Analyses of genetic data: an overview

T. Jombart

Imperial College London

Bogota
30-11-2010

Outline

Which data and which questions?
- Genetic data
- Objectives

Methodological approaches: an overview
- Different approaches
- Model-based approaches
- Exploratory approaches

Multivariate analysis
- Rationale
- Applications
- In practice

Outline

Which data and which questions?
- Genetic data
- Objectives

Methodological approaches: an overview
- Different approaches
- Model-based approaches
- Exploratory approaches

Multivariate analysis
- Rationale
- Applications
- In practice

Outline

Which data and which questions?
- Genetic data
- Objectives

Methodological approaches: an overview
- Different approaches
- Model-based approaches
- Exploratory approaches

Multivariate analysis
- Rationale
- Applications
- In practice

Genetic markers

Genetic markers:

- genetically inherited traits which vary between individuals
- items of a nucleic or proteic sequence which vary between individuals

Technically:

- a marker is a variable with different states; one marker = one *locus* (pl. *loci*)
- states are called *alleles*
- one individual can possess several alleles for one locus

Genetic markers

Genetic markers:

- genetically inherited traits which vary between individuals
- items of a nucleic or proteic sequence which vary between individuals

Technically:

- a marker is a variable with different states; one marker = one *locus* (pl. *loci*)
- states are called *alleles*
- one individual can possess several alleles for one locus

Microsatellites

microsatellites are chunks of DNA formed by the repetition of a short sequence of nucleotides. Variability lies in the number of repetitions.

Example:
Individual 1: `TTATTATTATTA` / `TTATTA`
Individual 2: `TTATTATTA` / `TTATTATTA`

In practice, recoded as:
Individual 1: 4/2
Individual 2: 3/3

Microsatellites can be highly variable (lots of alleles).

Single Nucleotide Polymorphism (SNP)

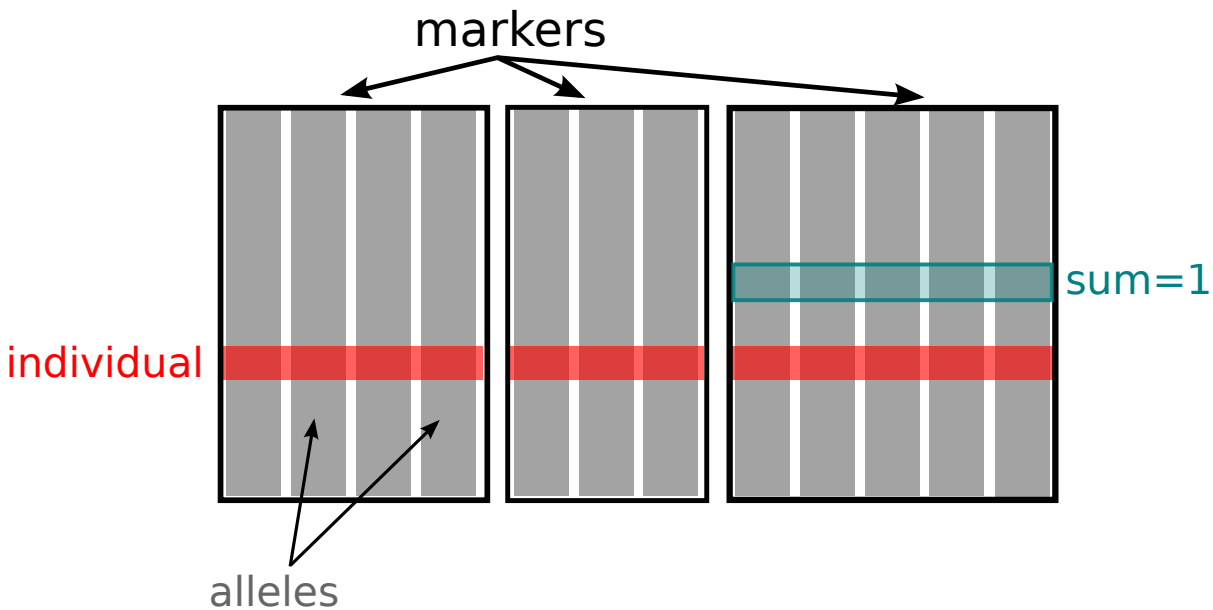
*SNP*s are nucleotides which vary in a set of aligned DNA sequences.

Example:
Individual 1: `ATGGGTATCTG`
Individual 2: `ATGCGTATCAG`

SNPs are often binary (two alleles), but analysed in large numbers (thousands).

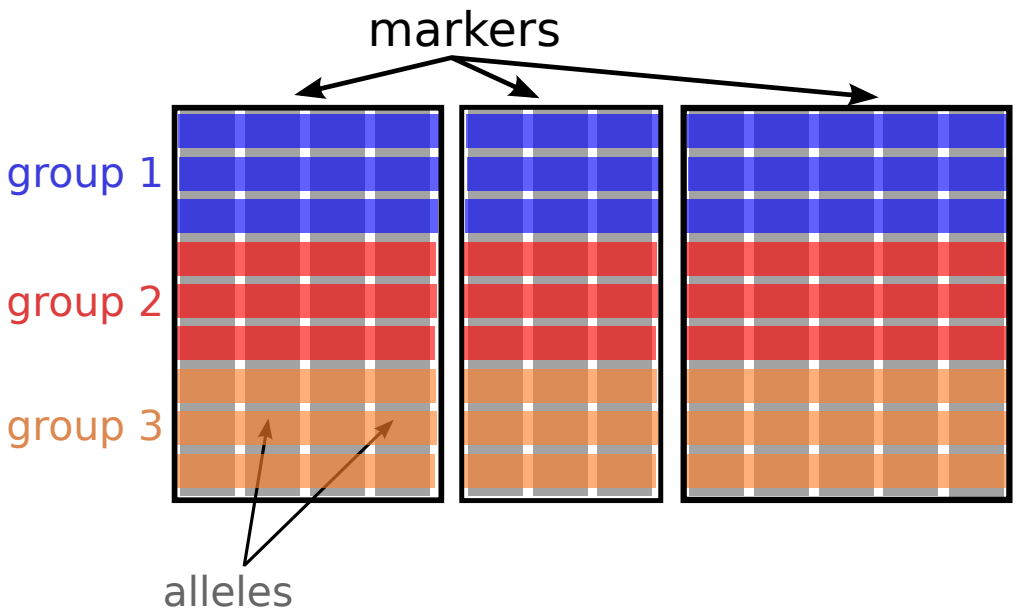
Data: general aspect

- in practice, genetic markers form tables of relative frequencies of alleles.
- data can be aggregated by groups



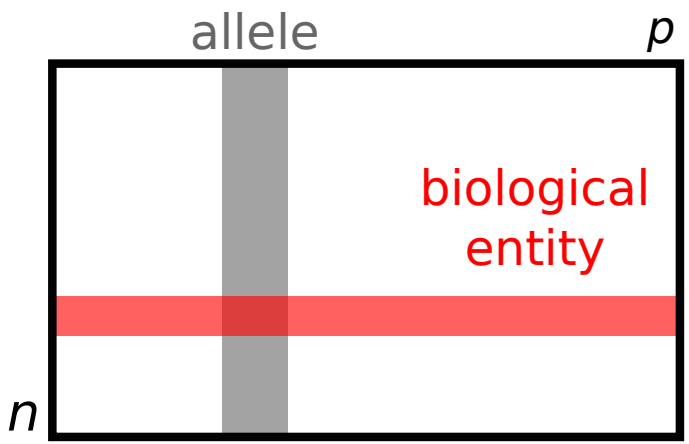
Data: general aspect

- in practice, genetic markers form tables of relative frequencies of alleles.
- data can be aggregated by groups



Data: general aspect

- in practice, genetic markers form tables of relative frequencies of alleles.
- data can be aggregated by groups



In general, we analyse allele frequencies of individuals or groups (*biological entities*).

Data coding

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:

	locus1.50	locus1.55	locus1.80	locus2.29	locus2.30
1	0.0	0.0	1.0	0.0	1.0
2	0.5	0.5	0.0	0.0	1.0
3	0.5	0.0	0.5	0.5	0.5
4	1.0	0.0	0.0	0.0	1.0
5	1.0	0.0	0.0	1.0	0.0

Data coding

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:

	locus1.50	locus1.55	locus1.80	locus2.29	locus2.30
1	0.0	0.0	1.0	0.0	1.0
2	0.5	0.5	0.0	0.0	1.0
3	0.5	0.0	0.5	0.5	0.5
4	1.0	0.0	0.0	0.0	1.0
5	1.0	0.0	0.0	1.0	0.0

