

# CONSTAX2: improved taxonomic classification of environmental DNA markers

Julian Liber<sup>1,2\*</sup>, Gregory Bonito<sup>3,4</sup>, Gian Maria Niccolò Benucci<sup>3,4</sup>

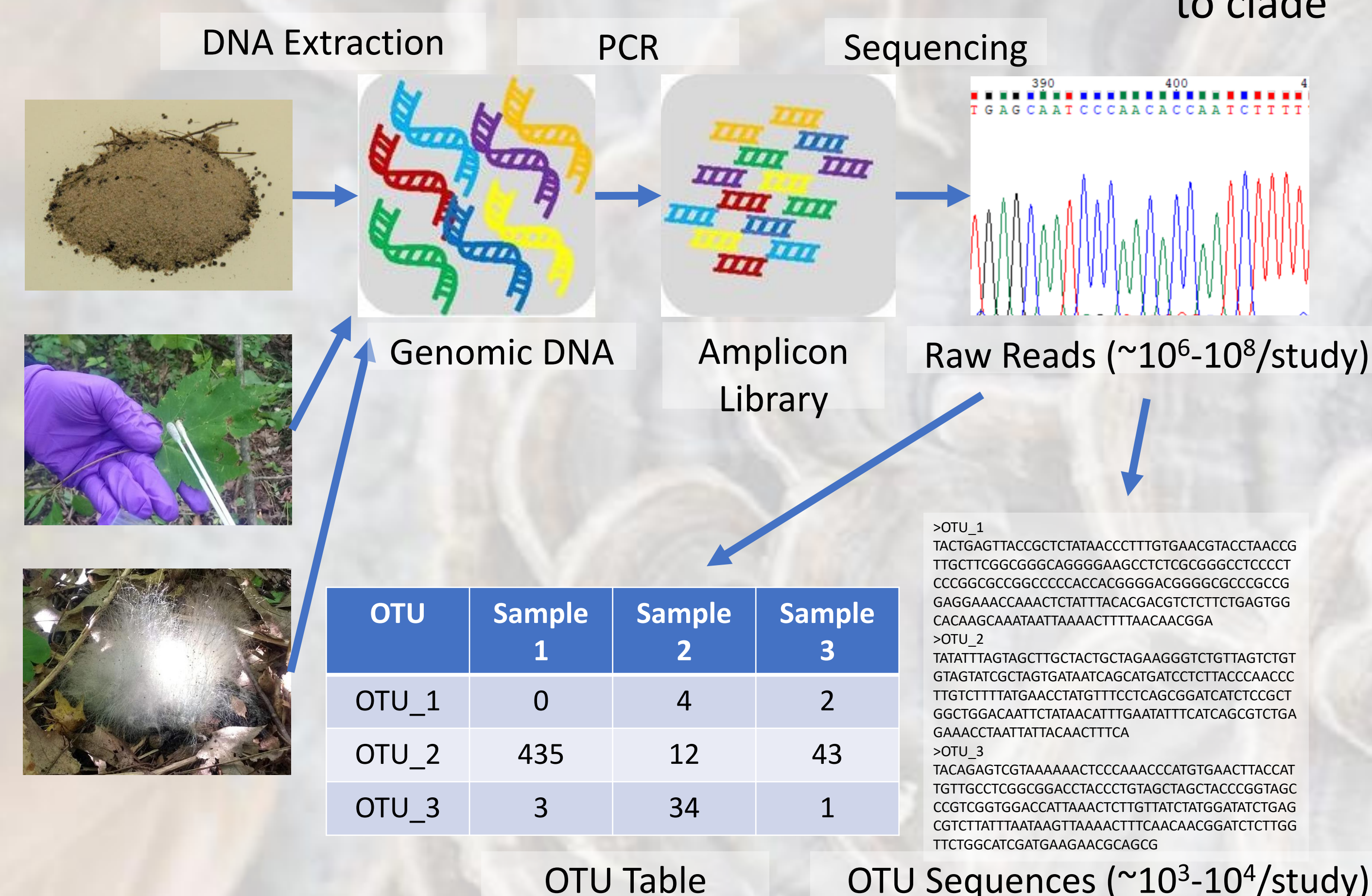
<sup>1</sup>Department of Plant Biology, Michigan State University; <sup>2</sup>Department of Biology, Duke University; <sup>3</sup>Department of Plant Soil and Microbial Sciences, Michigan State University; <sup>4</sup>Great Lakes Bioenergy Research Center, Michigan State University

