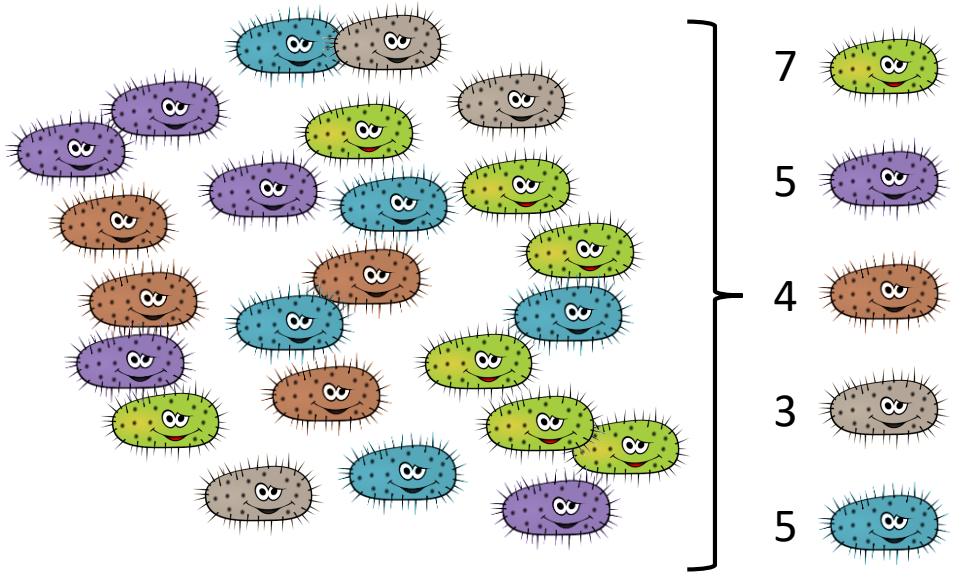
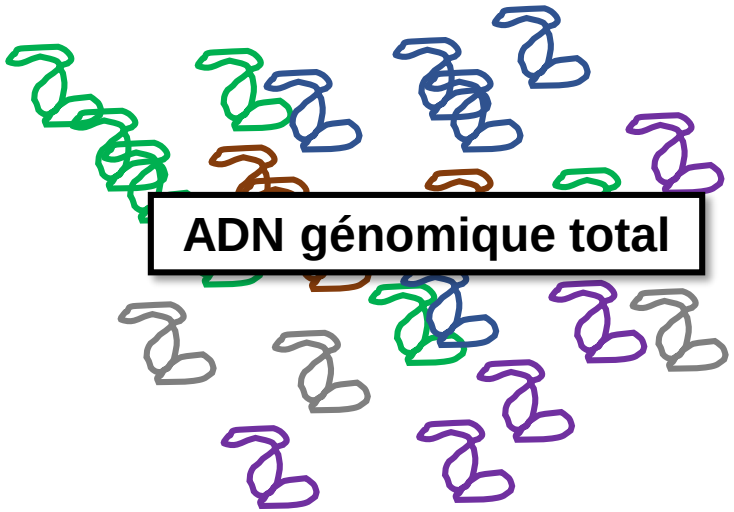
Analyse de variants exacts d'amplicons 16S avec **dada2** + **phyloseq**

