

Chapitre 4. Phylogénie moléculaire : retracer l'évolution à partir des séquences

Introduction à la bioinformatique (UE SSV3U15)
2025-2026

Jacques van Helden
Aix-Marseille Université
orcid.org/0000-0002-8799-8584

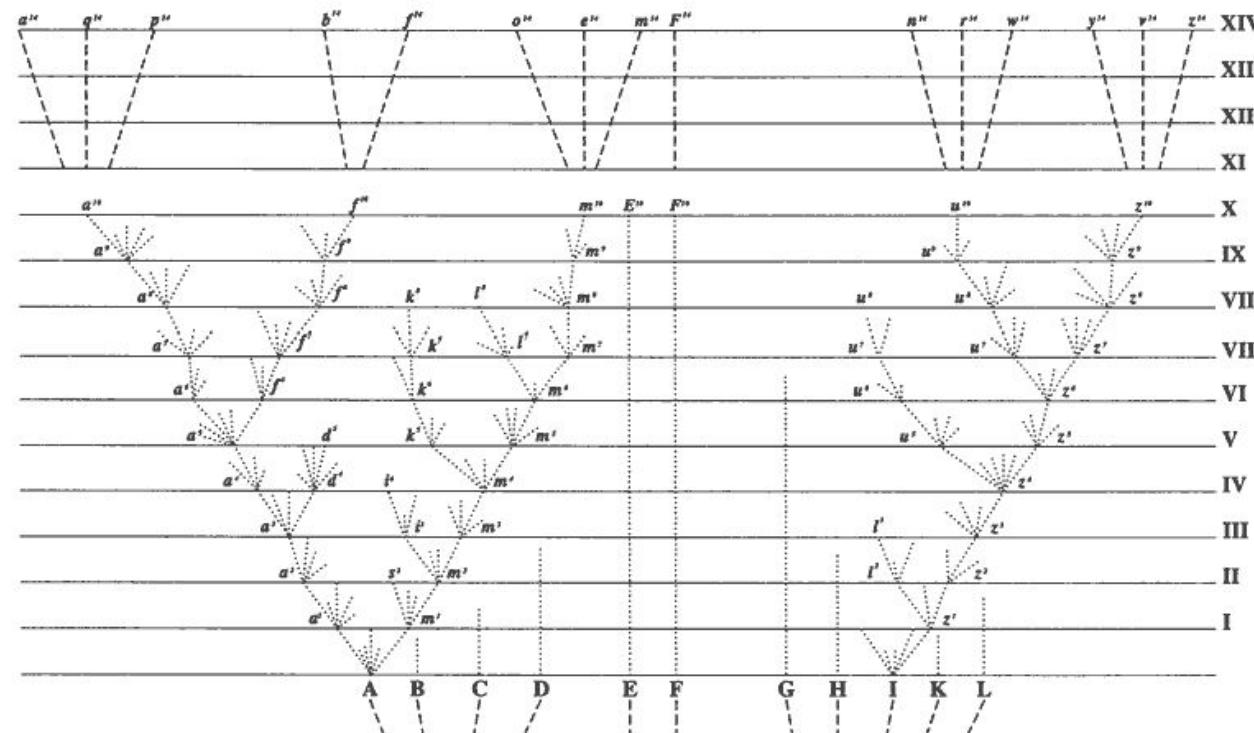
Contenu de ce chapitre

1. Représentations arborescentes de l'évolution
2. Concepts: homologie, analogie, paralogie, orthologie
3. Les duplications à l'origine de l'innovation
4. Phylogénomique : retracer l'évolution des espèces à partir des séquences génomiques
5. Retracer l'origine de SARS-CoV-2 dans les génomes des coronavirus
6. Pseudogènes ("gènes fossiles")
7. Quand les branches de l'arbre du vivant s'entrecroisent

Représentations arborescentes de l'évolution

La divergence des caractères

La seule figure de l'Origine des Espèces (C.Darwin, 1859) est une représentation conceptuelle de l'arbre de la vie.

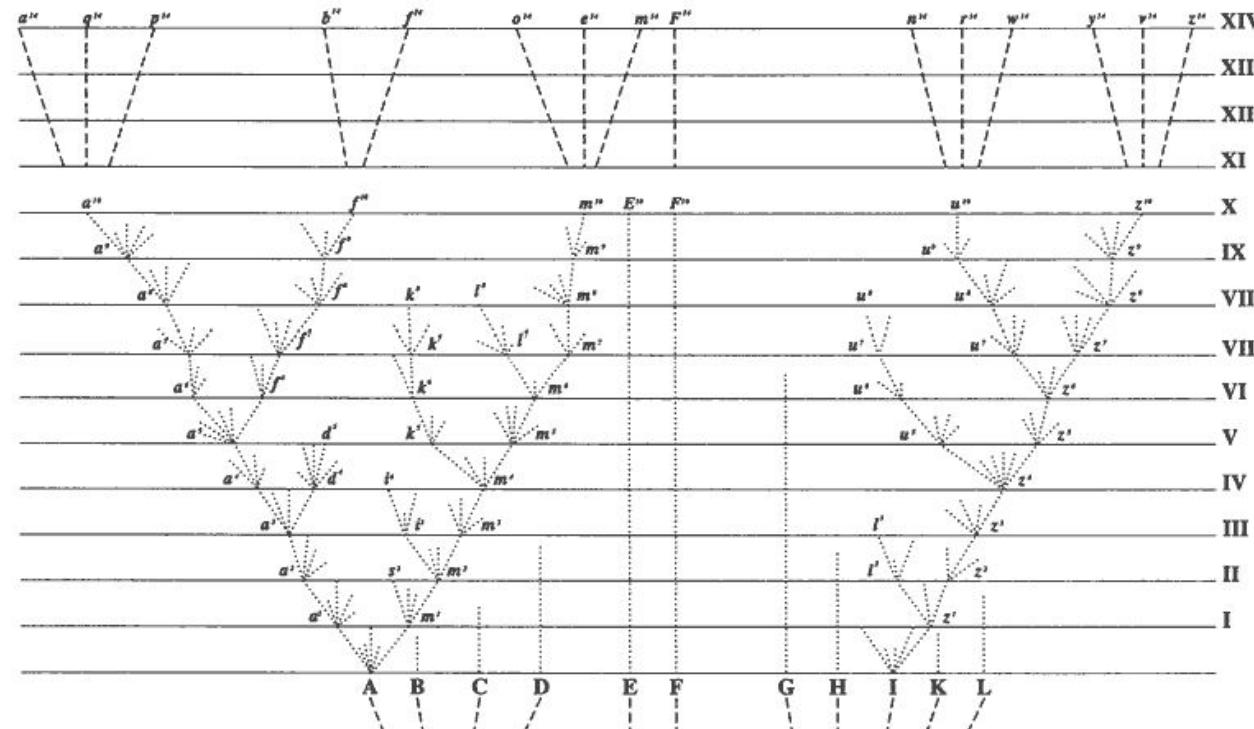


La divergence des caractères

Il s'agit d'un **arbre synchrone** :
chaque niveau horizontal
représente un moment donné.

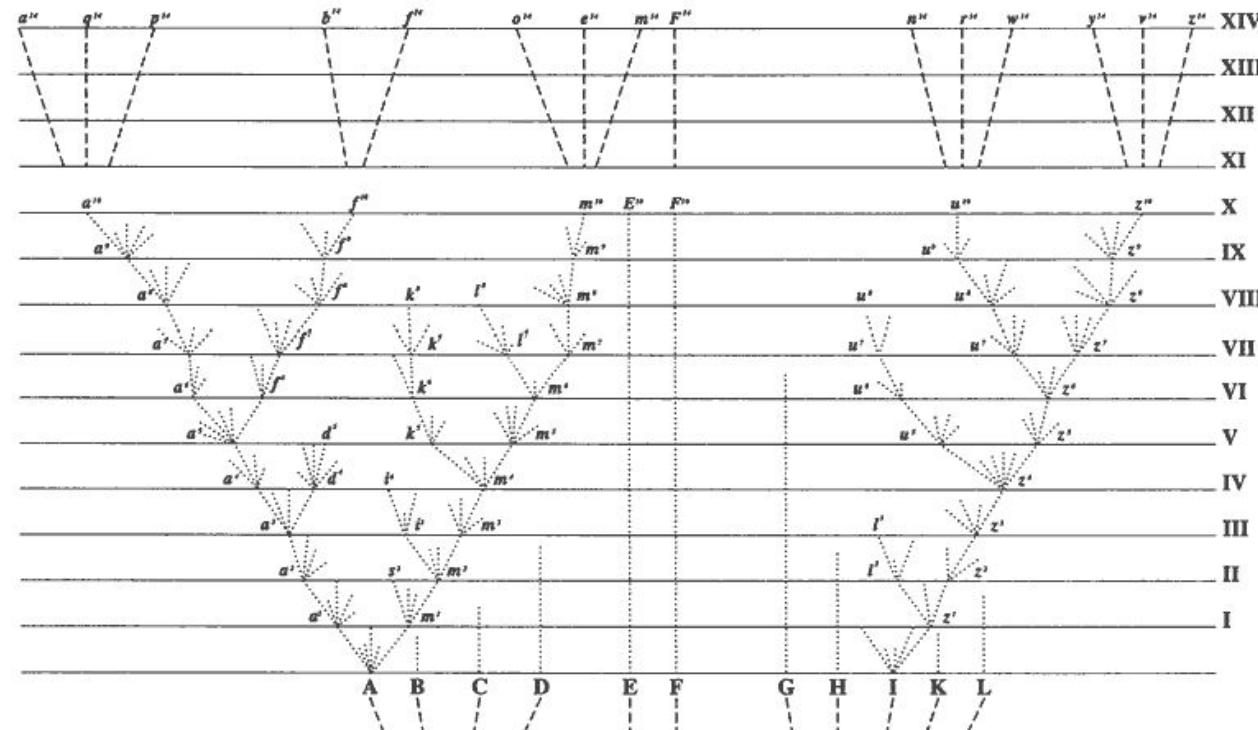
- La racine correspond aux époques les plus anciennes.
- Le niveau le plus élevé correspond au présent.
- A chaque époque on trouve des organismes de différents niveaux de complexité.

La hauteur ne représente donc pas une complexité ou un "niveau d'évolution", mais simplement l'écoulement du temps.



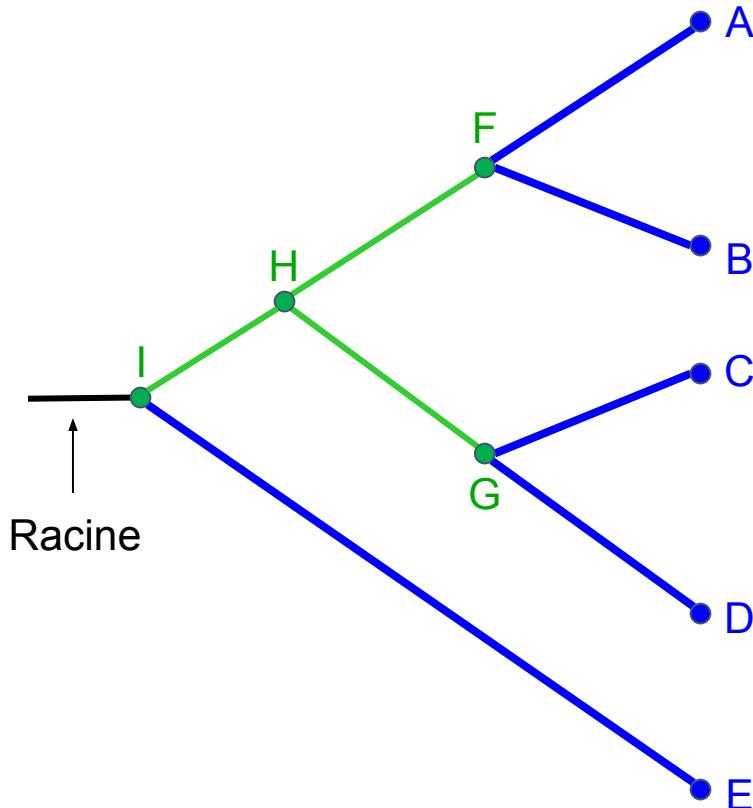
La divergence des caractères

- La plupart des branches sont abortives
- **Évolution graduelle** par accumulation de variations (mutations) le long des branches.
- Juste après un branchement, on a de très petites différences entre les variétés.
- Les observations dont on dispose sont généralement fragmentaires.
- Elles ne sont pas forcément placées sur une trajectoire linéaire depuis un ancêtre donné jusqu'aux espèces actuelles.



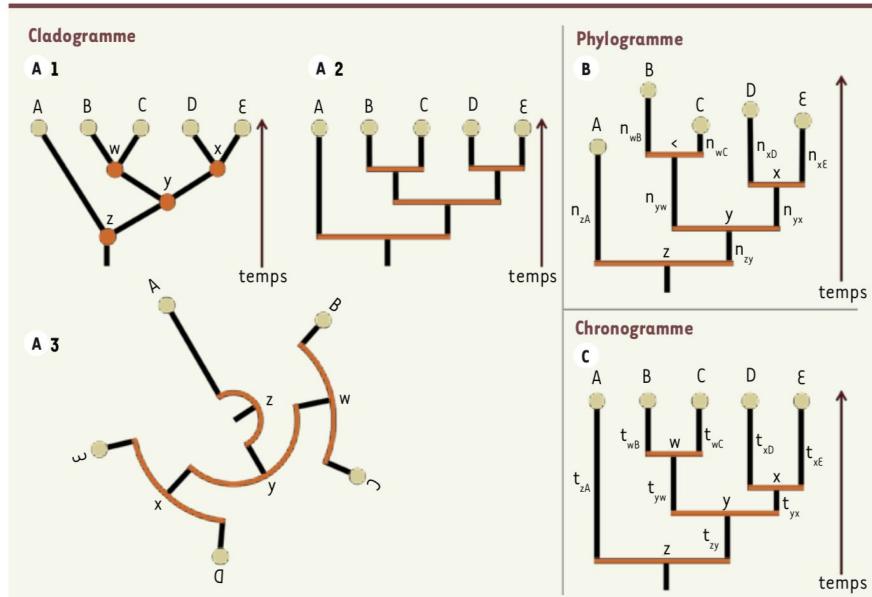
Unités taxonomiques opérationnelles (OTU) et hypothétiques (HTU)

- Les relations évolutives entre les objets étudiés (espèces, organes, séquences) sont représentées par des arbres phylogénétiques
- Les arbres sont des graphes composés de noeuds et de branches
 - **Noeuds = unités taxonomiques**
 - Feuilles ou **OTU = Unités Taxonomiques Opérationnelles** (A, B, C, D, E), pour lesquelles on dispose de données. Note : les OTU peuvent correspondre à des organismes existants ou éteints (données paléontologiques ou paléogénomiques).
 - Noeuds internes ou **HTU = Unités taxonomiques Hypothétiques** (F, G, H, I), pour lesquelles on ne dispose pas de données, et qui correspondent aux espèces ancestrales communes à plusieurs OTU.
 - Branches = relations de parenté(ancêtre/descendants) entre unités taxinomiques
 - Branches internes
 - Branches externes
- On appelle **topologie** l'ensemble des branchements de l'arbre.



Représentations arborescentes des histoires évolutives

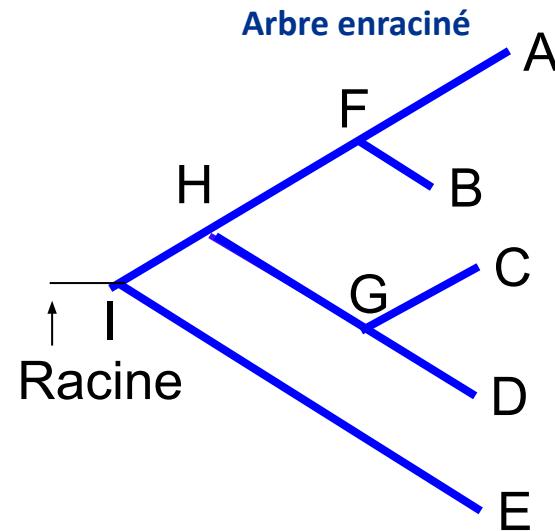
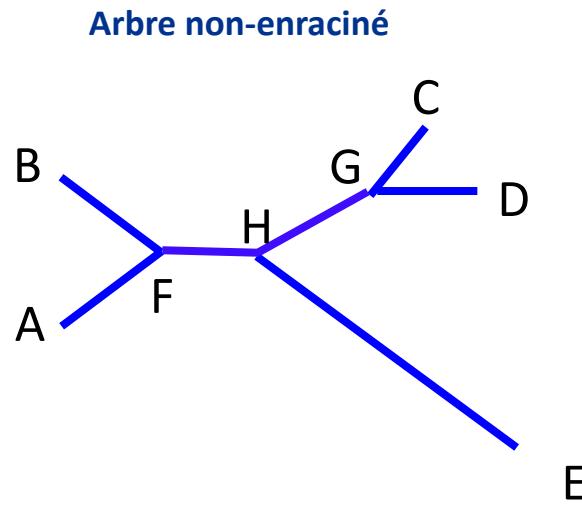
- On représente les histoires évolutives sous forme d'arbres
- Différents types de représentation peuvent être utilisés selon les cas.
 - Bifurcations triangulaires ou rectangulaires
 - Disposition radiale
- Dans un **phylogramme**, les longueurs des branches représentent le **nombre de différences génétiques ou morphologiques** entre deux espèces. Les feuilles de l'arbre (OTU) ne sont donc pas forcément alignées, car certaines branches peuvent évoluer plus rapidement que d'autres.
- Dans un chronogramme, la longueur des branches représente le **temps de divergence**. Les unités taxonomiques opérationnelles (OTU) peuvent être contemporaines (par exemple des séquences d'organismes actuels) auquel cas les feuilles sont alignées. Cependant, dans certains cas on dispose d'échantillons fossiles, et les feuilles peuvent alors occuper des hauteurs différentes.
- Le **cladogramme** indique les relations entre unités taxonomiques, sous forme de branchements successifs. La longueur des branches n'est indicative ni du temps écoulé ni du degré de divergence évolutive. On aligne généralement les OTU sur une même ligne, mais c'est une convention esthétique, qui n'associe aucune valeur numérique à la longueur des branches.



Casane, D. & Laurenti, P. Penser la biologie dans un cadre phylogénétique: L'exemple de l'évolution des vertébrés. Med Sci (Paris) 28, 1121–1127 (2012).
doi.org/10.1051/medsci/20122812024

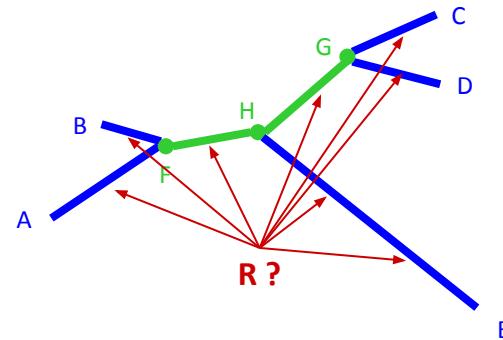
Arbres enracinés ou non enracinés

- Les arbres non-enracinés ne sont pas réellement des arbres phylogénétiques car ils n'ont pas de direction temporelle → indiquent les distances, mais pas les relations de parenté entre les noeuds.
- La **racine** définit une orientation de l'arbre, et donc un chemin évolutif unique vers chaque feuille.
- Elle symbolise le **dernier ancêtre commun** (l'ancêtre commun plus récent) de toutes les OTU.



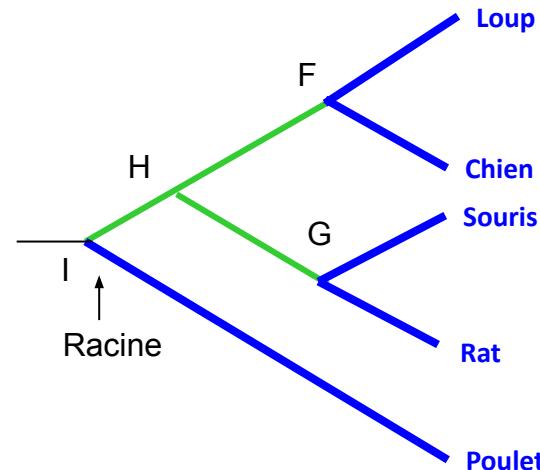
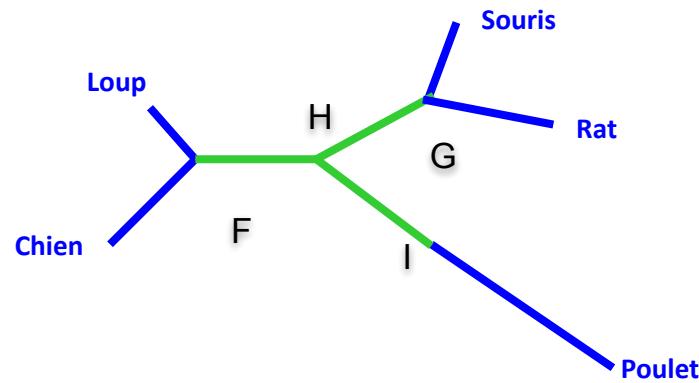
Comment enracer un arbre phylogénétique ?

