Cycle de formation NGS
Module Métagénomique

Partie 1 : Métagénomique ciblée

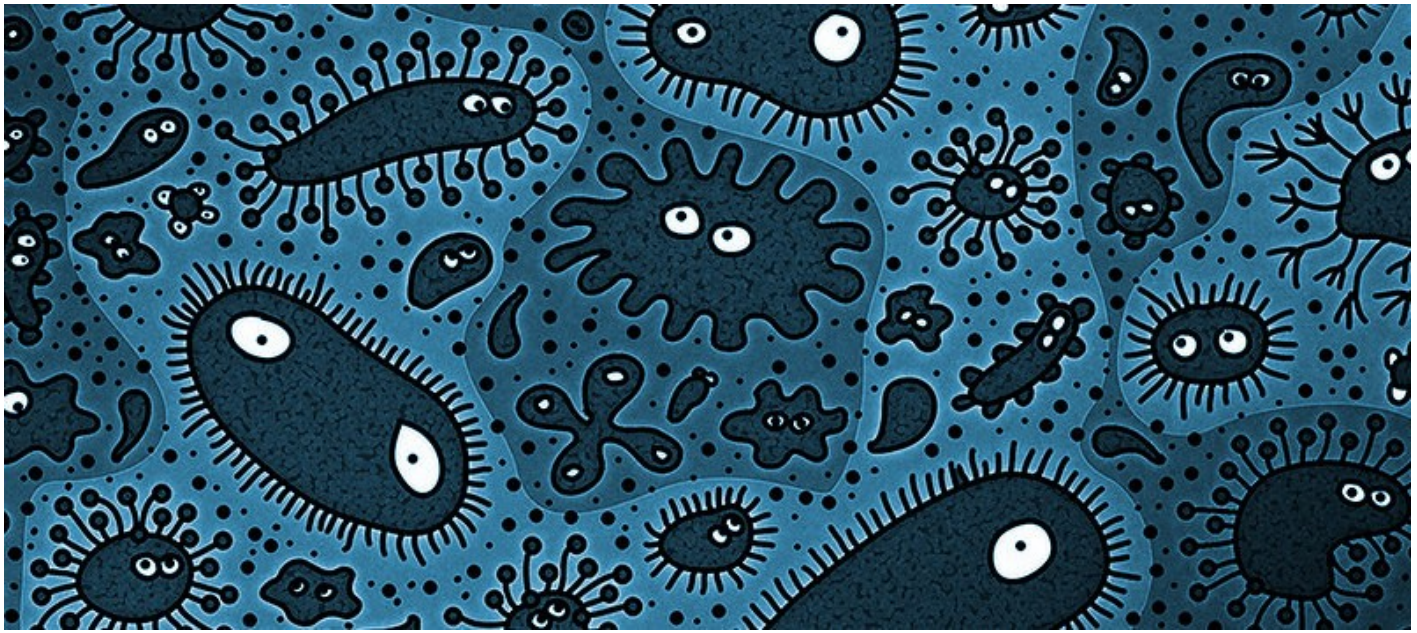
Ségolène Caboche (Université de Lille)
segolene.caboche@pasteur-lille.fr

&

Gaël Even (Gènes Diffusion)
g.even@genesdiffusion.com

26 et 27 mai 2021

Microbiote



- Le **microbiote** est l'ensemble des micro-organismes (bactéries, virus, champignons, levures) vivants dans un environnement spécifique appelé **microbiome**

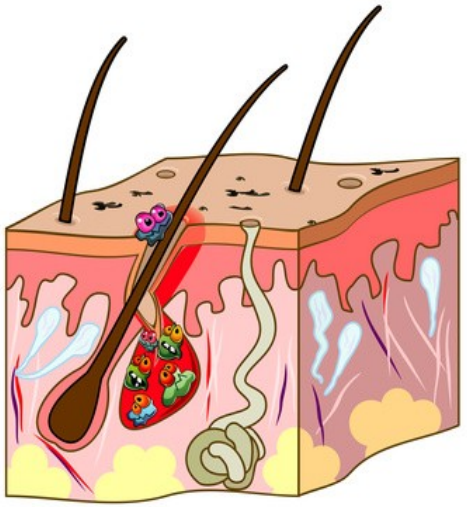
Le microbiote intestinal humain

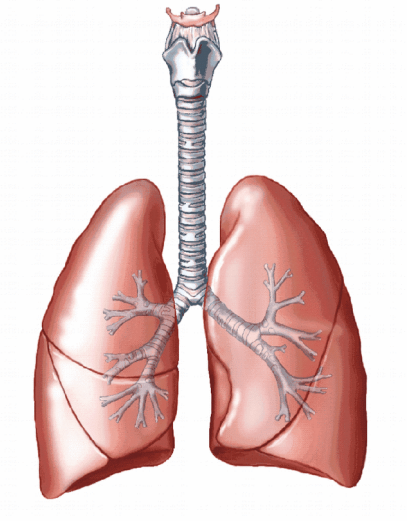
- des millions de micro-organismes formant une communauté en symbiose avec l’organisme
- nécessaire au bon fonctionnement du corps humain
- intervient dans de nombreuses voies métaboliques fondamentales
- Environ 1 kg de bactéries dans notre intestin à l'âge adulte
- Au moins 150 fois plus de gènes dans le microbiote que dans le génome humain

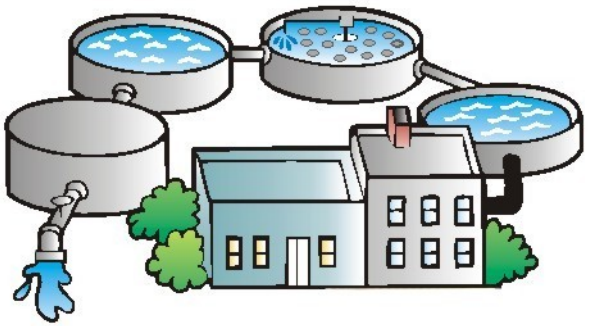
Le microbiote intestinal humain

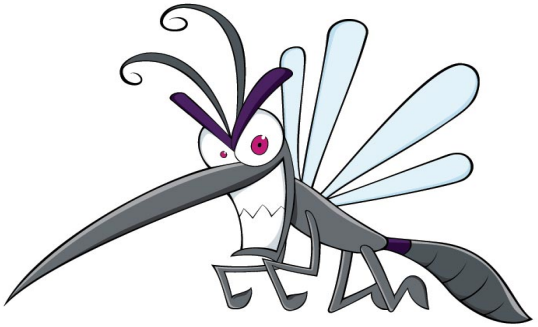
- rôle de barrière vis-à-vis d'autres agents microbiens pathogènes
- sa destruction par des antibiotiques peut être responsable par exemple des infections intestinales par *Clostridium difficile*
- Impact dans différentes pathologies :
  - Maladie de Crohn
  - Obésité
  - Diabète
  - Allergies
  - Cancer colorectal
  - ...

Autres exemples de microbiotes





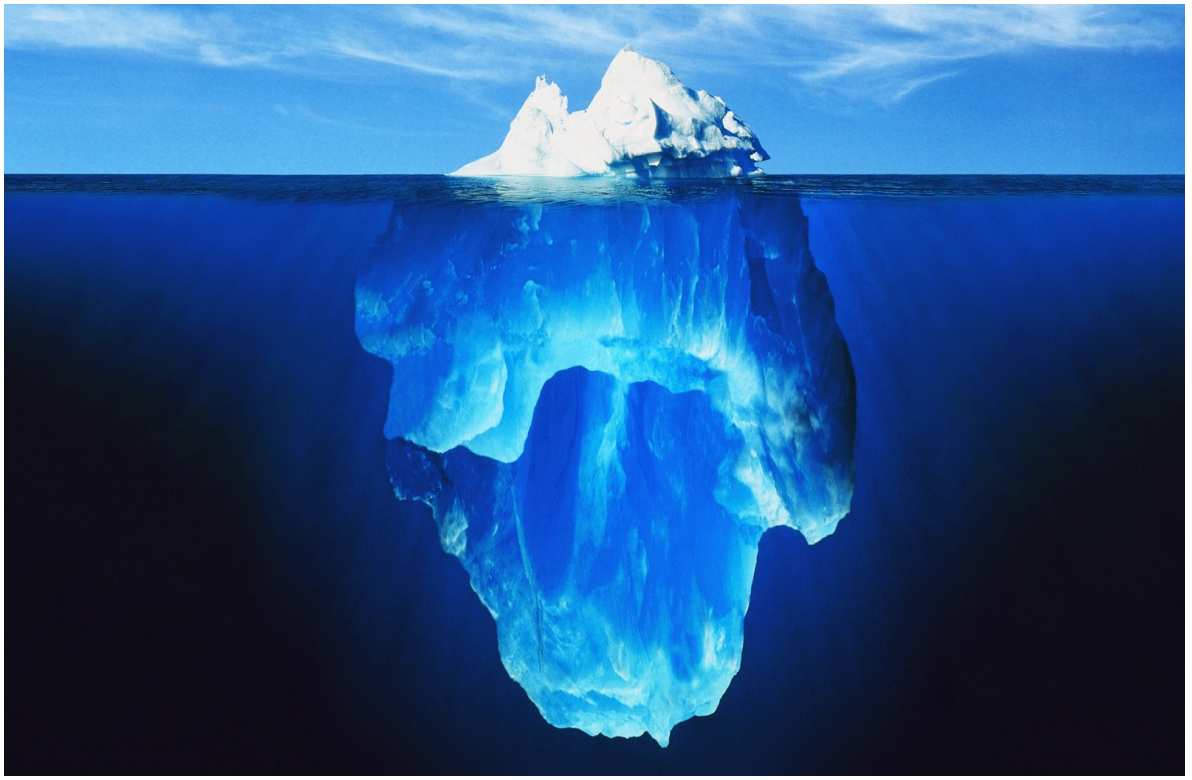




La métagénomique

- La **métagénomique** est la méthode d'étude du microbiote
- A l'inverse de la génomique qui consiste à séquencer un **unique** génome, la métagénomique séquence les génomes de **plusieurs espèces différentes** dans un milieu donné
- Une analyse typique de métagénomique donne la composition d'un microbiome (espèces présentes, leurs abondances et leurs diversités)
- C'est en partie grâce à l'évolution majeure des technologies de **séquençage haut-débit** et à la **bioinformatique**, que la métagénomique s'est si fortement développée

Pourquoi le séquençage haut-débit ?



- Seul 1% des bactéries de la plupart des environnements peut être cultivé *(Amann et al., 1990)*
- Une approche clonage + séquençage Sanger ne permet d'étudier qu'une diversité très limitée *(Sabree et al., 2009*)

Pourquoi le séquençage haut-débit ?

Le séquençage haut-débit s’affranchit du besoin de culture, et peut générer assez de séquences pour couvrir tous les organismes en présence dans plusieurs échantillons en un seul run.

Métagenomique : 3 questions

Qui ?

- Quels organismes ?
- Quelles proportions ?

www.compostinfo.com/tutorial/microbes.htm

Quoi ?

- Types de gènes
- Voies métaboliques

Comment ?

- Comparaison d'échantillons
- Corrélations avec les paramètres environnementaux

2 grands types de métagénomique

Microbiome samples

no isolation
no lab cultivation

Extract DNA

Amplicon sequencing

Whole-Genome shotgun sequencing

Identification of species and relative frequencies

Phylogenetic view of community composition

Identification of variants and polymorphisms

Functional information

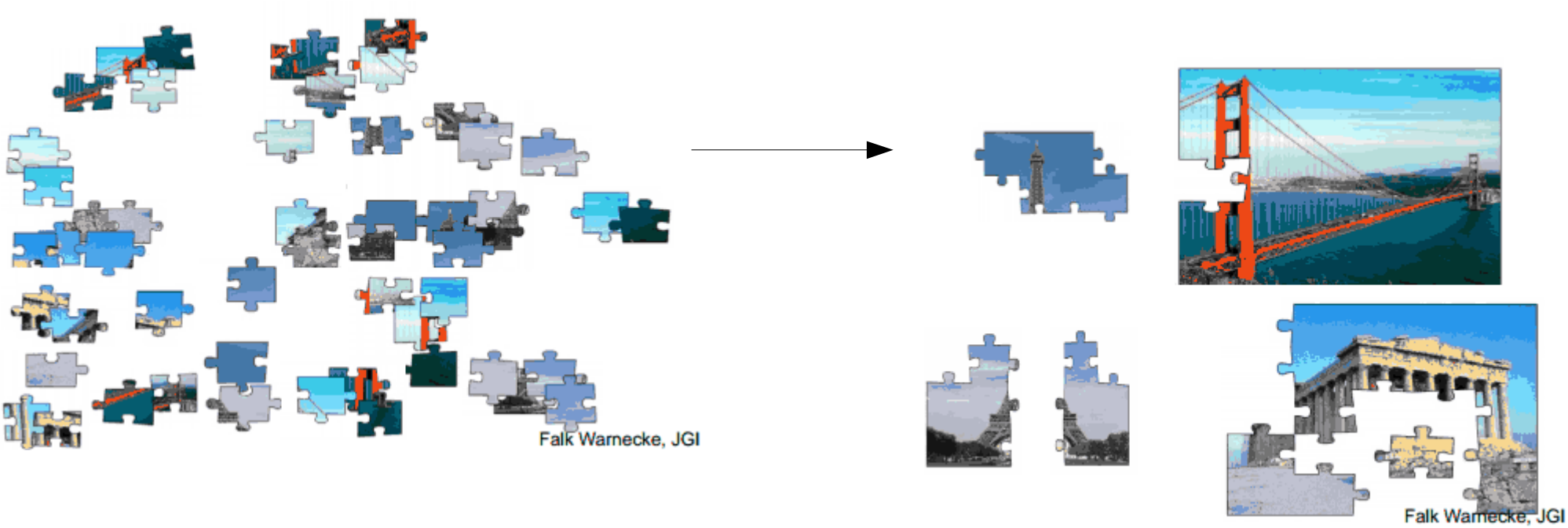
2 grands types de métagénomique

Whole Genome Shotgun

- Séquençage de l'ADN total
- Mélange de plusieurs génomes
- Pas d'*a priori*
- Information maximale
- Complexité d'analyse

2 grands types de métagénomique

Whole Genome Shotgun



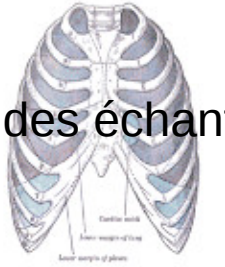
2 grands types de métagénomique

Métagénomique ciblée

- Choix d’un marqueur de la taxonomie
- 1 région spécifique séquencée et amplifiée
- Analyses plus rapides et facilitées
- Plusieurs méthodologies publiées
- Biais de choix de la région
- Biais quantitatif

La métagénomique ciblée

Plan d'expérience



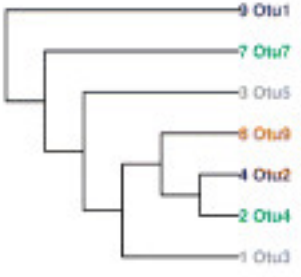
Collecte des échantillons

Extraction d'ADN

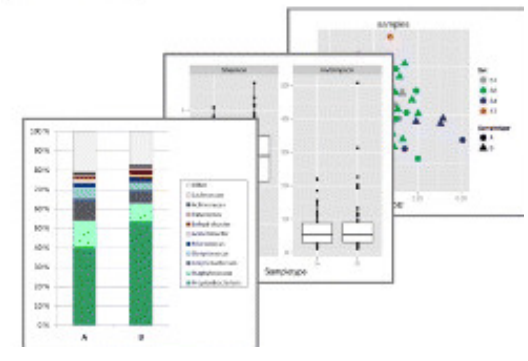
PCR

Séquençage

Analyse primaire
Assignation taxonomique



Analyse secondaire
Statistiques/Interprétation



Plan d’expérience

- Étape fondamentale de toute étude de métagénomique ciblée
- Définir **une question biologique** claire
- Il faut prendre en compte le nombre d'échantillons à séquencer, les variables à évaluer, la profondeur de séquençage, le nombre de réplicats => ils guident le choix des technologies, des protocoles techniques et des méthodes d'analyse
    - Par exemple, une étude cas/témoin doit être représentée par suffisamment d'échantillons pour être statistiquement valide
- Il est préférable de renforcer le plan d'expérience par l'**ajout de réplicats biologiques à un surséquençage** qui n'apportera pas plus d'information utile à l'analyse
- Intervention d'un bioinformaticien et/ou biostatisticien => validation d’un plan d'expérience robuste pour répondre à la question biologique initialement posée

Plan d’expérience

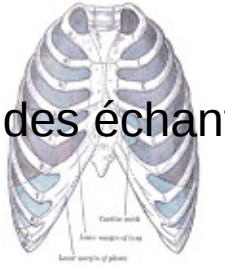
- Évaluation des connaissances actuelles sur les microbiotes d'intérêt (conditions environnementales, contaminants potentiels, caractéristiques cellulaires, …) pour guider le choix de protocoles techniques adaptés
- Déterminer les métadonnées à recueillir pour chaque échantillon qui permettront de les replacer dans un contexte biologique

=> ATTENTION : l’étape du plan d’expérience ne doit pas être sous-estimée

La métagénomique ciblée

Plan d'expérience

Collecte des échantillons

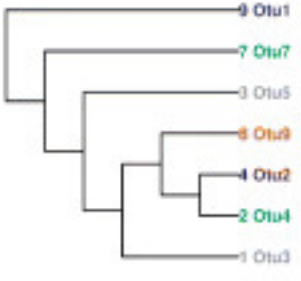


Extraction d'ADN

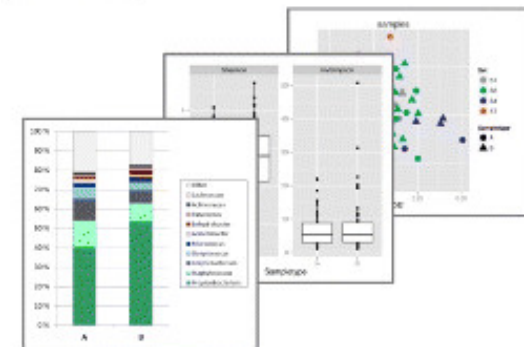
PCR

Séquençage

Analyse primaire
Assignation taxonomique



Analyse secondaire
Statistiques/Interprétation



Collecte des échantillons

- Chaque prélèvement doit être représentatif du milieu étudié
- Beaucoup de questions :

  - Peut-on transposer les conclusions tirées de l'étude d'un microbiote fécal à une flore intestinale?

  - Un sol n'étant pas homogène, quelle stratégie d'échantillonnage mettre en place pour être le plus représentatif de sa composition microbienne globale ?

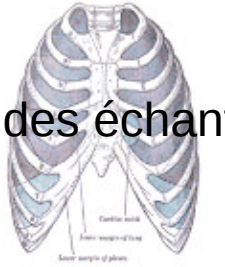
- L'idéal serait d'extraire l'ADN sitôt l'échantillon prélevé car la matrice étant vivante, elle est sujette à évolution après prélèvement
- Cependant, lors d'études à grande échelle, il est souvent nécessaire de **transporter et de stocker** les échantillons (par congélation à -80°C et ajout de préservateur par exemple) pour figer la composition du microbiote

Collecte des échantillons

- Les méthodes de préservation peuvent tout de même altérer la composition des échantillons : certaines études ont décrit une composition différente entre un échantillon congelé (avec ou sans préservateur) et un échantillon frais [Choo et al. 2015]
- Effectuer des prélèvements dans des conditions non stériles peut augmenter le risque d'introduction d'ADN exogène contaminant dans les échantillons d'intérêt

La métagénomique ciblée

Plan d'expérience



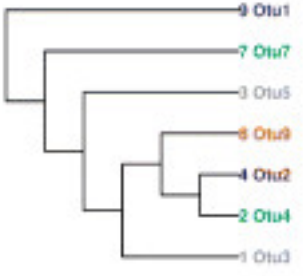
Collecte des échantillons

Extraction d'ADN

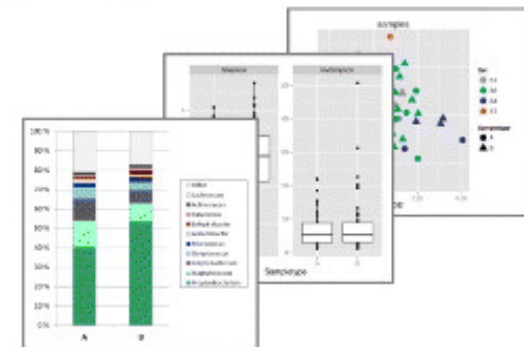
PCR

Séquençage

Analyse primaire
Assignation taxonomique



Analyse secondaire
Statistiques/Interprétation



Extraction d’ADN

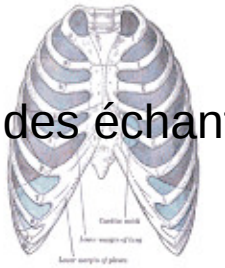
- L’étape d'extraction et de purification va permettre d'accéder à l'ADN des organismes en présence, idéalement **sans contamination** de l'ADN hôte
- L'ADN doit être extrait en quantité suffisante pour la préparation de la librairie de séquençage
- La **méthode d'extraction** doit être adaptée à la nature des cellules de ces derniers :
    - Par exemple, une approche enzymatique n'extraira pas l'ADN des cellules plus résistantes (archées et Gram+ par exemple), mais une approche mécanique risque de fragmenter l'ADN qui ne pourra pas être amplifié correctement

Extraction d’ADN

- L'ADN exogène issu de l'hôte ou de certains réactifs peut être à l'origine de contaminations, d'où l'importance de mettre en place des contrôles avec des **témoins négatifs** (témoins d’extraction, de PCR...)
- L'étape d'extraction ne permet pas de distinguer entre **cellules vivantes et cellules mortes** dans le milieu, ce qui est à garder à l'esprit dans le cas d'une étude d'inférence fonctionnelle complémentaire
- La co-extraction de molécules inhibitrices peut impacter l'efficacité même de la PCR, et donc porter préjudice à l'élaboration de la librairie de séquençage

La métagénomique ciblée

Plan d'expérience



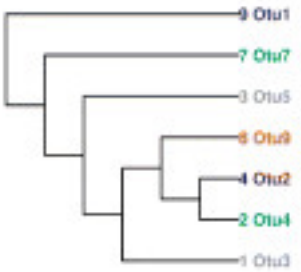
Collecte des échantillons

Extraction d'ADN

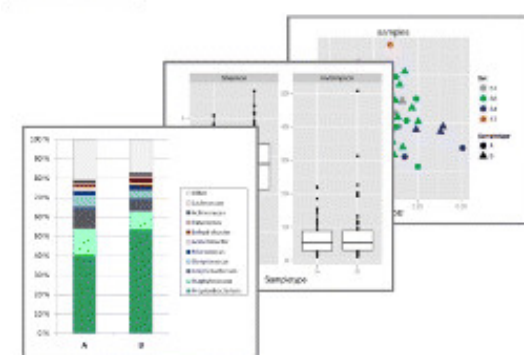
PCR

Séquençage

Analyse primaire
Assignation taxonomique



Analyse secondaire
Statistiques/Interprétation



Amplification de la cible

- **BUT** : cibler le locus génomique d'intérêt et en synthétiser une quantité suffisante pour le séquençage
- Un bon marqueur taxonomique doit être