Jeff Gauthier
Club Bioinfo IBIS
16 octobre 2018

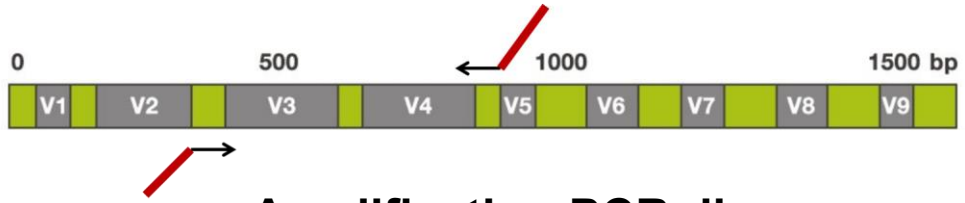


Communauté microbienne

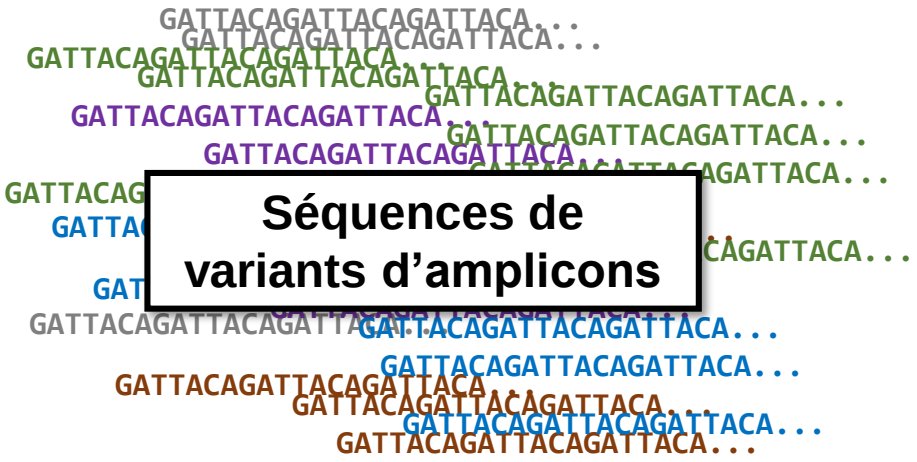
Extraction d'ADN



Amplification PCR d'un gène marqueur universel
(ex. Gène de l'ARNr 16S)

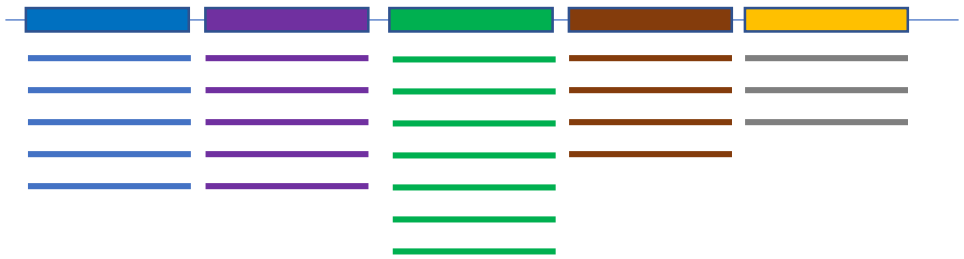


Séquençage massivement parallèle



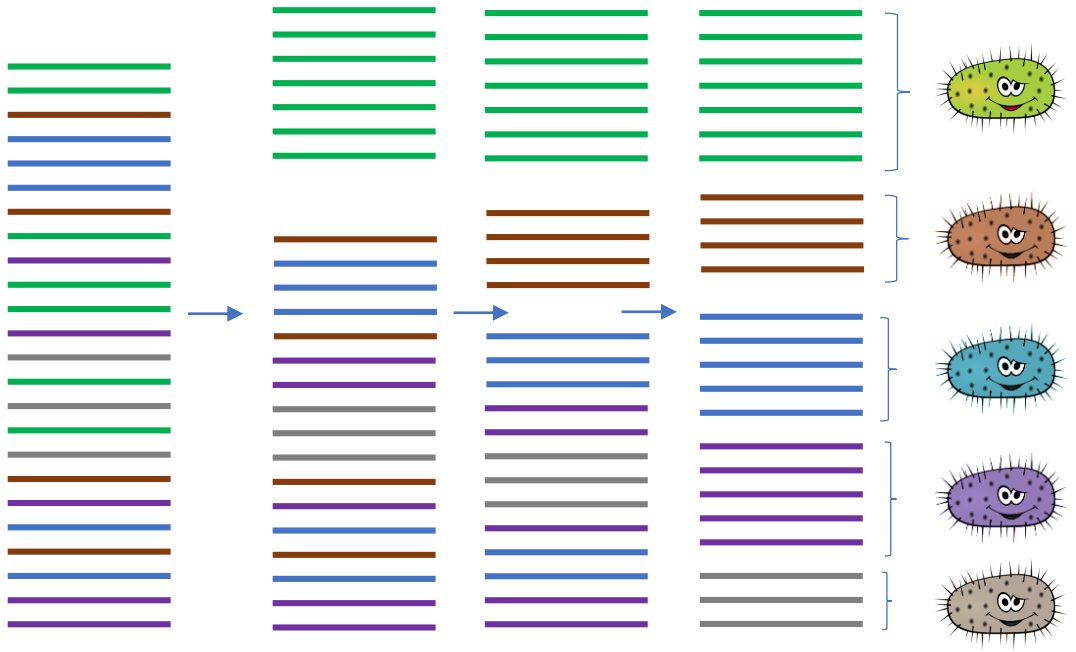
Le défi : reconstituer la communauté originale sans *a priori*

