Phylogenetic Comparative Methods

M. Mariadassou and H. Chiapello
with many slides courtesy of P. Bastide and F. Cerutti

MAIAGE, INRA,Université Paris-Saclay, 78350 Jouy-en-Josas, France

Maths for Genomics
7th March 2018

1 Motivation
- A Simple Example
- Another Example
- One Last Example (courtesy of P. Bastide)
- Take Home Message

2 Continuous Characters
- Brownian Motion
- Multivariate Brownian Motion
- Phylogenetic Correlation/Regression
- To Brownian Motion and Beyond

3 Discrete Characters
- Univariate Models
- Multivariate Characters

4 Summary

Outline

1. Motivation
   - A Simple Example
   - Another Example
   - One Last Example (courtesy of P. Bastide)
   - Take Home Message
2. Continuous Characters
3. Discrete Characters
4. Summary

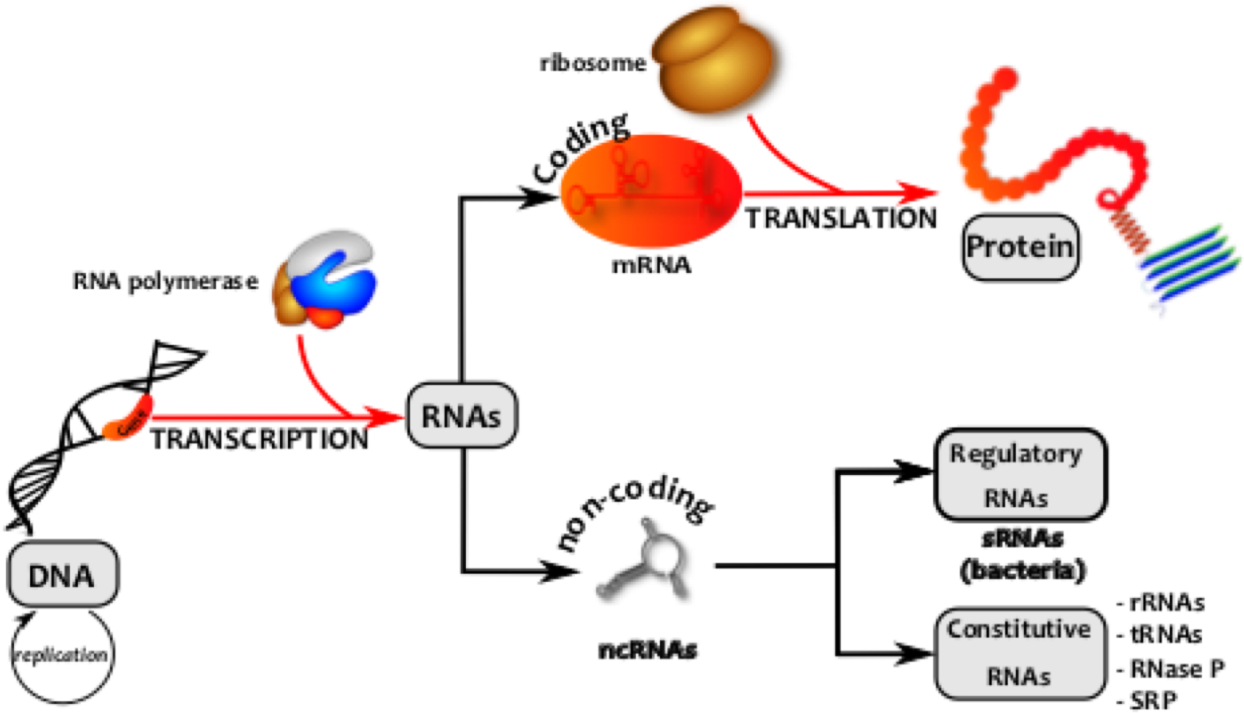
Motivation

A simple example: detection of coevolution between small regulatory RNAs and coding genes in a bacterial genus

With many slides/figures courtesy of F. Cerutti

Small regulatory RNAs

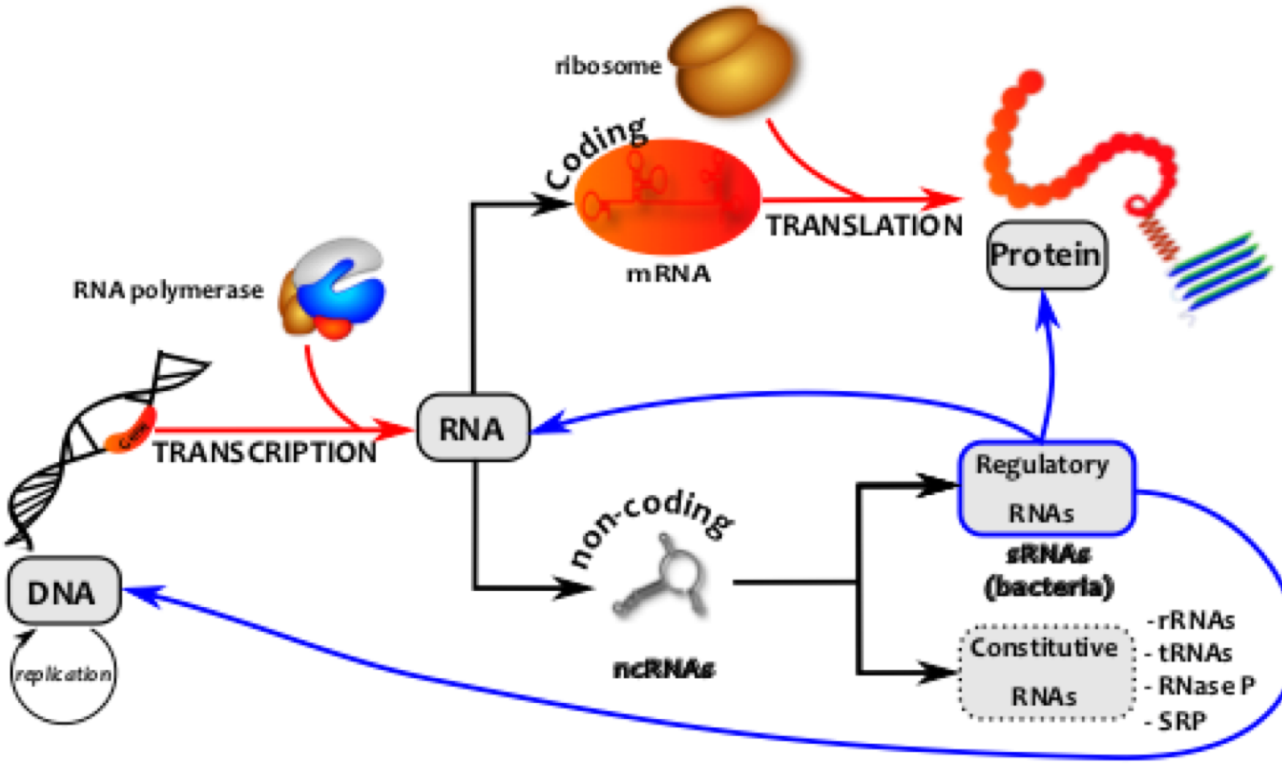
- Small **functional** RNAs, **transcripted** and generally **untranslated**
- **Widespread in all kingdoms**



Source: F. Cerutti PhD thesis

Small regulatory RNAs

- Small **functional** RNAs**, transcripted** and generally **untranslated**
- **Widespread in all kingdoms**

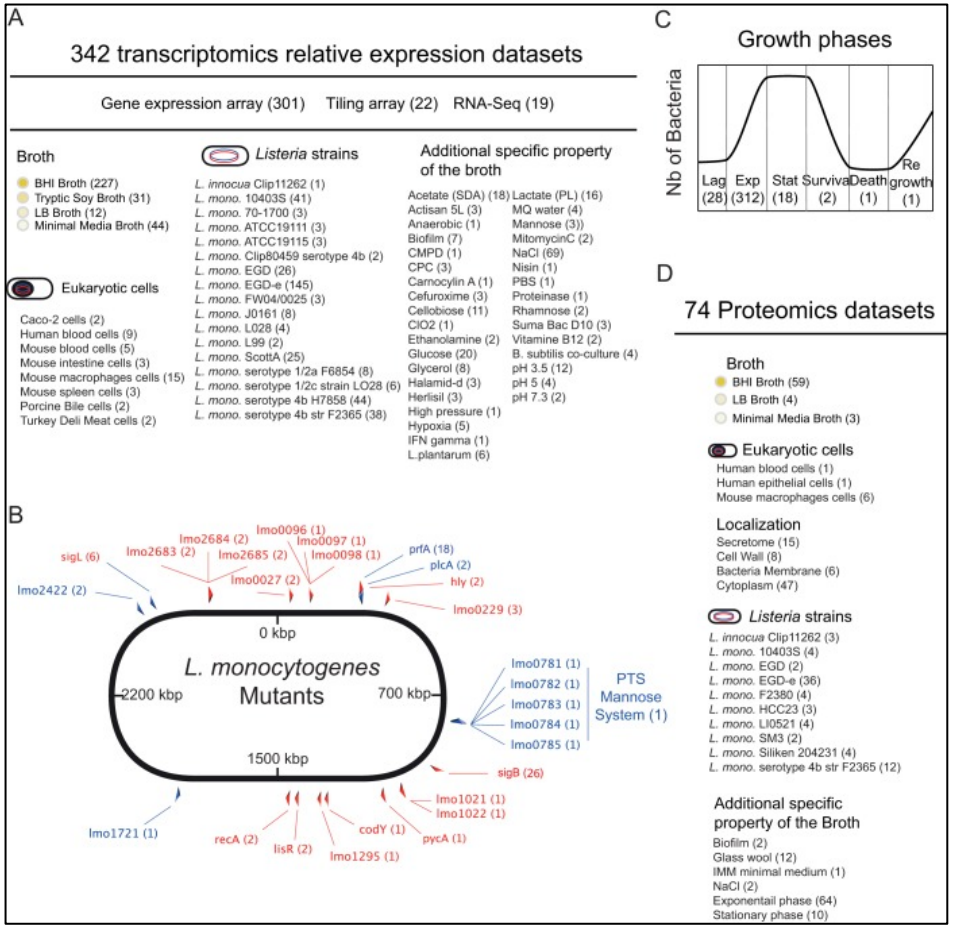


Source: F. Cerutti PhD thesis

sRNAs in bacteria

Continuous accumulation of **genomics and transcriptomics assays** (RNAseq, tilling array,..)

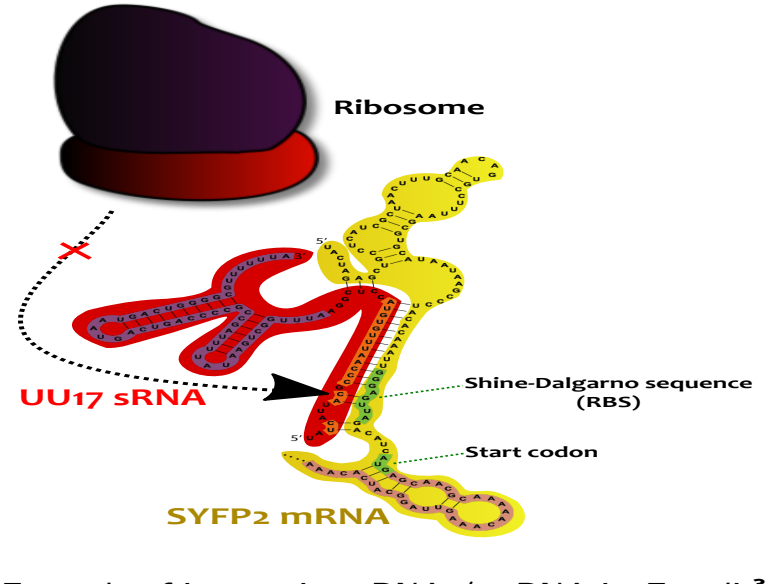
- **Increasing amount of non-coding regulatory RNAs** identified in several bacteria
- Size: 50-500 nt
- Annotated on a few strains
- **Many sRNAs with unknown functions**



Transcriptomic and proteomic data sets available in the Listeriomics database https://listeriomics.pasteur.fr
Becavin et al. mSystems. 2017

Key regulators of gene expression

- **Adaptive response relative to environmental changes** (stress response, quorum sensing, pathogenicity…)
- **Generally post-transcriptional action** (a sRNA acts by interacting with 5'/3'UTR region of a mRNA)
- **Mechanisms are not yet fully understood**

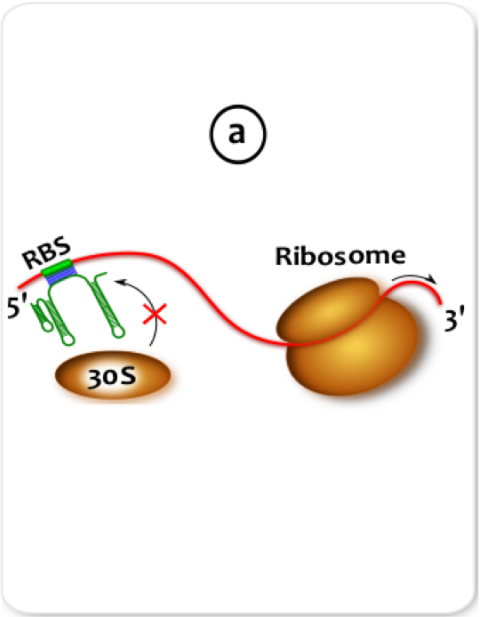


Example of a sRNA/mRNA interaction in *E. coli*
Source : http://2012.igem.org

Peer A and Margalit H, J. Bacteriol., 20112
Skippington E and Ragan MA, Genome Biol. Evol., 2012

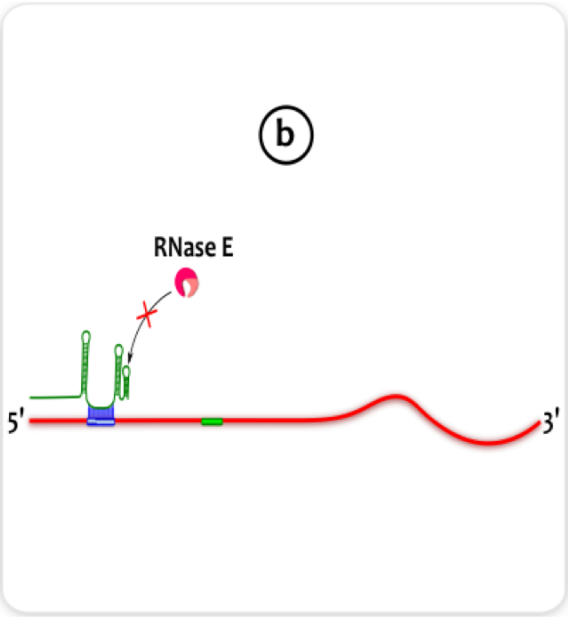
Examples of sRNA-mRNA interaction mechanisms

Masking the RBS



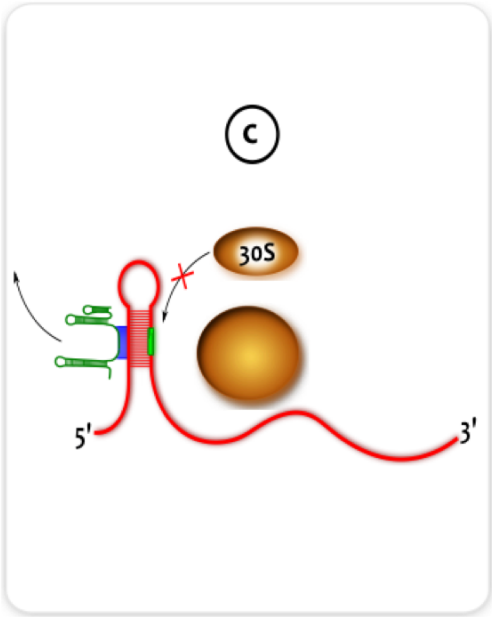
Inhibition of mRNA translation

Hiding a Rnase specific site



Stabilization of the mRNA

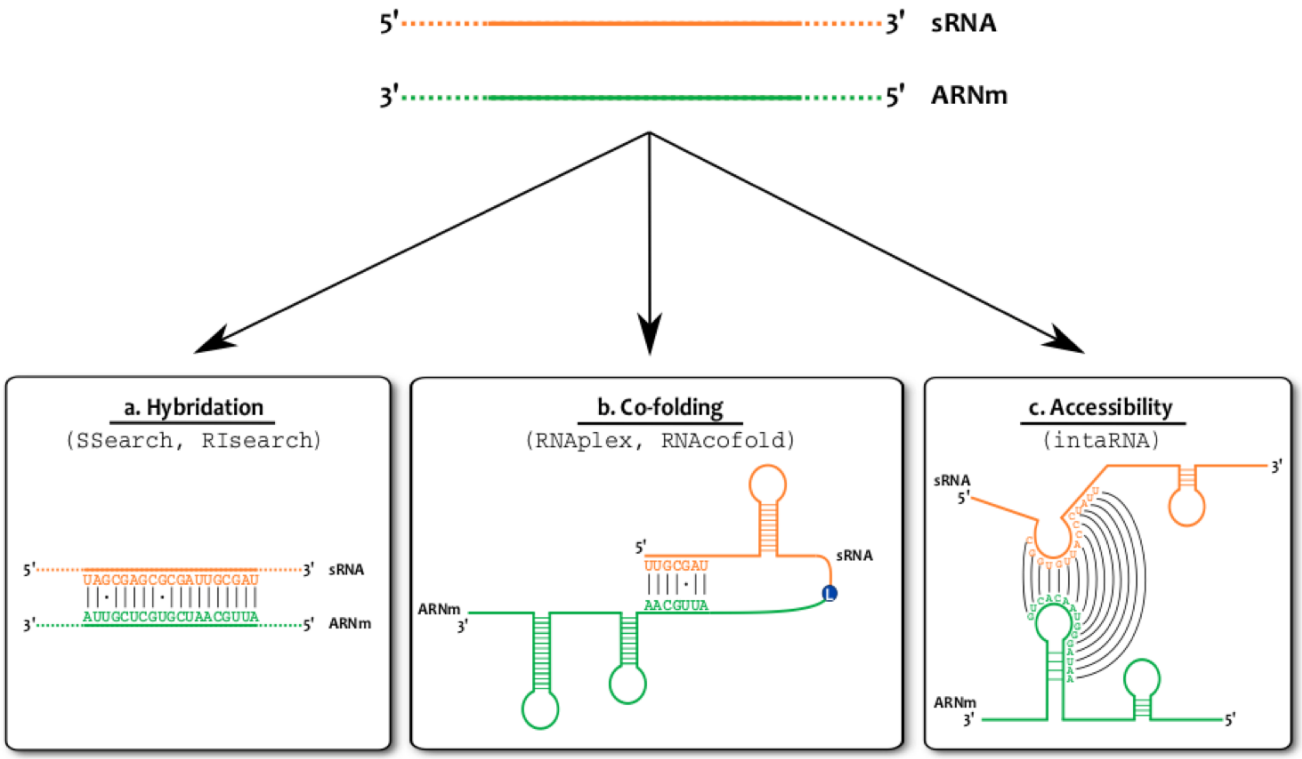
Release on a RBS trapped in a secondary structure element



Activation of the mRNA

Source: F. Cerutti PhD thesis

The 3 classes of methods to predict sRNA-mRNA pairs



Search for long streches of complementary between sRNA and mRNA

Search for optimal co-folding of sRNA and mRNA

Search for accessibility sites between sRNA and mRNA

Source: F. Cerutti PhD thesis

But mRNA targets are difficult to predict

- Prediction tools often yield a prohibitive number of candidates
- Lack of biological knowledge regarding the rules governing sRNA–mRNA interactions
- Length of sRNA-mRNA interacting region: 20 to 5 pb!

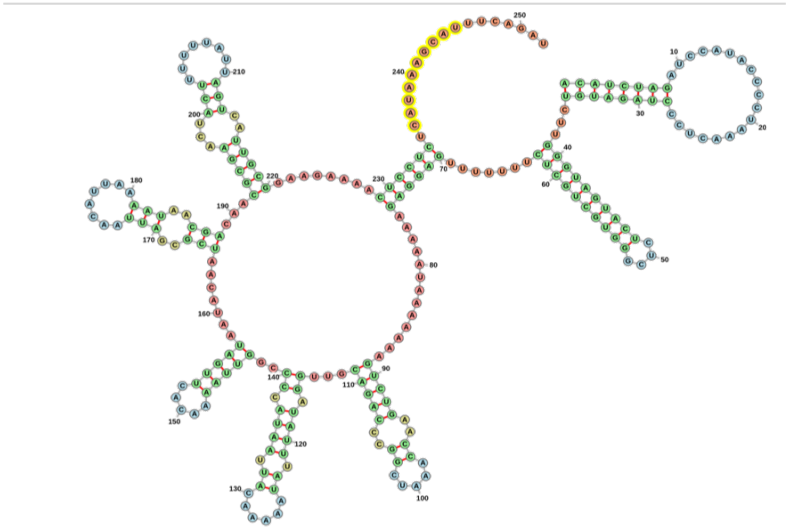
The integration of additional biological information can help to filter the results

Thébault et al. Brief Bioinformatics., 2015

Regulatory sRNAs in bacteria

Available annotated genomes and sRNA libraries allow to perform evolutionary studies, but:

- **sRNA content only available for a few reference strains**
- **Few studies** on sRNA evolution (in Gram - bacteria)
- No study on sRNA coevolution

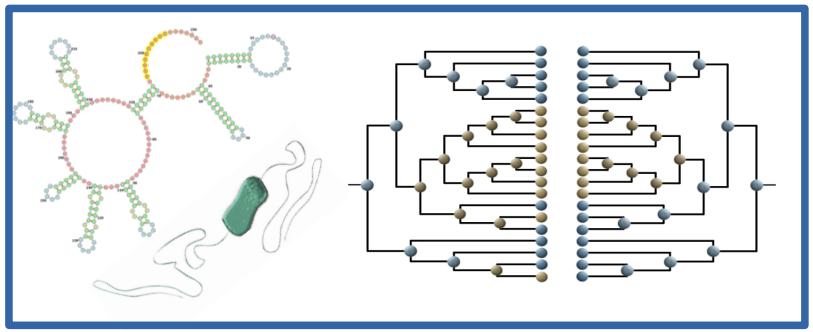


Peer A and Margalit H, J. Bacteriol., 2011
Skippington E and Ragan MA, Genome Biol. Evol., 2012

Addressed questions

How do **sRNAs evolve at a bacterial genus** level ?

Are there **coevolving** relationships between sRNAs and mRNAs regions ?



Does **coevolution patterns help to**

- Propose a **putative function** for some sRNAs ?
- Predict **mRNA targets** of some sRNAs ?

How detecting coevolution between sRNA and mRNA ?

