

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

Multivariate analysis in a nutshell

Applications to genetic data

Genetic diversity of pathogen populations

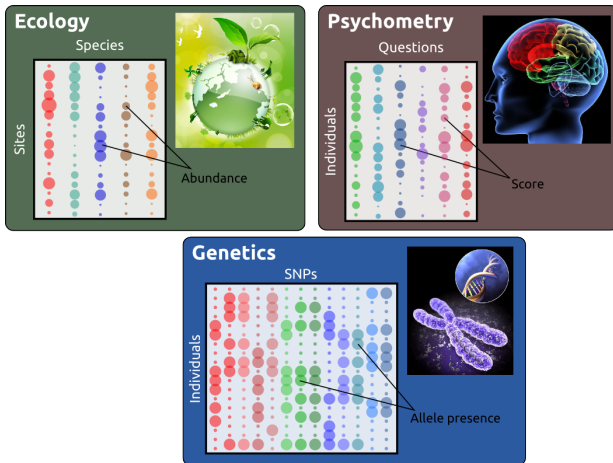
Outline

Multivariate analysis in a nutshell

Applications to genetic data

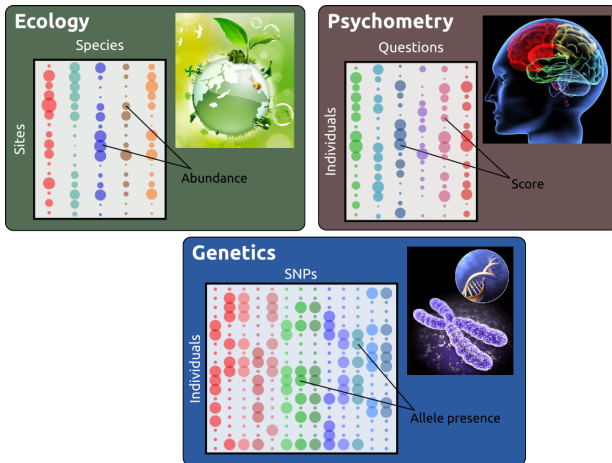
Genetic diversity of pathogen populations

Multivariate data: some examples



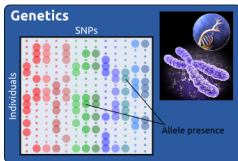
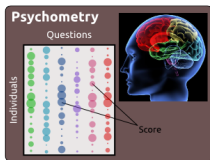
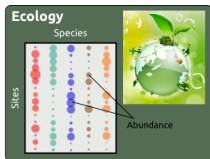
Association between individuals? Correlations between variables?

Multivariate data: some examples

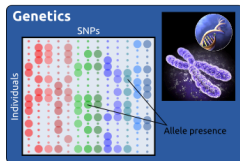
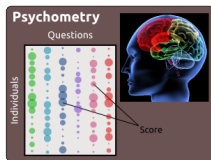
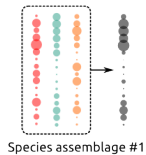
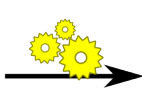
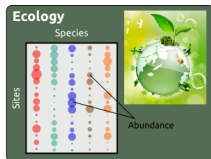


Association between individuals? Correlations between variables?

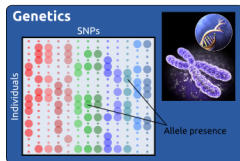
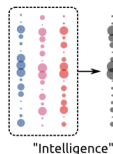
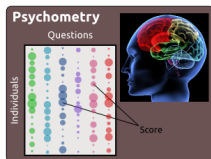
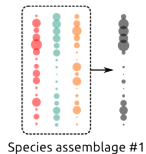
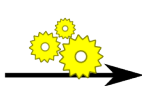
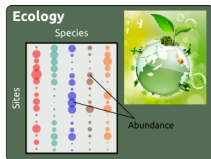
Multivariate analysis to summarize diversity



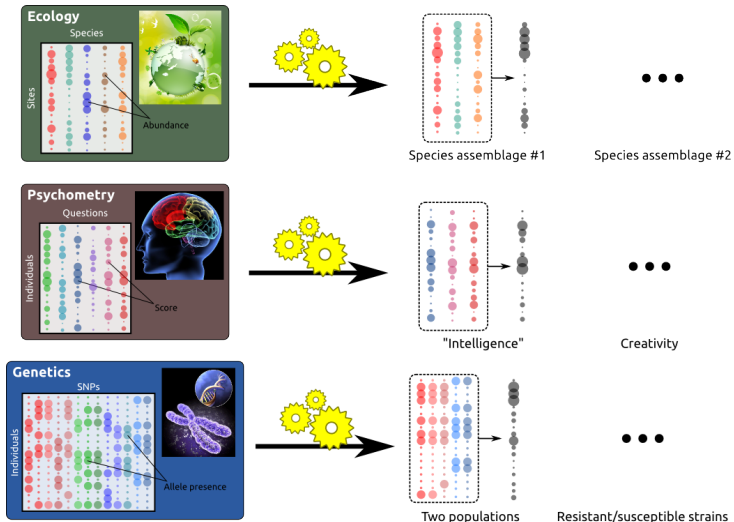
Multivariate analysis to summarize diversity



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.

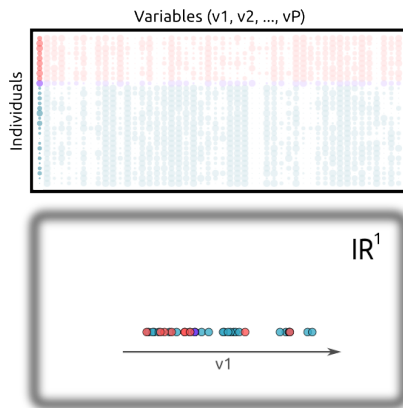
Most common methods

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)

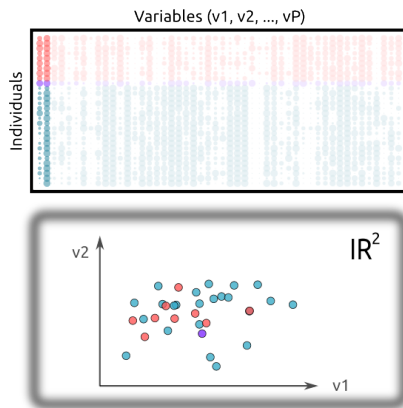
Many other methods for ≥ 2 data tables, spatial analysis, phylogenetic analysis, etc.

1 dimension, 2 dimensions, P dimensions



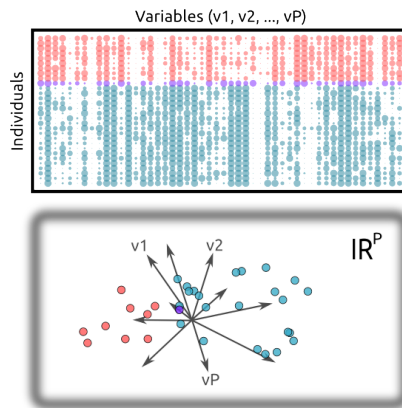
Need to find most informative directions in a P -dimensional space.