BD



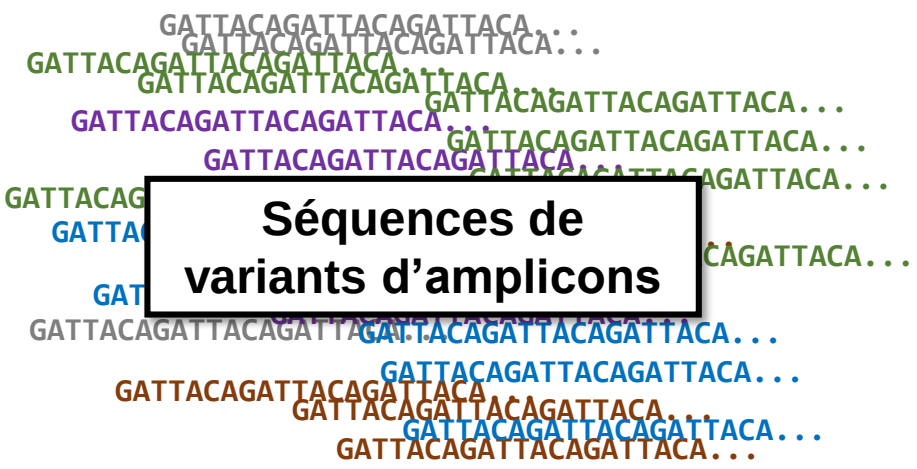
Approche 1
Alignement des seqs sur une BD de taxa connus (basé sur une référence)

Approche 2
Regroupement des séquences similaires AVANT annotation (classification *de novo*)



Le défi : reconstituer la communauté originale sans *a priori*

Approche 1
Alignement des seqs sur une BD de taxa connus (basé sur une référence)



Approche 2
Regroupement des séquences similaires AVANT annotation (classification *de novo*)

Dans les deux cas...

Seuil d'identité pour classifier les séquences

(97%)

ex. QIIME et Mothur

Matière à réflexion...

Updating the 97% identity threshold for 16S ribosomal RNA OTUs

Robert C Edgar ✉

Bioinformatics, Volume 34, Issue 14, 15 July 2018, Pages 2371–2375, https://doi.org/10.1093/bioinformatics/bty113

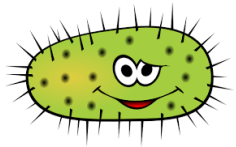
Published: 28 February 2018 **Article history** ▾

La problématique des seuils d'identité

1) Problème des "triplets"

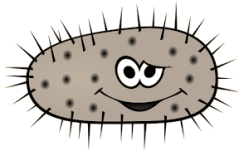
Edgar RC (2018) Bioinformatics

OTU 1 (97% id.)
Ref: GATTTCAGATTACA...



