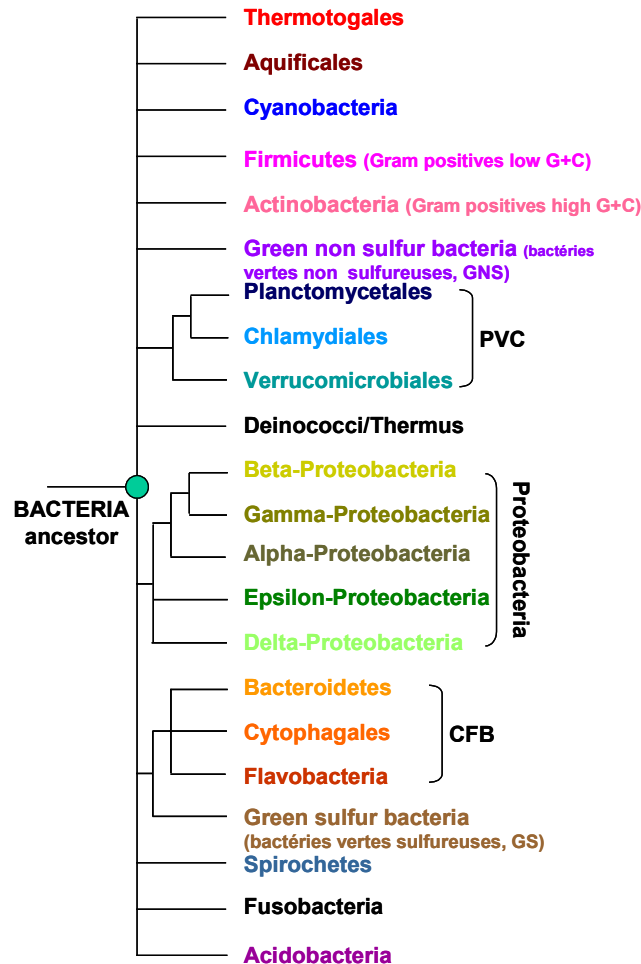


# Stratégie à suivre pour choisir un groupe d'étude et un groupe extérieur



1/ Définir le groupe d'étude de telle sorte qu'il représente une lignée taxonomique réelle. Ce doit être un groupe monophylétique.  
 ⇒ Choisir les séquences parmi l'ensemble de ces lignées de telle sorte que l'ensemble des lignées soit représentées.

2/ Définir un groupe extérieur constitué de l'ensemble des lignées de même niveau taxonomique que le groupe d'étude.  
 ⇒ Choisir les séquences parmi l'ensemble de ces lignées de telle sorte que l'ensemble des lignées soit représentées.

Groupes d'études valides: Cyanobacteria, Thermotogales, delta-Proteobacteria, PVC, Proteobacteria, beta-Proteobacteria+gamma-Proteobacteria, Bacteria, Archaea, organismes cellulaires (i.e. Archaea, Bacteria et Eucaryotes)...

Groupes d'études non valides: Cyanobacteria + Firmicutes, Alpha-Proteobacteria + Beta-Proteobacteria...

## Exemple n°1

Si le groupe d'étude est les Firmicutes, les lignées taxonomiques de niveau équivalents sont les autres phyla bactériens (i.e. Thermotogales, Aquificales, Cyanobacteria, Proteobacteria, PVC...)  
 ⇒ Solution: choisir des séquences provenant de chacun de ces phyla

## Exemple n°2

Si le groupe d'étude est les alpha-Proteobacteria, les lignées taxonomiques de niveau équivalents sont les autres Proteobacteria (i.e. beta, delta, epsilon et gamma)  
 ⇒ Solution: choisir des séquences provenant de chacun de ces groupes

## Exemple n°3

Si le groupe d'étude est les gamma-Proteobacteria, les lignées taxonomiques de niveau équivalents sont les autres beta-Proteobacteria  
 ⇒ Solution: choisir des séquences provenant de ce groupe

## Exemple n°4

Si le groupe d'étude est les Bacteria, les lignées taxonomiques de niveau équivalents sont les Archaea et les Eucaryotes  
 ⇒ Solution: choisir des séquences provenant de chacun de ces groupes

