Polymorphisme de taille des ITS

Impact sur les analyses « métabarcoding » et solutions pour détecter et conserver les ITS de grande longueur

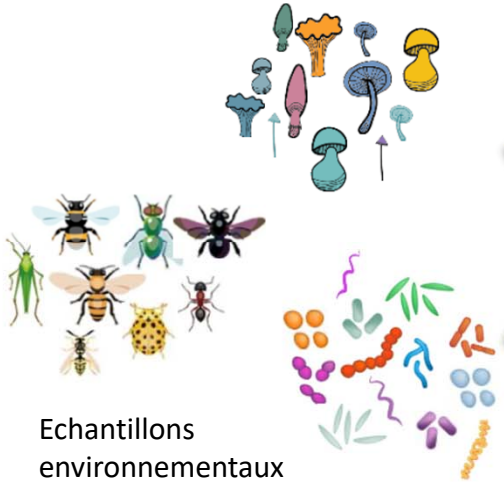
Maria Bernard, Olivier Rué, Mahendra Mariadassou et Géraldine Pascal
Présentation Lucas Auer

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

INRAE

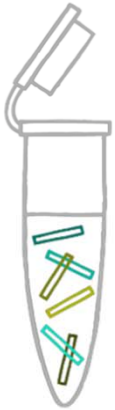
INTERACTIONS ARBRES-MICROORGANISMES · Genotoul Bioinfo · SIGENAE · GenPhySE · MaIAGE · GABI · migale

Principes du métabarcoding



Echantillons environnementaux

Extraction ADN

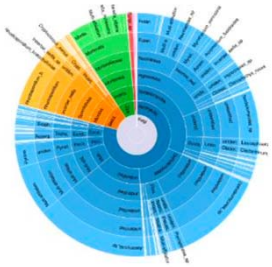


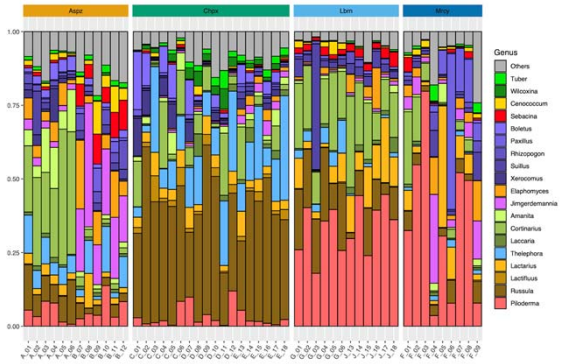
Amplification du marqueur

ADNr 16S/18S

Séquençage du marqueur

Tables d'abondances avec annotations taxonomiques



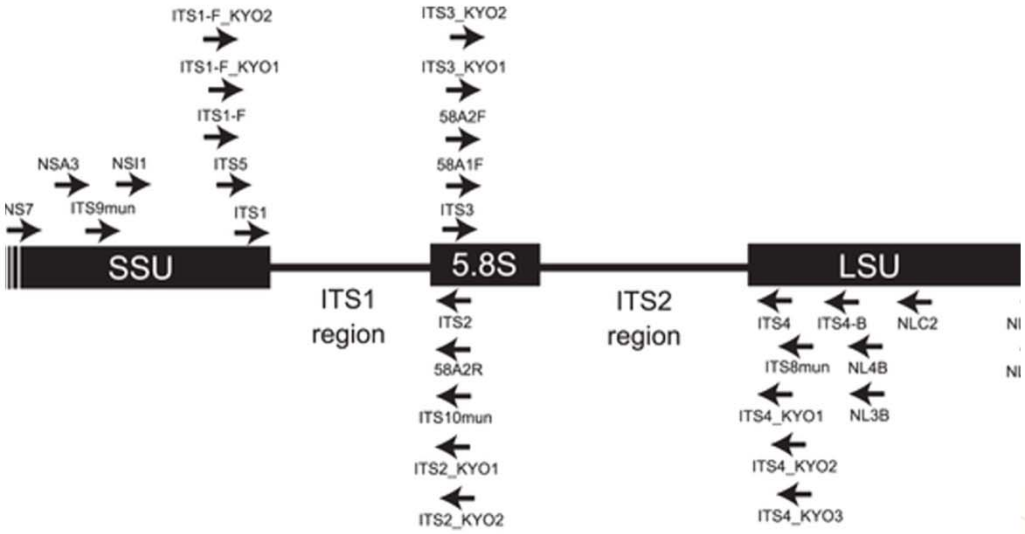


Traitement bio-informatique des données

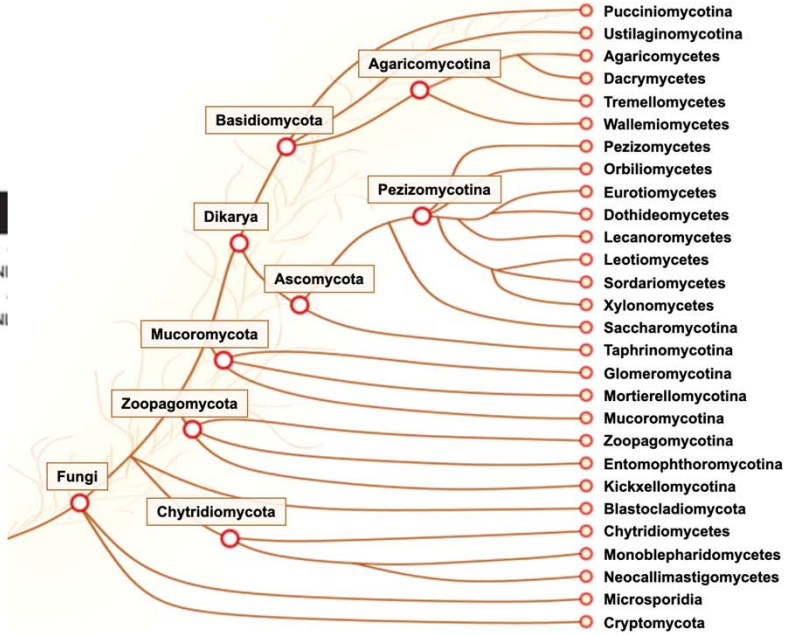
Marqueurs fongiques et ITS

- L'ADNr 18S est assez peu spécifique (trop peu de variations entres espèces)
- Les régions inter- sous-unités sont transcrites (***I**nternal **T**ranscribed **S**pacers*) et très variables

Map of nuclear ribosomal RNA genes and their ITS regions.



Toju H, Tanabe AS, Yamamoto S, Sato H (2012) High-Coverage ITS Primers for the DNA-Based Identification of Ascomycetes and Basidiomycetes in Environmental Samples. PLOS ONE 7(7): e40863. https://doi.org/10.1371/journal.pone.0040863



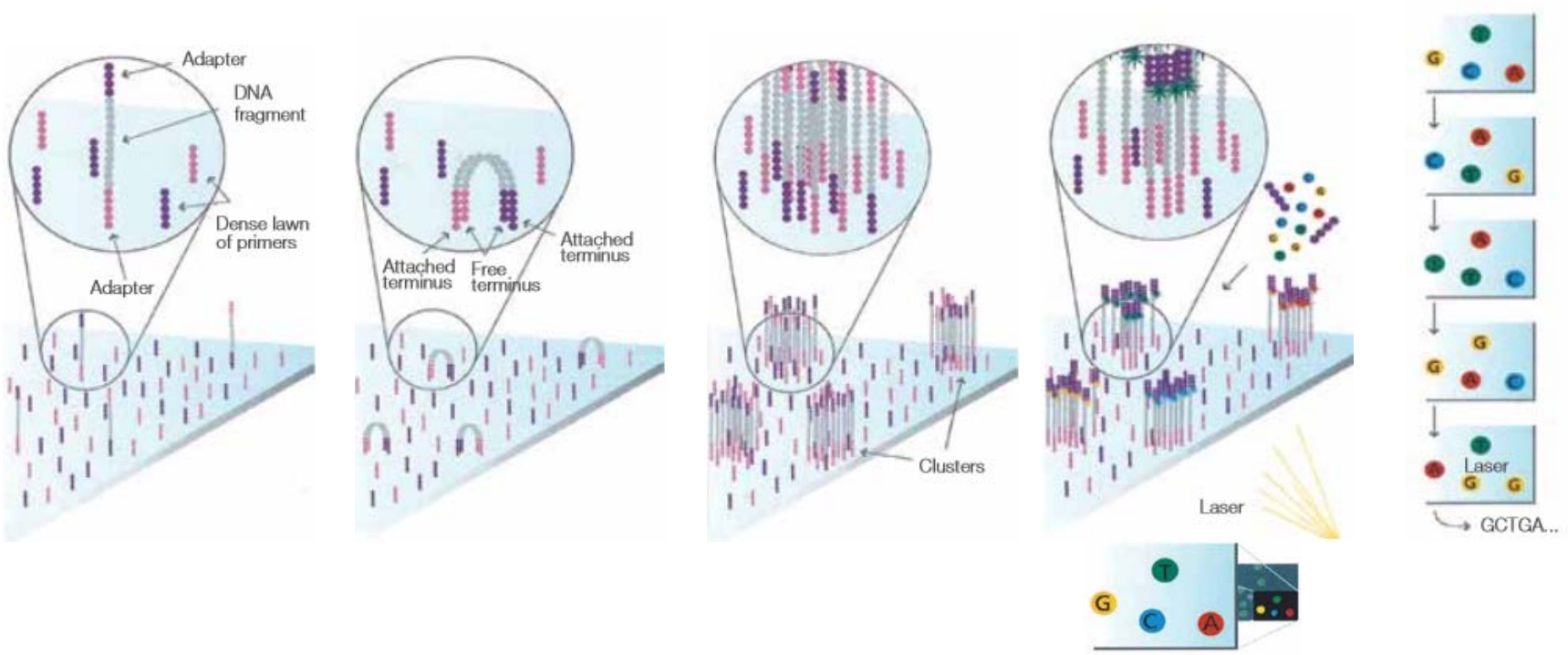
Amorces « universelles fongiques » pas si universelles que ça

