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

Introduction à la bioinformatique (UE SSV3U15)
Version allégée

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

Contenu de ce chapitre

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)

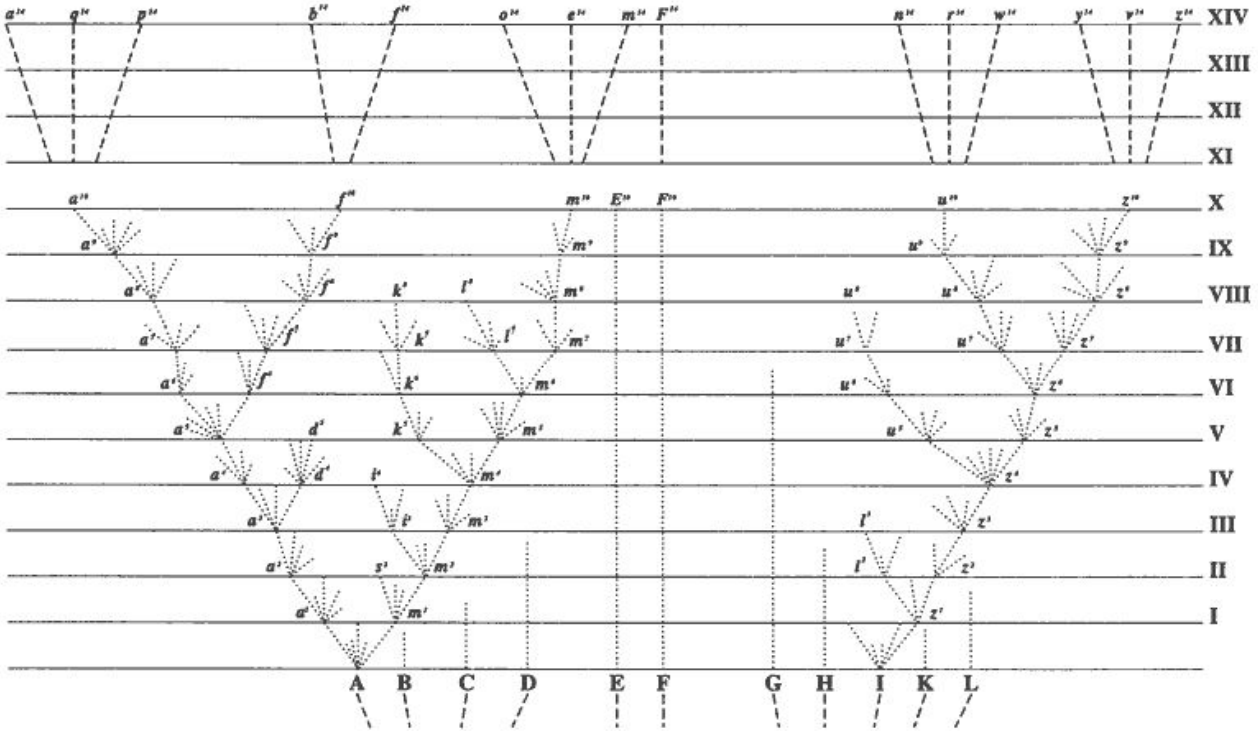
9. Retracer l'évolution de SARS-CoV-2 dans les génomes des coronavirus

Représentations arborescentes de l’évolution

La divergence des caractères

La seule figure de l'Origine des Espèces (C.Darwin, 1859)

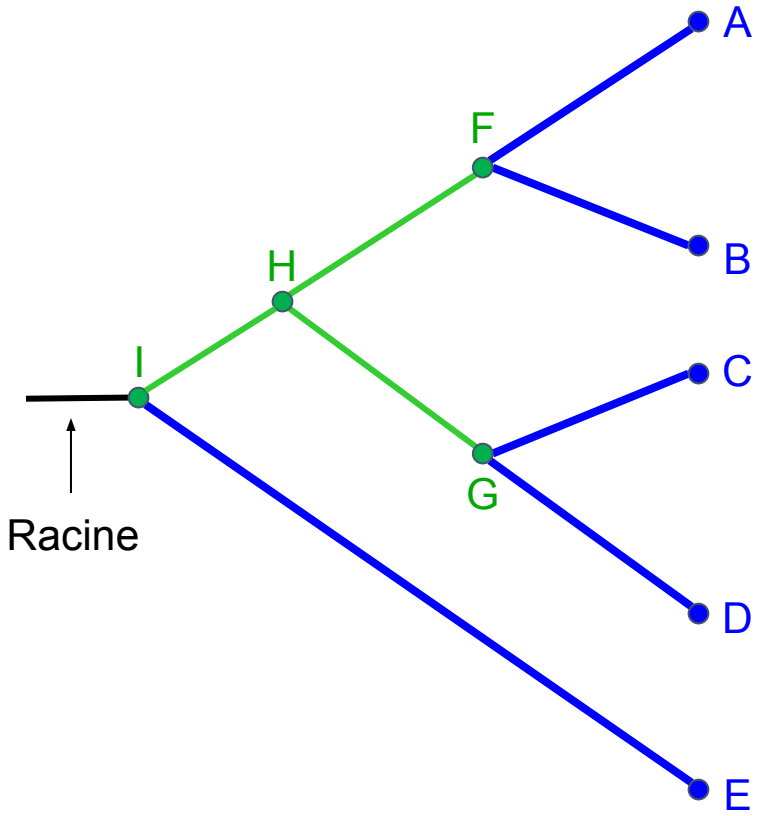
- Arbre de la vie
- Synchrone : hauteur = temps
- Evolution graduelle
- Branches abortives = espèces éteintes



Source: Darwin, C. (1859). L'origine des espèces.

Unités taxonomiques **opérationnelles** (OTU) et **hypothétiques** (HTU)

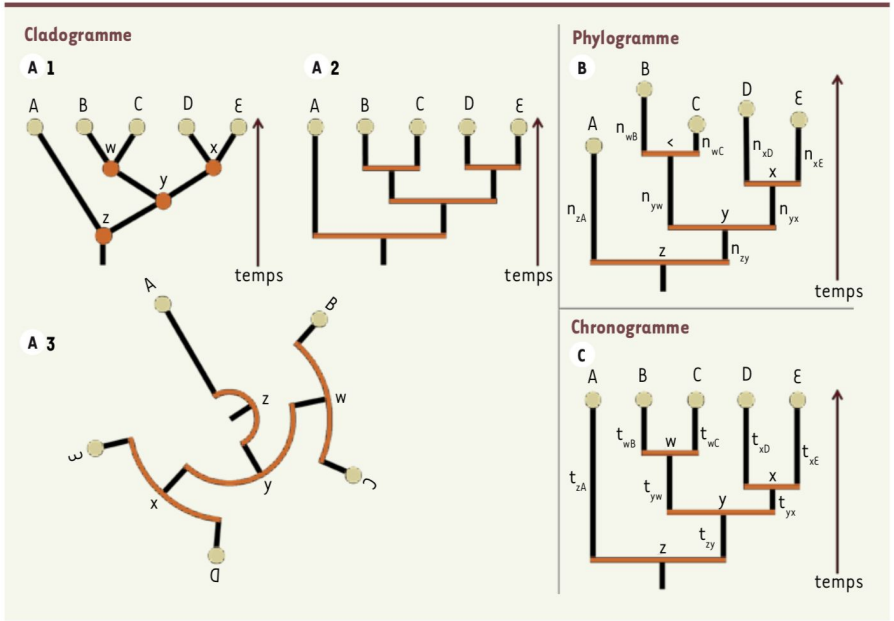
- **Topologie** = ensemble des branchements de l'arbre.
- **Noeuds = unités taxonomiques**
  - **OTU** = **Unités Taxonomiques Opérationnelles**
    - = Noeuds externes = feuille
    - espèces actuelles ou éteintes pour lesquelles on dispose de données
  - **HTU** = **Unités taxonomiques Hypothétiques** = noeuds internes
    - on ne dispose pas de données
    - correspondent aux dernier ancêtre commun entre plusieurs OTU.
- Branches = relations de parenté(ancêtre/descendants) entre unités taxinomiques
  - Branches internes
  - Branches externes



Source: Emese Meglézc

Représentations arborescentes des histoires évolutives

- **Cladogramme**
  - Représente uniquement les branchements successifs
  - Longueurs des branches non-informatives
- **Phylogramme**
  - Longueurs des branches: **distance évolutive** (différences génétiques ou morphologiques)
  - Certaines branches peuvent évoluer plus rapidement que d'autres → feuilles de l'arbre (OTU) pas forcément alignées
- **Chronogramme**
  - Longueur des branches: **temps de divergence**.
  - Si OTU contemporaines, les feuilles sont alignées

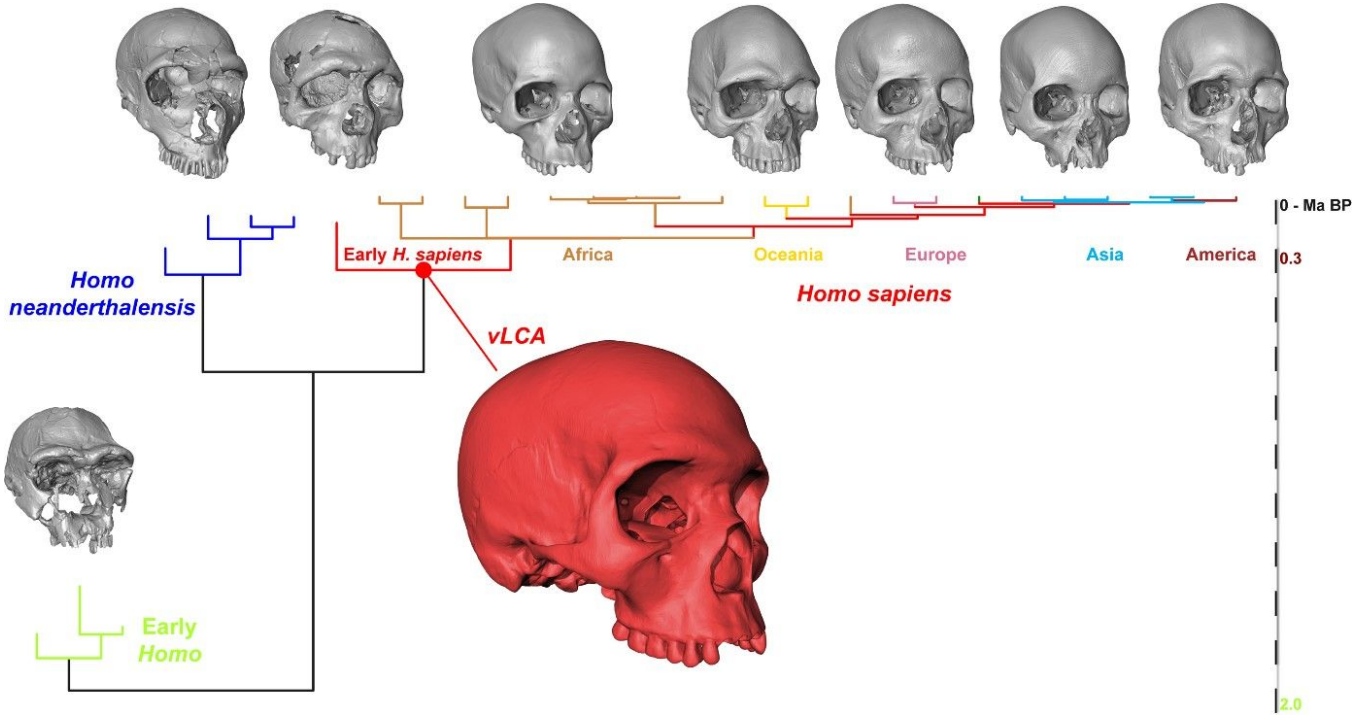


Casane, D. & Laurenti, P. Penser la biologie dans un cadre phylogénétique: L'exemple de l' évolution des vertébrés. Med Sci (Paris) 28, 1121–1127 (2012). doi.org/10.1051/medsci/20122812024

Arbre phylogénétique des espèces fossiles et actuelles d'Hominiens

- Axe vertical : temps (chronogramme)
- Crânes gris : OTU, pour lesquelles on dispose de spécimens.
- Crâne rouge
  - **reconstruction** du crâne de l'ancêtre commun de tous les *Homo sapiens*, inférée à partir de 29 spécimens actuels et fossiles.
  - → **Unité taxonomique hypothétique (HTU)**

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

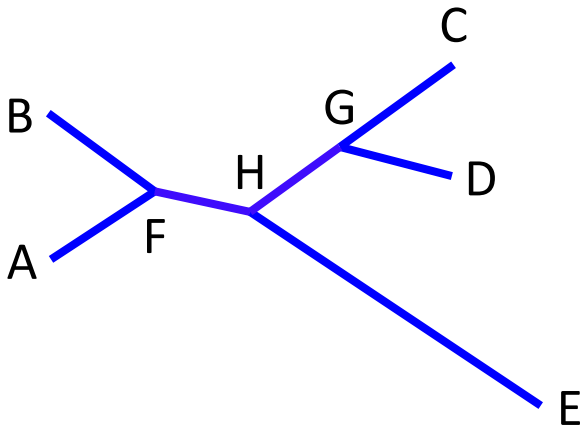


Source de la figure : Aurélien Mounier (CNRS/MNHN) et Véronique Etienne (CNRS communication)
https://www.cnrs.fr/en/press/what-noggin-modern-humans-ancestor-would-have-looked

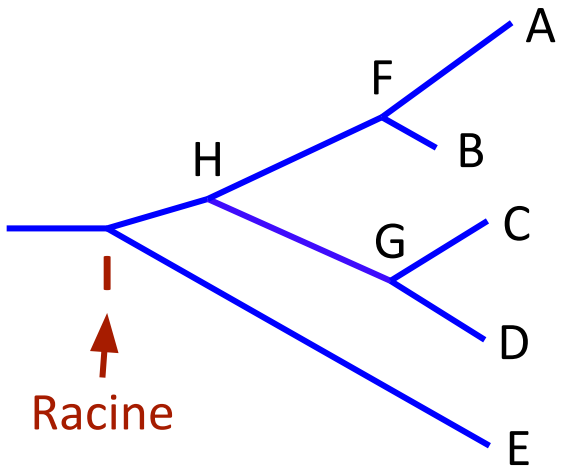
Arbres enracinés ou non enracinés

- **Arbre non-enraciné**
  - pas de direction temporelle
  - ⟶ indique les **distances, mais pas les relations de parenté** entre les noeuds.
- **Arbre racinés**
  - La **racine** symbolise le **dernier ancêtre commun** (l'ancêtre commun plus récent) de toutes les OTU.

Arbre non-enraciné

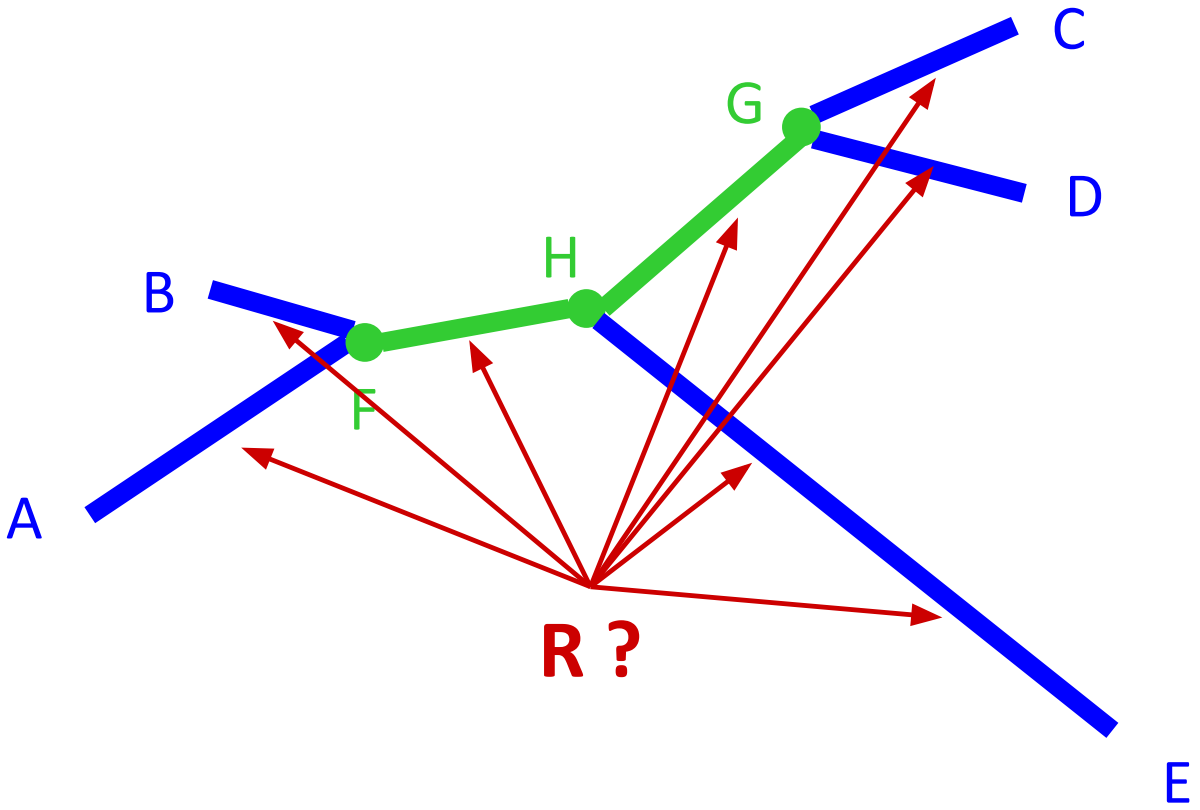


Arbre enraciné



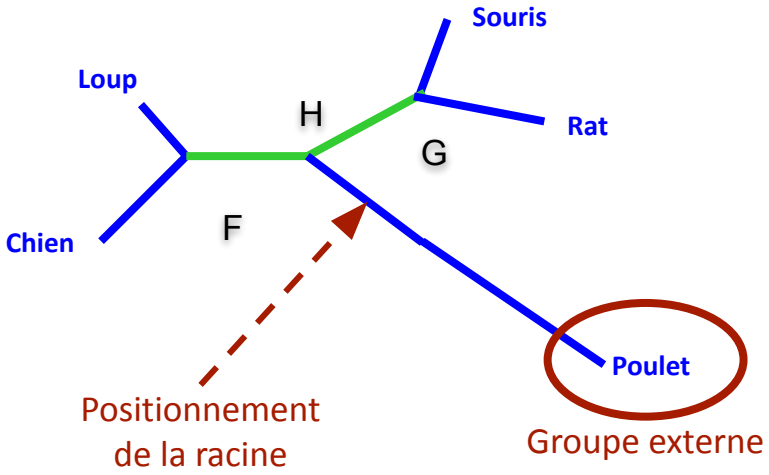
Comment enraciner un arbre phylogénétique ?

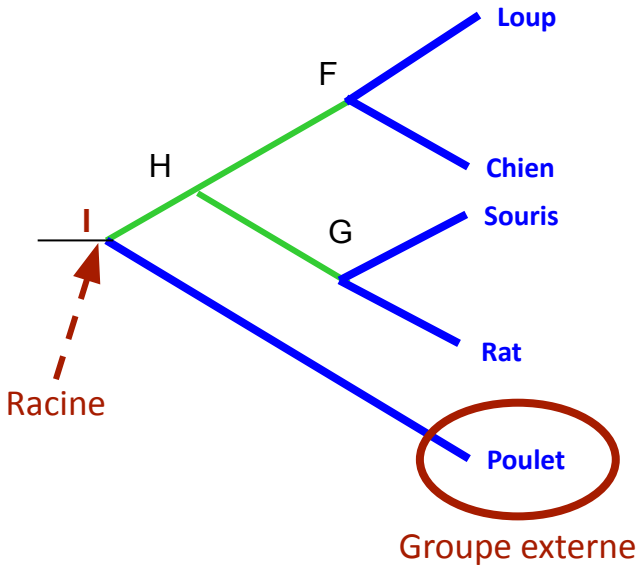
- A priori, la racine pourrait se situer sur à n'importe quelle position de n'importe quelle branche de l'arbre.
- Cependant, une seule de ces positions correspond à la véritable histoire évolutive.



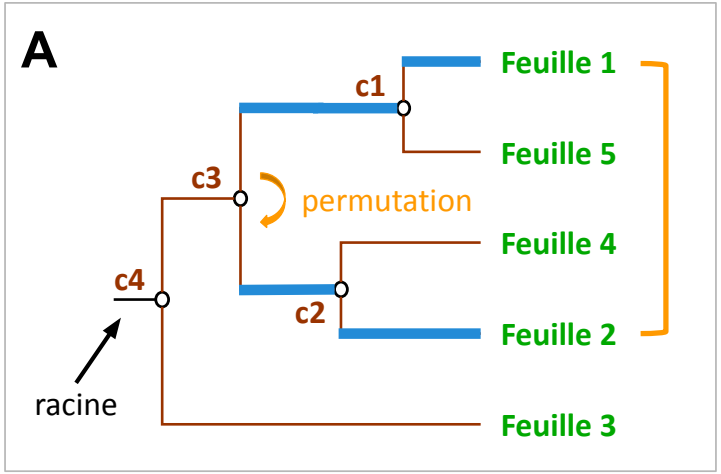
Approches pour enraciner un arbre phylogénétique

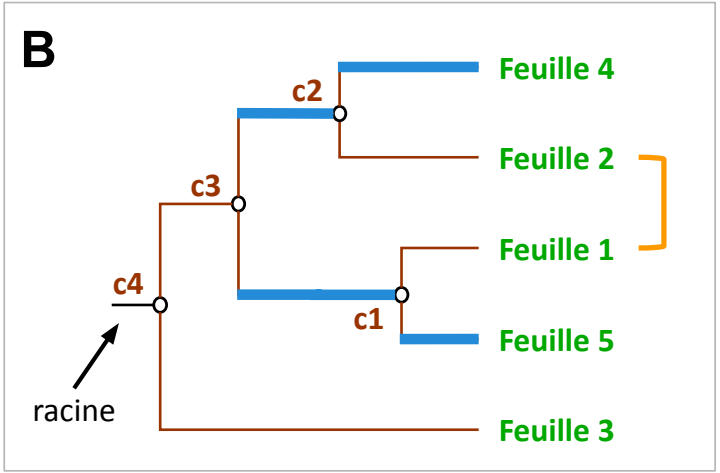
- **Enracinement au poids moyen**
  - Aucune garantie de pertinence
- Inclusion d'un **groupe externe** (**outgroup**) dans les données
  - Enracinement sur la branche qui sépare ce groupe de tous les autres





Isomorphisme des arbres phylogénétiques





- L'inversion verticale des branches d'une bifurcation ne modifie pas l'information.
  → Arbres ***isomorphiques***, équivalents
- **distances verticales** [crochets] ne correspondent **pas** à la distance évolutive
- distance évolutive = **somme des longueurs des branches horizontales**

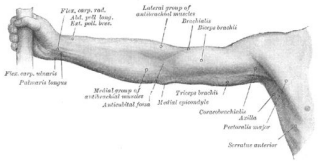
Homologie et analogie

Homologie

- **Homologie :** ressemblance entre des caractères (phénotypiques ou génétiques) qui s'explique par une **origine ancestrale commune**.
- Exemple : membres antérieurs des vertébrés terrestres
- **Caractère ancestral**, présent chez leur dernier ancêtre commun.
- **Évolution divergente :** accumulation de modifications différentes dans les lignées depuis leur séparation

Homologie

Humain (mammifère)



Chimpanzé (mammifère)

Pigeon (oiseau)

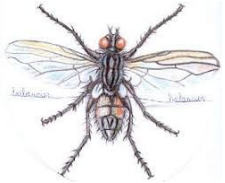
Analogie

- **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
  - Similarités de structure
- **Evolution convergente** : pression évolutive ⟶ modifications similaires
- **Caractères dérivés** : absents de l'ancêtre commun, apparus indépendamment dans chacune des deux lignées après leur séparation.

