Data handling and PCA

Fernando Racimo

Adelaide, January 2018

Today

- Experimental design
- Data handling
- PCA
- Spatial and isolation-by-distance methods

Today

- **Experimental design**
- Data handling
- PCA
- Spatial and isolation-by-distance methods

Real-life scenario

Let’s figure out the evolutionary history of X

YOU

Experimental design

Sample size	Per-sample depth
1,000	1X
500	2X
100	10X
20	50X

total depth is
1,000X

Fumagalli 2013

Experimental design

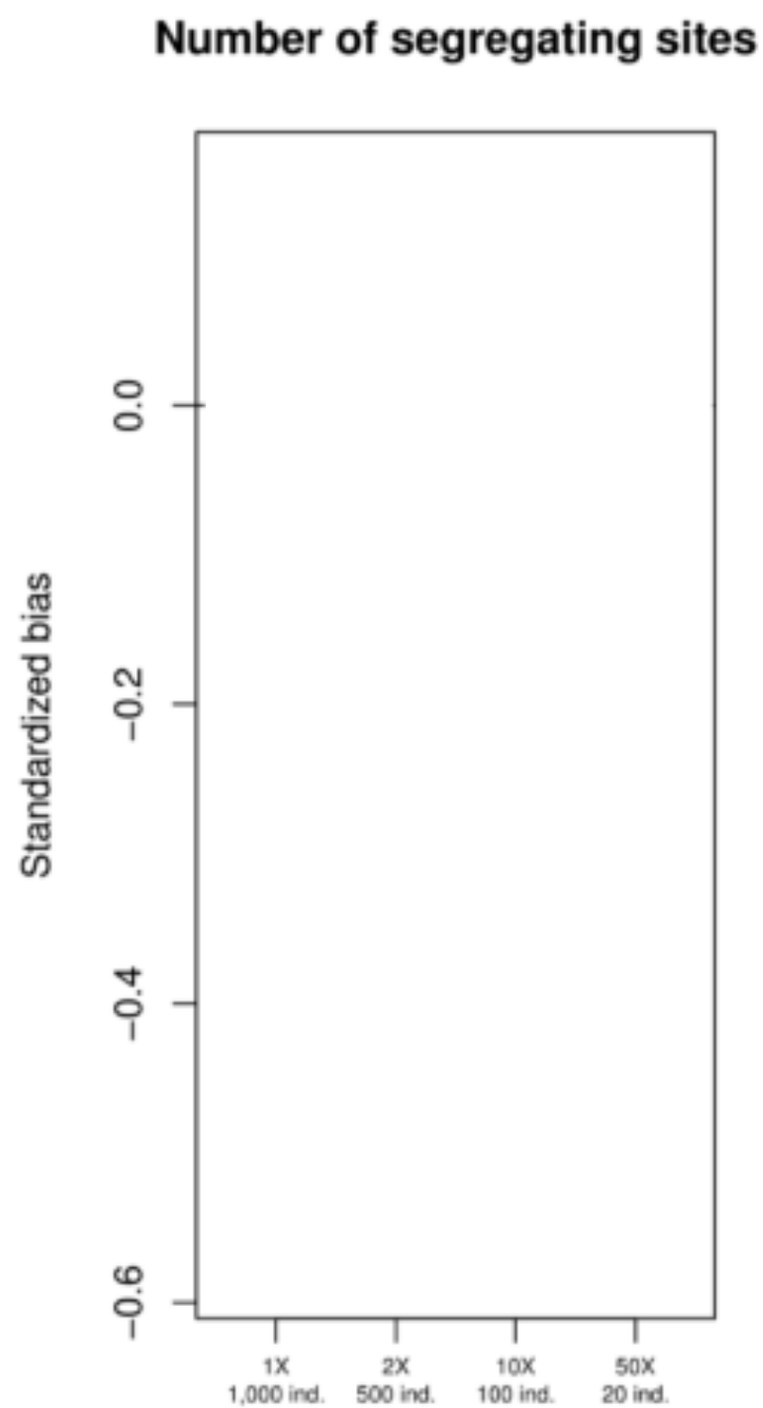
- Let's measure the bias of these set-ups for a set of summary statistics

- We'll compare the value we estimate from our data (S hat) against the true value of the **whole population** (S)

