User's guide: Manual for ITSx 1.0.9

ITSx is a Perl-based software tool to extract ITS1, 5.8S, and ITS2 – as well as full-length ITS sequences – from large Sanger as well as high-throughput sequencing datasets. ITSx uses hidden Markov models computed from large alignments of a total of 20 groups of eukaryotes, including fungi, metazoans, and plants, and the sequence extraction is based on the predicted positions of the ribosomal genes in the sequences. The software constitutes a completely rewritten version of the Fungal ITS Extractor; so different in its entire foundation that we feel it requires a change of name – ITSx (x for eXtractor).

This document is a guide on how to install and use the software utility. The software is written for Unix-like platforms, and should work on nearly all Linux-based systems, as well as MacOS X.

Contents of this manual

1. Detailed installation instructions

The README.txt file bundled with the package provides a quick installation guide.

In order to install certain packages, you might need to have superuser privileges. For installation on Mac, you will have to install the Apple Xcode package available on your MacOS X System DVD in order to be able to compile programs. Please talk to your system administrator if you feel unsure about these steps. Note that the packages are mandatory and that you should not proceed unless these criteria are fulfilled. [One user reported that a re-install of Xcode was needed to get HMMER to work.]

[If you don't have superuser privileges on your machine: Create a directory within your user directory, e.g. /home/user/bin/, and store all required binaries there. By adding this directory to your PATH, any software placed in the directory will behave as if installed for all users using superuser privileges. If you use the bash shell, you can usually add a bin directory to your PATH by adding the line "export PATH=$PATH:$HOME/bin/:." to the file .profile in your home directory. (The process of adding items to one's PATH, however, varies among systems and shells.) Close the terminal and open a new one for this change to take effect.]

Perl needs to be installed on the computer. Most Unix-based systems including Linux and MacOS X have Perl pre-installed. You can check this by opening a command line terminal and type "perl -v". In case Perl is not installed, you have to download (http://www.perl.org) and compile the program.

Download and install HMMER version 3 (http://hmmer.janelia.org/software). ITSx relies on HMMER version 3 and will ***not*** work with earlier versions of HMMER. Download the HMMER package source code to your preferred directory such as /home/user/. Open a command line terminal, move into the directory with "cd /home/user/" and unpack the

tarball with “tar -xvfz hmmer-3.0.tar.gz”. Now, you must compile HMMER from source files. To compile it from source, enter the new directory and follow the installation instructions in the file INSTALL.

If you have trouble compiling HMMER, you can try to use the pre-compiled binaries available at the HMMER home page. After download, and unpacking of the tarball, the binaries are located in the binaries directory contained within the newly created HMMER directory. Move into the binaries directory and move all of its contained files into your preferred bin directory (usually either /usr/local/bin/ or your own bin directory, e.g. /home/user/bin/). The HMMER package should now be installed on your computer; you can check this by typing “hmmscan -h” in the terminal and press enter; you should now see HMMER output.

Go to http://microbiology.se/software/itsx and download the ITSx package. Download it to your preferred directory. Unpack the downloaded tarball with “tar xvfz ITSx_1.0.tar.gz”. A directory called ITSx_1.0 will be created. You will see the following files and directories inside it: ITSx, the ITSx_db directory (containing the Hidden Markov Models), this user’s guide, the README.txt file, the license.txt file as well as test input files. Install ITSx by copying the ITSx file and the ITSx_db directory to your preferred bin directory. (You may also run ITSx from this directory by typing “./ITSx” or “perl ITSx” followed by the command line options.) When ITSx is successfully installed you should see its help message when typing the command “ITSx --help”.

For the user with more limited UNIX experience, the easiest thing to do is to copy the query (DNA sequence) file to the ITSx dir and process it from there. ITSx should then be evoked as “perl ITSx …”.