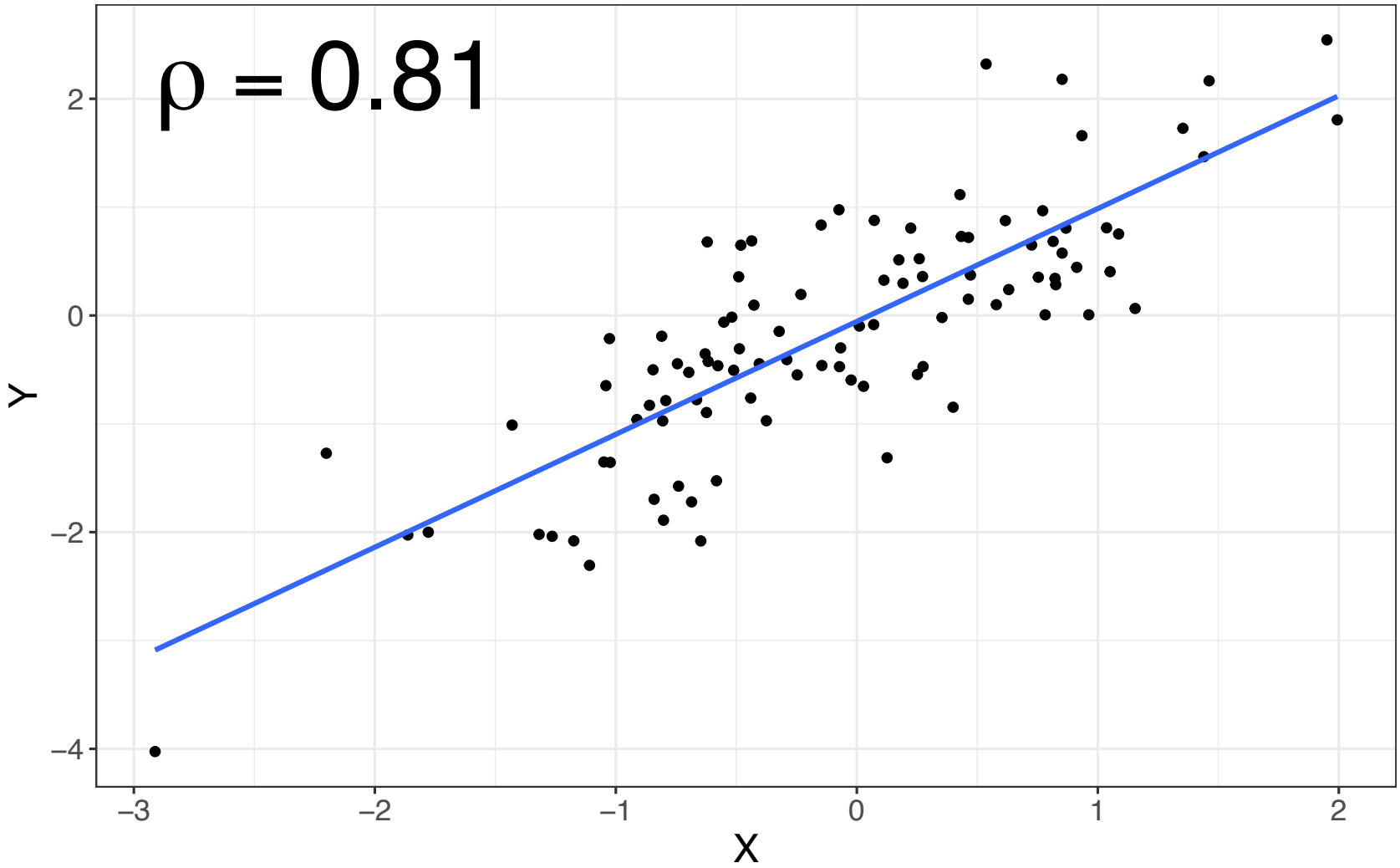
Methodological challenge

Develop a comparative genomics approach that:

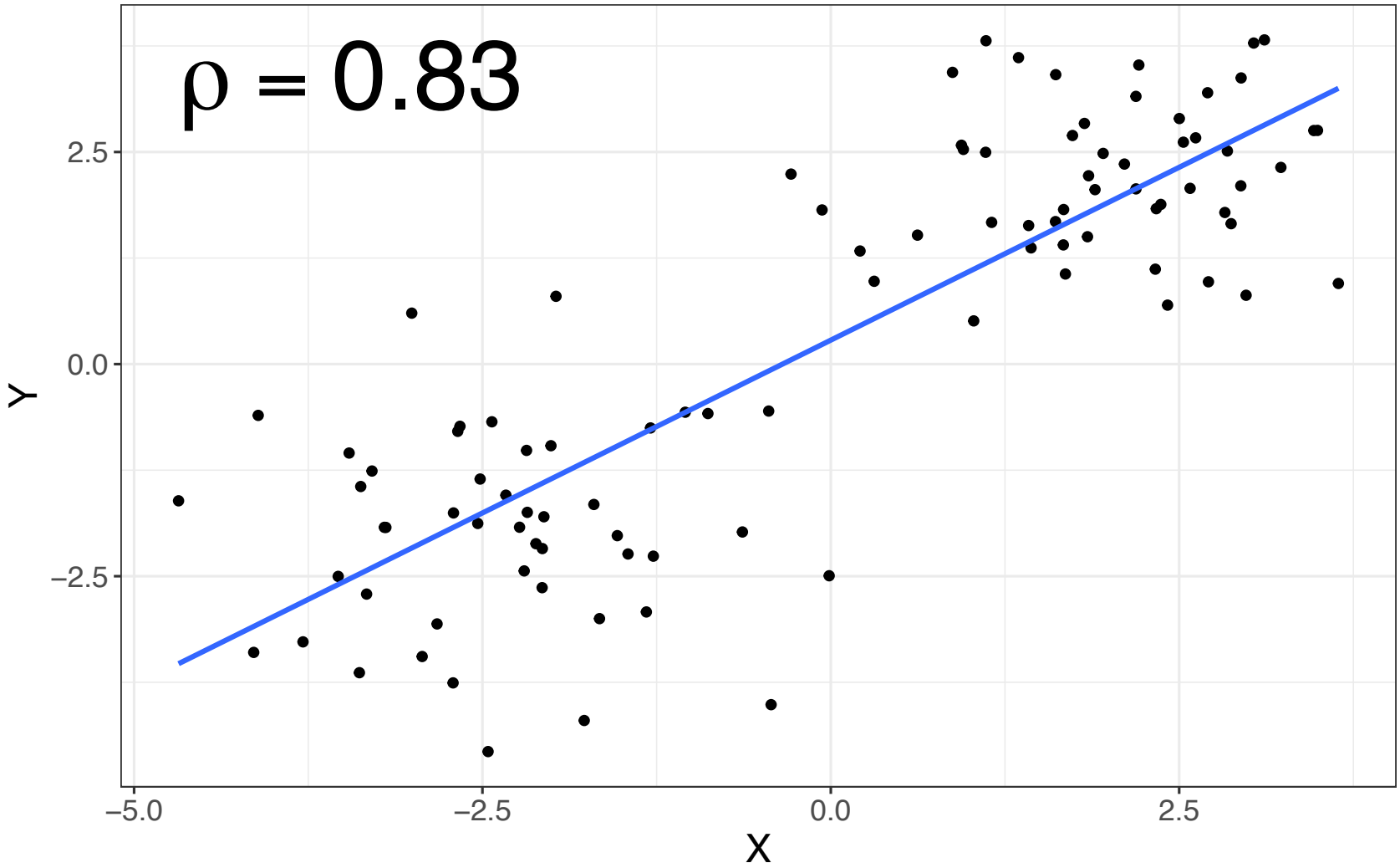
- reconstruct mRNAs and sRNA presence/absence patterns in a set of bacterial genomes (from a given genus)

- detect coevolution events from those patterns

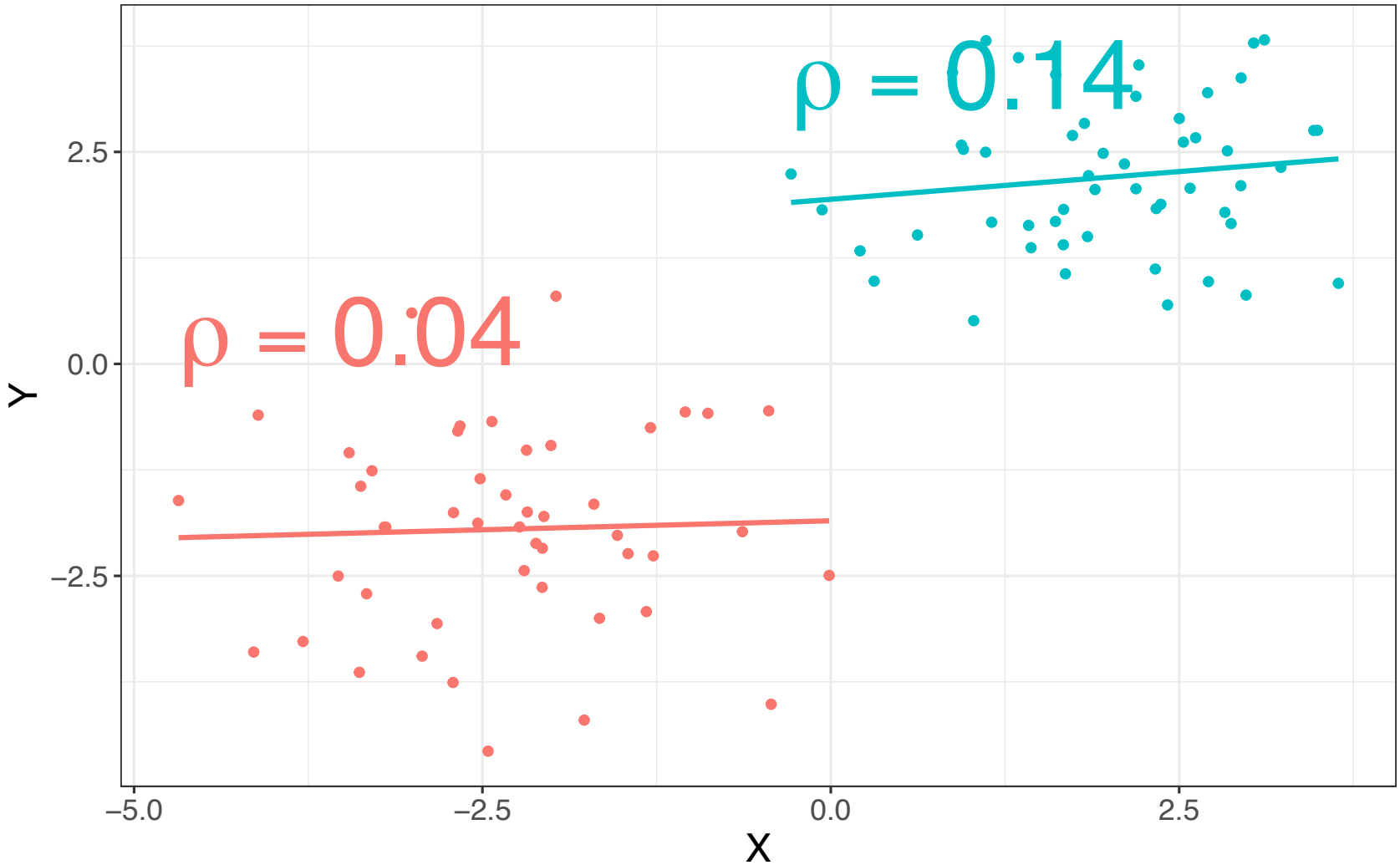
Comparing things



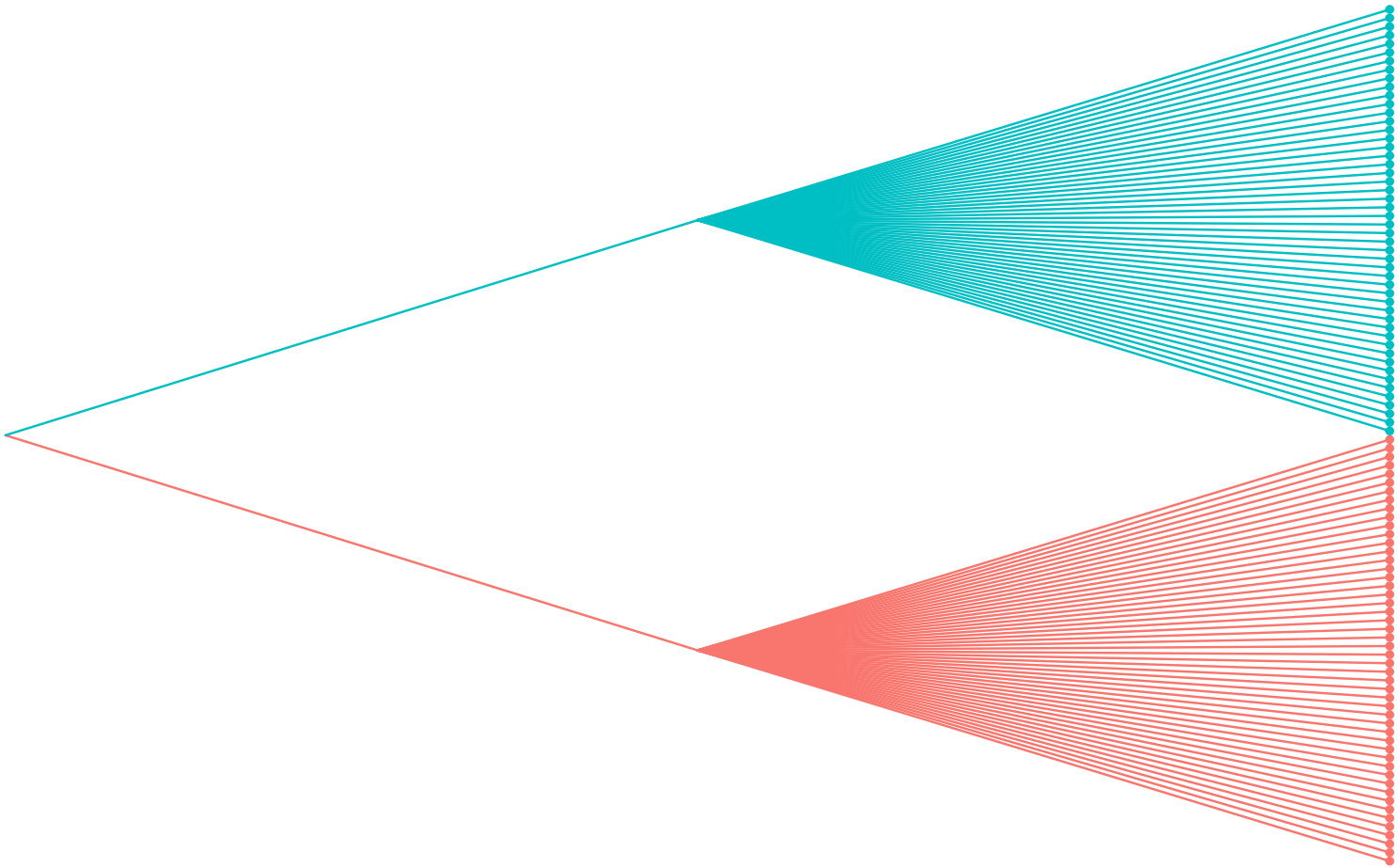
Comparing things (Cont'd)



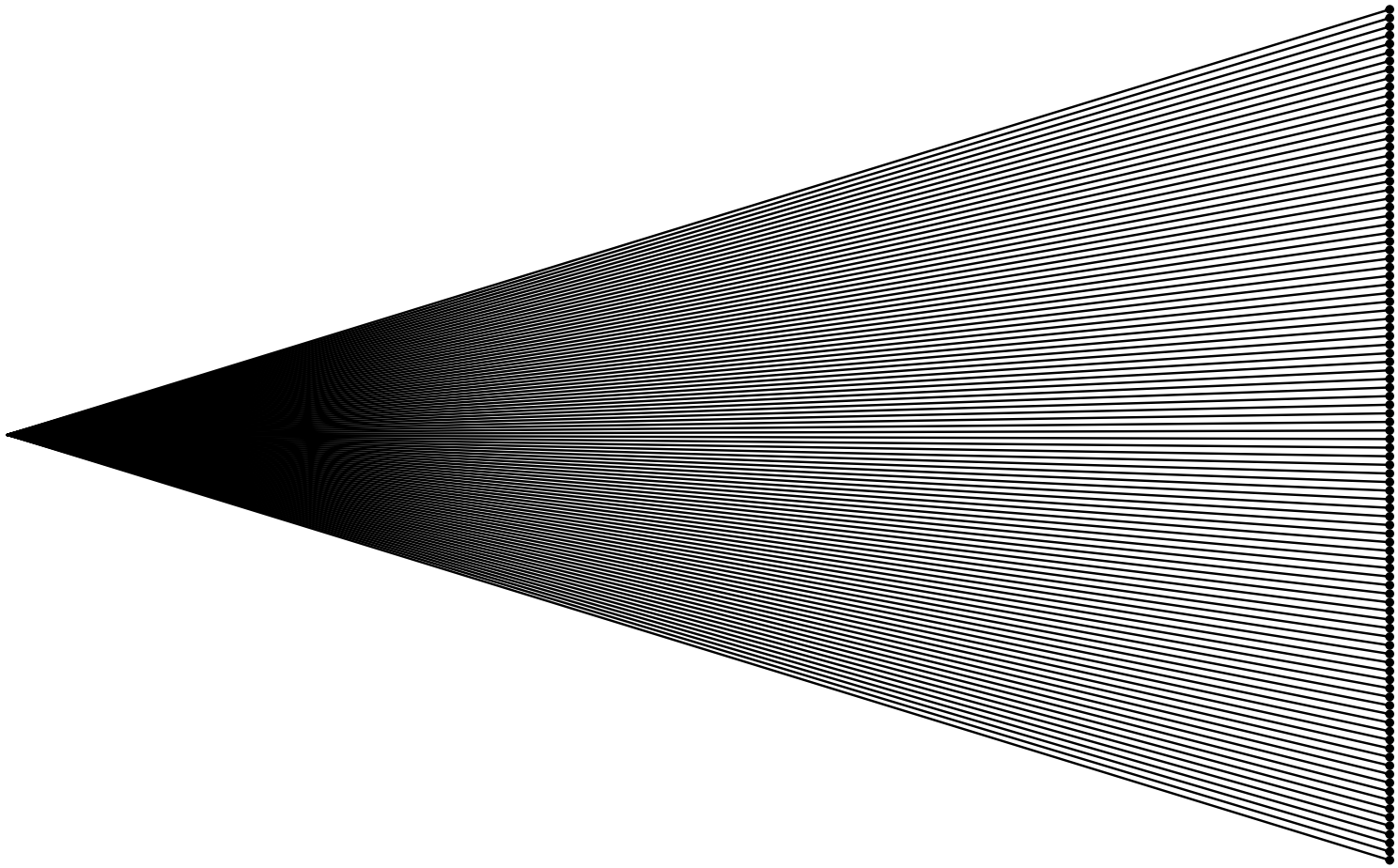
Comparing things (Cont'd)



Felsenstein's Worst Case Scenario



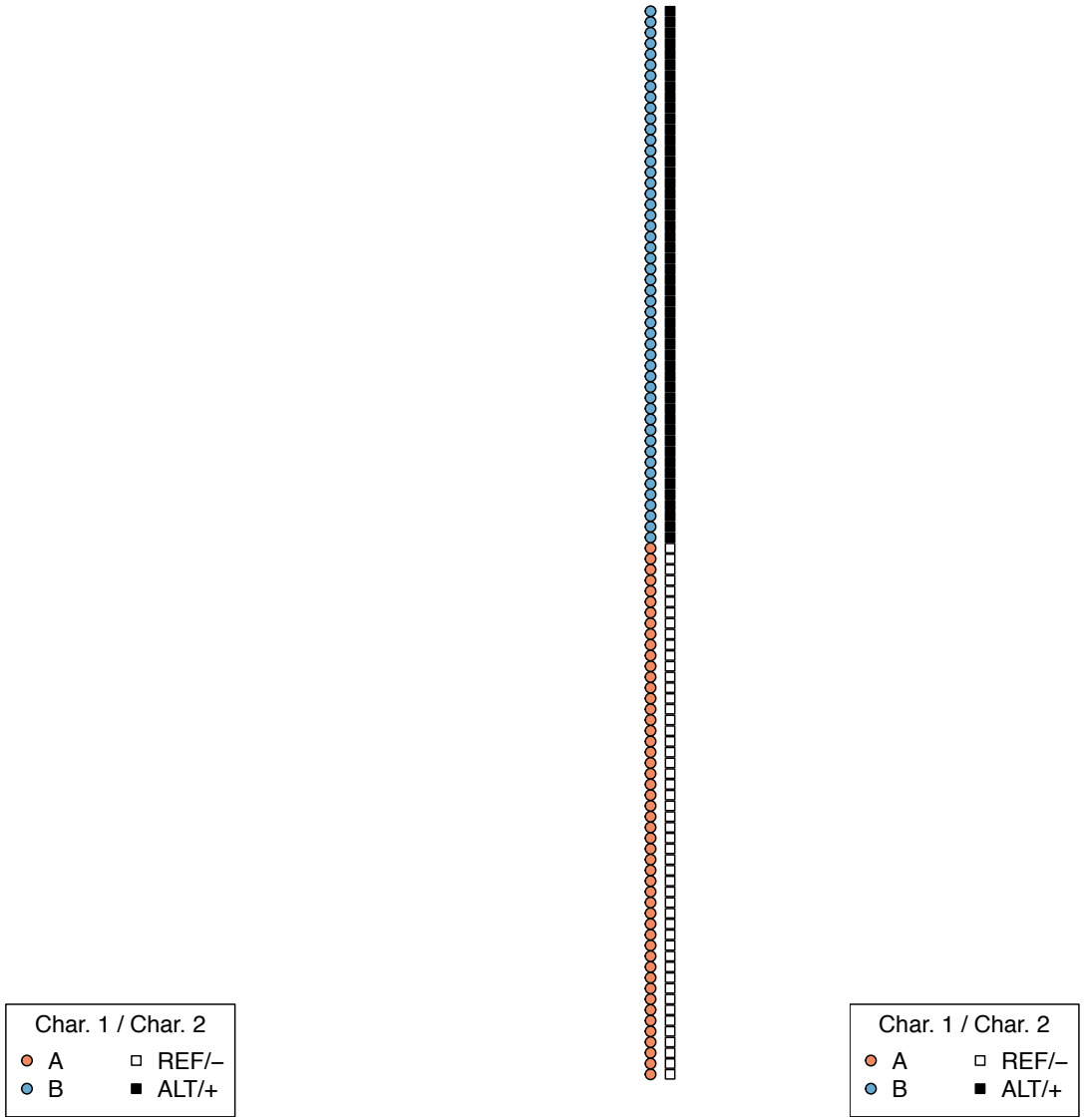
What You Had in Mind



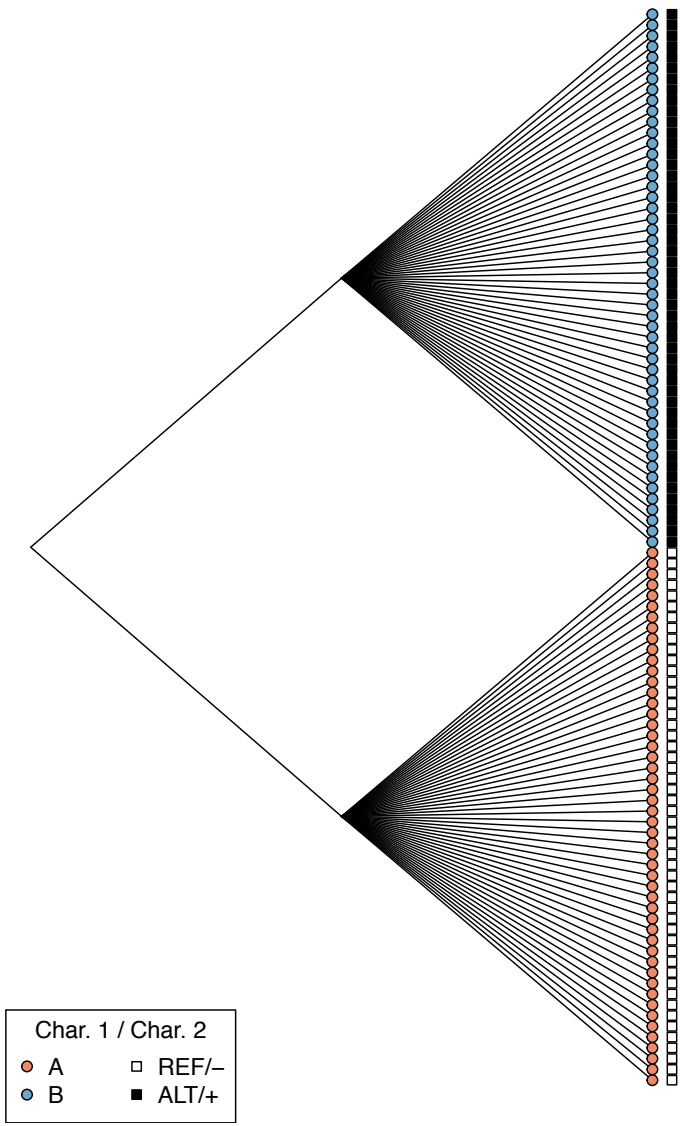
Outline

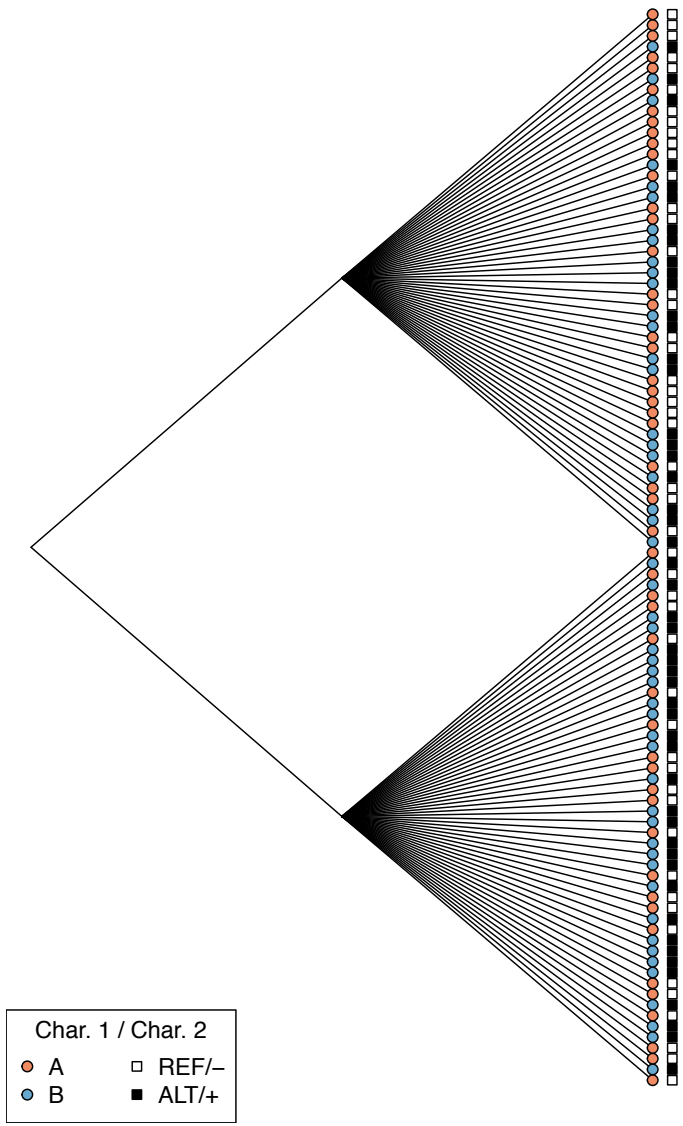
1. Motivation
   - A Simple Example
   - Another Example
   - One Last Example (courtesy of P. Bastide)
   - Take Home Message
2. Continuous Characters
3. Discrete Characters
4. Summary

Comparing things



Comparing things (Cont'd)

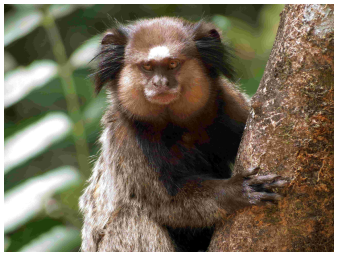




Outline

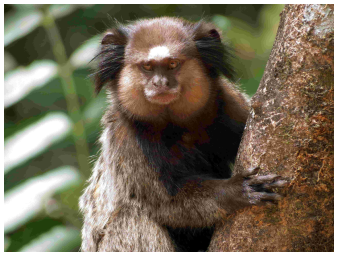
1. Motivation
   - A Simple Example
   - Another Example
   - One Last Example (courtesy of P. Bastide)
   - Take Home Message
2. Continuous Characters
3. Discrete Characters
4. Summary

