RDP: Data and Tools for Microbial Community Analysis

James Cole

August 19, 2014

Center for Microbial Ecology
Dept. of Plant, Soil and Microbial Sciences
Michigan State University

RDP Specialized Tools for 16S rRNA Analysis

Ribosomal Database Project: data and tools for high throughput rRNA analysis

James R. Cole, Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske and James M. Tiedje (2014) *Nucl. Acids Res.* 41:D633-D642.

RDP's Popular Online Tools

Interactive tools

Browsers -- browse and select from taxonomic hierarchy; powerful search and selection features

RDP Classifier -- places sequences into bacterial taxonomy; fast and accurate

ProbeMatch -- fast search algorithm, limit searches to specific regions

SeqMatch -- finds nearest neighbor, more accurate than BLAST

RDPipeline and FunGene tools for gene-targeted metagenomics

MIMARKS GoogleSheets -- helps organizing standards Compliant metadata

***myRDP* space** -- upload and analyze your own 16S sequences in your private space

http://rdp.cme.msu.edu/

RDP Specialized Tools for Fungal Analysis

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Aug. 2007, p. 5261–5267 Vol. 73, No. 16
0099-2240/07/$08.00+0 doi:10.1128/AEM.00062-07
Copyright © 2007, American Society for Microbiology. All Rights Reserved.

Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy▽†

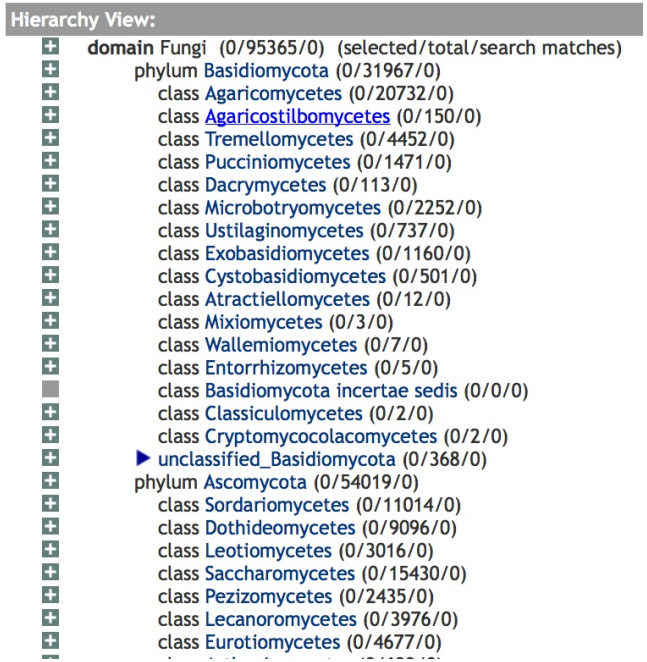
Qiong Wang,[1] George M. Garrity,[1,2] James M. Tiedje,[1,2] and James R. Cole[1]*

- Rapidly assigns sequences into bacterial, archaeal and fungal taxonomies
- Works well on partial or full-length sequences
- Bootstrap confidence estimates
- Can be easily trained on alternative taxonomies.
- Source code is freely available

RDP Fungal LSU Alignment & Classifier

RDP Fungal 28S Classifier (Liu *et al.*, 2012) was updated with a new training set providing increased coverage of the Glomeromycota, Chytridiomycota, and other basal lineages and better performance in separating fungal from non-fungal eukaryotes.

Fungal 28S Classifier is 460 times faster than BlastN and provides equal or superior accuracy (Liu *et al.*, 2012)

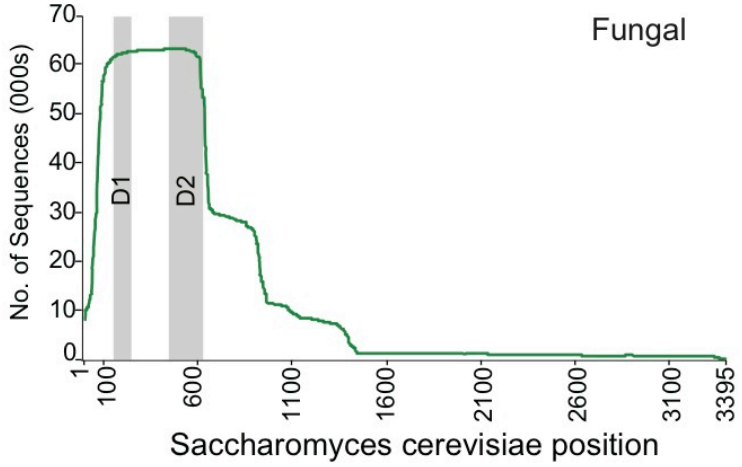


Fungal LSU Training Set No.11
Taxonomic Composition of Major Ranks

Rank	Count
Domain	7
Phylum	27
Class	72
Order	156
Family	475
Genus	1895

RDP Fungal LSU Aligner

- Uses a covariance model built from 183 LSU sequences from complete fungal genomes and the fungal sequence set from the Comparative RNA Web Site
- Sequences cover four major fungal phyla: Ascomycota, Basidiomycota, Chytridiomycota, and Blastocladiomycota
- The model includes the combined 5.8S and 28S sequences, useful separately or together
- The Fungal 28S Aligner and its model are available on the RDP site and the RDP GitHub repository, for easy inclusion in your local workflow



Cole, J. *et al.*, 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucl. Acids Res.* 41:D633-D642.

RDP Fungal ITS Classifier

- **DOE_SFA** set: published hand-curated set (Porras-Alfaro *et al.*, 2014). It contains lineage only to the genus level.
- **Warcup** set: provided by Paul Greenfield and Vinita Deshpande of the Australian Commonwealth Scientific and Industrial Research Organization. It also incorporates some training sequences from DOE_SFA and UNITE ref sets. It contains lineages to the species level.
- **UNITE** set: A set consisting of UNITE core sequences (excluding chimeric and low quality) for each dynamic species hypothesis (Koljalg *et al.*, 2013).

We tested the UNITE set twice, once using the UNITE taxonomy (**UNITE_name** set) and a second time, using a concatenation of the UNITE "Species hypotheses" accession code number and UNITE terminal taxon name to group sequences into terminal taxa (**UNITE_sh** set).

Example:

UNITE_name set: "Cortinarius caesiocortinatus"
UNITE_sh set: two terminal taxa "Cortinarius_caesiocortinatus|SH192002.06FU"
and "Cortinarius_caesiocortinatus|SH192062.06FU"

Taxonomic Composition of Major Ranks

Rank	Warcup	DOE_SFA	UNITE
domain (kingdom)	1	2	2
phylum	8	11	10
class	40	36	45
order	131	118	167
family	364	328	523
genus	1,620	1,134	2,135
species	8,967	NA	20,221*
Unique Sequences	17,923	6,889	145,019

