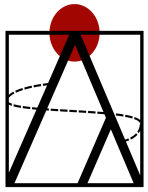
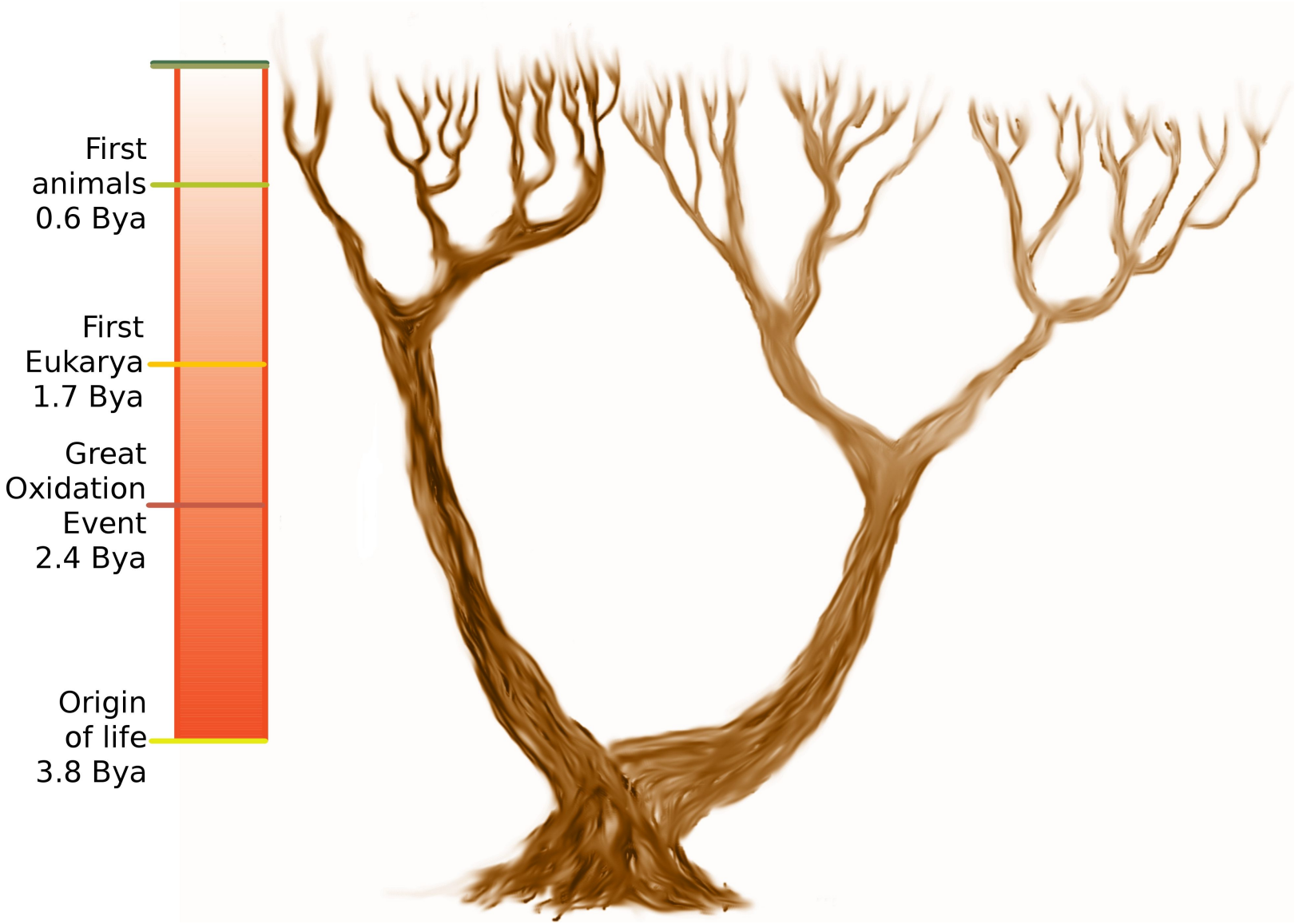



