Scaling up phylogenetic networks to genome-size data

Co-supervisor
70 %

Celine Scornavacca
CR2, ISE-M lab

Supervisor
30 %

Vincent Berry
Prof, LIRMM lab

Université Montpellier
Institut de Biologie Computationnelle (IBC)

Phylogenetic trees

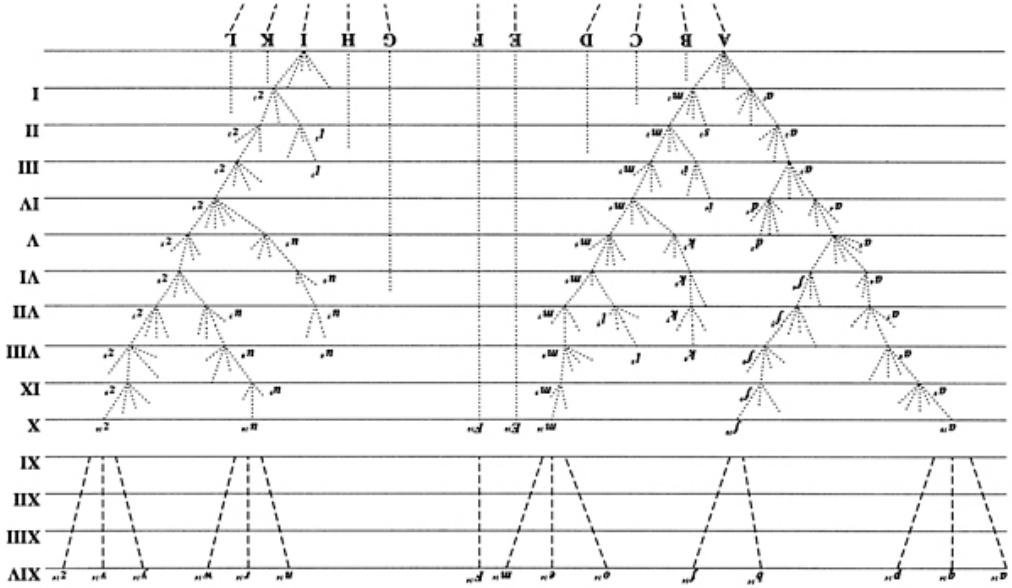
Rooted phylogenetic trees are used to depict the evolutionary history of a set of taxa, whose internal nodes represent speciation events.

out-branching trees with no indegree-1 outdegreee-1 nodes and whose leaves are each associated to a species or gene (taxa)

Phylogenetic trees

Rooted phylogenetic trees are used to depict the evolutionary history of a set of taxa, whose internal nodes represent speciation events.
But … Darwin described evolution as 'descent with modification'

(does not necessarily imply a tree representation…)



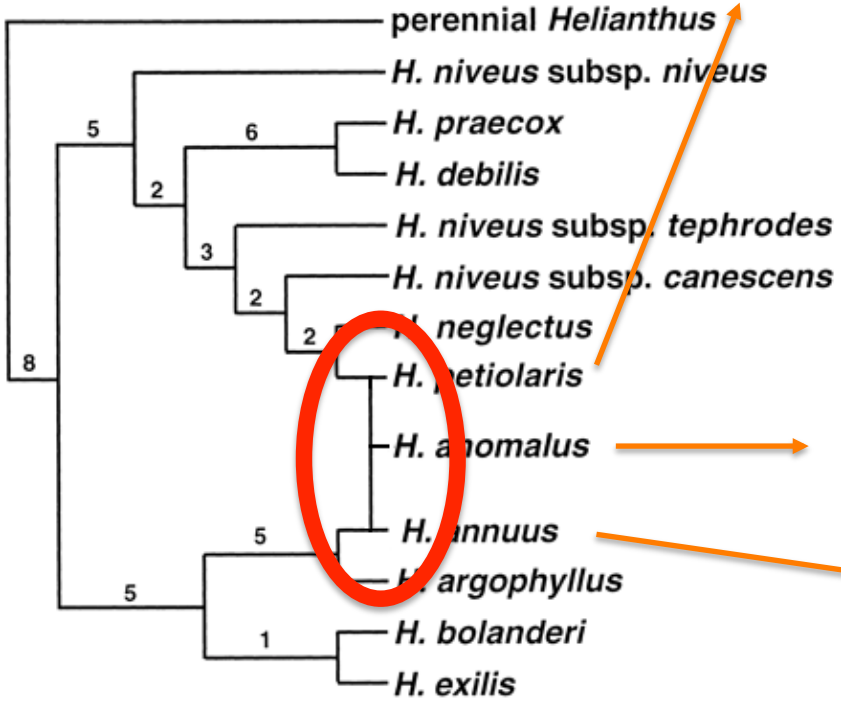
The Origin of Species (1859)