* The **UNITE_sh** has 20,221 species level taxa, the **UNITE_name** has 10,346

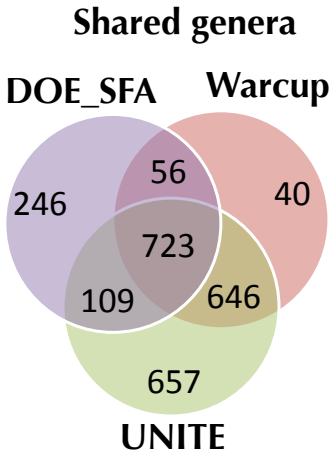
ITS Reference Set Composition

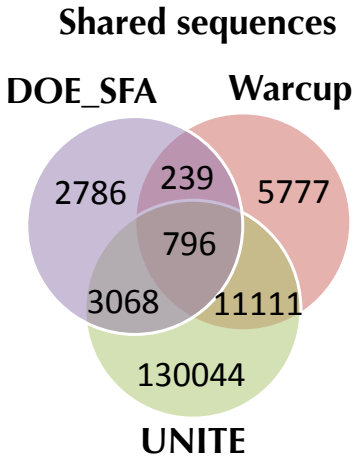
Table 2a: Shared genera

	Warcup	DOE_SFA	UNITE
Warcup		48%	85%
DOE_SFA	69%		73%
UNITE	64%	39%	

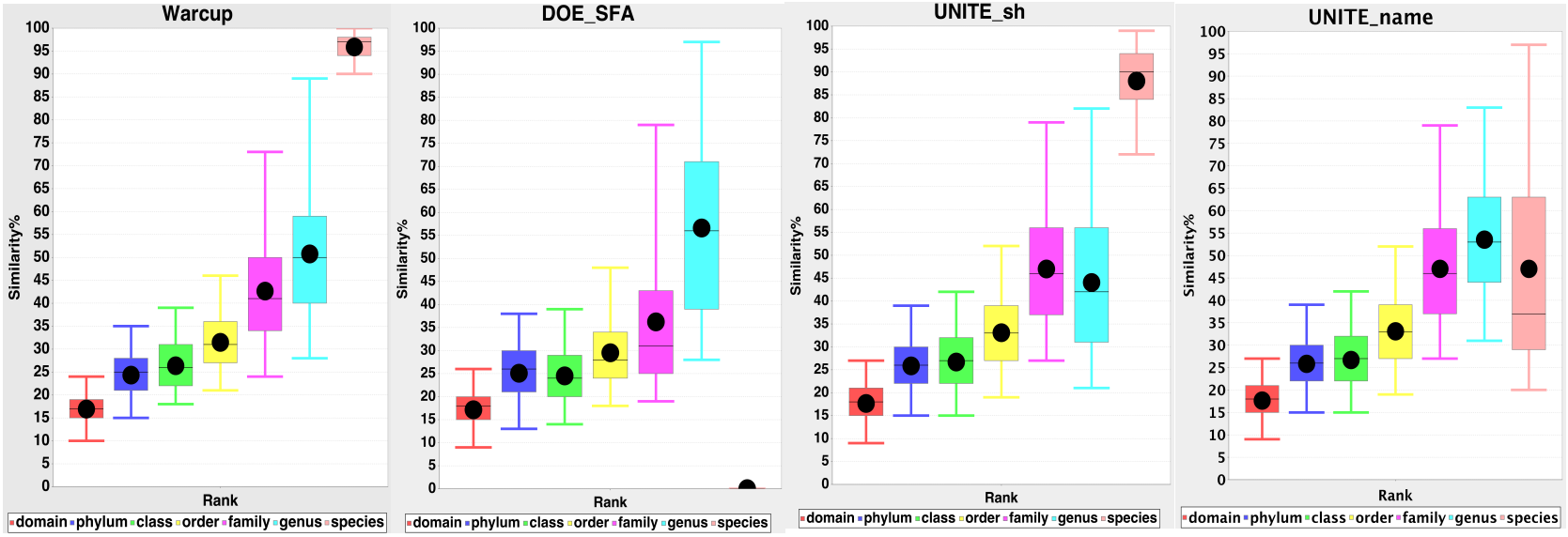
Table 2b: Shared Sequences

	Warcup	DOE_SFA	UNITE
Warcup		6%	66%
DOE_SFA	15%		56%
UNITE	8%	3%	



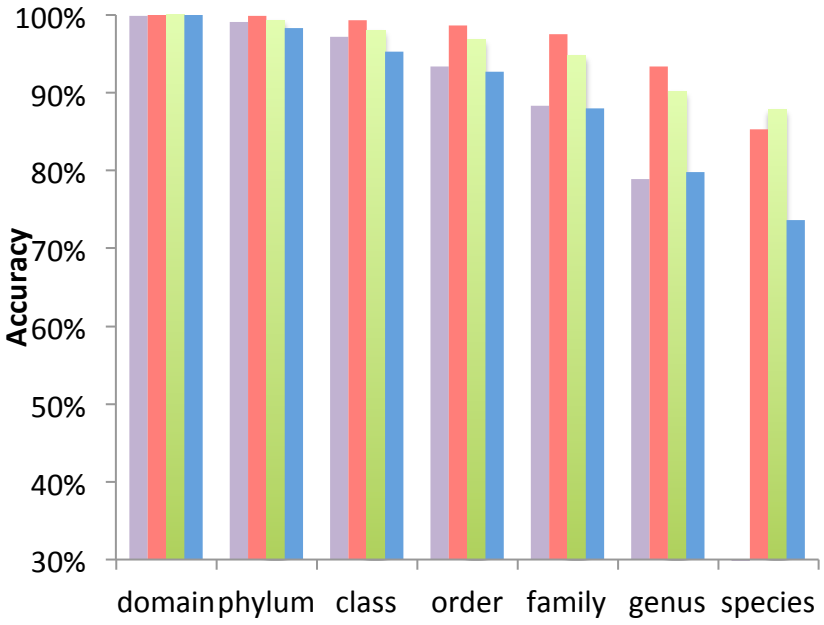


Taxon Similarity

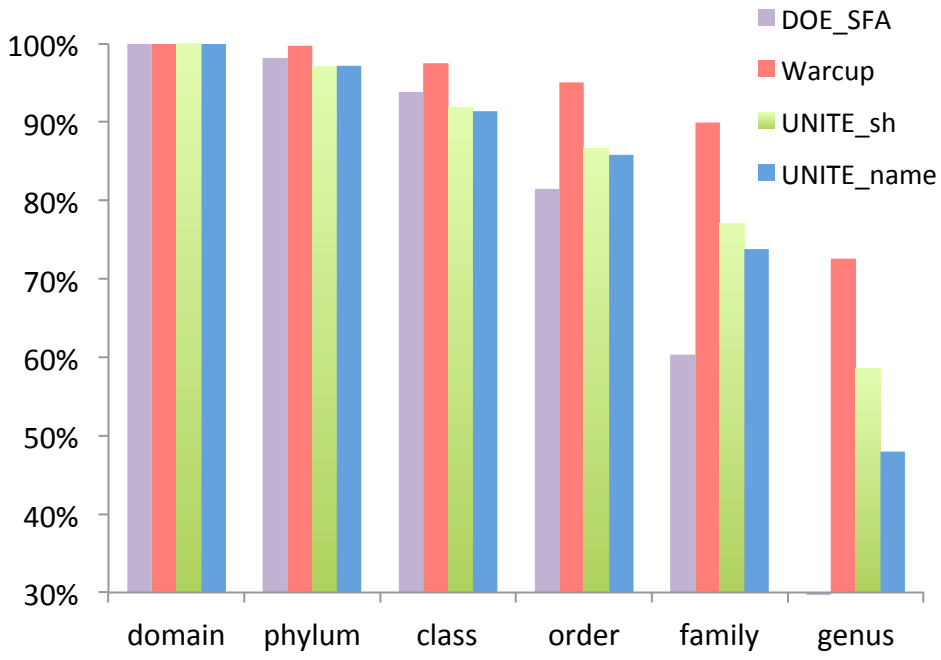


RDP Fungal ITS Classifier Accuracy

Leave-one-sequence-out



Leave-one-taxon-out



RDP Specialized Tools for Functional Gene Data Analysis

Genes Beyond rRNA

- Faster evolving phylogenetic markers
  - Full-length 16S rRNA resolves genera, not species
  - Species (still) defined as 70% DNA-DNA reassociation. Average Nucleotide Identity (ANI) has been suggested as a replacement
- Genes encoding important ecological functions (eco-functional genes)
  - Ecologically important genes (e.g., carbon and nitrogen cycling, biogeochemical processes)

SSU rRNA similarity vs. taxonomic rank

Species threshold: 98.7%

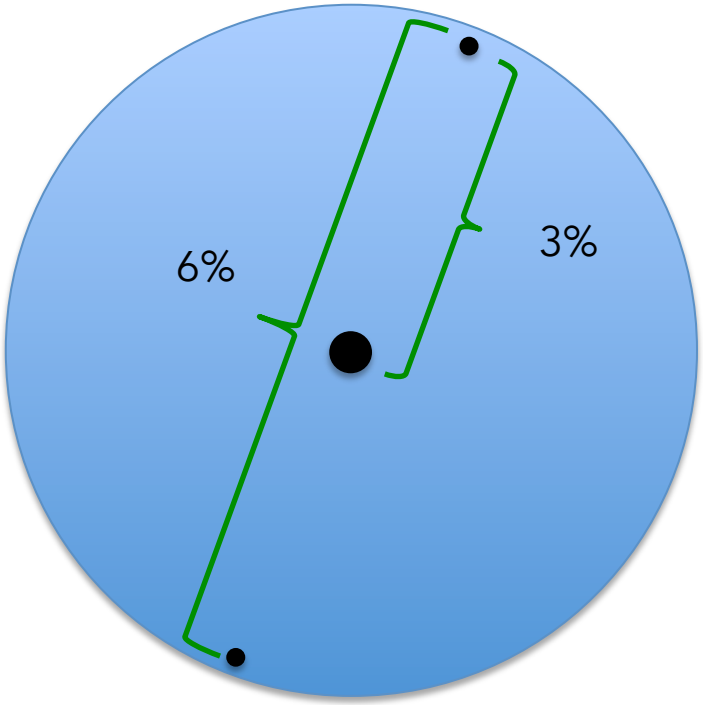
Stackebrandt & Ebers. *Microbiol. Today* **8**, 6–9 (2006).

