# Analysis of Population Structure

**Bertrand Servin**
INRA Animal Genetics
**Joseph Pickrell**
Harvard Medical School

Synbreed Winter School, Salzburg, November 2012

# Outline

1. What do we call *population structure* ?
2. Why do we care ?
3. How do we model / infer it ?

- Morning: review of two popular approaches
- Afternoon: Recent methodological developments and practical

# Outline

# Population Structure

## Definition

The genetic structure of a population characterizes the **distribution of genotypes** across individuals. A population is said to be structured when this distribution is **heterogeneous across individuals**.

# Population Structure

- Consider a sample of $N$ individuals genotyped $L$ markers (SNPs)
- $\mathbf{G}$ is the $N \times L$ matrix of observed genotypes

## Unstructured case

The homogeneous (unstructured case) corresponds to the situation where, at any locus $\ell$ on the genome, for all individuals:

$$g_\ell \sim \mathcal{B}inomial(2, p_\ell)$$

## Structured case

$p_\ell \rightarrow p_{i\ell} = f(\theta_i)$ varies across individuals, affects the whole genome.

# Causes of population structure

- Any phenomenon that affects gene flow between lineages through time will give rise to structure
  - Restricted mating: livestock breeds, closed germplasms ...
  - Geography: strong or soft (isolation by distance) barriers ...
  - ...
- Unstructured samples are the exception.

# Why do we care ?

## Inference of the history of a sample

- Today's structure is caused by (unknown) demographic events in the past
- Characterizing and modeling population structure informs on these events

## GWAS

- Sample of individuals with measured phenotype of interest
- **By chance** phenotype variation correlates with genotype heterogeneity at a locus (structure) : false (non causal) associations

## Null model for genotype distribution

- Help to pinpoint outlying (selected) regions in the genome

# Two common approaches to analyse population structure

**Model-based** the structure arises from latent sub-populations (clusters):

- within a cluster genotype distribution is homogeneous (Hardy-Weinberg)
- STRUCTURE model(s) (Pritchard, Stephens and Donnelly (2000))

**"Model-free"** Principal Component Analysis

# Outline

# Principle

- Underlying the observed genotypes, the sample is structured in latent (unobserved) groups = **clusters**.
- Simplest case: one individual belongs to one group ($z_i$)
- If we knew the cluster memberships:

$$g_{i\ell}|z_i = k \sim \mathcal{B}inomial(2, P_{k\ell})$$

  where $P_{k\ell}$ is the allele probability at locus $\ell$ in cluster $k$.
- This defines the data-augmented likelihood : $P(\mathbf{G}|\mathbf{Z}, \mathbf{P})$

# Inference

$$P(\mathbf{Z}, \mathbf{P}|\mathbf{G}) \propto P(\mathbf{G}|\mathbf{Z}, \mathbf{P})P(\mathbf{Z})P(\mathbf{P})$$

- Given this model, inferring population structure is :
  - Assigning individuals to clusters : $P(\mathbf{Z}|\mathbf{G})$ posterior probability of each cluster for each individual
  - Estimating allele frequencies in each group: $P(\mathbf{P}|\mathbf{G})$
- Technically, this is done using MCMC (in the original paper)

