Introduction to multivariate analysis — applications in genomics —

Thibaut Jombart
(`t.jombart@imperial.ac.uk`)

MRC Centre for Outbreak Analysis and Modelling
Imperial College London

MSc "*Modern epidemiology*"
22-03-2013

Outline

Multivariate analysis in a nutshell

Applications to genomic data

Genetic diversity of pathogen populations

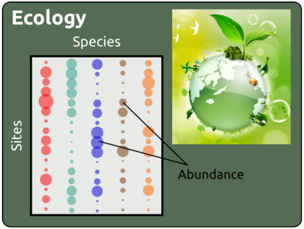
Outline

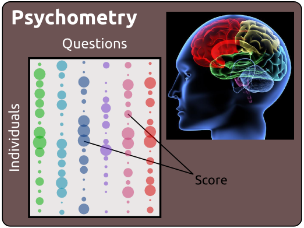
Multivariate analysis in a nutshell

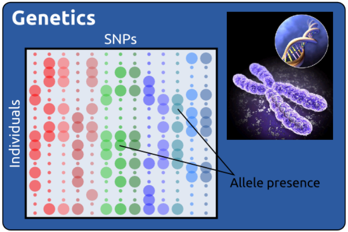
Applications to genomic data

Genetic diversity of pathogen populations

Multivariate data: some examples

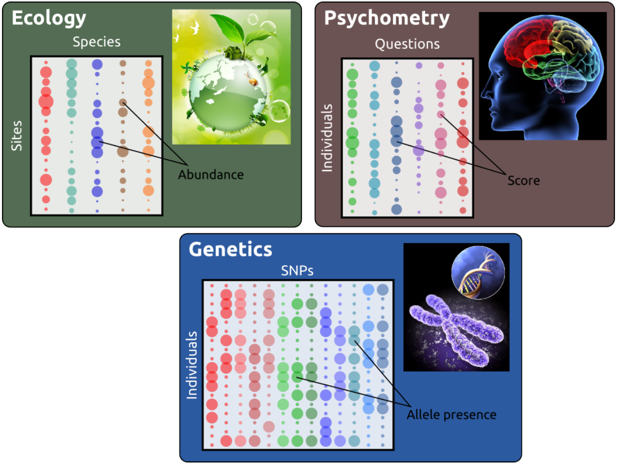






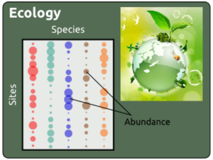
Association between individuals? Correlations between variables?

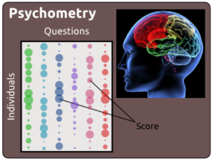
Multivariate data: some examples

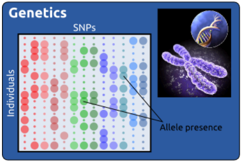


Association between individuals? Correlations between variables?

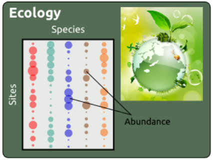
Multivariate analysis to summarize diversity

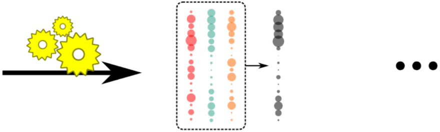




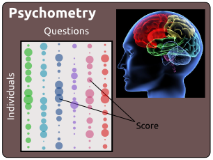


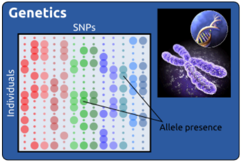
Multivariate analysis to summarize diversity



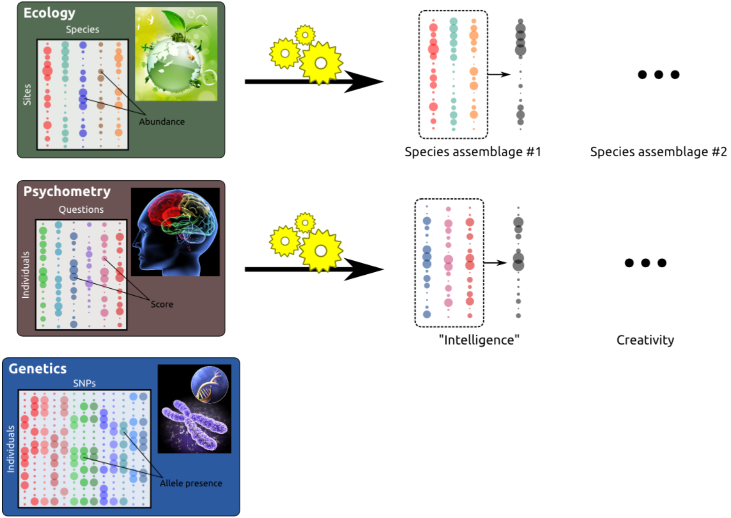


Species assemblage #1 Species assemblage #2

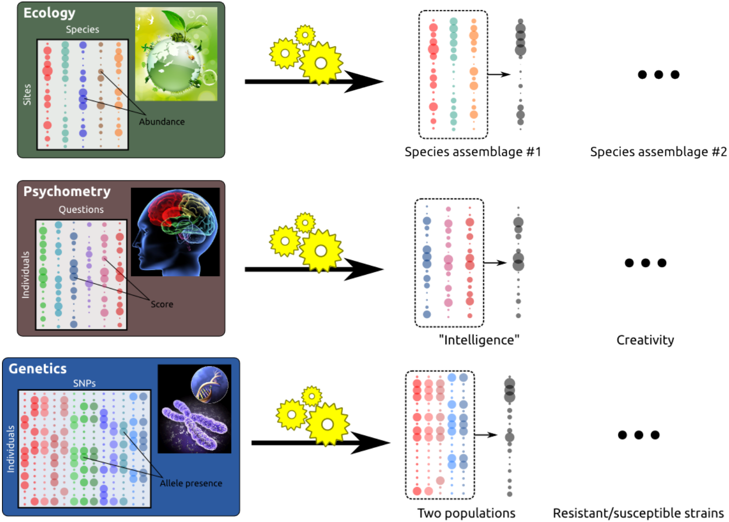




Multivariate analysis to summarize diversity



Multivariate analysis to summarize diversity



Multivariate analysis: an overview

Multivariate analysis, a.k.a:

- "*dimension reduction techniques*"
- "*ordinations in reduced space*"
- "*factorial methods*"

Purposes:

- summarize diversity amongst observations
- summarize correlations between variables

Multivariate analysis: an overview

Multivariate analysis, a.k.a:

- "*dimension reduction techniques*"
- "*ordinations in reduced space*"
- "*factorial methods*"

Purposes:

- summarize diversity amongst observations
- summarize correlations between variables

Most common methods

Differences lie in input data:

- quantitative/binary variables: *Principal Component Analysis* (PCA)
- 2 categorical variables: *Correspondance Analysis* (CA)
- >2 categorical variables: *Multiple Correspondance Analysis* (MCA)
- Euclidean distance matrix: *Principal Coordinates Analysis* (PCoA) / *Metric Multidimensional Scaling* (MDS)

Many other methods for ≥ 2 data tables, spatial analysis, phylogenetic analysis, etc.

Most common methods

Differences lie in input data:

- quantitative/binary variables: *Principal Component Analysis* (PCA)
- 2 categorical variables: *Correspondance Analysis* (CA)
- >2 categorical variables: *Multiple Correspondance Analysis* (MCA)
- Euclidean distance matrix: *Principal Coordinates Analysis* (PCoA) / *Metric Multidimensional Scaling* (MDS)

Many other methods for ≥ 2 data tables, spatial analysis, phylogenetic analysis, etc.

Most common methods

Differences lie in input data:

- quantitative/binary variables: *Principal Component Analysis* (PCA)
- 2 categorical variables: *Correspondance Analysis* (CA)
- $>$2 categorical variables: *Multiple Correspondance Analysis* (MCA)
- Euclidean distance matrix: *Principal Coordinates Analysis* (PCoA) / *Metric Multidimensional Scaling* (MDS)

Many other methods for ≥ 2 data tables, spatial analysis, phylogenetic analysis, etc.

Most common methods

Differences lie in input data:

- quantitative/binary variables: *Principal Component Analysis* (PCA)
- 2 categorical variables: *Correspondance Analysis* (CA)
- $>$2 categorical variables: *Multiple Correspondance Analysis* (MCA)
- Euclidean distance matrix: *Principal Coordinates Analysis* (PCoA) / *Metric Multidimensional Scaling* (MDS)

Many other methods for ≥ 2 data tables, spatial analysis, phylogenetic analysis, etc.

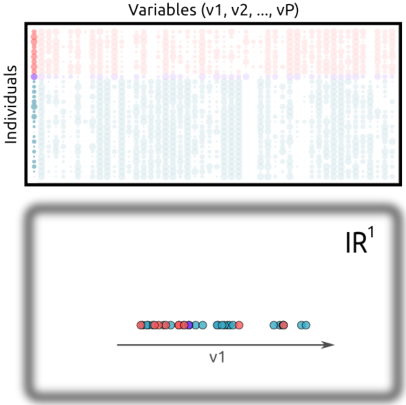
Most common methods

Differences lie in input data:

- quantitative/binary variables: *Principal Component Analysis* (PCA)
- 2 categorical variables: *Correspondance Analysis* (CA)
- $>$2 categorical variables: *Multiple Correspondance Analysis* (MCA)
- Euclidean distance matrix: *Principal Coordinates Analysis* (PCoA) / *Metric Multidimensional Scaling* (MDS)

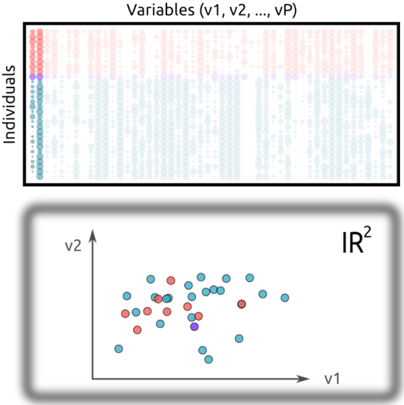
Many other methods for ≥ 2 data tables, spatial analysis, phylogenetic analysis, etc.

