




SOFTWARE TOOL ARTICLE

ITSxpress: Software to rapidly trim internally transcribed spacer sequences with quality scores for marker gene analysis [version 1; peer review: 2 approved]

Adam R. Rivers¹, Kyle C. Weber¹, Terrence G. Gardner ², Shuang Liu², Shalamar D. Armstrong³

¹Genomics and Bioinformatics Research Unit, USDA Agricultural Research Service, Gainesville, FL, 32608, USA

²Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC, 27695, USA

³Department of Agronomy, Purdue University, Purdue, IN, 47907, USA

v1 **First published:** 06 Sep 2018, 7:1418 (<https://doi.org/10.12688/f1000research.15704.1>)
Latest published: 06 Sep 2018, 7:1418 (<https://doi.org/10.12688/f1000research.15704.1>)

Abstract



The internally transcribed spacer (ITS) region between the small subunit ribosomal RNA gene and large subunit ribosomal RNA gene is a widely used phylogenetic marker for fungi and other taxa. The eukaryotic ITS contains the conserved 5.8S rRNA and is divided into the ITS1 and ITS2 hypervariable regions. These regions are variable in length and are amplified using primers complementary to the conserved regions of their flanking genes. Previous work has shown that removing the conserved regions results in more accurate taxonomic classification. An existing software program, ITSx, is capable of trimming FASTA sequences by matching hidden Markov model profiles to the ends of the conserved genes using the software suite HMMER. ITSxpress was developed to extend this technique from marker gene studies using Operational Taxonomic Units (OTU's) to studies using exact sequence variants; a method used by the software packages Dada2, Deblur, QIIME 2, and Unoise. The sequence variant approach uses the quality scores of each read to identify sequences that are statistically likely to represent real sequences. ITSxpress enables this by processing FASTQ rather than FASTA files. The software also speeds up the trimming of reads by a factor of 14-23 times on a 4-core computer by temporarily clustering highly similar sequences that are common in amplicon data and utilizing optimized parameters for Hmsearch. ITSxpress is available as a QIIME 2 plugin and a stand-alone application installable from the Python package index, Bioconda, and Github.



Keywords

Amplicon sequencing, marker gene sequencing, internally transcribed spacer, ITS, trimming, QIIME

Open Peer Review

Reviewer Status  

	Invited Reviewers	
	1	2
version 1 published 06 Sep 2018	 report	 report

- J. Gregory Caporaso** , Northern Arizona University, Flagstaff, USA
- Johanna B. Holm** , University of Maryland School of Medicine, Baltimore, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Adam R. Rivers (adam.rivers@ars.usda.gov)

Author roles: **Rivers AR:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Software, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Weber KC:** Software, Writing – Review & Editing; **Gardner TG:** Resources, Writing – Review & Editing; **Liu S:** Resources; **Armstrong SD:** Resources

Competing interests: No competing interests were disclosed.

Grant information: This research was funded by the United States Department of Agriculture (USDA), Agricultural Research Service (ARS) research project number 6066-21310-005-00-D and computational analysis using SCINet under project 0500-00093-001-00-D. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2018 Rivers AR *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

How to cite this article: Rivers AR, Weber KC, Gardner TG *et al.* **ITSxpress: Software to rapidly trim internally transcribed spacer sequences with quality scores for marker gene analysis [version 1; peer review: 2 approved]** F1000Research 2018, 7:1418 (<https://doi.org/10.12688/f1000research.15704.1>)

First published: 06 Sep 2018, 7:1418 (<https://doi.org/10.12688/f1000research.15704.1>)

Introduction

The internally transcribed spacer (ITS) between the small subunit (SSU/18S) ribosomal RNA gene and the large subunit (LSU/28S) ribosomal RNA gene is a commonly used phylogenetic marker. The Fungal Barcoding Consortium standardized the practice of ITS sequencing by adopting the region for its efforts (Schoch *et al.*, 2012), and the major fungal database UNITE uses the region as well (Köljalg *et al.*, 2013). It is a common practice to amplify the ITS1 or ITS2 region using primers located in the more conserved 18S/5.8S genes or the 5.8S/28S genes. Previous work has shown that leaving these more conserved regions on the ITS sequence creates miss-assignments. In one study of full length ITS sequences, 11% of the time the ITS1 and ITS2 regions matched one reference sequence but the full sequence including ITS1, ITS2 and the 5.8S did not (Nilsson *et al.*, 2009). The software package ITSx was developed and subsequently improved (Bengtsson-Palme *et al.*, 2013; Nilsson *et al.*, 2010) to accurately trim ITS sequences from longer reads. ITSx uses hidden Markov models (HMMs) created for fungi and 17 other groups of eukaryotes to identify the start and stop sites for the ITS region. The software used the HMMER package Hmmscan until version 1.1b when Hmmscan was substituted for increased speed (Eddy, 2011).

