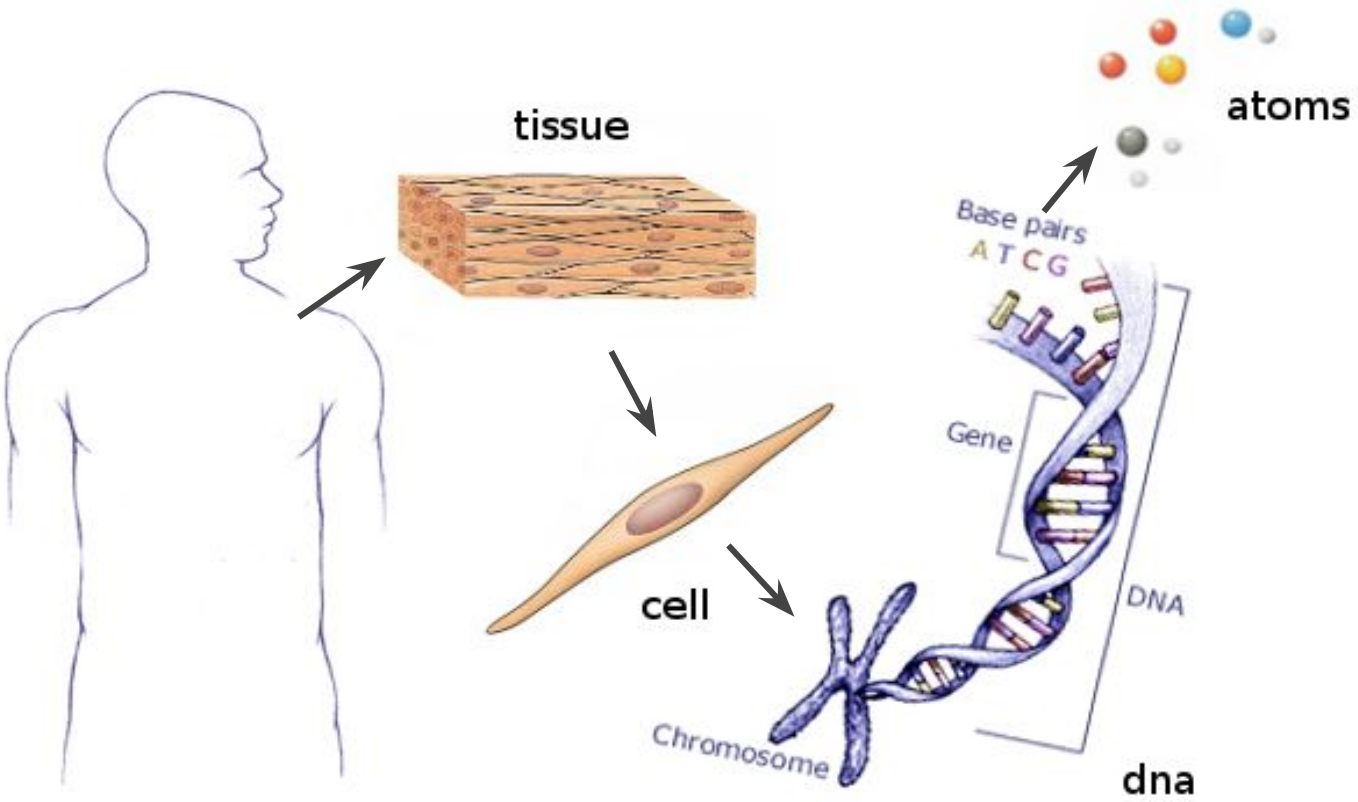
Étudier l’évolution des génomes avec la bioinfo

Emmanuel Noutahi
Étudiant au PhD en bioinfo
LBIT au DIRO, UdeM

Plan

1. Mécanismes d’évolution des génomes.
2. Bioinformatique ?
3. Exemple de problèmes étudiés au LBIT.

ADN et Génome



- ADN : très longue chaîne double brin constituée d'une suite de 4 bases complémentaires (A-T, G-C)
- presque l'entièreté de l'information génétique d'un individu est encodé dans l'ADN