1 dimension, 2 dimensions, P dimensions



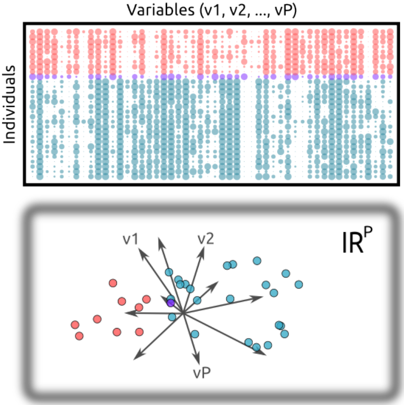
Need to find most informative directions in a P-dimensional space.

1 dimension, 2 dimensions, P dimensions



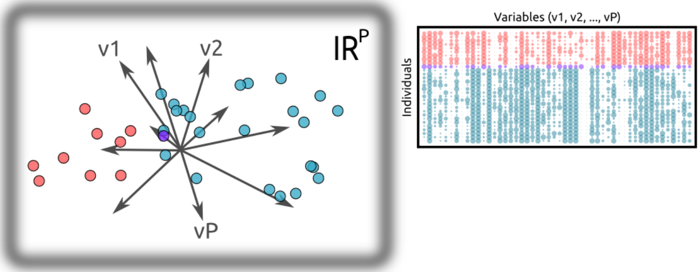
Need to find most informative directions in a P-dimensional space.

1 dimension, 2 dimensions, P dimensions



Need to find most informative directions in a P-dimensional space.

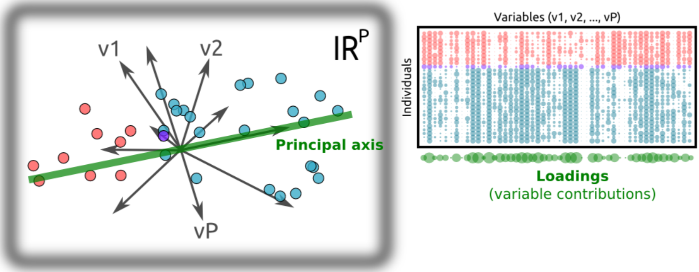
Reducing P dimensions into 1



- $\mathbf{X} \in \mathbb{R}^{N \times P}$; $\mathbf{X} = [\mathbf{x}_1 | \dots | \mathbf{x}_P]$: data matrix
- $\mathbf{u} \in \mathbb{R}^{P}$; $\mathbf{u} = [u_1, \dots, u_P]$: **principal axis** $(\|\mathbf{u}\|^2 = \sum_{j=1}^{P} u_j^2 = 1)$
- $\mathbf{v} \in \mathbb{R}^{N}$; $\mathbf{v} = \mathbf{X}\mathbf{u} = \sum_{j=1}^{P} u_j \mathbf{x}_j$: **principal component**

→ find $\mathbf{u}$ so that $\frac{1}{N}\|\mathbf{v}\|^2 = \text{var}(\mathbf{v})$ is maximum.

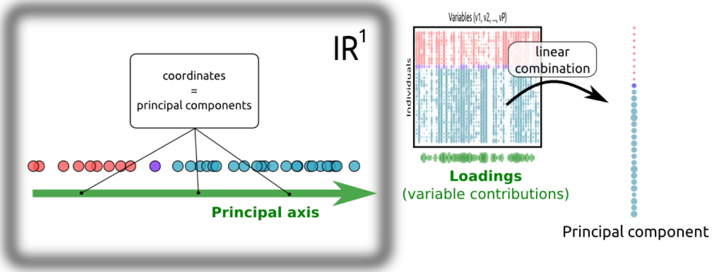
Reducing P dimensions into 1



- $\mathbf{X} \in \mathbb{R}^{N \times P}$; $\mathbf{X} = [\mathbf{x}_1 | \dots | \mathbf{x}_P]$: data matrix
- $\mathbf{u} \in \mathbb{R}^P$; $\mathbf{u} = [u_1, \dots, u_P]$: **principal axis** $(\|\mathbf{u}\|^2 = \sum_{j=1}^{P} u_j^2 = 1)$
- $\mathbf{v} \in \mathbb{R}^N$; $\mathbf{v} = \mathbf{X}\mathbf{u} = \sum_{j=1}^{P} u_j \mathbf{x}_j$: **principal component**

→ find $\mathbf{u}$ so that $\frac{1}{N}\|\mathbf{v}\|^2 = \text{var}(\mathbf{v})$ is maximum.

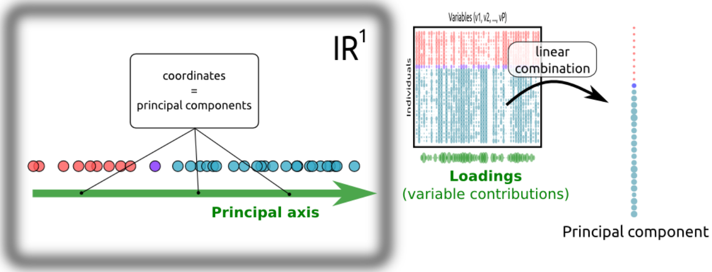
Reducing P dimensions into 1



- $\mathbf{X} \in \mathbb{R}^{N \times P}$; $\mathbf{X} = [\mathbf{x}_1 | \dots | \mathbf{x}_P]$: data matrix
- $\mathbf{u} \in \mathbb{R}^P$; $\mathbf{u} = [u_1, \dots, u_P]$: **principal axis** $(\|\mathbf{u}\|^2 = \sum_{j=1}^P u_j^2 = 1)$
- $\mathbf{v} \in \mathbb{R}^N$; $\mathbf{v} = \mathbf{X}\mathbf{u} = \sum_{j=1}^P u_j \mathbf{x}_j$: **principal component**

→ find $\mathbf{u}$ so that $\frac{1}{N}\|\mathbf{v}\|^2 = \text{var}(\mathbf{v})$ is maximum.

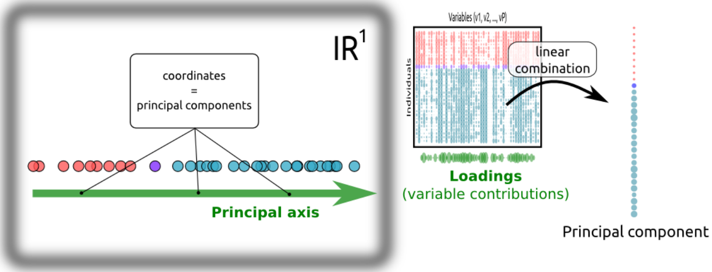
Reducing P dimensions into 1



- $\mathbf{X} \in \mathbb{R}^{N \times P}$; $\mathbf{X} = [\mathbf{x}_1 | \dots | \mathbf{x}_P]$: data matrix
- $\mathbf{u} \in \mathbb{R}^{P}$; $\mathbf{u} = [u_1, \dots, u_P]$: **principal axis** $(\|\mathbf{u}\|^2 = \sum_{j=1}^{P} u_j^2 = 1)$
- $\mathbf{v} \in \mathbb{R}^{N}$; $\mathbf{v} = \mathbf{X}\mathbf{u} = \sum_{j=1}^{P} u_j \mathbf{x}_j$: **principal component**

$\rightarrow$ find $\mathbf{u}$ so that $\frac{1}{N}\|\mathbf{v}\|^2 = \text{var}(\mathbf{v})$ is maximum.

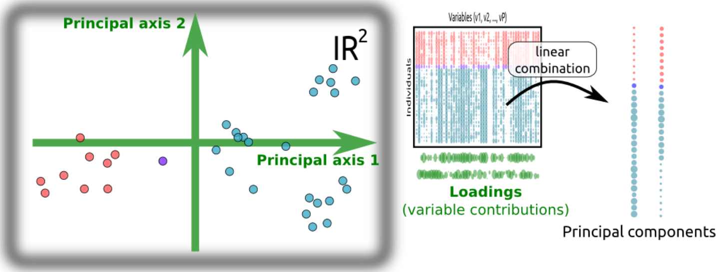
Keeping more than one principal component



- $\mathbf{u}_1$ and $\mathbf{v}_1$: **1st principal axis and component**
- $\mathbf{u}_2$ and $\mathbf{v}_2$: **2nd principal axis and component**

→ constraint: $\mathbf{u}_1 \perp \mathbf{u}_2$ $(\Longleftrightarrow \text{cor}(\mathbf{v}_1, \mathbf{v}_2) = 0)$
→ find $\mathbf{u}_2$ so that $\frac{1}{N}\|\mathbf{v}_2\|^2 = \text{var}(\mathbf{v}_2)$ is maximum

Keeping more than one principal component

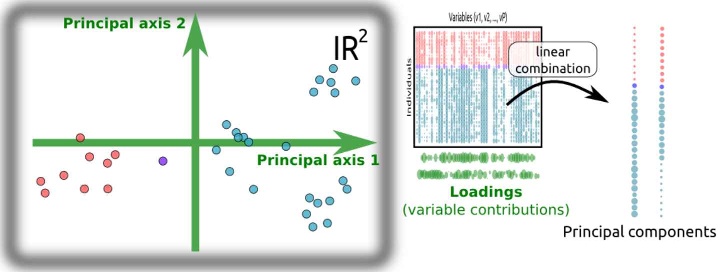


- $\mathbf{u}_1$ and $\mathbf{v}_1$: **1st principal axis and component**
- $\mathbf{u}_2$ and $\mathbf{v}_2$: **2nd principal axis and component**

→ constraint: $\mathbf{u}_1 \perp \mathbf{u}_2$ ($\Longleftrightarrow$ cor$(\mathbf{v}_1, \mathbf{v}_2) = 0$)