Analogie

Pigeon (oiseau)

Mouche (insecte)



Homologie et analogie : un exemple non trivial

- **Membres antérieurs** des chauves-souris sont homologues à ceux des oiseaux.
  - **Hérités du dernier ancêtre commun**
  - Caractère ancestral
- Cependant, ***en tant qu'organes de vol***, les **ailes** des chauves-souris sont **analogues** aux ailes des oiseaux
  - apparus indépendamment dans les deux lignées.
  - ***Caractères dérivés*** : absents chez le dernier ancêtre commun

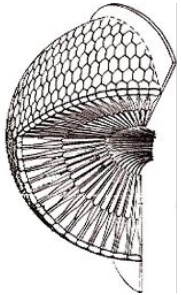
Pigeon (oiseau)

Chauve-souris (mammifère)

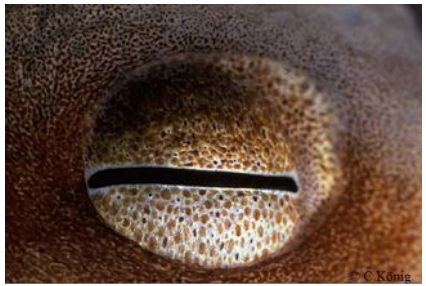
Structures analogues et convergence évolutive

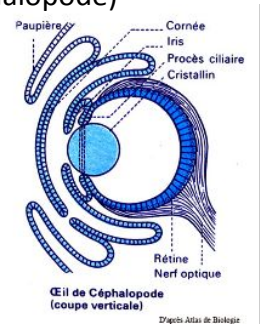
- Structures très diverses d'yeux
- L'oeil de pieuvre présente de fortes similarités de structures avec l'oeil humain
- Cependant, l'oeil de pieuvre et l'oeil humain résultent de **voies évolutives indépendantes**.
  - → ressemblance par **convergence évolutive** plutôt qu'origine commune.
  - Cas spectaculaire de **ressemblance par analogie**.
- Quelques différences notoires
  - **Similarités:** oeil sphérique, cornée, iris, cristallin, …
  - **Différences:** orientation des cellules rétiniennes: axones partent vers l'intérieur chez les vertébrés, vers l'extérieur chez les céphalopodes

Oeil de drosophile (insecte)

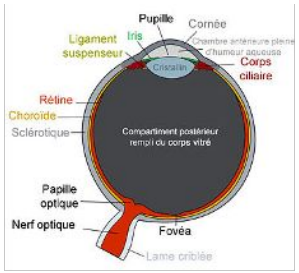


Oeil de pieuvre (mollusque céphalopode)

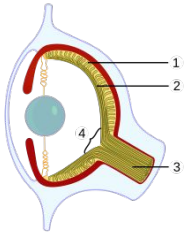




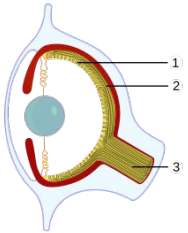
Oeil humain (mammifère)



Vertébré



Céphalopode



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

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

Evolution graduelle de l'oeil des mollusques

Yeux de différentes espèces de mollusques

→ grande diversité de structure et de complexité

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

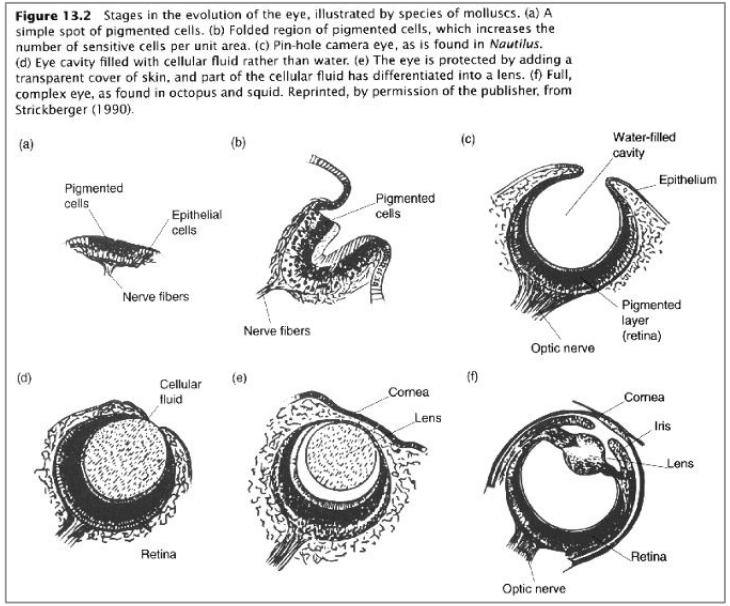
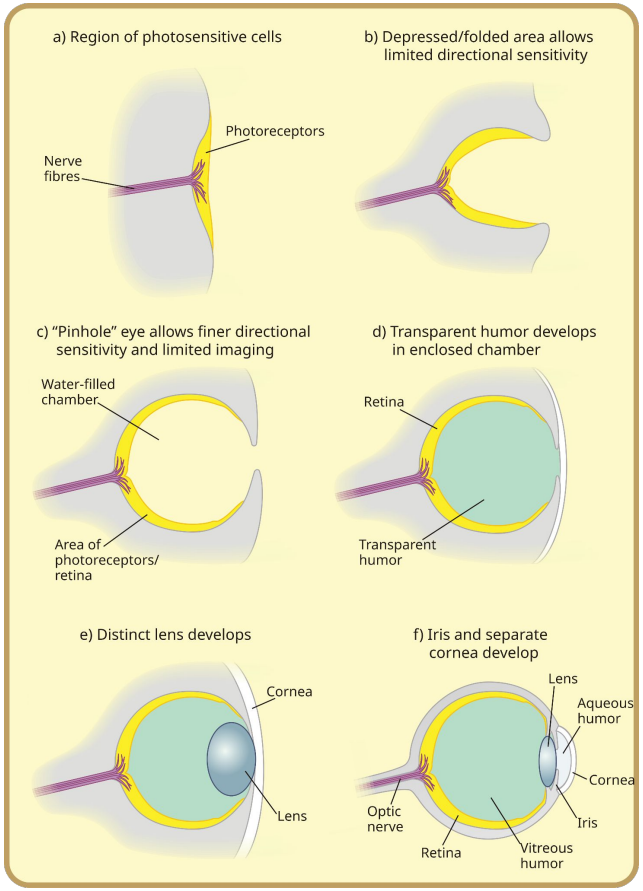


Figure 13.2 Stages in the evolution of the eye, illustrated by species of molluscs. (a) A simple spot of pigmented cells. (b) Folded region of pigmented cells, which increases the number of sensitive cells per unit area. (c) Pin-hole camera eye, as is found in *Nautilus*. (d) Eye cavity filled with cellular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



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

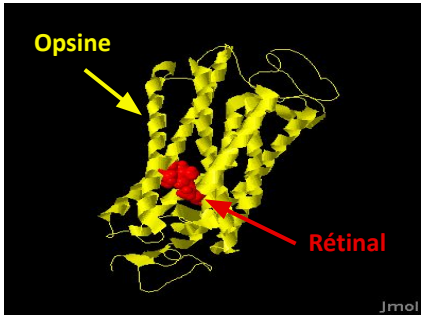
Les duplications à l’origine de l’innovation

rétinol

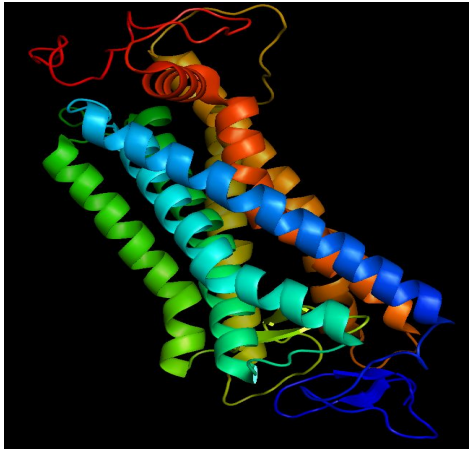
Structure d'une opsine

Modèle tridimensionnel du pigment des cônes bleus

(Structure PDB 1kpn affichée avec JMol)



(Structure PDB 1kpn affichée avec MacPyMol)



Perception de la lumière

- 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

https://www.rcsb.org/structure/removed/1KPN

La vision trichromatique des primates de l'ancien monde (y compris l'humain)

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** :

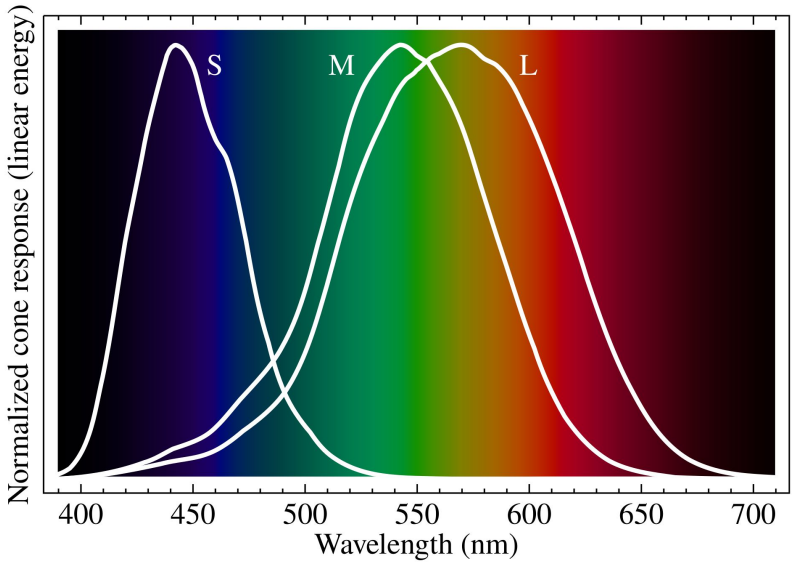
  Distance entre pics = 564 - 533 = 31 nm

  → **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

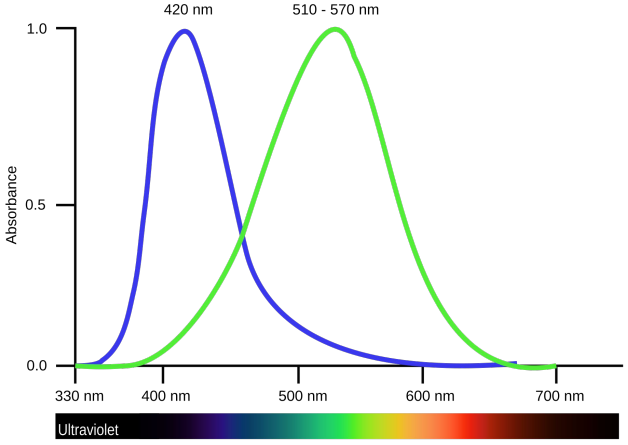


https://fr.wikipedia.org/wiki/C%C3%B4ne_(photor%C3%A9cepteur)
https://fr.wikipedia.org/wiki/C%C3%B4ne_(photor%C3%A9cepteur)#/media/Fichier:Spectre_absorption_des_cones.svg
https://fr.wikipedia.org/wiki/Vision_des_couleurs#/media/Fichier:Cone-fundamentals-with-srgb-spectrum.svg

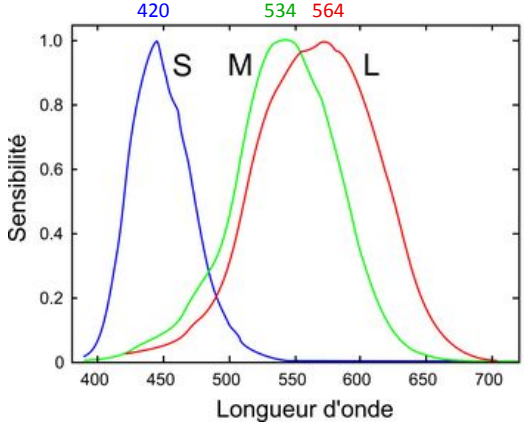
La vision dichromatique des autres mammifères

- 2 opsines
  - Ondes courtes (SWS), pic dans le bleu
  - Ondes moyennes-longues (MWS-LWS) : pic varie entre vert et rouge selon espèces
- Note : certaines formes de daltonisme humain reviennent à une vision dichromatique

Vision dichromatique des autres mammifères



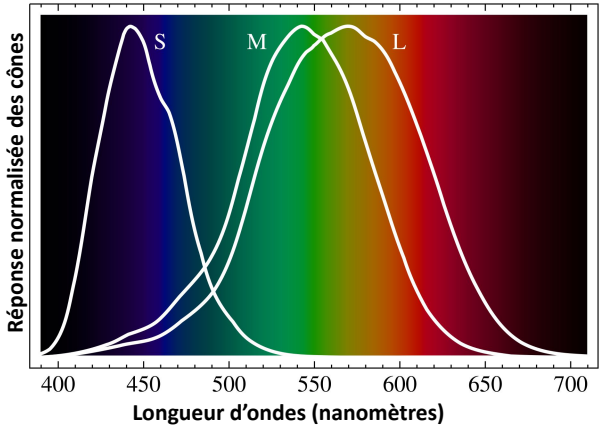
Vision trichromatique des primates de l'ancien monde

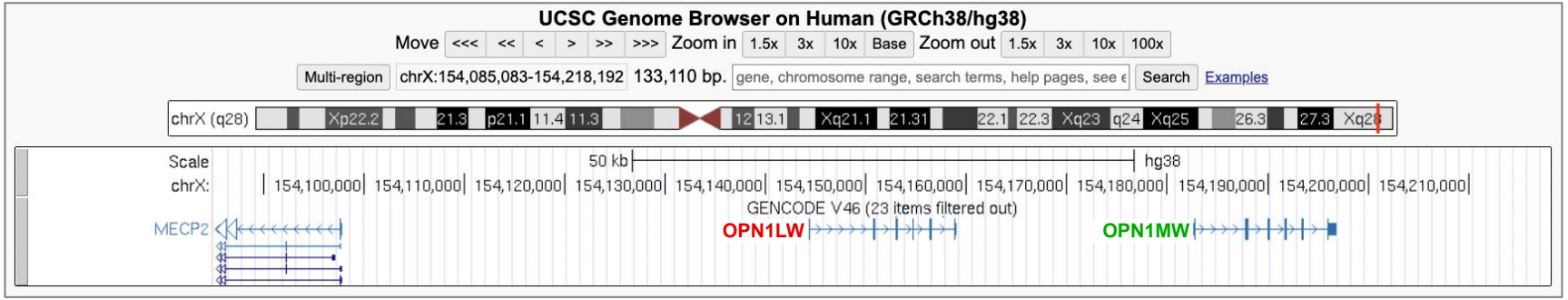


Origine de la vision trichromatique des primates de l'ancien monde

- Chromosome X: duplication du gène de l'opsine rouge-verte ⟶ 2 gènes en tandem
  - **OPN1LW** (Long Waves) ⟶ opsine **rouge**
  - **OPN1MW** (Medium Waves) ⟶ opsine **verte**
- Chromosome 7 humain : OPN1SW (Short Waves) ⟶ opsine **bleue**

Spectre de sensibilité des opsines humaines





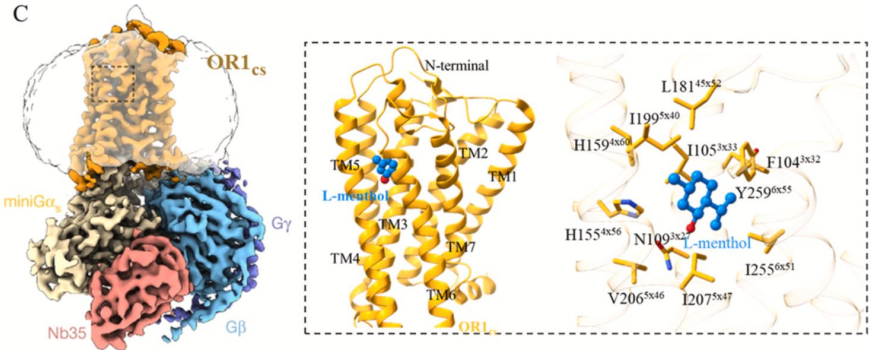
Une grande famille … de gènes

Récepteurs olfactifs

- La plus grande famille de gènes chez les métazoaires
  - Souris : ~800 gènes
  - Humain : ~ 400 gènes
- Expansion résulte de fréquentes duplications intra-chromosomiques.

Processus évolutif : duplication - divergence

Duplication
→ mutations
→ divergence de séquences
→ différences fonctionnelles (spécificité olfactive)



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

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 G, 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
- 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

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

Inférence d'un arbre phylogénétique à partir d'une famille de séquences homologues

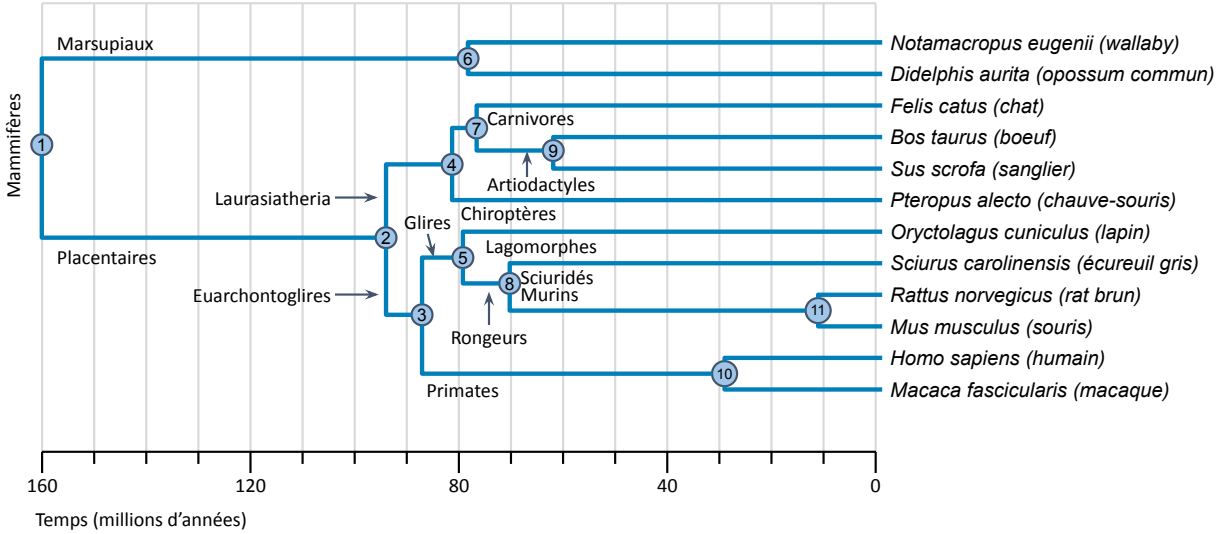
Phylogénie moléculaire : inférer l'histoire évolutive des organismes en s'appuyant sur l'analyse des séquences de macromolécules – acides nucléiques ou protéines.

Etapes

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

Collecte des séquences

- **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

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

Observations

- 2 groupes bien distincts (carrés orange - rouge)
- SW regroupées, séparées des MW+LW
- LW et MW mélangées

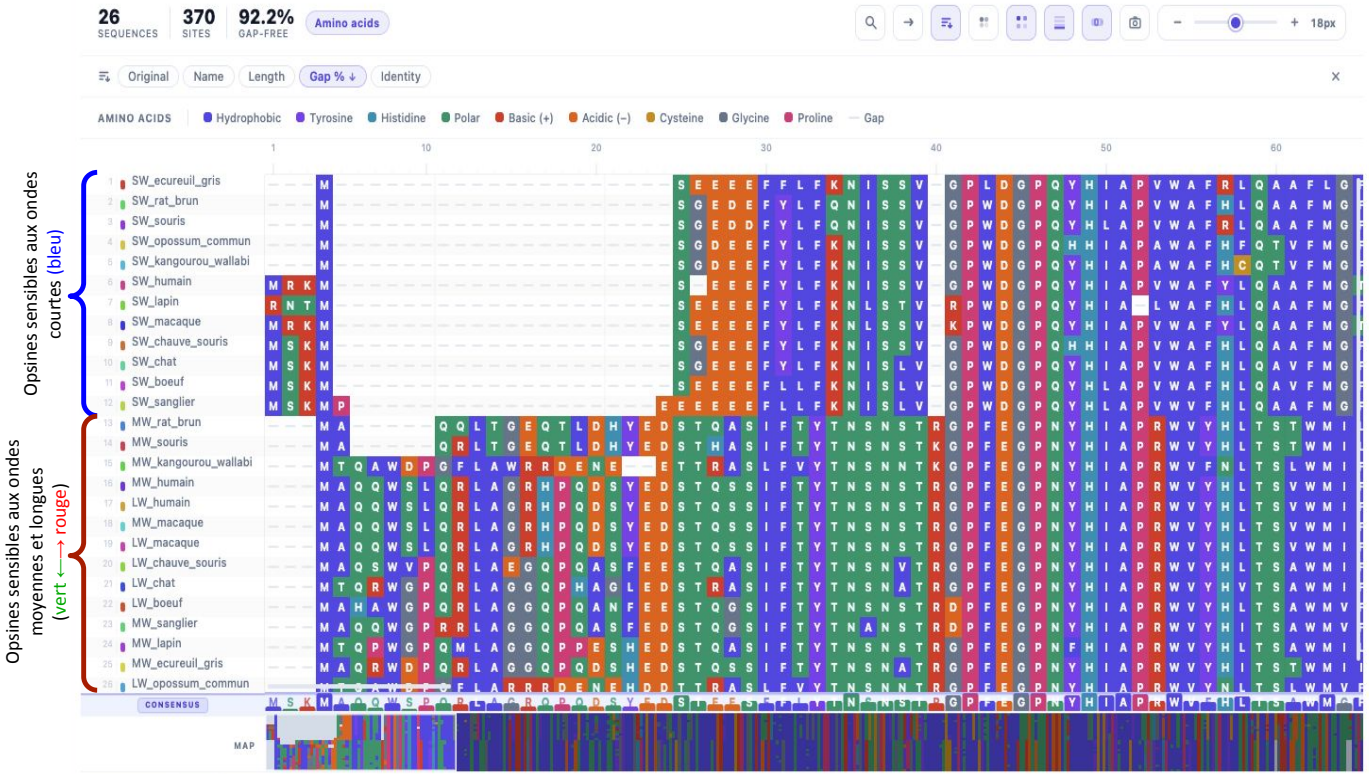
Opsines sensibles aux ondes courtes (violet – bleu)

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

	SW humain	SW macaque	SW souris	SW rat brun	SW ecureuil 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 ecureuil 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	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 ecureuil 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 ecureuil 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

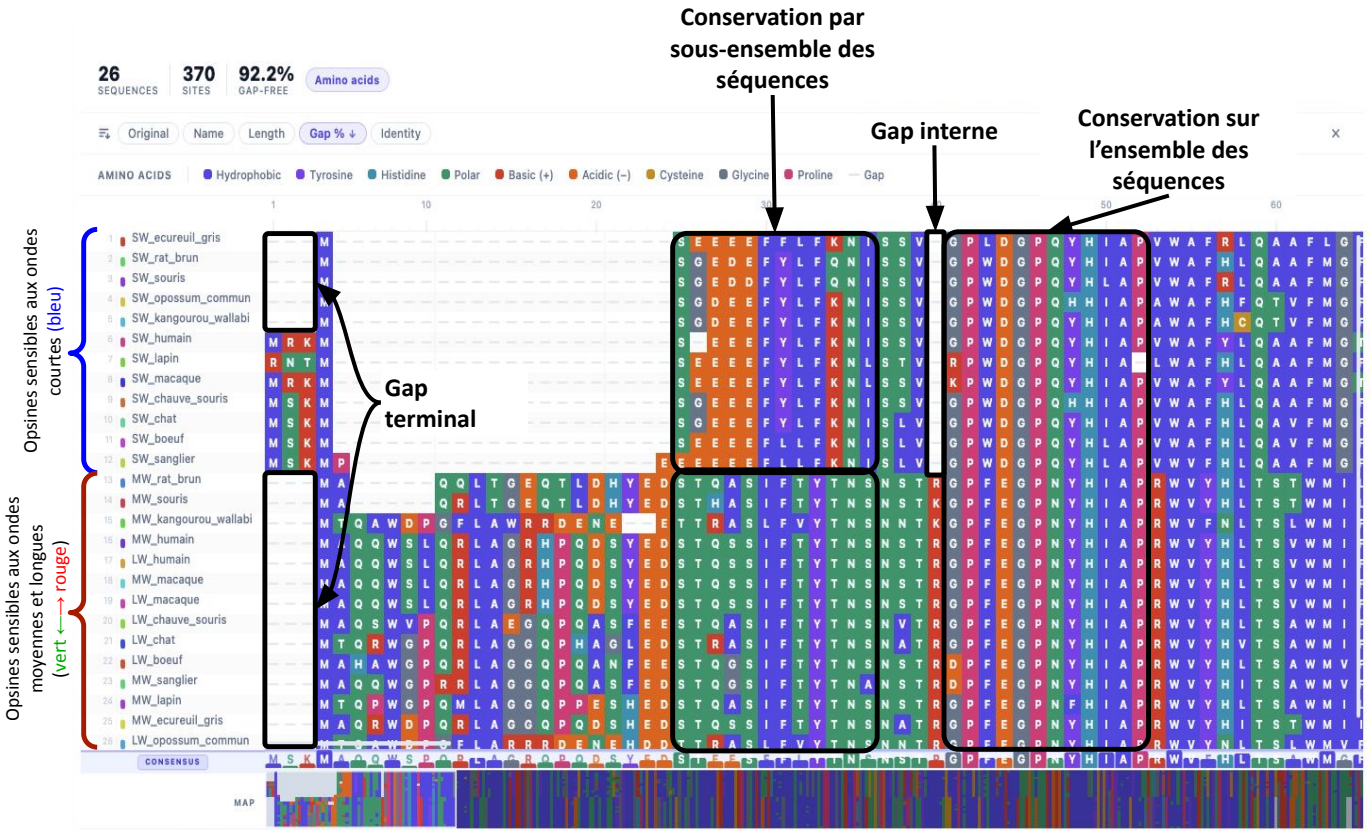
Alignement multiple des opsines de mammifère