- A priori, la racine pourrait se situer sur à n'importe quelle position sur n'importe quelle branche de l'arbre.
- Cependant, il n'y a qu'une seule de ces positions qui correspond à la véritable histoire évolutive.



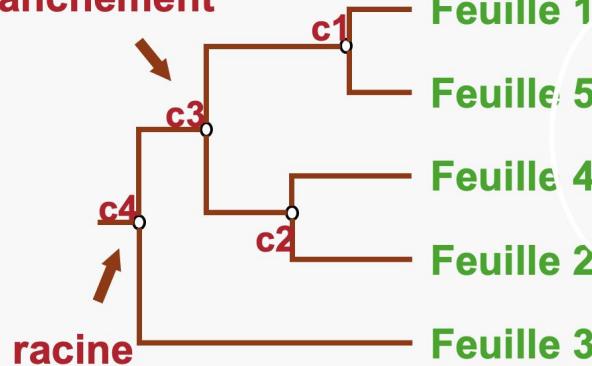
Approches pour enracer un arbre phylogénétique

- Dans certains cas, on peut s'appuyer sur une connaissance *a priori* de la feuille la plus externe parmi les OTU étudiées, qualifiée de **groupe extérieur (outgroup en anglais)**
 - Exemple : si un arbre contient chien, loup, souris, rat et poulet → sur base des connaissances biologiques, on décide que le **groupe extérieur est le poulet**
- En absence de connaissance *a priori* du OTU les plus externes parmi les OTU étudiées, on peut envisager un **enracinement au poids moyen** : on enrache l'arbre sur la branche qui minimise la moyenne des distances aux feuilles.
 - Note: ceci implique une hypothèse d'**horloge moléculaire**: on considère que le taux de mutation est constant au cours de l'évolution, et égal entre les branches. Cette hypothèse n'est généralement pas très réaliste, il s'agit d'une approximation.

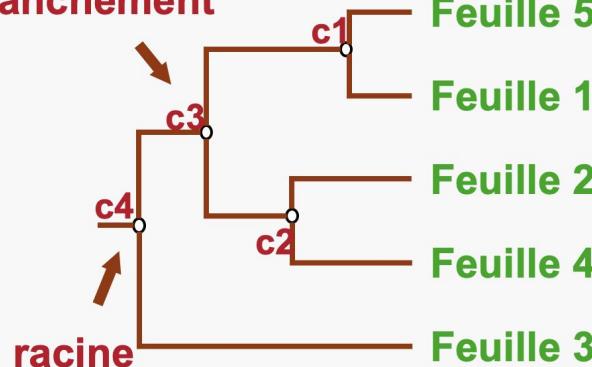


Isomorphisme des arbres phylogénétiques

branchement



branchement



- Dans un arbre, les deux enfants de chaque branche peuvent être interchangés.
- Le résultat est un arbre **isomorphe**, considéré équivalent à l'arbre initial.
- Les deux arbres de gauche sont équivalents.
- Cependant
 - Arbre du dessus: les feuilles 1 et 2 sont très éloignées.
 - Arbre du dessous: les feuilles 1 et 2 sont voisines.
- Les distances verticales entre deux nœuds ne reflètent pas leur distance réelle !
- La distance entre deux nœuds est la somme des longueurs des branches qui les séparent.

Concepts: homologie, analogie, paralogie, orthologie

Similarité de séquences, homologie et analogie

■ Homologie :

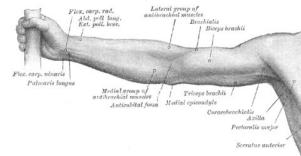
- ressemblance entre des caractères (phénotypiques ou génétiques) qui s'explique par une **origine ancestrale commune**
- des différences entre résultent de l'accumulation de modifications. Il s'agit d'une **évolution divergente** (les caractères deviennent de plus en plus différents avec le temps).

■ Analogie :

- ressemblance résultant de **trajectoires indépendantes**.
- Il s'agit d'une **évolution convergente** (les caractères se ressemblent de plus en plus avec le temps), qui peut éventuellement manifester l'effet d'une même pression évolutive.

Homologie

Humain (mammifère)



Chimpanzé (mammifère)



Analogie

Poule (oiseau)

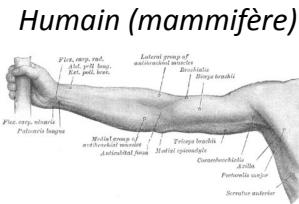


Mouche (insecte)



Homologie et analogie : un exemple non trivial

- Les **membres antérieurs** des chauves-souris sont homologues à ceux des oiseaux. Il s'agit d'un **caractère ancestral** (caractère hérité depuis le dernier ancêtre commun d'un taxon).
 - La présence de membres antérieurs est commune à tous les tétrapodes actuels, ils en ont hérité à partir des premiers vertébrés terrestres.
- Cependant, les **ailes** des chauves-souris sont analogues aux ailes des oiseaux, car les caractères qui permettent à ces tétrapodes de voler sont apparus indépendamment dans les deux lignées. Il s'agit de **caractères dérivés** (qui n'étaient pas présent chez leur ancêtre commun).



Humain (mammifère)

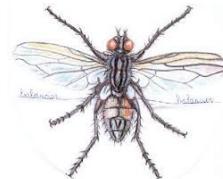


Chimpanzé (mammifère)

Poule (oiseau)



Mouche (insecte)



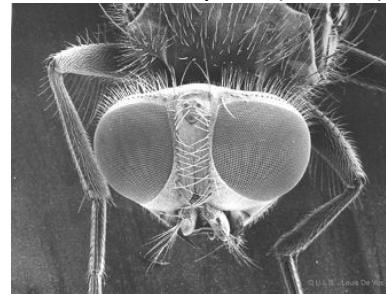
Chauve-souris (mammifère)



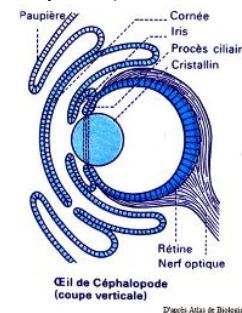
Structures analogues et convergence évolutive

- La vision des différents groupes d'animaux repose sur des yeux de structures très diverses.
- L'oeil à facettes des insectes est très différent de l'oeil des vertébrés
- L'oeil de pieuvre présente de fortes similarités de structures avec l'oeil humain, mais quelques différences notoires
 - **Similarités:** oeil sphérique, cornée, iris, cristallin, ...
 - **Défauts:** orientation des cellules rétinienques: les axones partent vers l'intérieur chez les vertébrés, vers l'extérieur chez les céphalopodes
- En dépit de leur ressemblance anatomique, l'oeil de pieuvre et l'oeil humain résultent de **voies évolutives indépendantes**. Leur ressemblance est due à une **convergence évolutive** plutôt qu'à une origine commune. Il s'agit d'un cas spectaculaire de ressemblance par **analogie**.

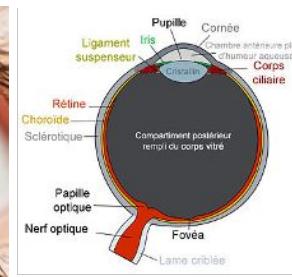
Oeil de drosophile (insecte)



Oeil de pieuvre (mollusque céphalopode)



Oeil humain (mammifère)



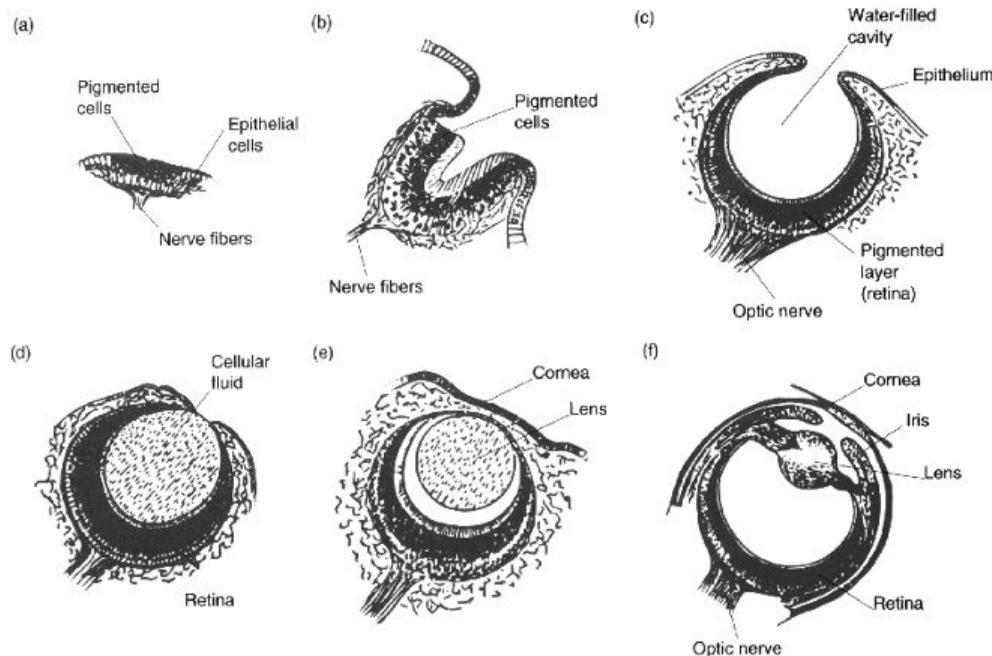
Figures:

<https://www.futura-sciences.com/sante/dossiers/medecine-oeil-vision-dela-vision-667/page/2/>

Chez différentes espèces de mollusques, on observe une grande diversité dans la structure de l'oeil, allant de formes très rudimentaires (quelques cellules photosensibles sur l'épiderme) à un oeil aussi complexe que celui des vertébrés, dans le cas de la pieuvre.

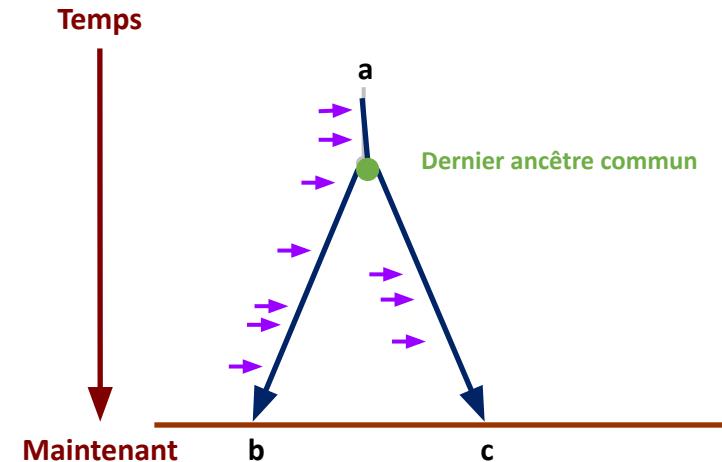
→ **une évolution graduelle est capable de produire des structures extrêmement complexes**

Figure 13.2 Stages in the evolution of the eye, illustrated by species of molluscs. (a) A simple spot of pigmented cells. (b) Folded region of pigmented cells, which increases the number of sensitive cells per unit area. (c) Pin-hole camera eye, as is found in *Nautilus*. (d) Eye cavity filled with cellular fluid rather than water. (e) The eye is protected by adding a transparent cover of skin, and part of the cellular fluid has differentiated into a lens. (f) Full, complex eye, as found in octopus and squid. Reprinted, by permission of the publisher, from Strickberger (1990).



Séquences homologues

- Au même titre que pour les caractères physiques, les ressemblances entre deux séquences macromoléculaires peuvent résulter de différentes causes.
- **Homologie**
 - ressemblances héritées d'une **séquence ancestrale commune**
 - évolution **divergente**: accumulation de mutations à partir de la séquence ancestrale commune



Les séquences b et c sont homologues

Evénements évolutifs générant des séquences homologues

- Pour l'analyse de la phylogénie moléculaire, nous porterons un intérêt tout particulier à deux événements évolutifs susceptibles de générer des séquences homologues: duplication et spéciation.
- **Duplication**
 - Une duplication est une mutation qui génère un dédoublement d'une partie de l'ADN génomique. La duplication peut recouvrir l'ensemble du génome (formation d'organismes polyploïdes), un chromosome entier, ou un fragment de chromosome de taille plus ou moins grande.
 - Les duplications peuvent éventuellement entraîner l'apparition de copies multiples d'un ou plusieurs gènes, provoquant ainsi une certaine redondance de l'information génétique.
 - Dans certains cas, l'une des copies dupliquées du gène acquiert, par accumulation de mutations, de nouvelles caractéristiques qui lui permettent d'assumer une nouvelle fonction. Ce mécanisme, appelé duplication-divergence, est en grande partie à l'origine de la diversification des fonctions biologiques.
- **Spéciation**
 - Processus évolutif qui résulte en la formation d'espèces distinctes à partir d'une espèce unique.
 - Les événements de duplication et spéciation suscitent l'apparition de copies multiples à partir d'une seule séquence, soit au sein d'une même espèce (duplication), soit au sein des espèces distinctes dérivées de la spéciation. Ces séquences, dont la similarité résulte d'une séquence ancestrale commune, sont dites **homologues**

- Avant d'affirmer que deux séquences sont homologues, nous devrions pouvoir retracer leur histoire jusqu'à leur ancêtre commun.
- Nous ne pouvons malheureusement pas disposer des séquences de toutes les espèces disparues. Il est donc impossible de démontrer formellement l'homologie.
- Cependant, nous pouvons appuyer l'hypothèse d'homologie sur une analyse de la vraisemblance d'un scénario évolutif (taux de mutations, niveaux de similarités).
 - Si des séquences très longues ont des taux très forts de similarité, on considérera qu'elles descendent *vraisemblablement* d'un ancêtre commun.
 - Pour des séquences courtes, une forte ressemblance, voire une identité parfaite, peuvent éventuellement provenir
- L'inférence d'homologie est toujours attachée à un certain ***risque de faux positifs***. Les modèles évolutifs nous permettent d'estimer ce risque.

L'homologie est une relation logique (soit vraie, soit fausse)

- Deux séquences sont homologues ou elles ne le sont pas.
- Homologie = le fait de posséder des caractères similaires parce qu'ils dérivent d'un caractère ancestral commun
- Il est donc complètement inappropriate de parler de « niveau d'homologie » ou « pourcentage d'homologie ».



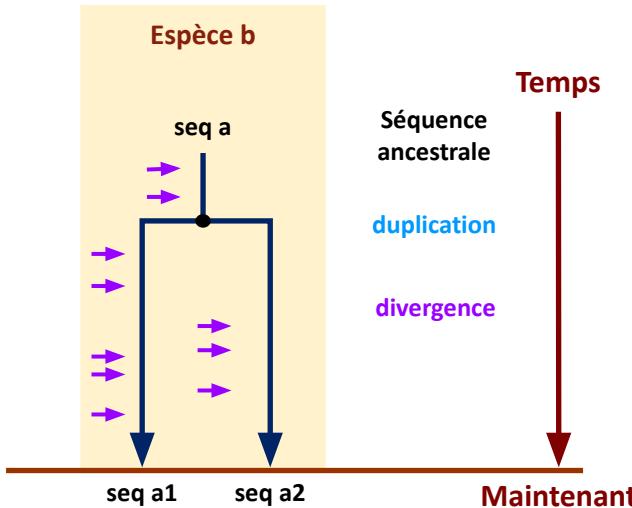
Formulation correcte

- On observe un certain niveau de similarité entre deux séquences (pourcentages de résidus identiques, pourcentages de résidus « similaires »).
- Sur cette base, on évalue deux scénarios évolutifs: cette similarité peut provenir d'une évolution convergente (analogie) ou divergente à partir d'un ancêtre commun (homologie).
- Si la deuxième hypothèse est la plus vraisemblable, on *infère* que les séquences sont homologues.

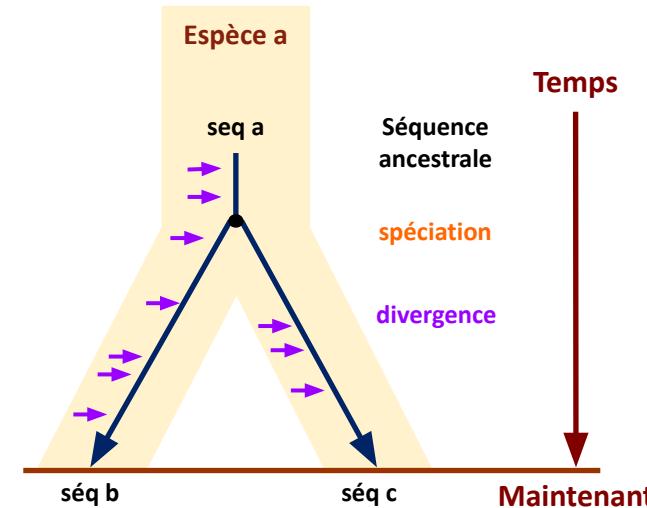
Scénarios évolutifs

- Nous disposons de deux séquences, et nous supposons qu'elles divergent d'un ancêtre commun.
- La divergence peut résulter
 - d'une **duplication** (création de deux copies du gène dans le même génome)
 - ou d'une **spéciation** (formation d'espèces séparées à partir d'une espèce unique).
- Les **flèches violettes** indiquent les mutations (substitutions, délétions, insertions) qui s'accumulent au sein d'une séquence particulière au cours de son histoire évolutive. Ces mutations sont à l'origine de la diversification des séquences, des structures et des fonctions.

Duplication



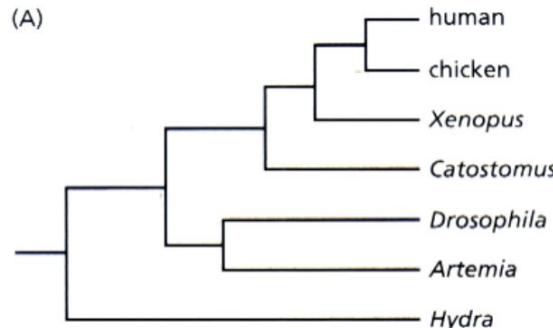
Spéciation



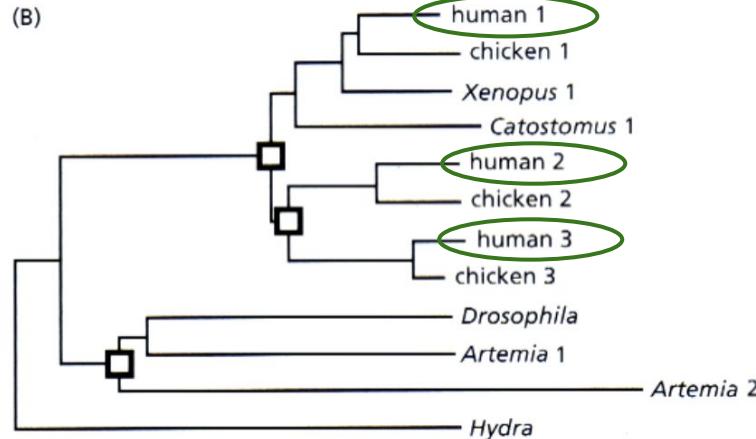
Arbre des espèces et arbre des molécules

- En partant d'une famille de séquences macromoléculaires (ADN, ARN, protéines), on peut construire des arbres phylogénétiques.
 - Arbre des espèces → chaque noeud interne représente une spéciation
 - Arbre des molécules → chaque noeud interne représente soit une duplication, soit une spéciation
- En comparant l'arbre des molécules et l'arbre des espèces, on peut inférer l'histoire évolutive de cette famille de séquences.
- Note : sur l'arbre des molécules, un même organisme peut éventuellement se retrouver plusieurs fois, s'il existe plusieurs molécules homologues dans son génome.

