

Handling Fungal data in MoBeDAC

Jason Stajich
UC Riverside

Fungal Taxonomy and naming undergoing a revolution



WG 1.3 – Fungi

More than just mushrooms

Print



Chair: Pedro Crous
CBS Fungal Biodiversity Centre,
Utrecht, Netherlands
Vice-chair: Keith Seifert,
Agriculture and Agri-food Canada,
Ottawa, Canada

Fungi have far higher diversity than land plants. There could be more than two million species but this estimate is very tentative because fungal taxonomy is so incomplete (only 50,000 species have been formally described).

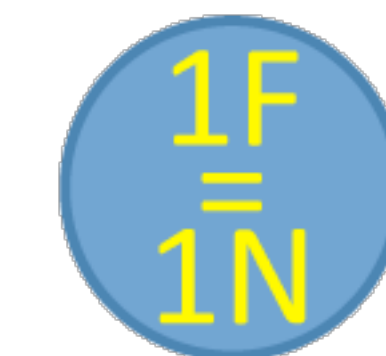
Although no fungal species are known to be endangered, members of this kingdom are important as disease agents, as food, as producers of antibiotics, as agents of fermentation and as the basis of much organic decay.

Past taxonomic work on fungi has been slowed by their morphological conservatism and by the small size of most species. Thus, DNA-based taxonomy will revolutionize our understanding of fungal diversity and enable, for the first time, the connection of their life stages. IBOL will register barcodes for at least 10,000 fungal species by 2014 with a particular focus on building barcode libraries for indoor fungi, for basidiomycetes (the "higher fungi") and for those fungi that are important pathogens of agriculture and forestry.

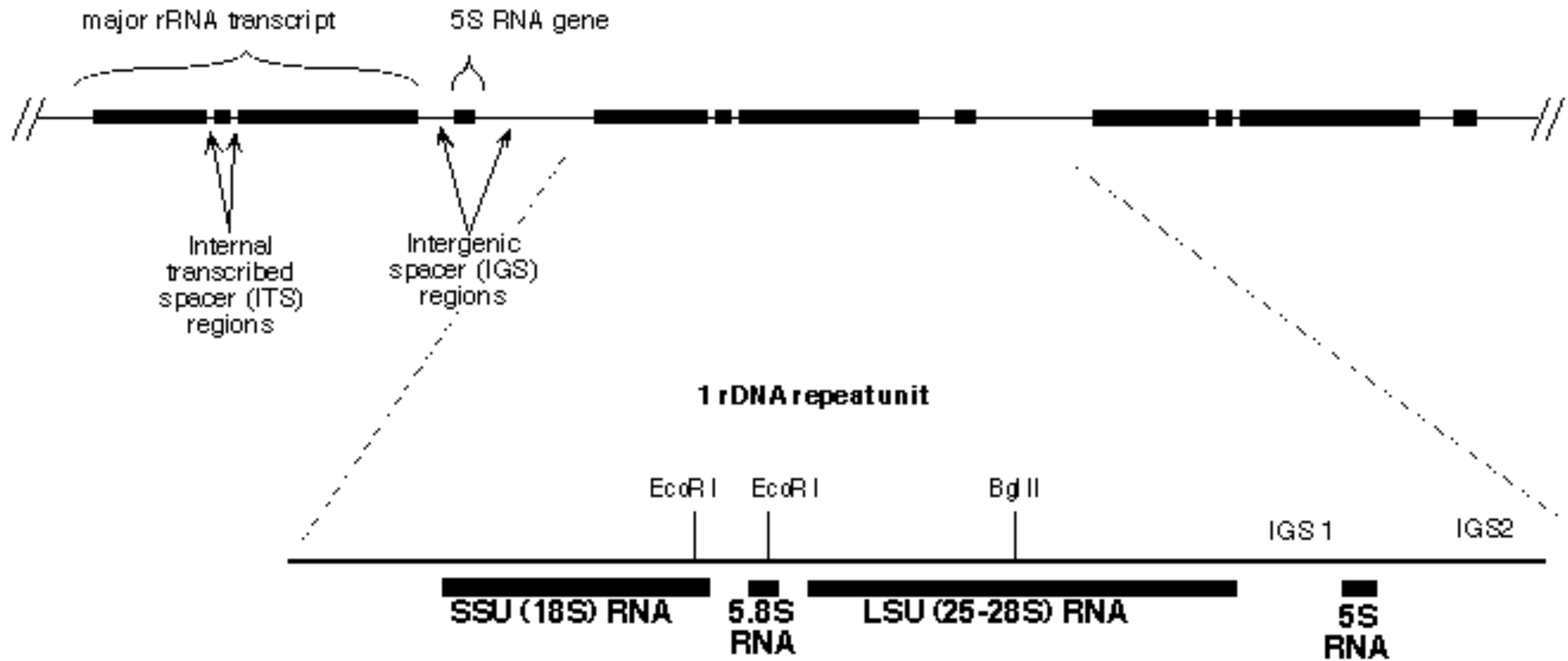
Goal
10,000 species

Also posted in [Barcode Library](#), [WG 1.3 - Fungi](#), [Working Group Profile](#)

One fungus, one name

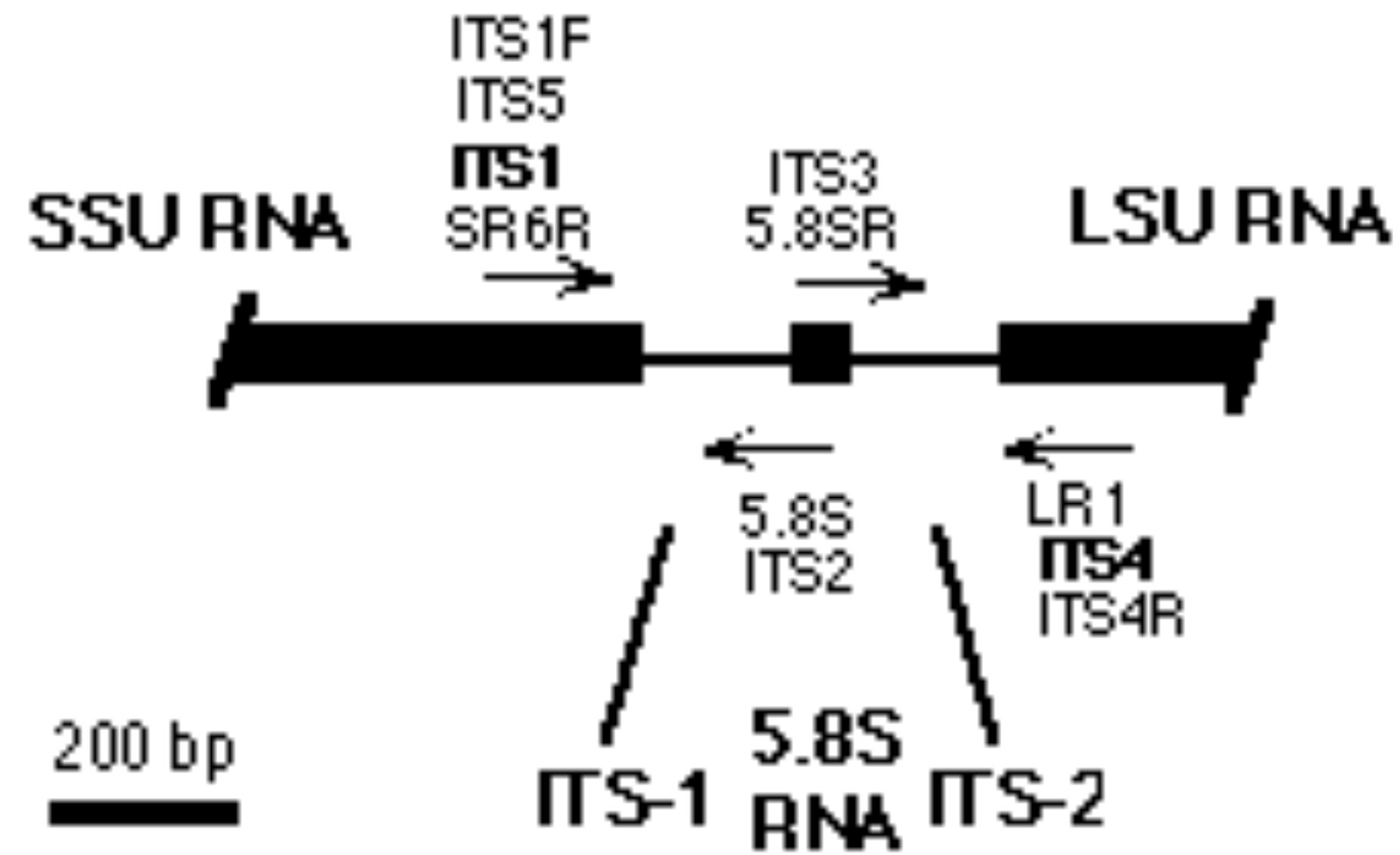


CBS Symposium



<http://www.biology.duke.edu/fungi/mycolab/primers.htm>

ITS primers



Primers for routine sequencing are shown in bold

<http://www.biology.duke.edu/fungi/mycolab/primers.htm>

Why is this hard?

- ITS is easier to amplify because it is multicopy and concerted evolution keeps the copies homogenized(*) and rRNA genes will change more slowly helping make universal primers possible(*)
- Curated sequence database of marker to taxonomy needs to be built
- Taxonomy specified to different depths for some lineages, especially early branching ones
- ITS cannot really be used to build trees - it is a good barcoding molecule as it changes rapidly. Though in some lineages not rapidly enough and resolving in those lineages requires another marker
- LSU is good for backbone and major grouping but hard to resolve species or even genus often with this molecule.