Outline

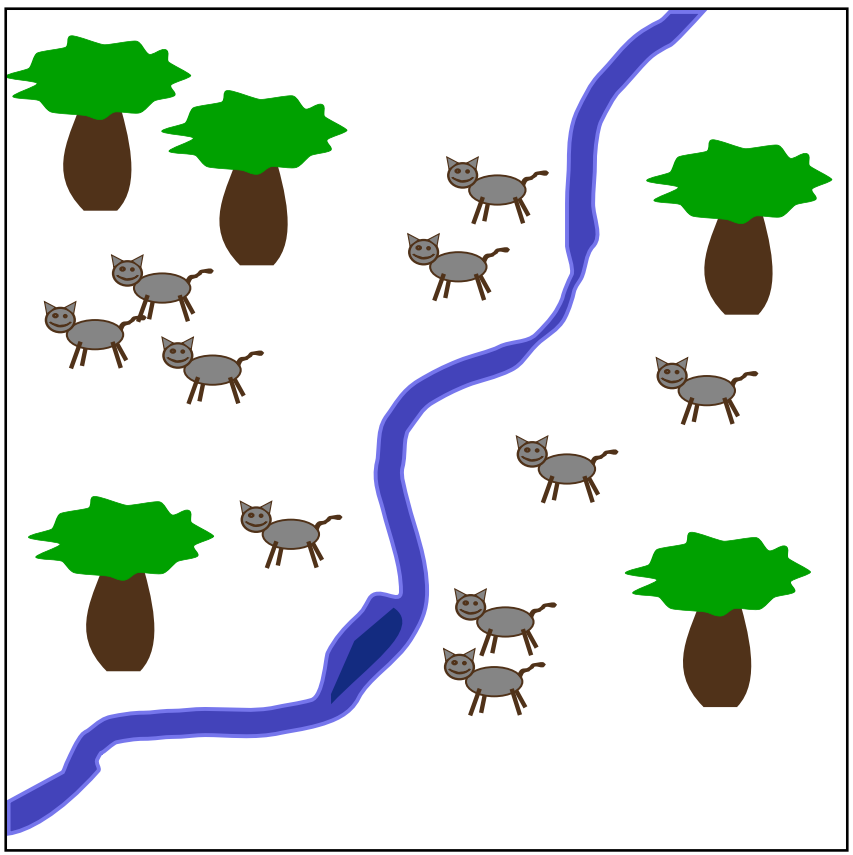
Which data and which questions?
- Genetic data
- Objectives

Methodological approaches: an overview
- Different approaches
- Model-based approaches
- Exploratory approaches

Multivariate analysis
- Rationale
- Applications
- In practice

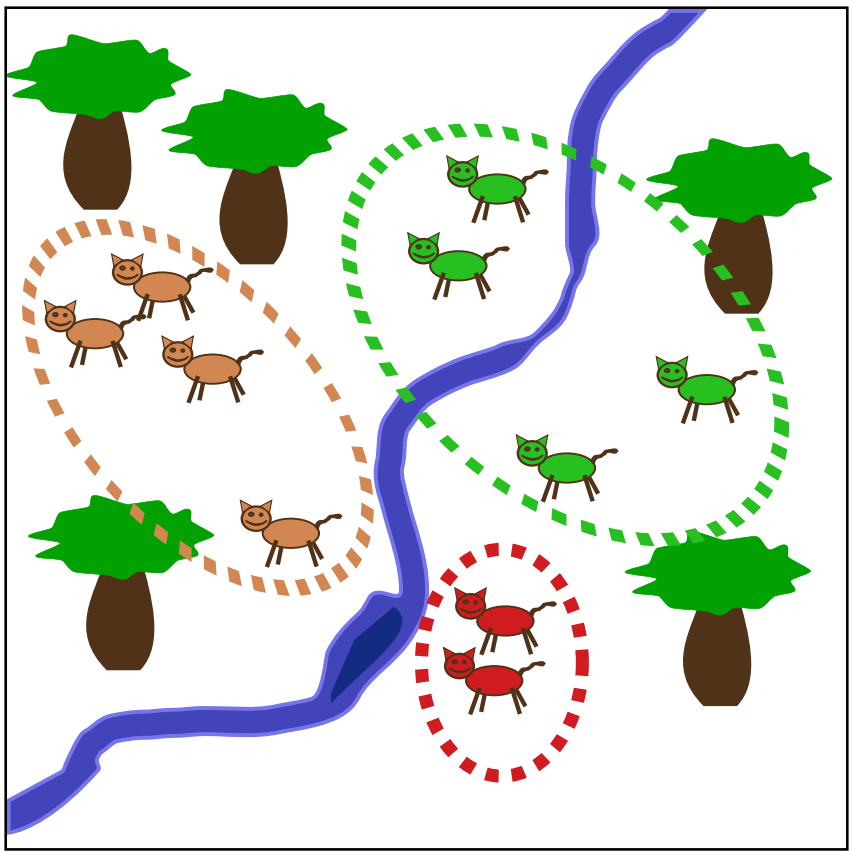
Identifying genetic groups

Genetic clusters can indicate populations, i.e. groups of random mating individuals.



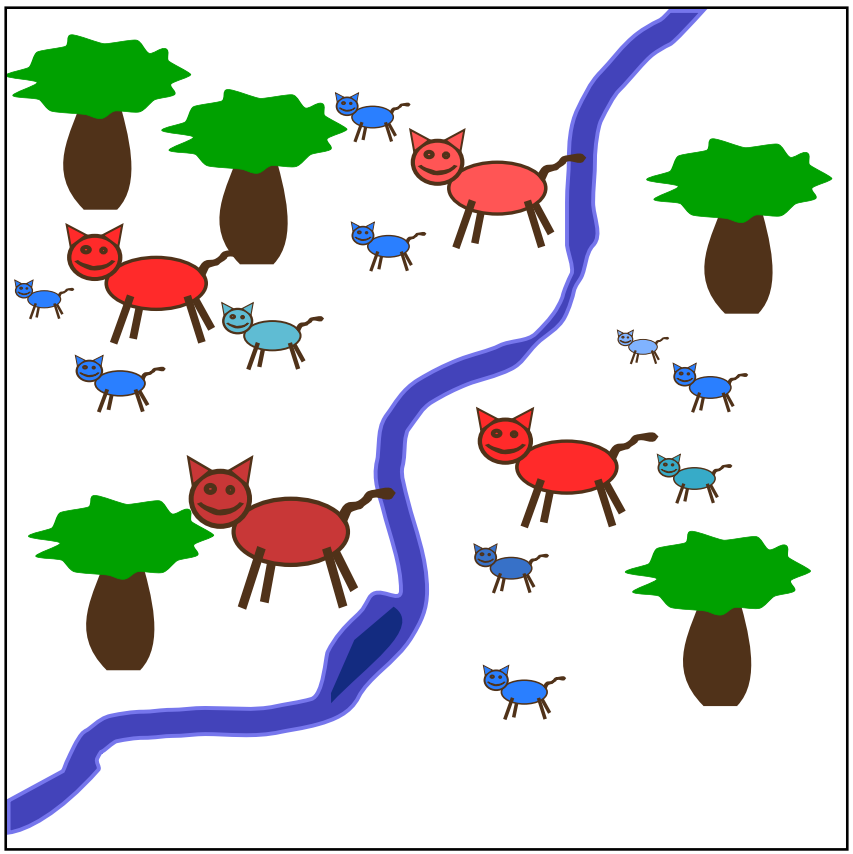
Identifying genetic groups

Genetic clusters can indicate populations, i.e. groups of random mating individuals.



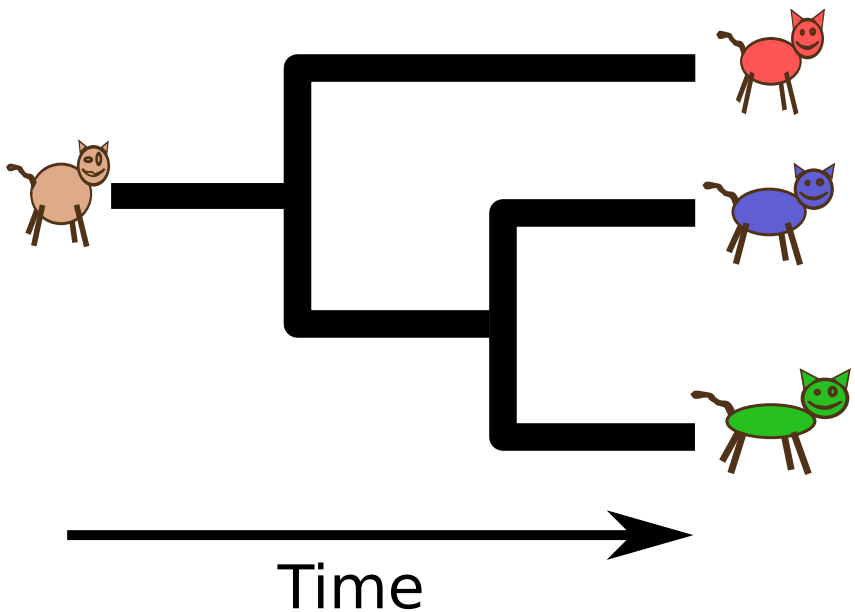
Correlating genotype and phenotype

Correlations between alleles and phenotypic features can indicate the genetic determinism of a trait.