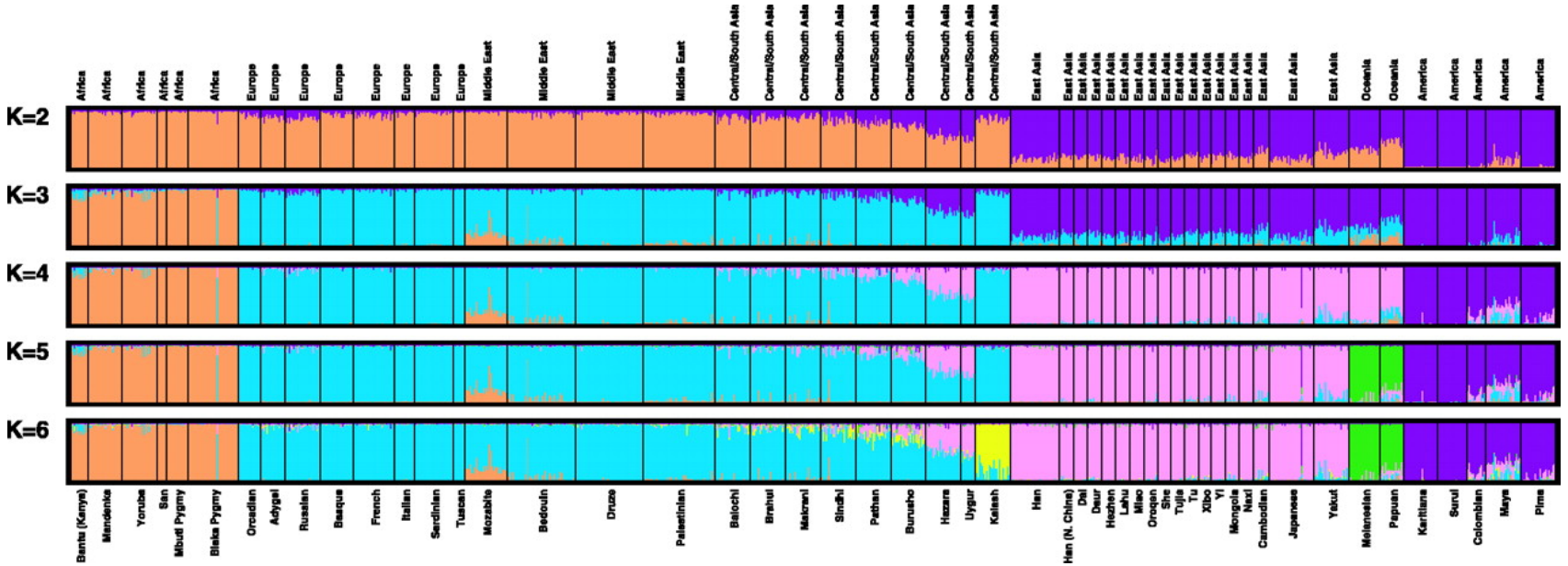
# Incorporating Admixture

- Simple model assumes one individual comes from one group.
- But one genome can be a mosaic from multiple populations = *admixture* (partial barrier to gene flow)
- $q_{ik}$, proportion of the genome of individual $i$ coming from cluster $k$.

$$g_{i\ell}|q_{ik} \sim \mathcal{B}inomial(2, p_{i\ell})$$

$$p_{i\ell} = \sum_{k=1}^{K} q_{ik} P_{kj}$$

# Example: HGDP Rosenberg *et al.* (2002)




Estimated $q_{ik}$ for each individual

# Incorporating Admixture Linkage Disequilibrium

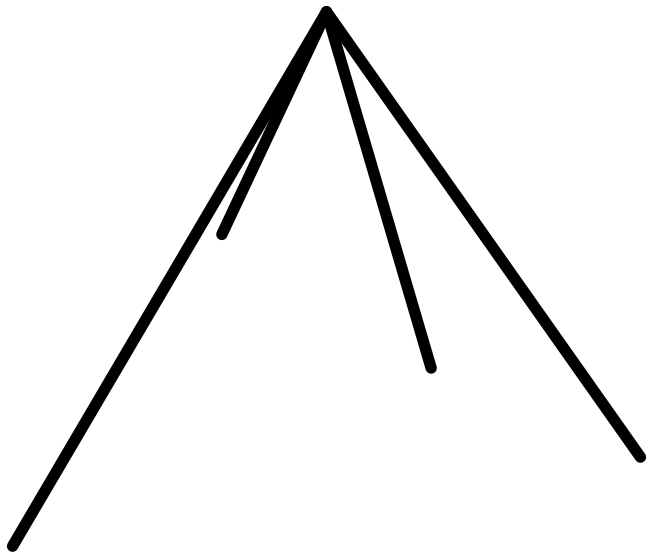
- After admixture events, whole chromosomes are inherited from a single "cluster"
- Through time, recombination will break these chromosomes
- However, close loci will tend to originate from the same "cluster".
- STRUCTURE's Linkage Model incorporates :

$$Pr(z_{\ell+1} = k' | z_\ell = k, r, Q)$$

  which depends on a recombination parameter $r$
- Technically becomes a Hidden Markov Model