* Assumptions that mostly/often hold true

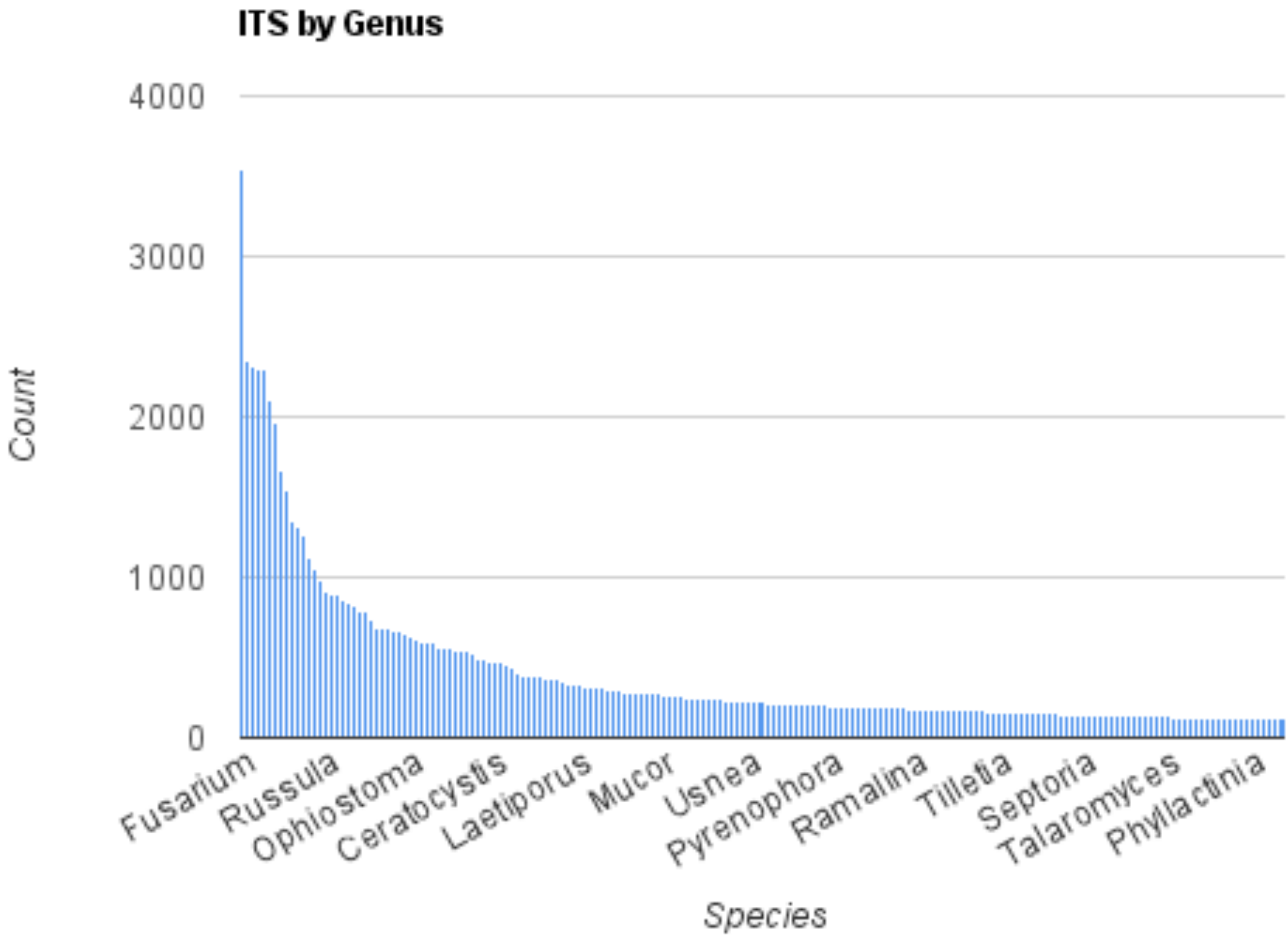
Speaking the same language

- Unified Taxonomy
- Multiple marker sequences
 - ITS, SSU, LSU
 - COI1
- Assembling the Fungal Tree of Life markers
 - (RPB1, RPB2, EF1alpha)
 - Phylogenomically chosen 40-60 protein coding genes.



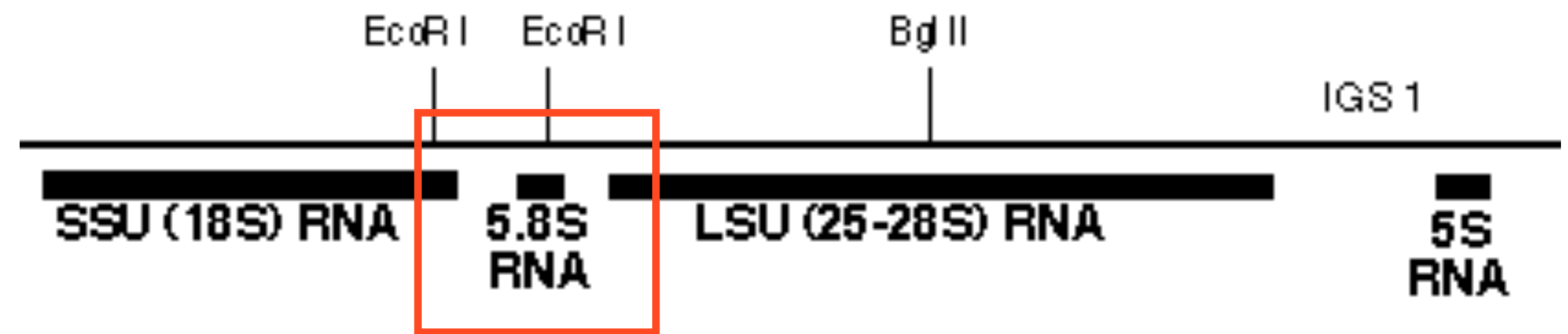
What's in GenBank for ITS? (ca 2010)

1-454 run



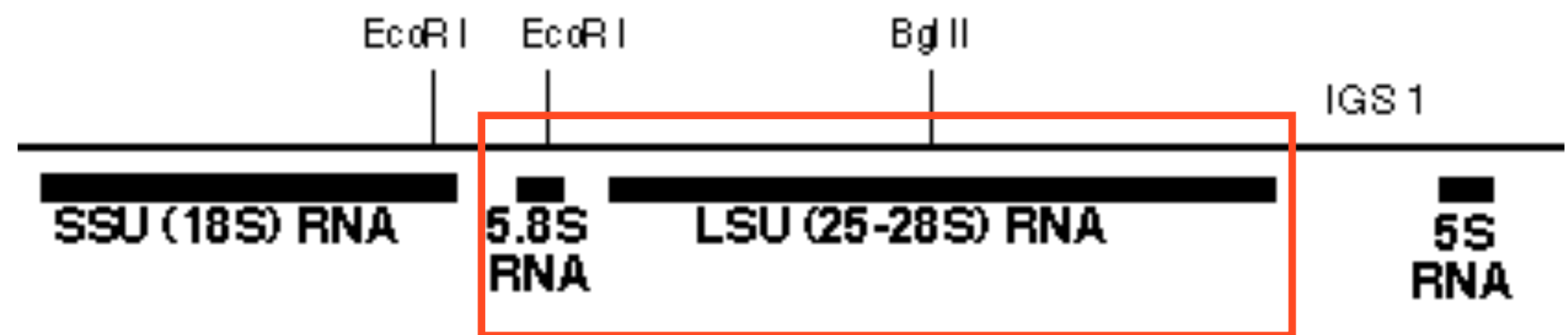
Neocallimastigomycota	267605
Ascomycota	79597
Basidiomycota	45393
Glomeromycota	3809
Chytridiomycota	485
Microsporidia	482
Blastocladiomycota	10

Some ITS databases



- UNITE (unite.ut.ee) - UNITE barcoding sequences: 3878 ITS sequences of 1508 species from 255 genera
Fungal ITS sequences in database (UNITE + INSD): 205,688
- Extracts from GenBank ~
- In case of uncurated data there are many mis-specified taxonomy assignments to sequence.
 - Worst are endophytic fungi that are assigned to plants or other improper specimen identification
- Need: Expert curation, collection. Much of this is being done very well by UNITE team and collaborators. We (Sloan funded project) do want to help

LSU & databases



- Ribosomal Database Project (MSU) has a Naive Bayesian classifier for Fungal LSU - <http://rdp.cme.msu.edu/classifier/>
- Still new to us to be able to test out (Liu et al 2011) but promising and quite fast as it doesn't have same alignment-based
- Vast majority of data being generated seems to be ITS, but sometimes there is a paired LSU study.

Playing with real data

- Amend et al PNAS 2010 “Indoor fungal composition is geographically patterned and more diverse in temperate zones than in the tropics.”
- 72 samples of fungi from 6 continents. Sampled ITS2 region and the D1-D2 region of LSU with 454-FLX
- Main finding of increasing species diversity with increasing latitude

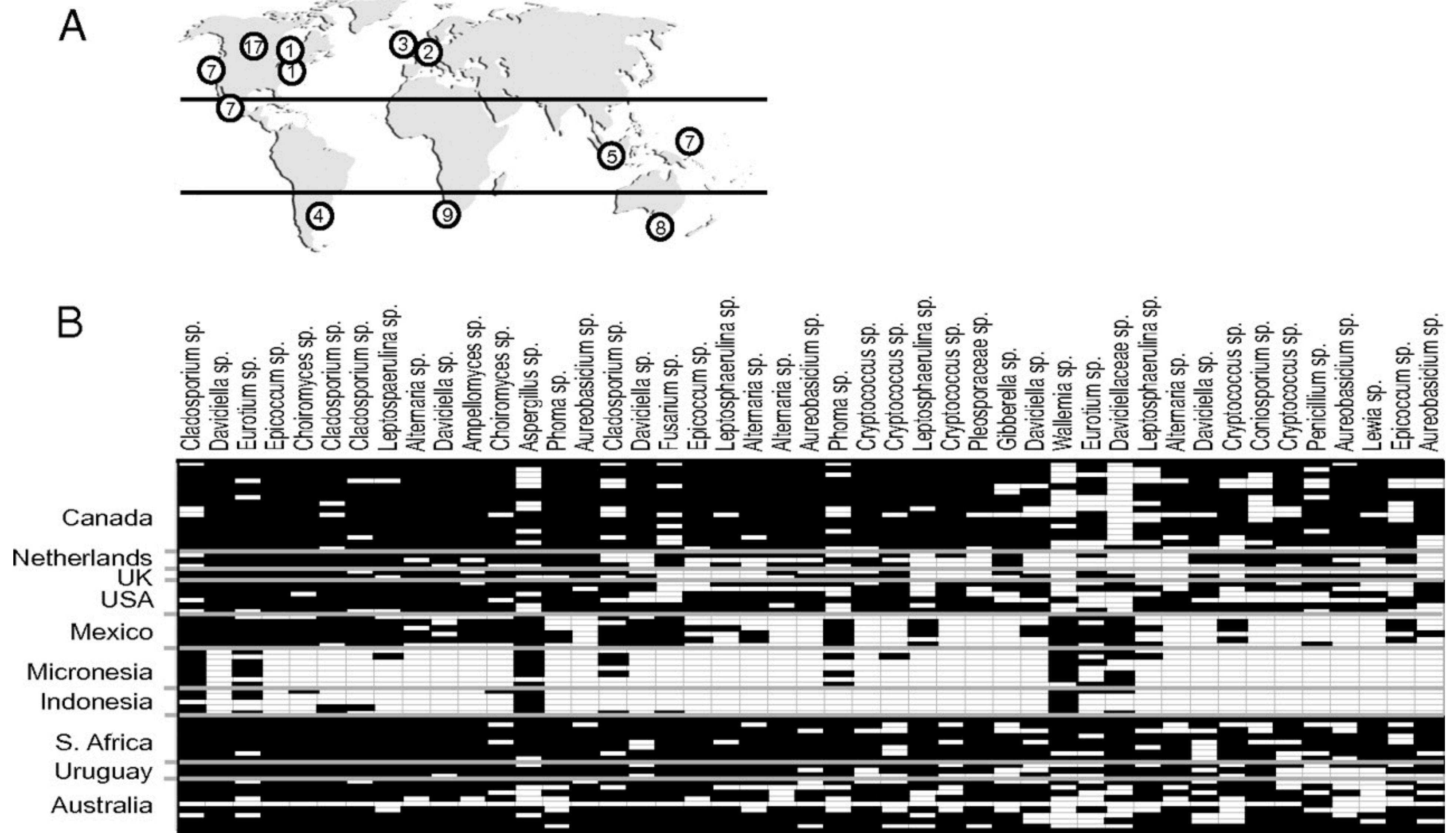


Fig 1. Amend et al 2010

MG-RAST with Fungal Data

Technical
Anthony Amend (UC Berkeley)
Plant and Microbial Biology, United States of America