ADN et Génome

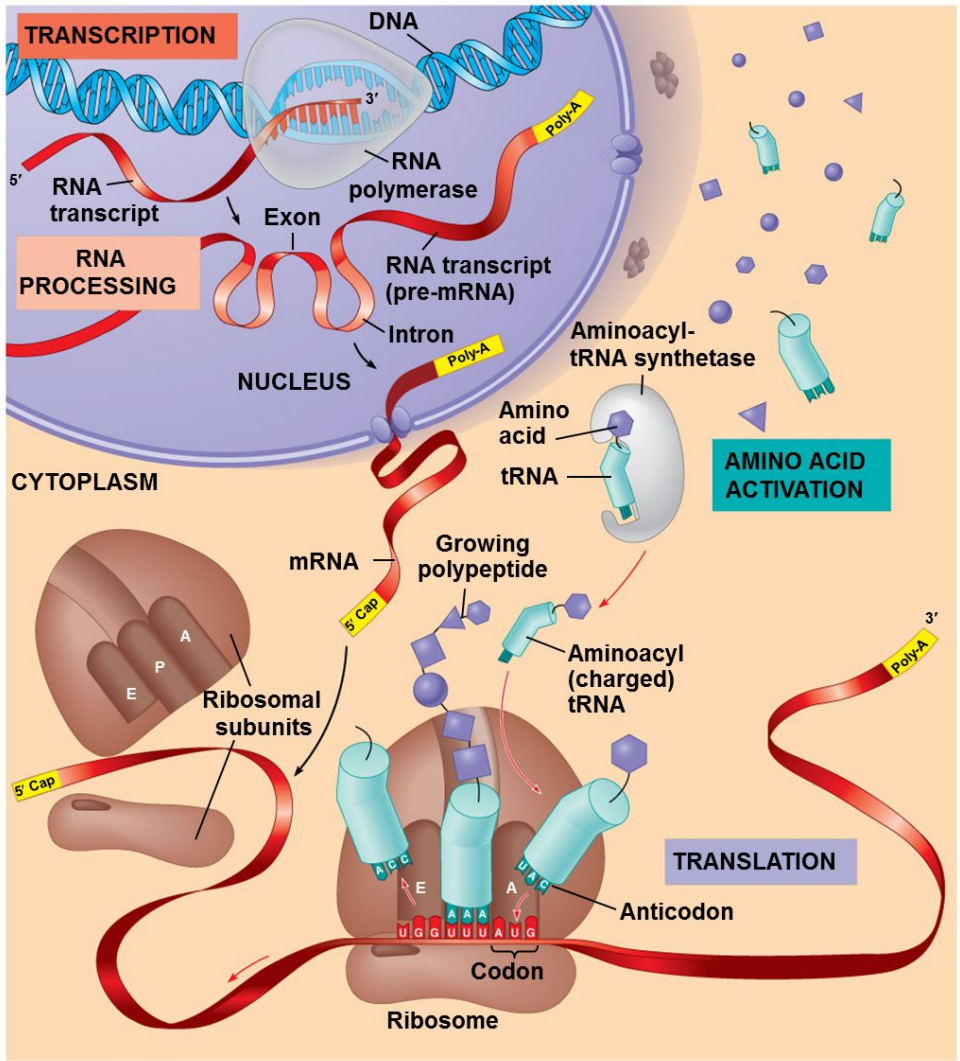
NCBI Reference Sequence: NT_086360.2

GenBank Graphics

```
>NT_086360.2 Homo sapiens chromosome 22 sequence, ENCODE region ENm004
CAAAAAAAATTAGTGGGGCATGGTGGCGGGCACCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGA
ATTACTTGAACCTGAGAGGCAGAGGTTGCAGTGAGCCGAGATCACACCACCGCACTCCAGCCTGGGTGAT
GGAGCGAGACTCCCCTTTTTTTCCTTCTTTTTTTTTTTTTTTTTTTTTTTGAGATAGAGTCTTGCTTCATT
GCCCAGGCTGGAGTGCAGTGGTGTGATCTCAGCTCACTGCAACCTCTGCCTCCTGGGTTCAAGCAATTCT
CCTGCCTCAGCCTCCTGAGGGAGCTGGGATTACAGGTGCCCGCCACTACGCCCCACTAATTTTTTTGTGT
ATTTTTAGTAGAGACAGGGTTTCACCATGTTGGCCAAGCTGGTCTTGATCTCCTGACCTCAGATGATCCG
CTCACCTTGGCCTCCCAAAGTACTGGGATTACAGGCTTGAGCCACCATGCCCAGCTGAGAATCCCTATGA
AAAAGCAAGGCCAGACACTGTGTCTCATGCCTATAATCCCAGCACTTTGGGAGACTGACGCGAGCAGATC
ACTTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACATGGTGAAAACTTGTCTCTACTAAAAATACATAA
AAAAAATTGGTGGGGCATGGTGGCGAGCACCTGTAATCCCAGCTGCTCAGGAGGCTGAGGTGCGAGAATC
GCTTGAACCTGGGAGGCGGACATTGCAATGAACCAAGATTGCACCACTGCACTTCAGCCTGGGCGATAGA
GCGAGACTCATTCTCAAAAAAAAAAAAAAAAAAAAAAAAAGATATTTTAAGTGATGTTTTAATAAGCTGCAT
CTGTTTTCAGTTCATTTATGTGTATAATTCAGTATTTGCCATACAGTAGACCCTTGGTATTTTTAGATTT
TATATTTGTTTTAACTGTTTTGACTGATTCTTAAGGGCAGTCAAATTAATAGTTTATAATTTTGCTGAGA
TGTAAATTTGAATCACATGTGCTGTGGGGTTGGTTTTTCAATCACATCATTAGGTACACTGAGTTCTGTG
AAGTTTTACCTTACTGCATTTCCTGGGCAGTTGAGCAGTCATTTTTATCTTTTTAAAATCTTTTAAAATT
TATTTATTTATTTTTATTATTTTTTGAGATGGAGTCTTACTCTGTCGCCCAGGCTGGCATAATCTCGGCT
CACTGCAACCTTCGACTGTCGGGTTCAAGCCATTCTTGTGCCTCAGCCTCCAGAGTAGCTGGGATTACAG
GTGTGTGCCACCATACCCGGCTAATTTTTGTATTTTTTGTACAGACGGGGTTTCACCATGTTGGCTAGGA
TGGTCTCGAACTCCTGAGCTCAAGCAATCTGCCTGCTTCGGCCTCTGGGAAAGTGCTGTGGAATTATAGG
TGTGAGTTACTGGGCCCAGCCTGAGCAGTCATTTTTTACTGGGGACATACATCAGAGTCACTTGTAAAGC
CTTTTGAAGTACAGATCCCTATCTGTTGGCCCTTCAAATCAGAATCTGCAGGGATGGGTTTCAGCTCTCT
ACATTTTAAAAGTTTTTCCAGGTGATTCGGTTGTGCCTGCTTAAGAACGGCTACAAAAATTTTTGTGAGT
TGGGGATTTGCAGAAATTGCAATATGTATCATATGTAAATATCAAGATTCAGCTGATACCTGCTGTAAGG
AATTTATATGGAGGCTCAGCCTTCTGAAAAGTGGGGGAACACTGAAAAATATTAATGTAGAATTGGCTGG
GCATGGTGGCTCACTCCTGTAATCCTAGTACTTTGGGAGGCTGAGGTGGGTGGATTGCCTGAGCTCAGGA
GTTCGAGACCAGCCTGGGCAACACGTCGAAACCCTGACTCTACTAAAATACAAAAAATTAGTCAGATATG
GTGGCATATGCCCTGTAGTCCCAGCTACTCTGGAGGCTGAGACAGGATAATTGCTAGAACCTGGGAGGCA
GAGGTTGCAGTGAGCCGAGATCACGCCACTGCATTCCAGCCTGGGTGGCAGAGTGAGACTCCGTCTTTTA
AAAACAAACAAACAAAAAATAATGTAGAATAATGGATTTTGGCTAGAAAACATGGGTTATAACCTATTAG
GACTATGACTTTGGACAAGTCACTTACCTGCTGTGAACTTCAGCTAGTATCCTTATCTCACAGGG ···
```

- ADN : très longue chaîne double brin constituée d'une suite de 4 bases complémentaires (A-T, G-C)
- presque l'entièreté de l'information génétique d'un individu est encodé dans l'ADN

De l'ADN aux protéines (unités fonctionnelles)



© 2011 Pearson Education, Inc.

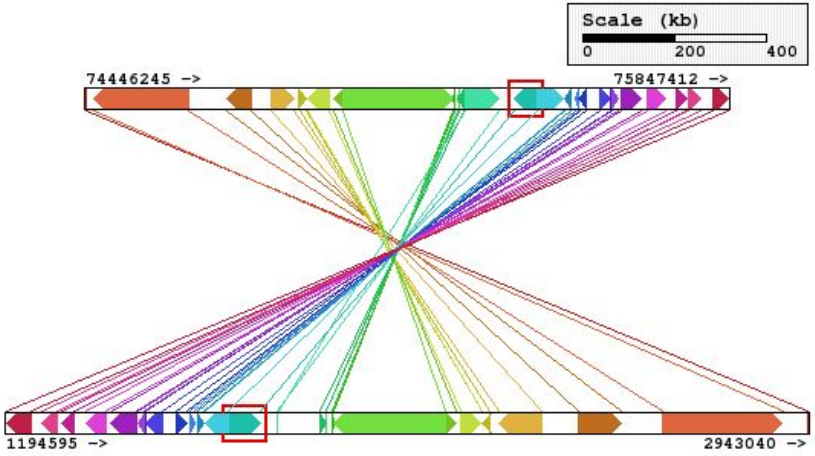
Second letter

First letter	U	C	A	G	Third letter
U	UUU Phe UUC UUA Leu UUG	UCU UCC UCA Ser UCG	UAU Tyr UAC UAA STOP UAG STOP	UGU Cys UGC UGA STOP UGG Trp	U C A G
C	CUU CUC CUA Leu CUG	CCU CCC CCA Pro CCG	CAU His CAC CAA Gln CAG	CGU CGC CGA Arg CGG	U C A G
A	AUU AUC Ile AUA AUG Met	ACU ACC ACA Thr ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G
G	GUU GUC GUA Val GUG	GCU GCC GCA Ala GCG	GAU Asp GAC GAA Glu GAG	GGU GGC GGA Gly GGG	U C A G

code génétique standard

Contenu, organisation et structure variable

- Structure et organisation variable entre espèces
  - **Circulaire et compact** (ex : Bactéries) vs **Linéaire** organisé en **chromosomes** (ex : Animaux)
  - **Pas de régions introniques** dans les gènes codants chez les **procaryotes**, etc …
  - Ordre différent