# Correlated Allele frequencies

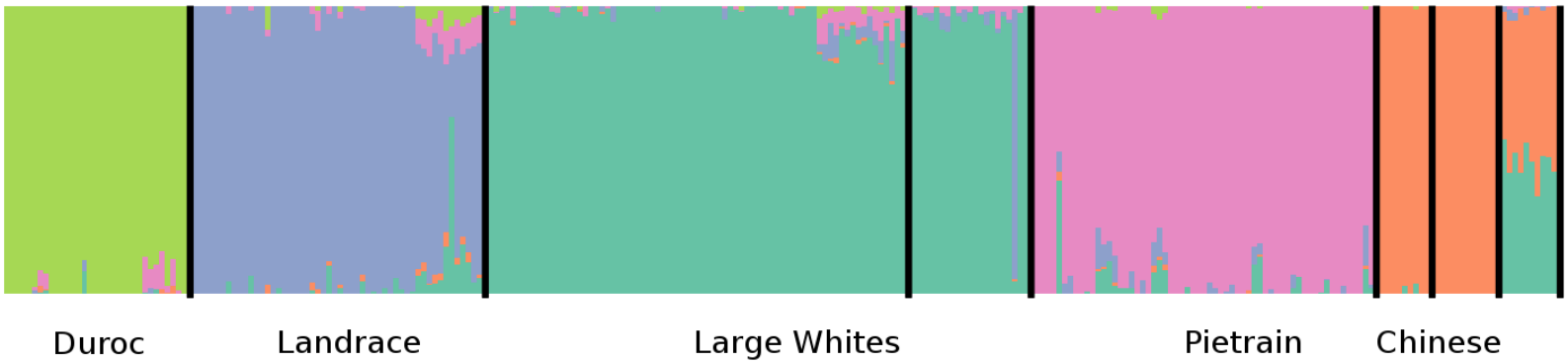
- Populations underlying data are not independant
- Simple model: arise from a common ancestral population, with a star-like phylogeny
- Consequence: correlation between allele frequencies
- Structure $F$ model.



# STRUCTURE on large SNP datasets

- In the original implementation of the Pritchard *et al.* model, parameter estimation is performed via MCMC
- Offers posterior distribution of (the numerous) model parameters
- For large SNP datasets, running times become (very) large
- Admixture Software (Alexander et al., 2009) offers a solution to quickly get parameter estimates (**Q** and **P**).
- Does not implement the linkage or F model

# Example of Pig breeds




- Clear breed structure
- mislabeled individuals
- small amount of admixture within pure breeds
- synthetic line shows expected contributions from founder origins

# Outline

# Principle

- Given in Patterson, Price and Reich (2006)
- Consider the matrix of genotypes $\mathbf{G}$
- Center and standardize each column (SNP genotypes) : $\mathbf{M}$
- $\mathbf{F} = \frac{1}{L}\mathbf{M}\mathbf{M}'$ is the kinship matrix between individuals, estimated from SNP data.
- Perform Eigen-decomposition of $\mathbf{F}$, *i.e.* find the matrix of eigen-vectors $\mathbf{U}$ and corresponding eigen-values $\boldsymbol{\lambda}$ satisfying:

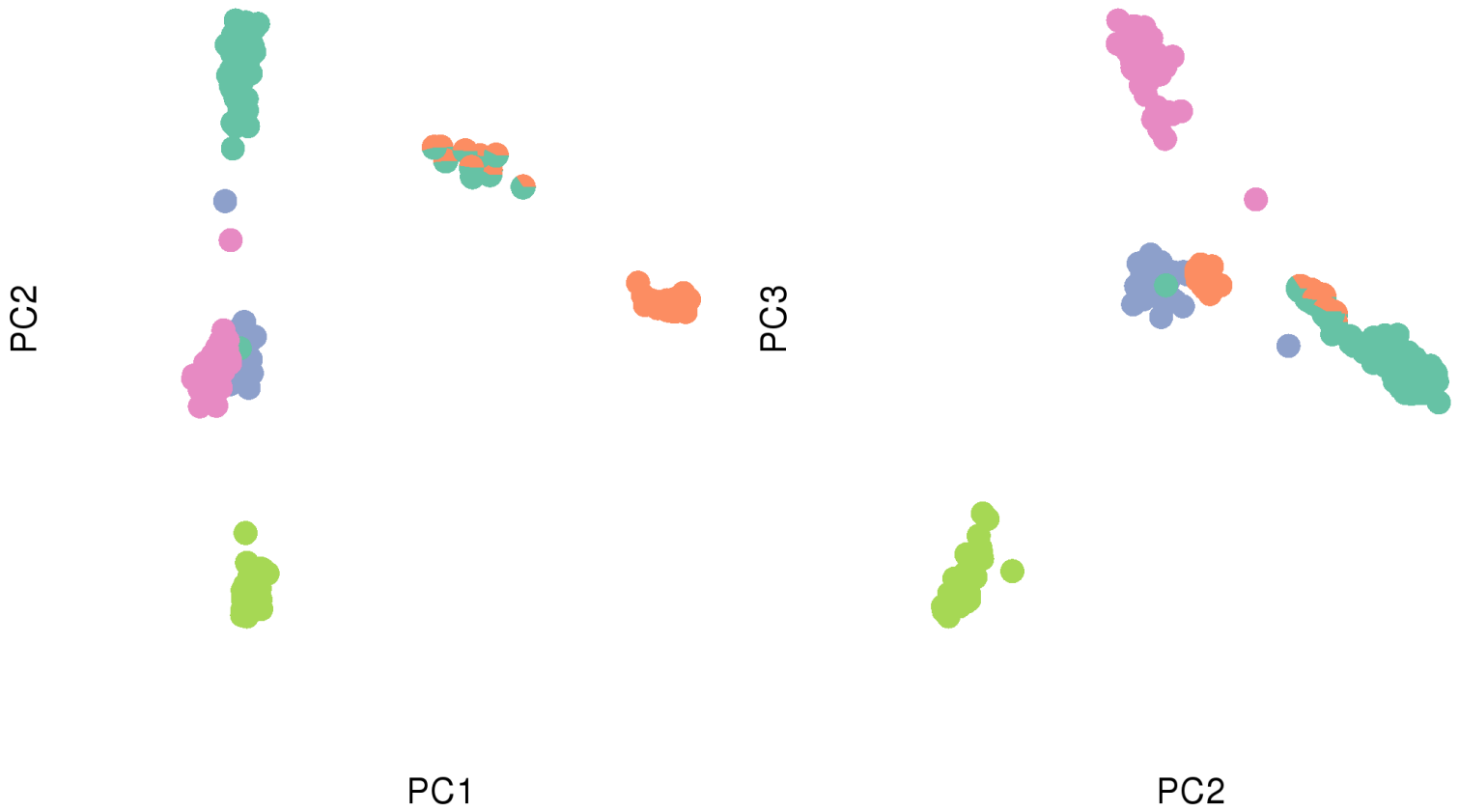
$$\mathbf{F} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix with $\lambda_i$ on the diagonal

# PCA captures axes of variation in kinship

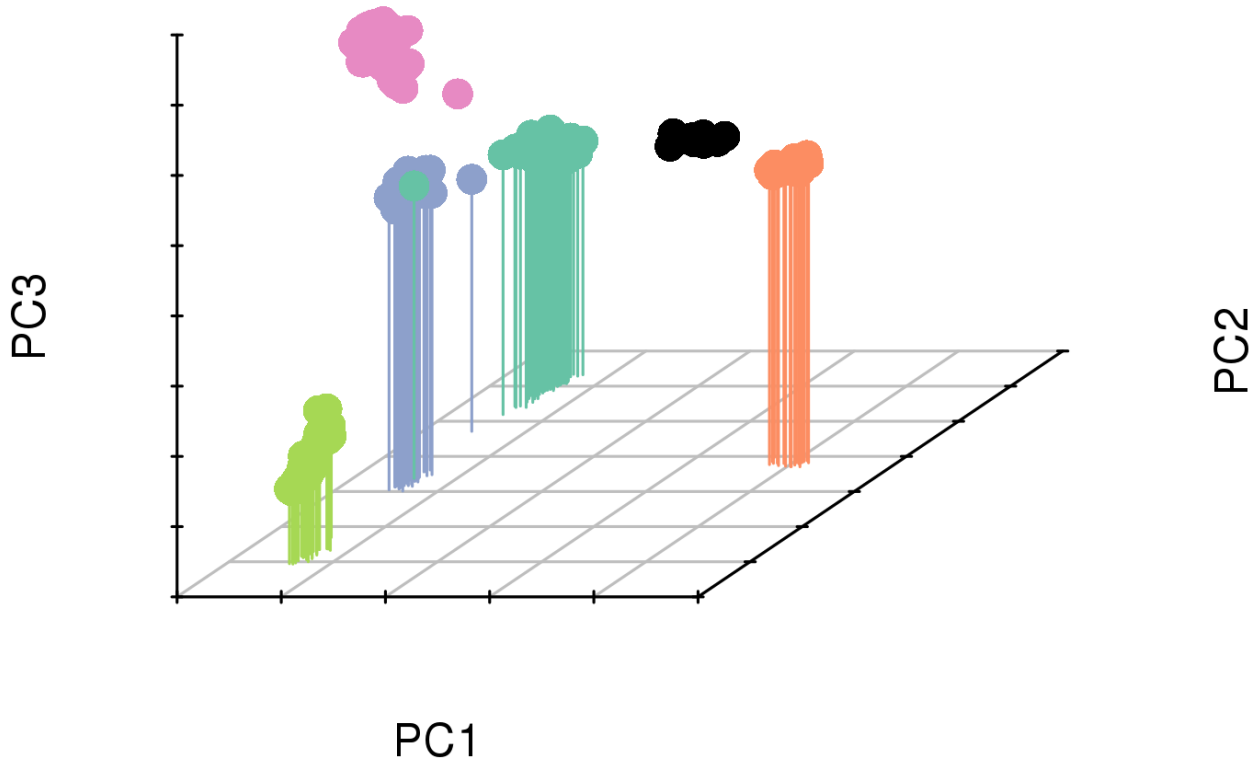
- The eigen-vectors (the principal components) define "axes of variations" in kinship
- Ordered by importance (large $\lambda$ values)
- The first principal components capture strong effects on the variation in genotypes = population structure effects.
- Indeed, there is a genealogical interpretation to PCA on genetic data (McVean 2009)