L'arbre des espèces

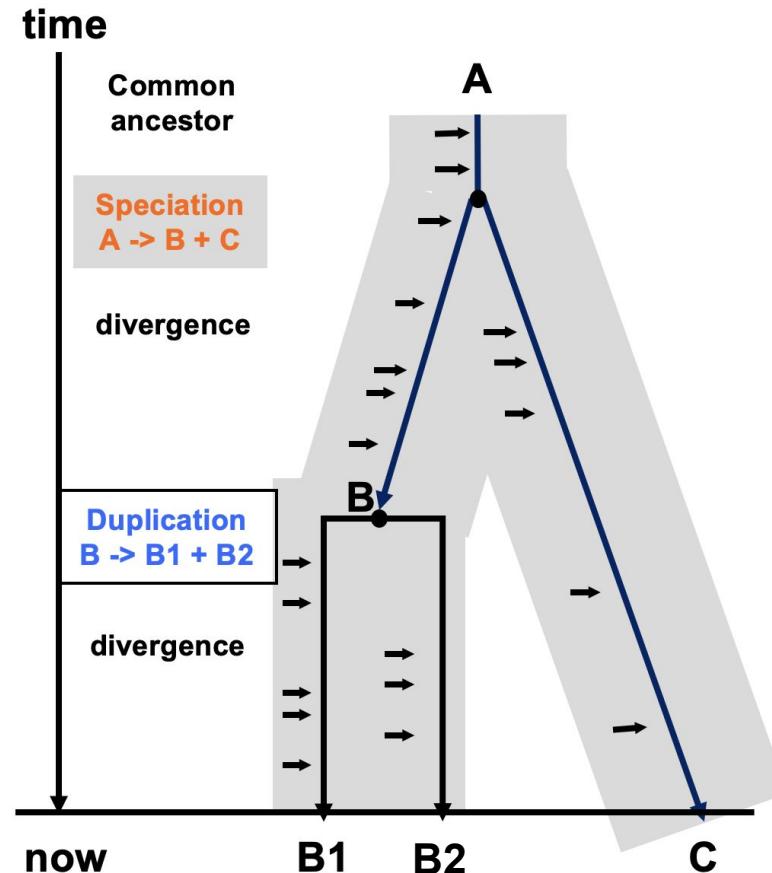


Arbre des molécules



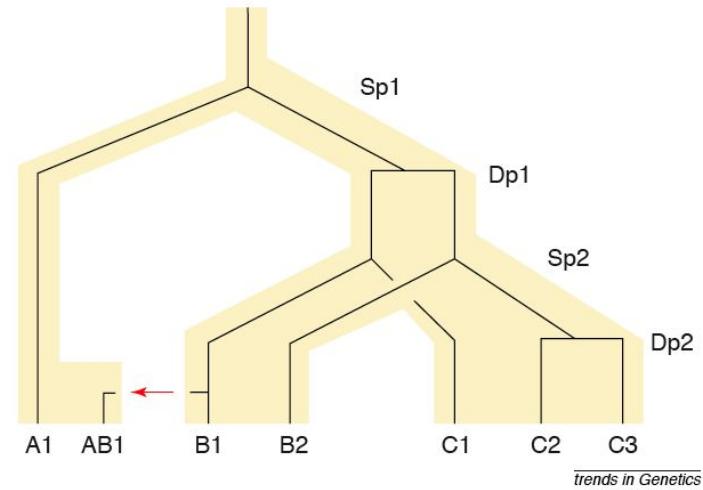
Orthologie versus paralogie

- Zvelebil & Baum (2000) fournissent une définition claire et opérationnelle des concepts d'orthologie et paralogie.
 - Orthologues:** séquences dont le dernier ancêtre commun précède immédiatement un événement de spéciation.
 - Paralogues** séquences dont le dernier ancêtre commun précède immédiatement un événement de duplication
- Exemples:
 - B et C sont **orthologues**, car leur dernier ancêtre commun (A) précède un événement de **spéciation** ($A \rightarrow B + C$).
 - B1 et B2 sont **paralogues** car le premier événement évolutif qui succède à leur dernier ancêtre commun (B) est une **duplication** ($B \rightarrow B1 + B2$).



Représentation détaillée des événements de spéciation / duplication

- La figure de droite combine deux niveaux de représentation
 - Les lignes noires fines représentent les relations évolutives entre molécules (arbre des molécules).
 - Les ombrages épais représentent l'arbre des espèces.
- Les **spéciations (Sp)** sont représentées par des branchements triangulaires sur l'arbre des espèces
 - En cas de spéciation, la molécule ancestrale se retrouve dans chacune des espèces dérivées.
- Les **duplications (Dp)** sont représentées par des branchements rectangulaires.
 - En cas de duplication, on retrouve au sein de la même espèce deux copies de la séquence ancestrale.



The idealized evolution of a gene (lines) is shown from a common ancestor in an ancestral population (the gray background), descending to three populations labelled A, B and C. There are two speciation events (Sp1 and Sp2), each occurring at the junctions shown as an upside down Y. There are also two gene-duplication events (Dp1 and Dp2), depicted by a horizontal bar. Two genes whose common ancestor resides at a Y junction (speciation) are orthologous. Two genes whose common ancestor resides at a horizontal bar junction (gene duplications) are paralogous. Thus, C2 and C3 are paralogous to each other but are orthologous to B2. Both are paralogous to B1 but orthologous to A1. The red arrow denotes the transfer of the B1 gene from species B to species A. As a result, the AB1 gene is xenologous to all six other genes. All three subtype relationships are reflexive, that is, $A1 \Rightarrow B1$ implies $B1 \Rightarrow A1$ where \Rightarrow should be read, for example, as 'is orthologous to.' However, the relationships are not transitive. Thus, $C2 \Rightarrow A1 \Rightarrow C3$ might be true, but it is not necessarily therefore true that $C2 \Rightarrow C3$, as indeed it is not in the figure if \Rightarrow is read as 'is orthologous to.' A different non-transitivity occurs for 'is paralogous to' with $B2 \Rightarrow C1 \Rightarrow C2$.

Définitions des concepts d'après Fitch (2000)

L'article de Fitch (2000) définit les concepts suivants.

- **Homologie**

- Owen (1843). « le même organe sous toutes ses variétés de forme et de fonction ».
- Fitch (2000). L'homologie est la relation entre toute paire de caractères qui descendent, généralement avec divergence, d'un caractère ancestral commun.
- Note: "caractère" peut se référer à un trait phénotypique, un site d'une séquence, à un gène entier, ...
- Application moléculaire: deux gènes sont homologues s'ils divergent d'un gène ancestral commun.

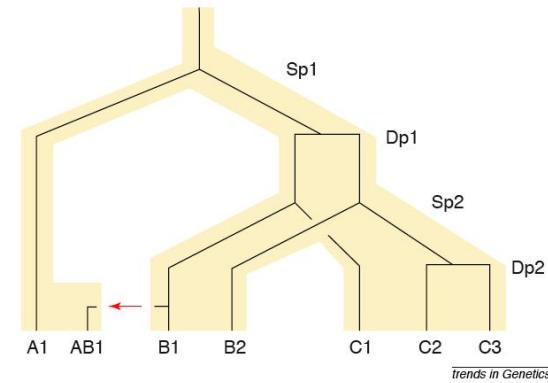
- **Analogie:** relation entre deux caractères qui se sont développés de façon convergente à partir d'ancêtres non-apparentés.

- **Cénancêtre:** l'ancêtre commun le plus récent pour les groupes taxonomiques considérés.

- **Orthologie:** relation entre deux caractères homologues dont l'ancêtre commun se trouve chez le cénancêtre des taxa à partir desquels les séquences ont été obtenues.

- **Paralogie:** relation entre deux caractères émanant d'une duplication de gène pour ce caractère.

- **Xénologie:** relation entre deux caractères dont l'histoire, depuis leur dernier ancêtre commun, inclut un transfert entre espèces (horizontal) du matériel génétique pour au moins l'un de ces caractères.



The idealized evolution of a gene (lines) is shown from a common ancestor in an ancestral population (the gray background), descending to three populations labelled A, B and C. There are two speciation events (Sp1 and Sp2), each occurring at the junctions shown as an upside down Y. There are also two gene-duplication events (Dp1 and Dp2), depicted by a horizontal bar. Two genes whose common ancestor resides at a Y junction (speciation) are orthologous. Two genes whose common ancestor resides at a horizontal bar junction (gene duplications) are paralogous. Thus, C2 and C3 are paralogous to each other but are orthologous to B2. Both are paralogous to B1 but orthologous to A1. The red arrow denotes the transfer of the B1 gene from species B to species A. As a result, the AB1 gene is xenologous to all six other genes. All three subtype relationships are reflexive, that is, $A1 \Rightarrow B1$ implies $B1 \Rightarrow A1$ where \Rightarrow should be read, for example, as 'is orthologous to.' However, the relationships are not transitive. Thus, $C2 \Rightarrow A1 \Rightarrow C3$ might be true, but it is not necessarily therefore true that $C2 \Rightarrow C3$, as indeed it is not in the figure if \Rightarrow is read as 'is xenologous to.' A different non-transitivity occurs for 'is paralogous to' with $B2 \Rightarrow C1 \Rightarrow C2$.

Analogie
Homologie

Paralogie

Xénologie ou non
(xénologues issus de paralogues)

Orthologie

Xénologie ou non
(xénologues issus d'orthologues)

Exercice: types d'homologie

Sur base des définitions de **Zvelebil & Baum** (paralogie et orthologie), et de **Fitch** (xénologie), qualifiez la relation entre chaque paire de gènes dans le schéma de Fitch (ci-contre).

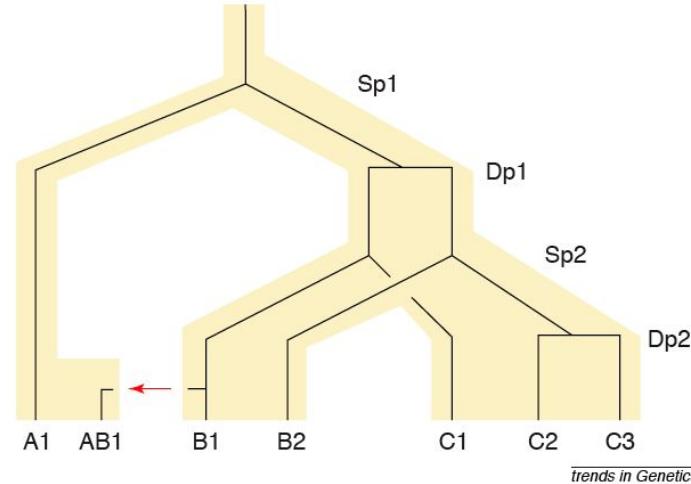
- P paralogie
- O orthologie
- X xenologie
- A analogie

Solutions dans un diaporama séparé

	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2							
C1							
C2							
C3							

Figure: Fitch, W. M. Homology a personal view on some of the problems. Trends Genet 16, 227–231 (2000). [doi.org/10.1016/s0168-9525\(00\)02005-9](https://doi.org/10.1016/s0168-9525(00)02005-9)

Définitions : Zvelebil, M. J. & Baum, J. O. Understanding bioinformatics. (Garland Science, 2008).



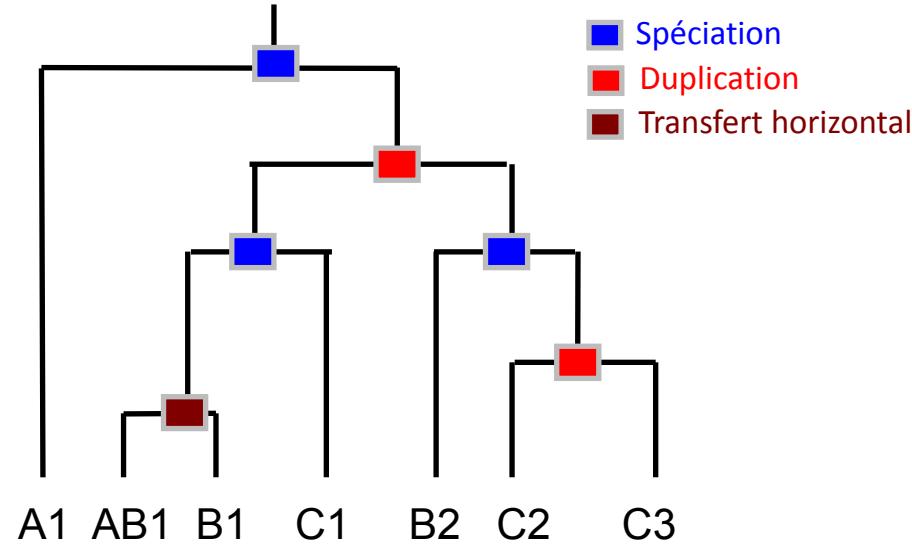
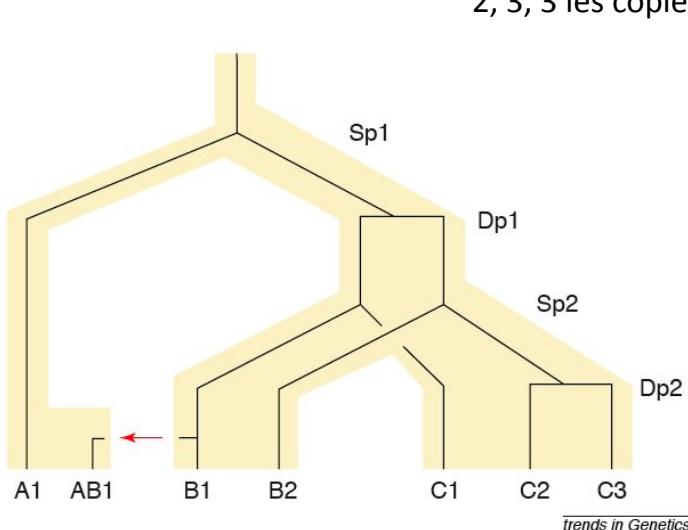
- Paire d'**orthologues**: paire de gènes dont le dernier ancêtre commun précède immédiatement un événement de spéciation (ex: a_1 and a_2). Source: Zvelebil & Baum, 2000.
- Paire de **paralogues**: paire de gènes dont le dernier ancêtre commun précède immédiatement une duplication génique (ex: b_2 and b_2'). Source: Zvelebil & Baum, 2000.
- **Xénologie**: relation entre deux caractères dont l'histoire, depuis leur dernier ancêtre commun, inclut un transfert entre espèces (horizontal) du matériel génétique pour au moins l'un de ces caractères. Source: Fitch, 2000.

Représentation classique des spéciations / duplications

La représentation de gauche et de droite sont équivalentes.

Celle de gauche (Fitch, 2000) montre bien la transmission parallèle des unités moléculaires (gènes, protéines) qui résultent d'une duplication (bifurcation rectangulaire) au fil des spéciations (bifurcations triangulaires).

La représentation de droite est celle utilisée par les bases de données telles [Ensembl genomes](#). Sur l'arbre des molécules, chaque noeud interne est marqué par une couleur différente selon qu'il correspond à une spéciation ou à une duplication.

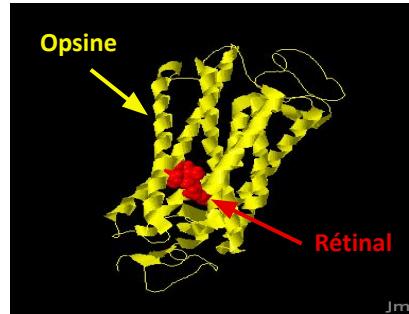


Les duplications à l'origine de l'innovation

Structure d'une opsine

Modèle tridimensionnel du pigment des cônes bleus

(Structure PDB 1kpn affichée avec Jmol)



(Structure PDB 1kpn affichée avec MacPyMol)

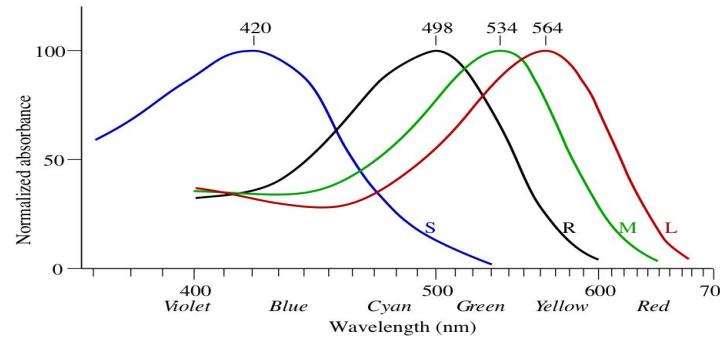


Perception de la lumière

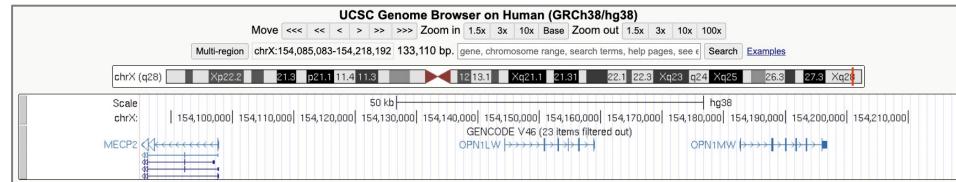
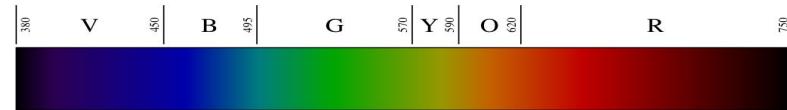
- **Rétinal** : petite molécule, absorbe la lumière
- **Opsine** :
 - protéine transmembranaire
- Complexe Opsine-Rétinal → perception de la lumière
- La séquence d'une opsine détermine le spectre de sensibilité → mutations peuvent modifier la longueur d'onde de sensibilité maximale

La vision trichromatique chez les primates de l'ancien monde

- Les primates de l'ancien monde (Afrique + Asie + Europe), y compris l'humain, ont une **vision trichromatique**, basée sur 3 pigments
 - Bleu (short-waves opsin, SW)
 - Vert (medium-waves opsin, MW)
 - Rouge (long-waves opsin, LW)
- Les autres mammifères, y compris les primates du nouveau monde (Amériques) ont une **vision dichromatique**.
 - Ils disposent d'une opsine sensible au bleu, et d'une autre sensible aux ondes vert-rouge,
 - Ils peuvent distinguer le bleu du vert ou du rouge, mais ne font pas la différence entre vert et rouge (équivalent au daltonisme humain).
- Chez les primates de l'ancien monde, la présence d'opsines distinctes avec une sensibilité "plutôt rouge" et "plutôt verte" résulte d'une **duplication du gène codant pour l'opsine rouge-verte**. En effet, on trouve sur le chromosome X 2 gènes en tandem
 - OPN1LW (Long Waves) : gène de l'opsine rouge
 - OPN1MW (Short Waves) : gène de l'opsine verte
- OPN1SW (Short Waves) : gène de l'opsine verte, sur chromosome 7 chez l'humain

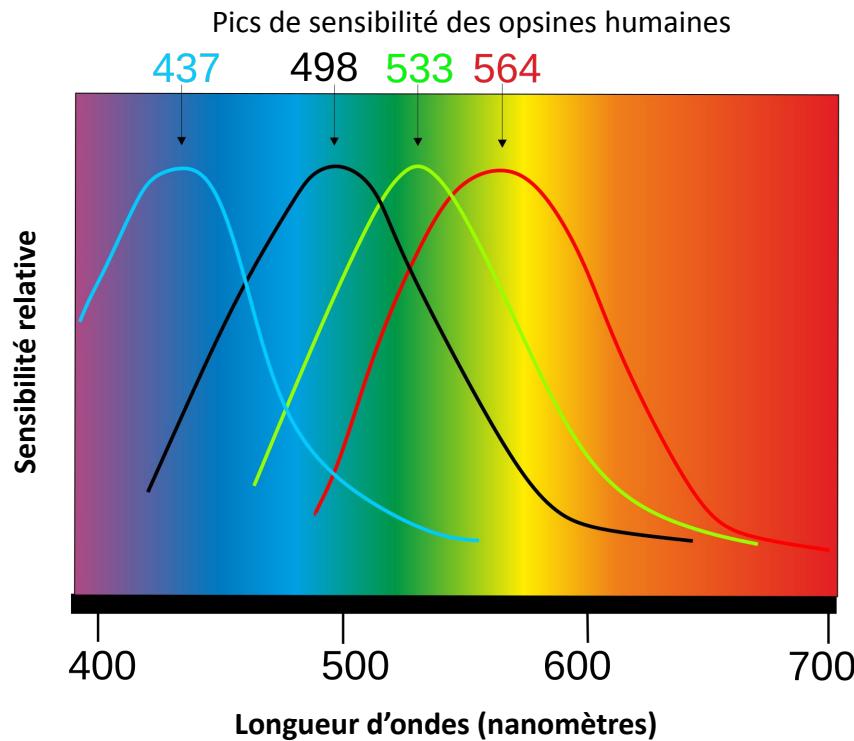


[http://fr.wikipedia.org/wiki/Cône_\(biologie\)](http://fr.wikipedia.org/wiki/Cône_(biologie))



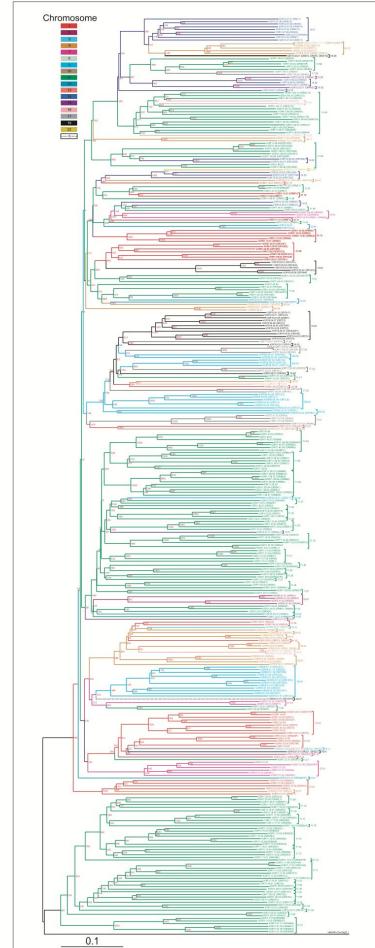
Le spectre visible

- Chaque cellule photoréceptrice perçoit une gamme spécifique de longueur d'ondes, avec un pic précis.
 - 420 nanomètres (nm) pour les **cônes sensibles au bleu** (*short-wave sensitive: SWS*)
 - 489 nm pour les bâtonnets
 - 534 nm pour les **cônes sensibles au vert** (*medium-wave sensitive: MWS*)
 - 564 nm pour les **cônes sensibles au rouge** (*long-wave sensitive: LWS*)
- Les cônes (cellules rétiniennes sensibles aux couleurs) expriment chacun une opsine différente. Le nombre d'opsines sensibles aux couleurs varie selon les espèces
 - 3 opsines -> vision **trichromatique** (primates de l'ancien monde)
 - 2 opsines -> vision **dichromatique** (la plupart des autres mammifères)
- Noter le fort recouvrement entre les spectres de sensibilité des opsinas verte et rouge, et la faible différence entre leurs pics (31 nm) par rapport à la différence entre opsinas bleue et verte (96 nm).