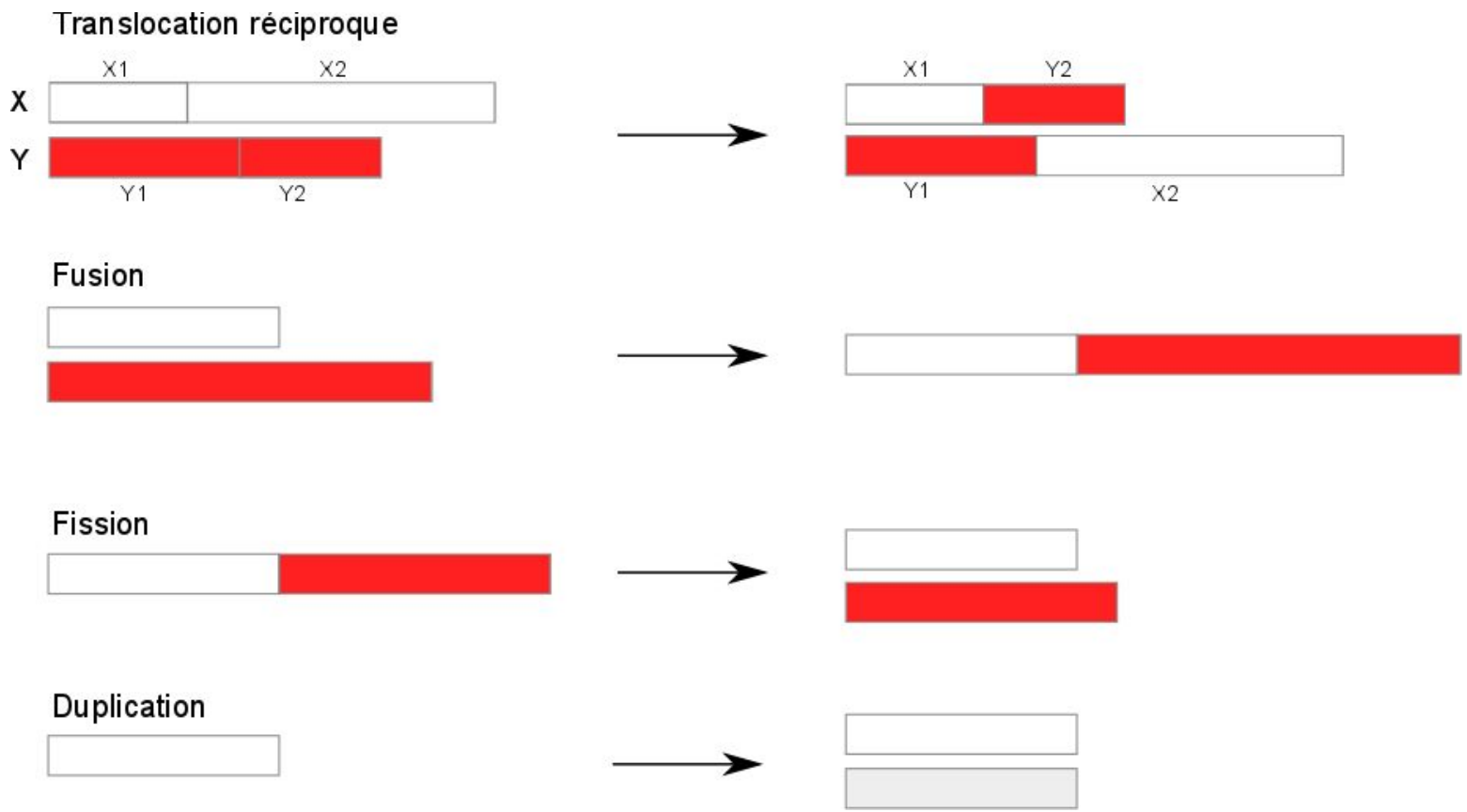


block synténique (RPA1), souris vs humain

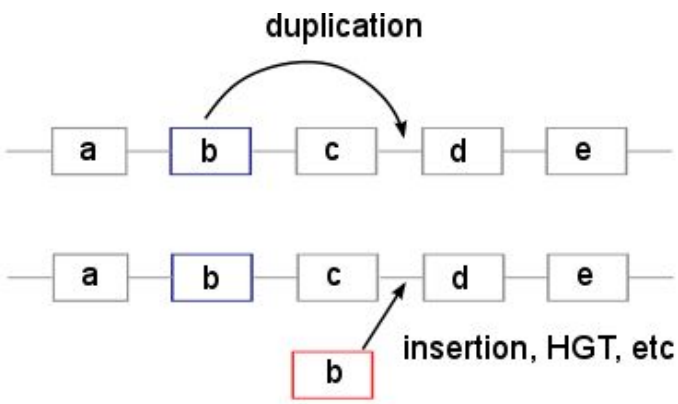
- Contenu
  - ~4,000 gènes chez *E. coli,* vs ~20,000 chez l'humain
  - **Différence inter et intra-espèces** dans la séquence et la distribution des gènes.

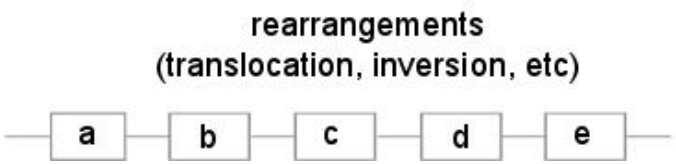
Les génomes évoluent … et parfois, s'adaptent.

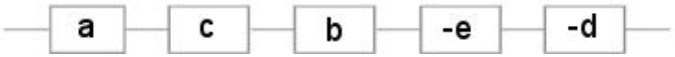
Mode d'évolution des génomes



Mode d'évolution au sein des génomes







Mutation à l’intérieur des séquences

```
P69905 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLSH-----GS 53
P68871 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN 58
P02144 -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKAS 59
        : *:  :   *   ****         * * *::   * *   *  *  *

P69905 AQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAH 113
P68871 PKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHH 118
P02144 EDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSK 119
         :* **  *  *:   :         :  *:: *  *  :    : ::   ::  *  :

P69905 LPAEFTPAVHASLDKFLASVSTVLTSKYR------ 142
P68871 FGKEFTPPVQAAYQKVVAGVANALAHKYH------ 147
P02144 HPGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
          :*    : : :*        :: :*:
```

L’alignement de séquences permet de mesurer la (dis)similarité entre gènes et peut servir à déterminer la distance évolutive les séparant.

Pourquoi s’intéresser à l’évolution des génomes ?

- Pourquoi pas ?
- Classification du vivant basée sur la réalité phylogénétique
- Inférence de la chronologie de l’évolution et identification des forces motrices à l’origine du processus évolutif
- Inférence des génomes ancestraux
- Meilleure compréhension des systèmes biologiques
- Applications diverses
  - génomique comparative
  - développement de médicaments
  - agriculture, etc

Étudier l’évolution par la phylogénie

koala

champignon

Arcella

chat

chauve-souris

blé

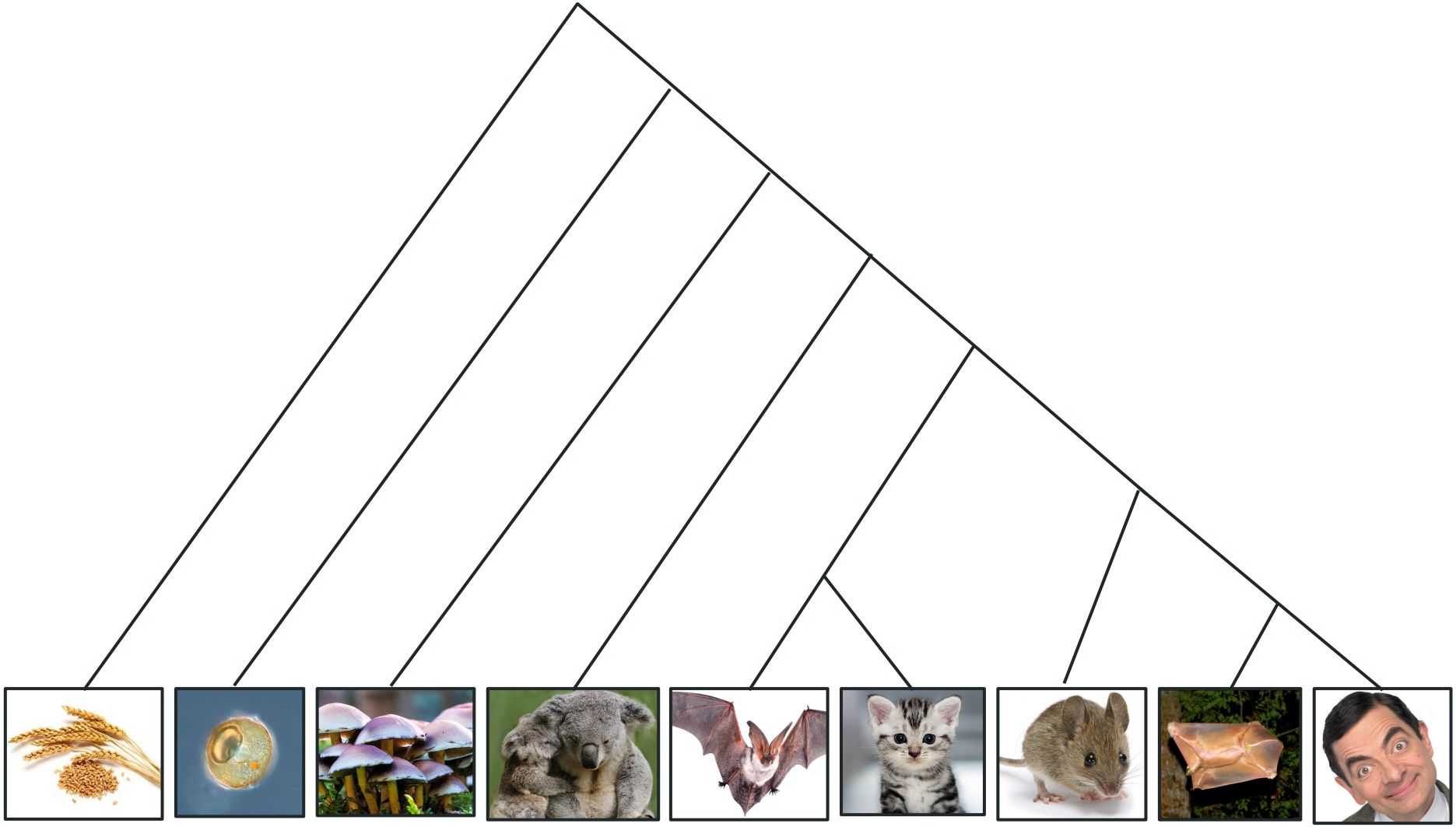
humain

OVNI ??

