

L'infection broncho-pulmonaire à *Pseudomonas aeruginosa* revisitée sous l'angle du microbiote respiratoire

Création d'un pipeline bioinformatique d'analyse des
séquences ARNr 16S appliqué au microbiote respiratoire

Sacha Schutz
Interne en génétique moléculaire
Bioinformaticien
RICAI 2016



Plan

- Introduction
- Matériel et méthodes
- ▼ Résultats
- Conclusion

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

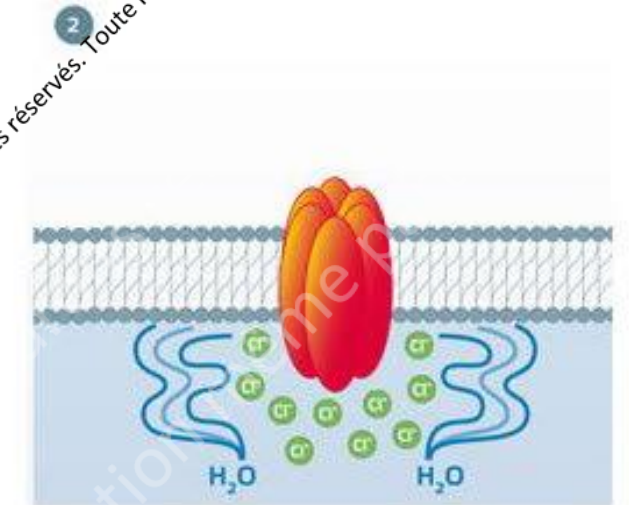
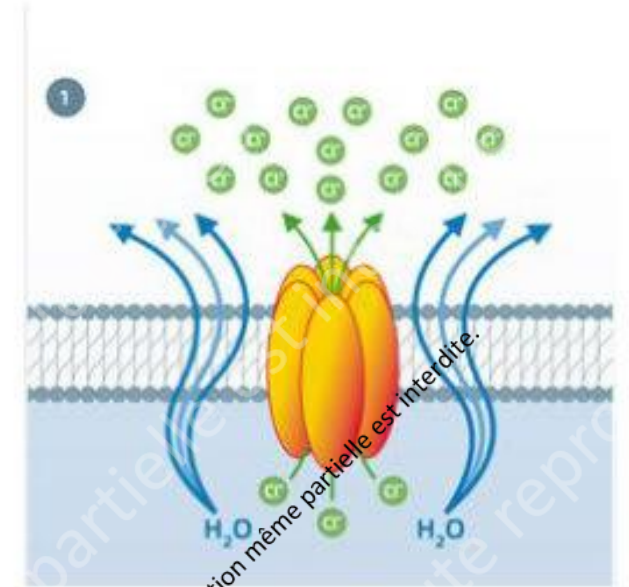
Introduction

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

La mucoviscidose

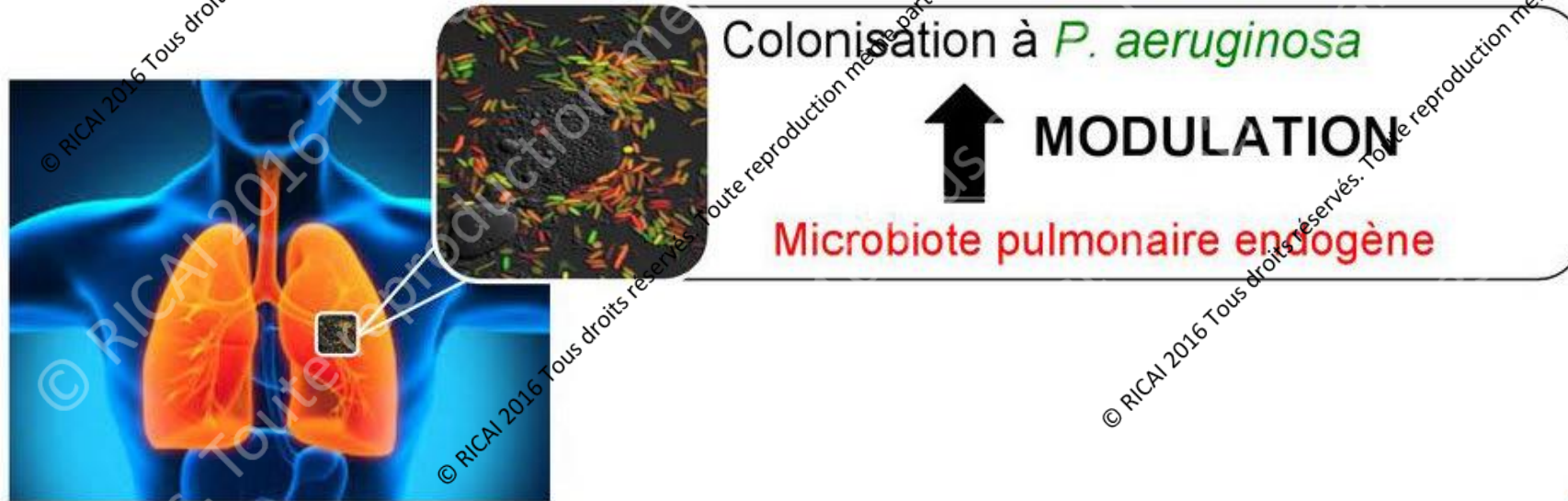
- Maladie génétique **autosomique récessive**
- 1ère maladie génétique de l'enfant
- Mutation du gène **CFTR**
- 1 personne sur 30 est porteuse de la mutation
- Sécrétion par l'épithélium bronchique d'un mucus visqueux propice aux infections bactériennes
- Colonisation à ***Pseudomonas aeruginosa* (PA)** : marque un tournant décisif dans la maladie
- Maladie **génétique** et **infectieuse**



Rationnel de l'étude Mucobiome

- Âge de la primocolonisation à *P. aeruginosa* variable
- Phénotypes extrêmes : < 4 ans; > 20 ans

Hypothèse :



- **Objectif :**

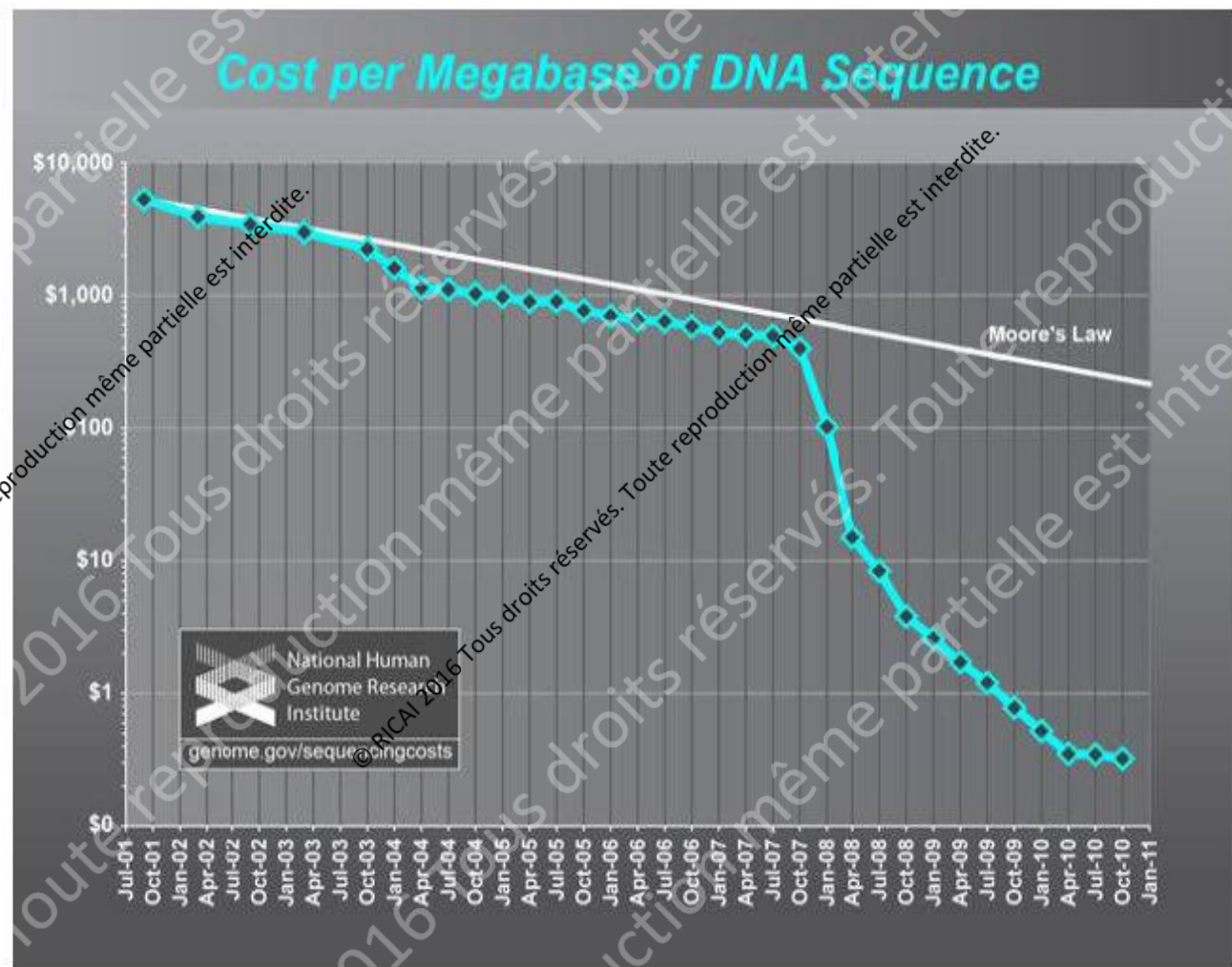
Identifier les déterminants microbiens de la primocolonisation à *P. aeruginosa*

Définitions

- Le **microbiote** est l'ensemble des micro-organismes (**bactéries**, levures, champignons, virus ...) vivant dans un environnement spécifique.
- Le **microbiome** est l'ensemble des **génomés** présents dans un environnement spécifique. (ex: le microbiome pulmonaire)
- La **métagénomique** est une méthode d'étude du microbiome par le **séquençage haut débit** de l'ensemble des ADNs présents dans un environnement spécifique.

Séquençage haut débit

- Le génome humain (3Gb)
 - 1990 - 2003
 - Technologie Sanger
 - 13 ans
 - 3 milliards de dollars
 - 2016
 - Technologie *Illumina HiSeq X*
 - 3 jours
 - 1000 dollars



© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

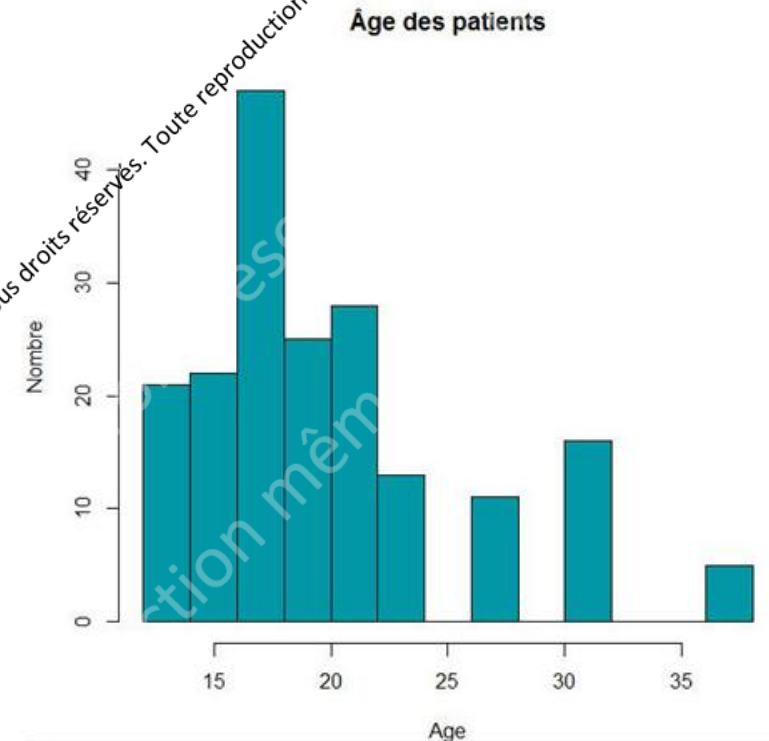
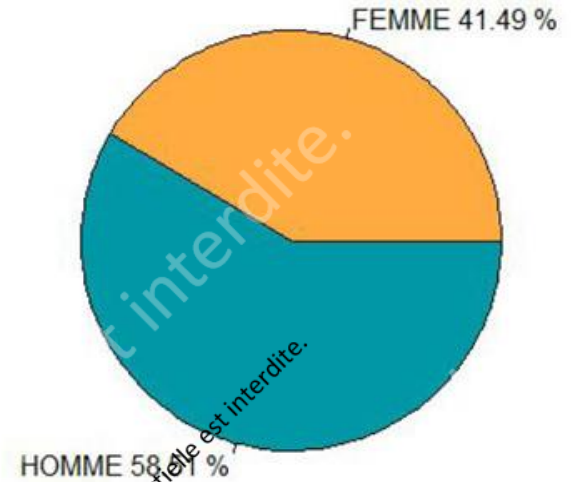
Matériel et Méthodes

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

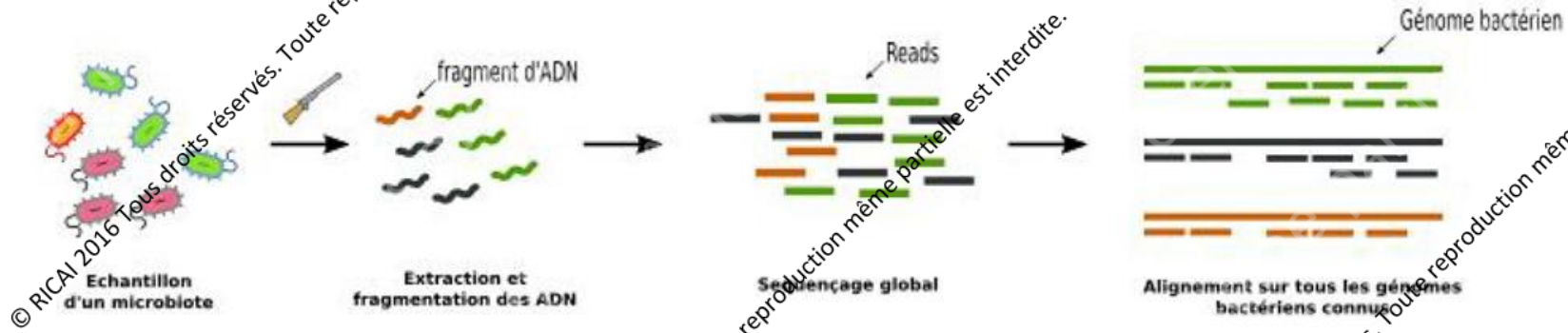
Protocole de l'étude (MUCOBIOME)

- Etude prospective (3 ans) multicentrique (Brest, Roscoff, Nantes)
- **61** patients atteints de mucoviscidose
- Critère d'inclusion : exempt de PA depuis au moins 1 an
- **188** expectorations recueillies (-80°C)
- Données clinico-biologiques :
 - Catégorie (Lee et al., 2003)
 - **Never**: Pas d'antécédent d'infection à PA
 - **Free**: Antécédent d'infection à PA mais datant de plus de 1 an
 - Âge (âge médian = 20)
 - Sexe, Indice de masse corporelle, Antibiothérapie
 - ...