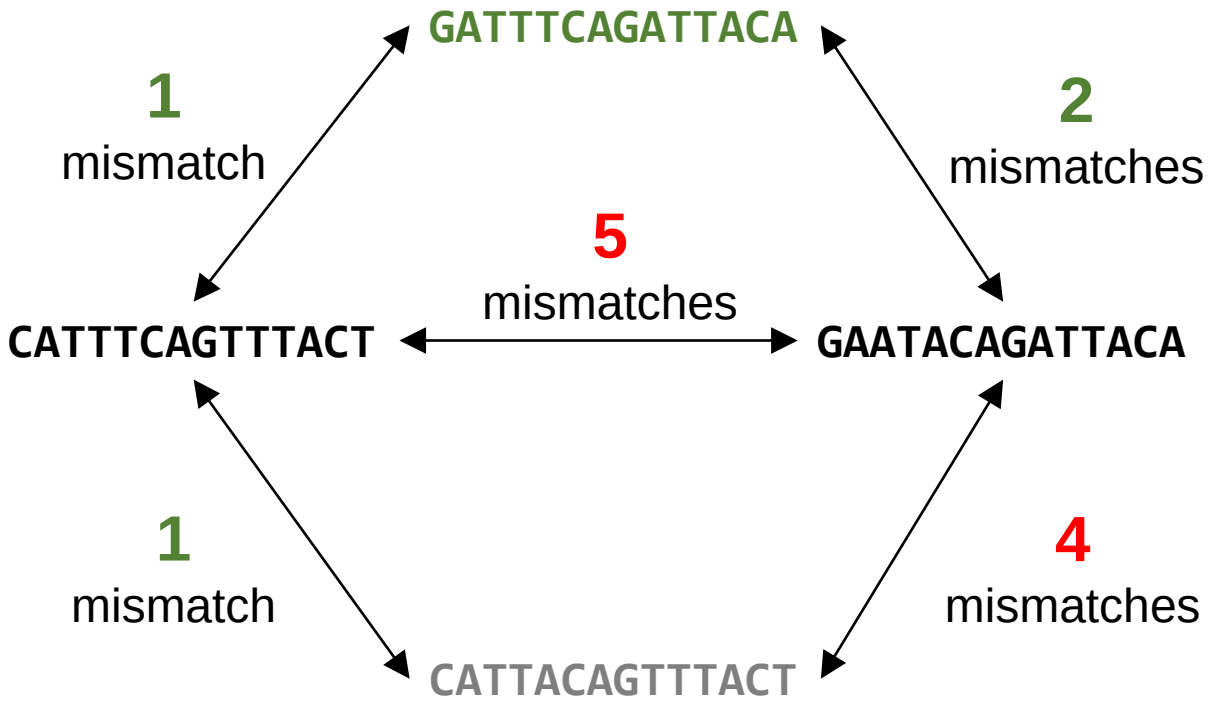
GATTACAGATTACA
GATTTCAGATTACA
GATTTCAGCTTACA
GATTTGAGATTACA
GTTTTCAGTTTACA
GAATACAGATTACA
CATTTCAGTTTACT

OTU 2 (97% id.)
Ref: CATTACAGTTTACT...



CATTTCCGTTTACT
CATTACACTTTACT
CATATCAGTTTACT

Seuil permis: 3 *mismatchs* ou moins par rapport à la référence
(ou au centroïde si de novo)

