

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

Introduction à la bioinformatique (UE SSV3U15)

Version détaillée

Jacques van Helden

Aix-Marseille Université

orcid.org/0000-0002-8799-8584

1. Représentations arborescentes de l'évolution
2. Homologie et analogie
3. Les duplications à l'origine de l'innovation
4. Phylogénie moléculaire : construction d'un arbre à partir d'une famille de des séquences
5. Relations entre séquences : homologie, analogie, paralogie, orthologie, xénologie
6. Phylogénomique : retracer l'évolution des espèces à partir des séquences génomiques
7. Pseudo-gènes ("gènes fossiles") : les reliques de gènes ancestraux dans les génomes d'espèces actuelles
8. Quand les branches de l'arbre du vivant s'entrecroisent

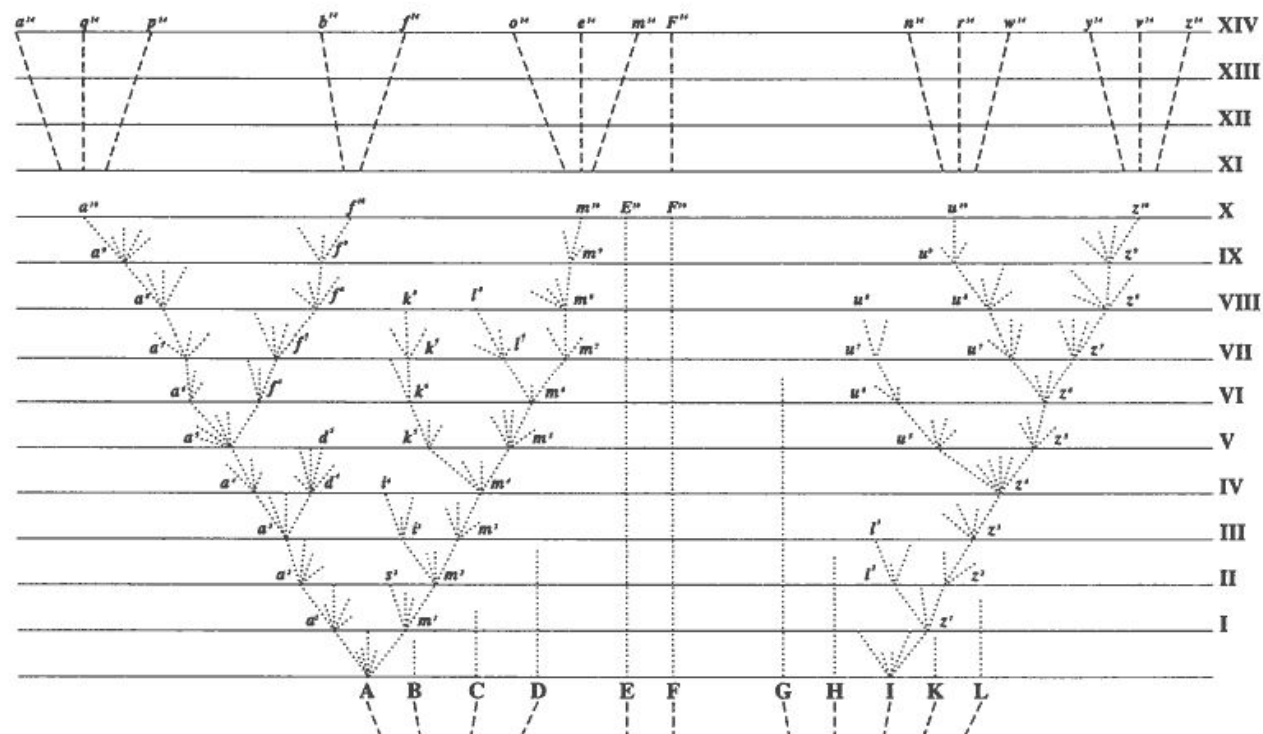
Matériel supplémentaire (ne fait pas partie de la matière d'examens)

9. Retracer l'évolution de SARS-CoV-2 dans les génomes des coronavirus

Représentations arborescentes de l'évolution

Origine des espèces – L'arbre de la vie de Darwin

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



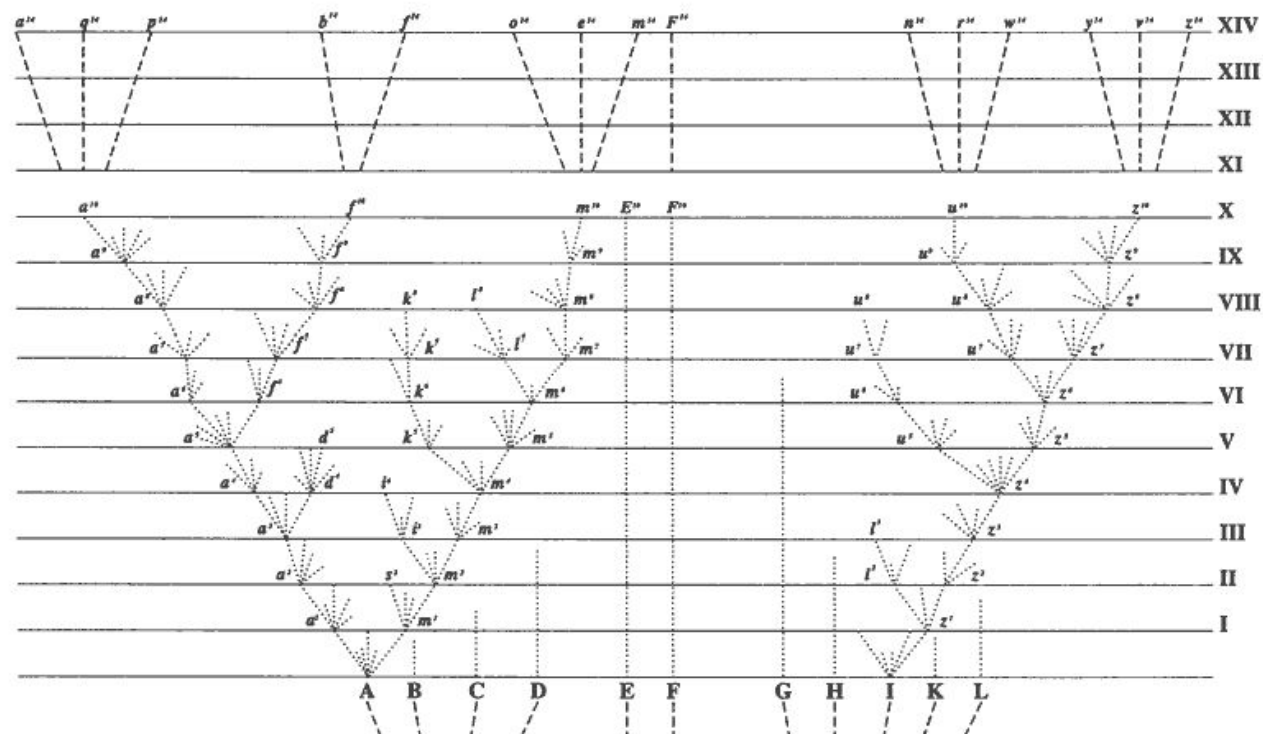
Origine des espèces – L'arbre de la vie de Darwin

Il s'agit d'un **arbre synchrone** :

chaque niveau horizontal représente un moment donné.

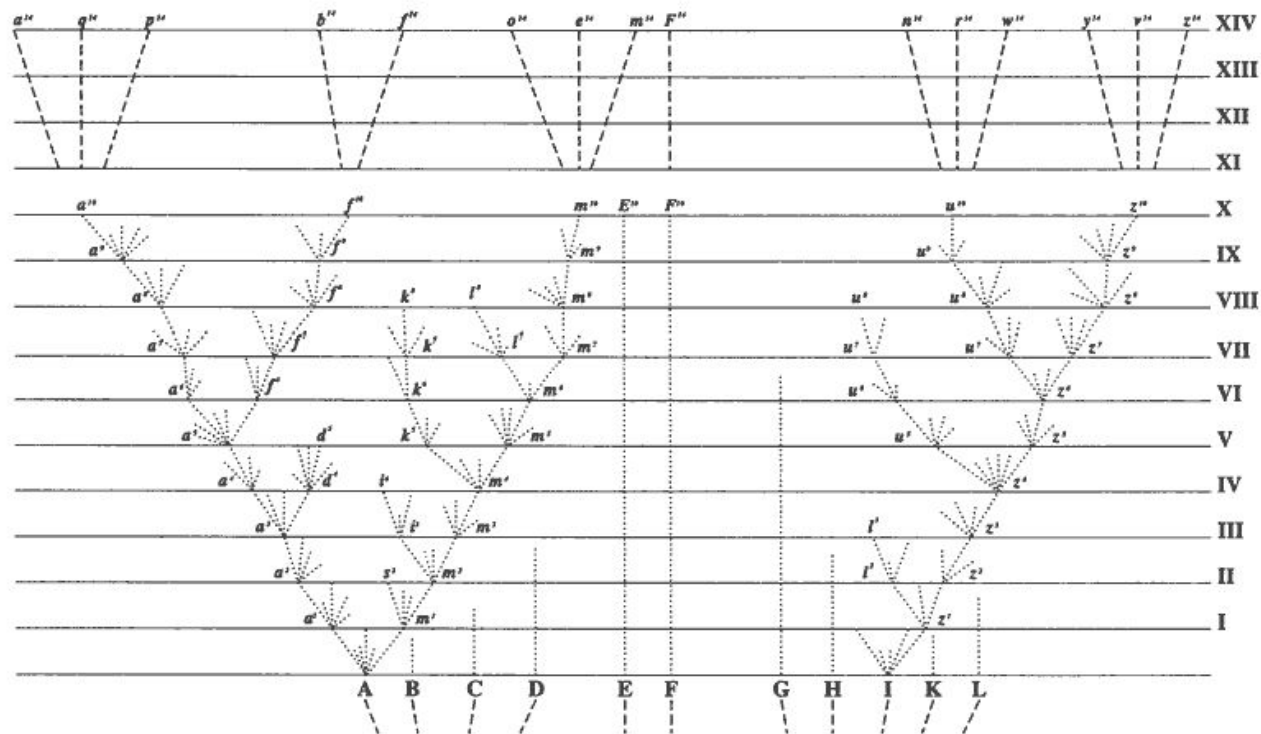
- La racine correspond aux époques les plus anciennes.
- Le niveau le plus élevé correspond au présent.
- A chaque époque on trouve des organismes de différents niveaux de complexité.

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



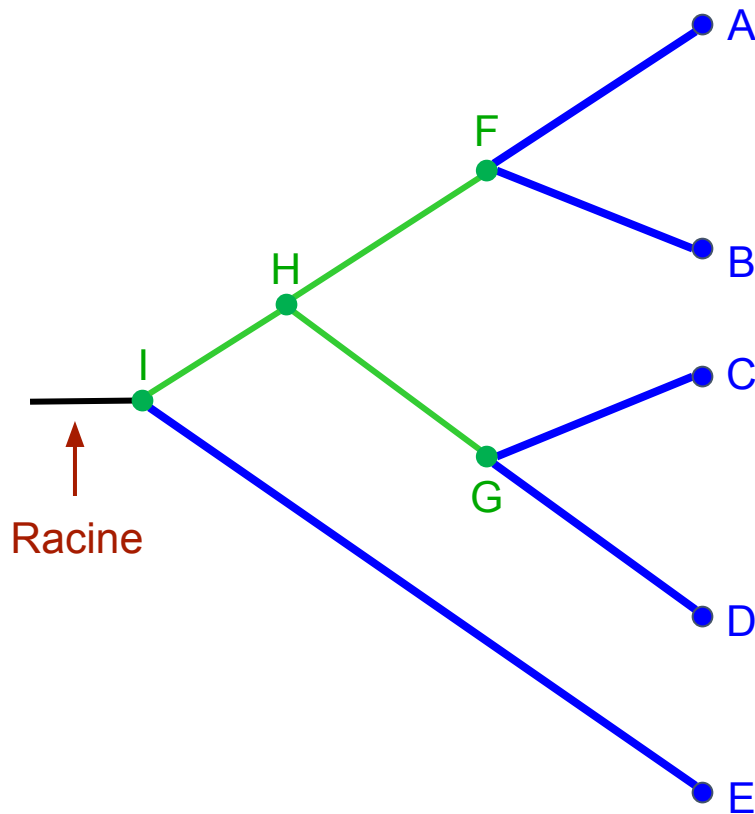
Origine des espèces – L'arbre de la vie de Darwin

- 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 = lignées évolutives (successions ancêtre/descendants) entre unités taxinomiques
 - Branches internes
 - Branches externes
- On appelle **topologie** l'ensemble des branchements de l'arbre.



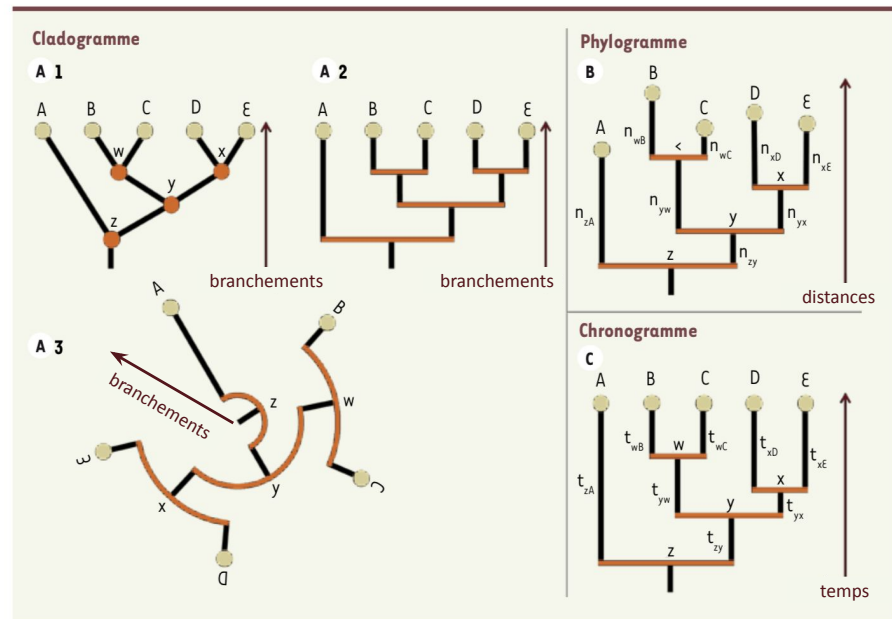
Représentations arborescentes des histoires évolutives

Différentes modalités de représentation des arbres

- Le **cladogramme** (Figure A) représente uniquement la **topologie de l'arbre**, c'est-à-dire l'**ordre des branchements** entre les unités taxonomiques. La longueur des branches n'a pas de signification particulière : elle ne représente ni le temps écoulé ni la quantité de divergence évolutive.
- Dans un **phylogramme** (Figure B), la longueur des branches est proportionnelle à la **quantité de changement évolutif estimée** le long de chaque lignée, par exemple le nombre de substitutions d'une séquence ou, dans le cas de caractères morphologiques, le nombre estimé de changements de caractères. 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** (Figure C), la longueur des branches est proportionnelle au **temps écoulé**. Lorsque toutes les OTU sont contemporaines, leurs feuilles sont alignées. En revanche, si certaines OTU correspondent à des fossiles, les feuilles peuvent se trouver à des hauteurs différentes, reflétant leur âge.

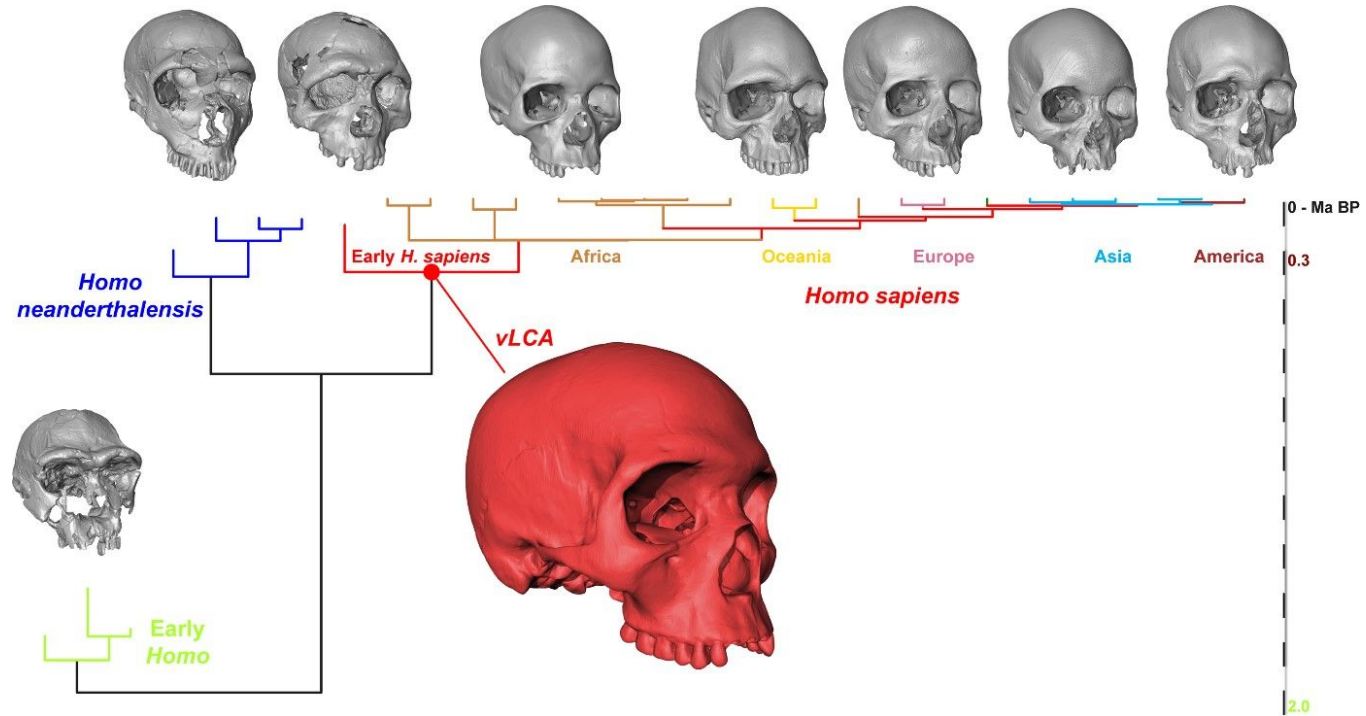
Les branchements peuvent également être organisés de différentes façons.

- Bifurcations triangulaires (A1) ou rectangulaires (A2)
- Disposition verticale (A1, A2), horizontale (non montré ici) ou radiale (A3)



Arbre phylogénétique des espèces fossiles et actuelles d'Hominiens

- L'axe vertical représente le temps. Il s'agit donc d'un chronogramme.
- Les crânes gris correspondent à des unités taxonomiques opérationnelles (OTU), pour lesquelles on dispose de spécimens.
- Le crâne rouge représente une reconstruction du crâne de l'ancêtre commun de tous les *Homo sapiens*. La forme de ce crâne a été inférée à partir de 29 spécimens de crânes d'*Homo sapiens* actuels et fossiles.

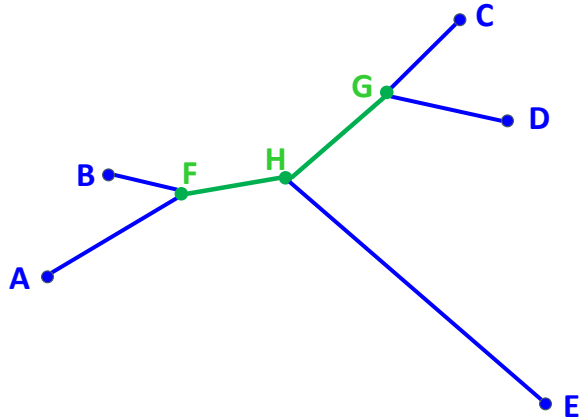


Note : l'existence d'un ancêtre commun ne fait aucun doute d'un point de vue scientifique. Ce qui est hypothétique est l'ordre des branchements et les détails morphologiques du crâne reconstruit

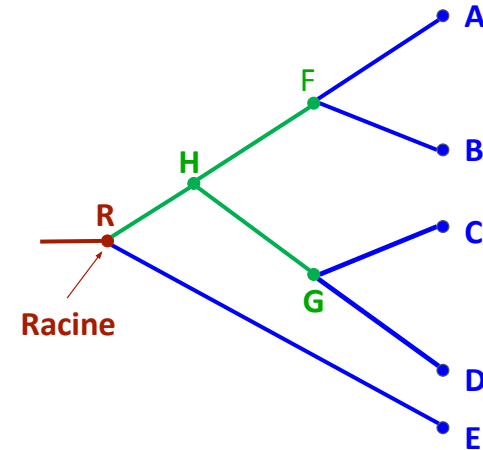
Arbres enracinés ou non

- La plupart des méthodes d'inférence phylogénétique produisent dans un premier temps des arbres **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.

Arbre non-enraciné

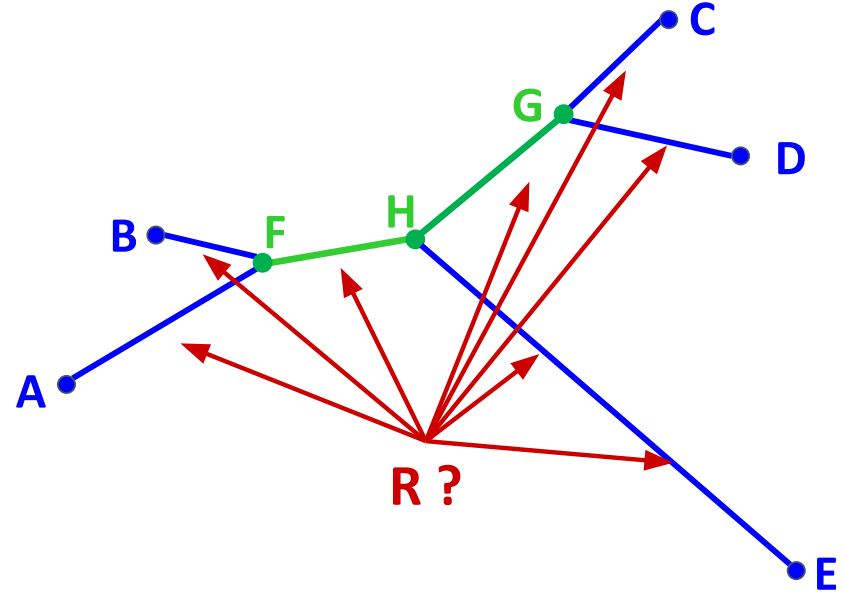


Arbre enraciné



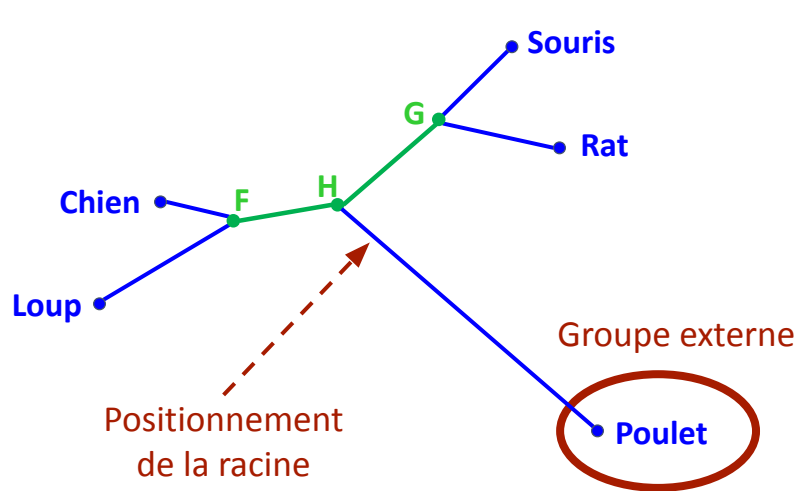
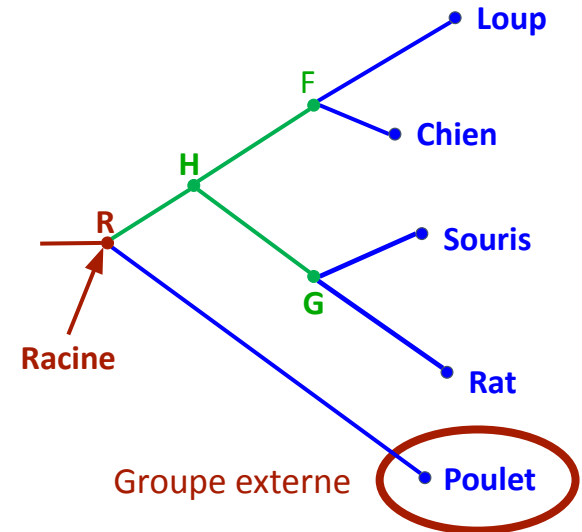
Comment enraciner un arbre phylogénétique ?

