Multivariate analysis of genetic data
— an introduction —

Thibaut Jombart

MRC Centre for Outbreak Analysis and Modelling
Imperial College London

Population genomics in R
Lausanne
23 Aug 2016

Outline

Multivariate analysis in a nutshell

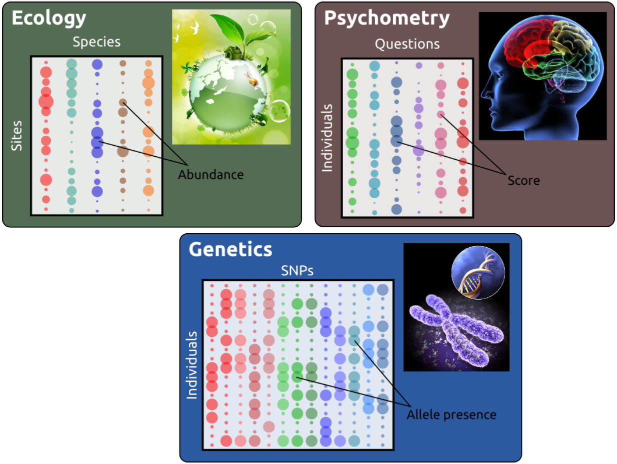
Applications to genetic data

Outline

Multivariate analysis in a nutshell

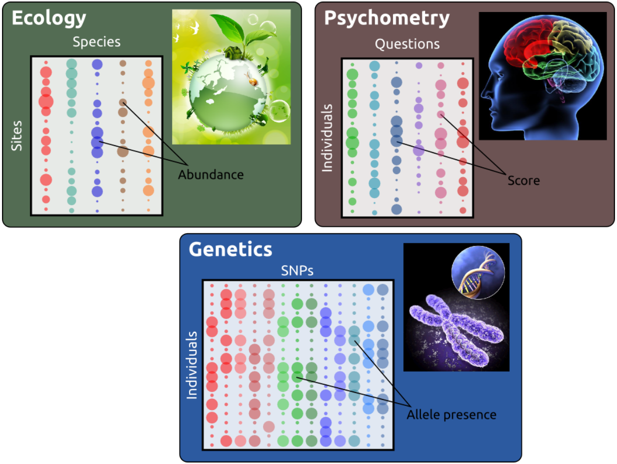
Applications to genetic data

Multivariate data: some examples



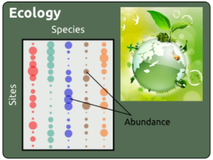
Association between individuals? Correlations between variables?

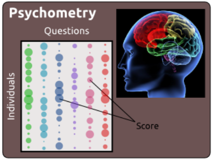
Multivariate data: some examples

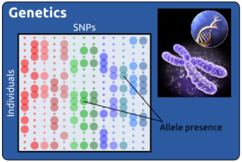


Association between individuals? Correlations between variables?

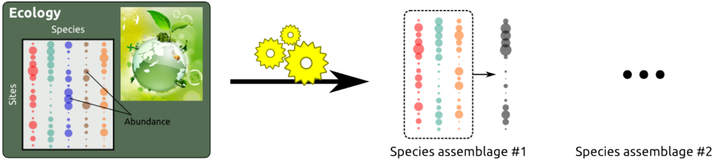
Multivariate analysis to summarize diversity

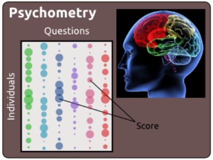


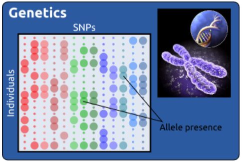




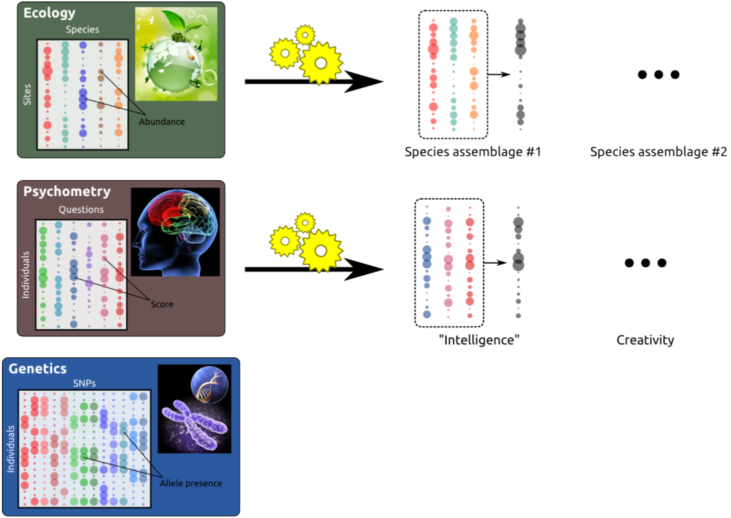
Multivariate analysis to summarize diversity



