RDP: Data and Tools for Microbial Community Analysis

Benli Chai[1]* (chaibenl@msu.edu), Yanni Sun,[1] C. Titus Brown,[2] James M. Tiedje,[1] and **James R. Cole**[1]

[1]Center for Microbial Ecology, Michigan State University, East Lansing. MI; [2]University of California-Davis, Davis, CA

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

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.

RDP's data collections include 3,224,600 16S rRNA and 108,901 fungal 28S rRNA sequences as of February 2016. Over the past year, RDP websites (Cole et al., 2014) were visited, on average, by 7805 researchers (unique IPs) in 15,464 analysis sessions each month.

During 2015, we updated the the RDP Classifier (Wang et al., 2007) and RDP Taxonomy three times to reflect recently discovered bacterial, archaeal, and fungal lineages and latest taxonomic emendations. The RDP Taxonomy now models over 2500 genera (including about 100 unofficial genera), an increase of over 800 genera, with over 13,000 training sequences. 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 264 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 1069 researchers per month in 1753 analysis sessions. During the past year, we updated FunGene data releases nine times from searches of the primary sequence databases.

We have made several performance improvements to the website to make access to sequence data much faster, especially for genes with a large number of data. In addition to optimizing existing gene models in N and C cycles, we have added more genes of environmental importance, such as the *acdS* gene which promotes microbe-mediated plant growth and tolerance to drought and salinity stresses, a group of key genes in metabolic responses to environmental toxicants such as polychlorinated dibenzo-*p*-dioxins (PCDDs), as well as antibiotic resistance gene data using Resfams reference sequences. In addition, we are leveraging FunGene to provide training data for our Xander gene-targeted assembler and for high-throughput qPCR gene-targeted primer and probe development.

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

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.