Imperial College London

MRC Centre for Outbreak Analysis and Modelling

Multivariate analysis of genetic data: an introduction

Thibaut Jombart

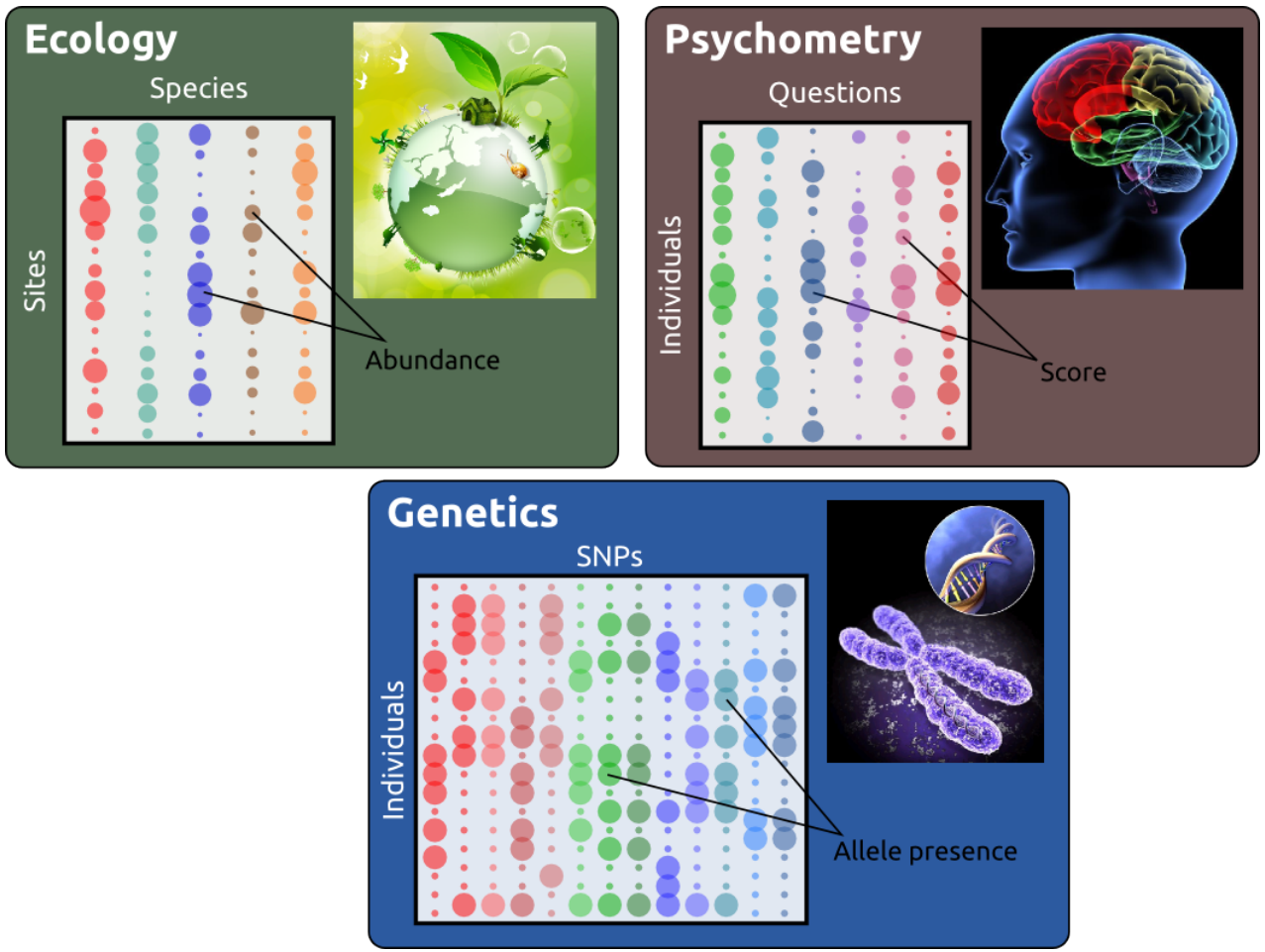
MRC Centre for Outbreak Analysis and Modelling
Imperial College London

XXIV Simposio Internacional De Estadística
Bogotá, 25th July 2014

Outline

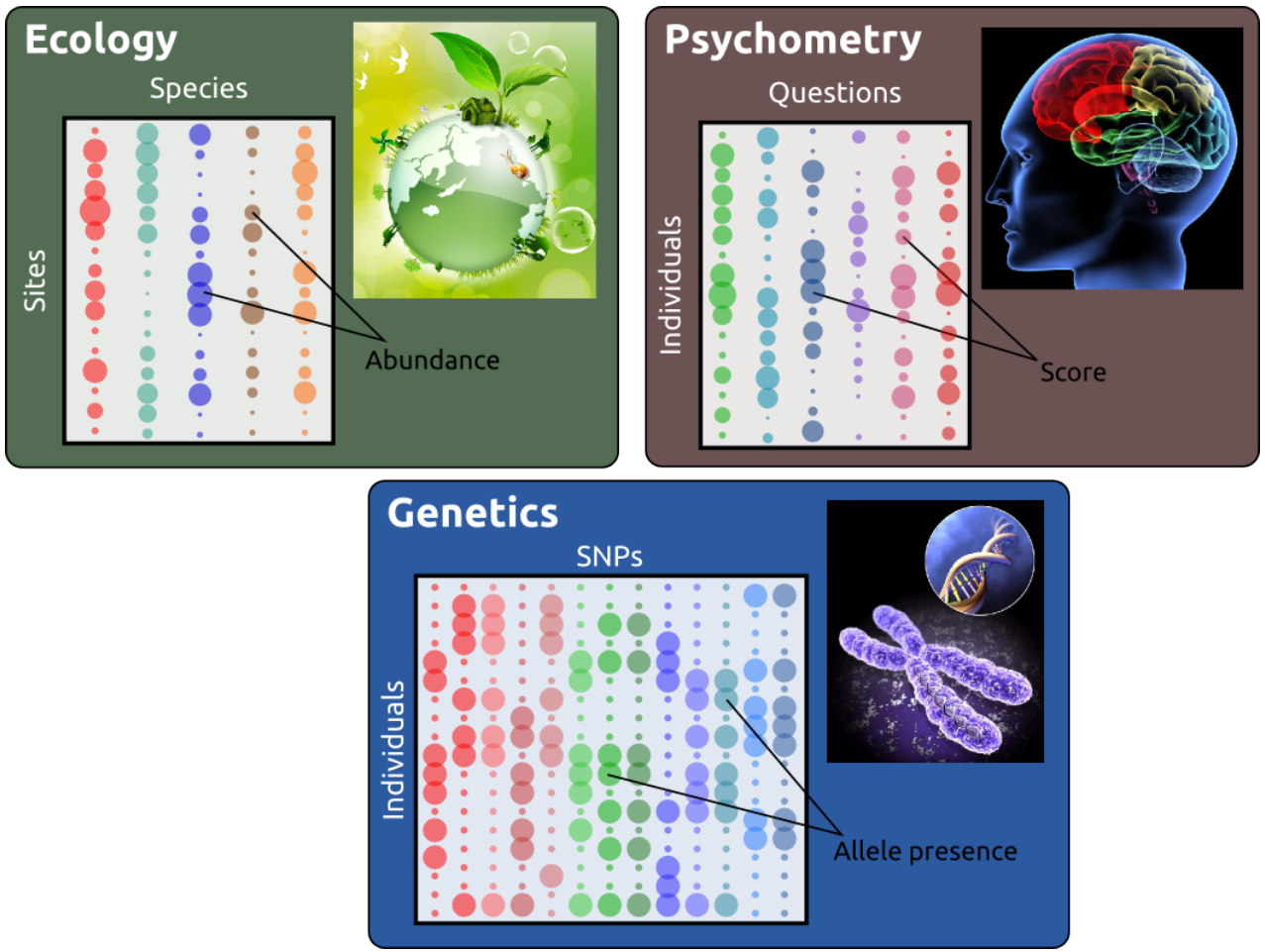
Outline

Multivariate data: some examples



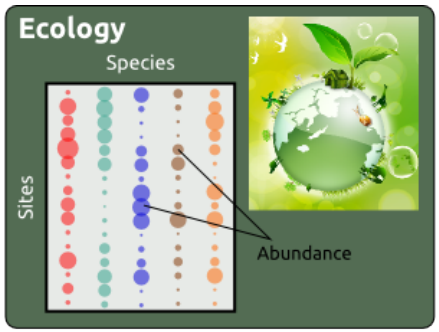
Association between individuals? Correlations between variables?

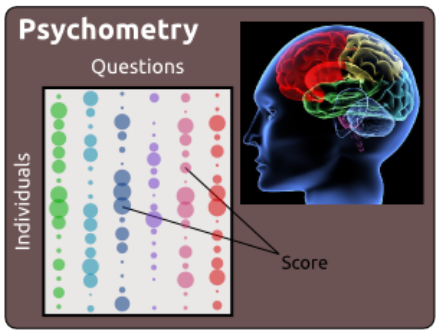
Multivariate data: some examples

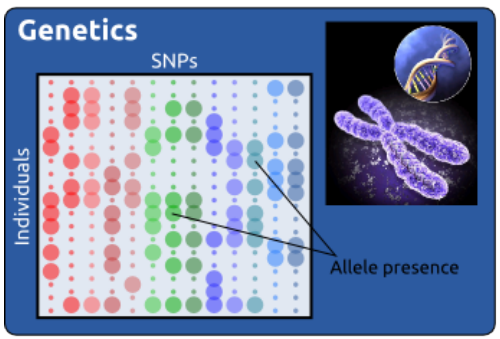


Association between individuals? Correlations between variables?

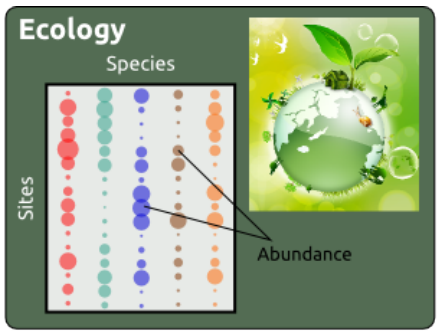
Multivariate analysis to summarize diversity

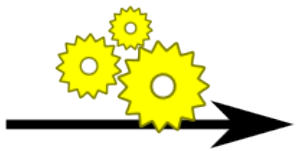


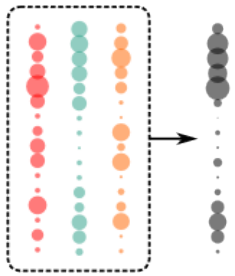




Multivariate analysis to summarize diversity

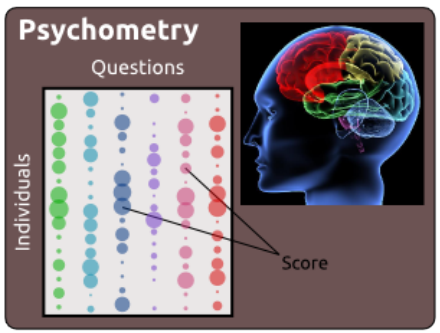


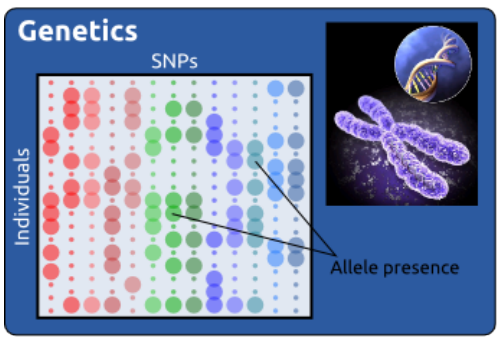




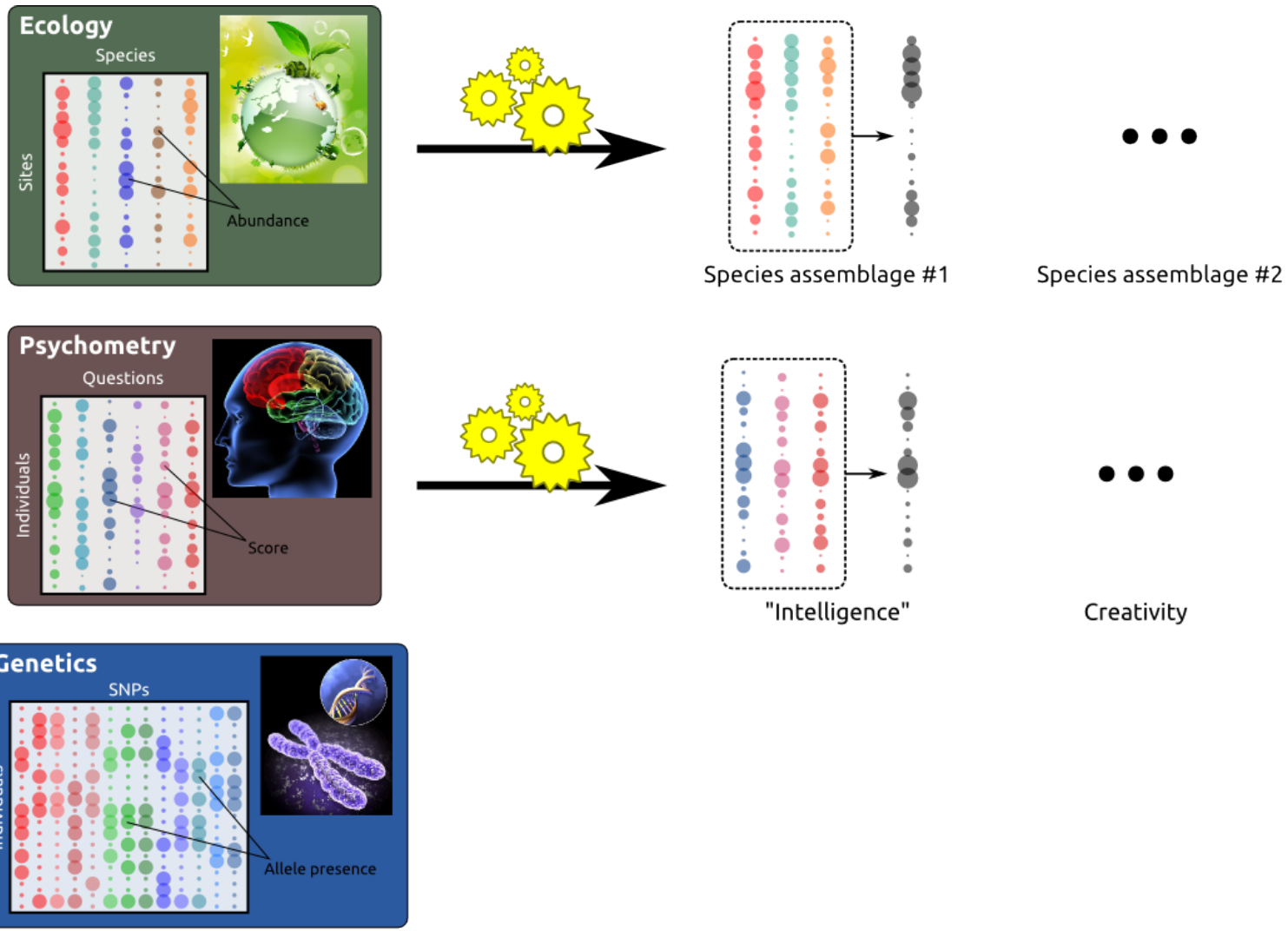
Species assemblage #1

Species assemblage #2

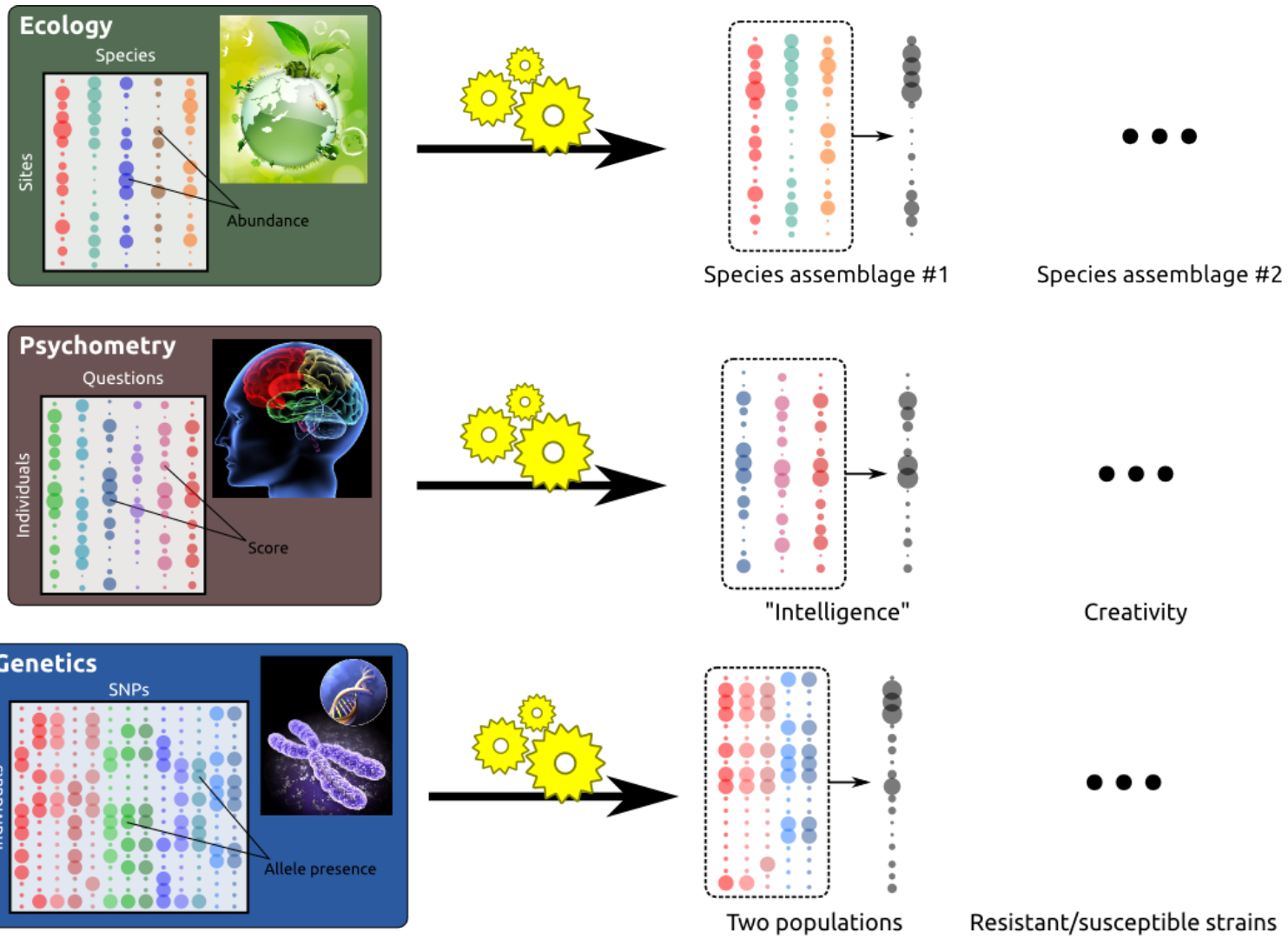




Multivariate analysis to summarize diversity



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.