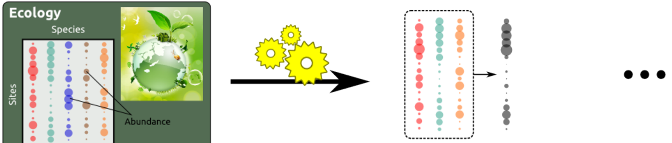




Multivariate analysis to summarize diversity

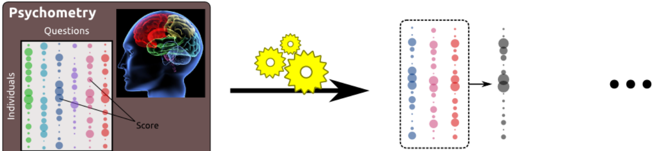


Multivariate analysis to summarize diversity



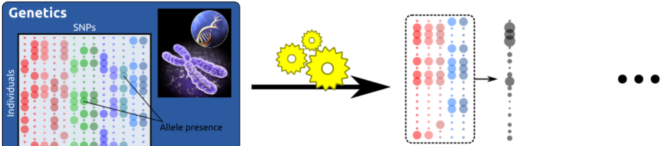
Species assemblage #1

Species assemblage #2



"Intelligence"

Creativity



Two populations

Resistant/susceptible strains

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.

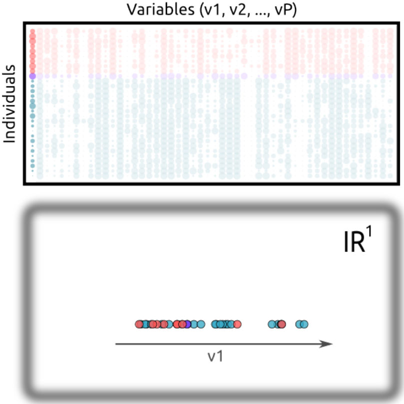
Most common methods

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)

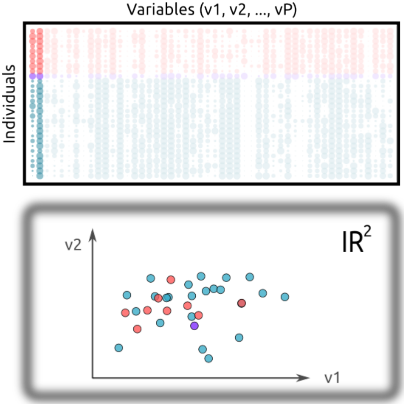
Many other methods for ≥ 2 data tables, spatial analysis, phylogenetic analysis, etc.

1 dimension, 2 dimensions, P dimensions



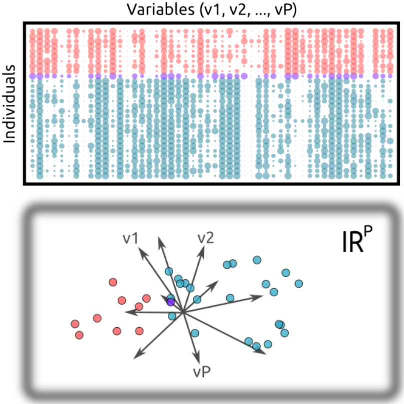
Need to find most informative directions in a P-dimensional space.

1 dimension, 2 dimensions, P dimensions



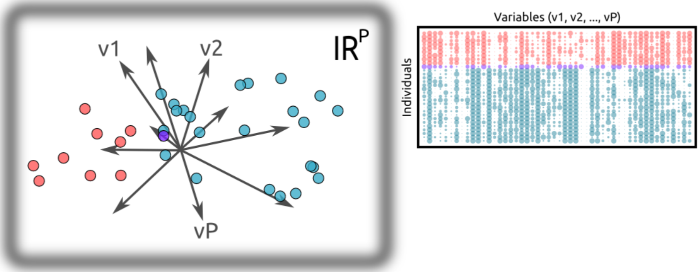
Need to find most informative directions in a P-dimensional space.

1 dimension, 2 dimensions, P dimensions



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

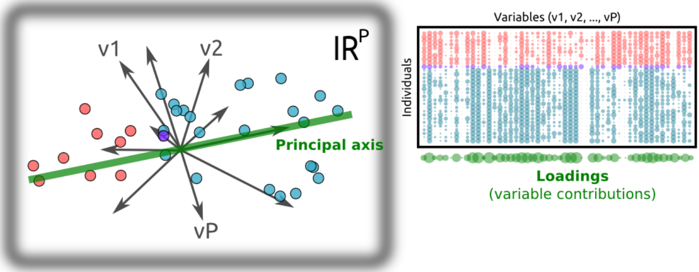
Reducing P dimensions into 1



- $\mathbf{X} \in \mathbb{R}^{N \times P}$; $\mathbf{X} = [\mathbf{x}_1 | \dots | \mathbf{x}_P]$: 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.

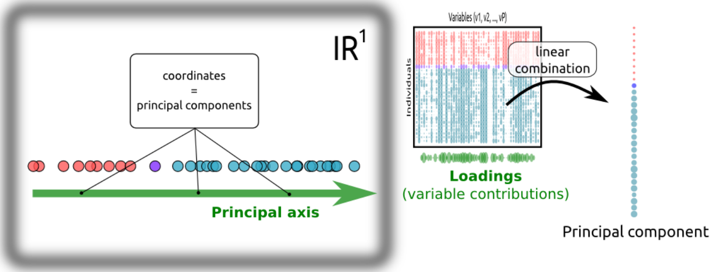
Reducing P dimensions into 1



- $\mathbf{X} \in \mathbb{R}^{N \times P}$; $\mathbf{X} = [\mathbf{x}_1 | \dots | \mathbf{x}_P]$: 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.

