

Analysis of metagenomics data

Elizabeth M. Glass^{1,2} and Folker Meyer^{*,1,2,3}

¹Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439. ²Computation Institute, The University of Chicago, 5735 South Ellis Avenue, Chicago, IL 60637. ³Institute for Genomics and Systems Biology, The University of Chicago, 900 East 57th Street, Chicago, IL 60637.

* Corresponding author: folker@mcs.anl.gov

Key words: bioinformatics, metagenome, comparative analysis, high-throughput computing, metadata

Abstract

Improved sampling of diverse environments and advances in the development and application of next-generation sequencing technologies is accelerating the rate at which new metagenomes are produced. Over the past few years, the major challenge associated with metagenomics has shifted from generating to analyzing sequences. Metagenomic analysis includes the identification, and functional and evolutionary analysis of the genomic sequences of a community of organisms. There are many challenges involved in the analysis of these data sets including sparse metadata, a high volume of sequence data, genomic heterogeneity and incomplete sequences. Due to the nature of metagenomic data, analysis is very complex and requires new approaches and significant compute resources. Recently, several computational systems and tools have been developed and applied to analyze their functional and phylogenetic composition.

The metagenomics RAST server (MG-RAST) is a high-throughput system that has been built to provide high-performance computing to researchers interested in analyzing metagenomic data. Automated functional assignments of sequences in the metagenome are generated by comparing both protein and nucleotide databases. Phylogenetic and functional summaries of the metagenomes are constructed, and statistical tools for comparative metagenomics are provided. MG-RAST provides a collaborative environment that allows for user privacy and management. In MG-RAST, all users retain full control of their data, and everything is available for download in a variety of formats. This service has removed one of the primary bottlenecks in metagenome sequence analysis, the availability of high-performance computing for annotating data.

1. Introduction

Studying uncultivable microorganisms has been a major obstacle to understanding natural microbial populations within the context of their environment. Metagenomics is expanding quickly as next-generation sequencing approaches become more widespread and applied to an increasing number of environments. It has bypassed the need for cloning and has enabled a new approach to comparative metagenomics (Ronaghi et al., 1996; Ronaghi et al., 1998; Margulies et al., 2005). Now sequence abundance can be used to contextualize datasets for driving pattern recognition and uncovering unique properties within natural microbial communities.

Regardless of the sequencing approach used to generate data, the first steps in analysis of any metagenome involve comparative analysis against various ribosomal and protein and nucleotide databases. These comparisons have a large computational cost but provide the basic data types for many subsequent analyses, including phylogenetic comparisons, functional annotations, binning of sequences, phylogenomic profiling, and metabolic reconstructions and modeling. Analysis of single metagenomes can provide a greater understanding of a microbial community, but the comparison of multiple metagenomes provides greater insight.

Sequence data, however, must be accompanied by enough contextual information (metadata), such as sample characteristics, to make individual investigations reproducible and enable valid interpretation (Field et al., 2009). Community-driven minimum information checklists (Taylor et al, 2007), common ontologies (Smith et al, 2007) and formats (Jones et al, 2007; Sansone et al, 2008) have major roles to play. Therefore, data

describing such information as a sample's environment, sample origin, isolation, and treatment are an important resource to link to sequence data in order to enable meaningful comparative analysis. The Genomics Standards Consortium (GSC) has defined the Minimum Information About a (Meta)Genome Sequence (MIGS/MIMS) (Kottmann et al., 2008), which describes core descriptors of environmental context (habitat). MIGS/MIMS extends the minimum information provided by the International Nucleotide Sequence Database Collaboration (INSDC) (Cochrane et al., 2010).

Recently, several computational systems and tools have been developed and applied to analyze their functional and phylogenetic composition. One such system, MG-RAST (Meyer et al., 2008), is available over the web to researchers, and access is not limited to specific groups or data types. This system has a scalable compute backend that has allowed to the analysis of over 10,000 metagenomes (as of January 2011) thus far.

2. Metagenomic Analysis

Metagenomic analysis is not straightforward. The data is much more complex than what has previously been seen in genomics. Metagenomic sequence data has lower sequence redundancy, lower sequence quality, short read lengths, increased polymorphisms and relative abundance (simple vs. complex communities). In addition to these inherent issues and the evolution of sequencing technologies and chemistries, the size of the data is changing. The scientific community has already seen the size of these data sets quickly move from Megabase pairs (Mbps) to Gigabasepairs (Gbps) and now Terabases, which require significant compute resources.