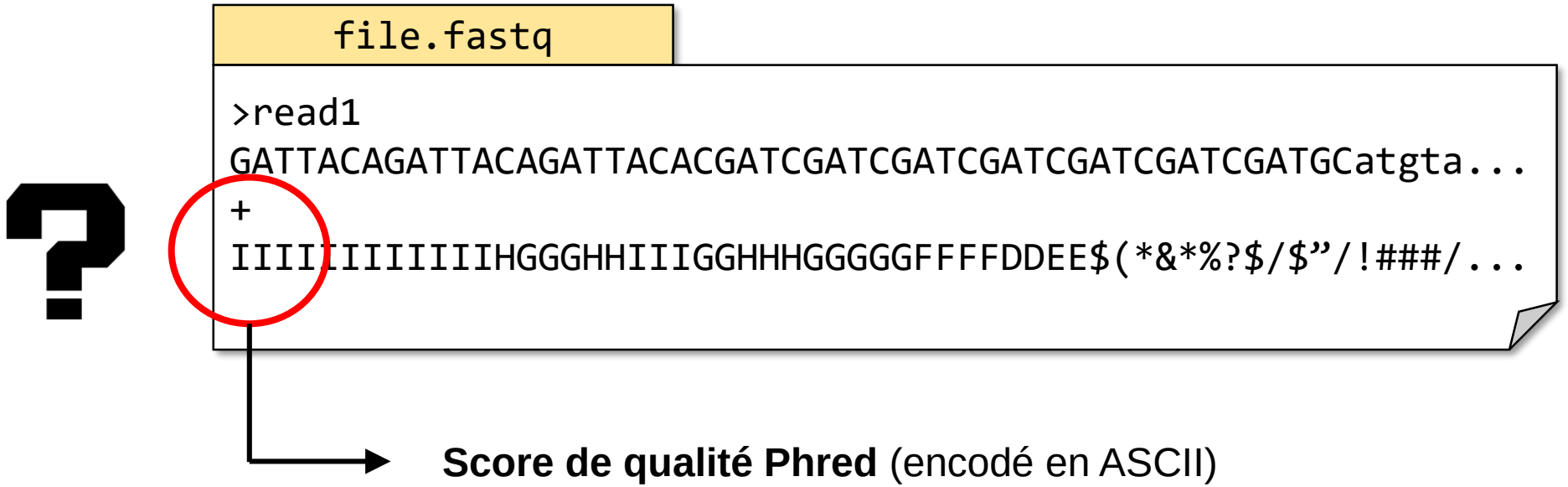


L'ordre de classification est important !

La problématique des seuils d'identité

2) Sous-utilisation des données de qualité

Edgar RC (2018) Bioinformatics



Score de qualité Phred (encodé en ASCII)

$$Q = -10 \times \log_{10} P$$

Ex.: un score Phred de **40** = **1 chance sur 10 000** qu'un nucléotide soit incorrectement identifié

La problématique des seuils d'identité

2) Sous-utilisation des données de qualité

Edgar RC (2018) Bioinformatics

file.fastq

```
>read1
GATTACAGATTACAGATTACACGATCGATCGATCGATCGATCGATCGATGCatgta...
+
IIIIIIIIIIIIHGGGHHIIIGGHHHGGGGGFFFFDDEE$(*&*%?$/$"/!###/...
```

Une *read* de **300** pb avec un score **Q** moyen de **30**:

(1/1000) + (1/1000) + ... (298 fois) = **30% de chances qu'il ait >= 1 erreur**

P(>= 2 erreurs/read) = 30% x 30% = **9%**
P(>= 3 erreurs/read) = 30% x 30% x 30% = **2.7%**
P(>= 4 erreurs/read) = 30% x 30% x 30% x 30% = **0.81%**

P(>= 9 erreurs/read) = (30%)^9 = 0.00196%

9 erreurs / 300 bp
= variation permise avec un seuil de 97% id.

