Rétrospective FROGS 2015–2025

Géraldine Pascal, Maria Bernard, Olivier Rué, Mahendra Mariadassou, Frédéric Escudié, Lucas Auer, Gabryelle Agoutin

Géraldine Pascal, Maria Bernard, Olivier Rué, Mahendra Mariadassou, Frédéric Escudié, et al.. Rétrospective FROGS 2015–2025. Symposium FROGSDays, INRAE Toulouse, Dec 2025, Castanet Tolosan, France. ⟨hal-05492401⟩

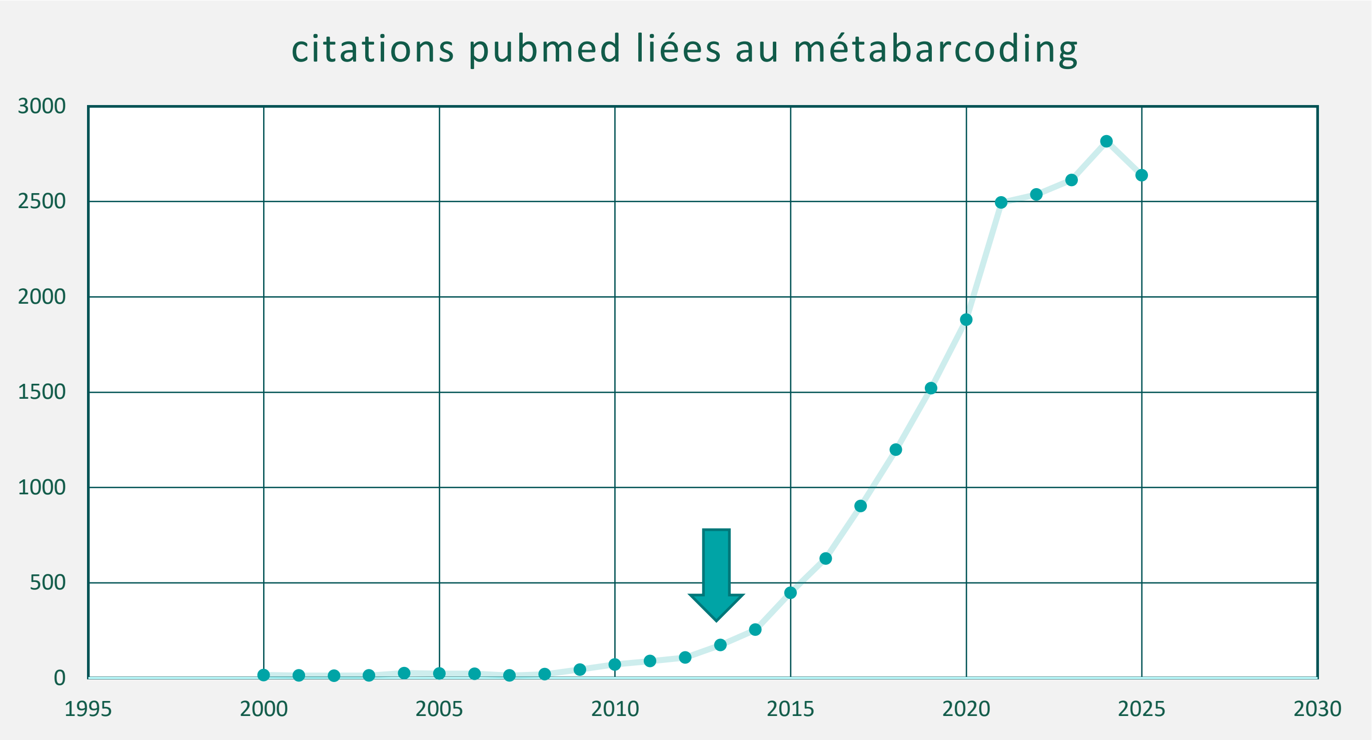
HAL Id: hal-05492401
https://hal.inrae.fr/hal-05492401v1

Submitted on 4 Feb 2026

Rétrospective FROGS 2015–2025

Maria Bernard, Olivier Rué, Mahendra Mariadassou, Frédéric Escudié, Lucas Auer, Gabryelle Agoutin, Géraldine PASCAL

Pourquoi FROGS ? Le point de départ



Pourquoi FROGS ? Le point de départ

454 Roche

- Avant 2015, c'était la référence historique (surtout pour le 16S).
- Longs reads (400–700 pb) mais :
- très cher,
- volumes faibles,
- erreurs dans les homopolymères.

Illumina MiSeq

- 2×250 pb
- très bonne mais : chute forte après 250–280 pb.
- chimères fréquentes
- Volume important

Pourquoi FROGS ? Le point de départ

QIIME

- Basé sur des OTUs (97% ou 99%)
- scripts éparpillés, faible transparence des paramètres, dépendances multiples et difficiles à installer.

MOTHUR

- Très robuste et très complet
- Communauté active
- analyse lente sur de gros volumes, lignes de commande, basé sur des OTUs 97%

USEARCH

- Basé sur des OTUs (97% ou 99%)
- n'était pas open-source, limité à 32 bits → bridait les gros jeux de données

Pourquoi FROGS ? Le point de départ

- Analyses basées quasi exclusivement sur des OTUs 97%.

Limites méthodologiques

- Installation souvent difficile : dépendances,
- versions Python ou R incompatibles,
- pas d'interface graphique

Limites techniques

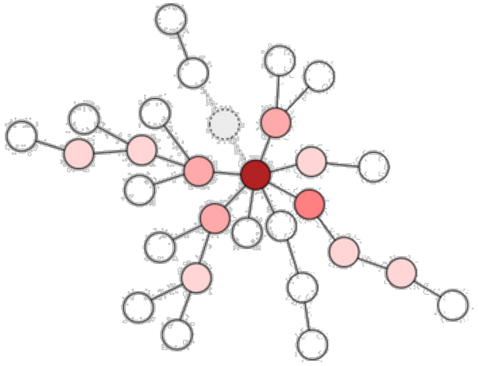
- Outils puissants mais pas pédagogiques.
- Documentation lacunaire.
- Gros besoin d'harmoniser les pratiques.

Limites côté utilisateurs

Les objectifs initiaux (2015)

Accessibilité pour les non-spécialistes

Interface graphique, plateforme web, ergonomique, sans installation



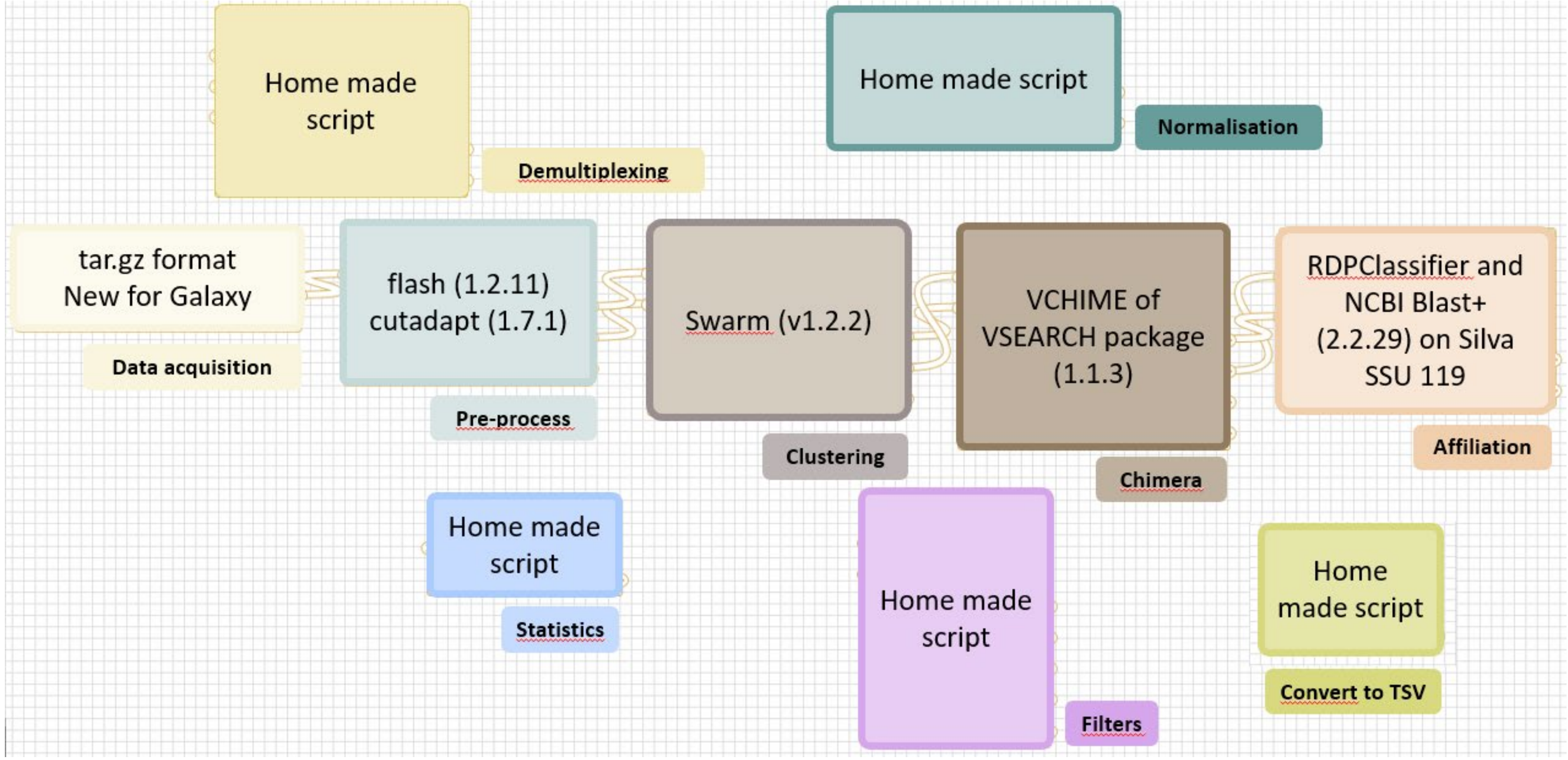
Pas d'OTUs à 97%, adoption du clustering haute résolution avec Swarm

Affiliation à l'espèce/ Multi-affiliations

Standardisation, paramètres transparents, codes libre

Former, expliquer

L’architecture en 2015



Home made script

Demultiplexing

Home made script

Normalisation

tar.gz format New for Galaxy

Data acquisition

flash (1.2.11) cutadapt (1.7.1)

Pre-process

Swarm (v1.2.2)

Clustering

VCHIME of VSEARCH package (1.1.3)

Chimera

RDPClassifier and NCBI Blast+ (2.2.29) on Silva SSU 119

Affiliation

Home made script

Statistics

Home made script

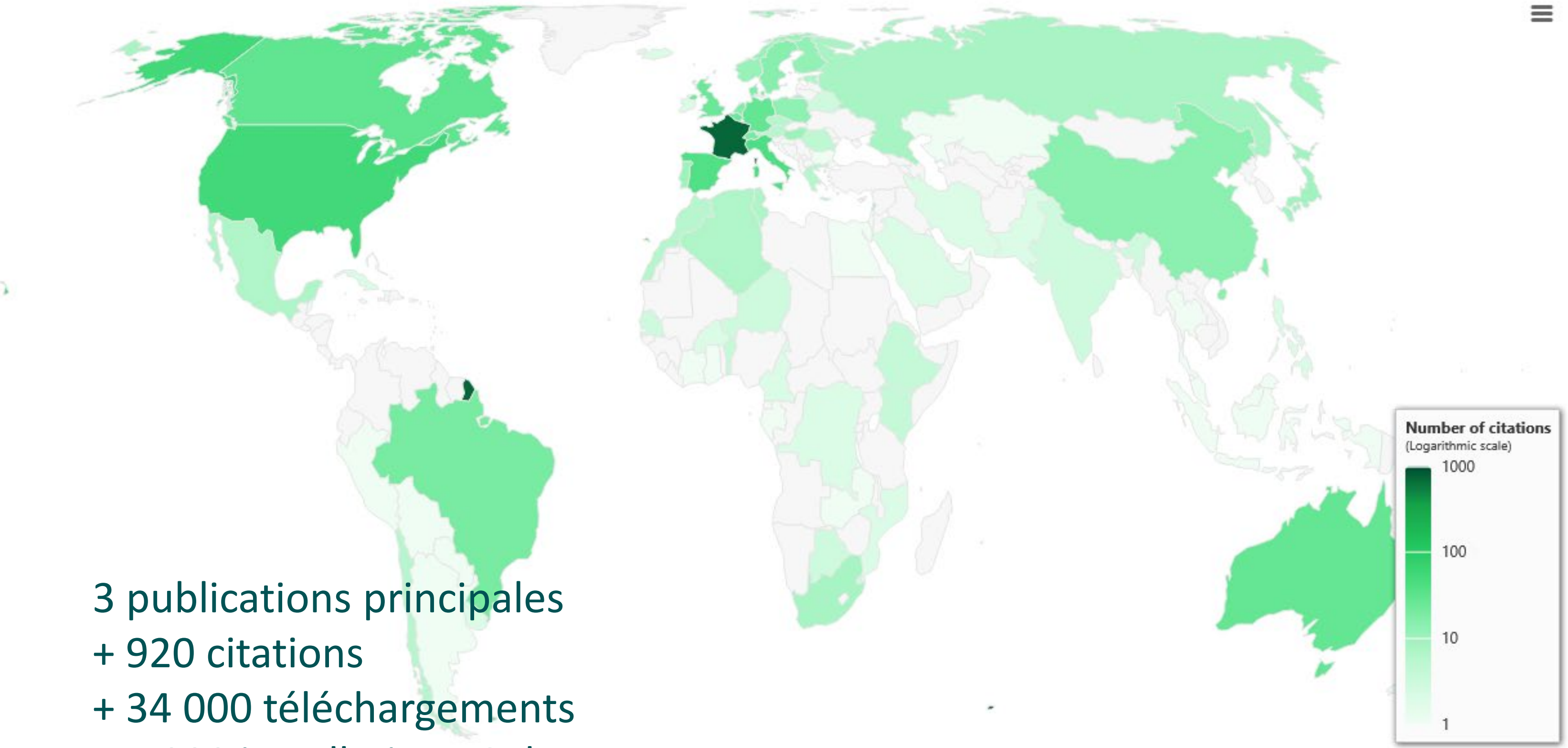
Filters

Home made script

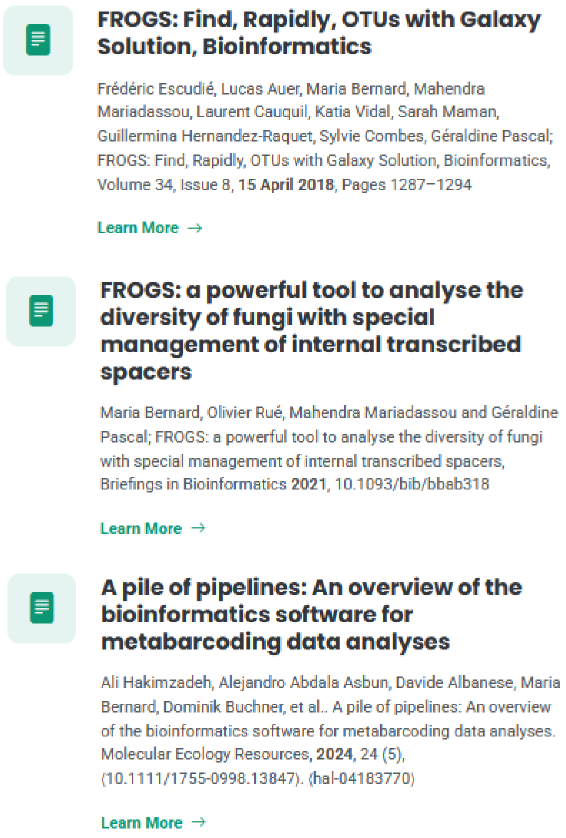
Convert to TSV

Pour données 454 et Illumina 16S et 18S – 6 banques de données.