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

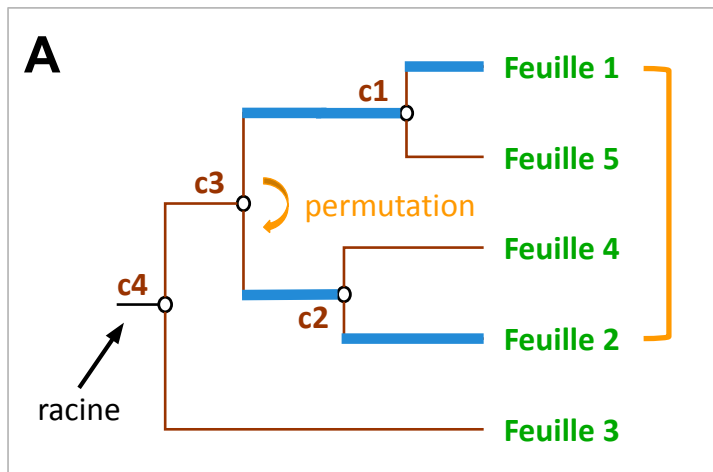
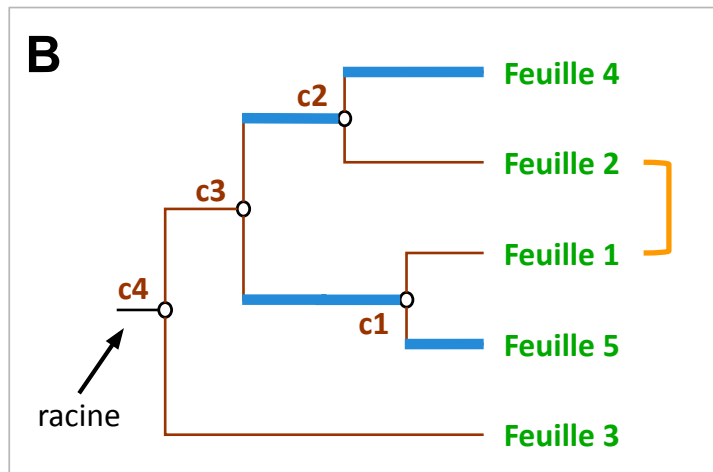


Approches pour enracer un arbre phylogénétique

- Dans certains cas, on peut s'appuyer sur la connaissance préalable d'un **groupe externe** (*outgroup* en anglais), dont on sait qu'il est plus éloigné de toutes les autres OTU qu'elles ne le sont entre elles.
 - Exemple : si un arbre contient chien, loup, souris, rat et poulet → sur base des connaissances biologiques, on décide que le **groupe externe** est le poulet
- En absence de connaissance *a priori* du OTU les plus externes parmi les OTU étudiées, on peut envisager un **enracinement au point milieu** : on identifie la paire de feuilles les plus distantes, et on place la racine à la moitié du chemin qui les relie.
 - **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



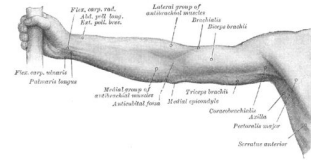
- Dans un arbre, les deux branches de chaque bifurcation peuvent être permutées.
- Le résultat est un arbre **isomorphe**, équivalent à l'arbre initial du point de vue de la topologie (succession des branchements).
- **Les arbres A et B sont équivalents**
- Cependant
 - Arbre du dessus: les feuilles 1 et 2 semblent très éloignées.
 - Arbre du dessous: les feuilles 1 et 2 semblent voisines.
- Les **distances verticales [crochets]** entre deux nœuds ne reflètent pas leur distance évolutive !
- Dans un phylogramme rectangulaire, la distance évolutive entre deux nœuds est obtenue en calculant la **somme des longueurs des branches horizontales** qui les séparent (**traits bleus épaissis**).

Homologie et analogie

- **Homologie** : ressemblance entre des caractères (phénotypiques ou génétiques) qui s'explique par une **origine ancestrale commune**.
- Exemple : les membres antérieurs des vertébrés terrestres sont homologues : ils sont hérités d'un ancêtre commun.
- Au sein des tétrapodes actuels, le membre antérieur constitue un **caractère ancestral**, présent chez leur dernier ancêtre commun.
- **Évolution divergente** : les différences entre ces organes résultent de l'accumulation de modifications différentes dans les lignées issues d'un ancêtre commun.

Homologie

Humain (mammifère)



Chimpanzé (mammifère)



Pigeon (oiseau)



- **Analogie** : ressemblance entre caractères (phénotypiques ou de génétiques) résultant d'**histoires évolutives indépendantes**.
- Exemple : les ailes des oiseaux et des mouches.
 - Fonction identique : le vol
 - Certaines similarités dans la structure (membranes plates, articulation musculaire permettant des mouvements verticaux).
- Suggère une **évolution convergente** : les similarités entre les caractères résultent d'une accumulation progressive de modifications, sous l'effet d'une même pression évolutive.
- Les caractères permettant le vol sont donc des **caractères dérivés**, absents de l'ancêtre commun mais apparus indépendamment dans chacune des deux lignées après leur séparation.

Analogie

Pigeon (oiseau)



Mouche (insecte)



Homologie et analogie : un exemple non trivial

- Les **membres antérieurs** des chauves-souris sont homologues à ceux des oiseaux.
 - Ils ont été hérités depuis des premiers vertébrés terrestres.
 - Ce caractère était présent chez l'ancêtre des tétrapodes actuels → **caractère ancestral.**
- Cependant, **en tant qu'organes de vol**, les ailes des chauves-souris sont **analogues** aux ailes des oiseaux
 - Les caractères qui permettent à ces deux groupes de tétrapodes de voler sont apparus indépendamment dans les deux lignées.
 - Il s'agit de **caractères dérivés**, qui n'étaient pas présents chez le dernier ancêtre commun des oiseaux et des mammifères.

Pigeon (oiseau)



Chauve-souris (mammifère)

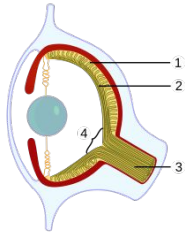
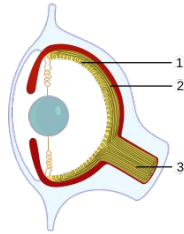


Structures analogues et convergence évolutive

- 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, ...
 - **Différences:** orientation des cellules rétinienne: les axones partent vers l'intérieur chez les vertébrés, vers l'extérieur chez les céphalopodes (figure du bas)
- 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**.

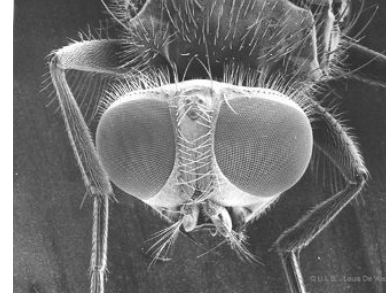
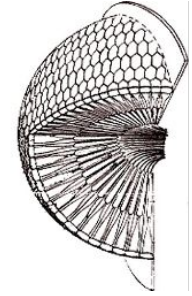
Vertébré

Céphalopode

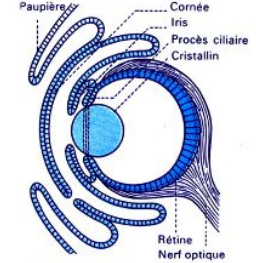


1. Rétine
2. Axones
3. Nerf optique
4. Tache aveugle chez les vertébrés

Oeil de drosophile (insecte)

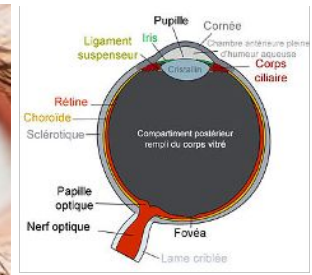


Oeil de pieuvre (mollusque céphalopode)



Oeil de Céphalopode (coupe verticale)
D'après Atlas de Biologie

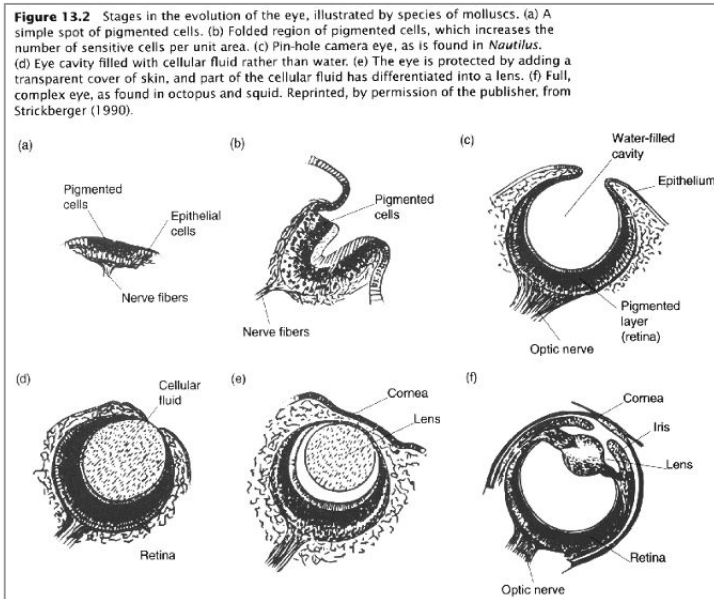
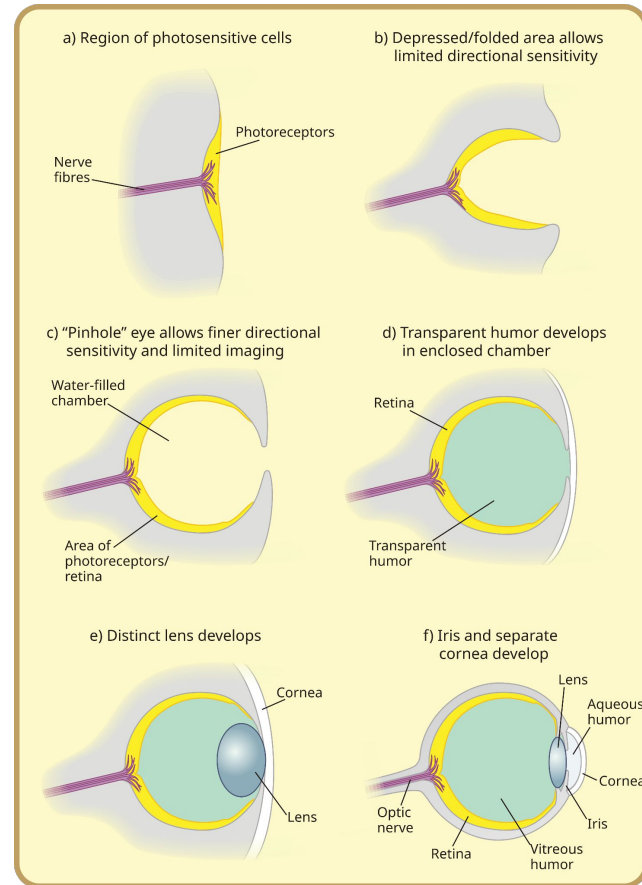
Oeil humain (mammifère)



Evolution graduelle de l'oeil des mollusques

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 pieuvres.

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



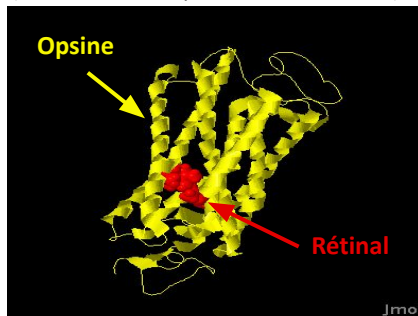
Figures :

- Carroll, S. B. The making of the fittest: DNA and the ultimate forensic record of evolution. (Quercus, 2009).
- https://en.wikipedia.org/wiki/Evolution_of_the_eye#/media/File:Diagram_of_eye_evolution.svg

Les duplications à l'origine de l'innovation

Modèle tridimensionnel du pigment des cônes bleus

(Structure PDB 1kpn affichée avec Jmol)



(Structure PDB 1kpn affichée avec MacPyMol)



Perception de la lumière

- **Opsine** : protéine transmembranaire
- **Rétinal** : petite molécule, absorbe la lumière
- 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 des primates de l'ancien monde

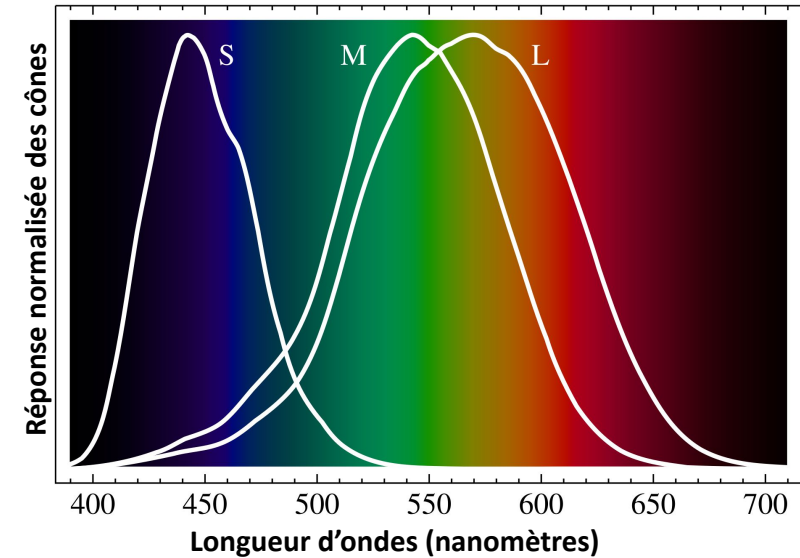
Spectre de sensibilité des cônes

- Cônes sensibles aux **ondes courtes** (short-wave sensitive : **SWS**) : 420 nm (autour du **violet-bleu**)
- Cônes sensibles aux **ondes moyennes** (medium-wave sensitive : **MWS**) : 534 nm (autour du **vert**)
- Cônes sensibles aux **ondes longues** (long-wave sensitive : **LWS**) : 564 nm (autour du **rouge**)

Distances entre pics

- Opsine **MWS** versus **LWS** :
Distance entre pics = $564 - 533 = 31$ nm
→ **pics très proches**
→ **fort recouvrement** entre spectres de sensibilité
- Opsine **SWS** versus **MWS** :
Distance entre pics = $533 - 437 = 96$ nm

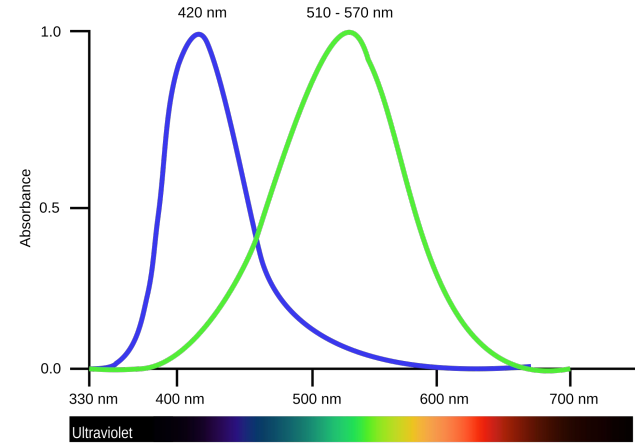
Spectre de sensibilité des opsines humaines



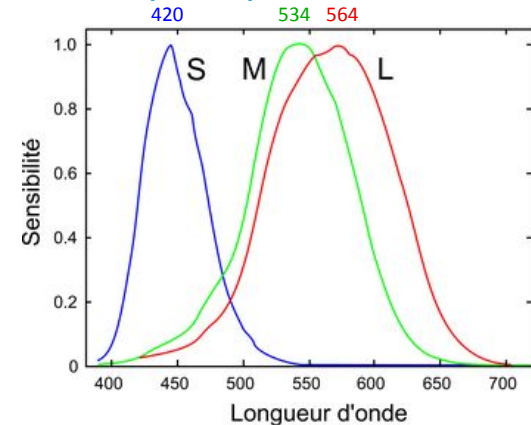
La vision dichromatique des autres mammifères

- La plupart des mammifères ont une **vision dichromatique**.
 - Ils disposent d'une opsine sensible au **bleu**, et d'une autre sensible aux ondes **vert-rouge**, avec un pic qui varie selon les espèces
 - Ils peuvent distinguer le bleu du vert ou du rouge, mais ne font pas la différence entre vert et rouge
- Note : certaines formes de daltonisme humain reviennent à une vision dichromatique

Vision dichromatique des autres mammifères



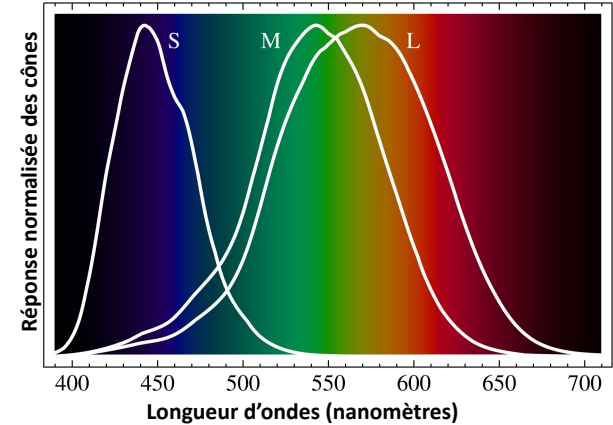
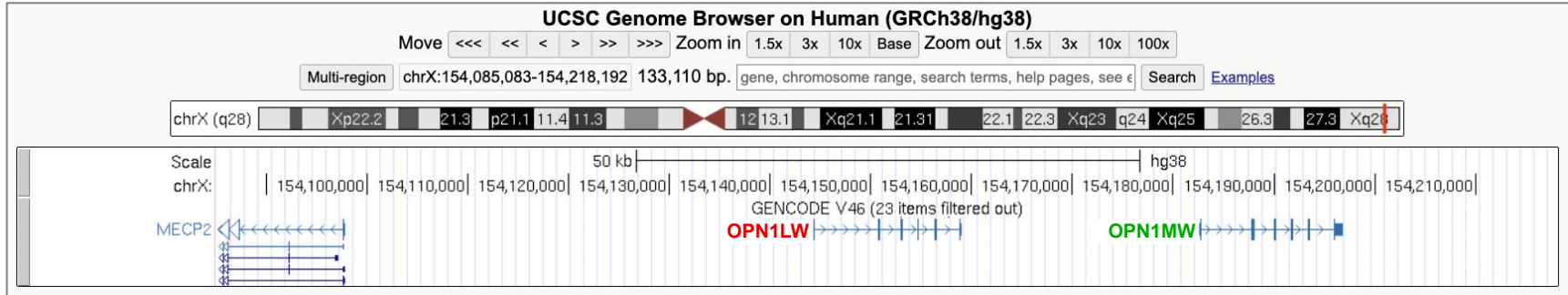
Vision trichromatique des primates de l'ancien monde



Origine de la vision trichromatique des primates de l'ancien monde

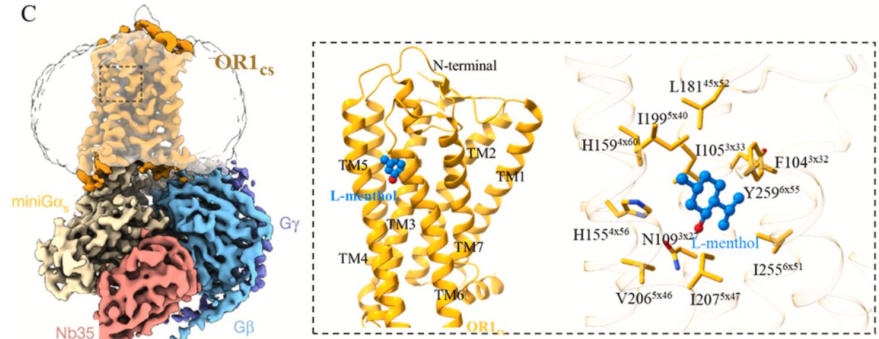
- Chez les primates de l'ancien monde, la présence d'opsines distinctes avec une sensibilité "plutôt verte" (MWS) et "plutôt rouge" (SWS) résulte d'une **duplication du gène codant pour l'opsine rouge-verte**.
- Sur le chromosome X de ces primates on trouve 2 gènes en tandem qui produisent 2 opsines différentes
 - OPN1LW** (Long Waves) → opsine **rouge**
 - OPN1MW** (Medium Waves) → opsine **verte**
- Par ailleurs, sur le chromosome 7 humain, on trouve le gène **OPN1SW** (Short Waves) → opsine **bleue**

Spectre de sensibilité des opsines humaines



Une grande famille ... de gènes

- La plus grande famille de gènes chez les métazoaires est celle des **récepteurs olfactifs**.
- Chaque gène code pour une protéine transmembranaire exprimée dans les neurones récepteurs de l'épithélium olfactif, dont le domaine N-terminal, extracellulaire, permet de détecter une gamme très spécifique de certains composés chimiques.
- L'expansion exceptionnelle de cette famille de gènes résulte de duplications répétées, qui induisent des **différences fonctionnelles** en modifiant la spécificité olfactive de chaque récepteur
→ processus de **duplication-divergence**
- On trouve également dans les génomes un grand nombre de loci récepteurs inactifs, suite à des mutations non-sens, substitutions, délétions partielles ou modifications des régions régulatrices. Il s'agit de **pseudogènes**.



Structure tridimensionnelle d'un récepteur olfactif lié à un odorant (le L-menthol). Gu et al. (2025). DOI : [10.1016/j.tifs.2025.104971](https://doi.org/10.1016/j.tifs.2025.104971)

Nombre de récepteurs gènes fonctionnels et de pseudogènes de récepteurs olfactifs chez l'humain et la souris. Barnes IHA et al. 2020. DOI: [10.1186/s12864-020-6583-3](https://doi.org/10.1186/s12864-020-6583-3)

Organisme	Humain	Souris
Nombre total de gènes codants (NCBI, 2026)	23 306	22 198
Loci correspondant à des récepteurs olfactifs	874	1 483
Gènes produisant potentiellement un récepteur olfactif fonctionnel	389	1 141
Pseudogènes de récepteurs olfactifs	485	342

- Gu Y, Zhou X, Sun B, Zhang L, Yang Y, Qi P, Li P, Sun S, Zhang Y. 2025. A systematic review of the structure and function of human olfactory receptors and key technologies involved. Trends in Food Science & Technology, 159: 104971. DOI : [10.1016/j.tifs.2025.104971](https://doi.org/10.1016/j.tifs.2025.104971)
- Barnes IHA et al. 2020. Expert curation of the human and mouse olfactory receptor gene repertoires identifies conserved coding regions split across two exons. BMC Genomics, 21: 196. DOI: [10.1186/s12864-020-6583-3](https://doi.org/10.1186/s12864-020-6583-3)

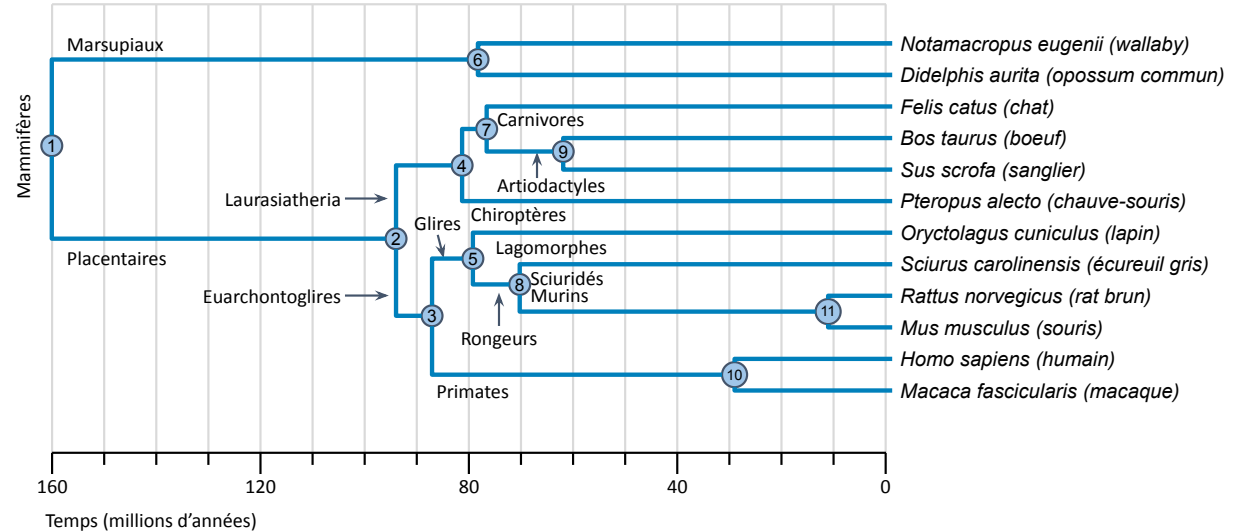
Phylogénie moléculaire : construction d'un arbre à partir d'une famille de des séquences

La **phylogénie moléculaire** consiste à inférer l'histoire évolutive des organismes en s'appuyant sur l'analyse des séquences de macromolécules – acides nucléiques ou protéines.