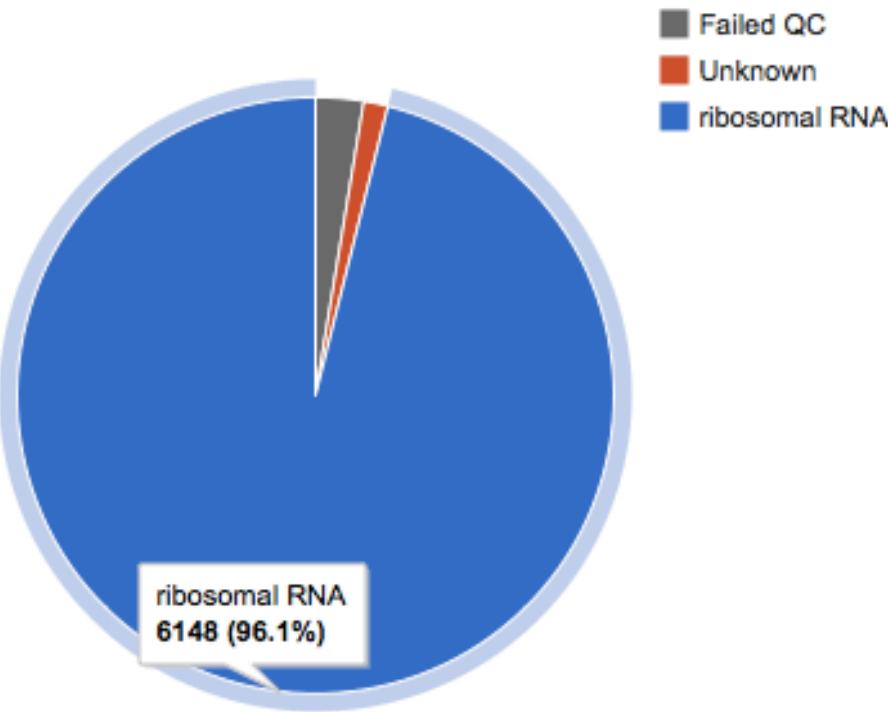
METAGENOMES

Export Jobs Table

display 50 items per page

«first «prev displaying 51 - 100 of 128 next» last»

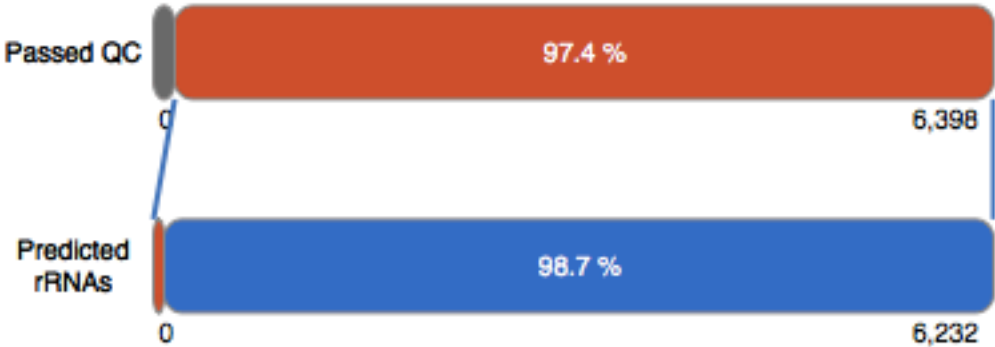
MG-RAST ID	Metagenome Name	Size (bp)	Blome	Location	Country	Sequence Type	Download
			city ; builc			Amplicc	
4484996.3	Reg90_ITS	2,916,194	city ; building ; dust	-	Canada	Amplicon	submitted metadata analysis
4484997.3	SW01_ITS	870,516	city ; building ; dust	-	United States of America	Amplicon	submitted metadata analysis
4484999.3	TT1UY_ITS	2,308,713	city ; building ; dust	-	Uruguay	Amplicon	submitted metadata analysis
4485000.3	TTW4_ITS	1,113,476	city ; building ; dust	-	Netherlands	Amplicon	submitted metadata analysis
4485001.3	VM1_ITS	2,345,057	city ; building ; dust	-	United States of America	Amplicon	submitted metadata analysis
4485002.3	VW01_ITS	1,217,526	city ; building ; dust	-	United States of America	Amplicon	submitted metadata analysis
4485003.3	WL1_ITS	2,440,816	city ; building ; dust	-	Micronesia	Amplicon	submitted metadata analysis
4485004.3	WL2_ITS	1,272,024	city ; building ; dust	-	Micronesia	Amplicon	submitted metadata analysis
4485005.3	WL3_ITS	3,025,402	city ; building ; dust	-	Micronesia	Amplicon	submitted metadata analysis
4485006.3	WL4_ITS	2,121,128	city ; building ; dust	-	Micronesia	Amplicon	submitted metadata analysis
4485007.3	WL5_ITS	2,345,057	city ; building ; dust	-	Micronesia	Amplicon	submitted metadata analysis
4485008.3	WL6_ITS	162,078	city ; building ; dust	-	Micronesia	Amplicon	submitted metadata analysis
4485009.3	WL7_ITS	1,899,453	city ; building ; dust	-	Micronesia	Amplicon	submitted metadata analysis
4485010.3	zZT1UY_ITS	530,191	city ; building ; dust	-	Uruguay	Amplicon	submitted metadata analysis
4485011.3	AA01_ITS	2,056,707	city ; building ; dust	-	United States of America	Amplicon	submitted metadata analysis



Note: Sequences containing multiple predicted features are only counted in one category. Currently downloading of sequences via chart slices is not enabled.

- Technical Data
 - Statistics
 - Metadata
 - Source Distribution
 - Sequence Length Histogram
 - Sequence GC Distribution

166 sequences failed quality control. Of the 6,232 sequences (totaling 2,100,463 bps) that passed quality control, 6,148 (98.7%) produced a total of 507 identified ribosomal RNAs.



Upload: Size	2,916,194 bp
Upload: Sequences Count	6,398
Upload: Mean Sequence Length	455 ± 106 bp
Upload: Mean GC percent	47 ± 4 %
Post QC: Size	2,100,463 bp
Post QC: Sequences Count	6,232
Post QC: Mean Sequence Length	337 ± 0 bp
Post QC: Mean GC percent	47 ± 0 %
Processed: Predicted rRNA Features	591
Alignment: Identified rRNA Features	507

PROJECT INFORMATION

This dataset is part of project [Indoor fungi from temperate to tropical buildings](#).

Pooled samples from multiple indoor environmental regions. Sequencing against ITS and 28S rRNA done for each sample. View paper here: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2922287>

There are 127 other metagenomes in this project

GSC MIXS INFO

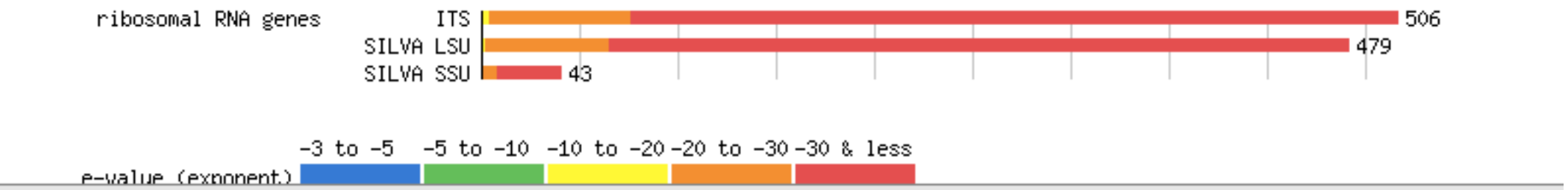
Investigation Type	Metagenome: Amplicon
Project Name	Indoor fungi from temperate to tropical buildings
Latitude and Longitude	50.45472, -104.60667
Country and/or Sea, Location	Canada
Collection Date	2010
Environment (Biome)	city
Environment (Feature)	building

SOURCE HITS DISTRIBUTION [?] HIDE

6,148 (96.1%) of reads had similarity to ribosomal RNA genes.

