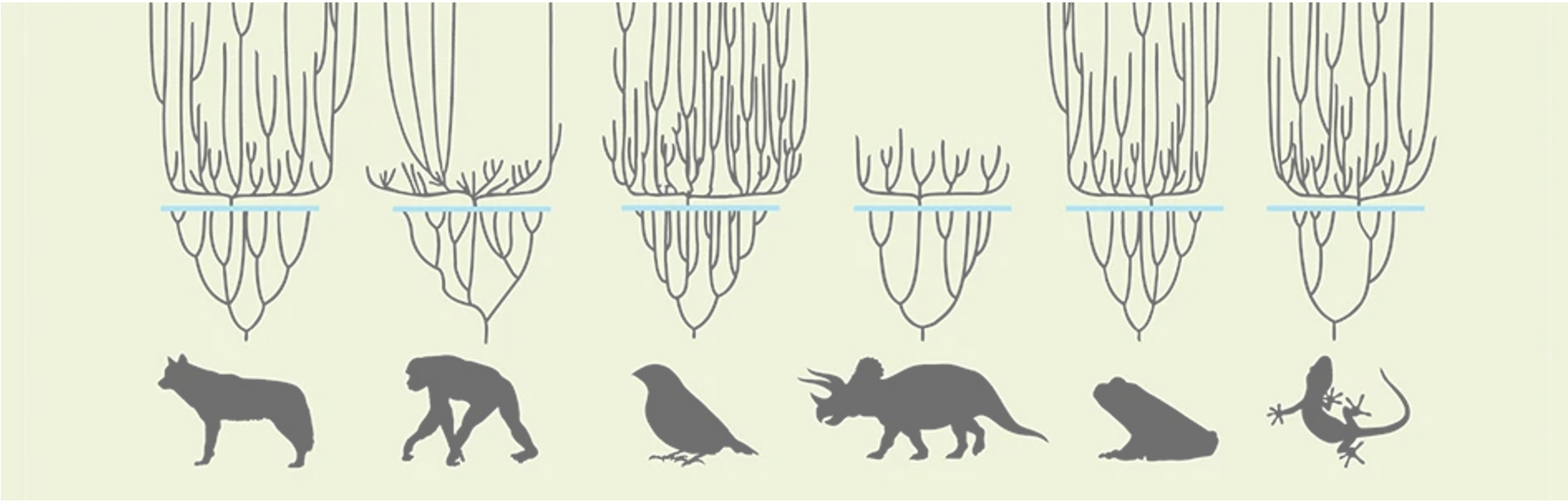
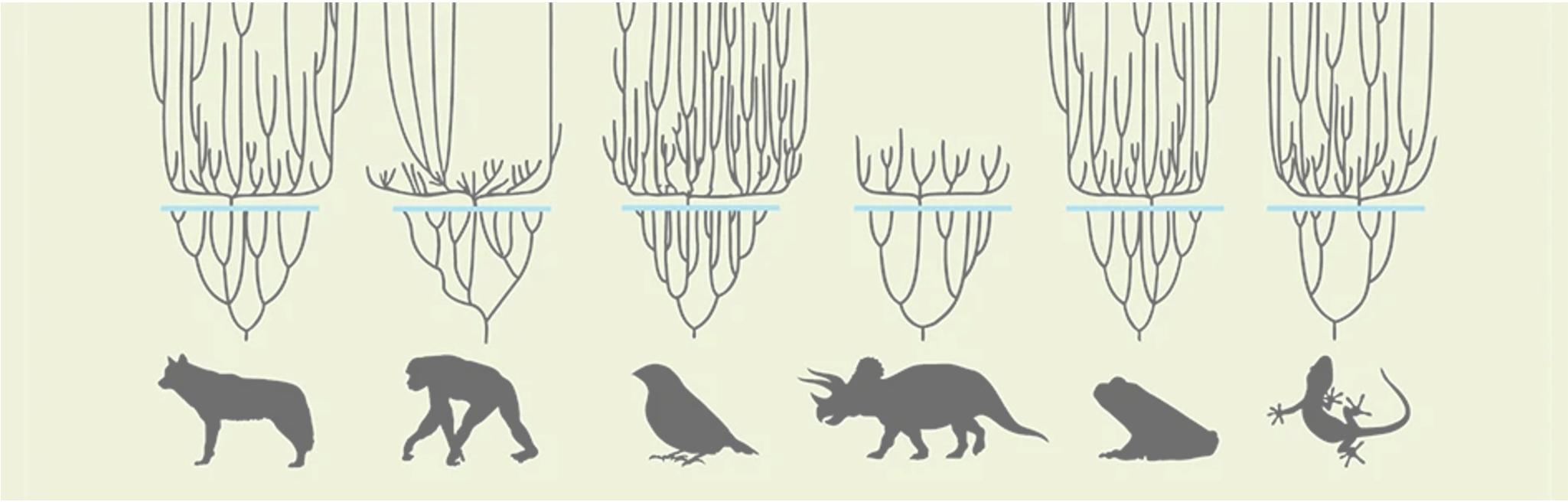
Introduction to phylogenetics

Cara Brook
Gwen Kettenburg
Slides adapted from Richard Ree and Andrew Hipp, University of Chicago

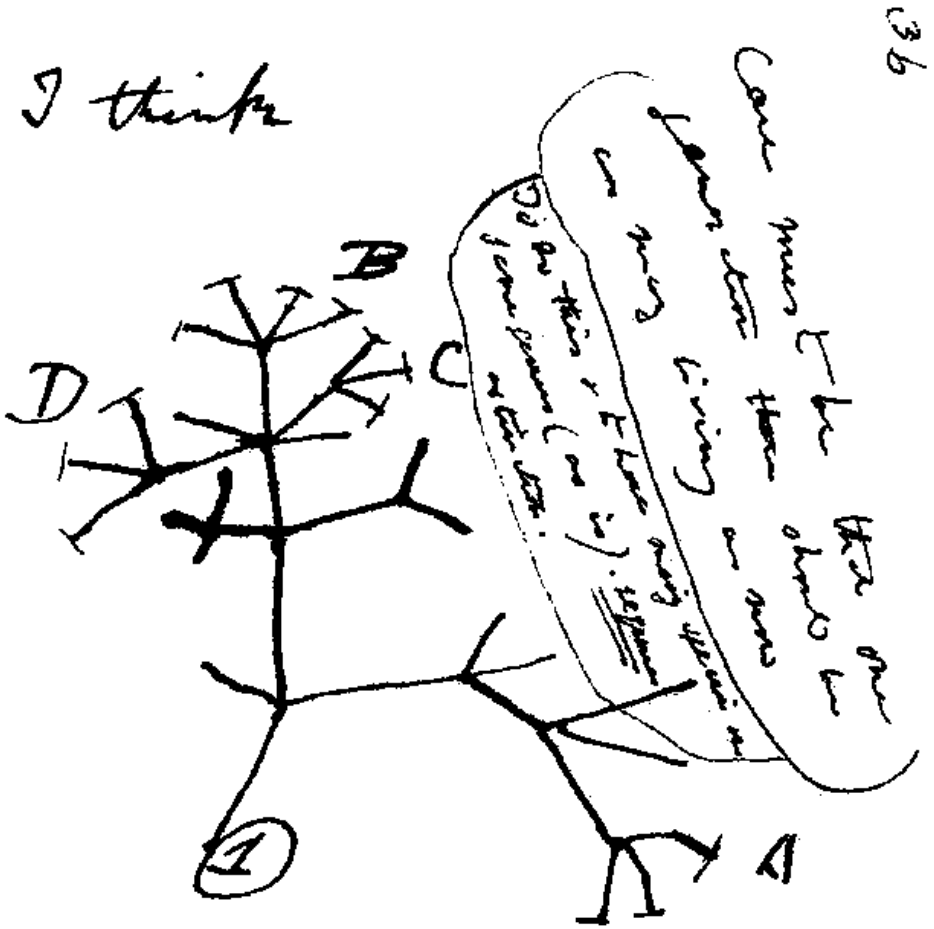


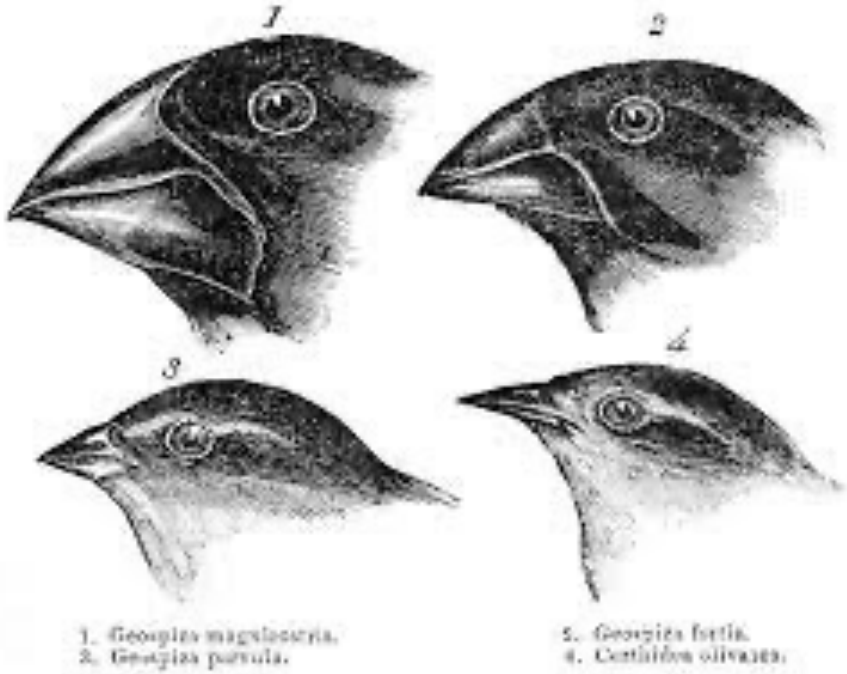
Introduction à la phylogénétique

Cara Brook
Gwen Kettenburg
Diapositives adaptées de Richard Ree et Andrew Hipp, Université de Chicago

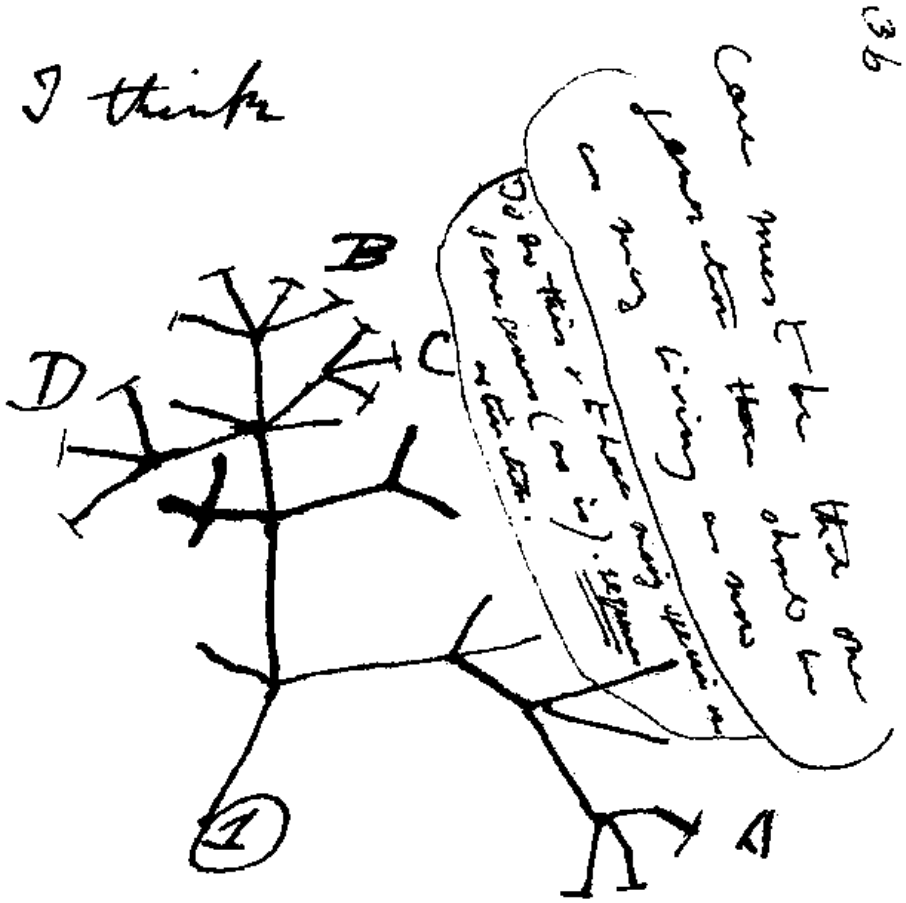


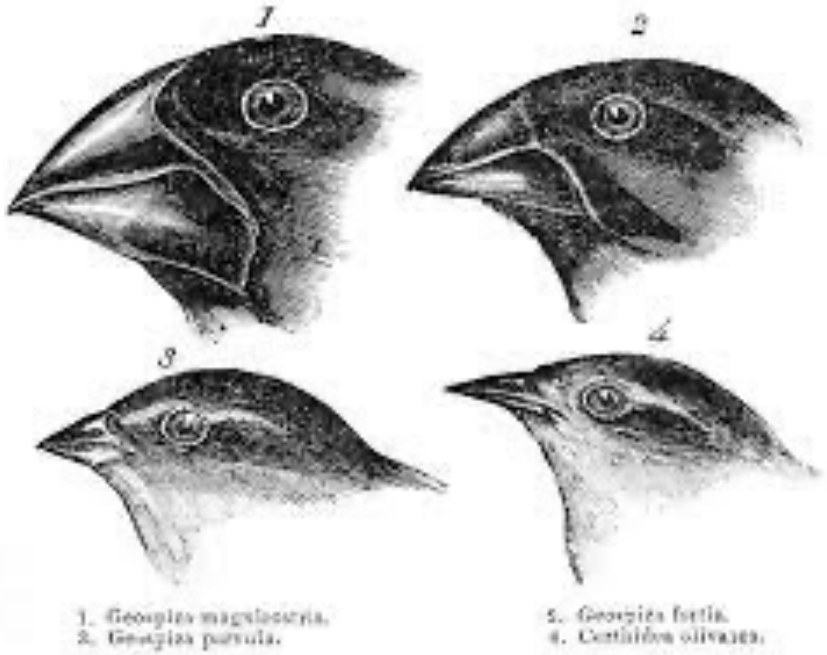
What is a phylogeny?



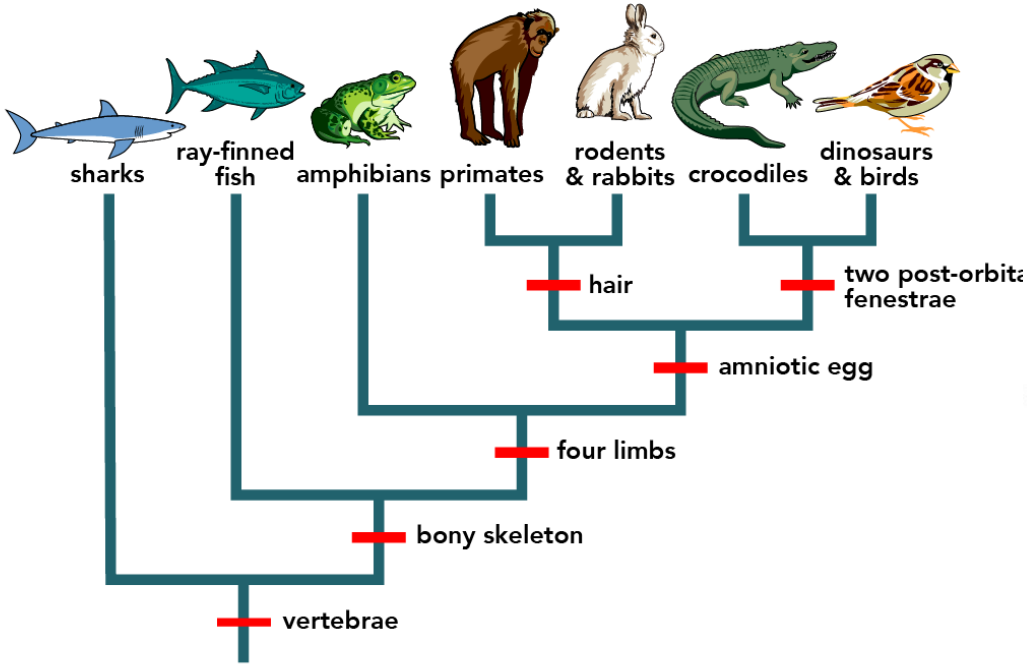


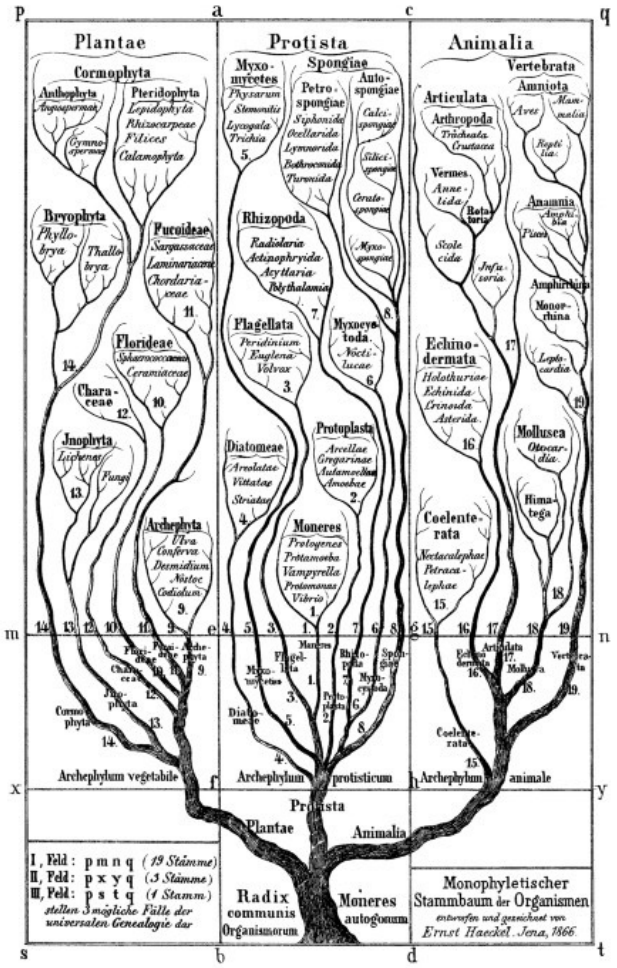
Qu’est-ce qu’une phylogénie ?





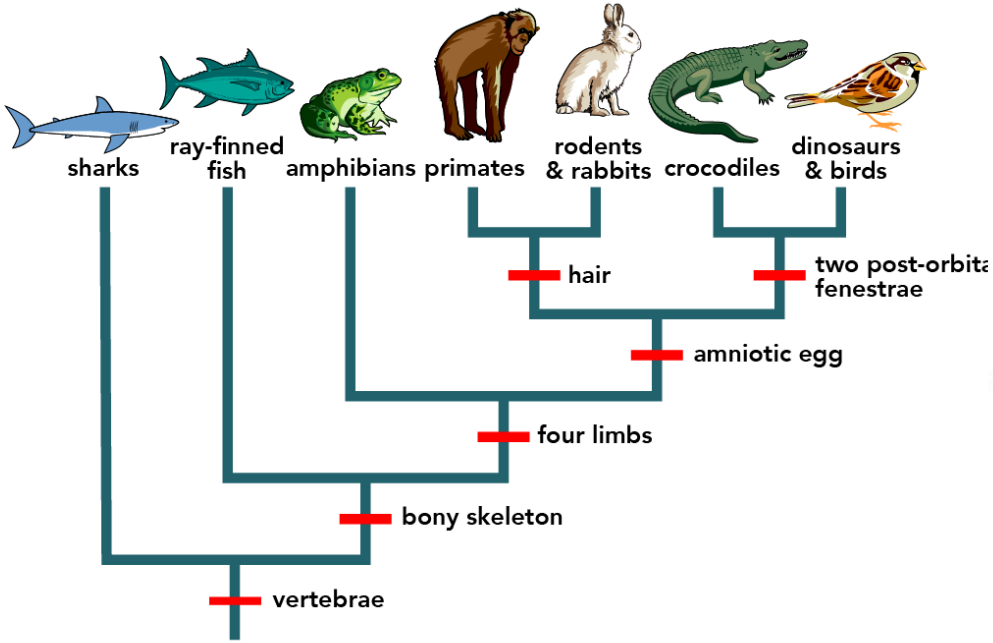
What is a phylogeny?

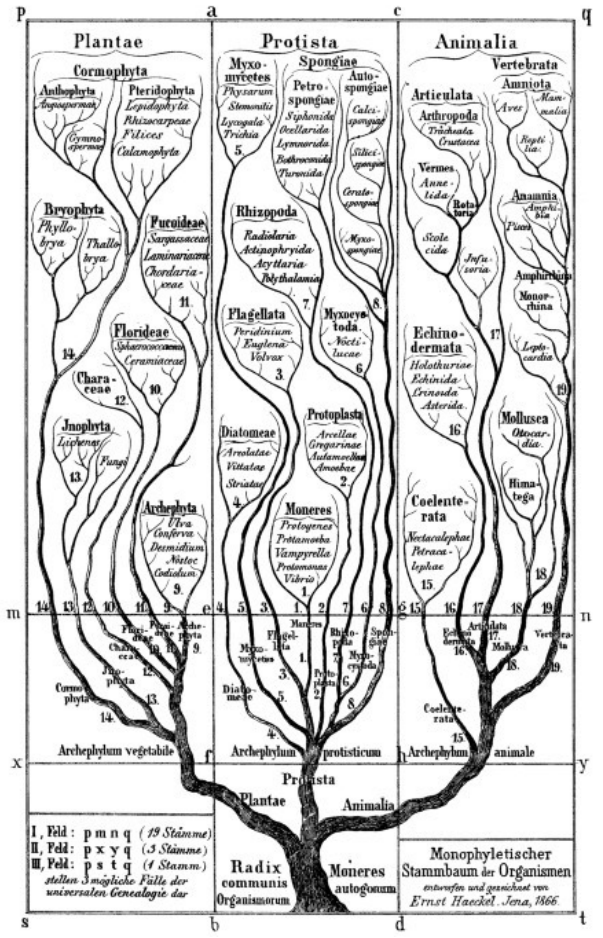




Baum et. al, Nature, 2008
Hossfeld and Levit, Nature, 2016

Qu’est-ce qu’une phylogénie ?

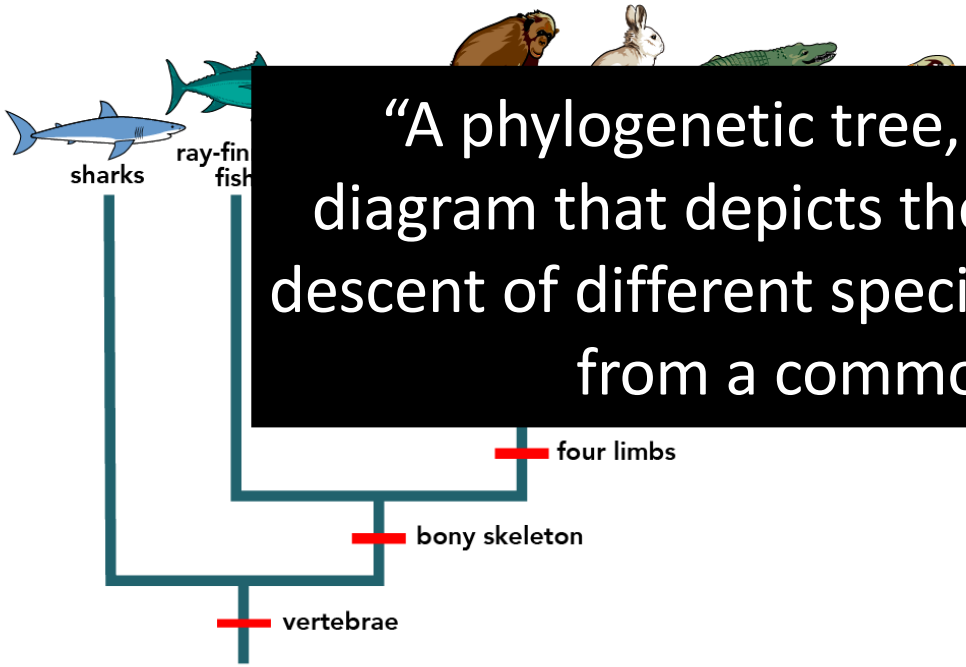




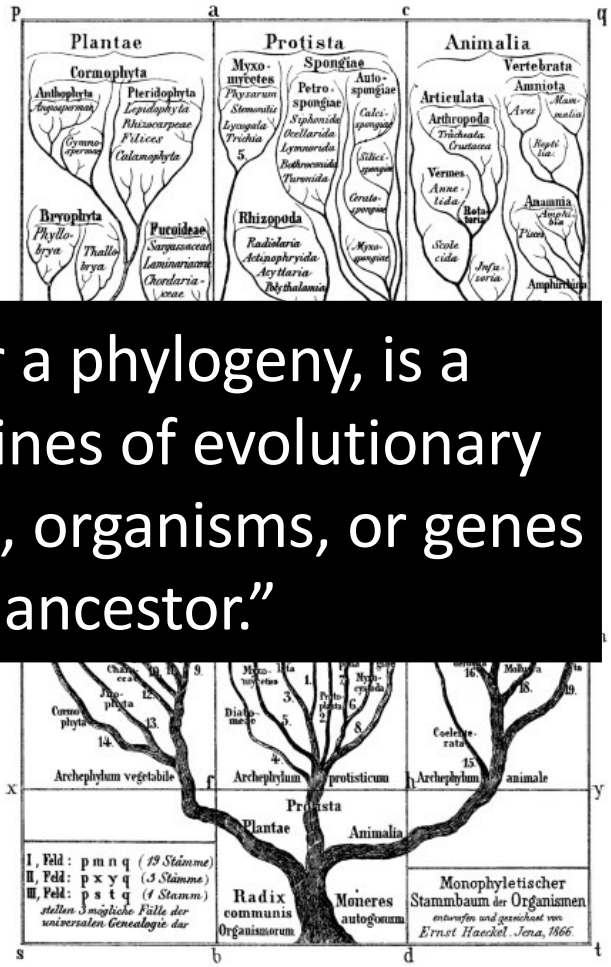
Baum et. al, Nature, 2008

Hossfeld and Levit, Nature, 2016

What is a phylogeny?



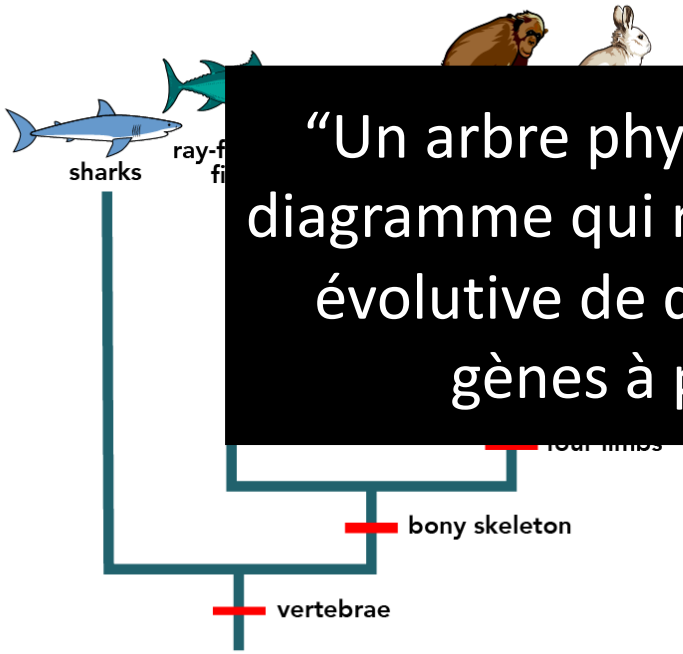
"A phylogenetic tree, or a phylogeny, is a diagram that depicts the lines of evolutionary descent of different species, organisms, or genes from a common ancestor."



Baum et. al, Nature, 2008

Hossfeld and Levit, Nature, 2016

Qu’est-ce qu’une phylogénie ?

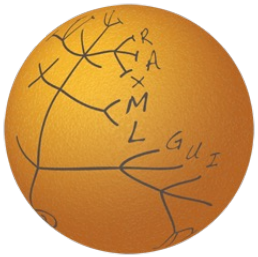


“Un arbre phylogénétique, ou phylogénie, est un diagramme qui représente les lignes de descendance évolutive de différentes espèces, organismes ou gènes à partir d’un ancêtre commun.”

Baum et. al, Nature, 2008

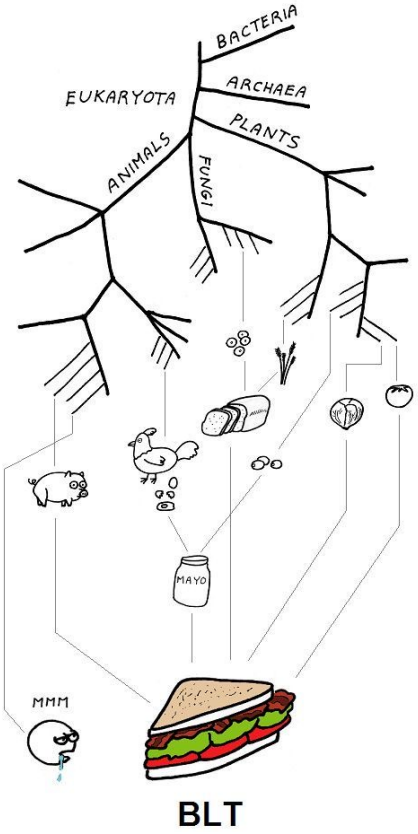
Hossfeld and Levit, Nature, 2016

Goals:

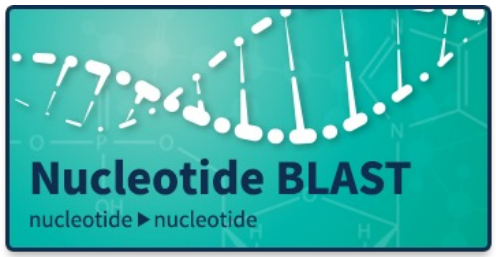


- Lecture component
  - Learn basics of what a phylogeny is
  - Learn how to read phylogenies
  - Basics of phylogenetic modeling
- Tutorial component
  - Learn how to make a phylogenetic tree from sequencing data
  - Using influenza protein sequences in MEGA software
  - Edit and visualize tree in R and FigTree

RStudio®



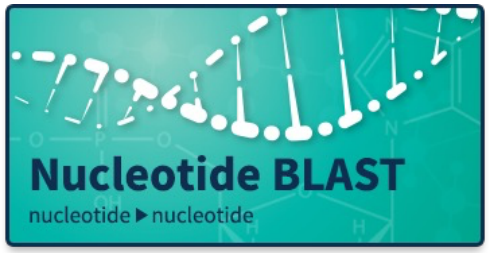
BLT

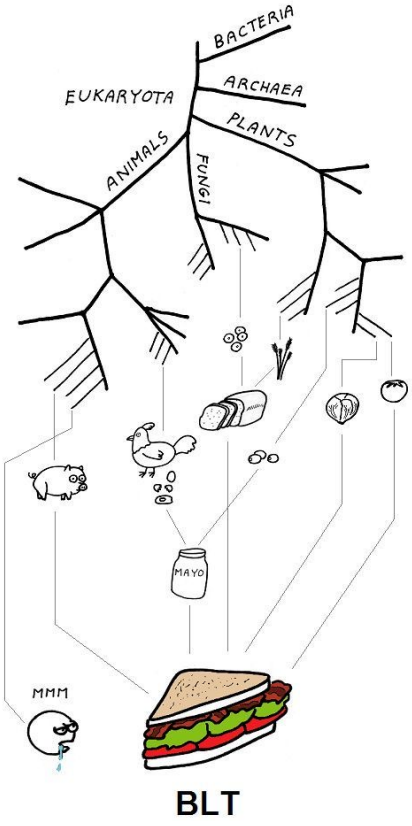


MEGA Molecular Evolutionary Genetics Analysis

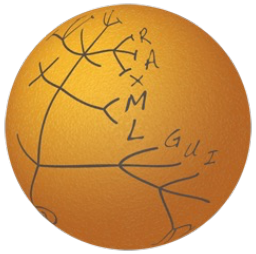
Objectifs:

- 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



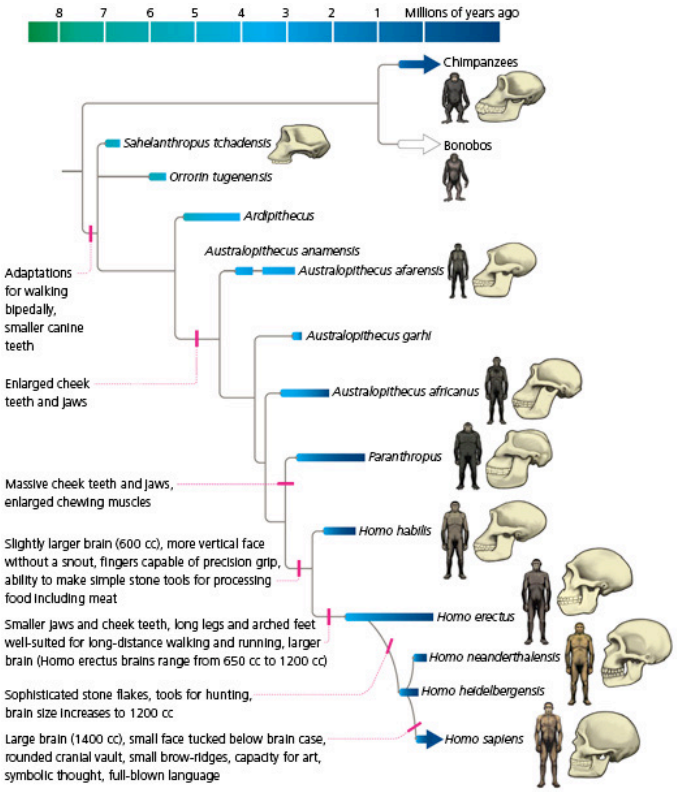


R Studio®

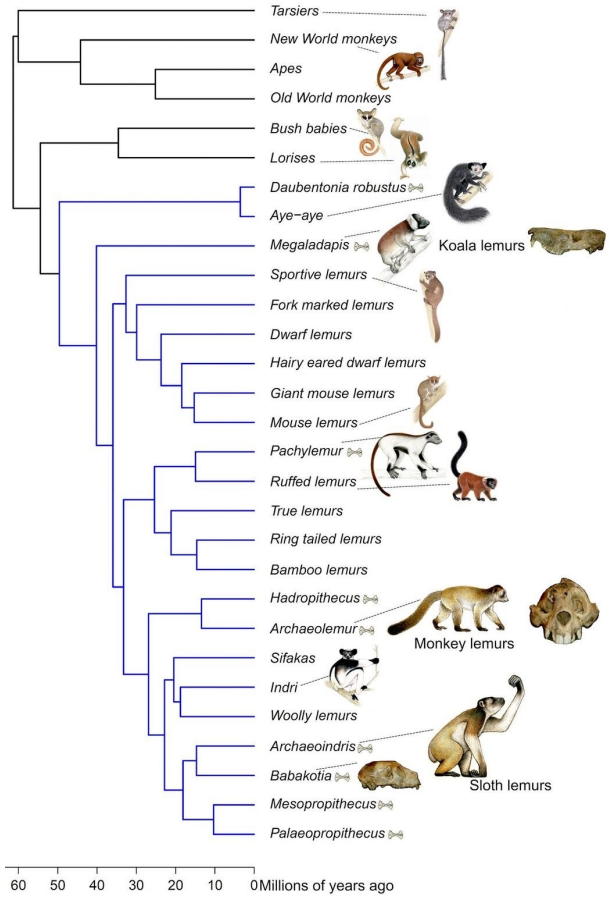


MEGA Molecular Evolutionary Genetics Analysis

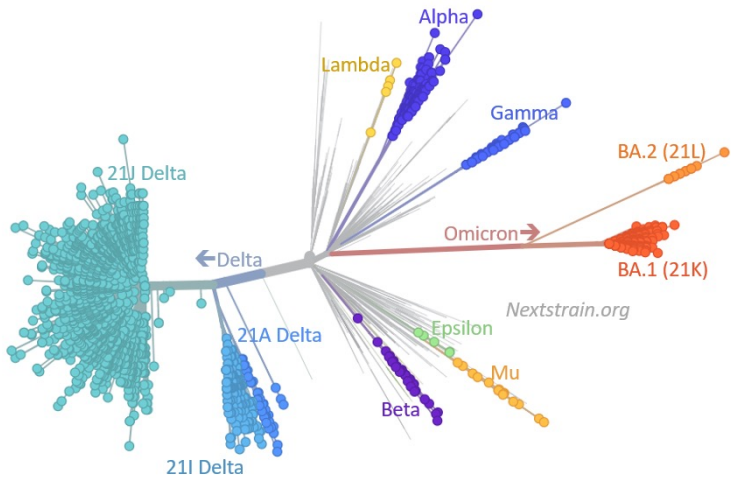
What can you do with phylogenies?