  - **universel** : présent dans tous les génomes extraits et contenant des séquences conservées pour ancrer la PCR

  - **spécifique** : contenant des séquences variables permettant de discriminer les taxons
- Différents loci se sont imposés comme des marqueurs de référence pour différents règnes, souvent présents dans l'opéron ribosomique

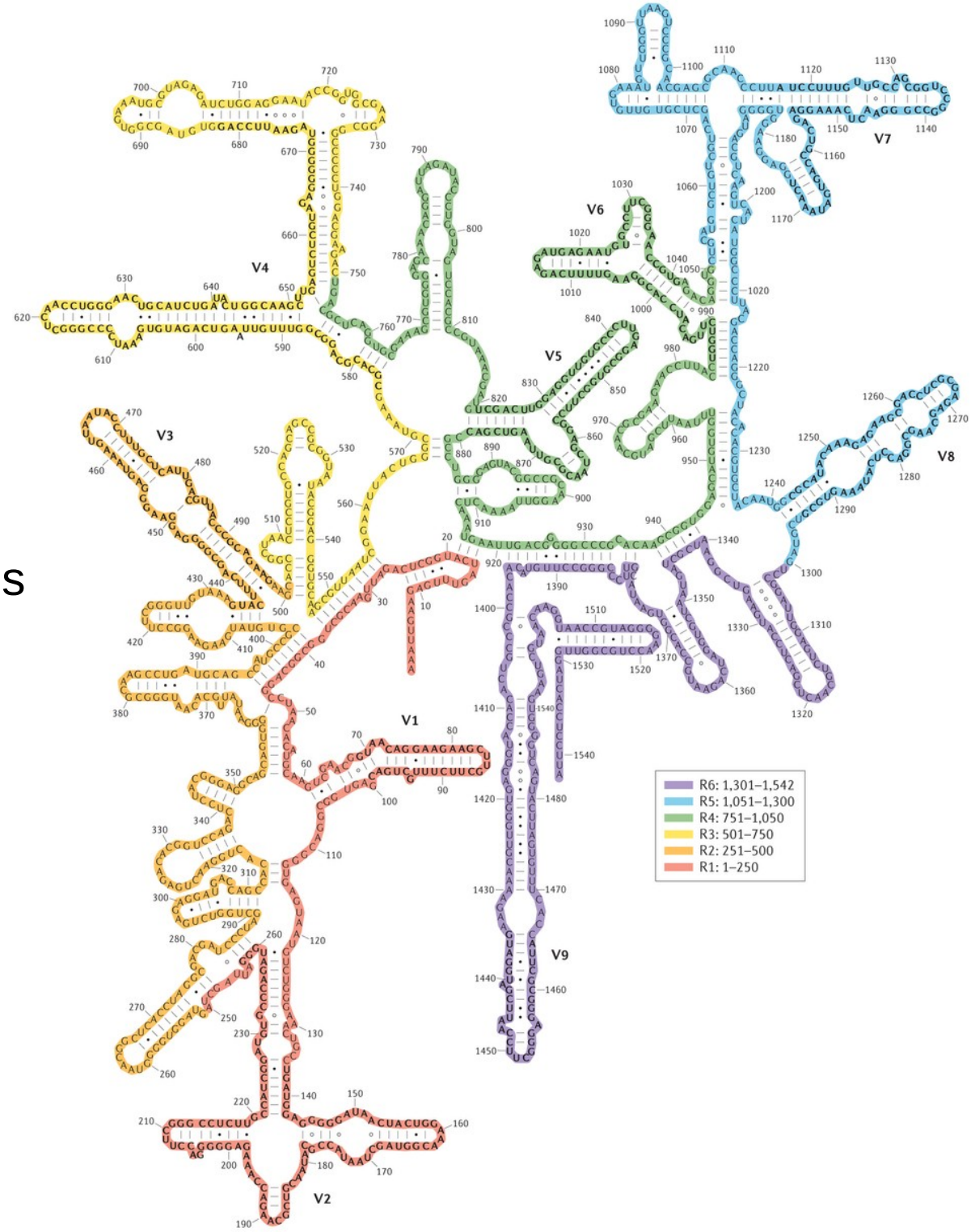
  - ADNr 16S pour les bactéries

  - ITS (Internal transcribed spacer) pour les champignons

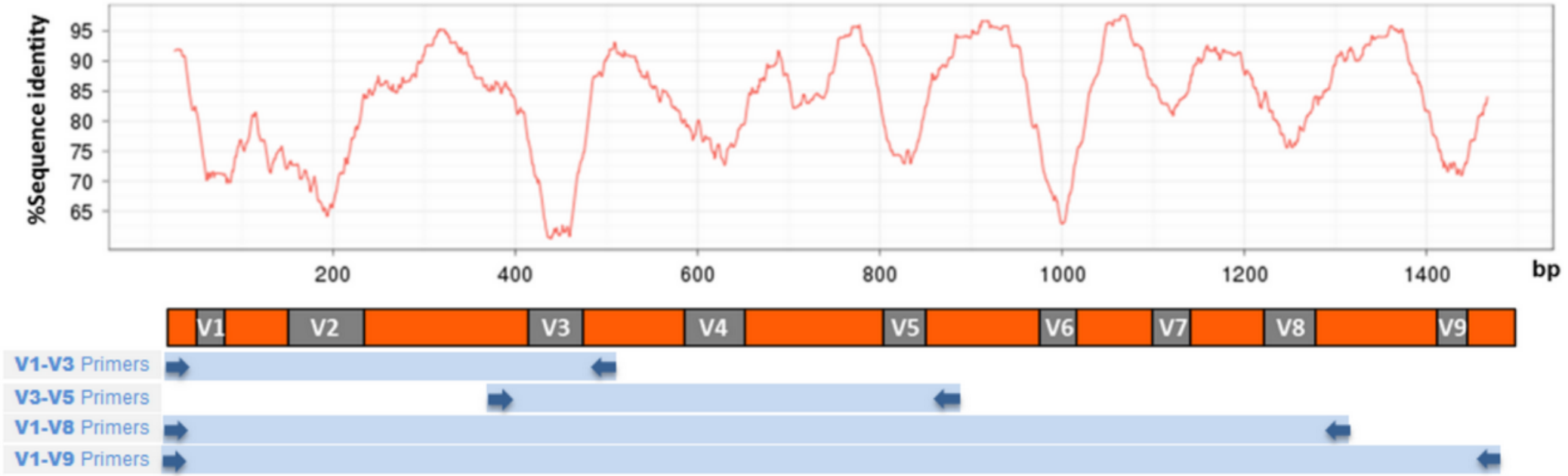
  - ADNr 18S pour les eucaryotes

ADNr 16S

- Les ribosomes sont formés de deux sous-unités permettant la traduction de l'ARN en protéine
- Chez les bactéries, la petite sous unité est formée de l'ARN 16S
- C’est un ARN non codant composé d'environ 1500 nucléotides possédant des régions conservées et des régions variables



ADNr 16S



- Les régions conservées : il est possible de construire des amorces pour une PCR afin de sélectionner la région d'intérêt et de capturer l'ensemble des séquences d'ADNr 16S
- Les régions variables : pas de rôle fonctionnel important et pouvant diverger au cours de l'évolution sous l'effet des mutations neutres. Elles vont permettre de discriminer les taxons bactériens

ADNr 16S : les biais

- Les méthodes actuelles de séquençage haut-débit ne permettent de couvrir qu'un fragment de la région génomique d'intérêt (450 à 550 nucléotides)

  => séquençage de deux à trois régions hypervariables au maximum donc moins de sites discriminants

- Différentes études se contredisent sur le choix de la région, indépendamment du type de microbiote : V2-V3 [Liu et al. 2008] ou V4-V6 [Yang et al. 2016]
- Certaines régions sont conseillées selon le microbiote étudié :
  - V1 – V2 pour l'étude des eaux usées [Guo et al. 2013]
  - V3 - V4 pour le microbiote intestinal humain [Nossa et al. 2010]

ADNr 16S : les biais

- L'opéron ribosomique est présent en un **nombre de copies variables** selon les génomes, pouvant aller de 1 à 15 copies selon l'espèce [Klappenbach et al. 2000]
- Les différentes copies peuvent être différentes ce qui rend plus complexe l'identification de ces organismes

  => certains génomes présentent une variation de séquences intragénomiques pouvant atteindre jusqu'à 11 %

- La variation du nombre de copies de l'ADNr 16S introduit un **biais d'estimation de l'abondance relative** des différents organismes dans l'échantillon : une espèce avec un grand nombre de copies sera favorablement amplifiée et représentée dans les séquences issues du séquençage
- Ainsi, la variation du nombre de copies du locus cible entre génomes provoque une estimation biaisée de la diversité microbienne

ADNr 16S : les amorces

- Essentielles pour capter un maximum d'organismes en présence, sans amplifier de l'ADN contaminant
- Le choix des amorces se fait souvent sur base bibliographique, en utilisant un couple déjà validé par des études précédentes sur le même type de microbiote
- Plusieurs études ont évalué la sensibilité des amorces les plus utilisées [Mao et al. 2012, Klindworth et al. 2013], qui est toutefois directement dépendante de la composition du microbiote
- L'ajout de bases dégénérées permet d'augmenter la sensibilité d'une amorce, aux dépens de sa spécificité : augmentation du risque d'amplification de séquences contaminantes (par exemple, amplification de séquences chloroplastiques ou mitochondriales)

ADNr 16S : les amorces

- La réaction d'amplification en elle-même est source de nombreux biais techniques: il est délicat d'assumer que les séquences amplifiées sont **représentatives des abondances initiales** des organismes
- Un nombre de cycles PCR trop élevé ou une concentration de l'ADN matrice trop importante augmente les risques de **déséquilibrer le ratio** des organismes en présence et la formation de séquences chimériques (lorsqu'un amplicon en cours de synthèse glisse vers une autre séquence matrice)
- Les **erreurs d'amplification** (substitutions causées par de mauvais appariements et délétions dues à des glissements de la polymérase) peuvent avoir un impact sur les résultats, si par exemple elles touchent un nucléotide discriminant entre deux taxons

  => utiliser une polymérase à activité correctrice peut limiter la quantité d'erreurs dans les amplicons, mais semble favoriser l'apparition de chimères [Ahn et al. 2012]

Autres marqueurs

- Eucaryotes (protistes et fungi)
  - 18S
  - ITS
- Bactéries
  - CPN60
  - ITS
  - RecA gene
- Virus
  - Gp23 pour les bactériophages T4Plike
  - RdRp pour les picornaviruses

Des marqueurs évoluant plus rapidement sont utilisés pour différencier des organismes au niveau de l'espèce

Autres marqueurs

- D'autres gènes à copie unique ont été proposés comme marqueurs pour la métagénomique ciblée pour estimer une meilleure abondance relative des différentes espèces bactériennes
- Certains gènes de ménage tels que ceux de la famille recA ou le gène rpoB sont universellement présents chez les bactéries et ont un degré de variation de copies moindre entre espèces, comparé à l'ADNr 16S
- Ces gènes semblent en outre discriminer la taxonomie à un niveau plus élevé, en présentant plus de variabilité [Thompson et al. 2004, Vos et al. 2012]

Autres marqueurs

- Le design des amorces universelles sur des gènes codant pour des protéines est plus difficile: des régions protéiques conservées n'impliquent pas que les régions génétiques correspondantes le soient également
- Les amorces doivent contenir plus de dégénérescences pour s'aligner sur un maximum d'organismes, sans garantir leur universalité
- Ces gènes sont également plus sensibles à des événements de transferts horizontaux
- La représentation de ces gènes dans les banques de séquences est minime comparée aux séquences d'ADNr 16S

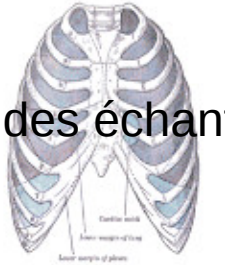
Autres marqueurs

- Plus une cible est étudiée, plus elle est référencée, plus on favorisera son étude ; ce cercle vicieux rend difficile le développement de nouvelles méthodes basées sur d'autres cibles
- Ce problème se présente par exemple pour les études de métagénomique fongique : les séquences ribosomiques interstitielles (ITS) sont un marqueur évolutif plus discriminant des Fungi que l'ADNr 18S [Schoch et al. 2012]. Or, il existe peu de banques spécifiques à ce marqueur
- Les séquences d'ITS qui s'y trouvent sont souvent incomplètes et/ou mal annotées. Le manque de connaissances sur ce marqueur freine ainsi le développement des applications de métagénétique fongique, qui sont pourtant précieuses dans l'étude de certaines pathologies humaines par exemple

La métagénomique ciblée

Plan d'étude

Collecte des échantillons

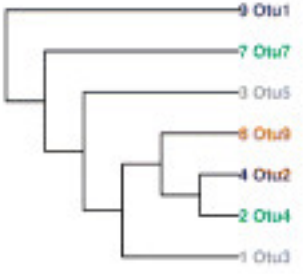


Extraction d'ADN

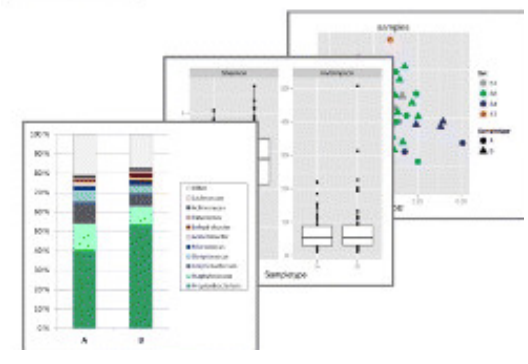
PCR

Séquençage

Analyse primaire
Assignation taxonomique



Analyse secondaire
Statistiques/Interprétation



Différents types de séquenceurs haut-débit ?

Séquenceurs très haut-débit

Séquenceurs de paillasses

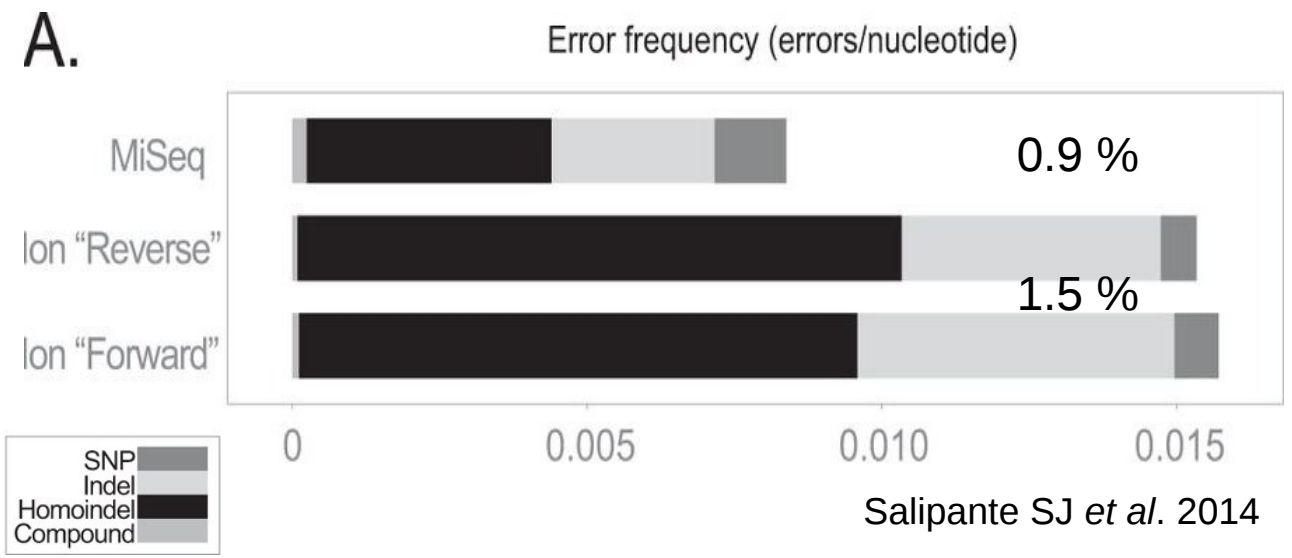
=> très bien adaptés à la métagénomique ciblée

Les séquenceurs de paillasse

- Séquenceur Illumina MiSeq
- Séquenceur Ion Torrent Personal Genome Machine (PGM)

Détection : fluorescense

Différence majeure : erreurs de séquençage



Salipante SJ *et al*. 2014

Détection : variation de PH

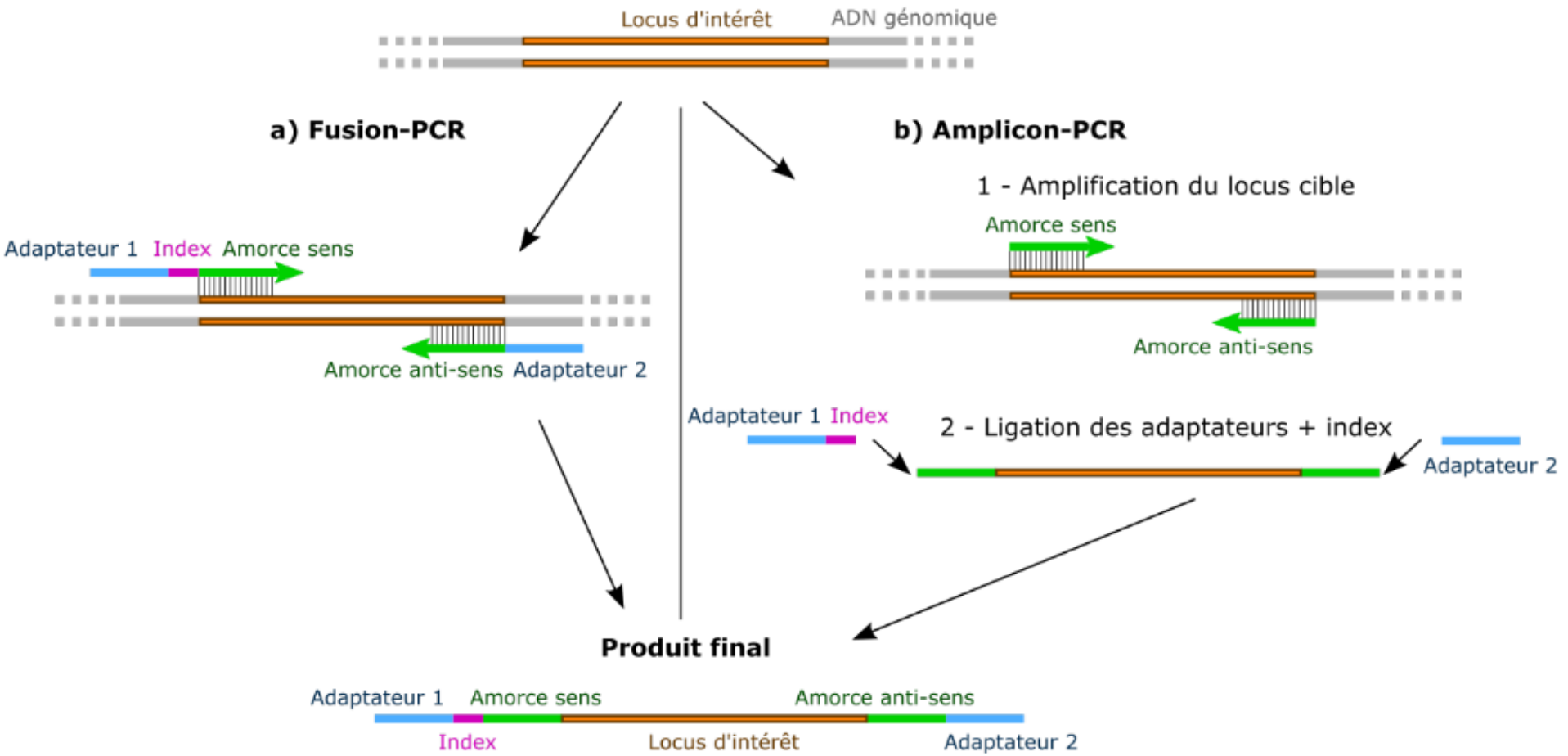
Les séquenceurs de paillasse

	454 GS Junior	Ion Torrent PGM	Illumina MiSeq
Date de sortie	Fin 2009	2010	2011
Méthode d'amplification	PCR par émulsion (hors séquenceur)	PCR par émulsion (hors séquenceur)	PCR en ponts (dans le séquenceur)
Mode de détection de la polymérisation	Pyroséquençage (détection de lumière)	Détection de variation de pH	Terminateur fluorescent réversible
Taille maximale des lectures	400 nt	450 nt	2 x 300 nt
Débit	35 Mb	Chip 314™ ~ 40 Mb Chip 316™ ~ 200 Mb Chip 318™ ~ 1 Gb	1,5 Gb
Temps d'un run (hors préparation des librairies)	8 h	3 h	27 h

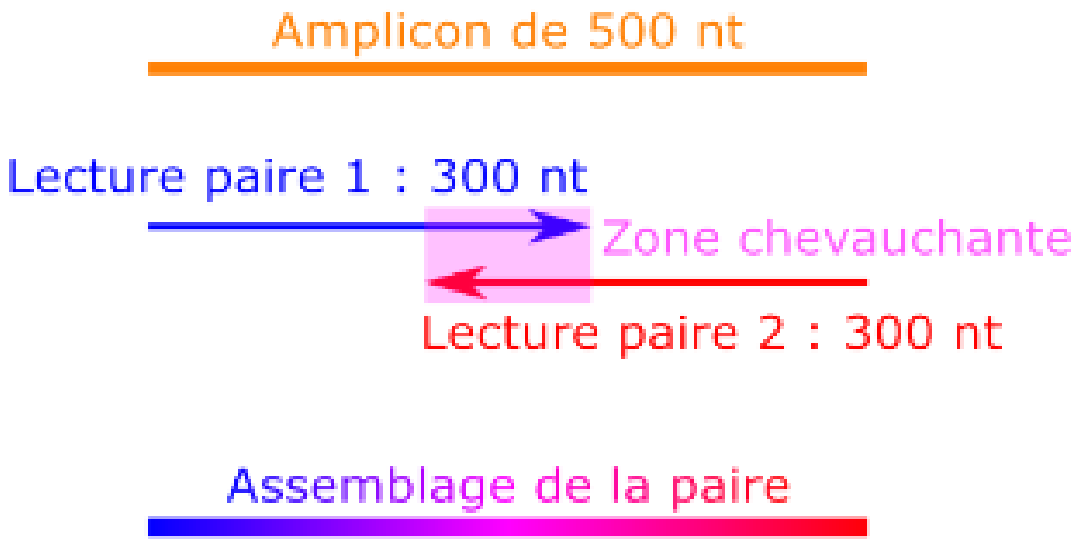
Préparation de la librairie

- Les fragments d'ADN à séquencer sont générés par PCR sur le mélange d'ADN génomique initial, en utilisant des d'amorces ciblant le locus d'intérêt
- Ces amplicons doivent également contenir à leurs extrémités des séquences artificielles nécessaires au séquençage :
  - des séquences d'adaptateur, permettant d'ancrer par complémentarité les fragments au support et d'initier le séquençage
  - des séquences d'index, codes-barres artificiels permettant de marquer les fragments selon leur échantillon d'origine, et ainsi de mélanger plusieurs échantillons en un seul run de séquençage (dit multiplex)

