EGC 2020 – Bruxelles
(Extraction et Gestion des Connaissances)
Atelier VIF
(Visualisation d'informations, interaction et fouille de données)

Combiner arbres phylogénétiques et visualisation d'ensembles

Jean-Baptiste Lamy, Flora Jay

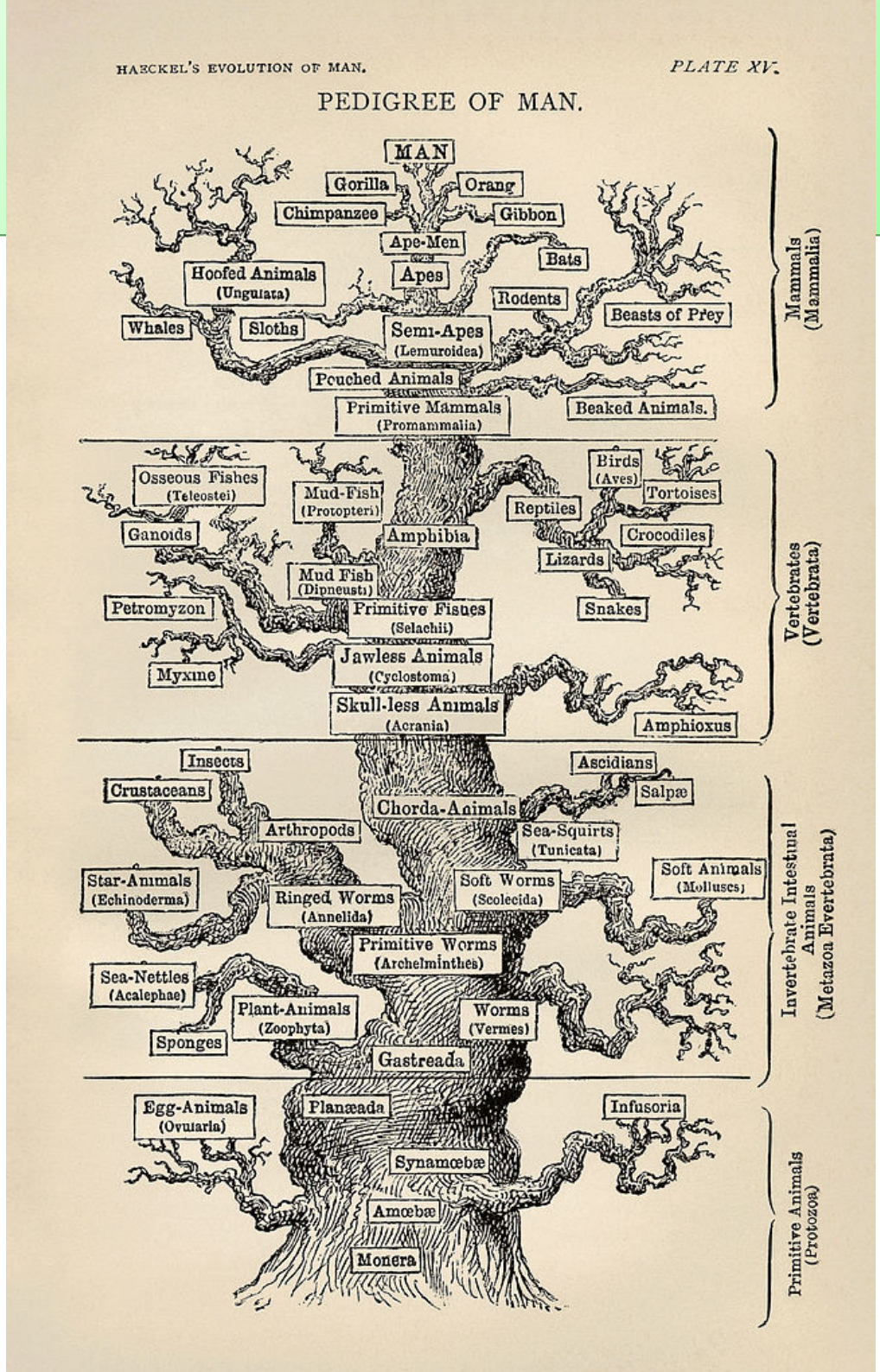
Laboratoire de Recherche en Informatique,CNRS/Université Paris-Sud/Université Paris-Saclay, Orsay, France

Laboratoire EcoAnthropologie et Ethnobiologie, CNRS/MNHN/Université Paris Diderot, Paris, France

LIMICS
Université Paris 13, 93017 Bobigny
Sorbonne Universités
INSERM UMRS 1142

Introduction

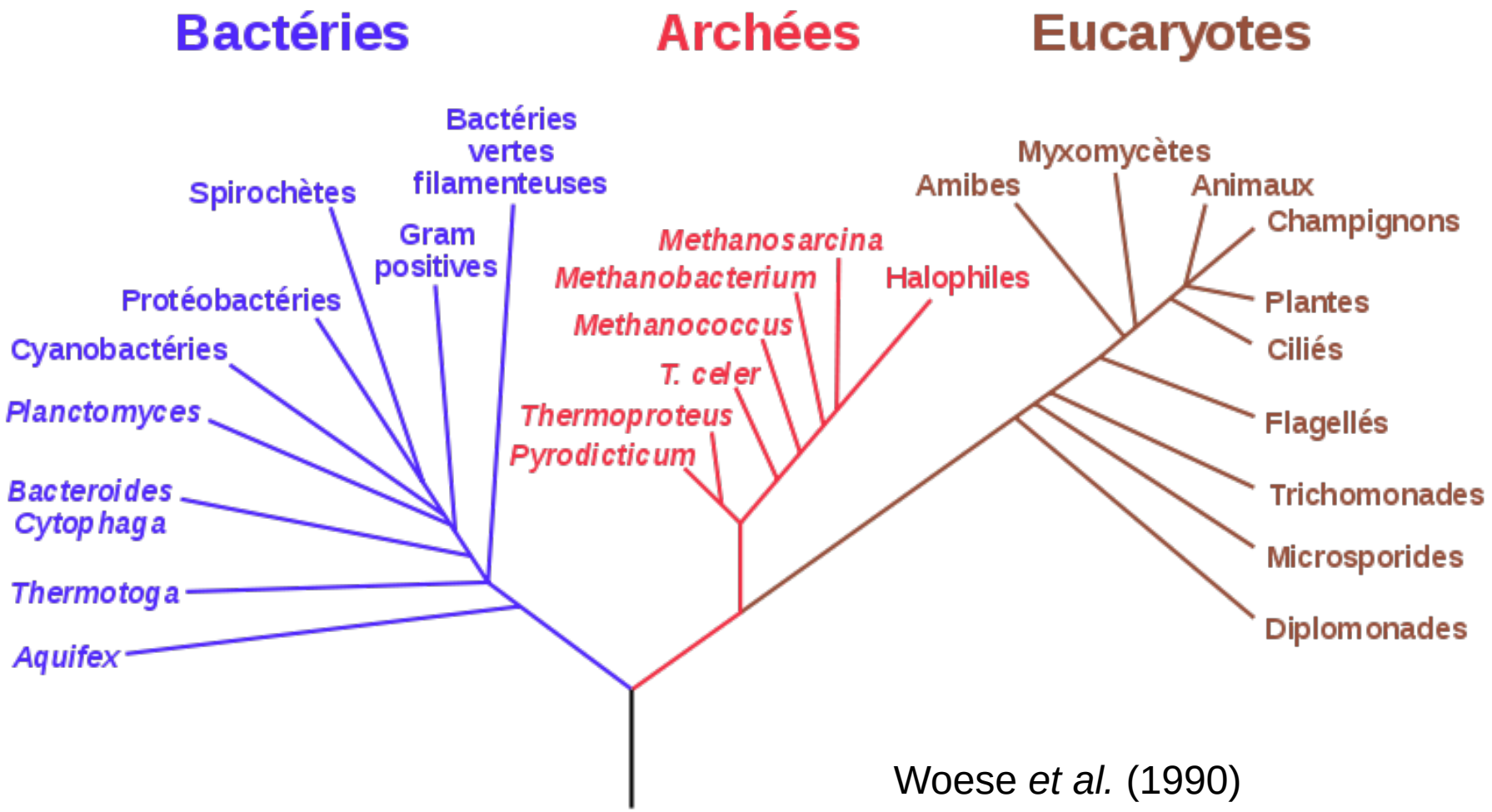
- **Phylogénie**
  - Science qui étudie les liens de parenté entre les être vivants :
    - Individus
    - Population
    - Espèces
  - et qui cherche à reconstituer l’évolution
- **Les arbres sont très largement utilisés pour représenter l’information phylogénétique**
  - Arbres phylogénétiques



