Imperial College London

BBE

MRC Medical Research Council

Exploring genetic diversity: multivariate approaches

Thibaut Jombart – Imperial College London

t.jombart@imperial.ac.uk

Content

I) Introduction: data and methods

2) Multivariate analysis in a nutshell

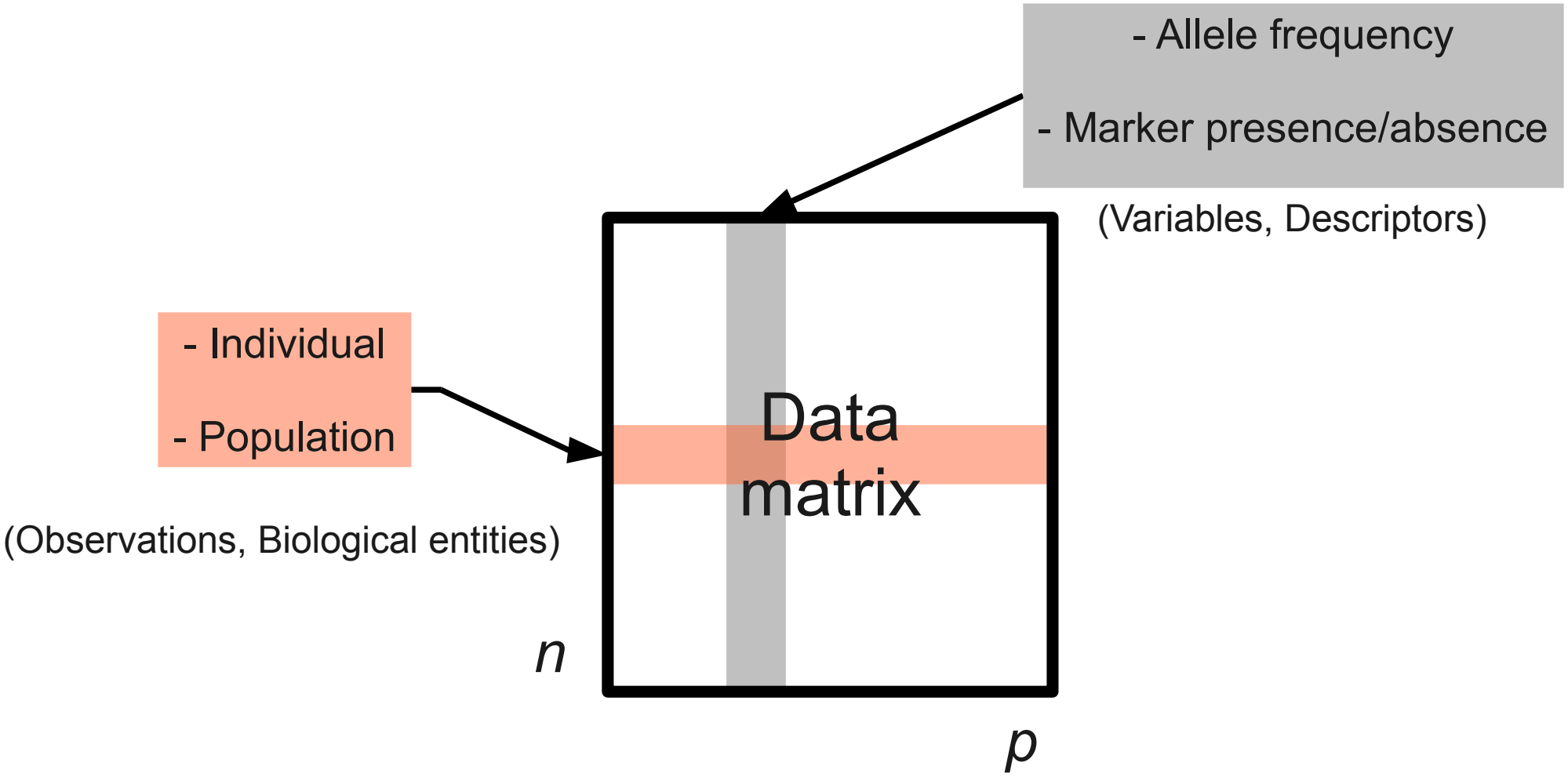
3) Some examples

4) Spatial genetic structures

5) In practice

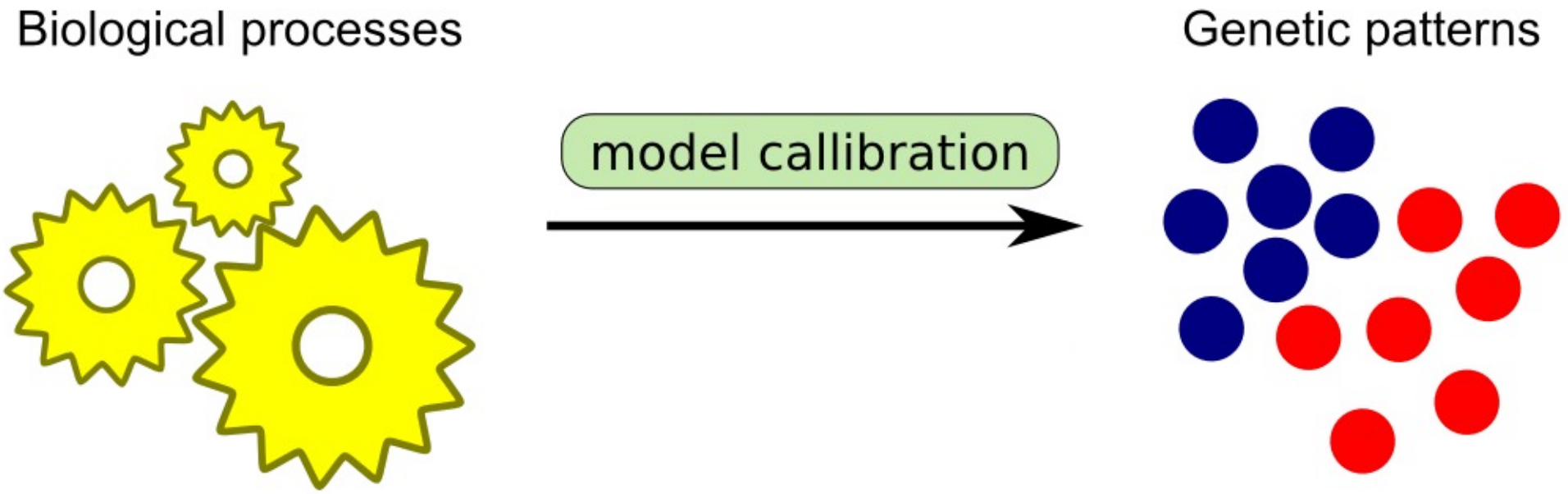
Content

Genetic data (genetic markers)



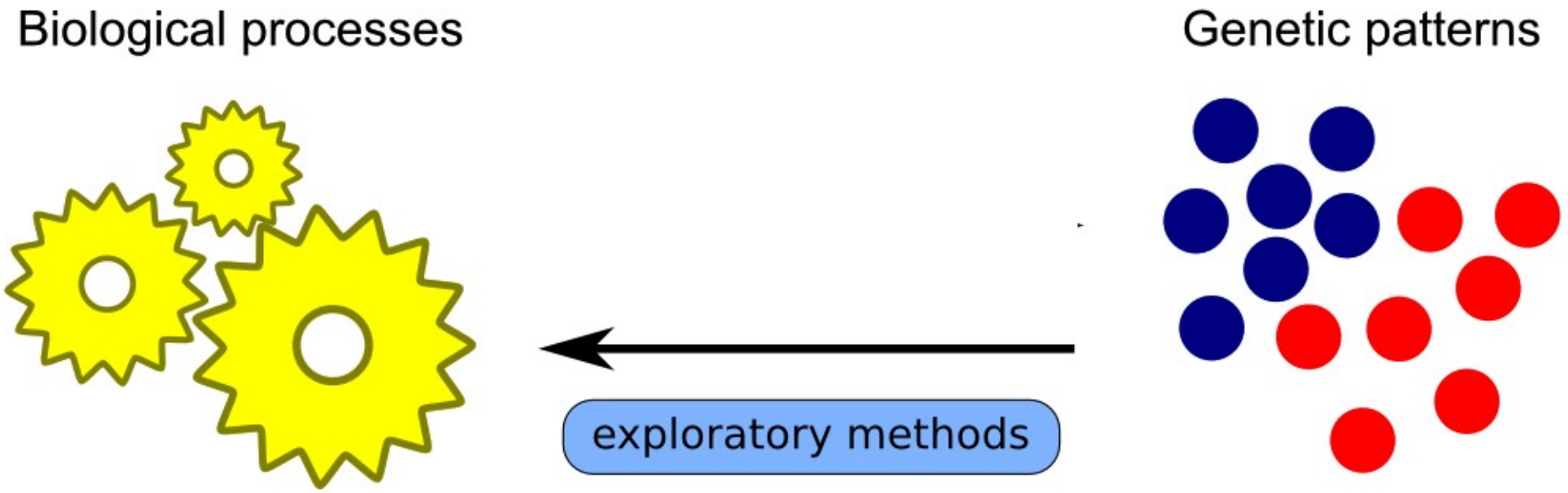
Different ways of exploiting this rich information

Two complementary approaches



- One or a small number of models (possible comparisons)
- Parameter estimation based on the data

Two complementary approaches



- No explicit model, little underlying assumptions
- Description of specific features (e.g. genetic diversity)

Different types of methods

- **Model-based approaches**
  - Bayesian clustering (e.g. STRUCTURE, BAPS)
  - Phylogenetic trees (e.g. PhyML, BEAST)
- **Exploratory approaches**
  - Distance-based trees (e.g. UPGMA, NJ)
  - Multivariate methods (e.g. PCA, PCoA)

Different methods

- **Model-based approaches**
  - Bayesian clustering (e.g. STRUCTURE, BAPS)
  - Phylogenetic trees (e.g. PhyML, BEAST)
- **Exploratory approaches**
  - Distance-based trees (e.g. UPGMA, NJ)
  - Multivariate methods (e.g. PCA, PCoA)

