Chapter 13

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

Kevin P. Keegan, Elizabeth M. Glass, and Folker Meyer

Abstract

Approaches in molecular biology, particularly those that deal with high-throughput sequencing of entire microbial communities (the field of metagenomics), are rapidly advancing our understanding of the composition and functional content of microbial communities involved in climate change, environmental pollution, human health, biotechnology, etc. Metagenomics provides researchers with the most complete picture of the taxonomic (i.e., *what organisms are there*) and functional (i.e., *what are those organisms doing*) composition of natively sampled microbial communities, making it possible to perform investigations that include organisms that were previously intractable to laboratory-controlled culturing; currently, these constitute the vast majority of all microbes on the planet. All organisms contained in environmental samples are sequenced in a culture-independent manner, most often with 16S ribosomal amplicon methods to investigate the taxonomic or whole-genome shotgun-based methods to investigate the functional content of sampled communities. Metagenomics allows researchers to characterize the community composition and functional content of microbial communities, but it cannot show which functional processes are active; however, near parallel developments in transcriptomics promise a dramatic increase in our knowledge in this area as well.

Since 2008, MG-RAST (Meyer et al., BMC Bioinformatics 9:386, 2008) has served as a public resource for annotation and analysis of metagenomic sequence data, providing a repository that currently houses more than 150,000 data sets (containing 60+ tera-base-pairs) with more than 23,000 publically available. MG-RAST, or the metagenomics RAST (rapid annotation using subsystems technology) server makes it possible for users to upload raw metagenomic sequence data in (preferably) fastq or fasta format. Assessments of sequence quality, annotation with respect to multiple reference databases, are performed automatically with minimal input from the user (*see* Subheading 4 at the end of this chapter for more details). Post-annotation analysis and visualization are also possible, directly through the web interface, or with tools like matR (metagenomic analysis tools for R, covered later in this chapter) that utilize the MG-RAST API (http://api.metagenomics.anl.gov/api.html) to easily download data from any stage in the MG-RAST processing pipeline. Over the years, MG-RAST has undergone substantial revisions to keep pace with the dramatic growth in the number, size, and types of sequence data that accompany constantly evolving developments in metagenomics and related -omic sciences (e.g., metatranscriptomics).

Key words Metagenomics, Comparative analysis, Sequence quality, Automated pipeline, High-throughput, matR

Francis Martin and Stéphane Uroz (eds.), *Microbial Environmental Genomics (MEG)*, Methods in Molecular Biology, vol. 1399, DOI 10.1007/978-1-4939-3369-3_13, © Springer Science+Business Media New York 2016

1 Introduction

Developing approaches in molecular biology are rapidly advancing our understanding of the composition and functional content of microbial communities. This has led to a much clearer picture of the pivotal role these communities play in phenomena as diverse as climate change, environmental pollution, human health, and developments in biotechnology. Metagenomics utilizes cutting-edge technology in sequencing and sequence characterization to create an inventory of microbial genes that are present in any given environment, including those contained in microbes intransigent to classical culture-based methods; currently, these constitute the vast majority (estimates typically report 99 % or more) of all microbes on the planet. All organisms contained in native microbial communities (also referred to as assemblages) are sequenced in a culture-independent manner, most often with 16S ribosomal amplicon methods to investigate the taxonomic or whole-genome shotgun-based methods to investigate the functional content of sampled communities. This makes it possible to create a much clearer picture of the composition (*who is there*) and potential functional content (*what can they do*) of microbial communities than was possible with previous methods. Metatranscriptomics extends this knowledge by providing us with a catalog that can link functions active in a community (*what are they doing*) to the temporal and conditional variables that drive them (*why are they doing …*).

For these kinds of sequence-dependent studies, the underlying quality of sequence data is a fundamental concern, complicated by the use of an ever-expanding assortment of methods, equipment, and software. Metagenomic analyses rely on the use of highly automated analysis tools; therefore, early identification of quality-related problems is essential to avoid wasteful use of limited computational resources as well as interpretation of fundamentally flawed data that can lead to erroneous biological inferences.

In regards to sequence quality, the scientific community faces another hurdle with the study of metagenomics data. Most researchers understand how critical it is for sequence data to exhibit the highest possible quality—especially in applications where a high level of functional or taxonomic resolution is desired—but do not possess the technical expertise to assess quality (i.e., independently from metrics provided by black-box vendor-specific software and/or sequencing centers that may not be using the most current or best practices). MG-RAST possesses multiple features that make it easy for users to assess sequence quality and address some of the most common issues (e.g., high error rates, contamination with adapter sequences, contamination with artificial duplicate reads, etc.).

In recent years, the sequencing costs have dramatically reduced whereas costs of computing have remained relatively stable. This has shifted the limiting factor in sequence-dependent investigations from data generation (i.e., sequencing) to data analysis (annotation and post-annotation analyses). Wilkening et al. [1] provide a real currency cost for the analysis of 100 giga-base-pairs of DNA sequence data using BLASTX on Amazon's EC2 service, $300,000. A more recent study by 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. As the size and number of sequence data sets continue to increase, costs related to their analysis will continue to rise.

In addition, metadata (data describing data—e.g., data that describe the temporal and environmental parameters for a sampled microbial community) provide an essential complement to experimental data, helping to answer questions about its source, mode of collection, and reliability as well as making it possible answer meaningful biological questions (e.g., what factor(s) cause a shift in the composition or functional content of a microbial community in a particular environment). 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] has served as both a repository and tool for the analysis of metagenomic data (and metadata)—annotation and post-annotation analyses. Currently, the system has analyzed more than 60 tera-base-pairs of data from more than 150,000 data sets, with more than 23,000 available to the public. Over the years, MG-RAST has undergone substantial revisions to keep pace with the dramatic growth in the number, size, and types of sequence data that accompany constantly evolving developments in metagenomics and related -omic sciences (e.g., metatranscriptomics). These include innovations in engineering as well as modifications to our 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 non-redundant protein database) [5] for annotation. The M5nr is an integration of many sequence databases into one single, indexed, searchable database. A single similarity search (using BLAST [6] or BLAT [7]) allows 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]

Computation of sequence similarities is becoming a limiting factor in metagenomic analyses. Sequence similarity search results encoded in an open, exchangeable format distributed with the sequence sets have the potential to limit the needs for computational re-analysis of data sets. A prerequisite for sharing of similarity results is a common reference—this is exactly what the M5nr provides, a commonly indexed database that contains all of databases noted above.