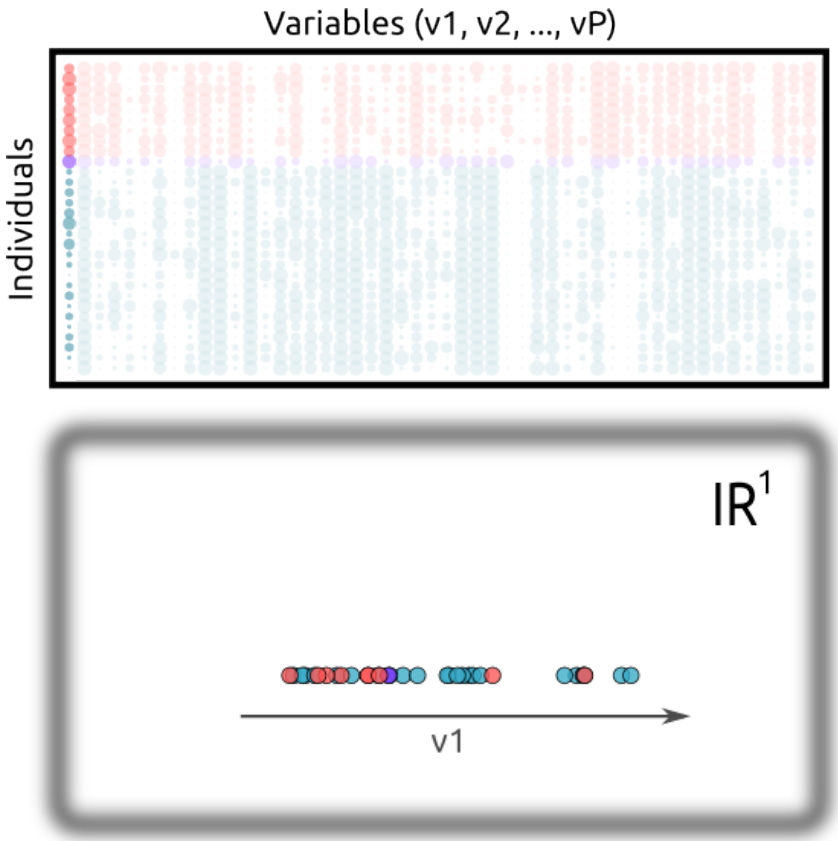
Most common methods

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)

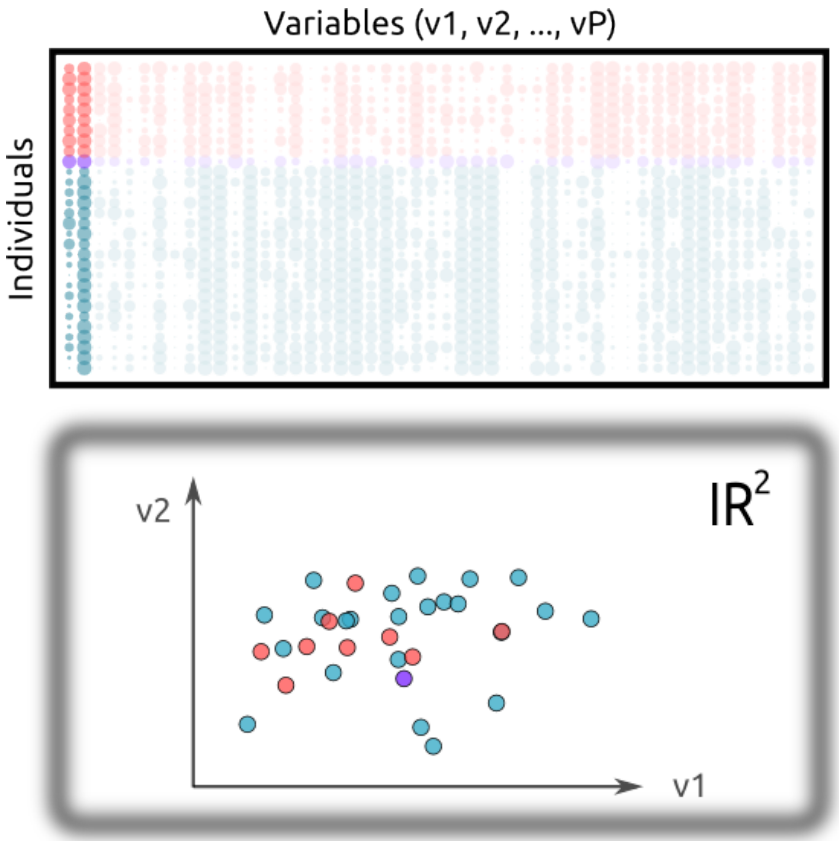
Many other methods for ≥ 2 data tables, spatial analysis, phylogenetic analysis, etc.

1 dimension, 2 dimensions, P dimensions



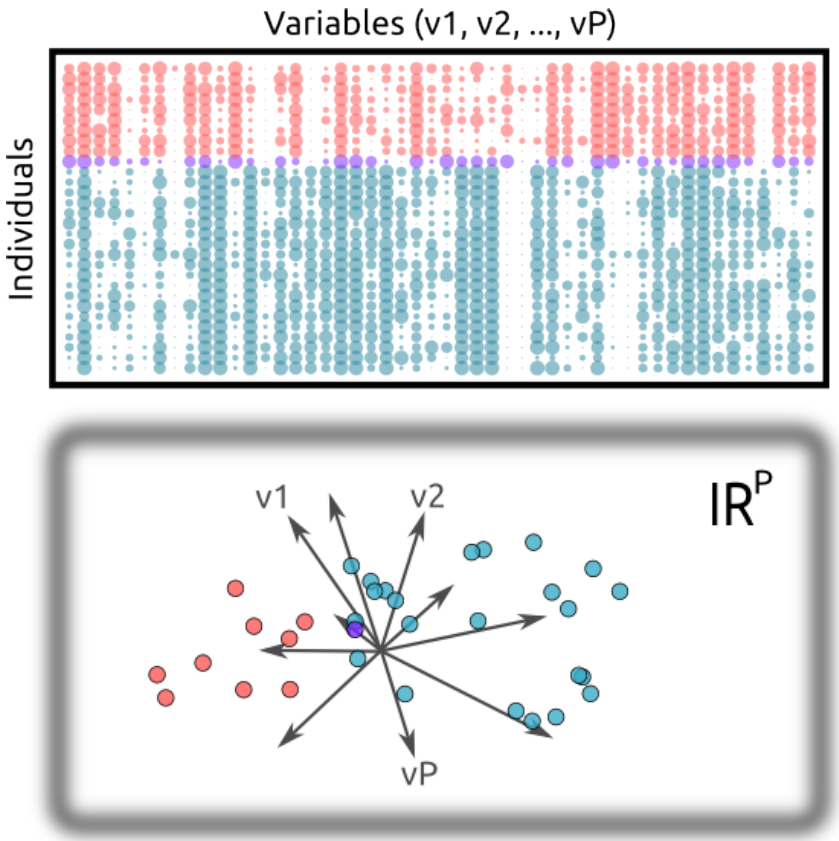
Need to find most informative directions in a P-dimensional space.

1 dimension, 2 dimensions, P dimensions



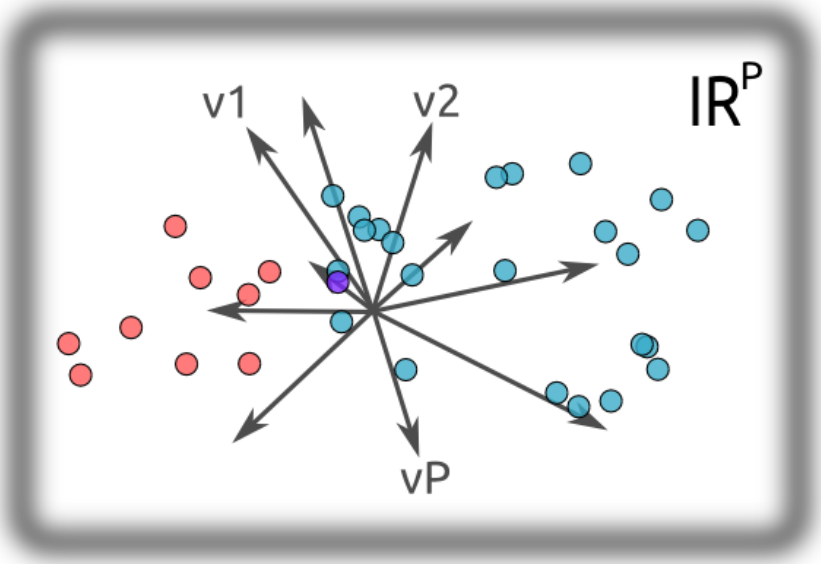
Need to find most informative directions in a P-dimensional space.

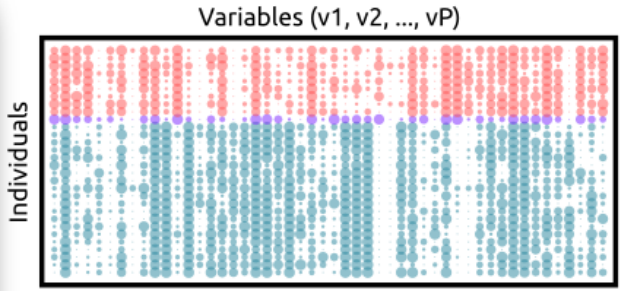
1 dimension, 2 dimensions, P dimensions



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

Reducing P dimensions into 1

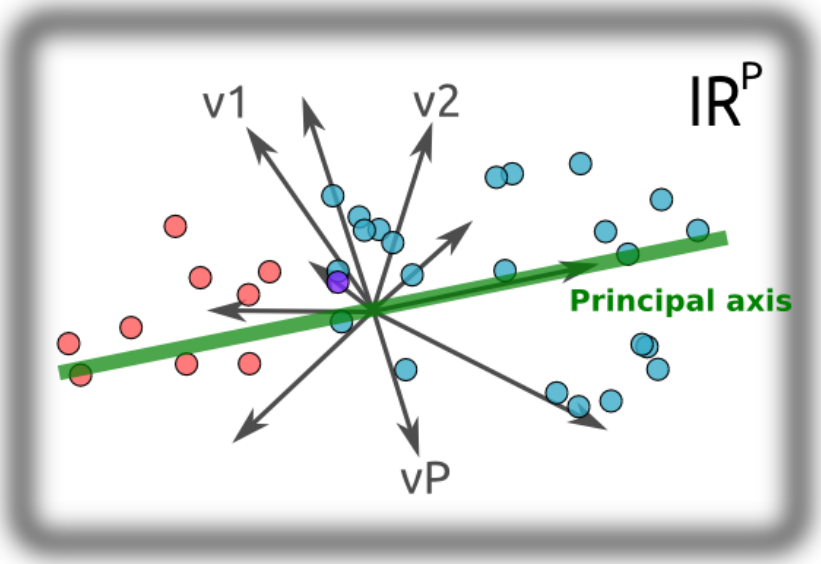


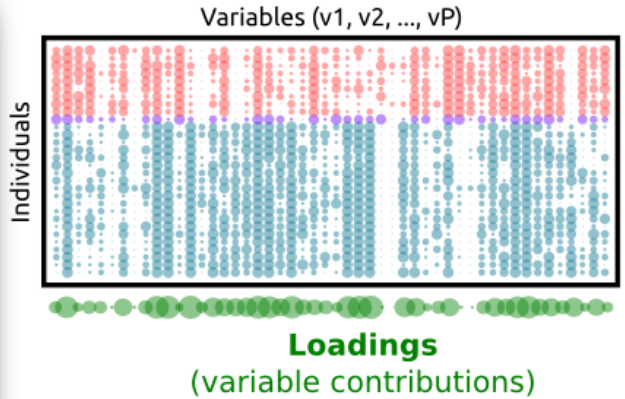


- $\mathbf{X} \in \mathbb{R}^{N \times P}$; $\mathbf{X} = [\mathbf{x}_1 | \ldots | \mathbf{x}_P]$: data matrix
- $\mathbf{Q} \in \mathbb{R}^{P \times P}$ metric in $\mathbb{R}^P$; $\mathbf{D} \in \mathbb{R}^{N \times N}$ metric in $\mathbb{R}^N$
- $\mathbf{u} \in \mathbb{R}^P$; $\mathbf{u} = [u_1, \ldots, u_P]$: **principal axis** $(\|\mathbf{u}\|^2_{\mathbf{Q}} = 1)$
- $\mathbf{v} \in \mathbb{R}^N$; $\mathbf{v} = \mathbf{XQu}$: **principal component**

$\rightarrow$ find $\mathbf{u}$ so that $\|\mathbf{v}\|^2_{\mathbf{D}}$ is maximum.

Reducing P dimensions into 1

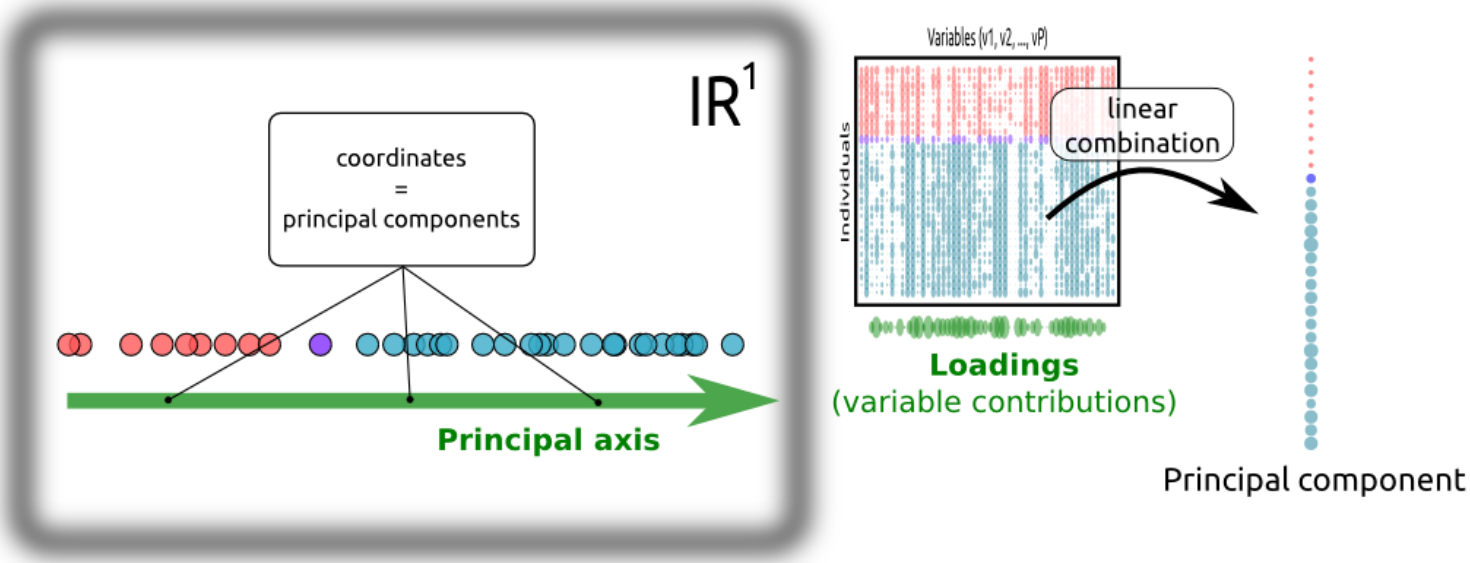




- $\mathbf{X} \in \mathbb{R}^{N \times P}$; $\mathbf{X} = [\mathbf{x}_1 | \ldots | \mathbf{x}_P]$: data matrix
- $\mathbf{Q} \in \mathbb{R}^{P \times P}$ metric in $\mathbb{R}^P$; $\mathbf{D} \in \mathbb{R}^{N \times N}$ metric in $\mathbb{R}^N$
- $\mathbf{u} \in \mathbb{R}^P$; $\mathbf{u} = [u_1, \ldots, u_P]$: **principal axis** $(\|\mathbf{u}\|^2_{\mathbf{Q}} = 1)$
- $\mathbf{v} \in \mathbb{R}^N$; $\mathbf{v} = \mathbf{XQu}$: **principal component**

$\rightarrow$ find $\mathbf{u}$ so that $\|\mathbf{v}\|^2_{\mathbf{D}}$ is maximum.

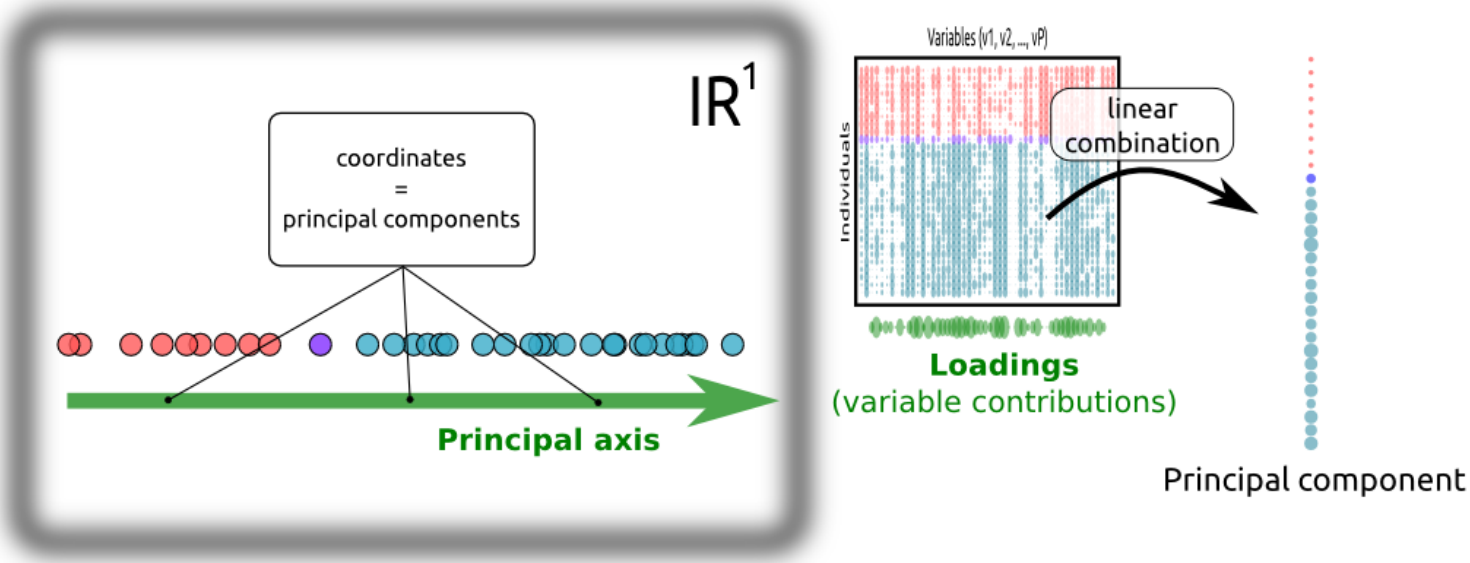
Reducing P dimensions into 1



- $\mathbf{X} \in \mathbb{R}^{N \times P}$; $\mathbf{X} = [\mathbf{x}_1 | \ldots | \mathbf{x}_P]$: data matrix
- $\mathbf{Q} \in \mathbb{R}^{P \times P}$ metric in $\mathbb{R}^P$; $\mathbf{D} \in \mathbb{R}^{N \times N}$ metric in $\mathbb{R}^N$
- $\mathbf{u} \in \mathbb{R}^P$; $\mathbf{u} = [u_1, \ldots, u_P]$: **principal axis** $(\|\mathbf{u}\|^2_{\mathbf{Q}} = 1)$
- $\mathbf{v} \in \mathbb{R}^N$; $\mathbf{v} = \mathbf{XQu}$: **principal component**

$\rightarrow$ find $\mathbf{u}$ so that $\|\mathbf{v}\|^2_{\mathbf{D}}$ is maximum.

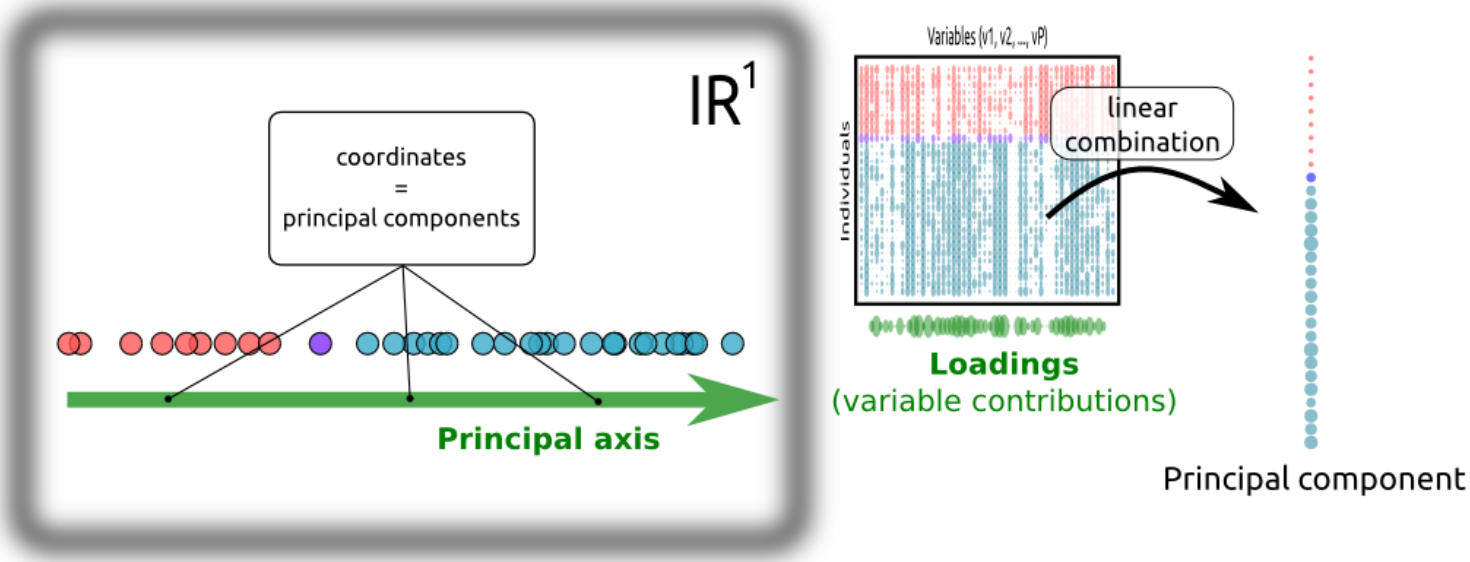
Reducing P dimensions into 1



- $\mathbf{X} \in \mathbb{R}^{N \times P}$; $\mathbf{X} = [\mathbf{x}_1 | \ldots | \mathbf{x}_P]$: data matrix
- $\mathbf{Q} \in \mathbb{R}^{P \times P}$ metric in $\mathbb{R}^P$; $\mathbf{D} \in \mathbb{R}^{N \times N}$ metric in $\mathbb{R}^N$
- $\mathbf{u} \in \mathbb{R}^P$; $\mathbf{u} = [u_1, \ldots, u_P]$: **principal axis** $(\|\mathbf{u}\|^2_{\mathbf{Q}} = 1)$
- $\mathbf{v} \in \mathbb{R}^N$; $\mathbf{v} = \mathbf{XQu}$: **principal component**

$\rightarrow$ find $\mathbf{u}$ so that $\|\mathbf{v}\|^2_{\mathbf{D}}$ is maximum.