	Genus	Family
Number of taxa	568	201
Median sequence identity	96.4% (96.2, 96.55)	92.25% (91.65, 92.9)
Minimum sequence identity	94.8% (94.55, 95.05)	87.65% (86.8, 88.4)
Threshold sequence identity	94.5%	86.5%

Modified from: Yarza et al. *Nature Reviews Microbiology* **12**, 635–645 (2014)

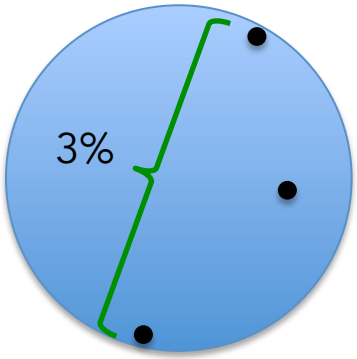
Common Cluster Methods

Reference based heuristic
97% identity Clusters



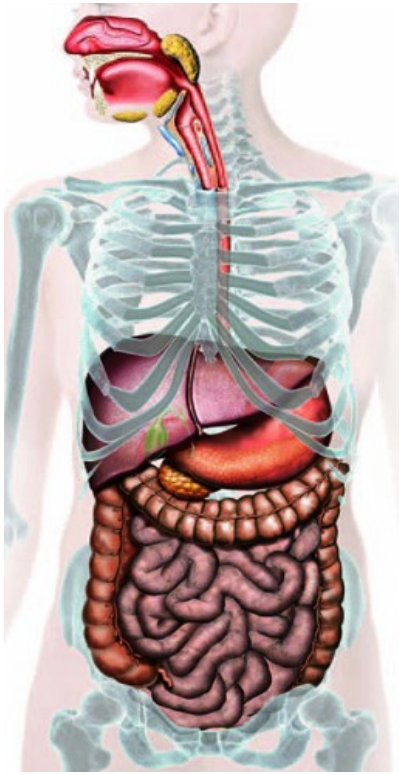
CD-HIT
UCLUST

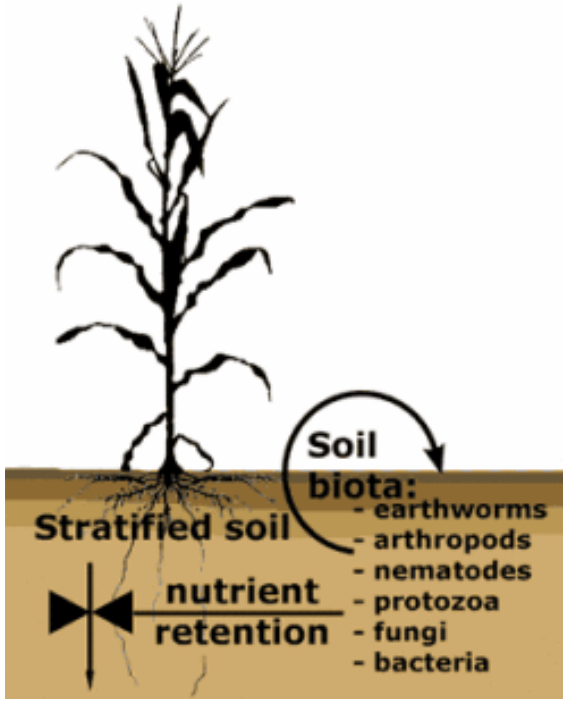
Complete linkage
97% identity Clusters

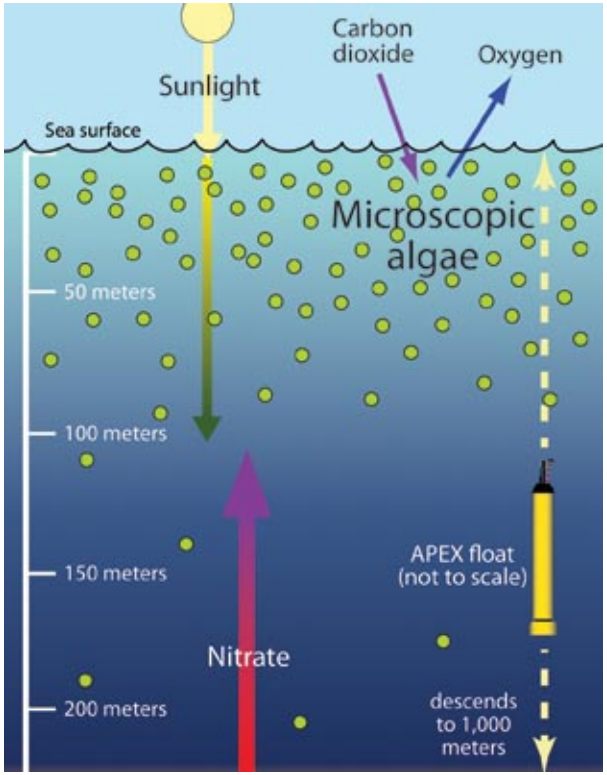


RDP mcCLUST
Qiime
Mothur

Important (eco)functional genes: environment-dependent







RDP Fungene

GENE CATEGORIES:

PHYLOGENETIC MARKERS
BIOGEOCHEMICAL CYCLES
BIODEGRADATION
PLANT PATHOGENICITY
METAL CYCLING
ANTIBIOTIC RESISTANCE

PIPELINE TOOLS:

PIPELINE INITIAL PROCESS
DEFINED COMMUNITY ANALYSIS
DEREPLICATOR
USEARCH CHIMERA CHECK
FRAMEBOT
ALIGNER
MCCLUST
EXPAND MAPPINGS
REPRESENTATIVE SEQUENCES
RAREFACTION
SHANNON & CHAO1 INDEX
JACCARD & SØRENSEN INDEX

- Offers databases of many common ecofunctional genes and proteins
- Integrated tools allow researchers to browse these collections and choose subsets for further analysis, build phylogenetic trees, test primers and probes for coverage, and download aligned sequences
- Specialized tools to process coding gene amplicon data

Fish, J. A., B. Chai, Q. Wang, Y. Sun, C. T. Brown, J. M. Tiedje, and J. R. Cole. 2013. FunGene: the Functional Gene Pipeline and Repository. *Front. Microbiol.* **4**:291

Ecological Patterns of *nifH* Genes in Four Terrestrial Climatic Zones Explored with Targeted Metagenomics Using FrameBot, a New Informatics Tool

Qiong Wang, John F. Quensen III, Jordan A. Fish, Tae Kwon Lee, Yanni Sun, James M. Tiedje, James R. Cole (2013)

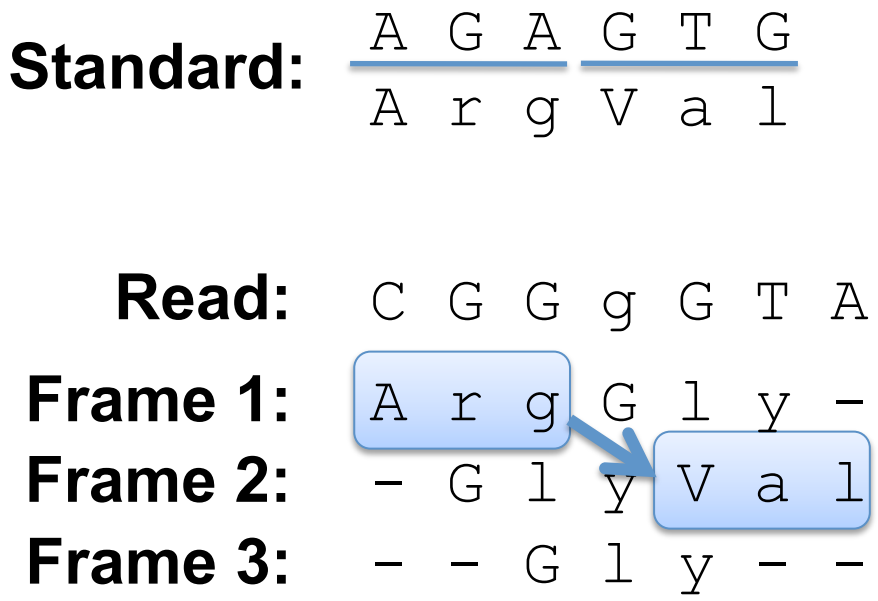
***mBio* 4(5): e00592-13, doi: 10.1128/mBio.00592-13**

FrameBot

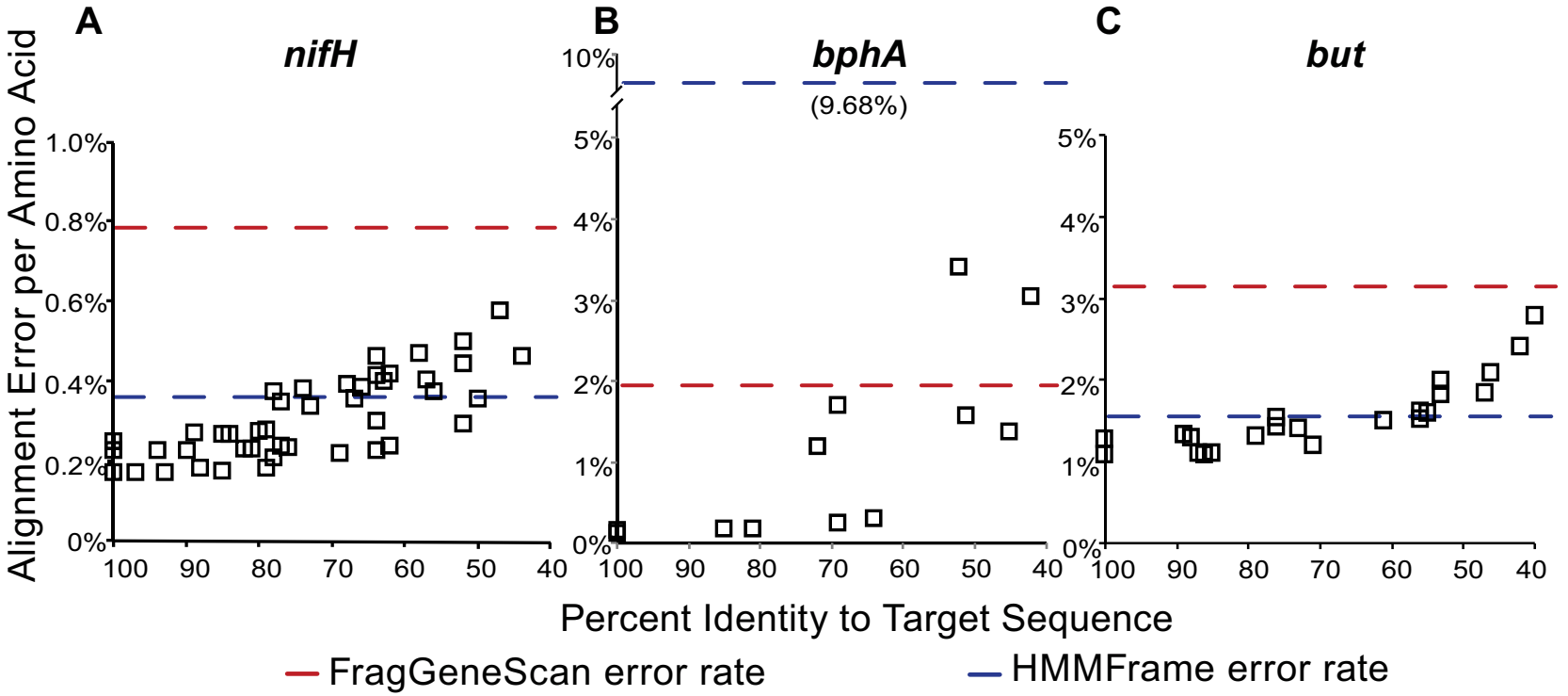
frameshift correction and nearest neighbor assignment