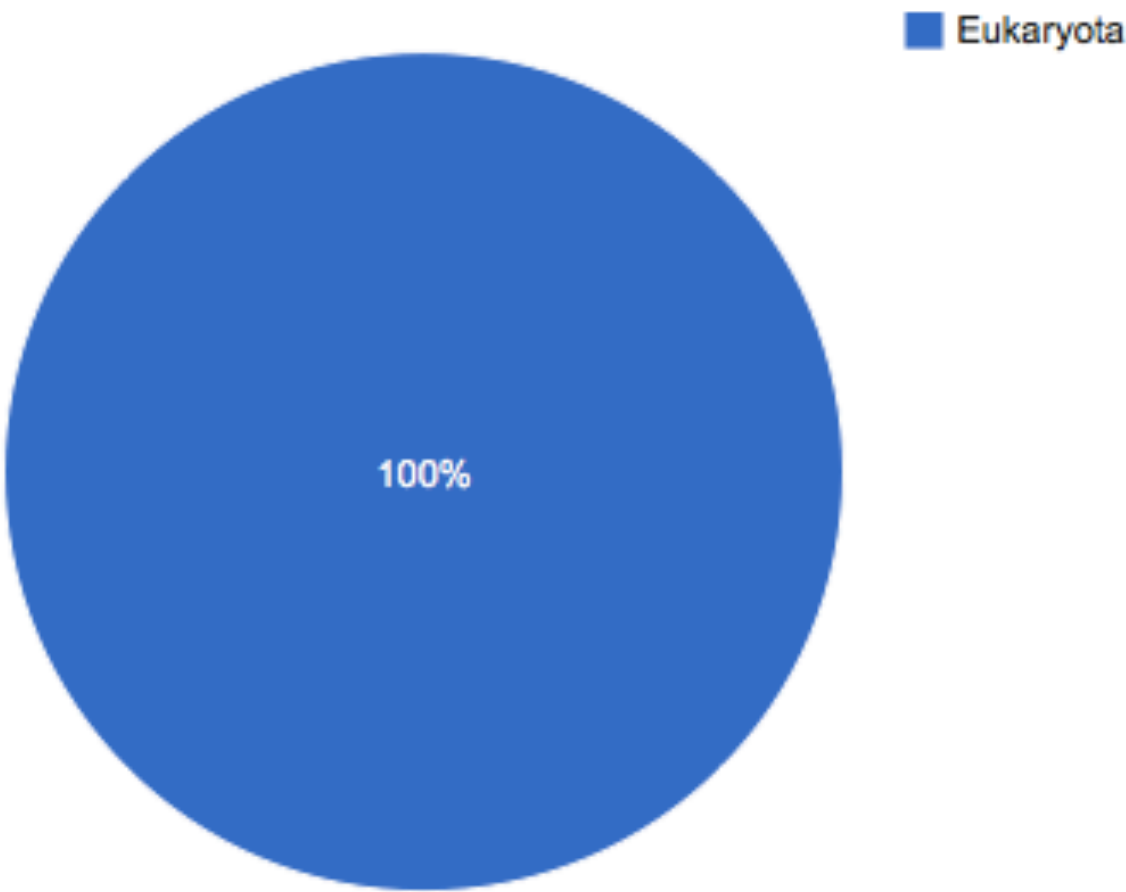
The graph below displays the number of features in this dataset that were annotated by the different databases below. These include databases with functional hierarchy information, and ribosomal RNA databases. The bars representing annotated reads are colored | databases have different numbers of hits, but can also have different types of annotation data.

Download chart data

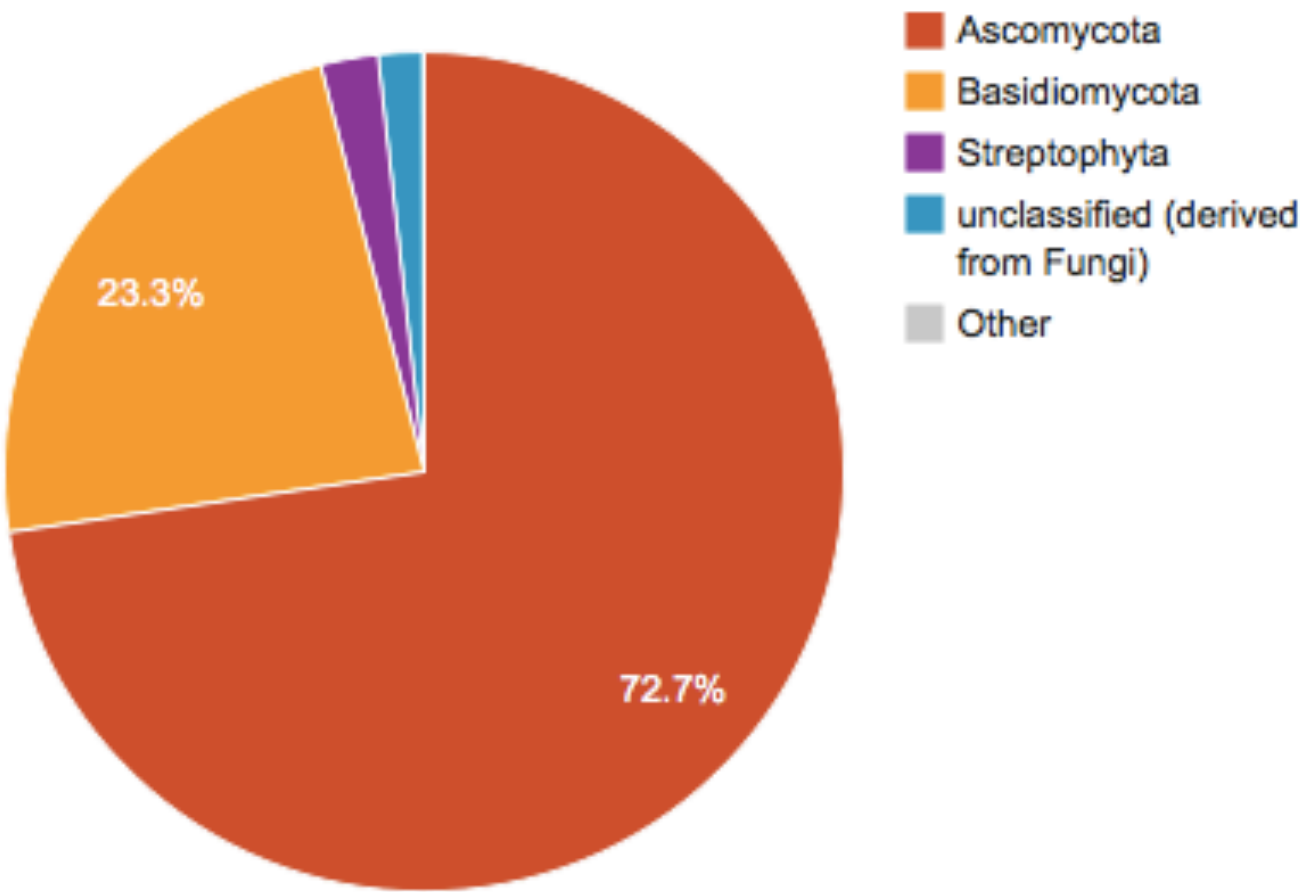


[View taxonomic interactive chart](#)

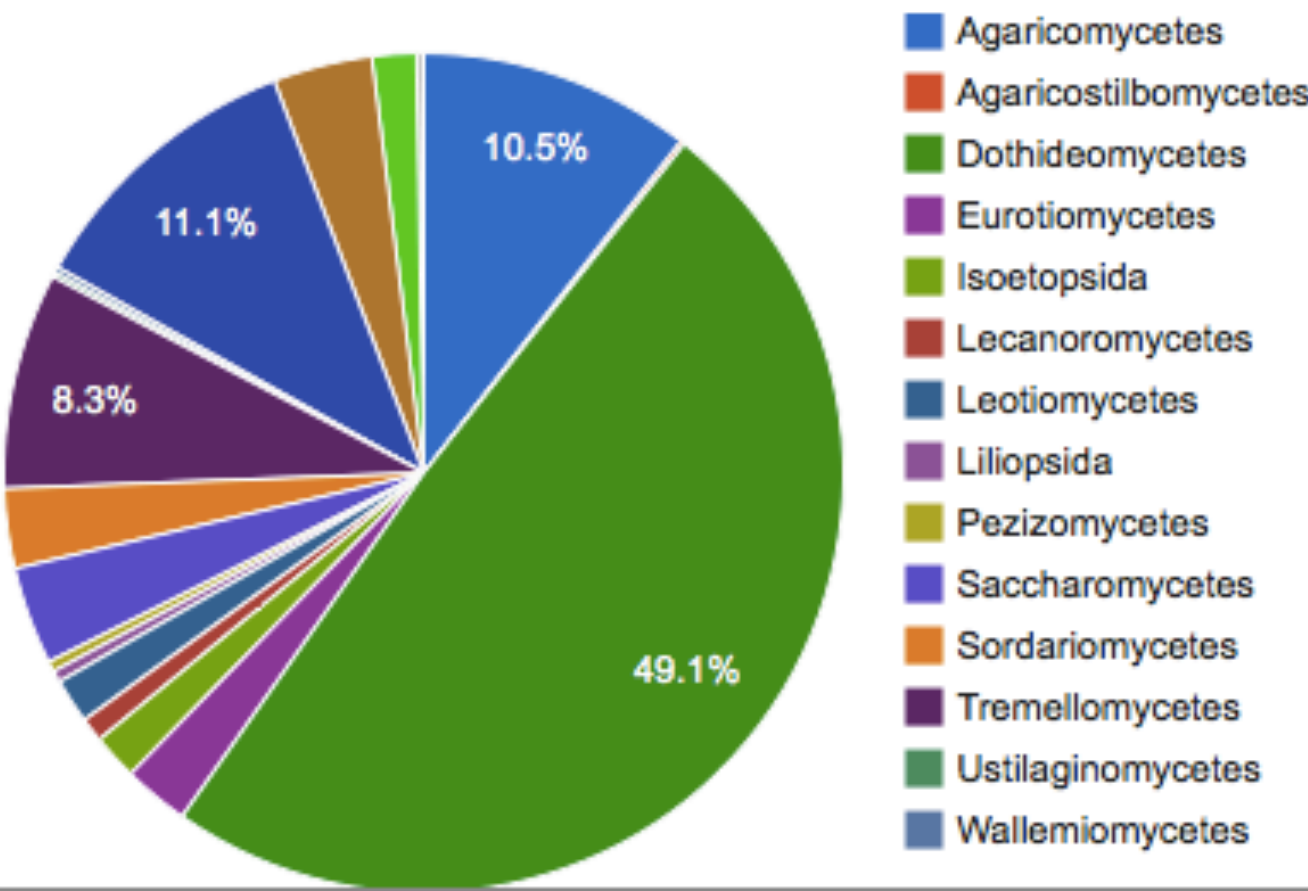
domain [Download chart data](#)