pour certains taxons basaux des champignons (dont les Glomeromycotina, endomycorhiziens symbiotiques fondamentaux dans certains écosystèmes)

Nombre de copies de l'opéron très élevé et très variable

14 à 1400, médiane 80 pour moyenne 113 estimé sur une centaine d'espèces

Séquençage Illumina paired-end



En fin de lecture R1, un cycle de « bridge amplification » a lieu pour synthétiser les séquences « sens reverse » puis réaliser les cycles de séquençage lecture R2.
On obtient deux lectures de 250 (ou 300 selon les kits de réactif), une par extrémité de la molécule séquencée.

Problématique de polymorphisme de longueur

Chevauchement complet (amplicon de longueur 250)
Cas idéal, permet d'améliorer la qualité des bases en comparant les lectures R1 et R2.
Région hypervariable 16/18S, taxons à ITS courts

Chevauchement partiel (amplicon <480)
Permet d'augmenter la longueur totale séquencée en joignant les extrémités terminales de R1 et R2
Doubles régions hypervariables (V3-V4 16S), ITS moyens

Des lectures non jointives sont dans ces deux cas signes de mauvaise qualité ou d'erreurs

Pas de chevauchement (amplicon >480)
Pas de possibilité de joindre les lectures R1 et R2 et d'obtenir la séquence complète de la molécule amplifiée
Couples d'amorces mal choisis (régions trop éloignées), ITS de grande longueur

Solutions classiques

Exclusion des lectures non jointives

Situation non attendue dans la plupart des pipelines d'analyse : avec des amorces bien choisies en 16S/18S/gyrB/etc cela ne doit pas arriver.

→ élimination de la paire de lectures
(Mothur, qiime, dada2, …)

Utilisation d'une seule des deux lectures R1 ou R2

Pour pallier à ce problème spécifique aux ITS, il est possible de ne conserver qu'une seule des deux lectures et de se limiter à l'information de 250-300 pb

→ élimination d'une lecture
(possible hors usage normal, requière quelques compétences)

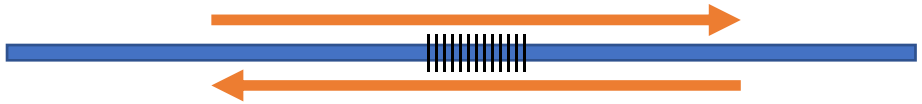
Utilisation de R1 quand R1 et R2 sont non jointives

Intermédiaire entre les deux cas précédents

(proposé par usearch)

Solution proposée par FROGS

Si lectures jointives → jointure classique :

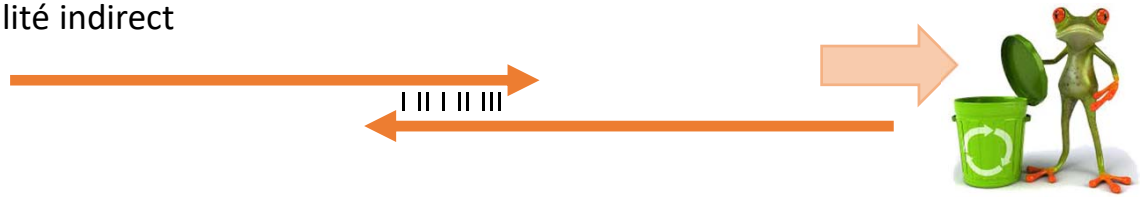


Si lectures non jointives → jointure artificielle :

Pour les lectures non jointives **uniquement**, production de séquences artificiellement combinées avec R1 et R2 reliées par une série de N.

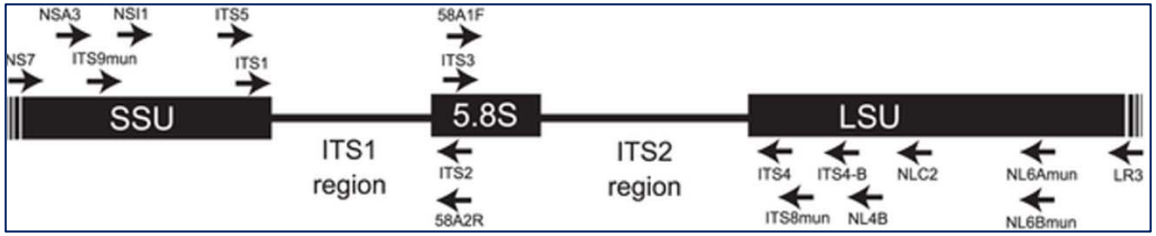
Conséquences sur le reste de l'analyse

Habituellement les séquences sont filtrées sur la bonne correspondance entre les fins de R1 et R2 (qui présentent souvent des baisses de qualité)
→ filtre de qualité indirect