M5nr mechanisms are used for automatically maintaining this comprehensive non-redundant protein database and creating a quarterly release of this resource. In addition, MG-RAST provides tools for translating similarity searches into many namespaces, e.g., KEGG, NOG, SEED Subsystems, NCBI's GenBank, etc.

2.2 Analysis Pipeline

The pipeline shown in Fig. 1 contains a number of improvements to previous MG-RAST versions. Several key algorithmic improvements were needed to support the flood of user-generated data. Initial analysis differentiates amplicon-based ribosomal 16S from whole-genome shotgun (WGS) samples and processed them separately (*see* Subheading 3.1 below for processing of WGS data and Subheading 3.2 for processing of amplicon 26s data). WGS samples are analyzed with dedicated software to perform gene prediction on nucleotide data prior to protein similarity-based annotation. This provides drastic runtime improvements over nucleotide similarity-based approaches. Clustering of predicted proteins at 90 % identity provides additional performance improvement while preserving biological signals. While protein-based annotation is used for proteins predicted from WGS samples, samples detected as 16S ribosomal data are annotated with nucleotide-based similarity.

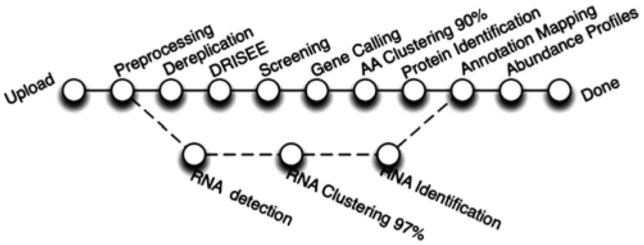


Fig. 1 Details of the analysis pipeline for MG-RAST. After upload, the pipeline diverges for amplicon and WGS data sets. Amplicon samples run through RNA detection, clustering, and identification. WGS samples undergo a number of additional processing steps to assess data quality prior to annotation

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

2.3 Compute Resources

While MG-RAST was originally built as a traditional, centrally located, cluster-based bioinformatics system, the most recent version embraces novel technologies that make it possible for it to utilize local and remote compute resources. MG-RAST data are stored in SHOCK [18] and computing is orchestrated by AWE [19]. 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 provided by individual users, and a number of traditional clusters. An installation of the pipeline exists at DOE's NERSC supercomputing center. In recent months, the system handles over 4 terabase-pairs of data per month. The use of Skyport [38] has enabled multi cloud workflows without introducing management overhead.

3 Methods

The pipeline diverges after upload for 16S ribosomal 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 functional abundance profiles. rRNA samples run through RNA detection, clustering, and identification, and the production of taxonomic abundance profiles. Subheading 4 found at the end of this chapter includes additional details.

3.1 The WGS Pipeline

1. *Preprocessing*. After upload, data are preprocessed by using SolexaQA [20] 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 [21]. All sequences submitted to the system are available, but discarded reads are not analyzed further.

2. *Dereplication*. For shotgun metagenome and shotgun metatranscriptome data sets, 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) [22]. Instead of simply discarding the ADRs, we set them aside and use them later as a means to assess sample quality. We note that dereplication is not suitable for amplicon data sets that are likely to share common prefixes.
3. *DRISEE*. MG-RAST v3 uses DRISEE (Duplicate Read Inferred Sequencing Error Estimation) [23] to analyze the sets of ADRs and determine the degree of variation among prefix-identical sequences derived from the same template. See below for details.
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 [24] (a fast, memory-efficient, short read aligner), and only reads that do not match the model organisms pass into the next stage of the annotation pipeline.

   Note that 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.
5. *Gene calling*. The previous version of MG-RAST used nucleotide-based similarity for annotation of WGS data, an approach that is significantly more expensive computationally than de novo gene prediction followed by protein similarity-based annotation. After an in-depth investigation of tool performance [25], we have moved to a machine learning approach that utilizes FragGeneScan [26] to predict proteins/protein fragments from de novo sequence data (FragGeneScan uses a well tested algorithm [25] to perform in silico translation of predicted protein coding nucleic acid sequences). Utilizing this approach, we can predict coding regions in DNA sequences that are 75 base pairs or longer. Our novel approach also enables the analysis of user-provided assembled contigs. We note that FragGeneScan is trained for prokaryotes only. While it will identify proteins for eukaryotic sequences, the results should be viewed critically.
6. *AA clustering*. MG-RAST builds clusters of proteins at the 90 % identity level using the uclust [27] implementation in QIIME [28], 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.

7. *Protein identification*. Once created, a representative (the longest sequence) for each cluster is subjected to similarity analysis. Functional identification of representative sequences does not use BLAST, instead we use a much more efficient algorithm, sBLAT, an implementation of the BLAT algorithm, which we parallelized using OpenMPI. We reconstruct the putative species composition of WGS data by looking at the phylogenetic origin of the database sequences hit by the protein-based similarity searches. Note that processing of rRNA 16S amplicon data is covered in Subheading 3.2 below.

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 recomputation. For example, COG and KEGG views can be displayed, which both show the relative abundances of histidine biosynthesis in a data set 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 for end users to determine the number of hits (occurrences of the input sequence in the database) in their samples.

   - *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 (e.g., a single feature may have 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 is better suited for comparisons of functional and taxonomic profiles of multiple 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 [29]. 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.

9. *Abundance profiles.* Abundance profiles (essentially tables that indicate detected taxa or functions and their relative abundance as determined by the methods described in Subheading 3.1, **step 8**—examples can be found in the MG-RAST user manual, *see* the "additional documentation" in Subheading 4 found at the end of this chapter) are the primary data product that the MG-RAST's user interface uses to display information in annotated data sets. Using the abundance profiles, the MG-RAST system defers to the user to select several parameters that will define their abundance data, *e*-value, percent identity, and minimal alignment length. As it is not possible to arbitrarily select thresholds suitable for all use cases, users can select their own thresholds for each of these values.

Taxonomic profiles use the NCBI taxonomy. All taxonomic information is projected against the NCBI taxonomy.

Functional profiles are available for data sources that provide hierarchical information. These currently include SEED subsystems, KEGG orthologs, and COGs. SEED subsystems represent an independent reannotation effort utilized by RAST [30] and MG-RAST. Manual curation of subsystems makes them an extremely valuable data source. The current subsystems hierarchy can be viewed at http://pubseed.theseed.org//SubsysEditor.cgi which allows browsing the subsystems.

