MG-RAST, a Metagenomics Service for the Analysis of Microbial Community Structure and Function

Elizabeth M. Glass and Folker Meyer

Abstract

Molecular ecology approaches are rapidly advancing our understanding of microbial communities involved in the synthesis and degradation of hydrophobic organics involved with major consequences for applications in climate change, environmental pollution, human health, and biotechnology. Metagenomics allows researchers to inventory microbial genes in various environments to understand the genetic potential of uncultured bacteria and archaea. Metagenomics enables us to sequence genomes from a complex assemblage of microbes in a culture-independent manner. Amplicon and WGS studies are the most widely employed methods to estimate "who is there" and "what they are doing." Metagenomics allows researchers to access the functional and metabolic diversity of microbial communities.

Since 2008, MG-RAST serves as a repository for metagenomic datasets and as an analysis provider. Currently, the system has analyzed and hosts over 130,000 datasets. Over the years, MG-RAST has undergone a significant number of revisions to accommodate the dramatic growth in dataset size, new data types, and wider system adoption among the research community.

Keywords Automated pipeline, Comparative analysis, High throughput, Metagenomics, Sequence quality

1 Introduction

Metagenomics allows researchers to inventory microbial genes in various environments to understand the genetic potential of uncultured bacteria and archaea. Metagenomics powers our ability to sequence genomes from a complex assemblage of microbes in a culture-independent manner. Amplicon (16 s, 18 s, ITS) and WGS (whole shotgun sequence) studies are the most widely employed methods to estimate "who is there" and "what they are doing." However, understanding the sequence quality is of primary importance. Metagenomics analyses rely on automated analysis tools, and therefore, the estimation of the quality of the sequences is of great importance. Trying to derive meaningful biological inferences from potentially questionable data would be detrimental.

T.J. McGenity et al. (eds.), *Hydrocarbon and Lipid Microbiology Protocols*, Springer Protocols Handbooks (2016) 69–87,
DOI 10.1007/8623_2015_119, © Springer-Verlag Berlin Heidelberg 2015, Published online: 26 July 2015

Concerns about sequence quality are not the only hurdle the scientific community faces with the study of metagenomics data. Another problem is the gap between the shrinking costs of sequencing and the more or less stable costs of computing. Analysis of the sequence data has become the limiting factor, not initial data generation, as seen in the past. Wilkening et al. [1] provide a real currency cost for the analysis of 100 gigabasepairs of DNA sequence data using BLASTX on Amazon's EC2 service: $300,000. A more recent study by the University of Maryland researchers [2] estimates the computation for a terabase of DNA shotgun data using their CLOVR metagenome analysis pipeline at over $5 million per terabase.

In order to be able to conduct meaningful comparative analyses of metagenomic samples, one must consider metadata (contextual data). This provides an essential complement to experimental data, helping to answer questions about its source, mode of collection, and reliability. Metadata collection and interpretation have become vital to the genomics and metagenomics community, but considerable challenges remain, including exchange, curation, and distribution.

Since 2008, MG-RAST [3] serves as a repository for metagenomic datasets and as an analysis provider. Currently, the system has analyzed and hosts over 130,000 datasets. Over the years, MG-RAST has undergone a significant number of revisions to accommodate the dramatic growth in dataset size, new data types, and wider system adoption among the research communities. We have made both substantial engineering changes and modifications to the bioinformatics pipeline to accommodate the evolving needs of novel sequencing technologies and growing data volumes. The MG-RAST system has been an early adopter of the minimal checklist standards and the expanded biome-specific environmental packages devised by the Genomics Standards Consortium [4] and provides an easy-to-use uploader for metadata capture at the time of data submission.

2 Materials

2.1 Database

The MG-RAST automated analysis pipeline uses the M5nr (MD5-based nonredundant protein database) [5] for annotation. The M5nr is an integration of many sequence databases into one single, searchable database (plus an index). A single similarity search (using BLAST [6] or BLAT [7]) will allow the user to retrieve similarities to several databases:

- EBI: European Bioinformatics Institute [8]
- GO: Gene Ontology [9]

- JGI: Joint Genome Institute [10]
- KEGG: Kyoto Encyclopedia of Genes and Genomes [11]
- NCBI: National Center for Biotechnology Information [12]
- Phantome: Phage Annotation Tools and Methods [13]
- SEED: The SEED Project [14]
- UniProt: UniProt Knowledgebase [15]
- VBI: Virginia Bioinformatics Institute [16]
- eggNOG: evolutionary genealogy of genes – non-supervised orthologous groups [17]

Computing of sequence similarity results is becoming a limiting factor in metagenome analysis. Sequence similarity search results encoded in an open, exchangeable format distributed with the sequence sets have the potential to limit the needs for computational reanalysis of datasets.

2.2 Analysis Pipeline

The pipeline shown in Fig. 1 contains a significant number of improvements to previous versions. We have transitioned from using just the SEED database to using the M5nr, made several key algorithmic improvements that were needed to support the flood of user-generated data, and used dedicated software to perform gene prediction instead of using a similarity-based approach which reduces runtime requirements. The additional clustering of proteins at 90% identity reduces data while preserving biological signals. We also restrict the pipeline annotations only to protein-coding genes and ribosomal RNA (rRNA) genes.

In Sect. 3, we describe each step of the pipeline in detail. All datasets generated by the individual stages of the processing pipeline are made available as downloads.

