Quel modèle et quelle base de données 16S pour l'analyse du microbiome urinaire bactérien?

Marthe Leoz[1,2], Johan Brière[1,2], Marie Leoz[1], Camille Villenave[1], Martine Pestel-Caron[3], and Sandrine Dahyot[3]

[1]Univ Rouen Normandie, Université de Caen Normandie, INSERM, Normandie Univ, DYNAMICURE UMR 1311, F-76000 Rouen, France; [2]DataScientest, 92800 Puteaux, France;
[3]Univ Rouen Normandie, Université de Caen Normandie, INSERM, Normandie Univ, DYNAMICURE UMR 1311, CHU Rouen, Department of Bacteriology, F-76000 Rouen, France
Corresponding Author: marthe.leoz@gmail.com

Contexte

L'analyse du microbiome urinaire par metabarcoding 16S pose des difficultés de spécificité et sensibilité. Les résultats obtenus peuvent varier d'une approche à l'autre, et certains genres / espèces d'intérêt peuvent être mal identifiés. Parmi eux, on note des uropathogènes (*Escherichia coli, Klebsiella pneumoniae…*) et des représentants importants du microbiome (dont les différents lactobacilles).
L'objectif de ce travail est d'optimiser un outil d'analyse des séquences des régions V3-V4 du gène 16S obtenues selon le protocole standard Illumina.

Jeux de référence et de test

STUDY → **Espèces observées dans le milieu urinaire (taxo NCBI 2025)**

Baddoo et al., Front Urol, 2024
Ugarcina et al., Microbiol Spectr, 2022
Dubourg et al., Front. Microbiol, 2020
Morand et al., J Clin Microbiol, 2019

Jeu de référence / d'entraînement des modèles
- Nombre min-max de représentants par espèce (ex: Urinary_1-20)
- 16S complet vs extraction V3-V4
- Filtrage (suppression des ids présents dans le jeu de test)

de 3370 à 16172 séquences selon les jeux

Jeu de test
- Extraction V3-V4
- Suppression des séquences doublons

843 séquences extraites aléatoirement (20% du plus petit jeu de référence)

Modèles

Alignment-based models

Rapid Fuzz : généraliste parcimonie

Blastn : nucléotide - nucléotide extension de k-mer

```
\AATGAAGATCAAAAGCTTATTCATCTGTTT
 |||||||||  | | | |||||  |||||  |||
\AATGAAGACCTAGA-CTTATCCATCTCTTT
```

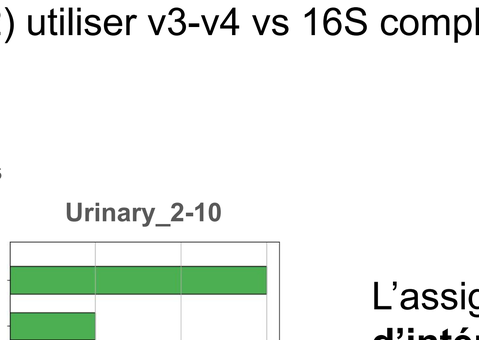
➤ Retournent des scores de similarité

Machine Learning (ML) models

Data Pre-processing

CTAGTTGGATAC
→ extraction k-mers
→ [CTAGTTG, TAGTTGG, AGTTGGA, GTTGGAT, …]
→ vectorisation
→ [3.256, -1.437, 6.524, …]

Cosine / knn



➤ Retourne la classe la plus proche (Cosine), ou la classe majoritaire parmi les K plus proches voisins (K-Nearest Neighbors, knn)

Naive Bayes (NB)

$$P_{classe|kmer} = \frac{P_{kmer|classe} \times P_{classe}}{P_{kmer}}$$

➤ Retourne la classe la plus probable

Résultats

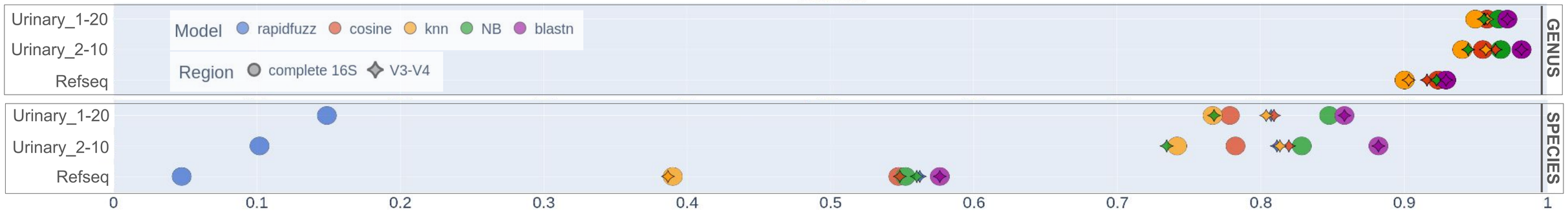


Fig1: F1-score moyen par modèle (couleur), région d'entraînement (symbole), jeu de référence/entraînement et niveau taxonomique.