Etapes

1. Collecte des séquences
2. Alignement multiple
3. Inférence phylogénétique

- **Choix des espèces**
- **Recherche de séquences par similarité** dans les bases de données de séquences (Ensembl, Uniprot, NCBI)
- Vérification de la **couverture des groupes taxonomiques d'intérêt**
- **Filtrage** : élimination de séquences redondantes et de biais de représentation



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

Procédure

- Données : séquences peptidiques
 - 12 opsines SWS
 - 14 opsines MW ou LW
- Alignement par paires
- Calcul du taux d'identités
- Ordonnancement des colonnes et lignes pour rapprocher les opsines les plus similaires

Observations

- 2 groupes bien distincts (carrés orange - rouge)
- Les opsines SW sont toutes regroupées, plus proches entre elles que des MW+LW
- Les opsines LW et MW sont mélangées, elles ne forment pas 2 blocs séparés

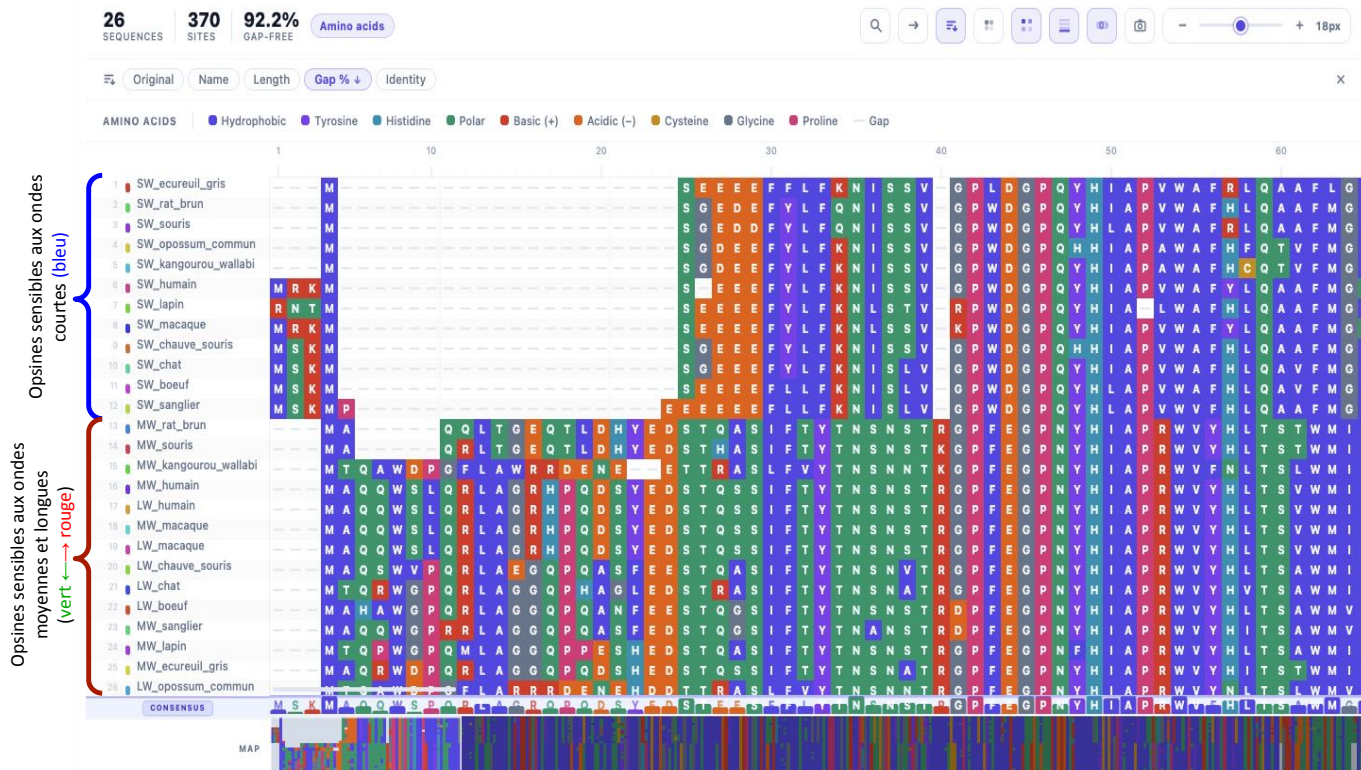
Opsines sensibles aux ondes courtes (violet – bleu)

Opsines sensibles aux ondes moyennes et longues (vert ← → rouge)

	SW humain	SW macaque	SW souris	SW rat brun	SW ecreuil gris	SW lapin	SW chauve souris	SW sanglier	SW boeuf	SW chat	SW opossum commun	SW kangourou wallaby	MW humain	LW humain	MW macaque	LW macaque	MW souris	MW rat brun	MW ecreuil gris	MW lapin	LW chauve souris	MW sanglier	LW boeuf	LW chat	LW opossum commun	MW kangourou wallaby	
SW humain	100	95	85	86	86	84	88	86	85	85	82	83	42	41	41	41	42	42	41	41	42	40	41	41	41	41	
SW macaque	95	100	85	87	87	85	88	85	86	86	83	84	42	41	41	41	41	41	41	41	42	40	41	41	41	40	41
SW souris	85	85	100	96	90	86	90	85	87	86	85	86	41	40	40	40	39	39	40	39	41	39	40	39	40	41	
SW rat brun	86	87	96	100	89	86	90	86	87	86	85	86	42	41	41	41	41	41	41	40	42	40	41	40	41	41	
SW ecreuil gris	86	87	90	89	100	88	87	86	87	87	84	86	42	42	42	42	41	41	41	41	42	41	41	41	41	42	
SW lapin	84	85	86	86	88	100	86	84	85	84	81	83	42	41	42	41	41	41	41	41	42	40	41	41	41	41	
SW chauve souris	88	88	90	90	87	86	100	87	91	91	86	87	39	38	39	38	38	38	38	39	39	38	38	38	38	39	
SW sanglier	86	85	85	86	86	84	87	100	92	86	81	82	42	41	41	41	40	40	40	40	42	40	40	40	40	40	
SW boeuf	85	86	87	87	87	85	91	92	100	91	83	86	43	41	42	41	42	42	41	41	42	40	41	41	41	42	
SW chat	85	86	86	86	87	84	91	86	91	100	83	85	42	41	41	41	40	40	40	40	41	39	40	40	41	42	
SW opossum commun	82	83	85	85	84	81	86	81	83	83	100	91	41	40	41	40	41	41	40	40	42	40	40	39	41	41	
SW kangourou wallaby	83	84	86	86	86	83	87	82	86	85	91	100	42	41	42	41	42	42	41	40	42	40	41	40	42	42	
MW humain	42	42	41	42	42	42	39	42	43	42	41	42	100	96	98	96	88	89	90	88	90	89	89	90	83	83	
LW humain	41	41	40	41	42	41	38	41	41	41	40	41	96	100	96	99	88	89	91	88	90	90	89	91	83	82	
MW macaque	41	41	40	41	42	42	39	41	42	41	41	42	98	96	100	96	89	90	90	87	90	90	89	90	84	84	
LW macaque	41	41	40	41	42	41	38	41	41	41	40	41	96	99	96	100	88	89	91	88	90	90	90	92	83	82	
MW souris	42	41	39	41	41	41	38	40	42	40	41	42	88	88	89	88	100	98	89	90	90	87	88	88	82	82	
MW rat brun	42	41	39	41	41	41	38	40	42	40	41	42	89	89	90	89	98	100	91	91	91	89	89	89	84	82	
MW ecreuil gris	41	41	40	41	41	41	38	40	41	40	40	41	90	91	90	91	89	91	100	92	92	90	90	90	82	80	
MW lapin	41	41	39	40	41	41	39	40	41	40	40	40	88	88	87	88	90	91	92	100	91	90	90	91	80	80	
LW chauve souris	42	42	41	42	42	42	39	42	42	41	42	42	90	90	90	90	90	91	92	91	100	92	93	92	83	82	
MW sanglier	40	40	39	40	41	40	38	40	40	39	40	40	89	90	90	90	87	89	90	90	92	100	97	93	80	79	
LW boeuf	41	41	40	41	41	41	38	40	41	40	40	41	89	89	89	90	88	89	90	90	93	97	100	93	81	80	
LW chat	41	41	39	40	41	41	38	40	41	40	39	40	90	91	90	92	88	89	90	91	92	93	93	100	83	83	
LW opossum commun	41	40	40	41	41	41	38	40	41	41	41	42	83	83	84	83	82	84	82	80	83	80	81	83	100	93	
MW kangourou wallaby	41	41	41	41	42	41	39	40	42	42	41	42	83	82	84	82	82	82	80	80	82	79	80	83	93	100	

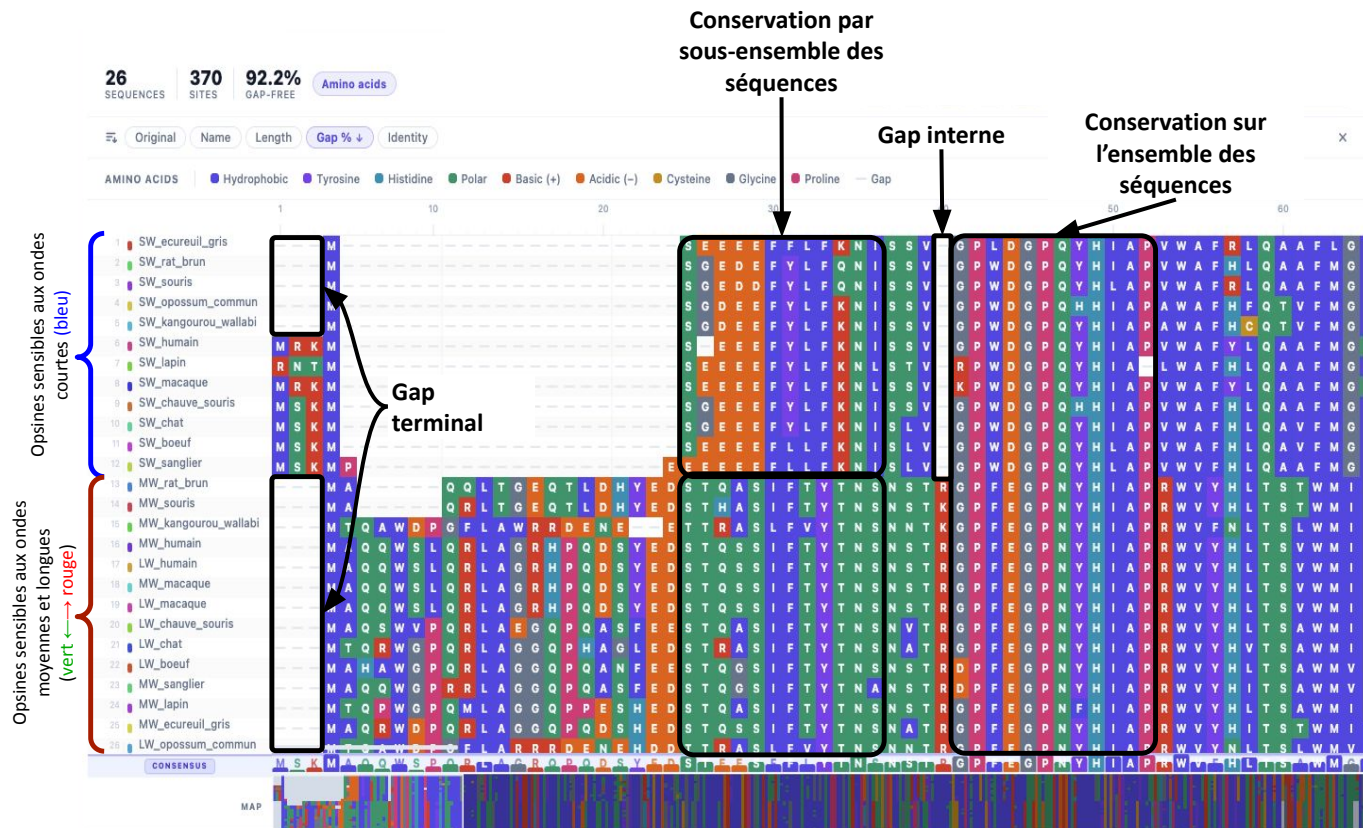
Alignement multiple des opsines de mammifère

- 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
- Les **couleurs** reflètent certaines **propriétés biochimiques** des acides aminés (voir légende) : hydrophobes polaire, basique, acide.



Alignement multiple des opsines de mammifère

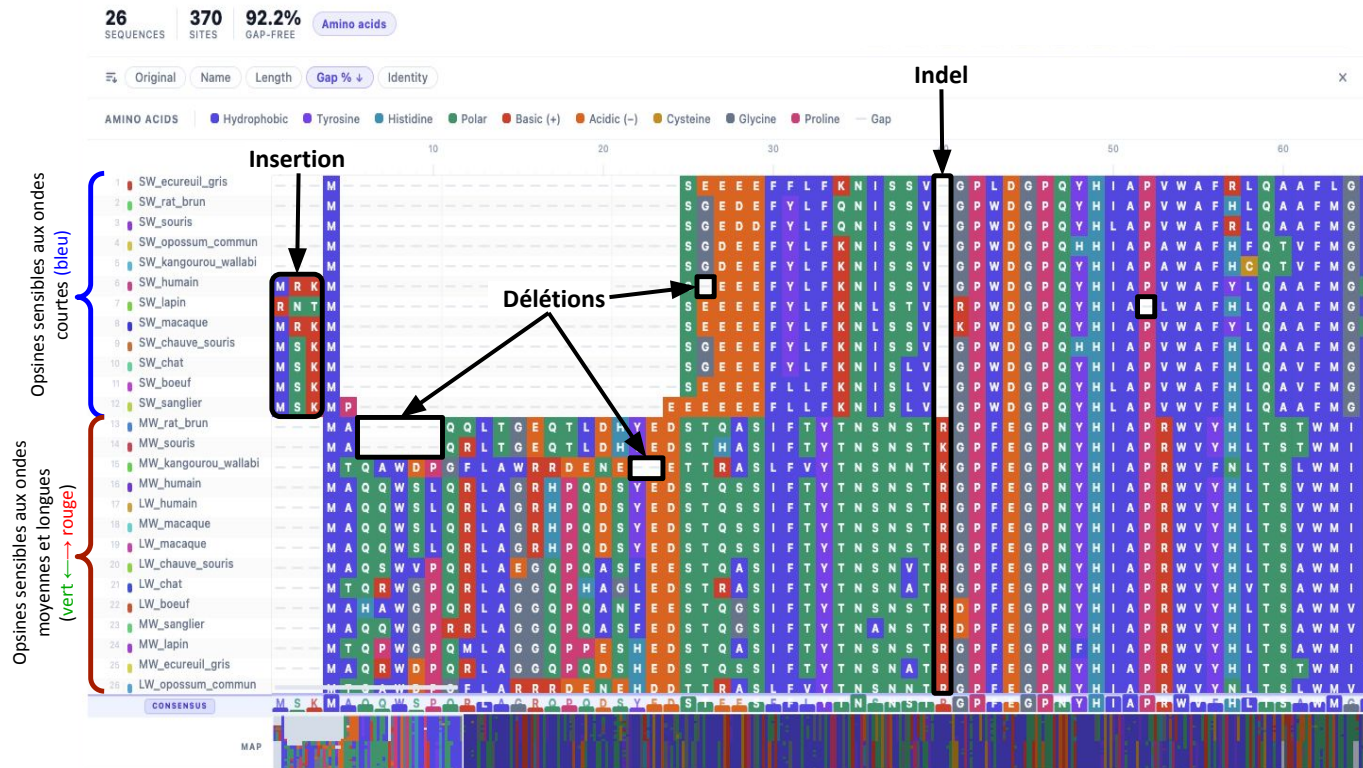
- L'alignement multiple permet de distinguer des **blocs conservés**, soit sur l'ensemble de l'alignement, soit sur un sous-ensemble des séquences.
- Les **gaps** (espacements, représentés par des "-") permettent d'aligner les séquences identiques ou similaires.



Alignement multiple des opsines de mammifère

- Contrairement aux alignements par paire, on peut *dans certains cas* distinguer les délétions des insertions
 - **Délétion**: 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
- Cependant, certaines positions restent non-résolues (**indel**) car les deux lignées ont soit le gap, soit un résidu.

Note : l'inférence de l'événement évolutif, délétion ou insertion, qui est à l'origine d'un gap, nécessite généralement d'interpréter les alignements en les comparant à l'arbre phylogénétique inféré.



Inférence d'un arbre des séquences 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 ont des avantages et inconvénients respectifs, et 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

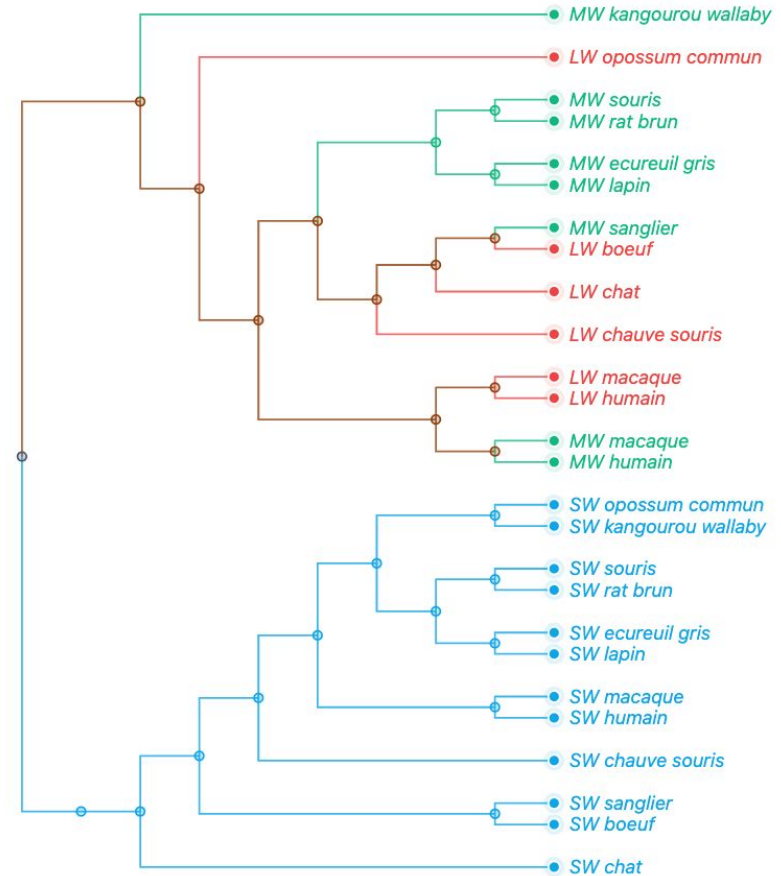
- Rapide
- Permet de traiter un grand nombre de séquences
- Permet d'estimer la longueur des branches (produit un phylogramme)
- Résultats moins fiables que par la méthode du maximum de vraisemblance

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)

Cladogramme des séquences

- Cladogramme généré avec l'outil en ligne Phylogeny.fr
- Coloration des branches → sensibilité des opsines
 - Bleu: SWS
 - Vert : MWS
 - Rouge: LWS
- Enracinement au point moyen
- Parfaite séparation entre SWS et les autres
- Par contre, les MWS et LWS sont mélangés



Phylogramme des séquences

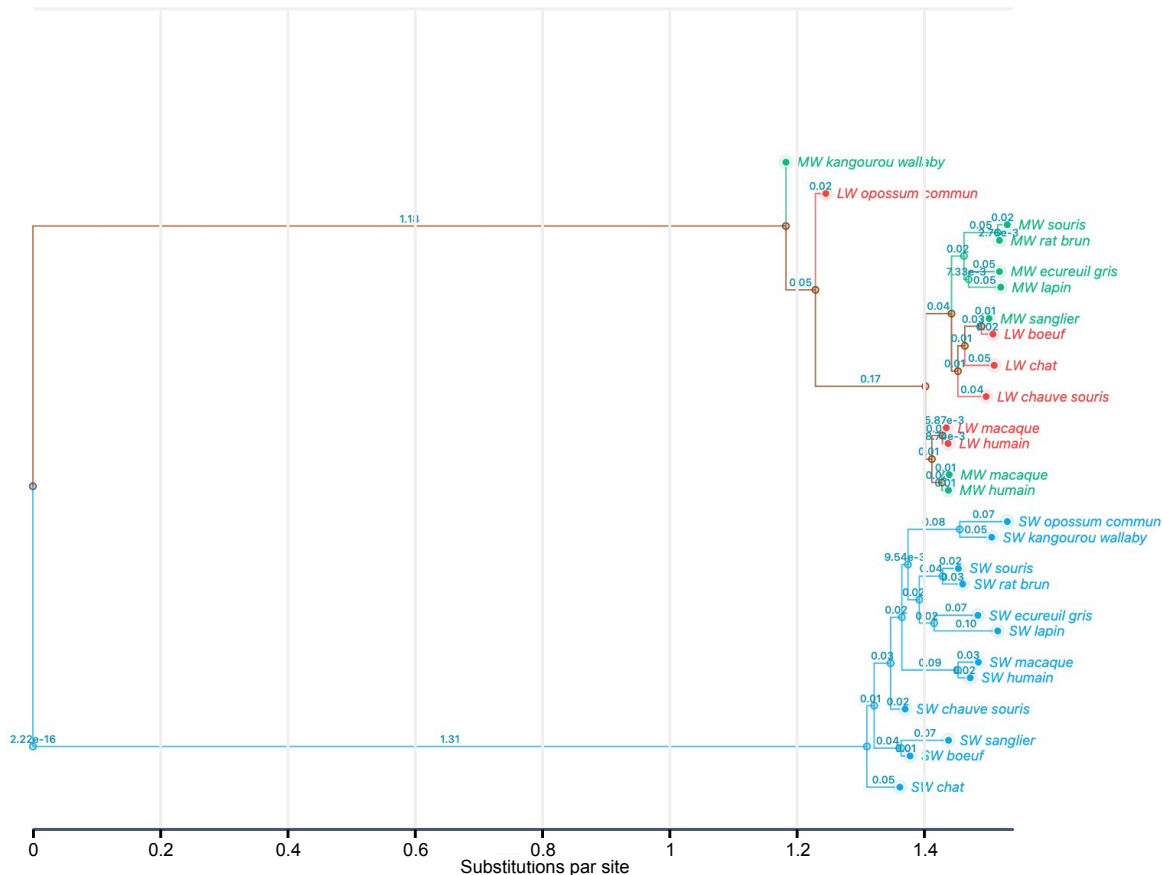
Méthode

- Phylogramme généré avec l'outil en ligne Phylogeny.fr
- Coloration des branches → sensibilité des opsines
 - Bleu: SWS
 - Vert : MWS
 - Rouge: LWS
- Enracinement au point moyen

Observations

- Parfaite séparation entre SWS et les autres
- Par contre les MWS et LWS sont mélangés
- Les **nombre**s associés aux **branches** indiquent le **nombre de substitutions par site**.
- **Grande distance** entre SWS et les autres : en moyenne, >1 substitution par site (colonne)

Phylogramme de 26 opsines de mammifère



Bootstrapping – à quel point les branchements sont-ils supportés par les données ?

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

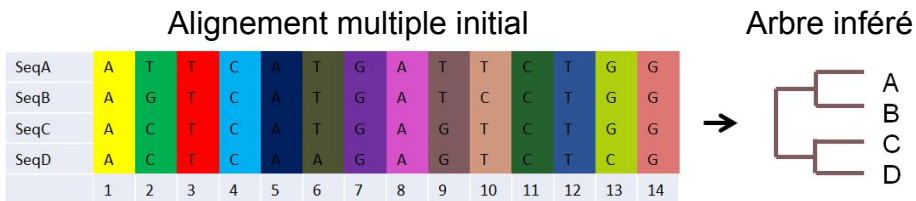
On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

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)



Bootstrapping – à quel point les branchements sont-ils supportés par les données ?