ITSxpress was created to extend the capabilities of ITSx from marker gene studies using operational taxonomic units (OTUs) to studies using exact sequence variants. Amplicon sequencing creates sequences with errors. In order to distinguish true sequences from sequencing errors, sequences have been clustered into OTUs by sorting reads by abundance then clustering them in a greedy fashion at a specified percent identity (often 97%). Recently, new methods (e.g. Dada2, Deblur and Unoise) have been published that use statistical models or information theoretic models to identify exact sequence variants that represent true biological sequences (Amir *et al.*, 2017; Callahan *et al.*, 2016; Caporaso *et al.*, 2010; Edgar, 2016). These methods require the error profiles of individual sequences, which requires trimming each FASTQ sequence (Cock *et al.*, 2010) to the ITS region of interest. ITSxpress trims FASTQ files for this purpose.

Methods

Implementation

ITSxpress rapidly merges and trims paired-end FASTQ sequences to the ITS region of interest for the identification of exact sequence variants. The software merges and error-corrects reads using BBMerge (Bushnell *et al.*, 2017). The merged FASTQ reads are then sorted by abundance and clustered by default at 99.5% identity to generate a representative set of sequences using VSEARCH (Rognes *et al.*, 2016). The user may also select dereplication from 98% to 100% identity. These unique sequences are compared to the HMMs used by ITSx version 1.1b (Bengtsson-Palme *et al.*, 2013) using Hmmscan (Eddy, 2011). Read filtering heuristics in Hmmscan are enabled and reports are filtered as well. The start and stop position of each cluster representative is then used to trim each sequence in the cluster and all original FASTQ sequences that could be merged are returned with the ends trimmed. All major steps

(merging, dereplication and Hmmscan) are multithreaded. The source code is version controlled and tested by continuous integration.

Operation

ITSxpress is an open source Python package that can be run on Linux or MacOS systems and does not require any special memory or processor configuration. It is available from Github, Pip, Bioconda and as a plugin for QIIME 2. The QIIME 2 package operates on native QIIME 2 .qza files. A typical workflow for an ITS sequencing project would take a set of paired-end FASTQ forward and reverse sequences and return a FASTQ file with merged, trimmed sequences and a log file. Uncompressed FASTQ or Gzip compressed FASTQ files can be used. The command line version of ITSxpress accepts interleaved, paired-end files, forward and reverse paired end files, and single-ended files. The QIIME 2 plugin version of ITSxpress accepts a .qza QIIME 2 artifact file of the type "PairedEndSequences-WithQuality" or "SequencesWithQuality" that contains one or more samples with single or paired data. It merges (if paired) and trims all samples and returns a QIIME 2 artifact file containing single-ended sequences with quality or paired-end sequences with quality that can be used for sequence variant calling by DADA2 or Deblur (Amir *et al.*, 2017; Callahan *et al.*, 2016).

Testing

To compare the speed and trimming results of ITSxpress, we compared the ITS1 and ITS2 sequences from 15 soil samples collected from the rhizosphere of maize in fields with different winter cover crops. ITS1 reads were amplified using the ITS1F/ITS2 primer set (Gardes & Bruns, 1993; White *et al.*, 1990). ITS2 reads were amplified using the ITS3/ITS4 primer set (White *et al.*, 1990). Reads were multiplexed and sequenced on an Illumina Miseq in 2x300bp run mode using version 3.0 chemistry.

Tests of ITSxpress and ITSx performance were run on single compute nodes with 2 x 10 core Intel Xeon Processors (E5-2670 v2 2.50GHz 25MB cache) with hyper-threading enabled, 128GB DDR3 ECC memory and two Intel DC S3500 Series SATA 6.0Gb/s SSDs. For the first test of trimming speed, 5 replicates were run where 15 ITS1 and 15 ITS2 samples were trimmed using ITSxpress and ITSx with 4 logical cores. Trimming was done using ITSxpress with default settings. ITSx was run with multithreading and heuristic filtering turned on and only the fungal database selected. The running times for ITSx and ITSxpress were plotted on a log scale, Figure 1. The number of total reads in each sample and reads remaining after clustering at 99.5% identity are shown on a log scale, Figure 2.