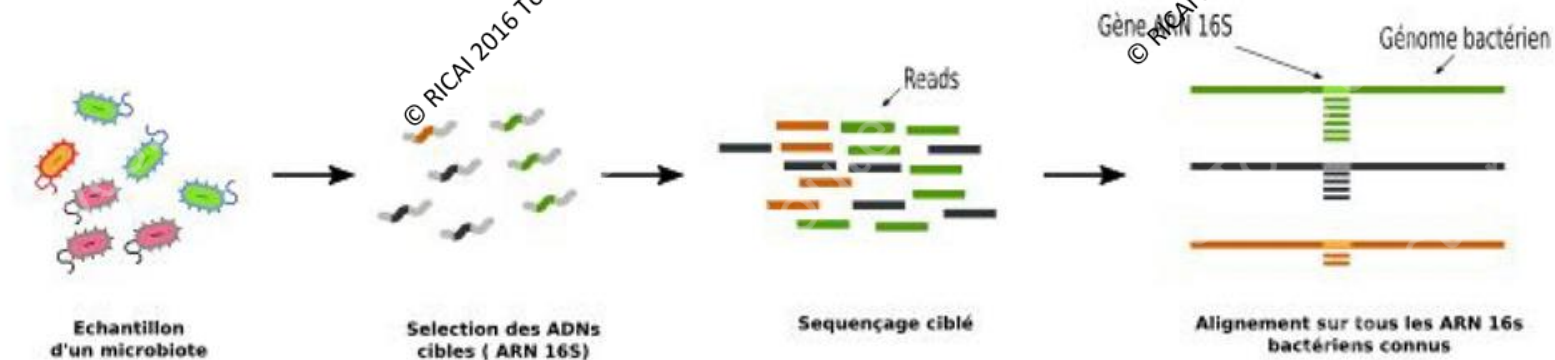


Métagénomique

- La métagénomique globale
Séquençage de l'ensemble des ADN

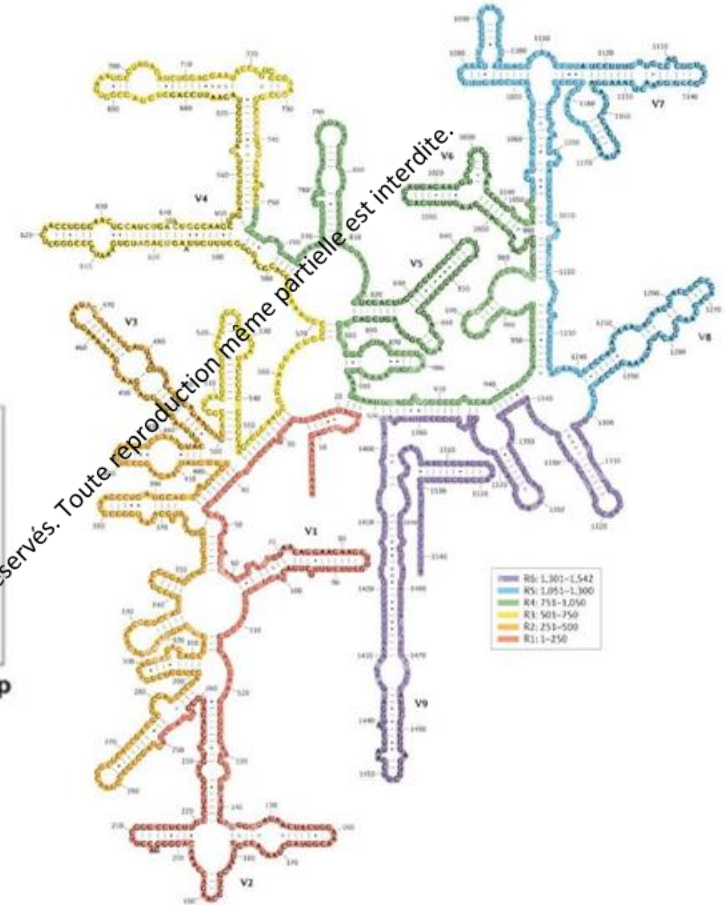
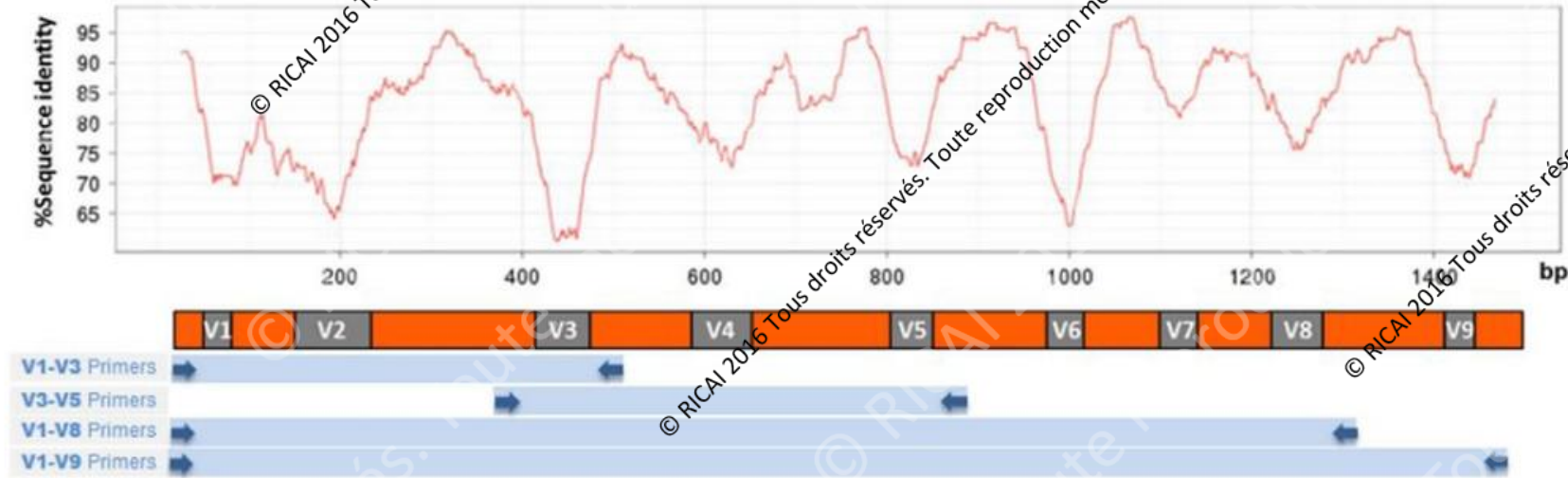


- La métagénomique ciblé
Séquençage d'une unique cible sur l'ADN



Métagénomique ciblée : ARN 16S

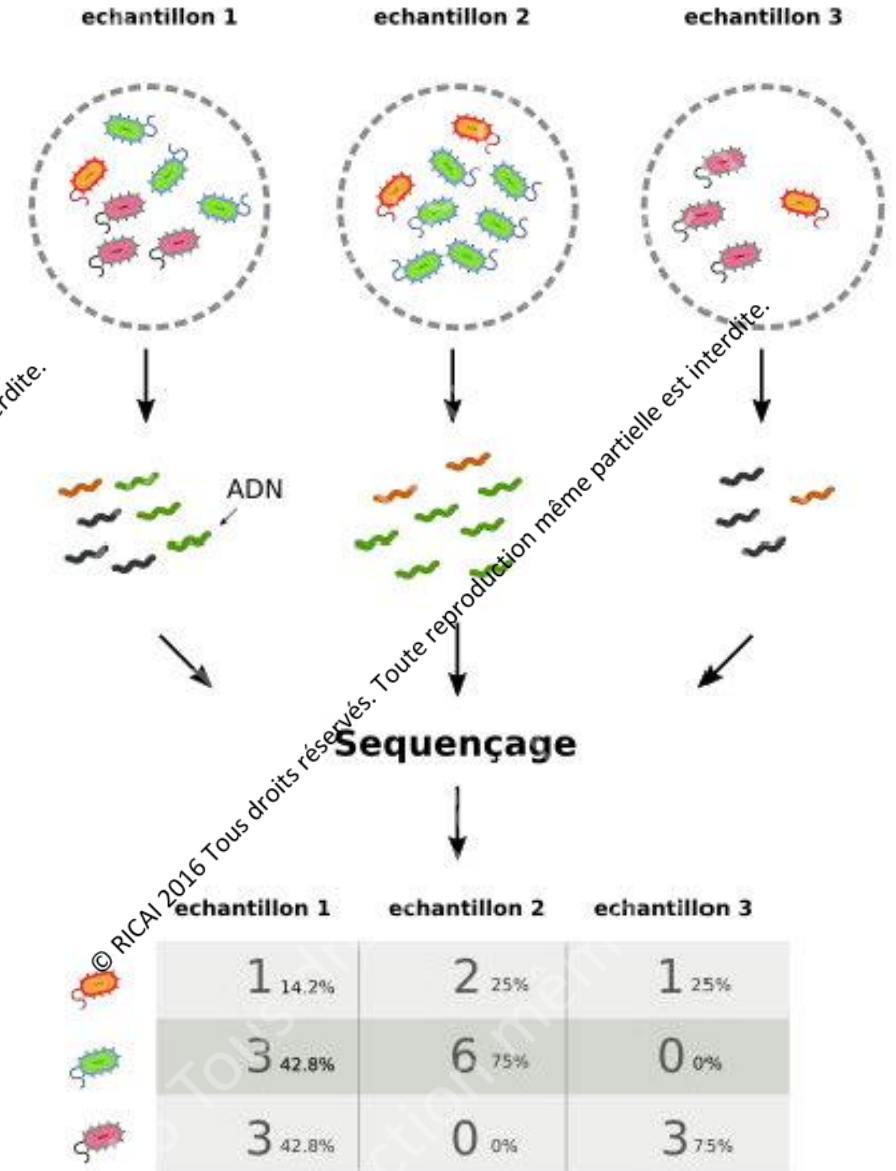
- 1500 nucléotides
- Régions variables: Spécificité de l'espèce
- Régions constantes: Cible des amorces



Analyse bioinformatique

Au final création de la **table des OTUs*** :

- **Echantillons x Taxons**
- Contient les **abondances** (nombre de séquences) par échantillon et par taxons
- Source des analyses multivariées

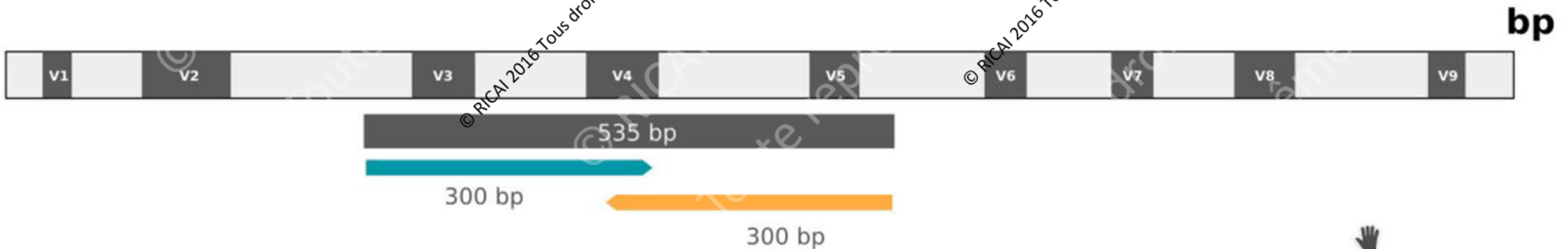
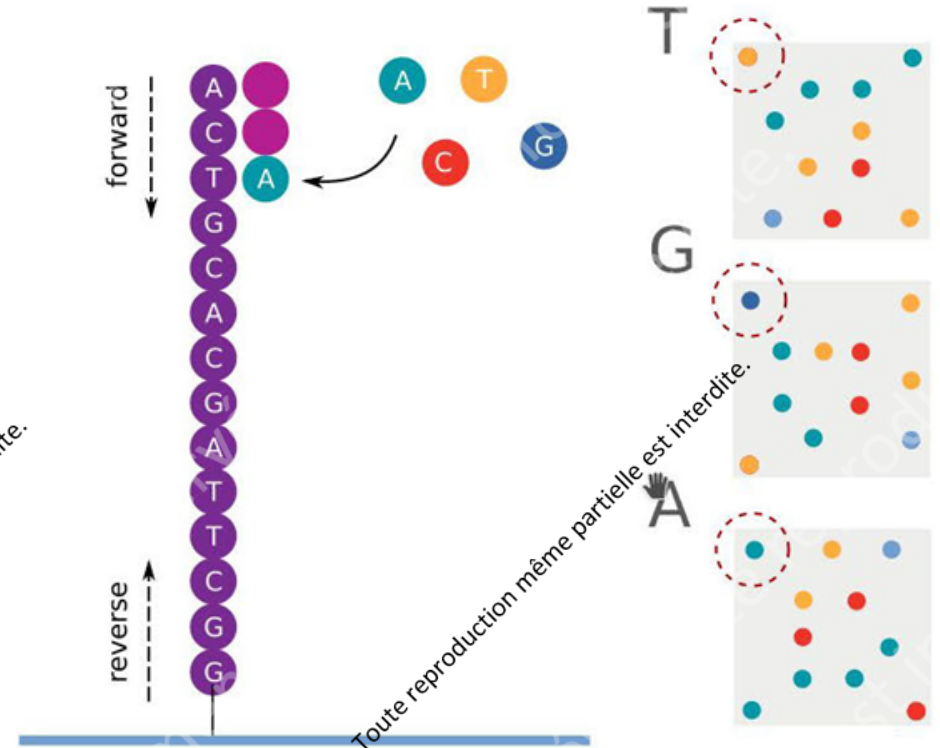


