milou : un package R pour construire des profils taxonomiques issus de métabarcoding multi-marqueurs

Journées métagenomiques – PEPI IBIS

Benoit Goutorbe

9 novembre 2022

MaIAGE

CRCM
Centre de Recherche en Cancérologie de Marseille

Alphabio
LABORATOIRE DE BIOLOGIE MEDICALE

Métagénomique, métabarcoding et besoin d'alternatives

Métabarcoding :

- ✓ Profils taxonomiques
- ✓ **Bon marché** et facile à mettre en œuvre
- ✓ Nombreux outils et bases de données
- × **Résolution** limitée
- × Couverture taxonomique limitée
- × Biais (amplification et nombre de copies)

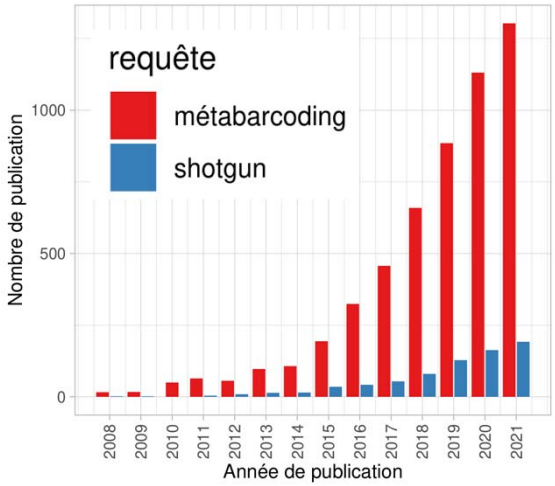
Métagénomique *shotgun* :

- ✓ Meilleure résolution taxonomique
- ✓ Caractérisation fonctionnelle
- ✓ Reconstruction de MAGs
- × **Coût** de séquençage élevé
- × Données massives, difficiles à traiter
- × Contamination par l'**ADN de l'hôte**

→ Besoin d'analyser un **grand nombre d'échantillons** (haute dimensionalité, grande variabilité inter-individuelle, écosystèmes dynamiques)

→ De nombreux projets ont recourt au métabarcoding malgré ses biais et limitations

→ Recherche d'**alternatives**



Métabarcoding : choix des marqueurs & limites associées

Caractéristiques attendues du marqueur	Biais & limitations associés
• Présent et amplifié chez tous les organismes • Une seule copie dans chaque génome • Suffisamment diversifié pour le pouvoir discriminant • Base de données de référence exhaustive	• Limité à un domaine du vivant **(bactéries)** • Biais d'universalité des amorces • Biais du nombre de copies • Faible résolution taxonomique

→ Chez les bactéries : l'immense majorité d'études utilise le gène codant pour l'**ARNr16S**

→ Des **alternatives** existent : gyrB, cpn60, ITS_{bact} , rpoB, pheS, ...

→ Les différents marqueurs apportent des informations **différentes** mais **complémentaires**

Métabarcoding multi-marqueurs

→ **Concept** : utiliser plusieurs marqueurs en parallèle

→ **Hypothèse** : on va pouvoir tirer parti des avantages de chaque marqueur (universalité, pouvoir discriminant, biais du nombre de copies) pour produire de meilleurs **profils taxonomiques**

→ **Avantages** :