Subsystems represent a four-level hierarchy,

1. Subsystem level 1—highest level
2. Subsystem level 2—intermediate level
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 to implement a four-level hierarchy. We note that KEGG data are no longer available for free download; therefore, we rely on the latest freely downloadable version of these 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

The rRNA pipeline starts with upload of rRNA reads and proceeds through the following steps:

1. *rRNA detection*. Reads are identified as rRNA through a simple rRNA detection. 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 differentiate samples containing solely rRNA data from other samples (e.g., WGS or transcriptomic samples).
2. *rRNA clustering*. The rRNA-similar reads are clustered at 97 % identity, and the longest sequence is picked as the cluster representative.
3. *rRNA identification*. A nucleotide BLAT similarity search for the longest cluster representative is performed against the M5rna database, integrating SILVA [31], Greengenes [32], and RDP [33].

3.3 Using the MG-RAST User Interface

The MG-RAST system provides a rich web user interface that covers all aspects of metagenome analysis, from data upload to ordination analysis of annotation abundances. The web interface can also be used for data discovery. Metagenomic data sets 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 may 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, principal component analyses, or raw abundance tables exported in tabular form. The raw or processed data can be recovered via download pages or with the matR package for R (*see* Subheading 4 below). 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 [34] (as well as domain-specific plug-ins for specialized environments not extending the minimal GSC standards); several projects, including TerraGenome, HMP, TARA, and EMP, use these GSC standards, enabling standardized queries that integrate new samples into these massive data sets.

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 recomputation of results.

3.3.1 Navigation

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

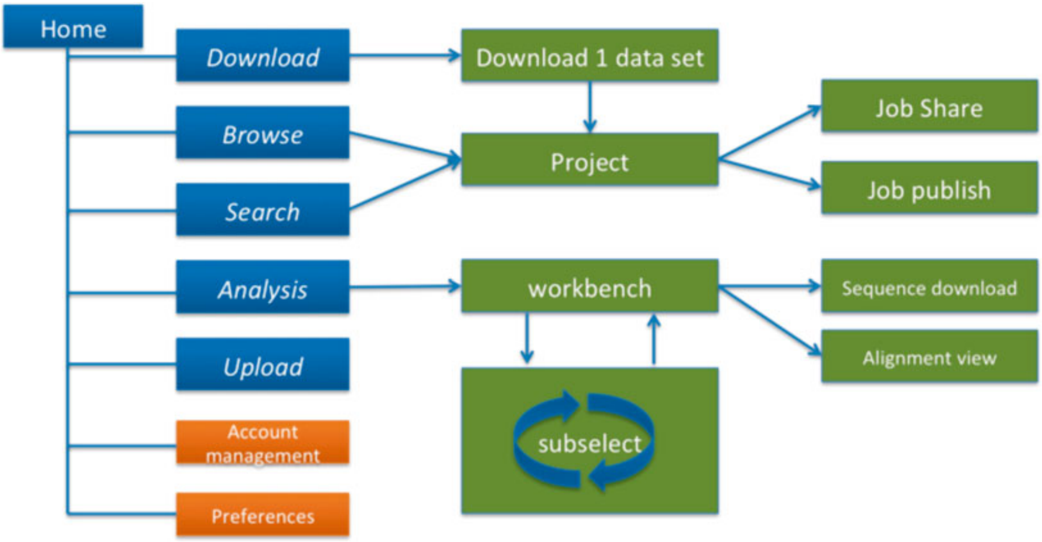


Fig. 2 Sitemap for the MG-RAST version 3 website. On the site map the main pages are shown in *blue*, management pages in *orange*. The *green boxes* represent pages that are not directly accessible from the home page

- Download page—lists all publicly available data for download. The data are structured into projects.
- Browse page—allows interactive browsing of all data sets 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 data sets.
- Upload page—allows users to provide their samples and metadata to MG-RAST.
- Home (Metagene Overview) page—provides an overview for each individual data set.

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 large files and also allows parallel uploads. MG-RAST supports data sets 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 data sets of over 100 giga-base-pairs, often resulting in file sizes in excess of 150 GB.

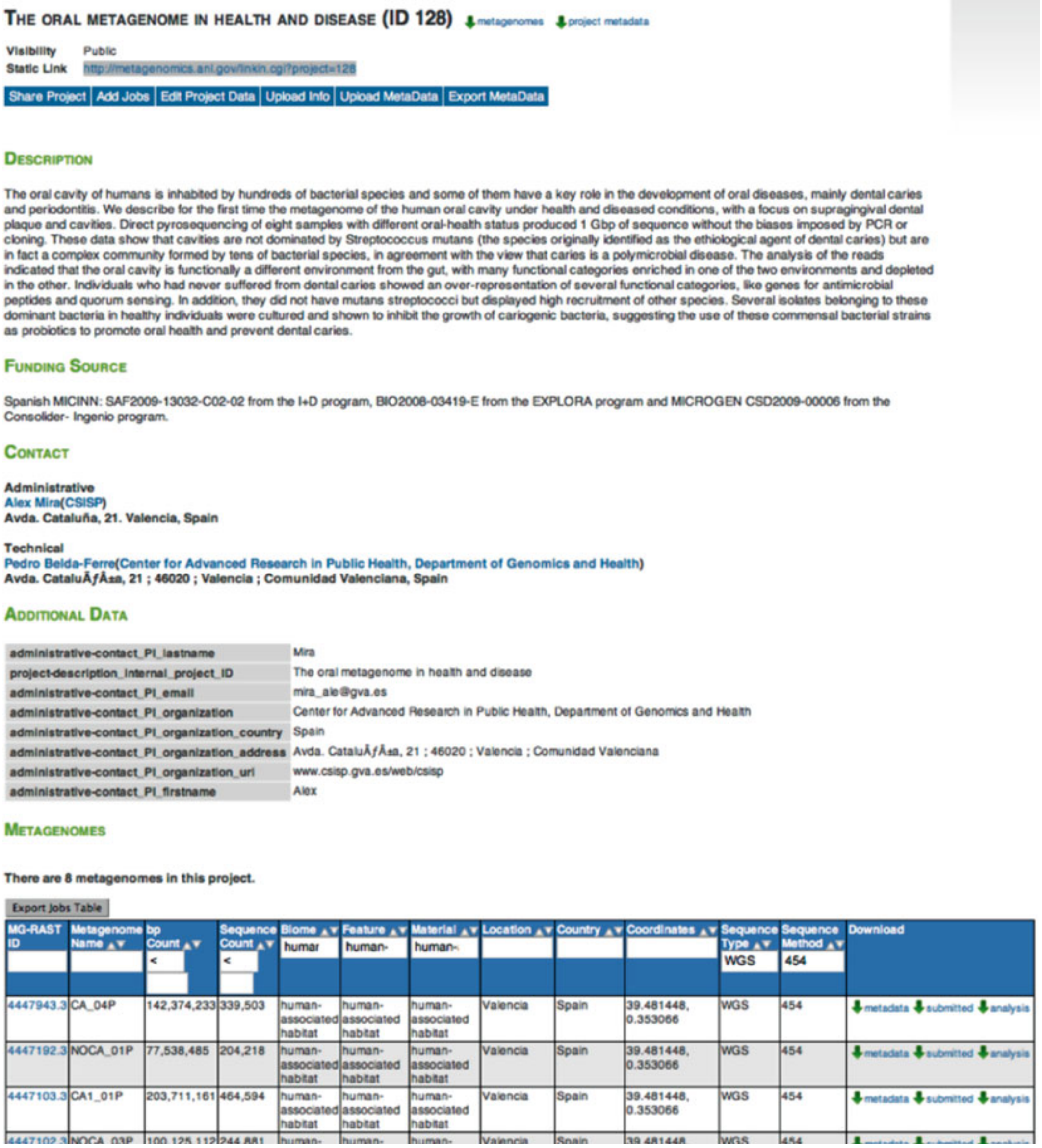
3.3.3 Browse Page: Metadata-Enabled Data Discovery

