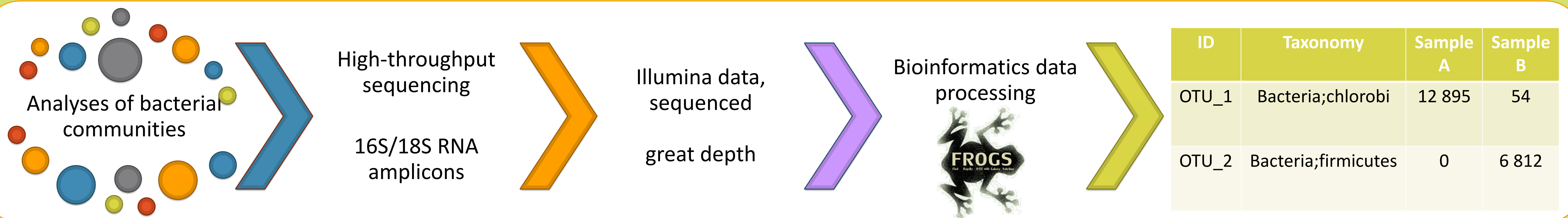


FROGS: Find Rapidly OTUs with Galaxy^[5] Solution

Frédéric ESCUDIE*, Lucas AUER*, Maria BERNARD, Laurent CAUQUIL, Katia VIDAL, Sarah MAMAN, Mahendra MARIADASSOU, Guillermina HERNANDEZ-RAQUET, Géraldine PASCAL

*These authors have contributed equally to the present work.