New World Monkeys (Aristide et al., 2016)

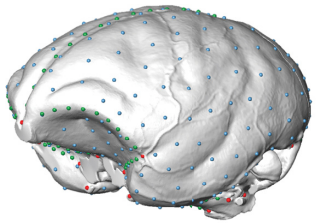


Callithrix penicillata

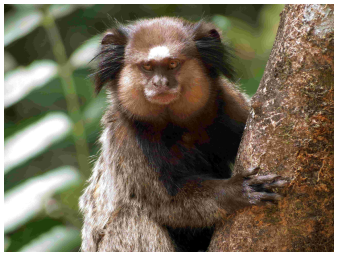
New World Monkeys (Aristide et al., 2016)



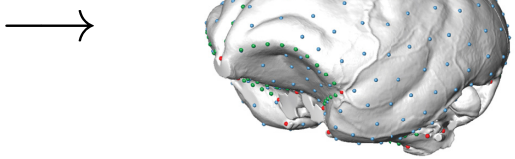
Callithrix penicillata

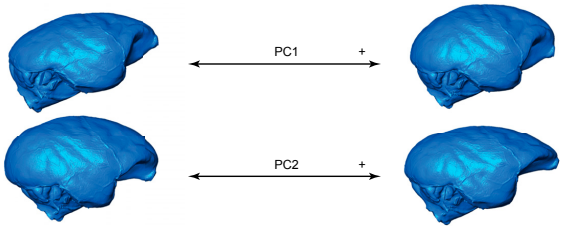


New World Monkeys (Aristide et al., 2016)

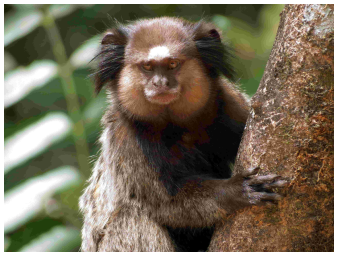


Callithrix penicillata

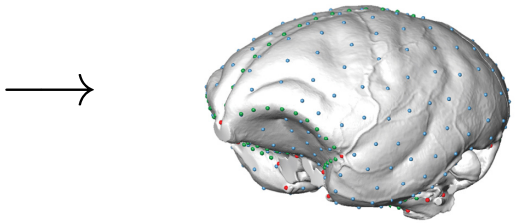


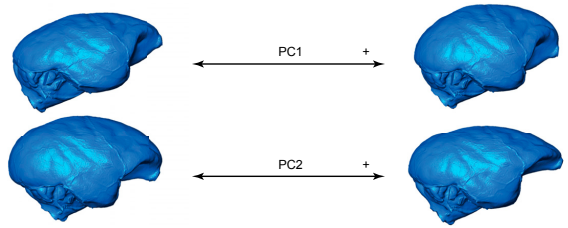


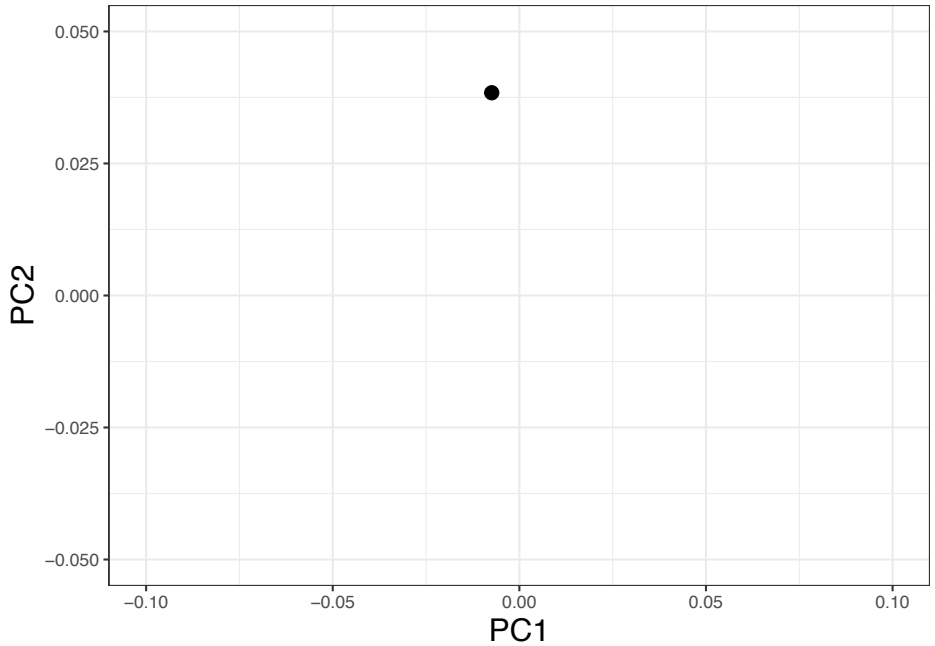
New World Monkeys (Aristide et al., 2016)



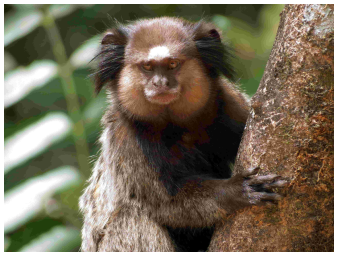
Callithrix penicillata



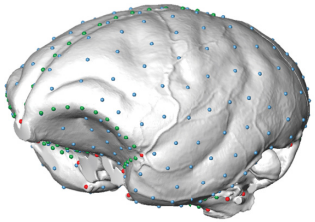


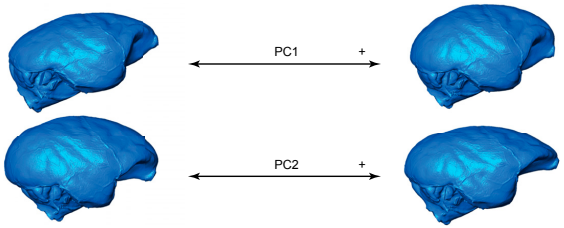


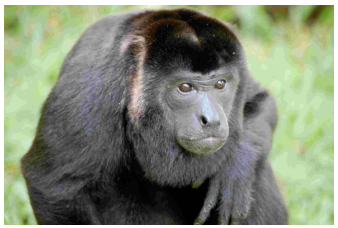
New World Monkeys (Aristide et al., 2016)



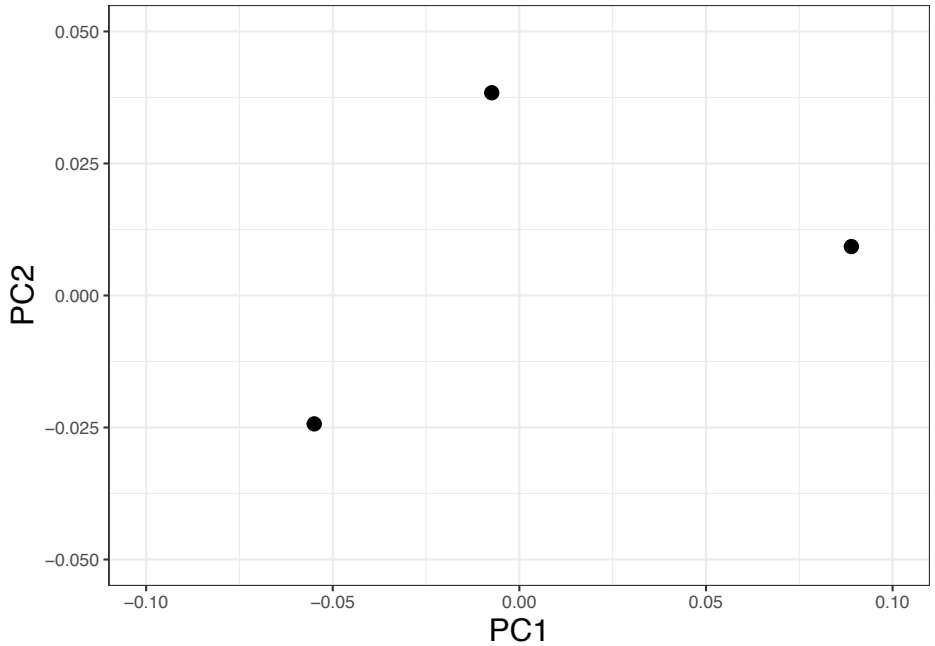
Callithrix penicillata

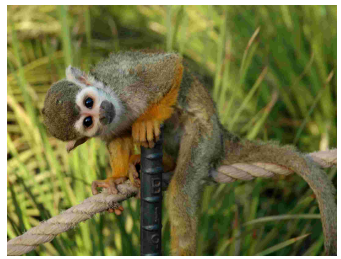






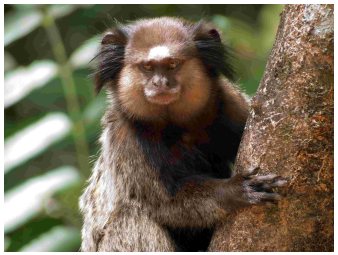
Alouatta palliata



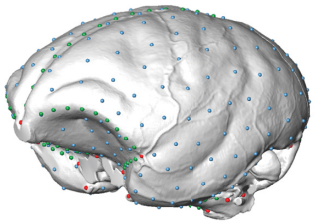


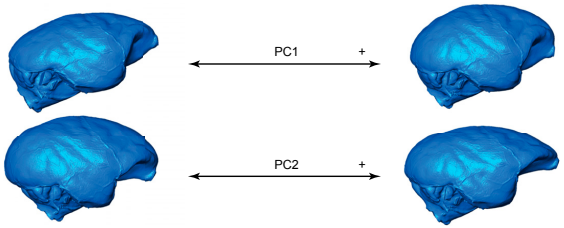
Saimiri sciureus

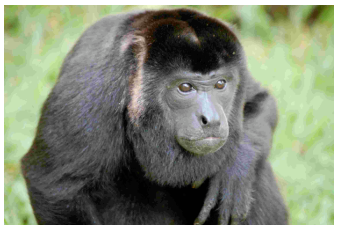
New World Monkeys (Aristide et al., 2016)



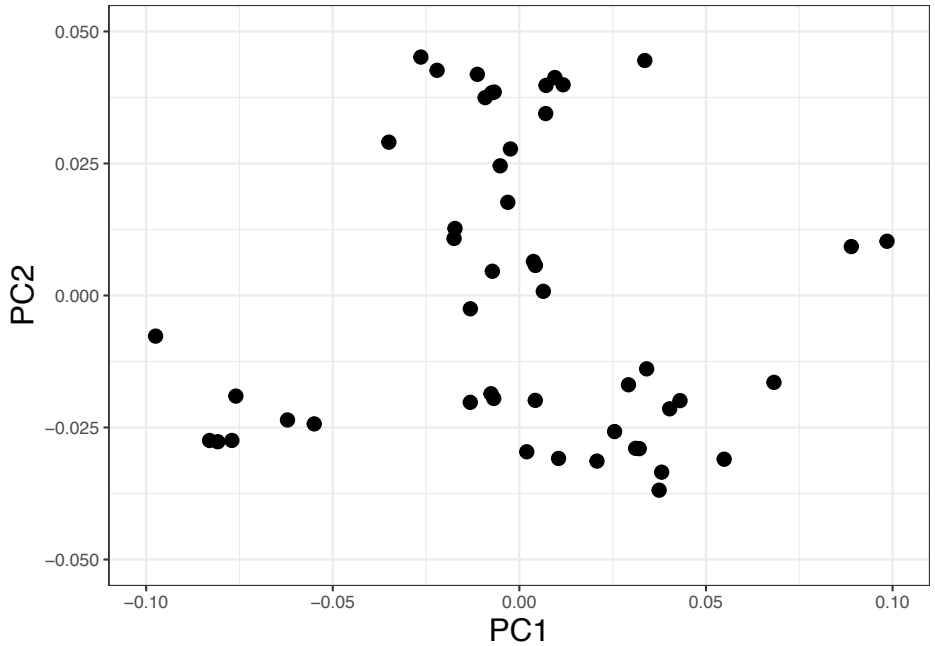
Callithrix penicillata

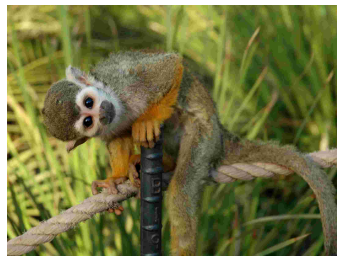






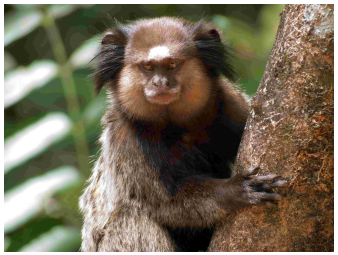
Alouatta palliata



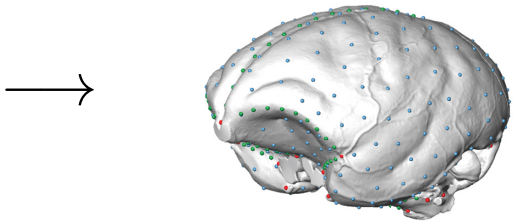


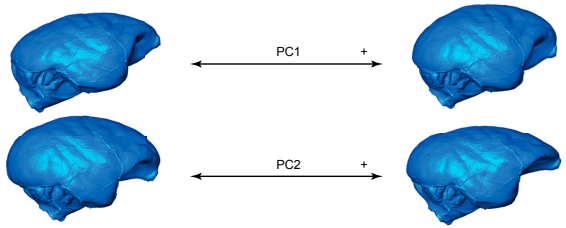
Saimiri sciureus

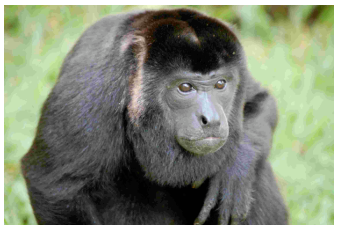
New World Monkeys (Aristide et al., 2016)



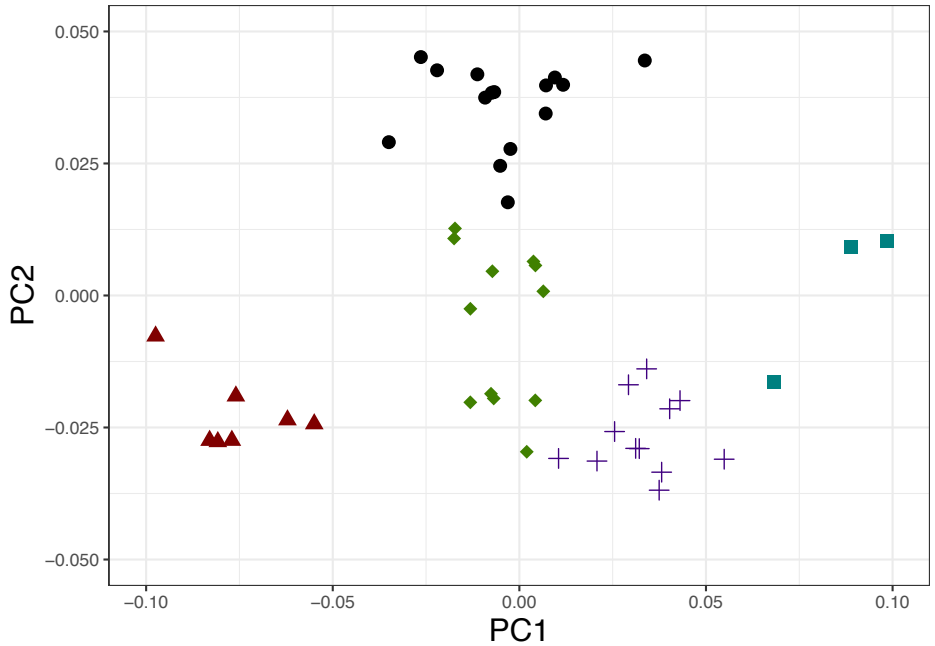
Callithrix penicillata

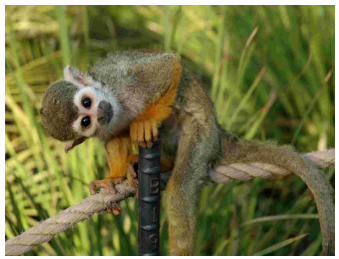






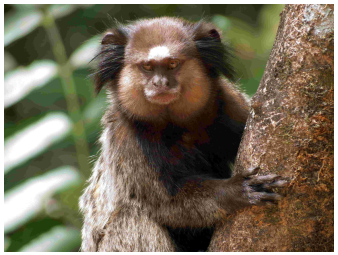
Alouatta palliata



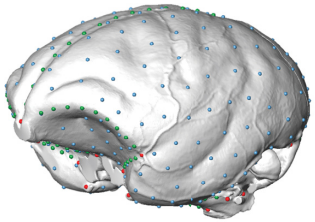


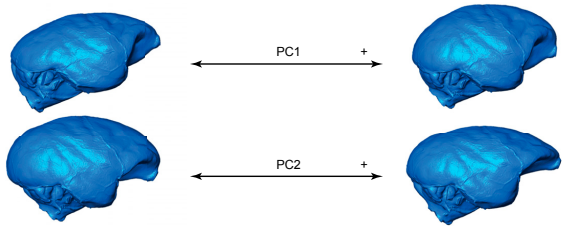
Saimiri sciureus

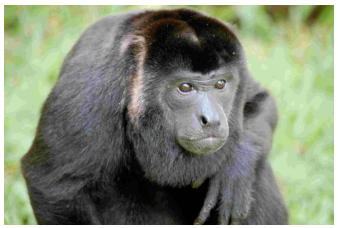
New World Monkeys (Aristide et al., 2016)



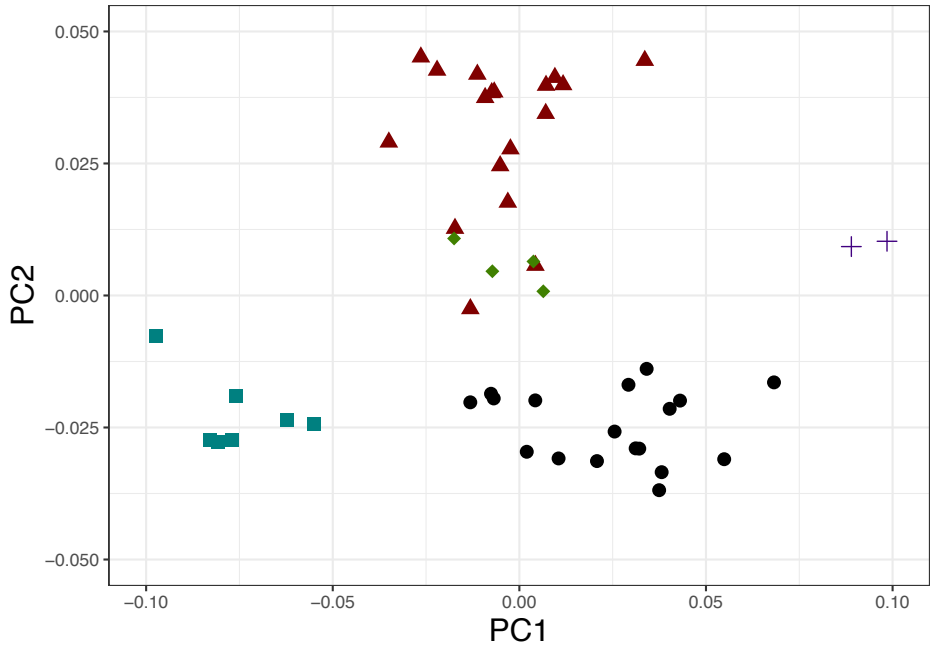
Callithrix penicillata

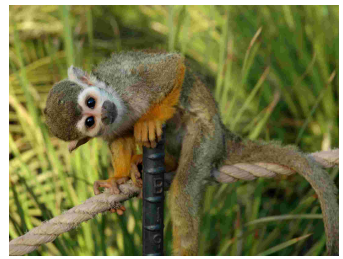






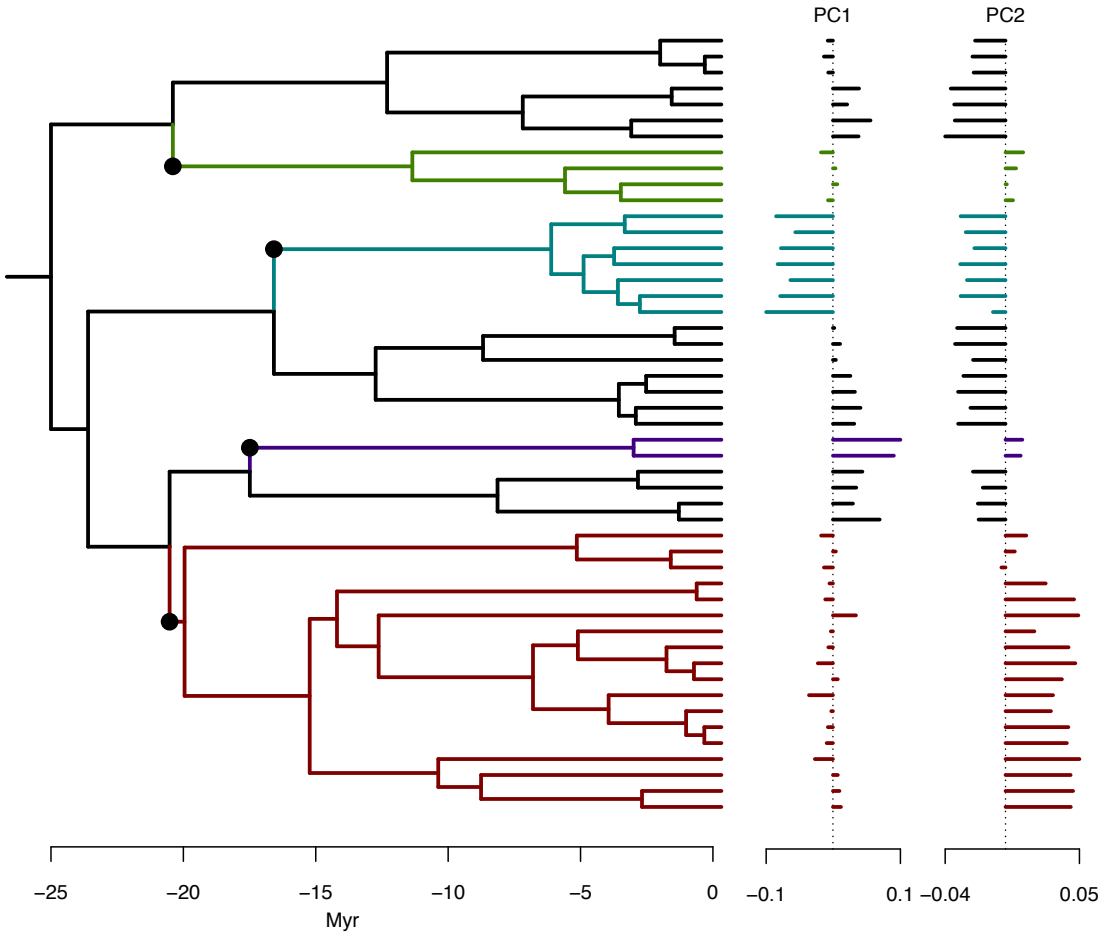
Alouatta palliata

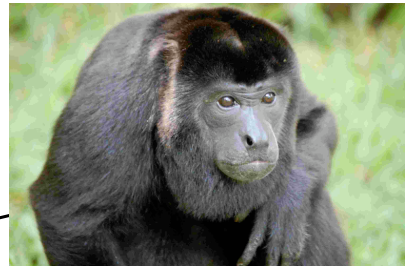




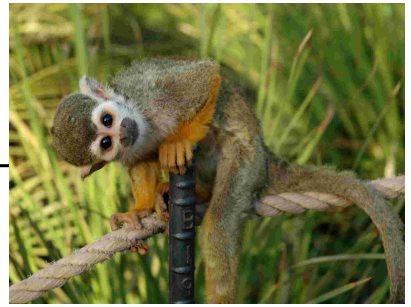
Saimiri sciureus

New World Monkeys (Aristide et al., 2016)

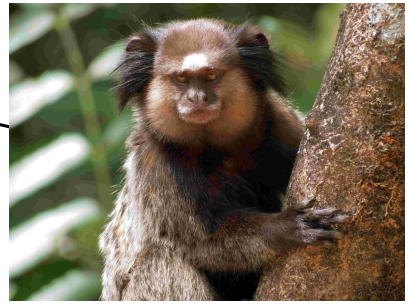




Alouatta palliata



Saimiri sciureus



Callithrix penicillata

Outline

1. Motivation
   - A Simple Example
   - Another Example
   - One Last Example (courtesy of P. Bastide)
   - Take Home Message
2. Continuous Characters
3. Discrete Characters
4. Summary

Tree Thinking

Evolution Matters!!

Tree Thinking

Evolution Matters!!