The Browse page lists all data sets visible to the user (the users own data sets as well as all public data and all data shared by other users). This page also provides an overview of the nonpublic data sets 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 data set size) or metadata (e.g., location or biome).

3.3.4 Project Page

The project page provides a list of data sets 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 data set owners, the Project page provides an editing capability using a number of menu entries at the top of the page. Figure 3 shows the available options.

- Share Project—make the data in this project available to third parties via sending them access tokens.
- Add Jobs—add additional data sets to this project.



THE ORAL METAGENOME IN HEALTH AND DISEASE (ID 128) metagenomes project metadata

Visibility Public
Static Link http://metagenomics.anl.gov/linkin.cgi?project=128

DESCRIPTION

The oral cavity of humans is inhabited by hundreds of bacterial species and some of them have a key role in the development of oral diseases, mainly dental caries and periodontitis. We describe for the first time the metagenome of the human oral cavity under health and diseased conditions, with a focus on supragingival dental plaque and cavities. Direct pyrosequencing of eight samples with different oral-health status produced 1 Gbp of sequence without the biases imposed by PCR or cloning. These data show that cavities are not dominated by Streptococcus mutans (the species originally identified as the ethiological agent of dental caries) but are in fact a complex community formed by tens of bacterial species, in agreement with the view that caries is a polymicrobial disease. The analysis of the reads indicated that the oral cavity is functionally a different environment from the gut, with many functional categories enriched in one of the two environments and depleted in the other. Individuals who had never suffered from dental caries showed an over-representation of several functional categories, like genes for antimicrobial peptides and quorum sensing. In addition, they did not have mutans streptococci but displayed high recruitment of other species. Several isolates belonging to these dominant bacteria in healthy individuals were cultured and shown to inhibit the growth of cariogenic bacteria, suggesting the use of these commensal bacterial strains as probiotics to promote oral health and prevent dental caries.

FUNDING SOURCE

Spanish MICINN: SAF2009-13032-C02-02 from the I+D program, BIO2008-03419-E from the EXPLORA program and MICROGEN CSD2009-00006 from the Consolider- Ingenio program.

CONTACT

Administrative
Alex Mira(CSISP)
Avda. Cataluña, 21. Valencia, Spain

Technical
Pedro Belda-Ferre(Center for Advanced Research in Public Health, Department of Genomics and Health)
Avda. CataluÃƒÂ±a, 21 ; 46020 ; Valencia ; Comunidad Valenciana, Spain

ADDITIONAL DATA

administrative-contact_PI_lastname	Mira
project-description_internal_project_ID	The oral metagenome in health and disease
administrative-contact_PI_email	mira_ale@gva.es
administrative-contact_PI_organization	Center for Advanced Research in Public Health, Department of Genomics and Health
administrative-contact_PI_organization_country	Spain
administrative-contact_PI_organization_address	Avda. CataluÃƒÂ±a, 21 ; 46020 ; Valencia ; Comunidad Valenciana
administrative-contact_PI_organization_url	www.csisp.gva.es/web/csisp
administrative-contact_PI_firstname	Alex

METAGENOMES

There are 8 metagenomes in this project.

MG-RAST ID	Metagenome Name	bp Count	Sequence Count	Biome	Feature	Material	Location	Country	Coordinates	Sequence Type	Sequence Method	Download
		<	<	humar	human-	human-				WGS	454	
4447943.3	CA_04P	142,374,233	339,503	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	metadata submitted analysis
4447192.3	NOCA_01P	77,538,485	204,218	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	metadata submitted analysis
4447103.3	CA1_01P	203,711,161	464,594	human-associated habitat	human-associated habitat	human-associated habitat	Valencia	Spain	39.481448, 0.353066	WGS	454	metadata submitted analysis
4447102.3	NOCA_03P	100,125,112	244,881	human-	human-	human-	Valencia	Spain	39.481448	WGS	454	metadata submitted analysis

Fig. 3 Project page, providing a summary of all data in the project and an interface for downloads

- 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 data set. This metagenome overview page provides a summary of the annotations for a single data set. 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 data set. It provides information regarding technical detail 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 data sets 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 data set, a unique identifier that is usable as an accession number for publications. Additional information, such as the name of the submitting PI, organization, and a user-provided metagenome name are displayed at the top of the page. A static URL for linking to the system that will be stable across changes to the MG-RAST web interface is provided as additional information (Fig. 7).

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. 4) 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.