Arbre de vie, par Ernst Haeckel dans *L'évolution de l'Homme* (1879)

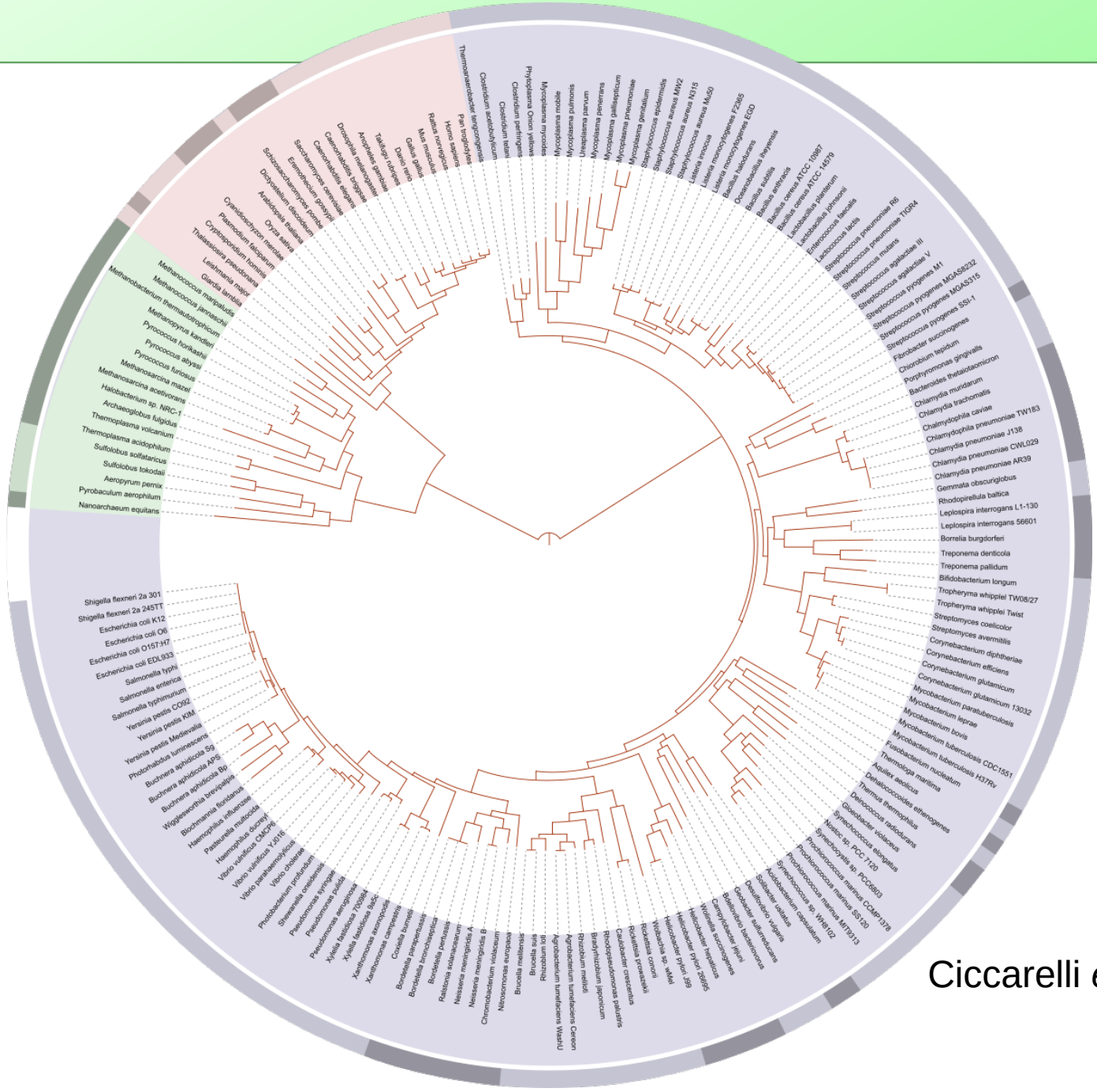
Introduction

Arbre phylogénétique de la vie



Woese *et al.* (1990)