L'histoire de la vie dans les génomes

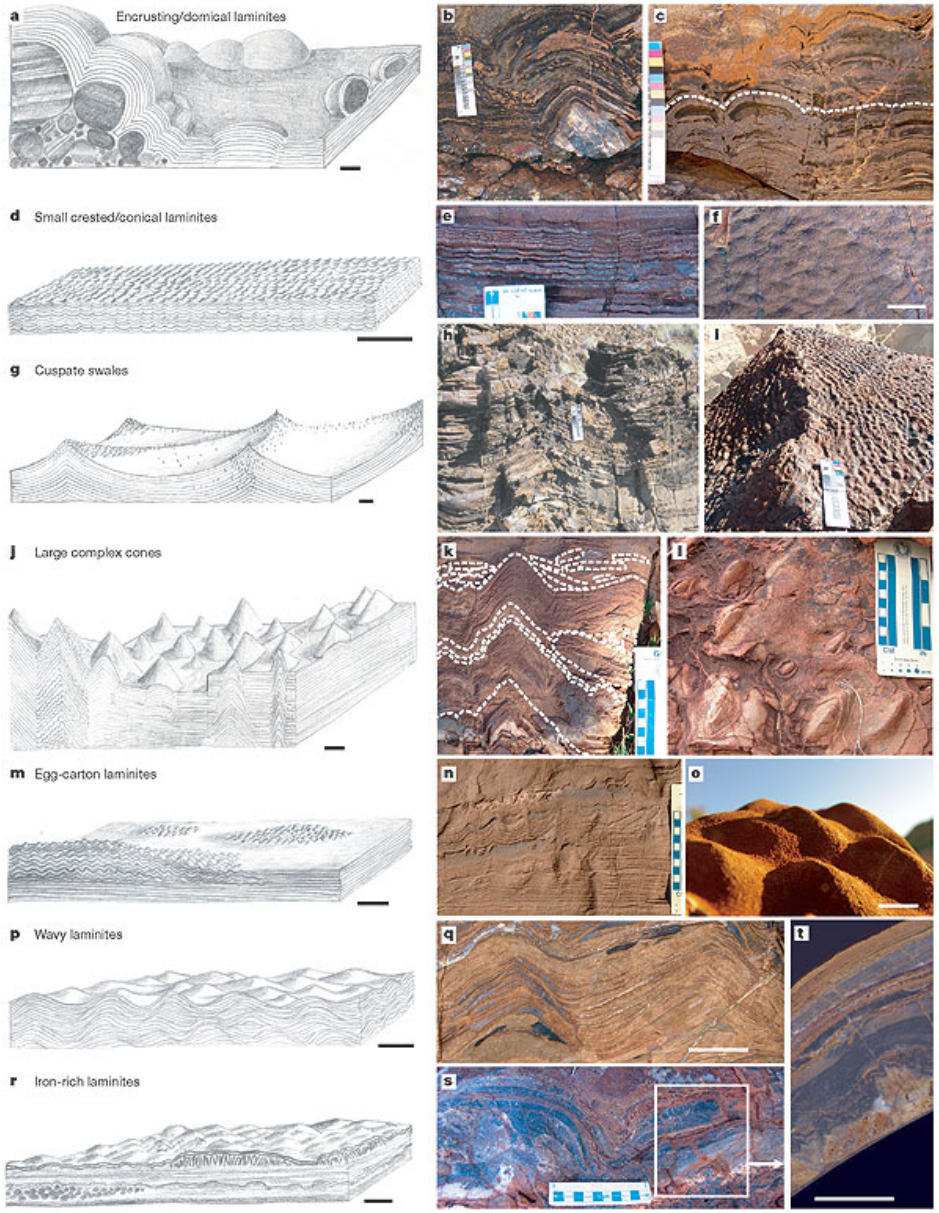
Bastien Boussau

LBBE, UMR5558 CNRS

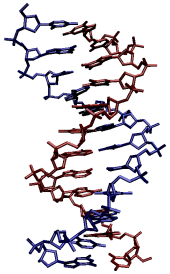
Evolution


First
animals
0.6 Bya
First
Eukarya
1.7 Bya
Great
Oxidation
Event
2.4 Bya
Origin
of life
3.8 Bya

Evolution in rocks



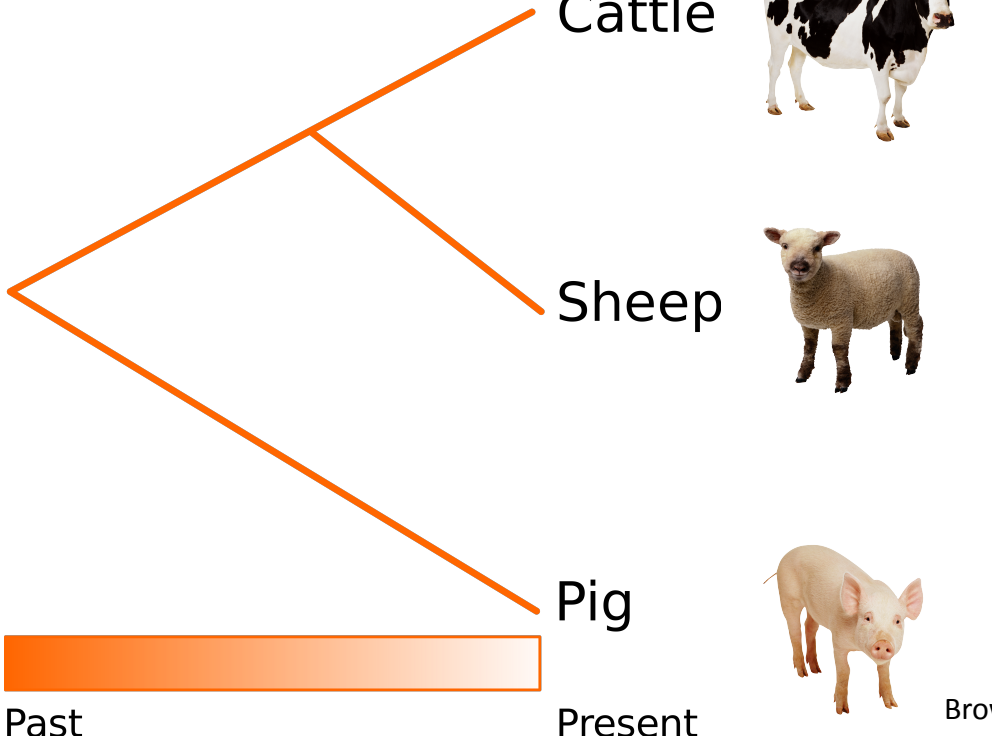
Allwood et al., Nature 2006.



Evolution in our genomes

Table 10. *The structure of the glycyl chains of pig and sheep insulins*

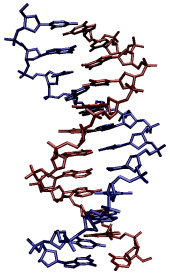
Structure of fraction *A* of cattle insulin	Gly	Ileu	Val	Glu	NH_2 Glu	$CySO_3H$		$CySO_3H$		Ala	Ser	Val	$CySO_3H$		Ser	Leu	Tyr	NH_2 Glu	Leu	Glu	NH_2 Asp	Tyr	$CySO_3H$	NH_2 Asp
	1	2	3	4	5	6		7		8	9	10	11		12	13	14	15	16	17	18	19	20	21
Structure of fraction *A* of sheep insulin	Gly	Ileu	Val	Glu	NH_2 Glu	$CySO_3H$		$CySO_3H$		Ala	Gly	Val	$CySO_3H$		Ser	Leu	Tyr	NH_2 Glu	Leu	Glu	NH_2 Asp	Tyr	$CySO_3H$	NH_2 Asp
Structure of fraction *A* of pig insulin	Gly	Ileu	Val	Glu	NH_2 Glu	$CySO_3H$		$CySO_3H$		Thr	Ser	Ileu	$CySO_3H$		Ser	Leu	Tyr	NH_2 Glu	Leu	Glu	NH_2 Asp	Tyr	$CySO_3H$	NH_2 Asp



Brown et al., Biochem. J. 1955.