2. Usage and commands

For the *very* impatient only: follow the brief installation instructions in the file README.txt. To extract ITS sequences from the file test.fasta, you would then type "ITSx -i test.fasta -o test" on the command line.

For all other users: ITSx accepts input in the FASTA format. As it pre-processes the input sequences, it is possible to input both aligned and unaligned FASTA files, containing both DNA and RNA sequences. By default, ITSx outputs six files; one summary file of the entire run, one more detailed table containing the positions in the respective sequences where the ITS subregions were found, one “semi-graphical” representation of hits, one FASTA file of all identified ITS sequences, one FASTA file for the ITS1 and ITS2 regions, respectively. In addition, if entries that did not contain any ITS region are found, a list of sequence IDs representing those entries is also written. If potentially chimeric sequences are found, those are also written to a separate file, and the same is true for entries that are problematic in some other respect. To list all the available options for ITSx, type “ITSx --help”. You can use the test.fasta file that comes bundled with the software for a test run. This file contains 50 randomly selected ITSx entries. In the simplest case, ITSx is run by “ITSx -i input_file -o output”. Below is a listing of all options ITSx accepts. Boolean options can be turned on with “T”, “true” or “1” and off using “F”, “false” or “0”.

Main options:

-i {file}	Nucleotide FASTA input file to investigate. ITSx accepts both aligned and unaligned FASTA. If no input is specified, ITSx will read sequences from standard input, which means that FASTA sequences can be piped into ITSx.

-o {file} — Base for the file names of the output files. Suffixes will be added automatically. Defaults to ITSx_out.

-p {directory} — A path to a directory containing HMM-profiles for ITSx. By default, the program assumes to find the databases in the ITSx_db directory, located in the same directory as ITSx itself.

--date {T or F} — Adds a date and time stamp to the output file. This can be useful e.g. if ITSx is part of a pipeline where input files with the same name could cause overwriting of important data. Off (F) by default.

--reset {T or F} — If enabled, this re-creates the HMM-database before the search is run. This is useful if HMMER has been updated, or if ITSx searches seem to fail entirely. Off (F) by default.

Sequence selection options:

-t {list of organism groups} — Set of profiles to use for the search (comma-separated). Accepts any list of sets, e.g. "rhodophyta,fungi", "h,f" or simply "all" to include all sets. Can be used to restrict the search to only a few organism groups types to save time, if one or more of the origins are not relevant to the dataset under study. Default is to use all (the "all" option). The list of options is given below, both with character codes and full names.

Character code	Full name	Alternative name
A	Alveolata	alveolates
B	Bryophyta	mosses
C	Bacillariophyta	diatoms
D	Amoebozoa	
E	Euglenozoa	
F	Fungi	
G	Chlorophyta	green-algae
H	Rhodophyta	red-algae
I	Phaeophyceae	brown-algae
L	Marchantiophyta	liverworts
M	Metazoa	animals
O	Oomycota	oomycetes
P	Haptophyceae	prymnesiophytes
Q	Raphidophyceae	raphidophytes
R	Rhizaria	
S	Synurophyceae	synurids
T	Tracheophyta	higher-plants
U	Eustigmatophyceae	eustigmatophytes
X	Apusozoa	
Y	Parabasalia	parabasalids
.	All	

-E {value} — Domain E-value cutoff a sequence must obtain in the HMMER-based step to be included in the output. Default = 1e-5.

-S {value} — Domain score cutoff that a sequence must obtain in the HMMER-based step to be included in the output. Default = 0.