To compare the performance of ITSxpress and ITSx as computer cores were added, tests were run on the ITS1 and ITS2 sample with the largest numbers of sequences (ITS1: n=100543 16% unique, ITS2: n=145499, 30% unique). The sample was processed 5 times with 1, 4, 8, 16, 30, and 40 virtual compute cores. The mean and standard error were plotted, Figure 3. Program settings were the same as in the first test.

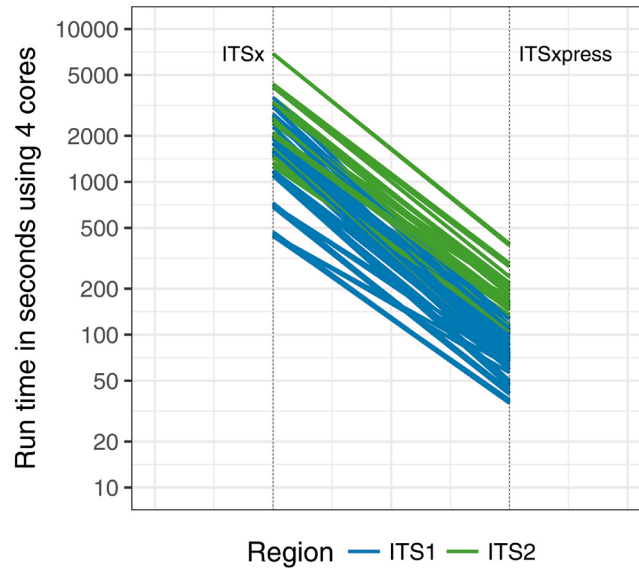


Figure 1. The run times for ITS1 and ITS2 samples processed using ITSx and ITSxpress using 4 logical compute cores. N=5 for each of the 30 samples.

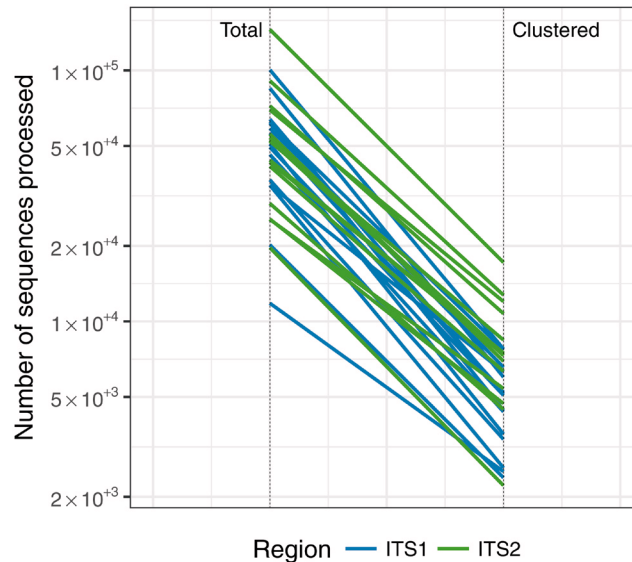


Figure 2. The number of total reads and the number or representative reads after clustering at 99.5% identity.

The trimming positions from ITSx and ITSxpress were compared for every ITS1 and ITS2 sequence. If a read was not trimmed identically by ITSx and ITSxpress, it was globally aligned and the start and stop positions were compared. Alignment was done using the Biopython Pairwise2 implementation of a global alignment function with the parameters (match score: 2, mismatch penalty: -1, gap opening penalty: -0.5, gap extension penalty: -0.1) (Cock *et al.*, 2009).

Results

When using 4 cores, ITSxpress trimmed ITS1 region samples a median of 23 times faster (Bayesian 95% Highest

Density Interval (HDI) 7 – 32) than ITSx, HDI interval calculated with the R package HDInterval (Meredith & Kruschke, 2018). ITSxpress trimmed the ITS2 region 14 times faster (95% HDI interval 8 – 24) than ITSx (Figure 1). Clustering at 99.5% identity reduced the number of reads used for Hmsearch by a median of 71 times (95% HDI 17 – 95) for ITS1 and 36 times (95% HDI 21 – 52) for ITS2, Figure 2.