→ find $\mathbf{u}_2$ so that $\frac{1}{N}\|\mathbf{v}_2\|^2 =$ var$(\mathbf{v}_2)$ is maximum

Keeping more than one principal component

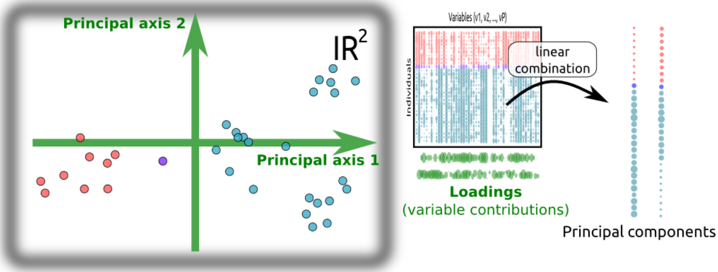


- $\mathbf{u}_1$ and $\mathbf{v}_1$: **1st principal axis and component**
- $\mathbf{u}_2$ and $\mathbf{v}_2$: **2nd principal axis and component**

→ constraint: $\mathbf{u}_1 \perp \mathbf{u}_2$ ($\Longleftrightarrow$ cor$(\mathbf{v}_1, \mathbf{v}_2) = 0$)

→ find $\mathbf{u}_2$ so that $\frac{1}{N}\|\mathbf{v}_2\|^2 =$ var$(\mathbf{v}_2)$ is maximum

Keeping more than one principal component

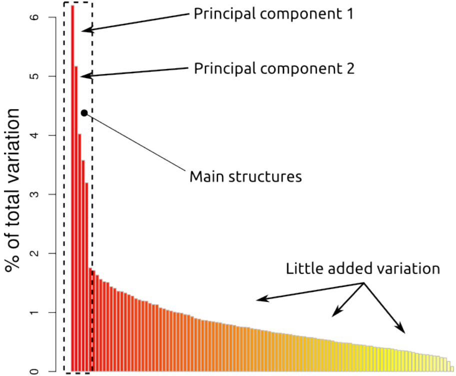


- $\mathbf{u}_1$ and $\mathbf{v}_1$: **1st principal axis and component**
- $\mathbf{u}_2$ and $\mathbf{v}_2$: **2nd principal axis and component**

→ constraint: $\mathbf{u}_1 \perp \mathbf{u}_2$ ($\Longleftrightarrow$ cor$(\mathbf{v}_1, \mathbf{v}_2) = 0$)
→ find $\mathbf{u}_2$ so that $\frac{1}{N}\|\mathbf{v}_2\|^2 =$ var$(\mathbf{v}_2)$ is maximum

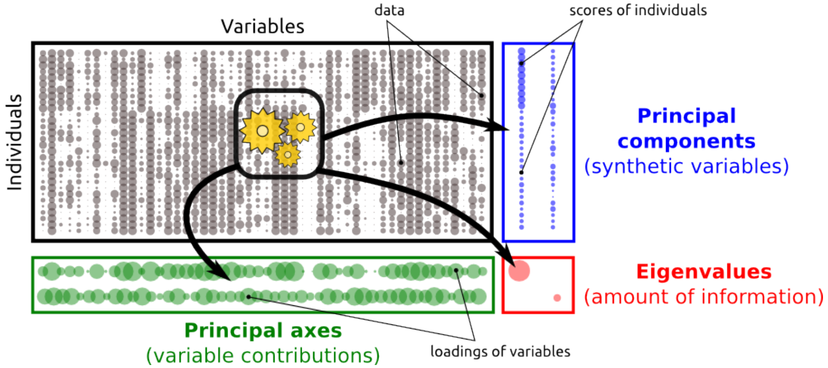
How many principal components to retain?

Choice based on "**screeplot**": barplot of eigenvalues



Retain only "significant" structures... but not trivial ones.

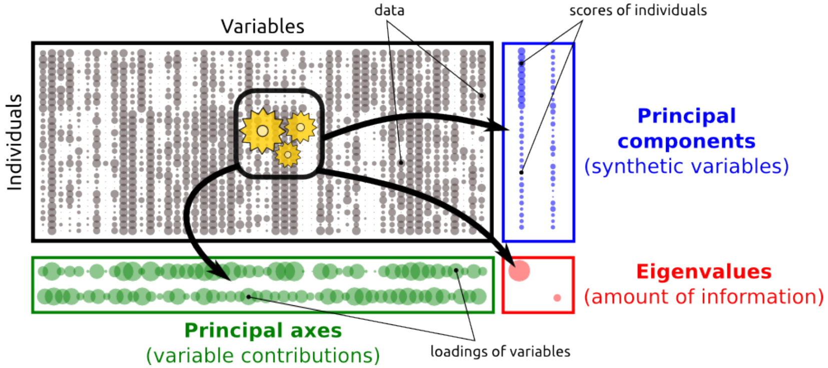
Outputs of multivariate analyses: an overview



Main outputs:

- **principal components**: diversity amongst individuals
- **principal axes**: nature of the structures
- **eigenvalues**: magnitude of structures

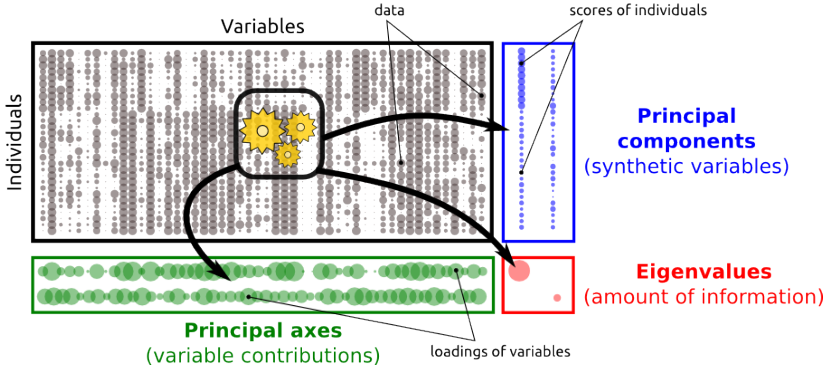
Outputs of multivariate analyses: an overview



Main outputs:

- **principal components**: diversity amongst individuals
- **principal axes**: nature of the structures
- **eigenvalues**: magnitude of structures

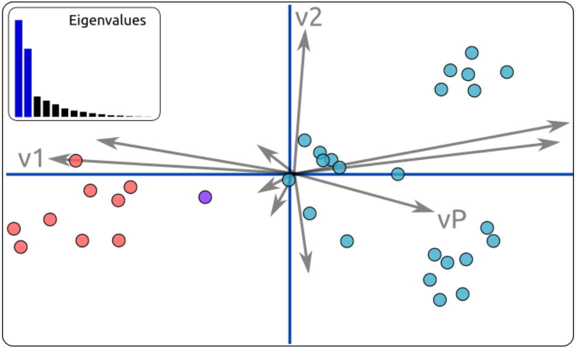
Outputs of multivariate analyses: an overview



Main outputs:

- **principal components**: diversity amongst individuals
- **principal axes**: nature of the structures
- **eigenvalues**: magnitude of structures

Usual summary of an analysis: the biplot



Biplot: principal components (points) + loadings (arrows)

- groups of individuals
- discriminating variables (longest arrows)
- magnitude of the structures

Multivariate analysis in a nutshell

- **variety of methods** for different types of variables
- **principal components** (PCs) summarize diversity
- **variable loadings** identify discriminating variables
- other uses of PCs: **maps** (spatial structures), **models** (response variables or predictors), ...

Multivariate analysis in a nutshell

- **variety of methods** for different types of variables
- **principal components** (PCs) summarize diversity
- **variable loadings** identify discriminating variables
- other uses of PCs: **maps** (spatial structures), **models** (response variables or predictors), ...

Multivariate analysis in a nutshell

- **variety of methods** for different types of variables
- **principal components** (PCs) summarize diversity
- **variable loadings** identify discriminating variables
- other uses of PCs: **maps** (spatial structures), **models** (response variables or predictors), ...

Multivariate analysis in a nutshell

- **variety of methods** for different types of variables
- **principal components** (PCs) summarize diversity
- **variable loadings** identify discriminating variables
- other uses of PCs: **maps** (spatial structures), **models** (response variables or predictors), ...