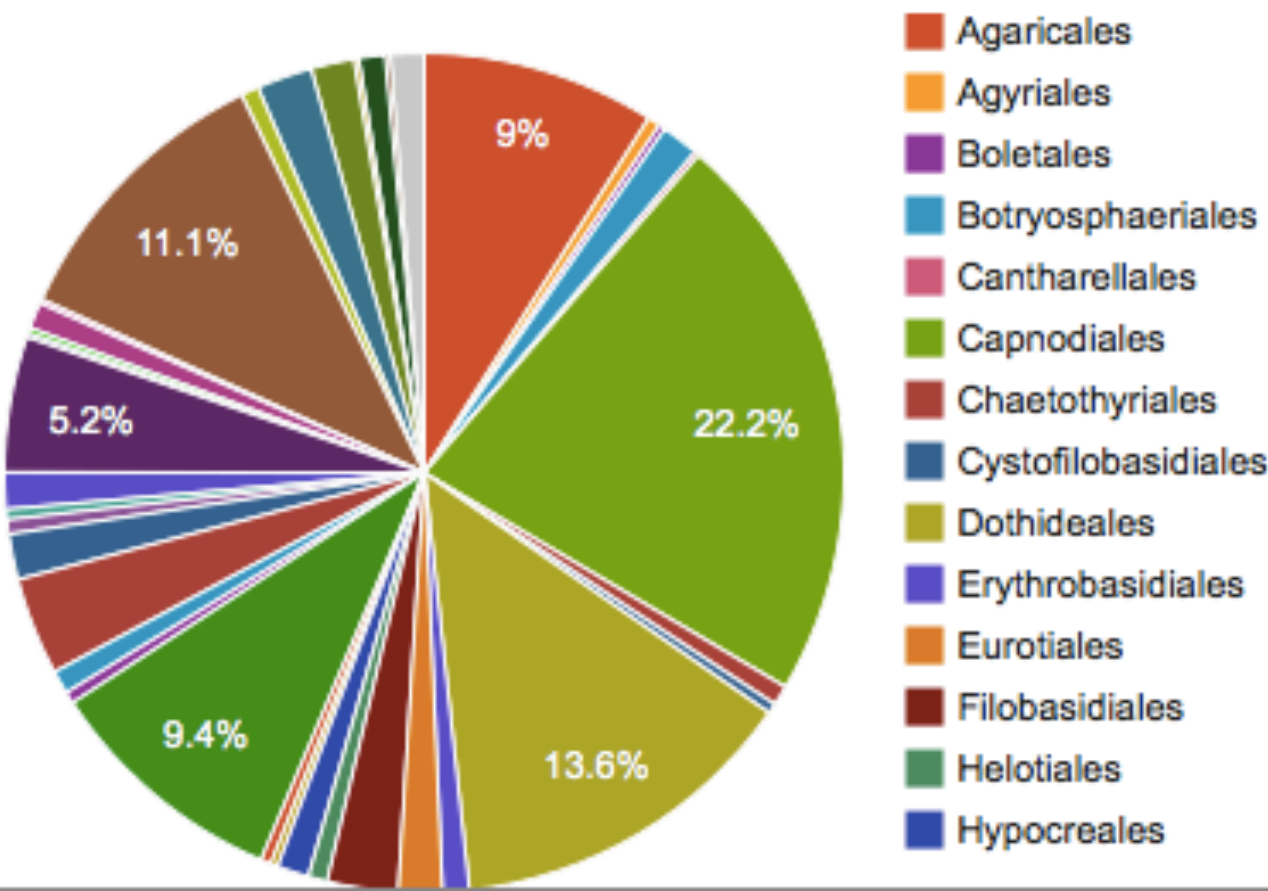
phylum [Download chart data](#)



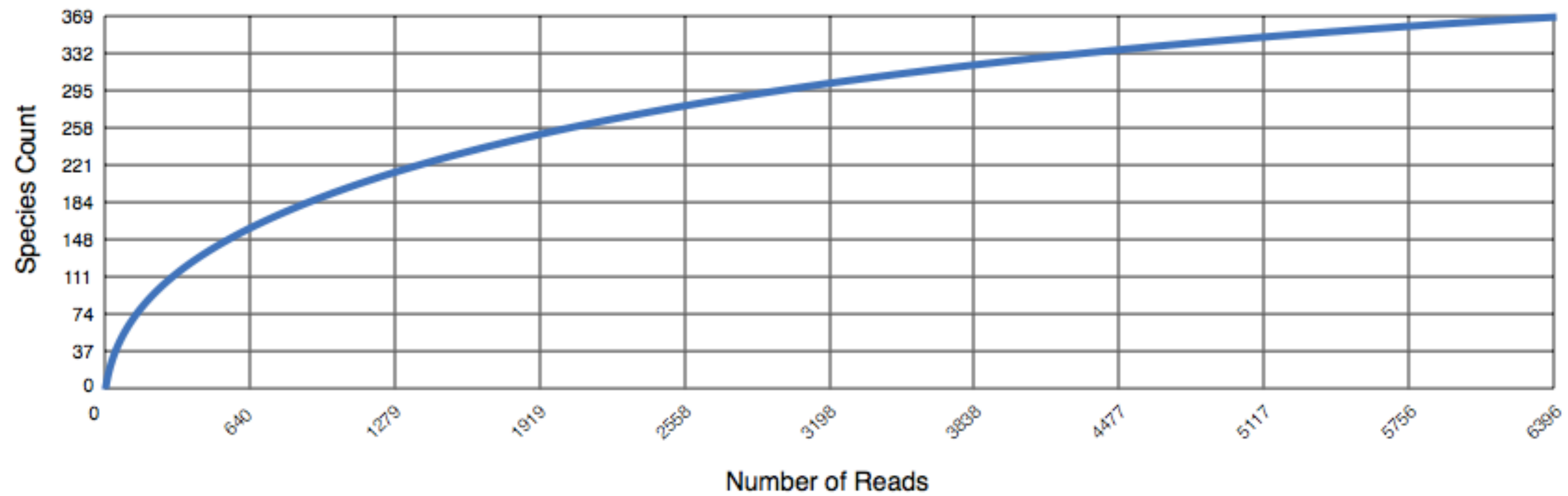
class [Download chart data](#)



order [Download chart data](#)



Hits summarized by different taxonomic levels



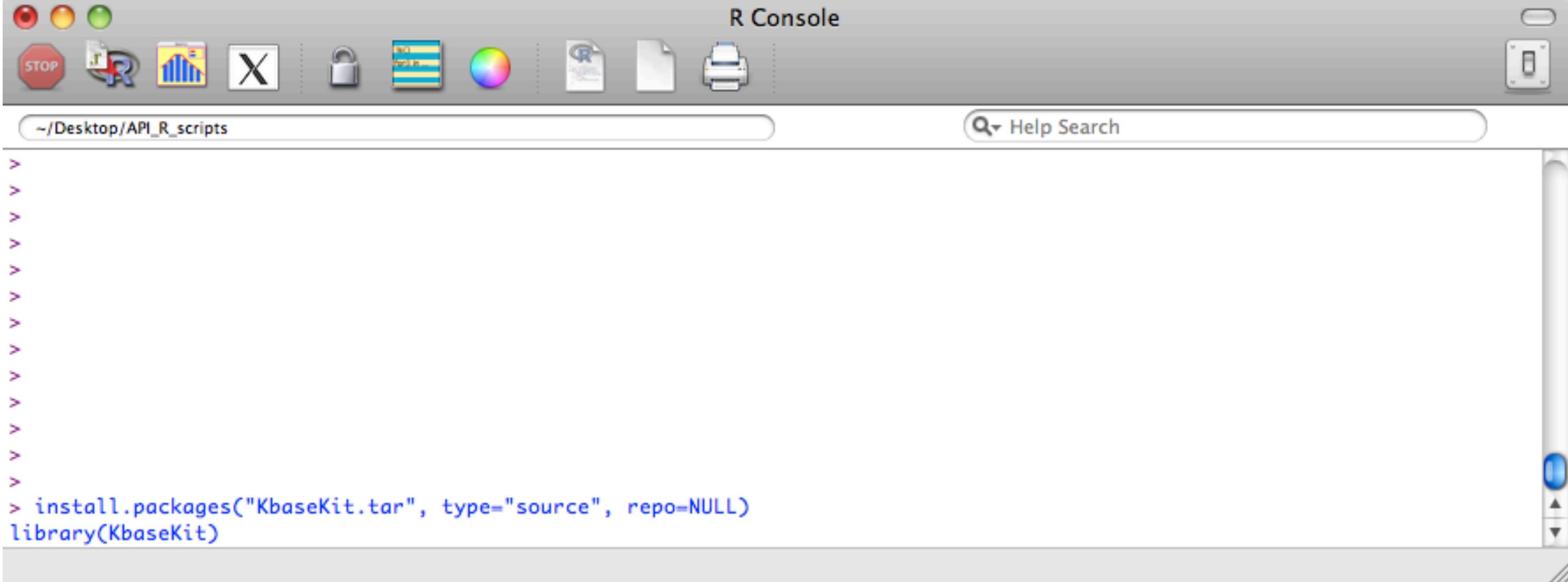
Rarefaction curve (1 sample)

Summary of Indoor Fungal Metagenomes using MG-RAST tools

Analyses performed with KbaseKit R package
(Kevin Keegan, Daniel Braithwaite)