*Operational Taxon Unit

Table des OTUs

Séquençage

- **Séquenceur Illumina MiSeq**
 - Reads 2 x 300 pb
 - 25 millions de reads
 - 15 Gb
- **MiSeq Reagent Kit v3**
 - Amplification de la région V3-V5



Données du séquençage

Pour chaque patient

1 couple de fichiers **fastq** :

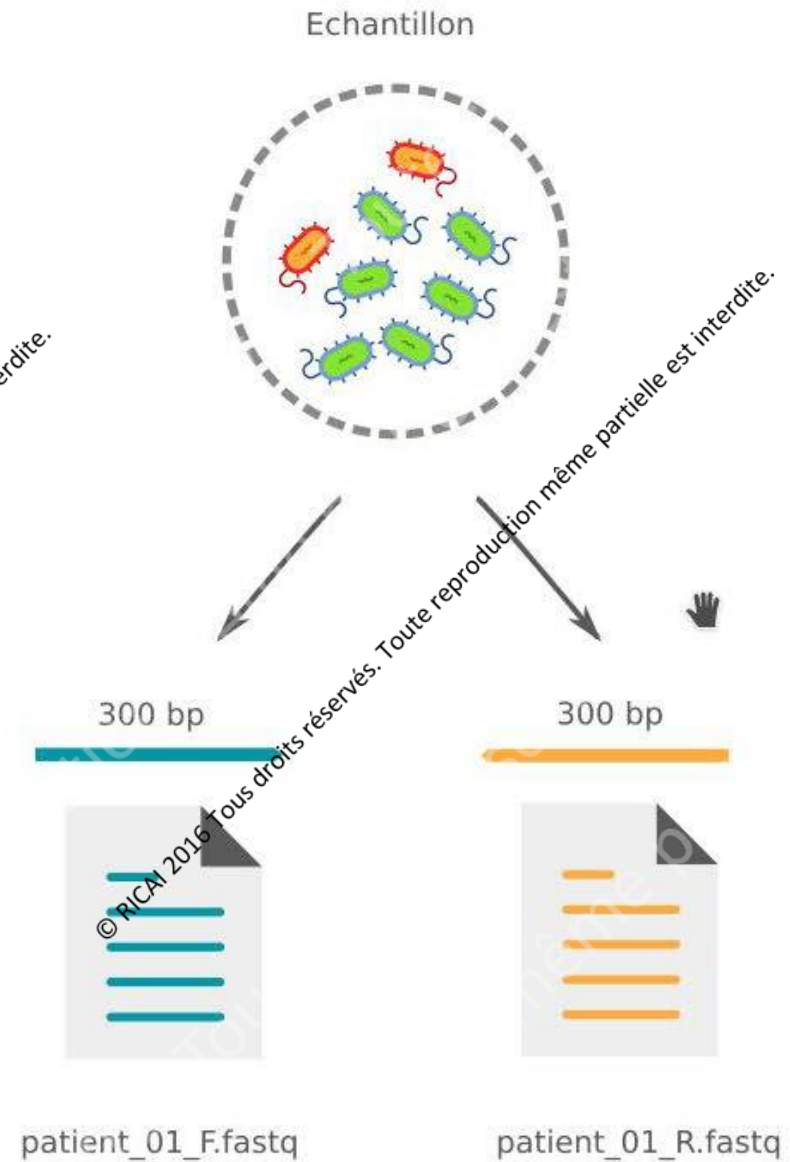
- Contient les séquences des reads
- Ainsi que la qualité du séquençage

```
@SEQ_ID
```

```
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCA
```

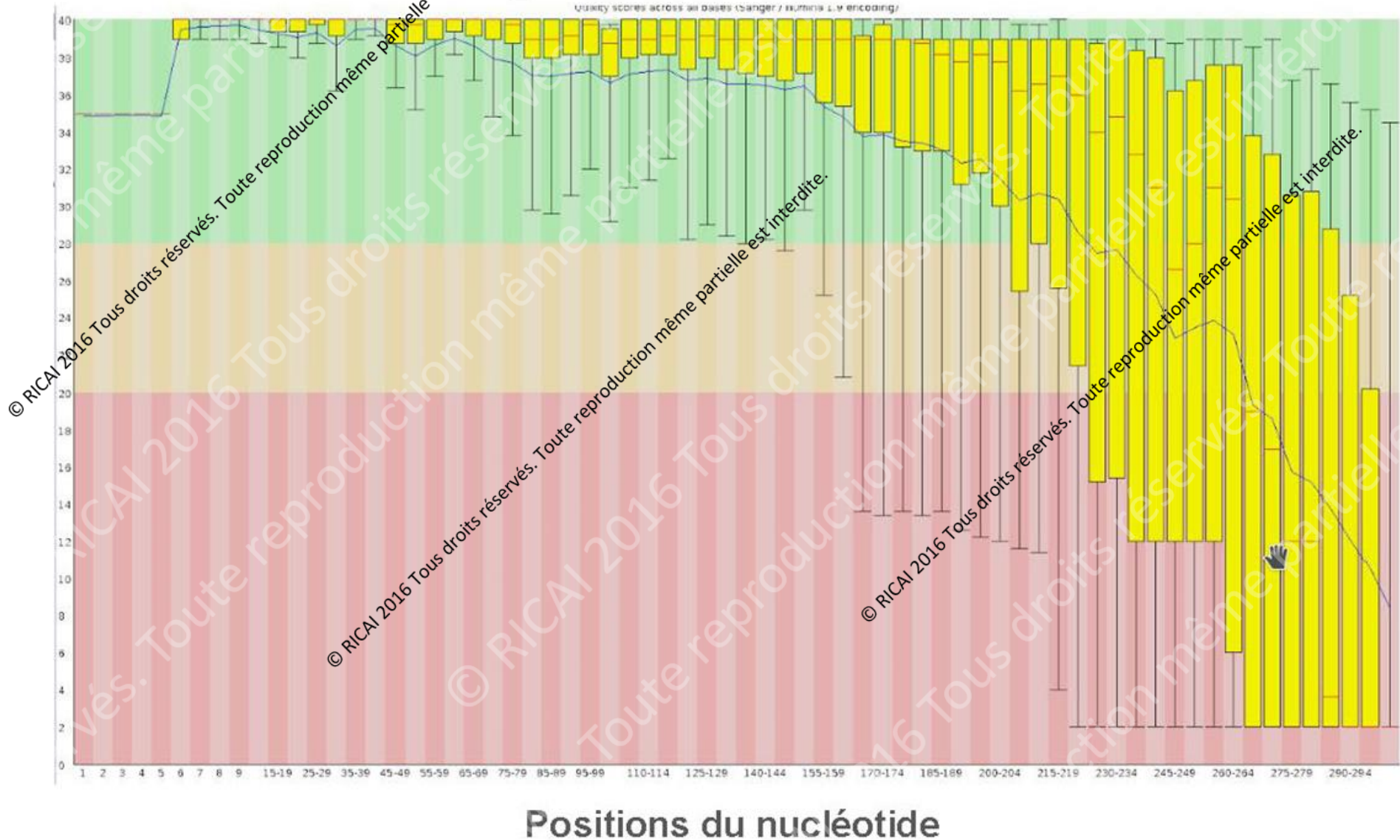
```
+
```

```
!"*((( (**+))%%%++)(%%%%)1**+*!))**55CCF>>>>>>
```



Qualité du séquençage

Qualité
(probabilité
d'erreur)

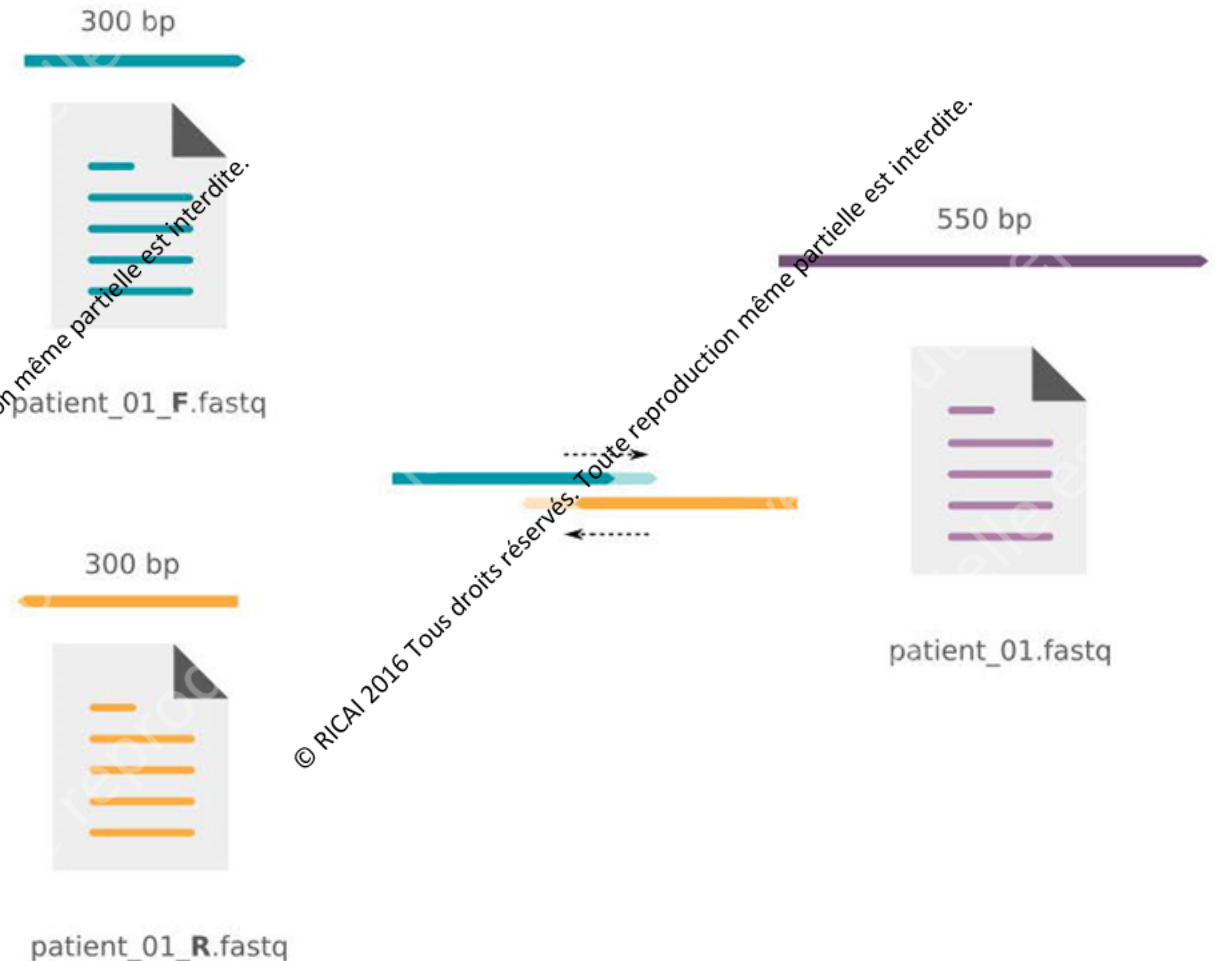


Fusion des paires de reads : *Merge pairs*

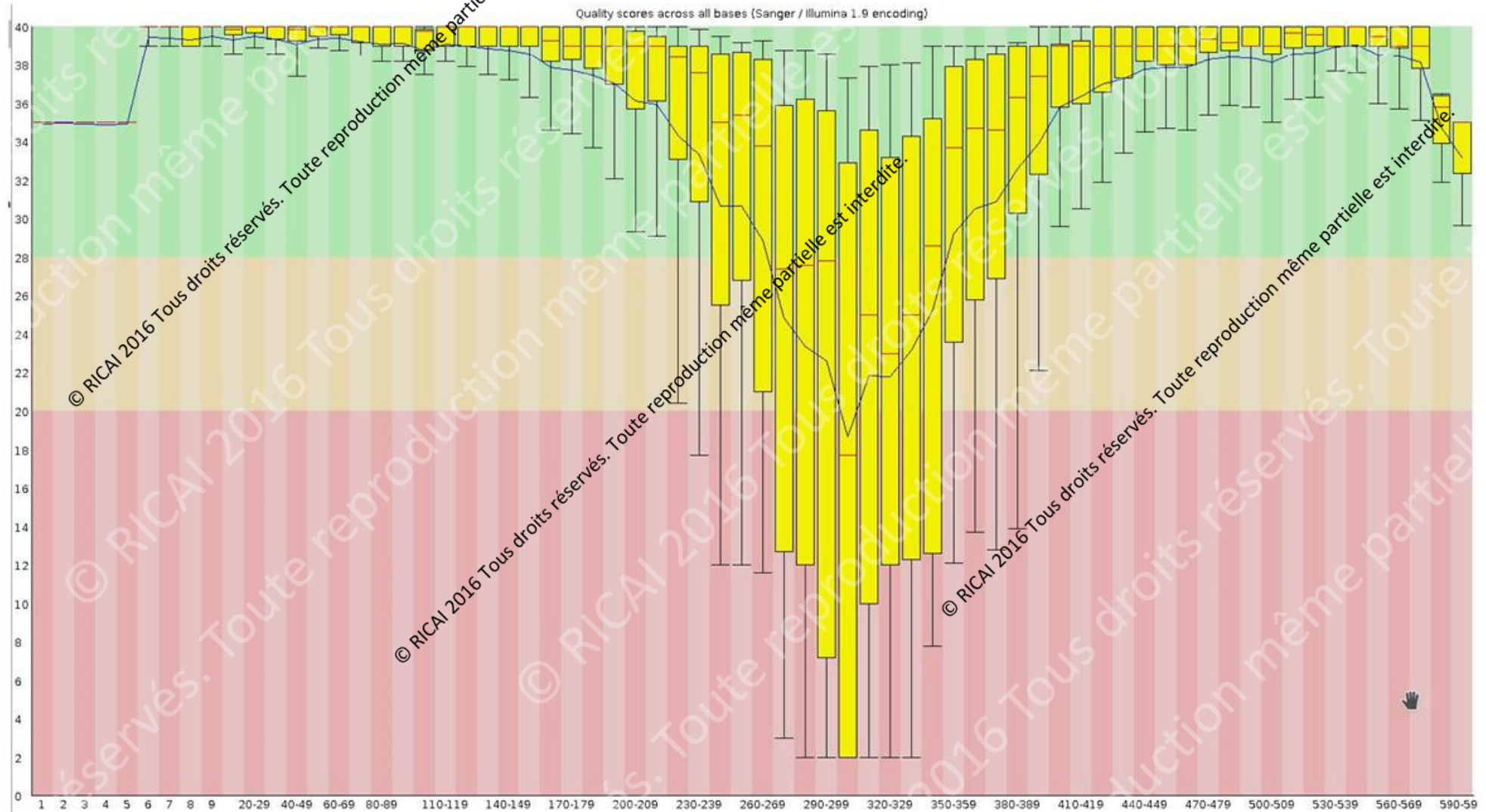
Recherche le chevauchement avec le minimum de différence.