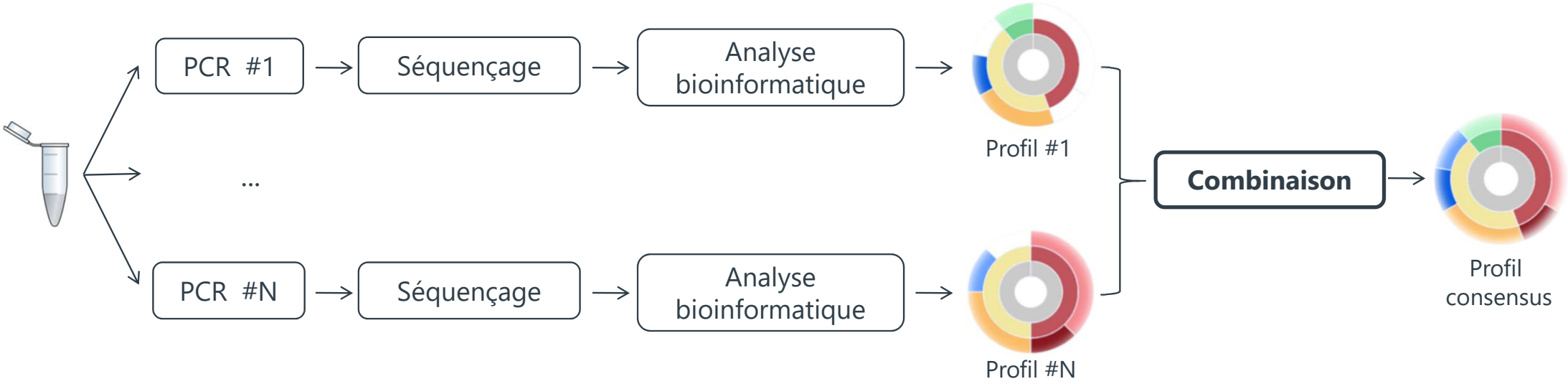
- Facilité d'utilisation
- Faible surcoût
- Meilleure confiance dans les résultats
- Possibilité d'analyse en deux temps

→ **Outils existants** :

- Combinaison différentes régions du gène 16S (MVRSION, SMURF, sidle)
- Da Silva *et al. 2018* : approche uniquement qualitative
- Stefanni *et al.* 2018 : approche semi-automatisée, pas généralisable

→ Pas de solution satisfaisante pour construire un profil taxonomique consensus

Métabarcoding multi-marqueurs : notre approche



Flexibilité :

- Choix des **marqueurs**
- Choix des **outils** et **bases de données**
- Tous types d'**écosystèmes**

Challenge :

➞ Construire un profil taxonomique consensus qui tire parti des avantages de chaque marqueur

Cas du microbiote vaginal

Composition : 5 *community state types* (CSTs)

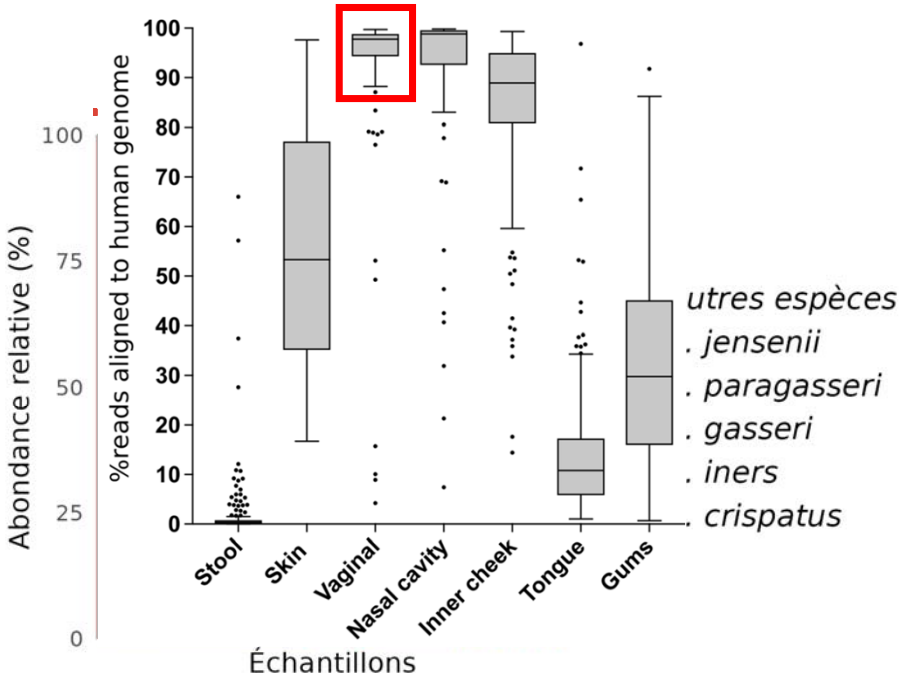
CST I	*L. crispatus*
CST II	*L. gasseri*
CST III	*L. iners*
CST IV	autres
CST V	*L. jensenii*