UC Museum of Paleontology

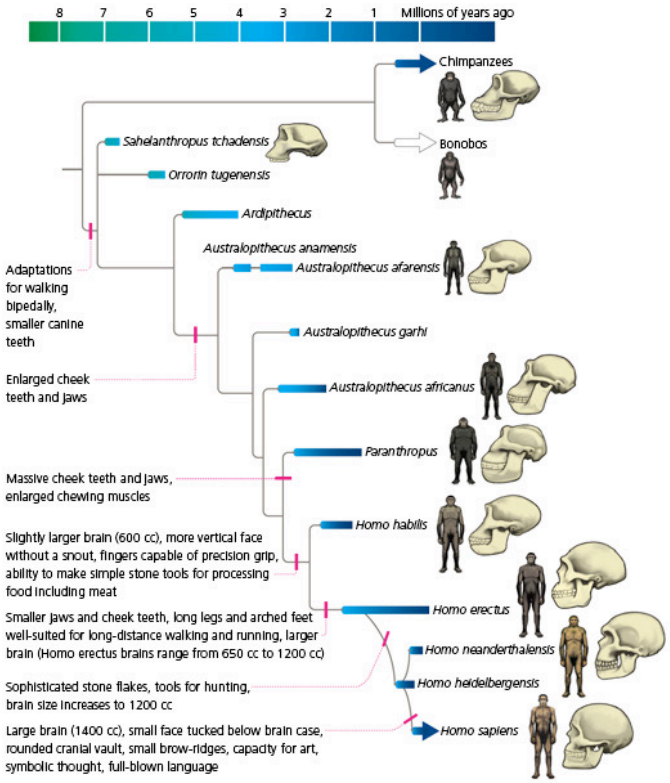


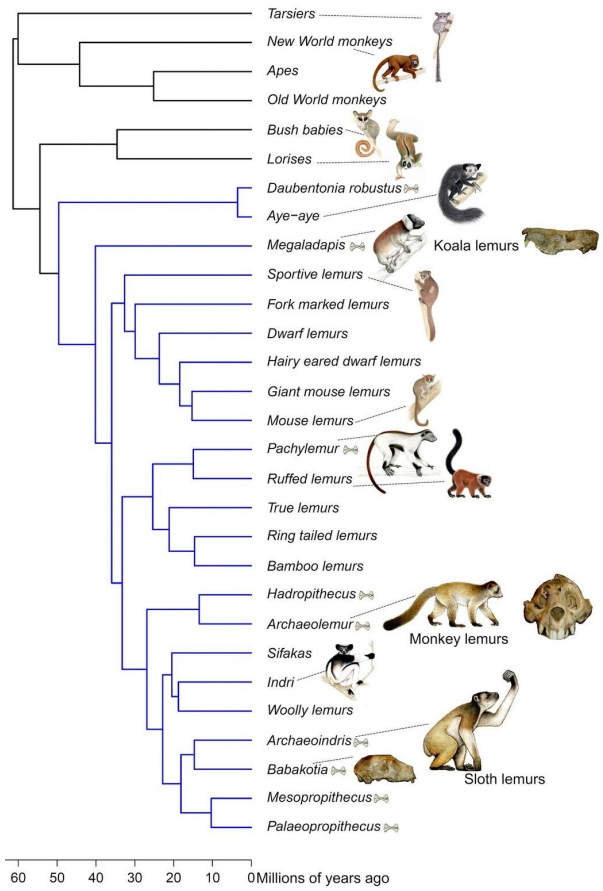
Herrera et al 2016

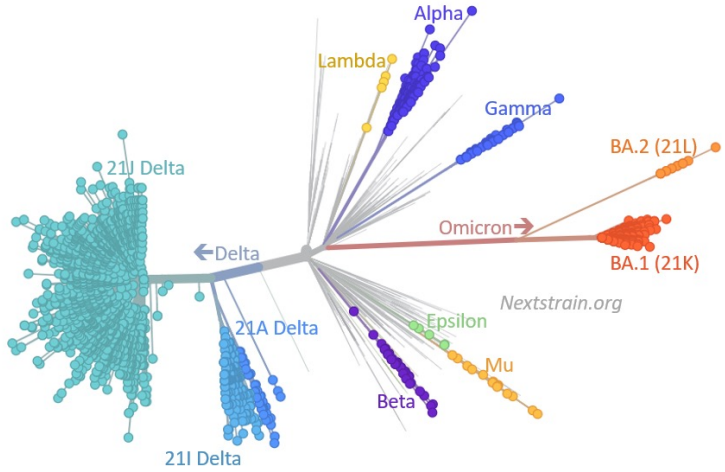


Lowe 2022

Que pouvez-vous faire avec les phylogénies ?





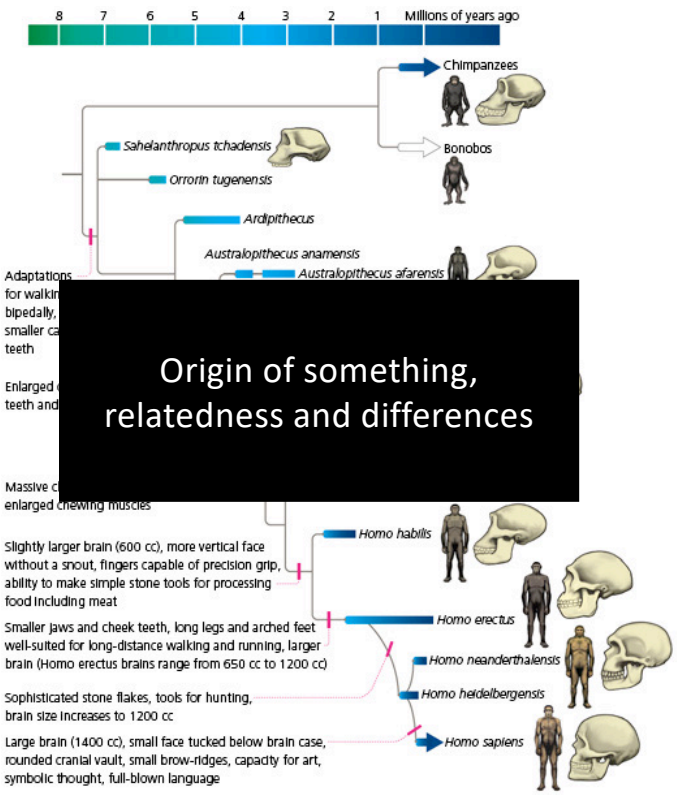


UC Museum of Paleontology

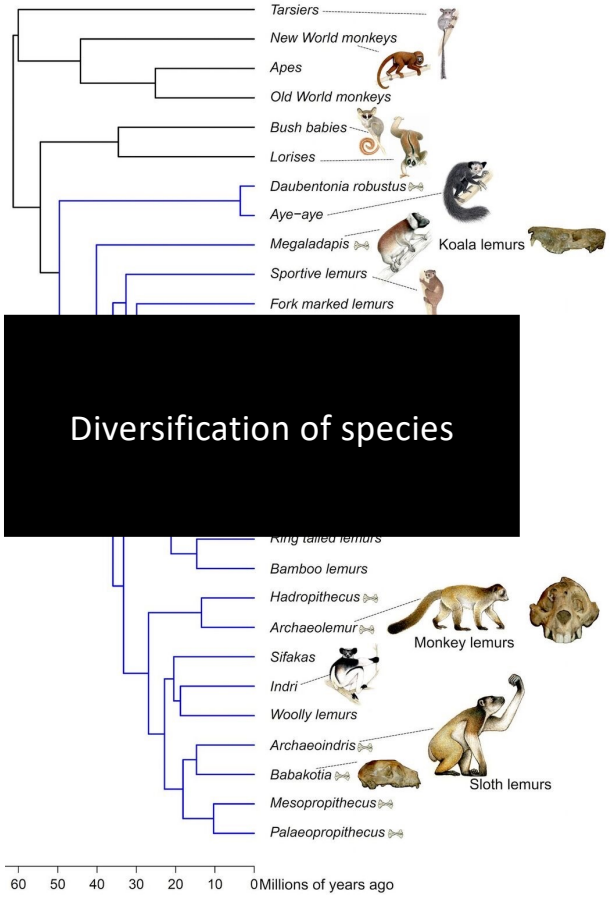
Herrera et al 2016

Lowe 2022

What can you do with phylogenies?



Origin of something, relatedness and differences



Diversification of species

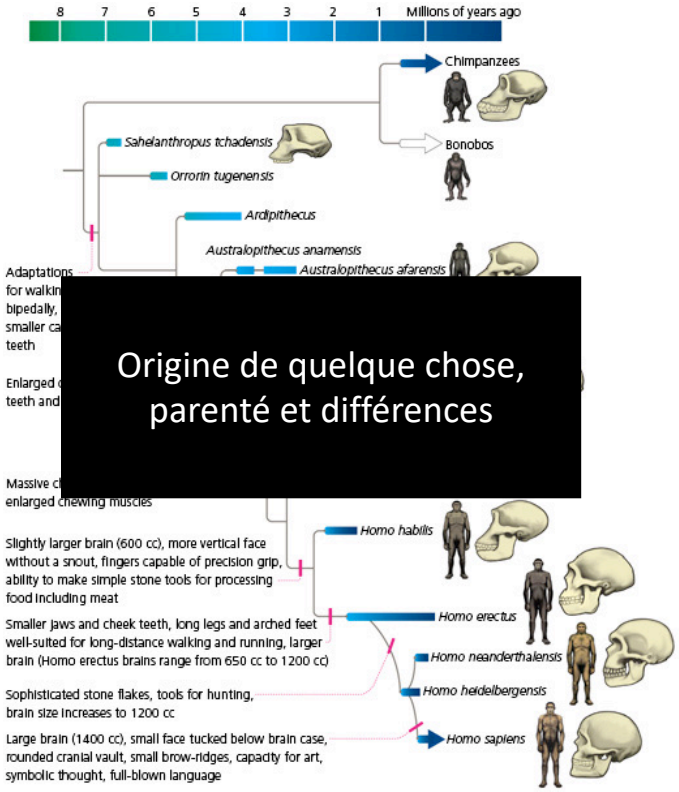
Changes in real time

UC Museum of Paleontology

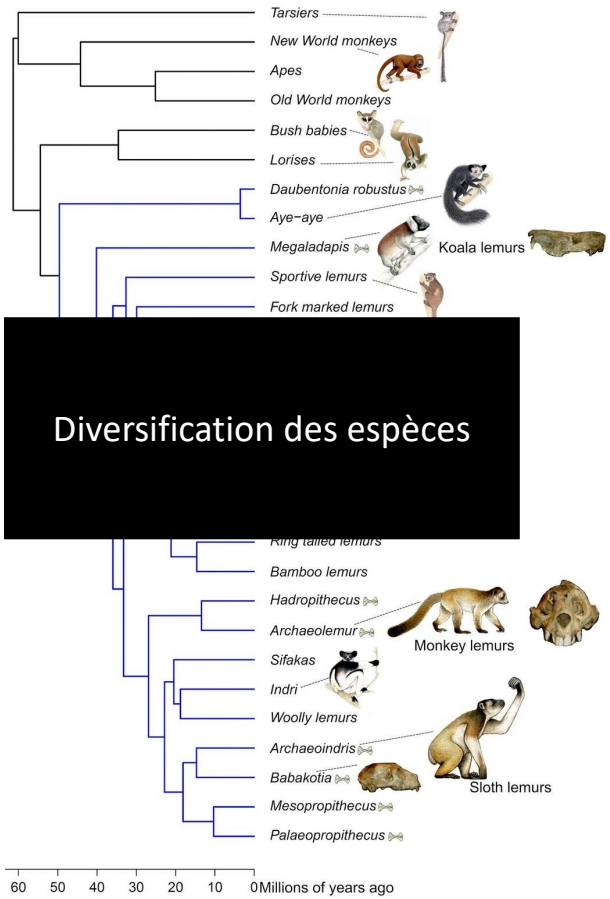
Herrera et al 2016

Lowe 2022

Que pouvez-vous faire avec les phylogénies ?



Origine de quelque chose, parenté et différences



Diversification des espèces

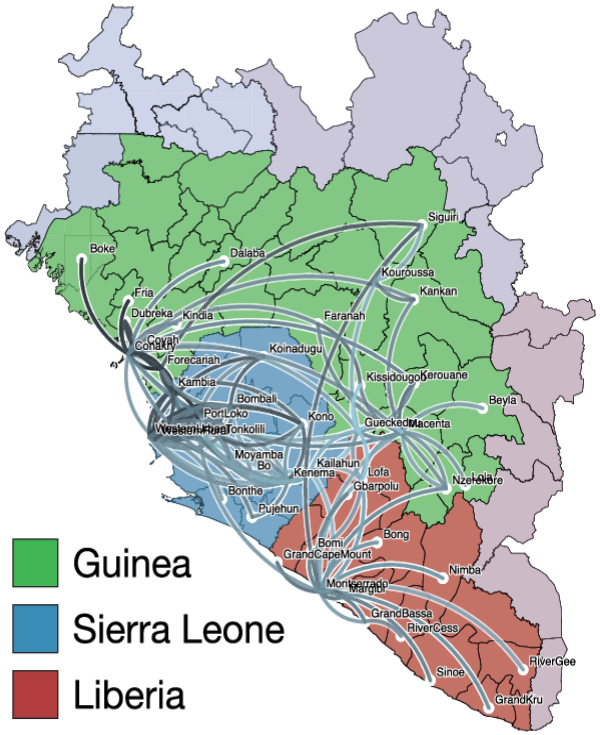
Changements en temps réel

UC Museum of Paleontology

Herrera et al 2016

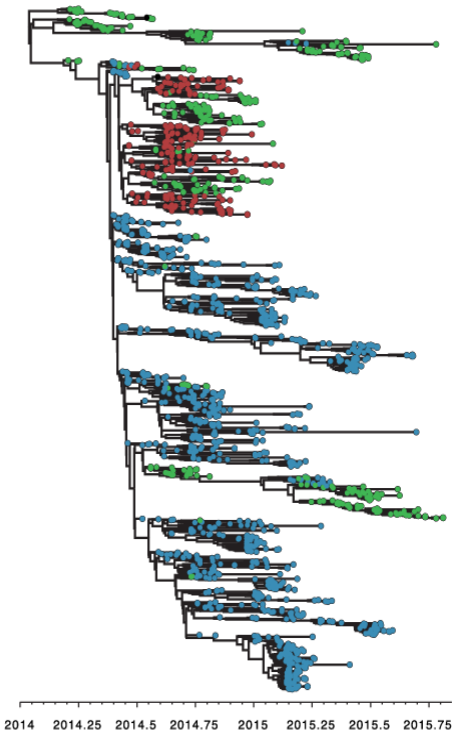
Lowe 2022

Phylodynamics



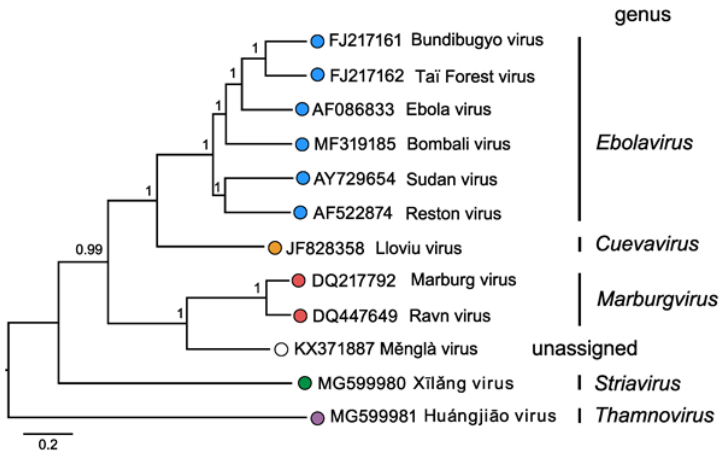
WHERE does it go?

Bayesian trees



WHEN is the most recent common ancestor?

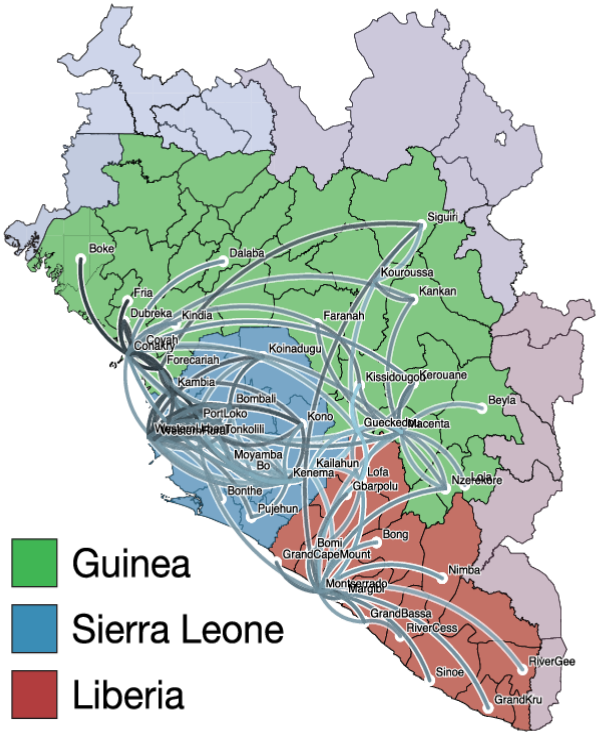
Maximum likelihood



HOW different is it to what's known?

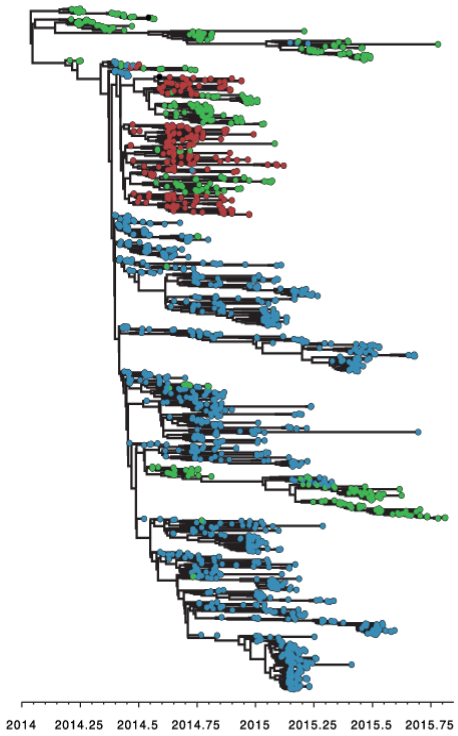
Suchard et al. 2018
Kuhn et al. 2019

Phylodynamique



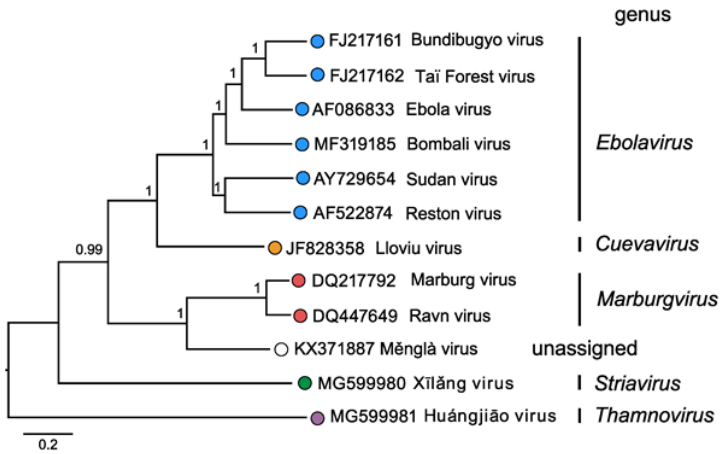
OÙ va-t-il ?

Arbres bayésiens



QUAND est l'ancêtre commun le plus récent ?

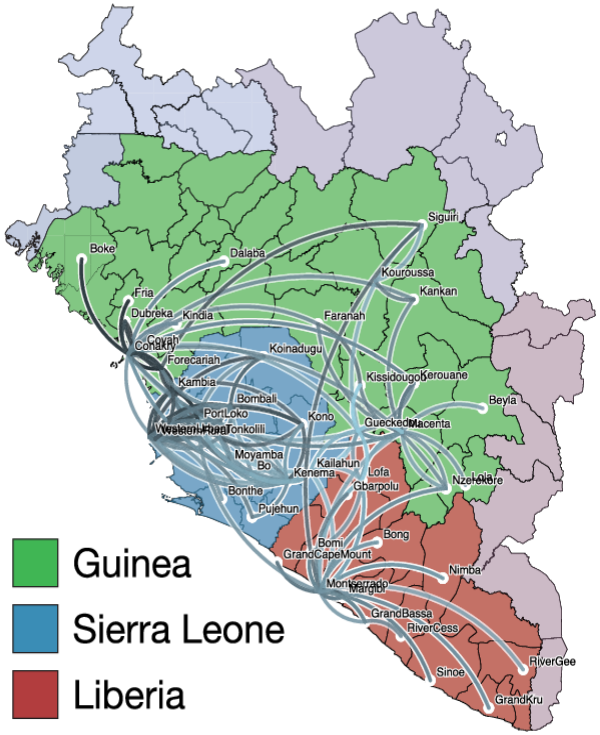
Maximum de vraisemblance



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

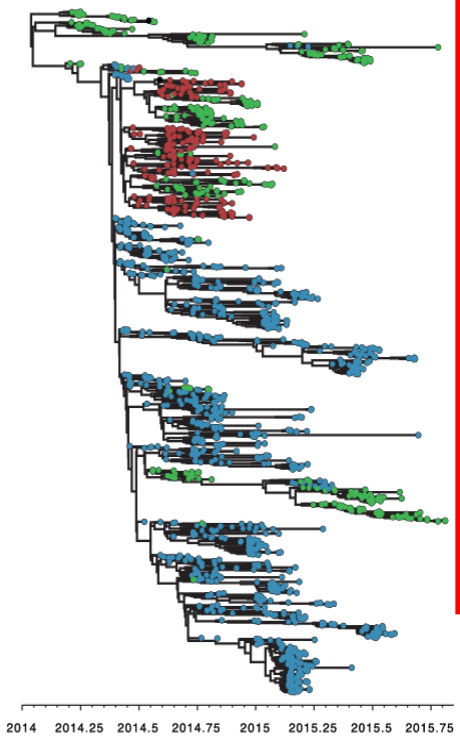
Suchard et al. 2018
Kuhn et al. 2019

Phylodynamics



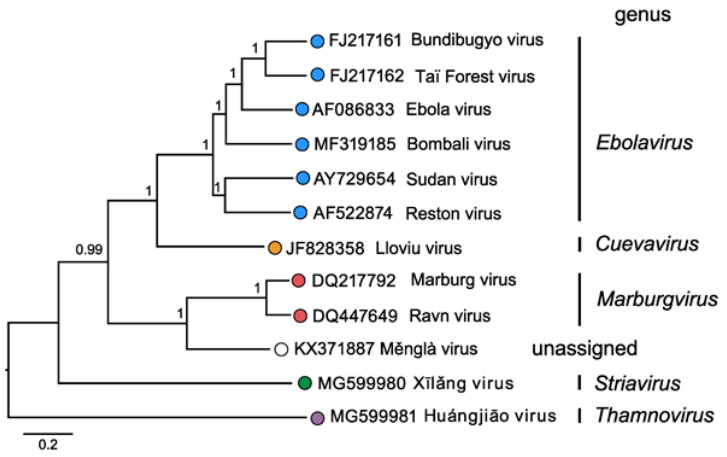
WHERE does it go?

Bayesian trees



WHEN is the most recent common ancestor?

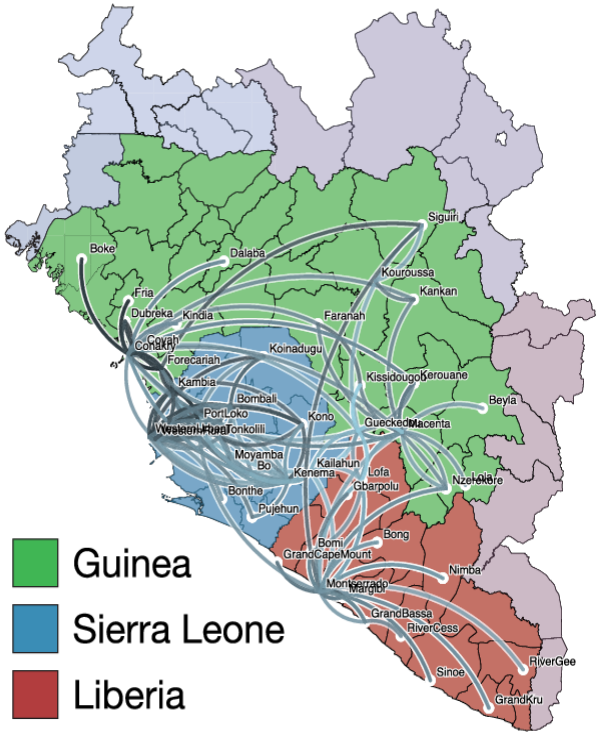
Maximum likelihood



HOW different is it to what's known?

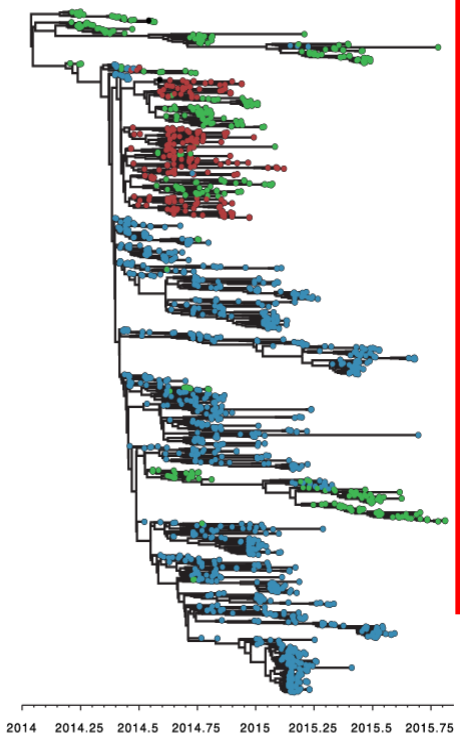
Suchard et al. 2018
Kuhn et al. 2019

Phylodynamique



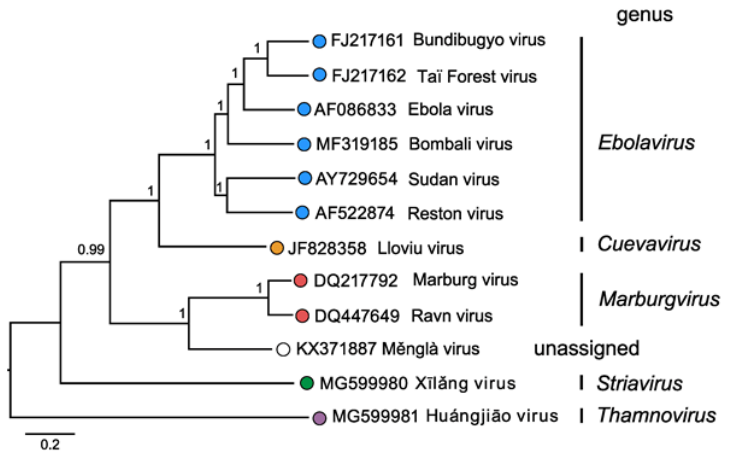
OÙ va-t-il ?

Arbres bayésiens



QUAND est l'ancêtre commun le plus récent ?

