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

Maria Bernard, Olivier Rué, Mahendra Mariadassou, Géraldine Pascal

FROGS: A POWERFUL TOOL TO ANALYSE THE DIVERSITY OF FUNGI WITH SPECIAL MANAGEMENT OF INTERNAL TRANSCRIBED SPACERS

FROGS efficiently manages ITS amplicons

Maria Bernard[1,2]*, Olivier Rué[3,4]*, Mahendra Mariadassou[3], Géraldine Pascal[5].

[1]Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350, Jouy-en-Josas, France

[2]INRAE, SIGENAE, 78350, Jouy-en-Josas, France

[3]Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

[4]Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, 78350, Jouy-en-Josas, France

[5]GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France.

corresponding author: Geraldine Pascal - Landline: +33 (0)5 61 28 51 05 - Email: geraldine.pascal@inrae.fr

*The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Maria Bernard is a bioinformatics engineer. She is a member of a platform team conducting NGS sequence analysis and designing software. She specializes in workflow development in particular for metabarcoding analysis.

Olivier Rué is a bioinformatics engineer. He is in charge of data analysis at the Migale bioinformatics facility. He specializes in the analysis of metabarcoding and metagenomics data.

Mahendra Mariadassou has a Ph.D. in statistics. He is involved in the development of new statistical methods and tools for metabarcoding analysis.

Géraldine Pascal has a Ph. D. in bioinformatics and coordinates the FROGS project. She is currently involved in designing solutions for long read problems, workflow development and metagenomics analysis.

ABSTRACT

Fungi are present in all environments. They fulfil important ecological functions and play a crucial role in the food industry. Their accurate characterization is thus indispensable, particularly through metabarcoding. The most frequently used markers to monitor fungi are ITSs. These markers are the best documented in public databases but have one main weakness: PCR amplification may produce non-overlapping reads in a significant fraction of the fungi. When these reads are filtered out, traditional metabarcoding pipelines lose part of the information and consequently produce biased pictures of the composition and structure of the environment under study. We developed a solution that enables processing of the entire set of reads including both overlapping and non-overlapping, thus providing a more accurate picture of fungal communities. Our comparative tests using simulated and real data demonstrated the effectiveness of our solution, which can be used by both experts and non-specialists on a command line or through the Galaxy-based web interface.

INTRODUCTION

Using amplicon sequencing to describe the microbial composition of an environment is a time saving and cost-effective strategy and can be used even for very large-scale surveys [1]. Most studies currently focus on the bacterial fraction of microbial communities but the fungal fraction is equally important, as fungi are ubiquitous and provide several ecosystem services [2]. Unfortunately, studying the fungal fraction using metabarcoding has its own challenges. Indeed, in fungi, there is no equivalent of the 16S rRNA gene, which is widely used and highly suitable for bacteria. The best candidates are internal transcribed spacers (ITS), but these are more difficult to manipulate. The main problem with ITS is size polymorphism, with a size range of 361 to 1475 bases in UNITE 7.1 [3] (unlike 16S where 95% of the sequences have a

length between 1205 and 1556 bases). Most studies describing ITS data analyses process either (i) paired-end reads but filter out non-overlapping, non-mergeable reads, thus systematically discarding taxa with longer ITS, or (ii) single-end reads, thus limiting taxonomic resolution and losing the benefit of information contained in longer sequences [4, 5]. In both cases, the tools used, which only support mergeable paired-end reads or single-end reads, are unable to process all the sequences produced. The new version of FROGS [6] solves these challenges. This metabarcoding analysis pipeline runs not only on the command line but also on the Galaxy platform, meaning it is easy to use for non-experts. FROGS tools make it possible to generate and explore the abundance of OTUs (Operational Taxonomic Units) in the target environment along with their taxonomic affiliation, FROGS also includes 7 tools dedicated to descriptive statistics and two tools for differential abundance analysis (supplemental data section 1). FROGS is widely used by the research community worldwide to study microbial environments i.e., water, skin, soil, gut, animal, human, food microbiomes [7-11]. Enriched with new features every year, FROGS is a very powerful tool to treat ITS amplicons. Here, we propose a smart method and the tools needed to deal with both overlapping and non-overlapping paired-end reads of, in particular, ITS.

AVAILABILITY AND IMPLEMENTATION

FROGS can be used both on the Galaxy platform and on a command line. Installing and updating FROGS are easy thanks to the FROGS repositories in CONDA (https://anaconda.org/bioconda/frogs) and Galaxy Toolshed (https://toolshed.g2.bx.psu.edu/view/frogs/frogs). Two github FROGS repositories are effective https://github.com/geraldinepascal/FROGS and

https://github.com/geraldinepascal/FROGS-wrappers. FROGS supports ITS since version 3.0 and is fully documented on the companion website http://frogs.toulouse.inrae.fr/.

FROGS: 7-STEP GUIDELINES FOR THE MANAGEMENT OF INTERNAL TRANSCRIBED SPACERS (ITS)

To process ITS data correctly, we recommend a seven-step process shown in Figure 1 (for command lines, see supplemental data section 2). Here, we only highlight the adaptations made to process ITS data and the novelties introduced since version 3.0 [6].

FROGS Preprocessing: After the paired-end amplification process, ITS reads may not overlap. In this step, merging tools such as PEAR [12], VSEARCH [13] or Flash [14] cannot merge these reads and they may thus be lost by the system. The solution we found to work best to avoid losing ITS amplicons is to both conventionally process mergeable data and to create artificial sequences with non-mergeable sequences. The idea is to join the two reads together with a stretch of 100 N letters. These "un-merged" sequences are tagged as "combined" sequences. This option is activated only if reads do not entirely cover the target sequence. Otherwise, noise is retained in analyses. FROGS "combined" sequences are artificial and present particular features, especially their size. For a MiSeq sequencing of 2x250pb, FROGS will combine non-overlapping reads into a single sequence 600 bp long. So, for ITS analysis to keep both merged and unmerged reads, choose *--keep-unmerged* parameter. We would normally advise using PEAR [12] to merge sequences because it produces the best results, but because non academics require a commercial license, we chose VSEARCH [13] as our default paired-end read merger.

FROGS Clustering: To cluster both classical amplicons and unmergeable amplicon as ITS, we advise to choose the parameters of distance equal to 1 and the fastidious option (*--distance 1 --fastidious*) as recommended in Swarm v2 [15]. We have changed the guidelines from those given in FROGS v.1., and these new parameters provide an equally well defined cluster but are much faster. Swarm handles "combined" sequences very well. We simply and temporarily change the stretch of 100 N into a stretch of 50 A and 50 C. Then, Swarm treats these combined sequences as mergeable sequences. The principle of Swarm is that it creates links between sequences two by two when they have only one difference between them (with distance parameter equals 1). Thus, Swarm does the same for combined sequences, the 50 A 50 C stretches not being discriminating between combined sequences since they are identical, these differences being elsewhere on the sequence.

FROGS Chimera removal: there is no impact on ITS analysis with this tool, use it without specificity.

FROGS OTU Filter: Users must remove too low abundance clusters that merge with clusters of artifact sequences. The problem is doubtless not a bioinformatics problem but rather the use of metabarcoding to detect rare microbes in a large population of microbes. The best way to remove these small noisy clusters is to apply a 0.005% *(--min-abundance*) abundance threshold and/or, if applicable, to keep the clusters only if they are present in a certain number of samples or replicated samples (*--min-sample-presence*). The user now has the possibility to provide a FASTA file containing potential contaminant sequences e.g., phiX, host sequences, mitochondrial sequences, OTUs that align with a contaminant sequence, are deleted.

FROGS ITSx: If the objective is to analyse fungal ITS, we advise only keeping sequences detected by ITSx [16] to have an ITS signature with no conserved flanking sequences truncation (*--check-its-only*). ITSx identifies the highly conserved neighbouring rRNA sequences SSU, 5.8S and/or LSU. If the targeted variable region (ITS1 or ITS2) is not detected, the sequence is discarded. The ITSx step is time consuming and has to be done after clustering and filtering steps.

