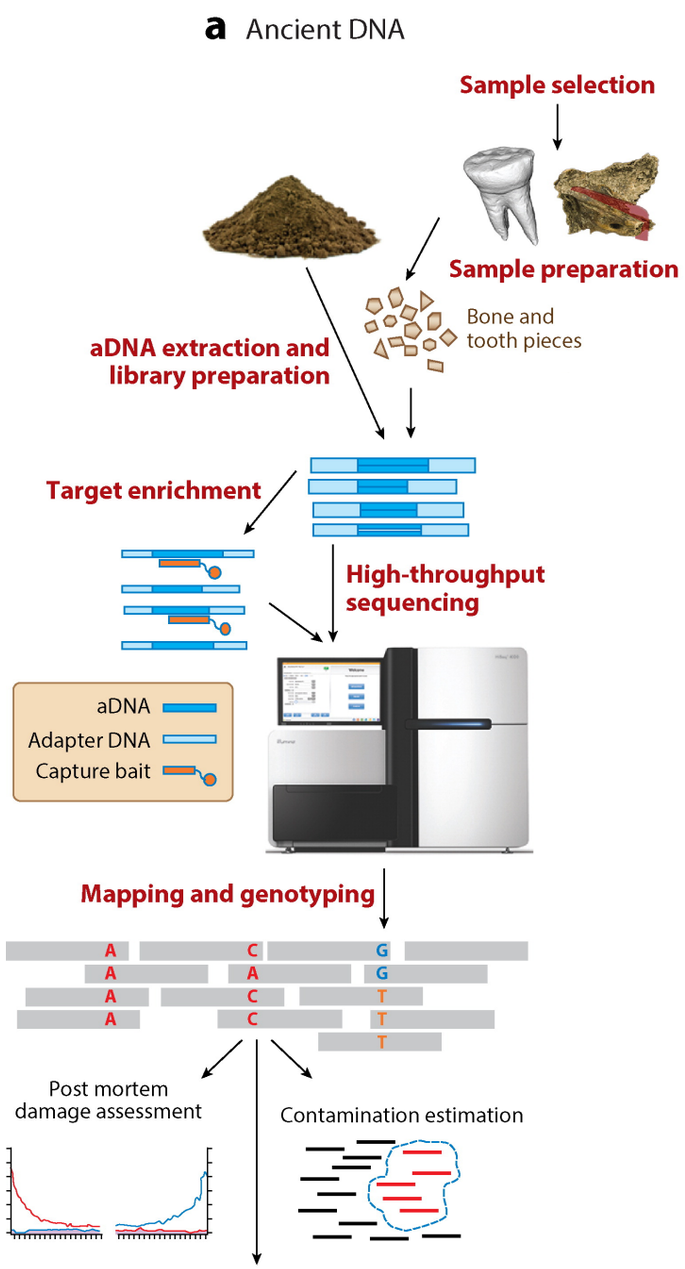
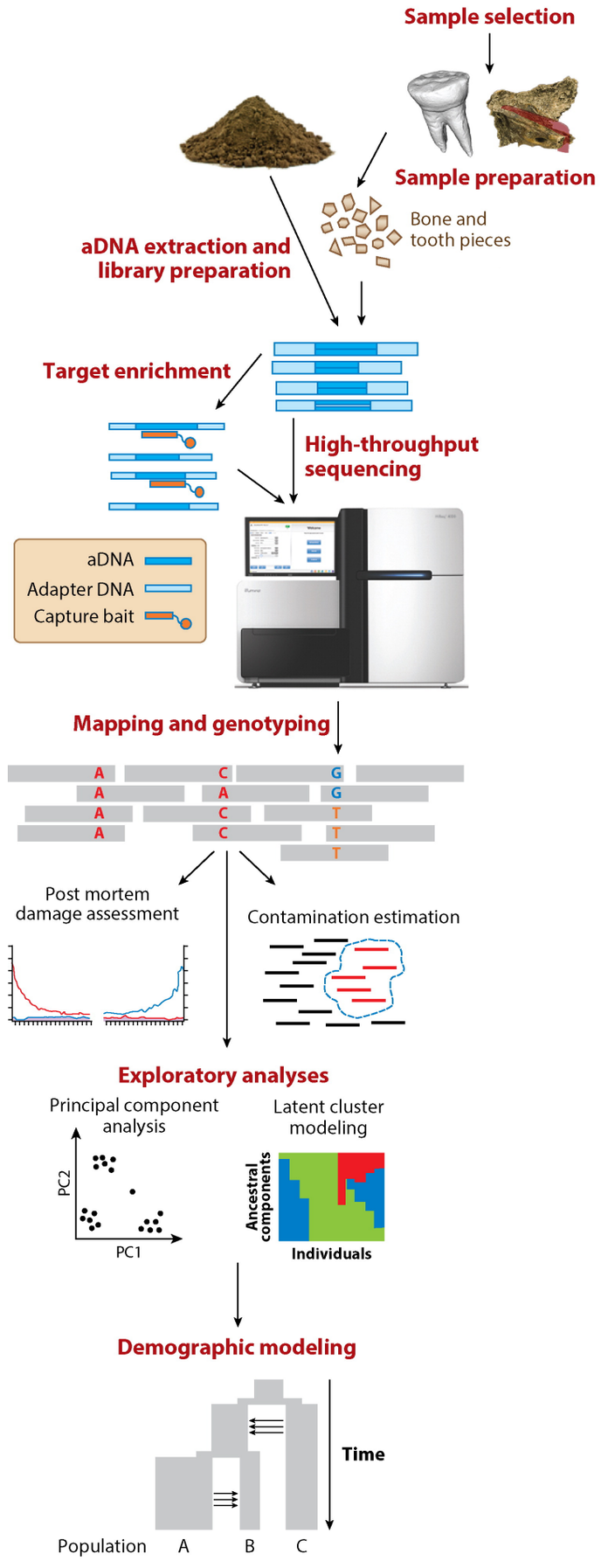
Population genetics 1: exploratory analyses

Fernando Racimo

Copenhagen, August 2019

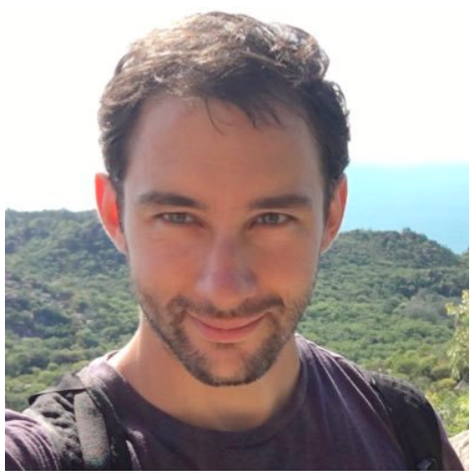


Cappellini et al. 2018



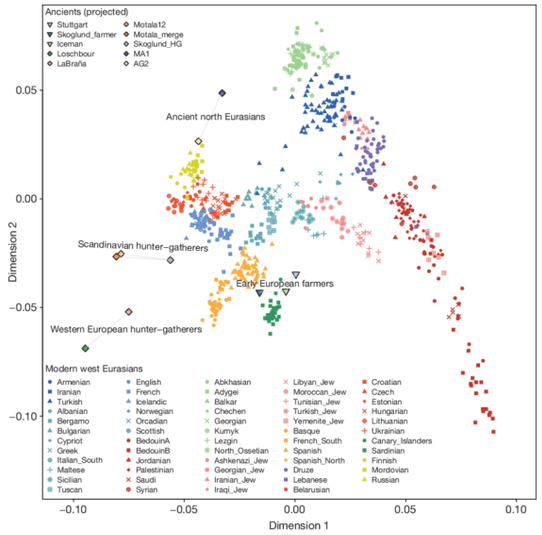
Population genetics

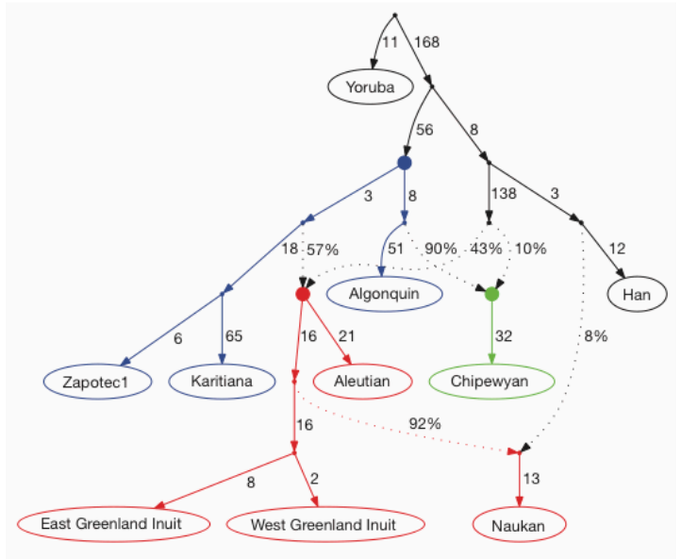
Cappellini et al. 2018

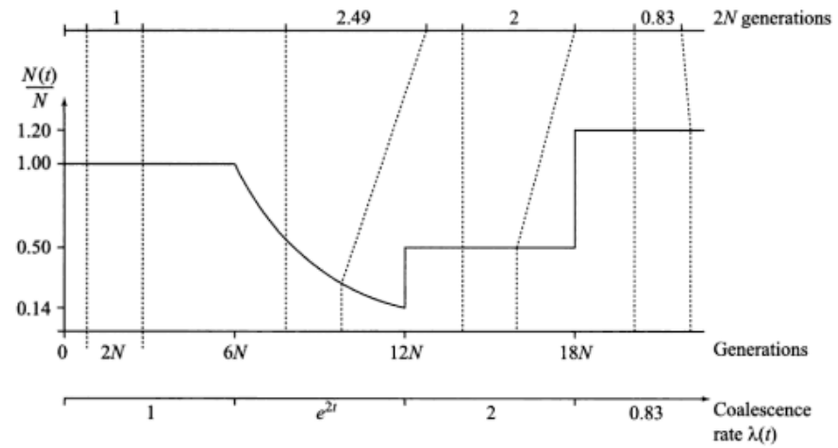


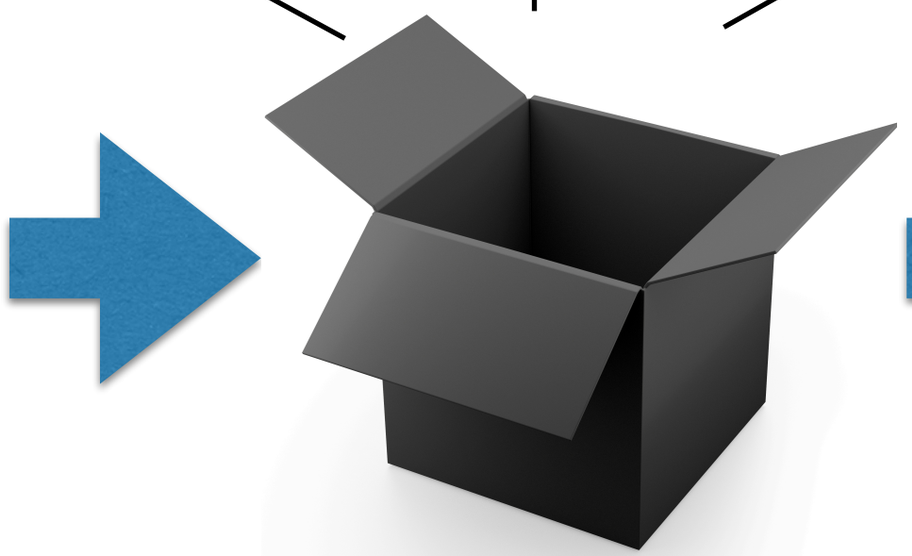

I’ll take over from here…


nature
NANOTECHNOLOGY
EARTHQUAKE RUPTURE
AVIAN INFLUENZA VIRUS
NATURE JOBS
NEANDERTHAL GENOMICS


Ancients (projected)
Stuttgart
Skoglund_farmer
Iceman
Loschbour
LaBraña
Motala12
Motala_merge
Skoglund_HG
MA1
AG2
Ancient north Eurasians
Scandinavian hunter−gatherers
Early European farmers
Western European hunter−gatherers
Modern west Eurasians
Armenian
Iranian
Turkish
Albanian
Bergamo
Bulgarian
Cypriot
Greek
Italian_South
Maltese
Sicilian
Tuscan
English
French
Icelandic
Norwegian
Orcadian
Scottish
BedouinA
BedouinB
Jordanian
Palestinian
Saudi
Syrian
Abkhasian
Adygei
Balkar
Chechen
Georgian
Kumyk
Lezgin
North_Ossetian
Ashkenazi_Jew
Georgian_Jew
Iranian_Jew
Iraqi_Jew
Libyan_Jew
Moroccan_Jew
Tunisian_Jew
Turkish_Jew
Yemenite_Jew
Basque
French_South
Spanish
Spanish_North
Druze
Lebanese
Belarusian
Croatian
Czech
Estonian
Hungarian
Lithuanian
Ukrainian
Canary_Islanders
Sardinian
Finnish
Mordovian
Russian
Dimension 1
Dimension 2
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
0.10

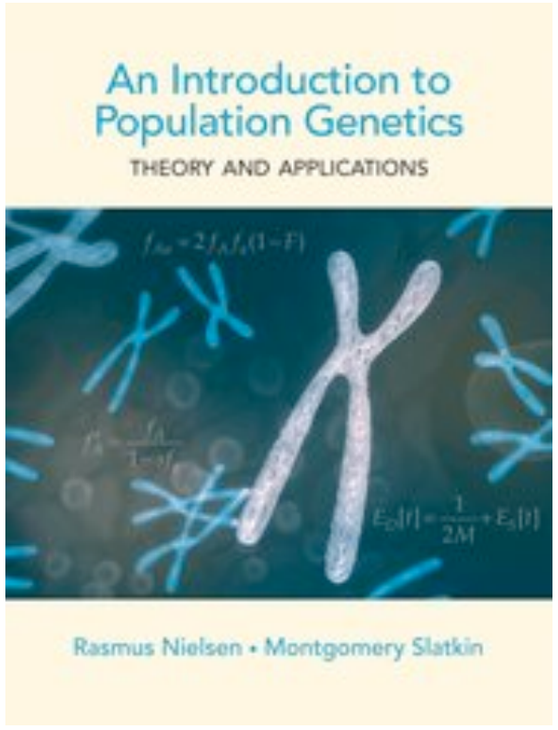

Yoruba
Zapotec1
Karitiana
Algonquin
Aleutian
Chipewyan
Han
East Greenland Inuit
West Greenland Inuit
Naukan
11
168
56
8
3
8
138
3
18
57%
51
90%
43%
10%
12
6
65
16
21
32
8%
16
92%
8
2
13

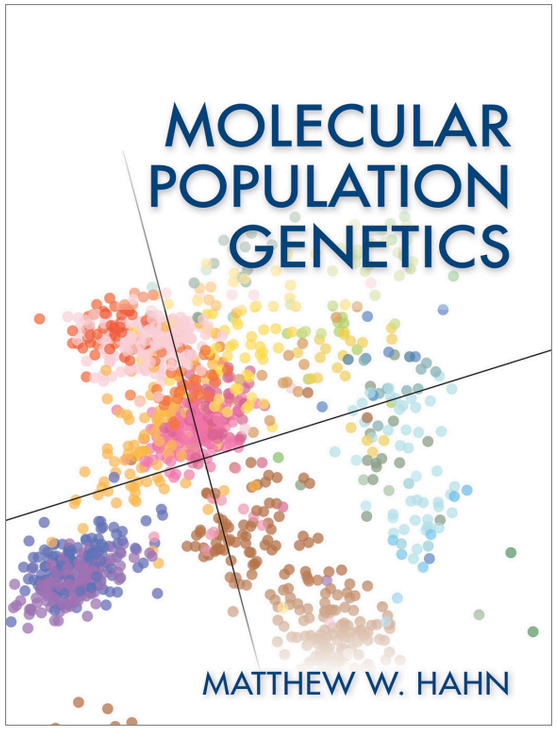

1
2.49
2
0.83
2N generations
N(t)/N
1.20
1.00
0.50
0.14
0
2N
6N
12N
18N
Generations
1
e^{2t}
2
0.83
Coalescence rate λ(t)

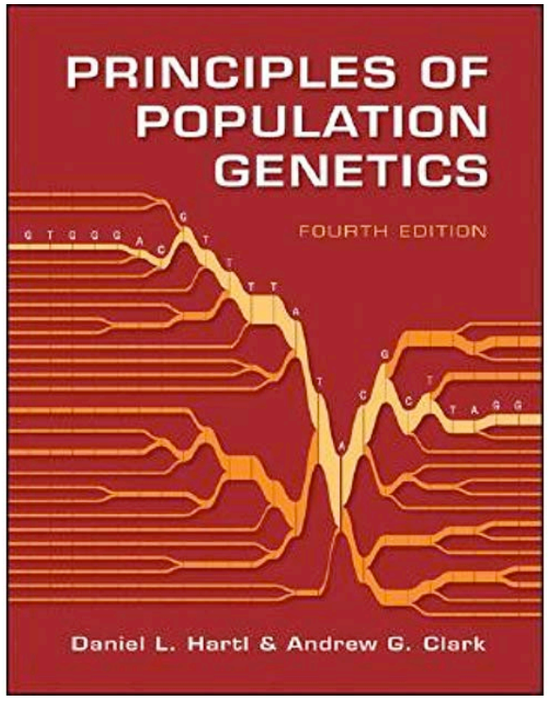


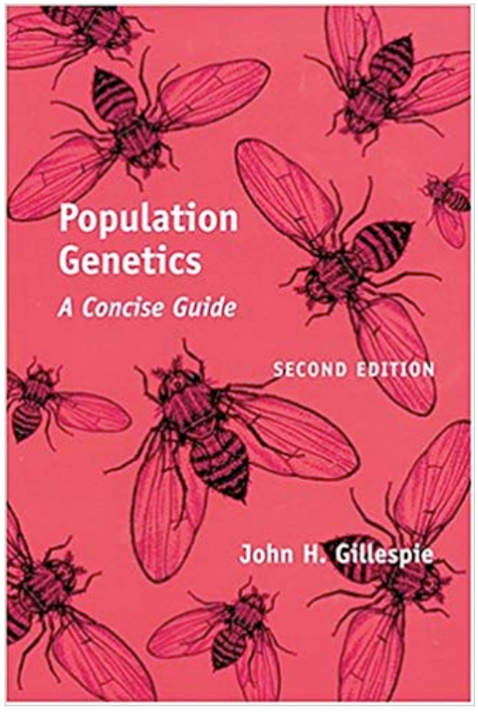

nature
NEANDERTHAL GENOMICS
A million DNA base pairs sequenced

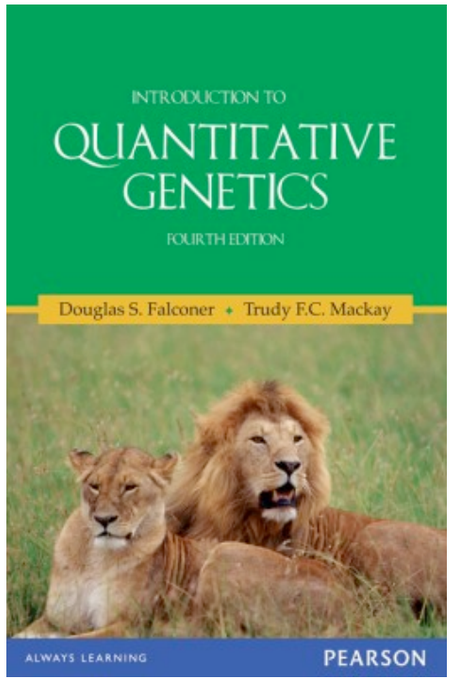
Introductory population and quantitative genetics books



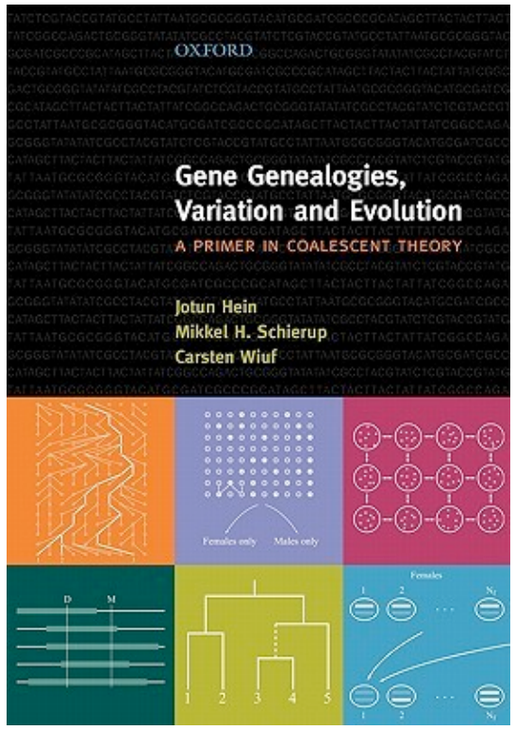


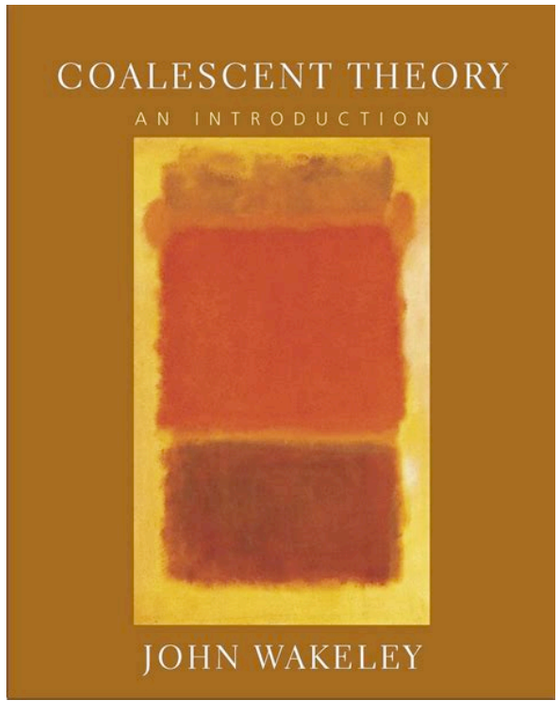


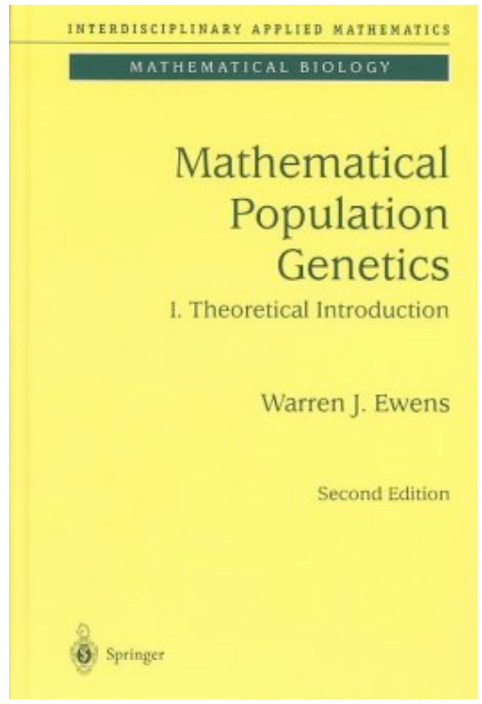


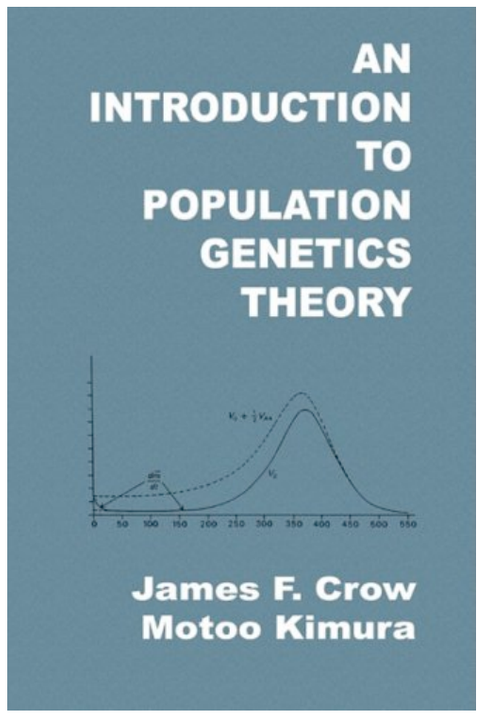


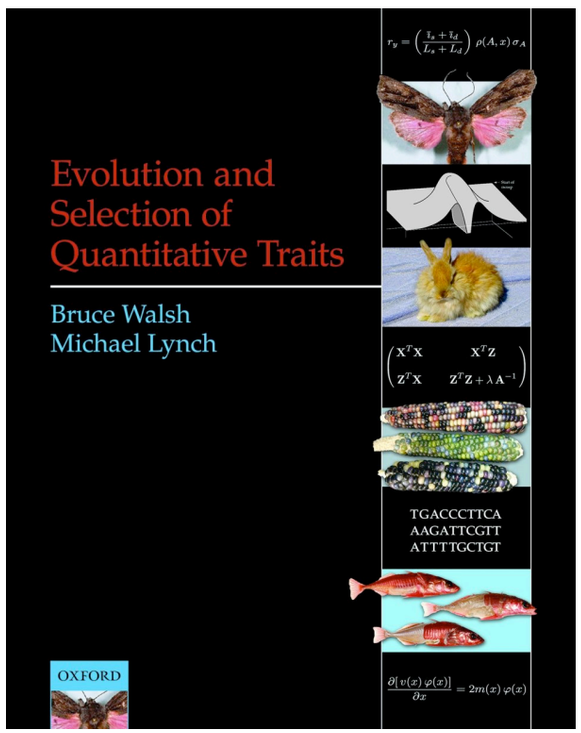
Advanced population and quantitative genetics books











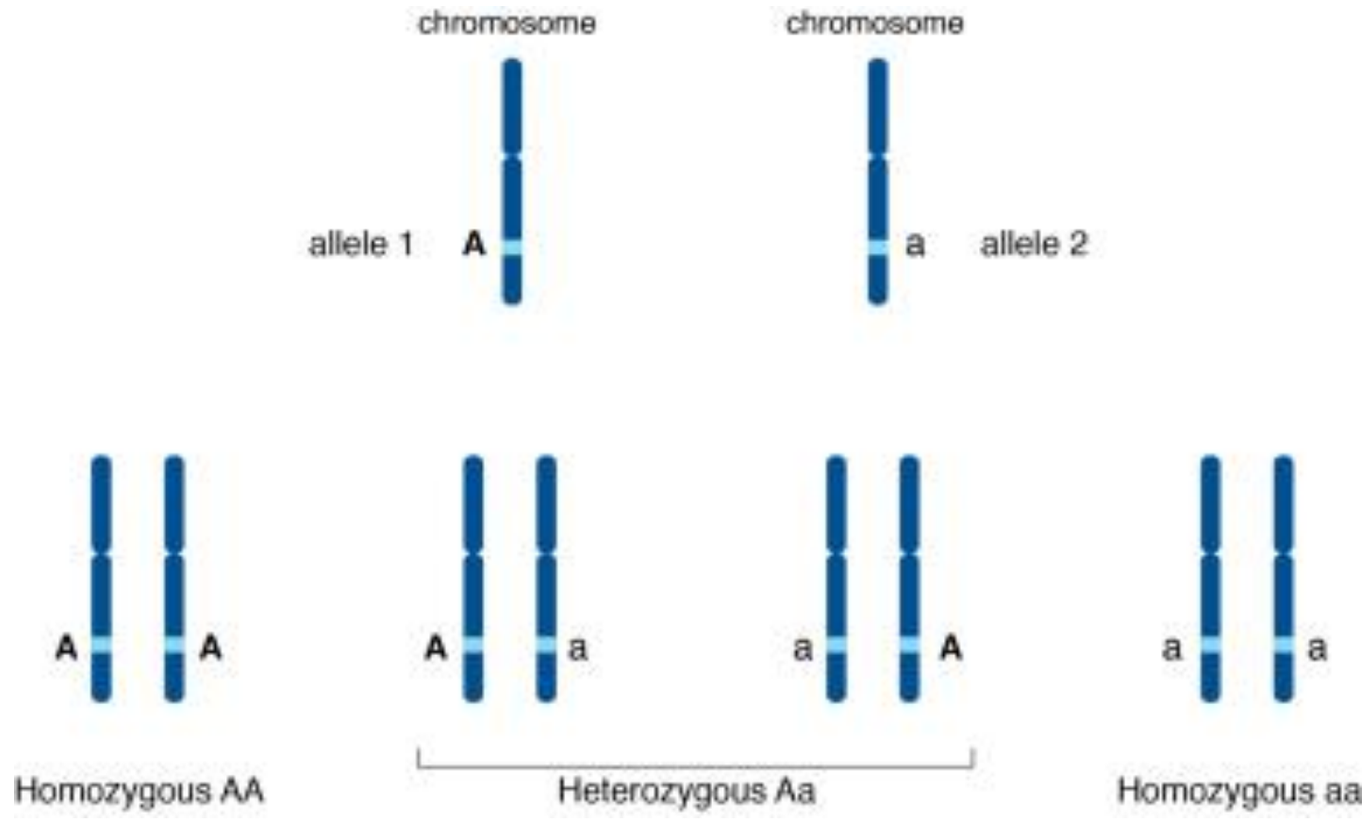
Today

- **Introduction to population genetics**
- Exploratory vs. hypothesis-driven analyses
- PCA
- Useful datasets for human paleogenomics
- Latent mixed-membership models (“Structure”)

Terminology

- **Allele:** one of two or more alternative forms of a genetic locus that reside at the same place on a chromosome.

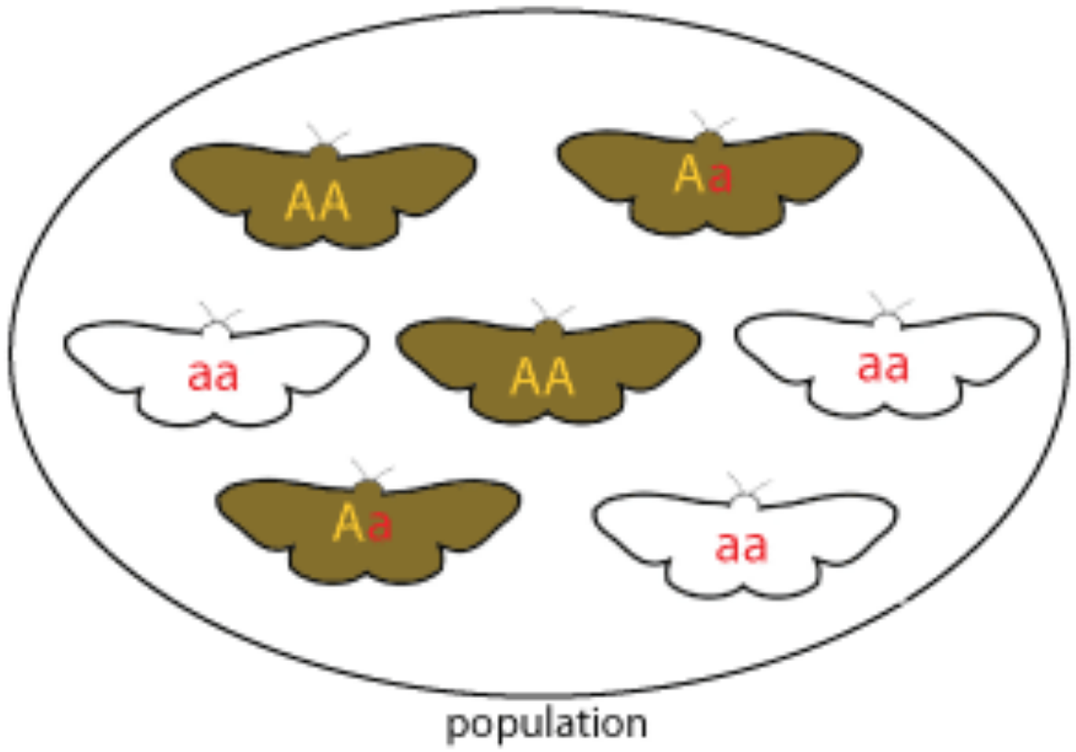
- **Genotype:** the set of alleles present at a genetic locus in an organism (two alleles if the organism is diploid).



Terminology

- **Allele frequency:** relative frequency of an allele in a population, expressed as the fraction of all chromosomes that carry that allele.