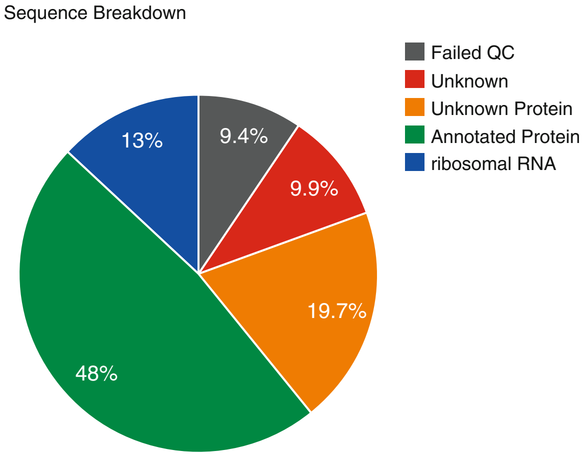


Fig. 4 Sequences to the pipeline are classified into one of five categories: *grey* = failed the QC, *red* = unknown sequences, *yellow* = unknown function but protein coding, *green* = protein coding with known function, and *blue* = ribosomal RNA. For this example, over 50 % of sequences were either 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 [35]) 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 data set to the extent that such information has been made available. Metadata enables other researchers to discover data sets 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 data set 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 are 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 the percentage of the predicted protein features annotated with similarity to a protein of known function per source database. In addition, ribosomal RNA genes are mapped to the rRNA databases.

In addition, this display will print 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. 5) 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 is 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 [36], 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 (*see* Fig. 6).

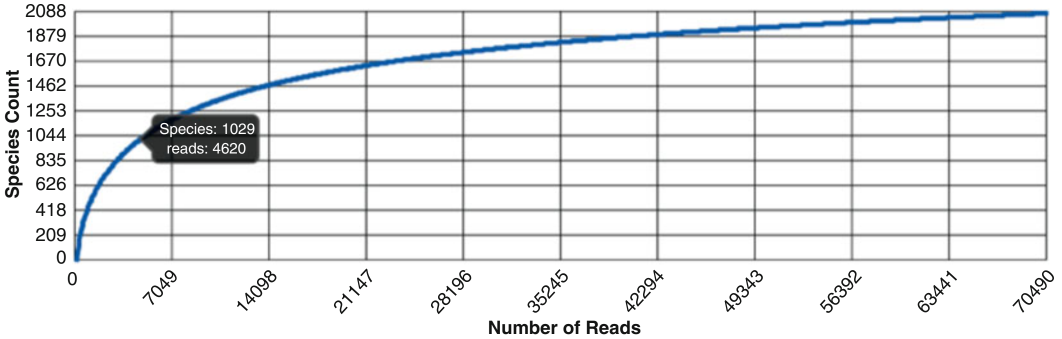


Fig. 5 Rarefaction plot showing a curve of annotated species richness (i.e., the number of unique species). This curve is a plot of the total number of distinct species annotations as a function of the number of sequences sampled

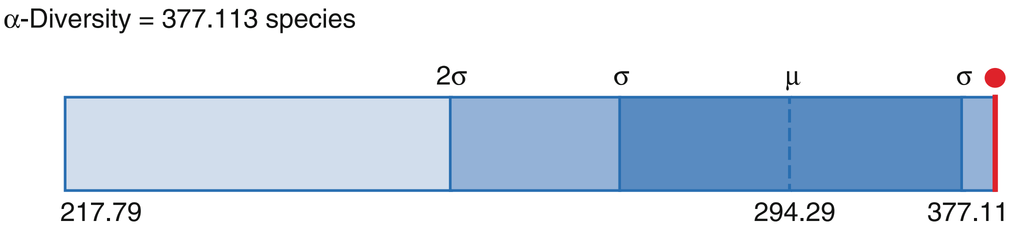


Fig. 6 Alpha diversity plot showing the range of -diversity values in the project the data set belongs to. The min, max, and mean values are shown, with the standard deviation ranges in different shades. The alpha-diversity of this metagenome is shown in *red*. The species-level annotations are from all the annotation source databases used by MG-RAST

The alpha diversity estimate is a single number that summarizes the distribution of species-level annotations in a data set. 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.

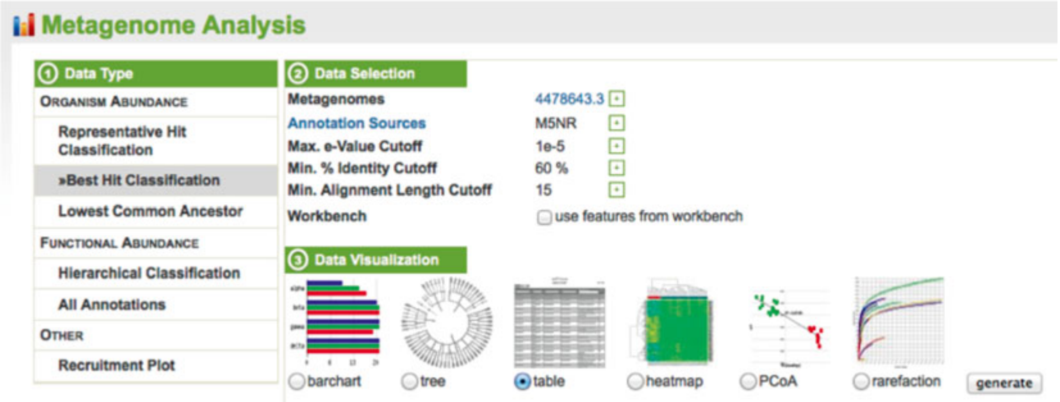
A more detailed functional analysis, allowing the user to manipulate parameters for sequence similarity matches, is available from the Analysis page.

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 (Fig. 7).

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.



Metagenome Analysis

1 Data Type

Organism Abundance
- Representative Hit Classification
- »Best Hit Classification
- Lowest Common Ancestor

Functional Abundance
- Hierarchical Classification
- All Annotations

Other
- Recruitment Plot

2 Data Selection

Metagenomes	4478643.3
Annotation Sources	M5NR
Max. e-Value Cutoff	1e-5
Min. % Identity Cutoff	60 %
Min. Alignment Length Cutoff	15
Workbench	□ use features from workbench

3 Data Visualization

○ barchart ○ tree ● table ○ heatmap ○ PCoA ○ rarefaction generate

Fig. 7 Three-step process in using the Analysis page: (*1*) select a profile and hit (see text) type; (*2*) select a list of metagenomes and set annotation source and similarity parameters; (*3*) choose a comparison

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
   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 barcharts, trees, spreadsheet-like tables, heatmaps, PCoA, rarefaction plots, circular recruitment plot, and KEGG maps.