$$Bias(S) = \frac{\hat{S} - S}{S}$$

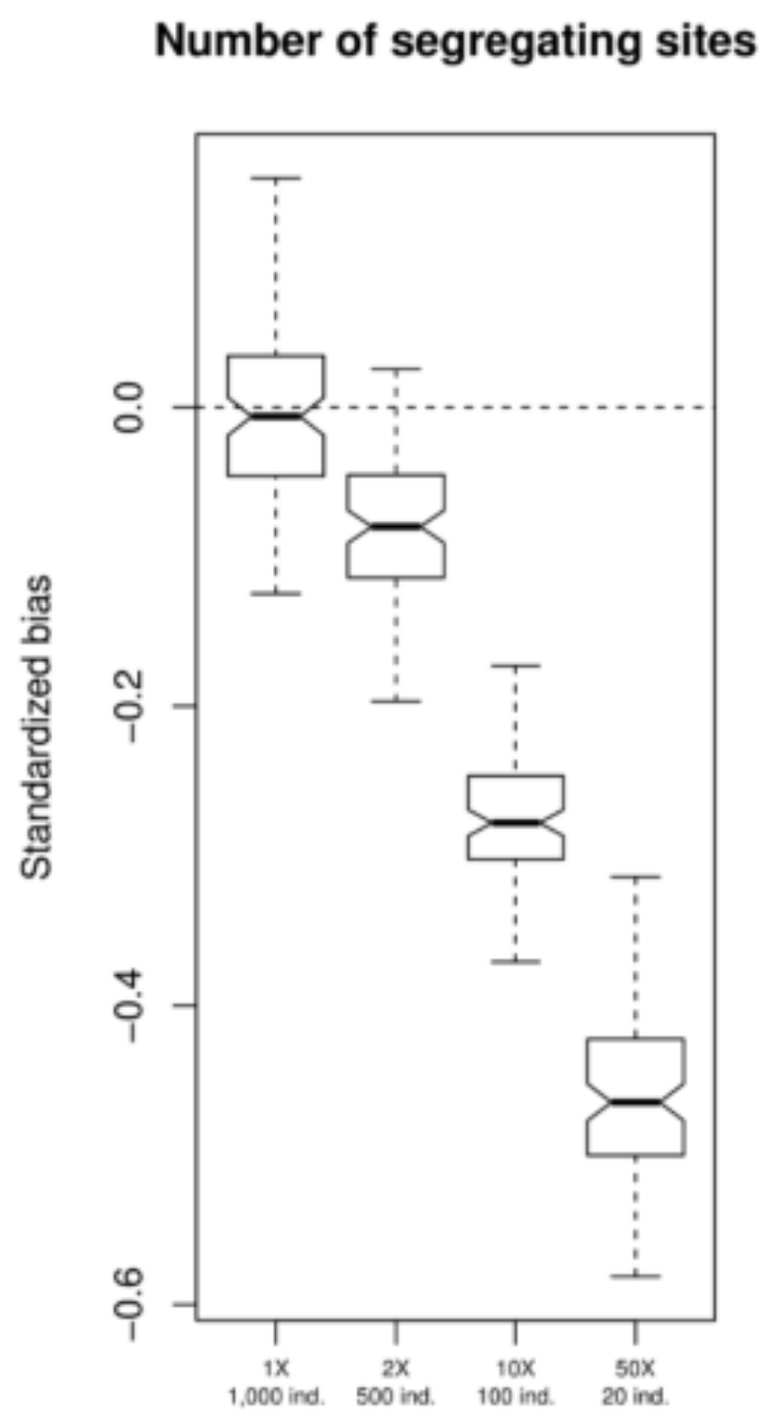
Fumagalli 2013

Experimental design



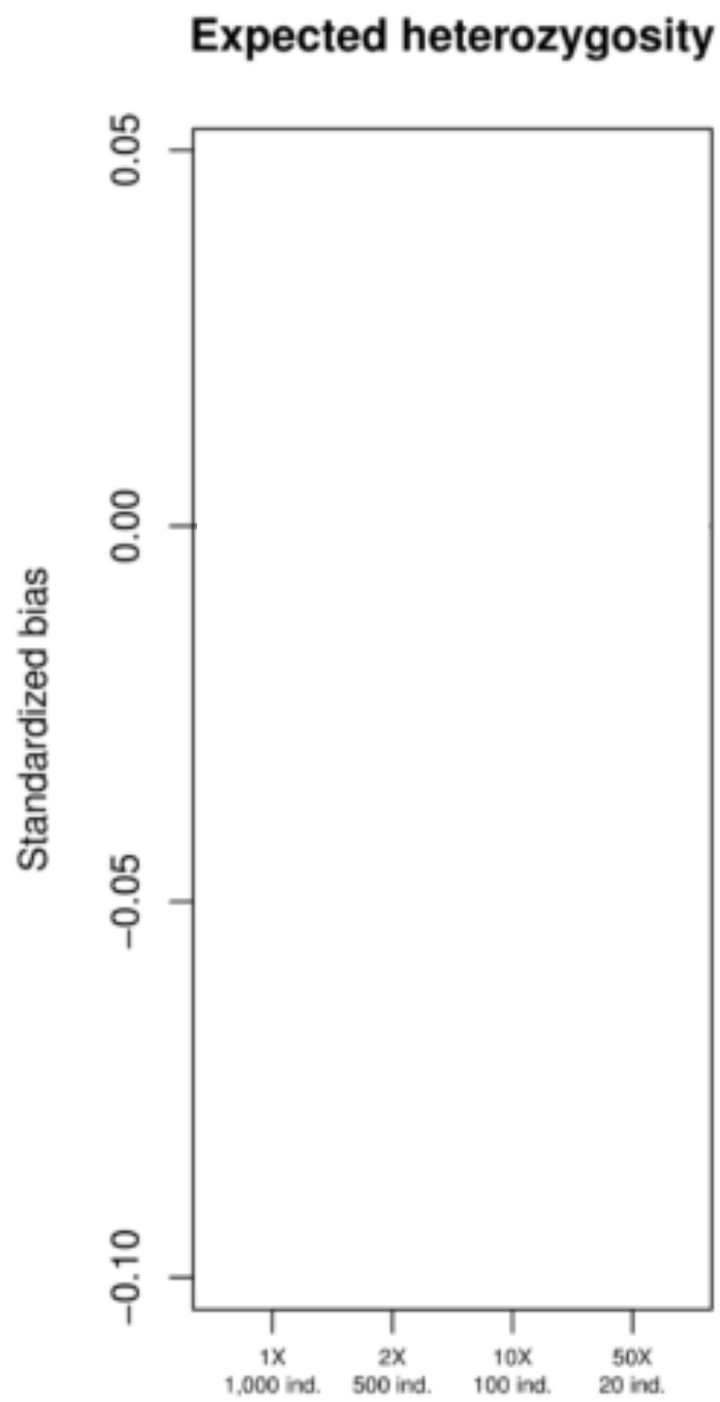
Fumagalli 2013

Experimental design



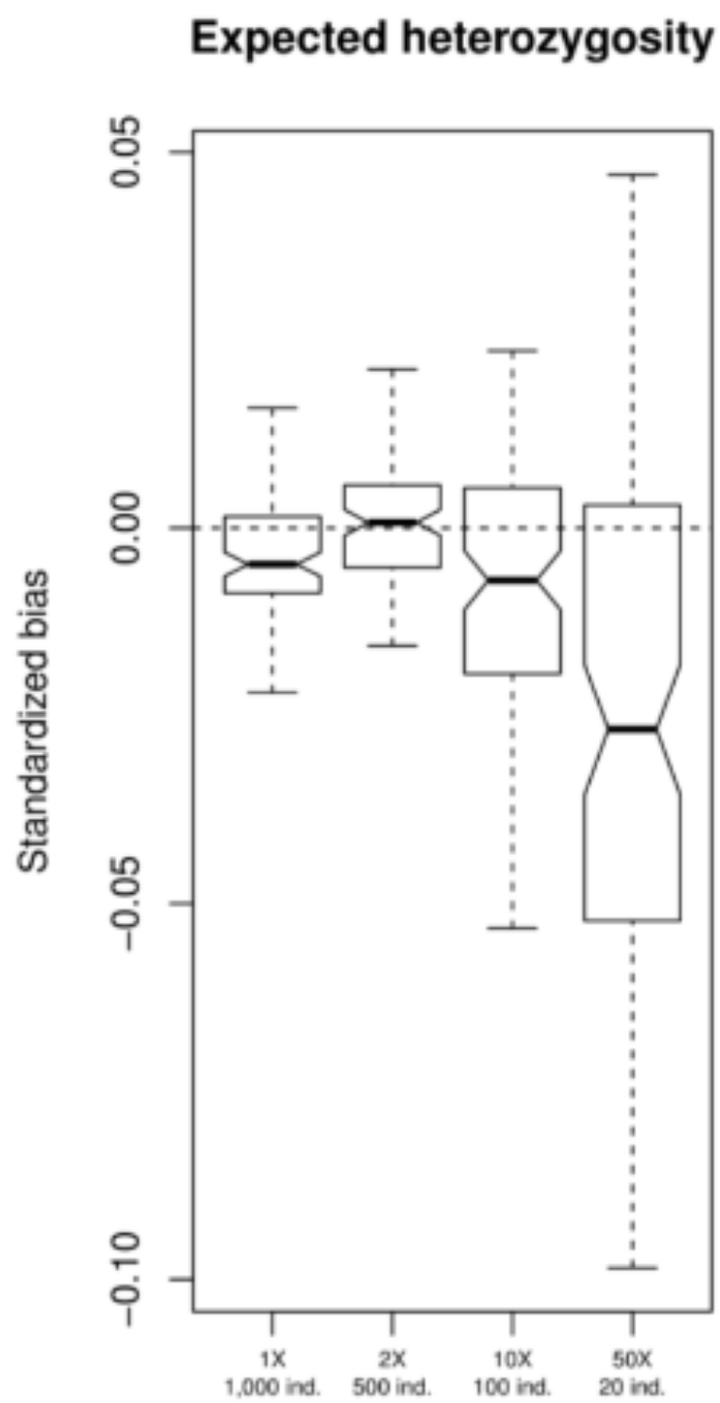
Fumagalli 2013

Experimental design



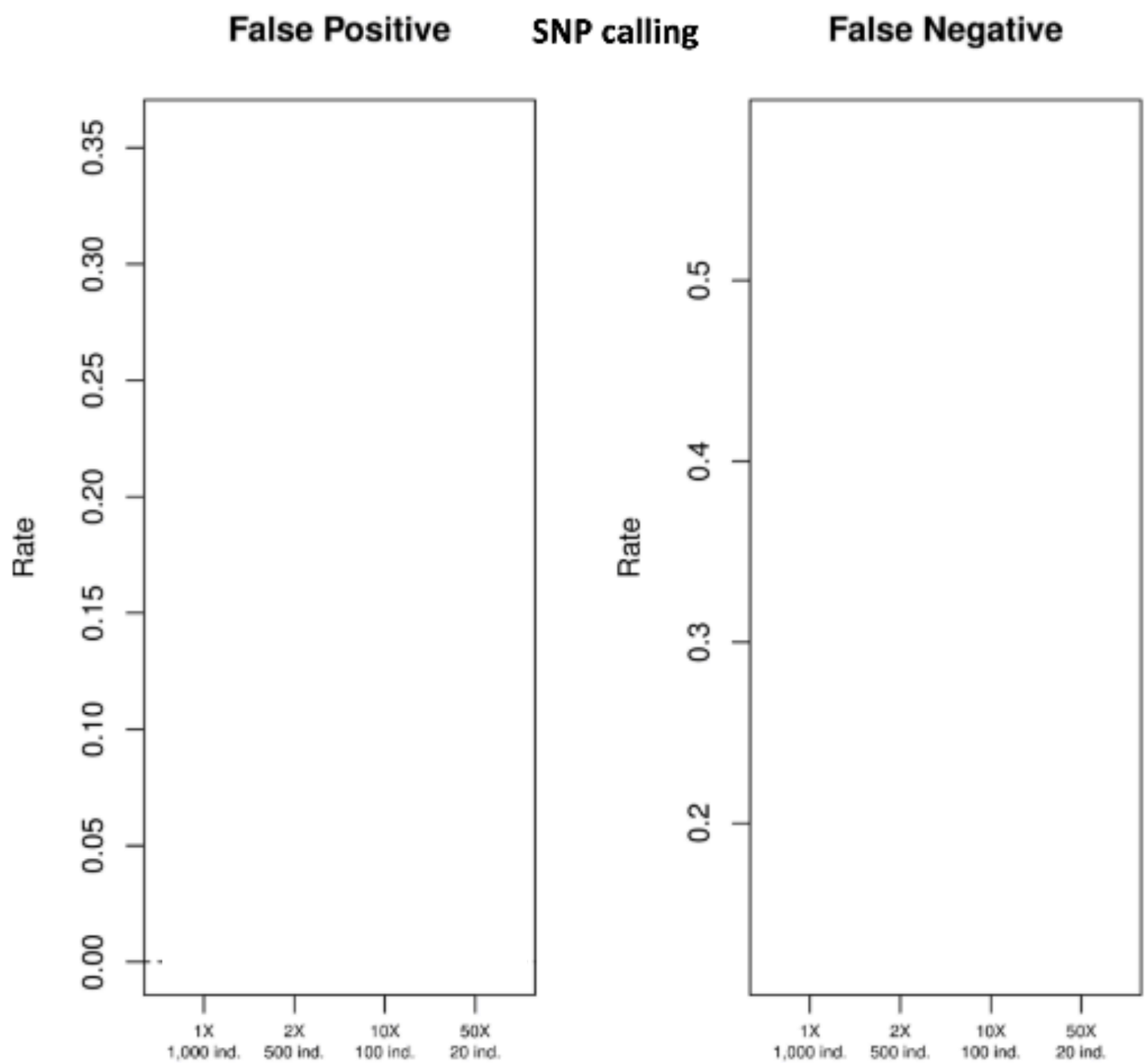
Fumagalli 2013

Experimental design



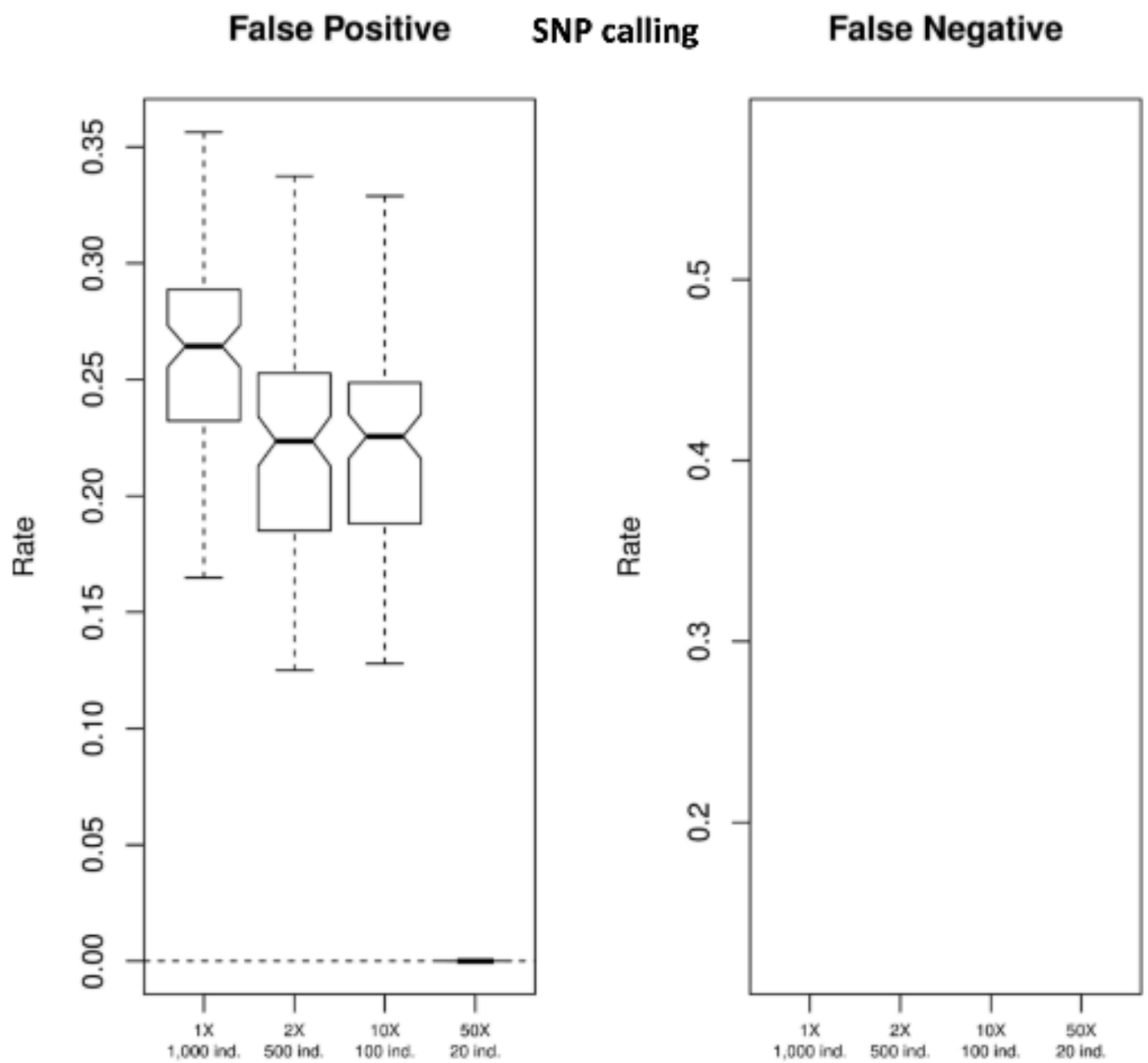
Fumagalli 2013

Experimental design



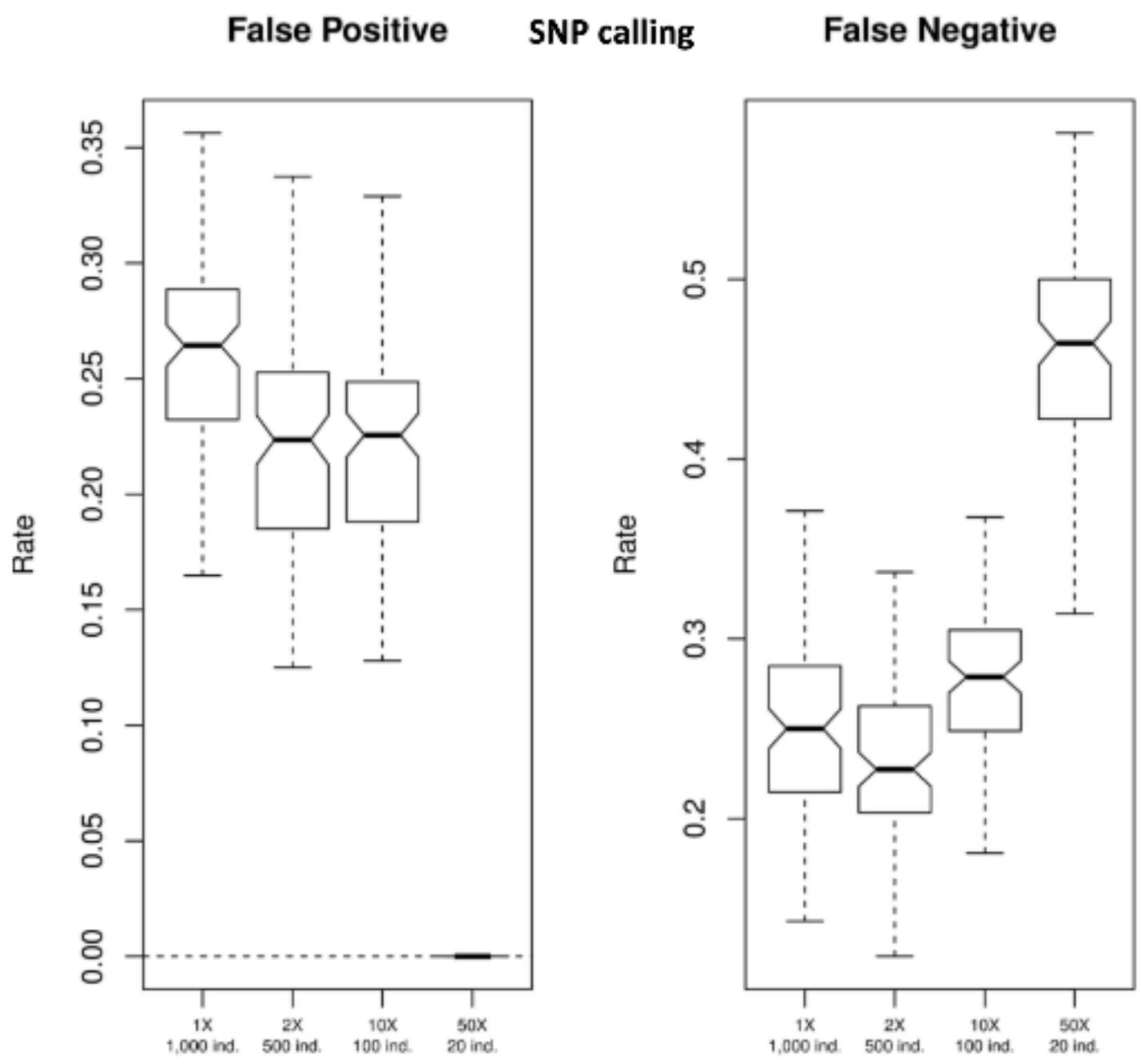
Fumagalli 2013

Experimental design



Fumagalli 2013

Experimental design

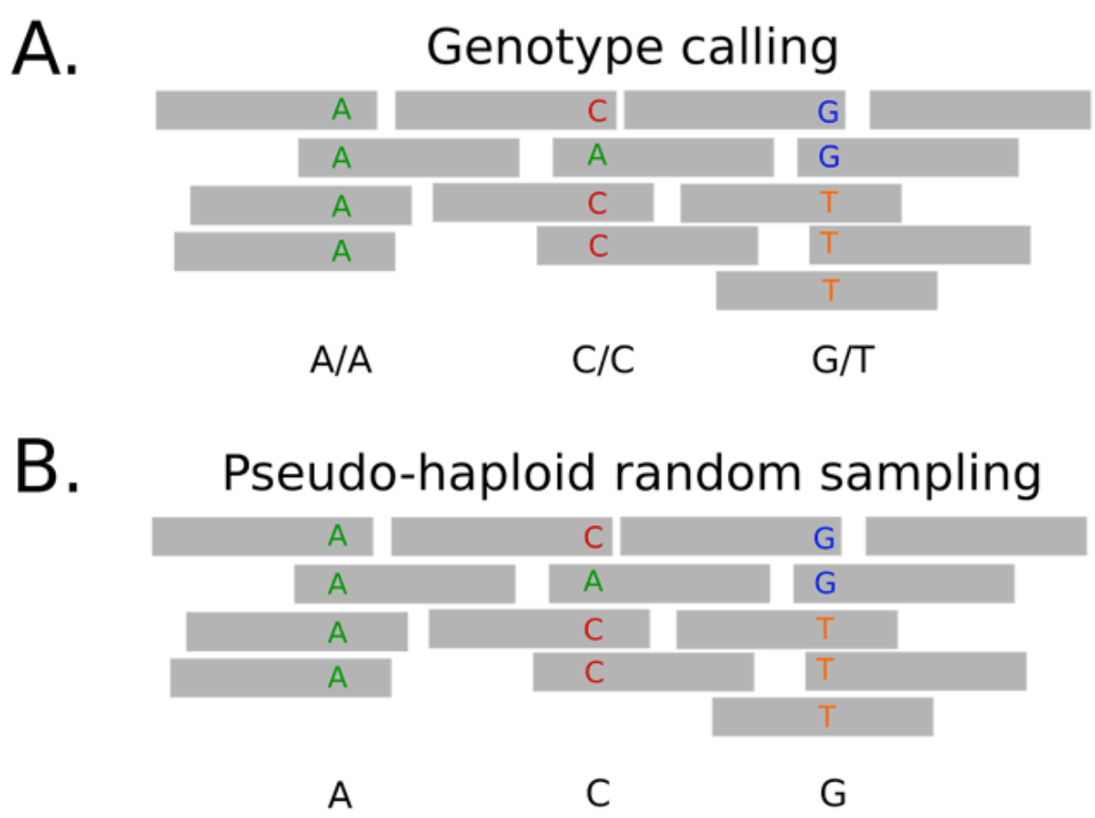


Fumagalli 2013

Today

- Experimental design
- **Data handling**
- PCA
- Spatial and isolation-by-distance methods

Alternative ways to deal with population genomic data



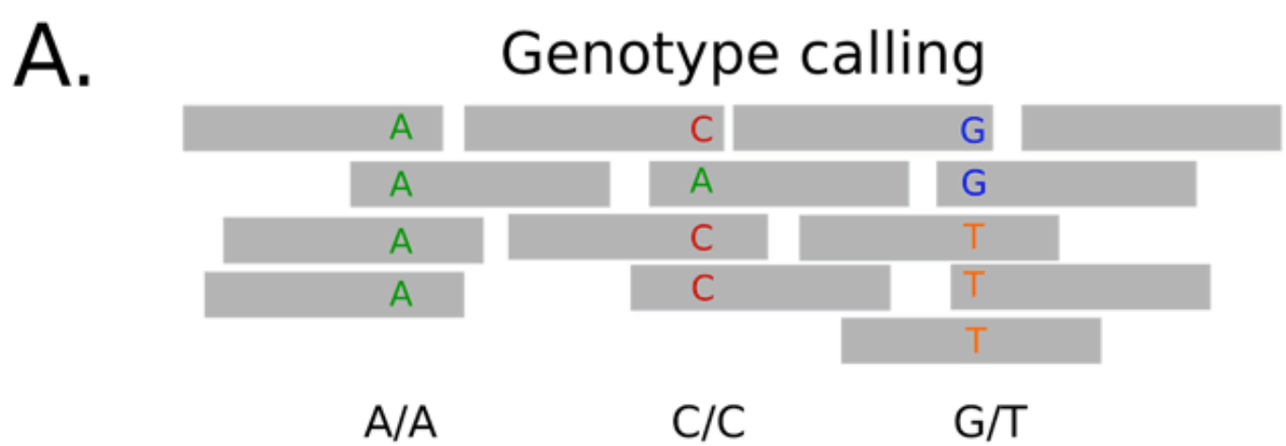
C.

Genotype likelihoods

Genotype	Likelihood	Genotype	Likelihood	Genotype	Likelihood
A/A	0.87	C/C	0.68	G/T	0.89
A/T	0.01	A/C	0.21	G/G	0.03
A/G	0.01	A/A	0.06	T/T	0.04
...		...		...	

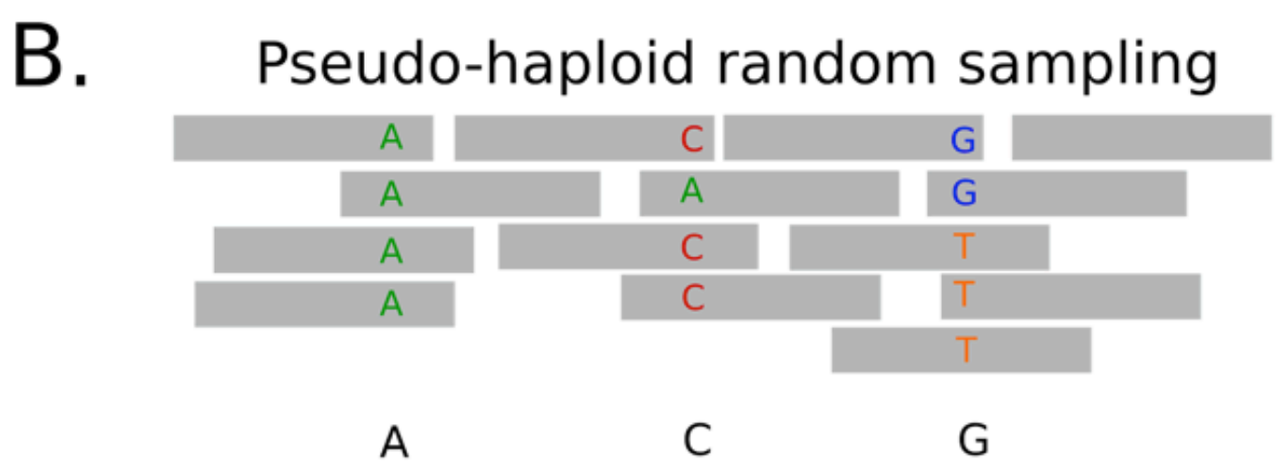
Genotype calling

- Generally requires high coverage data (> ~15X)
- Can lead to biases in comparisons with differences in coverage (more likely to over-call homozygous states on low-coverage data)
- Necessary for certain commonly used programs: PSMC, MSMC, etc.



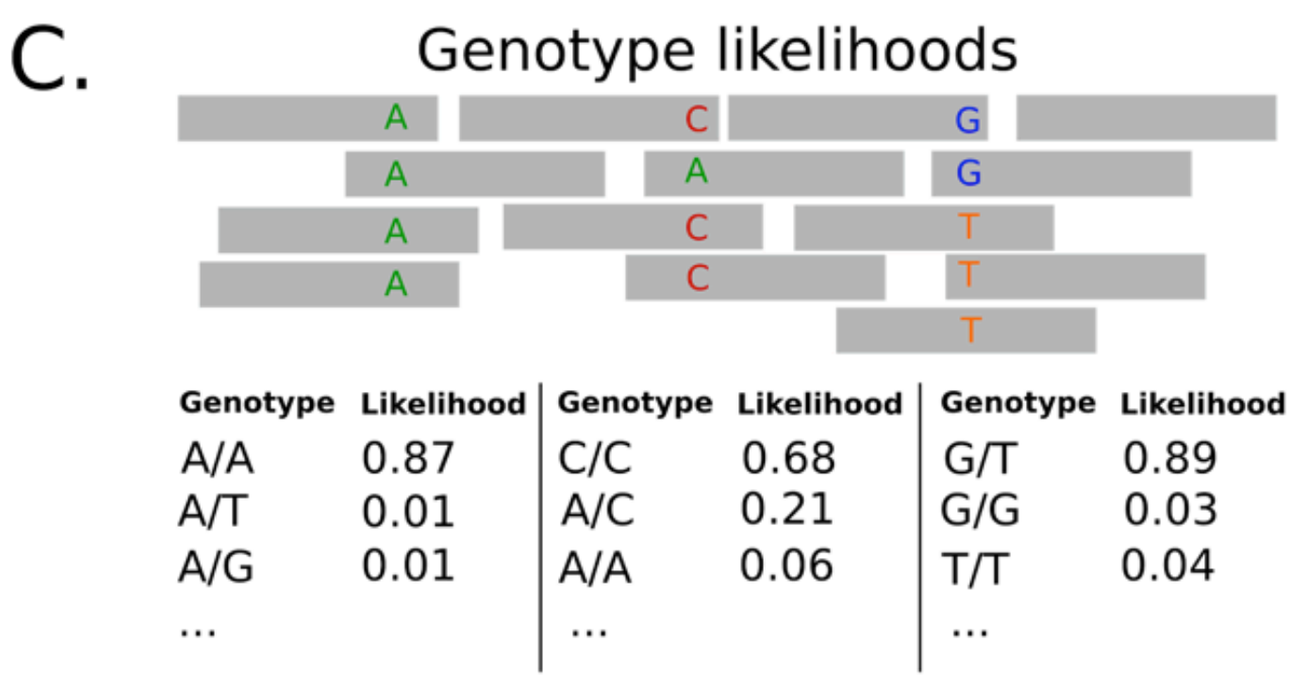
Pseudo-haploid sampling

- Unbiased with respect to differences in coverage
- Easy to produce and manipulate: we treat every diploid genome as haploid
- We lose information: we ignore all other reads that we do not sample!



Genotype likelihoods

- Uses the maximum amount of information
- Need a program to precompute likelihoods: ANGSD
- Need programs that can deal with genotype likelihoods: ngsAdmix, ngsTools, etc.
- Best for detecting selection (many individuals, low coverage data) -> good population allele frequency representation



Genotype likelihoods

- Genotype likelihood = P[data | a particular genotype]. The “log-likelihood” is the logarithm of the likelihood (easier to combine multiple probabilities: sums instead of products)

- 10 possible (unphased) genotypes: AA, AC, AG, AT, CC, CG, CT, GG, GT, TT

- Therefore, 10 log-likelihood values at each site, e.g. -10, -6.7, -8.3, -2.3, -3.5, -2.2, etc.

- Assuming we have M reads at a particular site:

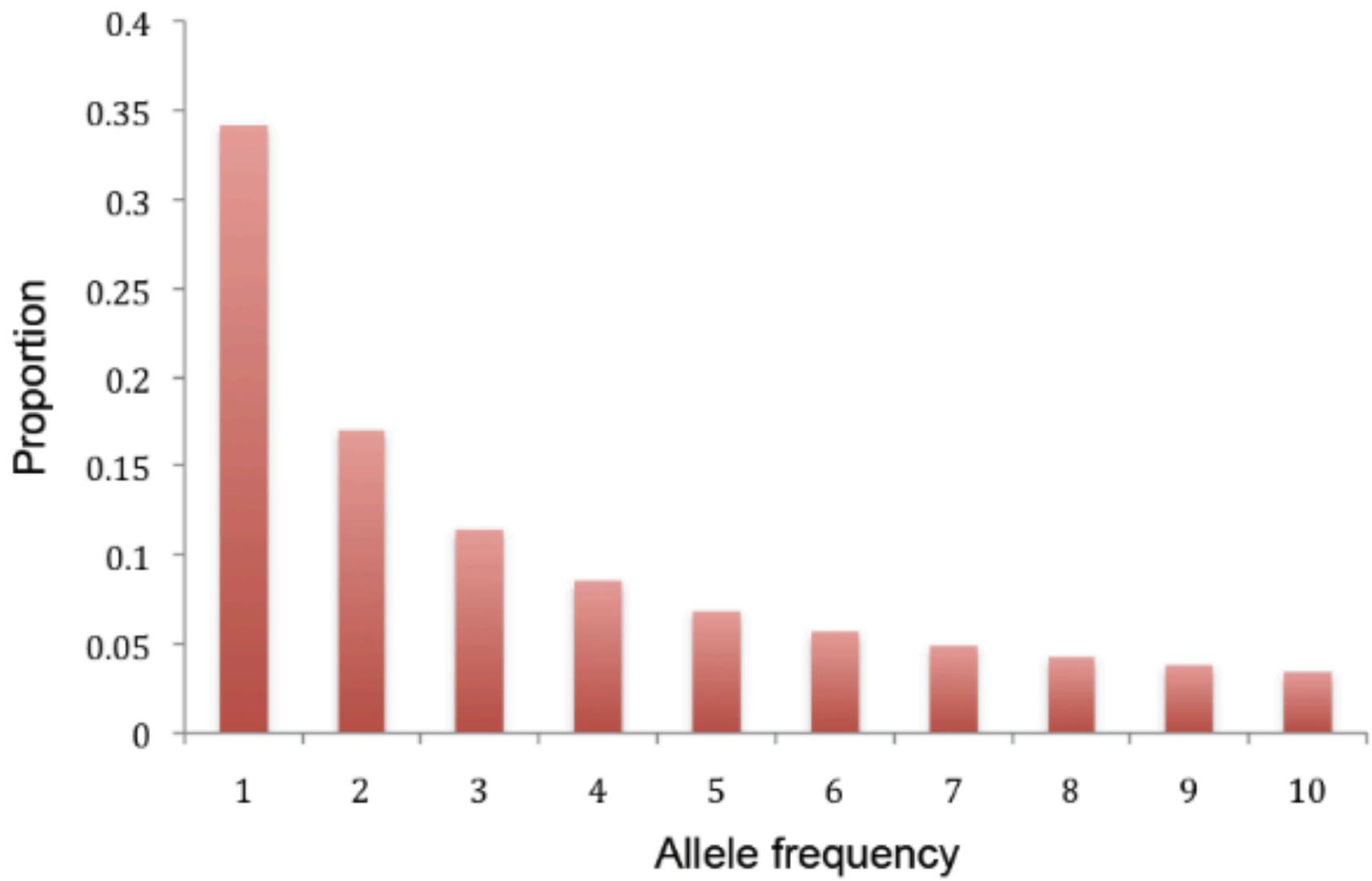
$$Pr\left(D|G=\{A_1,A_2\}\right)=\prod_{i=1}^{M}Pr\left(b_i|G=\{A_1,A_2\}\right)$$