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

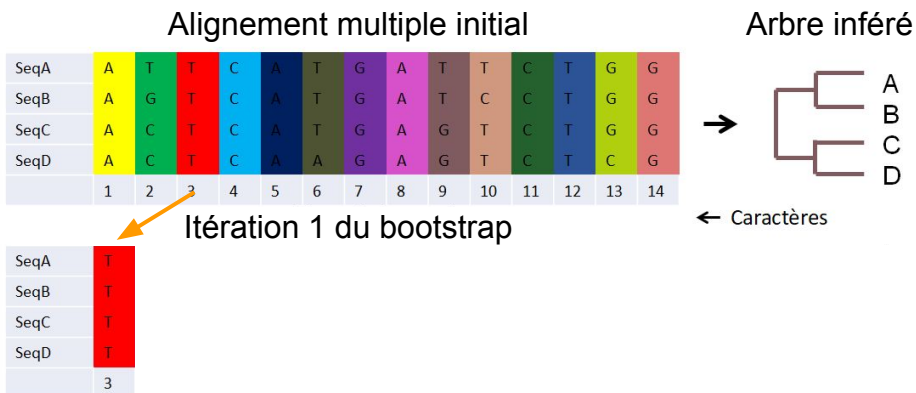
On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

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)



Bootstrapping – à quel point les branchements sont-ils supportés par les données ?

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

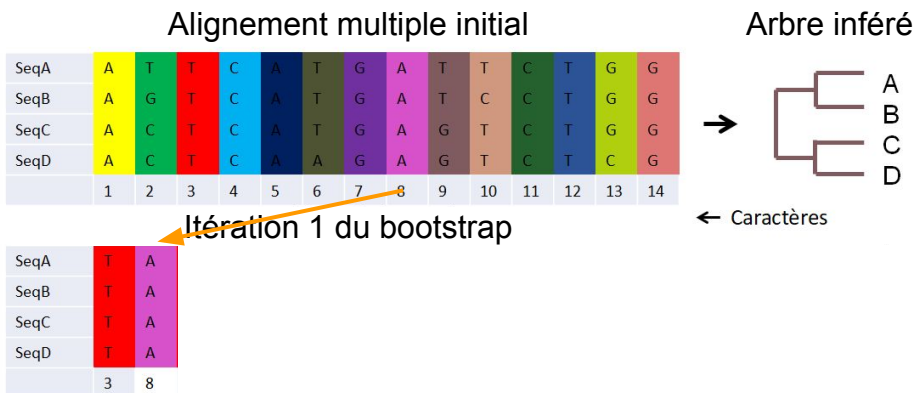
On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

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)



Bootstrapping – à quel point les branchements sont-ils supportés par les données ?

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

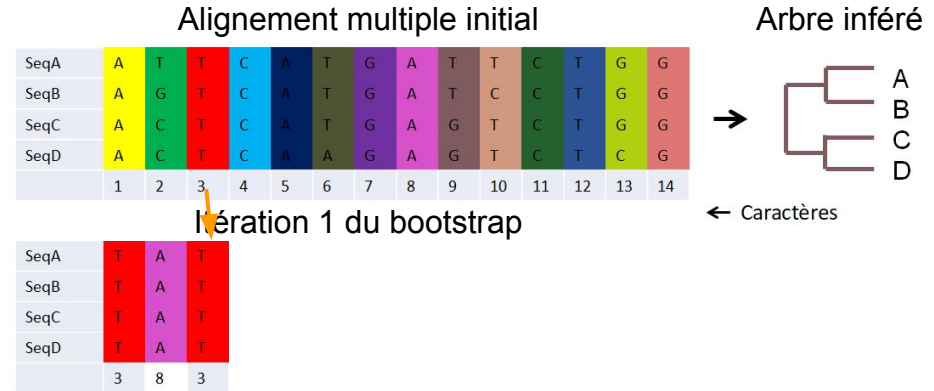
On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

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)



Bootstrapping – à quel point les branchements sont-ils supportés par les données ?

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

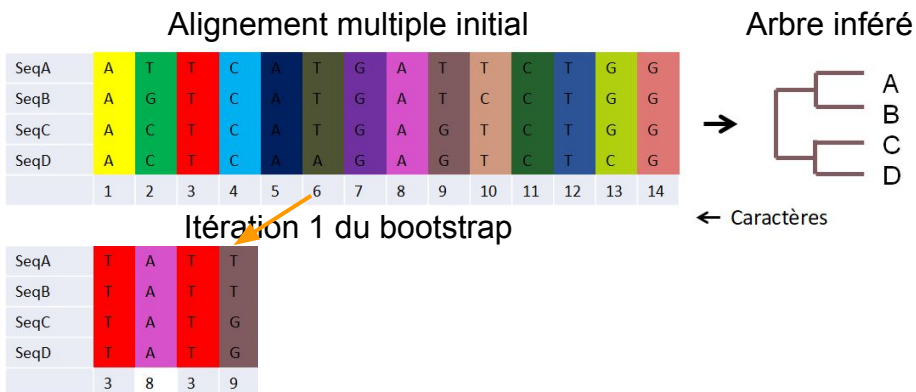
On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

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)



Bootstrapping – à quel point les branchements sont-ils supportés par les données ?

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

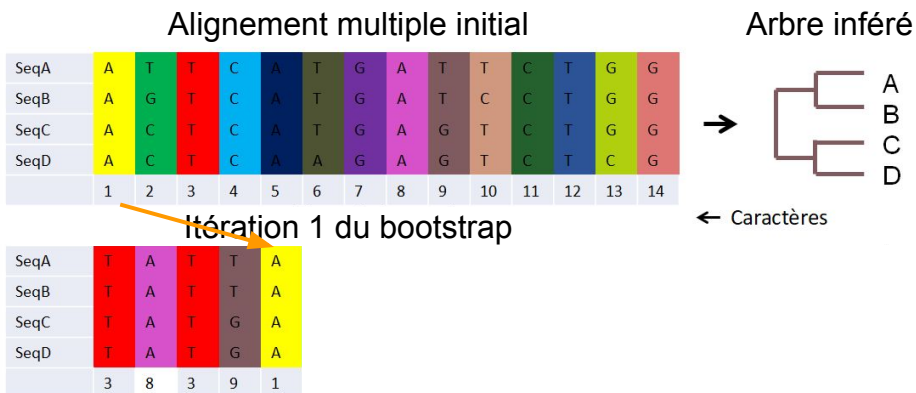
On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

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)



Bootstrapping – à quel point les branchements sont-ils supportés par les données ?

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

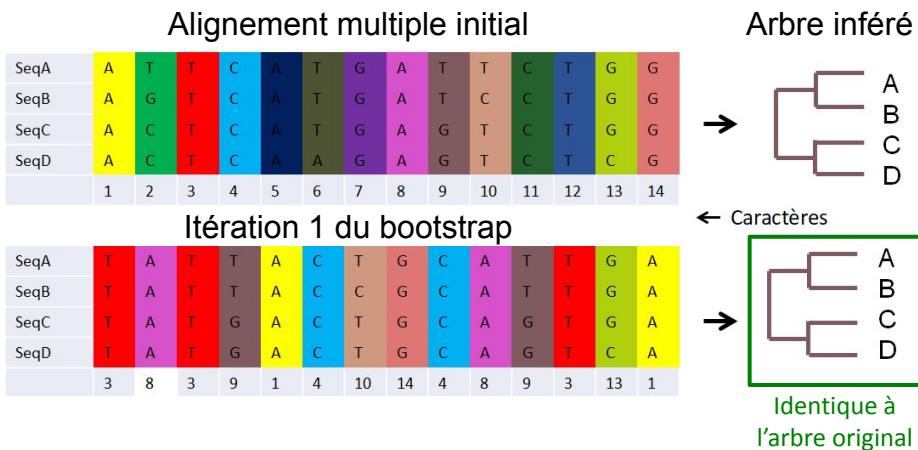
On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

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)



Bootstrapping – à quel point les branchements sont-ils supportés par les données ?

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

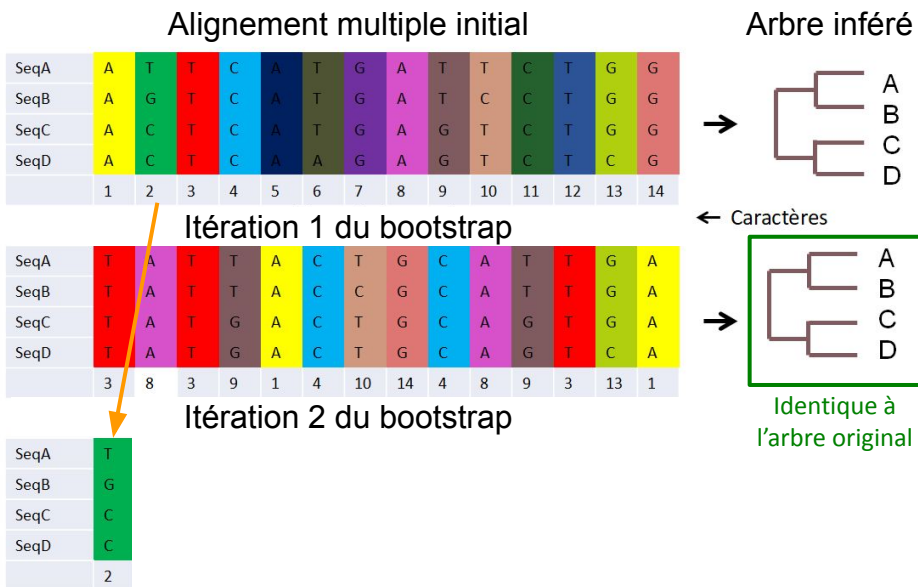
On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

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)



Bootstrapping – à quel point les branchements sont-ils supportés par les données ?

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

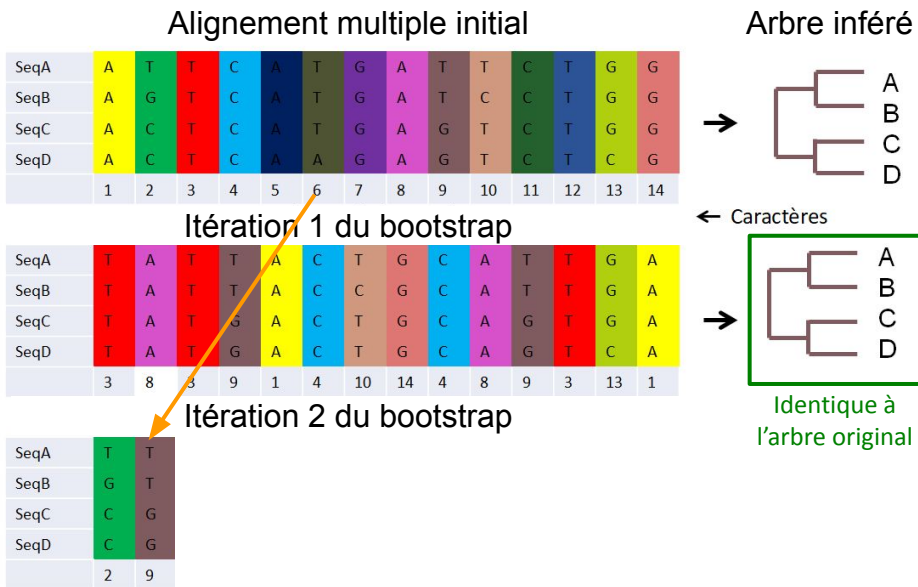
On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

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)



Bootstrapping – à quel point les branchements sont-ils supportés par les données ?

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

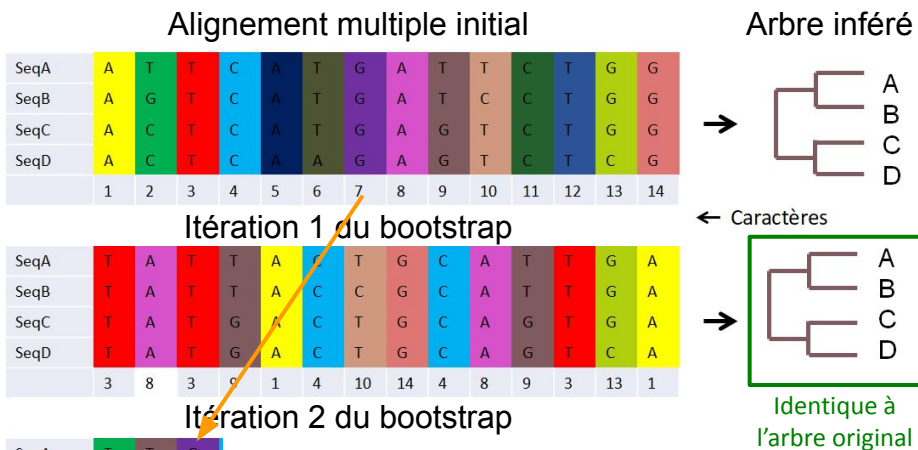
On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

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)



Bootstrapping – à quel point les branchements sont-ils supportés par les données ?

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

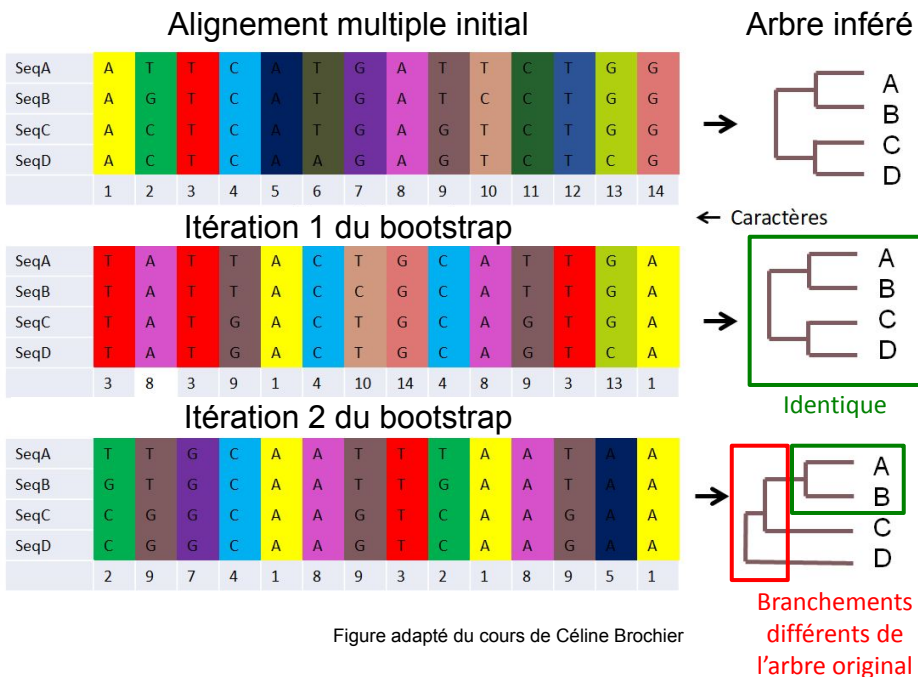
On peut s'interroger sur la fiabilité de cette inférence, qui dépend des séquences particulières dont on dispose dans l'échantillon analysé.

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)



Bootstrapping – à quel point les branchements sont-ils supportés par les données ?

- 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 au rééchantillonnage**
- **Valeur faible** → branchement sensible au rééchantillonnage

Exemple : 36/100

→ branchement identique seulement dans ~1/3 des bootstraps

→ dépend fortement du choix des colonnes

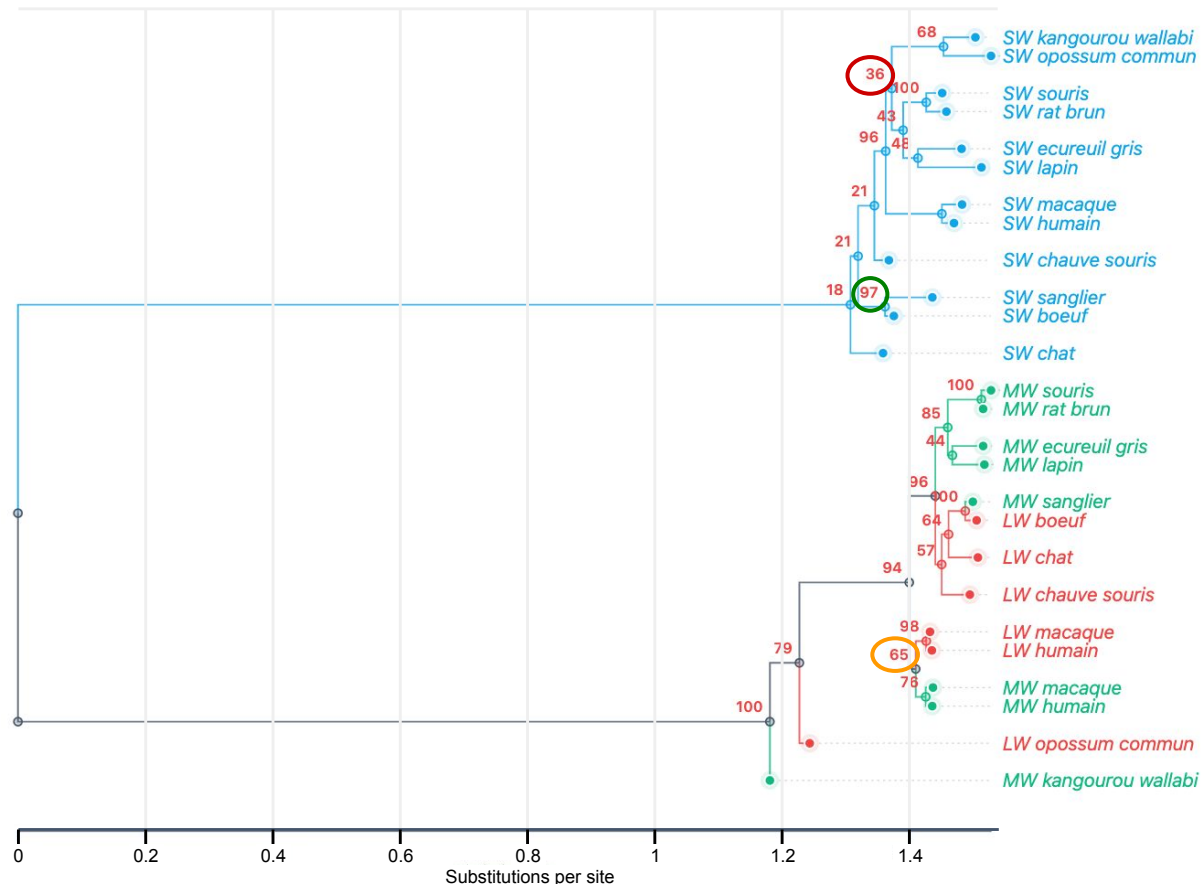
→ pas représentatif de l'alignement complet.

Attention ! Le bootstrap nous informe sur la robustesse de l'arbre inféré par rapport aux données, mais ne garantit pas la pertinence des données par rapport à l'histoire évolutive réelle ("arbre vrai").

Problèmes potentiels

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

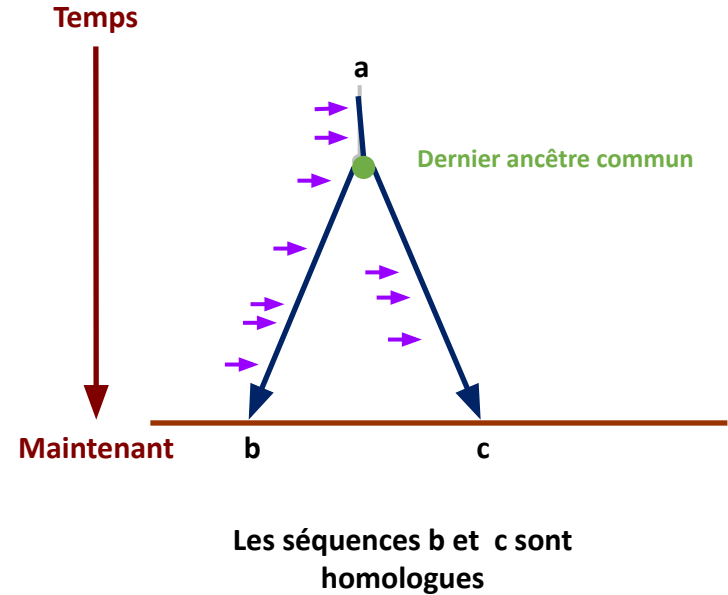
Phylogramme de 26 opsines de mammifère



Relations entre séquences : homologie, analogie, paralogie, orthologie, xénologie

Séquences homologues ou analogues

- Comme pour les caractères phénotypiques, les ressemblances entre deux séquences macromoléculaires peuvent résulter de différentes causes évolutives.
- **Séquences homologues**
 - Ressemblances héritées d'une **séquence ancestrale commune**
 - Évolution **divergente** : accumulation de mutations à partir de la séquence ancestrale commune
 - Constituent le type de données principal de la **phylogénie moléculaire**. Exemples développés ci-après.
- **Séquences analogues**
 - Ressemblances ne provenant pas d'une séquence ancestrale commune
 - Évolution **convergente** : les ressemblances sont apparues indépendamment dans les deux lignées, sous l'effet de facteurs de sélection communs.
 - Généralement limitée à de courtes séquences.
 - Exemple : éléments cis-régulateurs assurant la liaison d'un même facteur dans les régions régulatrices de différents gènes.



- 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.
 - **Attention, les pourcentages peuvent être trompeurs !** Pour des séquences courtes, une forte ressemblance, voire une identité parfaite, peuvent éventuellement provenir d'une évolution convergente
- L'inférence d'homologie est toujours attachée à un certain **risque de faux positifs**. Les modèles évolutifs nous permettent d'estimer ce risque.
 - E-valeur élevée (>1) ? → similarités vraisemblablement dues au hasard → Faux-positif
 - E-valeur très faible (ex : $1e-5$, $1e-100$, ...) ? → très significatif
→ très improbable que la similarité résulte du hasard
→ hypothèse alternative : la ressemblance entre les séquences résulte d'un héritage commun à partir d'une séquence ancestrale → **hypothèse d'homologie**

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 inapproprié** de parler de « niveau d'homologie » ou « pourcentage d'homologie ».