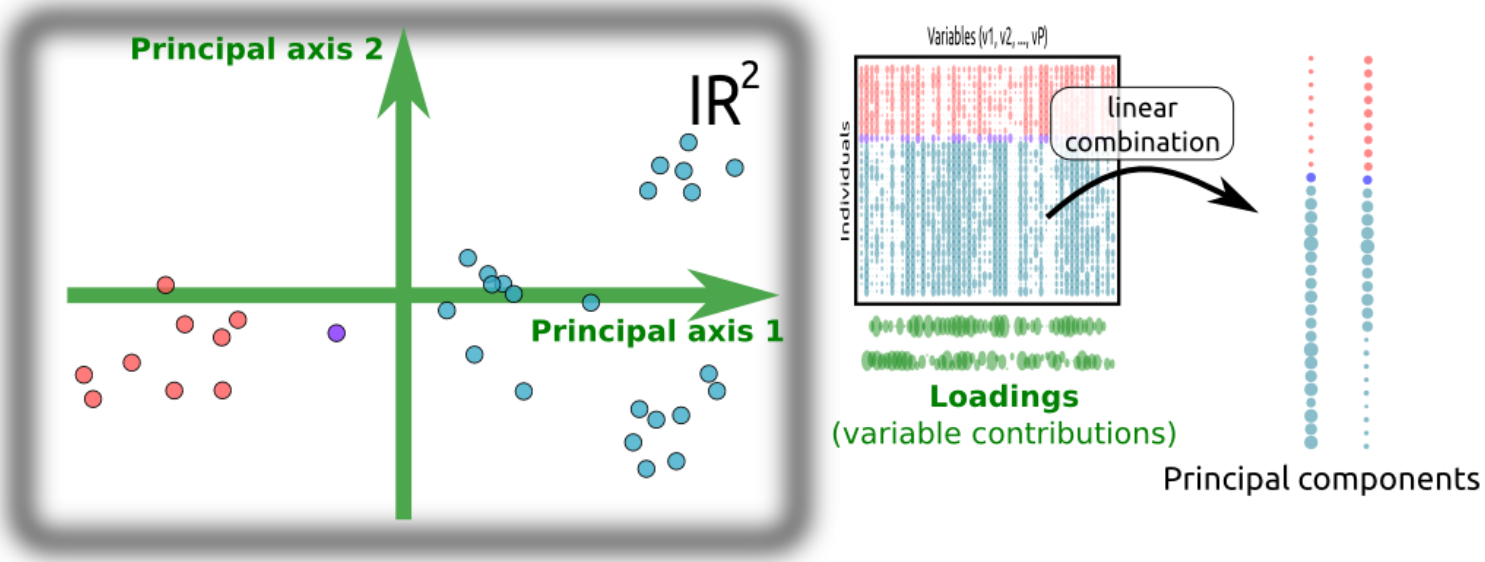
Keeping more than one principal component



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

$\rightarrow$ constraint: $\mathbf{u}_1 \perp \mathbf{u}_2$ (i.e., $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle_{\mathbf{Q}} = 0$)
$\rightarrow$ find $\mathbf{u}_2$ so that $\|\mathbf{v}_2\|^2_{\mathbf{D}}$ is maximum

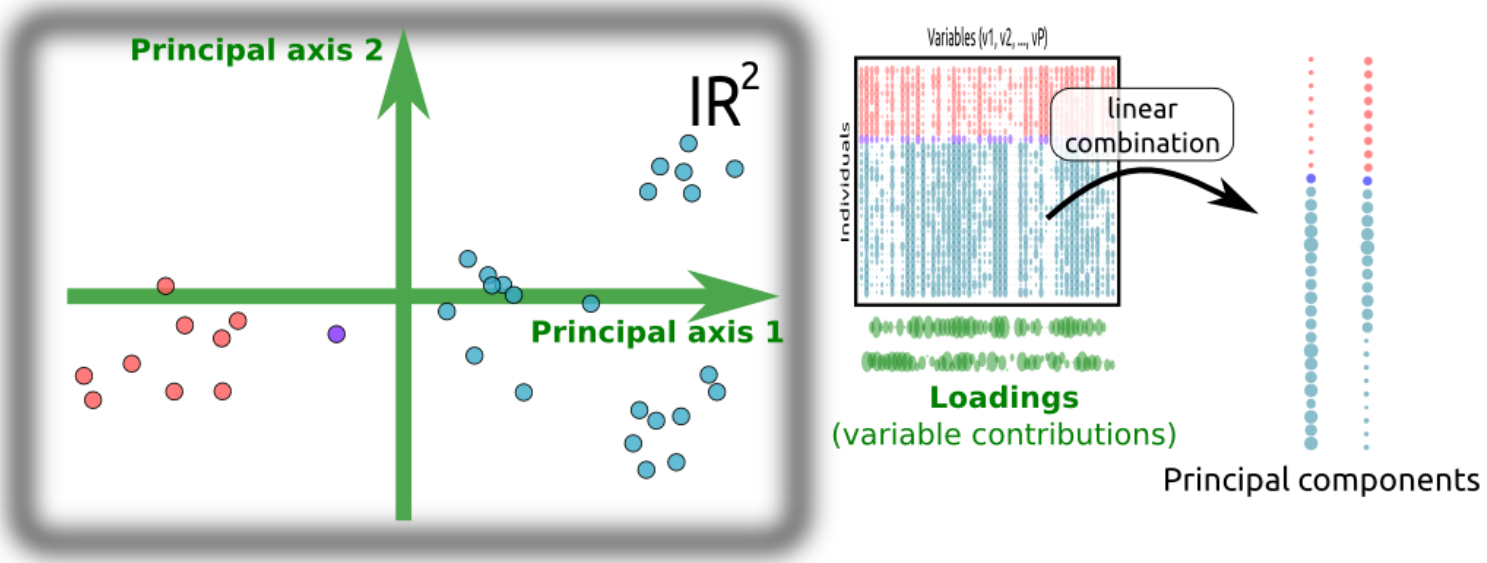
Keeping more than one principal component



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

$\rightarrow$ constraint: $\mathbf{u}_1 \perp \mathbf{u}_2$ (i.e., $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle_{\mathbf{Q}} = 0$)
$\rightarrow$ find $\mathbf{u}_2$ so that $\|\mathbf{v}_2\|^2_{\mathbf{D}}$ is maximum

Keeping more than one principal component

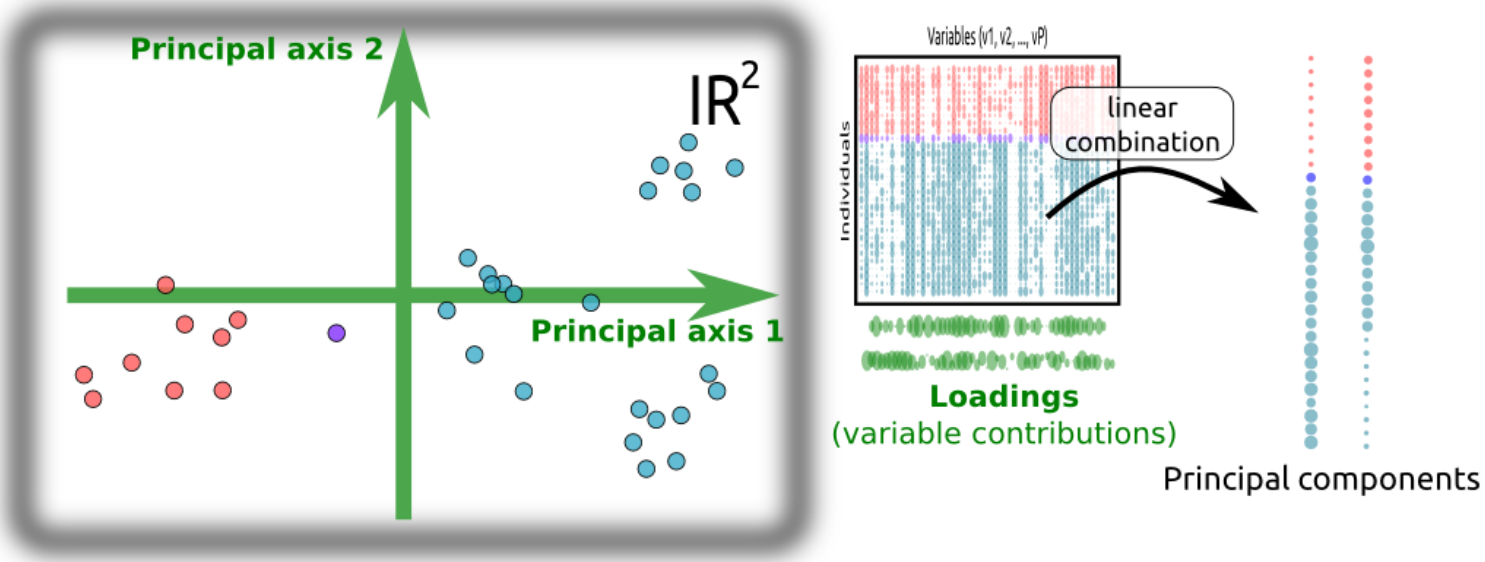


- $\mathbf{u}_1$ and $\mathbf{v}_1$: **1st principal axis and component**
- $\mathbf{u}_2$ and $\mathbf{v}_2$: **2nd principal axis and component**

$\rightarrow$ constraint: $\mathbf{u}_1 \perp \mathbf{u}_2$ (i.e., $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle_{\mathbf{Q}} = 0$)

$\rightarrow$ find $\mathbf{u}_2$ so that $\|\mathbf{v}_2\|^2_{\mathbf{D}}$ 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**

$\rightarrow$ constraint: $\mathbf{u}_1 \perp \mathbf{u}_2$ (i.e., $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle_{\mathbf{Q}} = 0$)
$\rightarrow$ find $\mathbf{u}_2$ so that $\|\mathbf{v}_2\|^2_{\mathbf{D}}$ is maximum

How do we do this?

Things that don't change:

- take $\mathbf{u}_i$ the i-th eigenvector of the $\mathbf{Q}$-symmetric matrix $\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}$
- (alternatively) take $\mathbf{v}_i$ the i-th eigenvector of the $\mathbf{D}$-symmetric matrix $\mathbf{X}\mathbf{Q}\mathbf{X}^T\mathbf{D}$

Things that change:

- pre-transformations of $\mathbf{X}$ (recoding, standardisation, etc.)
- metrics $\mathbf{Q}$ and $\mathbf{D}$ (implicitely distances in $\mathbb{R}^P$ and $\mathbb{R}^N$)
- most usual analyses are defined by $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

How do we do this?

Things that don't change:

- take $\mathbf{u}_i$ the i-th eigenvector of the $\mathbf{Q}$-symmetric matrix $\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}$
- (alternatively) take $\mathbf{v}_i$ the i-th eigenvector of the $\mathbf{D}$-symmetric matrix $\mathbf{X}\mathbf{Q}\mathbf{X}^T\mathbf{D}$

Things that change:

- pre-transformations of $\mathbf{X}$ (recoding, standardisation, etc.)
- metrics $\mathbf{Q}$ and $\mathbf{D}$ (implicitely distances in $\mathbb{R}^P$ and $\mathbb{R}^N$)
- most usual analyses are defined by $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

How do we do this?

Things that don't change:

- take $\mathbf{u}_i$ the i-th eigenvector of the $\mathbf{Q}$-symmetric matrix $\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}$
- (alternatively) take $\mathbf{v}_i$ the i-th eigenvector of the $\mathbf{D}$-symmetric matrix $\mathbf{X}\mathbf{Q}\mathbf{X}^T\mathbf{D}$

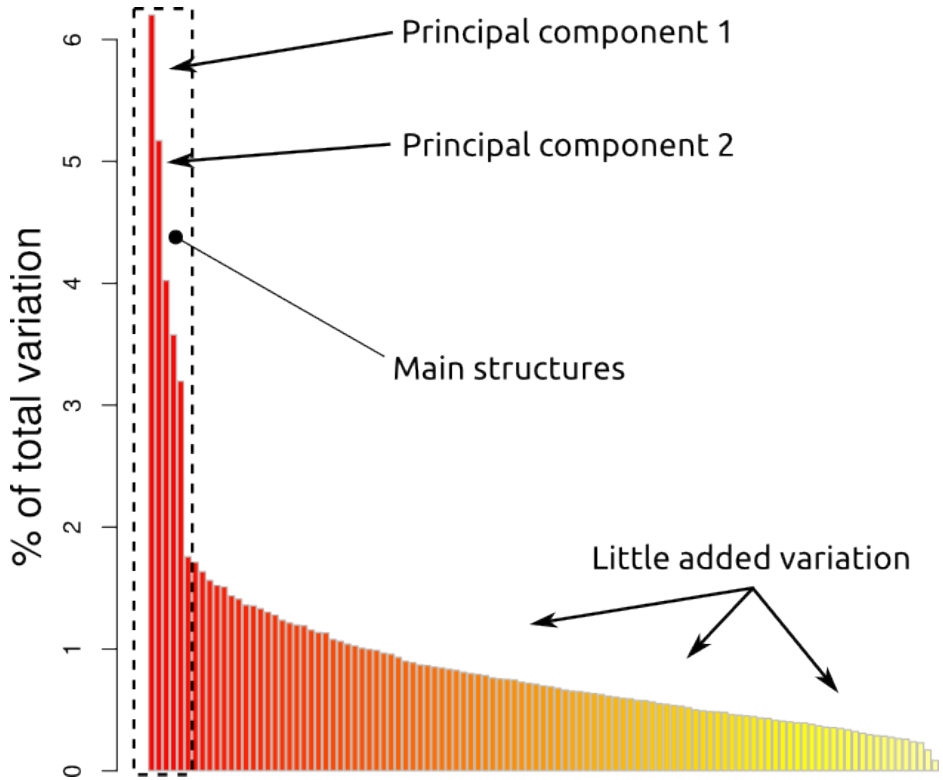
Things that change:

- pre-transformations of $\mathbf{X}$ (recoding, standardisation, etc.)
- metrics $\mathbf{Q}$ and $\mathbf{D}$ (implicitely distances in $\mathbb{R}^P$ and $\mathbb{R}^N$)
- most usual analyses are defined by $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$


packages: *ade4*, *vegan*

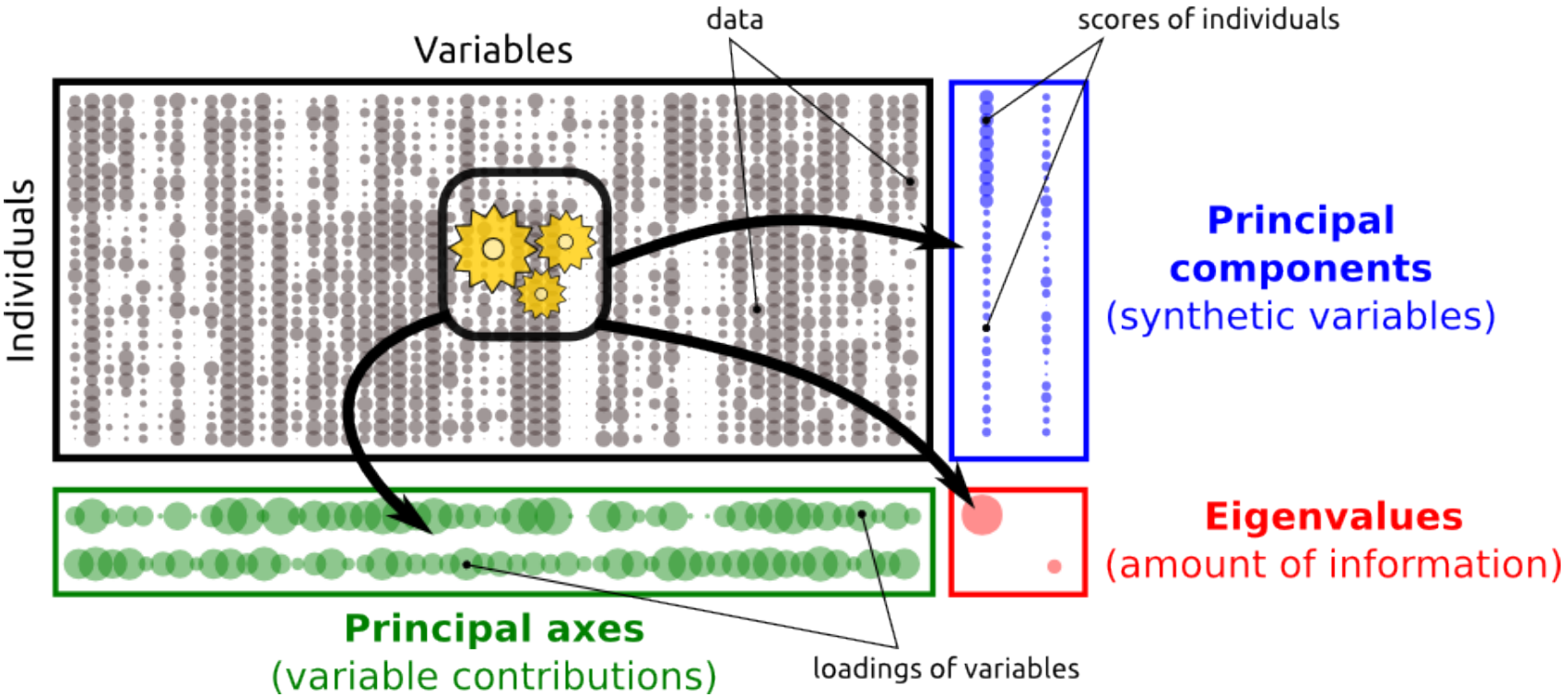
How many principal components to retain?

