

1 **CONSTAX2: Improved taxonomic classification of environmental DNA markers**

2 Julian Liber^{1*}, Gregory Bonito^{2,3}, Gian Maria Niccolò Benucci^{2,3}

3 Liber ORCID: 0000-0002-9941-8268

4 Bonito ORCID: 0000-0002-7262-8978

5 Benucci ORCID: 0000-0003-1589-947X

6

7 **(*) Corresponding author:** liberjul@msu.edu

8

9 **Affiliations:**

10 ¹Department of Plant Biology, Michigan State University, 612 Wilson Rd. East Lansing, MI, USA
11 48824

12 ²Department of Plant Soil and Microbial Sciences, Michigan State University, 1066 Bogue St.
13 East Lansing, MI, USA 48824

14 ³Great Lakes Bioenergy Research Center, Michigan State University, 1129 Farm Ln. East
15 Lansing, MI, USA 48824

16

17 **Summary**

18 CONSTAX - the CONSensus TAXonomy classifier - was developed for accurate and
19 reproducible taxonomic annotation of fungal rDNA amplicons and is based upon a consensus
20 approach of RDP, SINTAX and UTAX algorithms. CONSTAX2 can be used to classify
21 prokaryotes and incorporates BLAST-based classifiers to reduce classification errors.

22 Additionally, CONSTAX2 implements a conda-installable, command line tool with improved
23 classification metrics, faster training, multithreading support, capacity to incorporate external
24 taxonomic databases, new isolate matching and high-level taxonomy tools, replete with
25 documentation and example tutorials.

26

27 **Availability and Implementation**

28 CONSTAX2 is available at <https://github.com/liberjul/CONSTAXv2>, and is packaged for Linux
29 and MacOS from Bioconda. A tutorial and documentation are available at
30 <https://constax.readthedocs.io/en/latest/>.

31

32 **Introduction**

33 High-throughput sequencing has revolutionized metagenomics and microbiome sciences (Di
34 Bella *et al.*, 2013). These culture-independent methods have revealed previously unrecognized
35 microbial diversity and has allowed researchers to detect organisms occurring at extremely low
36 abundances (Brown *et al.*, 2015). Amplicon-based sequencing, which relies on amplification and
37 sequencing of conserved genetic markers such as the rRNA operon or protein-coding genes, is
38 an extremely popular technique for studying microbiomes and microbial communities. Following
39 sequencing, quality control, and demultiplexing, amplicon reads are clustered and
40 representative sequences are classified taxonomically. Many algorithms have been developed
41 to conduct the task of assigning taxonomy to environmental sequences. Some of the most
42 popular include BLAST-based tools (Altschul *et al.*, 1990; Bokulich *et al.*, 2018), the Ribosomal
43 Database Project (RDP) naive Bayesian classifier (Wang *et al.*, 2007), and the USEARCH
44 algorithms SINTAX (Edgar, 2016) and UTAX (Edgar, 2013).

45 While each of these tools can be implemented independently to assign taxonomy, a consensus-
46 based approach was demonstrated to increase the number of sequences with taxonomic
47 assignments (Gdanetz *et al.*, 2017). Since the original release of the CONSTAX classifier, we
48 have realized the need for improved ease of use, updated software compatibility, simpler
49 installation, improved accuracy and adaptability, and application to bacteria or other organisms.
50 To address these needs, an updated version, CONSTAX2, has been developed.

51 **Implementation**

52 CONSTAX2 (referred to hereafter as “CONSTAX”) is released as a conda-installable command-
53 line tool, available from the bioconda installation channel (Grüning *et al.*, 2018) for LinuxOS,
54 MacOS, and WSL systems. It is installed with the command “conda install -c bioconda constax”,
55 see <https://github.com/liberjul/CONSTAXv2>. CONSTAX requires two files: 1) “-i, --input” a
56 database file in FASTA format with header lines containing taxonomy of the sequences in
57 SILVA (Glöckner *et al.*, 2017) or UNITE (Nilsson *et al.*, 2019) style, and 2) “-d, --db” an input file
58 of user-submitted sequences in FASTA format. This version implements a BLAST classification
59 algorithm instead of the legacy UTX classifier if the “-b, --blast” flag is used.