1 dimension, 2 dimensions, P dimensions



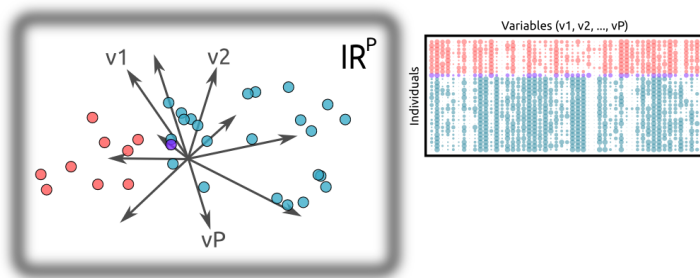
Need to find most informative directions in a P -dimensional space.

1 dimension, 2 dimensions, P dimensions



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

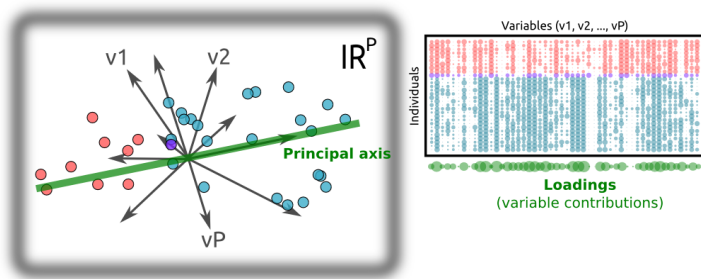
Reducing P dimensions into 1



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

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

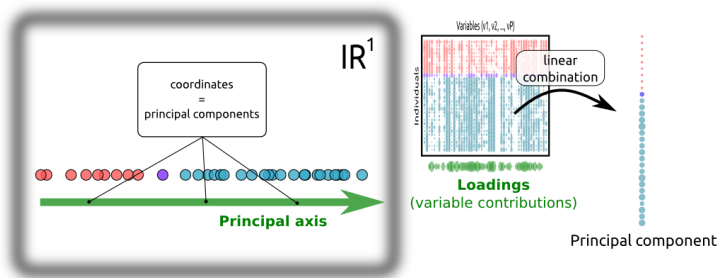
Reducing P dimensions into 1



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

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

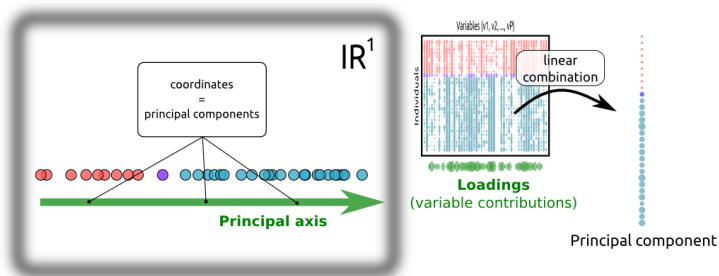
Reducing P dimensions into 1



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

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

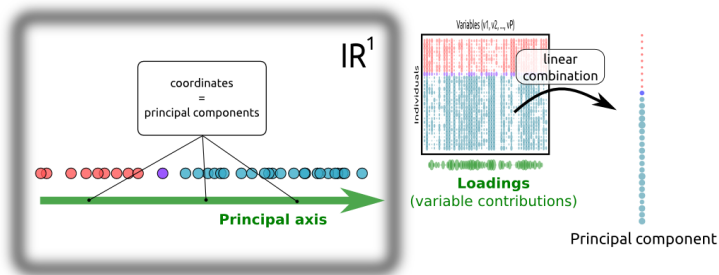
Reducing P dimensions into 1



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

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

Keeping more than one principal component

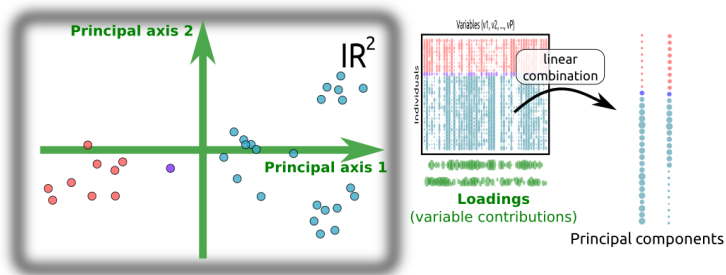


- \mathbf{u}_1 and \mathbf{v}_1 : **1st principal axis and component**
- \mathbf{u}_2 and \mathbf{v}_2 : **2nd principal axis and component**

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

→ find \mathbf{u}_2 so that $\|\mathbf{v}_2\|_{\mathbf{D}}^2$ is maximum

Keeping more than one principal component

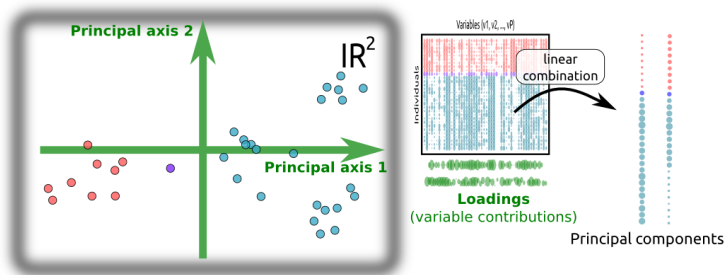


- \mathbf{u}_1 and \mathbf{v}_1 : **1st principal axis and component**
- \mathbf{u}_2 and \mathbf{v}_2 : **2nd principal axis and component**

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

→ find \mathbf{u}_2 so that $\|\mathbf{v}_2\|_{\mathbf{D}}^2$ is maximum

Keeping more than one principal component

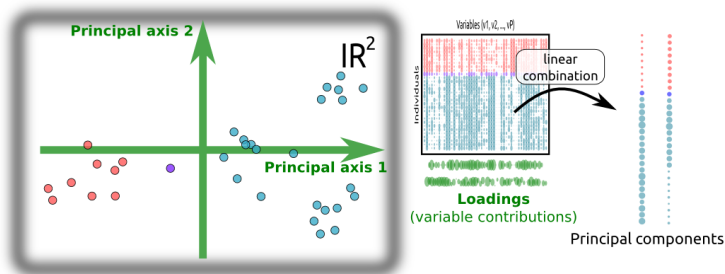


- \mathbf{u}_1 and \mathbf{v}_1 : **1st principal axis and component**
- \mathbf{u}_2 and \mathbf{v}_2 : **2nd principal axis and component**

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

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

→ find \mathbf{u}_2 so that $\|\mathbf{v}_2\|_{\mathbf{D}}^2$ 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} (implicitly 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} (implicitly 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:

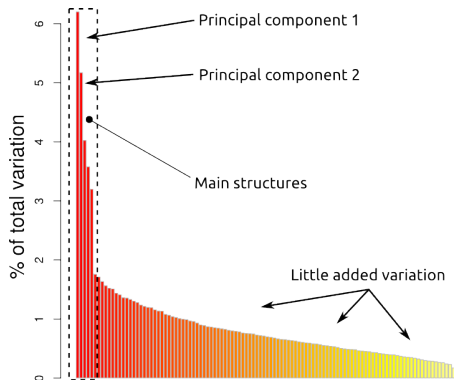
- pre-transformations of \mathbf{X} (recoding, standardisation, etc.)
- metrics \mathbf{Q} and \mathbf{D} (implicitly distances in \mathbb{R}^P and \mathbb{R}^N)
- most usual analyses are defined by $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$



packages: *ade4*, *vegan*

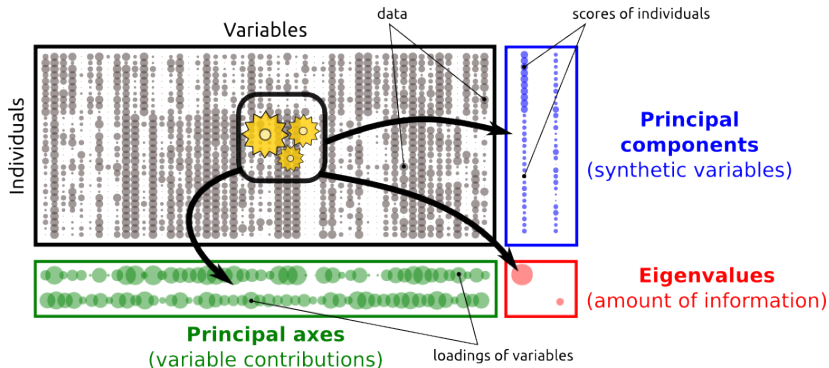
How many principal components to retain?

Choice based on “**screplot**”: barplot of eigenvalues



Retain only “significant” structures... but not trivial ones.

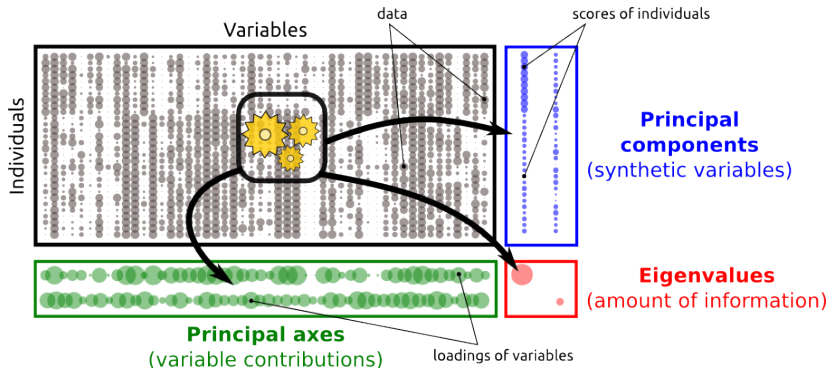
Outputs of multivariate analyses: an overview



Main outputs:

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

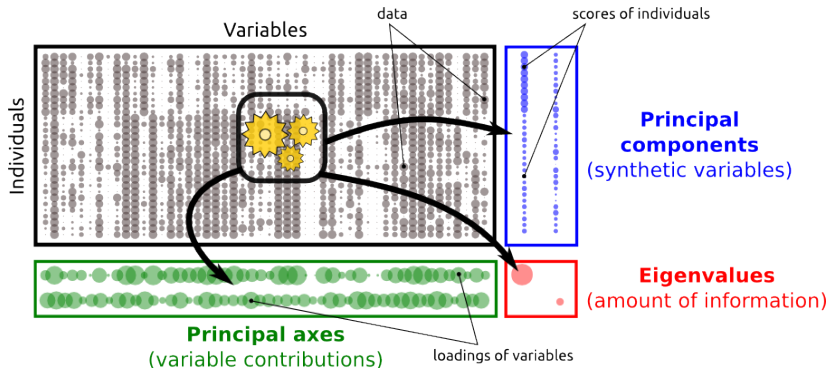
Outputs of multivariate analyses: an overview



Main outputs:

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

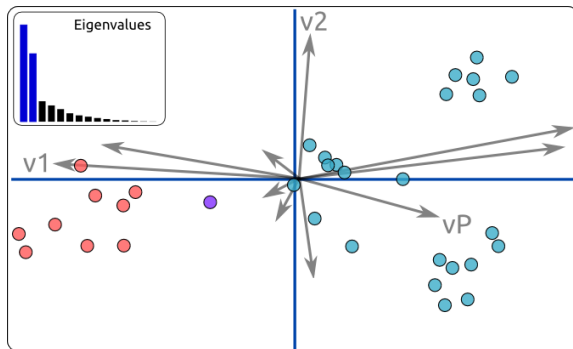
Outputs of multivariate analyses: an overview



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), ...