Molecular evolution and phylogenomics

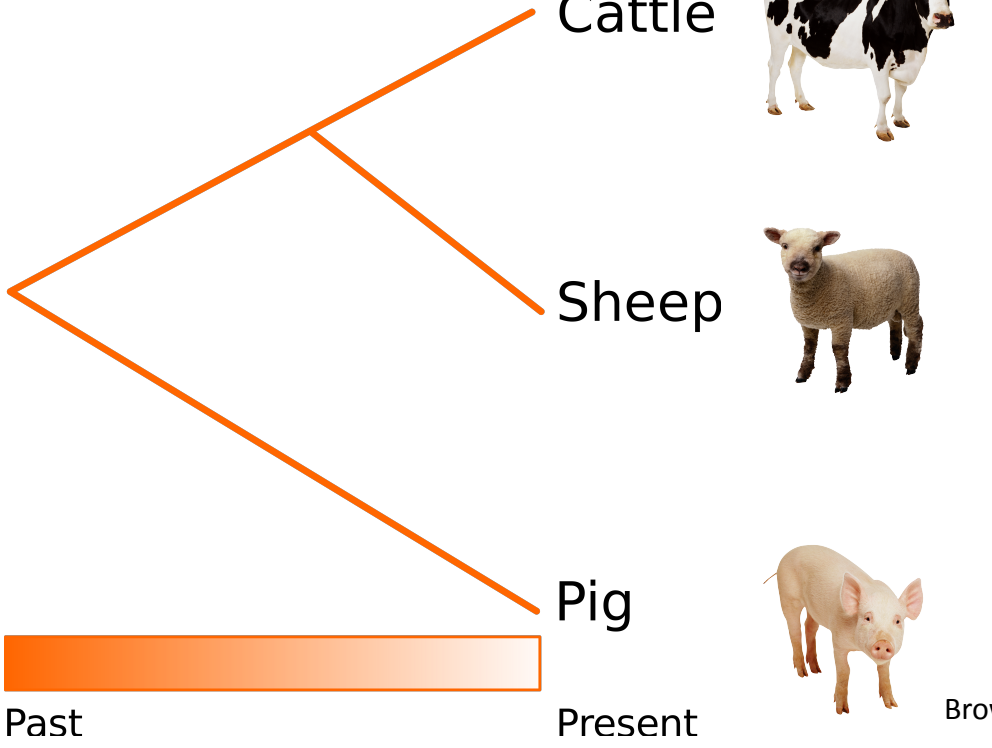
- Learning about the history of life
- Learning about the functions encoded in the genome:
  - What's important? What's not?
  - What is the function of a particular part of the genome?



Evolution in our genomes

Table 10. *The structure of the glycyl chains of pig and sheep insulins*

Structure of fraction *A* of cattle insulin	Gly	Ileu	Val	Glu	NH_2 Glu	$CySO_3H$		$CySO_3H$		Ala	Ser	Val	$CySO_3H$		Ser	Leu	Tyr	NH_2 Glu	Leu	Glu	NH_2 Asp	Tyr	$CySO_3H$	NH_2 Asp
	1	2	3	4	5	6		7		8	9	10	11		12	13	14	15	16	17	18	19	20	21
Structure of fraction *A* of sheep insulin	Gly	Ileu	Val	Glu	NH_2 Glu	$CySO_3H$		$CySO_3H$		Ala	Gly	Val	$CySO_3H$		Ser	Leu	Tyr	NH_2 Glu	Leu	Glu	NH_2 Asp	Tyr	$CySO_3H$	NH_2 Asp
Structure of fraction *A* of pig insulin	Gly	Ileu	Val	Glu	NH_2 Glu	$CySO_3H$		$CySO_3H$		Thr	Ser	Ileu	$CySO_3H$		Ser	Leu	Tyr	NH_2 Glu	Leu	Glu	NH_2 Asp	Tyr	$CySO_3H$	NH_2 Asp



Brown et al., Biochem. J. 1955.

Molecular evolution and phylogenomics

- **Learning about the history of life**
- Learning about the functions encoded in the genome:
  - What's important? What's not?
  - What is the function of a particular part of the genome?

Tree reconstruction: automation

- Based on a **probabilistic model** of sequence evolution
- Evaluation of a phylogenetic tree (e.g. insulin):
  - O(number of sequences*number of sites*(alphabet size) ^{2})
- Number of trees of size n sequences: (2n-3)!!:
  - E.g.: 50 sequences: as many trees as there are electrons in the visible universe

➔ Exploration heuristics, sampling methods (MCMC)

The age of the data

One genome = 500 to 50,000 genes
(Human genome: 30,000 genes, 3 Gb)

- 6887 complete genomes available

- Genome 10K: 10000 Vertebrate genomes

- i5K: 5000 insect genomes in the next 5 years

- 1KP: 1000 plant genomes

What we do in Lyon

1. Maintain databases of trees
   - The hogenom database
   - Associated developments
2. Study the evolution of life
   - The Ancestrome project
   - Associated developments

The Hogenom database (2013)

- Simon Penel
- 1470 complete genomes (1233 Bacteria, 97 Archaea, 140 Eukaryotes)
  - 27,000,000 sequences
  - 70 years of computation

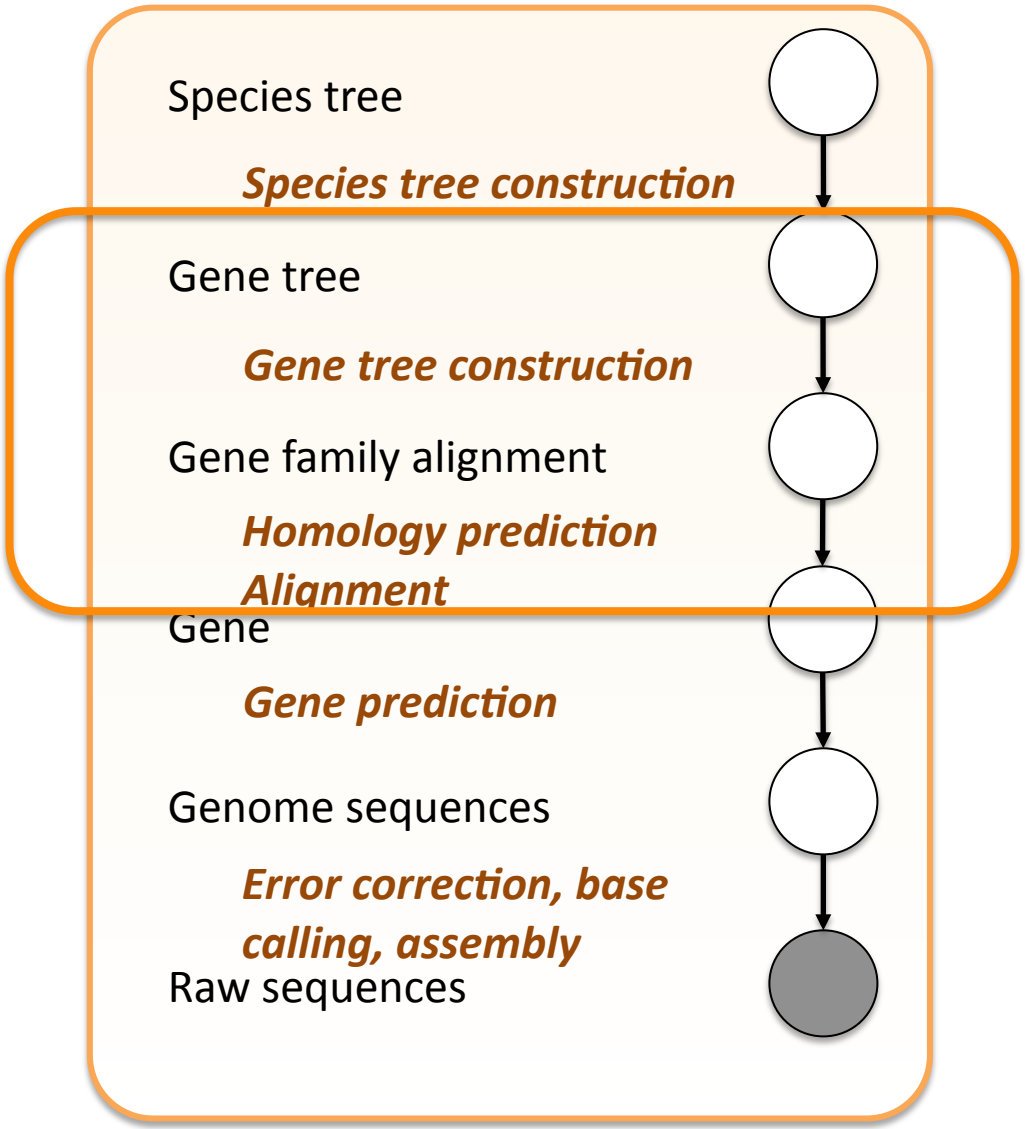
Historically:

Computations run at:

- the CC IN2P3 (through the grid)
- PRABI: mésocentre Tier-2 thématique Bioinformatique

We plan on continuing computations using the brand new machine of the « Fédération Lyonnaise de Modélisation et Sciences Numériques » (FLMSM) (Tier-2 lyonnais)

It takes 3 steps to build Hogenom



3 steps:

-**Family building**

-Family alignment

-**Family tree building**

Family building

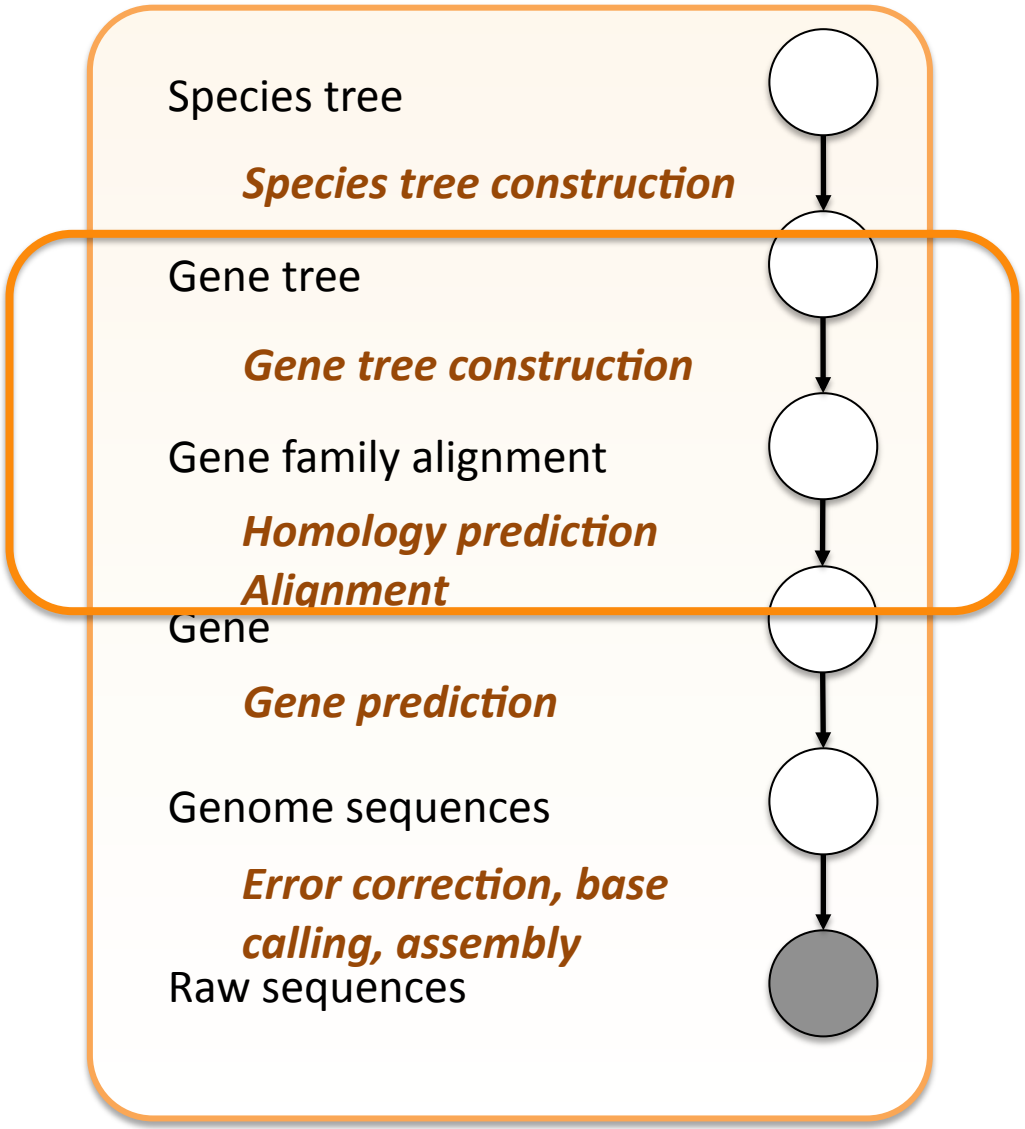
- Problem: We want to group cattle *insulin* with sheep *insulin*, not sheep *hemoglobin*
- ➔Grouping sequences by similarity

1. computing similarities between all pairs of sequences ((27 M)^2: very costly but now: incremental!)
2. group sequences according to their pattern of similarities: HiFix: 20h on our cluster (Mièle et al., Bioinformatics 2012)

Computing similarities between all pairs of sequences

Release	**Nombre de séquences**	**Ressources de calcul**	**Temps de calcul** (1 proc Intel Xeon.)	**Sorties**
1 (2009)	7 000 000	grille TIDRA (CCIN2P3)	13 ans	1,5 To
2 (2011)	13 000 000	grille TIDRA grille GRISBI PRABI	25 ans	4 To (cumulé 5,5 To)
3 (2013)	27 000 000	PRABI	70 ans	8 To (cumulé 13,5 To)