Outils :

- **Vsearch** : ~30%
- **illumina-utils** : ~30%
- **Flash** : ~60%

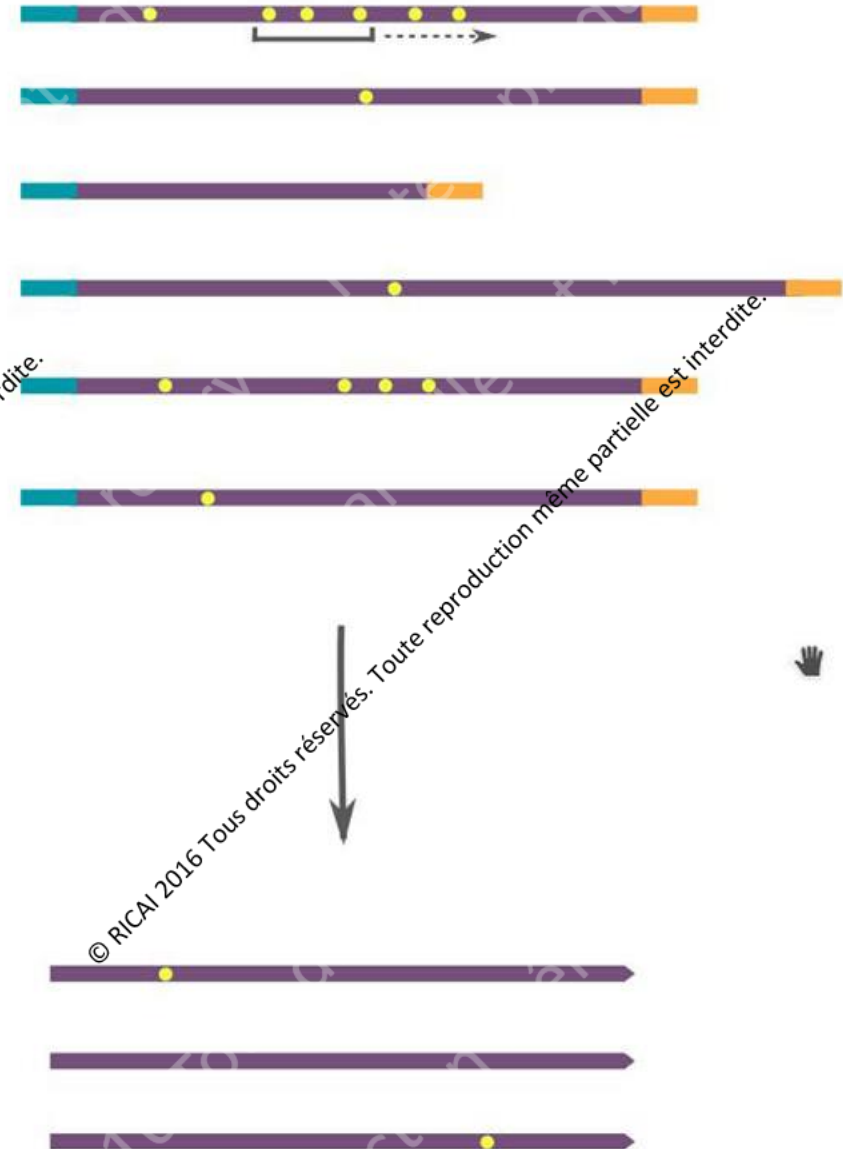


Fusion des paires de reads : *Merge pairs*

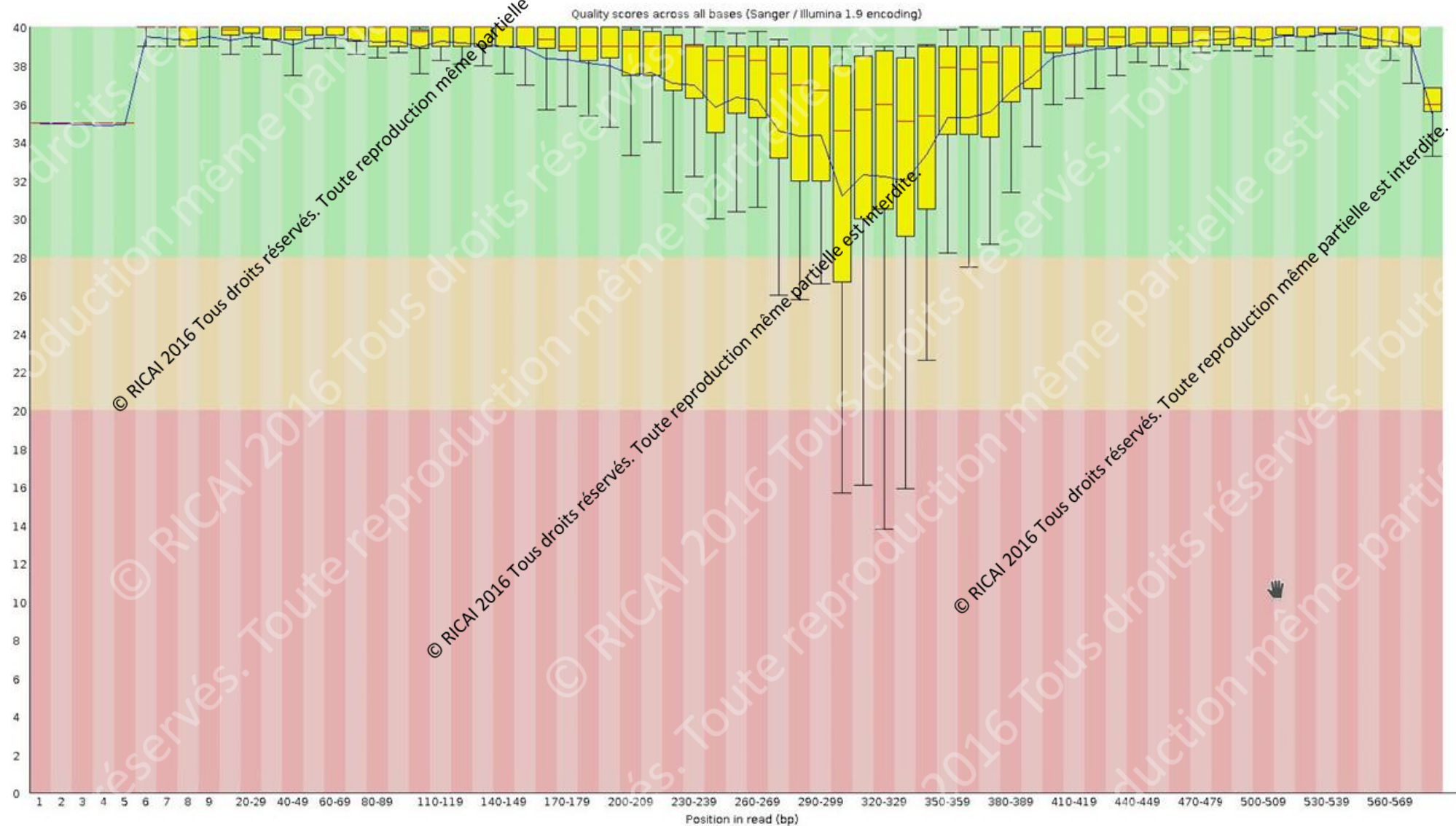


Filter des reads

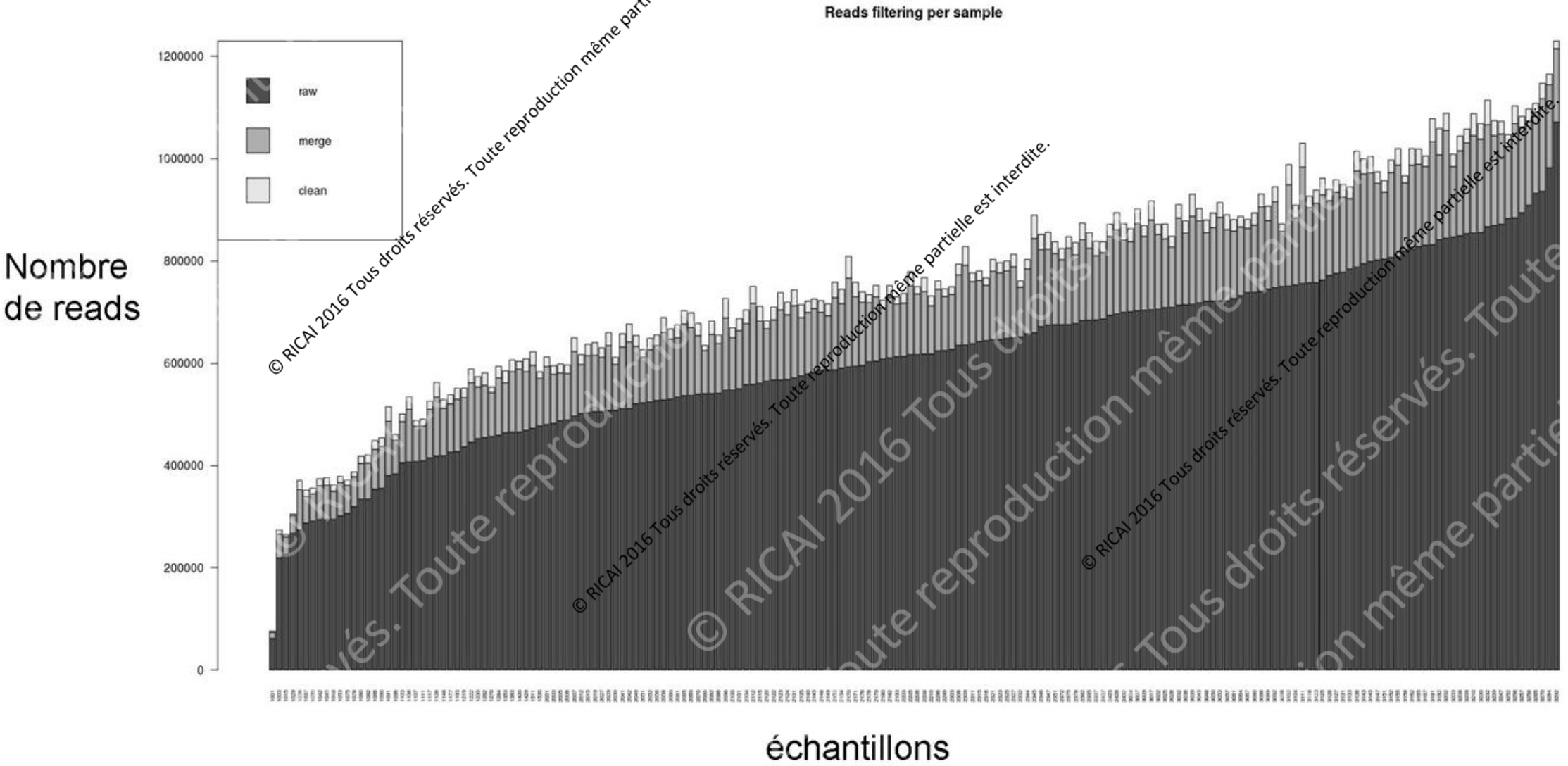
- Suppression des reads
 - Trop courts / Trop longs
 - Trop d'erreurs cumulées
- Découpe des reads
 - Suppression des amorces aux extrémités
- Outils
 - **Sickle**: fenêtre glissante de 20 pb. suppression si score moyen < 20



Filtere des reads



Pourcentage de reads analysables



Assignement taxonomique



• Assignement via une base de données : *Close reference picking*

- Les reads sont comparés à une **base de données** de séquences d'ARN 16s
- Seuil **97%**
- Alignement global
- Temps de **calcul** rapide
- Certains reads peuvent ne pas être classés
- **Greengene**, *Silva*, *RDP*



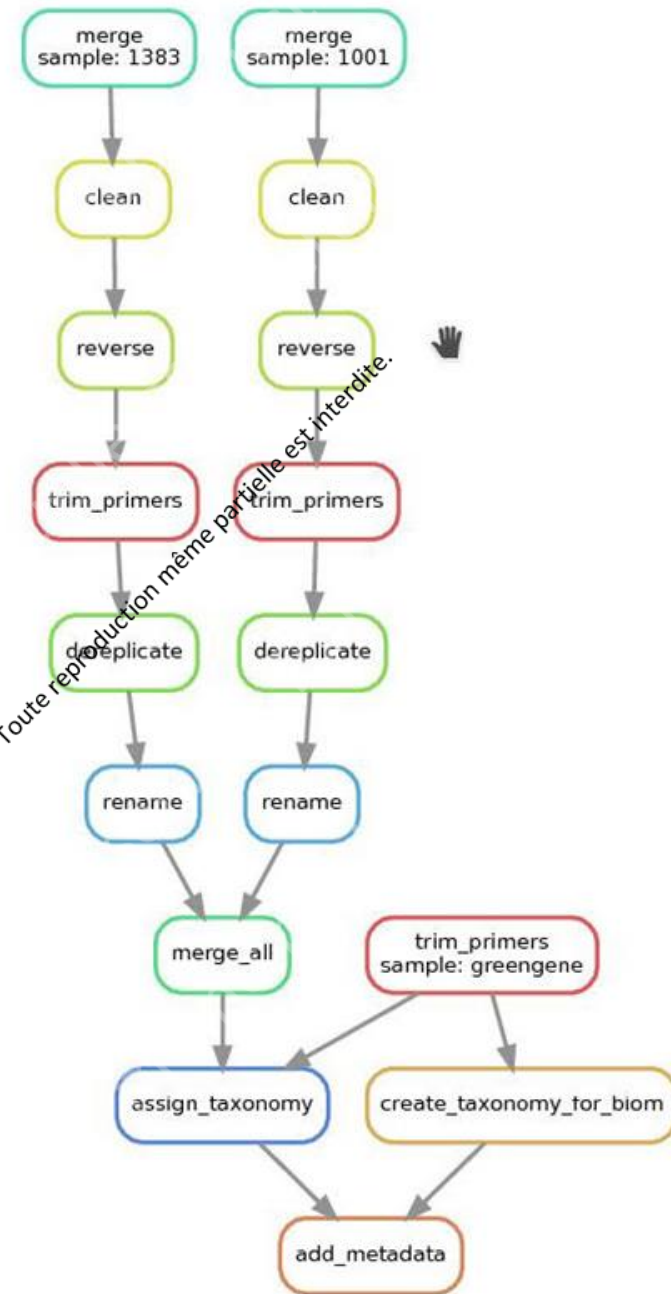
• Assignement *de novo* : *de novo picking*

