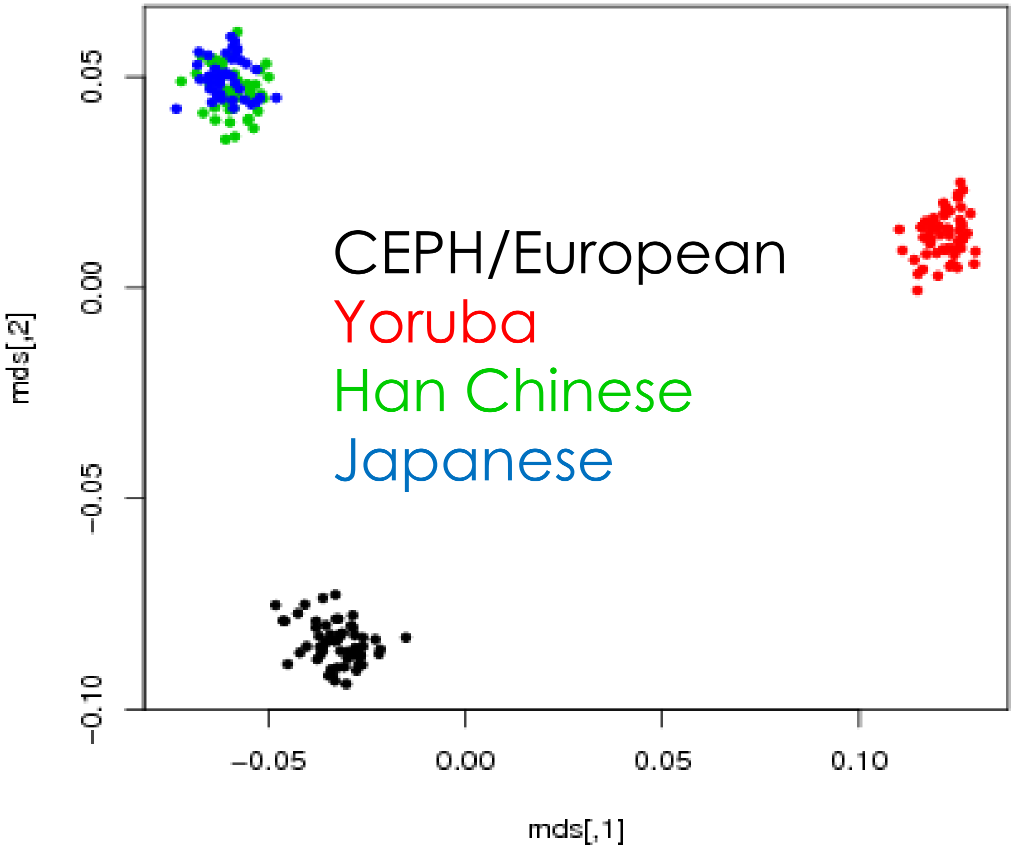
nature
genetics

Principal components analysis corrects for stratification in genome-wide association studies

Alkes L Price[1,2], Nick J Patterson[2], Robert M Plenge[2,3], Michael E Weinblatt[3], Nancy A Shadick[3] & David Reich[1,2]

Population stratification—allele frequency differences between cases and controls due to systematic ancestry differences—can cause spurious associations in disease studies. We describe a method that enables explicit detection and correction of population stratification on a genome-wide scale. Our method uses principal components analysis to explicitly model ancestry differences between cases and controls. The resulting correction is specific to a candidate marker's variation in frequency across ancestral populations, minimizing spurious associations while maximizing power to detect true associations. Our simple, efficient approach can easily be applied to disease studies with hundreds of thousands of markers.

Price et. al., *Nat. Genet* (2006)



Practical considerations about PCA?