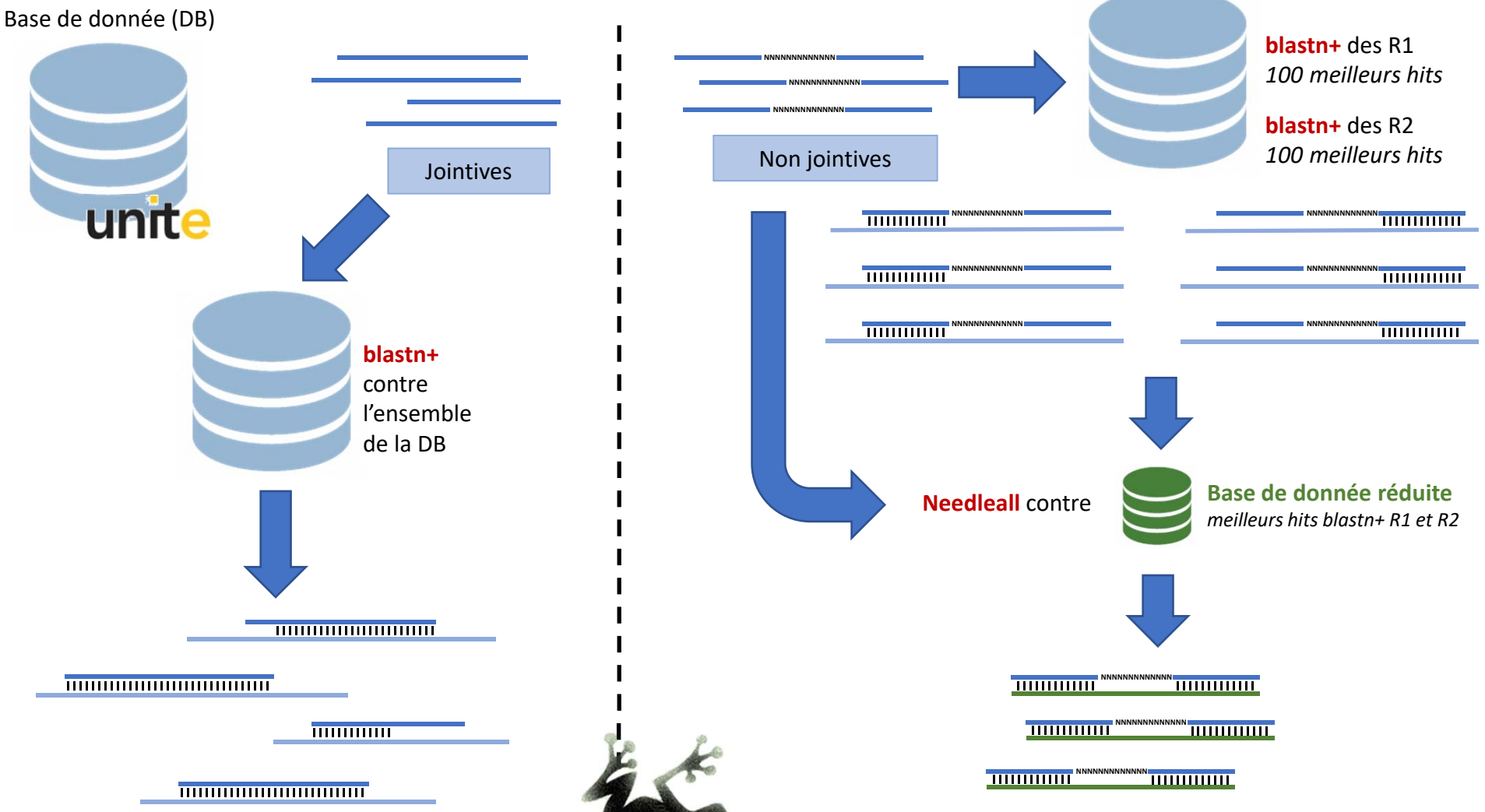


Pas de différence possible entre R1-R2 non jointives ou jointives avec trop d'erreurs

Compensation possible par un filtre ITSx pour éliminer toute séquence qui ne ressemble pas à une région ITS.



Affiliation taxonomique



Calcul des scores

Cas du % d'identité

$$\%id = \frac{\text{nombre correspondances}}{\text{nombre positions alignées}}$$

- R1/R2 jointives, longueur = 400

%id = 400 / 400 = 100%

- R1/R2 non jointives, longueur du meilleur hit = 500

%id = (250+250) / (500+100) = 83%

- R1/R2 non jointives, longueur du meilleur hit > 500

%id = (250+250) / (600+100) = 71%

Alignements parfaits mais mauvais scores… (pouvant faire rejeter l'annotation!)

FROGS

$$\%id_{\text{non jointives}} = \frac{\text{nombre correspondances}}{\text{nombre positions } \textbf{séquencées}}$$

%id = (250+250) / (500-100) = 100%

Evaluation – données simulées

Source unite *(génomes incertae ou avec bases ambiguës exclus)*

unite

Sous-échantillonnages de 20, 100 ou 500 séquences
maximisant la dispersion dans l'arbre phylogénétique

Ajout de 15 ITS « longs » (13 ITS1, 2 ITS2)

Extraction (ITSx) des régions ITS1 et ITS2

Simulation des régions amplifiées par les amorces

Simulation (***Grinder***) de 10 échantillons
de 100 000 lectures R1 et R2 par condition
2 distributions d'abondances, uniforme et loi de puissance
Introduction d'erreurs et de 2% de chimères.

Evaluation – outils comparés

- **DADA2-se** (v1.41.1)
- **DADA2-pe** (v1.41.1)
- **QIIME-se** (qiime2-2019.4)
- **QIIME-pe** (qiime2-2019.4)
- **USEARCH** (usearch11.0.667)
- **FROGS** (v3.2.0)

pe : paired-end, utilise les lectures R1 et R2 jointives
se : single-end, utilise uniquement le R1

Tool	Reads.processing	Clustering.denoising	Affiliation	Other
FROGS	ALL	clustering swarm	blast & needleall	ITSx
qiime-se	R1 only	open reference clustering (OTUs)	classifier (sklearn)	/
qiime-pe	merge R1 & R2	open reference clustering (OTUs)	classifier (sklearn)	/
dada2-se	R1 only	denoising (ASVs)	classifier (RDP)	/
dada2-pe	merge R1 & R2	denoising (ASVs)	classifier (RDP)	/
usearch	merge R1 & R2 + R1 only	denoising (zOTUs)	classifier (sintax)	/

Evaluation – métriques

Profondeur : nombre de lectures conservées après les différents filtres

Richesse : nombre d'OTU/clusters/zOTUs/ASV produits

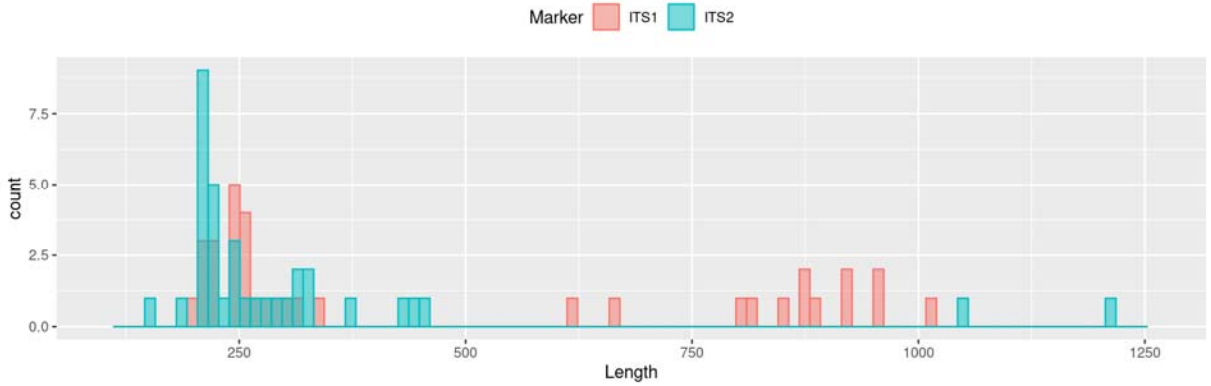
Divergence : distance de Bray-Curtis entre les abondances attendues et observées à un niveau taxonomique donné

FN : faux négatifs, nombre de taxons attendus et non détectés
FP : faux positifs, nombre de taxons non attendus et détectés
TP : vrais positifs, nombre de taxons attendus et détectés

Précision : $\frac{TP}{TP + FP}$

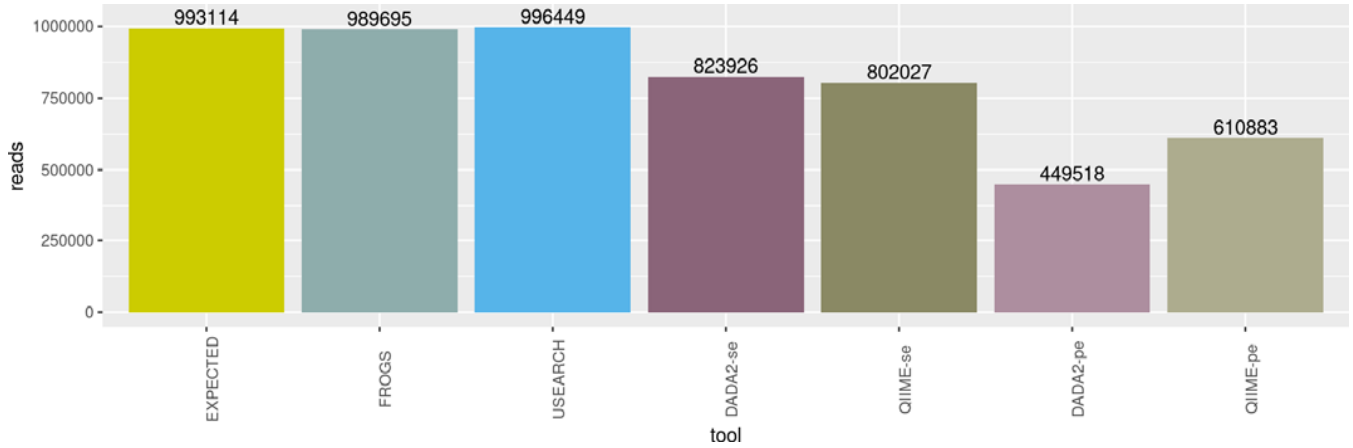
Sensibilité : $\frac{TP}{TP + FN}$

Jeu 35sp – longueurs des séquences :