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



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

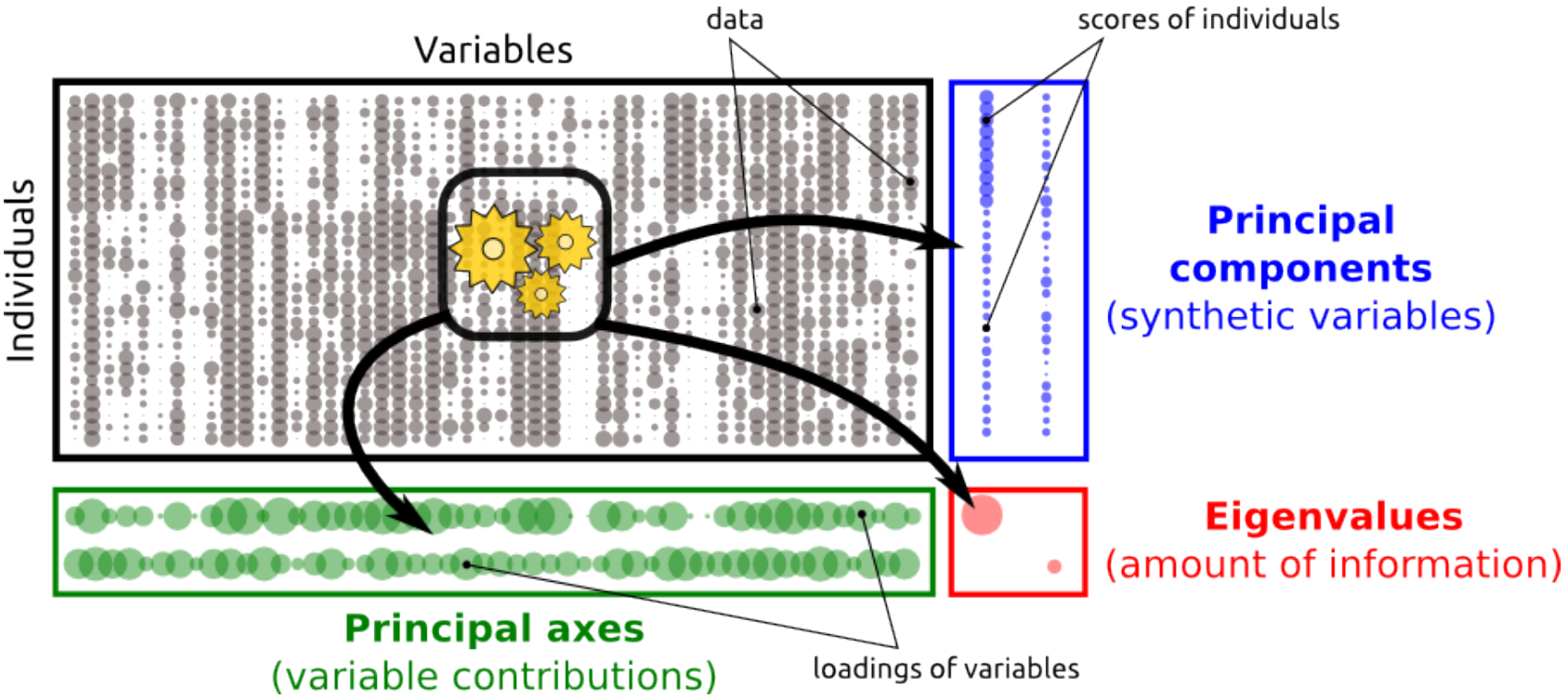
Outputs of multivariate analyses: an overview



Main outputs:

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

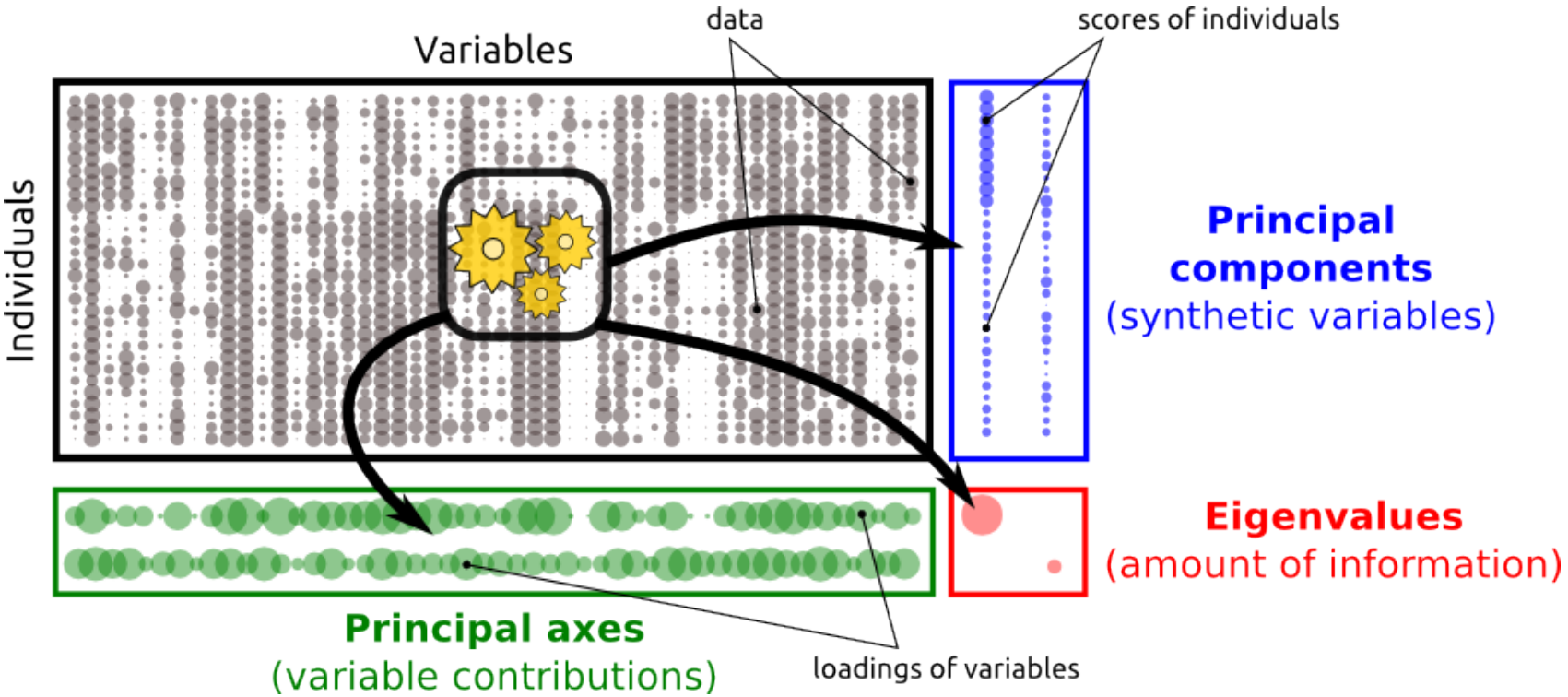
Outputs of multivariate analyses: an overview



Main outputs:

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

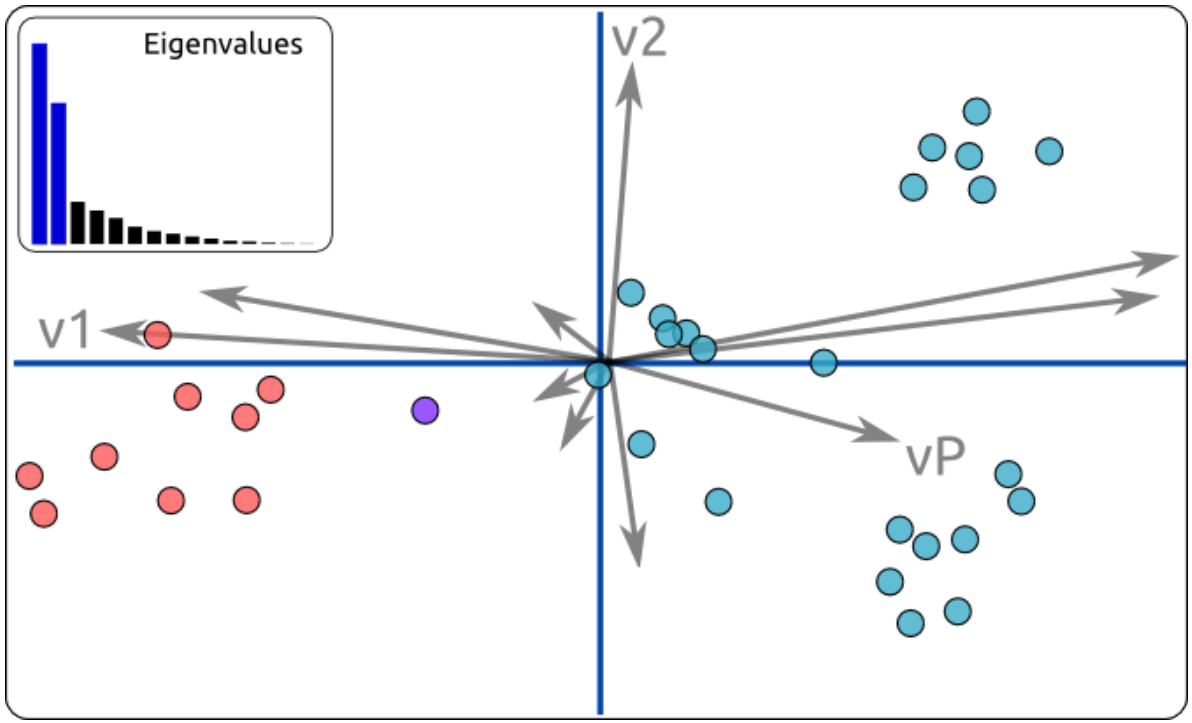
Outputs of multivariate analyses: an overview



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
- structuring 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

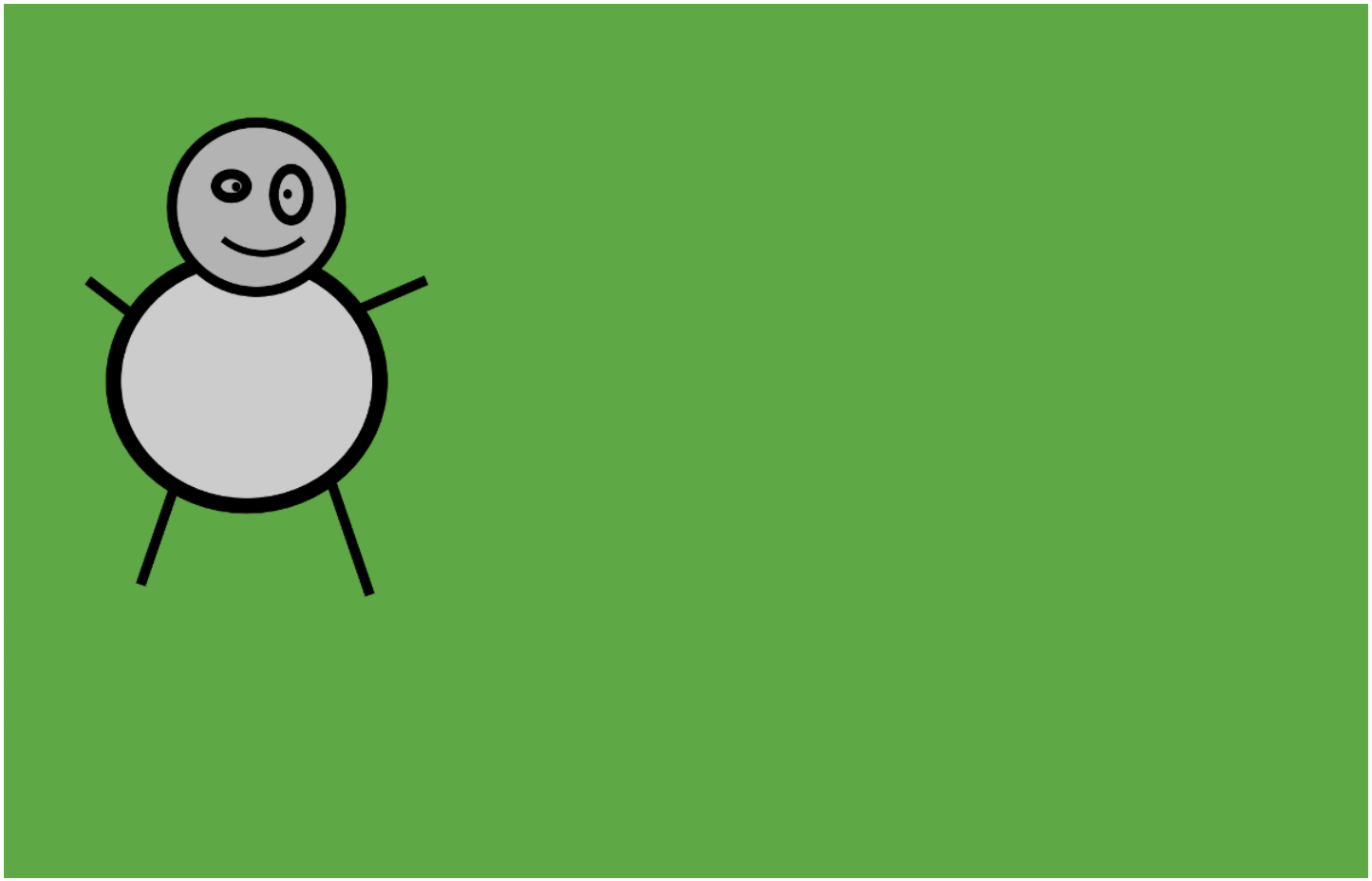
- **variety of methods** for different types of variables
- **principal components** (PCs) summarize diversity
- **variable loadings** identify discriminating variables
- other uses of PCs: **maps** (spatial structures), **models** (response variables or predictors), ...

Multivariate analysis in a nutshell

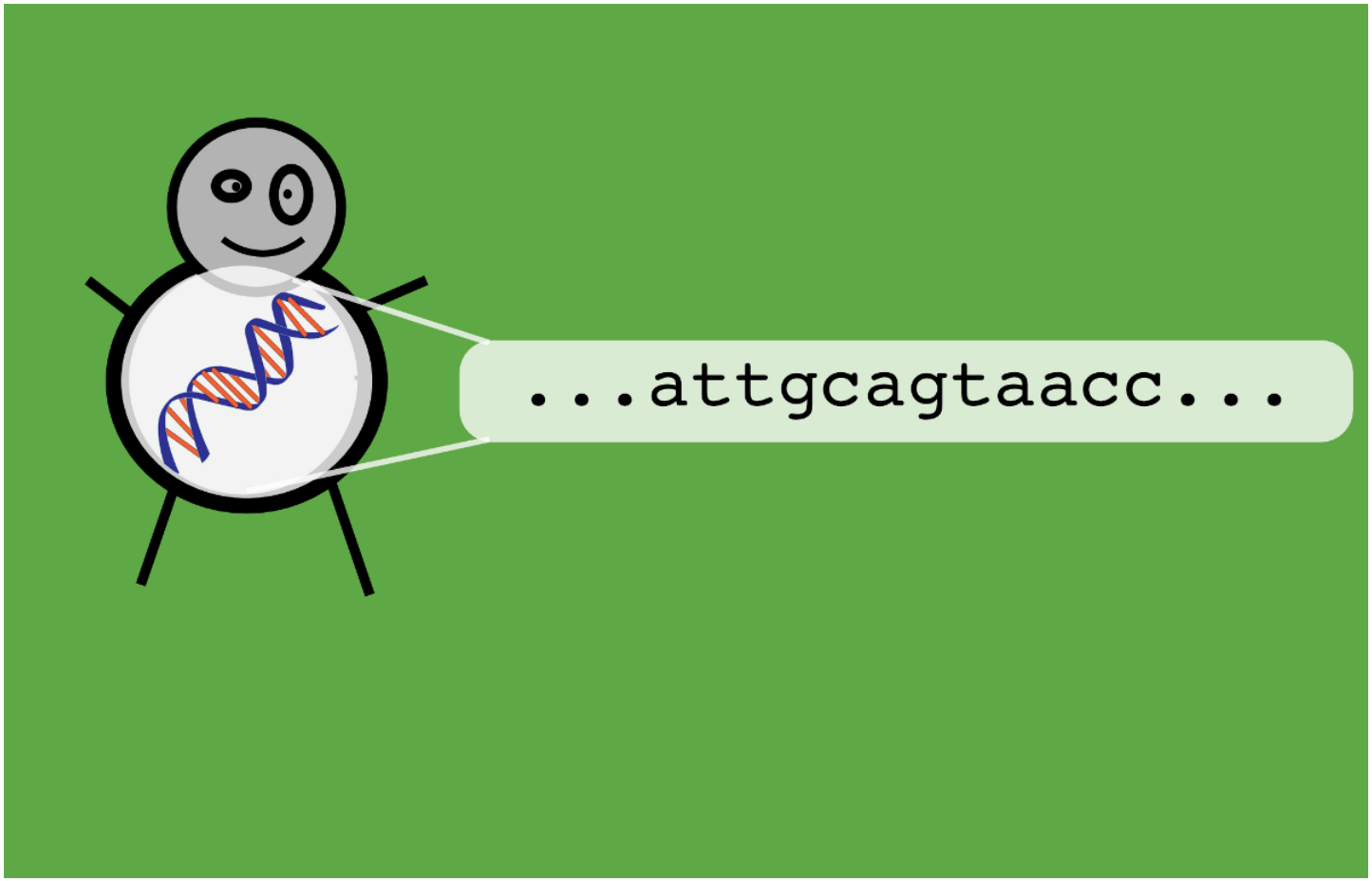
- **variety of methods** for different types of variables
- **principal components** (PCs) summarize diversity
- **variable loadings** identify discriminating variables
- other uses of PCs: **maps** (spatial structures), **models** (response variables or predictors), ...