The data selection dialog provides access to data sets in four ways. The four categories can be selected from a pulldown menu.

- private data—list of private or shared data sets 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 data sets grouped by the submitting user. The project name will be displayed.
- public data—display of all public data sets.

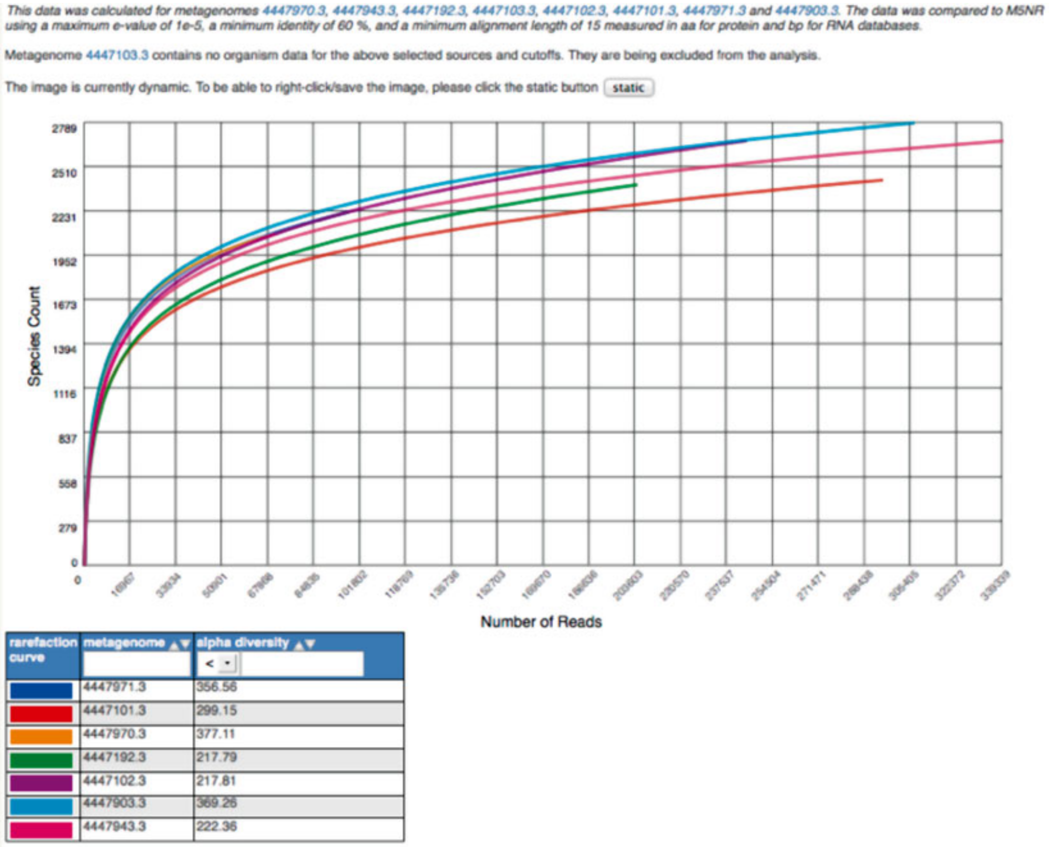
When using collections or projects, data can also be grouped into one set per collection or project and subsequently compared or added.

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 are transformed with a log transformation, the values exhibit a normal distribution. Log transformation of the counts 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 inter-sample 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. 8)



This data was calculated for metagenomes 4447970.3, 4447943.3, 4447192.3, 4447103.3, 4447102.3, 4447101.3, 4447971.3 and 4447903.3. The data was compared to M5NR using a maximum e-value of 1e-5, a minimum identity of 60 %, and a minimum alignment length of 15 measured in aa for protein and bp for RNA databases.

Metagenome 4447103.3 contains no organism data for the above selected sources and cutoffs. They are being excluded from the analysis.

The image is currently dynamic. To be able to right-click/save the image, please click the static button static

rarefaction curve	metagenome ▲▼	alpha diversity ▲▼ < ▾
🟦	4447971.3	356.56
🟥	4447101.3	299.15
🟧	4447970.3	377.11
🟩	4447192.3	217.79
🟪	4447102.3	217.81
🟦	4447903.3	369.26
🟥	4447943.3	222.36

Fig. 8 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

of the total number of distinct species annotations as a function of the number of sequences sampled. As shown in the figure, multiple data sets can be included.

The slope of the right-hand part of the curve is related to the fraction of sampled species that are rare. When the rarefaction curve is flat, more intensive sampling is likely to yield only a few additional species. 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 plot serves as a means of comparing species richness between samples in a way independent of the sampling depth. 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 is sampled, more intensive sampling is likely to yield only a few additional species. Sampling curves generally rise very quickly at first and then level off toward an asymptote as fewer new species are found per unit of individuals collected. These 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 data set.

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 Principle Coordinate Analysis (PCoA) to reduce the dimensionality of comparisons of multiple samples that consider functional or taxonomic annotations. Dimensionality reduction is a process that allows the complex variation found in a large data sets (e.g., the abundance values of thousands of functional roles or annotated species across dozens of metagenomic samples) to be reduced to a much smaller number of variables that can be visualized as simple two or three-dimensional scatter plots. The plots enable interpretation of the multidimensional data in a human-friendly presentation. Samples that exhibit similar abundance profiles (taxonomic or functional) group together, whereas those that differ are found farther apart.

A key feature of PCoA-based analyses is that users can compare components not just to each other but 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 sub-selection 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 data sets against a hierarchy (e.g., Subsystems or the NCBI taxonomy). The hierarchy is displayed as a rooted tree, and the abundance (normalized for data set size or raw) for each data set 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 [37] or can be exported in BIOM [24] 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 exist 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 data set publication. Metazen [39] is a web based tool for assisting end-users in the creation of metadata with the correct controlled vocabularies and in the correct format.

In addition to publishing, data and analysis can also be shared with specific users (Fig. 9). To share data, users simply enter their email address via clicking sharing on the Overview page.