FrameBot Algorithm

- Extends an existing dynamic programming algorithm
- Requires a reference sequence set
- Returns the frameshift-corrected protein and DNA query sequences
- Reports the nearest match protein with the best score

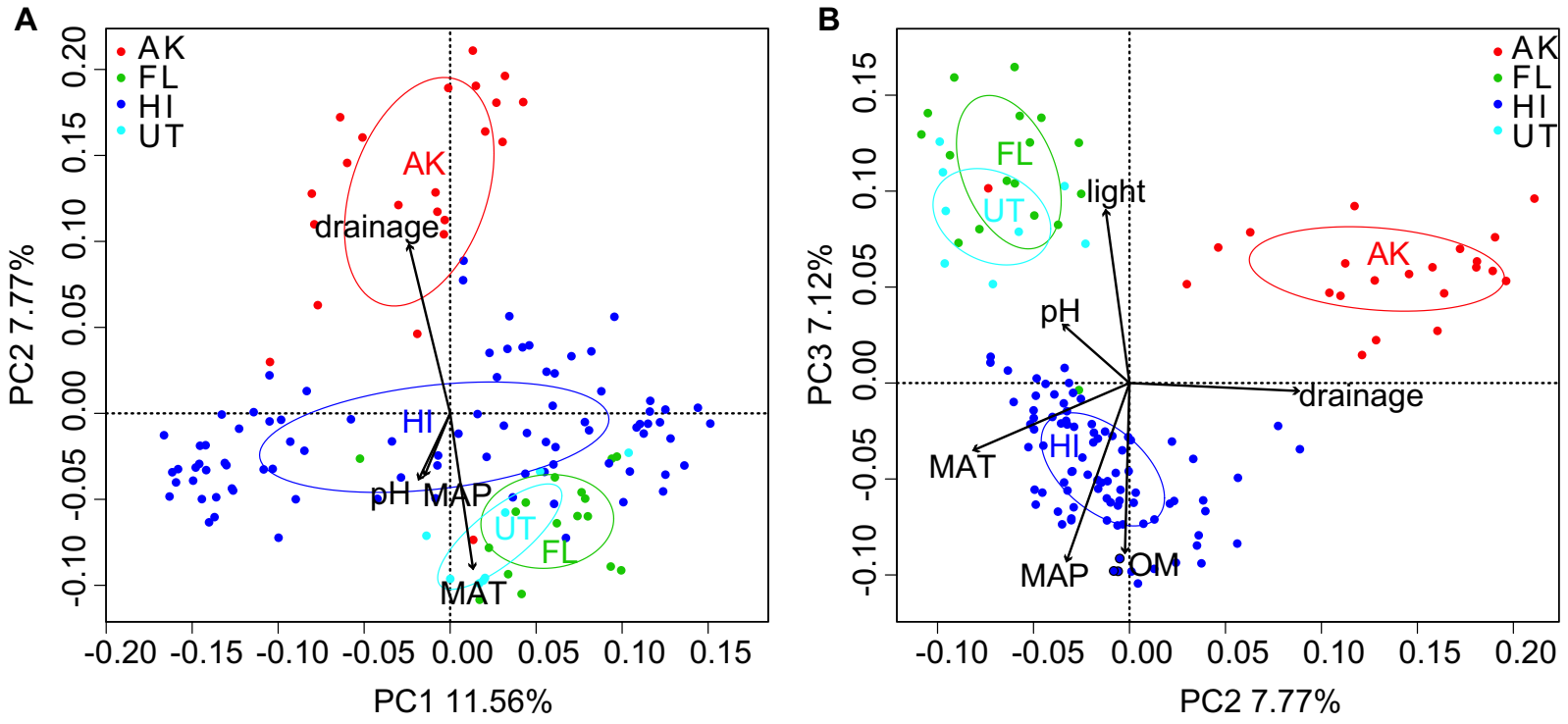


FrameBot Performance Using 454 Reads from the Mock Community



Wang, Q. et al., 2013. mBio 4:e00592-13

PCA Analysis Using FrameBot Nearest Neighbor Assignments



Wang, Q. et al., 2013. mBio 4:e00592-13

FrameBot Detected Indels in Illumina Read Pairs

- 37 samples amplifying 37 antibiotic resistance genes in swine agriculture
- 6.7 million raw MiSeq paired-end reads
- Frameshifts present in both strands of 0.75% to 11.7% of unique paired-reads (0.11% to 2.14% of the total reads)

Gene Name	% unique reads	% total reads
tetO	11.7%	2.1%
tetQ	8.5%	1.5%
tetW (set 1)	7.2%	1.4%
tetW (set 2)	3.7%	0.8%
tetB	7.6%	1.5%
tetG (set 1)	2.7%	0.7%
tetG (set 2)	0.8%	0.4%
tetL	8.3%	1.2%
tetR	3.0%	0.6%

RDP Specialized Tools for Amplicon Sequencing Data Analysis

RDP Pipelines

rdpipeline

Data Processing Steps:

- PIPELINE INITIAL PROCESS
- CLASSIFIER
- ALIGNER
- COMPLETE LINKAGE CLUSTERING

Formats for Common Programs:

- CLUSTER FILE FORMAT CONVERSION
- DISTANCE MATRIX

Analysis Tools:

- SHANNON & CHAO1 INDEX
- JACCARD & SØRENSEN INDEX
- RAREFACTION
- CHIMERA CHECK (powered by UCHIME)
- RDP LIB COMPARE
- DEFINED COMMUNITY ANALYSIS

Miscellaneous Utilities:

- ALIGNMENT MERGER
- REPRESENTATIVE SEQUENCE
- SEQUENCE SELECTION

NCBI/EBI Submission Tools:

- ENA SEQUENCE READ ARCHIVE
- FASTQ

Initial Process
barcode sorting and quality filtering raw reads, includes paired-end reads assembly

Aligner
Infernal model-based aligners for 16S and fungal 28S

Complete-Linkage Clustering
Cluster results in RDP, BIOM and other formats

Defined Community Analysis
evaluate error patterns and rates on reads amplified from a mix of known organisms

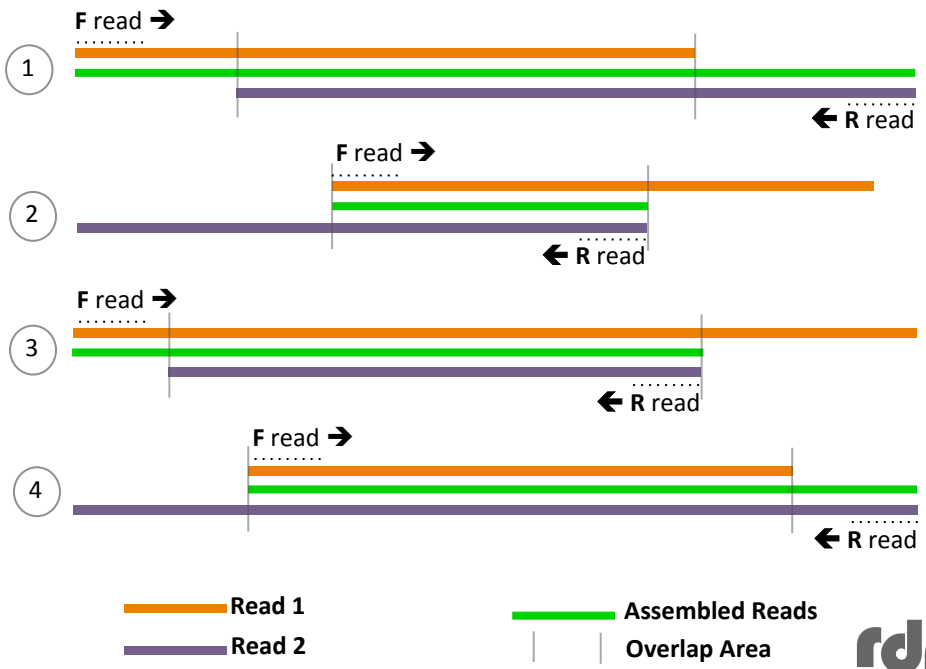
***Updated* RDP Classifier**
places sequences into bacterial and fungal taxonomies; easy to train for additional genes