Global alignment was used to compare the trimming results of ITSx and ITSxpress for reads that were not identical. When reads were clustered at 99.5% identity, the default behavior, ITSxpress and ITSx trimmed 99.822% (n=773021) of

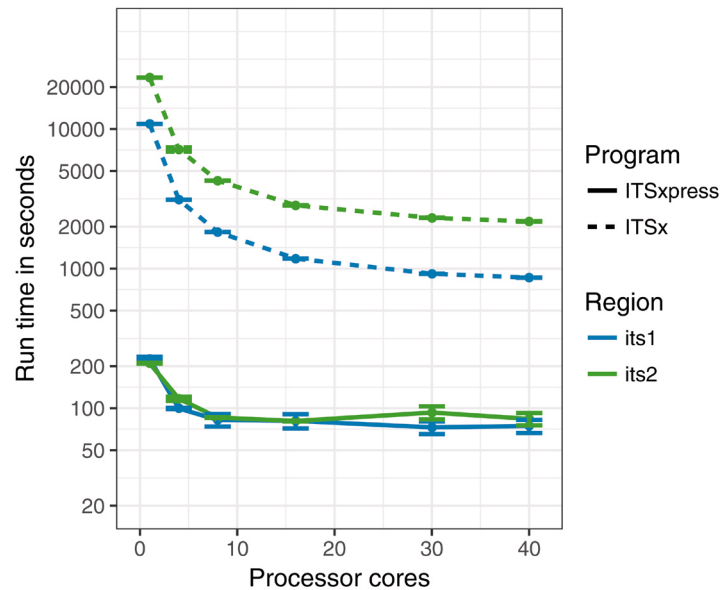


Figure 3. The mean and standard error of run times for of ITSx and ITSxpress on multiple logical cores. The largest samples from the ITS1 (n=100,543) and ITS2 (n=145,499) datasets were selected for analysis. N=5 for each core/sample combination.

reads in the ITS1 region within 2 bases of each other and 99.099% (n= 782385) of reads in the ITS2 region within 2 bases of each other. When reads were dereplicated at 100% identity ITSxpress and ITSx trimmed 99.992% (n= 773019) of reads in the ITS1 region within 2 bases of each other and 99.864% (n=782582) of reads in the ITS2 region within 2 bases of each other.

Discussion

ITSxpress increases the trimming speed of ITS sequences by clustering reads and optimizing the parameters for Hmmersearch. Most of the decrease in running time is attributable to clustering. Clustering at 99.5% identity resulted in reducing the number of sequences by a median of 71 to 36 times for the ITS 1 and ITS2 regions. The time complexity for Hmmersearch on a single core is approximately linear with the number of sequences so decreases in the number of sequences significantly decrease running time. The time required for clustering varies, for dereplication at 100% identity VSEARCH uses the rapid CityHash64 function (Rognes *et al.*, 2016). For clustering at less than 100% reads are sorted by abundance then clustered using greedy search. These steps take time but are faster than Hmmersearch and scale sub-linearly, resulting in median speed increases of 6-9x for the sequences dereplicated at 100% identity and 14-23x for the sequences clustered at 99.5% identity.

Both ITSx and ITSxpress use Hmmersearch, the same hmm models, and run using multiple cores. ITSxpress uses empirically tuned Hmmersearch heuristic values of 1×10^{-6} for F1, F2 and F3 which show increased speed and little loss of sensitivity. ITSx uses Hmmersearch's default values of 1×10^{-2} for F1, 1×10^{-3} for F2 and 1×10^{-5} for F3 when the "--heuristics" flag is set.

ITSx and ITSxpress scale differently as cores are added. ITSxpress spends about half its time clustering when the clustering identity is below 100%, and for a typical ITS sample this reduces the number of sequences to be analyzed by Hmmersearch to the point where parallelizing Hmmersearch does not result in large speed gains. This trait is beneficial for users using laptop or desktop computers because they can trim a typical ITS sample in less than a minute using 1–4 cores. Both programs use Hmmersearch for the most computationally intensive part of their workflows. ITSx benefits from Hmmersearch parallelization up to about 10 cores but then the increases decline; the nonlinear scaling of Hmmersearch is noted in the HMMER User Guide. (Eddy, 2011).