R package to download and analyze MG-RAST annotated data

Install the KbaseKit

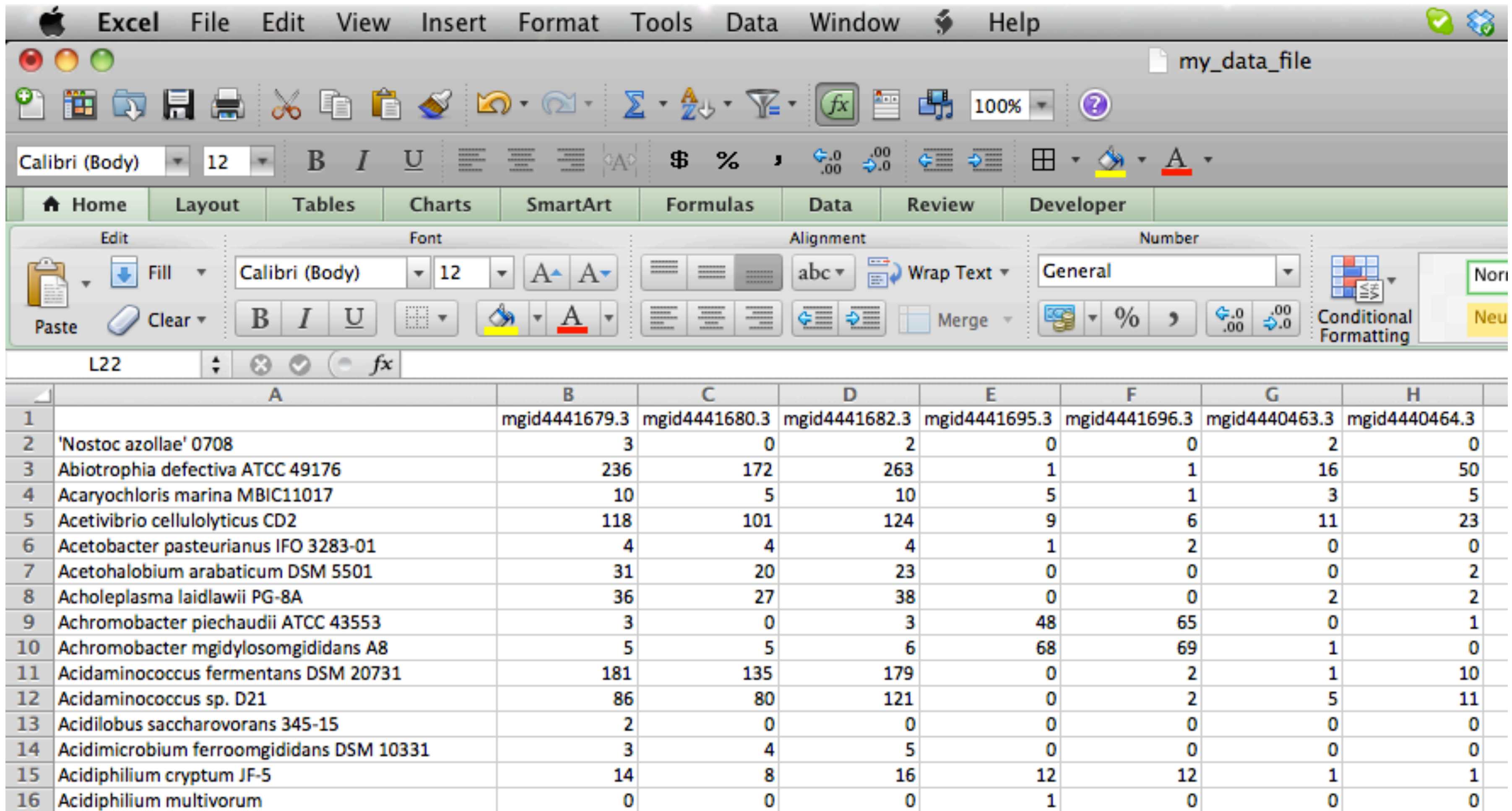


Batch download data from MG-RAST

```
> my_data <- kbGet( "4441679.3;4441680.3;4441682.3;4441695.3;4441696.3;4440463.3;4440464.3", "abundance", namespace="SEED", param="format/plain" )
```

Save in simple format for R, Matlab, Excel etc.

```
> write.table(my_data, file = "my_data_file", col.names=NA, row.names = TRUE, sep="\t", quote=FALSE)
```

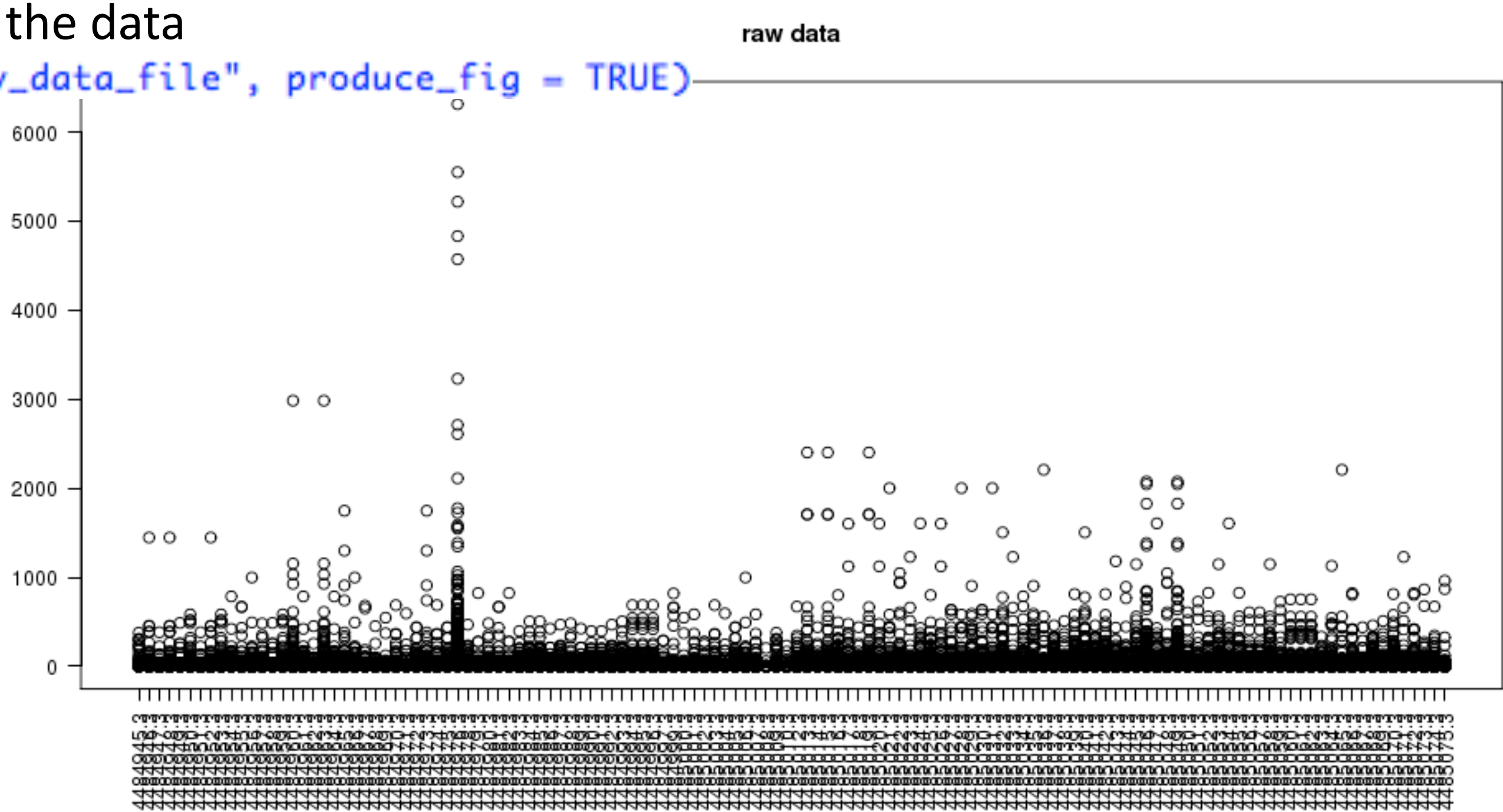


	A	B	C	D	E	F	G	H
1		mgid4441679.3	mgid4441680.3	mgid4441682.3	mgid4441695.3	mgid4441696.3	mgid4440463.3	mgid4440464.3
2	'Nostoc azollae' 0708	3	0	2	0	0	2	0
3	Abiotrophia defectiva ATCC 49176	236	172	263	1	1	16	50
4	Acaryochloris marina MBIC11017	10	5	10	5	1	3	5
5	Acetivibrio cellulolyticus CD2	118	101	124	9	6	11	23
6	Acetobacter pasteurianus IFO 3283-01	4	4	4	1	2	0	0
7	Acetohalobium arabaticum DSM 5501	31	20	23	0	0	0	2
8	Acholeplasma laidlawii PG-8A	36	27	38	0	0	2	2
9	Achromobacter piechaudii ATCC 43553	3	0	3	48	65	0	1
10	Achromobacter mgidylosomgididans A8	5	5	6	68	69	1	0
11	Acidaminococcus fermentans DSM 20731	181	135	179	0	2	1	10
12	Acidaminococcus sp. D21	86	80	121	0	2	5	11
13	Acidilobus saccharovorans 345-15	2	0	0	0	0	0	0
14	Acidimicrobium ferroomgididans DSM 10331	3	4	5	0	0	0	0
15	Acidiphilium cryptum JF-5	14	8	16	12	12	1	1
16	Acidiphilium multivorum	0	0	0	1	0	0	0

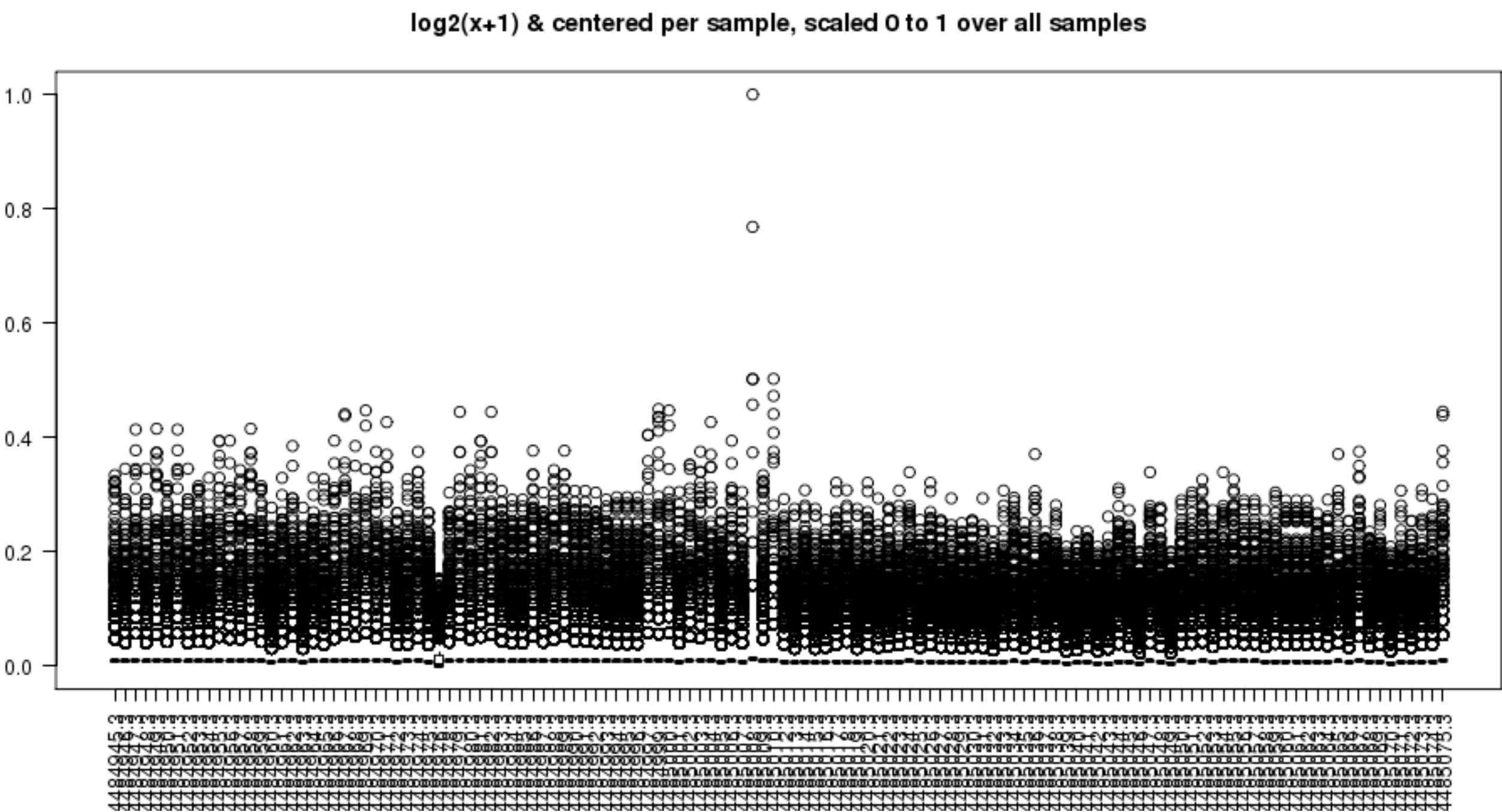
Preprocess (normalize and standardize the data)

```
> preprocessing(file_in = "my_data_file", produce_fig = TRUE)
```