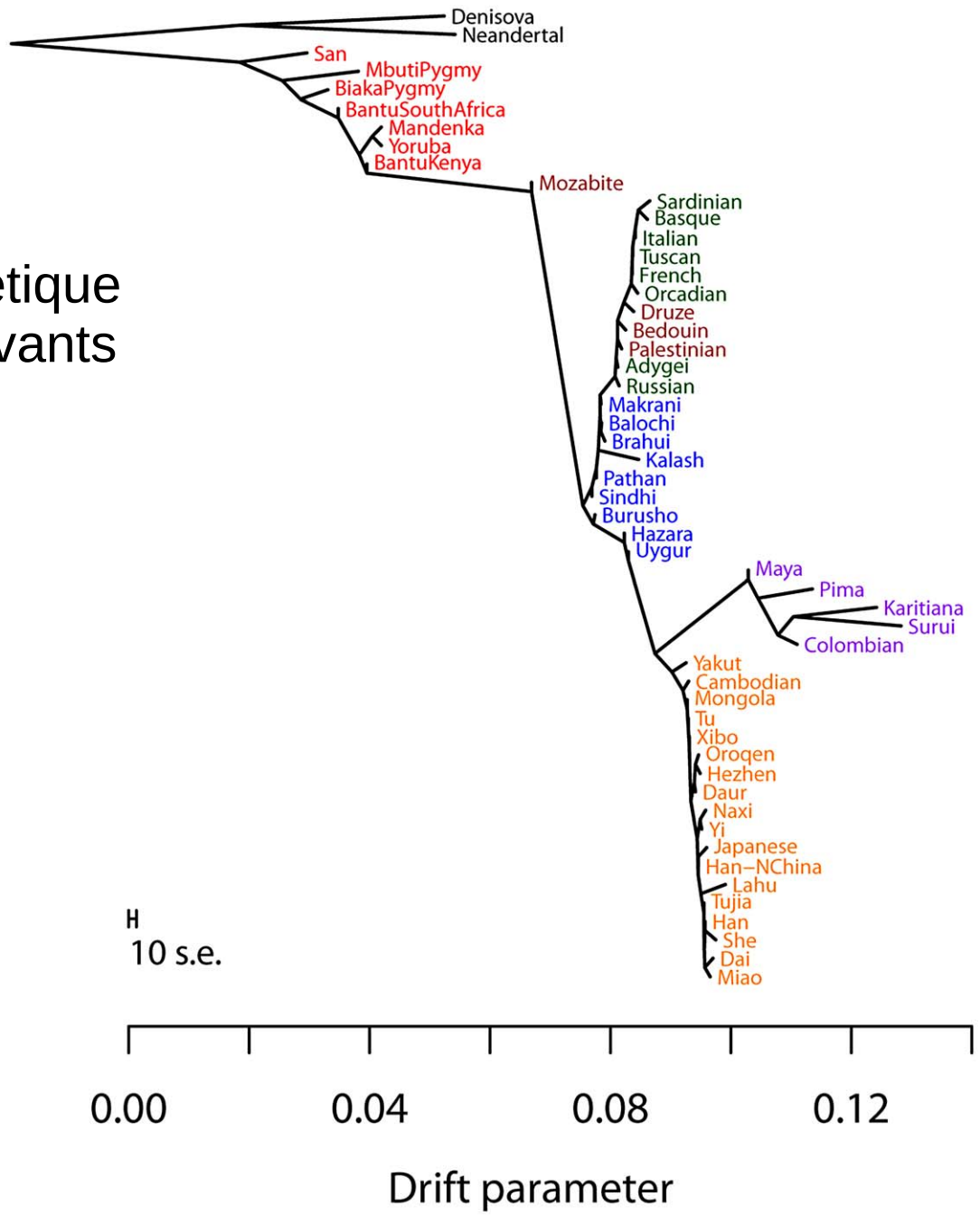
Introduction



Ciccarelli *et al.* (2006)

Introduction

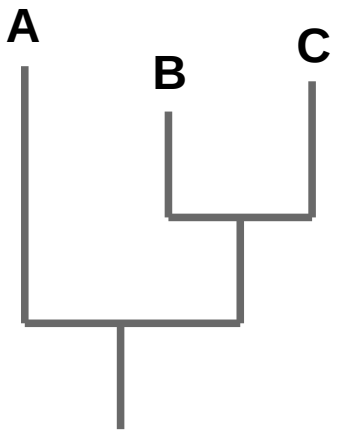
- **Génétique des populations**
  - Étude de la diversité génétique des populations d'êtres vivants et de leurs relations



JK Pickrell *et al.* (2012)

Introduction

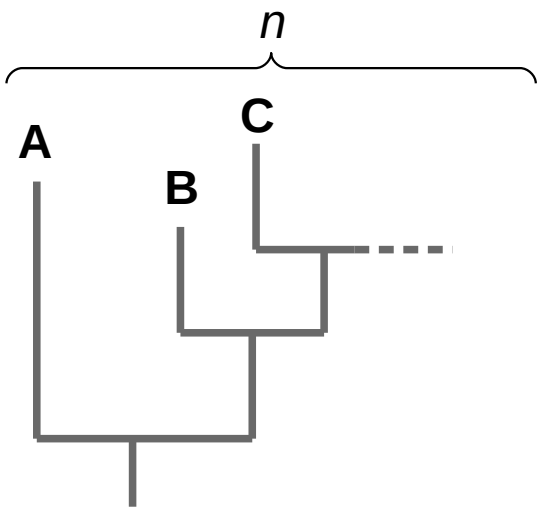
- **Problème**
  - Les arbres ne peuvent représenter qu'une partie des similarités observées entre les êtres vivants



- Cet arbre représente la similarité entre B et C, et entre A, B et C
- Mais il y a 4 similarités possibles ici : {A, B}, {A, C}, {B, C}, {A, B, C}
- Les éventuelles similarités existantes entre A et B, et entre A et C ne peuvent pas être représentées sur cet arbre !

Introduction

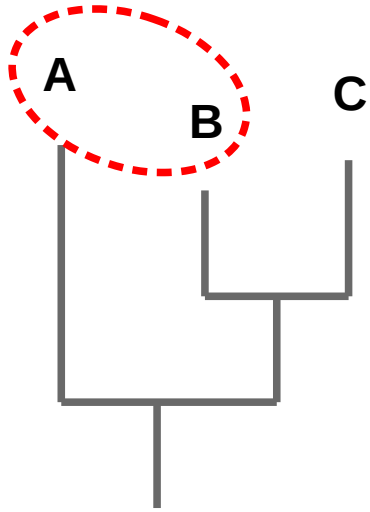
- **Pour un arbre à *n* feuilles :**
  - Il existe $2^n - n - 1$ similarités possibles, correspondant aux sous-ensembles d'au moins deux feuilles
  - Mais un arbre peut au mieux représenter $n - 1$ de ces similarités



- Plus *n* augmente, plus la vision donnée par l'arbre devient réductrice

Introduction

- **Pourquoi peut-il exister des similarités entre des branches distincts d'un arbre phylogénétique ?**
  - Échange de gènes entre individus d'espèces différentes : **transfert horizontal de gènes** (plasmides chez les bactéries, racines des plantes)
  - Reproduction entre espèces proches, notamment peu après la scission d'une espèce en deux ?
  - **Migration** en génétique des populations : mouvement d'un groupe d'individus qui quittent une population pour en rejoindre une autre, et lui apporte leur matériel génétique
    - Un arbre peut représenter la composante évolutive et démographique de la génétique des populations, mais pas la composante migratoire