- **Genotype frequency:** relative frequency of a genotype in a population, expressed as the fraction of all individuals that carry that genotype.

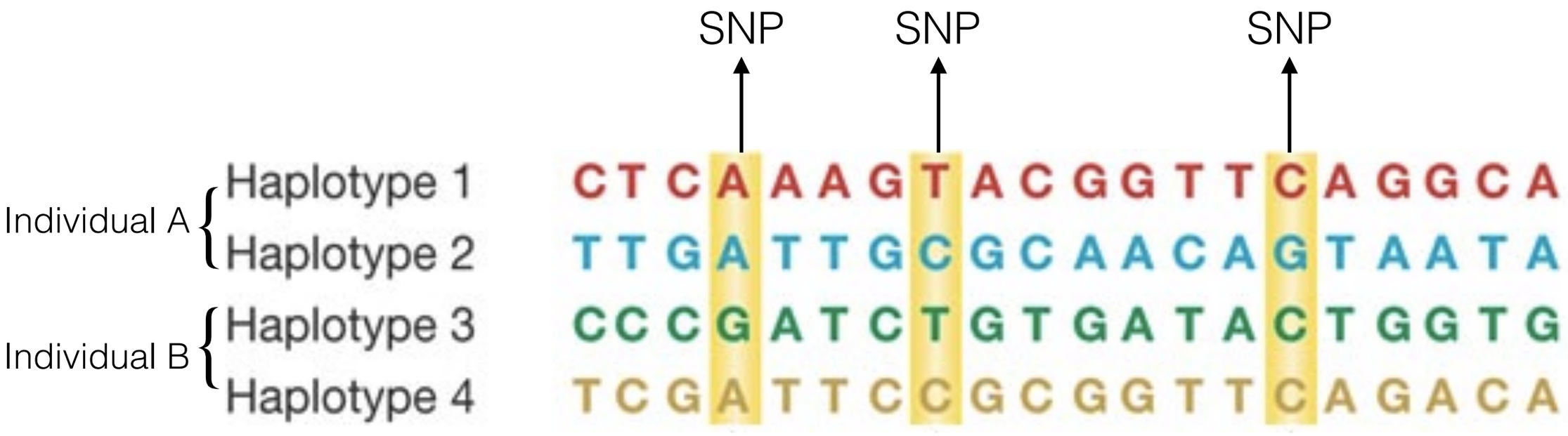


Terminology

- **Polymorphism:** a site with two or more alleles segregating in a population

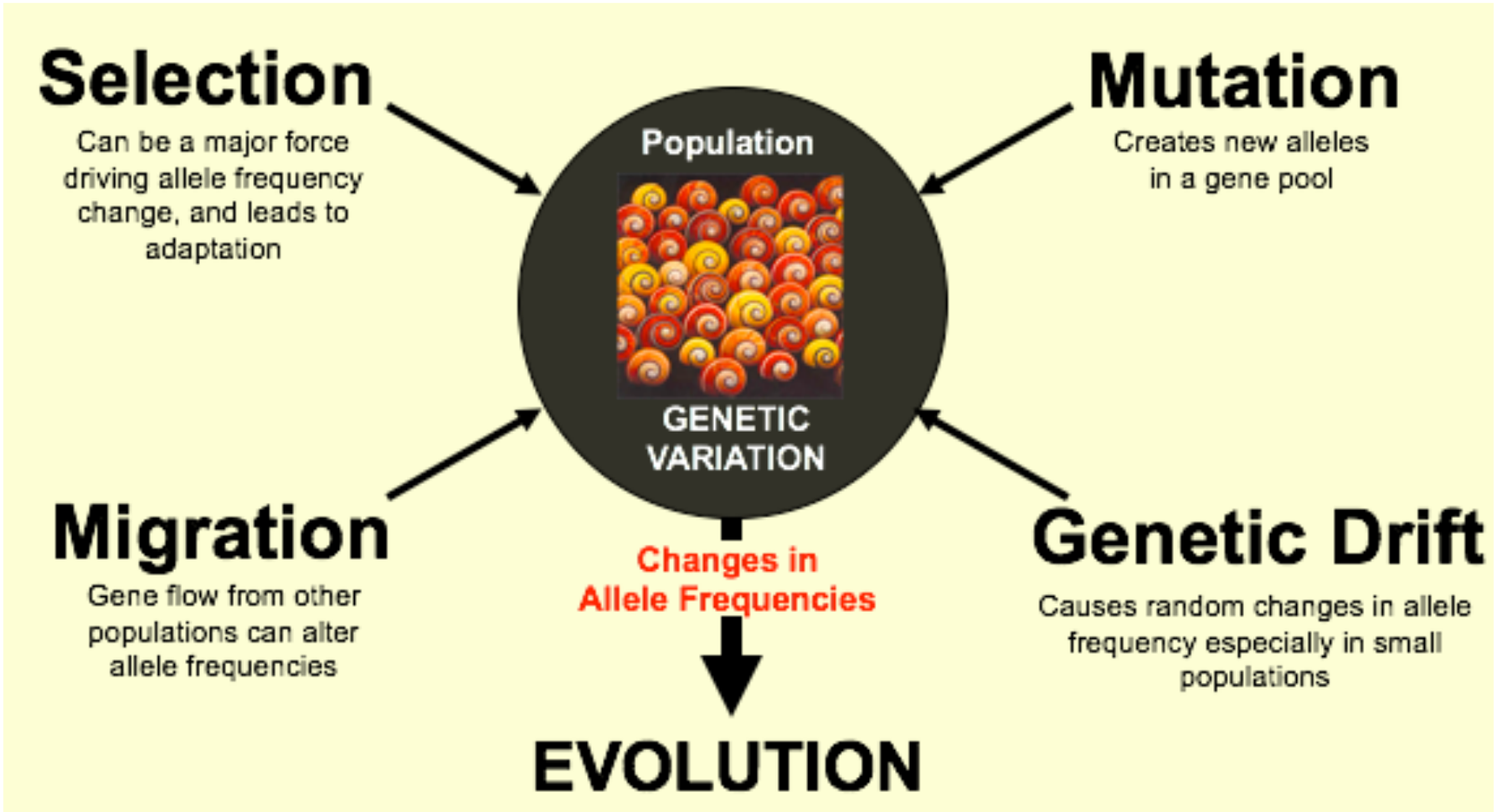
- **SNP:** single nucleotide polymorphism - a polymorphism in which a single nucleotide (A, C, T or G) differs among different members of the population

- Polymorphisms can be SNPs, small insertions, deletions, or larger structural variants (translocations, copy number variants, etc.)



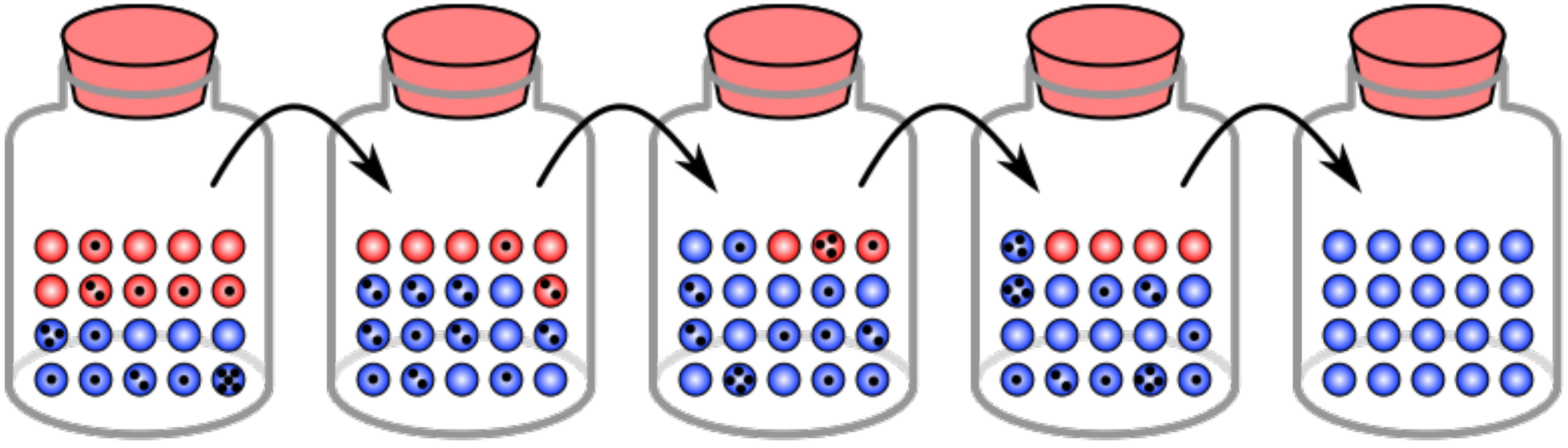
Population genetics

- **Population genetics** is the study of the genetic composition of populations, and of its changes across time and space.



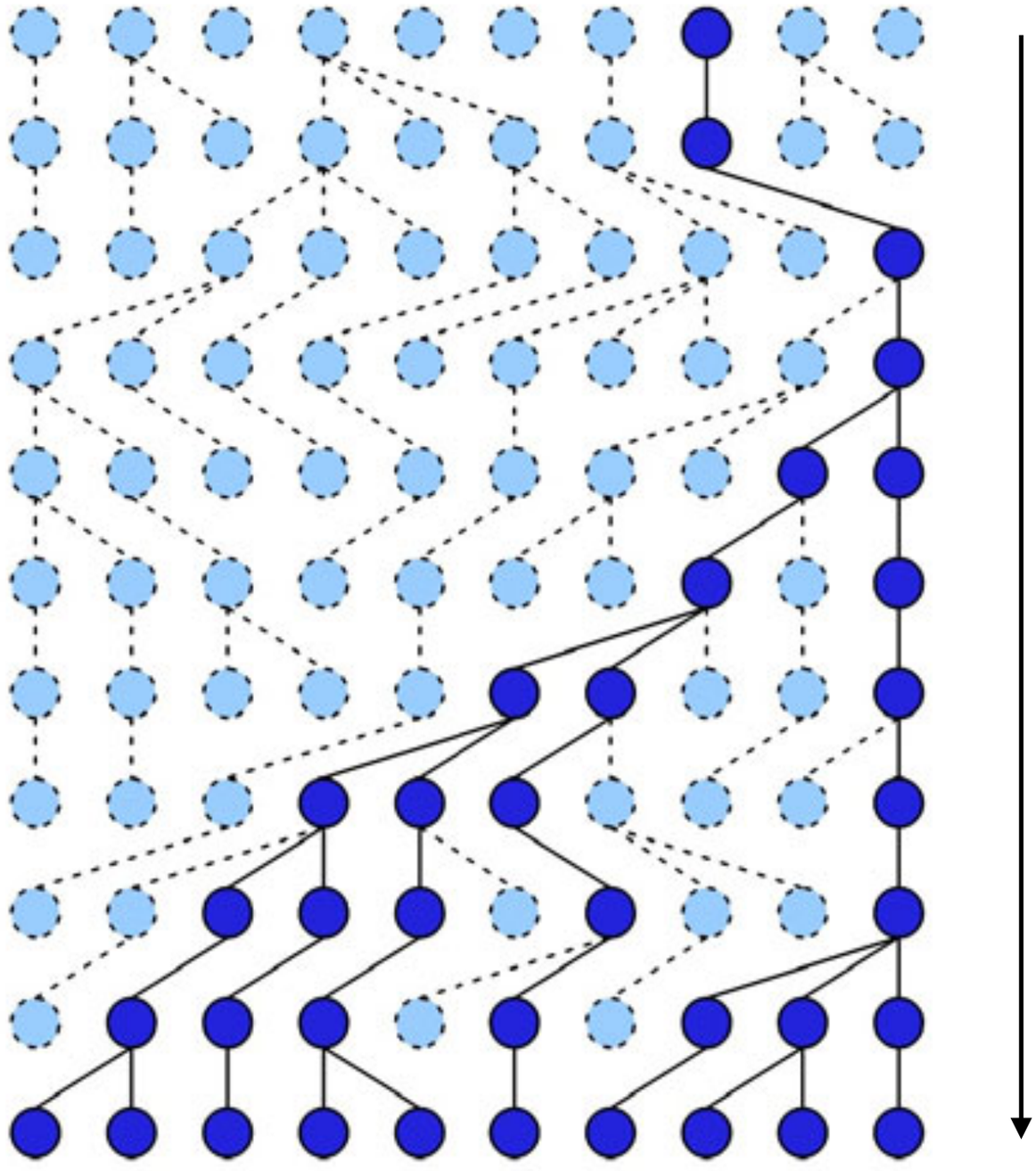
Population genetics

- **“Forwards” approach:** study allele frequencies in a population evolving forwards in time (more intuitive, but sometimes less useful).



Hardy-Weinberg model

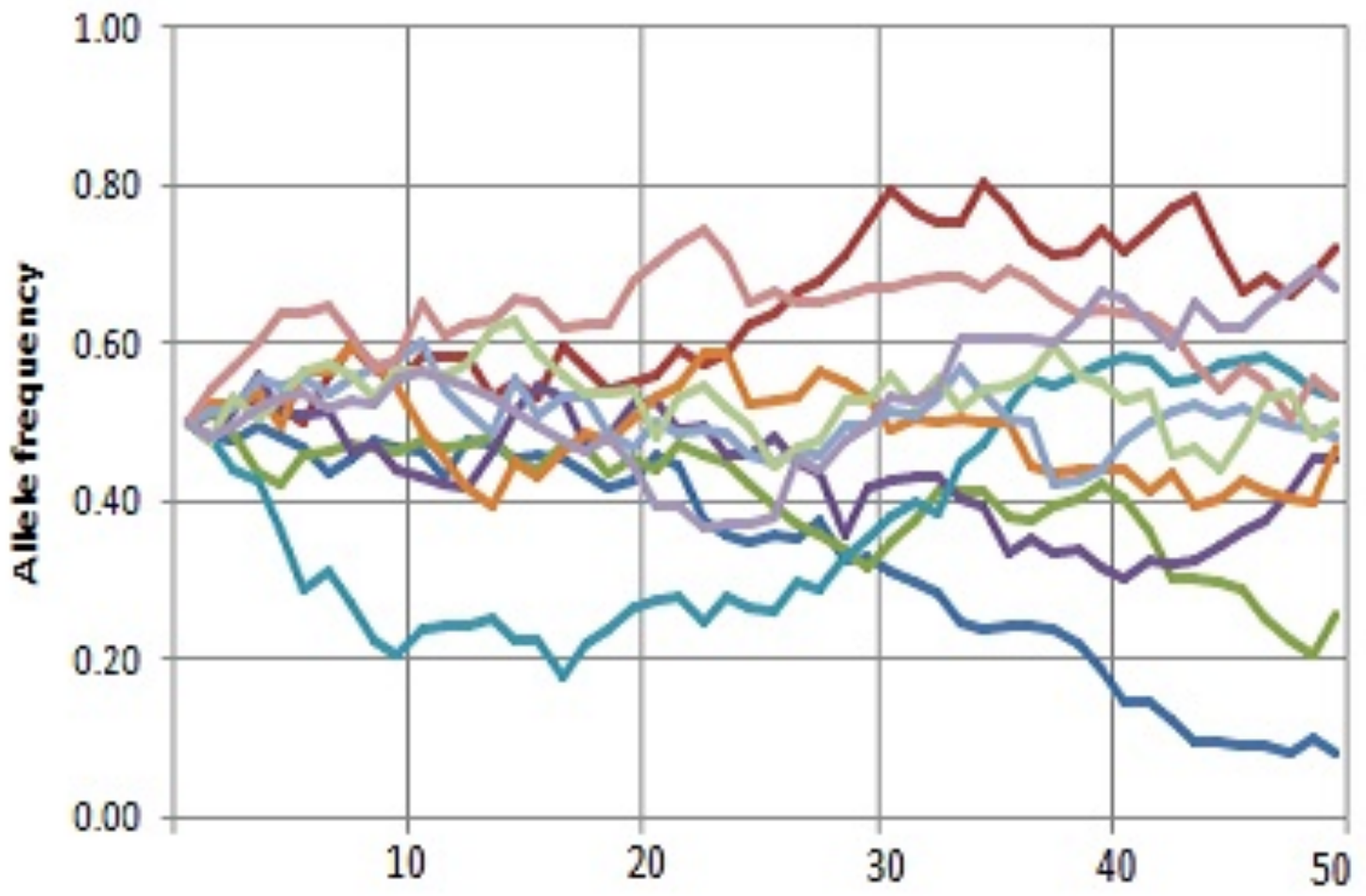
First generation



Time

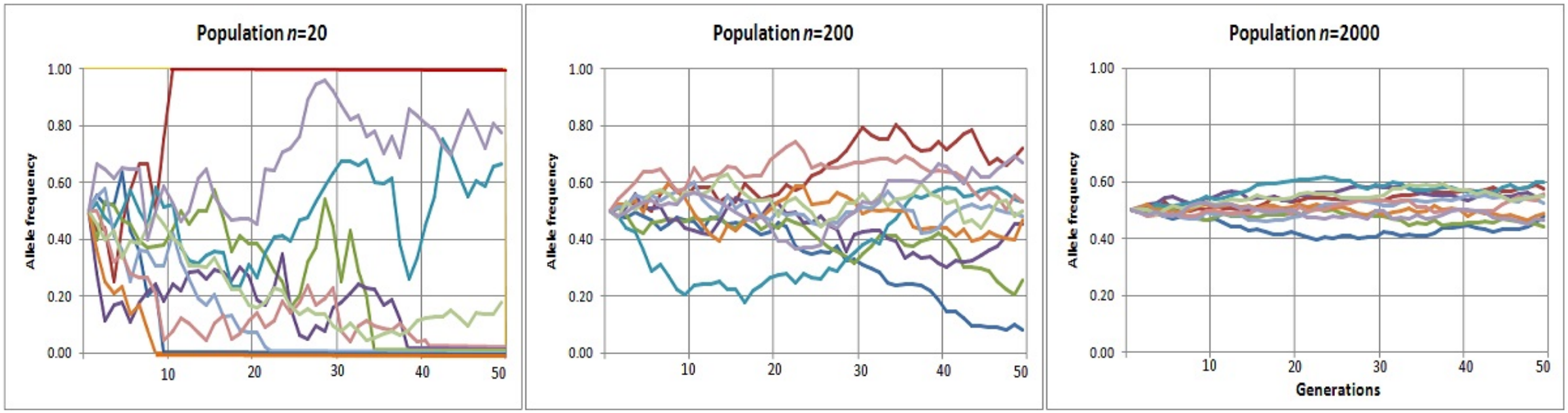
Present-day generation

Genetic drift



Genetic drift

- Random fluctuations in allele frequencies due to random sampling at each generation

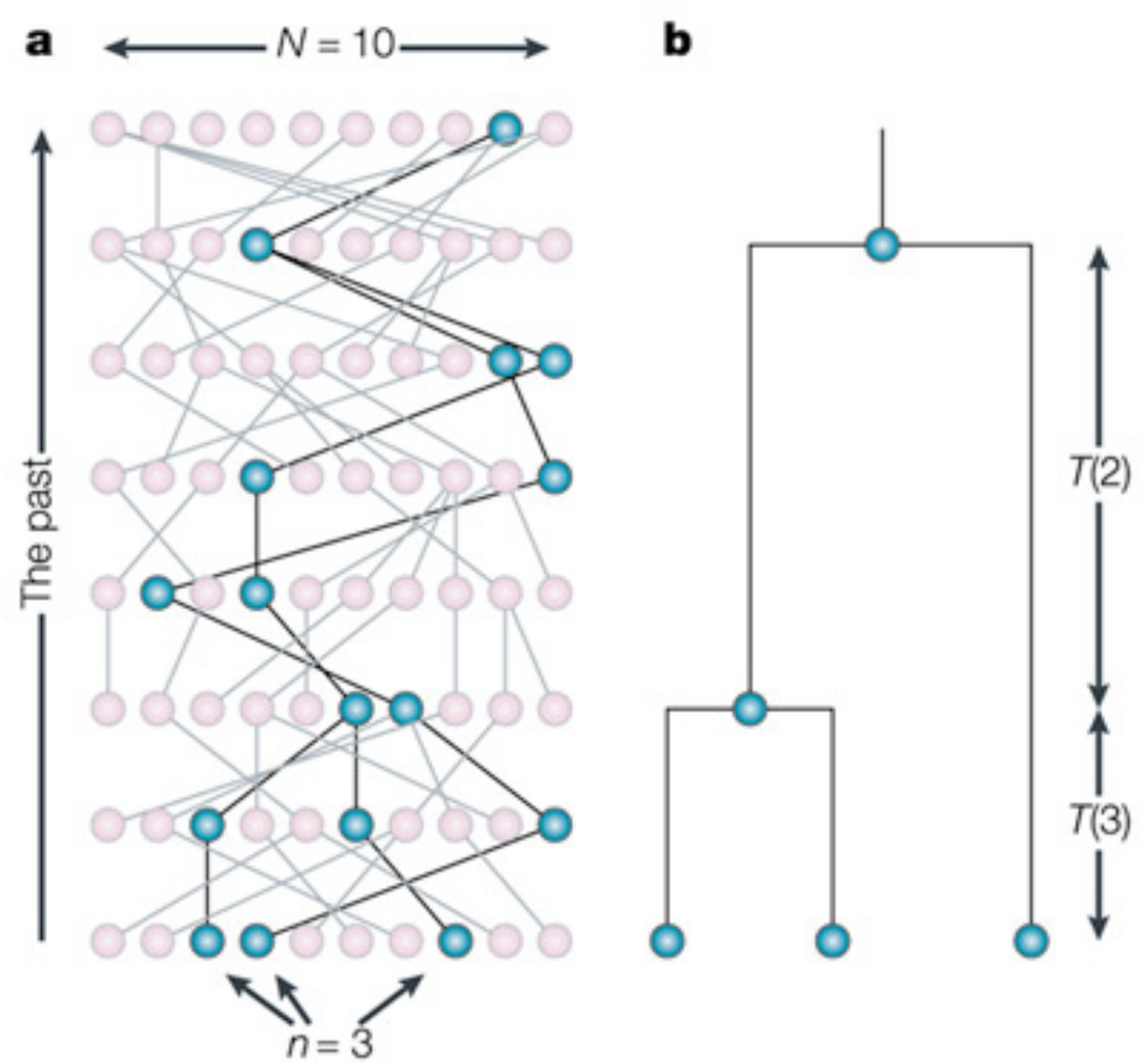


Exercises

- https://github.com/FerRacimo/Archaeomics/blob/master/IntroPopGen.md

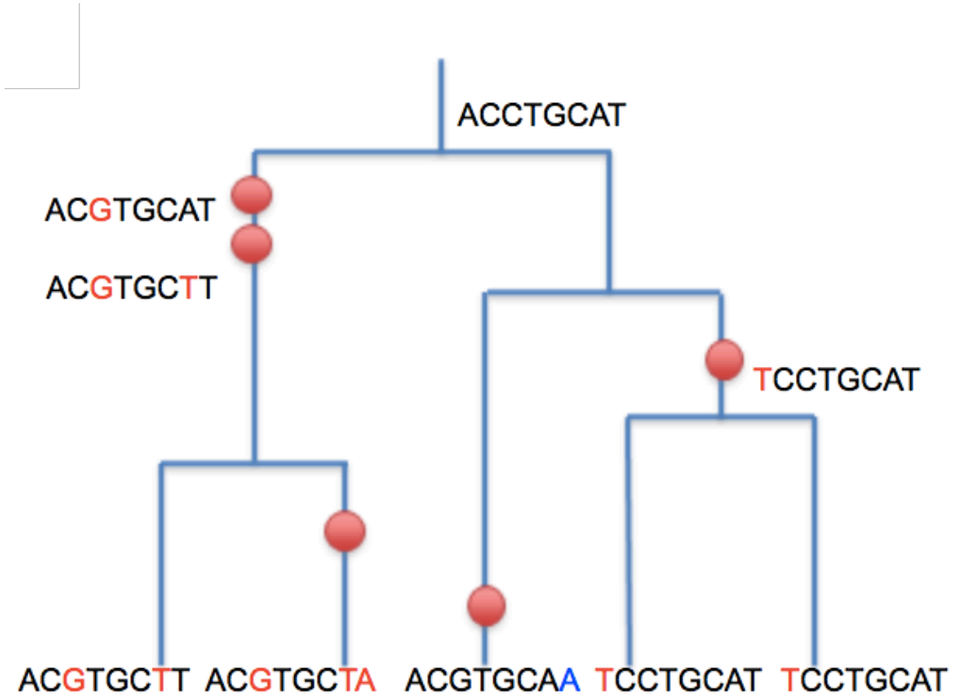
Population genetics

- **“Genealogical” / “coalescent” approach:** study genealogies of sampled individuals evolving backwards in time



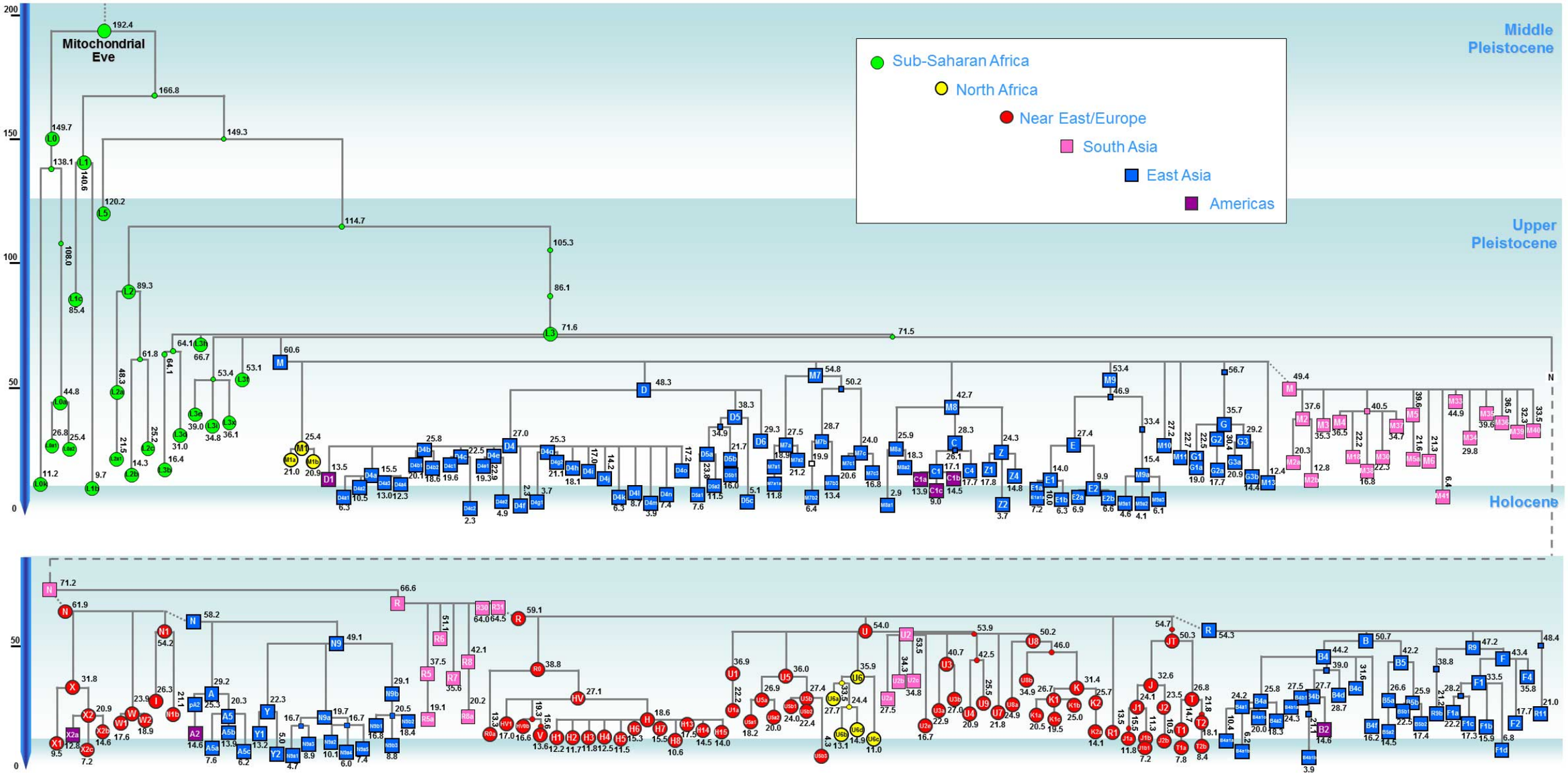
Coalescent approach

- Small regions of the genome (and non-recombining loci like mitochondria) are characterized by a single genealogy



ACGTGCTT
ACGTGCTA
ACGTGCAA
T CCTGCAT

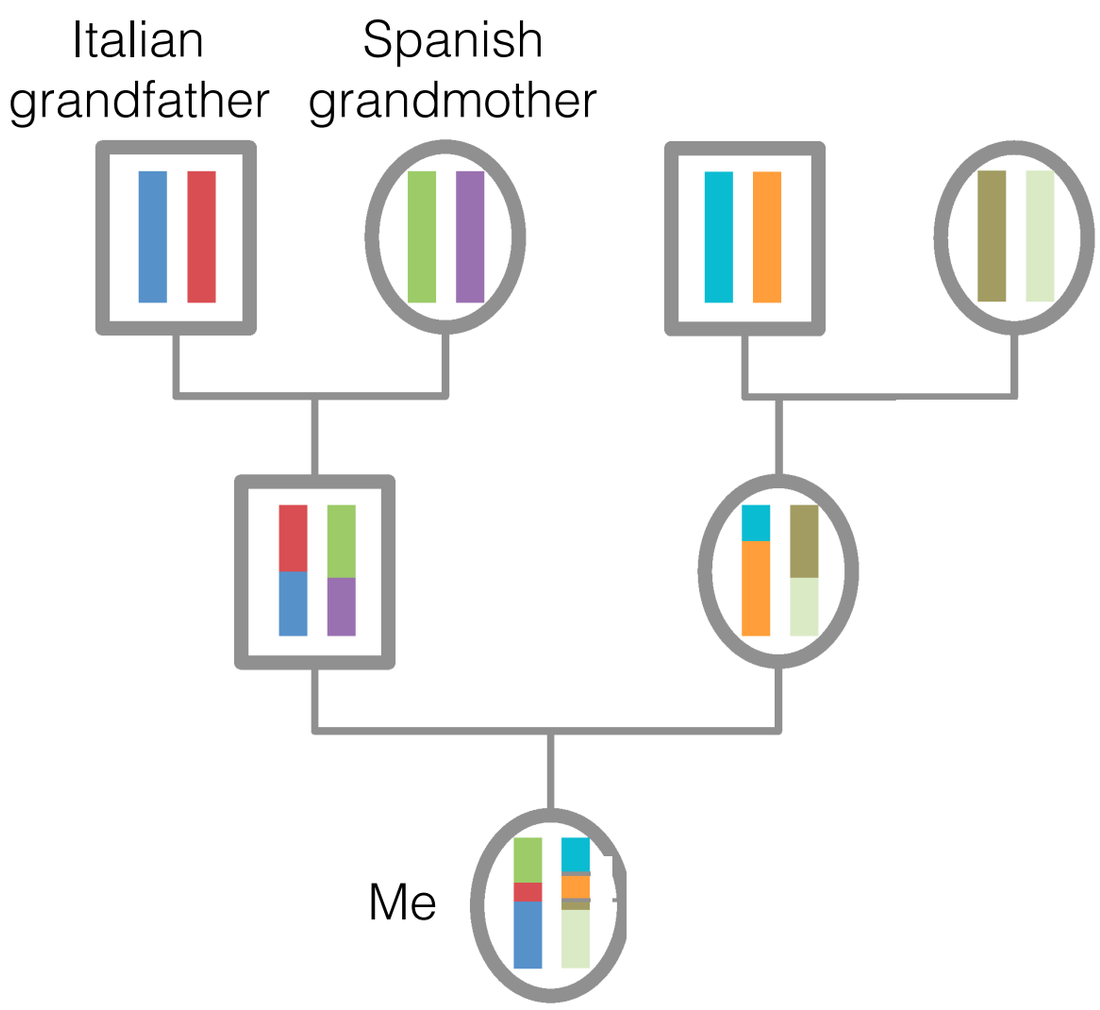
Mitochondrial tree



Coalescent trees are not the same as population histories!