Modeling Evolution

- Be careful when comparing traits on evolutionary-related organisms
- If traits don't respond instantaneously to natural selection, there is phylogenetic inertia
- Need to model trait evolution (along the tree)

1 Motivation
- A Simple Example
- Another Example
- One Last Example (courtesy of P. Bastide)
- Take Home Message

2 Continuous Characters
- Brownian Motion
- Multivariate Brownian Motion
- Phylogenetic Correlation/Regression
- To Brownian Motion and Beyond

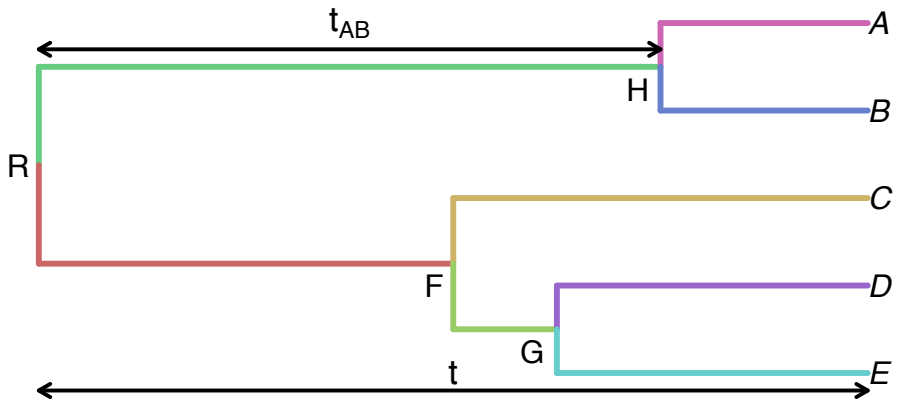
3 Discrete Characters
- Univariate Models
- Multivariate Characters

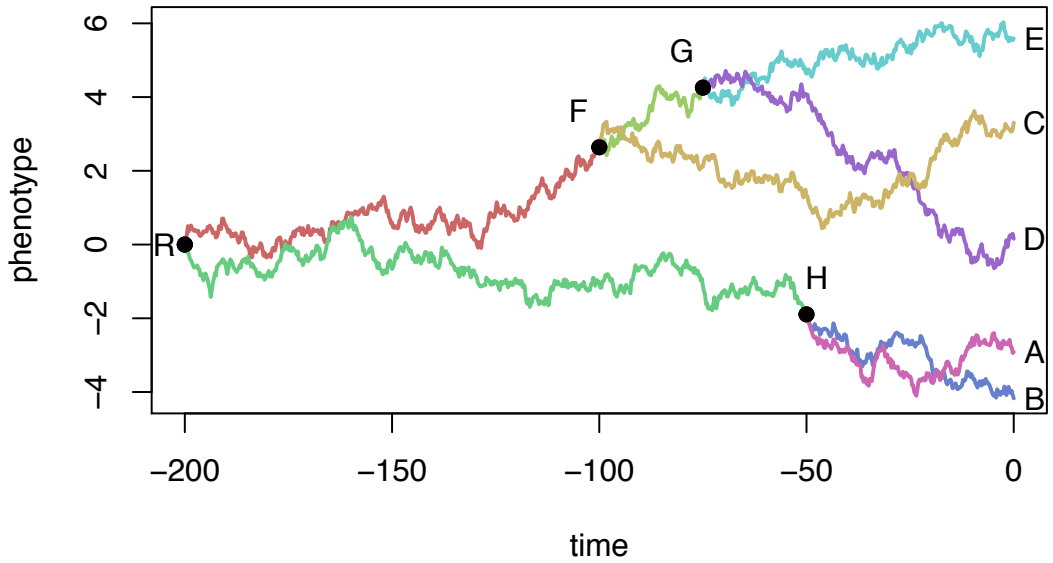
4 Summary

Outline

1. Motivation
2. Continuous Characters
   - Brownian Motion
   - Multivariate Brownian Motion
   - Phylogenetic Correlation/Regression
   - To Brownian Motion and Beyond
3. Discrete Characters
4. Summary

Stochastic Process on a Tree (Felsenstein, 1985)





The tree is **known**.

Only **tip** values are observed

Process described on a **single** branch.

Process **duplicated** at each node.

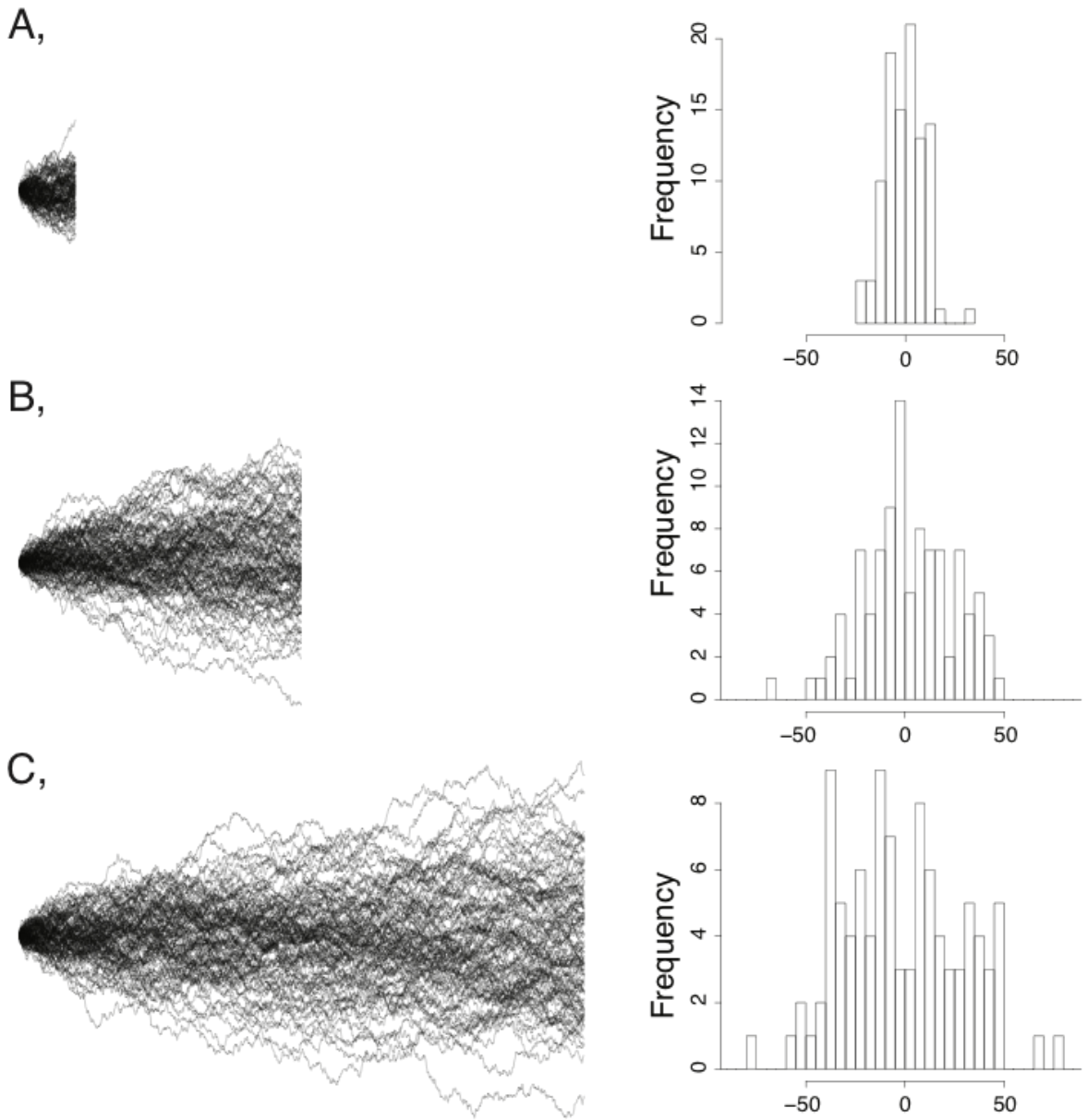
A bit of Mathematics

The BM is the solution is the solution to the stochastic differential equation:

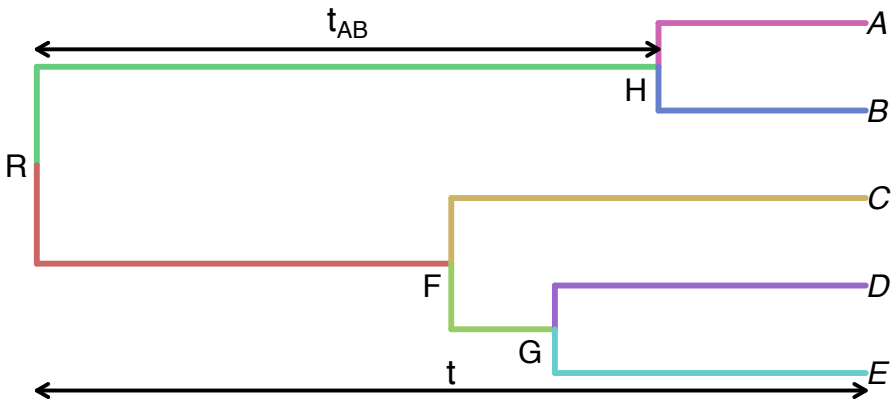
$$dX(t) = \sigma dB(t)$$

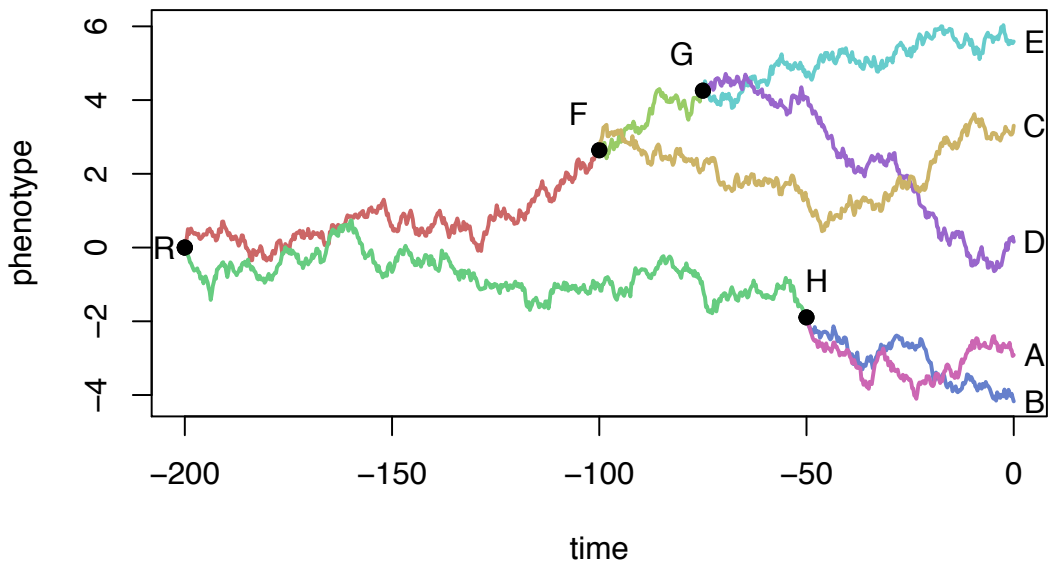
and satisfies:

- $E[X(0)] = \mu$
- The increments of X are independent
- $X(t) - X(0) \sim \mathcal{N}(0, \sigma^2 t)$


A,
B,
C,
Frequency
0
5
10
15
20
−50
0
50
Frequency
0
2
4
6
8
10
12
14
−50
0
50
Frequency
0
2
4
6
8
−50
0
50

Phylogenetic Correlation



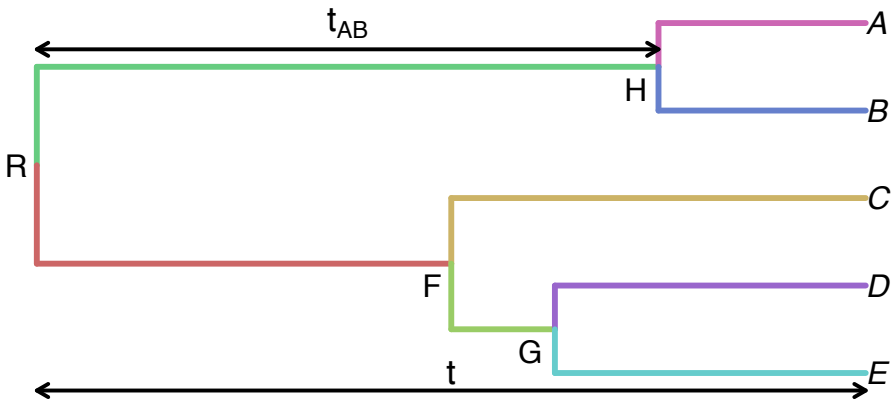


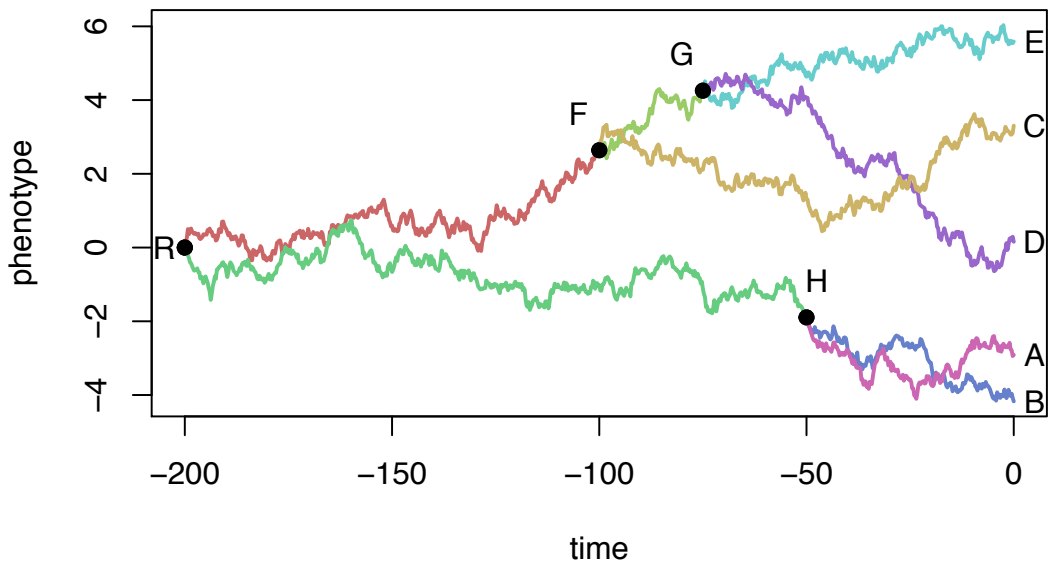
$$\mathbb{V}\text{ar}\left[A \mid R\right] = \sigma^2 t$$

$$A - R = (A - H) + (H - R)$$

$$B - R = (B - H) + (H - R)$$

Phylogenetic Correlation





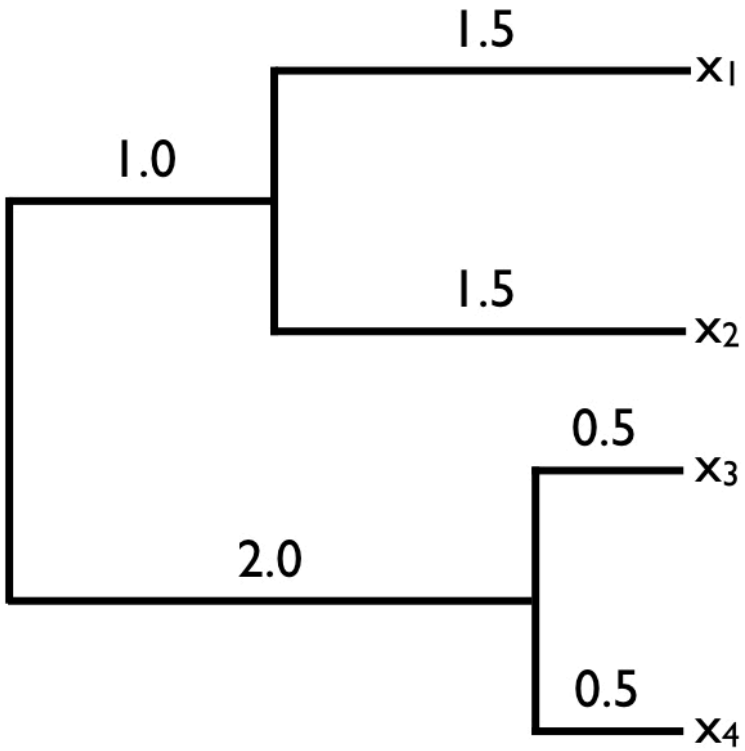
$$\mathbb{V}\text{ar}\left[A \mid R\right] = \sigma^2 t$$

$$A - R = (A - H) + (H - R)$$

$$B - R = (B - H) + (H - R)$$

$$\mathbb{C}\text{ov}\left[A; B \mid R\right] = \sigma^2 t_{AB}$$

Phylogenetic Correlation



$$C = \begin{bmatrix} 2.5 & 1.0 & 0 & 0 \\ 1.0 & 2.5 & 0 & 0 \\ 0 & 0 & 2.5 & 2.0 \\ 0 & 0 & 2.0 & 2.5 \end{bmatrix}$$

Model

Noting $\mathbf{Y}$ the size-n vector of observed values

$$\mathbf{Y} \sim \mathcal{N}(\mu \mathbf{1}_n, \sigma^2 \mathbf{C})$$

with $\mathbf{C}$ is fully determined by the tree.

Model

Noting $\mathbf{Y}$ the size-n vector of observed values

$$\mathbf{Y} \sim \mathcal{N}(\mu \mathbf{1}_n, \sigma^2 \mathbf{C})$$

with $\mathbf{C}$ is fully determined by the tree.

Likelihood

$$L(\mathbf{Y}; \mu, \sigma^2, \mathbf{C}) = -\frac{(\mathbf{Y} - \mu \mathbf{1_n})^\intercal \mathbf{C}^{-1} (\mathbf{Y} - \mu \mathbf{1}_n)}{2\sigma^2} - \frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \log |\mathbf{C}|$$

Model

Noting $\mathbf{Y}$ the size-n vector of observed values

$$\mathbf{Y} \sim \mathcal{N}(\mu \mathbf{1}_n, \sigma^2 \mathbf{C})$$

with $\mathbf{C}$ is fully determined by the tree.

Likelihood

$$L(\mathbf{Y}; \mu, \sigma^2, \mathbf{C}) = -\frac{(\mathbf{Y} - \mu \mathbf{1_n})^\intercal \mathbf{C}^{-1} (\mathbf{Y} - \mu \mathbf{1}_n)}{2\sigma^2} - \frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \log |\mathbf{C}|$$

Estimates

$$\hat{\mu} = (\mathbf{1}^\intercal \mathbf{C}^{-1} \mathbf{1})^{-1} (\mathbf{1}^\intercal \mathbf{C}^{-1} \mathbf{Y})$$

$$\hat{\sigma^2} = \frac{(\mathbf{Y} - \hat{\mu} \mathbf{1_n})^\intercal \mathbf{C}^{-1} (\mathbf{Y} - \hat{\mu} \mathbf{1}_n)}{n}$$

Remarks

- $\mathbf{1}^\intercal \mathbf{C}^{-1}$ act as a vector of **weights**
- $\mathbf{1}^\intercal \mathbf{C}^{-1}\mathbf{1}$ act as a an effective **sample size**
- Hence
  - $\hat{\mu}$ is a weighted average of $\mathbf{Y}$
  - $\hat{\sigma^2}$ is the usual norm of $(\mathbf{Y} - \hat{\mathbf{Y}})$ but under the $\mathbf{C}$-metric

Recap

- We can model a continuous trait on a tree :)

Recap

- We can model a continuous trait on a tree :)
- But we can't compare two traits yet :(

Outline

1. Motivation
2. Continuous Characters
   - Brownian Motion
   - Multivariate Brownian Motion
   - Phylogenetic Correlation/Regression
   - To Brownian Motion and Beyond
3. Discrete Characters
4. Summary

Multivariate Traits

Idea

- Model the joint evolution of many traits on a **single** branch...
- using a standard multivariate gaussian with rate matrix $\mathbf{R}$
- before adding a phylogenetic structure.

Multivariate Traits

Idea

- Model the joint evolution of many traits on a **single** branch...
- using a standard multivariate gaussian with rate matrix $\mathbf{R}$
- before adding a phylogenetic structure.

Idea

- $\mathbf{C}$ captures the variance across *organisms* due to shared history
- $\mathbf{R}$ captures the variance across *traits* due to coevolution

Multivariate Brownian Model (mvBM)

The mvBM is solution to the stochastic differential equation:

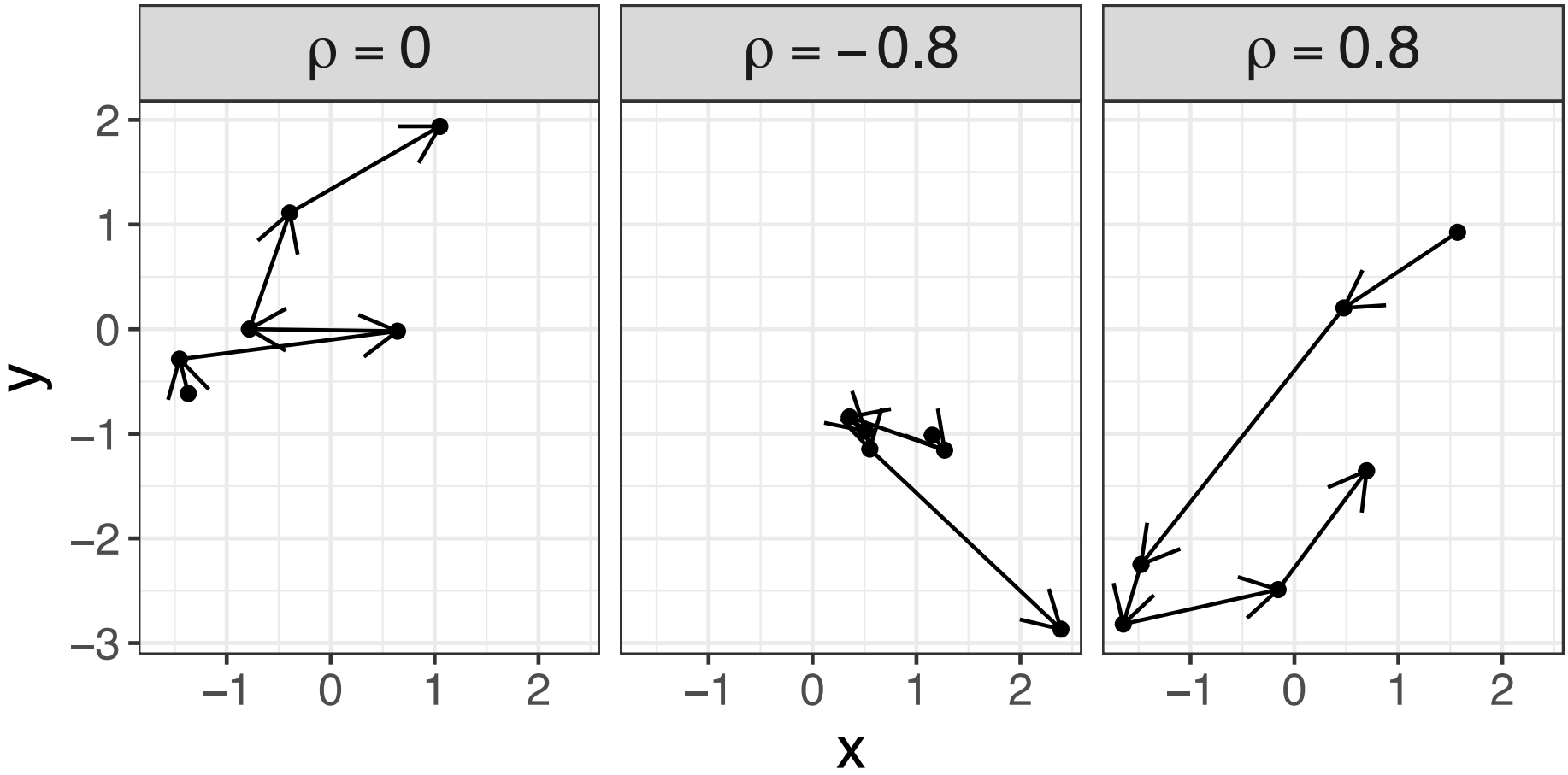
$$d\mathbf{X}(t) = \mathbf{R}^{1/2}\sigma d\mathbf{B}(t)$$

and satisfies:

- $E[\mathbf{X}(0)] = \boldsymbol{\mu}$
- The increments of $\mathbf{X}$ are independent
- $\mathbf{X}(t) - \mathbf{X}(0) \sim \mathcal{N}(\mathbf{0}, t\mathbf{R})$

Example: A Bivariate BM

Consider $\mathbf{R} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$



Correlation Structure

$\mathbf{Y}$ is now a $n \times p$ matrix (n organisms $\times$ p traits) satisfying

$$\mathbb{C}\text{ov}(Y_{ik}, Y_{jl}) = R_{kl} \times C_{ij}$$

Correlation Structure

$\mathbf{Y}$ is now a $n \times p$ matrix (n organisms $\times$ p traits) satisfying

$$\mathbb{C}\text{ov}(Y_{ik}, Y_{jl}) = R_{kl} \times C_{ij}$$

The covariance factors as the product of a phylogenetic component (C_{ij}) and an phenotypic one (R_{kl}).

Correlation Structure

$\mathbf{Y}$ is now a $n \times p$ matrix (n organisms $\times$ p traits) satisfying

$$\mathbb{C}\text{ov}(Y_{ik}, Y_{jl}) = R_{kl} \times C_{ij}$$

The covariance factors as the product of a phylogenetic component (C_{ij}) and an phenotypic one (R_{kl}).

In particular

$$\mathbf{V} = \mathbb{V}\text{ar}(\mathbf{Y}) = \mathbf{R} \otimes \mathbf{C}$$

Correlation Structure

$\mathbf{Y}$ is now a $n \times p$ matrix (n organisms $\times$ p traits) satisfying

$$\mathbb{C}\text{ov}(Y_{ik}, Y_{jl}) = R_{kl} \times C_{ij}$$

The covariance factors as the product of a phylogenetic component (C_{ij}) and an phenotypic one (R_{kl}).

In particular

$$\mathbf{V} = \mathbb{V}\text{ar}(\mathbf{Y}) = \mathbf{R} \otimes \mathbf{C}$$

$\mathbf{V}$ is $np \times np$ and captures the covariance of all across all species.