Outline

From DNA sequences to patterns of biological diversity

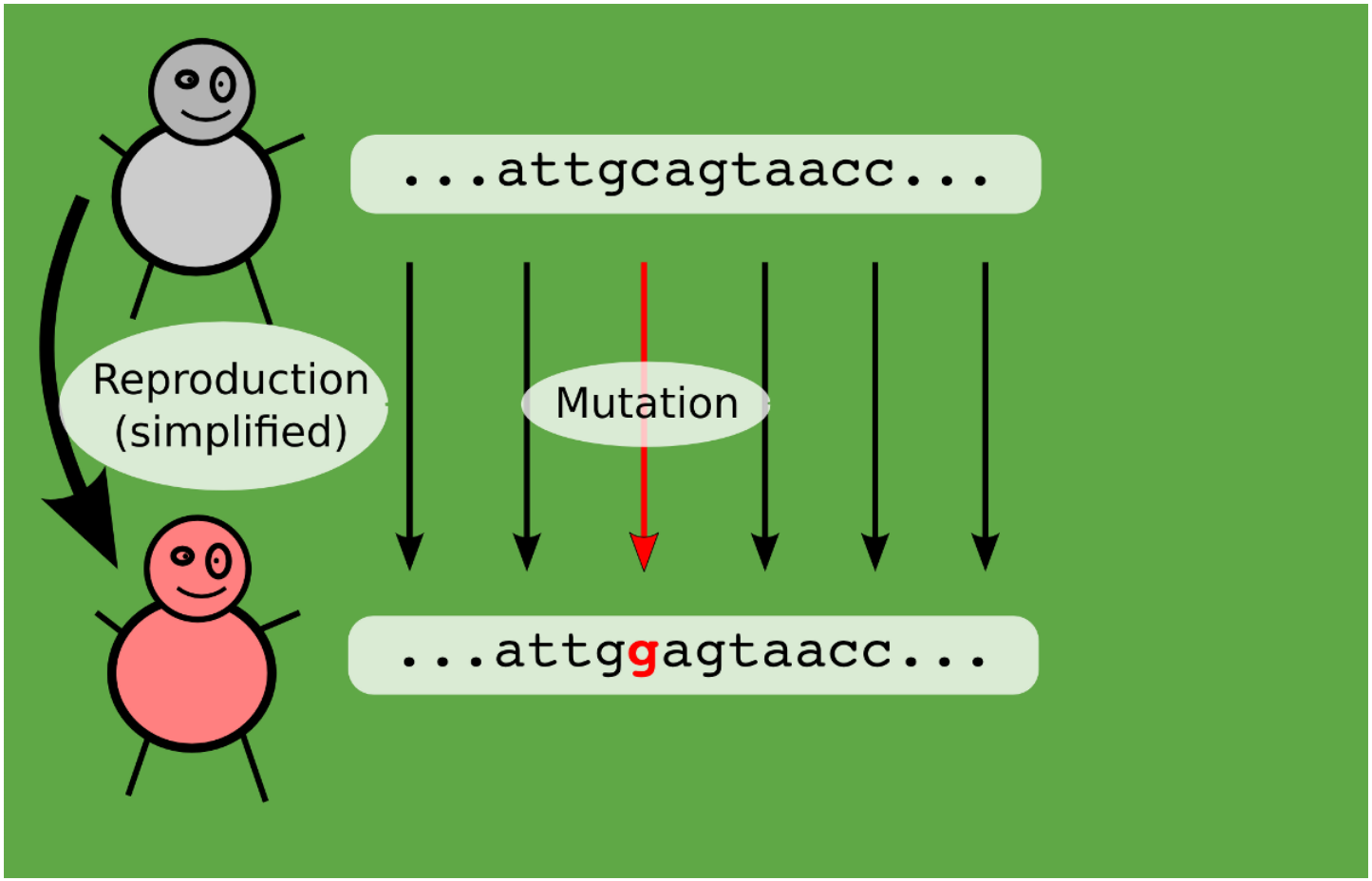


From DNA sequences to patterns of biological diversity

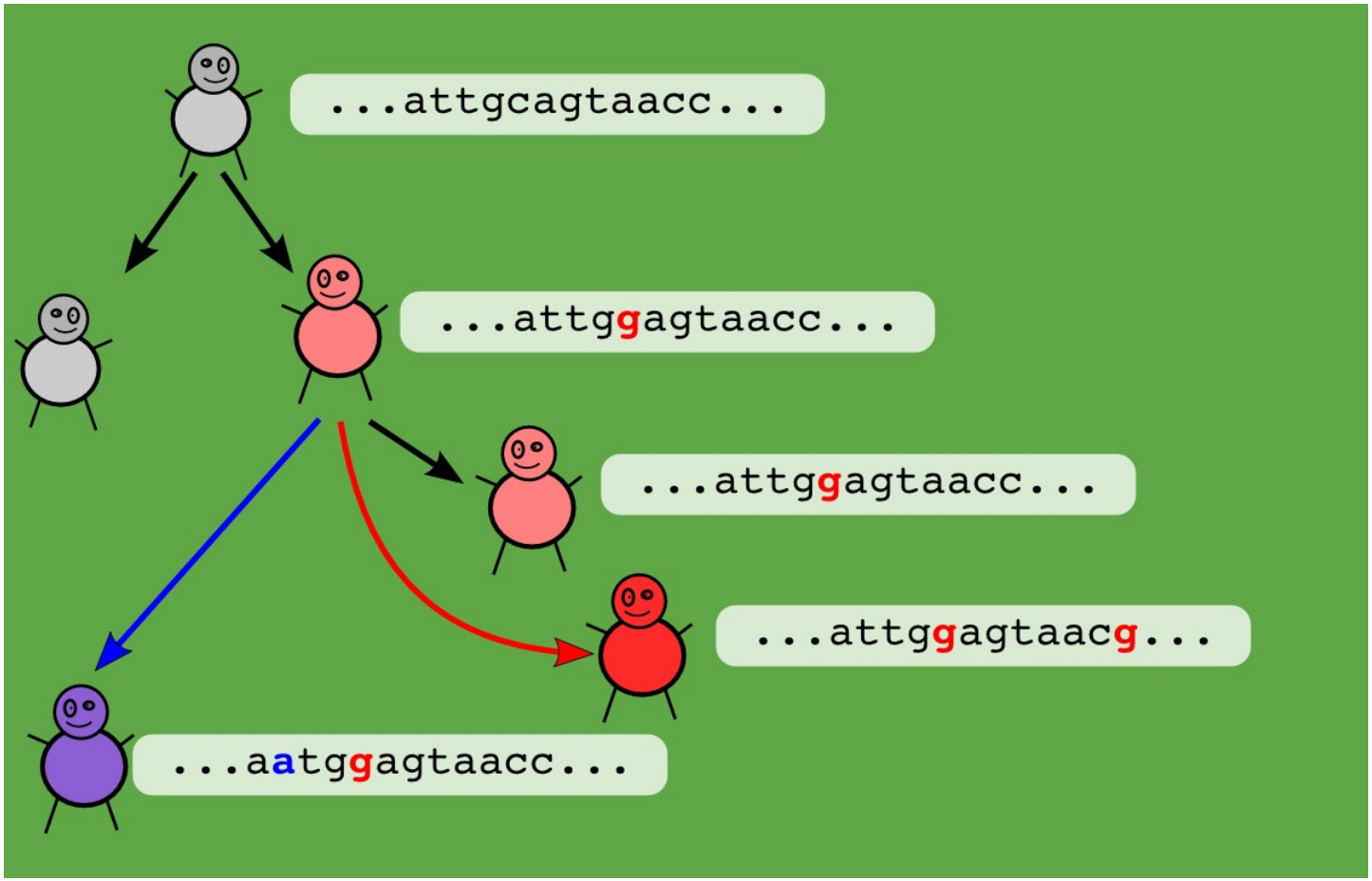


From DNA sequences to patterns of biological diversity

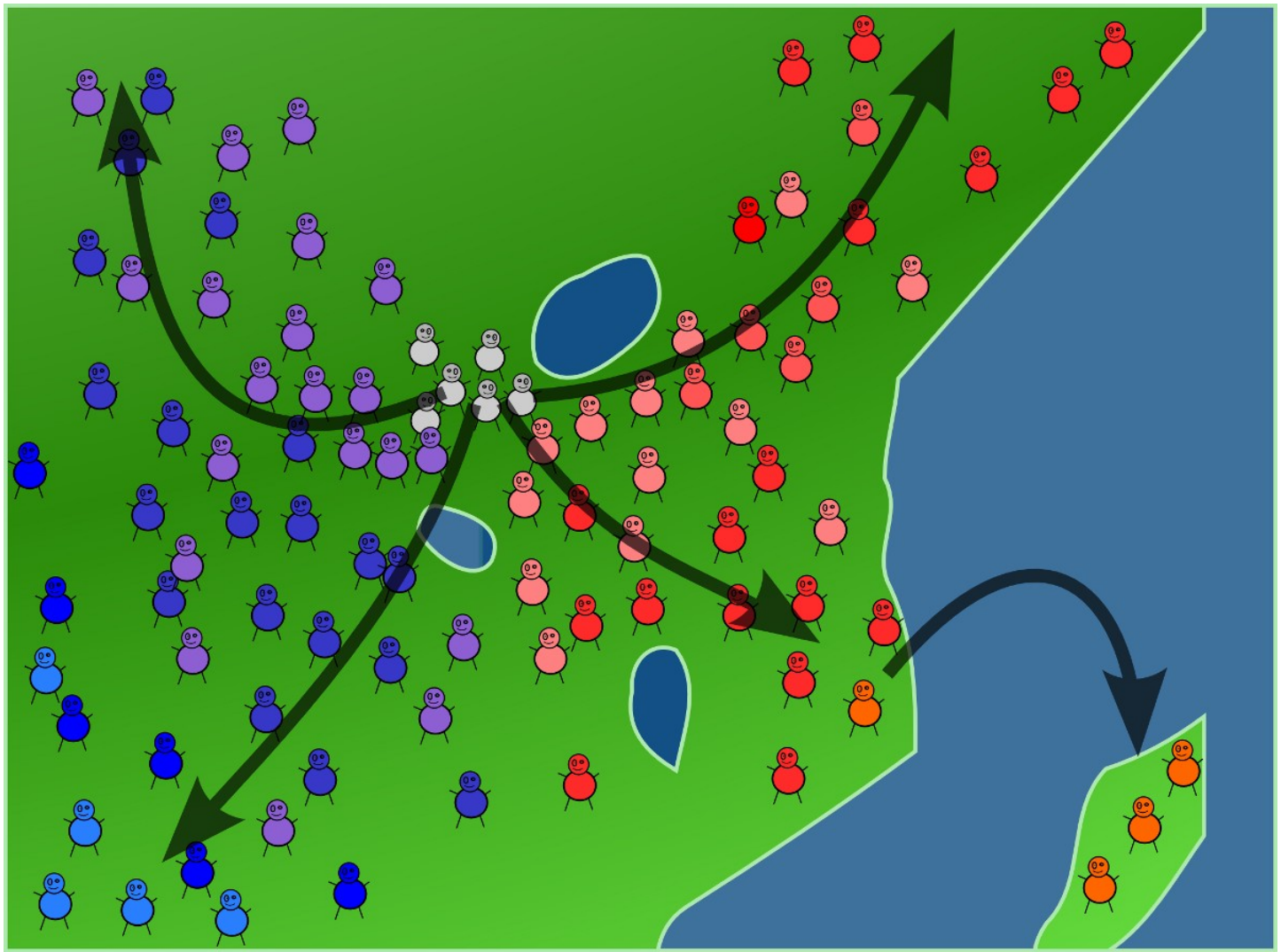
From DNA sequences to patterns of biological diversity



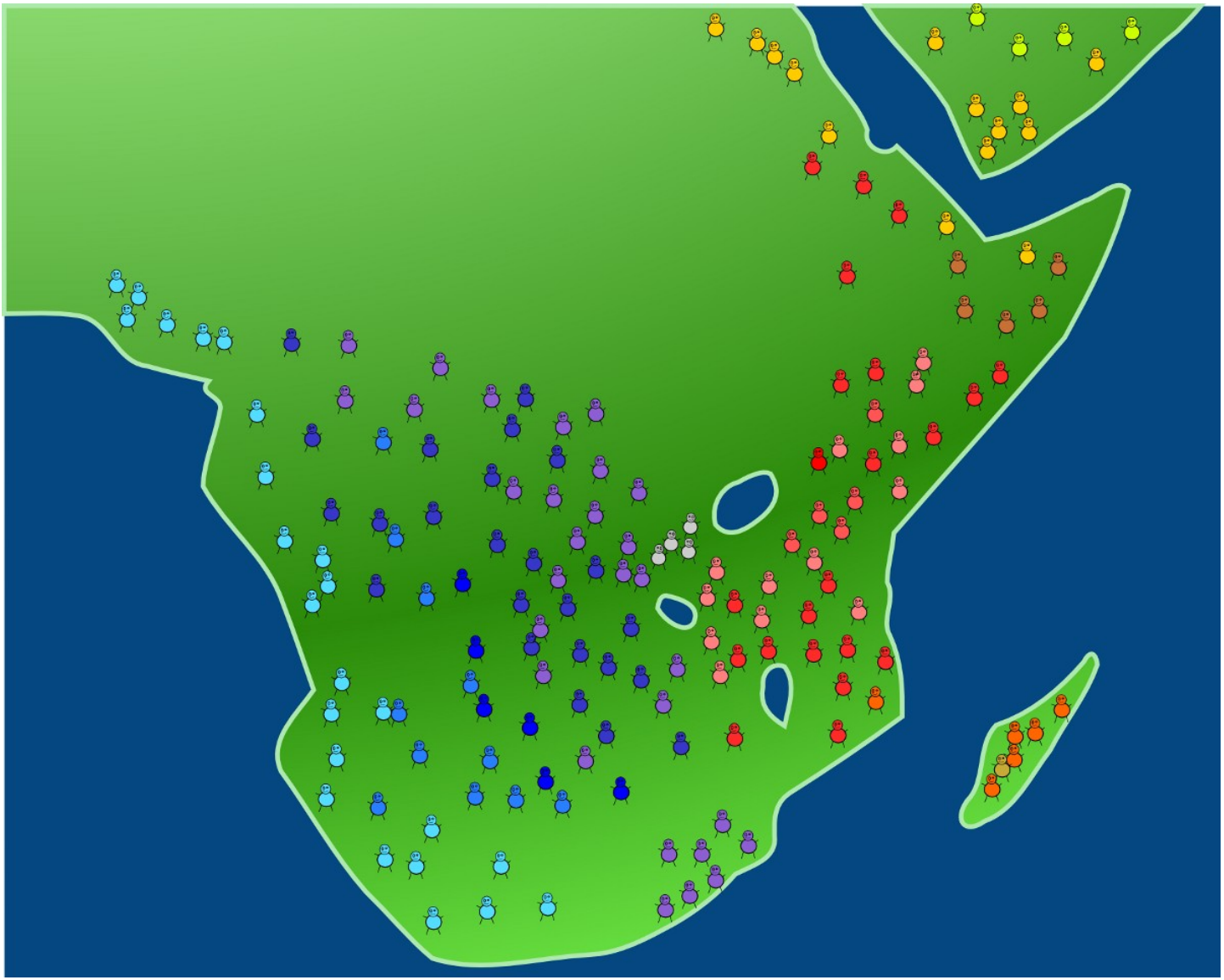
From DNA sequences to patterns of biological diversity



From DNA sequences to patterns of biological diversity

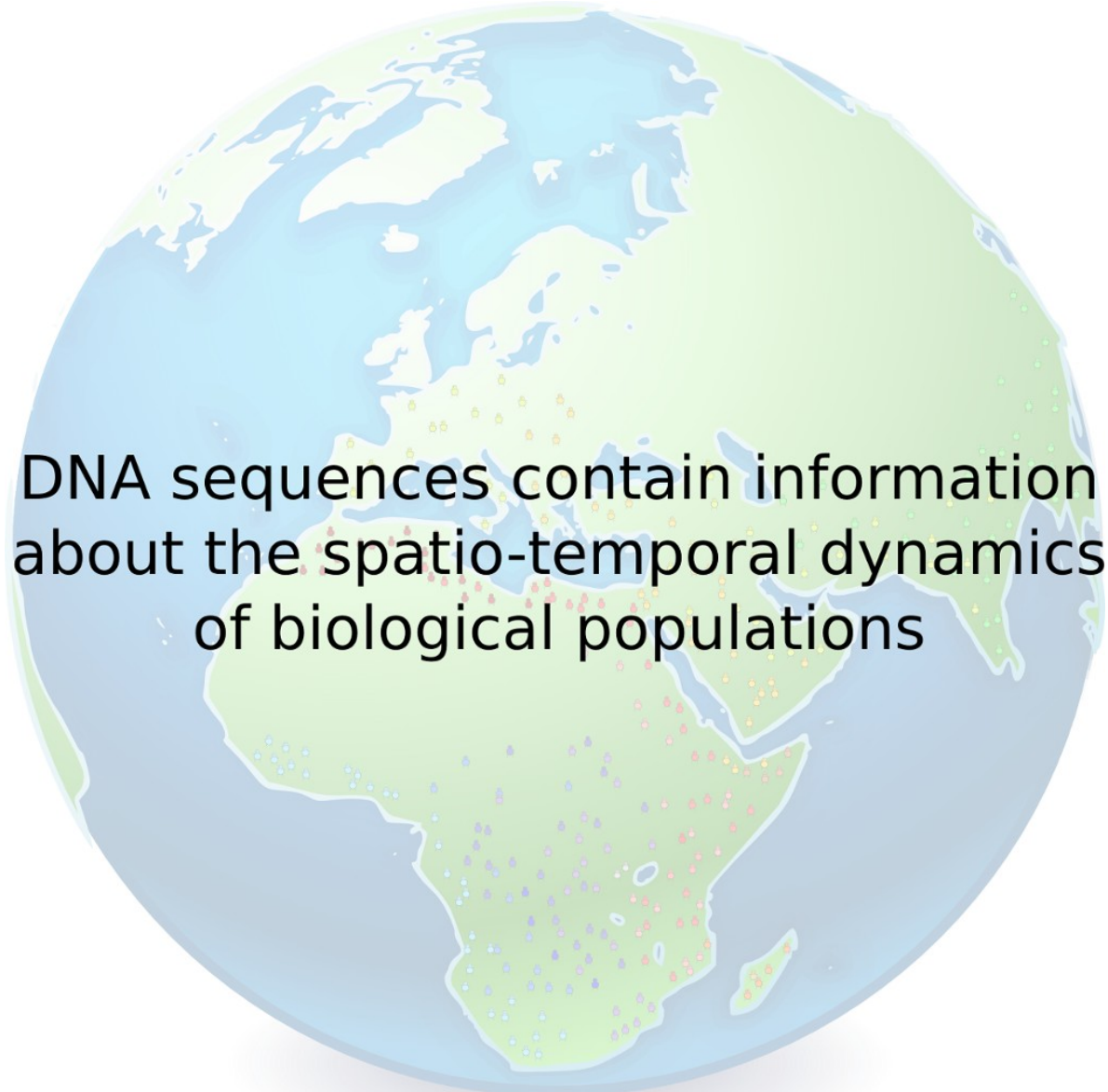


From DNA sequences to patterns of biological diversity

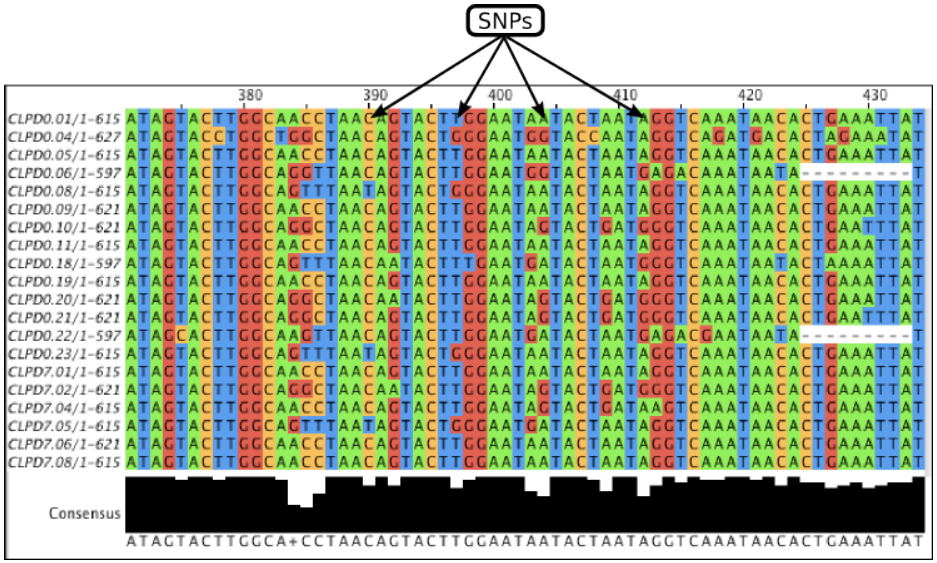


From DNA sequences to patterns of biological diversity

From DNA sequences to patterns of biological diversity



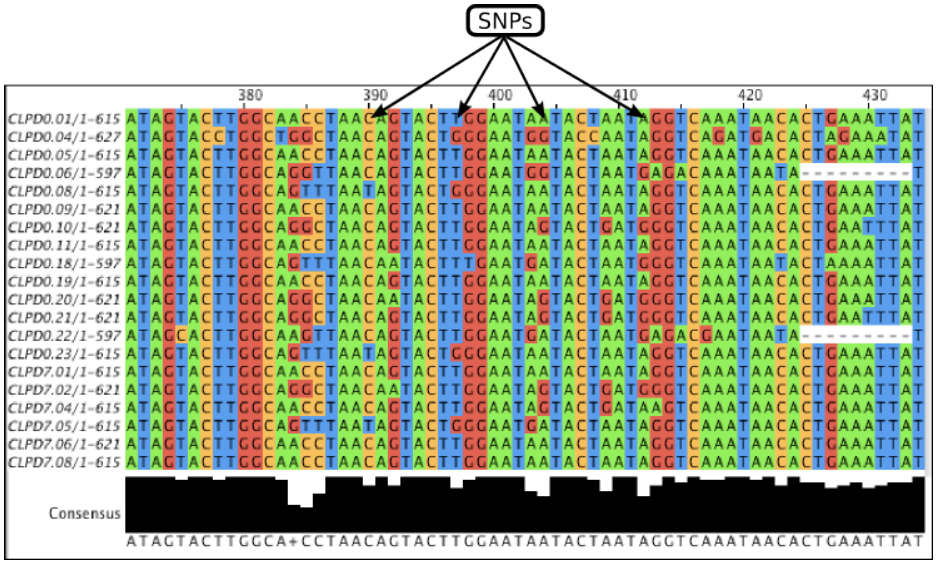
DNA sequences: a rich source of information



- hundreds/thousands individuals
- up to millions of single nucleotide polymorphism (SNPs)
- more generally, most genetic data can be treated as frequencies

$\Rightarrow$ **Multivariate analysis use to summarize genetic diversity.**

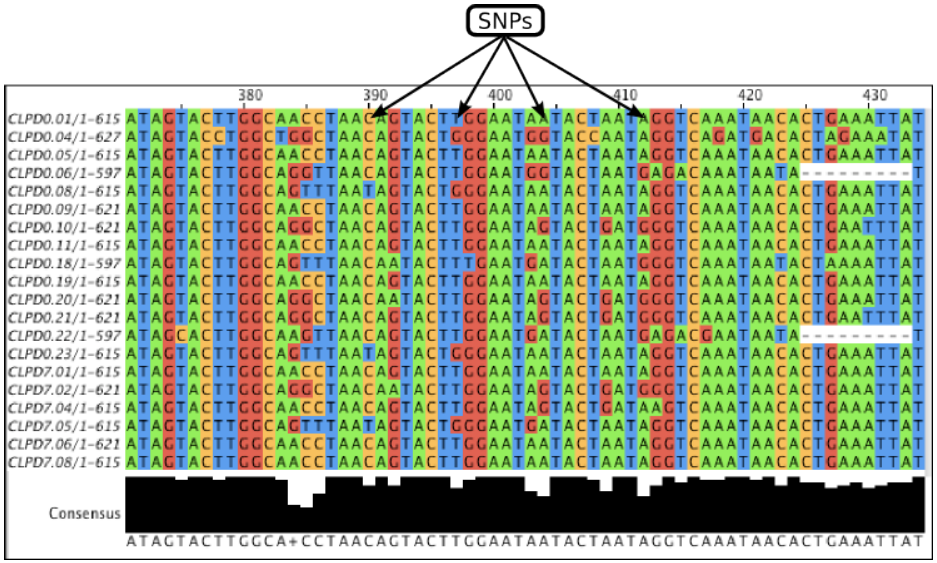
DNA sequences: a rich source of information