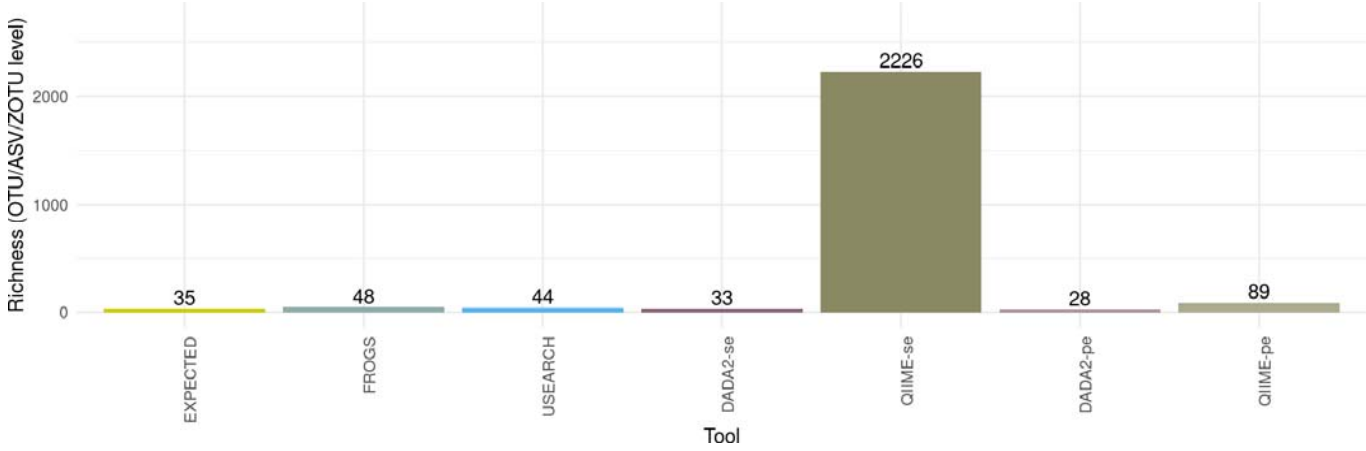
Evaluation – 35sp-PowerLaw

Profondeur et richesse



FROGS et **USEARCH** conservent quasiment toutes les lectures.
Les autres éliment les lectures non-jointives (pe) mais et parfois beaucoup d'autres (lectures singleton notamment)

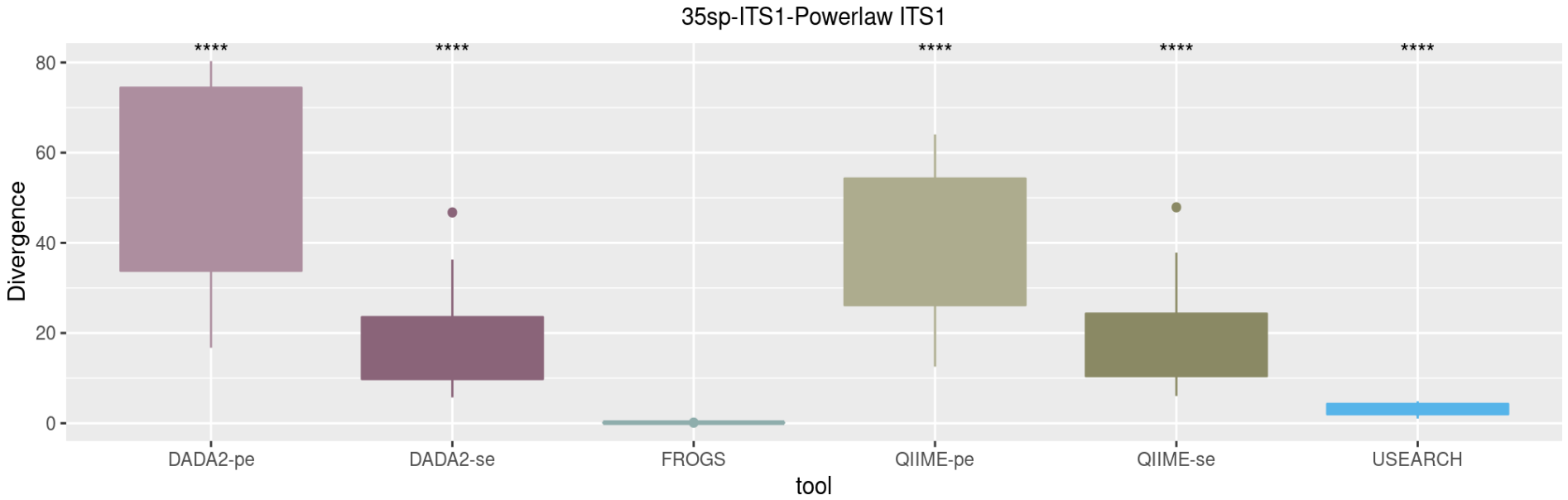
Un peu plus (**FROGS**, **USEARCH**) ou un peu moins **(DADA2-se**, **DADA-pe**) d'OTUs qu'attendu sauf pour **QIIME** qui explose.



Evaluation – 35sp-PowerLaw

Divergence

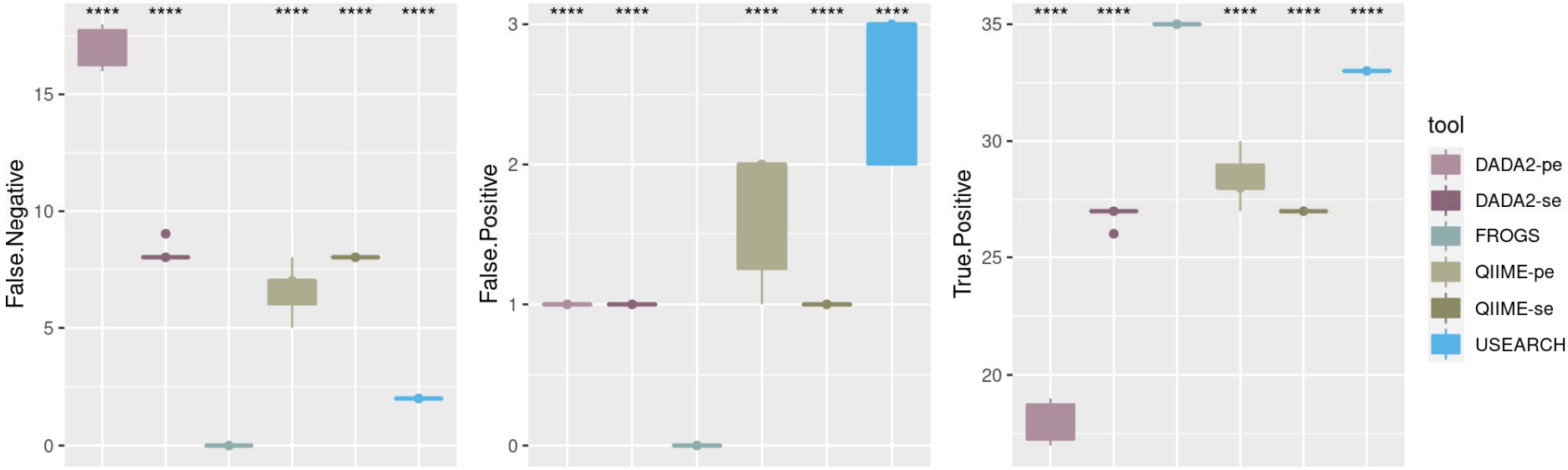
Bray-Curtis entre les abondances attendues et observées **(à l'espèce)**



Très forte divergence pour les approches paired-end (due à l'importance des ITS longs dans le jeu de données)
Les approches **single-end** sont meilleures mais restent avec des divergences significatives.