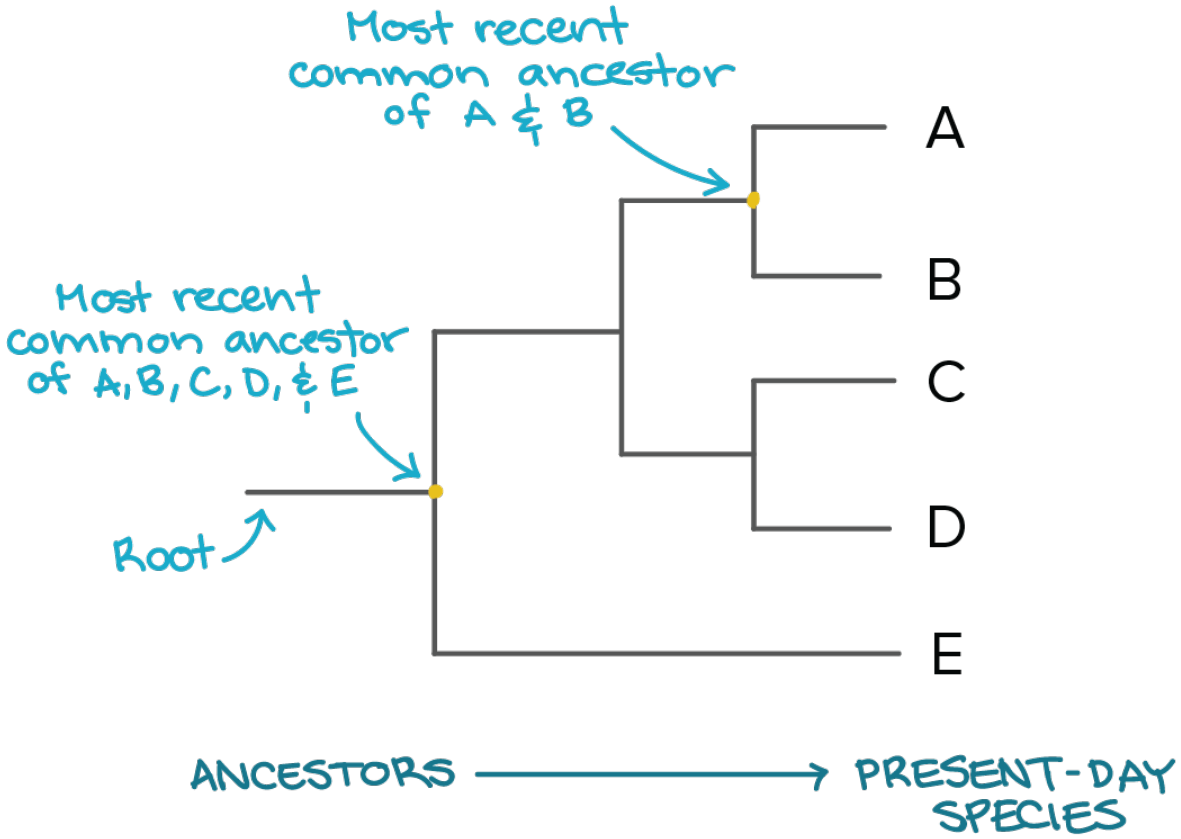
Maximum de vraisemblance



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

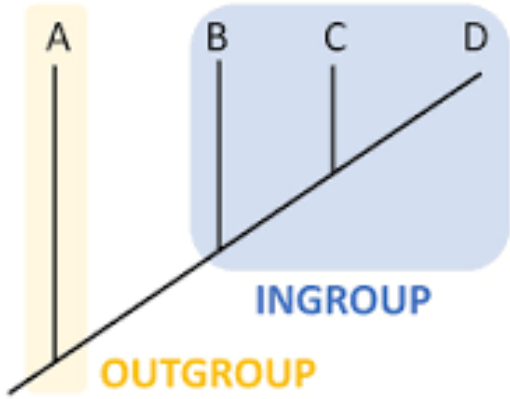
Suchard et al. 2018
Kuhn et al. 2019

Anatomy of a phylogeny

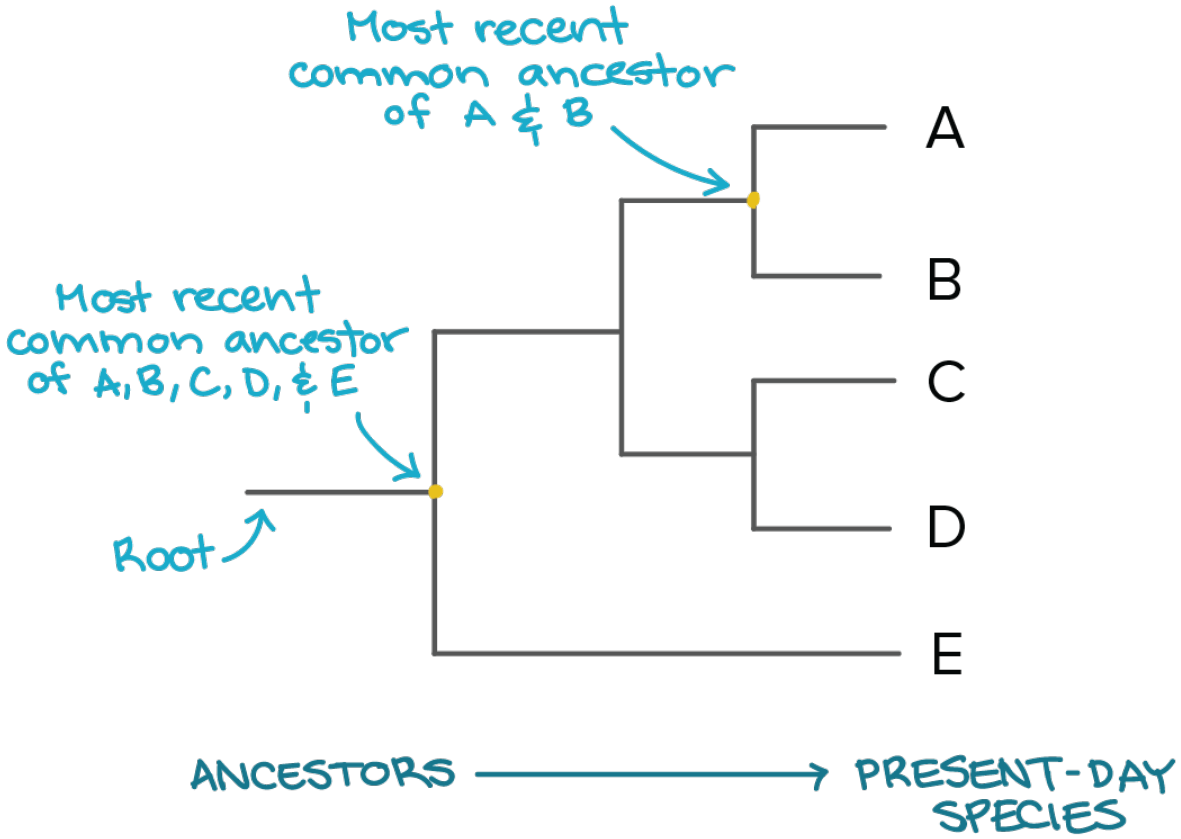


CONFIDENCE

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

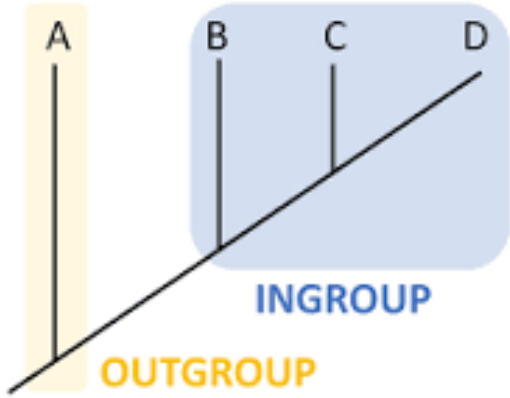


Anatomie d'une phylogénie



CONFIDENCE

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



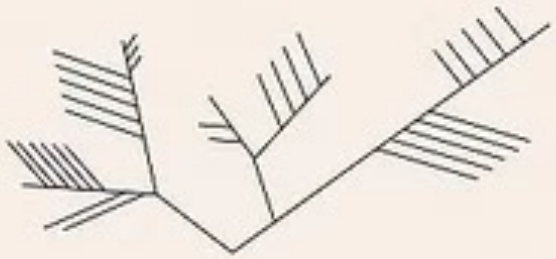
Cladogram versus phylogenetic tree

CLADOGRAM

– the relationships are ***hypothetical***

– you can easily make on your own

PHYLOGENETIC TREE

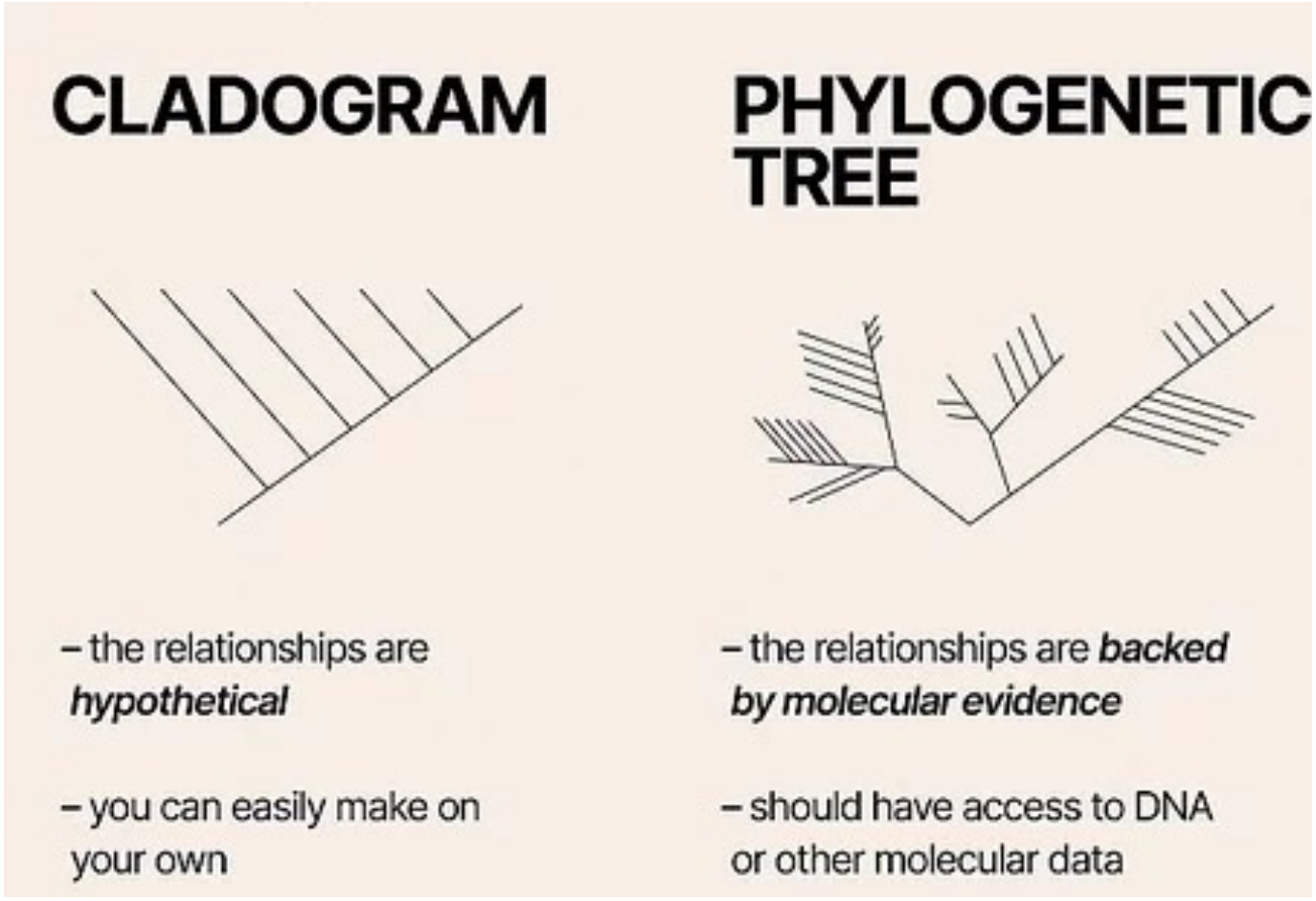


– the relationships are ***backed by molecular evidence***

– should have access to DNA or other molecular data

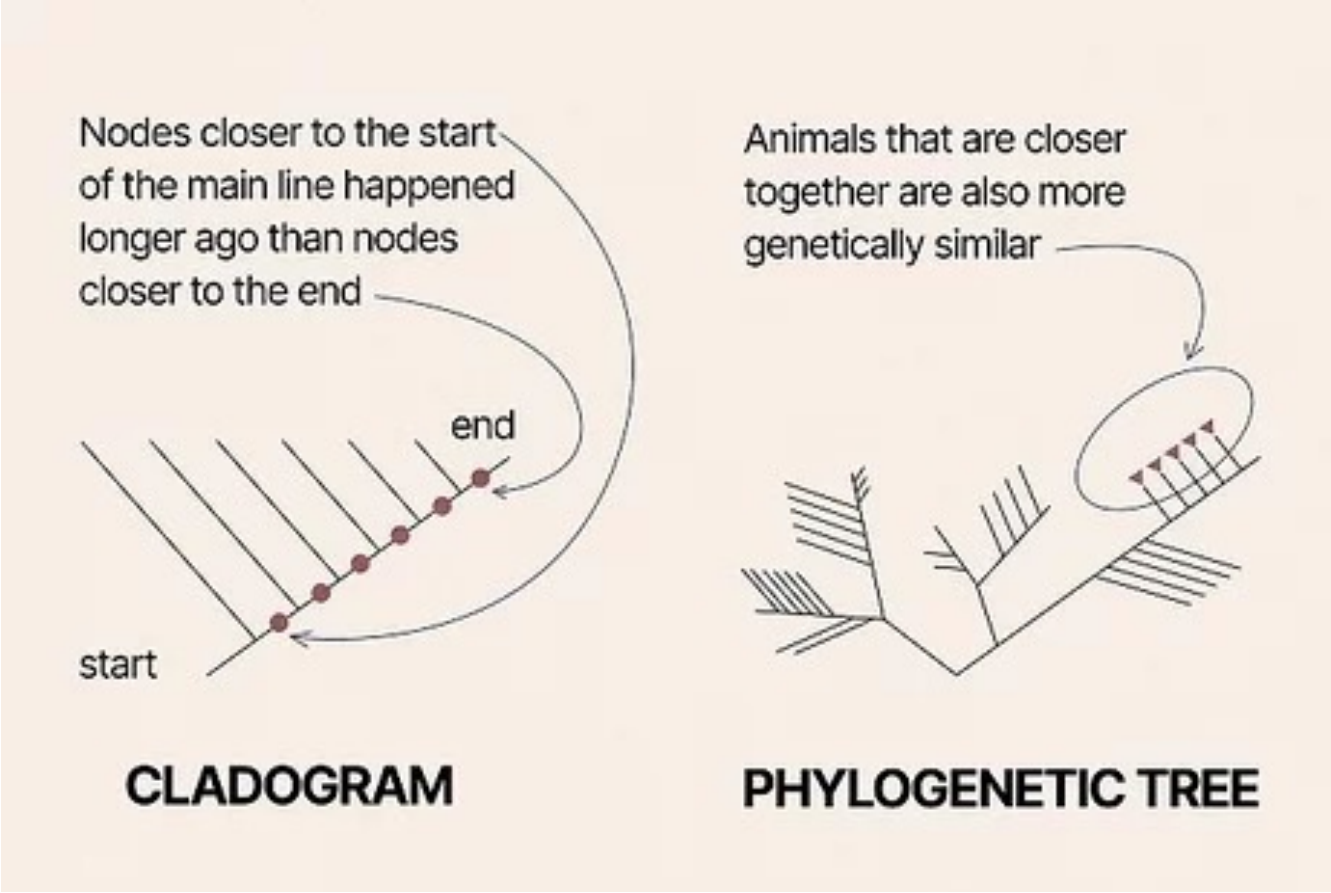
Wikipedia, 2022

Cladogramme versus arbre phylogénétique



Wikipédia, 2022

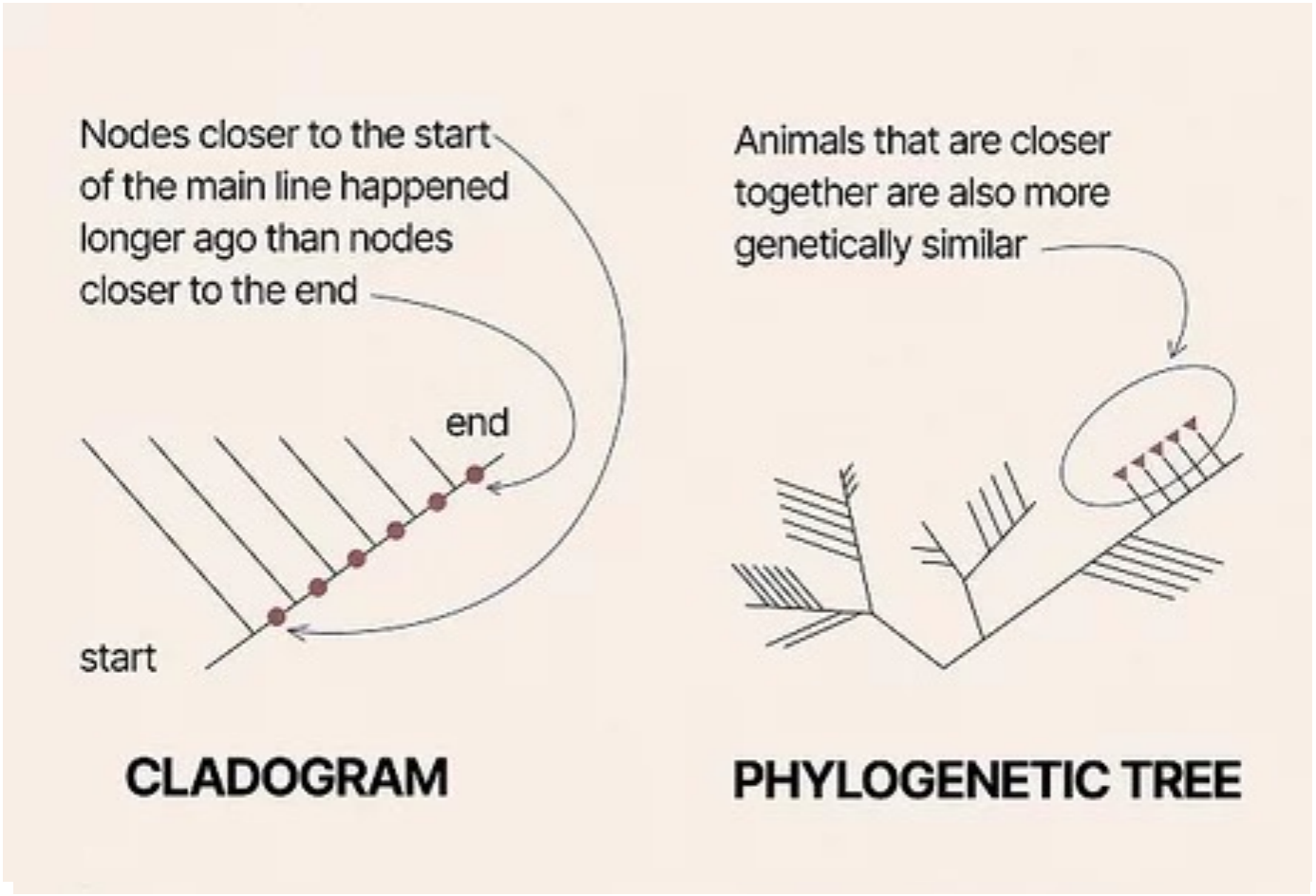
Cladogram versus phylogenetic tree



Why aren't we using cladograms for epidemiology instead of phylogenetic trees?

Wikipedia, 2022

Cladogramme versus arbre phylogénétique

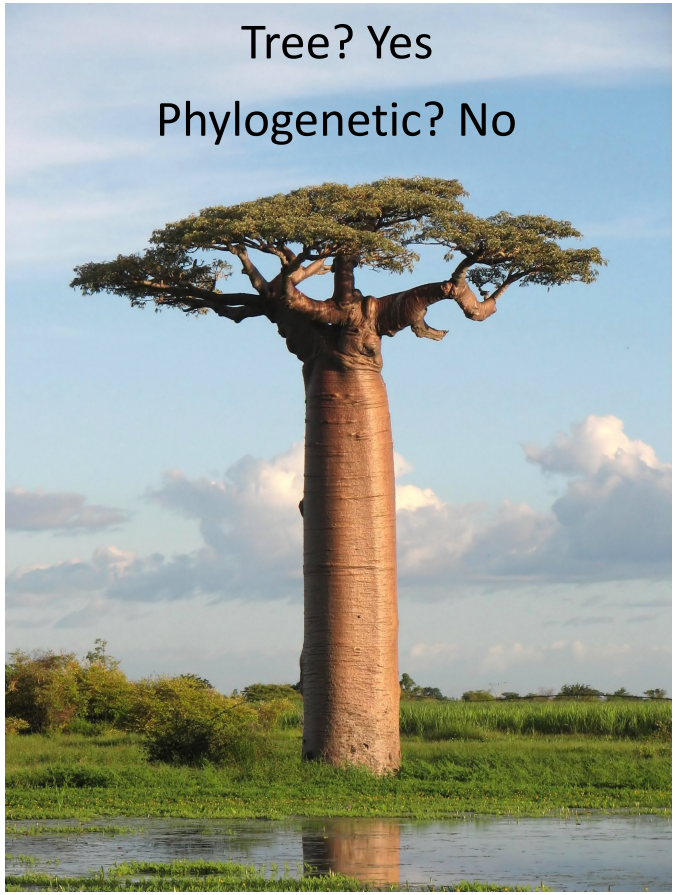


Pourquoi on utilize les phylogénies au lieu des cladogrammes pour l'epidémiologie?

Wikipédia, 2022

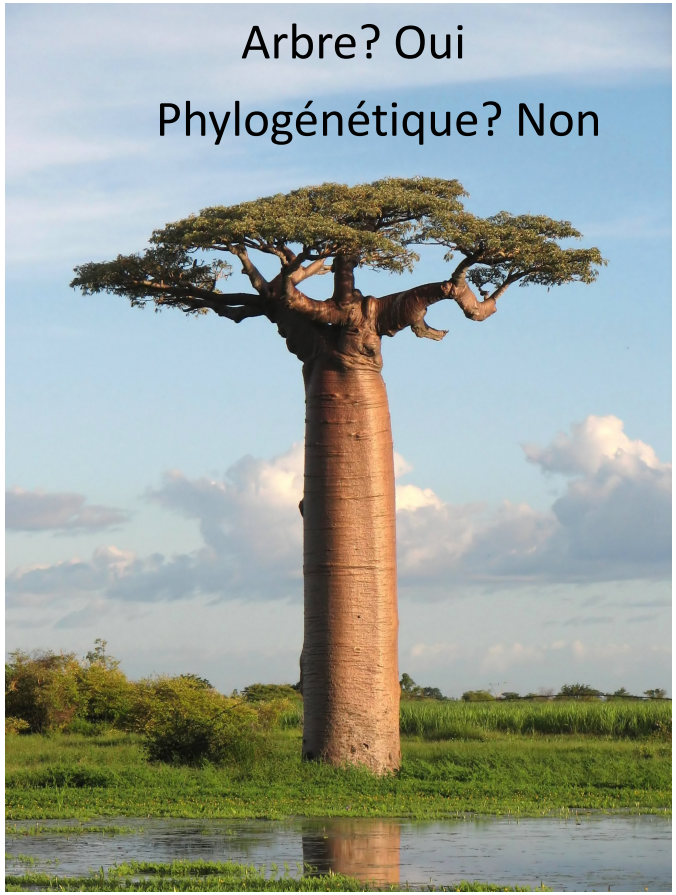
Checkpoint!

- Cladograms are good for generating hypotheses, phylogenies show genetic similarity
- What do bootstrap values show?
- What is a root?



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 ?

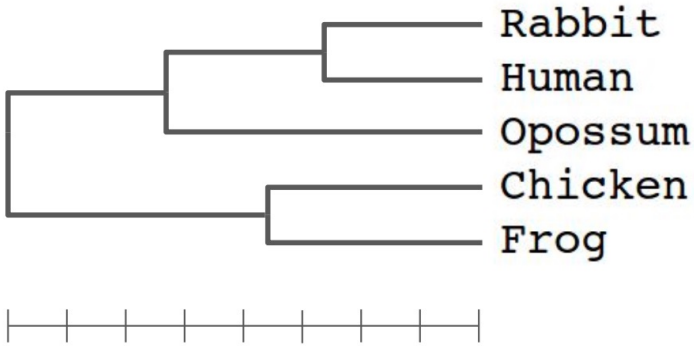


Parsimony versus likelihood

- Parsimony: minimum number of changes
- Likelihood: maximum probability of the data having evolved on the tree

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

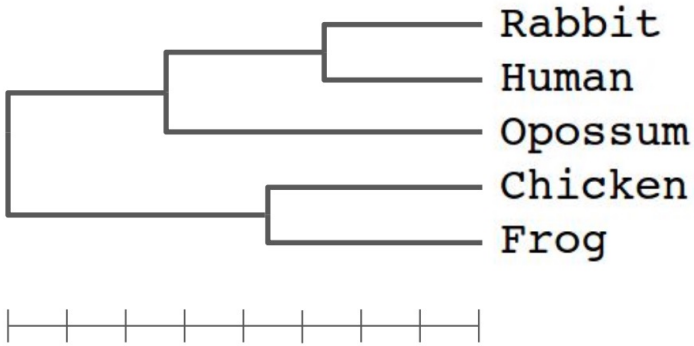
Adapted from Ree and Hipp, UChicago, 2021

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

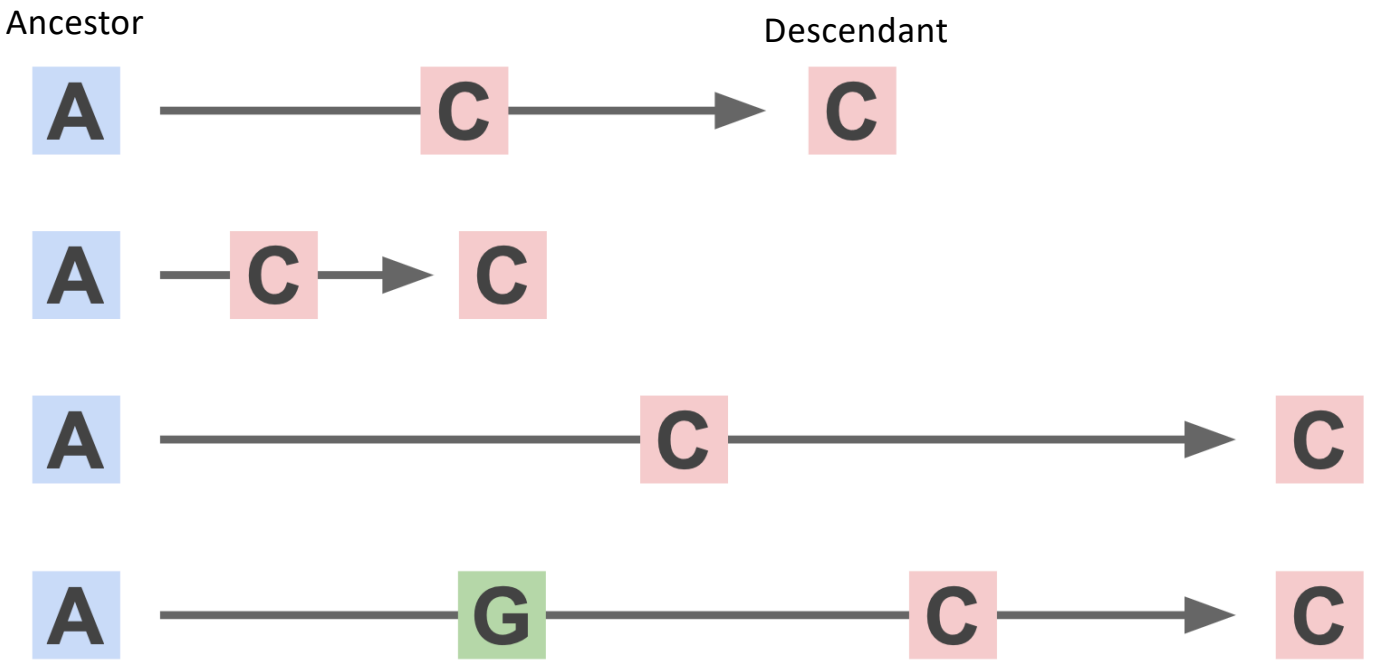


proposed tree has **branch lengths** in units of expected number of changes per site

Adapté de Ree et Hipp, UChicago, 2021

Parsimony: minimum number of changes regardless of time/opportunity

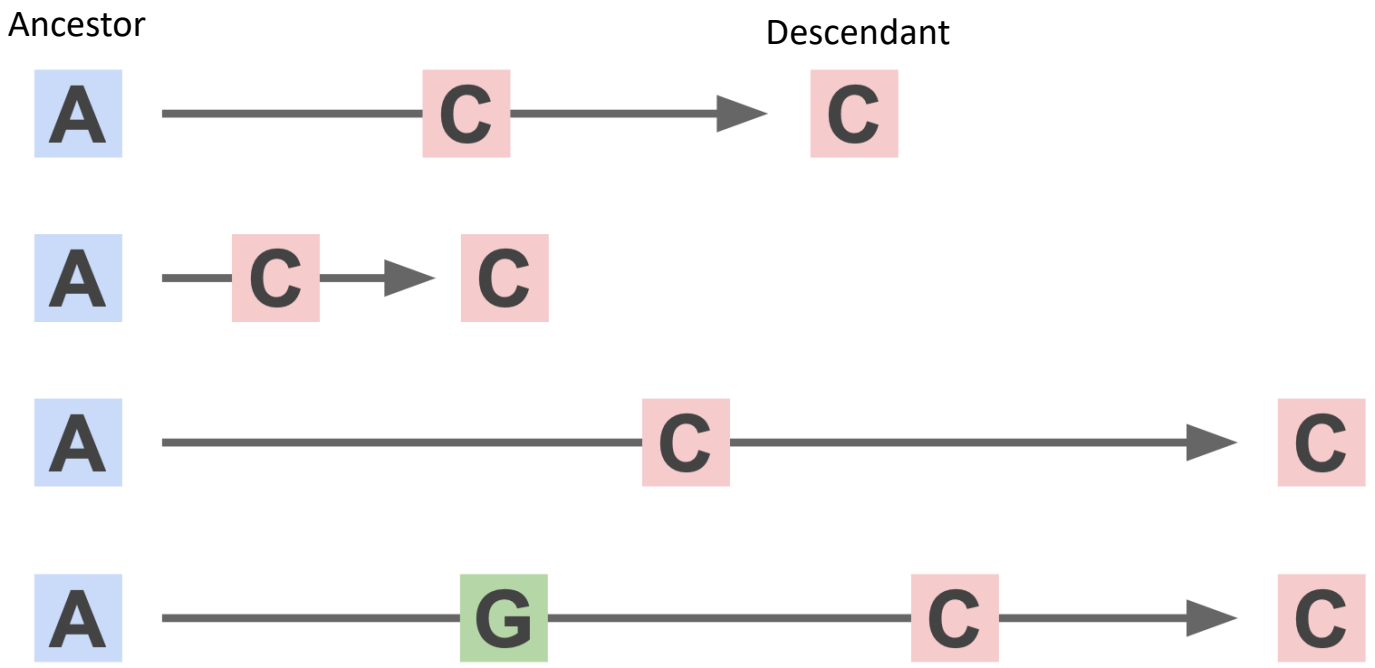
Which is the least parsimonious?



Adapted from Ree and Hipp, UChicago, 2021

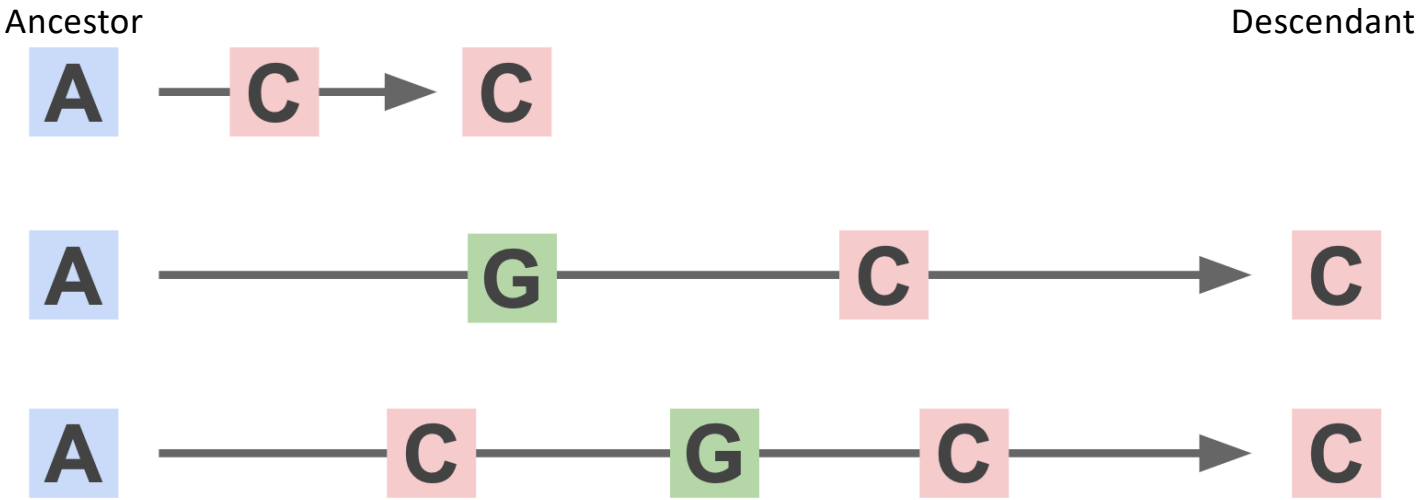
Parcimonie : nombre minimum de changements, quel que soit le moment ou l'opportunité.

Qu'est-ce qui est le moins parcimonieux ?



Adapté de Ree et Hipp, UChicago, 2021

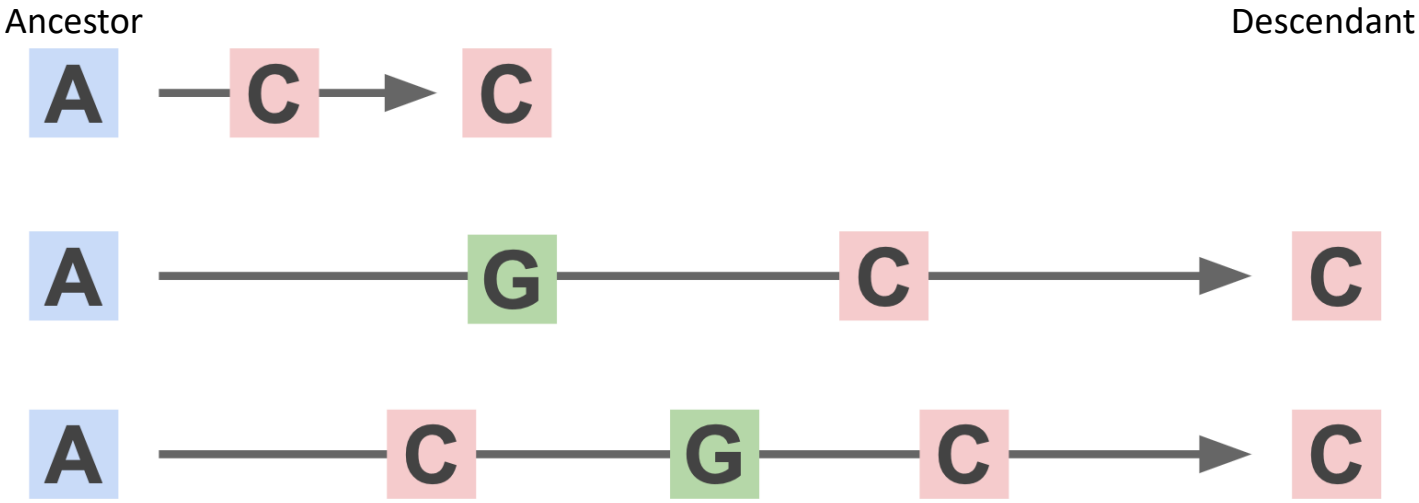
Likelihood: probability of ancestral and descendant status is a function of time (branch length)



We don't know what the actual history of the change is, so use a model of evolution to consider all possible histories (maximum likelihood)

Adapted from Ree and Hipp, UChicago, 2021

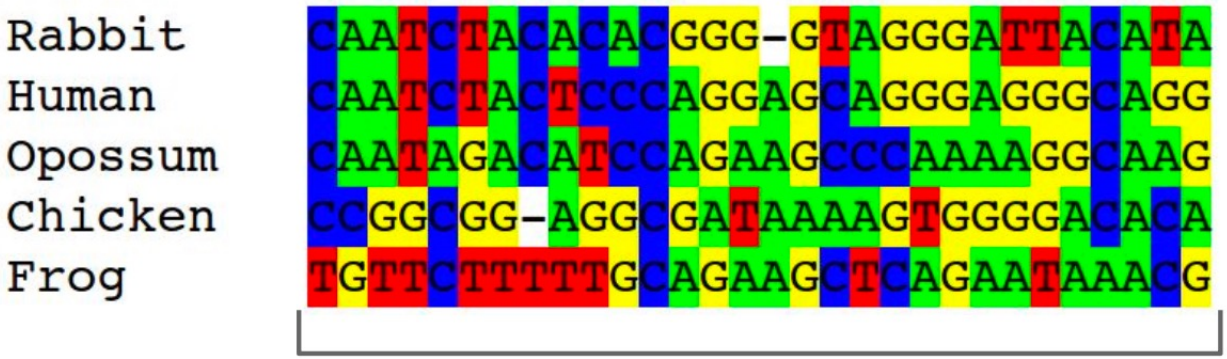
Vraisemblance : la probabilité d'un statut d'ancêtre et de descendant est fonction du temps (longueur de la branche)



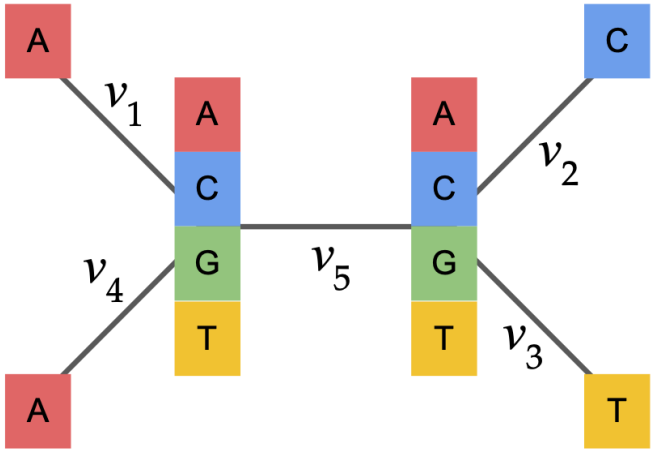
Nous ne savons pas quelle est l'histoire réelle du changement, alors utilisez un modèle d'évolution pour considérer toutes les histoires possibles (probabilité maximale)

Adapté de Ree et Hipp, UChicago, 2021

Likelihood cont'd.