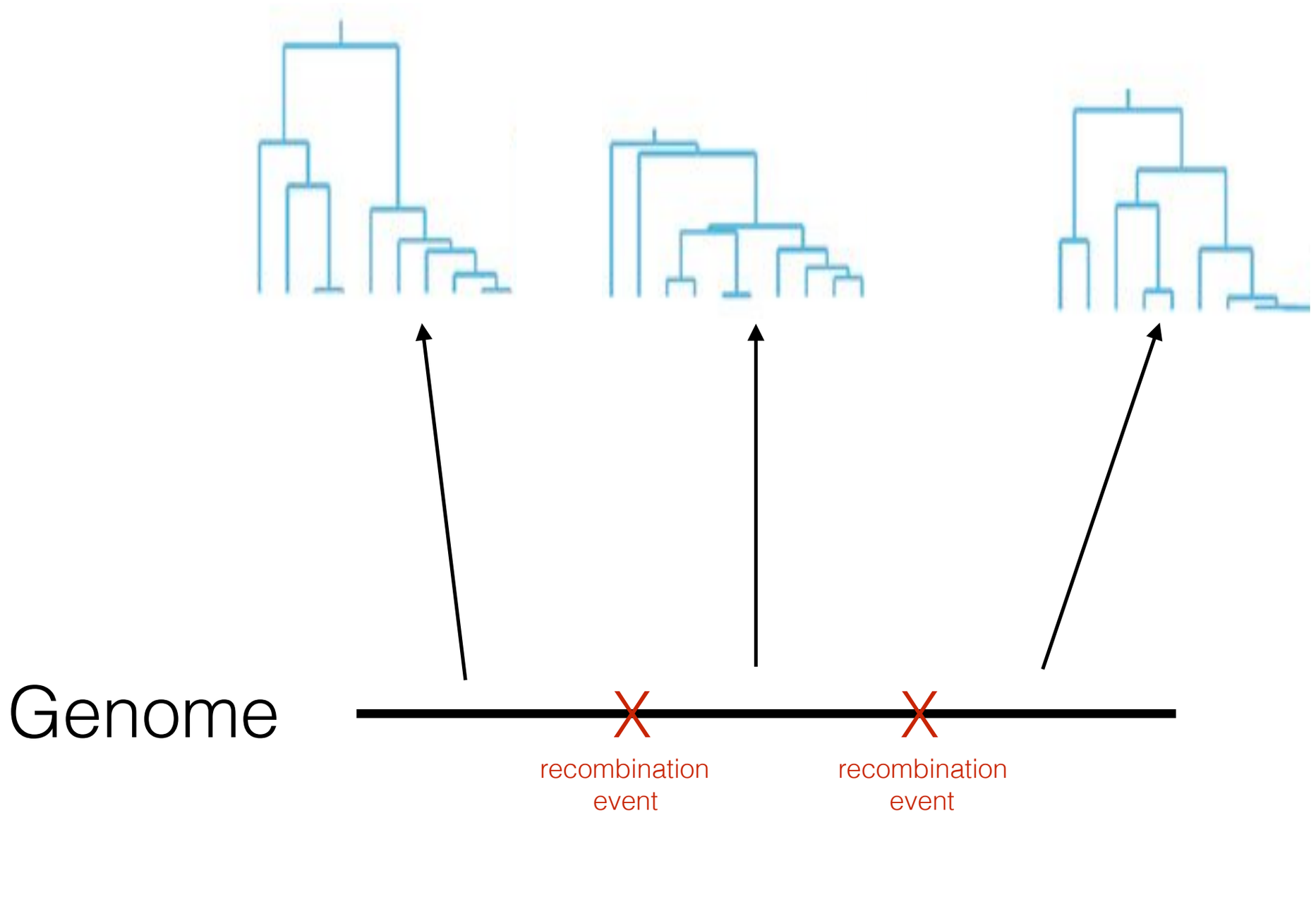
- If there were no recombination, the entire genome would be characterized by a single tree, and a single common ancestor.

- **This is not the case:** recombination creates different genealogies in different parts of the genome, with different ancestors as we go backwards in time.



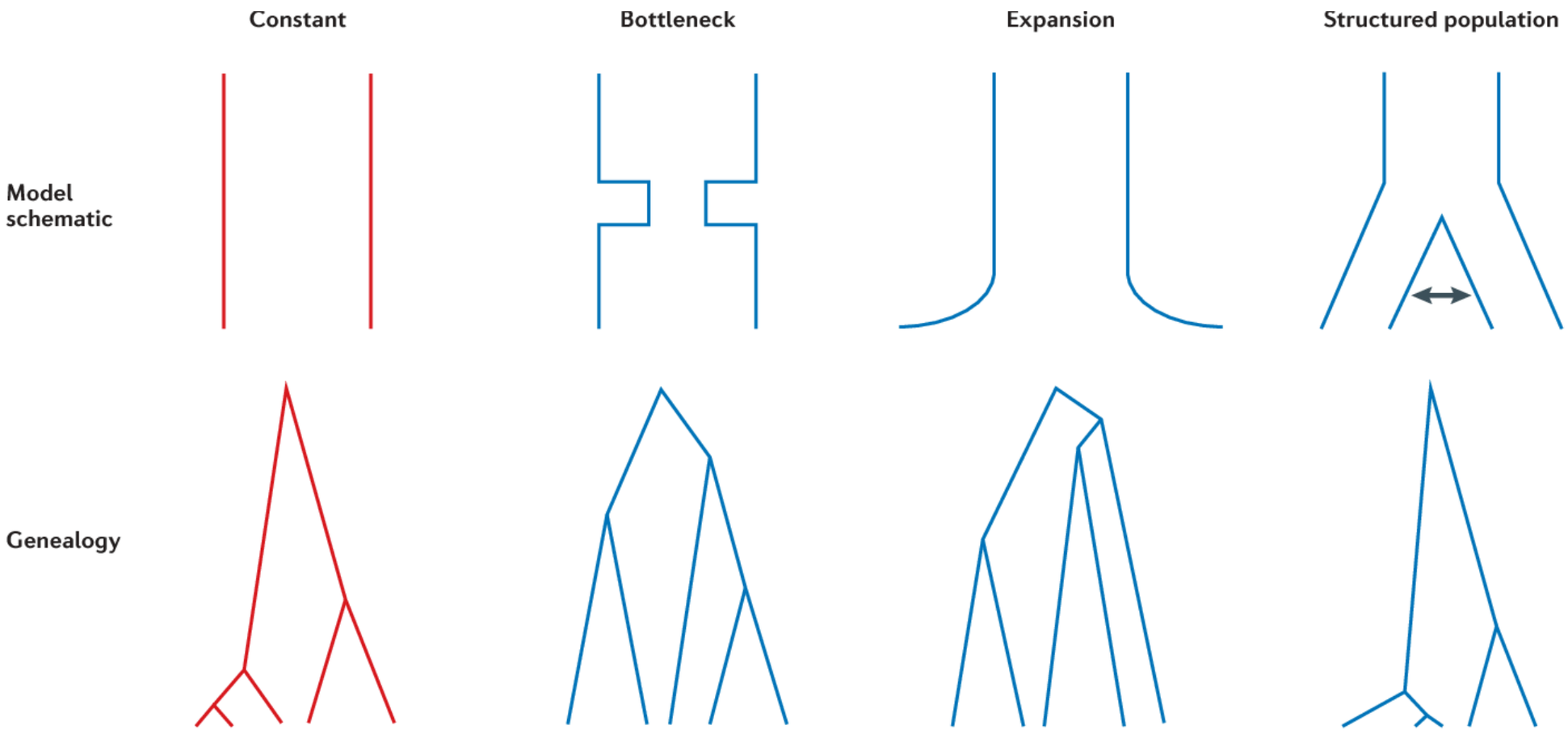
Genealogies across the genome

- The entire genome is characterized by a set of (correlated) trees that are different because of recombination events



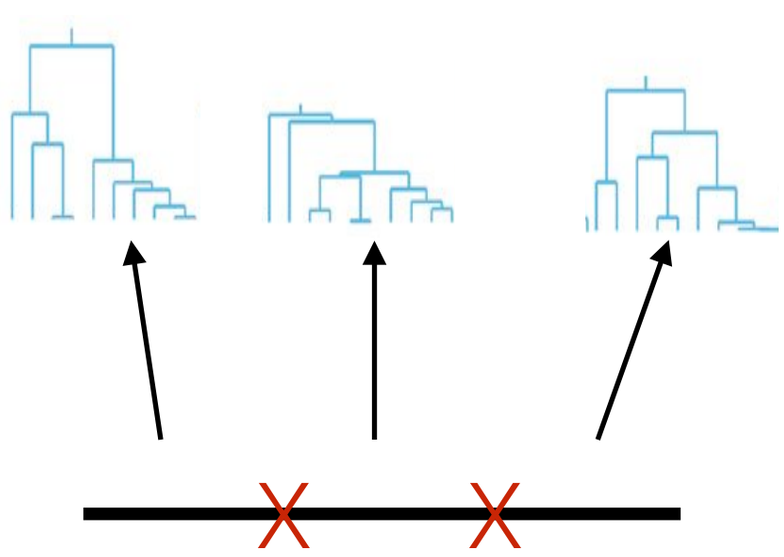
Genealogies across the genome

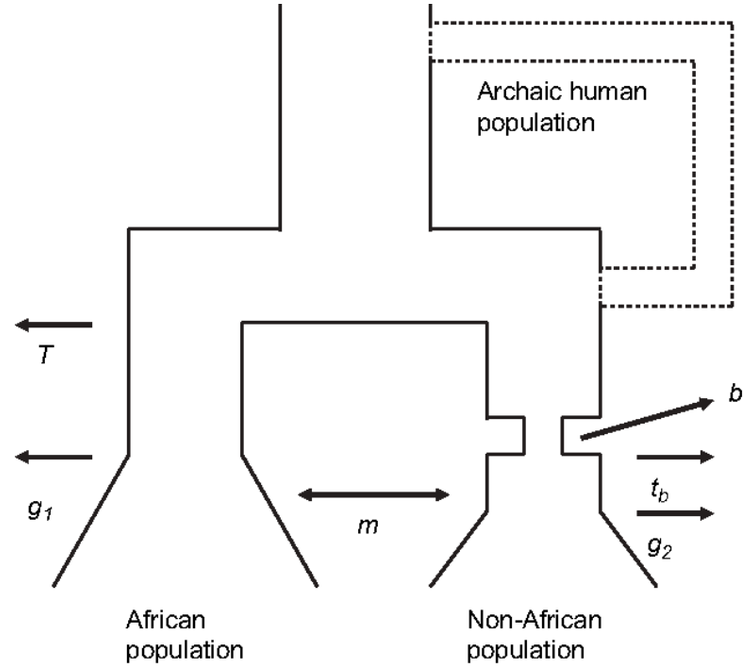
- The trees contain information about demographic history!



Schraiber and Akey 2015

Population genetic inference





Genomic sequences with mutations (**observed**)

Trees along the genome (**unknown**)

Demographic history (**unknown**)

Today

- Introduction to population genetics
- **Exploratory vs. hypothesis-driven analyses**
- PCA
- Useful datasets for human paleogenomics
- Latent mixed-membership models (“Structure”)

Exploratory analyses

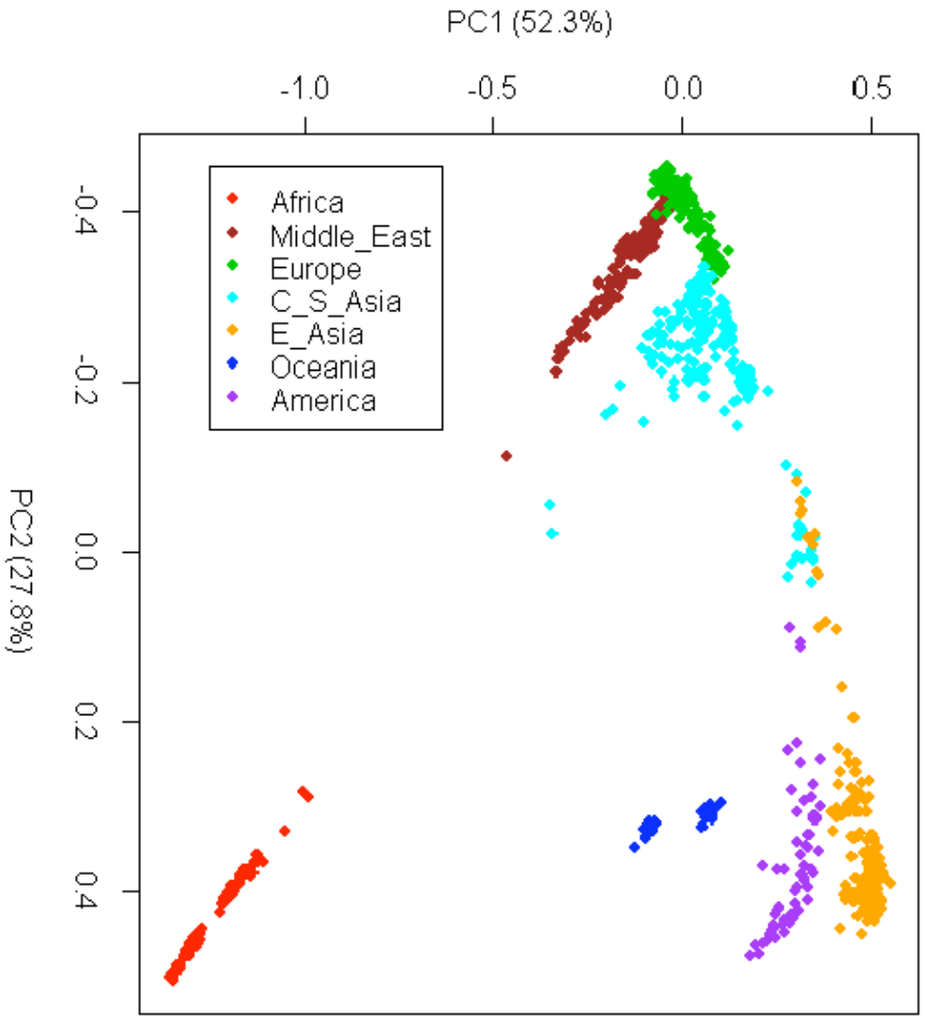
- When we've just gotten some population genomic data (ancient or modern) and don't know where to start with it.

- What are the general patterns of variation? How much structure is there in my data?

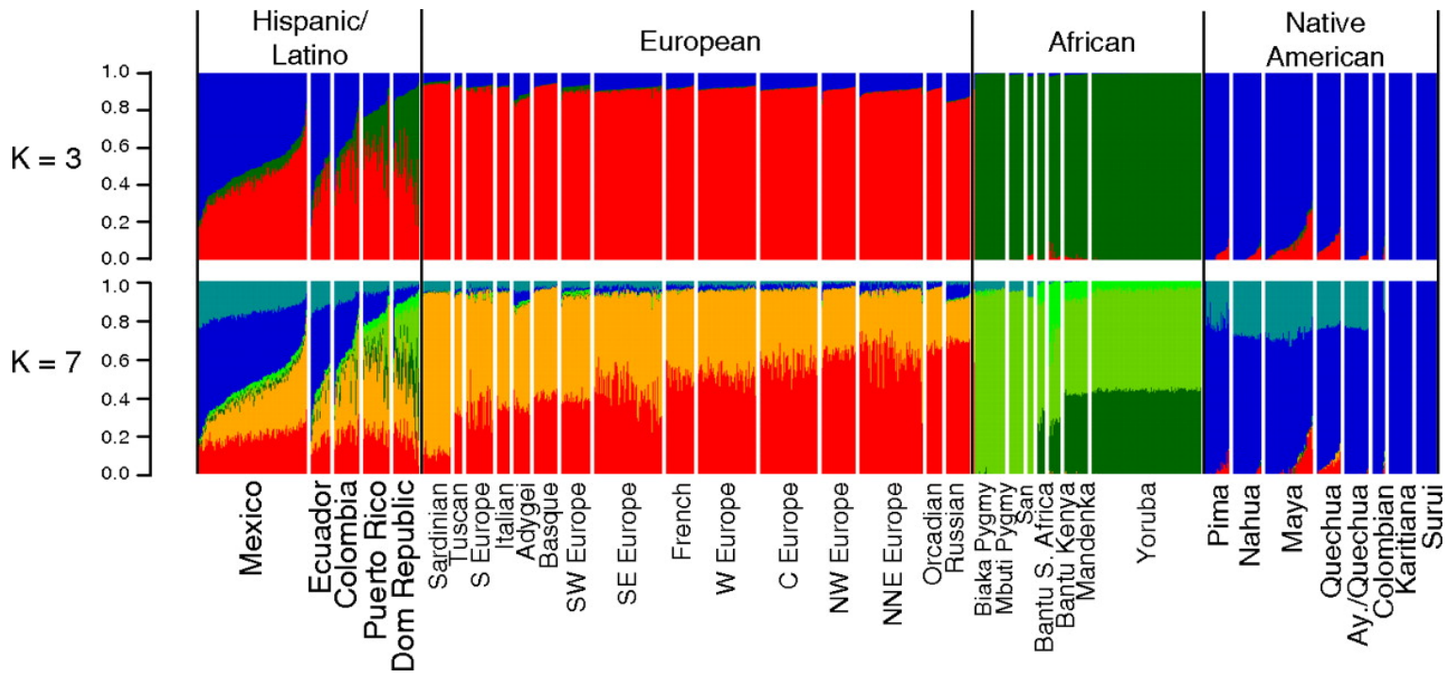
- Which groups can be clustered together? Which groups are best modeled as a mixture of other groups?

- Are certain samples particularly interesting?

Exploratory analyses



PCA



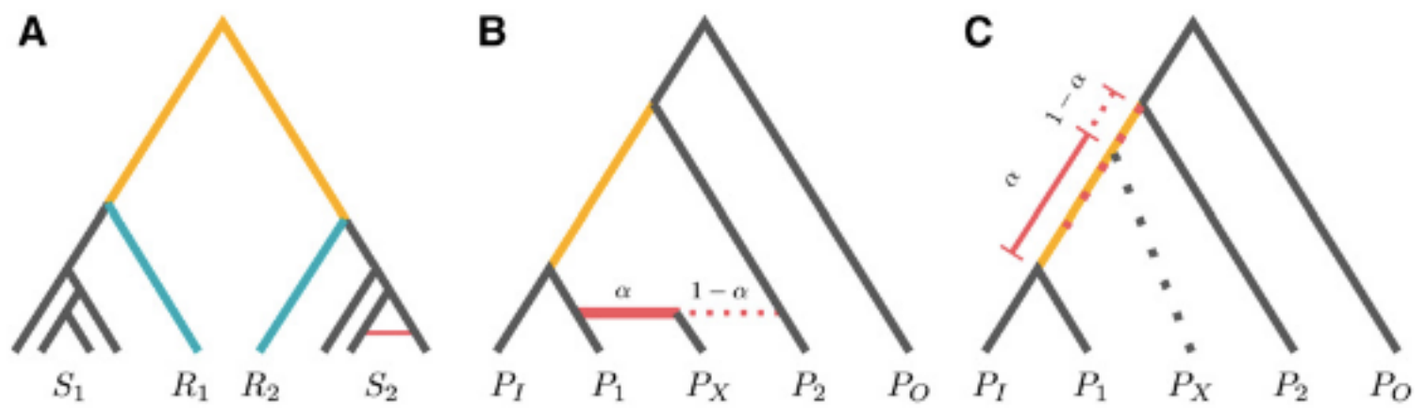
Latent mixed-membership models ("Structure")

Hypothesis-driven analyses & parameter estimation

- When we want to start building models of population history and testing particular hypotheses about the past.

- Is a particular population the result of an admixture event? What are the admixture proportions? When did the event happen?

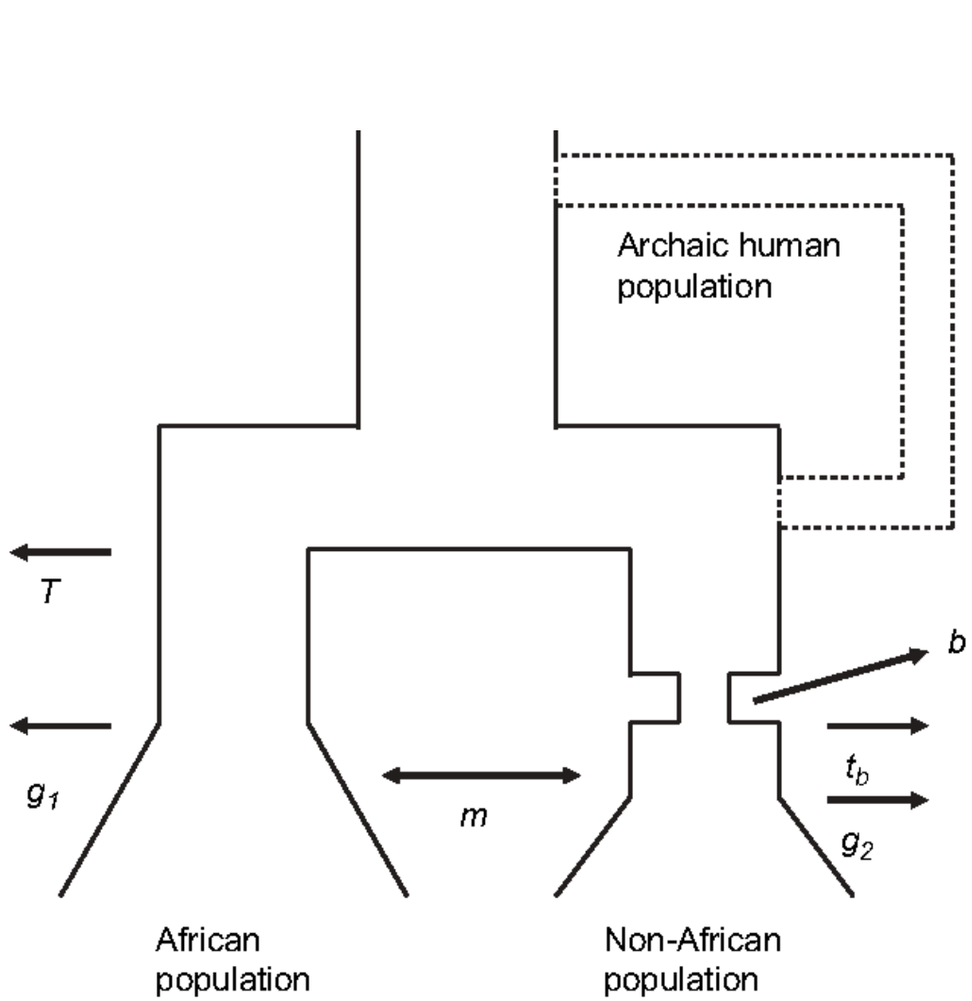
- When did two populations diverge? When did a population contract or expand?

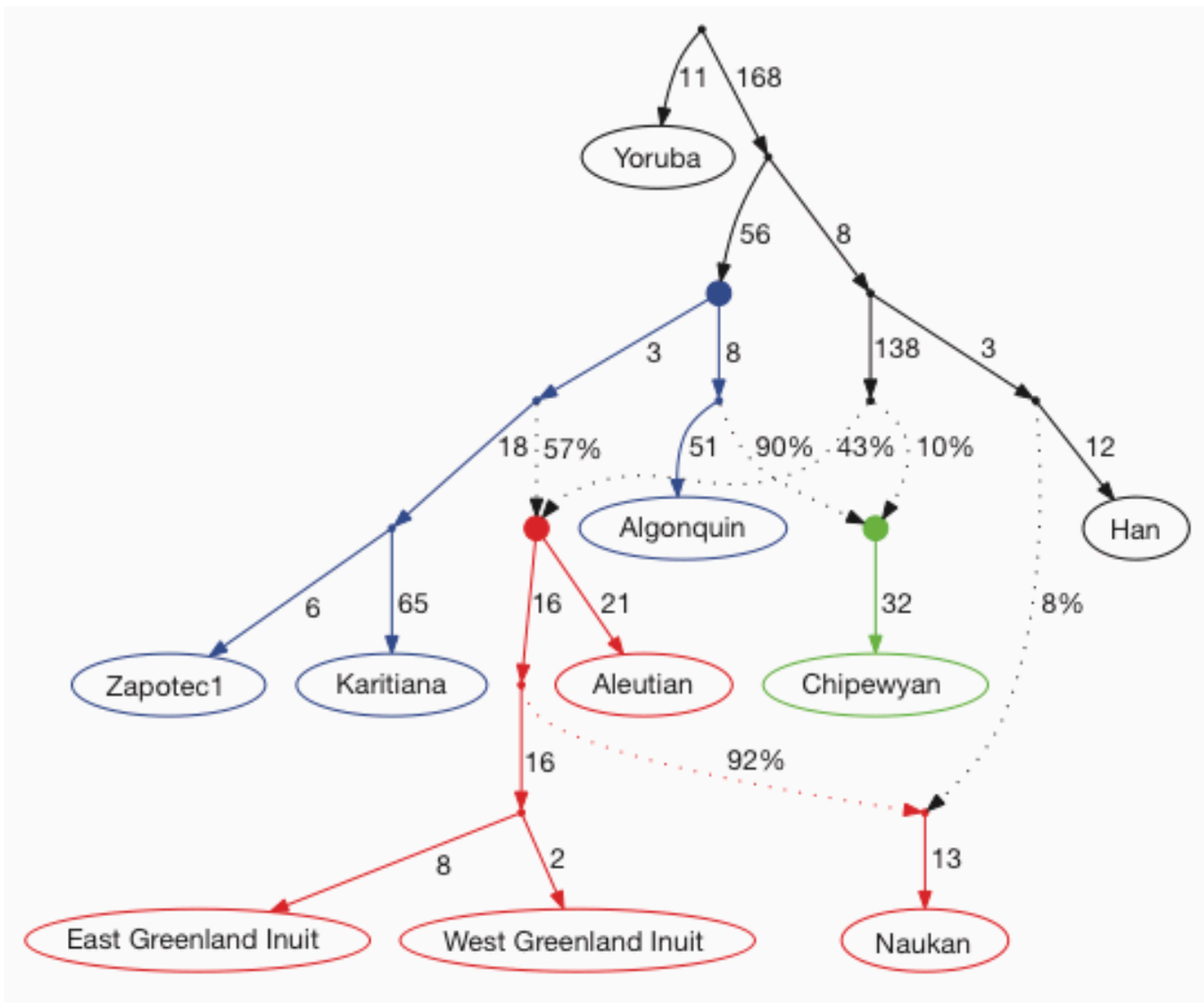


Peter et al. 2016

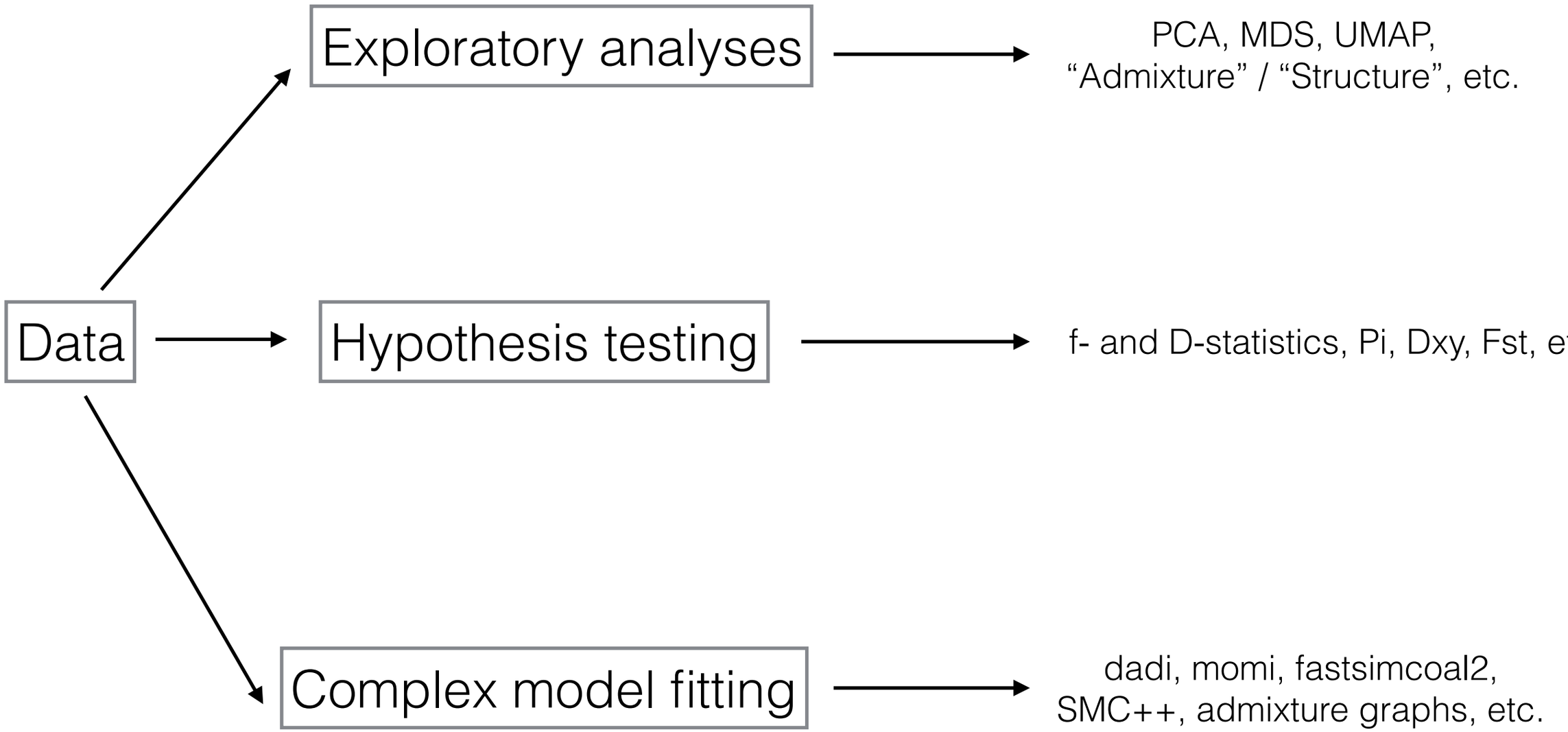
Inference of complex demographies

- What is the best history (or set of histories) that can best describe my data?





Paleogenomic workflow



Today

- Introduction to population genetics
- Exploratory vs. hypothesis-driven analyses
- **PCA**
- Useful datasets for human paleogenomics
- Latent mixed-membership models (“Structure”)

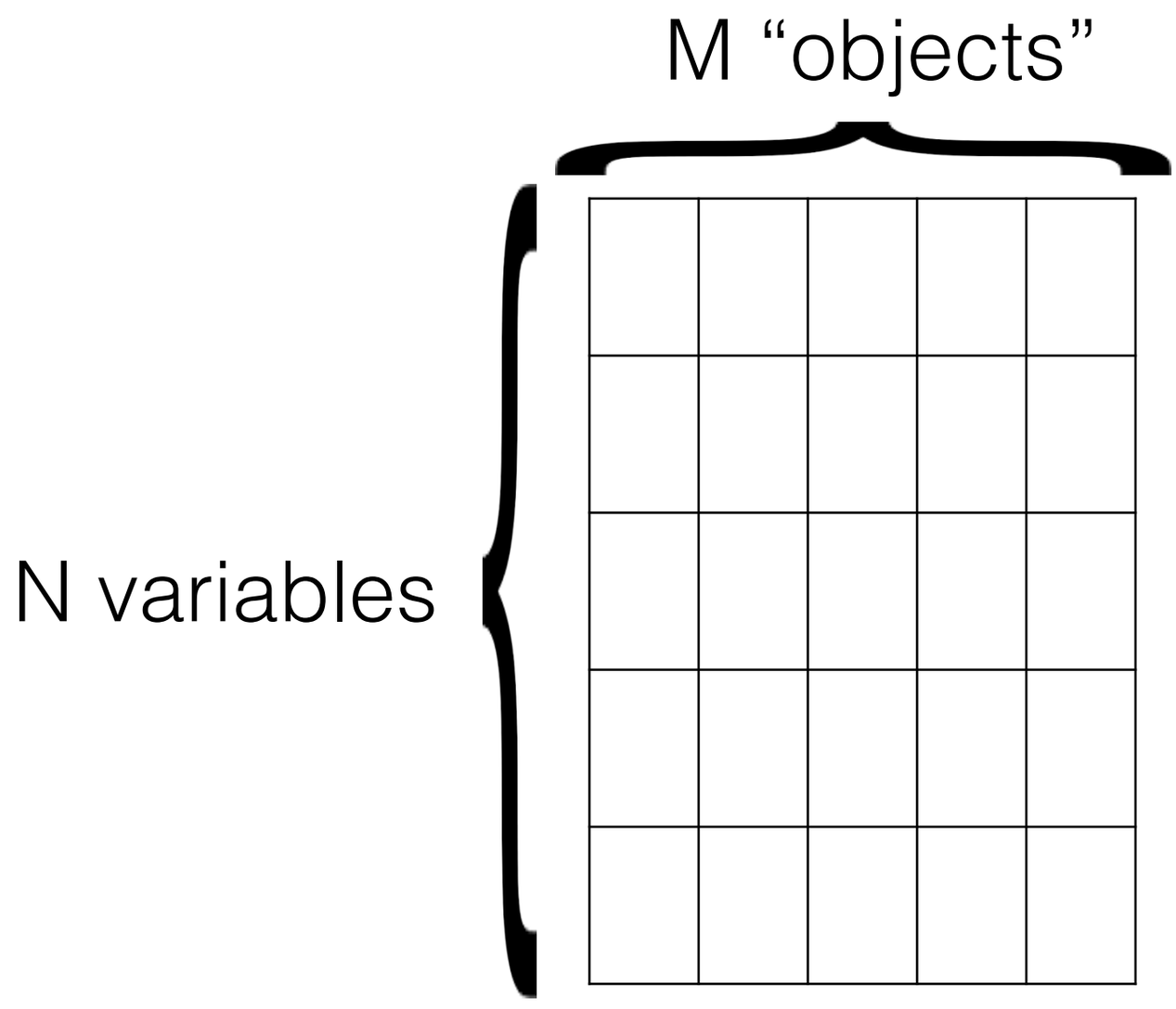
PCA

- Principal Component Analysis: an orthogonal transformation of a set of observations of correlated variables into a set of values of linearly uncorrelated variables (?????)