Evaluation – 35sp-PowerLaw

Faux négatifs – faux positifs – vrais positifs **(à l'espèce)**



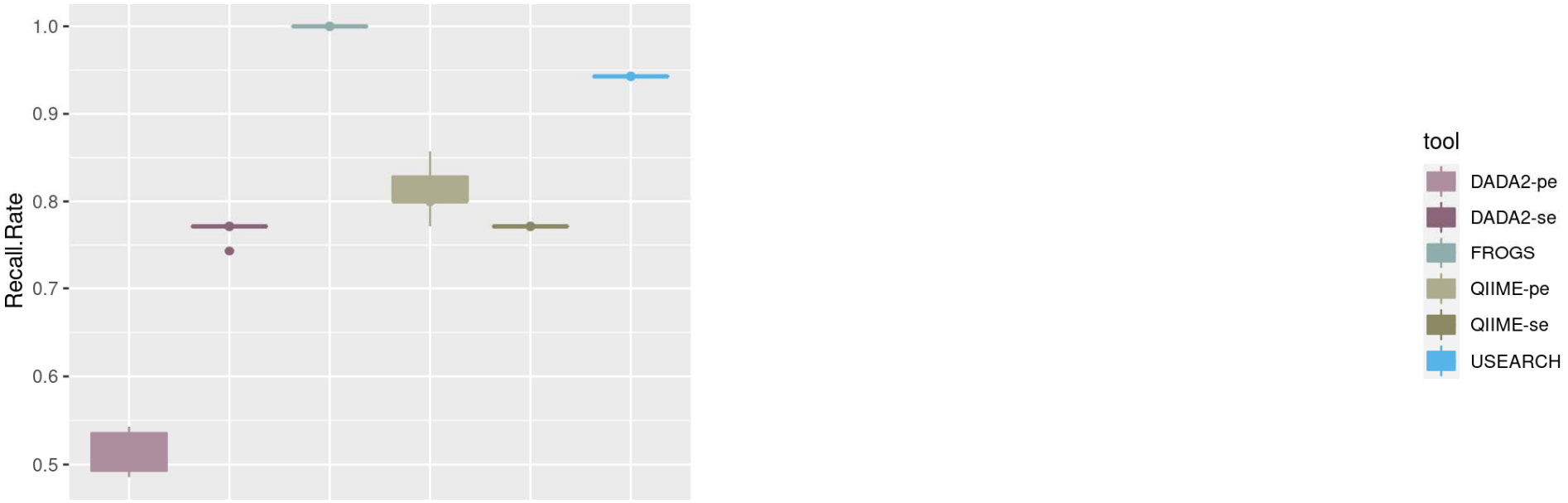
A l'espèce, **FROGS** a 0 faux négatifs et positifs pour 35 vrais positifs.
Tous les outils ne trouvent que peu de faux positifs (la richesse élevée de **QIIME** ne conduit qu'à quelques espèces non attendus → nombreux OTUs pour une seule espèce)

DADA2 a un nombre de faux négatifs (espèces attendues et non détectées) très élevé même en single-end.

Evaluation – 35sp-PowerLaw

Sensibilité (« recall rate ») et précision

Sensibilité : $\frac{TP}{TP + FP}$



Ratio entre espèces attendues détectées et espèces détectées + « manquées »

DADA2 a une assez bonne précision mais une faible sensibilité, surtout en **paired-end** (attendu du fait de la composition du jeu de données) mais aussi en **single-end**.

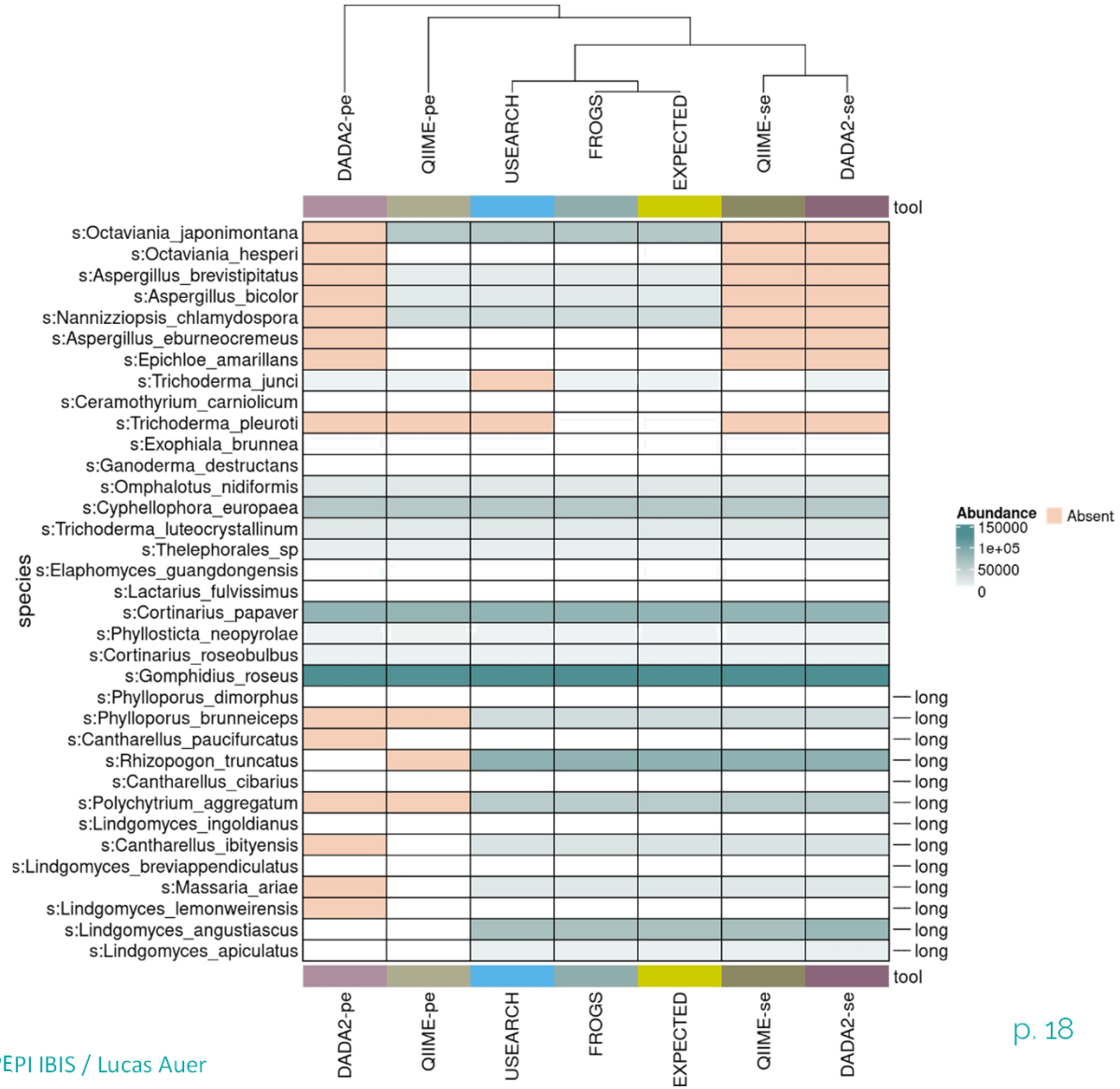
Evaluation – 35sp-PowerLaw

Présence-absence des taxons attendus

FROGS trouve les 35 espèces attendues, avec des abondances correspondant aux abondances attendues.

Comme attendu, les approches **paired-end** ratent la plupart des espèces à ITS longs. Mises à part celles-ci, tous les outils reconstituent assez bien les abondances des espèces détectées.

DADA2 rate plusieurs espèces, dont des espèces abondantes avec un ITS court.