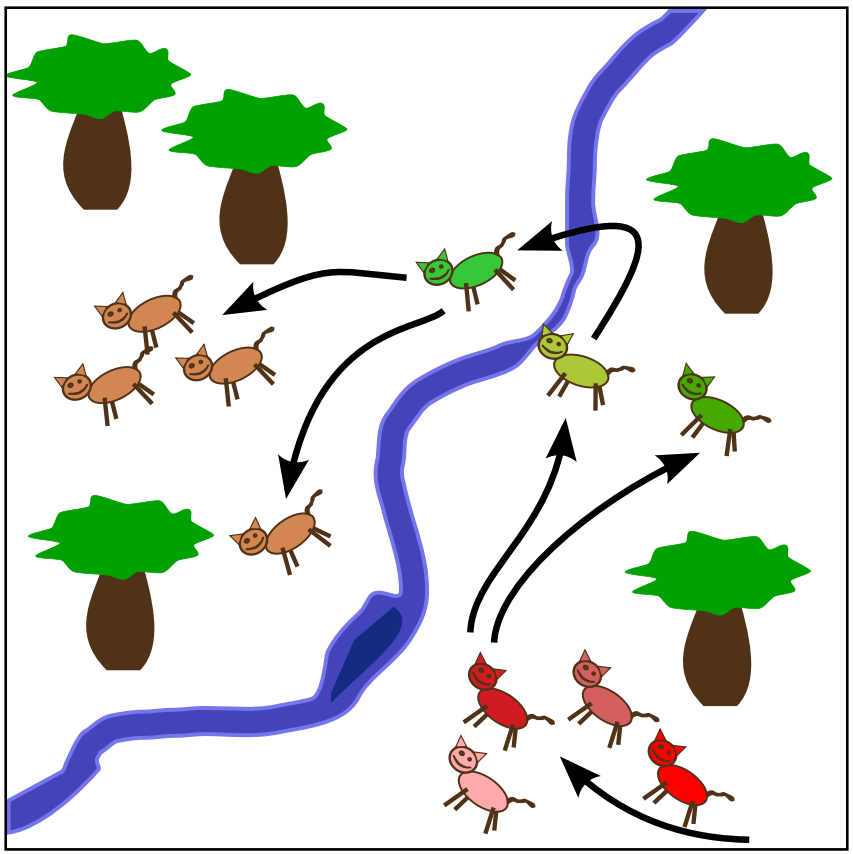
Reconstructing evolutionary history

Genetic distances can be used to infer the evolutionary history of a set of taxa.



Inferring migration routes

Examination of genetic diversity can be used to infer migration routes.



Many possible applications

- different types of genetic markers (microsatellites, SNPs, etc.)
- wide range of questions (identifying populations, reconstructing evolutionary history and migrations, etc.)

$\longrightarrow$ a large variety of methods.

Many possible applications

- different types of genetic markers (microsatellites, SNPs, etc.)
- wide range of questions (identifying populations, reconstructing evolutionary history and migrations, etc.)

$\longrightarrow$ a large variety of methods.

Many possible applications

- different types of genetic markers (microsatellites, SNPs, etc.)
- wide range of questions (identifying populations, reconstructing evolutionary history and migrations, etc.)

$\longrightarrow$ a large variety of methods.

Outline

Which data and which questions?
- Genetic data
- Objectives

Methodological approaches: an overview
- Different approaches
- Model-based approaches
- Exploratory approaches

Multivariate analysis
- Rationale
- Applications
- In practice

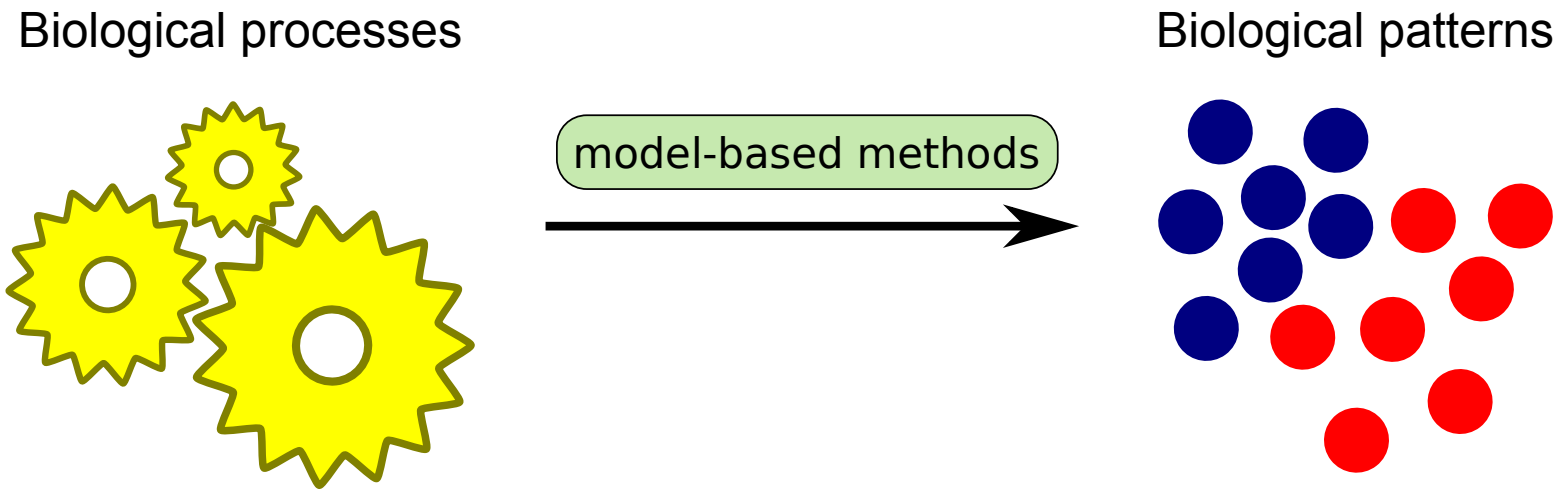
Outline

Which data and which questions?
- Genetic data
- Objectives

Methodological approaches: an overview
- Different approaches
- Model-based approaches
- Exploratory approaches

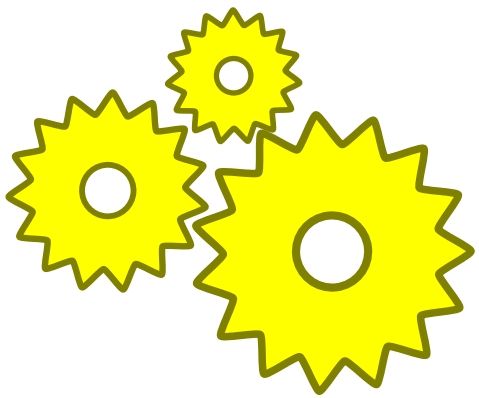
Multivariate analysis
- Rationale
- Applications
- In practice

Two complementary approaches



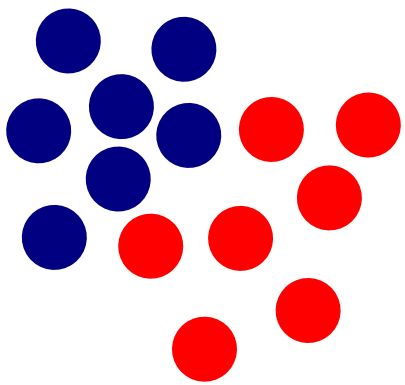
Two complementary approaches

Biological processes



exploratory methods

Biological patterns



A wide variety of methods

Model-based approaches:

- Bayesian clustering (e.g. STRUCTURE, BAPS)
- phylogenetic reconstruction (e.g. phangorn, BEAST)

Exploratory approaches:

- distance-based trees (e.g. NJ, UPGMA)
- multivariate methods (e.g. PCA, PCoA)

A wide variety of methods

Model-based approaches:

- Bayesian clustering (e.g. STRUCTURE, BAPS)
- phylogenetic reconstruction (e.g. phangorn, BEAST)

Exploratory approaches:

- distance-based trees (e.g. NJ, UPGMA)
- multivariate methods (e.g. PCA, PCoA)

Outline

Which data and which questions?
- Genetic data
- Objectives

Methodological approaches: an overview
- Different approaches
- Model-based approaches
- Exploratory approaches

Multivariate analysis
- Rationale
- Applications
- In practice

Bayesian clustering

- formulate a model of the data y with parameters θ (e.g. group membership) and prior distribution ($p(\theta)$)
- formulate a likelihood $p(y|\theta)$ for a given set of parameters θ
- obtain the posterior distribution of the parameters $p(\theta|y)$

The posterior distribution is determined as (Bayes'formula):

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

or in practice, since $p(y)$ is often unknown:

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

In other words:

$$\text{param. given data} \propto \text{data given param.} \times \text{prior param.}$$

Bayesian clustering

- formulate a model of the data y with parameters θ (e.g. group membership) and prior distribution ($p(\theta)$)
- formulate a likelihood $p(y|\theta)$ for a given set of parameters θ
- obtain the posterior distribution of the parameters $p(\theta|y)$

The posterior distribution is determined as (Bayes'formula):

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

or in practice, since $p(y)$ is often unknown:

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

In other words:

$$\text{param. given data} \propto \text{data given param.} \times \text{prior param.}$$

Bayesian clustering

- formulate a model of the data y with parameters θ (e.g. group membership) and prior distribution ($p(\theta)$)
- formulate a likelihood $p(y|\theta)$ for a given set of parameters θ
- obtain the posterior distribution of the parameters $p(\theta|y)$

The posterior distribution is determined as (Bayes'formula):

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

or in practice, since $p(y)$ is often unknown:

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

In other words:

$$\text{param. given data} \propto \text{data given param.} \times \text{prior param.}$$

Bayesian clustering

- formulate a model of the data y with parameters θ (e.g. group membership) and prior distribution ($p(\theta)$)
- formulate a likelihood $p(y|\theta)$ for a given set of parameters θ
- obtain the posterior distribution of the parameters $p(\theta|y)$

The posterior distribution is determined as (Bayes'formula):

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

or in practice, since $p(y)$ is often unknown:

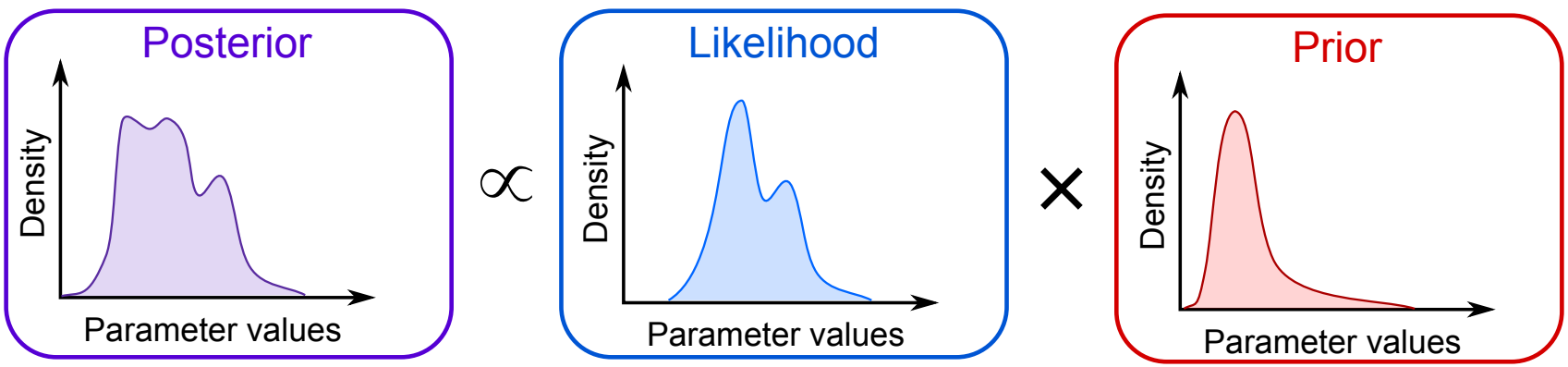
$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

In other words:

$$\text{param. given data} \propto \text{data given param.} \times \text{prior param.}$$

Bayesian clustering

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$



Bayesian clustering

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

In practice, we sample from $p(\theta|y)$ using Markov Chain Monte Carlo (MCMC) algorithms.

MCMC (at step t):

1. simulate candidate values of the parameters θ^t
2. compute $p(\theta^t|y)$
3. accept/reject θ^t as a part of the posterior distribution based on the comparison of $p(\theta^t|y)$ with previous step $p(\theta^{t-1}|y)$
4. increase t, return to step 1

Bayesian clustering

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

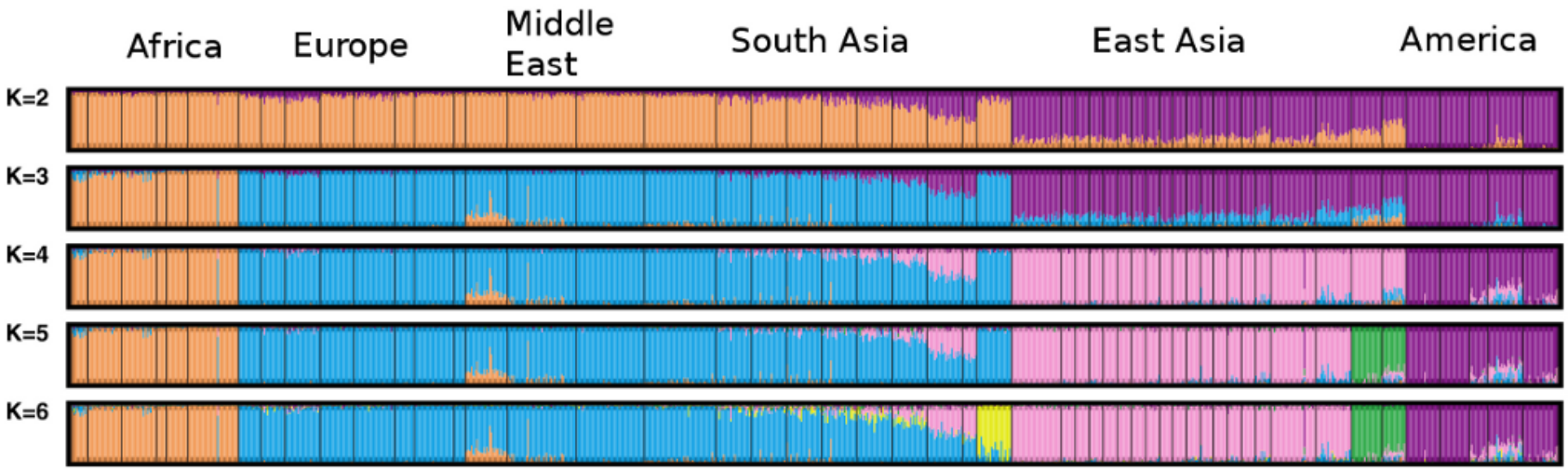
In practice, we sample from $p(\theta|y)$ using Markov Chain Monte Carlo (MCMC) algorithms.

MCMC (at step t):

1. simulate candidate values of the parameters θ^t
2. compute $p(\theta^t|y)$
3. accept/reject θ^t as a part of the posterior distribution based on the comparison of $p(\theta^t|y)$ with previous step $p(\theta^{t-1}|y)$
4. increase t, return to step 1

Bayesian clustering: example

STRUCTURE analysis of Human data (377 microsatellites, 52 populations, 1056 individuals)



(Rosenberg et al 2002, *Science*)

Phylogenetic reconstruction

Recontruct a tree (tips=observations, nodes=inferred ancestors) which fulfills a given criterion.

Different approaches:

- maximum *parsimony*: minimal number of changes between ancestors and descendents
- maximum *likelihood*: find θ so that $p(y|\theta)$ is maximum
- Bayesian approach: estimate $p(\theta|y)$ based on Bayes' formula

Phylogenetic reconstruction

Recontruct a tree (tips=observations, nodes=inferred ancestors) which fulfills a given criterion.

Different approaches:

- maximum *parsimony*: minimal number of changes between ancestors and descendents
- maximum *likelihood*: find θ so that $p(y|\theta)$ is maximum
- Bayesian approach: estimate $p(\theta|y)$ based on Bayes' formula

Phylogenetic reconstruction

Recontruct a tree (tips=observations, nodes=inferred ancestors) which fulfills a given criterion.

Different approaches:

- maximum *parsimony*: minimal number of changes between ancestors and descendents
- maximum *likelihood*: find θ so that $p(y|\theta)$ is maximum
- Bayesian approach: estimate $p(\theta|y)$ based on Bayes' formula

Phylogenetic reconstruction

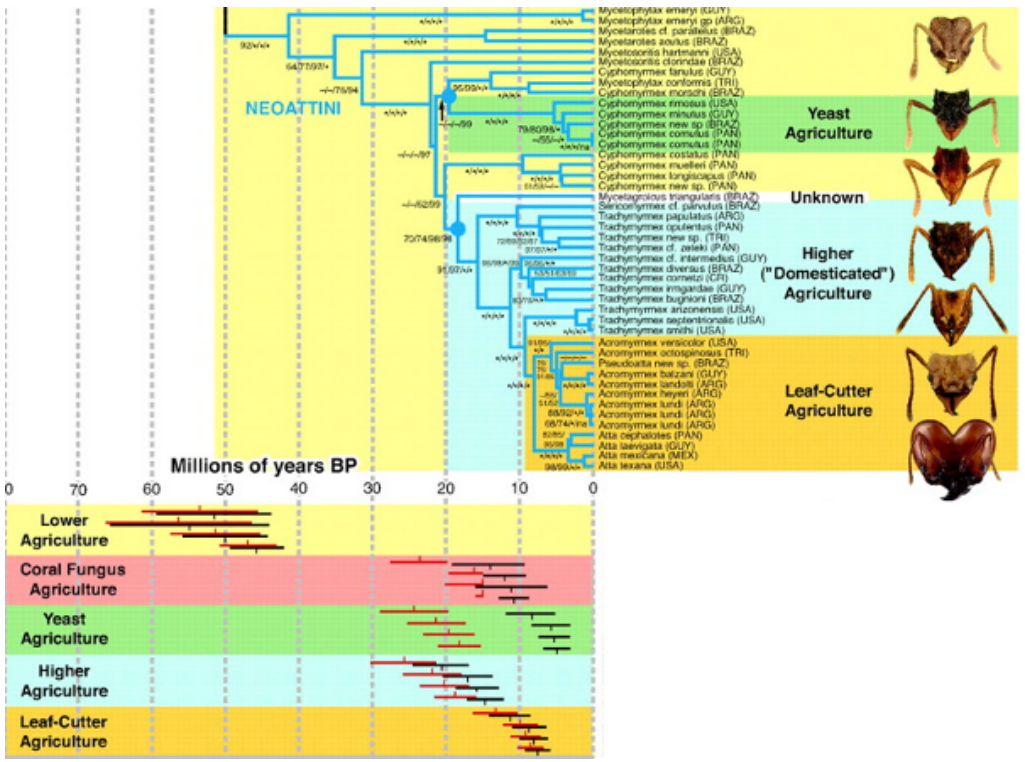
Recontruct a tree (tips=observations, nodes=inferred ancestors) which fulfills a given criterion.

Different approaches:

- maximum *parsimony*: minimal number of changes between ancestors and descendents
- maximum *likelihood*: find θ so that $p(y|\theta)$ is maximum
- Bayesian approach: estimate $p(\theta|y)$ based on Bayes' formula

Phylogenetic reconstruction: example

Time-callibrated phylogeny of ants species using BEAST.