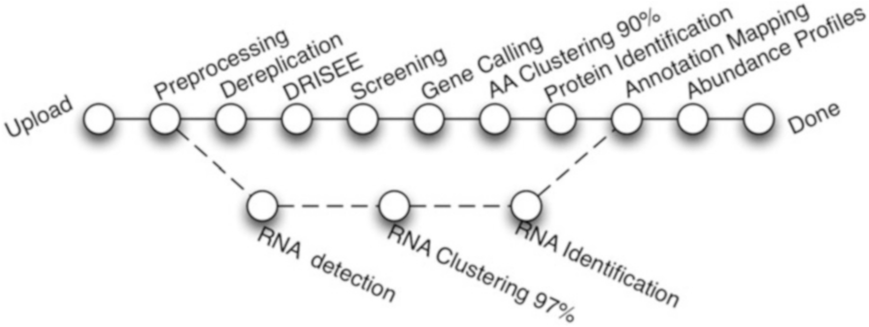

Fig. 1 Details of the analysis pipeline for MG-RAST. After upload, the pipeline diverges for amplicon and WGS datasets. Amplicon samples run through RNA detection, clustering, and identification, while WGS has more steps that start with preprocessing and data quality assessment through annotation and abundance profile creation

2.3 Compute Resources

While originally MG-RAST has been built like a traditional cluster-based bioinformatics system, the most recent version of MG-RAST embraces cloud technologies to enable the use of computer resources anywhere. MG-RAST data is stored in SHOCK and computing is orchestrated by AWE [18]. These technologies were developed to enable execution on a variety of computational platforms; currently, computational resources are contributed by the DOE Magellan cloud at Argonne National Laboratory, Amazon EC2 Web Services, and finally a number of traditional clusters. An installation of the pipeline exists at DOE's NERSC supercomputing center. Currently, the system handles over 2 terabasepairs per month.

3 Methods

The pipeline diverges after upload for amplicon and whole genome shotgun (WGS) samples. The WGS pipeline is composed of several steps from the removal of low-quality reads, dereplication, gene calling, and annotation to creation of abundance profiles. rRNA samples run through RNA detection, clustering, and identification.

3.1 The WGS Pipeline

3.1.1 Preprocessing

After upload, data is preprocessed by using SolexaQA [19] to trim low-quality regions from FASTQ data. Platform-specific approaches are used for 454 data submitted in FASTA format: reads more than two standard deviations away from the mean read length are discarded thereafter [20]. All sequences submitted to the system are available, but discarded reads will not be analyzed further.

3.1.2 Dereplication

For shotgun metagenome and shotgun metatranscriptome datasets, we perform a dereplication step. We use a simple k-mer approach to rapidly identify all 20 character prefix identical sequences. This step is required in order to remove artificial duplicate reads (ADRs) [21]. Instead of simply discarding the ADRs, we set them aside and use them later. We note that dereplication is not suitable for amplicon datasets that are likely to share common prefixes.

3.1.3 DRISEE

MG-RAST v3 uses DRISEE (duplicate read inferred sequencing error estimation) [22] to analyze the sets of ADRs and determine the degree of variation among prefix-identical sequences derived from the same template. DRISEE provides positional error estimates that can be used to inform read trimming within a sample. It also provides global (whole sample) error estimates that can be used to identify samples with high or varying levels of sequencing error that may confound downstream analyses, particularly in the case of studies that utilize data from multiple sequencing samples.

3.1.4 Screening

The pipeline provides the option of removing reads that are near-exact matches to the genomes of a handful of model organisms, including fly, mouse, cow, and human. The screening stage uses Bowtie (a fast, memory-efficient, short read aligner) [23], and only reads that do not match the model organisms pass into the next stage of the annotation pipeline. This option will remove all reads similar to the human genome and render them inaccessible. This decision was made in order to avoid storing any human DNA on MG-RAST.

3.1.5 Gene Calling

The previous version of MG-RAST used similarity-based gene predictions, an approach that is significantly more expensive computationally than de novo gene prediction. After an in-depth investigation of tool performance [24], we have moved to a machine learning approach: FragGeneScan [25]. Using this approach, we can now predict coding regions in DNA sequences of 75 bp and longer. Our novel approach also enables the analysis of user-provided assembled contigs. FragGeneScan is trained for prokaryotes only. While it will identify proteins for eukaryotic sequences, the results should be viewed as more or less random.

3.1.6 AA Clustering

MG-RAST builds clusters of proteins at the 90% identity level using the UCLUST [26] implementation in QIIME [27] preserving the relative abundances. These clusters greatly reduce the computational burden of comparing all pairs of short reads, while clustering at 90% identity preserves sufficient biological signals.

3.1.7 Protein Identification

Once created, a representative (the longest sequence) for each cluster is subjected to similarity analysis. Instead of BLAST, we use sBLAT, an implementation of the BLAT algorithm, which we parallelized using Open MPI for this work.

Once the similarities are computed, we present reconstructions of the species content of the sample based on the similarity results. We reconstruct the putative species composition of the sample by looking at the phylogenetic origin of the database sequences hit by the similarity searches.

3.1.8 Annotation Mapping

Sequence similarity searches are computed against a protein database derived from the M5nr, which provides nonredundant integration of many databases. Users can easily change views without recomputating data. For example, COG and KEGG views can be displayed, which both show the relative abundances of histidine biosynthesis in a dataset of four cow rumen metagenomes.

Help in interpreting results: MG-RAST searches the nonredundant M5nr and M5RNA databases in which each sequence is unique. These two databases are built from multiple sequence database sources, and the individual sequences may occur multiple times in different strains and species (and sometimes genera) with 100% identity. In these circumstances, choosing the "right taxonomic information" is not a straightforward process. To optimally serve a number of different use cases, we have implemented three methods – best hit, representative hit, and lowest common ancestor – for end users to determine the number of hits (occurrences of the input sequence in the database) reported for a given sequence in their dataset.