Préparation de la librairie



Illumina MiSeq



Séquençage Illumina paired-end :
Le séquençage de la cible est effectué dans les deux sens, générant 2 lectures par amplicon qui se chevauchent, permettant de les assembler en une lecture plus longue sur la base de cette zone chevauchante

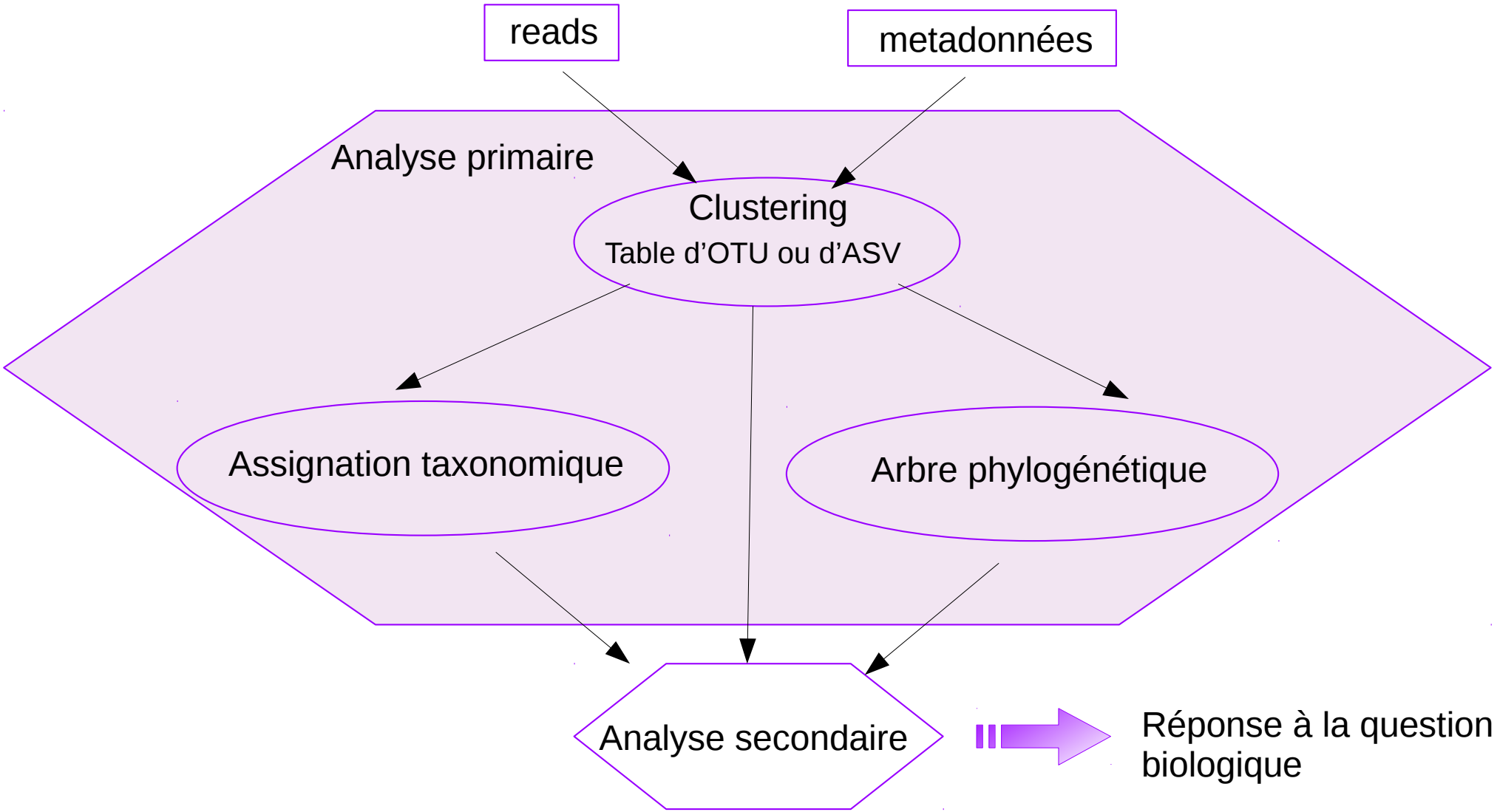
A la sortie du séquenceur

```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
@SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
BBCBBBBBB@@BAB?BBBBCBC>BBBAA8>BBBAA@
```

Fichiers FASTQ contenant les reads démultiplexés (ou non):

- 1 fichier par échantillon pour Ion Torrent
- 2 fichiers (reads1 et reads2) par échantillon pour Illumina

Analyse des données

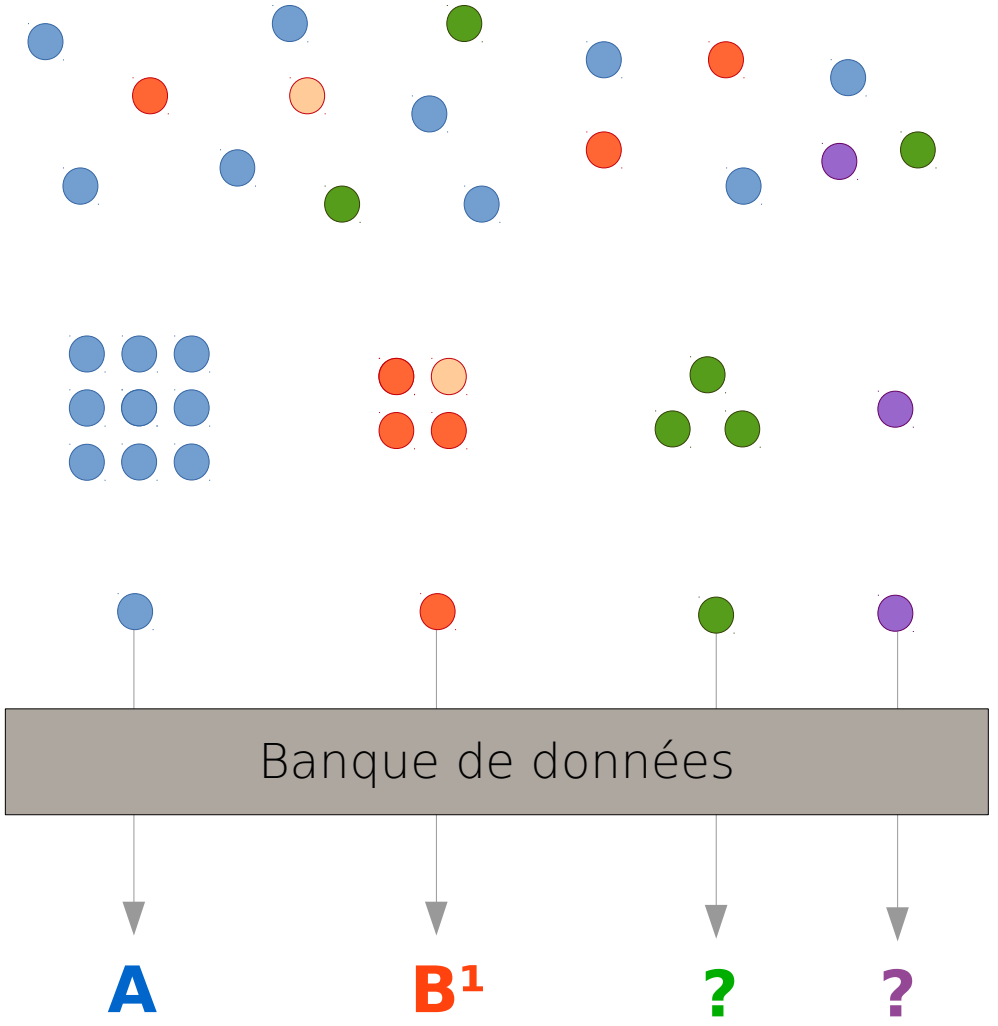


Analyse primaire

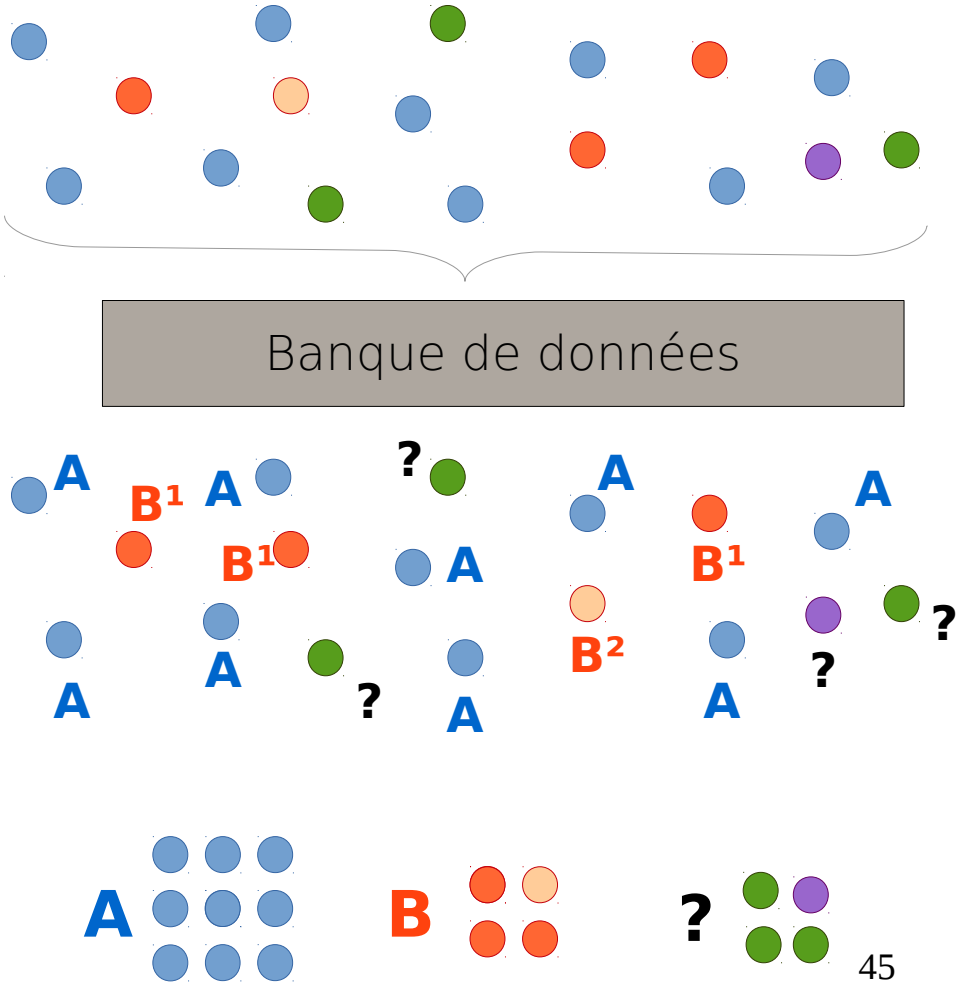
- L’analyse primaire a pour but de regrouper ensemble les lectures qui proviennent du même organisme (= qui se ressemblent)
- 2 grands types d’approches :
  - Approches taxonomiques : développées pour la métagénomique WGS
  - Approches par clustering : développées et **utilisées classiquement** pour la métagénomique ciblée (OTU) => évolution vers les variants (ASV)

Les 2 types d'approches

Approches par clustering



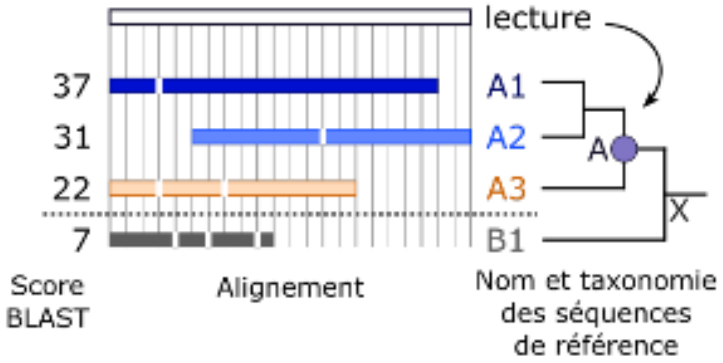
Approches taxonomiques



Approches taxonomiques

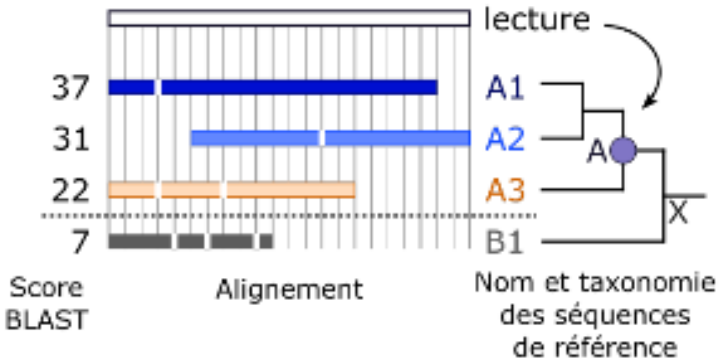
- Premières approches utilisées pour la métagénomique ciblée mais vite remplacées car trop chronophage
- Approches **développées pour la métagénomique WGS**
- Annotation individuelle de chaque lecture, avant de les regrouper par taxon sur la base de ces annotations
- Beaucoup de ces approches ne sont pas applicables à des données amplicon
- Des approches basées sur des comparaisons de motifs entre les lectures et des banques de séquences de référence peuvent être appliquées à l'analyse de données de métagénomique ciblée, même si elles n'ont jamais été utilisées dans ce contexte

Alignement



- Alignement de chaque lecture contre une banque de séquences de référence (historiquement par BLAST) et annotation de chaque lecture par le taxon dont il est le plus similaire
- Cette méthode ne prend pas en compte la possibilité d'une divergence entre une lecture (issue d'un génome inconnu par exemple) et les génomes présents dans la banque de référence, pouvant généraliser de **fausses assignations taxonomiques trop précises**
- La lecture est assignée à l'espèce A1 avec laquelle elle présente le meilleur alignement en termes de score mais la lecture pourrait être assigné à l'espèce A2: le score plus bas pouvant simplement être causé par le fait que A2 est tronquée dans la banque, et ne couvre pas le début du read

Alignement

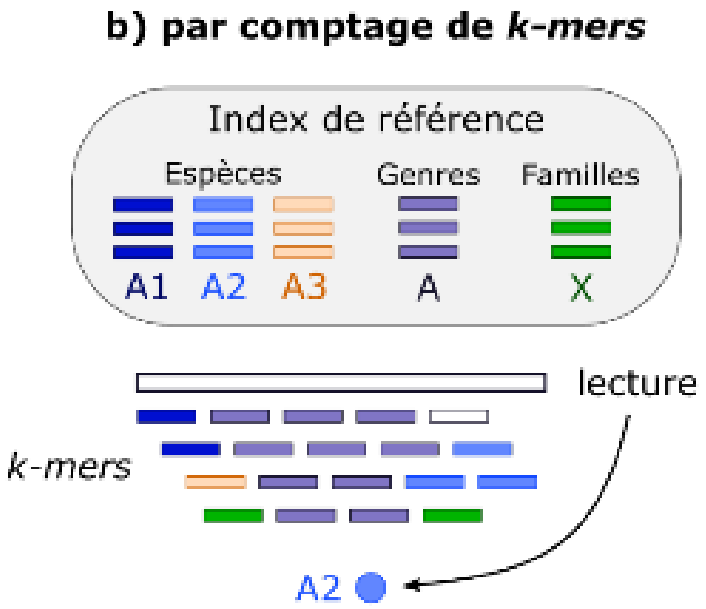


- Introduction de l'algorithme LCA (*Lowest Common Ancestor*), introduit dans le logiciel MEGAN, et intégré par la suite dans de nombreux pipelines
- Cet algorithme interprète, pour chaque lecture, une sélection de plusieurs hits BLAST validés comme étant significatifs sur la base de leur score
- Les hits significatifs sont les séquences des espèces A1, A2 et A3. La lecture est assignée à leur ancêtre commun, le genre A.
- La principale limite de ces méthodes est l'étape d'alignement, demandant **beaucoup de temps de calcul** pour de gros jeux de données

k-mers

- Nouvelles méthodes : comparaison des lectures à une banque de référence **sans alignement**
- Basées sur l'étude de la composition des lectures en k-mers pour retrouver des signatures spécifiques des génomes auxquels elles appartiennent
- Par exemple, **kraken** [Wood et al.2014] construit d'abord un index de tous les k-mers trouvés dans la banque de référence, et assigne à chacun d'eux l'ancêtre commun à tous les organismes contenant ce k-mer. Chaque lecture est alors elle-même découpée en k-mers et comparée à cet index.

kmers



- La lecture partage le plus de k-mers avec le genre A, elle est ainsi au moins assignée à ce genre.
- En évaluant les k-mers partagés entre la lecture et les espèces du genre A, la lecture a une majorité de k-mers similaires à l'espèce A2 : l'annotation peut être précisée en assignant la lecture à l'espèce A2.

L’approche taxonomique

- Ces pipelines retournent en résultat une assignation taxonomique par lectures ; l'utilisateur doit ensuite regrouper les lectures en taxons pour générer une table.
- Ces méthodes pouvant traiter plusieurs millions de lectures par minute sont **bien plus rapides que les méthodes par alignement** ; elles sont ainsi de plus en plus utilisées dans l'assignation taxonomique de données métagénomiques WGS
- Limitation : basée sur une banque de séquences de référence, représentant un *a priori* de connaissance

  => ce type d’approche ne peut **pas être appliqué sur des microbiotes peu décrits** dans la littérature et peu représentés dans les banques de référence

L’approche taxonomique

- Ces pipeline peuvent tout de même être utilisés sur des données de métagénomique ciblée, puisque les banques de référence utilisées par ces pipelines contiennent aussi des séquences d'ADNr 16S
- Une telle utilisation n'a toutefois jamais été décrite dans la littérature
- Nous avons montré que ces pipelines peuvent être utilisés en métagénomique ciblée [Siegwald et al. 2017]
  - Très rapides (quelques secondes/minutes)
  - Bon résultats dans le cas de microbiotes bien référencés
  - Permettent d’obtenir une première image du microbiote d’intérêt : contaminations, espèces en présence …
  - Peut être complémentaire à une approche classique par clustering

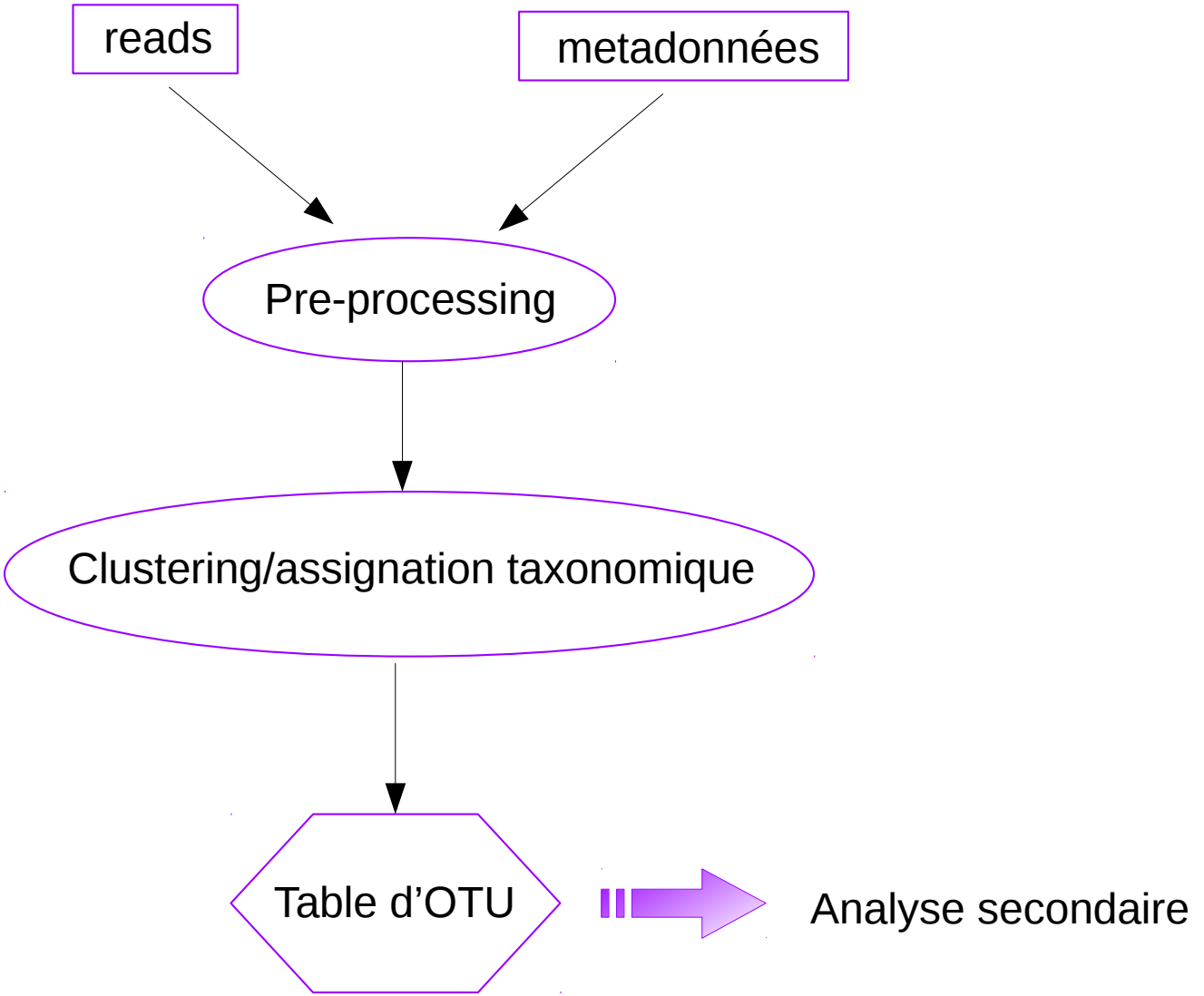
Approches par clustering

- Ces approches reposent sur la formation de clusters = OTU (Operational Taxonomic Unit) Unité Taxonomique Opérationnelle