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

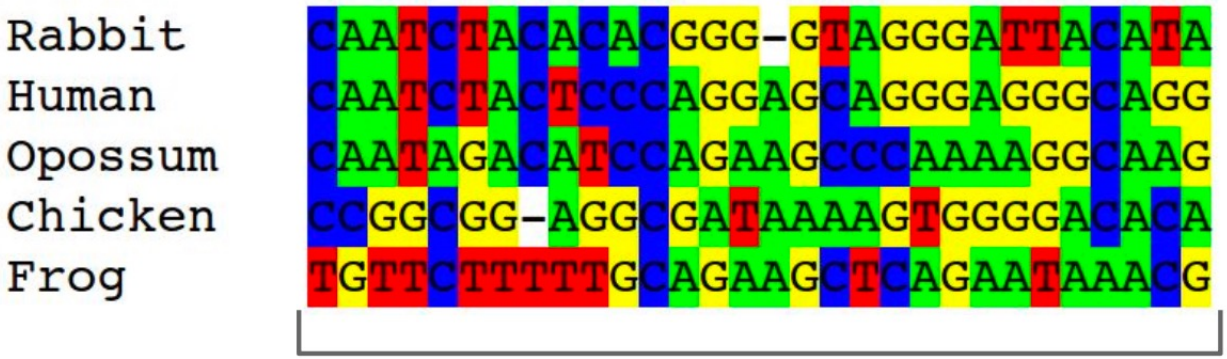


Consider *all possible ancestral states* at internal nodes, and calculate their contribution to the overall likelihood.*

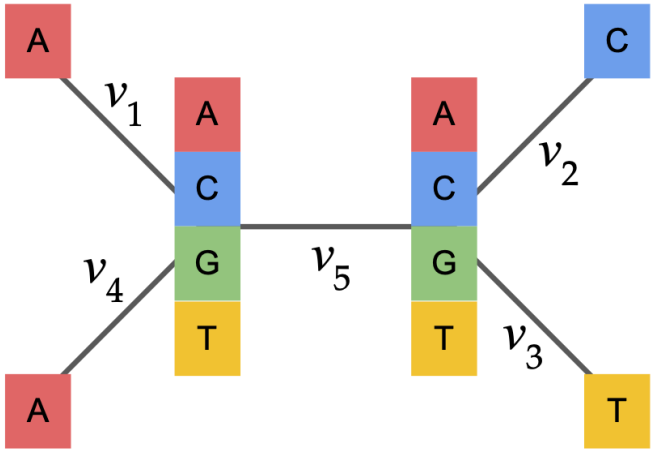
Parameters: tree topology, branch lengths, substitution rates estimated to maximize likelihood of data

Adapted from Ree and Hipp, UChicago, 2021

Vraisemblance (suite).



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



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

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

Adapté de Ree et Hipp, UChicago, 2021

Models of DNA evolution

- Markov models that describe relative rates of different changes
  - JC69 (Jukes and Cantor 1969)
  - K80 model (Kimura 1980)
  - K81 model (Kimura 1981)
  - F81 (Felsenstein 1981)
  - HKY85 model (Hasegawa, Kishino and Yano 1985)
  - T92 model (Tamura 1992)
  - TN93 model (Tamura and Nei 1993)
  - GTR model (Tavaré 1986)
  - Yep there's a lot of them! How do I know what's best for my data?

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 ?

Good news, most people don't need to know the mathematical specifics of these models

JC69 model (Jukes and Cantor 1969) [edit]

JC69, the Jukes and Cantor 1969 model,[2] 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 μ, the overall substitution rate. As previously mentioned, this variable becomes a constant when we normalize the mean-rate to 1.

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

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

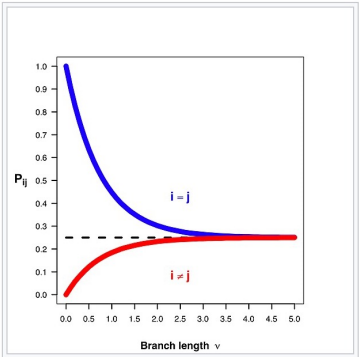
When branch length, ν, is measured in the expected number of changes per site then:

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

It is worth noticing that $\nu = \frac{3}{4}t\mu = (\frac{\mu}{4} + \frac{\mu}{4} + \frac{\mu}{4})t$ 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 μ.

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

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



Probability P_{ij} of changing from initial state i to final state j as a function of the branch length (ν) 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).

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,[2] 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 μ, the overall substitution rate. As previously mentioned, this variable becomes a constant when we normalize the mean-rate to 1.

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

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

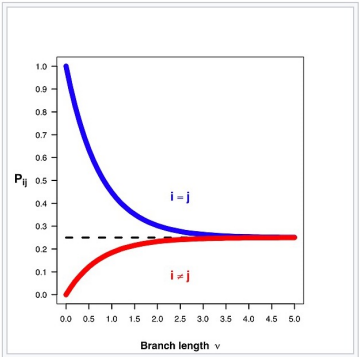
When branch length, ν, is measured in the expected number of changes per site then:

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

It is worth noticing that $\nu = \frac{3}{4}t\mu = (\frac{\mu}{4} + \frac{\mu}{4} + \frac{\mu}{4})t$ 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 μ.

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

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



Probability P_{ij} of changing from initial state i to final state j as a function of the branch length (ν) 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).

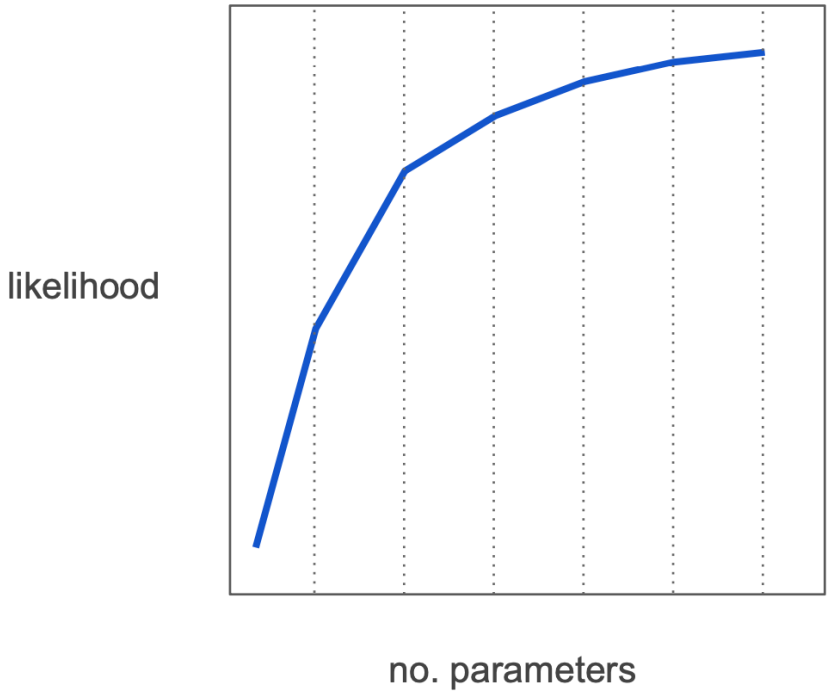
Model selection

More parameters means higher likelihood, but is the increase in likelihood necessary? Adds much more complexity

- Programs will use statistical methods to answer this question using Akaike Information Criteria (AIC), Bayesian Information Criterion (BIC), likelihood ratio tests, etc.

Model testing will give you BIC and AIC score

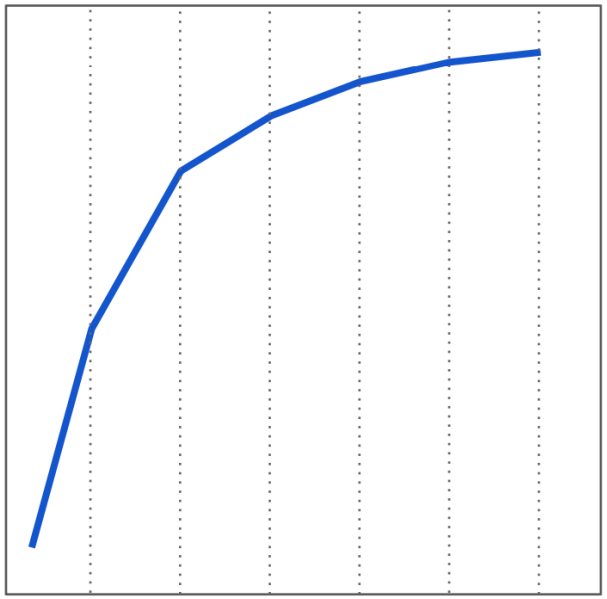
- AIC score: tries to select the model that most adequately describes an unknown, high dimensional reality
- BIC score: tries to find the TRUE model among the set of candidates



Adapted from Ree and Hipp, UChicago, 2021

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

Adapté de Ree et Hipp, UChicago, 2021

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 (Jukes and Cantor, 1969).	000000
F81	3	Equal rates but unequal base freq. (Felsenstein, 1981).	000000
K80 or K2P	1	Unequal transition/transversion rates and equal base freq. (Kimura, 1980).	010010
HKY or HKY85	4	Unequal transition/transversion rates and unequal base freq. (Hasegawa, Kishino and Yano, 1985).	010010
TN or TN93	5	Like `HKY` but unequal purine/pyrimidine rates (Tamura and Nei, 1993).	010020
TNe	2	Like `TN` but equal base freq.	010020
K81 or K3P	2	Three substitution types model and equal base freq. (Kimura, 1981).	012210
K81u	5	Like `K81` but unequal base freq.	012210
TPM2	2	AC=AT, AG=CT, CG=GT and equal base freq.	010212
TPM2u	5	Like `TPM2` but unequal base freq.	010212
TPM3	2	AC=CG, AG=CT, AT=GT and equal base freq.	012012
TPM3u	5	Like `TPM3` but unequal base freq.	012012
TIM	6	Transition model, AC=GT, AT=CG and unequal base freq.	012230
TIMe	3	Like `TIM` but equal base freq.	012230
TIM2	6	AC=AT, CG=GT and unequal base freq.	010232
TIM2e	3	Like `TIM2` but equal base freq.	010232
TIM3	6	AC=CG, AT=GT and unequal base freq.	012032
TIM3e	3	Like `TIM3` but equal base freq.	012032
TVM	7	Transversion model, AG=CT and unequal base freq.	012314
TVMe	4	Like `TVM` but equal base freq.	012314
SYM	5	Symmetric model with unequal rates but equal base freq. (Zharkikh, 1994).	012345
GTR	8	General time reversible model with unequal rates and unequal base freq. (Tavare, 1986).	012345

Bootstrapping

- Specify number of replicates: how many times does the test replicate the original sequence alignment?
- Standard in MEGA is 500 replicates, 1000 is better but takes longer

Original
sequence alignment

Site number

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Species A	A	A	T	G	C	T	A	G	T	G	G	T	G	A	T
Species B	A	A	G	C	T	A	T	G	G	T	G	A	T	C	G
Species C	A	G	C	C	T	A	T	G	T	G	G	A	A	C	G
Species D	A	A	C	C	C	A	T	T	G	G	G	T	G	A	T

Original alignment
tree

Bootstrap
pseudo-replicate #1

	5	3	3	1	12	9	2	4	11	13	10	14	8	11	13
Species A	C	T	T	A	T	T	A	G	G	G	A	G	G	G	
Species B	T	G	G	A	A	G	A	C	G	T	T	C	G	G	T
Species C	T	C	C	A	A	T	G	C	G	A	G	C	G	G	A
Species D	C	C	C	A	T	T	A	C	G	G	G	A	T	G	G

Bootstrap tree #1

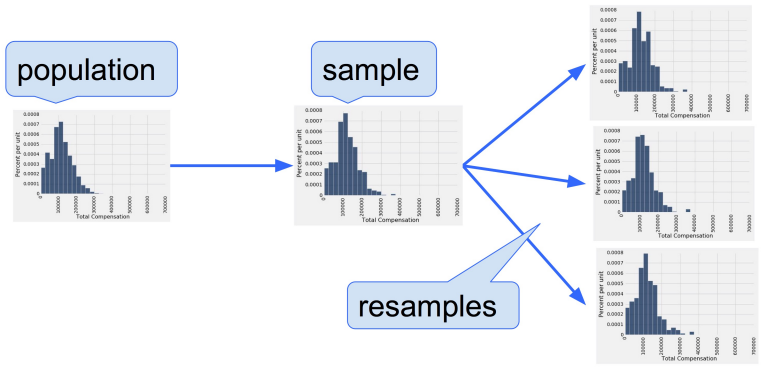
Bootstrap
pseudo-replicate #2

	9	7	12	5	2	4	2	6	14	9	4	9	7	2	1
Species A	T	A	T	C	A	G	A	T	A	T	G	T	A	A	A
Species B	G	T	A	T	A	C	A	A	C	G	C	G	T	A	A
Species C	T	T	A	T	G	C	G	A	C	T	C	T	T	G	A
Species D	T	T	T	C	A	C	A	A	A	T	C	T	T	A	A

Bootstrap tree #2

Bootstrapping

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



Original
sequence alignment

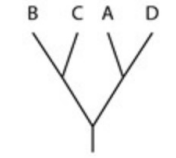
Site number

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Species A	A	A	T	G	C	T	A	G	T	G	G	T	G	A	T
Species B	A	A	G	C	T	A	T	G	G	T	G	A	T	C	G
Species C	A	G	C	C	T	A	T	G	T	G	G	A	A	C	G
Species D	A	A	C	C	C	A	T	T	G	G	G	T	G	A	T

Original alignment tree

Bootstrap
pseudo-replicate #1

	5	3	3	1	12	9	2	4	11	13	10	14	8	11	13
Species A	C	T	T	A	T	T	A	G	G	G	A	G	G	G	
Species B	T	G	G	A	A	G	A	C	G	T	T	C	G	G	T
Species C	T	C	C	A	A	T	G	C	G	A	G	C	G	G	A
Species D	C	C	C	A	T	T	A	C	G	G	G	A	T	G	G



Bootstrap tree #1

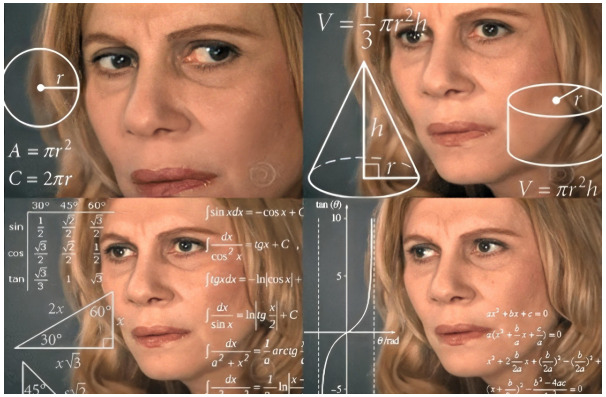
Bootstrap
pseudo-replicate #2

	9	7	12	5	2	4	2	6	14	9	4	9	7	2	1
Species A	T	A	T	C	A	G	A	T	A	T	G	T	A	A	A
Species B	G	T	A	T	A	C	A	A	C	G	C	G	T	A	A
Species C	T	T	A	T	G	C	G	A	C	T	C	T	T	G	A
Species D	T	T	T	C	A	C	A	A	A	T	C	T	T	A	A

Bootstrap tree #2

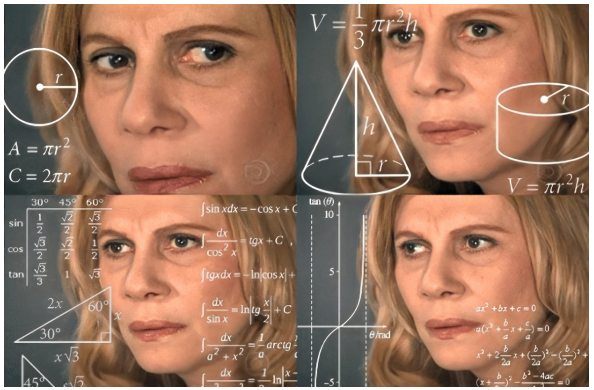
Checkpoint

- More parameters = higher or lower likelihood?
- Low bootstrap values means low confidence in tree topology
- Not all sequences undergo evolution the same way, models account for this change



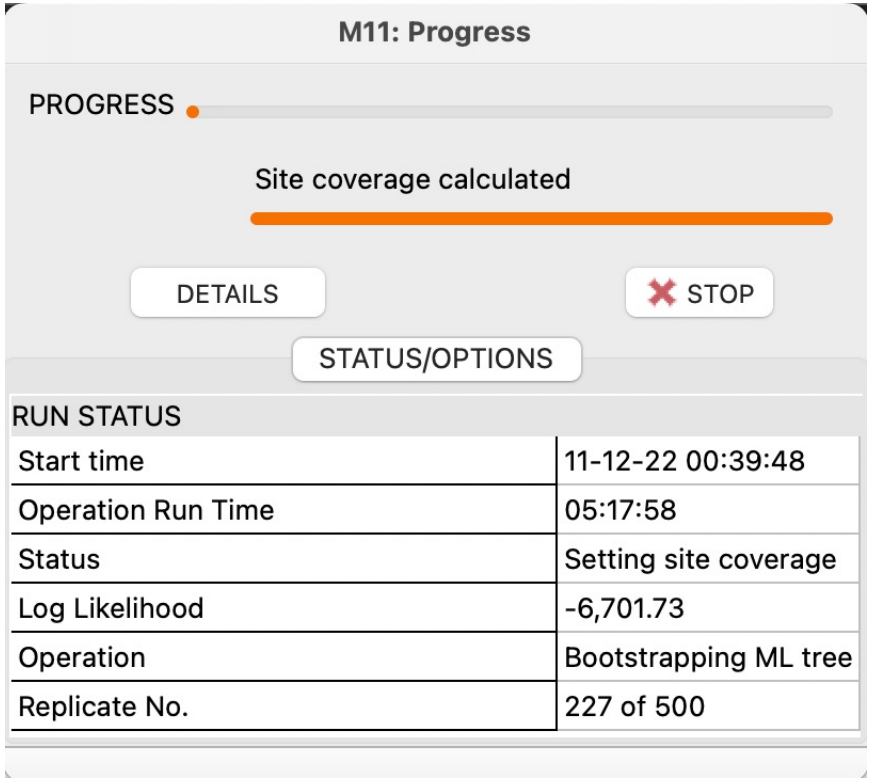
Point de contrôle!

- 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



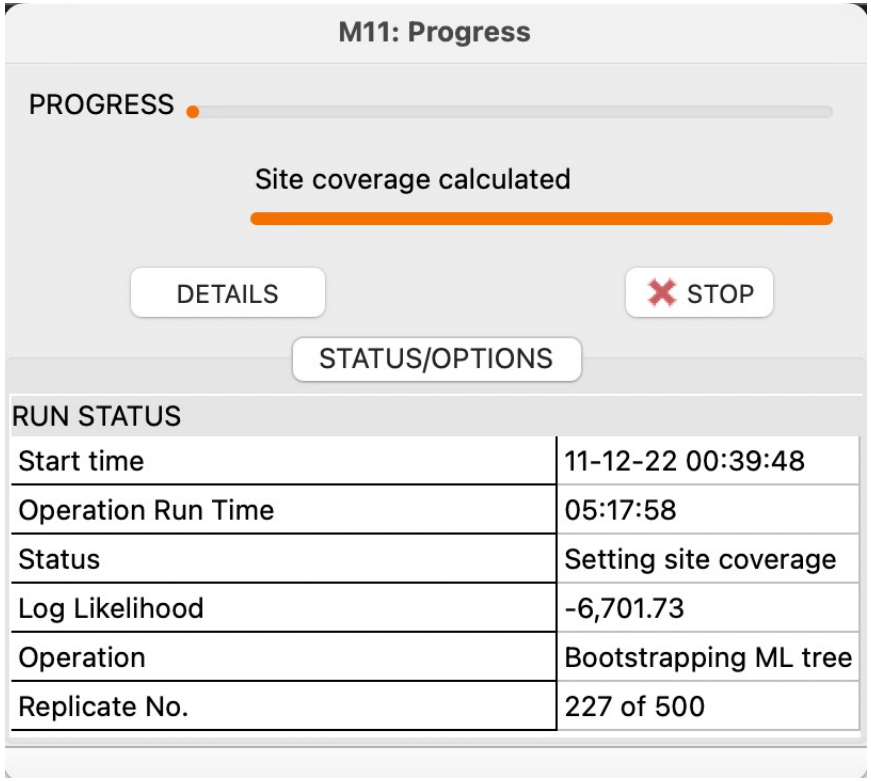
Warnings and limitations

- Building phylogenies takes a LONG time, larger ones can take up to a week to run and Bayesian phylogenies can run for months, so a computing cluster is almost necessary for this
- Without a proper outgroup or root, a phylogeny doesn't tell you much about order of descent
- **Why was it so hard to make phylogenies for SARS-CoV-2?**



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



JOURNAL ARTICLE

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.

Putting your tree in context is important, without a control, can you really infer anything?

JOURNAL ARTICLE

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

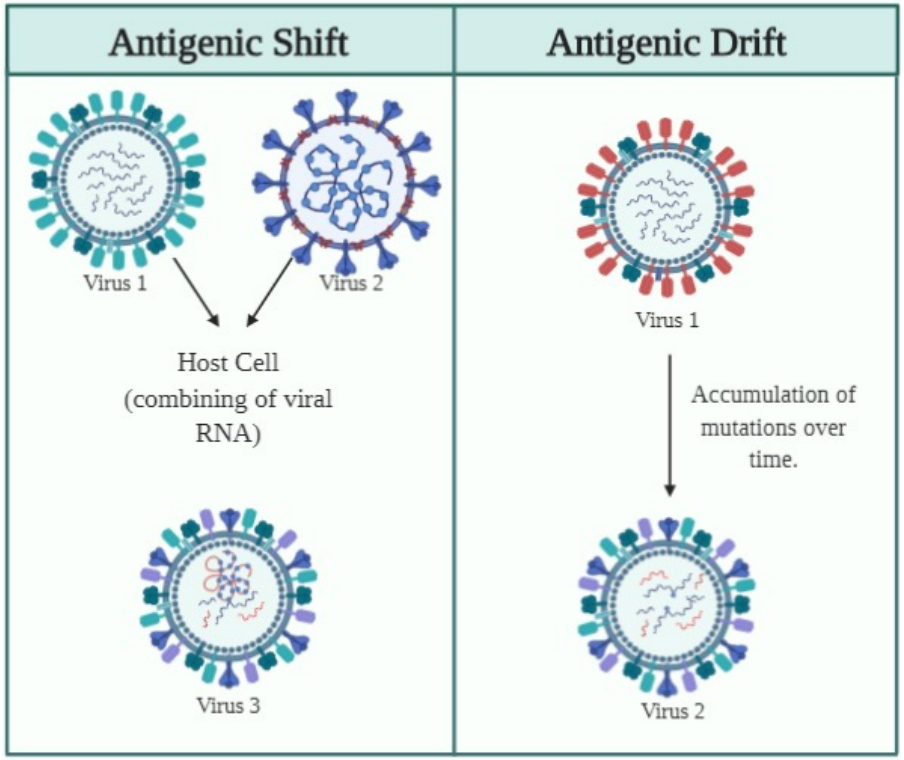
PDF Split View Cite Permissions Share ▾

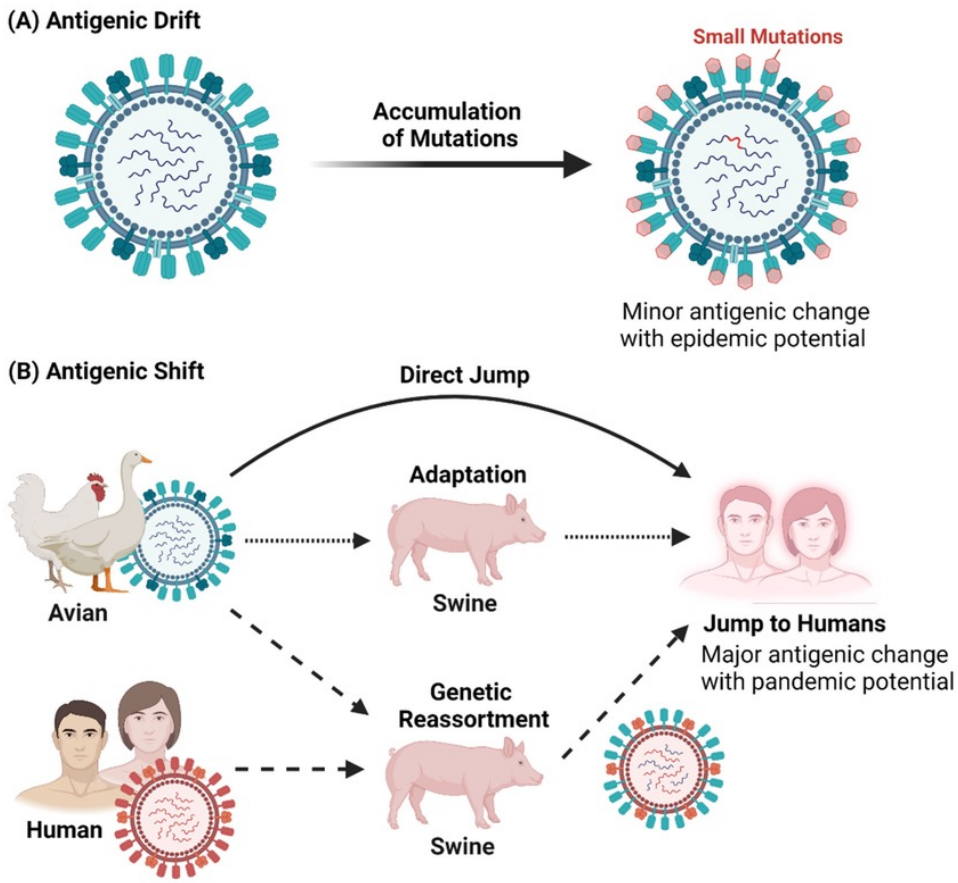
Abstract

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?

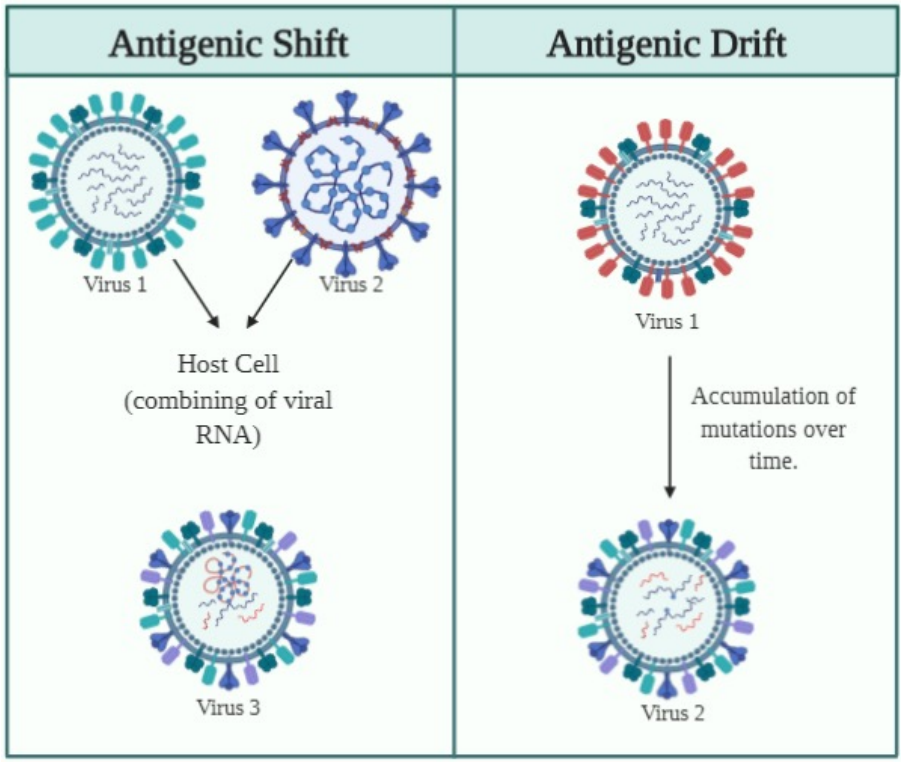
Case study: influenza – antigenic shift and drift

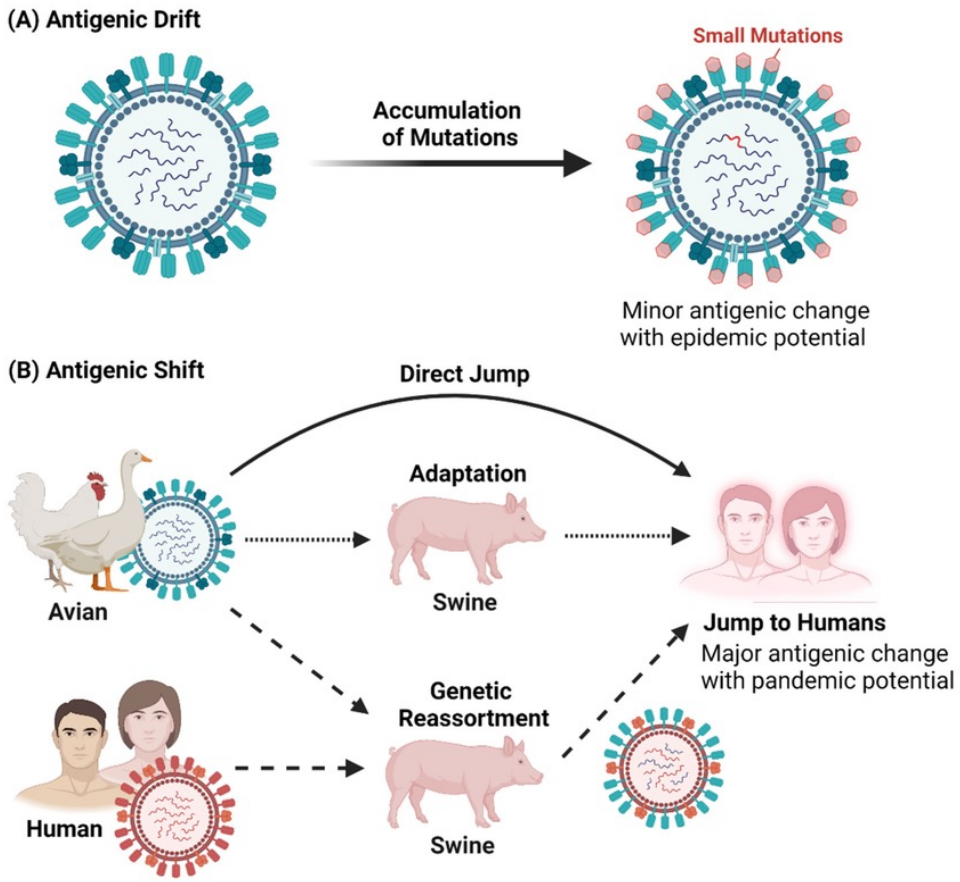




Yeo et al. 2021

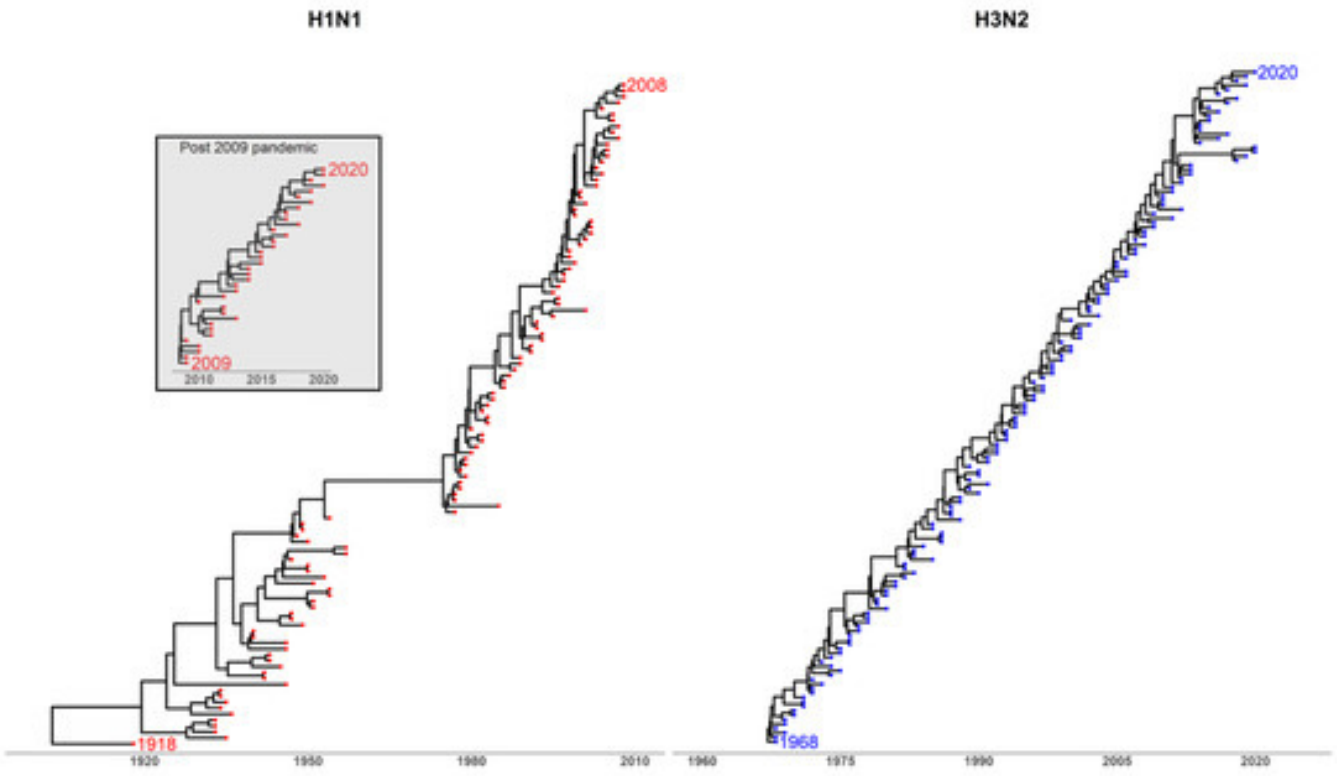
Étude de cas : grippe – décalage et dérive antigéniques





Yeo et al. 2021

Case study: influenza – antigenic shift and drift

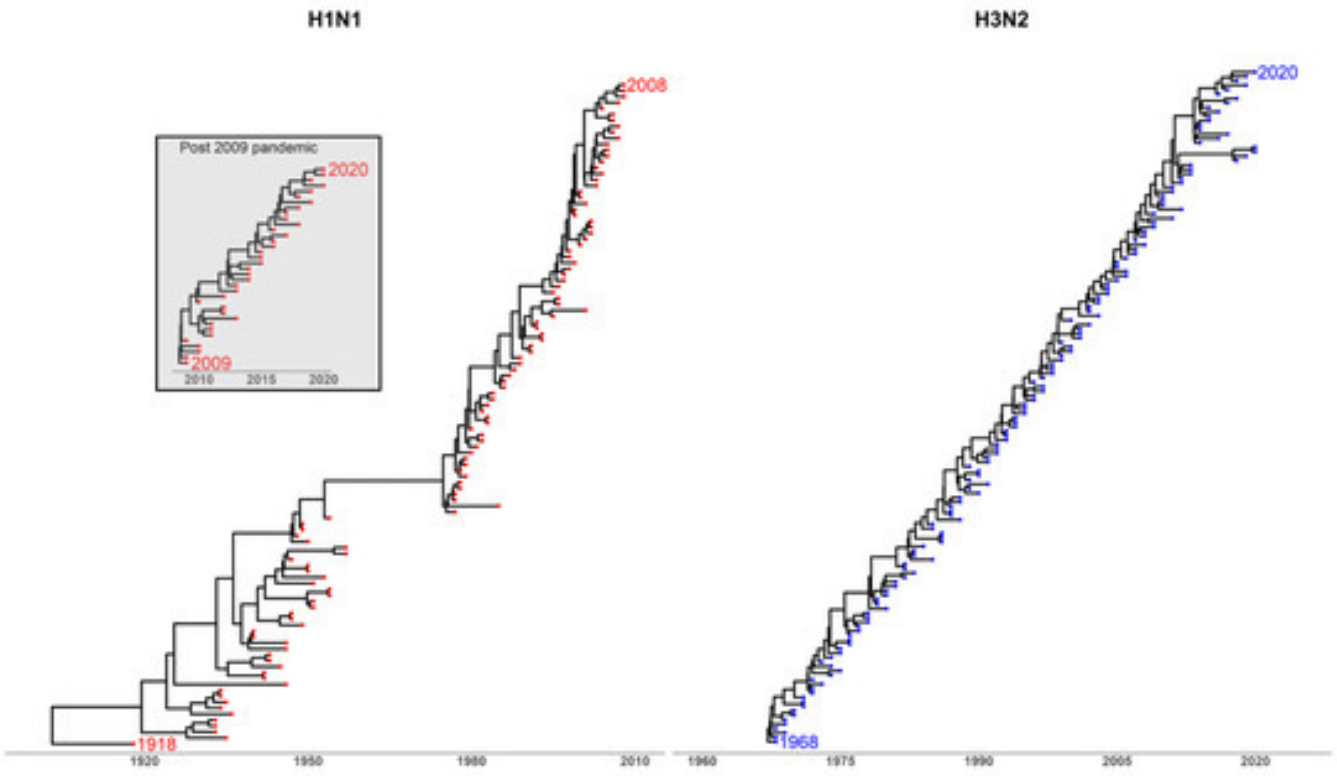


Bolton et al. 2021

Is drift or shift happening here?