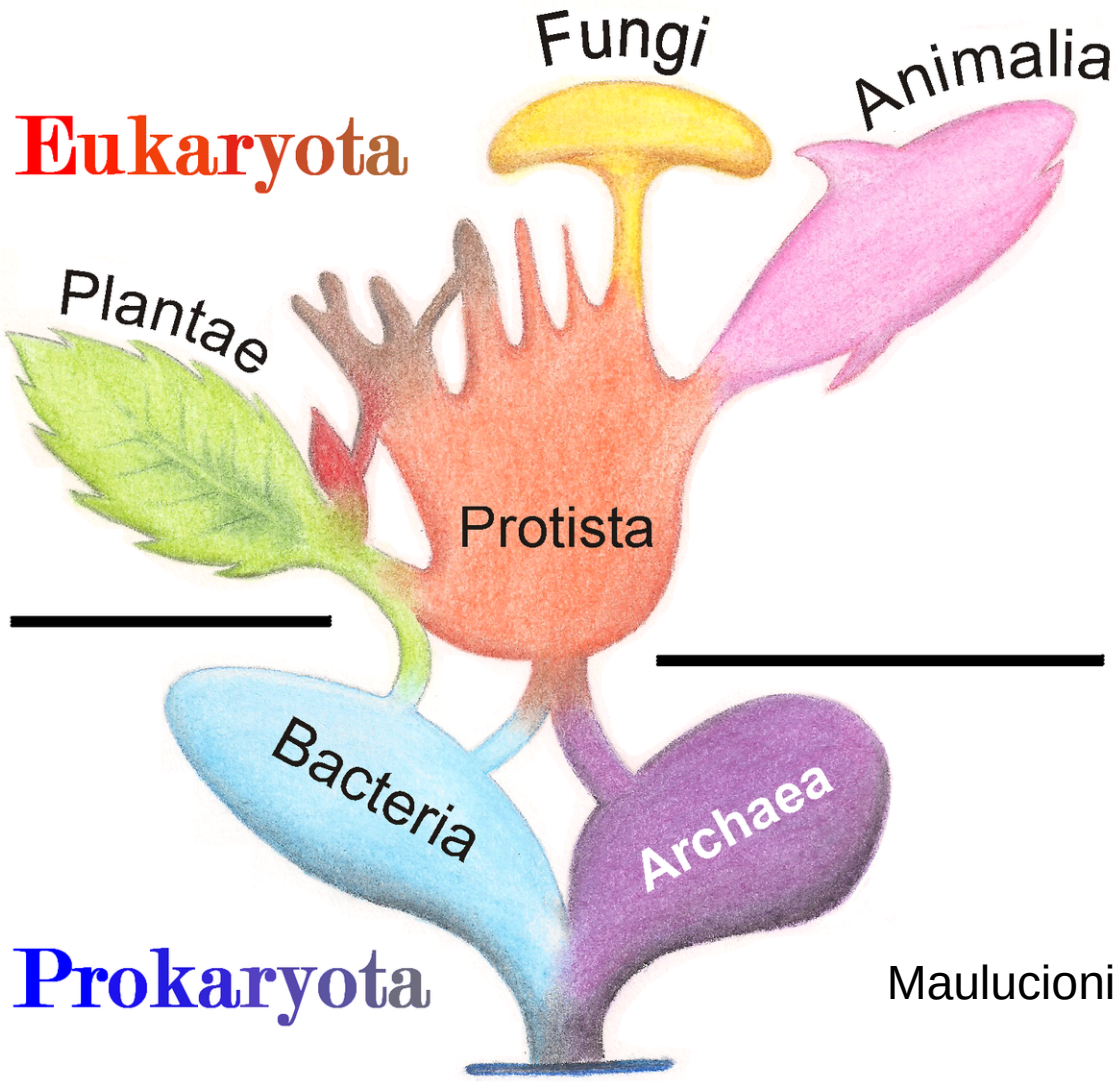


Introduction

- **=> Nécessité d’une représentation plus complexe que les arbres !**

Introduction

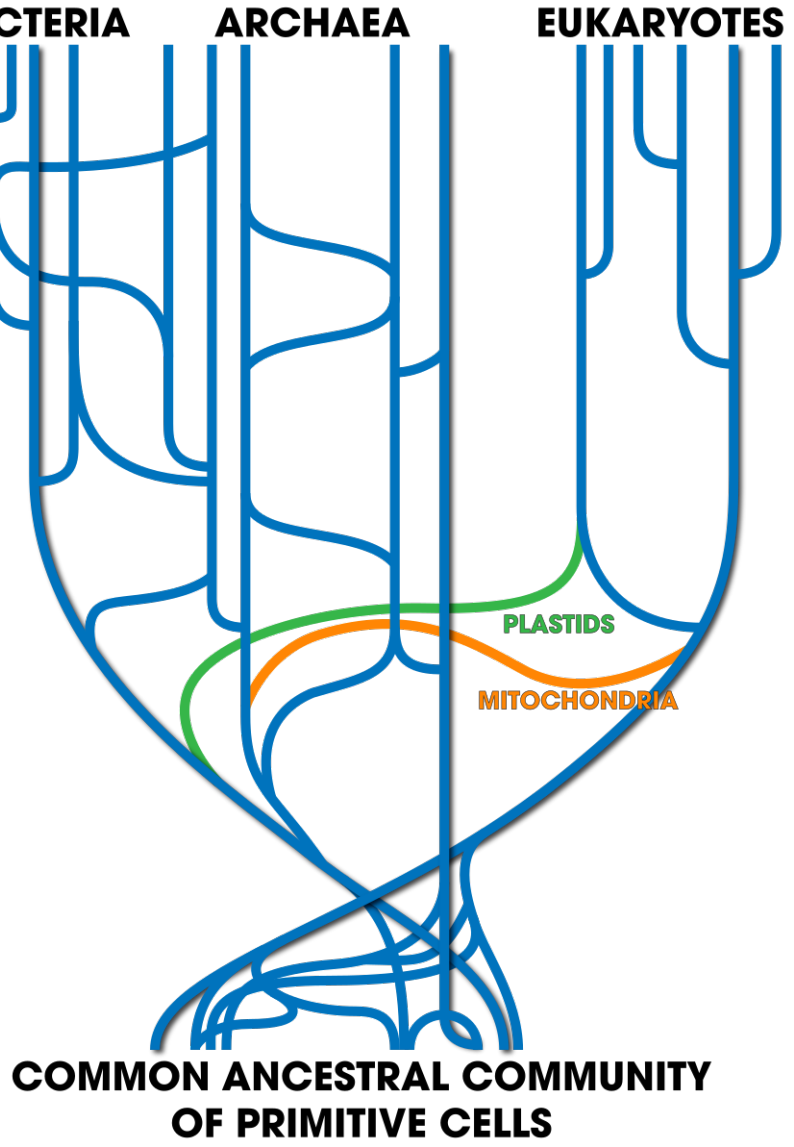
- **Arbre “à bulle”**



Maulucioni y Doridí, 2013