Une grande famille ... de gènes

- La plus grande famille de gènes chez les métazoaires est celle des **récepteurs olfactifs**.
- génome de la souris: ~800 gènes codant pour des récepteurs olfactifs
- génome humain: ~ 400 gènes codant pour des récepteurs olfactifs
- Cette énorme famille de gènes résulte de fréquentes duplications.
- Les duplications se produisent fréquemment au sein d'un chromosome: les groupes de paralogues proches se retrouvent sur le même chromosome.
- Les mutations subséquentes provoquent des **divergences entre séquences** des paralogues, qui induisent des **différences fonctionnelles** (spécificité olfactive de chaque récepteur).



Alignment multiple

Matrice de pourcentages d'identité (opsines de mammifère, export de clustalx)

Groupe extérieur: 2 monotrèmes
(ornithorynque et echidné)

Opsines sensible au rouge et au vert

Alignement multiple des opsines de mammifère

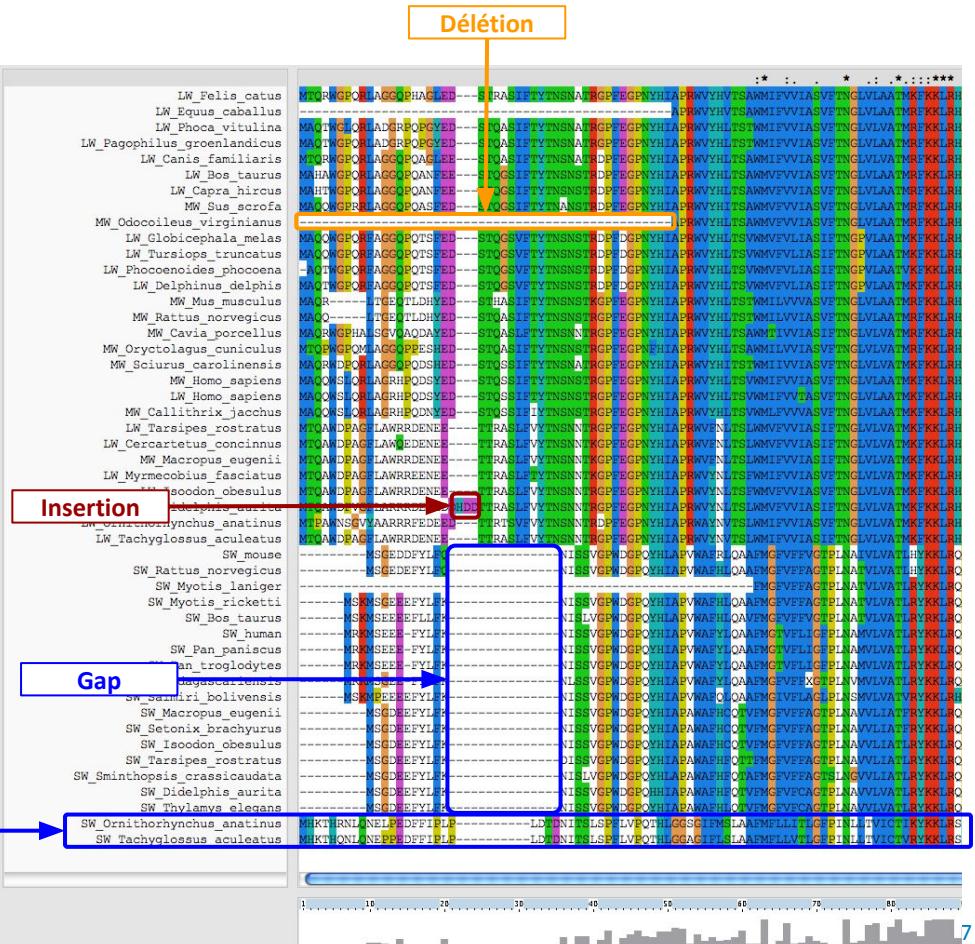
- Pour inférer un arbre phylogénétique à partir d'une famille de séquences, on part toujours d'un alignement multiple.
- Figure : première partie d'un alignement multiple entre 50 opsines de mammifères.
- A l'œil nu, on distingue déjà 2 groupes évidents.
- Dessus: opsines sensibles aux ondes moyennes (vert) ou longues (rouge)
- Dessous: opsines sensibles aux ondes courtes (bleu)



Lecture d'un alignement multiple

- Un alignement multiple consiste à aligner entre elles un ensemble de séquences similaires, de façon à maximiser les correspondances entre résidus.
- Le résultat peut être affiché de façon graphique, avec
 - Une ligne par séquence
 - Une colonne par position de l'alignement multiple
- L'alignement multiple permet de distinguer des **blocs conservés**, soit sur l'ensemble de l'alignement, soit sur un sous-ensemble des séquences.
- Les **gaps** (espacements, représentés par des "-") permettent d'ajuster les régions correspondantes à gauche et à droite.
- Contrairement aux alignements par paire, on peut dans certains cas distinguer les déletions des insertions
 - Déletion: fragment de séquence absent d'une ou de quelques séquences mais présent ailleurs
 - Insertion : fragment de séquence présent dans une ou quelques séquences mais absent ailleurs
- La présence d'un **groupe extérieur** (ici *Ornithorhynchus* et *Tachyglossus*) permet de raciner l'arbre, et de lever certaines ambiguïtés.
- Cependant, l'inférence de l'événement évolutif, déletion ou insertion, qui est à l'origine d'un gap, nécessite généralement d'interpréter les alignements en les comparant à l'arbre phylogénétique inféré.

Groupe extérieur (outgroup)



Inférence d'un arbre des molécules à partir de l'alignement multiple

Inférence d'un arbre des molécules à partir de l'alignement multiple

Il existe plusieurs approches pour inférer un arbre phylogénétique à partir d'un alignement multiple.

Les algorithmes sous-jacents ne sont pas présentés au cours, mais il est utile de savoir que ces méthodes s'appliquent dans des situations différentes. Le choix de l'algorithme dépend essentiellement du type de données.

Attention: ces méthodes peuvent donner des arbres différents, tant par la topologie (succession des branchements) que par la longueur des branches (estimation des distances évolutives).

Il est donc important de

- **savoir choisir l'algorithme en fonction des données ;**
- **pouvoir estimer la fiabilité d'un arbre phylogénétique** inféré à partir d'un jeu de séquences.

Maximum de vraisemblance (Maximum likelihood)

- Considérée comme la plus fiable des méthodes
- Permet d'estimer la longueur des branches (produit un phylogramme)
- Coûteuse en temps
- Ne permet pas de traiter des grandes familles de séquences

Neighbour Joining

- Rapide
- Permet de traiter un grand nombre de séquences
- Permet d'estimer la longueur des branches (produit un phylogramme)
- Résultats moins fiables que par la méthode du maximum de vraisemblance

Parcimonie

- Permet d'inférer les caractères ancestraux (pour les séquences: émettre une hypothèse concernant le résidu chaque noeud ancestral à chaque position de l'alignement)
- Très coûteux en temps
- Limité à un petit nombre de séquences
- Ne permet pas d'estimer la longueur des branches (produit un cladogramme)

Méthode de bootstrap pour estimer la robustesse des arbres

En phylogénie moléculaire, on infère un arbre phylogénétique à partir d'un alignement multiple (après avoir supprimé les colonnes qui comportent des gaps).

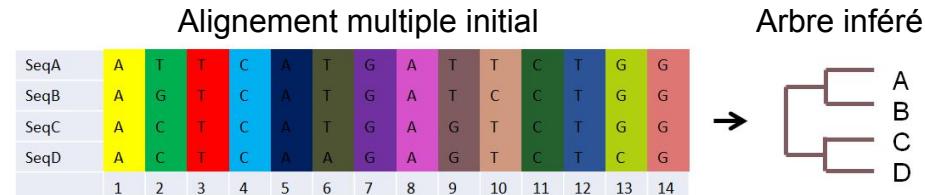
On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

Pour évaluer la **robustesse de l'inférence par rapport aux données**, on peut appliquer la méthode du **bootstrapping**.

1. Etant donné un alignement de N séquences et M colonnes, on effectue une sélection aléatoire de M colonnes **avec remise**. Chaque colonne peut donc être tirée 0, 1 ou plusieurs fois.
2. On **calcule un arbre** avec ces colonnes ré-échantillonnées.
3. On **répète l'opération** un bon nombre de fois (ex: 1000)
4. On assigne à chaque branchement de l'arbre initial une **valeur de bootstrap** = le nombre de fois où ce branchement se retrouve à l'identique dans les N arbres produits.

La valeur de bootstrap est un **indice de la robustesse** de l'arbre phylogénétique par rapport aux fluctuations d'échantillonnage.

! Ne garantit pas que l'arbre reflète l'histoire évolutive (les données peuvent être erronées ou biaisées)



Méthode de bootstrap pour estimer la robustesse des arbres

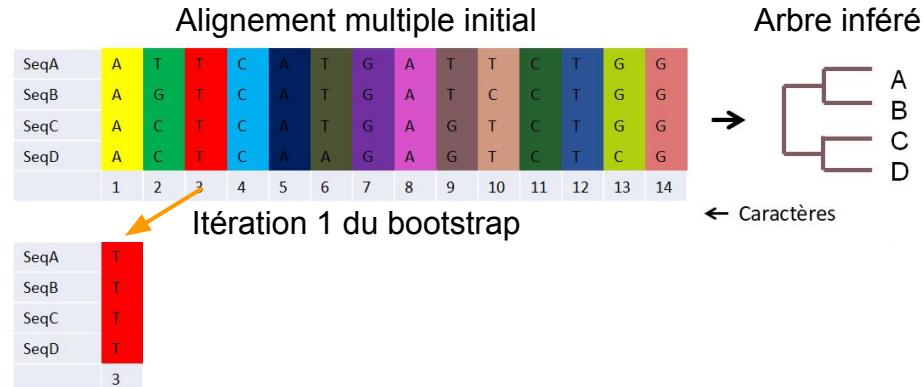
En phylogénie moléculaire, on infère un arbre phylogénétique à partir d'un alignement multiple (après avoir supprimé les colonnes qui comportent des gaps).

On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

Pour évaluer la fiabilité de l'inférence, on peut appliquer la méthode du **bootstrapping**.

1. Etant donné un alignement de N séquences et M colonnes, on effectue une sélection aléatoire de M colonnes **avec remise**. Chaque colonne peut donc être tirée 0, 1 ou plusieurs fois.
2. On **calcule un arbre** avec ces colonnes ré-échantillonnées.
3. On **répète l'opération** un bon nombre de fois (ex: 1000)
4. On assigne à chaque branchement de l'arbre initial une **valeur de bootstrap** = le nombre de fois où ce branchement se retrouve à l'identique dans les N arbres produits.

La valeur de bootstrap est un **indice de la robustesse** de l'arbre phylogénétique par rapport aux fluctuations d'échantillonnage.



Méthode de bootstrap pour estimer la robustesse des arbres

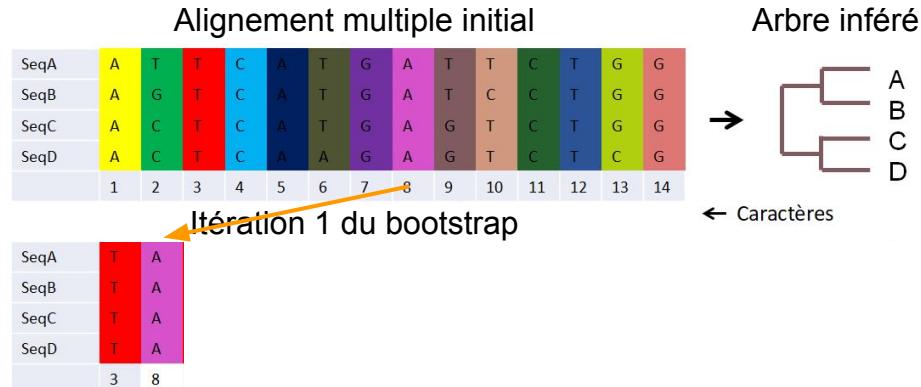
En phylogénie moléculaire, on infère un arbre phylogénétique à partir d'un alignement multiple (après avoir supprimé les colonnes qui comportent des gaps).

On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

Pour évaluer la fiabilité de l'inférence, on peut appliquer la méthode du **bootstrapping**.

1. Etant donné un alignement de N séquences et M colonnes, on effectue une sélection aléatoire de M colonnes **avec remise**. Chaque colonne peut donc être tirée 0, 1 ou plusieurs fois.
2. On **calcule un arbre** avec ces colonnes ré-échantillonnées.
3. On **répète l'opération** un bon nombre de fois (ex: 1000)
4. On assigne à chaque branchement de l'arbre initial une **valeur de bootstrap** = le nombre de fois où ce branchement se retrouve à l'identique dans les N arbres produits.

La valeur de bootstrap est un **indice de la robustesse** de l'arbre phylogénétique par rapport aux fluctuations d'échantillonnage.



Méthode de bootstrap pour estimer la robustesse des arbres

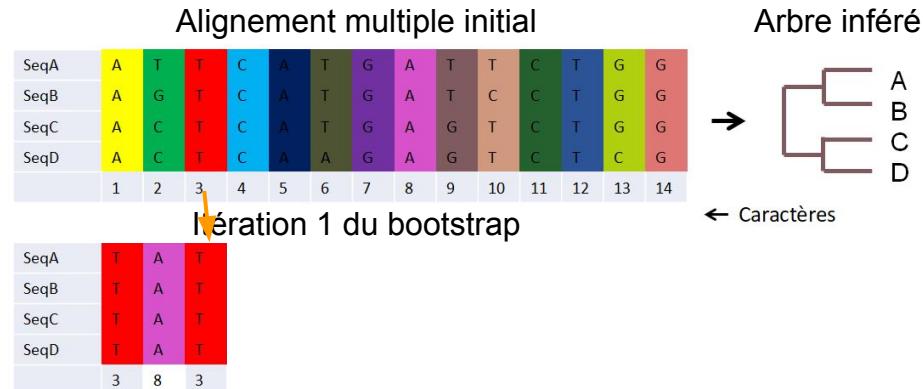
En phylogénie moléculaire, on infère un arbre phylogénétique à partir d'un alignement multiple (après avoir supprimé les colonnes qui comportent des gaps).

On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

Pour évaluer la fiabilité de l'inférence, on peut appliquer la méthode du **bootstrapping**.

1. Etant donné un alignement de N séquences et M colonnes, on effectue une sélection aléatoire de M colonnes **avec remise**. Chaque colonne peut donc être tirée 0, 1 ou plusieurs fois.
2. On **calcule un arbre** avec ces colonnes ré-échantillonnées.
3. On **répète l'opération** un bon nombre de fois (ex: 1000)
4. On assigne à chaque branchement de l'arbre initial une **valeur de bootstrap** = le nombre de fois où ce branchement se retrouve à l'identique dans les N arbres produits.

La valeur de bootstrap est un **indice de la robustesse** de l'arbre phylogénétique par rapport aux fluctuations d'échantillonnage.



Méthode de bootstrap pour estimer la robustesse des arbres

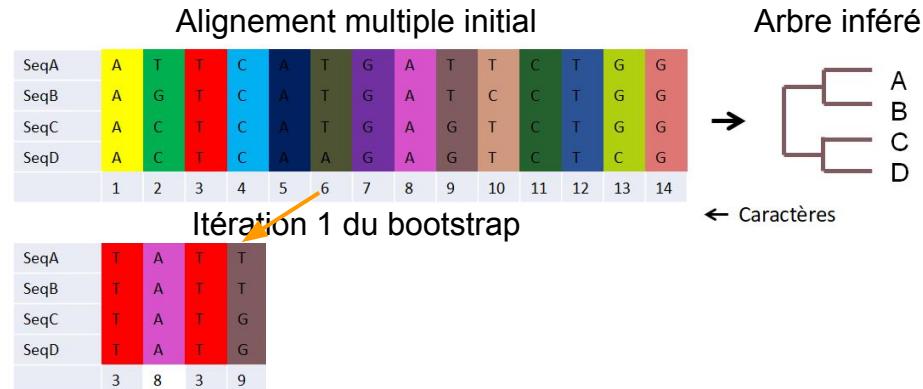
En phylogénie moléculaire, on infère un arbre phylogénétique à partir d'un alignement multiple (après avoir supprimé les colonnes qui comportent des gaps).

On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

Pour évaluer la fiabilité de l'inférence, on peut appliquer la méthode du **bootstrapping**.

1. Etant donné un alignement de N séquences et M colonnes, on effectue une sélection aléatoire de M colonnes **avec remise**. Chaque colonne peut donc être tirée 0, 1 ou plusieurs fois.
2. On **calcule un arbre** avec ces colonnes ré-échantillonnées.
3. On **répète l'opération** un bon nombre de fois (ex: 1000)
4. On assigne à chaque branchement de l'arbre initial une **valeur de bootstrap** = le nombre de fois où ce branchement se retrouve à l'identique dans les N arbres produits.

La valeur de bootstrap est un **indice de la robustesse** de l'arbre phylogénétique par rapport aux fluctuations d'échantillonnage.



Méthode de bootstrap pour estimer la robustesse des arbres

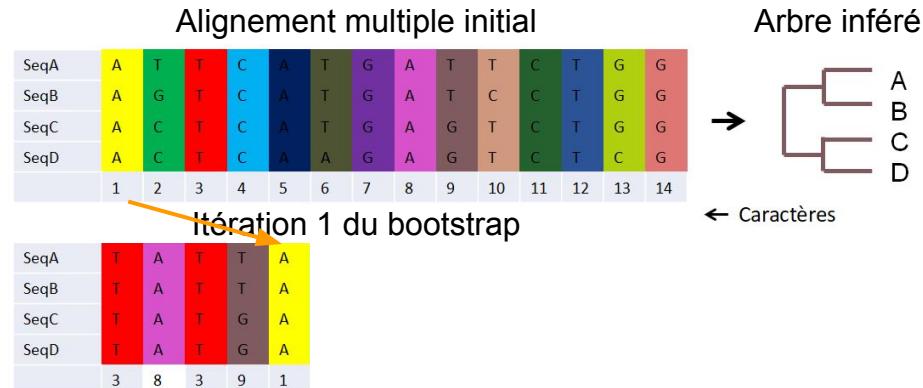
En phylogénie moléculaire, on infère un arbre phylogénétique à partir d'un alignement multiple (après avoir supprimé les colonnes qui comportent des gaps).

On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

Pour évaluer la fiabilité de l'inférence, on peut appliquer la méthode du **bootstrapping**.

1. Etant donné un alignement de N séquences et M colonnes, on effectue une sélection aléatoire de M colonnes **avec remise**. Chaque colonne peut donc être tirée 0, 1 ou plusieurs fois.
2. On **calcule un arbre** avec ces colonnes ré-échantillonnées.
3. On **répète l'opération** un bon nombre de fois (ex: 1000)
4. On assigne à chaque branchement de l'arbre initial une **valeur de bootstrap** = le nombre de fois où ce branchement se retrouve à l'identique dans les N arbres produits.