- *Best hit*: The best hit classification reports the functional and taxonomic annotation of the best hit in the M5nr for each feature. In those cases where the similarity search yields multiple same-scoring hits for a feature, we do not choose any single correct label. For this reason MG-RAST double counts all annotations with identical match properties and leaves determination of truth to our users. While this approach aims to inform about the functional and taxonomic potential of a microbial community by preserving all information, subsequent analysis can be biased because of a single feature having multiple annotations, leading to inflated hit counts. For users looking for a specific species or function in their results, the best hit classification is likely what is wanted.
- *Representative hit*: The representative hit classification selects a single, unambiguous annotation for each feature. The annotation is based on the first hit in the homology search and the first annotation for that hit in the database. This approach makes counts additive across functional and taxonomic levels and thus allows, for example, the comparison of functional and taxonomic profiles of different metagenomes.
- *Lowest common ancestor* (*LCA*): To avoid the problem of multiple taxonomic annotations for a single feature, MG-RAST provides taxonomic annotations based on the widely used LCA method introduced by MEGAN [28]. In this method all hits are collected that have a bit score close to the bit score of the best hit. The taxonomic annotation of the feature is then determined by computing the LCA of all species in this set. This replaces all taxonomic annotations from ambiguous hits with a single higher-level annotation in the NCBI taxonomy tree.

3.1.9 Abundance Profiles

Abundance profiles are the primary data product that MG-RAST's user interface uses to display information on the datasets. Using the abundance profiles, the MG-RAST system defers making a decision on when to transfer annotations. Since there is no well-defined threshold that is acceptable for all use cases, the abundance profiles contain all similarities and require their users to set cutoff values. The threshold for annotation transfer can be set by using the following parameters: e-value, percent identity, and minimal alignment length.

The taxonomic profiles use the NCBI taxonomy. All taxonomic information is projected against this data. The functional profiles are available for data sources that provide hierarchical information. These currently comprise SEED subsystems, KEGG orthologs, and COGs.

The SEED subsystems represent an independent reannotation effort that powers, for example, the RAST [29] effort. Manual curation of subsystems makes them an extremely valuable data source. The page at http://pubseed.theseed.org//SubsysEditor.cgi allows browsing the subsystems.

Subsystems represent a four-level hierarchy:

1. Subsystem level 1 – highest level
2. Subsystem level 2
3. Subsystem level 3 – similar to a KEGG pathway
4. Subsystem level 4 – actual functional assignment to the feature in question

KEGG Orthologs. MG-RAST uses the KEGG enzyme number hierarchy to implement a four-level hierarchy. We note that KEGG data is no longer available for free download. We thus have to rely on using the latest freely downloadable version of the data:

1. KEGG level 1 – first digit of the EC number (EC:X.*.*.*)
2. KEGG level 2 – first two digits of the EC number (EC:X.Y.*.*)
3. KEGG level 3 – first three digits of the EC number (EC:X:Y:Z:.*)
4. KEGG level 4 – entire four digits of the EC number

The high-level KEGG categories are as follows:

1. Cellular processes
2. Environmental information processing
3. Genetic information processing
4. Human diseases
5. Metabolism
6. Organizational systems

COG and eggNOG Categories. The high-level COG and eggNOG categories are as follows:

1. Cellular processes
2. Information storage and processing
3. Metabolism
4. Poorly characterized

3.2 The rRNA Pipeline

This pipeline takes rRNA sequence data or WGS data (annotates only RNAs in WGS) that proceeds through the following steps.

3.2.1 Preprocessing

After upload, data is preprocessed by using SolexaQA to trim low-quality regions from FASTQ data. Platform-specific approaches are used for 454 data submitted in FASTA format: reads more than two standard deviations away from the mean read length are discarded following. All sequences submitted to the system are available, but discarded reads will not be analyzed further.

3.2.2 rRNA Detection

Reads are identified as rRNA through a simple rRNA detection step. An initial BLAT search against a reduced RNA database efficiently identifies RNA. The reduced database is a 90% identity clustered version of the SILVA database and is used merely to rapidly identify sequences with similarities to ribosomal RNA.

3.2.3 rRNA Clustering

The rRNA-similar reads are then clustered, using UCLUST, at 97% identity, and the longest sequence is picked as the cluster representative (abundance values are retained, however).

3.2.4 rRNA Identification

A BLAT similarity search for the longest cluster representative is performed against the M5RNA database, integrating SILVA [30], Greengenes [31], and RDP [32].

3.3 Using the MG-RAST User Interface

The MG-RAST system provides a rich web user interface that covers all aspects of the metagenome analysis, from data upload to ordination analysis. The web interface can also be used for data discovery. Metagenomic datasets can be easily selected individually or on the basis of filters such as technology (including read length), quality, sample type, and keyword, with dynamic filtering of results based on similarity to known reference proteins or taxonomy. For example, a user might want to perform a search such as "phylum eq. 'actinobacteria' and function in KEGG pathway Lysine Biosynthesis and sample in 'Ocean'" to extract sets of reads matching the appropriate functions and taxa across metagenomes. The results can be displayed in familiar formats, including bar charts, trees that incorporate abundance information, heatmaps, or principal component analyses, or exported in tabular form. The raw or processed data can be recovered via download pages. Metabolic reconstructions based on mapping to KEGG pathways are also provided.