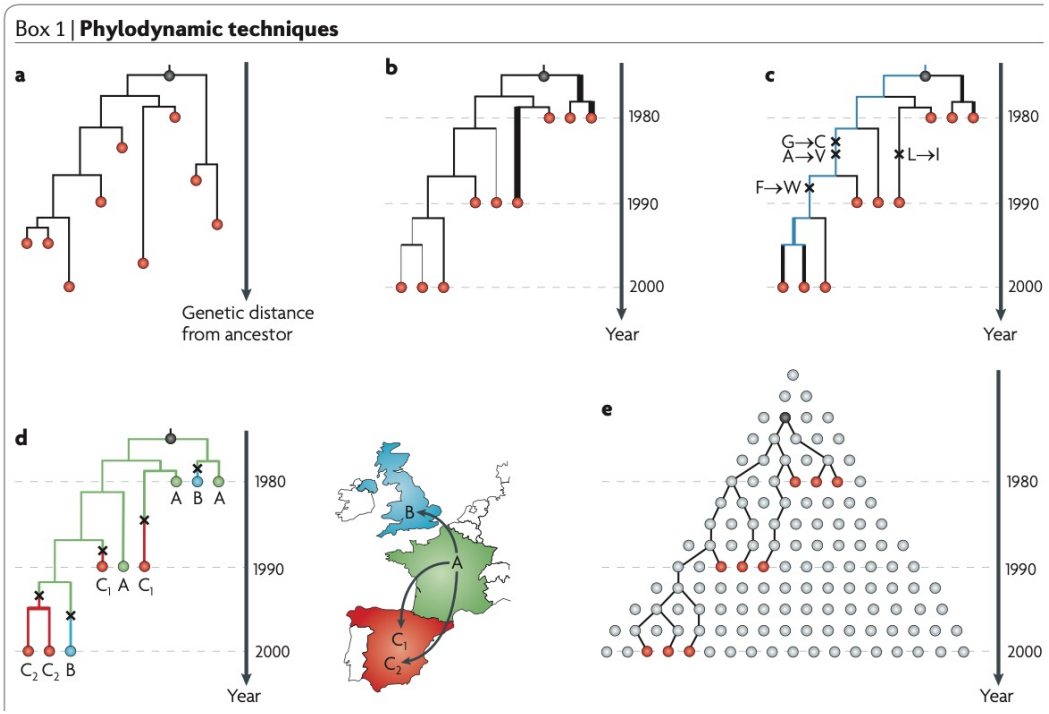
How does this impact vaccine design?

Étude de cas : grippe – décalage et dérive antigéniques



Bolton et al. 2021

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

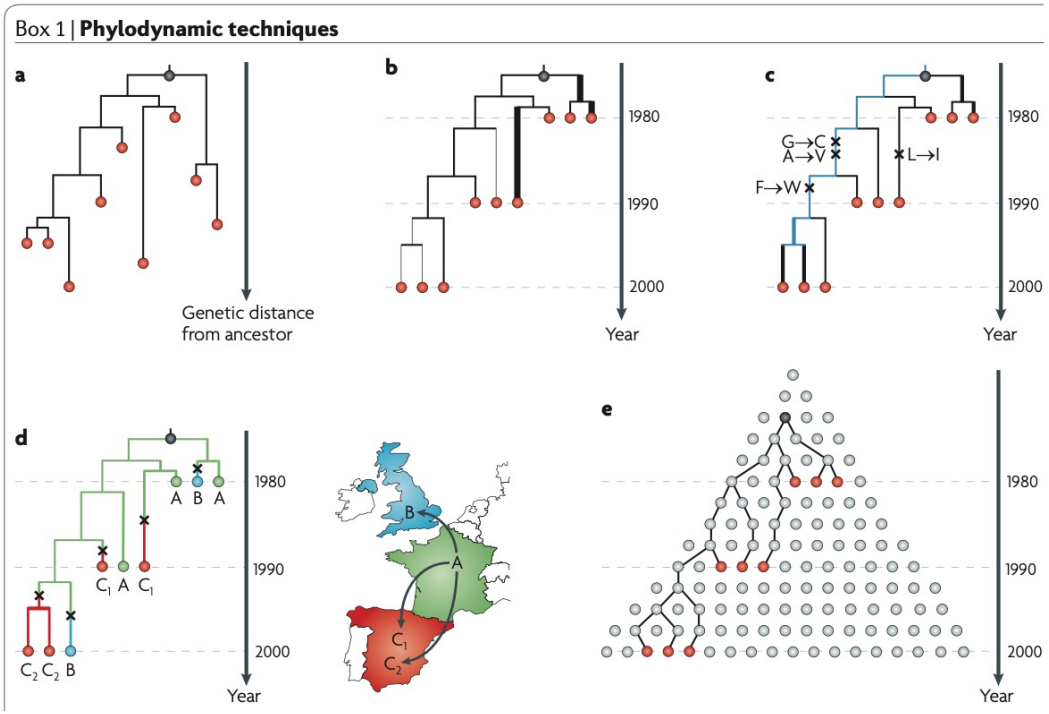


Units matter

Grenfell et al. 2004
Pybus and Rambaut 2009

Shapes matter

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



L'importance des unités

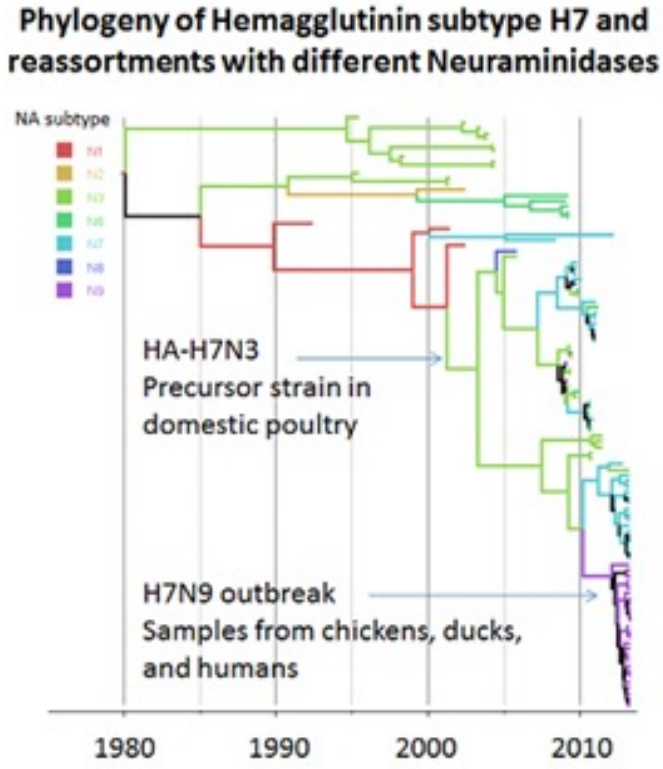
Grenfell et al. 2004
Pybus and Rambaut 2009

L'importance des formes

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

You can do a lot with phylogenies…

- Phylodynamics
- Can look at phylogeny in context of other factors
  - Time (how long ago did this virus diverge)
  - Location (how did a virus change as it spread?)
  - Host (how did a virus change in different hosts?)
- Maximum likelihood phylogenies good for:
  - How different/similar is one thing in comparison to known things

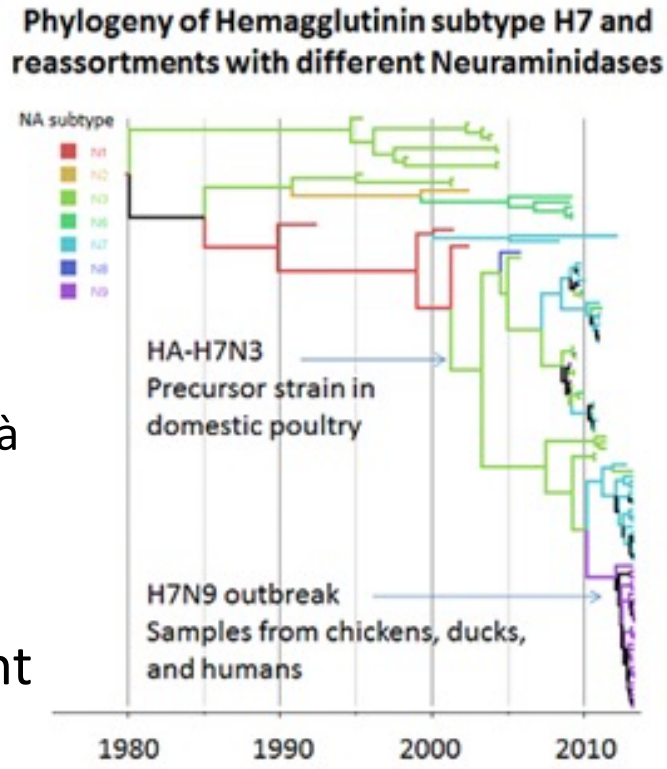


BEAST

Bayesian Evolutionary Analysis Sampling Trees

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

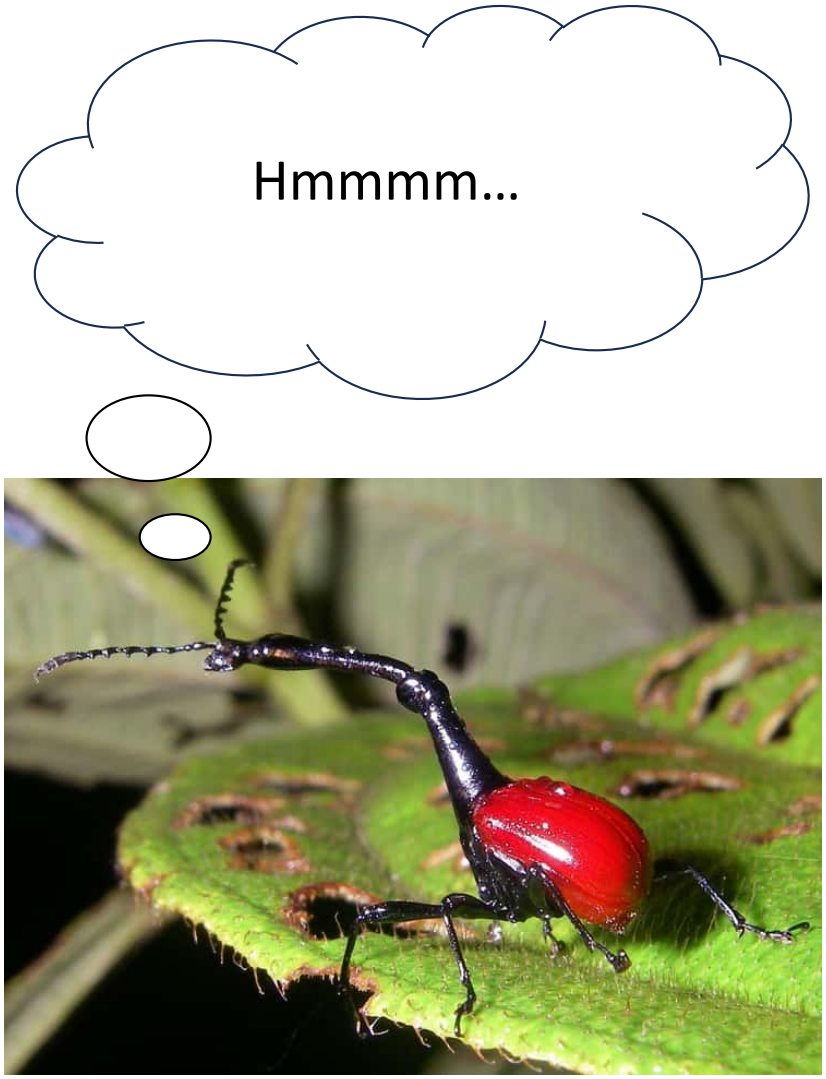


BEAST

Bayesian Evolutionary Analysis Sampling Trees

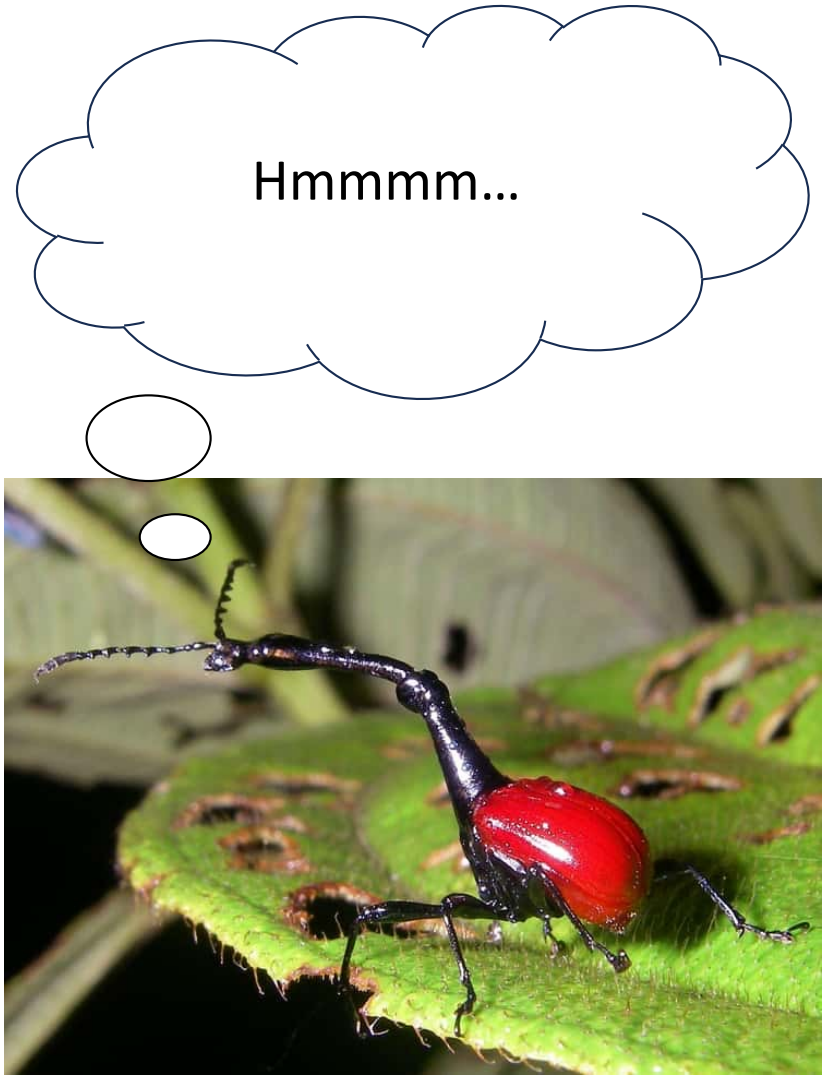
Checkpoint

- Phylogenetics is useful but computationally intensive
- Why do we need a root in phylogenetic trees?
- Can shape of a phylogeny tell us anything about the subject?



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 ?



So you have your sequences, now what?

- Get some reference sequences from NCBI
- Get an outgroup from NCBI
- Align them (use a software like MEGA or online like MAFFT)
- Pick the best model (use a software like MEGA or ModelTest-NG)
- Run the phylogeny using your aligned sequences and chosen model (use a software like MEGA or RAxML)
- Visualize/edit tree in either R or FigTree

All of this listed is free to use ☺

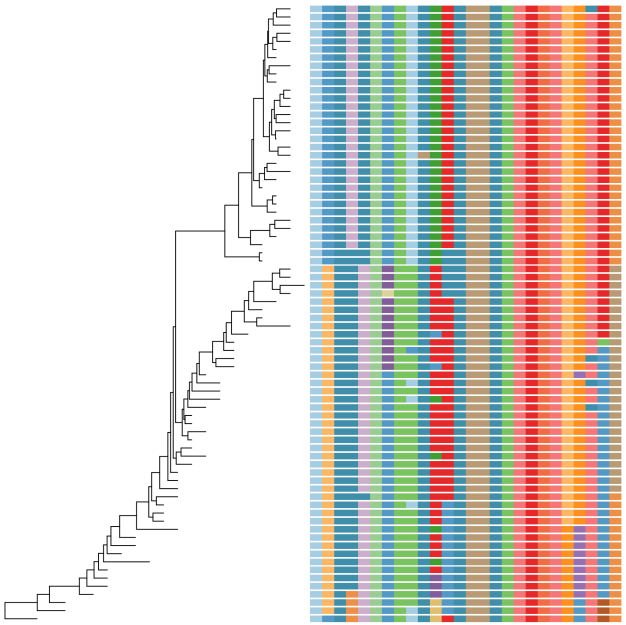
Vous avez donc vos séquences, et maintenant?

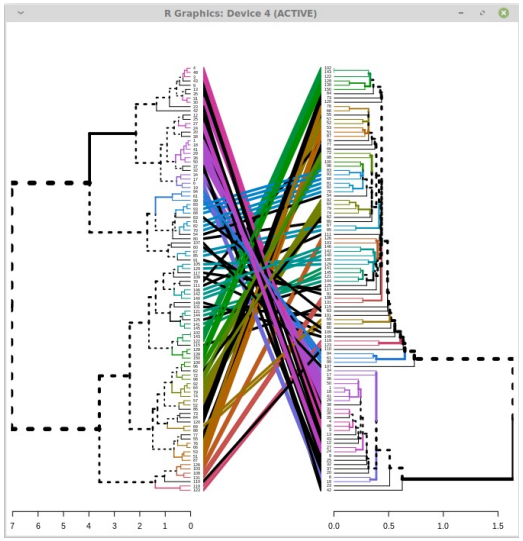
- Obtenir des séquences de référence du NCBI
- Obtenir un groupe externe de NCBI
- Alignez-les (utilisez un logiciel comme MEGA ou en ligne, vous avez donc vos séquences, et maintenant ? comme MAFFT)
- Choisissez le meilleur modèle (utilisez un logiciel comme MEGA ou ModelTest-NG)
- Exécutez la phylogénie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

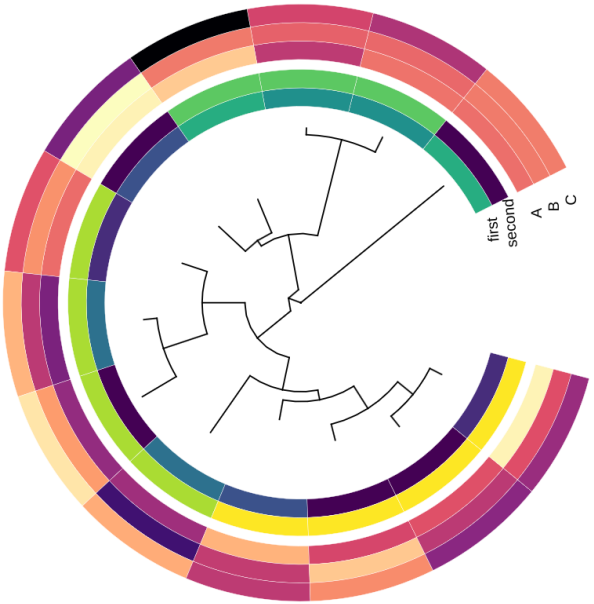
Tout ce qui est répertorié est gratuit à utiliser

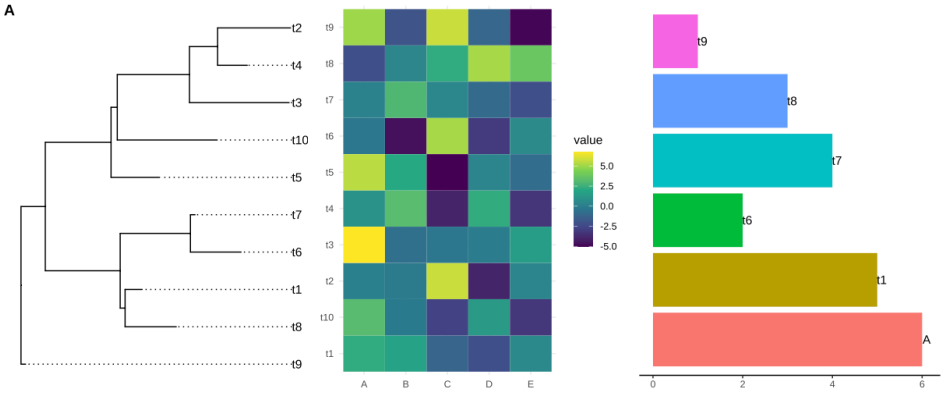
Potential for pretty figures

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



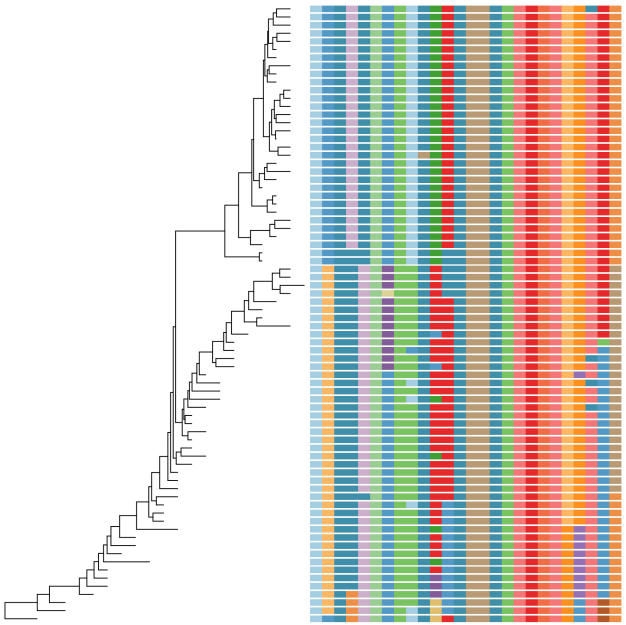


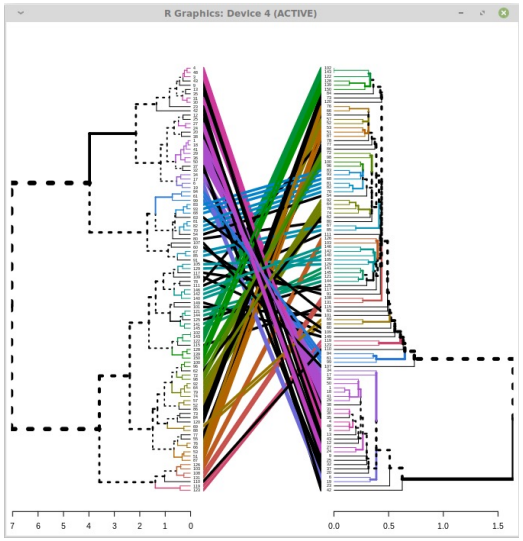


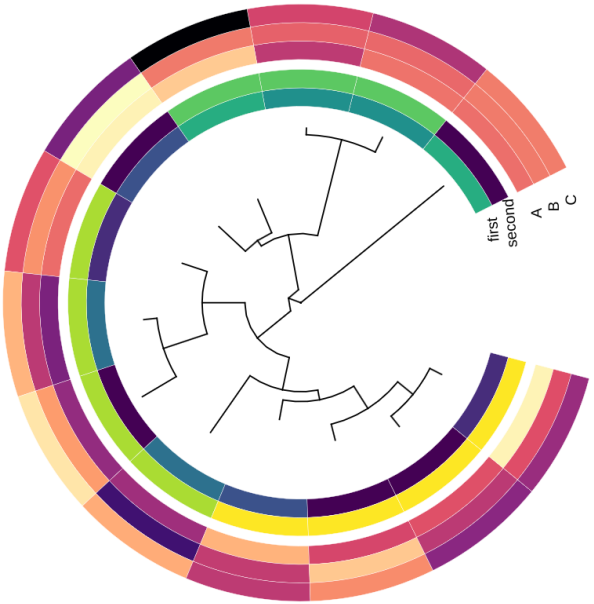


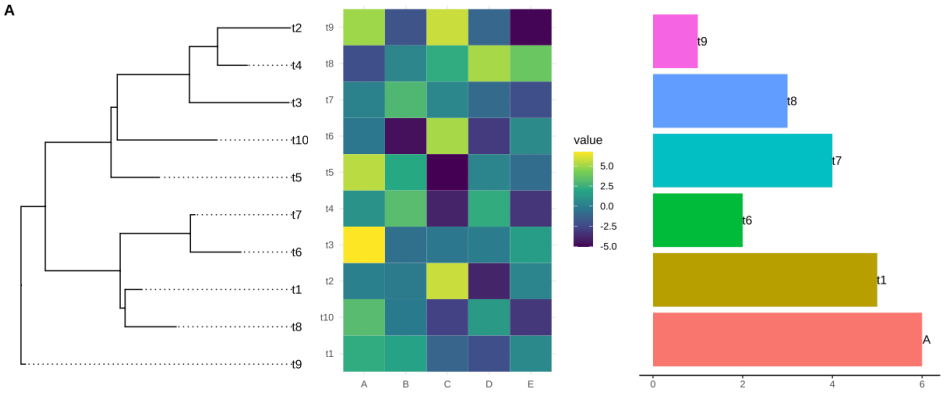
Potentiel pour de jolies figurines

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



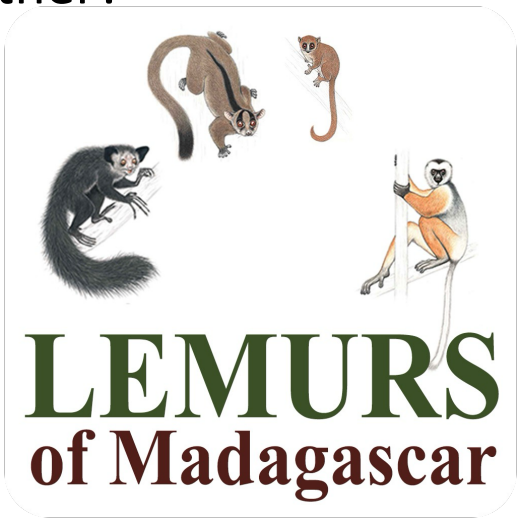






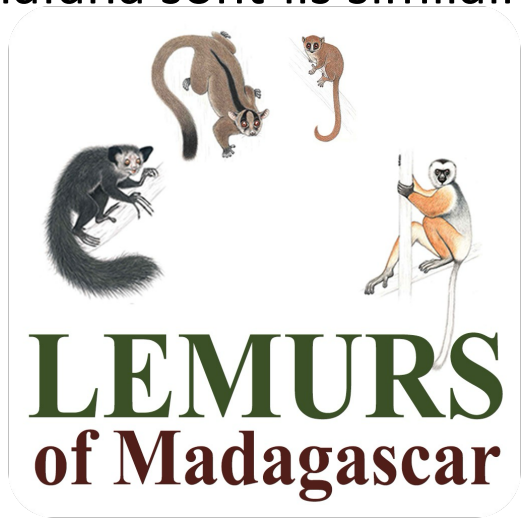
Lemurs of Ranomafana national park - TUTORIAL

- Cytochrome B
  - Used a lot in species identification, limited variability within and much greater variation between species
- Prompt: how similar are the lemurs that can be found in Ranomafana National Park to each other?



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 ?



Final checkpoint

- Phylogenetic trees has a wide range of applications to many research topics!
- Uses modeling and stats “behind the scenes”
- **Can anyone give an example of how they could use a phylogeny for their own research? ☺**

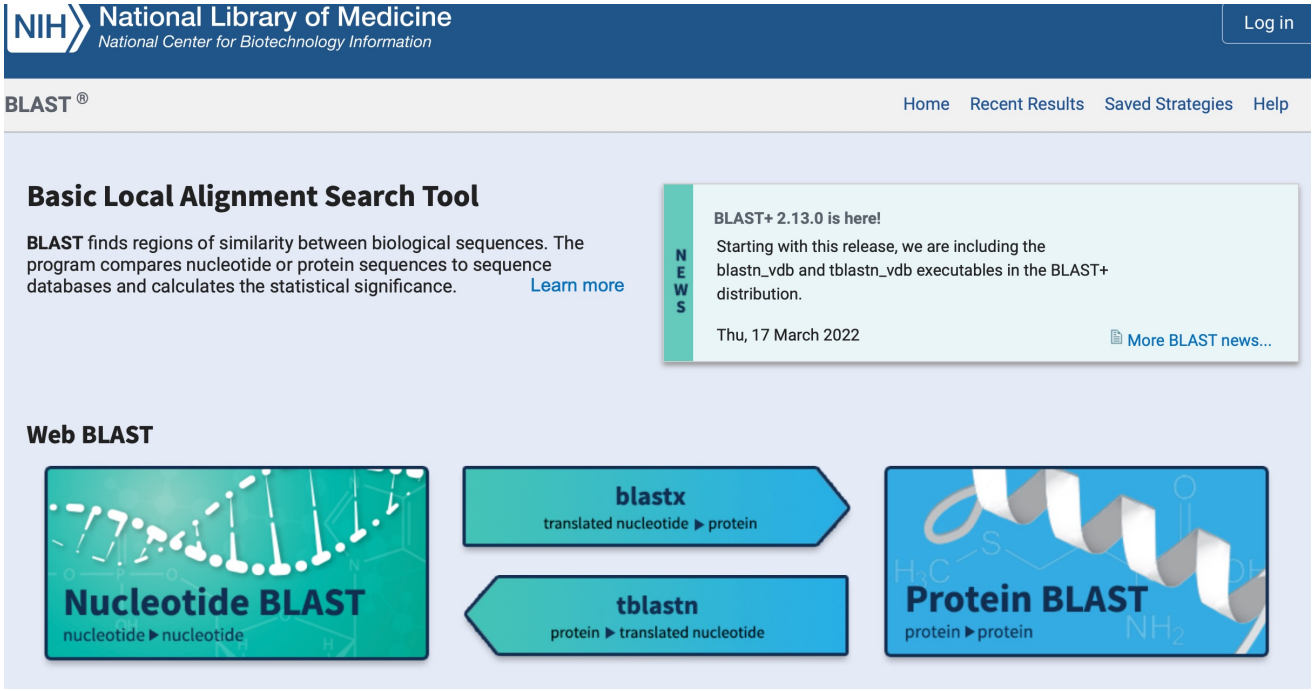
Point de contrôle final!