Contact: geraldine.pascal@toulouse.inra.fr



User-friendly

+

Accuracy

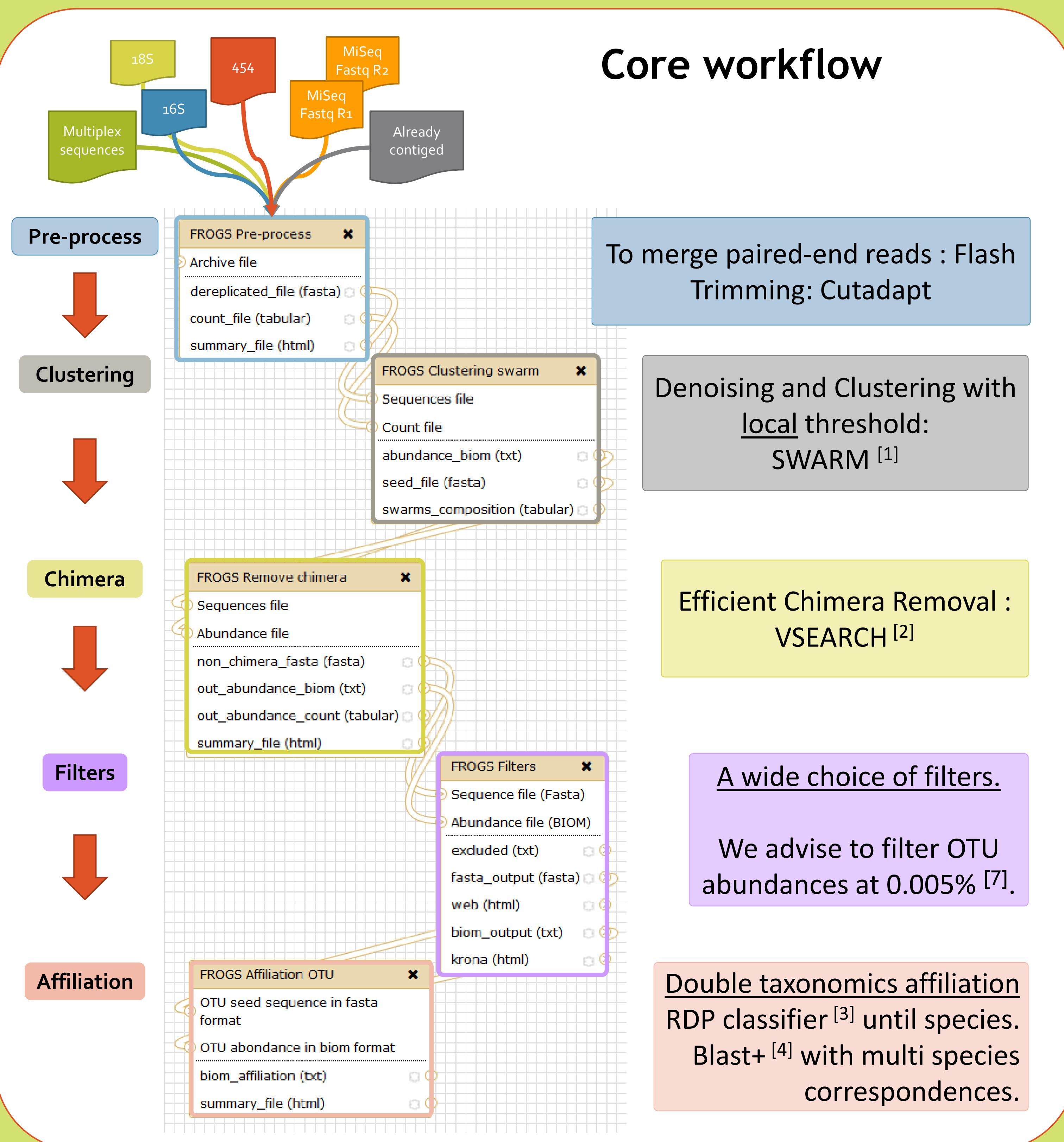
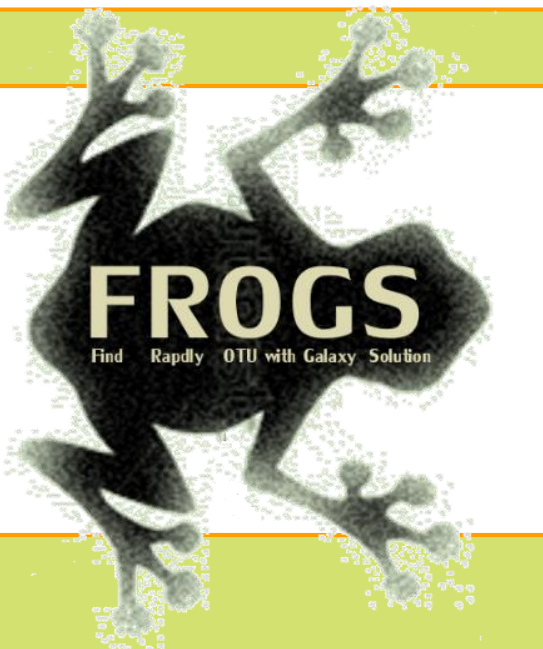
+