souris

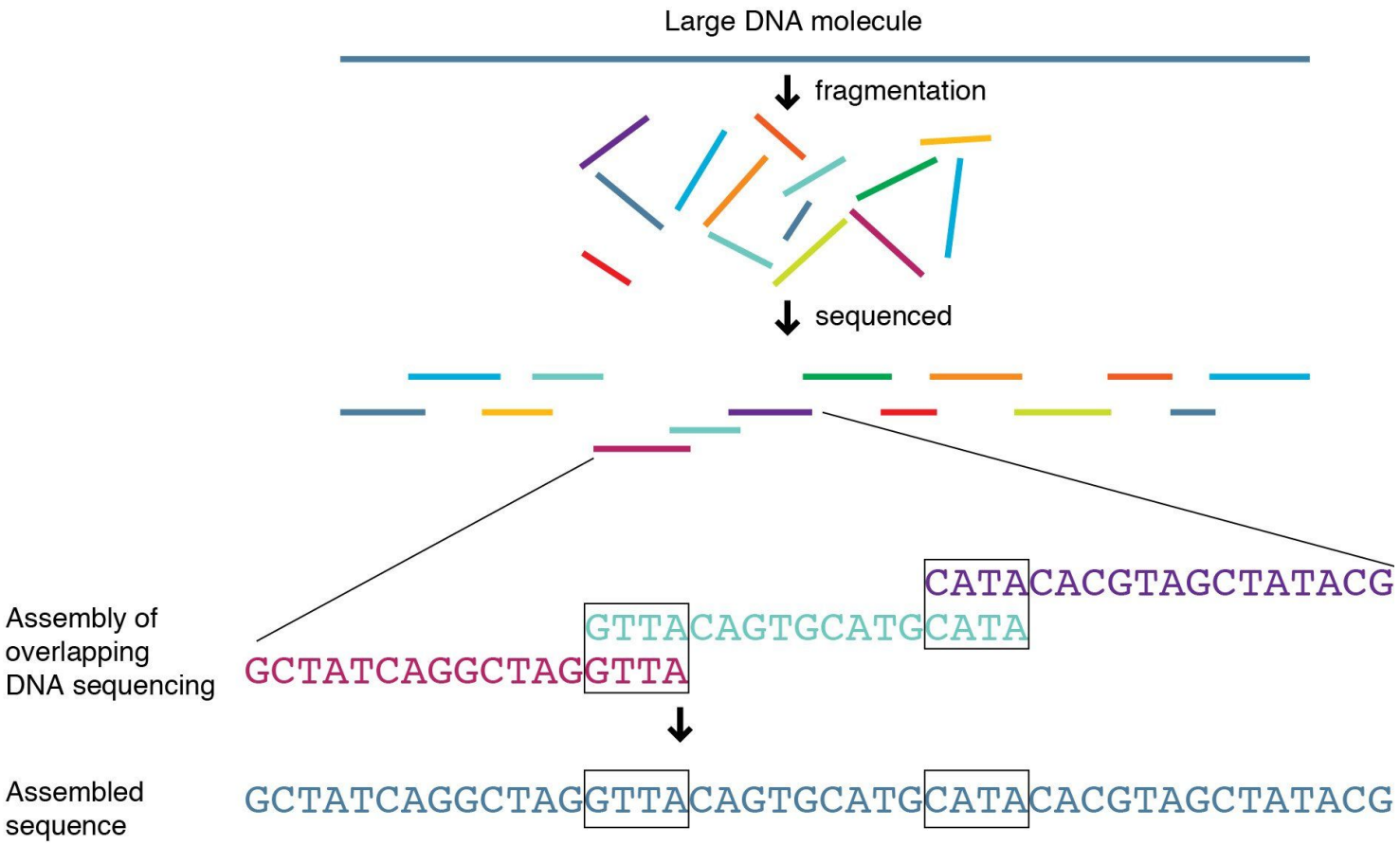
Quel est la relation de parenté entre ces organismes ?

- Quel espèce est plus proche de l’homme ?
- Quel espèce est plus distante des autres ?

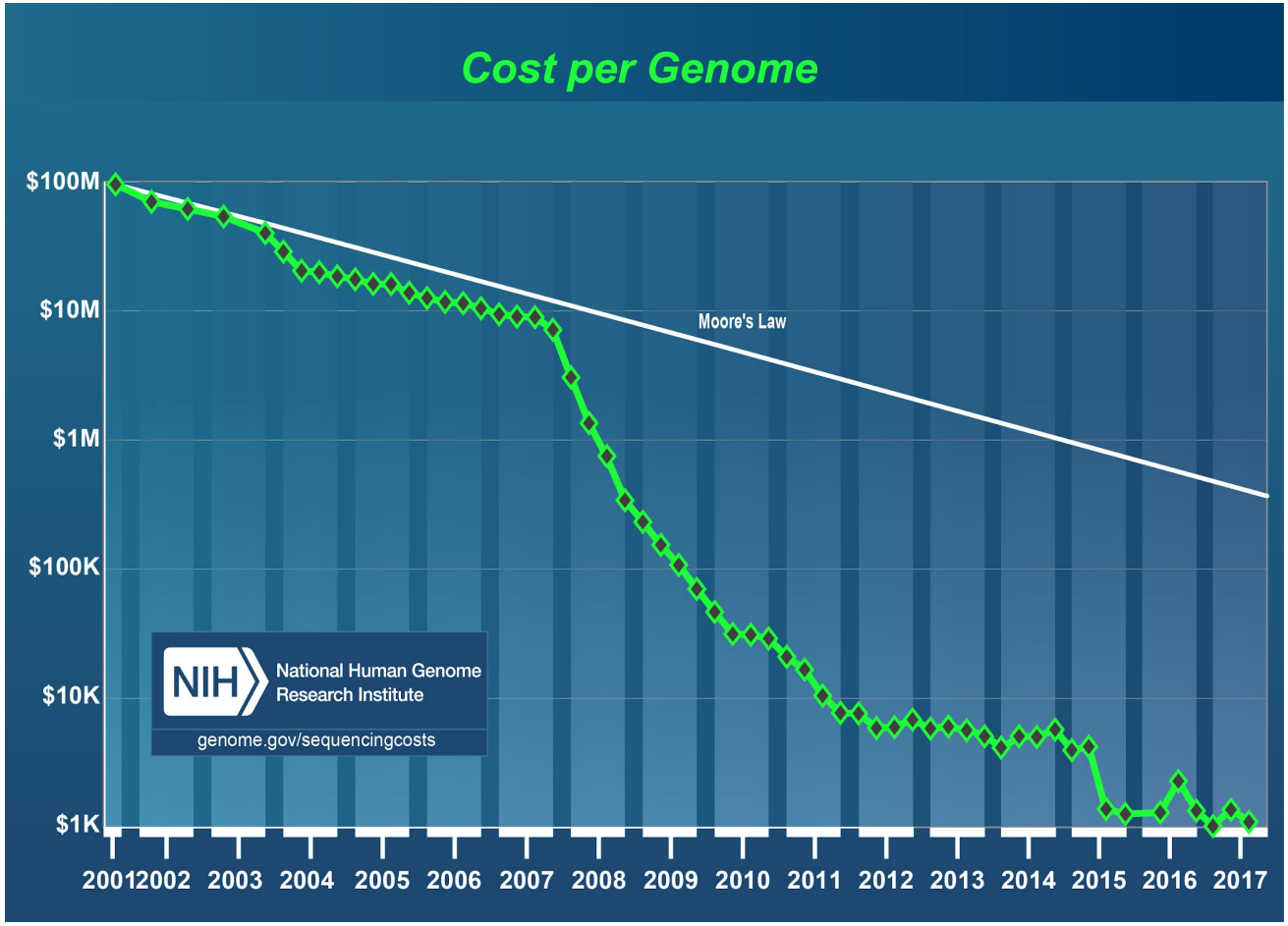


les choses ne sont pas forcément comme elles paraissent !