60 The user may designate several additional parameters, including confidence threshold for
61 assignment (“-c, --conf”), BLAST classifier parameters, and whether to use a conservative
62 consensus rule (“--conservative”), which requires agreement of two (instead of one) non-null
63 assignments to assign a taxonomy at the given rank. CONSTAX offers multithreaded
64 classification with the argument, “-n, --num_threads”.

65 CONSTAX generates three directories while running: 1) training files directory (“-f, --trainfile”),
66 taxonomy assignments directory (“-x, --tax”), and an output directory (“-o, --output”). Prior to
67 classifying sequences, training must be performed on any newly used database file with the “-t,
68 --train” flag. After initial training, generated training files can be used in any later run by
69 designating the same training files directory. When training is performed, CONSTAX will
70 automatically generate formatted database files required by each classifier, as long as the
71 supplied database has SILVA or UNITE header formatting. Following training, the classification
72 or search command is performed for each classifier, and files are output to the taxonomic
73 assignments directory. Finally, each classification output is reformatted into a standard format
74 and used to generate a consensus hierarchical taxonomy, and stored in the output directory as
75 tab-separated value files.

76 CONSTAX2 offers two additional features: 1) the ability to match input sequences to isolates
77 using the "--isolates" option; and 2) the ability to determine higher-level taxonomy using
78 representative databases with the "--high_level_db" option. Both approaches implement the
79 BLAST algorithm to associate input sequences with hits from the respective databases,
80 returning a single best hit. Cutoffs for query coverage and percent identity can be specified.
81 Isolate matching streamlines culture-dependent and culture-independent analyses, and can also
82 be used to implicate potential contamination by association with known isolates previously
83 worked with in the laboratory or sequencing facility where the samples were processed. Higher-
84 level taxonomy designations are also useful in filtering host, organelles, or non-target taxa,
85 which may show up in rDNA surveys. For 16S rDNA prokaryote datasets the SILVA NR99
86 database is recommended, while the latest UNITE Eukaryotes database is recommended for
87 ITS studies of Fungi.

88 **Results**

89 *Algorithm speed*

90 The implementation of the BLAST algorithm as a third classifier and replacement of UTAX
91 provides crucial speedup of the training step (Fig 1A), facilitating the use of the much larger
92 SILVA database. For 16,000 sequences randomly sampled from the SILVA database, the
93 BLAST implementation (including SINTAX, RDP, and BLAST) trained 370 ± 32.1 sequences * s⁻¹
94 (mean \pm SD), while the UTAX implementation (including SINTAX, RDP, and UTAX) trained
95 41.9 ± 0.911 sequences * s⁻¹, an approximately 9-fold improvement. Furthermore, the BLAST
96 implementation trains faster per sequence at larger database sizes.

97 Although the BLAST implementation is faster for training, classification is faster with the UTAX
98 implementation (Fig 1B). The maximum classification speed was achieved at 32 threads for the
99 BLAST implementation and between 4 and 8 threads for the UTAX implementation, depending
100 on the number of query sequences classified, which minorly affected per-sequence rates. At

101 4000 query sequences, the BLAST implementation classified at a speed of 16.349 ± 0.298
102 sequences * s^{-1} on 32 threads, while the UTX implementation classified at a speed of
103 34.449 ± 0.611 sequences * s^{-1} on 4 threads.

104 *Algorithm performance*

105 Clade partitioned cross-validation and classification metrics from SINTAX (Edgar, 2016) were
106 used (Supplementary Data) on each of the classifiers and consensus taxonomy assignments
107 were compared for genus and family partitions as well as for full length ITS1-5.8S-ITS2 or 16S
108 regions (accounting for the commonly used subregions ITS1, ITS2, V4, V3-4) with errors per
109 query, over-classification, and misclassification, for 5 query-reference paired datasets (Fig 1C-
110 D, Table S1). The popular mothur knn and Wang classifiers (Schloss *et al.*, 2009) and qiime q2-
111 feature-classifier plugin (Bokulich *et al.*, 2018) classifiers were compared using the same
112 protocol. CONSTAX with the non-conservative consensus with BLAST had the fewest errors per
113 query (EPQ) for any classifier (0.236-0.248, 95% CI for all regions and partition levels), or tied
114 for fewest with the UTX consensus, across the UNITE dataset. Alternatively, CONSTAX with
115 the conservative consensus with BLAST had the fewest errors for all classifications in the SILVA
116 dataset (EPQ=0.214-0.259). The BLAST implementation was valuable in decreasing
117 misclassifications, but this was generally associated with increased (erroneous) over-
118 classifications.

