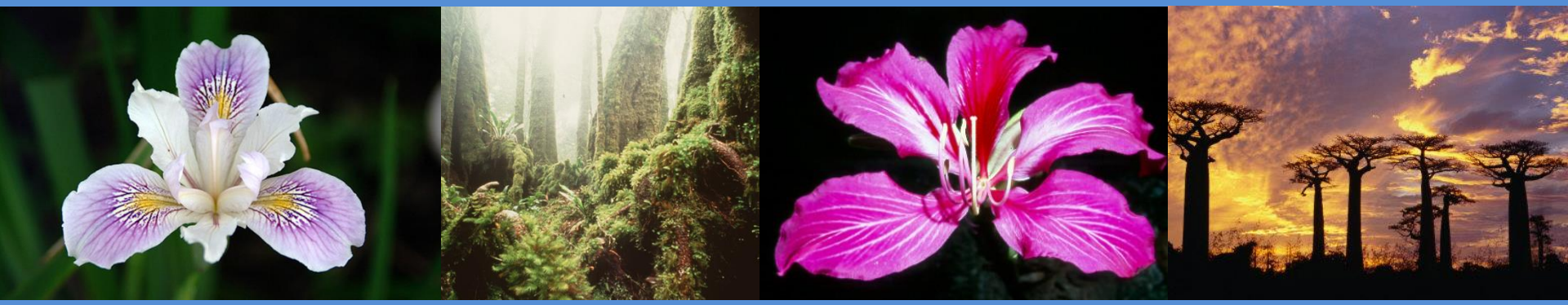
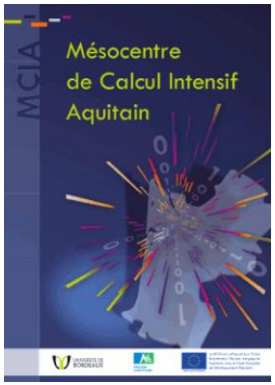


Comment l'utilisation d'Avakas et de la grille ont fait évoluer nos façons de travailler en biodiversité

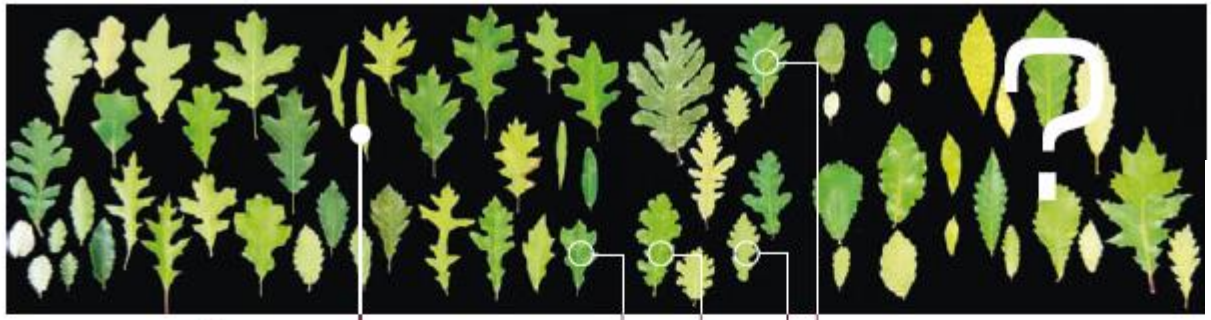
Yec'han Laizet, Philippe Chaumeil, Jean-Marc Frigerio, Stéphanie Mariette, Sophie Gerber, Alain Franc & al,
INRA, BioGeCo - Bordeaux