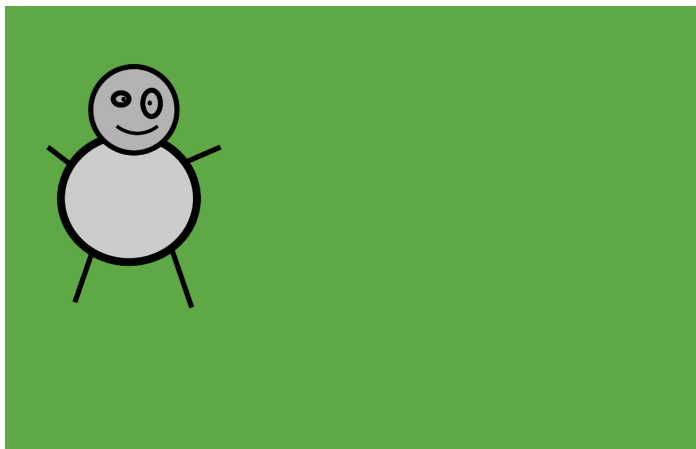
Outline

Multivariate analysis in a nutshell

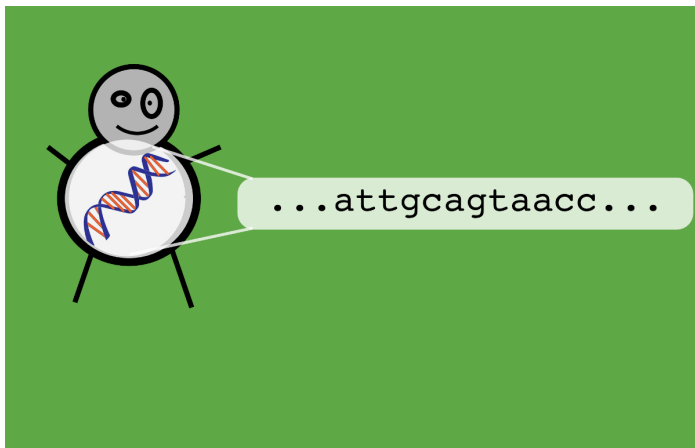
Applications to genetic data

Genetic diversity of pathogen populations

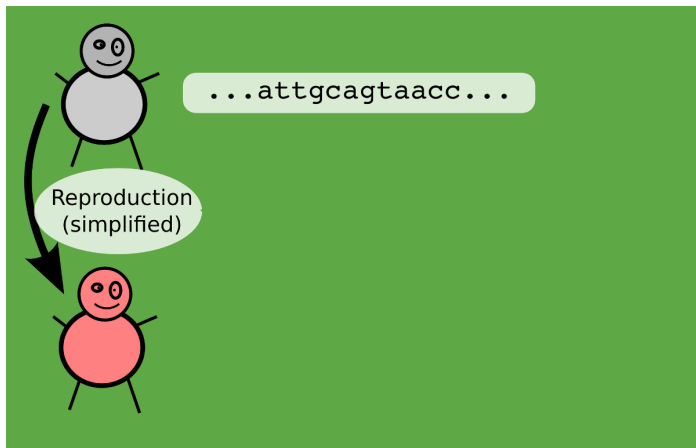
From DNA sequences to patterns of biological diversity



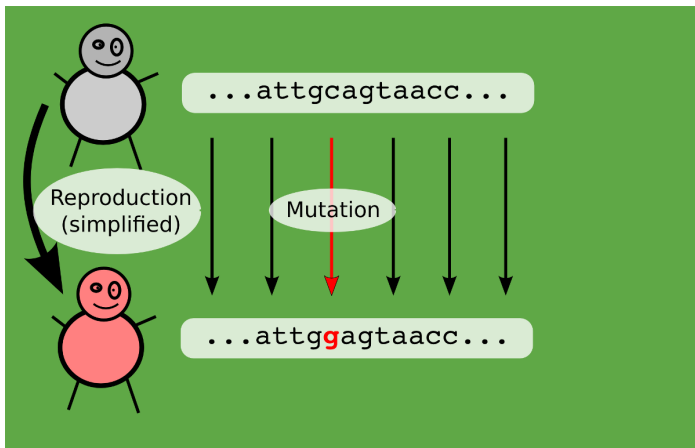
From DNA sequences to patterns of biological diversity



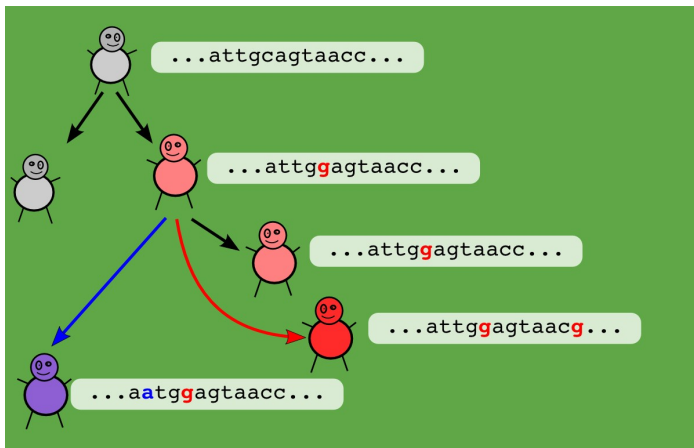
From DNA sequences to patterns of biological diversity



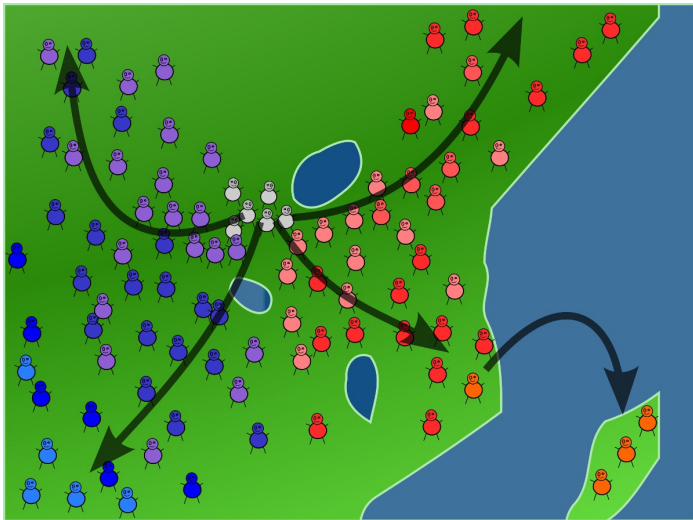
From DNA sequences to patterns of biological diversity



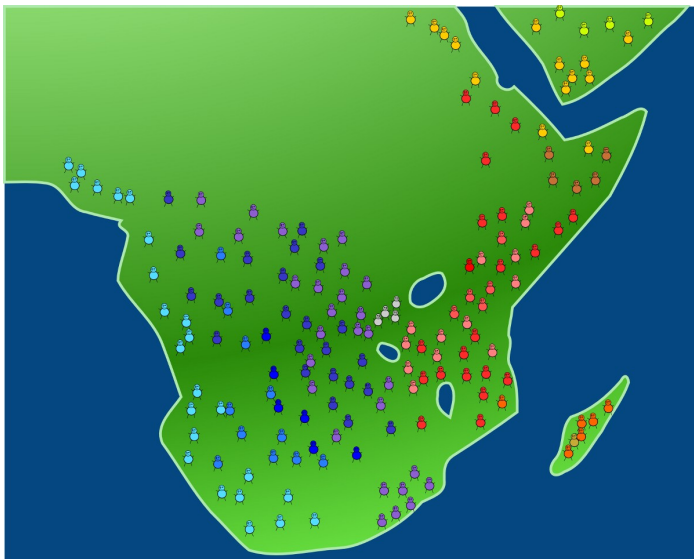
From DNA sequences to patterns of biological diversity



From DNA sequences to patterns of biological diversity



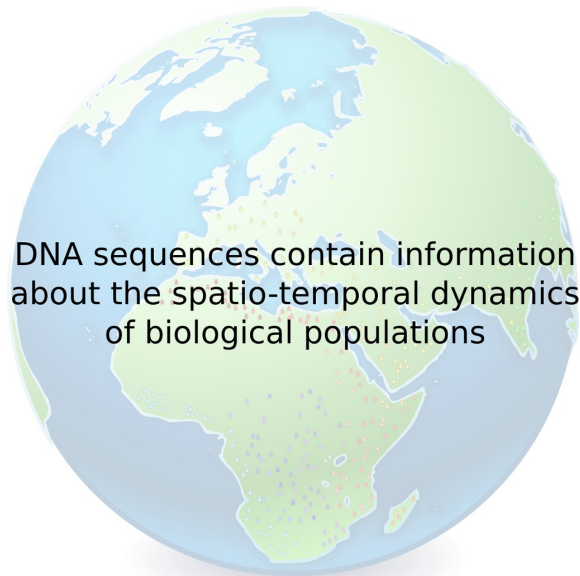
From DNA sequences to patterns of biological diversity