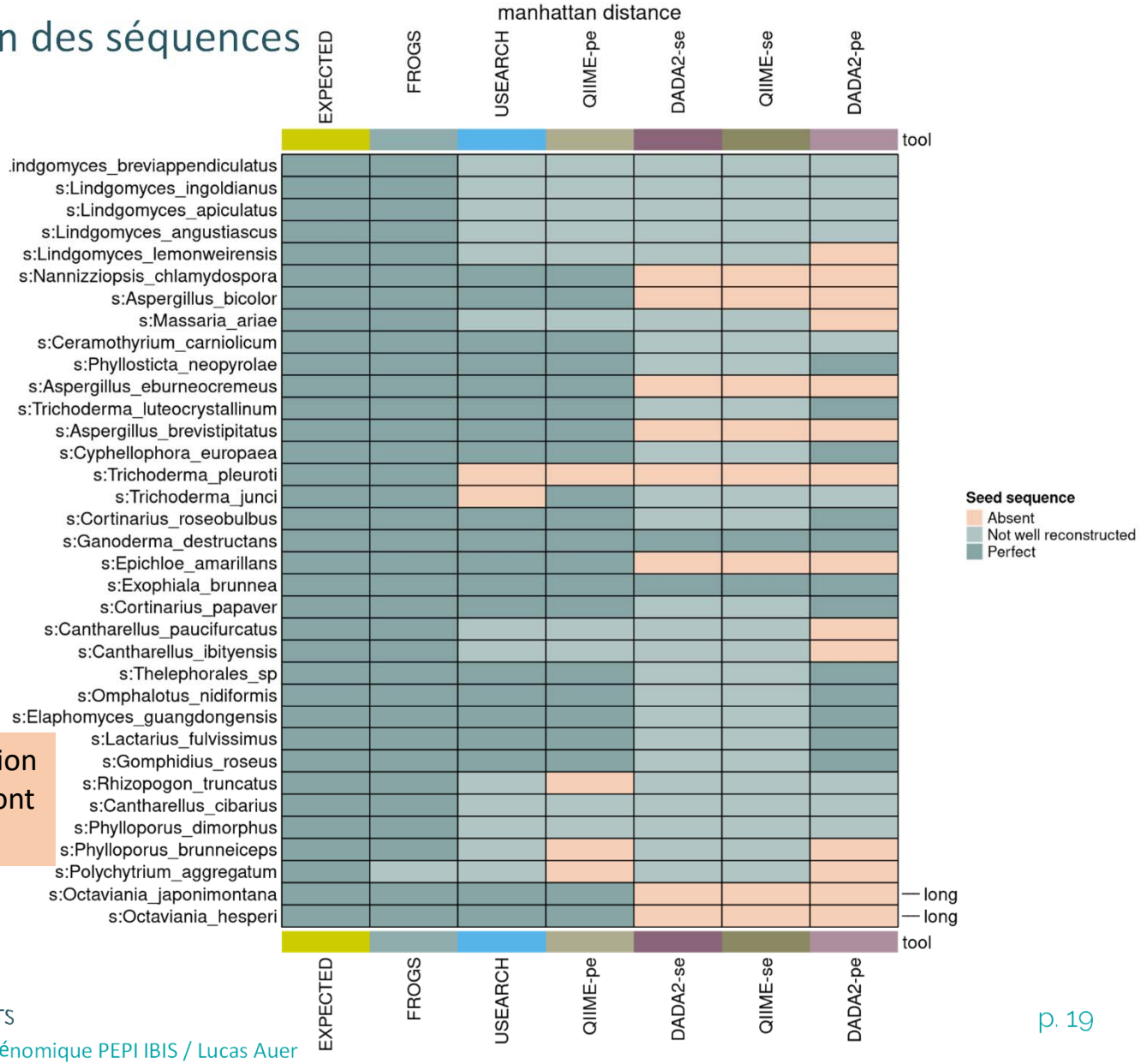
Evaluation – 35sp-PowerLaw

Reconstruction des séquences

FROGS retrouve les séquences exactes de presque toutes les espèces (notamment grâce à l'utilisation de PEAR pour le merging)

Les autres outils font largement moins bien, dont étonnamment DADA2.

Très problématique pour l'assignation taxonomique et si les séquences sont déposées dans des banques !

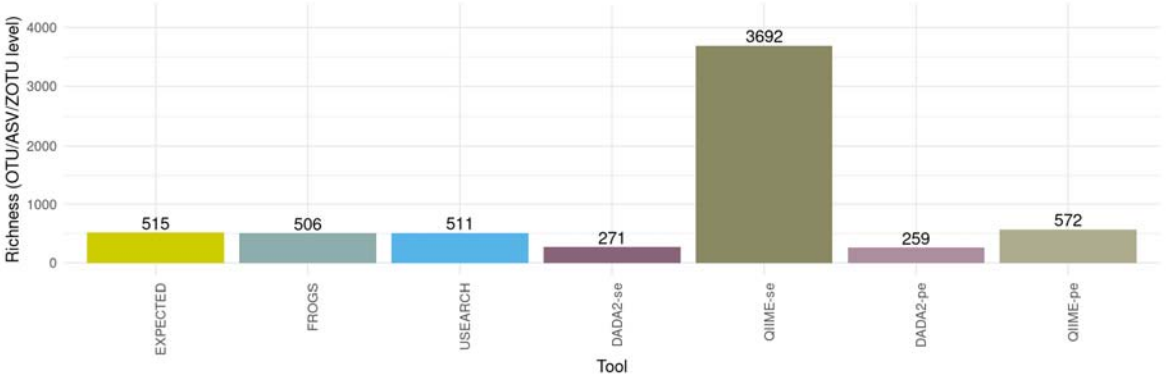


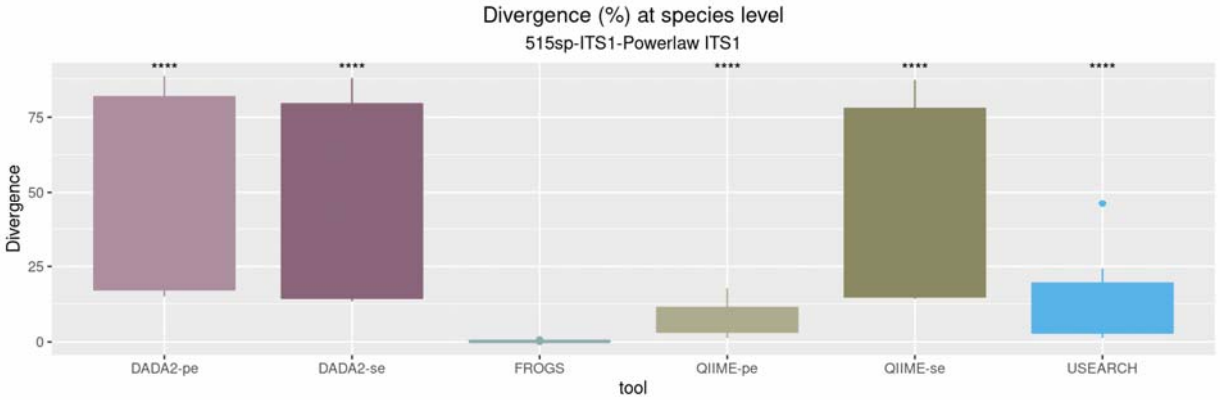
Evaluation – 515sp-PowerLaw

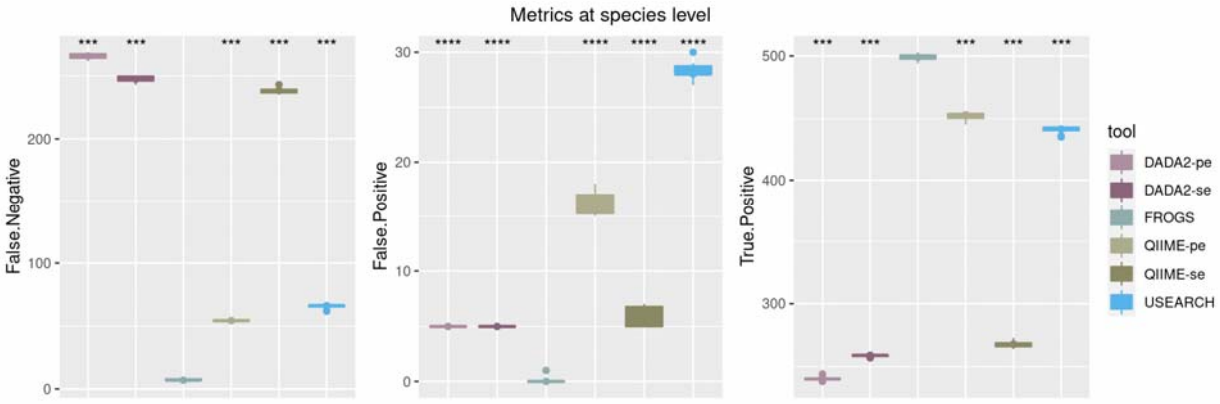
Avec une richesse attendue plus forte et beaucoup d'abondances faibles, **FROGS** a des performances plutôt stables (précision et sensibilité proches de 1).

DADA2 produit des divergences très élevées (autant que **QIIME-se**) liées à un taux de faux négatifs très élevé (sans doute parce qu'il traite les échantillons indépendamment).

USEARCH présente plutôt de bonnes performances malgré un taux de faux négatifs assez élevé.