Speed

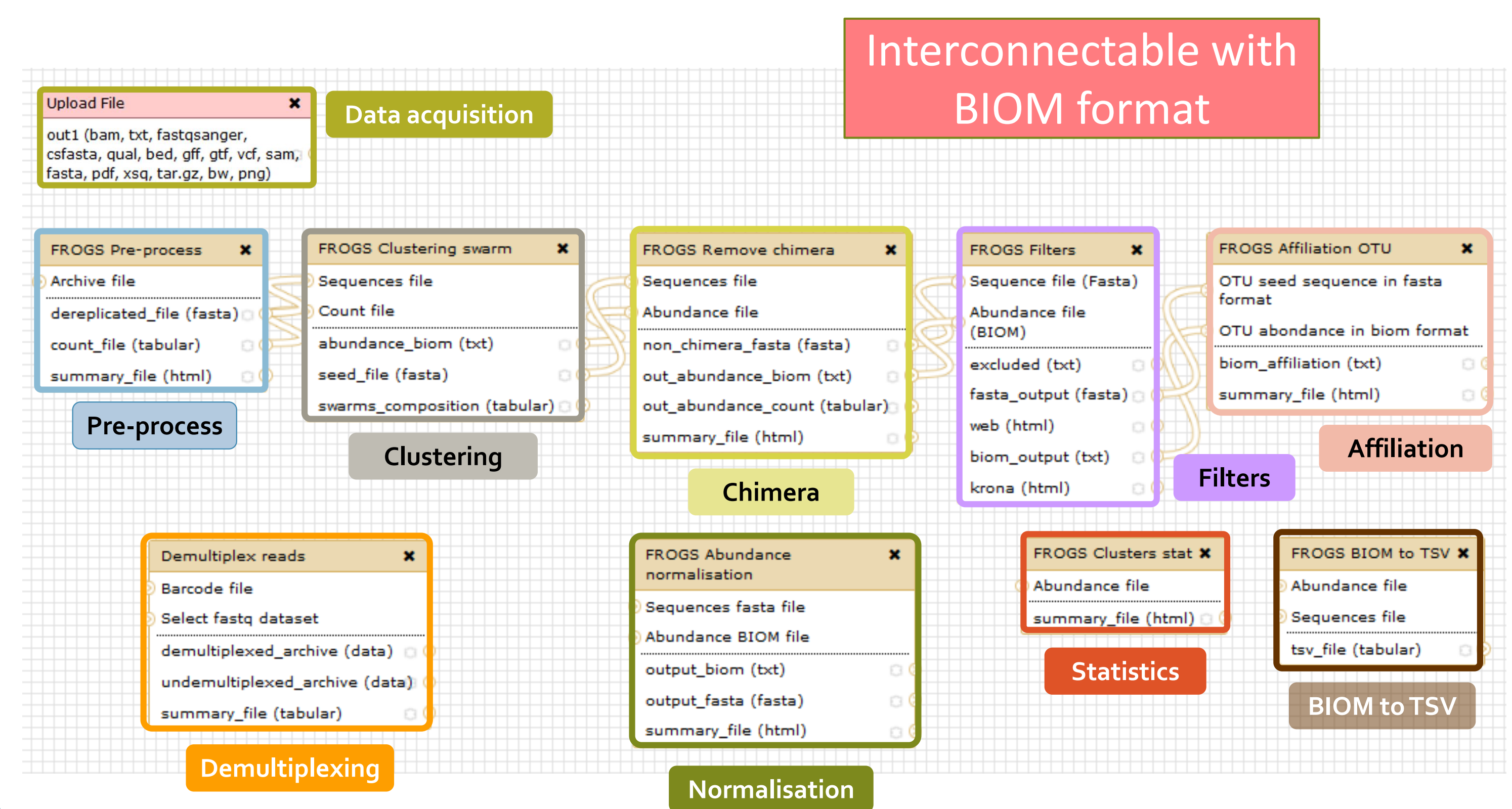
+

Scalability

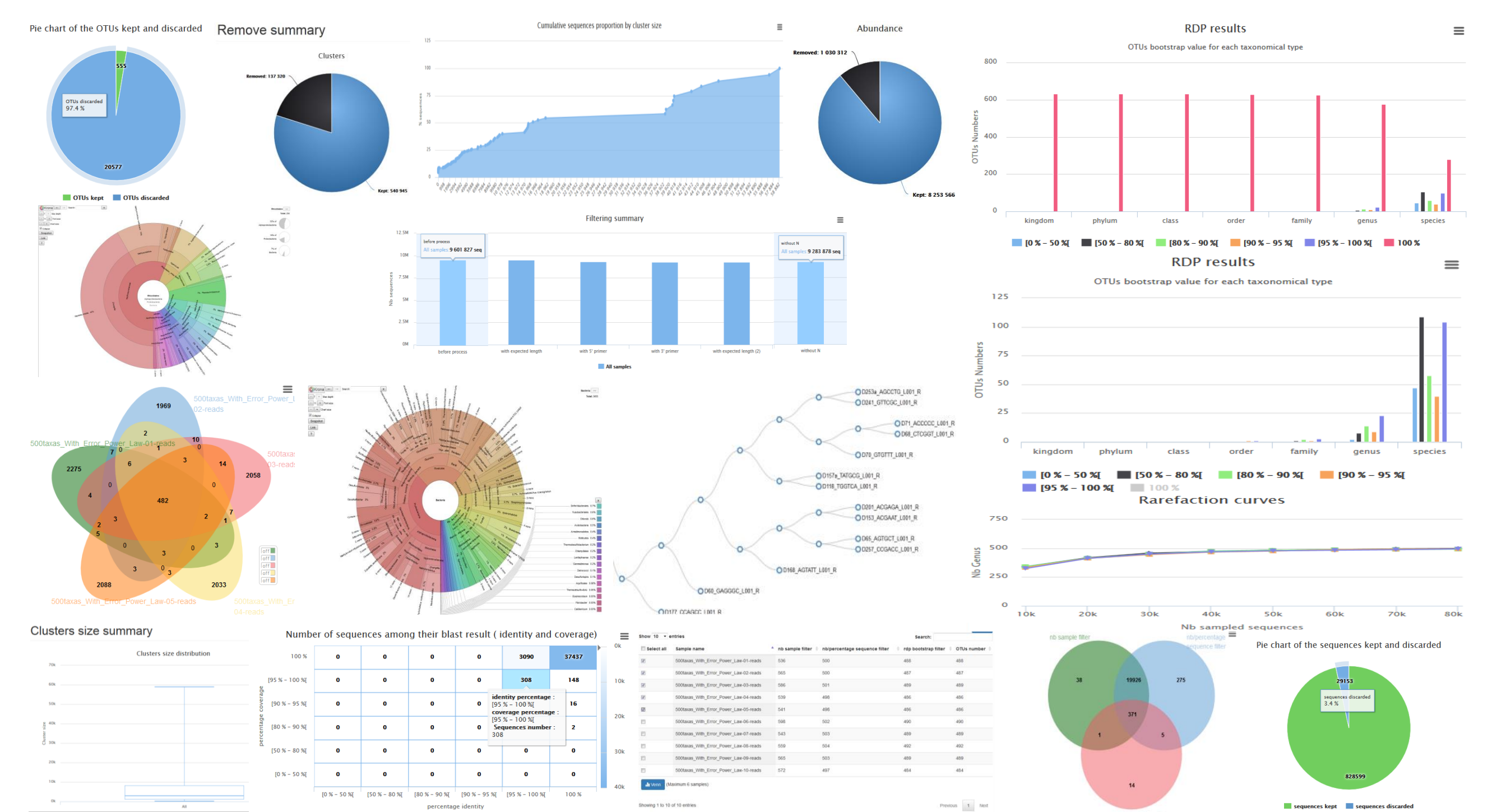
2



FROGS = complete analyze

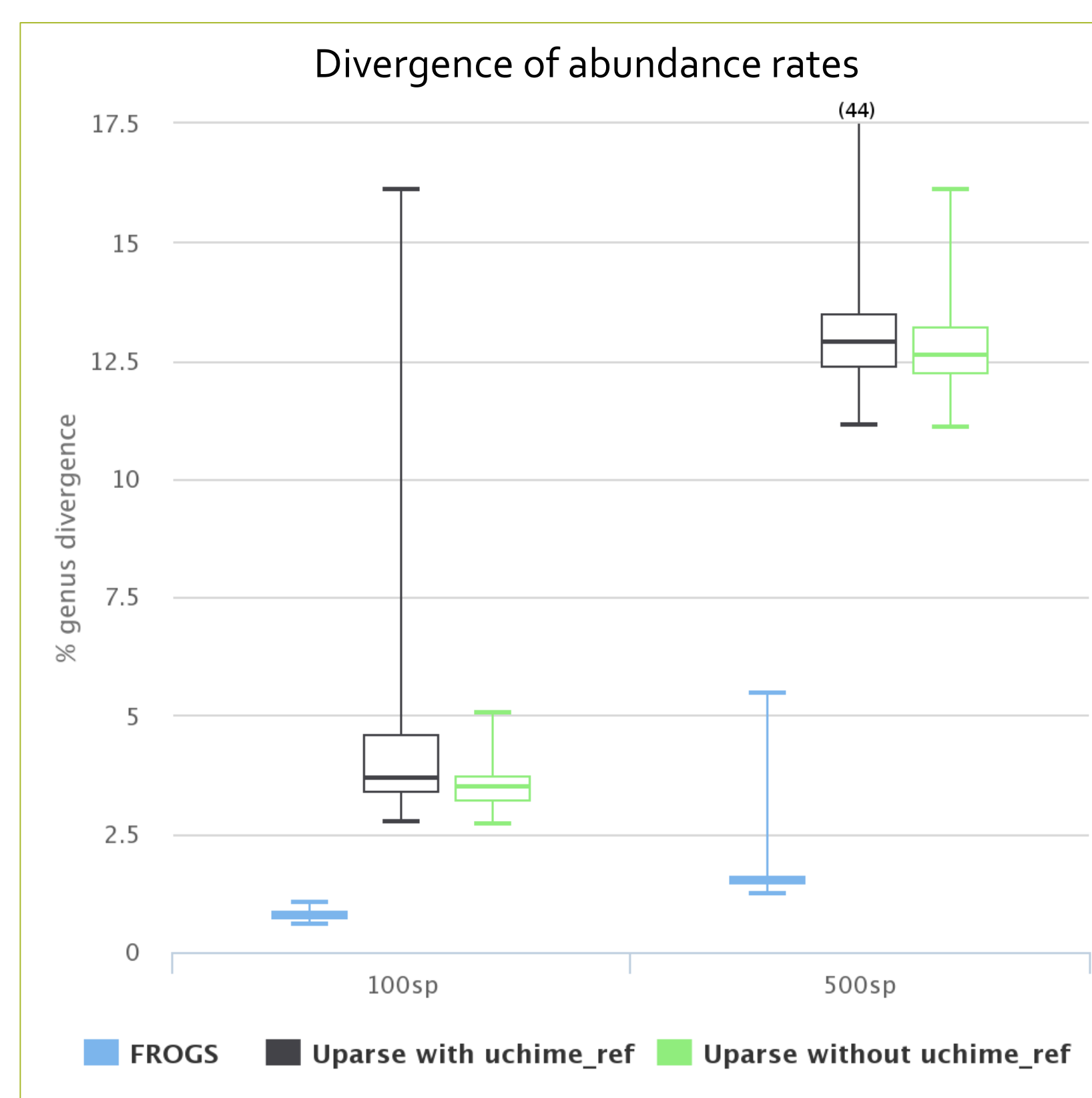


Great graphics outputs



Accuracy on simulated data set

For the tests we use two sets of species manually extracted from LTP 115. These species were chosen to obtain near and distant species in the same dataset. For each set of species we generate 5 artificial runs of 10 samples. In each sample the abundance of species following a power law. The ranking of species abundance is randomly determined in each sample. Then, an Illumina error profile is applied on reads and 20 % of chimera are added. All the runs are processed after FROGS preprocess by FROGS and UPARSE [6].



Divergences on expected affiliations		