- Useful for **exploratory data analysis** and widely used in many fields.

- When variables are **numerous** and may be **correlated in unknown ways**

Multivariate data



Genotype data

M genomes

N loci

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

Motivation

- Order objects in a way that **similar objects are near each other** and dissimilar objects are farther from each other

- Reduce data to a few axes of variation (dimensionality-reduction) to facilitate **recognition of patterns**

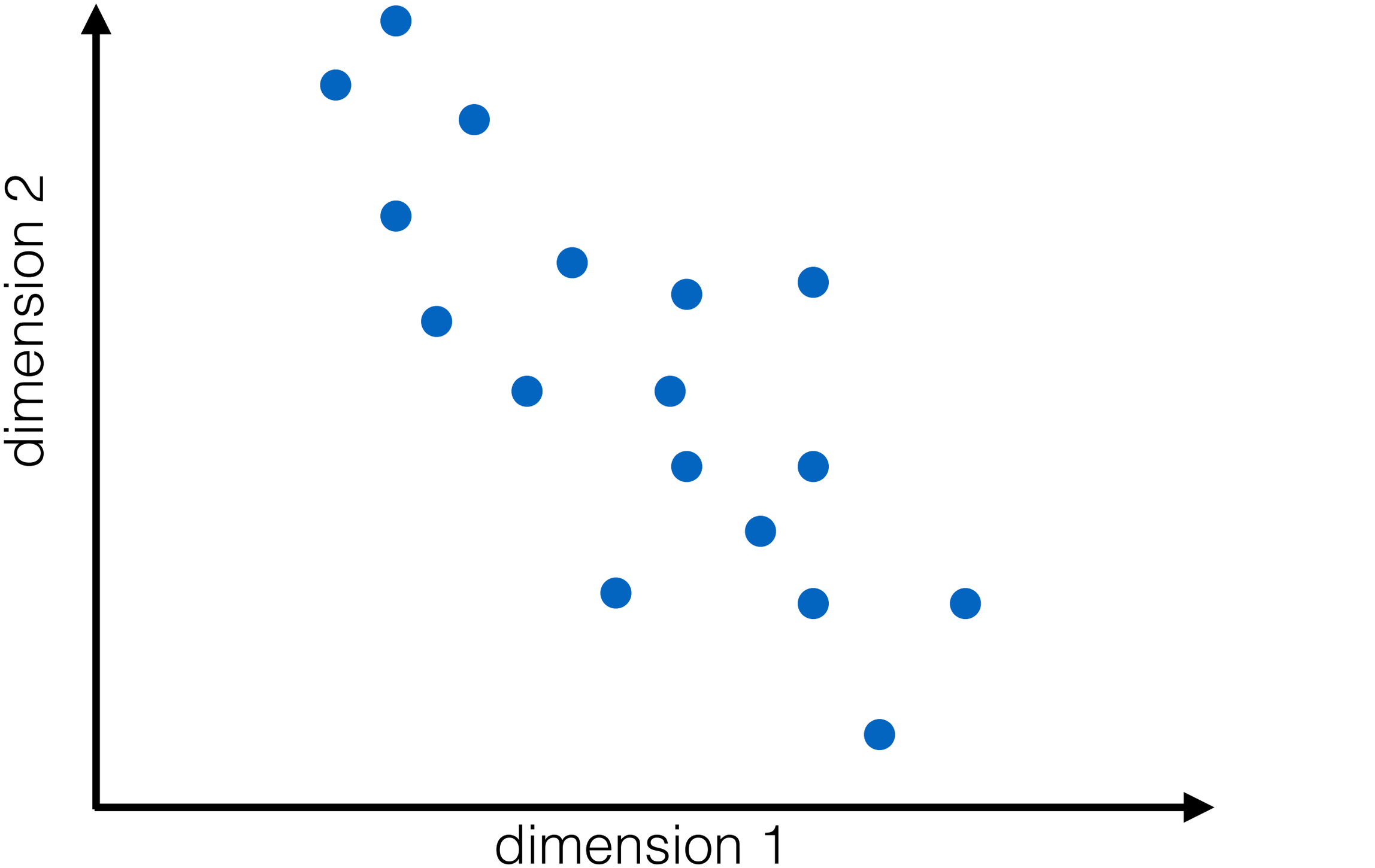
- Gradients reflect **underlying factors or processes**

Motivation in population genetics

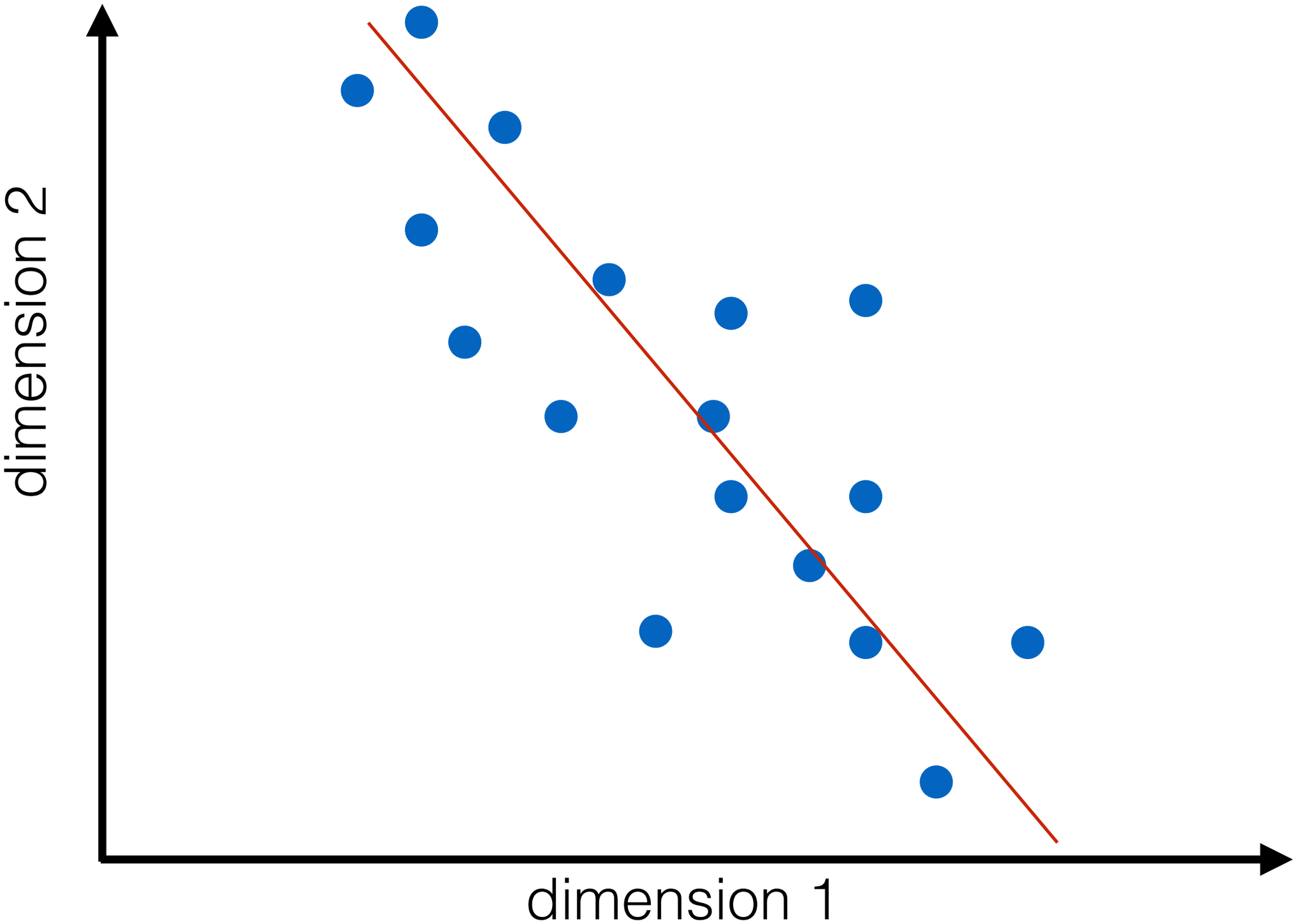
- Order individuals into groups with similar genetic histories

- Find the strongest axes of variation among different populations

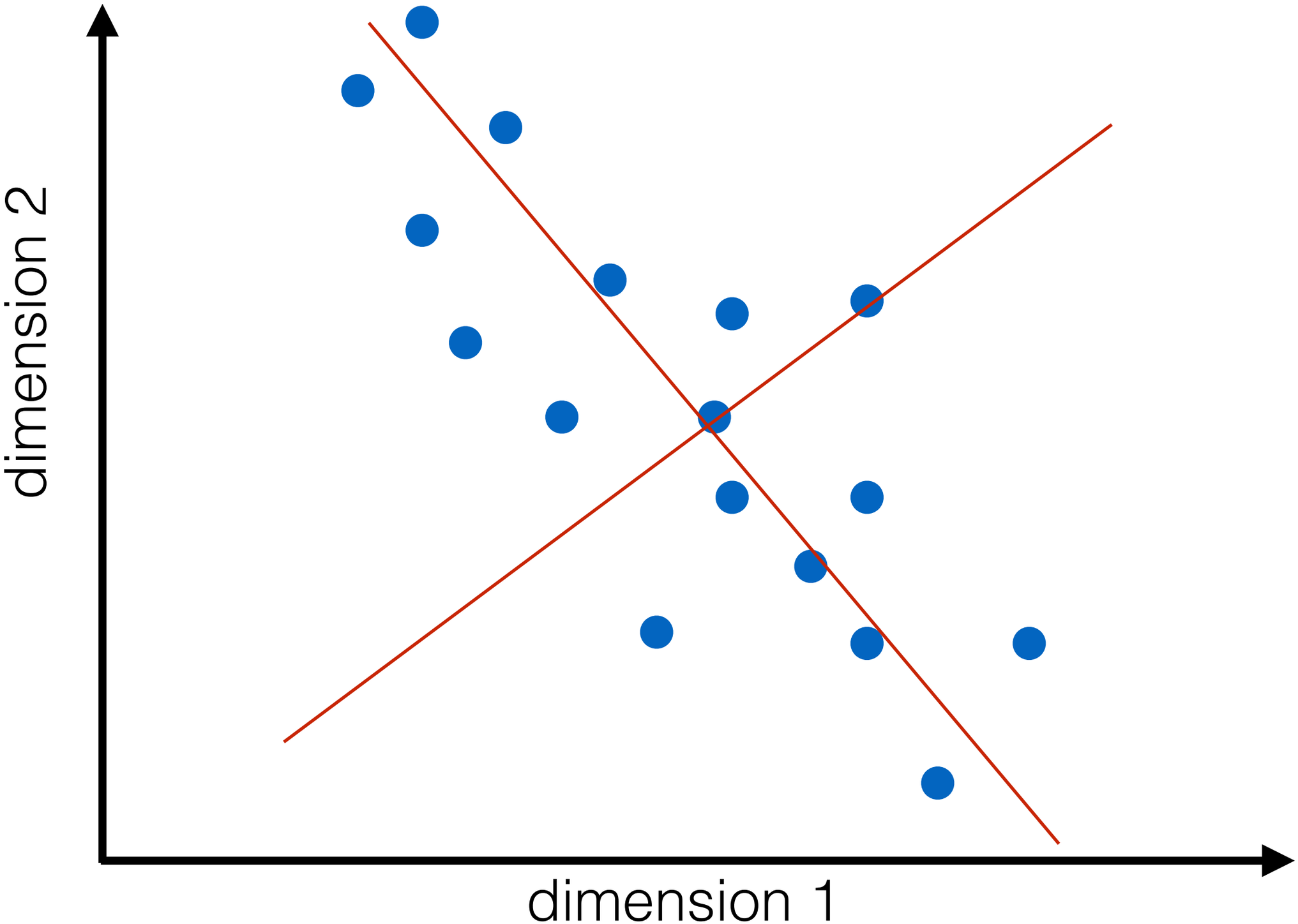
Finding the best **orthogonal** axes of variation



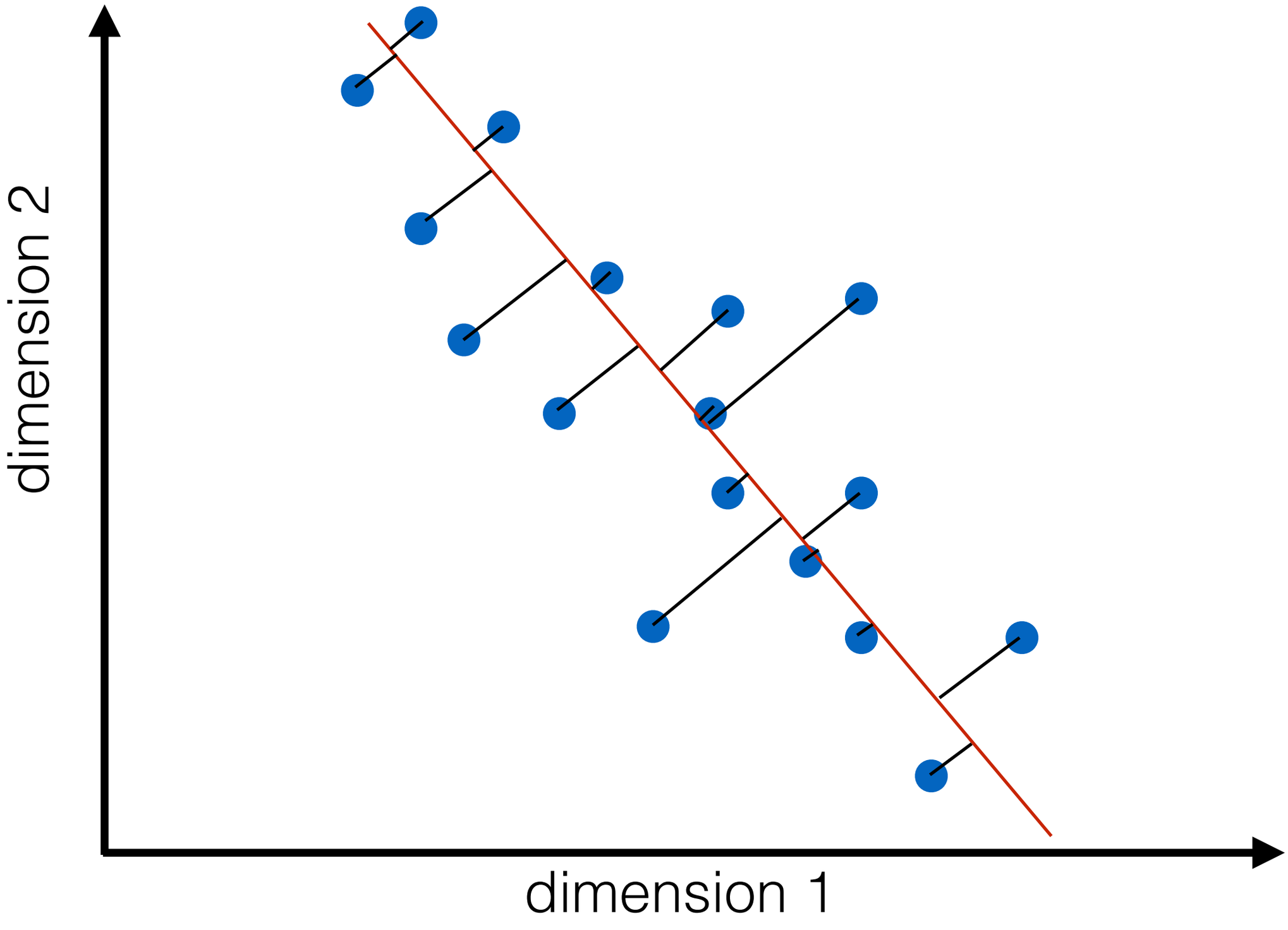
Finding the best **orthogonal** axes of variation



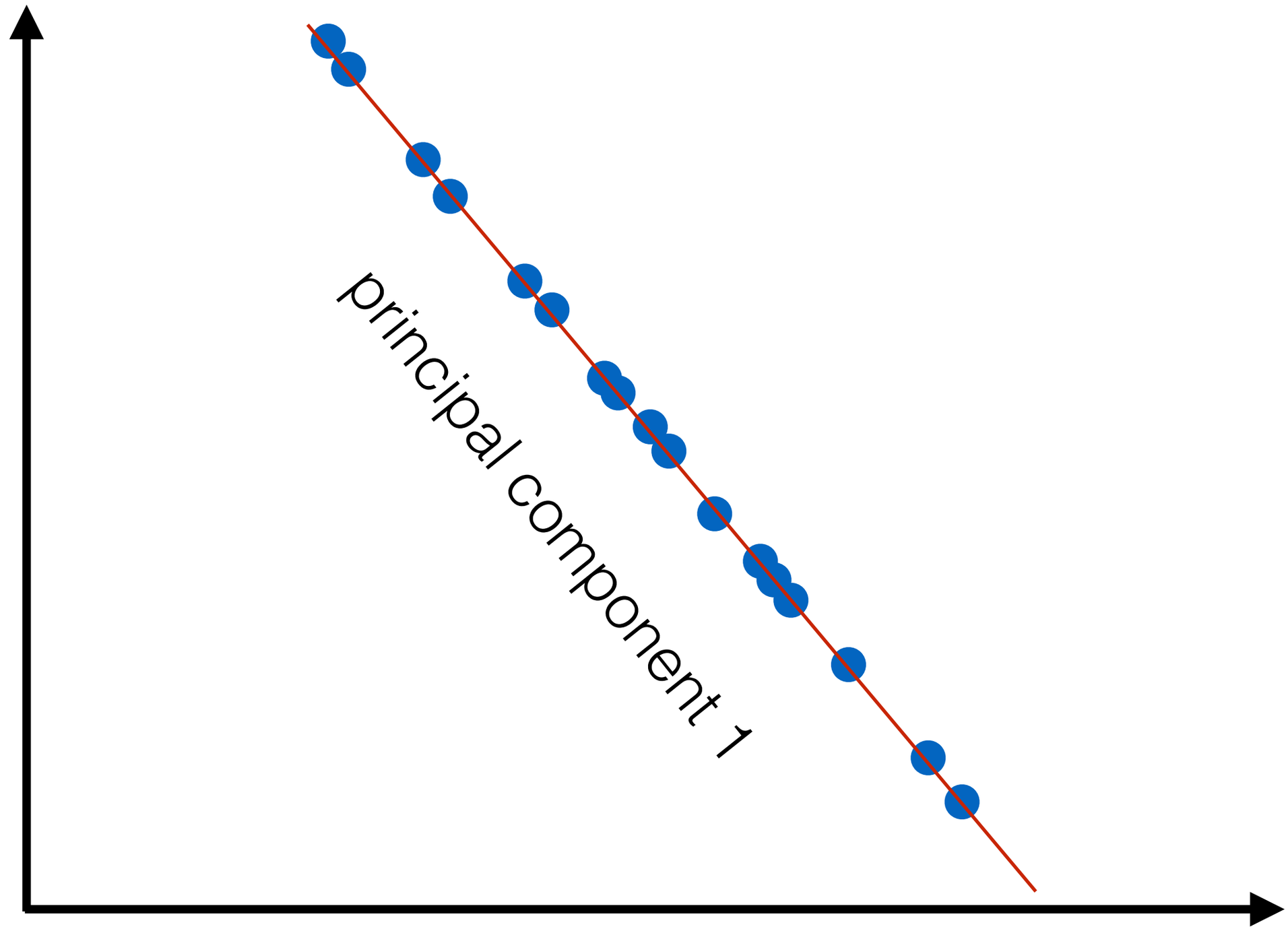
Finding the best **orthogonal** axes of variation



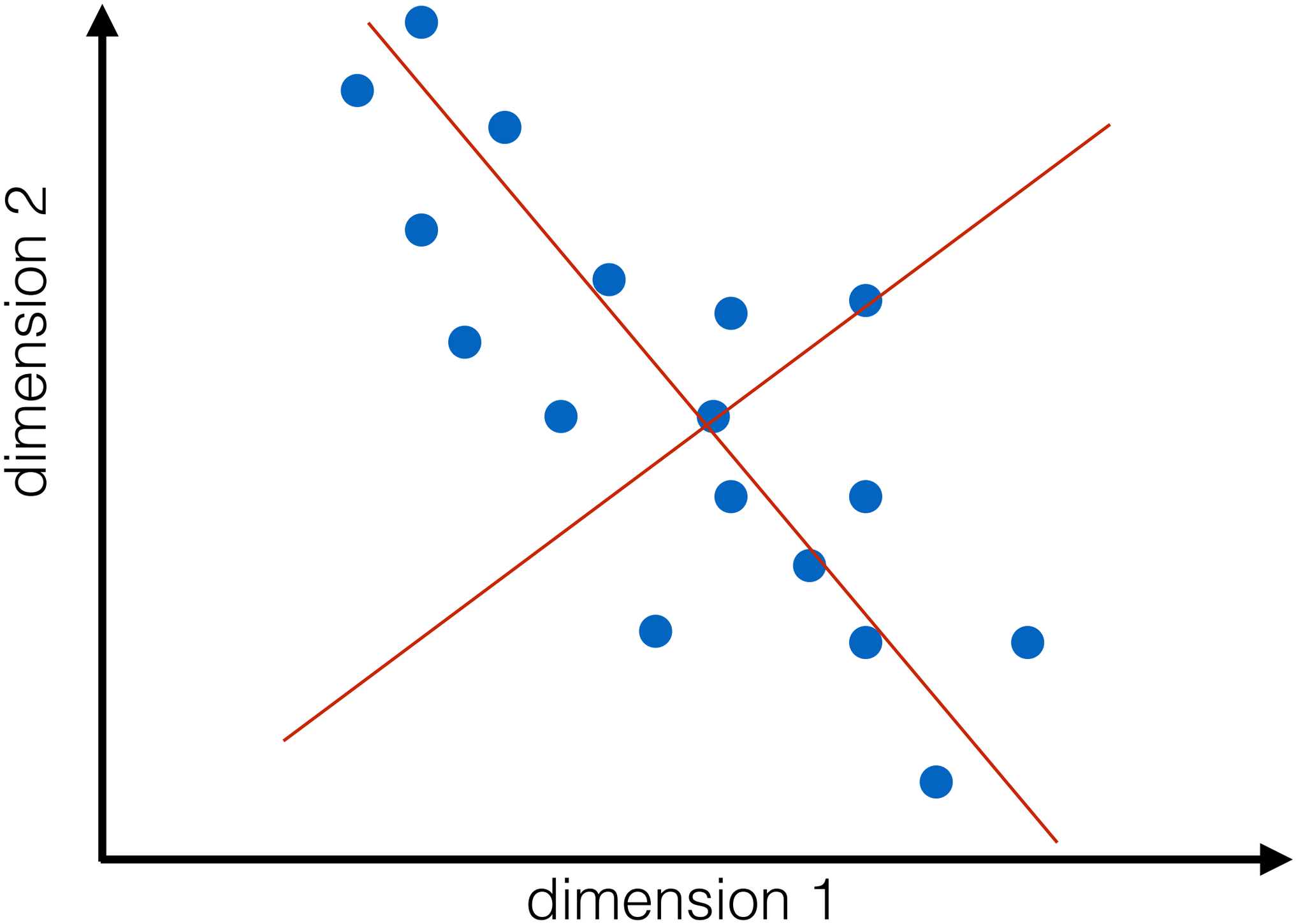
Projecting data onto **orthogonal** axes



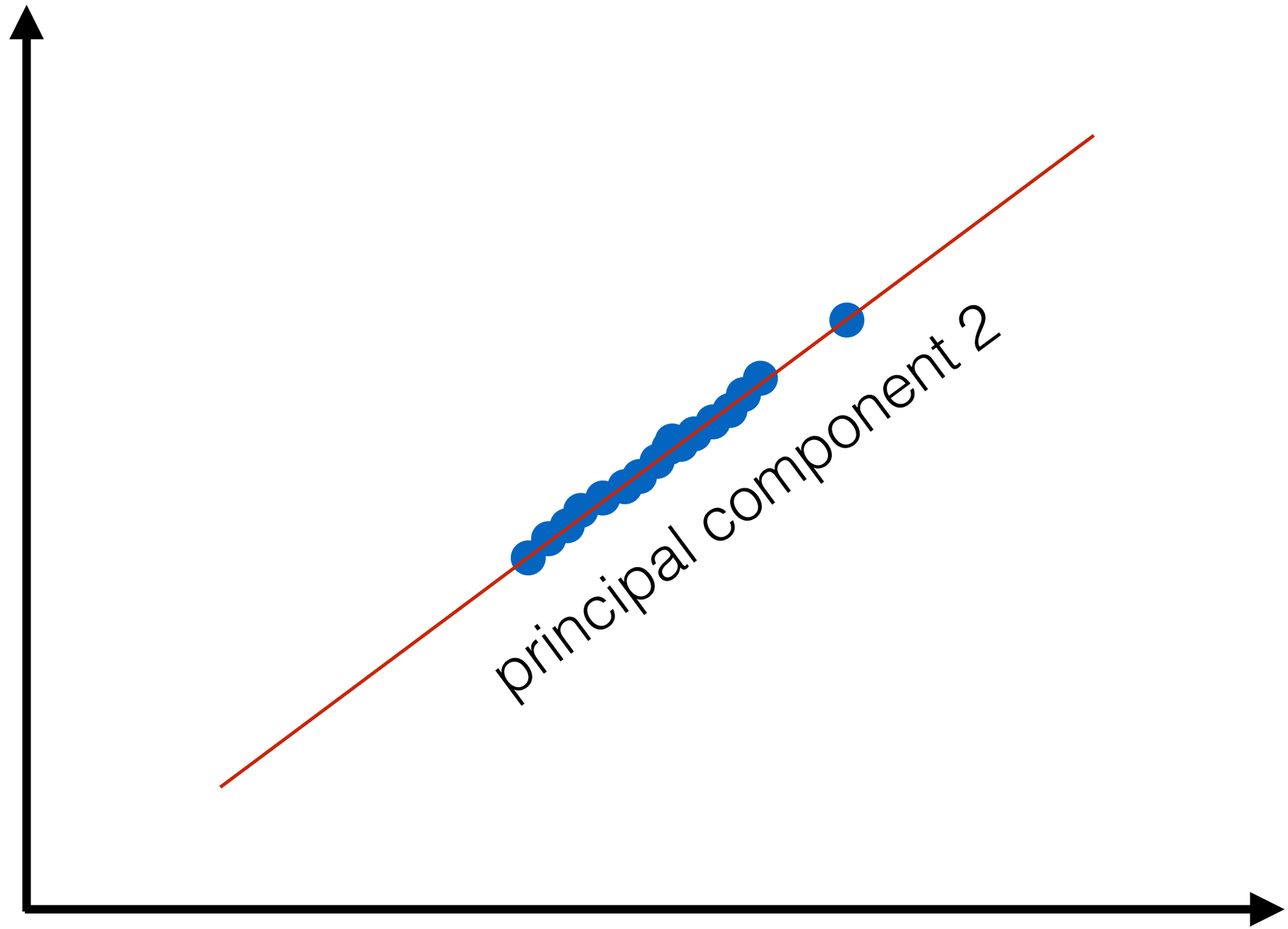
Projecting data onto **orthogonal** axes



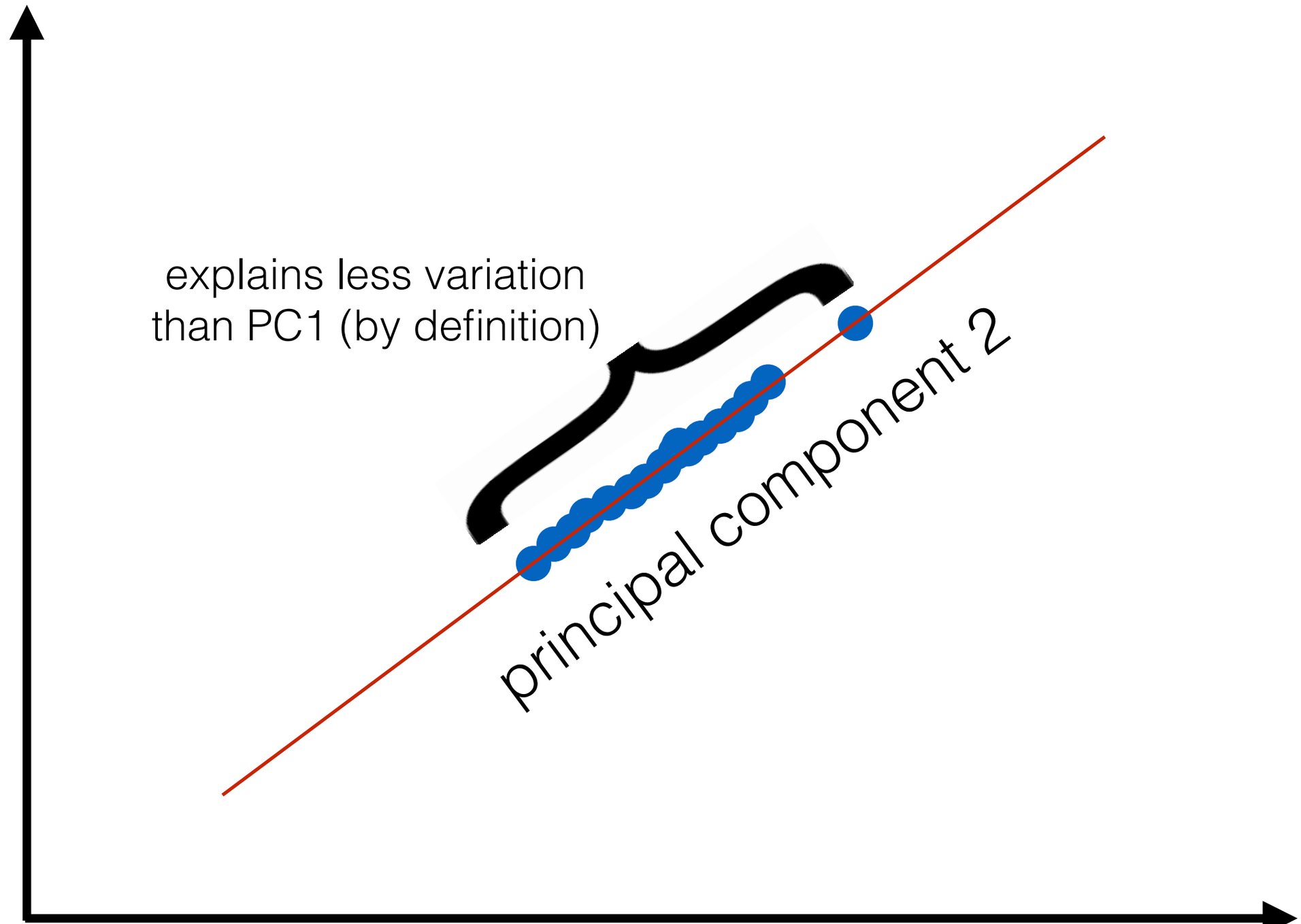
Projecting data onto **orthogonal** axes