Diffusion et adoption



3 publications principales
+ 920 citations
+ 34 000 téléchargements
+ 1 000 installations Galaxy



FROGS: Find, Rapidly, OTUs with Galaxy Solution, Bioinformatics

Frédéric Escudié, Lucas Auer, Maria Bernard, Mahendra Mariadassou, Laurent Cauquil, Katia Vidal, Sarah Maman, Guillermina Hernandez-Raquet, Sylvie Combes, Géraldine Pascal; FROGS: Find, Rapidly, OTUs with Galaxy Solution, Bioinformatics, Volume 34, Issue 8, **15 April 2018**, Pages 1287–1294

Learn More →

FROGS: a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers

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

Learn More →

A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses

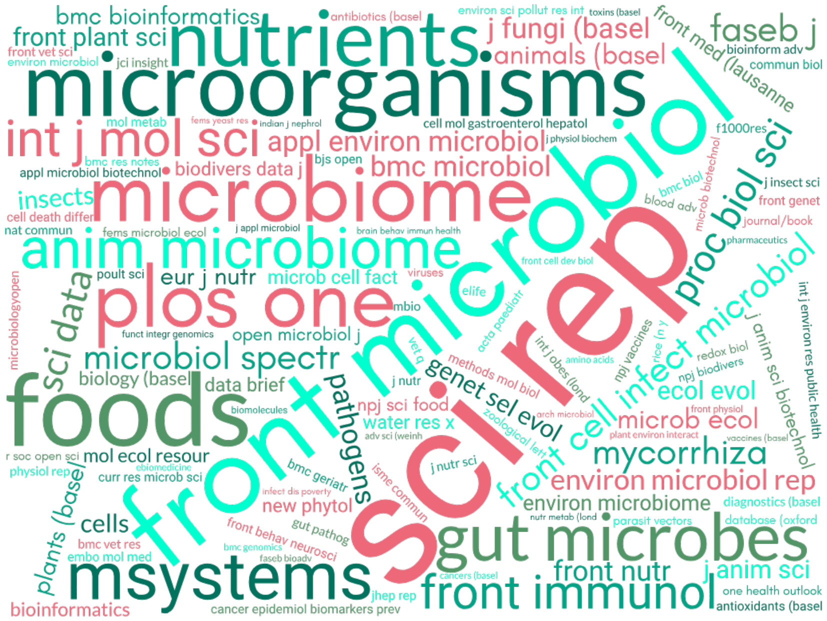
Ali Hakimzadeh, Alejandro Abdala Asbun, Davide Albanese, Maria Bernard, Dominik Buchner, et al.. A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses. Molecular Ecology Resources, **2024**, 24 (5), (10.1111/1755-0998.13847). (hal-04183770)

Learn More →

Diffusion et adoption

Extraction des principaux mot clés des articles citant FROGS

Élevage / Animaux domestiques	piglets, sows, broilers, chickens, hens, rabbits, dogs, cattle, cow, sheep, ducks, geese
Animaux de laboratoire	mice, mouse, rats, hamsters
Plantes & Fungi	plant, populus, tree, forest, fungi, yeast
Humains / Santé	human, patients, infants, disease, infection, immune, fecal, intestinal
Environnement / alimentation	soil, fermentation, cheese, milk, grappe, wine, extreme environments, bioengineering and bioprocesses
Milieux aquatiques	oyster, protiste, ocean, trout, aquaculture



Les grandes évolutions du pipeline

V2.0.0
2017

Intégration des statistiques de bases (α-, β-, …)

V3.0.0
2019

Prise en charge des ITS (méthode, outil, banques)

+ 3.1: packaging sous conda

V4.0.0
2022

Outils d'inférence fonctionnelle

V4.1.0
2023

Prise en charge des lectures long-reads

+

OTUs → ASVs

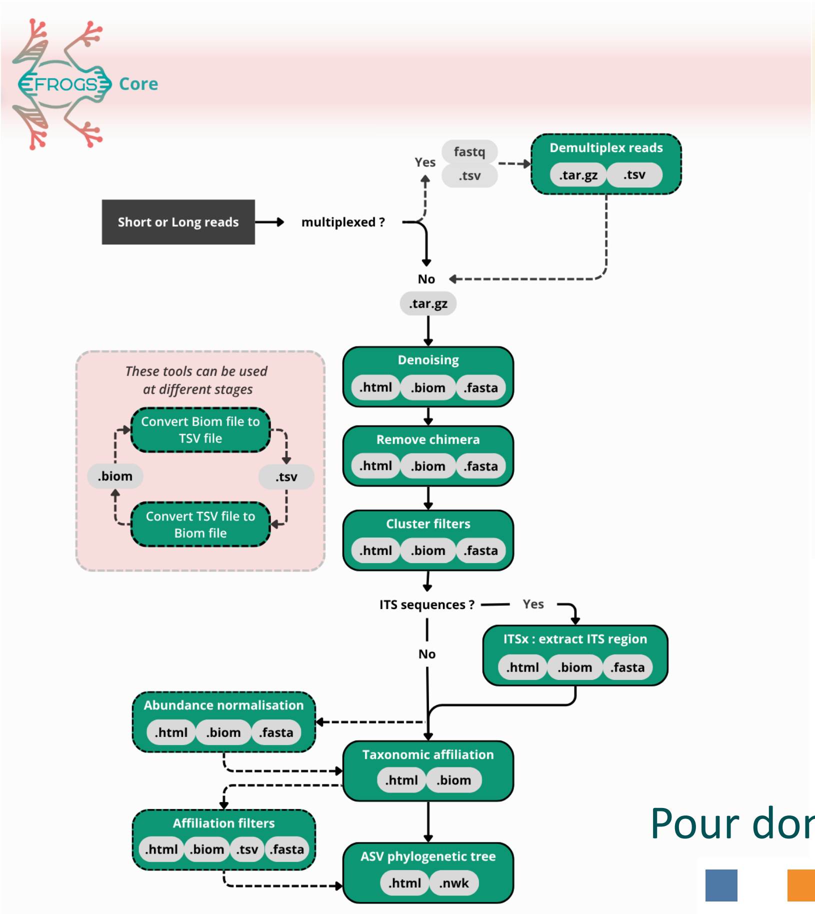
V5.0.0
2024

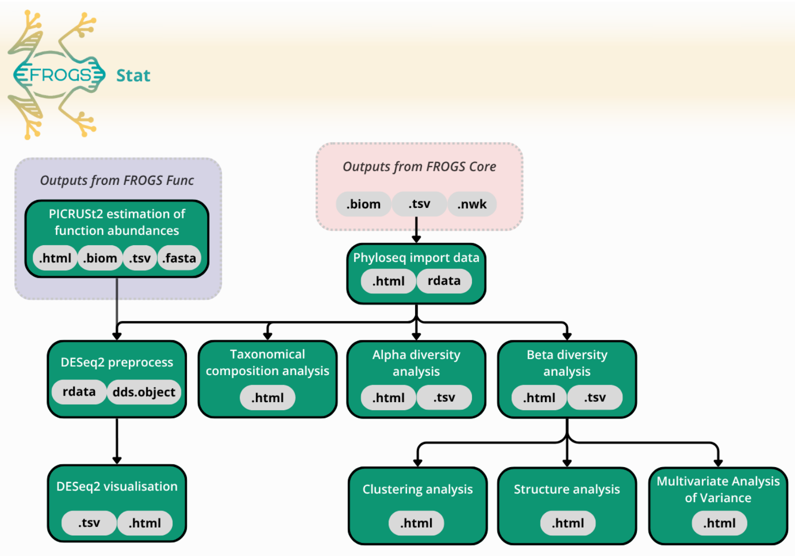
Construction des ASV par DADA2

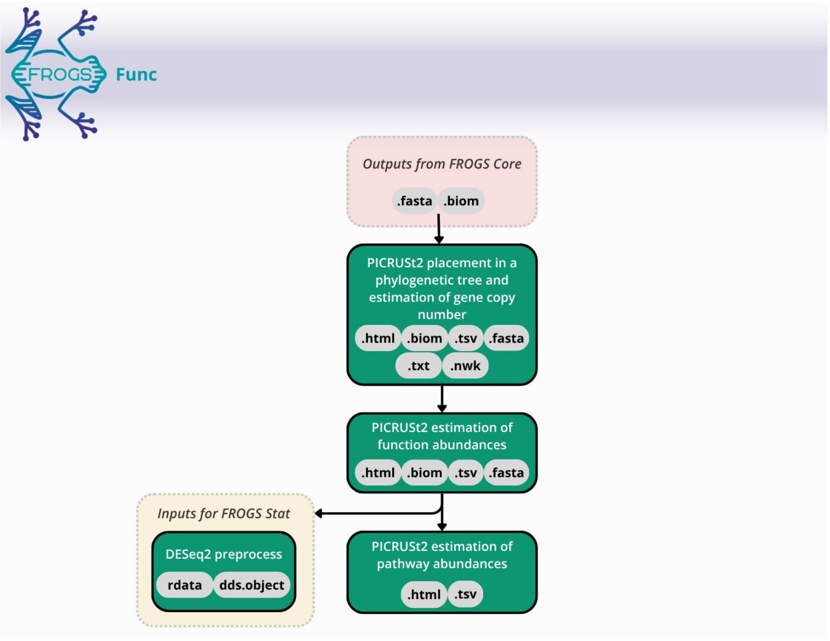
+

Totale restructuration

L’architecture en 2025





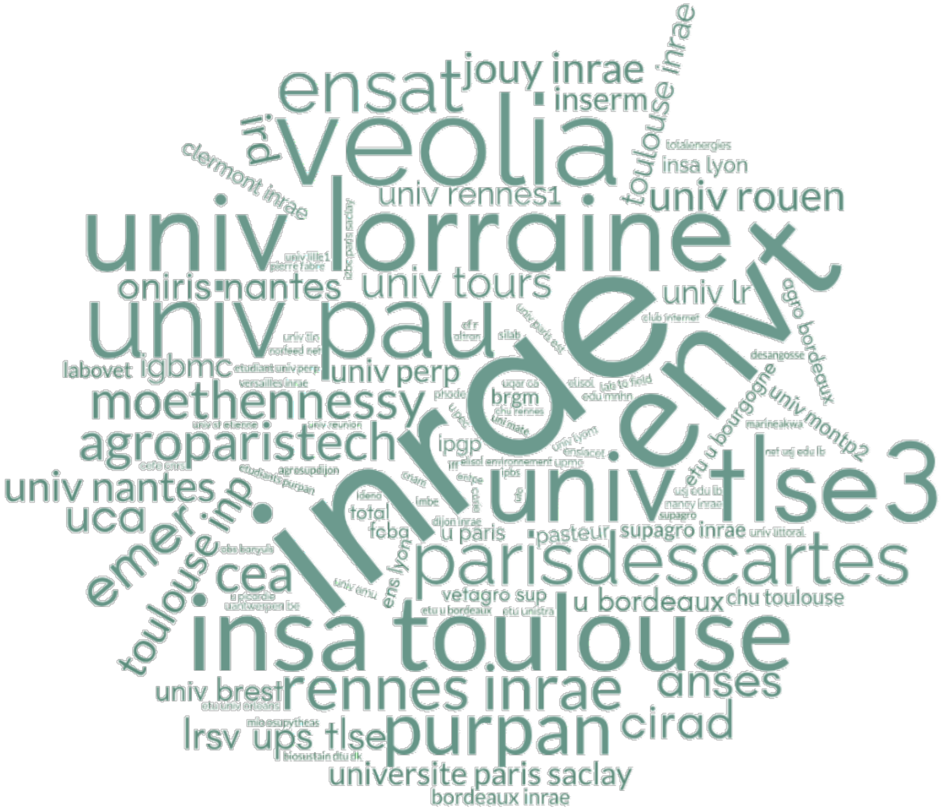


Pour données 454, Illumina, AVITI, PACBIO, ONT – 17 marqueurs - 127 banques de données.

🟦 12S 🟧 16S 🟥 16S-ITS-23S 🟩 18S 🟩 23S 🟨 28S 🟪 COI 🟥 EF1,18S 🟫 ITS ⬜ ITS2 🟦 SSU-ITS-LSU 🟧 gyrb 🟥 matK 🟩 rbcL 🟩 rpoB 🟨 trnH 🟪 trnl

Les formations FROGS

- Formations principales
  - En présentiel à Jouy en Josas – 1 fois/an
  - À distanciel – 2 à 3 fois /an
  - À façon pour groupes spécifiques
  - Les mardis de la grenouille
- + 400 formés
- Tuto et documents de formation en ligne



La communauté FROGS (le vrai moteur)

Retours d’expérience lors des formations

Retours supports utilisateurs

Retours de nos labo

frogs-support@inrae.fr

Retours issues github

Vers quoi allons-nous ?

