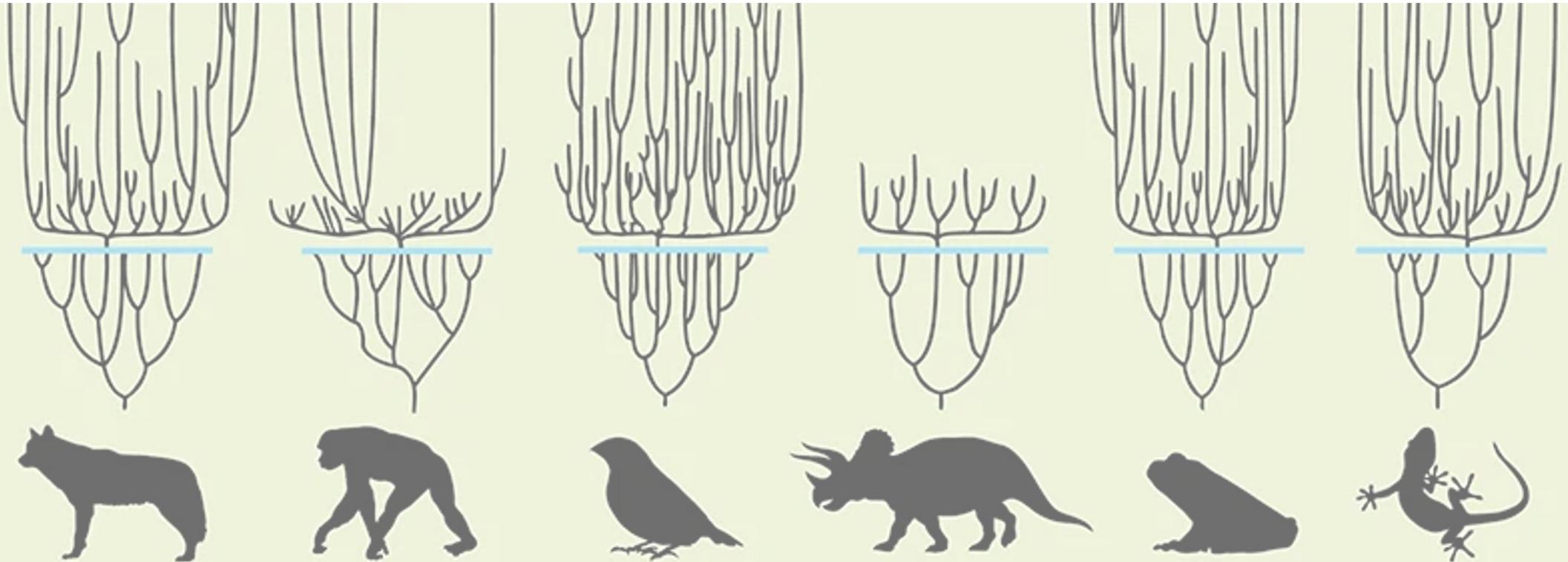


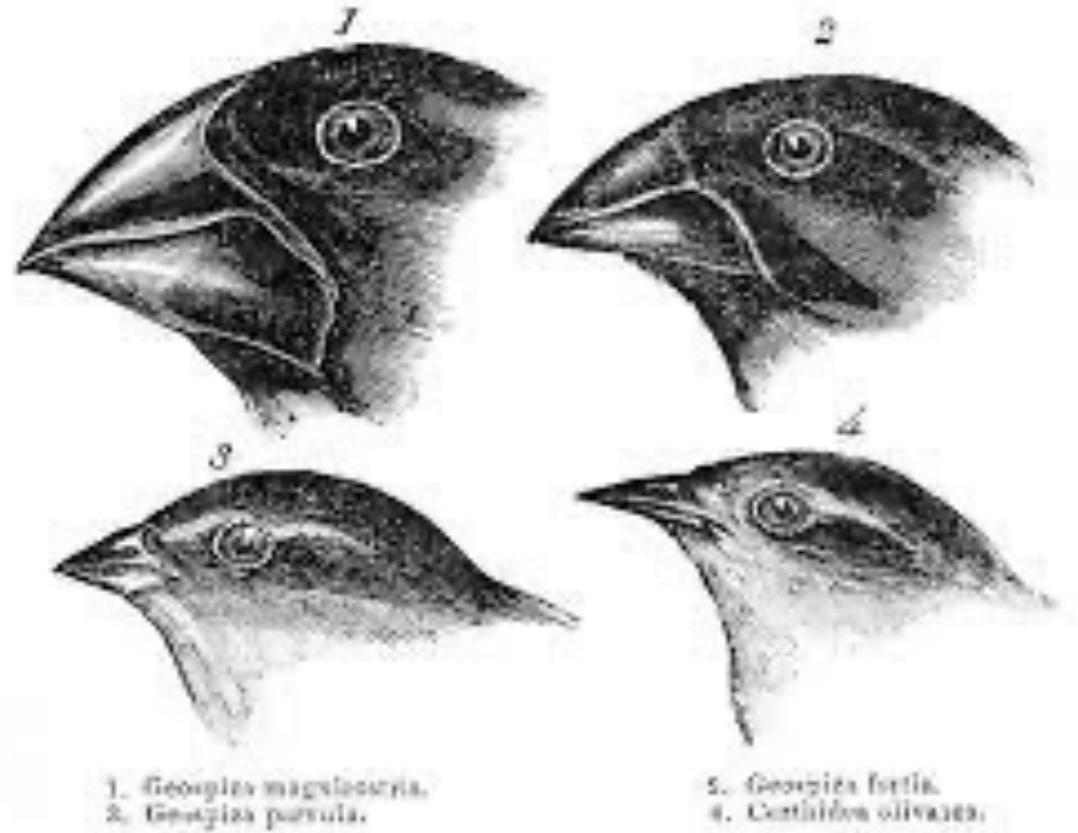
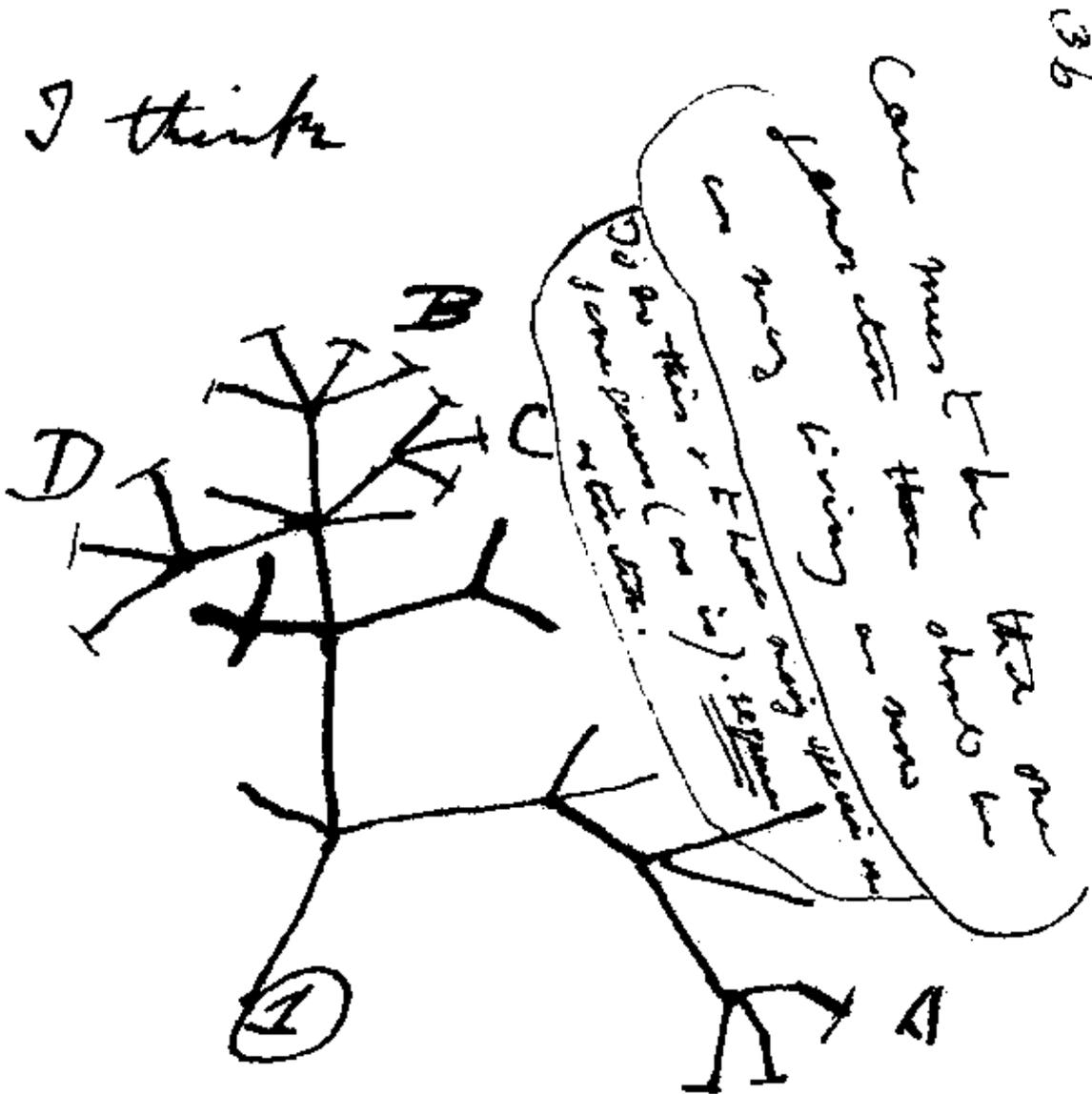
# Introduction à la phylogénétique

Gwen Kettenburg

Diapositives adaptées de Richard Ree et Andrew Hipp, Université de Chicago

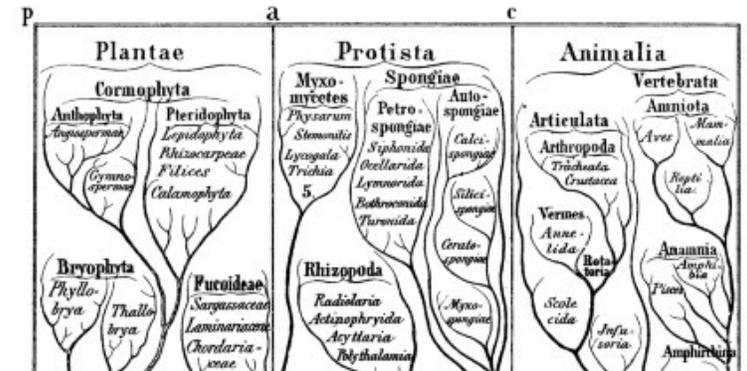


# Qu'est-ce qu'une phylogénie ?

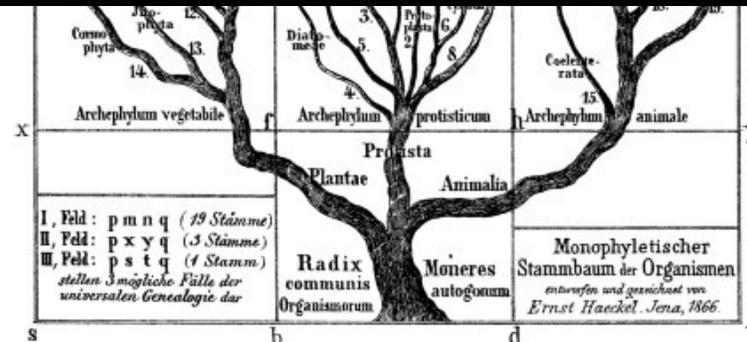
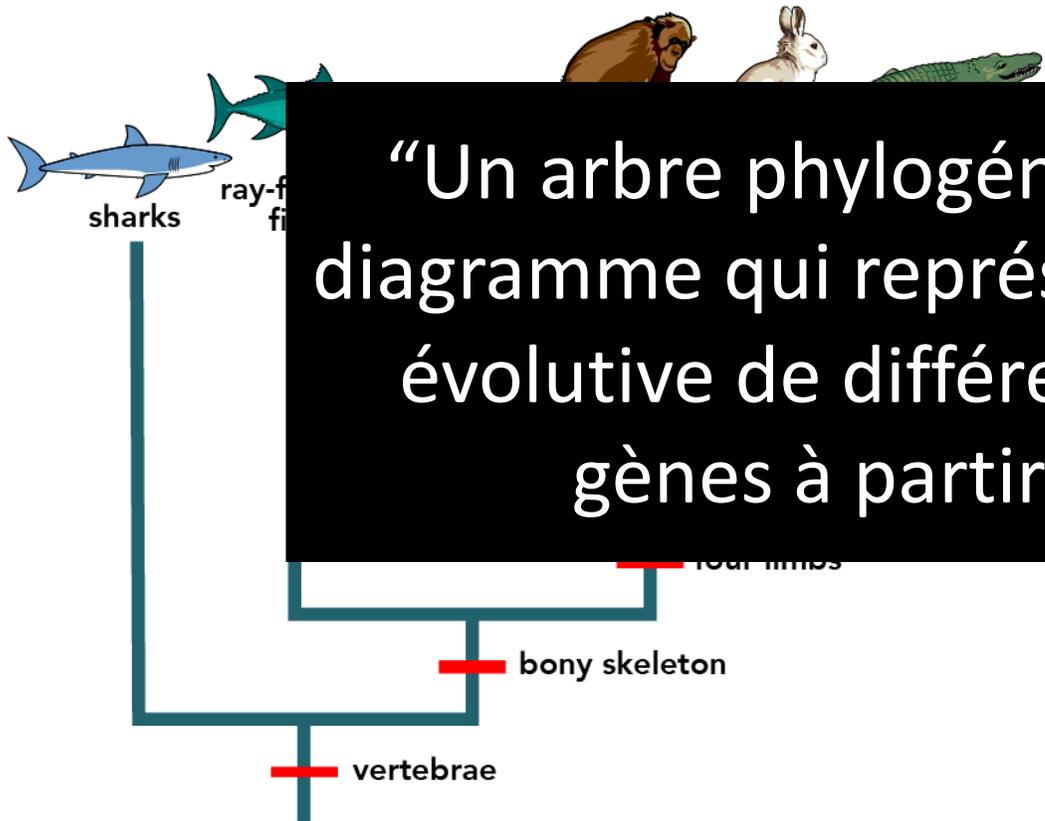




# Qu'est-ce qu'une phylogénie ?

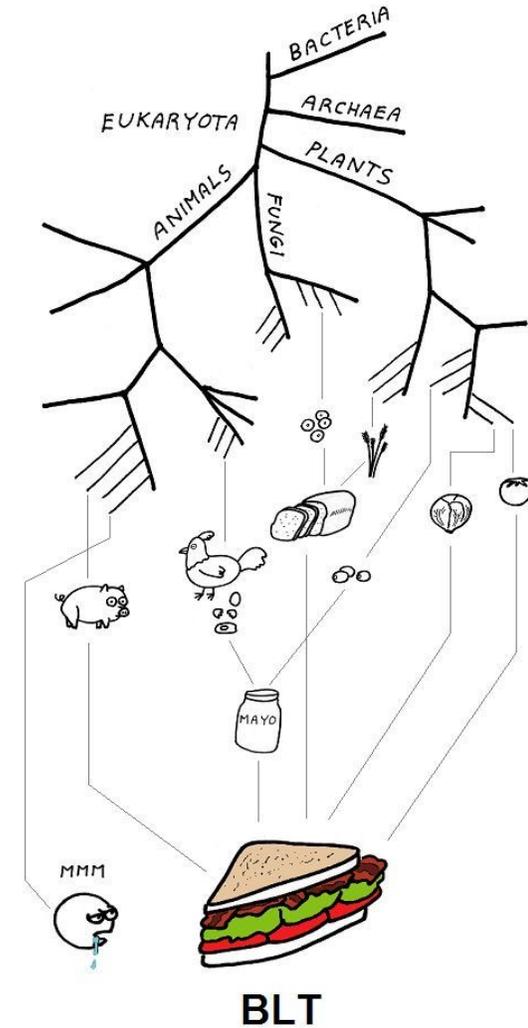
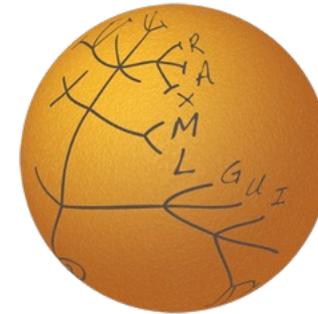


“Un arbre phylogénétique, ou phylogénie, est un diagramme qui représente les lignes de descendance évolutive de différentes espèces, organismes ou gènes à partir d'un ancêtre commun.”



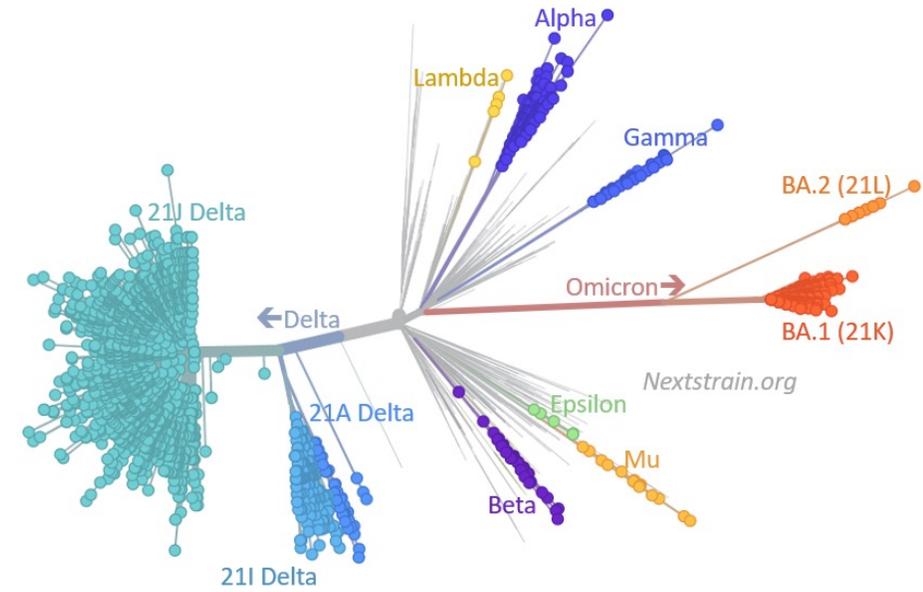
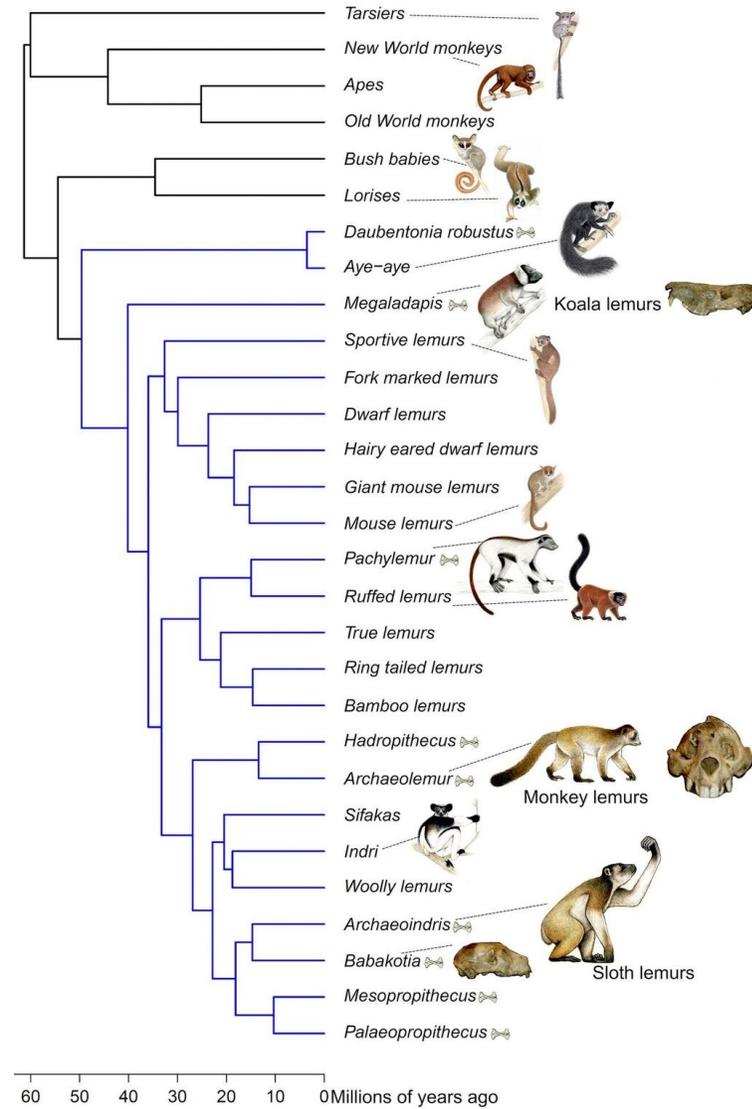
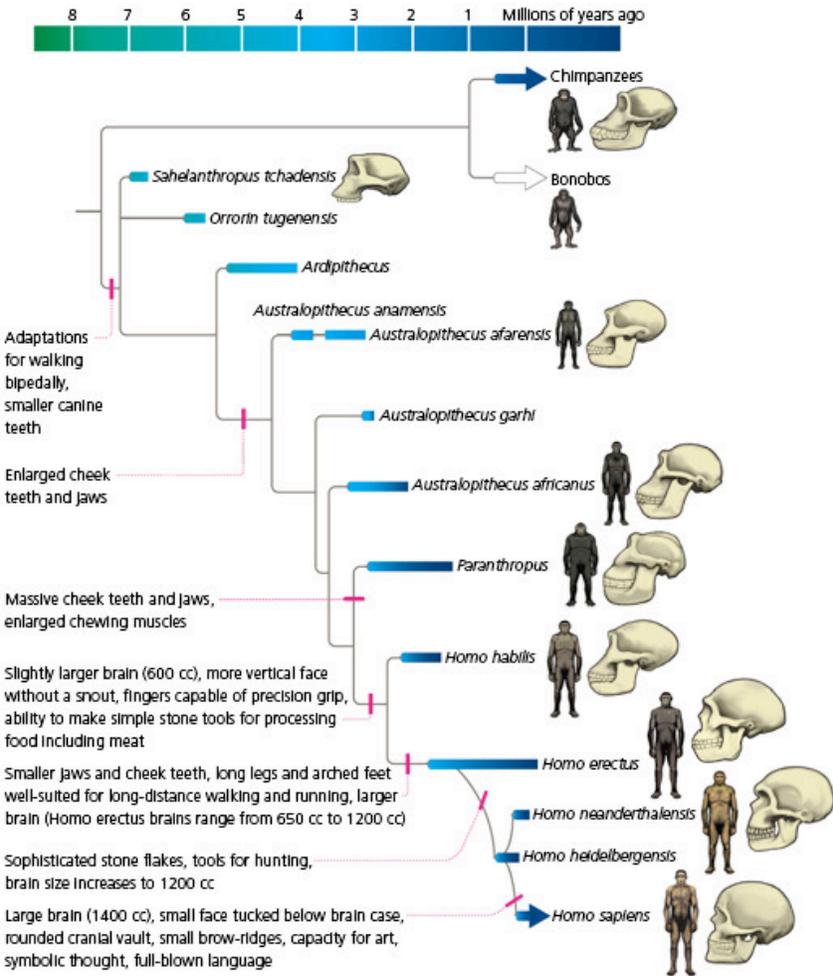
# Objectifs:

- Volet cours magistral
  - Apprenez les bases de ce qu'est une phylogénie
  - Apprendre à lire les phylogénies
  - Bases de la modélisation phylogénétique
- Composant didacticiel
  - Apprenez à créer un arbre phylogénétique à partir de données de séquençage
  - Utilisation de séquences de protéines du cytochrome B du lémurien dans le logiciel RAxML
  - Modifier et visualiser l'arborescence dans R et FigTree

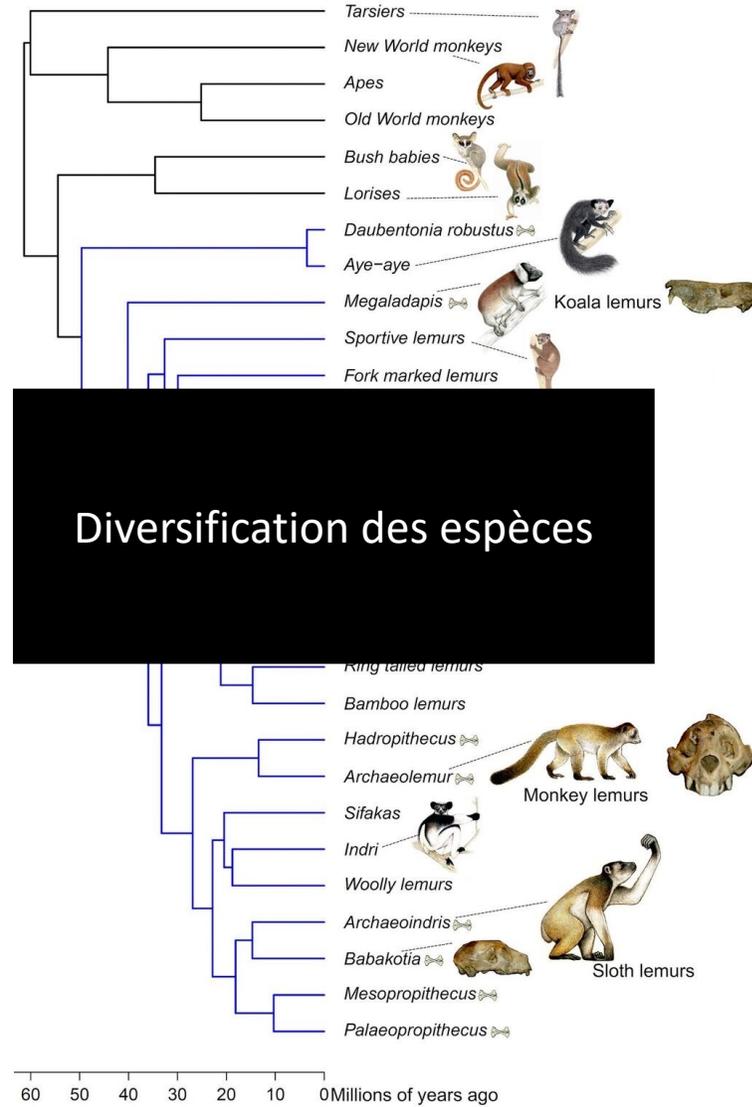
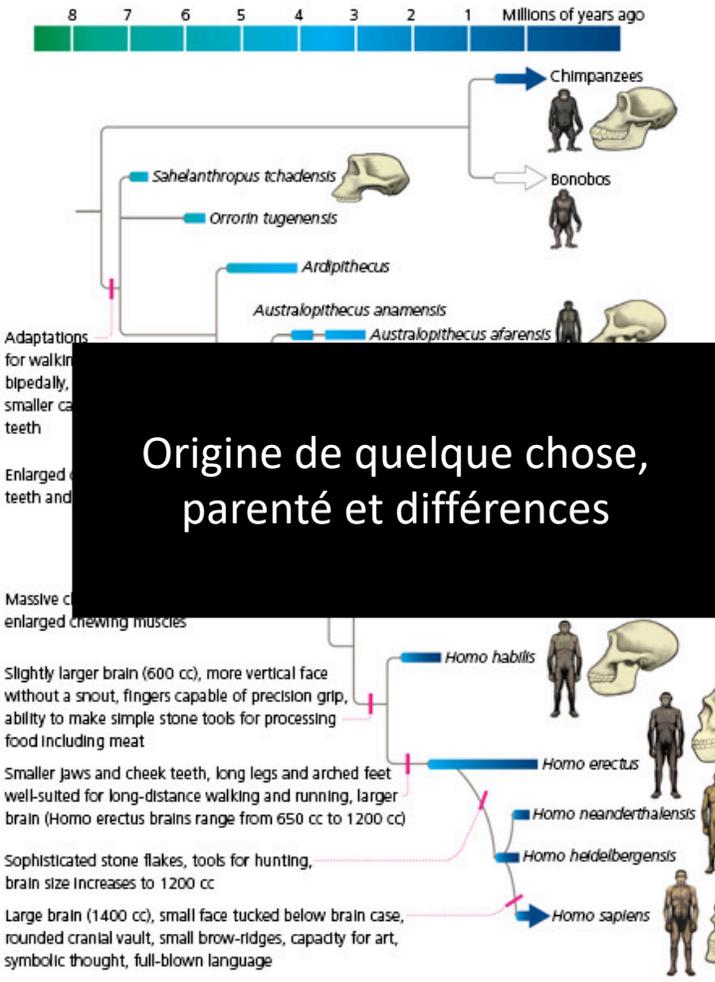


Molecular Evolutionary  
Genetics Analysis

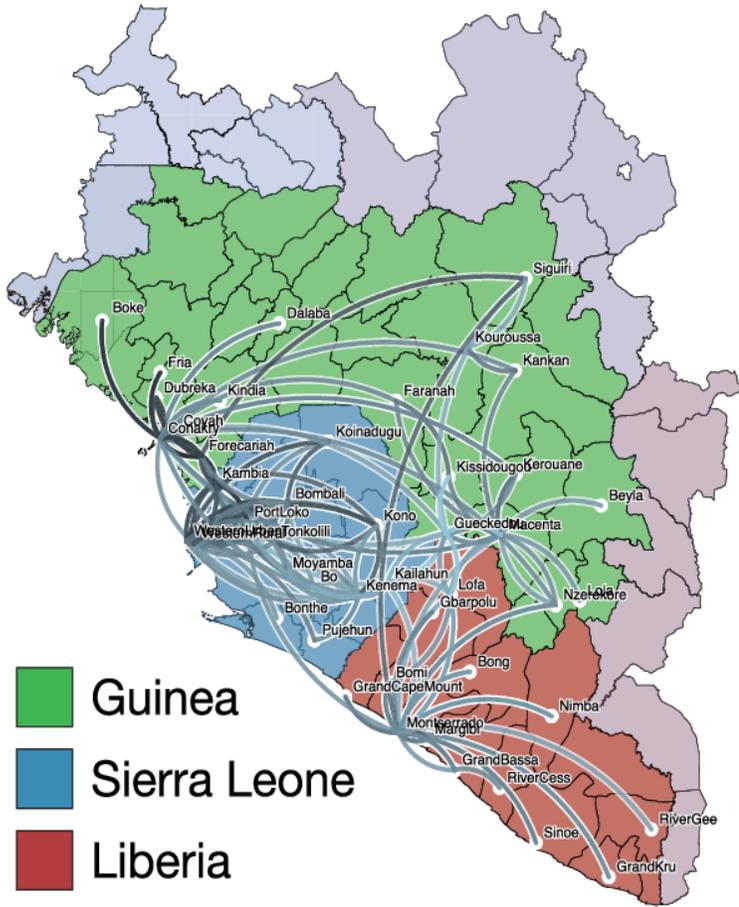
# Que pouvez-vous faire avec les phylogénies ?



# Que pouvez-vous faire avec les phylogénies ?

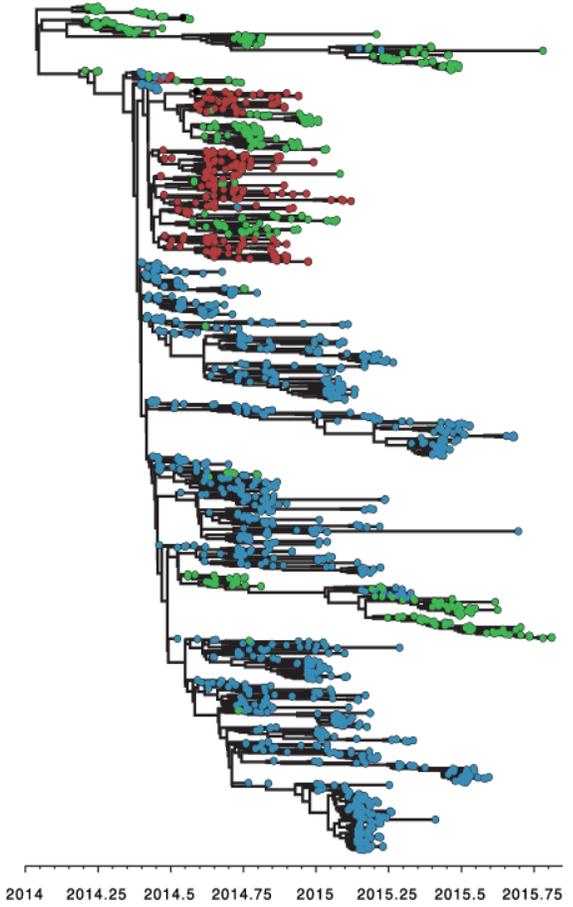


# Phylodynamique



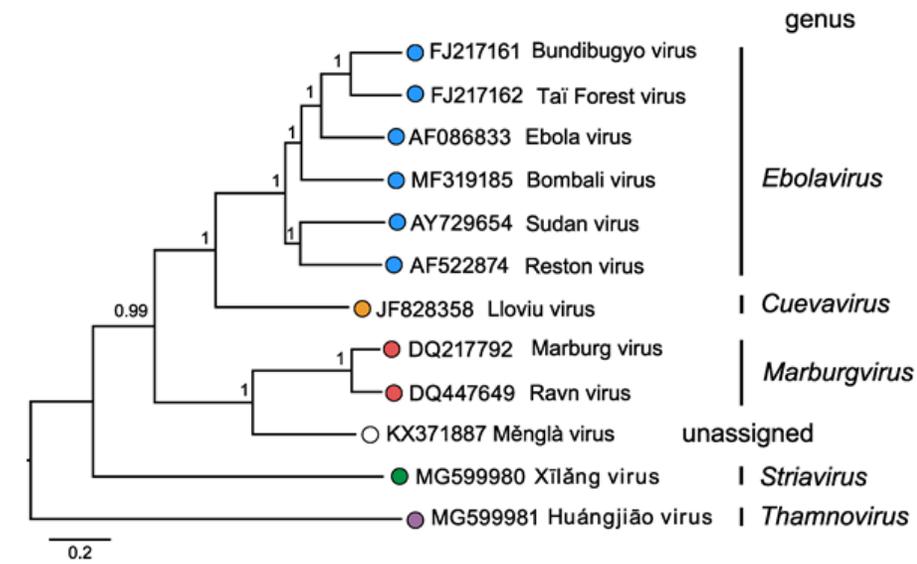
OÙ va-t-il ?

# Arbres bayésiens



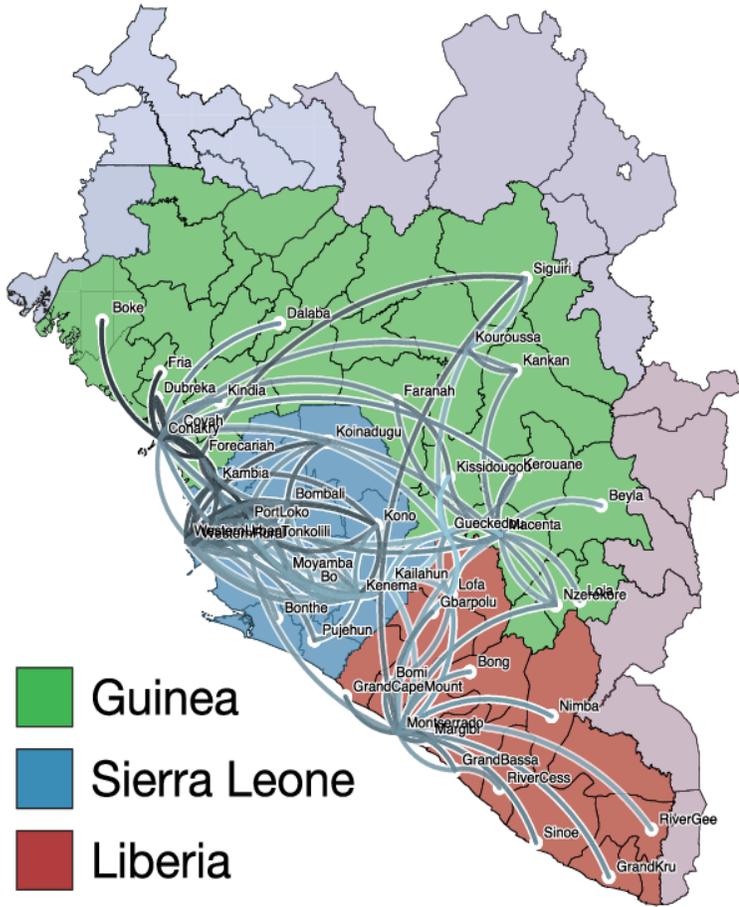
QUAND est l'ancêtre commun le plus récent ?

# Maximum de vraisemblance



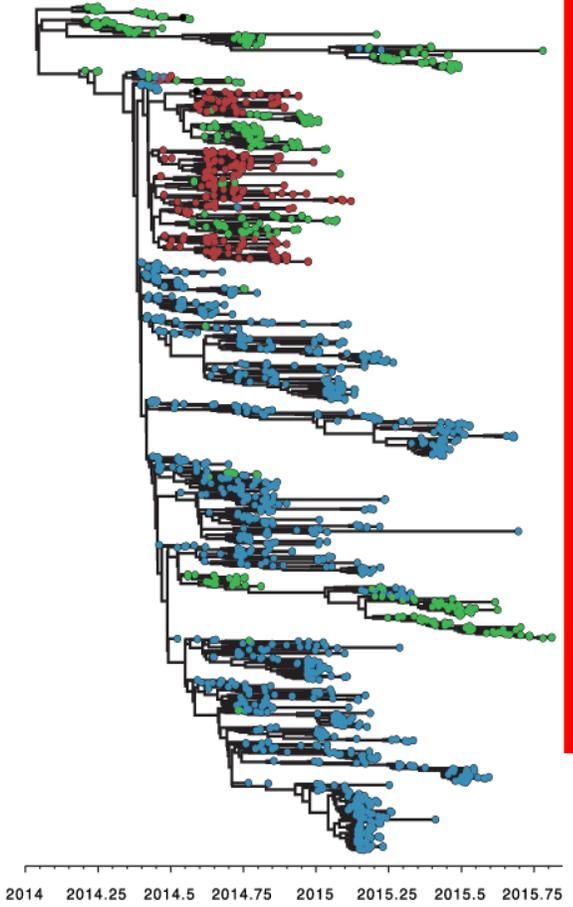
À quel point est-ce différent de ce que l'on connaît ?

# Phylodynamique



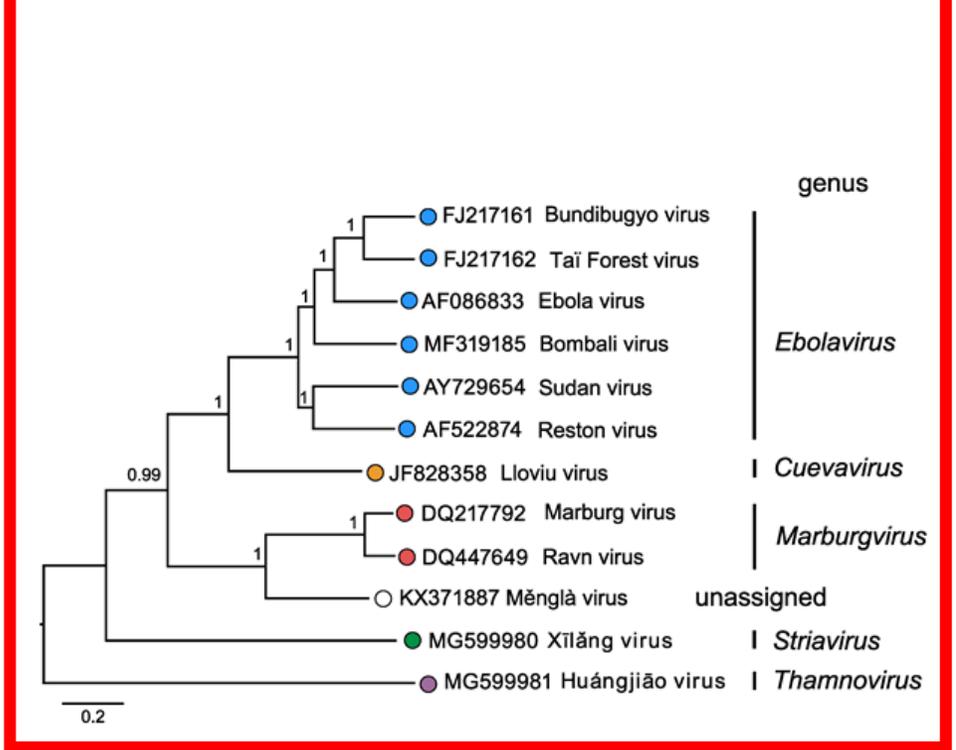
OÙ va-t-il ?