-N {value} — The minimum number of domains (different HMM gene profiles) that must match a sequence for it to be included in

	the output (detected as an ITS sequence). Setting the value lower than two will increase the number of false positives, while increasing it above two will decrease ITSx detection abilities on fragmentary data. Default = 2.
--selection_priority {sum, domains, eval, score}	Determines what will be of highest priority when assessing the origin of the sequence. Options are: - sum, which sums the scores for each profile match and divides the sum by the number of profiles of the given type - domains, which uses the number of domains retrieved of a given type - eval, which uses the average E-value of the found hits - score, which uses the average score of the found hits Default is to use sum (sum of scores).
--search_eval {value}	The actual E-value cutoff used in the HMMER search. High numbers may slow down the process. Should never be set to a lower value than the -E option. Cannot be used in combination with the --search_score option. Default is 0.01.
--search_score {value}	The score cutoff used in the HMMER search. Low numbers may slow down the process. Should never be set to a higher number than the -S option. Cannot be used in combination with the --search_eval option. Default is to use E-value cutoff (see --search_eval above), not score.
--allow_single_domain {e-value,score or F}	Allow inclusion of sequences that only find a single domain, given that they meet the more stringent E-value and score thresholds specified. By default, single domains are allowed, with E-value cutoff 1e-9 and score cutoff 0 (“1e-9,0”).
--allow_reorder {T or F}	Allows profiles not to be in the expected order on the extracted sequences. If turned off, a file of potentially chimeric sequences (with profile matches in the wrong order) is written, allowing for,e.g., rudimentary assembly chimera detection. Off (F) by default.
--complement {T or F}	If on, ITSx checks both DNA strands for matches to HMM-profiles. On (T) by default.
--cpu {value}	The number of CPU threads to use. ITSx performs significantly faster using more CPUs/cores. Default is 1.
--multi_thread {T or F}	Multi-thread the HMMER-search. On (T) by default if the number of CPUs/cores is larger than one (--cpu option > 1), else off (F).
--heuristics {T or F}	Selects whether to use HMMER's heuristic filtering, off (F) by default. Leave this setting off for higher precision.

Output options:

--summary {T or F}	If on, ITSx outputs a summary of results. File suffix is “.summary.txt”. On (T) by default.
--graphical {T or F}	If on, ITSx outputs “graphical” text representations of where in each sequence the conserved domains were found. File suffix is “.graph”. On (T) by default.

--fasta {T or F}	If on, FASTA-formatted files containing the extracted ITS sequences are written. One file for each region specified using the --save_regions option is written, along with a file with all full-length ITS regions identified. On (T) by default.
--preserve {T or F}	If on, ITSx will preserve the sequence headers from the input file instead of replacing them with ITSx headers in the output. Off (F) by default.
--save_regions {SSU,ITS1,5.8S,ITS2,LSU,all,none}	A comma separated list of regions to output separate FASTA files for. Outputs only the actual ITS sequences (ITS1,ITS2) by default.
--anchor {integer or HMM} :	If set to a number, ITSx saves an additional number of bases before and after each extracted region, corresponding to that number, for anchoring in e.g. multiple sequence alignments. If set to 'HMM', all bases matching the corresponding HMMs before and after the two ITS regions will be outputted. As of version 1.0.9, ITSx also saves these "anchor sequences" to the full-length output file (suffix ".full.fasta"). By default, no additional base pairs will be written (same behaviour as previous ITSx versions).
--only_full {T or F}	If true, the output is limited to full-length ITS1 and ITS2 regions only. Off (F) by default.
--partial {value}	If larger than 0, ITSx will save additional FASTA-files for full and partial ITS sequences longer than the specified cutoff value. By default, this setting is left to 0 (zero), which means off.
--concat {T or F}	Saves a FASTA-file with concatenated ITS sequences (with 5.8S pruned). Off (F) by default. Usage of this option is *not endorsed,* as it will essentially produce chimeric ITS1+ITS2 sequences!
--minlen {value}	Minimum length the ITS regions must be to be outputted in the concatenated file (see --concat above). Default is zero (0).
--table {T or F}	If on, ITSx saves table format output of probable ITSx sequences. Off (F) by default.
--detailed_results {T or F}	If on, ITSx saves a table of detailed results for all probable ITSx sequences. Off (F) by default.
--not_found {T or F}	If on, ITSx outputs a list of entries that do *not* seem to contain any ITS sequence. File suffix is "_not_found.txt". On (T) by default.
--truncate {T or F}	Removes ends of ITS sequences if they are outside of the ITS region. If off, the whole input sequence is saved. On (T) by default.
--silent {T or F}	Suppresses printing of progress info to screen. Off (F) by default.
--graph_scale {value}	Sets the scale of the graphical output. If the provided value is zero, a percentage view is shown. Default is 0.
--save_raw {T or F}	Saves all raw data for searches etc. instead of removing it when finished. Saves data to a directory with the suffix "_ITSx_raw_output". Off (F) by default.