\*Send correspondence to [liberjul@msu.edu](mailto:liberjul@msu.edu)

[@LiberJulian](https://twitter.com/LiberJulian) on Twitter; [liberjul.github.io](https://github.com/liberjul)

## Metabarcoding Studies

“Barcode”  
sequence  
(16S/ITS) unique  
to clade



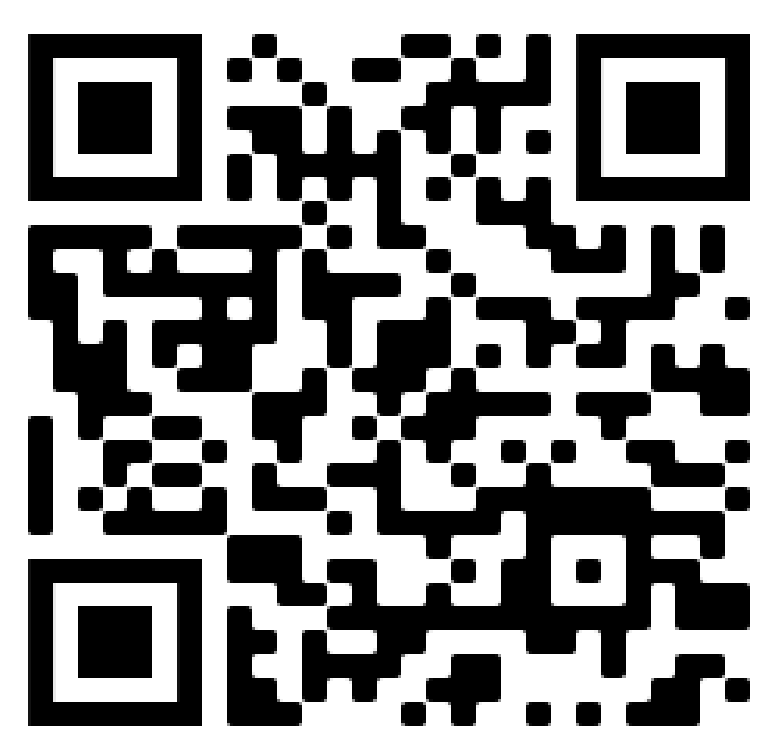
OTU = Operational Taxonomic Unit; a stand-in for species

## CONSTAX Development

- What is taxonomic assignment for?
  - Gives a name to an observed population of organisms
  - Allows for connection to literature
  - Communicates reliable information about the organism
- CONSTAX originally developed by Gdanetz et al. (2017)
  - Designed for fungal studies
  - CONsensus TAXonomy (CONS-TAX) – combines multiple classifier results to improve accuracy
- Goals for CONSTAX2
  - Easy installation and usage (conda-installable, tutorials, documentation)
  - Applicable to fungi, bacteria, archaea, and other organisms
  - Improve accuracy and speed
  - Use current/non-legacy software tools
  - Compare sequences to isolates (cultured/voucher representatives) and determine high-level taxonomy (remove organelle or host sequences)