Ecological Measurement
Jaccard & Sørensen Indices, Shannon & Chao1 Indices, Rarefaction

rdp

RDP Paired-end Read Assembler

- Extends PANDAseq (Masella *et al.*, 2012; Cole *et al.*, 2014), performs modified statistical analysis to find the most likely overlap, computes assembled Q scores for the read overlap region
- Built-in filtering to remove low quality assembled reads
- Handles complex overlap layouts, such as reads past the 5' end of the primer
- Can run with multiple threads: 1.4 hrs to assemble over 16 million reads from one MiSeq run using a single CPU
- Integrated into the RDPipeline Initial Process. Command line version downloadable from the RDP site



Quality-Based Overlap Detection

1) Use Q score derived probabilities p & q to estimate if two calls are from matching bases.

$$P(x=y \mid p,q) = \begin{cases} (1 - p) * (1 - q) + p * q / 3 \text{ when } \hat{x} = \hat{y} \\ (1 - p) * q / 3 + (1 - q) * p / 3 + 2 * p * q / 9 \text{ when } \hat{x} \neq \hat{y} \end{cases}$$

2) Chose overlap most likely to be in-phase (compare to null model).

$$\text{Bits saved} = \log_2 \prod_{i=1}^{n} \frac{P_i(X = Y)}{P(null)}$$

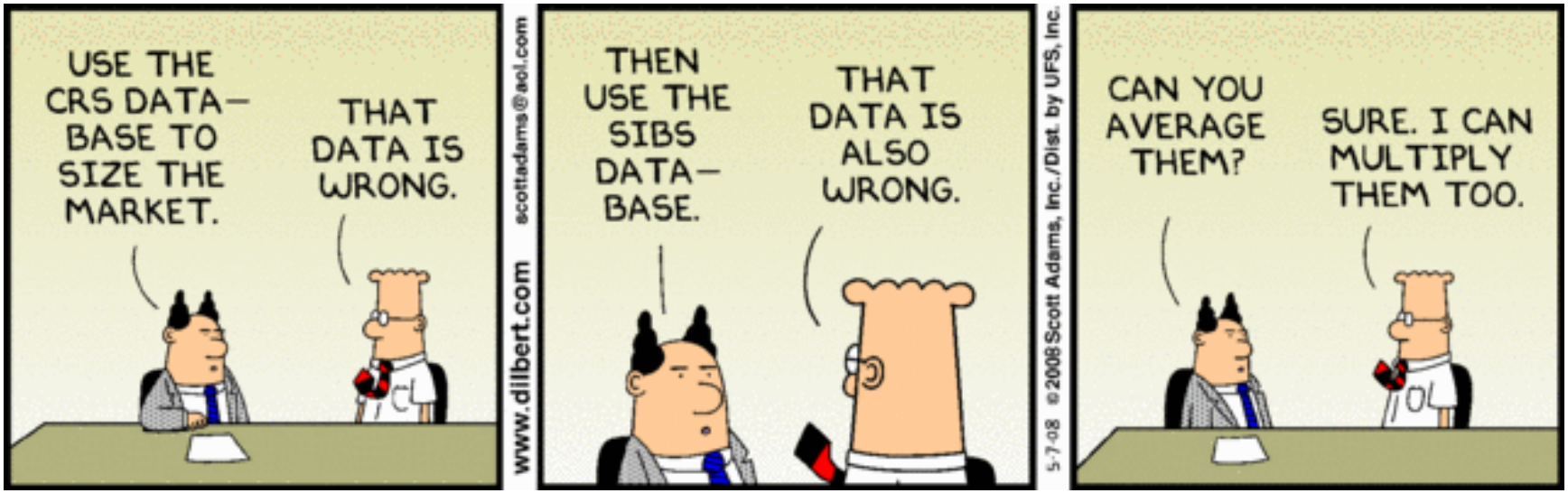
3) Derive new base Q scores for merged reads.

$$P = (\min(p,q) - p*q/3) / (p + q - 4/3 * p*q); \; Q = -10*\log_2 P$$

4) Calculate overall "read Q score" for merged read.

$$Q_r = -10 * \log_2 (1/l * \Sigma P)$$

What if my run is bad, can Informatics help?



Error Rates and Patterns

Platform	Primary Errors	Single-pass Error Rate	Final Error Rate
ABS 3730 (Sanger)	Substitution	0.1-1	0.1-1
454	Indel	1	1
Illumina	Substitution	~0.1	~0.1
Ion Torrent	Indel	1	1
SOLiD	A-T bias	~5	≤ 0.1
Oxford Nanopore	Deletions	≥ 4	4
PacBio RS	CG deletions	~15	≤ 15

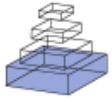
Based on company's sources

Defined Community Analysis

- Use RDP's Defined Community Analysis tool (Fish *et al.*, 2013)
- Computes a global alignment between the read and each reference sequence with the Needleman–Wunsch algorithm
- Returns the pairwise alignment with the closest match
- Calculates errors: substitution, insertion, deletion
- Error frequency is presented per-base, per-read and per-reference. These data can be input into Excel to easily produce graphical representations of the effects of different filtering strategies (as shown in the following slides)

frontiers in MICROBIOLOGY

METHODS ARTICLE
published: 01 October 2013
doi: 10.3389/fmicb.2013.00291



FunGene: the functional gene pipeline and repository

Jordan A. Fish, Benli Chai, Qiong Wang, Yanni Sun, C. Titus Brown, James M. Tiedje, and James R. Cole

MiSeq Paired-end Reads Sequencing

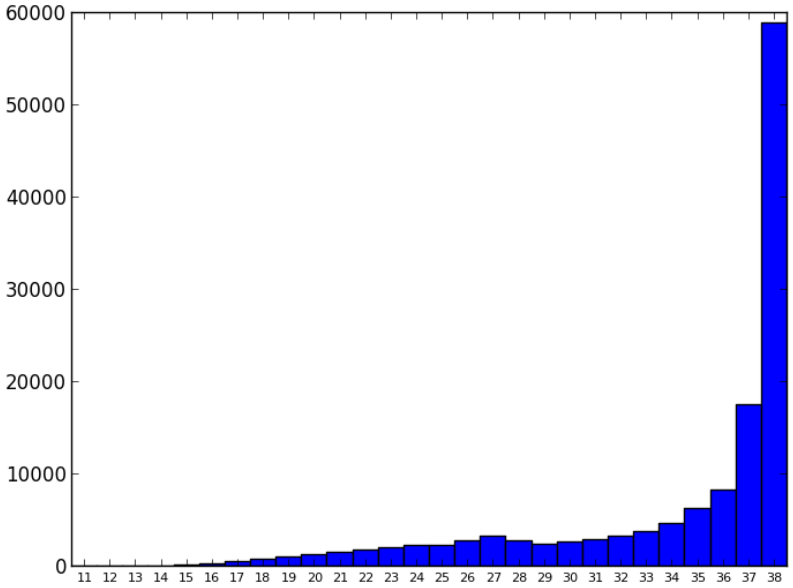
- Defined community containing 21 bacterial strains with known 16S sequence (HMP mock)
- Pat Schloss's protocol for MiSeq 16S rRNA, (Kozich *et al.*, *Appl Environ Microbiol*, 2013)
- Two MiSeq runs, amplifications.
- Expected amplicon length: 253 bp
- Paired-end read length: 250 bp
- Reads do not contain primers

Analysis Steps

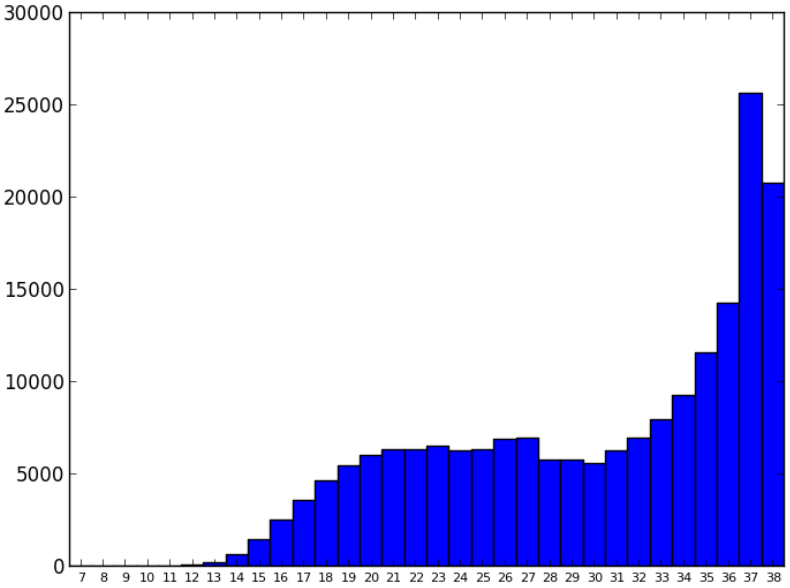
- Assemble (RDP Assembler)
- Chimera Check (UCHIME)
- Contamination Check (RDP SeqMatch)
- Error Analysis (RDP Defined Community Analysis Tool)