Journées Scientifiques
du Mésocentre
Pau, 10 février 2014

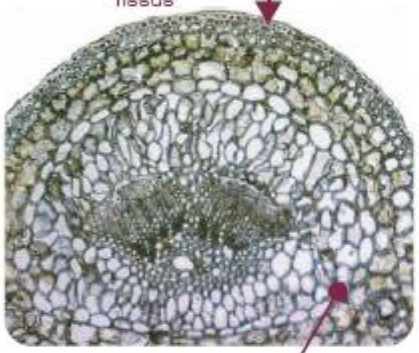


Molecular based systematics: from molecular phylogenies to metabarcoding

Ph. Chaumeil, Y. Laizet, J.-M. Frigerio, F. Rimet, A. Bouchez & Alain Franc

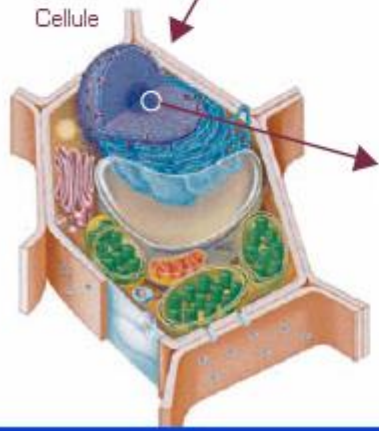


Tissus

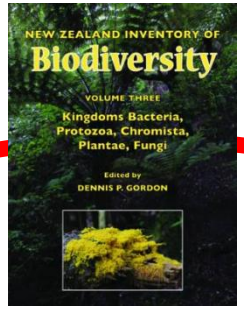


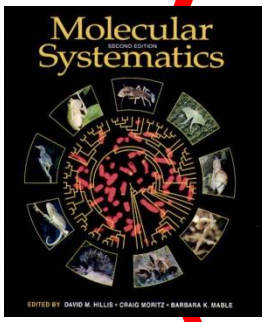
ACGTGTGCTAT ▶ *Quercus petraea*
ACGCGTGCTAT ▶ *Quercus robur*
ACGT--GCTAT ▶ *Quercus pubescens*
ACGCAGTCTAT ▶ *Quercus cocinea*

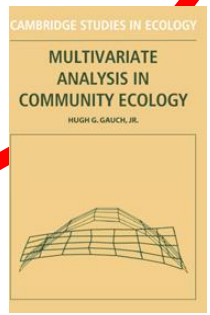
Cellule



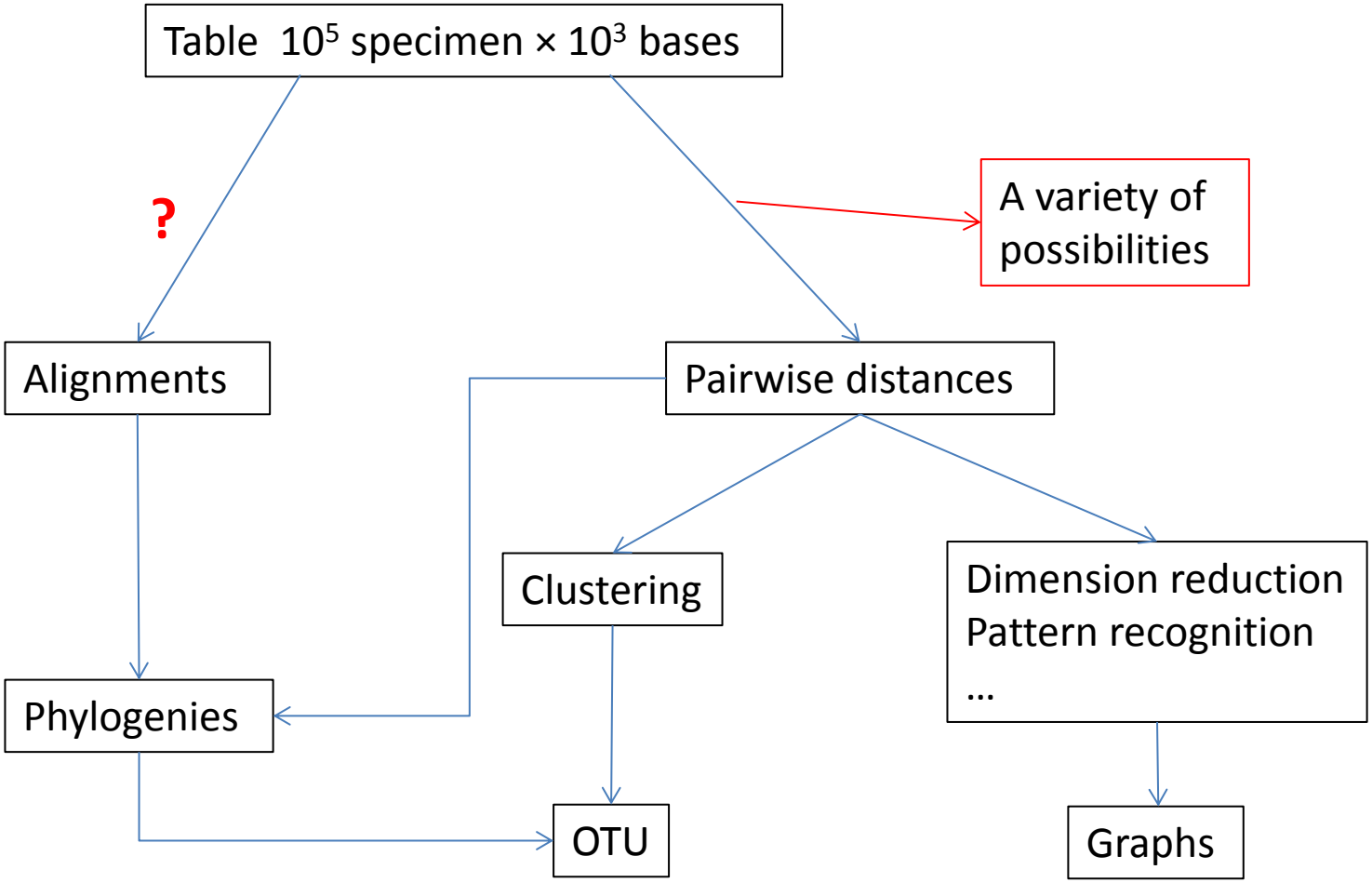
Extraction,
Amplification,
Séquençage
ADN