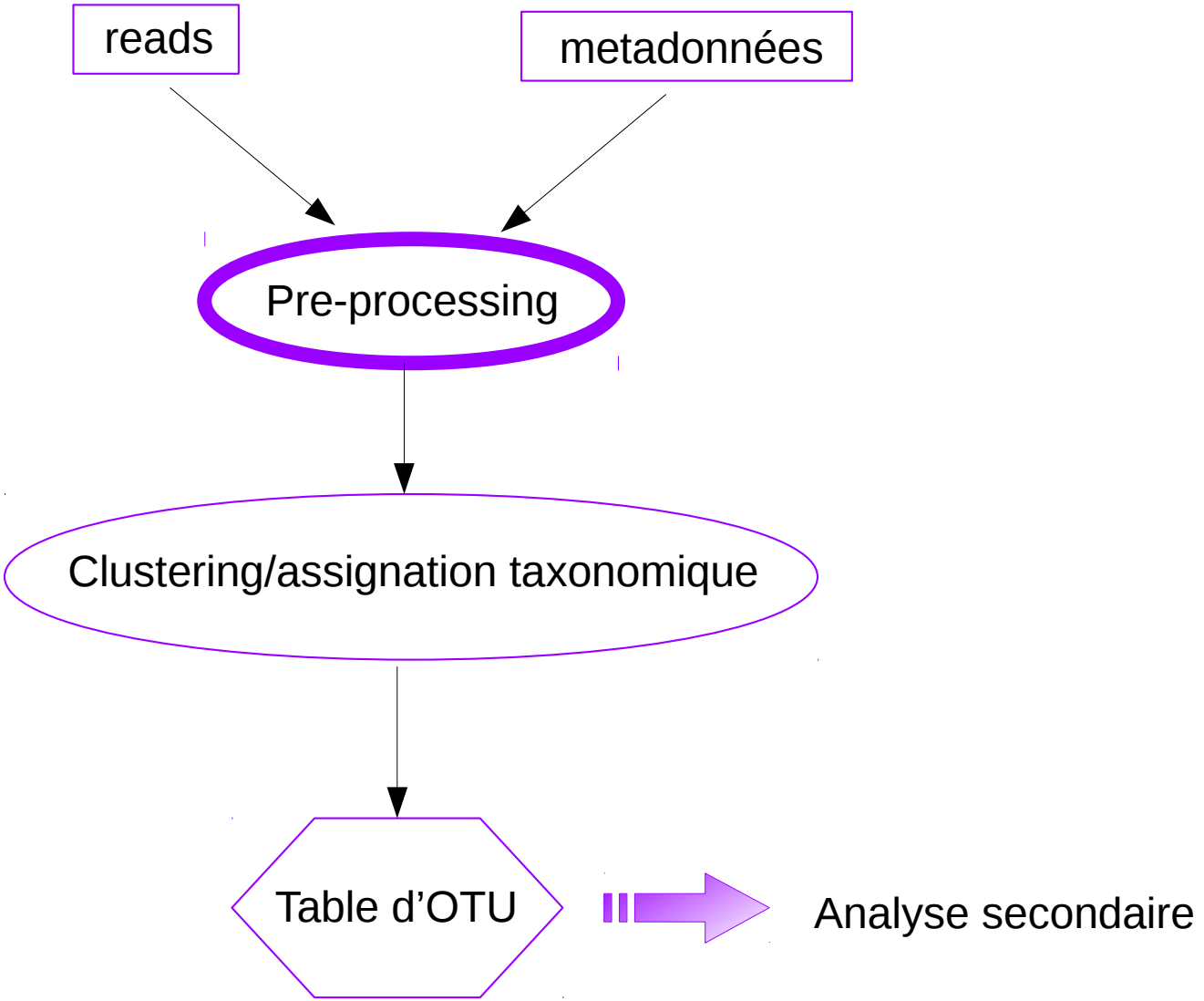
  => **évolution vers les ASV**

- Regroupement de séquences sur la base d'un seuil de similarité
- Plusieurs étapes avant le clustering

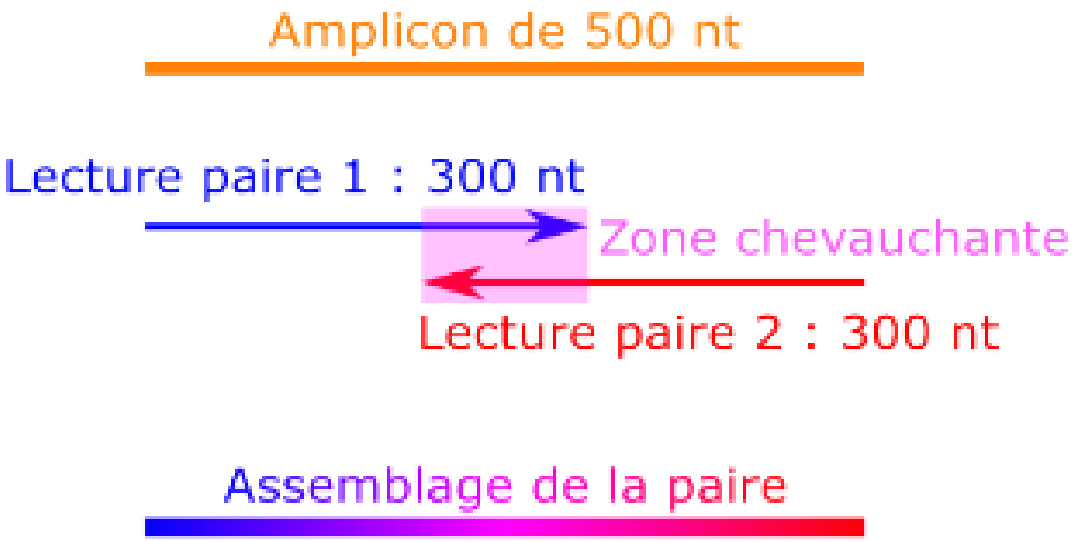
Analyse primaire par OTU



Analyse primaire par OTU

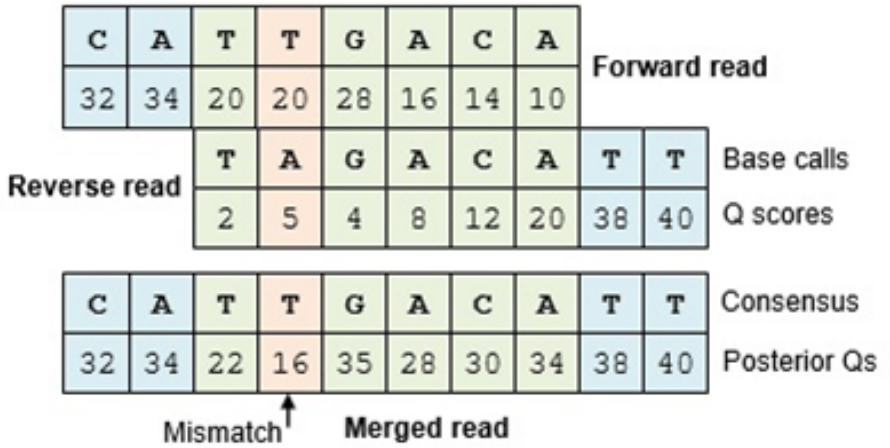


Assemblage des reads paired-end



Assemblage des reads paired-end

- Principe :



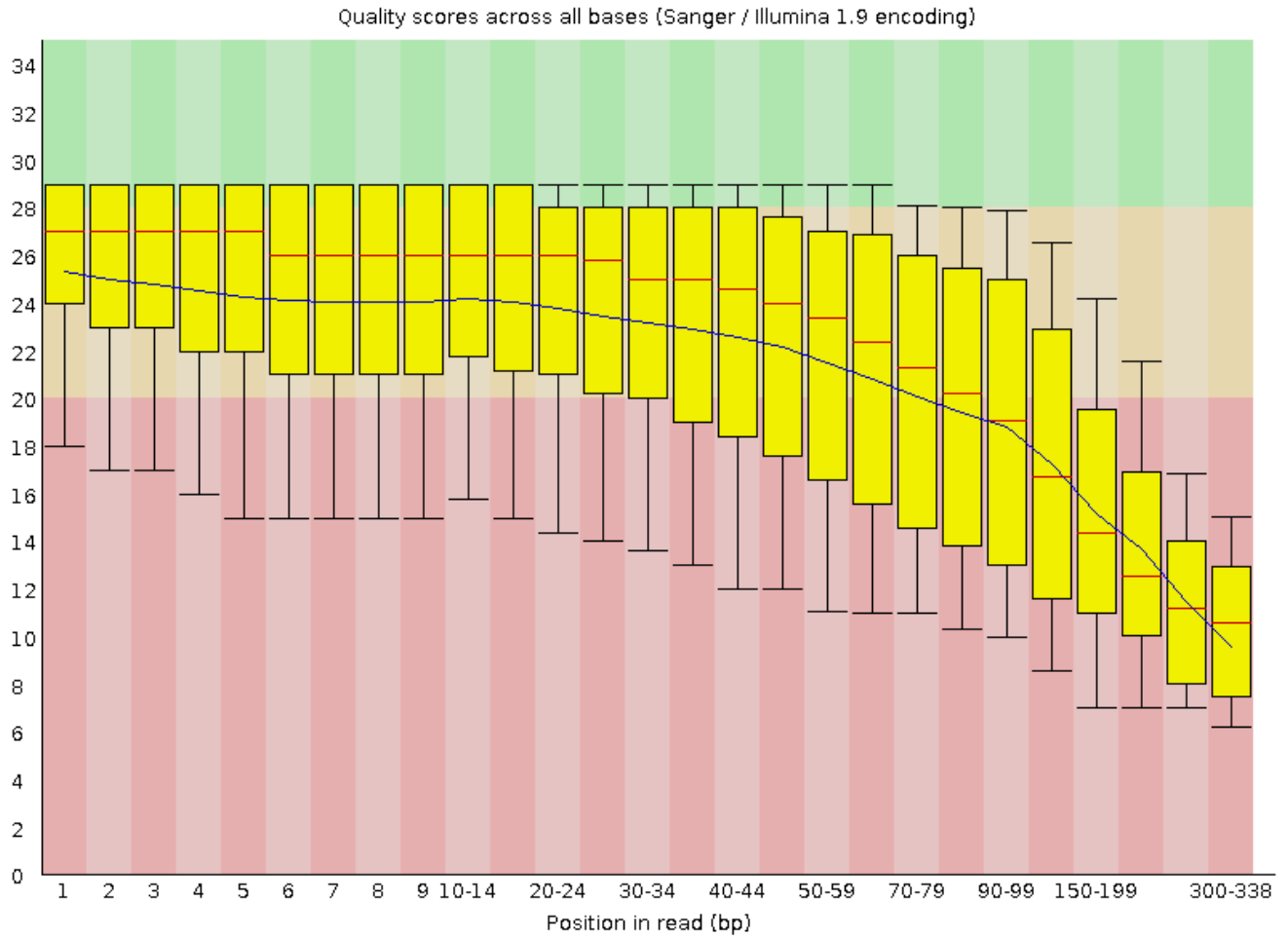
- Résultat : 1 seul lecture = 1 seul fichier FASTQ
- Exemple de logiciel : PEAR, make.contigs (Mothur),….

Trimming/Filtrage

- Trimming/Filtrage des lectures selon plusieurs critères

  => bases ambiguës, primers, longueur, qualité ...

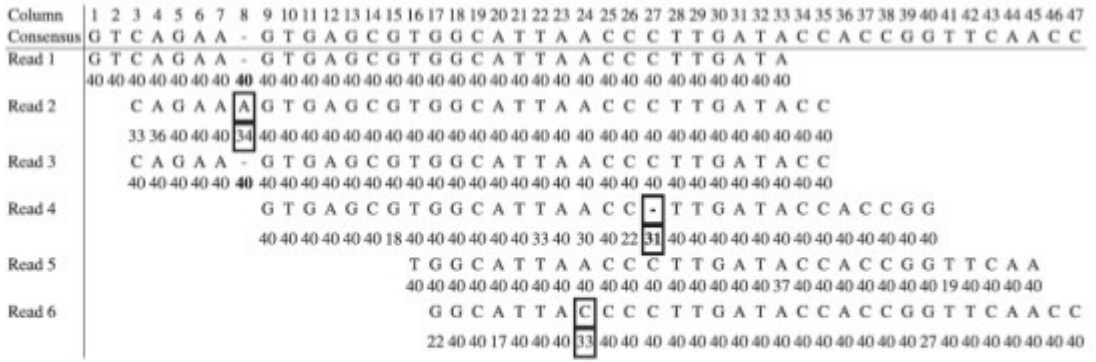
FastQC



Débruitage

- Débruitage (denoising) : élimination des erreurs de séquençage
- Selon le modèle d'erreurs de la technologie de séquençage il est possible de corriger *in silico* des erreurs de séquençage (modèle mathématiques , probabilités, …)

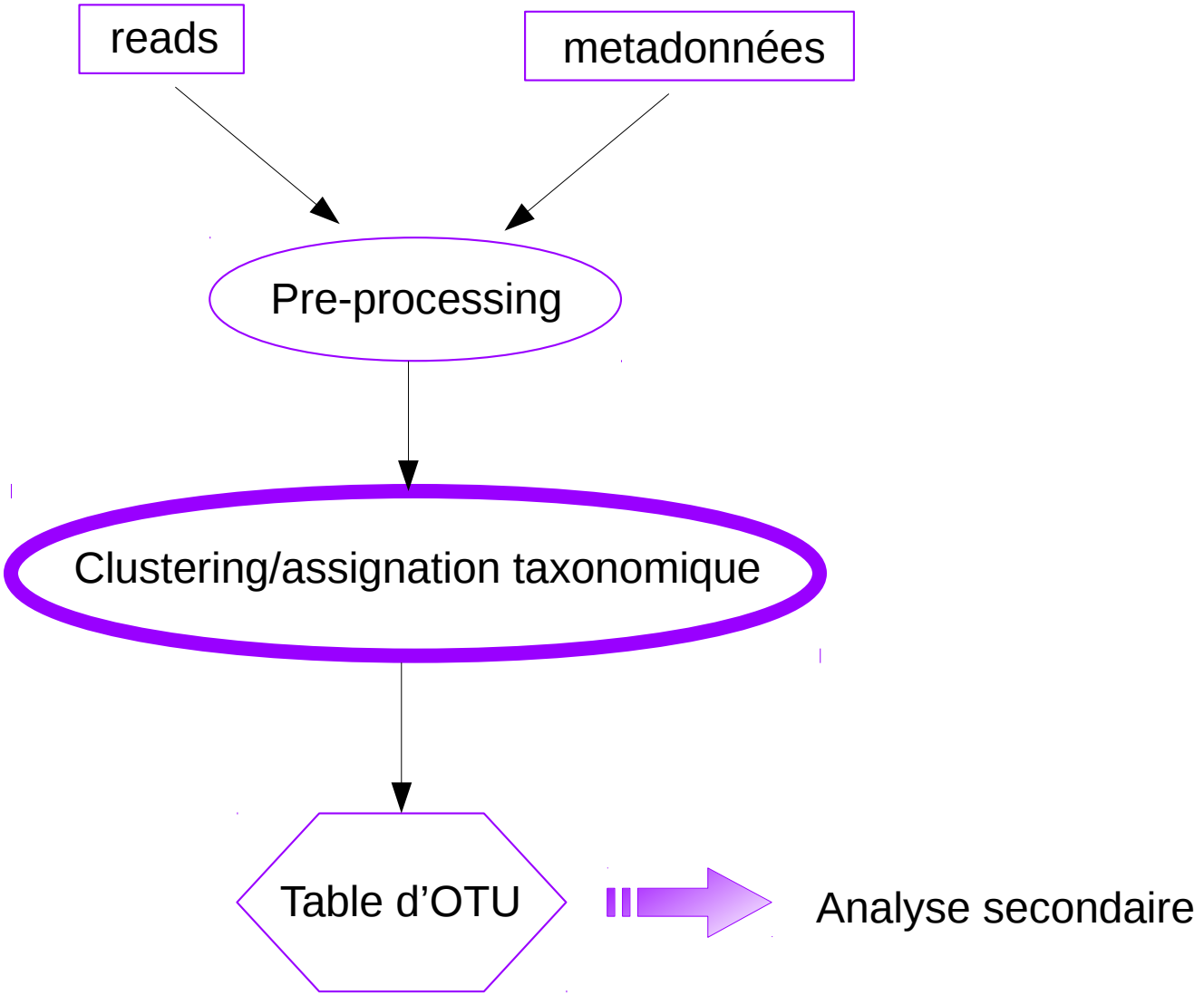
Error Correction



Salamela and Schroder 2011 Bioinformatics

- Étape importante pour le 454

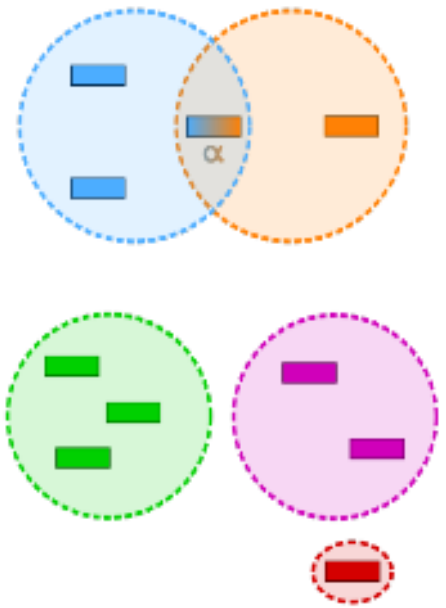
Analyse primaire



Le clustering

- 3 grandes approches
  - Clustering de novo
  - Clustering closed-reference
  - Clustering open-reference

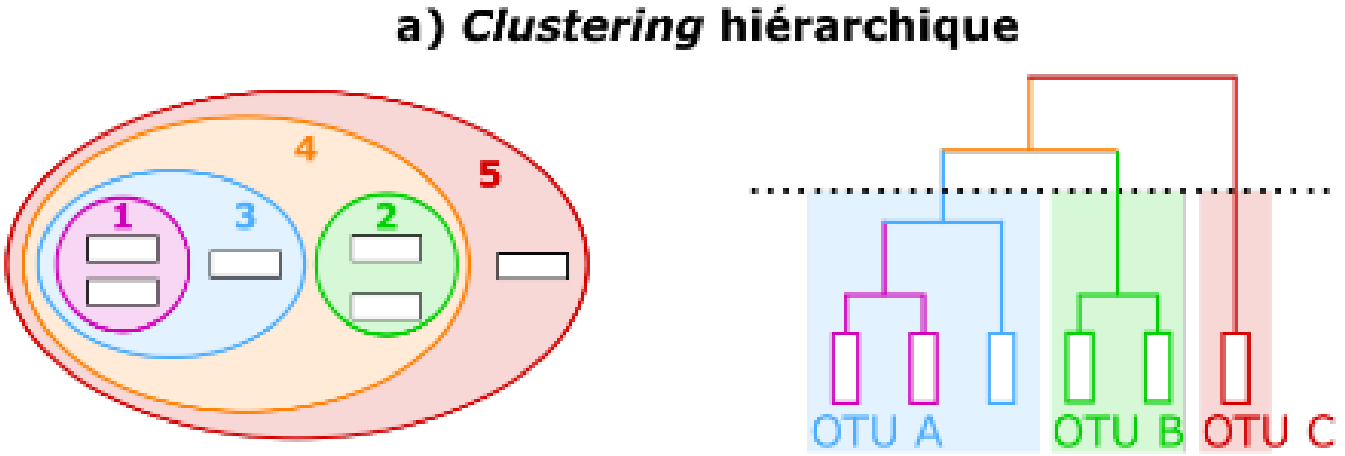
Clustering de novo



- Basé uniquement sur les propriétés intrinsèques des séquences (similarité entre les séquences)
- Comparaison 2 à 2 de toutes les séquences
- 2 types de méthodes
  - clustering hiérarchique
  - clustering par centroïdes

séquence brute | séquence d'une banque de référence | seuil de similarité (ex : 97%) | cluster

Clustering hiérarchique



- Comparaison 2 à 2 des lectures (par exemple, Needleman-Wunsch) => matrices de distances

- regrouper les lectures de manière itérative sur la base de la matrice de distances, au départ, chaque lecture est un cluster isolé.

- Le clustering hiérarchique va regrouper les clusters les plus proches en un seul cluster, de manière itérative jusqu'à ce qu'il n'y ait plus qu'un seul cluster contenant toutes les lectures