Formulation correcte

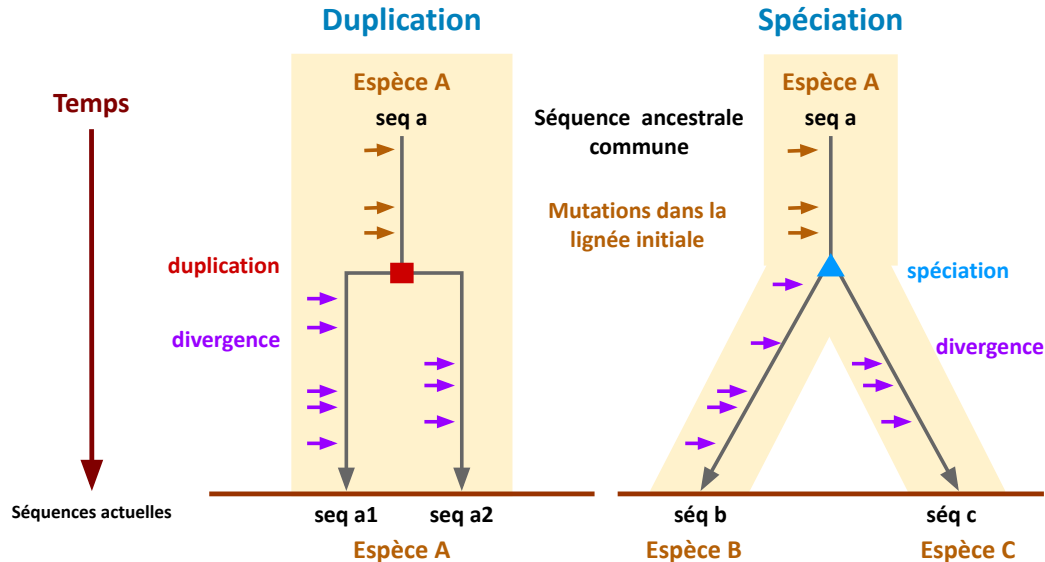
- On observe un certain niveau de similarité entre deux séquences (pourcentages de résidus identiques, pourcentages de résidus « similaires »).
- Sur cette base, on évalue deux scénarios évolutifs: cette similarité peut provenir d'une évolution convergente (analogie) ou divergente à partir d'un ancêtre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

A l'origine des séquences homologues : duplications et spéciations

- En phylogénie moléculaire, deux événements évolutifs jouent un rôle particulièrement important dans l'apparition de séquences homologues : les duplications et les spéciations.
- **Duplication**
 - Une duplication est une mutation qui provoque un dédoublement d'une partie de l'ADN génomique. Elle peut concerner 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, créant ainsi une 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 aboutit à la formation d'espèces distinctes à partir d'une lignée évolutive unique.
 - Au moment de la séparation, les deux lignées héritent des mêmes gènes
 - Les mutations s'accumulent ensuite indépendamment dans les deux lignées
- Les duplications et spéciations 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**

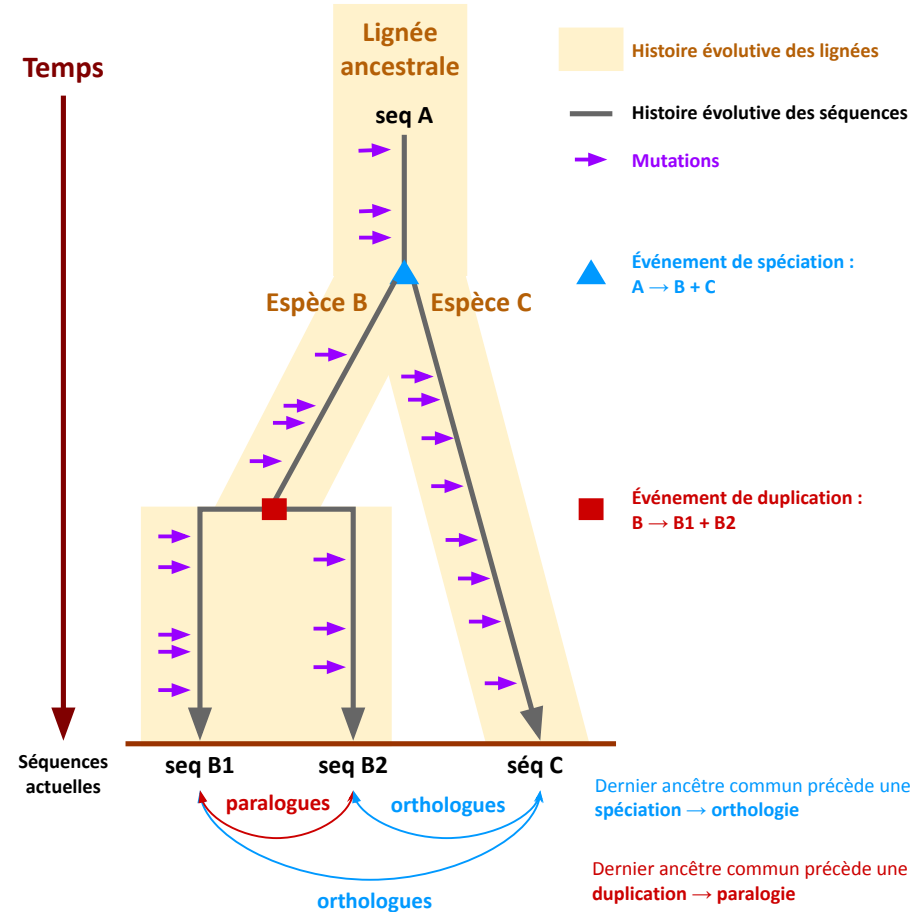
Scénarios évolutifs

- Nous disposons de deux séquences, et nous supposons qu'elles divergent d'un ancêtre commun.
- A priori, la divergence entre ces deux séquences peut résulter
 - d'une **spéciation** (formation d'espèces séparées à partir d'une espèce unique).
 - d'une **duplication** (création de deux copies du gène dans le même génome)
- Les flèches horizontales indiquent les mutations (substitutions, délétions, insertions) qui s'accumulent au sein des séquences au cours de leur histoire évolutive. Ces mutations se produisent déjà dans la lignée initiale de séquence (**flèches ocre**), mais suite à la spéciation ou à la divergence elles provoquent la diversification progressive des séquences, des structures et des fonctions (**flèches violettes**).



Orthologie versus paralogie

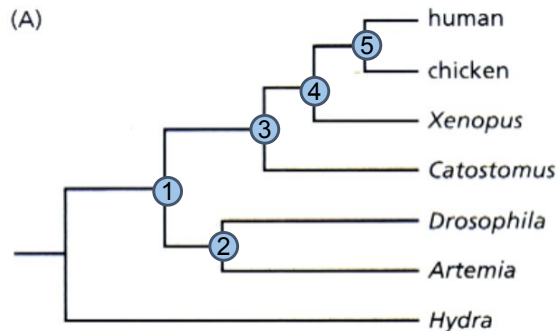
- Zvelebil & Baum (2000) fournissent une définition claire et opérationnelle des concepts d'orthologie et de paralogie.
 - **Orthologues** : séquences dont le dernier ancêtre commun précède immédiatement un événement de **spéciation**.
 - **Paralogues** : séquences dont le dernier ancêtre commun précède immédiatement un événement de **duplication**
- Exemples:
 - C et B1 sont **orthologues**, car leur dernier ancêtre commun (A) précède immédiatement un événement de **spéciation** ($A \rightarrow B + C$).
 - C et B2 sont également **orthologues**. Une séquence peut donc avoir plusieurs orthologues chez une autre espèce.
 - B1 et B2 sont **paralogues** car leur dernier ancêtre commun (B) précède immédiatement un événement de **duplication** ($B \rightarrow B1 + B2$).



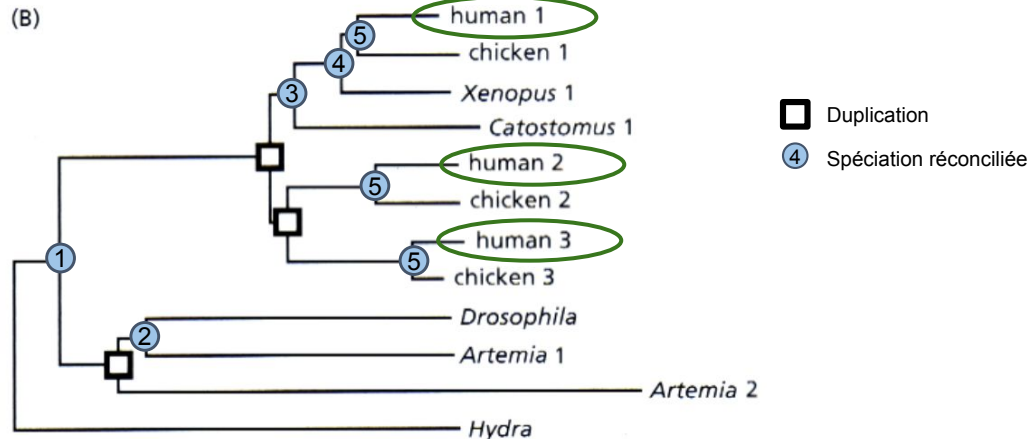
Réconciliation phylogénétique : combiner les arbres des espèces et des molécules

- 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 séquences → chaque noeud interne représente soit une duplication, soit une spéciation
- En comparant l'arbre des séquences et l'arbre des espèces, on peut inférer l'histoire évolutive de cette famille de séquences.
- Note : sur l'arbre des séquences, un même organisme peut éventuellement se retrouver plusieurs fois, s'il existe plusieurs molécules homologues dans son génome, suite à des duplications.
- Après une duplication, l'une des copies peut disparaître de certaines lignées (ex : pas de *Drosophila* 2)

Arbre des espèces



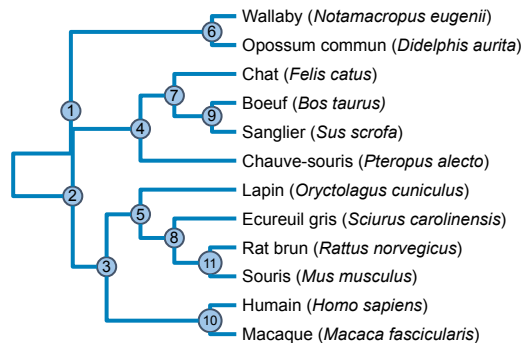
Arbre des séquences



La réconciliation ne fonctionne pas toujours

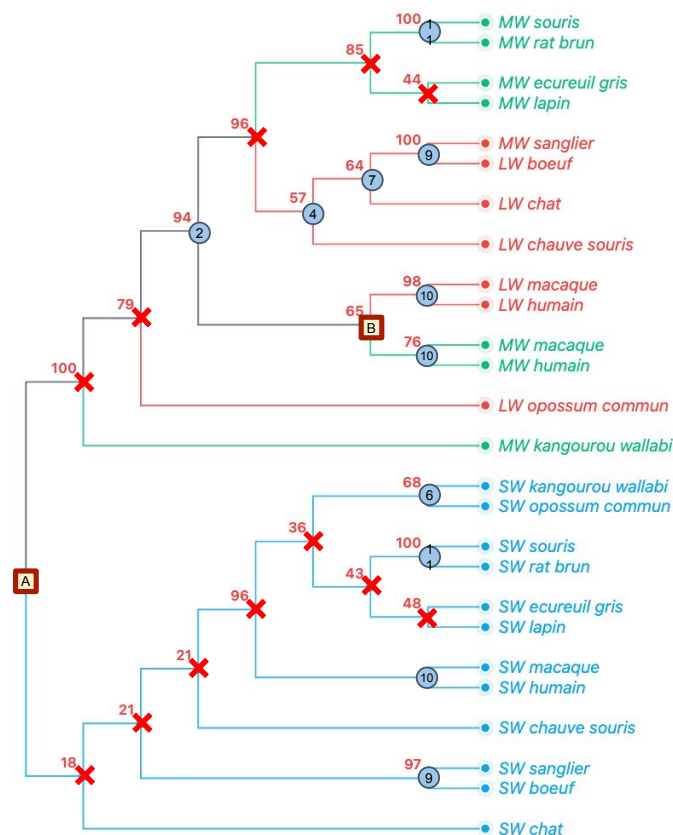
- **Arbre des espèces** : 12 espèces de mammifères
- **Arbre des séquences** : 26 opsines
- **Duplication initiale [A]** à l'origine de la vision dichromatique (SWS versus LWS/MWS)
- **Seconde duplication [B]** chez un ancêtre commun des primates de l'ancien monde, à l'origine de la vision trichromatique
- Les cercles numérotés indiquent **noeuds cohérents** entre arbre des séquences et arbre des espèces (cercles numérotés).
- Il existe aussi des **noeuds irréconciliables (X)**, où l'arbre des séquences est incohérent avec celui des espèces. Ces noeuds correspondent souvent (mais pas toujours) à une valeur très faible de bootstrap.

Arbre des espèces



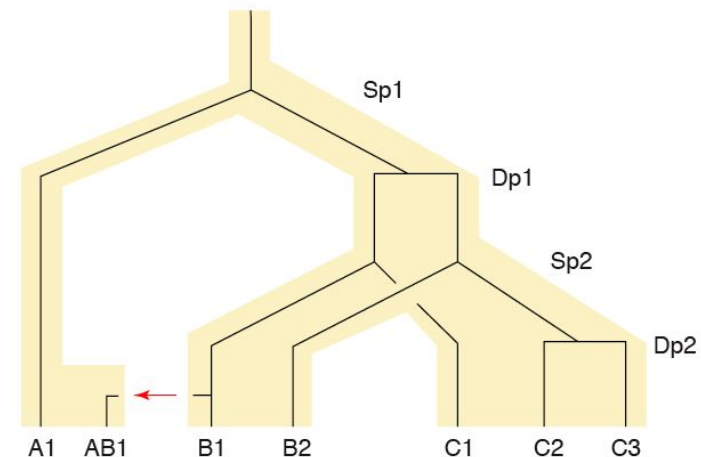
- Duplication
- ④ Spéciation réconciliée
- ✗ Noeud non réconciliable

Arbre des séquences (opsines)



Représentation détaillée des événements de spéciation / duplication

- La figure de droite combine deux niveaux de représentation
 - Les ombrages épais représentent l'**arbre des espèces**.
 - Les lignes noires fines représentent l'**arbre des séquences**, qui retrace les relations évolutives entre elles au fil des événements évolutifs (spéciations, duplications).
- 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.



trends in Genetics

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 $B2 \Rightarrow C1 \Rightarrow C2$.

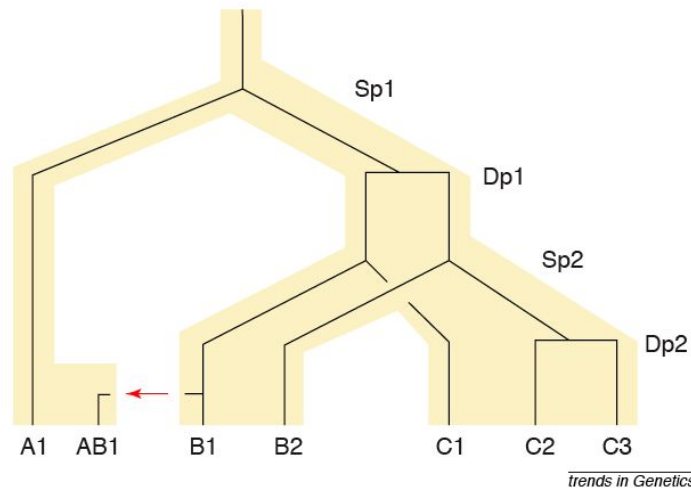
Exercice: types d'homologie

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 xénologie
A analogie

Les solutions sont fournies à la fin de ce diaporama

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



- Paire d'**orthologues**: paire de séquences 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 séquences 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.*

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).

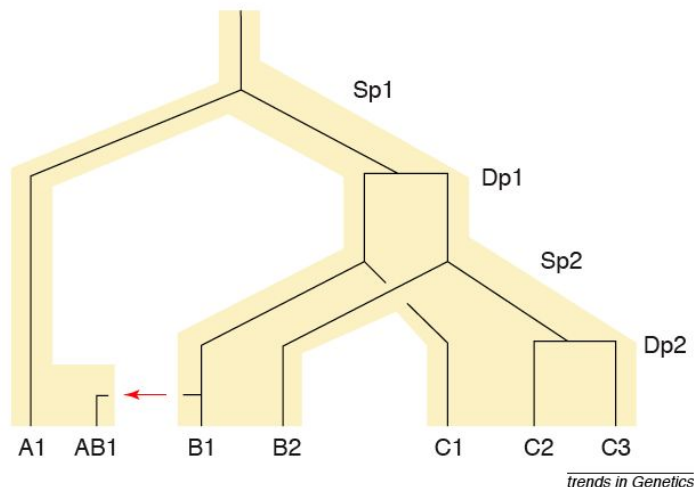
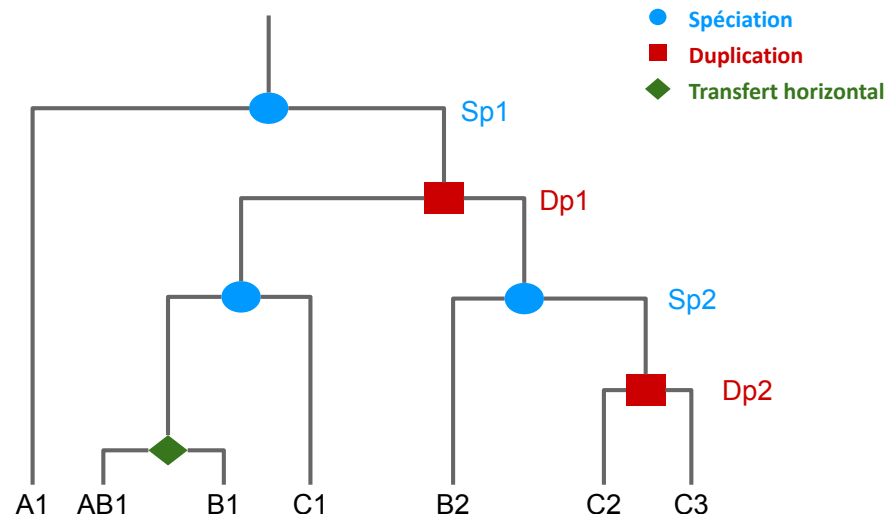
Représentation classique des spéciations / duplications

Les représentations de gauche et de droite sont équivalentes.

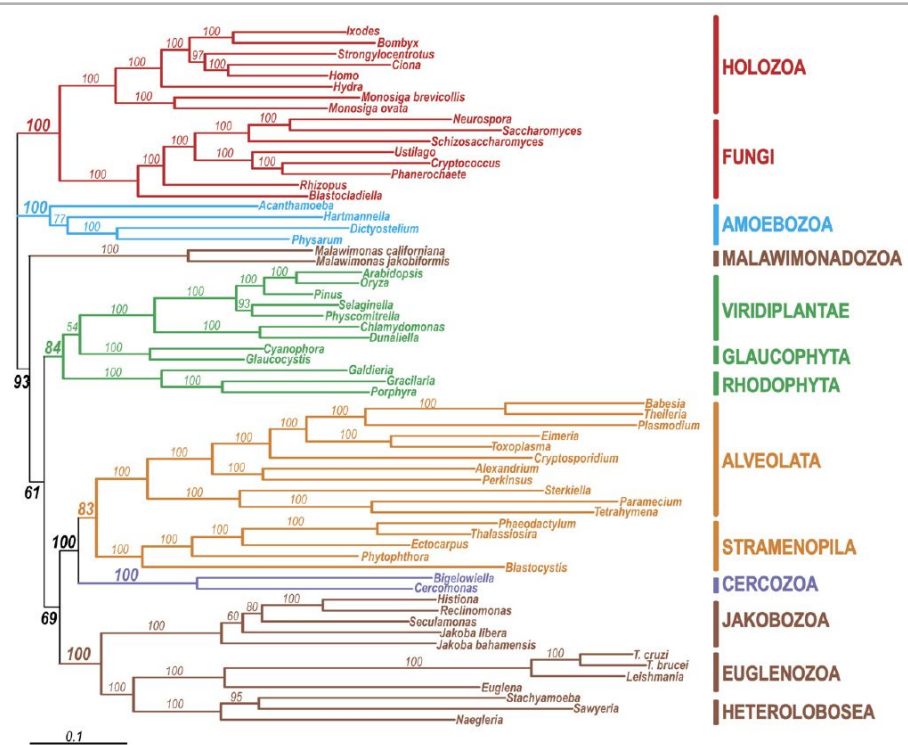
- Celle de gauche (Fitch, 2000) montre bien la transmission parallèle des unités moléculaires (gènes, protéines) qui résultent d'une duplication (bifurcation rectangulaire) au fil des spéciations (bifurcations triangulaires). Le fond jaune représente l'histoire évolutive des lignées, les lignes noires celle des séquences.
- La représentation de droite est celle utilisée par les bases de données telles qu'[Ensembl genomes](#). Sur l'arbre des séquences, chaque noeud interne est marqué par une couleur différente selon qu'il correspond à une spéciation ou à une duplication.

Les lettres **A, B, C** représentent les **espèces**.

Les **numéros 1, 2, 3** distinguent les copies des **gènes** au sein d'une même espèce



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



- 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

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. Ils étaient donc classés dans le même ordre (insectivores) mais dans une famille séparée.

Description des hérissons et des tenrecs dans l'encyclopédie Bordas de 1981.

Hérisson

classe	Mammifères
ordre	Insectivores
famille	Erinacéidés
genre et espèce	Erinaceus europaeus et autres

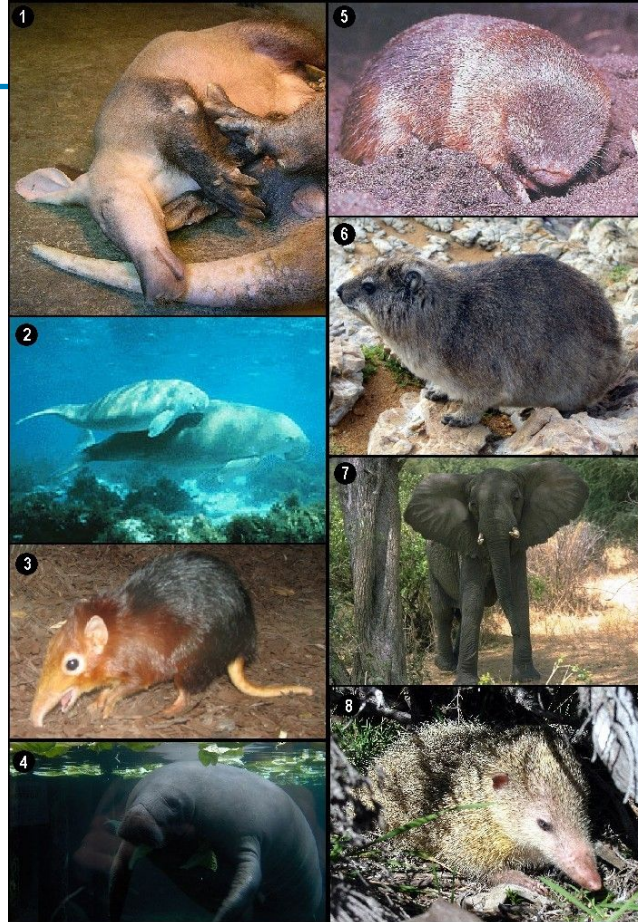
Tenrec

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



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, lamentin, dugong, oryctérope, taupe dorée, daman, ...