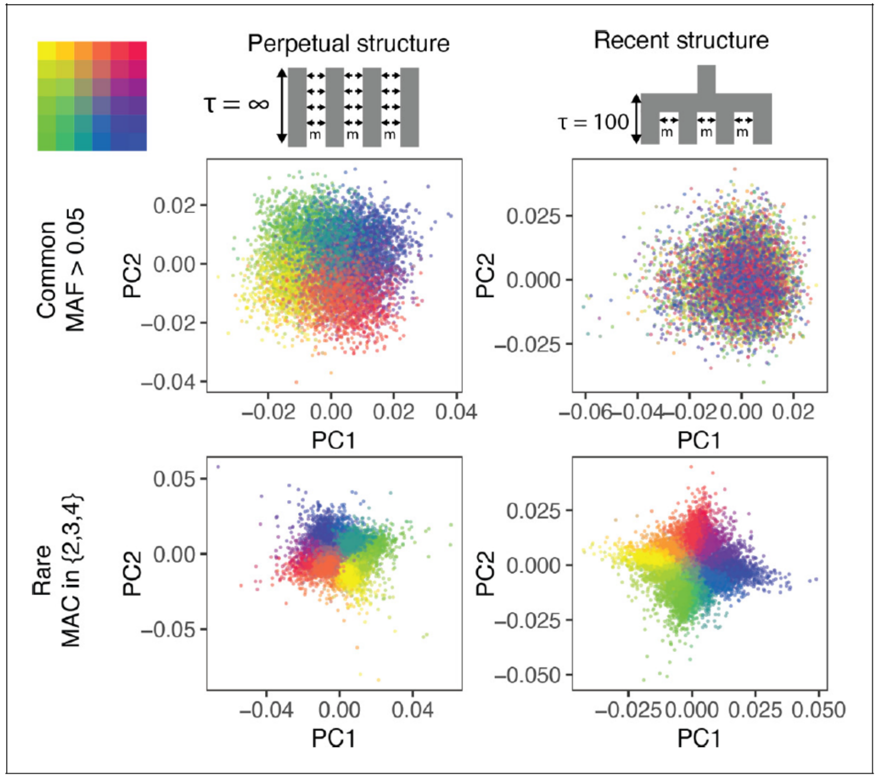
Caveats

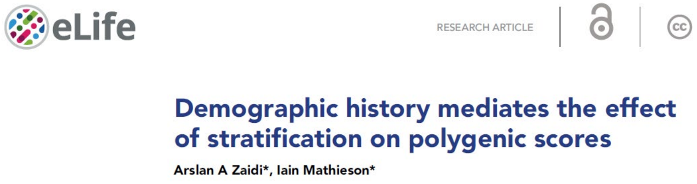
- Allele frequency threshold
- Impact of sample size
- Impact of LD
- Projected PCA analysis

Caveats

- Allele frequency threshold
- Impact of sample size
- Impact of LD
- Projected PCA analysis

Impact of allele frequencies





eLife

RESEARCH ARTICLE

Demographic history mediates the effect of stratification on polygenic scores

Arslan A Zaidi*, Iain Mathieson*

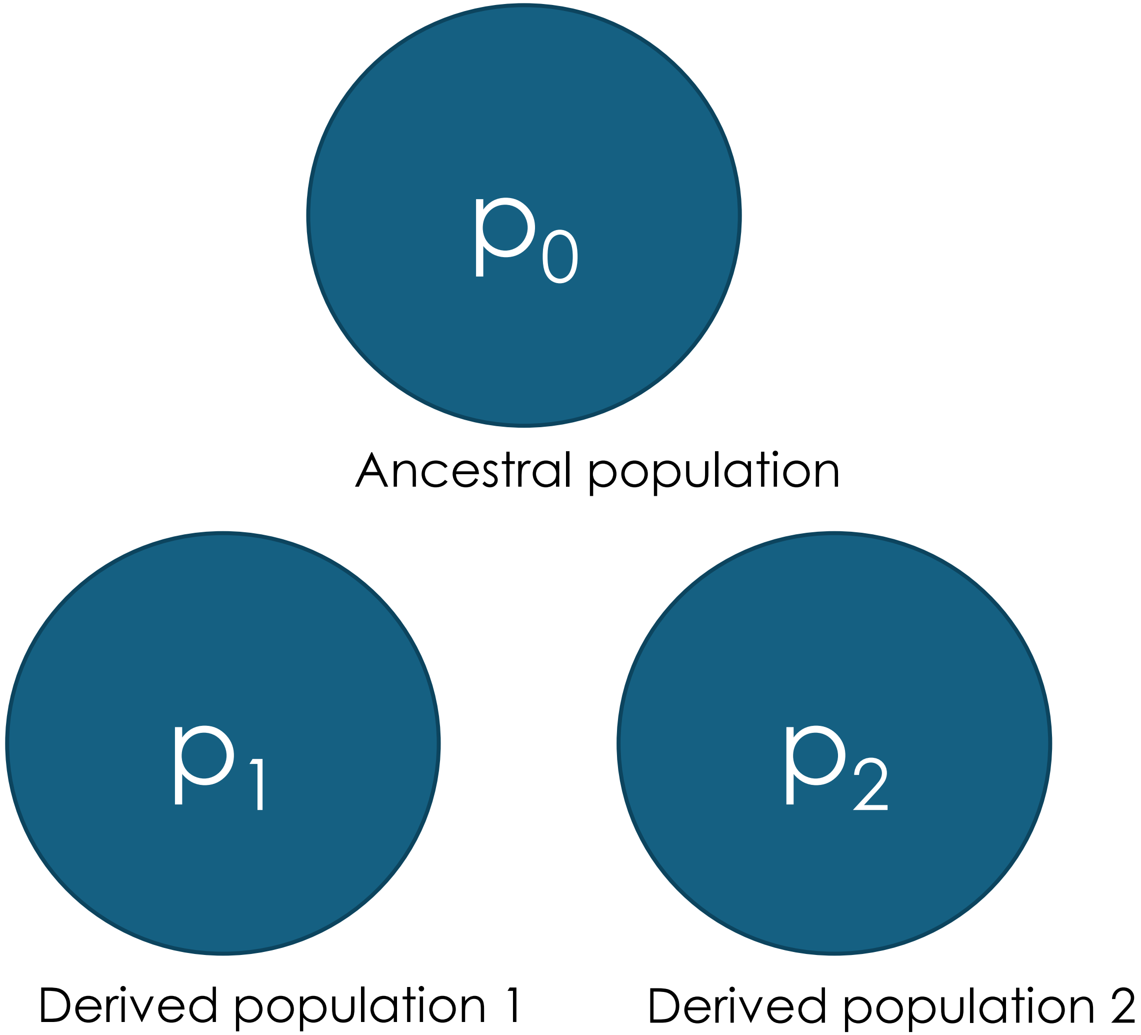
(1) Match frequency spectrum between SNPs tested for association and those used in PCA

Caveats

- Allele frequency threshold
- Impact of sample size
- Impact of LD
- Projected PCA analysis