$$=\prod_{i=1}^{M}\left(\frac{1}{2}p\left(b_i|A_1\right)+\frac{1}{2}p\left(b_i|A_2\right)\right),$$

$$p\left(b|A\right)=\begin{cases}\frac{e}{3} & b\neq A\\ 1-e & b=A.\end{cases}$$

Nielsen et al. 2012
Korneliussen et al. 2014

SFS likelihoods

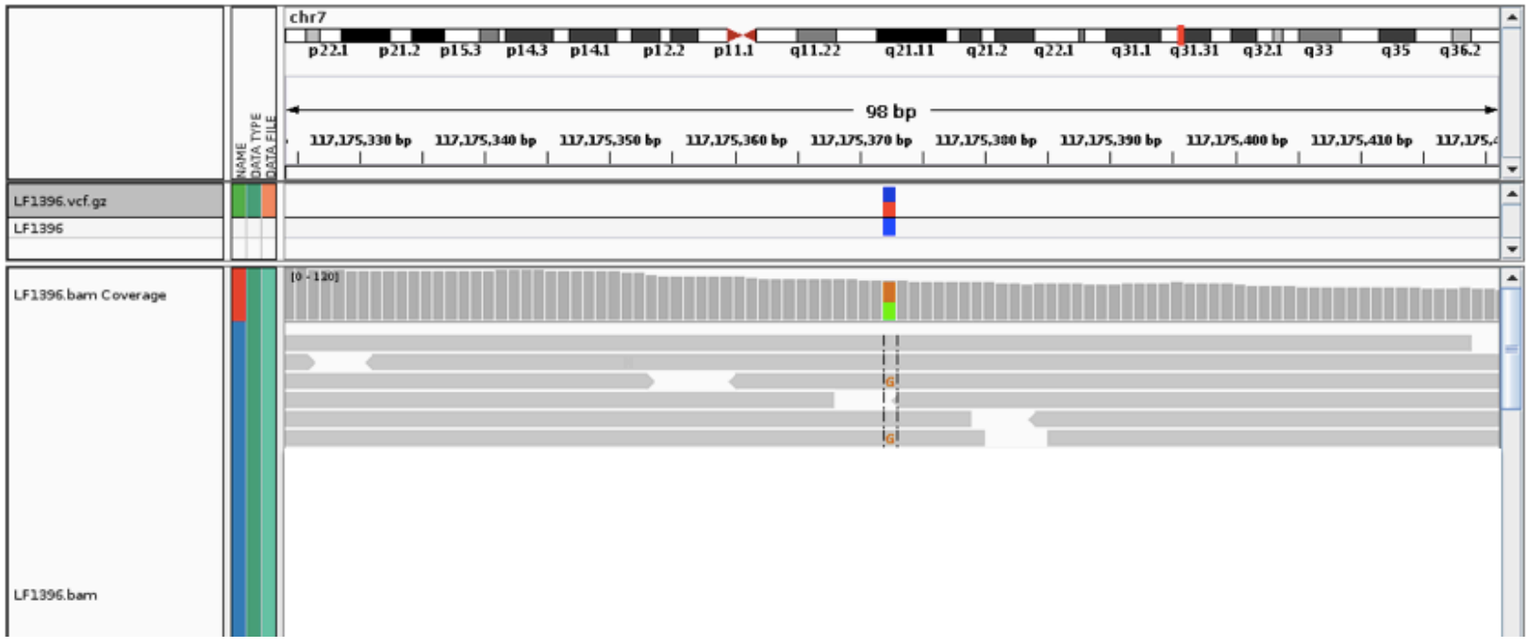


SFS likelihoods

- With low-coverage data, we don't have genotypes, so we cannot simply add up derived alleles to compute the SFS

- We can instead compute a likelihood **for each bin in the site-frequency spectrum**, given a set of reads from multiple individuals in a panel

- This approach is implemented in ANGSD and ngsTools



Fumagalli et al. 2014
Korneliussen et al. 2014

SFS likelihoods

- Let X be the sequencing data for our entire genome (all sites with ancestral and/or derived reads).
- X_s is the number of ancestral and derived reads at a particular site s.
- For 1 population, the SFS is a 1-dimensional vector $\vec{\gamma}$ with entries γ_i:
- $L(X|\gamma) = \prod_{s=1}^{N} L(X_s|\vec{\gamma}) = \prod_{s=1}^{N} \sum_{i=0}^{2n} \gamma_i P[X_s|D = i]$
- Then, we can use likelihood maximization algorithms to find a maximum likelihood estimate for each entry of the SFS (the values γ_i)

Nielsen et al. 2012
Korneliussen et al. 2014

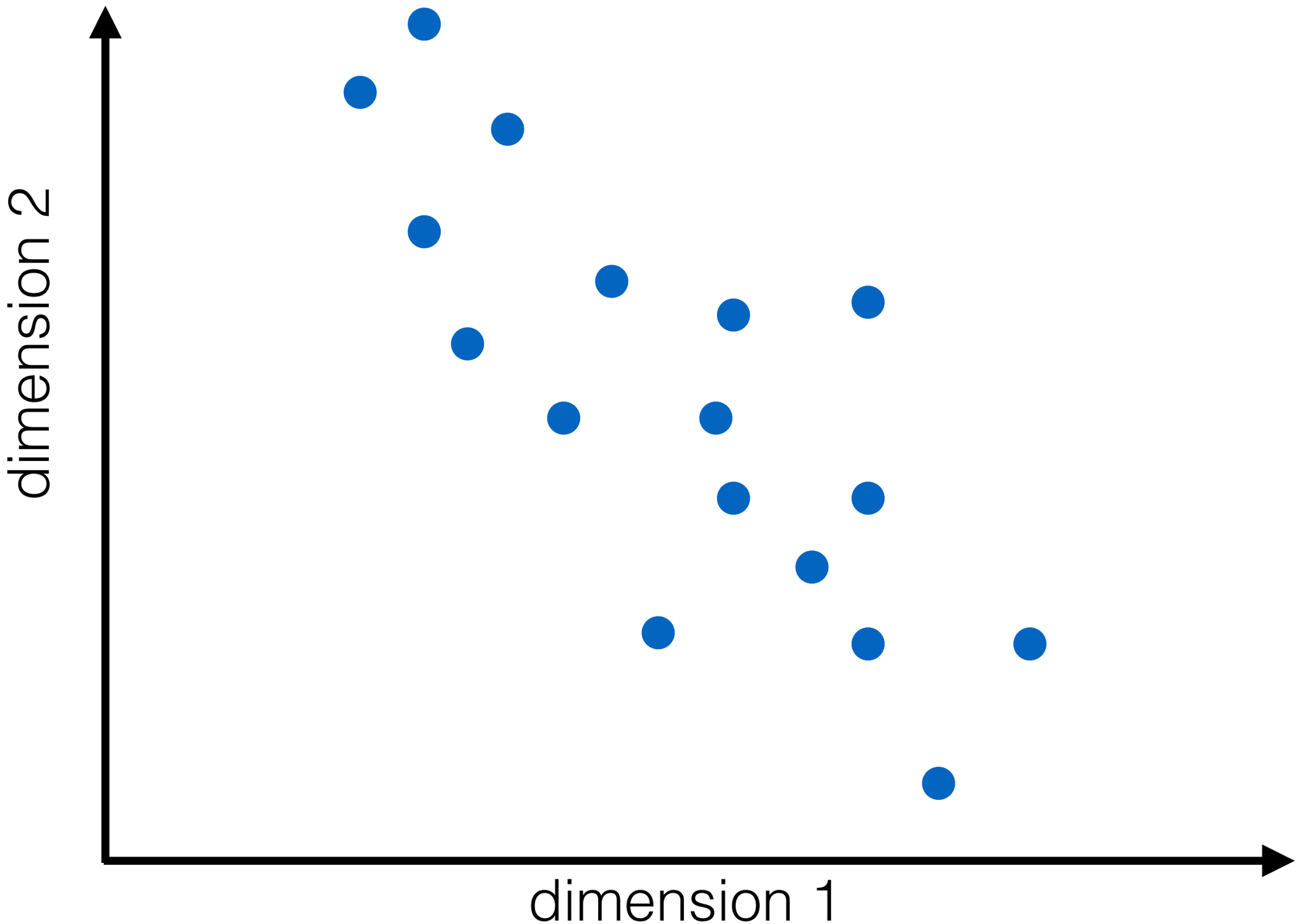
Today

- Experimental design
- Data handling
- **PCA**
- Spatial and isolation-by-distance methods

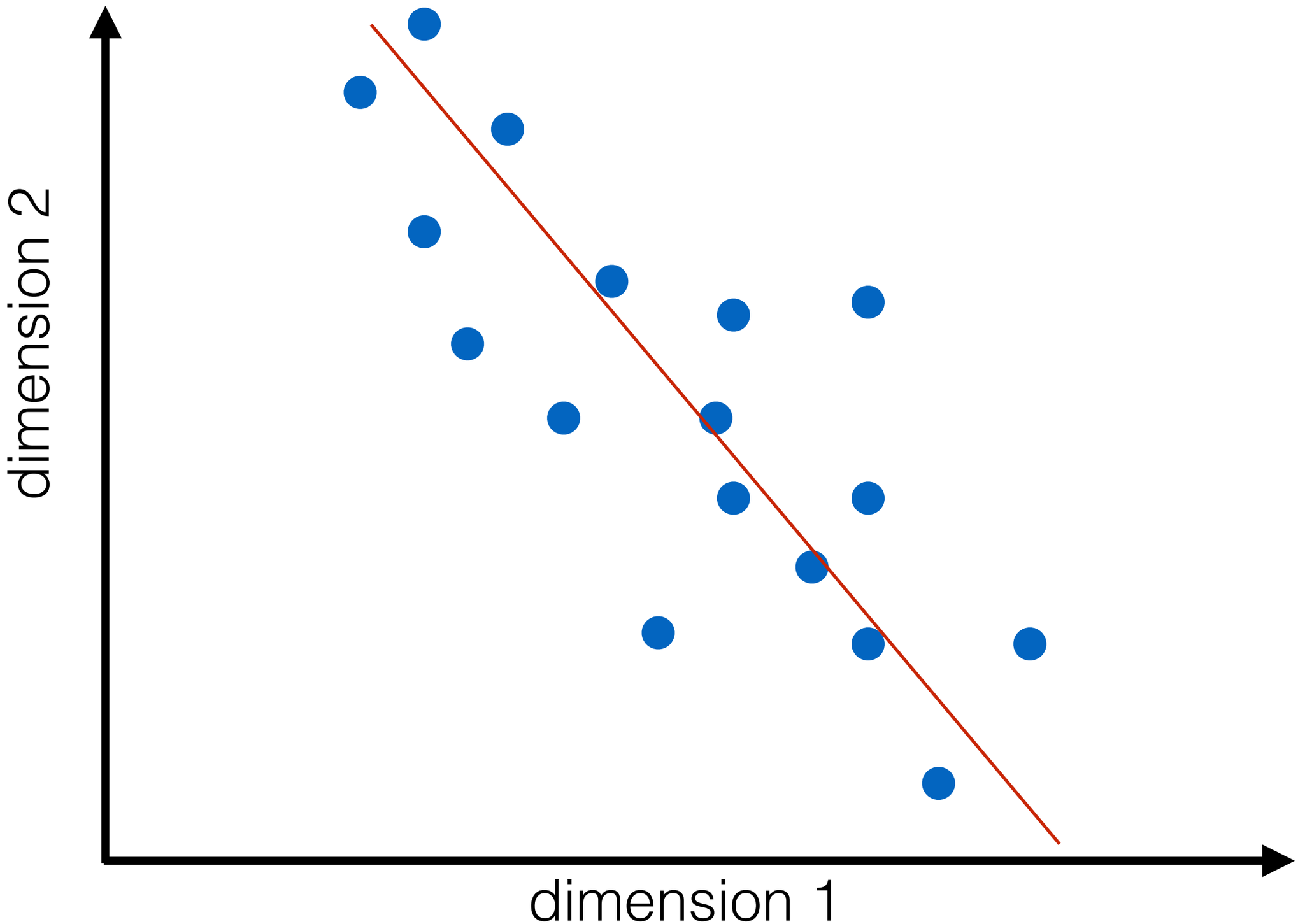
What is PCA?

- Principal Component Analysis: an orthogonal transformation of a set of observations of correlated variables into a set of values of linearly uncorrelated variables
- A technique for dimensionality reduction
- A technique for extracting the principal axes of variation in a dataset

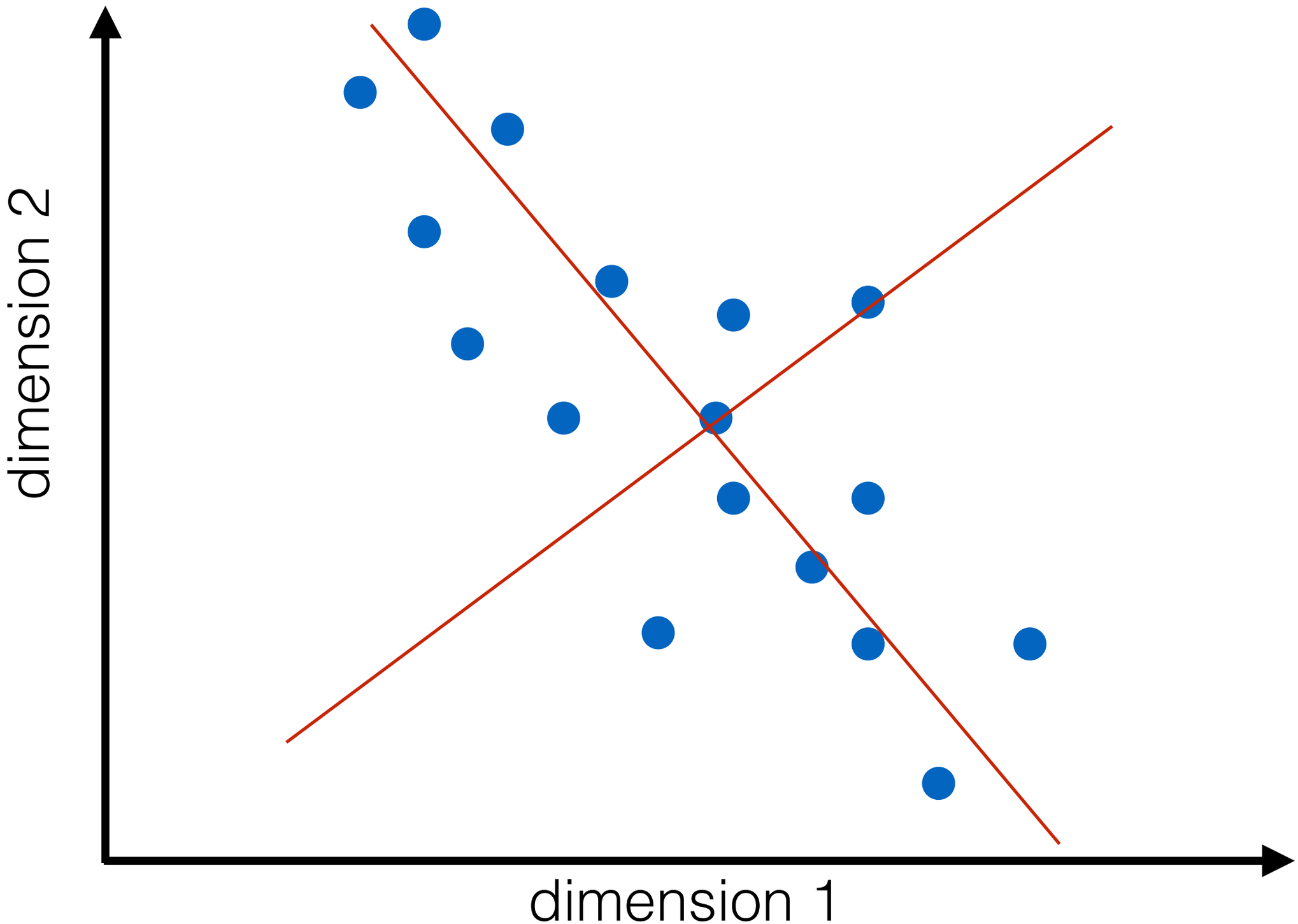
Finding the best **orthogonal** axes of variation



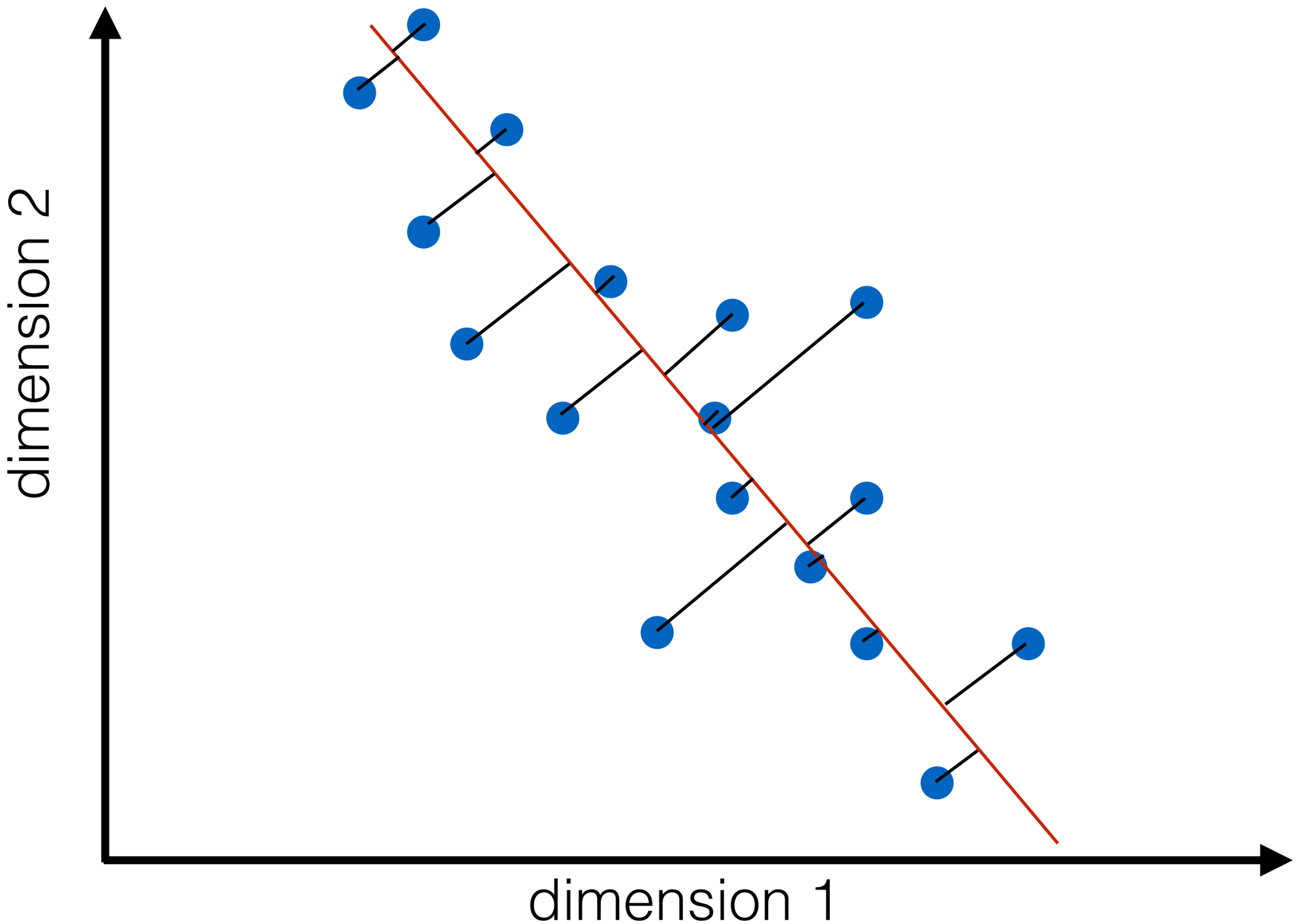
Finding the best **orthogonal** axes of variation