Model

Noting $\mathbf{Y}$ the $n \times p$ vector of observed values (in vector format)

$$\mathbf{Y} \sim \mathcal{N}_m(\boldsymbol{\mu} \otimes \mathbf{1}_n, \mathbf{R} \otimes \mathbf{C})$$

with $\mathbf{C}$ is fully determined by the tree.

Model

Noting $\mathbf{Y}$ the $n \times p$ vector of observed values (in vector format)

$$\mathbf{Y} \sim \mathcal{N}_m(\boldsymbol{\mu} \otimes \mathbf{1}_n, \mathbf{R} \otimes \mathbf{C})$$

with $\mathbf{C}$ is fully determined by the tree.

Likelihood

$$L(\mathbf{Y}; \boldsymbol{\mu}, \mathbf{R}, \mathbf{C}) = -\frac{(\mathbf{Y} - \boldsymbol{\mu} \otimes \mathbf{1_n})^\intercal \mathbf{R}^{-1} \otimes \mathbf{C}^{-1} (\mathbf{Y} - \boldsymbol{\mu} \otimes \mathbf{1}_n)}{2} - \frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{R} \otimes \mathbf{C}|$$

Model

Noting $\mathbf{Y}$ the $n \times p$ vector of observed values (in vector format)

$$\mathbf{Y} \sim \mathcal{N}_m(\boldsymbol{\mu} \otimes \mathbf{1}_n, \mathbf{R} \otimes \mathbf{C})$$

with $\mathbf{C}$ is fully determined by the tree.

Likelihood

$$L(\mathbf{Y}; \boldsymbol{\mu}, \mathbf{R}, \mathbf{C}) = -\frac{(\mathbf{Y} - \boldsymbol{\mu} \otimes \mathbf{1_n})^\intercal \mathbf{R}^{-1} \otimes \mathbf{C}^{-1} (\mathbf{Y} - \boldsymbol{\mu} \otimes \mathbf{1}_n)}{2} - \frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{R} \otimes \mathbf{C}|$$

Estimates

$$\hat{\boldsymbol{\mu}} = (\mathbf{1}^\intercal \mathbf{C}^{-1} \mathbf{1})^{-1} (\mathbf{1}^\intercal \mathbf{C}^{-1} \mathbf{Y})^\intercal$$

$$\hat{\mathbf{R}} = \frac{(\mathbf{Y} - \mathbf{1_n} \hat{\boldsymbol{\mu}}^\intercal)^\intercal \mathbf{C}^{-1} (\mathbf{Y} - \mathbf{1}_n \hat{\boldsymbol{\mu}}^\intercal)}{n}$$

Outline

1. Motivation
2. Continuous Characters
   - Brownian Motion
   - Multivariate Brownian Motion
   - Phylogenetic Correlation/Regression
   - To Brownian Motion and Beyond
3. Discrete Characters
4. Summary

Evolutionary Correlation

$\hat{R}_{kl}$ is the estimated evolutionary correlation between traits k and l

Evolutionary Correlation

$\hat{R}_{kl}$ is the estimated evolutionary correlation between traits k and l

One can fall back on standard statistics and test

- $R_{kl} = 0$ against
- $R_{kl} \neq 0$

using a Likelihood Ratio Test (for example) or integrate those models in a Bayesian framework.

Phylogenetic Regression

Why Stop With Correlation?

- Correlation is symmetric by nature

Phylogenetic Regression

Why Stop With Correlation?

- Correlation is symmetric by nature
- To compute the effect of one (or more) traits on another, we usually use **regression**

Phylogenetic Regression

Why Stop With Correlation?

- Correlation is symmetric by nature
- To compute the effect of one (or more) traits on another, we usually use **regression**
- Let's define a phylogenetic regression!!

Phylogenetic Regression

Why Stop With Correlation?

- Correlation is symmetric by nature
- To compute the effect of one (or more) traits on another, we usually use **regression**
- Let's define a phylogenetic regression!!

Phylogenetic Regression

We consider the following (phylogenetic) regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{E} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{C})$$

where we assume that:

- $\mathbf{Y}$ has a phylogenetic structure;
- $\mathbf{E}$ has a phylogenetic structure;
- $\mathbf{X}$ may or may not have a phylogenetic structure

Phylogenetic Regression (Cont'd)

Usual Estimates

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^\intercal \mathbf{C}^{-1} \mathbf{X})^{-1} (\mathbf{X}^\intercal \mathbf{C}^{-1} \mathbf{Y})^\intercal$$

$$\hat{\sigma^2} = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}})^\intercal \mathbf{C}^{-1} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}})}{n}$$

Phylogenetic Regression (Cont'd)

Usual Estimates

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^\intercal \mathbf{C}^{-1} \mathbf{X})^{-1} (\mathbf{X}^\intercal \mathbf{C}^{-1} \mathbf{Y})^\intercal$$

$$\hat{\sigma^2} = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}})^\intercal \mathbf{C}^{-1} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}})}{n}$$

Remarks

- Phylogenetic regression is a special case of GLS
- Phylogenetic Independent Constrasts (PICs) are a special case of Phylogenetic Regression
- We can import many developments from linear models in this framework.

Outline

1. Motivation
2. Continuous Characters
   - Brownian Motion
   - Multivariate Brownian Motion
   - Phylogenetic Correlation/Regression
   - To Brownian Motion and Beyond
3. Discrete Characters
4. Summary

Extensions

Extensions

- Replace BM by Ornstein-Uhlenbeck process (optimal values)
- Replace BM by Levy process (Simpsonian evolution)
- Add discrete shifts (singular events)
- Replace σ^2 / $\mathbf{R}$ by a time-varying function (accelerating/decelerating evolution)
- Add diversity-dependence (trait value impacts diversification rates)
- ...

Extensions

Extensions

- Replace BM by Ornstein-Uhlenbeck process (optimal values)
- Replace BM by Levy process (Simpsonian evolution)
- Add discrete shifts (singular events)
- Replace σ^2 / $\mathbf{R}$ by a time-varying function (accelerating/decelerating evolution)
- Add diversity-dependence (trait value impacts diversification rates)
- ...

Caveats

- All these extensions make computations (a lot) harder.

Extensions

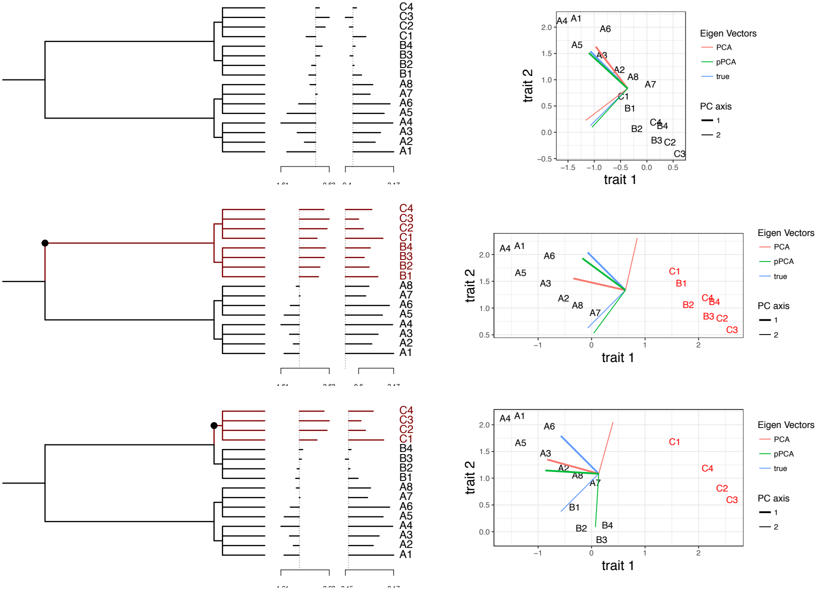
Extensions

- Replace BM by Ornstein-Uhlenbeck process (optimal values)
- Replace BM by Levy process (Simpsonian evolution)
- Add discrete shifts (singular events)
- Replace σ^2 / $\mathbf{R}$ by a time-varying function (accelerating/decelerating evolution)
- Add diversity-dependence (trait value impacts diversification rates)
- ...

Caveats

- All these extensions make computations (a lot) harder.
- But neglecting them can lead to serious mistakes...

Problem with singular events (Bastide et al., 2018)



1 Motivation
- A Simple Example
- Another Example
- One Last Example (courtesy of P. Bastide)
- Take Home Message

2 Continuous Characters
- Brownian Motion
- Multivariate Brownian Motion
- Phylogenetic Correlation/Regression
- To Brownian Motion and Beyond

3 Discrete Characters
- Univariate Models
- Multivariate Characters

4 Summary

Outline

1. Motivation
2. Continuous Characters
3. Discrete Characters
   - Univariate Models
   - Multivariate Characters
4. Summary

One character on one branch

The Mk (Markov) Model

- Discrete character evolves according to a **Markov** model
- Transition to state j depends only on **current** state i
- If the transition rate from i to j is q_{ij} then

$$\mathbb{P}\left[X(t+\Delta t)=j \mid X(t)=i\right] \simeq q_{ij}\Delta t$$

One character on one branch

The Mk (Markov) Model

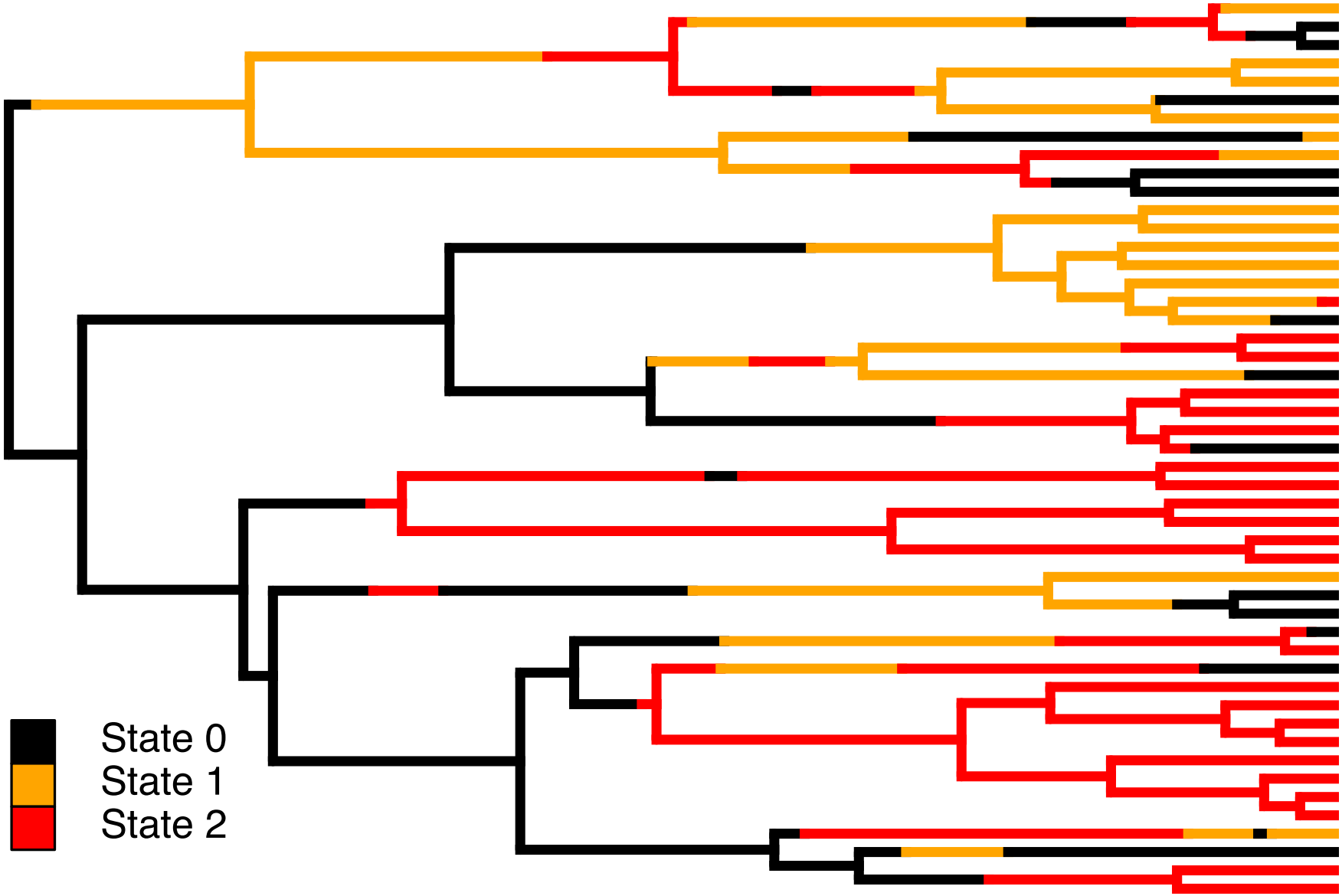
- Discrete character evolves according to a **Markov** model
- Transition to state j depends only on **current** state i
- If the transition rate from i to j is q_{ij} then

$$\mathbb{P}\left[X(t+\Delta t)=j \mid X(t)=i\right] \simeq q_{ij}\Delta t$$

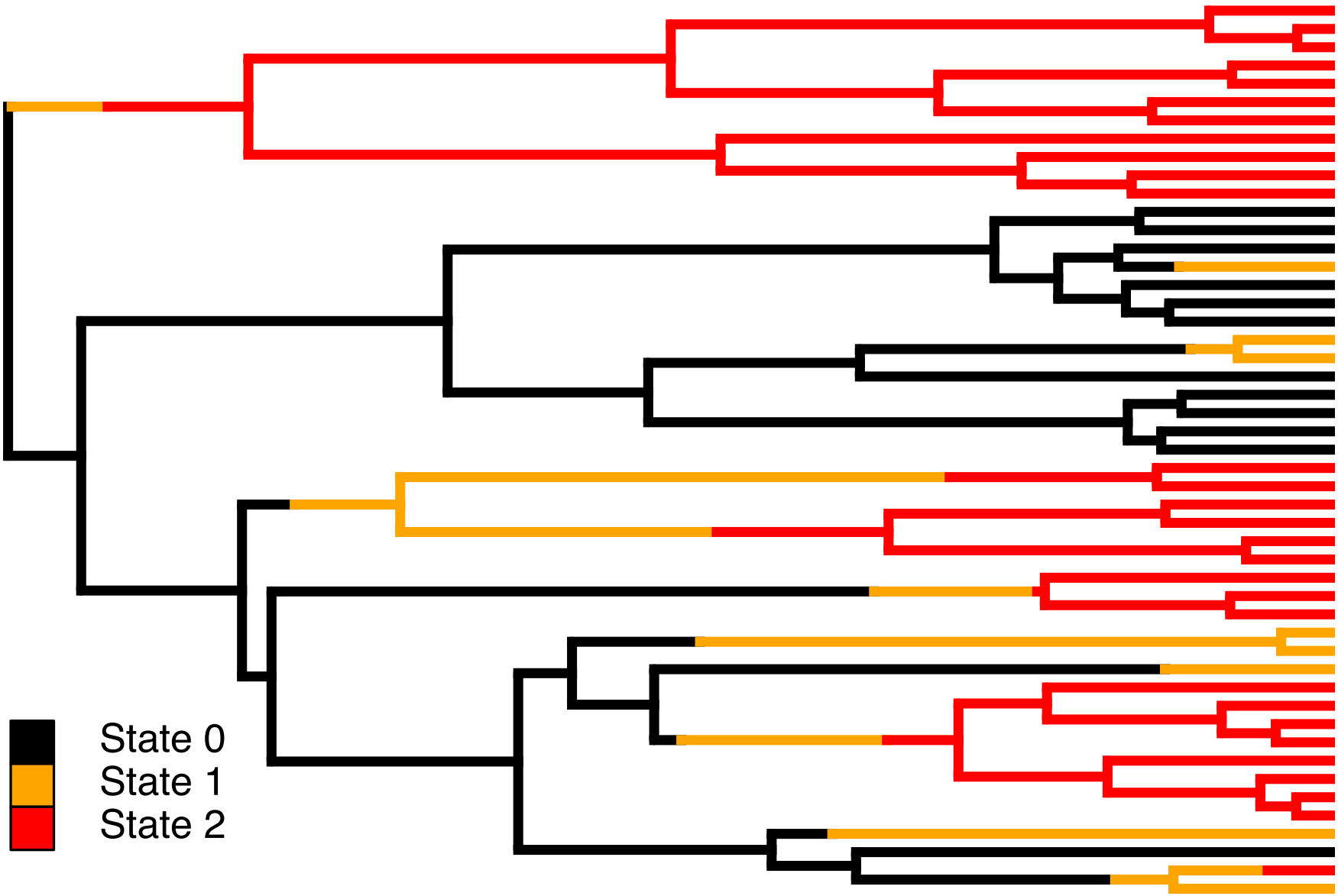
Rate Matrix

$$\mathbf{Q}=\begin{bmatrix} \bullet & q_{12} & \cdots & q_{1k} \\ q_{21} & \bullet & \cdots & q_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ q_{k1} & q_{k2} & \cdots & \bullet \end{bmatrix}$$

One character on a Tree (I)



One character on a Tree (II)



Likelihood on a Branch

When Both Ends are Known

$$P_{ij}(t) = \mathbb{P}\left[X(t) = j \mid X(0) = i\right] = (e^{t\mathbf{Q}})_{ij}$$

Computing $e^{t\mathbf{Q}}$ is costly!!

Likelihood on a Branch

When Both Ends are Known

$$P_{ij}(t) = \mathbb{P}\left[X(t) = j \mid X(0) = i\right] = (e^{t\mathbf{Q}})_{ij}$$

Computing $e^{t\mathbf{Q}}$ is costly!!

Otherwise

$$P_{\bullet j}(t) = \mathbb{P}\left[X(t) = j\right] = \sum_{i=1}^{k} \mathbb{P}\left[X(0) = i\right] (e^{t\mathbf{Q}})_{ij}$$

Likelihood on a Tree: Felsenstein's Pruning Algorithm

(Felsentstein, 1983)

Pruning Algorithm

- An example of dynamic programming
- Similar to Message Passing (but exact because we have a tree)
- Based on a recursion formula for conditional likelihoods on subtrees: $L_N(i)$: the probability to obtain the observed data at the tips given that the subtree rooted at N is in state i

Likelihood on a Tree: Felsenstein's Pruning Algorithm

(Felsentstein, 1983)

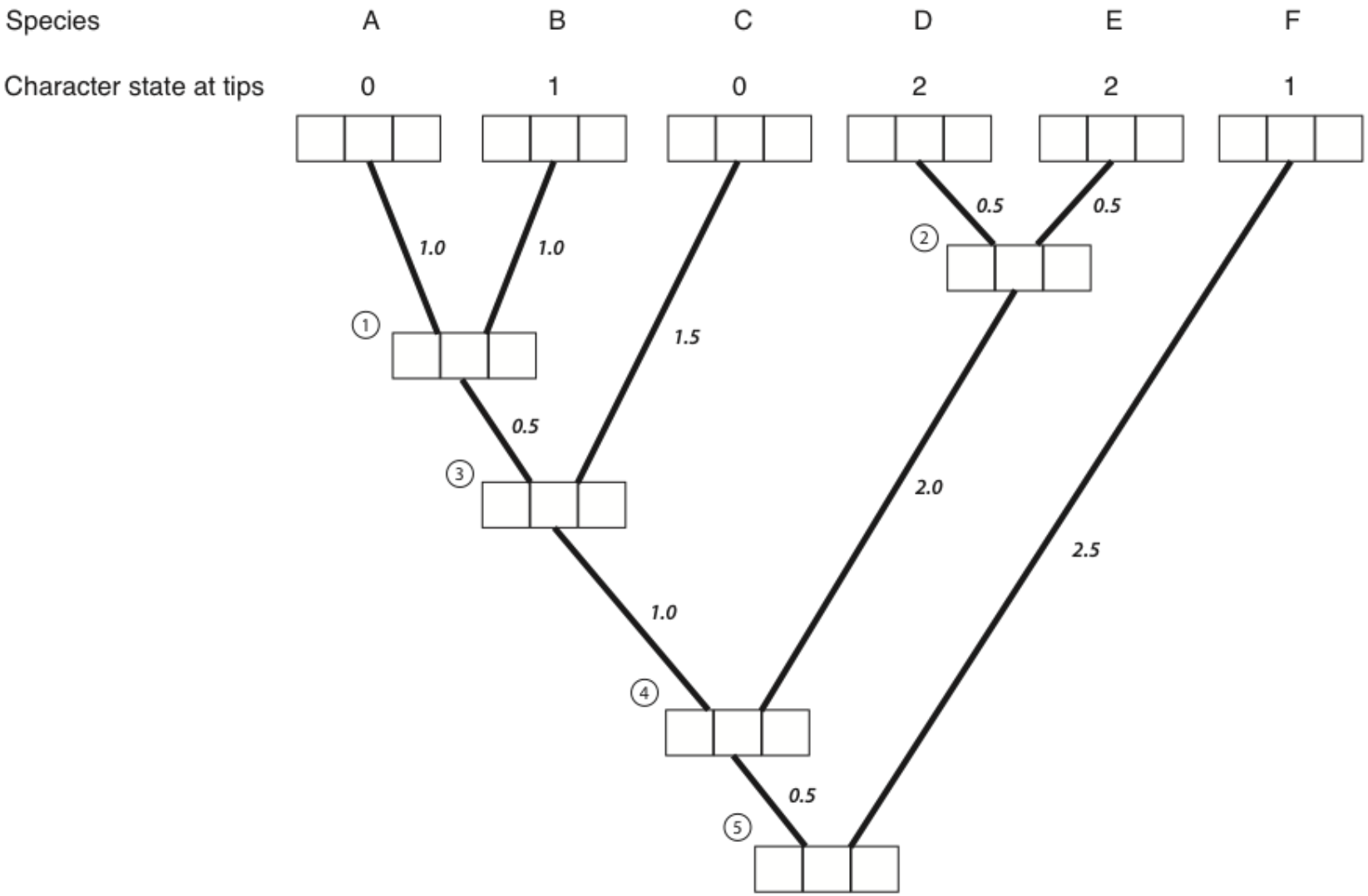
Pruning Algorithm

- An example of dynamic programming
- Similar to Message Passing (but exact because we have a tree)
- Based on a recursion formula for conditional likelihoods on subtrees: $L_N(i)$: the probability to obtain the observed data at the tips given that the subtree rooted at N is in state i

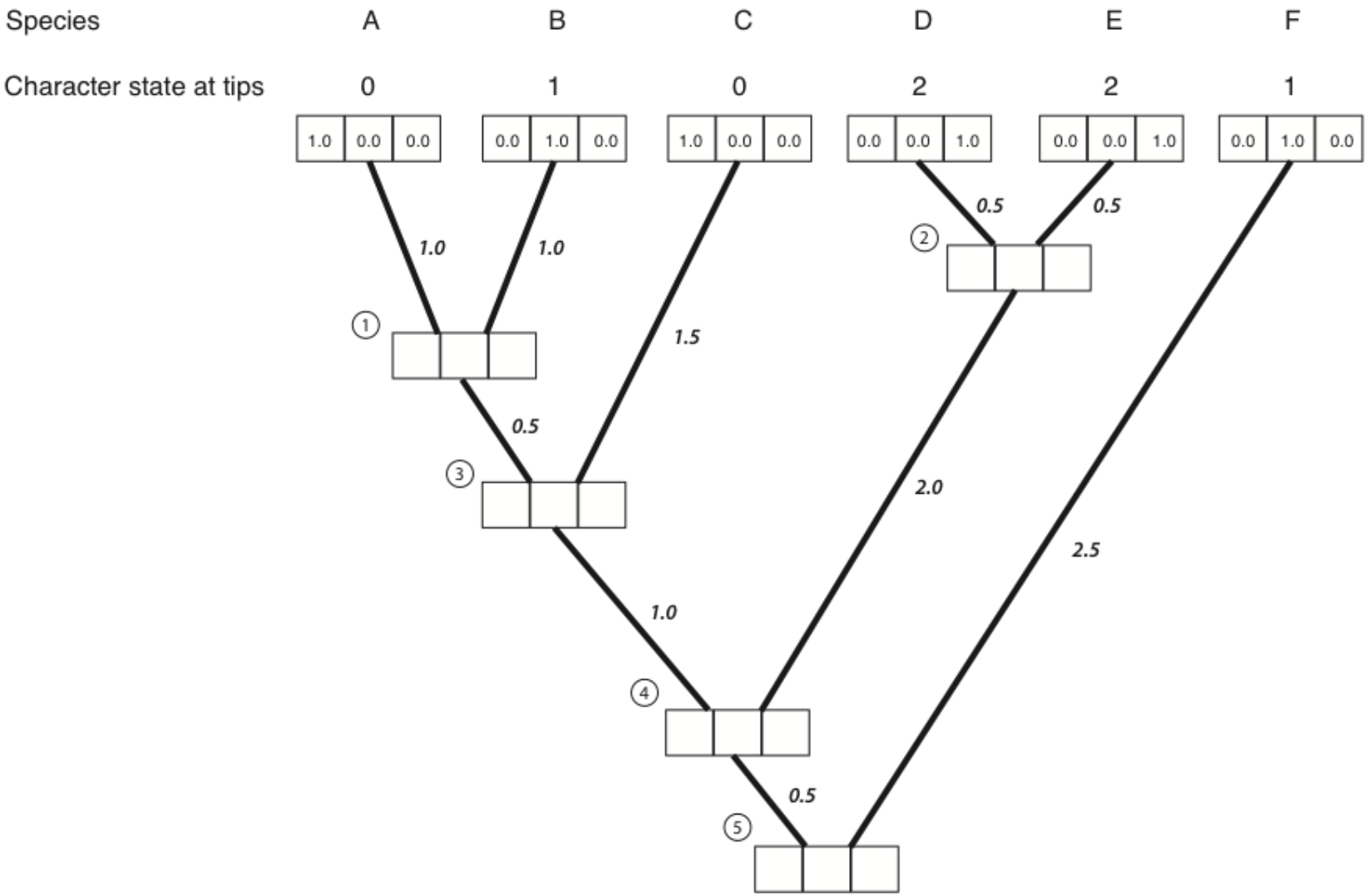
Recursion Formula

$$L_N(i) = \left(\sum_x \mathbb{P}\left[X_L = x \mid X_P = i\right] L_L(x) \right) \times \left(\sum_y \mathbb{P}\left[X_R = y \mid X_P = i\right] L_R(y) \right)$$