## Exemple n°5

Si le groupe d'étude est les Bacteria et qu'aucune séquence d'Archaea et/ou d'Eucaryotes  
 ⇒ Solution: travailler sur un arbre non raciné

## Stratégie à suivre pour sélectionner les séquences

Il est impératif de choisir les séquences qui vont constituer le groupe d'études de manière à représenter la diversité de ce groupe, c'est-à-dire provenant de l'ensemble des sous-groupes le constituant. Voir le site du tree of life (<http://www.tolweb.org/tree/>) ou du NCBI pour identifier les sous-groupes constituant le groupe extérieur (<http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi>).

Suivre la même stratégie pour choisir les séquences constituant le groupe extérieur.

### Exemple

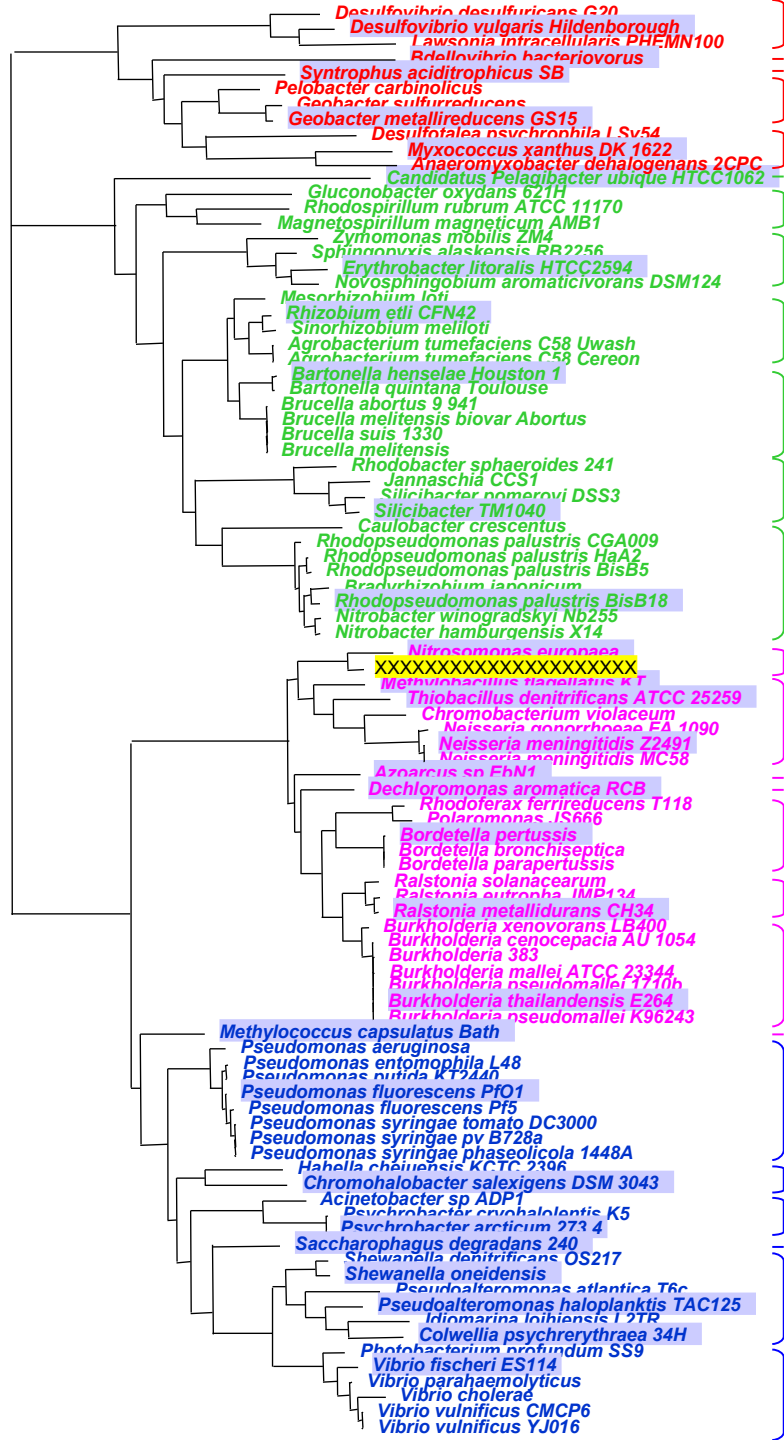
Le groupe d'étude est le groupe rose + le groupe bleu, la séquence inconnue est indiquée par un fond jaune, le groupe extérieur est le groupe vert + groupe rouge.