Techniques d'analyse :

- Observations microscopiques, qPCR
- Métabarcoding 16S manque de résolution (très conservé chez les *Lactobacillus)*
- Métagénomique *shotgun* très coûteuse, car fortement contaminée par l'ADN humain.

Enjeux :

Identifier et quantifier les espèces de *Lactobacillus* afin de déterminer les CSTs



Taux d'ADN humain en métagénomique *shotgun*, Marrotz *et al.* 2018

Travaux préliminaires : choix des marqueurs

Matériel

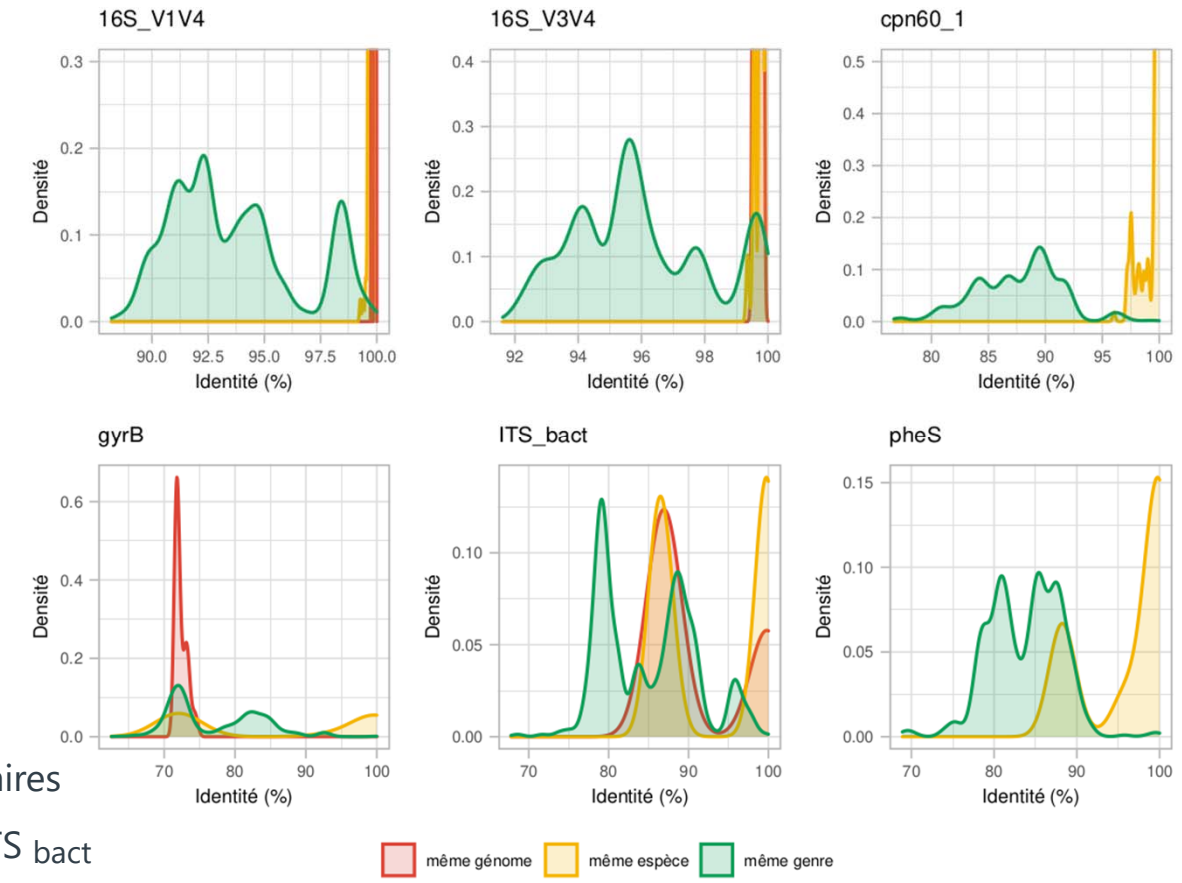
- 7 Marqueurs
- 94 génomes d'intérêt

Critères

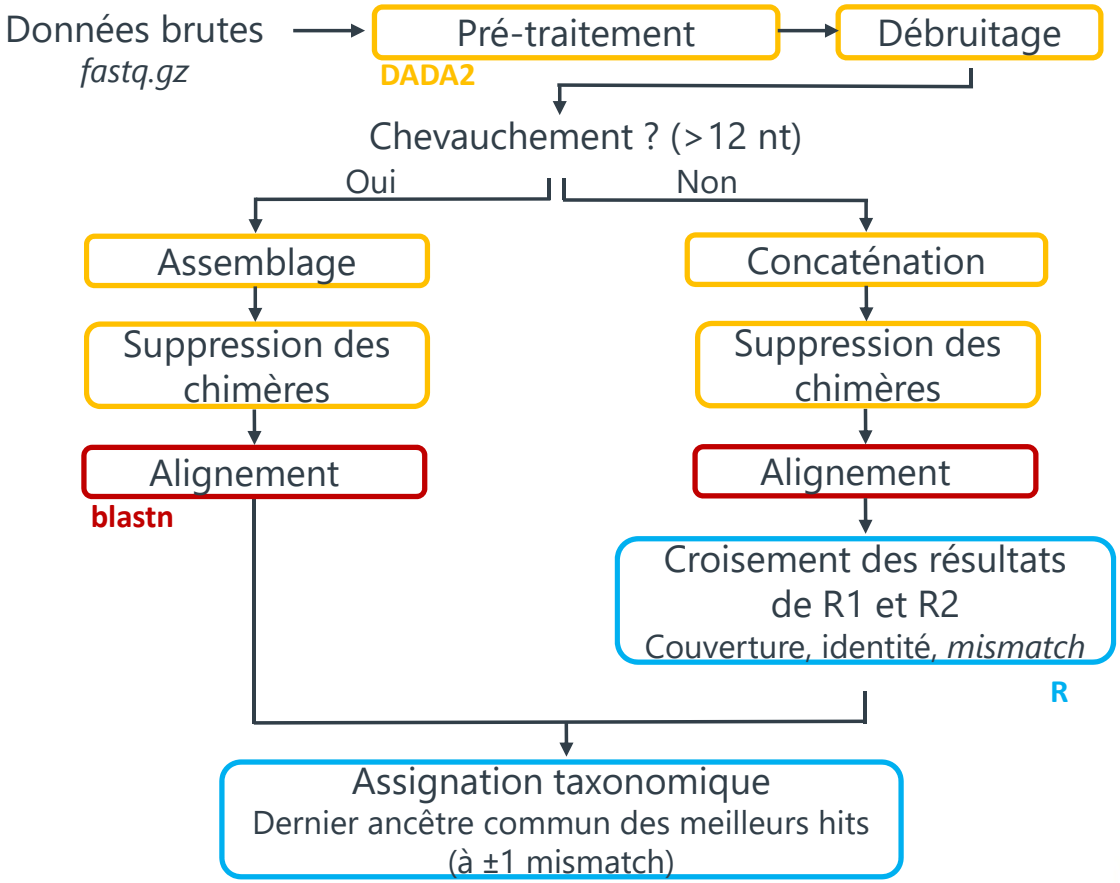
- Amplification
- Nombre de copies
- Taille de l'amplicon
- Pouvoir discriminant