Content

I) Introduction: data and methods

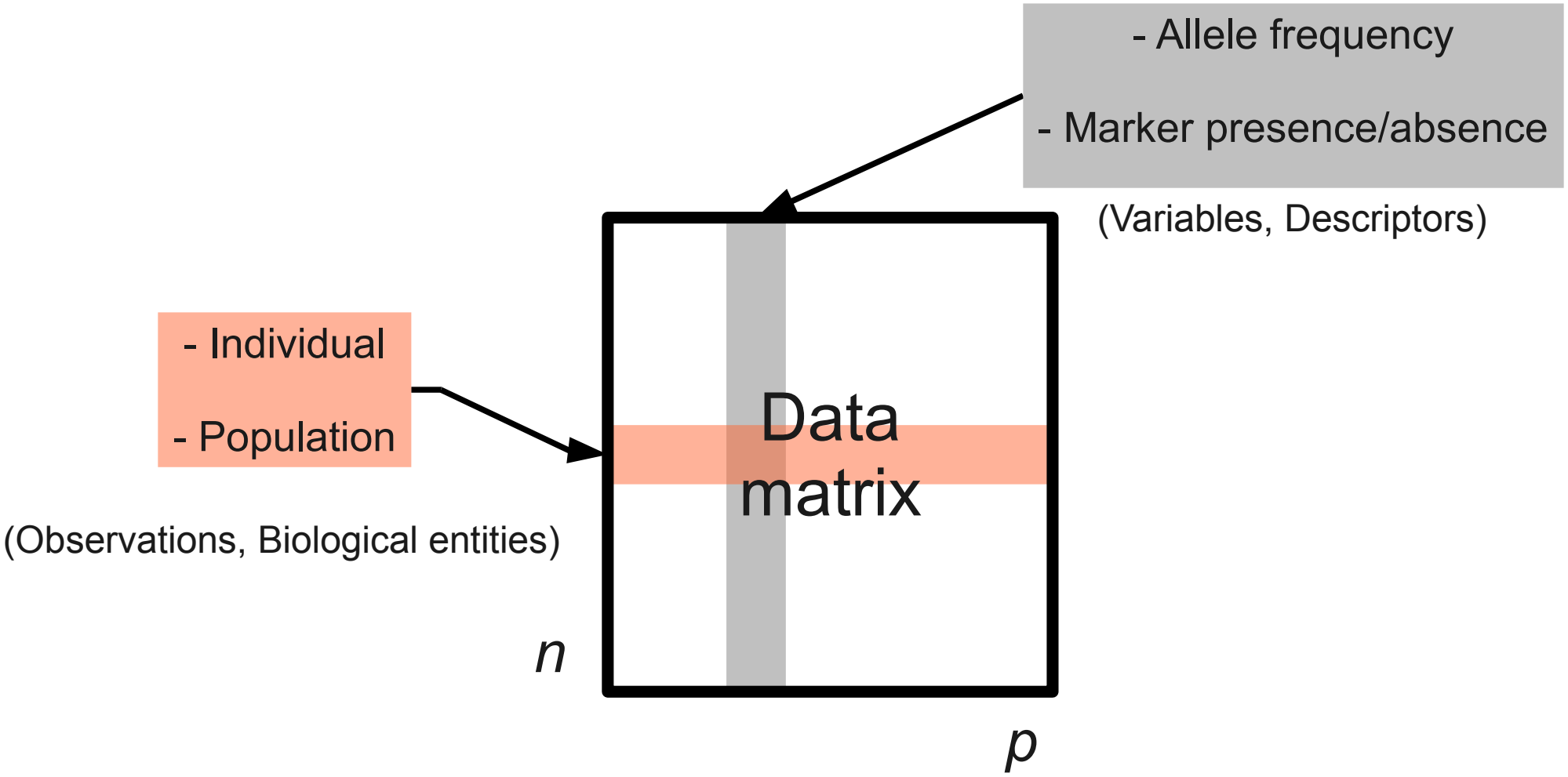
2) Multivariate analysis in a nutshell

3) Some examples

4) Spatial genetic structures

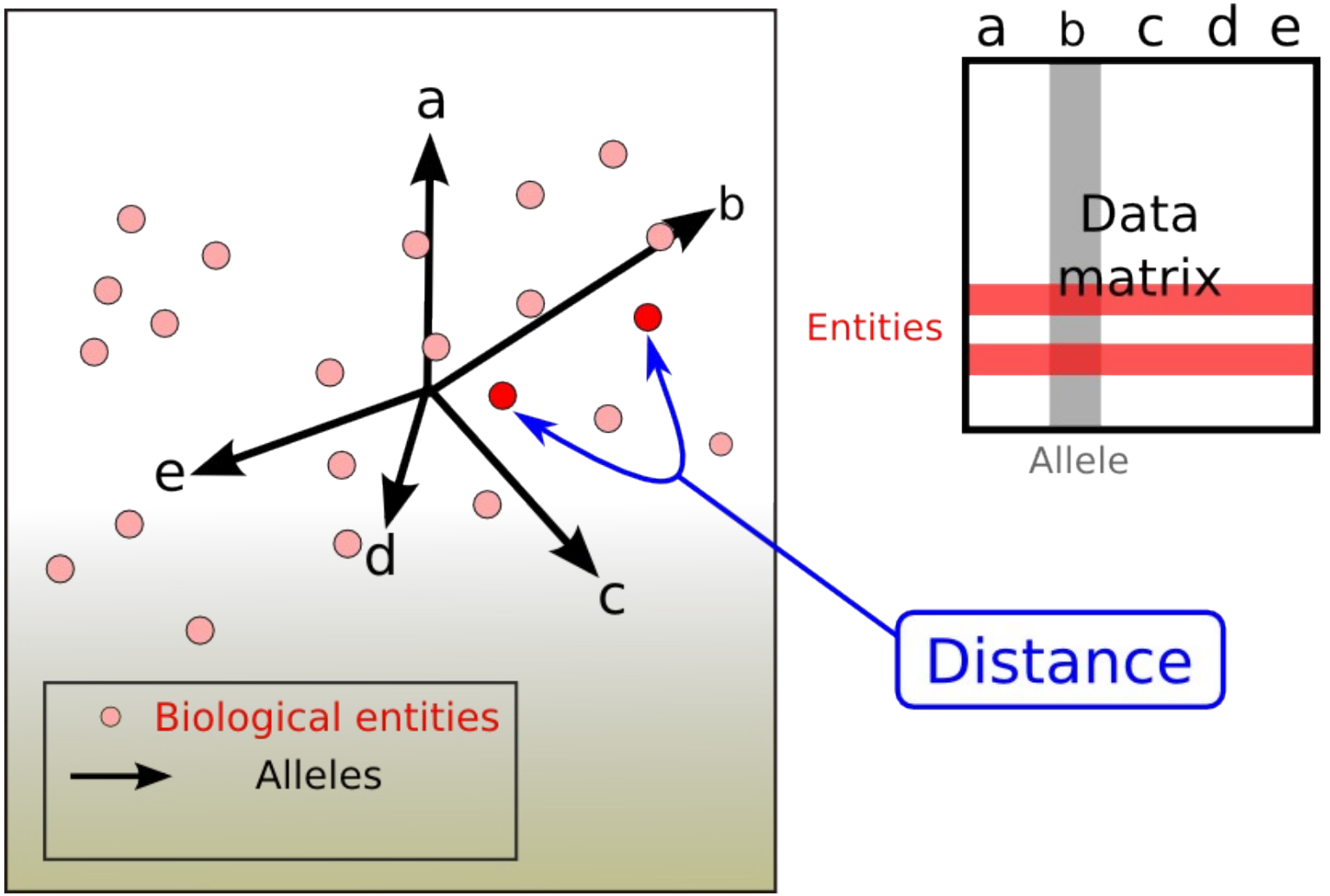
5) In practice

Genetic data (again)



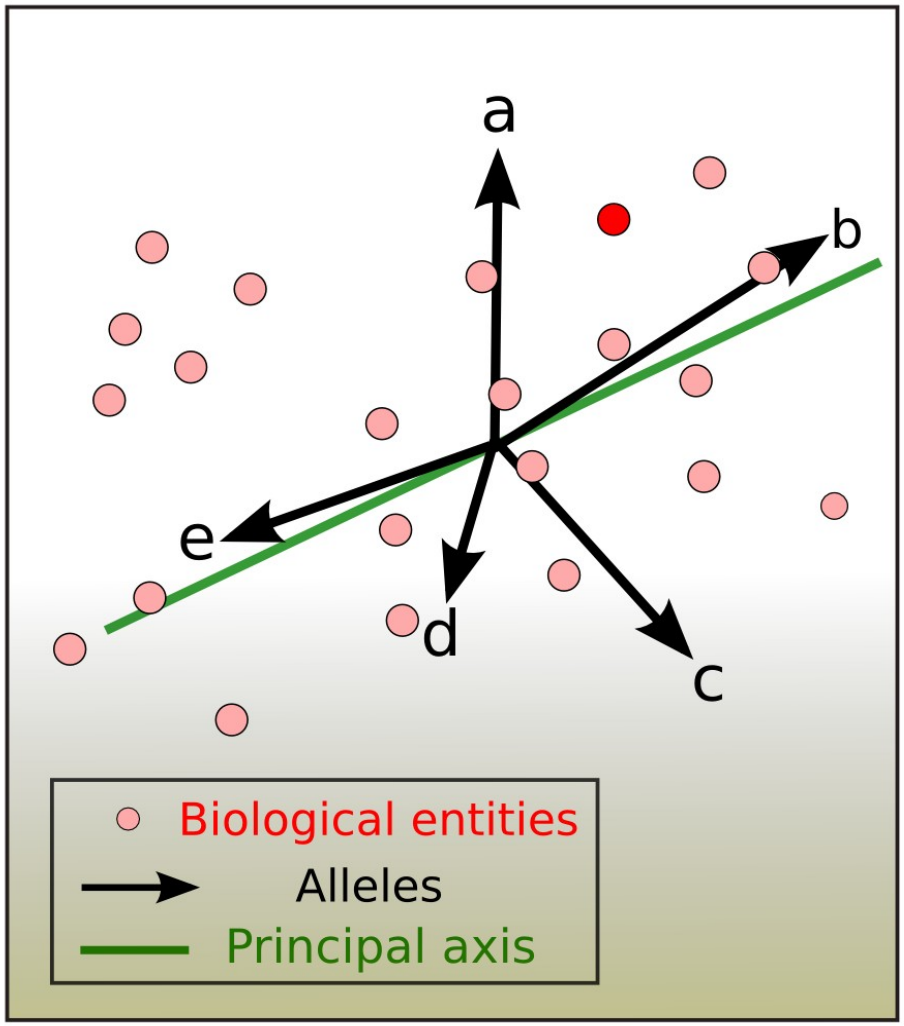
This information can be considered from the geometric point of view.