NEW ZEALAND INVENTORY OF
Biodiversity
VOLUME THREE
Kingdoms Bacteria, Protozoa, Chromista, Plantae, Fungi
Edited by
DENNIS P. GORDON


Molecular
Systematics
SECOND EDITION
EDITED BY DAVID M. HILLIS • CRAIG MORITZ • BARBARA K. MABLE

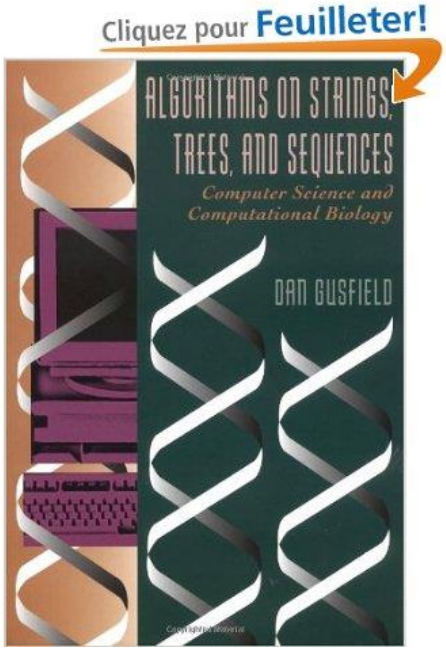

CAMBRIDGE STUDIES IN ECOLOGY
MULTIVARIATE ANALYSIS IN COMMUNITY ECOLOGY
HUGH G. GAUCH, JR.

Que faire ?
Comment connaitre l'inconnu ?



Evolution des techniques de séquençage et méthodes exactes

Technologie Sanger	Séquences de 500 bp
Technologie 454	10^5 reads de longueur 450 à 600
Technologie Illumina	10^6 reads de longueur 100 bp
Projets actuels (Tara)	10^7 reads de longueur 100 à 400



Une explosion
de la taille des jeux de données

Quels traitements ?
Quels algorithmes exacts ?
Quelles heuristiques ?

Tara @ http://oceans.taraexpeditions.org/

Taxonomy on Edit distance

Definition: The edit distance between two strings is defined as
the minimum number of edits needed to transform one string into the other,
with the allowable edit operations being
insertion, deletion, or substitution of a single character.

SOVIET PHYSICS–DOKLADY VOL. 10, NO. 8 FEBRUARY, 1966

CYBERNETICS AND CONTROL THEORY

BINARY CODES CAPABLE OF CORRECTING DELETIONS, INSERTIONS, AND REVERSALS

V. I. Levenshtein

(Presented by Academician P. S. Novikov, January 4, 1965)
Translated from Doklady Akademii Nauk SSSR, Vol. 163, No. 4,
pp. 845–848, August, 1965
Original article submitted January 2, 1965

