169. RDP: Data and Tools for Studying Structure and Function of Microbial Communities

Benli Chai[1*] (chaibenl@msu.edu), Jordan A. Fish[1], Qiong Wang[1], Yanni Sun[3], C. Titus Brown[2,3], James M. Tiedje[1,2] and **James R. Cole**[1]

[1]Center for Microbial Ecology, [2]Microbiology and Molecular Genetics, [3]Computer Science and Engineering, Michigan State University, East Lansing, MI 48824

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

Project Goals: The Ribosomal Database Project (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, greenhouse gas production, and environmental bioremediation.

In the current release (October 2013), RDP offers 2,809,406 aligned and annotated quality- controlled public bacterial, archaeal, and fungal rRNA sequences. Over the past year, the RDP website was visited, on average, by **10,000 users** (unique IP) in **22,000 analysis sessions** each month. RDP recently released new alignments of bacterial and archaeal 16S rRNA gene sequence alignments and a fungal 28S gene sequence alignment using the latest Infernal 1.1 aligner with specially-tuned covariance models (CMs). As part of RDP's efforts to support the fungal research community following the release of the Fungal 28S RDP Classifier (Liu et al., 2011), most RDP tools, including the RDP Hierarchy Browser, Sequence Match, Probe Match, and RDPipeline, have been updated to work with the new fungal 28S sequences.

The **new RDPipeline** (Cole et al., 2013) expands upon our existing high-throughput tool offerings and is designed to accommodate the latest benchtop high-throughput sequencing technologies. RDPipeline integrates with researchers' existing *my*RDP accounts for streamlined analysis job submission and monitoring. The new RDPipeline includes both improved performance in optimizing back-end job load distribution and increased capacity for larger datasets. It also provides additional user-friendly features such as a "my jobs" page for each user to track the job status, download results, and retrieve process parameters for past analysis tasks submitted to RDPipeline. Other enhancements include optimized paired-end read assembly (Assembler). Tested on Illumina MiSeq paired-end data, this tool outperformed its peers in selectively filtering out error-containing sequence reads, and also better handles different types of paired-end overlaps. A new data validation mechanism implemented in RDPipeline provides feedback if incorrect data input is submitted before an analysis job starts running--a feature especially valuable for inexperienced users.

FunGene, RDP's Functional Gene Pipeline and Repository (Fish et al., 2013), offers databases of many common ecofunctional genes and proteins, as well as integrated tools that 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. Additional FunGene tools are specialized to process coding gene amplicon data. For example, **RDP FrameBot** (Wang et al., 2013) produces frameshift-corrected protein and DNA sequences from raw reads while finding the most closely related protein reference sequence. These tools can help provide better insight into microbial communities by directly

studying key genes involved in important ecological processes. Over the past year, RDP FunGene **usage increased 1.8-fold** to 1997 researchers in 2823 analysis sessions per month.

Porting RDP tools to KBase will provide an opportunity for RDP to reach the broader research community. To learn how to develop RDP tools into KBase services (modules). RDP staff hosted a two-day **KBase Bootcamp at MSU** in May 2013. Two instructors from Argonne National Laboratory, four RDP staff members and one additional student from MSU, three graduate students and one post-doc from the University of Oklahoma, and one graduate student from Georgia Institute of Technology all participated in the KBase Bootcamp. Following the bootcamp all nine participants completed a survey. Overall, they all agreed or agreed strongly with the statement "I found the KBase bootcamp useful". Five of the participants said they would apply for KBase developer accounts, while the other four indicated they would be developing tools that will use KBase services.

In addition to web-based services, RDP now distributes many of its process/analysis tools as stand-alone, open-source versions through https://github.com/rdpstaff. Tutorials are provided to guide researchers through the otherwise complex data processing steps in well-defined, task- oriented workflows with detailed instructions. RDP's mission includes user support; email rdpstaff@msu.edu or call +1(517)432-4997.

References:

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

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

Liu K.L., Porras-Alfaro A., Kuske C.R., Eichorst S.A., and Xie G. Accurate, rapid taxonomic classification of fungal large-subunit rRNA genes. *Appl. Environ. Microbiol.* 78:1523- 1533 (2012).

Wang Q., Quensen III J.F., Fish J.A., Lee T.-K., Sun Y., Tiedje J.M., and Cole J.R. Ecological patterns of nifH genes in four terrestrial climatic zones explored with targeted metagenomics using FrameBot, a new informatics tool. *mBio* 4:e00592-13 (2013).

Ribosomal Database Project (RDP) is supported by the Office of Science (Biological and Environmental Research), U.S. Department of Energy grant DE-FG02-99ER62848.