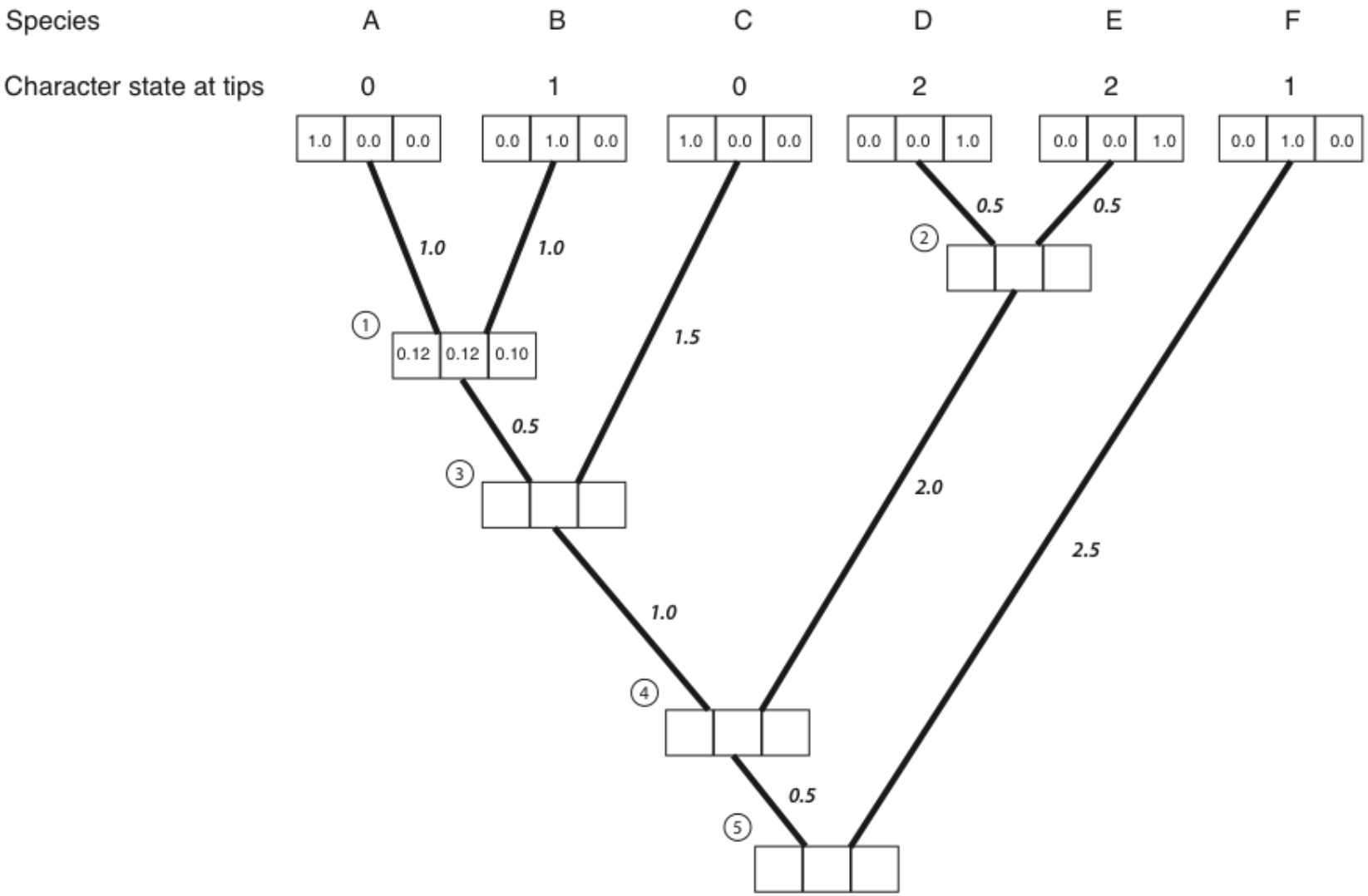
Likelihood on a Tree: Felsenstein's Pruning Algorithm



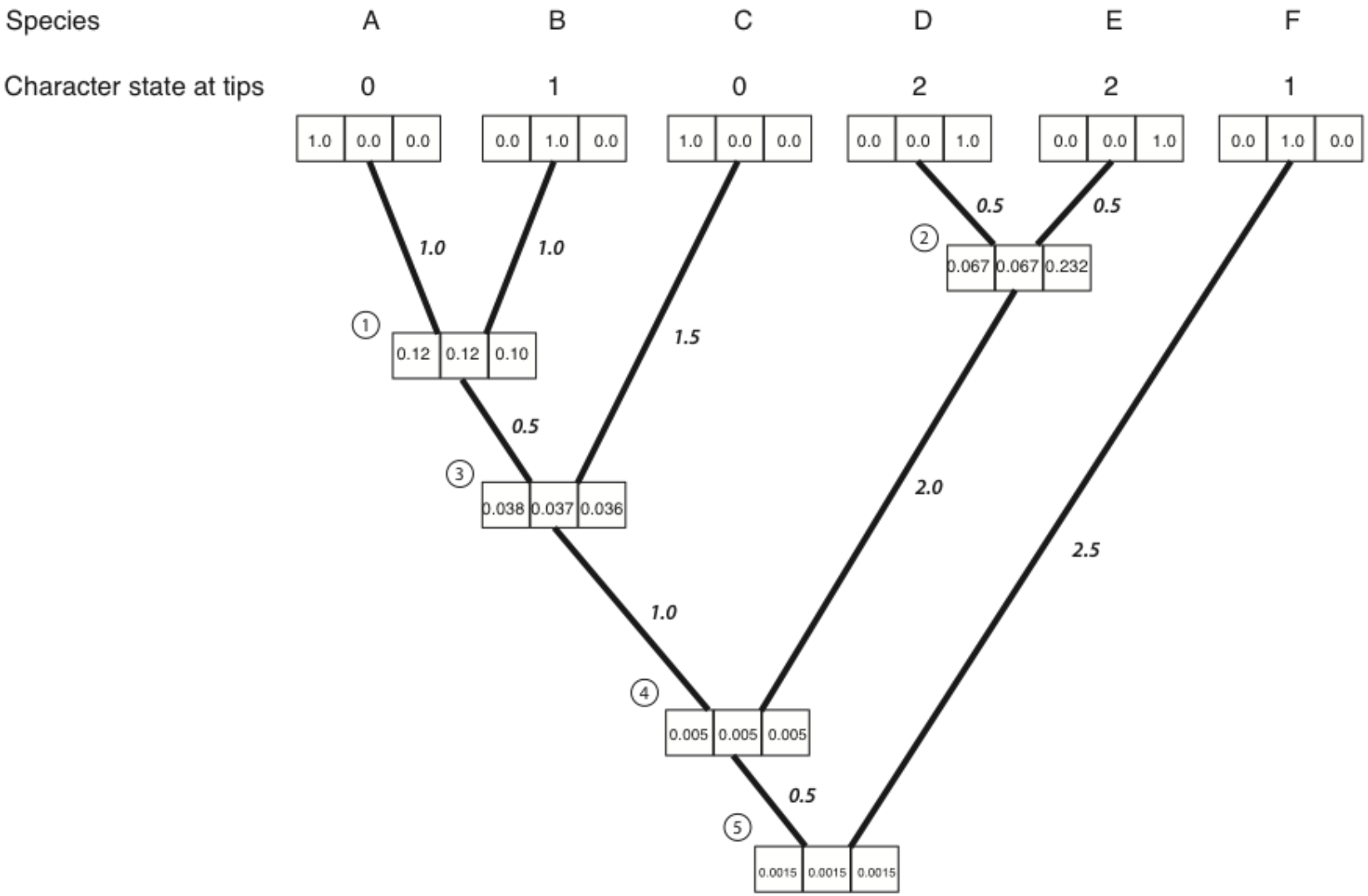
Likelihood on a Tree: Felsenstein's Pruning Algorithm



Likelihood on a Tree: Felsenstein's Pruning Algorithm



Likelihood on a Tree: Felsenstein's Pruning Algorithm



Likelihood on a Tree: Felsenstein's Pruning Algorithm

We can compute the likelihood and estimate (numerically) $\mathbf{Q}$

Outline

1. Motivation
2. Continuous Characters
3. Discrete Characters
   - Univariate Models
   - Multivariate Characters
4. Summary

Two Binary Characters (I) (Pagel, 1994)

For independent characters X_1 and X_2:

$$\mathbb{P}\left[X_1(t) = j, X_2(t) = l \mid X_1(0) = i, X_2(t) = k\right] = \mathbb{P}\left[X_1(t) = j \mid X_1(0) = i\right] \times \mathbb{P}\left[X_2(t) = l \mid X_2(t) = k\right]$$

Two Binary Characters (I) (Pagel, 1994)

Independent Characters

Let

$$\mathbf{Q}_A = \begin{array}{c} \begin{matrix} \text{a} & \text{A} \end{matrix} \\ \begin{bmatrix} \bullet & q_A \\ q_a & \bullet \end{bmatrix} \end{array} \begin{matrix} \text{a} \\ \text{A} \end{matrix} \quad \text{and} \quad \mathbf{Q}_B = \begin{array}{c} \begin{matrix} \text{b} & \text{B} \end{matrix} \\ \begin{bmatrix} \bullet & q_B \\ q_b & \bullet \end{bmatrix} \end{array} \begin{matrix} \text{b} \\ \text{B} \end{matrix}$$

Then

$$\mathbf{Q}_{AB} = \begin{array}{c} \begin{matrix} \text{ab} & \text{aB} & \text{Ab} & \text{AB} \end{matrix} \\ \begin{bmatrix} \bullet & q_B & q_A & \cdot \\ q_b & \bullet & \cdot & q_A \\ q_a & \cdot & \bullet & q_B \\ \cdot & q_a & q_b & \bullet \end{bmatrix} \end{array} \begin{matrix} \text{ab} \\ \text{aB} \\ \text{Ab} \\ \text{AB} \end{matrix}$$

Two Binary Characters (II) (Pagel, 1994)

Dependent Characters

In general for dependent characters, we don't have

$$q_{Ab,AB} = q_{ab,aB} = q_B \quad \text{and} \quad q_{AB,Ab} = q_{aB,ab} = q_b$$

$$q_{ab,Ab} = q_{aB,AB} = q_A \quad \text{and} \quad q_{AB,aB} = q_{Ab,ab} = q_a$$

Two Binary Characters (II) (Pagel, 1994)

Dependent Characters

In general for dependent characters, we don't have

$$q_{Ab,AB} = q_{ab,aB} = q_B \quad \text{and} \quad q_{AB,Ab} = q_{aB,ab} = q_b$$
$$q_{ab,Ab} = q_{aB,AB} = q_A \quad \text{and} \quad q_{AB,aB} = q_{Ab,ab} = q_a$$

Special Dependence: b/B depends on a/A

If character b/B depends on character a/A

$$q_{Ab,AB} \neq q_{ab,aB} \quad \text{and/or} \quad q_{AB,Ab} \neq q_{aB,ab}$$

Testing correlation

One can test for evolutionary correlation by testing

- H_0: the rates satisfy the 4 previous equalities (independence)
- H_1: they don't (dependence)

Using Likelihood Ratio Test (4 parameters under H_0, 8 in general under H_1) or integrating the previous model in a Bayesian framework.

Tree Thinking

Evolution Matters!!

Tree Thinking

Evolution Matters!!

Modeling Evolution

- Many (many) models exist for the coevolution of discrete and/or continuous traits.
- Correcting for phylogeneically induced correlation is possible (and should be done).
- Prevents you from drawing spurrious conclusions from the data.

References I

Aristide, L., dos Reis, S. F., Machado, A. C., Lima, I., Lopes, R. T., and Perez, S. I. (2016). Brain shape convergence in the adaptive radiation of New World monkeys. *Proceedings of the National Academy of Sciences*, 113(8):2158–2163.

Bastide, P., Ané, C., Robin, S., and Mariadassou, M. (2018). Inference of adaptive shifts for multivariate correlated traits. *Systematic Biology*, page syy005.

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):1–15.

Felsentstein, J. (1983). Statistical inference of phylogenies. *J.R. Statist. Soc.*, 146(3):246–272.

Pagel, M. (1994). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings: Biological Sciences*, 255(1342):37–45.

Photo Credits:
- Miguelrangeljr - Own work, CC BY-SA 3.0, `https://commons.wikimedia.org/w/index.php?curid=28294644`
- Steven G. Johnson - Own work, CC BY-SA 3.0, `https://commons.wikimedia.org/w/index.php?curid=4372777`
- Braboowi at the English language Wikipedia, CC BY-SA 3.0,
`https://commons.wikimedia.org/w/index.php?curid=7069103`
- Xiphophorus Genetic Stock Center, Texas State University,
`http://www.xiphophorus.txstate.edu/resources/galleries/comprehensive.html`
- ”Lonesome George in profile” by Mike Weston - Flickr: Lonesome George 2. Licensed under CC BY 2.0 via Wikimedia Commons

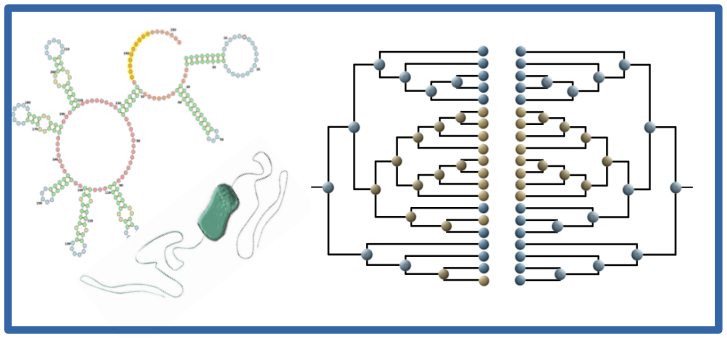
Detection of coevolution between small regulatory RNAs and coding genes in a bacterial genus

With many slides/figures courtesy of F. Cerutti

Our work

Develop a **phylogenomics strategy** to study **bacterial sRNAs and coding genes evolution and coevolution**

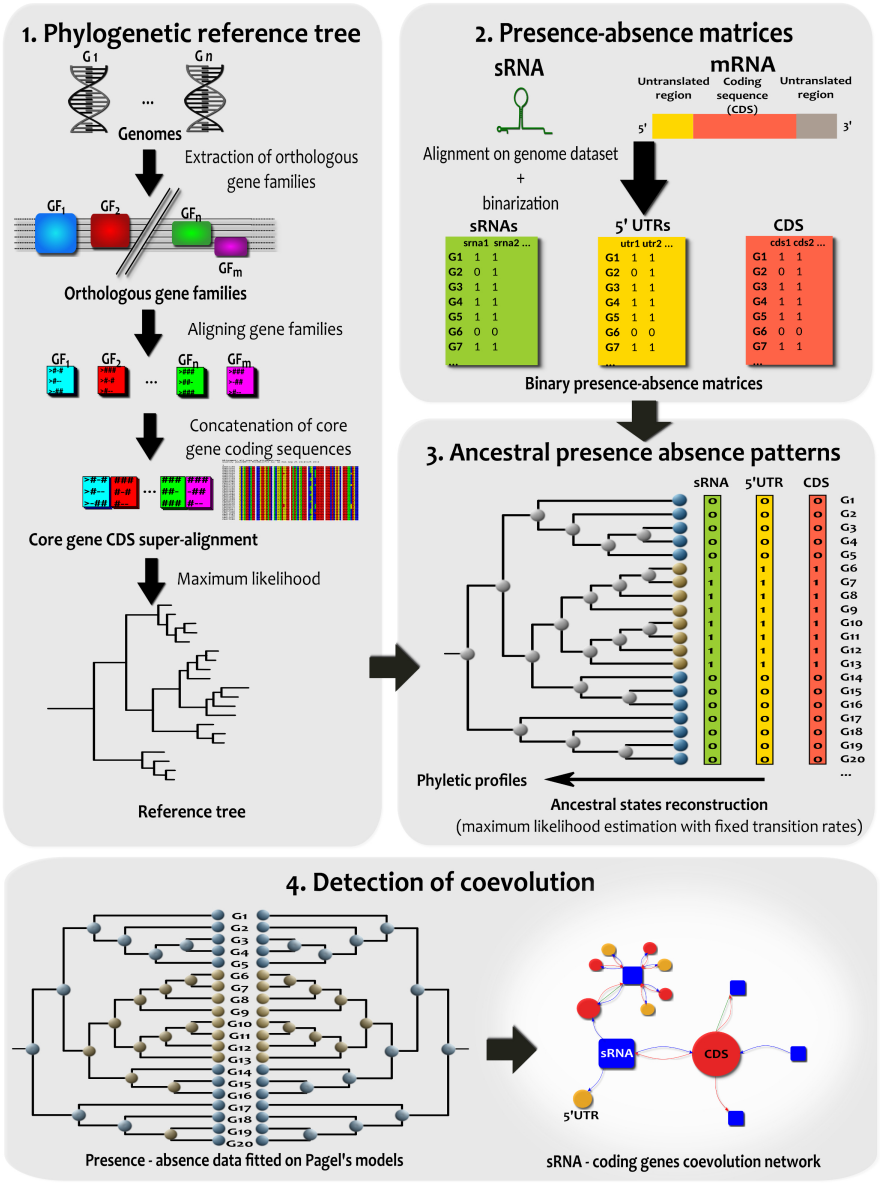
Input
- *a sRNA dataset in a reference genome*
- *a list of annotated genomes of a bacterial genus*



Output: *a list of sRNAs and coding gene regions coevolution pairs + a coevolution network*

Apply it on a dataset of ***Listeria*** **sRNAs and coding gene regions (5'UTR and CDS)**

Results: Strategy overview



Phylogenomics based approach

4 main steps to detect coevolution

*Snakemake** workflow including Python and R scripts and many other tools!

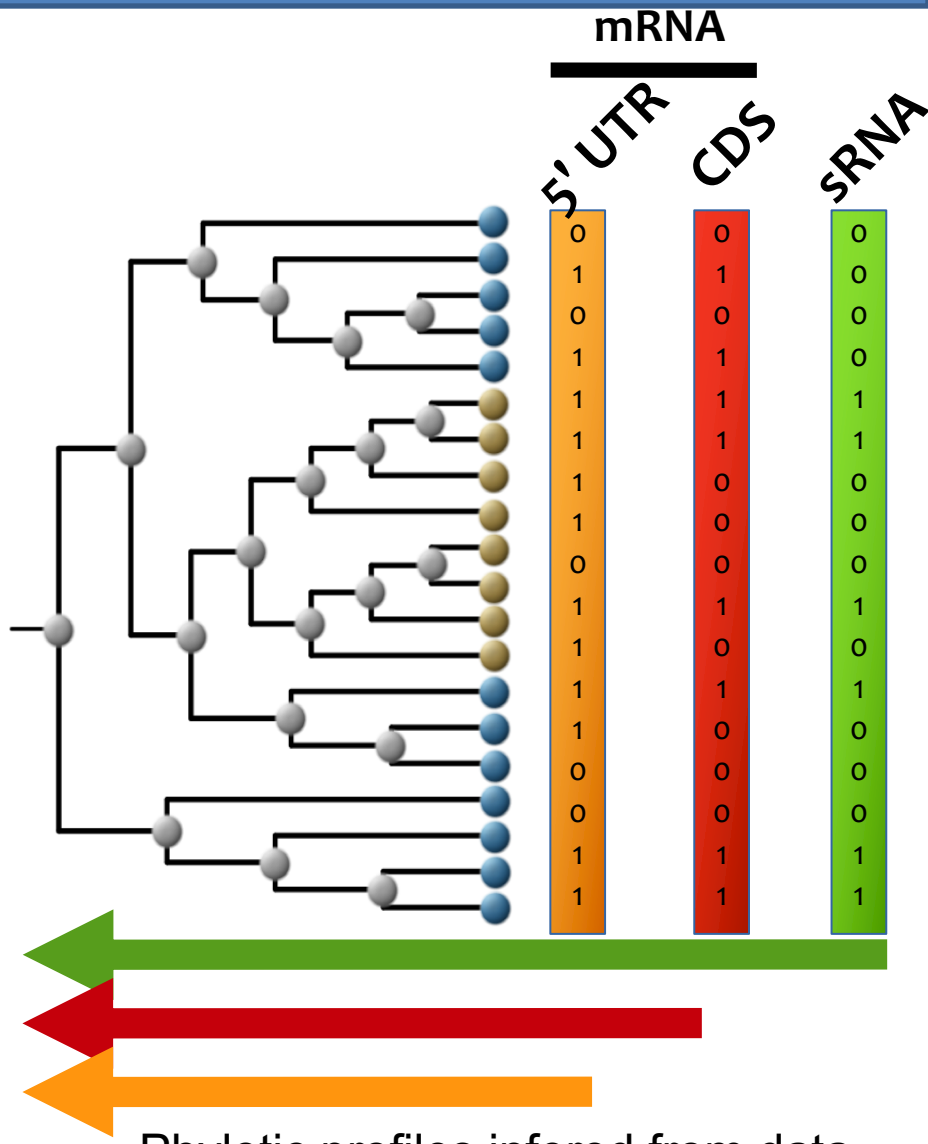
*Köster J, Rahmann S. Snakemake–a scalable bioinformatics workflow engine. Bioinformatics 2012;28:2520–2.

Method: presence/absence states reconstruction

Maximum likelihood ancestral states reconstruction [1] :

- **More likely profiles** built using *'rayDISC'* function from *'corHMM'* R package **[2]**
- **Gain / loss events** predicted

Phyletic profiles computed for each **sRNA** and mRNA **CDS** and **5'UTR**



[1] Mark Pagel, Syst. Biol., 1999
[2] Beaulieu JM et al., Syst. Biol., 2013

Coevolution detection principle

Based on **Pagel[1] method**

- **Takes into account Phylogeny** (**a reference tree**) and 3 binary traits: presence/absence of a **sRNA, a 5'UTR and a CDS**

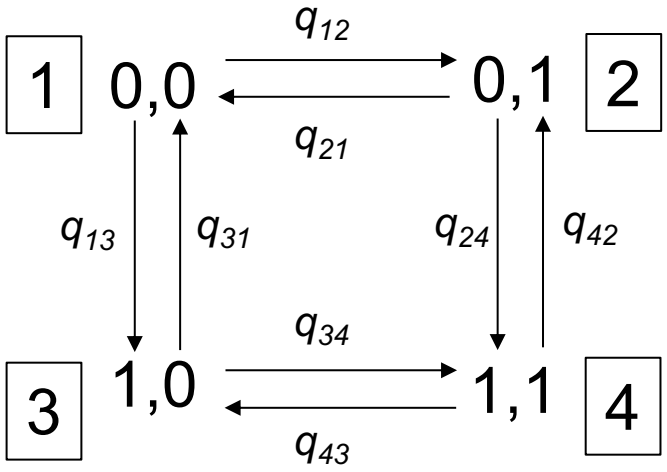
- Use **continuous-time Markov models and ancestral states** to describe trait evolution and compare **the statistical likehood** of two models:
    - One in which **two traits are allowed to evolved independently** on the tree (H_0)
    - One in which **two traits are allowed to evolve in a correlate** fashion (H_1)

[1]Pagel M. 2005 Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters. Proceedings of the Royal Society B: Biological Sciences.