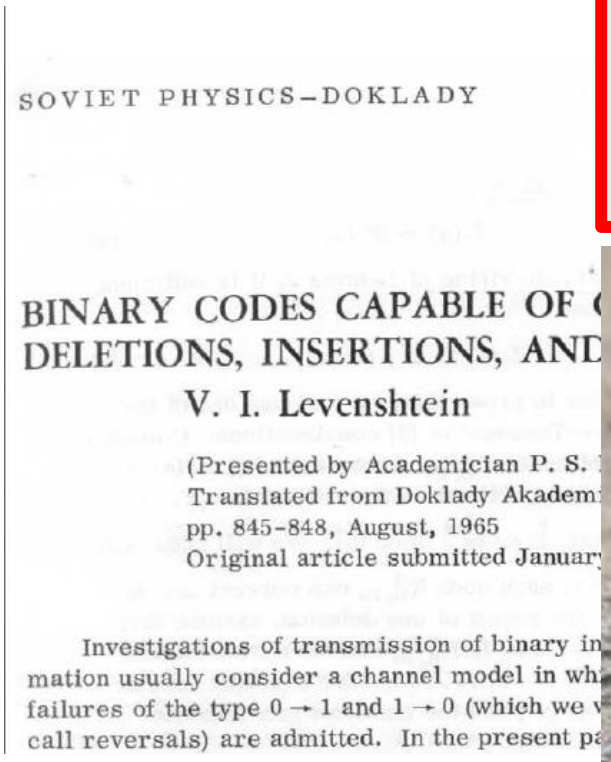
Investigations of transmission of binary information usually consider a channel model in which failures of the type $0 \rightarrow 1$ and $1 \rightarrow 0$ (which we will call reversals) are admitted. In the present paper

were inserted (deleted) from at least one of the words x or y to obtain z are deleted from (inserted into) the word z, then, as we can easily see, we obtain a word that can be obtained from both x and y

Taxonomy on Edit distance

Definition: The edit distance between two strings is defined as
the minimum number of edits needed to transform one string into the other,
with the allowable edit operations being
insertion, deletion, or substitution of a single character.

kitten → **s**itten (substitution of 'k' with 's')
sitt**e**n → sitt**i**n (substitution of 'e' with 'i')
sittin → sittin**g** (insert 'g' at the end).


SOVIET PHYSICS–DOKLADY

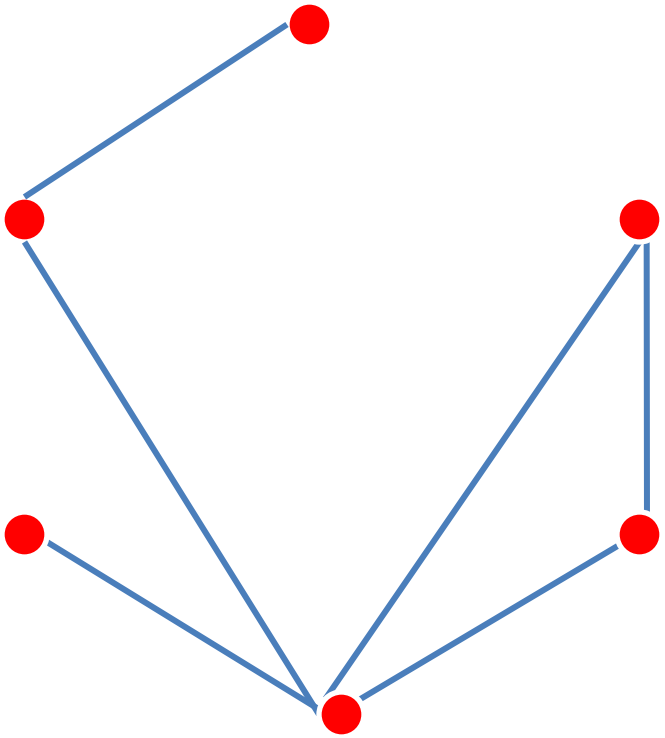
BINARY CODES CAPABLE OF
DELETIONS, INSERTIONS, AND

V. I. Levenshtein

(Presented by Academician P. S.
Translated from Doklady Akadem
pp. 845–848, August, 1965
Original article submitted Januar