4 matR, Metagenomic analysis tools for R

We have recently produced a package for the R environment for statistical computing (www.r-project.org/) that provides accessory analytical capabilities to complement those already available through the MG-RAST website. The matR package is primarily designed for download and analysis of MG-RAST-based annotation abundance profiles. It makes it possible to download annotation abundance data from MG-RAST into R friendly data objects suitable for analysis with included analysis functions. We note that matR has been specifically designed to perform large-scale analyses on abundance profiles from dozens to thousands of data sets with suitable pre-processing, normalization, statistics, and visualization

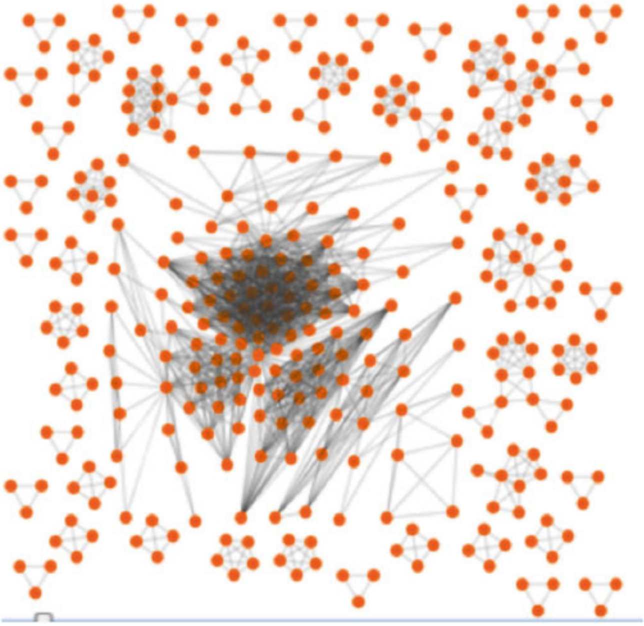


Fig. 9 Data sets shared in MG-RAST by users (*orange dots*), shown as connecting edges

tools. Users can utilize these built-in tools or any of the enormous variety of tools available within the R universe. The release version of matR is available through CRAN (http://cran.r-project.org/web/packages/matR/); pre-release and development versions are available on github (https://github.com/MG-RAST/matR/); a google group is available (https://groups.google.com/forum/#!forum/matr-forum); a publication demonstrating the ease with which matR can be used to conduct large-scale analyses is forthcoming.

4 Notes

Typical analysis parameters

MG-RAST utilizes a number of tools and analyses to generate annotation abundance data, and subsequent visualizations of abundance data, from raw sequence data. Users have the option to vary several of the parameters that define several aspects of how MG-RAST performs. While the default settings have been selected to perform *well* in most circumstances, it is not possible to find a single collection of analysis parameter values that will perform optimally on all data submitted to MG-RAST. Here we briefly mention some of the most important parameters, discussion when a user may want to alter the default values, and how they can go about selecting the best values for their analyses in a methodological fashion. We divide these parameters into two sections—pipeline options that users must be specific before their data are processed through MG-RAST and data options that define the annotation abundance data that are returned to the user,

Pipeline options

Pipeline options are specified by the user during upload and prior to annotations with MG-RAST. These options are used to filter data with a number of metrics that characterize the quality of each individual read in a sample. Users can modify these key parameters from their default values,

Assembled (NOT checked by default)

Select this option if your data are the product of any assembly-based tools applied to the data before upload to MG-RAST. *Note that we recommend upload of raw reads with their accompanying quality data (i.e., fastq files)*. This allows MG-RAST to directly sequence QC information when processing reads. When assembled data are used, a great deal of abundance information is lost; with assembled data MG-RAST can only provide abundance that indicates the number of times a feature is observed in the assembly. Relative abundance information contained in the original reads is lost. In addition, assembly might introduce chimeric artifacts that will confound subsequent annotation and analysis. Users should select this option if their data have undergone any assembly prior to upload.

Dereplication (CHECKED by default)

A well-known artifact in NGS sequencing is the production of artificial replicate sequences. These are identical (or nearly identical) reads that occur with extremely high abundances (see http://www.nature.com/ismej/journal/v3/n11/full/ismej200972a.html). While the exact causes for such sequences are not well understood, it has been posited that they are due in part to inclusion of low complexity and/or adapter sequences (sequences ligated onto reads to facilitate processing with NGS that SHOULD NEVER appear in output). Artificial duplicates are utilized by MG-RAST as the basis for DRISEE-based error estimates (http://www.ncbi.nlm.nih.gov/pubmed/22685393); we maintain that the inclusion of such sequences constitutes a clear sequencing error/artifact, but this is a contested notion that remains to be resolved (http://www.ncbi.nlm.nih.gov/pubmed/23698723). However, what metagenomicists can agree on is that such sequences appear frequently, when they are not expected, and are present even after raw sequence data have been treated with vendor-specific tools to remove them. Dereplication should always be turned on for WGS sequencing—but, users may want to deselect this option for amplicon-based data (if reads start with a high conserved region, they could be misinterpreted as artificial replicates) or any other data sets where an extremely high level of replication among reads is expected.

Screening (set to H. sapiens, NCBI v36 by default)

It is common for metagenomic NGS data to contain contaminant sequences from an undesired organism (e.g., human sequences in a human gut sampled microbiome). Screening makes it possible for these sequences to be identified and removed from subsequent analysis. Users have a number of other organisms that can be used to screen the data. Currently, MG-RAST supports filtering against a single contaminant organism. Users can select the most appropriate organism, or select "none." If a user wants to maintain all sequence data, they should select "none." Note that MG-RAST is not designed to annotate eukaryotic data, screening is used as a means to remove such data (assumed to be host sequence data) from a number of known sources. Note that the currently collection of organisms against which screening is possible can be expanded. Users should contact MG-RAST if they want add additional organisms to those that can be screened.

Dynamic Trimming (CHECKED by default with 15 as the minimum retained phred score and 5 as the maximum number of bases allowed in each read that can contain a base lower than the phred minimum)

Dynamic trimming is only possible if users have uploaded data in fastq format that contains sequence quality information. Dynamic trimming is used in place of length filtering and ambiguous-based filtering when a fastq is uploaded. More stringent values (e.g., higher phred, and lower allowance for base that do not meet the phred threshold) will reduce the length and amount of reads that MG-RAST processes, but can increase their overall quality. Users should increase stringency if they want to reduce annotated data to more constrained, higher quality sequences. We do not recommend reducing stringency; this will lead to the inclusion of low-quality data that will most likely produce no or extremely unreliable annotation.