- **Alignement multiple** :
  - aligner un ensemble de séquences similaires
  - maximiser les correspondances entre résidus.
- Résultat
  - **Une ligne par séquence**
  - **Une colonne par position**
  - **Couleurs :** propriétés biochimiques des acides aminés



Alignement multiple des opsines de mammifère

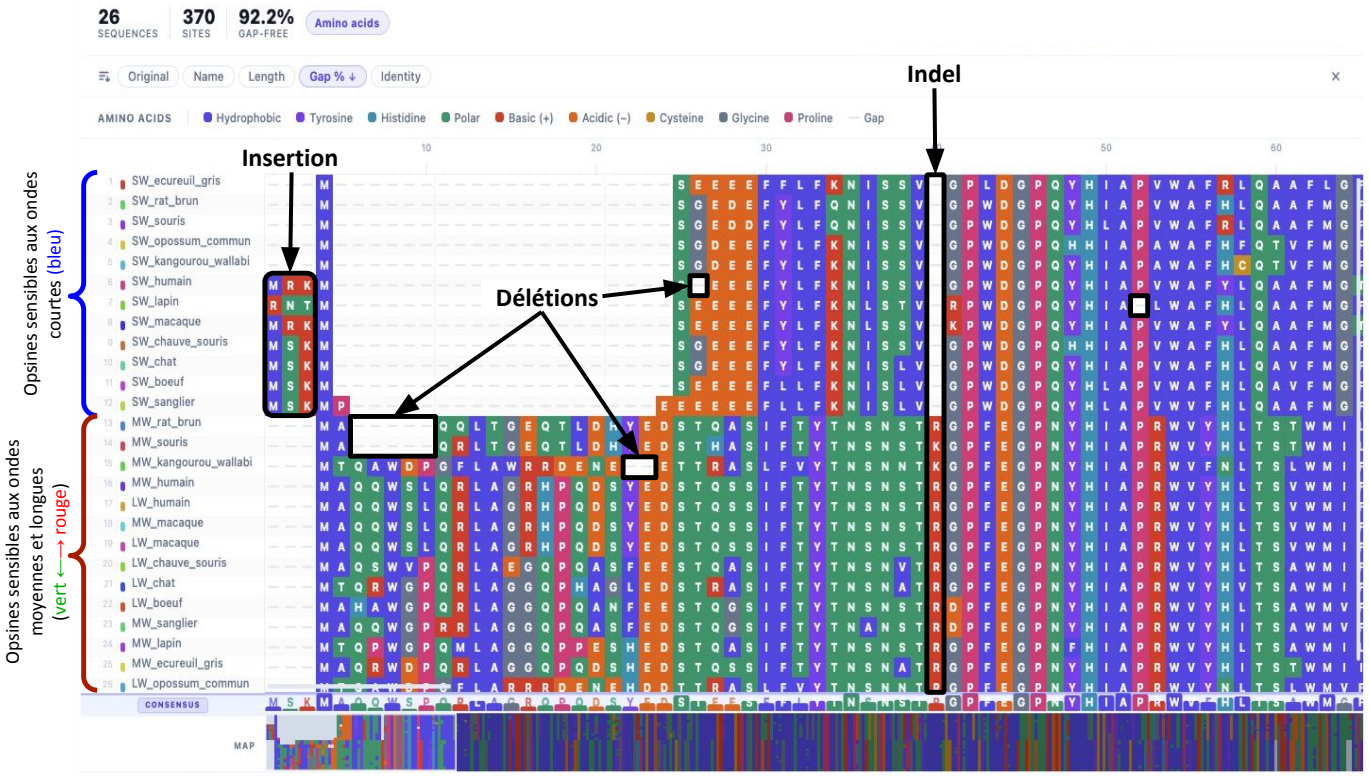
- **Blocs conservés**
  - sur l'ensemble de l'alignement
  - sur un sous-ensemble des séquences.
- **Gaps**
  - Internes
  - Terminaux



Alignement multiple des opsines de mammifère

- **Délétion ou insertion ?**
  - **Délétions :** fragments absents d'une ou de quelques séquences mais présents ailleurs
  - **Insertions :** fragments présents dans une ou quelques séquences mais absents ailleurs
  - **Indels :** positions non-résolues

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 molécules à partir de l'alignement multiple

Méthodes alternatives

- Algorithmes non présentés au cours
- ***A connaître : avantages et inconvénients (à droite)***

  → comment choisir l'algorithme approprié en fonction des jeux de données ?

Peuvent donner des résultats différents

- topologie de l'arbre (branchements)
- longueurs des branches

  → comment estimer la fiabilité d'un arbre ?

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

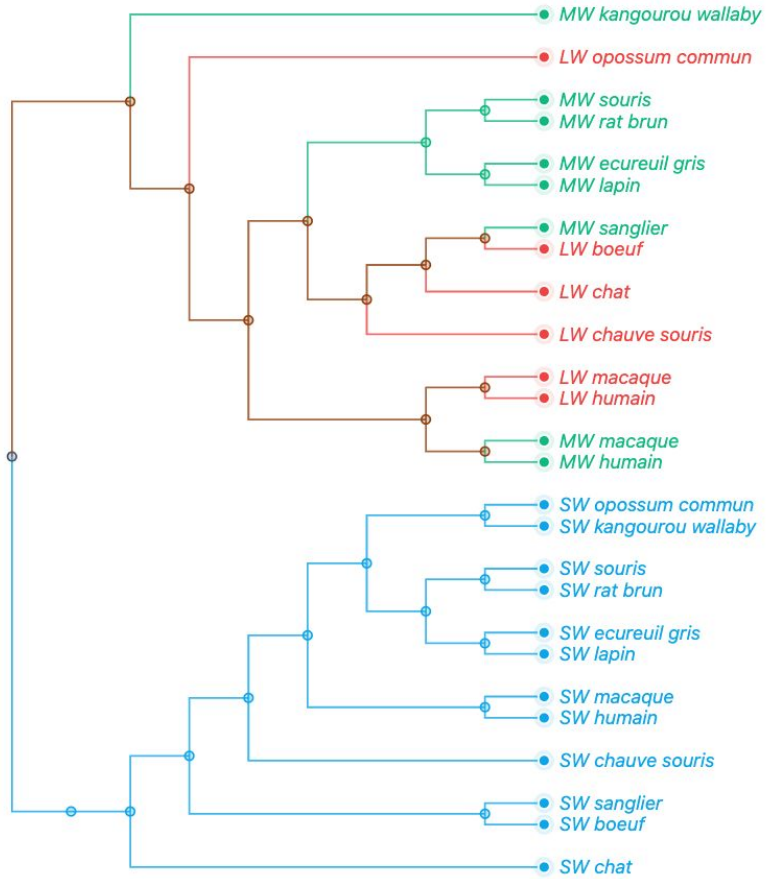
- Rapide
- Permet de traiter un grand nombre de séquences
- Permet d'estimer la longueur des branches (produit un phylogramme)
- Résultats moins fiables que par la méthode du maximum de vraisemblance

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



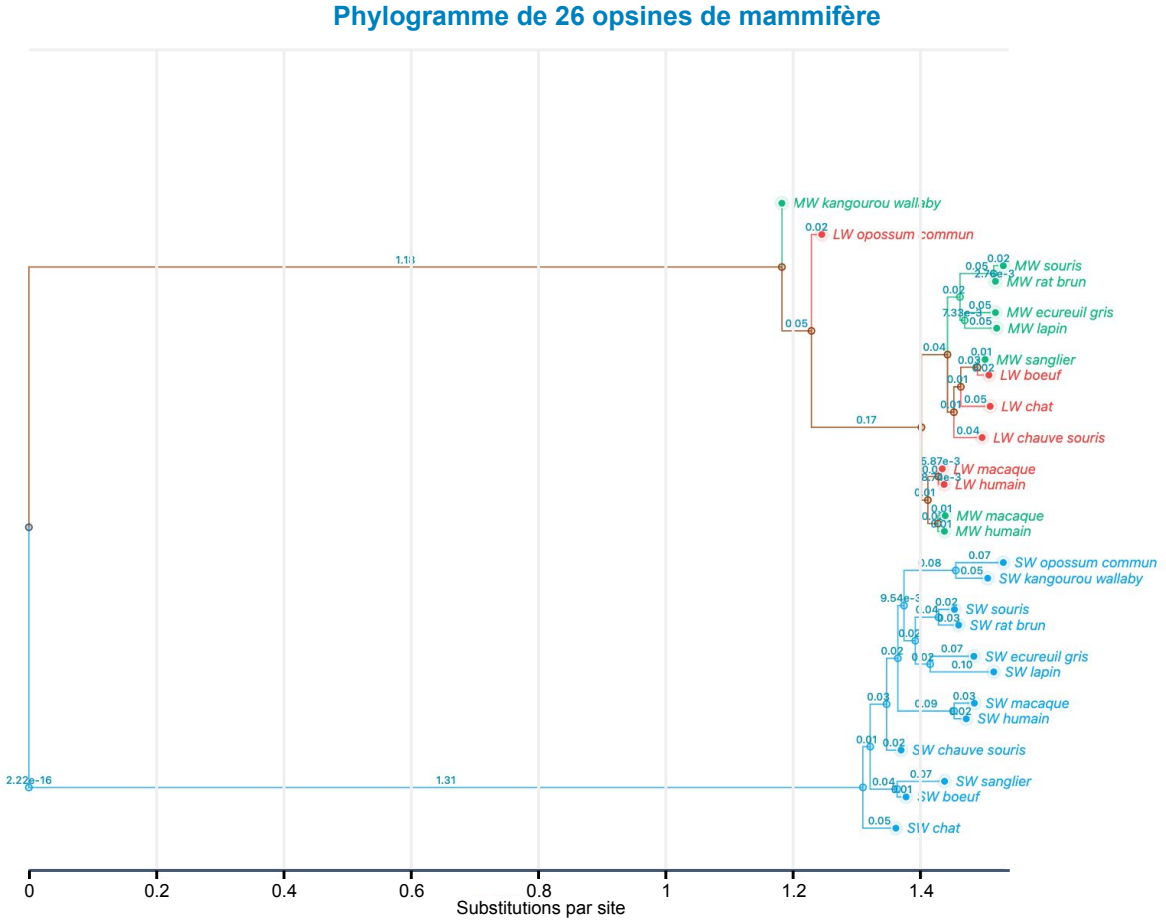
Phylogramme des séquences

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 **nombres associés aux branches** indiquent le **nombre de substitutions par site**.
- **Grande distance** entre SWS et les autres : en moyenne, >1 substitution par site (colonne)

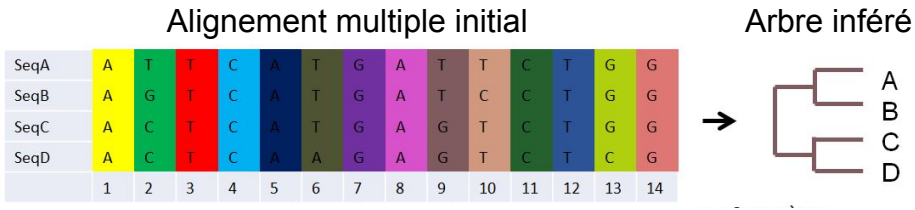


Bootstrapping – à quel point les branchements sont-ils supportés par les données ?

Alignement multiple complet

- M colonnes (positions d'alignement
- S lignes (séquences)

→ inférence d'un l'arbre phylogénétique



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

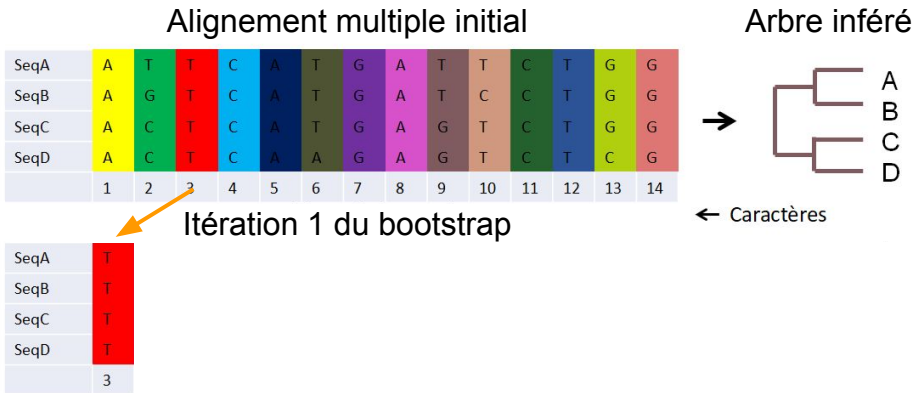
Alignement multiple complet

- M colonnes (positions d'alignement
- S lignes (séquences)

→ inférence d'un l'arbre phylogénétique

Estimation de la robustesse : bootstrapping.

- Sélection aléatoire de M colonnes ***avec remise***
  → Chaque colonne peut être tirée 0, 1 ou plusieurs fois.



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

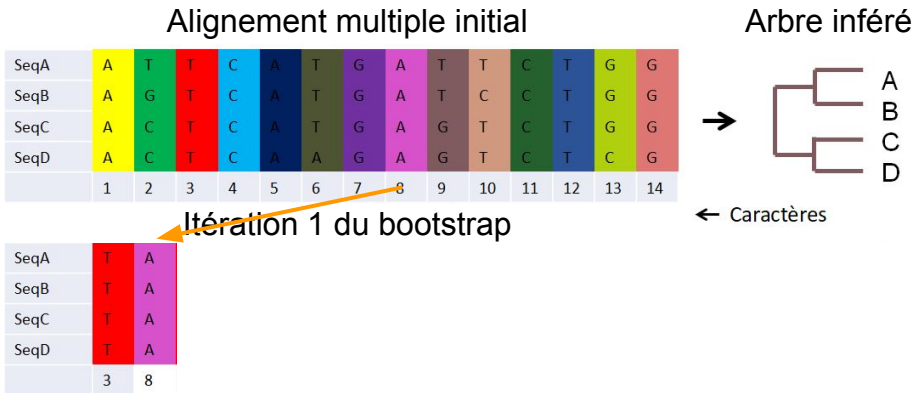
Alignement multiple complet

- M colonnes (positions d'alignement
- S lignes (séquences)

→ inférence d'un l'arbre phylogénétique

Estimation de la robustesse : bootstrapping.

- Sélection aléatoire de M colonnes ***avec remise***
  → Chaque colonne peut être tirée 0, 1 ou plusieurs fois.



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

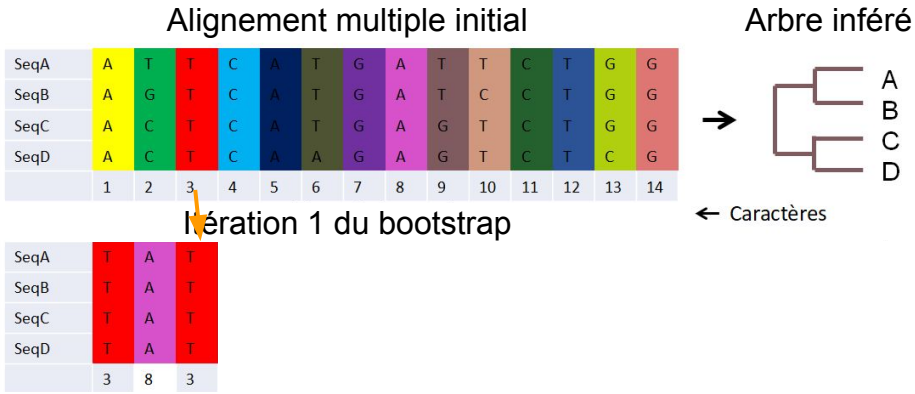
Alignement multiple complet

- M colonnes (positions d'alignement
- S lignes (séquences)

→ inférence d'un l'arbre phylogénétique

Estimation de la robustesse : bootstrapping.

- Sélection aléatoire de M colonnes ***avec remise***
  → Chaque colonne peut être tirée 0, 1 ou plusieurs fois.



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

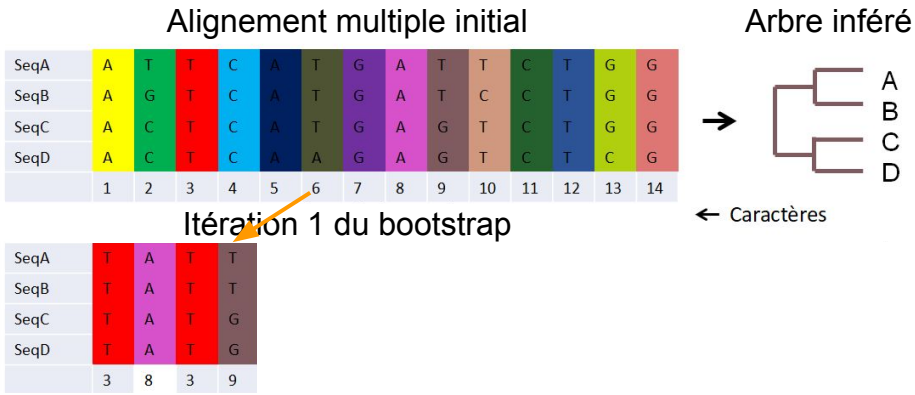
Alignement multiple complet

- M colonnes (positions d'alignement
- S lignes (séquences)

→ inférence d'un l'arbre phylogénétique

Estimation de la robustesse : bootstrapping.

- Sélection aléatoire de M colonnes ***avec remise***
  → Chaque colonne peut être tirée 0, 1 ou plusieurs fois.



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

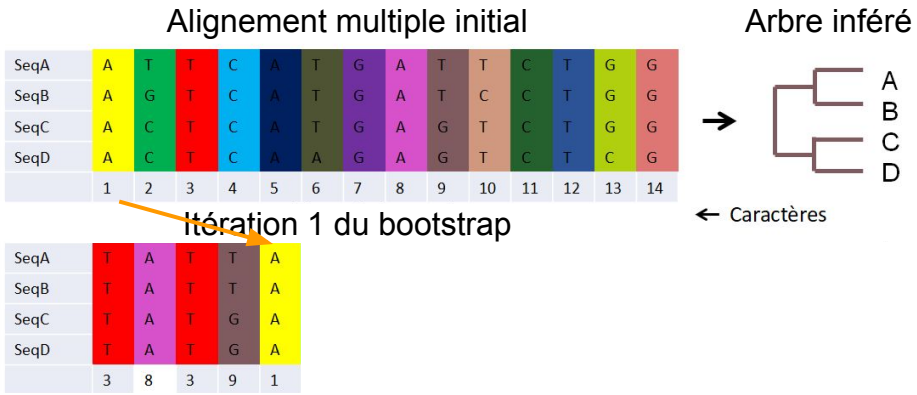
Alignement multiple complet

- M colonnes (positions d'alignement
- S lignes (séquences)

→ inférence d'un l'arbre phylogénétique

Estimation de la robustesse : bootstrapping.

- Sélection aléatoire de M colonnes ***avec remise***
  → Chaque colonne peut être tirée 0, 1 ou plusieurs fois.



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

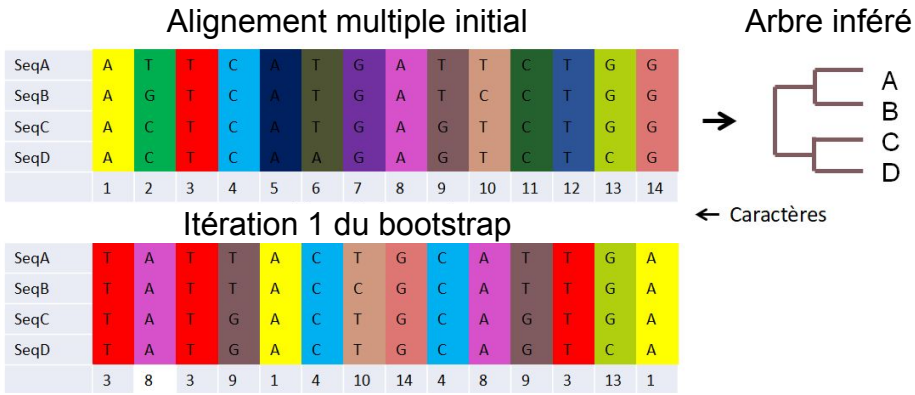
Alignement multiple complet

- M colonnes (positions d'alignement
- S lignes (séquences)

→ inférence d'un l'arbre phylogénétique

Estimation de la robustesse : bootstrapping.

- Sélection aléatoire de M colonnes ***avec remise***
  → Chaque colonne peut être tirée 0, 1 ou plusieurs fois.



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

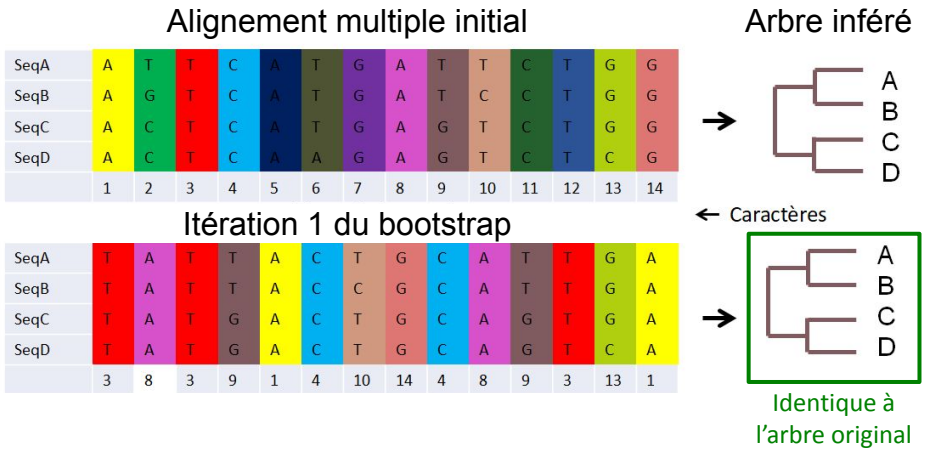
Alignement multiple complet

- M colonnes (positions d'alignement
- S lignes (séquences)

→ inférence d'un l'arbre phylogénétique

Estimation de la robustesse : bootstrapping.

- Sélection aléatoire de M colonnes ***avec remise***
  → Chaque colonne peut être tirée 0, 1 ou plusieurs fois.
- Inférence d'un arbre avec ces colonnes ré-échantillonnées.



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

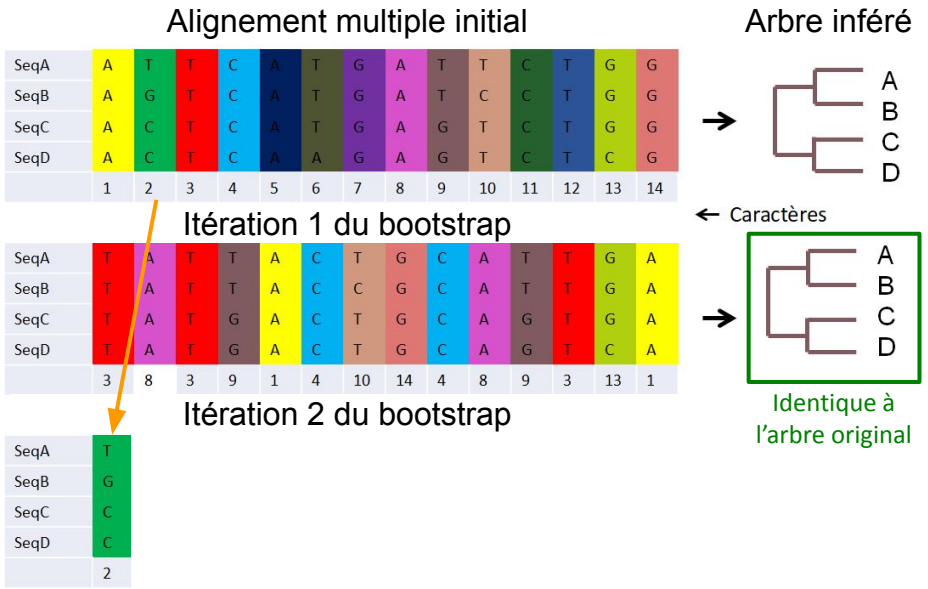
Alignement multiple complet

- M colonnes (positions d'alignement
- S lignes (séquences)

→ inférence d'un l'arbre phylogénétique

Estimation de la robustesse : bootstrapping.

- N iterations (ex: N=100 ou N=1000)
  - Sélection aléatoire de M colonnes ***avec remise***
    → Chaque colonne peut être tirée 0, 1 ou plusieurs fois.
  - Inférence d'un arbre avec ces colonnes ré-échantillonnées.