(Schultz & Brady 2008, *PNAS*)

Outline

Which data and which questions?
- Genetic data
- Objectives

Methodological approaches: an overview
- Different approaches
- Model-based approaches
- Exploratory approaches

Multivariate analysis
- Rationale
- Applications
- In practice

Distance-based trees

Often confused with phylogenies. In fact, these trees do not generally try to reconstruct the evolutionary history. They just represent genetic distances.

General approach:

- compute a genetic distance between biological entities (e.g. individuals, populations, species)
- use hierarchical clustering algorithm on the genetic distance matrix
- use bootstrap to evaluate uncertainties in the bifurcations

Distance-based trees

Often confused with phylogenies. In fact, these trees do not generally try to reconstruct the evolutionary history. They just represent genetic distances.

General approach:

- compute a genetic distance between biological entities (e.g. individuals, populations, species)
- use hierarchical clustering algorithm on the genetic distance matrix
- use bootstrap to evaluate uncertainties in the bifurcations

Distance-based trees

Often confused with phylogenies. In fact, these trees do not generally try to reconstruct the evolutionary history. They just represent genetic distances.

General approach:

- compute a genetic distance between biological entities (e.g. individuals, populations, species)
- use hierarchical clustering algorithm on the genetic distance matrix
- use bootstrap to evaluate uncertainties in the bifurcations

Distance-based trees

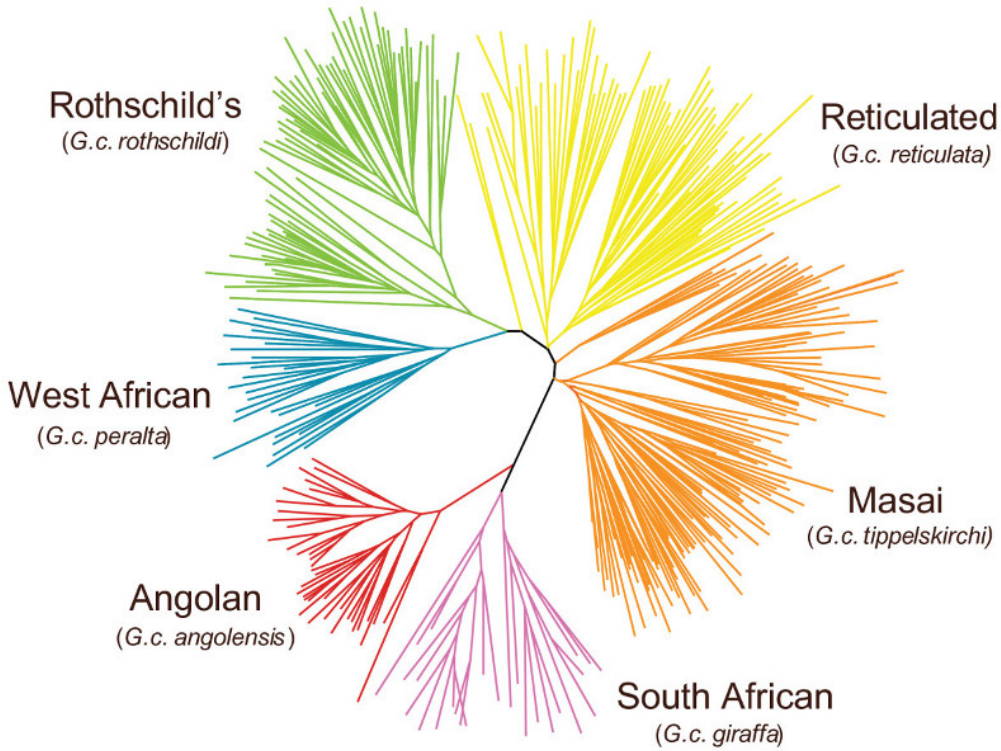
Often confused with phylogenies. In fact, these trees do not generally try to reconstruct the evolutionary history. They just represent genetic distances.

General approach:

- compute a genetic distance between biological entities (e.g. individuals, populations, species)
- use hierarchical clustering algorithm on the genetic distance matrix
- use bootstrap to evaluate uncertainties in the bifurcations

Distance-based trees: example

Unexpected biodiversity in the giraffe based on 14 microsatellites (allele-sharing distance + NJ).



(Brown et al. 2007, *BMC Biology*)

Multivariate analysis

Represent genetic distances using a small number of *synthetic variables*.

How do they work?
What can they be used for?
What software to use?
...

Outline

Which data and which questions?
- Genetic data
- Objectives

Methodological approaches: an overview
- Different approaches
- Model-based approaches
- Exploratory approaches

Multivariate analysis
- Rationale
- Applications
- In practice

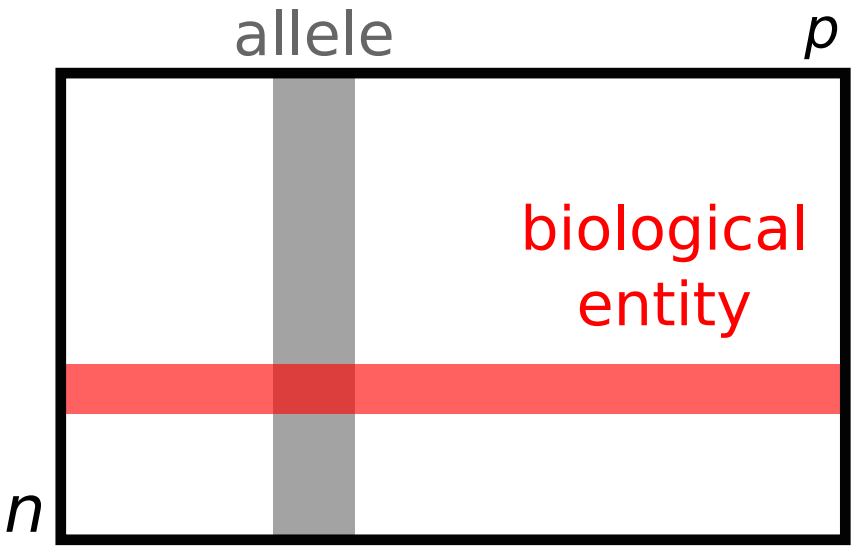
Outline

Which data and which questions?
- Genetic data
- Objectives

Methodological approaches: an overview
- Different approaches
- Model-based approaches
- Exploratory approaches

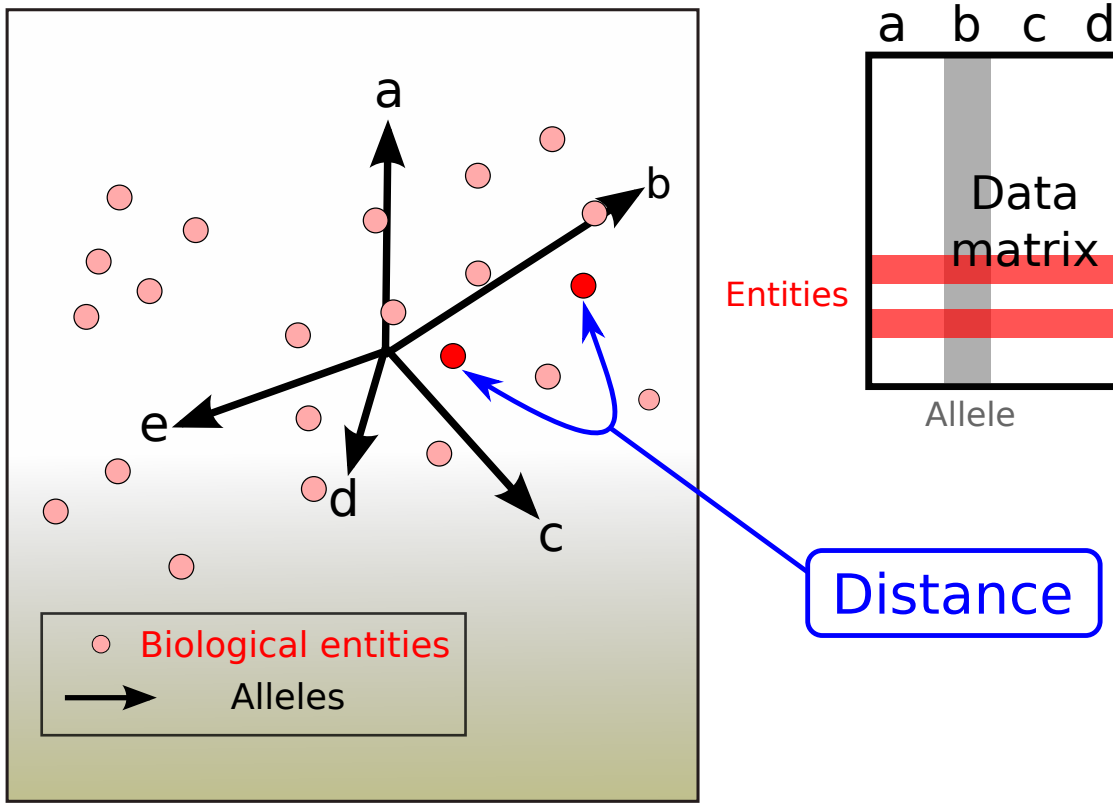
Multivariate analysis
- Rationale
- Applications
- In practice