- Méthodologie spéciale long-read
  avec construction de pseudo-souches

- Amélioration des FROGS Func pour les ITS
  avec bases de données de références augmentées

- Meilleure gestion du support utilisateurs
  avec mise en place de tickets

- Augmenter notre visibilité avec Galaxy Europe

Remerciements

Developers

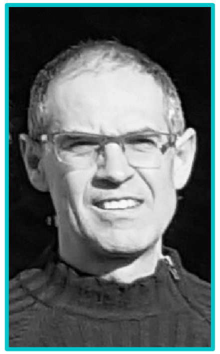

Frédéric ESCUDIÉ


Maria BERNARD


OLIVIER RUÉ


VINCENT DARBOT

Biology experts


Laurent CAUQUIL


Sylvie COMBES

Galaxy support


Patrice DÉHAIS


Sarah MAMAN


Lucas AUER


Mahendra MARIADASSOU

Statistical experts


Gabryelle AGOUTIN

Databank manager


Maelle POMIÈS

Web master


Géraldine PASCAL

Coordinator


FROGS

Simulated datasets, for testing FROGS' Accuracy

- 500 species, covering all bacterial phyla
- Power Law distribution of the species abundances
- Error rate calibrated with real sequencing runs
- 20% chimeras
- 10 samples of 100 000 sequences each (1M sequences)

Simulated dataset : 1M sequences

↓

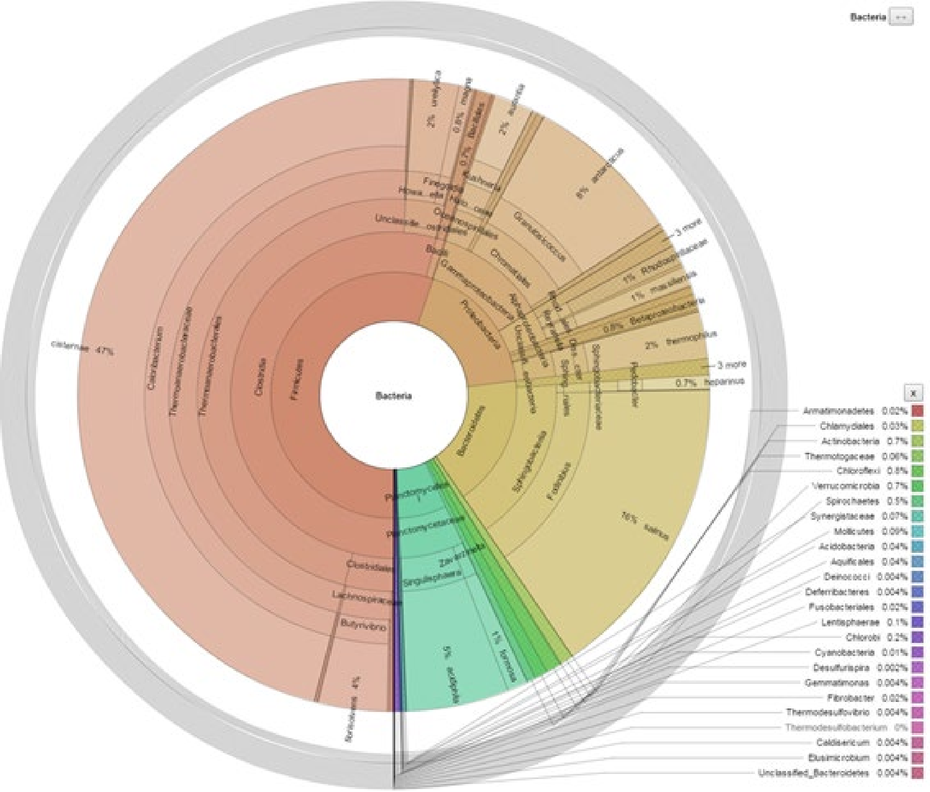
SWARM : 109 000 clusters

↓

VSEARCH: 21 000 clusters

↓

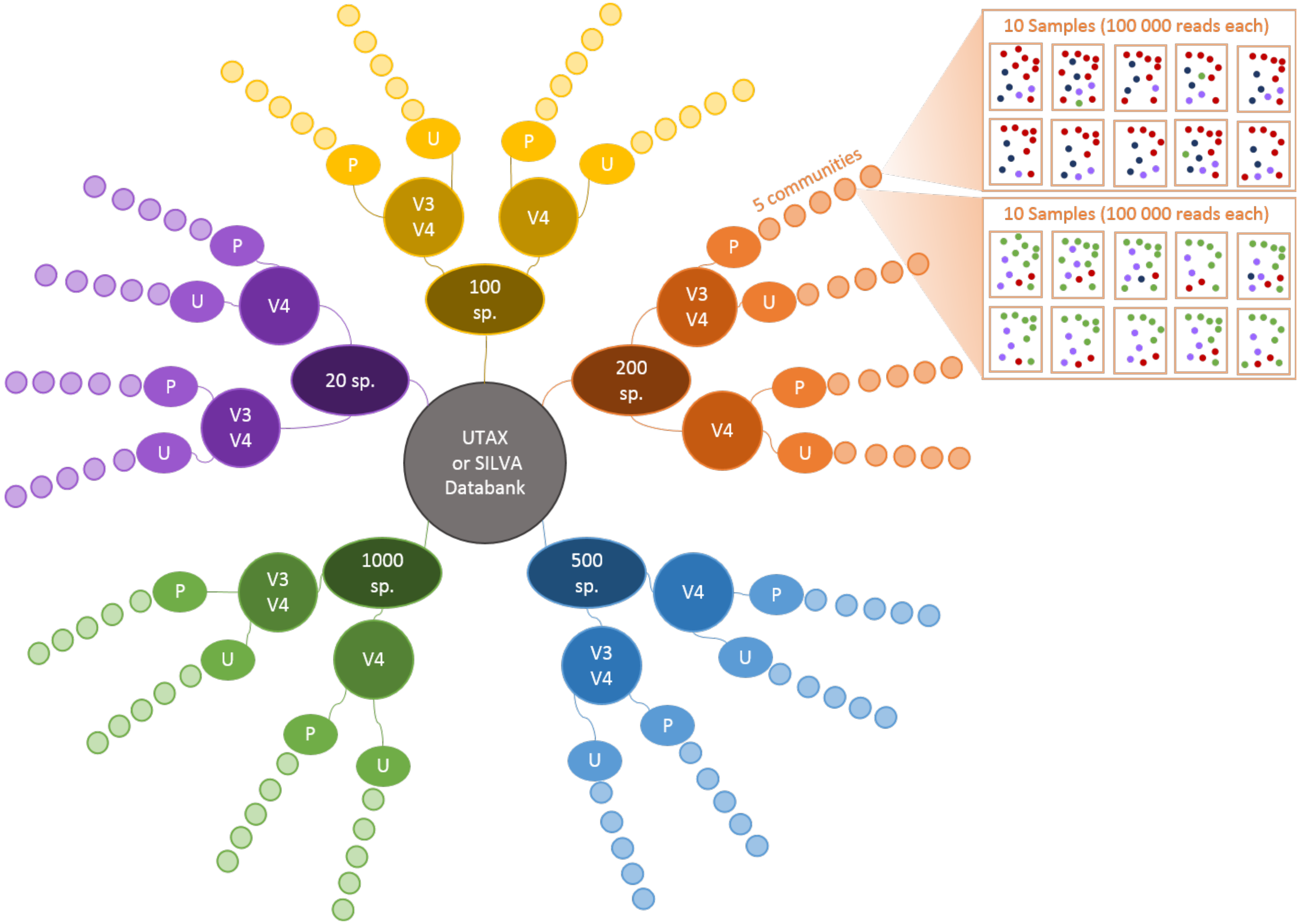
filters : 0.005% **505 OTUs**



FROGS’ Accuracy

- 1.10^{+8} synthetic sequences were treated with FROGS, UPARSE and MOTHUR, QIIME, with their guidelines, to compare their performances
- 20, 100, 200, 500 or 1000 different species
- power law or a uniform distribution
- 5 to 20% of chimera

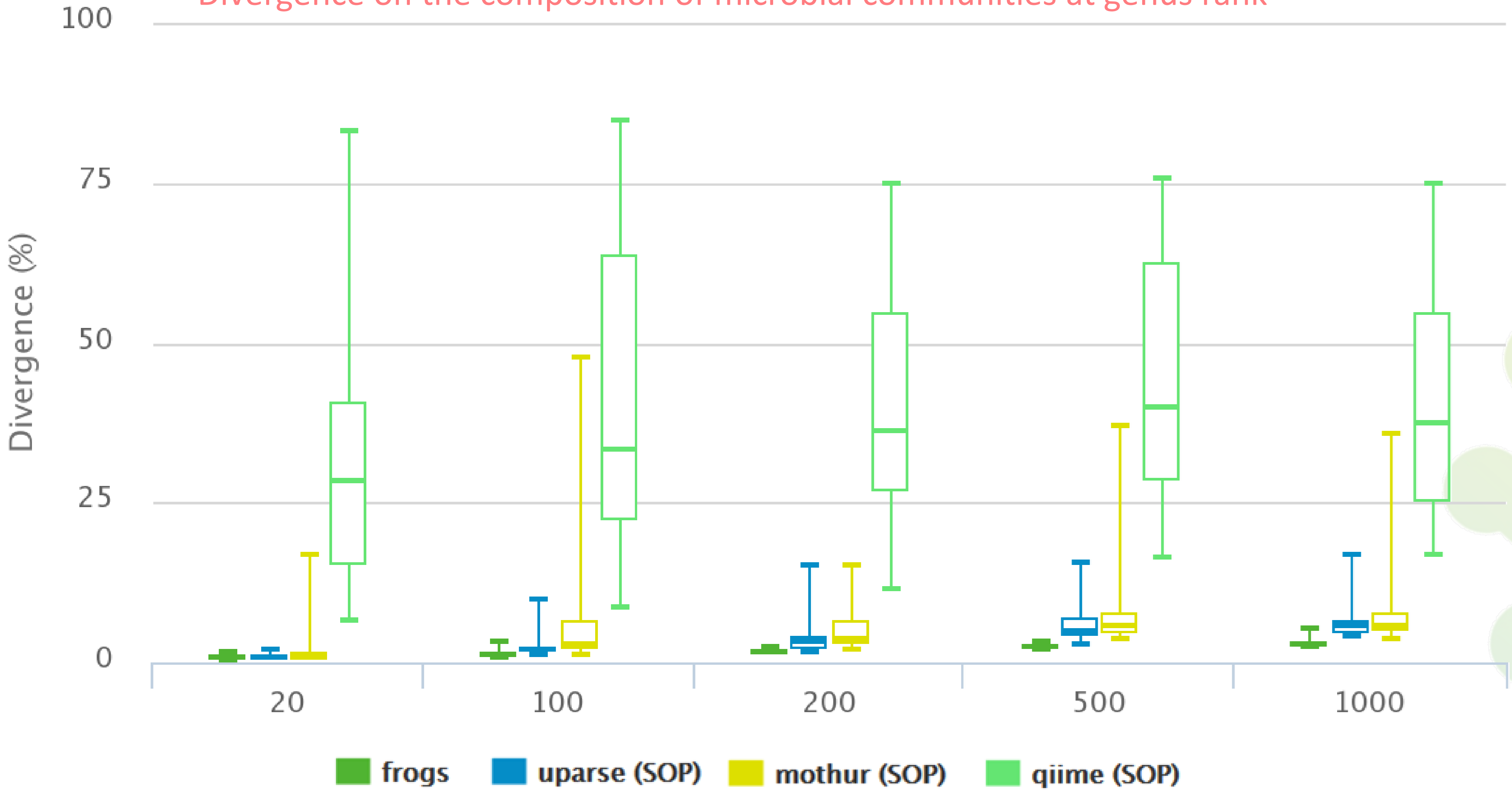
→ Divergence on the composition of microbial communities at the different taxonomic ranks



V3V4 Power Law

Affiliations divergence

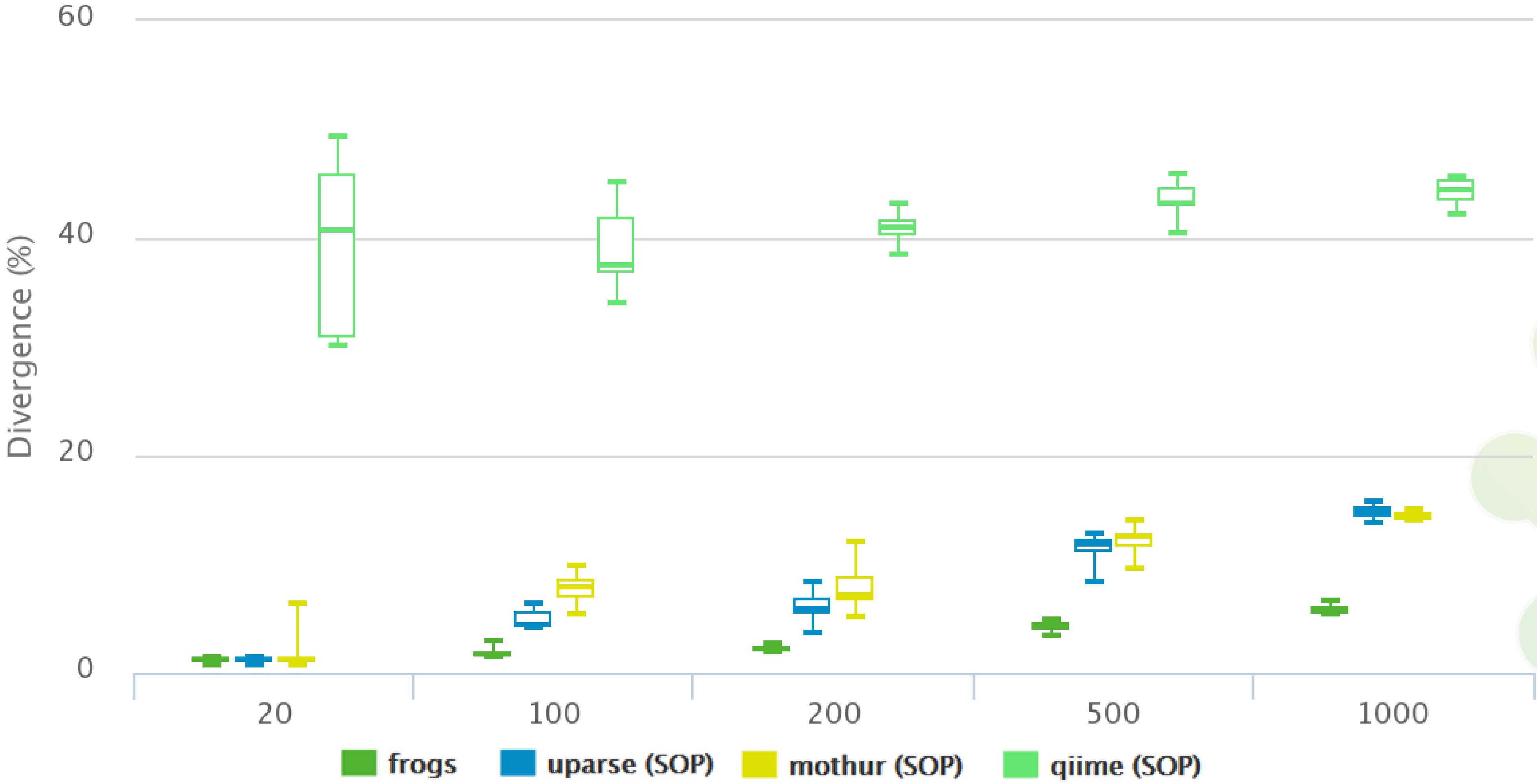
Divergence on the composition of microbial communities at genus rank