Bilan

- Aucun marqueur n'est parfait
- Les marqueurs sont complémentaires
- 16S V1V4, 16S V3V4 gyrB, cpn60, ITS bact



Travaux préliminaires : pipeline pour chaque marqueur



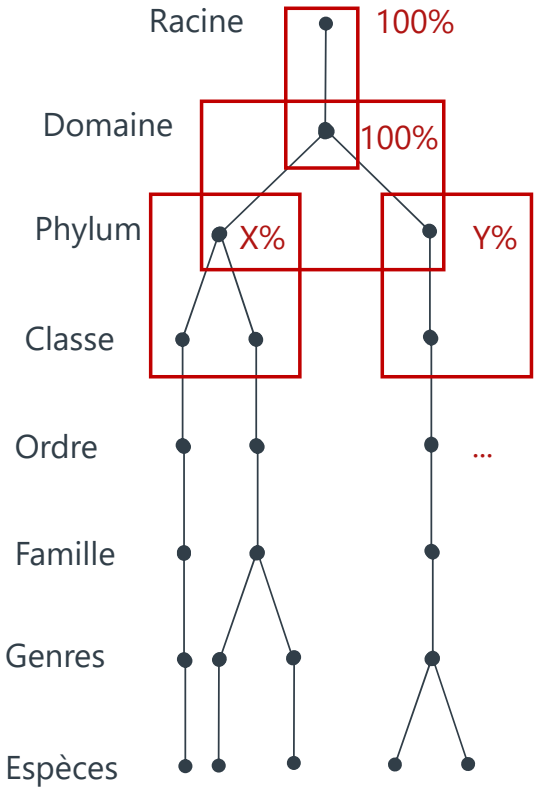
→ Largement inspiré de FROGS (Escudié *et al.* 2018)

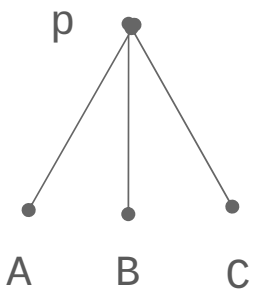
Bases de données :

- 16S : Silva 138
- gyrB : Poirier *et al.* 2018
- cpn60 : cpnDB (Kryachko *et al.* 2017)
- ITS_{bact} : Milani *et al.* 2020

→ Transposition à la **taxonomie à jour** du NCBI

Méthode consensus : approche récursive





Sources de discordance entre marqueurs :

- Pouvoir résolutif
- Biais d'universalité des amorces
- Biais du nombre de copies

	Marqueur 1 (m1)	Marqueur 2 (m2)	Marqueur 3 (m3)
parent P	3500	2000	2500
enfant A	2000 (57%)	1200 (60%)	1000 (40%)
enfant B	500 (14%)	0	500 (20%)
enfant C	1000 (28%)	500 (25%)	1000 (40%)

Méthode consensus : estimation des abondances relatives

	m1	m2	m3
P	3500	2000	2500
A	2000 (57%)	1200 (60%)	1000 (40%)
B	500 (14%)	0	500 (20%)
C	1000 (28%)	500 (25%)	1000 (40%)

Ratio entre chaque paire d'enfants →

log2ratio	m1	m2	m3	Moyenne
A/B	2	NA	1	1.5
A/C	1	1.26	0	0.75
B/C	-1	NA	-1	-1

→ Retour aux abondances relatives

Consensus (C1)

A	B	C
0.51	0.17	0.32

→ Alternatives :

- Méthode C2 : maximum des ratios enfant/parent
- Méthode C3 : moyenne après transformation par *centered log ratio* (clr)