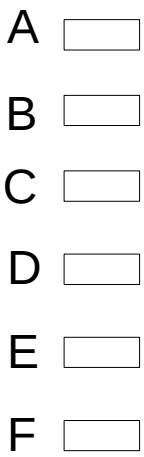
Clustering hiérarchique

Reads

A
B
C
D
E
F

	A	B	C	D	E	F
A	0	0,6	0,4	0,2	0,3	0,4
B		0	0,9	0,6	0,8	0,7
C			0	0,5	0,1	0,3
D				0	0,4	0,5
E					0	0,3
F						0

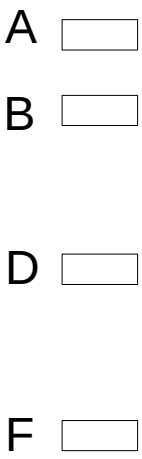
Clustering hiérarchique



	A	B	C	D	E	F
A	0	0,6	0,4	0,2	0,3	0,4
B		0	0,9	0,6	0,8	0,7
C			0	0,5	**0,1**	0,3
D				0	0,4	0,5
E					0	0,3
F						0

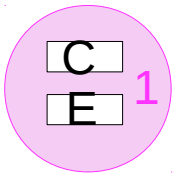
C et E => cluster 1

Clustering hiérarchique



	A	B	C + E	D	F
A	0	0,6	**0,35**	0,2	0,4
B		0	**0,85**	0,6	0,7
C + E			0	**0,45**	**0,3**
D				0	0,5
F					0

(0,5+0,4)/2



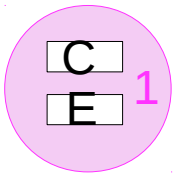
Clustering hiérarchique

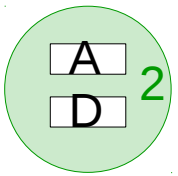
B

F

	A	B	C + E	D	F
A	0	0,6	0,35	**0,2**	0,4
B		0	0,85	0,6	0,7
C + E			0	0,45	0,3
D				0	0,5
F					0

A et D => cluster 2



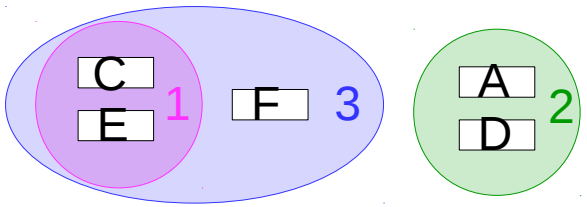


Clustering hiérarchique

B

	A + D	B	C + E	F
A + D	0	0,6	0,4	0,45
B		0	0,85	0,7
C + E			0	**0,3**
F				0

C+E et F => cluster 3

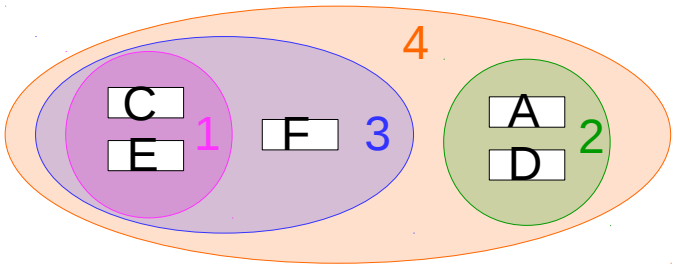


Clustering hiérarchique

B

	A + D	B	(C+E)+ F
A + D	0	0,6	**0,425**
B		0	0,775
(C+E)+ F			0

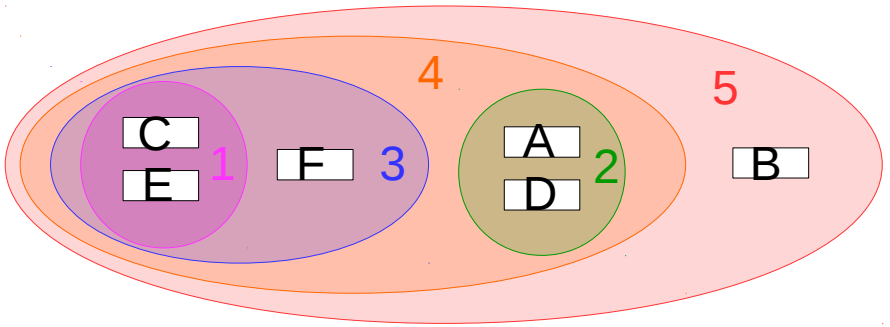
((C+E)+F) + (A+D) => cluster 4



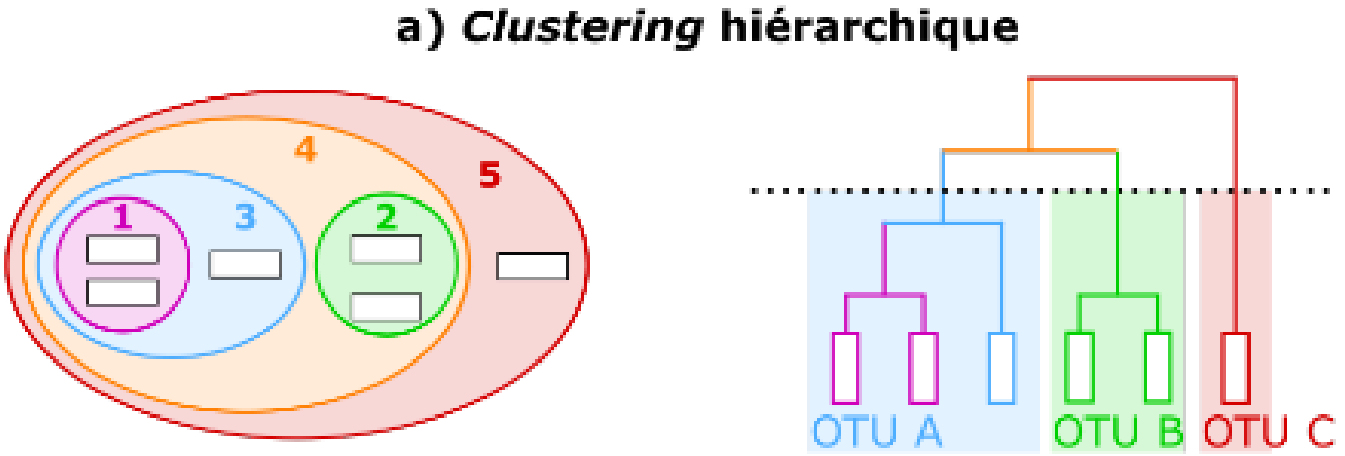
Clustering hiérarchique

	(A+D)+ ((C+E) +F)	B
(A+D)+ ((C+E) +F)	0	**0,6875**
B		0

(((C+E)+F) + (A+D)) + B => cluster 5



Clustering hiérarchique



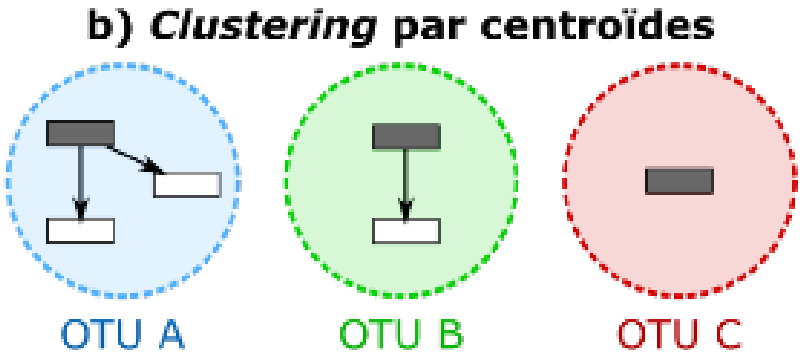
- Le clustering hiérarchique peut être représenté sous la forme d'un dendrogramme
- Les OTU sont ensuite définis en sélectionnant un seuil de pourcentage de similarité (par exemple 97 %), correspondant à un niveau du dendrogramme où chaque branche forme un OTU contenant des séquences similaires à au moins ce seuil

Clustering hiérarchique

- Complexité au moins quadratique => difficilement applicables à de grands jeux de données, génèrent d'immenses matrices de distance
- Utilisation de méthodes de réduction de données :
  - logiciels de débruitage des lectures pour éliminer les erreurs
  - étape de dé-réplication des lectures et/ou de pré-clustering optimisé, pour réduire au maximum la quantité de lectures à comparer entre elles
- Utilisation des distances entre mots (k-mers) au lieu des distances entre lectures

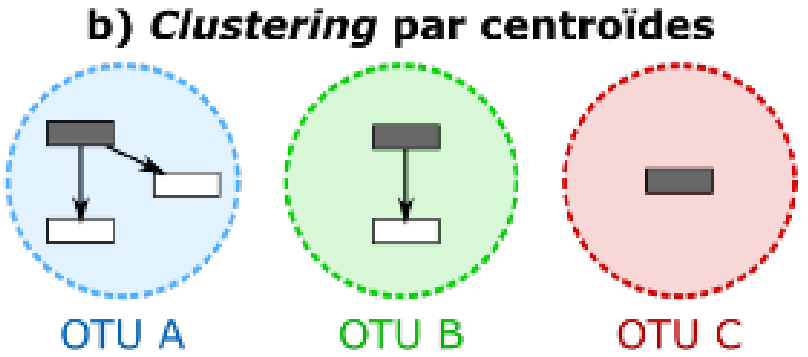
  => ESPRIT [Sun et al. 2009] permet ainsi d'éliminer les k-mers redondants entre séquences, mais a toujours une complexité quadratique

Clustering par centroïde



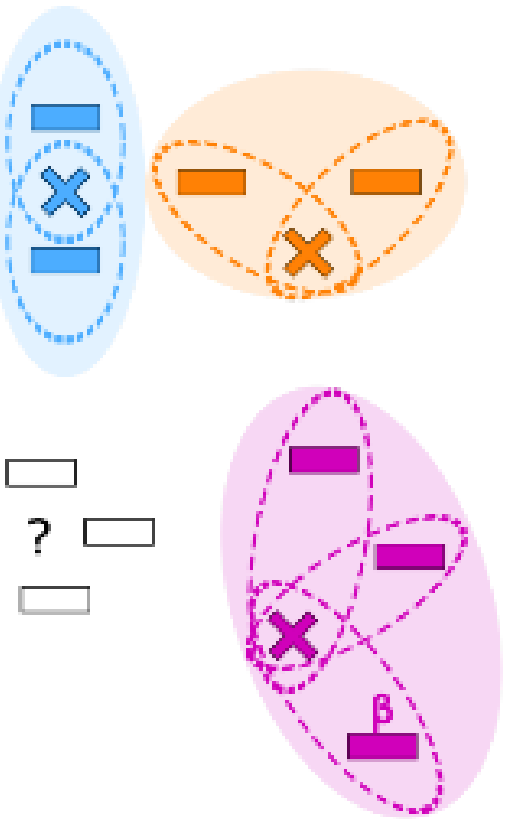
- **Approche heuristique** : les lectures sont d'abord triés par longueur et/ou par abondance décroissante, partant de l'hypothèse que les séquences les plus longues et les plus abondantes contiennent un signal biologique fort
- La première lecture de la liste est considérée comme étant le centroïde du premier cluster
- La lecture suivante est comparée à ce centroïde : si leur identité est supérieure au seuil choisi pour définir un cluster (par exemple 97 %), alors la lecture est ajoutée au cluster existant. Sinon, elle devient le centroïde d'un nouveau cluster

Clustering par centroïde



- Toutes les lectures sont ainsi comparées aux centroïdes définis au fur et à mesure => évite la comparaison de toutes les séquences entre elles
- **Fortement dépendant de l'ordre des séquences**: une lecture peut être assignée à un centroïde alors qu'elle présente plus de similarité avec un centroïde qui sera créé ultérieurement car plus loin dans la liste des séquences
- Exemples : UCLUST [Edgar 2010] et son équivalent libre VSEARCH [Rognes et al. 2016], CD-HIT [Fu et al. 2012] ou encore Sumaclust

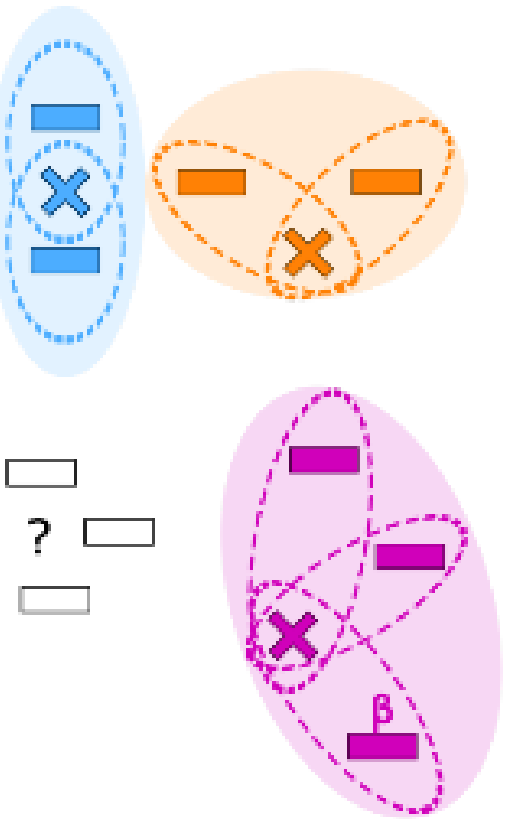
Approche closed-reference



- Similaire à l'approche *de novo* par centroïdes
- Supervisée, en utilisant non plus des lectures du jeu de données comme centroïdes, mais les séquences d'une **banque de référence**
- Chaque lecture est ainsi comparée à toutes les séquences de la banque considérées comme centroïde, et est assignée au cluster dont le centroïde y est le plus similaire
- Exemples : UCLUST_ref [Edgar 2010], SortMeRNA [Kopylova et al. 2012]).

séquence brute | séquence d'une banque de référence | seuil de similarité (ex : 97%) | cluster

Approche closed-reference

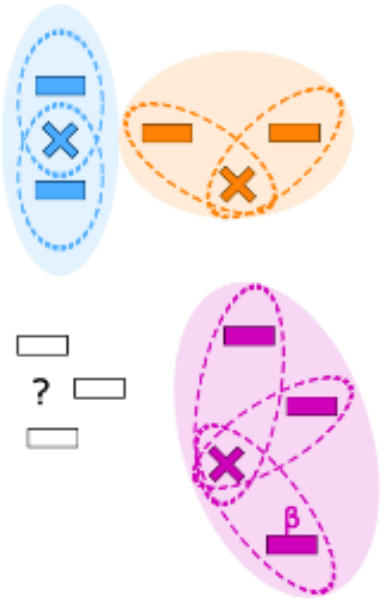


- Les centroïdes sont **plus robustes** dans cette approche, car ils correspondent à une référence biologique sans être dépendants du jeu de données
- Permet une **comparaison directe** des OTU entre différents jeux de données sur la base des centroïdes communs, ce qui est impossible dans le cas d'un clustering *de novo*
- Permet une **annotation taxonomique directe** des reads, chaque OTU pouvant être annoté avec l'annotation de sa séquence centroïde

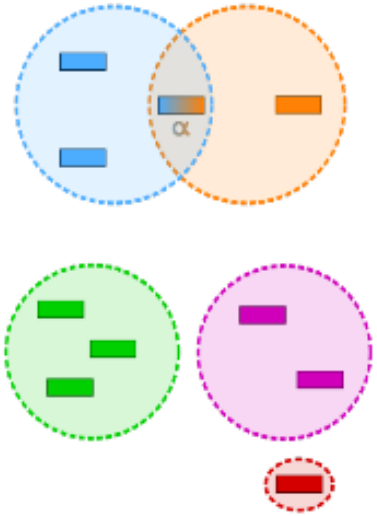
séquence brute | séquence d'une banque de référence | seuil de similarité (ex : 97%) | cluster

Limites

Closed-reference

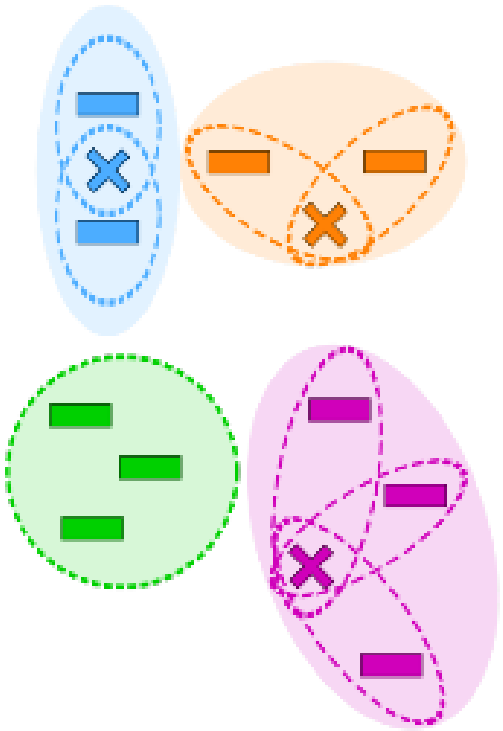


de novo



- La définition d'un cluster n'est pas la même dans l'approche closed-reference que dans l'approche *de novo* : dans l'approche closed-reference, les séquences d'un même cluster peuvent être à une plus grande distance l'une de l'autre qu'elles le sont de la séquence de référence
- la séquence β est à 97 % de similarité de la séquence de référence mais similarité inférieure des autres séquences du même cluster
- cette séquence est isolée (en rouge) par approche *de novo*
- Inconvénient majeur de l'approche closed-reference : **dépendante d'une banque** de référence (*a priori* de connaissance)
- environnement peu décrit dans les banques, des séquences non référencées (en blanc) ne pourront pas être assignées à un OTU

Approche open-reference



- Mélange des deux approches précédentes : d'abord une analyse closed-reference, puis une analyse *de novo* sur les séquences qui ne s'alignent pas avec la banque de référence
- ATTENTION : les OTU générés ne reposent pas sur les mêmes définitions, ce qui ne permet pas de les comparer entre eux

Limites du clustering

- Nécessité de fixer un seuil de similarité pour définir un OTU
- Le seuil de 97 % s'est imposé comme un standard universel pour représenter différentes espèces bactériennes, sans correspondre à une réalité taxonomique
- Ce seuil a été établi sur la base d'études d'hybridation ADN-ADN, sur lesquelles reposaient la définition de nouvelles espèces bactériennes, et la similarité des séquences d'ADNr correspondantes [Stackebrandt & Goebel 1994].
- La valeur de 97 % d'identité de séquence a été généralisée pour définir que deux séquences d'ADNr 16S appartiennent à la même espèce
- Ces conclusions ont été tirées de séquences d'ADNr 16S complètes, et non de fragments de gènes tels qu'étudiés en métagénomique ciblée

Limites du clustering

- Différentes régions hypervariables de l'ADNr 16S présentant différents taux de variabilité [Baker et al. 2003], ce seuil fixe de 97 % ne peut être universel pour toutes les régions étudiées
- Ce seuil est ainsi régulièrement remis en question, les OTU générés sur ce critère ne pouvant pas être mis en corrélation avec un niveau taxonomique donné.

  => les lignées bactériennes évoluant à différents rythmes, aucun seuil fixe ne peut représenter la séparation entre toutes les espèces bactériennes [Clarridge 2004, Koeppel & Wu 2013, Rossi-Tamisier et al. 2015].

- Ce seuil ne peut être appliqué à l'étude d'autres cibles génomiques.

  => par exemple, dans des études de métagénomique fongique, les régions ITS utilisées comme marqueurs taxonomiques varient entre 76 % et 99 % d'identité entre espèces

Limites du clustering

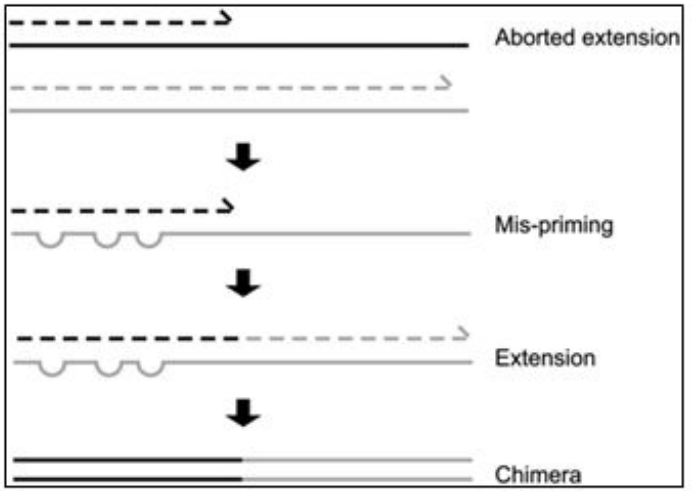
- De nouveaux algorithmes de clustering *de novo* cherchent à contrer cette définition de seuil d’identité en évitant l'utilisation d'un seuil fixe d'identité globale
- Par exemple, SWARM [Mahé et al. 2014] se base sur un nombre maximal de différences locales entre amplicons, ce qui lui permet une formation d'OTU plus fine et les résultats sont indépendants de l'ordre des séquences d'entrée, mais les OTU générés ne peuvent toutefois pas être mis en corrélation avec un rang taxonomique précis

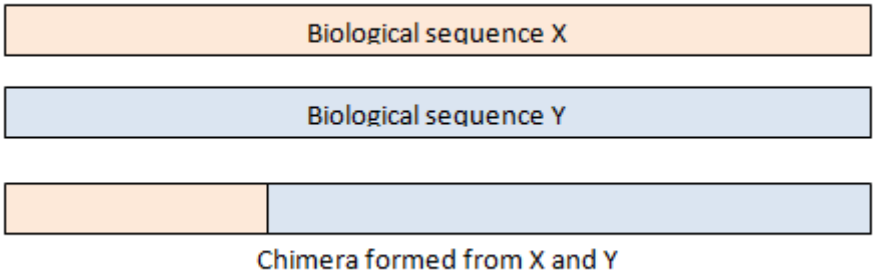
Filtrage après le clustering

- Après avoir formé les OTU, il est possible de filtrer les résultats :
  - En fonction de la taille des OTU : filtrer les singletons (= les OTU qui ne contiennent qu'une seule lecture) ou encore les OTU qui contiennent 2 lectures, 3 lectures ….
  - Filtrage des chimères

Élimination des chimères

- Séquences fusionnées lors des amplifications
- Par alignement sur des banques de référence (Uchime, Decipher, ChimeraSlayer...)
- De novo (Uchime de novo, Perseus...)

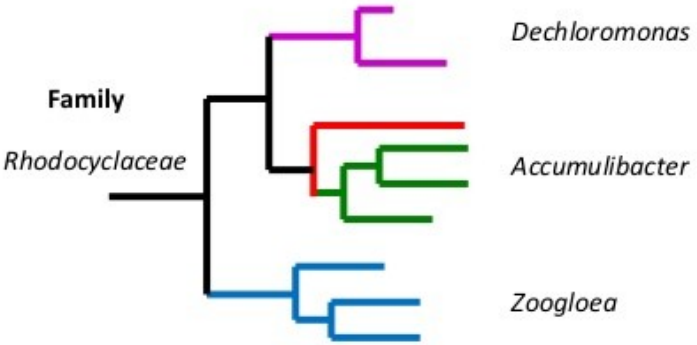




- Pas nécessaire pour la technologie Ion Torrent

Assignation taxonomique

Taxonomic assignment



Class	Order	Family	Genus
Betaproteobacteria	*Rhodocyclales*	*Rhodocyclaceae*	***Accumulibacter***

CENTER FOR MICROBIAL COMMUNITIES | AALBORG UNIVERSITY

- Une fois les OTU formés, une **séquence représentative** est sélectionnée pour chacun d'eux afin de les annoter en comparant cette séquence à **une banque de référence**, cette annotation étant étendue à toutes les lectures appartenant à l'OTU
- Cette comparaison peut être effectuée par un alignement de séquences (de type BLAST), ou encore par une comparaison de k-mers (par exemple avec RDP Classifier [Wang et al. 2007])

Assignation taxonomique

- La séquence représentative pour chaque OTU peut être :
  - la plus longue (contenant le plus de sites informatifs)
  - la plus représentée (contenant potentiellement moins d'erreurs)
  - le centroïde de l'OTU s'il a été défini
  - Une séquence consensus entre toutes les lectures de l'OTU
  - Aléatoire
  - …
- Ce choix dépend fortement de la méthode d'assignation taxonomique envisagée.