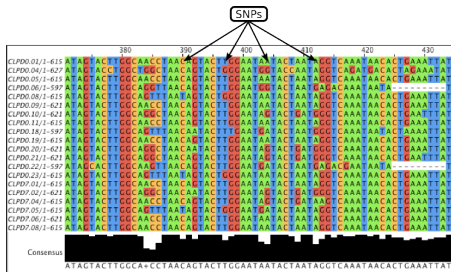
From DNA sequences to patterns of biological diversity



From DNA sequences to patterns of biological diversity



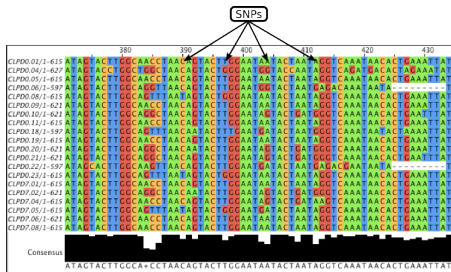
DNA sequences: a rich source of information



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

⇒ Multivariate analysis use to summarize genetic diversity.

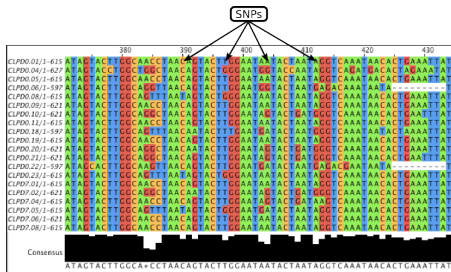
DNA sequences: a rich source of information



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

⇒ Multivariate analysis use to summarize genetic diversity.

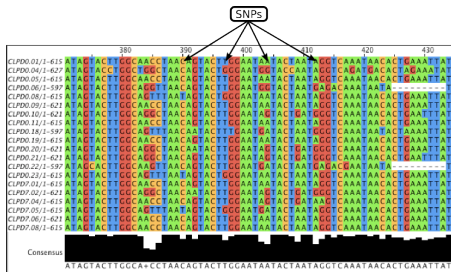
DNA sequences: a rich source of information



- hundreds/thousands individuals
- up to millions of **s**ingle **n**ucleotide **p**olymorphism (**SNPs**)
- more generally, most genetic data can be treated as **frequencies**

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

DNA sequences: a rich source of information

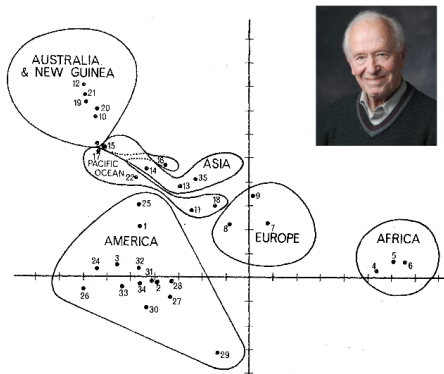


- hundreds/thousands individuals
- up to millions of **s**ingle **n**ucleotide **p**olymorphism (**SNPs**)
- more generally, most genetic data can be treated as **frequencies**

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

First application of multivariate analysis in genetics

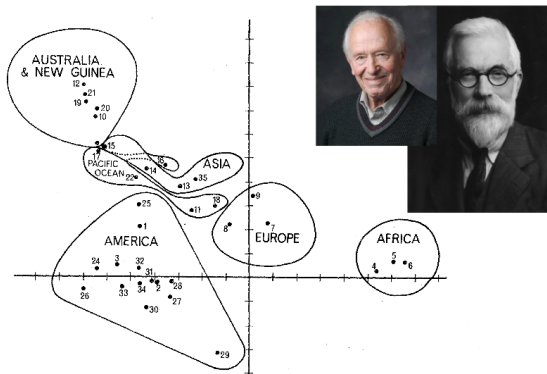
PCA of genetic data, native human populations (Cavalli-Sforza 1966, *Proc B*)



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*)

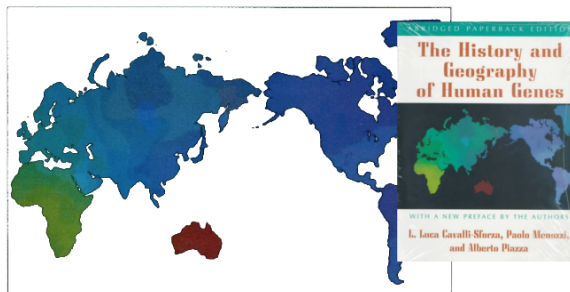


First 2 principal components separate populations into continents.

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: *adegetnet*, *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

Multivariate analysis in a nutshell

Applications to genetic 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)



Large scale

Small scale

Report

Plasmodium falciparum Accompanied the Human Expansion out of Africa

Kariyuki Tanabe,^{1,2} Toshihiro Mita,³ Thibaut Jeancoat,⁴ Andrew Ekinov,⁵ Sheng Huo,⁶ Rebecca Baxendale,⁷ Lisa Harland-Carrington,⁸ Arwen Davis,⁹ Nicola Sironi,¹⁰ Michael Omer,¹¹ Raulo Kariyuki,¹² Francis Proye,¹³ Andrea Ekinov,¹⁴ Anna Flinn,¹⁵ Alina Kozlov,¹⁶ Toshihiro Mita,¹⁷ Andrew Rafter,¹⁸ Nicola Sironi,¹⁹ and Francis Baxendale²⁰

¹Department of Molecular Pathology Research Institute for Malaria Diseases, Osaka University, Suita 565-0871, Japan

²Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

³Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

⁴Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

⁵Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

⁶Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

⁷Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

⁸Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

⁹Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

¹⁰Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

¹¹Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

¹²Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

¹³Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

¹⁴Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

¹⁵Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

¹⁶Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

¹⁷Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

¹⁸Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

¹⁹Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

²⁰Department of Microbiology and Immunology, Osaka University, Suita 565-0871, Japan

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

B. S. Gager¹, G. F. Munday², S. P. Stone³, C. C. Kibbler⁴, J. A. Roberts⁵, G. Duckworth⁶, R. Lal⁷, and S. Eickhoff⁸

¹Department of Medical Microbiology and Immunology, University of London, 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 7HT, United Kingdom; ²Department of Biomedical Sciences, University of London, 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 7HT, United Kingdom; ³Department of Biomedical Sciences, University of London, 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 7HT, United Kingdom; ⁴Department of Biomedical Sciences, University of London, 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 7HT, United Kingdom; ⁵Department of Biomedical Sciences, University of London, 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 7HT, United Kingdom; ⁶Department of Biomedical Sciences, University of London, 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 7HT, United Kingdom; ⁷Department of Biomedical Sciences, University of London, 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 7HT, United Kingdom; ⁸Department of Biomedical Sciences, University of London, 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 7HT, United Kingdom