Le séquençage, révolution de la biologie moléculaire

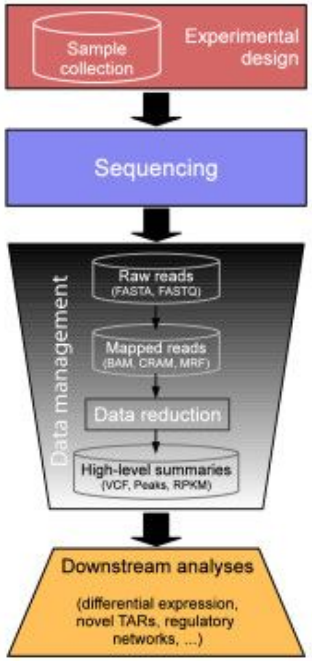


Séquencer son génome, moins chère qu'un iphone X ?

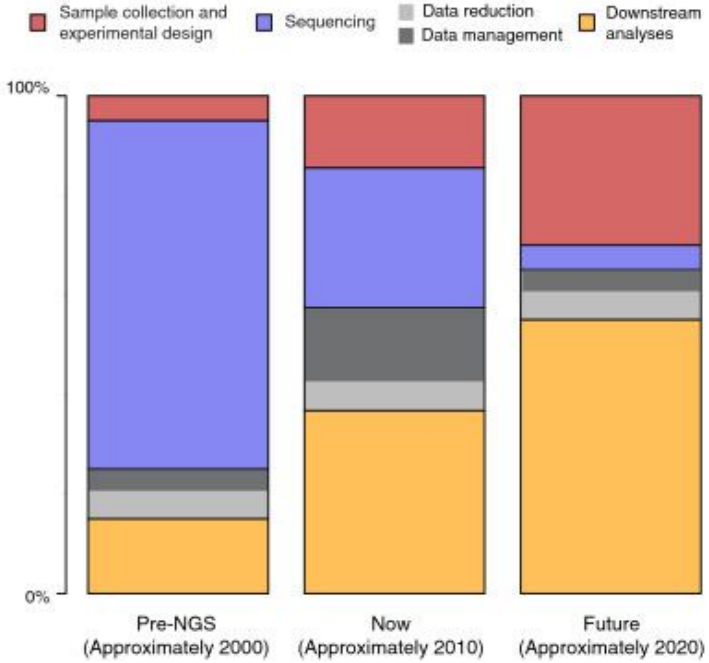


2017 : Illumina promet le séquençage à 100$

Nouveaux challenges en génomique !



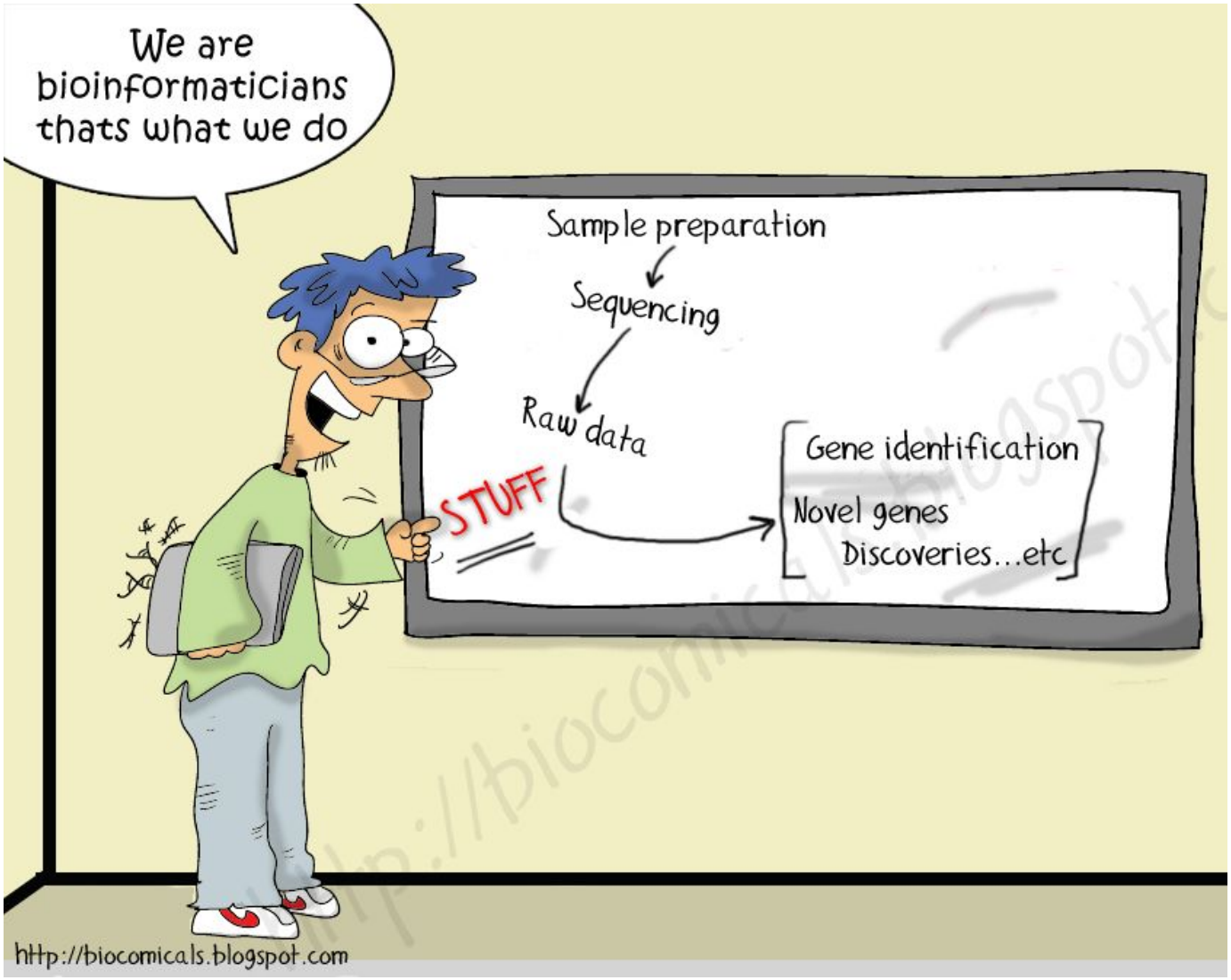
Sboner et al. 2011
doi:10.1186/gb-2011-12-8-125



- Beaucoups de données qui doivent être stockées de façons effective pour faciliter les requêtes
- Algorithmes **efficaces** et **ultra-rapides** pour analyser ces données
- Faible ratio signal/bruit : besoin de méthodes précises
- etc

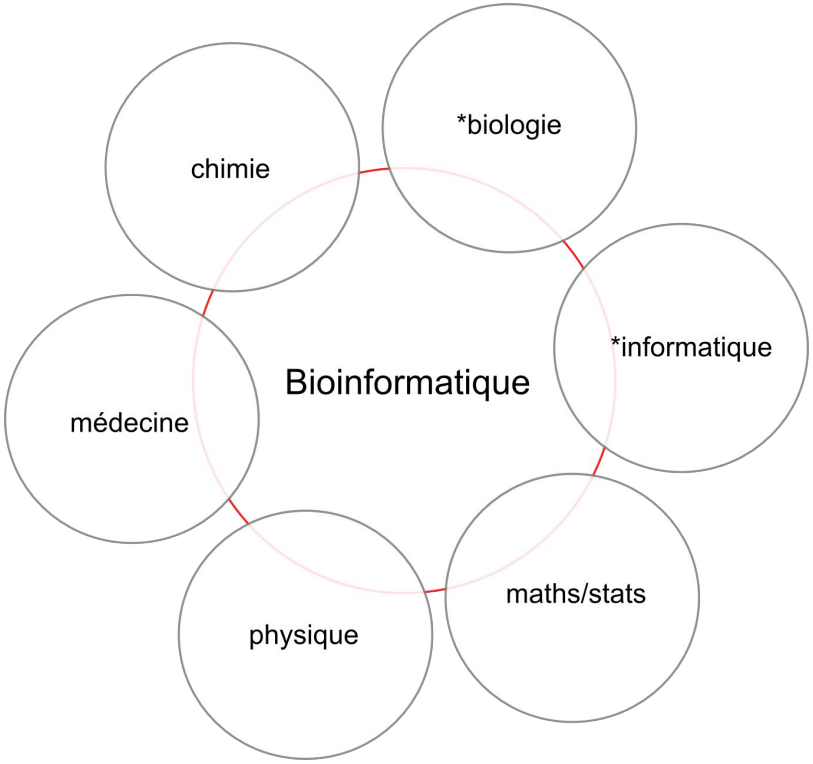
Analyser les données biologiques devient plus coûteux que de les générer !

...AGTAGTGTCTCGGTGGCTGTGGGACGTGGTGAC...