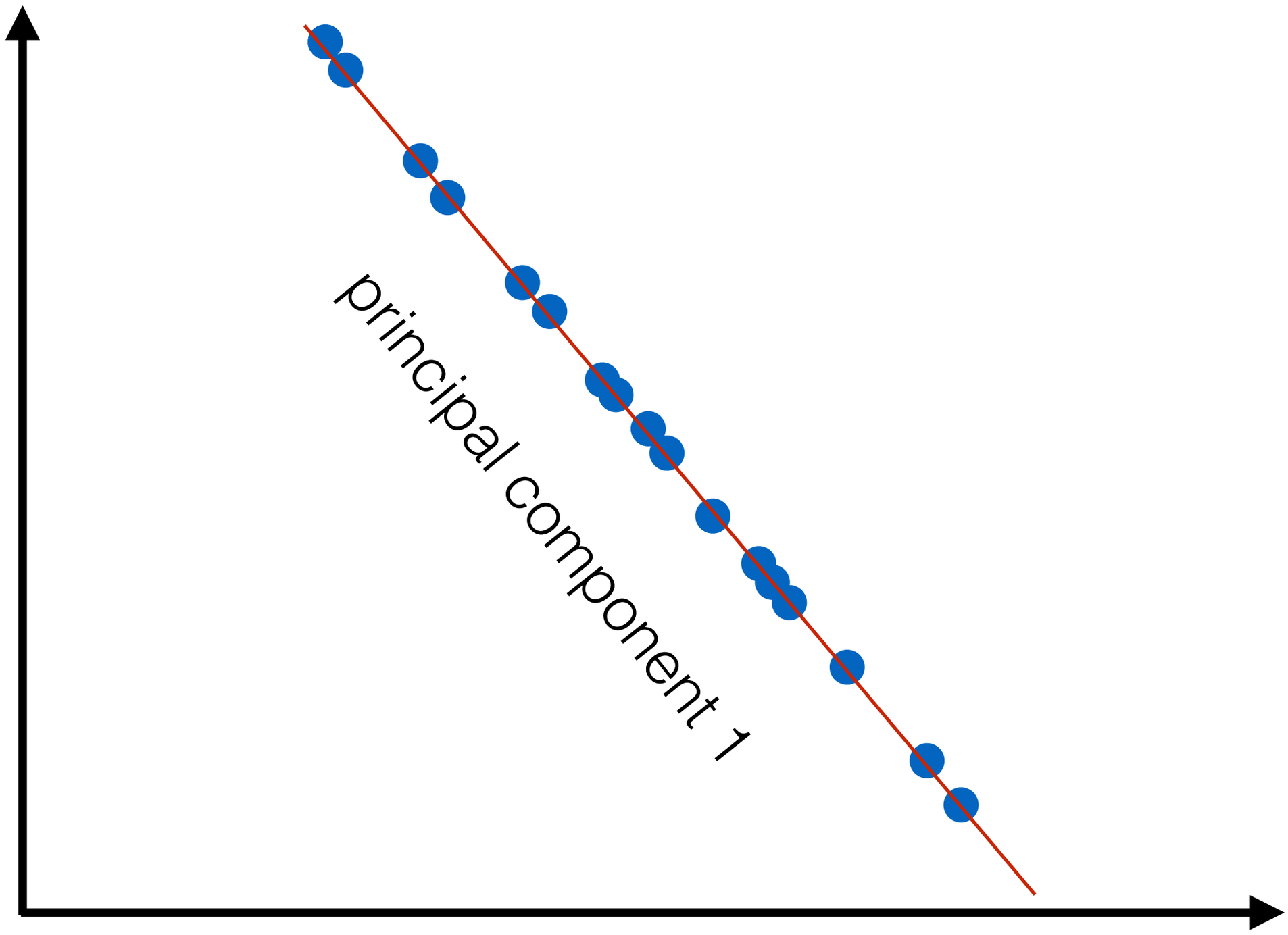
Finding the best **orthogonal** axes of variation



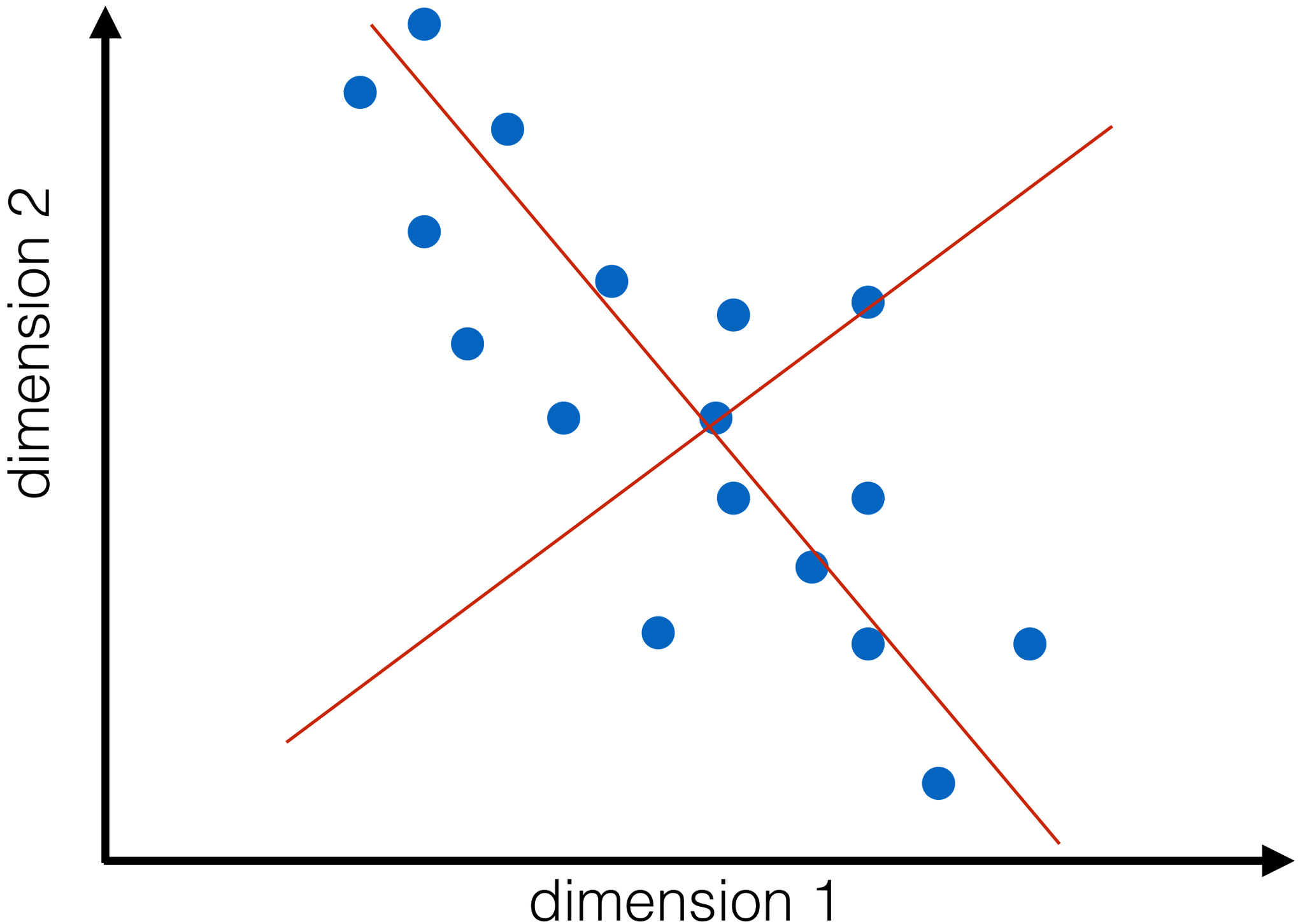
Projecting data onto **orthogonal** axes



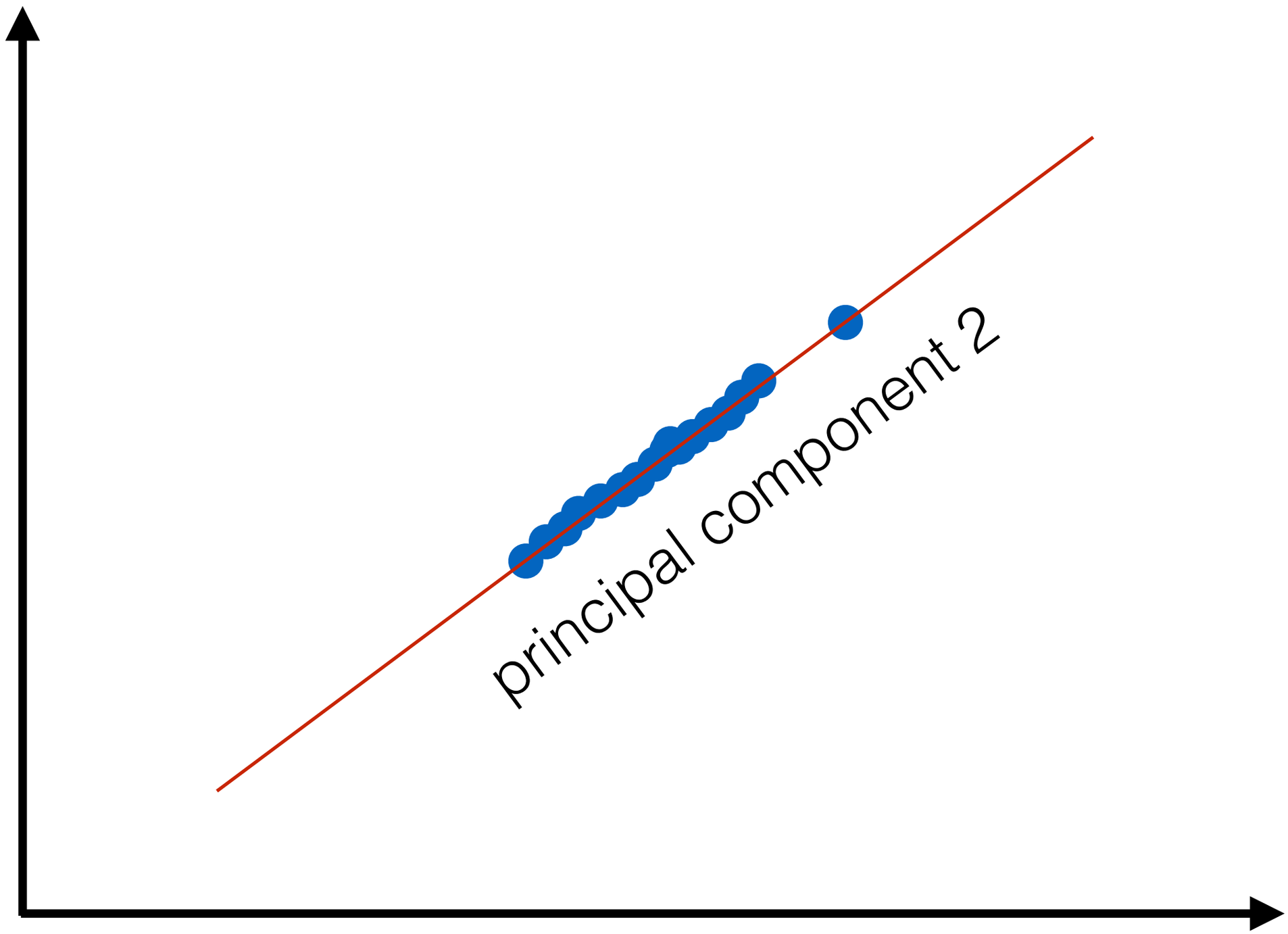
Projecting data onto **orthogonal** axes



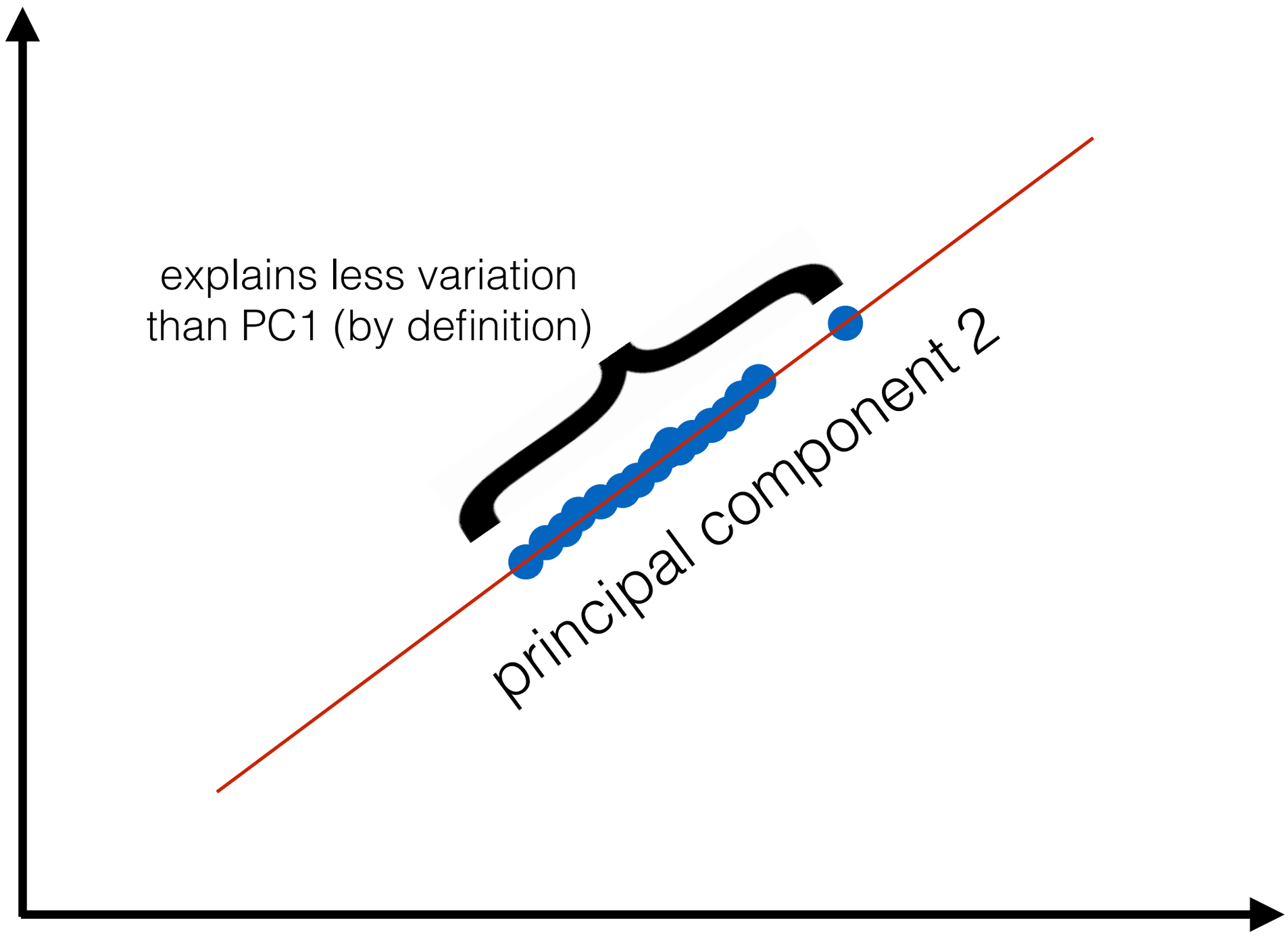
Projecting data onto **orthogonal** axes



Projecting data onto **orthogonal** axes



Projecting data onto **orthogonal** axes



Genotype data are multi-dimensional

- Each SNP is a dimension!

M individuals

N SNPs

$$
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
0 & 1 & 2 & 1 & 2 \\
2 & 1 & 1 & 0 & 1 \\
0 & 0 & 1 & 2 & 2 \\
2 & 1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 \\
2 & 2 & 1 & 1 & 0
\end{bmatrix}
\xrightarrow[\text{each SNP}]{\text{Mean-center}}
\begin{bmatrix}
0.4 & 0.4 & 0.4 & -0.6 & -0.6 \\
-1.2 & -0.2 & 0.8 & -0.2 & 0.8 \\
1.0 & 0.0 & 0.0 & -1.0 & 0.0 \\
-1.0 & -1.0 & 0.0 & 1.0 & 1.0 \\
1.2 & 0.2 & 0.2 & -0.8 & -0.8 \\
-0.6 & -0.6 & 0.4 & 0.4 & 0.4 \\
0.8 & 0.8 & -0.2 & -0.2 & -1.2
\end{bmatrix}
= \mathbf{X}
$$

Solution: eigen-decomposition of covariance matrix

Solution: eigen-decomposition of covariance matrix

1) Multiply **X** by itself: $\mathbf{X}^T\mathbf{X} = \mathbf{C} \longrightarrow$ covariance matrix (M x M)

Solution: eigen-decomposition of covariance matrix

1) Multiply **X** by itself: $\mathbf{X^T X = C} \longrightarrow$ covariance matrix (M x M)

$$\begin{bmatrix} V_a & C_{a,b} & C_{a,c} & C_{a,d} & C_{a,e} \\ C_{a,b} & V_b & C_{b,c} & C_{b,d} & C_{b,e} \\ C_{a,c} & C_{b,c} & V_c & C_{c,d} & C_{c,e} \\ C_{a,d} & C_{b,d} & C_{c,d} & V_d & C_{d,e} \\ C_{a,e} & C_{b,e} & C_{c,e} & C_{d,e} & V_e \end{bmatrix}$$

Solution: eigen-decomposition of covariance matrix

1) Multiply **X** by itself: $\mathbf{X}^T\mathbf{X} = \mathbf{C} \longrightarrow$ covariance matrix (M x M)

Solution: eigen-decomposition of covariance matrix

1) Multiply **X** by itself: $\mathbf{X}^T\mathbf{X} = \mathbf{C}$ → covariance matrix (M x M)

2) Eigen-decompose **C**: $\mathbf{C} = \mathbf{V}\mathbf{D}\mathbf{V}^T$

Solution: eigen-decomposition of covariance matrix

1) Multiply **X** by itself: $\mathbf{X^T X = C} \longrightarrow$ covariance matrix (M x M)

2) Eigen-decompose **C**: $\mathbf{C = V D V^T}$

$$\mathbf{X} = \begin{matrix} 2 & -2 \\ 1 & -1 \\ 0 & 0 \\ -1 & 1 \\ -2 & 2 \end{matrix} \qquad \mathbf{C} = \begin{matrix} 10 & -10 \\ -10 & 10 \end{matrix} = \underbrace{\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} 20 & 0 \\ 0 & 0 \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V^T}}$$

Solution: eigen-decomposition of covariance matrix

1) Multiply **X** by itself: $\mathbf{X^T X = C} \longrightarrow$ covariance matrix (M x M)

2) Eigen-decompose **C**: $\mathbf{C = V D V^T}$

$$\mathbf{X} = \begin{matrix} 2 & -2 \\ 1 & -1 \\ 0 & 0 \\ -1 & 1 \\ -2 & 2 \end{matrix} \qquad \mathbf{C} = \begin{matrix} 10 & -10 \\ -10 & 10 \end{matrix} = \underbrace{\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} 20 & 0 \\ 0 & 0 \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V^T}}$$