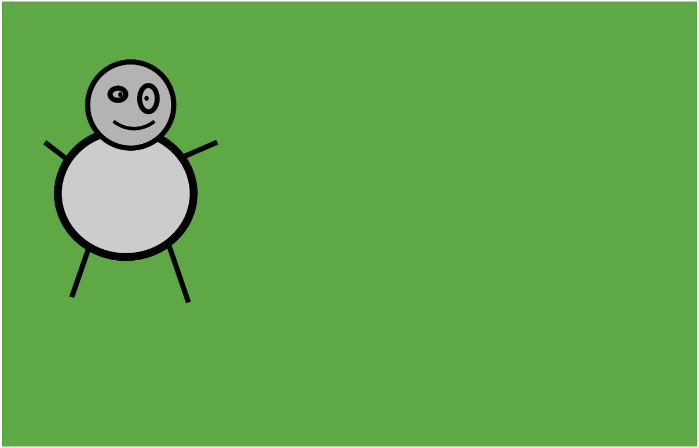
Outline

Multivariate analysis in a nutshell

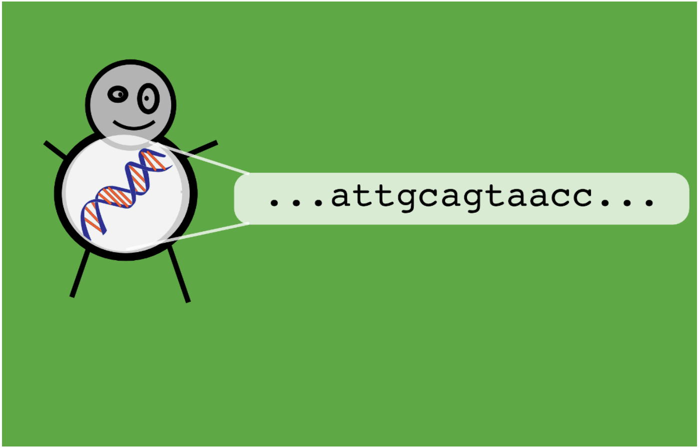
Applications to genomic data

Genetic diversity of pathogen populations

From DNA sequences to patterns of biological diversity

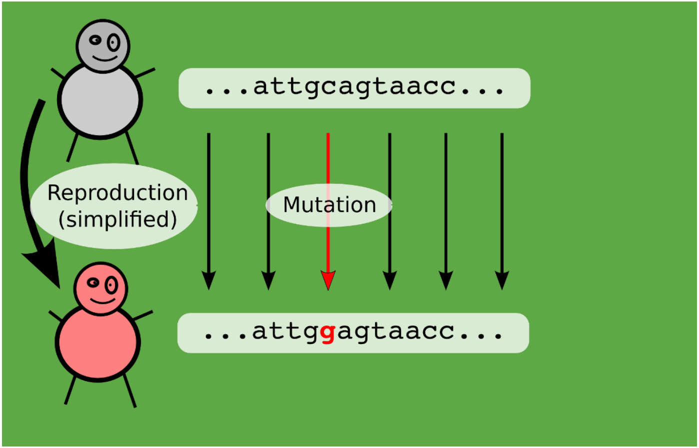


From DNA sequences to patterns of biological diversity

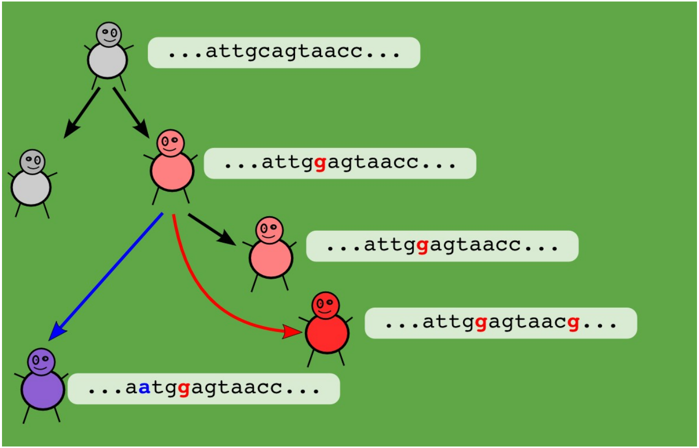


From DNA sequences to patterns of biological diversity

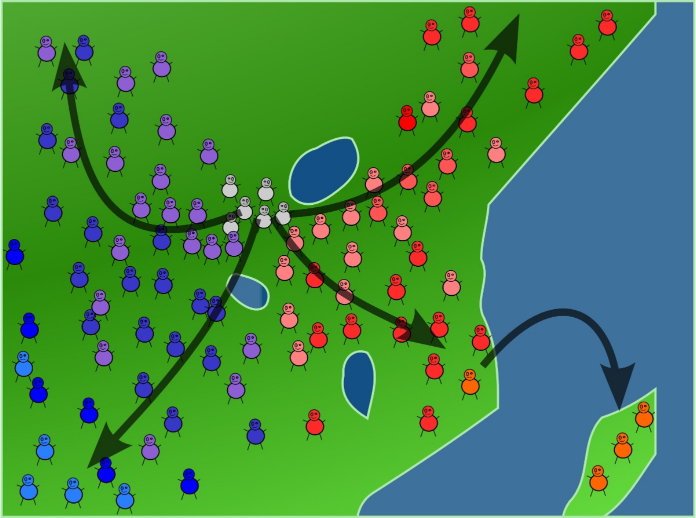
From DNA sequences to patterns of biological diversity



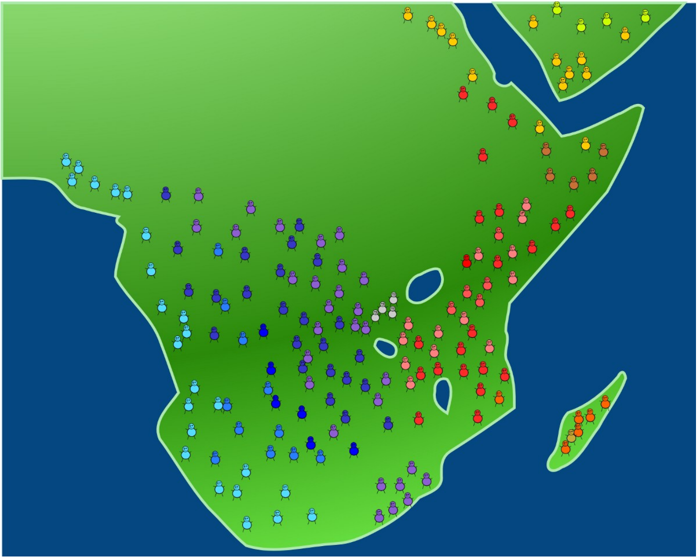
From DNA sequences to patterns of biological diversity



From DNA sequences to patterns of biological diversity

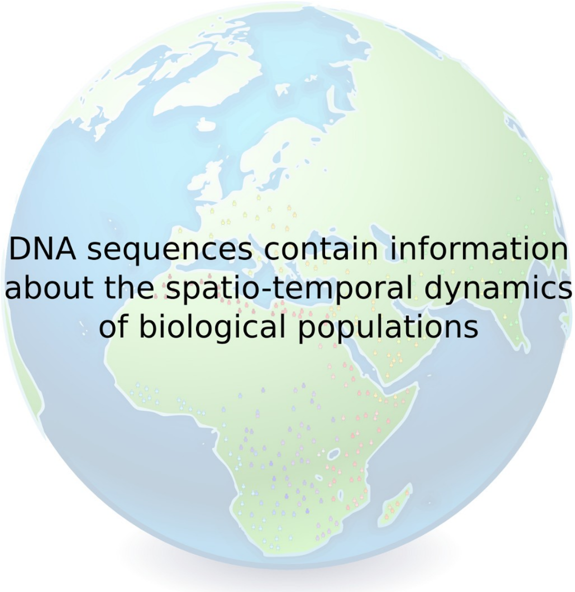


From DNA sequences to patterns of biological diversity



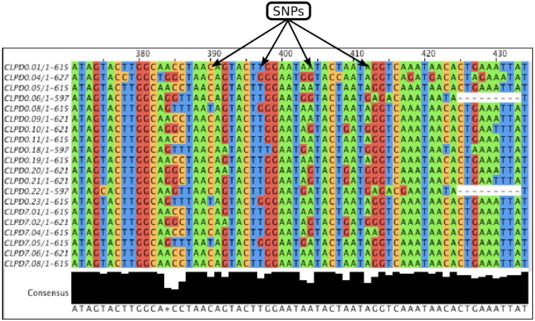
From DNA sequences to patterns of biological diversity

From DNA sequences to patterns of biological diversity



DNA sequences contain information about the spatio-temporal dynamics of biological populations

DNA sequences: a rich source of information

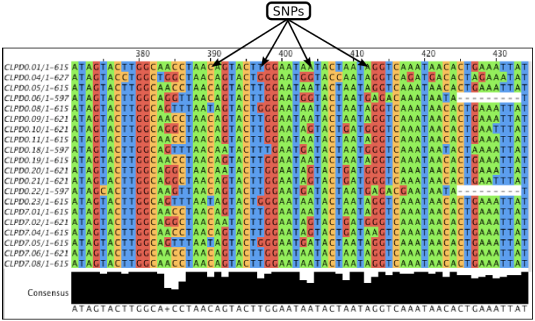


```
                       380       390       400       410       420       430
CLPD0.01/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.04/1-627 ATAGTACCTGGCTGGCTAACAGTACTGGGAATGGTACCAATAGGTCAGATGACACTAGAAATAT
CLPD0.05/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.06/1-597 ATAGTACTTGGCAGGTTAACAGTACTTGGAATGGTACTAATGAGACAAATAATA---------T
CLPD0.08/1-615 ATAGTACTTGGCAGTTTAATAGTACTGGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.09/1-621 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.10/1-621 ATAGTACTTGGCAGGCTAACAGTACTTGGAATAGTACTGATGGGTCAAATAACACTGAATTTAT
CLPD0.11/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.18/1-597 ATAGTACTTGGCAGTTTAACAATACTTTGAATGATACTAATGGGTCAAATAATACTAAAATTAT
CLPD0.19/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.20/1-621 ATAGTACTTGGCAGGCTAACAATACTTGGAATAGTACTGATGGGTCAAATAACACTGAAATTAT
CLPD0.21/1-621 ATAGTACTTGGCAGGCTAACAGTACTTGGAATAGTACTGATGGGTCAAATAACACTGAATTTAT
CLPD0.22/1-597 ATAGCACTTGGCAAGTTAACAGTACTTGGAATGATACTAATGAGACGAATAATA---------T
CLPD0.23/1-615 ATAGTACTTGGCAGTTTAATAGTACTGGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD7.01/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD7.02/1-621 ATAGTACTTGGCAGGCTAACAATACTTGGAATAGTACTGATGGGTCAAATAACACTGAAATTAT
CLPD7.04/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAGTACTGATAAGTCAAATAACACTGAAATTAT
CLPD7.05/1-615 ATAGTACTTGGCAGTTTAATAGTACTGGGAATGATACTAATAGGTCAAATAACACTGAAATTAT
CLPD7.06/1-621 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD7.08/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT

     Consensus
               ATAGTACTTGGCA+CCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
```

- hundreds/thousands individuals
- up to millions of single nucleotide polymorphism (SNPs)