⇒ Pour constituer le groupe d'étude, prendre ~15-20 séquences dans chacun des sous-groupes désignés par une accolade rose ou bleue

⇒ Pour constituer le groupe extérieur, prendre ~5-10 séquences dans chacun des sous-groupes désignés par une accolade rouge ou verte

Un bon choix de séquences correspondrait aux séquences indiquées par un fond violet clair. La phylogénie résultante montrera la séquence **XXXXXXXXXXXX** nichée au sein du groupe rose, à proximité de la séquence de *Nitrosomonas europaea*. La conclusion sera que la séquence d'intérêt provient d'un organisme appartenant au groupe rose, et probablement au même sous-groupe que *Nitrosomonas europaea*.

⇒ Les pages suivantes illustrent quelques cas d'école. Les exemples présentés reflètent les situations non conventionnelles les plus classiques auxquelles vous serez confrontés, mais ne constituent en aucun cas une liste exhaustive. En effet, il existe une infinité de situations, mais la plupart peuvent être interprétées logiquement.



## Informations utiles

Actuellement, les génomes complets des groupes suivants sont accessibles au NCBI.

### Bacteria

Alpha-Proteobacteria	79
Beta-Proteobacteria	48
Gamma-Proteobacteria	145
Epsilon-Proteobacteria	19
Delta-Proteobacteria	18
Firmicutes	129
Actinobacteria	48
Acidobacteria	2
Aquificales	1
Thermotogales	6
Cyanobacteria	30
Bacteroidetes/Chlorobi	17
PVC	13
Chloroflexi	8
Thermus/Deinococci	4
Fusobacteria	1
Spirochaetes	9
Other	2

### Archaea

Crenarchaeota	15
Euryarchaeota	33
Nanoarchaeota	1

Bien évidemment, de nombreuses séquences provenant d'autres bactéries ou archéobactéries sont accessibles dans SWISS-PROT ou dans *nr*.

Néanmoins ces informations peuvent être utiles.

En effet, si une étude conduite avec un groupe d'étude constitué des gamma-Proteobacteria révèle que seuls quelques représentant de ce groupe possèdent un homologue alors que plus de 145 génomes complets ont été séquencés, la plus grande prudence est nécessaire pour interpréter les phylogénies obtenues.

Elles peuvent révéler que des transferts horizontaux de gènes ou que des pertes massives de gènes ont eus lieu au sein de ce groupe.