- Regroupement des reads par similarité **Clustering**
- Seuil **97%**
- Temps de **calcul élevé**
- Tous les reads sont classés



Pipeline

- Mise en place et optimisation d'un pipeline original
- OpenSource : github.com/dridk/mucobiome
- Outils: *snakemake*, *bash*, *vsearch*, *sickle*, *cutadapt*, *seqtk*, *R*, *phyloseq*, *vegan*
- >25h de calcul sur 40 coeurs pour 188 échantillons
- Création finale d'un fichier **biom** contenant :
 - La **table des OTUs**
 - La **taxonomie** des OTU (*Domain, Kingdom, Phylum, Class...*)
 - Les **métadatas** associées aux échantillons (*Free, Never Age ...*)
- Statistique descriptive de la table des OTUs
 - Diversité alpha
 - Diversité bêta



© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

Résultats et Discussion

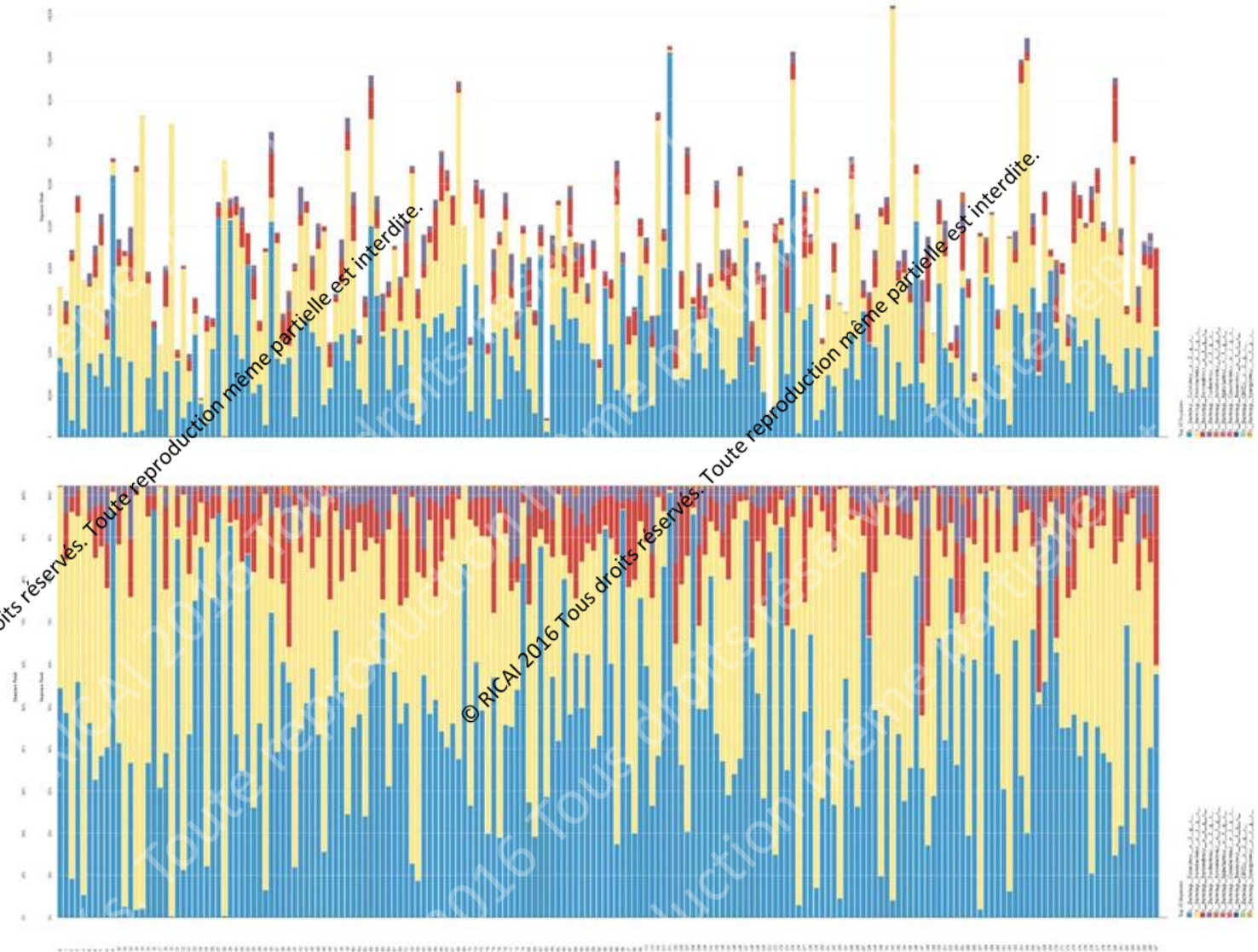
© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

Abondances

Phylum

- **Firmicutes**
 - ex: *Streptococcus*
- **Proteobacteria**
 - ex: *Haemophilus*
- **Bacteroidetes**
 - ex: *Prevotella*
- **Fusobacteria**
 - ex: *Fusobacterium*



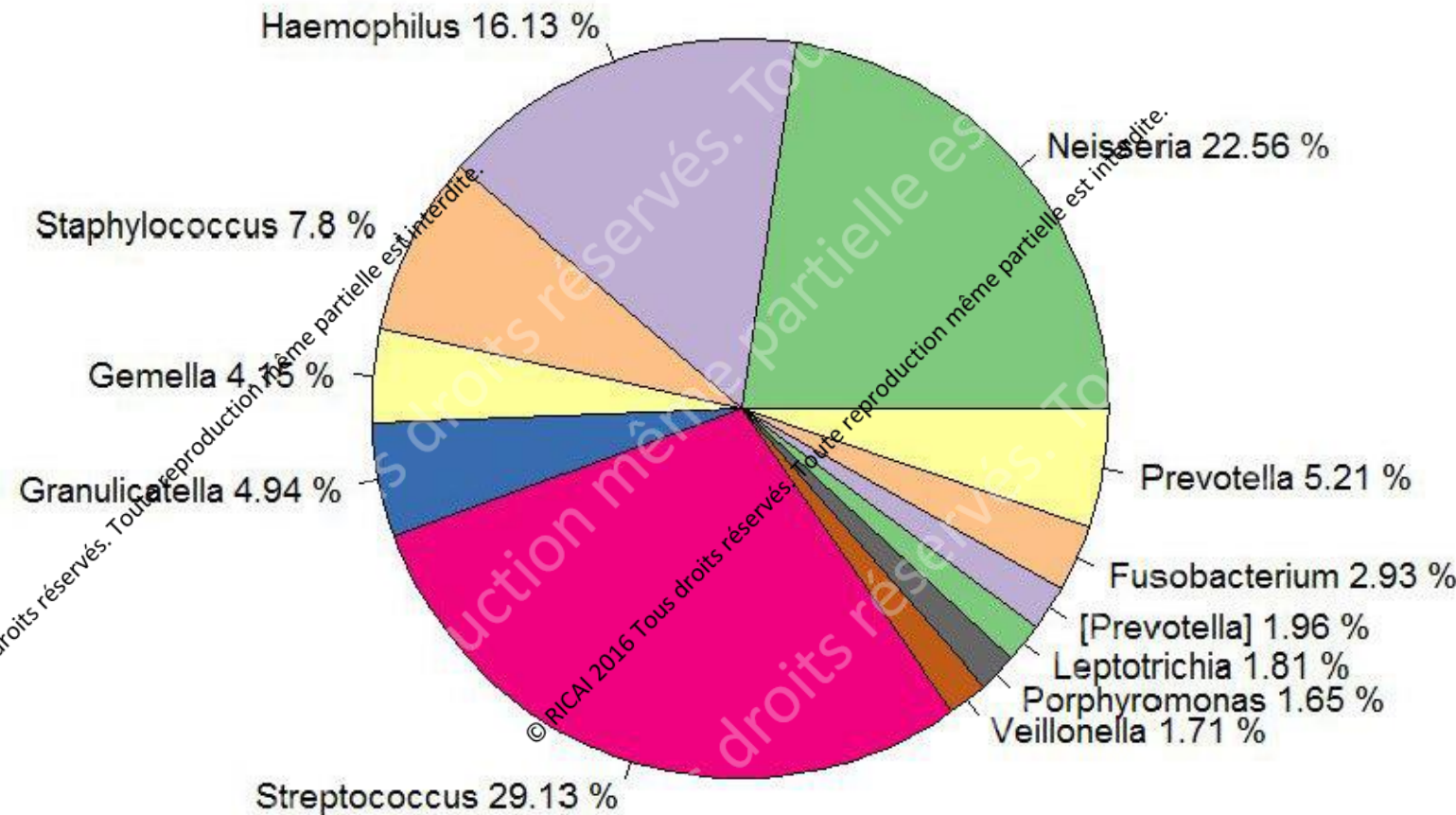
Core microbiota¹

Genres bactériens présent :

Avec une abondance > 1%
Dans > 50% des échantillons

*Un core anaérobie*² (15%) :

*Prevotella, Fusobacterium,
Leptotrichia, Porphyromonas*



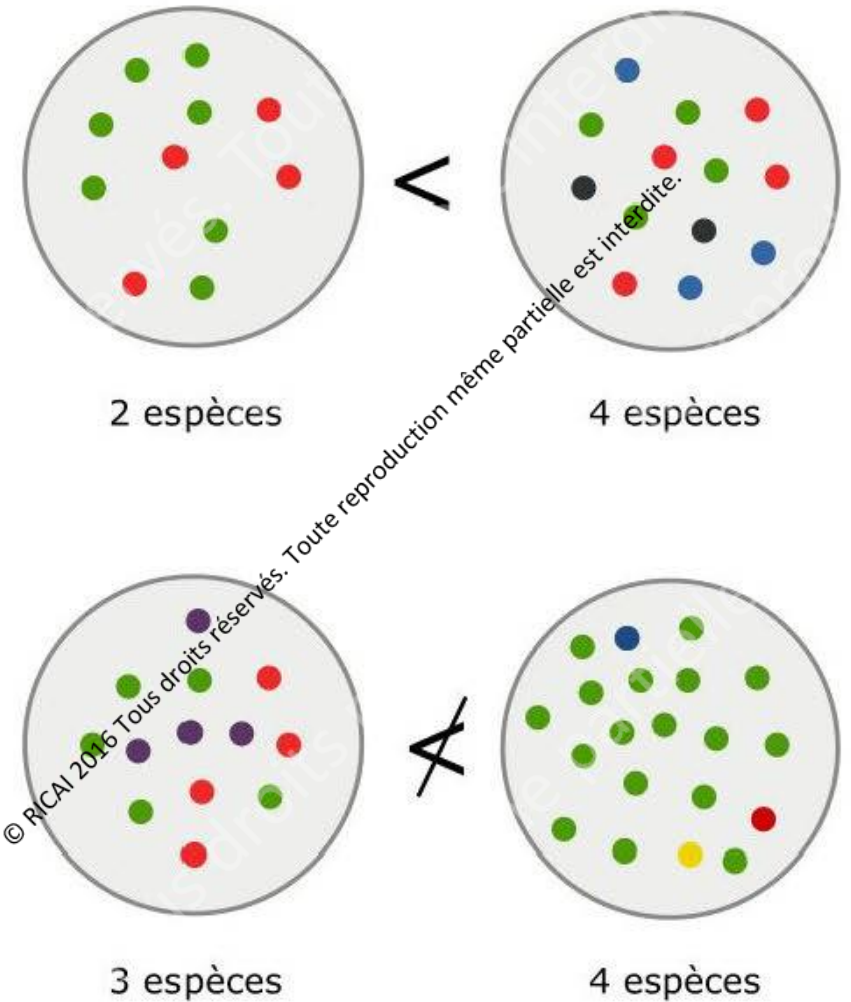
¹Lung microbiota across age and disease stage in cystic fibrosis
Scientific Reports |5:10241| DOI: 10.1038/srep10241

²Insights into the respiratory tract microbiota of patients with cystic
fibrosis during early *Pseudomonas aeruginosa* colonization. Springerplus.
2015 Aug 9;4:405. doi: 10.1186/s40064-015-1207-0.

Diversité Alpha

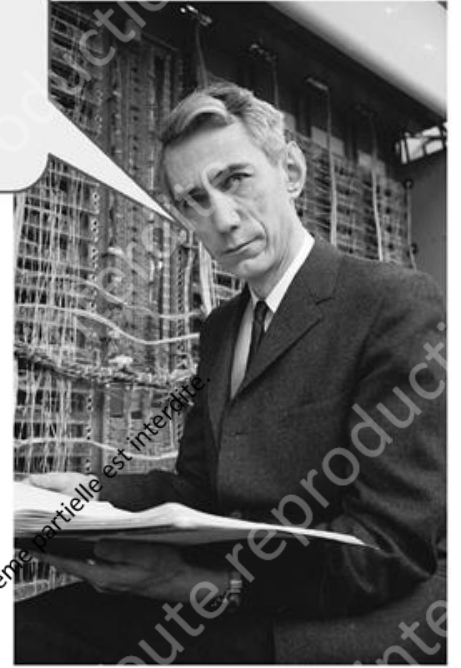
Mesure de la diversité dans **un** échantillon:

- Nombre d'espèces observées
- Indice de Chao1
- Indice de Simpson
- **Indice de Shannon**
- Indice de Fisher
- Indice ACE

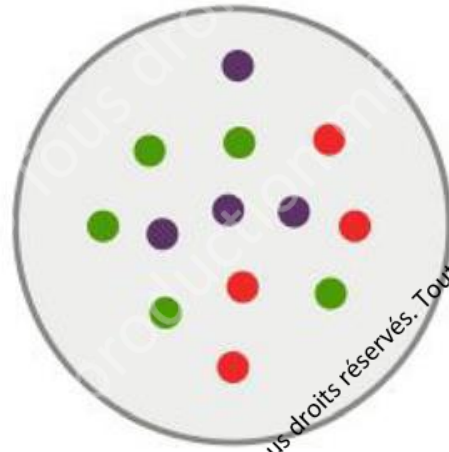


Indice de Shannon

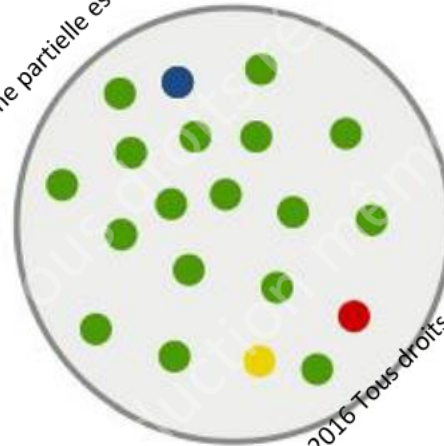
Votre téléphone portable, c'est grâce à moi !