Given sufficient compute resources, there are several different approaches that can be taken with raw sequence reads. The analysis “path” and the tools you choose can

influence your results. There is no “one size fits all” tool or best practice established for analysis metagenomic data sets. Various approaches have strengths and weaknesses and are constantly evolving.

However, the major metagenomic analysis pipelines such as MG-RAST, IMG/M (Markowitz et al., 2008), and CAMERA (Sun et al., 2011), provide compelling analysis strategies and features as well as distinct implementations of common operations. The MG-RAST server is the most widely used tool for the analysis of shotgun metagenomics and provides a basis for sequence analysis of large, complex data sets. Over 4,000 users have submitted data sets and several hundred users work on the system each day.

The MG-RAST system accepts shotgun metagenomic DNA sequence data in different formats and from a variety of platforms, providing initial quality control and normalization of the data. The pipeline also accepts assembled sequences in fasta format. Sequence data may be compressed by one of several common computer programs to speed upload. Users may choose to upload raw unassembled reads or assembled contigs. The system also provides a GSC compliant metadata editor to enter relevant information about a sample. This information is then incorporated into the analysis and querying capabilities. The server provides several methods to access different data types, including phylogenetic and metabolic reconstructions, and the ability to compare the metabolism and annotations of one or more metagenomes and genomes. In addition, the server offers browsing of data and a comprehensive search capability. Access to the data is password protected, and all data generated by the automated pipeline is available for download and analysis in variety of common formats. One of the more widely used features is the ability to share data prior to publication, leading to networks of shared data sets.

2.1 MetaData

It is apparent that the full potential of comparative metagenome analysis can be achieved only in the context of the metadata (information describing the sample). The selection of samples based on rich metadata is crucial for understanding large-scale patterns when multiple metagenomes are compared. The GSC has proposed a minimal set of data, called the Minimum Information about a (Meta)Genome Sequence (MIGS/MIMS) that should be collected with every metagenome sequence. Although this is an evolving standard, the MG-RAST server is MIGS/MIMS-compliant. Metadata is requested from the user at the time of sequence submission to MG-RAST. Metadata can be added to at any point after submission and a minimal set is required for sharing or publishing (making public). This data is stored with the user's data and is made available to them.

2.2 Preprocessing

Preprocessing of sequence reads before analysis (assembly, gene prediction, and annotation) is an overlooked aspect of metagenomic analysis. Preprocessing includes steps in filtering data for vectors, host contaminants and quality trimming. Mistakes in any of these steps can have significant downstream affect on analyses (Koonin et al., 2007). MG-RAST employs a normalization step, generating unique internal IDs, and removing duplicate sequences. Users can select filtering for contaminants. It also includes a runtime-efficient method for obtaining a quality estimate for each sample and removal of sequencing artifacts.

2.4 Identifying Genes

Sequence length is an important factor in determining an approach to gene calling. Shorter reads lengths pose an obvious and significant challenge. The most commonly used method for identifying genes in metagenomic reads are via similarity searches using metagenomic sequences against databases of known proteins. BLAST (Altschul et al., 1997) has become too costly in terms of computation. Faster alternatives like BLAT (Kent, 2002), doing assembly and feature prediction greatly reduce the computational burden of comparing all pairs of short reads. MG-RAST relies on BLAT to perform sequence similarity searches as it provides significant speed-ups over BLAST, offers very similar results and loses little sensitivity in our tests.

MG-RAST screens for potential protein encoding genes (PEGs) via a BLAT search against the MG-RAST nonredundant database. This strategy will reveal already known genes that are present in the metagenome. A drawback to using this approach as a sole means to identify genes is that many genes are most probably not present in the databases due to the bias towards culturable organisms. Therefore MG-RAST performs feature prediction using FragGeneScan (Rho et al., 2010), before running similarity searches. FragGeneScan predicts coding regions in sequences that are greater than or equal to 80bp.