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

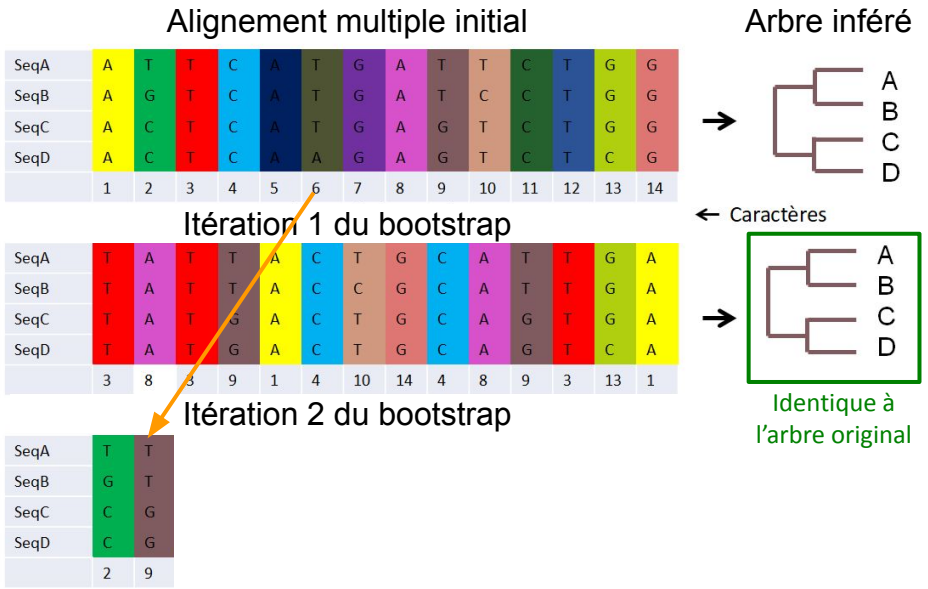
Alignement multiple complet

- M colonnes (positions d'alignement
- S lignes (séquences)

→ inférence d'un l'arbre phylogénétique

Estimation de la robustesse : bootstrapping.

- N iterations (ex: N=100 ou N=1000)
  - Sélection aléatoire de M colonnes ***avec remise***
    → Chaque colonne peut être tirée 0, 1 ou plusieurs fois.
  - Inférence d'un arbre avec ces colonnes ré-échantillonnées.



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

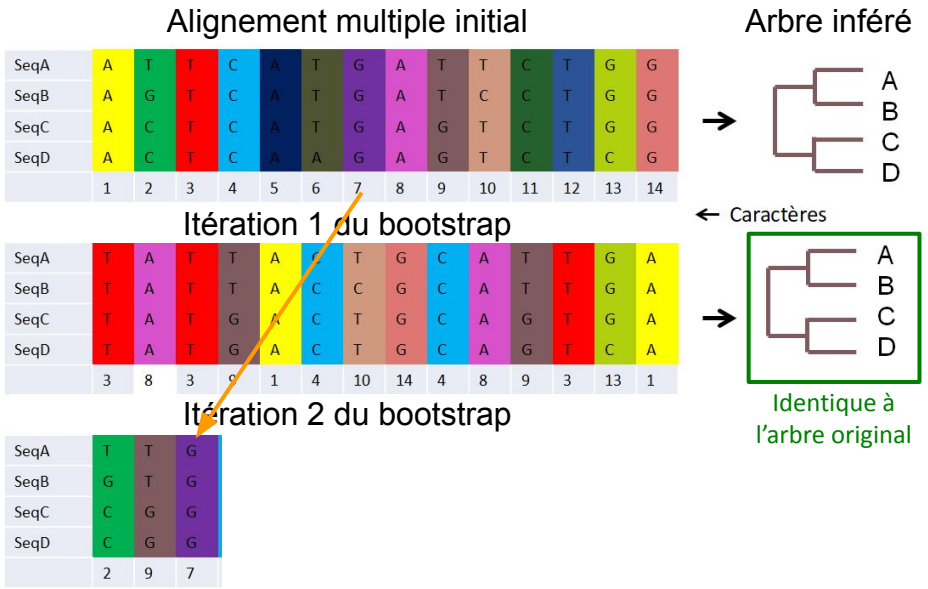
Alignement multiple complet

- M colonnes (positions d'alignement
- S lignes (séquences)

→ inférence d'un l'arbre phylogénétique

Estimation de la robustesse : bootstrapping.

- N iterations (ex: N=100 ou N=1000)
  - Sélection aléatoire de M colonnes ***avec remise***
    → Chaque colonne peut être tirée 0, 1 ou plusieurs fois.
  - Inférence d'un arbre avec ces colonnes ré-échantillonnées.



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

Alignement multiple complet

- M colonnes (positions d'alignement
- S lignes (séquences)

→ inférence d'un l'arbre phylogénétique

Estimation de la robustesse : bootstrapping.

- N iterations (ex: N=100 ou N=1000)
  - Sélection aléatoire de M colonnes ***avec remise***
    → Chaque colonne peut être tirée 0, 1 ou plusieurs fois.
  - Inférence d'un arbre avec ces colonnes ré-échantillonnées.

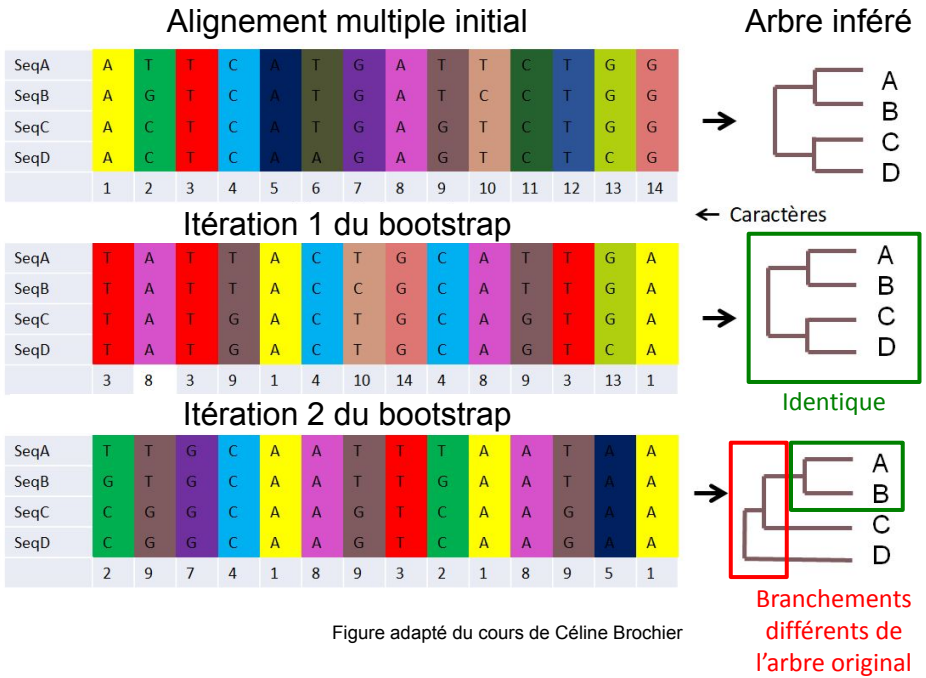


Figure adapté du cours de Céline Brochier

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

Alignement multiple complet

- M colonnes (positions d'alignement
- S lignes (séquences)

→ inférence d'un l'arbre phylogénétique

Estimation de la robustesse : bootstrapping.

- N iterations (ex: N=100 ou N=1000)
  - Sélection aléatoire de M colonnes ***avec remise***
    → Chaque colonne peut être tirée 0, 1 ou plusieurs fois.
  - Inférence d'un arbre avec ces colonnes ré-échantillonnées.
- **Valeur de bootstrap** de chaque branchement
  - nombre de fois où ce branchement se retrouve à l'identique dans les N arbres produits.
  - → indique la robustesse de l'arbre par rapport aux données
  - **/!** Ne garantit pas que l'arbre reflète l'histoire évolutive (les données peuvent être erronées ou biaisées)

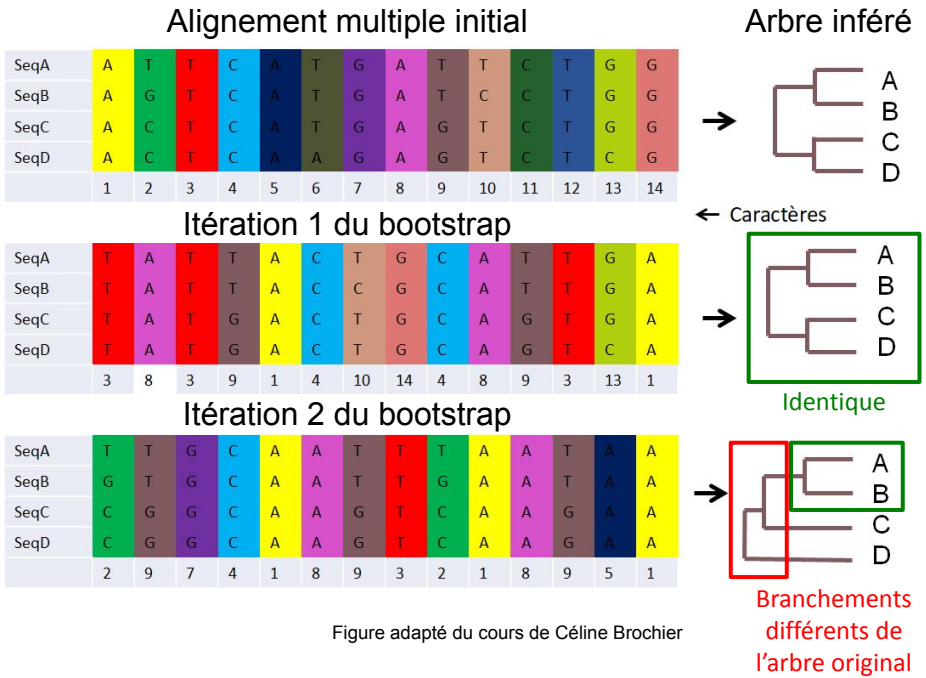


Figure adapté du cours de Céline Brochier

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

Valeur de bootstrap

- Assignée à chaque branchement.
- Indique le nombre de fois où ce branchement se retrouve à l'identique, parmi les N arbres de bootstrap.
- **Valeur élevée** → **robuste**
- **Valeur faible** → sensible

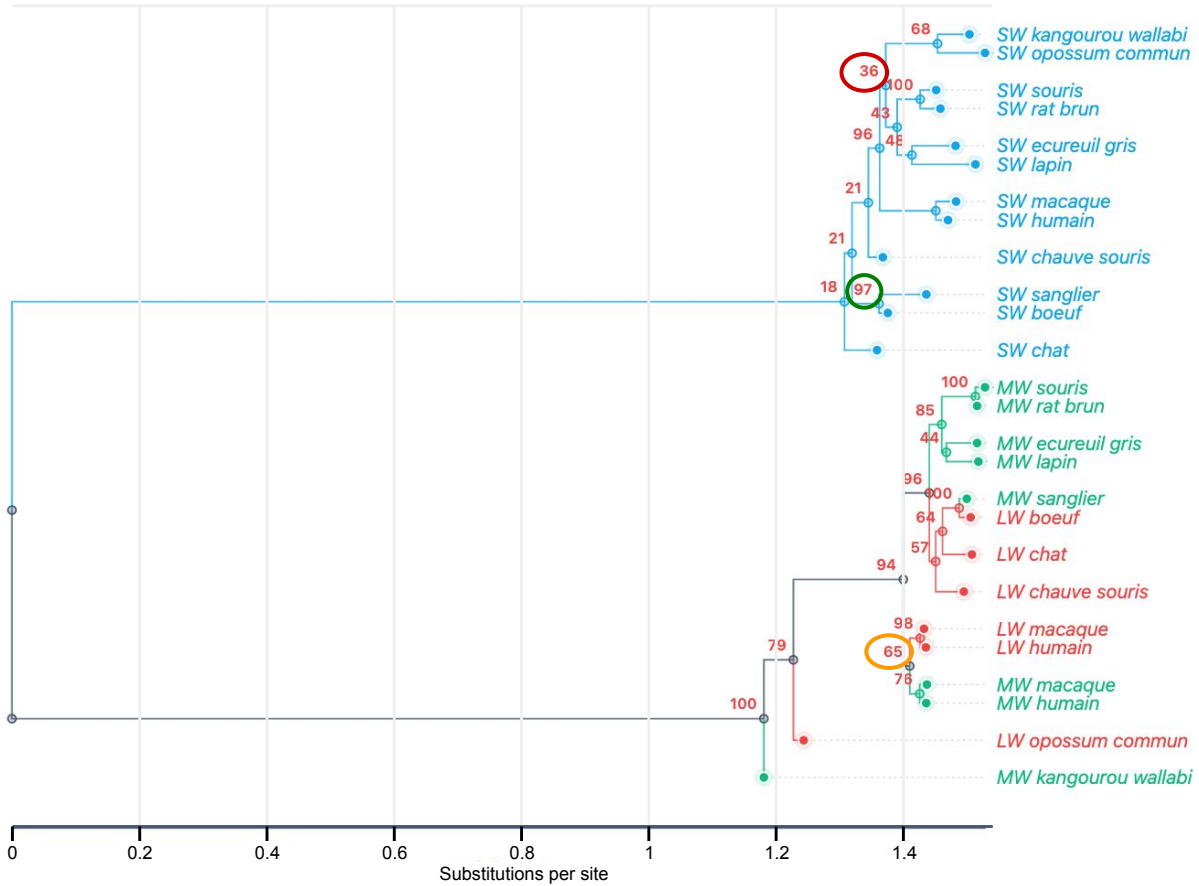
Exemple : 36/100
→ branchement identique dans ~1/3 des bootstraps
→ dépend fortement des colonnes
→ pas représentatif

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*

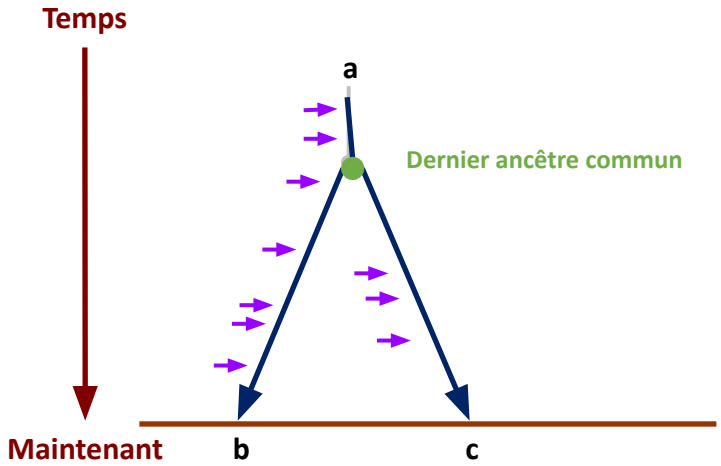
Phylogramme de 26 opsines de mammifère



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

Séquences homologues ou analogues

- **Homologie**
  - Ressemblances héritées d'une **séquence ancestrale commune**
  - Évolution **divergente :** accumulation de mutations à partir de la séquence ancestrale commune
- **Séquences analogues**
  - Ressemblances ne provenant pas d'une séquence ancestrale commune
  - Évolution **convergente** : ressemblances apparues indépendamment dans les deux lignées
  - Facteurs de sélection communs.
  - Exemple : éléments cis-régulateurs.



Les séquences b et c sont homologues

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

- Homologie = tout ou rien

  → **Formulations inappropriées**
  - *“Ces séquences ont 50% d’homologie”*
  - *“Niveau d’homologie”*

Formulation correcte
  - Ces séquences sont vraisemblablement homologues car elles ont taux de similarité trop significatif pour s’expliquer par une coïncidence ou par une convergence évolutive.

Inférence d'homologie

- Nous ne disposons que des séquences actuelles → on ne peut pas prouver l'homologie
- On peut par contre *inférer* l'homologie sur base d'une estimation de la vraisemblance

Alignement de séquences

Calcul des taux d'identités / similarité :

Attention, les pourcentages peuvent être trompeurs sur des petites séquences

→ on calcule la **significativité statistique**

- 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**

Evénements évolutifs générant des séquences homologues

Duplication

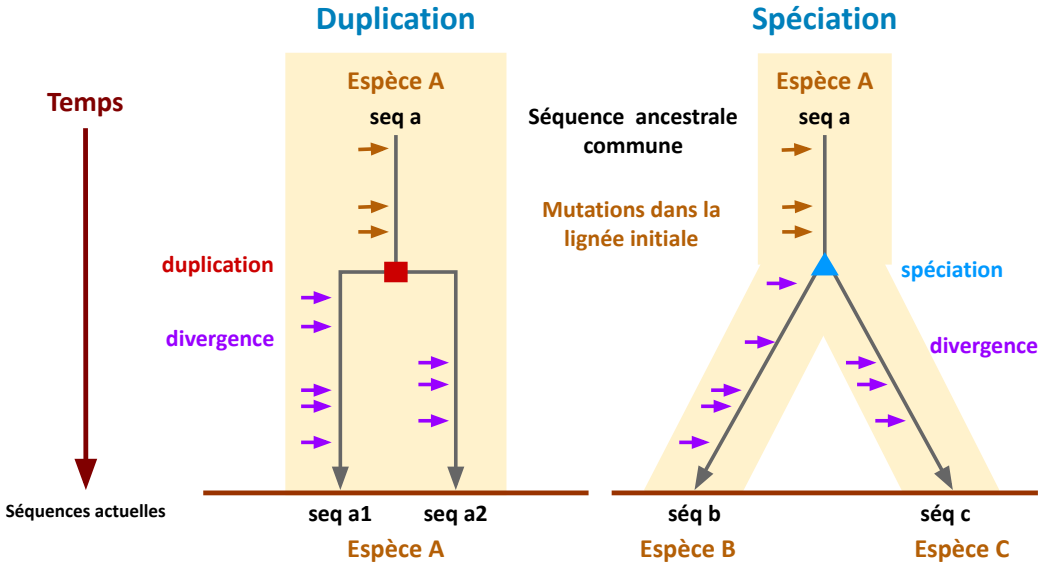
- Dédoublement d'une partie de l'ADN génomique.
- Taille variable : quelques nucléotides, un ou plusieurs gènes, un chromosome entier, le génome entier
- Crée de la redondance → plusieurs suites possibles
  - Délétion d'une copie
  - Mutations → on trouve dans le génome des traces d'un gène non-fonctionnel ("gène fossile", pseudo-gène)
  - Mutations → divergence, les deux copies restent fonctionnelles et se spécialisent

Spéciation

- Formation d'espèces distinctes à partir d'une espèce unique.
- Les deux lignées-filles héritent de tous les gènes
- Ensuite, accumulation de mutations indépendamment dans les deux lignées

Scénarios évolutifs

- Séquences communes
- Deux scénarios alternatifs
  - ***duplication*** (création de deux copies du gène dans le même génome)
  - ***spéciation*** (formation d'espèces séparées à partir d'une espèce unique).
- Accumulation de mutations → diversification des séquences, des structures et des fonctions

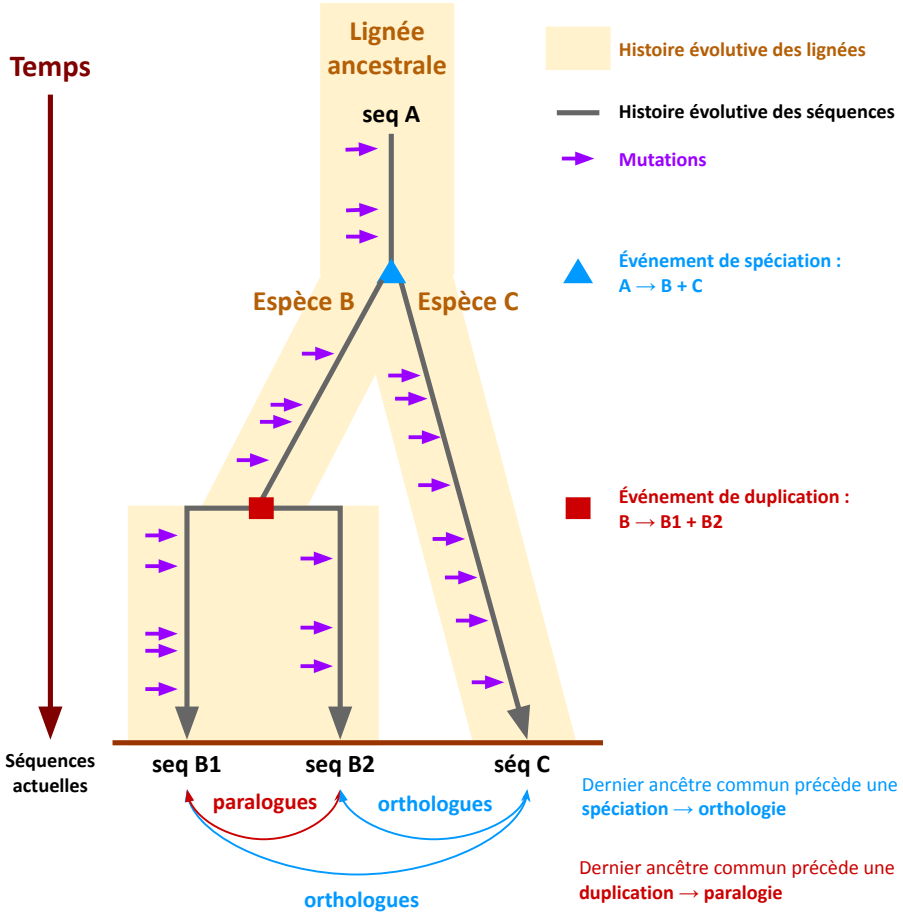


Orthologie versus paralogie

Définitions de Zvelebil (claires et opérationnelles)

- **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***
  - C et B2 sont ***orthologues***
  - B1 et B2 sont ***paralogues***

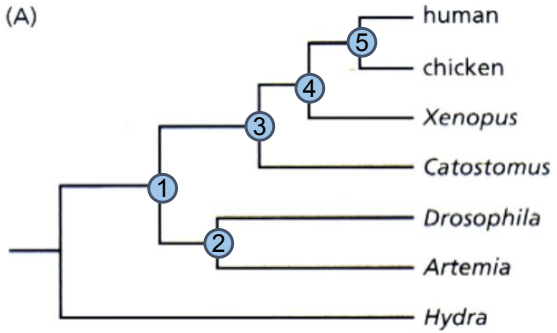


Zvelebil, M. J. & Baum, J. O. Understanding bioinformatics. (Garland Science, 2008).

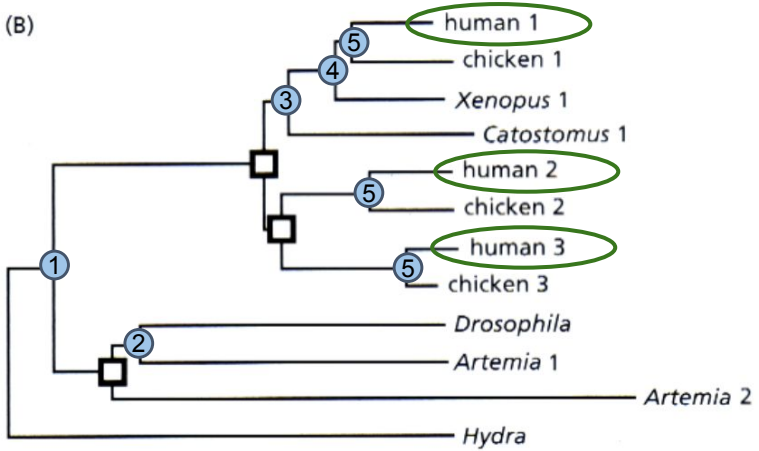
Arbre des espèces et arbre des molécules

- Arbre des espèces → chaque noeud interne représente une spéciation
- Arbre des molécules → chaque noeud interne représente soit une duplication, soit une spéciation

Arbre des espèces



Arbre des séquences

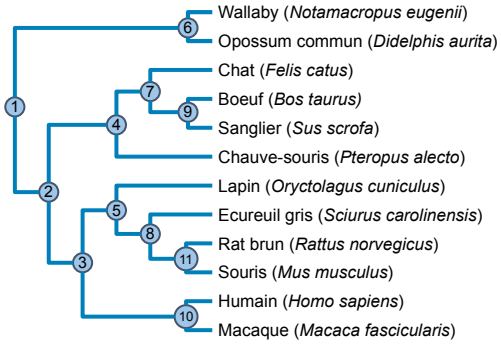


Source: Zvelebil, M.J. and Baum, J.O. (2008) Understanding Bioinformatics. Garland Science, New York and London.

La réconciliation ne fonctionne pas toujours

- 12 espèces de mammifères
- 26 séquences d' opsines
- **Duplication initiale [A]** → vision dichromatique (SWS versus LWS/MWS)
- **Seconde duplication [B]** → vision trichromatique des primates
- Cercles numérotés : **noeuds cohérents** entre arbre des espèces et arbre des molécules
- **X** : **noeuds non réconciliables**
  - Souvent valeur très faible de bootstrap (mais pas toujours).