Sample selection is crucial for understanding large-scale patterns when multiple metagenomes are compared. Accordingly, MG-RAST supports MIxS and MIMARKS [33] (as well as domain-specific plug-ins for specialized environments not extending the minimal GSC standards); several projects, including TerraGenome, HMP (Human Microbiome Project), TARA (Oceans Science), and EMP (Earth Microbiome Project), use these GSC standards, enabling standardized queries that integrate new samples into these massive datasets.

One key aspect of the MG-RAST approach is the creation of smart data products enabling the user at the time of analysis to determine the best parameters for, for example, a comparison between samples. This is done without the need for recomputing results.

3.3.1 Navigation

The MG-RAST web site is rich with functionality and offers a lot of different options. The site at http://metagenomics.anl.gov has five main pages and a home page, shown in blue in Fig. 2:

- Download page – lists all publicly available data for download. The data is structured into projects.
- Browse page – allows interactive browsing of all datasets and is powered by metadata.
- Search page – allows identifier, taxonomy, and function-driven searches against all public data.
- Analysis page – enables in-depth analyses and comparisons between datasets.
- Upload page – allows users to provide their samples and metadata to MG-RAST.
- Home (overview) page – provides an overview for each individual dataset.

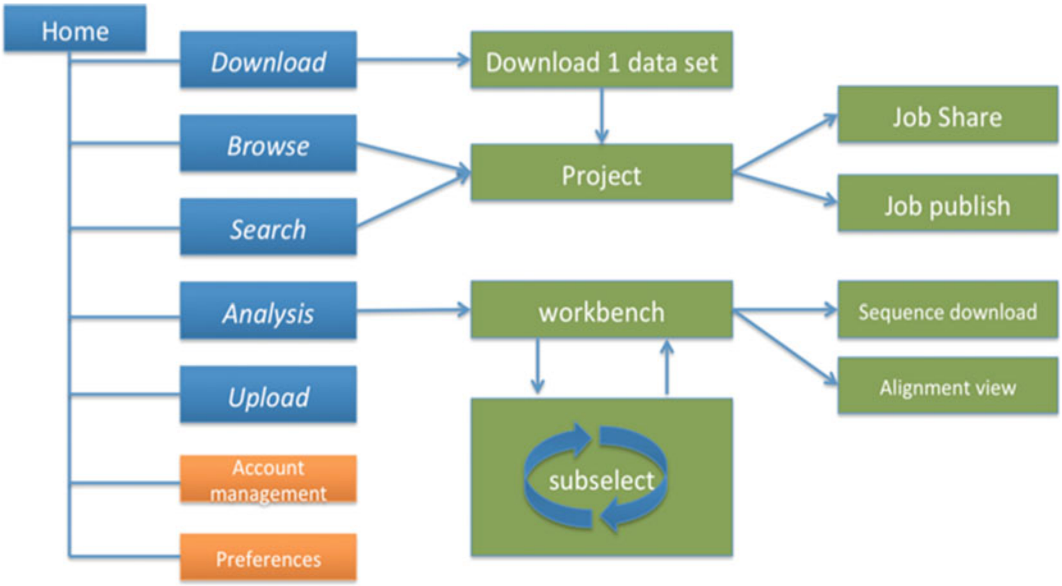

Fig. 2 Sitemap for the MG-RAST web site. On the site map, the main pages are shown in *blue* and management pages in *orange*. The *green boxes* represent pages that are not directly accessible from the home page

3.3.2 Upload Page

Data and metadata can be uploaded in the form of spreadsheets along with the sequence data by using both the ftp and the http protocols. The web uploader will automatically split larger files and allow parallel uploads. MG-RAST supports datasets that are augmented with rich metadata using the standards and technology developed by the GSC. Each user has a temporary storage location inside the MG-RAST system. This inbox provides temporary storage for data and metadata to be submitted to the system. Using the inbox, users can extract compressed files, convert a number of vendor-specific formats to MG-RAST submission-compliant formats, and obtain an MD5 checksum for verifying that transmission to MG-RAST has not altered the data. The web uploader has been optimized for large datasets of over 100 gigabasepairs, often resulting in file sizes in excess of 150 GB.

3.3.3 Browse Page: Metadata-Enabled Data Discovery

The browse page lists all datasets visible to the user. This page also provides an overview of the nonpublic datasets submitted by the user or shared with users. The interactive metagenome browse table provides an interactive graphical means to discover data based on technical data (e.g., sequence type or dataset size) or metadata (e.g., location or biome).

3.3.4 Project Page

The project page provides a list of datasets and metadata for a project. The table at the bottom of the project page provides access to the individual metagenomes by clicking on the identifiers in the first column. In addition, the final column provides downloads for metadata, submitted data, and the analysis results via the three labeled arrows. For the dataset owners, the project page provides an editing capability using a number of menu entries at the top of the page:

- Share Project – make the data in this project available to third parties via sending them access tokens.
- Add Jobs – add additional datasets to this project.
- Edit Project Data – edit the contents of this page.
- Upload Info – upload information to be displayed on this page.
- Upload MetaData – upload a metadata spreadsheet for the project.
- Export MetaData2 – export the metadata spreadsheet for this project.

3.3.5 Overview Page

MG-RAST automatically creates an individual summary page for each dataset. This metagenome overview page provides a summary of the annotations for a single dataset. The page is made available by the automated pipeline once the computation is finished. This page is a good starting point for looking at a particular dataset. It provides a significant amount of information on technical details and biological content. The page is intended as a single point of reference for metadata, quality, and data. It also provides an initial overview of the analysis results for individual datasets with default parameters. Further analyses are available on the analysis page.