# Back to pigs



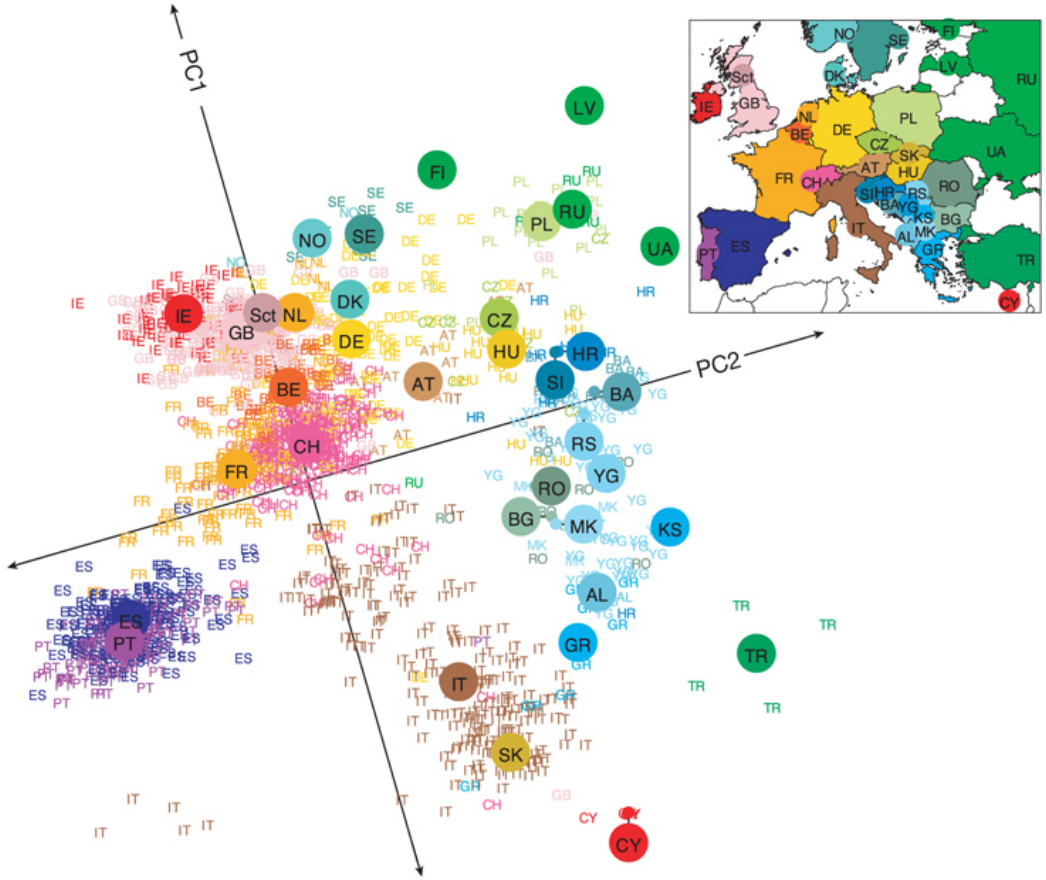

PCA captures the structure in individual breeds, and places admixed individuals between the contributing populations.