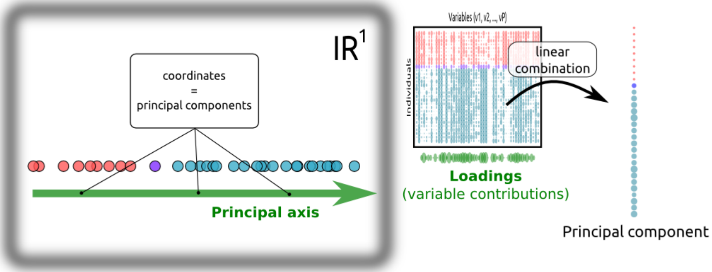
Reducing P dimensions into 1



- $\mathbf{X} \in \mathbb{R}^{N \times P}$; $\mathbf{X} = [\mathbf{x}_1 | \ldots | \mathbf{x}_P]$: data matrix
- $\mathbf{u} \in \mathbb{R}^P$; $\mathbf{u} = [u_1, \ldots, u_P]$: **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.

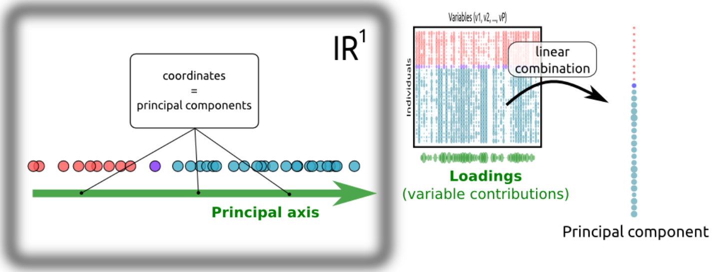
Reducing P dimensions into 1



- $\mathbf{X} \in \mathbb{R}^{N \times P}$; $\mathbf{X} = [\mathbf{x}_1 | \ldots | \mathbf{x}_P]$: data matrix
- $\mathbf{u} \in \mathbb{R}^{P}$; $\mathbf{u} = [u_1, \ldots, u_P]$: **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.

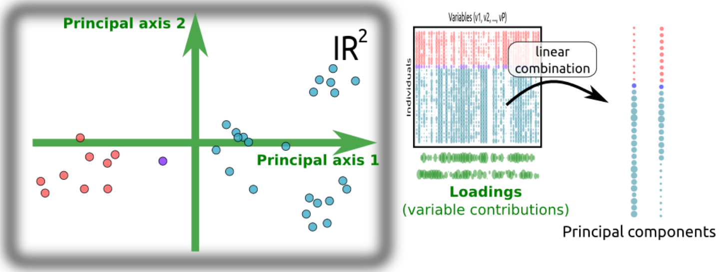
Keeping more than one principal component



- $\mathbf{u}_1$ 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

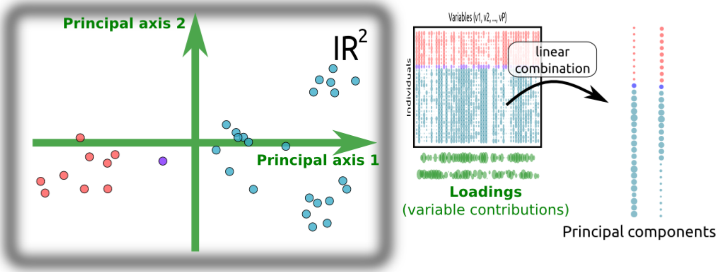
Keeping more than one principal component



- $\mathbf{u}_1$ 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

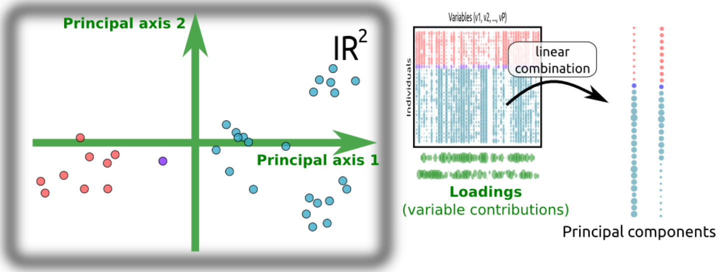


- $\mathbf{u}_1$ and $\mathbf{v}_1$: **1st principal axis and component**
- $\mathbf{u}_2$ and $\mathbf{v}_2$: **2nd principal axis and component**

→ 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

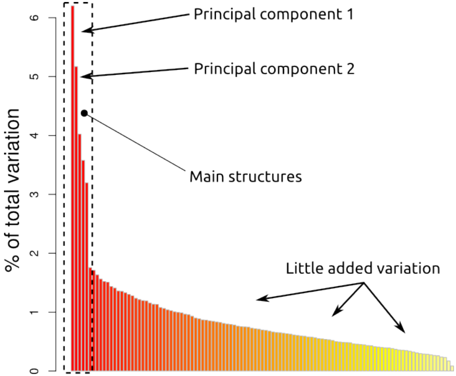


- $\mathbf{u}_1$ and $\mathbf{v}_1$: **1st principal axis and component**
- $\mathbf{u}_2$ and $\mathbf{v}_2$: **2nd principal axis and component**