# Erreurs classiques lors de la sélection des séquences

Exemple: Choix de séquences ne représentant pas la diversité du groupe d'étude

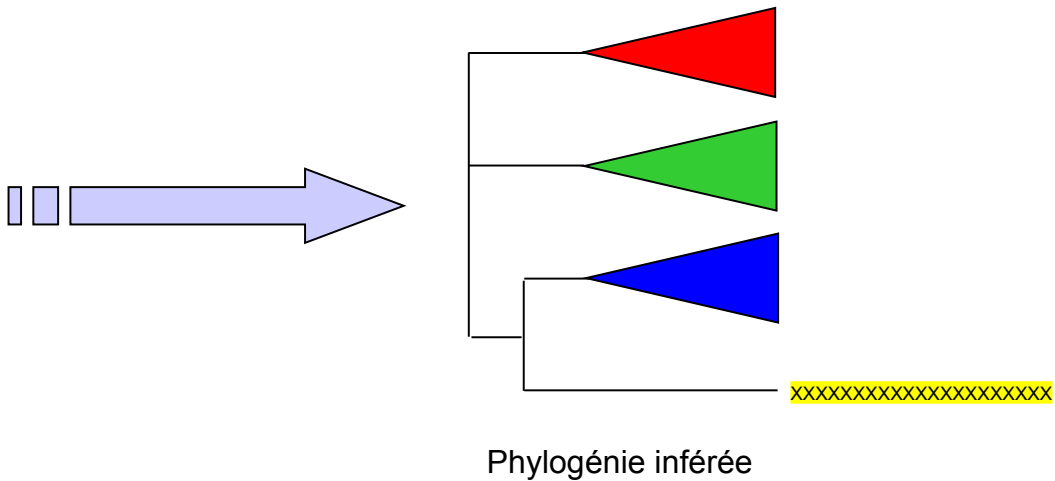
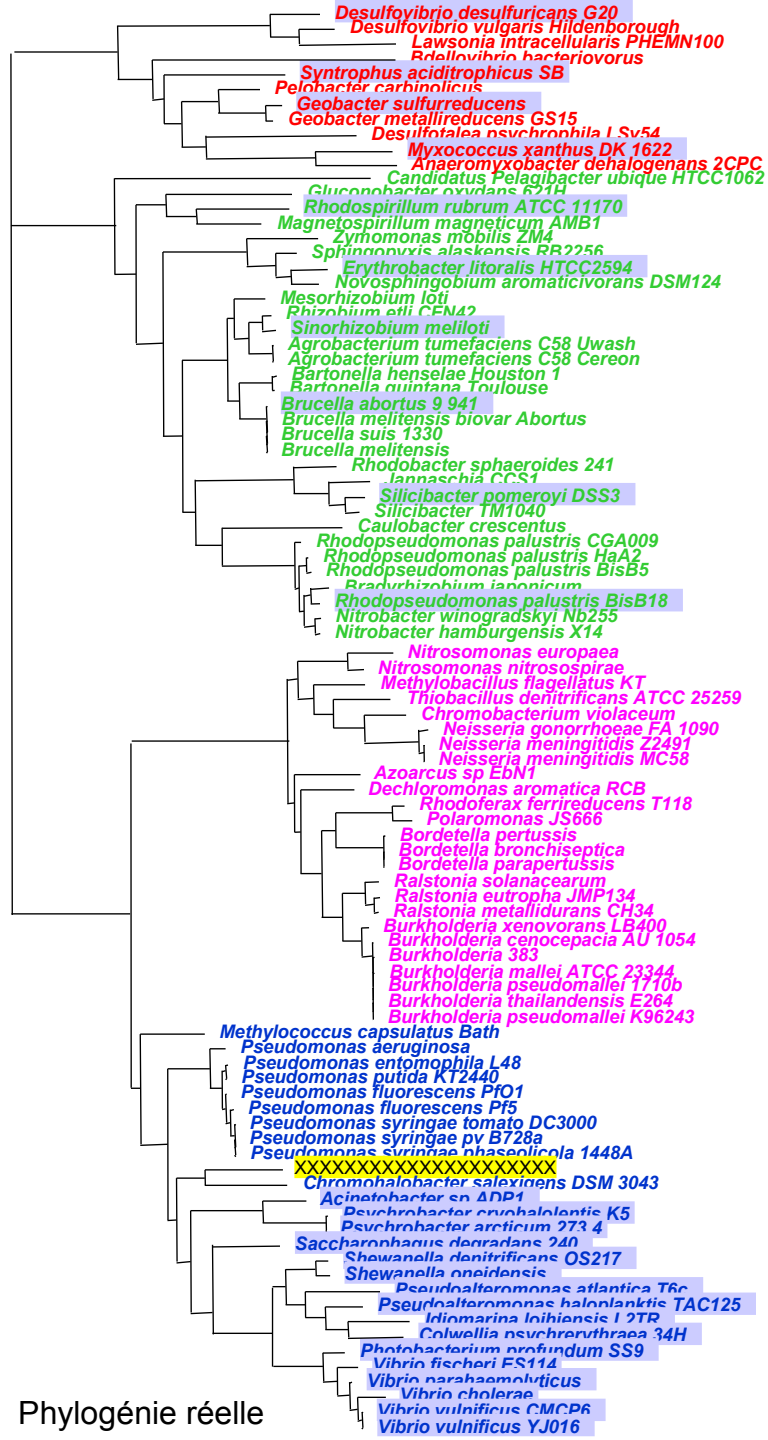
Le groupe d'étude est le groupe rose + le groupe bleu, la séquence inconnue est indiquée par un fond jaune, le groupe extérieur est le groupe vert + groupe rouge.