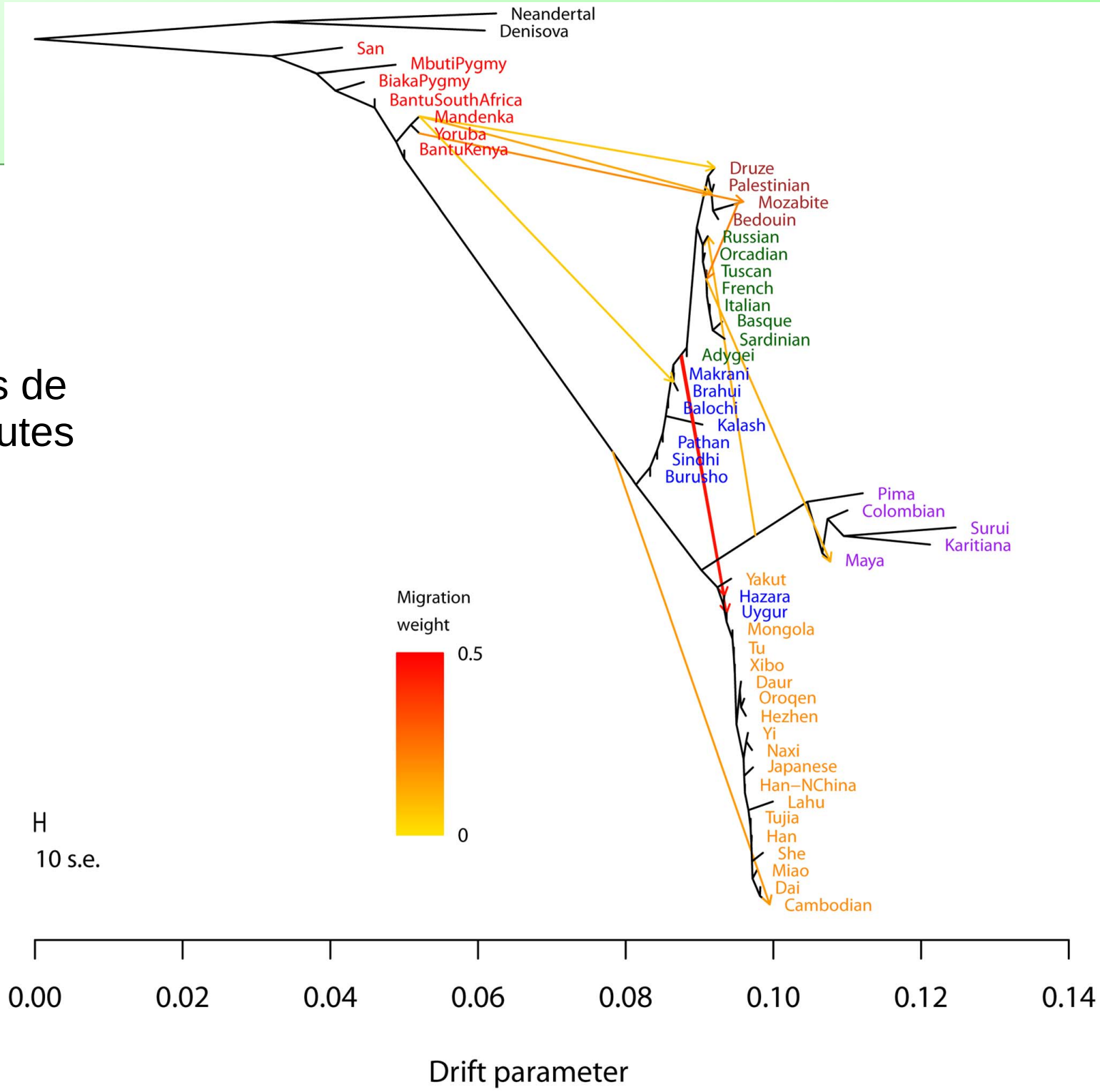
Introduction

- **Arbre en réseaux**

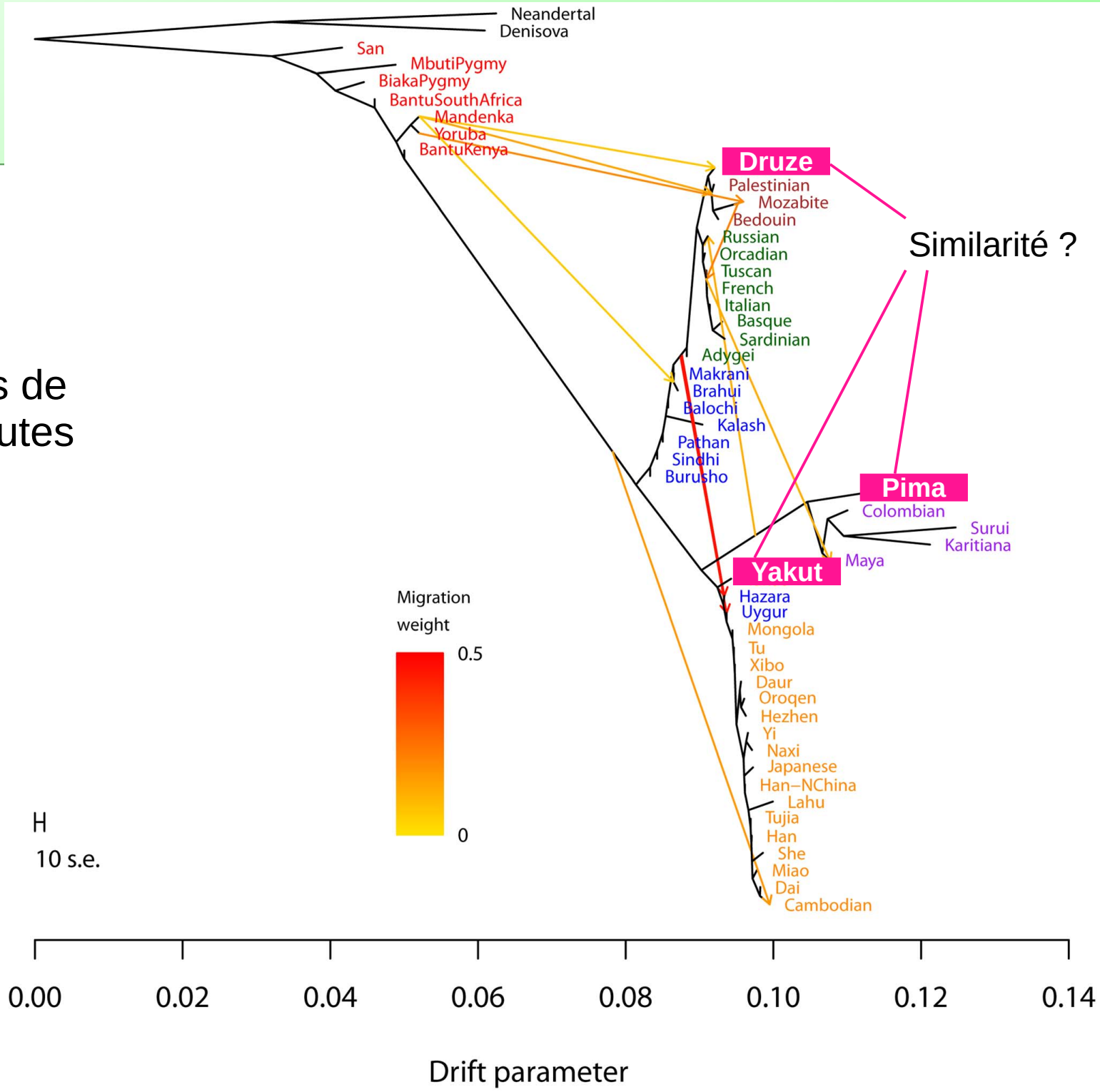


BF Smets (2005), CC-BY-SA

- **Arbre + flèches**
  - Peu lisible
  - Ne permet pas de représenter toutes les similarités