## Documentation and Use

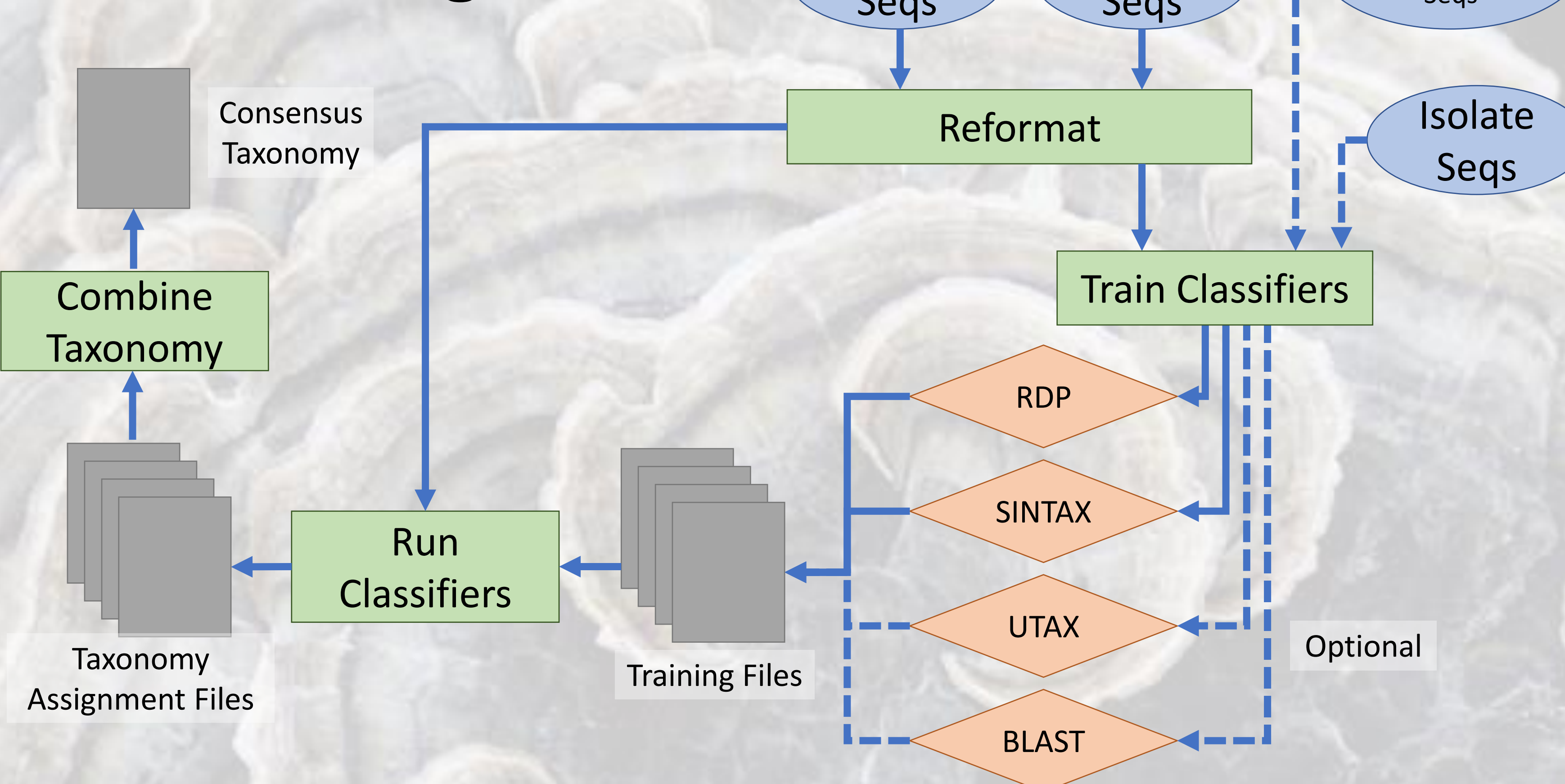
- Available installation from the Bioconda installation channel
- Tutorials and documentation on ReadTheDocs
- Issues and feedback via GitHub



BIOCONDA

<https://github.com/liberjul/CONSTAXv2>  
<https://constax.readthedocs.io/>

## CONSTAX Algorithm



## Results

- ~9-fold increase in training speed
  - Allows for use of SILVA dataset (20x larger than UNITE)
  - Trains faster per sequence at larger database sizes
- About half the classification speed, but allows for parallelization
- Fewer errors per query for CONSTAX2 (balance of over- and misclassification) than other classifiers
  - Consistent across regions, partition level, dataset
  - “Conservative” rule requires two classifiers to agree for assignment – reduces over-classification but can increase misclassification