Un mauvais choix de séquences représentant le groupe d'étude est indiqué par un fond violet clair.

La phylogénie résultante montrera la séquence **XXXXXXXXXXXX** émergeant entre le groupe extérieur et le groupe d'étude.

⇒ L'émergence de la séquence d'intérêt entre le groupe d'étude et le groupe extérieur traduit très fréquemment (i) un mauvais choix des séquences représentant le groupe d'étude et/ou le groupe extérieur, ou (ii) une mauvaise définition du groupe d'étude et/ou du groupe extérieur.

⇒ Solution: Refaire une sélection plus rationnelle des séquences et/ou redéfinir les groupes d'étude et extérieur.



Phylogénie inférée

Phylogénie réelle

## Cas difficiles d'interprétation (I)

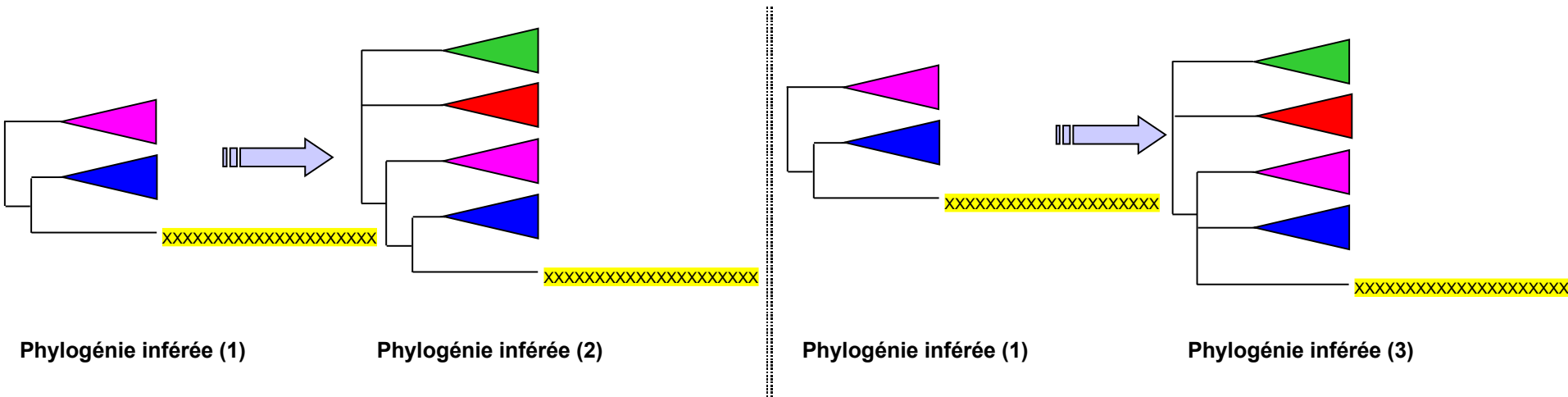
### Exemple:

Le groupe d'étude est le groupe bleu, groupe extérieur est le groupe rose, la séquence inconnue est indiquée par un fond jaune. Malgré un bon choix de séquences représentant le groupe d'étude et le groupe d'étude, la phylogénie obtenue montre la séquence **XXXXXXXXXXXX** émergeant entre le groupe extérieur et le groupe d'étude dans la phylogénie inférée (1).

Conclusion: On ne peut pas conclure définitivement que la séquence appartient bien au groupe d'étude. Pour pouvoir conclure, il faut élargir le groupe extérieur et refaire une phylogénie.