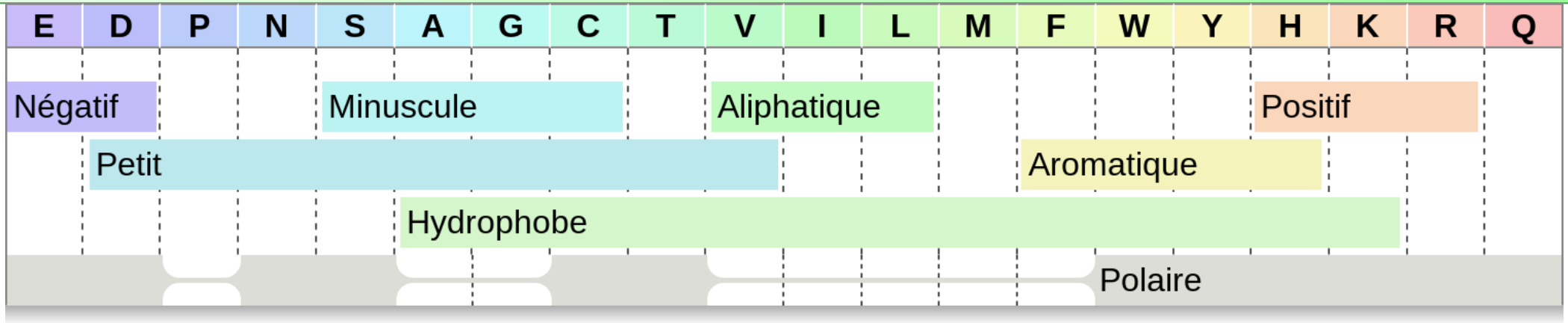
- **Arbre + flèches**
  - Peu lisible
  - Ne permet pas de représenter toutes les similarités



Objectif

- **Proposer une visualisation plus complète de l'information phylogénétique que celle proposée par les arbres**
- **En s'appuyant sur la visualisation d'ensembles**
  - Et plus particulièrement les boîtes arc-en-ciel (JB Lamy 2017, 2019)

Les boîtes arc-en-ciel

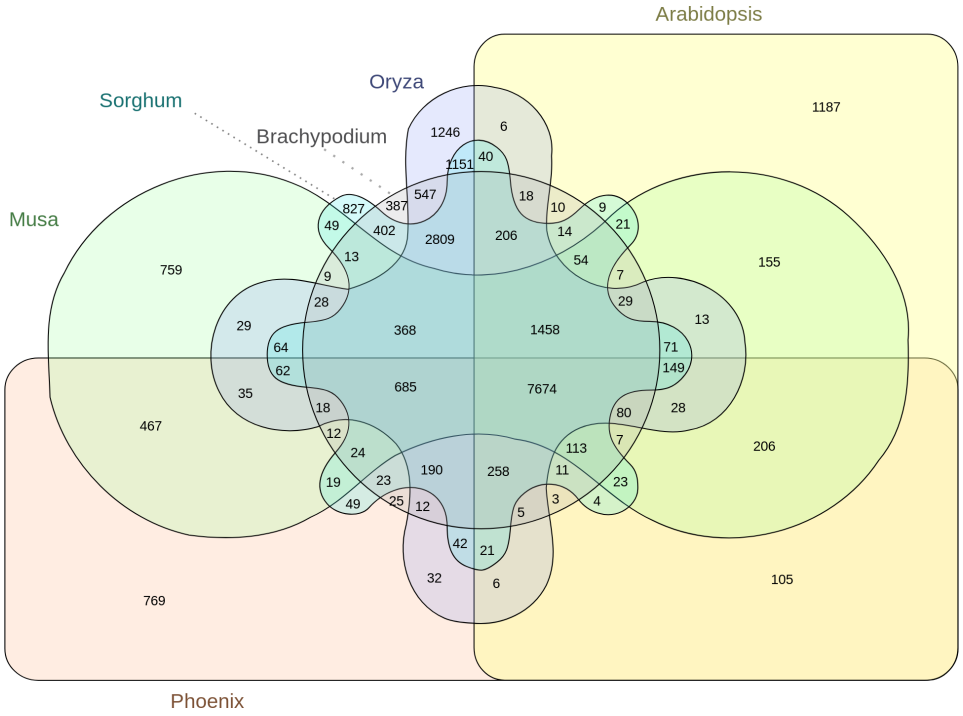


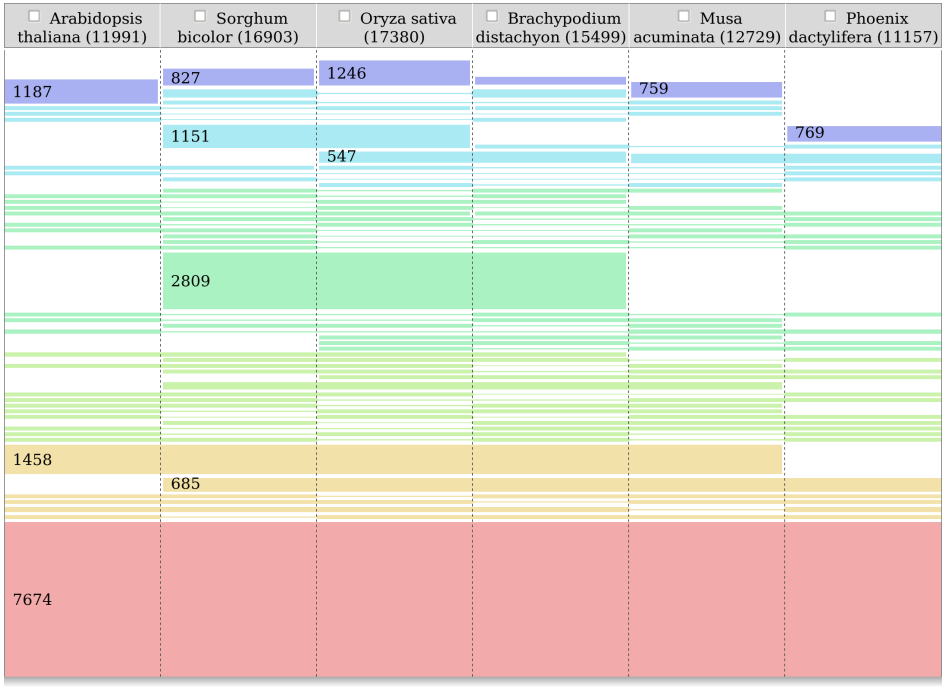
- **Une technique récente pour visualiser des ensembles**
  - éléments => colonnes
  - ensembles => boîtes rectangulaires
  - Les éléments sont ordonnés par un algorithme métaheuristique (Artificial Feeding Birds) de sorte à placer côte à côte les éléments qui appartiennent aux mêmes ensembles
  - Lorsque cela n'est pas possible, des "trous" sont présents dans les boîtes
  - Les boîtes sont empilées avec les plus grandes en bas

[Lamy JB et al. Rainbow boxes: a new technique for overlapping set visualization and two applications in the biomedical domain. **Journal of Visual Language and Computing** 2017]

Les boîtes arc-en-ciel

- **Variante proportionnelle des boîtes arc-en-ciel : RainBio**
  - La hauteur des boîtes représente la cardinalité des ensembles
  - Seule technique permettant la visualisation proportionnelle exacte jusqu'à 6 ensembles (=64 intersections) :