# Arbres bayésiens



QUAND est l'ancêtre commun le plus récent ?

# Maximum de vraisemblance

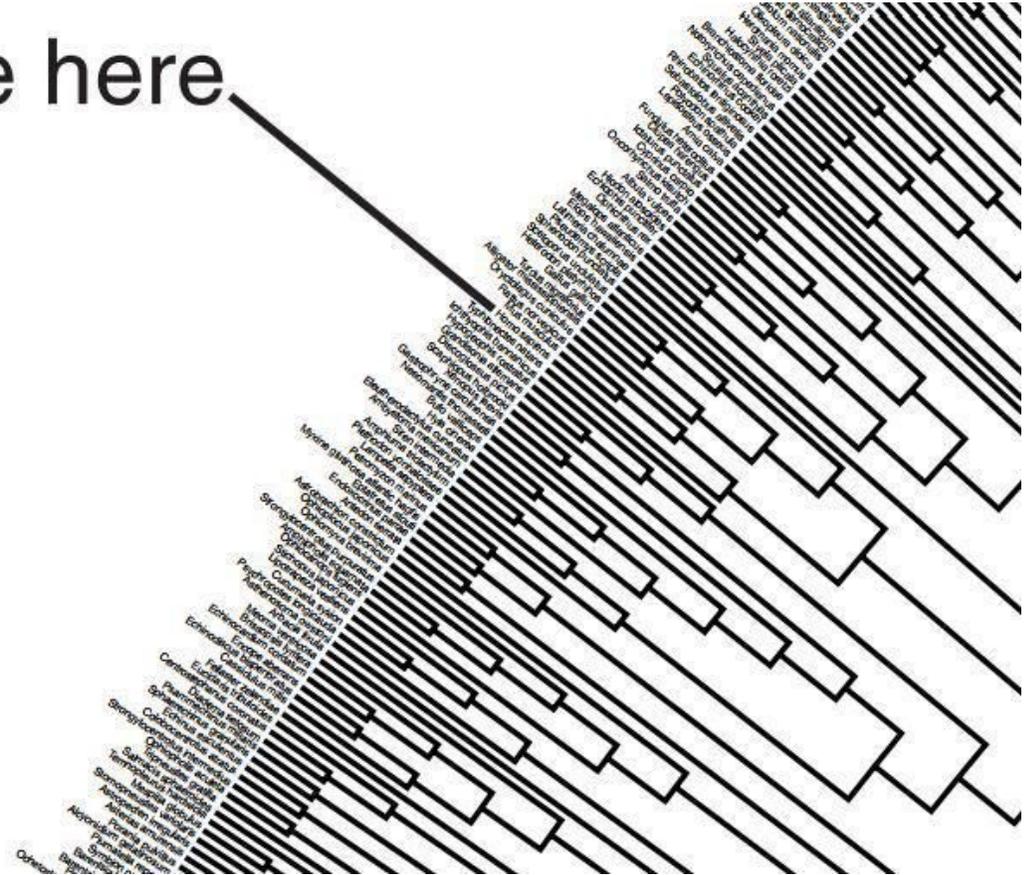


À quel point est-ce différent de ce que l'on connaît ?

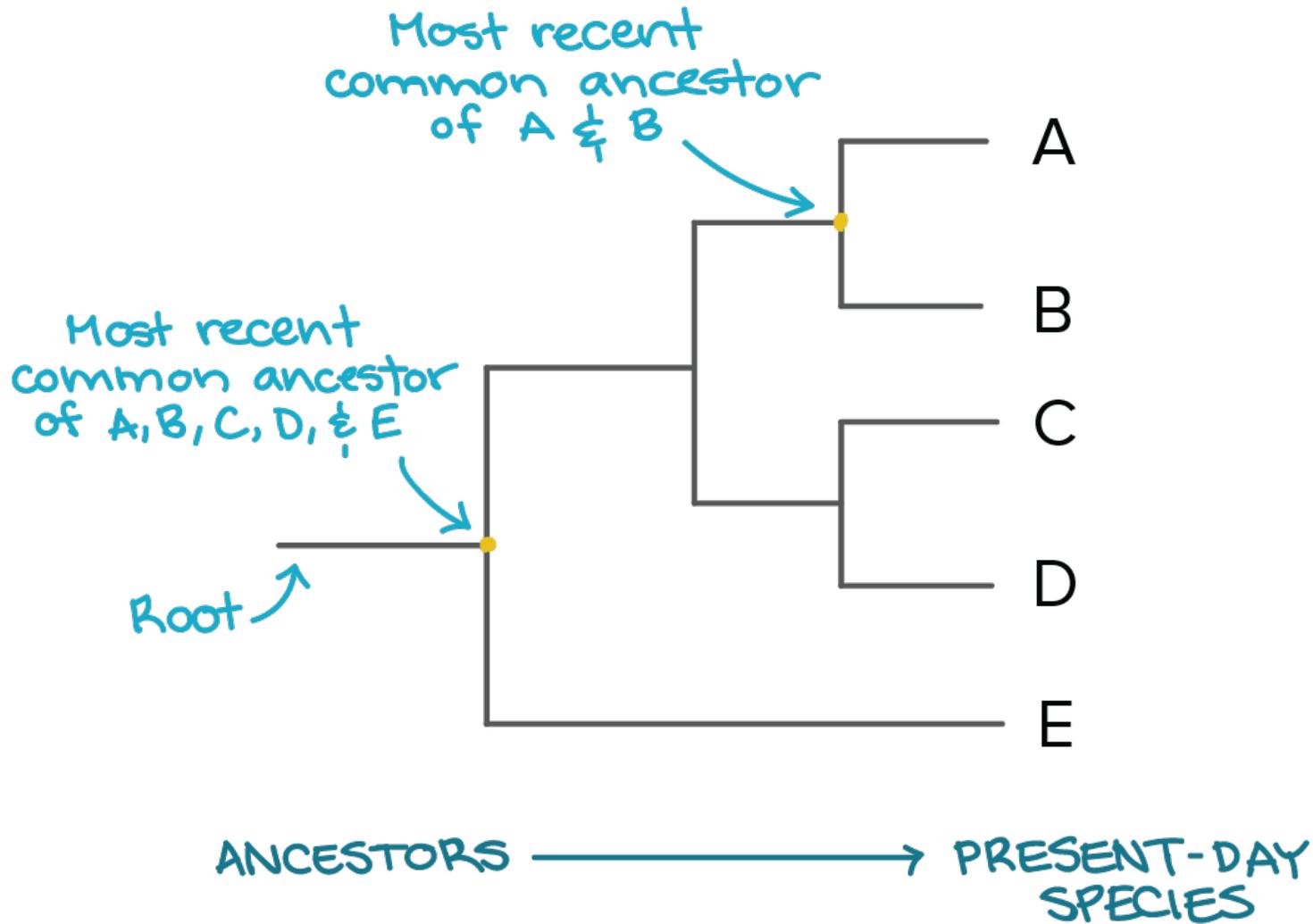
# Point de contrôle!

- Qu'est-ce qu'une phylogénie ?
- Peut-on utiliser une analyse phylogénétique avec des données temporelles ?
- Pouvez-vous utiliser une analyse phylogénétique pour voir à quel point quelque chose est similaire à un autre ?

You are here



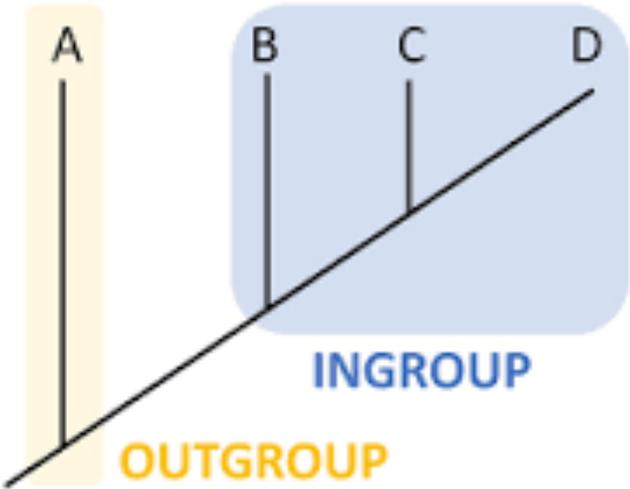
# Anatomie d'une phylogénie



## CONFIDENCE

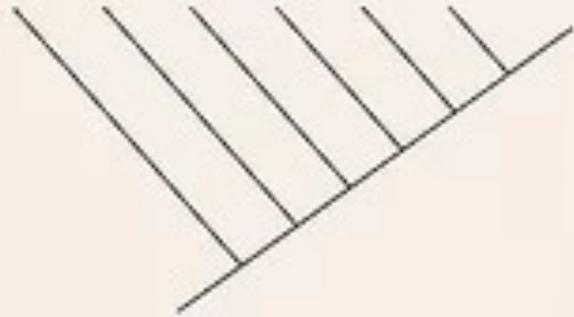
### BOOTSTRAP VALUE

STRONGLY SUPPORTED	>90%
WELL SUPPORTED	70%-90%
WEAKLY SUPPORTED	50%-70%
NOT SUPPORTED	<50%



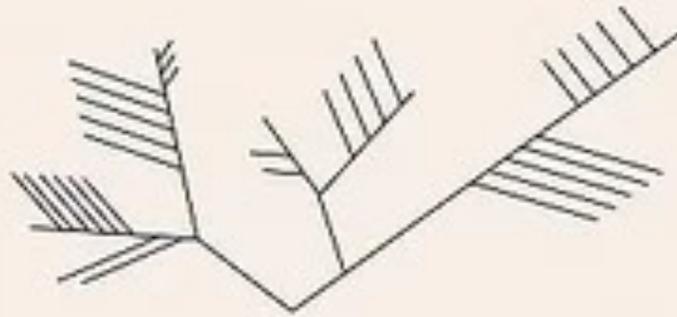
# Cladogramme versus arbre phylogénétique

## CLADOGRAM



- the relationships are *hypothetical*
- you can easily make on your own

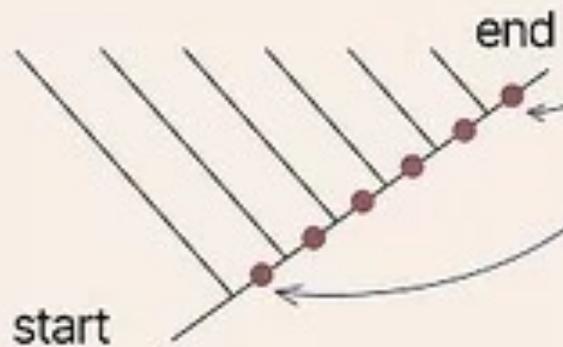
## PHYLOGENETIC TREE



- the relationships are *backed by molecular evidence*
- should have access to DNA or other molecular data

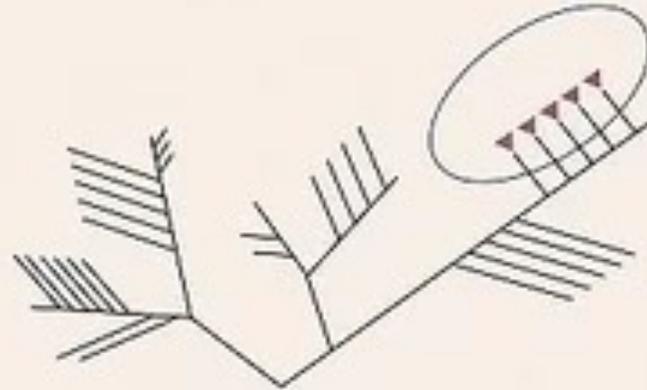
# Cladogram versus phylogenetic tree

Nodes closer to the start of the main line happened longer ago than nodes closer to the end



**CLADOGRAM**

Animals that are closer together are also more genetically similar



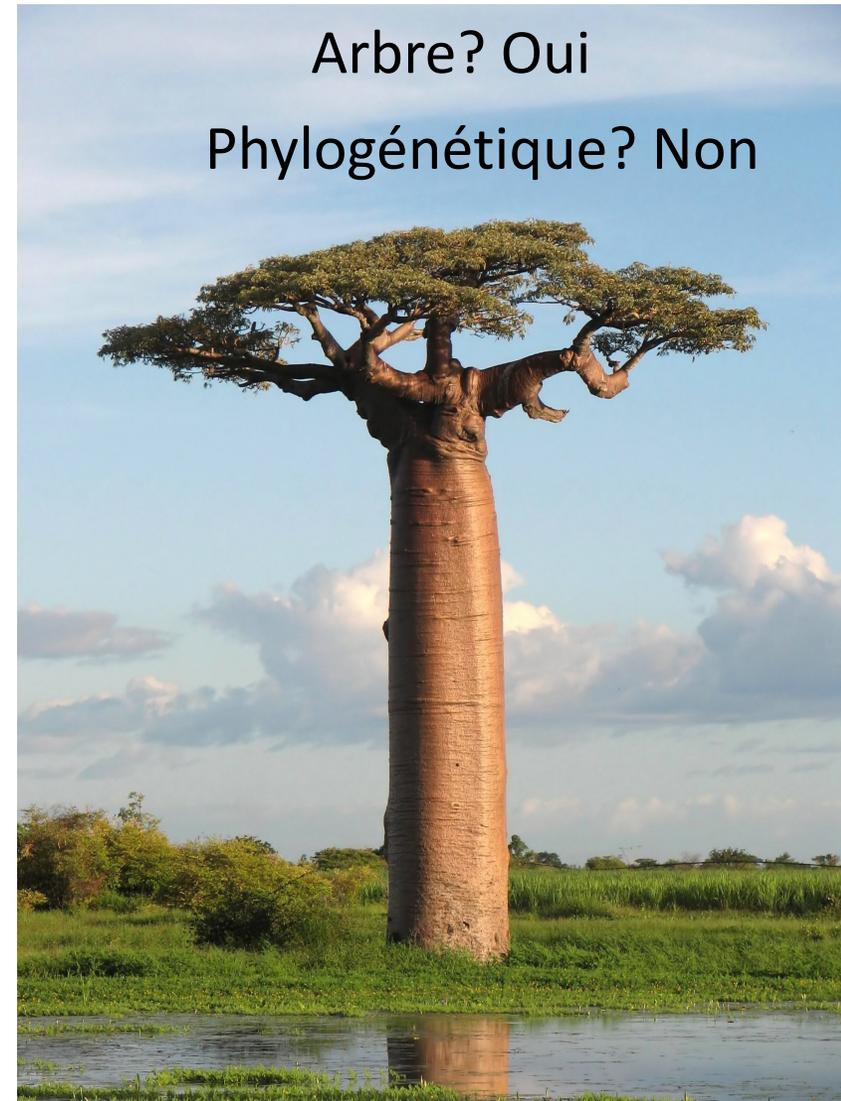
**PHYLOGENETIC TREE**

**Utiliseriez-vous un cladogramme ou un arbre phylogénétique pour :**

- **Hypothèses sur les ancêtres de l'homme**
- **Traquer un nouvel agent pathogène**

# Point de contrôle!

- Les cladogrammes sont bons pour générer des hypothèses, les phylogénies montrent des similitudes génétiques
- Que montrent les valeurs d'amorçage ?
- Qu'est-ce qu'une racine ?



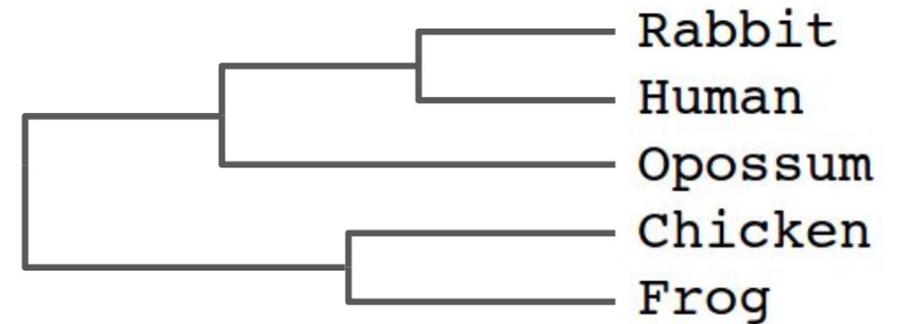
# Parcimonie vs. Vraisemblance

- Parcimonie : nombre minimum de modifications
- Vraisemblance : probabilité maximale que les données aient évolué sur l'arbre



**branch length** can mean different things:

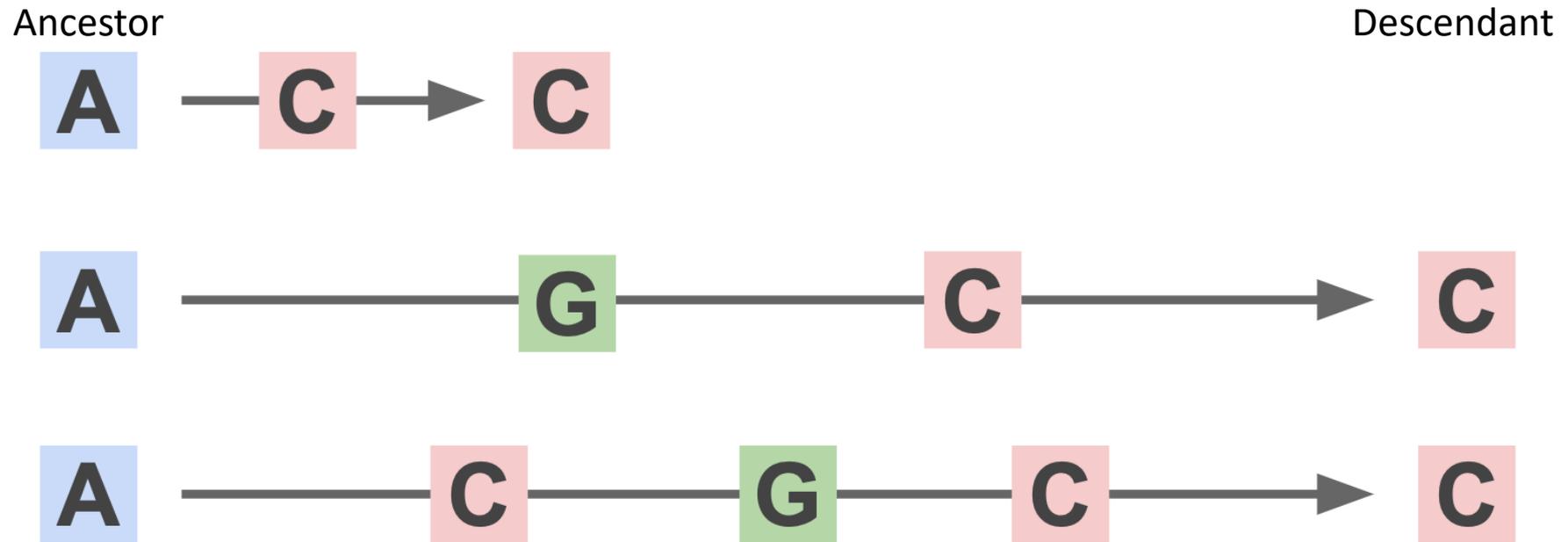
- minimum number of changes (parsimony)
- time; opportunity for change
- expected number of changes, given a model of evolution



proposed tree has **branch lengths** in units of expected number of changes per site



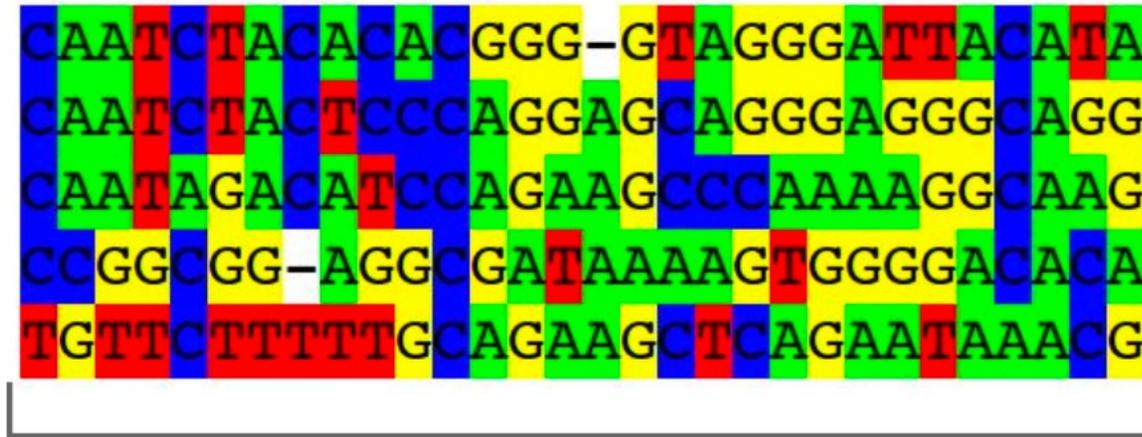
Vraisemblance : la probabilité d'un statut d'ancêtre et de descendant est fonction du temps (longueur de la branche)



Nous ne savons pas quelle est l'histoire réelle du changement, alors utilisez un modèle d'évolution pour considérer toutes les histoires possibles (probabilité maximale)

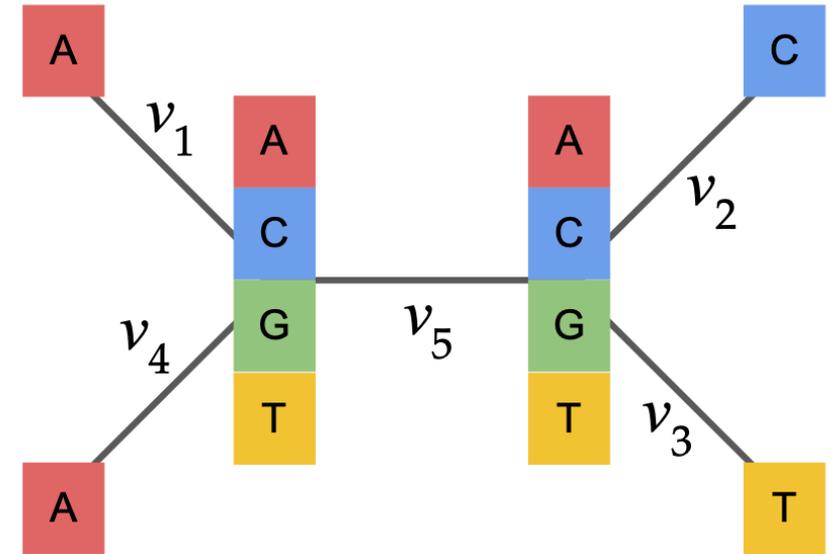
# Vraisemblance (suite).