La valeur de bootstrap est un **indice de la robustesse** de l'arbre phylogénétique par rapport aux fluctuations d'échantillonnage.



Méthode de bootstrap pour estimer la robustesse des arbres

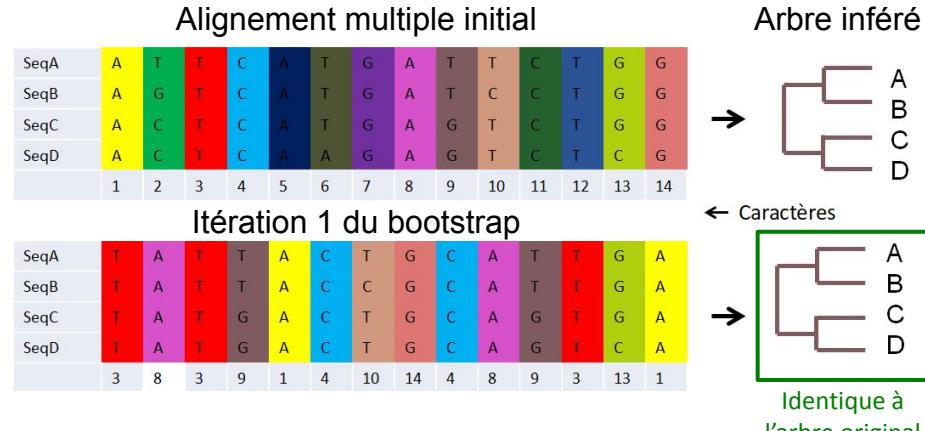
En phylogénie moléculaire, on infère un arbre phylogénétique à partir d'un alignement multiple (après avoir supprimé les colonnes qui comportent des gaps).

On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

Pour évaluer la fiabilité de l'inférence, on peut appliquer la méthode du **bootstrapping**.

1. Etant donné un alignement de N séquences et M colonnes, on effectue une sélection aléatoire de M colonnes **avec remise**. Chaque colonne peut donc être tirée 0, 1 ou plusieurs fois.
2. On **calcule un arbre** avec ces colonnes ré-échantillonnées.
3. On **répète l'opération** un bon nombre de fois (ex: 1000)
4. On assigne à chaque branchement de l'arbre initial une **valeur de bootstrap** = le nombre de fois où ce branchement se retrouve à l'identique dans les N arbres produits.

La valeur de bootstrap est un **indice de la robustesse** de l'arbre phylogénétique par rapport aux fluctuations d'échantillonnage.



Méthode de bootstrap pour estimer la robustesse des arbres

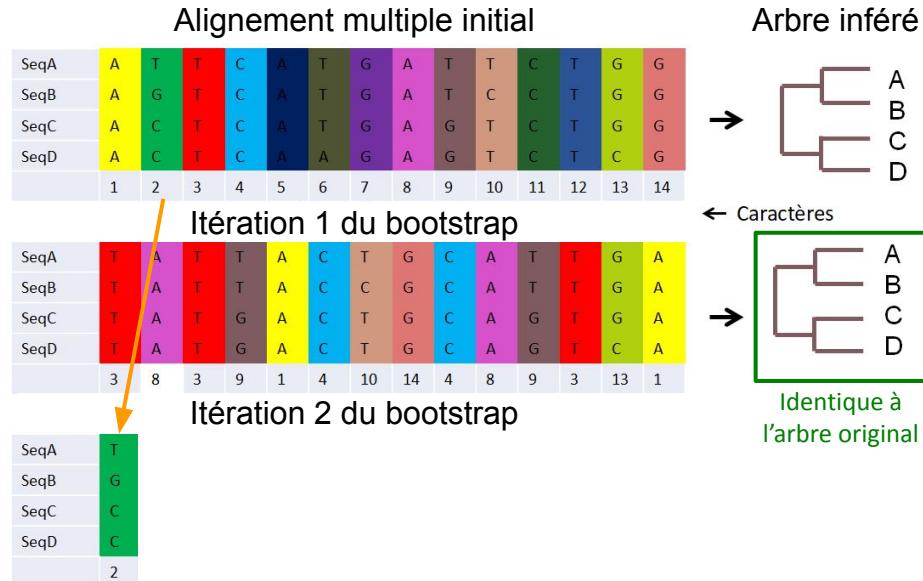
En phylogénie moléculaire, on infère un arbre phylogénétique à partir d'un alignement multiple (après avoir supprimé les colonnes qui comportent des gaps).

On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

Pour évaluer la fiabilité de l'inférence, on peut appliquer la méthode du **bootstrapping**.

1. Etant donné un alignement de N séquences et M colonnes, on effectue une sélection aléatoire de M colonnes **avec remise**. Chaque colonne peut donc être tirée 0, 1 ou plusieurs fois.
2. On **calcule un arbre** avec ces colonnes ré-échantillonnées.
3. On **répète l'opération** un bon nombre de fois (ex: 1000)
4. On assigne à chaque branchement de l'arbre initial une **valeur de bootstrap** = le nombre de fois où ce branchement se retrouve à l'identique dans les N arbres produits.

La valeur de bootstrap est un **indice de la robustesse** de l'arbre phylogénétique par rapport aux fluctuations d'échantillonnage.



Méthode de bootstrap pour estimer la robustesse des arbres

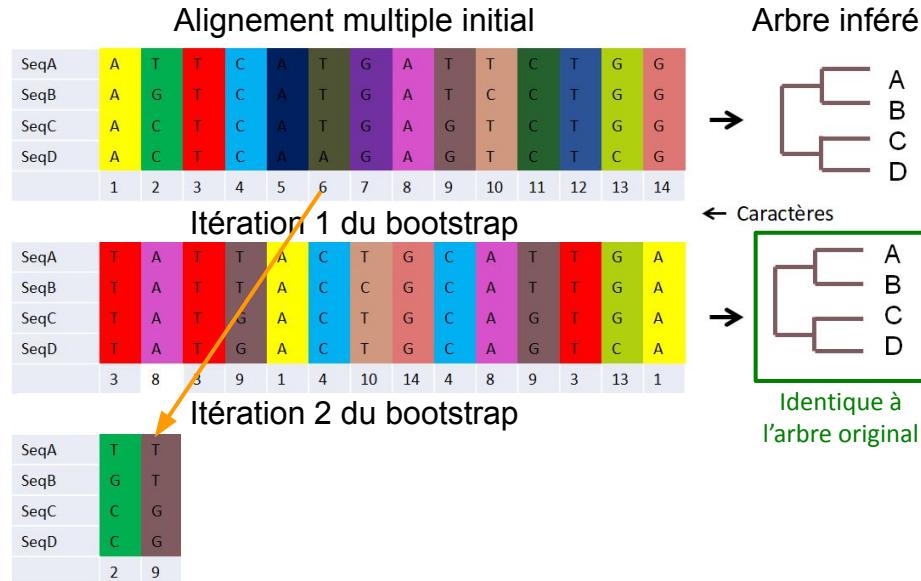
En phylogénie moléculaire, on infère un arbre phylogénétique à partir d'un alignement multiple (après avoir supprimé les colonnes qui comportent des gaps).

On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

Pour évaluer la fiabilité de l'inférence, on peut appliquer la méthode du **bootstrapping**.

1. Etant donné un alignement de N séquences et M colonnes, on effectue une sélection aléatoire de M colonnes **avec remise**. Chaque colonne peut donc être tirée 0, 1 ou plusieurs fois.
2. On **calcule un arbre** avec ces colonnes ré-échantillonnées.
3. On **répète l'opération** un bon nombre de fois (ex: 1000)
4. On assigne à chaque branchement de l'arbre initial une **valeur de bootstrap** = le nombre de fois où ce branchement se retrouve à l'identique dans les N arbres produits.

La valeur de bootstrap est un **indice de la robustesse** de l'arbre phylogénétique par rapport aux fluctuations d'échantillonnage.



Méthode de bootstrap pour estimer la robustesse des arbres

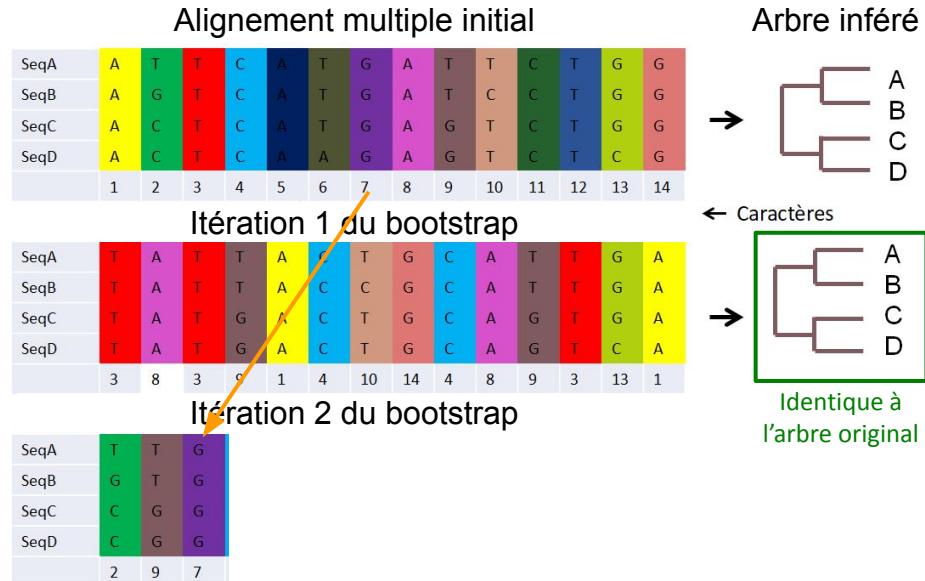
En phylogénie moléculaire, on infère un arbre phylogénétique à partir d'un alignement multiple (après avoir supprimé les colonnes qui comportent des gaps).

On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

Pour évaluer la fiabilité de l'inférence, on peut appliquer la méthode du **bootstrapping**.

1. Etant donné un alignement de N séquences et M colonnes, on effectue une sélection aléatoire de M colonnes **avec remise**. Chaque colonne peut donc être tirée 0, 1 ou plusieurs fois.
2. On **calcule un arbre** avec ces colonnes ré-échantillonnées.
3. On **répète l'opération** un bon nombre de fois (ex: 1000)
4. On assigne à chaque branchement de l'arbre initial une **valeur de bootstrap** = le nombre de fois où ce branchement se retrouve à l'identique dans les N arbres produits.

La valeur de bootstrap est un **indice de la robustesse** de l'arbre phylogénétique par rapport aux fluctuations d'échantillonnage.



Méthode de bootstrap pour estimer la robustesse des arbres

En phylogénie moléculaire, on infère un arbre phylogénétique à partir d'un alignement multiple (après avoir supprimé les colonnes qui comportent des gaps).

On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

Pour évaluer la fiabilité de l'inférence, on peut appliquer la méthode du **bootstrapping**.

1. Etant donné un alignement de N séquences et M colonnes, on effectue une sélection aléatoire de M colonnes **avec remise**. Chaque colonne peut donc être tirée 0, 1 ou plusieurs fois.
2. On **calcule un arbre** avec ces colonnes ré-échantillonnées.
3. On **répète l'opération** un bon nombre de fois (ex: 1000)
4. On assigne à chaque branchement de l'arbre initial une **valeur de bootstrap** = le nombre de fois où ce branchement se retrouve à l'identique dans les N arbres produits.

La valeur de bootstrap est un **indice de la robustesse** de l'arbre phylogénétique par rapport aux fluctuations d'échantillonnage.

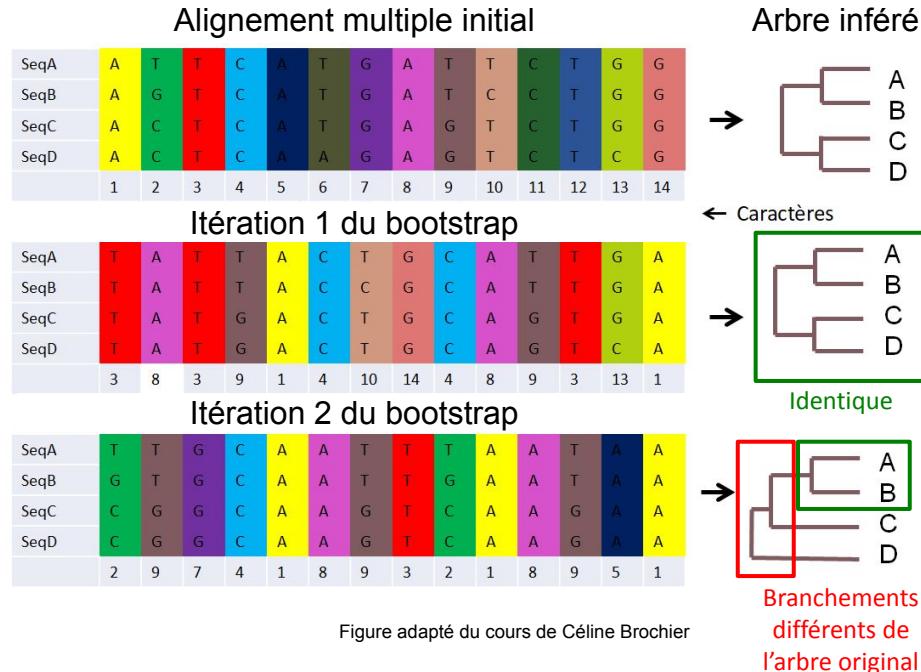


Figure adapté du cours de Céline Brochier

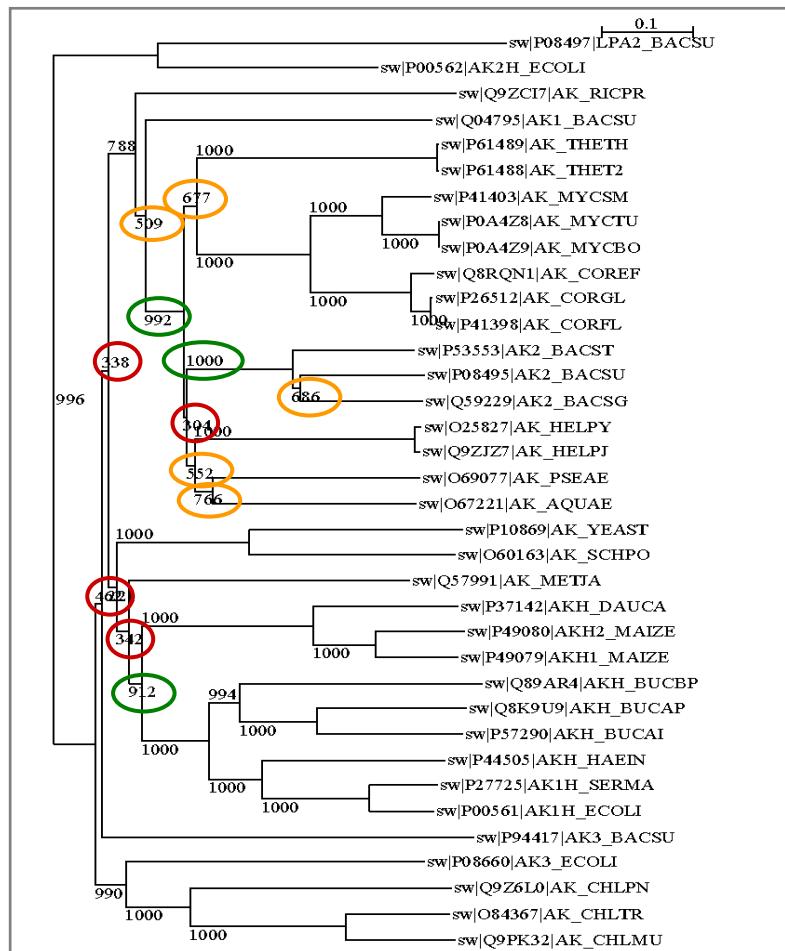
Bootstrapping

- Sur un arbre phylogénétique, une **valeur de bootstrap** est assignée à chaque branchement pour indiquer nombre de fois où ce branchement se retrouve à l'identique dans les N arbres de bootstrap.
- Valeur élevée** → branchement **robuste aux fluctuations d'échantillonnage**
- Valeur faible** → branchement sensible aux fluctuations d'échantillonnage
 - Exemple : 338/1000 signifie que ce branchement n'est présent que dans ~1/3 des bootstraps ; il dépend donc fortement d'un sous-ensemble des colonnes plutôt que de représenter l'alignement complet.

Attention ! La valeur de bootstrap nous informe sur la robustesse de l'arbre inféré par rapport aux données, mais ceci n'est en aucun cas une garantie de pertinence de ces données par rapport à l'histoire évolutive réelle ("arbre vrai").

Problèmes potentiels

- Biais d'échantillonnage**
- Erreurs dans les données**



Phylogénomique : retracer l'évolution des espèces à partir des séquences génomiques

Phylogénomique

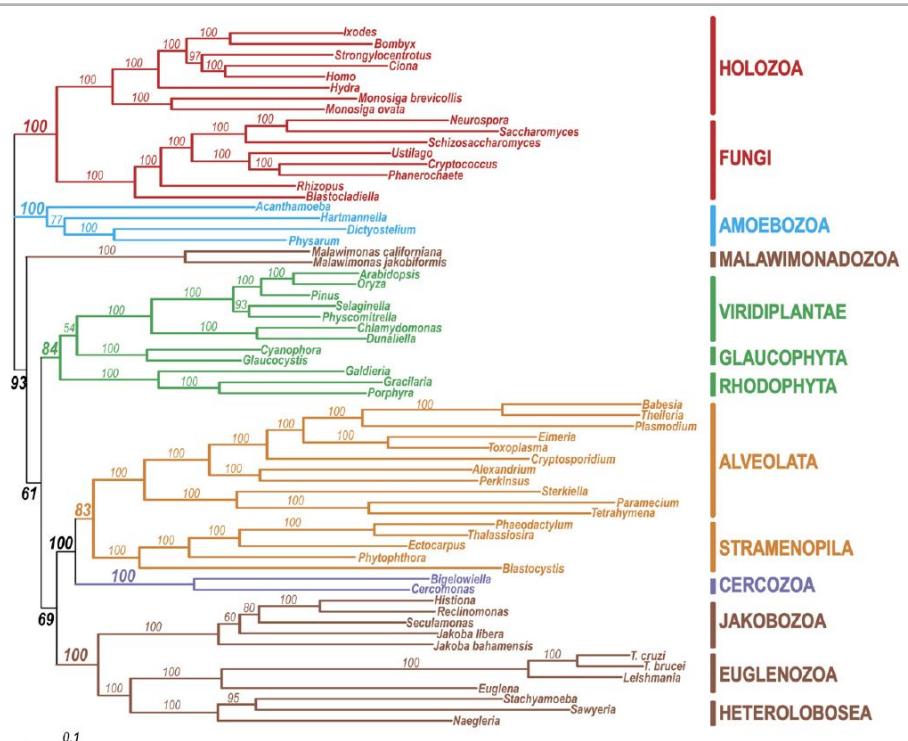


Figure 1. Maximum-Likelihood Tree of Eukaryotes

The tree includes 64 species and is based on 143 concatenated nucleus-encoded proteins (31,604 amino acid positions). Numbers indicate support values of RaxML analysis (100 replicates) with the WAG + F + I' model. Posterior probabilities obtained in the Bayesian Inference with MrBayes are 1.0 for all branches. The scale bar denotes the estimated number of amino acid substitutions per site. The tree was rooted according to a gene fusion [13, 16].

- En phylogénie moléculaire, une approche classique consiste à se concentrer sur un gène considéré comme représentatif de l'évolution de la famille de gènes homologues, et à construire un arbre sur base de la divergence de séquence de ce gène.
- Ces approches peuvent maintenant être généralisées en comparant les séquences de **plusieurs centaines de gènes ou de protéines**
- Elles permettent d'inférer des phylogénies entre organismes très éloignés (règnes différents), et d'établir ainsi des scénarios concernant les premières étapes de la diversification des êtres vivants.

Exemple (figure de gauche)

- Arbre basé sur 143 familles de protéines
- Grands groupes d'eucaryotes

Source: Rodríguez-Ezpeleta et al. Curr Biol (2007) vol. 17 (16) pp. 1420-5

Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans.

Suffit-il de manger des insectes pour être un Insectivore ?

Au 20^e siècle, les **tenrecs** étaient considérés comme une famille d'insectivores, présentant une ressemblance morphologique mais des différences anatomiques importantes par rapport aux hérissons.