[Lamy JB et al. RainBio : Proportional visualization of large sets in biology. **IEEE Transactions on Visualisation and Computer Graphics** 2019]

Approche 1 : ensembles

- **Première approche : visualiser l'information phylogénétique sous forme d'ensembles**
  - Jeu de données extrait du "*1000 Genomes Project*"
    - 26 populations
    - 2500 personnes correspondant à 5000 matériels génétiques
      - Pour chacun, 750 valeurs booléennes correspondant à des mutations sur des positions marqueurs génétiques
      - La valeur 0 représente la valeur ancestrale
      - La valeur 1 représente la valeur mutée
  - 1 population correspond à plusieurs matériels génétiques
    - Ex la population A a le marqueur 1 présent chez 70% de ces individus
    - => nécessité de discrétiser pour arriver à des ensembles

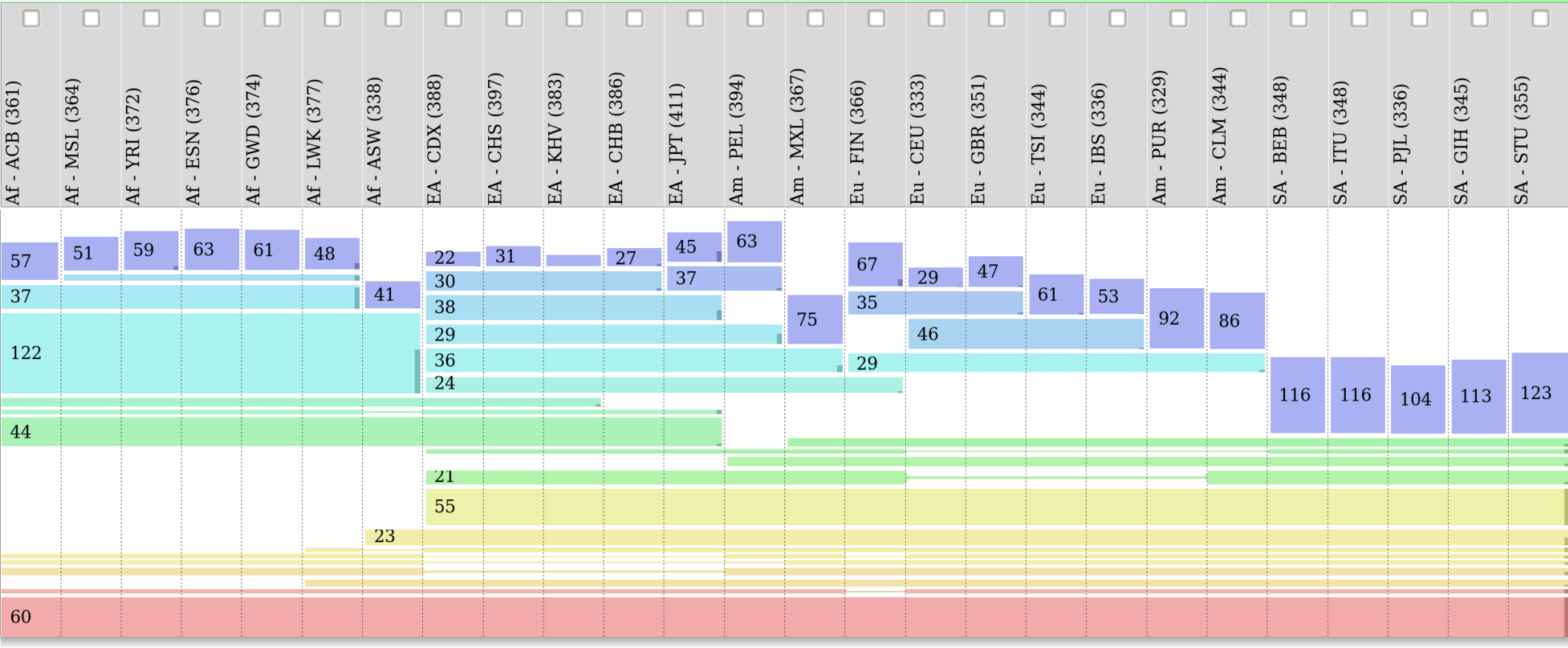
Approche 1 : ensembles

- 26 populations
- 2500 personnes correspondant à 5000 matériels génétiques
  - Pour chacun, 750 valeur booléenne correspondant à des mutations sur des positions marqueurs génétiques

- 1 population = 1 élément
- 1 marqueur booléen = 10 ensembles
  - Ensemble des populations ayant une moyenne > 95% pour le marqueur
  - Ensemble des populations ayant une moyenne > 85% pour le marqueur
  - Ensemble des populations ayant une moyenne > 75% pour le marqueur
  - Ensemble des populations ayant une moyenne > 65% pour le marqueur
  - …

- Visualisation avec RainBio
  - Limité à 64 intersections (les autres sont clusterisés)

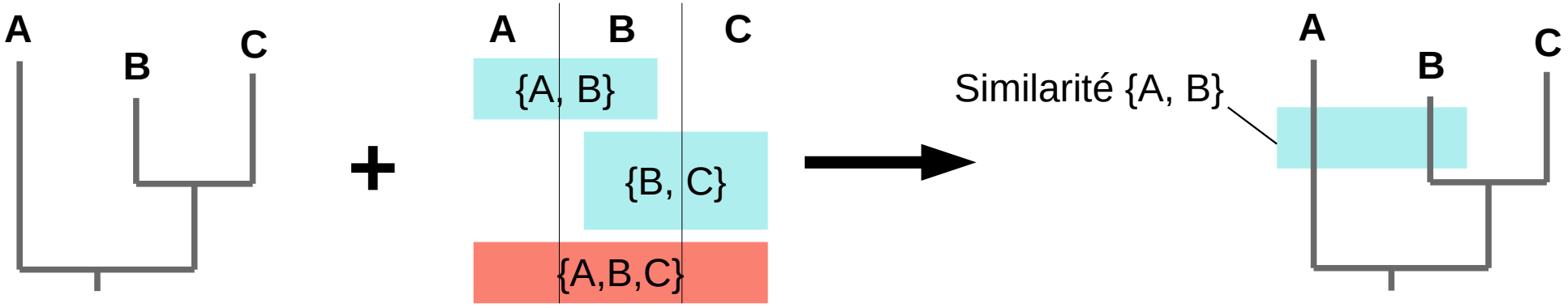
Approche 1 : ensembles



- **Des similarités intéressantes**
  - Mais l'on ne retrouve pas toutes les informations de l'arbre

Approche 2 : arbre + ensembles

- **Seconde approche : visualiser l'information phylogénétique sous forme d'un arbre avec des boîtes tracées par dessus**
  - Permet de garder l'arbre tout en l'enrichissant avec des similarités qu'il ne peut pas représenter