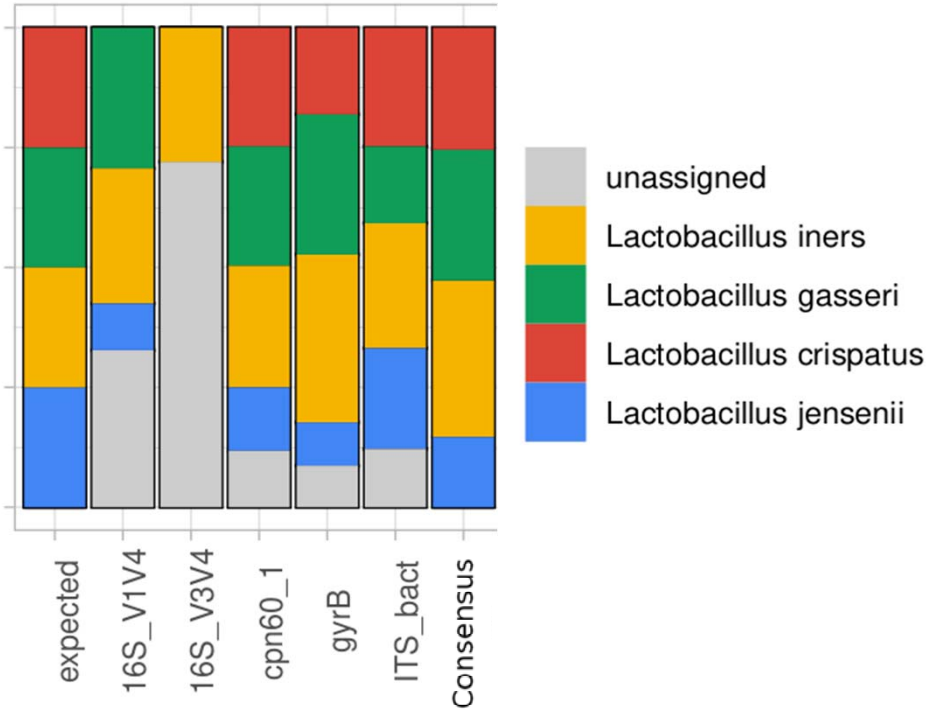
→ Système de pondération pour donner plus de poids aux marqueurs plus résolutifs

L' approche multi-marqueurs améliore l'identification des espèces de *Lactobacillus*

Matériel :

- Mélange équilibré des 4 espèces de *Lactobacillus* d'intérêt
- 5 marqueurs

- Mélange des génomes de RefSeq
- 10K *reads /* échantillon / marqueur
- 2*251bp avec modèle d'erreur empirique



Abondances relatives des espèces estimées par les différentes méthodes

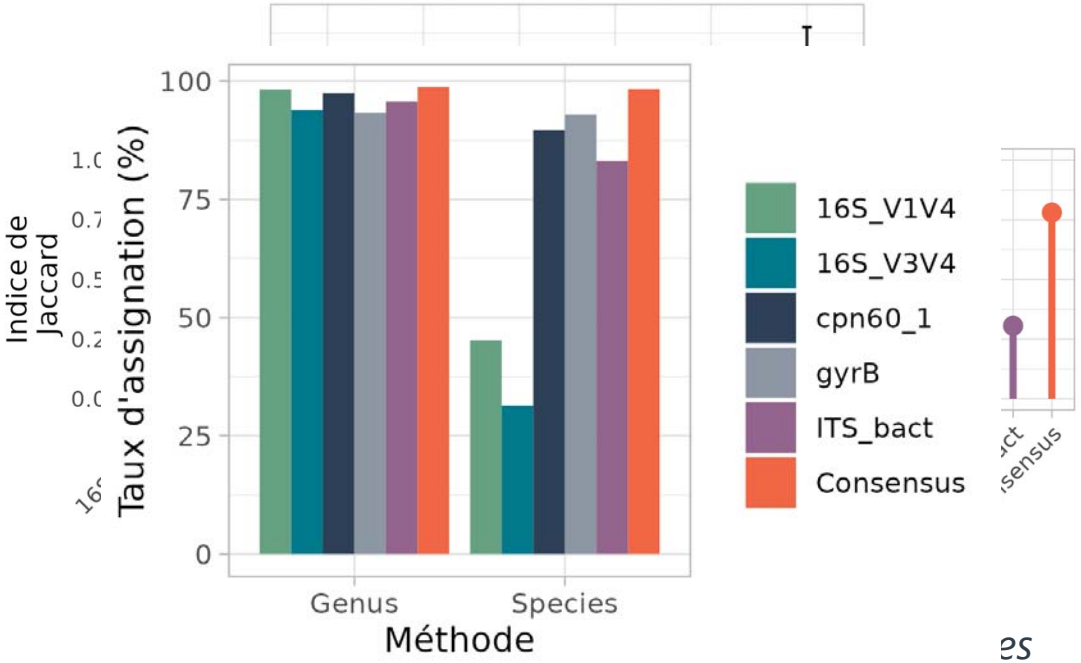
Simulations de microbiotes vaginaux réalistes