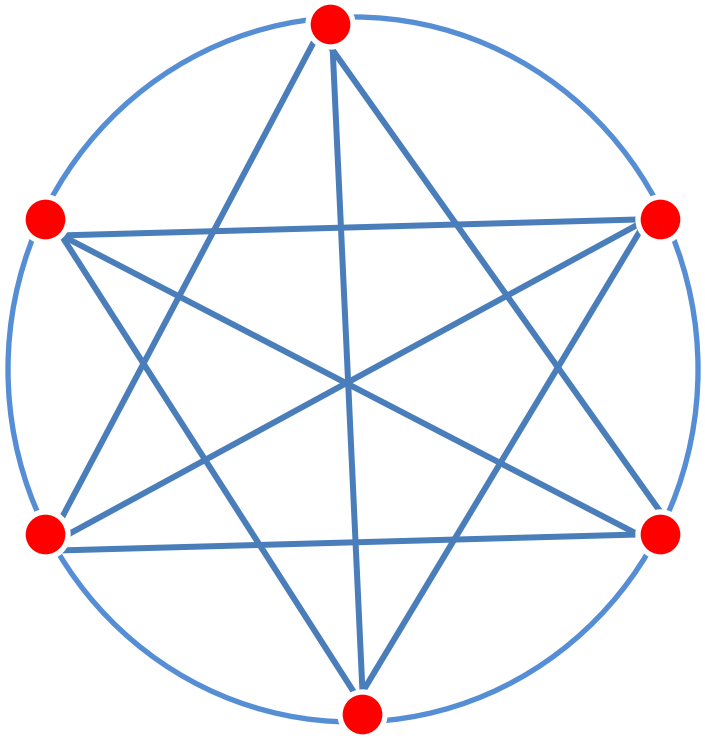
Investigations of transmission of binary in
mation usually consider a channel model in wh
failures of the type 0 → 1 and 1 → 0 (which we
call reversals) are admitted. In the present pa

Two useful notions on graphs



Connex component

Basis for BLAST
Finding: easy



Clique

Basis for Phylogenetics
Ultrametrics
Finding: hard (NP complete)

Quelques algorithmes …

Algorithme	Distance	Application
disseq	S.-W.	base/base
What'syst	Idem	Diagnostic
Read'syst	Idem	Inventaire automatique
Declic	<disseq>	Visualiser une base

Très facilement distribuable sur une grille de calcul (ou un cluster)

- Calcul matriciel
- Algorithmes sur graphe (cc, cliques, community)
- Programmation dynamique
- Mathématiques discrètes
- Visualisation

Projet e-biothon (mécenat IBM)

SYSFERA

Inria
INVENTEURS DU MONDE NUMÉRIQUE

cnrs
dépasser les frontières

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Un changement
d'échelle

Des millions de read
Calcul exact, sans heuristiques
(scores d'alignements)

Des flux de plusieurs To
par semaine

Une interface « à la Galaxy »

Puis les outils de diagnostic
et d'écologie des communautés

Un projet du e-biothon (metabarcoding)
ouvert à la communauté
d'inventaires et écologie des communautés microbiennes

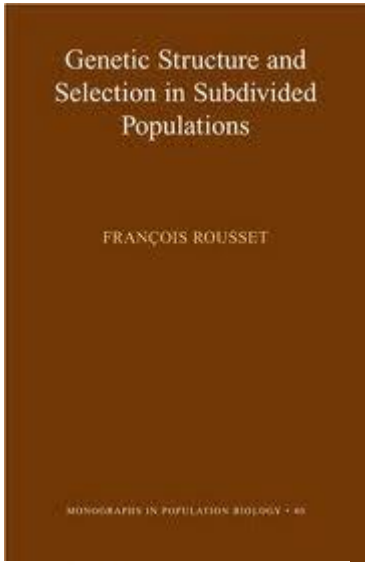
Population genetics: From one simulation to in silico numerical design

Stéphanie Mariette & al,

BioGeCo & BFP

INRA Bordeaux

Introduction to population differentiation



Natural populations are subdivided

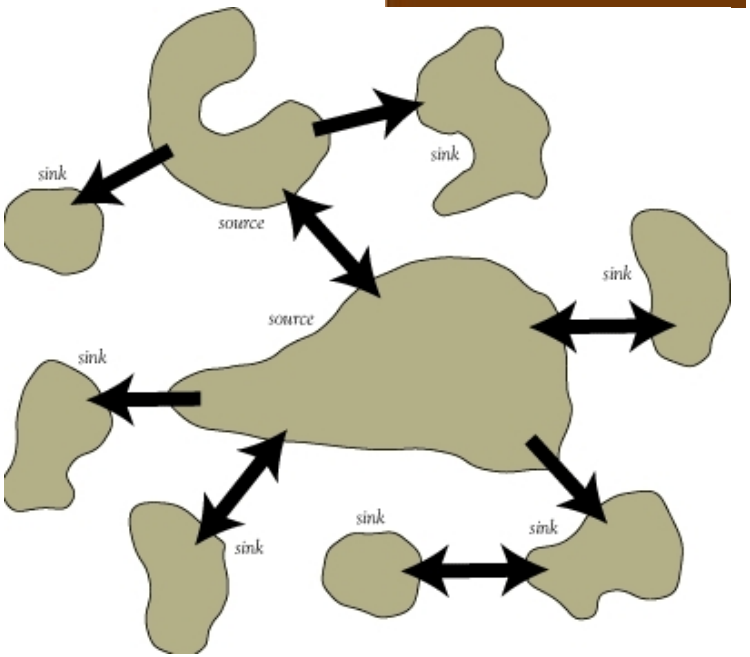
(Landscape, geography, behavior, …)