⇒ **Multivariate analysis use to summarize genetic diversity.**

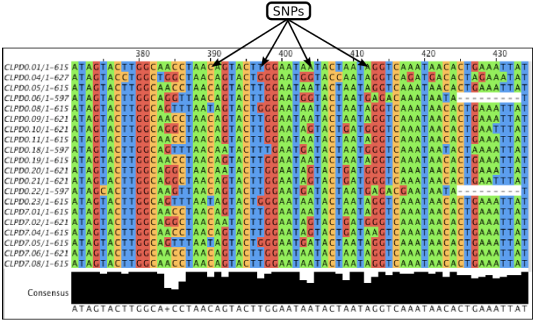
DNA sequences: a rich source of information



- hundreds/thousands individuals
- up to millions of single nucleotide polymorphism (SNPs)

⇒ **Multivariate analysis use to summarize genetic diversity.**

DNA sequences: a rich source of information

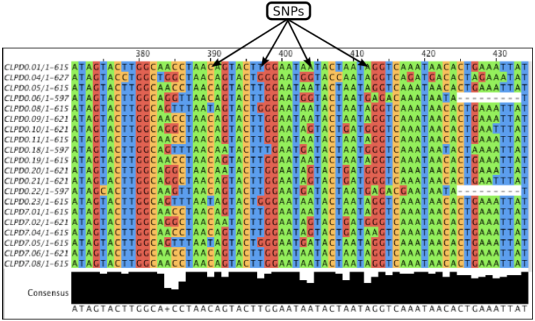


```
CLPD0.01/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.04/1-627 ATAGTACCTGGCTGGCTAACAGTACTGGGAATGGTACCAATAGGTCAGATGACACTAGAAATAT
CLPD0.05/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.06/1-597 ATAGTACTTGGCAGGTTAACAGTACTTGGAATGGTACTAATGAGACAAATAATA---------T
CLPD0.08/1-615 ATAGTACTTGGCAGTTTAATAGTACTGGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.09/1-621 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.10/1-621 ATAGTACTTGGCAGGCTAACAGTACTTGGAATAGTACTGATGGGTCAAATAACACTGAATTTAT
CLPD0.11/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.18/1-597 ATAGTACTTGGCAGTTTAACAATACTTTGAATGATACTAATGGGTCAAATAATACTAAAATTAT
CLPD0.19/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.20/1-621 ATAGTACTTGGCAGGCTAACAATACTTGGAATAGTACTGATGGGTCAAATAACACTGAAATTAT
CLPD0.21/1-621 ATAGTACTTGGCAGGCTAACAGTACTTGGAATAGTACTGATGGGTCAAATAACACTGAATTTAT
CLPD0.22/1-597 ATAGCACTTGGCAAGTTAACAGTACTTGGAATGATACTAATGAGACGAATAATA---------T
CLPD0.23/1-615 ATAGTACTTGGCAGTTTAATAGTACTGGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD7.01/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD7.02/1-621 ATAGTACTTGGCAGGCTAACAATACTTGGAATAGTACTGATGGGTCAAATAACACTGAAATTAT
CLPD7.04/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAGTACTGATAAGTCAAATAACACTGAAATTAT
CLPD7.05/1-615 ATAGTACTTGGCAGTTTAATAGTACTGGGAATGATACTAATAGGTCAAATAACACTGAAATTAT
CLPD7.06/1-621 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD7.08/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
     Consensus ATAGTACTTGGCA+CCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
```

- hundreds/thousands individuals
- up to millions of single nucleotide polymorphism (SNPs)

⇒ **Multivariate analysis use to summarize genetic diversity.**

DNA sequences: a rich source of information



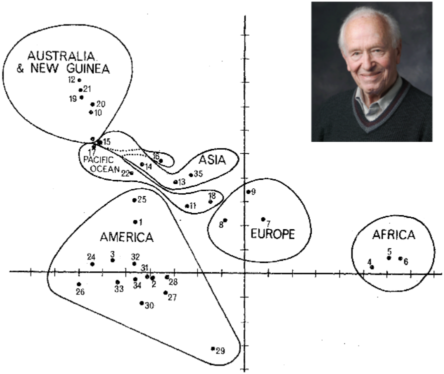
```
CLPD0.01/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.04/1-627 ATAGTACCTGGCTGGCTAACAGTACTGGGAATGGTACCAATAGGTCAGATGACACTAGAAATAT
CLPD0.05/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.06/1-597 ATAGTACTTGGCAGGTTAACAGTACTTGGAATGGTACTAATGAGACAAATAATA---------T
CLPD0.08/1-615 ATAGTACTTGGCAGTTTAATAGTACTGGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.09/1-621 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.10/1-621 ATAGTACTTGGCAGGCTAACAGTACTTGGAATAGTACTGATGGGTCAAATAACACTGAATTTAT
CLPD0.11/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.18/1-597 ATAGTACTTGGCAGTTTAACAATACTTTGAATGATACTAATGGGTCAAATAATACTAAAATTAT
CLPD0.19/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD0.20/1-621 ATAGTACTTGGCAGGCTAACAATACTTGGAATAGTACTGATGGGTCAAATAACACTGAAATTAT
CLPD0.21/1-621 ATAGTACTTGGCAGGCTAACAGTACTTGGAATAGTACTGATGGGTCAAATAACACTGAATTTAT
CLPD0.22/1-597 ATAGCACTTGGCAAGTTAACAGTACTTGGAATGATACTAATGAGACGAATAATA---------T
CLPD0.23/1-615 ATAGTACTTGGCAGTTTAATAGTACTGGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD7.01/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD7.02/1-621 ATAGTACTTGGCAGGCTAACAATACTTGGAATAGTACTGATGGGTCAAATAACACTGAAATTAT
CLPD7.04/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAGTACTGATAAGTCAAATAACACTGAAATTAT
CLPD7.05/1-615 ATAGTACTTGGCAGTTTAATAGTACTGGGAATGATACTAATAGGTCAAATAACACTGAAATTAT
CLPD7.06/1-621 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
CLPD7.08/1-615 ATAGTACTTGGCAACCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT

     Consensus ATAGTACTTGGCA+CCTAACAGTACTTGGAATAATACTAATAGGTCAAATAACACTGAAATTAT
```

- hundreds/thousands individuals
- up to millions of single nucleotide polymorphism (SNPs)

⇒ **Multivariate analysis use to summarize genetic diversity.**

First application of multivariate analysis in genetics

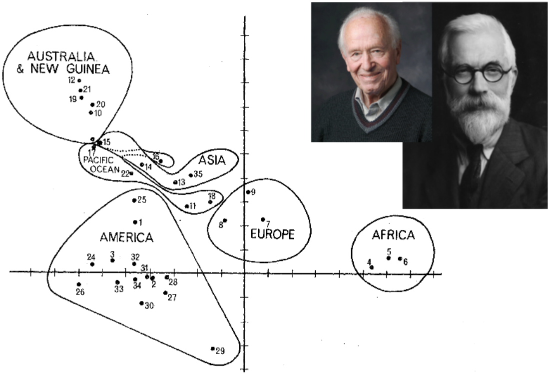
PCA of genetic data, native human populations (Cavalli-Sforza 1966, *Proc B*)



First 2 principal components separate populations into continents.

First application of multivariate analysis in genetics

PCA of genetic data, native human populations (Cavalli-Sforza 1966, *Proc B*)

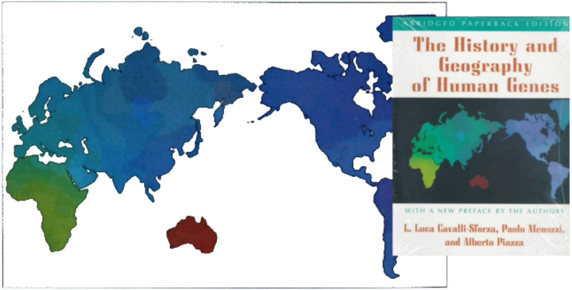


First 2 principal components separate populations into continents.

Applications: some examples

PCA of genetic data + colored maps of principal components

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



Signatures of Human expansion out-of-Africa.

Since then...

Multivariate methods used in genetics

- Principal Component Analysis (PCA)
- Principal Coordinates Analysis (PCoA) / Metric Multidimensional Scaling (MDS)
- Correspondance Analysis (CA)
- Discriminant Analysis (DA)
- Canonical Correlation Analysis (CCA)
- ...

Since then...

Applications

- reveal spatial structures (historical spread)
- explore genetic diversity
- identify cryptic species
- discover genotype-phenotype association
- ...
- review in Jombart et al. 2009, *Heredity* **102**: 330-341

Applications in genetics of pathogen populations.

Since then...

Applications

- reveal spatial structures (historical spread)
- explore genetic diversity
- identify cryptic species
- discover genotype-phenotype association
- ...
- review in Jombart et al. 2009, *Heredity* **102**: 330-341

Applications in genetics of pathogen populations.

Outline

Multivariate analysis in a nutshell

Applications to genomic data

Genetic diversity of pathogen populations

Why investigate the diversity of pathogen populations?

Genetic data: increasingly important in infectious disease epidemiology

Purposes

- classify pathogens, describe their relationships
- assess the spatio-temporal dynamics of infectious diseases
- reconstruct epidemiological processes (transmission)

Why investigate the diversity of pathogen populations?

Genetic data: increasingly important in infectious disease epidemiology

Purposes

- classify pathogens, describe their relationships
- assess the spatio-temporal dynamics of infectious diseases
- reconstruct epidemiological processes (transmission)

Why investigate the diversity of pathogen populations?