FROGS Affiliation: users have to select one of available ITS databanks. The affiliation process is specific for ITS, adapted to un-merged data thanks to a mix of local and global alignments. Because of the inclusion of the 100 N stretch in the “combined” sequences, the affiliation of the OTU seed sequences against the databases by BLASTN+ [17], as in FROGS v1, was not possible. BLASTN+ aligns the sequences locally and, by definition, only keeps the best local alignment, i.e., either the 5' sequence or the 3' sequence, but in no case the entire “combined” sequence, since the N stretch collapses the BLAST score. We thus chose a double strategy. All normally merged sequences are treated like in FROGS v1 by BLASTN+ against a databank (e.g. UNITE) while the combined sequences, which are automatically detected by the system, are treated with needleall [18], a global sequence aligner. To align with needleall takes too long, to reduce this time, we chose to align combined sequences not against the entire databank (e.g. UNITE) but against a drastically reduced bank. To this end, first, the two sequence parts (read1 and read2) are aligned with BLASTN+ versus the chosen database. For each read, reference sequences with score included in the 100 best scores are kept. A reduced database is constructed with all kept references among all read1 and read2 of all combined OTU BLASTN+ alignments. Second, combined sequences are aligned with needleall versus this small new databank. For both the combined sequences and the others, only the best hits with the same best score are reported. For each alignment returned, several metrics are computed:

identity percentage, coverage percentage, and alignment length. Please note that we format and add on demand databases that are absent from FROGS but necessary for our users. The list of currently available databases can be found here: http://genoweb.toulouse.inra.fr/frogs_databanks/assignation/readme.txt

FROGS Affiliation Filter: users can mask or delete taxonomic affiliations of OTUs according to several criteria. For BLAST taxonomies: minimum identity percentage, minimum coverage percentage, minimum alignment length, maximum e-value for BLAST taxonomies; for RDP (Ribosomal Database Project) taxonomies: a minimum bootstrap threshold at a specific rank; and for all: the absence of the full or partial taxon name. Two modes are possible to process these impacted OTUs. In the masking mode (*--mask*), the taxonomic information is replaced by NA. In this case, affiliations with insufficient characteristics are not retained. In delete mode (*--delete*), the OTUs are deleted according to the filter settings. Among other things, this makes it possible to delete chimeric OTUs (too low percentage coverage). Another feature enabled by the tool is being able to mask or delete taxonomic information by specifying particular key words (e.g. unknown species). This functionality is particularly useful to remove contaminants that can be detected using taxonomic information.

DATASETS AND TESTS TO ASSESS THE MANAGEMENT OF ITS:

To test our methodology, we generated increasingly complex simulated datasets, available on https://doi.org/10.15454/AOT7UL (supplemental data section 3, supplemental Figure 1) and used a mock biological community (supplemental data section 4). These data were processed by FROGS but also by USEARCH, QIIME2 and DADA2. Since QIIME2 and DADA2 filter out non-overlapping reads, we used QIIME2 and DADA2 in single-end and paired-end modes. For USEARCH, we followed the original guidelines by taking into account merged sequences and

5‘ R1 reads of non-overlapping paired-end sequences. All benchmarking command lines are available on http://frogs.toulouse.inrae.fr/ITS/frogs-its.html#Command_lines. We used four metrics to assess our results: (i) the divergence rate: computed as the Bray-Curtis distance between expected and observed abundance profiles at a given taxonomic level (observed composition is the abundance resulting from each tool process), (ii) the number of false-negative taxa (FN): the number of expected taxa that were not recovered by the method, (iii) the number of false positive taxa (FP): the number of recovered taxa that were not expected, and (iv) the number of true positive taxa (TP): the number of recovered taxa that were expected. We computed the precision (TP/(TP+FP)) and the recall rate (TP/(TP+FN)) of all methods, on all datasets: 35, 115 and 515 species, ITS1, ITS2, power law or uniform distribution of abundances, of simulated and real data. These two metrics allow us to evaluate the performance of each tool based on FP, TP, and FN metrics.

RESULTS AND DISCUSSION

On simulated data, whatever the complexity of the microbial diversity of the samples (35 (Figure 2), 115 or 515 species (see companion website http://frogs.toulouse.inrae.fr/ITS/frogs-its.html#35_species, #115_species #515_species and supplemental Figures 2 to 6)), FROGS accurately recovered the species originally present in the samples. FROGS missed very few fungal sequences, had very few or no FP and very few or no FN. Precision (min. 0.991 max. 1) and the recall rates (min. 0.976 max. 1) were excellent (supplemental Figure 2 and supplementary Table), and better than those of the other tools. As expected, species with low abundance rates were the most difficult to recover. The reference reconstruction and the number of OTUs with the expected length was very good with FROGS. The results of comparison with other tools showed, as expected, that tools that

process paired-end reads (DADA2-pe, QIIME-pe) missed unmergeable ITS and tools that only process R1 reads (DADA2-se and QIIME-se) lost too much sensitivity and had more difficulty correctly identifying the species in the samples. This calls into question the strategy of processing only R1 reads. Despite good precision, DADA2 combined a low recall rate (supplemental Figure 2), mis-reconstruction of the reference (Figure 2D) and a high divergence rate. USEARCH reconstructed the references quite well and correctly identified the expected species, but at the price of more FP resulting in very limited precision (Figure 2A, B and supplemental Figures 3 to 6). The USEARCH filter applied to the bootstrap value at the end of the process, as recommended, removed some TP ZOTUs (Zero-radius OTUs). The great strength of FROGS is its ability to handle both mergeable and unmergeable reads but still being easy to use. However, it should be noted that the quality of the unmergeable sequences is lower because they do not benefit from the correction enabled by the overlapping parts. We can therefore expect less good identification of fungi with unmergeable ITS amplicons than of mergeable ones. Moreover, the processing unmergeable sequences requires global sequence alignment, which increases FROGS computing time.

On real data (http://frogs.toulouse.inrae.fr/ITS/frogs-its-meat.html, supplemental Figure 7 and supplemental Table), FROGS produced very good results. The divergence rates of all the tool were worse than on simulated data because they were calculated on expected data while it is difficult to match theoretical and real quantities when preparing samples. The greatest divergences could therefore be due to this. Concerning FP, TP and FN, FROGS showed high precision and a satisfactory recall rate (median 0.99 and 0.92 respectively), even if, as expected, FROGS was not able to differentiate between species that are too similar, e.g. Penicillium. Nevertheless, considering the taxonomies found, FROGS was the tool that

detected the most expected species. FROGS also reconstructed the expected sequences best (100% identity and 100% coverage of the entire sequence).

CONCLUSION

FROGS is an easy-to-use suite that is perfectly suited and efficient not only for the processing of internal transcribed spacer amplicon sequences, but also of all other unmergeable amplicons i.e. rpb2 [19] and D1-D2 [20]. The test of FROGS data processing demonstrated excellent recall rates and precision thanks to smart ITS data management while accurately reconstructing the sequences as they are expected. FROGS can be used both on the command line and on the Galaxy interface, making it easy for everyone to access.

KEY POINTS

- FROGS processes ITS amplicons (and any other non-overlapping reads) with a smart method.
- FROGS was compared to many popular metabarcoding tools and proved to be very sensitive with confident OTU sequence reconstruction.
- FROGS produces very few false positives and false negatives, has an excellent recall rate and high precision.
- FROGS can be run both on the command line and on the Galaxy interface.

KEYWORDS

Fungi, ITS, metabarcoding, workflow, amplicon, metagenomics

ACKNOWLEDGMENTS

The authors are grateful to the genotoul bioinformatics platform Occitanie (doi: 10.15454/1.5572369328961167E12) and Migale bioinformatics facility (doi: 10.15454/1.5572390655343293E12) for providing help, computing and storage resources. The authors thank Valentin Marcon, Patrick Durand, Laure Quintric and Olivier Inizan for the packaging of FROGS under CONDA. The authors thank Ta Thi Ngan for her involvement in the development of FROGSSTAT tools. The authors also thank the members of the METABARFOOD project for allowing us to use the data generated as part of the project, particularly Monika Coton, Jérôme Mounier and Audrey Pawtowski who were in charge of the fermented meat samples. The authors are grateful to Daphne Goodfellow for attention to the English-language version.