Information options:

-h	Displays the help message.

--help	Displays the help message.
--bugs	Displays the bug fixes and known bugs in this version of ITSx.
--license	Displays licensing information.

3. Output files

ITSx outputs a number of files, depending on what is selected by the user (see Usage and Commands above). By default, three FASTA-files, a table of positions for the regions recovered, a file containing graphical representation of putative ITS sequences, a list of entries not found to contain ITS sequences, and a summary file are written. In addition, tables of putative ITS sequences, and additional FASTA files, can be written on request by the user. There is also an option to preserve all the intermediate data generated by the HMMER searches.

FASTA-output

ITSx generates one FASTA file for each ITS region (and additional files for the SSU, LSU, and 5.8S if specified), and one file with the full-length ITS sequences identified. Sequences in these files are marked according to their putative origin. However, ITSx is not designed to make accurate predictions on organism groups, and no double-checking of this prediction is performed. Therefore, the ITS sequences extracted should be further examined using e.g. BLAST searches.

Note that ITSx adds the type of the ITS sequence ("fungi ITS sequence") to the definition line in the example below:

```
>gi|17298408|gb|AF394527.1.N|F   fungi   ITS   sequence   (515   bp)
AGCAGAGCGATTTGGGGTTTAATATGTATGTATACATTACGTTCGAAGGATCGATTGGCTTTGGTGA…
```

Summary

A summary of the ITSx run is written to a file with the suffix ".summary.txt". In this file the statistics of the run are collected, as are the starting and ending times for the run. Also, lists of the identifiers of extracted ITS sequences are written to this file. An example of a summary file is shown below:

```
ITSx run started at Mon Sep 17 16:27:31 2012.
----------------------------------------------------------------
Number of sequences in input file:              100
Sequences detected as ITS by ITSx:      95
  On main strand:               95
  On complementary strand:      0
Sequences detected as chimeric by ITSx: 3
ITS sequences by preliminary origin:
  Alveolates:           1
  Amoebozoa:            0
  Bacillariophyta:      0
  Brown algae:          0
  Bryophytes:           0
  Euglenozoa:           6
  Eustigmatophytes:     0
  Fungi:                35
  Green algae:          1
  Liverworts:           8
  Metazoa:              17
```

```
  Microsporidia:         0
  Oomycetes:             0
  Prymnesiophytes:       0
  Raphidophytes:         0
  Red algae:             20
  Rhizaria:              6
  Synurophyceae:         0
  Tracheophyta:          1
--------------------------------------------------------------------
ITSx run finished at Mon Sep 17 16:31:06 2012.
```

Positions file

ITSx writes the positions in which the SSU, ITS1, 5.8S, ITS2 and LSU regions are found in each query sequence to a file with the suffix ".positions.txt". This file is organized as a tab separated text file, with the following columns: Sequence ID, Length of the sequence, SSU range, ITS1 range, 5.8S range, ITS2 range, and LSU range.