Matériel :

- 96 échantillons disponibles dans CuratedMetagenomicData
- 167 espèces au total
- Biais d'universalité et biais du nombre de copies

Résultats :

- Meilleure assignation taxonomique
- Meilleure estimation de la composition en espèces

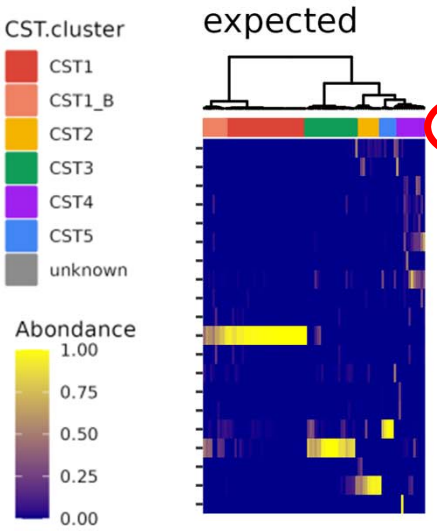


Taux d'assignations taxonomiques des différentes méthodes

Distances de Jaccard entre les compositions en espèces attendues et estimées par les différentes méthodes

Identification des CSTs

→ Clustering hiérarchique basé sur les distances de Bray-Curtis entre les échantillons



Clustering et identification des CSTs par les différentes méthodes

Cas problématiques

	attendue
Fusobacteria	1.4
Tenericutes	1.4
Actinobacteria	3.6
Bacteroidetes	1.4
Firmicutes	92.1
unassigned	.0
Total	100.0

Bilan

- Les profils taxonomiques multi-marqueurs **tirent parti** de l'**universalité** et du **pouvoir discriminant** de chaque marqueur pour donner un profil taxonomique plus proche des profils théoriques (sur simulation)
- Dans le cadre du **microbiote vaginal**, les approches multi-marqueurs permettent de mieux estimer la composition en espèces des échantillons, d'identifier toutes les espèces de *Lactobacillus* d'intérêt, et ainsi d'identifier correctement les CSTs.
- Il vaut mieux être **conservateur** dans l'**assignation taxonomique**, pour limiter les faux positifs
- Package **R** : milou (***m****ulti*-***l****oci metabarc****o****ding consens****u****s finder*) *[en cours]*

  > remotes::install_gitlab(repo = "benoit.goutorbe/milou", host = "forgemia.inra.fr")

Perspectives

- **Optimisation de coûts** : nombre de marqueurs séquencés et profondeur de séquençage
- **Validations biologiques** : *mock community* et échantillons vaginaux *[en cours]*
- **Publication** de la méthode *[en cours]*
- Applications à d'**autres écosystèmes**
- Améliorations/enrichissements de la méthode :
  - → Système de pondération (évaluer l'universalité des marqueurs)
  - → Analyses de diversité (α et β)
  - → Information phylogénétique
  - → …

Merci

MaIAGE

Sophie Schbath

Anne-Laure Abraham

Mahendra Mariadassou

Plateforme Migale

CRCM

Ghislain Bidaut

Laboratoire Alphabio

Philippe Halfon

Anne Plauzolles

Marion Bonnet

Sabrine Bellabes