3) Columns of V are the eigenvectors of C (PCs): $\mathbf{C} * \mathbf{V_1} = D_{1,1} * \mathbf{V_1}$

Solution: eigen-decomposition of covariance matrix

1) Multiply **X** by itself: $\mathbf{X^T X = C} \longrightarrow$ covariance matrix (M x M)

2) Eigen-decompose **C**: $\mathbf{C = V D V^T}$

$$\mathbf{X} = \begin{matrix} 2 & -2 \\ 1 & -1 \\ 0 & 0 \\ -1 & 1 \\ -2 & 2 \end{matrix} \qquad \mathbf{C} = \begin{matrix} 10 & -10 \\ -10 & 10 \end{matrix} = \underbrace{\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} 20 & 0 \\ 0 & 0 \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V^T}}$$

3) Columns of V are the eigenvectors of C (PCs): $\mathbf{C} * \mathbf{V_1} = D_{1,1} * \mathbf{V_1}$

4) Diagonal entries of D are the eigenvalues of C

Solution: eigen-decomposition of covariance matrix

1) Multiply **X** by itself: $\mathbf{X^T X = C} \longrightarrow$ covariance matrix (M x M)

2) Eigen-decompose **C**: $\mathbf{C = V D V^T}$

$$\mathbf{X} = \begin{matrix} 2 & -2 \\ 1 & -1 \\ 0 & 0 \\ -1 & 1 \\ -2 & 2 \end{matrix} \qquad \mathbf{C} = \begin{matrix} 10 & -10 \\ -10 & 10 \end{matrix} = \underbrace{\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} 20 & 0 \\ 0 & 0 \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V^T}}$$

3) Columns of V are the eigenvectors of C (PCs): $\mathbf{C} * \mathbf{V_1} = D_{1,1} * \mathbf{V_1}$

4) Diagonal entries of D are the eigenvalues of C

5) Eigenvectors with largest eigenvalues are top PCs

Solution: eigen-decomposition of covariance matrix

1) Multiply **X** by itself: $\mathbf{X^T X = C} \longrightarrow$ covariance matrix (M x M)

2) Eigen-decompose **C**: $\mathbf{C = V D V^T}$

$$\mathbf{X} = \begin{matrix} 2 & -2 \\ 1 & -1 \\ 0 & 0 \\ -1 & 1 \\ -2 & 2 \end{matrix} \qquad \mathbf{C} = \begin{matrix} 10 & -10 \\ -10 & 10 \end{matrix} = \underbrace{\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} 20 & 0 \\ 0 & 0 \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V^T}}$$

PC1

3) Columns of V are the eigenvectors of C (PCs): $\mathbf{C} * \mathbf{V_1} = D_{1,1} * \mathbf{V_1}$

4) Diagonal entries of D are the eigenvalues of C

5) Eigenvectors with largest eigenvalues are top PCs

Solution: eigen-decomposition of covariance matrix

1) Multiply **X** by itself: $\mathbf{X^T X = C} \longrightarrow$ covariance matrix (M x M)

2) Eigen-decompose **C**: $\mathbf{C = V D V^T}$

$$\mathbf{X} = \begin{matrix} 2 & -2 \\ 1 & -1 \\ 0 & 0 \\ -1 & 1 \\ -2 & 2 \end{matrix} \qquad \mathbf{C} = \begin{matrix} 10 & -10 \\ -10 & 10 \end{matrix} = \underbrace{\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} 20 & 0 \\ 0 & 0 \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V^T}}$$

PC1 **PC2**

3) Columns of V are the eigenvectors of C (PCs): $\mathbf{C} * \mathbf{V_1} = D_{1,1} * \mathbf{V_1}$

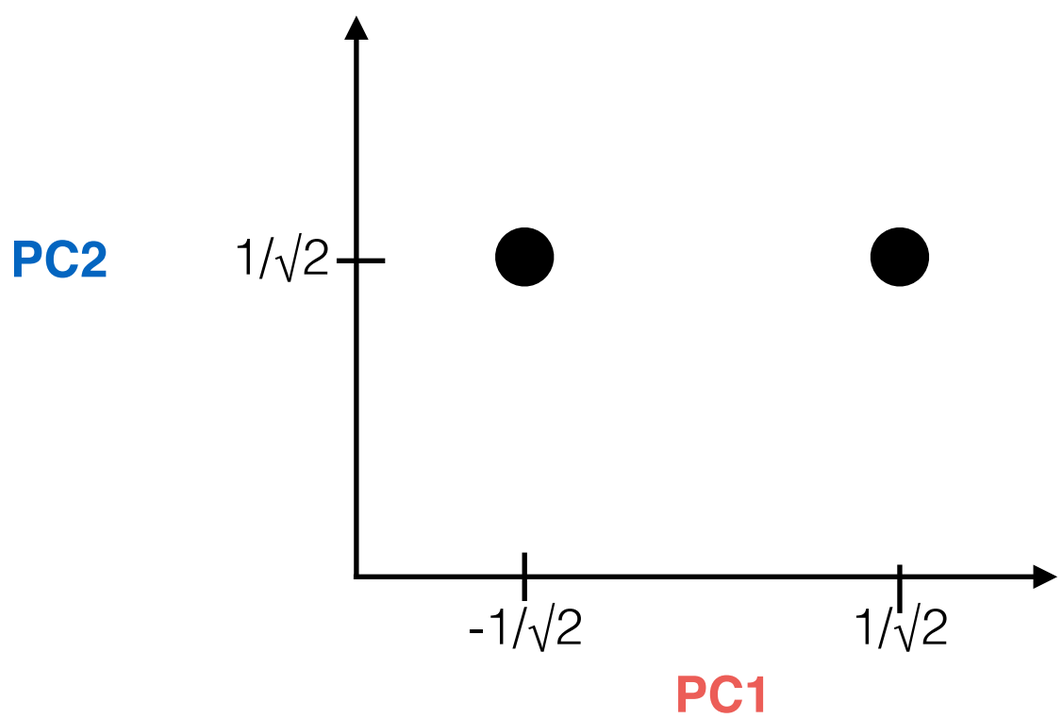
4) Diagonal entries of D are the eigenvalues of C

5) Eigenvectors with largest eigenvalues are top PCs

Solution: eigen-decomposition of covariance matrix

$$\mathbf{X} = \begin{matrix} 2 & -2 \\ 1 & -1 \\ 0 & 0 \\ -1 & 1 \\ -2 & 2 \end{matrix} \qquad \mathbf{C} = \begin{matrix} 10 & -10 \\ -10 & 10 \end{matrix} = \underbrace{\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} 20 & 0 \\ 0 & 0 \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V}^{\mathbf{T}}}$$

PC1 **PC2**

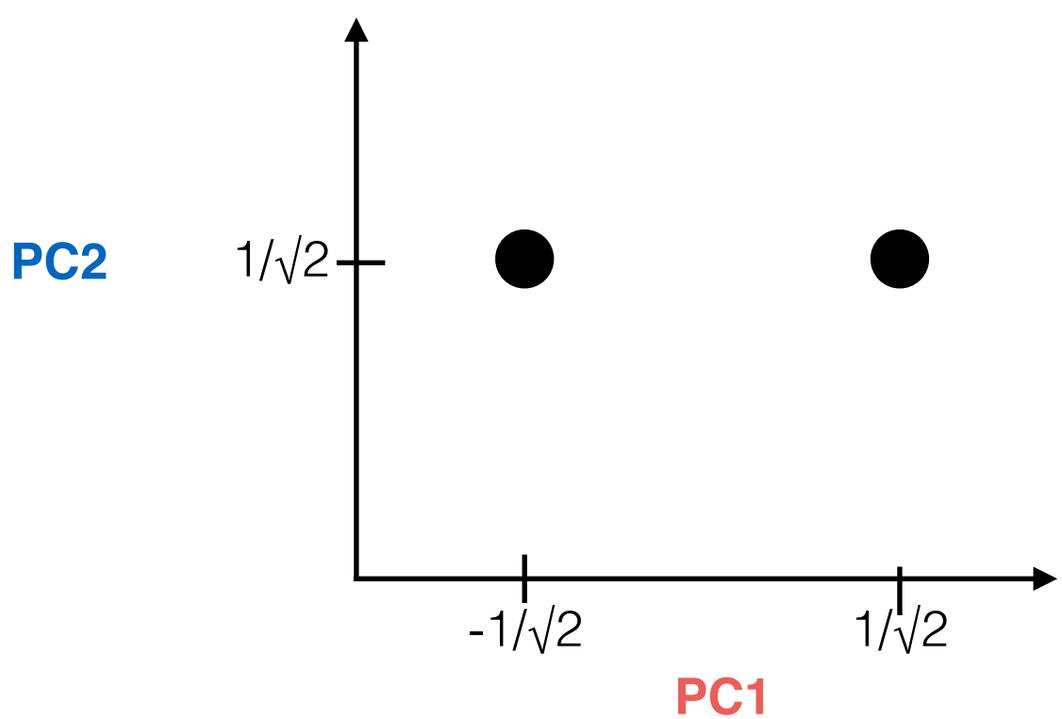


Solution: eigen-decomposition of covariance matrix

$$\mathbf{X} = \begin{matrix} 2 & -2 \\ 1 & -1 \\ 0 & 0 \\ -1 & 1 \\ -2 & 2 \end{matrix} \qquad \mathbf{C} = \begin{matrix} 10 & -10 \\ -10 & 10 \end{matrix} = \underbrace{\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} 20 & 0 \\ 0 & 0 \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V}^{\mathbf{T}}}$$

PC1 **PC2**

The **relative sizes** of the eigenvalues indicate the **proportion of total variance** each eigenvector explains

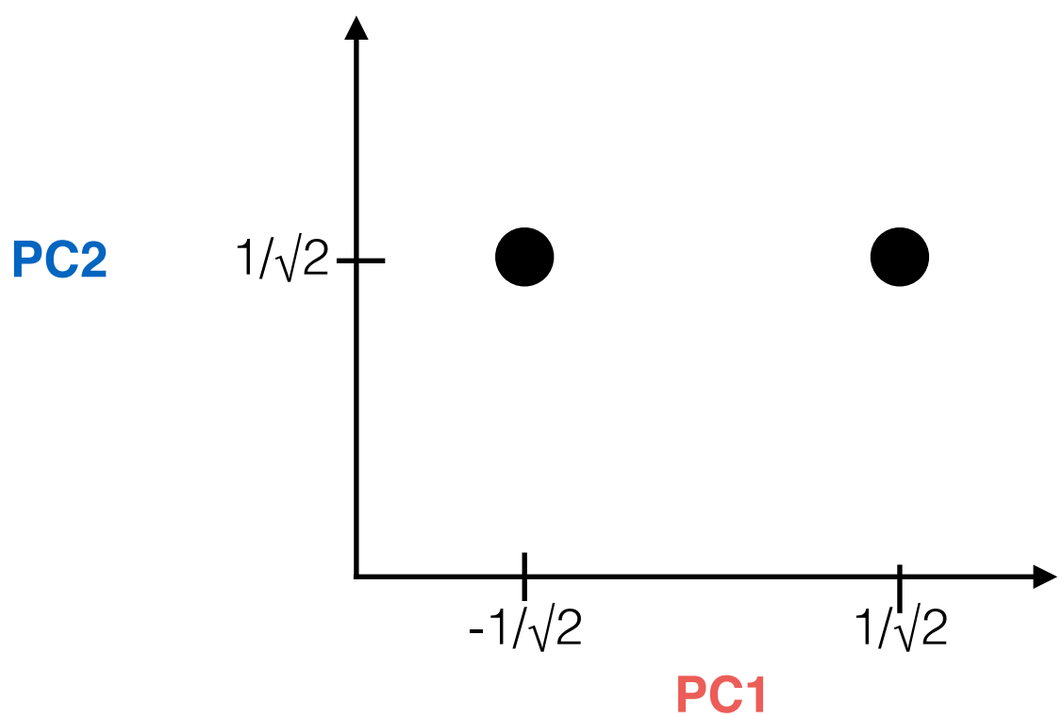


Solution: eigen-decomposition of covariance matrix

$$\mathbf{X} = \begin{matrix} 2 & -2 \\ 1 & -1 \\ 0 & 0 \\ -1 & 1 \\ -2 & 2 \end{matrix} \qquad \mathbf{C} = \begin{matrix} 10 & -10 \\ -10 & 10 \end{matrix} = \underbrace{\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} 20 & 0 \\ 0 & 0 \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}}_{\mathbf{V}^{\mathbf{T}}}$$

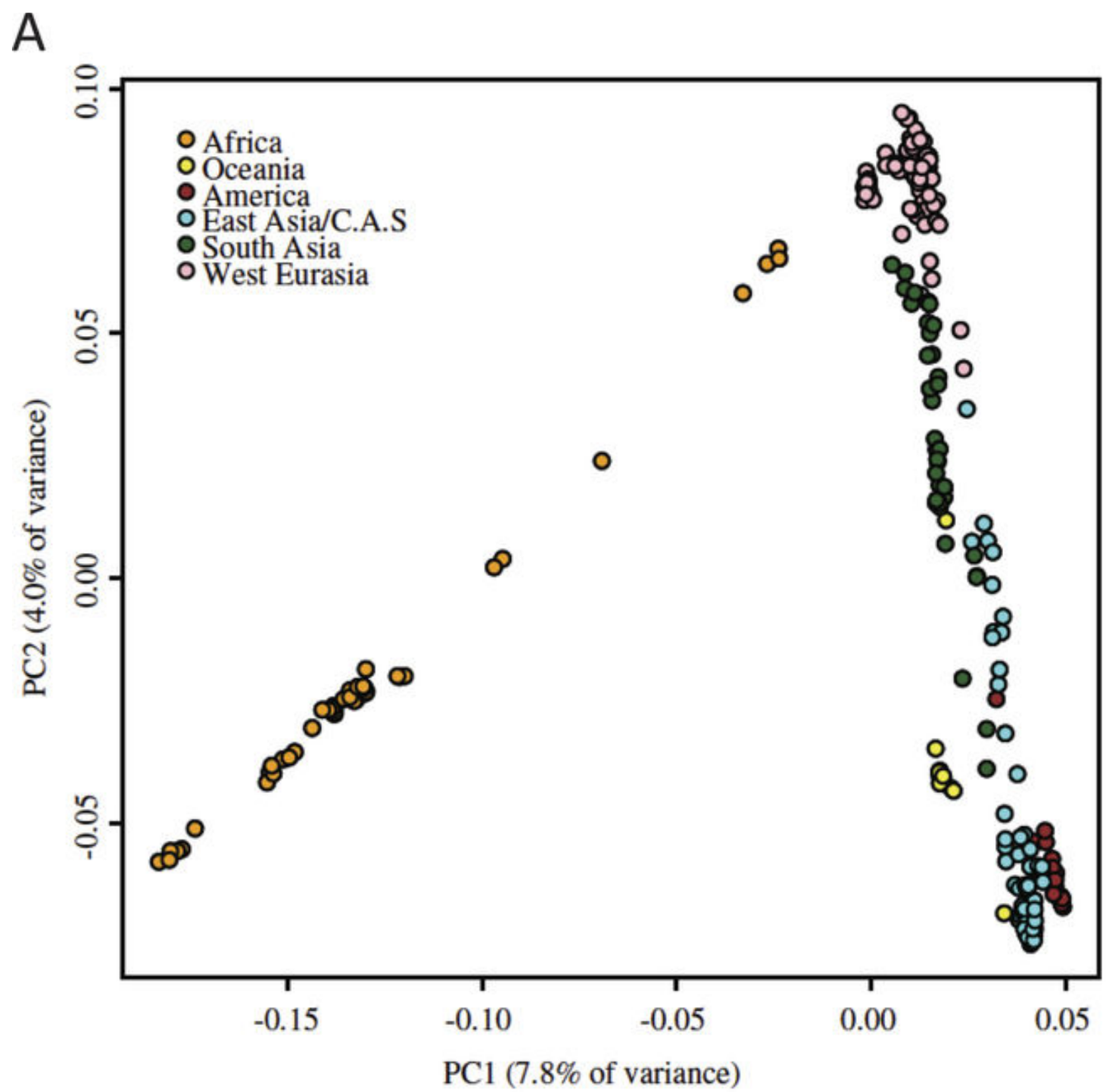
PC1 **PC2**

The **relative sizes** of the eigenvalues indicate the **proportion of total variance** each eigenvector explains