REFERENCES

1. Ibarbalz FM, Henry N, Brandao MC et al. Global Trends in Marine Plankton Diversity across Kingdoms of Life, Cell 2019;179:1084-1097 e1021.
2. Naranjo-Ortiz MA, Gabaldon T. Fungal evolution: diversity, taxonomy and phylogeny of the Fungi, Biol Rev Camb Philos Soc 2019;94:2101-2137.
3. Koljalg U, Nilsson RH, Abarenkov K et al. Towards a unified paradigm for sequence-based identification of fungi, Mol Ecol 2013;22:5271-5277.
4. Charlie Pauvert MB, Valérie Laval, Véronique Edel-Hermann, Laure Fauchery, Angélique Gautier, Isabelle Lesur, Jessica Vallance, Corinne Vacher,. Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline,, Fungal Ecology 2019:23-33.
5. Nguyen NH, Smith D, Peay K et al. Parsing ecological signal from noise in next generation amplicon sequencing, New Phytol 2015;205:1389-1393.
6. Escudie F, Auer L, Bernard M et al. FROGS: Find, Rapidly, OTUs with Galaxy Solution, Bioinformatics 2018;34:1287-1294.
7. de Lorgeril J, Lucasson A, Petton B et al. Immune-suppression by OsHV-1 viral infection causes fatal bacteraemia in Pacific oysters, Nat Commun 2018;9:4215.
8. Teyssier A, Rouffaer LO, Saleh Hudin N et al. Inside the guts of the city: Urban-induced alterations of the gut microbiota in a wild passerine, Sci Total Environ 2018;612:1276-1286.
9. Frere L, Maignien L, Chalopin M et al. Microplastic bacterial communities in the Bay of Brest: Influence of polymer type and size, Environ Pollut 2018;242:614-625.

10. Hoyles L, Fernandez-Real JM, Federici M et al. Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women, Nat Med 2018;24:1070-1080.
11. Duron O, Morel O, Noel V et al. Tick-Bacteria Mutualism Depends on B Vitamin Synthesis Pathways, Curr Biol 2018;28:1896-1902 e1895.
12. Zhang J, Kobert K, Flouri T et al. PEAR: a fast and accurate Illumina Paired-End reAd mergeR, Bioinformatics 2014;30:614-620.
13. Rognes T, Flouri T, Nichols B et al. VSEARCH: a versatile open source tool for metagenomics, PeerJ 2016;4:e2584.
14. Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies, Bioinformatics 2011;27:2957-2963.
15. Mahe F, Rognes T, Quince C et al. Swarm v2: highly-scalable and high-resolution amplicon clustering, PeerJ 2015;3:e1420.
16. Bengtsson-Palme J, Ryberg M, Hartmann M et al. Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data, Methods in Ecology and Evolution 2013;4:914-919.
17. Camacho C, Coulouris G, Avagyan V et al. BLAST+: architecture and applications, BMC Bioinformatics 2009;10:421.
18. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins, J Mol Biol 1970;48:443-453.
19. Větrovský T, Kolařík M, Žifčáková L et al. The rpb2 gene represents a viable alternative molecular marker for the analysis of environmental fungal communities, Molecular Ecology Resources 2016;16:388-401.
20. Sonnenberg R, Nolte, A.W. & Tautz, D. An evaluation of LSU rDNA D1-D2 sequences for their use in species identification., Front Zool 2007;4.

FIGURES:

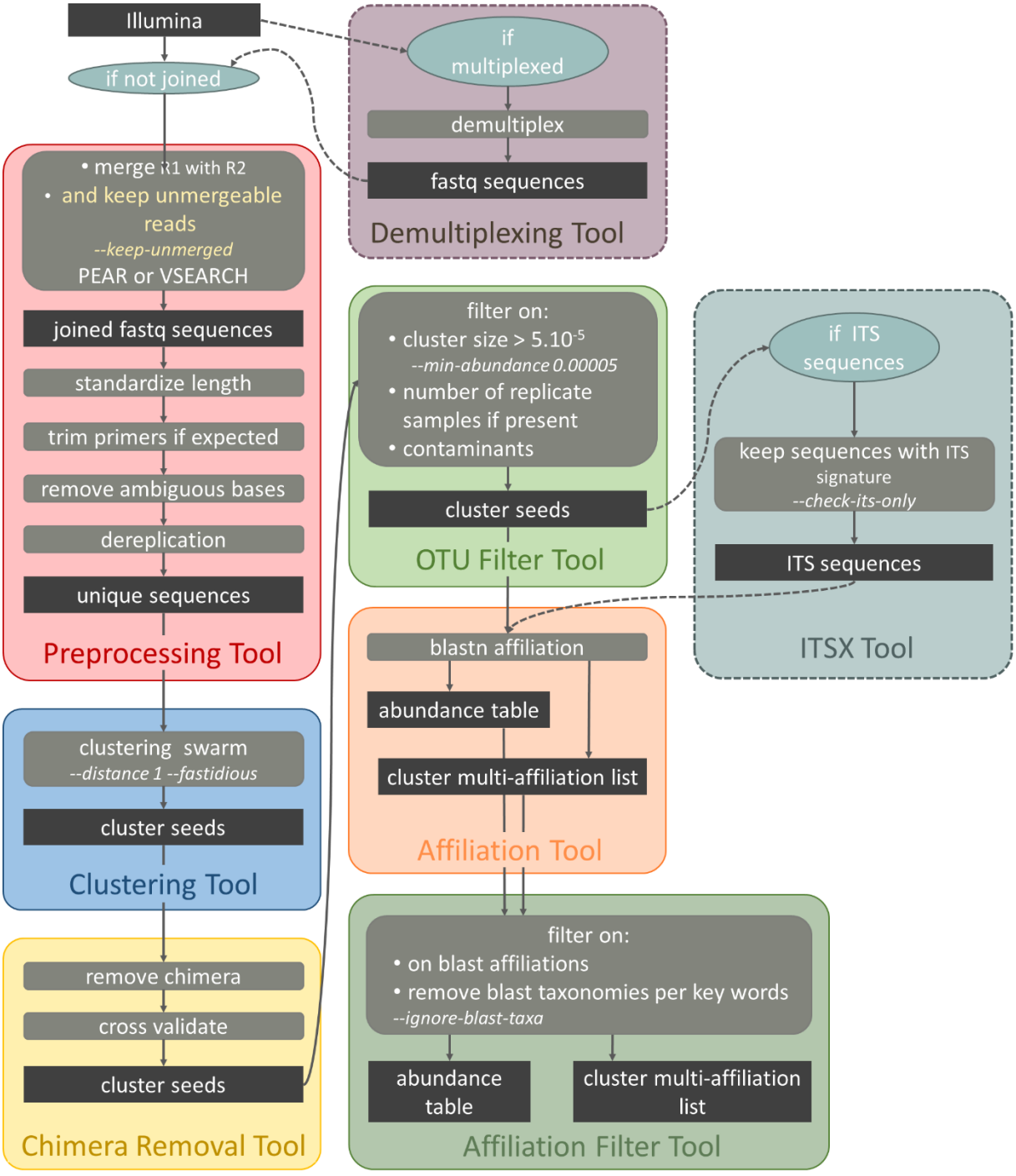


Figure 1: Standard operation procedure for long and unmergeable reads from metabarcoding sequencing i.e. ITS, rpb2, D1-D2