ITSx and ITSxpress trim most sequences exactly the same. At 100% identity one in 12,500 ITS1 sequences and one in 735 ITS2 sequences differ by more than two bases. This may be caused by differences in the heuristic settings for Hmmersearch. At 99.5% identity clustering the differences are greater, with one in 560 ITS1 sequences and one in 110 ITS2 sequences differing by more than two bases. At 99.5% identity, sequences from 600-800bp can be 3bp different, but be clustered together. Substitutions do not affect the trimming position, but insertions or deletions do, accounting for some of the difference. The clustering identity can be set to as low as 98% identity to accommodate special uses but lowering the identity below 99.5% is not generally recommended since ITSxpress is quite fast even at 100% identity.

ITSxpress quickly merges reads and trims the selected ITS region from a range of amplicon samples. It trims FASTQ files allowing for the use of newer sequence variant methods of exact sequence clustering and is available as a command line application and as a plugin for QIIME 2.

Software and data availability

Software

The source code for the stand-alone version of ITSxpress version 1.6.1 used for this manuscript is available from: <https://doi.org/10.5281/zenodo.1317575> (Rivers, 2018a). This software is available under the terms of the [Creative Commons Zero “No rights reserved” data waiver](#) (CC0 1.0 Public domain dedication).

Updated versions of the ITSxpress software are available from:

- Github: <https://github.com/USDA-ARS-GBRU/itsxpress>
- The Python Package index: <https://pypi.org/project/itsxpress/>
- Bioconda: <https://bioconda.github.io/recipes/itsxpress/README.html>

The QIIME 2 plugin for ITSxpress is available from: <http://doi.org/10.5281/zenodo.1317579> (Weber & Rivers, 2018); Github (https://github.com/USDA-ARS-GBRU/q2_itsxpress); and the Python Package index (https://pypi.org/project/q2_itsxpress/). This software is available under the terms of the CC0 1.0 Public domain dedication.

The computer code used to benchmark the software and generate the figures in this paper is available

at: <http://doi.org/10.5281/zenodo.1317585> (Rivers, 2018b); and Github (<https://github.com/USDA-ARS-GBRU/itsxpress-paper>). The code is also available under the terms of the CC0 1.0 Public domain dedication.

Data

Data used in this study are deposited in the NCBI Sequence Read Archive under the accessions listed in NCBI BioProject Accession [PRJNA483055](#).

Grant information

This research was funded by the United States Department of Agriculture (USDA), Agricultural Research Service (ARS) research project number 6066-21310-005-00-D and computational analysis using SCINet under project 0500-00093-001-00-D. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA.

The opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not reflect specific views of the USDA.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

- Amir A, McDonald D, Navas-Molina JA, *et al.*: **Deblur rapidly resolves single-nucleotide community sequence patterns.** *mSystems*. 2017; 2(2): pii: e00191-16.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 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 Ecol Evol*. 2013; 4(10): 914–919.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Bushnell B, Rood J, Singer E: **BBMerge - Accurate paired shotgun read merging via overlap.** *PLoS One*. 2017; 12(10): e0185056.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Callahan BJ, McMurdie PJ, Rosen MJ, *et al.*: **DADA2: High-resolution sample inference from Illumina amplicon data.** *Nat Methods*. 2016; 13(7): 581–583.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Caporaso JG, Kuczynski J, Stombaugh J, *et al.*: **QIIME allows analysis of high-throughput community sequencing data.** *Nat Methods*. 2010; 7(5): 335–336.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Cock PJ, Antao T, Chang JT, *et al.*: **Biopython: freely available Python tools for computational molecular biology and bioinformatics.** *Bioinformatics*. 2009; 25(11): 1422–1423.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Cock PJ, Fields CJ, Goto N, *et al.*: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.** *Nucleic Acids Res*. 2010; 38(6): 1767–1771.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Eddy SR: **Accelerated profile HMM searches.** *PLoS Comput Biol*. 2011; 7(10): e1002195.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Edgar RC: **UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing.** *bioRxiv*. 2016; 081257.
[Publisher Full Text](#)
- Gardes M, Bruns TD: **ITS primers with enhanced specificity for basidiomycetes—application to the identification of mycorrhizae and rusts.** *Mol Ecol*. 1993; 2(2): 113–118.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Köjalg U, Nilsson RH, Abarenkov K, *et al.*: **Towards a unified paradigm for sequence-based identification of fungi.** *Mol Ecol*. 2013; 22(21): 5271–7.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Meredith M, Kruschke J: **HDInterval: Highest (posterior) density intervals.** 2018.
[Reference Source](#)
- Nilsson RH, Ryberg M, Abarenkov K, *et al.*: **The ITS region as a target for characterization of fungal communities using emerging sequencing technologies.** *FEMS Microbiol Lett*. 2009; 296(1): 97–101.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Nilsson RH, Veldre V, Hartmann M, *et al.*: **An open source software package for automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology.** *Fungal Ecol*. 2010; 3(4): 284–287.
[Publisher Full Text](#)
- Rivers AR: **ITSxpress.** [software repository]. 2018a.
<http://www.doi.org/10.5281/zenodo.1317575>
- Rivers AR: **ITSxpress: Software to rapidly trim internally transcribed spacer sequences with quality scores for marker gene analysis [data set].** 2018b.
<http://www.doi.org/10.5281/zenodo.1317585>
- Rognes T, Flouri T, Nichols B, *et al.*: **VSEARCH: a versatile open source tool for metagenomics.** *PeerJ*. 2016; 4: e2584.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Schoch CL, Seifert KA, Huhndorf S, *et al.*: **Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi.** *Proc Natl Acad Sci U S A*. 2012; 109(16): 6241–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Weber KC, Rivers AR: **Q2-ITSxpress [software repository].** 2018.
<http://www.doi.org/10.5281/zenodo.1317579>
- White TJ, Bruns T, Lee S, *et al.*: **Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics.** In *PCR Protocols: A Guide to Methods and Applications*. San Diego. 1990; 315–322.
[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status:  