PCA performance depends on sample size

F_{ST} model: Balding-Nichols



p_0

Ancestral population

p_1

Derived population 1

p_2

Derived population 2

$$E[p_i \,|\, p_0] = p_0$$

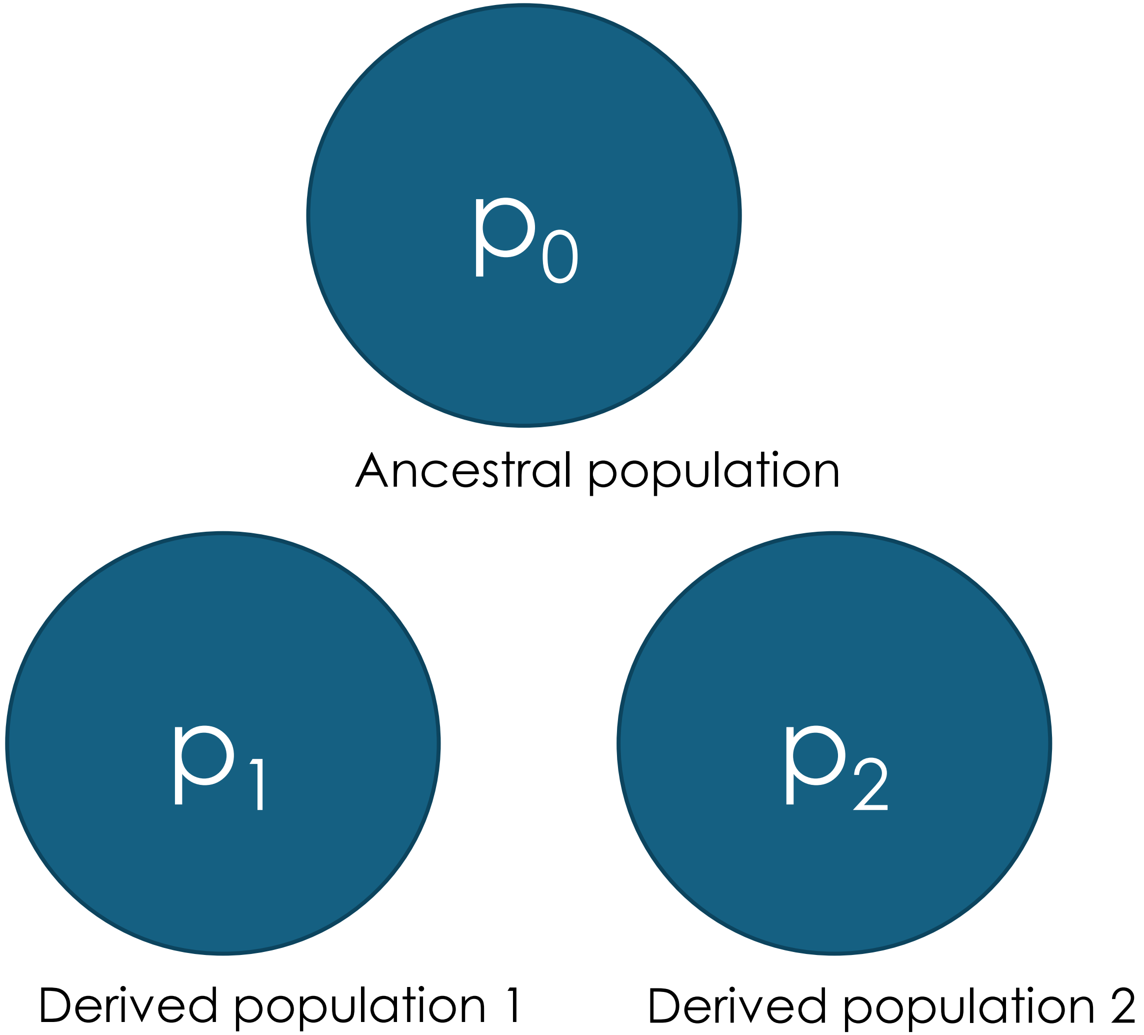
$$var[p_i \,|\, p_0] = F_{ST}\, p_0(1- p_0)$$