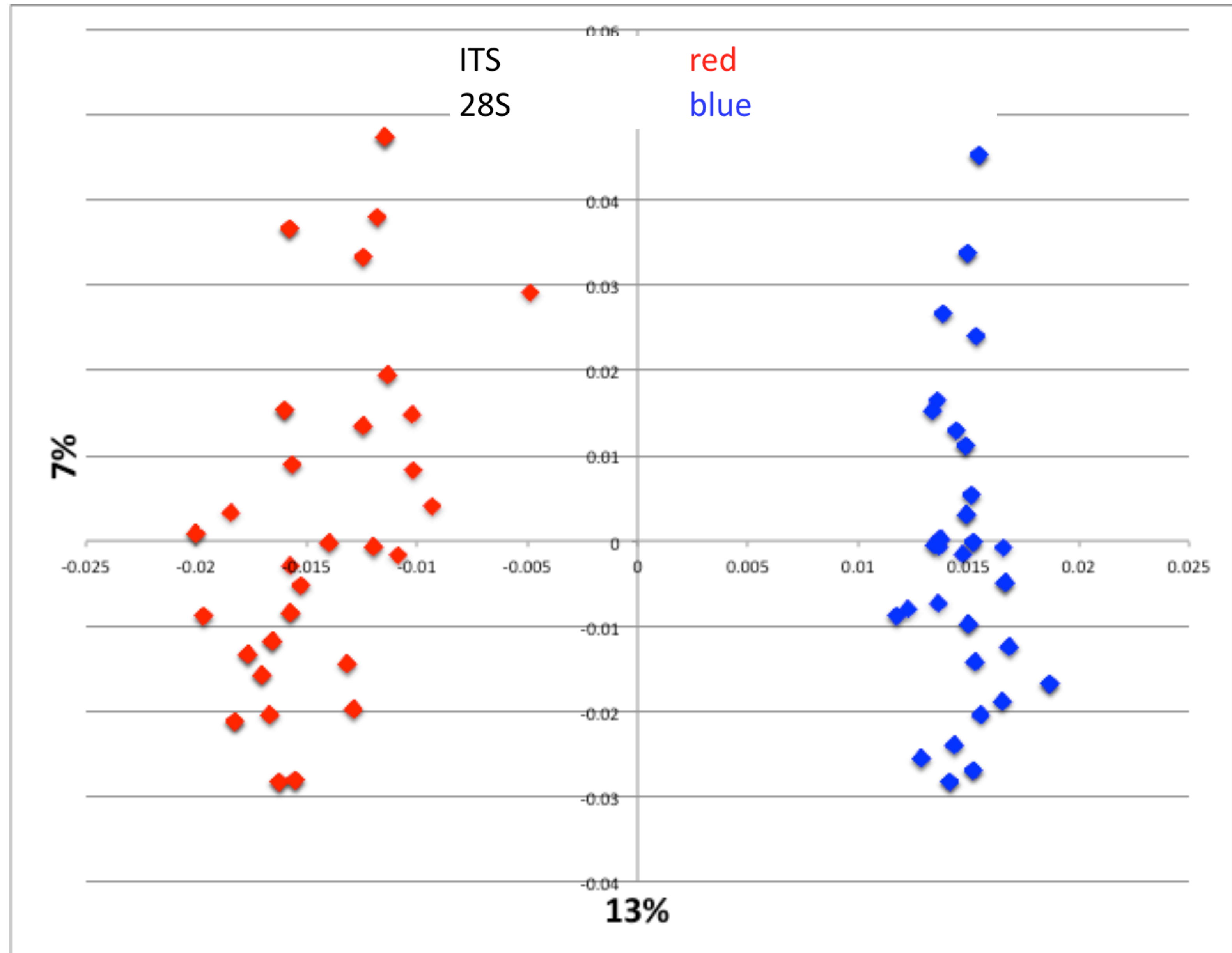
Distribution of species level taxonomic abundances for 128 samples



After normalization and standardization, data are more comparable, but non-normal




```
> plot_pco(file_in = "my_data_file.preprocessed_data", dist_method = "euclidean", header = 1 )
```

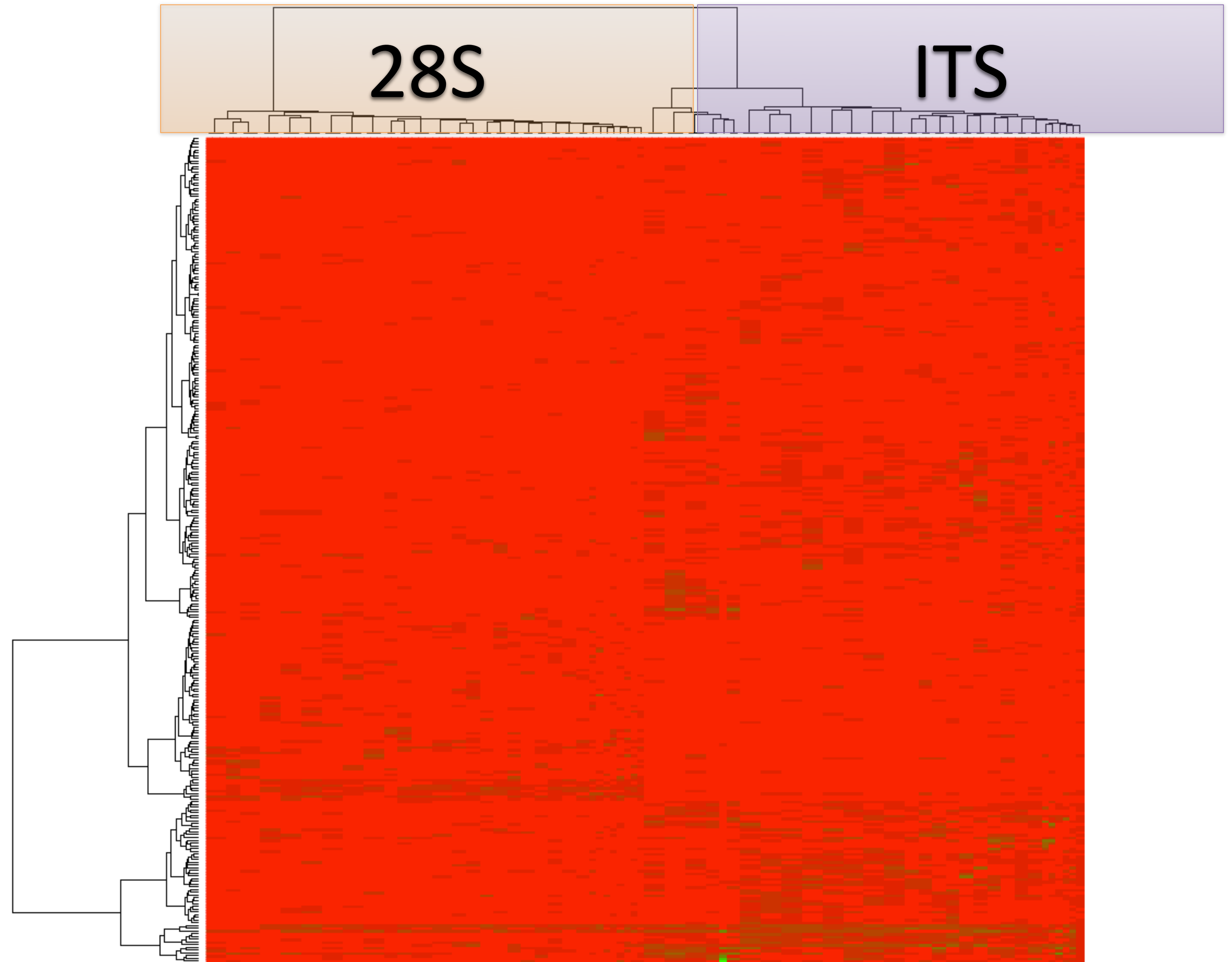


PCA of normalized counts – Painted by rRNA type

```
> heatmap_dendrogram(file_in = "my_data_file.preprocessed_data")
```

Abundance
Red(low) -> Green(high)

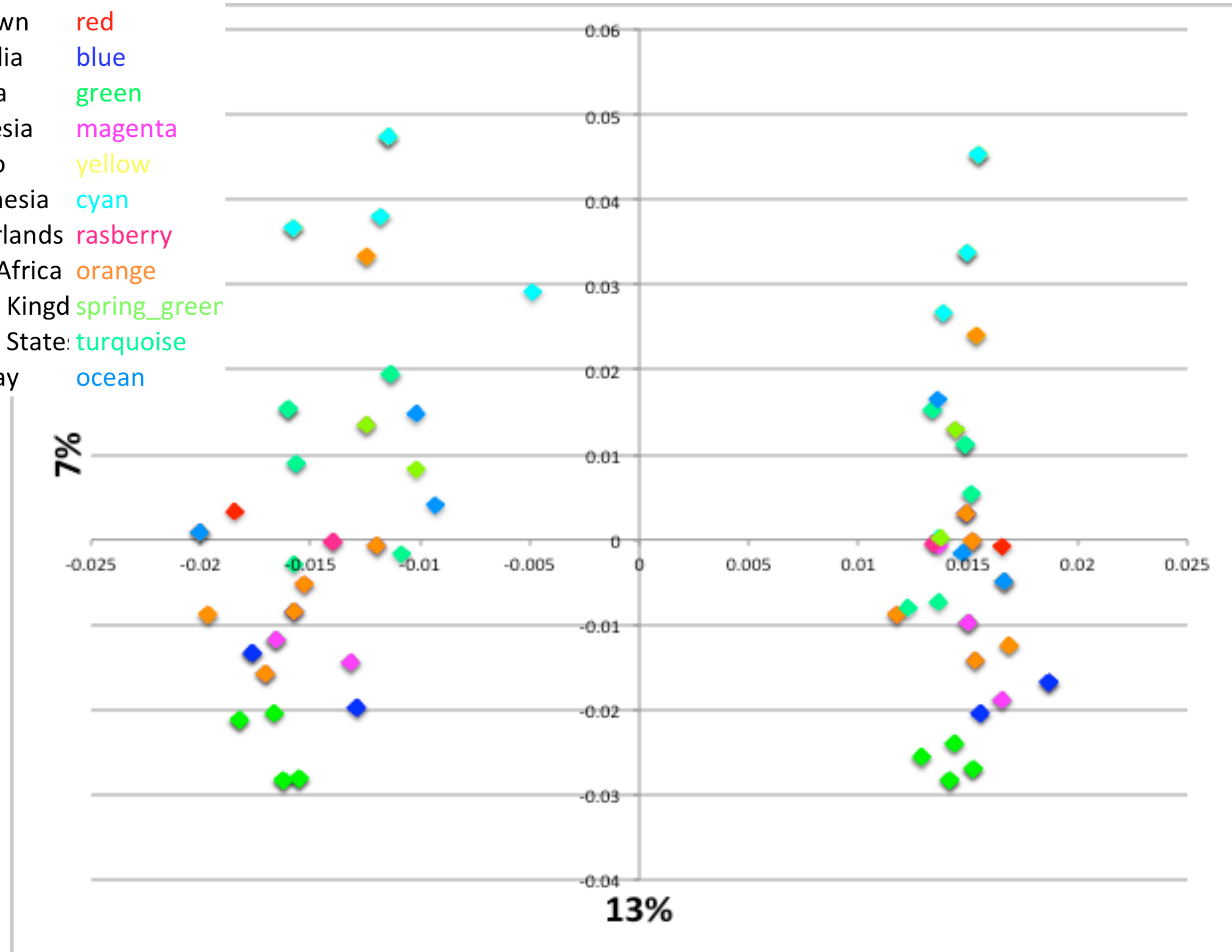
10% most abundant taxa
that are significantly different
between 28S and ITS
(Mann Whitney test –
Bonferroni adjusted p-value
0.05)