Recall: the data



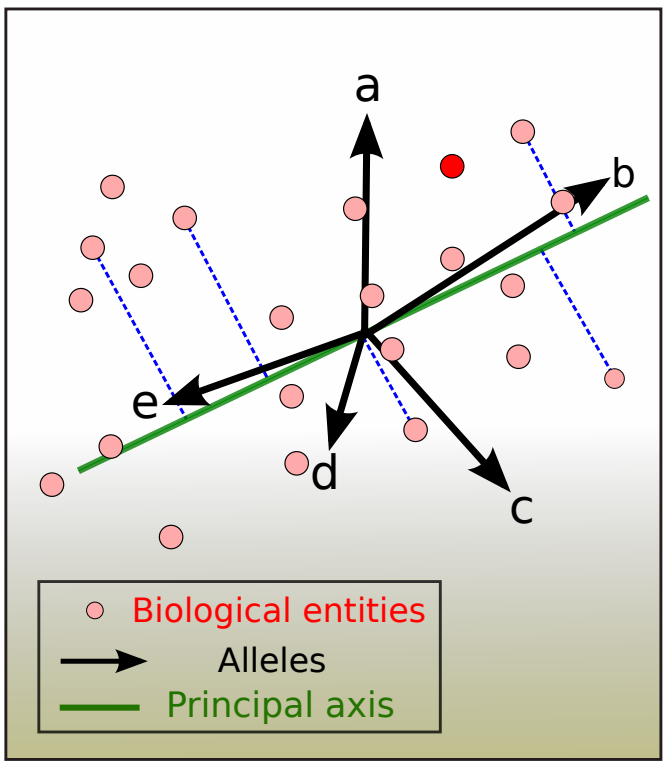
Geometric approach of the data

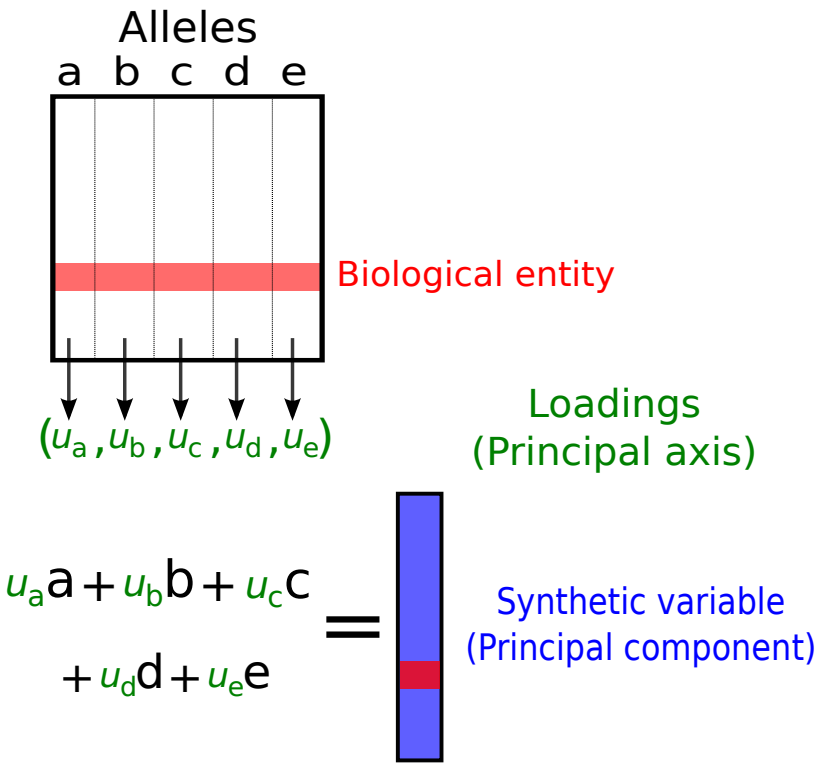
Data form a could of points in a multivariate space:



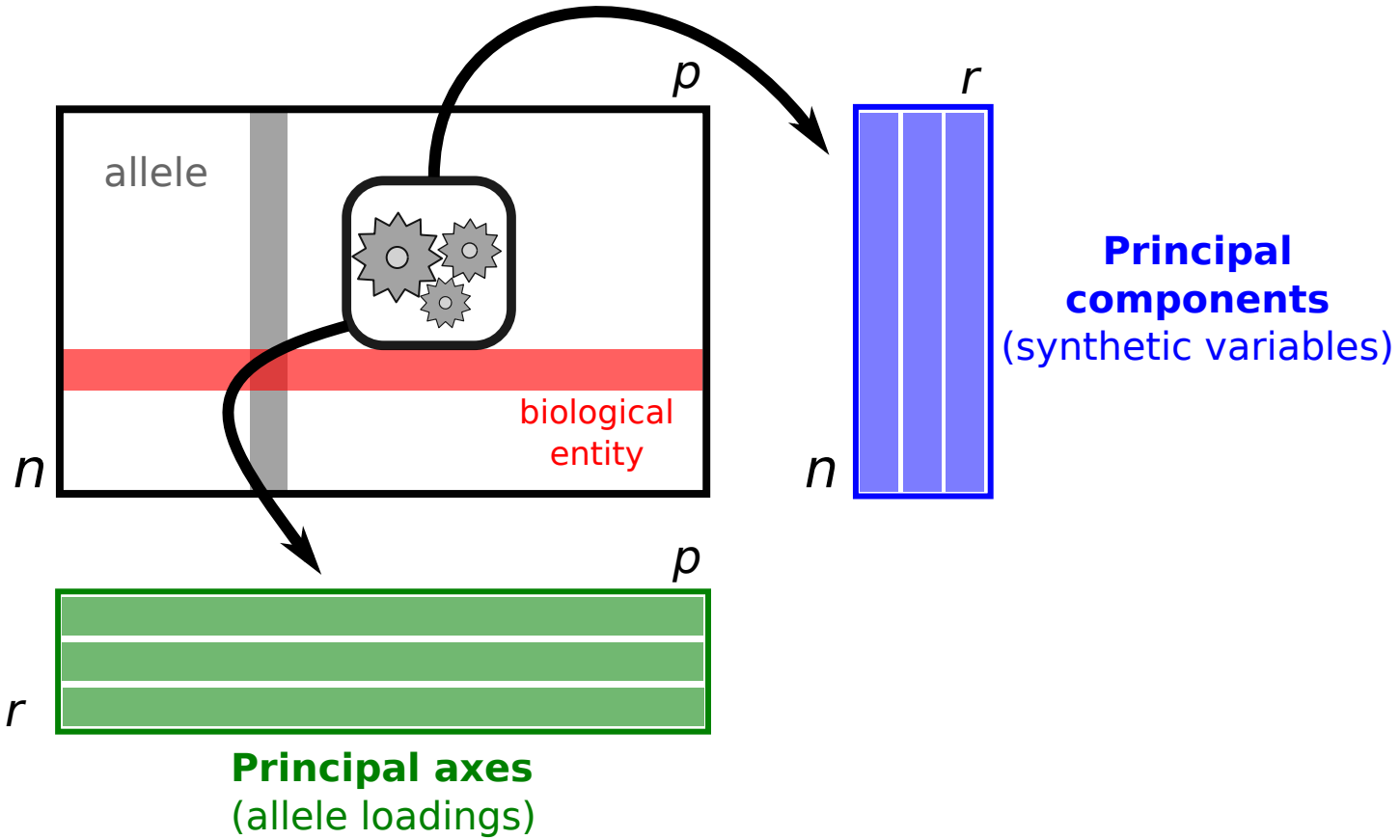
Rationale of multivariate analyses

Finding the directions showing most information.





Outputs of multivariate analyses



- principal components: summarize genetic diversity
- principal axes: allele contributions

Multivariate analyses: computations

Notations:

- n: number of entities
- p: number of alleles
- $\mathbf{X} \in \mathbb{R}^{n \times p}$: transformed data matrix (e.g. centred/scaled allele frequencies)
- $\mathbf{D}$: a metric in $\mathbb{R}^n$ (weights for entities)
- $\mathbf{u}$: a normed vector of $\mathbb{R}^p$, i.e. $\|\mathbf{u}\|^2 = 1$ (loadings)
- $\mathbf{Xu} \in \mathbb{R}^n$: a linear combination of alleles

Problem: find $\mathbf{u}$ so that $\mathbf{Xu}$ are most scattered
$\Leftrightarrow$ $\mathbf{u}$ so that $\|\mathbf{Xu}\|^2_{\mathbf{D}}$ is maximum.

Multivariate analyses: computations

Notations:

- n: number of entities
- p: number of alleles
- $\mathbf{X} \in \mathbb{R}^{n \times p}$: transformed data matrix (e.g. centred/scaled allele frequencies)
- $\mathbf{D}$: a metric in $\mathbb{R}^n$ (weights for entities)
- $\mathbf{u}$: a normed vector of $\mathbb{R}^p$, i.e. $\|\mathbf{u}\|^2 = 1$ (loadings)
- $\mathbf{Xu} \in \mathbb{R}^n$: a linear combination of alleles

Problem: find $\mathbf{u}$ so that $\mathbf{Xu}$ are most scattered
$\Leftrightarrow$ $\mathbf{u}$ so that $\|\mathbf{Xu}\|^2_{\mathbf{D}}$ is maximum.

Multivariate analyses: computations

Problem: $\mathbf{u}$ so that $\|\mathbf{Xu}\|^2_{\mathbf{D}}$ is maximum.

Solution:

- take $\mathbf{u_1}$ the first eigenvector of $\mathbf{X}^T\mathbf{DX}$ associated to the eigenvalue λ_1; then $\|\mathbf{Xu_1}\|^2_{\mathbf{D}}$ is maximum and equates λ_1
- with the constraint $\mathbf{u}_1 \perp \mathbf{u}_2$, $\|\mathbf{Xu}_2\|^2_{\mathbf{D}}$ is maximum and equates λ_2
- further axes $\mathbf{u}_i$ maximise $\|\mathbf{Xu}_i\|^2_{\mathbf{D}}$ and are orthogonal to previous axes (i.e., $\mathbf{u}_i \perp \mathbf{u}_j \ \forall \ j < i$)

Multivariate analyses: computations

Problem: $\mathbf{u}$ so that $\|\mathbf{Xu}\|^2_{\mathbf{D}}$ is maximum.