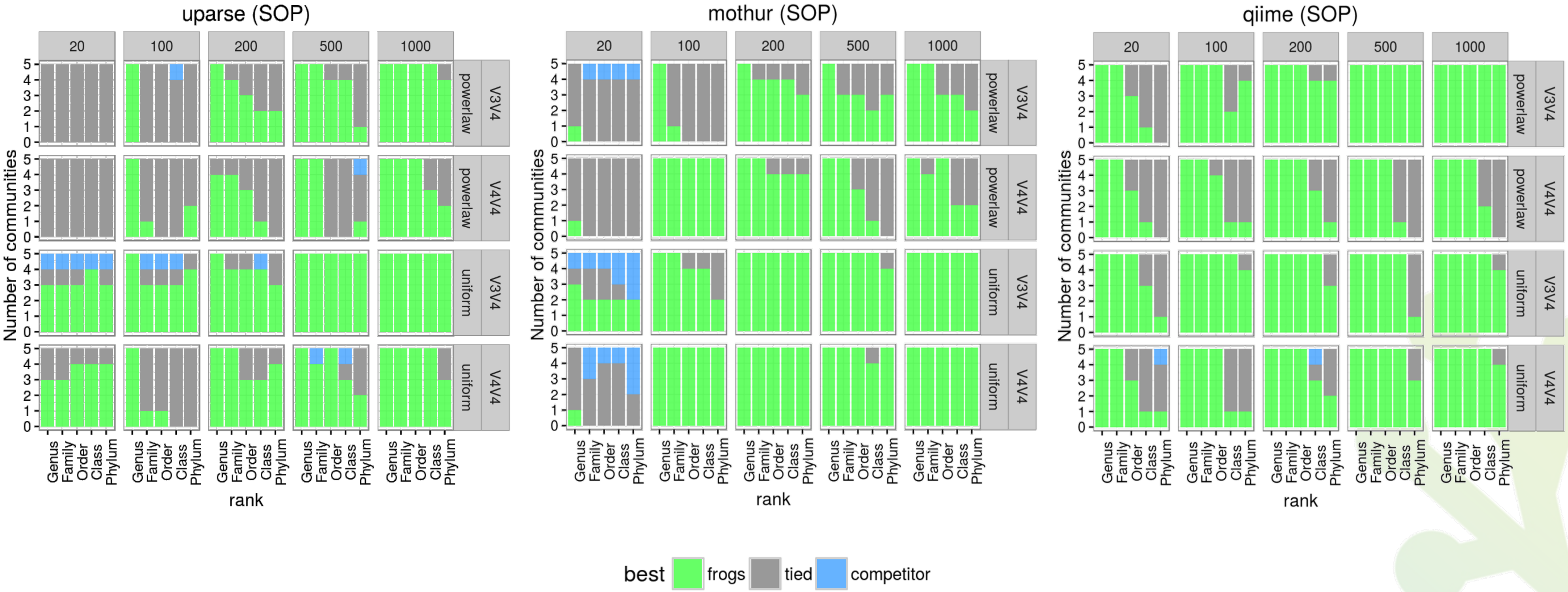
V3V4 Uniform

Affiliations divergence

Divergence on the composition of microbial communities at genus rank



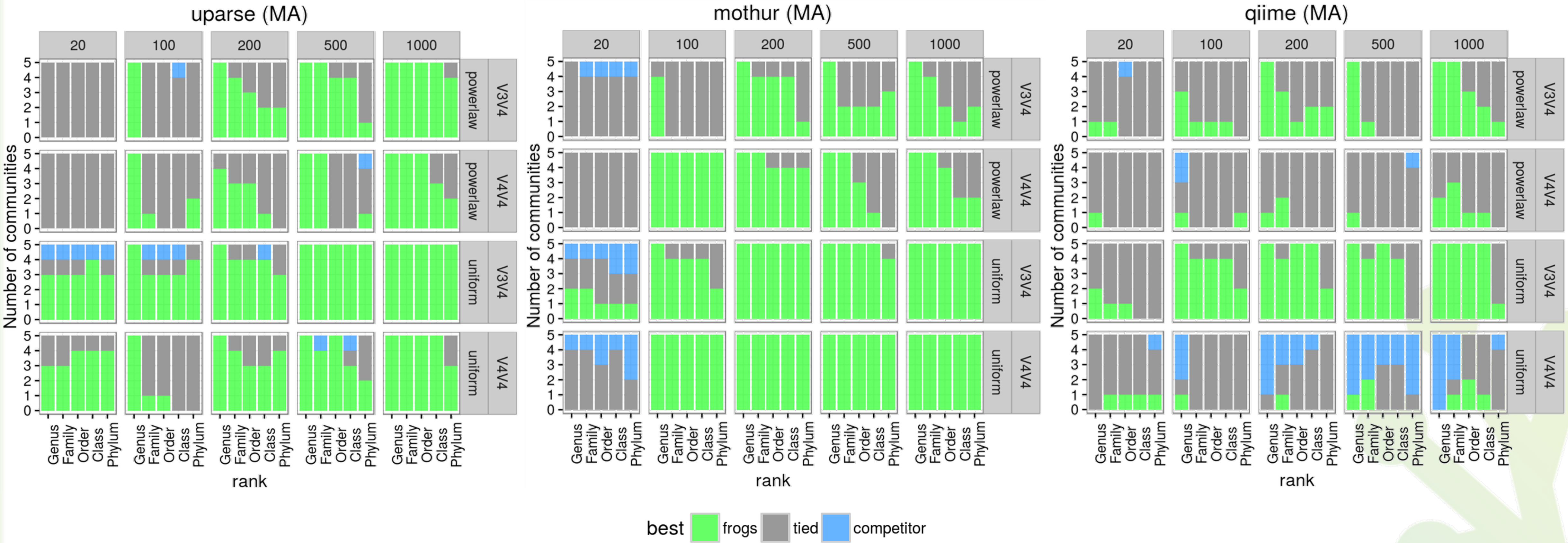
The results of non-parametric paired tests (signed rank test) of Affiliation divergence on simulated data from UTAX



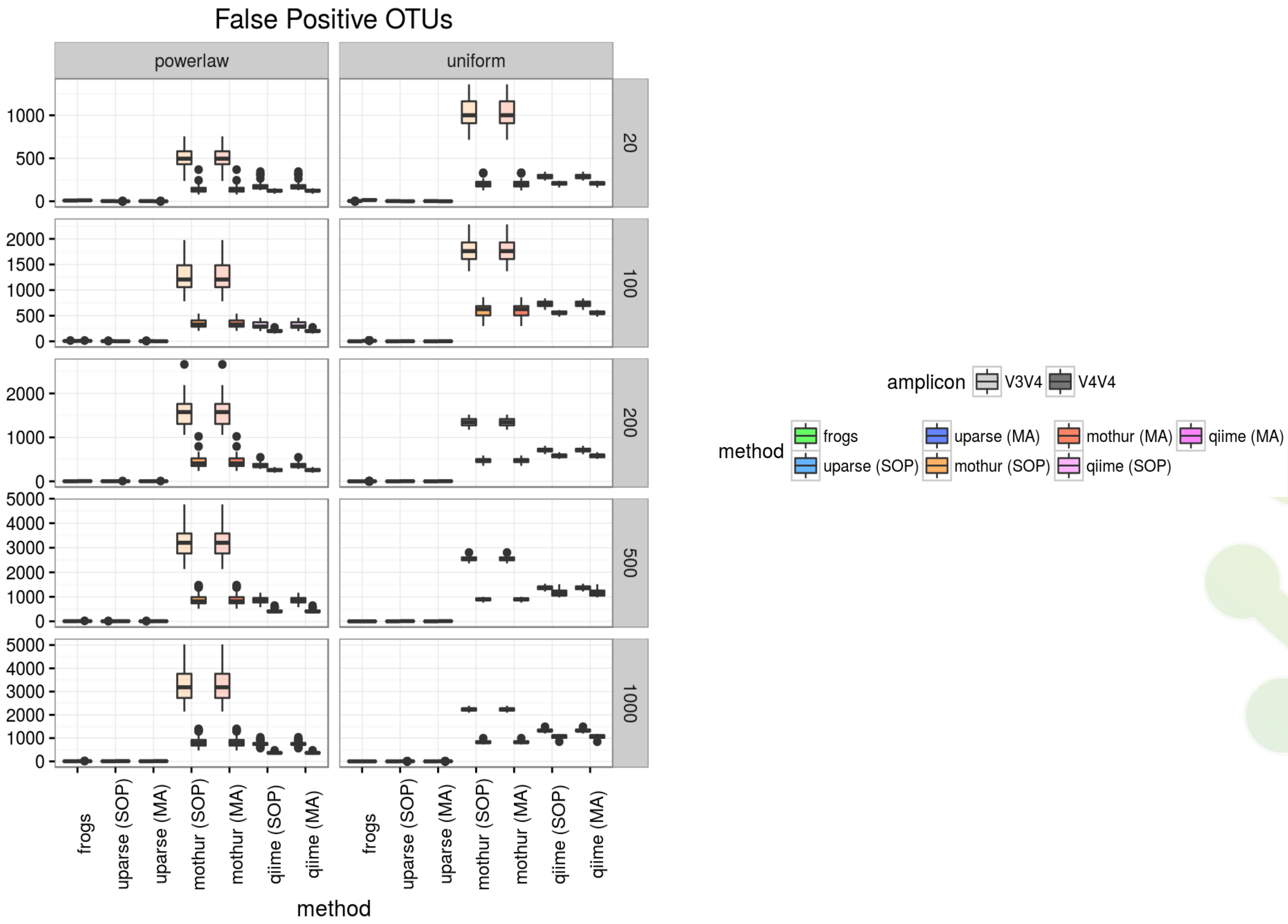
FROGS performs as well as or better than UPARSE, MOTHUR and QIIME in most settings. The only condition in which FROGS does worse than UPARSE and MOTHUR is small community size (20).

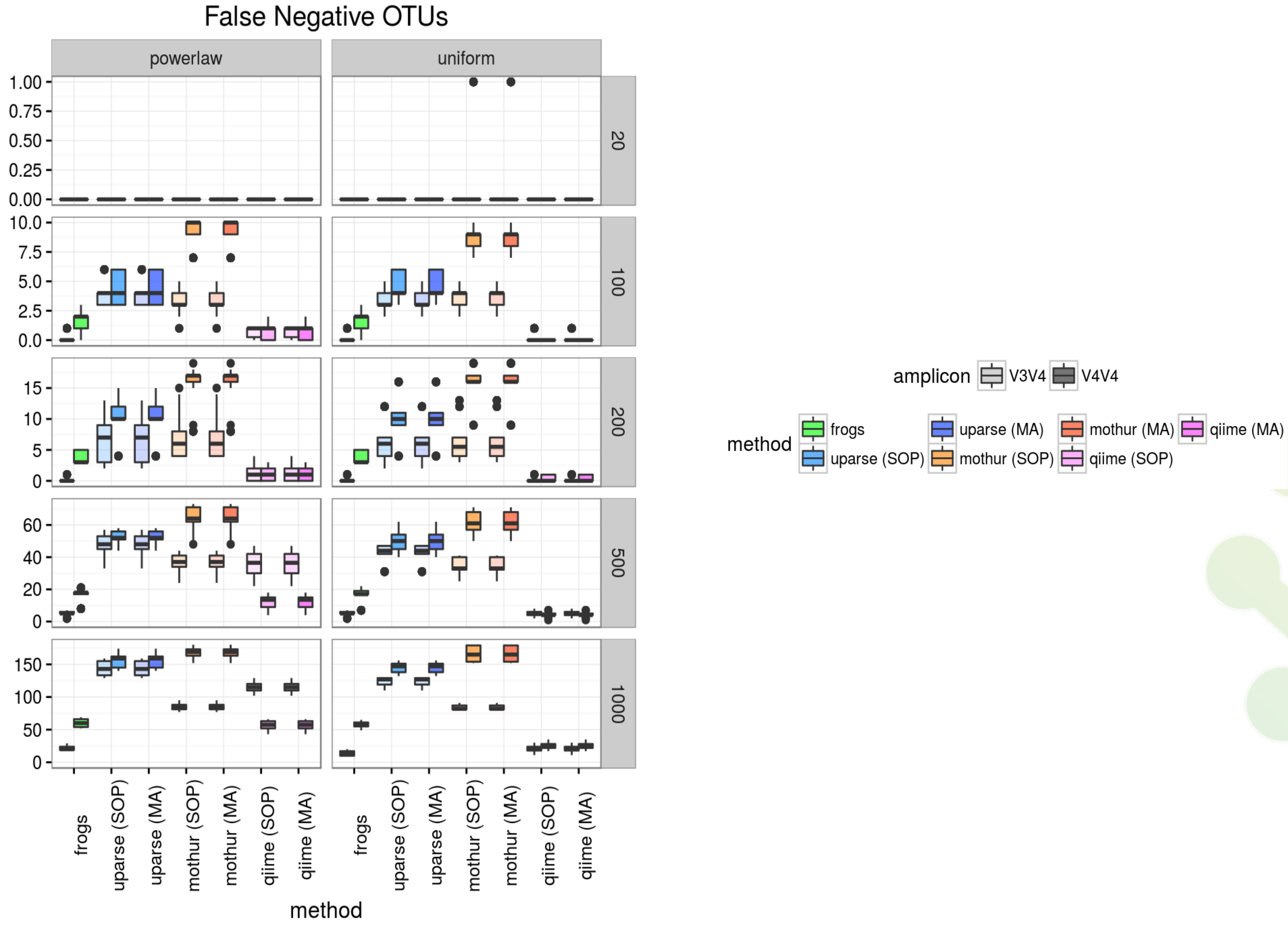
The results of non-parametric paired tests (signed rank test) of Affiliation divergence on simulated data from UTAX