Hérisson

classe	Mammifères
ordre	Insectivores
famille	Erinacéidés
genre	<i>Erinaceus europaeus</i>
et espèce	<i>et autres</i>

Tenrec

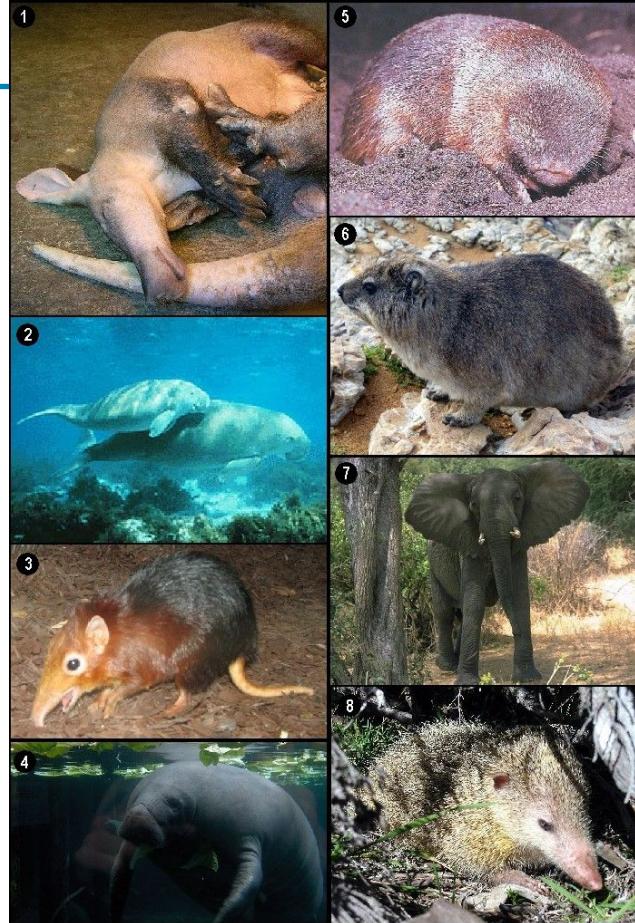
classe	Mammifères
ordre	Insectivores
famille	Tenrécidés
genres	<i>Centetes ecaudatus</i>
et espèces	(tenrec commun)
	<i>Hemicentetes semispinosus</i>
	(tenrec à bandes ou tenrec strié)
	<i>Limnogale mergulus</i>
	(tenrec à pieds palmés)
	<i>Microgale longicaudata</i>
	(tenrec à longue queue)
	<i>Oryzorictes hova</i>
	(tenrec des rizières)
	<i>Setifer setosus</i>
	(tenrec-hérisson)
	<i>et autres</i>



Les Afrotheria

Quels sont les plus proches cousins du tenrec ?

- A la fin des années 1990, l'inférence phylogénétique est mise à contribution pour identifier les premiers moments de la radiation des mammifères.
- Cette analyse révèle qu'une série d'espèces qu'il était extrêmement difficile de classer proviennent d'un branchement précoce au sein des mammifères.
- C'est notamment le cas du tenrec (8, 8b) dont l'apparence rappelle celle du hérisson, mais l'anatomie en diffère fortement.
- C'est sur cette base qu'a été constitué le **super-ordre des Afrotheria**, qui rassemble une série de mammifères aux morphologies les plus diverses : Autres Afrotheria: éléphant, lacentin, dugong, oryctérope, taupe dorée, daman, ...

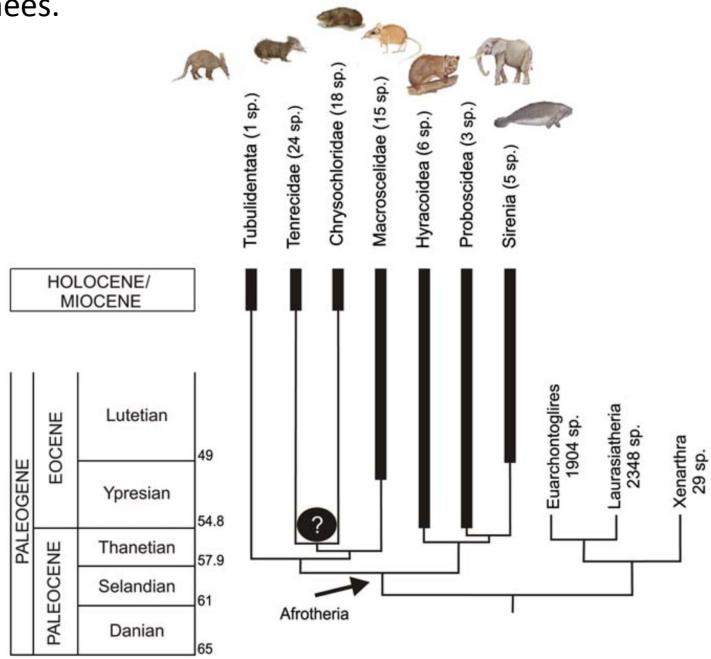


- Murphy, W. J. et al. Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenetics. *Science* 294, 2348–2351 (2001). doi.org/10.1126/science.1067179
- Tabuce, R., Asher, R. J. & Lehmann, T. Afrotherian mammals: a review of current data. *mammalia* 72, (2008). doi.org/10.1515/MAMM.2008.004

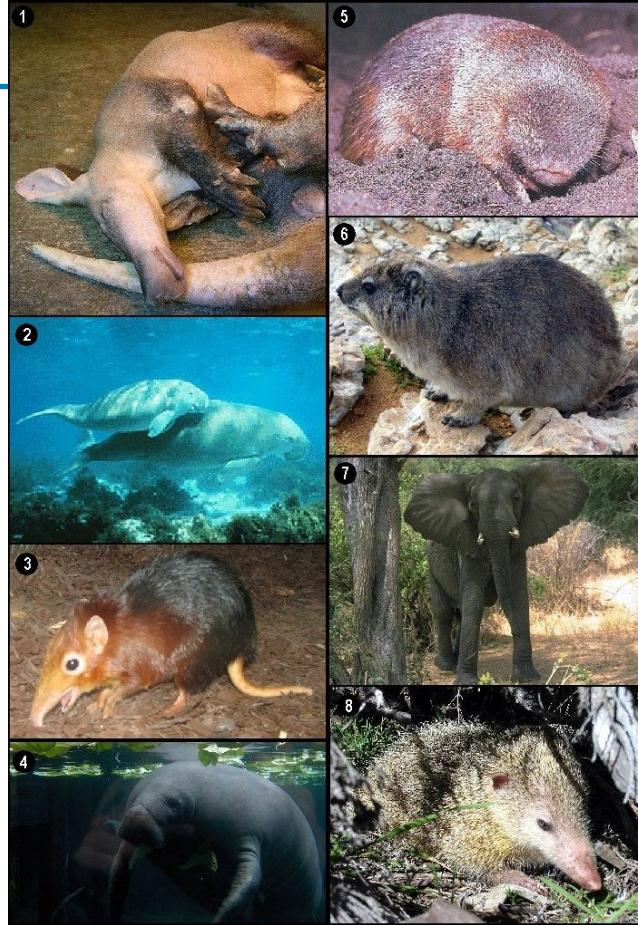
1. Oryctérope du Cap; 2. Dugongs; 3. Macroscéléide de Peters ; 4. Lamantin; 5. Taupe dorée; 6. Daman du Cap; 7. Éléphant de savane d'Afrique ; 8. Tangue ("tailless tenrec"); 8, 8b: Tenrec
<http://upload.wikimedia.org/wikipedia/commons/0/01/Kleiner-igeltanrek-a.jpg>

Les Afrotheria

- Sur base d'analyse de l'ADN, on estime que la divergence entre Afrotheria remonte à 60-55 millions d'années.



- Murphy, W. J. et al. Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenetics. *Science* 294, 2348–2351 (2001). doi.org/10.1126/science.1067179
- Tabuce, R., Asher, R. J. & Lehmann, T. Afrotherian mammals: a review of current data. *mammalia* 72, (2008). doi.org/10.1515/MAMM.2008.004



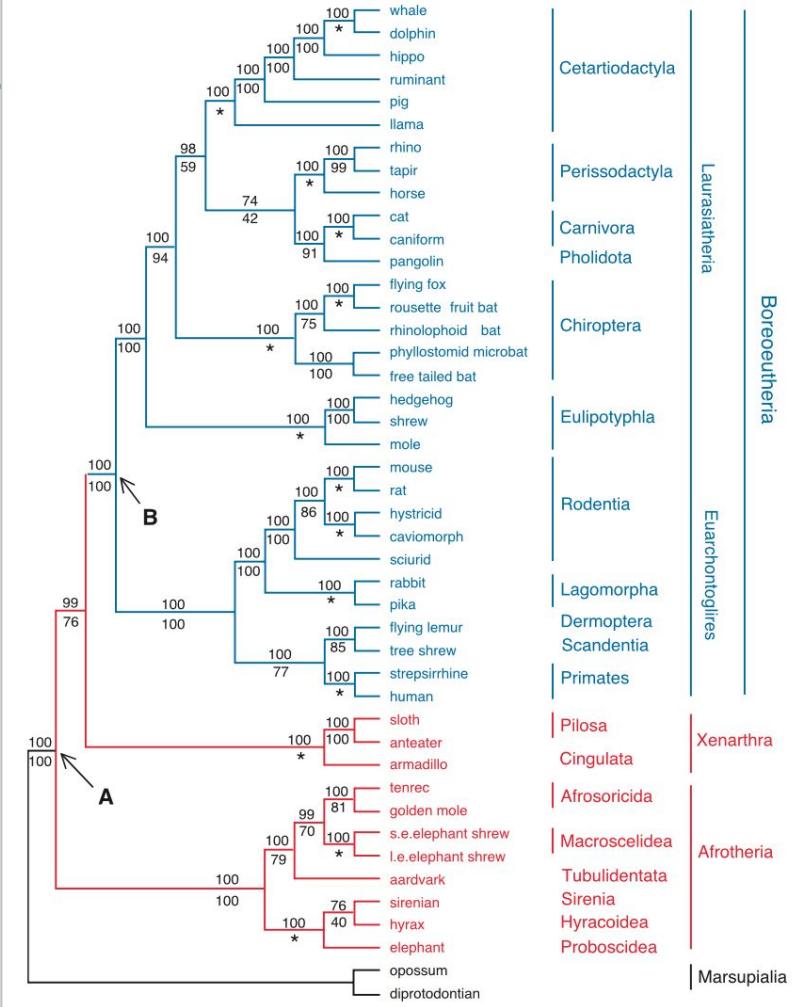
1. Oryctérope du Cap; 2. Dugongs; 3. Macroscéléide de Peters ; 4. Lamantin; 5. Taupe dorée; 6. Daman du Cap; 7. Éléphant de savane d'Afrique ; 8. Tangue ("tailless tenrec"); 8, 8b: Tenrec
<http://upload.wikimedia.org/wikipedia/commons/0/01/Kleiner-igeltanrek-a.jpg>

Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenetics

William J. Murphy,^{1,*} Eduardo Eizirik,^{1,2*} Stephen J. O'Brien,^{1†}
 Ole Madsen,³ Mark Scally,^{4,5} Christophe J. Douady,^{4,5}
 Emma Teeling,^{4,5} Oliver A. Ryder,⁶ Michael J. Stanhope,^{5,7}
 Wilfried W. de Jong,^{3,8} Mark S. Springer^{4†}

Molecular phylogenetic studies have resolved placental mammals into four major groups, but have not established the full hierarchy of interordinal relationships, including the position of the root. The latter is critical for understanding the early biogeographic history of placentals. We investigated placental phylogeny using Bayesian and maximum-likelihood methods and a 16.4-kilobase molecular data set. Interordinal relationships are almost entirely resolved. The basal split is between Afrotheria and other placentals, at about 103 million years, and may be accounted for by the separation of South America and Africa in the Cretaceous. Crown-group Eutheria may have their most recent common ancestry in the Southern Hemisphere (Gondwana).

Fig. 1. Phylogeny of living placental mammals reconstructed using a Bayesian phylogenetic approach. An identical topology was obtained with maximum likelihood [$-\ln L = 211110.54$; see (15) for methodological details]. The number above each branch refers to the Bayesian posterior probability (shown as percentages; i.e., 95 represents a posterior probability of 0.95) of the node derived from 26,250 MCMC sampled trees on the basis of the complete 16.4-kb data. Additional analyses with the full data set and with data sets that varied taxon sampling (i.e., jackknifing single outgroup taxa) and character sampling (nuclear only and nuclear coding loci only) produced similarly high posterior probabilities (15). Values below branches represent percent support in maximum likelihood ($\text{GTR} + \Gamma + I$) nonparametric bootstrap. An asterisk indicates nodes constrained in the ML nonparametric bootstrap analysis. (A) Bifurcation between Afrotheria and Xenarthra + Boreoeutheria at approximately 103 million years, which corresponds to the vicariant event that separated Africa and South America (Fig. 2B). (B) Branch where dispersal from South America to Laurasia is hypothesized to have occurred (15). Blue, monophyletic Northern Hemisphere group (i.e., Boreoeutheria); red, paraphyletic Southern Hemisphere group (i.e., Xenarthra + Afrotheria); black, outgroups.



Retracer l'origine de SARS-CoV-2 dans les génomes des coronavirus

La publication du génome de SARS-CoV-2

3 février 2020: publication du **génome complet de SARS-CoV-2**

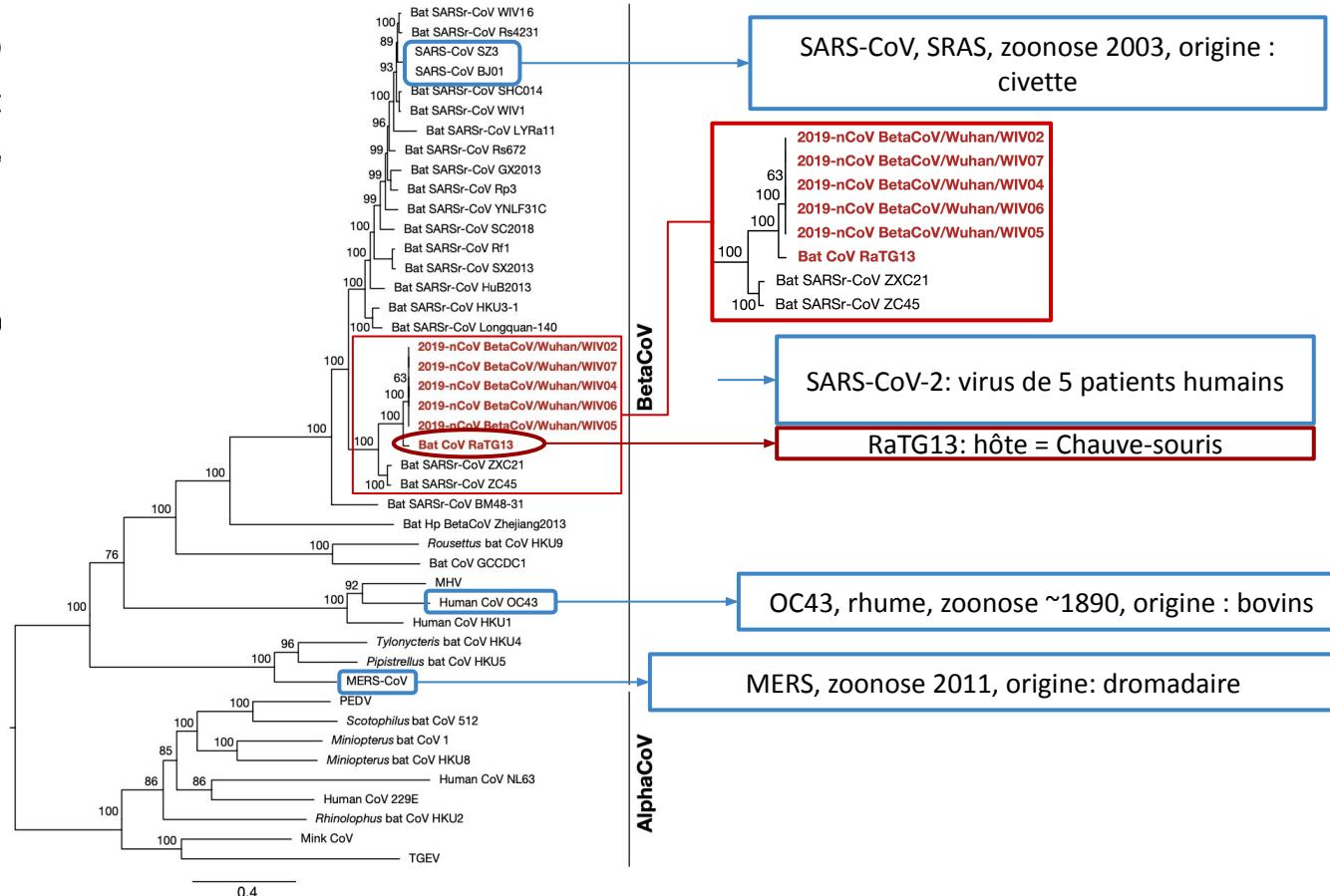
Recherche de virus similaires dans les bases de données de séquence. Les virus les plus proches sont des virus de chauves-souris (Bat CoV ZC45)

Dans le même article, les auteurs décrivent un **nouveau génome de virus de chauve-souris : RaTG13**

- virus connu le plus proche de SARS-CoV-2
- 96.2% d'identité sur l'ensemble du génome

Notes:

- Ce taux d'identité correspond à une divergence évolutive de 4 à 7 décennies. Il ne s'agit donc pas d'un parent direct de SARS-CoV-2 mais d'un cousin très éloigné.



Un virus synthétique avec des bouts de HIV ?

Le 17 avril 2020, le Professeur Luc Montagnier, Prix Nobel de médecine pour sa contribution à la découverte du HIV (le virus responsable du SIDA), défraie la chronique en annonçant sur plusieurs médias (Pourquoi Docteur, CNEWS) que le génome du coronavirus SARS-CoV-2, agent de la pandémie COVID-19, comporte quatre fragments de séquences provenant du HIV. De plus, il affirme que la présence de ces séquences ne résulte pas d'une recombinaison naturelle (fréquente chez les virus) ou d'un accident, mais d'un vrai travail d'ingénieur, effectué intentionnellement, vraisemblablement dans le cadre de recherches visant à développer des vaccins contre le HIV.

Pour appuyer sa théorie, Luc Montagnier cite deux études :

- le travail d'un collègue mathématicien, Jean-Claude Perez, qui "a fouillé les moindres détails de la séquence",
- une analyse des séquences génomiques et protéiques des coronavirus préalablement publiée par une équipe indienne, qui a, selon lui, "été forcée de rétracter" sa publication.