Genetic data: increasingly important in infectious disease epidemiology

Purposes

- classify pathogens, describe their relationships
- assess the spatio-temporal dynamics of infectious diseases
- reconstruct epidemiological processes (transmission)

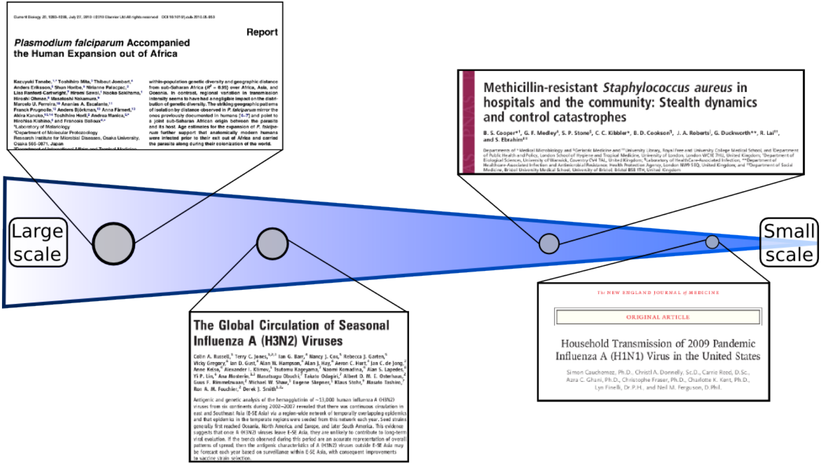
Why investigate the diversity of pathogen populations?

Genetic data: increasingly important in infectious disease epidemiology

Purposes

- classify pathogens, describe their relationships
- assess the spatio-temporal dynamics of infectious diseases
- reconstruct epidemiological processes (transmission)

Different questions at different scales



Current Biology 20, 1283–1288, July 27, 2010 ©2010 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2010.05.053

Report

***Plasmodium falciparum* Accompanied the Human Expansion out of Africa**

Kazuyuki Tanabe,[1,*] Toshihiro Mita,[3] Thibaut Jombart,[4] Anders Eriksson,[5] Shun Horibe,[6] Nirianne Palacpac,[2] Lisa Ranford-Cartwright,[7] Hiromi Sawai,[1] Naoko Sakihama,[1] Hiroshi Ohmae,[8] Masatoshi Nakamura,[9] Marcelo U. Ferreira,[10] Ananias A. Escalante,[11] Franck Prugnolle,[12] Anders Björkman,[13] Anna Färnert,[13] Akira Kaneko,[13,14] Toshihiro Horii,[2] Andrea Manica,[5,*] Hirohisa Kishino,[6] and Francois Balloux[4,*]

[1]Laboratory of Malariology
[2]Department of Molecular Protozoology
Research Institute for Microbial Diseases, Osaka University, Osaka 565-0871, Japan
[3]Department of International Affairs and Tropical Medicine

within-population genetic diversity and geographic distance from sub-Saharan Africa ($R^2 = 0.95$) over Africa, Asia, and Oceania. In contrast, regional variation in transmission intensity seems to have had a negligible impact on the distribution of genetic diversity. The striking geographic patterns of isolation by distance observed in *P. falciparum* mirror the ones previously documented in humans [4–7] and point to a joint sub-Saharan African origin between the parasite and its host. Age estimates for the expansion of *P. falciparum* further support that anatomically modern humans were infected prior to their exit out of Africa and carried the parasite along during their colonization of the world.

Methicillin-resistant *Staphylococcus aureus* in hospitals and the community: Stealth dynamics and control catastrophes

B. S. Cooper[*,†], G. F. Medley[‡], S. P. Stone[§], C. C. Kibbler[*], B. D. Cookson[¶], J. A. Roberts[‖], G. Duckworth[**], R. Lai[††], and S. Ebrahim[‡‡]

Departments of [*]Medical Microbiology and [§]Geriatric Medicine and [††]University Library, Royal Free and University College Medical School, and [‖]Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, University of London, London WC1E 7HU, United Kingdom; [‡]Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, United Kingdom; [¶]Laboratory of HealthCare-Associated Infection, [**]Department of Healthcare-Associated Infection and Antimicrobial Resistance, Health Protection Agency, London NW9 5EQ, United Kingdom; and [‡‡]Department of Social Medicine, Bristol University Medical School, University of Bristol, Bristol BS8 1TH, United Kingdom

The Global Circulation of Seasonal Influenza A (H3N2) Viruses

Colin A. Russell,[1] Terry C. Jones,[1,2,3] Ian G. Barr,[4] Nancy J. Cox,[5] Rebecca J. Garten,[5] Vicky Gregory,[6] Ian D. Gust,[4] Alan W. Hampson,[4] Alan J. Hay,[6] Aeron C. Hurt,[4] Jan C. de Jong,[2] Anne Kelso,[4] Alexander I. Klimov,[5] Tsutomu Kageyama,[7] Naomi Komadina,[4] Alan S. Lapedes,[8] Yi P. Lin,[6] Ana Mosterin,[1,3] Masatsugu Obuchi,[7] Takato Odagiri,[7] Albert D. M. E. Osterhaus,[2] Guus F. Rimmelzwaan,[2] Michael W. Shaw,[5] Eugene Skepner,[1] Klaus Stohr,[9] Masato Tashiro,[7] Ron A. M. Fouchier,[2] Derek J. Smith[1,2,*]

Antigenic and genetic analysis of the hemagglutinin of ~13,000 human influenza A (H3N2) viruses from six continents during 2002–2007 revealed that there was continuous circulation in east and Southeast Asia (E-SE Asia) via a region-wide network of temporally overlapping epidemics and that epidemics in the temperate regions were seeded from this network each year. Seed strains generally first reached Oceania, North America, and Europe, and later South America. This evidence suggests that once A (H3N2) viruses leave E-SE Asia, they are unlikely to contribute to long-term viral evolution. If the trends observed during this period are an accurate representation of overall patterns of spread, then the antigenic characteristics of A (H3N2) viruses outside E-SE Asia may be forecast each year based on surveillance within E-SE Asia, with consequent improvements to vaccine strain selection.

The NEW ENGLAND JOURNAL of MEDICINE

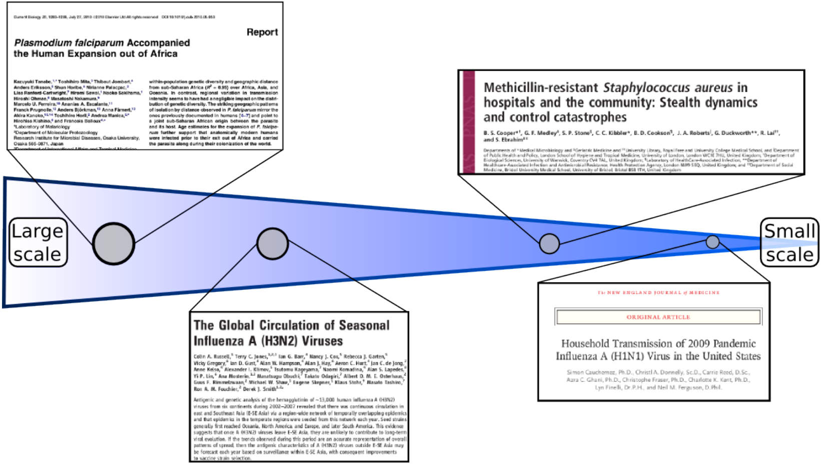
ORIGINAL ARTICLE

Household Transmission of 2009 Pandemic Influenza A (H1N1) Virus in the United States

Simon Cauchemez, Ph.D., Christl A. Donnelly, Sc.D., Carrie Reed, D.Sc., Azra C. Ghani, Ph.D., Christophe Fraser, Ph.D., Charlotte K. Kent, Ph.D., Lyn Finelli, Dr.P.H., and Neil M. Ferguson, D.Phil.

Where and how can multivariate analysis of pathogen genetic data be useful?

Different questions at different scales



Current Biology 20, 1283–1288, July 27, 2010 ©2010 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2010.05.053

Report

Plasmodium falciparum Accompanied the Human Expansion out of Africa

Kazuyuki Tanabe,[1,*] Toshihiro Mita,[2] Thibaut Jombart,[4] Anders Eriksson,[5] Shun Horibe,[6] Nirianne Palacpac,[2] Lisa Ranford-Cartwright,[7] Hiromi Sawai,[1] Naoko Sakihama,[1] Hiroshi Ohmae,[8] Masatoshi Nakamura,[9] Marcelo U. Ferreira,[10] Ananias A. Escalante,[11] Franck Prugnolle,[12] Anders Björkman,[13] Anna Färnert,[13] Akira Kaneko,[13,14] Toshihiro Horii,[2] Andrea Manica,[5,*] Hirohisa Kishino,[6] and Francois Balloux[4,*]
[1]Laboratory of Malariology
[2]Department of Molecular Protozoology
Research Institute for Microbial Diseases, Osaka University, Osaka 565-0871, Japan
[3]Department of International Affairs and Tropical Medicine

within-population genetic diversity and geographic distance from sub-Saharan Africa ($R^2 = 0.95$) over Africa, Asia, and Oceania. In contrast, regional variation in transmission intensity seems to have had a negligible impact on the distribution of genetic diversity. The striking geographic patterns of isolation by distance observed in *P. falciparum* mirror the ones previously documented in humans [4–7] and point to a joint sub-Saharan African origin between the parasite and its host. Age estimates for the expansion of *P. falciparum* further support that anatomically modern humans were infected prior to their exit out of Africa and carried the parasite along during their colonization of the world.