Technical Details on Sequencing and Analysis

The overview page provides the MG-RAST ID for a dataset, a unique identifier that is usable as an accession number for publications. Additional information such as the name of the submitting PI and organization and a user-provided metagenome name are displayed at the top of the page as well. A static URL for linking to the system that will be stable across changes to the MG-RAST web interface is provided in the MG-RAST user manual found at: ftp://ftp.metagenomics.anl.gov/data/manual/mg-rast-manual.pdf.

MG-RAST provides an automatically generated paragraph of text describing the submitted data and the results computed by the pipeline. By means of the project information, we display additional information provided by the data submitters at the time of submission or later.

One of the key diagrams in MG-RAST is the sequence breakdown pie chart (Fig. 3) classifying the submitted sequences submitted into several categories according to their annotation status. As detailed in the description of the MG-RAST v3 pipeline above, the features annotated in MG-RAST are protein-coding genes and ribosomal proteins.

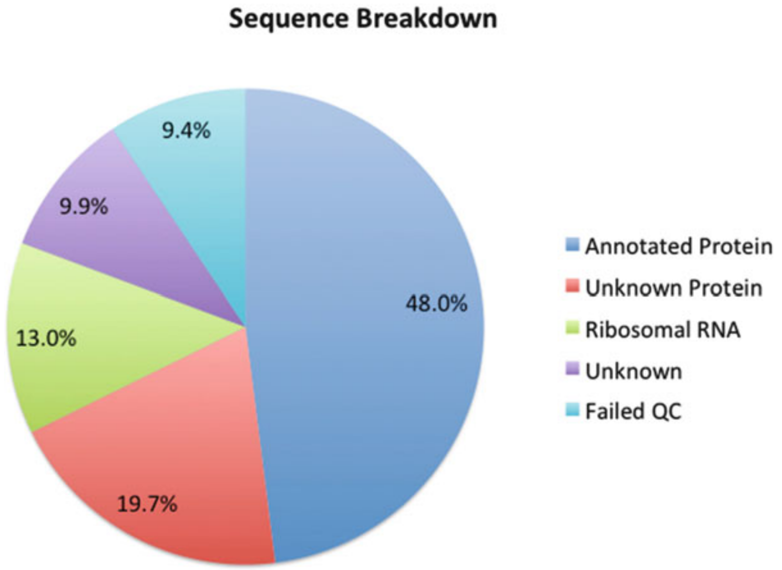

Fig. 3 Sequences to the pipeline are classified into one of five categories: *gray* = failed the QC, *red* = unknown sequences, *yellow* = unknown function but protein coding, *green* = protein coding with known function, and *blue* = ribosomal RNA. For this example, over 50% of sequences were either filtered by QC or failed to be recognized as either protein coding or ribosomal

Note that for performance reasons no other sequence features are annotated by the default pipeline. Other feature types such as small RNAs or regulatory motifs (e.g., CRISPRs [34]) not only will require significantly higher computational resources but also are frequently not supported by the unassembled short reads that constitute the vast majority of today's metagenomic data in MG-RAST.

The quality of the sequence data coming from next-generation instruments requires careful design of experiments, lest the sensitivity of the methods is greater than the signal-to-noise ratio the data supports.

The overview page also provides metadata for each dataset to the extent that such information has been made available. Metadata enables other researchers to discover datasets and compare annotations. MG-RAST requires standard metadata for data sharing and data publication. This is implemented using the standards developed by the Genomics Standards Consortium.

All metadata stored for a specific dataset is available in MG-RAST; we merely display a standardized subset in this table. A link at the bottom of the table ("More Metadata") provides access to a table with the complete metadata. This enables users to provide extended metadata going beyond the GSC minimal standards. A mechanism to provide community consensus extensions to the minimal checklists and the environmental packages is explicitly encouraged but not required when using MG-RAST.

Metagenome Quality Control

The analysis flowchart and analysis statistics provide an overview of the number of sequences at each stage in the pipeline. The text block next to the analysis flowchart presents the numbers next to their definitions.

Source Hits Distribution

The source hits distribution shows what percentage of the predicted protein features could be annotated with similarity to a protein of known function and which database those functions were from. In addition, ribosomal RNA genes are mapped to the rRNA databases.

The display will show the number of records in the M5nr protein database and in the M5RNA ribosomal databases.

Other Statistics

MG-RAST also provides a quick link to other statistics. For example, the analysis statistics and analysis flowchart provide sequence statistics for the main steps in the pipeline from raw data to annotation, describing the transformation of the data between steps. Sequence length and GC histograms display the distribution before and after quality control steps. Metadata is presented in a searchable table that contains contextual metadata describing sample location, acquisition, library construction, and sequencing using GSC compliant metadata. All metadata can be downloaded from the table.

3.3.6 Biological Part of the Overview Page

The taxonomic hit distribution display divides taxonomic units into a series of pie charts of all the annotations grouped at various taxonomic ranks (domain, phylum, class, order, family, genus). The subsets are selectable for downstream analysis; this also enables downloads of subsets of reads, for example, those hitting a specific taxonomic unit.

Rank Abundance

The rank abundance plot provides a rank-ordered list of taxonomic units at a user-defined taxonomic level, ordered by their abundance in the annotations.

Rarefaction

The rarefaction curve of annotated species richness is a plot (*see* Fig. 4) of the total number of distinct species annotations as a function of the number of sequences sampled. The slope of the right-hand part of the curve is related to the fraction of sampled species that are rare. On the left, a steep slope indicates that a large fraction of the species diversity remains to be discovered. If the curve becomes flatter to the right, a reasonable number of individuals are sampled: more intensive sampling is likely to yield only few additional species. Sampling curves generally rise quickly at first and then level off toward an asymptote as fewer new species are found per unit of individuals collected.