In parallel with feature prediction and BLAT similarity searches against the protein database, the sequence data is also compared to other databases by using the appropriate algorithms and significance selection criteria. These databases include several ribosomal databases, including GREENGENES (DeSantis et al., 2006), RDP-II (Cole et al., 2007), and Silva (Pruesse et al., 2007). The search criteria are specific for each database. For example, using Sblat against the rDNA databases enables users to screen for ribosomal

RNA genes, but much more stringent selection criteria are used to identify candidate RNA genes than for identifying protein-encoding genes (by default, the similarity must exceed 50 bp in length and have an expect value less than 1×10^{-5}). Lastly, these matches to the MG-RAST database and ribosomal databases are used to compute the derived data. A phylogenomic reconstruction of the sample is computed by using both the phylogenetic information contained in the non-redundant database and the similarities to the ribosomal RNA databases. Functional classifications of the PEGs are computed by projecting against protein functional annotations based on these similarity searches. These annotations become the raw input to an automatically generated initial metabolic reconstruction of the sample, as well as subsequent metabolic model for the sample by providing suggestions for metabolic fluxes and flows, reactions, and enzymes.

While the existing version relied on sequence comparison to the non-redundant database provided by the SEED (Overbeek et al., 2005) and SEED subsystems solely, the new version is based on a database emanating from the Genomics Standards Consortiums M5 platform. This non-redundant database provides a non-redundant integration of many databases (e.g INSD, SEED, IMG, KEGG (Kanehisa et al., 2004), EGGNOGs (Jensen et al., 2008)), thus allowing supporting multiple different views on the data with one similarity search.

2.5 Multiple supported classification schemes.

A number of competing naming schemes to described functional classification of genes and proteins exist. While the use of consistent SEED subsystem based annotations provides many advantages other databases provide different functional hierarchies (e.g. SEED subsystems, IMG, COG/NOGs) or ontologies (GO (Barrell et al., 2009)). Enabled

by this protein database, we provide the ability to “on-the-fly” switch between different annotation resources. Allowing users to view their data through mapping to different classification schemes allowing them to tease out differences or similarities between metagenomic data sets not visible otherwise.

The user interface for MG-RAST was designed to provide easy navigation and use of comparative tools. There are multiple views for browsing and analysis of the data, as well as a means to download all result tables and the sequences for every subset displayed.

Users are also enabled to modify the displayed results by modifying search parameters used to compute the functional, metabolic, and phylogenetic reconstruction. This allows more stringent match criteria (e.g., expectation value, score, overall percent identity, length of match, and number of mismatches); and, by restricting the matches, the derived data is dynamically changed. The default parameters have been chosen by empirical testing and represent a tradeoff between accuracy and specificity.

2.6 Annotations

Users can view and search their annotated metagenome based on annotation source (see description of MG-RAST database) through various avenues. Metagenome Overview provides a summary of the sequence and annotation statistics against the various databases. More details are presented in the Sequence Profiles, which display the metabolic and phylogenetic distributions in a given sample. Views are in the form of charts and tables and data is downloadable for each profile. Like all analyses in MG-RAST, the user can modify inclusion parameters and export results. Each metabolic or phylogenetic/phylogenomic profile can also be viewed singularly or compared with other metagenomes using a circular tree comparison tool.

2.7 Comparative Metagenomics

Considering that comparative analysis is the core driver for discovery-based biology, MG-RAST enables more than just views of the analysis results of a given metagenome, the system supports comparative analysis. Therefore, comparative metagenomics tools are central to the utility of the MG-RAST platform. Several tools have been developed and integrated into the MG-RAST framework, allowing users to compare a metagenome to either (1) other metagenomes, (2) individual genomes, or (3) both metagenomes and genomes.

2.7.1 Comparative Heat Maps

Metabolic: PEGs identified to have functions belonging to a SEED subsystem(s) and KEGG pathways are mapped to that subsystem/pathway. When these functional roles are linked to specific genes across metagenomes and a populated subsystem emerges. The utility of this organization is extended by subsystem connections that allow linkage of genes between subsystems.

Each subsystem present in a sample is scored by counting the number of sequences that are similar to a protein in each subsystem. This score is divided by the total number of sequences from the sample that are similar to any protein in a subsystem, to give a fraction of sequences in subsystems that are in a given subsystem. This approach allows comparisons between samples that have different numbers of sequences. Since the fractions tend to be small, the scores can be factored for display purposes. Moreover, the display can be limited or expanded to include various levels in the subsystem hierarchy, to specific areas of metabolism, or other subsystem groups, as chosen by the user.

