RDP: Data and Tools for Microbial Community Analysis

Benli Chai[1]* (chaibenl@msu.edu), Santosh Gunturu,[1] Leo Tift,[1] Yanni Sun,[1] C. Titus Brown,[2] James Tiedje,[1] **James Cole**[1]

[1]Michigan State University, East Lansing, Michigan 48824; [2]University of California-Davis, California 95616

Project Goals:
RDP offers aligned and annotated rRNA and important ecofunctional gene sequences with related analysis services to the research community. These services help researchers with the discovery and characterization of microbes important to bioenergy production, biogeochemical cycles, climate change, greenhouse gas production, and environmental bioremediation.

http://rdp.cme.msu.edu
http://fungene.cme.msu.edu
https://github.com/rdpstaff

RDP's data collections include 3,356,809 16S rRNA and 125,525 fungal 28S rRNA sequences as of December 2016. Over the past year, RDP websites (Cole et al., 2014) were visited, on average, by 8412 researchers (unique IPs) in 16,639 analysis sessions each month.

During 2016, we updated the RDP Classifier (Wang et al., 2007) and the underlying RDP Taxonomy two times to reflect recently discovered bacterial, archaeal, and fungal lineages and latest taxonomic emendations. The RDP Taxonomy now models over 2500 bacteria and archaea genera (including about 100 unofficial genera), with over 13,000 training sequences. RDP Classifier has updated its "Warcup" training set (Deshpande et al., 2016) with the latest, much improved version for rapid fungal classification using Internal Transcribed Spacer (ITS) sequences. Besides the "Warcup set", RDP Classifier also offers UNITE set trained fungal classification. Most RDP tools are now available as open source command-line versions through RDP's GitHub repository (https://github.com/rdpstaff). This includes our recently published Xander software package (Wang et al., 2015). Xander incorporates our novel method for assembling protein-coding sequences for genes of interest from a metagenomic dataset. In addition to the software packages, the repository includes additional resources including examples, documentation and tutorials. These command- line tools provide researchers with an independent method to analyze their own data, including high-throughput data and many of these tools are already used in third-party pipelines. These stand-alone versions of our tools have been created for easy porting to KBase in the future.

RDP's FunGene Pipeline & Repository (Fish et al., 2013) provides databases for 270 protein coding genes useful as phylogenetic markers and for following important ecological functions. In addition to the aligned and annotated gene and protein sequences, FunGene provides online analysis functions and tools for selecting subsets of sequences for download and further analysis. Use of the FunGene web, on average, was 778 researchers per month in 1402

analysis sessions. During the past year, we updated FunGene data releases five times from searches of the primary sequence databases.

We have continued improving the website for better performance in accessing reference sequences and analyzing amplicon data. In addition to optimizing existing gene models in N and C cycles, we have added more genes of environmental importance, such as an additional N cycle gene model for improved homolog detection and characterization, five new models for major classes of bacterial phytase genes (HAPs, PTPs, PAPs, and BPPs) and an alkaline phosphatase gene (ALP), which play an important role in general phosphorus cycle in different environments in catalyzing the cleavage of phosphate groups from the indigestible organic forms and make it bioavailable.

References:

Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42 (Database issue): D633-D642. doi: 10.1093/nar/gkt1244

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. doi: 10.3389/fmicb.2013.00291

Wang, Q., J. A. Fish, M. Gilman, Y. Sun, C. T. Brown, J. M. Tiedje, and J. R. Cole. (2015). Xander: employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome* 3: 32 doi: 10.1186/s40168-015-0093-6

Wang, Q, G. M. Garrity, J. M. Tiedje, and J. R. Cole. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl Environ Microbiol.* 73: 5261-5267. doi: 10.1128/AEM.00062-07

Deshpande, V., Q. Wang, P. Greenfield, M. Charleston, A. Porras-Alfaro, C. R. Kuske, J. R. Cole, D. J. Midgley, and N. Tran-Dinh. (2016). Fungal identification using a Bayesian Classifier and the 'Warcup' training set of Internal Transcribed Spacer sequences. *Mycologia* 108(1):1-5. doi: 10.3852/14-293

This research was supported by the Office of Science (BER), U.S. Department of Energy Grant No. DE-FG02-99ER62848, with contributions from Office of Science (BER), U.S. Department of Energy Grant Nos. DE-SC0014108, DE-SC0010715, DE-FC02- 07ER64494, NIEHS Superfund Research Program Award 5P42ES004911-23 and National Science Foundation Award No. DBI-1356380.