- hundreds/thousands individuals
- up to millions of single nucleotide polymorphism (SNPs)
- more generally, most genetic data can be treated as frequencies

$\Rightarrow$ **Multivariate analysis use to summarize genetic diversity.**

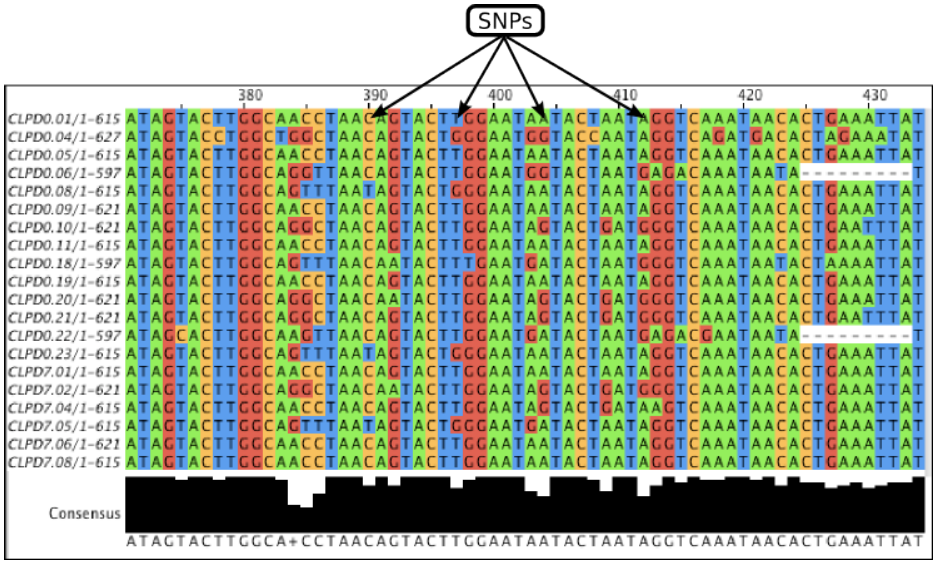
DNA sequences: a rich source of information



- hundreds/thousands individuals
- up to millions of single nucleotide polymorphism (SNPs)
- more generally, most genetic data can be treated as frequencies

$\Rightarrow$ **Multivariate analysis use to summarize genetic diversity.**

DNA sequences: a rich source of information

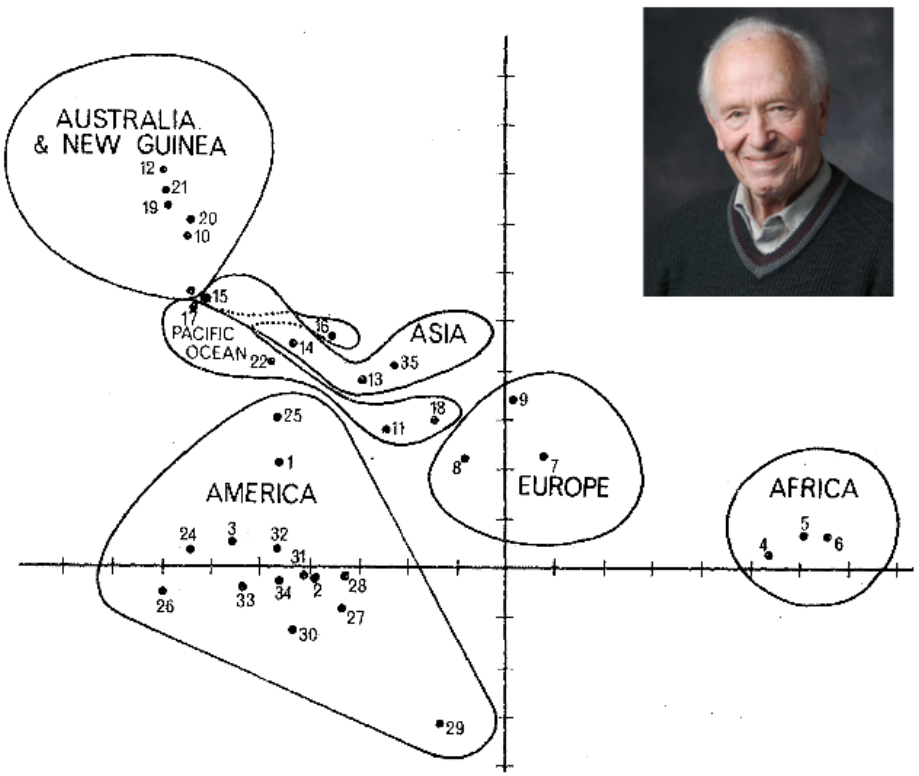


- hundreds/thousands individuals
- up to millions of single nucleotide polymorphism (SNPs)
- more generally, most genetic data can be treated as frequencies

$\Rightarrow$ **Multivariate analysis use to summarize genetic diversity.**

First application of multivariate analysis in genetics

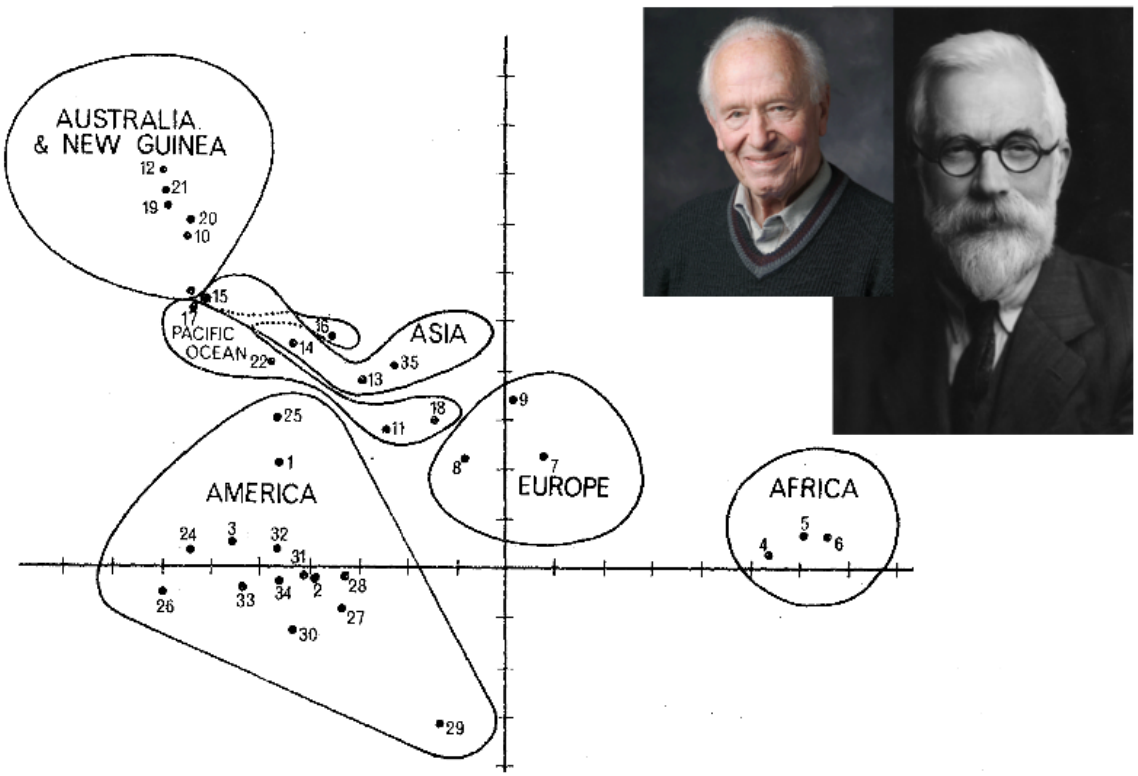
PCA of genetic data, native human populations (Cavalli-Sforza 1966, *Proc B*)



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...

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)
- ...

 packages: *adegenet*, *ade4*, *pegas*

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

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)

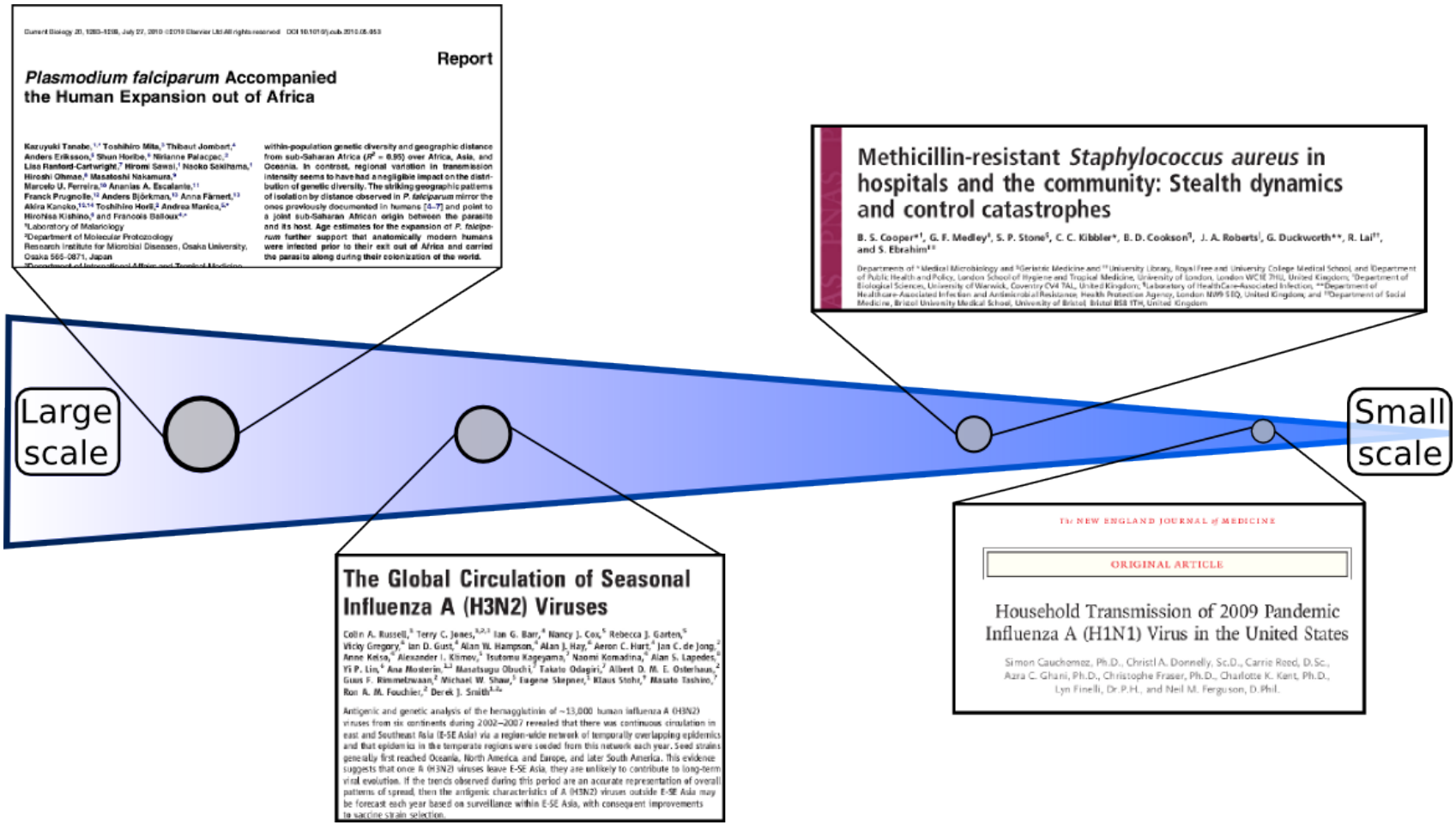
Why investigate the diversity of pathogen populations?

Genetic data: increasingly important in infectious disease epidemiology

Purposes

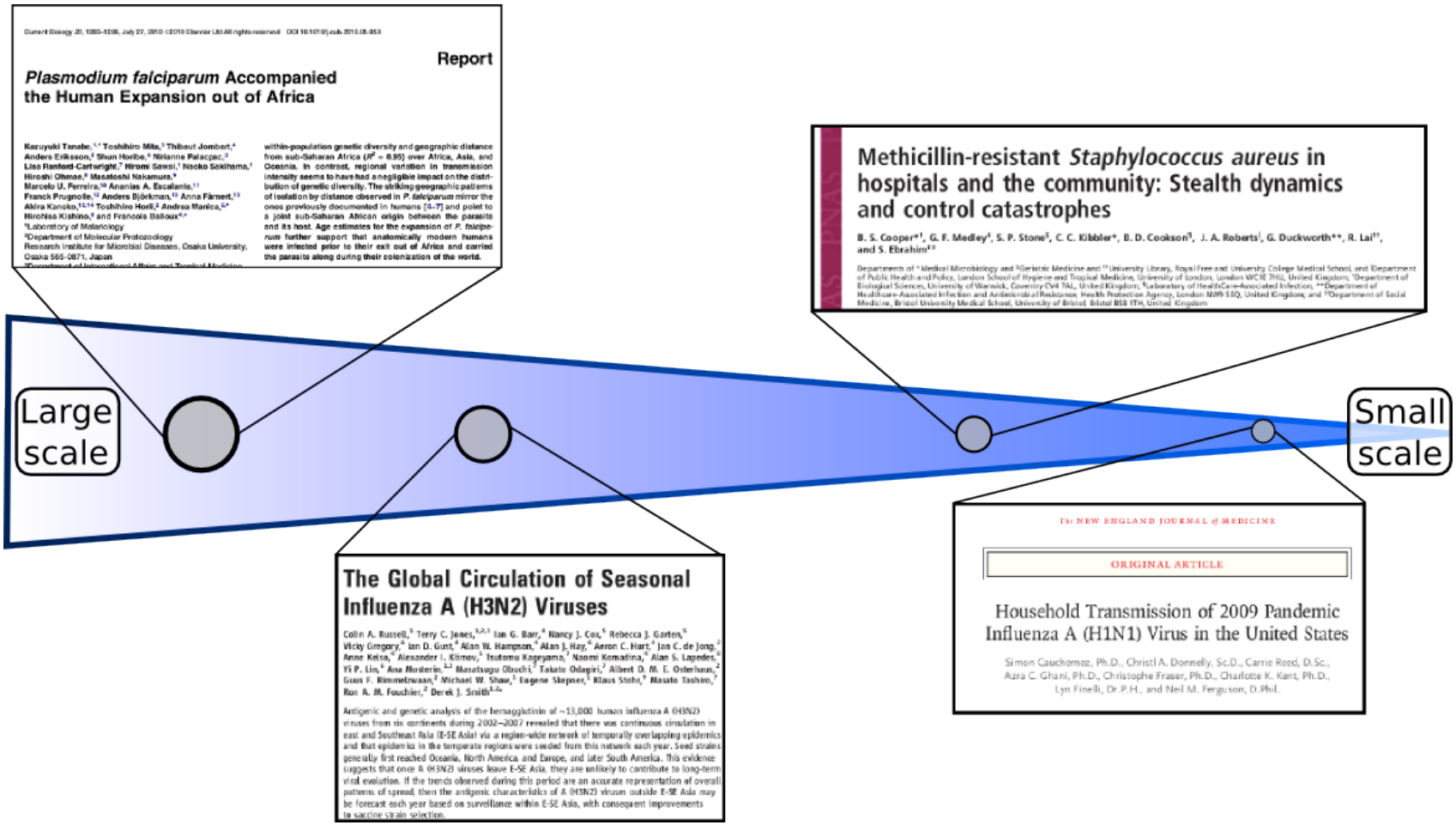
- classify pathogens, describe their relationships
- assess the spatio-temporal dynamics of infectious diseases
- reconstruct epidemiological processes (transmission)

Different questions at different scales



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

Different questions at different scales



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).

$\Rightarrow$ 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).

$\Rightarrow$ 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).

$\Rightarrow$ 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).

$\Rightarrow$ 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).

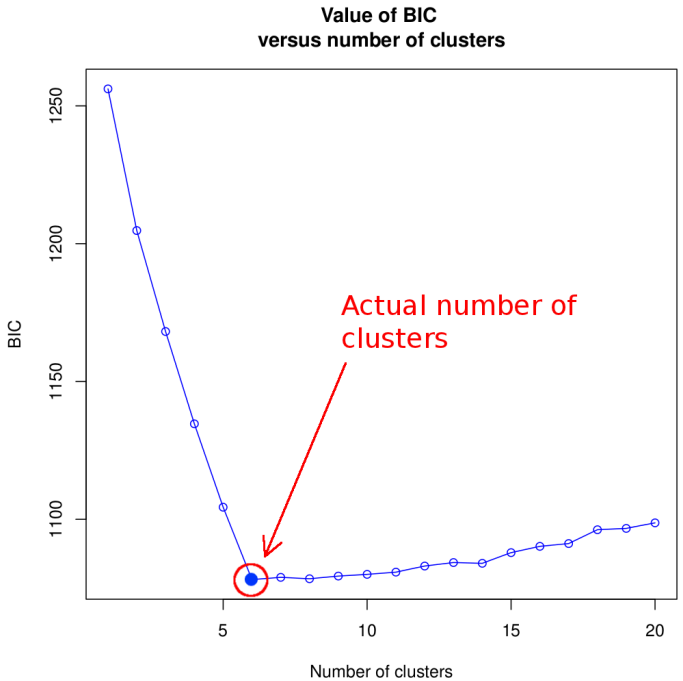
$\Rightarrow$ aim: identify and describe genetic clusters

Genetic clustering using K-means & BIC

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

Variance partitioning model (ANOVA):

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



Performances:

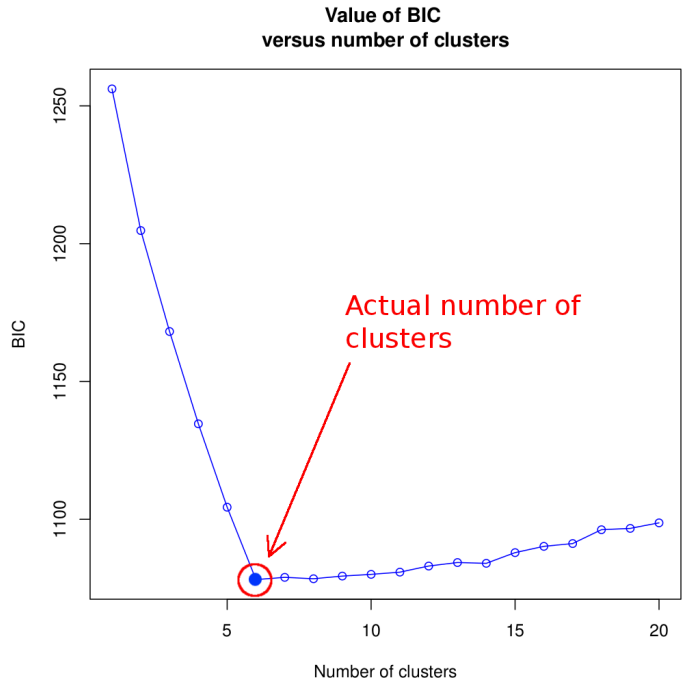
- K-means $\geq$ STRUCTURE on simulated data (various island and stepping stone models)
- orders of magnitude faster (seconds vs hours/days)

Genetic clustering using K-means & BIC

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

Variance partitioning model (ANOVA):

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



Performances:

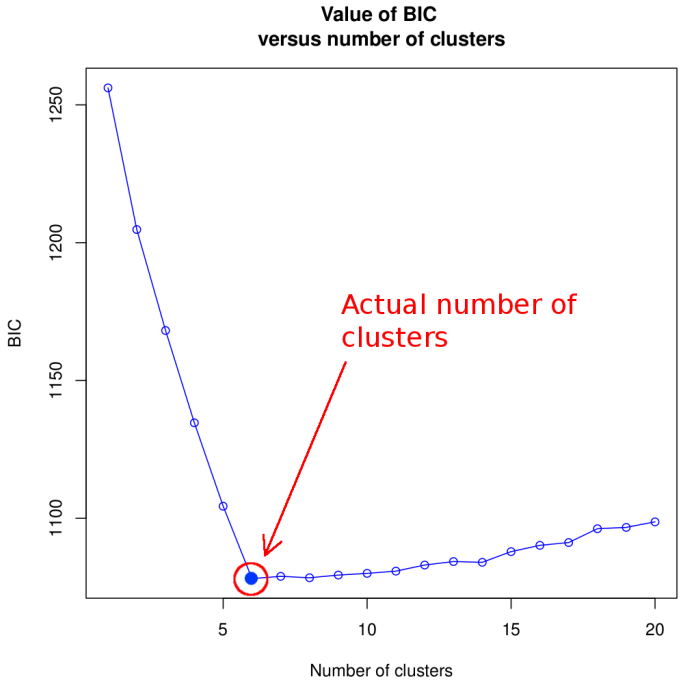
- K-means $\geq$ STRUCTURE on simulated data (various island and stepping stone models)
- orders of magnitude faster (seconds vs hours/days)