  => Par exemple, une séquence consensus n'est pas compatible avec une assignation taxonomique par BLAST, qui ne prend pas en compte les nucléotides dégénérés

Les banques de données

- Le choix d'une banque de séquences de référence est crucial : cette banque doit être :
  - adaptée au locus cible d'intérêt
  - correctement annotée
  - aussi exhaustive que possible
  - doit suivre une taxonomie standardisée
- Il existe trois banques principales de référence pour l'ADNr 16S bactérien

silva
high quality ribosomal RNA databases

GREENGENES
The 16S rRNA Gene Database and Tools

rdp

Les banques de données

	SILVA SSU Parc	SILVA SSU Ref	Greengenes	RDP
Version actuelle	128 (septembre 2016)		13.5 (mai 2013)	11.5 (septembre 2016)
Organismes	Bactéries, archées, eucaryotes		Bactéries, archées	Bactéries, archées
Origine des séquences	European Nucleotide Archive		Genbank	European Nucleotide Archive
Nombre de séquences	5 616 941	1 922 213	1 262 986	3 356 809
Taille minimale des séquences	300	1200 (bactéries/archées) 900 (eucaryotes)	1250	500
Sélection & validation des séquences	Alignement ≥ 50 % d'identité avec au moins une autre séquence de la banque	Alignement ≥ 70 % d'identité avec au moins une autre séquence de la banque	Score d'alignement positif avec au moins une autre séquence de la banque + élimination des séquences chimériques	Au moins 30 % de 7-mers partagés avec une autre séquence de la banque + score d'alignement positif sur un alignement de référence
Taxonomie	SILVA [Yilmaz *et al.* 2013]		Greengenes [McDonald *et al.* 2012]	RDP [Cole *et al.* 2014]
Licence	Utilisation gratuite académique / non-commerciale Licence payante non-académique / commerciale		Creative Commons BY-SA 3.0	Creative Commons BY-SA 3.0
Référence	[Quast *et al.* 2013]		[DeSantis *et al.* 2006]	[Cole *et al.* 2014]

ATTENTION : taxonomies différentes

Eubacterium

SILVA :
Eubacteriaceae

Greengenes :
Erysipelotrichaceae

high quality ribosomal RNA databases

- SILVA propose un ensemble de séquences alignées et annotées des gènes codant pour la grande et la petite sous-unité d'ADNr chez les bactéries, archées et eucaryotes
- Deux versions pour la petite sous-unité (SSU) de l'ADNr :
  - SSU Parc contient plus de 9 millions de séquences
  - SSU Ref contient une sélection curée d'un peu moins de 1 millions de séquences sélectionnées pour leur grande taille et haute qualité d'alignement
  - L'utilisation de cette dernière version permet une plus grande confiance dans les séquences de la banque (qui sont plus longues donc ayant plus de chances de couvrir le locus d'intérêt), aux dépens d'une certaine exhaustivité
  - SILVA met à jour sa banque en ajoutant incrémentalement les nouvelles séquences à l'alignement de la version existante, et à l'arbre taxonomique associé

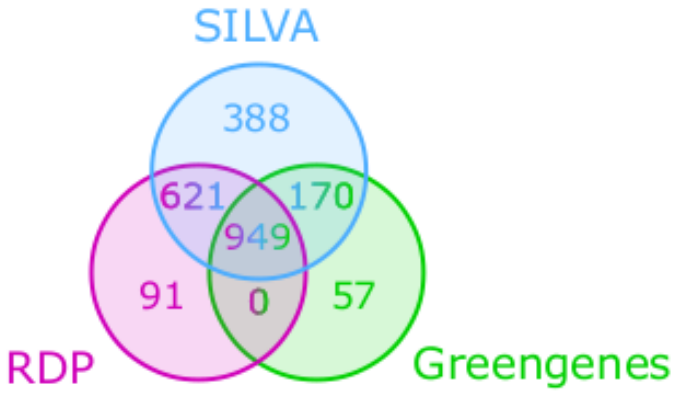
GREENGENES
The 16S rRNA Gene Database and Tools

- Greengene est l'équivalent américain de SILVA proposant des séquences d'ADNr 16S uniquement issues de Genbank contenant un peu plus d'un million de séquences de taille supérieure à 1 250 nucléotides
- Chaque nouvelle version de la banque est accompagnée d'un nouvel alignement complet de toutes ses séquences : cette approche permet de prendre en compte toutes les informations évolutives incluses dans les nouvelles séquences, mais est plus sensible à des séquences erronées ou de moins bonne qualité
- Greengenes a pour particularité de vérifier que chaque nouvelle séquence ajoutée à la banque n'est pas une séquence chimérique (au moins 3 % des séquences publiques d'ADNr seraient en réalité des séquences chimériques !)
- Il est important de noter que Greengenes n'**a pas été mise à jour depuis 2013**, et ne contient ainsi aucune séquence découverte au-delà de cette année

- RDP (Ribosomal Database Project) est une autre initiative américaine proposant dans sa version actuelle un peu plus de 3,3 millions de séquences d'ADNr 16S dont environ 90 000 d'archées
- RDP a la particularité d'avoir un alignement et un arbre taxonomique de référence, basés sur un ensemble restreint de 10 000 séquences issues du séquençage de souches types
- RDP utilise la structure secondaire des séquences d'ARN associées à ces séquences de référence pour guider l'alignement de nouvelles séquences ajoutées à la banque
- Toutefois, **l'assignation taxonomique des séquences de cette banque est limitée au genre**. À noter que RDP propose également une autre banque d'ADNr 28S fongique

Les banques de données

Nombre de genres bactériens partagés



- 2/3 des genres bactériens communs MAIS de nombreux taxons sont spécifiques à chaque banque
- Chaque banque utilise sa **propre taxonomie** et méthode pour classifier de nouvelles séquences, surtout pour les divisions dites « candidates »
- Ces dernières sont des groupes taxonomiques sans souche type cultivable, souvent constituées de séquences environnementales. Leurs séquences d'ADNr 16S sont suffisamment divergentes des séquences existantes pour être considérés comme de potentielles nouvelles branches taxonomiques

Choix de la banque

- Les pipelines sont associés à une banque privilégiée
- Le choix de la banque est guidé par les développeurs des pipelines selon leurs méthodes algorithmiques et/ou des préférences personnelles
- D’autres banques 16S :
  - NCBI 16S database (sous database du NCBI)
  - EzBioCloud 16S database (Gratuit sur demande pour les académiques)
  - ...

Table de comptage (OTU ou ASV)

Phyla	Genera	S1	S2	S3	S4	S5	S6	S7	S8	S9
Firmicutes	Megasphaera	56	32	231	194	37	75	49	71	104
Firmicutes	Acidaminococcus	81	61	80	332	74	57	202	198	3580
Firmicutes	Mitsuokella	519	1	1	0	1	14	2098	840	0
Firmicutes	Selenomonas	0	1	3	1	0	0	3	0	0
Firmicutes	Veillonella	1	1	2	3	9	6		22	0
Firmicutes	Dialister	51	589	1400	29		11	59	233	0
Firmicutes	Ruminococcaceae	415	372	2907	4985	668	1686	1658	1239	2520
Firmicutes	Subdoligranulum	178	55	986	2577	628	448	1796	609	77
Firmicutes	Faecalibacterium	1346	397	20395	31260	4464	1908	6539	7312	7536
Firmicutes	Acetivibrio	18	3	7	7	21	53	22	29	2
Firmicutes	Anaerotruncus	14	43	593	480	134	163	117	130	224
Firmicutes	Ruminococcus	1202	4560	10018	3962	1545	396	626	3121	122

Format BIOM

- Biological Observation Matrix (extension .biom)
- Faciliter le stockage des tables de comptages (optimisation de l’espace mémoire)
- Encapsulation des données de l’étude (table + métadonnées) dans un seul fichier
- Faciliter l’échange entre les différents outils (standardisation du format de sortie)

Les pipelines basés sur les OTU

- Mothur

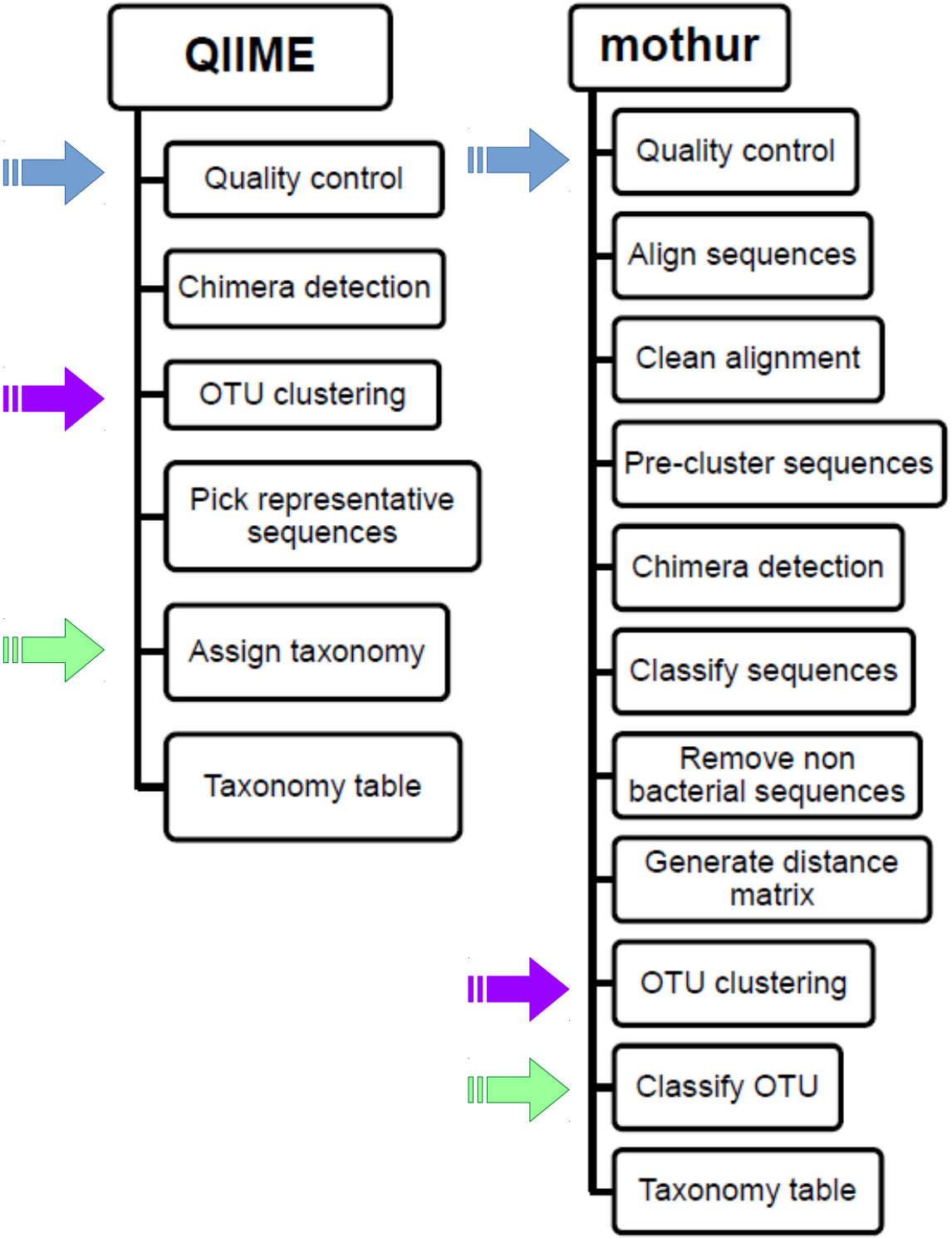
- Un seul programme avec peu de dépendances
- Installation facile
- Pas très flexible

- QIIME

Quantitative Insights Into Microbial Ecology

qiime

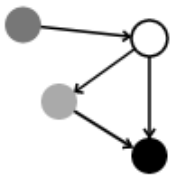
- Interface qui lie beaucoup de programmes
- Beaucoup de dépendances
- Flexible (intégration de scripts personnels)


QIIME
Quality control
Chimera detection
OTU clustering
Pick representative sequences
Assign taxonomy
Taxonomy table
mothur
Quality control
Align sequences
Clean alignment
Pre-cluster sequences
Chimera detection
Classify sequences
Remove non bacterial sequences
Generate distance matrix
OTU clustering
Classify OTU
Taxonomy table


lime
2

QIIME2

- Remplace QIIME1 qui n’est plus maintenu depuis janvier 2018 [Bolyen *et al.* 2019]
- Introduction de nouveaux concepts

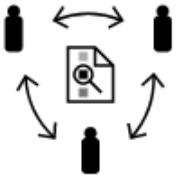


Automatically track your analyses with decentralized data provenance — no more guesswork on what commands were run!

Provenance des fichiers

Interactively explore your data with beautiful visualizations that provide new perspectives.

Visualisation interactive



Easily share results with your team, even those members without QIIME 2 installed.

Partage facilité des résultats

Plugin-based system — your favorite microbiome methods all in one place.

Organisation en plugins

QIIME2

Choose the interface that fits your needs

q2cli the command line interface

2. ~ (zsh)

```
$ qiime info
System versions
Python version: 3.5.3
QIIME 2 release: 2017.6
QIIME 2 version: 2017.6.0
q2cli version: 2017.6.0

Installed plugins
alignment 2017.6.0
composition 2017.6.0
dada2 2017.6.0
```

Artifact API the data scientist's interface

Untitled - idle

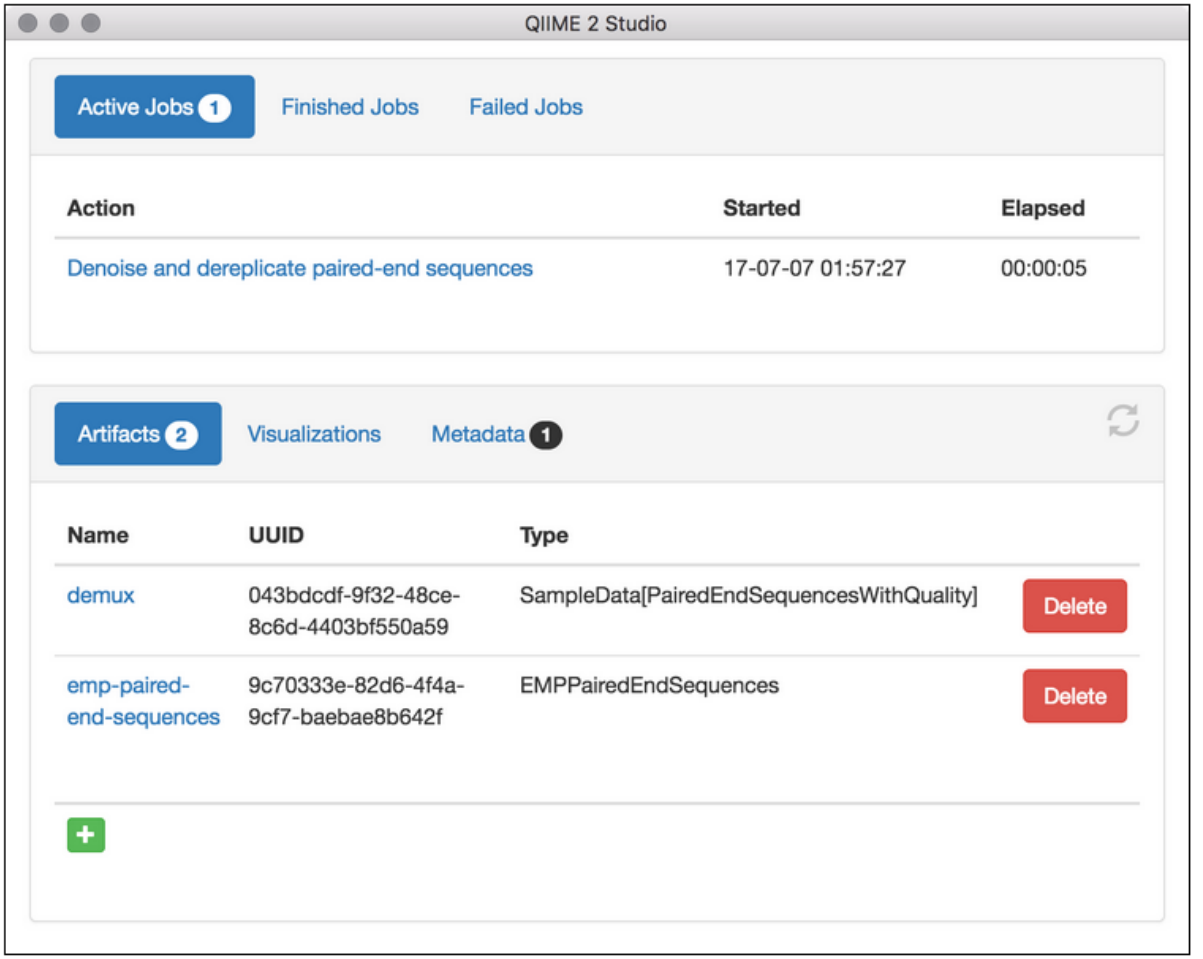
```
[1]  import pandas as pd
     from qiime2 import Artifact

[2]  t = Artifact.load('table.qza')
     t.view(pd.DataFrame)
```

	4b5eeb300368260019c1fbc7a3c718fc
L1S105	2222.0
L1S140	0.0
L1S208	0.0
L1S257	0.0
L1S281	0.0

Python 3 | idle Not saved yet

q2studio the graphical user interface (PROTOTYPE)

q2studio is a functional prototype of a graphical user interface for QIIME 2, and is not necessarily feature-complete with respect to q2cli and the Artifact API.



Concepts

- Fichiers dans QIIME2 : les artefacts
  - Les fichiers artefacts contiennent les données et les métadonnées qui décrivent les données comme leur type, leur format et leur provenance
  - L'extension de ces fichiers est **.qza**
- Besoin d'importer toutes les données en .qza
- Reproductibilité: on sait exactement comment le fichier a été généré

Concepts

- Fichiers de visualisation : extension **.qzv**
- Contiennent des métadonnées similaires aux fichiers artefacts sur la provenance + les données pour la visualisation
- Les fichiers .qzv peuvent être archivés et partagés entre les collaborateurs
- Contrairement aux artefacts, les fichiers de visualisation sont des fichiers de sortie : ils ne peuvent pas être utilisés en entrée pour les analyses

Concepts

- Interface de visualisation dynamique
- **https://view.qiime2.org**
- Permet d’explorer les fichiers QIIME2 (.QZA et .QZV) sans installation de logiciels
- Facilite le partage des données entre collaborateurs

Visualisation

Applications Emplacements Système

denoising-stats × demux.qzv | QII × QIIME 2 View ×

Sécurisé | https://view.qiime2.org

qiime2view

This interface can view .qza and .qzv files directly in your browser without uploading to a server. Click here to learn more.

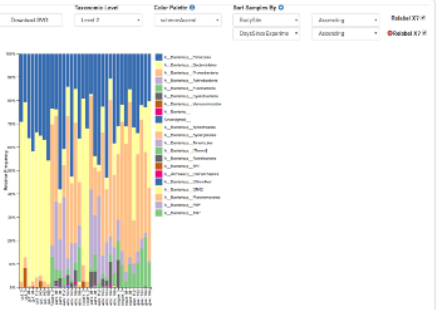
Drag and drop or click here

to view a QIIME 2 Artifact or Visualization (.qza/.qzv) from your computer.

You can also provide a link to a file on Dropbox or a file from the web.

Gallery

Don't have a QIIME 2 result of your own to view? Try one of these!



Taxonomic Bar Plots

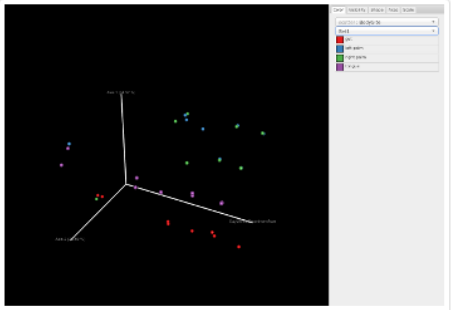
Explore the taxonomy of samples in the Moving Pictures Tutorial. Try selecting different taxonomic levels and metadata-based sample sorting.

Try it!

Explore Sampling Depth

Preview the impact of rarefying your data by manipulating the sampling depth to determine which samples or sample groups would be filtered.

Try it!



3D PCoA with Emperor

View the differences between sample composition using unweighted UniFrac in ordination space. Color the samples by different metadata columns.

Try it!

Visualisation

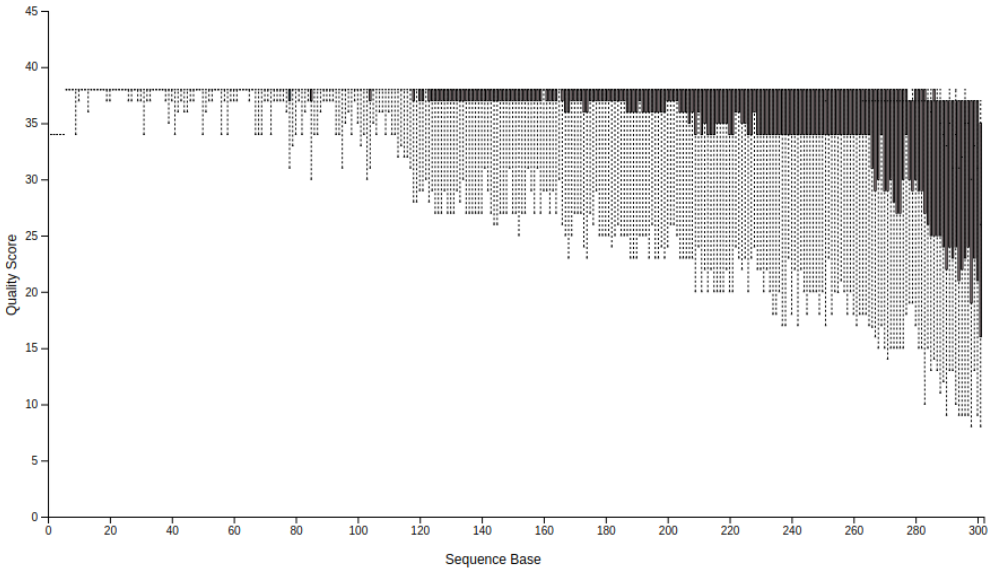
Sécurisé | https://view.qiime2.org/visualization/?type=html&src=64e4f5d0-f03e-4260-8195-f8b236364c2b

qiime2view File: **demux.qzv** Visualization Peek Provenance

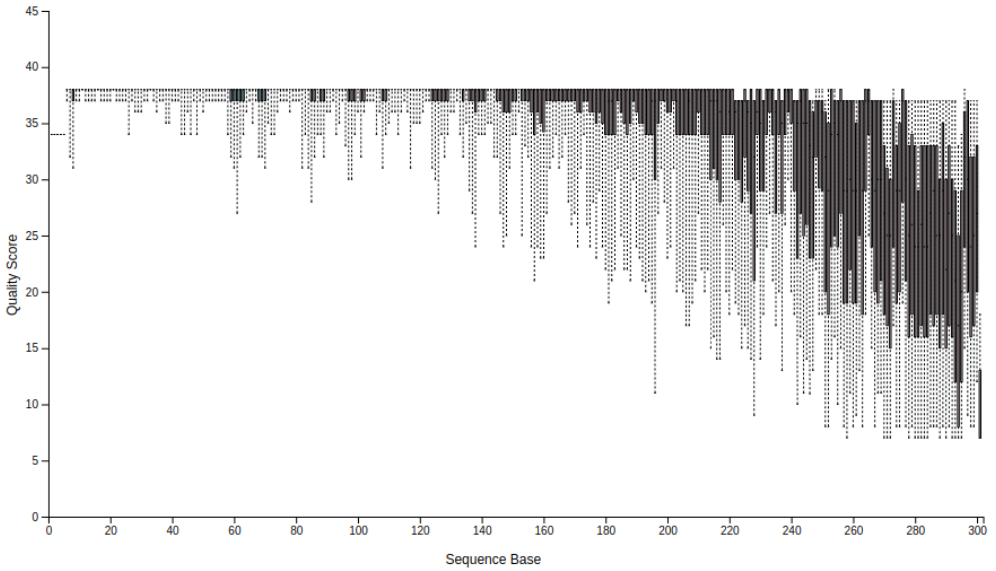
Overview | Interactive Quality Plot

Click and drag on plot to zoom in. Double click to zoom back out to full size. Hover over a box to see the parametric seven-number summary of the quality scores at the corresponding position.

Forward Reads



Reverse Reads



The plot at position 178 was generated using a random sampling of 9983 out of 558772 sequences without replacement. This position (178) is greater than the minimum sequence length observed during subsampling (82 bases). As a result, the plot at this position is not based on data from all of the sequences, so it should be interpreted with caution when compared to plots for other positions. Outlier quality scores are not shown in box plots for clarity.