La problématique des seuils d'identité

3) Perte de résolution taxonomique

Edgar RC (2018) Bioinformatics

Famille	Genre	Espèce	Souche	Séquence (supposons 100 pb)
Enterobacteriaceae	Escherichia	coli	K12	...GATTACAGATTACAGATTACAGATTACAGATTACAGATTA...
Enterobacteriaceae	Escherichia	coli	O157:H7	...GATTACAGATTACAGATTACAGAATACAGATTACAGATTA...
Enterobacteriaceae	Escherichia	albertii	-	...GATTAAAGATTACAGATTACAGAATACAGAGTACAGATTA...
Enterobacteriaceae	Enterococcus	faecalis	-	...GATTAAAGTTTACAGATTAGAGAATTCAGAGTACAGATTA...
Enterobacteriaceae	Citrobacter	freundii	-	...CATTTAAGTTTACAGATTAGAGATTTCAGAGTACAGATTA...

Si 100% ID (aucun mismatch permis)

La problématique des seuils d'identité

3) Perte de résolution taxonomique

Edgar RC (2018) Bioinformatics

Famille	Genre	Espèce	Souche	Séquence (supposons 100 pb)
Enterobacteriaceae	Escherichia	coli	?	...GATTACAGATTACAGATTACAGATTACAGATTACAGATTA...
Enterobacteriaceae	Escherichia	coli	?	...GATTACAGATTACAGATTACAGAATACAGATTACAGATTA...
Enterobacteriaceae	Escherichia	albertii	-	...GATTAAAGATTACAGATTACAGAATACAGAGTACAGATTA...
Enterobacteriaceae	Enterococcus	faecalis	-	...GATTAAAGTTTACAGATTAGAGAATTCAGAGTACAGATTA...
Enterobacteriaceae	Citrobacter	freundii	-	...CATTTAAGTTTACAGATTAGAGATTTCAGAGTACAGATTA...

Si 99% ID (1 mismatch permis)

La problématique des seuils d'identité

3) Perte de résolution taxonomique

Edgar RC (2018) Bioinformatics

Famille	Genre	Espèce	Souche	Séquence (supposons 100 pb)
Enterobacteriaceae	Escherichia	?	?	...GATTACAGATTACAGATTACAGATTACAGATTACAGATTA...
Enterobacteriaceae	Escherichia	?	?	...GATTACAGATTACAGATTACAGAATACAGATTACAGATTA...
Enterobacteriaceae	Escherichia	?	-	...GATTAAAGATTACAGATTACAGAATACAGAGTACAGATTA...
Enterobacteriaceae	Enterococcus	faecalis	-	...GATTAAAGTTTACAGATTAGAGAATTCAGAGTACAGATTA...
Enterobacteriaceae	Citrobacter	freundii	-	...CATTTAAGTTTACAGATTAGAGATTTCAGAGTACAGATTA...

Si 98% ID (**2** mismatches permis)

La problématique des seuils d'identité

3) Perte de résolution taxonomique

Edgar RC (2018) Bioinformatics

Famille	Genre	Espèce	Souche	Séquence (supposons 100 pb)
Enterobacteriaceae	?	?	?	...GATTACAGATTACAGATTACAGATTACAGATTACAGATTA...
Enterobacteriaceae	?	?	?	...GATTACAGATTACAGATTACAGAATACAGATTACAGATTA...
Enterobacteriaceae	?	?	-	...GATTAAAGATTACAGATTACAGAATACAGAGTACAGATTA...
Enterobacteriaceae	?	?	-	...GATTAAAGTTTACAGATTAGAGAATTCAGAGTACAGATTA...
Enterobacteriaceae	Citrobacter	freundii	-	...CATTTAAGTTTACAGATTAGAGATTTCAGAGTACAGATTA...

Si 97% ID (**3** mismatches permis)

Alternative au seuil d'identité

...ne pas en utiliser!

Autrement dit...

Classifier les séquences à **100%** d'identité

Mais...

À **100%** id., on assume que les reads n'ont pas d'erreurs (***faux***).