Phylogenetic: The taxonomic heat map works in an analogous fashion but highlights the different taxonomic profiles in each sample, as determined by the phylogenetic or phylogenomic approaches selected by the end user (e.g., 16S comparisons, phylogenomics from BLAT results). Again, samples may be grouped in a nonquantitative fashion to rapidly highlight particular phylogenetic groups that predominate in different samples.

2.7.2 Principal component analysis

Many comparative analyses use multivariate statistics when several metagenomic datasets are involved, or when several types of factors are thought to affect the observed compositions of the communities. MG-RAST has incorporated a R-based PCA (Principle Component Analysis) to its suite of comparative tools.

2.7.3 Recruitment Plot

The recruitment plot tool is set up to provide a selected sequenced microbial genome as a scaffold to map metagenome-derived sequences to. As in the heat map, sequences that have been annotated from a metagenome are used as the queries. The initial view provides a ranked list of microbial genomes that contain the most number of matched sequences from the metagenome. This gives an indication of the relative representations in terms of genomic content found within the metagenome.

2.8 Metabolic Reconstructions and Models

Metagenomics also has the potential provide insights into the critical biochemical mechanisms in each environment. Models in the MG-RAST are based on the initially assembled metabolic reconstructions. The functional roles from the reconstruction are then mapped to reactions in the SEED and KEGG biochemistry databases, and this

mapping is used to assemble a reaction list for the model. Models are based on a steady state and undergo flux balance analysis.

3. Results and Discussion

Improved sampling of diverse environments, combined with advances in the development and application of next-generation sequencing technologies, is accelerating the pace at which new metagenomes are generated. In fact, the amount of sequence data being produced will quickly outpace the ability of scientists to analyze it. Analysis of metagenomic data needs to incorporate scalable computing resources.

The process of building MG-RAST is the result of several years of planning and engineering. The system provides integration of metagenome data, microbial genomics, and manually curated annotations. The metagenomics analysis pipeline was designed to allow for interactive analysis and the system as a whole has been built by using an extensible format allowing the integration of new datasets and algorithms without a need for recomputation of existing results. The system has been restructured to be scalable. This means MG-RAST uses cloud computing, which decouples it from a particular dataset and allows vast compute resources, to conduct the analysis.

The MG-RAST server handles both assembled and unassembled data. Each approach has advantages that should be considered when comparing metagenomes. For example, a case where sequences should be assembled is when comparisons between samples is being calculated, as the assembly process loses the frequency information critical for determining differences between samples. In contrast, assembled sequences tend to be longer and therefore more likely to accurately identify gene function or phylogenetic source from binning (McHardy et al., 2007).

The analytical methods integrated into MG-RAST provide core annotations and analysis tools to compare and contrast sets of metagenomes (Edwards et al., 2006; Fierer et al., 2007; Mou et al., 2008). The approach underlying the subsystems-based functional analysis of metagenomes has been validated with 90 different samples from nine major biomes. The analysis demonstrated that the biomes could clearly be separated by their functional composition (Dinsdale et al., 2008). All of the metagenomes present in that study are included in the publicly available datasets visible on the MG-RAST server. Although the service contains core functionality for the annotation and analysis of metagenomes, many of the techniques traditionally used for genome analysis either do not work with metagenomes or show significant performance degradation (Krause et al., 2006). Therefore, new analytical methods are needed to fully understand metagenomics data. The most obvious problem is with the large number of unknown sequences in any sample. Others and we are developing new binning, clustering, and coding region prediction tools to handle these unknown sequences, and effective tools will be incorporated into the pipeline when available. Another problem is that the rapid pace with which sequence data is being generated outpaces increases in computational speed, and therefore improvements in common search algorithms are required to ensure that sequence space can be accurately and efficiently searched.

4. Internet Resources

MG-RAST (<http://metagenomics.anl.gov>)

Acknowledgements

This work was supported in part by the U.S. Department of Energy, under Contract DE-AC02-06CH11357.