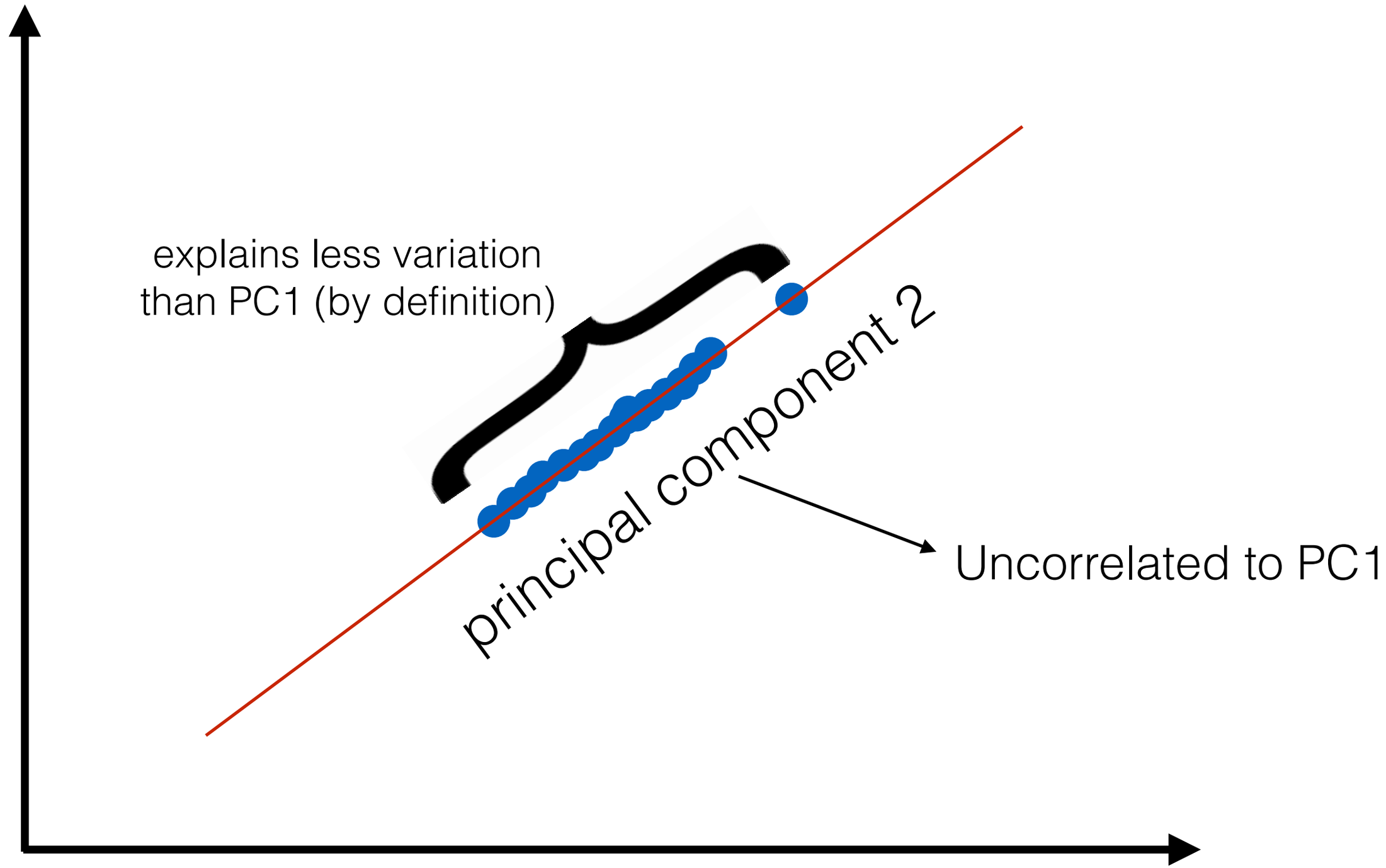
Projecting data onto **orthogonal** axes



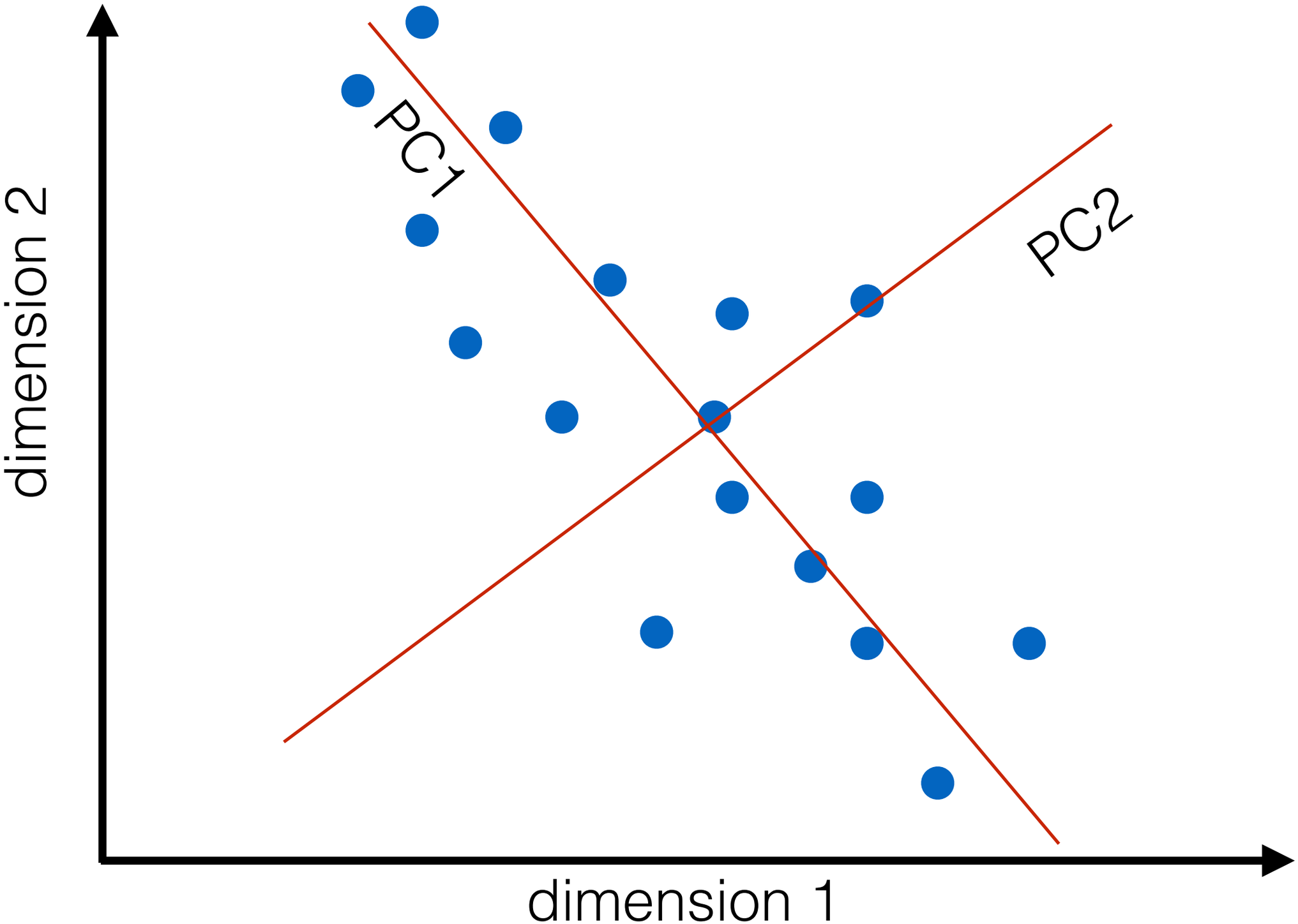
Projecting data onto **orthogonal** axes



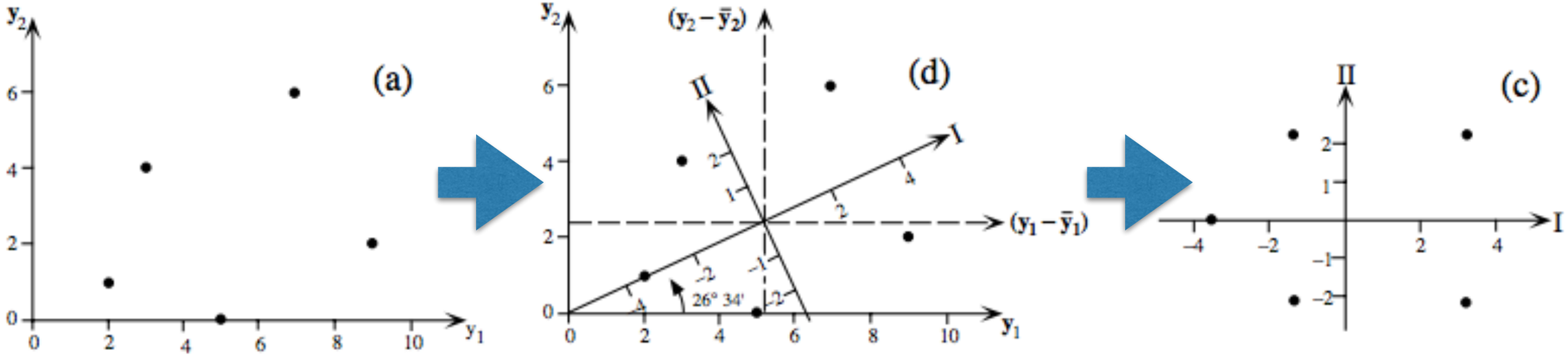
Projecting data onto **orthogonal** axes



New set of axes



PCA as a **centered rotation** in N-dimensional space



Rotation in N-dimensional space

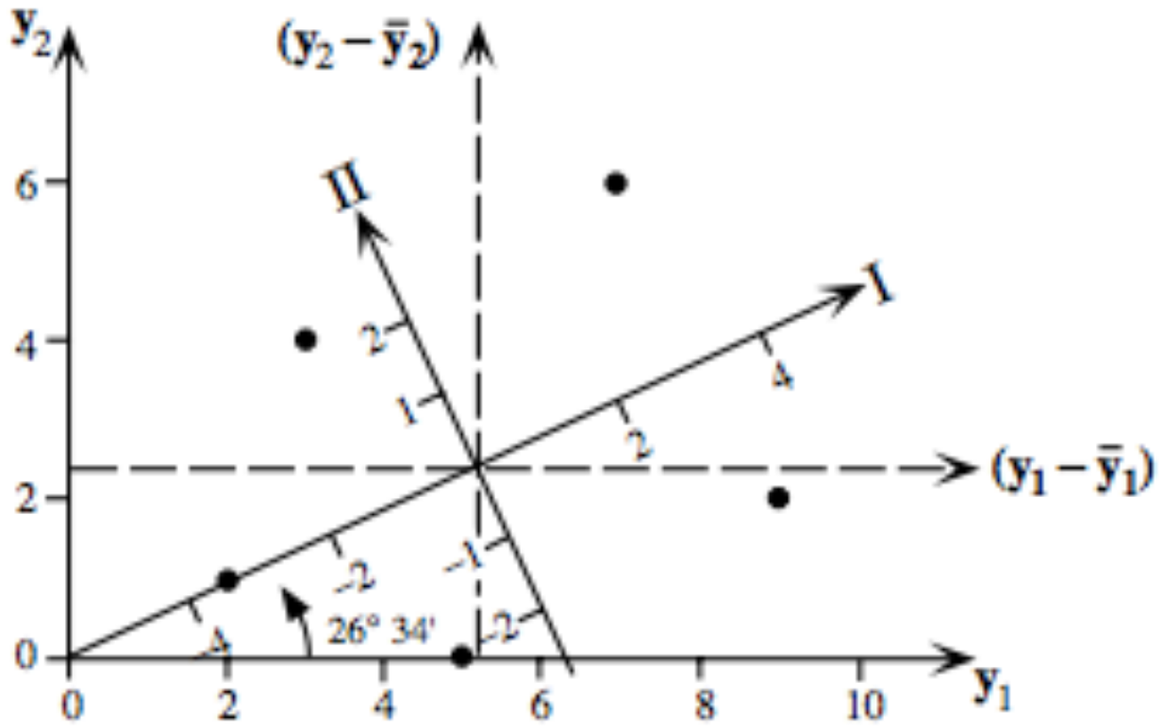
- It is easy to see which axes are the ones that explain the most variation in 2-dimensional space

- It is **much harder to do this (visually) when N is large** (multi-dimensional data)

Rotation in N-dimensional space

2 variables

1	1	1	0	0
0	1	2	1	2



Rotation in N-dimensional space

7 variables

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

?

Rotation in N-dimensional space

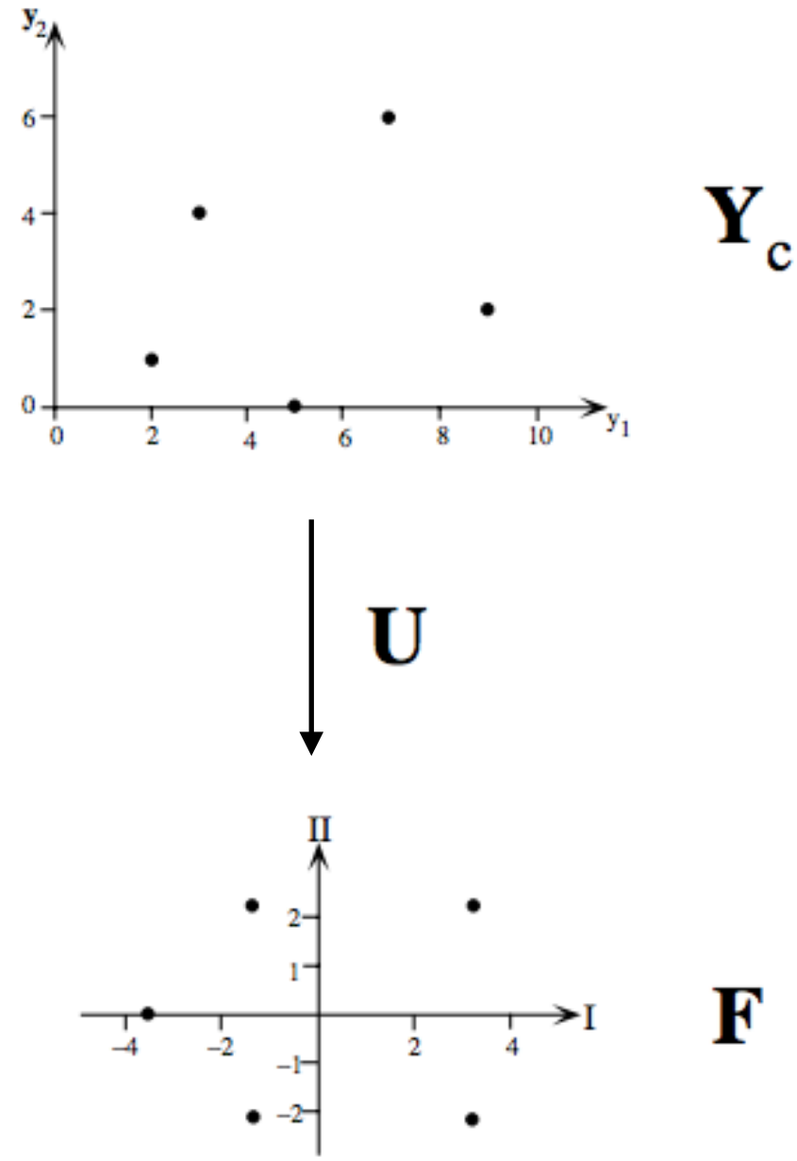
$$\mathbf{F} = \mathbf{Y}_c\,\mathbf{U}$$

$$\mathbf{F} = \begin{bmatrix} -3.2 & -1.6 \\ -2.2 & 1.4 \\ -0.2 & -2.6 \\ 1.8 & 3.4 \\ 3.8 & -0.6 \end{bmatrix} \begin{bmatrix} 0.8944 & -0.4472 \\ 0.4472 & 0.8944 \end{bmatrix} = \begin{bmatrix} -3.578 & 0 \\ -1.342 & 2.236 \\ -1.342 & -2.236 \\ 3.130 & 2.236 \\ 3.130 & -2.236 \end{bmatrix}$$

original data

new axes (eigenvectors of covariance matrix)

data in PCA space



Proportion of variance explained

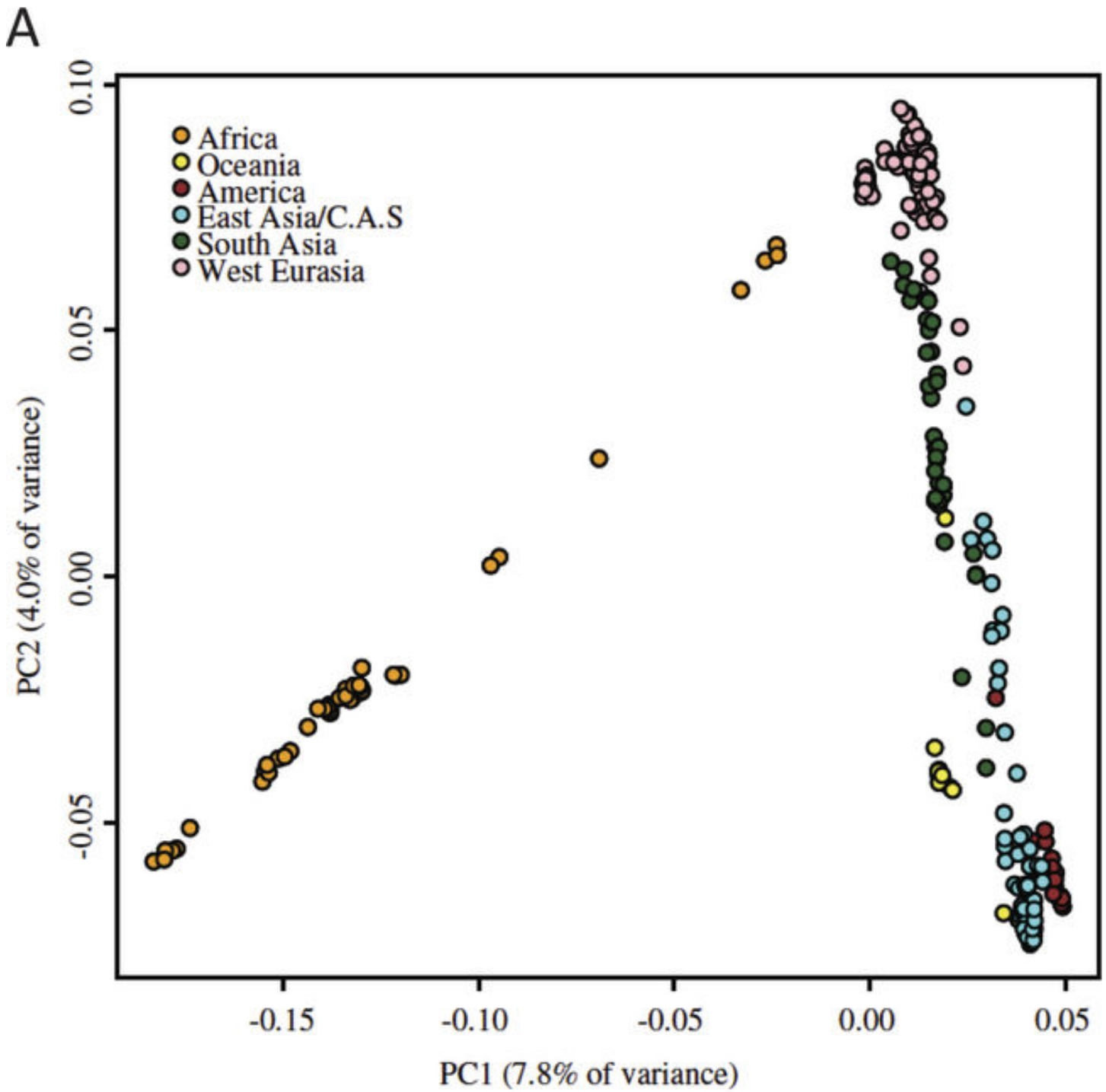
- In a PCA, each eigenvector has a corresponding eigenvalue

- The **largest eigenvalues** correspond to the **eigenvectors that explain the most variation**

- Percent of variance explained by eigenvector k = eigenvalue k / (sum of all eigenvalues)

- **Largest eigenvalue -> largest axis of variation**

PCA of worldwide human genomes



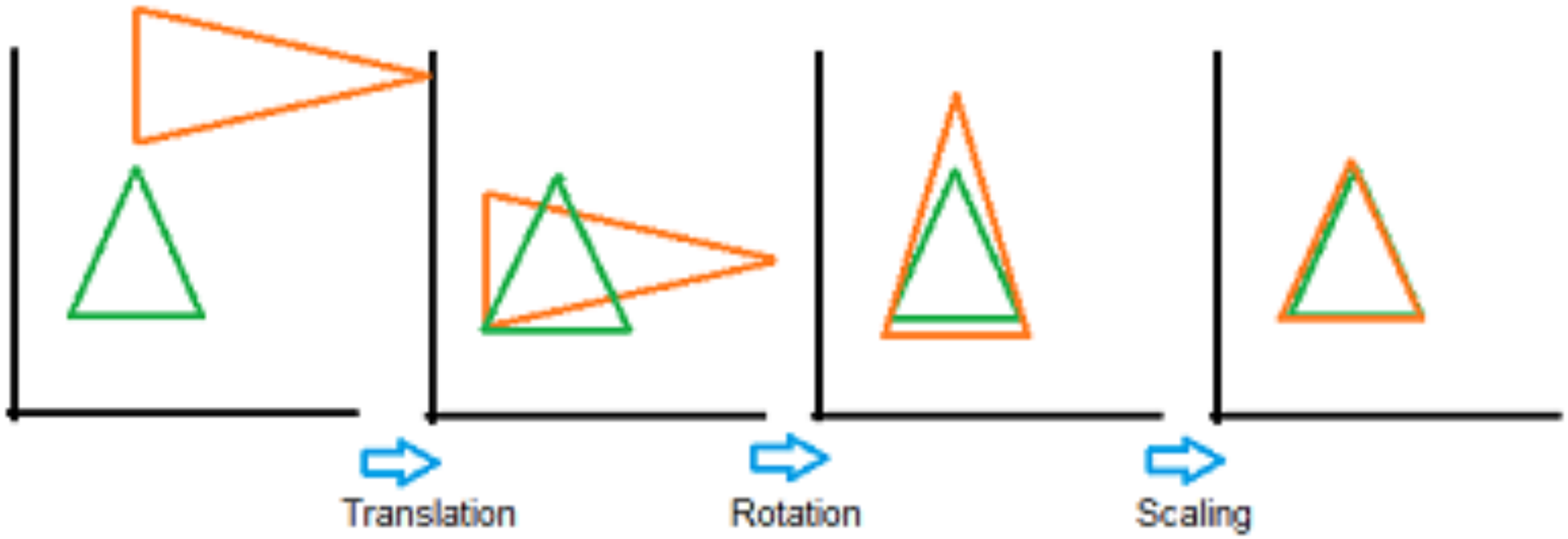
Dealing with missing data: Procrustes transformation

- 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 means very few sites with overlap in coverage across all of them

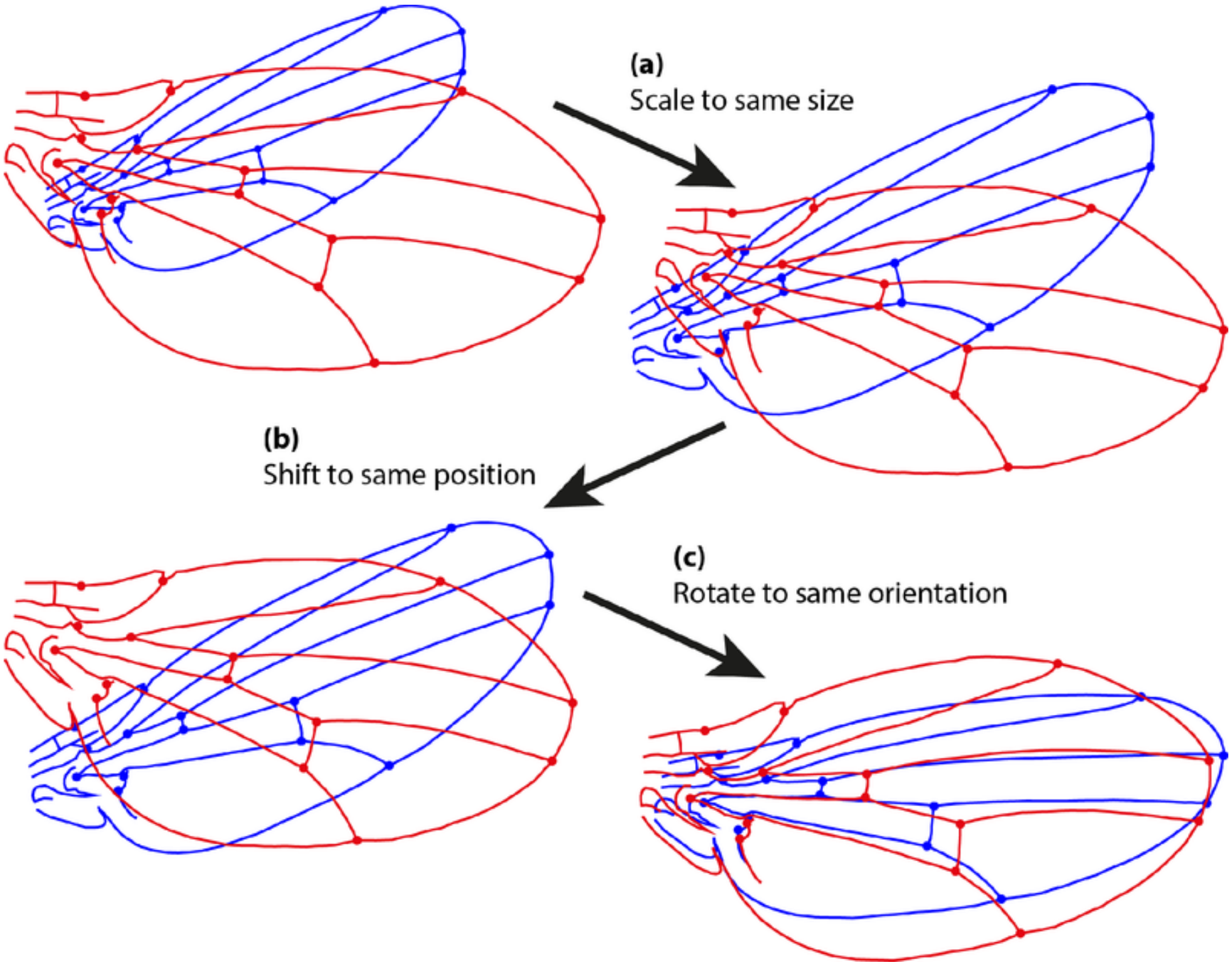
Dealing with missing data: Procrustes transformation

- 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

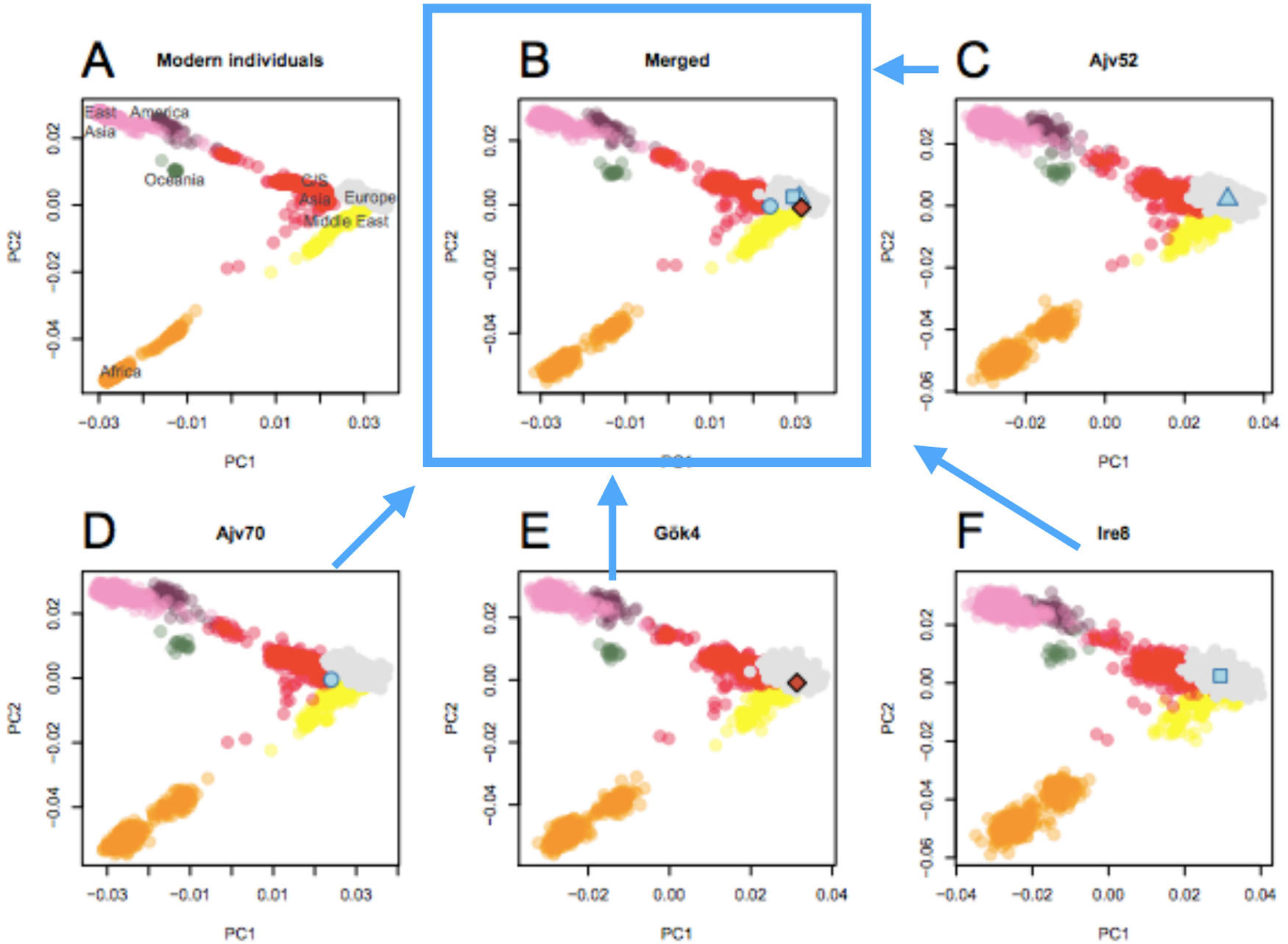
Shape-preserving Procrustes transformation