Solution:

- take $\mathbf{u_1}$ the first eigenvector of $\mathbf{X}^T\mathbf{DX}$ associated to the eigenvalue λ_1; then $\|\mathbf{Xu_1}\|^2_{\mathbf{D}}$ is maximum and equates λ_1
- with the constraint $\mathbf{u}_1 \perp \mathbf{u}_2$, $\|\mathbf{Xu}_2\|^2_{\mathbf{D}}$ is maximum and equates λ_2
- further axes $\mathbf{u}_i$ maximise $\|\mathbf{Xu}_i\|^2_{\mathbf{D}}$ and are orthogonal to previous axes (i.e., $\mathbf{u}_i \perp \mathbf{u}_j \ \forall \ j < i$)

Multivariate analyses: computations

Problem: $\mathbf{u}$ so that $\|\mathbf{Xu}\|^2_{\mathbf{D}}$ is maximum.

Solution:

- take $\mathbf{u_1}$ the first eigenvector of $\mathbf{X}^T\mathbf{DX}$ associated to the eigenvalue λ_1; then $\|\mathbf{Xu_1}\|^2_{\mathbf{D}}$ is maximum and equates λ_1
- with the constraint $\mathbf{u}_1 \perp \mathbf{u}_2$, $\|\mathbf{Xu}_2\|^2_{\mathbf{D}}$ is maximum and equates λ_2
- further axes $\mathbf{u}_i$ maximise $\|\mathbf{Xu}_i\|^2_{\mathbf{D}}$ and are orthogonal to previous axes (i.e., $\mathbf{u}_i \perp \mathbf{u}_j \ \forall \ j < i$)

Multivariate analyses: computations

Problem: $\mathbf{u}$ so that $\|\mathbf{Xu}\|^2_{\mathbf{D}}$ is maximum.

Solution:

- take $\mathbf{u_1}$ the first eigenvector of $\mathbf{X}^T\mathbf{DX}$ associated to the eigenvalue λ_1; then $\|\mathbf{Xu_1}\|^2_{\mathbf{D}}$ is maximum and equates λ_1
- with the constraint $\mathbf{u}_1 \perp \mathbf{u}_2$, $\|\mathbf{Xu}_2\|^2_{\mathbf{D}}$ is maximum and equates λ_2
- further axes $\mathbf{u}_i$ maximise $\|\mathbf{Xu}_i\|^2_{\mathbf{D}}$ and are orthogonal to previous axes (i.e., $\mathbf{u}_i \perp \mathbf{u}_j \ \forall \ j < i$)

Outline

Which data and which questions?
- Genetic data
- Objectives

Methodological approaches: an overview
- Different approaches
- Model-based approaches
- Exploratory approaches

Multivariate analysis
- Rationale
- Applications
- In practice

In practice: many methods

Methods applied in genetics:

- Principal Component Analysis (PCA):
  - centred / not centred / fancy centring
  - scaled / not scaled / fancy scaled
  - special transformation for frequency data
- Principal Coordinates Analysis (PCoA) / Metric Multidimensional Scaling (MDS):
  - many genetic distances available
- Correspondance Analysis (CA)
- Discriminant Analysis (DA)
- ... review in Jombart et al. 2009, *Heredity* **102**: 330-341

In practice: many methods

Methods applied in genetics:

- Principal Component Analysis (PCA):
  - centred / not centred / fancy centring
  - scaled / not scaled / fancy scaled
  - special transformation for frequency data
- Principal Coordinates Analysis (PCoA) / Metric Multidimensional Scaling (MDS):
  - many genetic distances available
- Correspondance Analysis (CA)
- Discriminant Analysis (DA)
- ... review in Jombart et al. 2009, *Heredity* **102**: 330-341

In practice: many methods

Methods applied in genetics:

- Principal Component Analysis (PCA):
  - centred / not centred / fancy centring
  - scaled / not scaled / fancy scaled
  - special transformation for frequency data
- Principal Coordinates Analysis (PCoA) / Metric Multidimensional Scaling (MDS):
  - many genetic distances available
- Correspondance Analysis (CA)
- Discriminant Analysis (DA)
- ... review in Jombart et al. 2009, *Heredity* **102**: 330-341

In practice: many methods

Methods applied in genetics:

- Principal Component Analysis (PCA):
  - centred / not centred / fancy centring
  - scaled / not scaled / fancy scaled
  - special transformation for frequency data
- Principal Coordinates Analysis (PCoA) / Metric Multidimensional Scaling (MDS):
  - many genetic distances available
- Correspondance Analysis (CA)
- Discriminant Analysis (DA)
- ... review in Jombart et al. 2009, *Heredity* **102**: 330-341

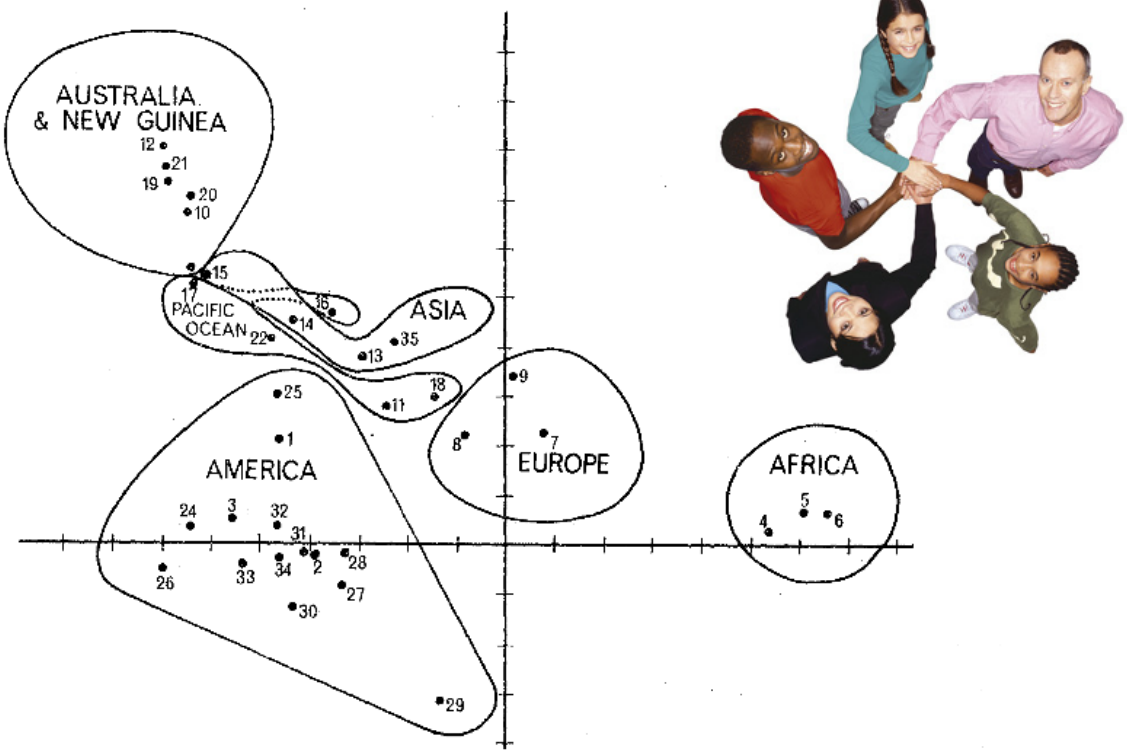
In practice: many methods

Methods applied in genetics:

- Principal Component Analysis (PCA):
  - centred / not centred / fancy centring
  - scaled / not scaled / fancy scaled
  - special transformation for frequency data
- Principal Coordinates Analysis (PCoA) / Metric Multidimensional Scaling (MDS):
  - many genetic distances available
- Correspondance Analysis (CA)
- Discriminant Analysis (DA)
- ... review in Jombart et al. 2009, *Heredity* **102**: 330-341

Applications: some examples

PCA; genetic diversity in native human populations



(Cavalli-Sforza 1966, *Proc B*)