Graphical representations

ITSx writes graphical (ASCII) representations of where in each sequence the various conserved regions were found to a text file with the suffix ".graph". Separate graphs are written for each origin and strand, which means that each sequence entry may be present more than once in this file, if it has matches to HMM-profiles from more than one origin. This makes it possible to manually inspect how ITSx has evaluated each sequence. The graphical representations looks something like this:

```
H matches on main strand:
>> gi|17999611|gb|AY029388.1.H  2114 bp
------SSU----------------5.8--End-------------------LSU--------
**********************************************************
```

The first row shows the type of the entries below, as well as the strand they are found on. Each entry begins with the characters ">> ", followed by the sequence identifier and its length. Below the identifier row, the sequence graph is shown. By default, all sequences are scaled so that they are of equal length, and the domains are placed according to their *relative* position in the sequence. The characters that are used in the graphical representations are explained in the table below.

Feature	Description
-	Part of the sequence without any conserved domain (variable region).
SSU	Start of a conserved domain.
>	Indicates that one conserved domain goes into the next. Domains are normally not overlapping, so this could be an indication of a compromised input sequence.

The line of asterisks indicates the end of one set of matches. Note that the graph should be viewed with a non-proportional font, such as Courier, if loaded into, e.g., Word.

Detailed results table

The full results of the ITSx extraction is saved to a file with the suffix ".extraction.results". This file consists of tab-separated columns containing various information on each ITS sequence found. The file can be easily imported into programs such as Excel. The contents of the columns (from left to right) are explained in this table:

Column	Description
ID	The identifier of the query sequence.
Length	The length of the query sequence.
Origin	A one-letter abbreviation of the sequence origin (see table under the –t option in the *Usage and commands* section for complete list).
Strand	A zero (0) if the ITS was found on the main strand, a one (1) if it was found on the complementary strand.
Domains	The number of conserved domains for the most likely origin that was found in the sequence.
Average E-value	The average E-value for these domains.
Average score	The average score for these domains.
Score sum	The average sum of scores for these domains.
Start	The starting position of the first domain.
End	The ending position of the last domain.
First domain	The domain that is located first in the sequence.
Last domain	The domain that is located last in the sequence.
Chimera	The word "Chimeric" if the sequence was marked as a potential assembly chimera. Empty if not. Sequences will only be marked as chimeric if the --allow_reorder option is turned off. Note that this is *not* a robust measure against chimeras of all kinds.
Specific origin information	A collection of information of *all* possible origins for the given query. Each entry is a space-separated list, containing the origin type, the number of domains of that type, the average E-value, and the average score, e.g. "N: 4 8.2e-11 43.475"

Extraction results table

If table output is turned on, ITSx will save statistics of every profile set that the sequence in question matches to in a file with the suffix ".hmmer.table". This file consists of tab-separated columns containing information on the ITS sequence found. The contents of the columns (from left to right) are explained in this table:

Column	Description
ID	The identifier of the query sequence.
Length	The length of the query sequence.
List of hits	Each new column contains information of a profile match. Each column is organised as follows: "<starting position> - <ending position>: <name of matching profile> (<score>, <E-value>)".

As in the graphical output file, the table file is divided into sections. Each section represents one group of sequences, and begins with the line "X matches on main strand:", and ends with a line of asterisks.

List of sequences not containing any ITS region

If not-found output is turned on, ITSx will write a list of sequences for which no ITS regions could be found to a file with the suffix "_no_detections.txt". The file contains only the identifiers of the non-ITS sequences, and is written only if there are entries without any detected ITS region.

Chimeric sequences

If the option --allow_reorder is turned off, ITSx will save an additional FASTA file containing sequences that are suspected to be chimeric, given that such sequences are found. These are sequences with domains located in the wrong order. This is useful on full-length or near full-length data sets, but should not be used on short reads as it could increase the number of false negatives when run on short sequences.

Raw data

If the option to save all raw data is turned on, ITSx will save all data from the pre-processing, HMMER-search, as well as a file of raw statistics into a directory with the suffix "_ITSx_raw_output".