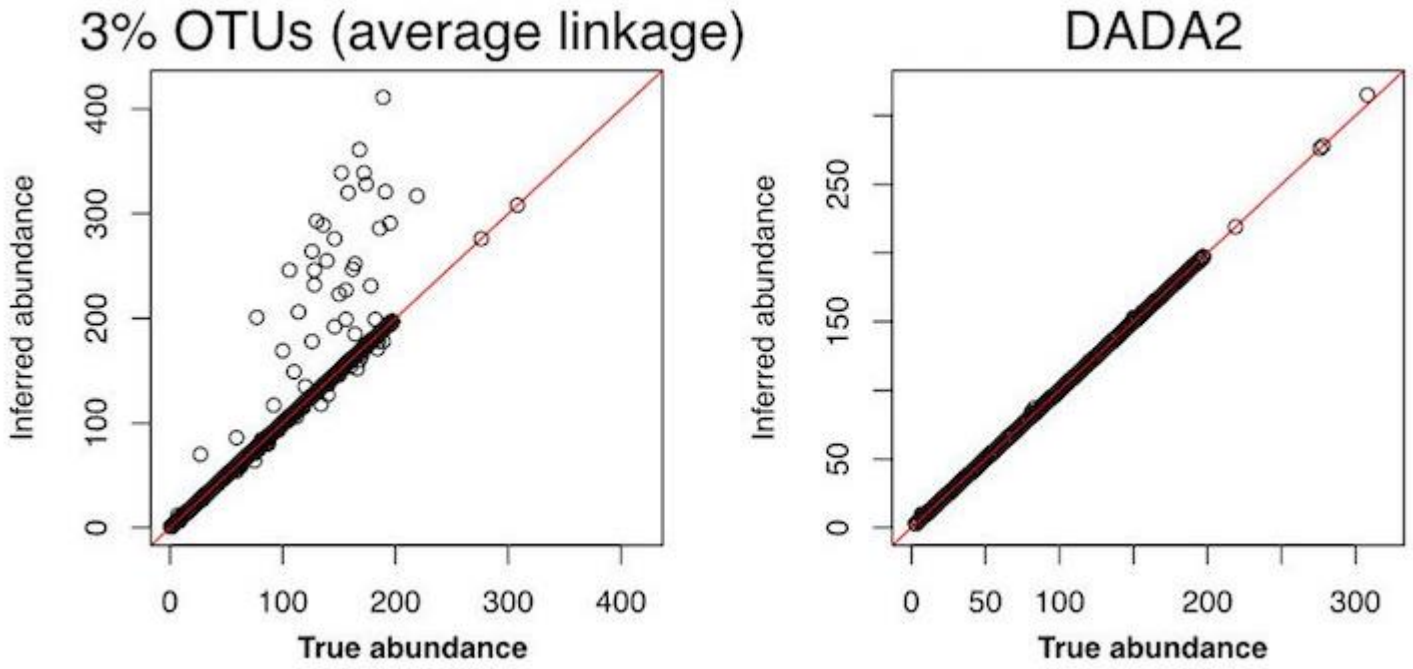
```
>read1
GATTACAGATTACAGATTACACGATCGATCGATCGATCGATCGATCGATGCatgta...
+
IIIIIIIIIIIHGGGHHIIIGGHHHGGGGGFFFFDDEE$(*&*%?$/$"/!###/...
```

Et si on pouvait utiliser cette information?

Callahan *et al.* (2016) Nature

dada2

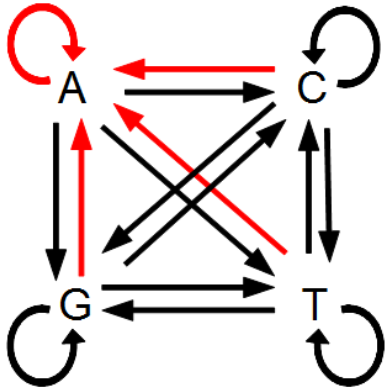
(divisive amplicon denoising algorithm v2)