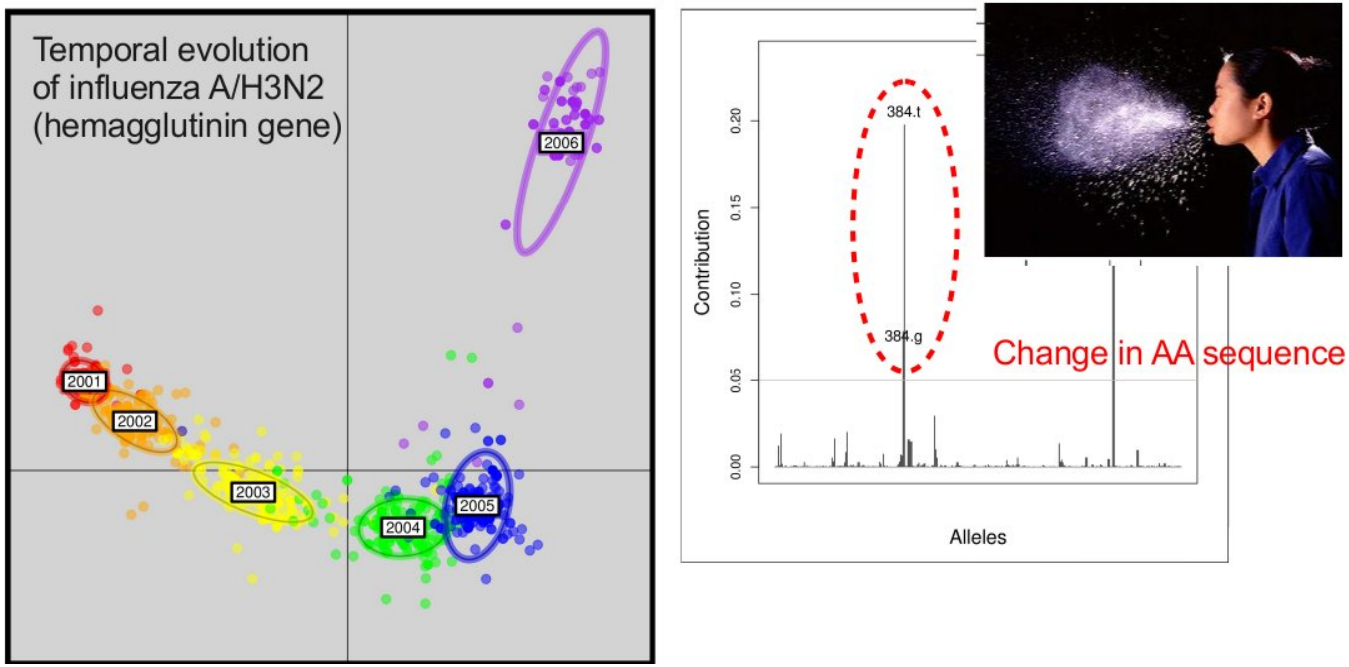
Applications: some examples

PCA + mapping of PC; genetic diversity in native human populations

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

Applications: some examples

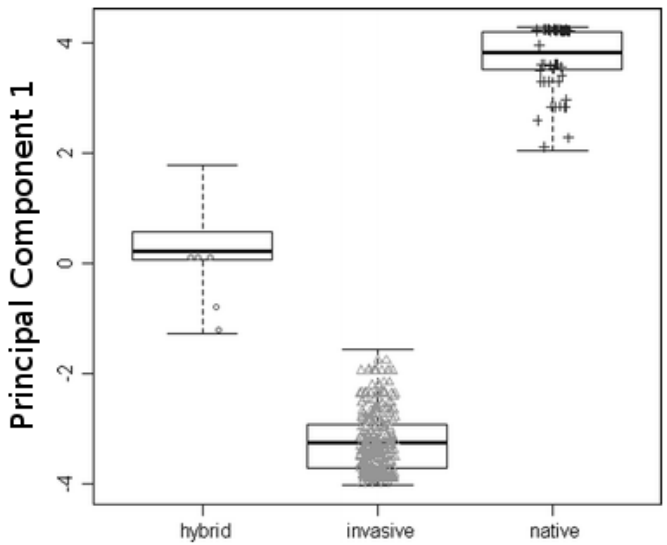
DAPC; genetic evolution of influenza A (H3N2)



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

Applications: some examples

PCA; hybridization in common reed *Phragmites australis*



(Paul et al. 2010, *Biological Invasions*)

Outline

Which data and which questions?
- Genetic data
- Objectives

Methodological approaches: an overview
- Different approaches
- Model-based approaches
- Exploratory approaches

Multivariate analysis
- Rationale
- Applications
- In practice

The scary picture

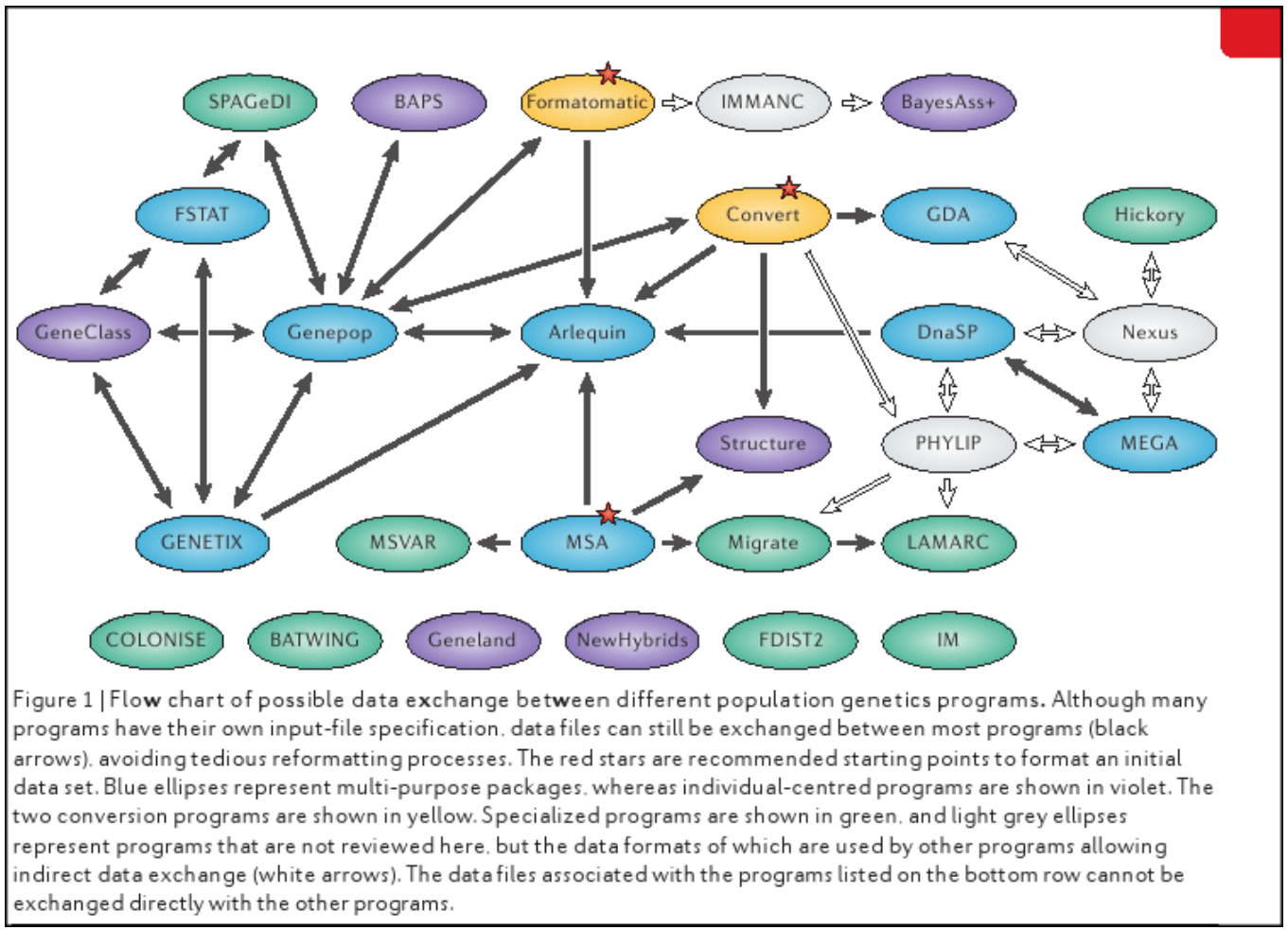


Figure 1 | Flow chart of possible data exchange between different population genetics programs. Although many programs have their own input-file specification, data files can still be exchanged between most programs (black arrows), avoiding tedious reformatting processes. The red stars are recommended starting points to format an initial data set. Blue ellipses represent multi-purpose packages, whereas individual-centred programs are shown in violet. The two conversion programs are shown in yellow. Specialized programs are shown in green, and light grey ellipses represent programs that are not reviewed here, but the data formats of which are used by other programs allowing indirect data exchange (white arrows). The data files associated with the programs listed on the bottom row cannot be exchanged directly with the other programs.

(Excoffier & Heckel 2006 *Nature Reviews Genetics*)

The scary picture

"*In a perfect world, research teams would be able to develop analysis tools to address their specific problem, but in practice they have to make their data fit the available tools, leading to obvious discrepancies between the initial goals and the results.*"

Taking genetic data into

Usual population genetic software:

- poor data interoperability
- very few multivariate methods
- not adaptable
- not programmable

The software:

- good data interoperability
- most multivariate methods
- lots of statistical tools
- several genetic/phylogenetic packages
- programming language

Taking genetic data into

Usual population genetic software:

- poor data interoperability
- very few multivariate methods
- not adaptable
- not programmable

The software:

- good data interoperability
- most multivariate methods
- lots of statistical tools
- several genetic/phylogenetic packages
- programming language

Analysing genetic data in

- *adegenet*: multivariate analyses of genetic data, simulations, spatial genetics
- *pegas*, *hierfstat*,*genetix*: classical population genetics
- *ape*, *phangorn*, *adephylo*, *picante*: phylogenetics

Analysing genetic data in

- *adegenet*: multivariate analyses of genetic data, simulations, spatial genetics
- *pegas*, *hierfstat*,*genetix*: classical population genetics
- *ape*, *phangorn*, *adephylo*, *picante*: phylogenetics

Analysing genetic data in

- *adegenet*: multivariate analyses of genetic data, simulations, spatial genetics
- *pegas*, *hierfstat*,*genetix*: classical population genetics
- *ape*, *phangorn*, *adephylo*, *picante*: phylogenetics