Shape-preserving Procrustes transformation

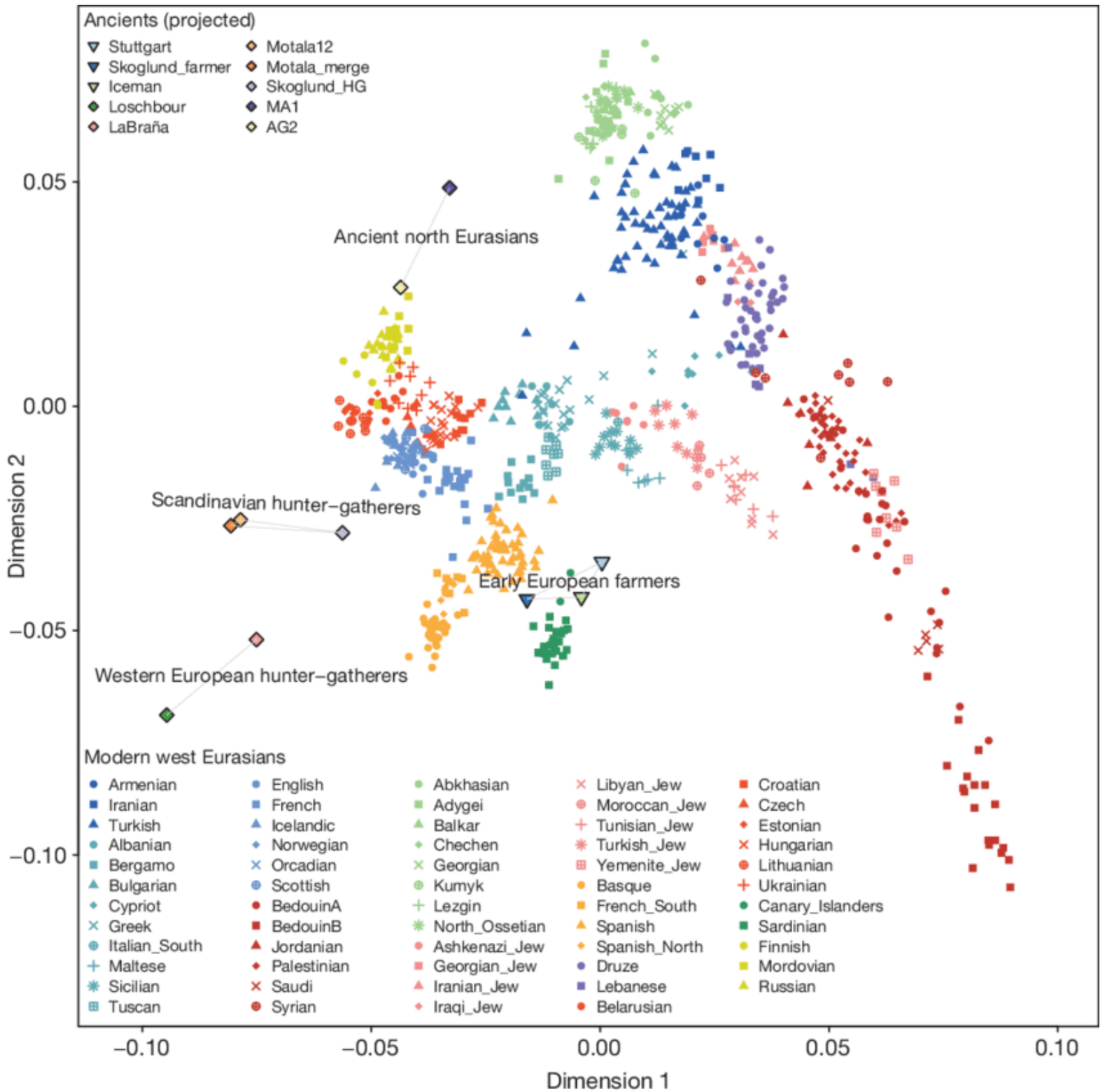


Use a Procrustes transformation using a **high-coverage reference PCA**



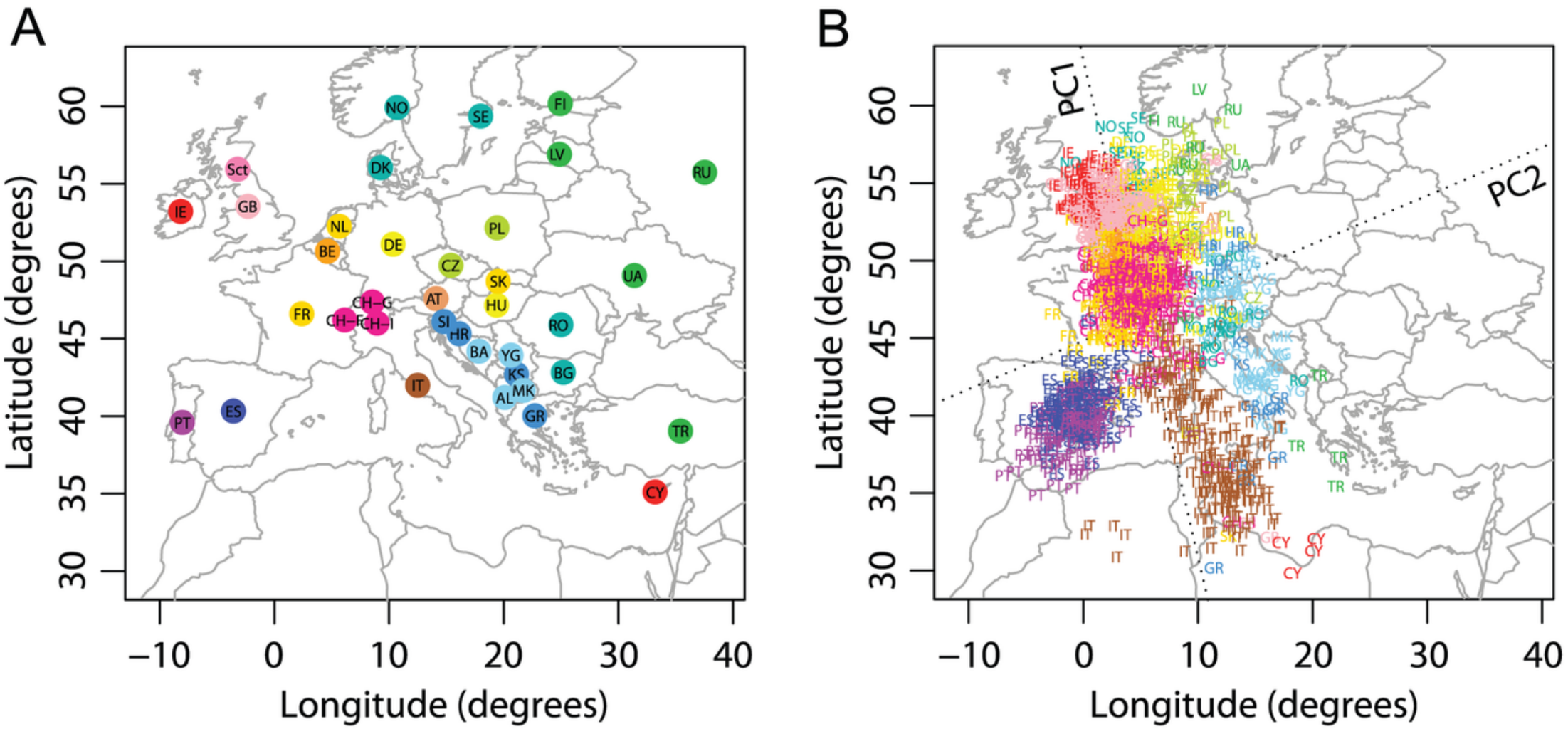
Skoglund et al. 2012

PCA of (projected) ancient genomes



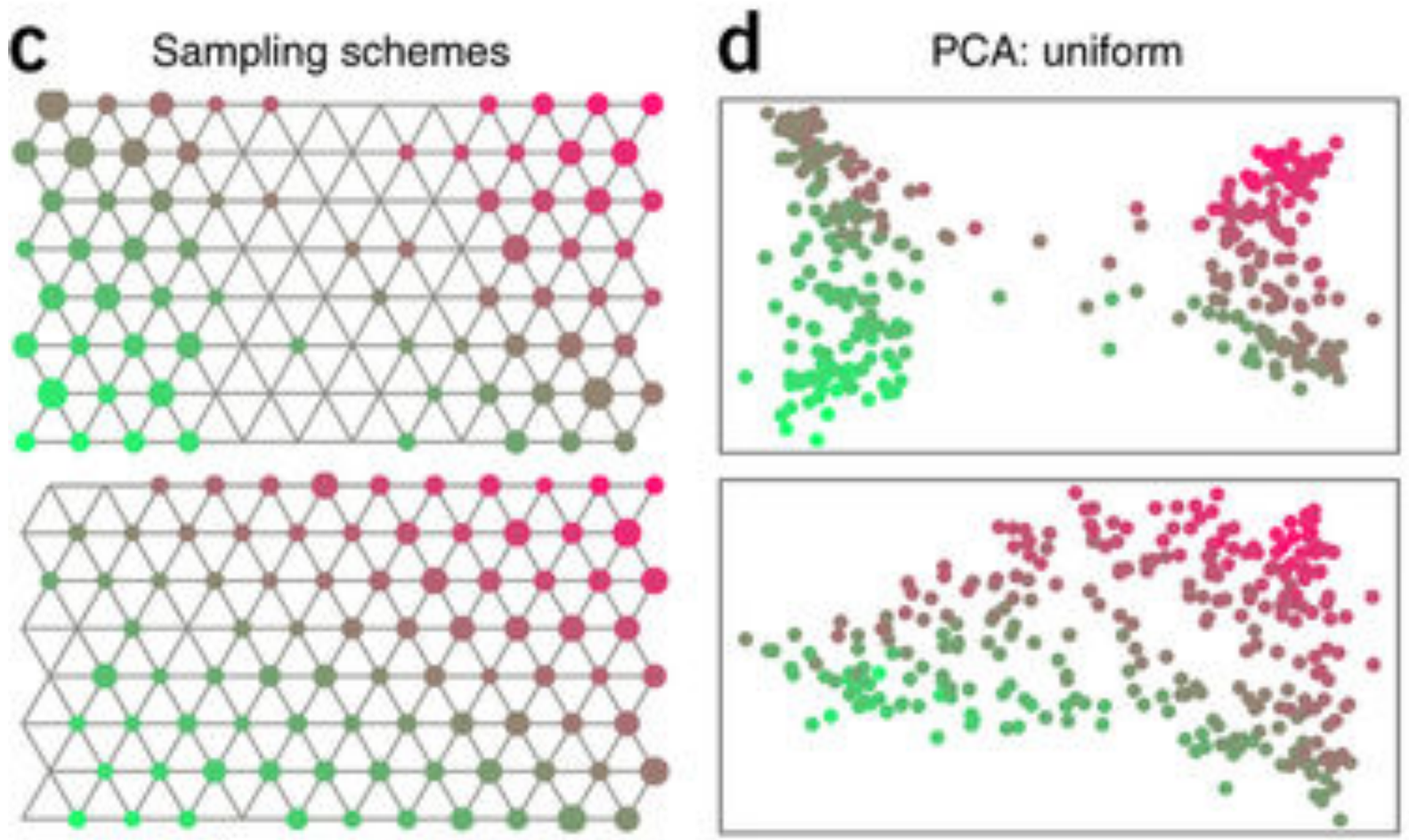
Lazaridis et al. 2014

Present-day PCA recovers signals of “isolation-by-distance”

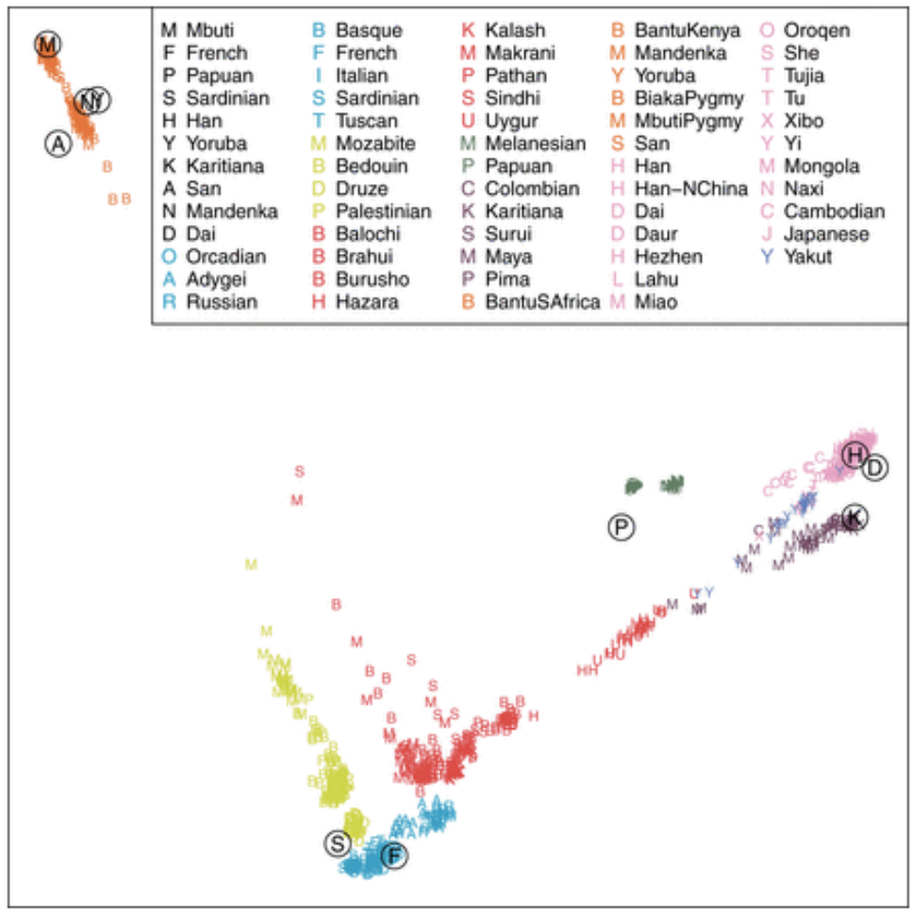


Novembre et al. 2008

Sampling scheme can be misleading



Other methods for dimensionality reduction



MDS

UMAP

Malaspinas et al. 2014
Diaz-Papkovich et al. 2018

Exercises

- https://github.com/FerRacimo/Archaeomics/blob/master/PopGen1_PCAandAdmixture.md

Today

- Introduction to population genetics
- Exploratory vs. hypothesis-driven analyses
- PCA
- **Useful datasets for human paleogenomics**
- Latent mixed-membership models (“Structure”)

The 1000 Genomes Project

- 2,504 present-day genomes from 26 global populations (low coverage shotgun + exome sequencing + microarray genotyping + imputation)

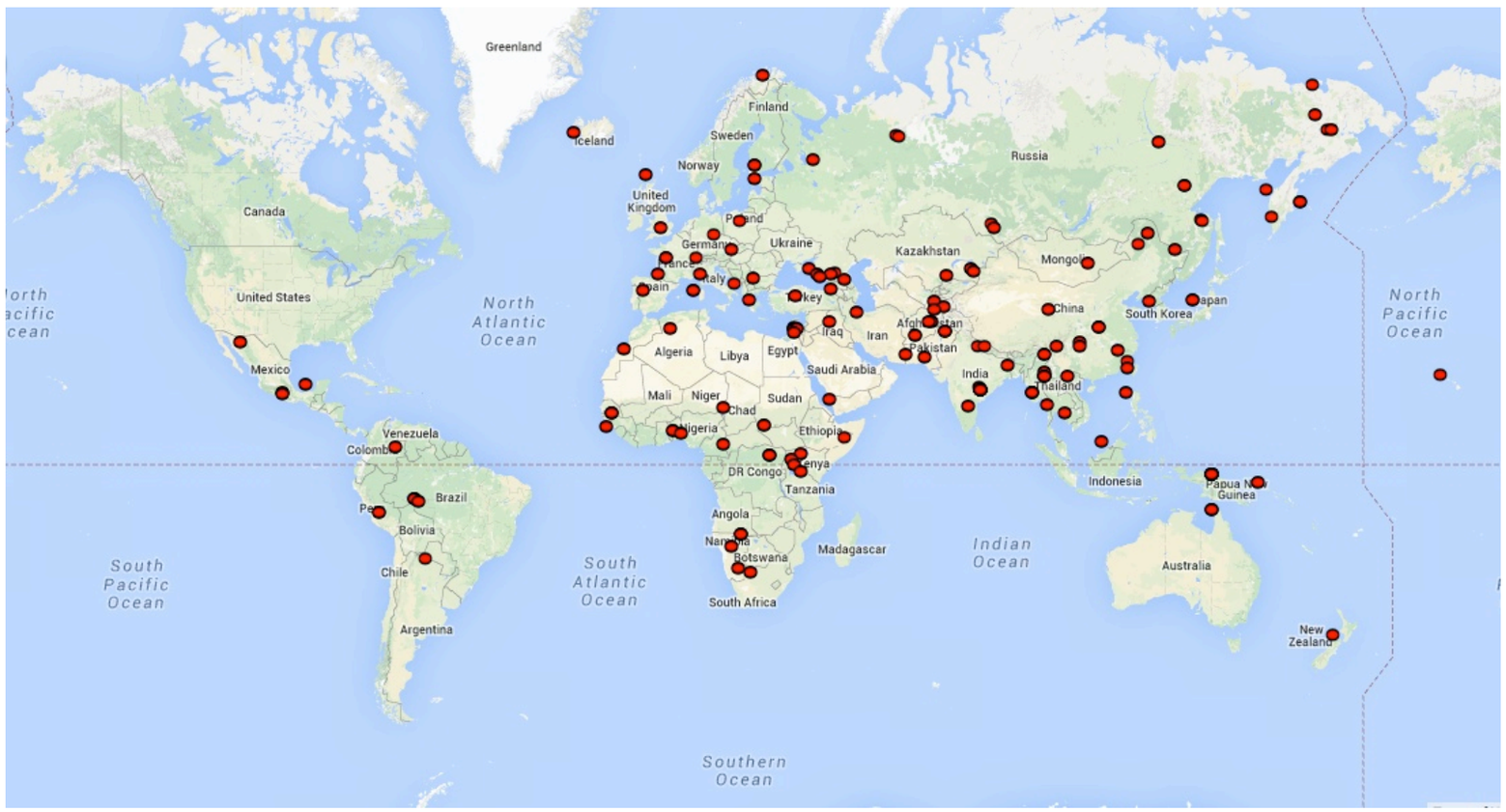
- http://www.internationalgenome.org/

The 1000 Genomes Project Consortium 2015

Simons Genome Diversity Panel (SGDP)

- 250 present-day genomes (with at least 30X coverage) from 125 diverse populations

- https://www.simonsfoundation.org/simons-genome-diversity-project/



Mallick et al. 2017

Human Genome Diversity Panel (HGDP)

- 929 present-day genomes (now at high coverage!) from 54 diverse populations

- ftp://ngs.sanger.ac.uk/production/hgdp

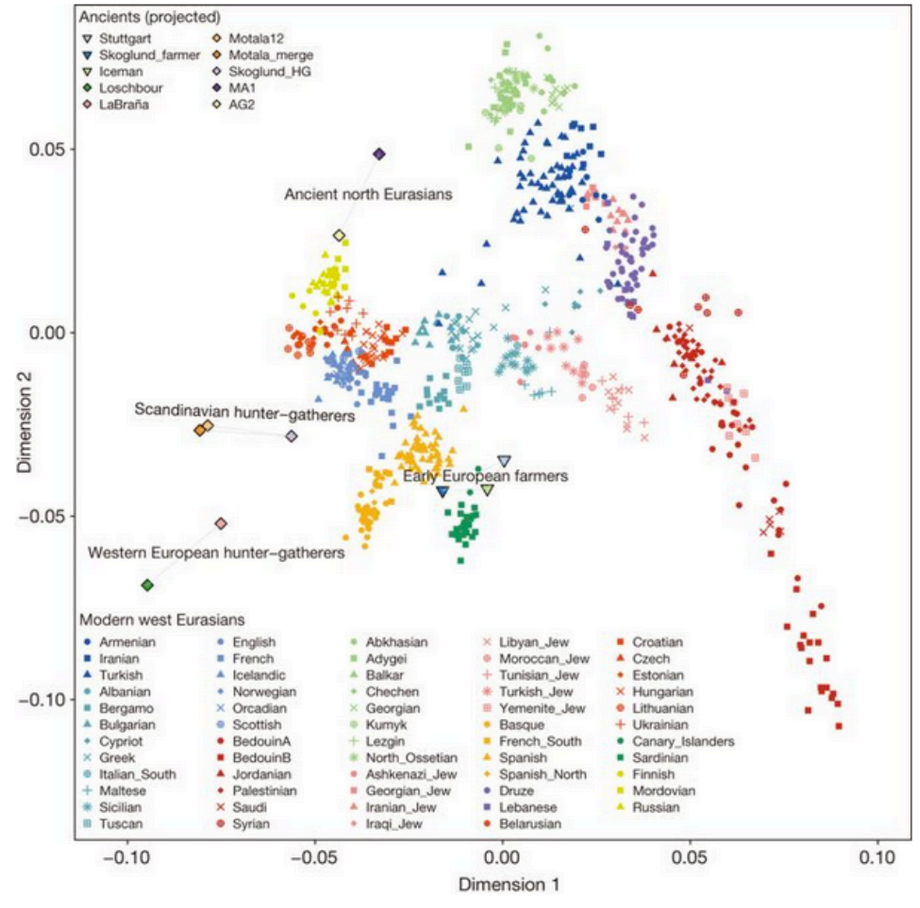
Bergstrom et al. 2019

Affymetrix Human Origins SNP Panel

- SNP capture of 2,918 individuals from diverse present-day populations across the world

- Typed at ~600,000 SNPs across the genome

- https://reich.hms.harvard.edu/datasets



Lazaridis et al. 2014, 2016, Skoglund et al. 2016, Patterson et al. 2012

Genome Aggregation Database (gnomAD)

- Includes 125,748 exomes and 15,708 whole genomes from selected populations around the world

- Note: not processed uniformly.

gnomAD

genome aggregation database

Search by gene, region, or variant

Examples - Gene: PCSK9, Variant: 1-55516888-G-GA

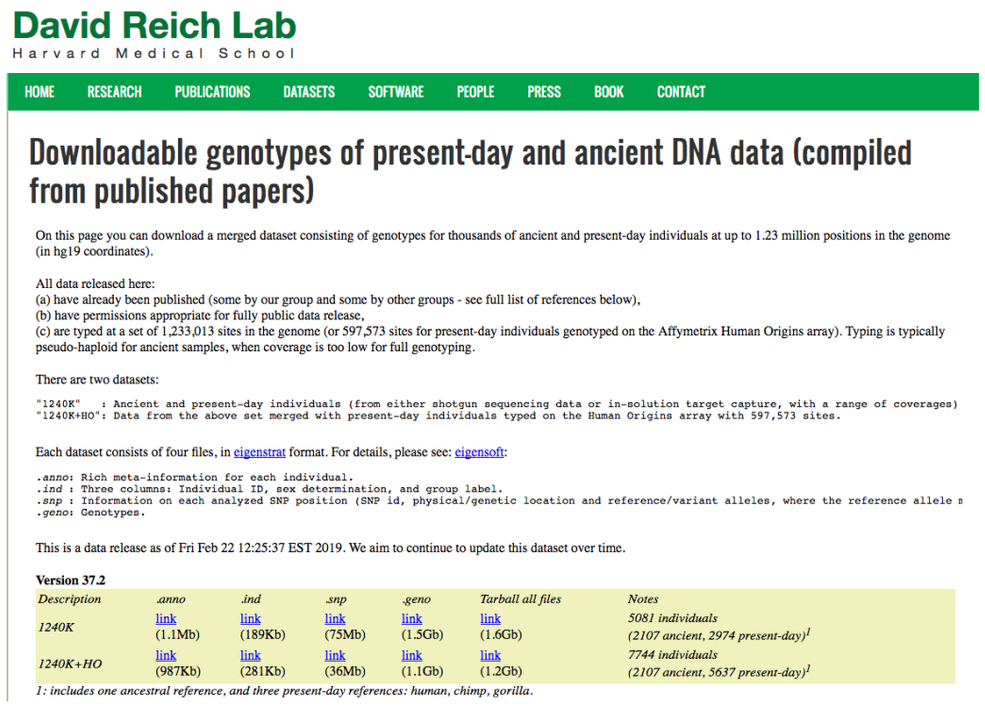
The Genome Aggregation Database (gnomAD) is a resource developed by an international coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects, and making summary data available for the wider scientific community.

The data set provided on this website spans 125,748 exome sequences and 15,708 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies. The gnomAD Principal Investigators and groups that have contributed data to the current release are listed here.

All data here are released for the benefit of the wider biomedical community, without restriction on use - see the terms of use here. Sign up for our mailing list for future release announcements here.

Reich lab ancient genome database

- All ancient genomes published to date
- Only sites typed at the Human Origins SNP array (600,000-SNP version or 1.24 million SNP version)
- https://reich.hms.harvard.edu/downloadable-genotypes-present-day-and-ancient-dna-data-compiled-published-papers



David Reich Lab
Harvard Medical School

HOME RESEARCH PUBLICATIONS DATASETS SOFTWARE PEOPLE PRESS BOOK CONTACT

Downloadable genotypes of present-day and ancient DNA data (compiled from published papers)

On this page you can download a merged dataset consisting of genotypes for thousands of ancient and present-day individuals at up to 1.23 million positions in the genome (in hg19 coordinates).

All data released here:
(a) have already been published (some by our group and some by other groups - see full list of references below),
(b) have permissions appropriate for fully public data release,
(c) are typed at a set of 1,233,013 sites in the genome (or 597,573 sites for present-day individuals genotyped on the Affymetrix Human Origins array). Typing is typically pseudo-haploid for ancient samples, when coverage is too low for full genotyping.

There are two datasets:

```
"1240K"   : Ancient and present-day individuals (from either shotgun sequencing data or in-solution target capture, with a range of coverages)
"1240K+HO": Data from the above set merged with present-day individuals typed on the Human Origins array with 597,573 sites.
```

Each dataset consists of four files, in eigenstrat format. For details, please see: eigensoft:

```
.anno: Rich meta-information for each individual.
.ind : Three columns: Individual ID, sex determination, and group label.
.snp : Information on each analyzed SNP position (SNP id, physical/genetic location and reference/variant alleles, where the reference allele m
.geno: Genotypes.
```

This is a data release as of Fri Feb 22 12:25:37 EST 2019. We aim to continue to update this dataset over time.

Version 37.2

Description	*.anno*	*.ind*	*.snp*	*.geno*	*Tarball all files*	*Notes*
1240K	link (1.1Mb)	link (189Kb)	link (75Mb)	link (1.5Gb)	link (1.6Gb)	*5081 individuals (2107 ancient, 2974 present-day)[1]*
1240K+HO	link (987Kb)	link (281Kb)	link (36Mb)	link (1.1Gb)	link (1.2Gb)	*7744 individuals (2107 ancient, 5637 present-day)[1]*

1: includes one ancestral reference, and three present-day references: human, chimp, gorilla.

Different types of data

	Targeted SNP capture
Data characteristics	
Genomic coverage	Targeted SNPs and alleles
Typical enrichment range	45–13,000 (6)
Best use scenario	Low endogenous DNA; low/medium complexity
Analyses characteristics[a]	
Diploid genotyping	Possible with high coverage, potential for allelic bias
Ascertainment bias	Specific to targeted SNP panel
Suitability for merging with reference variant sets	Only variants overlapping with capture panel
Basic population structure and admixture analyses	Yes
Demographic inference	Methods not sensitive to ascertainment bias and/or allowing for correction
Rare variant analyses	Only variants overlapping with capture panel
Recovery of host-associated pathogens	Only if targeted with capture probes

Cappellini et al. 2018

Different types of data

	Targeted SNP capture	Whole-genome capture
Data characteristics		
Genomic coverage	Targeted SNPs and alleles	Genome-wide
Typical enrichment range	45–13,000 (6)	2–13× (79)
Best use scenario	Low endogenous DNA; low/medium complexity	Low endogenous DNA; high complexity
Analyses characteristics[a]		
Diploid genotyping	Possible with high coverage, potential for allelic bias	Possible with high coverage, potential for allelic bias
Ascertainment bias	Specific to targeted SNP panel	None
Suitability for merging with reference variant sets	Only variants overlapping with capture panel	All variants
Basic population structure and admixture analyses	Yes	Yes
Demographic inference	Methods not sensitive to ascertainment bias and/or allowing for correction	Yes
Rare variant analyses	Only variants overlapping with capture panel	Yes
Recovery of host-associated pathogens	Only if targeted with capture probes	Only if targeted with capture probes

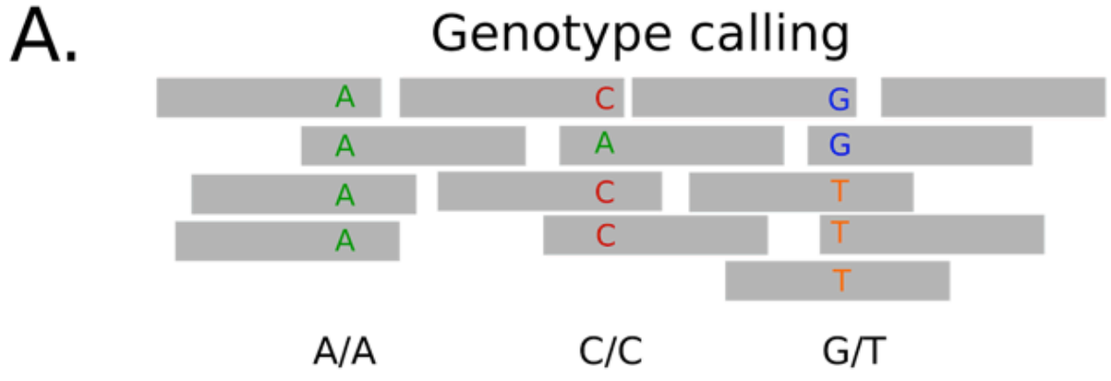
Cappellini et al. 2018

Different types of data

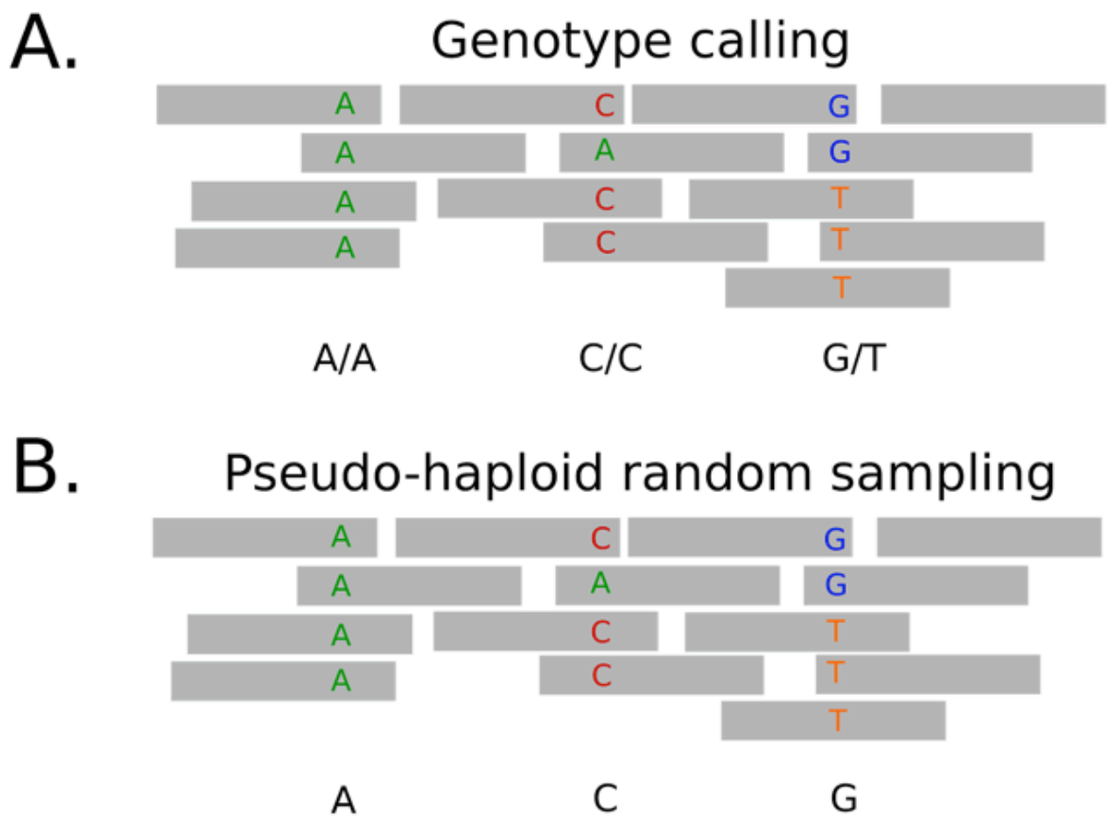
	Targeted SNP capture	Whole-genome capture	Whole-genome shotgun
Data characteristics			
Genomic coverage	Targeted SNPs and alleles	Genome-wide	Genome-wide
Typical enrichment range	45–13,000 (6)	2–13× (79)	None
Best use scenario	Low endogenous DNA; low/medium complexity	Low endogenous DNA; high complexity	Medium/high endogenous DNA; high complexity
Analyses characteristics[a]			
Diploid genotyping	Possible with high coverage, potential for allelic bias	Possible with high coverage, potential for allelic bias	Possible with high coverage
Ascertainment bias	Specific to targeted SNP panel	None	None
Suitability for merging with reference variant sets	Only variants overlapping with capture panel	All variants	All variants
Basic population structure and admixture analyses	Yes	Yes	Yes
Demographic inference	Methods not sensitive to ascertainment bias and/or allowing for correction	Yes	Yes
Rare variant analyses	Only variants overlapping with capture panel	Yes	Yes
Recovery of host-associated pathogens	Only if targeted with capture probes	Only if targeted with capture probes	Yes

Cappellini et al. 2018

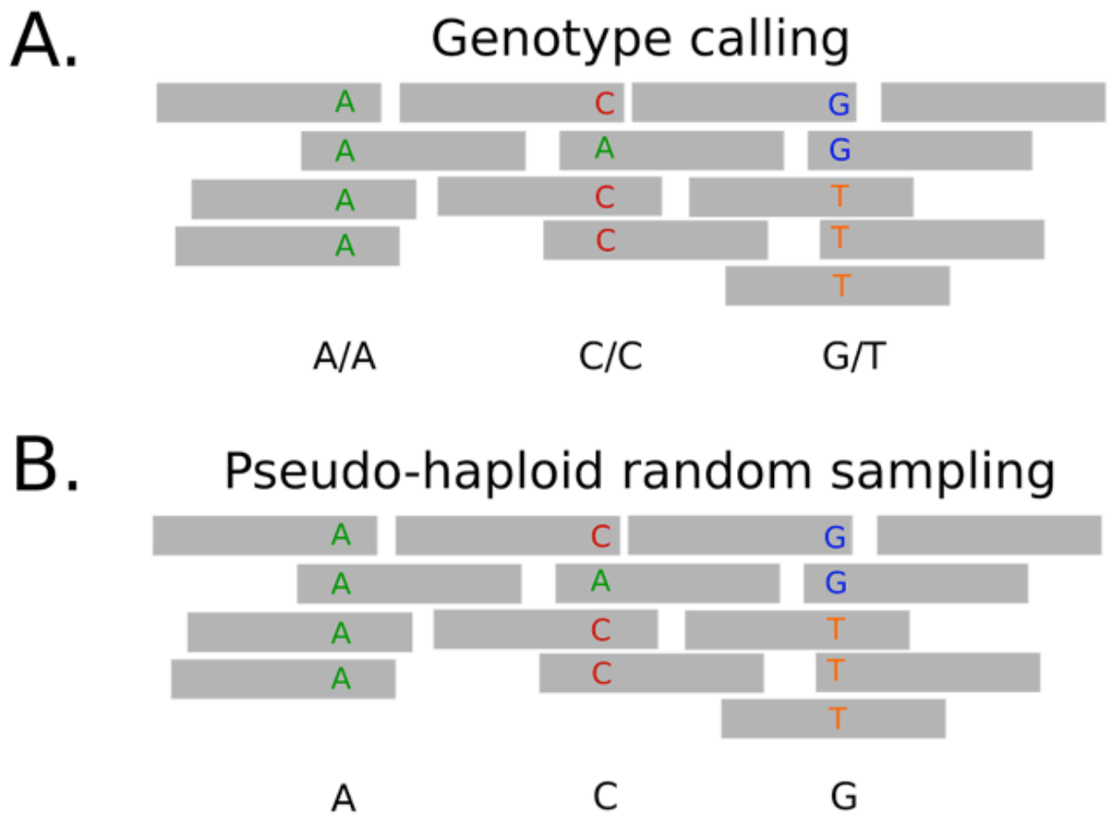
Different types of data **processing**



Different types of data **processing**



Different types of data **processing**



Today

- Introduction to population genetics
- Exploratory vs. hypothesis-driven analyses
- PCA
- Useful datasets for human paleogenomics
- **Latent mixed-membership models (“Structure”)**

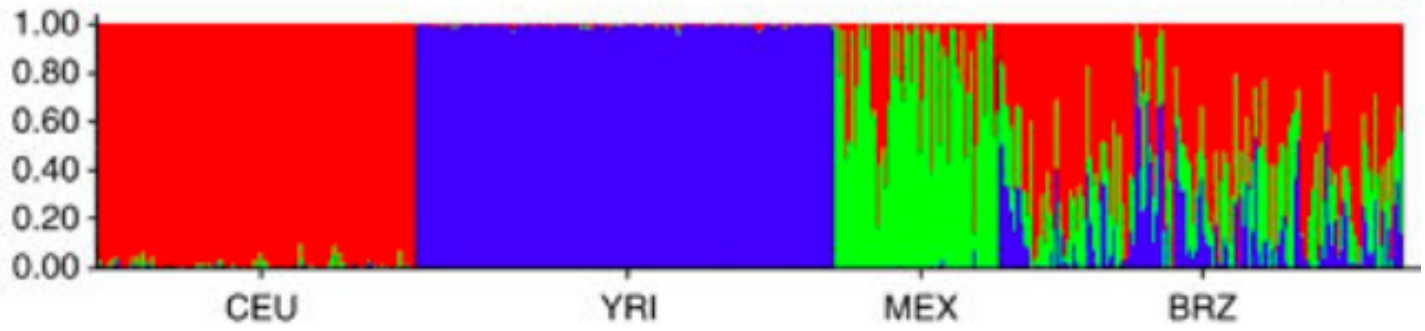
Population structure

Fernando Racimo

Copenhagen, August 2019

Questions

- Is there population structure in a population?
- Can we identify subpopulation clusters of shared ancestry?
- Are individuals best modeled as mixtures of ancestral populations?
- How much admixture was passed on from each population?