The implicit assumption of using trees is that, at a macroevolutionary scale, each (current or extinct) species or gene *only descends from one ancestor*

Reticulate evolution

However, at a larger scale, genomes sometimes inherit from multiple ancestors, because of reticulate events, e.g:

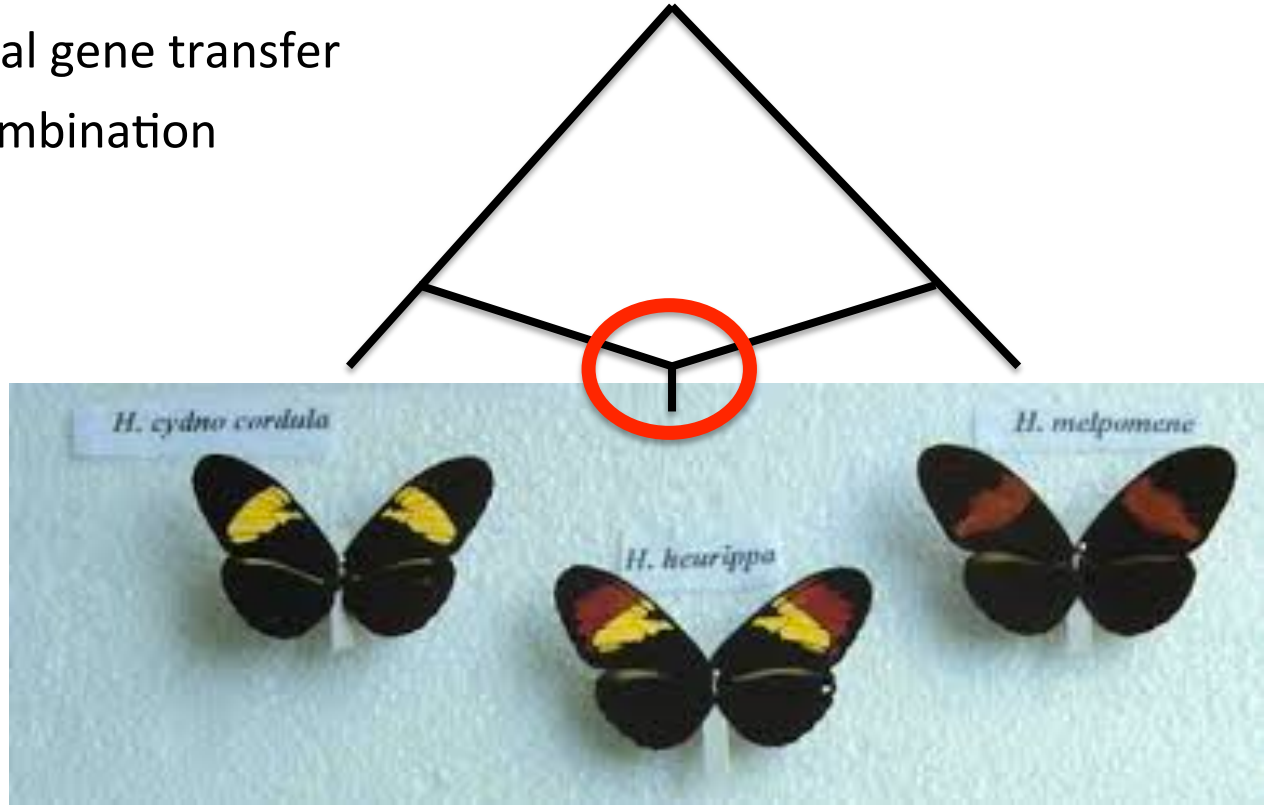
1) **Hybrid speciation**
2) Lateral gene transfer
3) Recombination



Reticulate evolution

However, at a larger scale, genomes sometimes inherit from multiple ancestors, because of reticulate events, e.g:

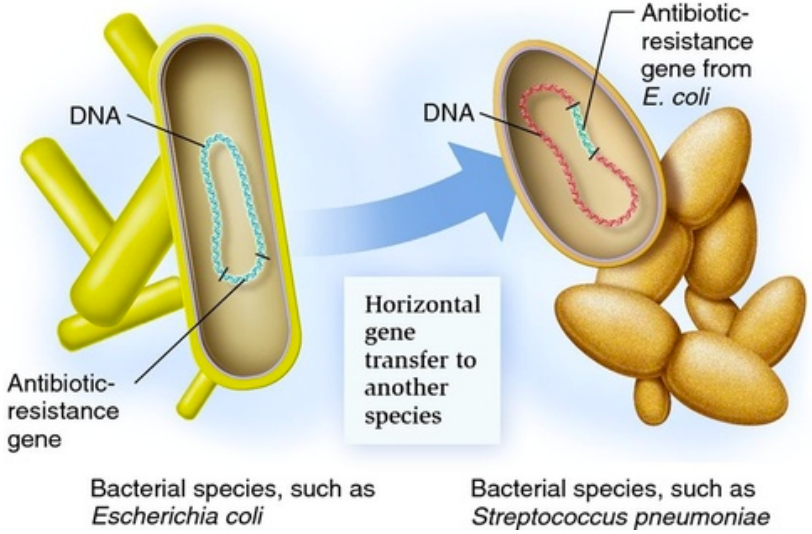
1) **Hybrid speciation**
2) Lateral gene transfer
3) Recombination

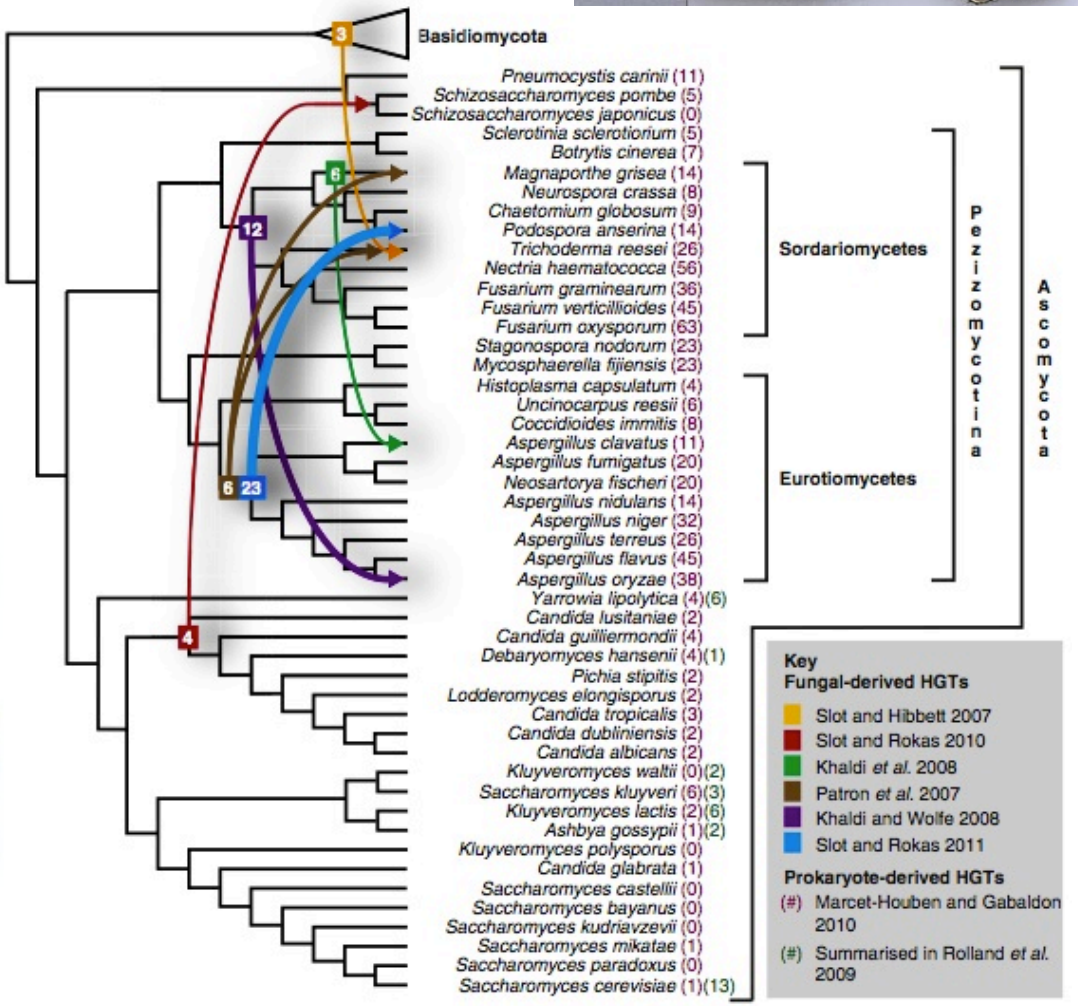


Reticulate evolution

However, at a larger scale, genomes sometimes inherit from multiple ancestors, because of reticulate events, e.g:

1) Hybrid speciation
2) **Lateral gene transfer**
3) Recombination

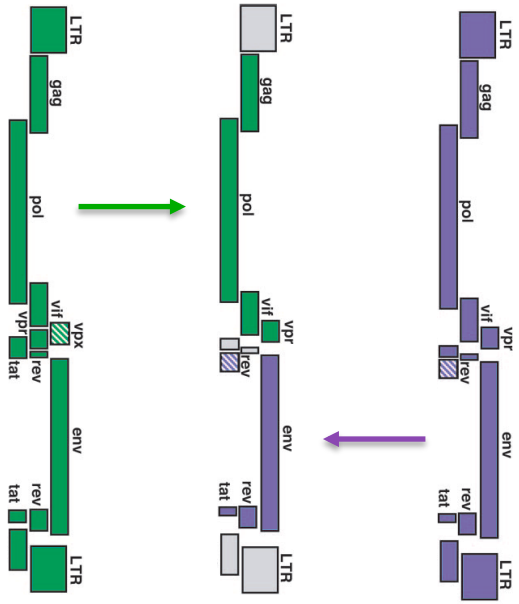




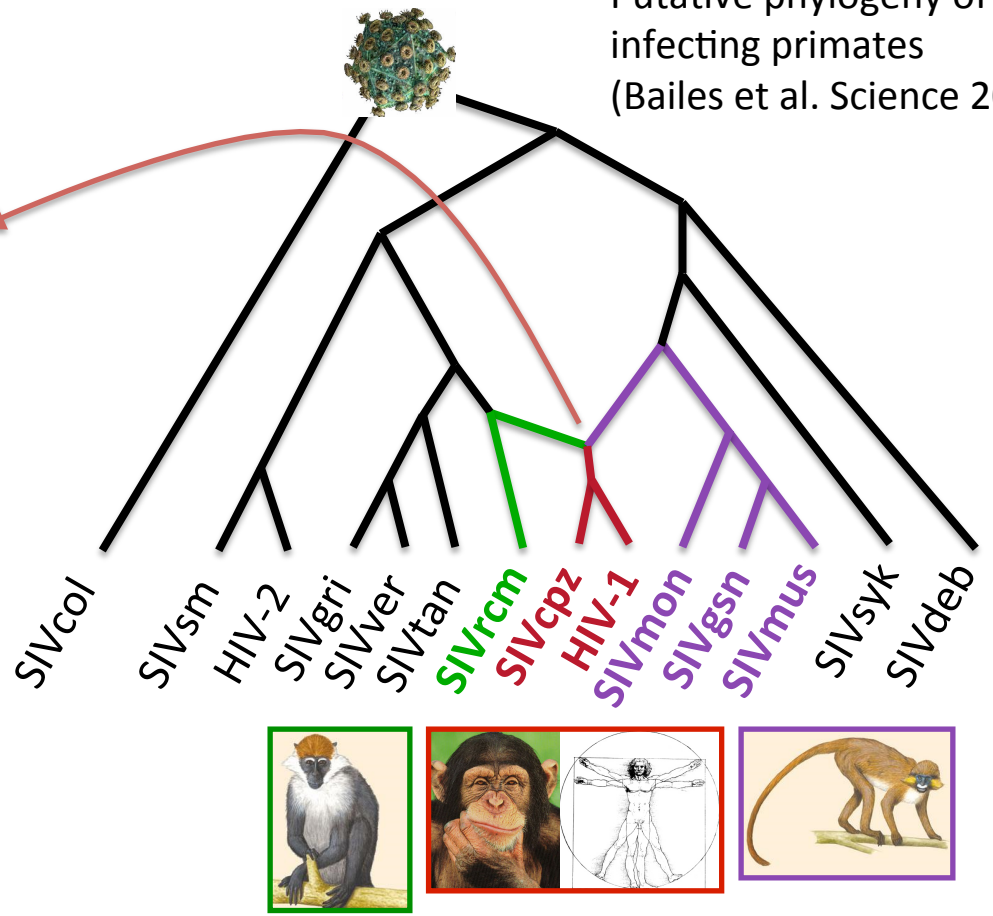
Reticulate evolution

However, at a larger scale, genomes sometimes inherit from multiple ancestors, because of reticulate events, e.g:

1) Hybrid speciation
2) Lateral gene transfer
3) **Recombination**



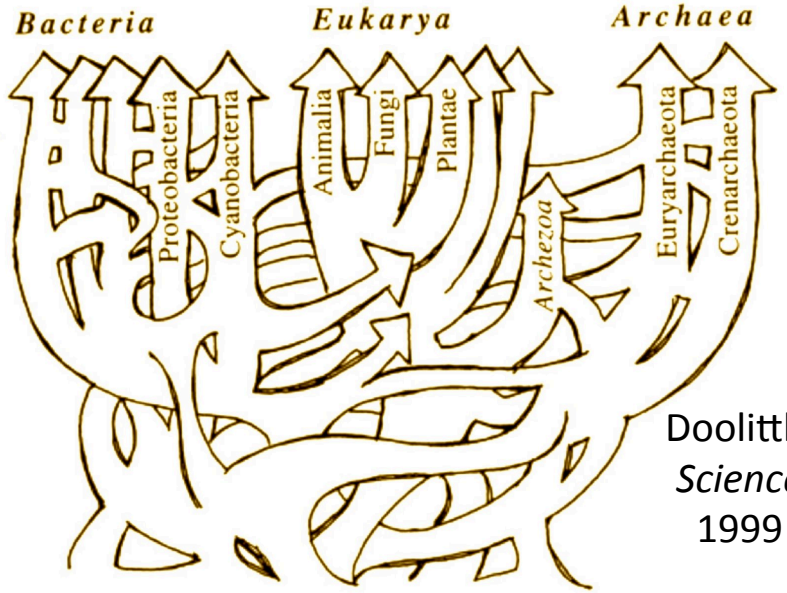
Putative phylogeny of HIV/SIV infecting primates (Bailes et al. Science 2003)



Phylogenetic networks

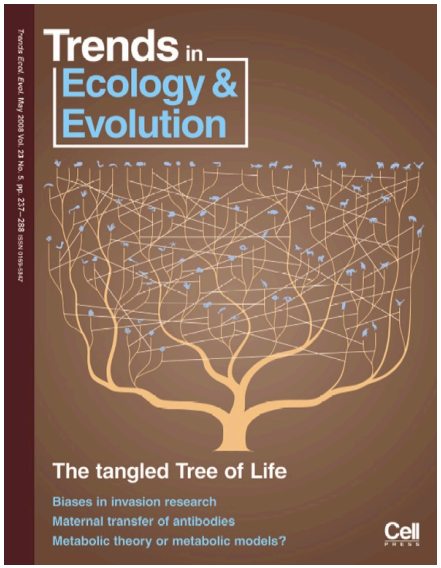
In the presence of reticulate events, phylogenies are networks (DAGs), not trees

The study of phylogenetic networks is a recent interdisciplinary field: maths, CS, biology…