With FROGS multi-Affiliation

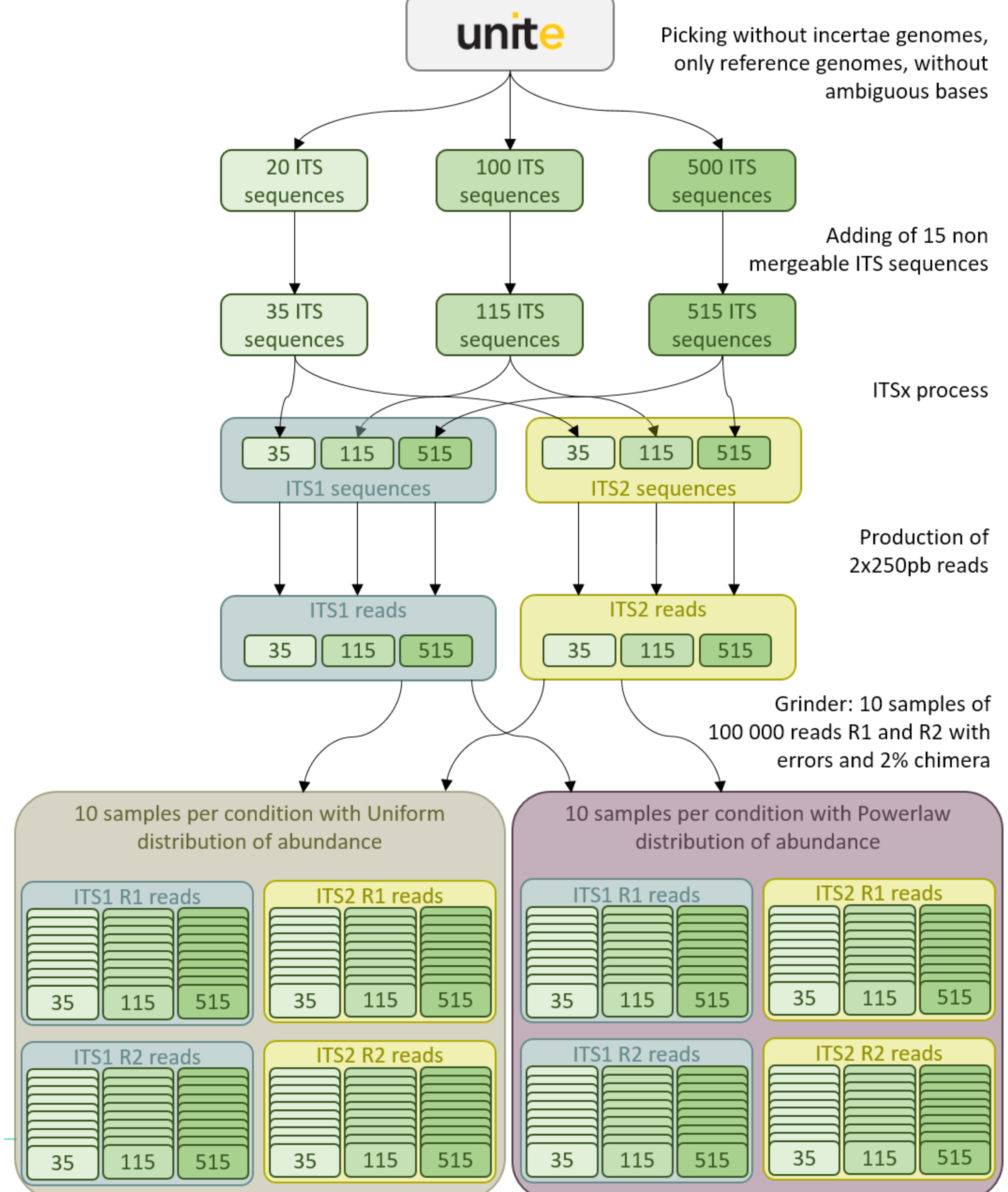


QIIME (MA) with large communities (size > 200) with uniform abundance using the V4 region is better than FROGS. The differences, although significant, are small in that case: less than 2 percentage points in all cases and most marked at the Genus level where the divergences of both FROGS and QIIME (MA) are already quite moderate (6~10%).


False Positive OTUs
powerlaw
uniform
20
100
200
500
1000
1000
500
0
2000
1500
1000
500
0
2000
1000
0
5000
4000
3000
2000
1000
0
5000
4000
3000
2000
1000
0
frogs
uparse (SOP)
uparse (MA)
mothur (SOP)
mothur (MA)
qiime (SOP)
qiime (MA)
method
amplicon
V3V4
V4V4
method
frogs
uparse (SOP)
uparse (MA)
mothur (SOP)
mothur (MA)
qiime (SOP)
qiime (MA)


False Negative OTUs
powerlaw
uniform
20
100
200
500
1000
1.00
0.75
0.50
0.25
0.00
10.0
7.5
5.0
2.5
0.0
15
10
5
0
60
40
20
0
150
100
50
0
frogs
uparse (SOP)
uparse (MA)
mothur (SOP)
mothur (MA)
qiime (SOP)
qiime (MA)
method
amplicon
V3V4
V4V4
method
frogs
uparse (SOP)
uparse (MA)
mothur (SOP)
mothur (MA)
qiime (SOP)
qiime (MA)

Benchmarking ITS



unite

Picking without incertae genomes, only reference genomes, without ambiguous bases

20 ITS sequences

100 ITS sequences

500 ITS sequences

Adding of 15 non mergeable ITS sequences

35 ITS sequences

115 ITS sequences

515 ITS sequences

ITSx process

ITS1 sequences: 35, 115, 515

ITS2 sequences: 35, 115, 515

Production of 2x250pb reads

ITS1 reads: 35, 115, 515

ITS2 reads: 35, 115, 515

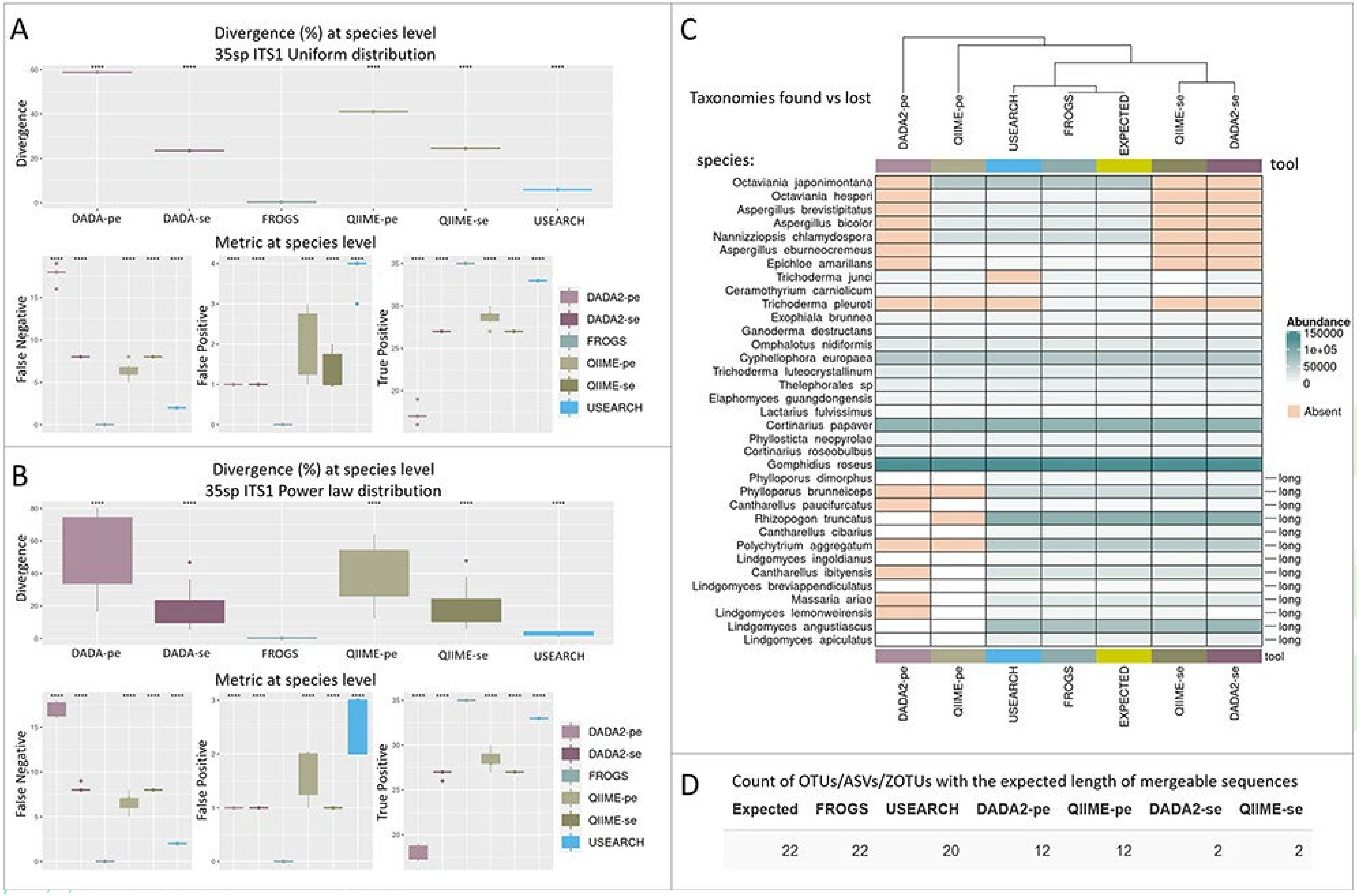
Grinder: 10 samples of 100 000 reads R1 and R2 with errors and 2% chimera

10 samples per condition with Uniform distribution of abundance

- ITS1 R1 reads: 35, 115, 515
- ITS2 R1 reads: 35, 115, 515
- ITS1 R2 reads: 35, 115, 515
- ITS2 R2 reads: 35, 115, 515

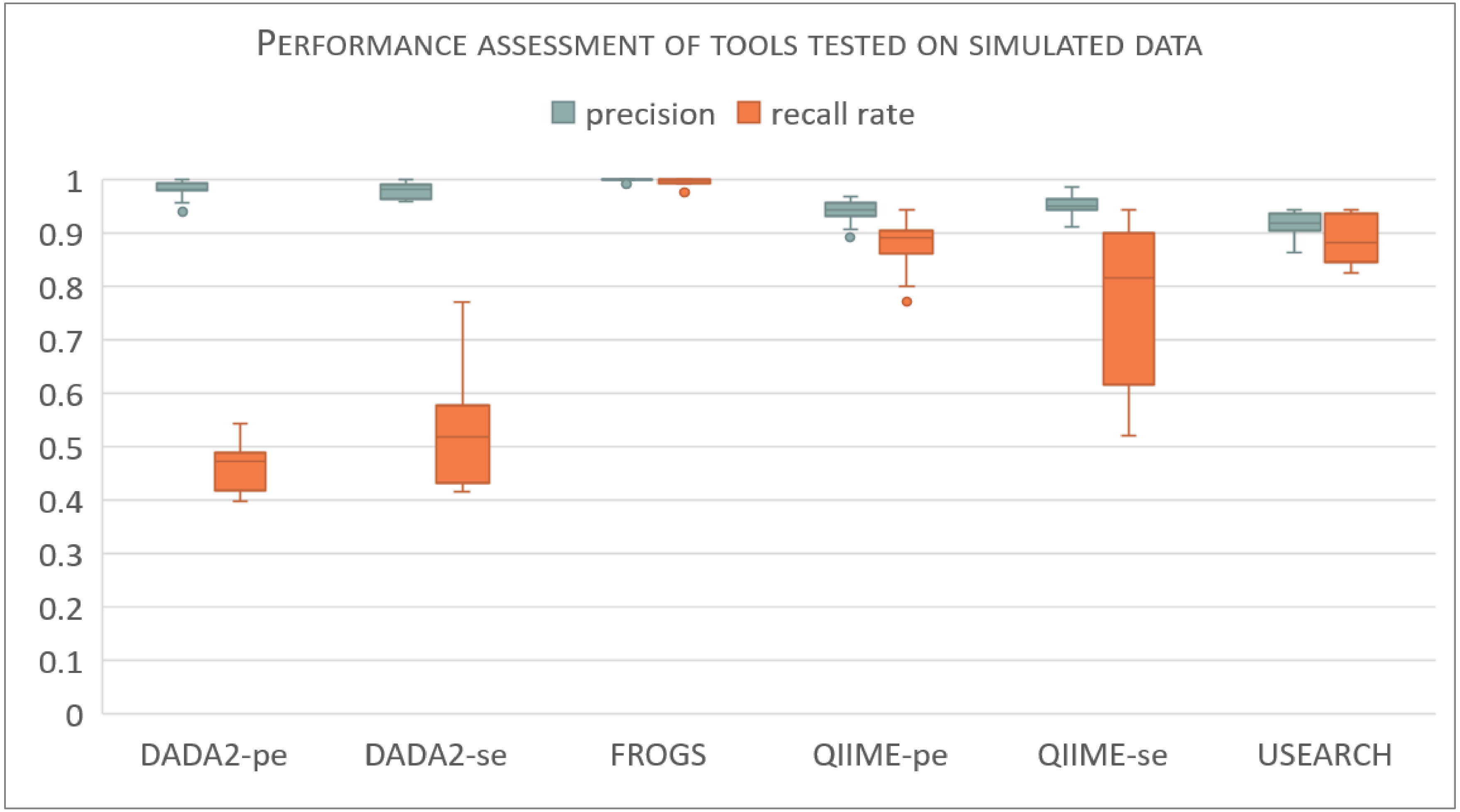
10 samples per condition with Powerlaw distribution of abundance

- ITS1 R1 reads: 35, 115, 515
- ITS2 R1 reads: 35, 115, 515
- ITS1 R2 reads: 35, 115, 515
- ITS2 R2 reads: 35, 115, 515