- 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?** ☺

SLIDES TO REVISIT ON YOUR OWN TIME

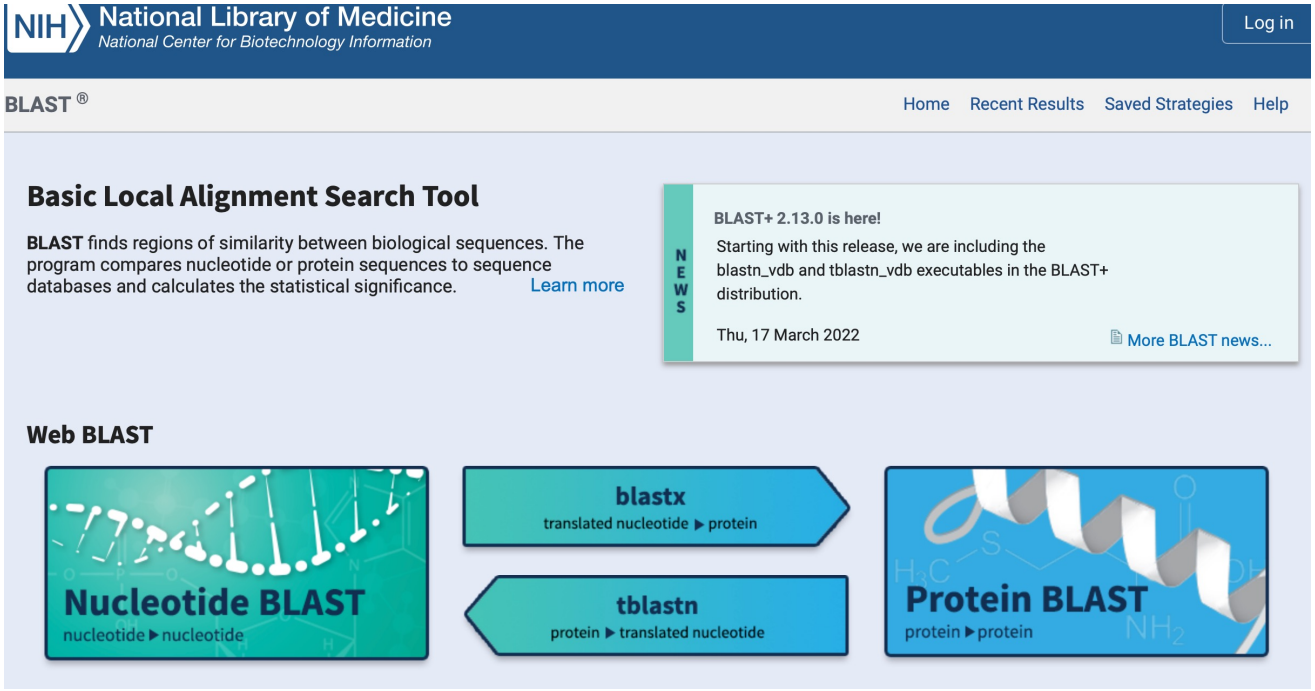
DIAPOSITIVES À REVOIR À VOTRE RYTHME

Steps to revisit later

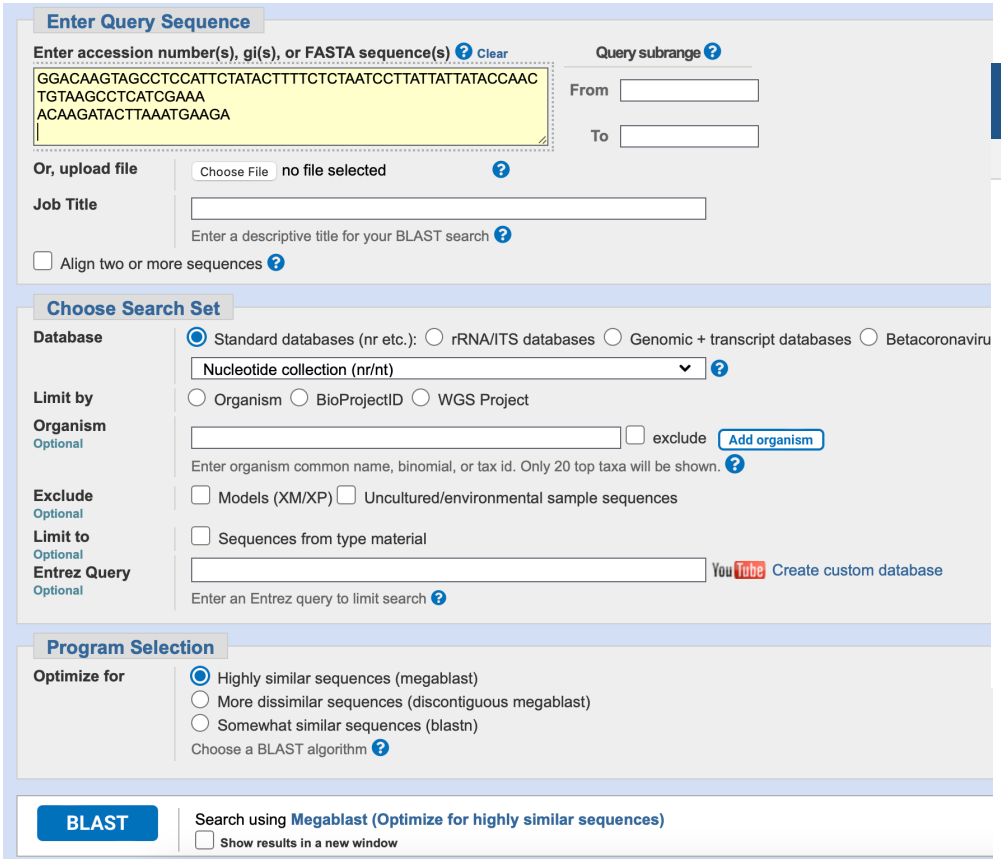


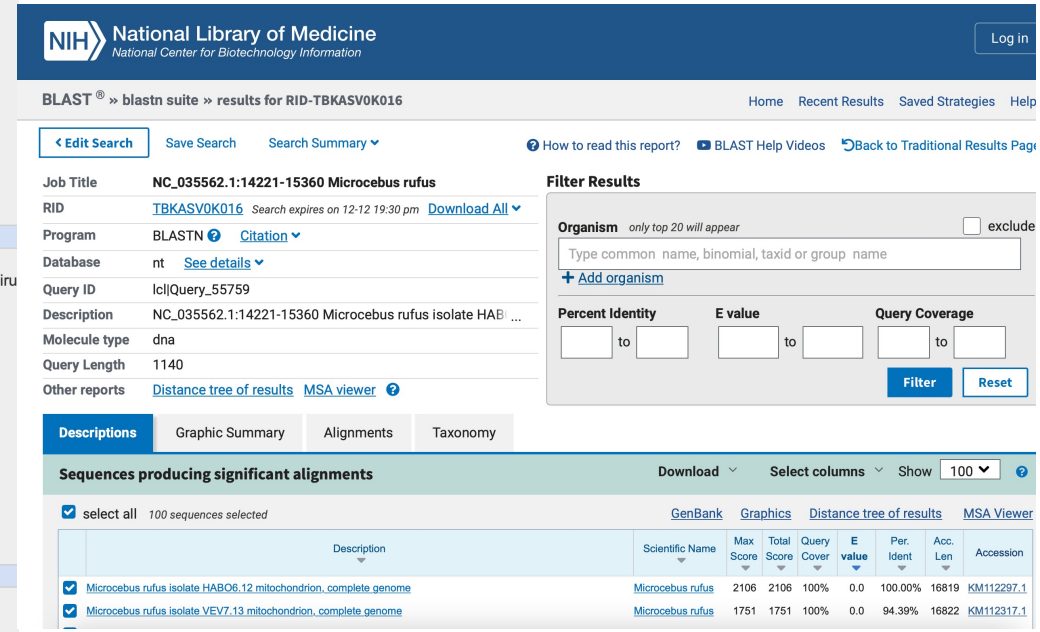
If you have a new sequence and need to figure out what it is…go to BLAST

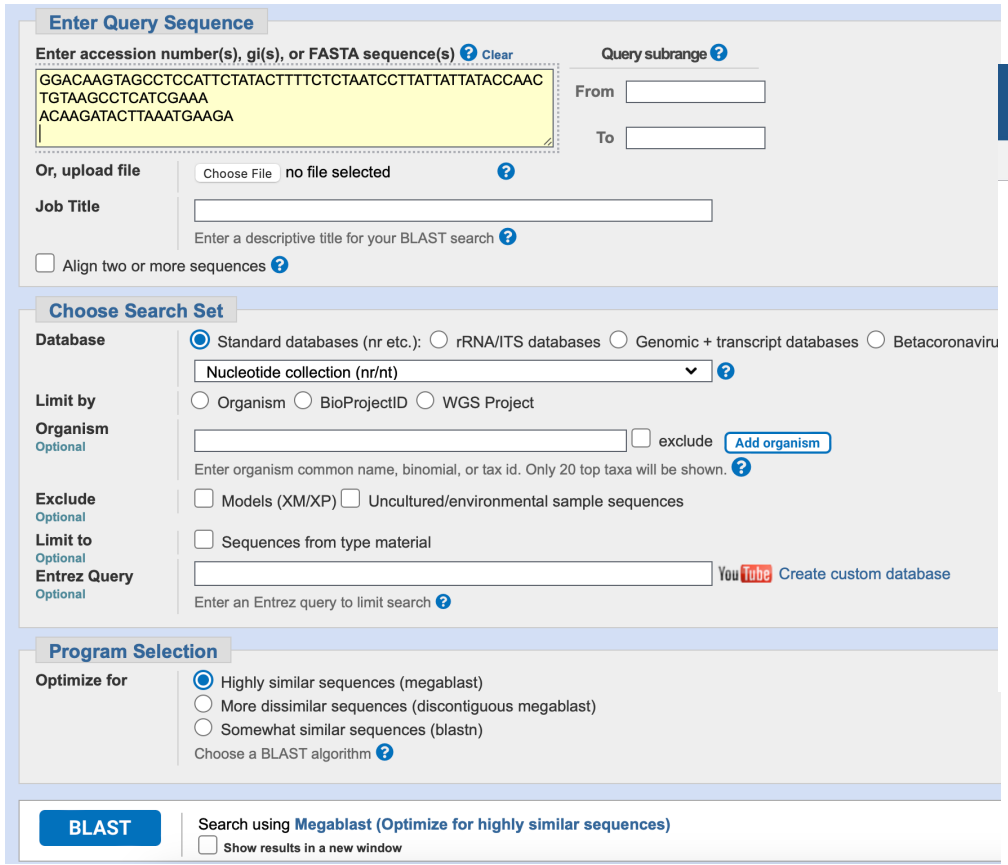
Étapes à revoir plus tard

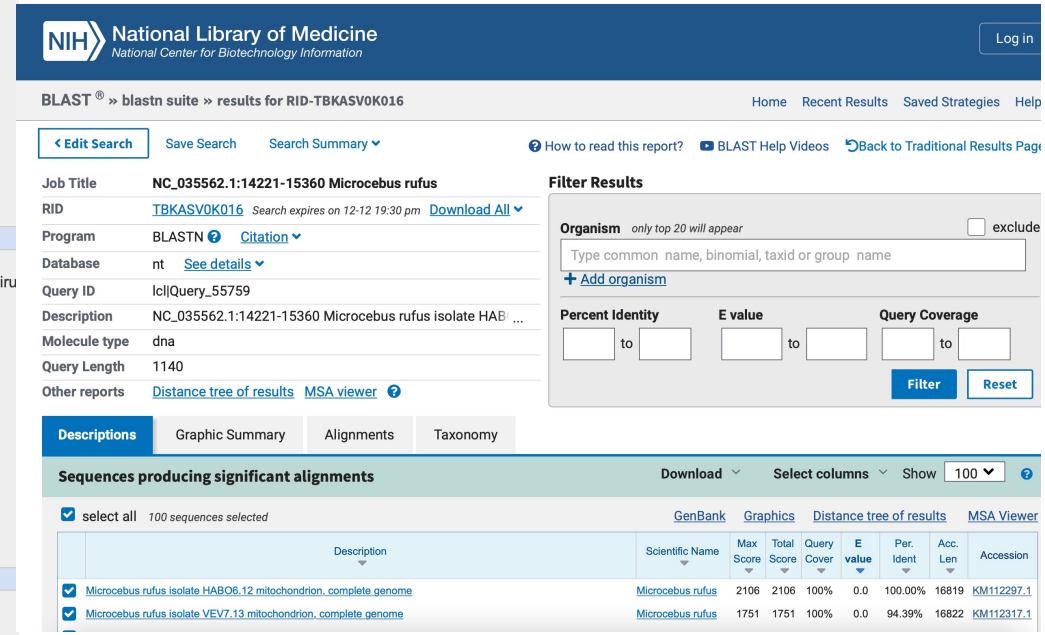


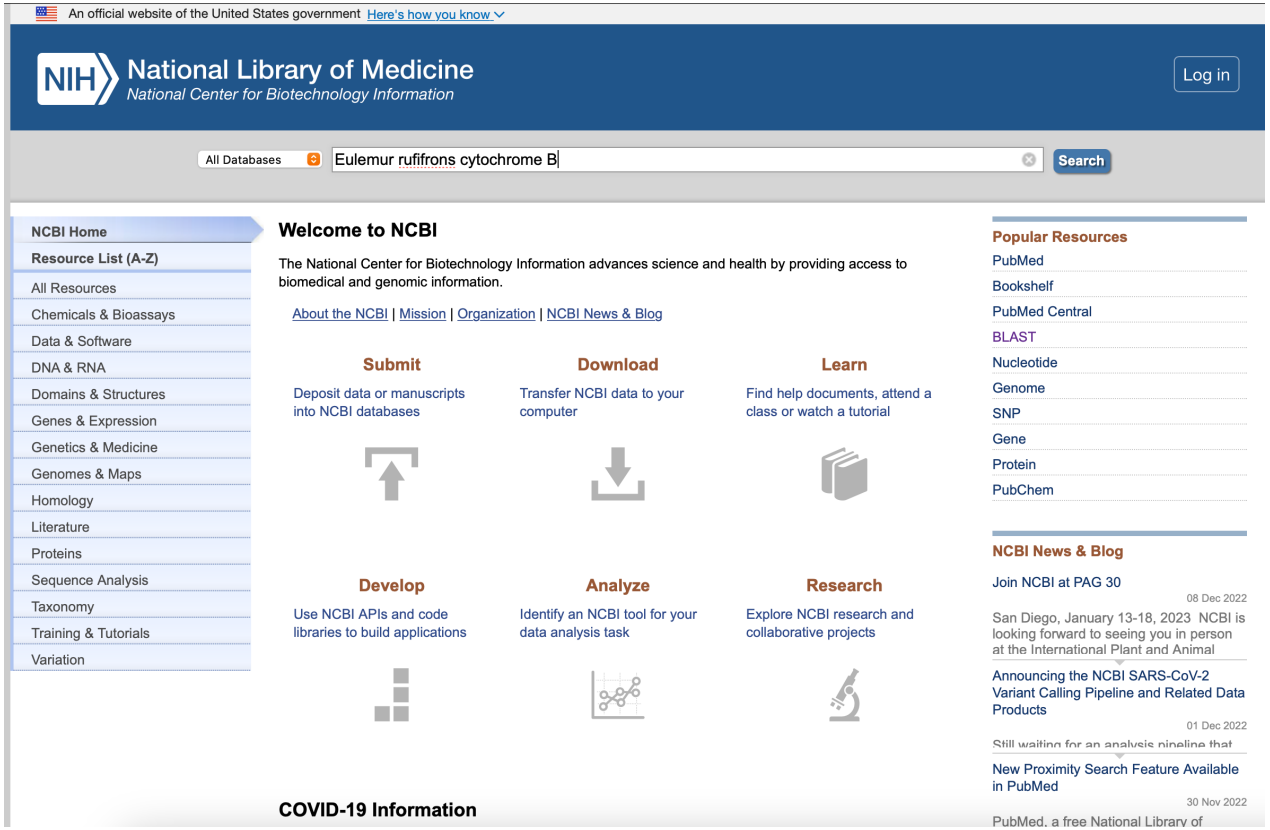
Vérifiez à quel type de séquence vous avez affaire en effectuant une recherche BLAST


Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
GGACAAGTAGCCTCCATTCTATACTTTTCTCTAATCCTTATTATTATACCAAC
TGTAAGCCTCATCGAAA
ACAAGATACTTAAATGAAGA
Query subrange
From
To
Or, upload file
Choose 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
Optional
exclude
Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Limit to
Optional
Sequences from type material
Entrez Query
Optional
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

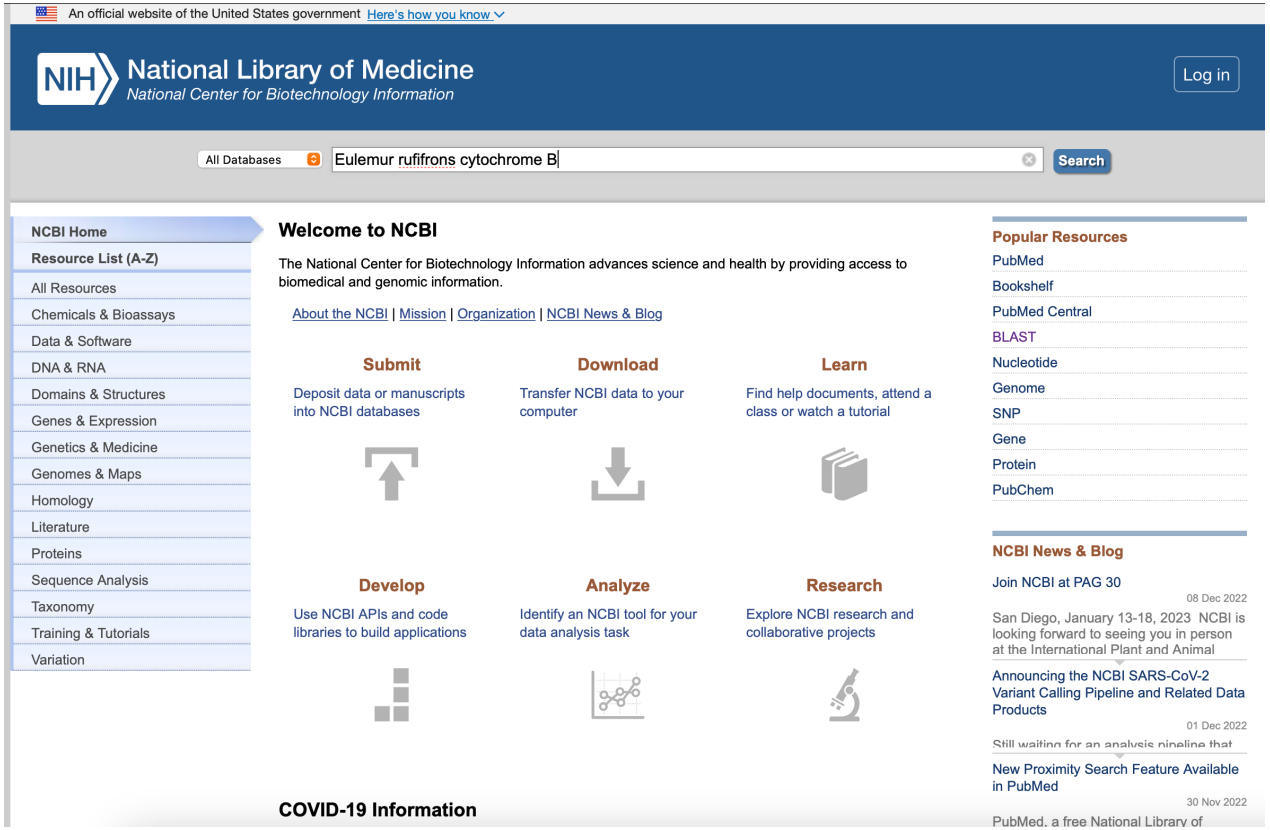

National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST ® » blastn suite » results for RID-TBKASV0K016
Home
Recent Results
Saved Strategies
Help
Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title
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
lcl|Query_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
Type common name, binomial, taxid or group name
Add organism
Percent Identity
E value
Query Coverage
to
Filter
Reset
Descriptions
Graphic Summary
Alignments
Taxonomy
Sequences producing significant alignments
Download
Select columns
Show
100
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
Microcebus rufus isolate HABO6.12 mitochondrion, complete genome
Microcebus rufus
2106
2106
100%
0.0
100.00%
16819
KM112297.1
Microcebus rufus isolate VEV7.13 mitochondrion, complete genome
Microcebus rufus
1751
1751
100%
0.0
94.39%
16822
KM112317.1


Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) Clear
GGACAAGTAGCCTCCATTCTATACTTTTCTCTAATCCTTATTATTATACCAAC
TGTAAGCCTCATCGAAA
ACAAGATACTTAAATGAAGA
Query subrange
From
To
Or, upload file
Choose 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
Optional
exclude
Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude
Optional
Models (XM/XP) Uncultured/environmental sample sequences
Limit to
Optional
Sequences from type material
Entrez Query
Optional
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

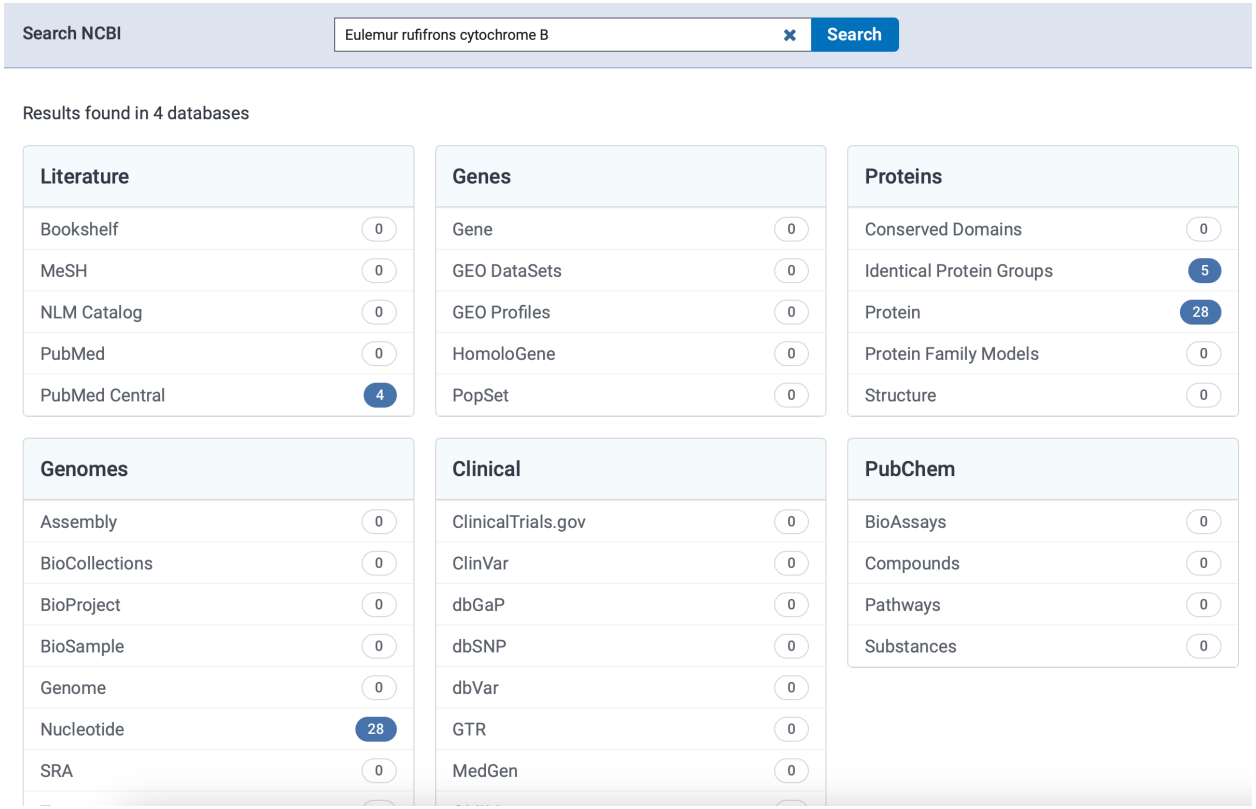

National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST ® » blastn suite » results for RID-TBKASV0K016
Home Recent Results Saved Strategies Help
Edit Search
Save Search
Search Summary
How to read this report?
BLAST Help Videos
Back to Traditional Results Page
Job Title 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 lcl|Query_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
Type common name, binomial, taxid or group name
Add organism
Percent Identity
E value
Query Coverage
to
Filter
Reset
Descriptions
Graphic Summary
Alignments
Taxonomy
Sequences producing significant alignments
Download
Select columns
Show 100
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
Microcebus rufus isolate HABO6.12 mitochondrion, complete genome
Microcebus rufus
2106 2106 100% 0.0 100.00% 16819 KM112297.1
Microcebus rufus isolate VEV7.13 mitochondrion, complete genome
Microcebus rufus
1751 1751 100% 0.0 94.39% 16822 KM112317.1



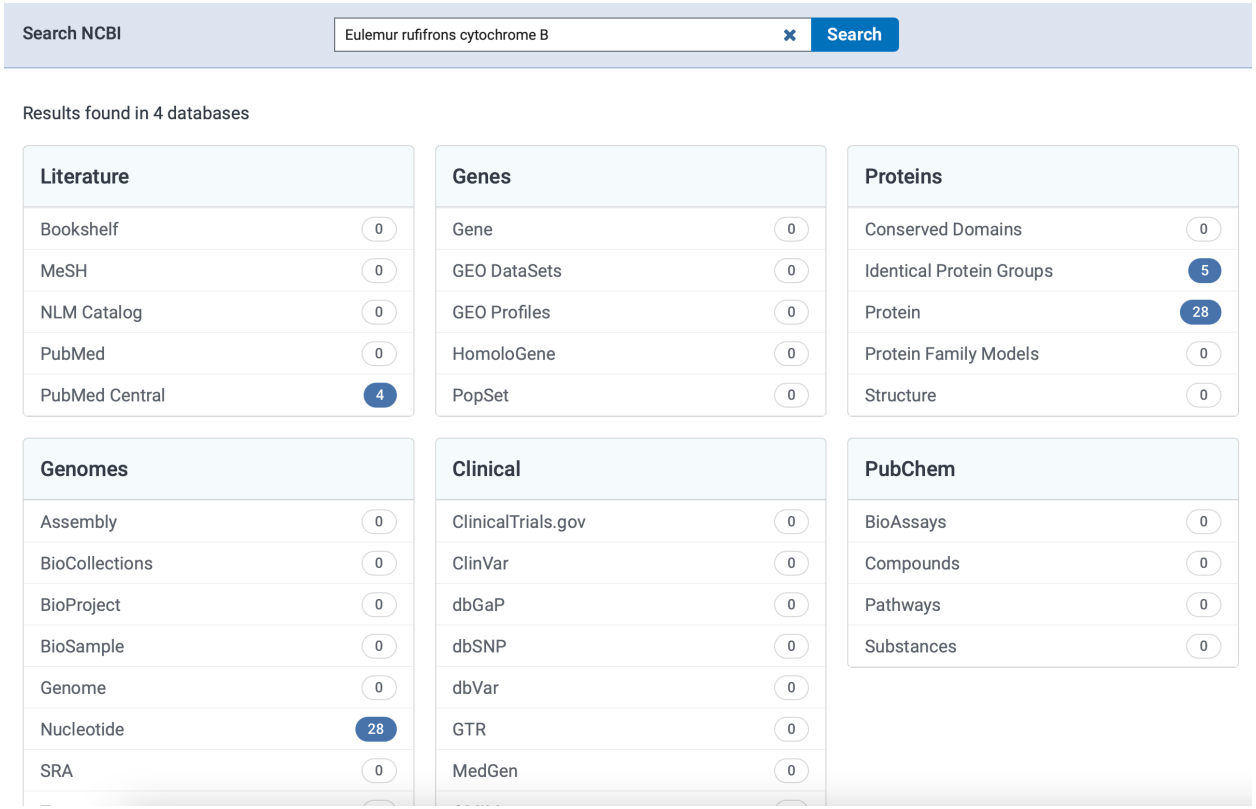
1. Go to NCBI, and search for the thing you want to build a phylogeny for, in our case cytochrome B of lemurs in Ranomafana national park



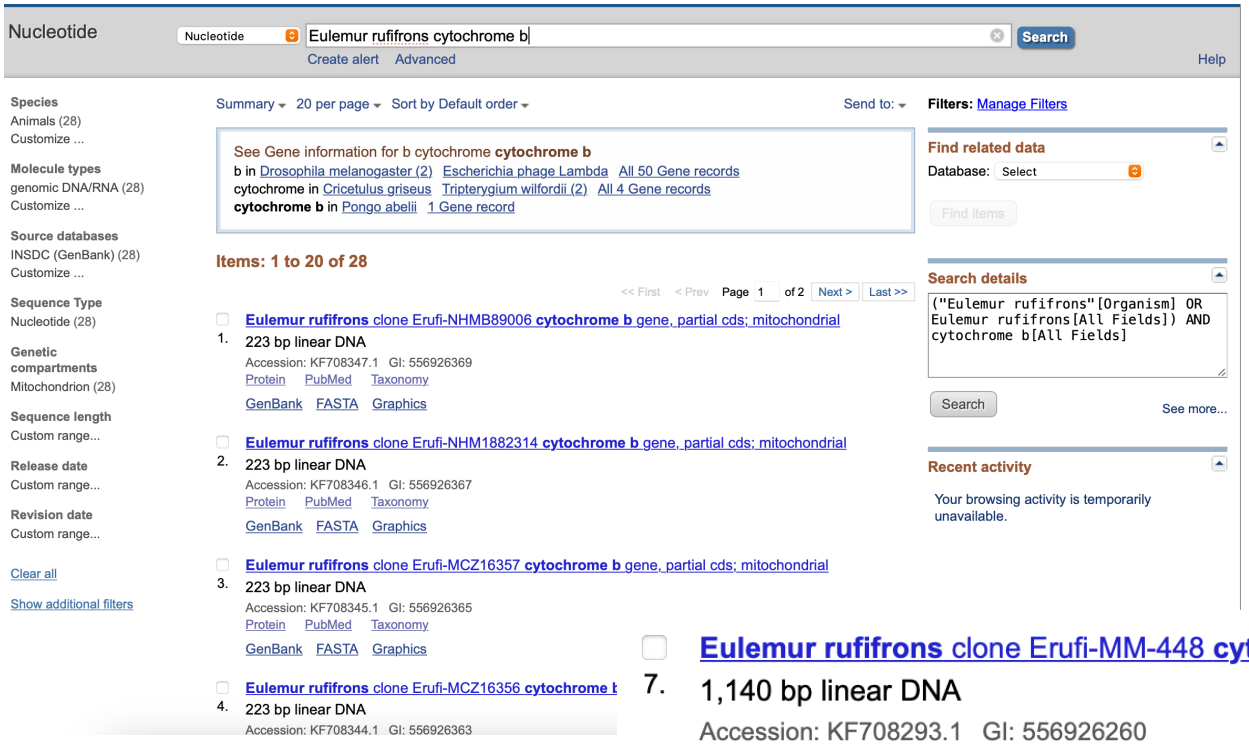
Allez au NCBI, et recherchez la chose pour laquelle vous voulez construire une phylogénie, dans notre cas le cytochrome B des lémuriens dans le parc national de Ranomafana

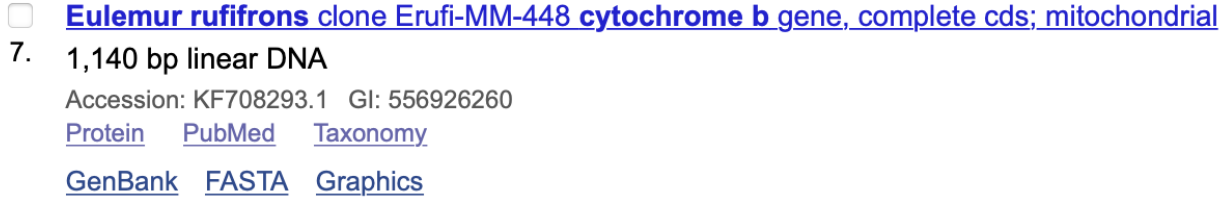


This is what it will look like, you can go to Nucleotide under the genome category and click on that



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



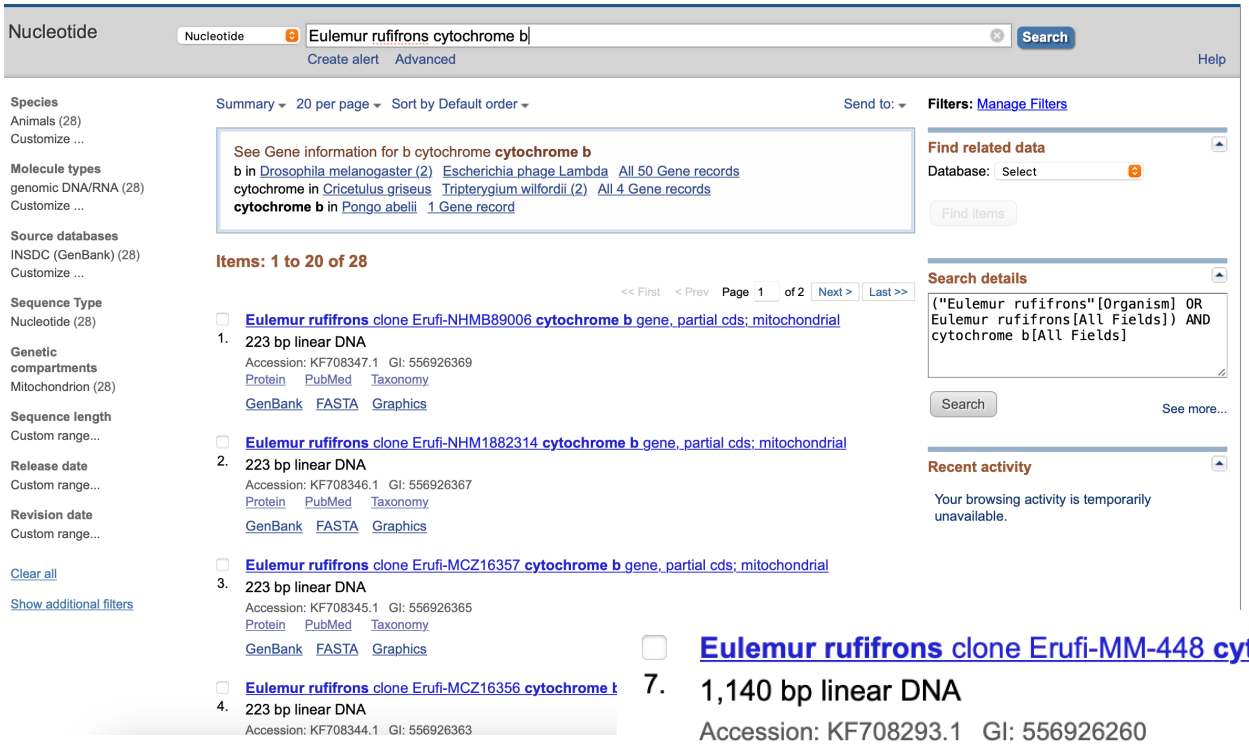


Pick the sequence of what you're interested in, in our case we want a complete cds

We might want partial cds if we have a partial sequence of interest, but right now we're just building a tree with known data, so complete cds is best