Length filtering (CHECKED by default, with a standard deviation multiplier set to 2; only applies to fasta data; it is not used on fastq data)

Only applied to fasta data, or fastq data with no quality information (essentially, a surrogate for the dynamic trimming applied to fastq data that include quality information), length trimming calculates the average sequence length for all reads in a data set and removes those that are,

Longer than than sample_mean + (standard_deviation_multiplier * standard_deviation)

or shorter than sample_mean − (standard_deviation_multiplier * standard_deviation)

This is an attempt to remove sequences that exhibit outlier lengths and are likely to be sequencing artifacts.

For fasta data we recommend that users always use length filtering, and that they do not use a standard deviation multiplier less than the default of 2. Users may want to increase the standard deviation multiplier if their reads exhibit a large degree of variation with respect to read length.

Ambiguous base filtering (CHECKED by default with a maximum number of allowed ambiguous bases per read set to 5; only applies to fasta data or fastq without quality information)

Sequences frequently produce “ambiguous bases” (bases that are not A, T, C, or G), these represent bases for which the sequences could not make a definitive call for the identity of the base. Ambiguous bases are expected at the end of sequenced reads, and are generally considered to be an indication of low quality if they are found anywhere else in the read, particularly at the start. MG-RAST will reject any read that contains more than the specified number of ambiguous bases. We recommend that users should not use values less stringent (i.e., larger) than the default of 5. Users are free to specify more stringent criteria (smaller number of allowed ambiguous bases)—this will reduce the number of annotated reads, but will produce a set of reads that have a higher overall quality (less likely to contain artifacts that could lead to erroneous annotations).

Data options

After processing through MG-RAST, users have a number of options that they can use to filter annotation abundance data (accessed via the). Chief among these are the parameters described briefly below,

Annotation source (default m5NR)

MG-RAST provides users with the unique ability to provide annotations for analyzed data from multiple source databases. By default, the m5NR is used—but users are free to use any one of the numerous additional annotation sources. As the m5NR represents a non-redundant union of all annotation databases contained within MG-RAST, we generally recommend its use over any of the individual databases.

Max e-value cutoff (default 1e–5)

The max *e*-value cutoff indicates the largest (least stringent) *e*-value for an annotation to be included in the output annotations. We generally recommend that users not use larger (less stringent) *e*-value cutoff. The use of smaller (more stringent) *e*-values will ensure that annotations exhibit higher statistical fidelity; however, this will come at the cost of a smaller overall number of annotated features. We suggest that users experiment with multiple *e*-values until they arrive at one that produces enough annotations to address the hypothesis(es) in question while minimizing the number of spurious (false positive) annotations.

Min % identity cutoff (default 60 %)

The minimum percent identity represents the lower bound threshold for annotations to be returned to the user. Matches between the query and selected annotation that match or exceed this value are retained, all others are rejected. We recommend that users not select a value any lower than the default. Users may choose larger values to return annotations only if they meet more stringent match criteria. An increase in the minimum percent identity cutoff will produce annotations that exhibit closer matches to the reference database at the cost of a lower overall number of annotations. Once again, we encourage users to experiment with this threshold until it produce the desired number of annotations at an acceptable level of stringency.

Min alignment length cutoff (default 15)

The minimum length cutoff represent the lower bound threshold for alignment lengths to be included in output annotations. As with the *e*-value and percent identity cutoffs, we recommend that users do not attempt to select a less stringent (smaller) value. They can select larger (more stringent) values to produce a smaller set of longer matches (generally considered to be synonymous with higher quality and increased accurate). Once again, users need to experiment with this value to find an optimal balance between stringency and number of annotations.

Additional Documentation

This chapter is intended as an introduction to MG-RAST, and necessarily treats topics as concisely as possible; other topics were omitted entirely for the sake of brevity. Users interested in a much more detailed description of MG-RAST are referred to the MG-RAST user manual. The manual can be downloaded from metagenomics.anl.gov/; simply click on the link for the *MG-RAST manual*.

References

1. Wilkening J, Wilke A, Desai N et al (2009) Using clouds for metagenomics. A case study. In: IEEE Cluster, 2009
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 9:386
4. Field D, Amaral-Zettler L, Cochrane G et al (2011) The genomic standards consortium. PLoS Biol 9: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 12:656–664
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–D25
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: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–D538
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–D42
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 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 33:5691–5702
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–D406
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–D254
18. Tang W, Wilkening J, Desai N, Gerlach W, Wilke A, Meyer F (2013) A scalable data analysis platform for metagenomics. Proceedings of the 2013 International Conference on Big Data
19. Bischof, J., Wilke, A., Gerlach, W., Harrison, T., Paczian, T., Tang, W., Trimble, W., Wilkening, J., Desai, N. and Meyer, F. (2015), Shock: Active Storage for Multicloud Streaming

Data Analysis, 2nd IEEE/ACM International Symposium on Big Data Computing, Limassol, Cyprus, 2015

20. Cox MP, Peterson DA, Biggs PJ (2010) Solexaqa: at-a-glance quality assessment of illumina second-generation sequencing data. BMC Bioinformatics 11:485
21. Huse SM, Huber JA, Morrison HG et al (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8:R143
22. Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. ISME J 3:1314–1317
23. 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:e1002541
24. 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:R25
25. 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 13:183
26. Rho M, Tang H, Ye Y (2009) Fraggenescan, Predicting genes in short and error prone reads. Nucleic Acids Res 38:e191
27. Edgar RC (2010) Search and clustering orders of magnitude faster than blast. Bioinformatics 26:2460–2461
28. Caporaso JG, Kuczynski J, Stombaugh J et al (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336
29. Huson DH, Auch AF, Qi J et al (2007) Megan analysis of metagenomic data. Genome Res 17:377–386
30. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75. doi:10.1186/1471-2164-9-75
31. 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 35:7188–7196
32. 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 72:5069–5072
33. 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 31:442–443
34. 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 29:415–420
35. Bolotin A, Quinquis B, Sorokin A et al (2005) Clustered regularly interspaced short palindrome repeats (CRISPRS) have spacers of extrachromosomal origin. Microbiology 151:2551–2561
36. Reeder J, Knight R (2009) The 'rare biosphere', a reality check. Nat Methods 6:636–637
37. Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a web browser. BMC Bioinformatics 12:385
38. Gerlach, W., Tang, W., Keegan, K., Harrison, T., Wilke, A., Bischof, J., D'Souza, M., Devoid, S., Murphy-Olson, D., and Desai, N. (2014) Skyport – Container-based execution environment management for multi-cloud scientific workflows. In Proc. 5th Int'l Workshop on Data-Intensive Computing in the Clouds. IEEE Press, pp. 25–32