Parametric seven-number summary for **position 178**

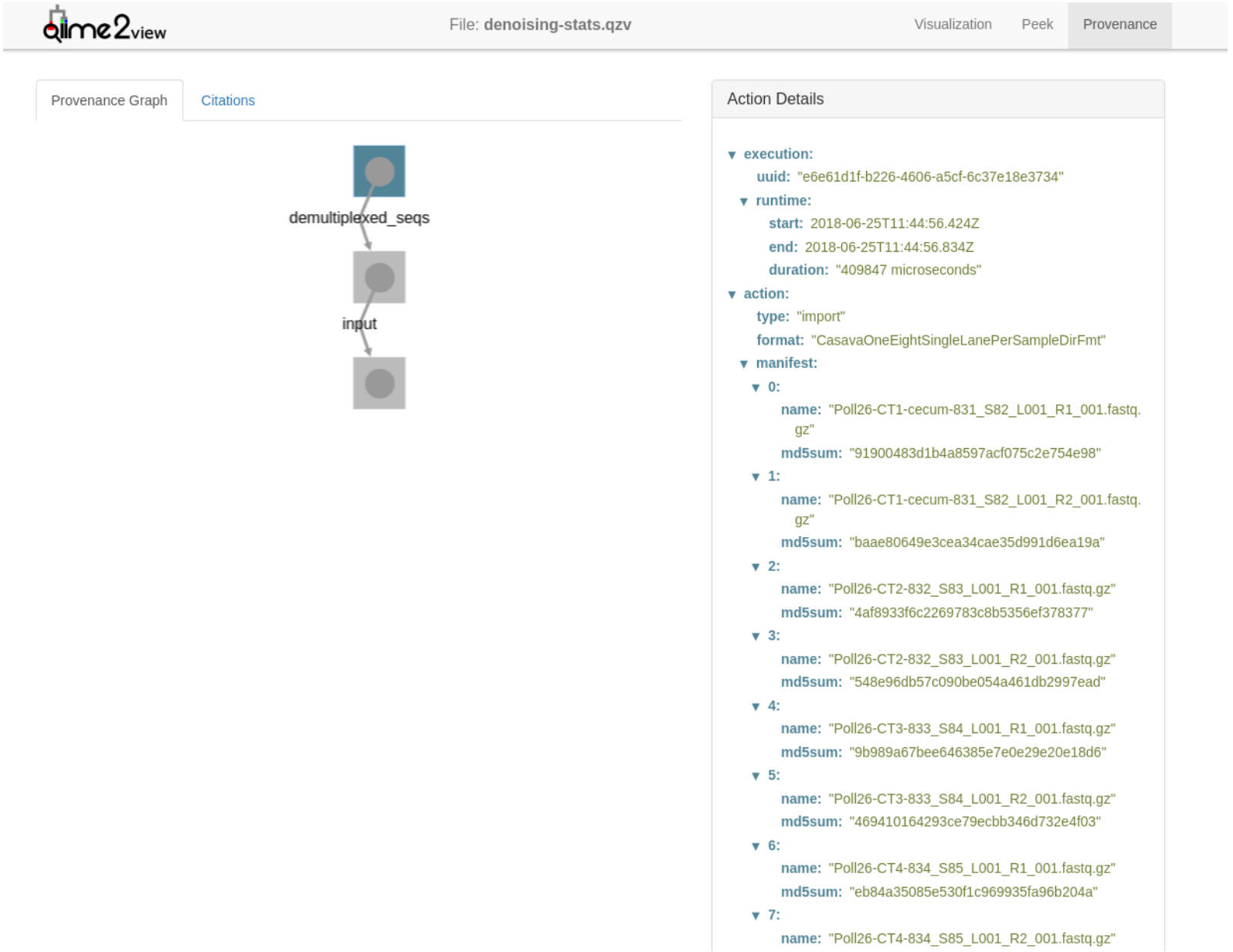
Box plot feature	Percentile	Quality score
(Not shown in box plot)	2nd	9
Lower Whisker	9th	25

These plots were generated using a random sampling of 10000 out of 558772 sequences without replacement. The minimum sequence length identified during subsampling was 82 bases. Outlier quality scores are not shown in box plots for clarity.

Parametric seven-number summary

Box plot feature	Percentile	Quality score
(Not shown in box plot)	2nd	...
Lower Whisker	9th	...
Bottom of Box	25th	...

Visualisation



Les métadonnées

	A	B	C	D	E	F	G	H	I	J	K
1	#SampleID	BarcodeSequence	LinkerPrimerSequence	BodySite	Year	Month	Day	Subject	ReportedAntibioticUsage	DaysSinceExperimentStart	Description
2	#q2:types	categorical	categorical	categorical	numeric	numeric	numeric	categorical	categorical	numeric	categorical
3	L1S8	AGCTGACTAGTC	GTGCCAGCMGCCGCG	gut	2008	10	28	subject-1	Yes	0	subject-1.gut.2008-10-28
4	L1S57	ACACACTATGGC	GTGCCAGCMGCCGCG	gut	2009	1	20	subject-1	No	84	subject-1.gut.2009-1-20
5	L1S76	ACTACGTGTGGT	GTGCCAGCMGCCGCG	gut	2009	2	17	subject-1	No	112	subject-1.gut.2009-2-17
6	L1S105	AGTGCGATGCGT	GTGCCAGCMGCCGCG	gut	2009	3	17	subject-1	No	140	subject-1.gut.2009-3-17
7	L2S155	ACGATGCGACCA	GTGCCAGCMGCCGCG	left palm	2009	1	20	subject-1	No	84	subject-1.left-palm.2009-1-20
8	L2S175	AGCTATCCACGA	GTGCCAGCMGCCGCG	left palm	2009	2	17	subject-1	No	112	subject-1.left-palm.2009-2-17
9	L2S204	ATGCAGCTCAGT	GTGCCAGCMGCCGCG	left palm	2009	3	17	subject-1	No	140	subject-1.left-palm.2009-3-17
10	L2S222	CACGTGACATGT	GTGCCAGCMGCCGCG	left palm	2009	4	14	subject-1	No	168	subject-1.left-palm.2009-4-14
11	L3S242	ACAGTTGCGCGA	GTGCCAGCMGCCGCG	right palm	2008	10	28	subject-1	Yes	0	subject-1.right-palm.2008-10-28
12	L3S294	CACGACAGGCTA	GTGCCAGCMGCCGCG	right palm	2009	1	20	subject-1	No	84	subject-1.right-palm.2009-1-20
13	L3S313	AGTGTCACGGTG	GTGCCAGCMGCCGCG	right palm	2009	2	17	subject-1	No	112	subject-1.right-palm.2009-2-17
14	L3S341	CAAGTGAGAGAG	GTGCCAGCMGCCGCG	right palm	2009	3	17	subject-1	No	140	subject-1.right-palm.2009-3-17
15	L3S360	CATCGTATCAAC	GTGCCAGCMGCCGCG	right palm	2009	4	14	subject-1	No	168	subject-1.right-palm.2009-4-14
16	L5S104	CAGTGTCAGGAC	GTGCCAGCMGCCGCG	tongue	2008	10	28	subject-1	Yes	0	subject-1.tongue.2008-10-28
17	L5S155	ATCTTAGACTGC	GTGCCAGCMGCCGCG	tongue	2009	1	20	subject-1	No	84	subject-1.tongue.2009-1-20
18	L5S174	CAGACATTGCGT	GTGCCAGCMGCCGCG	tongue	2009	2	17	subject-1	No	112	subject-1.tongue.2009-2-17
19	L5S203	CGATGCACCAGA	GTGCCAGCMGCCGCG	tongue	2009	3	17	subject-1	No	140	subject-1.tongue.2009-3-17
20	L5S222	CTAGAGACTCTT	GTGCCAGCMGCCGCG	tongue	2009	4	14	subject-1	No	168	subject-1.tongue.2009-4-14
21	L1S140	ATGGCAGCTCTA	GTGCCAGCMGCCGCG	gut	2008	10	28	subject-2	Yes	0	subject-2.gut.2008-10-28
22	L1S208	CTGAGATACGCG	GTGCCAGCMGCCGCG	gut	2009	1	20	subject-2	No	84	subject-2.gut.2009-1-20
23	L1S257	CCGACTGAGATG	GTGCCAGCMGCCGCG	gut	2009	3	17	subject-2	No	140	subject-2.gut.2009-3-17
24	L1S281	CCTCTCGTGATC	GTGCCAGCMGCCGCG	gut	2009	4	14	subject-2	No	168	subject-2.gut.2009-4-14
25	L2S240	CATATCGCAGTT	GTGCCAGCMGCCGCG	left palm	2008	10	28	subject-2	Yes	0	subject-2.left-palm.2008-10-28
26	L2S309	CGTGCATTATCA	GTGCCAGCMGCCGCG	left palm	2009	1	20	subject-2	No	84	subject-2.left-palm.2009-1-20
27	L2S357	CTAACGCAGTCA	GTGCCAGCMGCCGCG	left palm	2009	3	17	subject-2	No	140	subject-2.left-palm.2009-3-17
28	L2S382	CTCAATGACTCA	GTGCCAGCMGCCGCG	left palm	2009	4	14	subject-2	No	168	subject-2.left-palm.2009-4-14
29	L3S378	ATCGATCTGTGG	GTGCCAGCMGCCGCG	right palm	2008	10	28	subject-2	Yes	0	subject-2.right-palm.2008-10-28
30	L4S63	CTCGTGGAGTAG	GTGCCAGCMGCCGCG	right palm	2009	1	20	subject-2	No	84	subject-2.right-palm.2009-1-20
31	L4S112	GCGTTACACACA	GTGCCAGCMGCCGCG	right palm	2009	3	17	subject-2	No	140	subject-2.right-palm.2009-3-17
32	L4S137	GAACTGTATCTC	GTGCCAGCMGCCGCG	right palm	2009	4	14	subject-2	No	168	subject-2.right-palm.2009-4-14
33	L5S240	CTGGACTCATAG	GTGCCAGCMGCCGCG	tongue	2008	10	28	subject-2	Yes	0	subject-2.tongue.2008-10-28
34	L6S20	GAGGCTCATCAT	GTGCCAGCMGCCGCG	tongue	2009	1	20	subject-2	No	84	subject-2.tongue.2009-1-20
35	L6S68	GATACGTCCTGA	GTGCCAGCMGCCGCG	tongue	2009	3	17	subject-2	No	140	subject-2.tongue.2009-3-17
36	L6S93	GATTAGCACTCT	GTGCCAGCMGCCGCG	tongue	2009	4	14	subject-2	No	168	subject-2.tongue.2009-4-14
37											

Format= TSV texte séparé par des tabulations

Colonne identifier (ID) [obligatoire]

#SampleID #Sample ID #OTUID

#OTU ID sample_name

	A	B	C	D	E	F	G	H	I	J	K
1	#SampleID	BarcodeSequence	LinkerPrimerSequence	BodySite	Year	Month	Day	Subject	ReportedAntibioticUsage	DaysSinceExperimentStart	Description
2	#q2:types	categorical	categorical	categorical	numeric	numeric	numeric	categorical	categorical	numeric	categorical
3	L1S8	AGCTGACTAGTC	GTGCCAGCMGCCGC	gut	2008	10	28	subject-1	Yes	0	subject-1.gut.2008-10-28
4	L1S57	ACACACTATGGC	GTGCCAGCMGCCGC	gut	2009	1	20	subject-1	No	84	subject-1.gut.2009-1-20
5	L1S76	ACTACGTG									
6	L1S105	AGTGCGAT									
7	L2S155	ACGATGCG									
8	L2S175	AGCTATCC									
9	L2S204	ATGCAGCT									
10	L2S222	CACGTGAC									
11	L3S242	ACAGTTGC									
12	L3S294	CACGACAG									
13	L3S313	AGTGTCAC									
14	L3S341	CAAGTGAG									
15	L3S360	CATCGTAT									
16	L5S104	CAGTGTCA									
17	L5S155	ATCTTAGA									
18	L5S174	CAGACATT									
19	L5S203	CGATGCAC									
20	L5S222	CTAGAGAC									
21	L1S140	ATGGCAGC									
22	L1S208	CTGAGATA									
23	L1S257	CCGACTGA									
24	L1S281	CCTCTCGT									
25	L2S240	CATATCGCA									
26	L2S309	CGTGCATTA									
27	L2S357	CTAACGCA									
28	L2S382	CTCAATGA									
29	L3S378	ATCGATCT									
30	L4S63	CTCGTGGA									
31	L4S112	GCGTTACA									
32	L4S137	GAACTGTA									
33	L5S240	CTGGACTC									
34	L6S20	GAGGCTCA									
35	L6S68	GATACGTC									
36	L6S93	GATTAGCACTCT	GTGCCAGCMGCCGC	tongue	2009	4	14	subject-2	No	168	subject-2.tongue.2009-4-14
37											

- les ID ne doivent pas commencer par # (symbole utilisé pour une ligne de commentaire et ignorée)
- les ID ne peuvent pas être vide (au moins 1 caractère)
- les ID doivent être uniques

- 36 caractères ou moins
- seulement des caractères alpha-numériques: [a-z], [A-Z], or [0-9]), (.) ou (-)

Les métadonnées

	A	B	C	D	E	F	G	H	I	J	K
1	#SampleID	BarcodeSequence	LinkerPrimerSequence	BodySite	Year	Month	Day	Subject	ReportedAntibioticUsage	DaysSinceExperimentStart	Description
2	#q2:types	categorical	categorical	categorical	numeric	numeric	numeric	categorical	categorical	numeric	categorical
3	L1S8	AGCTGACTAGTC	GTGCCAGCMGCCGC	gut	2008	10	28	subject-1	Yes	0	subject-1.gut.2008-10-28
4	L1S57	ACACACTATGGC	GTGCCAGCMGCCGC	gut	2009	1	20	subject-1	No	84	subject-1.gut.2009-1-20
5	L1S76	ACTACGTGTGGT	GTGCCAGCMGCCGC	gut	2009	2	17	subject-1	No	112	subject-1.gut.2009-2-17

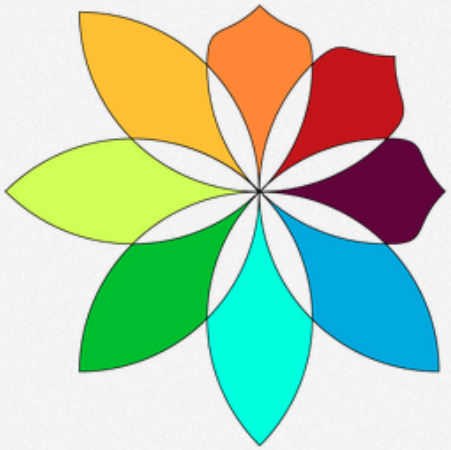
D’autres colonnes pour ajouter des métadonnées, 0 ou autant qu’on le souhaite

Les cellules vide sont considérées comme des données manquantes
Attention : NA n’est pas interprété comme des données manquantes

La seconde ligne sert à indiquer le type des métadonnées :
La première cellule doit être **#q2:types** pour indiquer que la ligne est un commentaire contenant des directives.

Deux valeurs possibles :

- **categorical** : les métadonnées sont des catégories
- **numeric** : les métadonnées sont des valeurs numériques



Keemei: Validate tabular bioinformatics file formats in Google Sheets

Keemei (canonically pronounced *key may*) is an open source Google Sheets add-on for validating tabular bioinformatics file formats, including QIIME 2 metadata files.

Keemei supports validating the following file formats:

- QIIME 2 metadata files
- QIIME 1 mapping files
- Qiita sample information files
- SRGD files (e.g., for use with geneGIS)

If you use Keemei for any published research, please include the following citation:

> **Keemei: cloud-based validation of tabular bioinformatics file formats in Google Sheets.**
> Rideout JR, Chase JH, Bolyen E, Ackermann G, González A, Knight R, Caporaso JG.
> GigaScience. 2016;5:27. http://dx.doi.org/10.1186/s13742-016-0133-6

Find the Keemei paper here.

https://keemei.qiime2.org/

QIIME2

- Intégration de nouvelles méthodes pour les différentes étapes (assignation taxonomique, analyses secondaires...)
- Intégration d’un nouveau concept : le clustering (les OTU) est remplacé par des méthodes basées sur le débruitage (les ASV)

Plugins

Version: 2019.10

Table of Contents

Quick search

Go

Plugins

The following pages describe what QIIME 2 plugins are available and how to develop a new plugin.

- Available plugins
  - alignment: Plugin for generating and manipulating alignments.
  - composition: Plugin for compositional data analysis.
  - cutadapt: Plugin for removing adapter sequences, primers, and other unwanted sequence from sequence data.
  - dada2: Plugin for sequence quality control with DADA2.
  - deblur: Plugin for sequence quality control with Deblur.
  - demux: Plugin for demultiplexing & viewing sequence quality.
  - diversity: Plugin for exploring community diversity.
  - emperor: Plugin for ordination plotting with Emperor.
  - feature-classifier: Plugin for taxonomic classification.
  - feature-table: Plugin for working with sample by feature tables.
  - fragment-insertion: Plugin for extending phylogenies.
  - gneiss: Plugin for building compositional models.
  - longitudinal: Plugin for paired sample and time series analyses.
  - metadata: Plugin for working with Metadata.
  - phylogeny: Plugin for generating and manipulating phylogenies.
  - quality-control: Plugin for quality control of feature and sequence data.
  - quality-filter: Plugin for PHRED-based filtering and trimming.
  - sample-classifier: Plugin for machine learning prediction of sample metadata.
  - taxa: Plugin for working with feature taxonomy annotations.
  - types: Plugin defining types for microbiome analysis.
  - vsearch: Plugin for clustering and dereplicating with vsearch.
- Future plugins
- Developing a QIIME 2 plugin
  - Overview
  - Plugin components
  - Testing your plugin with q2cli during development
  - Plugin testing
  - Advanced plugin development
  - Example plugins

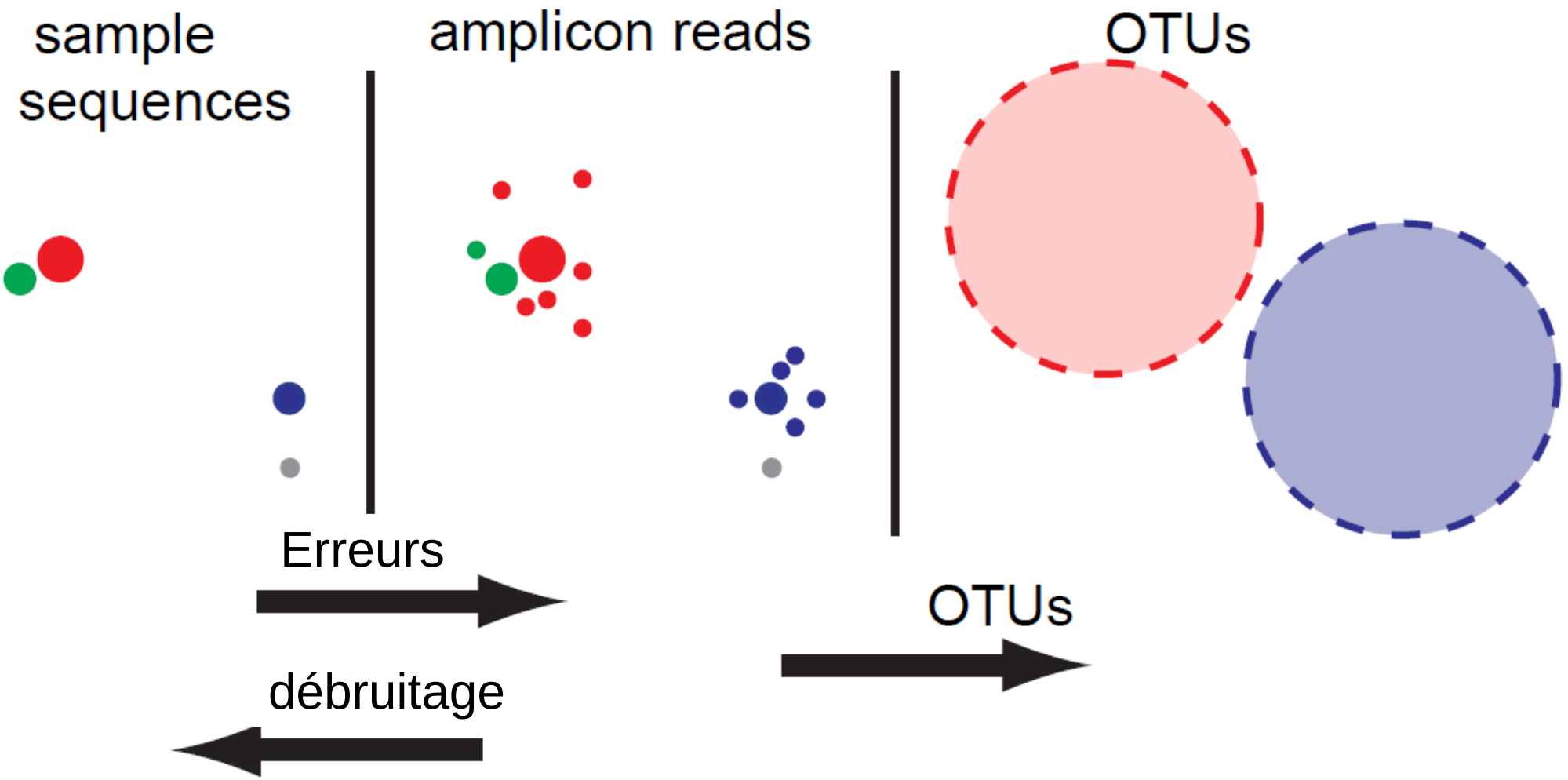

Partie
Exercice

Vers les ASV

- Les erreurs de séquençage impliquent de nombreux biais lors de la génération des OTU
- Idée : corriger les reads en fonction du modèle d'erreur de la technologie de séquençage

  => inférer statistiquement les séquences « réelles=corrigées » de l'échantillon
- Les lectures, ainsi que leur variants, sont regroupés en **ASV = Amplicon Sequence Variant** un taxon étant ensuite assigné à chaque ASV

Vers les ASV


sample
sequences
amplicon reads
OTUs
Erreurs
débruitage
OTUs

Vers les ASV

- Ces méthodes sont plus sensibles et permettent d'atteindre une meilleure résolution taxonomique tout en générant moins de faux positifs et en s'affranchissant de la notion de seuil de similarité
- Parmi ces méthodes, **DADA2** [Callahan et al., 2016] et **Deblur** [Amir et al., 2017] sont déjà largement utilisés pour l'analyse des données de métagénomique ciblée

DADA2

- DADA2 est un pipeline pour **détecter** et **corriger** les séquences d'amplicon Illumina.
- Filtrage des reads phiX (marqueur de contrôle qualité utilisé pour la technologie Illumina)
- Filtrage des **séquences chimériques**

DADA2

- DADA2 implémente un nouvel algorithme en **modélisant les erreurs** et utilise ce modèle pour inférer la **composition réelle de l'échantillon**
- DADA2 remplace le traditionnel OTU-picking en produisant à la place des tables de plus haute résolution de variants de séquences d'amplicon (ASV).
- La méthode est **plus sensible et spécifique** que les OTU : DADA2 détecte plus de variations biologiques réelles non identifiées avec les OTU alors qu'ils retourne moins de fausses séquences
- Données d'entrée : fichiers FASTQ demultiplexés

DADA2 : algorithme

- Divisive Amplicon Denoising Algorithm
- Première étape : les séquences sont comparées entre elles
- Heuristiques :
  - Calcul des distances entre les k-mers
    - si plus de 10 % de mismatches pas d'alignement
    - sinon alignement dans une bande (cf heuristique FASTA)