Arbre des espèces

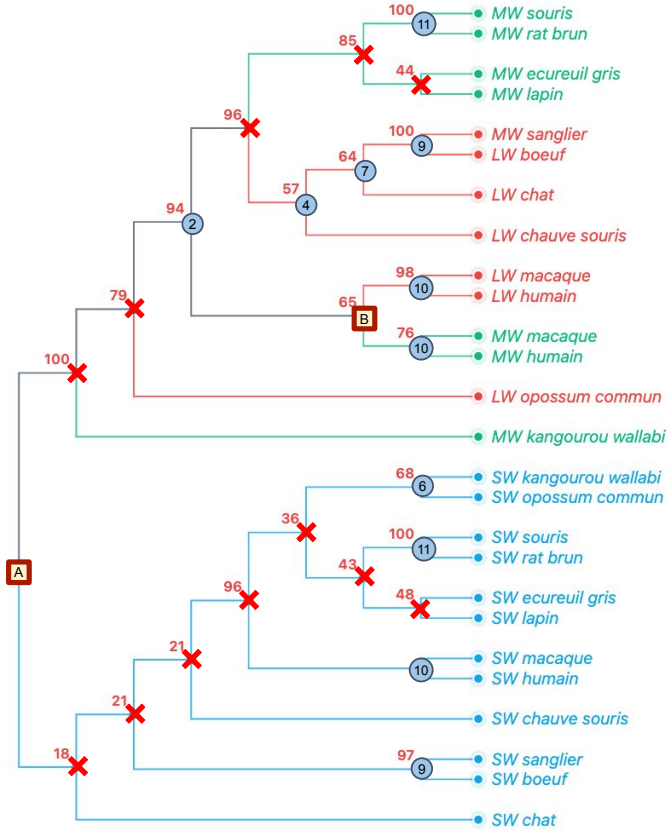


Duplication

Spéciation réconciliée

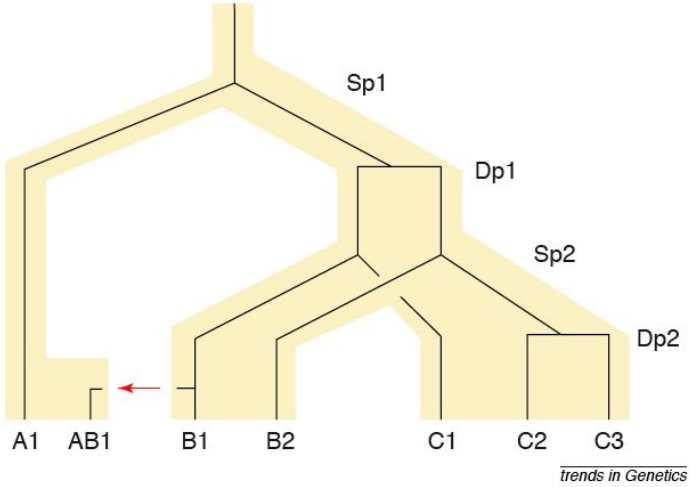
Noeud non réconciliable

Arbre des séquences (opsines)



Représentation détaillée des événements de spéciation / duplication

- **Arbre des molécules** : lignes noires
- **Arbre des espèces** : fond jaune
- **Spéciations (Sp)** : branchements triangulaires
- **Duplications (Dp)** branchements rectangulaires



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=>B1 implies B1=>A1 where => should be read, for example, as 'is orthologous to.' However, the relationships are not transitive. Thus, C2=>A1=>C3 might be true, but it is not necessarily therefore true that C2=>C3, as indeed it is not in the figure if => is read as 'is orthologous to.' A different non-transitivity occurs for 'is paralogous to' with B2=>C1=>C2.

- 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

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

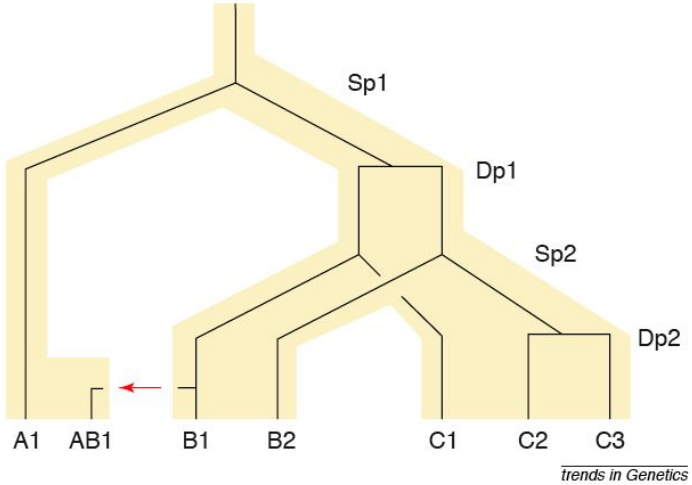
Exercice: types d'homologie

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

P paralogie
O orthologie
X xenologie
A analogie

Solutions 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 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.*

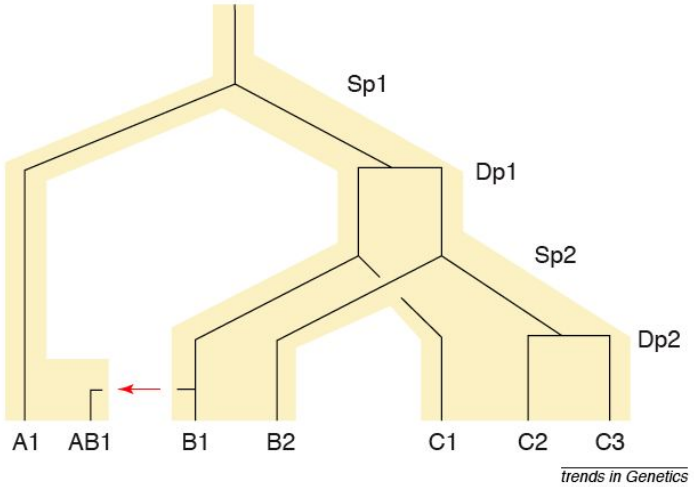
Figure: Fitch, W. M. Homology a personal view on some of the problems. Trends Genet 16, 227–231 (2000). doi.org/10.1016/s0168-9525(00)02005-9
Définitions : Zvelebil, M. J. & Baum, J. O. Understanding bioinformatics. (Garland Science, 2008).

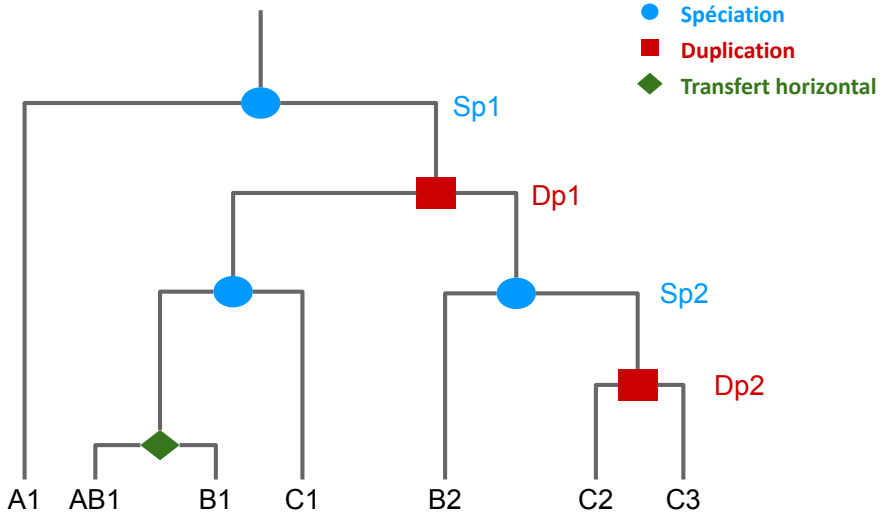
Représentation classique des spéciations / duplications

Représentations de la même histoire évolutive.

- Gauche (Fitch, 2000): arbre moléculaire (lignes noires) **superposé** à l'arbre des espèces (fond jaune)
- Droite (du type Ensembl genomes) : arbre des molécules, avec coloration des noeuds internes selon le type d'événement évolutif

Lettres A, B, C → espèces
1, 2, 3 → copies des gènes





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

Phylogénomique

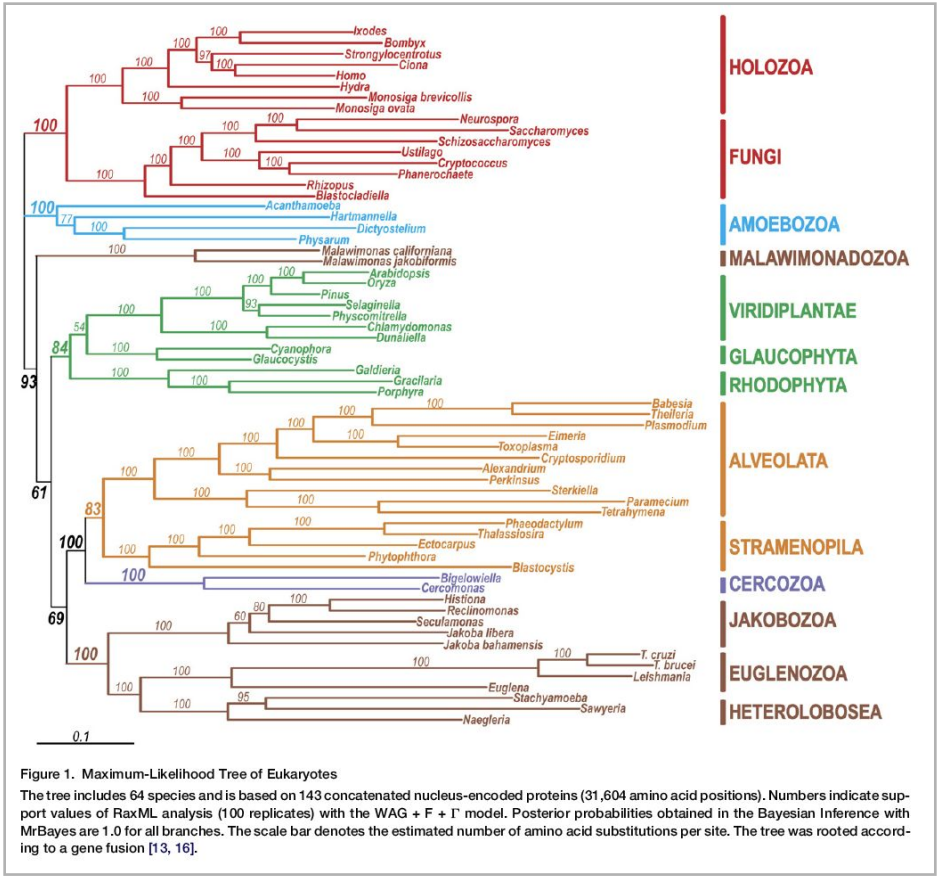


Figure 1. Maximum-Likelihood Tree of Eukaryotes

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

- Phylogénie moléculaire "classique"
  - → **arbre d'une famille de molécule**
- Phylogénomique
  - Inférence d'une **arbre des espèces**
  - Données: séquences de **plusieurs centaines de gènes ou protéines**
- → phylogénies entre organismes très éloignés (règnes différents)
- scénarios concernant les premières étapes de la diversification des êtres vivants.

Exemple (figure de gauche)

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

Source: Rodríguez-Ezpeleta et al. Curr Biol (2007) vol. 17 (16) pp. 1420-5
Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. `

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

Classification des **tenrecs** au 20è siècle

- ressemblance morphologique avec les hérissons → même ordre (insectivores)
- différences anatomiques importantes → famille séparée (tenrécidés)

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	**Tenrécidés**
genres et espèces	**Centetes ecaudatus** *(tenrec commun)* **Hemicentetes semispinosus** *(tenrec à bandes ou tenrec strié)* **Limnogale mergulus** *(tenrec à pieds palmés)* **Microgale longicaudata** *(tenrec à longue queue)* **Oryzorictes hova** *(tenrec des rizières)* **Setifer setosus** *(tenrec-hérisson)* *et autres*

Caratini, R. La Vie animale. (Bordas, 1981). ISBN: 978-2-04-011918-8

Les Afrotheria

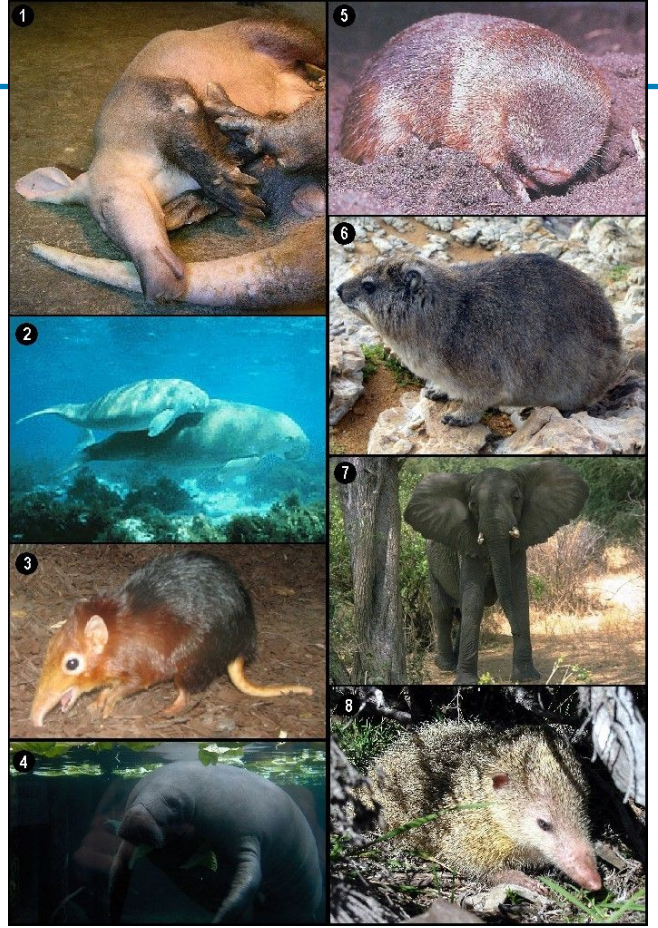
Quels sont les plus proches cousins du tenrec ?

Inférence phylogénétique (fin des années 1990)

→ regroupement d'espèces très différentes

→ définition du **super-ordre des Afrotheria**

Autres Afrotheria: éléphant, lamentin, dugong, oryctérope, taupe dorée, daman, …



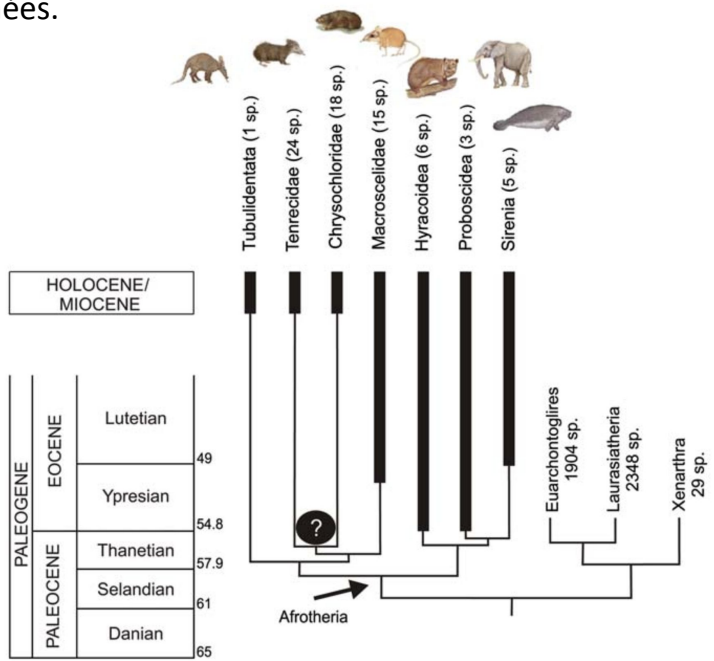
1. Oryctérope du Cap; 2. Dugongs; 3. Macroscélide de Peters ; 4. Lamantin; 5. Taupe dorée; 6. Daman du Cap; 7. Éléphant de savane d'Afrique ; 8. Tangue ("tailless tenrec"); 8, 8b: Tenrec

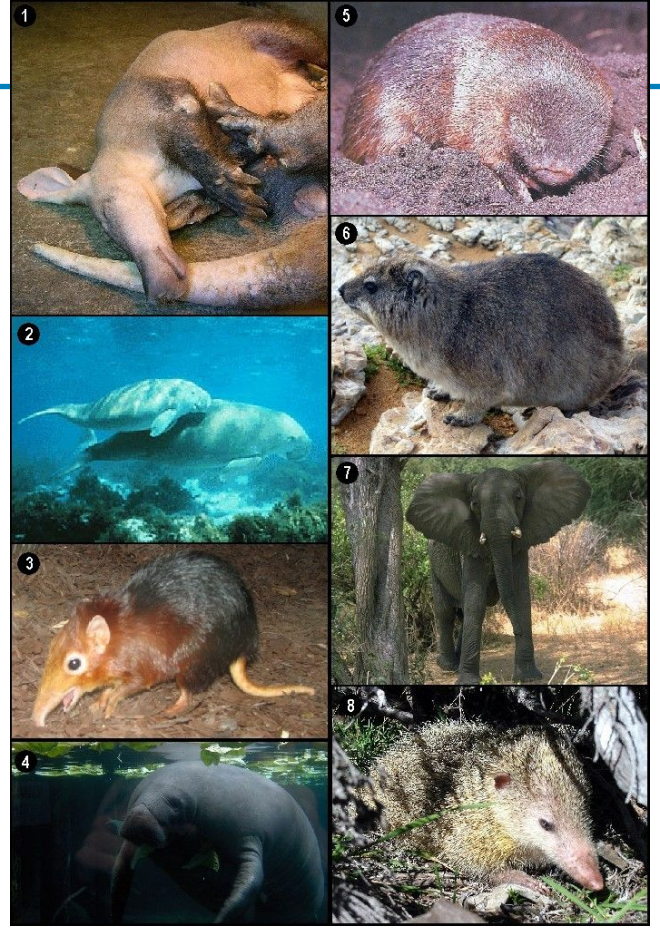
http://upload.wikimedia.org/wikipedia/commons/0/01/Kleiner-igeltanrek-a.jpg

- 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. Tangue ("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

Radiation des mammifères

Resolution of the Early Placental Mammal Radiation Using Bayesian Phylogenetics

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

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

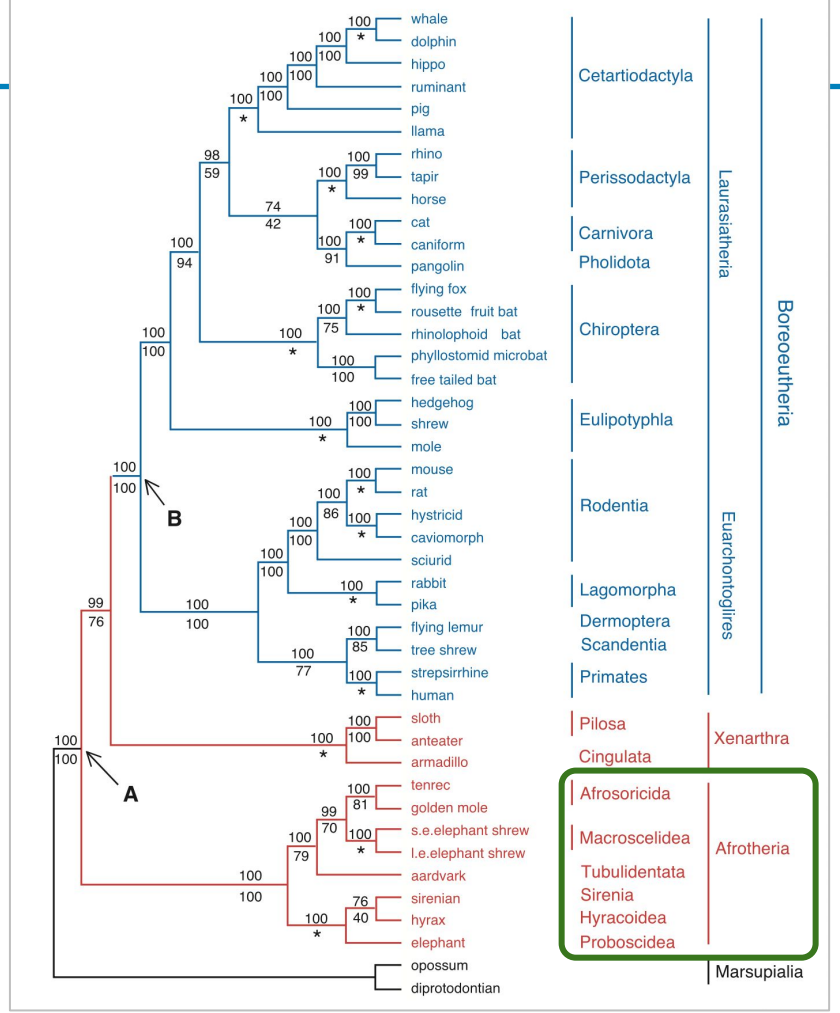


Fig. 1. Phylogeny of living placental mammals reconstructed using a Bayesian phylogenetic approach. An identical topology was obtained with maximum likelihood [–ln L = 211110.54; see (*15*) for methodological details]. The number above each branch refers to the Bayesian posterior probability (shown as percentages; i.e., 95 represents a posterior probability of 0.95) of the node derived from 26,250 MCMC sampled trees on the basis of the complete 16.4-kb data. Additional analyses with the full data set and with data sets that varied taxon sampling (i.e., jackknifing single outgroup taxa) and character sampling (nuclear only and nuclear coding loci only) produced similarly high posterior probabilities (*15*). Values below branches represent percent support in maximum likelihood (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 (*15*). Blue, monophyletic Northern Hemisphere group (i.e., Boreoeutheria); red, paraphyletic Southern Hemisphere group (i.e., Xenarthra + Afrotheria); black, outgroups.

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

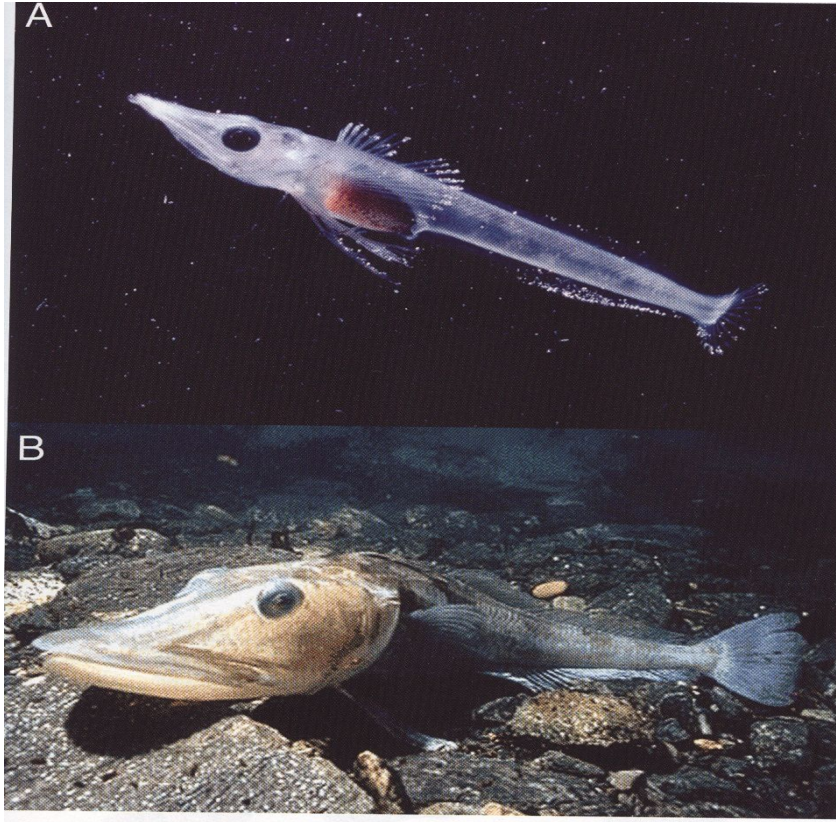
Pseudo-gènes (“gènes fossiles”) : les reliques de gènes ancestraux dans les génomes d’espèces actuelles

Les pseudogènes (“Gènes fossiles”)

- **Pseudogènes**
  - Gènes défectifs (ont perdu leur activité), mais dont on retrouve des traces dans les génomes.
  - Aussi appelés “gènes fossiles” mais risque de confusion avec les gènes fonctionnels dans les génomes fossiles (ex: *Homo neandertalis*)
- On trouve généralement de nombreux codons stops dans les séquences codantes.
- Succède à une relaxation de la pression sélective:
  - Mutations aléatoires sur toutes les parties du génome
    - Perturbent la fonction → éliminées (sélection naturelle « purificatrice »).
    - Occasionnellement, un gène devient dispensable → sélection relâchée → mutations s’accumulent
- Exemple : plusieurs espèces de levure (indépendamment)
  - perte de la capacité de digérer le galactose
  - Fossilisation de *chacun* des 7 gènes GAL (devenus dispensables)

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

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 mackeral icefish, *Champsocephalus gunnari*.