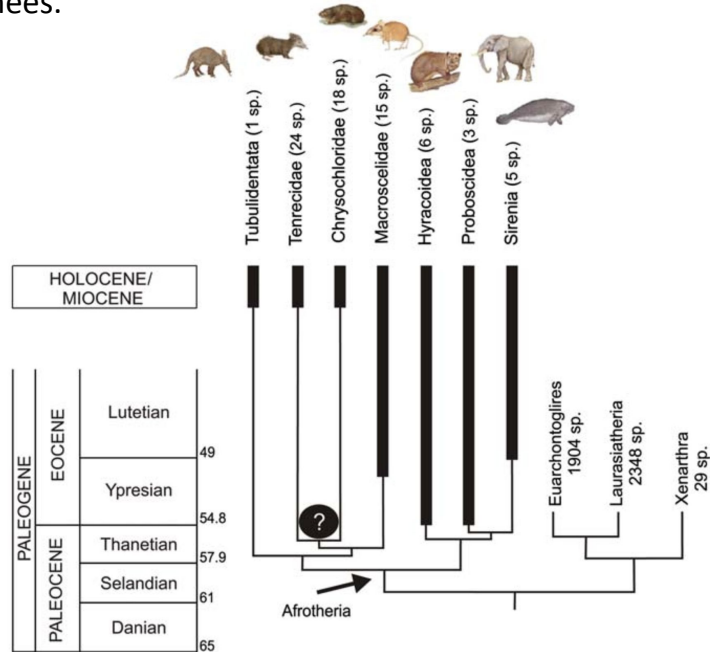
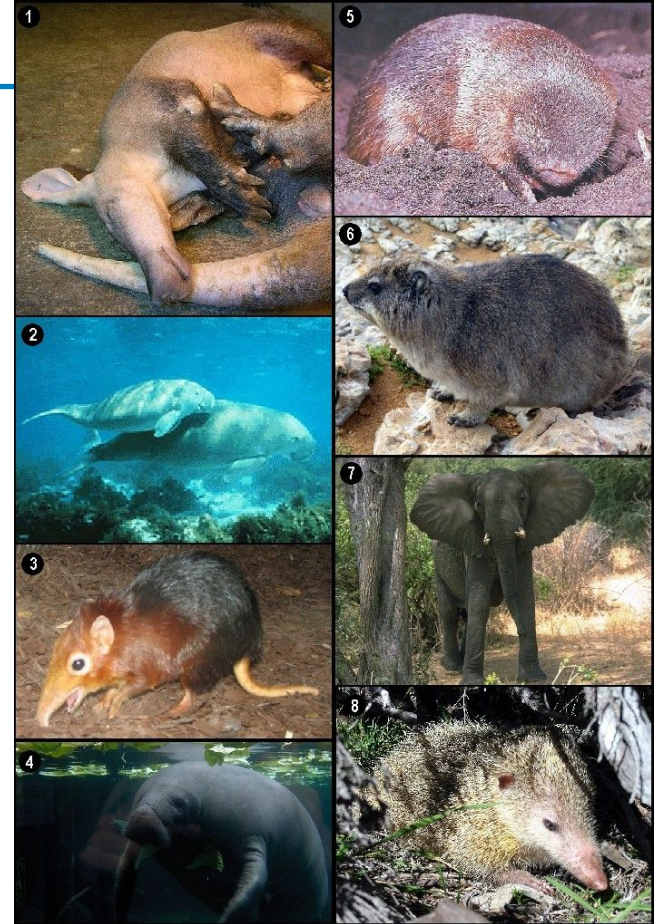
1. Oryctérope du Cap; 2. Dugongs; 3. Macroscélide de Peters ; 4. Lamantin; 5. Taupe dorée; 6. Daman du Cap; 7. Éléphant de savane d'Afrique ; 8. Tangué ("tailless tenrec"); 8, 8b: Tenrec

<http://upload.wikimedia.org/wikipedia/commons/0/01/Kleiner-igeltanrek-a.jpg>

- 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

Les Afrotheria

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



1. Oryctérope du Cap; 2. Dugongs; 3. Macroscélide de Peters ; 4. Lamantin; 5. Taupe dorée; 6. Daman du Cap; 7. Éléphant de savane d'Afrique ; 8. Tangué ("tailless tenrec"); 8b: Tenrec

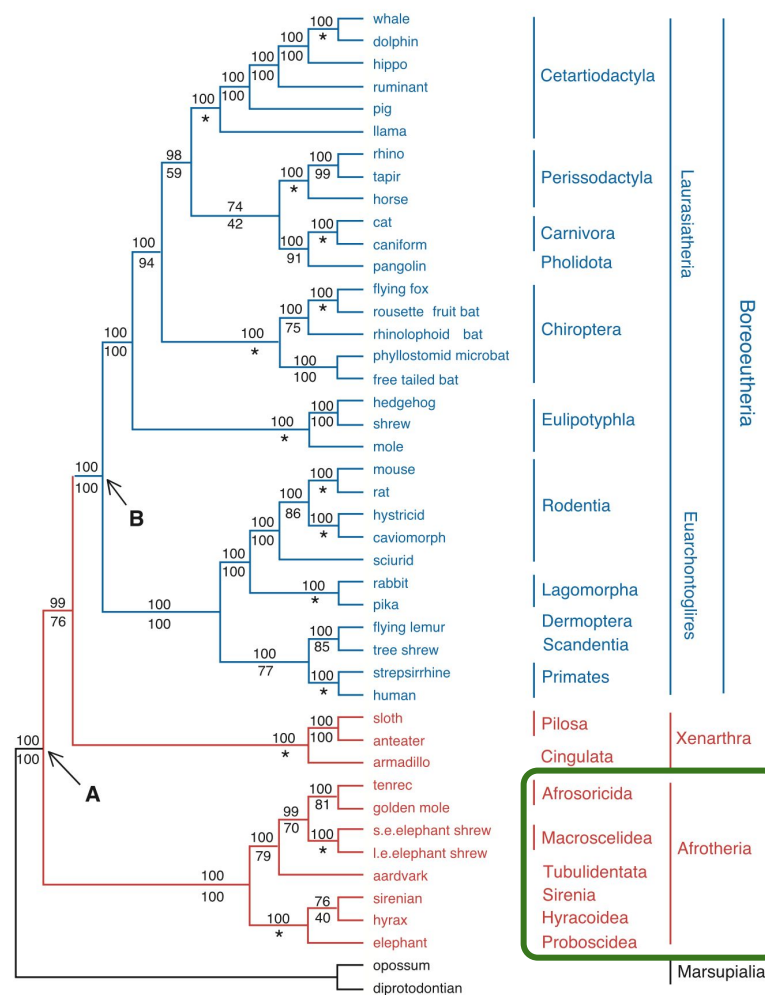
- Murphy, W. J. et al. Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenetics. Science 294, 2348–2351 (2001). doi.org/10.1126/science.1067179
- Tabuce, R., Asher, R. J. & Lehmann, T. Afrotherian mammals: a review of current data. mammalia 72, (2008). doi.org/10.1515/MAMM.2008.004

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 (75) 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 (75). Values below branches represent percent support in maximum likelihood (GTR + Γ + 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 (75). Blue, monophyletic Northern Hemisphere group (i.e., Boreoeutheria); red, paraphyletic Southern Hemisphere group (i.e., Xenarthra + Afrotheria); black, outgroups.

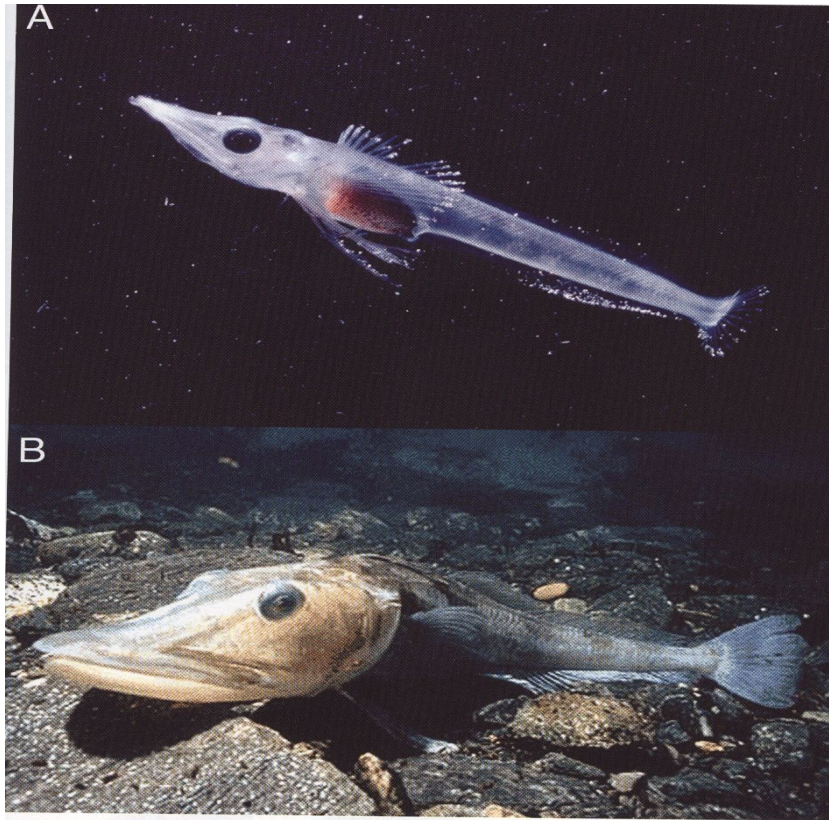


Pseudo-gènes (“gènes fossiles”) : les reliques de gènes ancestraux dans les génomes d’espèces actuelles

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 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 antigel**, 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 pseudogénisés (“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

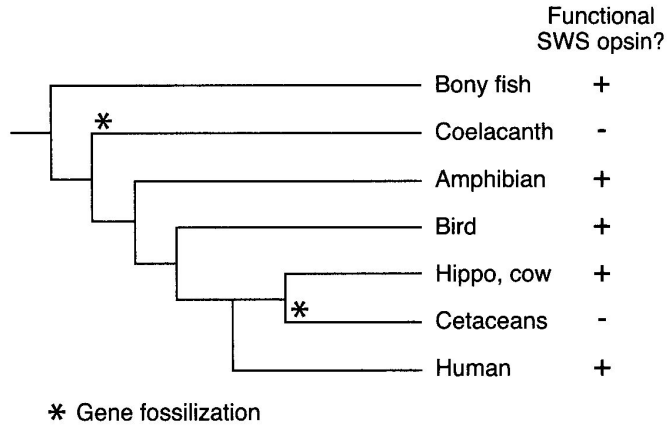


FIG. 5.2. **The same opsin gene has been fossilized twice.** The distribution of different mutations found in the coelacanth and cetacean SWS opsins and the evolutionary relationship of these species indicates that the SWS opsin was fossilized at least twice (asterisk). *Figure by Jamie Carroll.*

- Coelacanth et les cétacés ont perdu la perception des couleurs
- Perte indépendante chez les coelacanth 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).

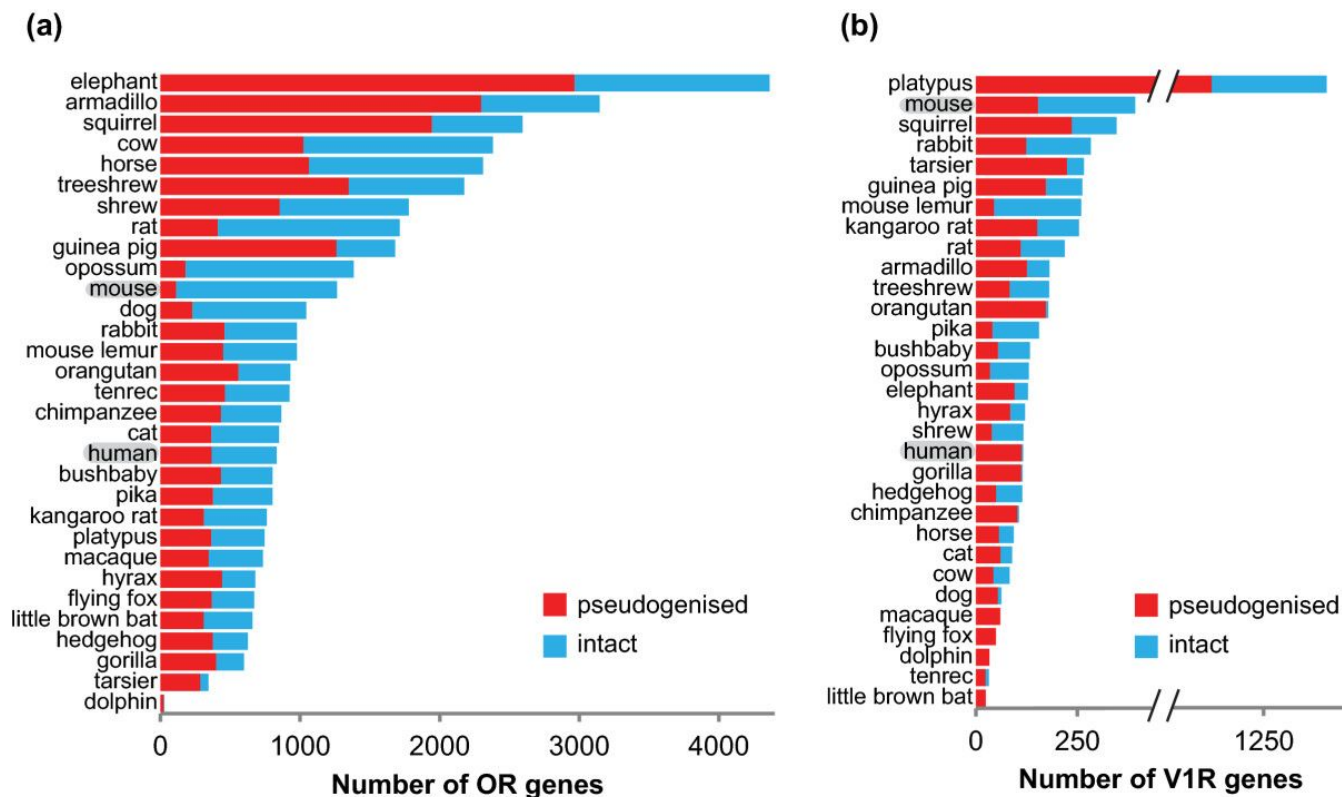
La pseudogénisation des récepteurs olfactifs (OR) et des vomérorécepteurs (VR)

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 (V1R)

- 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.



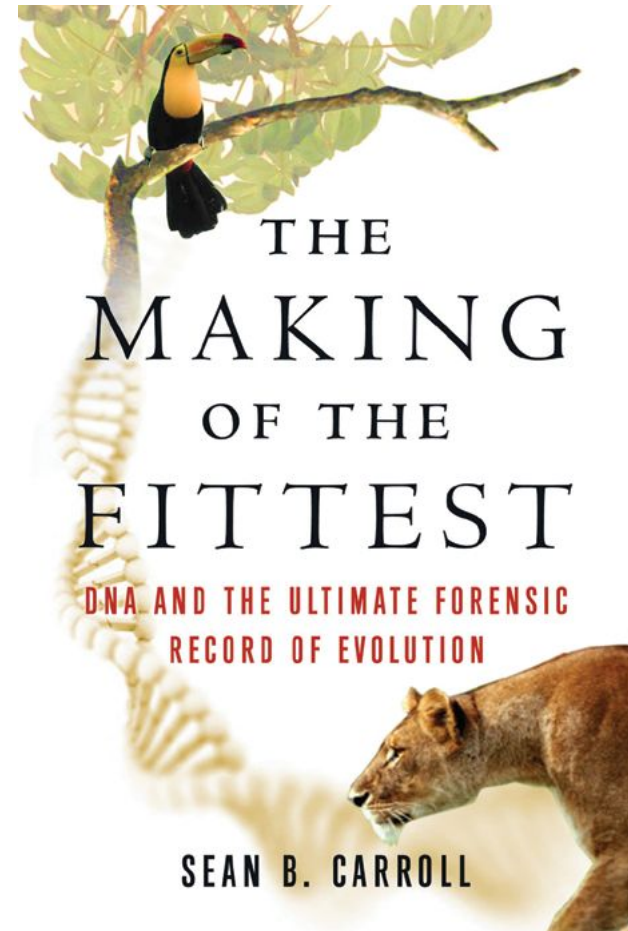
- *Mycobacterium tuberculosis*
 - Extracellulaire
 - 4.189 gènes codant pour des protéines
- *Mycobacterium leprae*
 - Intra-cellulaire
 - 1.605 gènes codant pour des protéines
 - A perdu (par délétions) ~1000 gènes présents chez *M.tuberculosis*
 - On trouve également plus de 1.000 gènes fossiles.

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' through a¹⁰) 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).

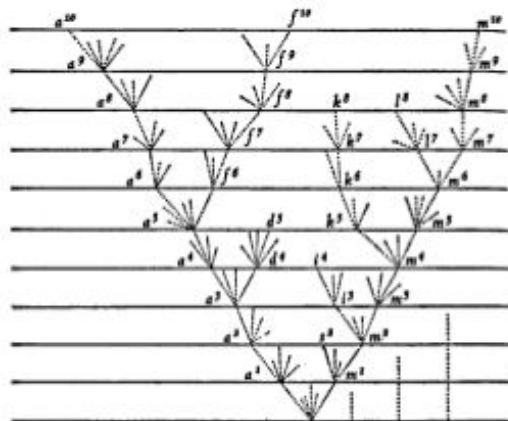
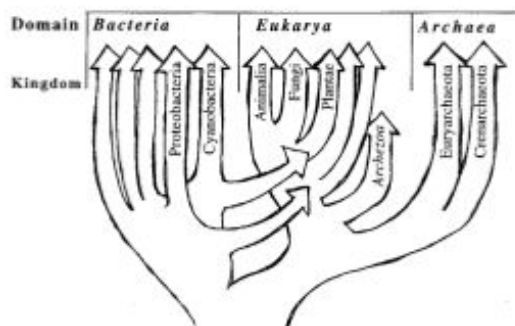


Fig. 2. The current consensus or standard model. Only a few of the "kingdoms" of the "domain" Bacteria are shown. Branching orders of several kingdoms within Bacteria and Eukarya remain in dispute. Mitochondrial and chloroplast endosymbioses are indicated by lower and upper diagonal arrows, respectively.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

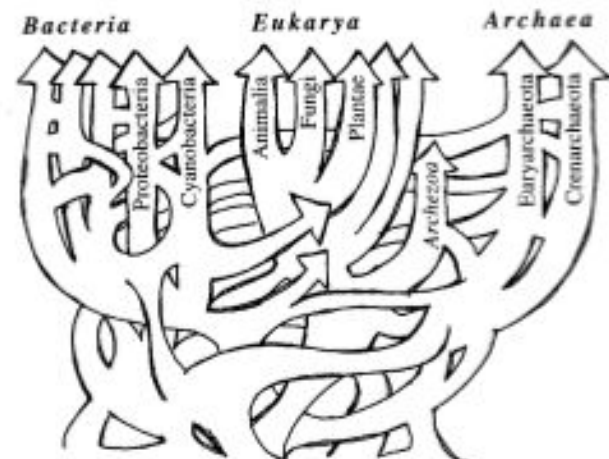


L'arbre de la vie de Darwin (Fig 1) est revisité par Doolittle (1999) pour tenir compte

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.

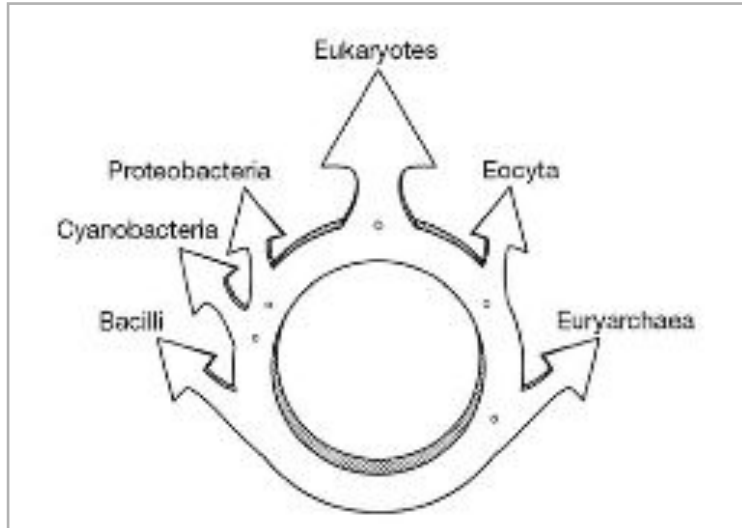


The ring of life provides evidence for a genome fusion origin of eukaryotes

Maria C. Rivera^{1,3,4} & James A. Lake^{1,2,4}

¹Molecular Biology Institute, ²MCD Biology, ³Human Genetics, ⁴IGPP and ⁵Astrobiology 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**.

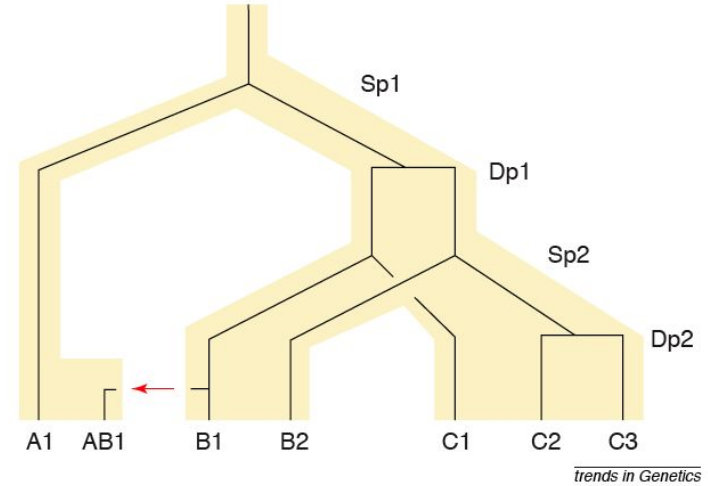
Solution de l'exercice orthologies / paralogies / xénologies (Fitch)

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 xénologie
A analogie

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



- 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.*

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).

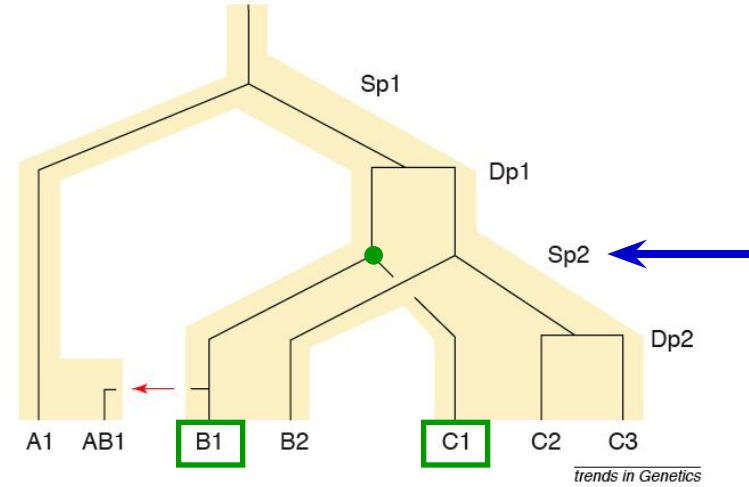
Exercice: types d'homologie

Exemple: B1 versus C1

- Les deux séquences (B1 and C1) appartiennent respectivement aux taxa B and C.
- Le cénancêtre est le dernier gène ancestral commun (**point vert**) entre B1 et C2.
- Ce cénancêtre précède immédiatement l'évènement de spéciation Sp2 entre les taxa B et C (**flèche bleue**)

→ **B1 et C1 sont orthologues**

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



- 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.*

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).

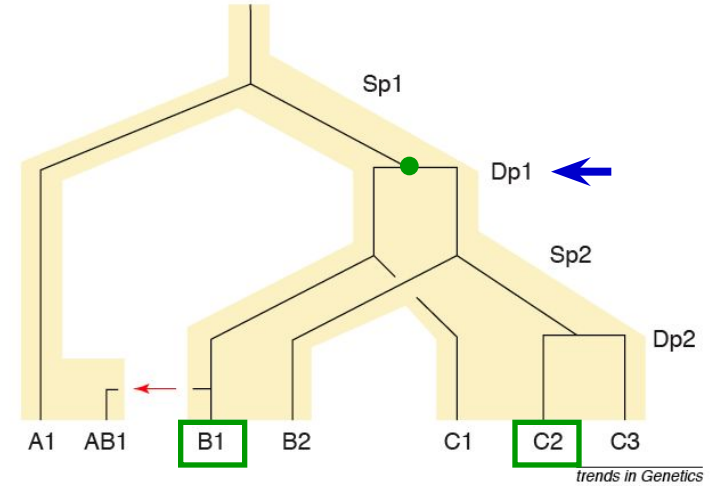
Exercice: types d'homologie

Exemple: B1 versus C2

- Les deux séquences (B1 and C2) appartiennent respectivement aux taxa B and C.
- Le dernier gène ancestral commun (**point vert**) est celui qui précède immédiatement la duplication Dp1 (**flèche bleue**).

→ **B1 et C2 sont paralogues**

	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2							
C1			O				
C2			P				
C3							



- 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.*

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).

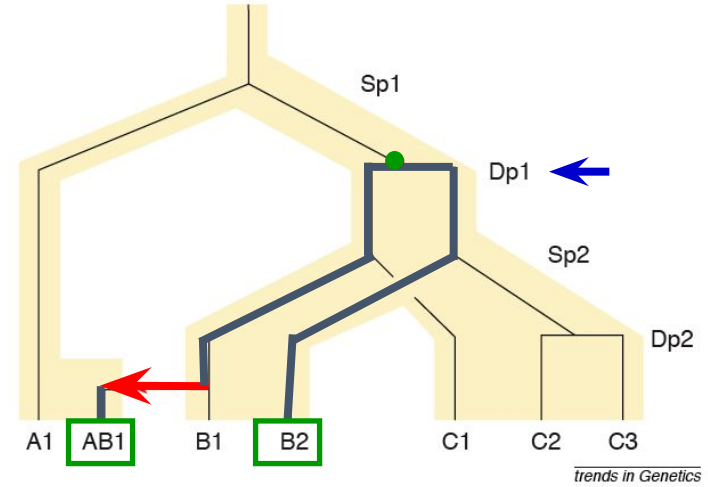
Exercice: types d'homologie

Exemple: AB1 versus B2

- Le dernier gène ancestral commun (**point vert**) est celui qui précède immédiatement la duplication Dp1 (**flèche bleue**).
- Le chemin qui relie AB1 et B2 inclut par ailleurs un transfert d'espèce (**flèche rouge**).

→ **AB1 et B2 sont d'une part paralogues, et d'autre part xénologues**

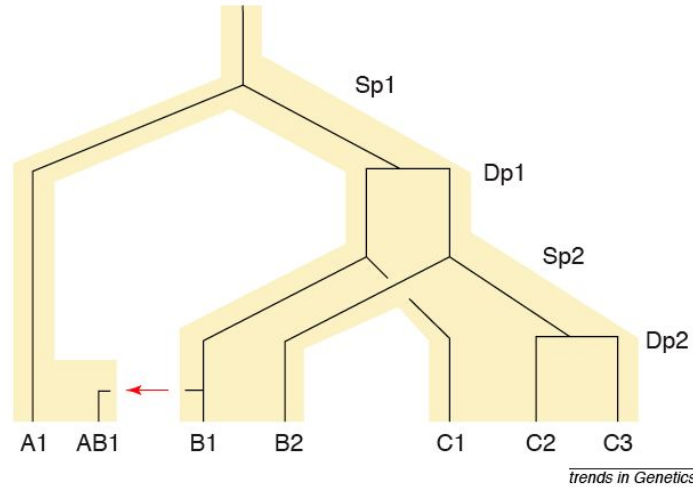
	A1	AB1	B1	B2	C1	C2	C3
A1							
AB1							
B1							
B2		P+X					
C1			O				
C2			P				
C3							



- 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.*

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).