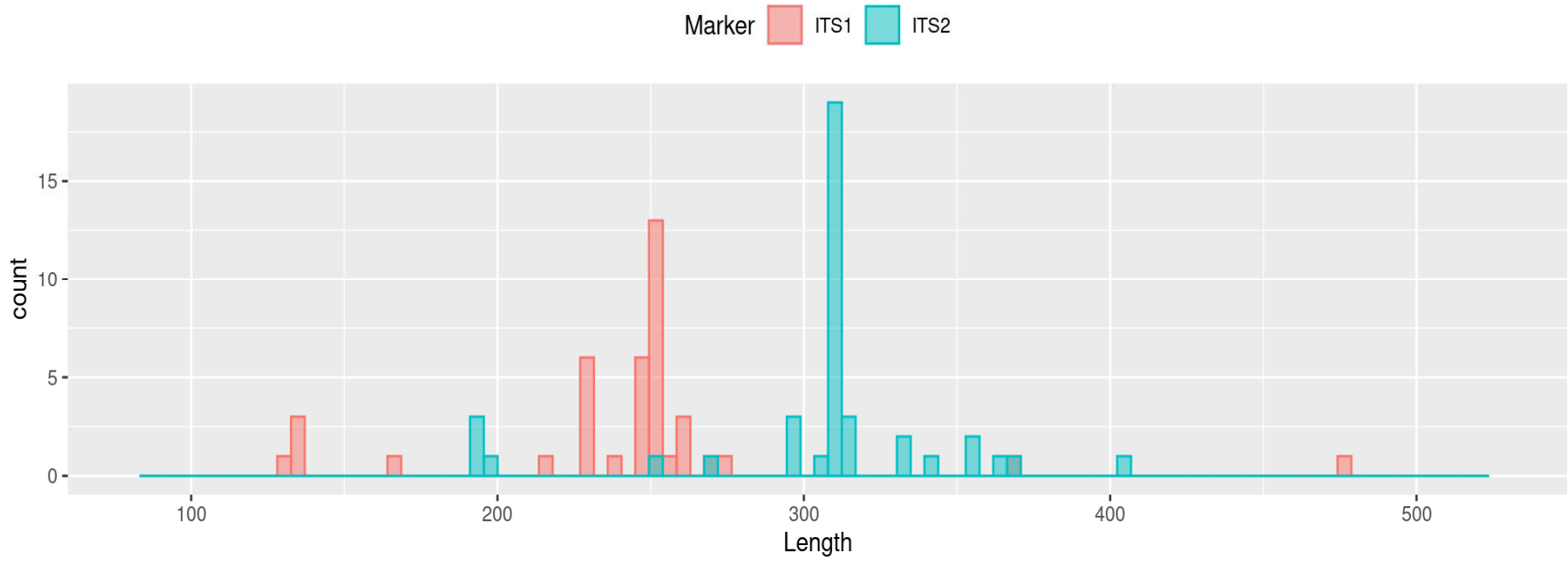
Evaluation – données réelles

Communauté synthétique

Mélange de 40 espèces provenant de produits alimentaires fermentés (projet METABARFOOD)
Séquences référence connues.
Mélanges **équimolaires** :

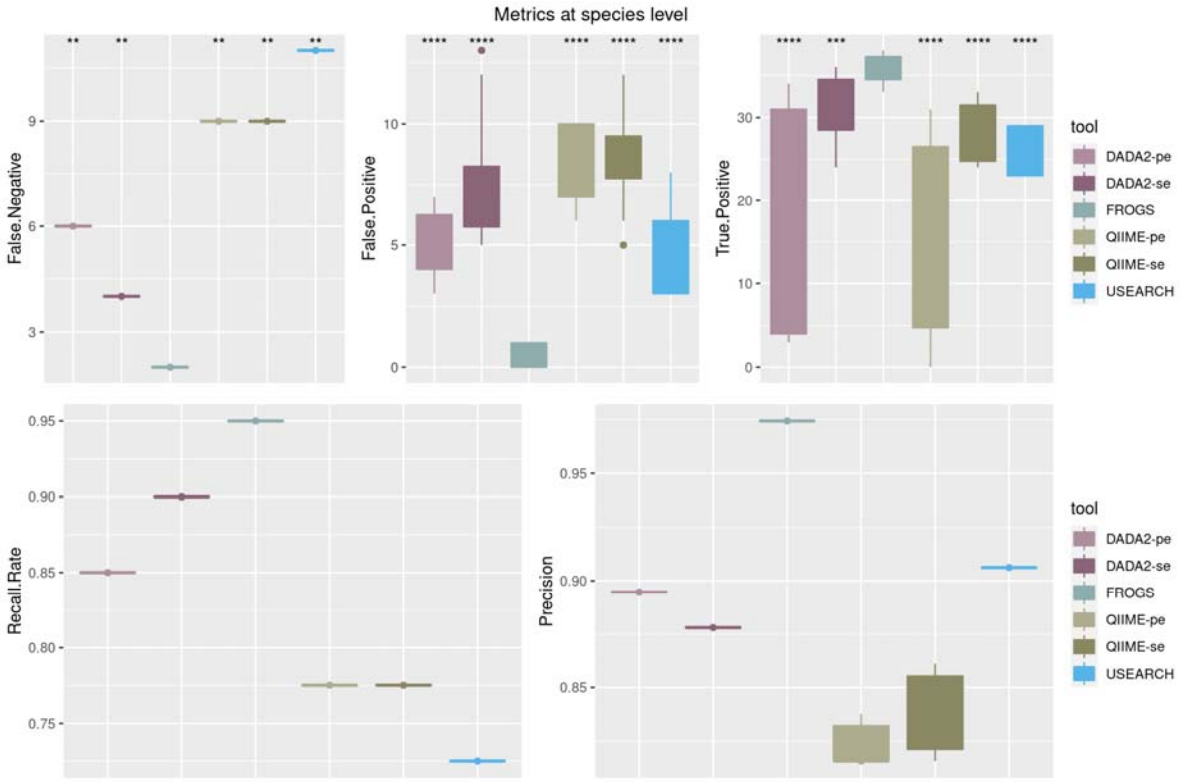
- d'**ADN génomique**
- de **produit de PCR**

Amplification ITS1 et ITS2, 3 réplicats chacun soit 12 échantillons séquencés en Illumina MiSeq 2x250.



Evaluation – données réelles

Résultats sur PCR – ITS1 : métriques



Plusieurs espèces non détectées par **FROGS**, mais qui a tout de même les meilleures performances (et les plus répétables) par rapport aux autres solutions.

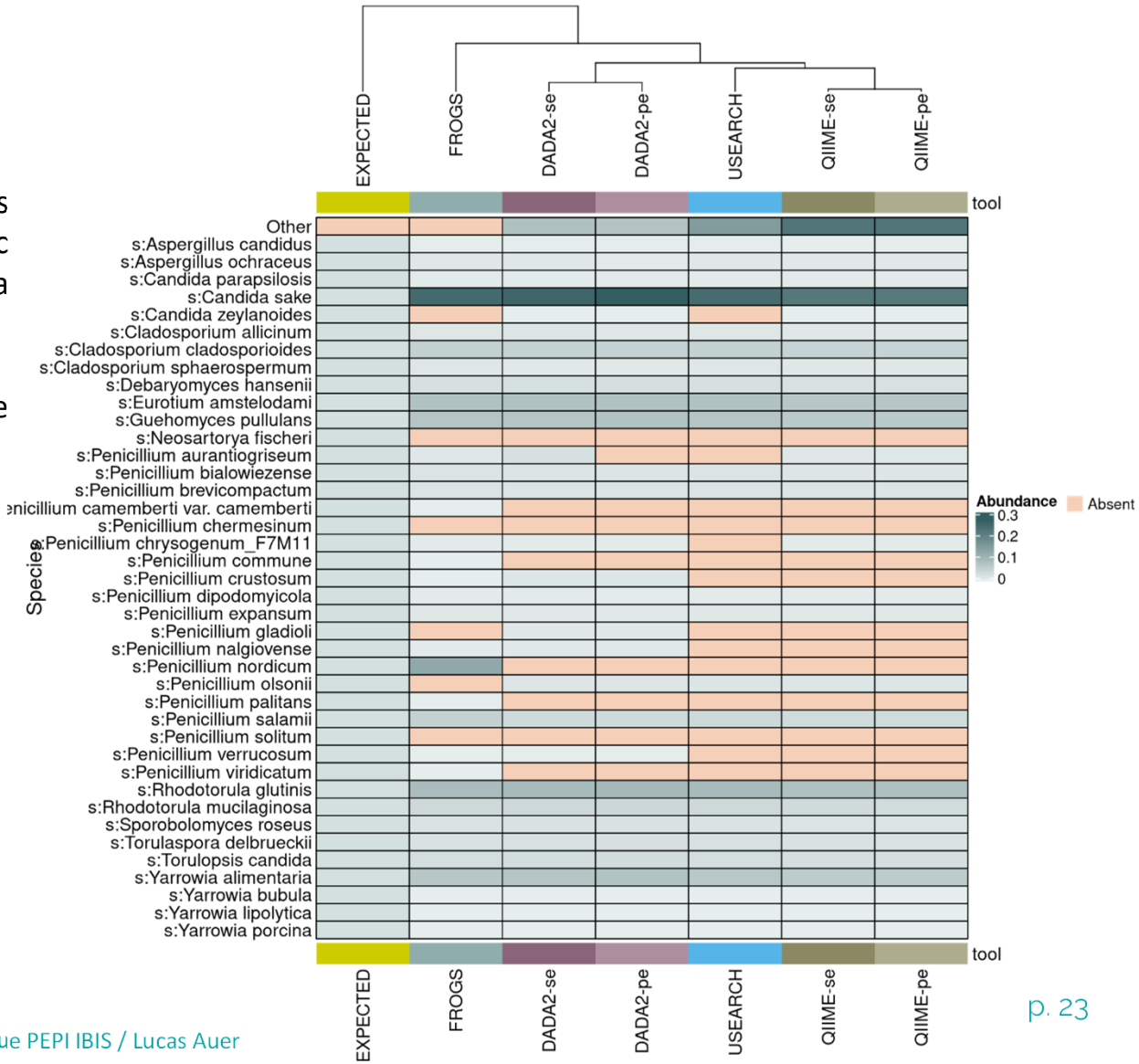
DADA2 est assez performant également, avec des précisions et sensibilités intermédiaires entre **FROGS** et **QIIME** / **USEARCH**