Read Q Score Distribution

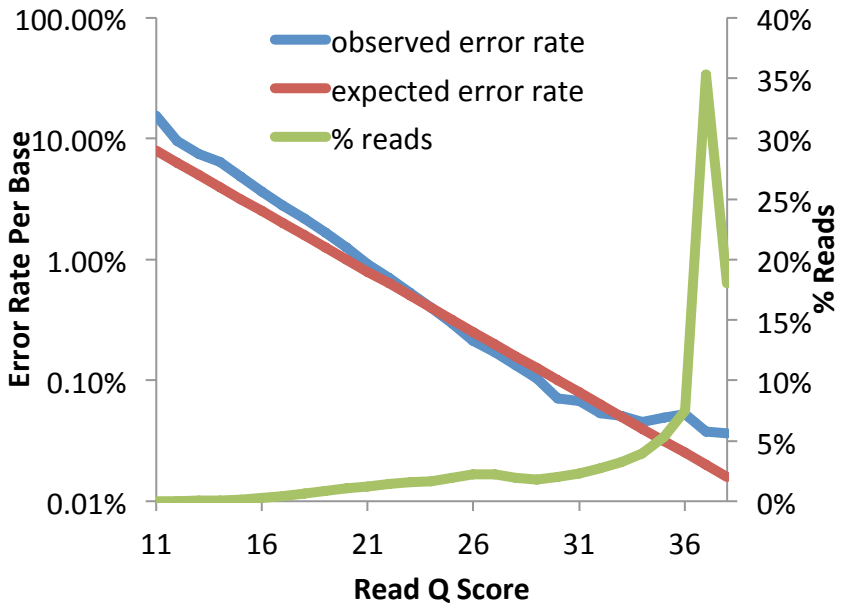
Mock A

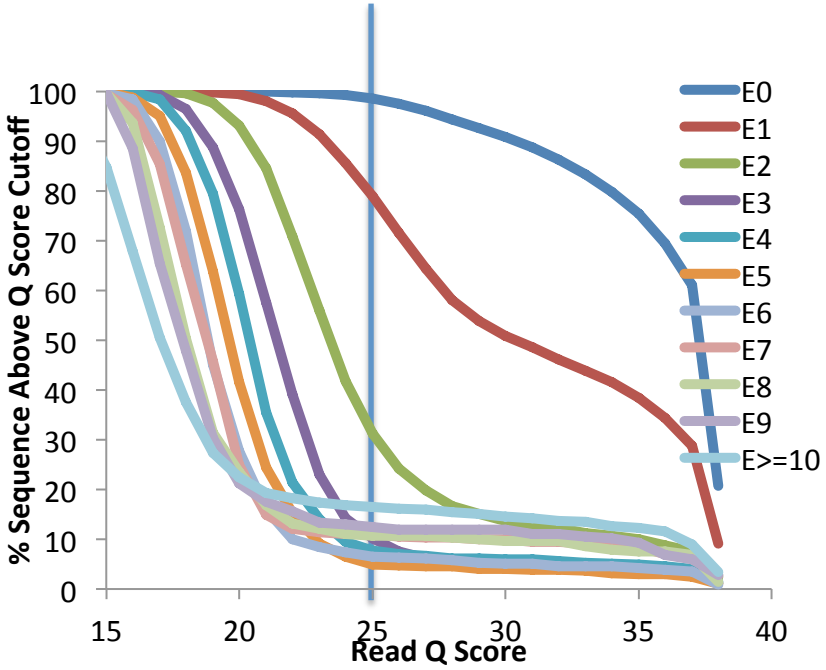


Mock 0819



Explore Read Q Score Filter (Mock A)