Rabbit  
Human  
Opossum  
Chicken  
Frog



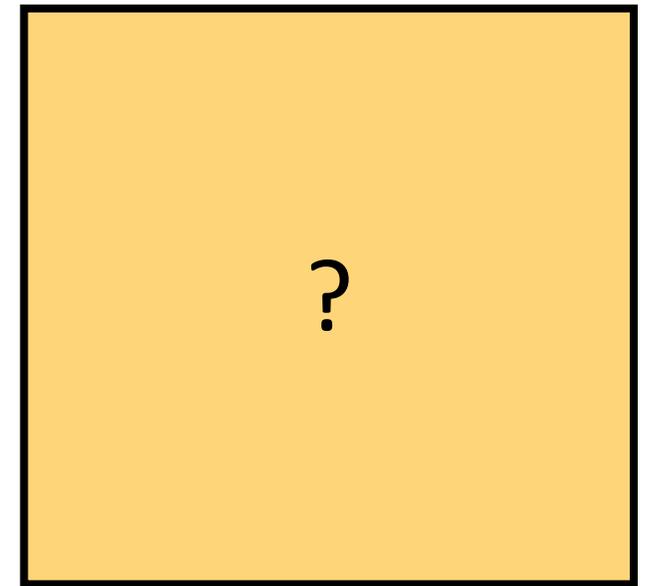
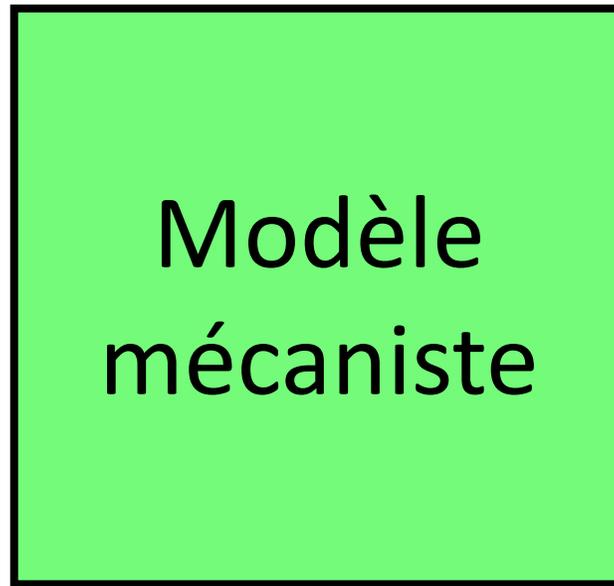
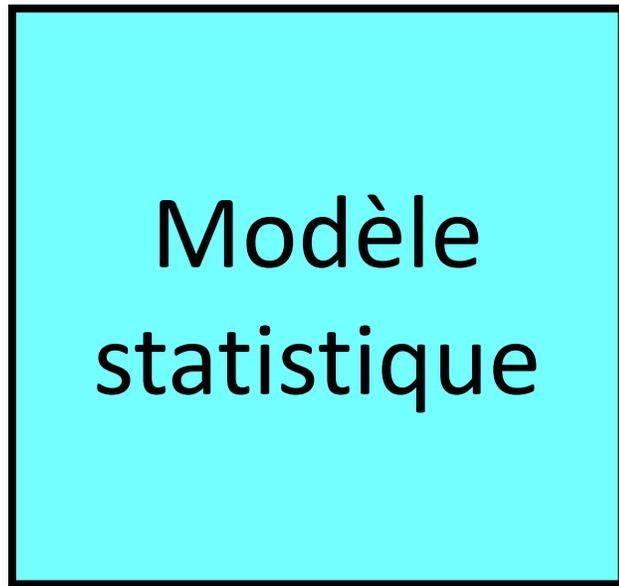
overall likelihood is the product of likelihoods across characters (sites)

Paramètres : topologie arborescente, longueurs des branches, taux de substitution estimés pour maximiser la vraisemblance des données



Consider *all possible ancestral states* at internal nodes, and calculate their contribution to the overall likelihood.\*

Jusqu'à présent...



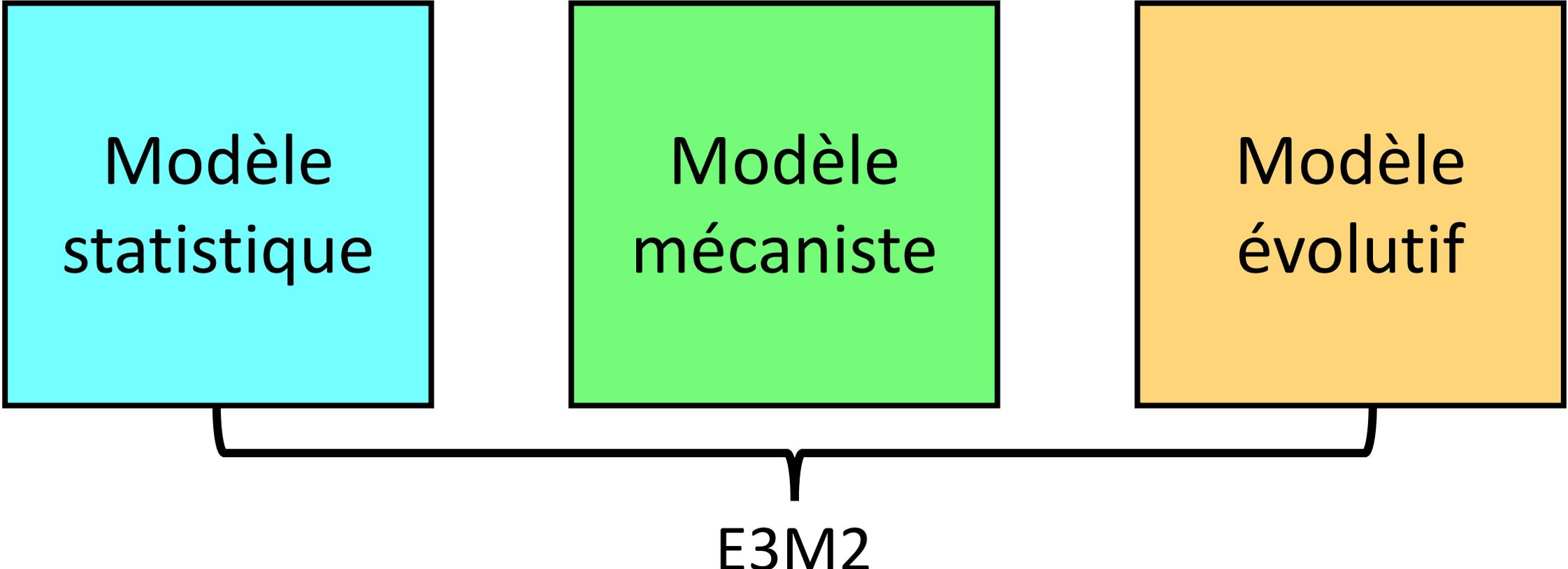
E2M2

Aujourd'hui...

Modèle  
statistique

Modèle  
mécaniste

Modèle  
évolutif



E3M2

# Modèles de l'évolution de l'ADN

- Modèles de Markov qui décrivent les taux relatifs de différents changements
  - JC69 (Jukes et Cantor 1969)
  - Modèle K80 (Kimura 1980)
  - Modèle K81 (Kimura 1981)
  - F81 (Felsenstein 1981)
  - Modèle HKY85 (Hasegawa, Kishino et Yano 1985)
  - Modèle T92 (Tamura 1992)
  - Modèle TN93 (Tamura et Nei, 1993)
  - Modèle GTR (Tavaré 1986)
  - Oui, il y en a beaucoup ! Comment puis-je savoir ce qui est le mieux pour mes données ?

# Bonne nouvelle, la plupart des gens n'ont pas besoin de connaître les spécificités mathématiques de ces modèles

## JC69 model (Jukes and Cantor 1969) [\[ edit \]](#)

JC69, the [Jukes](#) and [Cantor](#) 1969 model,<sup>[2]</sup> is the simplest [substitution model](#). There are several assumptions. It assumes equal base frequencies

$\left( \pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4} \right)$  and equal [mutation rates](#). The only parameter of this model is therefore  $\mu$ , the overall substitution rate. As previously

mentioned, this variable becomes a constant when we normalize the mean-rate to 1.

$$Q = \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix}$$

$$P = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} \end{pmatrix}$$

When branch length,  $\nu$ , is measured in the expected number of changes per site then:

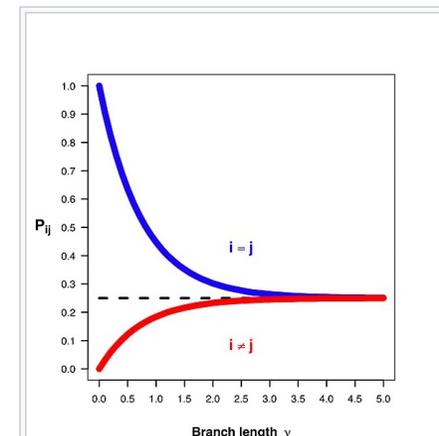
$$P_{ij}(\nu) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4\nu/3} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4\nu/3} & \text{if } i \neq j \end{cases}$$

It is worth noticing that  $\nu = \frac{3}{4}t\mu = \left( \frac{\mu}{4} + \frac{\mu}{4} + \frac{\mu}{4} \right)t$  what stands for sum of any column (or row) of matrix

$Q$  multiplied by time and thus means expected number of substitutions in time  $t$  (branch duration) for each particular site (per site) when the rate of substitution equals  $\mu$ .

Given the proportion  $p$  of sites that differ between the two sequences the Jukes-Cantor estimate of the evolutionary distance (in terms of the expected number of changes) between two sequences is given by

$$\hat{d} = -\frac{3}{4} \ln\left(1 - \frac{4}{3}p\right) = \hat{\nu}$$



Probability  $P_{ij}$  of changing from initial state  $i$  to final state  $j$  as a function of the branch length ( $\nu$ ) for JC69. Red curve: nucleotide states  $i$  and  $j$  are different. Blue curve: initial and final states are the same. After a long time, probabilities tend to the nucleotide equilibrium frequencies (0.25; dashed line).

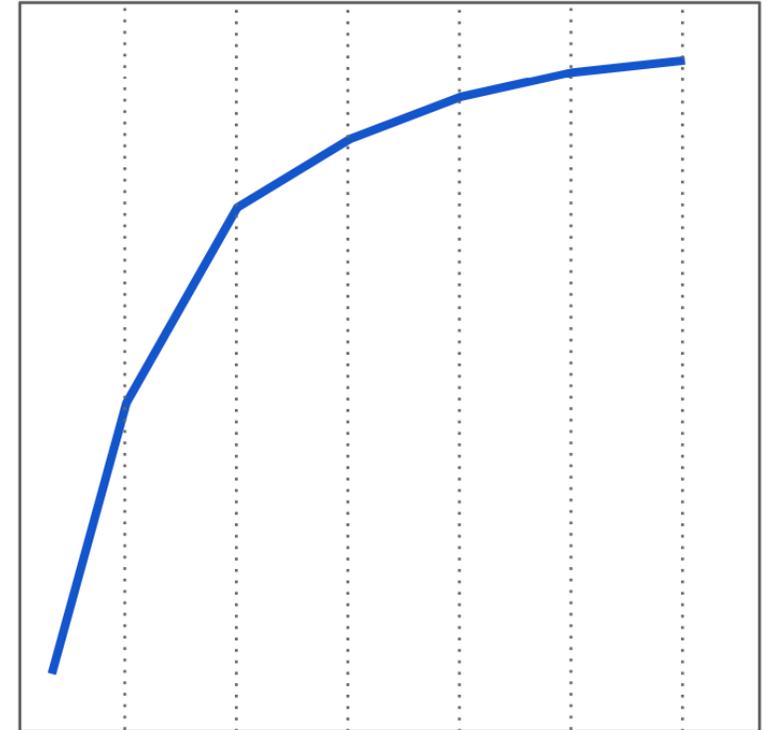
# Sélection du modèle

Plus il y a de paramètres, plus la probabilité est élevée, mais l'augmentation de la probabilité est-elle nécessaire? Ajoute beaucoup plus de complexité

Les tests de modèle vous donneront deux scores

- Score AIC : tente de sélectionner le modèle qui décrit le mieux une réalité inconnue et de haute dimension.
- Score BIC : tente de trouver le modèle VRAI parmi l'ensemble des candidats

likelihood



no. parameters

# DNA models

## Base substitution rates

IQ-TREE includes all common DNA models (ordered by complexity):

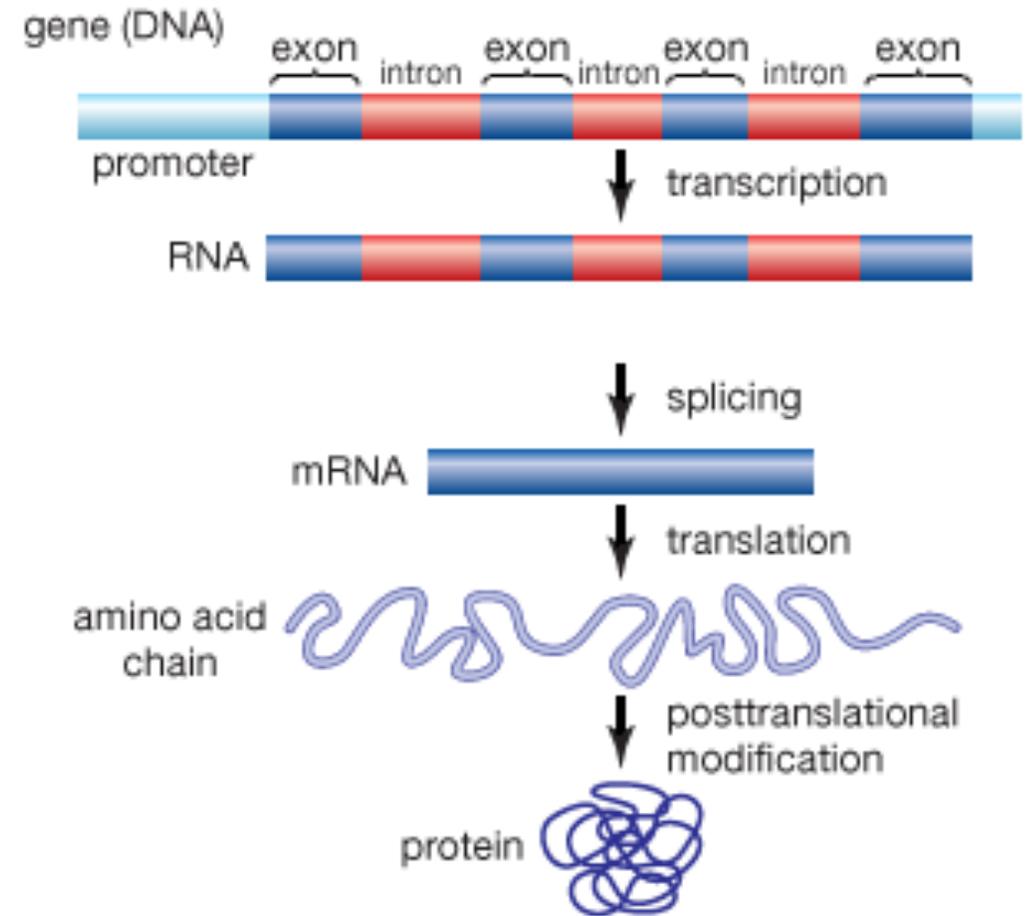
Model	df	Explanation	Code
JC or JC69	0	Equal substitution rates and equal base frequencies ( <a href="#">Jukes and Cantor, 1969</a> ).	000000
F81	3	Equal rates but unequal base freq. ( <a href="#">Felsenstein, 1981</a> ).	000000
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. ( <a href="#">Kimura, 1980</a> ).	010010
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. ( <a href="#">Hasegawa, Kishino and Yano, 1985</a> ).	010010
TN or TN93	5	Like <b>HKY</b> but unequal purine/pyrimidine rates ( <a href="#">Tamura and Nei, 1993</a> ).	010020
TNe	2	Like <b>TN</b> but equal base freq.	010020
K81 or K3P	2	Three substitution types model and equal base freq. ( <a href="#">Kimura, 1981</a> ).	012210
K81u	5	Like <b>K81</b> but unequal base freq.	012210
TPM2	2	AC=AT, AG=CT, CG=GT and equal base freq.	010212
TPM2u	5	Like <b>TPM2</b> but unequal base freq.	010212
TPM3	2	AC=CG, AG=CT, AT=GT and equal base freq.	012012
TPM3u	5	Like <b>TPM3</b> but unequal base freq.	012012
TIM	6	Transition model, AC=GT, AT=CG and unequal base freq.	012230
TIME	3	Like <b>TIM</b> but equal base freq.	012230

TIM2	6	AC=AT, CG=GT and unequal base freq.	010232
TIM2e	3	Like <b>TIM2</b> but equal base freq.	010232
TIM3	6	AC=CG, AT=GT and unequal base freq.	012032
TIM3e	3	Like <b>TIM3</b> but equal base freq.	012032
TVM	7	Transversion model, AG=CT and unequal base freq.	012314
TVMe	4	Like <b>TVM</b> but equal base freq.	012314
SYM	5	Symmetric model with unequal rates but equal base freq. ( <a href="#">Zharkikh, 1994</a> ).	012345
GTR	8	General time reversible model with unequal rates and unequal base freq. ( <a href="#">Tavare, 1986</a> ).	012345