Evaluation – données réelles

Résultats sur **ADN** – ITS1 : abondances

Certaines espèces présentent de très fortes différences d'abondance avec l'attendu, similaires quelle que soit la méthode…

Reflète les biais d'amplification ou de nombre de copie.

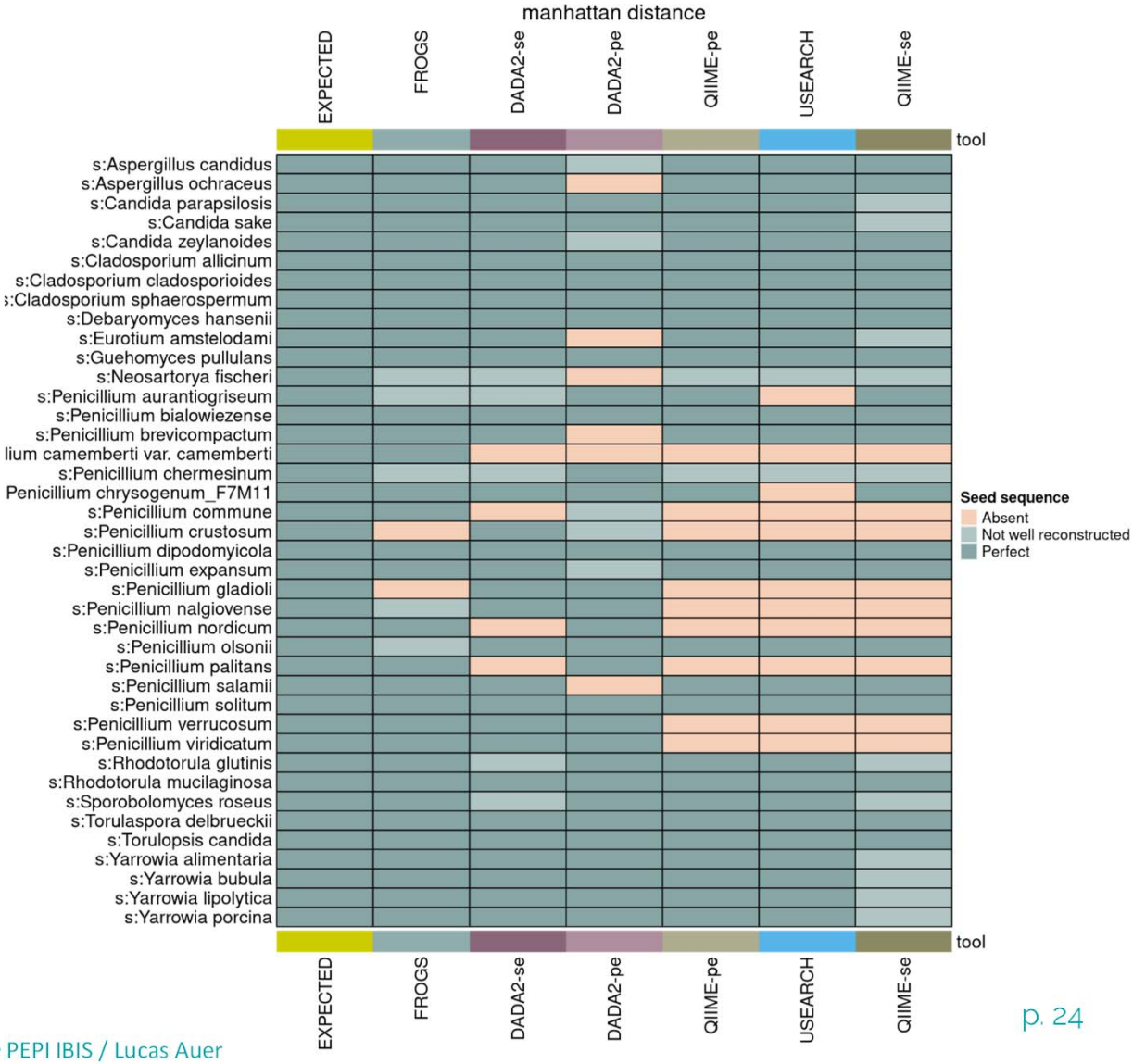


Evaluation – données réelles

Résultats sur PCR – ITS1 : reconstruction des séquences

FROGS et **DADA2-se** sont les outils qui s'en sortent le mieux.

De manière inattendue, **DADA2-pe** rate des séquences par rapport à **DADA2-se** alors que les longueurs attendues ne sont pas très élevées.
(probables erreurs à l'étape de merging)



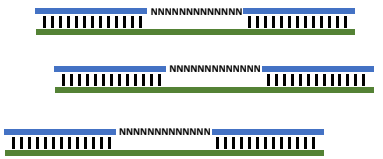
Résumé

Le polymorphisme de longueur des ITS fongiques est un problème pour les études de métabarcoding avec les technologies de séquençage actuelles.

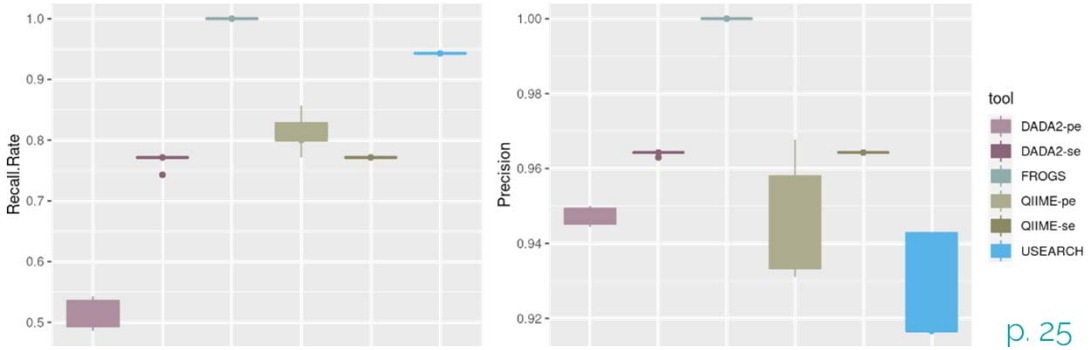
Et utiliser la moitié de l'information fait perdre en spécificité.

FROGS permet de conserver les lectures non jointives et de les annoter avec blastn+ et needleall.

Needleall contre **Base de donnée réduite**
meilleurs hits blastn+ R1 et R2



Cette stratégie permet d'obtenir des sensibilités et précision de 1 malgré des proportions élevées d'espèces à ITS longs.



Remerciements

Les FROGS des ITS :

Maria BERNARD

Olivier RUÉ

Mahendra MARIADASSOU

Géraldine PASCAL

Maria Bernard, Olivier Rué, Mahendra Mariadassou and Géraldine Pascal;
FROGS: a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers, ***Briefings in Bioinformatics*** 2021, 10.1093/bib/bbab318

Contacts :
frogs-support@inrae.fr
geraldine.pascal@inrae.fr