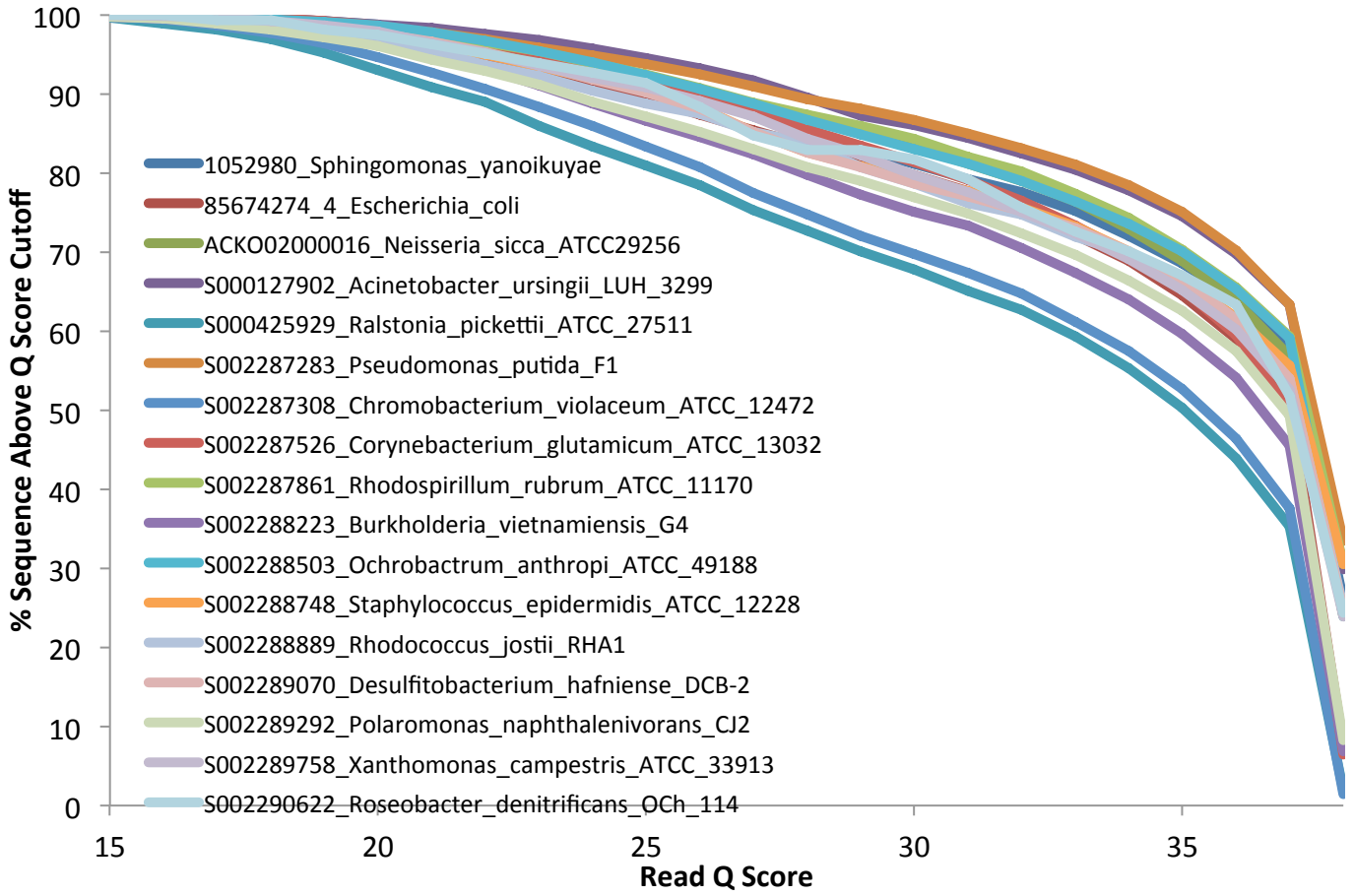




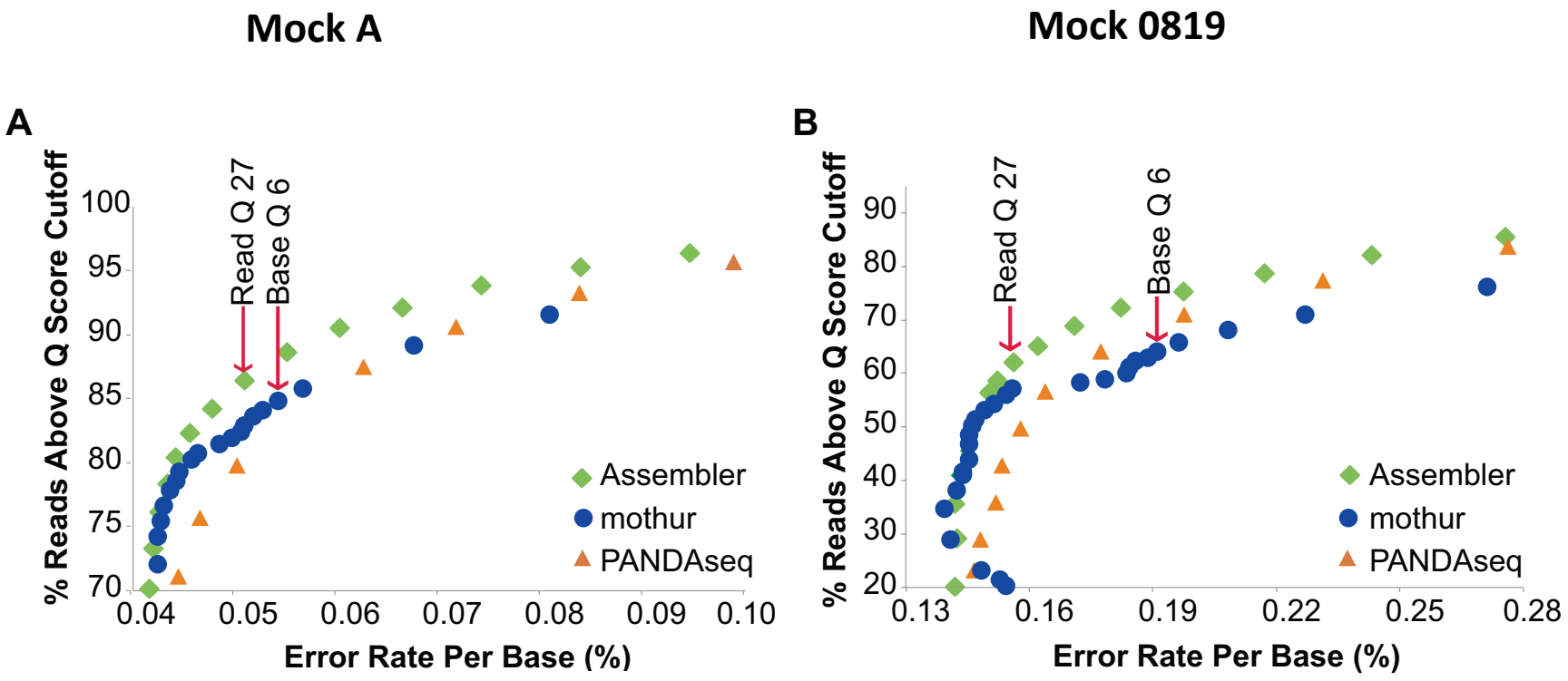
Observed error rate by read Q score is very close to expected error rate.

Applying read Q score of 25 can effectively remove reads with high errors. The number after E in the legend stands for number of errors found in a read. E0 means no errors.

Taxonomic Bias by Read Q Score Filter



Assembler Performance



Cole, J. R. et. al., 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis Nucl. Acids Res. 41(Database issue):D633-D642.

RDP Standards Support

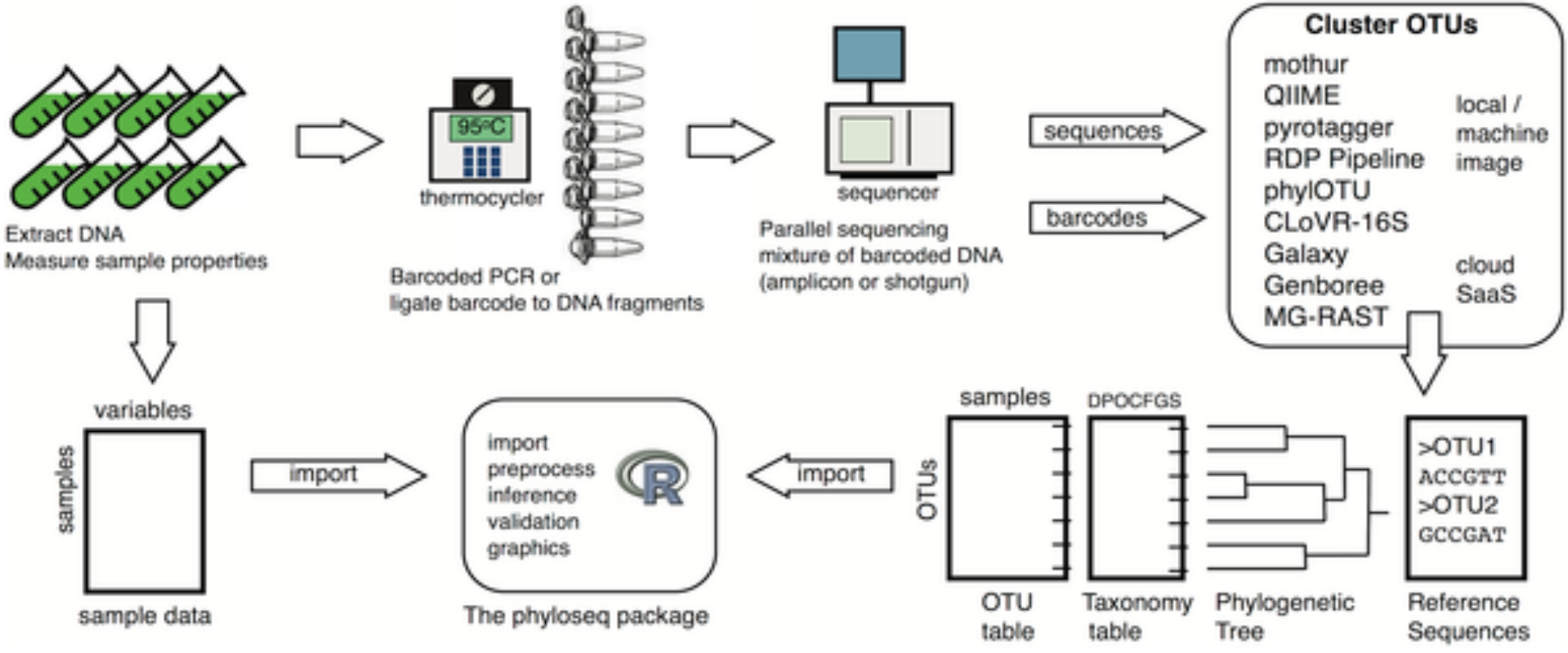
The mission of the GSC is to work with the wider community towards:

- the implementation of new genomic standards
- methods of capturing and exchanging metadata
- harmonization of metadata collection and analysis efforts across the wider genomics community

RDP Supports BIOM / PhyloSeq

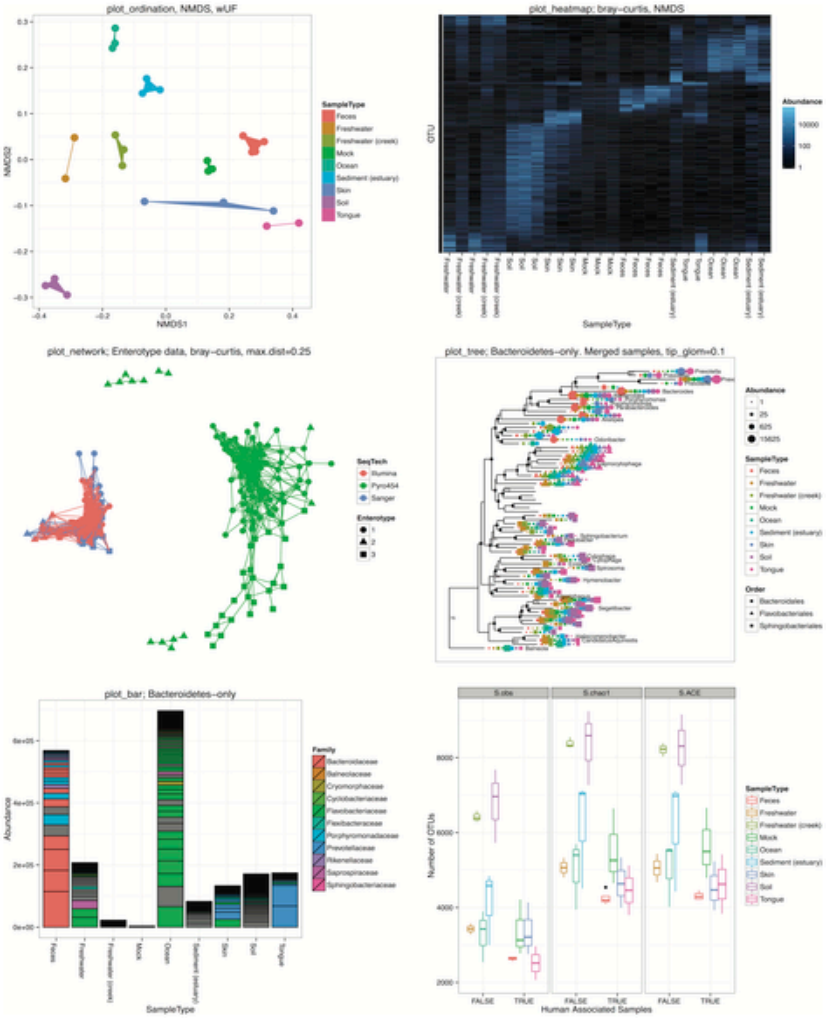
- The RDP Clustering tools produce minimal dense BIOM files as part of the results
- RDP Classifier (RDPipeline site and command-line) can now take a BIOM file as input with an optional metadata file, and produces a rich dense BIOM file
- The resulting BIOM file can be used by third-party tools such as phyloseq
- Step-by-step tutorials for PhyloSeq analysis of RDP processed data