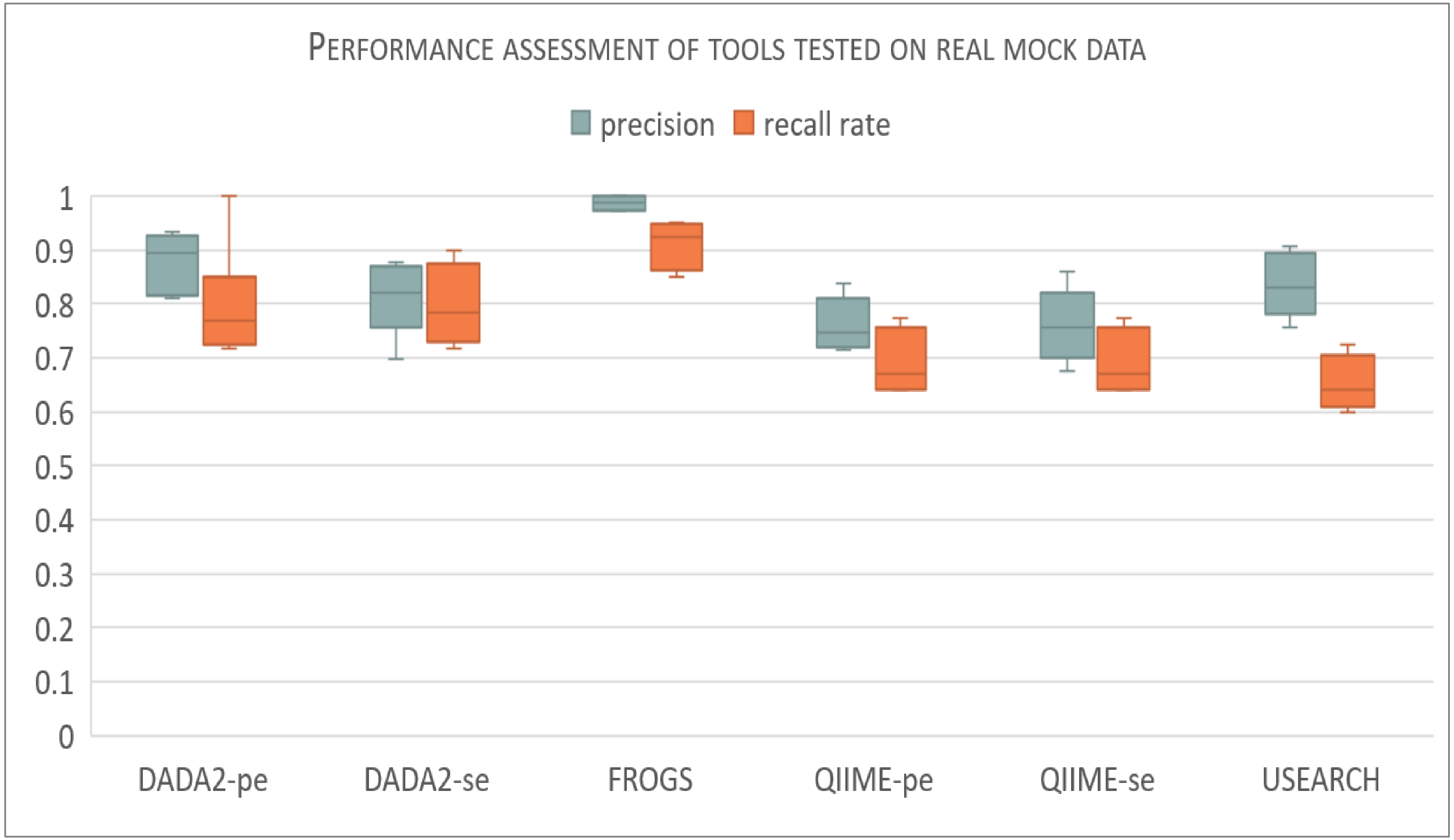


Count of OTUs/ASVs/ZOTUs with the expected length of mergeable sequences

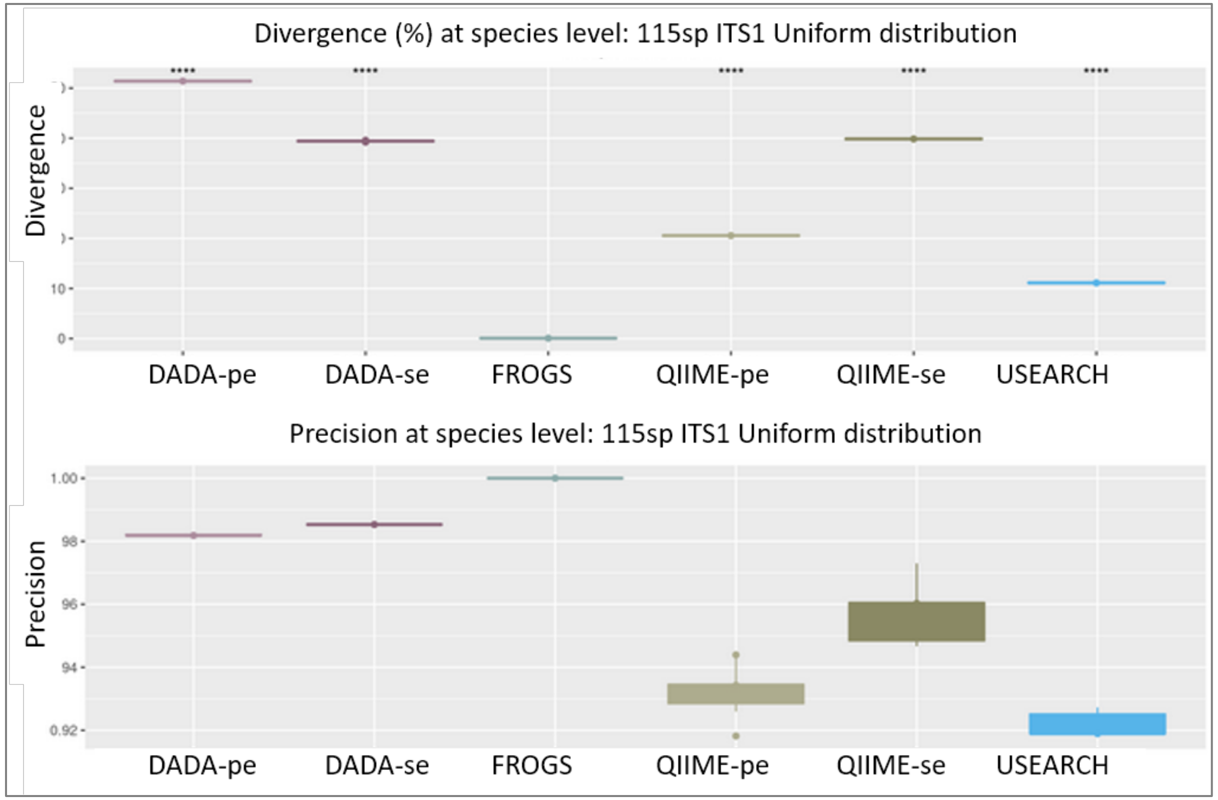
Expected	FROGS	USEARCH	DADA2-pe	QIIME-pe	DADA2-se	QIIME-se
22	22	20	12	12	2	2



Boxplot showing the overall precision (blue) and recall rate (orange) of each of the tools used to process test datasets (all datasets: 35, 115 and 515 species, ITS1 and ITS2, power law and uniform abundance distribution).

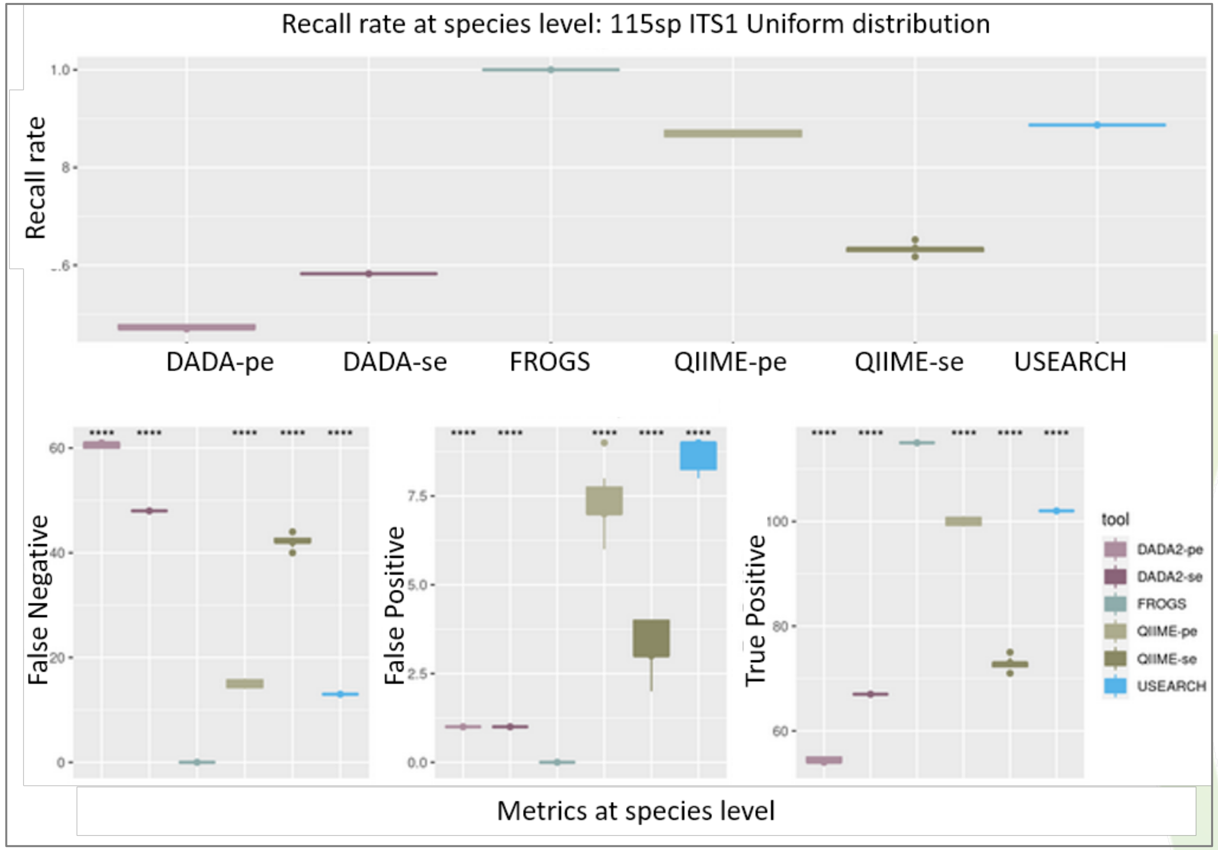


boxplot showing the overall precision (blue) and recall rate (orange) of each of the tools used to process biological datasets (ITS1 and ITS2 datasets with ADN or PCR products).



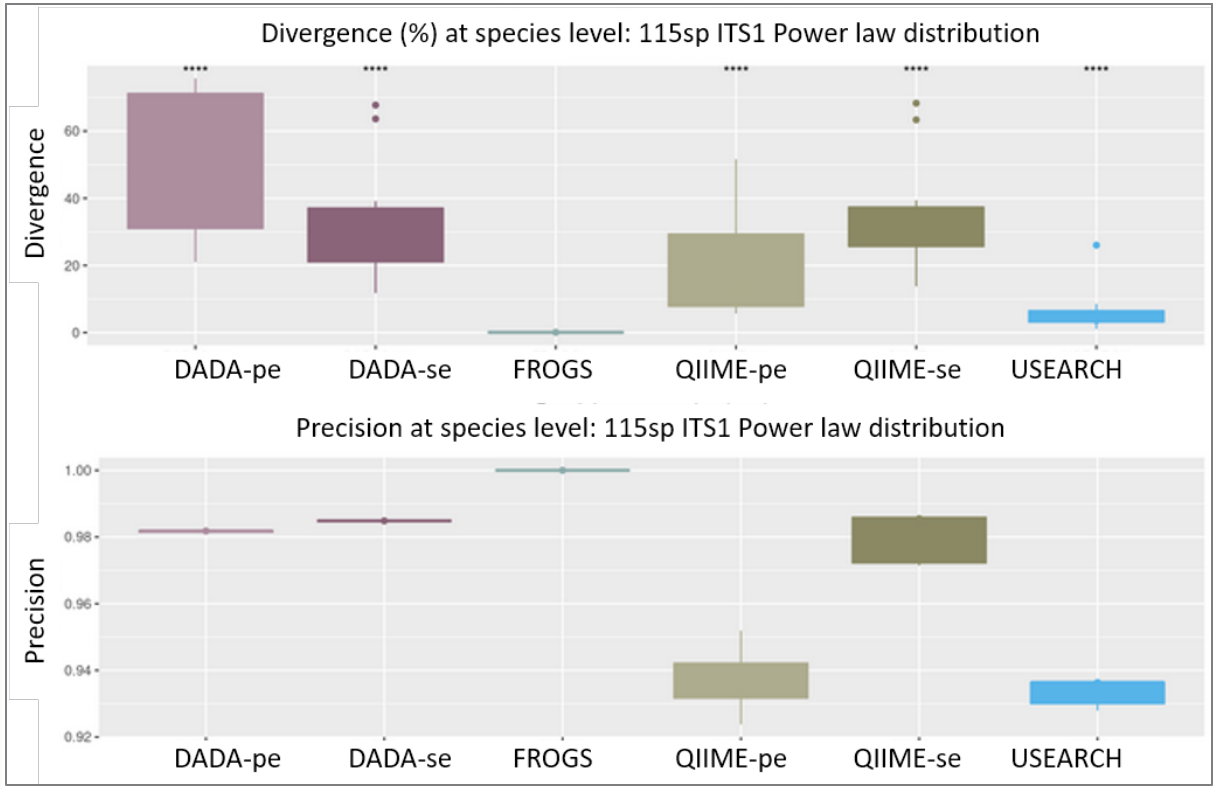
Divergence (%) at species level: 115sp ITS1 Uniform distribution

Precision at species level: 115sp ITS1 Uniform distribution



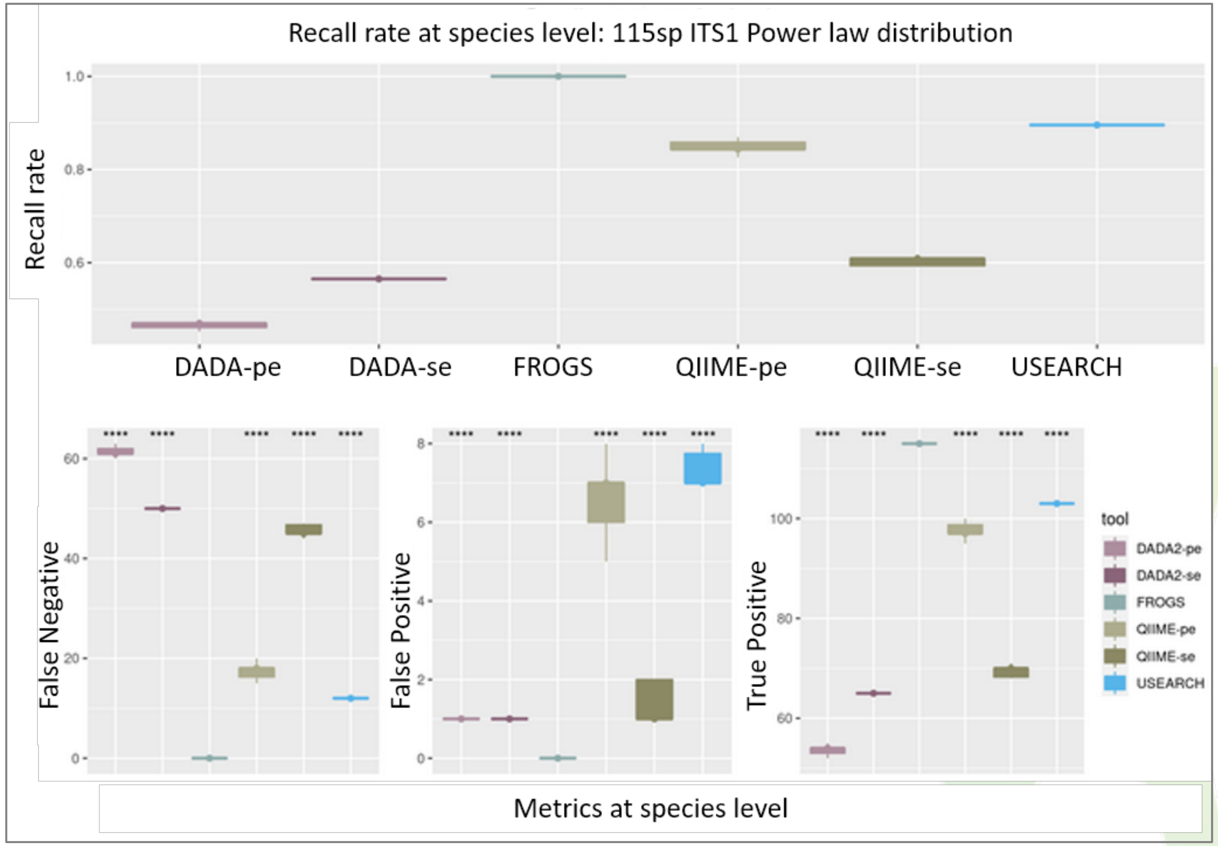
Recall rate at species level: 115sp ITS1 Uniform distribution