Version 1

Reviewer Report 24 October 2018

<https://doi.org/10.5256/f1000research.17138.r38810>

© 2018 Holm J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Johanna B. Holm 

Institute for Genome Sciences, Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD, USA

This report introduces an updated version of ITSx making the tool applicable to exact sequence variant analyses as opposed to operational taxonomic unit analyses. The tool is meant to be implemented upstream of sequence variant calling algorithms, producing dereplicated and merged reads. These reads can then be used in the deblur or dada2 algorithms for calling the exact sequence variants.

Minor Issues:

- **Methods/Operation:** Because there is clear instruction for use of the ITSxpress product in qiime2, it would be helpful to add a sentence regarding the use of the product in the dada2 workflow, as most users are accustomed to running the forward and reverse reads through the dada2 algorithm separately, and merging afterwards. Would the correct methodology be to run dada2(ITSxpress file) followed by sequence table production (skipping the merge step)?
- **Methods/Implementation:** replace "sorted by abundance and clustered" with "sorted by abundance and dereplicated", for clarity.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 19 September 2018

<https://doi.org/10.5256/f1000research.17138.r38059>

© 2018 Caporaso J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



J. Gregory Caporaso 

The Pathogen and Microbiome Institute, Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA

The authors present ITSxpress, an approach for trimming flanking rRNA regions from ITS sequence data. This is effectively a replacement for ITSx, which performed this process on fasta files after sequence clustering. By working with fastq data, ITSxpress allows this trimming to be applied to fastq files, which in turn allows for the application of modern amplicon analysis workflows.

Major points:

The user may also select dereplication from 98% to 100% identity.

The above sentence should probably be replaced with the following, since dereplication usually refers to clustering at 100% identity: *The user may also choose to cluster at between 98% and 100% identity.*

The start and stop position of each cluster representative is then used to trim each sequence in the cluster

How does this work if there were insertions and/or deletions between the cluster representative sequence and each cluster member? Wouldn't the position numbers be incorrect in that case?

How does a lower percent identity threshold for clustering impact accuracy (as compared to ITSx) and runtime? I'm wondering if it's worth it, for example, to run the clustering step at 98%, or maybe even lower, for quicker run time. Exploring accuracy and runtime as functions of percent identity seems like a missing piece of this study since in some cases the runtimes can be fairly long (e.g., 80 minutes) for a pre-processing step.

Minor points:

miss-assignments should be *mis-assignments*

OTU's is used in several places where *OTUs* should be used

It merges (if paired) and trims all samples

Should that say "trims all sequences"?

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: I am the PI on the QIIME 2 project, and this work describes software that is currently accessible as a QIIME 2 plugin.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research