$$s = (1- F_{ST})/ F_{ST}$$

$$p_i \sim Beta[sp_0, s(1- p_0)]$$

PCA performance depends on sample size

F_{ST} model: Balding-Nichols


p_0

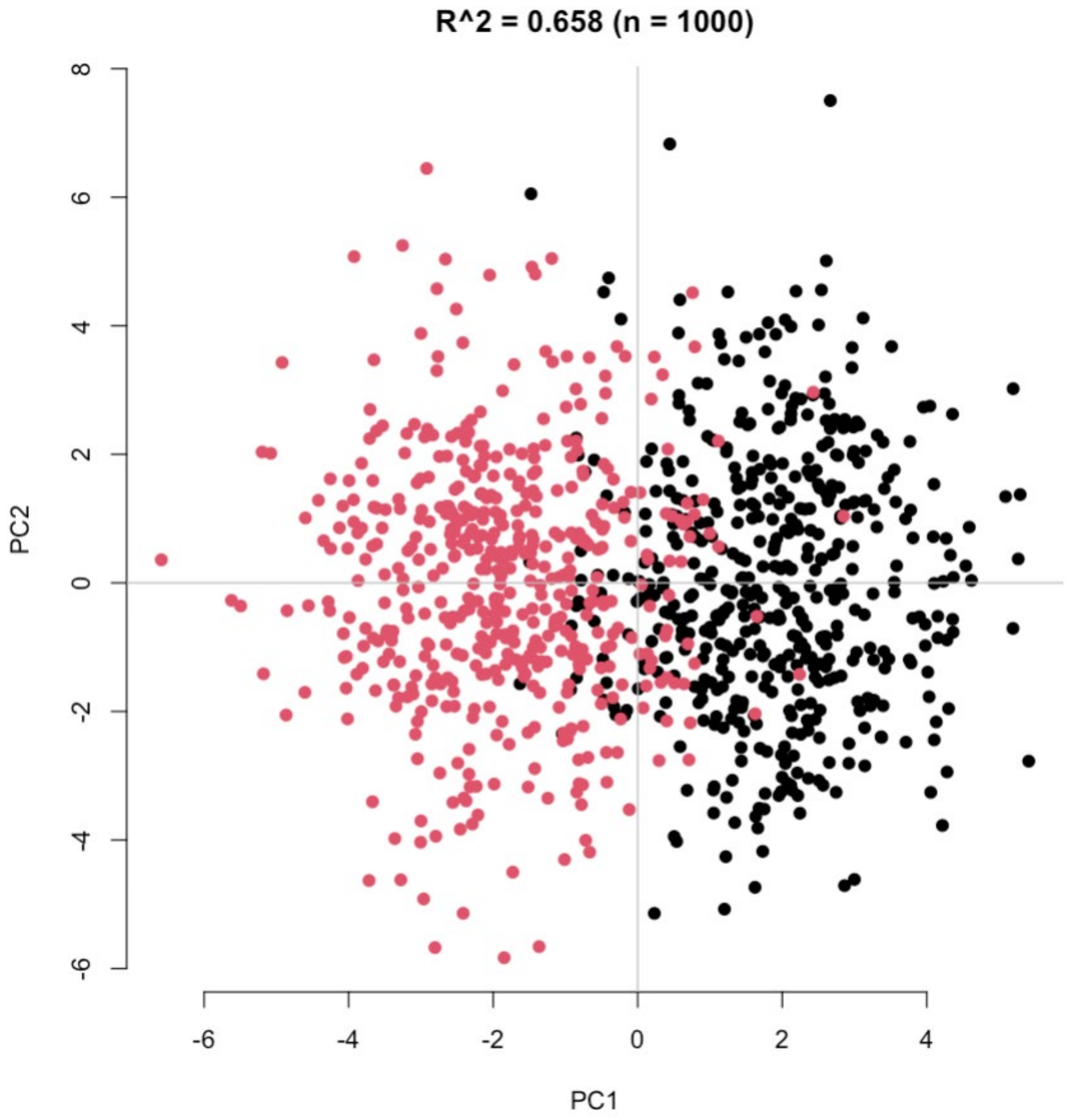
Ancestral population

p_1