L'analyse des F1-scores (Fig.1) indique un **bénéfice des jeux de référence / d'entraînement spécifiques du milieu urinaire,** avec peu ou pas d'avantage à (1) équilibrer le nombre de représentants par espèce; (2) utiliser v3-v4 vs 16S complet.
A ce stade, RapidFuzz < modèles ML < Blastn.

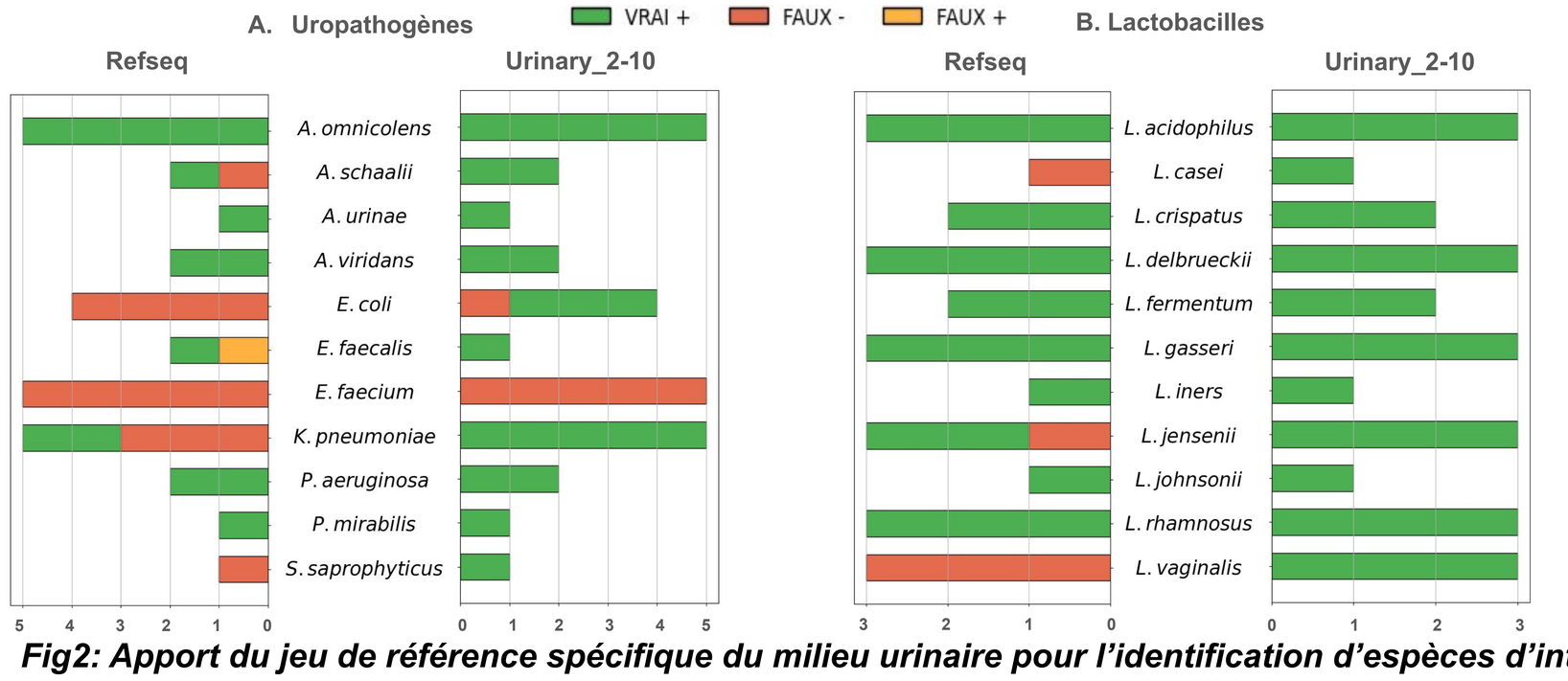


Fig2: Apport du jeu de référence spécifique du milieu urinaire pour l'identification d'espèces d'intérêt avec Blastn.

L'assignation taxonomique des **espèces d'intérêt** est améliorée par l'utilisation d'un jeu de référence spécifique du milieu urinaire (Fig.2):
- **Uropathogènes:**
  - Moins de faux+ et faux- (notamment chez *E.coli* et *K.pneumoniae*)
  - *E.faecium* à améliorer
- **Lactobacilles:**
  - Plus aucun faux- sur ce jeu test

Discussion et perspectives

- Optimisation des modèles ML: classes mieux équilibrées (enrichir les sous-représentées), varier les tailles de k-mers et les jeux de test, le nombre de voisins (knn), l'estimateur alpha (NB), tester le modèle de vectorisation DNA-BERT
- Limite jeu de référence urinaire: pas de possibilité de découverte de nouvelle espèce