	A1	AB1	B1	B2	C1	C2	C3
A1		O+X	O	O	O	O	O
AB1	O+X		X	P+X	O+X	P+X	P+X
B1	O	X		P	O	P	P
B2	O	P+X	P		P	O	O
C1	O	O+X	O	P		P	P
C2	O	P+X	P	O	P		P
C3	O	P+X	P	O	P	P	

Notes concernant la xénologie et les autres relations d'homologie.

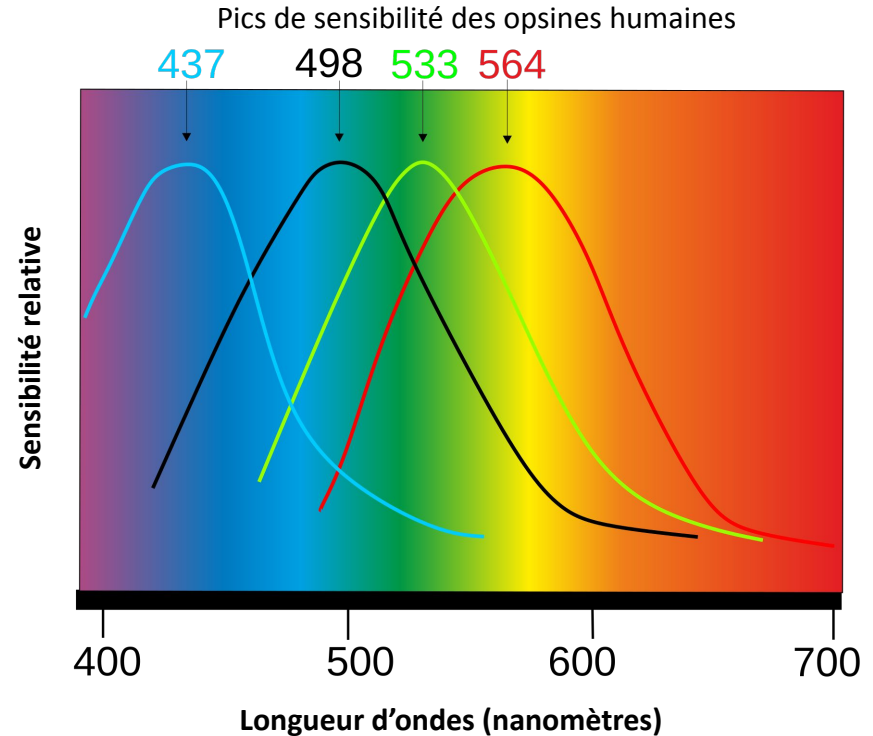
- La relation de xénologie, telles que définies par Fitch, s'applique dès qu'un transfert horizontal s'est produit depuis le dernier ancêtre commun entre deux molécules. Ceci s'applique donc à toutes les relations entre AB1 et une des autres séquences.
- Cette qualification de xénologue se place à un autre niveau que les statuts d'orthologie ou de paralogie, qui dépendent du type d'événement évolutif (spéciation ou duplication) qui a immédiatement succédé au dernier ancêtre commun.
- Nous pouvons donc considérer que les relations impliquant AB1 sont à la fois de xénologie et d'orthologie ou de paralogie selon le cas.

- 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.*

Matériel supplémentaire (anciennes diapo)
Ne fait pas partie de la matière d'examen

Le spectre visible

- Chaque cellule photoréceptrice perçoit une gamme spécifique de longueur d'ondes, avec un pic précis.
 - Cônes sensibles aux **ondes courtes** (short-wave sensitive: **SWS**): 420 nm (autour du **violet-bleu**)
 - Bâtonnets: 489 nm
 - Cônes sensibles aux **ondes moyennes** (medium-wave sensitive: **MWS**) : 534 nm (autour du **vert**)
 - Cônes sensibles aux **ondes longues** (long-wave sensitive: **LWS**) : **564 nm**
- Les cônes (cellules rétinienne sensibles aux couleurs) expriment chacun une opsin 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 opsines MWS et LWS
 - **faible différence entre leurs pics** (31 nm) par rapport à la différence entre opsines bleue et verte (96 nm).



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

Mammalian opsins - percent identity matrix generated by clustalx																																																					
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48					
	1W_Felis_catus	1W_Equus_caballus	1W_Phoca_vitulina	1W_Pagophilus_groenlandicus	1W_Canis_familiaris	1W_Bos_taurus	1W_Capra_hircus	1W_Sus_scrofa	1W_Odocoileus_virginianus	1W_Globicephala_melas	1W_Tursiops_truncatus	1W_Phocoenoides_phocoena	1W_Delphinus_delphis	1W_Mus_musculus	1W_Rattus_norvegicus	1W_Cavia_porcellus	1W_Oryctolagus_cuniculus	1W_Sciurus_carolinensis	1W_Homo_sapiens	1W_Homo_sapiens	1W_Callithrix_jacchus	1W_Tarsipes_rostratus	1W_Cercartetus_cuniculus	1W_Macropus_eugenii	1W_Myrmecobius_fasciatus	1W_Isodon_obesulus	1W_Didelphis_aurita	1W_Ornithorhynchus_anatinus	1W_Tachyglossus_acleatus	1W_Mouse	1W_Rattus_norvegicus	1W_Myotis_laniger	1W_Myotis_ricketti	1W_Bos_taurus	1W_Human	1W_Pan_paniscus	1W_Pan_troglodytes	1W_Daubentonius_madagascariensis	1W_Saimiri_bolivensis	1W_Macropus_eugenii	1W_Setonix_brachyurus	1W_Isodon_obesulus	1W_Tarsipes_rostratus	1W_Sminthopsis_cassicaudata	1W_Didelphis_aurita	1W_Thylamys_elegans	1W_Ornithorhynchus_anatinus	1W_Tachyglossus_acleatus					
1W_Felis_catus	100	97	92	92	95	93	93	93	95	90	89	90	88	89	87	91	90	90	91	91	91	91	84	84	86	85	84	86	85	41	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42	42				
1W_Equus_caballus	97	100	95	95	96	96	96	96	96	92	93	92	92	91	92	90	93	91	93	91	92	85	86	87	88	88	88	88	88	41	42	43	44	42	43	42	43	42	43	42	43	44	44	45	42	43	42	42	42	40			
1W_Phoca_vitulina	92	95	100	99	94	92	92	91	92	88	88	88	88	89	90	88	90	91	90	91	90	82	82	82	84	83	83	83	83	41	42	43	42	42	42	42	42	42	42	42	43	43	43	43	43	42	42	42	42	41			
1W_Pagophilus_groenlandicus	92	95	99	100	93	92	92	91	92	88	88	88	88	89	90	88	90	91	89	91	89	81	81	82	83	83	82	82	40	41	42	42	42	41	41	41	41	41	41	42	42	43	43	43	43	43	42	42	41	41			
1W_Canis_familiaris	95	96	94	93	100	95	95	94	95	90	89	90	90	89	90	90	93	92	90	89	90	83	83	83	85	84	83	84	84	41	43	43	42	42	42	42	42	42	42	42	42	42	43	43	43	44	42	43	41	41	41		
1W_Bos_taurus	93	96	92	92	95	100	99	97	97	93	94	93	93	88	89	87	90	89	90	89	91	81	82	81	83	83	82	83	82	41	43	42	43	42	42	42	42	42	42	42	42	42	43	43	43	44	41	42	41	41			
1W_Capra_hircus	93	96	92	92	95	100	97	97	97	93	94	93	93	88	89	87	90	89	90	89	91	81	82	81	83	83	82	82	41	43	42	43	42	42	42	42	42	42	42	42	42	42	43	43	43	44	41	42	41	41			
1W_Sus_scrofa	93	96	91	91	94	97	97	100	97	94	94	93	93	88	89	87	90	89	90	90	90	81	80	82	82	81	82	82	41	42	41	42	41	41	41	41	41	41	41	41	41	42	42	43	43	44	41	42	40	40			
1W_Odocoileus_virginianus	95	96	92	92	95	97	97	97	100	93	93	93	93	89	89	87	89	89	91	89	91	89	82	84	85	85	86	85	86	85	41	42	42	42	42	42	42	42	42	42	42	42	42	43	42	43	45	42	42	42	42		
1W_Globicephala_melas	90	92	88	88	90	93	93	94	93	100	99	98	99	86	87	86	88	88	88	88	89	79	80	79	81	80	81	81	80	41	41	40	42	41	41	41	41	41	41	41	41	41	42	42	43	43	44	41	42	40	39		
1W_Tursiops_truncatus	90	93	88	88	90	94	94	94	93	99	100	99	99	86	87	86	88	88	88	88	89	79	80	79	81	80	81	81	80	41	41	40	42	41	41	41	41	41	41	41	41	41	41	41	42	42	43	43	44	41	42	40	39
1W_Phocoenoides_phocoena	89	92	88	88	89	93	93	93	93	98	99	100	98	85	86	85	88	87	87	88	78	79	78	79	80	80	80	80	79	40	41	40	41	41	41	41	41	41	41	41	41	41	41	41	41	42	42	43	41	42	40	39	
1W_Delphinus_delphis	90	92	88	88	90	93	93	93	93	99	99	98	100	85	86	85	88	87	87	88	79	79	79	80	80	80	80	80	40	41	39	41	41	41	41	41	41	41	41	41	41	41	41	41	41	41	42	42	43	40	41	40	39
1W_Mus_musculus	88	91	89	89	89	88	88	88	88	86	86	85	85	100	98	91	90	88	88	89	82	83	82	84	83	83	83	82	41	43	42	43	43	43	43	43	43	43	43	44	44	44	44	42	43	42	43	42	41	41			
1W_Rattus_norvegicus	89	92	90	90	90	89	89	89	89	87	87	86	86	98	100	92	91	91	90	89	80	83	83	83	84	84	84	83	41	43	42	43	43	43	43	43	43	43	43	43	44	44	44	44	42	43	43	42	41				
1W_Cavia_porcellus	87	90	88	88	90	87	87	87	87	86	86	85	85	91	92	100	91	91	89	87	88	82	83	83	84	84	83	83	41	43	43	43	42	42	42	42	42	42	42	42	42	43	43	43	44	45	42	43	42	41			
1W_Oryctolagus_cuniculus	91	93	90	89	93	90	90	90	91	88	88	88	88	90	91	91	100	92	88	88	87	80	80	80	82	81	80	81	81	41	42	42	42	42	42	42	42	42	42	42	42	42	42	43	43	44	41	42	41	41			
1W_Sciurus_carolinensis	90	91	91	91	92	90	90	90	89	88	88	87	87	90	91	91	92	100	90	91	90	80	81	80	82	81	80	82	41	43	42	42	42	42	42	42	42	42	42	42	42	42	42	43	43	44	45	42	43	41	40		
1W_Homo_sapiens	90	93	90	90	89	89	89	89	89	88	88	87	87	88	90	89	88	90	100	96	96	82	82	83	84	84	84	83	42	43	43	44	44	43	43	43	43	43	43	43	44	44	44	44	44	45	43	44	42	41			
1W_Homo_sapiens	91	91	91	89	89	89	89	89	88	88	87	87	88	89	87	88	91	96	100	96	82	82	82	84	84	83	85	83	41	42	41	42	42	42	42	42	42	42	42	42	42	42	42	43	43	44	42	42	42	41			
1W_Callithrix_jacchus	91	92	90	89	90	91	90	92	89	89	88	88	89	90	88	87	90	96	96	100	83	83	83	85	85	85	86	84	41	43	42	43	42	42	42	42	42	42	42	42	42	42	42	43	43	43	45	42	42	41			
21W_Tarsipes_rostratus	83	85	82	81	83	81	81	80	84	80	79	78	79	82	83	82	80	80	82	82	83	100	98	96	95	96	96	90	92	42	43	43	42	42	42	42	42	42	42	42	42	42	42	43	43	44	43	42	42	41			
22W_Cercartetus_cuniculus	84	86	82	81	83	82	81	85	80	80	79	79	83	83	83	80	81	82	82	83	98	100	96	95	95	95	89	91	42	43	43	42	42	42	42	42	42	42	42	42	42	42	42	43	44	43	42	41					
23W_Macropus_eugenii	84	87	82	82	83	81	81	80	85	80	79	78	79	82	83	83	80	80	82	83	96	100	97	95	94	90	92	42	42	43	42	42	42	42	42	42	42	42	42	42	42	42	43	44	43	44	42	42	40				
24W_Myrmecobius_fasciatus	86	89	84	83	85	83	83	82	87	81	81	80	80	84	85	84	82	82	85	84	95	95	95	97	100	97	94	91	92	42	42	43	42	42	42	42	42	42	42	42	42	42	43	44	43	44	42	41	41				
25W_Isodon_obesulus	85	88	83	83	84	83	83	82	86	81	80	80	80	83	84	83	81	82	84	83	85	96	95	95	97	100	95	91	92	42	42	42	42	42	41	41	41	41	41	41	41	41	41	41	41	41	41	41	40	40			
27W_Didelphis_aurita	84	86	83	82	83	82	82	81	85	81	81	80	80	83	84	83	80	82	84	83	85	96	95	94	94	95	100	90	91	42	42	42	42	42	41	41	41	41	41	41	41	41	41	41	41	41	41	41	40	40			
28W_Ornithorhynchus_anatinus	86	89	83	82	84	83	82	86	81	81	80	8																																									

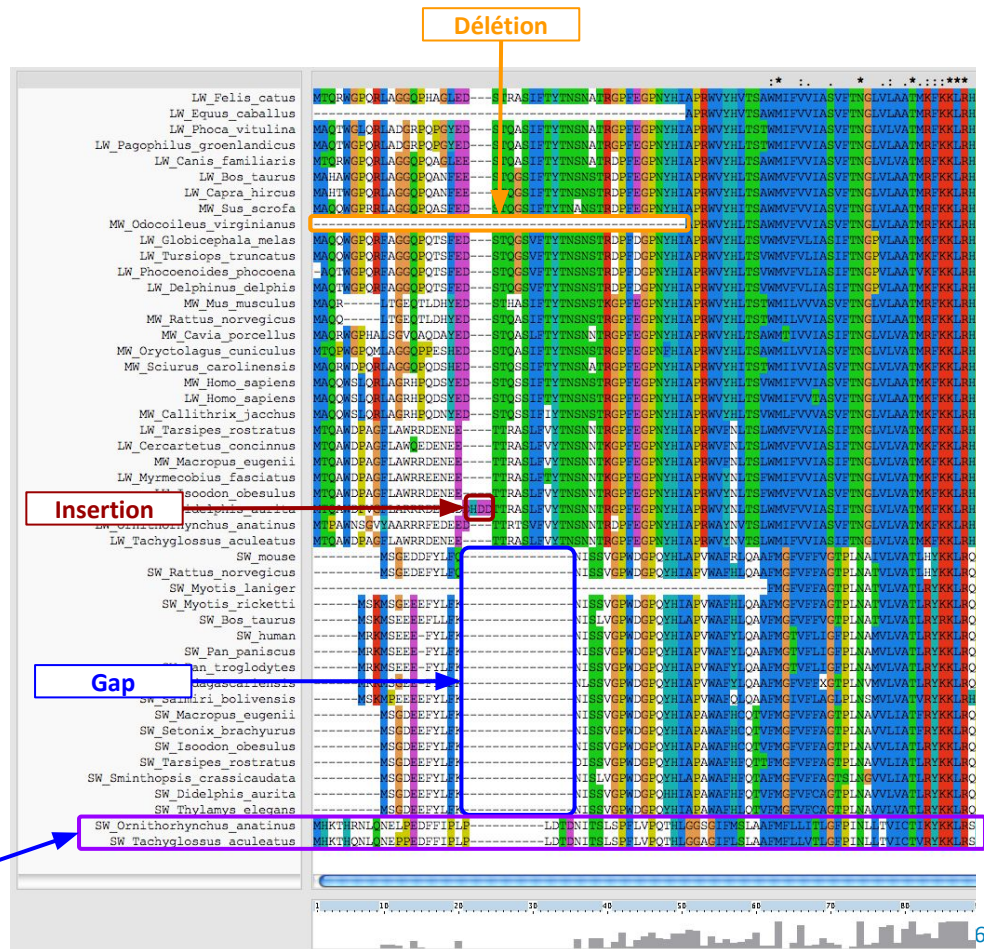
Alignement multiple des opsines de mammifère

- 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 délétions des insertions
 - Délétion: 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 externe** (ici Ornithorhynchus et Tachyglossus) permet d'enraciner l'arbre, et de lever certaines ambiguïtés.
- Cependant, l'inférence de l'événement évolutif, délétion ou insertion, qui est à l'origine d'un gap, nécessite généralement d'interpréter les alignements en les comparant à l'arbre phylogénétique inféré.



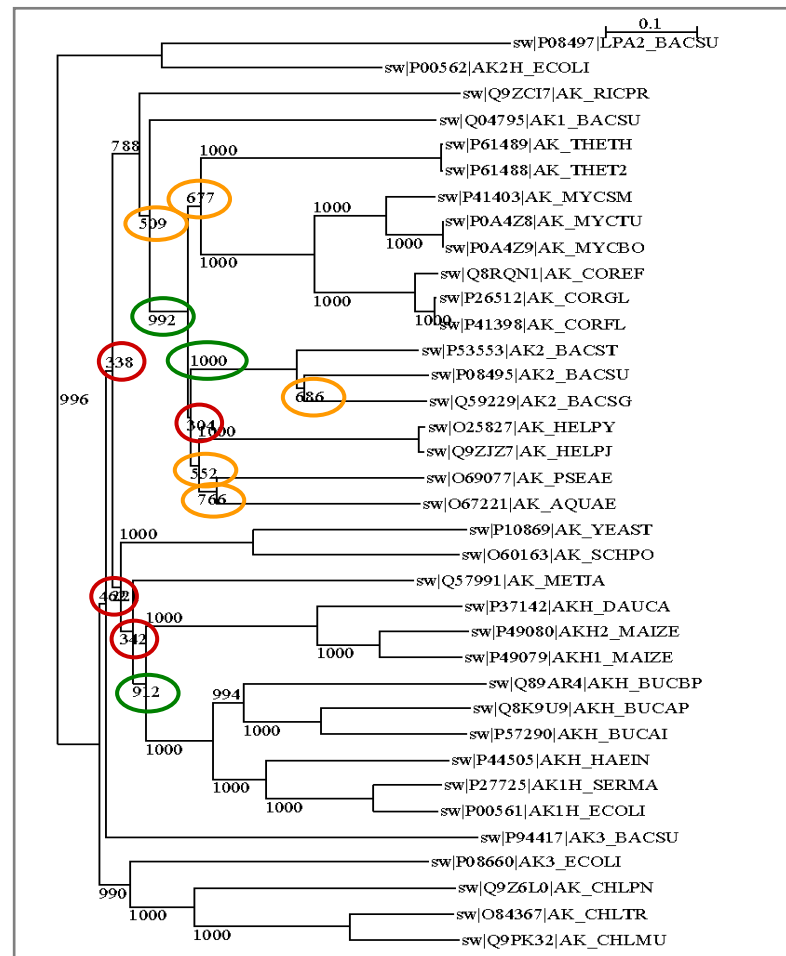
Bootstrapping

- 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**



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.

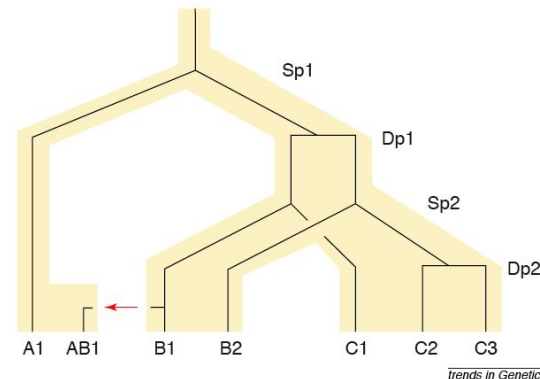
- **Analogie:** relation entre deux caractères qui se sont développés de façon convergente à partir d'ancêtres non-apparentés.

- **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 orthologous 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)

Retracer l'évolution de SARS-CoV-2 dans les génomes des coronavirus
Ne fait pas partie de la matière d'examen

La publication du génome de SARS-CoV-2

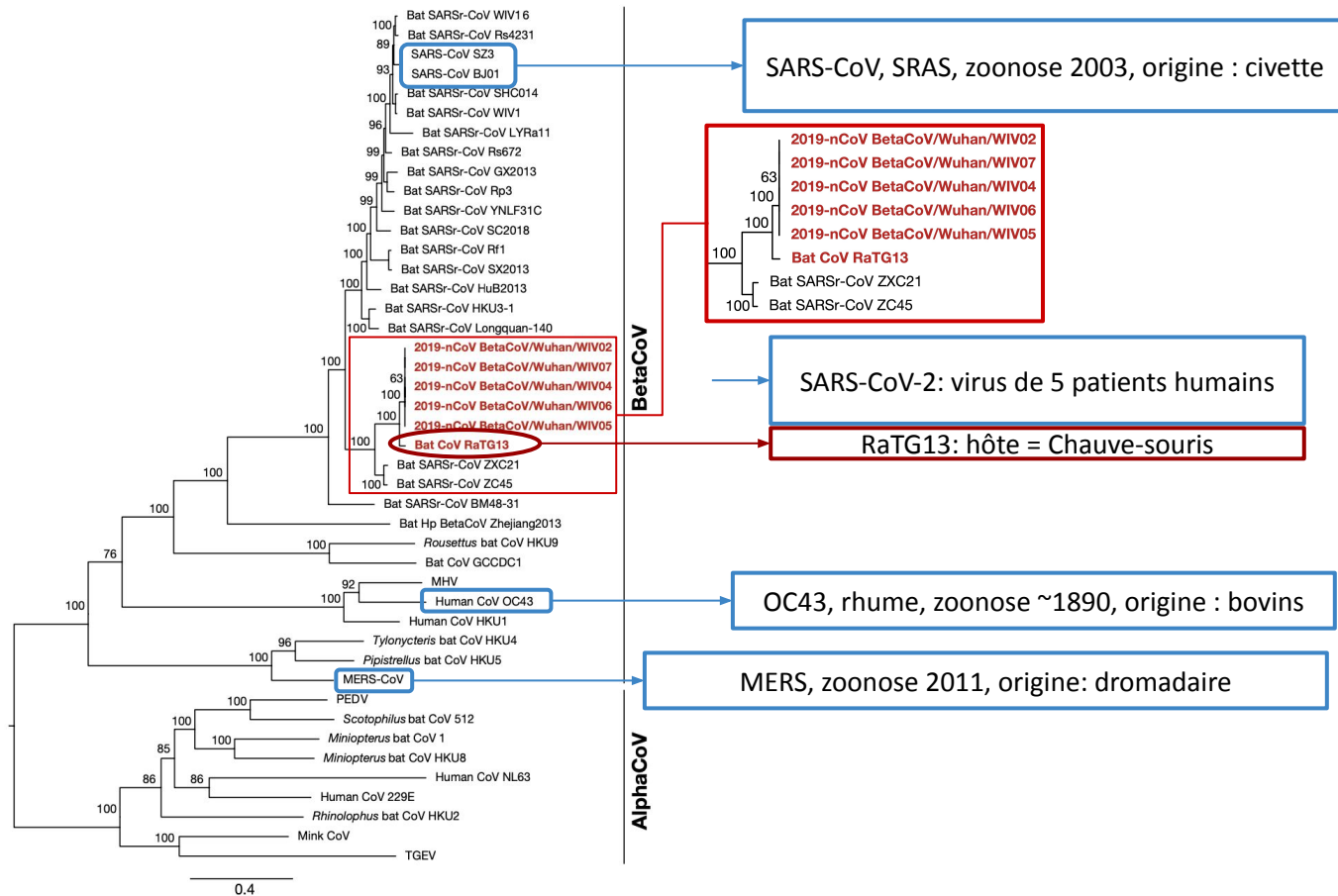
3 février 2020: publication du **génom**e complet de SARS-CoV-2
Recherche de virus similaires dans les bases de données de séquence. Les virus les plus proches sont des virus de chauves-souris (Bat CoV ZC45)