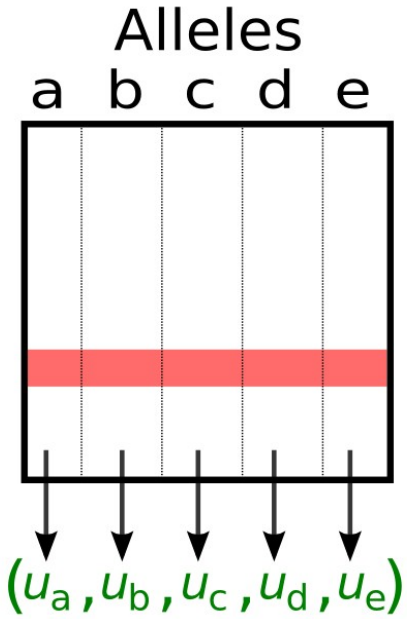
Multivariate analysis – rationale (1/3)



Which directions are the most informative?

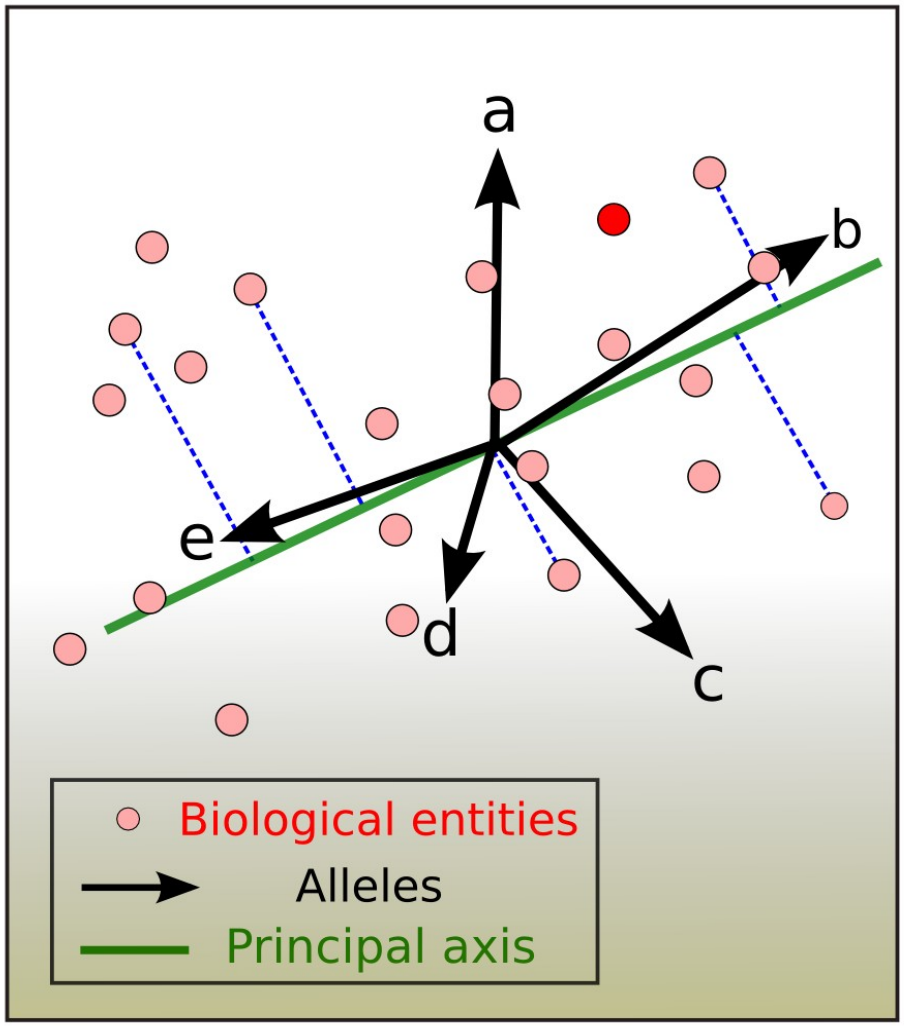
Multivariate analysis – rationale (2/3)

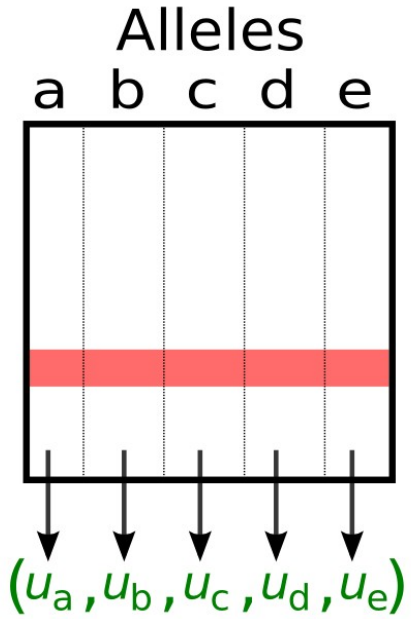




Loadings
(Principal axis)

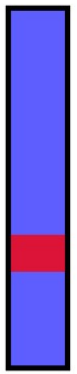
Multivariate analysis – rationale (2/3)





Loadings
(Principal axis)

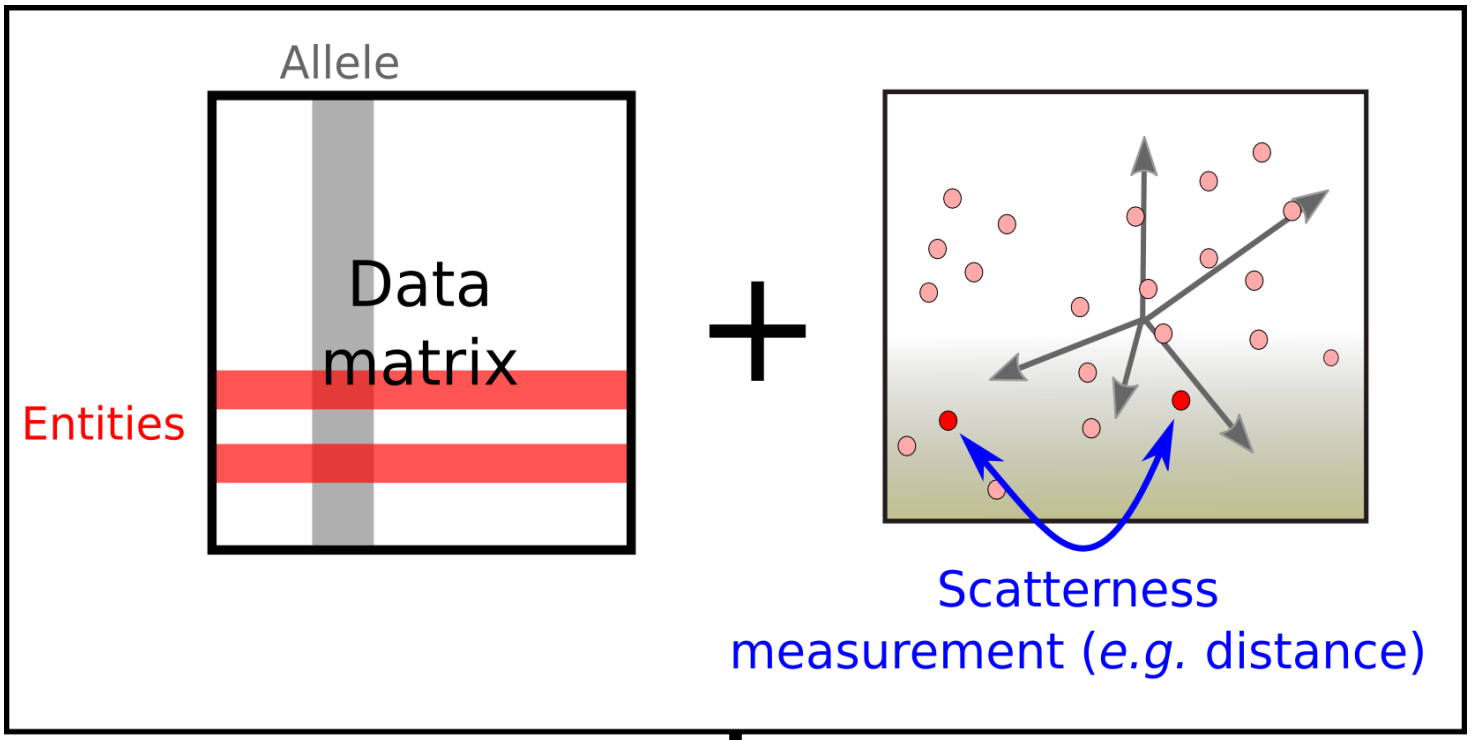
$u_a a + u_b b + u_c c + u_d d + u_e e$ =



Synthetic variable
(Principal component)

Summarise the genetic diversity among individuals / populations

Multivariate analysis – rationale (3/3)



Principal axes (alleles contributions)
Principal components (max. diversity, uncorrelated)

In practice, lots of methods are used

To name a few (used in genetics):

- Principal Component Analysis (PCA)
  - centred / not centred / fancy centring
  - scaled / not scaled / fancy scaling
  - transformed for compositional data

- Principal Coordinates Analysis (PCoA),
  aka (Metric) Multidimensional Scaling (MDS)
  - many genetic distances

- Correspondence Analysis (CA)

- Discriminant Analysis (DA)

- … (review in Jombart *et al.* 2009, *Heredity* **102**, 330-341)