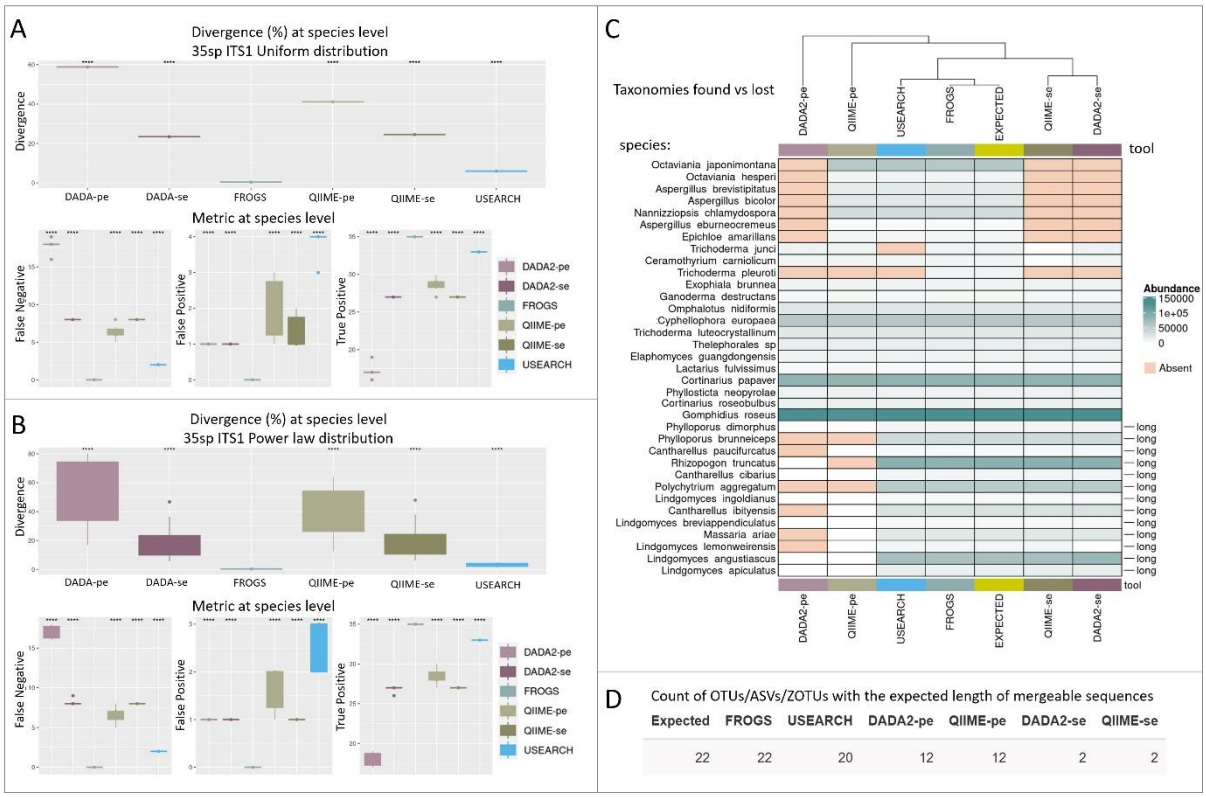


Figure 2: Results obtained on the 35 species ITS1 dataset (N.B. these results are a representative set of results). ***A*** *and* ***B:*** *Boxplots of the results of the processing of the synthetic datasets (uniform and power law abundance distribution, respectively) of the 4 tools tested (DADA2 paired-end (pe), DADA2 single-end (se) (pale and dark purple), FROGS (bluish green), QIIME-pe, QIIME-se (pale and dark khaki) and USEARCH (cyan)). 4 metrics (Abundance divergence, FP, FN and TP rates) are calculated in relation to the expected. Affiliations and associated abundances are taken into account. Each tool was compared to FROGS using a Wilcoxon signed-rank test, ****: p <= 0.0001; ***: p <= 0.001; **: p <= 0.01; *: p <= 0.05; ns: p > 0.05).* ***C:*** *The heatmap (columns are clustered using the Canberra distance) represents the taxonomy lost or found relative to the expected dataset (Expected: test dataset on 35 species, ITS1, power law distribution). The colour gradient represents the abundance of the species.* ***D:*** *Number of OTUs/ASVs with the expected length (OTUs: Operational Taxonomic Units built by FROGS and QIIME; ASVs: Amplicon Sequence Variants built by DADA2 and ZOTUs: Zero-radius OTUs built by USEARCH).*

SUPPLEMENTAL DATA

1. NINE FROGS STATISTICAL TOOLS

The analysis of OTU (Operational Taxonomic Unit) tables with respect to the associated metadata (pH, temperature, weight, treatment, etc.) requires the use of statistical methods dedicated to compositional data. The OTU abundance table produced by FROGS contains the abundance information for each OTU in each sample as well as its taxonomic affiliation. Questions such as "How many species are there in each sample i.e., species richness?" and "Which OTUs are abundant in my samples, i.e. structure" or "How does the microbial community vary with the different metadata?" can only be based on the calculation of distances, indices and descriptive statistics. FROGS therefore has seven tools for the statistical analysis of samples using the phyloseq R package, extended R functions in Rmarkdown scripts. The first tool imports data and transforms data in the abundance table and metadata into an Rdata object. The second tool uses phyloseq and a custom R function to build two plots to visualize the composition of the sample. The third tool builds richness plots to visualize the alpha diversity of the sample. The fourth computes the beta distance matrixes and displays them on a heatmap. The fifth builds a heatmap plot and an ordination plot using ordination methods to visualize data structure. The sixth builds the clustering plot for cluster analysis with different linkage methods, and the seventh uses phyloseq and vegan functions in R to perform multivariate analysis of variance (MANOVA).

As a preliminary step, users can obtain a phylogenetic tree based on OTUs using the FROGS Tree tool. This tool creates a multiple alignment of OTUs with Mafft and reconstructs phylogenetic tree with FastTree and phangorn R package.

Finally, FROGS include also two statistical tools dedicated to differential abundance analysis. Like previous tools, is uses Rmarkdown scripts, but based on DESeq2 R package. The first tool produces a DESeq2 object including the statistical model, and the second tool exports the comparison results between chosen conditions by producing differentially abundant OTU table, pie chart, MA plot and volcano plot.

As for bioinformatics tools, almost all tools return HTML reports with a variety of graphics and tables.

2. FROGS COMMAND LINES TO PROCESS ITS READS

```
#!/bin/bash


# To use FROGS, first you need to activate your environment
```