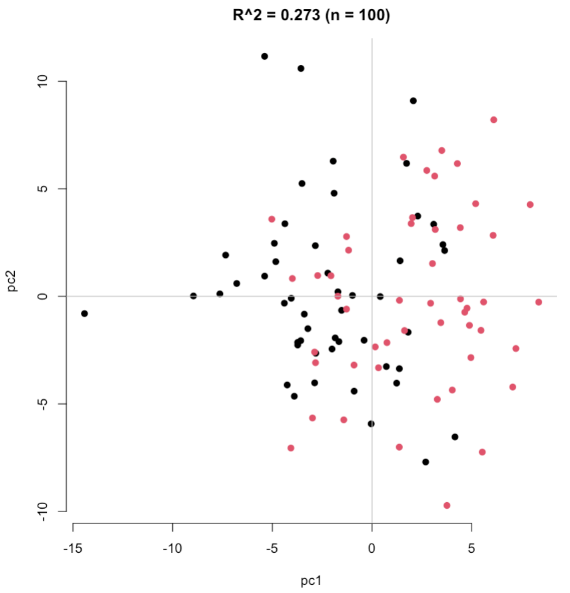
Derived population 1

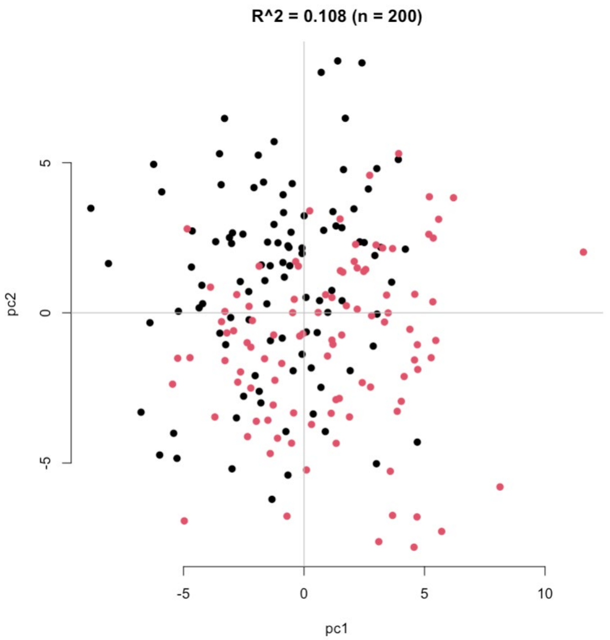
p_2

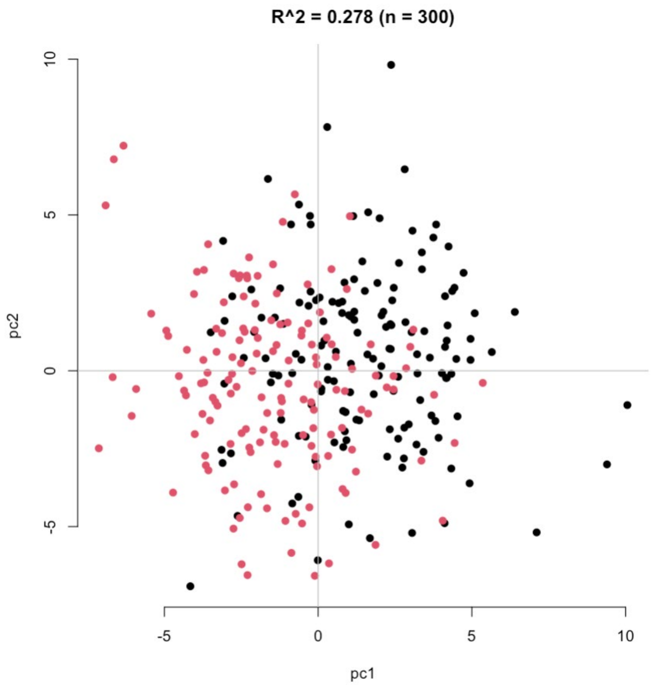
Derived population 2

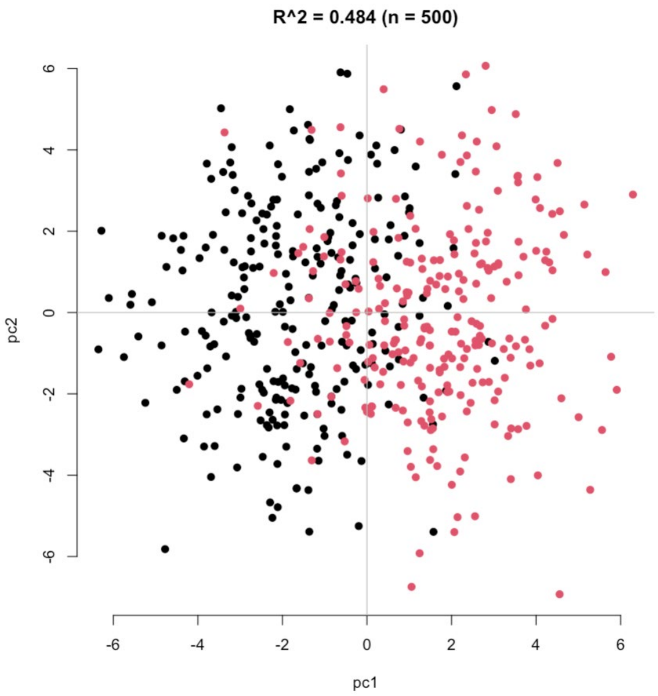

R^2 = 0.658 (n = 1000)