# Évaluer l'hétérogénéité entre les sites

- Évolution de l'hétérogénéité des taux :
- Positions des codons
- Exons (régions codantes) et introns (régions non codantes)
- Gènes d'entretien ménager et gènes non fonctionnels
- Structure de l'ARN (tiges et boucles)

Nous pouvons faire des déductions sur la sélection à partir de ces valeurs, mais cela rend les choses beaucoup plus compliquées

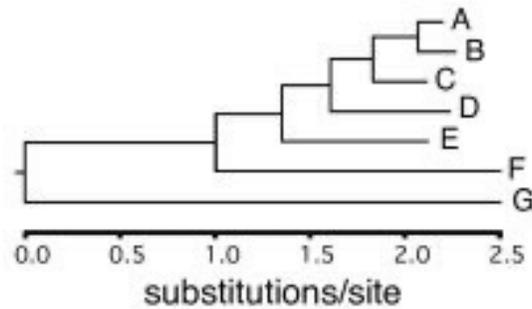


# Évaluer l'hétérogénéité entre les sites

## AVEC hétérogénéité de taux

```

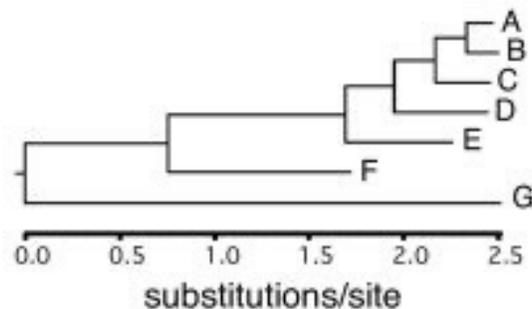
A TTEQGIKSSSTSLPAPQLPNWSGOYHEWVLKS---FQNEVK----KTLHCALSQGTATQSVLDELHADVWALLASSEVCYARPCGOVKPELAFLRKRA
B TAGAGEKTGTSLPAPNLPNWSGOYHEWVLKS---ERADVI----KTMHCRA LSDGI TATQSVLDELHFNVWALLASSEVCYARPCGDQKPELIYMKYKQA
C VAGGCEKAGTSLPAPYLPN--SGOYGEWVLKS---LSTHVI----KHMICEDLSDSDTITQSVLDELHGERWALLOSSEVCYARPCGOEKRVLHEHECYKRA
D CAGQAEKTGTSLPALHLPNWA--QYGEWVLKS---FPSQPV----MPIQCVPLSDARTAAQSVLDELHVESDALLDSSEVCYARPCGA-RHDLKFCVYSKA
E DEGLTQNTGTSLPALALPNWSGOYHEWVLKS---YG---FGQGGGAHCKPLSGDKTSIHSVLELHVAVLAALLMSGVCYALPCGAYKKALEFKCYLKA
F GEGFIKKTGTSAPAVALPDAEQYDEWP LKS---LAYGRVNF----AAVPGAYLSDPFGTSSHVLELHONHAALLLSSEVCFAPPCGESKGA LVVV CYSHA
G NDGPHIKKTS SPANL PNDPI DYDEWDLKS---CEAKSI----NGSNWKPLSGKYTGLDVAVLELHAMKDALLHATEVCLAPPCGY-TADLKAALGPA
  
```



## SANS hétérogénéité de taux

```

A H----GENYFC--SQVAKYLAF-YSHNLL--EALLRHATLEIQH--KSNSEHGTLGEPESA--EPR--VPAQNEKLLGKVVNNEFSAPGL----IKNP
B Q----GENYFC--GGVAKYLAW-VGHNLL--EALLRHATMEINR--KSKKEQNGLDGPEGA--EPR--IPAGGEKLLGMVHNMF GAAGL----VKMP
C S----GENYFC--PQVAKYLAW-MSNNLL--HAFITQAKLEIER--KRNQVEHGCGLDGNQD--EPR--IKNSGQXLLGGY--KELKNPGL----FVKP
D Q----AHQYFC--SHIGKYFAW-VANAFH--HVLLRYAKLEVER--KRTADHSTDLVAQNGA--KSV--LLPQPDALL--RHHKFIISTPLA----FIKT
E E----EDQYFC--KEMAKYLAW-VGHGYLERAHLSKHAKLATEK--KTEADHNTKLETAEQ--LVFC--IPPLPDRVAIYANTFFSAQHL----FIKT
F Q----GKKDLC--ENLN----TW-MQNRW--QALHK-TITVVQHDGKSSMGDHCCKAIDSKAS--LSPC--VSSGGYVQKSNQIDFFVSNVTV----YLKS
G NNDFSKPFLFCNYTGIL---ILQCAG-----YLDGETMIGRFQ--STQVGLYSERLFDPRYKCMGETHKATNNTDTFGDRKAPKRRVSVKAFKQQTAPQ
  
```



# Rate heterogeneity across sites

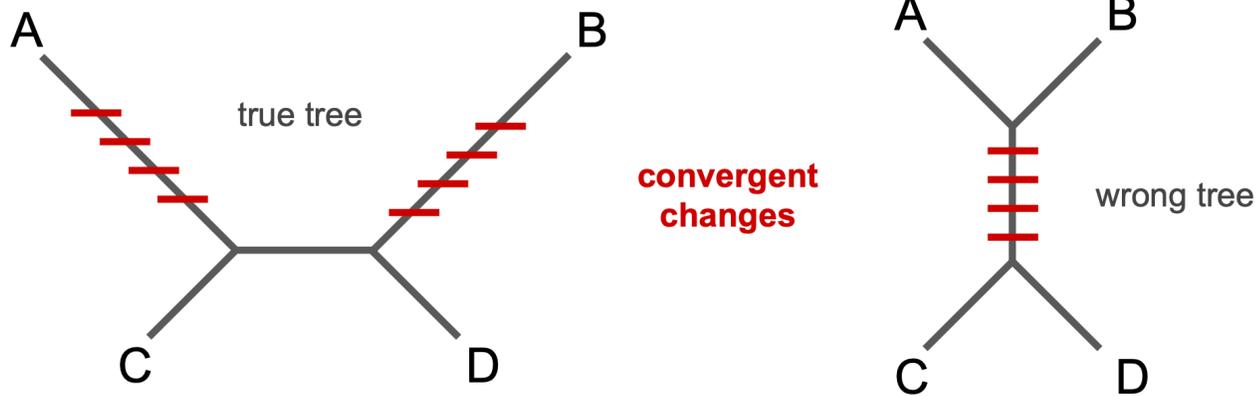
---

IQ-TREE supports all common rate heterogeneity across sites models:

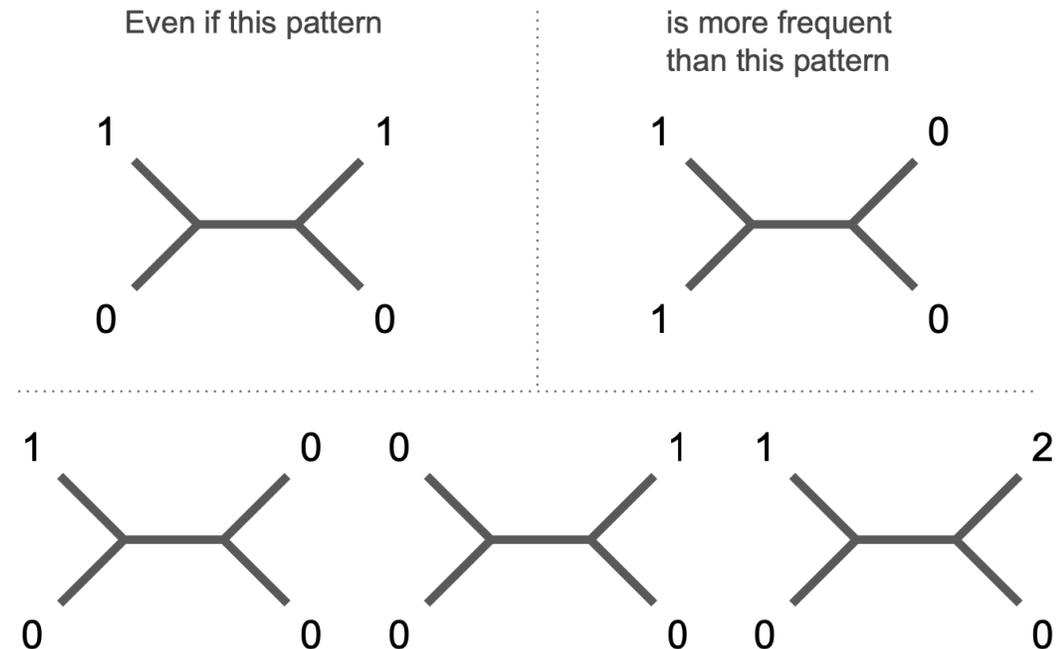
RateType	Explanation
+I	allowing for a proportion of invariable sites.
+G	discrete Gamma model (Yang, 1994) with default 4 rate categories. The number of categories can be changed with e.g. +G8 .
+GC	continuous Gamma model (Yang, 1994) (for AliSim only).
+I+G	invariable site plus discrete Gamma model (Gu et al., 1995).
+R	FreeRate model (Yang, 1995; Soubrier et al., 2012) that generalizes the +G model by relaxing the assumption of Gamma-distributed rates. The number of categories can be specified with e.g. +R6 (default 4 categories if not specified). The FreeRate model typically fits data better than the +G model and is recommended for analysis of large data sets.
+I+R	invariable site plus FreeRate model.

# Zone de Felsenstein

- Les longueurs de branche pour lesquelles la parcimonie déduit avec certitude une topologie incorrecte, celles-ci peuvent affecter les valeurs d'amorçage



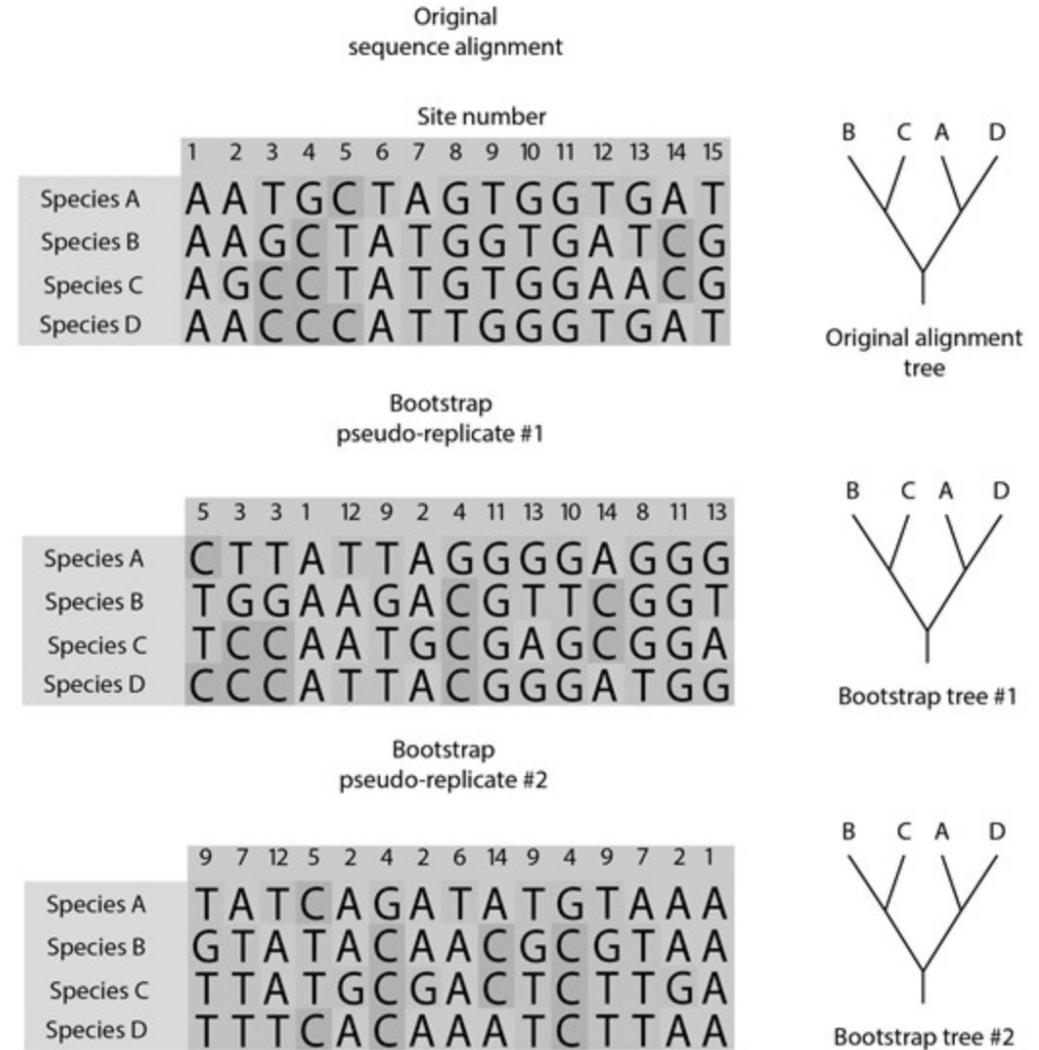
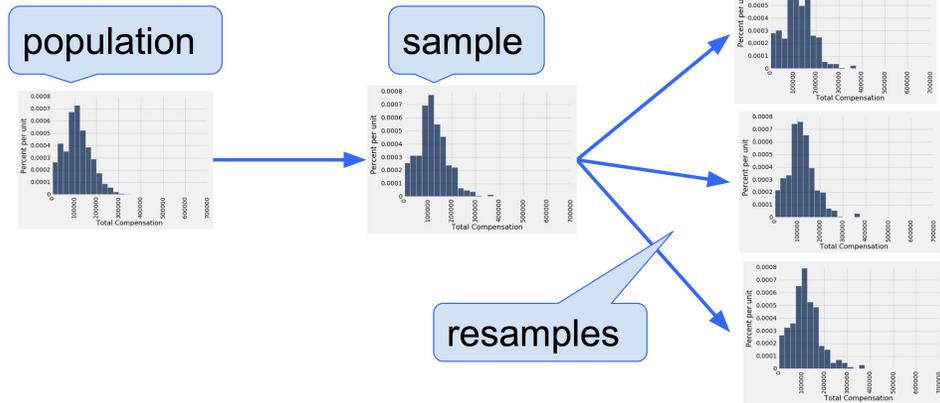
likelihood is a **consistent estimator** of tree topology because it converges on the correct value with increasing data



likelihood will correctly infer the true tree if these patterns are sufficiently frequent to allow accurate branch length estimation

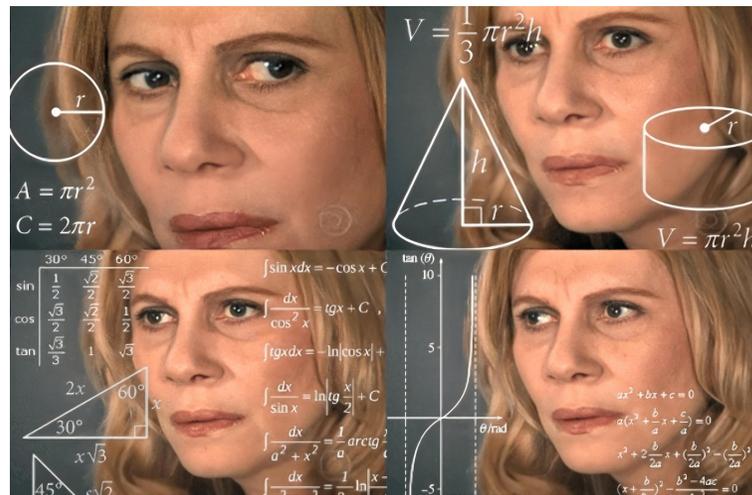
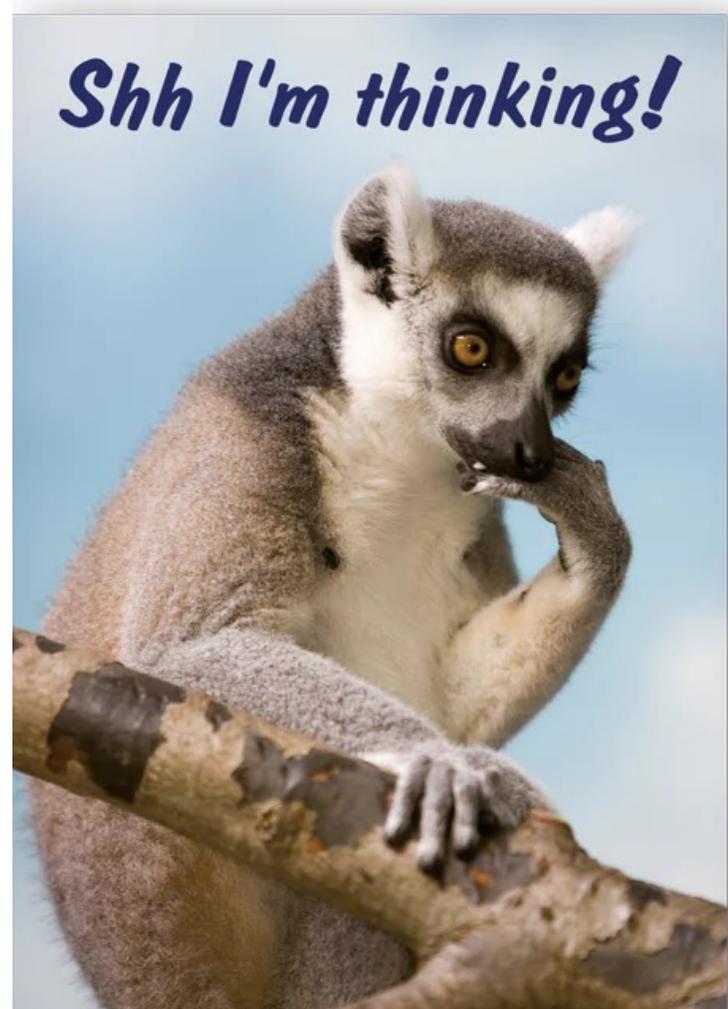
# Bootstrapping

- Spécifier le nombre de répétitions : combien de fois le test réplique-t-il l'alignement de la séquence d'origine?