The rarefaction curve is derived from the protein taxonomic annotations and is subject to problems stemming from technical artifacts. These artifacts can be similar to the ones affecting amplicon sequencing [35], but the process of inferring species from protein similarities may introduce additional uncertainty.

Alpha Diversity

In this section we display an estimate of the alpha diversity based on the taxonomic annotations for the predicted proteins. The alpha diversity is presented in context of other metagenomes in the same project.

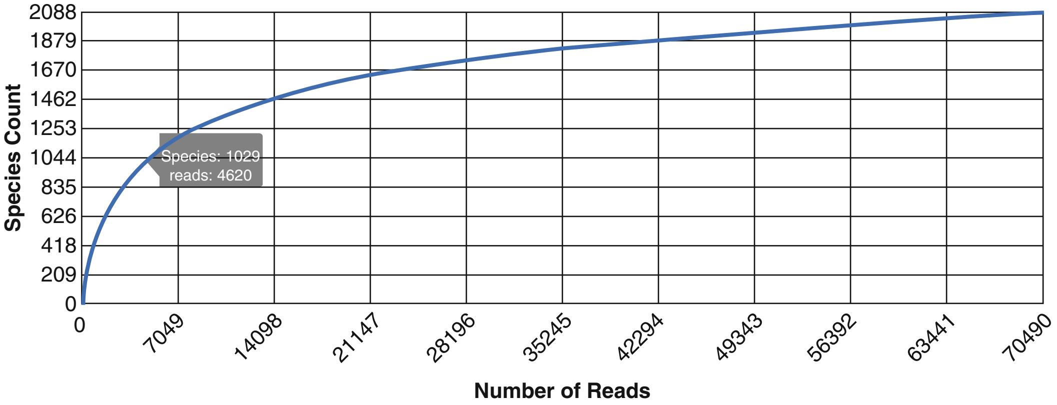

Fig. 4 Rarefaction plot showing a curve of annotated species richness. This curve is a plot of the total number of distinct species annotations as a function of the number of sequences sampled

The alpha diversity estimate is a single number that summarizes the distribution of species-level annotations in a dataset. The Shannon diversity index is an abundance-weighted average of the logarithm of the relative abundances of annotated species. We compute the species richness as the antilog of the Shannon diversity.

Functional Categories

This section contains four pie charts providing a breakdown of the functional categories for KEGG, COG, SEED subsystems, and eggNOGs. Clicking on the individual pie chart slices will save the respective sequences to the workbench. The relative abundance of sequences per functional category can be downloaded as a spreadsheet, and users can browse the functional breakdowns.

A more detailed functional analysis, allowing the user to manipulate parameters for sequence similarity matches, is available from the analysis page.

3.3.7 Analysis Page

The MG-RAST annotation pipeline produces a set of annotations for each sample; these annotations can be interpreted as functional or taxonomic abundance profiles. The analysis page can be used to view these profiles for a single metagenome or to compare profiles from multiple metagenomes using various visualizations (e.g., heatmap) and statistics (e.g., PCoA, normalization).

The page is divided into three parts following a typical workflow:

1. Data type
   Selection of an MG-RAST analysis scheme, that is, selection of a particular taxonomic or functional abundance profile mapping. For taxonomic annotations, since there is not always a unique mapping from hit to annotation, we provide three interpretations: best hit, representative hit, and lowest common ancestor. When choosing the LCA annotations, not all downstream tools are available. The reason is the fact that for the LCA annotations, not all sequences will be annotated to the same level: classifications are returned on different taxonomic levels.

   Functional annotations can be grouped into mappings to functional hierarchies or can be displayed without a hierarchy. In addition, the recruitment plot displays the recruitment of protein sequences against a reference genome. Each selected data type has data selections and data visualizations specific for it.
2. Data selection of sample and parameters. This dialog allows the selection of multiple metagenomes that can be compared individually or selected and compared as groups. Comparison is always relative to the annotation source, e-value, and percent identity cutoffs selectable in this section. In addition to the metagenomes available in MG-RAST, sets of sequences previously saved in the workbench can be selected for visualization.

3. Data visualization. Data visualization and comparison. Depending on the selected profile type, the profiles for the metagenomes can be visualized and compared by using bar charts, trees, spreadsheet-like tables, heatmaps, PCoA, rarefaction plots, circular recruitment plot, and KEGG maps.

The data selection dialog provides access to datasets in four ways. The four categories can be selected from a pull-down menu:

- Private data – list of private or shared datasets for browsing under available metagenomes.
- Collections – defined sets of metagenomes grouped for easier analysis. This is the recommended way of working with the analysis page.
- Projects – global groups of datasets grouped by the submitting user. The project name will be displayed.
- Public data – display of all public datasets.

When using collections or projects, data can also be grouped into one set per collection or project and subsequently compared or added.

Normalization

Normalization refers to a transformation that attempts to reshape an underlying distribution. A large number of biological variables exhibit a log-normal distribution, meaning that when the data is transformed with a log transformation, the values exhibit a normal distribution. Log transformation of the count data makes a normalized data product that is more likely to satisfy the assumptions of additional downstream tests such as ANOVA or t-tests. Standardization is a transformation applied to each distribution in a group of distributions so that all distributions exhibit the same mean and the same standard deviation. This removes some aspects of intersample variability and can make data more comparable. This sort of procedure is analogous to commonly practiced scaling procedures but is more robust in that it controls for both scale and location.

Rarefaction

The rarefaction view is available only for taxonomic data. The rarefaction curve of annotated species richness is a plot (*see* Fig. 5) of the total number of distinct species annotations as a function of the number of sequences sampled. As shown in Fig. 5, multiple datasets can be included.