Metrics at species level



Divergence (%) at species level: 115sp ITS1 Power law distribution

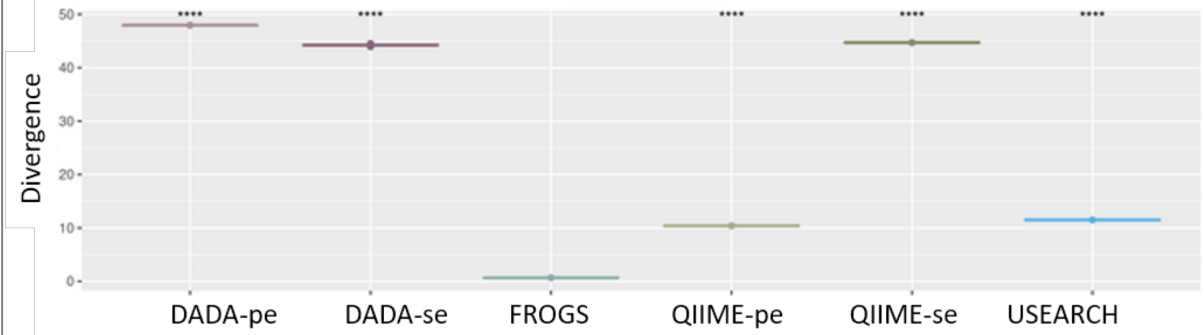
Precision at species level: 115sp ITS1 Power law distribution



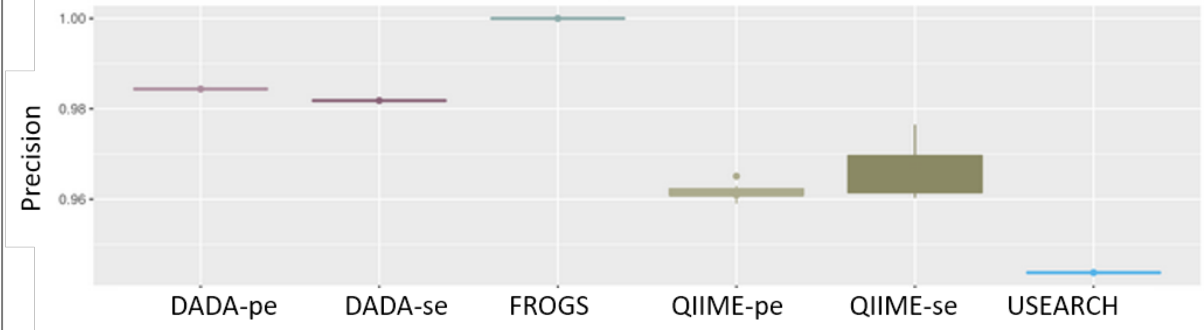
Recall rate at species level: 115sp ITS1 Power law distribution

Metrics at species level

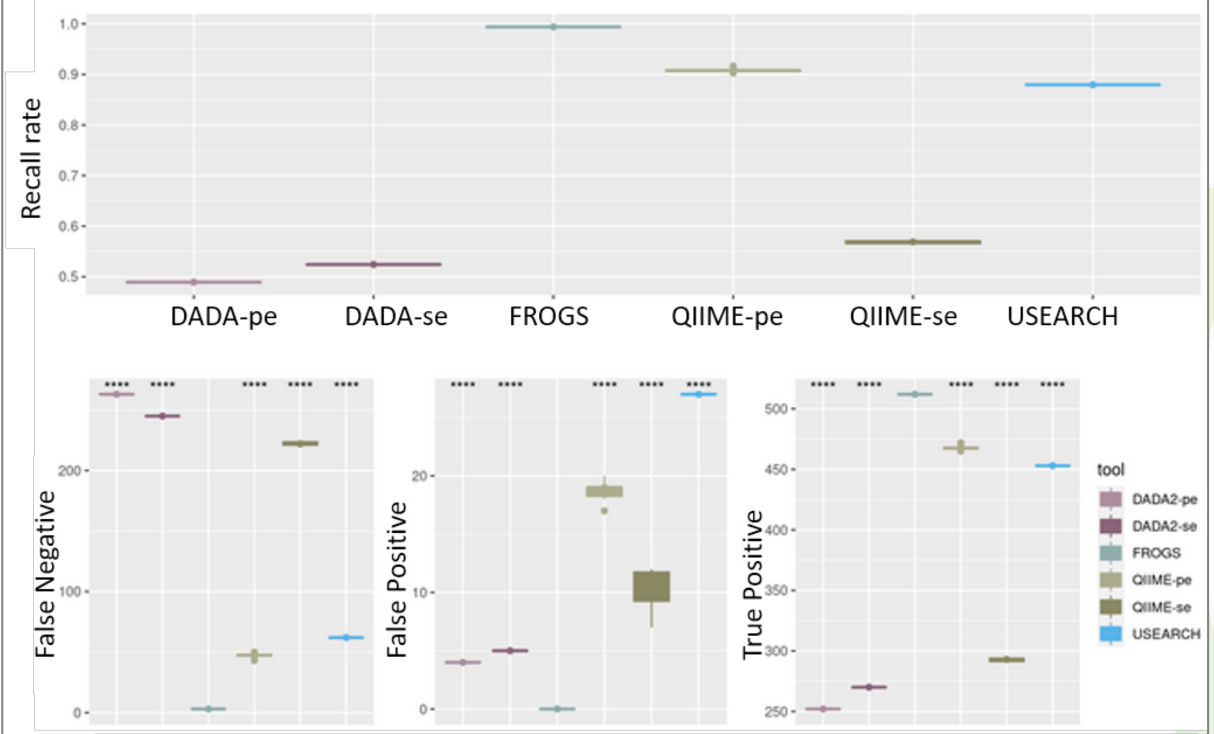
Divergence (%) at species level: 515sp ITS1 Uniform distribution



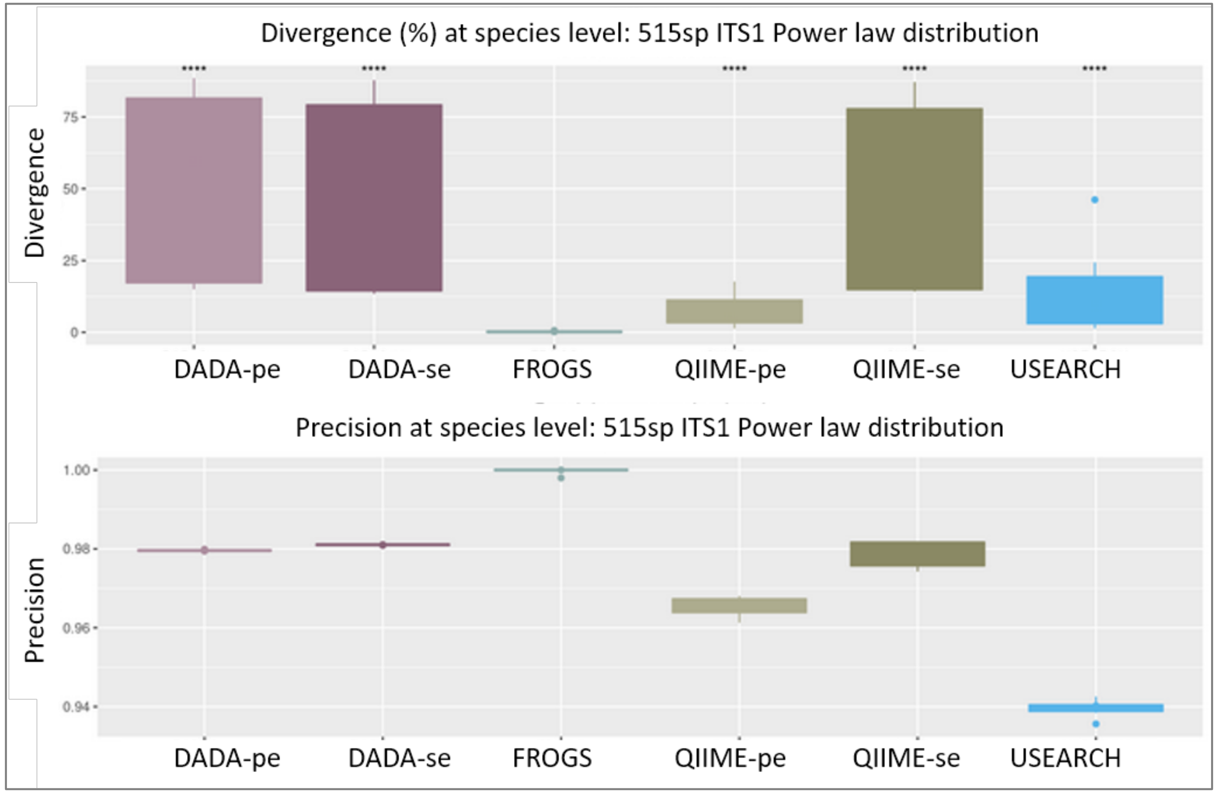
Precision at species level: 515sp ITS1 Uniform distribution

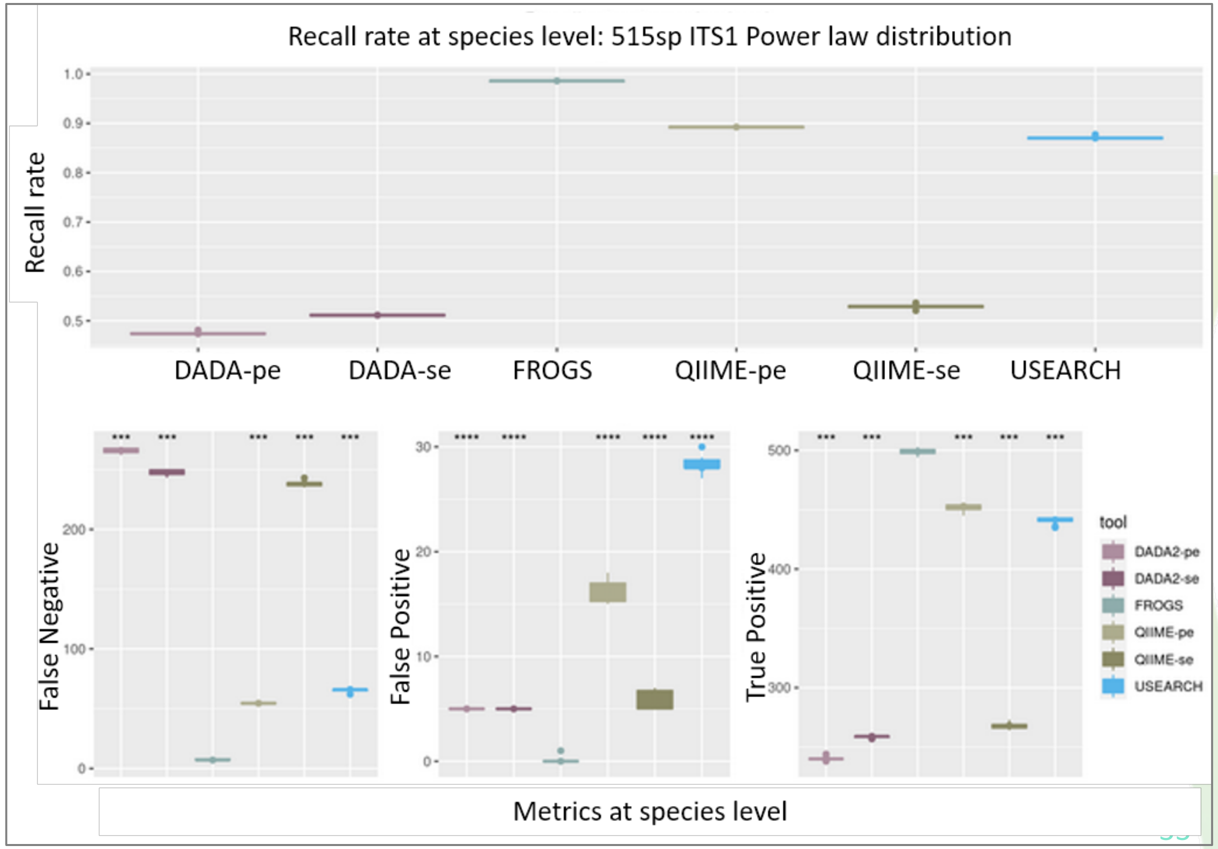


Recall rate at species level: 515sp ITS1 Uniform distribution



Metrics at species level


Divergence (%) at species level: 515sp ITS1 Power law distribution
Divergence
75
50
25
0
****
****
****
****
****
DADA-pe
DADA-se
FROGS
QIIME-pe
QIIME-se
USEARCH
Precision at species level: 515sp ITS1 Power law distribution
Precision
1.00
0.98
0.96
0.94
DADA-pe
DADA-se
FROGS
QIIME-pe
QIIME-se
USEARCH


Recall rate at species level: 515sp ITS1 Power law distribution
Recall rate
1.0
0.9
0.8
0.7
0.6
0.5
DADA-pe
DADA-se
FROGS
QIIME-pe
QIIME-se
USEARCH
False Negative
200
100
0
False Positive
30
20
10
0
True Positive
500
400
300
tool
DADA2-pe
DADA2-se
FROGS
QIIME-pe
QIIME-se
USEARCH
Metrics at species level