4. Algorithm and implementation

The main design goal for ITSx is to achieve fast and accurate extraction of ITS regions in large data sets, without introducing a large number of false positives. To be able to reach a high speed, ITSx relies on the HMMER3 software, which allows for extremely fast comparisons of HMM-profiles to a sequence set. To achieve high detection accuracy, ITSx uses multiple HMM-profiles built from the conserved domains flanking the ITS regions (SSU, 5.8S and LSU), representing a large number of species groups. This enables ITSx to extract ITS regions from all eukaryote lineages for which a substantial number of reference ITS sequences were available as of 2012.

While the default settings of ITSx should be usable in most situations, you should consider if they suitable for your purposes and for your data set. If the data set is small, this can be done by running the software multiple times on the data, with different settings, and analyse the outcome. On larger data sets, it might be more feasible to only run ITSx on a subset of the sequences for testing. The graphical output is very useful for determining whether ITSx performs as desired on the data, as the positions of the found conserved domains can be easily investigated. If domains are missing, the criteria might be set to be too stringent. If they are not in sequential order (from SSU to LSU with the 5.8S in between), that might be an indication that there is something wrong with the input sequences.

The HMMER program hmmscan, used by ITSx, normally uses heuristic filters to increase the search speed. ITSx runs hmmscan with the "--max" option in order to turn off all heuristic filters. This increases detection power at the cost of speed. However, the time requirement of the HMMER search is generally not an issue with ITSx, while accuracy is, and thus the heuristic filters are not used. The heuristic filters can be turned on by using the "--heuristics T" option.

5. Notes on PCR primers and short input sequences

One obvious use of ITSx would be to run it on sequence datasets generated by PCR amplicon studies of the ITS region, to extract ITS1 and/or ITS2 regions, and to sort out non-target sequences in the data prior to further analysis. ITSx uses the conserved SSU, 5S and LSU genes to located and orient the ITS regions. To do this, the software requires at least ~20 bp. (of at least one) of those genes to be present for each input sequence – preferably 25 bp. However, some primer pairs targeting the ITS

region will not include sufficiently large portions of these genes to be detected with the accuracy required by ITSx by default. This may lead to that fewer, or even none, of the input sequences are recognized as ITS containing.

If the dataset is known to contain only ITS sequences, a remedy to this problem can be to lower the stringency of ITSx, using the -E option. By default, this is set to 1e-5, but this can be increased to say 0.01, or even 1 to allow for detection of shorter portions of the conserved genes – down to some 15 bases. This feature comes at the price of an increase proportion of false-positive matches, but in the case of ITS-only datasets this will be less of a concern. Note that this should normally be done only for datasets that are known to contain only ITS sequences, and that caution needs to be taken in the downstream analysis so that false-positive extractions can be avoided (e.g. investigate spuriously long or short ITS sequences with a sound degree of scepticism). Conversely, going for very stringent settings may come at the price of sensitivity as sequences with deviant genes (or of reduced read quality) may be missed. The default settings of the software are calibrated with environmental datasets in mind, to keep false-positive extractions at a minimum.

If the problem instead is that only one of those genes is present on the input sequence, another, even more polished, solution exists. In order to score a sequence as an ITS sequence, ITSx prefers to see that the sequence produces matches to at least two HMMs (such as 3' SSU and 5' 5.8S). This helps ITSx to keep the number of false-positive identifications low. However, the software also allows sequences that produce a match to only one HMM (such as 3' SSU) to be scored as ITS sequences, provided that the match is particularly stringent. The stringency of this parameter is controlled through the --allow_single_domain switch (see above). The default value is 1e-9. If ITSx is used on very short sequences (e.g. 3' SSU + 100 bp. ITS1), then only one HMM can be expected to match, and the E-value of that match will be compared to the value of this parameter. Therefore, if the dataset at hand contains sequences which should only contain a single conserved region, using e.g. --allow_single_domain "1e-5,0" can be enough to go from no matches to all matches. However, as with the -E option above, increasing the E-value of this parameter comes at the expense of higher risk for false positives.

6. License information