Content

I) Introduction: data and methods

2) Multivariate analysis in a nutshell

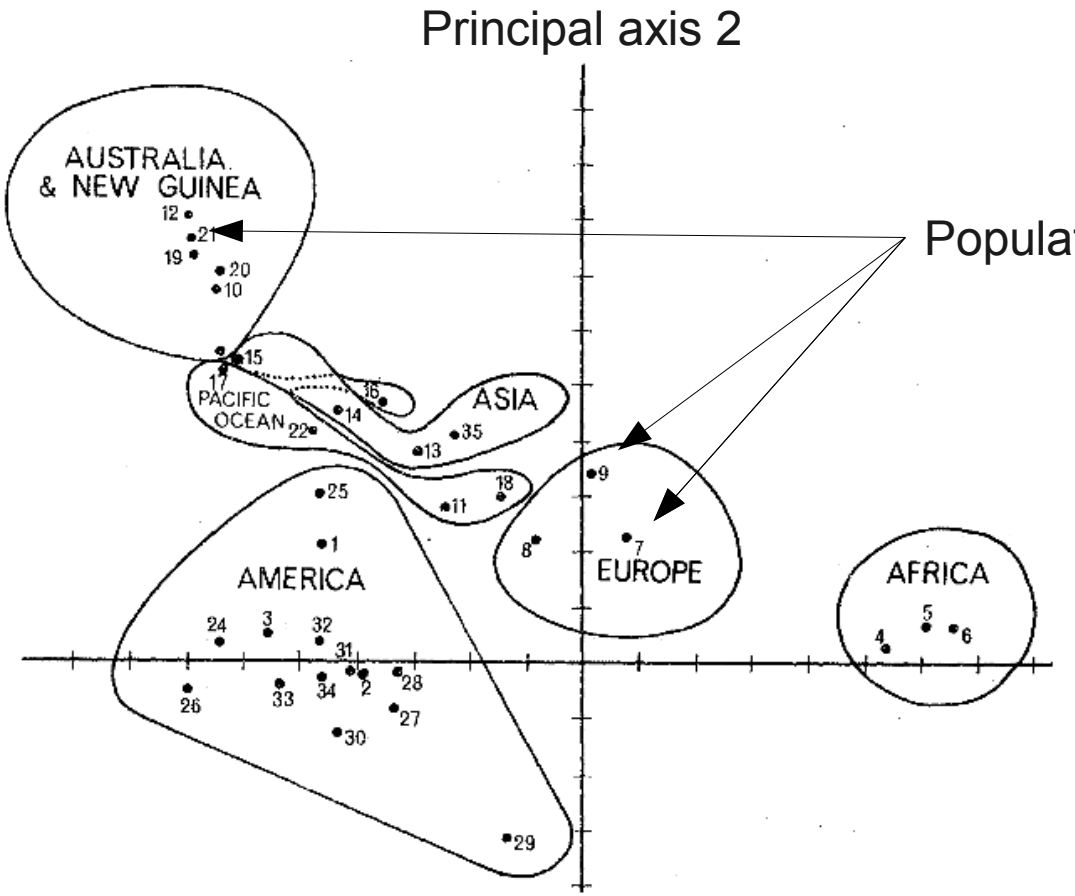
3) Some examples

4) Spatial genetic structures

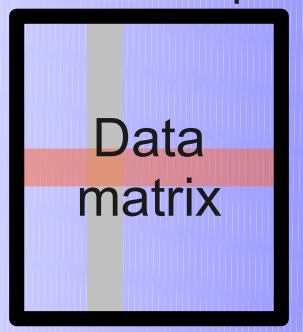
5) In practice

Multivariate analyses: some examples (1/6)

Getting a picture of the genetic diversity

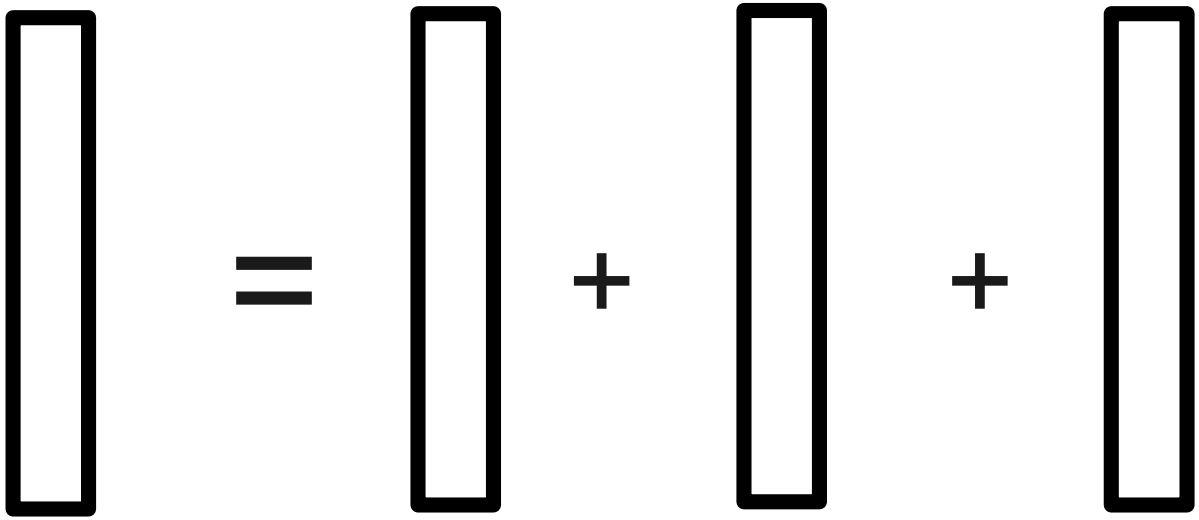


(Cavalli-Sforza 1966, *Proc. B*)



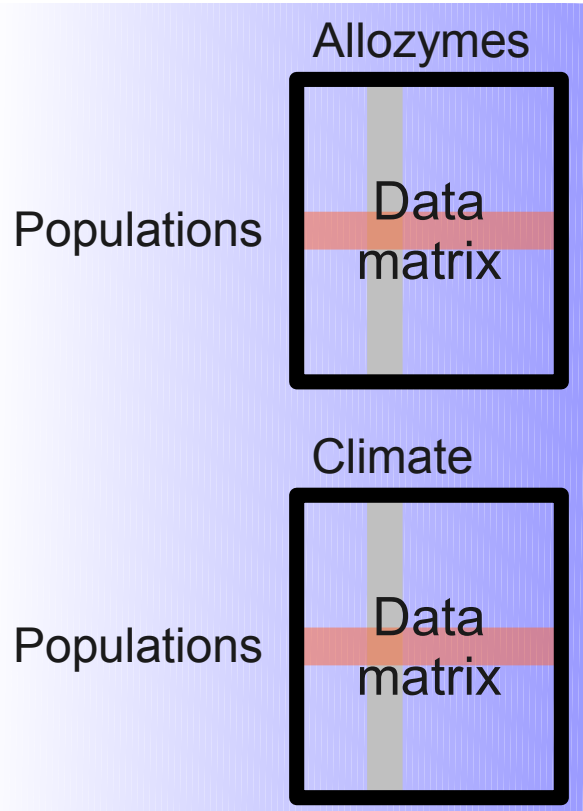
Multivariate analyses: some examples (2/6)

Using PC as response/predictor in models



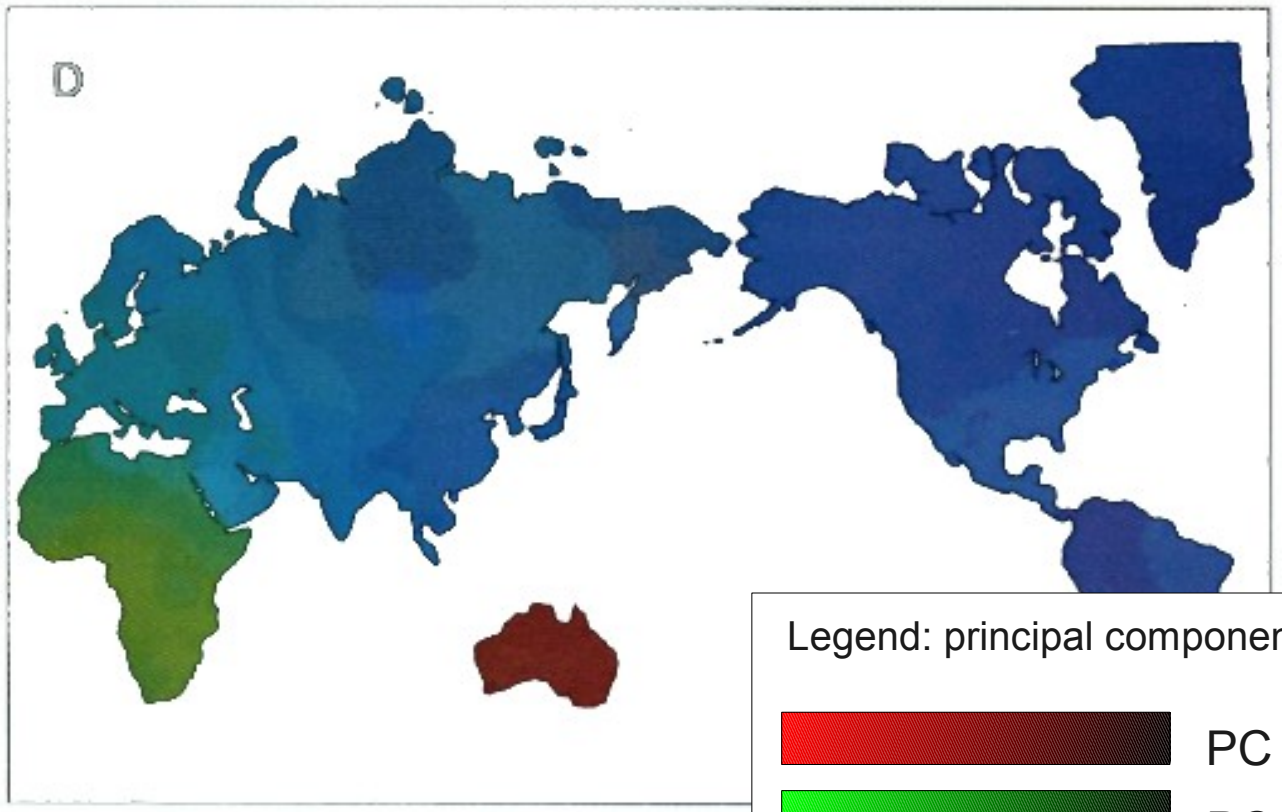
Genetic features (Principal components) | Covariates | Environmental variables | Residuals