→ 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

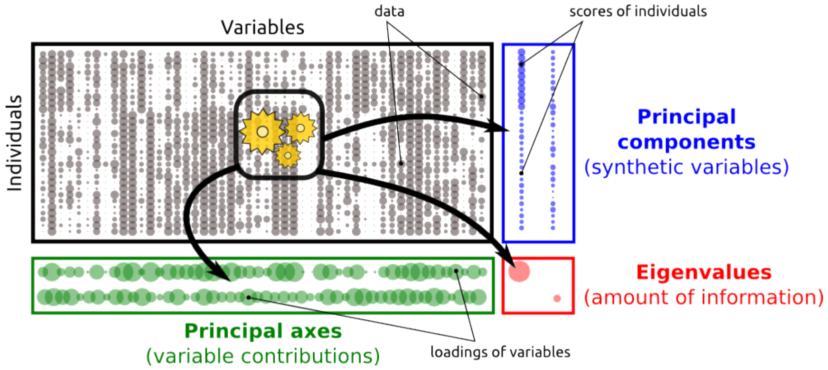
How many principal components to retain?

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



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

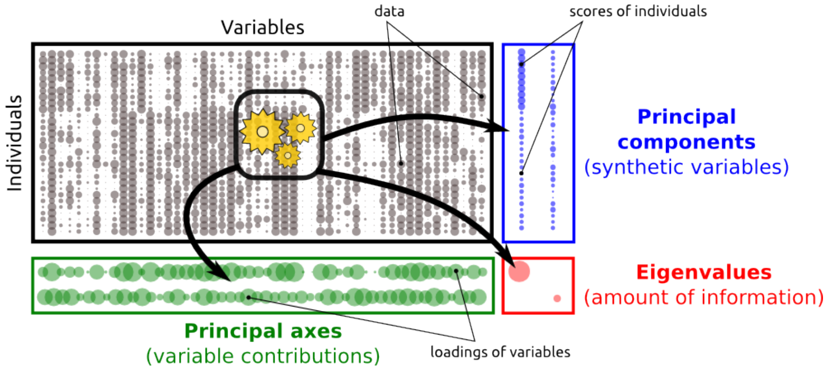
Outputs of multivariate analyses: an overview



Main outputs:

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

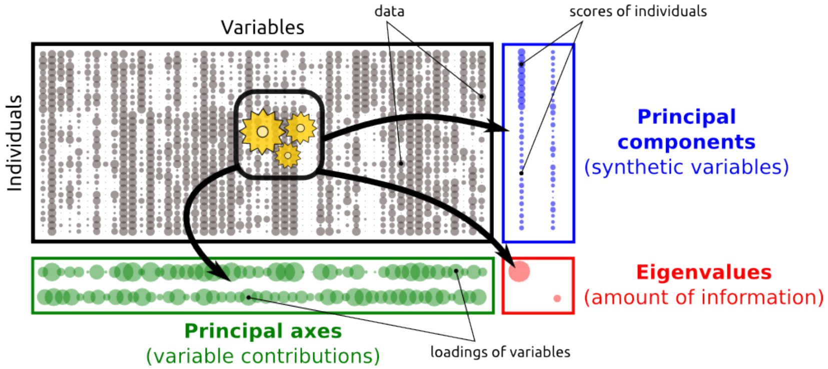
Outputs of multivariate analyses: an overview



Main outputs:

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

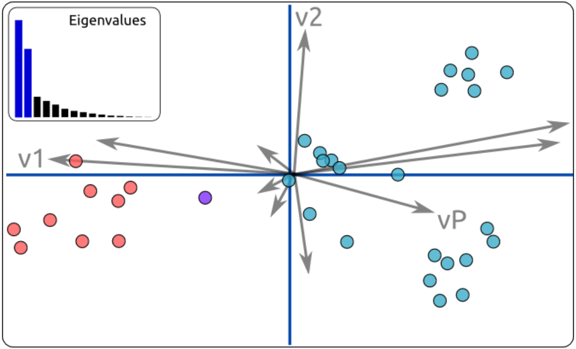
Outputs of multivariate analyses: an overview



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

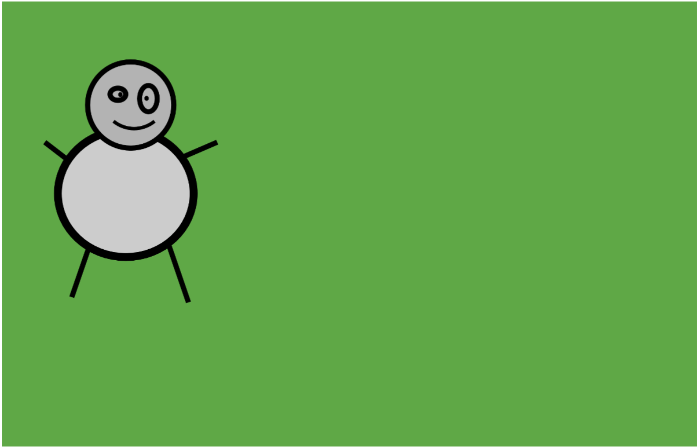
- **variety of methods** for different types of variables
- **principal components** (PCs) summarize diversity
- **variable loadings** identify discriminating variables
- other uses of PCs: **maps** (spatial structures), **models** (response variables or predictors), ...