- Poissons des glaces (Channichthyidae)
- Océan Arctique, eaux de 4°C à -2°C (salée → liquide à ces T°).
- **Protéines antigels** dans le sang
- **Absence de globules rouges** !
- Échanges d’oxygène à travers la peau
- Ni hémoglobine ni myoglobine fonctionnelle
- Présence dans le **génome de 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

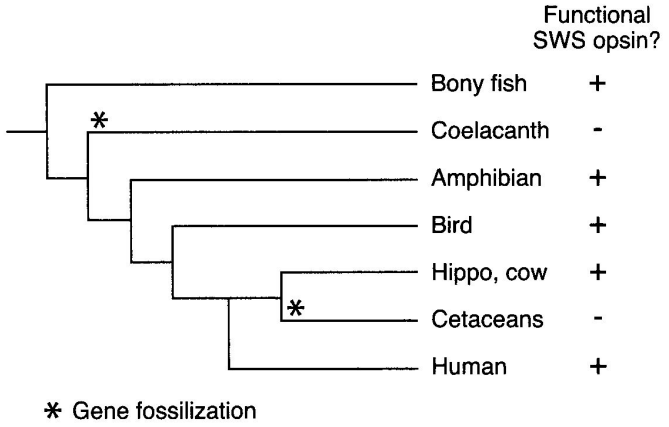


FIG. 5.2. **The same opsin gene has been fossilized twice.** The distribution of different mutations found in the coelacanth and cetacean SWS opsins and the evolutionary relationship of these species indicates that the SWS opsin was fossilized at least twice (asterisk). *Figure by Jamie Carroll.*

- Chez les coelacanthes et les cétacés, la perception des couleurs a été perdue.
- Perte indépendante chez les coelacanthes et chez les cétacés.
- Gènes des "c-opsines" (perception des couleurs) toujours présents, mais plus fonctionnels
- Grand nombre de mutations, y compris codons stops

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

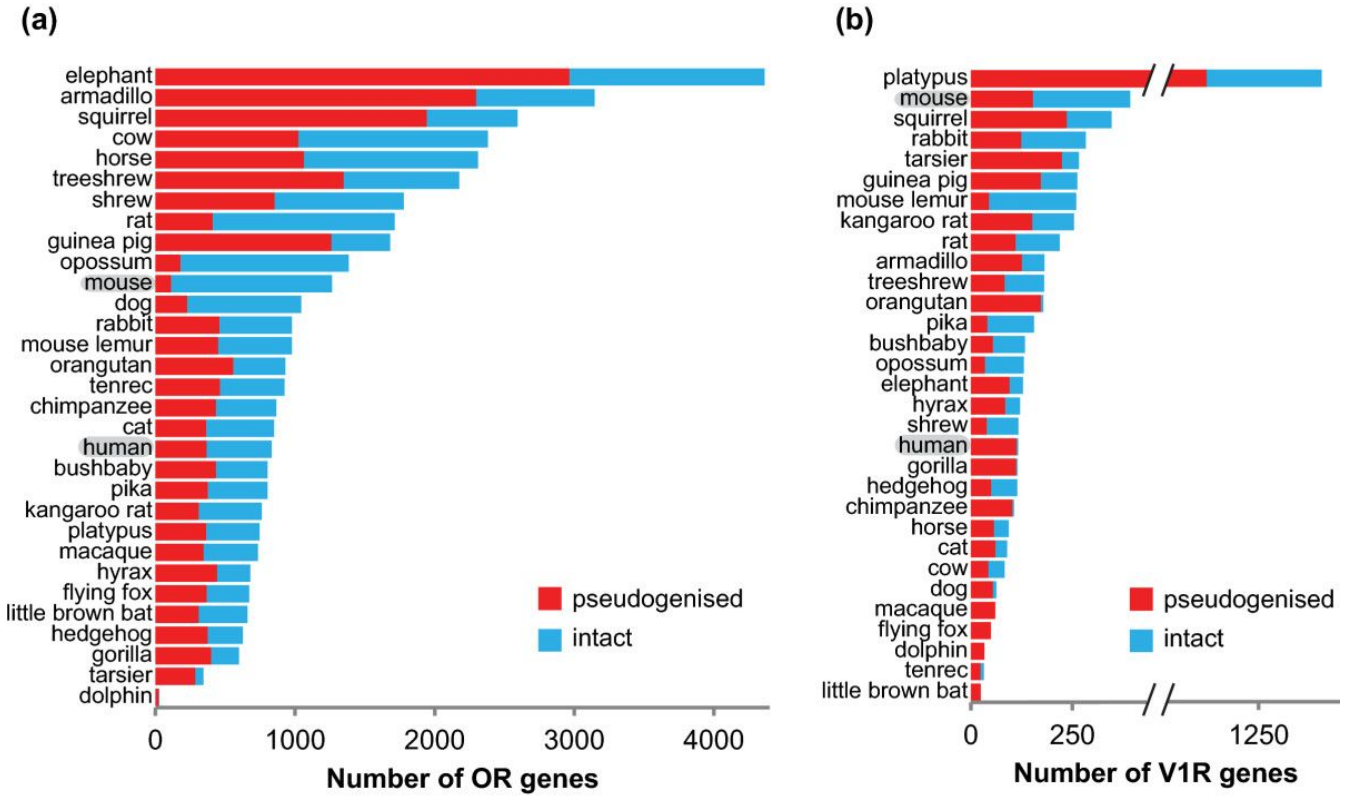
La pseudogénisation des récepteurs olfactifs (OR) et des vomérorécepteurs (VR)

Gauche: récepteurs olfactifs (OR)

- Souris:
  - 25.000 gènes
  - 1.400 récepteurs olfactifs.
- Humain
  - 21.000 gènes
  - 400 récepteurs olfactifs
  - Perte massive de récepteurs olfactifs : fossilisés, non-fonctionnels.
- Idem chez les autres primates de l'ancien monde, qui ont acquis la vision trichromatique.

Droite: vomérorécepteurs (V1R)

- Homologues des récepteurs olfactifs
- Perception des phéromones.
- Humain et gorille :
  - (quasiment) pas de communication par phéromones
  - génome comporte une centaine de vomérorécepteurs pseudogénisés.



Keller. Different noses for different mice and men. BMC Biol (2012) vol. 10 pp. 75. doi.org/10.1186/1741-7007-10-75

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

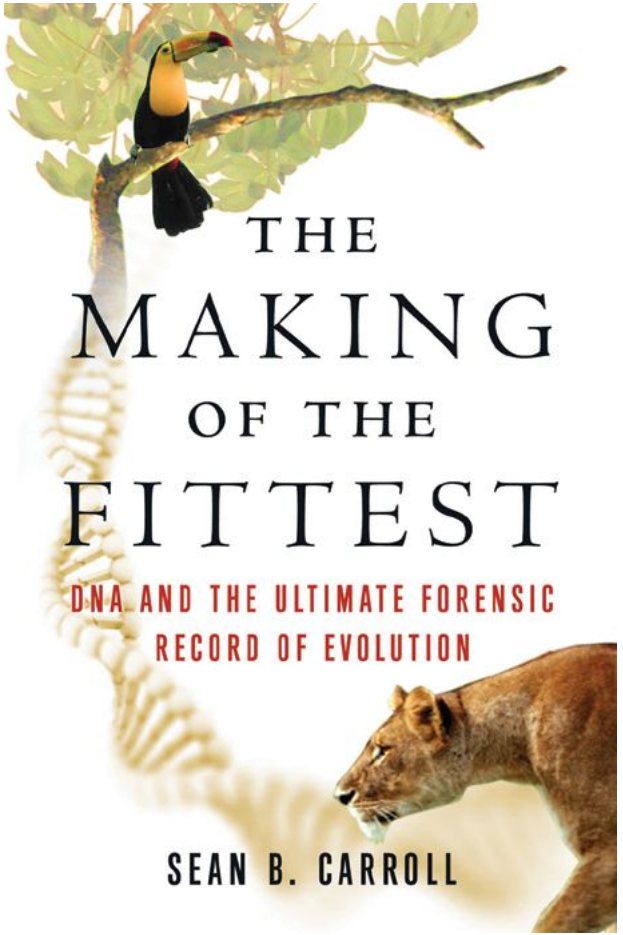
- *Mycobacterium tuberculosis*
  - Extracellulaire
  - 4.189 gènes codant pour des protéines
- *Mycobacterium leprae*
  - Intra-cellulaire
  - 1.605 gènes codant pour des protéines
  - A perdu (par délétions) ~1000 gènes présents chez *M.tuberculosis*
  - On trouve également plus de 1.000 gènes fossiles.

La fossilisation est une voie de non-retour

Dès que la pression sélective est relâchée, la probabilité de fossilisation d'un gène est très élevée.

Par contre, la probabilité que de revenir à une copie fonctionnelle à partir d'un gène fossilisé est quasiment nulle.

Ces probabilités sont discutées par S. Carroll.



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

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

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

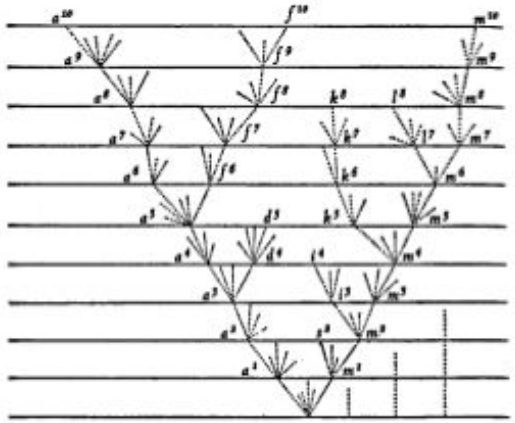


Fig. 1. Part of the only figure in the *Origin of Species*. Darwin first uses it to represent the divergence of variants within a species, showing successively more difference in a single lineage (a^1 through a^{10}) and splitting into multiple lineages (*m*, *s*, *i*, and so forth), some of which will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

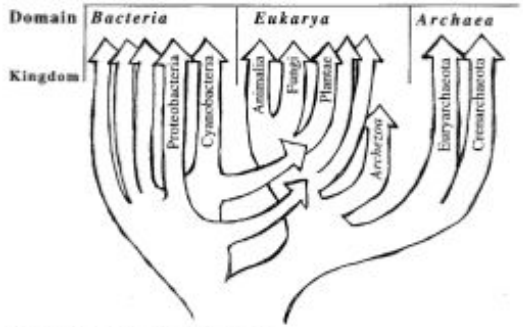


Fig. 2. The current consensus or standard model. Only a few of the "kingdoms" of the "domain" Bacteria are shown. Branching orders of several kingdoms within Bacteria and Eukarya remain in dispute. Mitochondrial and chloroplast endosymbioses are indicated by lower and upper diagonal arrows, respectively.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

Fig 1: arbre de la vie de Darwin (Fig 1)

Fiig 2-3: revisité par Doolittle (1999)

- Fig 2: événements d'**endosymbiose** liés à l'apparition des organelles des eucaryotes (mitochondrie et chloroplaste).
- Fig 3: **transferts horizontaux** entre génomes de procaryotes.

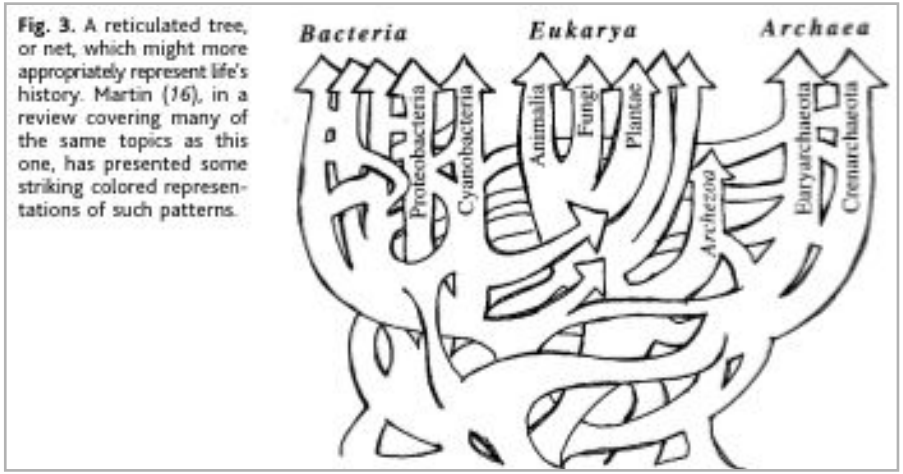


Fig. 3. A reticulated tree, or net, which might more appropriately represent life's history. Martin (*16*), in a review covering many of the same topics as this one, has presented some striking colored representations of such patterns.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. Science 284, 2124-9. doi.org/10.1126/science.284.5423.2124

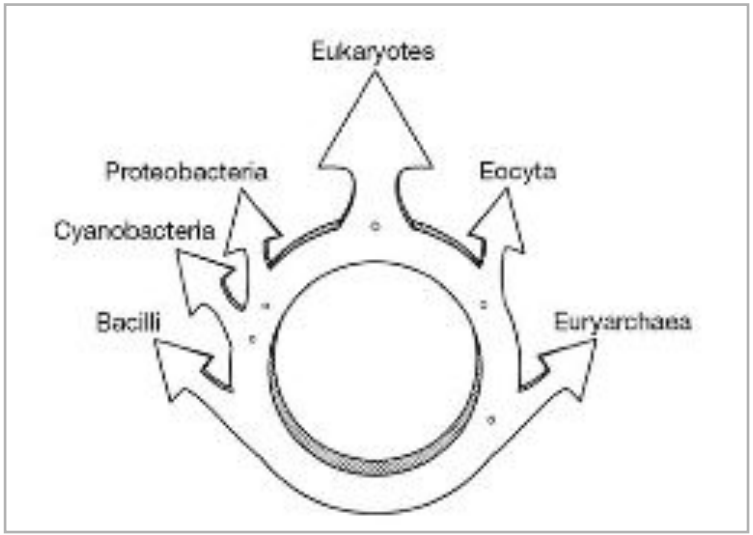
L’anneau de la vie

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]

[1]Molecular Biology Institute, MCD Biology, [2]Human Genetics, [3]IGPP, and [4]Astrobiology 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.



- Relations entre tous les gènes d’eucaryotes, d’eubactéries, et d’archées.
- → génomes eucaryotes résulteraient d’une **fusion entre bactérie et archée**.
- Gènes hérités des **archées** : majoritairement **maintien de la cellule** (réplication, transcription, régulation).
- Gènes hérités des **bactéries** : majoritairement **métabolisme**.

Rivera, M. C. and Lake, J. A. (2004). The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431, 152-5. doi.org/10.1038/nature02848

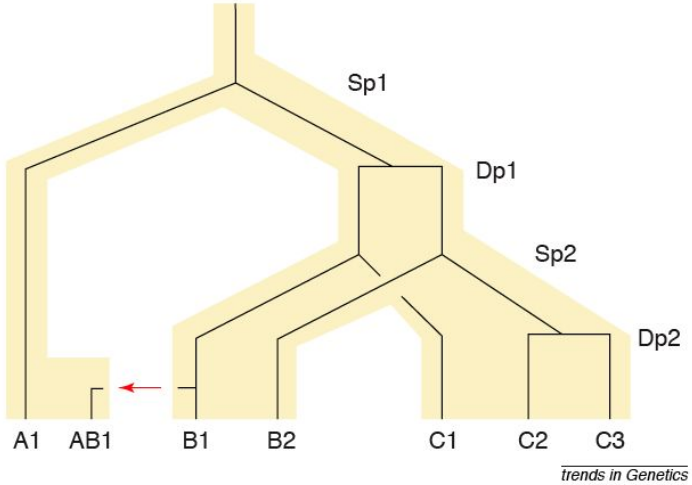
Solution de l’exercice orthologies / paralogies / xénologies (Fitch)

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

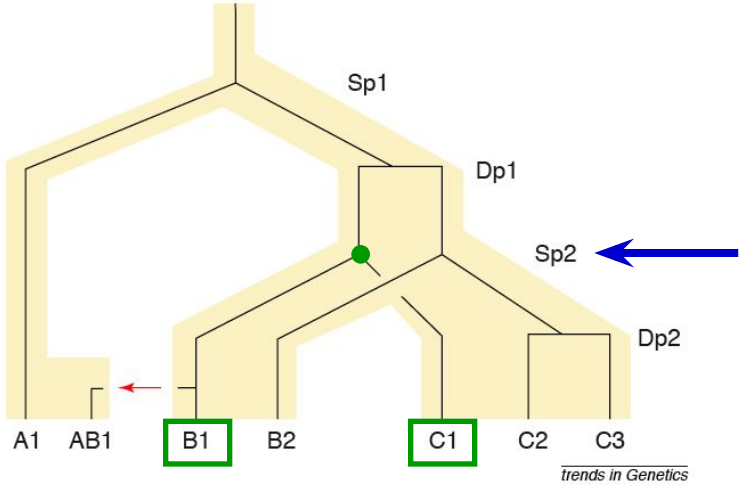
Exercice: types d'homologie

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

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

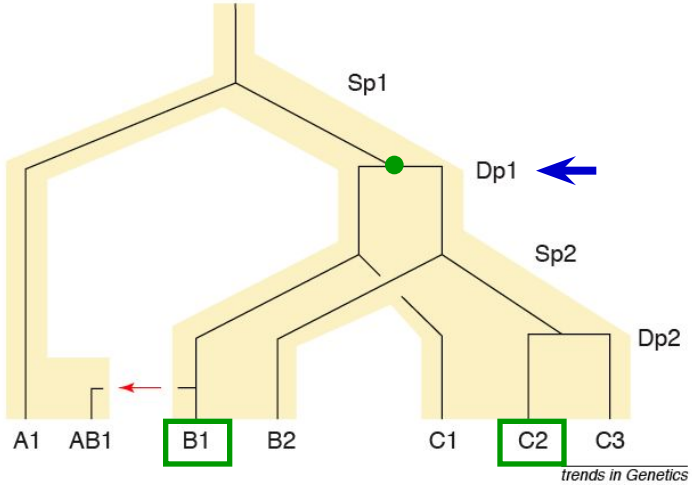
Exercice: types d'homologie

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

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

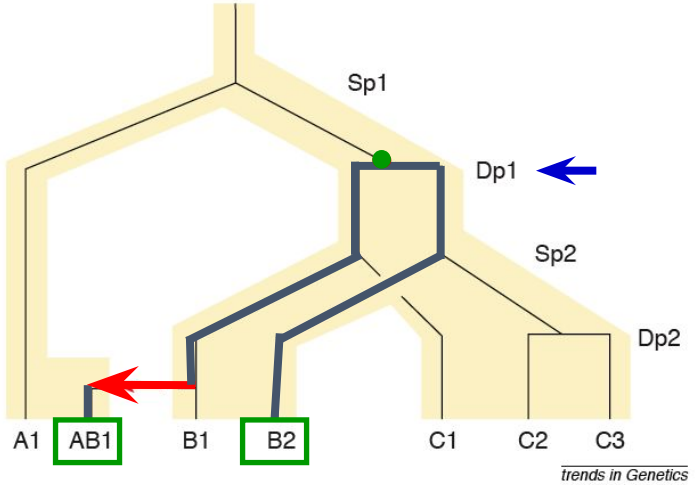
Exercice: types d'homologie

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							

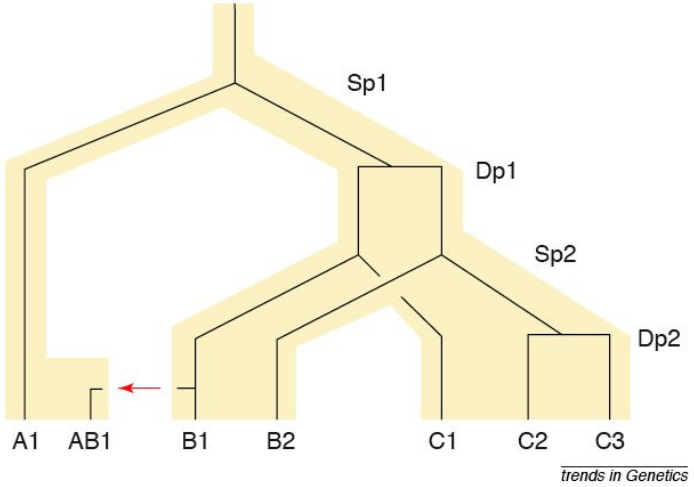


- Paire d'**orthologues:** paire de gènes dont le dernier ancêtre commun précède immédiatement un événement de spéciation (ex: a_1 and a_2). *Source: Zvelebil & Baum, 2000.*
- Paire de **paralogues**: paire de gènes dont le dernier ancêtre commun précède immédiatement une duplication génique (ex: b_2 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

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

Solution de l'exercice



	A1	AB1	B1	B2	C1	C2	C3
A1		O+X	O	O	O	O	O
AB1	0+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.

- S'applique dès qu'un transfert horizontal s'est produit depuis le dernier ancêtre commun entre deux molécules → toutes les relations entre AB1 et une des autres séquences.
- Qualification à un autre niveau que les statuts orthologie / paralogie, qui dépendent du type d'événement évolutif (spéciation ou duplication) qui a immédiatement succédé au dernier ancêtre commun.

→ On peut considérer que les relations impliquant AB1 sont *à la fois* de xénologie et d'orthologie / 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.*

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

Matériel supplémentaire (anciennes diapo)

Ne fait pas partie de la matière d'examen

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

Opsines sensibles au rouge et au vert

Opsines sensibles au bleu

Groupe extérieur: 2 monotrèmes (ornithorynque et echidné)