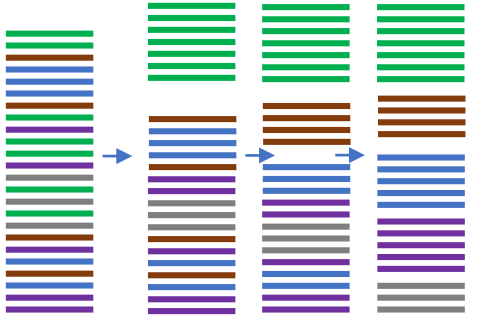
Callahan *et al.* (2016)
Nature

1) Filtration et *trimming*

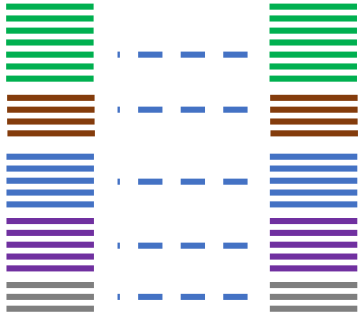
2) Apprentissage des erreurs



3) Inférence des variants

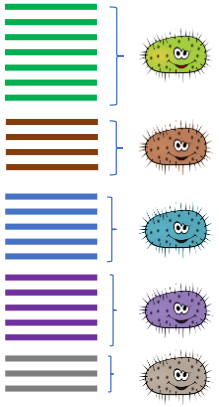


4) Assemblages des variants paired-end



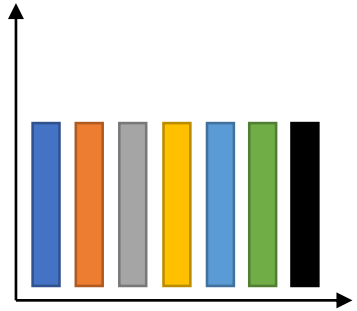
5) Filtration des "bimères"

6) Assignation taxonomique

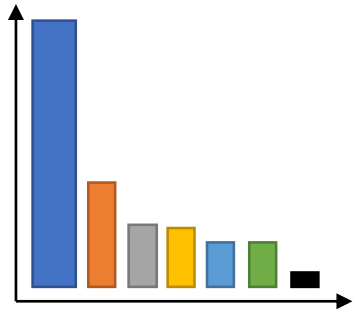


Tests de performance sur des données simulées

Callahan *et al.* (2016) Nature



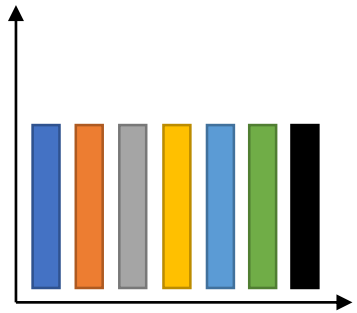
Données "BALANCED"
57 taxa ayant une distribution uniforme



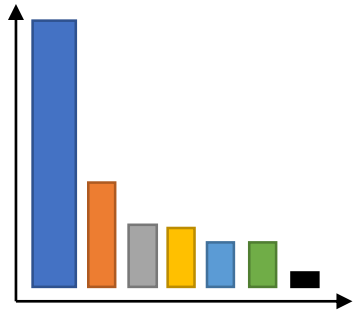
Données "EXTREME"
21 taxa dont l'abondance fluctue sous plusieurs ordres de grandeur

Tests de performance sur des données simulées

Callahan *et al.* (2016) Nature



Outil	Nb. Variants	Vrais positifs (100% id)	Faux positifs (1 err. ou +)	# Références identifiées
DADA2	87	86	1	57
Mothur (avg. linkage)	108	69	39	54
QIIME (uclust)	170	73	97	47



Outil	Nb. Variants	Vrais positifs (100% id)	Faux positifs (1 err. ou +)	# Références identifiées
DADA2	25	25	0	21
Mothur (avg. linkage)	44	37	7	23
QIIME (uclust)	36	27	9	19

Démonstration dada2 (+phyloseq)

Microbiote intestinal d'une souris femelle après sevrage

avec **RStudio**