Multivariate analysis in a nutshell

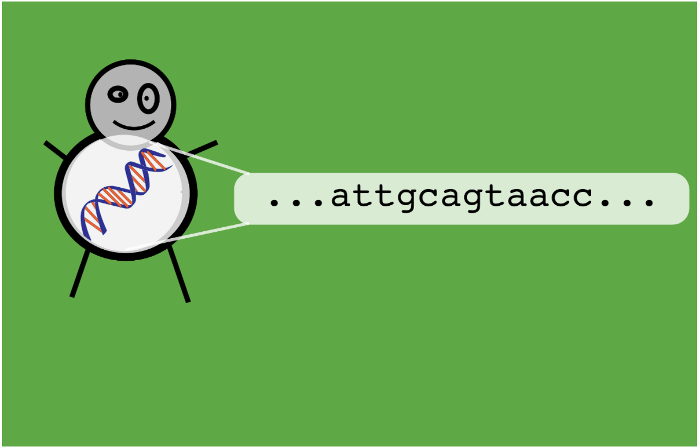
- **variety of methods** for different types of variables
- **principal components** (PCs) summarize diversity
- **variable loadings** identify discriminating variables
- other uses of PCs: **maps** (spatial structures), **models** (response variables or predictors), ...

Outline

From DNA sequences to patterns of biological diversity

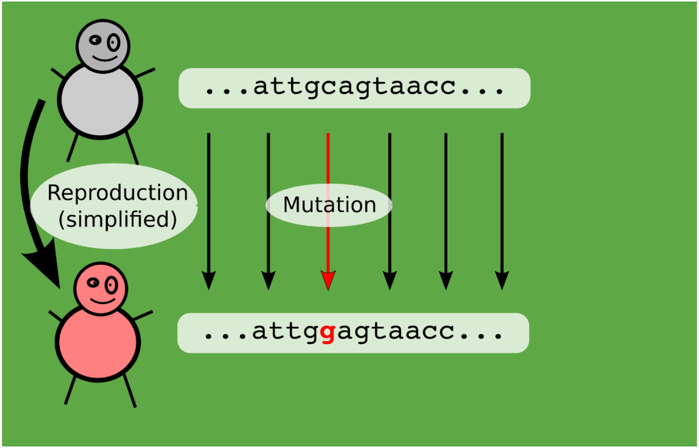


From DNA sequences to patterns of biological diversity

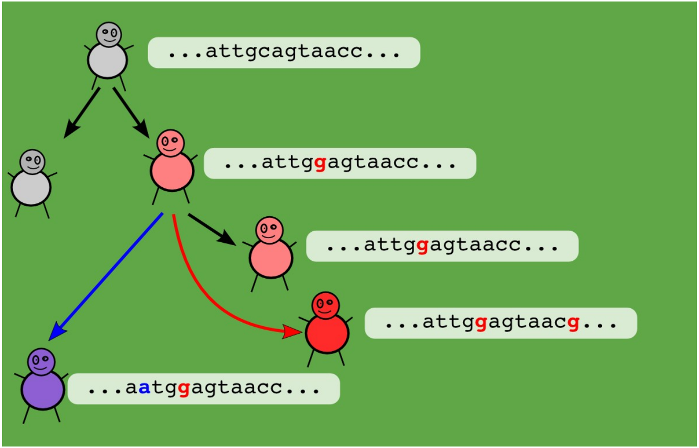


From DNA sequences to patterns of biological diversity

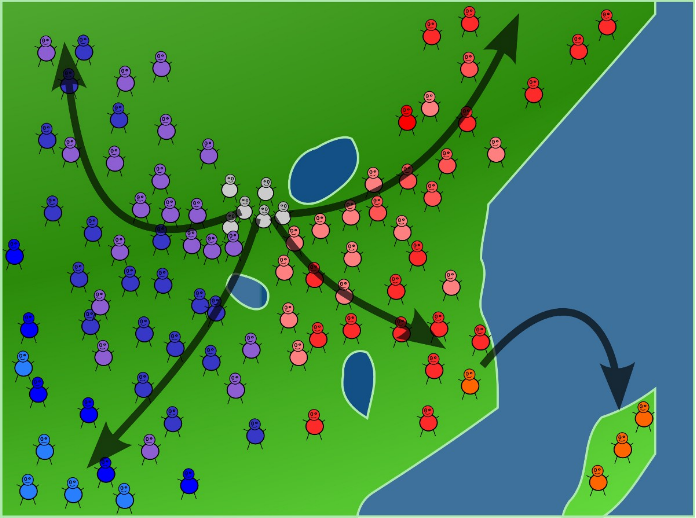
From DNA sequences to patterns of biological diversity