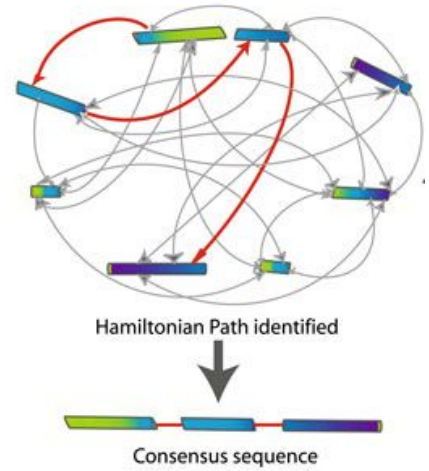
Qu’est ce que la bioinformatique ?

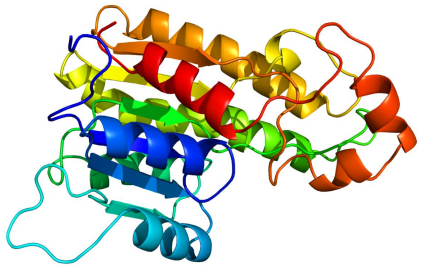
“ **Science** interdisciplinaire dont le but est de développer et appliquer des méthodes et technologies **informatiques** pour étudier des **problèmes biologiques** ”

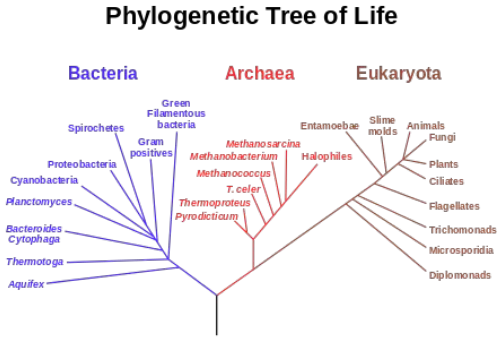


Application bioinformatique

- Assemblage de génome
  - graphe de De-Bruijn, cycle hamiltonien/eulérien, kmer, etc
- Annotation de génomes
  - alignement de séquences, chaîne de markov, modèle de covariance, etc
- Prédiction de structure secondaire/tertiaire
- Prédiction d'interaction de biomolécules
- Analyse d'expression de gènes
- Analyse d'images de microscopie, etc
- Science judiciaire
- Changement climatiques
- **Analyse phylogénétique**
  - vraisemblance, clustering, inférence bayésienne, etc

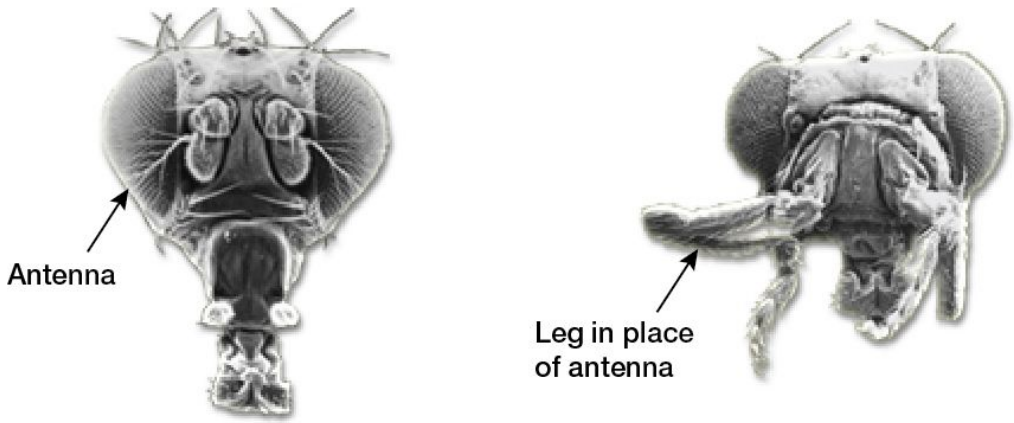






Exemple de problèmes étudiés au LBIT

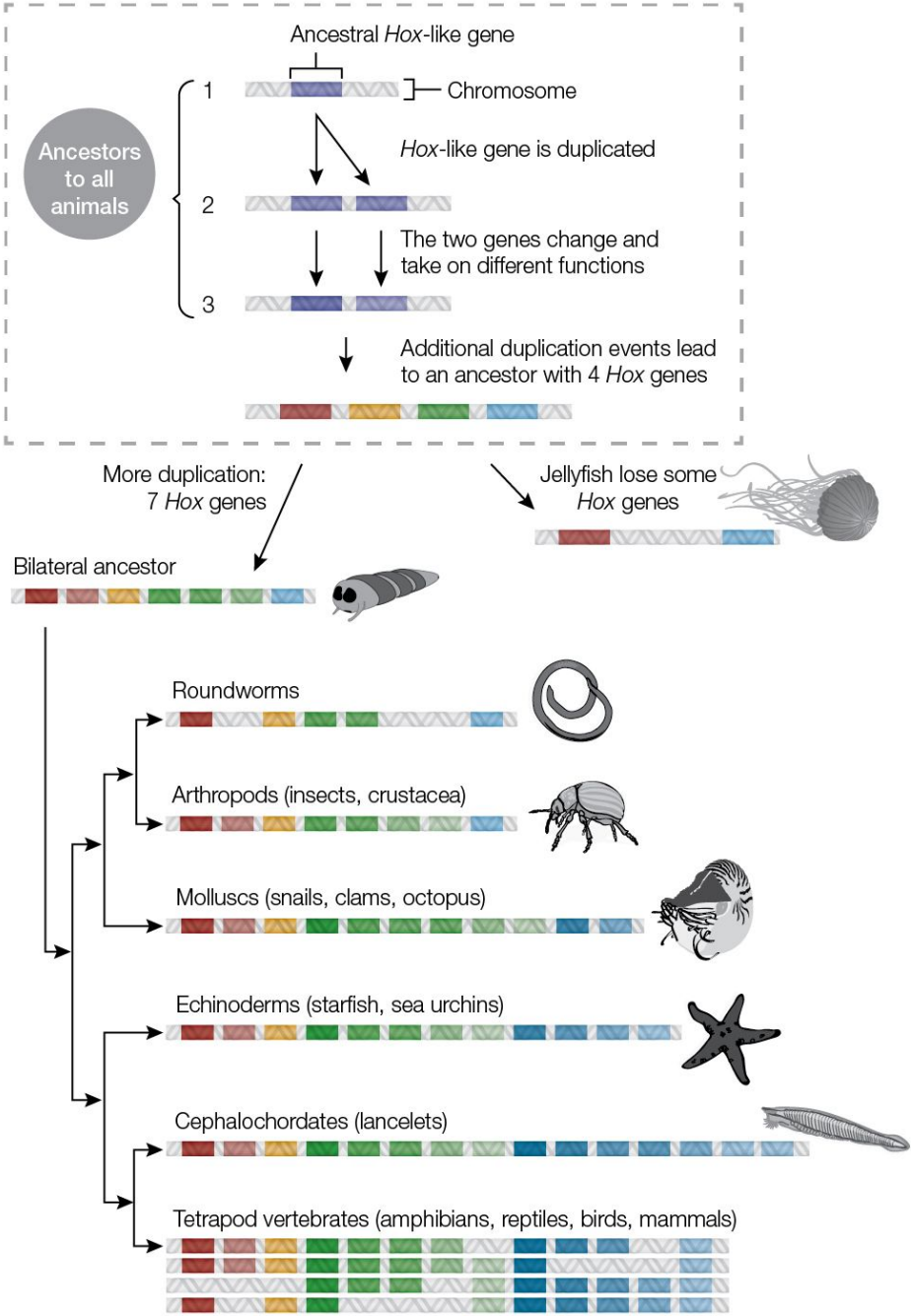
Évolution des gènes au sein des génomes: exemple des gènes HOX



- responsable de la détermination de l'identité des segments morphologiques
- présents chez tous les animaux bilatéraux.