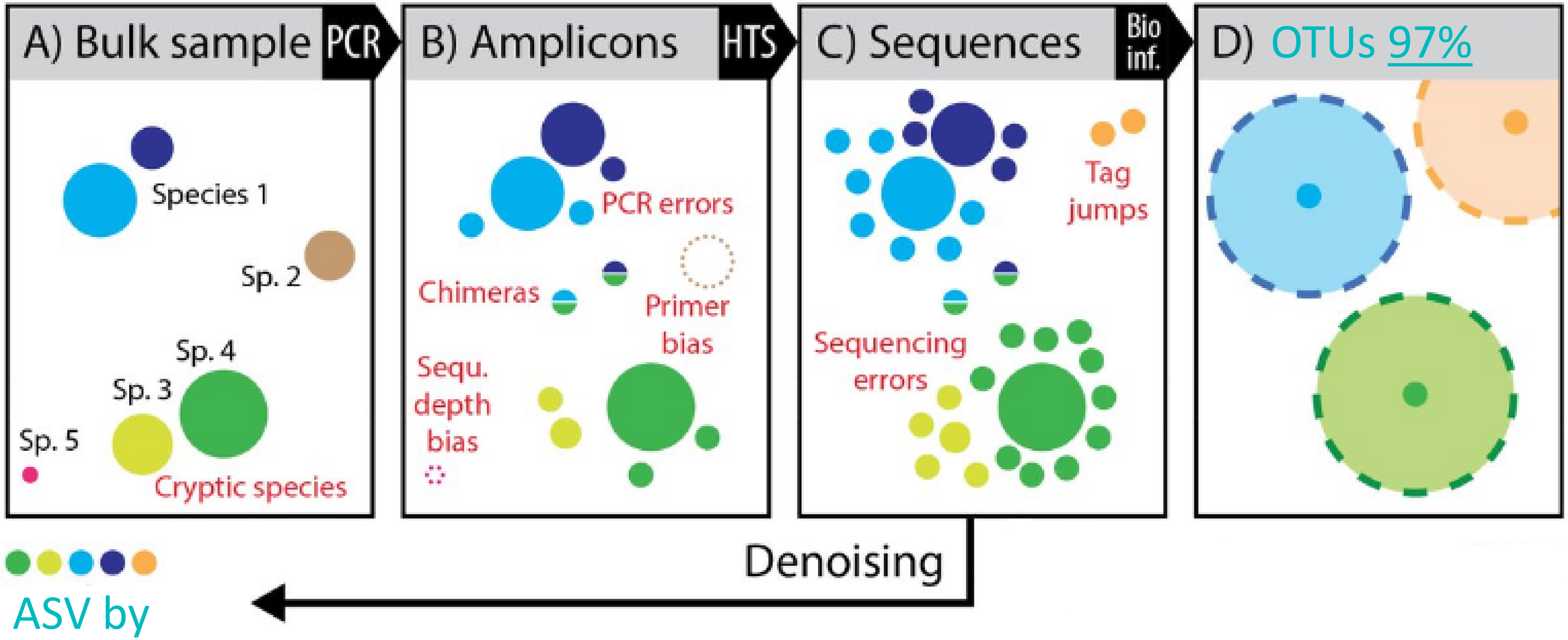
OTU → ASV

A long-standing discussion

- The ASV vs OTU debate launched by the arrival of dada2 is not so new and had been bothering us for several months/years.
- In fact, the debate largely preceded the term "ASV", and is precisely what made us opt for Swarm in FROGS (just under 10 years ago).
- To quote the author of swarm:
  “The traditional term "OTU" is negatively charged nowadays. The ASV vs OTU debate is creating confusion in the community and some users now think that all methods producing "OTUs" use a fixed clustering threshold (i.e. 97%-similarity) and are inherently bad. Of course, this is not the case and there are several methods published before the ASV term was coined that produce ASV-like clusters, swarm included.” To avoid that confusion, swarm's manual now only uses the generic term "cluster".
  https://github.com/torognes/swarm/commit/0bb491f9bf646c22a5363c27dc31a6d4b2ad335d “

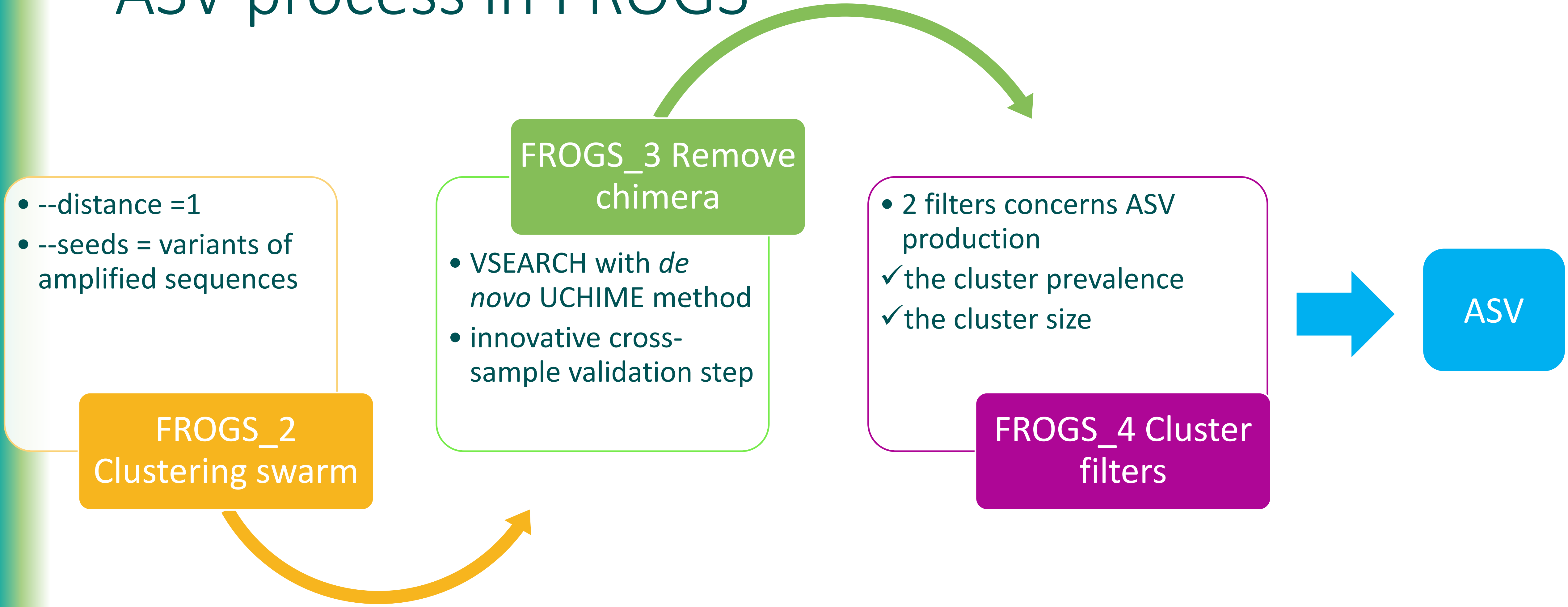
A question of vocabulary

- A few years ago, the semantic problem was the opposite, and any method that didn't produce OTUs was questioned or even disqualified.
- At the start of FROGS, we therefore chose to call our clusters "OTUs" at the end of the analysis (once the filters had been applied), but it's only a question of vocabulary, and the clusters produced by FROGS/swarm are very close to ASV in their construction.
- In any case, they look much more like ASVs than "fixed threshold" OTUs. The best thing would have been to use a new term, but Fréderic Mahé didn't make that choice at the time introducing a new term could have led to confusion.
- Since version 4.1.0 of FROGS, we have changed our vocabulary and all OTU terms have been changed to **cluster** or **ASV** in FROGS tools and outputs.



ASV by dada2 or FROGS

ASV process in FROGS



- --distance =1
- --seeds = variants of amplified sequences

FROGS_2 Clustering swarm

FROGS_3 Remove chimera

- VSEARCH with *de novo* UCHIME method
- innovative cross-sample validation step

- 2 filters concerns ASV production

✓the cluster prevalence
✓the cluster size

FROGS_4 Cluster filters

ASV

Swarm --seeds produces:
variants of amplified sequences.
"Variants" because the output sequences are all different; but with no constraints on the extent of variation - one nucleotide to infinity.

Received: 10 February 2023 | Revised: 5 June 2023 | Accepted: 6 July 2023
DOI: 10.1111/1755-0998.13847

FROM THE COVER

MOLECULAR ECOLOGY RESOURCES WILEY

A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses

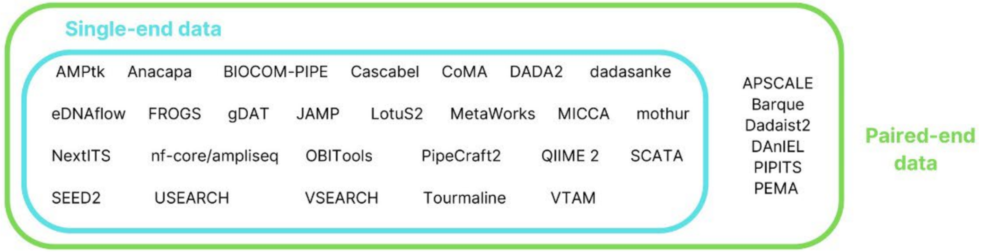
Ali Hakimzadeh[1] | Alejandro Abdala Asbun[2] | Davide Albanese[3] | Maria Bernard[4,5] | Dominik Buchner[6] | Benjamin Callahan[7] | J. Gregory Caporaso[8] | Emily Curd[9] | Christophe Djemiel[10] | Mikael Brandström Durling[11] | Vasco Elbrecht[6] | Zachary Gold[12] | Hyun S. Gweon[13,14] | Mehrdad Hajibabaei[15] | Falk Hildebrand[16,17] | Vladimir Mikryukov[1] | Eric Normandeau[18] | Ezgi Özkurt[16,17] | Jonathan M. Palmer[19] | Géraldine Pascal[20] | Teresita M. Porter[15] | Daniel Straub[21] | Martti Vasar[1] | Tomáš Větrovský[22] | Haris Zafeiropoulos[23] | Sten Anslan[1,24]

[1]Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia
[2]Department of Marine Microbiology and Biogeochemistry, NIOZ Royal Netherlands Institute for Sea Research, Texel, Netherlands
[3]Unit of Computational Biology, Research and Innovation Centre, Fondazione Edmund Mach, Italy
[4]INRAE, AgroParisTech, GABI, Université Paris-Saclay, Jouy-en-Josas, France
[5]INRAE, SIGENAE, Jouy-en-Josas, France
[6]Aquatic Ecosystem Research, University of Duisburg-Essen, Essen, Germany
[7]Department of Population Health and Pathobiology, College of Veterinary Medicine and Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA
[8]Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, Arizona, USA
[9]Vermont Biomedical Research Network, University of Vermont, Burlington, Vermont, USA
[10]Agroécologie, INRAE, Institut Agro, Univ. Bourgogne Franche-Comté, Dijon, France
[11]Department of Forest Mycology and Plant Pathology, Swedish University of Agricultural Sciences, Uppsala, Sweden
[12]Zachary Gold, NOAA Pacific Marine Environmental Laboratory, Seattle, Washington, USA
[13]UK Centre for Ecology & Hydrology, Oxfordshire, UK
[14]School of Biological Sciences, University of Reading, Reading, UK
[15]Department of Integrative Biology and Centre for Biodiversity Genomics, University of Guelph, Guelph, Ontario, Canada
[16]Gut Microbes & Health, Quadram Institute Bioscience, Norfolk, UK
[17]Earlham Institute, Norwich Research Park, Norfolk, UK
[18]Institut de Biologie Intégrative et des Systèmes, Université Laval, Québec, Québec, Canada
[19]Center for Forest Mycology Research, Northern Research Station, US Forest Service, Madison, Wisconsin, USA
[20]GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan, France
[21]Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany
[22]Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of Sciences, Praha, Czech Republic
[23]KU Leuven, Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, Laboratory of Molecular Bacteriology, Leuven, Belgium
[24]Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, Finland

Correspondence
Sten Anslan, Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia.
Email: sten.anslan@ut.ee

Present address
Jonathan M. Palmer, Genencor Technology Center, IFF, Palo Alto, California, USA

Mol Ecol Resour. 2023;00:1–17. wileyonlinelibrary.com/journal/men | 1

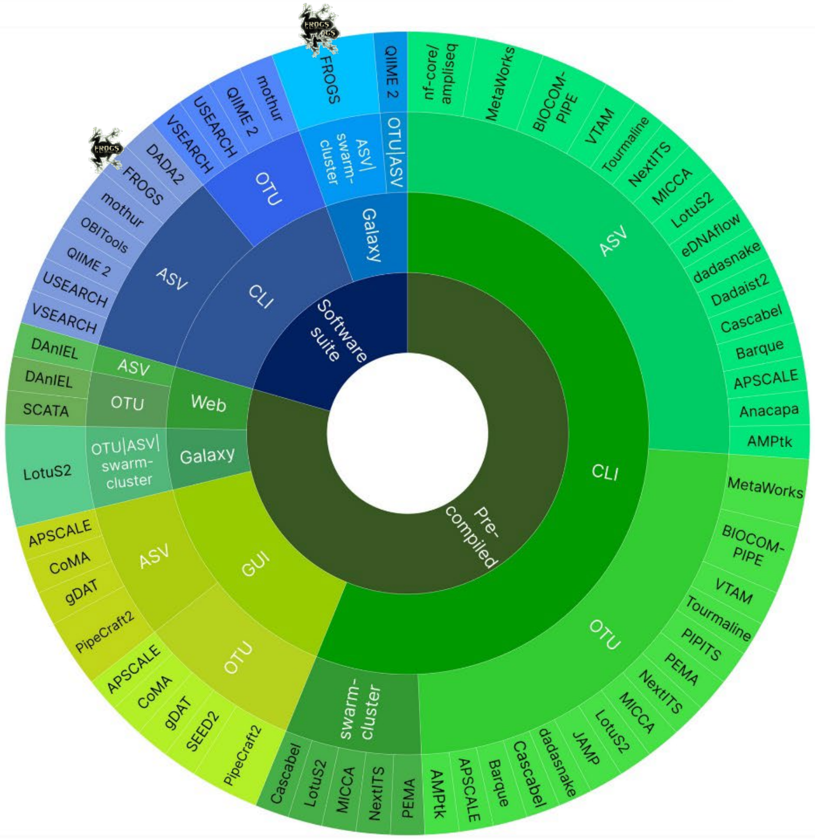


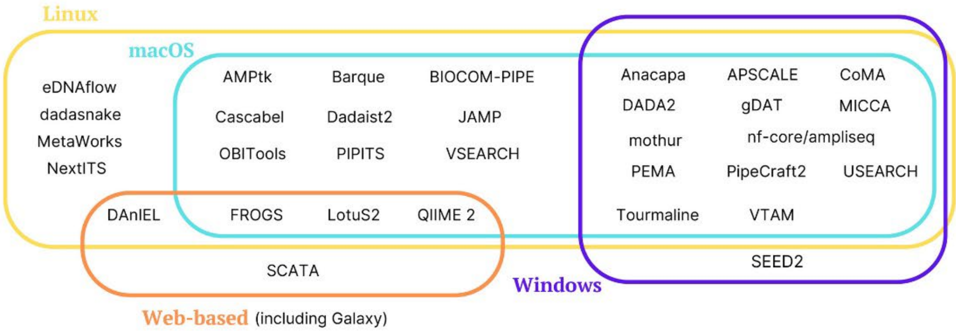
Single-end data

AMPtk Anacapa BIOCOM-PIPE Cascabel CoMA DADA2 dadasnake
eDNAflow FROGS gDAT JAMP LotuS2 MetaWorks MICCA mothur
NextITS nf-core/ampliseq OBITools PipeCraft2 QIIME 2 SCATA
SEED2 USEARCH VSEARCH Tourmaline VTAM

APSCALE
Barque
Dadaist2
DAnIEL
PIPITS
PEMA

Paired-end data





Linux

eDNAflow
dadasnake
MetaWorks
NextITS

macOS

AMPtk Barque BIOCOM-PIPE
Cascabel Dadaist2 JAMP
OBITools PIPITS VSEARCH

Anacapa APSCALE CoMA
DADA2 gDAT MICCA
mothur nf-core/ampliseq
PEMA PipeCraft2 USEARCH
Tourmaline VTAM

DAnIEL FROGS LotuS2 QIIME 2

SCATA

SEED2

Windows

Web-based (including Galaxy)