# Point de contrôle!

- Plus de paramètres = plus ou moins de probabilité?
- Des valeurs d'amorçage faibles signifient une faible confiance dans la topologie de l'arborescence
- Toutes les séquences ne subissent pas l'évolution de la même manière, les modèles rendent compte de ce changement



# Avertissements et limitations

- La construction de phylogénies prend beaucoup de temps, les plus grandes peuvent prendre jusqu'à une semaine pour s'exécuter et les phylogénies bayésiennes peuvent fonctionner pendant des mois, donc un cluster de calcul est presque nécessaire pour cela
- En l'absence d'un exogroupe ou d'une racine appropriée, une phylogénie ne vous dit pas grand-chose sur l'ordre de descendance

**M11: Progress**

PROGRESS

Site coverage calculated

DETAILS ✖ STOP

STATUS/OPTIONS

RUN STATUS	
Start time	11-12-22 00:39:48
Operation Run Time	05:17:58
Status	Setting site coverage
Log Likelihood	-6,701.73
Operation	Bootstrapping ML tree
Replicate No.	227 of 500

# Assessing Uncertainty in the Rooting of the SARS-CoV-2 Phylogeny

Lenore Pipes, Hongru Wang, John P Huelsenbeck , Rasmus Nielsen 

*Molecular Biology and Evolution*, Volume 38, Issue 4, April 2021, Pages 1537–1543,

<https://doi.org/10.1093/molbev/msaa316>

**Published:** 09 December 2020

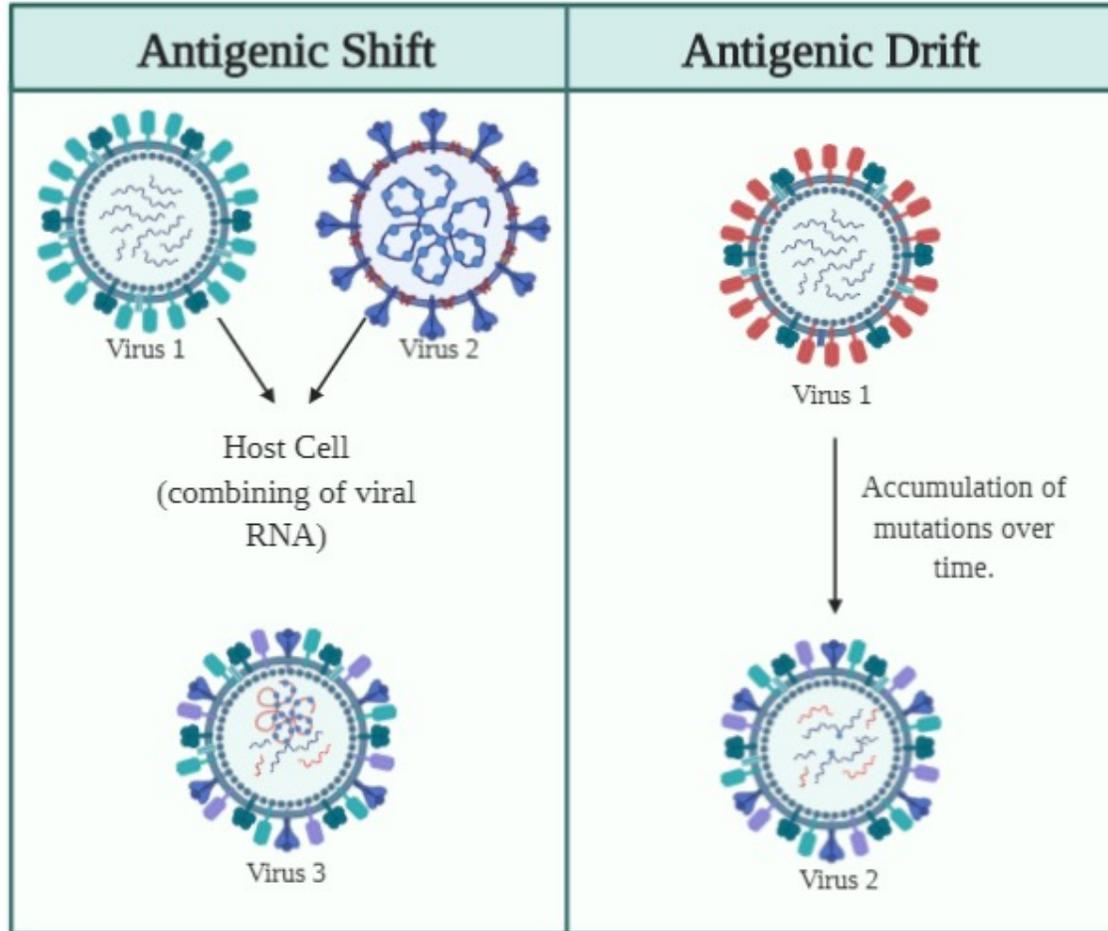
 PDF  Split View  Cite  Permissions  Share ▼

## Abstract

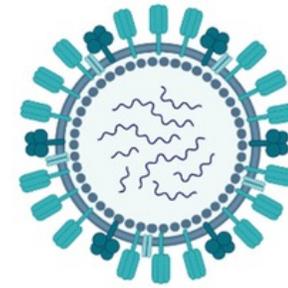
The rooting of the SARS-CoV-2 phylogeny is important for understanding the origin and early spread of the virus. Previously published phylogenies have used different rootings that do not always provide consistent results. We investigate several different strategies for rooting the SARS-CoV-2 tree and provide measures of statistical uncertainty for all methods. We show that methods based on the molecular clock tend to place the root in the B clade, whereas methods based on outgroup rooting tend to place the root in the A clade. The results from the two approaches are statistically incompatible, possibly as a consequence of deviations from a molecular clock or excess back-mutations. We also show that none of the methods provide strong statistical support for the placement of the root in any particular edge of the tree. These results suggest that phylogenetic evidence alone is unlikely to identify the origin of the SARS-CoV-2 virus and we caution against strong inferences regarding the early spread of the virus based solely on such evidence.

**Il est important de  
replacer votre  
arbre dans son  
contexte, sans  
contrôle, pouvez-  
vous vraiment en  
déduire quoi que  
ce soit?**

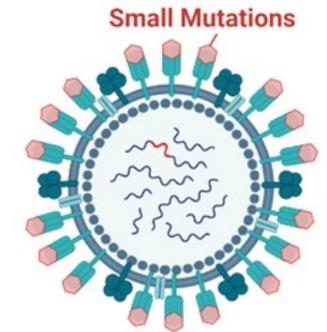
# Étude de cas : grippe – décalage et dérive antigéniques



(A) Antigenic Drift

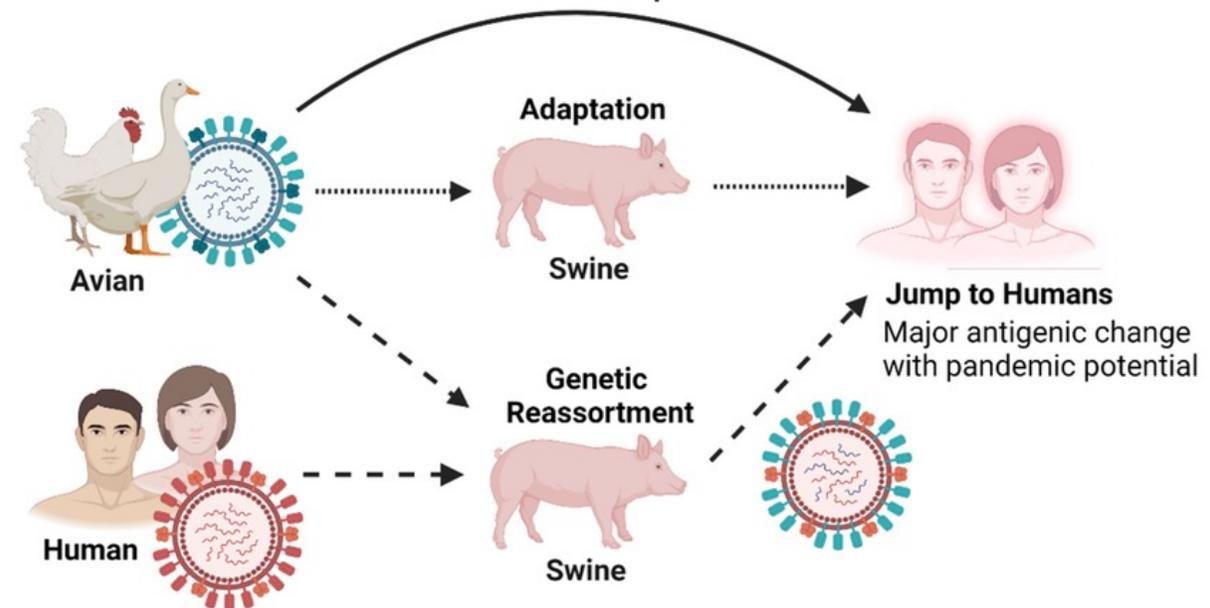


Accumulation of Mutations



Minor antigenic change with epidemic potential

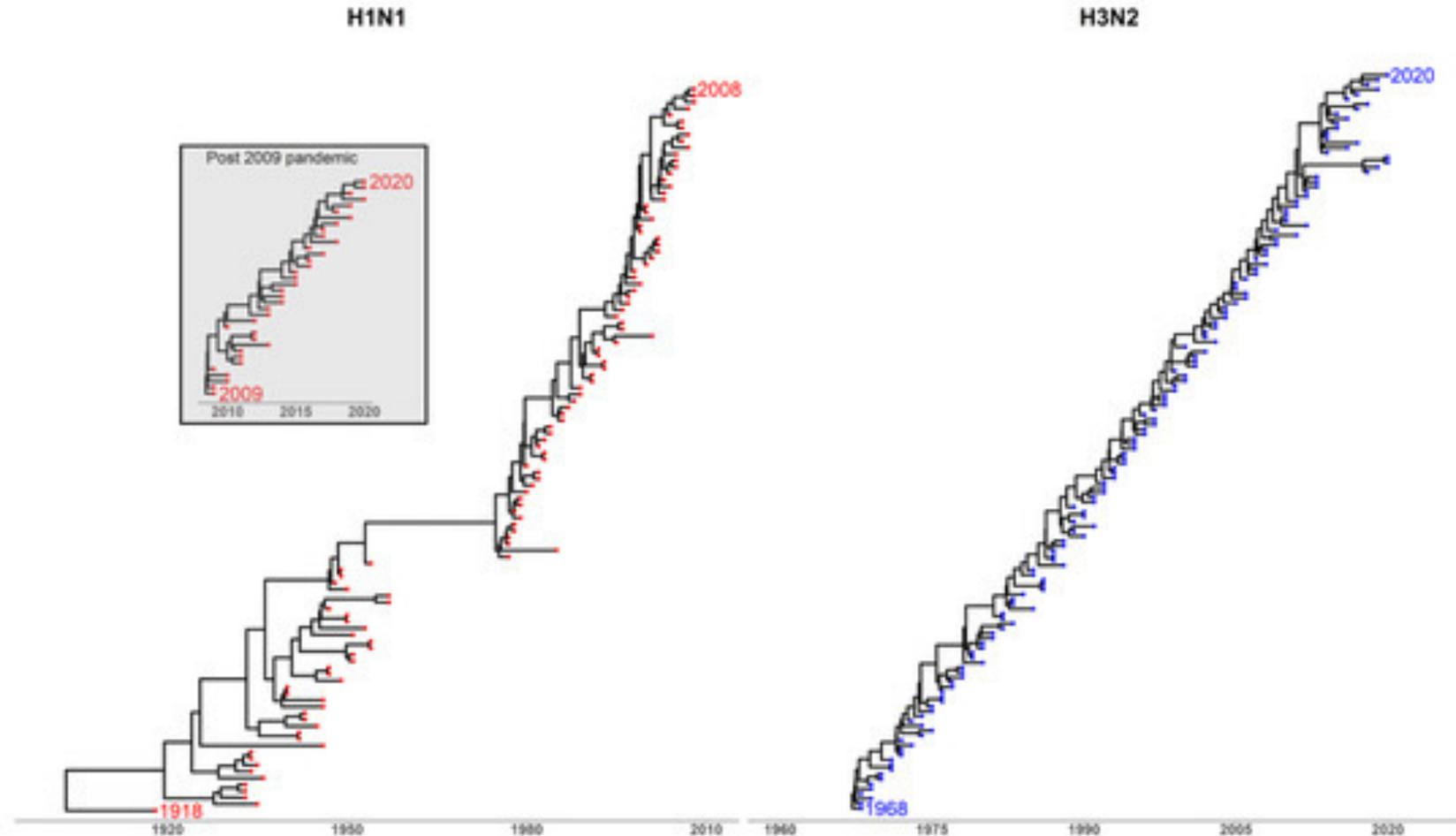
(B) Antigenic Shift



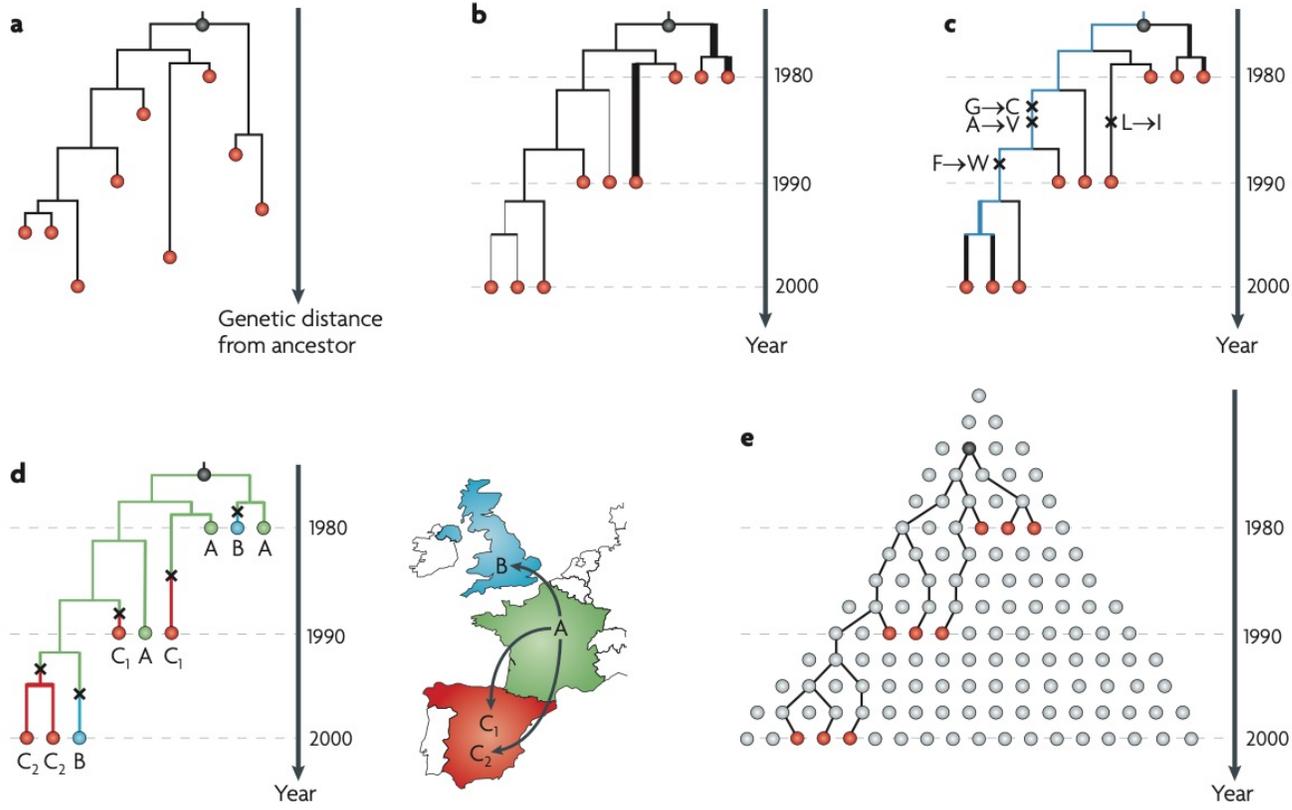
# Étude de cas : grippe – décalage et dérive antigéniques



**Y a-t-il une dérive ou un décalage ici ?  
Quel impact cela a-t-il sur la conception des vaccins ?**



Box 1 | **Phylodynamic techniques**



# L'importance des formes

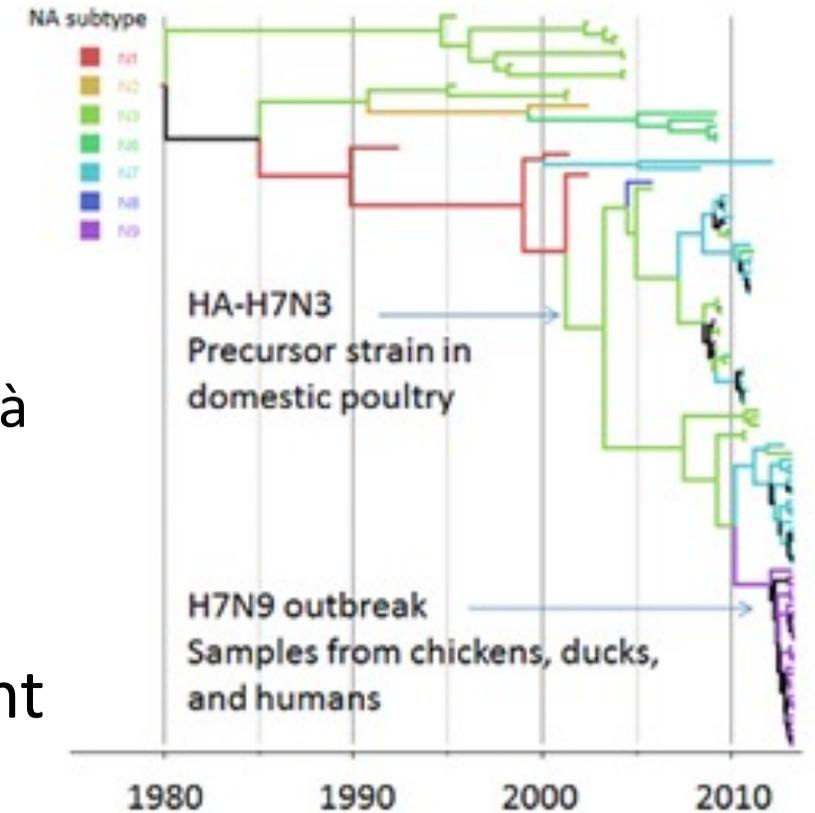
	Continual Immune Selection	Weak or Absent Immune Selection	
		Tree shape controlled by non-selective population dynamic processes	
Idealized Phylogeny Shapes	<p>Time →</p>	Population size dynamics	Spatial dynamics
		<p><b>Exponential growth</b></p> <p><b>Constant size</b></p>	<p><b>Strong spatial structure</b></p> <p><b>Weak spatial structure</b></p>
<b>Examples</b>	Human influenza A virus intra-host HIV	inter-host HIV inter-host HCV	Measles, rabies inter-host HIV
<b>Tree Inferences</b>	Detection of antigenic escape mutations	Estimation of population growth rates	Estimation of population migration rates

# L'importance des unités

# On peut faire beaucoup de choses avec les phylogénies...

- Phylodynamique
- Peut examiner la phylogénie dans le contexte d'autres facteurs
  - Temps (depuis combien de temps ce virus a-t-il divergé)
  - Emplacement (comment un virus a-t-il changé au fur et à mesure qu'il se propage ?)
  - Hôte (comment un virus a-t-il changé dans différents hôtes ?)
- Les phylogénies du maximum de vraisemblance sont bonnes pour :
  - À quel point une chose est différente/similaire par rapport aux choses connues

Phylogeny of Hemagglutinin subtype H7 and reassortments with different Neuraminidases

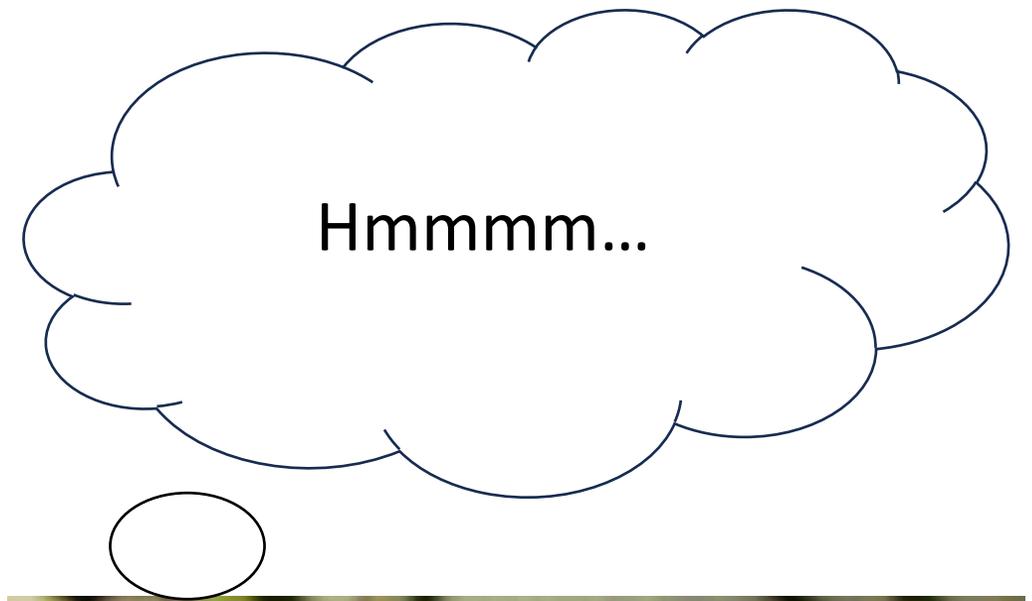


BEAST

Bayesian Evolutionary Analysis Sampling Trees

# Point de contrôle!

- La phylogénétique est utile, mais elle demande beaucoup de calculs
- Pourquoi avons-nous besoin d'une racine dans les arbres phylogénétiques ?
- La forme d'une phylogénie peut-elle nous dire quelque chose sur le sujet ?