Cds: protein coding sequence

Then download the fastas

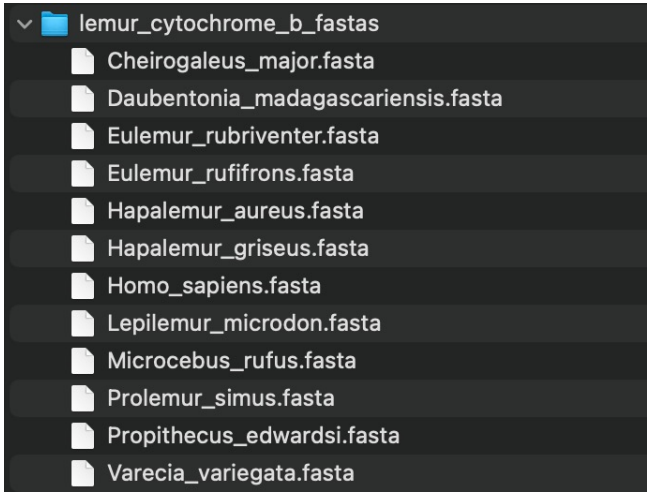


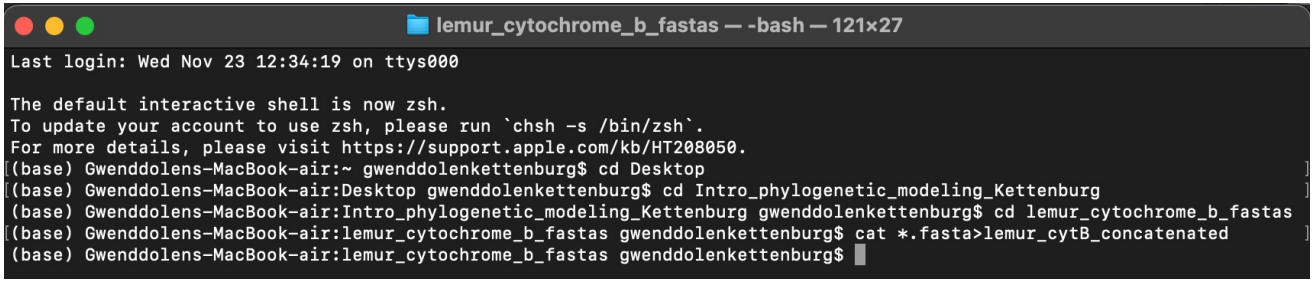
Choisissez la séquence de ce qui vous intéresse, dans notre cas nous voulons un cds complet

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





Step 2: when you have all your sequences of interest and your outgroup, you need to concatenate the sequences into one file, you can do this by making a text/edit file and pasting each sequence in, otherwise follow instructions on command line (mac) or powershell (windows) to do this

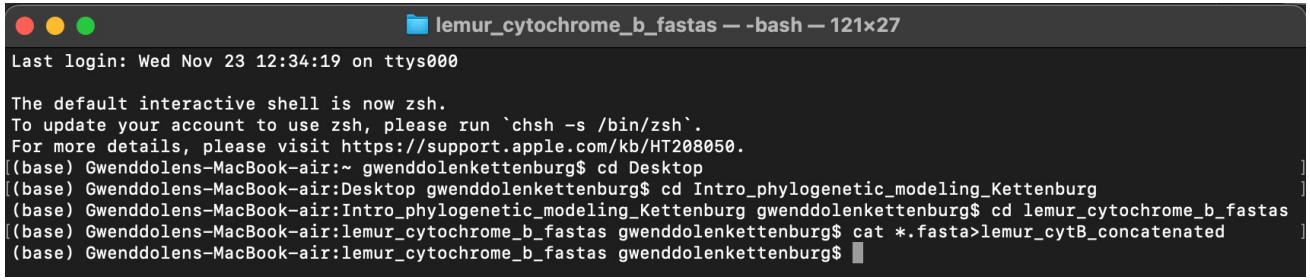
Example 1: Merge with file names (This will merge file1.csv & file2.csv to create concat.csv)

```
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
```