It takes 3 steps to build Hogenom



3 steps:
-Family building
-Family alignment
-Family tree building

PHYLDOG

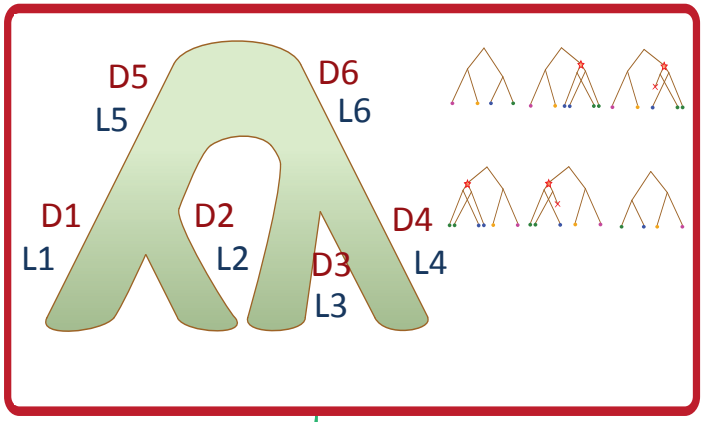
- We want good gene trees and a good species tree
- Usually they are estimated separately
- PHYLDOG jointly estimates the species tree and gene trees

➔ *Model of the dependence between gene trees and the species tree*

1: Databases

PHYLDOG

Rooted species tree,
numbers of duplications
and losses,
rooted gene trees



Joint reconstruction of
the species tree,
gene trees, and
numbers of duplications and losses

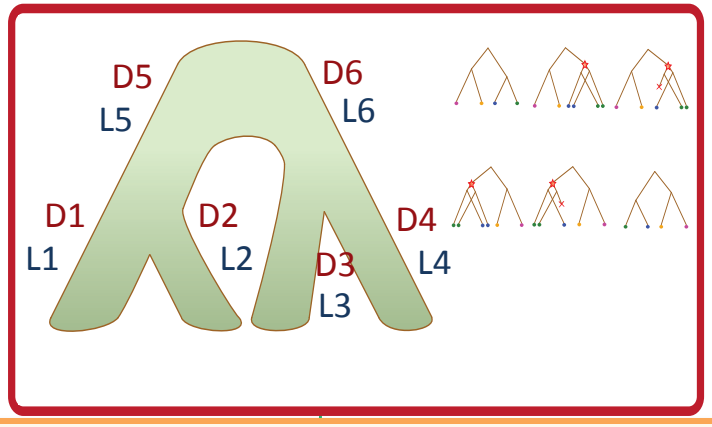
All gene families
(thousands of alignments)

Boussau et al., Genome Research 2013

1: Databases

PHYLDOG

Rooted species tree,
numbers of duplications
and losses,
rooted gene trees

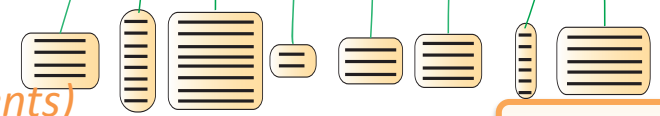


$$P(\text{Alignments}|\text{Gene trees}, \text{Species tree}, DL) = \prod_{G_i \in \{\text{Gene families}\}} P(G_i)$$

$$P(G_i) = P(\text{Alignment}_i|\text{Gene tree}_i) \times P(\text{Gene tree}_i|\text{Species tree}, DL)$$

gene trees, and
numbers of duplications and losses

All gene families
(thousands of alignments)