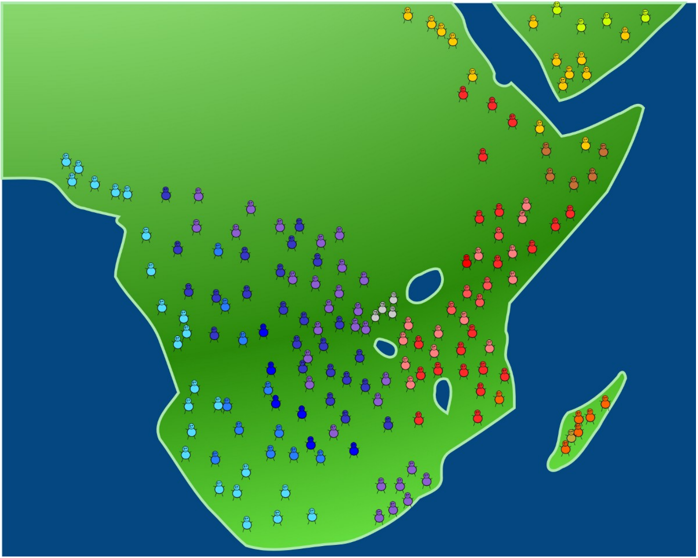
From DNA sequences to patterns of biological diversity



From DNA sequences to patterns of biological diversity

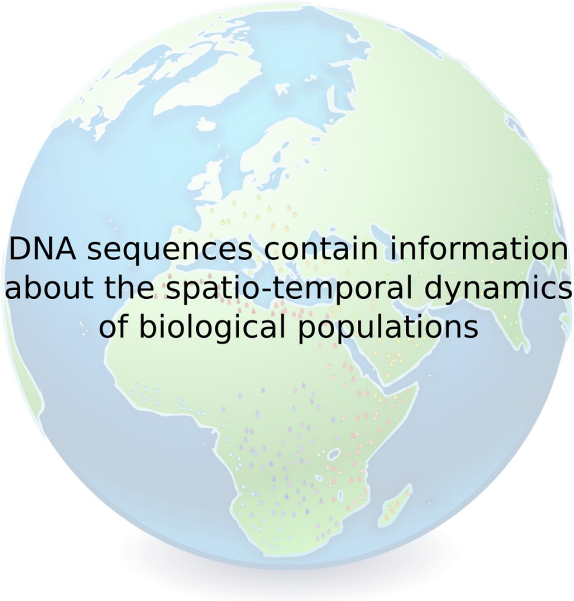


From DNA sequences to patterns of biological diversity



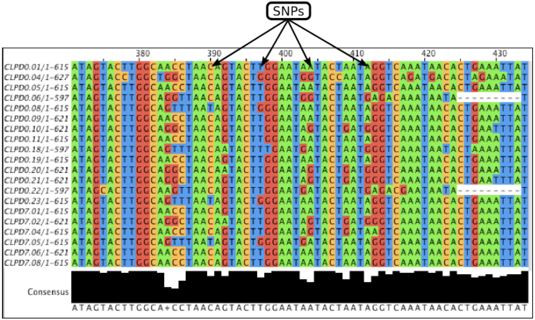
From DNA sequences to patterns of biological diversity

From DNA sequences to patterns of biological diversity



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

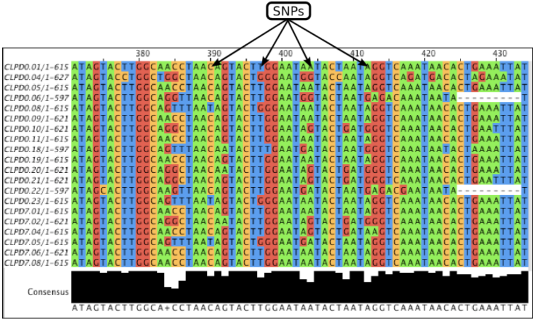
DNA sequences: a rich source of information



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

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

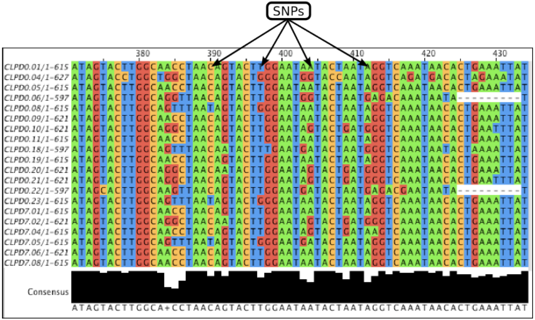
DNA sequences: a rich source of information



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

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

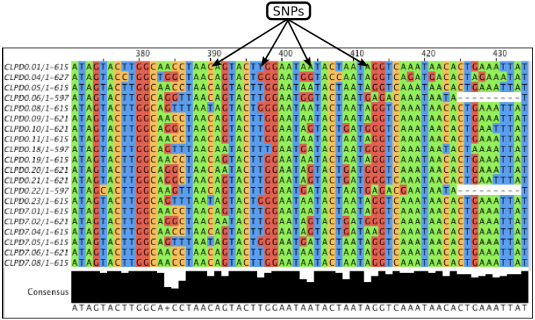
DNA sequences: a rich source of information