Séquence i : ATCGTACTGCTCTGA
Séquence j : ATTGTACTG - TCTGA

DADA2 : algorithme

- Taux d'erreurs : taux auquel un amplicon avec la séquence i est produit à partir de la séquence j = produit des probabilités de transition entre les L nucléotides alignés

$$\lambda_{ji} = \prod_{l=0}^{L} p(j(l) \rightarrow i(l), q_i(l))$$

Séquence i : ATCGTACTGCTCTGA
Séquence j : ATTGTACTG - TCTGA

p(T→ C, 35)

DADA2 peut estimer cette matrice de transition 16X41 à partir des donnée

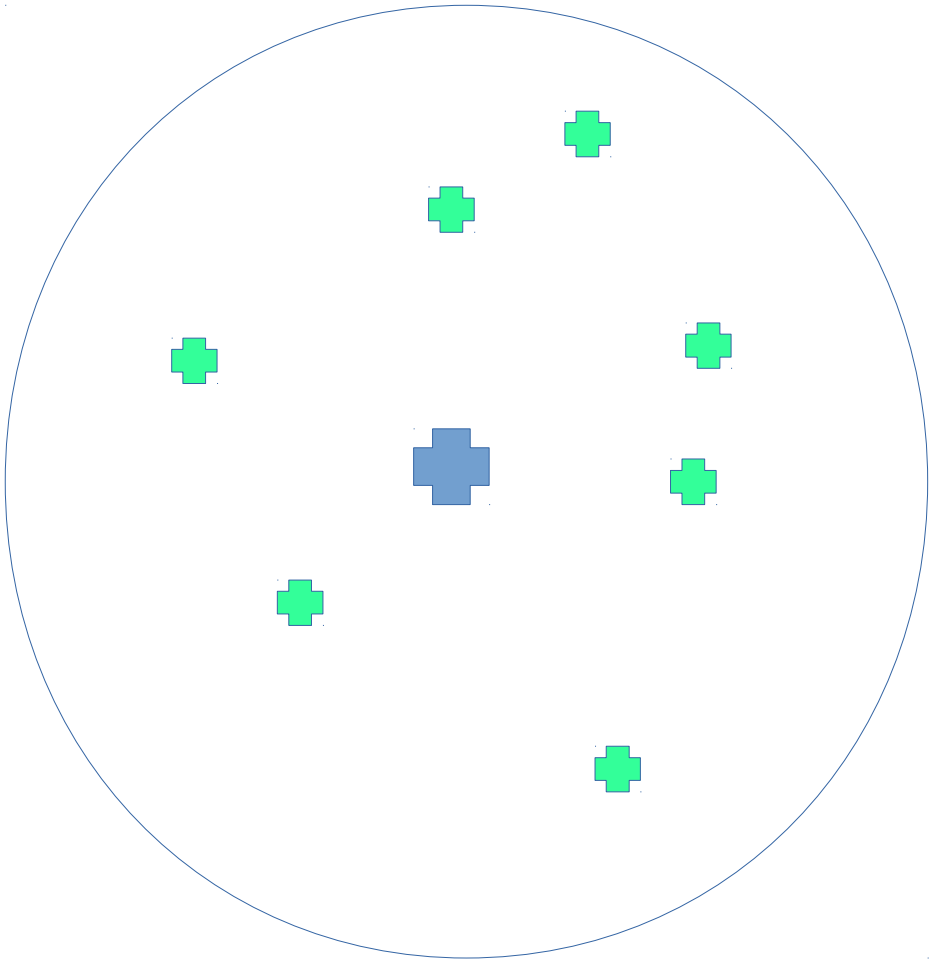
DADA2 : algorithme

- P-valeur d'abondance : quantifier la notion qu'une séquence i est trop abondante pour être expliquée par les erreurs de séquençage
- Les erreurs de séquençage sont modélisées par une loi de Poisson
- Une faible Pa(j → i) signifie qu'il y a plus de reads de séquence i que pouvant être expliquée par les erreurs de séquençage des nj copies de la séquence j

  = > Si Pa est faible la lecture correspond certainement un autre amplicon sinon c'est un variant

$$p_A(j \to i) = \frac{1}{1 - \rho_{\text{pois}}(n_j \lambda_{ji}, 0)} \sum_{a=a_i}^{\infty} \rho_{\text{pois}}(n_j \lambda_{ji}, a)$$

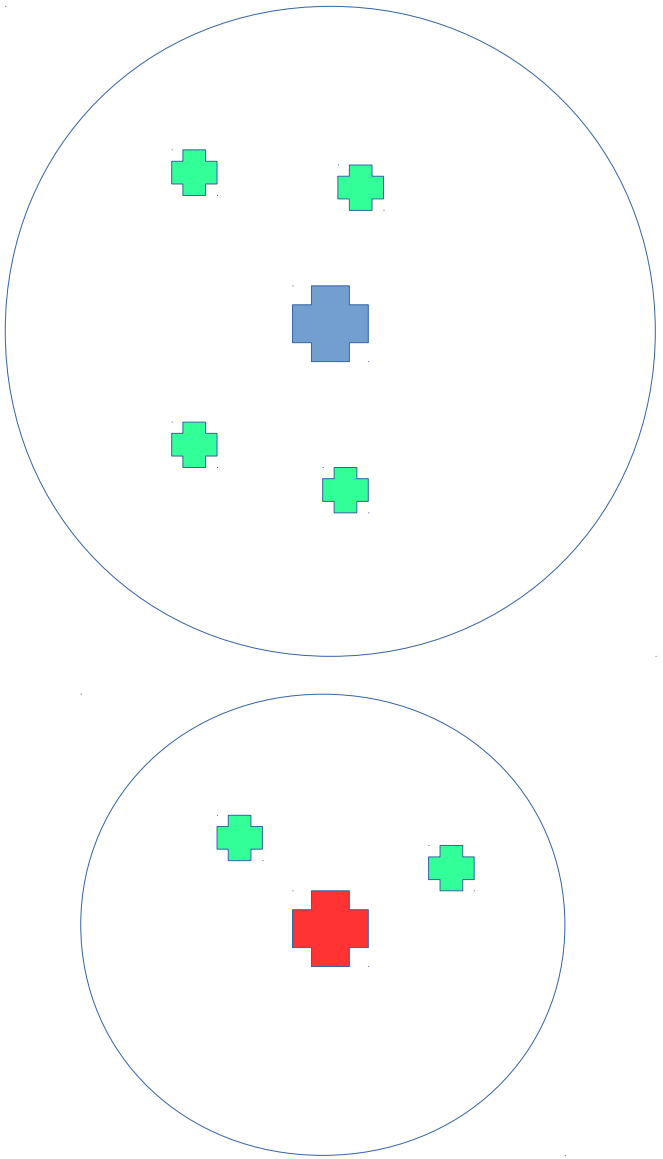
DADA2 : algorithme



Centroïde
= read le plus abondant

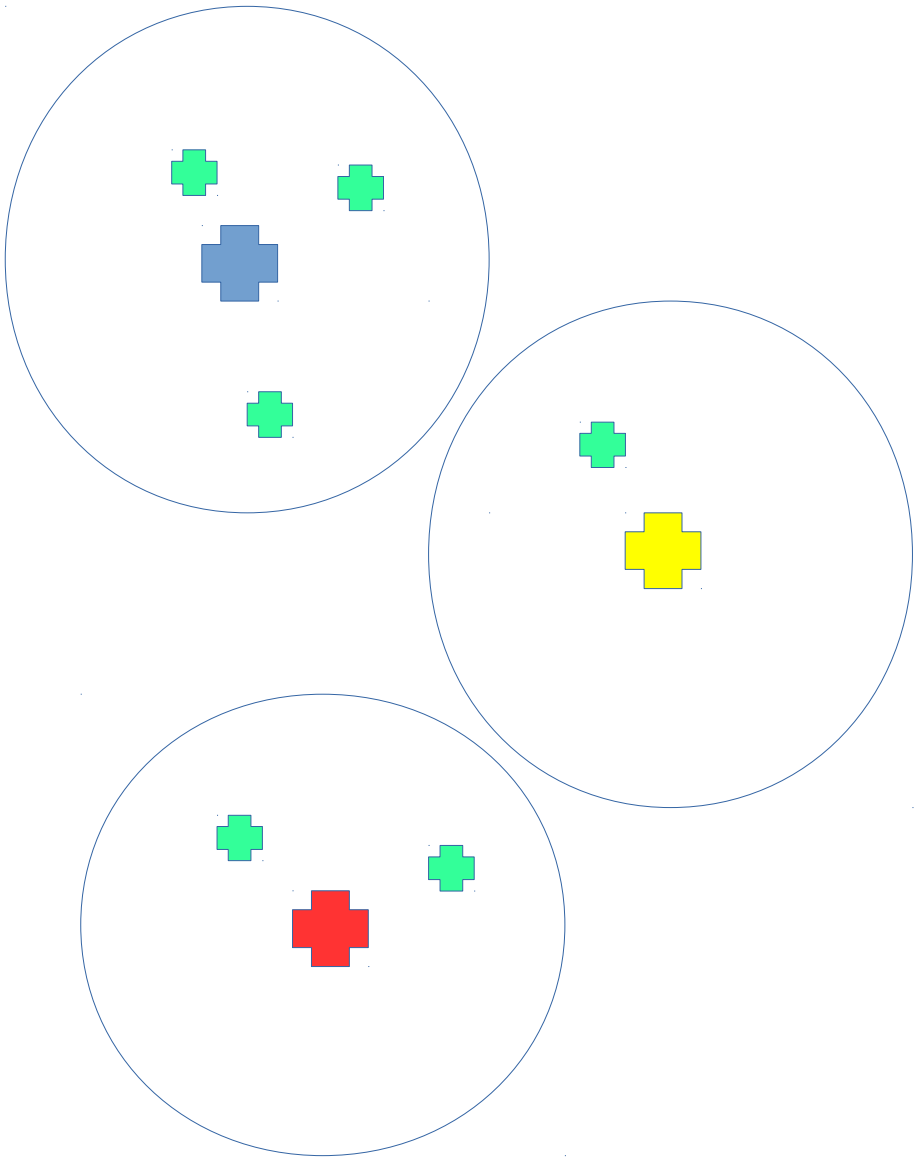
- Une seule partition
- Toutes les séquences sont comparées au centroïde
- Les taux d'erreurs et les probabilités d'abondance sont calculés
- Si la plus petite Pa est inférieure au seuil une nouvelle partition est créée

DADA2 : algorithme



- Toutes les séquences sont comparées au centroïde de la nouvelle partition et rejoignent la partition la plus proche
- Itération (plus petite p-valeur d'abondance, nouvelle partition, nouvelle répartition)

DADA2 : algorithme



- FIN : plus de p-valeurs d'abondance inférieure au seuil => toutes les partitions représentent un ensemble de reads provenant de la même séquence (=variants dus aux erreurs de séquençage)
- Chaque read est débruité avec son centroïde
- La composition de l'échantillon est l'ensemble des centroïdes avec l'abondance totale correspondant à la taille de la partition

DADA2 : warning

- Les reads paired-end ne sont pas mergés à cause de l'utilisation de la qualité base par base, ils seront mergés après
- Il faut que la séquence soit observée au moins une fois donc les singletons ont une probabilité d'abondance de 1 et ne peuvent pas formé leur propre partition donc **aucun singleton n'est inféré**

DADA2: pipeline

1) Filter and trim: filterAndTrim()
2) Dereplicate: derepFastq()
3) Learn error rates: learnErrors()
4) Infer sample composition: dada()
5) Merge paired reads: mergePairs()
6) Make sequence table: makeSequenceTable()
7) Remove chimeras: removeBimeraDenovo()

DADA2: pipeline

1) Filter and trim: filterAndTrim()

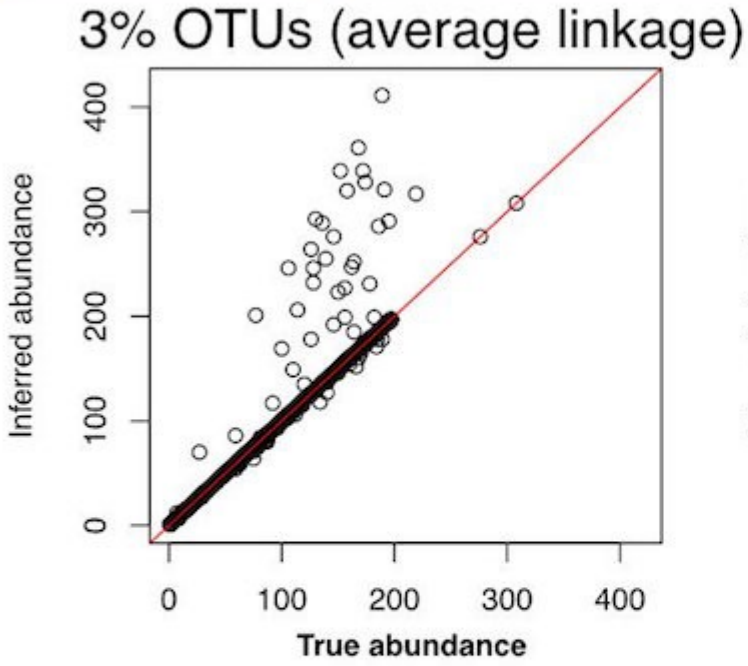
- Trimme les reads à une taille donnée
- Filtrage des reads trop petits
- Filtrage des reads contenant des bases ambiguës
- Filtrage des reads en fonction de la qualité

DADA2: pipeline

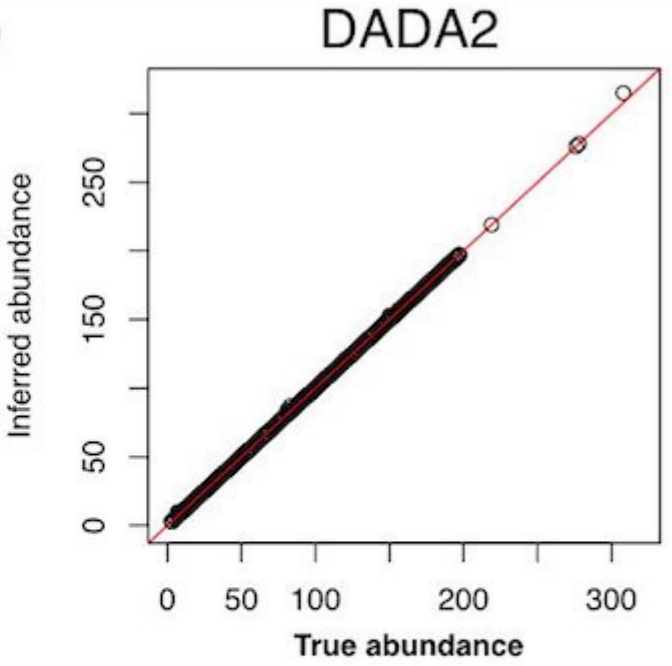
1) Filter and trim: filterAndTrim()
2) Dereplicate: derepFastq()
3) Learn error rates: learnErrors()
4) Infer sample composition: dada()
5) Merge paired reads: mergePairs()
6) Make sequence table: makeSequenceTable()
7) Remove chimeras: removeBimeraDenovo()

DADA2 : précision

Accuracy: Simulated data



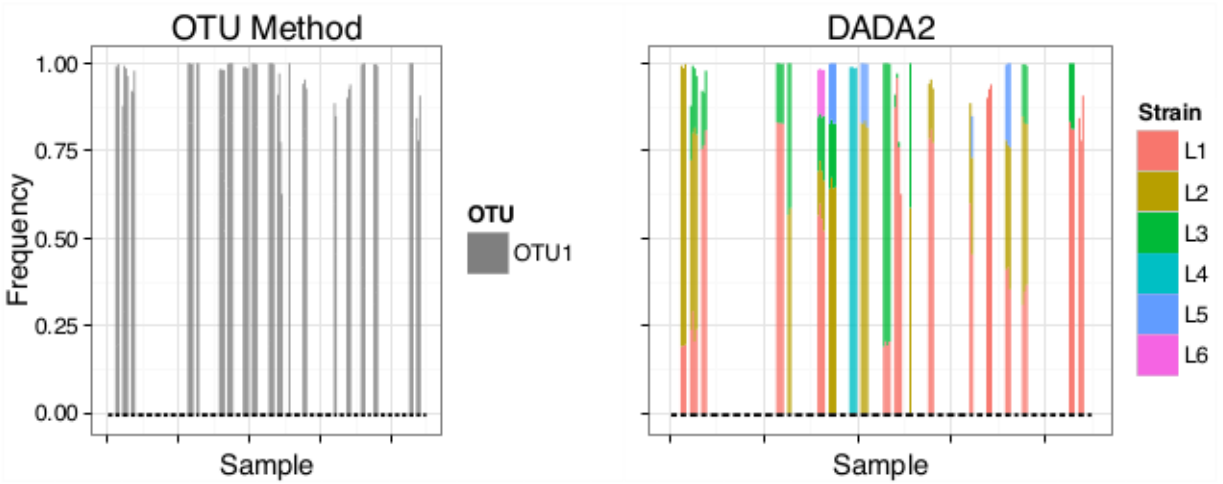
TP: 978
FP: 272
FN: 77
cor: 0.935



TP: 1042
FP: 0
FN: 13
cor: 0.999

Data: Kopylova, et al. mSystems, 2016.

DADA2 : sensibilité



- Etude d'échantillons de flore vaginale = flore la moins diverse du corps humain souvent dominée par un seul OTU Lactobacillus
- *L. crispatus* est l'espèce la plus abondante et synonyme de bonne santé
- L'analyse des données avec DADA2 a montré que la communauté *L. crispatus* est plus complexe qu'admis jusque là avec **6 variants différents** de 1 ou 2 nucléotides (L1 à L6) imperceptibles avec les OTU

reads_1.fastq reads_2.fastq

reads_1.fastq.gz reads_2.fastq.gz

pre-process

Merging des reads
PEAR

Evaluation des reads et détermination des seuils (FASTQC)

Formatage des fichiers
Re-multiplexage des reads FASTA

Filtrage et trimming
split_library.py

Détection des reads chimériques
USEARCH

pre-process

Importation des données
.qza

demux summarize

analyse primaire

OTU picking
pick_open_reference_otus.py (-m sortmerna_sumaclust)
greenGenes 97%
=> assignation taxonomique
=> sequences représentatives
=> arbres phylogénétiques

Filtrage des singletons
=> BIOM

Normalisation
CSS => BIOM normalisé

analyse primaire

Contrôle qualité et constructuction de la table d'ASV
=> table
=> rep_seq
=> stats

arbres phylogénétiques
=> unrooted tree
=> rooted tree

Assignation taxonomique

analyse secondaire

alpha-diversité
- Indice de richesse Chao1
- Indices de diversité Shannon & Simpson
- Comparaison des groupes (KRUSKAL-WALLIS pour significativité entre les groupes et MWW entre 2 groupes)

beta-diversité
- UniFrac: weighted et unweighted
- test statistique pour évaluer la différences entre les communauté (ADONIS)

Étude de la composition taxonomique
(au niveau de la famille et du genre)
- dans QIIME
- comparaison de 2 groupes avec le testWelch's test two-side et une correction benjamini FDR(STAMP)

analyse secondaire

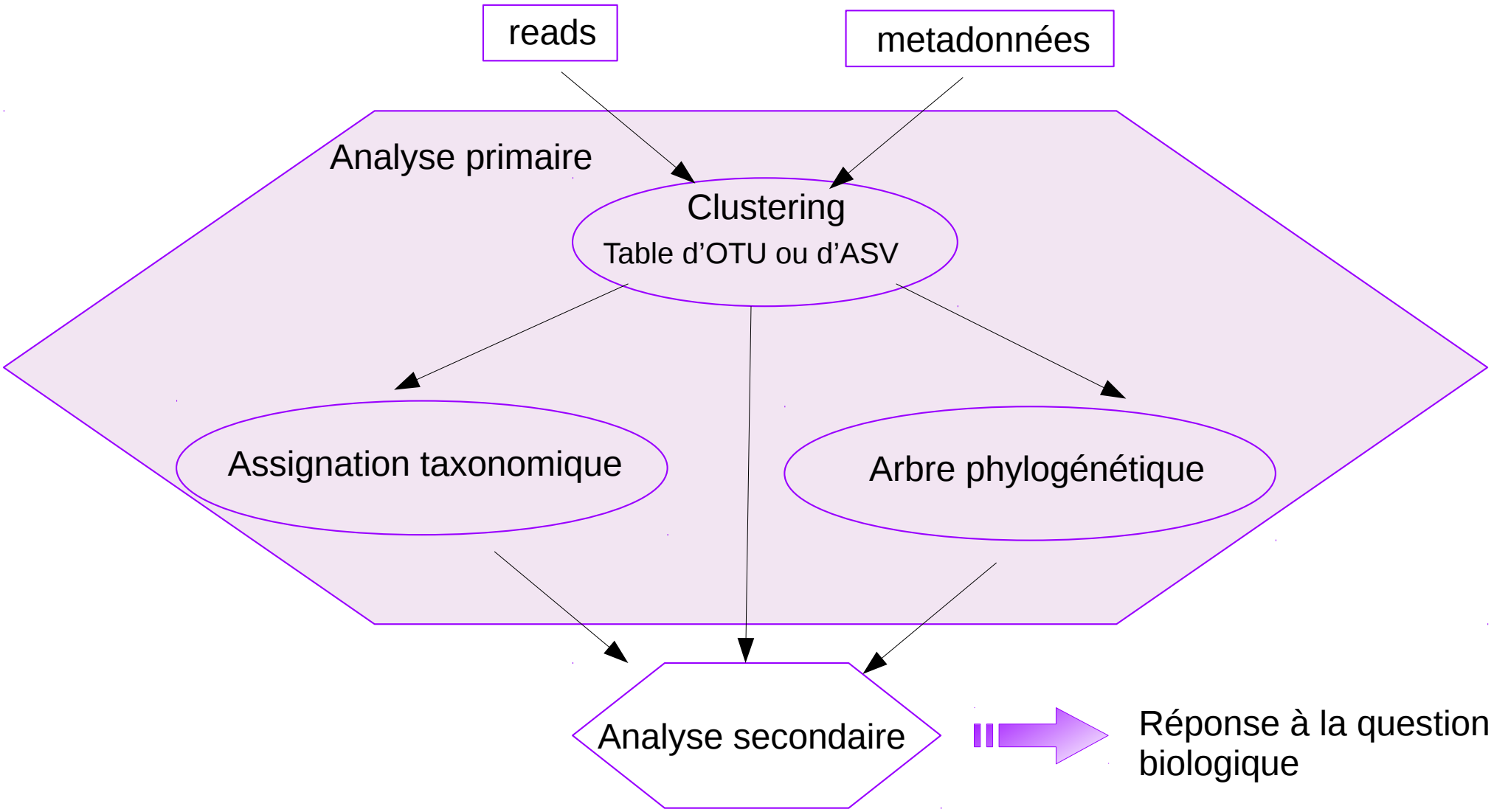
Alpha et beta diversité
core-metrics-phylogenetic

Alpha rarefaction plotting

Taxonomic analysis

Differential abundance testing
with ANCOM

Analyse des données




Partie
Exercice

références

- http://dridk.me/metagenomique.html
- Manuscrit de thèse de Léa Siegwald