# Back to pigs




PCA captures the structure in individual breeds, and places admixed individuals between the contributing populations.

# PCA mirrors geography in Europe (Novembre et al. 2008)



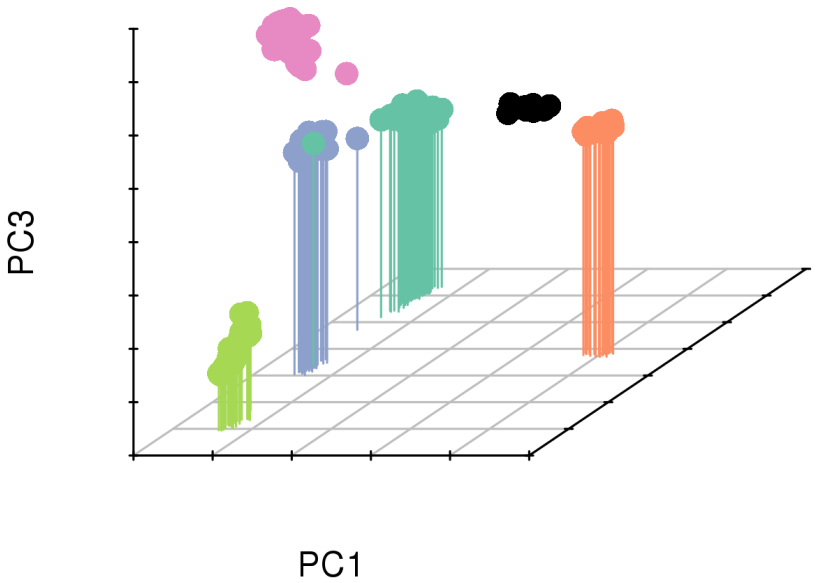

In these data, the main axes of variation correspond to space (geography)

# Conclusions on PCA

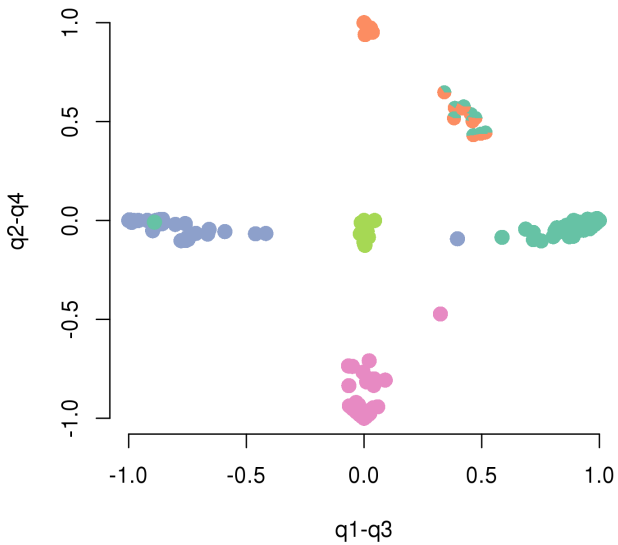
- Fast
- Can capture continuous variation (clines, isolation by distance ...) better than original STRUCTURE models
- Lack the interpretability of inference based on population genetics models.
- Influenced by sample sizes