- 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

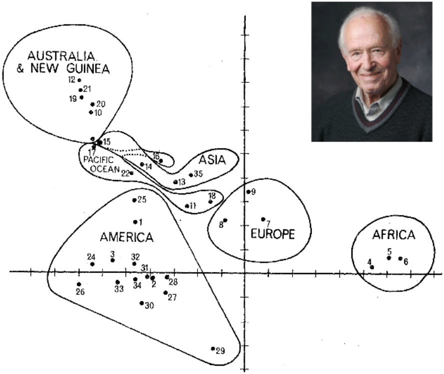


- hundreds/thousands individuals
- up to millions of single nucleotide polymorphism (SNPs)

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

First application of multivariate analysis in genetics

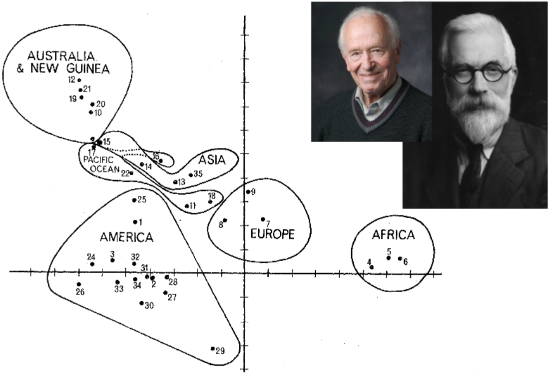
PCA of genetic data, native human populations (Cavalli-Sforza 1966, *Proc B*)



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*)

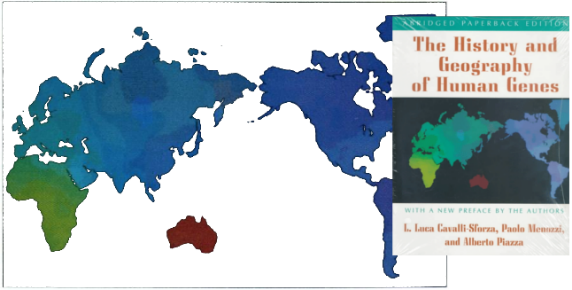


First 2 principal components separate populations into continents.

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

In practice

Multivariate analysis of genetic data using R

Usual workflow

1. read data in (*adegenet*)
2. convert data into numeric values (*adegenet*)
3. replace missing values (*adegenet*)
4. use "classical" methods (*ade4*/*adegenet*)
5. graphics and interpretation (*ade4*/*adegenet*)

In practice

Multivariate analysis of genetic data using R

Usual workflow

1. read data in (*adegenet*)
2. convert data into numeric values (*adegenet*)
3. replace missing values (*adegenet*)
4. use "classical" methods (*ade4*/*adegenet*)
5. graphics and interpretation (*ade4*/*adegenet*)

In practice

Multivariate analysis of genetic data using R

Usual workflow

1. read data in (*adegenet*)
2. convert data into numeric values (*adegenet*)
3. replace missing values (*adegenet*)
4. use "classical" methods (*ade4*/*adegenet*)
5. graphics and interpretation (*ade4*/*adegenet*)

In practice

Multivariate analysis of genetic data using R

Usual workflow

1. read data in (*adegenet*)
2. convert data into numeric values (*adegenet*)
3. replace missing values (*adegenet*)
4. use “classical” methods (*ade4*/*adegenet*)
5. graphics and interpretation (*ade4*/*adegenet*)

In practice

Multivariate analysis of genetic data using R

Usual workflow

1. read data in (*adegenet*)
2. convert data into numeric values (*adegenet*)
3. replace missing values (*adegenet*)
4. use “classical” methods (*ade4*/*adegenet*)
5. graphics and interpretation (*ade4*/*adegenet*)

Recoding data numerically

- Presence/absence (e.g. RFLP, AFLP) and SNPs: binary coding
- Multiallelic data (e.g. microsatellites) are recoded as counts/frequencies

Example using microsatellites:

Raw data:

```
  locus1 locus2
1  80/80  30/30
2  50/55  30/30
3  80/50  29/30
4  50/50  30/30
5  50/50  29/29
```

Recoded data (allele counts):

```
  locus1.50 locus1.55 locus1.80 locus2.29 locus2.30
1         0         0         2         0         2
2         1         1         0         0         2
3         1         0         1         1         1
4         2         0         0         0         2
5         2         0         0         2         0
```

Recoding data numerically

- Presence/absence (e.g. RFLP, AFLP) and SNPs: binary coding
- Multiallelic data (e.g. microsatellites) are recoded as counts/frequencies

Example using microsatellites:

Raw data:

```
  locus1 locus2
1  80/80  30/30
2  50/55  30/30
3  80/50  29/30
4  50/50  30/30
5  50/50  29/29
```

Recoded data (allele counts):

```
  locus1.50 locus1.55 locus1.80 locus2.29 locus2.30
1         0         0         2         0         2
2         1         1         0         0         2
3         1         0         1         1         1
4         2         0         0         0         2
5         2         0         0         2         0
```

Recoding data numerically

- Presence/absence (e.g. RFLP, AFLP) and SNPs: binary coding
- Multiallelic data (e.g. microsatellites) are recoded as counts/frequencies