→Subdivisions (demes) maintain genetic cohesion through an exchange of reproducing migrants

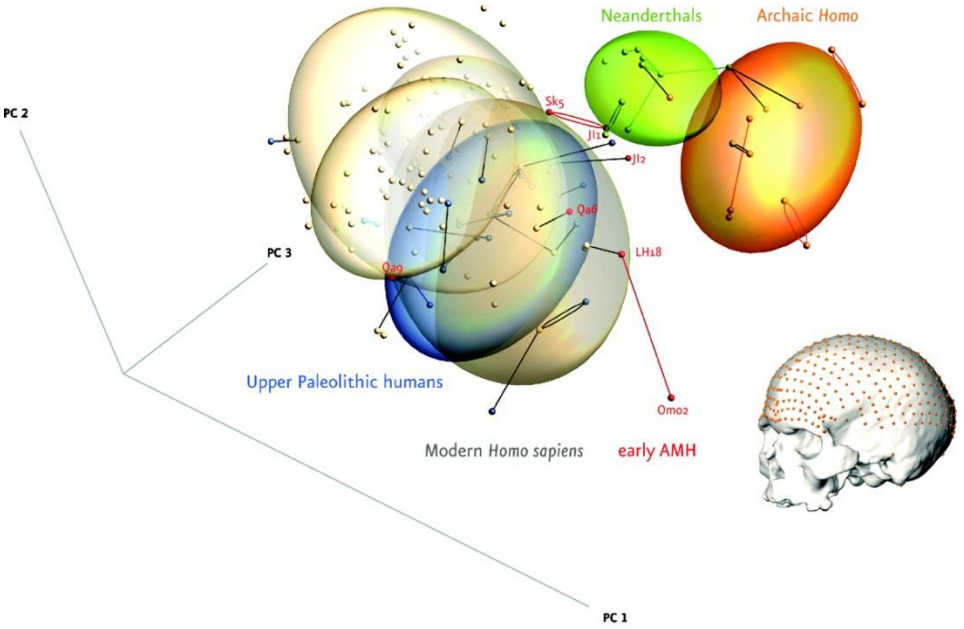
→ Pattern of migration and timing of separation between demes will have a profound effect on the degree of differentiation between demes

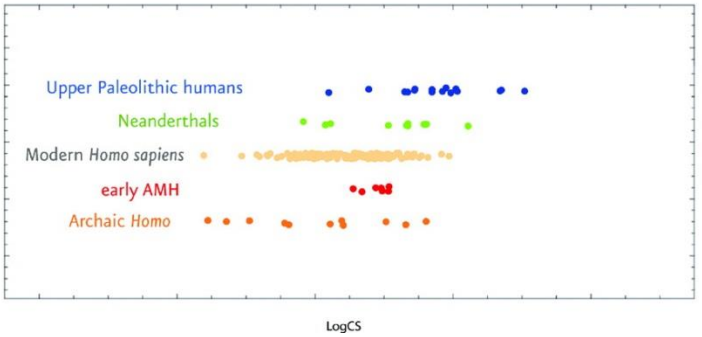
→ Subdivisions may be hierarchically arranged, at different spatial or geographical scales

(definitions after Excoffier)



Anatomically modern humans and archaic forms of Homo in shape space.