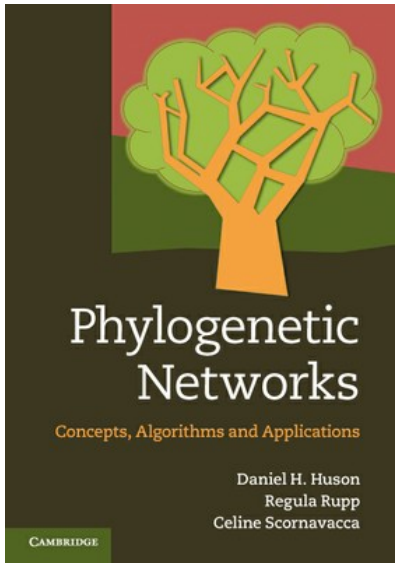


Doolittle
Science
1999

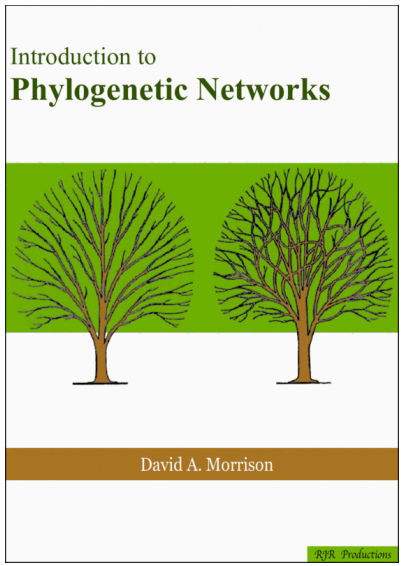
2008



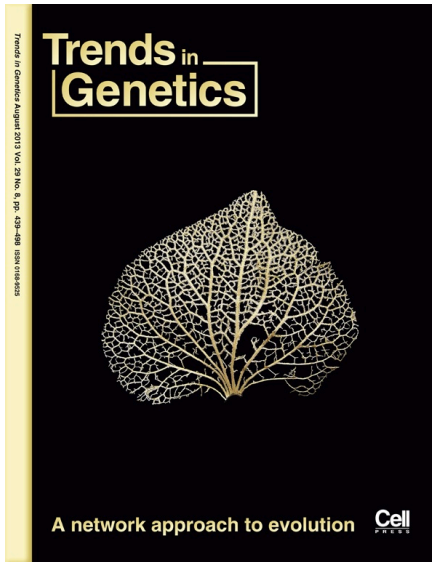
2010



2011

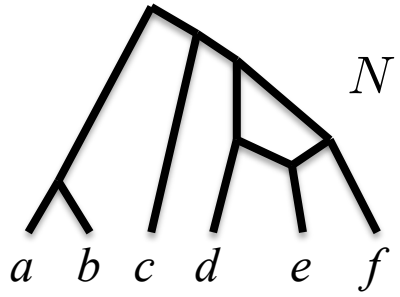


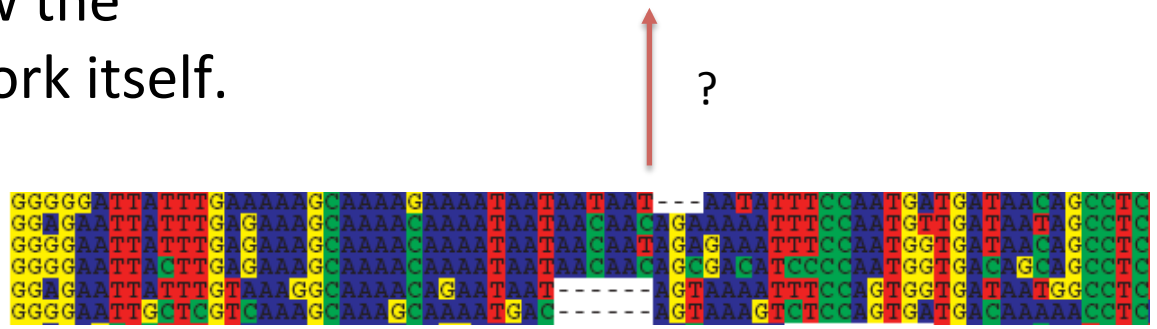
2013



Phylogenetic network inference

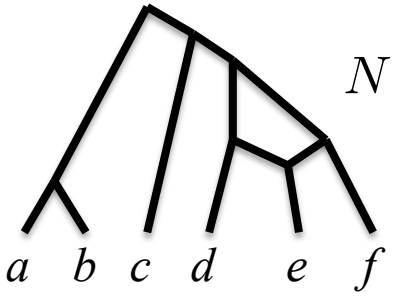
The phylogenetic network community considers that the ratio data/reticulation events is ‘large enough’ to allow the inference of the network itself.

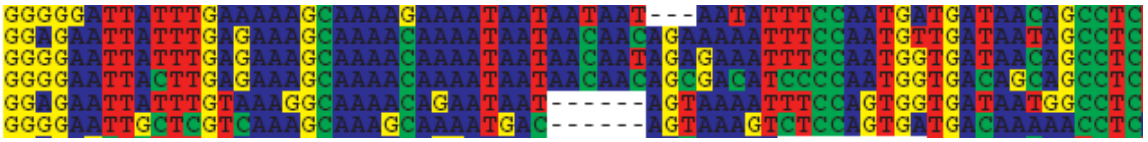




Phylogenetic network inference

The phylogenetic network community considers that the ratio data/reticulation events is 'large enough' to allow the inference of the network itself.