# Outline

# Are they so different ? Pigs

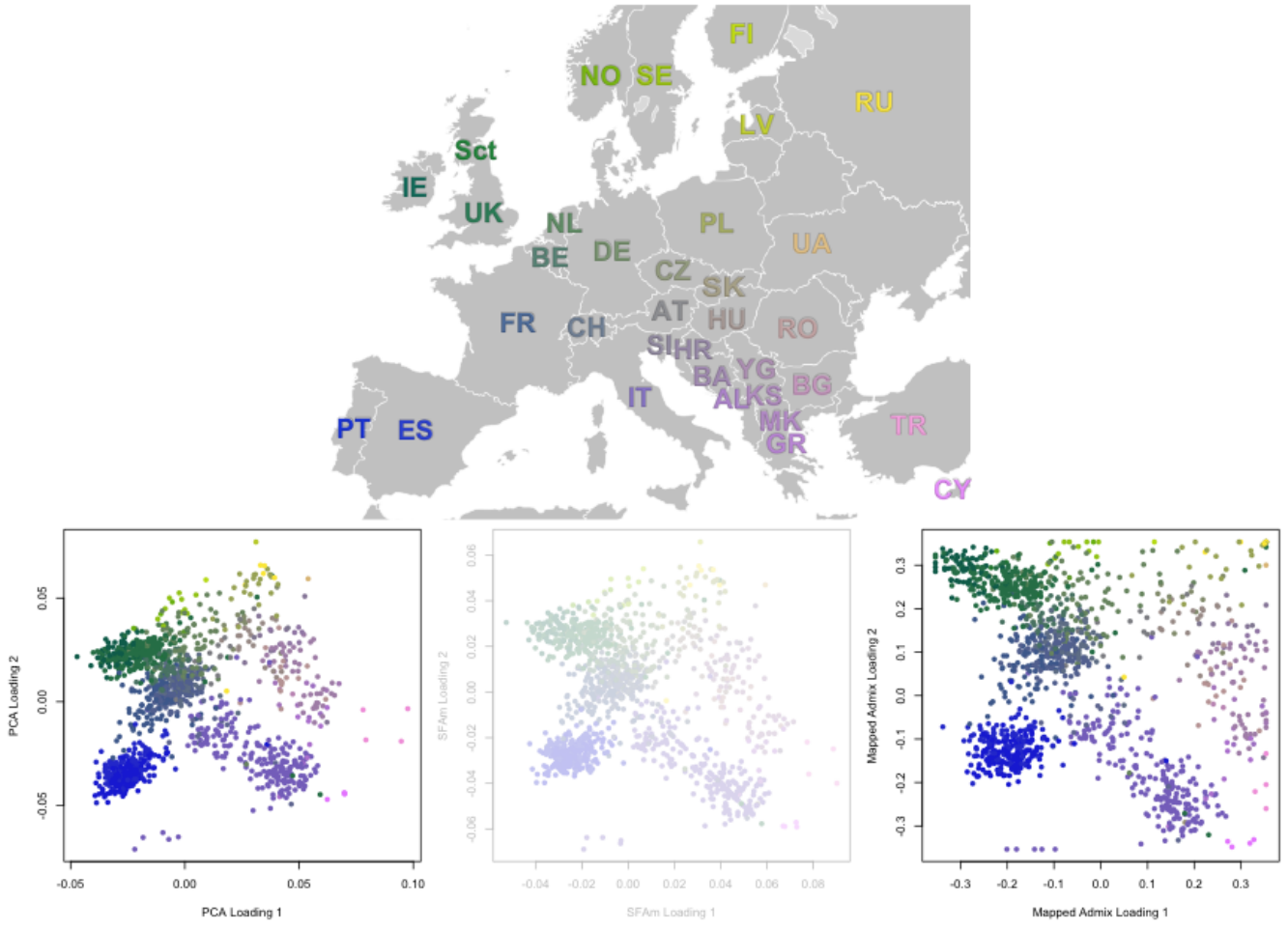



PCA




STRUCTURE

# Are they so different ? Europeans




Engelhardt and Stephens (2010)

# A unified framework

Engelhardt and Stephens (2010) show both approaches can be seen as modeling:

$$E[g_{ij}] = \sum_{k=1}^{K} \Lambda_{i,k} F_{k,j}$$

- $k$ indexes PCs (PCA) or clusters (STRUCTURE)
- $\Lambda_{i\bullet}$: PCA loadings , STRUCTURE's admixture proportions
- $F_{\bullet j}$: PCA factors, STRUCTURE's cluster mean allele frequencies ($\times 2$)

while imposing different constraints on $\mathbf{\Lambda}$ and $\mathbf{F}$. Opens the floor to new models ...

# Going further

- PCA and STRUCTURE models do not incorporate explicitly hierarchical structure of populations: how to infer structure with trees / graphs ?
- How do we map back admixture on the genome ? identify admixture on individual chromosome segments
- Come back this afternoon to know.

# Cited Papers / Suggested Readings

## Theory


Pritchard et al. (2000) Genetics. Original STRUCTURE paper.

Patterson et al. (2006) PLoS Genetics. Eigen-analysis of population structure

McVean (2009) PLoS Genetics. Genealogical interpretation of PCA.

Engelhardt and Stephens (2010) PLoS Genetics. Unified framework linking PCA and STRUCTURE.


## Data Application


Rosenberg et al. (2002) Science. Structure in the Human Genome Diversity Project

Novembre et al. (2008) Nature. PCA mirrors geography in Europe.