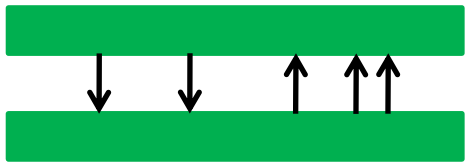




Gunz P et al. PNAS 2009;106:6094-6098

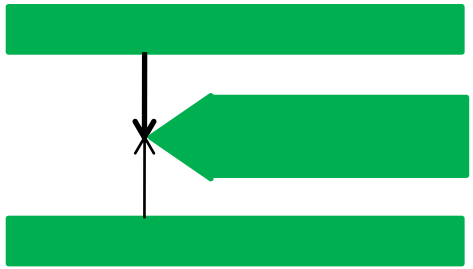
©2009 by National Academy of Sciences

Genetic evolutionary processes causing population differentiation



Restricted gene flow and genetic drift

Population splits and isolation

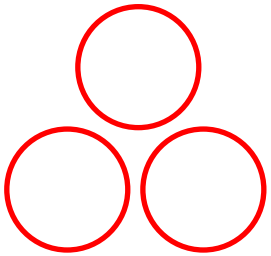


Admixture

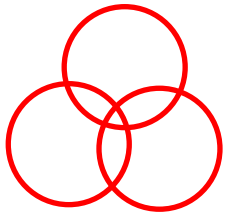
Possible interaction with selection:

➡ creates or diminish population differentiation

Continuum of population differentiation

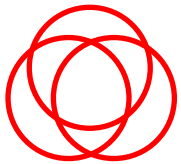


Complete independence

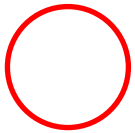


Modest connectivity

Pattern recognition …



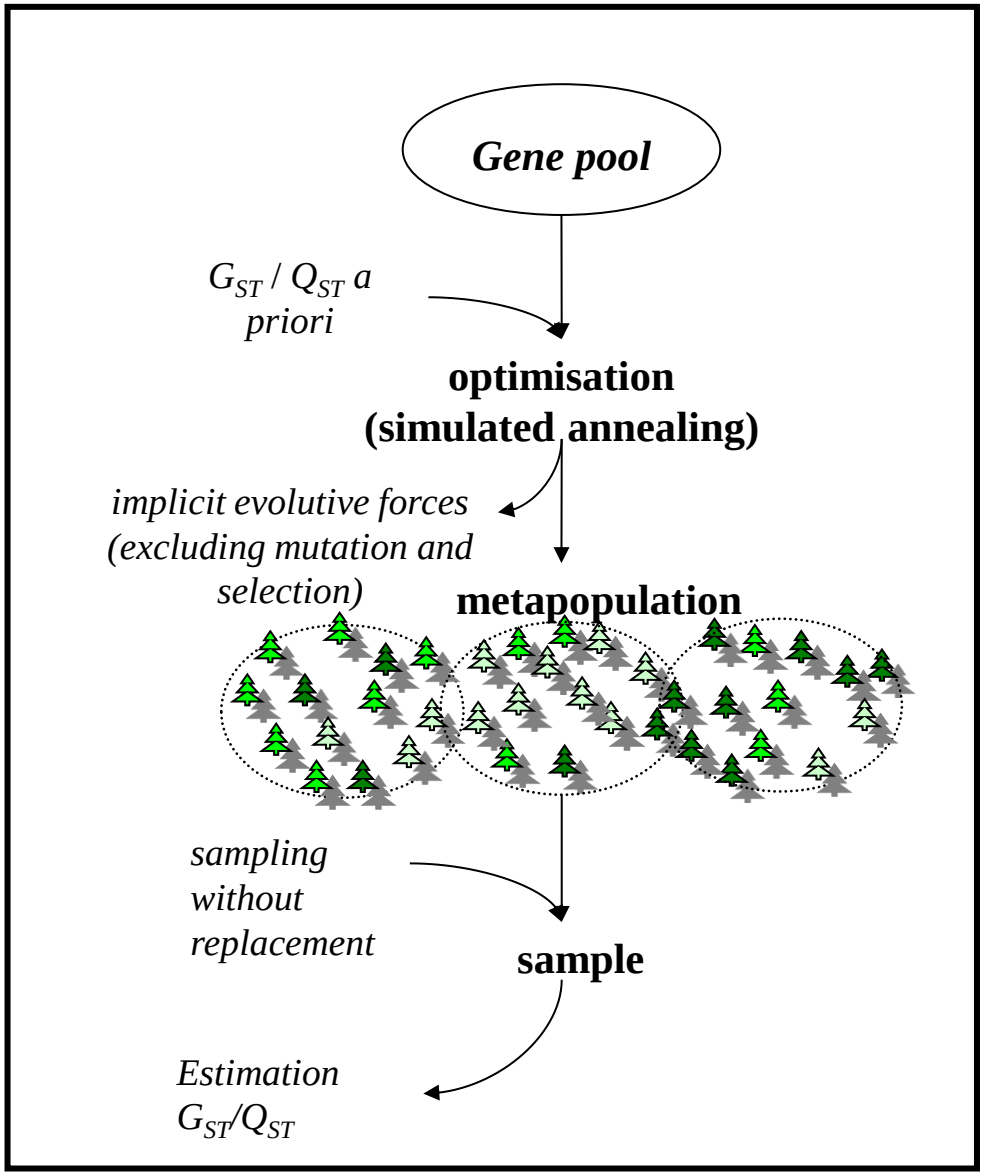
Substantial connectivity



Panmixia (subpopulations are completely congruent)