(Mulley *et al.* 1979, *Biochemical Genetics*)



Multivariate analyses: some examples (3/6)

Mapping the genetic differentiation



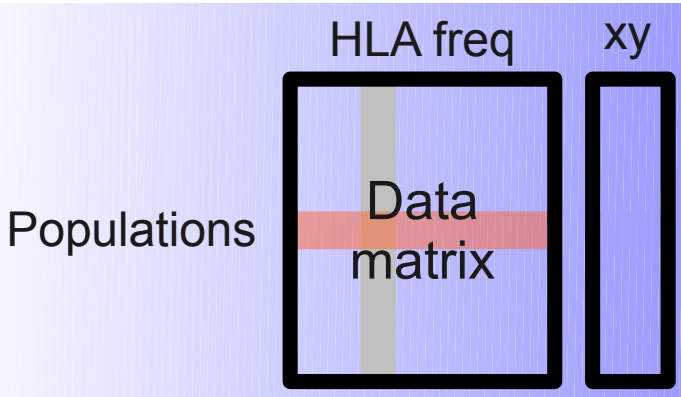
Legend: principal components

PC 1

PC 2

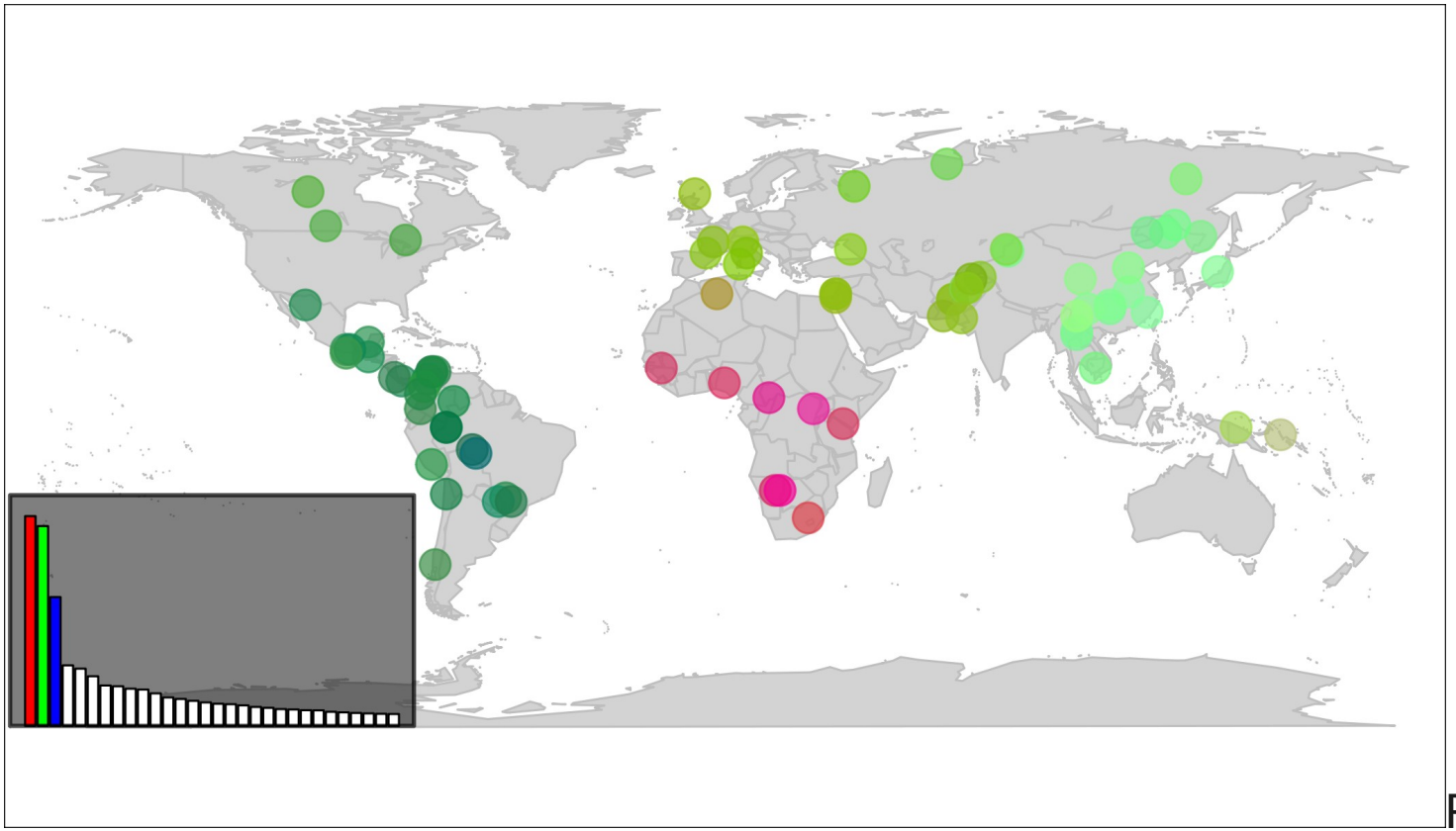
PC3

(Cavalli-Sforza *et al.* 1993, *Science*)

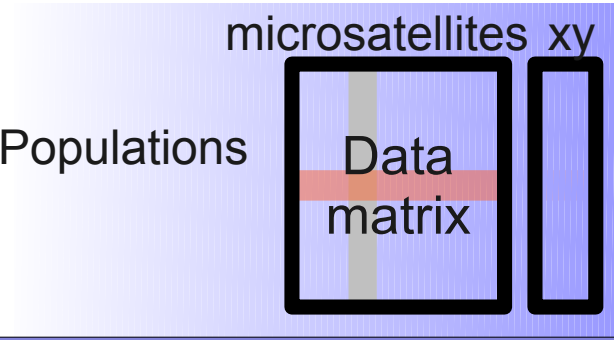


Multivariate analyses: some examples (4/6)

Mapping the genetic differentiation (again)

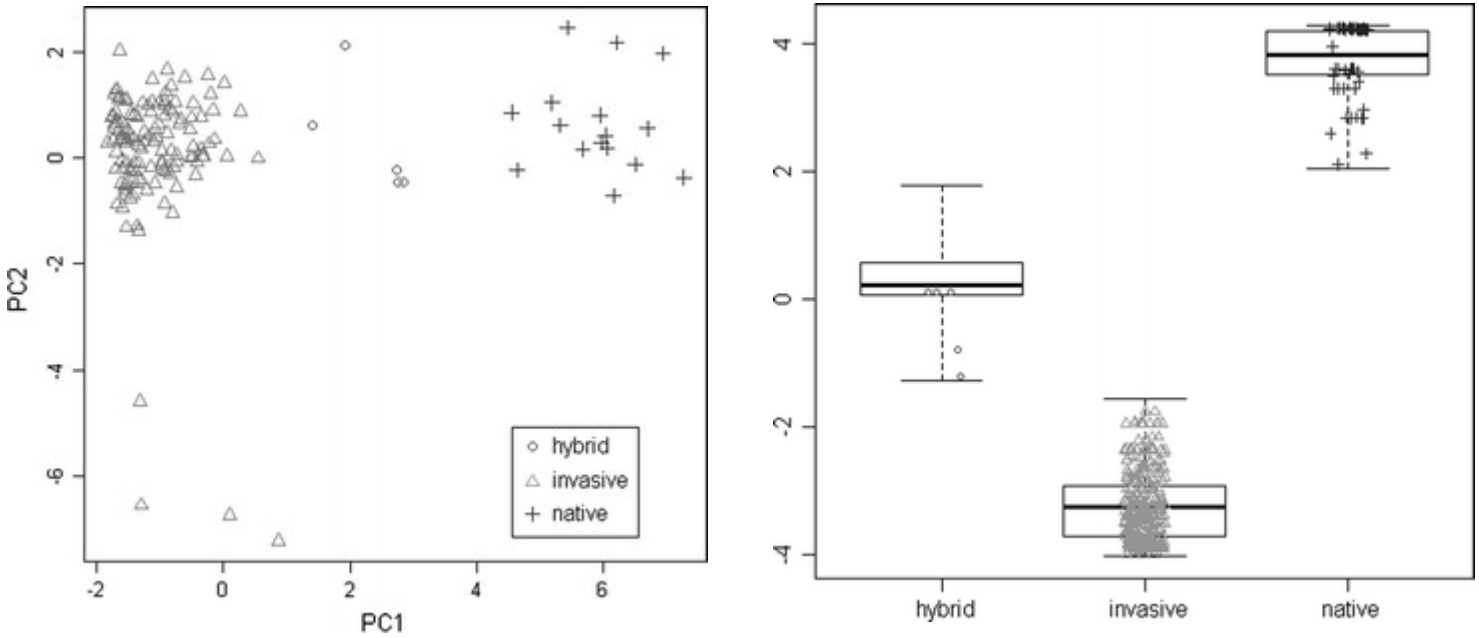


(Jombart et al. submitted)

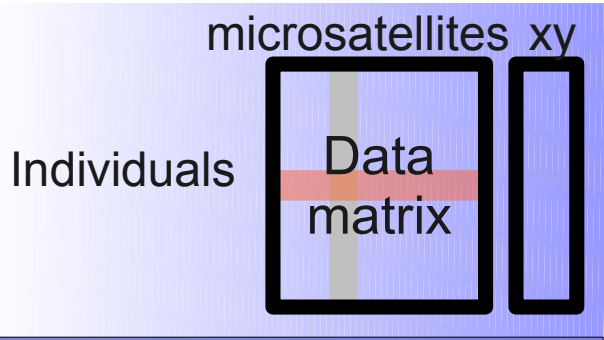


Multivariate analyses: some examples (5/6)