$$H[p] = -\sum_{i=1}^k p_i \log p_i$$



indice de Shannon
1.09

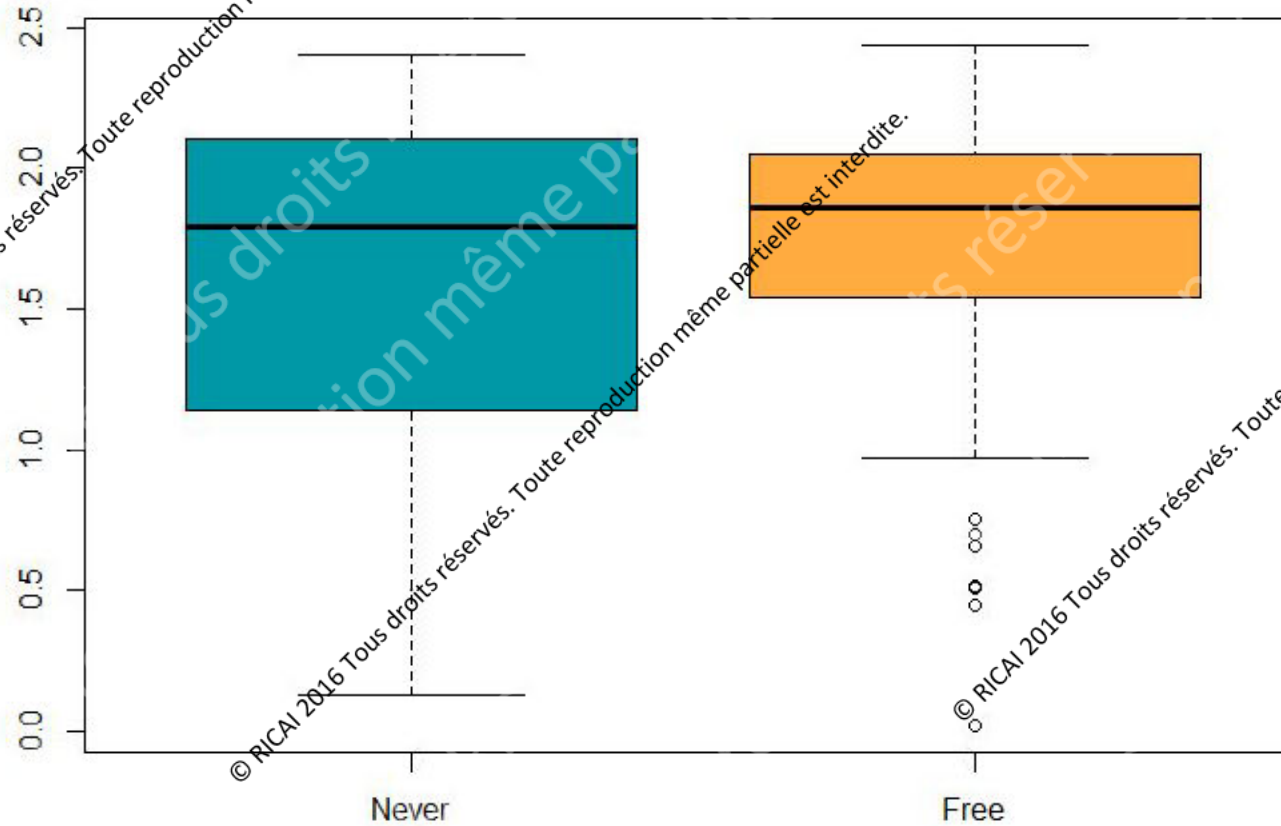


indice de Shannon
0.39

$$-\left(\frac{4}{13} \log\left(\frac{4}{13}\right) + \frac{4}{13} \log\left(\frac{4}{13}\right) + \frac{5}{13} \log\left(\frac{5}{13}\right)\right) = 1.09$$

Alpha diversité

Indice de Shannon des échantillons Never et Free

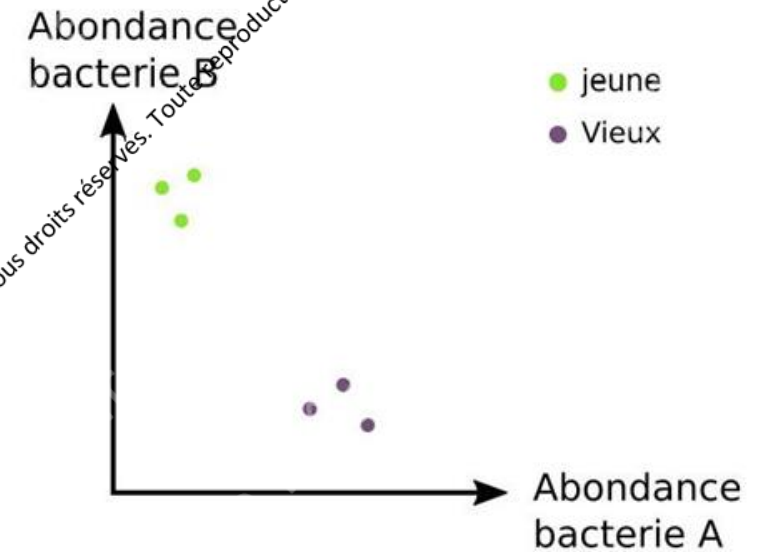
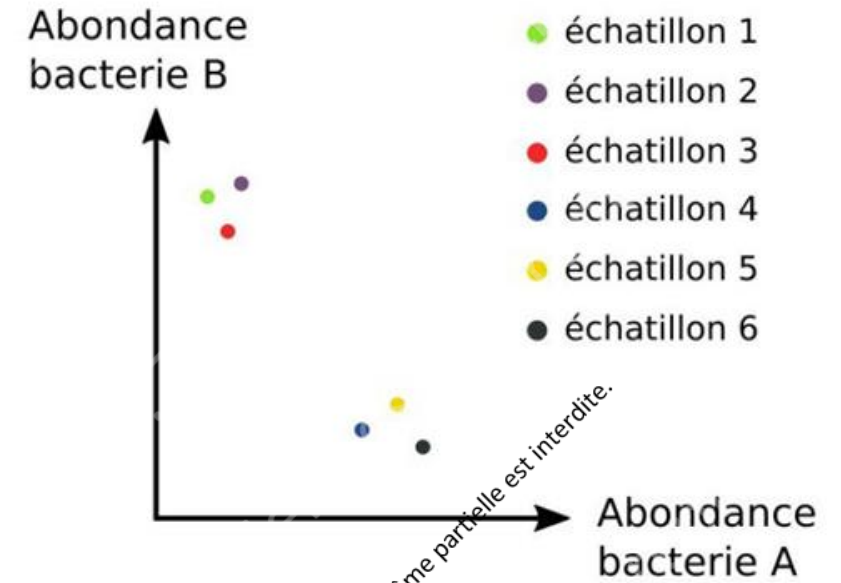


La variance des échantillons Never semble plus importante que les échantillons Free.

Diversité Bêta

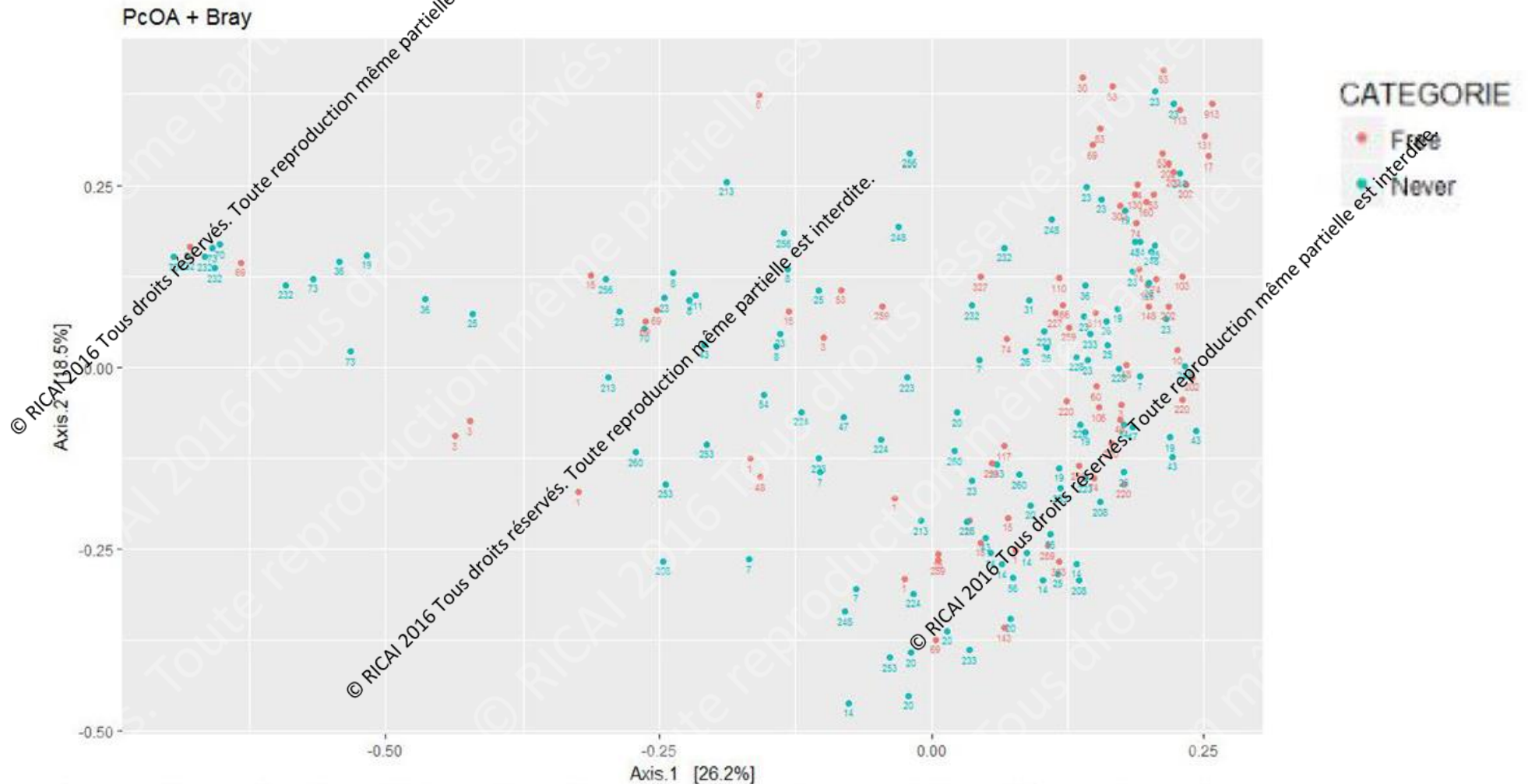
Mesure de diversité **entre plusieurs** échantillons

- Analyse multivariée :
 - Projections de **n** dimensions sur **2 axes** qui maximisent la **variabilité**
- Méthodes d'ordination :
 - Principal Component Analysis (PCA)
 - Principal Coordinate Analysis (PCoA)
 - Non-metric Multi Dimensional Scaling (NMDS)



Par catégorie d'âge

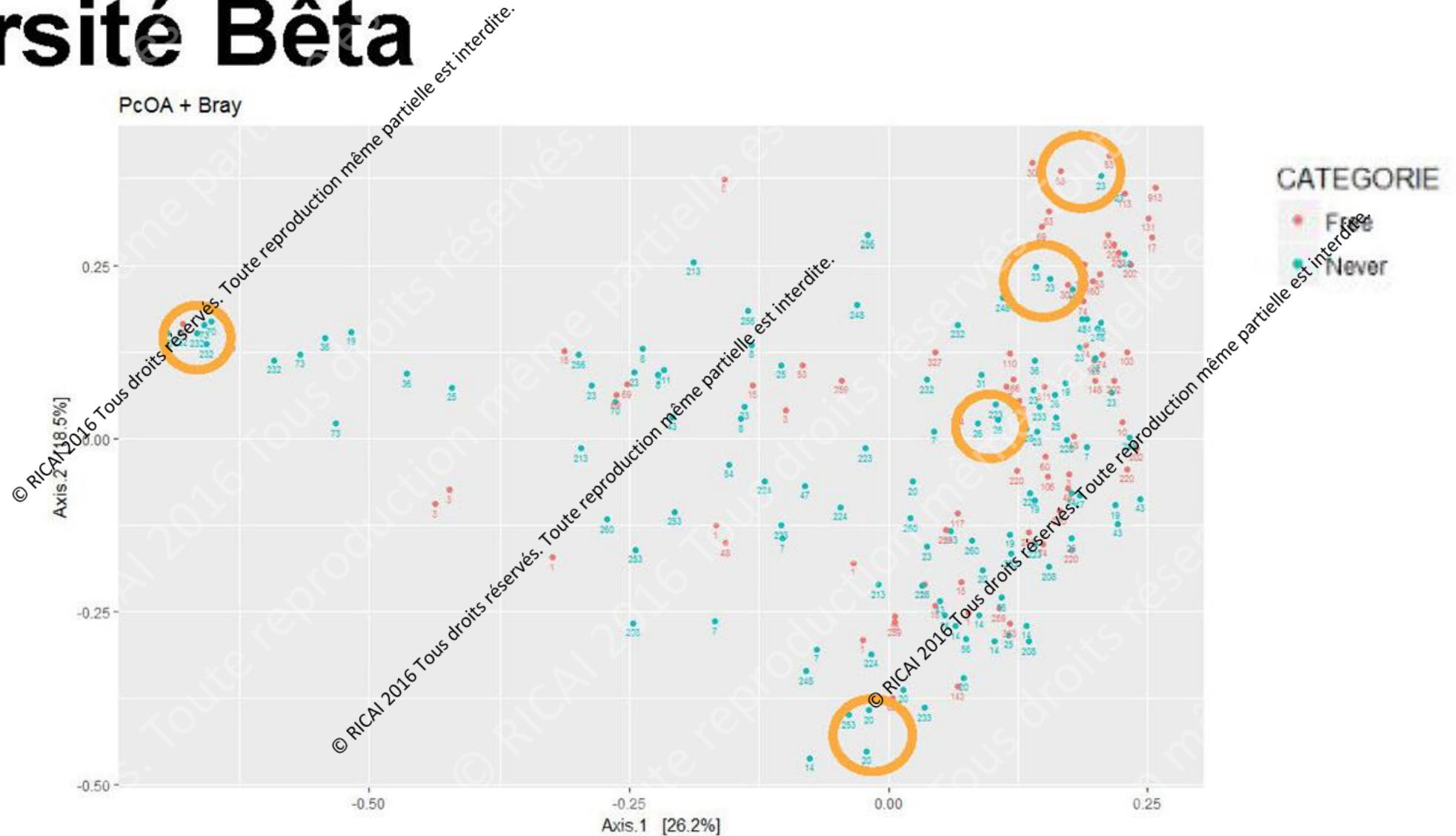
Diversité Bêta



bêta diversité des 188 échantillons labellisés par l'identifiant du patient.