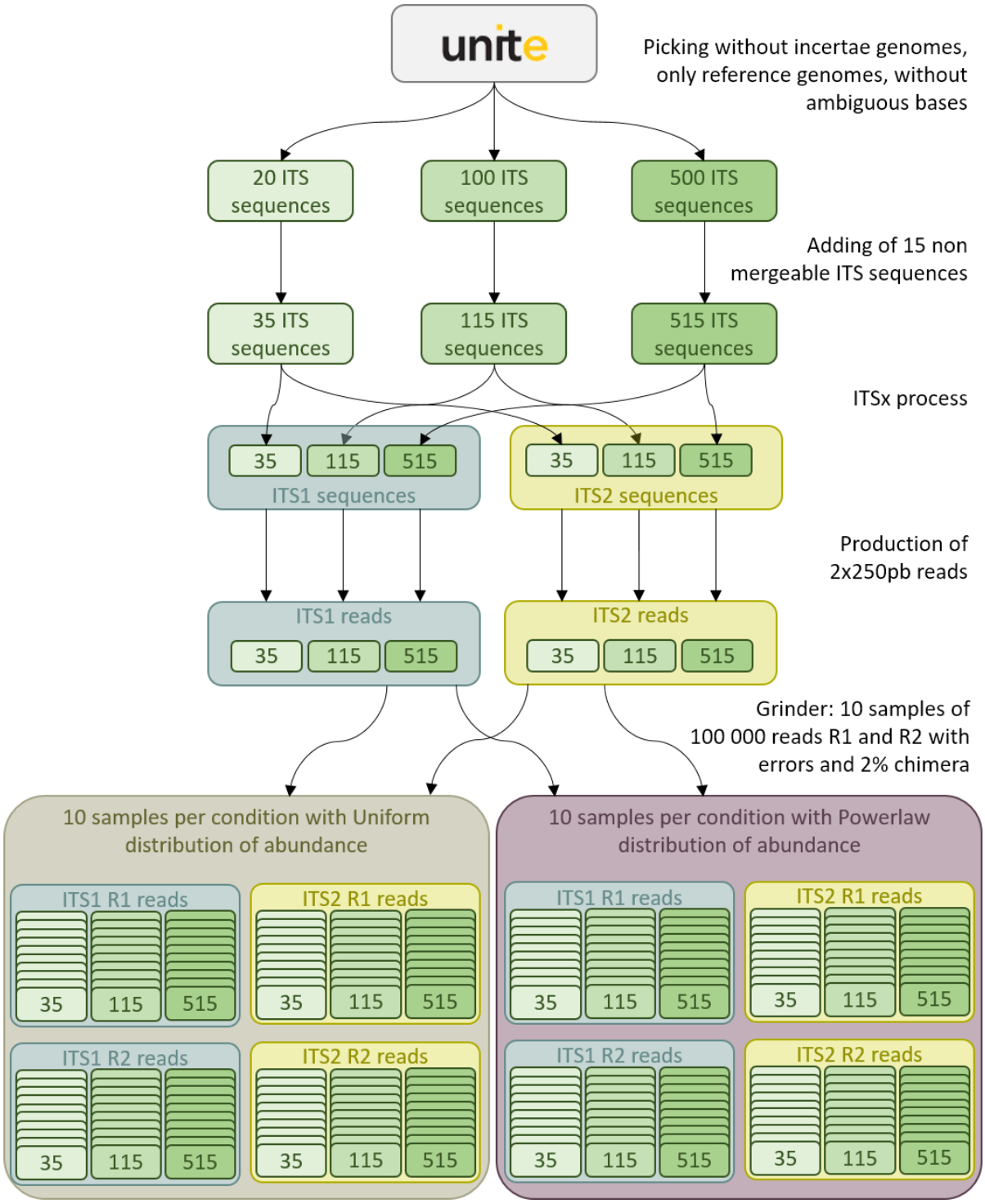
```
conda create --name __frogs@3.2.0 frogs=3.2.0

conda activate __frogs@3.2.0


# Archive.tar.gz contains all fastq files of R1 and R2 ITS (ITS1 or ITS2)

ARCHIVE_FILE=”./Archive.tar.gz”


# Database of reference for affiliation step

mkdir Unite_ref

cd Unite_ref

wget
http://genoweb.toulouse.inra.fr/frogs_databanks/assignation/Unite/Unite_s_7.1_2011
2016_ITS.tar.gz

tar -xvzf Unite_s_7.1_20112016_ITS.tar.gz

cd ..

REFERENCE="./Unite_ref/Unite_s_7.1_20112016_ITS.fasta"


# Number of threads to use

CPU=1


# Preprocess

preprocess.py illumina --input-archive $ARCHIVE_FILE --keep-unmerged --merge-
software pear --min-amplicon-size 20 --max-amplicon-size 490 --without-primers --
R1-size 250 --R2-size 250 --nb-cpus $CPU --output-dereplicated preprocess.fasta --
output-count preprocess.tsv --summary preprocess.html --log-file preprocess.log


# Clustering

clustering.py --input-fasta preprocess.fasta --input-count preprocess.tsv --
fastidious --distance 1 --nb-cpus $CPU --output-biom clustering.biom --output-fasta
clustering.fasta --log-file clustering.log


# Remove Chimera

remove_chimera.py --input-fasta clustering.fasta --input-biom clustering.biom --nb-
cpus $CPU --out-abundance remove_chimera.biom --non-chimera remove_chimera.fasta -
-log-file remove_chimera.log --summary remove_chimera.html


# Abundance filter process
```

```
otu-filters.py --min-abundance 0.00005 --nb-cpus $CPU --input-fasta remove_chimera.fasta --input-biom remove_chimera.biom --output-biom filters.biom --output-fasta filters.fasta --log-file filters.log --summary filters.html


# ITSX filter

itsx.py --nb-cpus $CPU --region $its --check-its-only --input-fasta filters.fasta --input-biom filters.biom --out-fasta itsx.fasta --out-abundance itsx.biom --summary itsx.html --log-file itsx.log


# Affiliation versus Unite

affiliation_OTU.py --nb-cpus $CPU --reference $REFERENCE --input-biom itsx.biom --input-fasta itsx.fasta --output-biom frogs.biom --summary affiliation.html --log-file affiliation.log


# Tabular abundance table

biom_to_tsv.py --input-biom frogs.biom --input-fasta itsx.fasta --output-tsv frogs.tsv --output-multi-affi multiaff.tsv --log-file biom_to_tsv.log


# Affiliation filter: parameters is very dependent on your dataset and questions

./affiliation_filters.py --input-biom frogs.biom --input-fasta itsx.fasta --output-frogs_affi_filter.biom --output-fasta frogs_affi_filter.fasta --summary frogs_affi_filter.html --impacted frogs_affi_filter_impacted_OTU.tsv --impacted-multihit frogs_affi_filter_impacted_OTU_multihit.tsv --log-file affi_filter.log \

--delete / --mask        # choose behavior

# add blast affiliation filter \

#--min-blast-identity 0.8 --min-blast-coverage 0.8 --ignore-blast-taxa "known_contaminant_terme or key_species"
```

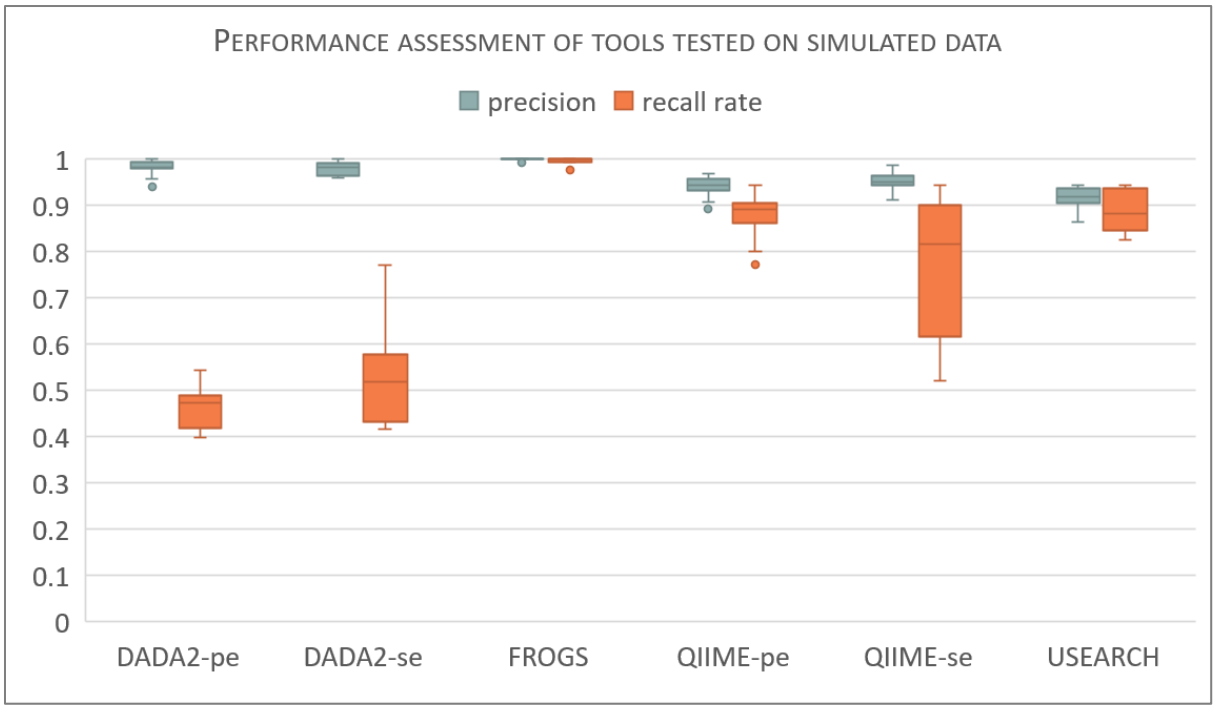
3. TEST DATASET FOR ASSESSMENT OF THE FROGS ITS METHODOLOGY

Our objective was to create datasets which make it possible, among other things, to test methods and tools to process data on fungi for metabarcoding. Consequently, we obtained computer-generated paired reads 2 x 250 nucleotides in length from artificial PCRs produced with Grinder software [1] on ITS1 or ITS2 from fungi which are as diverse as possible. Datasets are available on https://doi.org/10.15454/AOT7UL. We extracted the reference sequences from the UNITE v7.1 database centred on the eukaryotic nuclear ribosomal ITS region, we selected species without ambiguous key words i.e. without unknown and *incertae* key words in their name. UNITE 7.1 from general FASTA releases contains 54 568 sequences composed