The rarefaction curve is derived from the protein taxonomic annotations and is subject to problems stemming from technical artifacts. These artifacts can be similar to the ones affecting amplicon sequencing [31], but the process of inferring species from protein similarities may introduce additional uncertainty.

On the analysis page, the rarefaction curves are calculated from the table of species abundance. The curves represent the average number of different species annotations for subsamples of the complete dataset.

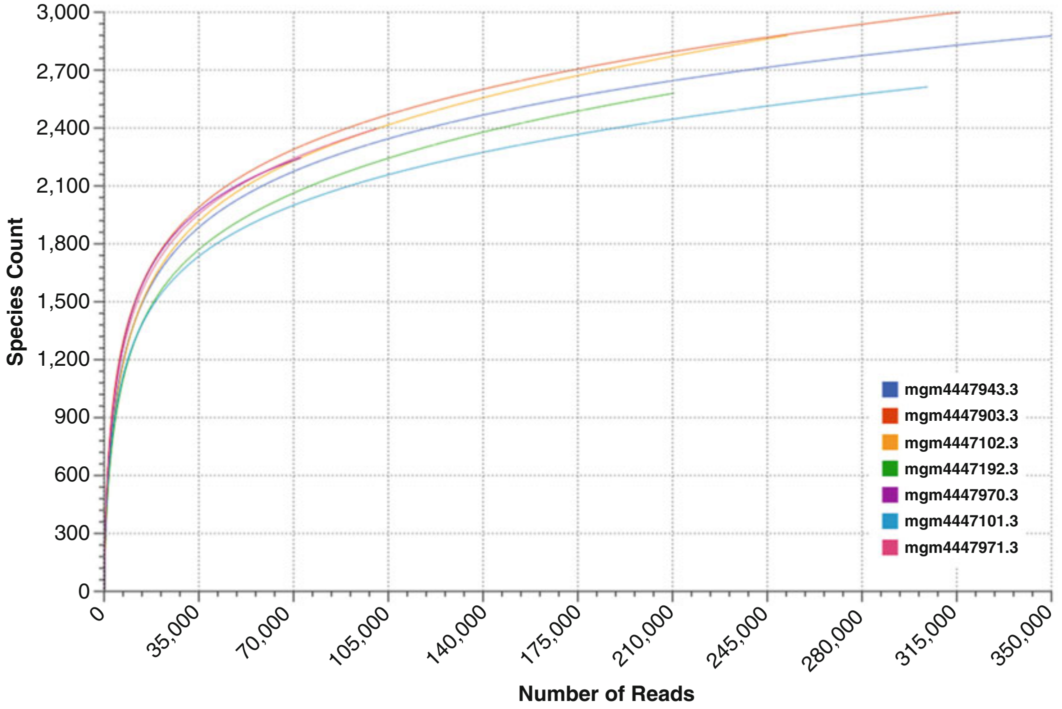

Fig. 5 Rarefaction plot showing a curve of annotated species richness. This curve is a plot of the total number of distinct species annotations as a function of the number of sequences sampled

Heatmap/Dendrogram

The heatmap/dendrogram allows an enormous amount of information to be presented in a visual form that is amenable to human interpretation. Dendrograms are trees that indicate similarities between annotation vectors. The MG-RAST heatmap/dendrogram has two dendrograms, one indicating the similarity/dissimilarity among metagenomic samples (x-axis dendrogram) and another indicating the similarity/dissimilarity among annotation categories (e.g., functional roles, the y-axis dendrogram). A distance metric is evaluated between every possible pair of sample abundance profiles. A clustering algorithm (e.g., ward-based clustering) then produces the dendrogram trees. Each square in the heatmap dendrogram represents the abundance level of a single category in a single sample. The values used to generate the heatmap/dendrogram figure can be downloaded as a table by clicking on the download button.

Ordination

MG-RAST uses principal coordinate analysis (PCoA) to reduce the dimensionality of comparisons of multiple samples that consider functional or taxonomic annotations. A key feature of PCoA-based analyses is that users can compare components not just to each other but also to metadata-recorded variables (e.g., sample pH, biome, DNA extraction protocol) to reveal correlations between extracted variation and metadata-defined characteristics of the samples.

It is also possible to couple PCoA with higher-resolution statistical methods in order to identify individual sample features (taxa or functions) that drive correlations observed in PCoA visualizations.

This coupling can be accomplished with permutation-based statistics applied directly to the data before calculation of distance measures used to produce PCoAs; alternatively, one can apply conventional statistical approaches (e.g., ANOVA or Kruskal-Wallis test) to groups observed in PCoA-based visualizations.

Bar Charts

The bar chart visualization option on the analysis page has a built-in ability to drill down by clicking on a specific category. You can expand the categories to show the normalized abundance (adjusted for sample sizes) at various levels. The abundance information displayed can be downloaded into a local spreadsheet. Once a subselection has been made (e.g., the domain *Bacteria* selected), data can be sent to the workbench for detailed analysis. In addition, reads from a specific level can be added into the workbench.

Tree Diagram

The tree diagram allows comparison of datasets against a hierarchy (e.g., subsystems or the NCBI taxonomy). The hierarchy is displayed as a rooted tree, and the abundance (normalized for dataset size or raw) for each dataset in the various categories is displayed as a bar chart for each category. By clicking on a category (inside the circle), detailed information can be requested for that node.

Table

The table tool creates a spreadsheet-based abundance table that can be searched and restricted by the user. Tables can be generated at user-selected levels of phylogenetic or functional resolution. Table data can be visualized by using Krona [36] or can be exported in BIOM format to be used in other tools (e.g., QIIME). The tables also can be exported as tab-separated text. Abundance tables serve as the basis for all comparative analysis tools in MG-RAST, from PCoA to heatmap/dendrograms.