Studying hybridization

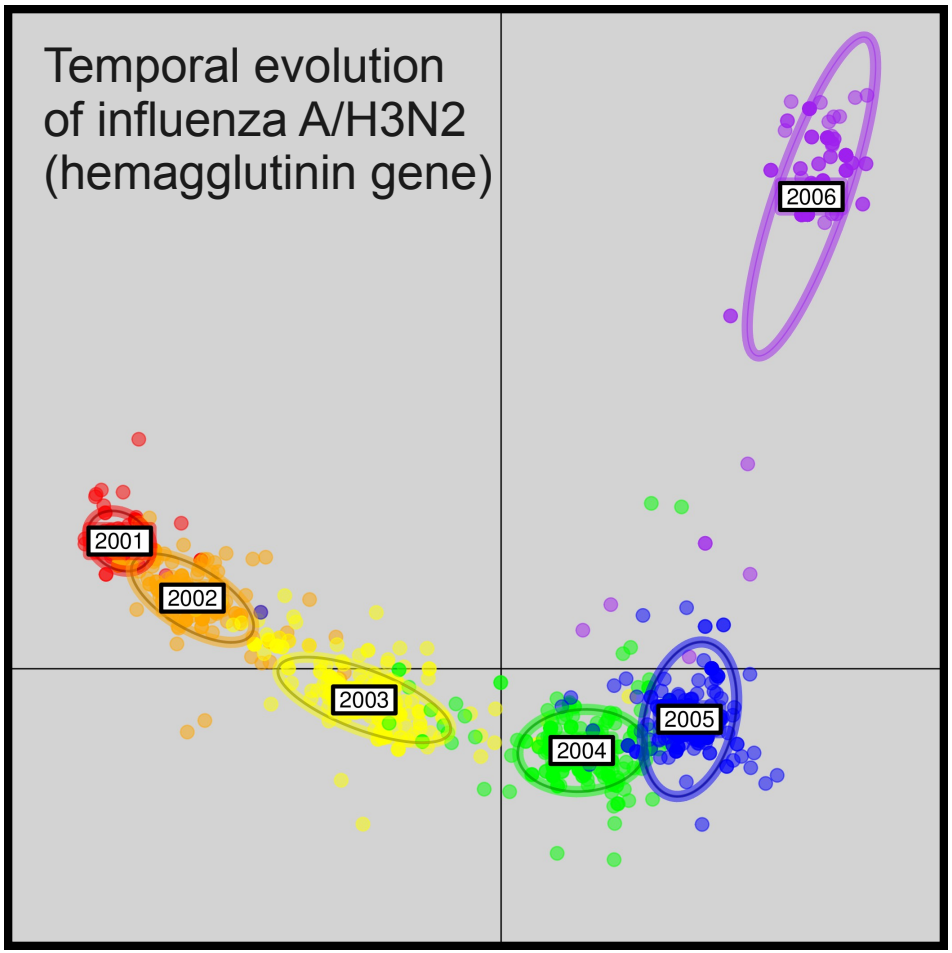


(Paul et al. 2010, *Biological Invasions*)

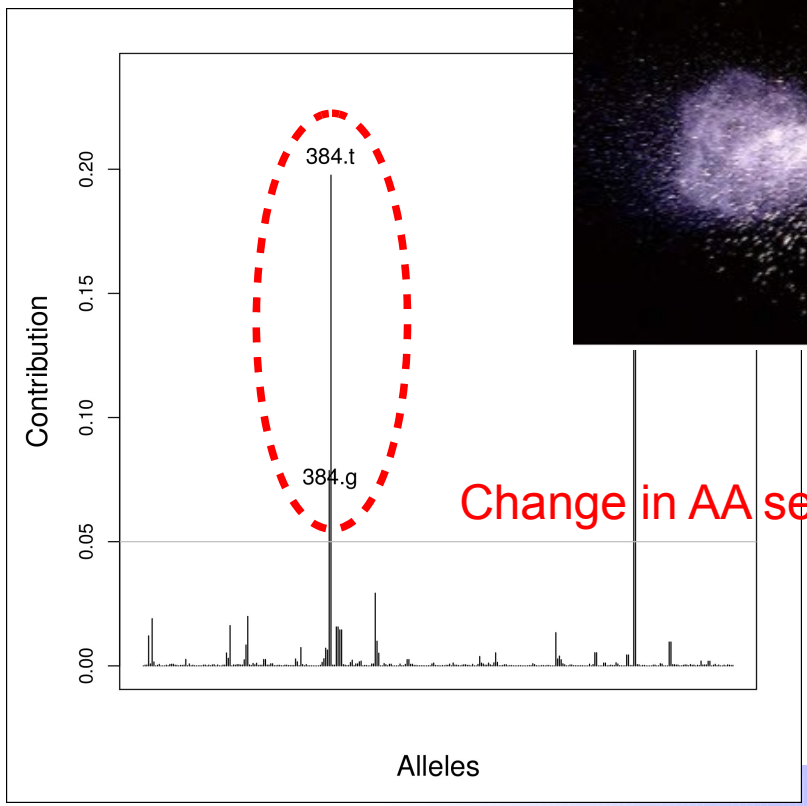


Multivariate analyses: some examples (6/6)

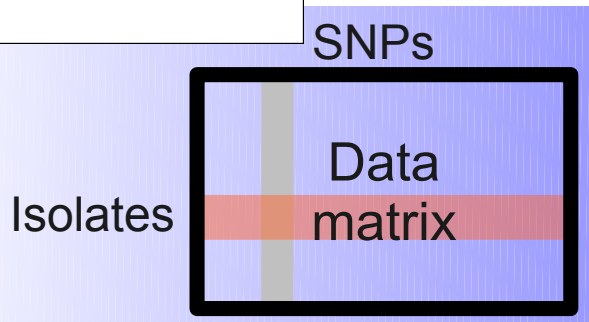
Analysing the temporal evolution of pathogens



(Jombart et al. submitted)



Change in AA sequence



A wide range of possible applications

- Many different data and questions
- Wide range of existing methods
- Adaptability to specific problems
  (e.g. non-even sampling, different ploidy levels, heterogeneous variation amongst loci)
- Datasets are growing bigger and more complex: more multivariate analyses to come...

Content

I) Introduction: data and methods

2) Multivariate analysis in a nutshell

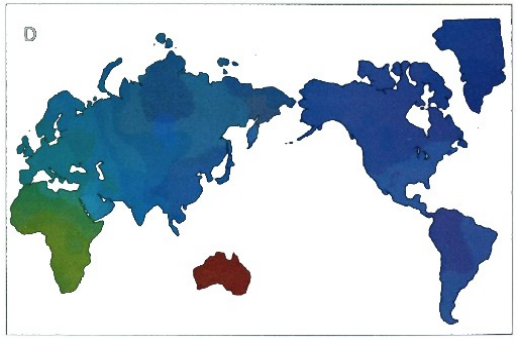
3) Some examples

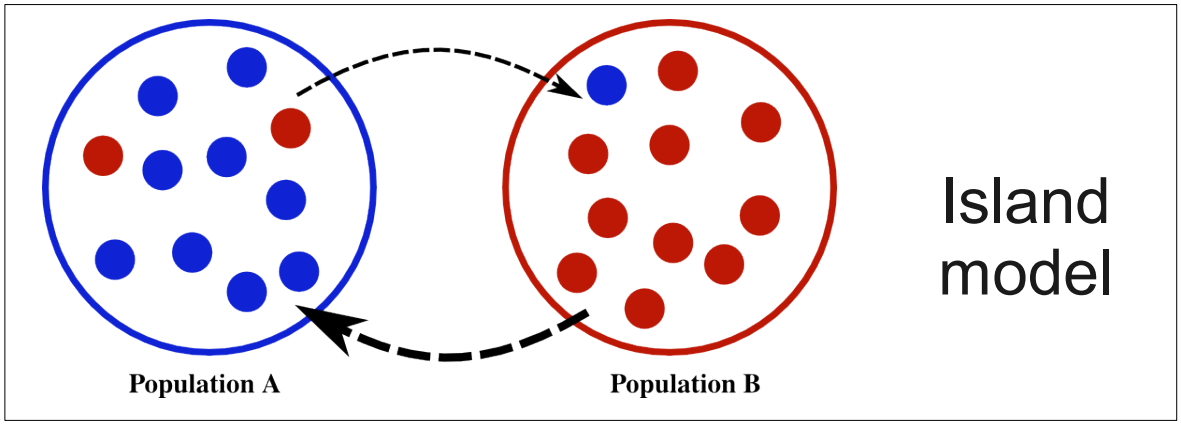
4) Spatial genetic structures

5) In practice

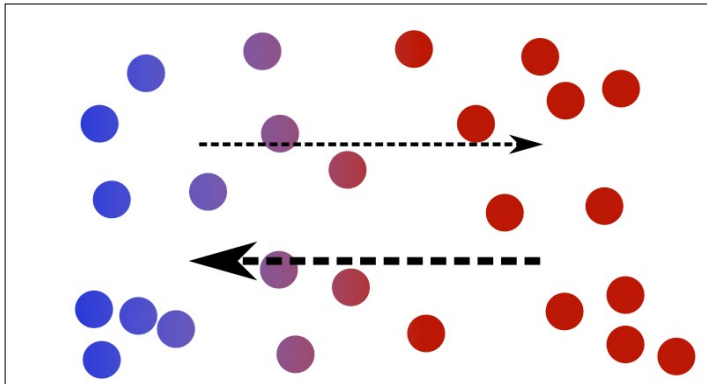
Why look for spatial genetic structures?