Selon Luc Montagnier, le virus covid19 est une manipulation humaine (17 avril 2020)

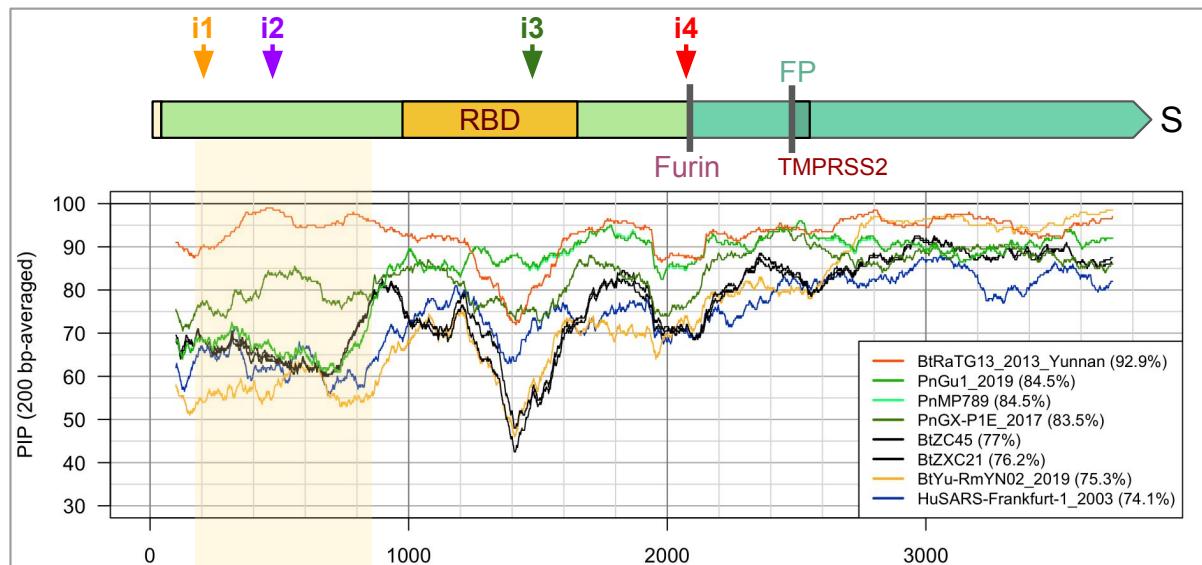
<https://www.youtube.com/watch?v=qSWCLHIOiMo>
(devenue inaccessible depuis lors)

"Je suis arrivé à la conclusion qu'il y avait eu une manipulation de ce virus. [...] Il y a un modèle qui est évidemment le virus classique, et là c'était un modèle venant de la chauve-souris, et là, à ce modèle on a par-dessus ajouté les séquences du VIH, du SIDA. ... Non, ce n'est pas naturel, c'était un travail de professionnel, de biologiste moléculaire, très minutieux, on peut dire d'horloger, au niveau des séquences. Dans quel but ce n'est pas clair. Mon travail c'est d'exposer les faits, c'est tout. Je n'accuse personne, je ne sais pas qui a fait ça et pourquoi. La possibilité c'est qu'on a voulu faire un vaccin contre le SIDA. Donc on a pris des petites séquences du virus [HIV] et on les a installées dans la séquence plus grande du coronavirus. [...] Il y a quand même une volonté d'étouffement, nous ne sommes pas les premiers. Un groupe de chercheurs indiens très renommés avaient publié la même chose, on les a forcés à rétracter. Si vous regardez leur publication vous voyez une grande bande "annulé". "

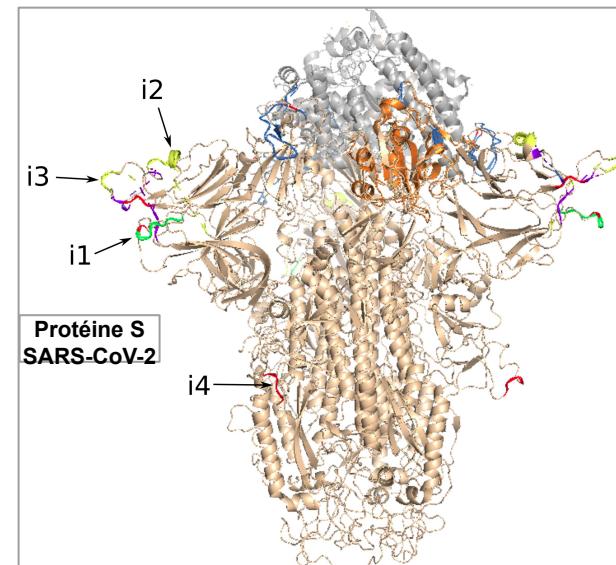
Quatre insertions dans le gène S de SARS-CoV-2

- Les flèches indiquent la position des 4 insertions sur le gène S (gauche) et sur la protéine spicule (droite).
- Les 3 premières sont situées à l'extérieur de la protéine, dans des régions “exposées” (boucles).

Structure et profil de conservation du gène S



Structure de la protéine spike



Alignement de séquences de SARS-CoV-2 sur le génome du HIV

Haut: fragment le plus significatif de l'alignement de la séquence du gène S sur le génome du VIH. Noter le score Expect = 7.5. Or ce score n'est considéré significatif que s'il est nettement inférieur à 1.

Bas: fragment le plus significatif de l'alignement d'une séquence aléatoire sur le génome du VIH. Noter le score Expect = 2.1, supérieur à 1 et donc non-significatif (comme on s'y attend, puisque la séquence est aléatoire).

Conclusion: l'alignement sur lequel s'appuient Perez et Luc Montagnier correspond à ce qu'on s'attend à trouver par hasard en alignant des séquences de cette taille.

→ les similarités ne permettent pas d'affirmer qu'on a inséré des séquences de HIV dans le génome de SARS-CoV-2.

HIV-1 isolate 19828.PPH11 from Netherlands envelope glycoprotein (env) gene, partial cds				
Sequence ID: HQ644953.1		Length: 1143	Number of Matches: 1	Range 1: 967 to 994
Score	Expect	Identities	Gaps	Strand
38.3 bits(41)	7.5	25/28(89%)	0/28(0%)	Plus/Plus
Query 86	AATGGTACTAAGAGGTTGATAACCTG	113		
Sbjct 967	AATGGTACTAAAGGTTAGATAACACTG	994		

Des insertions bizarres?

Figure de Pradhan et al (2020), initialement déposée sur bioRxiv mais ensuite retirée par les auteurs suite aux critiques qui leur avaient été adressées.

- Ce qu'ils qualifient d' "alignement multiple" est en fait un alignement d'une paire de séquences (avec une troisième ligne pour le consensus).
- L'alignement est inconsistante avec un alignement multiple obtenu avec d'autres séquences homologues.
- L'alignement par paire ne permet en aucun cas d'évaluer l'ancienneté de l'événement, ni de savoir s'il s'agit d'une insertion chez SARS-CoV-2, ou chez un ancêtre plus éloigné, ou bien d'une délétion dans la lignée du virus SRAS de 2003.

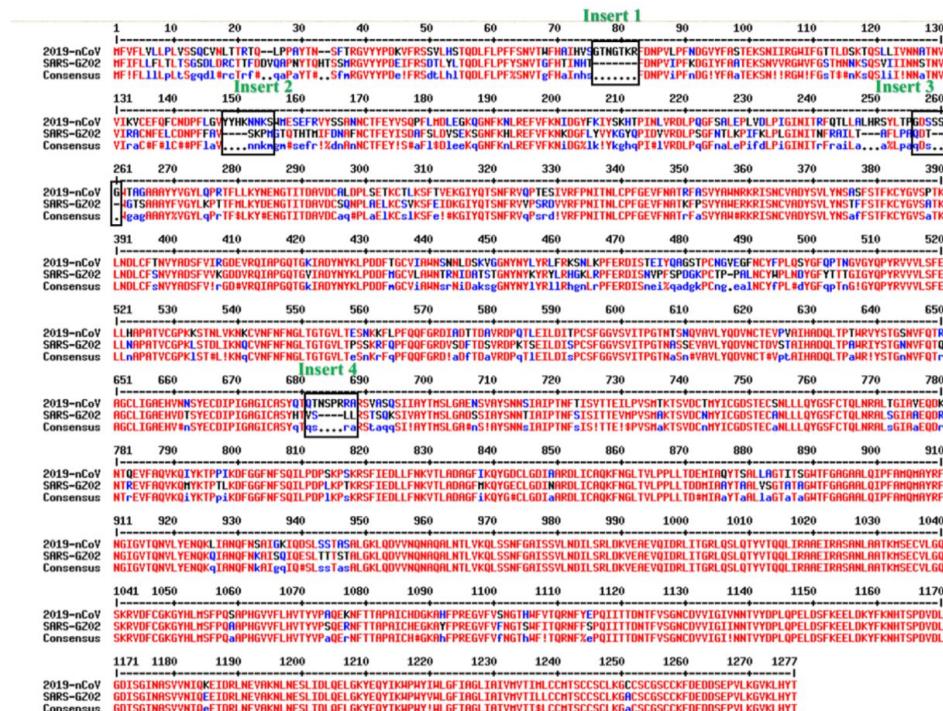
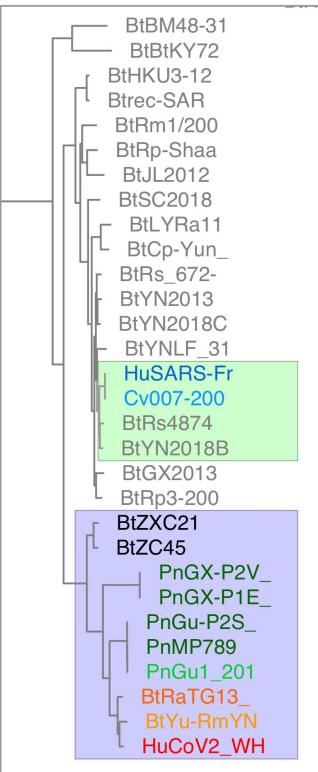


Figure 2: Multiple sequence alignment between spike proteins of 2019-nCoV and SARS. The sequences of spike proteins of 2019-nCoV (Wuhan-HU-1, Accession NC_045512) and of SARS CoV (GZ02, Accession AY390556) were aligned using MultiAlin software. The sites of difference are highlighted in boxes.

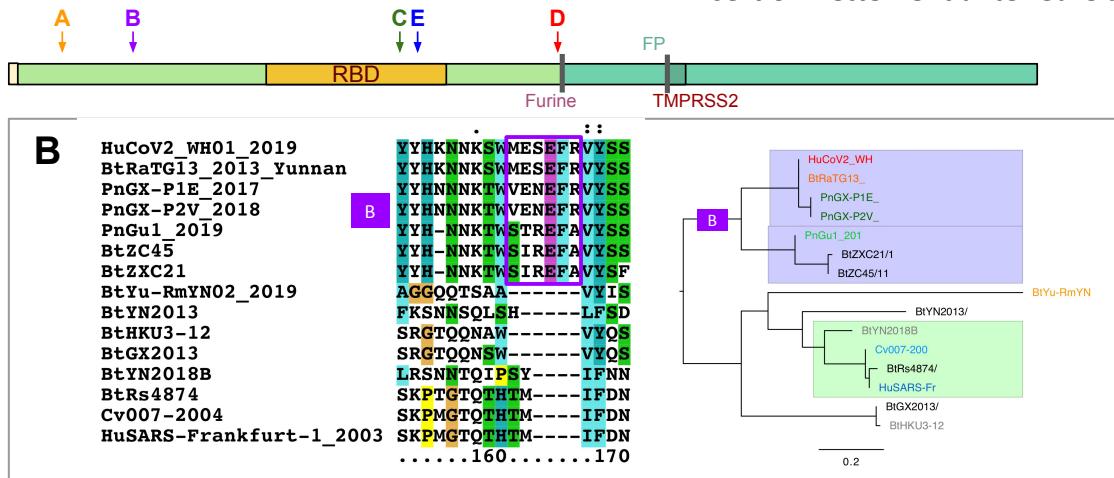
Insertion B : partagée entre plusieurs espèces de coronavirus

Arbre des génomes



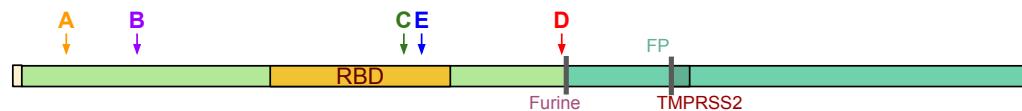
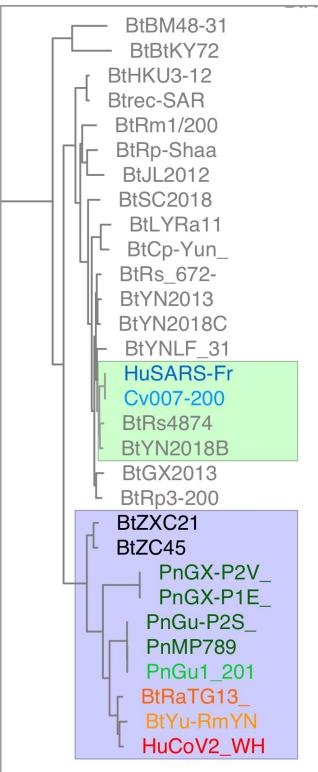
- B : insertion présente chez autres virus de chauve-souris + pangolin
- Séquences identiques entre SARS-CoV-2 et RaTG13, malgré 40 à 70 ans de divergence
- Séquences très proches entre virus de humain, de chauve-souris, et de pangolin
- Substitution secondaire, au sein de l'insertion :
 - MES dans 3 génomes
 - STR dans 3 autres

→ Insertion nettement antérieure à l'émergence de SARS-CoV-2

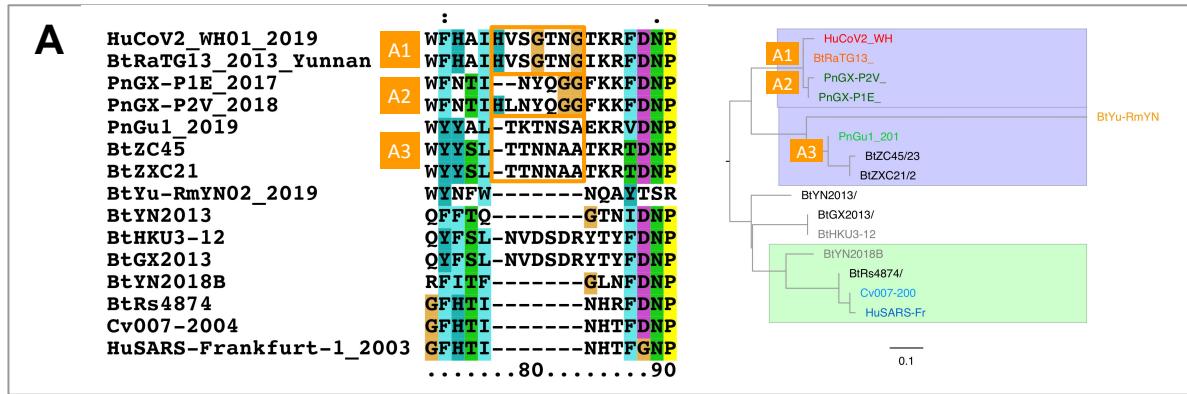


Insertion A : partagée entre SARS-CoV-2 et RaTG13

Arbre des génomes

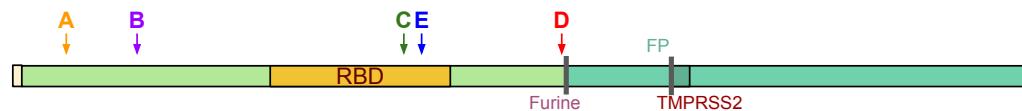
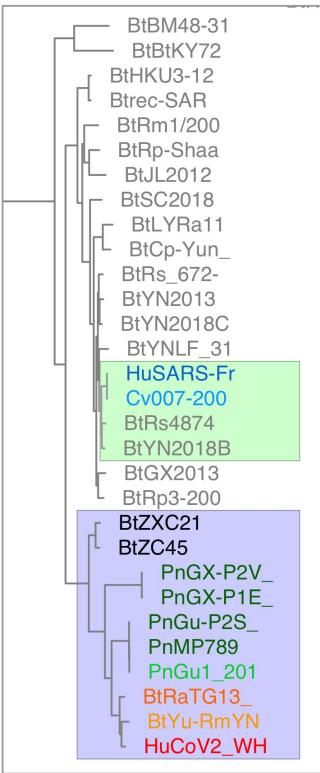


- A1 : Insertion identique entre SARS-CoV-2 et RaTG13, malgré 40 à 70 ans de divergence
→ **Insertion nettement antérieure à l'émergence de SARS-CoV-2**
- A2 , A3 : Insertions au même site, dans des sous-groupes séparés → événements indépendants



Insertion C

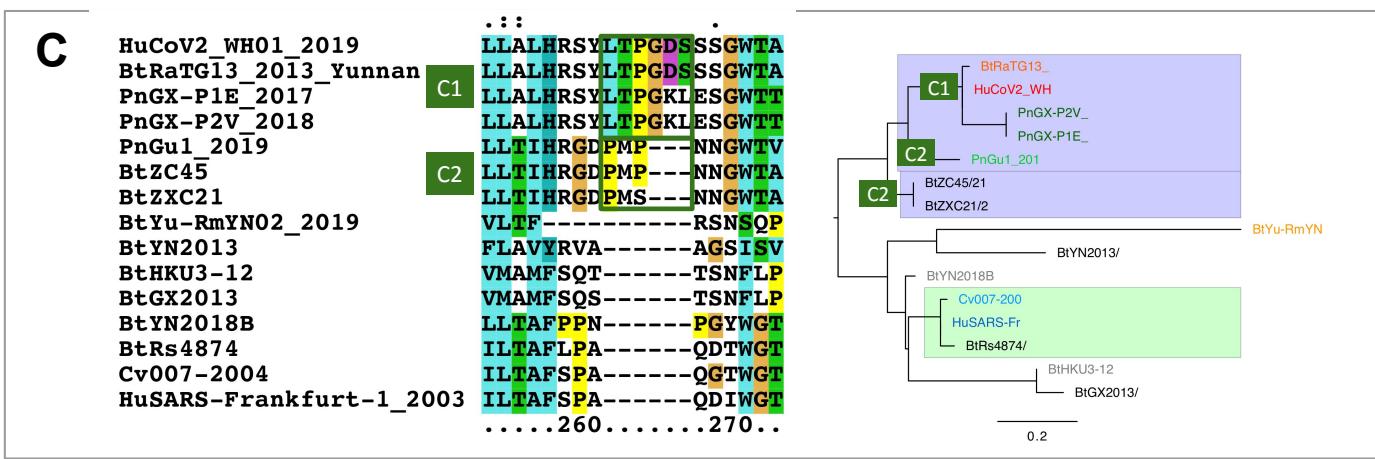
Arbre des génomes



- C1 : insertion partagée entre 4 des souches de virus apparentés, parmi les souches analysées

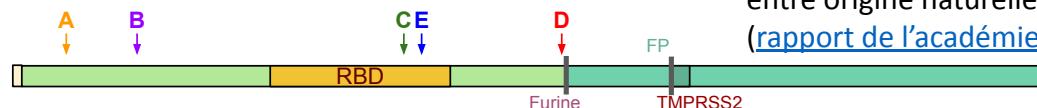
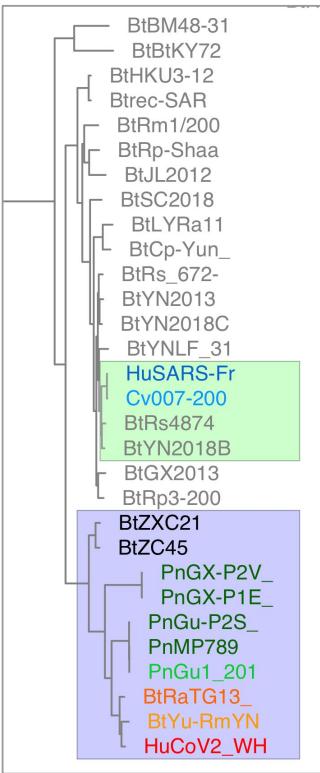
→ Insertion nettement antérieure à l'émergence de SARS-CoV-2