The Global Circulation of Seasonal Influenza A (H3N2) Viruses

Colin A. Russell,¹ Tony C. Jones,^{2,3,4,5} Ian G. Barr,⁶ Nancy J. Cox,⁷ Rebecca J. Gair,⁸ Vicky Gregory,⁹ Ian G. Barr,¹⁰ Alan H. Hay,¹¹ John C. De Jong,¹² Anne Kelso,¹³ Alexander C. Ewer,¹⁴ Isabella Kogej,¹⁵ Anne B. Smith,¹⁶ John C. De Jong,¹⁷ Anne Kelso,¹⁸ Alexander C. Ewer,¹⁹ Isabella Kogej,²⁰ Anne B. Smith,²¹ John C. De Jong,²² Anne Kelso,²³ Alexander C. Ewer,²⁴ Isabella Kogej,²⁵ Anne B. Smith,²⁶ John C. De Jong,²⁷ Anne Kelso,²⁸ Alexander C. Ewer,²⁹ Isabella Kogej,³⁰ Anne B. Smith,³¹ John C. De Jong,³² Anne Kelso,³³ Alexander C. Ewer,³⁴ Isabella Kogej,³⁵ Anne B. Smith,³⁶ John C. De Jong,³⁷ Anne Kelso,³⁸ Alexander C. Ewer,³⁹ Isabella Kogej,⁴⁰ Anne B. Smith,⁴¹ John C. De Jong,⁴² Anne Kelso,⁴³ Alexander C. Ewer,⁴⁴ Isabella Kogej,⁴⁵ Anne B. Smith,⁴⁶ John C. De Jong,⁴⁷ Anne Kelso,⁴⁸ Alexander C. Ewer,⁴⁹ Isabella Kogej,⁵⁰ Anne B. Smith,⁵¹ John C. De Jong,⁵² Anne Kelso,⁵³ Alexander C. Ewer,⁵⁴ Isabella Kogej,⁵⁵ Anne B. Smith,⁵⁶ John C. De Jong,⁵⁷ Anne Kelso,⁵⁸ Alexander C. Ewer,⁵⁹ Isabella Kogej,⁶⁰ Anne B. Smith,⁶¹ John C. De Jong,⁶² Anne Kelso,⁶³ Alexander C. Ewer,⁶⁴ Isabella Kogej,⁶⁵ Anne B. Smith,⁶⁶ John C. De Jong,⁶⁷ Anne Kelso,⁶⁸ Alexander C. Ewer,⁶⁹ Isabella Kogej,⁷⁰ Anne B. Smith,⁷¹ John C. De Jong,⁷² Anne Kelso,⁷³ Alexander C. Ewer,⁷⁴ Isabella Kogej,⁷⁵ Anne B. Smith,⁷⁶ John C. De Jong,⁷⁷ Anne Kelso,⁷⁸ Alexander C. Ewer,⁷⁹ Isabella Kogej,⁸⁰ Anne B. Smith,⁸¹ John C. De Jong,⁸² Anne Kelso,⁸³ Alexander C. Ewer,⁸⁴ Isabella Kogej,⁸⁵ Anne B. Smith,⁸⁶ John C. De Jong,⁸⁷ Anne Kelso,⁸⁸ Alexander C. Ewer,⁸⁹ Isabella Kogej,⁹⁰ Anne B. Smith,⁹¹ John C. De Jong,⁹² Anne Kelso,⁹³ Alexander C. Ewer,⁹⁴ Isabella Kogej,⁹⁵ Anne B. Smith,⁹⁶ John C. De Jong,⁹⁷ Anne Kelso,⁹⁸ Alexander C. Ewer,⁹⁹ Isabella Kogej,¹⁰⁰ Anne B. Smith,¹⁰¹ John C. De Jong,¹⁰² Anne Kelso,¹⁰³ Alexander C. Ewer,¹⁰⁴ Isabella Kogej,¹⁰⁵ Anne B. Smith,¹⁰⁶ John C. De Jong,¹⁰⁷ Anne Kelso,¹⁰⁸ Alexander C. Ewer,¹⁰⁹ Isabella Kogej,¹¹⁰ Anne B. Smith,¹¹¹ John C. De Jong,¹¹² Anne Kelso,¹¹³ Alexander C. Ewer,¹¹⁴ Isabella Kogej,¹¹⁵ Anne B. Smith,¹¹⁶ John C. De Jong,¹¹⁷ Anne Kelso,¹¹⁸ Alexander C. Ewer,¹¹⁹ Isabella Kogej,¹²⁰ Anne B. Smith,¹²¹ John C. De Jong,¹²² Anne Kelso,¹²³ Alexander C. Ewer,¹²⁴ Isabella Kogej,¹²⁵ Anne B. Smith,¹²⁶ John C. De Jong,¹²⁷ Anne Kelso,¹²⁸ Alexander C. Ewer,¹²⁹ Isabella Kogej,¹³⁰ Anne B. Smith,¹³¹ John C. De Jong,¹³² Anne Kelso,¹³³ Alexander C. Ewer,¹³⁴ Isabella Kogej,¹³⁵ Anne B. Smith,¹³⁶ John C. De Jong,¹³⁷ Anne Kelso,¹³⁸ Alexander C. Ewer,¹³⁹ Isabella Kogej,¹⁴⁰ Anne B. Smith,¹⁴¹ John C. De Jong,¹⁴² Anne Kelso,¹⁴³ Alexander C. Ewer,¹⁴⁴ Isabella Kogej,¹⁴⁵ Anne B. Smith,¹⁴⁶ John C. De Jong,¹⁴⁷ Anne Kelso,¹⁴⁸ Alexander C. Ewer,¹⁴⁹ Isabella Kogej,¹⁵⁰ Anne B. Smith,¹⁵¹ John C. De Jong,¹⁵² Anne Kelso,¹⁵³ Alexander C. Ewer,¹⁵⁴ Isabella Kogej,¹⁵⁵ Anne B. Smith,¹⁵⁶ John C. De Jong,¹⁵⁷ Anne Kelso,¹⁵⁸ Alexander C. Ewer,¹⁵⁹ Isabella Kogej,¹⁶⁰ Anne B. Smith,¹⁶¹ John C. De Jong,¹⁶² Anne Kelso,¹⁶³ Alexander C. Ewer,¹⁶⁴ Isabella Kogej,¹⁶⁵ Anne B. Smith,¹⁶⁶ John C. De Jong,¹⁶⁷ Anne Kelso,¹⁶⁸ Alexander C. Ewer,¹⁶⁹ Isabella Kogej,¹⁷⁰ Anne B. Smith,¹⁷¹ John C. De Jong,¹⁷² Anne Kelso,¹⁷³ Alexander C. Ewer,¹⁷⁴ Isabella Kogej,¹⁷⁵ Anne B. Smith,¹⁷⁶ John C. De Jong,¹⁷⁷ Anne Kelso,¹⁷⁸ Alexander C. Ewer,¹⁷⁹ Isabella Kogej,¹⁸⁰ Anne B. Smith,¹⁸¹ John C. De Jong,¹⁸² Anne Kelso,¹⁸³ Alexander C. Ewer,¹⁸⁴ Isabella Kogej,¹⁸⁵

25/34

25/34

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).