Most genetic models predict that genetic diversity should be spatially structured:

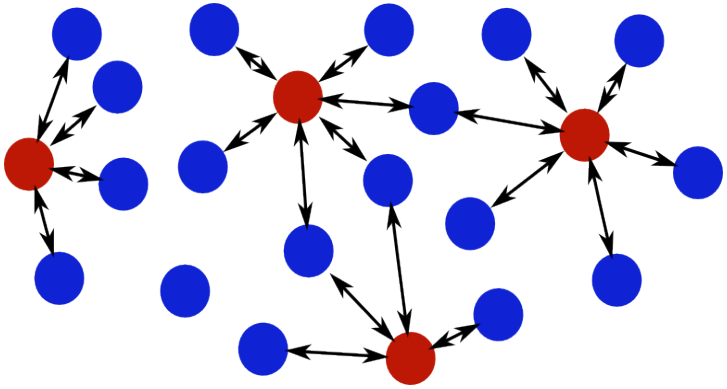




Island model

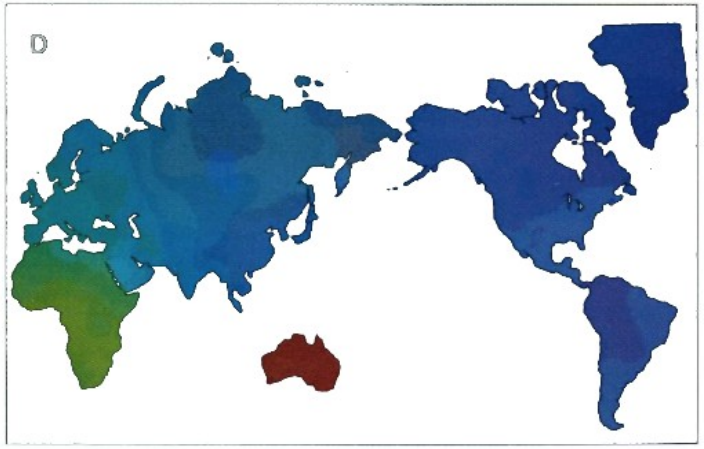


Isolation by distance



Inbreeding avoidance

Taking spatial information into account

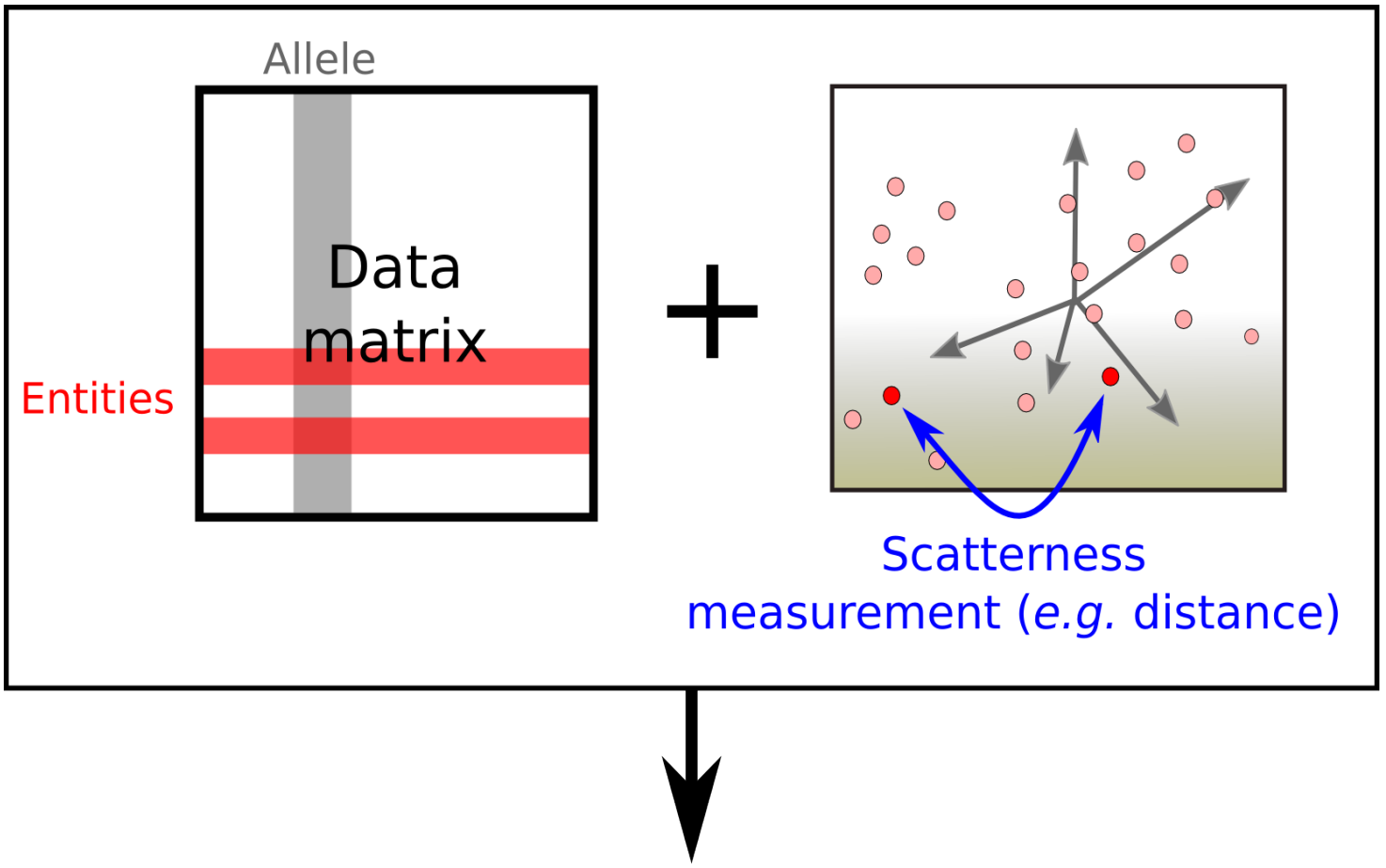


- Usual multivariate methods do not use spatial information.

- They can reveal `obvious' spatial patterns, but will overlook more subtle structures.

- To seek spatial genetic structures, we must find the part of the **genetic variability related to spatial proximity** between individuals/populations.

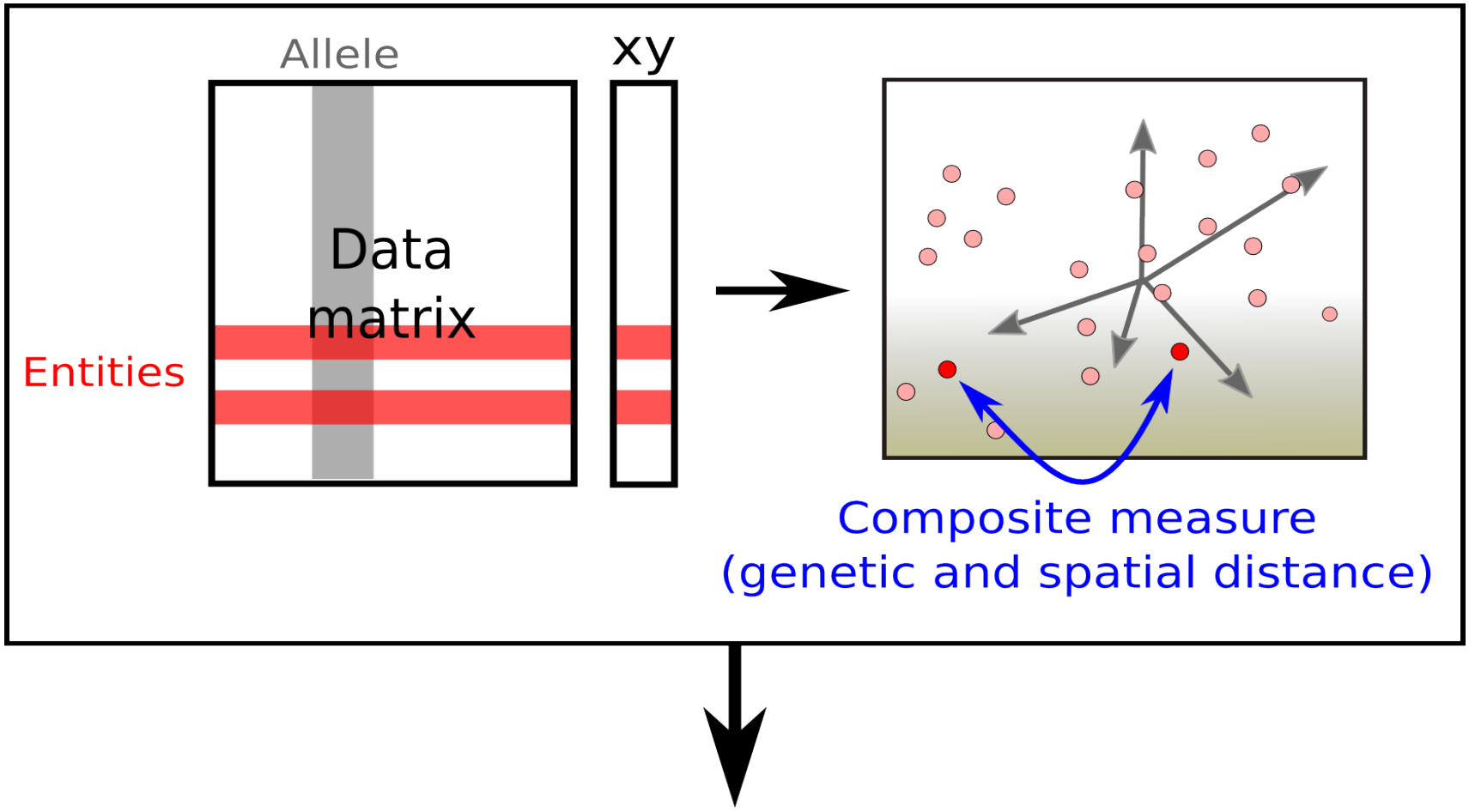
Usual multivariate analyses (recall)



Principal axes (alleles contributions)
Principal components (max. diversity, uncorrelated)

Spatial Principal Component Analysis (sPCA)

(Jombart *et al.* 2009)



Principal axes (alleles contributions)
Principal components
(spatially structured genetic variability)