Methicillin-resistant *Staphylococcus aureus* in hospitals and the community: Stealth dynamics and control catastrophes

B. S. Cooper[*†], G. F. Medley[‡], S. P. Stone[§], C. C. Kibbler[*], B. D. Cookson[¶], J. A. Roberts[‖], G. Duckworth[**], R. Lai[††], and S. Ebrahim[‡‡]

Departments of *Medical Microbiology and §Geriatric Medicine and ††University Library, Royal Free and University College Medical School, and ‖Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, University of London, London WC1E 7HU, United Kingdom; ‡Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, United Kingdom; ¶Laboratory of HealthCare-Associated Infection, **Department of Healthcare-Associated Infection and Antimicrobial Resistance, Health Protection Agency, London NW9 5EQ, United Kingdom; and ‡‡Department of Social Medicine, Bristol University Medical School, University of Bristol, Bristol BS8 1TH, United Kingdom

The Global Circulation of Seasonal Influenza A (H3N2) Viruses

Colin A. Russell,[1] Terry C. Jones,[1,2,3] Ian G. Barr,[4] Nancy J. Cox,[5] Rebecca J. Garten,[5] Vicky Gregory,[6] Ian D. Gust,[4] Alan W. Hampson,[4] Alan J. Hay,[6] Aeron C. Hurt,[4] Jan C. de Jong,[2] Anne Kelso,[4] Alexander I. Klimov,[5] Tsutomu Kageyama,[7] Naomi Komadina,[4] Alan S. Lapedes,[8] Yi P. Lin,[6] Ana Mosterin,[1,3] Masatsugu Obuchi,[7] Takato Odagiri,[7] Albert D. M. E. Osterhaus,[2] Guus F. Rimmelzwaan,[2] Michael W. Shaw,[5] Eugene Skepner,[1] Klaus Stohr,[9] Masato Tashiro,[7] Ron A. M. Fouchier,[2] Derek J. Smith[1,2,*]

Antigenic and genetic analysis of the hemagglutinin of ~13,000 human influenza A (H3N2) viruses from six continents during 2002–2007 revealed that there was continuous circulation in east and Southeast Asia (E-SE Asia) via a region-wide network of temporally overlapping epidemics and that epidemics in the temperate regions were seeded from this network each year. Seed strains generally first reached Oceania, North America, and Europe, and later South America. This evidence suggests that once A (H3N2) viruses leave E-SE Asia, they are unlikely to contribute to long-term viral evolution. If the trends observed during this period are an accurate representation of overall patterns of spread, then the antigenic characteristics of A (H3N2) viruses outside E-SE Asia may be forecast each year based on surveillance within E-SE Asia, with consequent improvements to vaccine strain selection.

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Household Transmission of 2009 Pandemic Influenza A (H1N1) Virus in the United States

Simon Cauchemez, Ph.D., Christl A. Donnelly, Sc.D., Carrie Reed, D.Sc., Azra C. Ghani, Ph.D., Christophe Fraser, Ph.D., Charlotte K. Kent, Ph.D., Lyn Finelli, Dr.P.H., and Neil M. Ferguson, D.Phil.

Where and how can multivariate analysis of pathogen genetic data be useful?

Describing pathogen populations

Population genetics: identify populations of organisms and describe their relationships

What is a population?

- *Usual definition*: set of organisms mating at random
- *Problem*: no "mating" in most pathogens (e.g. viruses, bacteria)
- **Genetic clusters**: set of genetically related pathogens (e.g. same outbreak, same epidemic).

⇒ aim: identify and describe genetic clusters

Describing pathogen populations

Population genetics: identify populations of organisms and describe their relationships

What is a population?

- *Usual definition*: set of organisms mating at random
- *Problem*: no "mating" in most pathogens (e.g. viruses, bacteria)
- **Genetic clusters**: set of genetically related pathogens (e.g. same outbreak, same epidemic).

⇒ aim: identify and describe genetic clusters

Describing pathogen populations

Population genetics: identify populations of organisms and describe their relationships

What is a population?

- *Usual definition*: set of organisms mating at random
- *Problem*: no “mating” in most pathogens (e.g. viruses, bacteria)
- **Genetic clusters**: set of genetically related pathogens (e.g. same outbreak, same epidemic).

⇒ aim: identify and describe genetic clusters

Describing pathogen populations

Population genetics: identify populations of organisms and describe their relationships

What is a population?

- *Usual definition*: set of organisms mating at random
- *Problem*: no "mating" in most pathogens (e.g. viruses, bacteria)
- **Genetic clusters**: set of genetically related pathogens (e.g. same outbreak, same epidemic).

⇒ aim: identify and describe genetic clusters

Describing pathogen populations

Population genetics: identify populations of organisms and describe their relationships

What is a population?

- *Usual definition*: set of organisms mating at random
- *Problem*: no "mating" in most pathogens (e.g. viruses, bacteria)
- **Genetic clusters**: set of genetically related pathogens (e.g. same outbreak, same epidemic).

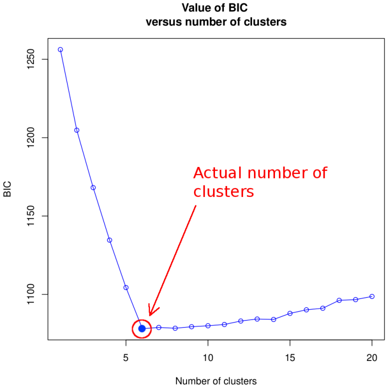
⇒ aim: identify and describe genetic clusters

Genetic clustering using K-means & BIC

(Jombart *et al.* 2010, *BMC Genetics*)

Variance partitioning model (ANOVA):

$$tot.\ variance = (bet.\ groups) + (wit.\ groups)$$



Performances:

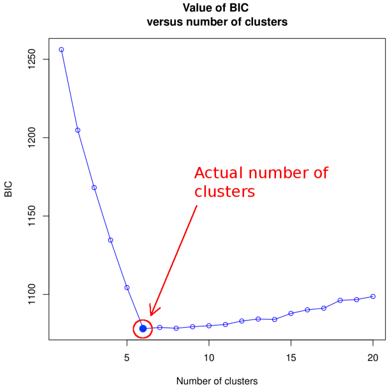
- K-means $\geq$ STRUCTURE on simulated data (various island and stepping stone models)
- orders of magnitude faster (seconds vs hours/days)

Genetic clustering using K-means & BIC

(Jombart *et al.* 2010, *BMC Genetics*)

Variance partitioning model (ANOVA):

$$tot.\ variance = (bet.\ groups) + (wit.\ groups)$$

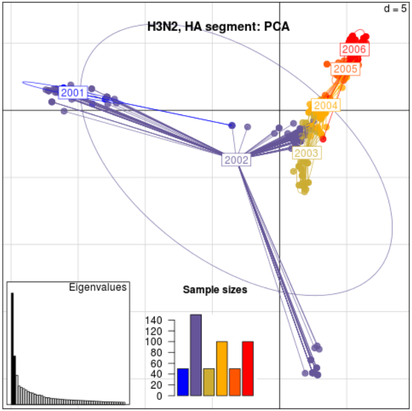


Performances:

- K-means $\geq$ STRUCTURE on simulated data (various island and stepping stone models)
- orders of magnitude faster (seconds vs hours/days)

PCA of seasonal influenza (A/H3N2) data

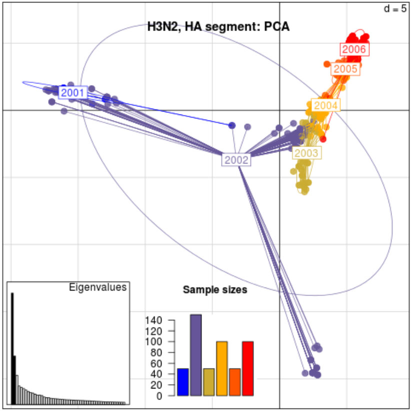
Data: seasonal influenza (A/H3N2), 500 HA segments.



Little temporal evolution, burst of diversity in 2002??

PCA of seasonal influenza (A/H3N2) data

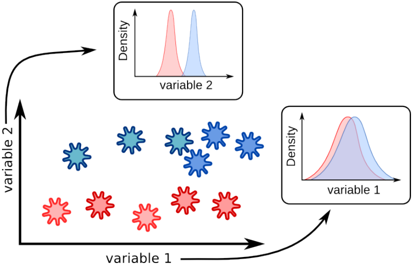
Data: seasonal influenza (A/H3N2), 500 HA segments.



Little temporal evolution, burst of diversity in 2002??

Which diversity to represent?

Total diversity not relevant to analyse clusters.



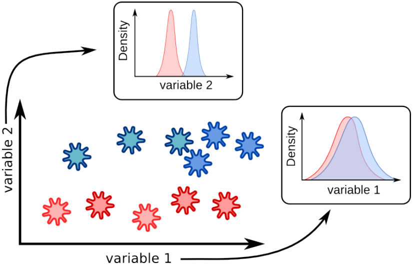
Discriminant Analysis of Principal Components (DAPC):

(Jombart *et al.* 2010, *BMC Genetics*)

- maximizes group discrimination ("$between/within$" ratio)
- provides group membership probabilities (prediction possible)
- as computer-efficient as PCA

Which diversity to represent?

Total diversity not relevant to analyse clusters.

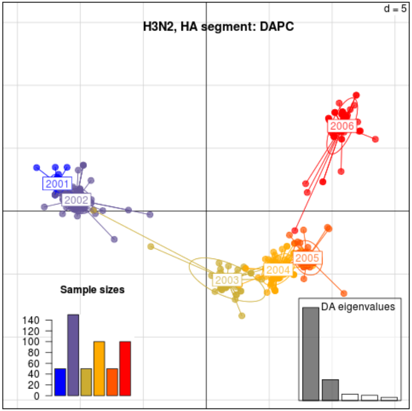


Discriminant Analysis of Principal Components (DAPC):

(Jombart *et al.* 2010, *BMC Genetics*)

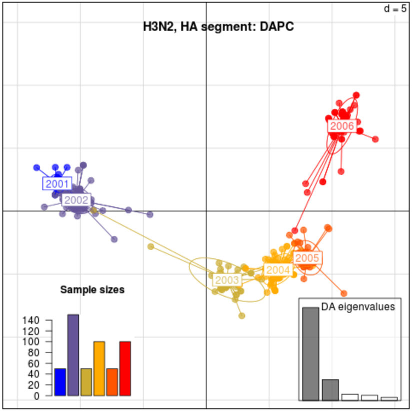
- maximizes group discrimination ("$between/within$" ratio)
- provides group membership probabilities (prediction possible)
- as computer-efficient as PCA

DAPC of seasonal influenza (A/H3N2) data



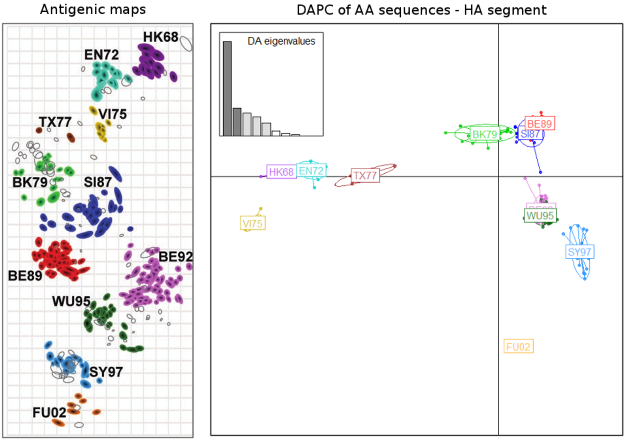
Strong temporal signal, originality of 2006 isolates (new alleles).