Example using microsatellites:

Raw data:

```
  locus1 locus2
1  80/80  30/30
2  50/55  30/30
3  80/50  29/30
4  50/50  30/30
5  50/50  29/29
```

Recoded data (allele counts):

```
  locus1.50 locus1.55 locus1.80 locus2.29 locus2.30
1         0         0         2         0         2
2         1         1         0         0         2
3         1         0         1         1         1
4         2         0         0         0         2
5         2         0         0         2         0
```

Recoding data numerically

- Presence/absence (e.g. RFLP, AFLP) and SNPs: binary coding
- Multiallelic data (e.g. microsatellites) are recoded as counts/frequencies

Example using microsatellites:

Raw data:

```
  locus1 locus2
1  80/80  30/30
2  50/55  30/30
3  80/50  29/30
4  50/50  30/30
5  50/50  29/29
```

Recoded data (allele counts):

```
  locus1.50 locus1.55 locus1.80 locus2.29 locus2.30
1         0         0         2         0         2
2         1         1         0         0         2
3         1         0         1         1         1
4         2         0         0         0         2
5         2         0         0         2         0
```

How are data handled in *adegenet*?

Types of data:

- codominant markers (e.g. microsatellites) with any ploidy level → allele counts
- dominant markers (e.g. RAPD) → presence/absence
- nucleotide / amino-acids variation → allele counts
- purely biallelic SNPs → binary data (bits)

Formats:

- software: *GENETIX*, *Fstat*, *Genepop*, *STRUCTURE*, *PLINK*
- `data.frame` of raw allelic data
- `data.frame` of allelic frequencies
- SNPs/amino-acids extracted from DNA/protein alignments

How are data handled in *adegenet*?

Types of data:

- codominant markers (e.g. microsatellites) with any ploidy level → allele counts
- dominant markers (e.g. RAPD) → presence/absence
- nucleotide / amino-acids variation → allele counts
- purely biallelic SNPs → binary data (bits)

Formats:

- software: *GENETIX*, *Fstat*, *Genepop*, *STRUCTURE*, *PLINK*
- `data.frame` of raw allelic data
- `data.frame` of allelic frequencies
- SNPs/amino-acids extracted from DNA/protein alignments

How are data handled in *adegenet*?

Types of data:

- codominant markers (e.g. microsatellites) with any ploidy level → allele counts
- dominant markers (e.g. RAPD) → presence/absence
- nucleotide / amino-acids variation → allele counts
- purely biallelic SNPs → binary data (bits)

Formats:

- software: *GENETIX*, *Fstat*, *Genepop*, *STRUCTURE*, *PLINK*
- `data.frame` of raw allelic data
- `data.frame` of allelic frequencies
- SNPs/amino-acids extracted from DNA/protein alignments

How are data handled in *adegenet*?

Types of data:

- codominant markers (e.g. microsatellites) with any ploidy level → allele counts
- dominant markers (e.g. RAPD) → presence/absence
- nucleotide / amino-acids variation → allele counts
- purely biallelic SNPs → binary data (bits)

Formats:

- software: *GENETIX*, *Fstat*, *Genepop*, *STRUCTURE*, *PLINK*
- `data.frame` of raw allelic data
- `data.frame` of allelic frequencies
- SNPs/amino-acids extracted from DNA/protein alignments

How are data handled in *adegenet*?

Types of data:

- codominant markers (e.g. microsatellites) with any ploidy level → allele counts
- dominant markers (e.g. RAPD) → presence/absence
- nucleotide / amino-acids variation → allele counts
- purely biallelic SNPs → binary data (bits)

Formats:

- software: *GENETIX*, *Fstat*, *Genepop*, *STRUCTURE*, *PLINK*
- `data.frame` of raw allelic data
- `data.frame` of allelic frequencies
- SNPs/amino-acids extracted from DNA/protein alignments

How are data handled in *adegenet*?

Types of data:

- codominant markers (e.g. microsatellites) with any ploidy level → allele counts
- dominant markers (e.g. RAPD) → presence/absence
- nucleotide / amino-acids variation → allele counts
- purely biallelic SNPs → binary data (bits)

Formats:

- software: *GENETIX*, *Fstat*, *Genepop*, *STRUCTURE*, *PLINK*
- `data.frame` of raw allelic data
- `data.frame` of allelic frequencies
- SNPs/amino-acids extracted from DNA/protein alignments

How are data handled in *adegenet*?

Types of data:

- codominant markers (e.g. microsatellites) with any ploidy level → allele counts
- dominant markers (e.g. RAPD) → presence/absence
- nucleotide / amino-acids variation → allele counts
- purely biallelic SNPs → binary data (bits)

Formats:

- software: *GENETIX*, *Fstat*, *Genepop*, *STRUCTURE*, *PLINK*
- `data.frame` of raw allelic data
- `data.frame` of allelic frequencies
- SNPs/amino-acids extracted from DNA/protein alignments

(Almost) time to get your hands dirty!

And after lunch, the pdf of the practical is online:

`http://adegenet.r-forge.r-project.org/`

or

Google → adegenet → documents → "Lausanne August 2016"