References

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.
- Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R. The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.* 2009 Jan;37(Database issue):D396-403.
- Cochrane G, Karsch-Mizrachi I, Nakamura Y; International Nucleotide Sequence Database Collaboration. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* 2011 Jan;39(Database issue):D15-8.
- Cole JR, Chai B, Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, et al. 2007. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res.* 35(Database issue):D169-72.
- DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol.* 72(7):5069-72.
- Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. 2008. Functional metagenomic profiling of nine biomes. *Nature.* 452(7187):629-32.
- Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, et al. 2006. Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics.* 7:57.
- Field D, Sansone SA, Collis A, Booth T, Dukes P, Gregurick SK, et al. Megascience. 'Omics data sharing. *Science.* 2009;326(5950):234-6.
- Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, et al. 2007. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl Environ Microbiol.* 73(21):7059-7066.
- Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* 2008 Jan;36(Database issue):D250-4.
- Jones AR, Miller M, Aebersold R, Apweiler R, Ball CA, Brazma A, Degreef J, Hardy N, Hermjakob H, Hubbard SJ, Hussey P, Igra M, Jenkins H, Julian RK Jr, Laursen K, Oliver SG, Paton NW, Sansone SA, Sarkans U, Stoeckert CJ Jr, Taylor CF, Whetzel PL, White JA, Spellman P, Pizarro A. (2007). The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nat Biotechnol.* 25(10):1127-1133.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 1;32(Database issue):D277-80.

Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res.* 2002 Apr;12(4):656-64.

Koonin EV. Metagenomic sorcery and the expanding protein universe. *Nat Biotechnol.* 2007 May;25(5):540-2.

Kottmann R, Gray T, Murphy S, Kagan L, Kravitz S, et al. 2008. A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). 12(2):115-21.

Krause L, Diaz NN, Bartels D, Edwards RA, Puhler A, et al. 2006. Finding novel genes in bacterial communities isolated from the environment. *Bioinformatics.* 22(14):e281-289.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380.

Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IM, Grechkin Y, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Hugenholtz P, Kyrpides NC. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* 2008 Jan;36(Database issue):D534-8.

McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. 2007. Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods.* 4(1):63-72.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics.* 2008 Sep 19;9:386.

Mou XSS, Edwards RA, Hodson RE, Moran MA. 2008. Bacterial carbon processing by generalist species in the coastal ocean. *Nature.* 451(7179):708-11.

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, et al. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 7;33(17):5691-702.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 35(21):7188-96.

Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 2010 Nov 1;38(20):e191.

Ronaghi M, Uhlen M, Nyren P. 1998. A sequencing method based on real-time pyrophosphate. *Science* 281: 363, 365.

Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* 242: 84-89.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. 2010. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 38(Database issue):D5-16.

Sansone SA, Rocca-Serra P, Brandizi M, Brazma A, Field D, Fostel J, Garrow AG, Gilbert J, Goodsaid F, Hardy N, Jones P, Lister A, Miller M, Morrison N, Rayner T, Sklyar N, Taylor C, Tong W, Warner G, Wiemann S. (2008). The First MGED RSBI (ISA-TAB) Workshop: "Can a Simple Format Work for Complex Studies?". *OMICS*. 12(2):143-9.

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg JL, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenger A, Sansone SA, Shah N, Whetzel PL, Suzanna L. (2007). The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol.* 25(11):1251-5.

Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J, Wooley J. Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res.* 2011 Jan;39(Database issue):D546-51.

Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T, Brazma A, Brinkman RR, Michael Clark A, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, Julian RK Jr, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Le Novère N, Leebens-Mack J, Lewis SE, Lord P, Mallon AM, Marthandan N, Masuya H, McNally R, Mehrle A, Morrison N, Orchard S, Quackenbush J, Reecy JM, Robertson DG, Rocca-Serra P, Rodriguez H, Rosenfelder H, Santoyo-Lopez J, Scheuermann RH, Schober D, Smith B, Snape J, Stoeckert CJ Jr, Tipton K, Sterk P, Untergasser A, Vandesompele J, Wiemann S. (2008) MIBBI: A Minimum Information Checklist Resource. *Nat Biotechnol* 26(8):889-96.

Figure legends

Figure 1. Sequences are compared to the MG-RAST protein database that provides a non-redundant integration of many databases (INSDC, SEED, IMG, KEGG, and EGGNOGs),

supporting many complementary views into the data with one similarity search. Show are functional distribution based on COG annotations.

Figure 2. An example of a comparative view in MG-RAST. A circular tree representing phylogenetic profiles from four samples is compared. Each node can be expanded to get detailed information about the distribution for each sample. Color shading of the family names indicates class membership.