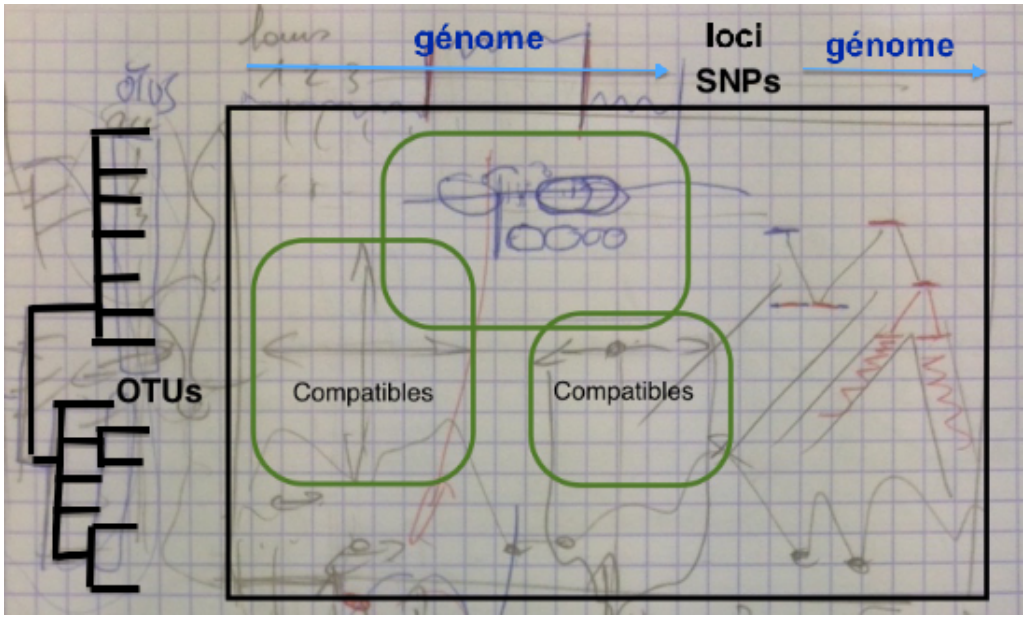


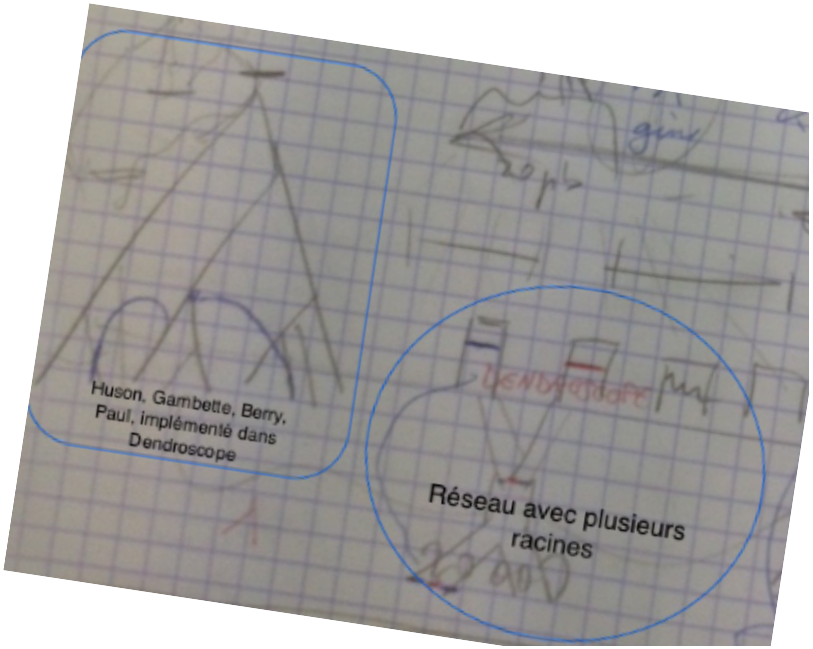
However, most biological literature still uses trees even when a network would be more suitable because network methods developed so far

- **do not yet take the full biological complexity into account**, and
- **do not scale up to genomic data** (based on optimization problems that are computationally hard, often even to approximate).

Training period at ISEM/LIRMM

- Phylogenetic networks have been intensively studied from a mathematical and computational perspective in the last years; the bibliographic part of the stage will thus focus on the literature on phylogenetic networks published since the appearance of [1], and will lead to the production of a report updating the survey provided in [Scornavacca et al 2012]
- A possible approach is to identify subsets of varieties that encompass maximum diversity and for which largest regions of consecutive loci in the genome have a tree-like evolutionary history. These ground trees will then serve to compute a phylogenetic network [1] representing hybridization events through which these trees were obtained from the initial founders.