RDP: Data and Tools for Microbial Community Analysis



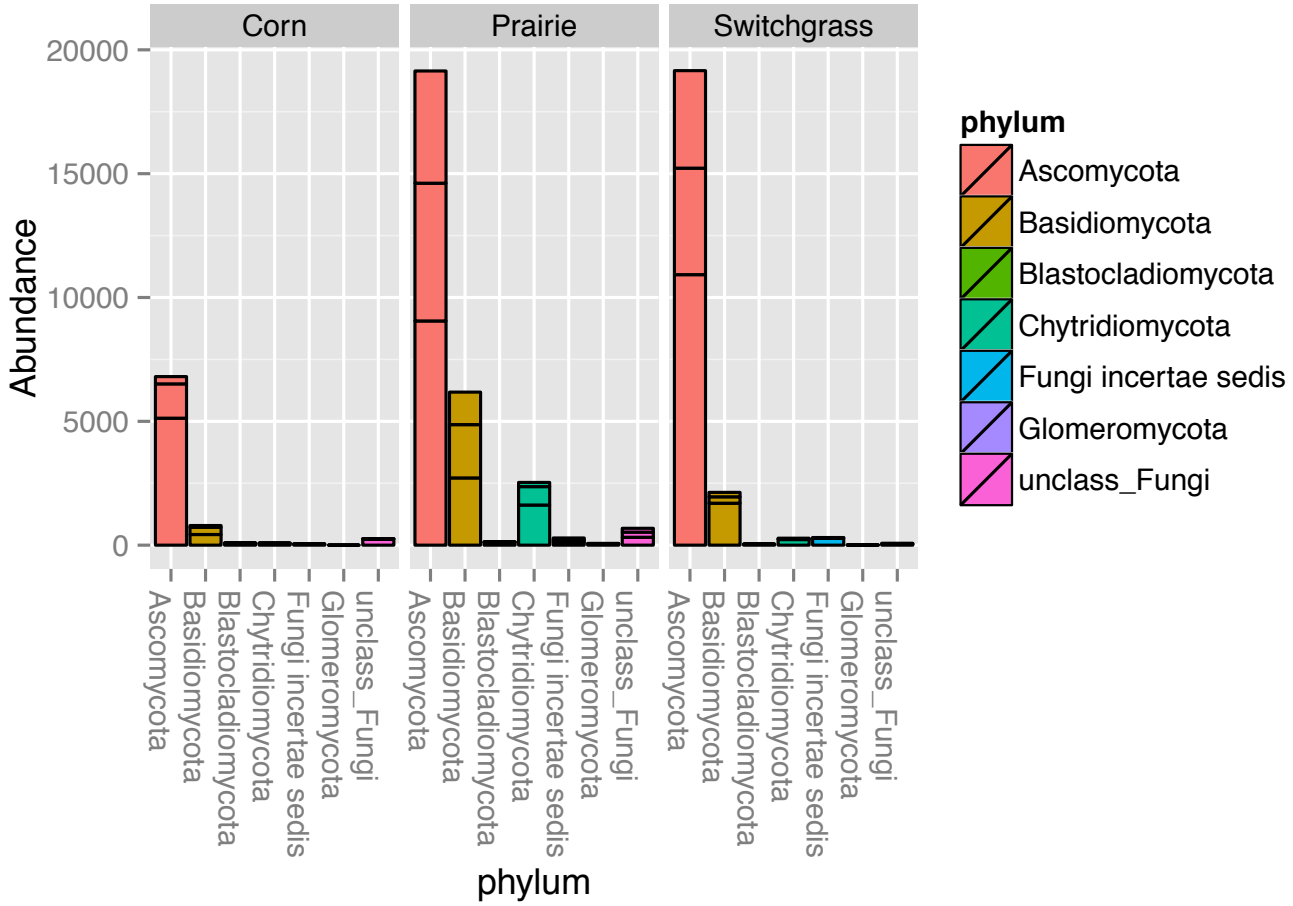
McMurdie PJ, Holmes S (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE 8(4): e61217. doi:10.1371/journal.pone.0061217
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0061217

RDP: Data and Tools for Microbial Community Analysis

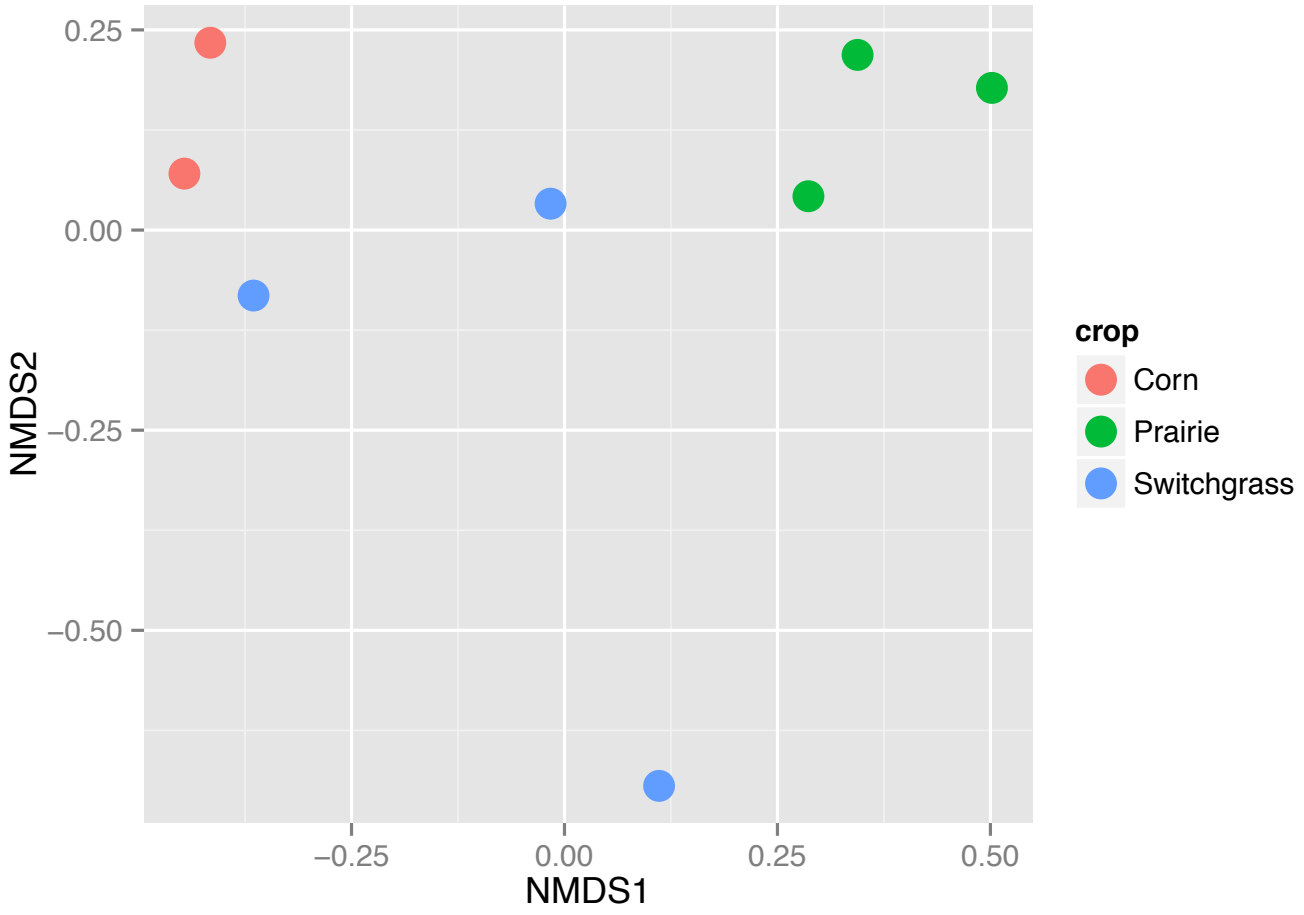


McMurdie PJ, Holmes S (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE 8(4): e61217. doi:10.1371/journal.pone.0061217
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0061217

Fungal LSU Classifier Results via PhyloSeq



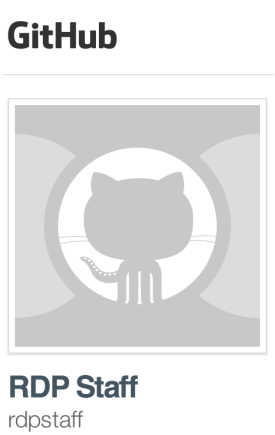
Fungal LSU Classifier Results via PhyloSeq



Integrate RDP Tools to Your Local Workflow

In addition to web-based services, RDP now distributes many of its process/ analysis tools as stand-alone, open-source versions from the RDPStaff GitHub page https://github.com/rdpstaff/

- Readily integrated as modules in customized workflow for individual tasks
- Many more processing/analysis options
- Supported with detailed instructions and tutorials
- Help desk and hotline available to help
- Easy to set up (requires Java)
  - $ git submodule init
  - $ git submodule update
  - $ make



Acknowledgements

http://rdp.cme.msu.edu

MSU

Jim Tiedje
Yanni Sun
C. Titus Brown
John Quensen

LANL

Cheryl R. Kuske
Gary Xie
Kuan-Liang Liu
Patrick Chain

WIU

Andrea Porras-Alfaro
Stephanie Eichorst

CME Bioinformatics Group:

Benli Chai
Jordan Fish
Donna McGarrell
Qiong Wang

CSIRO

Paul Greenfield
David Midgley
Nai Tran-Dinh

Sydney

Michael Charleston
Vinita Deshpande

Support from DOE, NSF, USDA, NIEHS