Mammalian opsins - percent identify matrix generated by clustalx

	1 LW_Felis_catus	2 LW_Equus_caballus	3 LW_Phoca_vitulina	4 LW_Pagophilus_groenlandicus	5 LW_Canis_familiaris	6 LW_Bos_taurus	7 LW_Capra_hircus	8 MW_Sus_scrofa	9 MW_Odocoileus_virginianus	10 LW_Globicephala_melas	11 LW_Tursiops_truncatus	12 LW_Phocoenoides_phocoena	13 LW_Delphinus_delphis	14 MW_Mus_musculus	15 MW_Rattus_norvegicus	16 MW_Cavia_porcellus	17 MW_Oryctolagus_cuniculus	18 MW_Sciurus_carolinensis	19 MW_Homo_sapiens	20 LW_Homo_sapiens	21 LW_Callithrix_jacchus	22 LW_Tarsipes_rostratus	23 LW_Cercartetus_concinnus	24 MW_Macropus_eugenii	25 LW_Myrmecobius_fasciatus	26 LW_Isoodon_obesulus	27 LW_Didelphis_aurita	28 LW_Ornithorhynchus_anatinus	29 LW_Tachyglossus_aculeatus	30 SW_mouse	31 SW_Rattus_norvegicus	32 SW_Myotis_laniger	33 SW_Myotis_ricketti	34 SW_Bos_taurus	35 SW_human	36 SW_Pan_paniscus	37 SW_Pan_troglodytes	38 SW_Daubentonia_madagascariensis	39 SW_Saimiri_bolivensis	40 SW_Macropus_eugenii	41 SW_Setonix_brachyurus	42 SW_Isoodon_obesulus	43 SW_Tarsipes_rostratus	44 SW_Sminthopsis_crassicaudata	45 SW_Didelphis_aurita	46 SW_Thylamys_elegans	47 SW_Ornithorhynchus_anatinus	48 SW_Tachyglossus_aculeatus
1 LW_Felis_catus	100	97	92	92	95	93	93	93	95	90	90	89	90	88	89	87	91	90	90	91	91	83	84	84	86	85	84	86	85	41	42	42	42	42	42	42	42	41	42	42	43	43	42	44	41	42	41	40
2 LW_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	89	88	86	89	88	41	42	43	44	42	43	43	43	42	43	43	44	43	44	45	42	43	42	42
3 LW_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	43	43	43	43	43	42	43	43	42	42	41
4 LW_Pagophilus_groenlandicus	92	95	99	100	93	92	92	91	92	88	88	88	88	89	90	88	89	91	90	91	89	81	81	82	83	83	82	82	82	40	41	42	42	42	41	41	41	41	42	42	43	43	43	43	41	42	41	41
5 LW_Canis_familiaris	95	96	94	93	100	95	95	94	95	90	90	89	90	89	90	90	93	92	90	89	90	83	83	83	85	84	83	84	84	41	43	43	43	42	42	42	42	42	42	42	43	43	43	44	42	43	41	41
6 LW_Bos_taurus	93	96	92	92	95	100	99	97	97	93	94	93	93	88	89	87	90	90	89	89	91	81	82	81	83	83	82	83	82	41	43	42	43	42	42	42	42	42	42	43	43	43	43	44	41	43	41	40
7 LW_Capra_hircus	93	96	92	92	95	99	100	97	97	93	94	93	93	88	89	87	90	90	89	89	90	81	82	81	83	83	82	82	82	41	42	42	43	42	42	42	42	42	42	43	43	43	43	44	41	43	41	40
8 MW_Sus_scrofa	93	96	91	91	94	97	97	100	97	94	94	93	93	88	89	87	90	90	89	90	90	80	81	80	82	82	81	82	82	41	42	41	42	42	41	41	41	42	42	42	43	43	43	44	41	42	40	40
9 MW_Odocoileus_virginianus	95	96	92	92	95	97	97	97	100	93	93	93	93	89	89	89	91	89	91	89	92	84	85	85	87	86	85	86	85	41	42	42	44	42	42	42	42	42	42	42	43	43	43	45	42	42	42	42
10 LW_Globicephala_melas	90	92	88	88	90	93	93	94	93	100	99	98	99	86	87	86	88	88	88	88	89	80	80	80	81	81	81	81	80	41	41	40	42	42	41	41	41	41	42	42	43	42	43	44	42	42	40	39
11 LW_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	42	41	41	41	41	42	42	43	43	43	44	41	42	40	39
12 LW_Phocoenoides_phocoena	89	92	88	88	89	93	93	93	93	98	99	100	98	85	86	85	88	87	87	87	88	78	79	78	80	80	80	80	79	40	41	40	41	41	41	41	41	41	42	42	42	42	42	43	41	42	40	39
13 LW_Delphinus_delphis	90	92	88	88	90	93	93	93	93	99	99	98	100	85	86	85	88	87	87	87	88	79	79	79	80	80	80	80	80	40	41	39	41	41	41	41	41	41	41	42	42	42	43	44	40	41	40	39
14 MW_Mus_musculus	88	91	89	89	89	88	88	88	89	86	86	85	85	100	98	91	90	90	88	88	89	82	83	82	84	83	83	83	82	41	43	42	43	43	43	43	43	43	44	43	44	44	44	44	42	43	43	42
15 MW_Rattus_norvegicus	89	92	90	90	90	89	89	89	89	87	87	86	86	98	100	92	91	91	90	89	90	83	83	83	85	84	84	84	83	41	43	42	43	43	43	43	43	43	44	43	44	44	44	44	42	43	43	42
16 MW_Cavia_porcellus	87	90	88	88	90	87	87	87	89	86	86	85	85	91	92	100	91	91	89	87	88	82	83	83	84	83	83	83	83	41	43	43	43	43	42	42	42	42	43	43	43	43	44	45	42	43	42	41
17 MW_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	41	42	42	43	43	44	41	42	41	41	41
18 MW_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	82	82	82	82	41	43	43	42	42	42	42	42	42	43	43	43	44	45	42	43	41	40	40
19 MW_Homo_sapiens	90	93	90	90	90	89	89	89	91	88	88	87	87	88	90	89	88	90	100	96	96	82	82	83	85	84	84	85	83	42	43	43	44	44	43	43	43	43	44	44	44	44	45	43	44	43	42	42
20 LW_Homo_sapiens	91	91	91	91	89	89	89	90	89	88	88	87	87	88	89	87	88	91	96	100	96	82	82	82	84	83	83	85	83	41	42	41	42	42	42	42	42	41	42	42	43	43	43	44	42	43	42	42
21 MW_Callithrix_jacchus	91	92	90	89	90	91	90	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	43	42	42	42	42	43	43	43	43	45	42	43	42	41	42
22 LW_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	43	43	44	43	43	42	44	42	41	41
23 LW_Cercartetus_concinnus	84	86	82	81	83	82	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	43	43	42	42	42	42	43	43	44	43	43	44	42	41	41	41
24 MW_Macropus_eugenii	84	87	82	82	83	81	81	80	85	80	79	78	79	82	83	83	80	80	83	82	83	96	96	100	97	95	94	90	92	42	42	43	42	42	42	42	42	42	43	43	44	43	44	43	42	44	42	42
25 LW_Myrmecobius_fasciatus	86	89	84	83	85	83	83	82	87	81	81	80	80	84	85	84	82	82	85	84	85	95	95	97	100	97	94	91	92	42	42	43	42	42	42	42	42	42	43	43	44	43	43	43	42	44	41	40
26 LW_Isoodon_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	42	42	43	43	42	43	42	41	43	40
27 LW_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	42	42	43	43	43	44	43	42	43	41
28 LW_Ornithorhynchus_anatinus	86	89	83	82	84	83	82	82	86	81	81	80	80	83	84	83	81	82	85	85	86	90	89	90	91	91	90	100	93	40	41	42	42	42	41	41	41	41	41	42	43	43	44	43	43	42	43	41
29 LW_Tachyglossus_aculeatus	85	88	83	82	84	82	82	82	85	80	80	79	80	82	83	83	81	82	83	83	84	92	91	92	92	92	91	91	93	100	39	40	42	41	41	41	41	41	41	42	43	43	43	42	41	42	40	39
30 SW_mouse	41	41	41	40	41	41	41	41	41	41	41	40	40	41	41	41	41	42	42	41	41	42	42	42	42	42	42	40	39	100	96	92	90	87	86	86	86	88	85	86	85	87	85	86	85	87	51	51
31 SW_Rattus_norvegicus	42	42	42	41	43	43	42	42	42	42	41	41	41	43	43	43	42	43	43	42	43	43	42	42	42	42	42	41	40	96	100	91	91	87	87	87	87	88	86	86	85	88	86	85	85	87	51	50
32 SW_Myotis_laniger	42	43	43	42	43	43	42	42	41	42	40	40	40	39	42	42	43	42	42	43	42	41	42	43	43	43	43	42	42	42	92	91	100	99	92	90	90	90	93	87	89	88	91	91	90	88	91	51
33 SW_Myotis_ricketti	42	44	42	42	43	43	43	42	44	42	42	41	41	41	43	43	43	42	42	44	42	43	42	43	42	42	42	42	41	90	91	99	100	91	89	89	89	91	88	88	88	89	89	88	86	90	49	49
34 SW_Bos_taurus	42	42	42	42	42	42	42	42	42	42	42	41	41	43	43	43	42	42	44	42	43	42	43	42	42	42	42	42	41	87	87	92	91	100	86	86	86	87	86	86	86	86	86	86	83	86	49	48
35 SW_human	42	43	42	41	42	42	42	41	42	41	41	41	41	43	43	42	42	42	43	42	42	42	42	42	42	42	41	41	41	86	87	90	89	86	100	100	100	91	92	83	83	85	83	85	82	86	52	51
36 SW_Pan_paniscus	42	43	42	41	42	42	42	41	42	41	41	41	41	43	43	42	42	42	43	42	42	42	42	42	42	42	41	41	41	86	87	90	89	86	100	100	100	91	92	83	83	85	83	85	82	86	52	51
37 SW_Pan_troglodytes	42	43	42	41	42	42	42	41	42	41	41	41	41	43	43	42	42	42	43	42	42	42	42	42	42	42	41	41	41	86	87	90	89	86	100	100	100	91	92	83	83	85	83	85	82	86	52	51
38 SW_Daubentonia_madagascariensis	41	42	42	41	42	42	42	42	42	41	41	41	41	43	43	42	41	42	43	41	42	42	42	42	42	42	41	42	41	88	88	93	91	87	91	91	91	100	88	85	85	86	86	86	84	87	50	49
39 SW_Saimiri_bolivensis	42	43	43	42	42	42	42	42	42	42	42	42	41	44	44	43	42	42	44	42	43	43	43	43	43	42	42	42	41	85	86	87	88	86	92	92	92	88	100	84	83	84	84	84	82	84	52	51
40 SW_Macropus_eugenii	42	43	43	42	42	43	43	42	42	42	42	42	41	43	43	43	42	43	44	42	43	43	43	43	43	43	42	42	42	86	86	89	88	86	83	83	83	85	84	100	99	95	94	93	91	94	51	51
41 SW_Setonix_brachyurus	43	44	43	43	43	43	43	43	43	43	43	42	42	44	44	43	42	43	44	43	43	44	44	44	44	43	43	44	43	85	85	88	88	86	83	83	83	85	83	99	100	95	95	93	90	94	50	50
42 SW_Isoodon_obesulus	43	43	43	43	43	43	43	42	42	43	42	42	43	43	43	43	43	44	43	43	43	43	43	43	43	43	43	43	43	87	88	91	89	86	85	85	85	86	84	95	95	100	95	94	91	95	50	51
43 SW_Tarsipes_rostratus	42	44	43	43	43	43	43	43	43	43	43	42	42	44	44	44	43	43	44	43	43	43	43	44	44	43	43	44	43	85	86	91	89	86	83	83	83	86	84	94	95	95	100	94	90	93	51	51
44 SW_Sminthopsis_crassicaudata	44	45	43	43	44	44	44	44	45	44	44	43	43	44	44	45	44	45	45	44	45	43	43	43	43	42	43	43	42	86	85	90	88	86	85	85	85	86	84	93	93	94	94	100	90	92	51	51
45 SW_Didelphis_aurita	41	42	42	41	42	41	41	41	41	42	41	41	41	40	42	42	42	41	42	43	42	42	42	42	43	42	41	42	42	85	85	88	86	83	82	82	82	84	82	91	90	91	90	90	100	94	49	49
46 SW_Thylamys_elegans	42	43	43	42	43	43	43	42	42	42	42	42	41	43	43	43	42	43	44	43	43	44	44	44	44	43	43	43	42	87	87	91	90	86	86	86	86	87	84	94	94	95	93	92	94	100	50	50
47 SW_Ornithorhynchus_anatinus	41	42	42	41	41	41	41	40	42	40	40	40	40	43	43	42	41	41	43	42	42	42	42	42	41	40	41	41	40	51	51	51	49	49	52	52	52	50	52	51	50	50	51	51	49	50	100	96
48 SW_Tachyglossus_aculeatus	40	42	41	41	41	40	40	40	42	39	39	39	39	42	42	41	41	40	42	42	41	41	41	42	40	40	40	40	39	51	50	51	49	48	51	51	51	49	51	51	50	51	51	51	49	50	96	100

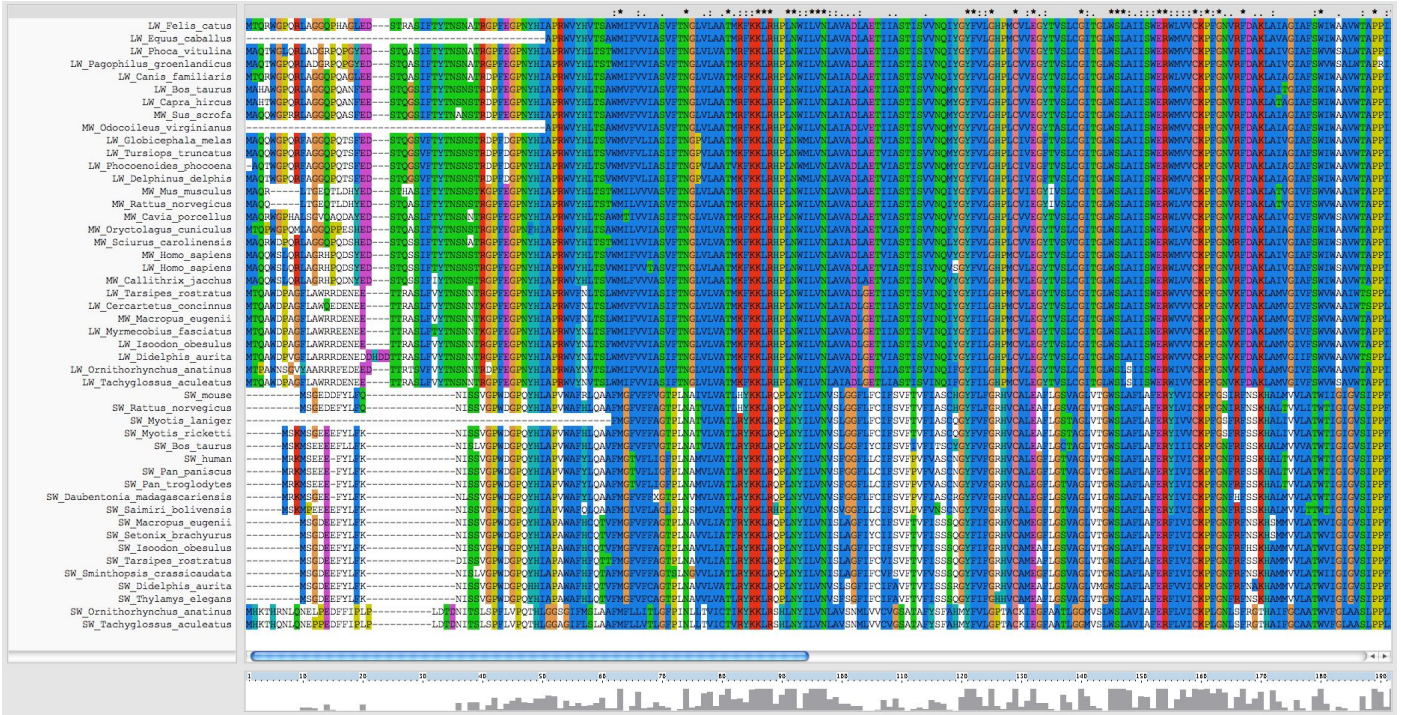
Alignement multiple des opsines de mammifère

Figure

- 50 opsines de différents mammifères (début d'alignement)

2 groupes évidents

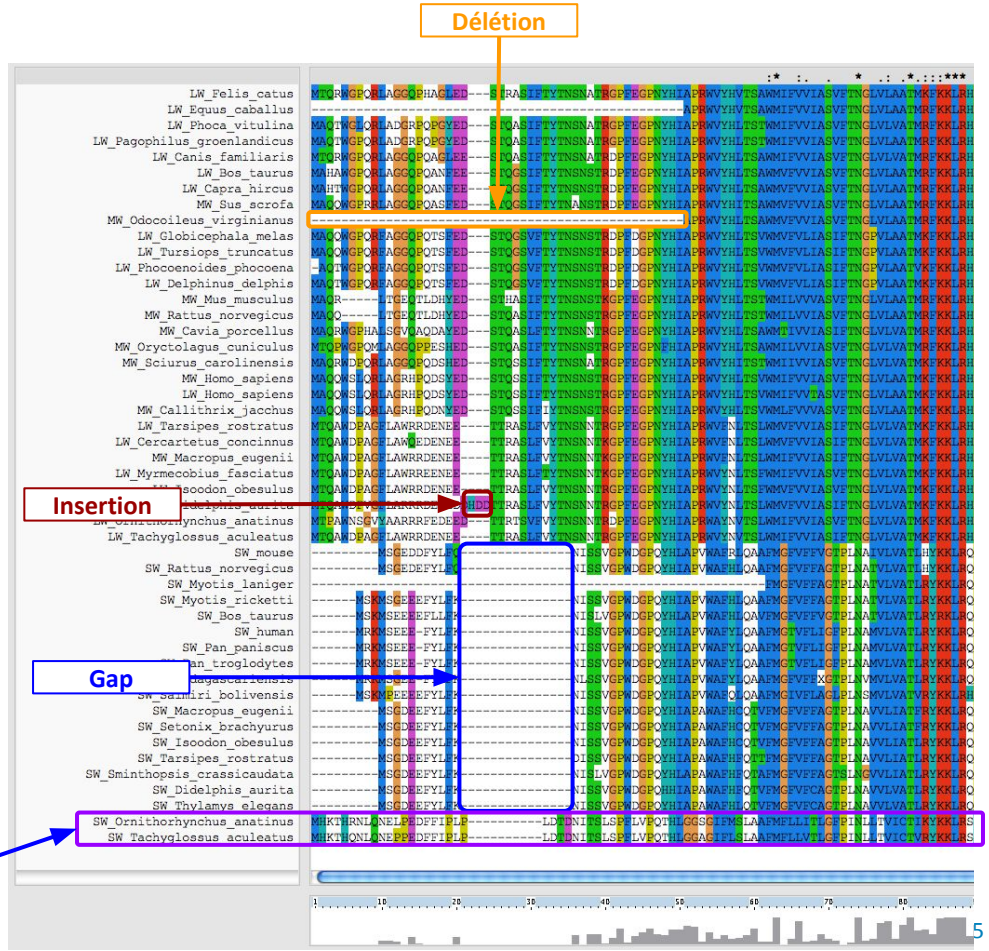
- Haut : opsines vertes (MWS) et rouges (LWS)
- Bas : opsines bleues (SWS)



Lecture d’un alignement multiple

- Une ligne par séquence
- **Groupe extérieur** : 2 monotrèmes (Ornithorhynchus et Tachyglossus)
- Une colonne par position de l’alignement multiple
- **Blocs conservés**
  - sur l’ensemble de l’alignement
  - sur un sous-ensemble des séquences
- **Gaps**
  - délétions : fragments de séquence absents d’une ou de quelques séquences mais présents ailleurs
  - insertions : fragment de séquence présents dans une ou quelques séquences mais absents ailleurs

Pour départager insertion ou délétion il faut généralement comparer l’alignement multiple à l’arbre des molécules (inférence phylogénétique)



Bootstrapping

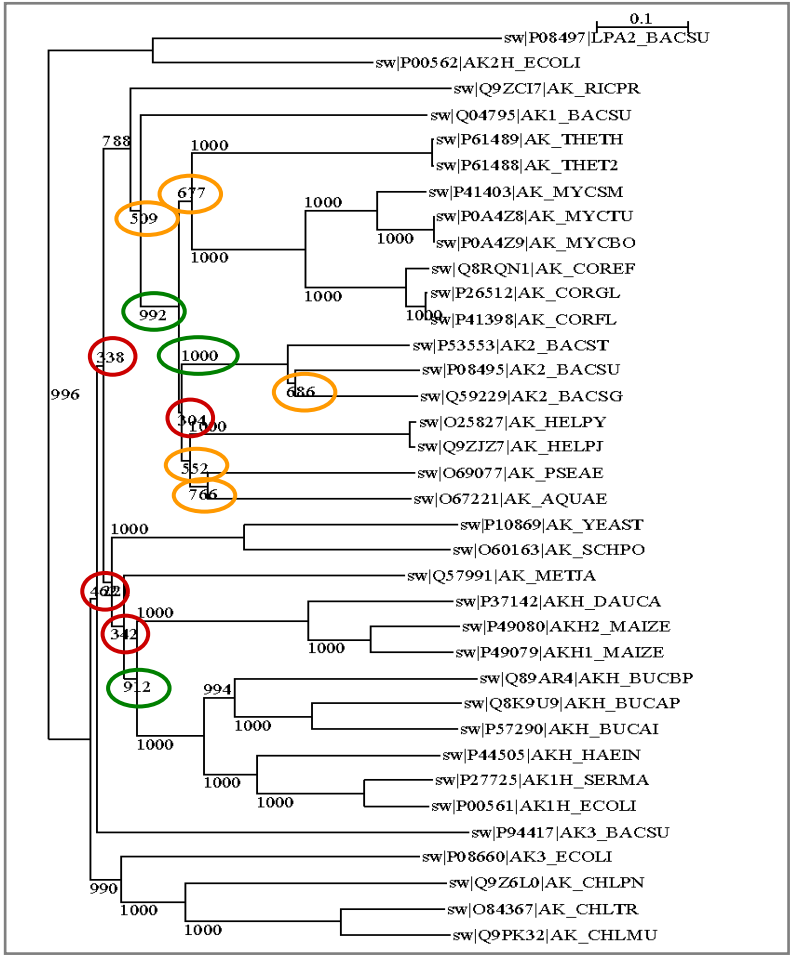
- **Valeur de bootstrap** assignée à chaque branchement
- Indique le nombre de fois où ce branchement se retrouve à l'identique dans les N arbres de bootstrap
  - **Valeur élevée** → **robuste aux fluctuations d' échantillonnage**
  - **Valeur faible** → sensible.

Attention ! La valeur de bootstrap

- **indique la robustesse par rapport aux données,**
- **ne garantit pas la fiabilité par rapport à l'histoire évolutive réelle ("arbre vrai")**

Problèmes potentiels

- Biais d'échantillonnage
- Données erronées



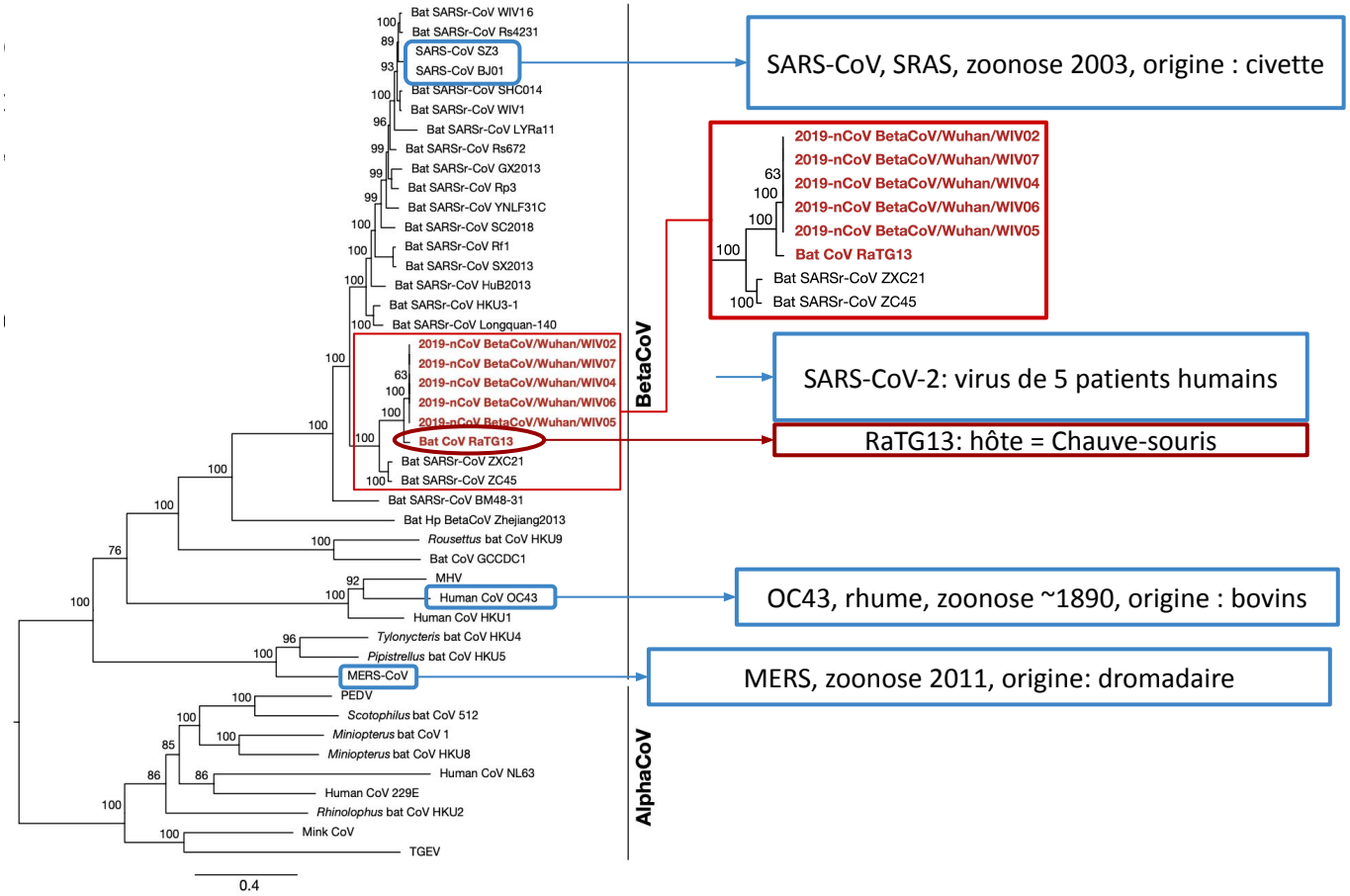
Retracer l’évolution de SARS-CoV-2 dans les génomes des coronavirus

Ne fait pas partie de la matière d’examen

La publication du génome de SARS-CoV-2

3 février 2020 : génomes complets publiés

- **SARS-CoV-2 (COVID-19)**
- **RaTG13:** virus de chauve-souris, le plus proche de SARS-CoV-2
- 96.2% d'identité sur l'ensemble du génome
  ± 40 à 70 ans de divergence
  → cousin éloigné plutôt qu'ascendant direct de SARS-CoV-2



Zhou P, Yang X-L, Wang X-G, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020; 579: 270–3. DOI: 10.1038/s41586-020-2012-7

Un virus synthétique avec des bouts de HIV ?

Le 17 avril 2020: Luc Montagnier (prix Nobel de médecine pour sa contribution à la découverte du HIV), affirme que SARS-CoV-2 résulte de manipulations : insertion de fragments de HIV dans un génome de coronavirus, et fuite accidentelle du labo.