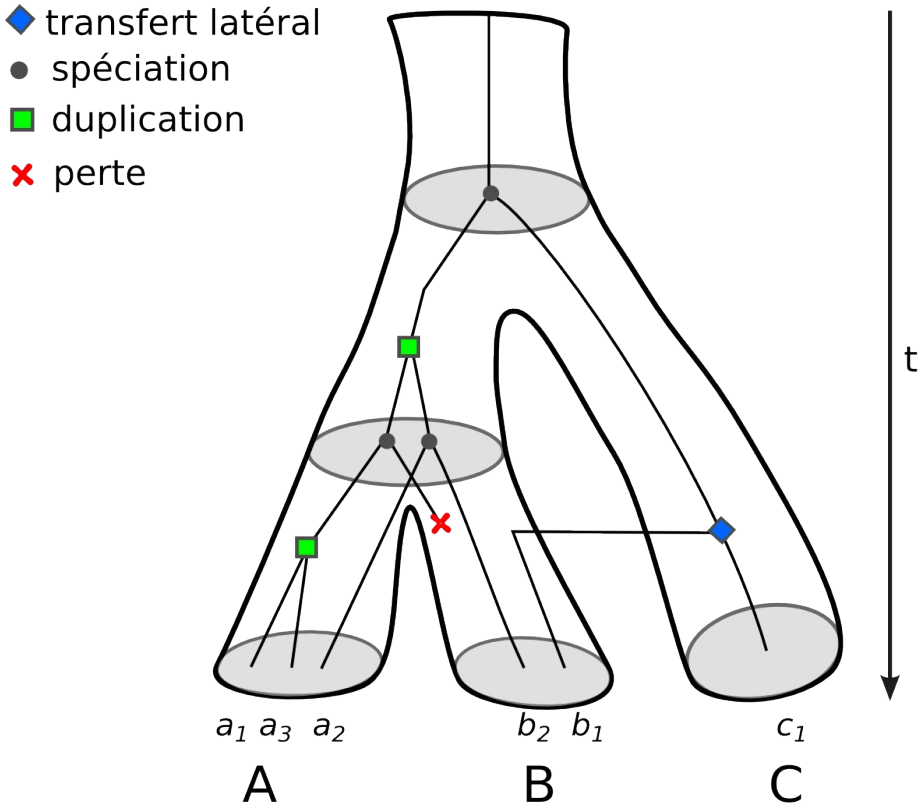
Évolution des familles de gènes

- Histoire évolutive des familles de gènes est marquée par plusieurs évènements de **duplication** et **pertes**
- Gènes homologues: descendent d'un même gène ancestral commun
  - orthologues
  - paralogues
  - xénologues
- Arbre phylogénétique de gènes pour mieux comprendre leur évolution

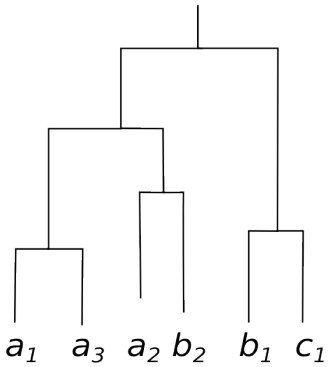


http://learn.genetics.utah.edu/content/basics/hoxgenes/

Inférence d'arbre phylogénétique de gènes



Une réconciliation de G
avec l'arbre des espèces



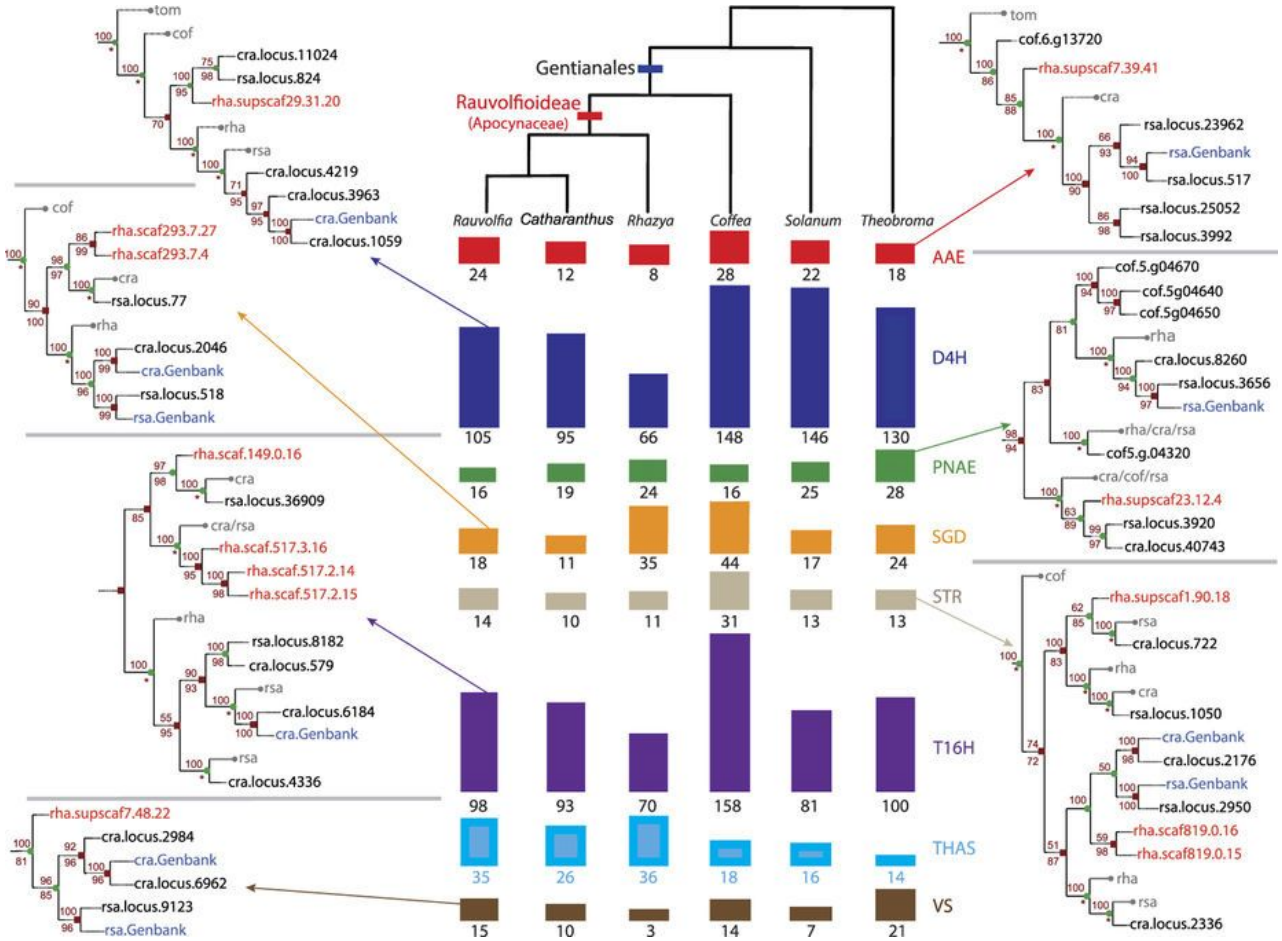
Arbre de gènes G

```
a1: MIKKPG ... ATINQSD*
a3: MLKKPG ... ASINQSD*
a2: MLR-PD ... ATINQSD*
b2: MLR-PE ... AT-NQSD*
b1: MLHK-D ... VTINNSD*
c1: MMHKPG ... VTINQSD*
```

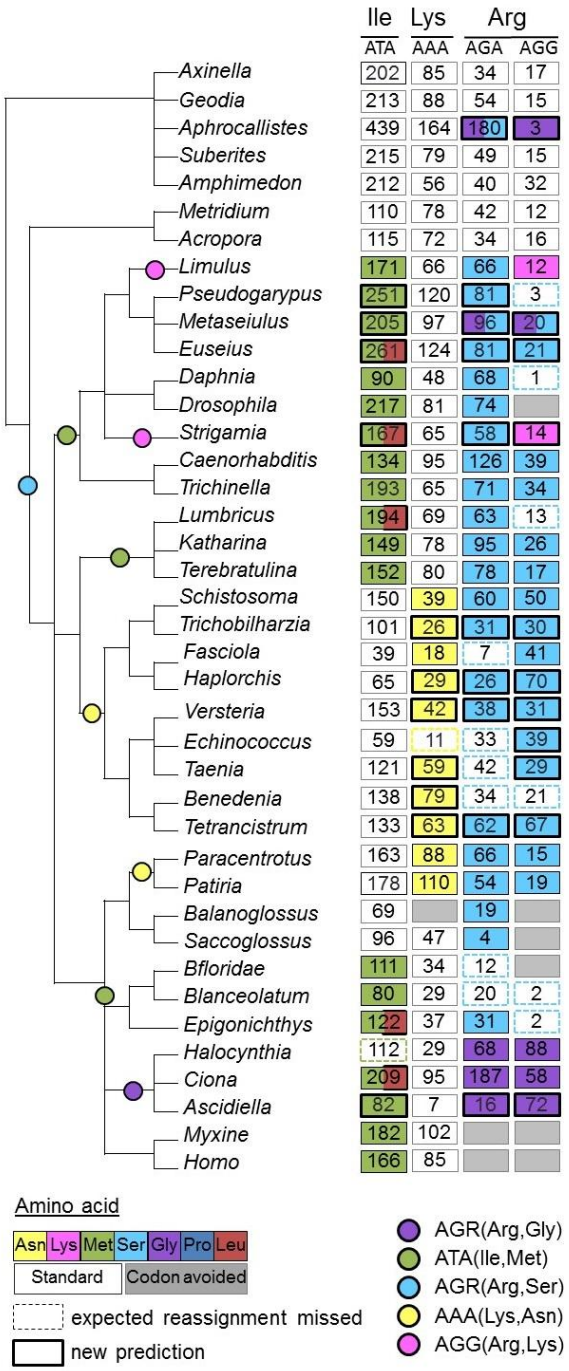
Alignement

L'information provenant des séquences uniquement n'est pas suffisante !