## Conclusions

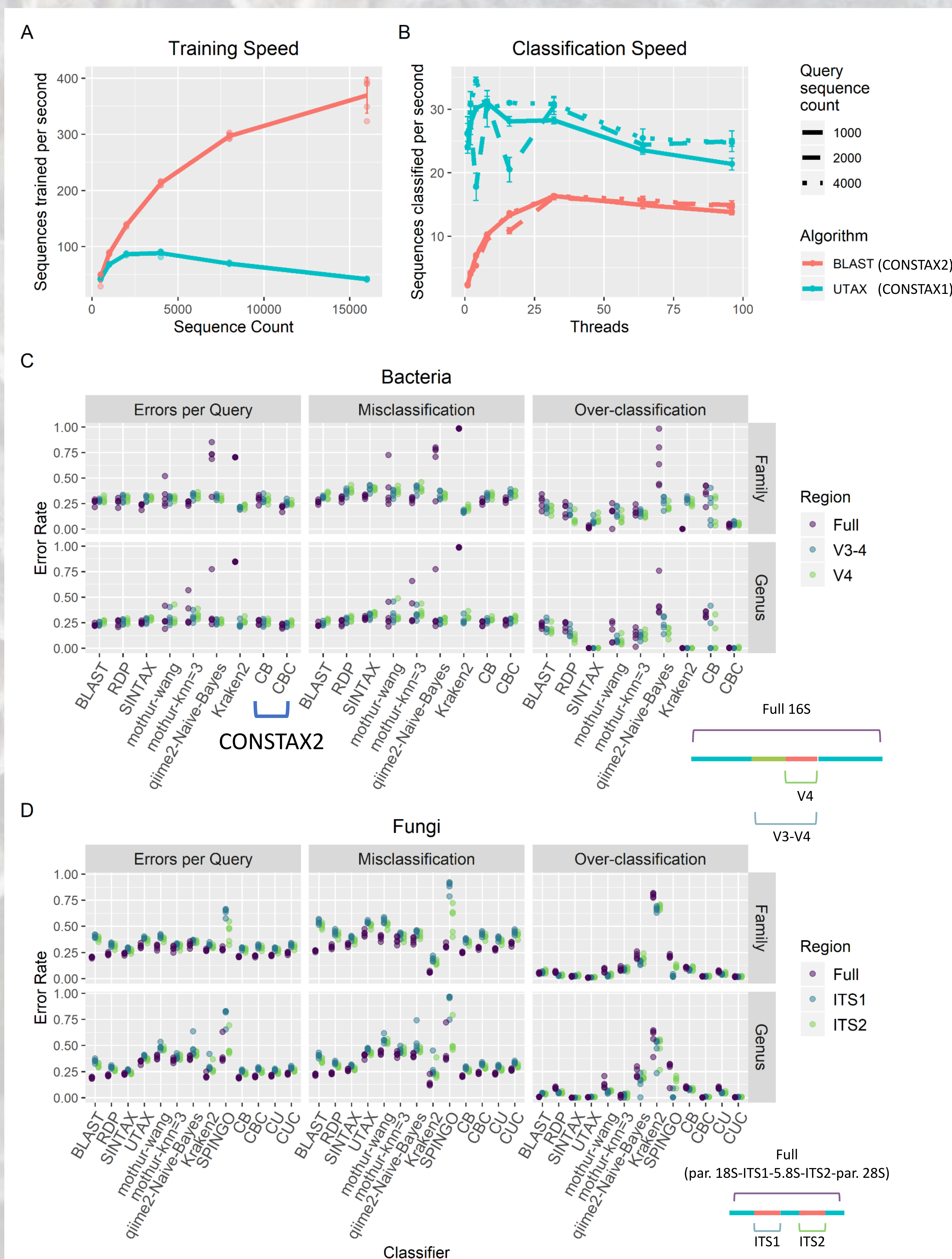
- Compared to CONSTAX1, CONSTAX2 shows
  - Greatly improved training speed
  - New ability to classify bacteria and other groups
  - Improved accuracy and simple use
  - New isolate-matching and high-level taxonomy
- Improvements and features to still make
  - Runtime and memory requirements
  - Webserver hosting

## Acknowledgements

- We thank Zachary Noel, Reid Longley, Acer VanWallendael, and Shay Shemanski for helping test the software. We thank Natalie Vande Pol for assistance in the version transition.
- This work was supported by the US National Science Foundation DEB 1737898 to GB and JL and through the Great Lakes Bioenergy Research Center, U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under award number DE-SC0018409 to GB and GMNB.

## References

- Liber JA, Bonito G, Benucci GMN (2021) CONSTAX2: improved taxonomic classification of environmental DNA markers. *Bioinformatics* doi: 10.1093/bioinformatics/btab347
- Gdanetz K, Benucci GMN, Vande Pol N, Bonito G (2017) CONSTAX: a tool for improved taxonomic resolution of environmental fungal ITS sequences. *BMC Bioinformatics* 18:538 doi 10.1186/s12859-017-1952-x



CB - CONSTAX with BLAST, CBC - CONSTAX with BLAST and conservative rule, CU - CONSTAX with UTAX, CUC - CONSTAX with UTAX and conservative rule.