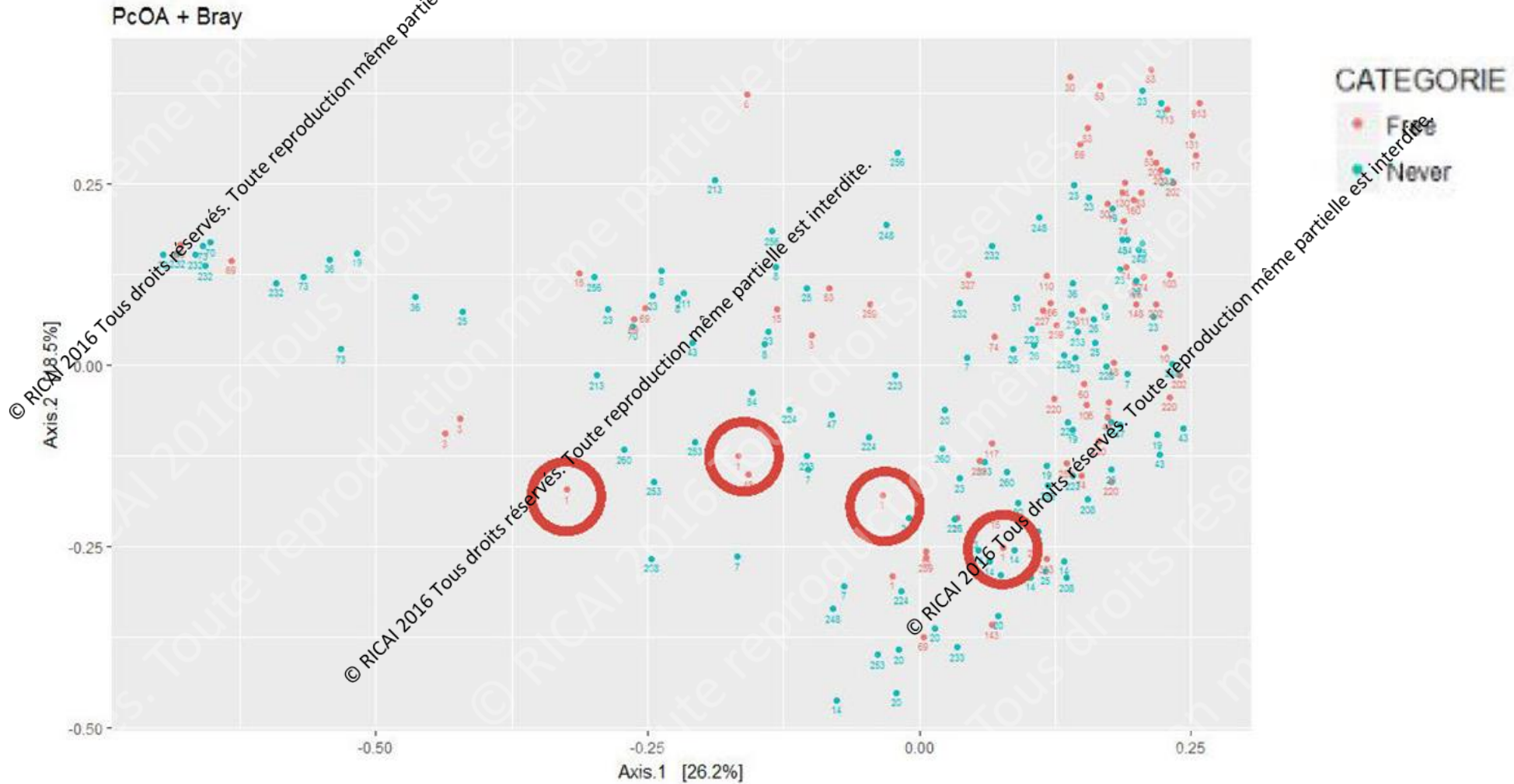
Les couleurs représentent les échantillons Free et Never

Diversité Bêta



Certains patients (les Never ?) = Un microbiote stable

Diversité Bêta



D'autres patients (Free ?) : microbiote instable

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

Conclusion

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

Amélioration du pipeline

- Prendre en compte les biais à la fois **biologique** et **bioinformatique**
- Analyser plus de reads. Seulement **10%** actuellement
- Cumuler d'**autres méthodes** d'analyse
 - **Phylogénie**
 - *Minimum entropy decomposition*
- Nécessite des ressources informatiques très importantes

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

Questions scientifiques auxquelles ce pipeline bioinformatique va nous permettre de répondre

- Impact de l'antibiothérapie anti-*Pseudomonas* sur le microbiote ?
- Il y a t'il des biomarqueurs pronostiques de l'infection à PA ?
- Il y a t'il des liens avec le génome de référence (cad de l'hôte ?)
 - Mutations impliquées dans la mucoviscidose
 - Relation avec le système immunitaire (CMH)

Le séquençage de 4ème génération



Bibliographie

Papier

Lung microbiota across age and disease stage in cystic fibrosis Scientific Reports [5:10241] **10.1038/srep10241**

Getting Started with Microbiome Analysis: Sample Acquisition to Bioinformatics **10.1002/0471142905.hg1808s82.Getting**

FLASH: Fast length adjustment of short reads to improve genome assemblies **10.1093/bioinformatics/btr507**

PEAR: A fast and accurate Illumina Paired-End reAd mergeR **10.1093/bioinformatics/btt593**

16S classifier: A tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets

10.1371/journal.pone.0196106

Insights into the respiratory tract microbiota of patients with cystic fibrosis during early *Pseudomonas aeruginosa* colonization **10.1186/s40064-015-1207-0**

Site Web

<https://github.com/torognes/vsearch>

<https://github.com/najoshi/sickle>

<https://ccb.jhu.edu/software/FLASH/>

<http://greengenes.lbl.gov/cgi-bin/nph-index.cgi>

Remerciements

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

Autres méthodes

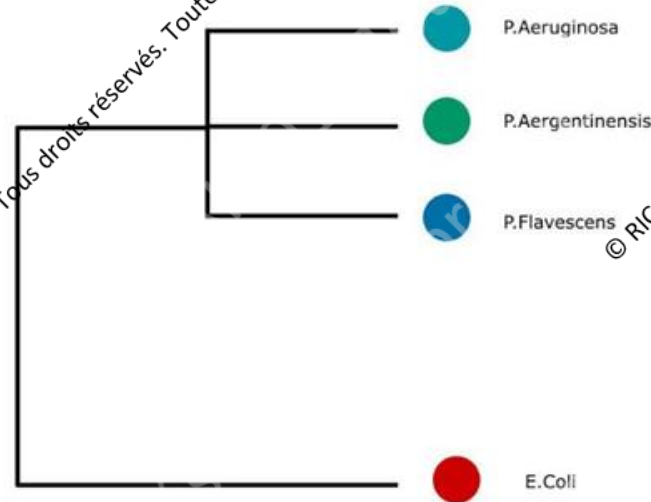
Prendre en compte la **phylogénie** des espèces

- **Echantillon A**

- *Pseudomonas aeruginosa*
- *Pseudomonas aeruginensis*
- *Pseudomonas flavescens*

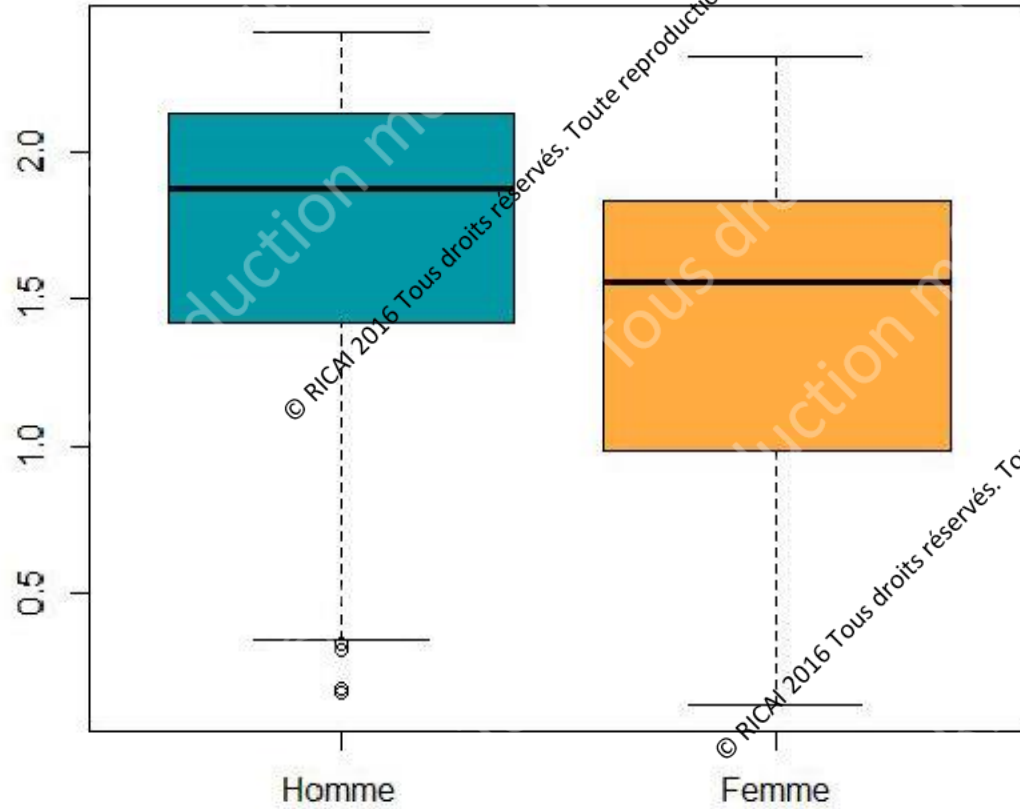
- **Echantillon B**

- *Pseudomonas aeruginosa*
- *Pseudomonas aeruginensis*
- *Escherichia coli*

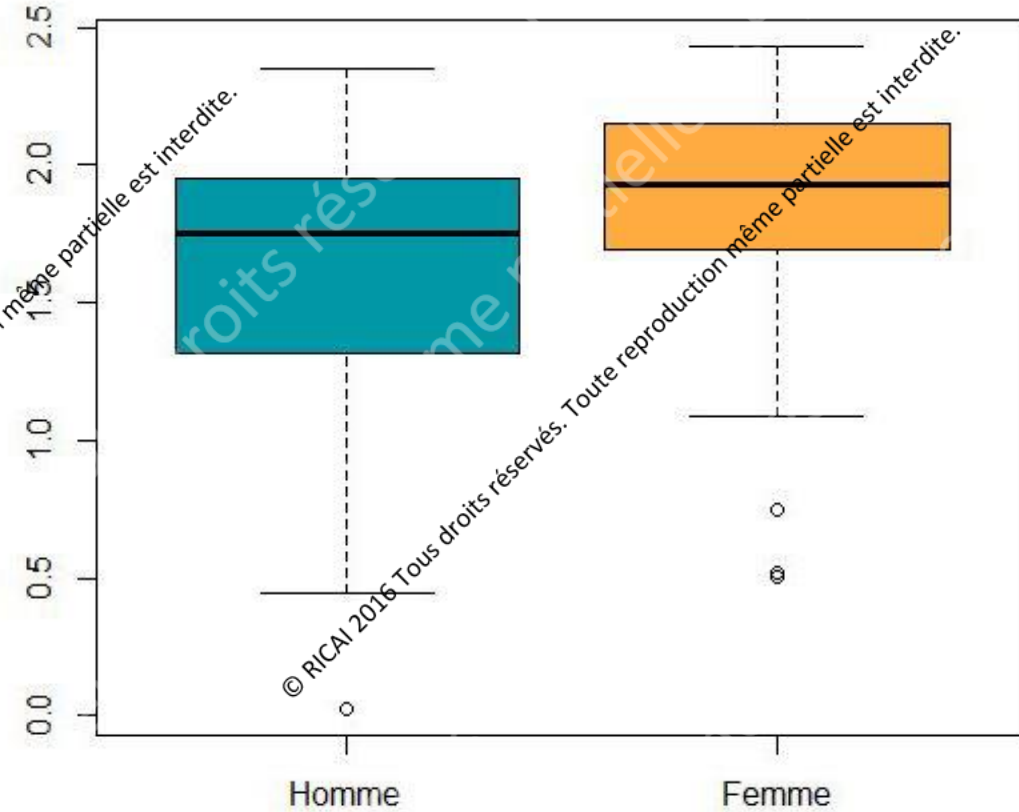


Gender Gap

Indice de Shannon des échantillons Never

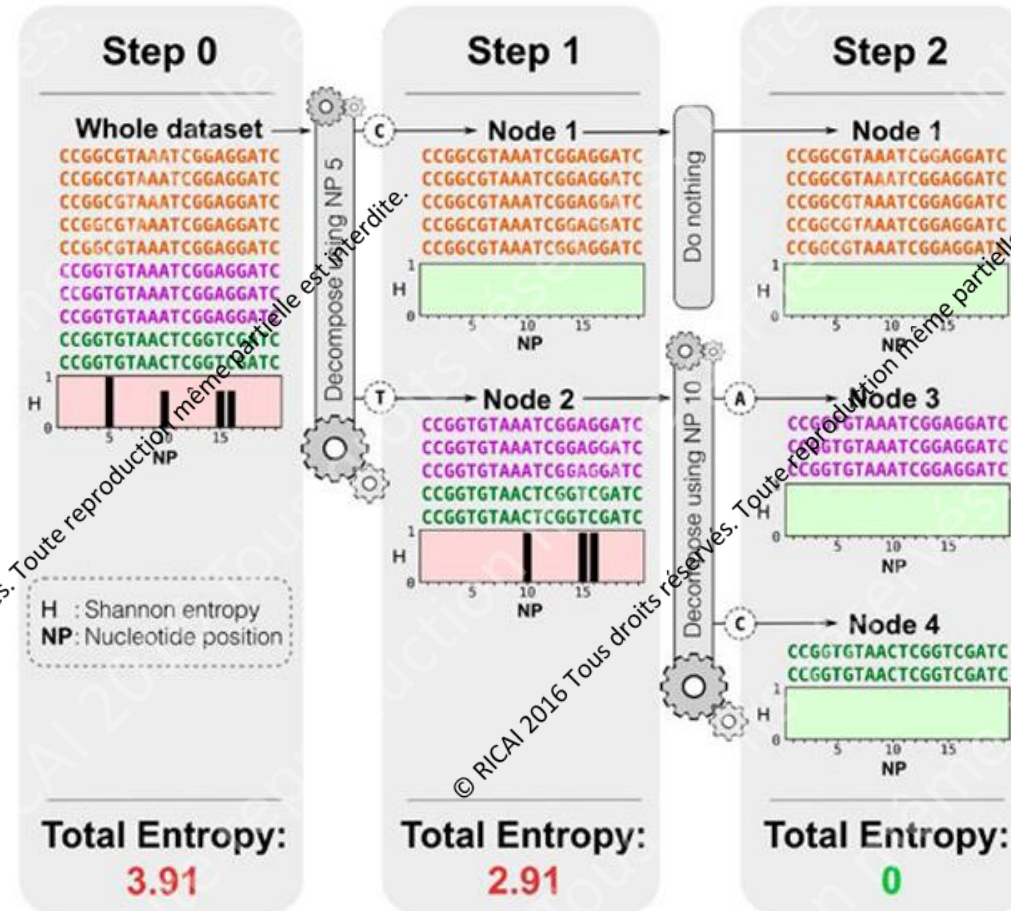


Indice de Shannon des échantillons Free

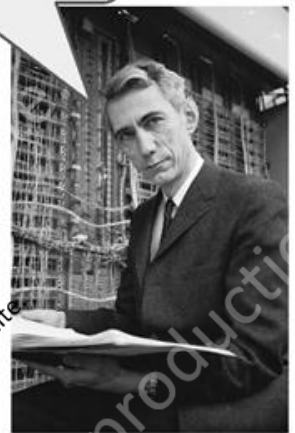


Autres méthodes

Décomposition par entropie minimum : Évite d'utiliser le seuil de 97%



encore moi !