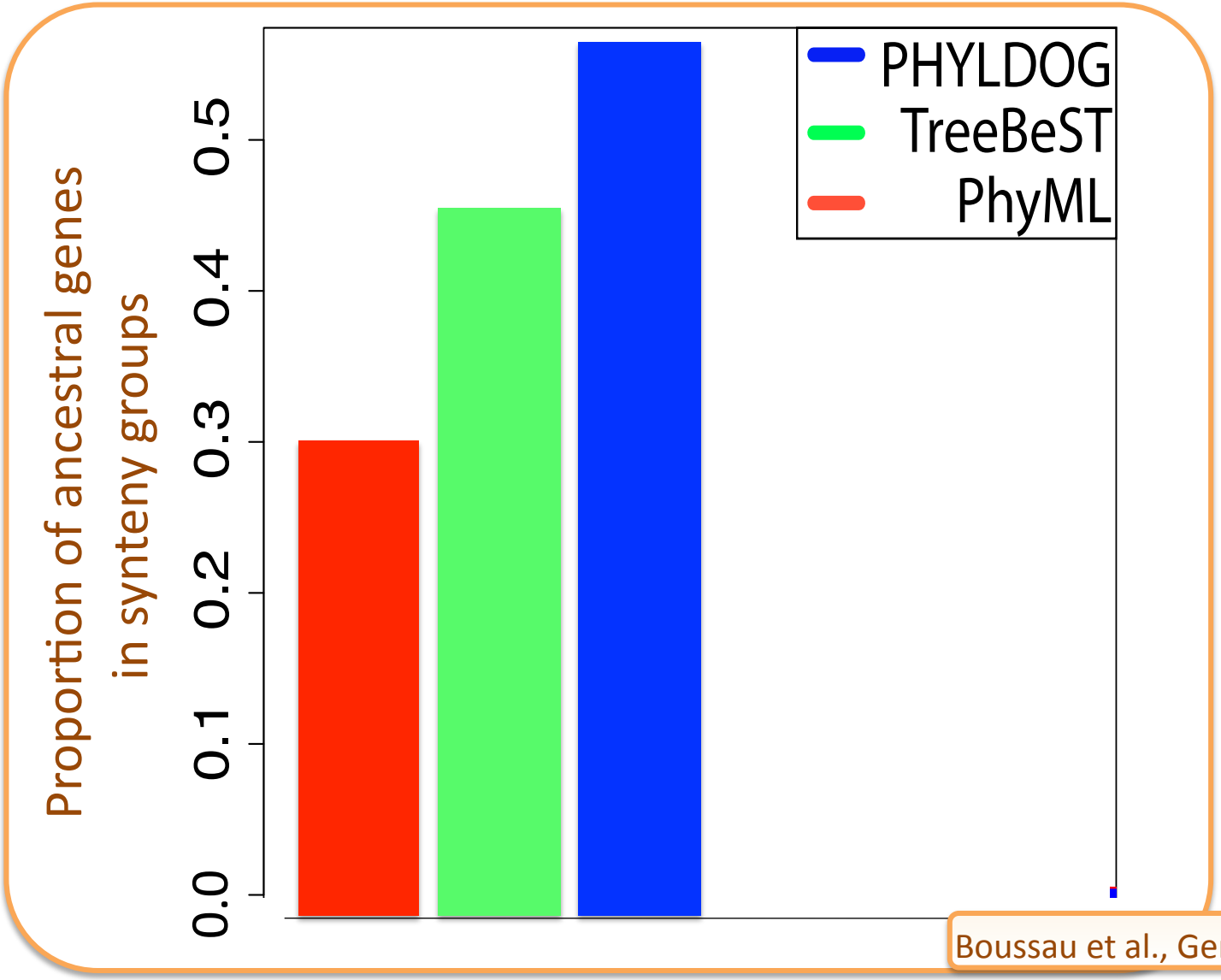


Boussau et al., Genome Research 2013

Computations

- Server-client architecture
- C++
- MPI
- Testing run on the PRABI cluster (Tier-2)
- Computations run on Jade (Tier-1) thanks to the GENCI
- 36 genomes, 7000 gene families:
  - 10 days, 3000 processors → 720,000 hours

Results

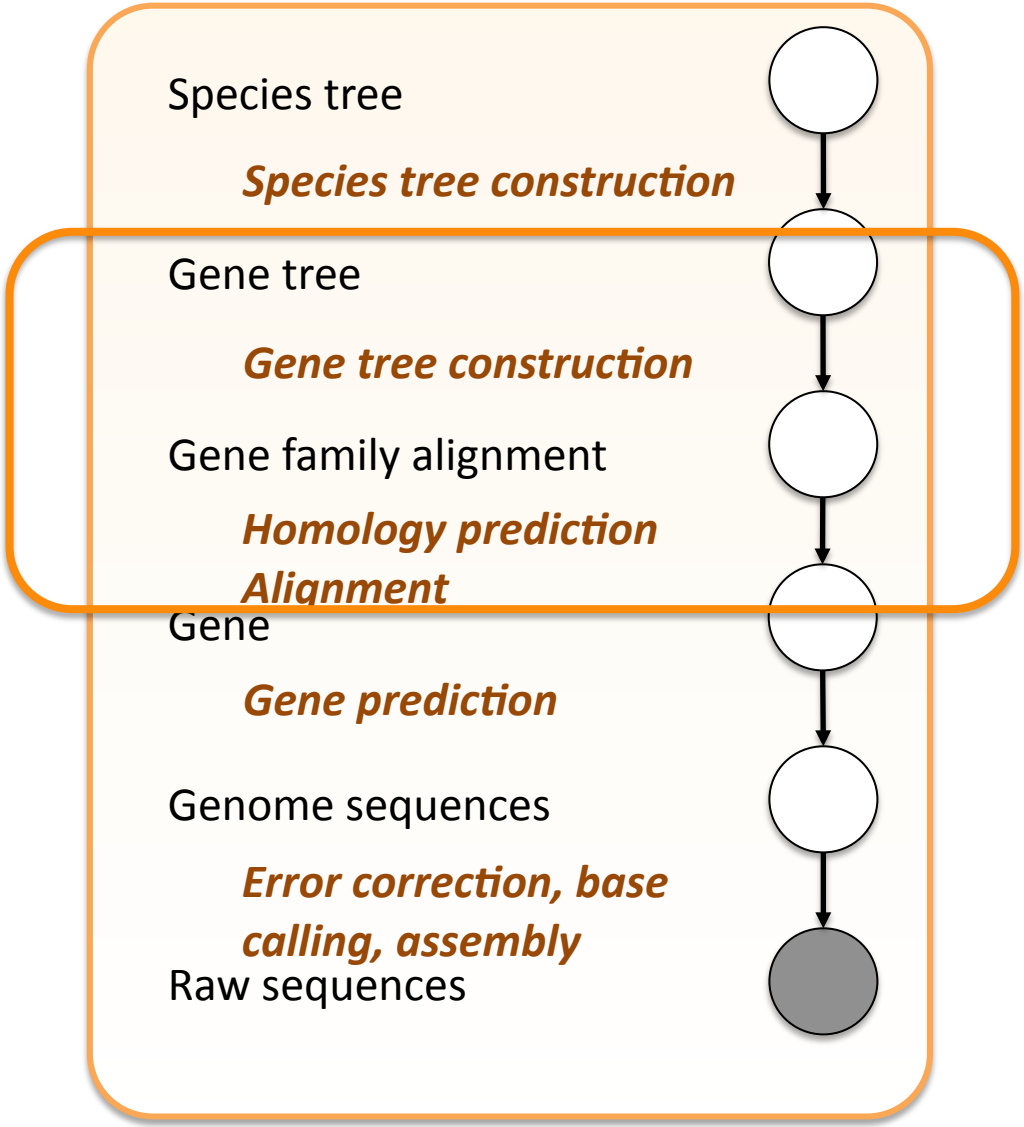


Boussau et al., Genome Research 2013

Future developments

- Better algorithms
- Better load balance
- openMP
- Making it incremental

And we still need to look at gene alignments…



3 steps:

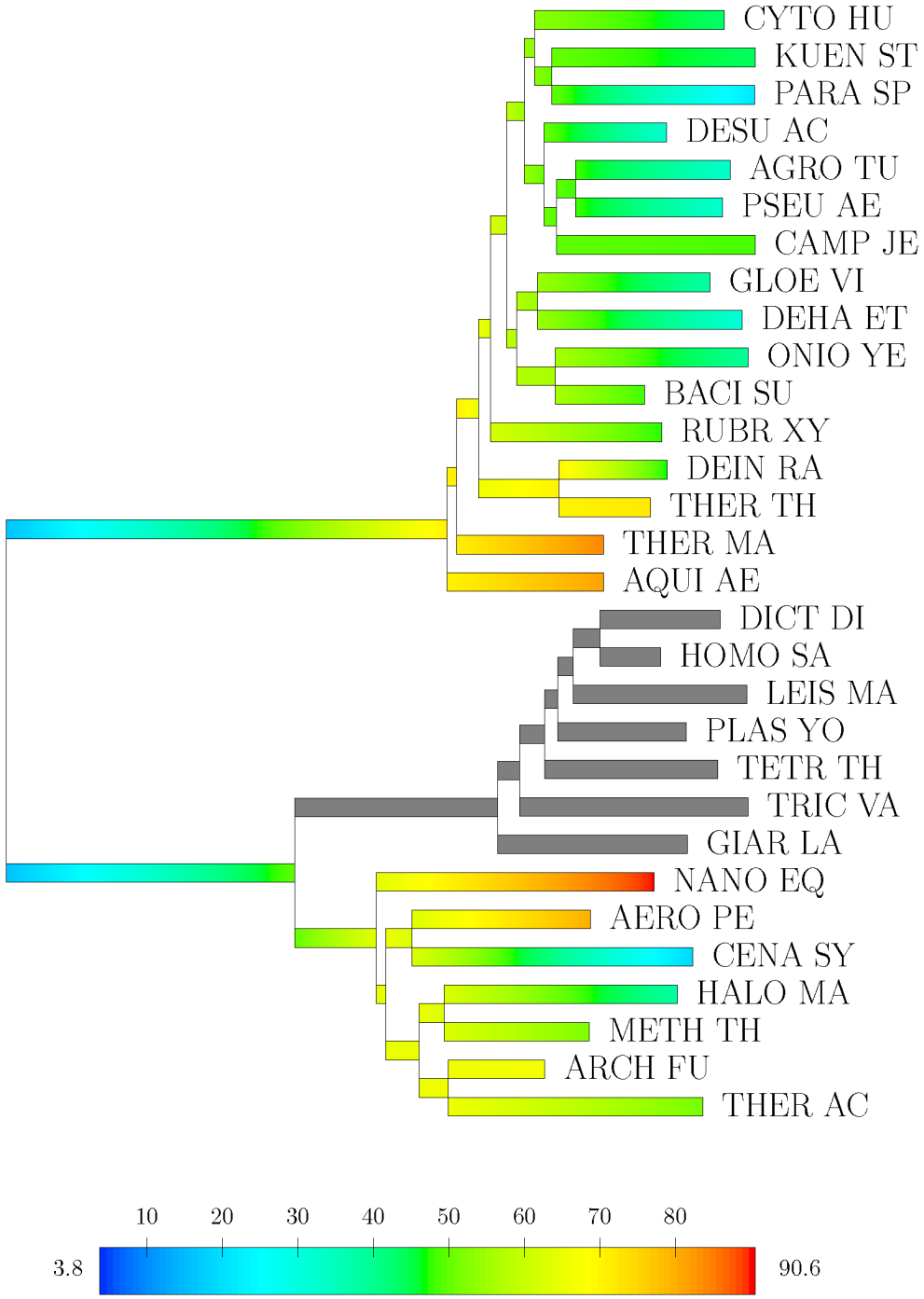
-Family building

-**Family alignment**

-Family tree building

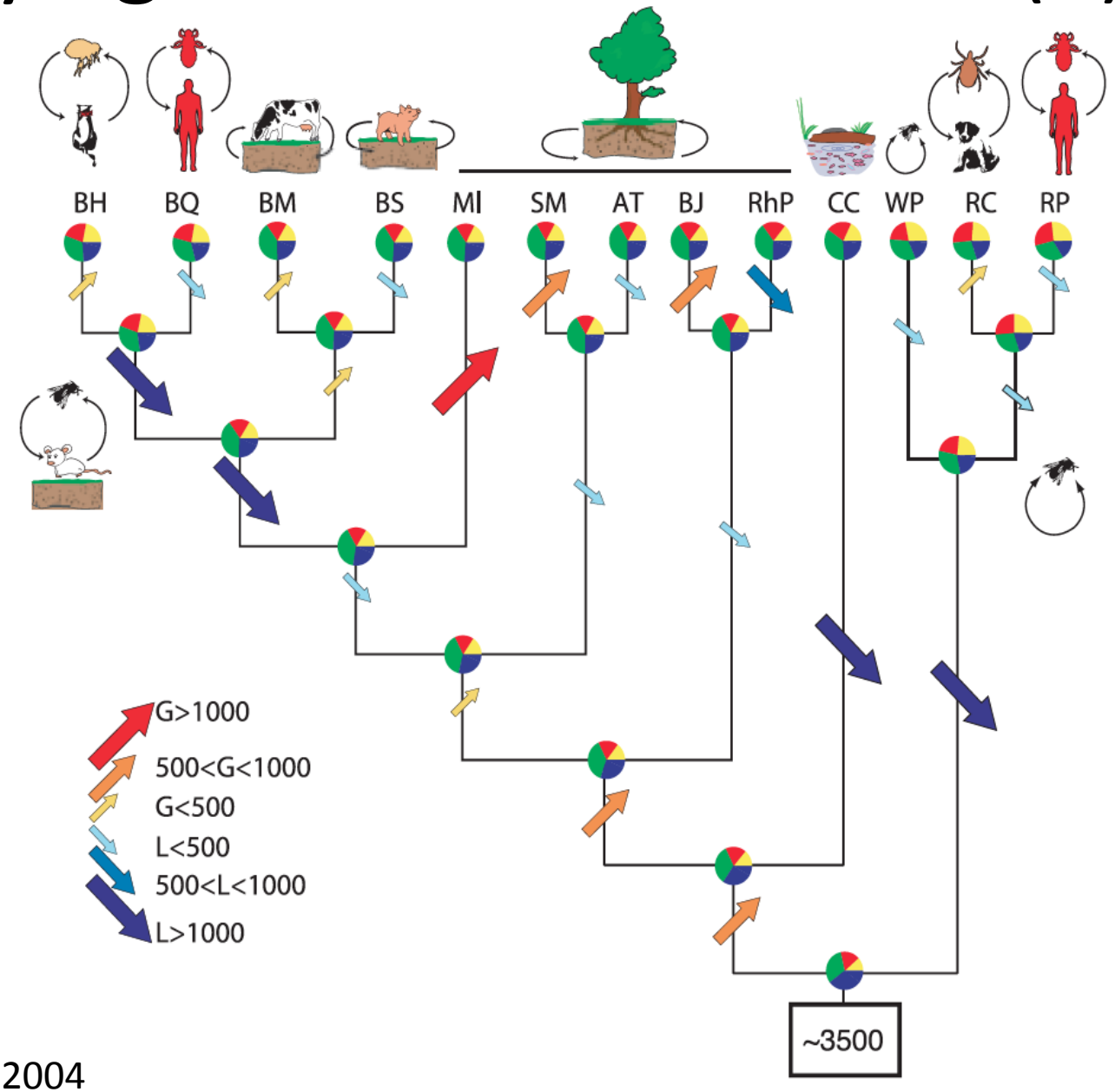
Lots of work needed to make them incremental, and to make them scale…