DAPC of seasonal influenza (A/H3N2) data



Strong temporal signal, originality of 2006 isolates (new alleles).

Identifying antigenic clusters in influenza (A/H3N2)


Antigenic maps

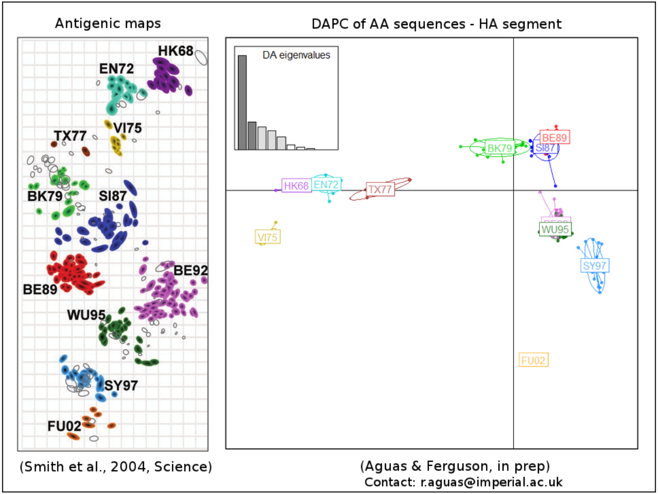
DAPC of AA sequences - HA segment

(Smith et al., 2004, Science)

(Aguas & Ferguson, in prep)
Contact: r.aguas@imperial.ac.uk

Antigenic clusters identified directly from AA sequences.

Identifying antigenic clusters in influenza (A/H3N2)



(Smith et al., 2004, Science)

(Aguas & Ferguson, in prep)
Contact: r.aguas@imperial.ac.uk

Antigenic clusters identified directly from AA sequences.

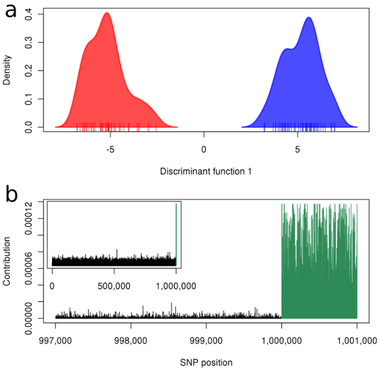
DAPC to identify structuring alleles

DAPC finds combinations of alleles most differing between groups.

Simulated data:

(Jombart & Ahmed 2011, *Bioinformatics*)

- 2 clusters, 50 isolates each
- 1,000,000 non structured SNPs
- 1,000 structured SNPs
  (i.e. different frequencies between groups)



Possible applications to pathogen GWAS (e.g. SNPs related to antibiotic resistance in bacteria).

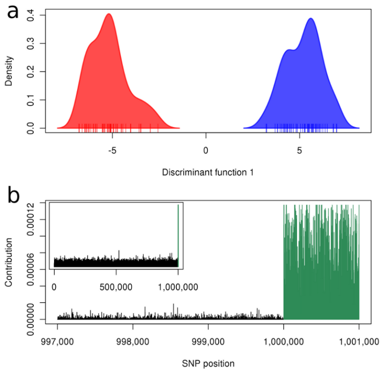
DAPC to identify structuring alleles

DAPC finds combinations of alleles most differing between groups.

Simulated data:

(Jombart & Ahmed 2011, *Bioinformatics*)

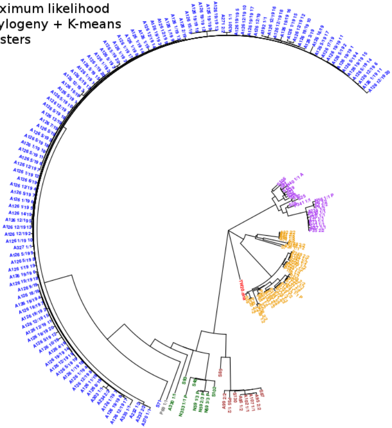
- 2 clusters, 50 isolates each
- 1,000,000 non structured SNPs
- 1,000 structured SNPs (i.e. different frequencies between groups)



Possible applications to pathogen GWAS (e.g. SNPs related to antibiotic resistance in bacteria).

Limits of multivariate analysis

Methicillin-resistant *Staphylococcus aureus* (MRSA) outbreak within hospital, Thailand. ~ 200 full-genome sequences. $\sim 1,000$ SNPs.



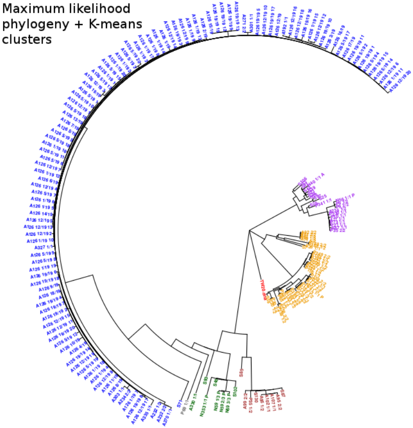
Maximum likelihood phylogeny + K-means clusters

Observations:

- greater diversity than expected
- genetic clusters can be defined
- transmissions at within-cluster level
- multivariate analysis = loss of information

Limits of multivariate analysis

Methicillin-resistant *Staphylococcus aureus* (MRSA) outbreak within hospital, Thailand. ~ 200 full-genome sequences. $\sim 1,000$ SNPs.

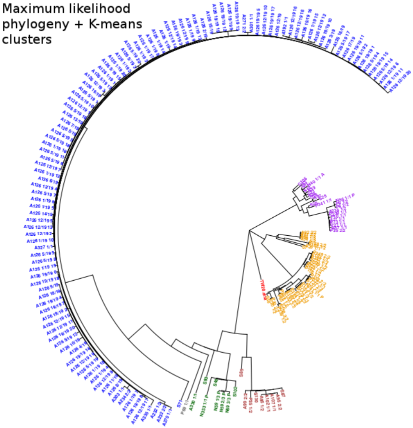


Observations:

- greater diversity than expected
- genetic clusters can be defined
- transmissions at within-cluster level
- multivariate analysis = loss of information

Limits of multivariate analysis

Methicillin-resistant *Staphylococcus aureus* (MRSA) outbreak within hospital, Thailand. ~ 200 full-genome sequences. $\sim 1,000$ SNPs.

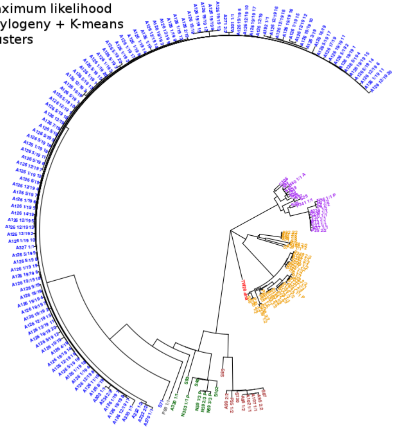


Observations:

- greater diversity than expected
- genetic clusters can be defined
- transmissions at within-cluster level
- multivariate analysis = loss of information

Limits of multivariate analysis

Methicillin-resistant *Staphylococcus aureus* (MRSA) outbreak within hospital, Thailand. ~ 200 full-genome sequences. $\sim 1,000$ SNPs.



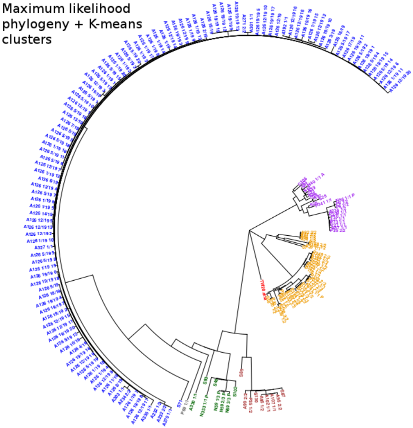
Maximum likelihood phylogeny + K-means clusters

Observations:

- greater diversity than expected
- genetic clusters can be defined
- transmissions at within-cluster level
- multivariate analysis = loss of information

Limits of multivariate analysis

Methicillin-resistant *Staphylococcus aureus* (MRSA) outbreak within hospital, Thailand. ~ 200 full-genome sequences. $\sim 1,000$ SNPs.



Observations:

- greater diversity than expected
- genetic clusters can be defined
- transmissions at within-cluster level
- multivariate analysis = loss of information

Multivariate analysis usually not informative on small-scale processes.

Summary

- multivariate analysis used for ~ 50 years in genetics, still an active field for methodological development
- increasingly useful as datasets grow
- specific application to pathogen genetic data
- limits reached when reconstructing fine-scale processes

Summary

- multivariate analysis used for ~ 50 years in genetics, still an active field for methodological development
- increasingly useful as datasets grow
- specific application to pathogen genetic data
- limits reached when reconstructing fine-scale processes

Summary

- multivariate analysis used for ~ 50 years in genetics, still an active field for methodological development
- increasingly useful as datasets grow
- specific application to pathogen genetic data
- limits reached when reconstructing fine-scale processes

Summary

- multivariate analysis used for ~ 50 years in genetics, still an active field for methodological development
- increasingly useful as datasets grow
- specific application to pathogen genetic data
- limits reached when reconstructing fine-scale processes