PCA performance depends on sample size

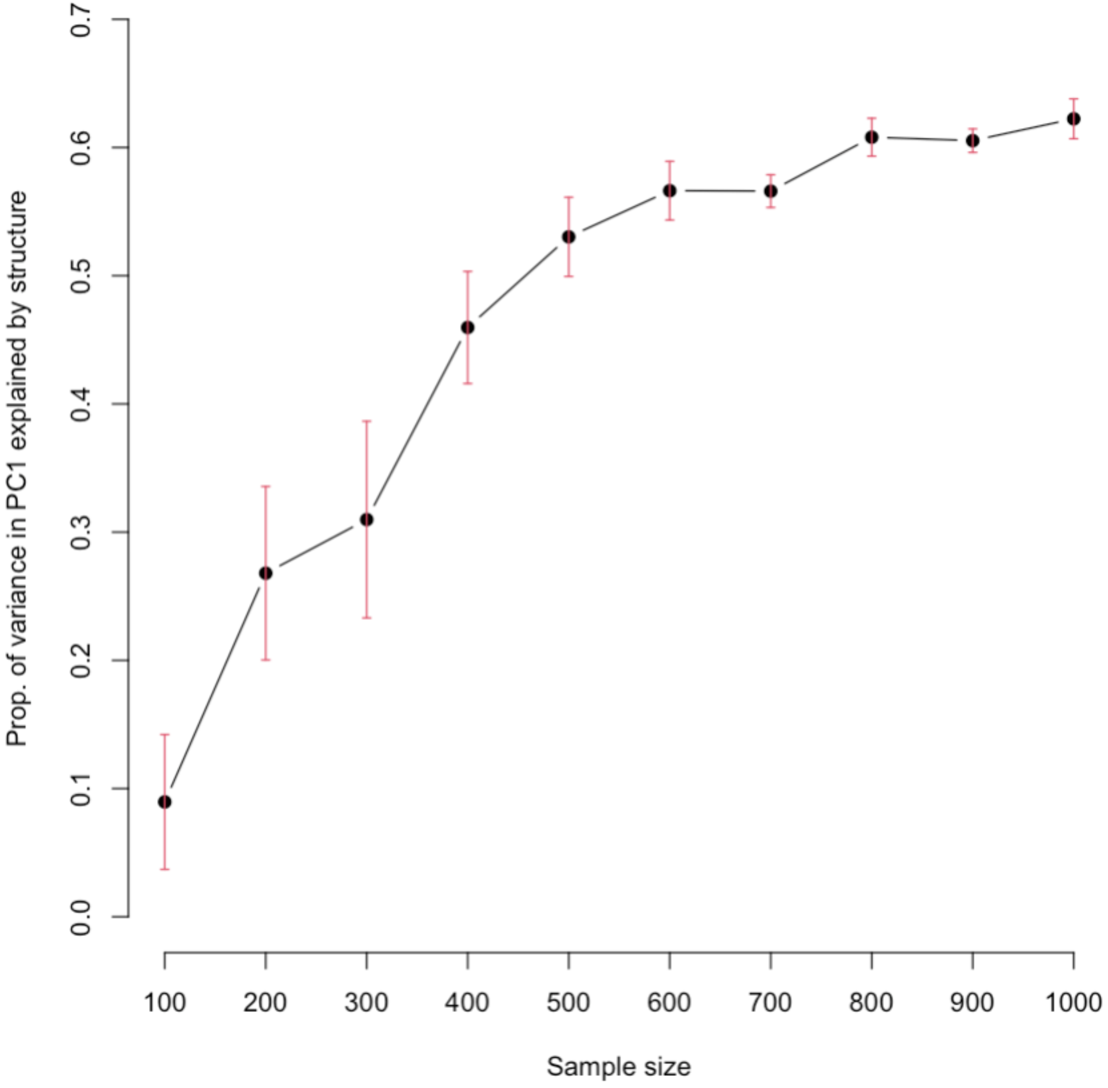








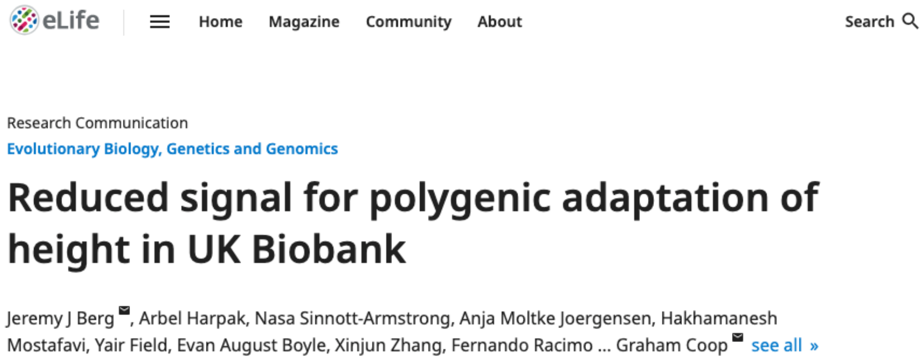
PCA performance depends on sample size



Implications

1) Structures are easier to detect in large samples
2) Adjustment for PC in small samples in sub-optimal