Workbench

The workbench was designed to allow users to select subsets of the data for comparison or export. Specifically, the workbench supports selecting sequence features and submitting them to further analysis or other analysis. A number of use cases are described below. An important limitation with the current implementation is that data sent to the workbench exists only until the current session is closed.

3.3.8 Metadata, Publishing, and Sharing

MG-RAST is both an analytical platform and a data integration system. To enable data reuse, for example, for meta-analyses, we require that all data being made available to third parties contain at least minimal metadata. The MG-RAST team has decided to follow the minimal checklist approach used by the GSC.

MG-RAST provides a mechanism to make data and analyses publicly accessible. Only the submitting user can make data public on MG-RAST. As stated above, metadata is mandatory for dataset publication.

In addition to publishing, data and analysis can also be shared with specific users. To share data, users simply enter their email address via clicking sharing on the overview page.

References

1. Wilkening J, Wilke A, Desai N, Meyer F (2009) Using clouds for metagenomics: a case study. In: Cluster. IEEE Computer Society, pp. 1–6. ISBN: 978-1-4244-5012-1
2. Angiuoli S, Matalka M, Gussman A et al (2011) Clovr: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinformatics 12:356
3. Meyer F, Paarmann D, D'Souza M et al (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9(1):386
4. Field D, Amaral-Zettler L, Cochrane G et al (2011) The genomic standards consortium. PLoS Biol 9(6), e1001088
5. Wilke A, Harrison T, Wilkening J et al (2012) The m5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. BMC Bioinformatics 13:141
6. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410
7. Kent WJ (2002) Blat–the blast-like alignment tool. Genome Res 12(4):656–64
8. Brooksbank C, Bergman MT, Apweiler R et al (2014) The European bioinformatics Institute's data resources 2014. Nucleic Acids Res 42(Database issue):D18–25
9. Reference Genome Group of the Gene Ontology Consortium (2009) The gene ontology's reference genome project: a unified framework for functional annotation across species. PLoS Comput Biol 5(7):e1000431
10. Markowitz VM, Ivanova NN, Szeto E et al (2008) IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36(Database issue):D534–538
11. Kanehisa M (2002) The KEGG database. Novartis Found Symp 247:91–101
12. Benson DA, Cavanaugh M, Clark K (2013) Genbank. Nucleic Acids Res 41(Database issue):D36–42
13. Dwivedi B, Schmieder R, Goldsmith DB et al (2012) PhiSiGns: an online tool to identify signature genes in phages and design PCR primers for examining phage diversity. BMC Bioinformatics 4(13):37
14. Overbeek R, Begley T, Butler RM et al (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33(17):5691–5702
15. Magrane M, Uniprot Consortium (2011) UniProt knowledgebase: a hub of integrated protein data. Database (Oxford). doi:10.1093/database/bar009
16. Snyder EE, Kampanya N, Lu J et al (2007) PATRIC: the VBI pathosystems resource integration center. Nucleic Acids Res 35(Database issue):D401–406
17. Jensen LJ, Julien P, Kuhn M et al (2008) Eggnog: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res 36(Database issue):D250–4
18. Tang W, Wilkening J, Desai N, Gerlach W, Wilke A, Meyer F (2013) A scalable data analysis platform for metagenomics. In: IEEE international conference on Big Data, IEEE, pp. 21–26
19. Cox MP, Peterson DA, Biggs PJ (2010) Solexaqa: at-a-glance quality assessment of illumina second-generation sequencing data. BMC Bioinformatics 11:485
20. Huse SM, Huber JA, Morrison HG et al (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8(7):R143
21. Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. ISME J 3(11):1314–1317
22. Keegan KP, Trimble WL, Wilkening J et al (2012) A platform-independent method for detecting errors in metagenomic sequencing data: drisee. PLoS Comput Biol 8(6), e1002541
23. Langmead B, Trapnell C, Pop M et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25

24. Trimble WL, Keegan KP, D'Souza M et al (2012) Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. BMC Bioinformatics 13(1):183

25. Rho M, Tang H, Ye Y (2009) Fraggenescan: predicting genes in short and error prone reads. Nucleic Acids Res 38(20), e191

26. Edgar RC (2010) Search and clustering orders of magnitude faster than blast. Bioinformatics 26(19):2460–2461

27. Caporaso JG, Kuczynski J, Stombaugh J et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7(5):335–336

28. Huson DH, Auch AF, Qi J et al (2007) Megan analysis of metagenomic data. Genome Res 17 (3):377–86

29. Aziz R, Bartels B, Best A et al (2008) The RAST server: rapid annotations using subsystems technology. BMC Genomics 9(1):75

30. Pruesse E, Quast C, Knittel K 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):7188–7196

31. DeSantis TZ, Hugenholtz P, Larsen N et al (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72 (7):5069–5072

32. Cole JR, Chai B, Marsh TL et al (2003) The ribosomal database project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucleic Acids Res 31(1):442–443

33. Yilmaz P, Kottmann R, Field D et al (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 29(5):415–420

34. Bolotin A, Quinquis B, Sorokin A et al (2005) Clustered regularly interspaced short palindrome repeats (CRISPRS) have spacers of extrachromosomal origin. Microbiology 151 (Pt 8):2551–2561

35. Reeder J, Knight R (2009) The 'rare biosphere': a reality check. Nat Methods 6 (9):636–637

36. Ondov BD, Bergman NH, Phillippy AM (2011) z. BMC Bioinformatics 12:385