Genetic clustering using K-means & BIC

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

Variance partitioning model (ANOVA):

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



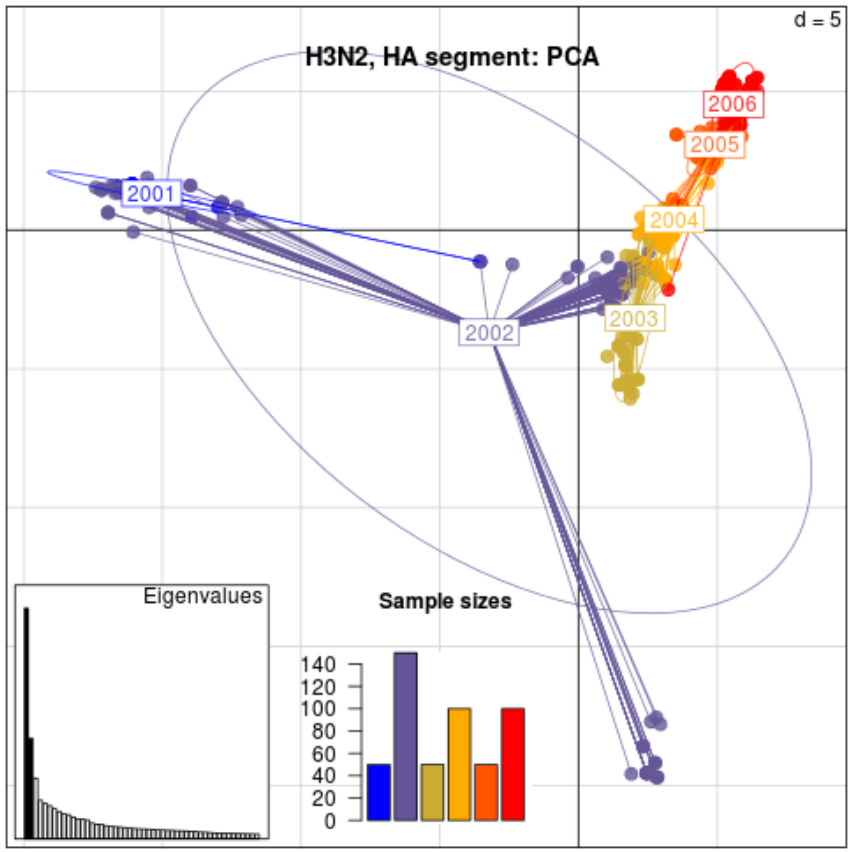
Performances:

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

package: *adegenet*, function `find.clusters`

PCA of seasonal influenza (A/H3N2) data

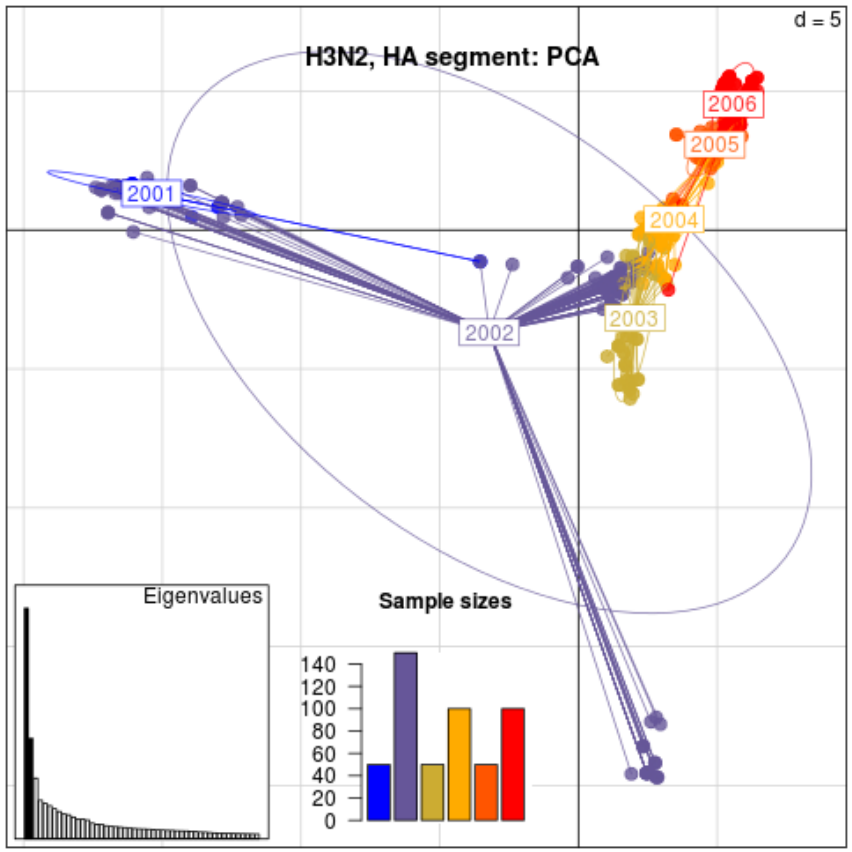
Data: seasonal influenza (A/H3N2), 500 HA segments.



Little temporal evolution, burst of diversity in 2002??

PCA of seasonal influenza (A/H3N2) data

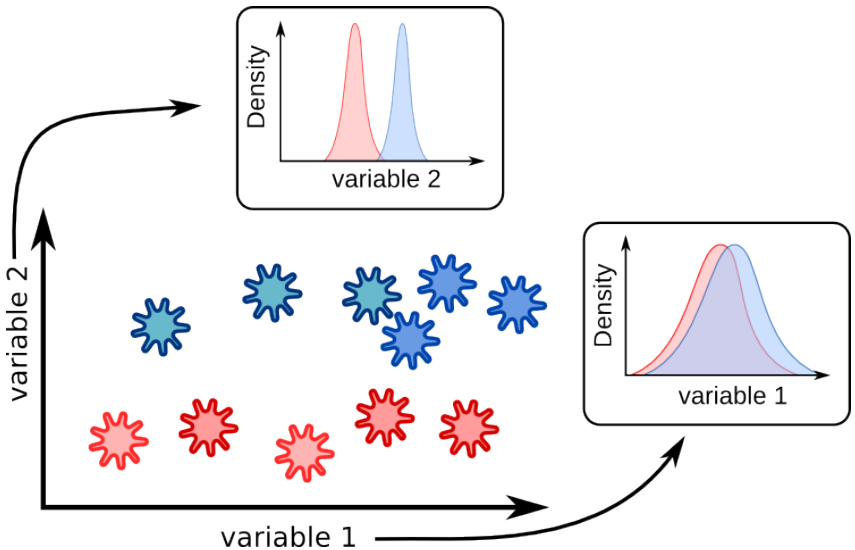
Data: seasonal influenza (A/H3N2), 500 HA segments.



Little temporal evolution, burst of diversity in 2002??

Which diversity to represent?

Total diversity not relevant to analyse clusters.



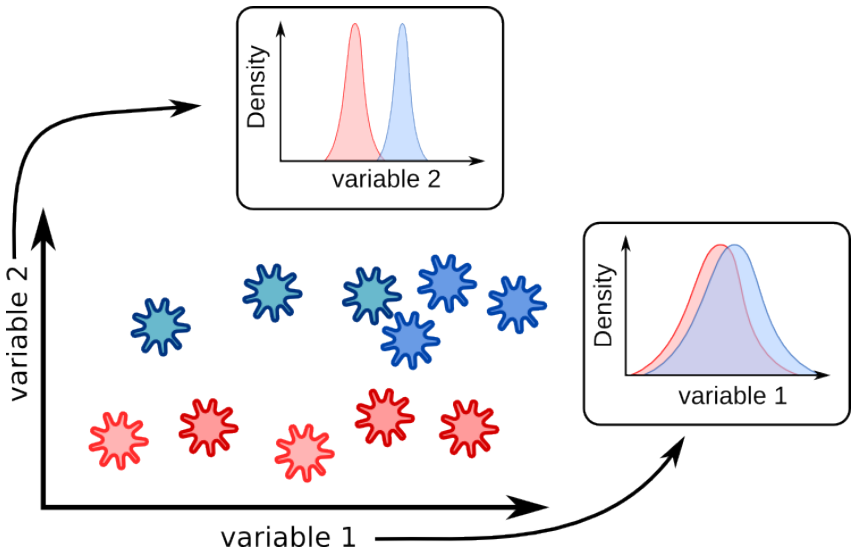
Discriminant Analysis of Principal Components (DAPC):

(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.



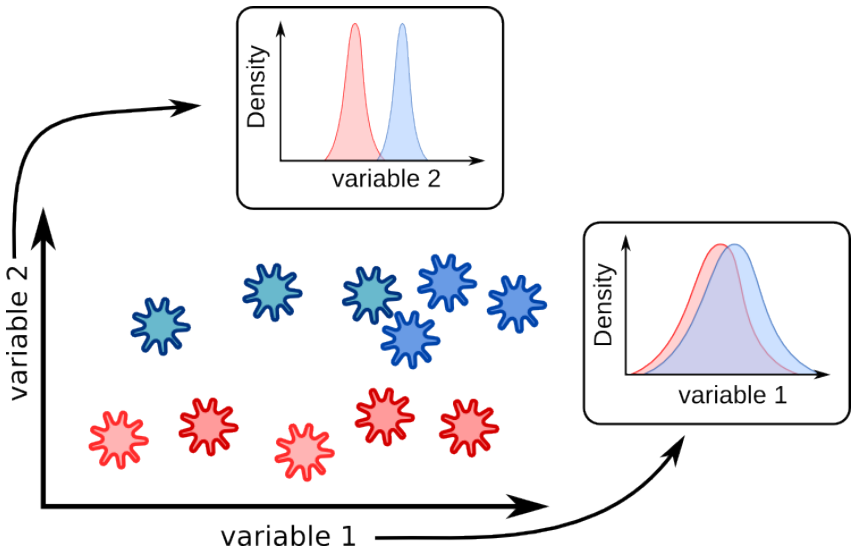
Discriminant Analysis of Principal Components (DAPC):

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

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

Which diversity to represent?

Total diversity not relevant to analyse clusters.



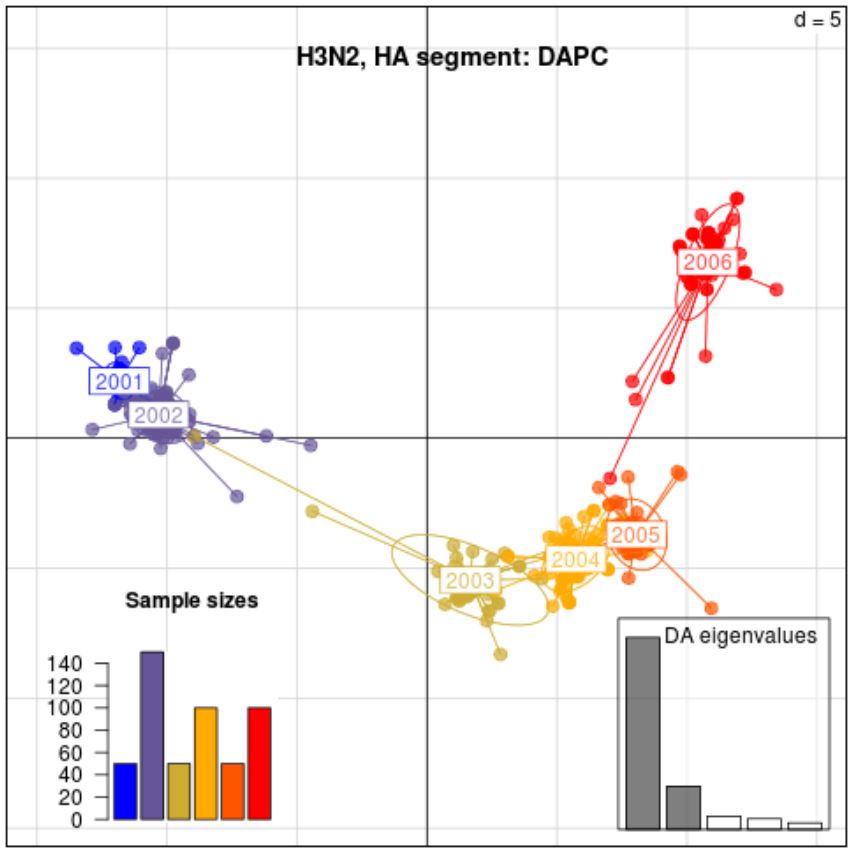
Discriminant Analysis of Principal Components (DAPC):

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

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

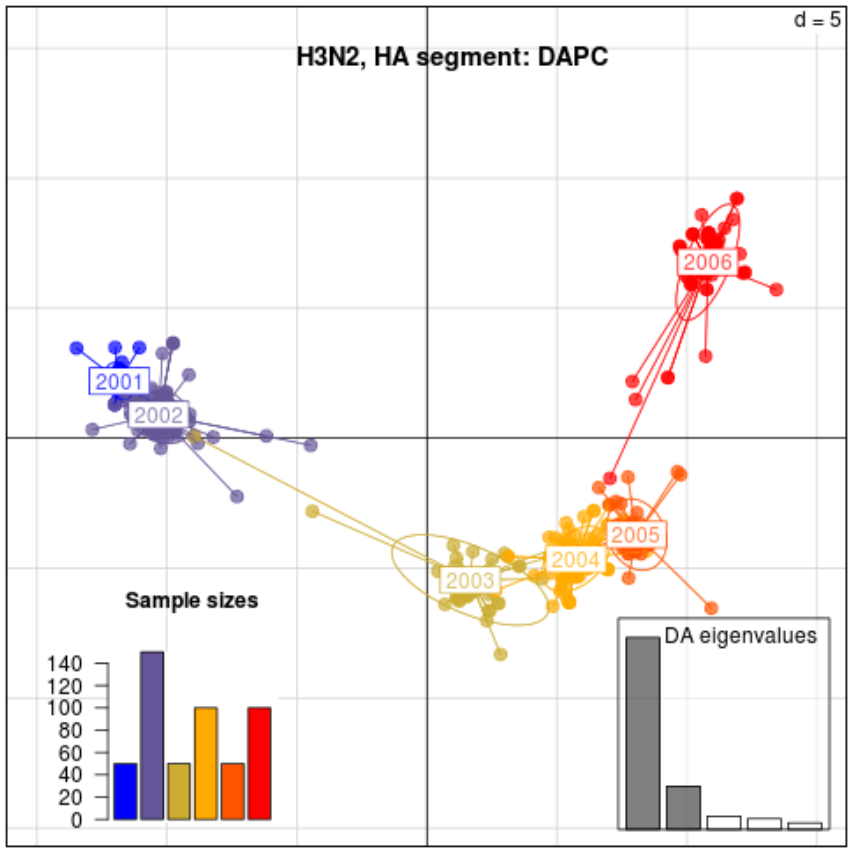
package: *adegenet*, function dapc

DAPC of seasonal influenza (A/H3N2) data



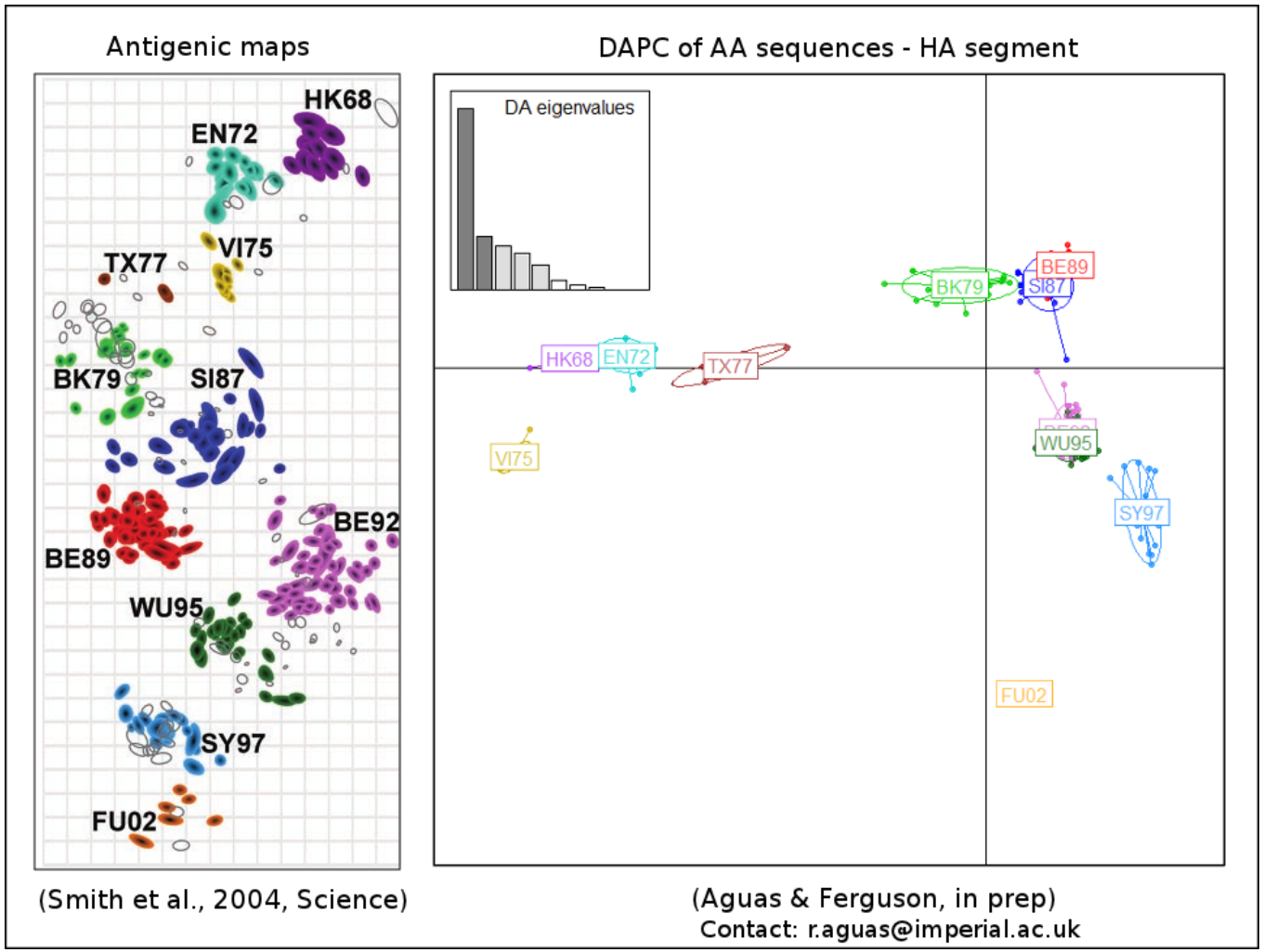
Strong temporal signal, originality of 2006 isolates (new alleles).