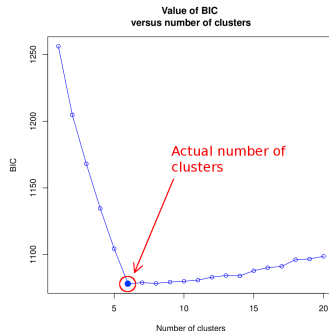
⇒ aim: **identify** and **describe** genetic clusters

Genetic clustering using K-means & BIC

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

Variance partitioning model (ANOVA):

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



Performances:

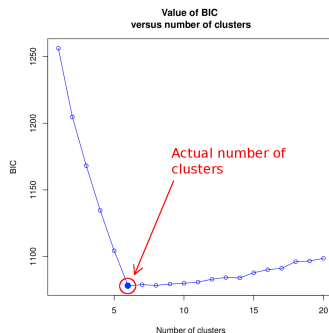
- K-means \geq STRUCTURE on simulated data (various island and stepping stone models)
- orders of magnitude faster (seconds vs hours/days)

Genetic clustering using K-means & BIC

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

Variance partitioning model (ANOVA):

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



Performances:

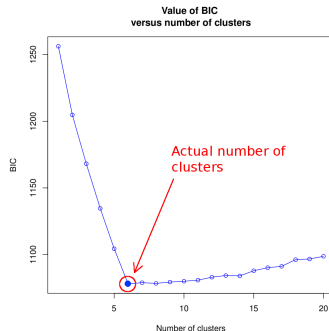
- K-means \geq STRUCTURE on simulated data (various island and stepping stone models)
- orders of magnitude faster (seconds vs hours/days)

Genetic clustering using K-means & BIC

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

Variance partitioning model (ANOVA):

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



Performances:

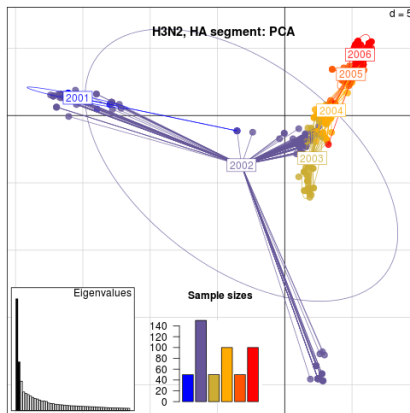
- K-means \geq STRUCTURE on simulated data (various island and stepping stone models)
- orders of magnitude faster (seconds vs hours/days)



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

PCA of seasonal influenza (A/H3N2) data

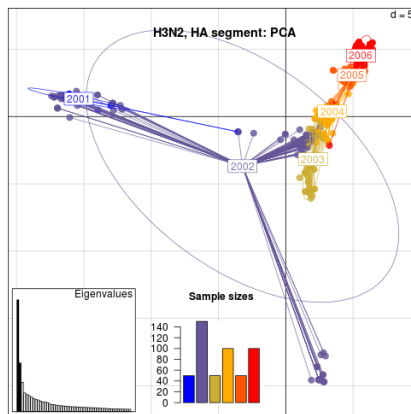
Data: seasonal influenza (A/H3N2), 500 HA segments.



Little temporal evolution, burst of diversity in 2002??

PCA of seasonal influenza (A/H3N2) data

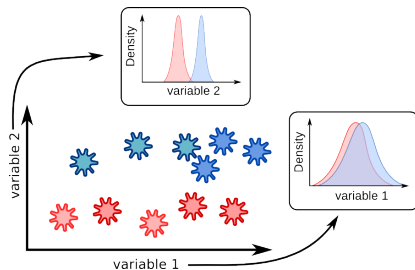
Data: seasonal influenza (A/H3N2), 500 HA segments.



Little temporal evolution, burst of diversity in 2002??

Which diversity to represent?

Total diversity not relevant to analyse clusters.



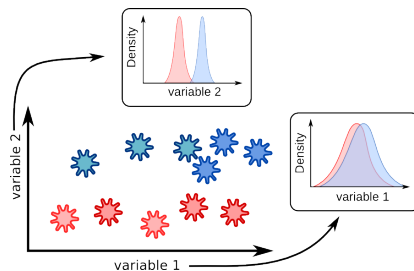
Discriminant Analysis of Principal Components (DAPC):

(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.



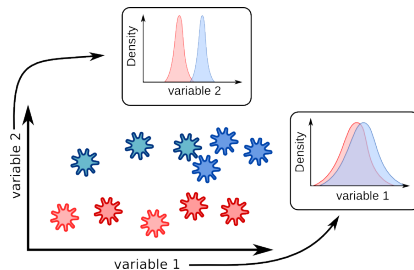
Discriminant Analysis of Principal Components (DAPC):

(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):

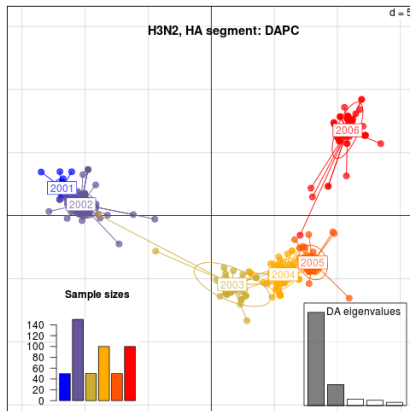
(Jombart *et al.* 2010, *BMC Genetics*)

- maximizes group discrimination (“*between/within*” ratio)
- provides group membership probabilities (prediction possible)
- as computer-efficient as PCA



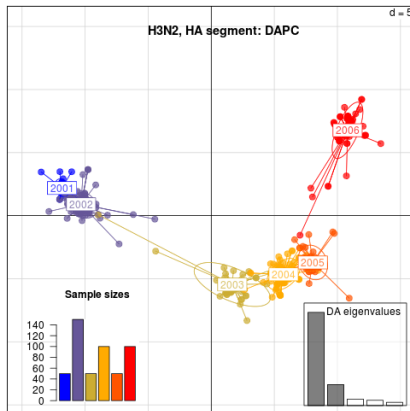
package: *adegenet*, function `dapc`

DAPC of seasonal influenza (A/H3N2) data



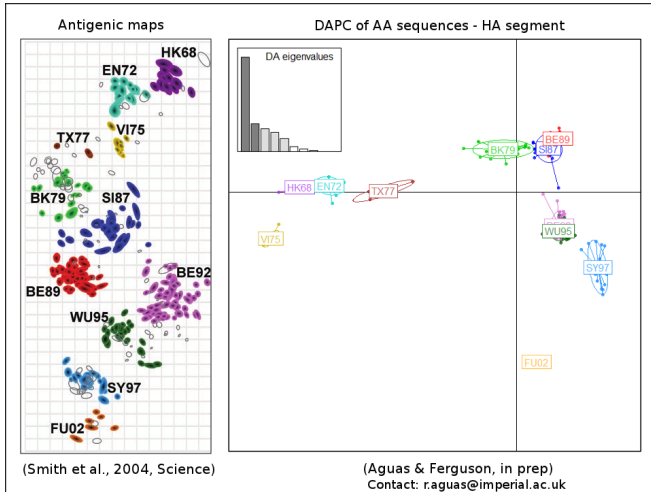
Strong temporal signal, originality of 2006 isolates (new alleles).