Studying the evolution of life (1)



Boussau et al., Nature 2008

Studying the evolution of life (2)



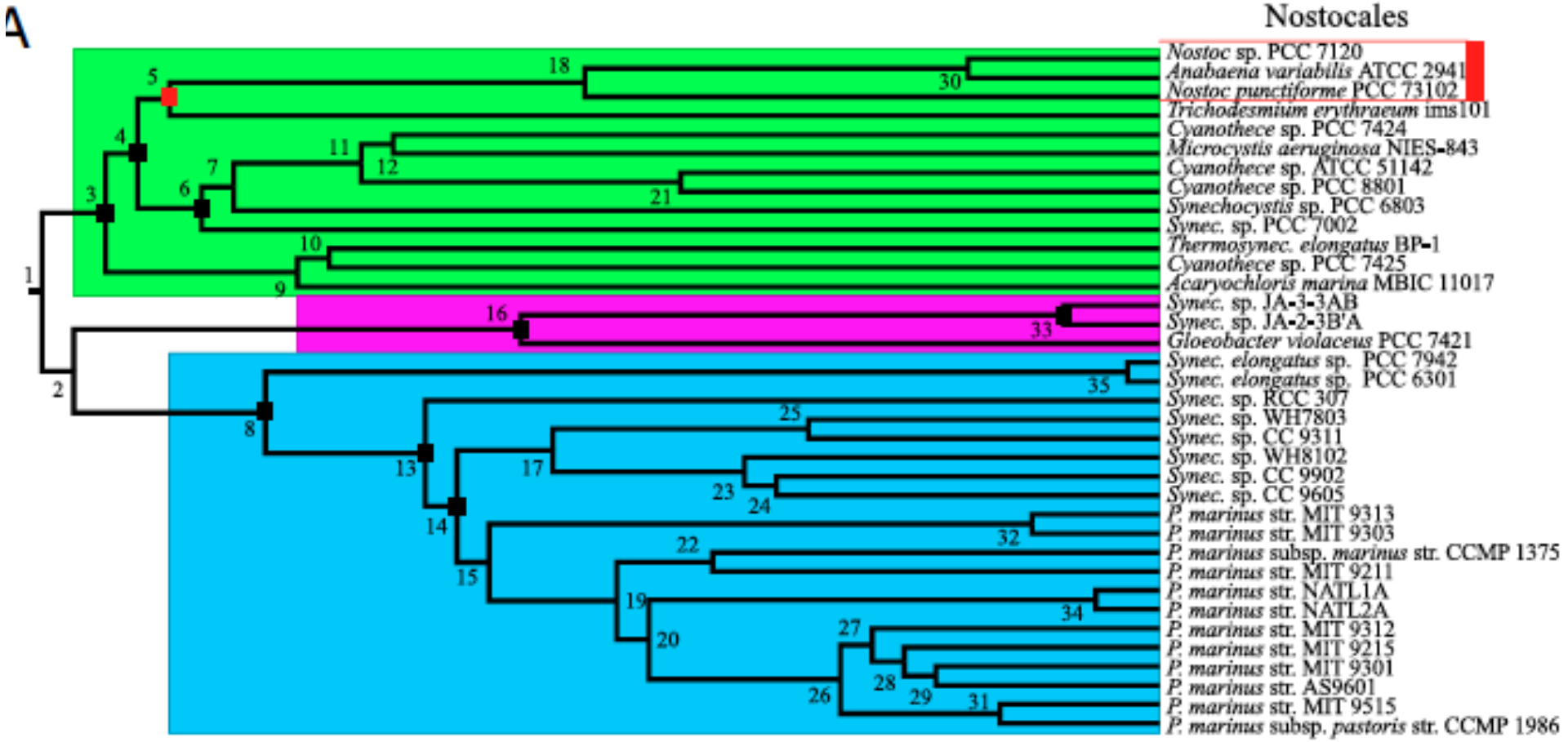
Boussau et al., PNAS 2004

The ancestrome project

- Investissement d'avenir
- 4 laboratoires, coordonné par Vincent Daubin au LBBE (Lyon)
- 5 ans
- 2.2M euros

➔ *Develop methods and use them to reconstruct ancestral genomes and infer ancient lifestyles*

A model for dating events more than 3 billion years ago



Szöllősi et al., PNAS 2012

Access to Curie through Prace

- Tier-0
- Last year: preparatory access: 4 million hours
- This year: 34 million hours
- **Our project**: reconstruct and date the tree of life and tens of thousands of gene trees with 102 genomes, using an accurate probabilistic model

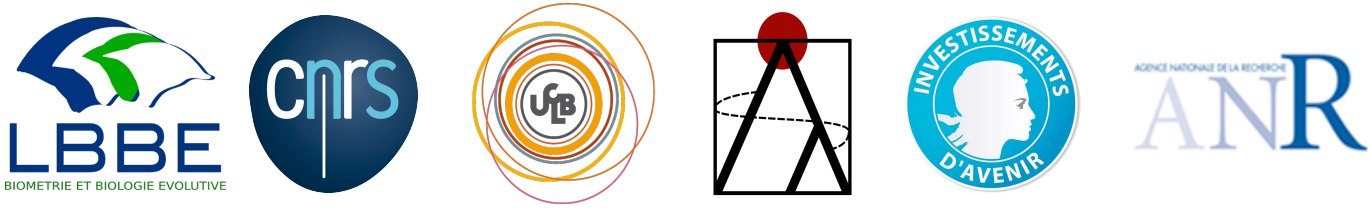
Summary

- We have used the entire pyramid:
  - Tier-2 for database updates and code debugging
  - Tier-1 for production runs
  - Tier-0 for extraordinary projects
- In the future:
  - Keep doing the same!
  - In phylogenomics, the need for heavy computations is not going to decrease

Thanks!

Vincent Daubin, Stéphane Delmotte, Laurent Duret, Manolo Gouy, Lionel Humblot, Vincent Mièle, Simon Penel, Bruno Spataro, Gergely Szöllősi, Eric Tannier

Funding and lab:



Computing resources:

HiFix: family building

- Problem: We want to group cattle *insulin* with sheep *insulin*, not sheep *hemoglobin*
- ➔Grouping sequences by similarity

1. computing similarities between all pairs of sequences ((7 M)$^{2:}$ very costly but now: incremental!)
2. group sequences according to their pattern of similarities: HiFix: 20h on our cluster

Mièle et al., Bioinformatics 2012