⇒ Si la phylogénie inférée avec un groupe extérieur plus large ressemble à la phylogénie inférée (2), c'est-à-dire qu'elle montre que la séquence **XXXXXXXXXXXX** est reliée spécifiquement au groupe d'étude (sans émerger à l'intérieur), il est légitime de conclure que la séquence **XXXXXXXXXXXX** est proche du groupe d'étude, même si on ne peut pas conclure qu'elle appartient au groupe d'étude. Il s'agit soit du représentant d'une lignée inconnue appartenant au groupe d'étude, soit du représentant d'une lignée inconnue proche parente du groupe d'étude.

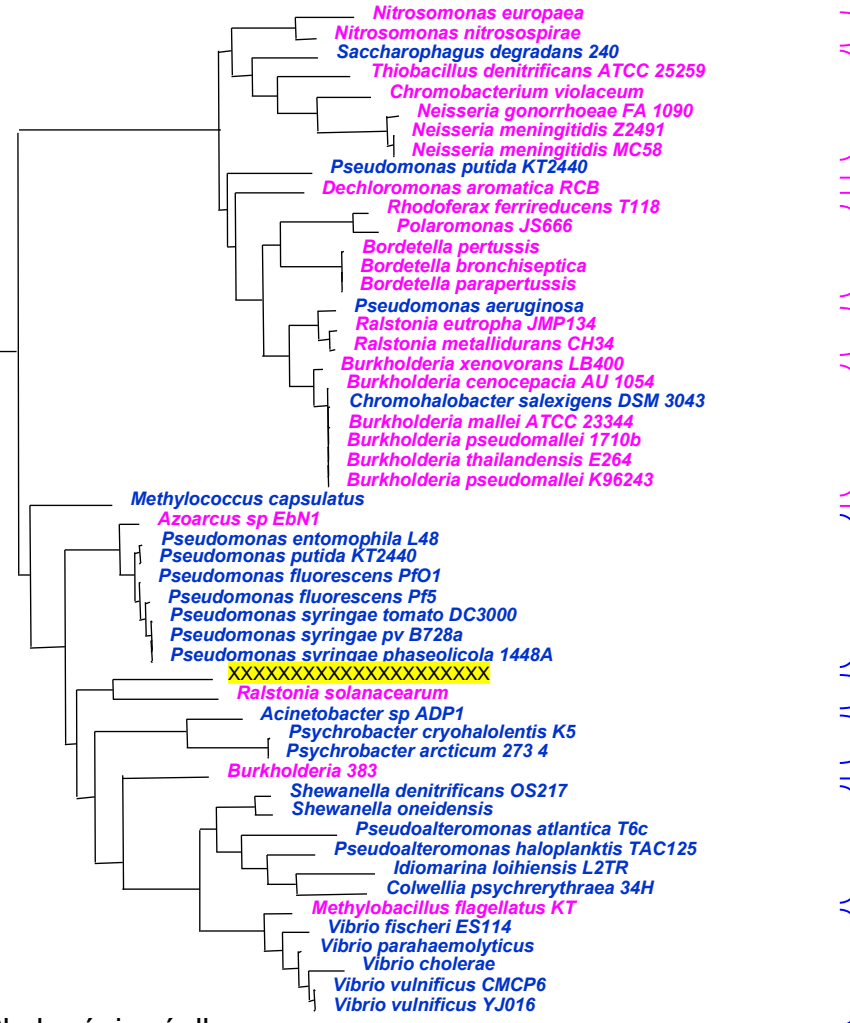
⇒ Si la phylogénie inférée avec le groupe extérieur plus large ressemble à la phylogénie inférée (3), c'est-à-dire qu'elle montre que la séquence **XXXXXXXXXXXX** n'est pas spécifiquement reliée au groupe d'étude, il est légitime de conclure qu'elle provient d'un représentant d'une lignée inconnue non spécifiquement apparentée au groupe d'étude.



## Cas difficiles d'interprétation (II)

La phylogénie obtenue montre des mélanges entre (i) les séquences des lignées représentant le groupe d'étude, (ii) les séquences des lignées du groupe extérieur et du groupe d'étude, (iii) les séquences des lignées représentant le groupe extérieur.