DAPC of seasonal influenza (A/H3N2) data



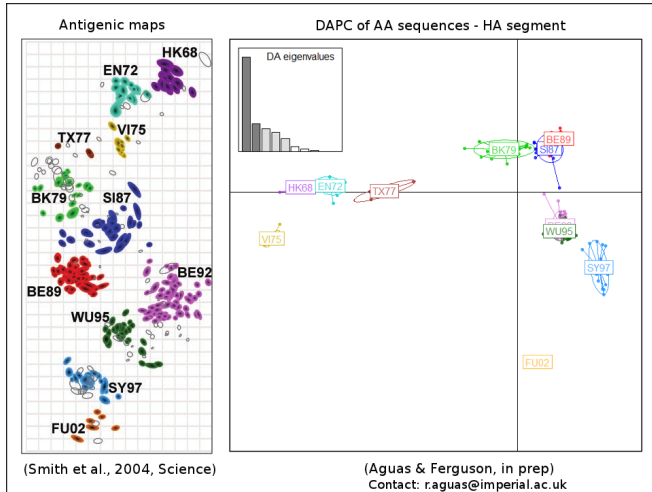
Strong temporal signal, originality of 2006 isolates (new alleles).

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.

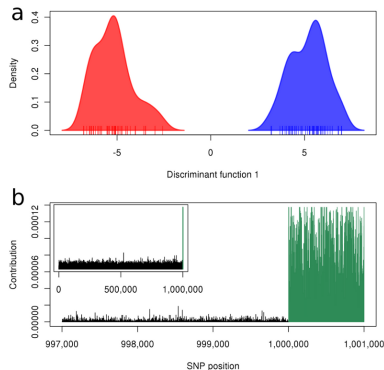
DAPC to identify structuring alleles

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).

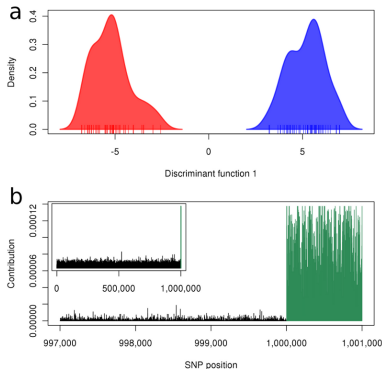
DAPC to identify structuring alleles

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)

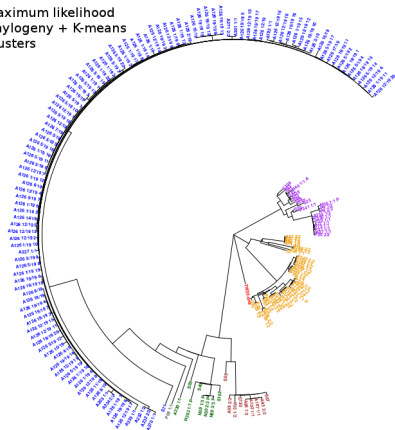


Possible applications to pathogen GWAS (e.g. SNPs related to antibiotic resistance in bacteria).

Limits of multivariate analysis

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

Maximum likelihood
phylogeny + K-means
clusters



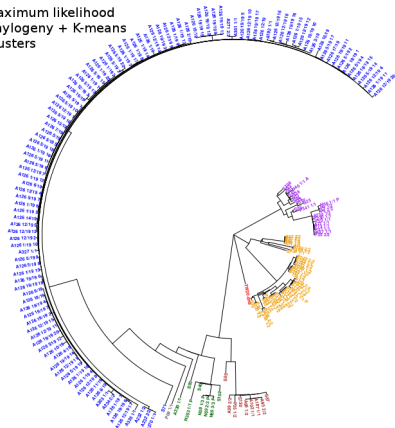
Observations:

- 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. ~ 1,000 SNPs.

Maximum likelihood
phylogeny + K-means
clusters



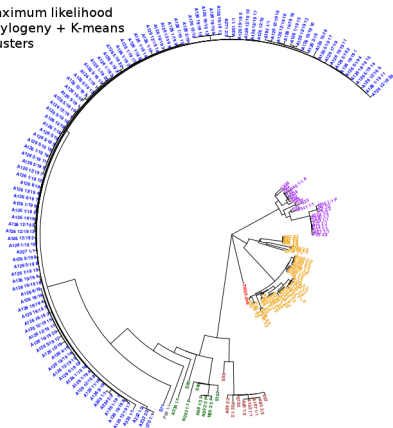
Observations:

- 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. ~ 1,000 SNPs.

Maximum likelihood
phylogeny + K-means
clusters



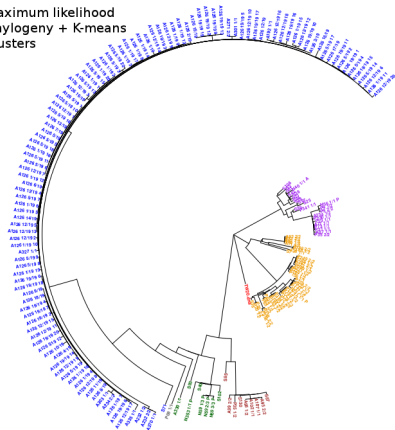
Observations:

- 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. ~ 1,000 SNPs.

Maximum likelihood
phylogeny + K-means
clusters



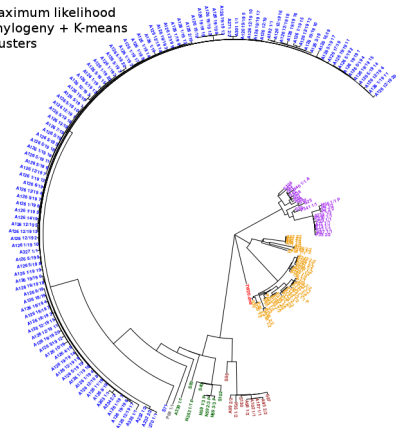
Observations:

- 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. ~ 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

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/>