Eléments cités par Mongagnier

- Publication de Jean-Claude Perez
- Préprint d'un groupe de chercheurs indiens "forcés à rétracter"

Selon Luc Montagnier, le virus covid19 est une manipulation humaine (17 avril 2020)
https://www.youtube.com/watch?v=qSWCLHIOiMo
(devenue inaccessible depuis lors)

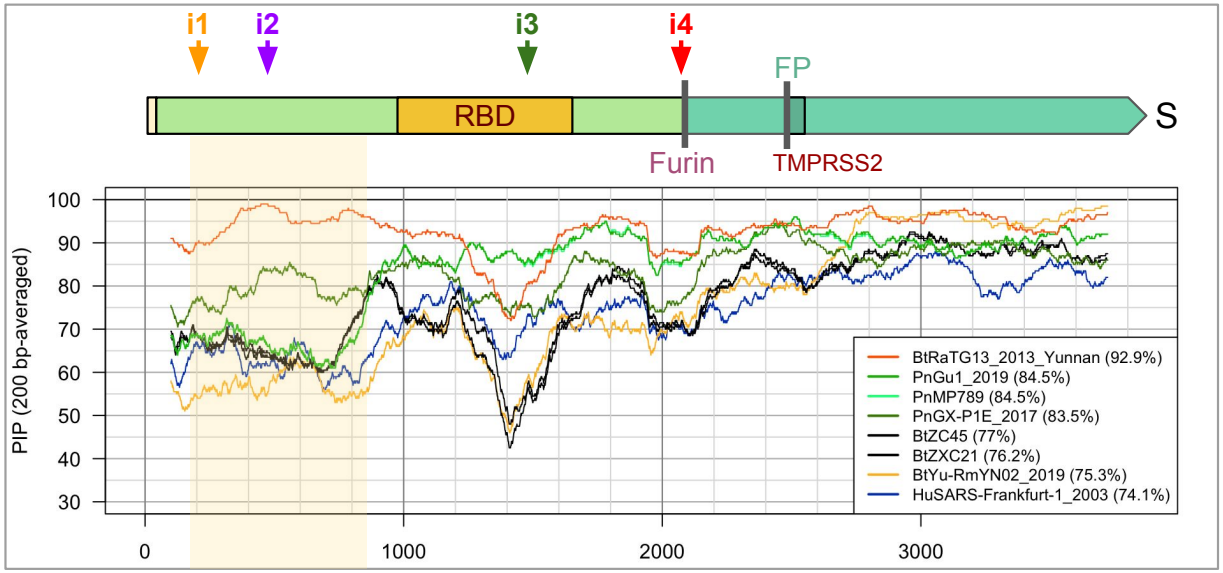
"Je suis arrivé à la conclusion qu'il y avait eu une manipulation de ce virus. [...] Il y a un modèle qui est évidemment le virus classique, et là c'était un modèle venant de la chauve-souris, et là, à ce modèle on a par-dessus ajouté les séquences du VIH, du SIDA. … Non, ce n'est pas naturel, c'était 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". "

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

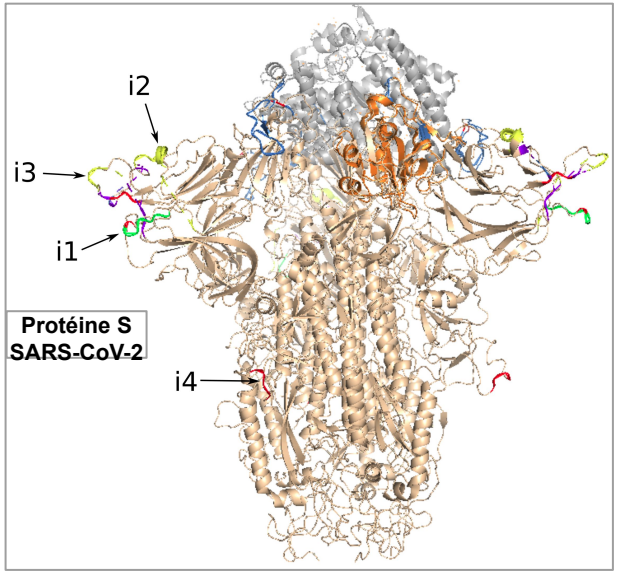
Quatre insertions dans le gène S de SARS-CoV-2

- Flèches: 4 insertions
- Note: i1, i2, i3 dans des régions “exposées” (boucles externes)

Structure et profil de conservation du gène S



Structure de la protéine spike



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

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

Haut: alignement par paire:

- Query: gène S
- Subject: génome complet de VIH (10kb)
- **Expect = 7.5 → pas du tout significatif**

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   AATGGTACTAAGAGGTTTGATAACCCTG  113
           ||||||||||| ||||| |||||| |||
Sbjct  967  AATGGTACTAAAAGGTTAGATAACACTG  994
```

Bas: contrôle négatif

- Query : séquence aléatoire
- Subject : génome complet de HIV
- Expect = 2.1 → meilleur alignement que gène S , mais toujours pas significatif (comme attendu, la séquence étant aléatoire !)

Conclusion: 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 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   CCTAAAAGTTCTTTGTAATAACTGTATTATT  381
             |||||||||||||||||||  | ||| |||
Sbjct  2523  CCTAAAAGTTCTTTGTAATATTTCTATAATT  2493
```

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

Des insertions bizarres?

Figure de Pradhan et al (2020), initialement sur bioRxiv, puis retirée par les auteurs suite à des critiques sur leur rigueur scientifique.

- alignement par paire, même si la légende indique “Multiple sequence alignment”)
- Insuffisant pour inférer si insertions ou délétion
- Insuffisant pour évaluer l’ancienneté de l’événement

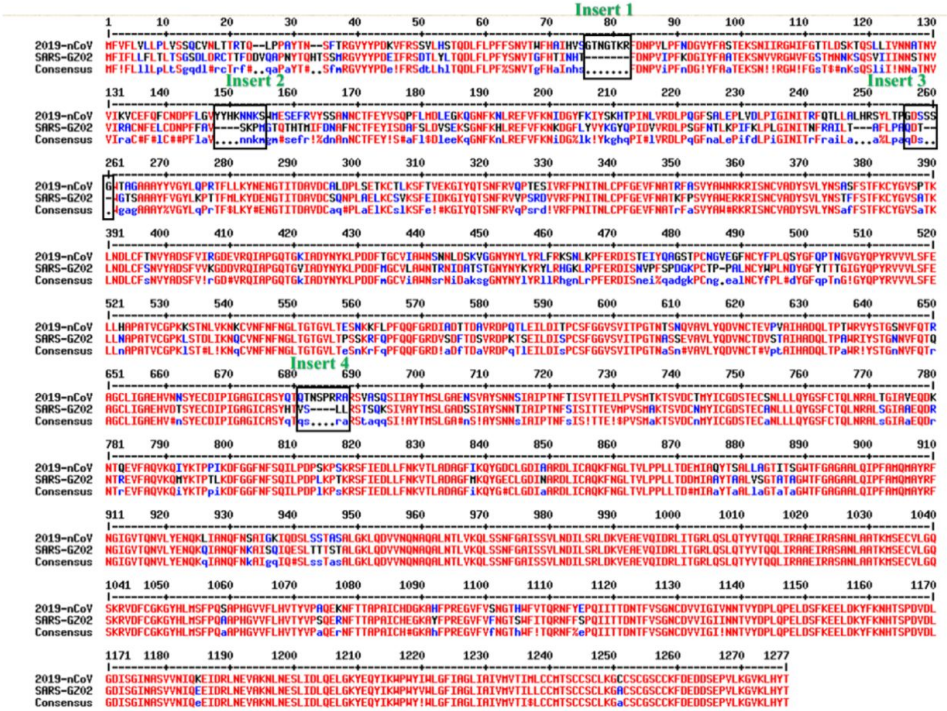
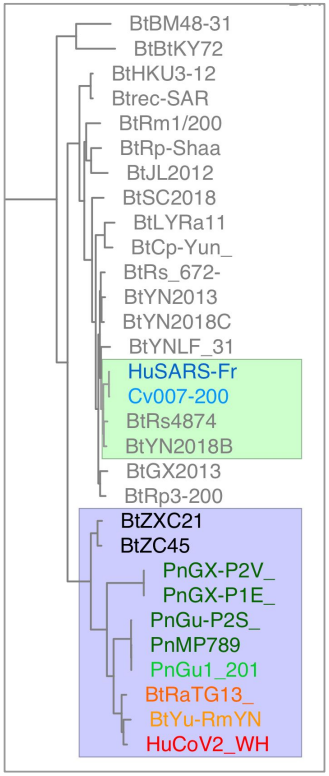


Figure 2: Multiple sequence alignment between spike proteins of 2019-nCoV and SARS. The sequences of spike proteins of 2019-nCoV (Wuhan-HU-1, Accession NC_045512) and of SARS CoV (GZ02, Accession AY390556) were aligned using MultiAlin software. The sites of difference are highlighted in boxes.

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

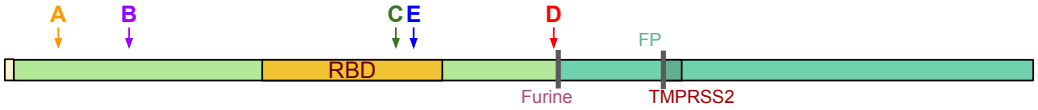
Insertion B : partagée entre plusieurs espèces de coronavirus

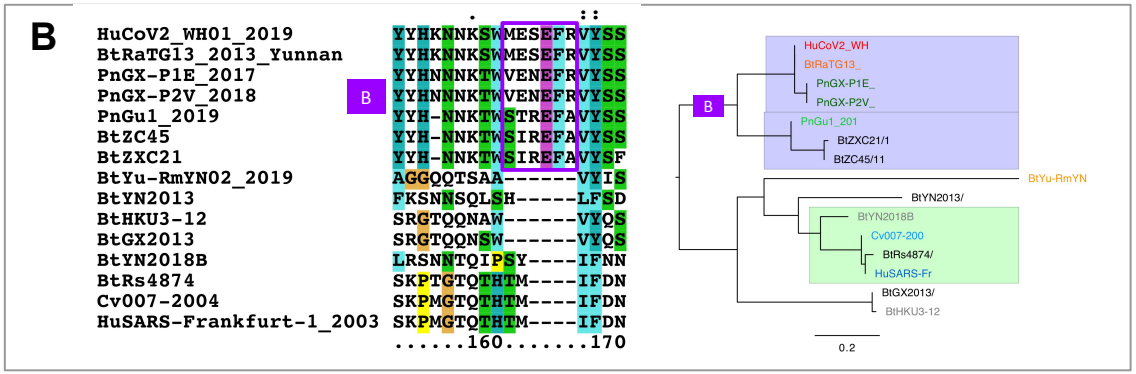
Arbre des génomes



- 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**

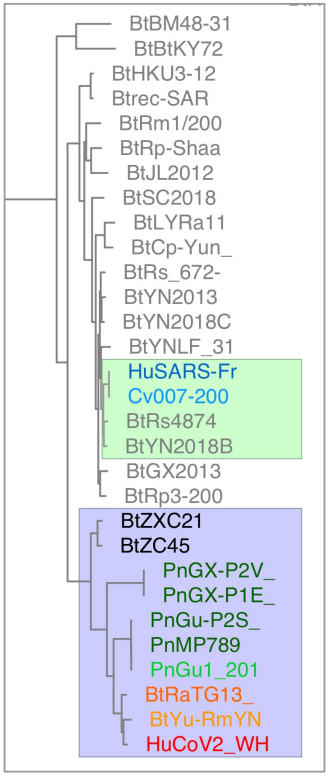




Sallard, E., Halloy, J., Casane, D., van Helden, J., Decroly, E. Retrouver les origines du SARS-COV-2 dans les phylogénies de coronavirus. Médecine/Sciences 36, Sept-Août 2020. https://doi.org/10.1051/medsci/2020123

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

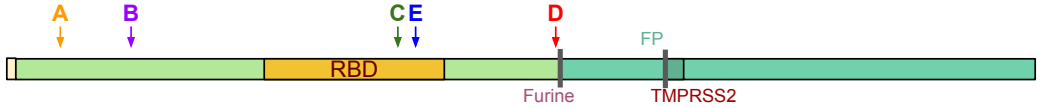
Arbre des génomes

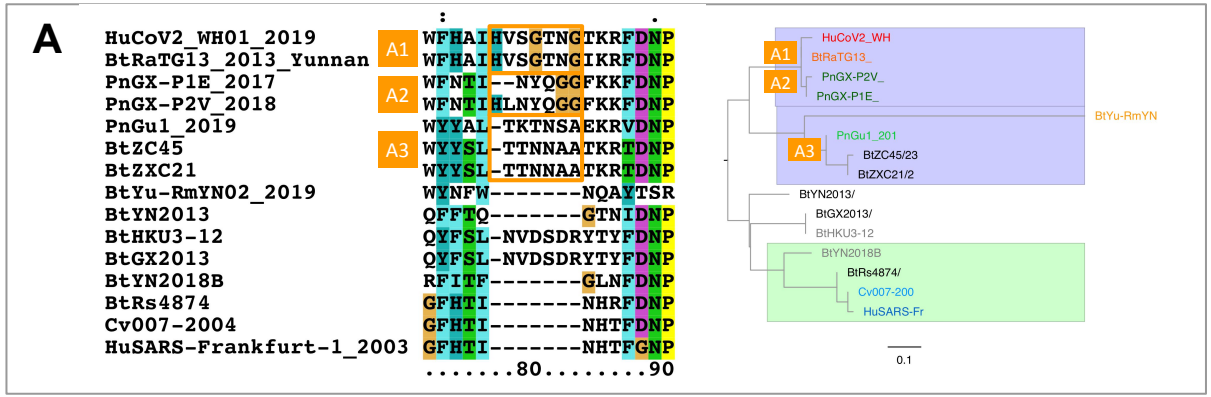


- A1 : Insertion identique entre SARS-CoV-2 et RaTG13, malgré 40 à 70 ans de divergence

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

- A2, A3 : Insertions au même site, dans des sous-groupes séparés → événements indépendants

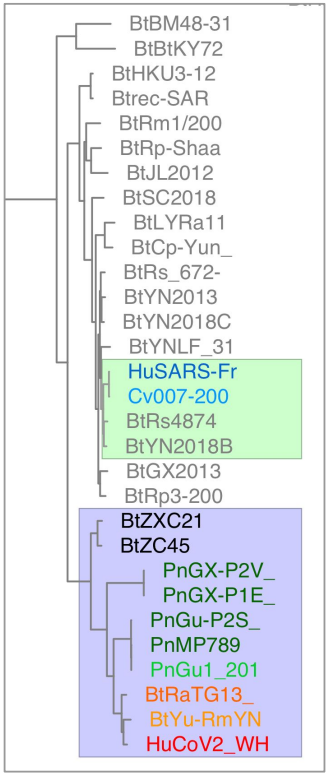




Sallard, E., Halloy, J., Casane, D., van Helden, J., Decroly, E. Retrouver les origines du SARS-COV-2 dans les phylogénies de coronavirus. Médecine/Sciences 36, Sept-Août 2020. https://doi.org/10.1051/medsci/2020123

Insertion C

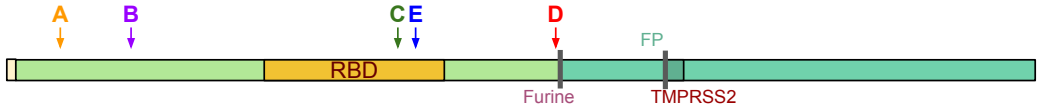
Arbre des génomes

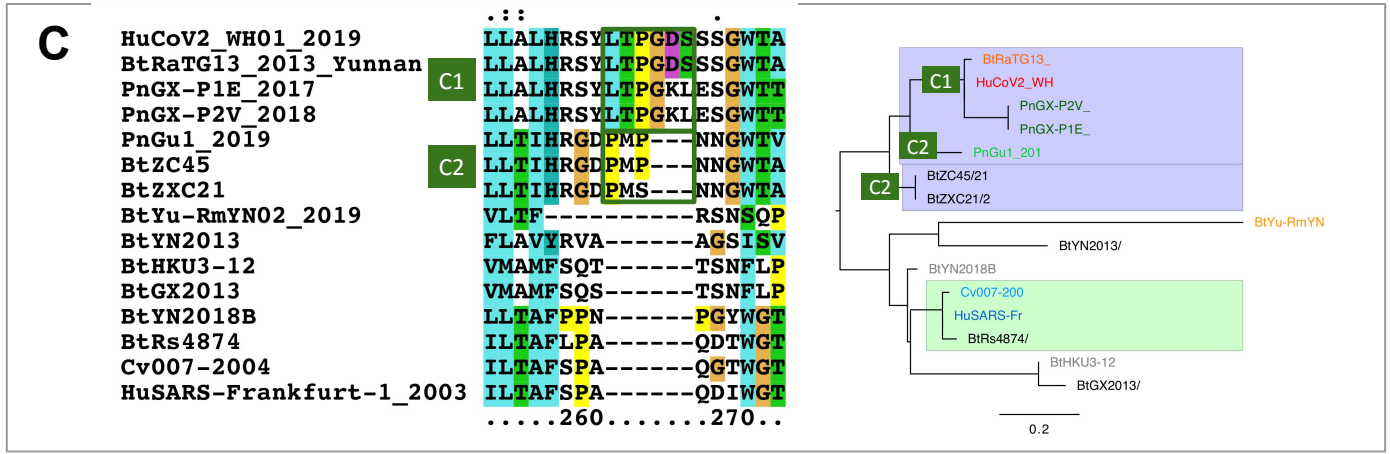


- C1 : insertion partagée entre 4 des souches de virus apparentés, parmi les souches analysées

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

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

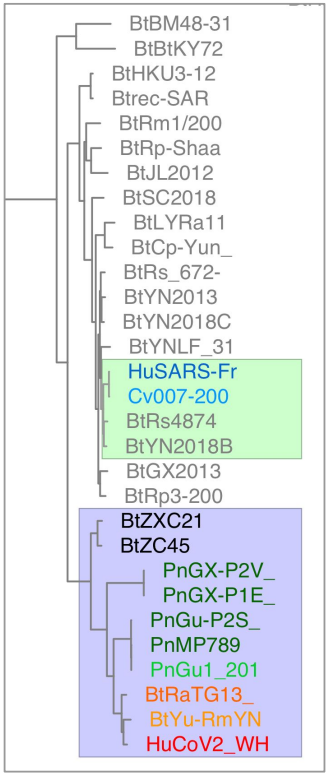




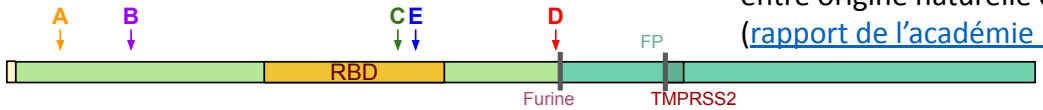
Sallard, E., Halloy, J., Casane, D., van Helden, J., Decroly, E. Retrouver les origines du SARS-COV-2 dans les phylogénies de coronavirus. Médecine/Sciences 36, Sept-Août 2020. https://doi.org/10.1051/medsci/2020123

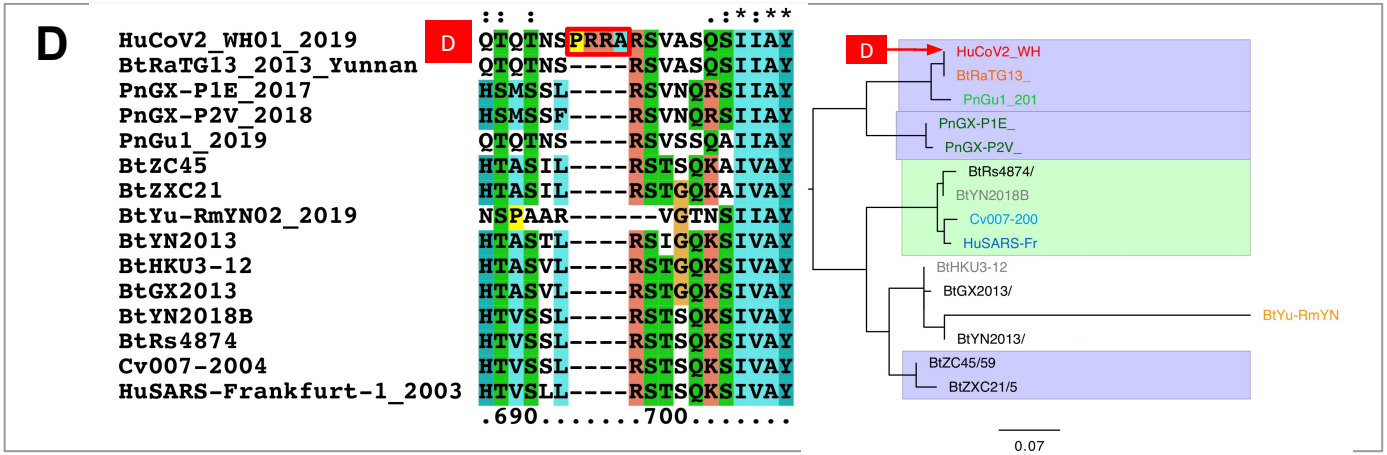
Insertion D

Arbre des génomes



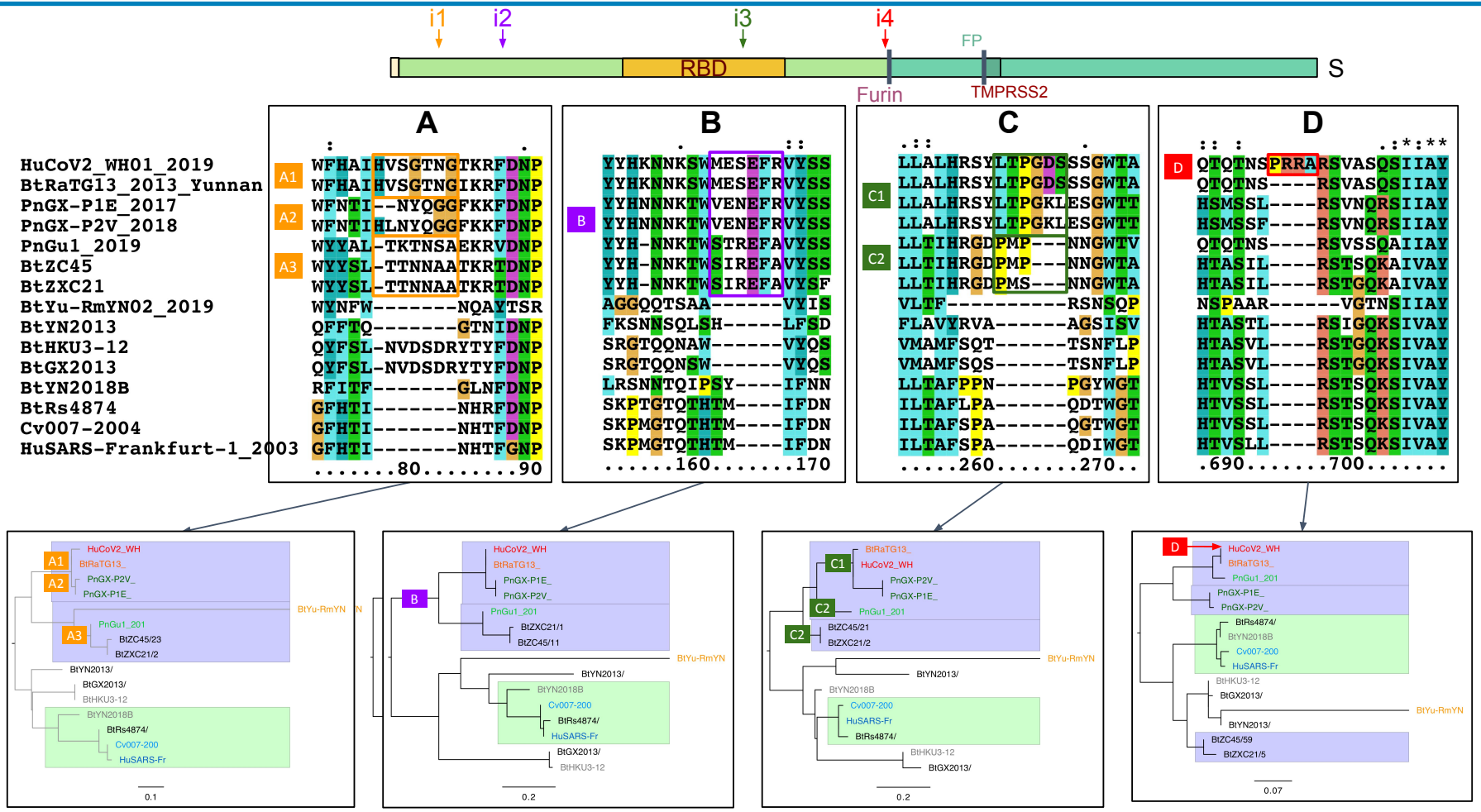
- D : insertion très courte, sans similarité significative avec HIV
- Insertion unique à SARS-CoV-2
- Crée un site furine, qui augmente la capacité de la protéine à se lier aux cellules humaines.
- Les données disponibles ne permettent pas de statuer entre origine naturelle ou artificielle de cette insertion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





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

Alignements multiples et arbres phylogénétiques des insertions de la protéine S



Points-clés : origine de SARS-CoV-2

- L'alignement multiple → blocs de séquences où un fragment est présent / absent.
- Comparaison avec arbre phylogénétique → inférence des événements évolutifs (insertion, délétion, substitution) et estimation de l'ancienneté
- L'hypothèse de Montagnier (insertion de séquences HIV dans un coronavirus) ne tient pas la route.
  - méconnaissance des méthodes bioinformatiques et des scores statistiques (E-valeur)
  - incompatible avec la présence de ces mêmes insertions dans plusieurs génomes
    - échantillons collectés à des dates et endroits indépendants
    - séquences publiées bien avant la pandémie.

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