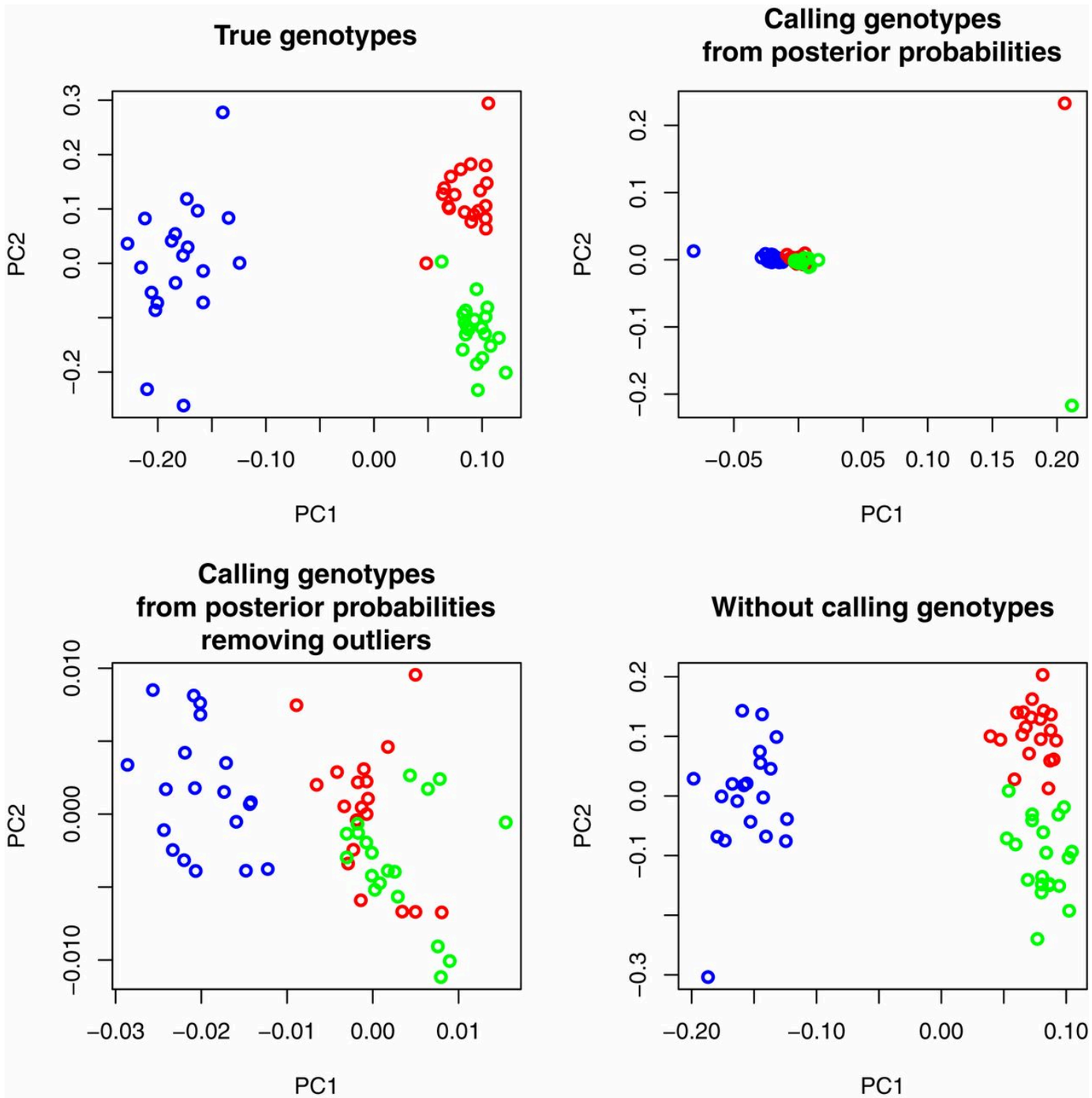


In this toy example, PC1 explains ALL of the variance

PCA of worldwide human genomes



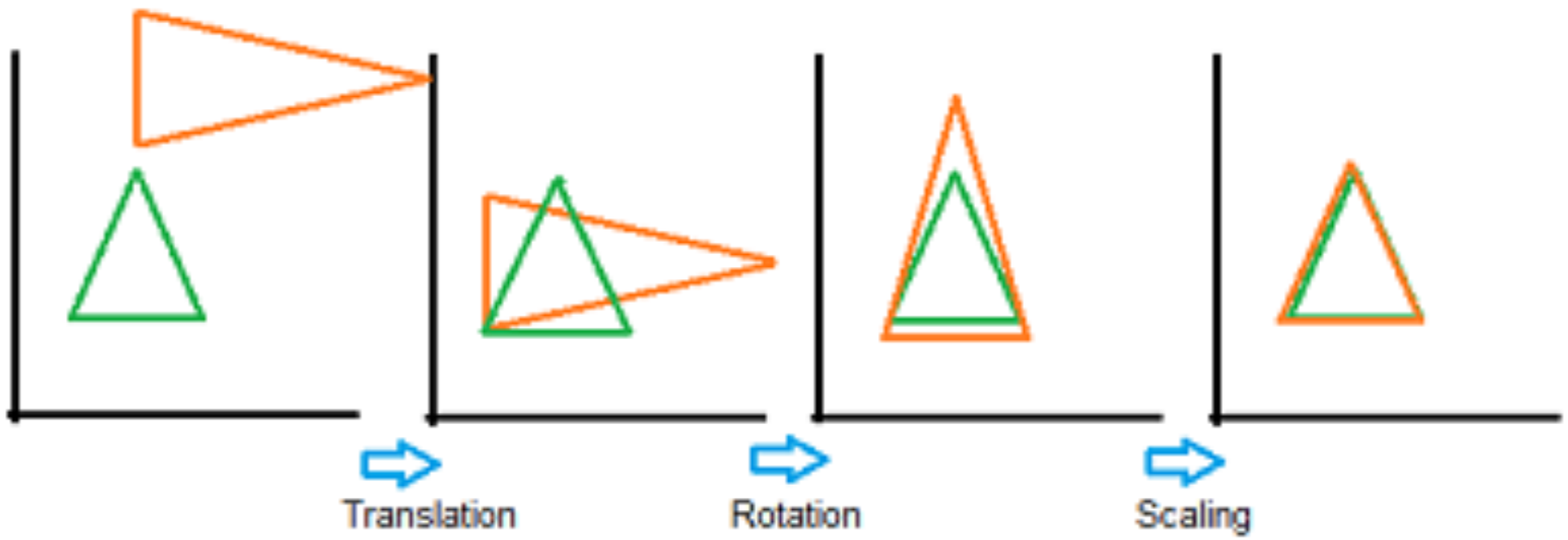
PCA from genotype likelihoods



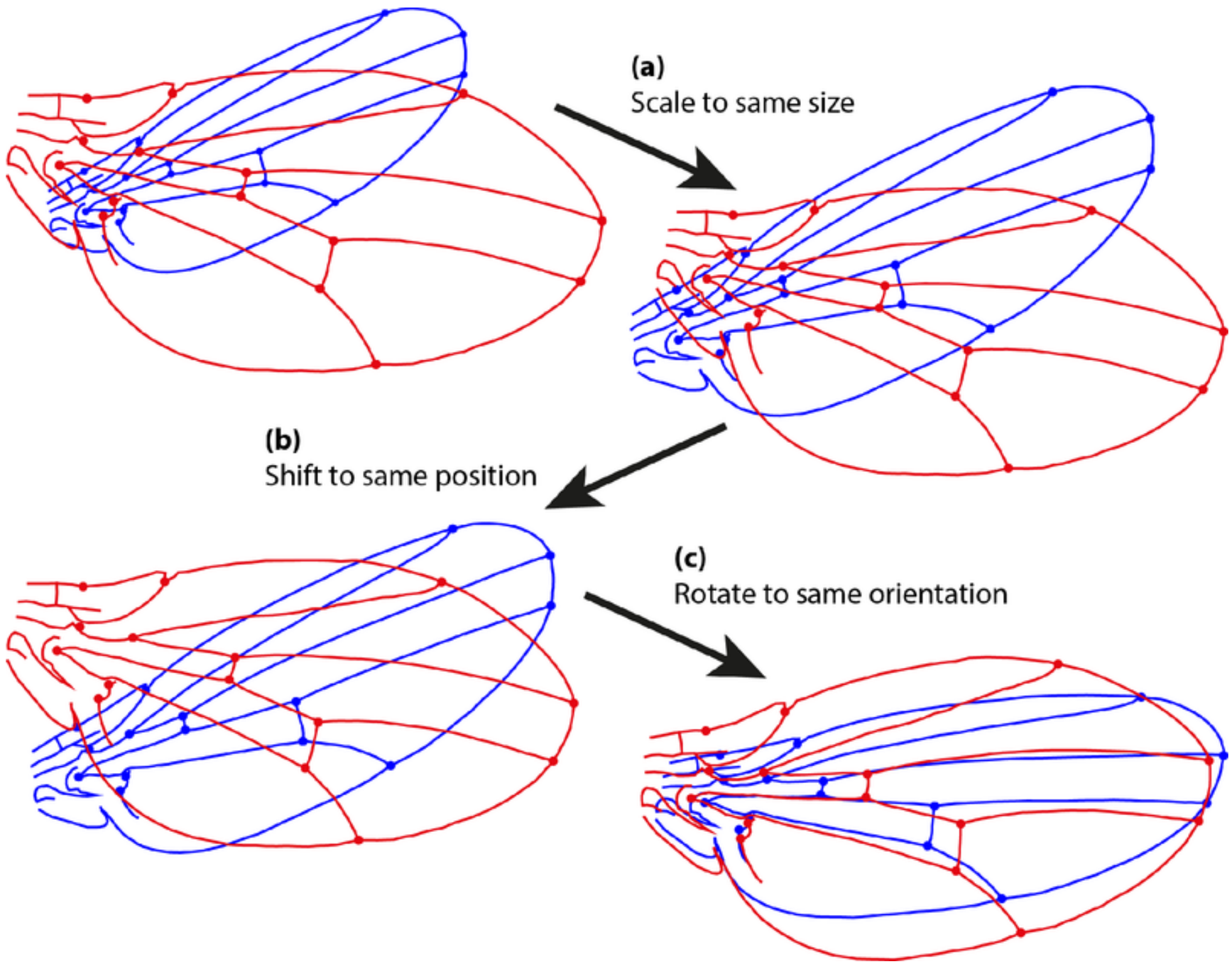
Dealing with missing data: Procrustes transformation

- SNPs in which at least 1 sample has missing data are unusable in a PCA
- Problem: low coverage genomes -> many sites with missing data
- Even bigger problem: combination of many low-coverage genomes -> very few sites with overlap in coverage across all of them
- Solution (Skoglund et al. 2012):
    - For each low-coverage genome, run 1 PCA (with many high-coverage genomes included)
    - Combine loadings from each individual PCA into an overall-PCA, using Procrustes transformation

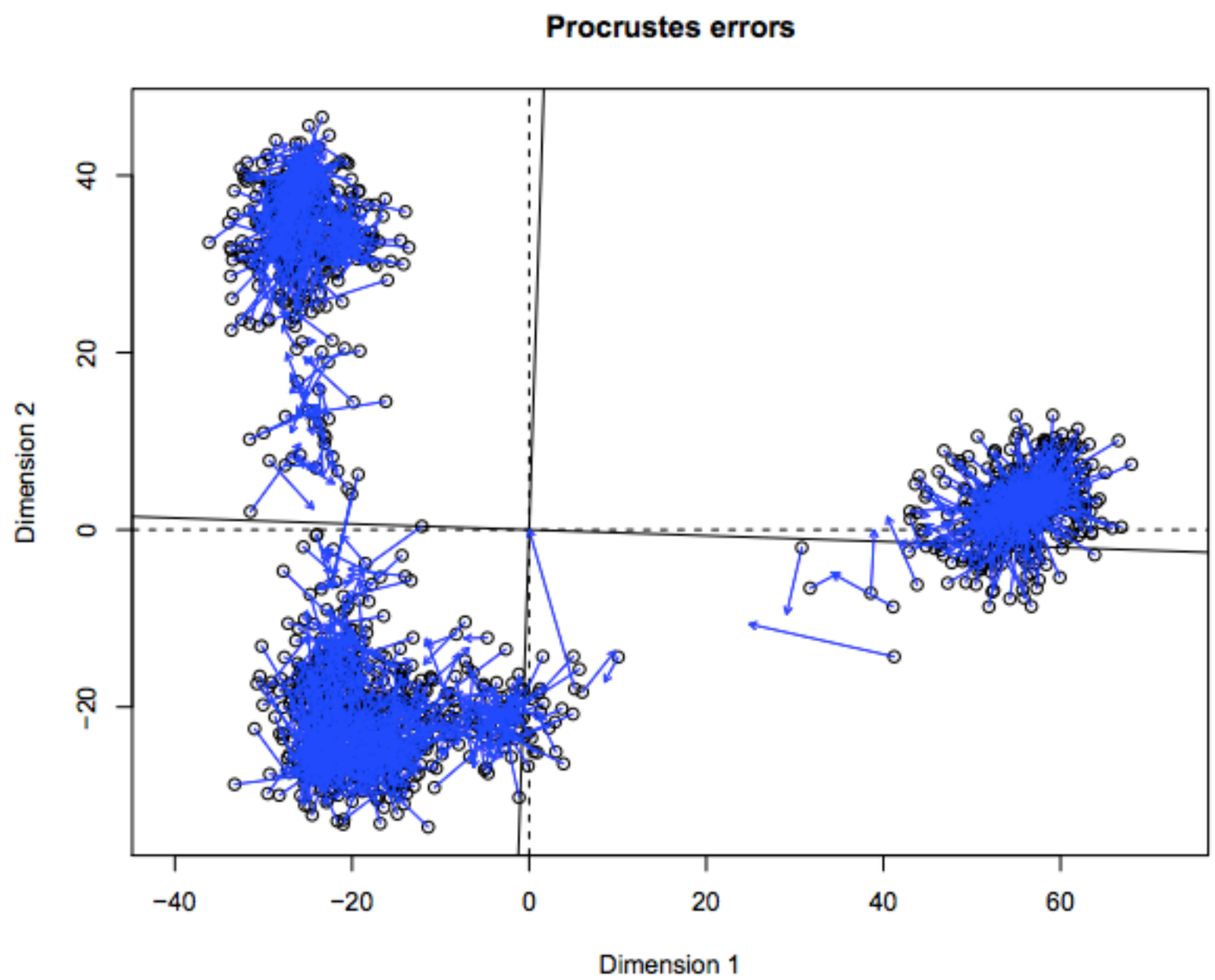
Shape-preserving Procrustes transformation