- C2 : insertion indépendante à la même position chez d'autres virus

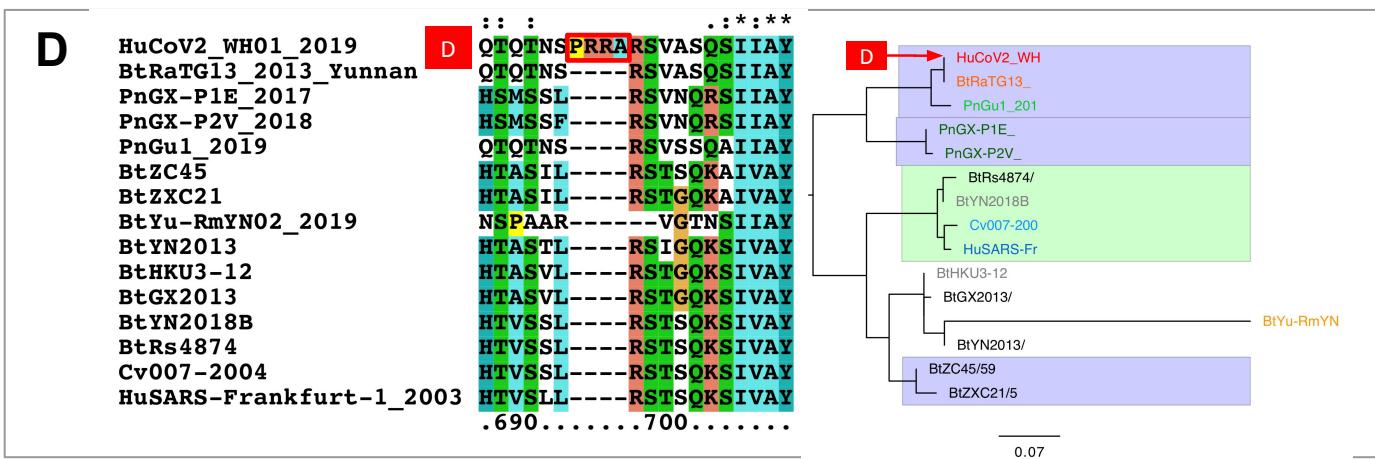


Insertion D

Arbre des génomes

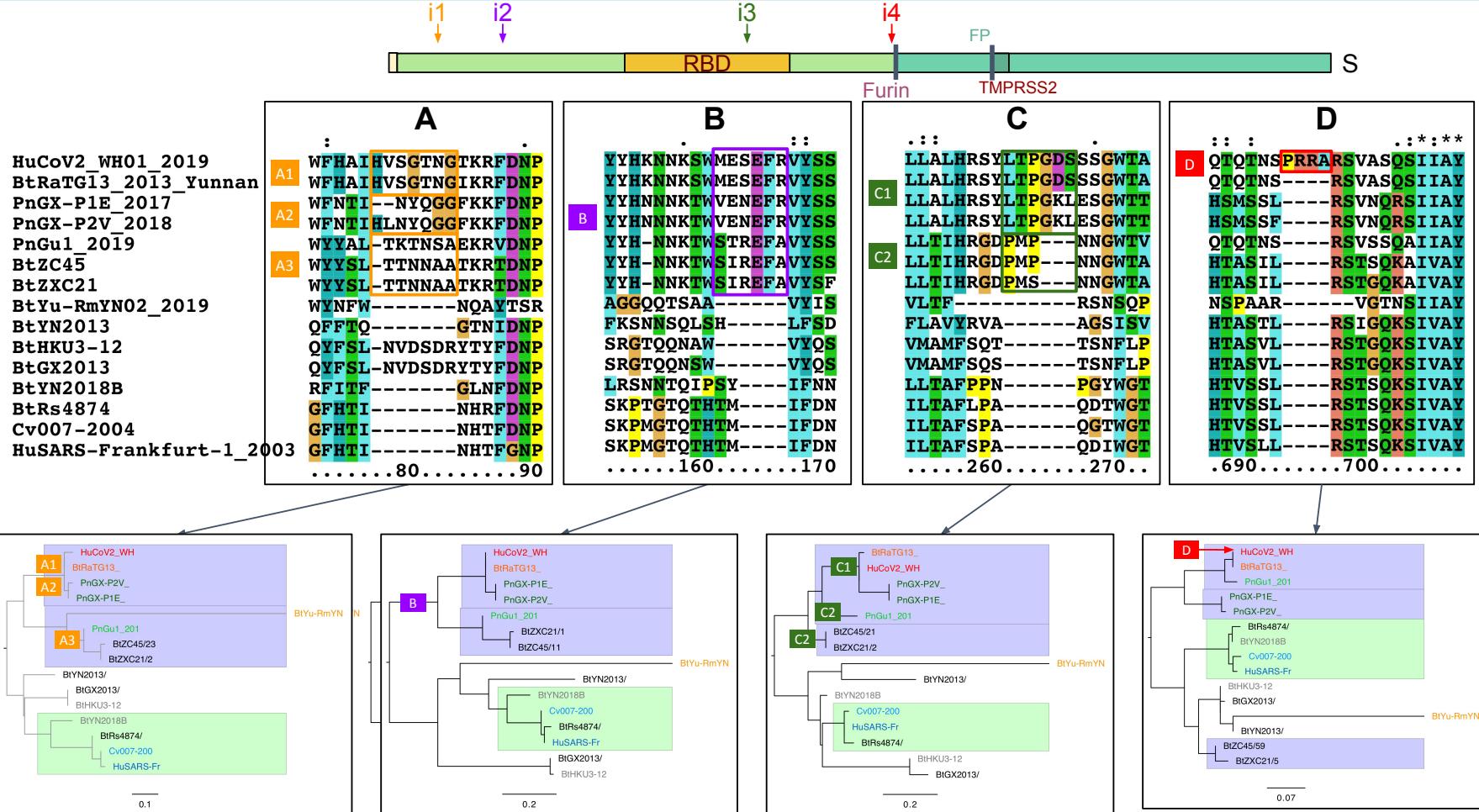


- **D** : insertion très courte, sans similarité significative avec HIV
- Insertion unique à SARS-CoV-2
- Crée un site furine, qui augmente la capacité de la protéine à se lier aux cellules humaines.
- Les données disponibles ne permettent pas de statuer entre origine naturelle ou artificielle de cette insertion ([rapport de l'académie nationale de médecine, 2025](#))



- Sallard, E., Halloy, J., Casane, D., van Helden, J., Decroly, E. Retrouver les origines du SARS-CoV-2 dans les phylogénies de coronavirus. Médecine/Sciences 36, Sept-Août 2020. <https://doi.org/10.1051/medsci/2020123>
- [De l'Origine du SARS-CoV-2 aux risques de zoonoses et de manipulations dangereuses de virus](#), rapport de l'Académie Nationale de Médecine (2025)

Alignements multiples et arbres phylogénétiques des insertions de la protéine S



Conclusion concernant l'hypothèse de Luc Montagnier

Selon Luc Montagnier, le virus covid19 est une manipulation humaine (17 avril 2020)

<https://www.youtube.com/watch?v=qSWCLHIOiMo>
(devenue inaccessible depuis lors)

"Je suis arrivé à la conclusion qu'il y avait eu une manipulation de ce virus. [...] Il y a un modèle qui est évidemment le virus classique, et là c'était un modèle venant de la chauve-souris, et là, à ce modèle on a par-dessus ajouté les séquences du VIH, du SIDA. ... Non, ce n'est pas naturel, c'était un travail de professionnel, de biologiste moléculaire, très minutieux, on peut dire d'horloger, au niveau des séquences. Dans quel but ce n'est pas clair. Mon travail c'est d'exposer les faits, c'est tout. Je n'accuse personne, je ne sais pas qui a fait ça et pourquoi. La possibilité c'est qu'on a voulu faire un vaccin contre le SIDA. Donc on a pris des petites séquences du virus [HIV] et on les a installées dans la séquence plus grande du coronavirus. [...] Il y a quand même une volonté d'étouffement, nous ne sommes pas les premiers. Un groupe de chercheurs indiens très renommés avaient publié la même chose, on les a forcés à rétracter. Si vous regardez leur publication vous voyez une grande bande "annulé"."

Nos analyses démontrent que **l'hypothèse d'une insertion de fragments de VIH dans un châssis de coronavirus ne tient pas la route.**

Elle reposait sur une méconnaissance des méthodes bioinformatiques et des indicateurs statistiques d'analyse de séquences.

Elle est totalement incompatible avec la présence de ces mêmes insertions dans plusieurs génomes, obtenus à partir d'échantillons collectés à des dates et endroits indépendants, et dont la séquence avait été publiée bien avant la pandémie.

Points-clés : origine de SARS-CoV-2

- L'alignement multiple permet d'identifier les blocs de séquences où un fragment est présent / absent.
- La comparaison avec l'arbre phylogénétique permet d'inférer les événements évolutifs (insertion, délétion, substitution) et d'estimer leur ancienneté
- Nos analyses démontrent que l'hypothèse d'une insertion de fragments de VIH dans un châssis de coronavirus ne tient pas la route.
 - Elle reposait sur une méconnaissance des méthodes bioinformatiques et des indicateurs statistiques d'analyse de séquences.
 - Elle est totalement incompatible avec la présence de ces mêmes insertions dans plusieurs génomes, obtenus à partir d'échantillons collectés à des dates et endroits indépendants, et dont la séquence avait été publiée bien avant la pandémie.

Pseudo-gènes (“gènes fossiles”)

Les pseudogènes (“Gènes fossiles”)

- Sean Carroll (2006) présente une série de cas de gènes « fossiles »: gènes qui ont perdu leur activité, mais dont on retrouve des traces dans les génomes.
- La “fossilisation” se manifeste généralement par la présence de nombreux codons stops dans les séquences codantes.
- Selon Carroll, cette fossilisation succède à une relaxation de la pression sélective:
 - Tous les gènes sont en permanence soumis au bombardement des mutations.
 - Les mutations qui perturbent la fonction d'un gène sont éliminées par la sélection naturelle (sélection « purificatrice »).
 - Si, pour une raison ou une autre, un gène devient dispensable pour un organisme donné dans un environnement donné, cette sélection est relâchée, et les mutations s'accumulent.
- Exemple: plusieurs espèces de levure ont perdu, de façon indépendante, la capacité de digérer le galactose. Dans chaque cas, chacun des 7 gènes GAL (devenus dispensables) est fossilisé.
- Note: **on dénomme actuellement ces régions génomiques “pseudogènes”**, d'une part parce qu'il ne s'agit pas à proprement parler de gènes (ils sont non fonctionnels) et d'autre part pour éviter la confusion avec les gènes (fonctionnels) identifiés dans les génomes d'organismes fossiles (par exemple les gènes d'*Homo neandertalis*).

La fossilisation des gènes de l'hémoglobine chez le poisson des glaces.

A



B



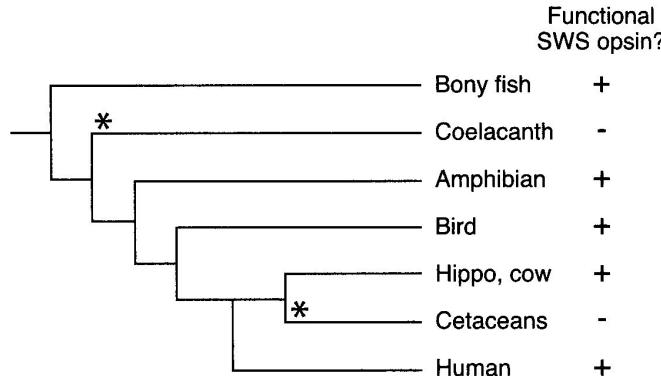
A Juvenile icefish. The transparent appearance is due to evolutionary loss of scales and red blood cells. (Photograph by Flip Micklin.)

B Adult mackerel icefish, *Champsocephalus gunnari*.

- Les poissons des glaces (famille des Channichthyidae) vivent dans l'Océan Arctique, dans des eaux dont la température varie de 4°C à -2°C (du fait de la salinité, elle est liquide).
- Leur sang contient des protéines « antigels », composées de motifs répétitifs.
- Le poisson des glaces n'a pas de globules rouges !
- Les échanges d'oxygène sont assurés à travers la peau, et l'O₂ dissous dans le sang est transféré aux organes.
- Leur corps ne contient pas non plus d'hémoglobine ni de myoglobine fonctionnelle.
- Cependant, on trouve dans leur génome des gènes fossilisés pour les deux chaînes de l'hémoglobine.

Carroll, S. B. (2006). The Making of the Fittest. DNA and the Ultimate Forensic Record of Evolution. Norton.

La perte de la perception des couleurs



* Gene fossilization

FIG. 5.2. **The same opsin gene has been fossilized twice.** The distribution of different mutations found in the coelacanth and cetacean SWS opsins and the evolutionary relationship of these species indicates that the SWS opsin was fossilized at least twice (asterisk). *Figure by Jamie Carroll.*

- Coelacanthes et les cétacés ont perdu la perception des couleurs
- Perte indépendante chez les coelacanthes et chez les cétacés.
- Les gènes des "c-opsines" (opsines responsables de la perception des couleurs) sont toujours présents, mais ils ne sont plus fonctionnels, du fait d'un grand nombre de mutations (y compris des codons stops).

La pseudogénisation des récepteurs olfactifs (OR) et des vomérorécepteurs (VR)

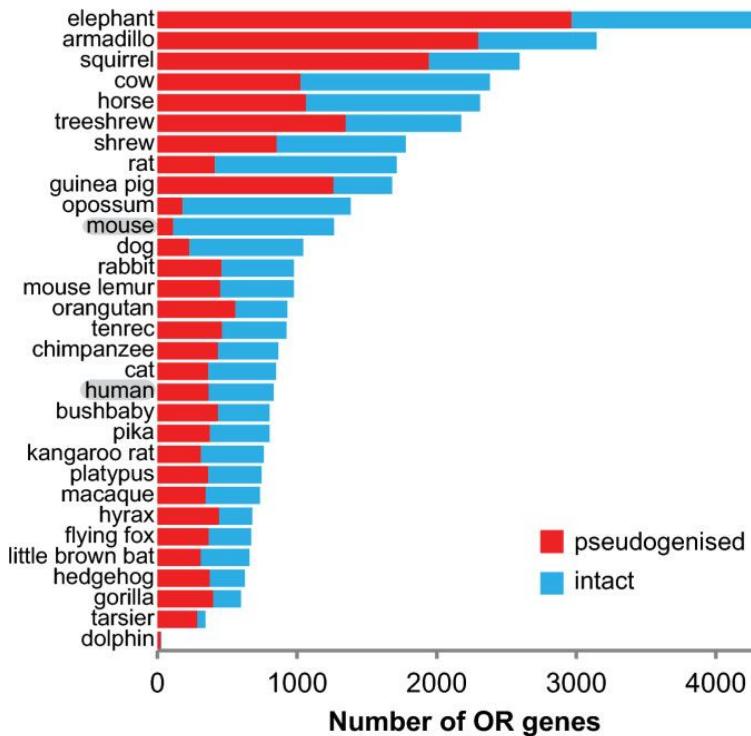
Gauche: récepteurs olfactifs (OR)

- Sur les 25.000 gènes de la souris, 1.400 codent pour des récepteurs olfactifs.
- Chez l'humain, la moitié de ces gènes sont fossilisés, et ne peuvent plus produire de récepteurs fonctionnels.
- L'analyse des génomes d'autres mammifères montre que la perte massive de récepteurs olfactifs se retrouve chez les primates de l'ancien monde, autrement dit ceux qui ont acquis la vision trichromatique.

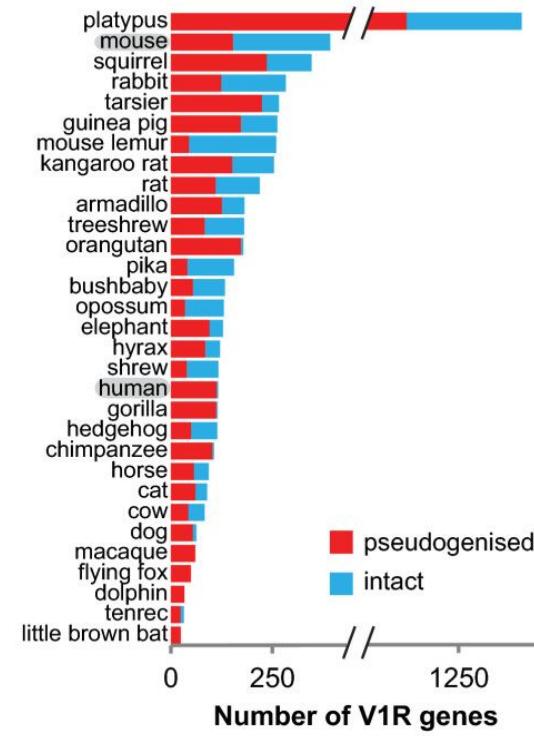
Droite: vomérorécepteurs (VR)

- Homologues des récepteurs olfactifs impliqués dans la perception des phéromones.
- Humain et gorille : (quasiment) pas de communication par phéromones, mais le génome comporte une centaine de vomérorécepteurs pseudogénisés.

(a)



(b)



L'évolution en marche: la fossilisation massive des gènes de *Mycobacterium leprae*

- *Mycobacterium tuberculosis*
 - Extracellulaire
 - 4.189 gènes codant pour des protéines
- *Mycobacterium leprae*
 - Intra-cellulaire
 - 1.605 gènes codant pour des protéines
 - A perdu (par délétions) ~1000 gènes présents chez *M.tuberculosis*
 - On trouve également plus de 1.000 gènes fossiles.

La fossilisation est une voie de non-retour

Dès que la pression sélective est relâchée, la probabilité de fossilisation d'un gène est très élevée. Par contre, la probabilité que de revenir à une copie fonctionnelle à partir d'un gène fossilisé est quasiment nulle.

Ces probabilités sont discutées par S. Carroll.

Quand les branches de l'arbre du vivant s'entrecroisent

Quand les branches de l'arbre du vivant s'entrecroisent

Fig. 1. Part of the only figure in the *Origin of Species*. Darwin first uses it to represent the divergence of variants within a species, showing successively more difference in a single lineage (a^1 through a^{10}) and splitting into multiple lineages (m , s , i , and so forth), some of which will become new species. Later, he expands the tree metaphor, explaining that "limbs divided into great branches ... were themselves once, when the tree was small, budding twigs; and this connection of the former and present buds by ramifying branches may well represent the classification of all extinct and living species in groups subordinate to groups" (3, p. 171).

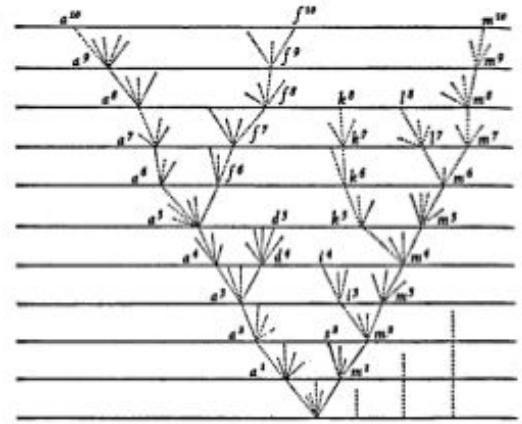
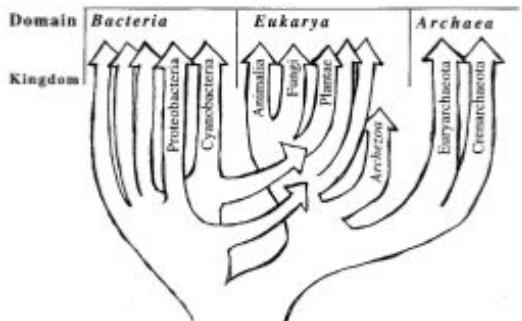


Fig. 2. The current consensus or standard model. Only a few of the "kingdoms" of the "domain" Bacteria are shown. Branching orders of several kingdoms within Bacteria and Eukarya remain in dispute. Mitochondrial and chloroplast endosymbioses are indicated by lower and upper diagonal arrows, respectively. Archaea, as a subkingdom composed of primitively amitochondrial protists, may be extinct. For SSU rRNA trees with much more detail, see (5).

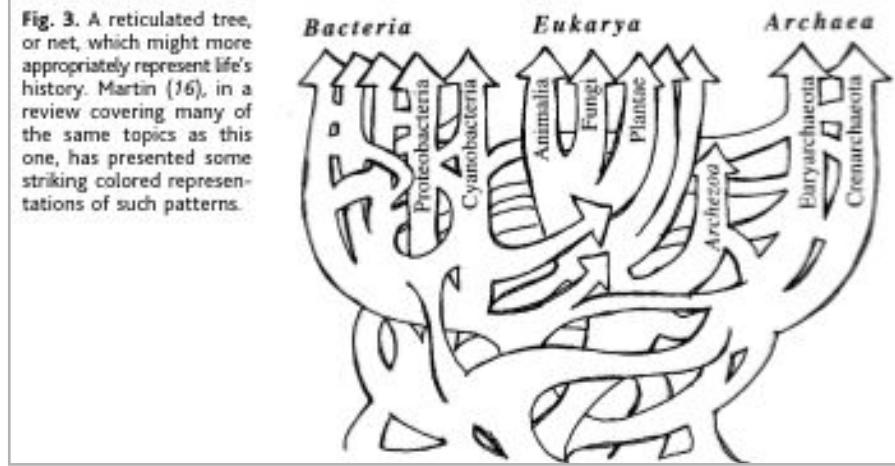


L'arbre de la vie de Darwin (Fig 1) est revisité par Doolittle (1999) pour tenir compte

Fig 2: des événements d'**endosymbiose** liés à l'apparition des organelles des eucaryotes (mitochondrie et chloroplaste).

Fig 3: des **transferts horizontaux** entre génomes de procaryotes.

Fig. 3. A reticulated tree, or net, which might more appropriately represent life's history. Martin (16), in a review covering many of the same topics as this one, has presented some striking colored representations of such patterns.



The ring of life provides evidence for a genome fusion origin of eukaryotes

Maria C. Rivera^{1,3,4} & James A. Lake^{1,3,4}

¹Molecular Biology Institute, MCD Biology, ²Human Genetics, ³IGPP, and ⁴Astrobiology Institute, University of California, Los Angeles 90095, USA

Genomes hold within them the record of the evolution of life on Earth. But genome fusions and horizontal gene transfer seem to have obscured sufficiently the gene sequence record such that it is difficult to reconstruct the phylogenetic tree of life. Here we determine the general outline of the tree using complete genome data from representative prokaryotes and eukaryotes and a new genome analysis method that makes it possible to reconstruct ancient genome fusions and phylogenetic trees. Our analyses indicate that the eukaryotic genome resulted from a fusion of two diverse prokaryotic genomes, and therefore at the deepest levels linking prokaryotes and eukaryotes, the tree of life is actually a ring of life. One fusion partner branches from deep within an ancient photosynthetic clade, and the other is related to the archaeal prokaryotes. The eubacterial organism is either a proteobacterium, or a member of a larger photosynthetic clade that includes the Cyanobacteria and the Proteobacteria.

- Rivera & Lake (2004) analysent les relations entre tous les gènes d'eucaryotes, d'eubactéries, et d'archées.
- Leur analyse suggère que les génomes eucaryotes résulteraient d'une fusion entre un génome de bactérie et un génome d'archée.
- Les gènes provenant des archées sont majoritairement impliqués dans des fonctions de maintien de la cellule (réplication, transcription et sa régulation).
- Les gènes provenant des bactéries sont majoritairement impliqués dans le métabolisme.