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

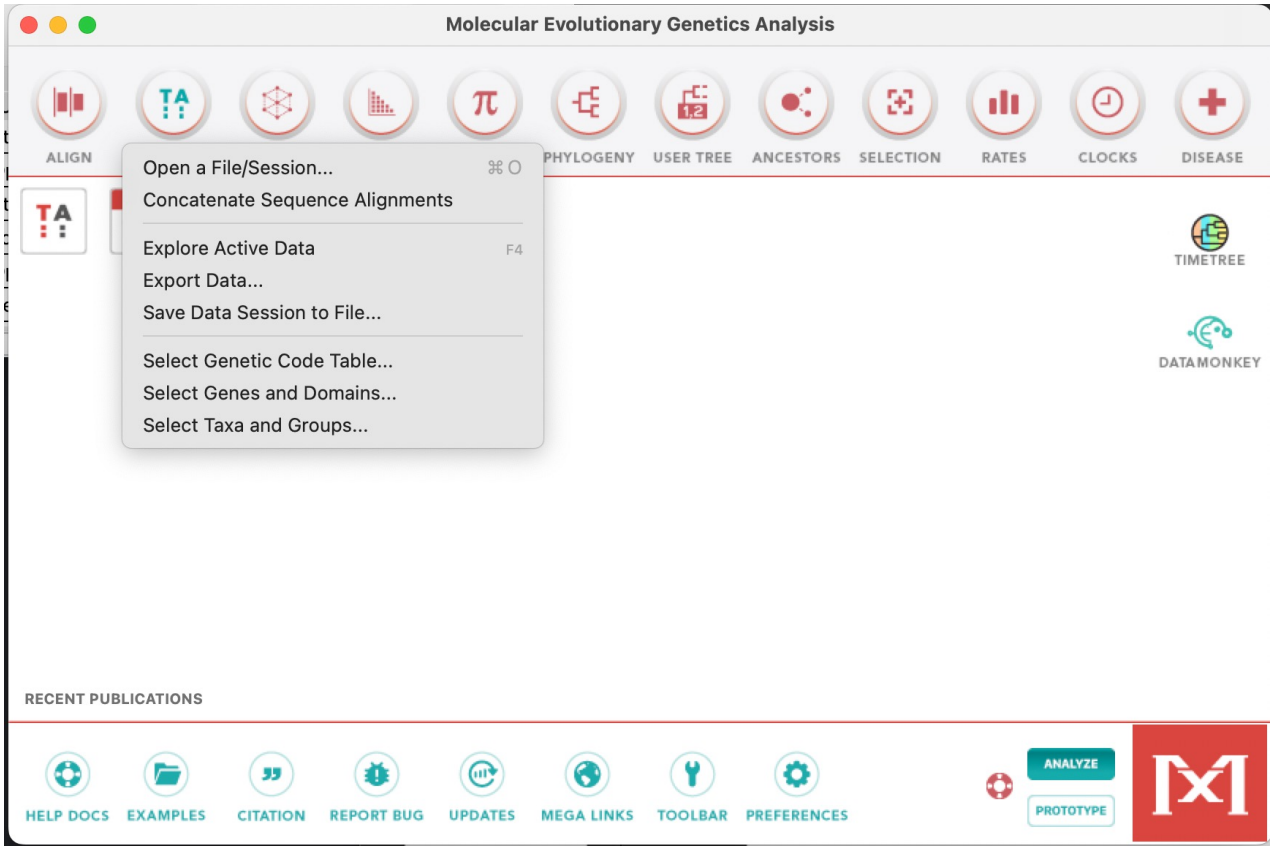
Example 1: Merge with file names (This will merge file1.csv & file2.csv to create concat.csv)

```
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

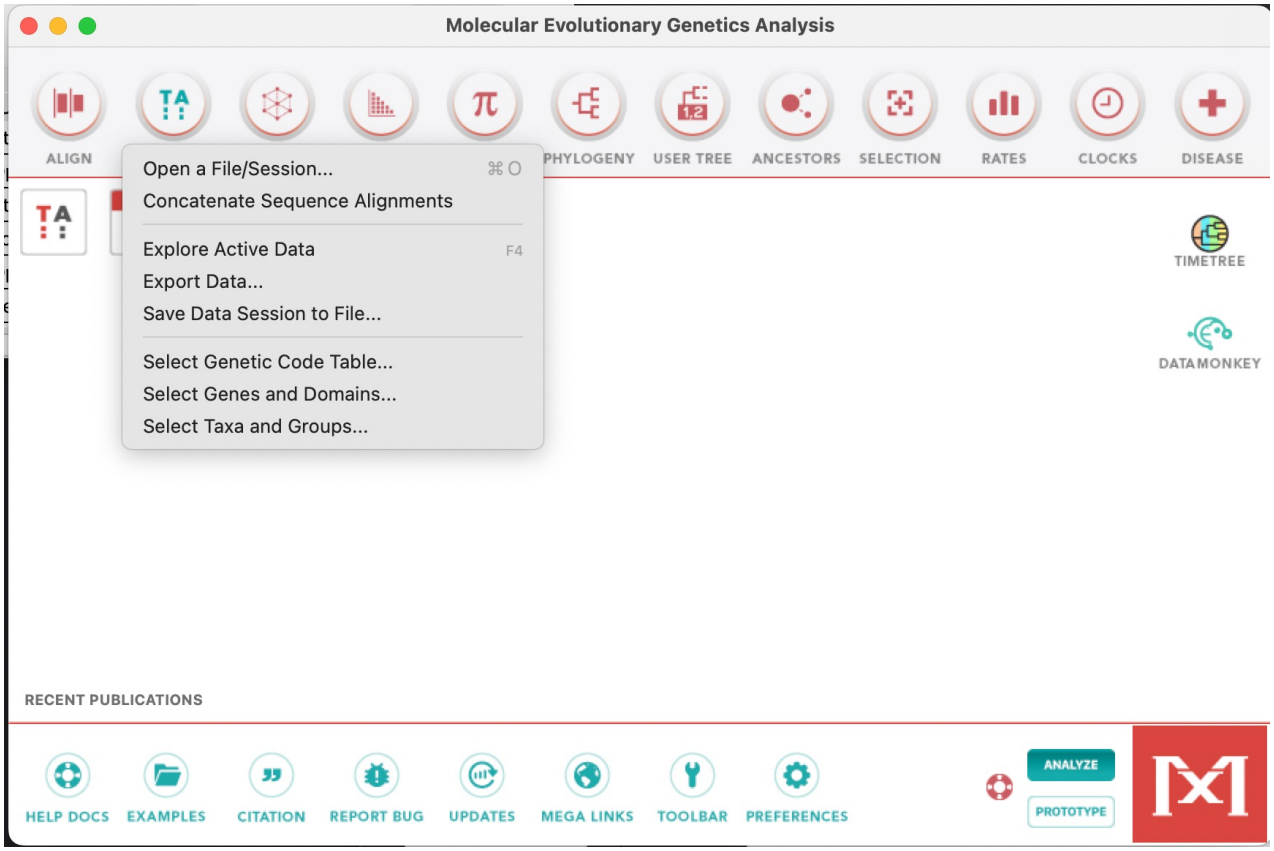
```
type  *.csv > concat_csv.txt
```



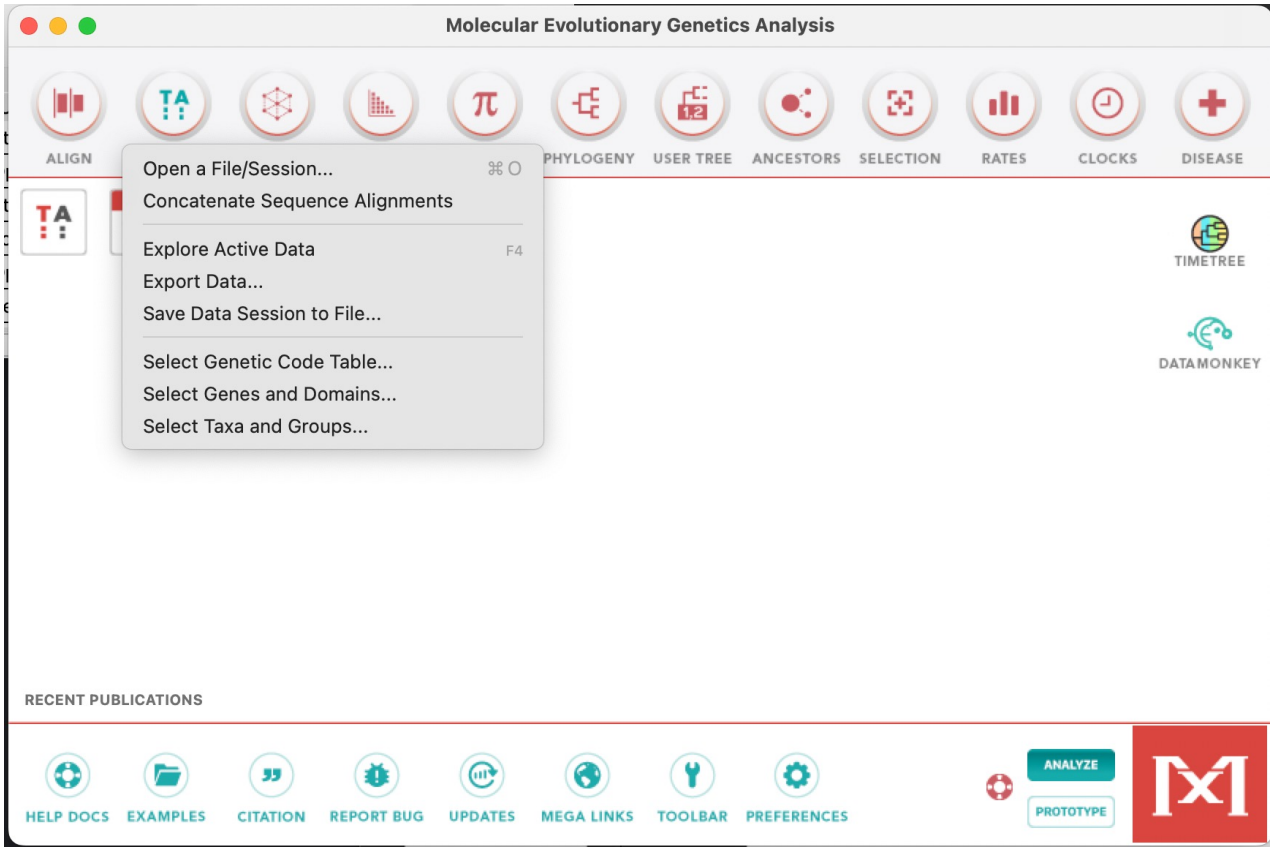
We need to align our sequences, so the analysis is comparing sequences in the same region of the genome

1. Open MEGA, and open a file/session, select your concatenated fasta file

2. MEGA will ask if you want to align or analyze, click on align

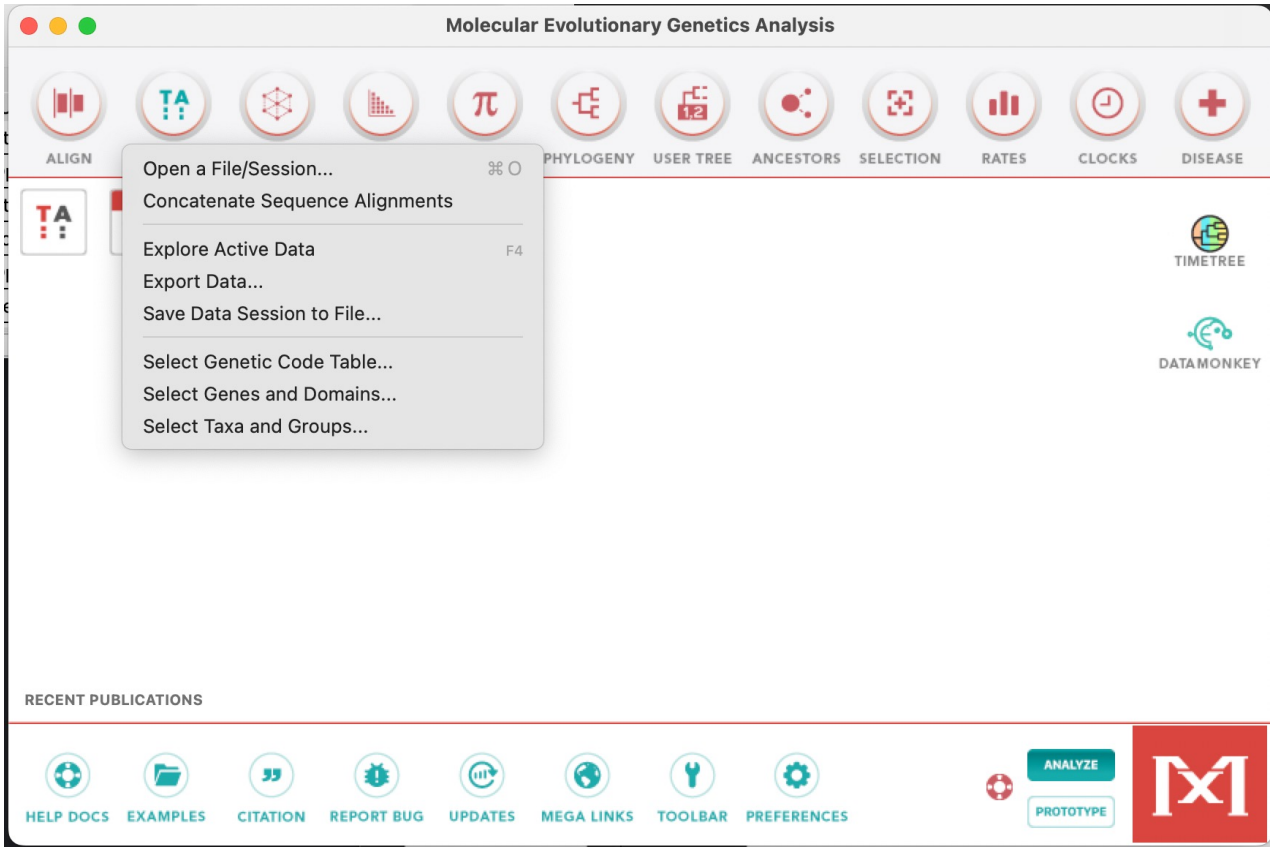


Ouvrez MEGA, et ouvrez un fichier/session, sélectionnez votre fichier fasta concaténé MEGA vous demandera si vous souhaitez aligner ou analyser, cliquez sur aligner



Step 3: open MEGA, and open a file/session, select your concatenated fasta file

MEGA will ask if you want to align or analyze, click on align



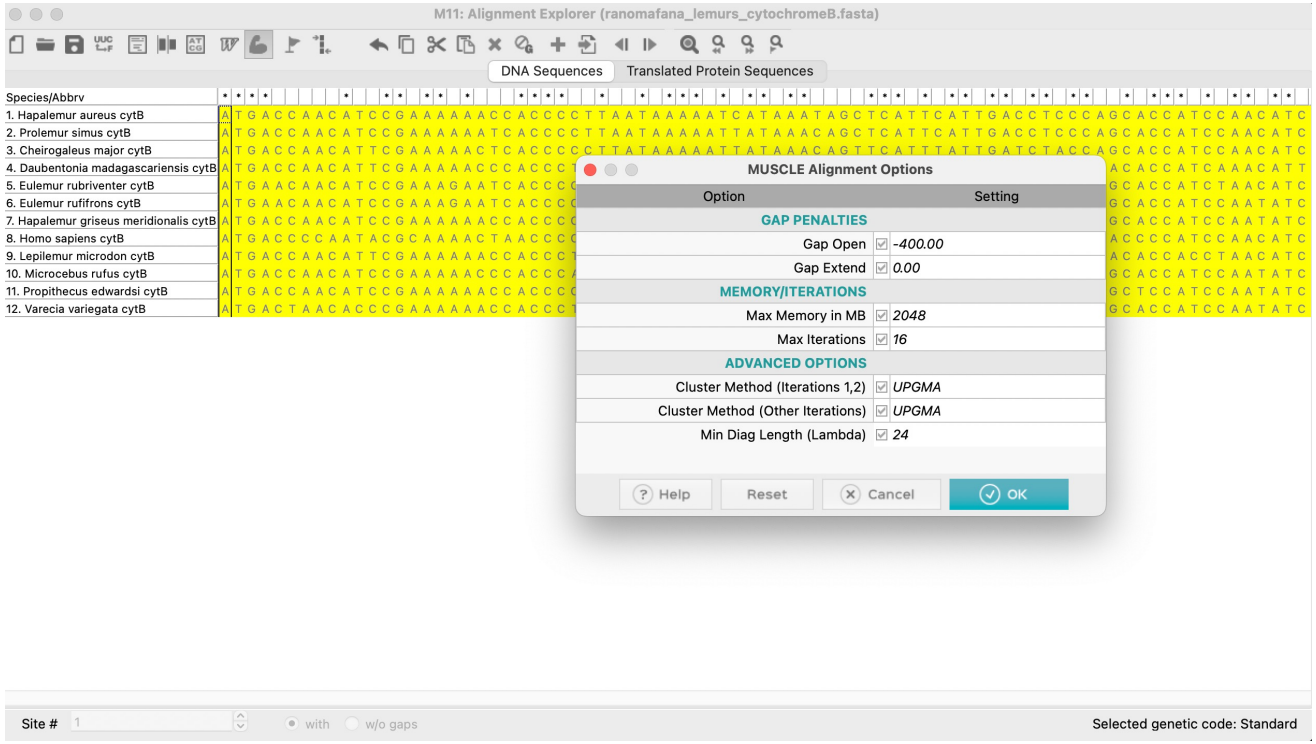
Ouvrez MEGA, et ouvrez un fichier/session, sélectionnez votre fichier fasta concaténé MEGA vous demandera si vous souhaitez aligner ou analyser, cliquez sur aligner

So the sequences are loaded into MEGA like this:

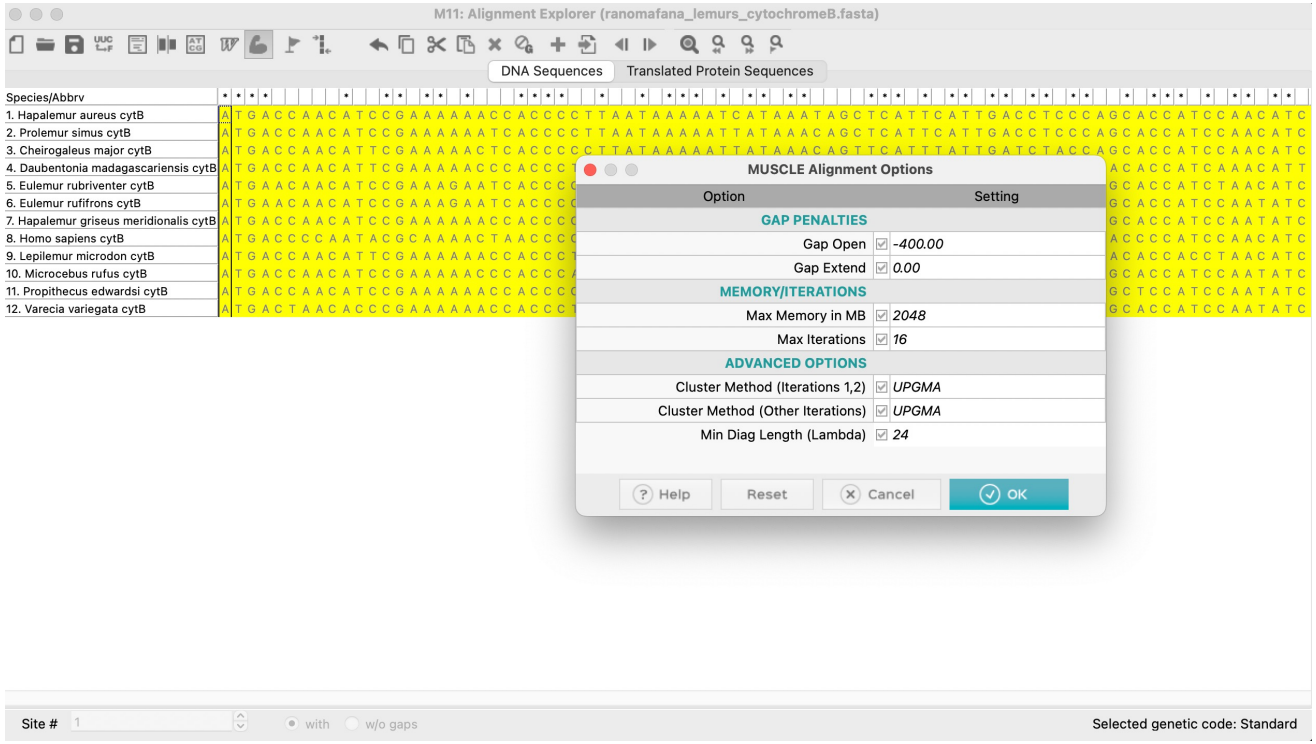
Step 4: click on the muscle arm to align with MUSCLE program

Les séquences sont donc chargées dans MEGA comme ceci :

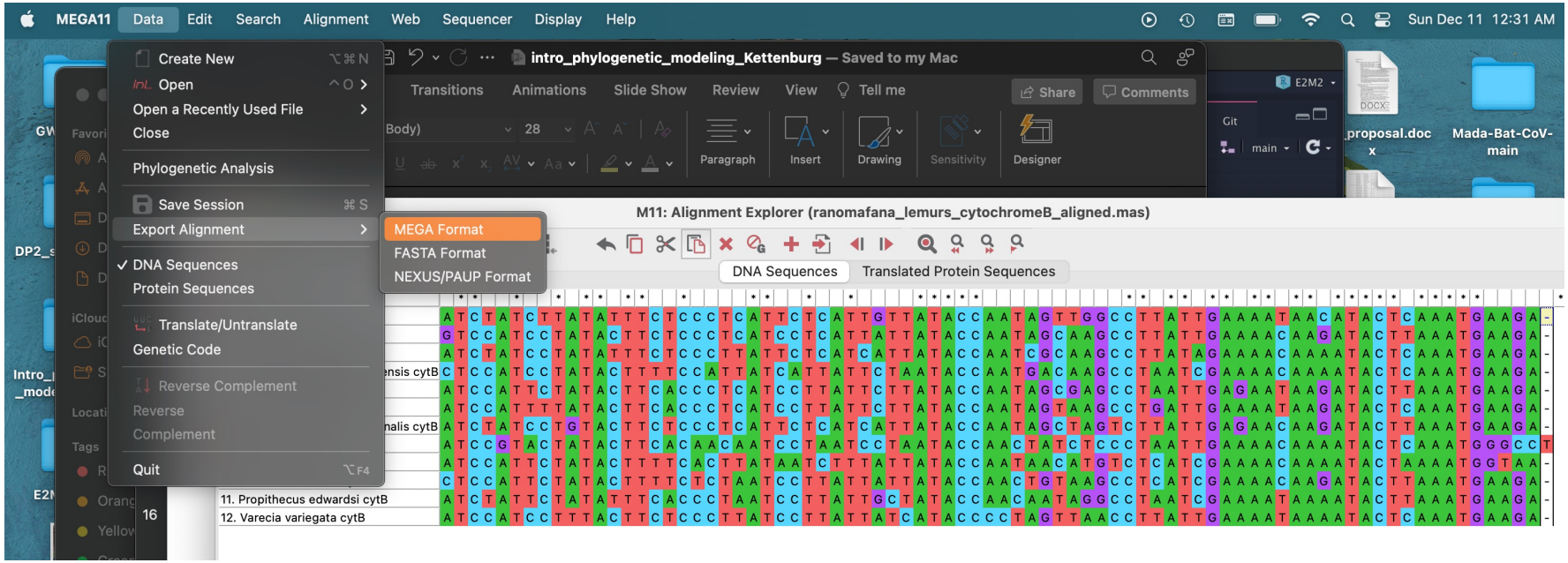
Cliquez sur le bras musculaire pour l'aligner sur le programme MUSCLE



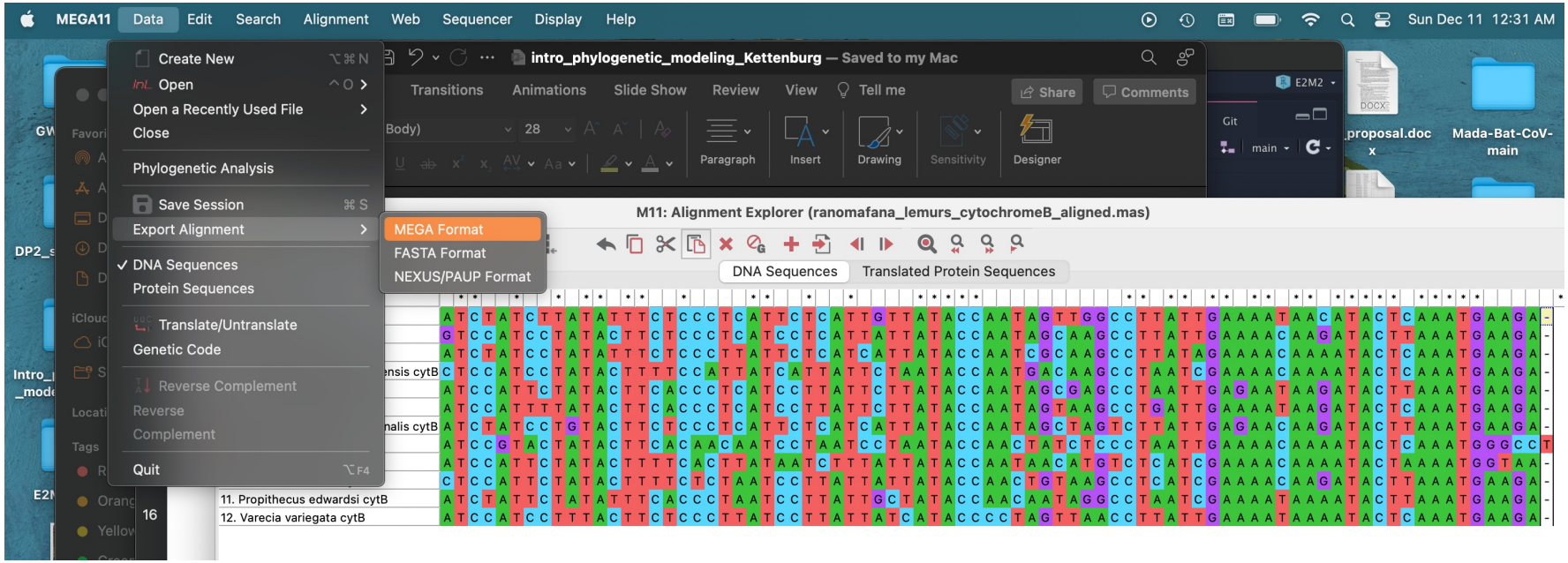
Go with suggested options, then hit okay



Suivez les options suggérées, puis appuyez sur OK

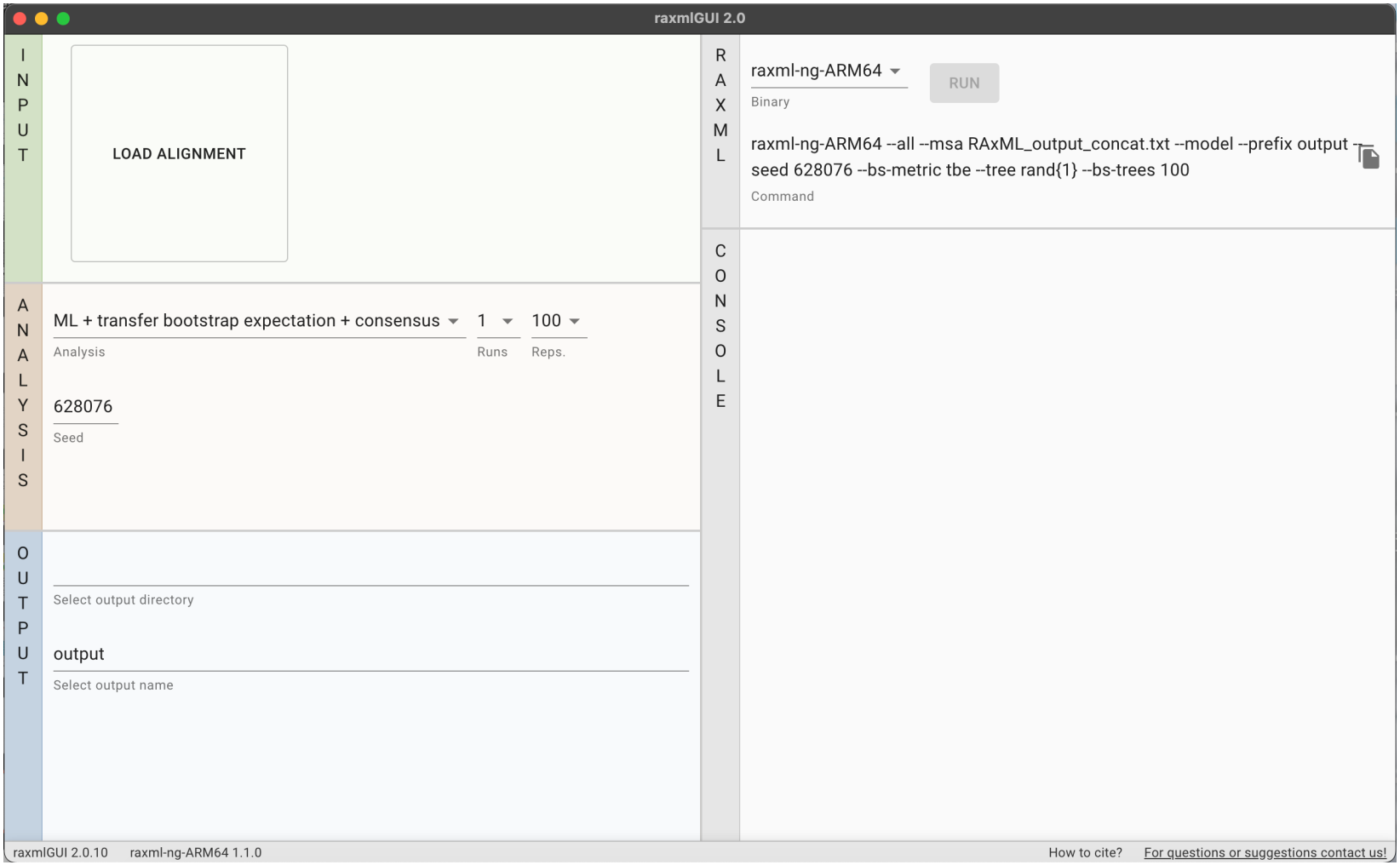


Save the aligned file, then we will proceed to model selection

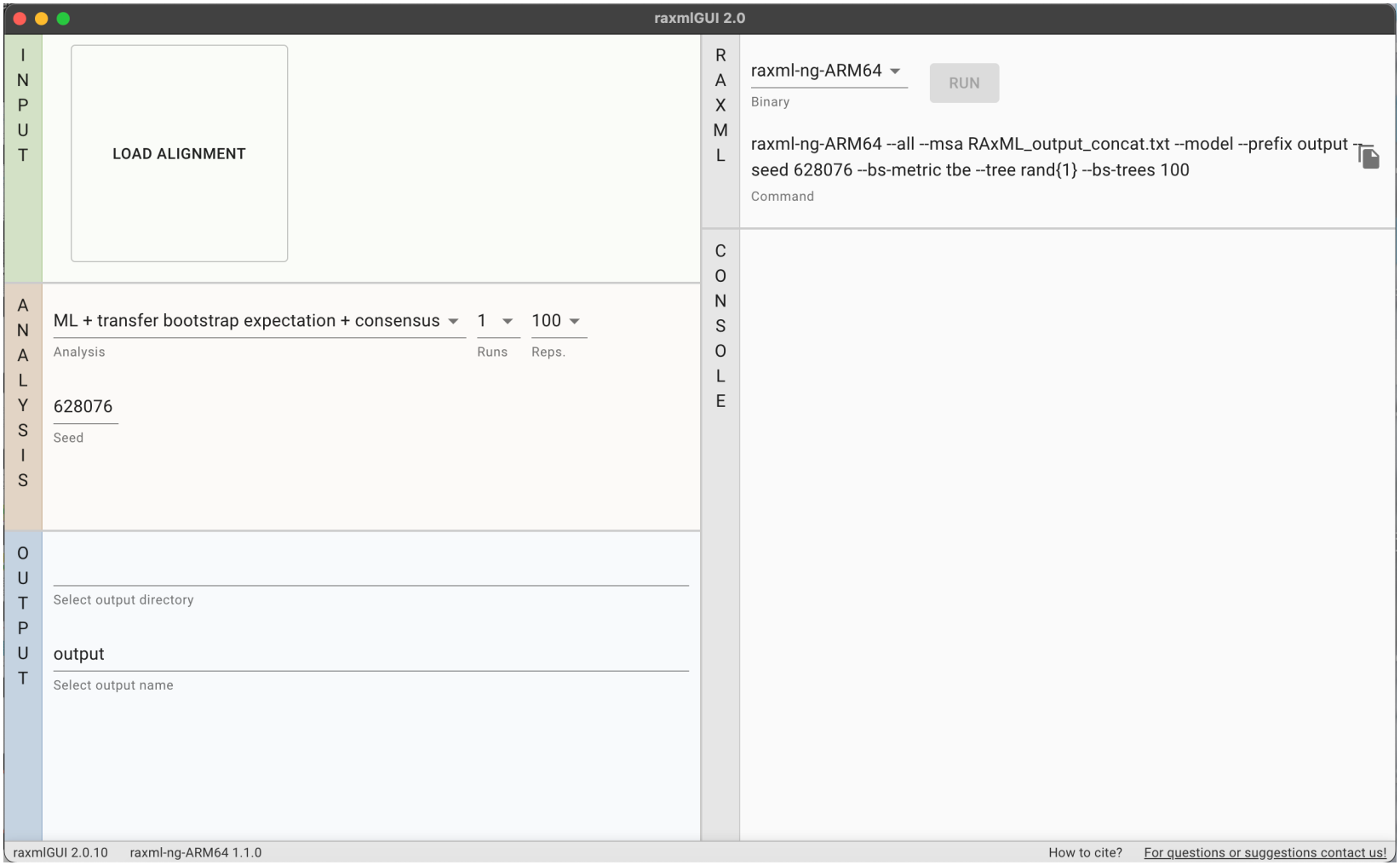


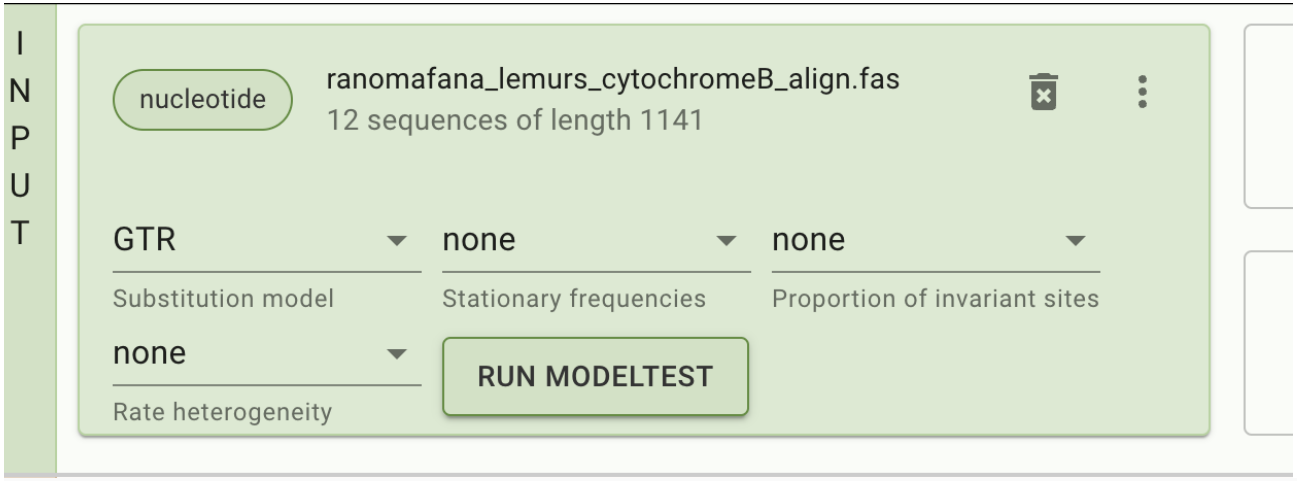
Enregistrez le fichier aligné, puis nous procéderons à la sélection du modèle

Open up RAxML GUI



Ouvrir l’interface graphique RAxML

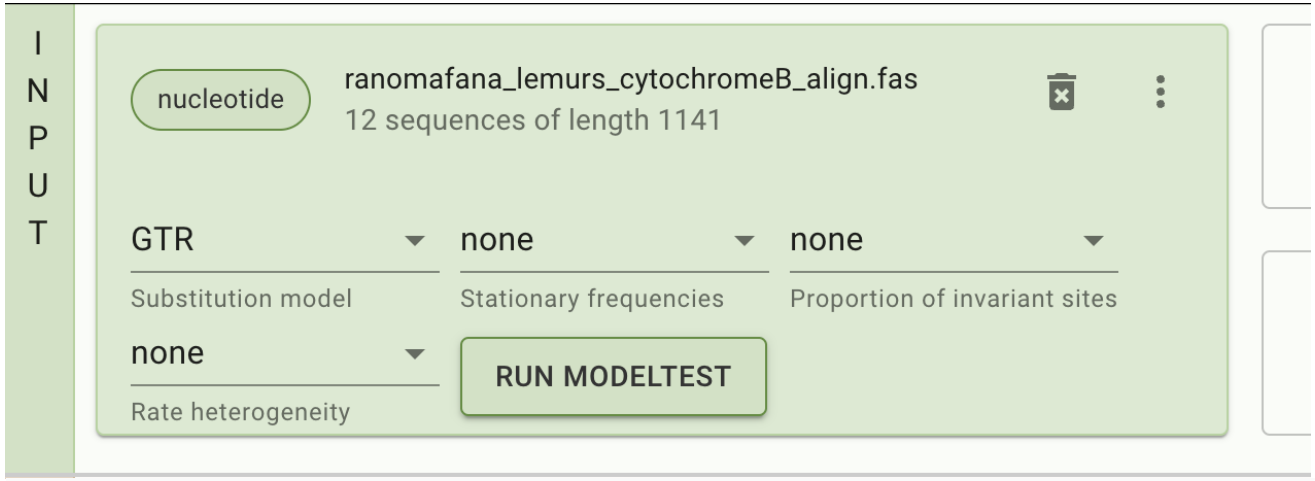




Partition 1/1:

	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

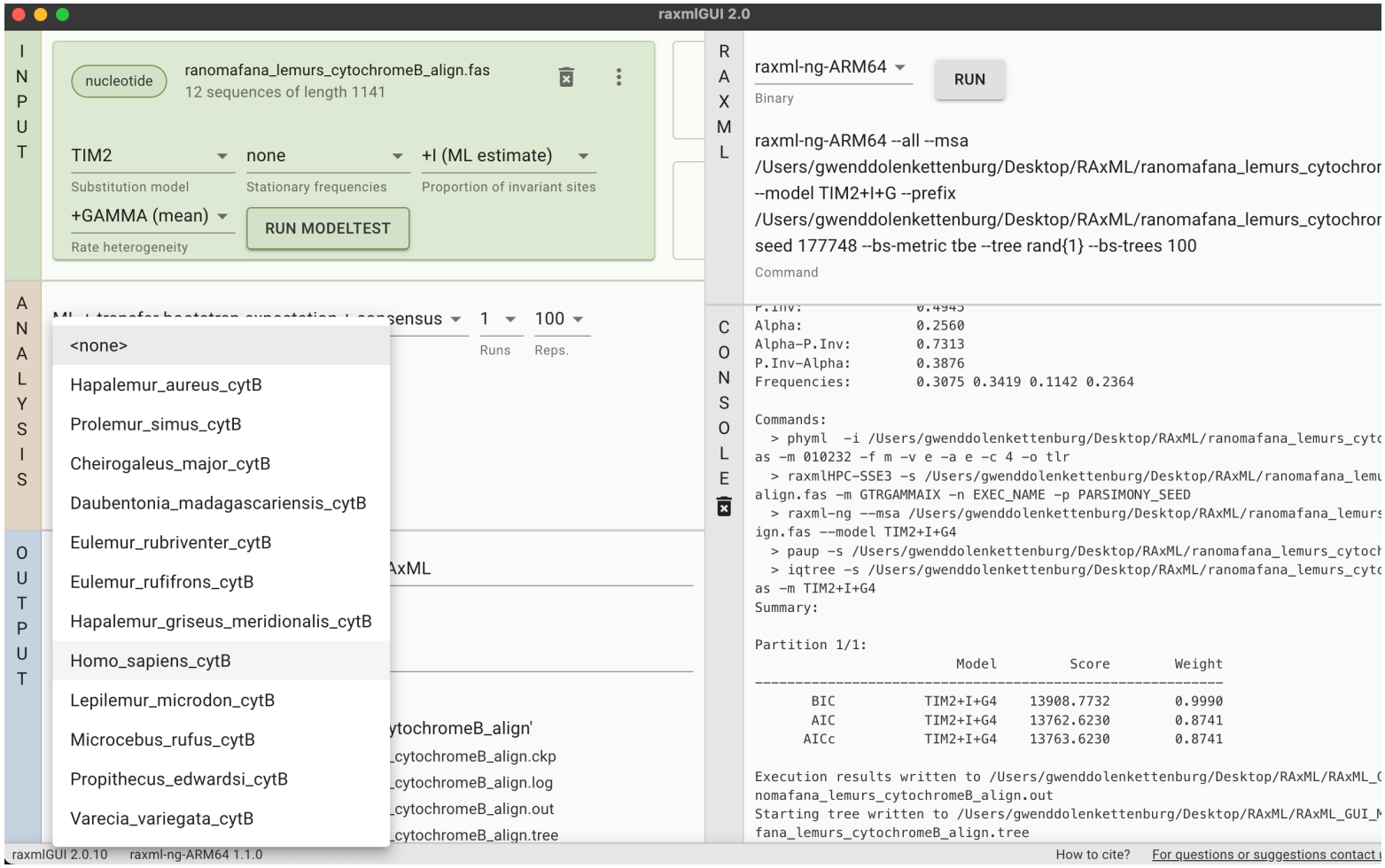
Load the aligned file and perform modeltest, for model selection go with the BIC score, it will spit out a report that saves to your files



Partition 1/1:

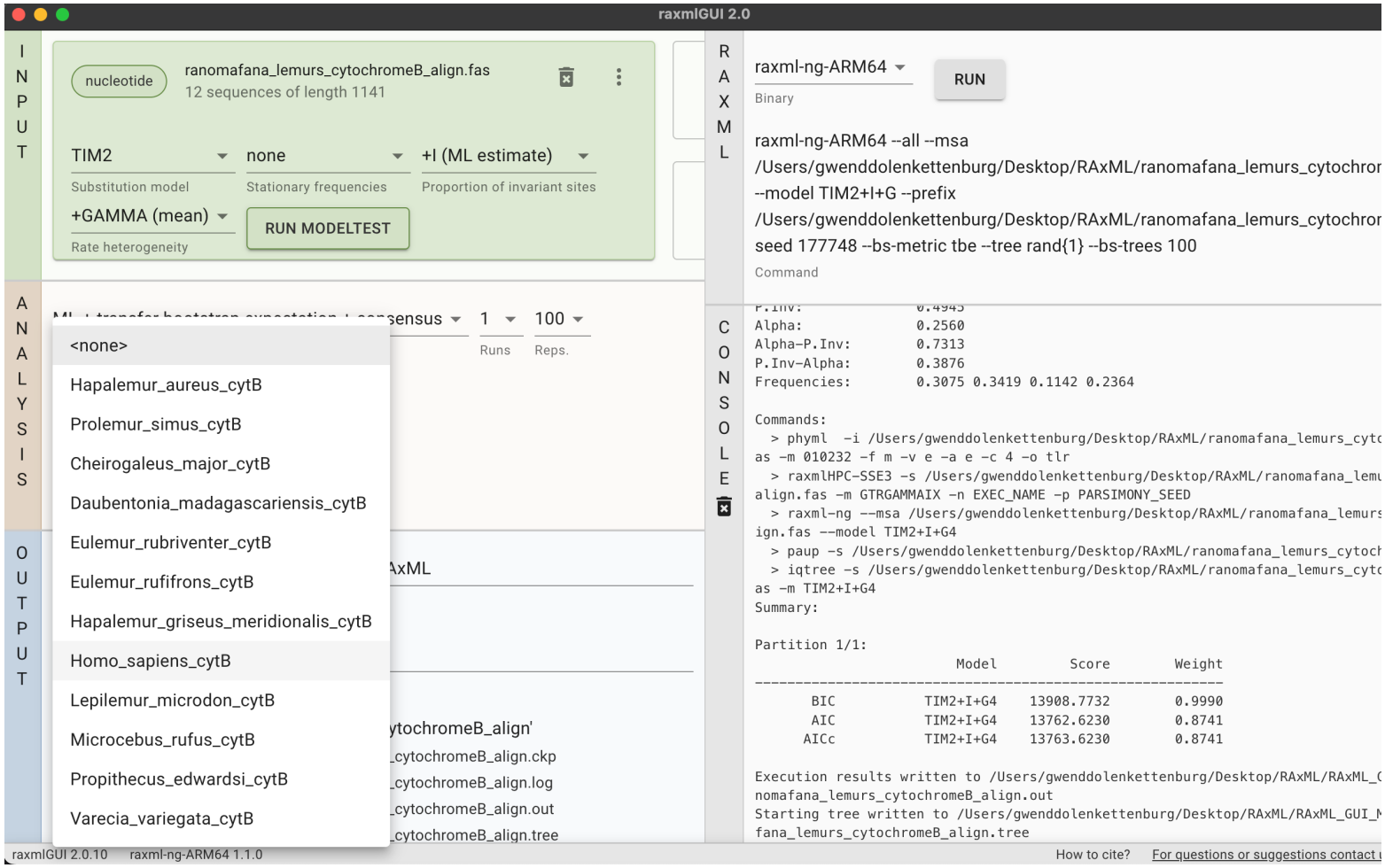
	Model	Score	Weight
BIC	TIM2+I+G4	13908.7732	0.9990
AIC	TIM2+I+G4	13762.6230	0.8741
AICc	TIM2+I+G4	13763.6230	0.8741

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



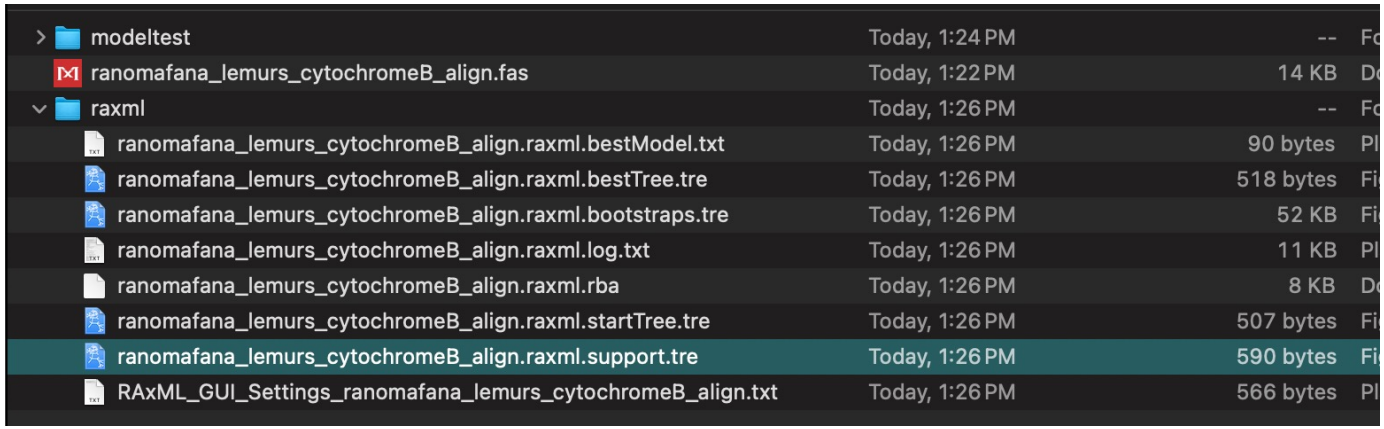
Using the model from modeltest, change as needed in the input box, then in analysis set the outgroup and then hit run in the RAxML section

It will also spit out some files…



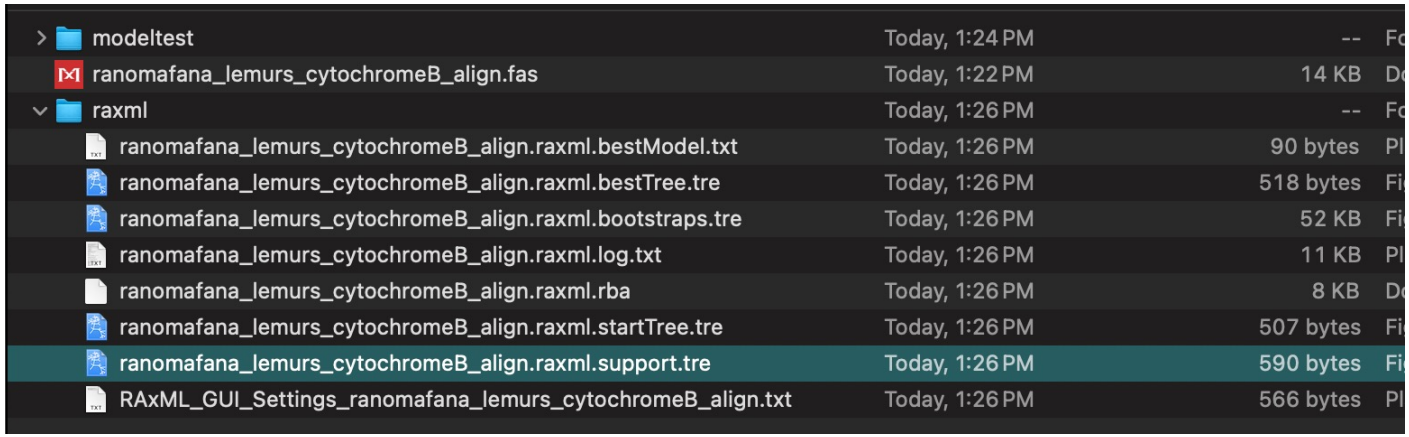
À l'aide du modèle de modeltest, modifiez selon vos besoins dans la zone de saisie, puis dans l'analyse, définissez l'outgroup, puis appuyez sur exécuter dans la section RAxML

Il va aussi recracher quelques fichiers...



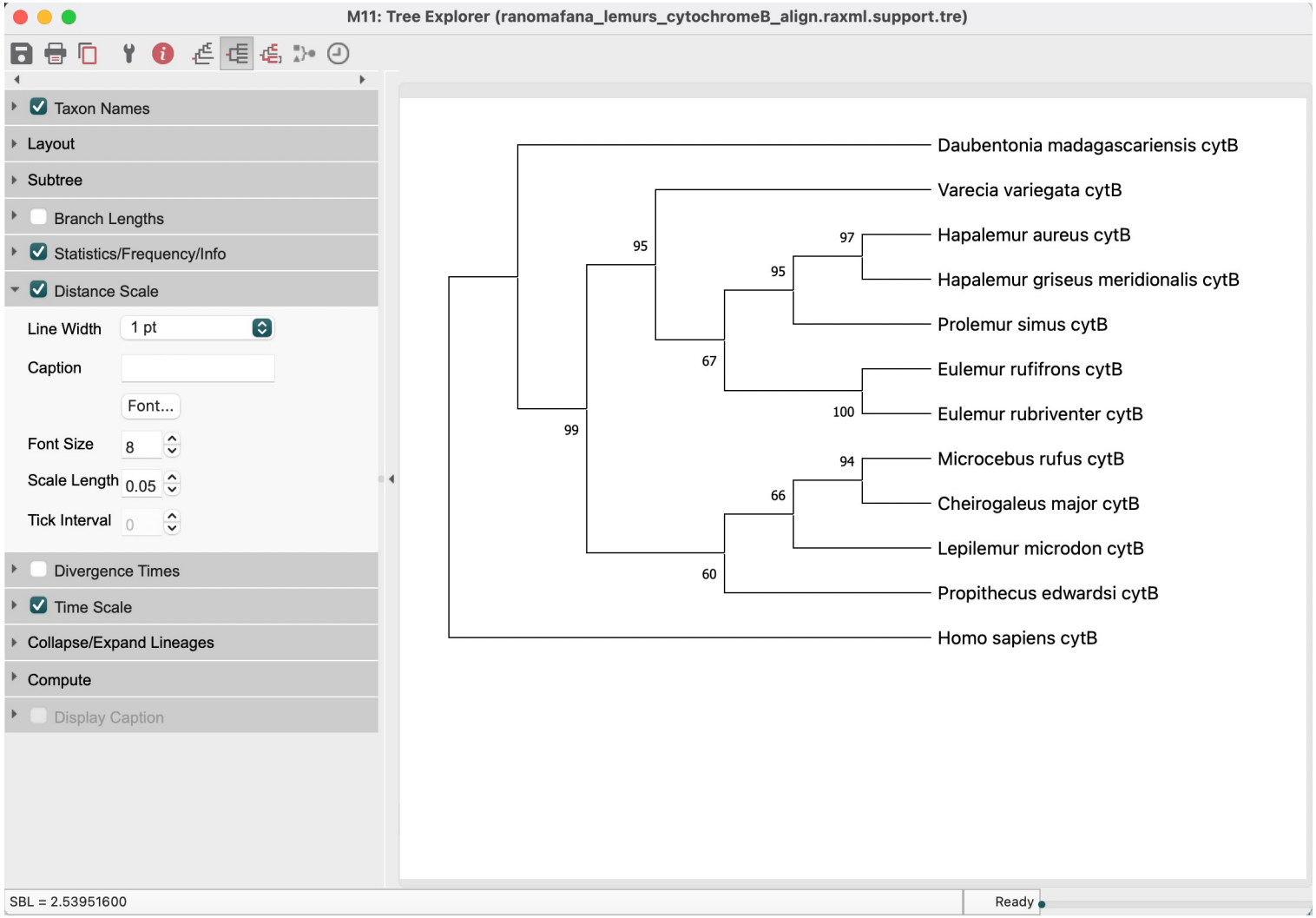
We're interested in the .support.tre file

This will include the tree topology and bootstrap support values, open in MEGA



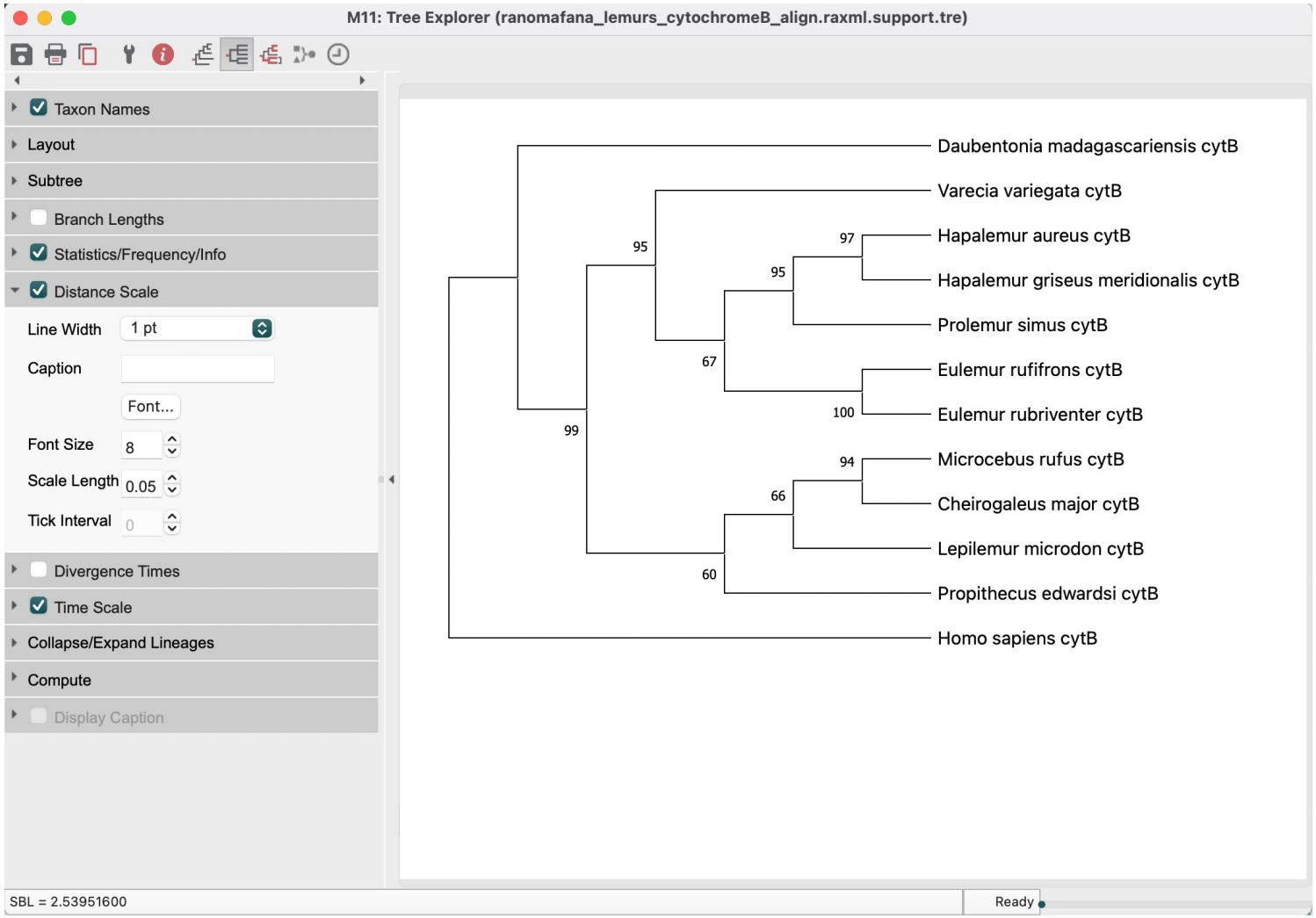
Nous sommes intéressés par le fichier .support.tre

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



We want a file of the tree saved that can be read in R, so newick. Click export trees…choose format newick, and customize in R

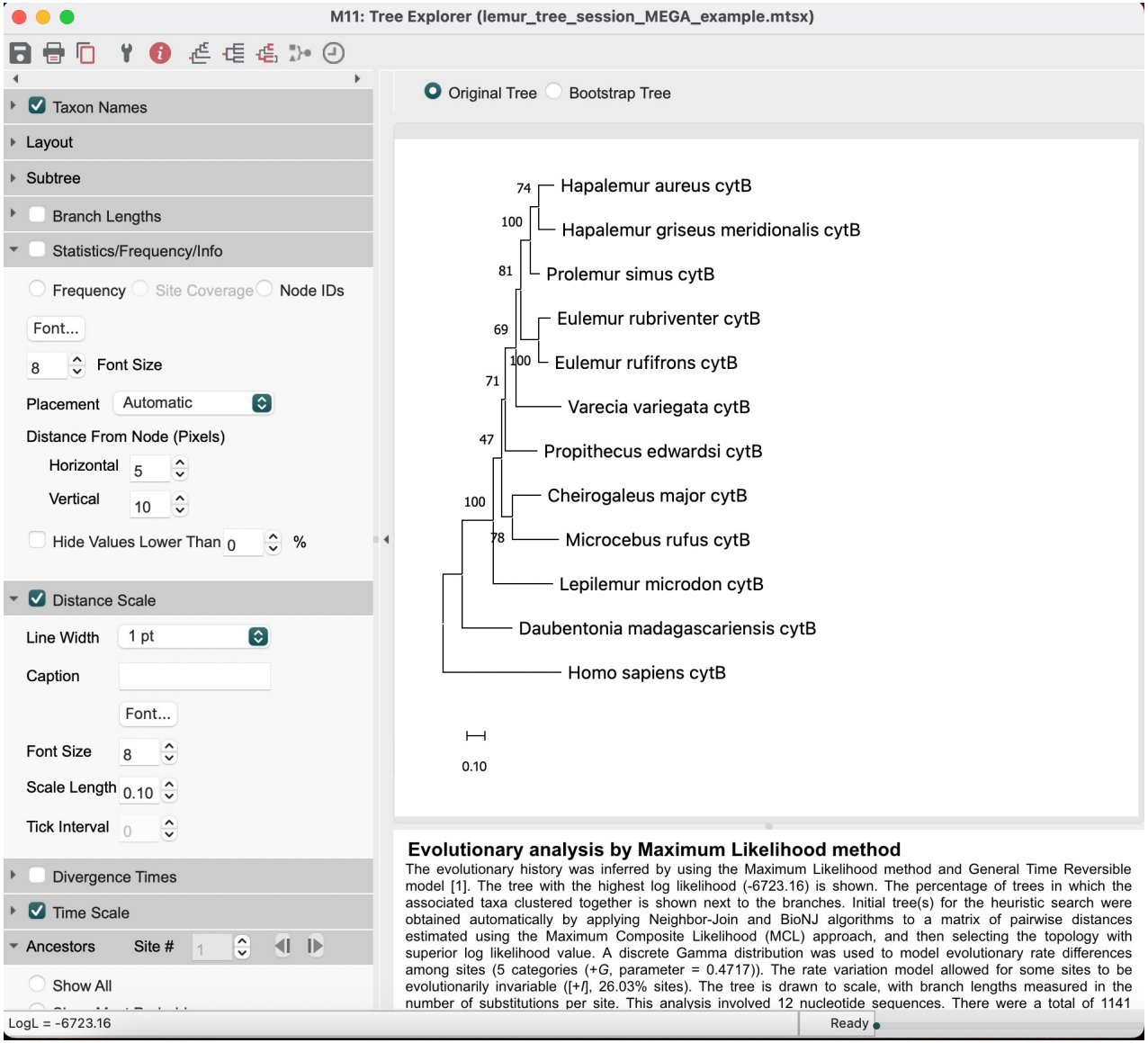
You can customize in MEGA too…it's just more limited



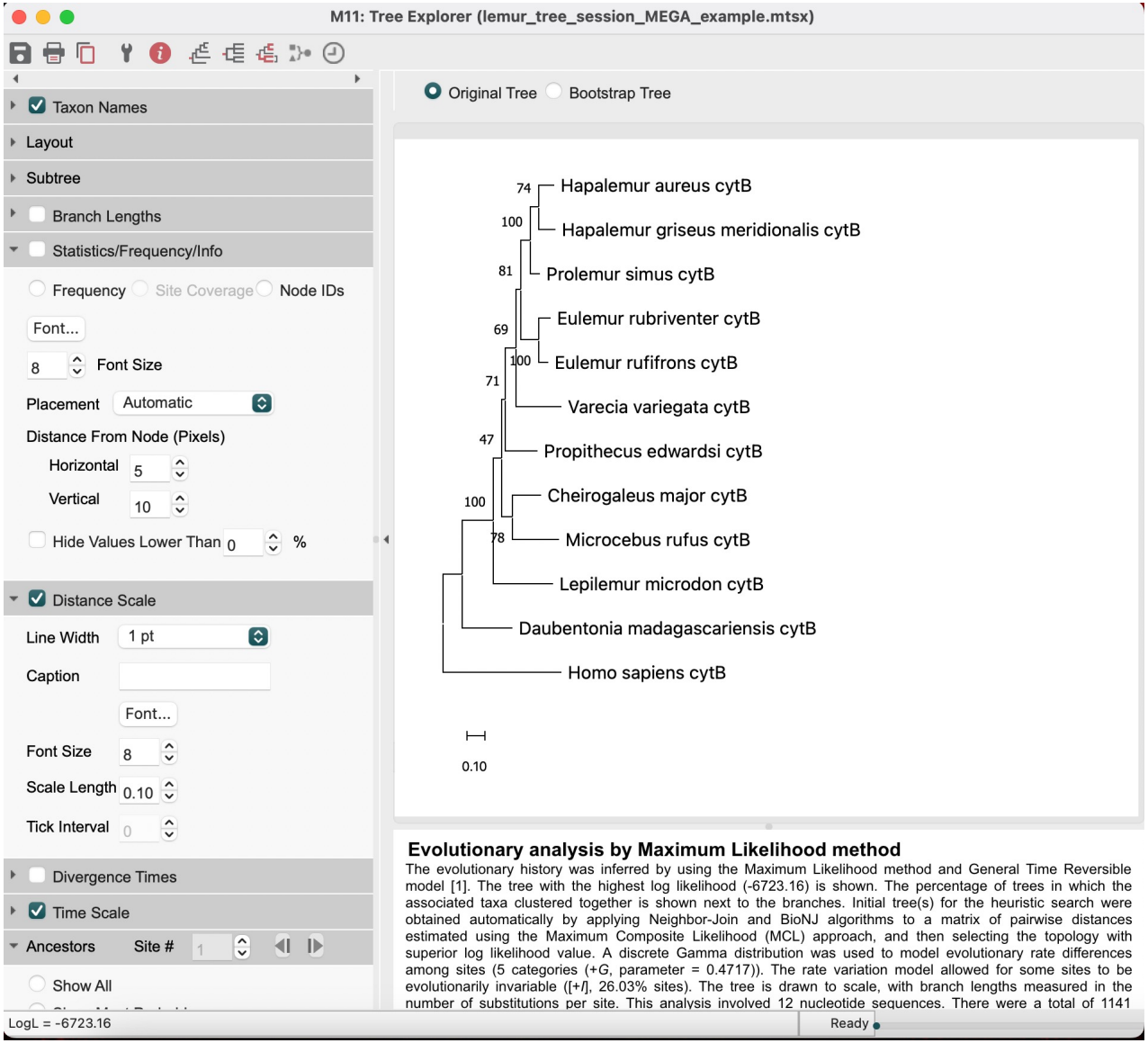
Nous voulons un fichier de l'arbre enregistré qui peut être lu dans R, donc newick. Cliquez sur Exporter les arbres... choisissez le format newick et personnalisez-le dans R

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

We could have done modeltest and RAxML in MEGA too…but takes forever!



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



Then make pretty in R!

- Follow instructions in lemur_tree_editing.R file

Alors faites joli en R !

- Suivez les instructions de lemur_tree_editing. Fichier R