⇒ Explications: La séquence étudiée est sujette à des transferts horizontaux de gènes. Il n'est donc pas possible de conclure formellement quant à l'appartenance de la séquence **XXXXXXXXXXXX** au groupe d'étude. La seule conclusion possible dans le cas illustré ci-dessous est que la séquence est proche de *Ralstonia solanacearum* sans qu'il soit possible de dire si elle appartient au groupe rose ou au groupe bleu.

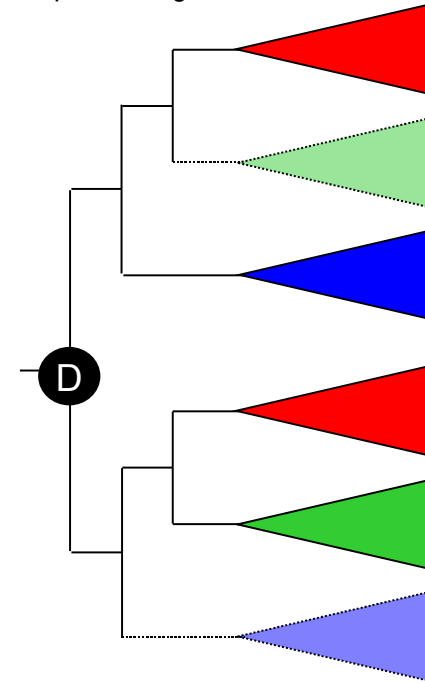
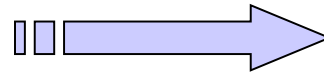
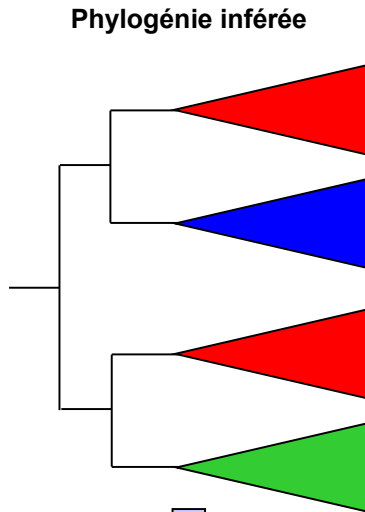
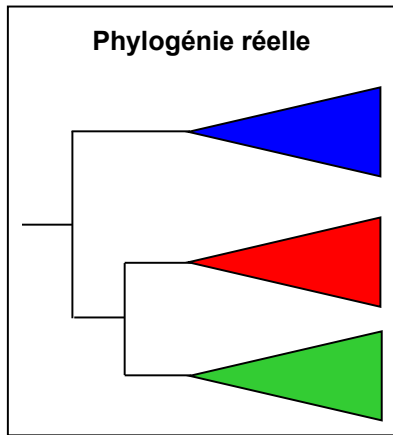


Phylogénie réelle

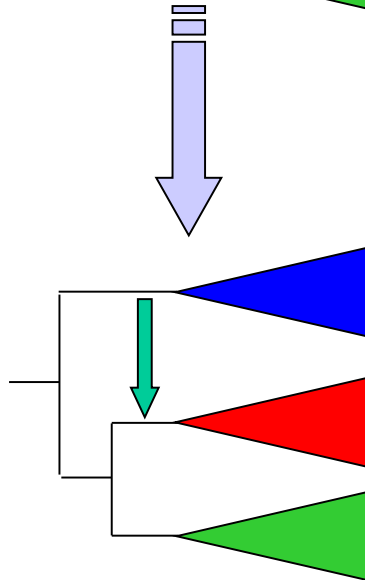
## Cas difficiles d'interprétation (III)

La phylogénie inférée montre des mélanges entre les groupes étudiés (se référer aux phylogénies de référence pour visualiser les relations de parenté conventionnelles entre les différents groupes taxonomiques).

Ceci peut indiquer l'existence de transferts horizontaux de gènes et/ou de duplications de gènes suivies de pertes de gènes.



**Histoire évolutive réelle (duplication (D) suivie de pertes différentielles dans les différentes lignées)**



**Histoire évolutive réelle transfert horizontal de gènes du groupe bleu vers certains membres du groupe rouge**