	FROGS guidelines*	UPARSE*
	500sp	500sp
Kingdom	0.00	0.00
Phylum	0.43	7.64
Class	0.65	10.11
Order	0.88	11.17
Family	1.12	12.12
Genus	1.52	12.61
Species	5.24	15.18

* in percentage

FROGS guidelines ⁺			UPARSE ⁺	
OTUs	Observed	Missing	Observed	Missing
500 sp.				
Median	553	8	548	10
100 sp.				
Median	124	0	126	0

* in OTU number

Speed on real datasets

9 600 000 sequences of a complete MiSeq run

Preprocess : 9 300 000 sequences	16 min
----------------------------------	--------

Swarm clustering : 680 000 clusters 7 hours

Chimera removal : 540 000 non-chimeric cl.

Small OTUs filtering : 20 000 OTUs

PhiX removal ~8 min

RDP affiliation ~20 min

Blast affiliation ~40 min

FROGS : 1400 OTUs

~ 8 hours

Upcoming

- Evaluate FROGS on others metrics and datasets (mock community, real already known community).
- Add multi-best-hit information with blast (different species with the same amplicon sequence).
- Add FROGS in the toolshed and open a github repository.
- Add databases for affiliation (greengenes, ITS).

References

- [1] Mahé & co. (2014) Swarm: robust and fast clustering method for amplicon-based studies.
- [2] VSEARCH GitHub repository, doi 10.5281/zenodo.15524.
- [3] Wang & co. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.
- [4] Altschul & co. (1990) Basic local alignment search tool.
- [5] Goecks & co. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.
- [6] Edgar. (2013) UPARSE: Highly accurate OTU sequences from microbial amplicon reads.
- [7] Bokulich (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing