Multivariate analysis of genetic markers as a tool to explore the genetic diversity

Thibaut Jombart – Imperial College London
tjombart@imperial.ac.uk
Why analysing genetic markers?

**Biological processes**
(demography, dispersal, selection)

**Spatial genetic structures**
Why analysing genetic markers?

Biological processes
(demography, dispersal, selection)

Spatial genetic structures

Identify genetic patterns to infer biological processes structuring the biodiversity
Genetic marker data

- Individual
- Population

Data matrix

- Allele frequency
- Marker presence/absence

Variables, Descriptors

(Observations, Biological entities)

Summarise the genetic diversity among individuals / populations
Multivariate analysis – rationale (1/2)

- Biological entities
- Alleles

Multivariate analysis of genetic markers
Multivariate analysis – rationale (1/2)

[Diagram of biological entities and alleles, showing principal axis and loadings.]
Summarise the genetic diversity among individuals / populations

Multivariate analysis – rationale (1/2)

\[ u_a a + u_b b + u_c c + u_d d + u_e e = \]

(Biological entity)

(Loadings (Principal axis))

(Synthetic variable (Principal component))
Multivariate analysis – rationale (2/2)

Data matrix

Entities

Allele

Scatterness measurement (e.g. distance)

Principal axes (alleles contributions)
Principal components (max. diversity, uncorrelated)
Multivariate analyses: a large family

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)
What to do with principal components? (1/3)

Get a picture of genetic diversity

(Cavalli-Sforza 1966)
What to do with principal components? (2/3)

Use as response/predictor in models

\[
\text{Genetic features} + \text{Covariates} + \text{Environmental variables} + \text{Residuals} = \text{Data matrix}
\]

(Mulley et al. 1979)
What to do with principal components? (3/3)

Make geographical maps of the genetic differentiation

(Cavalli-Sforza et al. 1993)
Why look for spatial genetic structures?

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

- **Island model**
  - Population A
  - Population B

- **Isolation by distance**
  - Population A
  - Population B

- **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.
Spatial Principal Component Analysis (sPCA)
(Jombart et al. 2009)

- **Data matrix**
  - Entities
  - Allele

- **Composite measure**
  - Genetic and spatial distance

- **Principal axes**
  - Alleles contributions
  - Principal components
    - Spatially structured genetic variability
`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

Population genetic software:
• very few multivariate methods
• no plasticity
• poor data interoperability

The R software:
• many multivariate methods
• total plasticity
• tons of statistic methods (tests, modeling, Monte-Carlo)
• great graphics
• great interoperability (e.g., GIS)
• programming language
• free software
Purpose:

- take genetic markers into a suitable format
- adapt multivariate methods to genetic markers
- provide advanced data handling
- implement novel methods (e.g., sPCA)
The adegenet package for R (2/3)

Multivariate analysis of genetic markers  T. Jombart - 09/09/09
Where to get information:

- **reference:** Jombart (2008) Bioinformatics 24: 1403-1405

- **adegenet website:** [http://adegenet.r-forge.r-project.org/](http://adegenet.r-forge.r-project.org/)

- **adegenet forum:** [adegenet-forum@lists.r-forge.r-project.org](mailto:adegenet-forum@lists.r-forge.r-project.org)

- Here and now!