PhD Thesis plan

1. **Speed up the phylogenetic networks reconstruction**:
   - Design **models** taking more into account the biological complexity : duplication, loss, transfers in gene family evolution ; syntenies in genome architecture ; regulatory networks ; …
   - More factors -> **reduced combinatorics**
   - Leverage these new features to design **algorithms** with reasonable running times.
2. **Obtain a realistic picture of ancestral genomes'** composition (which genes, on which chromosomes,...) **for ancestral species involved in reticulate evolution:**
   - Extend available methods designed for trees to networks, while limiting the combinatorial explosion
3. **Apply methods** on **plant real data** (Oryza, Banana, …) to explain the composition of current genomes through large-scale evolutionary events (duplications or losses of chromosome fragments).

In practice

- PhD thesis on **combinatorial algorithms** (combinatorial modelization, graphs, parameterized complexity, approximation algorithms), with an **applicative side** (programming, real data analysis)
- Hosted inside the ANCESTROME (ANR) & *GenomeHarvest* (Agropolis fondation) projects

Scaling up phylogenetic networks to genome-size data

Co-supervisor
70 %
Celine Scornavacca
CR2 ISE-M

Supervisor
30 %
Vincent Berry
Prof. LIRMM

Recent publications

- **2015** Inferring gene duplications, transfers and losses can be done in a discrete framework
V. Ranwez, S. **Scornavacca**, J.-P. Doyon, V. **Berry, *Journal of Mathematical Biology***
- **2015** Thu-Hien To, Edwin Jacox, Vincent Ranwez, Celine **Scornavacca**. A Fast Method for Calculating Reliable Event Supports in Tree Reconciliations via Pareto optimality**. BMC Bioinformatics**
- **2015** Mareike Fischer, Leo van Iersel, Steven Kelk, and Celine **Scornavacca**. On Computing the Maximum Parsimony Score of a Phylogenetic Network**. SIAM Journal on Discrete Mathematics** 29, no. 1, pp. 559-585.
- **2015** Fabio Pardi, Celine **Scornavacca**. Reconstructible phylogenetic networks: do not distinguish the indistinguishable. **PLOS Computational Biology.** (11), no 4:e1004135. doi:10.1371/journal.pcbi.1004135.
- **2015** Thu-Hien To and Celine **Scornavacca**. Efficient algorithms for reconciling gene trees and species networks via duplication and loss events. **BMC genomics** 2015(16) no. 10, S6
- **2015** Katharina Huber, Leo van Iersel, Vincent Moulton, Celine **Scornavacca**, Taoyang Wu. Reconstructing phylogenetic level-1 networks from nondense binet and trinet sets. **Algorithmica**
- **2015** Ancestral gene synteny reconstruction improves extant species scaffolding
Y. Anselmetti, V. **Berry**, C. Chauve, A. Chateau, E. Tannier and S. Berard *BMC Genomics, 16 Suppl 10:S11*,
- **2015** François Chevenet, Jean-Philippe Doyon, Celine **Scornavacca**, Emmanuelle Jousselin, Vincent **Berry**. SylvX: a viewer for phylogenetic tree reconciliations. **Bioinformatics**
- **2014** Yao-ban Chan, Vincent Ranwez, Celine Scornavacca. Exploring the space of gene/species reconciliations with transfers. **Journal of Mathematical Biology**
- **2014** Celine Scornavacca, Edwin Jacox and Gergely Szöllősi. *Joint amalgamation of most parsimonious reconciled gene trees*. **Bioinformatics**
- **2013** Support Measures to Estimate the Reliability of Evolutionary Events Predicted by Reconciliation Methods,
Nguyen T-H, Ranwez V, **Berry** V, **Scornavacca** C, *PLoS ONE* 8(10)
- **2013** Reconciliation and local gene tree rearrangement can be of mutual profit.
T.H. Nguyen, J.-P. Doyon, S. Pointet, A.-M. Arigon Chifolleau, V. Ranwez, V. **Berry, Algorithms for Molecular Biology**, 8:12.
- **2013** Representing a set of reconciliations in a compact way
C. **Scornavacca**, V. **Berry**, V. Ranwez, *Journal of Bioinformatics and Computational Biology* Vol. 11, No. 2