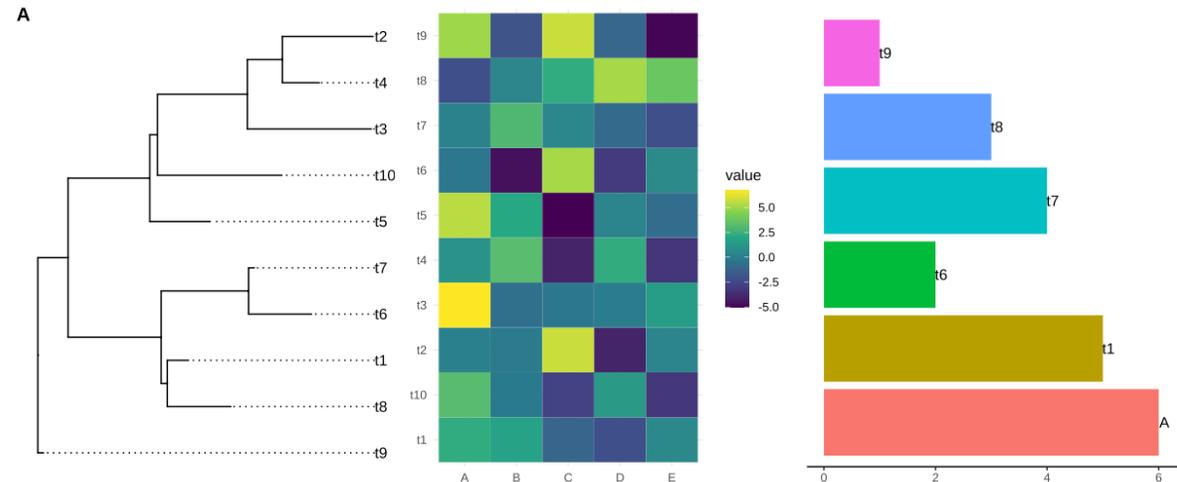
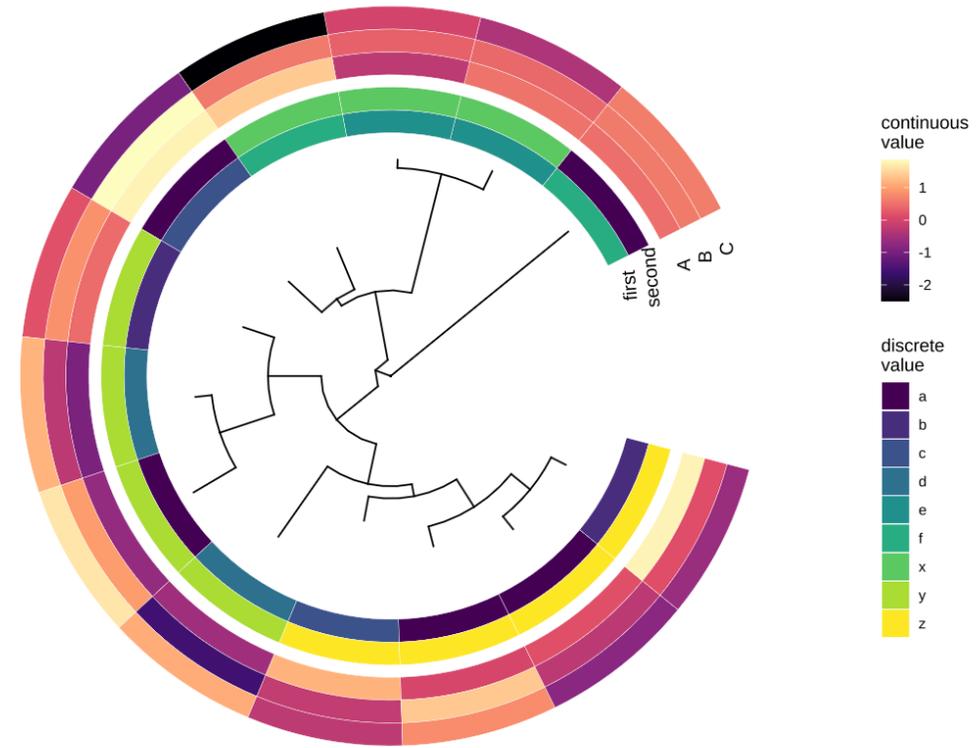
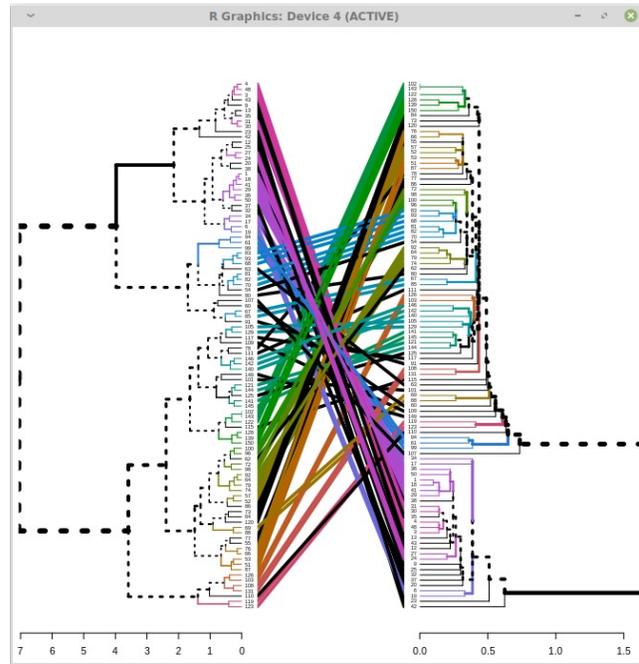
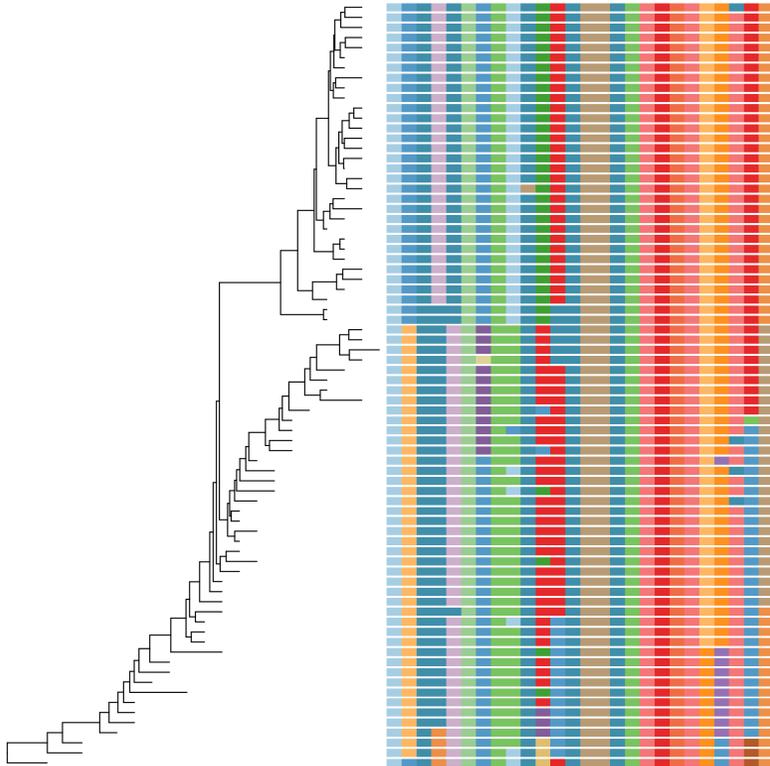
# Vous avez donc vos séquences, et maintenant?

- Obtenir des séquences de référence du NCBI
- Obtenir un groupe externe de NCBI
- Alignez-les (utilisez un logiciel comme MEGA ou en ligne, vous avez donc vos séquences, et maintenant ? comme MAFFT)
- Choisissez le meilleur modèle (utilisez un logiciel comme MEGA ou ModelTest-NG)
- Exécutez la phylogénie en utilisant les séquences alignées et le modèle choisi (utilisez un logiciel comme MEGA ou RAxML)
- Visualiser/modifier l'arborescence dans R ou MEGA

Tout ce qui est répertorié est gratuit à utiliser

# Potentiel pour de jolies figurines

**ggtree: an R package for visualization of tree and annotation data**



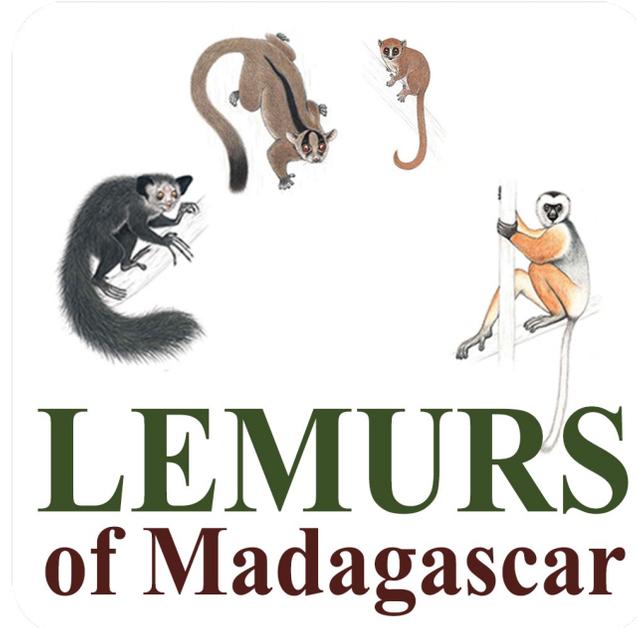
# Point de contrôle final!

- Les arbres phylogénétiques ont un large éventail d'applications à de nombreux sujets de recherche!
- Utilise la modélisation et les statistiques "en coulisses"
- **Quelqu'un peut-il donner un exemple de la façon dont il pourrait utiliser une phylogénie pour ses propres recherches? 😊**



# Lémuriens du parc national de Ranomafana

- Cytochrome B
  - Utilisé beaucoup dans l'identification des espèces, variabilité limitée à l'intérieur et variation beaucoup plus grande entre les espèces
- Question : à quel point les lémuriens que l'on peut trouver dans le parc national de Ranomafana sont-ils similaires les uns aux autres ?



DIAPPOSITIVES À REVOIR À VOTRE RYTHME

# Étapes à revoir plus tard

**NIH** National Library of Medicine  
National Center for Biotechnology Information

Log in

BLAST® Home Recent Results Saved Strategies Help

### Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

**NEWS**

**BLAST+ 2.13.0 is here!**  
Starting with this release, we are including the blastn\_vdb and tblastn\_vdb executables in the BLAST+ distribution.  
Thu, 17 March 2022 [More BLAST news...](#)

### Web BLAST

**Nucleotide BLAST**  
nucleotide ► nucleotide

**blastx**  
translated nucleotide ► protein

**tblastn**  
protein ► translated nucleotide

**Protein BLAST**  
protein ► protein

Vérifiez à quel type de séquence vous avez affaire en effectuant une recherche BLAST

### Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

```
GGACAAGTAGCCTCCATTCTATACTTTTCTCTAATCCTTATTATTATACCAAC
TGTAAGCCTCATCGAAA
ACAAGATACTTAAATGAAGA
```

Query subrange [?](#)

From

To

Or, upload file  no file selected [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

### Choose Search Set

Database  Standard databases (nr etc.):  rRNA/ITS databases  Genomic + transcript databases  Betacoronaviru

Nucleotide collection (nr/nt)  [?](#)

Limit by  Organism  BioProjectID  WGS Project

Organism   exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Exclude  Models (XM/XP)  Uncultured/environmental sample sequences

Limit to  Sequences from type material

Entrez Query  [YouTube](#) [Create custom database](#)

Enter an Entrez query to limit search [?](#)

### Program Selection

Optimize for  Highly similar sequences (megablast)  
 More dissimilar sequences (discontiguous megablast)  
 Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)

**BLAST**

Search using **Megablast (Optimize for highly similar sequences)**

Show results in a new window

[< Edit Search](#) [Save Search](#) [Search Summary](#)

[?](#) How to read this report? [BLAST Help Videos](#) [Back to Traditional Results Page](#)

Job Title **NC\_035562.1:14221-15360 Microcebus rufus**  
RID [TBKASV0K016](#) Search expires on 12-12 19:30 pm [Download All](#)   
Program BLASTN [?](#) [Citation](#)   
Database nt [See details](#)   
Query ID lclQuery\_55759  
Description NC\_035562.1:14221-15360 Microcebus rufus isolate HAB ...  
Molecule type dna  
Query Length 1140  
Other reports [Distance tree of results](#) [MSA viewer](#) [?](#)

### Filter Results

Organism *only top 20 will appear*  exclude

[+ Add organism](#)

Percent Identity  to  E value  to  Query Coverage  to

**Descriptions** [Graphic Summary](#) [Alignments](#) [Taxonomy](#)

### Sequences producing significant alignments

[Download](#)  [Select columns](#)  Show  [?](#)

select all *100 sequences selected* [GenBank](#) [Graphics](#) [Distance tree of results](#) [MSA Viewer](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	<a href="#">Microcebus rufus isolate_HAB06.12_mitochondrion_complete_genome</a>	<a href="#">Microcebus rufus</a>	2106	2106	100%	0.0	100.00%	16819	<a href="#">KM112297.1</a>
<input checked="" type="checkbox"/>	<a href="#">Microcebus rufus isolate_VEV7.13_mitochondrion_complete_genome</a>	<a href="#">Microcebus rufus</a>	1751	1751	100%	0.0	94.39%	16822	<a href="#">KM112317.1</a>

All Databases  Search

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

### Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

#### Submit

Deposit data or manuscripts into NCBI databases



#### Download

Transfer NCBI data to your computer



#### Learn

Find help documents, attend a class or watch a tutorial



#### Develop

Use NCBI APIs and code libraries to build applications



#### Analyze

Identify an NCBI tool for your data analysis task



#### Research

Explore NCBI research and collaborative projects



### COVID-19 Information

### Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

### NCBI News & Blog

- Join NCBI at PAG 30 08 Dec 2022  
San Diego, January 13-18, 2023 NCBI is looking forward to seeing you in person at the International Plant and Animal
- Announcing the NCBI SARS-CoV-2 Variant Calling Pipeline and Related Data Products 01 Dec 2022  
Still waiting for an analysis pipeline that
- New Proximity Search Feature Available in PubMed 30 Nov 2022  
PubMed, a free National Library of

Allez au NCBI, et recherchez la chose pour laquelle vous voulez construire une phylogénie, dans notre cas le cytochrome B des lémurien dans le parc national de Ranomafana

Search NCBI

Eulemur rufifrons cytochrome B



Search

Results found in 4 databases

Literature	Genes	Proteins
Bookshelf 0	Gene 0	Conserved Domains 0
MeSH 0	GEO DataSets 0	Identical Protein Groups 5
NLM Catalog 0	GEO Profiles 0	Protein 28
PubMed 0	HomoloGene 0	Protein Family Models 0
PubMed Central 4	PopSet 0	Structure 0

Genomes	Clinical	PubChem
Assembly 0	ClinicalTrials.gov 0	BioAssays 0
BioCollections 0	ClinVar 0	Compounds 0
BioProject 0	dbGaP 0	Pathways 0
BioSample 0	dbSNP 0	Substances 0
Genome 0	dbVar 0	
Nucleotide 28	GTR 0	
SRA 0	MedGen 0	

Voici à quoi cela ressemblera, vous pouvez aller à Nucléotide dans la catégorie génome et cliquer dessus

Nucleotide   [Help](#)

[Create alert](#) [Advanced](#)

Species: Animals (28) [Customize ...](#)

Molecule types: genomic DNA/RNA (28) [Customize ...](#)

Source databases: INSDC (GenBank) (28) [Customize ...](#)

Sequence Type: Nucleotide (28)

Genetic compartments: Mitochondrion (28)

Sequence length: Custom range...

Release date: Custom range...

Revision date: Custom range...

[Clear all](#)

[Show additional filters](#)

Summary  Sort by Default order  [Filters: Manage Filters](#)

See Gene information for b cytochrome **cytochrome b**

b in [Drosophila melanogaster \(2\)](#) [Escherichia phage Lambda](#) [All 50 Gene records](#)

cytochrome in [Cricetulus griseus](#) [Tripterygium wilfordii \(2\)](#) [All 4 Gene records](#)

cytochrome b in [Pongo abelii](#) [1 Gene record](#)

Items: 1 to 20 of 28

<< First < Prev Page 1 of 2 Next > Last >>

- [Eulemur rufifrons clone Erufi-NHMB89006 cytochrome b gene, partial cds; mitochondrial](#)
1. 223 bp linear DNA  
Accession: KF708347.1 GI: 556926369  
[Protein](#) [PubMed](#) [Taxonomy](#)  
[GenBank](#) [FASTA](#) [Graphics](#)
- [Eulemur rufifrons clone Erufi-NHM1882314 cytochrome b gene, partial cds; mitochondrial](#)
2. 223 bp linear DNA  
Accession: KF708346.1 GI: 556926367  
[Protein](#) [PubMed](#) [Taxonomy](#)  
[GenBank](#) [FASTA](#) [Graphics](#)
- [Eulemur rufifrons clone Erufi-MCZ16357 cytochrome b gene, partial cds; mitochondrial](#)
3. 223 bp linear DNA  
Accession: KF708345.1 GI: 556926365  
[Protein](#) [PubMed](#) [Taxonomy](#)  
[GenBank](#) [FASTA](#) [Graphics](#)
- [Eulemur rufifrons clone Erufi-MCZ16356 cytochrome b gene, partial cds; mitochondrial](#)
4. 223 bp linear DNA  
Accession: KF708344.1 GI: 556926363

[Eulemur rufifrons clone Erufi-MM-448 cytochrome b gene, complete cds; mitochondrial](#)

7. 1,140 bp linear DNA  
Accession: KF708293.1 GI: 556926260  
[Protein](#) [PubMed](#) [Taxonomy](#)  
[GenBank](#) [FASTA](#) [Graphics](#)

**Find related data**

Database:

**Search details**

("Eulemur rufifrons"[Organism] OR Eulemur rufifrons[All Fields]) AND cytochrome b[All Fields]

[See more...](#)

**Recent activity**

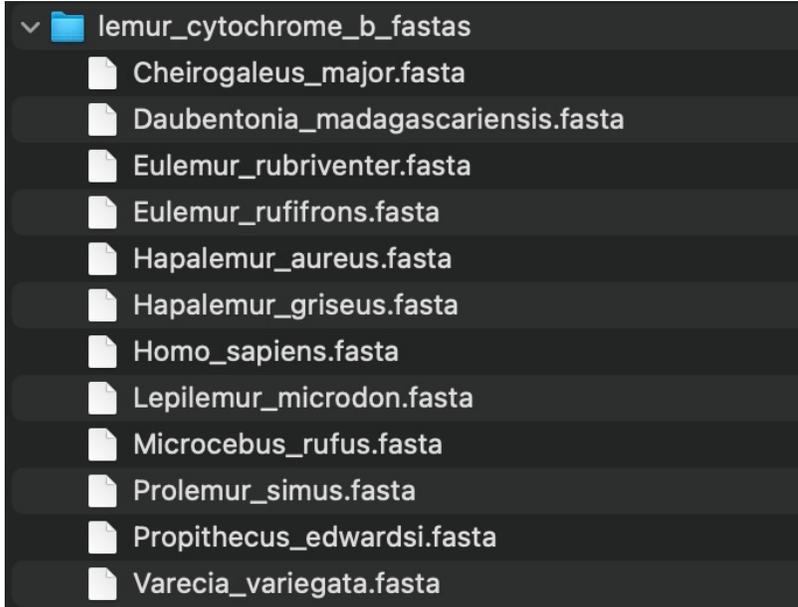
Your browsing activity is temporarily unavailable.