Shape-preserving Procrustes transformation

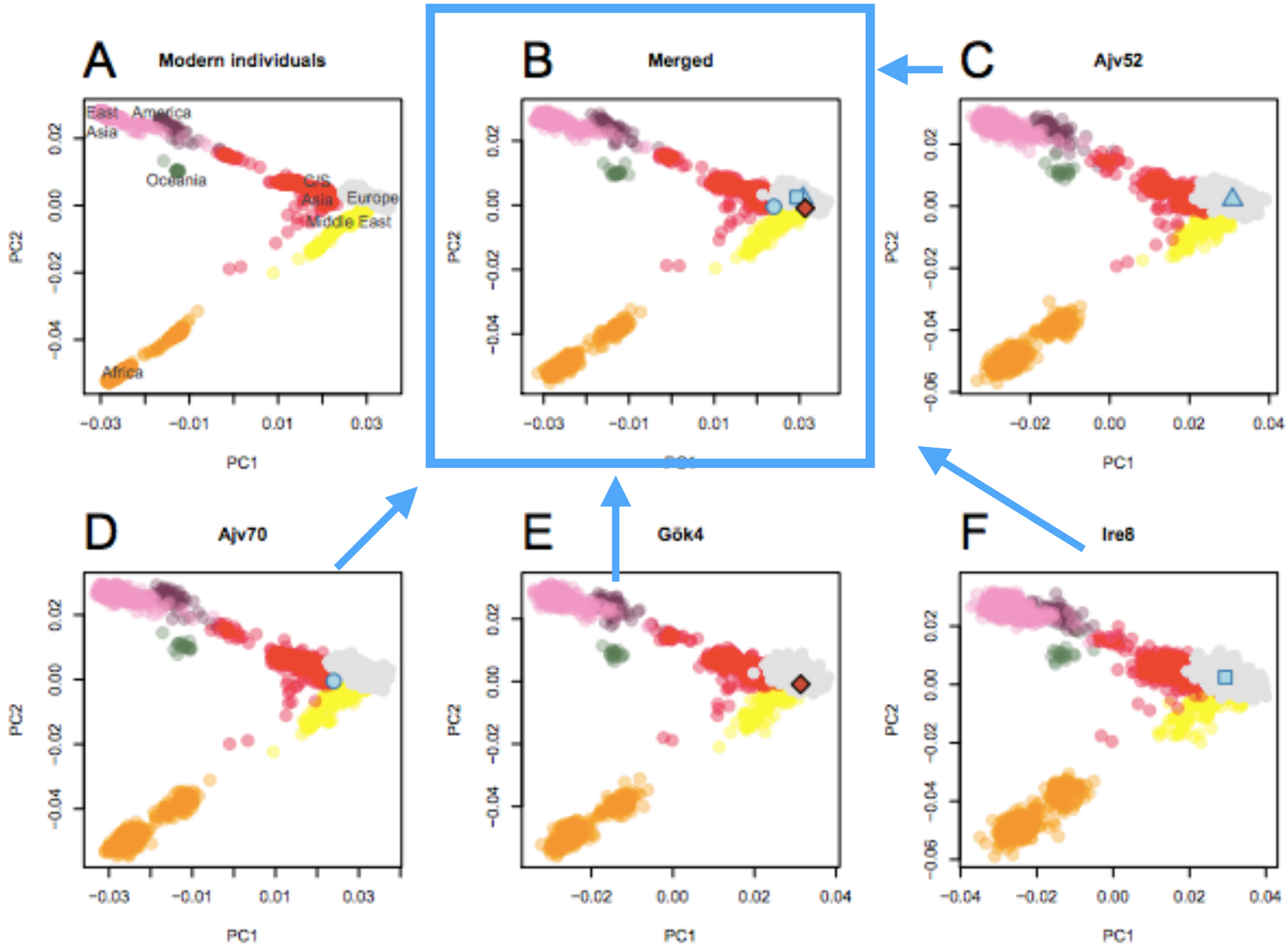


Use a Procrustes transformation using a **high-coverage reference PCA**



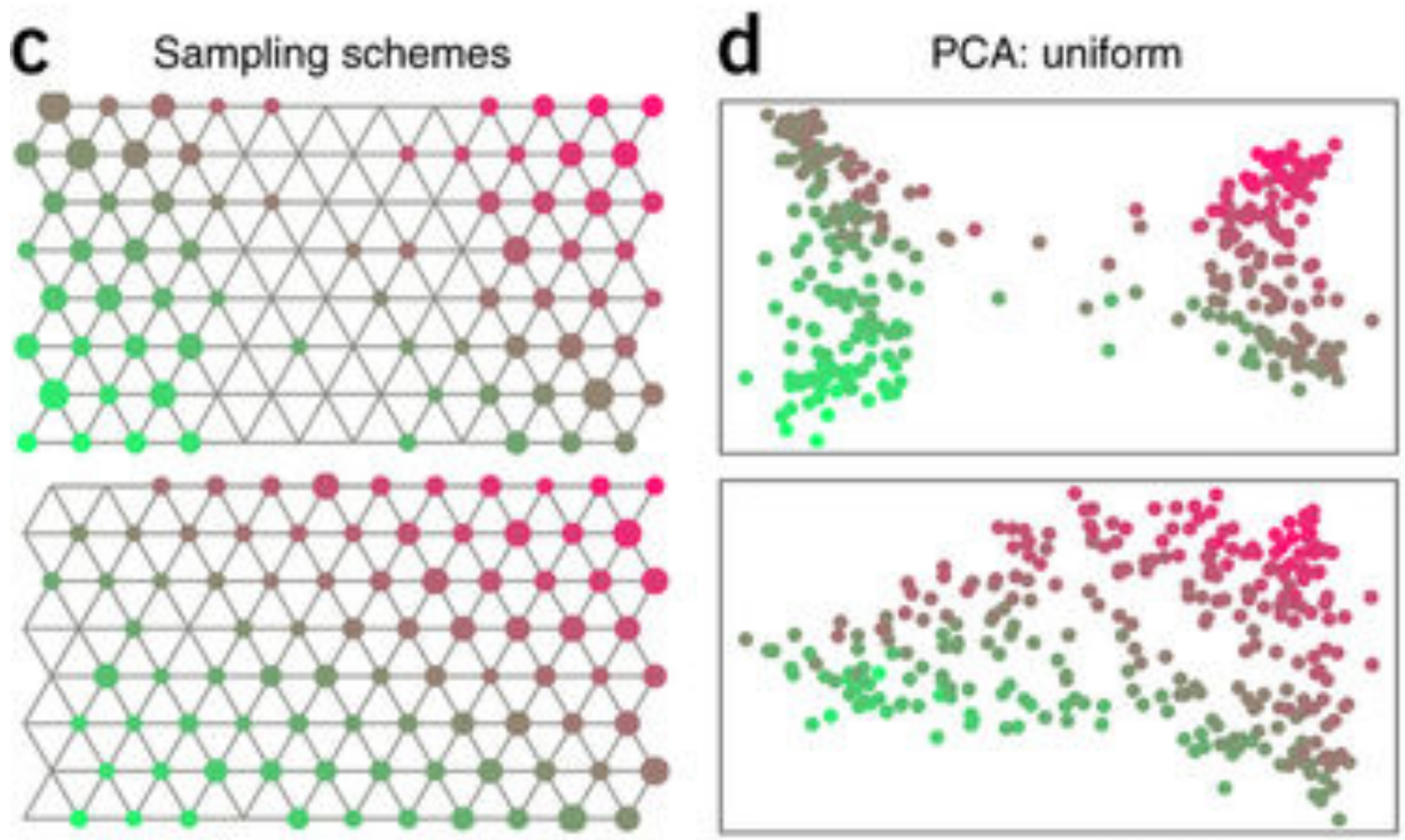
Skoglund et al. 2012

Procrustes transformation

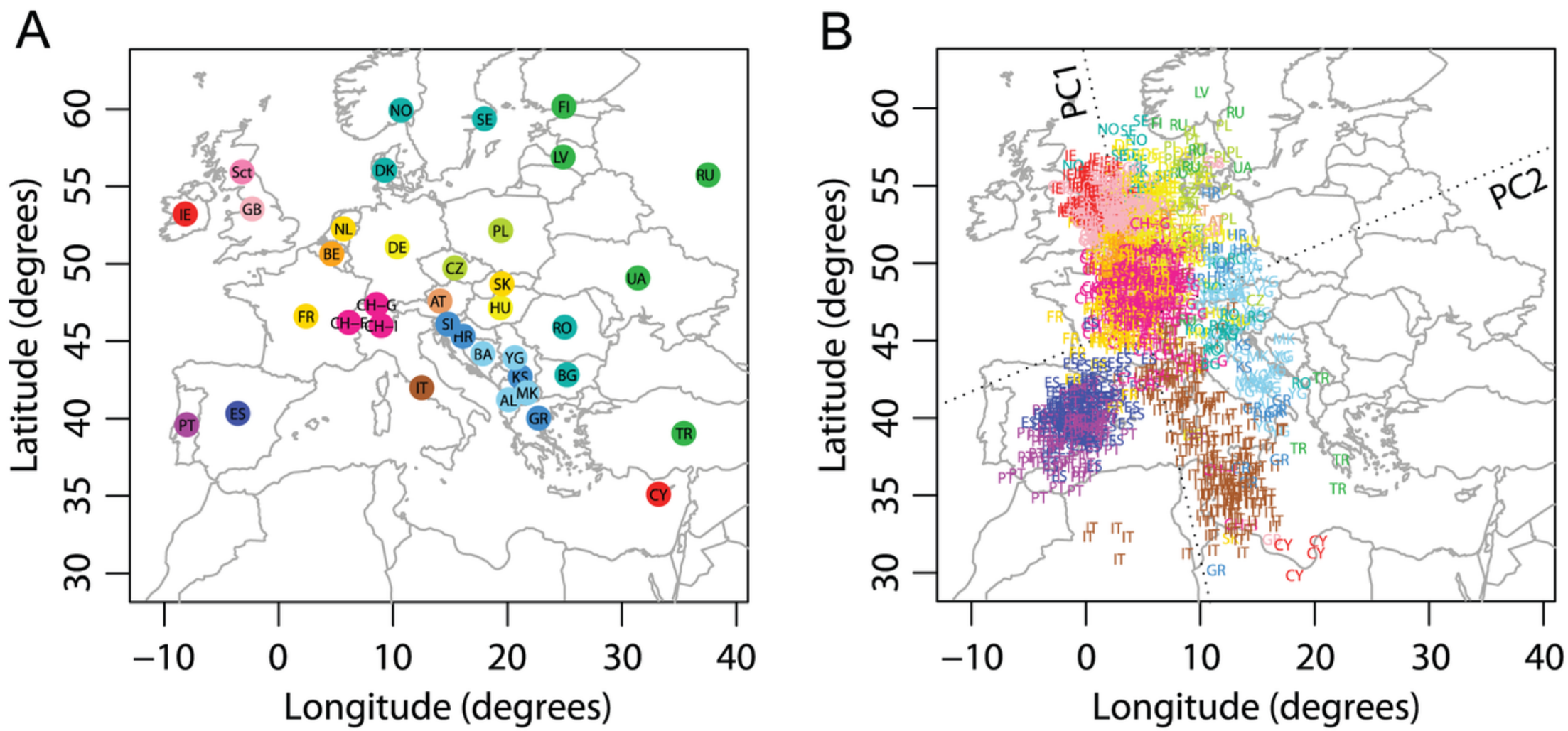


Skoglund et al. 2012

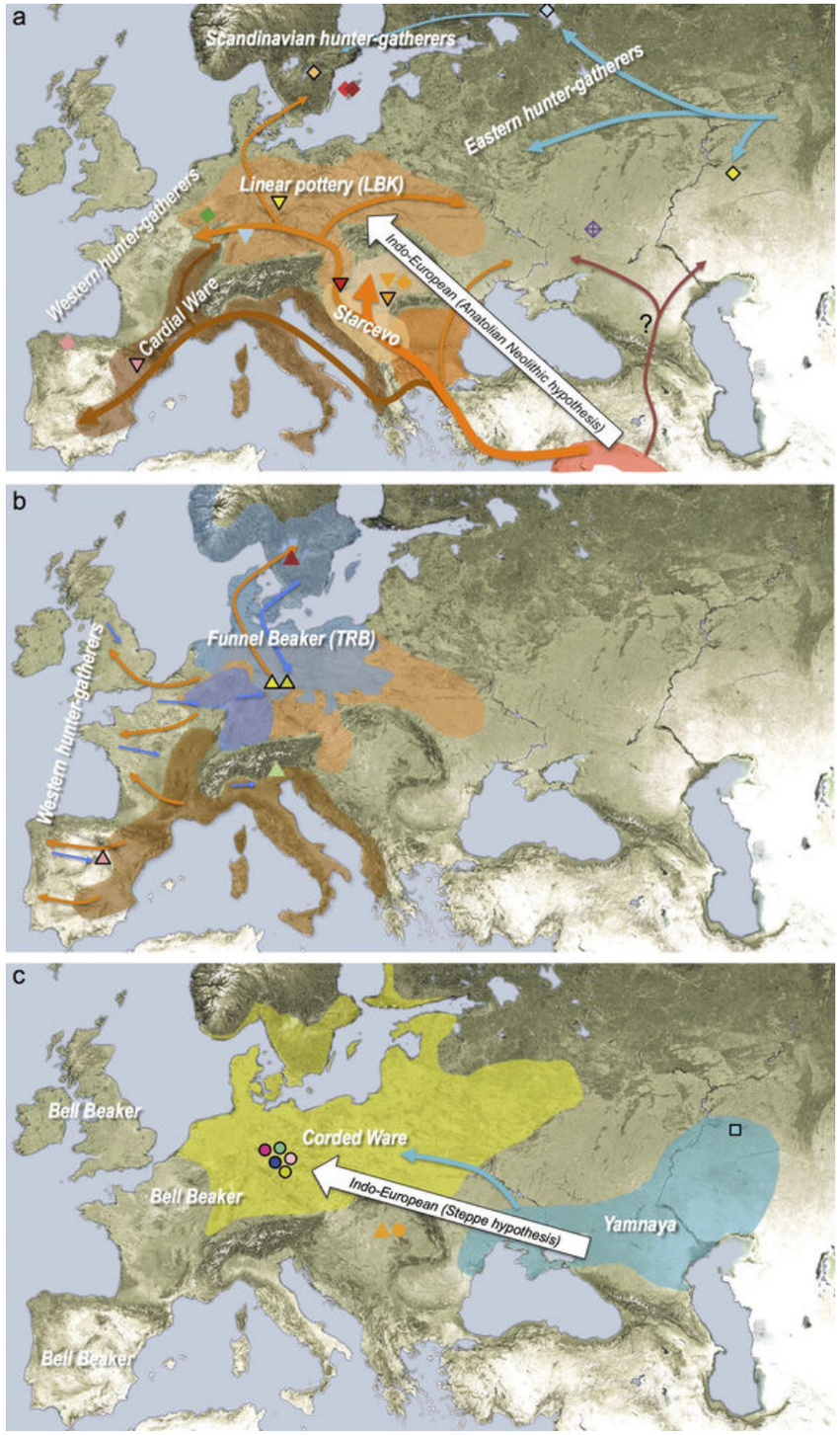
Sampling scheme can be misleading



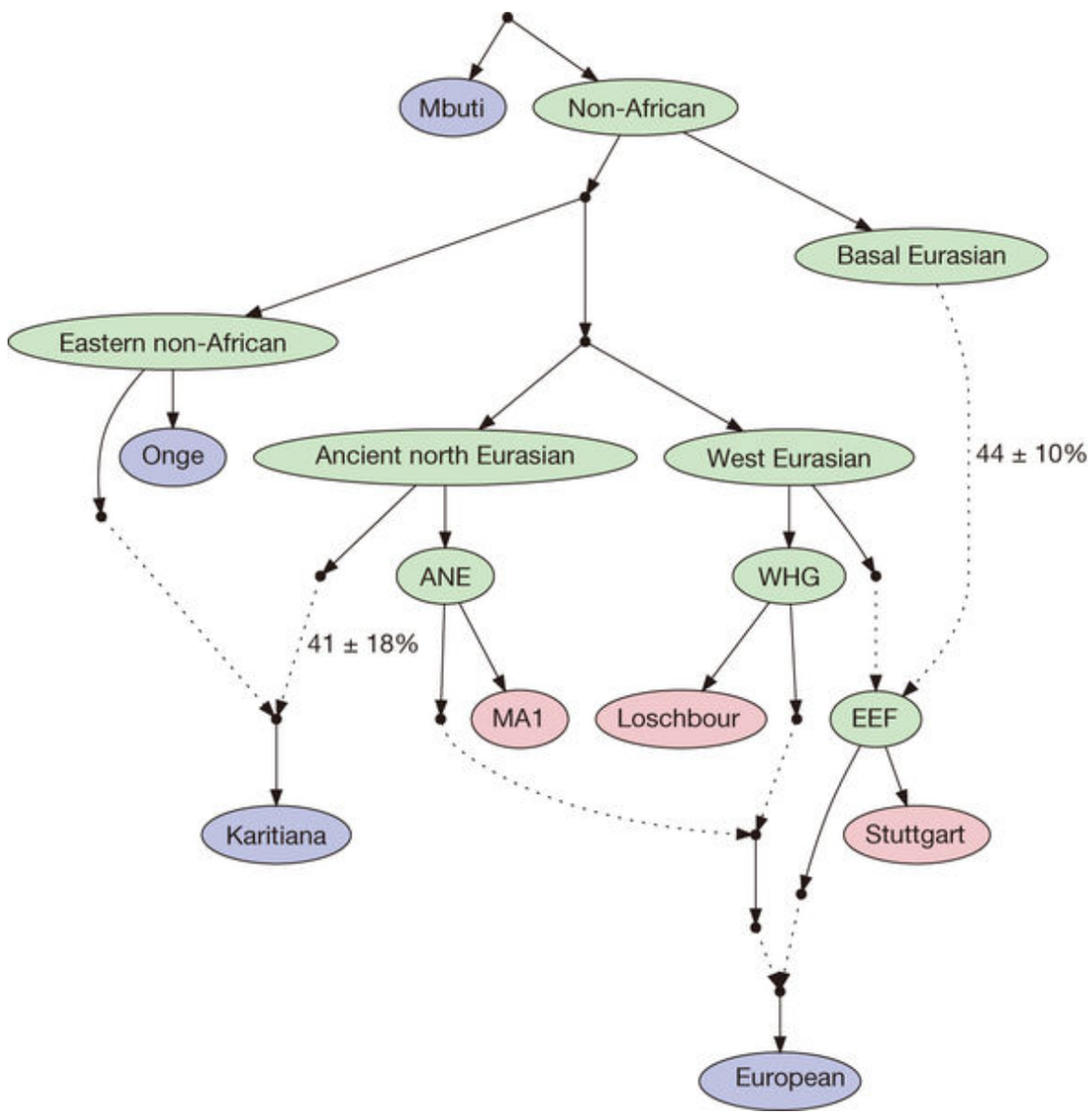
PCA can be misinterpreted!



PCA can be misinterpreted!



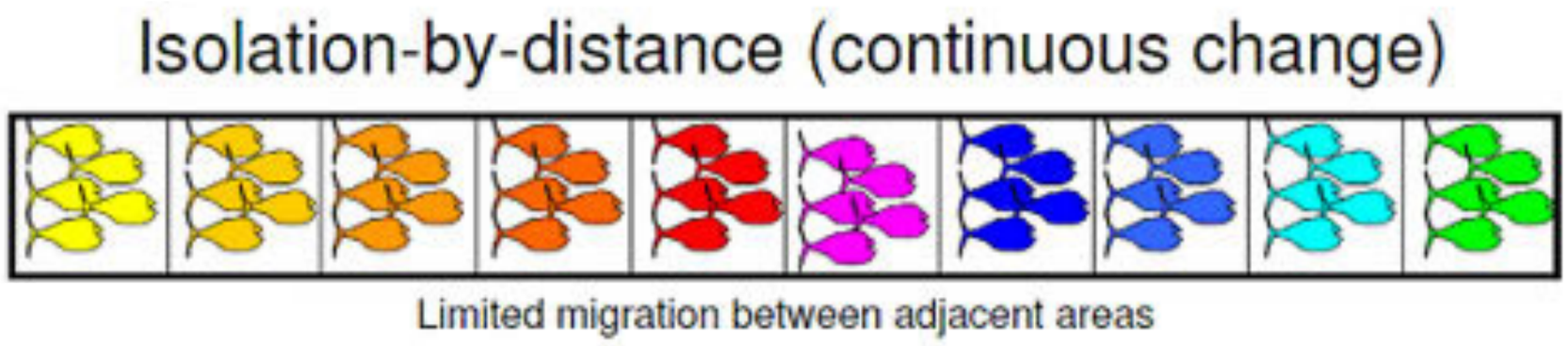
PCA can be misinterpreted!



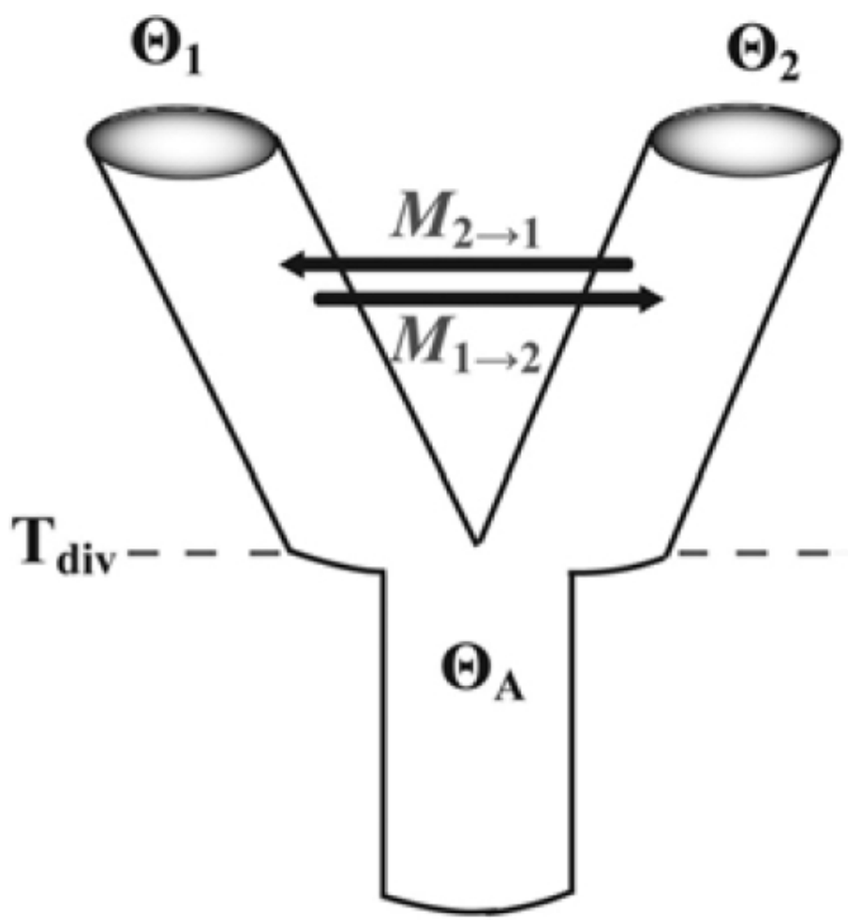
Today

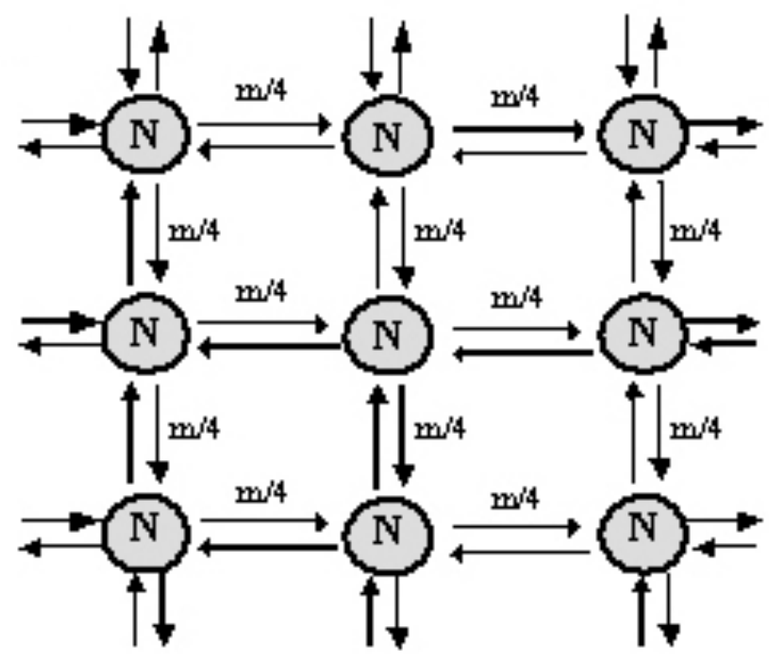
- Experimental design
- Data handling
- PCA
- **Spatial and isolation-by-distance methods**

Isolation-by-distance

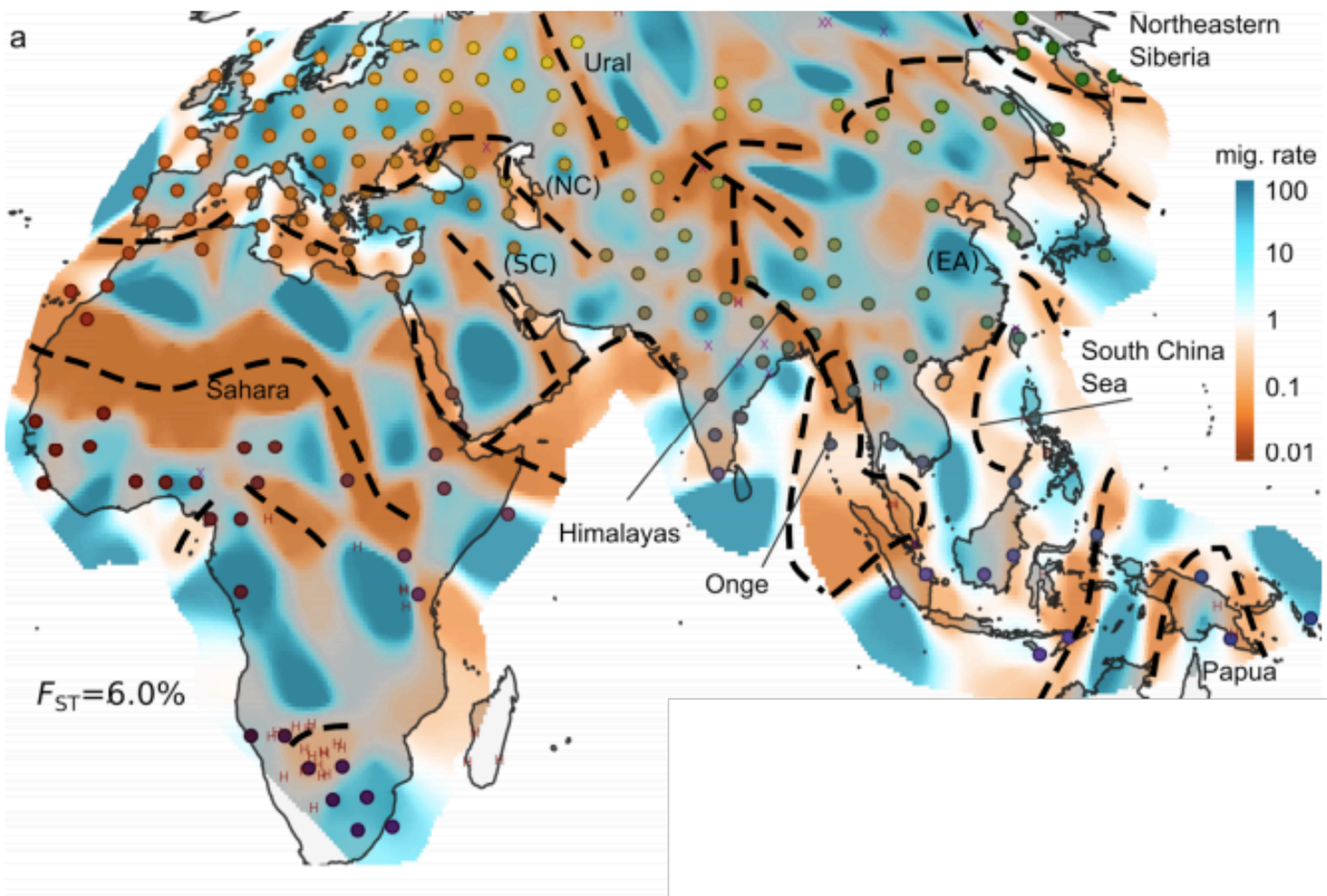


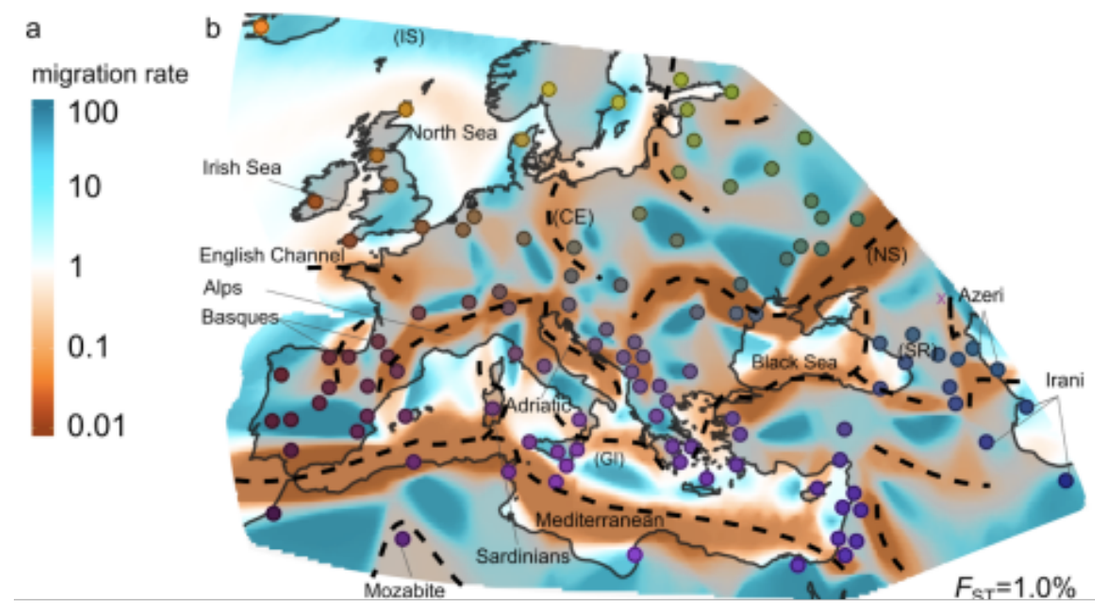
Long-range admixture vs. isolation-by-distance