After Waples and Gaggiotti, 2006, Molecular Ecology

Metatrom model (collaboration with L. Sanchez INRA Orléans)



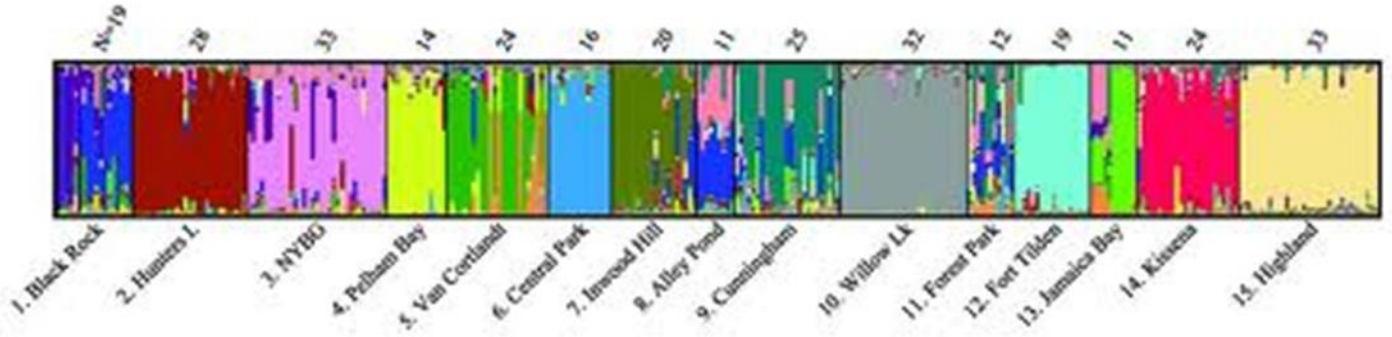
Pop. number between 10 and 50
Ind. number=200
Locus number between 20 and 400

Use of Avakas to test several sampling scenario (but not parallel)

One run: intensive
Distribution: Sampling, Bootstrap

Objective & intuitions of the Structure method

- Identify clusters of individuals, who are genetically more similar to each other than to ones from another cluster
- Allele frequencies differ among reproductively isolated populations
- Perfectly informative marker: an allele that occurs in only one population
- Bayesian method



Genetic ancestry analyses

- **Structure** software (*Quercus* sp., *Prunus* sp.)
- **Frappe** software (Tang's lab at http://med.stanford.edu/tanglab/software/frappe.html)

→ Not parallel analyses

→ But see a R package to run STRUCTURE in parallel (ParallelStructure) – a Python package for fast-Structure

One run: intensive
Distribution: Sampling, Bootstrap

Tools

search tools

Upload Your Data

ADDED TOOLS

MIAB

MAFFT alignment (basic options)

BMGE block selection

Fasta to Phylip conversion

Fasta to Nexus conversion

PHYML wrapper (maximum likelihood phylogenies)

Draw ascii tree (newick)

BEAST wrapper

DECLIC

Disseq

Declic Graph

Declic MDS

DIAGNO-SYST (HIDDEN TOOLS)

Diagno-Syst char format
Formats charfile

Diagno-Syst

Disseq reads to taxo

Taxonomic inventory from NGS reads

The Genome Transcriptome Facility of Bordeaux Pierroton is pleased to provide a Galaxy server to its users.

The service is provided AS IS WITHOUT ANY WARRANTY.

For questions, requests or technical issues, please contact us at contact.pgtp@pierroton.inra.fr .

Quota: 50GB (registred). For additional disk space, contact us

Deleted datasets and libraries will be automatically purged after 30 days.

Upload files > 2Go via SFTP

Upload your file via your favourite FTP utility (filezilla, ...)

- host: sftp://galaxy-pgtp.pierroton.inra.fr
- login: email address you used to register at this Galaxy server

History

Unnamed history

156.3 KB

3: Nitzschia_char.txt

2: Nitzschia_rbcL_dissw.txt

1: Nitzschia_rbcL.fas