DAPC of seasonal influenza (A/H3N2) data



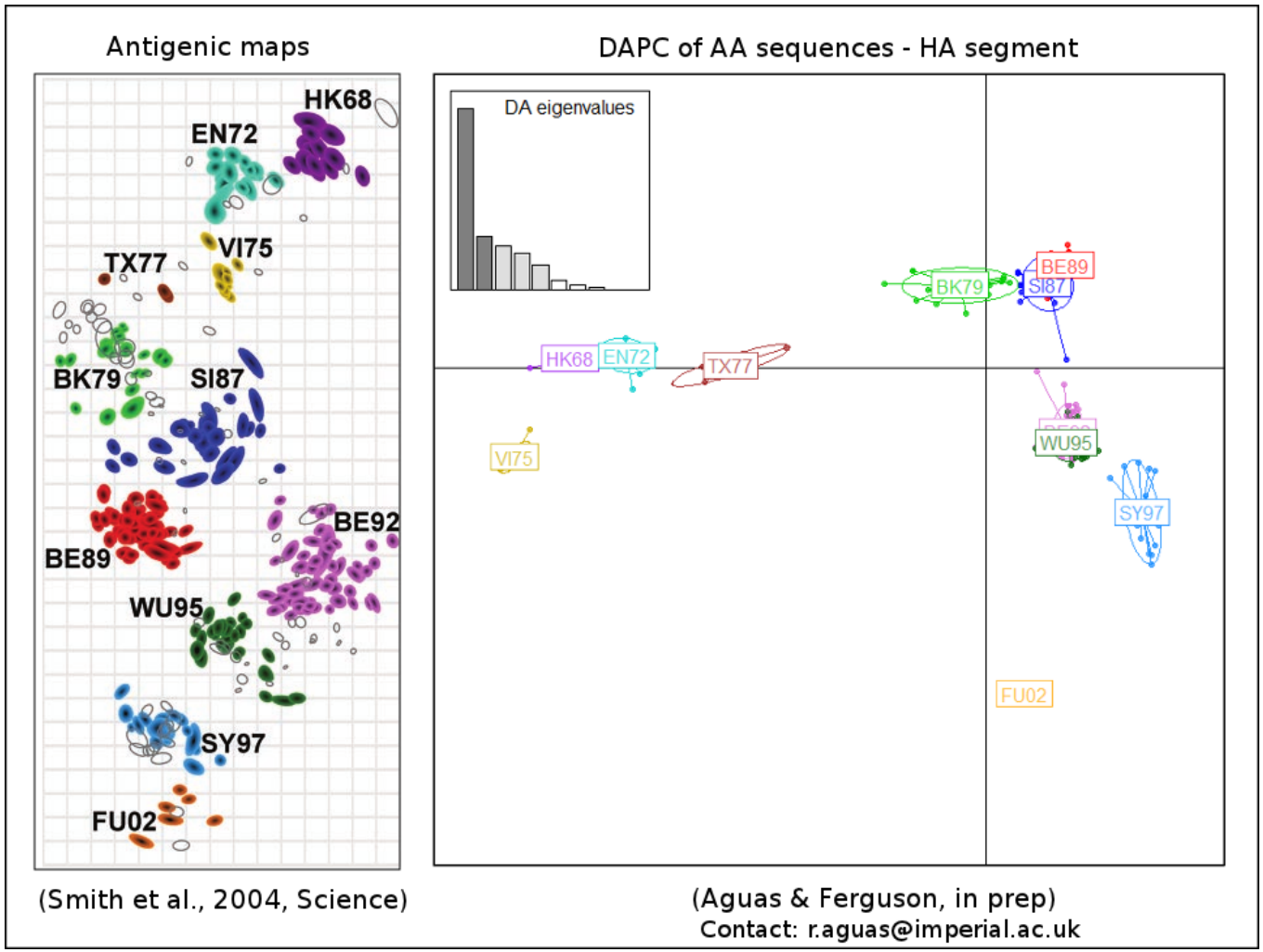
Strong temporal signal, originality of 2006 isolates (new alleles).

Identifying antigenic clusters in influenza (A/H3N2)



Antigenic clusters identified directly from AA sequences.

Identifying antigenic clusters in influenza (A/H3N2)



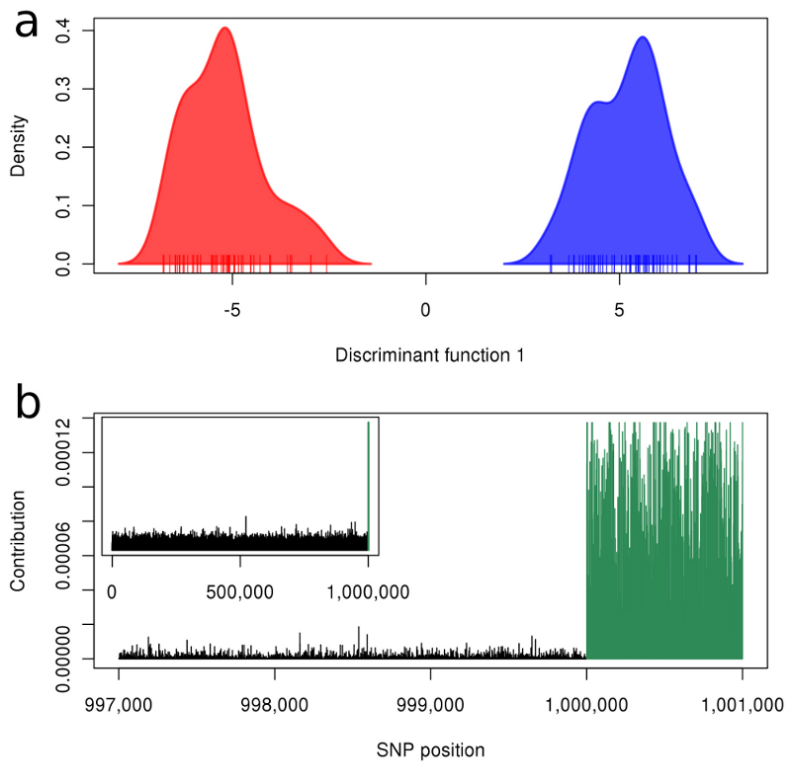
Antigenic clusters identified directly from AA sequences.

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)



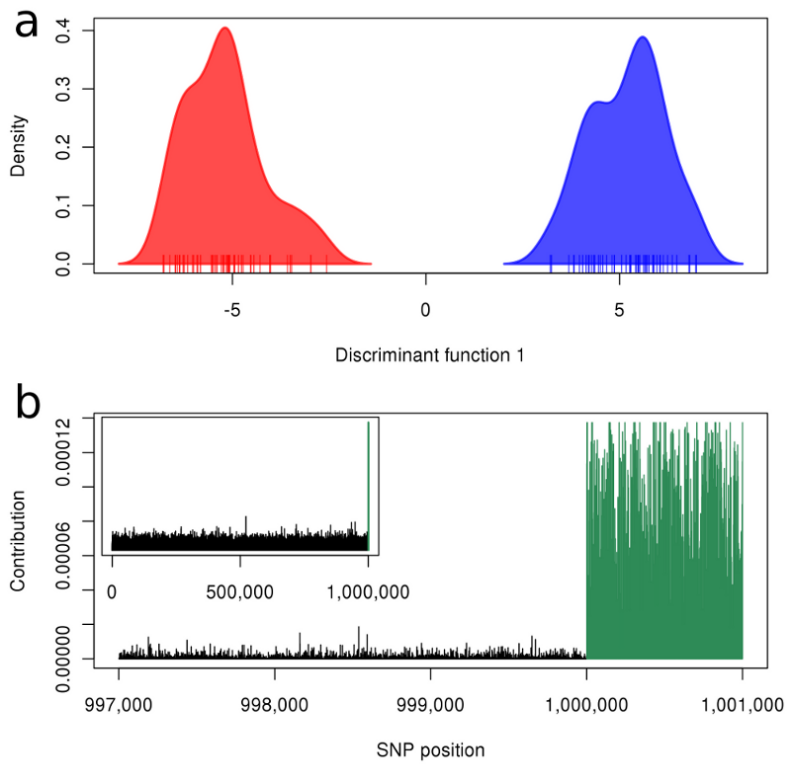
Possible applications to pathogen GWAS (e.g. SNPs related to antibiotic resistance in bacteria).

DAPC to identify structuring alleles

DAPC finds combinations of alleles most differing between groups.

Simulated data:
(Jombart & Ahmed 2011, *Bioinformatics*)

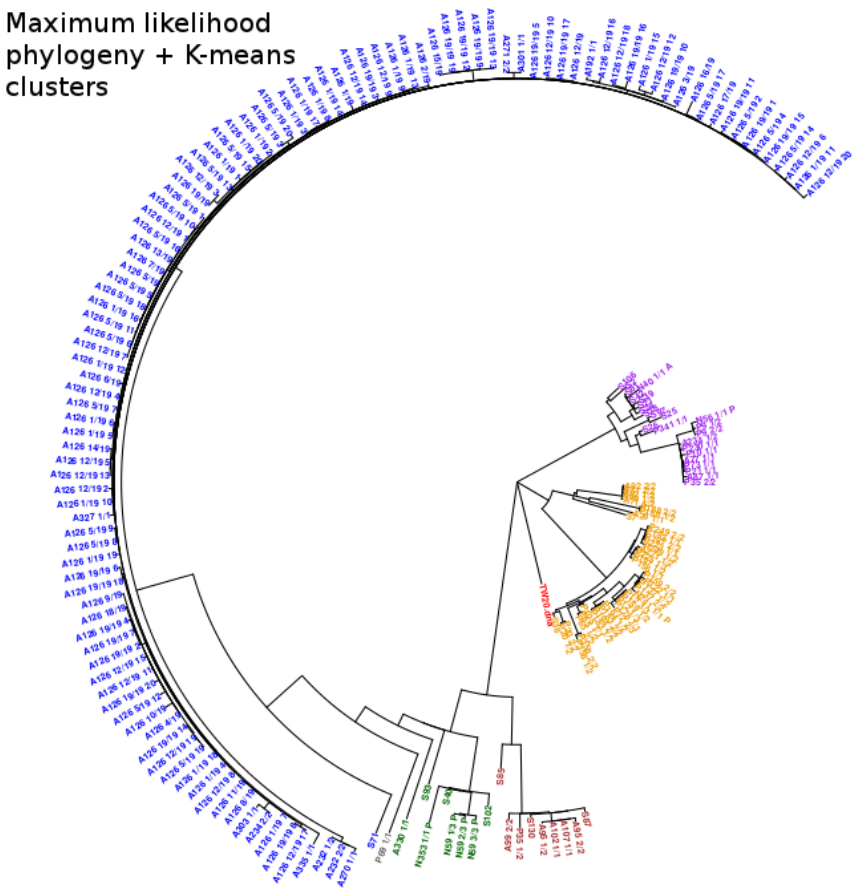
- 2 clusters, 50 isolates each
- 1,000,000 non structured SNPs
- 1,000 structured SNPs
  (i.e. different frequencies between groups)



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.

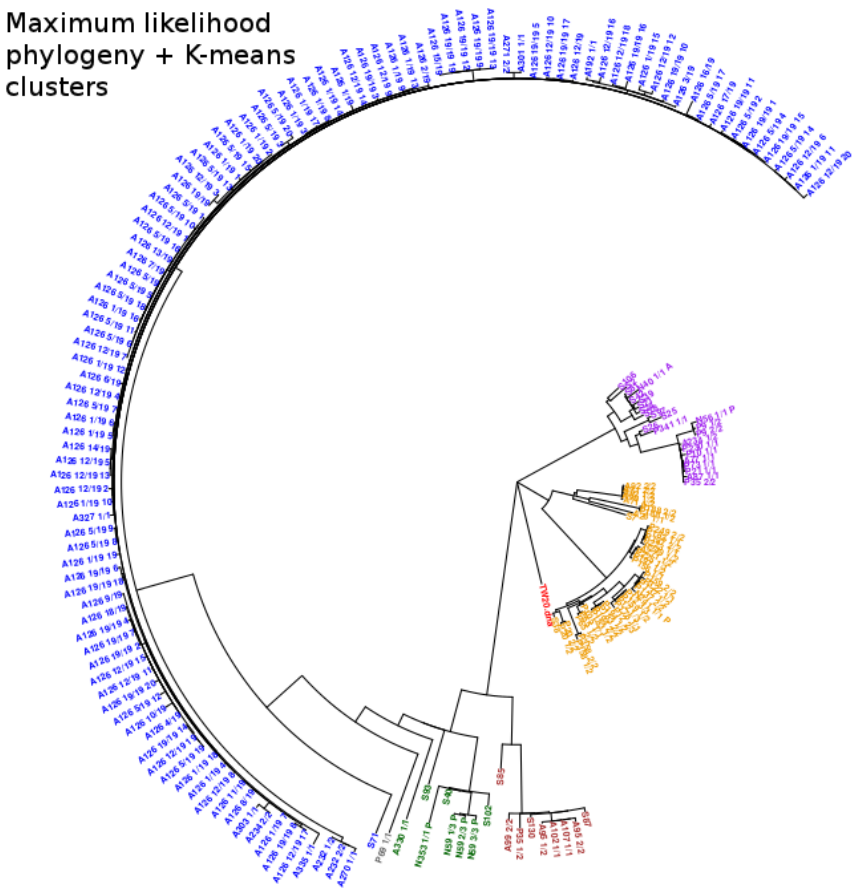


Observations:

- 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.

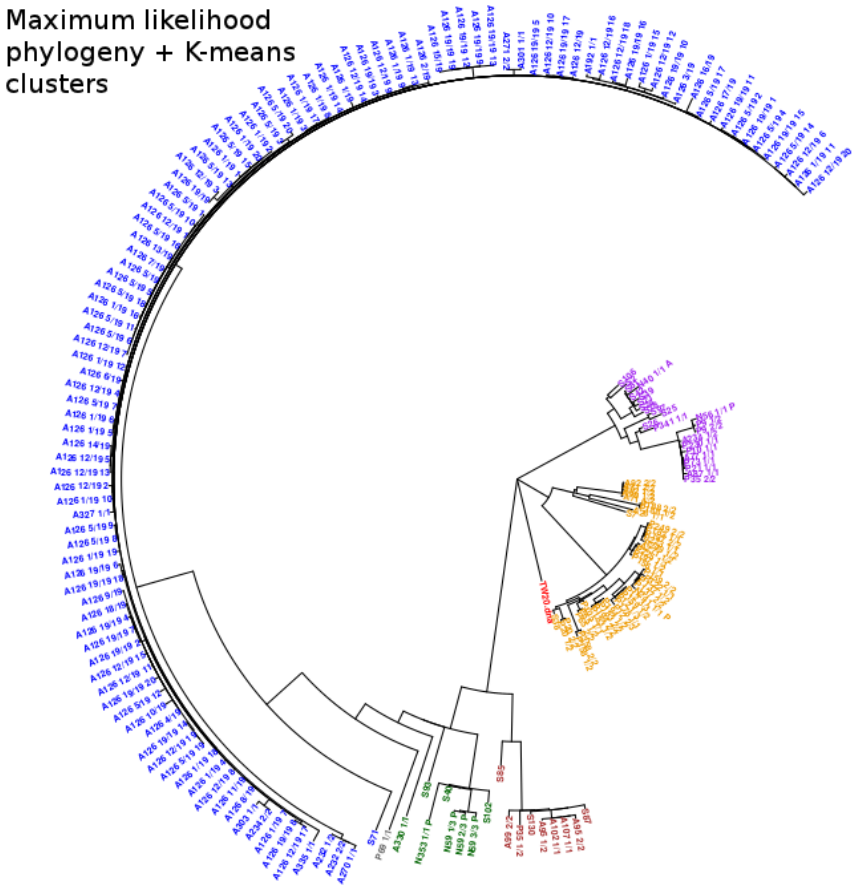


Observations:

- 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.

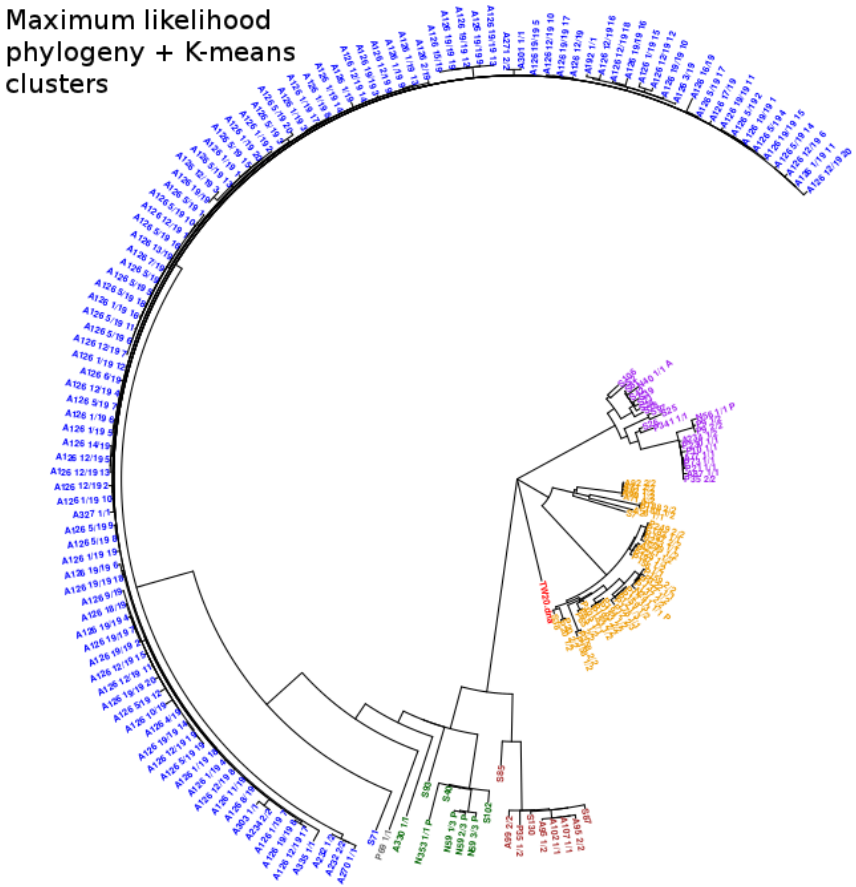


Observations:

- 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.

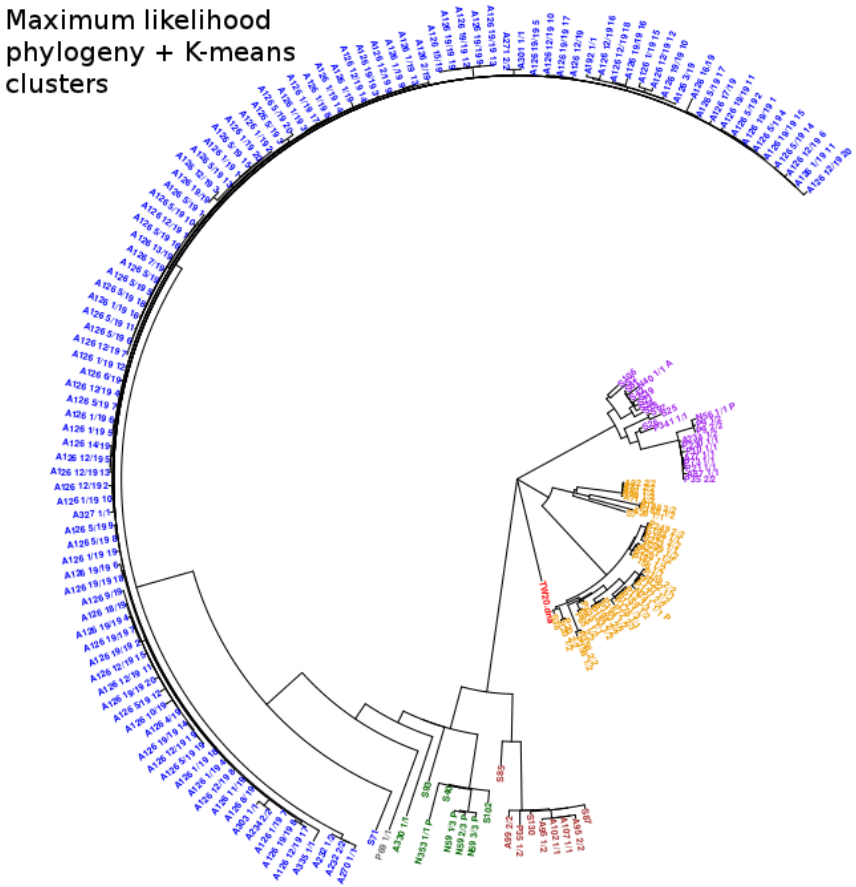


Observations:

- 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 applications to pathogen genetic data
- limits reached when reconstructing fine-scale processes
- more at: `http://adegenet.r-forge.r-project.org/`

Summary

- multivariate analysis used for ~ 50 years in genetics, still an active field for methodological development
- increasingly useful as datasets grow
- specific applications to pathogen genetic data
- limits reached when reconstructing fine-scale processes
- more at: `http://adegenet.r-forge.r-project.org/`

Summary

- multivariate analysis used for ~ 50 years in genetics, still an active field for methodological development
- increasingly useful as datasets grow
- specific applications to pathogen genetic data
- limits reached when reconstructing fine-scale processes
- more at: `http://adegenet.r-forge.r-project.org/`

Summary

- multivariate analysis used for ~ 50 years in genetics, still an active field for methodological development
- increasingly useful as datasets grow
- specific applications to pathogen genetic data
- limits reached when reconstructing fine-scale processes
- more at: `http://adegenet.r-forge.r-project.org/`

Summary

- multivariate analysis used for ~ 50 years in genetics, still an active field for methodological development
- increasingly useful as datasets grow
- specific applications to pathogen genetic data
- limits reached when reconstructing fine-scale processes
- more at: `http://adegenet.r-forge.r-project.org/`