Trait evolution and coevolution modeling

Pagel models principle

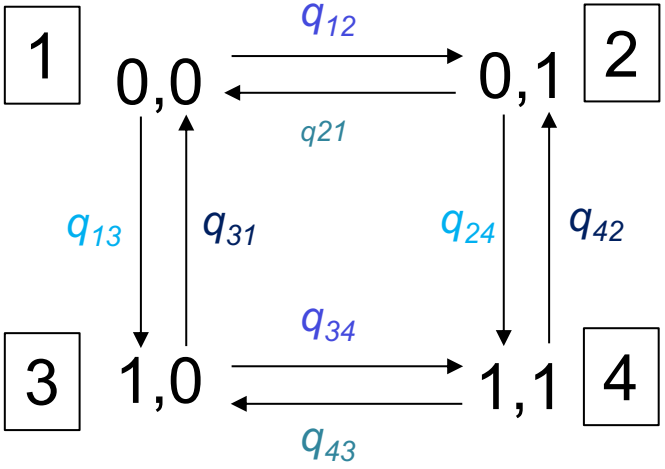
- Two binary traits can produce **4 different pairs of states**, corresponding to the **pairings of presence or absence in two genomic elements**



q_{ij}: rates of transitions between two states of two genes

- These rates of transitions q_{ij} are inferred from the implied **number of times the events represented by the rate coefficients** have occurred on the **reference tree**

Pagel models principle



q_{ij}: rates of transitions between two states of two genes

- **Model 1 (H_0) : If two traits evolve independently**, the rate of change between two states of a gene will not depend upon the other gene is present or absent, *i.e.*

  $q_{1,2}=q_{3,4}$ & $q_{1,3}=q_{2,4}$ & $q_{4,2}=q_{3,1}$ & $q_{4,3}=q_{2,1}$ & $q_{1,2}=q_{3,4}$ => **4 parameters**
- **Model 2 (H_1): If two traits have correlated evolution**, some of these pairs of transition rates differ **=> 8 parameters**

Pagel models principle

- Likelihoods of the two models are compared (LRT)

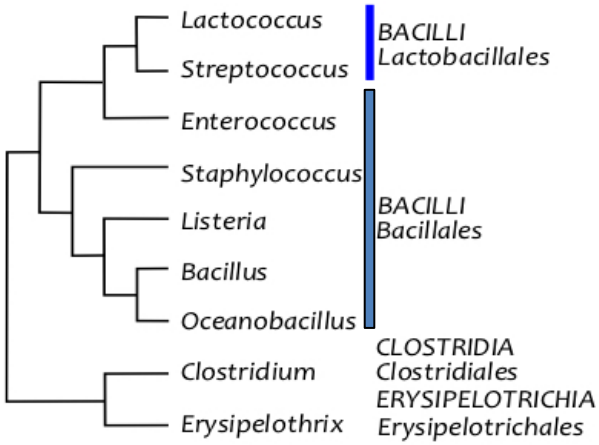
$$LR = -2 \times \log\left(\frac{maxL(I)}{maxL(D)}\right)$$

  - Pvalue corrected for multiple testing using Benjamini-Hochberg (BH) **false discovery rate correction**
  - Pvalue+BH < 0.01

The *Listeria* genus

- ***Gram+ bacteria**, firmicutes division, Bacilli class*
- Includes **foodborn opportunistic pathogens** infecting human and cattle (listeriosis)
- Model for **host-pathogen interaction**
- Available sets in *L.monocytogenes EGD-e*:

  sRNAs (tiling array, RNAseq) and **5'UTR regions**



Firmicutes phylum

Listeria invading an epithelial cell

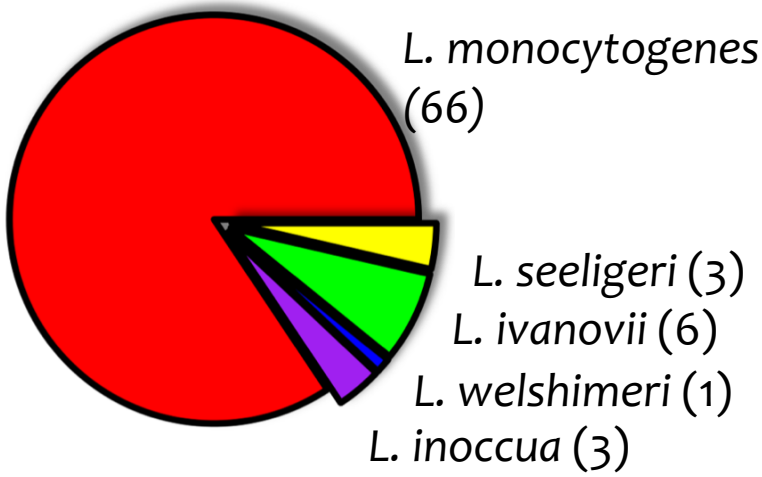
Datasets

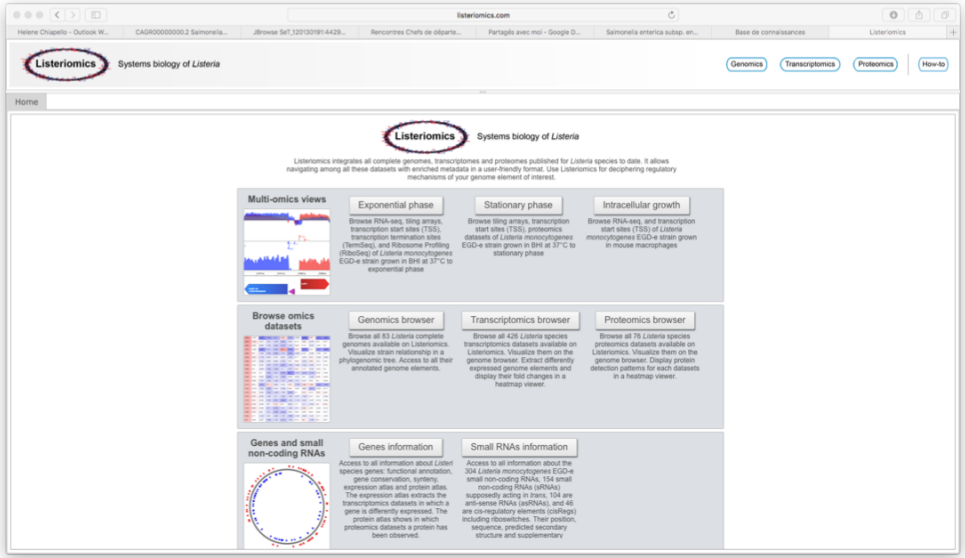
(1) 79 complete and draft ***Listeria* genomes**

- A reference strain *L. monocytogenes EGD-e*
- *5 different species*

(2) A cleaned dataset of *L. monocytogenes* EGDe regulatory **sRNA**

- 125 regulatory sRNA experimentally identified in EGD-e, source : *Listeriomics*[1], small ORF excluded
- After cleaning and merging : a total of **112 sRNA (97 singletons and 15 sRNA loci)**

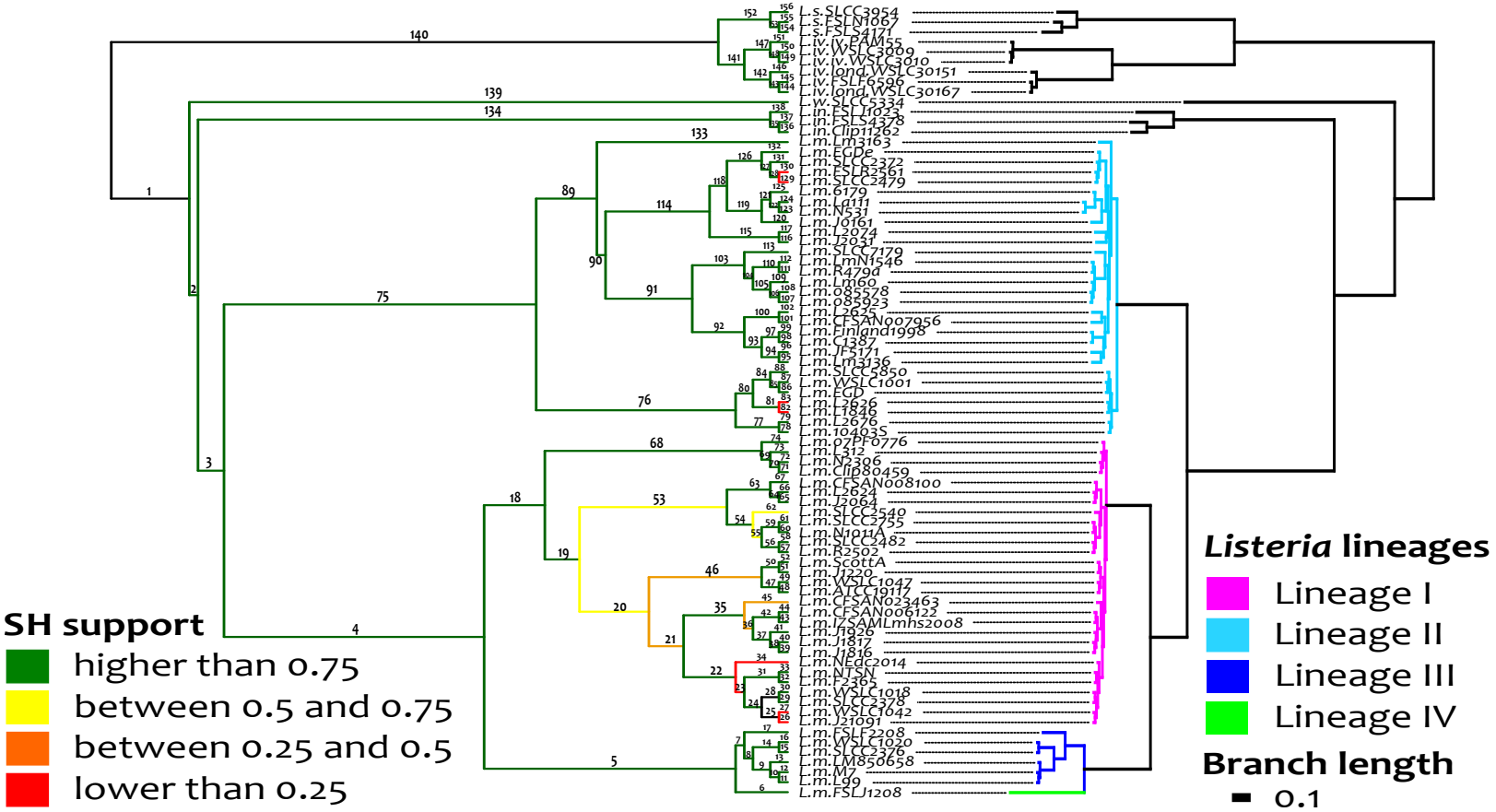




[1]Becavin *et al.*, Listeriomics: An Interactive Web Platform for Systems Biology of Listeria ». mSystems 2017. www.listeriomics.com

The *Listeria* reference tree

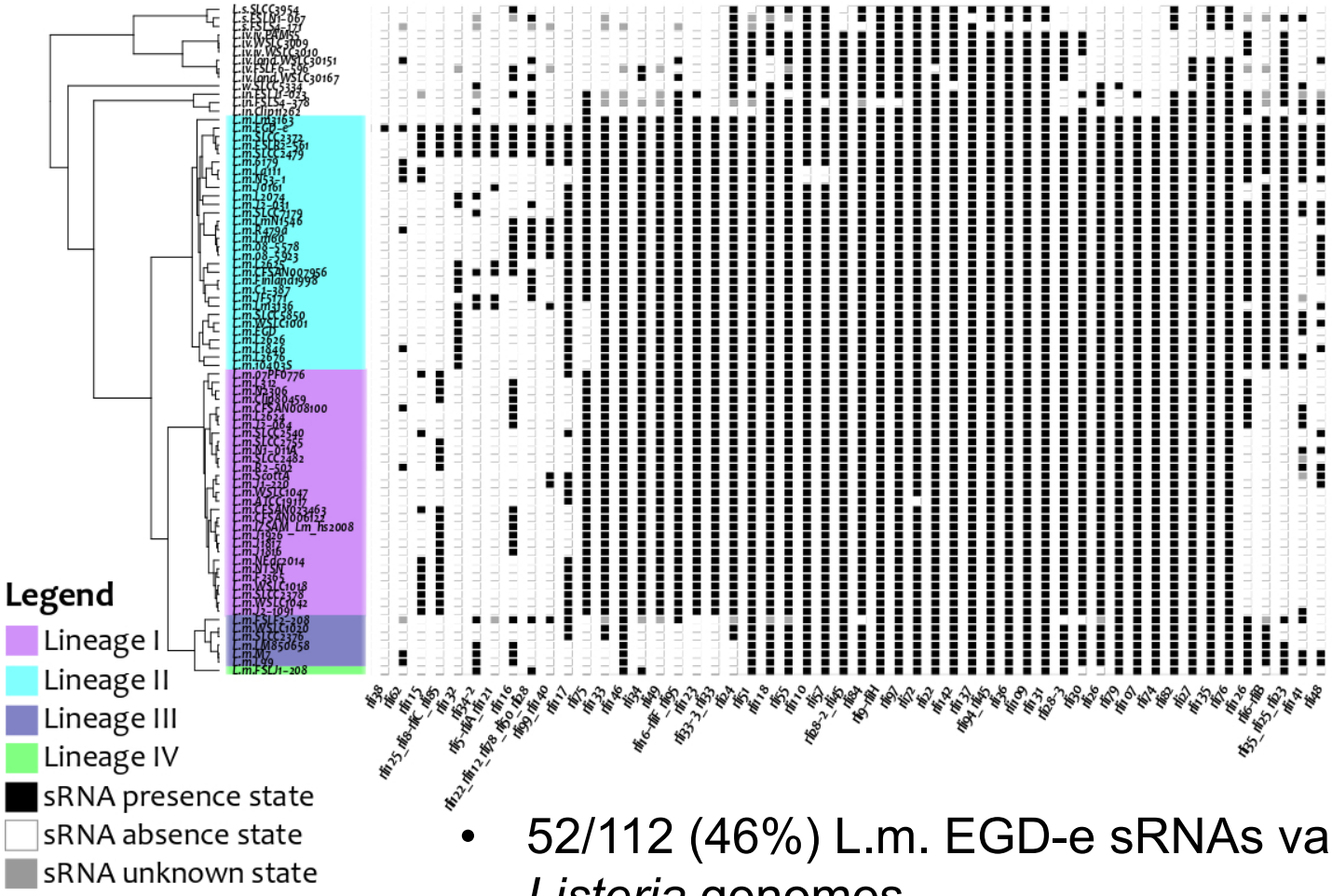
Maximum Likelihood tree (GTR model, superalignement of 1399 core genes)



- Most clades are well supported
- Consistent with known *Listeria* phylogenies and lineages

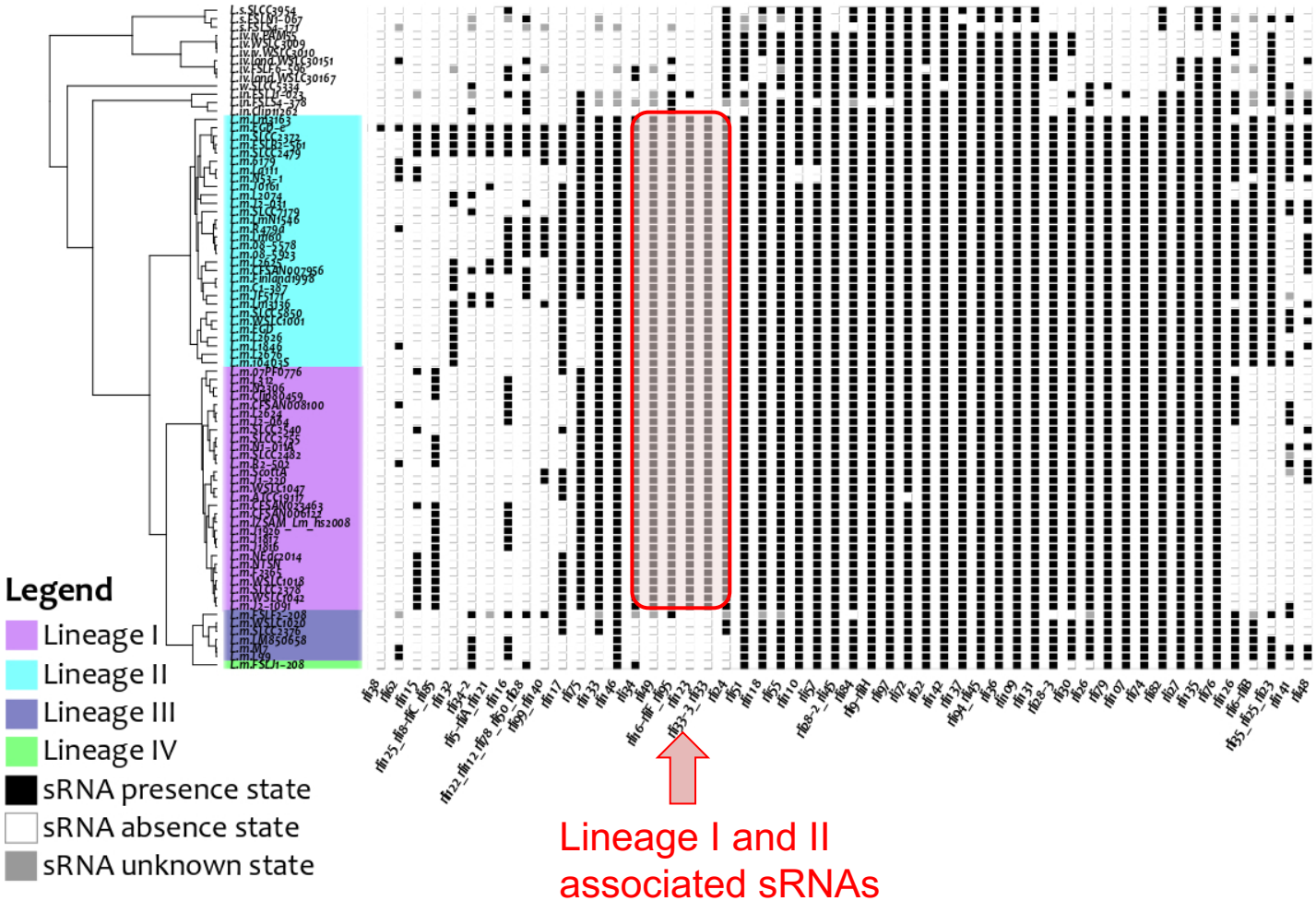
[1] Renato H. Orsi et al. , International Journal of Medical Micobiology, 2011

Phylogenetic distribution of *L. m EGD-e* sRNAs

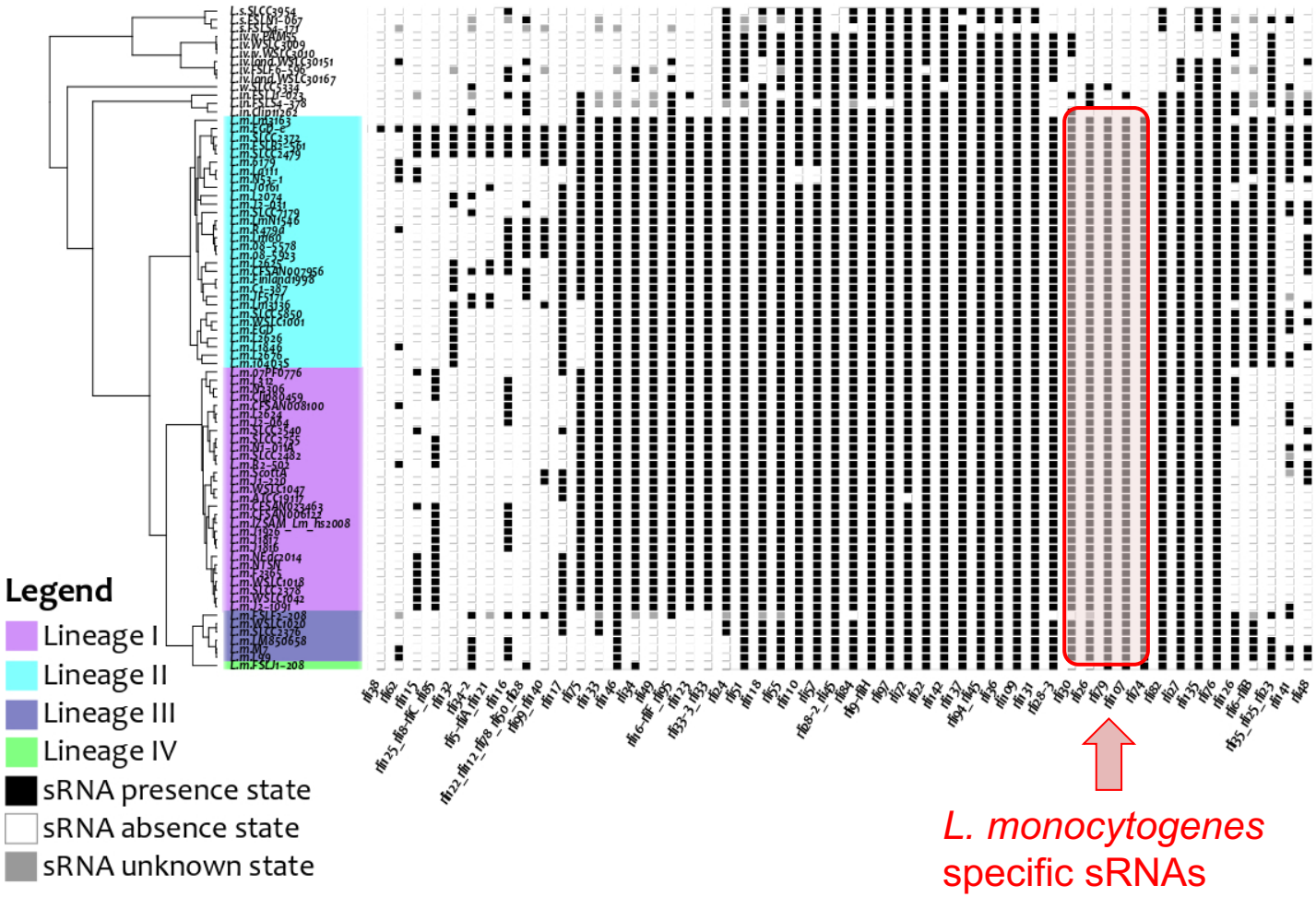


- 52/112 (46%) L.m. EGD-e sRNAs variable in *Listeria* genomes

A link between sRNA content and strain evolutionary distance ?

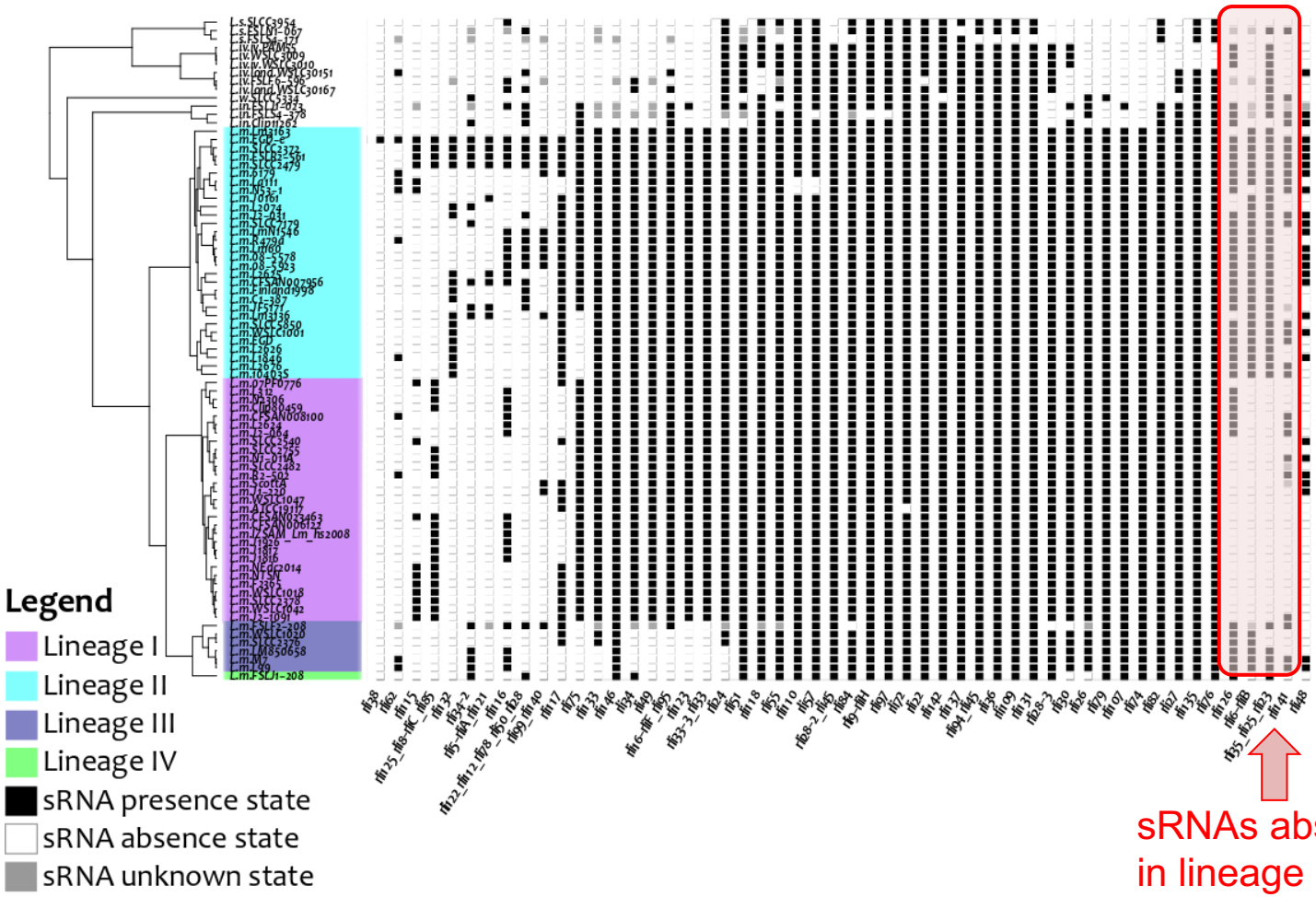


A link between sRNA content and strain evolutionary distance ?



L. monocytogenes specific sRNAs

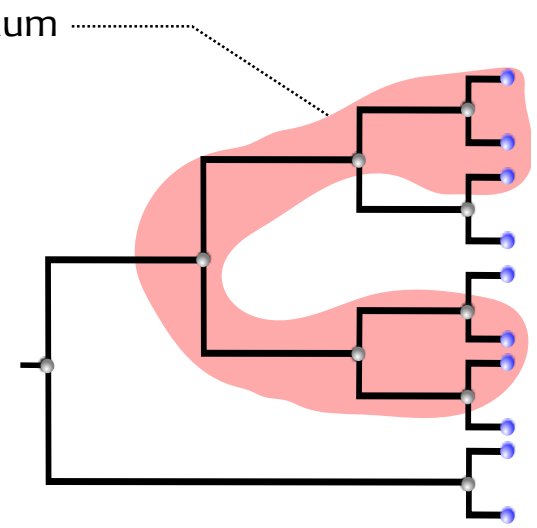
A link between sRNA content and strain evolutionary distance ?