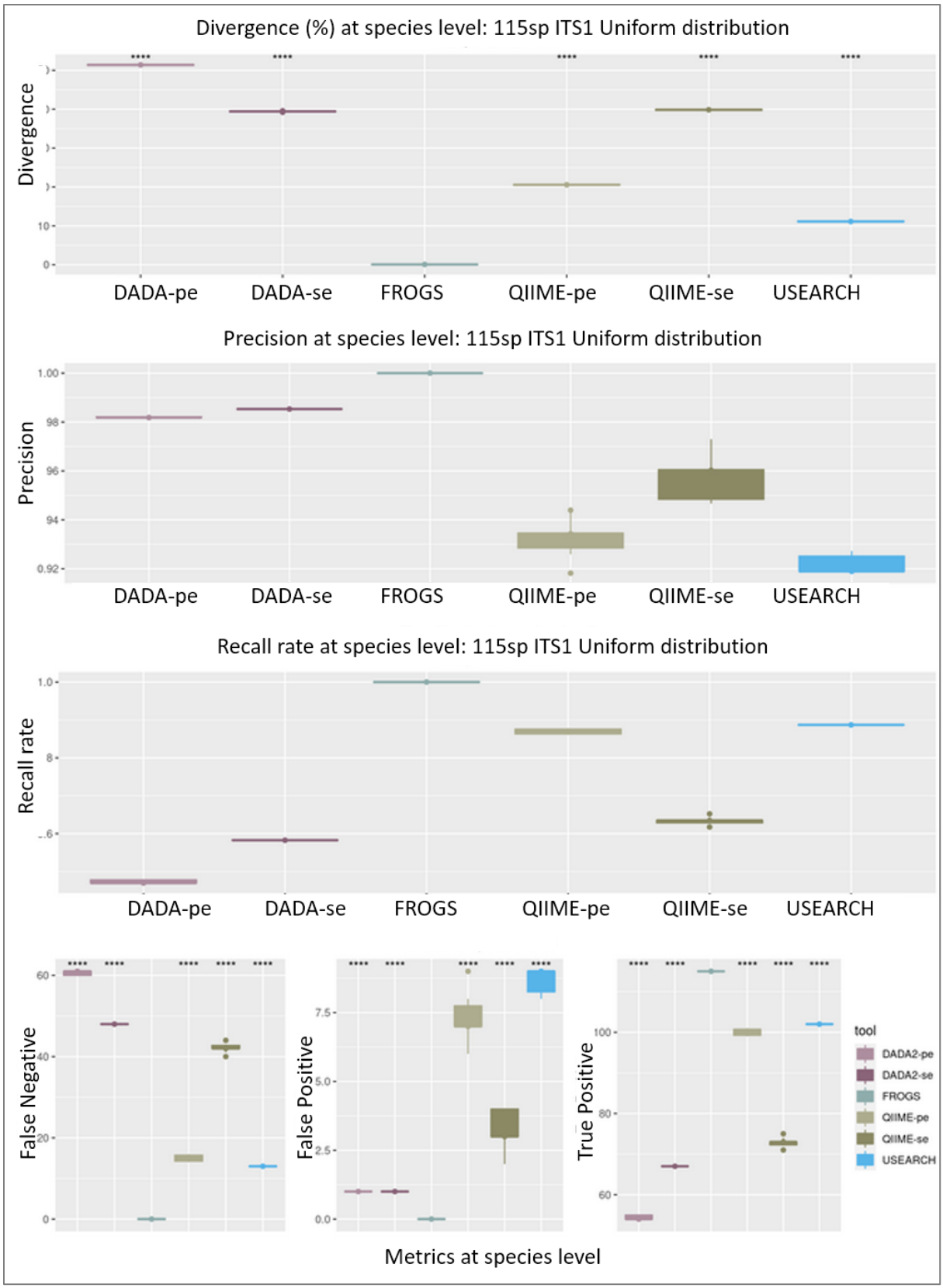
of 8 179 and 46 389 sequences from respectively reference and representative species. It is composed of 54 291 Fungi with 10 phyla, 52 classes and 185 orders. We produced simulated data reflecting this distribution. The sequences used to support artificial amplification by *in silico* PCR were taken from the UNITE v7.1 database thanks to the python script treeSampling.py (https://github.com/geraldinepascal/FROGS/tree/master/assessment/bin) and see supplemental Figure 1 that describe the following procedure. We picked 35, 115, or 515 species of fungi to create three different types of diversity. We used Grinder, which is, among other things, a programme designed to create random amplicon sequence libraries based on DNA reference sequences. We produced paired-end reads R1 and R2, type Illumina Miseq. We incorporated sequencing errors. Reads were simulated by increasing the number of substitutions and indels linearly along the reads (Grinder parameters: *-mutation_dist* linear 2.54e-1 2.79e-1 and *-mutation_ratio* 98.6 1.4). We generated samples with two types of abundance distribution that follow either a power law or a uniform law. For each type of diversity, we merged these first datasets with exclusive datasets of chimera (3 types of chimera - Grinder parameters: *-chimera_dist* 314 38 1) to obtain an overall chimera rate of 2% per dataset. We focused on including reads from ITS1 and ITS2 of long lengths, not overlaid by two paired 250-nucleotide reads (min:132 nucl. to max 763 nucl. *c.f.* sections #Amplicon length distribution and #List of non-mergeable amplicons available from http://frogs.toulouse.inrae.fr/ITS/frogs-its.html#35_species, #115_species, #515_species). Thus, the resulting paired-end reads become unmergeable using sequence merging tools such as FLASH, VSEARCH or PEAR and reflect the reality of the sequences obtained in vitro.



***Supplemental figure 1**: Diagram showing the in silico construction of the 240 simulated fastq files of 100 000 sequences each from the UNITE v7.1.*

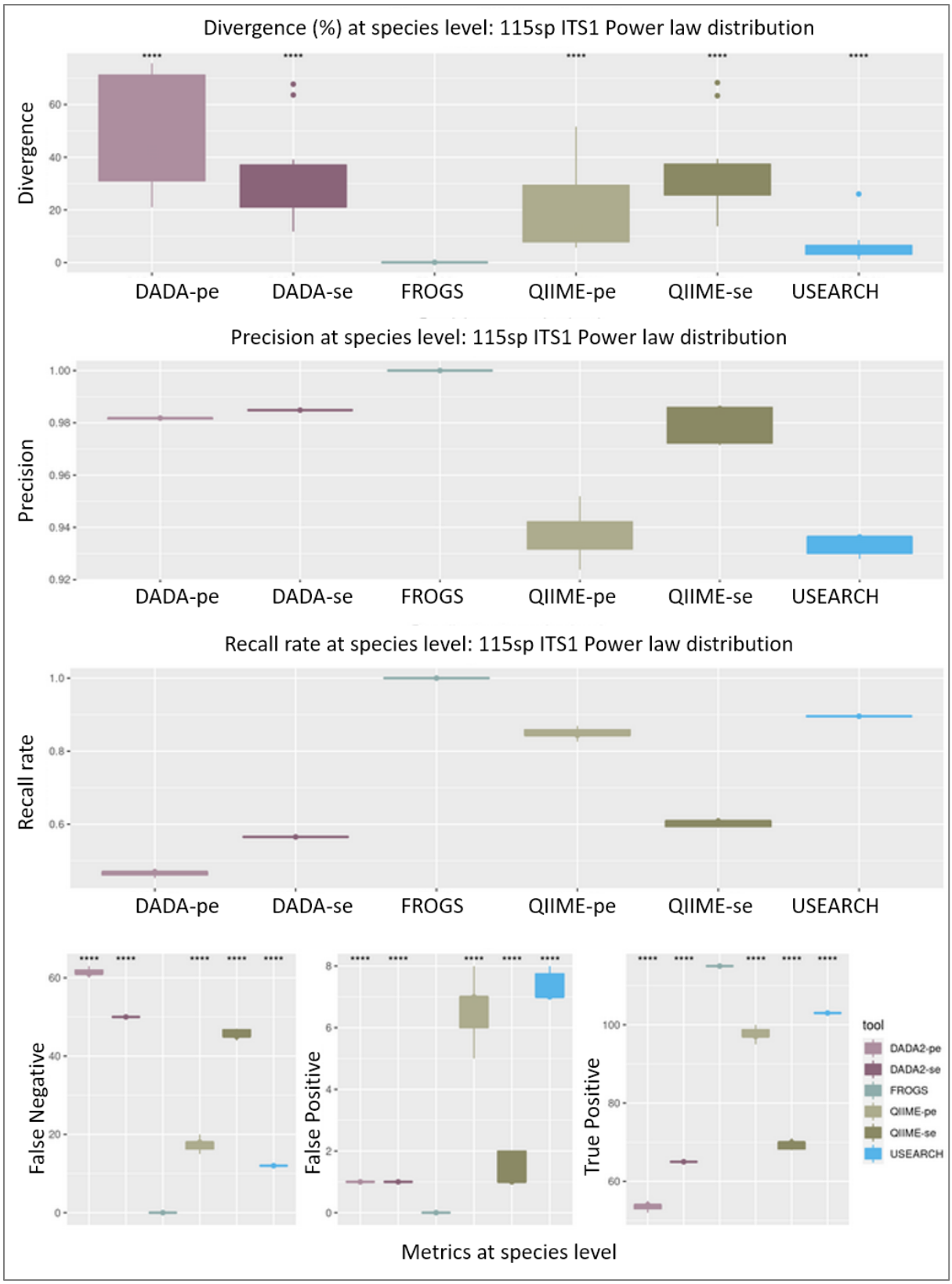


Supplemental figure 2: *Boxplot showing the overall precision (blue) and recall rate (orange) of each of the tools used to process test datasets (all datasets: 35, 115 and 515 species, ITS1 and ITS2, power law and uniform abundance distribution).*