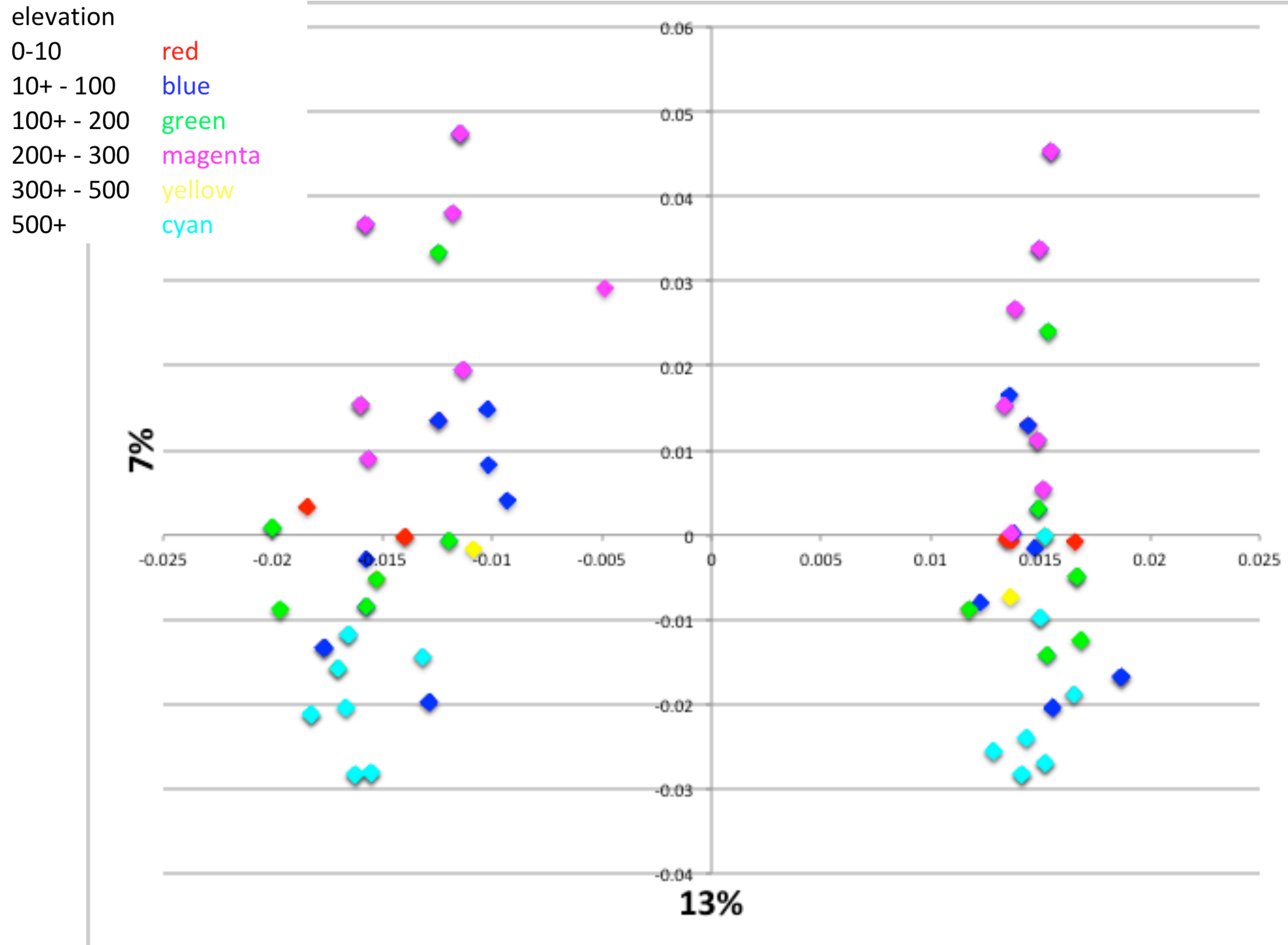
Investigated ITS and 28S samples to determine taxa that exhibit the most significant differences

PCA of normalized counts – Painted by sampled country

unknown red
 Australia blue
 Canada green
 Indonesia magenta
 Mexico yellow
 Micronesia cyan
 Netherlands raspberry
 South Africa orange
 United Kingd spring_green
 United State: turquoise
 Uruguay ocean

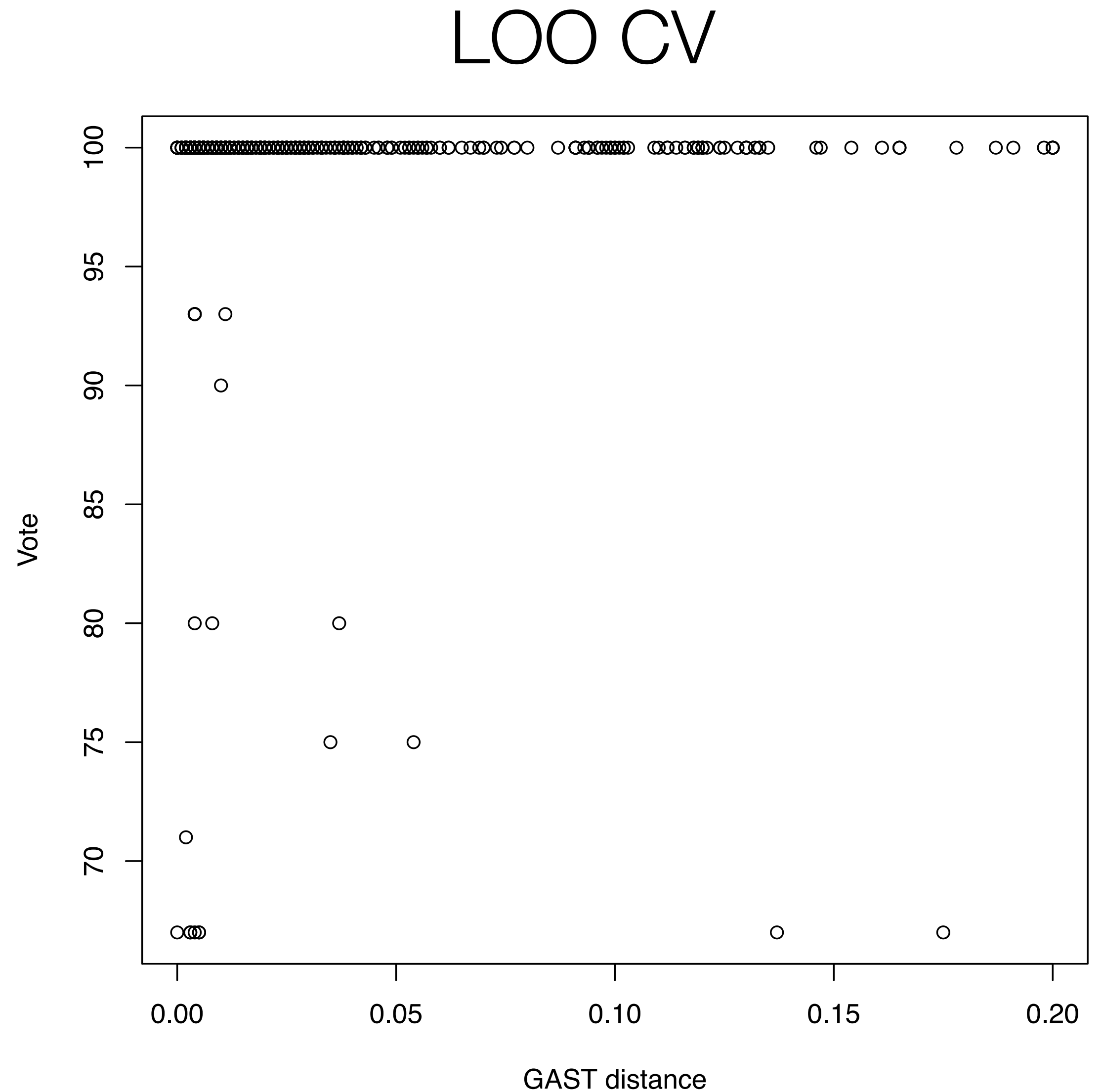


PCA of normalized counts – Painted by sampled elevation



VAMPS with Fungal data

- Testing the use of GAST and the UNITE ITS database on the Amend et al data.
- Good recall for this dataset - 8% of data is unknown, but still evaluating correctness of assigned taxa
- Have also tested leave-one-out cross validation with test ITS data and there is reasonable ability to recall taxa.
- With MBL team, be testing an integration of ITS data into the standard VAMPS analyses.



Thanks

UCR

Steven Ahrendt

Daniel Borchering
Raghu Ramamurthy (FungiDB)

VAMPS-MBL

Sue Huse

Anna Shipunova

Mitch Sogin

QIIME

Gail Ackerman

Jesse Stombaugh

Rob Knight

MG-RAST

Daniel Braithwaite

Travis Harrison

Kevin Keegan

Andreas Wilke