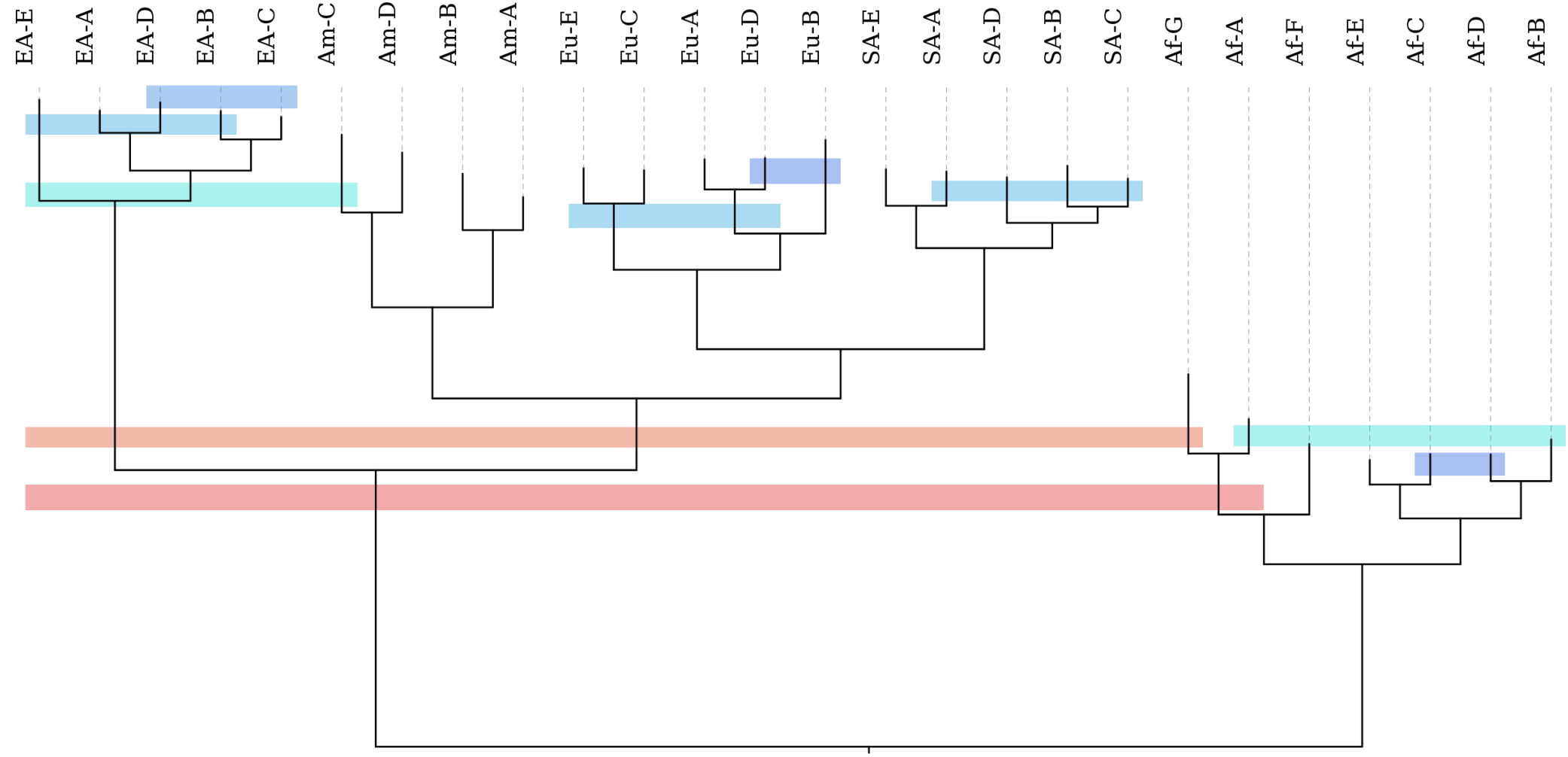


- Les boîtes correspondant à des similarités présentes sur l'arbre sont retirées
- Seules les boîtes les plus grandes sont conservées
- Les boîtes sont ajoutées sur l'arbre
  - La boîte est placé au moins aussi haut que les branches qu'elle regroupe

Approche 2 : arbre + ensembles

- **Problème : les distances sur l'arbre et la hauteur des boîtes ne sont pas dans des unités comparables**
  - 3 solutions possibles :
    - Construire l'arbre à partir des similarités
    - Calculer la hauteur des boîtes dans l'unité de l'arbre
    - "Étalonner" les distances et les similarités

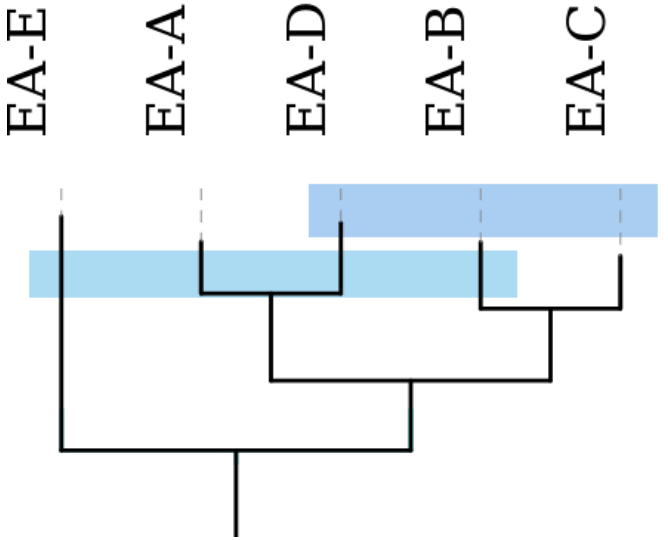
Approche 2 : arbre + ensembles



Discussion et conclusion

- **Nous avons proposé deux approches pour visualiser l'information phylogénétique au-delà des arbres**
  - En s'appuyant sur la visualisation d'ensembles
  - Résultats préliminaires
- **Les boîtes montrent des similarités observées mais n'indiquent pas leur origine**
  - En génétique des populations : dans quel sens s'est fait la migration ? De A vers B ou de B vers A ?
- **Un nombre important de boîtes contiennent beaucoup de population (plus de la moitié), ce qui est peu pertinent en génétique des populations**
  - On recherche plutôt des migrations d'une population vers une autre

Discussion et conclusion



- **Les résultats en génétique des populations restent à analyser et à interpréter**
- **Interprétation pas toujours simple**
  - Une similarité entre E, B, et l'ancêtre commun A-D
    - Similarité "à 3" difficile à interpréter en terme de migration
  - Une similarité entre D et l'ancêtre commun B-C
    - Incompatibilité chronologique avec la précédente !

Références

Auton, A., L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, et et al. (2015). 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 526(7571), 68–74.

Lamy, J. B., H. Berthelot, C. Capron, et M. Favre (2017). Rainbow boxes : a new technique for overlapping set visualization and two applications in the biomedical domain. Journal of Visual Language and Computing 43, 71–82.

Lamy, J. B. et R. Tsopra (2019). RainBio : Proportional visualization of large sets in biology. IEEE Transactions on Visualisation and Computer Graphics accepted.

Pickrell, J. K. et J. K. Pritchard (2012). Inference of population splits and mixtures from genome-wide allele frequency data. PLoS genetics 8(11), e1002967.