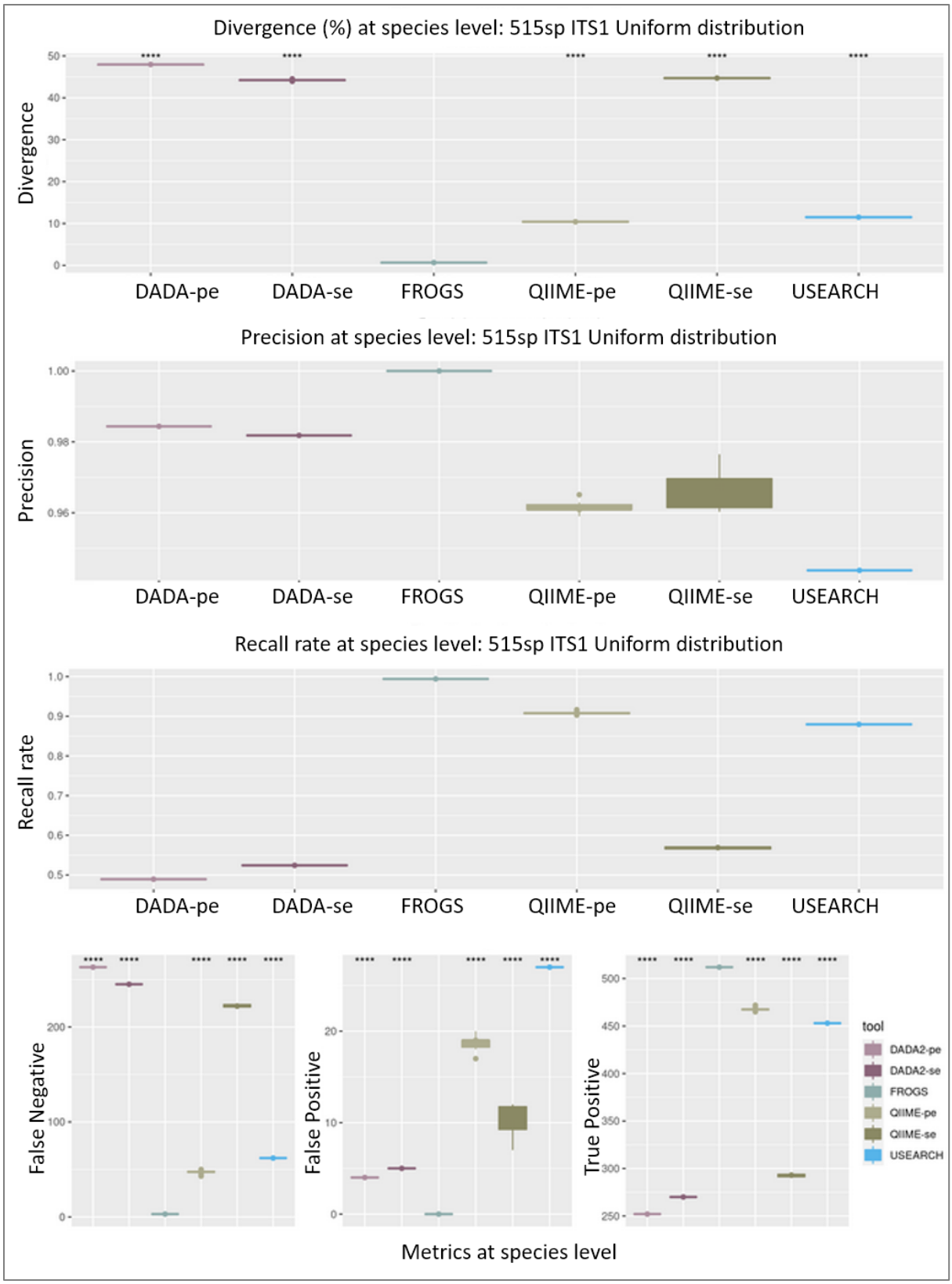
Supplemental figure 3*: Boxplots of the results of the processing of the synthetic datasets (**115 species**, ITS1, **uniform** abundance distribution) of the 4 tools tested (DADA2 paired-end (pe), DADA2 single-end (se) (pale and dark purple), FROGS (bluish green), QIIME-pe, QIIME-se (pale and dark khaki) and USEARCH (cyan)). Four metrics (Abundance divergence, False Positive,*

False Negative and True Positive rates) are calculated in relation to the expected. Each tool was compared to FROGS using a Wilcoxon signed-rank test, ****: p <= 0.0001; ***: p <= 0.001; **: p <= 0.01; *: p <= 0.05; ns: p > 0.05). *Affiliations and associated abundances are taken into account. (N.B. all results on ITS2 datasets are comparable).* See complete results on: http://frogs.toulouse.inrae.fr/ITS/frogs-its.html#Metrics_calculated_in_relation_to_the_expected19



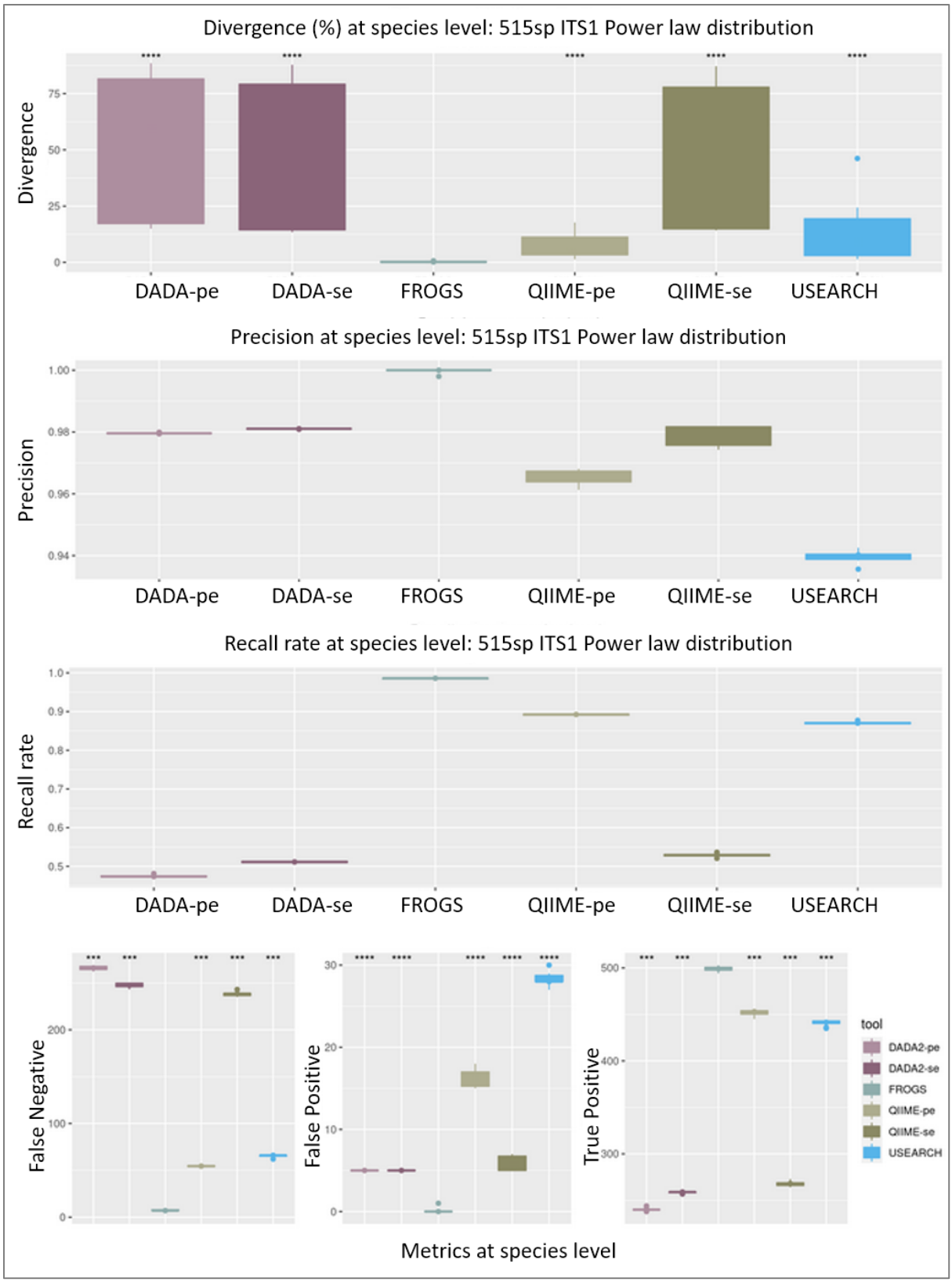
***Supplemental figure 4**: Boxplots of the results of the processing of the synthetic datasets (**115 species**, ITS1, **power law** abundance distribution) of the 4 tools tested (DADA2 paired-end (pe), DADA2 single-end (se) (pale and dark purple), FROGS (bluish green), QIIME-pe, QIIME-se (pale and dark khaki) and USEARCH (cyan)). Four metrics (Abundance divergence, False Positive,*