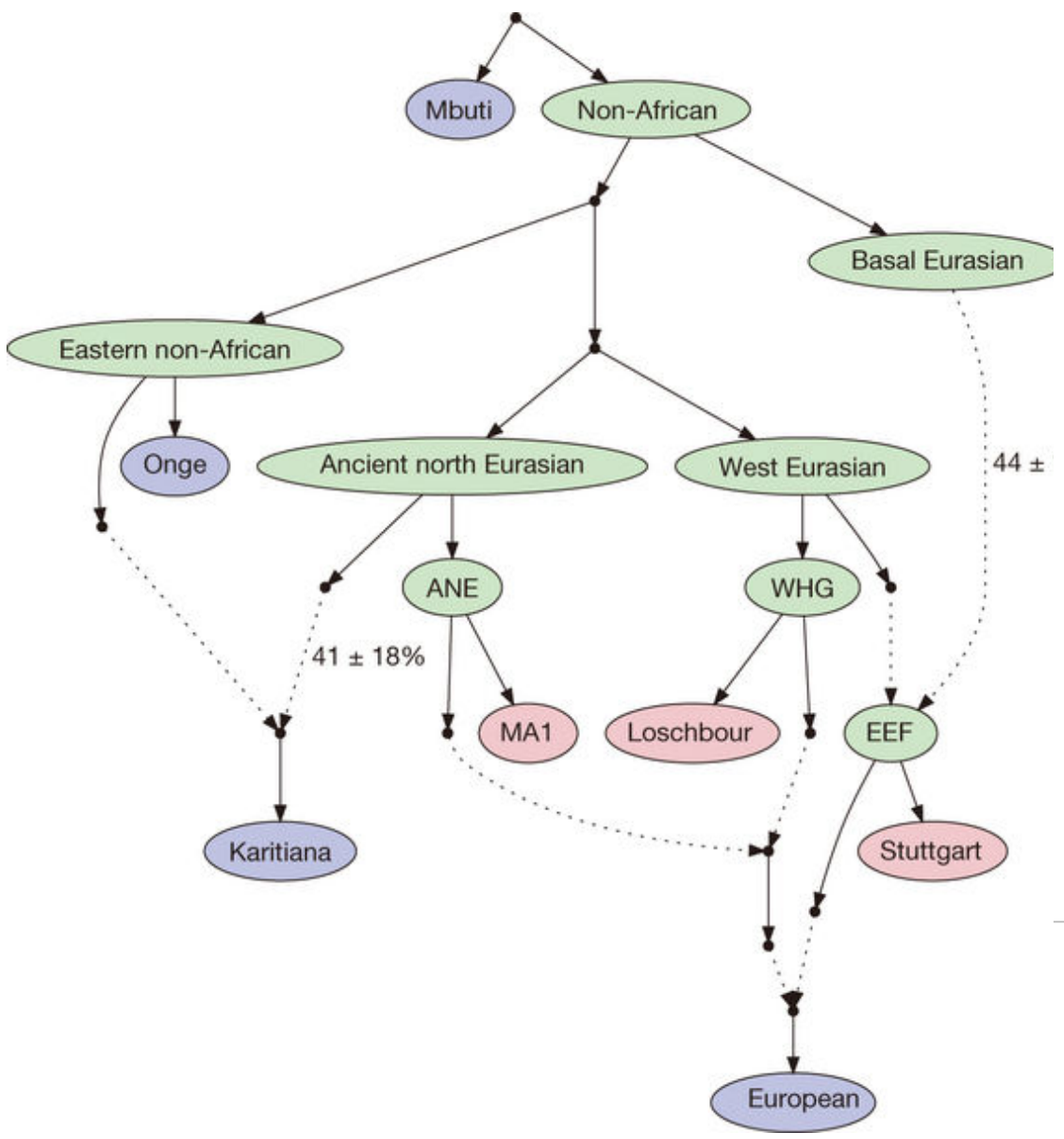
EEMS: a method to model isolation-by-distance

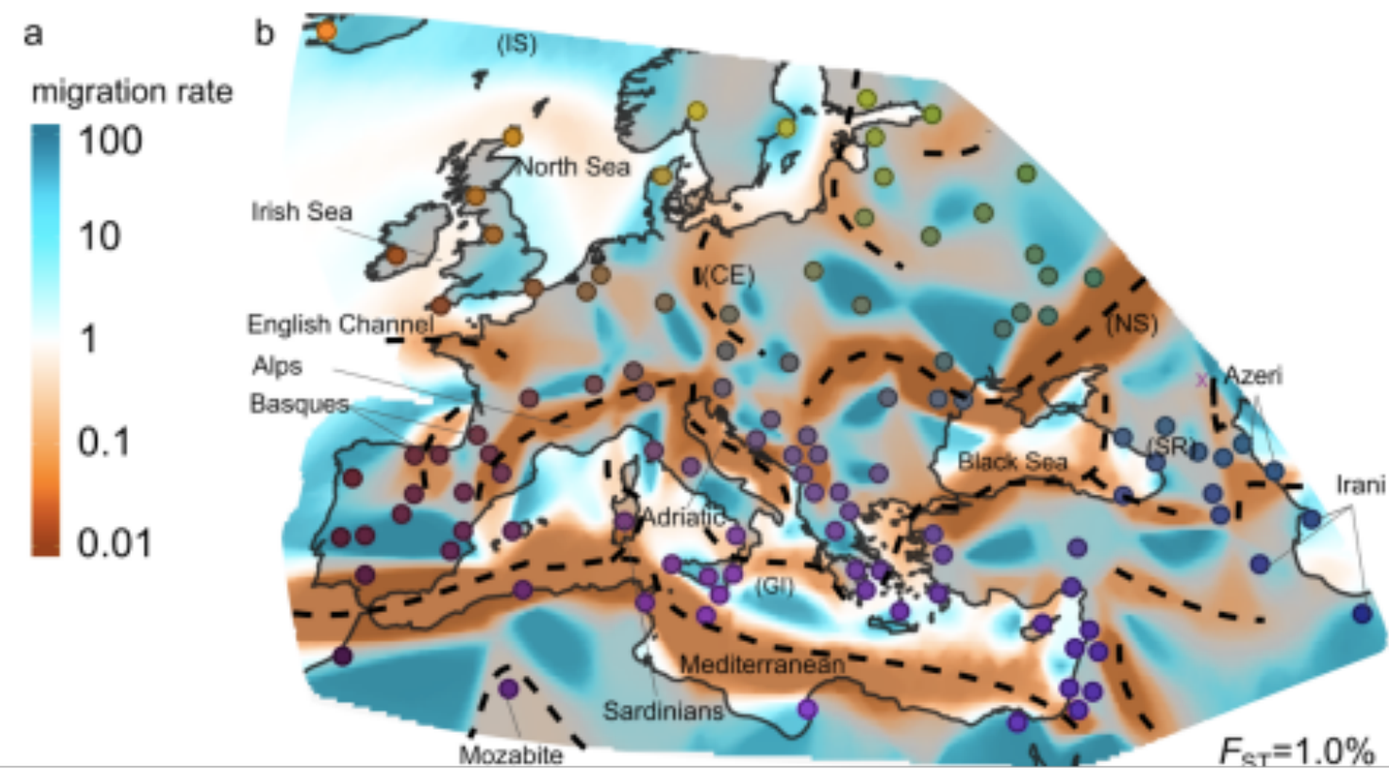




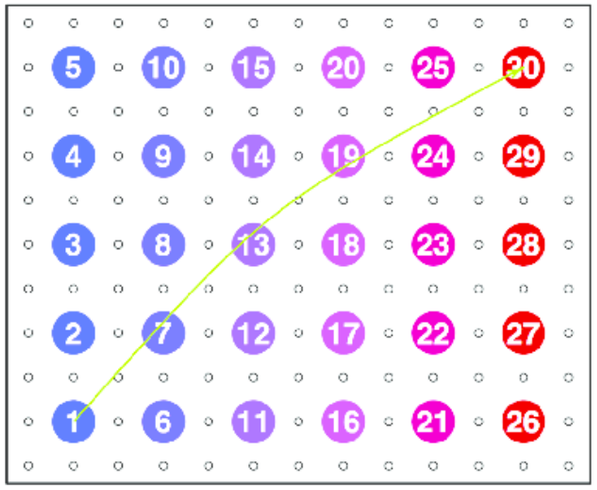
Petkova et al. (2016)
Peter et al. (2017)

Model assumptions are important

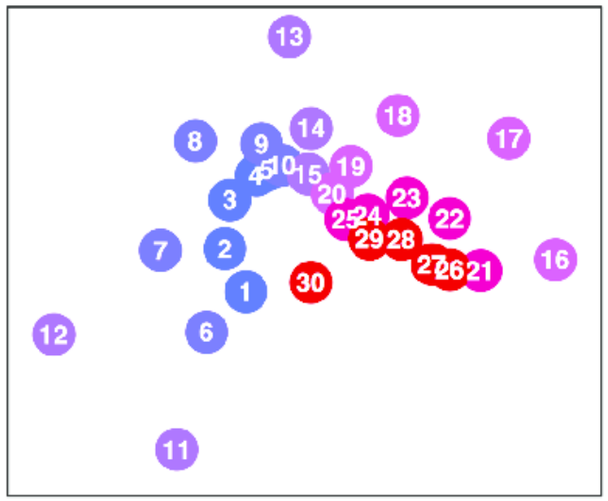




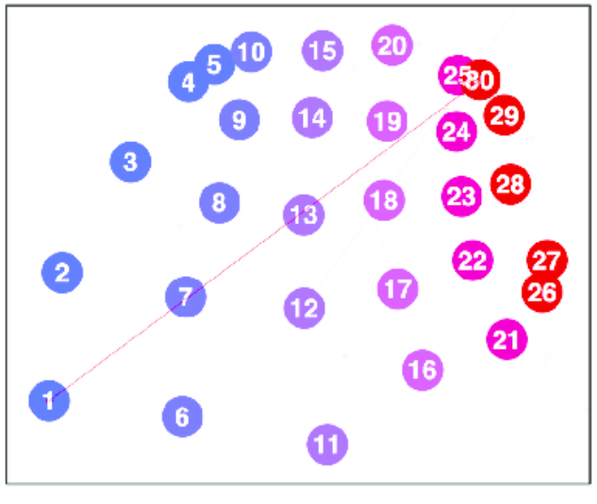
Long-range admixture + isolation-by-distance



(a) simulated lattice with admixture

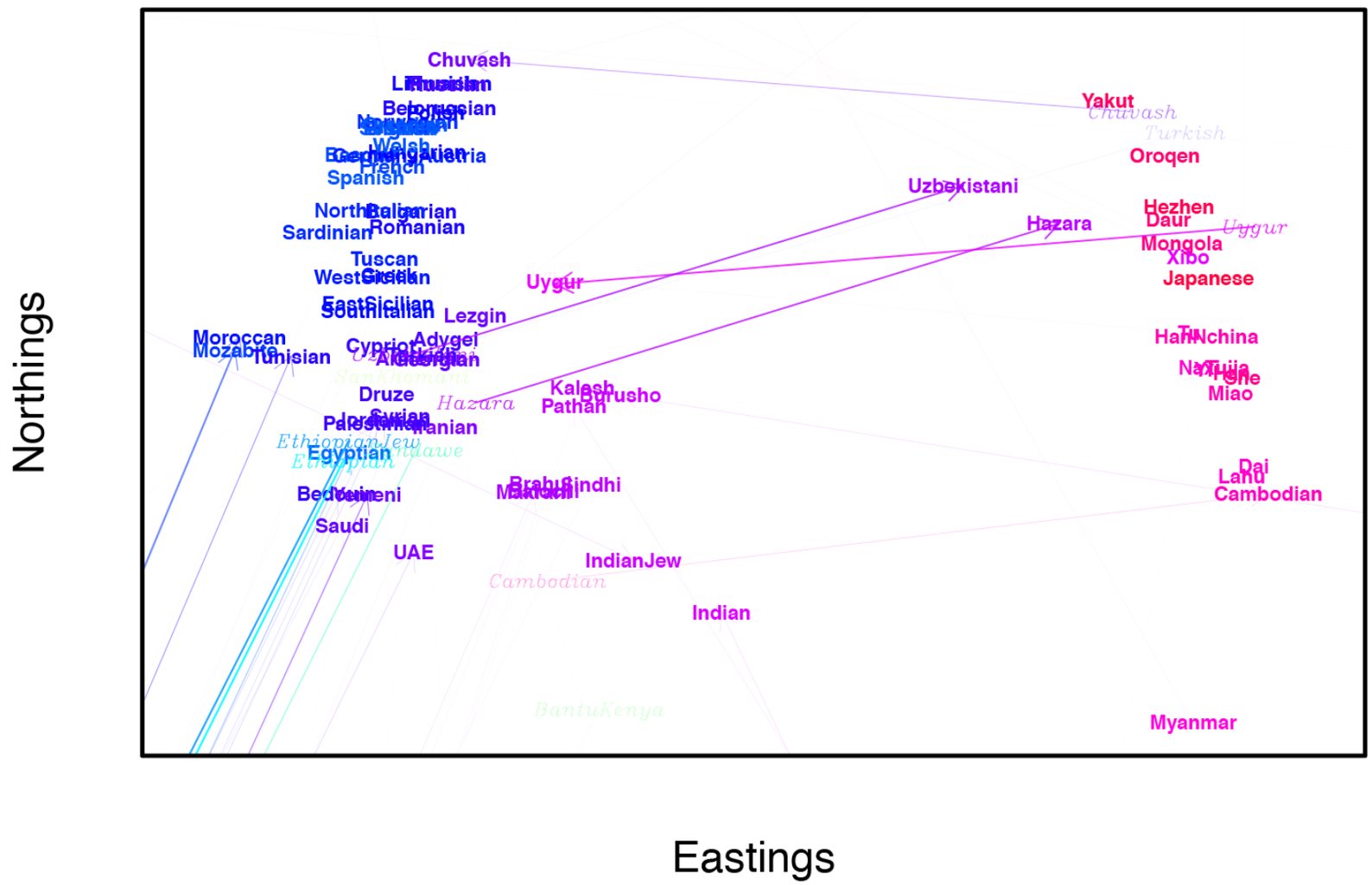


(b) geogenetic map without admixture inference



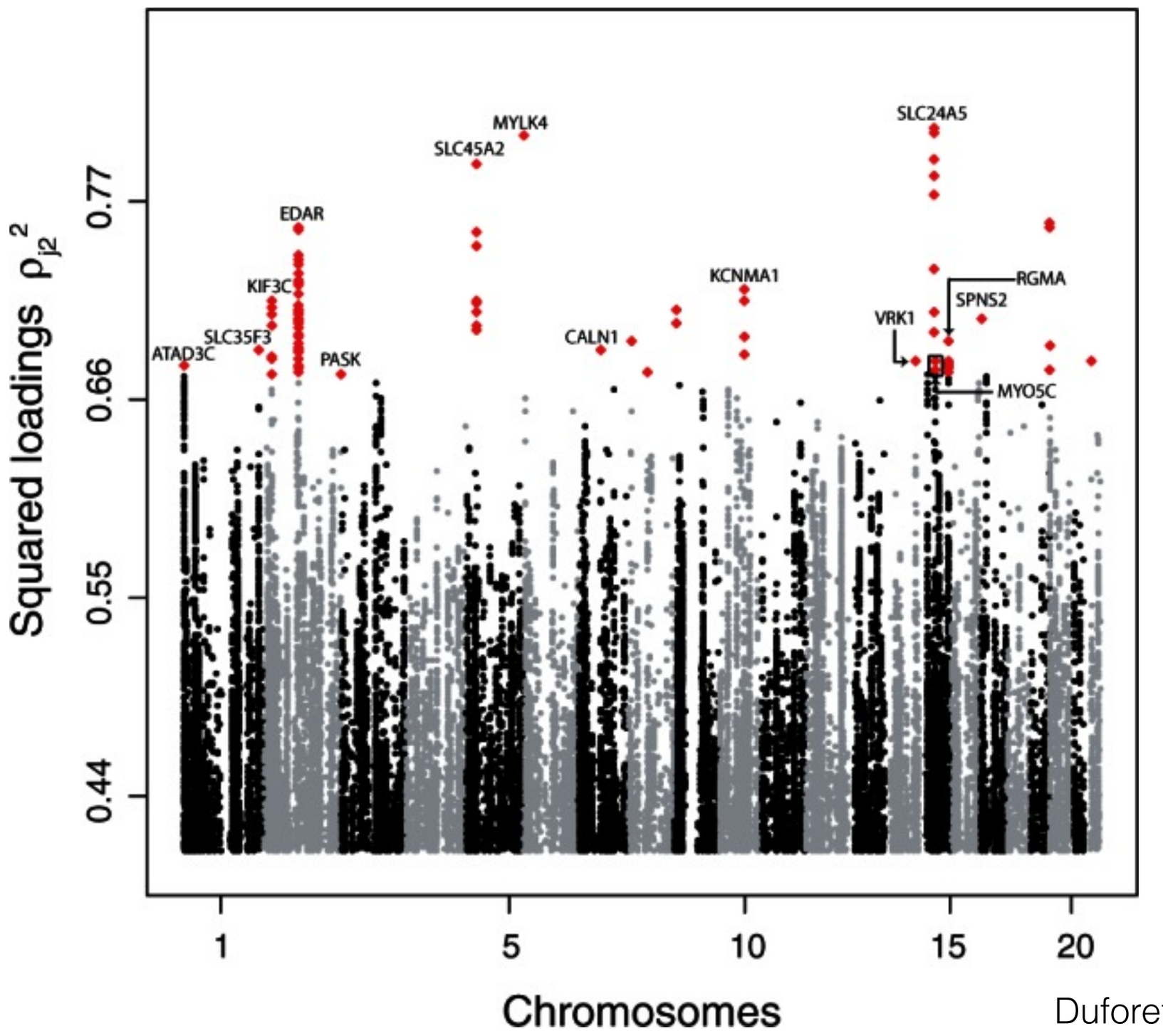
(c) geogenetic map with admixture inference

Long-range admixture + isolation-by-distance



(b) Close-up of Eurasian samples

Using PCA loadings to detect loci under selection



Duforet-Frebourg et al. (2016)