Évolution de gènes MIA (Monoterpenoid Indole alkaloid) chez les Gentianales avec ProfileNJ



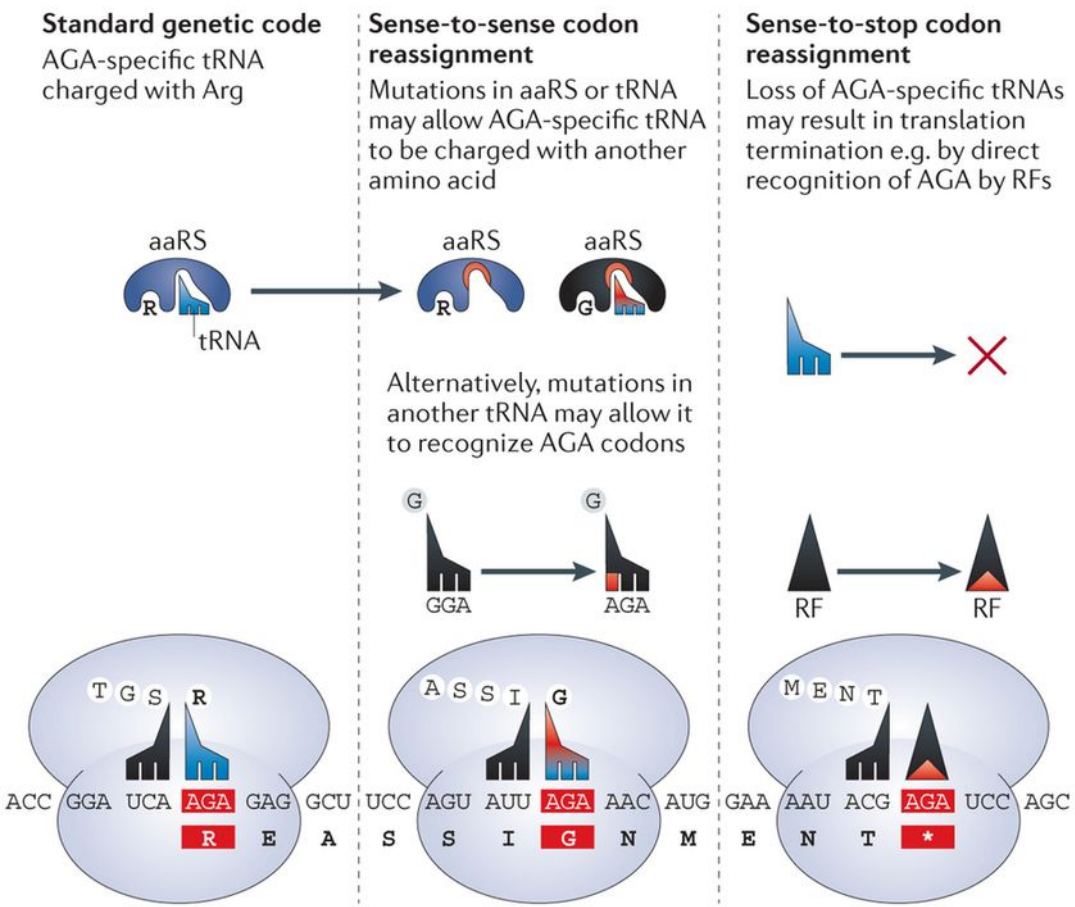
Sabir et al. 2016



Noutahi et al. 2017

“ Il n’y a pas que le génome qui évolue, son interprétation (code génétique) peut aussi évoluer ”

Mécanismes à l'origine des réaffectations d'identité de codons



Baranov et al. 2015

Une méthodologie pour prédire les réaffectations de codons.

CoreTracker: accurate codon reassignment prediction, applied to mitochondrial genomes

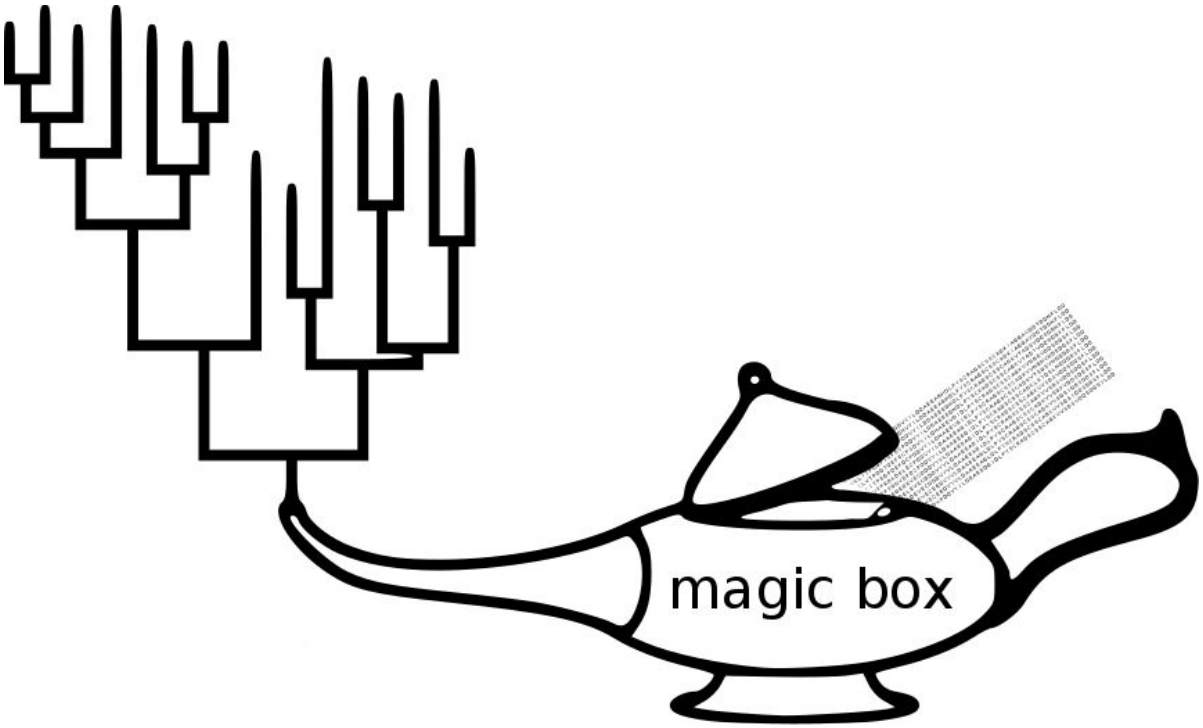
Emmanuel Noutahi, Virginie Calderon, Mathieu Blanchette, Franz B Lang, Nadia El-Mabrouk

Bioinformatics, Volume 33, Issue 21, 1 November 2017, Pages 3331–3339,
https://doi.org/10.1093/bioinformatics/btx421
Published: 26 June 2017 **Article history**

- Exploite la conservation de domaines à l'intérieur des séquences protéiques des gènes homologues.
- Tient compte du contexte évolutif des génomes.
- Tests statistiques et Random Forests pour prédire les réaffectations
- Valide les prédictions en mesurant leur impact sur les génomes
- Peut-être complémenté par l'étude de l'évolution des ARNts (Noutahi et al. 2018, en préparation)

Pipeline de CoreTracker

Merci !
Questions ?



ProfileNJ, CoreTracker, GATC, etc disponible à

- https://github.com/UdeM-LBIT/
- https://github.com/maclandrol/