Residual stratification in GWAS meta-analyses

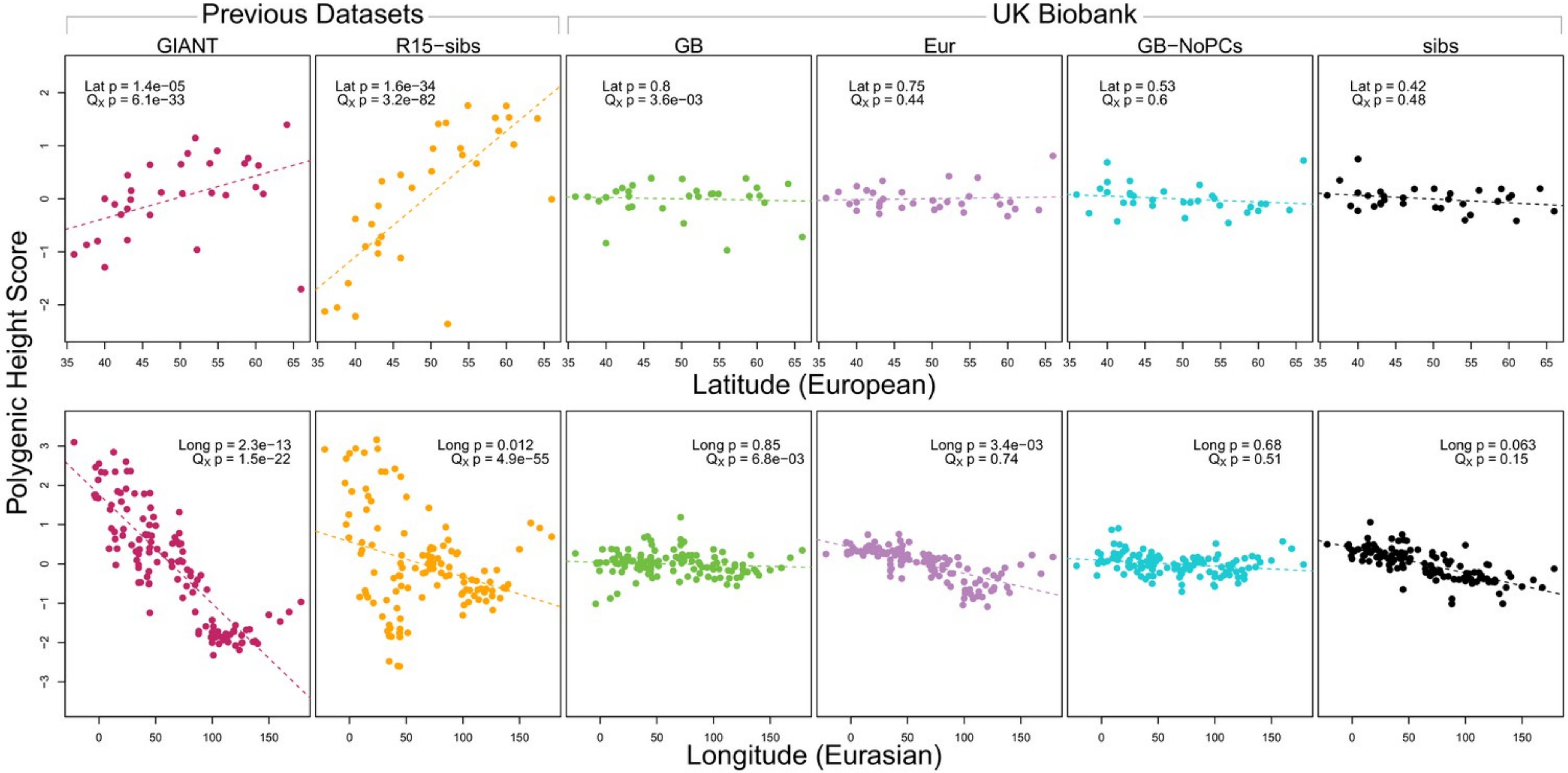


Research Communication

Evolutionary Biology, Genetics and Genomics

Reduced signal for polygenic adaptation of height in UK Biobank

Jeremy J Berg, Arbel Harpak, Nasa Sinnott-Armstrong, Anja Moltke Joergensen, Hakhamanesh Mostafavi, Yair Field, Evan August Boyle, Xinjun Zhang, Fernando Racimo ... Graham Coop see all »



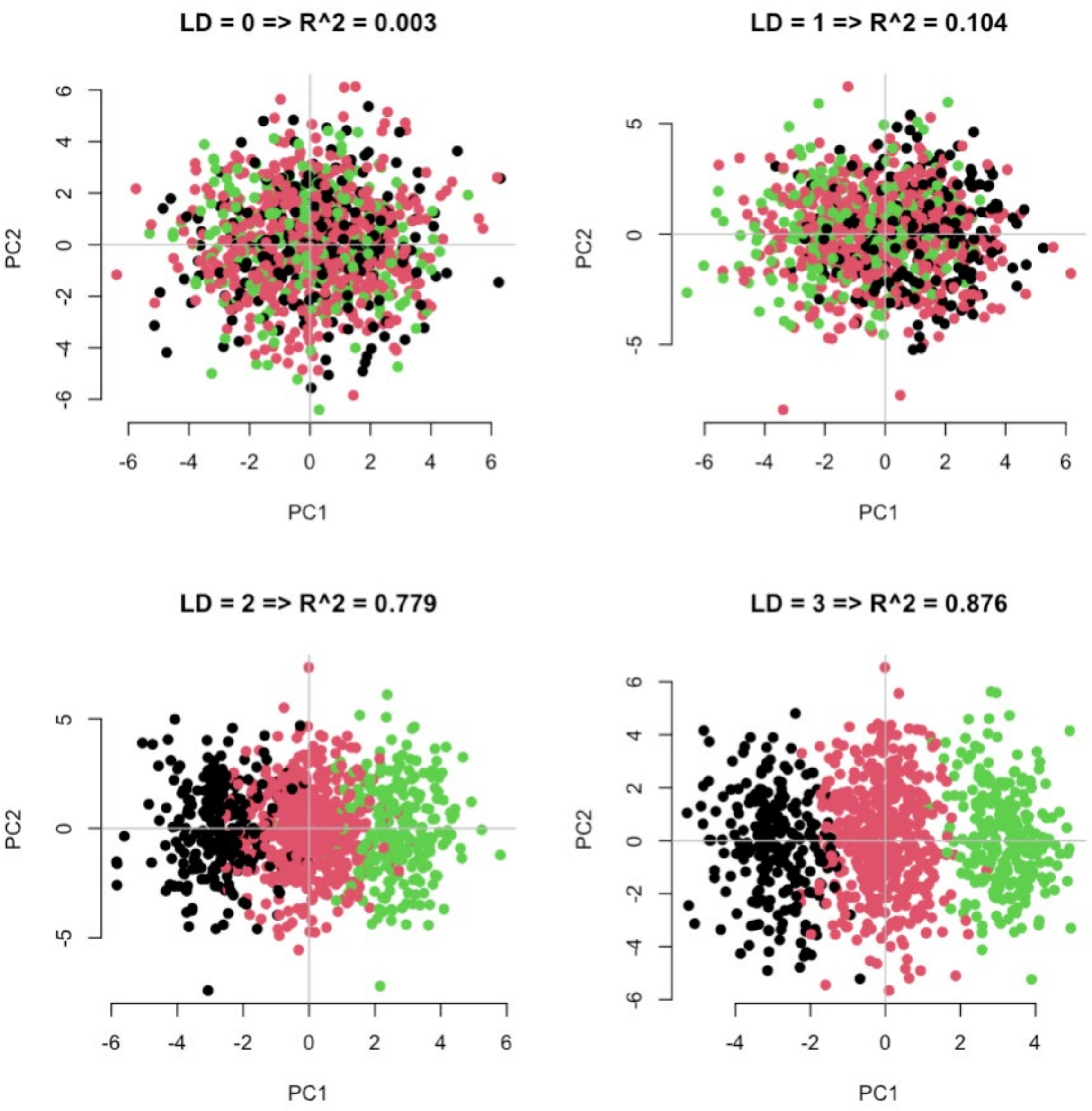
Caveats

- Allele frequency threshold
- Impact of sample size
- Impact of LD
- Projected PCA analysis

How much LD impacts PCA?

R DEMO

How much LD impacts PCA?



LD = 0 => R^2 = 0.003

LD = 1 => R^2 = 0.104

LD = 2 => R^2 = 0.779

LD = 3 => R^2 = 0.876

(1) PCA is skewed towards detecting structures within regions of the genome with high LD.

(2) Structures within these regions may not be relevant for your phenotype of interest.

(3) LD pruning reduces this bias and often improves the ability to correct for confounding.