Objectives

- Learn something about the pasty genetic history of a population under study
- Learn something about ourselves

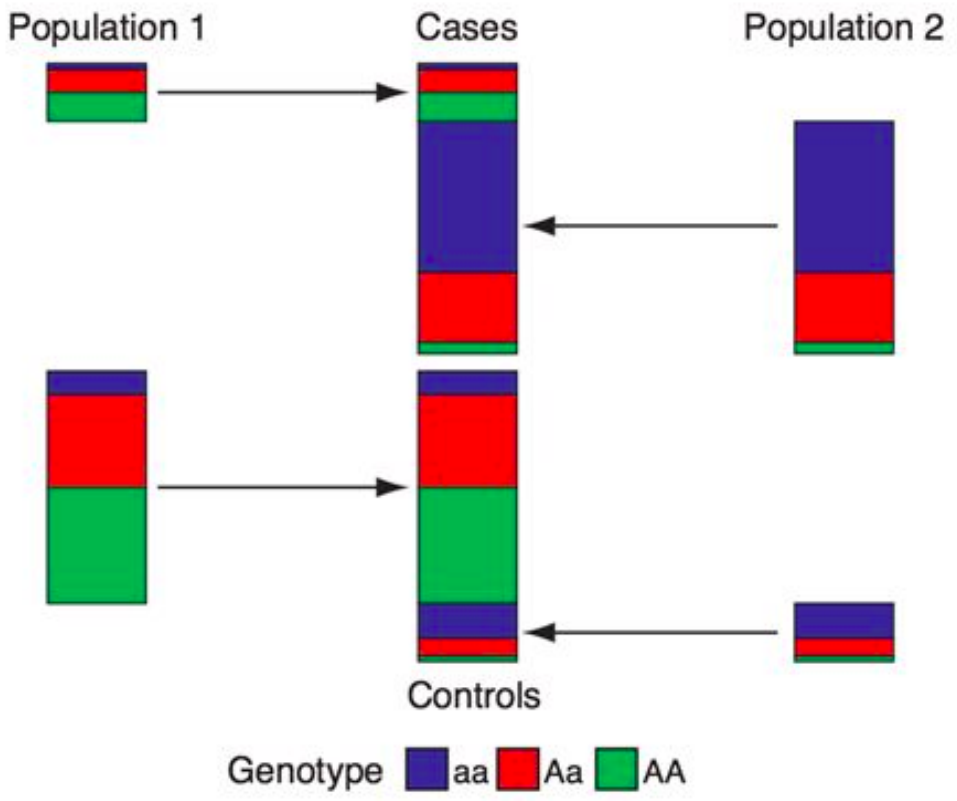
Fernando Racimo	100%
European	99.1% ›
Middle Eastern & North African	0.4%
South Asian	0.3%
Sub-Saharan African	0.1%
Unassigned	< 0.1%

See all 31 tested populations

European	99.1%
Southern European	68.4%
Iberian	38.5%
Italian	22.1%
Balkan	0.6%
Broadly Southern European	7.2%
Northwestern European	25.2%
French & German	13.4%
British & Irish	3.0%
Finnish	0.1%
Broadly Northwestern European	8.7%

Objectives

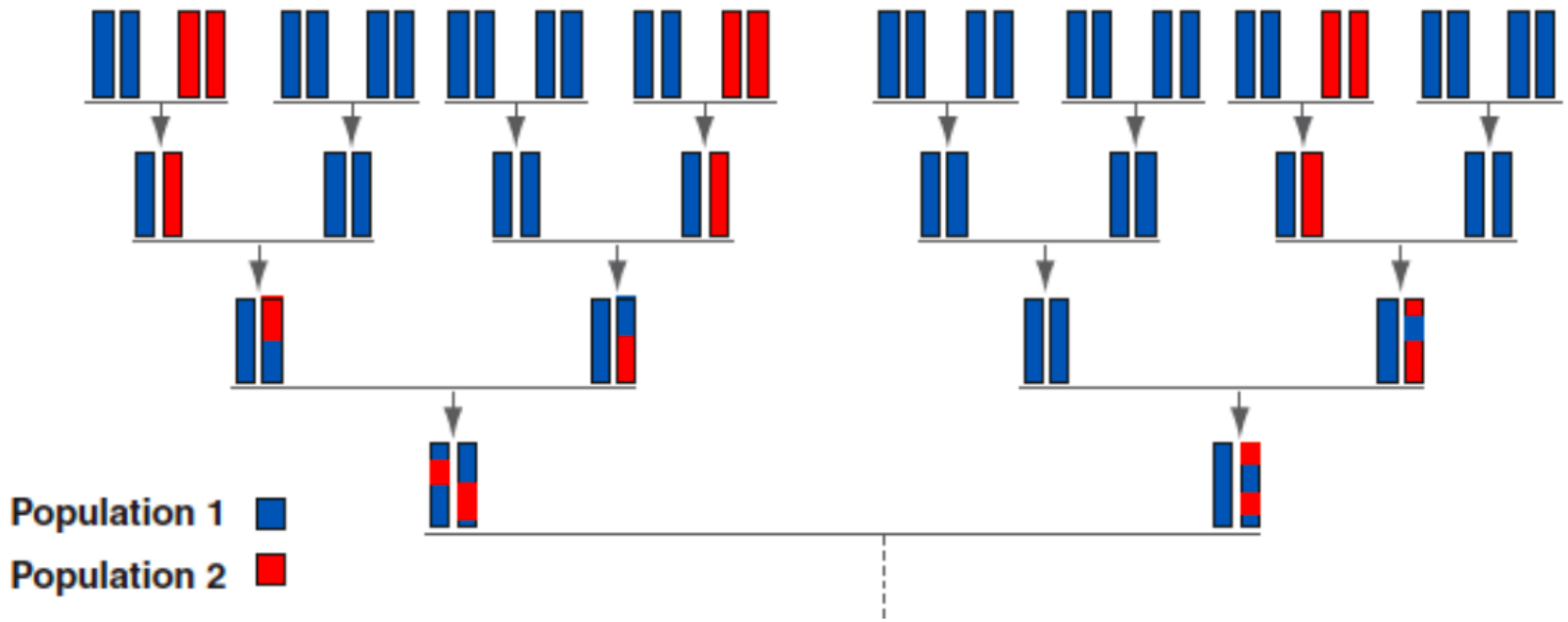
- Correct for population stratification in downstream analyses (e.g. GWAS)



The "Structure" model

- The original model was first proposed by Pritchard et al. (2000)
- **Assumption 1**: each individual can be modeled as a mixture of one or more ancestral **"source populations"**
- **Assumption 2**: each locus is independent
- The proportion of genetic matrerial from each source in each individual is called the **"admixture proportion"**
- **Problem 1**: we don't know the identity and number of these source populations
- **Problem 2**: we don't know the admixture proportions
- **Objective**: find best-fitting sources and their proportions

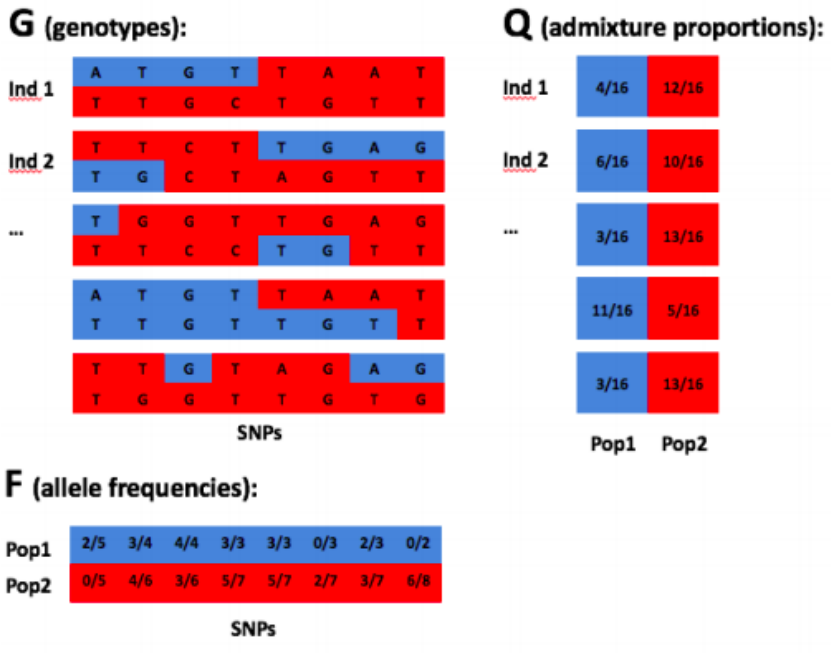
The "Structure" model



[0]Darvasi and Shifman 2005

The "Structure" model

- Known: genotypes (G)
- Unknown:
  - admixture proportions (Q)
  - allele frequencies in source populations (F)
- Need to estimate Q and F, given that we know G.
- Objective: Maximize likelihood function: $P[G|Q,F]$



[0] Ida Moltke pers. comm.

Likelihood function: one individual, one locus

- A locus j can have two possible alleles: A and a

Likelihood function: one individual, one locus

- A locus j can have two possible alleles: A and a
- We'll code a genotype as the number of alleles a: 0, 1 or 2.

Likelihood function: one individual, one locus

- A locus j can have two possible alleles: A and a
- We'll code a genotype as the number of alleles a: 0, 1 or 2.
- Let $G_{i,j}$ be the (diploid) genotype of locus j in individual i.

Likelihood function: one individual, one locus

- A locus j can have two possible alleles: A and a
- We'll code a genotype as the number of alleles a: 0, 1 or 2.
- Let $G_{i,j}$ be the (diploid) genotype of locus j in individual i.
- Let $F^j = (f^{j,1}, f^{j,2}, ..., f^{j,K})$ be the allele frequencies of allele a in the K "source populations"

Likelihood function: one individual, one locus

- A locus j can have two possible alleles: A and a
- We'll code a genotype as the number of alleles a: 0, 1 or 2.
- Let $G_{i,j}$ be the (diploid) genotype of locus j in individual i.
- Let $F^j = (f^{j,1}, f^{j,2}, ..., f^{j,K})$ be the allele frequencies of allele a in the K "source populations"
- Let $Q^i = (q^{i,1}, q^{i,2}, ..., q^{i,K})$ be the K admixture proportions of individual i.

Likelihood function: one individual, one locus

- A locus j can have two possible alleles: A and a
- We'll code a genotype as the number of alleles a: 0, 1 or 2.
- Let $G_{i,j}$ be the (diploid) genotype of locus j in individual i.
- Let $F^j = (f^{j,1}, f^{j,2}, ..., f^{j,K})$ be the allele frequencies of allele a in the K "source populations"
- Let $Q^i = (q^{i,1}, q^{i,2}, ..., q^{i,K})$ be the K admixture proportions of individual i.
- Then, for one of the allele copies in locus j in idividual i, the probability of randomly sampling the a allele is:
$P[a|Q, F] = q^{i,1} f^{j,1} + q^{i,2} f^{j,2} + ... + q^{i,K} f^{j,K} = h^{i,j}$

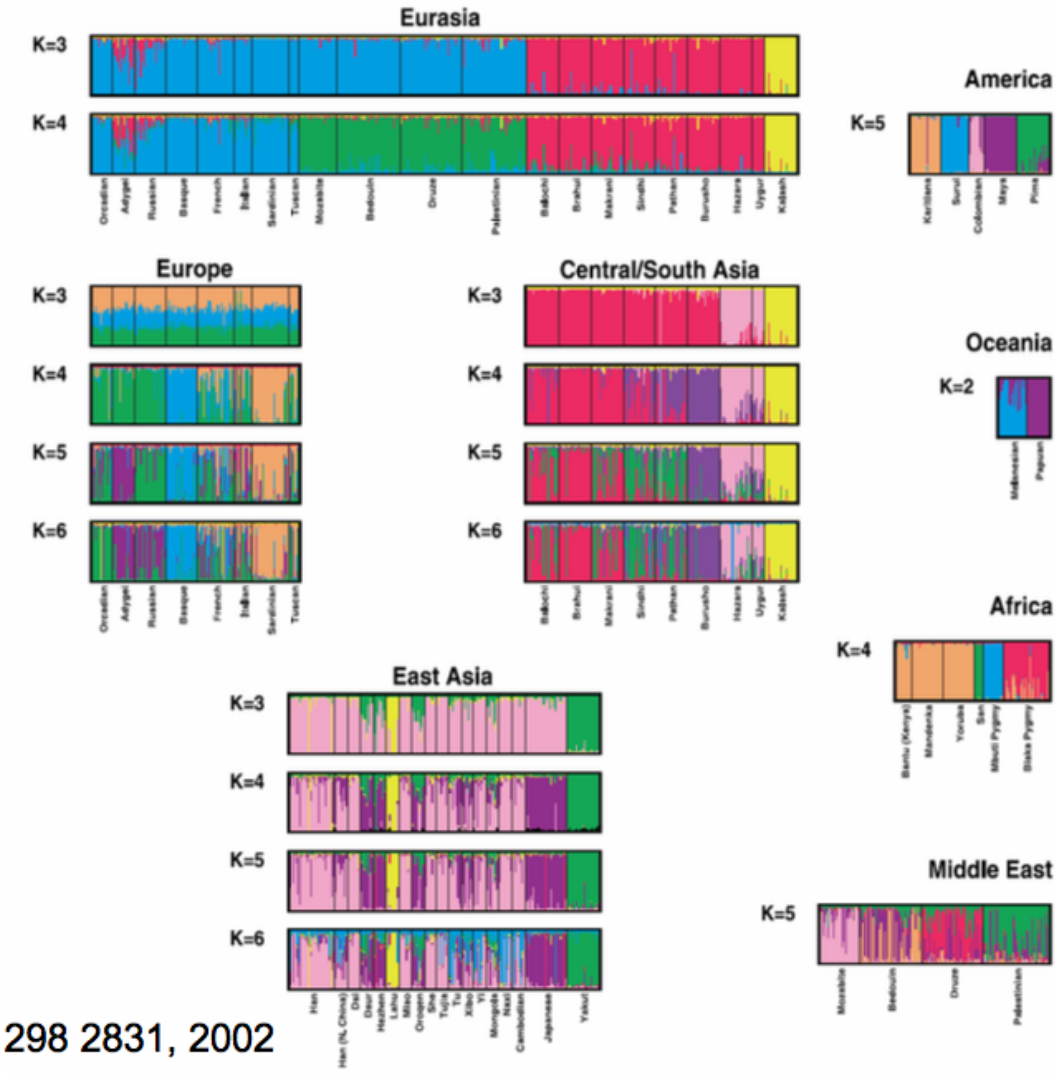
Likelihood function: one individual, one locus

- A locus j can have two possible alleles: A and a
- We'll code a genotype as the number of alleles a: 0, 1 or 2.
- Let $G_{i,j}$ be the (diploid) genotype of locus j in individual i.
- Let $F^j = (f^{j,1}, f^{j,2}, ..., f^{j,K})$ be the allele frequencies of allele a in the K "source populations"
- Let $Q^i = (q^{i,1}, q^{i,2}, ..., q^{i,K})$ be the K admixture proportions of individual i.
- Then, for one of the allele copies in locus j in idividual i, the probability of randomly sampling the a allele is:
  $P[a|Q, F] = q^{i,1} f^{j,1} + q^{i,2} f^{j,2} + ... + q^{i,K} f^{j,K} = h^{i,j}$
- Assuming Hardy-Weinberg equilibrium:
  $P[G_{ij} = 2|Q^i, F^j] = (h^{ij})^2$
  $P[G_{ij} = 1|Q^i, F^j] = 2(h^{ij})(1 - h^{ij})$
  $P[G_{ij} = 0|Q^i, F^j] = (1 - h^{ij})^2$

Likelihood function: N individuals, M loci

- Assuming loci are independent and individuals are unrelated...
- $P[G|Q,F] = \prod_i^N \prod_j^M P[G_{ij}|Q^i, Fj]$
- Q is a matrix of admixture proportions for each of the N individuals: $Q = (Q^1, Q^2, ..., Q^N)$
- F is a matrix of ancestral allele frequencies for each of the M loci: $F = (F^1, F^2, ..., F^M)$
- Structure-lke methods try to find the **parameters Q and F that maximize** $P[G|Q,F]$

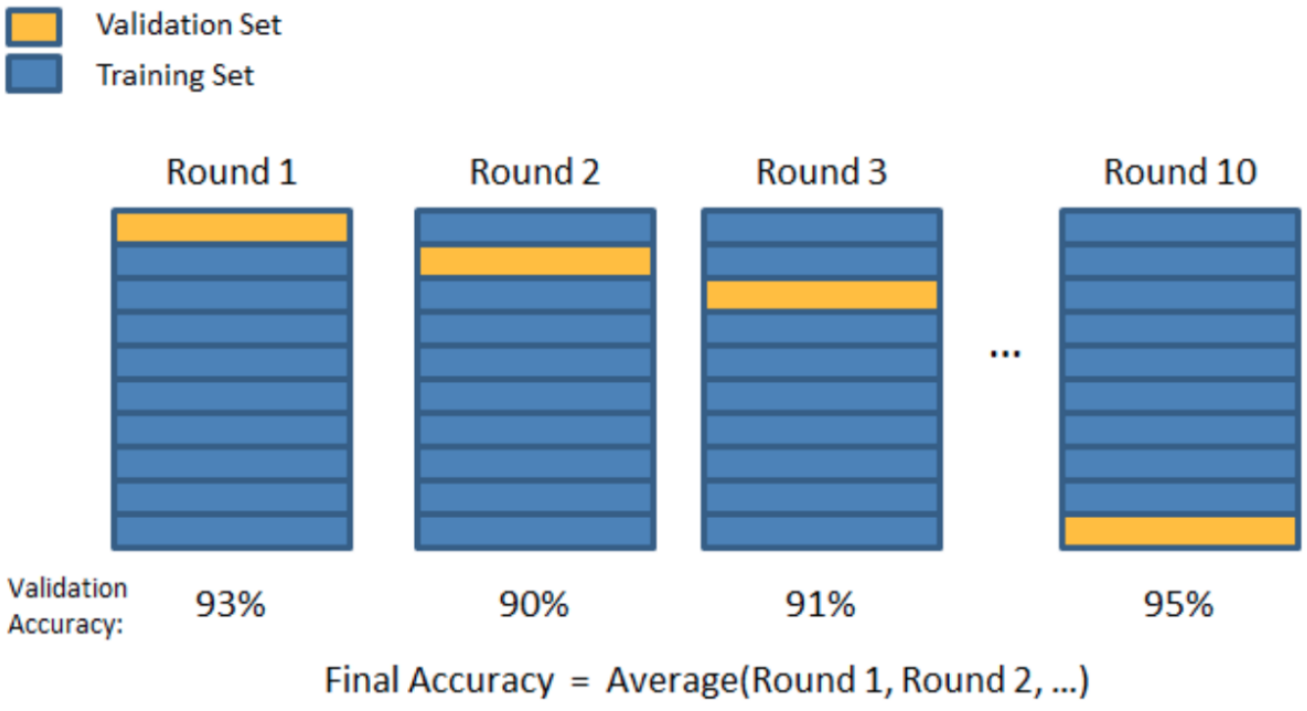
Structure model applied to human populations



Science 298 2831, 2002

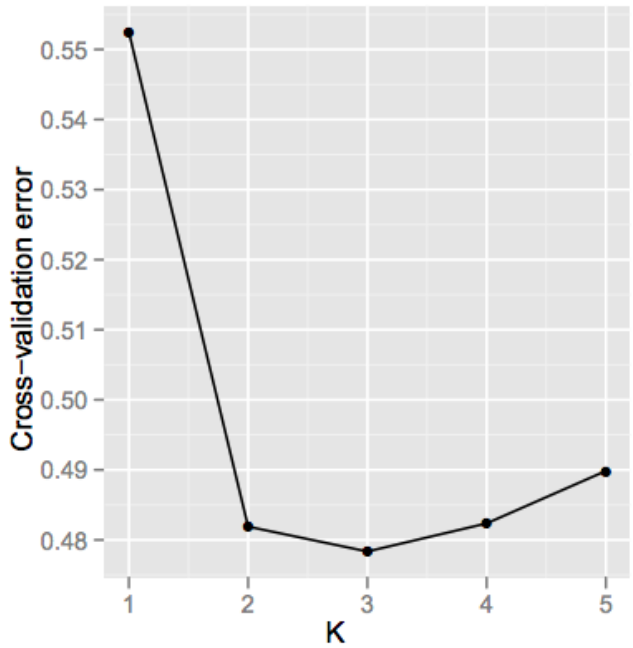
Choosing K

- We can use cross-validation to find a value of K that does not over-fit
- We leave some genotypes out and predict them based on their estimated ancestries
- **Important**: well-fitting parameter $\neq$ biologically meaningful parameter



Choosing K

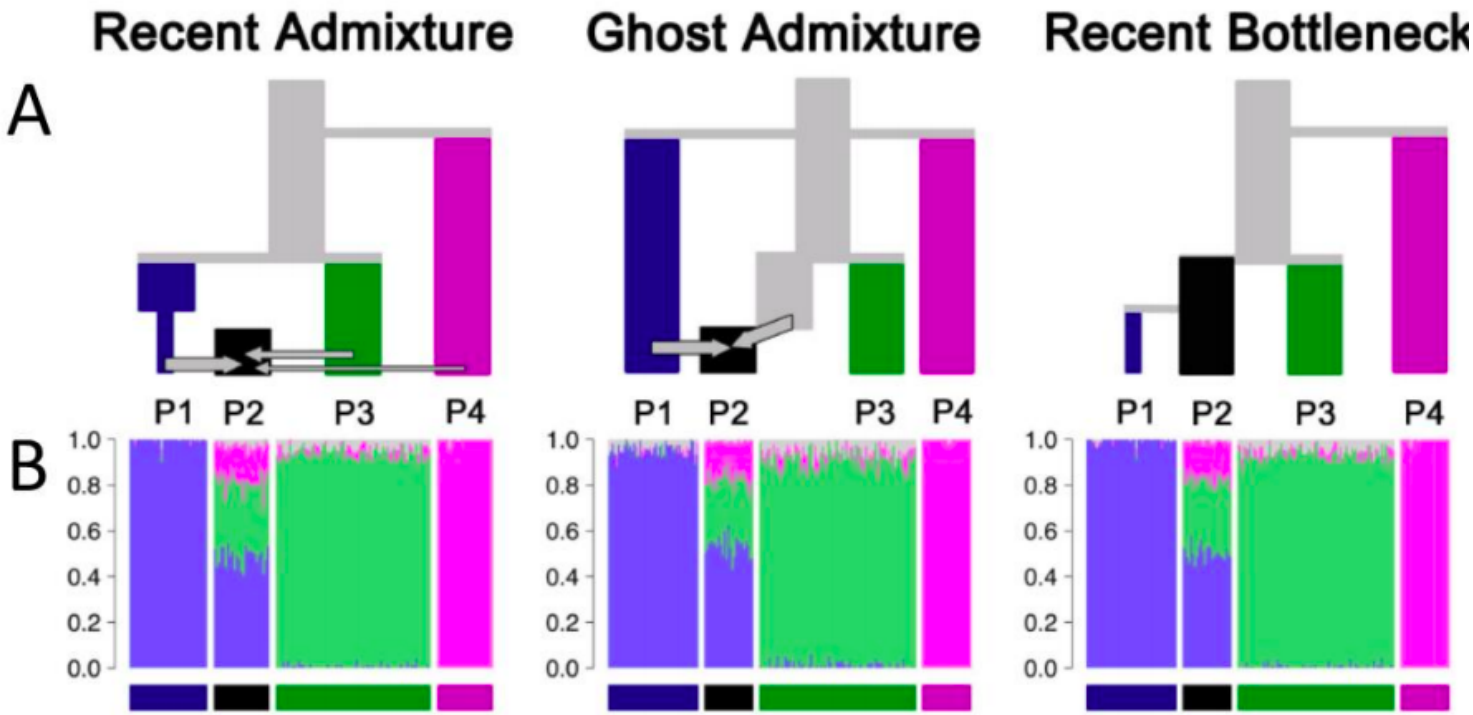
- We can use cross-validation to find a value of K that does not over-fit
- We leave some genotypes out and predict them based on their estimated ancestries
- **Important**: well-fitting parameter $\neq$ biologically meaningful parameter



[0] Alexander et al. 2011

Over-interpreting Structure results

- Structure does not necessarily pick up admixture events![1]
- "Source populations" need not be real populations that ever existed!
- A population that is highly drifted will be assigned its own cluster at high enough K

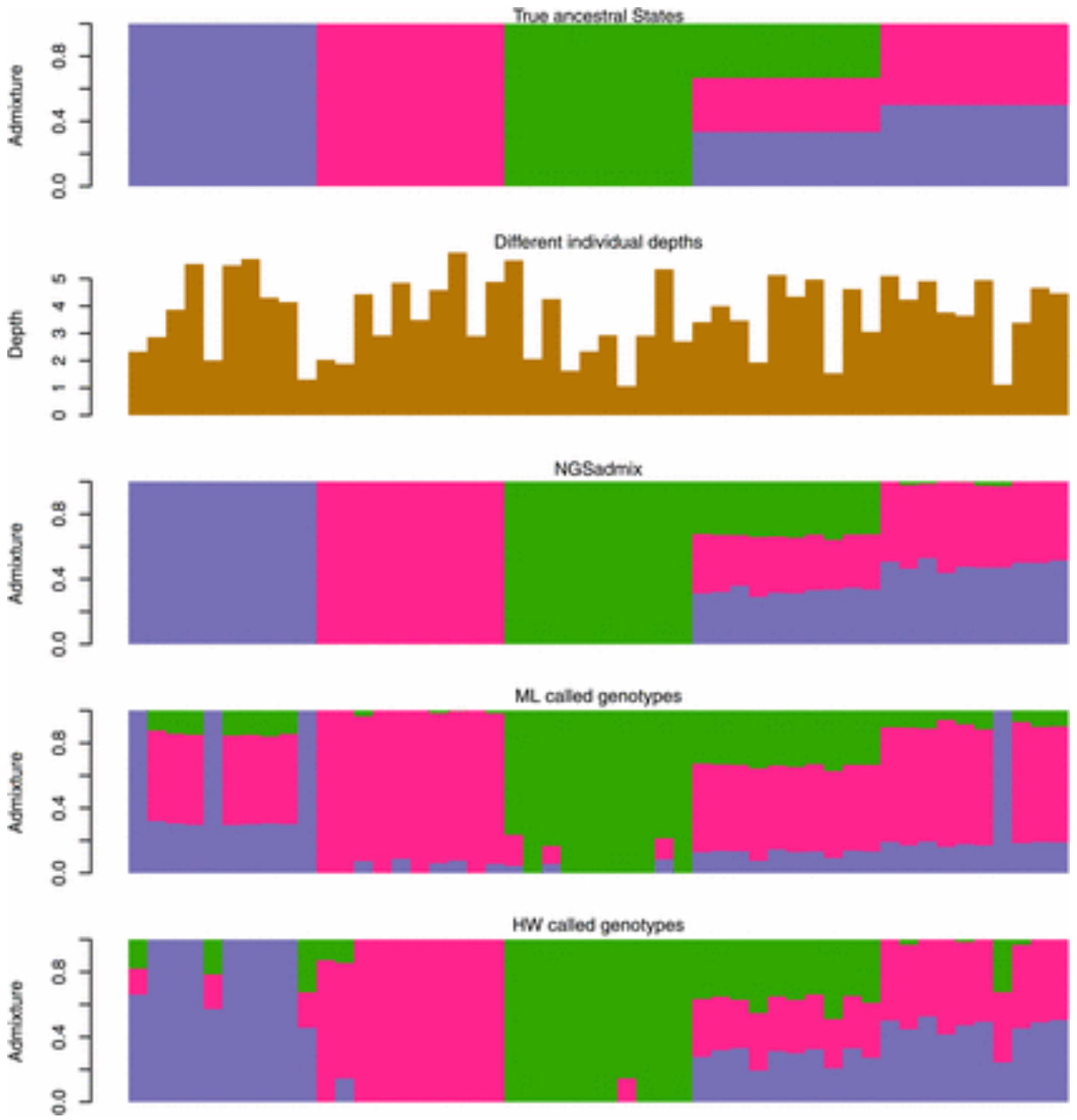


[1] Falush et al. 2016

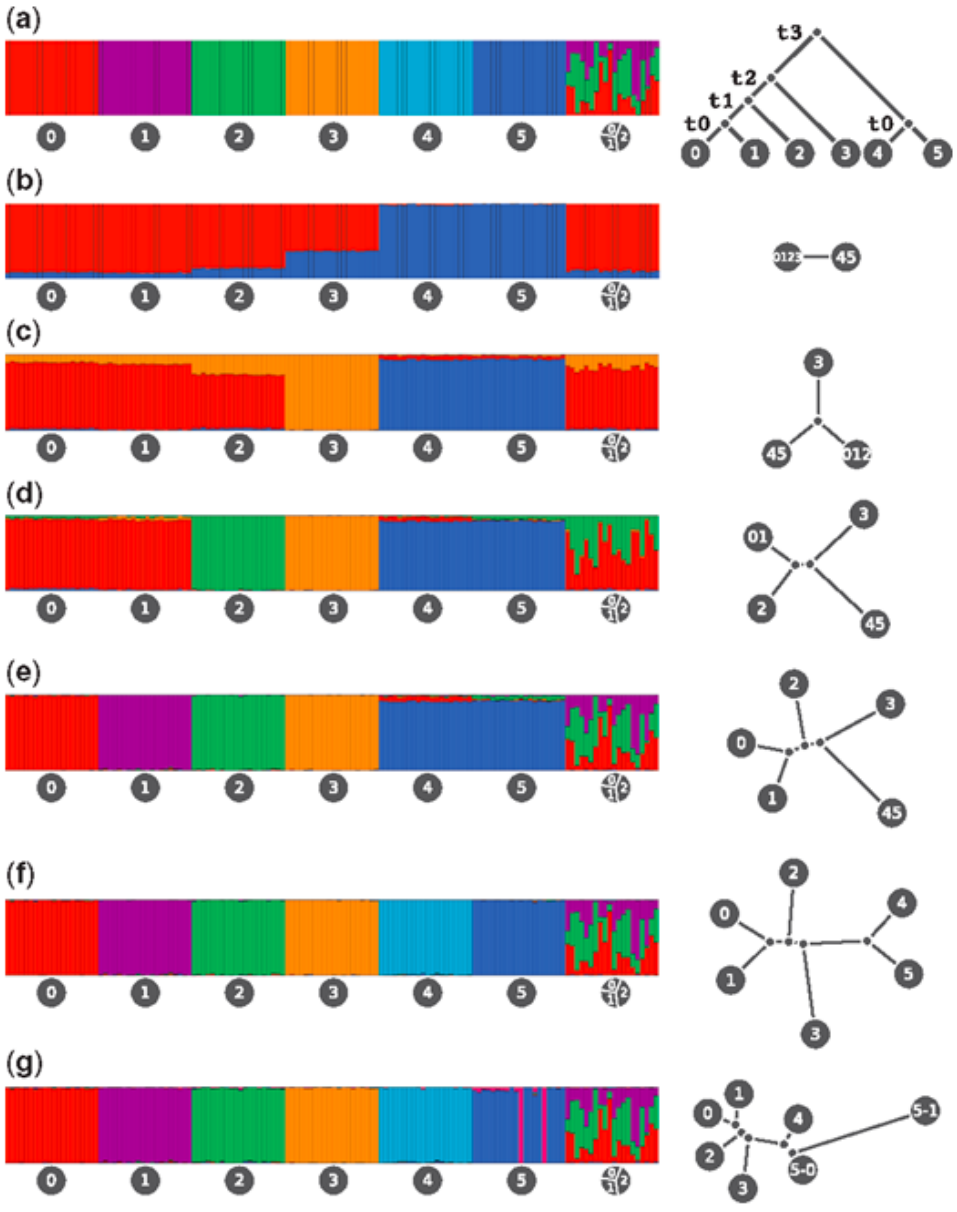
Variations on a theme...