Choisissez la séquence de ce qui vous intéresse, dans notre cas nous voulons un cds complet

Nous pourrions vouloir des cds partiels si nous avons une séquence partielle d'intérêt, mais pour l'instant, nous construisons simplement un arbre avec des données connues, donc les cds complets sont les meilleurs

Cds : séquence codante pour la protéine

Ensuite, téléchargez le fastas



Lorsque vous avez toutes vos séquences d'intérêt et votre exogroupe, vous devez concaténer les séquences en un seul fichier, vous pouvez le faire en créant un fichier texte/édition et en collant chaque séquence, sinon suivez les instructions sur la ligne de commande (mac) ou powershell (windows) pour le faire

```
lemur_cytochrome_b_fastas — -bash — 121x27
Last login: Wed Nov 23 12:34:19 on ttys000
The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
(base) Gwenddolens-MacBook-air:~ gwenddolenkettenburg$ cd Desktop
(base) Gwenddolens-MacBook-air:Desktop gwenddolenkettenburg$ cd Intro_phylogenetic_modeling_Kettenburg
(base) Gwenddolens-MacBook-air:Intro_phylogenetic_modeling_Kettenburg gwenddolenkettenburg$ cd lemur_cytochrome_b_fastas
(base) Gwenddolens-MacBook-air:lemur_cytochrome_b_fastas gwenddolenkettenburg$ cat *.fasta>lemur_cytB_concatenated
(base) Gwenddolens-MacBook-air:lemur_cytochrome_b_fastas gwenddolenkettenburg$
```

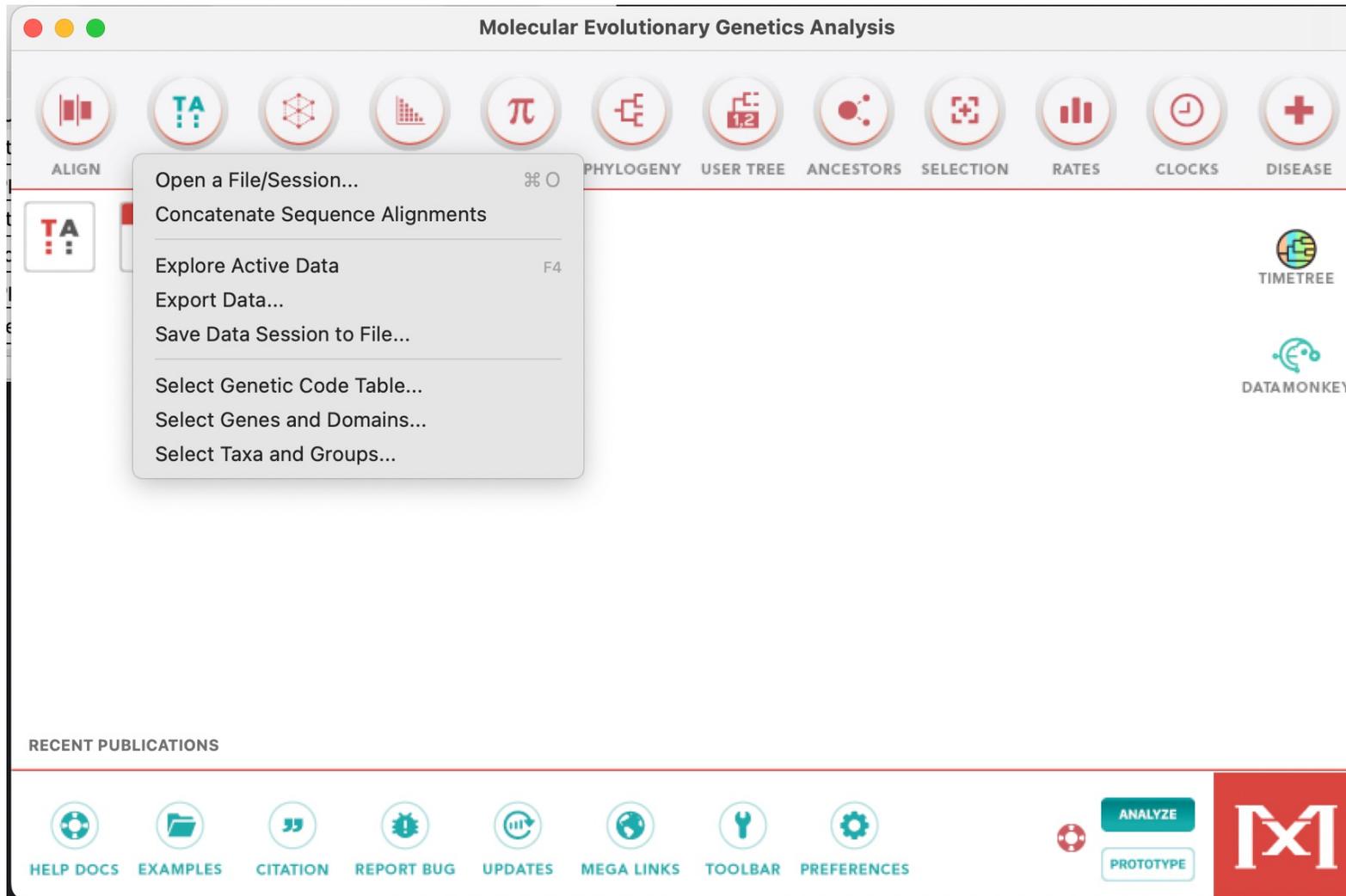
**Example 1:** Merge with file names (This will merge file1.csv & file2.csv to create concat.csv)

```
type file1.csv file2.csv > concat.csv
```

**Example 2:** Merge files with pattern (This will merge all files with csv extension and create concat.csv)

**When using asterisk(\*) to concatenate all files. Please DON'T use same extension for target file(Eg. .csv).** There should be some difference in pattern else target file will also be considered in concatenation

```
type *.csv > concat_csv.txt
```



Ouvrez MEGA, et ouvrez un fichier/session, sélectionnez votre fichier fasta concaténé MEGA vous demandera si vous souhaitez aligner ou analyser, cliquez sur aligner



M11: Alignment Explorer (ranomafana\_lemurs\_cytochromeB.fasta)

DNA Sequences Translated Protein Sequences

Species/Abbrv	* * * * *
1. Hapalemur aureus cytB	A T G A C C A A C A T C C G A A A A A A C C A C C C C T T A A T A A A A A T C A T A A A T A G C T C A T T C A T T G A C C T C C C A G C A C C A T C C A A C A T C
2. Prolemur simus cytB	A T G A C C A A C A T C C G A A A A A A T C A C C C C T T A A T A A A A A T T A T A A A C A G C T C A T T C A T T G A C C T C C C A G C A C C A T C C A A C A T C
3. Cheirogaleus major cytB	A T G A C C A A C A T T C G A A A A A C T C A C C C C T T A T A A A A A A T T A T A A A C A G T T C A T T T A T T G A T C T A C C A G C A C C A T C C A A C A T C
4. Daubentonia madagascariensis cytB	A T G A C C A A C A T T C G A A A A A C C C A C C C T
5. Eulemur rubriventer cytB	A T G A A C A A C A T C C G A A A G A A T C A C C C C
6. Eulemur rufifrons cytB	A T G A A C A A C A T C C G A A A G A A T C A C C C C
7. Hapalemur griseus meridionalis cytB	A T G A C C A A C A T C C G A A A A A A C C A C C C C
8. Homo sapiens cytB	A T G A C C C C A A T A C G C A A A A C T A A C C C C
9. Lepilemur microdon cytB	A T G A C C A A C A T T C G A A A A A A C C A C C C C
10. Microcebus rufus cytB	A T G A C C A A C A T C C G A A A A A C C C A C C C A
11. Propithecus edwardsi cytB	A T G A C C A A C A T C C G A A A A A A C C A C C C C
12. Varecia variegata cytB	A T G A C T A A C A C C G A A A A A A C C A C C C C

**MUSCLE Alignment Options**

Option	Setting
<b>GAP PENALTIES</b>	
Gap Open	<input checked="" type="checkbox"/> -400.00
Gap Extend	<input checked="" type="checkbox"/> 0.00
<b>MEMORY/ITERATIONS</b>	
Max Memory in MB	<input checked="" type="checkbox"/> 2048
Max Iterations	<input checked="" type="checkbox"/> 16
<b>ADVANCED OPTIONS</b>	
Cluster Method (Iterations 1,2)	<input checked="" type="checkbox"/> UPGMA
Cluster Method (Other Iterations)	<input checked="" type="checkbox"/> UPGMA
Min Diag Length (Lambda)	<input checked="" type="checkbox"/> 24

Site # 1  with  w/o gaps Selected genetic code: Standard

Suivez les options suggérées, puis appuyez sur OK



The screenshot displays the MEGA11 software interface. The main window shows a multiple sequence alignment of cytochrome B (cytB) genes. The alignment is titled "M11: Alignment Explorer (ranomafana\_lemurs\_cytochromeB\_aligned.mas)". The alignment is displayed in a grid format with columns representing positions and rows representing individual sequences. The sequences are color-coded by nucleotide: Adenine (A) in red, Thymine (T) in blue, Cytosine (C) in green, and Guanine (G) in purple. The alignment is shown in the "DNA Sequences" tab. The "Export Alignment" menu is open, and the "MEGA Format" option is selected. Other options in the menu include "FASTA Format" and "NEXUS/PAUP Format". The background shows a macOS desktop with a file explorer window open, displaying a folder named "intro\_phylogenetic\_modeling\_Kettenburg" and a file named "proposal.doc".

Enregistrez le fichier aligné, puis nous procéderons à la sélection du modèle

# Ouvrir l'interface graphique RAxML

**INPUT**

LOAD ALIGNMENT

**ANALYSIS**

ML + transfer bootstrap expectation + consensus ▾ 1 ▾ 100 ▾  
Analysis Runs Reprs.

628076  
Seed

**OUTPUT**

Select output directory

output

Select output name

**RAxML**

raxml-ng-ARM64 ▾ RUN

Binary

Command  
raxml-ng-ARM64 --all --msa RAxML\_output\_concat.txt --model --prefix output seed 628076 --bs-metric tbe --tree rand{1} --bs-trees 100

**CONSOLE**

raxmlGUI 2.0.10 raxml-ng-ARM64 1.1.0

How to cite? [For questions or suggestions contact us!](#)

nucleotide

ranomafana\_lemurs\_cytochromeB\_align.fas



12 sequences of length 1141

GTR



none



none



Substitution model

Stationary frequencies

Proportion of invariant sites

none



Rate heterogeneity

RUN MODELTEST

Partition 1/1:

	Model	Score	Weight
BIC	TIM2+I+G4	13908.7732	0.9990
AIC	TIM2+I+G4	13762.6230	0.8741
AICc	TIM2+I+G4	13763.6230	0.8741

Chargez le fichier aligné et effectuez modeltest, pour la sélection du modèle, utilisez le score BIC, il crachera un rapport qui s'enregistrera dans vos fichiers

Input: nucleotide ranomafana\_lemurs\_cytochromeB\_align.fas (12 sequences of length 1141)

Substitution model: TIM2  
 Stationary frequencies: none  
 Proportion of invariant sites: +I (ML estimate)

+GAMMA (mean)  
 Rate heterogeneity

RUN MODELTEST

Analysis: ML + transfer bootstrap expectation + consensus (1 / 100)

- <none>
- Hapalemur\_aureus\_cytB
- Prolemur\_simus\_cytB
- Cheirogaleus\_major\_cytB
- Daubentonia\_madagascariensis\_cytB
- Eulemur\_rubrивenter\_cytB
- Eulemur\_rufifrons\_cytB
- Hapalemur\_griseus\_meridionalis\_cytB
- Homo\_sapiens\_cytB
- Lepilemur\_microdon\_cytB
- Microcebus\_rufus\_cytB
- Propithecus\_edwardsi\_cytB
- Varecia\_variegata\_cytB

Output: RAxML

cytochromeB\_align'  
 \_cytochromeB\_align.ckp  
 \_cytochromeB\_align.log  
 \_cytochromeB\_align.out  
 \_cytochromeB\_align.tree

RAXML: raxml-ng-ARM64 (Binary) RUN

```

raxml-ng-ARM64 --all --msa
/Users/gwenddolenkettenburg/Desktop/RAxML/ranomafana_lemurs_cytochr
--model TIM2+I+G --prefix
/Users/gwenddolenkettenburg/Desktop/RAxML/ranomafana_lemurs_cytochr
seed 177748 --bs-metric tbe --tree rand{1} --bs-trees 100
Command
P.Inv: 0.4945
Alpha: 0.2560
Alpha-P.Inv: 0.7313
P.Inv-Alpha: 0.3876
Frequencies: 0.3075 0.3419 0.1142 0.2364
Commands:
> phylml -i /Users/gwenddolenkettenburg/Desktop/RAxML/ranomafana_lemurs_cyt
as -m 010232 -f m -v e -a e -c 4 -o tlr
> raxmlHPC-SSE3 -s /Users/gwenddolenkettenburg/Desktop/RAxML/ranomafana_lem
align.fas -m GTRGAMMAIX -n EXEC_NAME -p PARSIMONY_SEED
> raxml-ng --msa /Users/gwenddolenkettenburg/Desktop/RAxML/ranomafana_lemurs
ign.fas --model TIM2+I+G4
> paup -s /Users/gwenddolenkettenburg/Desktop/RAxML/ranomafana_lemurs_cytoch
> iqtree -s /Users/gwenddolenkettenburg/Desktop/RAxML/ranomafana_lemurs_cyt
as -m TIM2+I+G4
Summary:
Partition 1/1:
Model Score Weight
-----
BIC TIM2+I+G4 13908.7732 0.9990
AIC TIM2+I+G4 13762.6230 0.8741
AICc TIM2+I+G4 13763.6230 0.8741
Execution results written to /Users/gwenddolenkettenburg/Desktop/RAxML/RAxML_C
nomafana_lemurs_cytochromeB_align.out
Starting tree written to /Users/gwenddolenkettenburg/Desktop/RAxML/RAxML_GUI_M
fana_lemurs_cytochromeB_align.tree

```

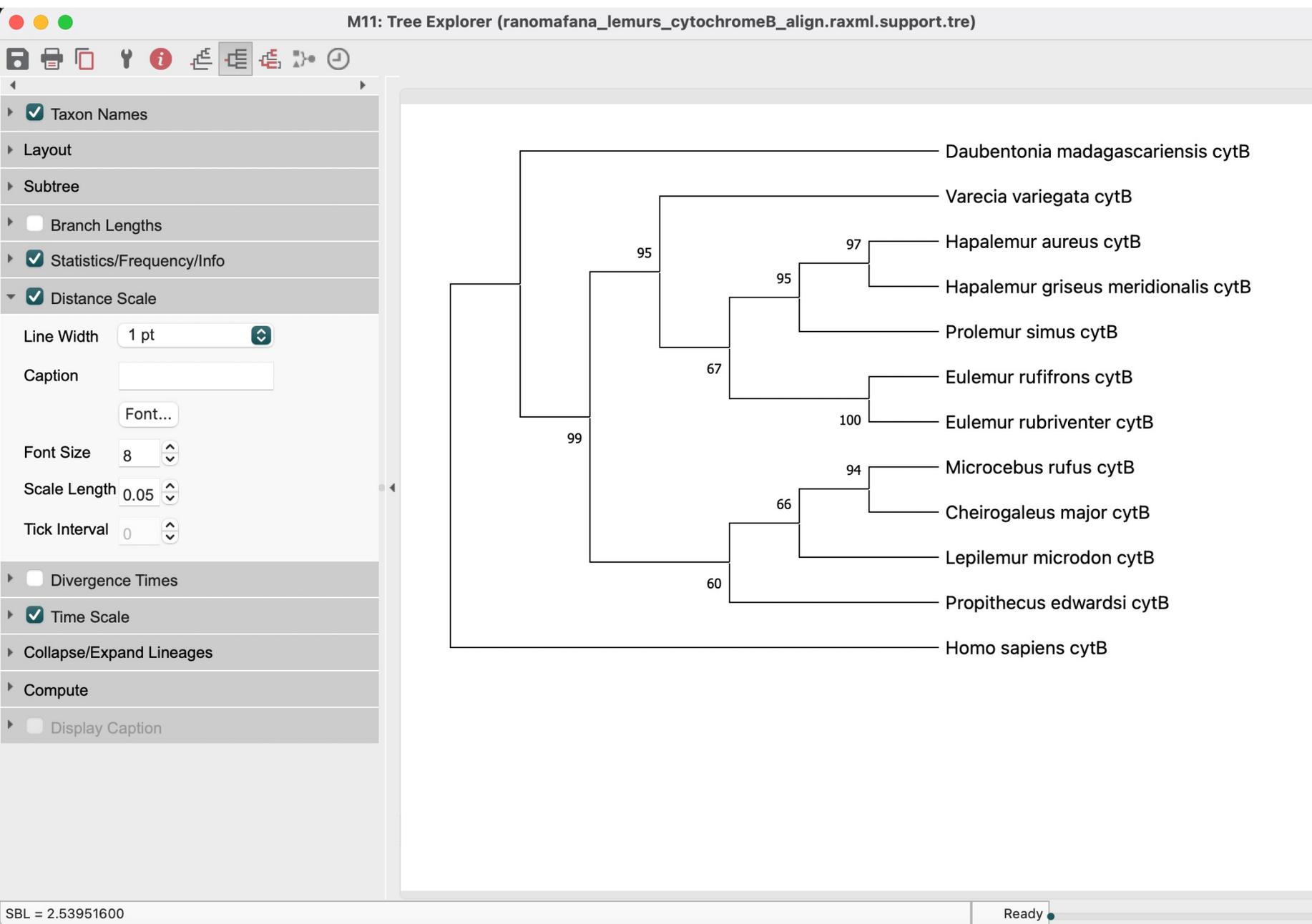
À l'aide du modèle de modeltest, modifiez selon vos besoins dans la zone de saisie, puis dans l'analyse, définissez l'outgroup, puis appuyez sur exécuter dans la section RAxML

Il va aussi recrachter quelques fichiers...

>	modeltest	Today, 1:24 PM	--	Fo
	ranomafana_lemurs_cytochromeB_align.fas	Today, 1:22 PM	14 KB	De
∨	raxml	Today, 1:26 PM	--	Fo
	ranomafana_lemurs_cytochromeB_align.raxml.bestModel.txt	Today, 1:26 PM	90 bytes	PI
	ranomafana_lemurs_cytochromeB_align.raxml.bestTree.tre	Today, 1:26 PM	518 bytes	Fi
	ranomafana_lemurs_cytochromeB_align.raxml.bootstraps.tre	Today, 1:26 PM	52 KB	Fi
	ranomafana_lemurs_cytochromeB_align.raxml.log.txt	Today, 1:26 PM	11 KB	PI
	ranomafana_lemurs_cytochromeB_align.raxml.rba	Today, 1:26 PM	8 KB	De
	ranomafana_lemurs_cytochromeB_align.raxml.startTree.tre	Today, 1:26 PM	507 bytes	Fi
	ranomafana_lemurs_cytochromeB_align.raxml.support.tre	Today, 1:26 PM	590 bytes	Fi
	RAXML_GUI_Settings_ranomafana_lemurs_cytochromeB_align.txt	Today, 1:26 PM	566 bytes	PI

Nous sommes intéressés par le fichier .support.tre

Cela inclura la topologie de l'arborescence et les valeurs de prise en charge des amorçages, ouvertes dans MEGA



Nous voulons un fichier de l'arbre enregistré qui peut être lu dans R, donc newick. Cliquez sur Exporter les arbres... choisissez le format newick et personnalisez-le dans R

Vous pouvez également personnaliser dans MEGA... c'est juste plus limité

Nous aurions pu faire du modeltest et du RAxML dans MEGA aussi... mais prend une éternité !

M11: Tree Explorer (lemur\_tree\_session\_MEGA\_example.mtsx)

Original Tree  Bootstrap Tree

74 Hapalemur aureus cytB  
100 Hapalemur griseus meridionalis cytB  
81 Prolemur simus cytB  
69 Eulemur rubriventer cytB  
100 Eulemur rufifrons cytB  
71 Varecia variegata cytB  
47 Propithecus edwardsi cytB  
100 Cheirogaleus major cytB  
78 Microcebus rufus cytB  
Lepilemur microdon cytB  
Daubentonia madagascariensis cytB  
Homo sapiens cytB

0.10

**Evolutionary analysis by Maximum Likelihood method**  
The evolutionary history was inferred by using the Maximum Likelihood method and General Time Reversible model [1]. The tree with the highest log likelihood (-6723.16) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.4717)). The rate variation model allowed for some sites to be evolutionarily invariable (+I, 26.03% sites). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. This analysis involved 12 nucleotide sequences. There were a total of 1141

LogL = -6723.16 Ready

# Alors faites joli en R !

- Suivez les instructions de `lemur_tree_editing`. Fichier R