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>

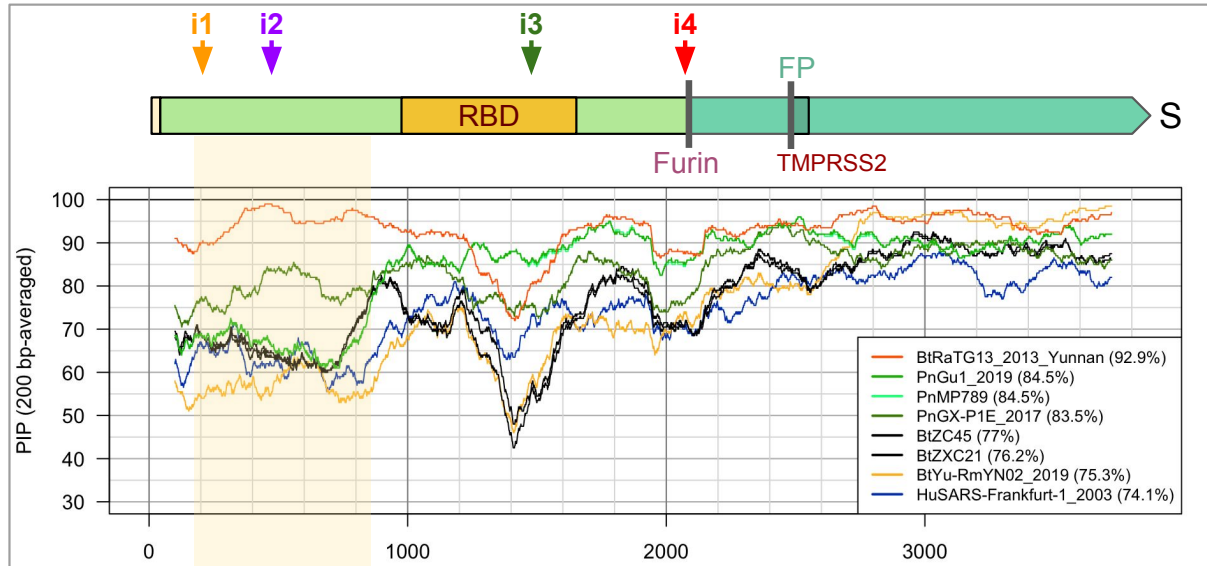
(devenue inaccessible depuis lors)

"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". "

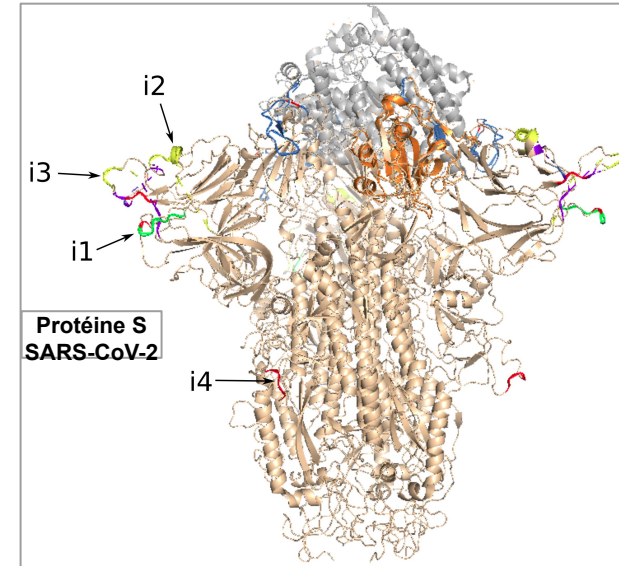
Quatre insertions dans le gène S de SARS-CoV-2

- 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 spicule



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	AATGGTACTAAGAGGTTTGATAACCTG	113	
Sbjct	967	AATGGTACTAAAAGGTTAGATAACACTG	994	

HIV-1 isolate patient B clone 16.3 from Netherlands envelope glycoprotein (env) gene, complete cds				
Sequence ID: HQ386166.1		Length: 2580	Number of Matches: 1	Range 1: 2493 to 2523
Score	Expect	Identities	Gaps	Strand
39.2 bits(42)	2.1	27/31(87%)	0/31(0%)	Plus/Minus
Query	351	CCTAAAAGTTCTTTGTAATACTGTATTATT	381	
Sbjct	2523	CCTAAAAGTTCTTTGTAATATTTCTATAATT	2493	

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 inconsistant 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.

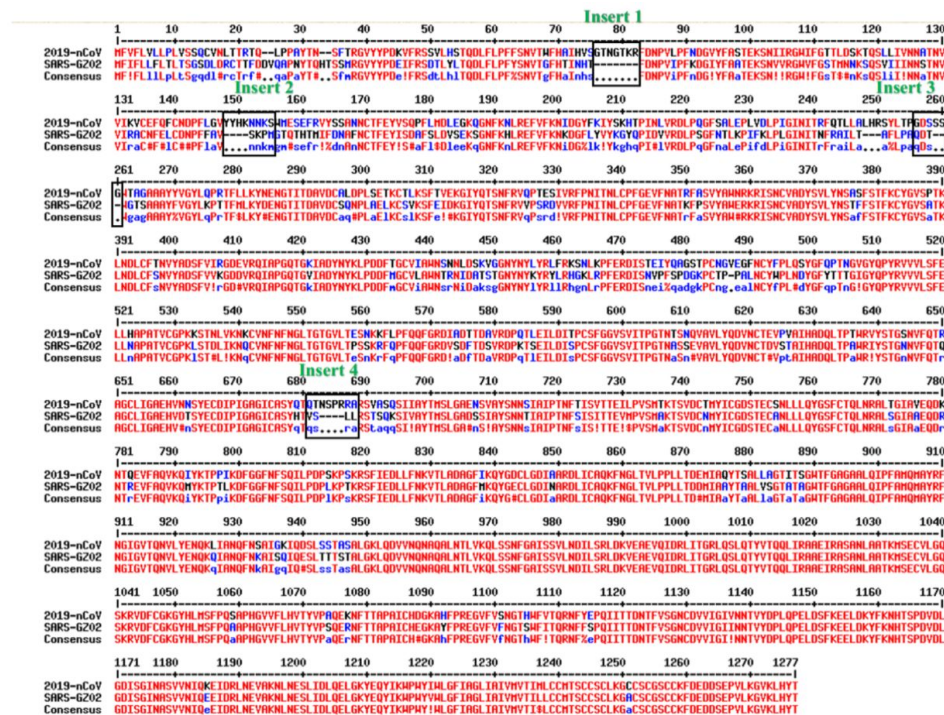
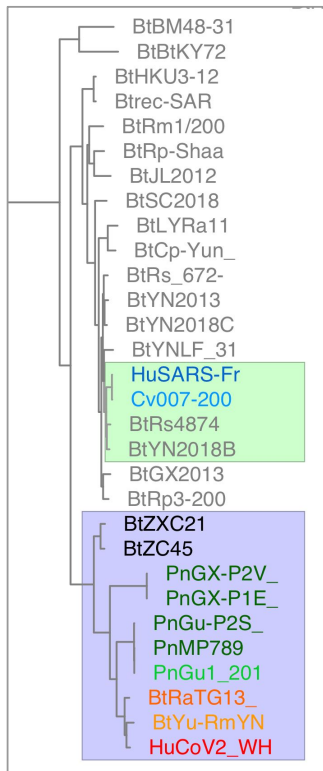


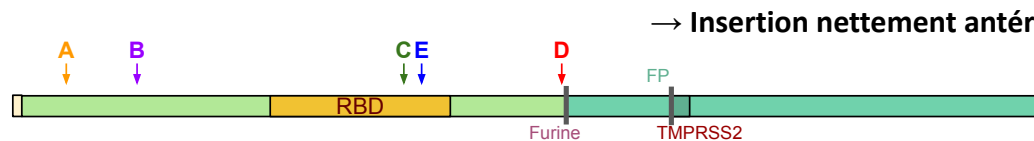
Figure 2: Multiple sequence alignment between spike proteins of 2019-nCoV and SARS. The sequences of spike proteins of 2019-nCoV (Wuhan-HU-1, Accession NC_045512) and of SARS CoV (GZ02, Accession AY390556) were aligned using MultiAlin software. The sites of difference are highlighted in boxes.

Insertion B : partagée entre plusieurs espèces de coronavirus

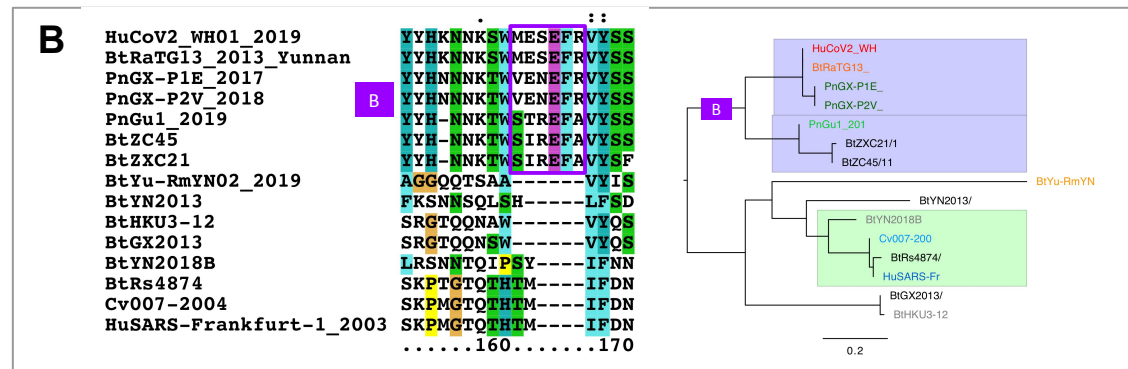
Arbre des génomes



- **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

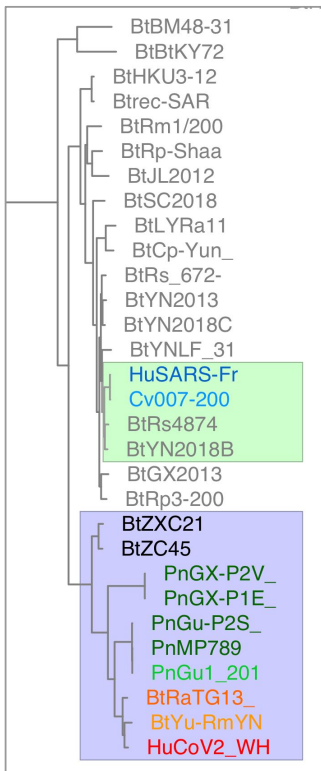


→ Insertion nettement antérieure à l'émergence de SARS-CoV-2

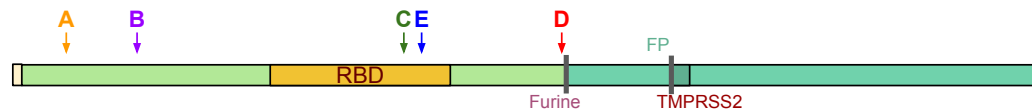
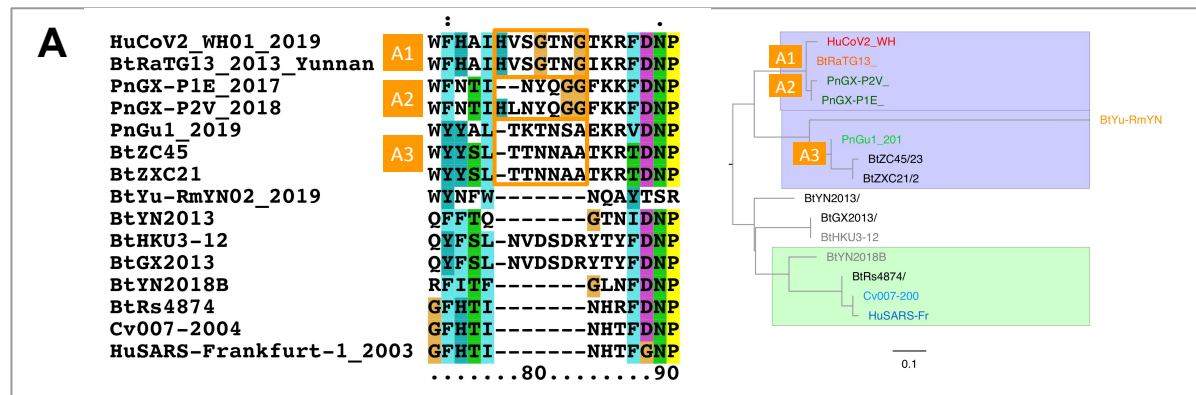


Insertion A : partagée entre SARS-CoV-2 et RaTG13

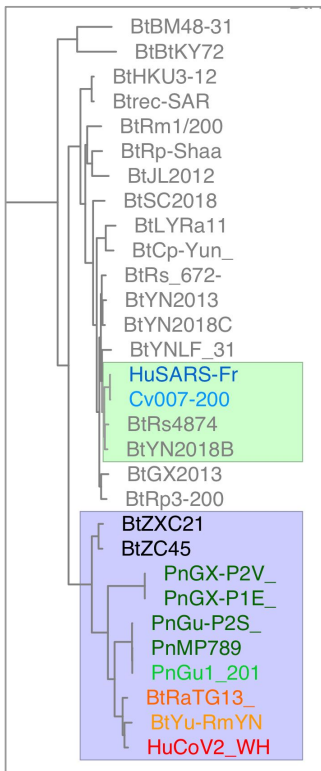
Arbre des génomes



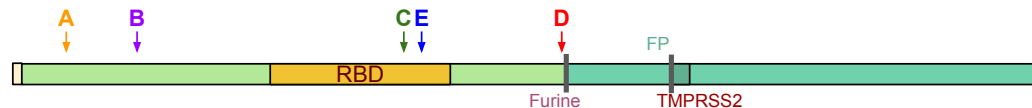
- **A1** : Insertion identique entre SARS-CoV-2 et RaTG13, malgré 40 à 70 ans de divergence
- **Insertion nettement antérieure à l'émergence de SARS-CoV-2**
- **A2**, **A3** : Insertions au même site, dans des sous-groupes séparés → événements indépendants



Arbre des génomes

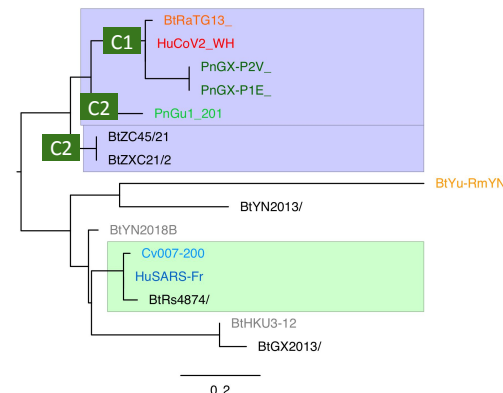


- **C1** : insertion partagée entre 4 des souches de virus apparentés, parmi les souches analysées
- Insertion nettement antérieure à l'émergence de SARS-CoV-2
- **C2** : insertion indépendante à la même position chez d'autres virus

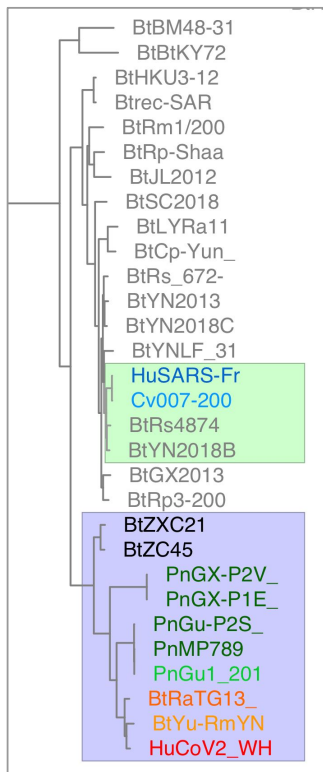


C

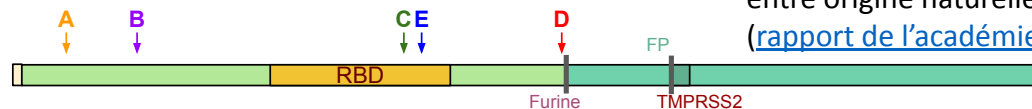
HuCoV2_WH01_2019	LLALHRSYLT	PGDSSSGWTA
BtRaTG13_2013_Yunnan	LLALHRSYLT	PGDSSSGWTA
PnGX-P1E_2017	LLALHRSYLT	PGKLESGWTT
PnGX-P2V_2018	LLALHRSYLT	PGKLESGWTT
PnGu1_2019	LLTIHRGDPMP	NNGWTV
BtZC45	LLTIHRGDPMP	NNGWTA
BtZXC21	LLTIHRGDPMS	NNGWTA
BtYu-RmYN02_2019	VLTF	RSNSQP
BtYN2013	FLAVYRVA	AGSISV
BtHKU3-12	VMAMFSQT	TSNFLP
BtGX2013	VMAMFSQS	TSNFLP
BtYN2018B	LLTAFPPN	PGYWGT
BtRs4874	ILTAFSPA	QDTWGT
Cv007-2004	ILTAFSPA	QDTWGT
HuSARS-Frankfurt-1_2003	ILTAFSPA	QDIWGT
	260.....	270..



Arbre des génomes



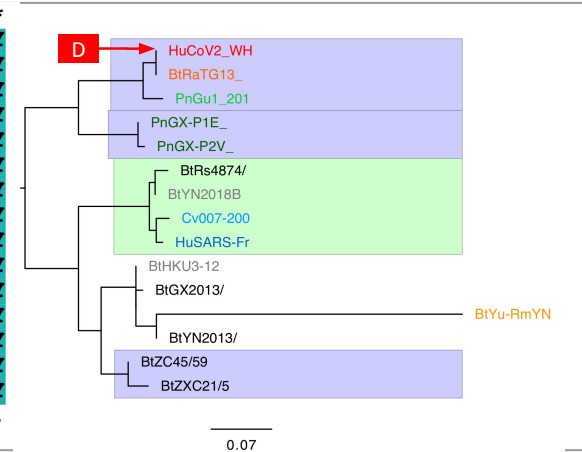
- **D** : insertion très courte, sans similarité significative avec HIV
- Insertion unique à SARS-CoV-2
- Crée un site furine, qui augmente la capacité de la protéine à se lier aux cellules humaines.
- Les données disponibles ne permettent pas de statuer entre origine naturelle ou artificielle de cette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



D

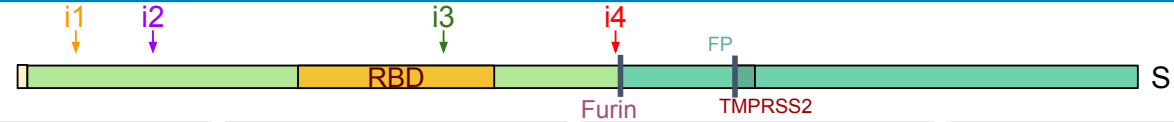
HuCoV2_WH01_2019	QTQTNS	PRRR	RSVASQ	SI	IAY
BtRaTG13_2013_Yunnan	QTQTNS	----	RSVASQ	SI	IAY
PnGX-P1E_2017	HSMSSL	----	RSVNQR	SI	IAY
PnGX-P2V_2018	HSMSSF	----	RSVNQR	SI	IAY
PnGu1_2019	QTQTNS	----	RSVSSQ	AI	IAY
BtZC45	HTASIL	----	RSTGOK	AI	IAY
BtZXC21	HTASIL	----	RSTGOK	AI	IAY
BtYu-RmYN02_2019	NSPAAR	----	VGTNSI	AI	IAY
BtYN2013	HTASTL	----	RSIGOK	SI	IAY
BtHKU3-12	HTASVL	----	RSTGOK	SI	IAY
BtGX2013	HTASVL	----	RSTGOK	SI	IAY
BtYN2018B	HTVSSL	----	RSTSOK	SI	IAY
BtRs4874	HTVSSL	----	RSTSOK	SI	IAY
Cv007-2004	HTVSSL	----	RSTSOK	SI	IAY
HuSARS-Frankfurt-1_2003	HTVSL	----	RSTSOK	SI	IAY

.690.....700.....

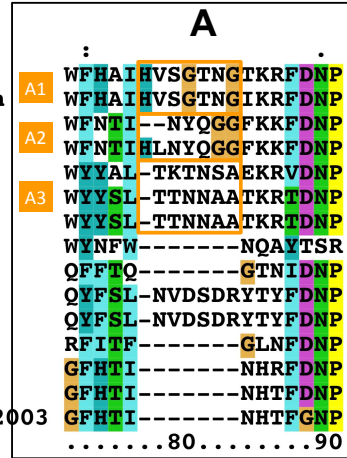
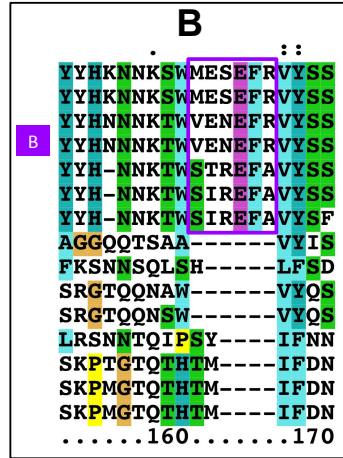
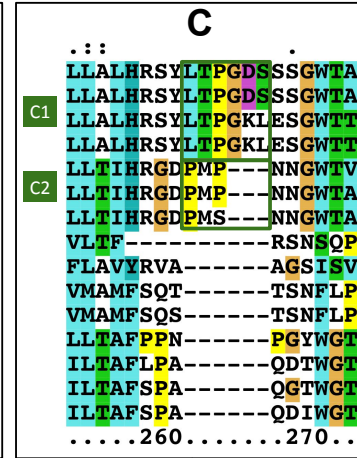
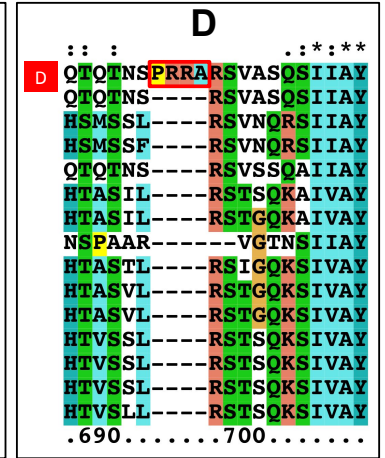
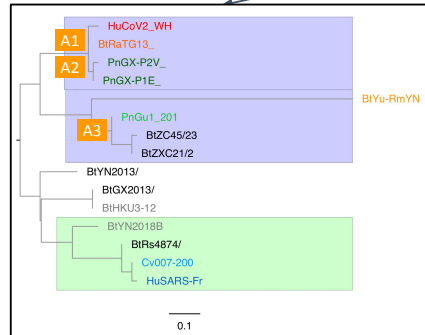
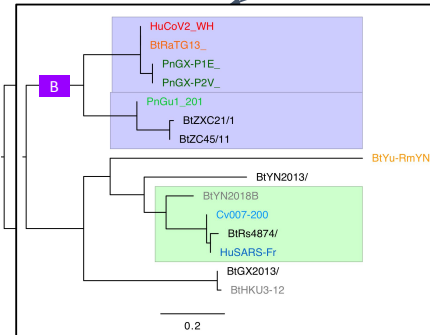
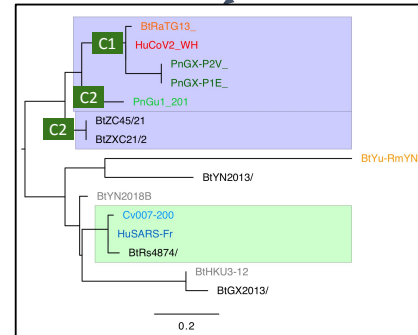
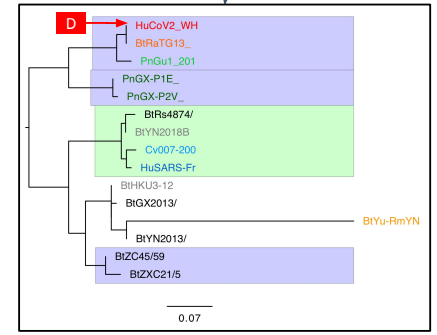


- Sallard, E., Halloy, J., Casane, D., van Helden, J., Decroly, E. Retrouver les origines du SARS-COV-2 dans les phylogénies de coronavirus. Médecine/Sciences 36, Sept-Août 2020. <https://doi.org/10.1051/medsci/2020123>
- [De l'Origine du SARS-CoV-2 aux risques de zoonoses et de manipulations dangereuses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

Alignements multiples et arbres phylogénétiques des insertions de la protéine S



HuCoV2_WH01_2019
BtRaTG13_2013_Yunnan
PnGX-P1E_2017
PnGX-P2V_2018
PnGu1_2019
BtZC45
BtZXC21
BtYu-RmYN02_2019
BtYN2013
BtHKU3-12
BtGX2013
BtYN2018B
BtRs4874
Cv007-2004
HuSARS-Frankfurt-1_2003



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.

- 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.