119 **Conclusion**

120 The newest implementation of CONSTAX offers improvement over its predecessor by ease of
121 use, and improved applicability and accuracy. Hierarchical taxonomy classification accuracy by
122 a consensus approach in CONSTAX2 is demonstrated to outperform commonly used classifiers
123 while remaining computationally feasible.

124 **Acknowledgements**

125 The authors thank Zachary Noel, Reid Longley, Acer VanWallendael, and Shay Shemanski for
126 helping test the software. We thank Natalie Vande Pol for assistance in the version transition.

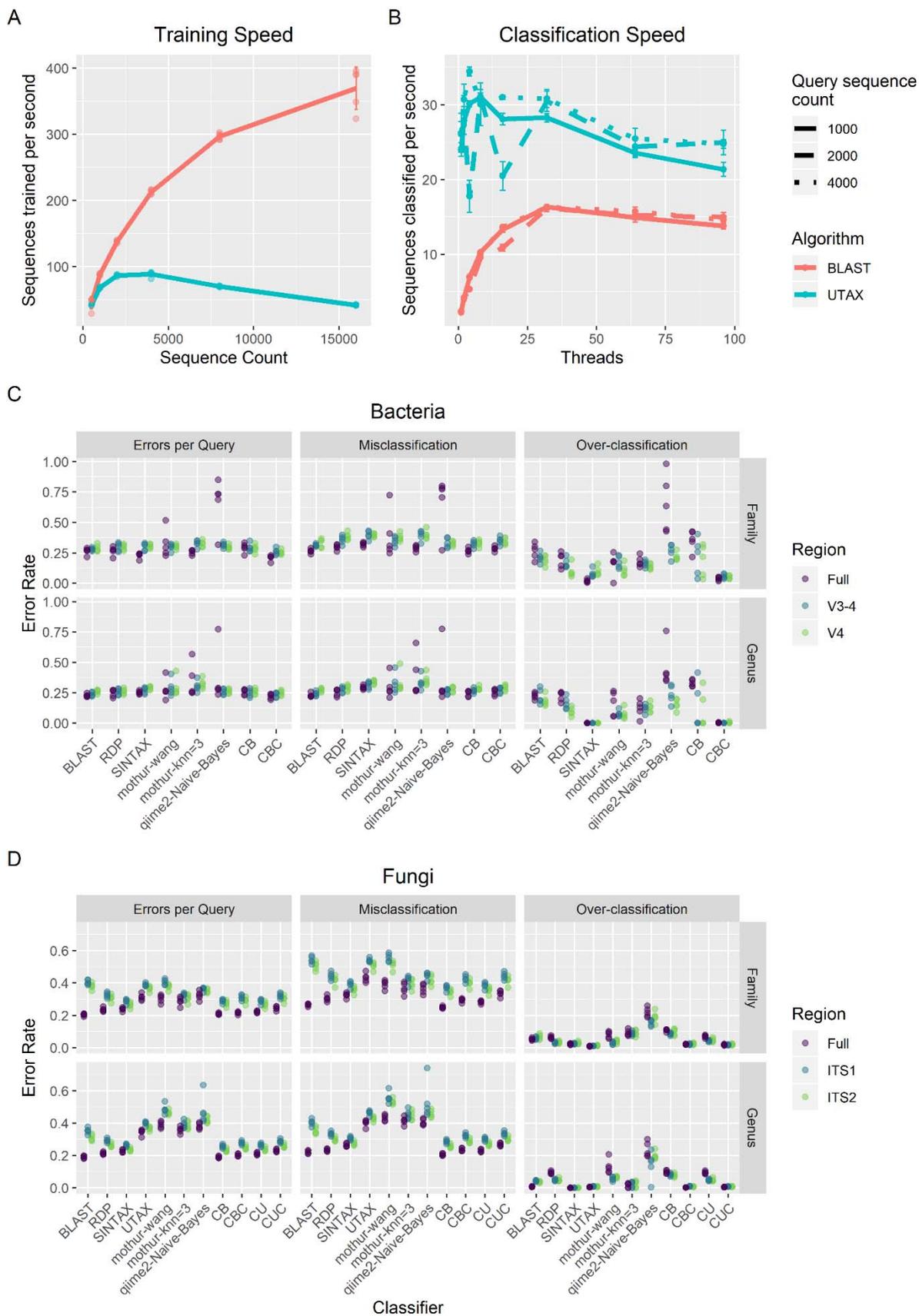
127 **Funding**

128 This work was supported by the US National Science Foundation DEB 1737898 to GB and JL
129 and through the Great Lakes Bioenergy Research Center, U.S. Department of Energy, Office of
130 Science, Office of Biological and Environmental Research, under award number DE-
131 SC0018409 to GB and GMNB.

132 *Conflict of Interest:* none declared.

133 **References**

- 134 Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
135 Bokulich,N.A. *et al.* (2018) Optimizing taxonomic classification of marker-gene amplicon
136 sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, **6**, 90.
137 Brown,S.P. *et al.* (2015) Scraping the bottom of the barrel: are rare high throughput sequences
138 artifacts? *Fungal Ecol.*, **13**, 221–225.
139 Di Bella,J.M. *et al.* (2013) High throughput sequencing methods and analysis for microbiome
140 research. *J. Microbiol. Methods*, **95**, 401–414.
141 Edgar,R.C. (2016) SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS
142 sequences. *bioRxiv*, 074161.
143 Edgar,R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads.
144 *Nat. Methods*, **10**, 996–998.
145 Gdanetz,K. *et al.* (2017) CONSTAX: a tool for improved taxonomic resolution of environmental
146 fungal ITS sequences. *BMC Bioinformatics*, **18**, 538.