The Listeria sRNA evolution patterns

52 sRNAs with **44 different reconstructed presence/absence profiles**

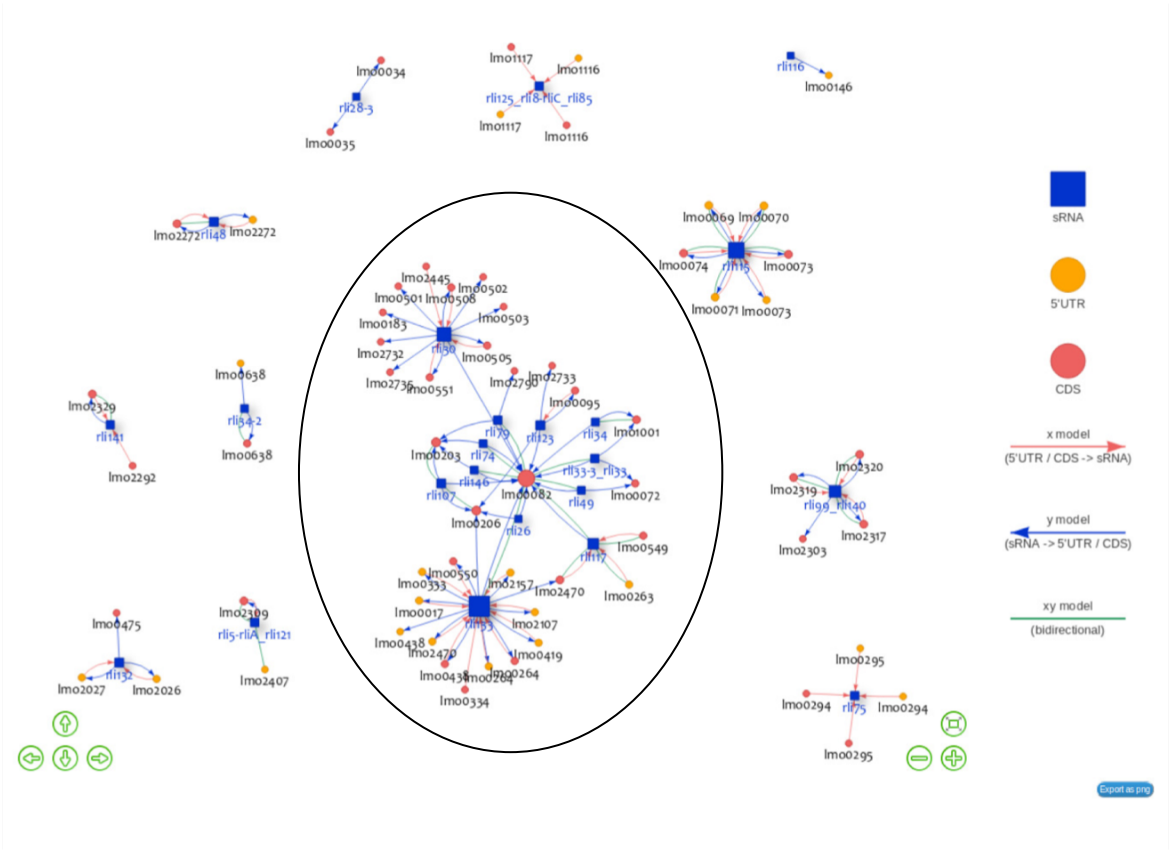
- 48/52 (92%) sRNAs exhibit **paraphyletic profiles**
- 47/52 (90%) sRNA were infered **present at the *Listeria* reference tree root**



http://genoweb.toulouse.inra.fr/Listeria_sRNA

The Listeria sRNA-coding genes coevolution network

23/52 sRNAs exhibit significant coevolutionary relationships with coding genes



136 (sRNA -5'UTR/CDS) significant coevolving pairs:

- 23 sRNAs
- 52 coding genes (23 5'UTR and 39 CDS)

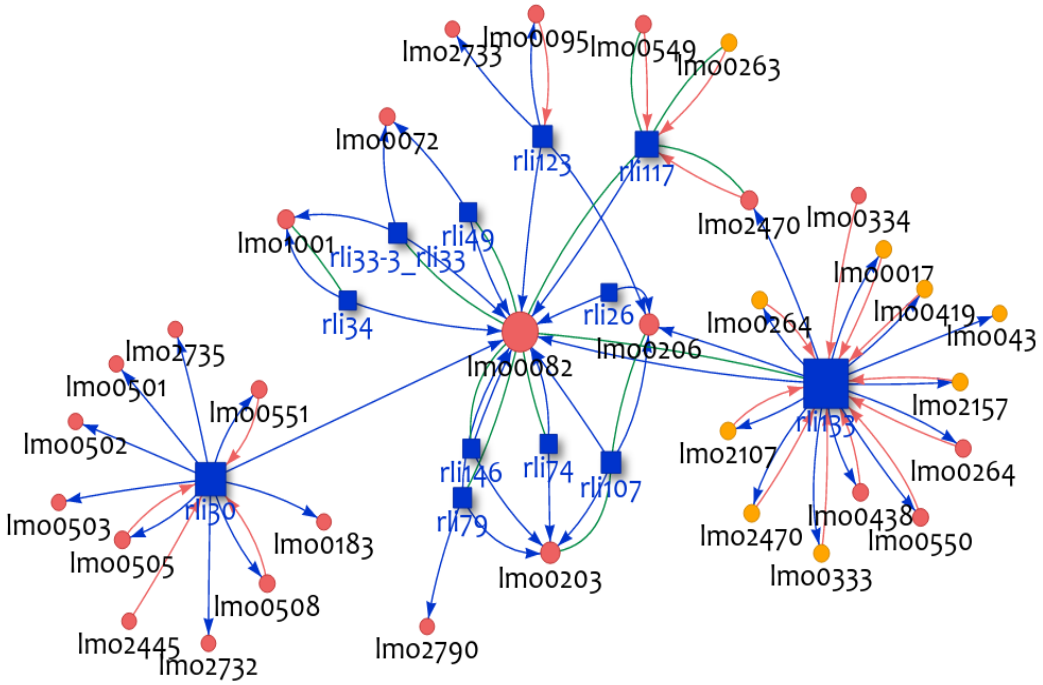
12 clusters including

- 11 individual clusters (1 sRNA) coevolving with a close gene (<8 kb)
- A hub of 12 sRNAs

http://genoweb.toulouse.inra.fr/Listeria_sRNA

The Listeria sRNA-coding genes coevolution network

23 putative sRNAs exhibit significant coevolutionary relationships with 52 coding genes



Presence of a a hub of 12 sRNAs coevolving with

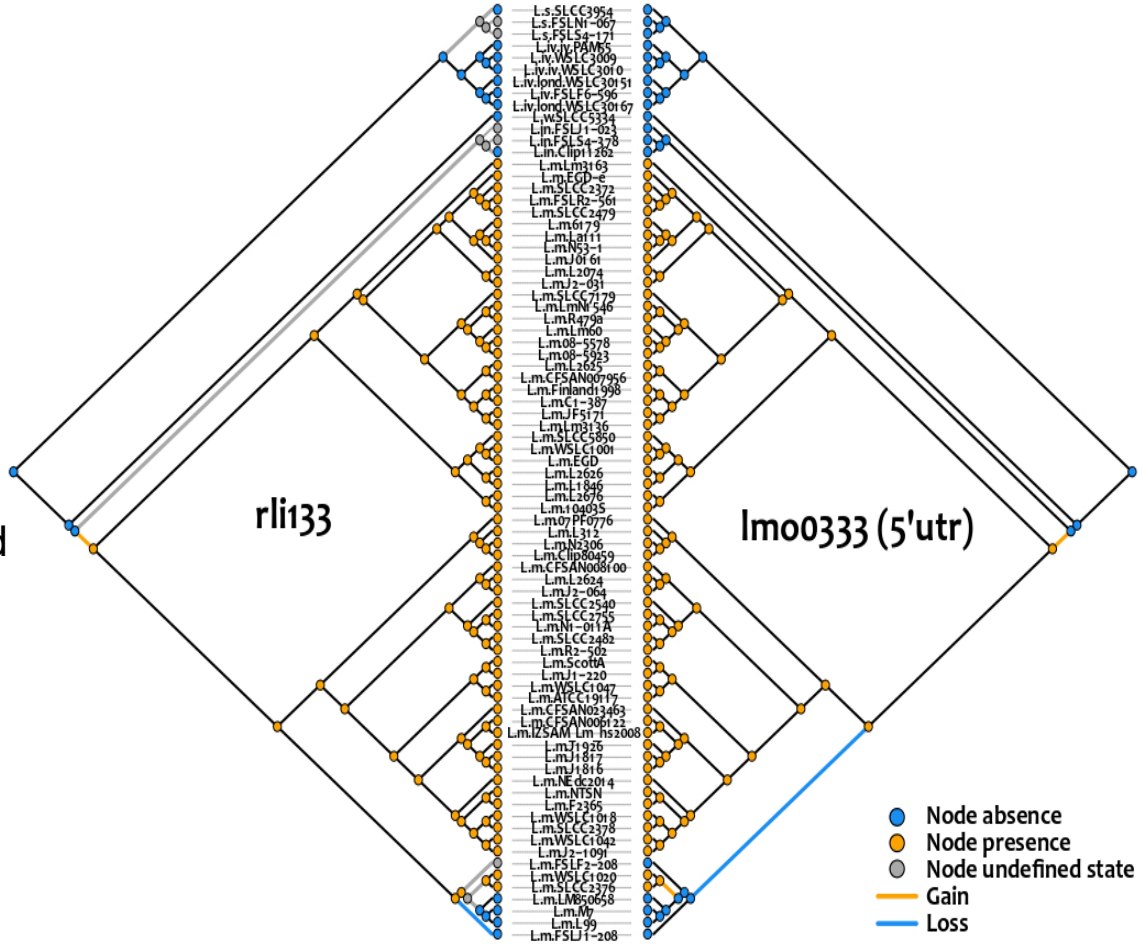
- mainly distantly related genes (>40kb)
- mainly genes related to **cell enveloppe** (internalins...), **secondary metabolism** and **pathogenicity**.

http://genoweb.toulouse.inra.fr/Listeria_sRNA

Case study: rli133

Rli133 presents coevolutionary relationships with 12 coding genes, 8 5' UTRs and 7 CDS regions including:

- 3 internalins:
  inlE → required for host tissue colonization1
  inlP → specific of placental tissue colonization2
  inlI (lmo0333) -> *unknown function*
- 1 virulence factor: ***IntA3***
- 1 protein of LIPI-1 pathogenic island and involved in survival in macrophage: ***orfX4***
- 1 stress-response gene involved in septum formation: ***sepA***



Case study: rli133

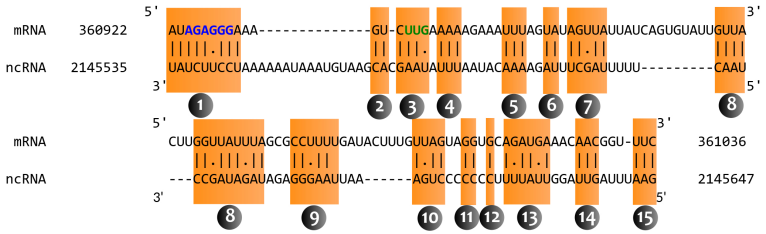
Does sRNA-coding gene coevolution mean putative physical interaction?

- *In silico* interaction predictions using a combination of tools included in the sRNA-TaBac server [1]
- **9 out of 12 coevolving genes** exhibit direct RNA-RNA interacting regions with rli133 compatible with a negative regulatory mechanism

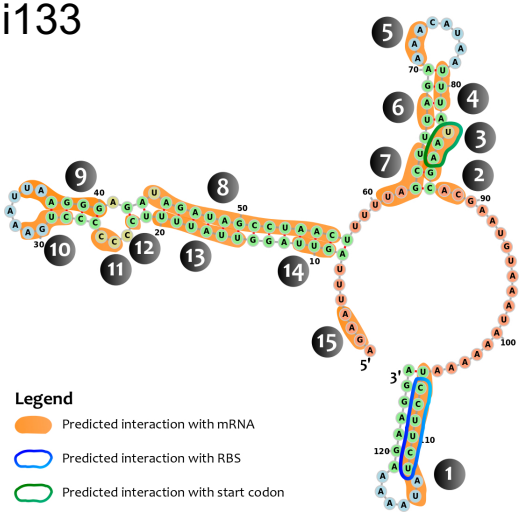
[1] http://srnatabac.toulouse.inra.fr:8080/

Case study: rli133

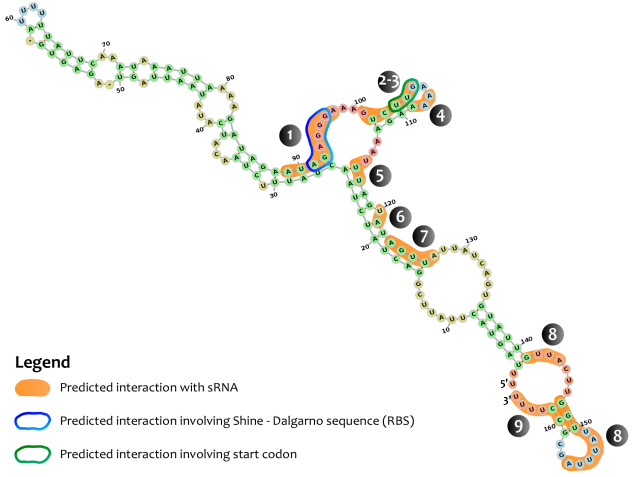
An example of predicted interaction betwenn rli133 and lmo0333 (inlL) 5’UTR region



rli133



inlL (5’UTR+60nt)



Conclusion

- A first **successful and generic strategy** proposed to study sRNA and coding genes evolution and coevolution
- Analysis of the *Listeria* sRNA-coding gene coevolving network enlights **a hub including many genes related to cell enveloppe, virulence and stress response**
- **Negative regulatory interaction mechanism of mRNA by sRNA** could be predicted for several coevolving groups
- The strategy is implemented in a **pipeline** (not yet packaged)

Perspectives

- **Strategy**
  - Adapt the approach
    - ❖ to different evolutionary scales
    - ❖ to higher number of genomes
  - Take into account paralogy of elements
  - Detect co-evolution in elements conserved in all genomes
  - Package and distribute the workflow

- **Application**
  - Evaluate the approach
    - ❖ on other type of elements
    - ❖ on other organisms

THANKS

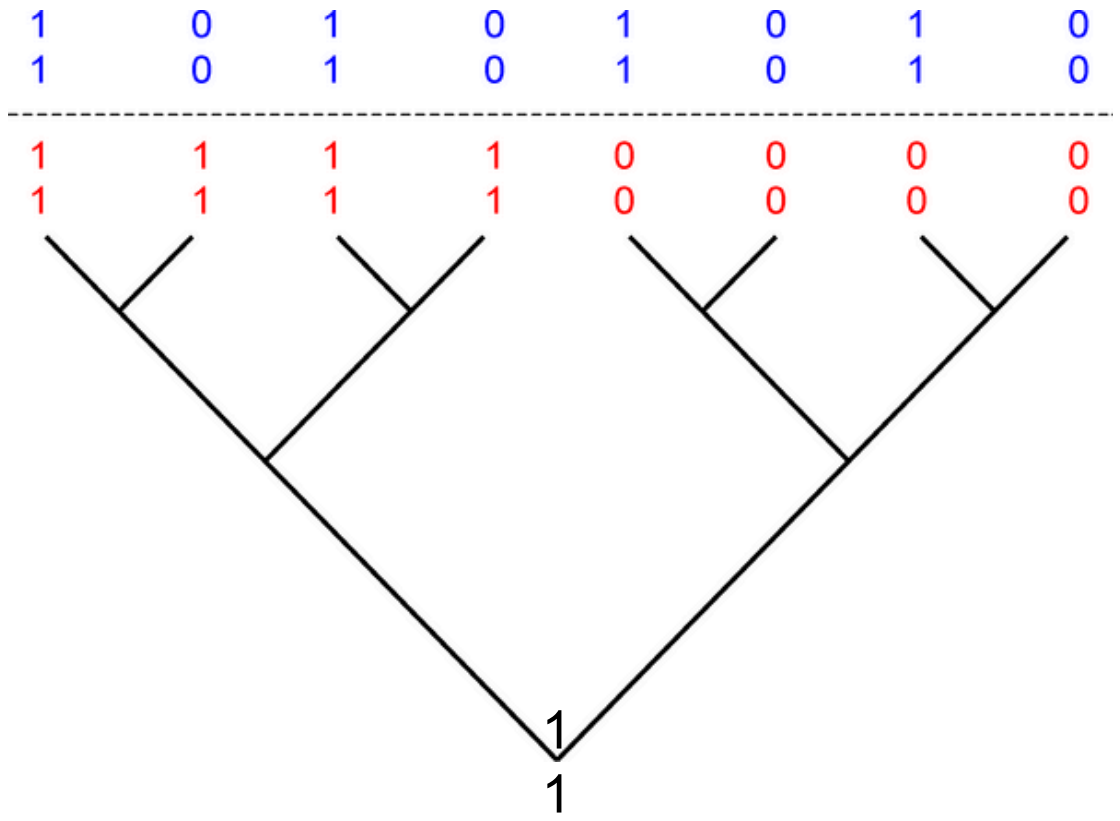
Franck Cerutti
Christine Gaspin
Claire Hoede
Ludovic Mallet
et al.

Cerutti F, Mallet L, Painset A, Hoede C, Moisan A, Bécavin C, Duval M, Dussurget O, Cossart P, Gaspin C, Chiapello H. Unraveling the evolution and coevolution of small regulatory RNAs and coding genes in Listeria. BMC Genomics. 2017 Nov 16;18(1):882.

Bacnet project coordinators : P. Cossart (I. Pasteur) & C. Gaspin (MIAT partner)

Coevolution: why do we need to take account phylogeny

Figure 1. Across-Species Correlation Confuses Shared Inheritance with Correlated Evolution but Phylogenetic Method Does Not



4 events

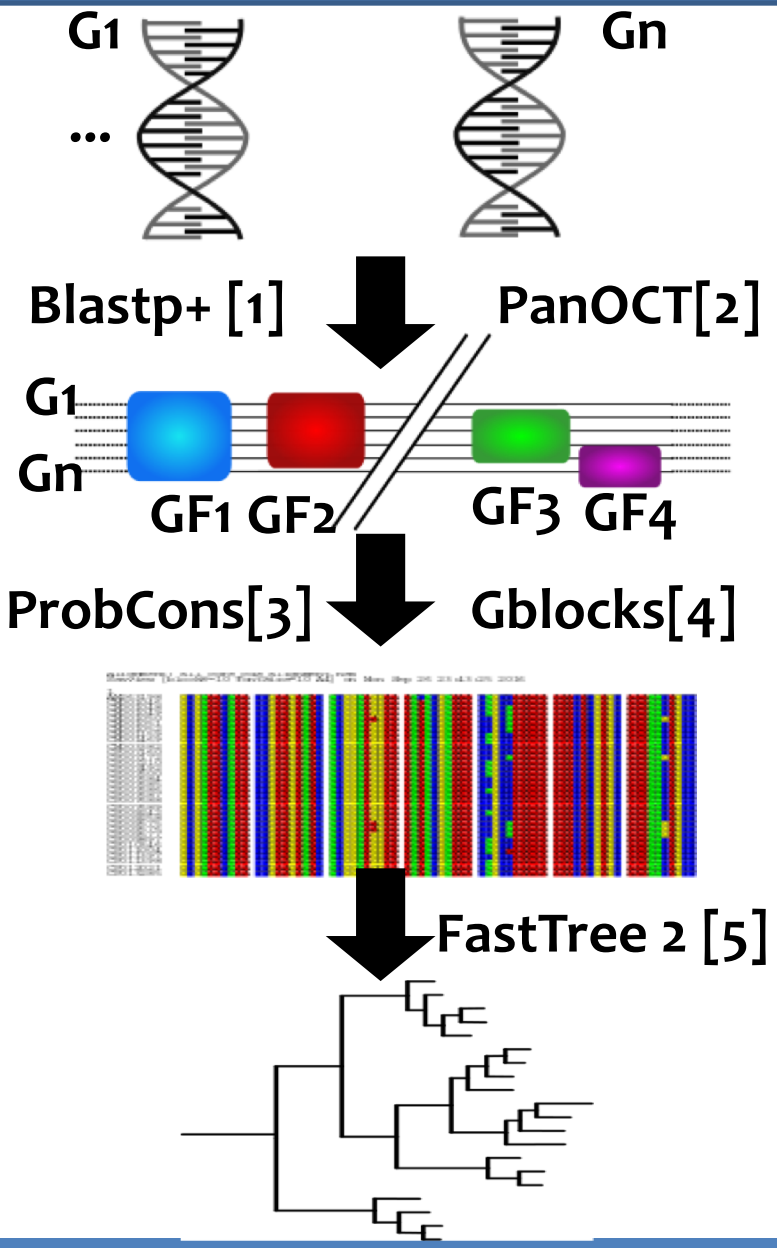
1 loss event

Barker D, Pagel M (2005) Predicting Functional Gene Links from Phylogenetic-Statistical Analyses of Whole Genomes. PLOS Computational Biology 1(1) http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.0010003

PLOS | COMPUTATIONAL BIOLOGY

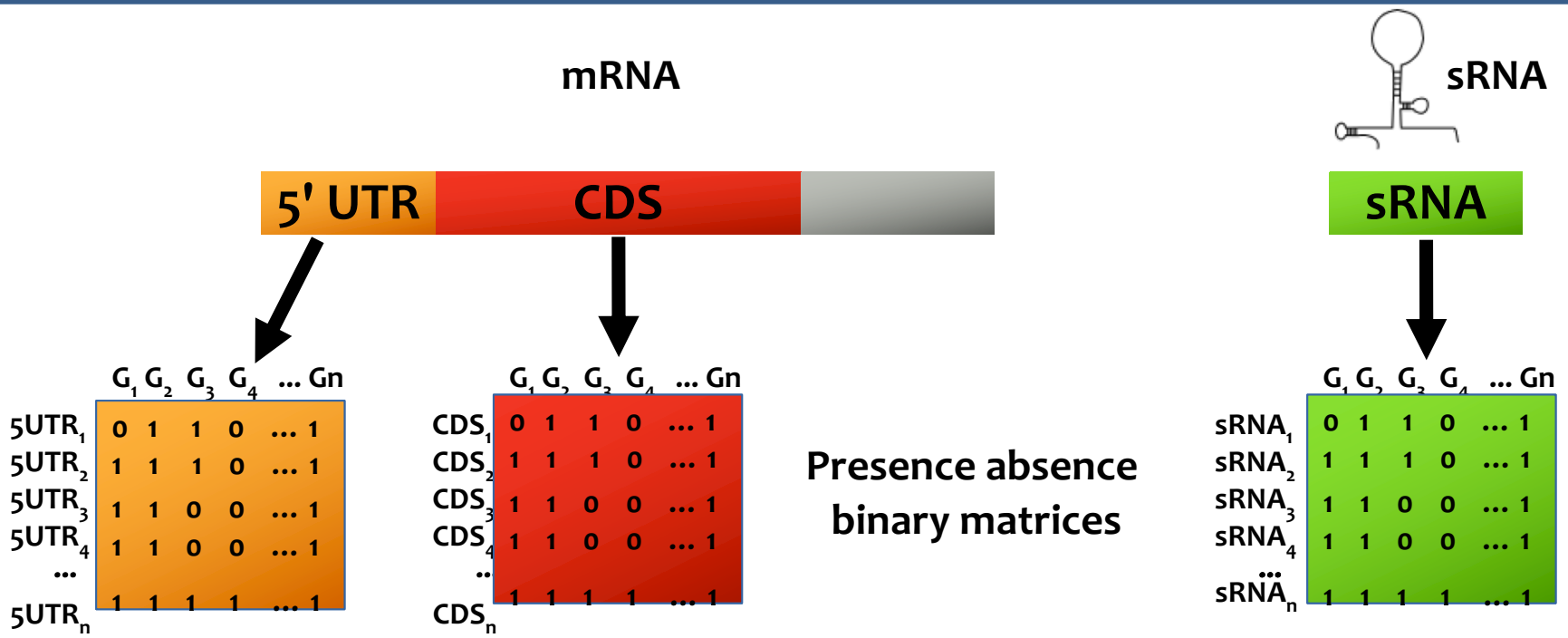
Method: reference tree

- **Superalignment of core genes**
  - All-vs-all proteins comparison
  - Ortholog protein families construction and alignment
  - Misaligned regions removed + protein alignments reverse translated
  - Nucleic alignment of orthologous families concatenated
- **Maximum likelihood tree** (GTR model) with SH support



[1] Camacho C et al., BMC Bioinformatics, 2009
[2] Fouts DE et al., Nucleic Aics Res., 2012
[3] Sievers F. and Higgins DG., Methods Mol Biol., 2014
[4] Talavera, G., and Castresana, Systematic Biol., 2007
[5] Price MN et al., PloS One, 2010

Method: presence/absence matrices



- **sRNAs & 5'UTR → Blastn+** (evalue ≤10^{-2}, coverage ≥ 70 % of query, word size = 7**)** + thresholds adapted to 5'UTR sequence length for predicted 5'UTR:
  - ✓ 15-20 nt : 90 % id. & 100 % cov. Min
  - ✓ 20-50 nt : 80 % id. & 80 % cov. Min
  - ✓ 50-100 nt : >80 % id. & 50 cov. Min
  - ✓ >100 nt : >80 % id. & 20% cov. Min
- **CDS → Blastp+** (default parameters)

[1] Camacho C et al., BMC Bioinformatics, 2009

Strategy overview : several recent method refinements

- **sRNA dataset**
  - Merge overlapping (sens **and antisens**) sRNA in **unique sRNA loci**
- **5'UTR and sRNA presence/absence profiles**
  - Take into account **missing data** before affecting absence profiles (important for draft genomes)
  - **Adapt presence thresholds according to sequence length** (5'UTR predicted) : (15-20 nt= 90 % id. &100 % cov. Min, 20-50 nt : 80 % id. & 80 % cov. Min, 50-100 nt : 80 % id. & 50 cov. Min, >100 80 % id. & 20% cov.)
- **Phyletic profiles comparison and coevolution detection**
  - use CorrHMM (ref) to identify significant correlations using FitPagel models (ref)
  - correct p-values to take into account multiple testing (n sRNAs loci => n tests).

https://cran.r-project.org/web/packages/corHMM/corHMM.pdf