- Structure (Pritchard et al. 2000): original model; uses Bayesian priors to obtain posterior estimates of Q and F
- Admixture (Alexander et al. 2011): faster than Structure; uses a maximum likelihood model rather than a Bayesian model; uses cross-validation to choose K
- fastStructure (Raj et al. 2014): faster than Structure; uses variational inference to choose K; can detect weak structure
- ngsAdmix (Skotte et al. 2013): can work with genotype likelihoods; better for low coverage data
- Ohana (Cheng et al. 2016): uses Gaussian approximation to model drift in each ancestry component; can detect selection by testing for local deviations from genome-wide model

ngsAdmix (Skotte et al. 2013)



Ohana (Cheng et al. 2017, 2019)



Ohana with selection (Cheng et al. 2019)

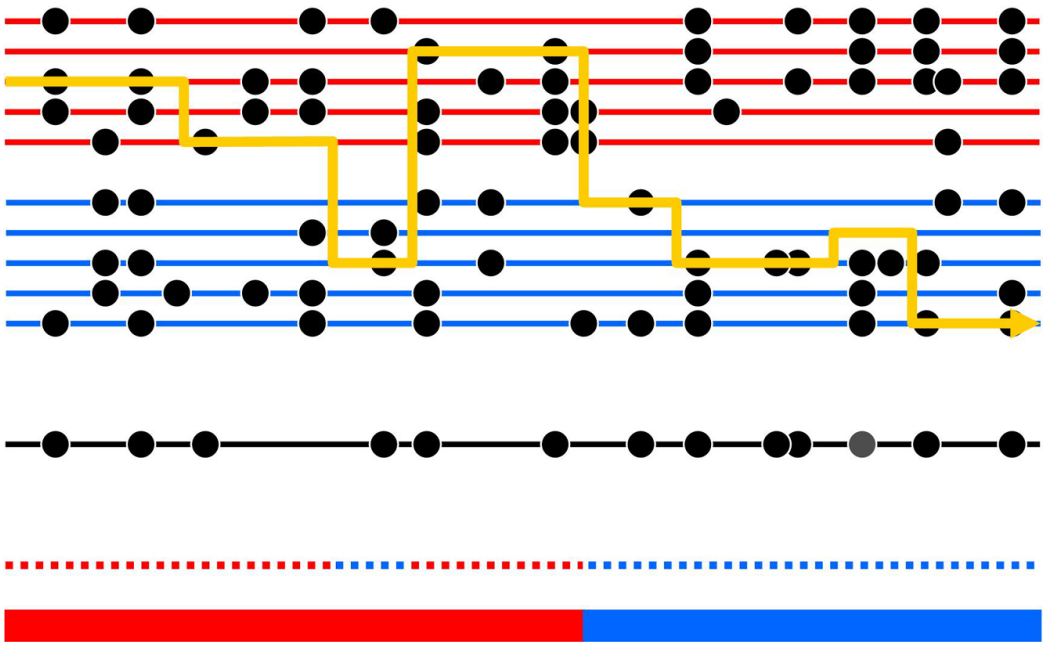
- Ohana can also be used to detect loci under positive selection
- The user can specify which ancestral population(s) one is interested in scanning for selection

SNP vs. Haplotype information

- All these methods ingnore the spatial distribution of SNPs along the genome
- They require LD pruning: a lot of SNPs sit in the same haplotype and have redundant information
- Advantage: can model each SNP independently (simple model)
- Disadvantage: we ignore haplotype information

Chromosome painting

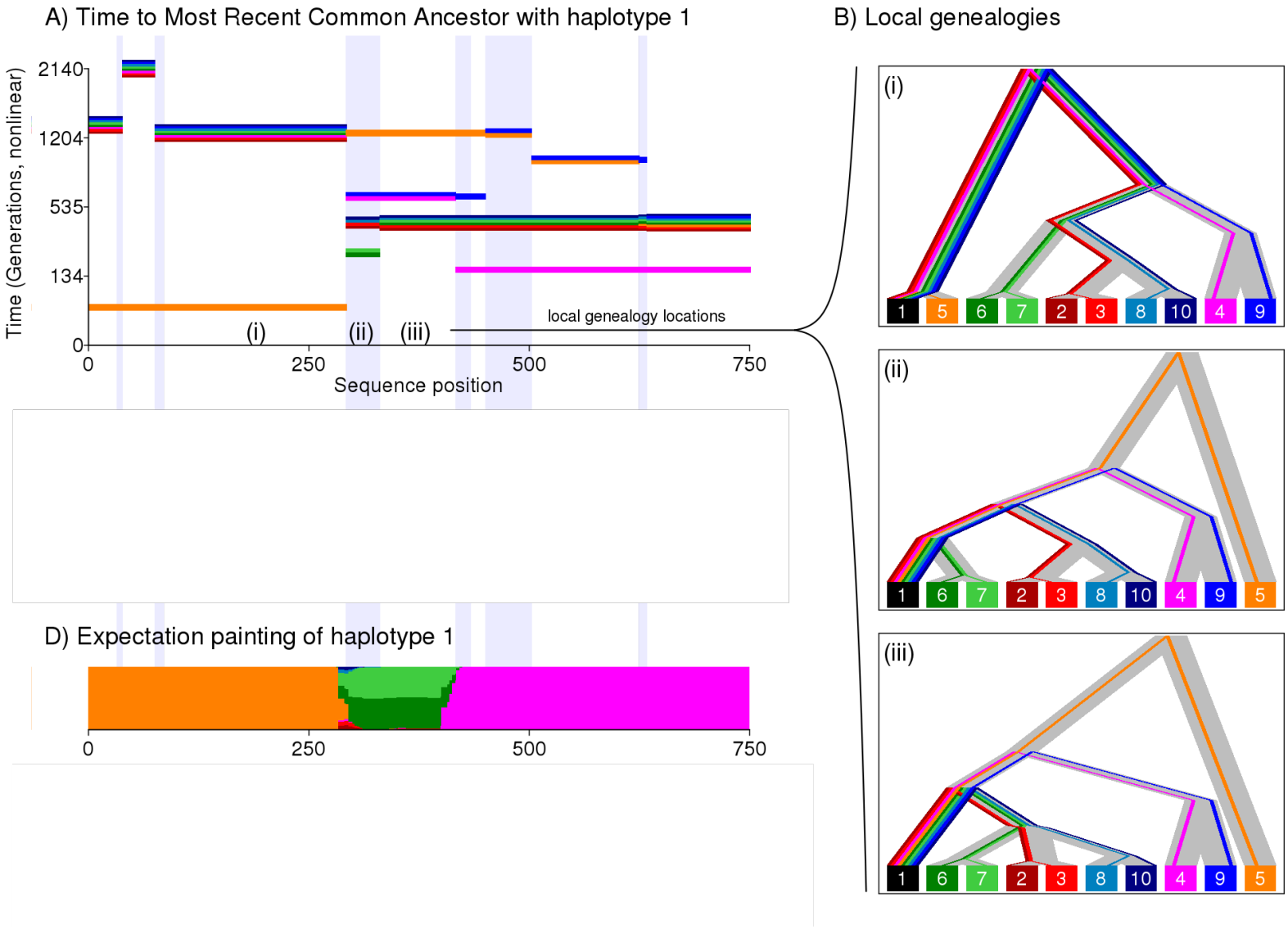
- Painting a chromosome by "copying" segments from other individuals[2]
- Can be done very fast, thanks to Li and Stephens algorithm [3]



[3]Price et al. 2009

[3]Li and Stephens 2003

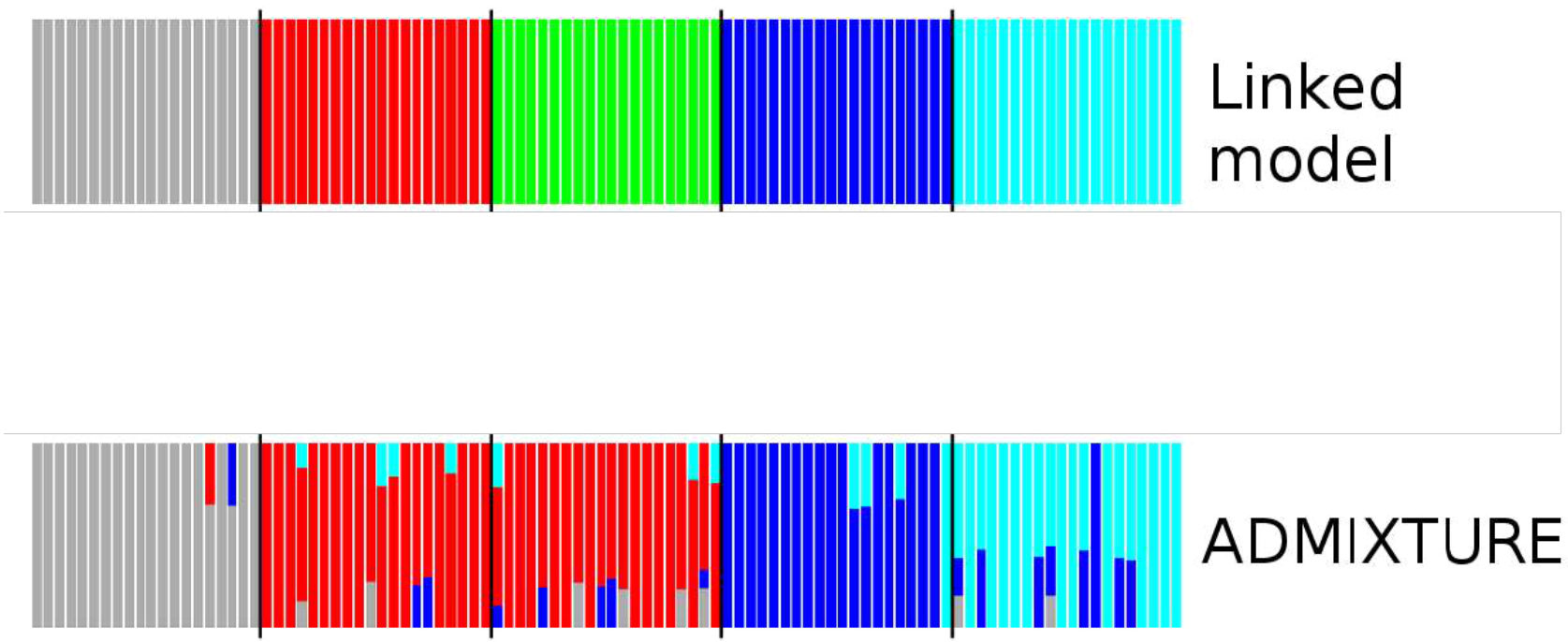
Chromosome painting



[3] Hellenthal et al. 2014

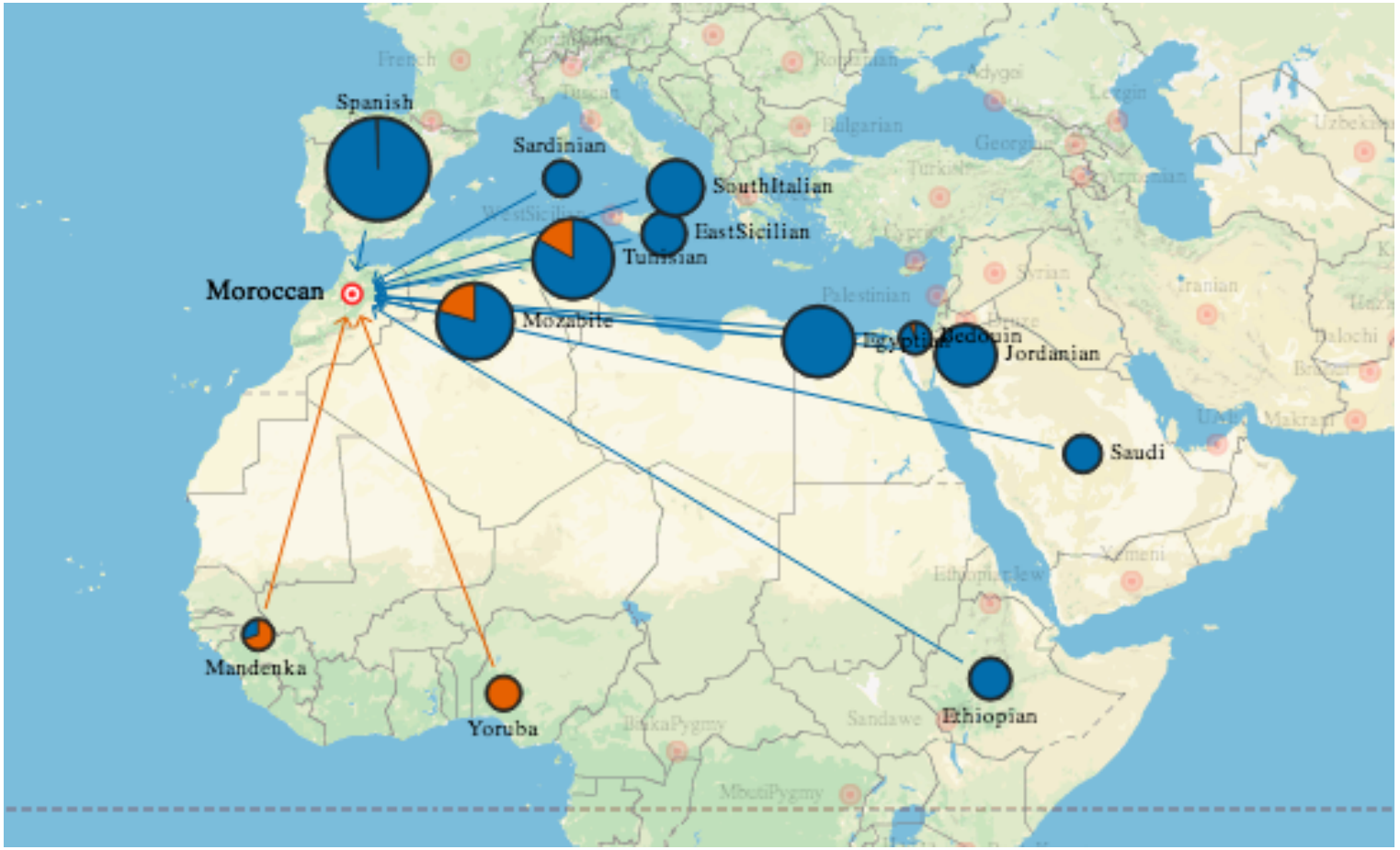
Chromosome "palettes"

- We can create a chromosome palette by aggregating the ancestry from each of the chromosome segments
- This is comparable (but not the same!) as a Structure / Admixture barplot



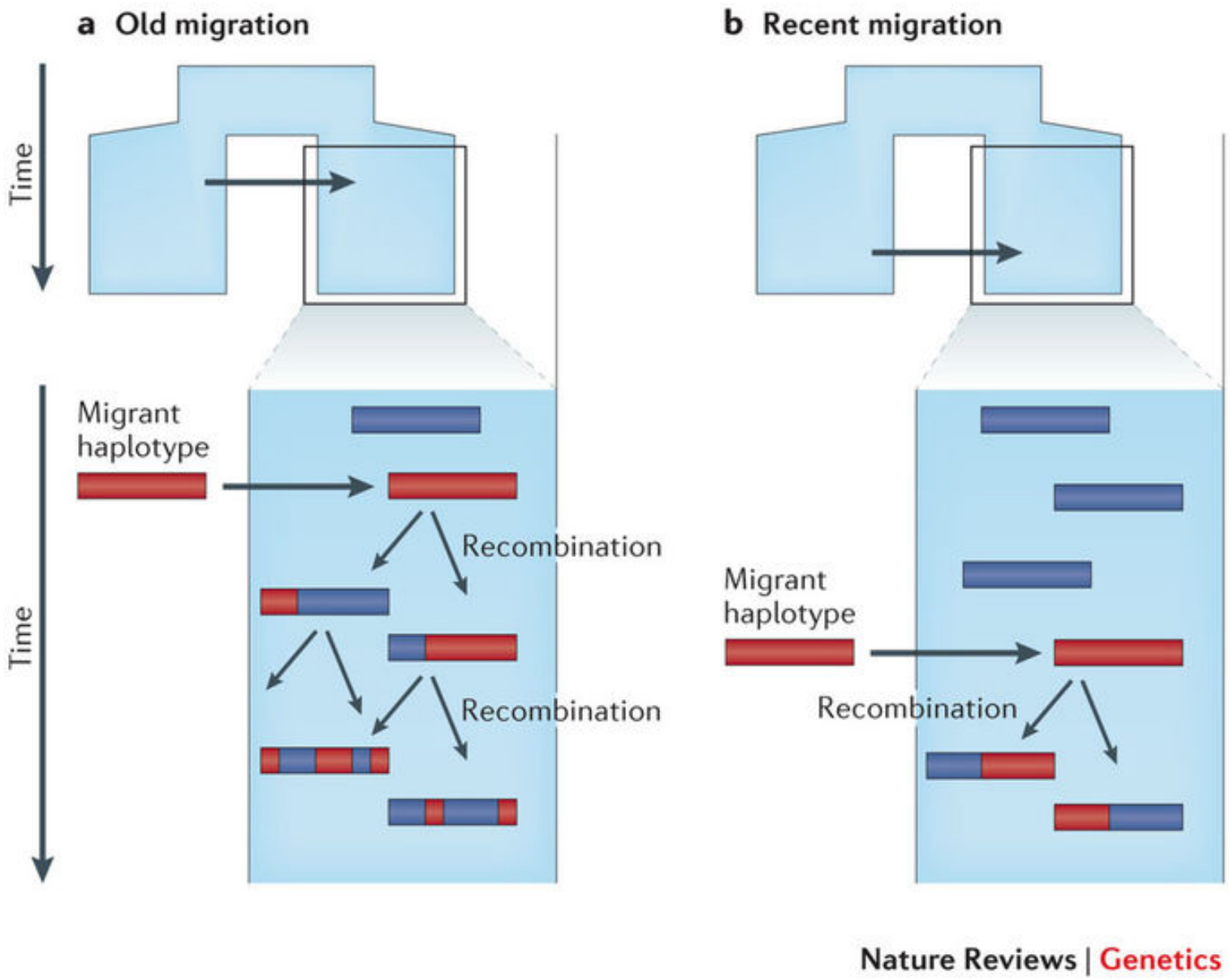
[3] Hellenthal et al. 2014

GlobeTrotter



[3] Hellenthal et al. 2014

Admixture date inference



[3] Souza and Hey 2013

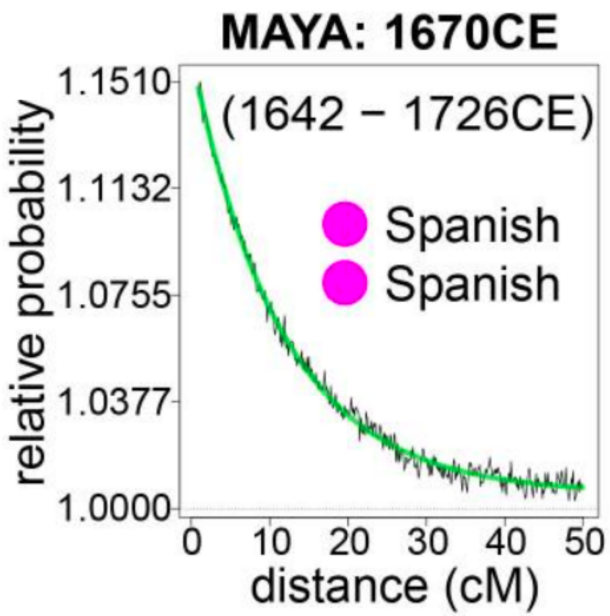
Admixture date inference

- Assume an admixture event happened at time λ between two populations A and B
- We're interested in modeling the length of tracts from A in an admixed genome
- Assume that the probability of no recombination between two points a distance g apart since admixture is $e^{-g\lambda}$ (where λ is the time since admixture scaled by the recombination rate)
- If the fraction of total ancestry from population A in the genome is α, then the probability that we'll find two loci with ancestry from A a distance g apart is:
- $p_{AA}(g) = \alpha(e^{-g\lambda} + (1 - e^{-g\lambda})\alpha) = \alpha^2 + \alpha(1 - \alpha)e^{-g\lambda}$
- This is just an exponential function of the admixture time λ! [4]

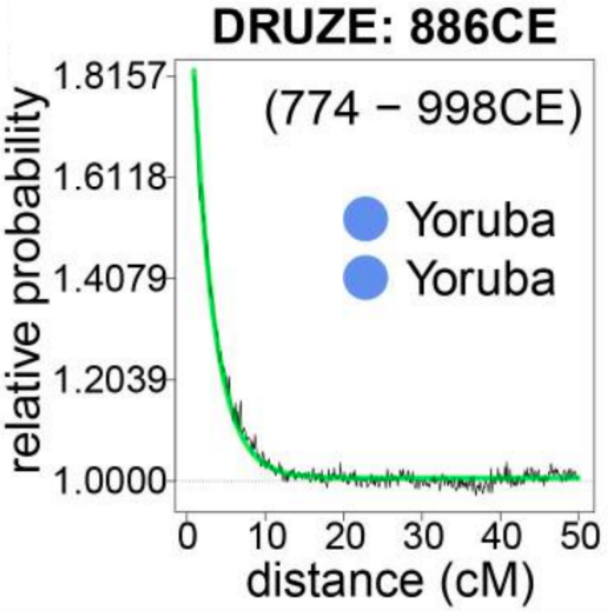
[4] Hellenthal et al. 2014

Admixture date inference

p_Spanish,Spanish as a function of distance in a Mayan genome

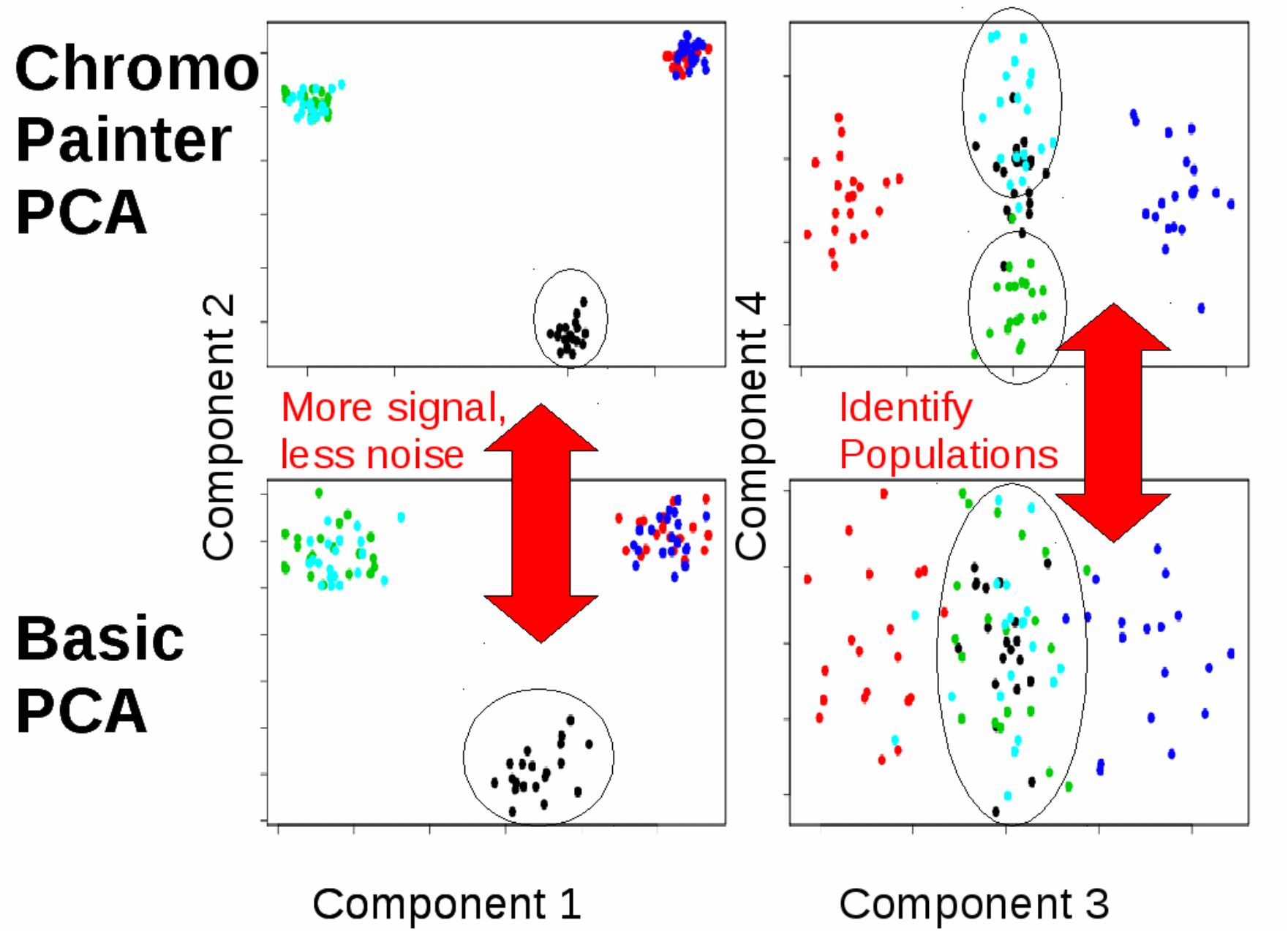


p_Yoruba,Yoruba as a function of distance in a Druze genome

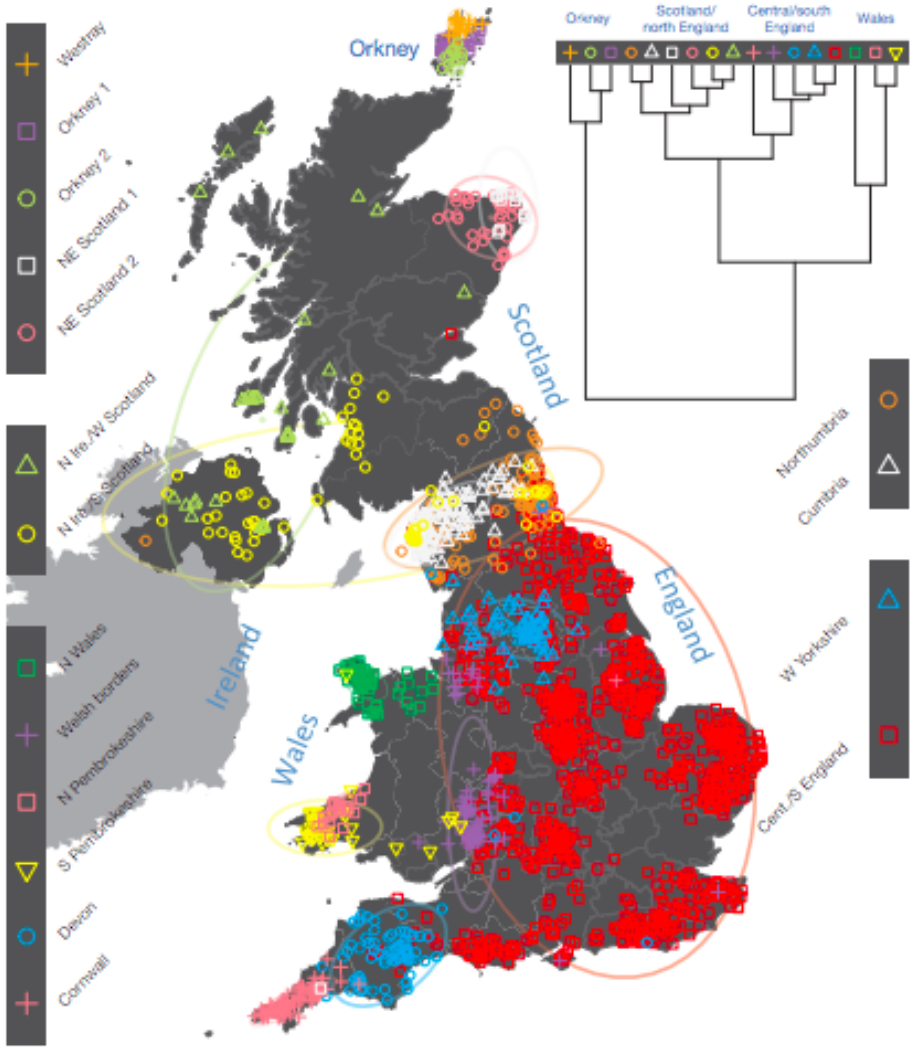


[4] Hellenthal et al. 2014

Haplotype data has more information than SNP data



Haplotype models can capture very recent history



[4] Leslie et al. 2015