False Negative and True Positive rates) are calculated in relation to the expected. Each tool was compared to FROGS using a Wilcoxon signed-rank test, ****: p <= 0.0001; ***: p <= 0.001; **: p <= 0.01; *: p <= 0.05; ns: p > 0.05). *Affiliations and associated abundances are taken into account. (N.B. all results on ITS2 datasets are comparable).* See complete results on: http://frogs.toulouse.inrae.fr/ITS/frogs-its.html#Metrics_calculated_in_relation_to_the_expected19



***Supplemental figure 5**: Boxplots of the results of the processing of the synthetic datasets (**515 species**, ITS1, **uniform** abundance distribution) of the 4 tools tested (DADA2 paired-end (pe), DADA2 single-end (se) (pale and dark purple), FROGS (bluish green), QIIME-pe, QIIME-se (pale and dark khaki) and USEARCH (cyan)). Four metrics (Abundance divergence, False Positive,*

False Negative and True Positive rates) are calculated in relation to the expected. Each tool was compared to FROGS using a Wilcoxon signed-rank test, ****: p <= 0.0001; ***: p <= 0.001; **: p <= 0.01; *: p <= 0.05; ns: p > 0.05). *Affiliations and associated abundances are taken into account. (N.B. all results on ITS2 datasets are comparable).* See complete results on: http://frogs.toulouse.inrae.fr/ITS/frogs-its.html#Metrics_calculated_in_relation_to_the_expected28

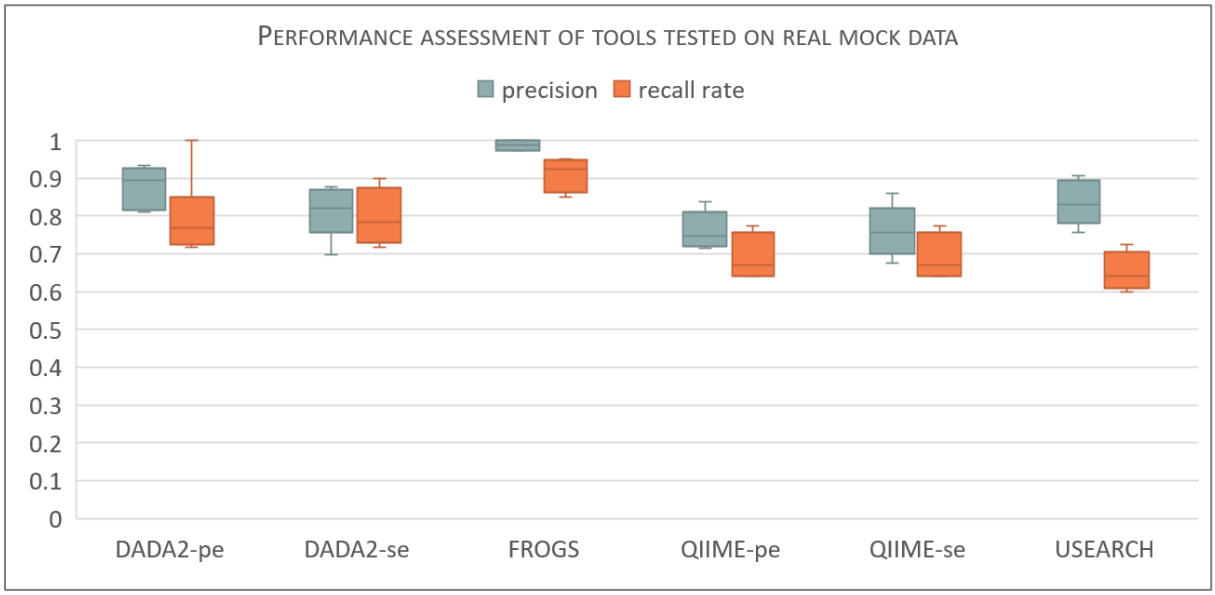


***Supplemental figure 6**: Boxplots of the results of the processing of the synthetic datasets (**515 species**, ITS1, **power law** abundance distribution) of the 4 tools tested (DADA2 paired-end (pe), DADA2 single-end (se) (pale and dark purple), FROGS (bluish green), QIIME-pe, QIIME-se (pale and dark khaki) and USEARCH (cyan)). Four metrics (Abundance divergence, False Positive,*

False Negative and True Positive rates) are calculated in relation to the expected. Each tool was compared to FROGS using a Wilcoxon signed-rank test, ****: p <= 0.0001; ***: p <= 0.001; **: p <= 0.01; *: p <= 0.05; ns: p > 0.05). *Affiliations and associated abundances are taken into account. (N.B. all results on ITS2 datasets are comparable).* See complete results on: http://frogs.toulouse.inrae.fr/ITS/frogs-its.html#Metrics_calculated_in_relation_to_the_expected28

4. REAL MOCK DATASET FOR ASSESSMENT OF THE FROGS ITS METHODOLOGY

For real datasets, we used MOCK samples of fermented meats generated in the METABARFOOD project (INRAE project aimed at developing standard metabarcoding and metagenomic analysis of fermented foods) (PRJNA685292: SAMN17081516 to SAMN17081531). DNA mocks correspond to equimolar mixtures of genomic DNA from 40 representative meat microbiota species. PCR mocks correspond to equimolar mixtures of PCR products for each target. For each sample, ITS1 (ITS1F 5'-CTTGGTCATTTAGAGGAAGTAA-3'; ITS2 5'-GCTGCGTTCTTCATCGATGC-3'), and ITS2 (ITS3 5'-GCATCGATGAAGAACGCAGC-3'; ITS4Kyo 5'-TCCTCCGCTTWTTGWTWTGC-3') regions were amplified. Sequencing was then performed using the Illumina MiSeq sequencing technology (2x250 bp paired-end reads).



***Supplemental figure 7**: 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).*

5. NOTES ON SUPPLEMENTAL TABLE

Supplemental table is joined to the document. This document contains two tables of data from the "simulated data" and the "biological data" used to test the tools and which made it possible to produce the *supplemental figure 2 and 7*. Respectively, the captions of these tables are:

"Companion table of the supplemental figure 2, indicating the performance (divergence rate, true positive, false negative counts and precision and recall rates) of the different methods (DADA2-pe, DADA2-se, FROGS, QIIME-pe, QIIME-se, USEARCH) used to process test datasets (all datasets: 35, 115 and 515 species, ITS1 and ITS2, power law and uniform abundance distribution). "

And

"Companion table of the supplemental figure 7, indicating the performance (divergence rate, true positive, false negative counts and precision and recall rates) of the different methods (DADA2-pe, DADA2-se, FROGS, QIIME-pe, QIIME-se, USEARCH) used to process meat datasets (ITS1 and ITS2 with ADN or PCR products)."

REFERENCE:

1. Angly FE, Willner D, Rohwer F et al. Grinder: a versatile amplicon and shotgun sequence simulator, Nucleic Acids Res 2012;40:e94.