UTOPIA: an automatically UpdaTed, cOmPlete and consistent ITS reference dAtabase

Sébastien Theil, Etienne Rifa

UTOPIA :
AN AUTOMATICALLY UPDATED, COMPLETE AND CONSISTENT ITS REFERENCE DATABASE

Sebastien Theil[1], Etienne Rifa[1]

[1]Université Clermont Auvergne, INRA, VetAgro Sup, UMRF Unité Mixte de Recherche sur le Fromage, 20 côte de Reyne, 15000 Aurillac, France

Context

Taxonomic assignment in metabarcoding analysis is a critical and challenging step. As more organisms being sequenced, taxonomy is evolving fast with multiple taxa rearrangement and thousand of new sequences uploaded each year. The internal transcribed spacer (ITS) is an ubiquitous sequence used as a barcode to identify fungi species in complex environmental samples.

Currently used databases like UNITE, offer a good and reliable reference, but update frequency is generally low, and new strain sequences can take several years to be integrated. UTOPIA provides a workflow that produce an updated ITS reference database directly from the NCBI genbank and taxonomy database.

UTOPIA

Fetch

NCBI → ~48h → GB

eFetch
internal transcribed spacer
+ complete sequence + Fungi

GENBANK
~2Gb

Extract

NCBI

Taxonomy

FASTA TSV

Accession ID
Sequences
7 Ranks taxonomy strings

Filter

PRINSEQ

100bp < length < 5000bp
No N's and no IUPAC

CD-HIT

Dereplication
&
clustering

Resolving multiple taxon per cluster :
Re-annotate to last common ancestors

IDTAXA

Formatting
Training

Database statistics

Taxon overview

Clustering threshold collapses species and few genus.
Number of sequences and taxons is comparable to UNITE.

	utopia Id90, cov90	utopia Id99, cov99	Unite Dynamic	Unite Dynamic_s	unite INSDC
kindom	1	2	1	1	1
phylum	9	10	20	21	21
class	63	64	93	102	105
order	233	236	273	301	321
family	809	856	804	918	1007
genus	3739	4978	3578	4512	5256
specie	13941	42709	17278	22253	33024
nb_seq	28538	126317	35667	71042	887397

Test on mock from Bakker *et al. 2018*

Metrics from **Edgar et al. 2018**:
TPR: True positive rate
MCR: Misclassification rate
UCR: under-classification rate
Acc: Accuracy

Utopia has best accuracy at Genus level, especially for ITS2.
Taxonomies from NCBI match better those used by Bakker et al.

ITS1 (26 sequences)

	Family				Genus			
	TPR	MCR	UCR	Acc	TPR	MCR	UCR	Acc
utopia id90, cov90	0,8	0,1	0,1	0,8	0,4	0,1	0,5	0,4
utopia id99, cov99	0,9	0,1	0	0,8	0,7	0,2	0,1	0,7
Unite Dynamic	0,7	0,1	0,2	0,7	0,6	0,1	0,3	0,5
Unite Dynamic_s	0,6	0,1	0,3	0,6	0,5	0,1	0,5	0,5
Unite INSDC	0,8	0	0,2	0,7	0,5	0,1	0,3	0,5

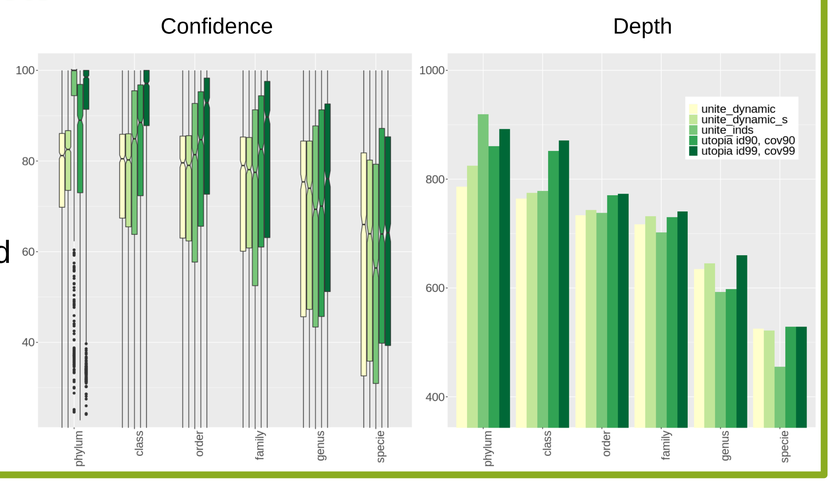
ITS2 (28 sequences)

	Family				Genus			
	TPR	MCR	UCR	Acc	TPR	MCR	UCR	Acc
utopia id90, cov90	0,7	0,2	0	0,7	0,7	0,2	0	0,7
utopia id99, cov99	0,7	0,1	0,2	0,7	0,6	0,1	0,2	0,6
Unite Dynamic	0,4	0	0,6	0,4	0,3	0	0,7	0,3
Unite Dynamic_s	0,5	0	0,5	0,5	0,2	0	0,8	0,2
Unite INSDC	0,8	0,1	0	0,8	0,6	0,1	0,3	0,6

Test on two real datasets

ITS1 dataset from raw milk:
1348 sequences

Utopia has a better median confidence at family level, and performs the same for genus and species.
Utopia99 and UNITE provide slightly the same annotation depth.



Conclusions

We provide to the community a protocol and scripts to build an updated ITS database locally from the very last release of NCBI. Our database is created, filtered, and clustered fully automatically. Taxonomy is processed to obtain a 7 ranks assignment for each sequence and formated to fit commonly used programs such as IDTAXA, or Phyloseq.

However we show here that UTOPIA is at least as good as UNITE in term of depth of annotation in real data and better in accuracy on the Bakker *et al.* mock community making our database a good alternative to assign fungus community with the latest sequences and taxonomies available.

Available on git repository at :
https://forgemia.inra.fr/umrf/utopia

UMRF
20, côte de Reyne
15000 AURILLAC
https://www6.ara.inra.fr/umrf

04.71.45.64.10