Métagénomique

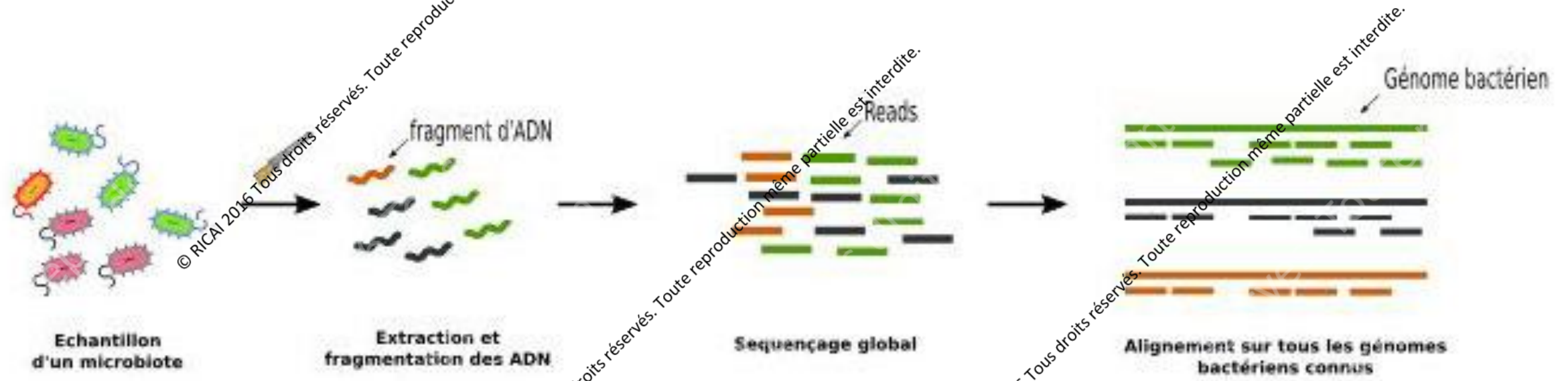
- Métagénomique globale

- Données massives
- Non-sélectif (Virus, humain ...)
- Algorithmes d'alignements complexes
- Vision de l'ensemble des gènes
- Taxonomie plus précise

- Métagénomique ciblé

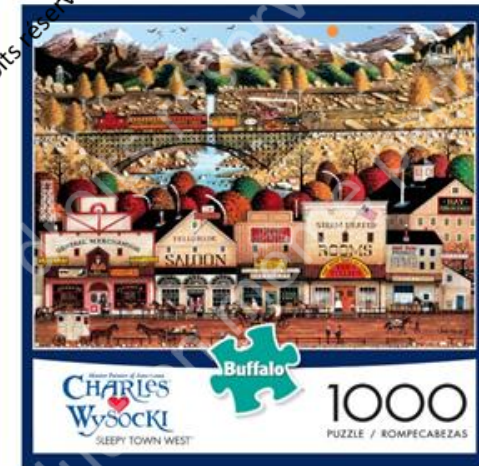
- Données plus faibles
- Sélectif (16S unique aux procaryotes / archéobactérie)
- Algorithmes plus simples
- Inférence fonctionnelle non réalisable

Métagénomique globale : *Shotgun sequencing*



L'ensemble des ADNs présents dans le milieu est séquencé

Métagénomique globale : *Shotgun sequencing*



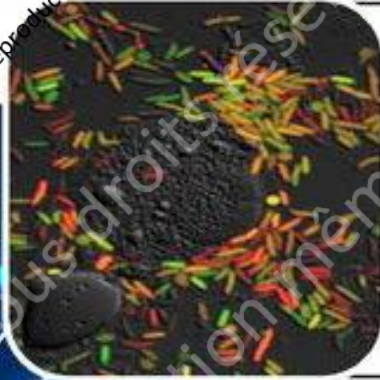
© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

© RICAI 2016 Tous droits réservés. Toute reproduction même partielle est interdite.

Objectif

Hypothèse :



Colonisation à *P. aeruginosa*

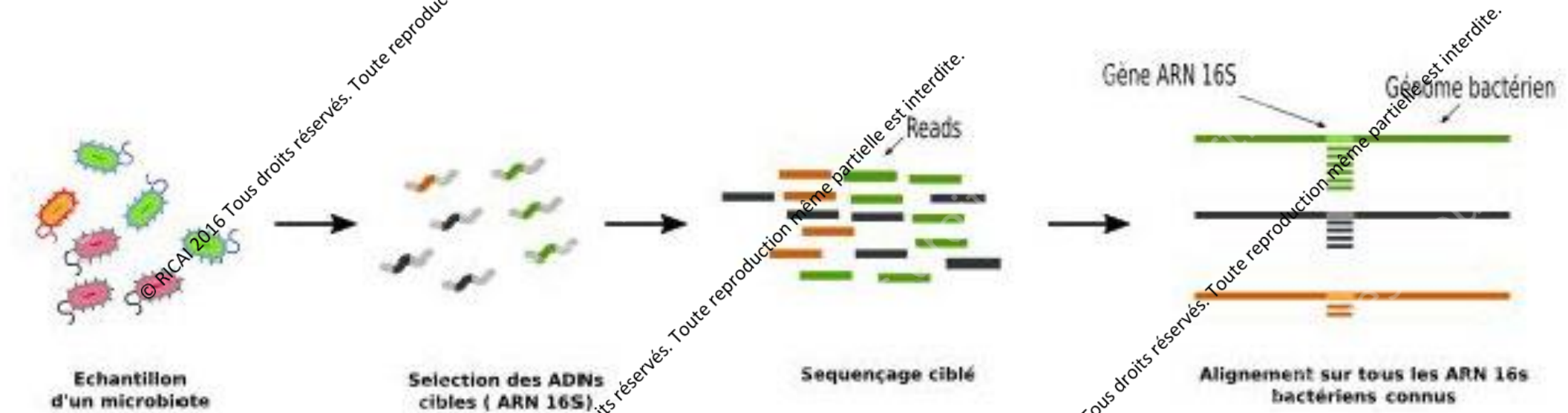


MODULATION

Microbiote pulmonaire endogène

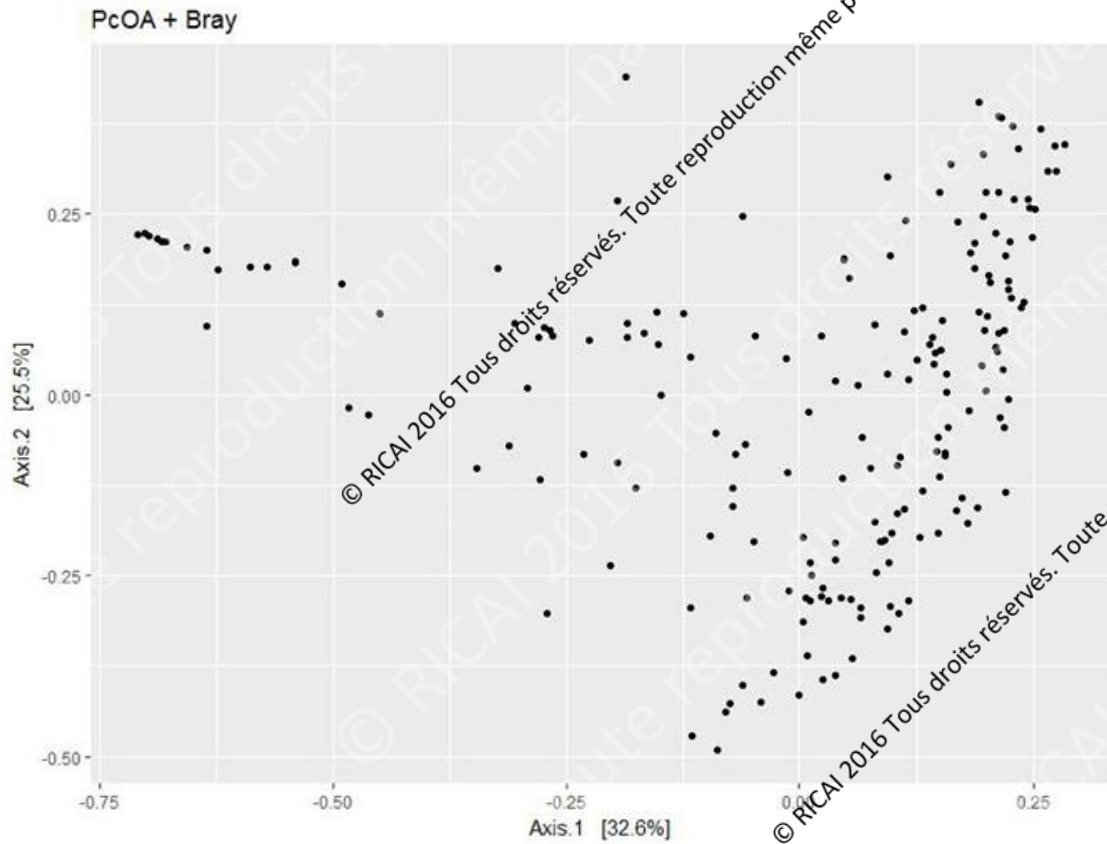
Identifier les déterminants microbiens de la primocolonisation à *P. aeruginosa*

Métagénomique ciblée : *ARN 16S*

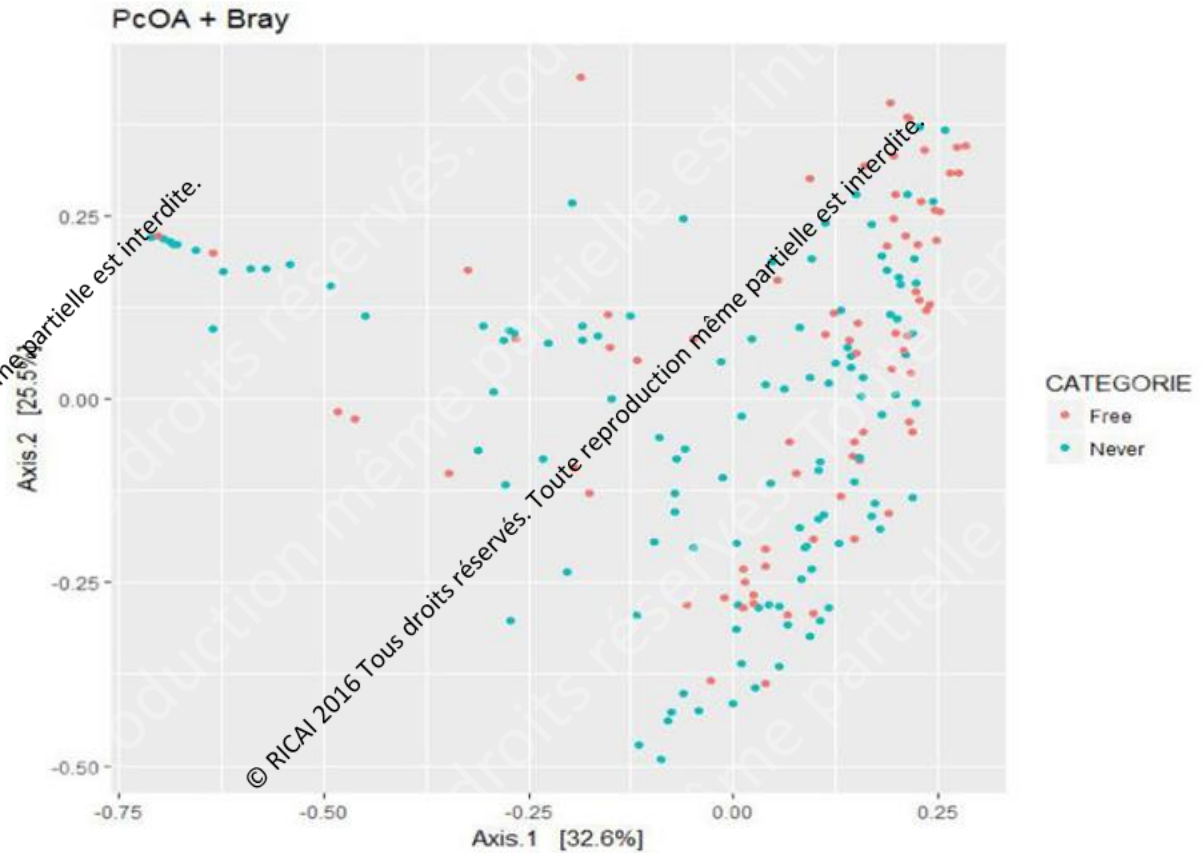


Seul la région cible est séquencée

Diversité Bêta



beta diversité des 188 échantillons



beta diversité des 188 échantillons
et leurs catégories Free ou Never