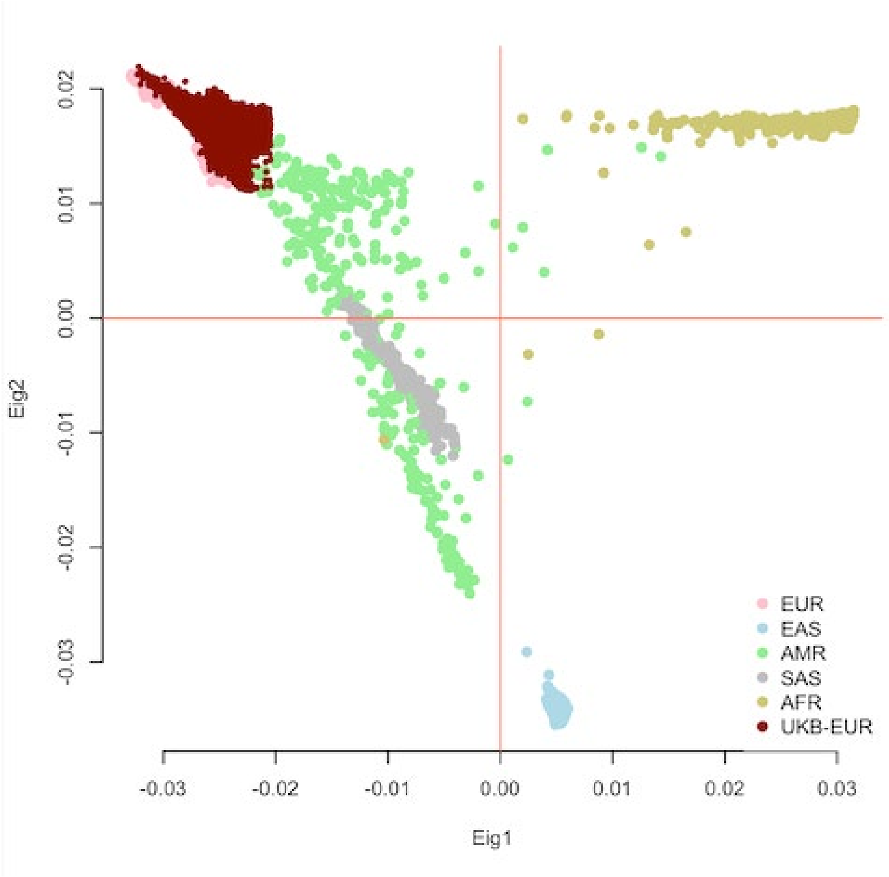
(4) Remove long-range LD loci (MHC, large inversion, etc.) .

Caveats

- Allele frequency threshold
- Impact of sample size
- Impact of LD
- Projected PCA analysis

Projected PCA

Sometimes it is better to "learn" the structures from a reference population then "project" your sample onto the resulting PC map.



Why?

1) Because you don't have enough genetic diversity in your sample!
2) Or to reduce computation.
3) To reduce biases due to relatedness

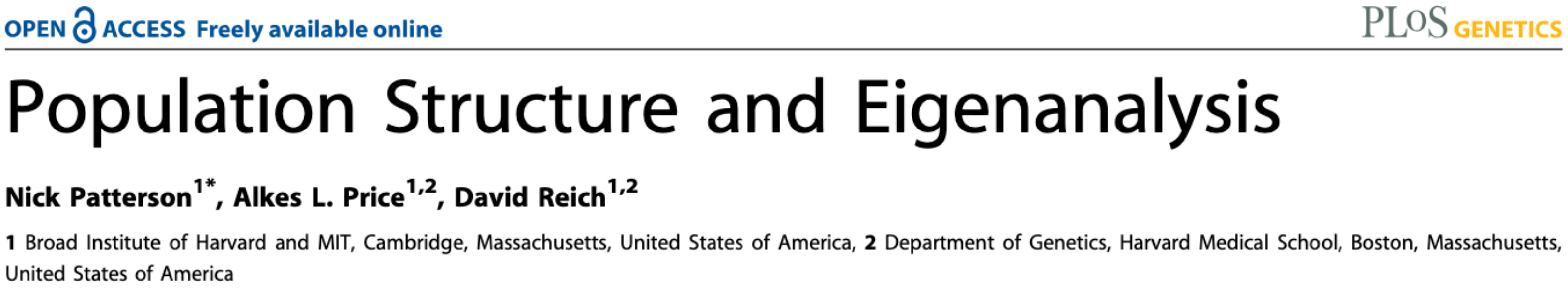
Project?

PC are linear combinations. So, if you know the loadings then you can represent anyone on the PC map.

More on PCA?

How many PCs should I use? 😅

Well, it depends…



OPEN ACCESS Freely available online PLoS GENETICS

Population Structure and Eigenanalysis

Nick Patterson[1*], Alkes L. Price[1,2], David Reich[1,2]

[1] Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America, [2] Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America

Now, if you use PCs in an association analysis then the answer is "all" (ala Linear Mixed Models)

Scaling PCA?

PCA can be a very computationally intensive. **Solutions**

1) Use subsets
2) Use random projection-based algorithms (PLINK) – good if N>5K

Many software tools available

EIGENSTRAT: from genotype only

PLINK: from genotype and GRM

GCTA: from GRM only

Or even R (if you can load the data; e.g., **BigSNPr** package)

Etc.

PCA checklist

- Run and visualize a PCA of genetic data once my life
- Use PCA to identify groups of individuals with similar ancestries
- Use PCA to correct biases due to population stratification