Content

I) Introduction: data and methods

2) Multivariate analysis in a nutshell

3) Some examples

4) Spatial genetic structures

5) In practice

Population genetic software – the scary picture

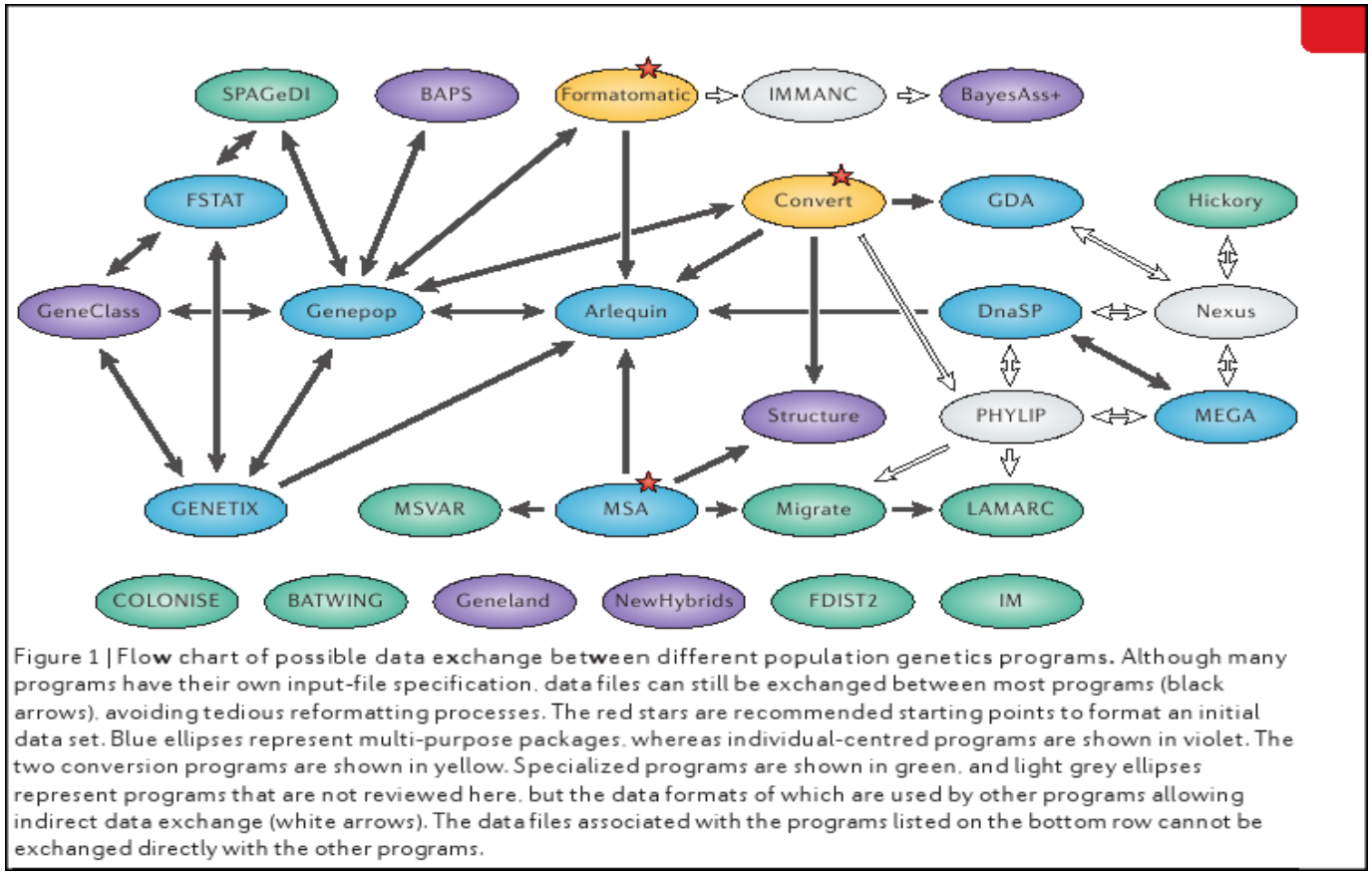


Figure 1 | Flow chart of possible data exchange between different population genetics programs. Although many programs have their own input-file specification, data files can still be exchanged between most programs (black arrows), avoiding tedious reformatting processes. The red stars are recommended starting points to format an initial data set. Blue ellipses represent multi-purpose packages, whereas individual-centred programs are shown in violet. The two conversion programs are shown in yellow. Specialized programs are shown in green, and light grey ellipses represent programs that are not reviewed here, but the data formats of which are used by other programs allowing indirect data exchange (white arrows). The data files associated with the programs listed on the bottom row cannot be exchanged directly with the other programs.

(Excoffier & Heckel 2006, *Nature Reviews Genetics*)

`In a perfect world, research teams would be able to develop analysis tools to address their specific problem, but in practice they have to make their data fit the available tools, leading to obvious discrepancies between the initial goals and the results'

Taking genetic markers into the field of (multivariate) statistics

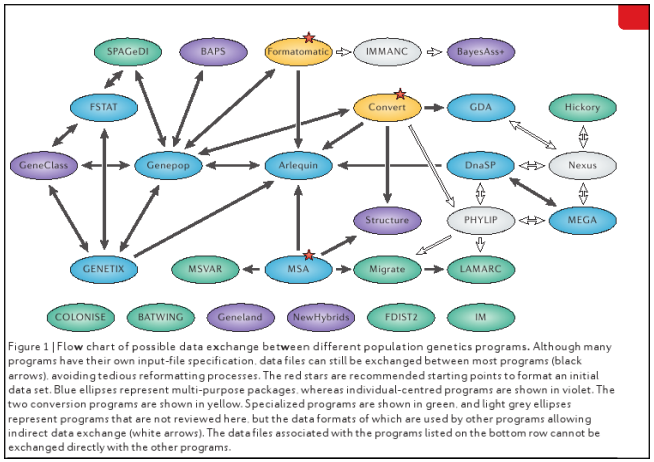


Figure 1 | Flow chart of possible data exchange between different population genetics programs. Although many programs have their own input-file specification, data files can still be exchanged between most programs (black arrows), avoiding tedious reformatting processes. The red stars are recommended starting points to format an initial data set. Blue ellipses represent multi-purpose packages, whereas individual-centred programs are shown in violet. The two conversion programs are shown in yellow. Specialized programs are shown in green, and light grey ellipses represent programs that are not reviewed here, but the data formats of which are used by other programs allowing indirect data exchange (white arrows). The data files associated with the programs listed on the bottom row cannot be exchanged directly with the other programs.

Population genetic software:

- very few multivariate methods
- no plasticity
- poor data interoperability

The R software:

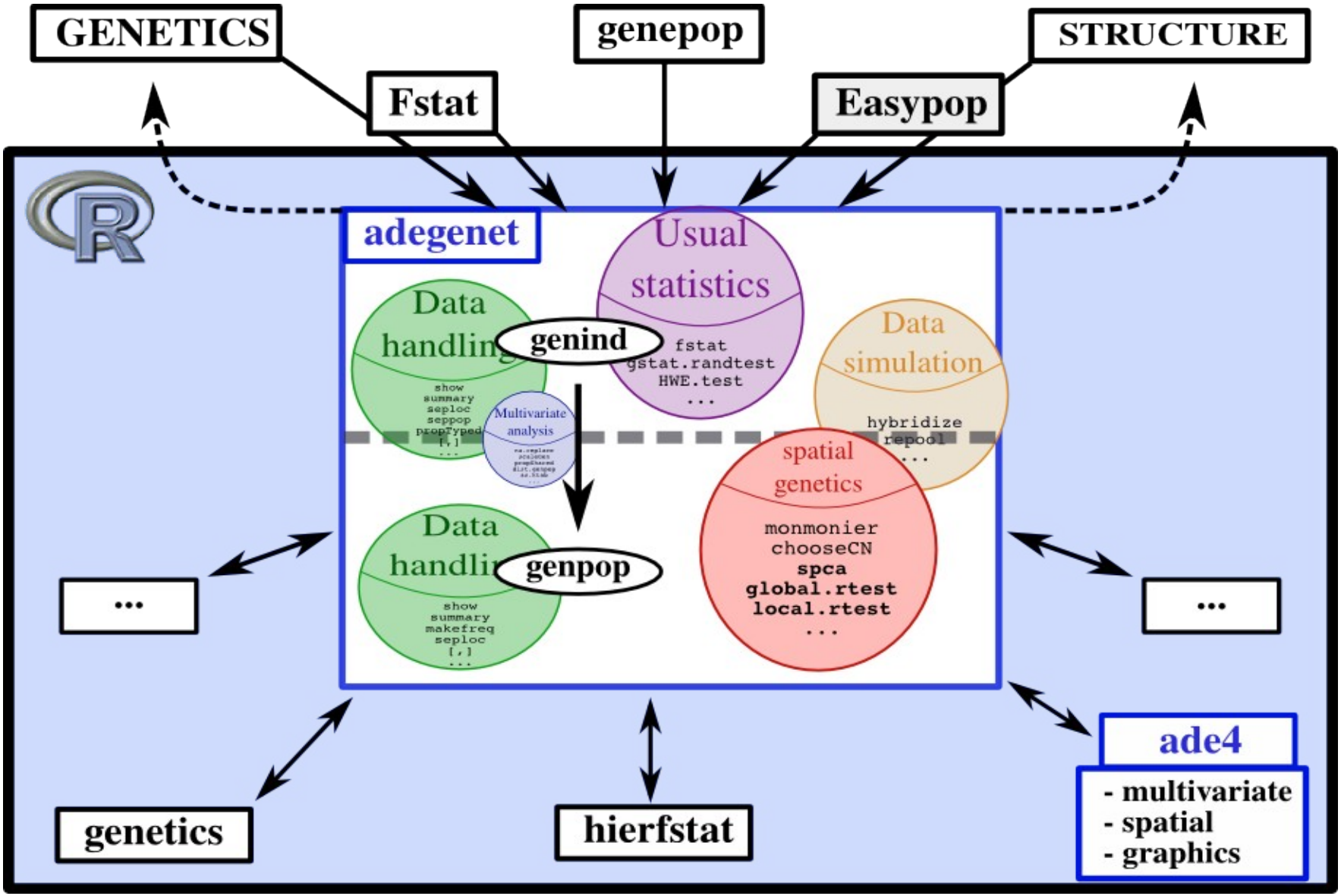
- many (most) multivariate methods
- total plasticity
- tons of statistical methods (tests, modelling, Monte-Carlo)
- great graphics
- great interoperability (*e.g.*, GIS)
- programming language
- free software

The *adegenet* package for R (1/3)
(Jombart 2008)

Purpose:

- take genetic markers into a suitable format
- adapt multivariate methods to genetic markers
- provide advanced data handling
- provide standard population genetics tools
- implement novel methods (*e.g.*, sPCA, DAPC, seqTrack)

The *adegenet* package for R (2/3)



The *adegenet* package for R (3/3)

Where to get information:

- reference: Jombart (2008) Bioinformatics 24: 1403-1405
- *adegenet* website: `http://adegenet.r-forge.r-project.org/` (tutorials and manuals for download)
- *adegenet* forum: `adegenet-forum@lists.r-forge.r-project.org`

- Here and now!