147 Glöckner,F.O. *et al.* (2017) 25 years of serving the community with ribosomal RNA gene
148 reference databases and tools. *J. Biotechnol.*, **261**, 169–176.
149 Grüning,B. *et al.* (2018) Bioconda: sustainable and comprehensive software distribution for the
150 life sciences. *Nat. Methods*, **15**, 475–476.
151 Nilsson,R.H. *et al.* (2019) The UNITE database for molecular identification of fungi: handling
152 dark taxa and parallel taxonomic classifications. *Nucleic Acids Res.*, **47**, D259–D264.
153 Schloss,P.D. *et al.* (2009) Introducing mothur: Open-Source, Platform-Independent,
154 Community-Supported Software for Describing and Comparing Microbial Communities.
155 *Appl. Environ. Microbiol.*, **75**, 7537–7541.
156 Wang,Q. *et al.* (2007) Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into
157 the New Bacterial Taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.



159 **Figure 1. Performance of the CONSTAX algorithm.** A) Reference sequences parsed per
160 second for training of the CONSTAX implementation with BLAST and UTAX, as a function of
161 the size of the training set. B) Sequences classified per second with BLAST and UTAX
162 implementations, as a function of query set size and threads used for parallelization. C-D)
163 Classification performance resulting from clade-partition cross-validation, at genus and family
164 partition ranks, for full and extracted regions, corresponding to each CONSTAX classifier and
165 other common classification tools, for C) Bacteria in the SILVA SSURef release 138 dataset and
166 D) Fungi in the UNITE RepS Feb 4 2020 general release. Errors per query, misclassification
167 rate, and over-classification rate are defined by Edgar (2016) and in Supplementary Data. CB -
168 CONSTAX with BLAST, CBC - CONSTAX with BLAST and conservative rule, CU - CONSTAX
169 with UTAX, CUC - CONSTAX with UTAX and conservative rule.