ANL/MCS-TM-333

MG-RAST
A Technical Report and Manual for Version 3.3.6 – rev1

Mathematics and Computer Science Division

About Argonne National Laboratory
Argonne is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC under contract DE-AC02-06CH11357. The Laboratory's main facility is outside Chicago, at 9700 South Cass Avenue, Argonne, Illinois 60439. For information about Argonne and its pioneering science and technology programs, see www.anl.gov.

Availability of This Report
This report is available, at no cost, at http://www.osti.gov/bridge. It is also available on paper to the U.S. Department of Energy and its contractors, for a processing fee, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
phone (865) 576-8401
fax (865) 576-5728
reports@adonis.osti.gov

ANL/MCS-TM-333

MG-RAST
A Technical Report and Manual for Version 3.3.6 – rev1

by
A. Wilke, E. M. Glass, J. Bischof, D. Braithwaite, M. DSouza, W. Gerlach, T. Harrison, K. Keegan, H. Matthews, T. Paczian, W. Tang, W. L. Trimble, J. Wilkening, N. Desai, F. Meyer
Mathematics and Computer Science Division, Argonne National Laboratory

May 31, 2013

MG-RAST
A technical report and manual
for version 3.3.6 – rev1

Andreas Wilke[1,2], Elizabeth M. Glass[1,2], Jared Bischof[2,1], Daniel Braithwaite[1,2], Mark DSouza[2,1], Wolfgang Gerlach[2,1], Travis Harrison[2,1], Kevin Keegan[1,2], Hunter Matthews[1,2], Tobias Paczian[2,1], Wei Tang[1,2], William L. Trimble[2,1], Jared Wilkening[1,2], Narayan Desai[1,2], and Folker Meyer[1,2]

[1]Argonne National Laboratory
[2]University of Chicago

May 17, 2013

Contents

Chapter 1

Introduction

The National Institute of Healths NHGRI publishes information (see Figure 1) describing the development of computing costs and DNA sequencing cost over time [25]. The dramatic gap between the shrinking sequencing cost and the more of less stable computing costs is a major challenge for biomedical researchers trying to use next generation DNA sequencing platforms to obtain information on microbial communities. Wilkening *et al.* [43] provide a real currency cost for the analysis of 100 gigabasepairs of DNA sequence data using BLASTX on Amazon's EC2 service: 900,000 US Dollars[1]. A more recent study by University of Maryland researchers [1] allows the computation of a real currency cost for a terabase of DNA shotgun data using complete metagenome analysis pipeline at over 5 million dollars per terabase.

However the growth in data enabled by next-generation sequencing platforms also provides an exciting opportunity for studying microbial communities, 99% of the microbes in which have not yet been cultured [33]. Cultivation free methods (often summarized as Metagenomics) offer novel insights into the biology of the vast majority of life on the planet [37].

Three types of metagenomics experiments are commonly used:

1. Environmental clone libraries (functional metagenomics)

   Frequently using Sanger sequencing instead of more cost efficient next generation sequencing.

2. Amplicon studies (single gene studies, 16s rDNA)

   Next generation sequencing of PCR amplified ribosomal genes providing a single reference gene based view of microbial community ecology.

3. Shotgun metagenomics

[1]This includes only the computation cost, no data transfer cost and was computed using 2009 prices.

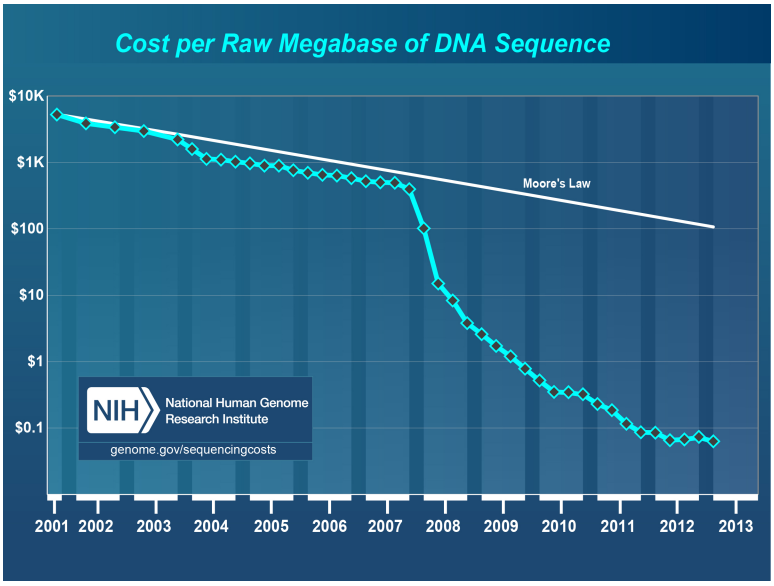


Figure 1.1: The cost for DNA sequencing is shriking. This comparison with Moore's law roughly describing the development of computing costs highlights the growing gap between sequence data and the available analysis resources. Source: NGHRI

The use of next generation technology applied directly to to environmental samples.

Each of these methods have strengths and weaknesses (see [37]) and so do the various sequencing technologies (see [26]).

To support user-driven analysis of all types of metagenomic data, we have provided MG-RAST [24] at `http://metagenomics.anl.gov`. MG-RAST enables researchers to study function and composition of microbial communities.

The MG-RAST portal offers automated quality control, annotation and comparative analysis services and archiving service. At the time of writing (May 30, 2013) MG-RAST has completed the analysis of over 25 Terabasepairs of DNA data in over 78,000 datasets contributed by thousands of researchers world-wide.

The MG-RAST system provides answers to the following scientific questions:

- Who is out there?

  Identifying the composition of microbial composition using either amplicon data for single genes or deriving community composition from shotgun metagenomic data using sequence similarities.

- What are they doing?

Using shotgun data (or metatranscriptomic data) derive the functional complement of a microbial community using similarity searches against a number of databases.

- Who is doing what?

  Based on sequence similarity searches, identify the organisms encoding specific functions.

1.1 Version history

Version 1

The original version of MG-RAST was developed in 2007 by Folker Meyer, Andreas Wilke, Daniel Paarman, Bob Olson and Rob Edwards. It relied heavily on the SEED [28] environment and allowed upload of pre-processed 454 and Sanger data.

Version 2

Version 2 released in 2008 had numerous improvements. It was optimized to handle full sized 454 data sets and is the first version of MG-RAST that was not fully SEED based. Version 2.0 used BLASTX analysis for both gene prediction and functional classification [24].

Version 3

While the previous version of MG-RAST (v2) was widely used, it was limited to datasets smaller than a few 100 Mbases and comparison of samples was limited to pairwise comparisons. Version 3 is not based on SEED technology, but uses the SEED subsystems as a preferred data source. Starting with version 3, MG-RAST moved to github.

1.2 Comparison of versions 2 and 3

In the new 3.0 version, datasets of 10s of gigabases can be annotated and comparison of taxa or functions that differed between samples is now limited by the available screen real estate. Figure 1.2 shows a comparison of the analytical and computational approaches used in MG-RAST v2 and v3. The major changes are the inclusion of a dedicated gene calling stage using FragGenescan [32], clustering of predicted proteins at 90% identify using uclust [9] and the use of BLAT [18] for the computation of similarities. Together with changes in the underlying infrastructure this has allowed dramatic scaling of the analysis with the limited hardware available.

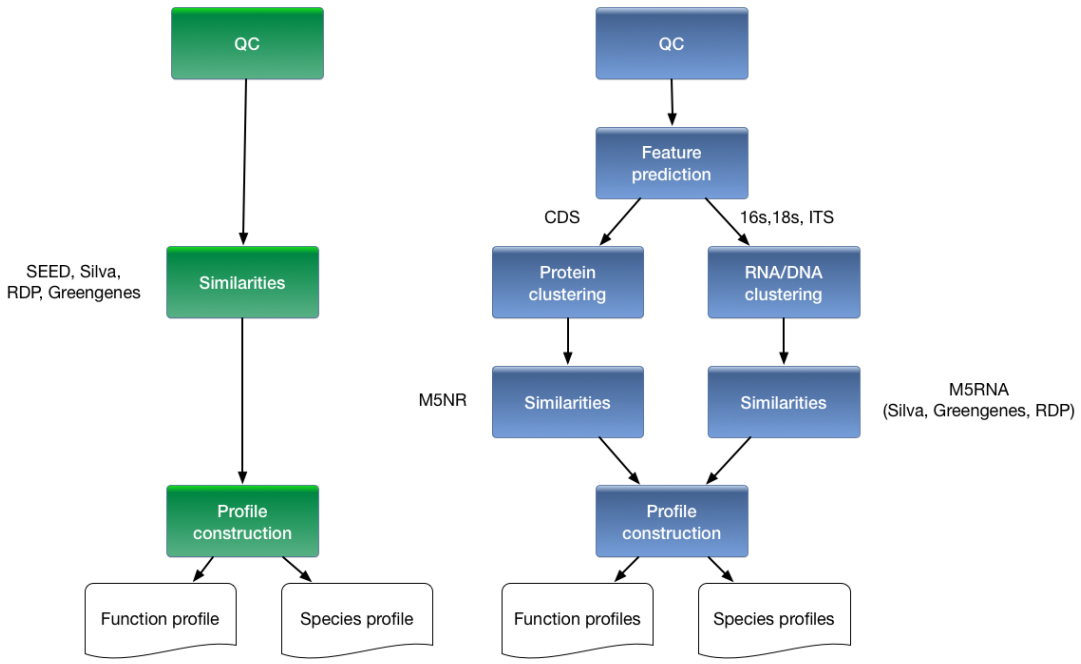


Figure 1.2: Overview of processing pipeline in (a) MG-RAST 2 and (b) MG-RAST 3. In the old pipeline, metadata was rudimentary, compute steps were performed on individual reads on a 40-node cluster that was tightly coupled to the system, and similarities were computed by BLAST to yield abundance profiles that could then be compared on a per-sample or per pair basis. In the new pipeline, rich metadata can be uploaded, normalization and feature prediction are performed, faster methods such as BLAT are used to compute similarities, and the resulting abundance profiles are fed into downstream pipelines on the cloud to perform community and metabolic reconstruction and to allow queries according to rich sample and functional metadata.

Similar to version 2.0, the new version of MG-RAST does not pretend to know the correct parameters for the transfer of annotations. Instead the user im empowered to choose the best parameters for their data sets.

The new version of MG-RAST represents a rethinking of core processes and data products, as well as new user interface metaphors and a redesigned computational infrastructure. MG-RAST supports a variety of user-driven analyses, including comparisons of many samples, previously too computationally intensive to support for an open user community.

Scaling to the new workload required changes in two areas: the underlying infrastructure needed to be re-thought and the analysis pipeline needed to be adapted to address the properties of the newest sequencing technologies.

1.3 The MG-RAST team

- Andreas Wilke
- Elizabeth M. Glass
- Jared Bischof
- Daniel Braithwaite
- Mark DSouza
- Wolfgang Gerlach
- Travis Harrison
- Kevin Keegan
- Hunter Matthews
- Tobias Paczian
- Wei Tang
- William L. Trimble
- Jared Wilkening
- Narayan Desai
- Folker Meyer

1.3.1 Contacting the MG-RAST team; the MG-RAST help desk

The MG-RAST project uses a ticket system to manage the interaction with our users.

Dr. Mark D'Souza is managing the help desk interaction with the users, please do not email him directly but use the address given in Figure 1.3.

We recommend including as much detail as possible into your emails to the help-desk, details like account names, MG-RAST identifiers will help us identify any issues and speed up resolving them.

Below is an example of the types of details we'd like to receive:

- your name

mg-rast@mcs.anl.gov

Figure 1.3: The email address for the MG-RAST project. Note that this is inserted into the document as an image, you will have to type it.

- a clear text description of your problem
- any MG-RAST identifiers (those are the 444xxxx.3 numbers)
- any project numbers
- what browser in which version are you using, if the problem relates to the web site
- what platform was your data created on
- if your data was a failure in the web site, what time did it occur

1.3.2 Past members of the MG-RAST team

The following people were associated with MG-RAST in the past:

- Daniel Paarmann, 2007-2008
- Rob Edwards, 2007-2008
- Mike Kubal, 2007-2008
- Alex Rodriguez, 2007-2008
- Bob Olson, 2007-2009
- Daniela Bartels, 2007-2011
- Yekaterina Dribinsky, 2011

MG-RAST was started by Rob Edwards and Folker Meyer in 2007.

Chapter 2

Under the hood: The MG-RAST technology platform and pipeline

2.1 System design aspects

One key aspect of scaling MG-RAST to large numbers of modern NGS datasets is the use of cloud [1]. computing which decouples MG-RAST from its previous dedicated hardware resources. Using our task server AWE [42] and the SHOCK data management system developed alongside it we have updated our underlying computational platform using purpose built software platform optimized for large scale sequence analysis.

The new analytical pipeline for MG-RAST version 3 (see Figure 2.3) is encapsulated and separated from the data store, enabling far greater scalability.

Combined the changes in infrastructure and pipeline had made the new MG-RAST version 750 times faster than version 2.

2.2 Data model

The MG-RAST data model (see Figure 2.1) is something that changes dramatically to handle the size of modern next generation sequencing data sets. We have made a number of choices that reduce the computational and storage burden.

It is important to mention that the size of the derived data products for a next generation data in MG-RAST set is typically about 10x the size of the actual data set. Individual data sets now reach

[1] We use the term *cloud* as a shortcut for Infrastructure as a Service (IaaS)

up to a Terabase[2] with the on disk footprint significantly larger than the basepair count due to the inefficient nature of FASTQ files, that basically double the on disk size for FASTQ representations.

- The use abundance profiles.

  Using abundance profiles where we count the number of occurences of function or taxon per metagenomic data set is one important factor that keeps the data sets managble. Instead of growing the data set sizes (often with several hundred million individual sequences per data set) the data products now are more or less static in size.

- Use a single similarity computing step per feature type.

  By running exactly one similarity computation for proteins and another one for rRNA features, we have limited the computional requirements.

- The use of clustering of features.

  By clustering features at 90% identity, we reduce the number of times we compute on similar proteins. Abundant features will be even more efficiently clustered leading to more compression among for abundant species.

As shown in Figure 2.1 MG-RAST relies on abundance profiles to capture information for each metagenome.

The following abundance profiles are calculated for every metagenome:

- MD5s

  This table represents the number of sequences (clusters) per database entry in the M5nr.

- functions

  This table represents a summary of all the MD5s that match a given function.

- ontologies

  This table represents a summary of all the MD5s that match a given hierarchy entry.

- organisms

  This table represents a summary of all MD5s that match a given taxon entry.

- LCAs

The static helper tables (show in blue in figure 2.2) help keep the main tables smaller, by normalizing and providing integer representations for the entities in the abundance profiles.

[2]This would be for several metagenomes that are part of the JGI Prairie pilot.

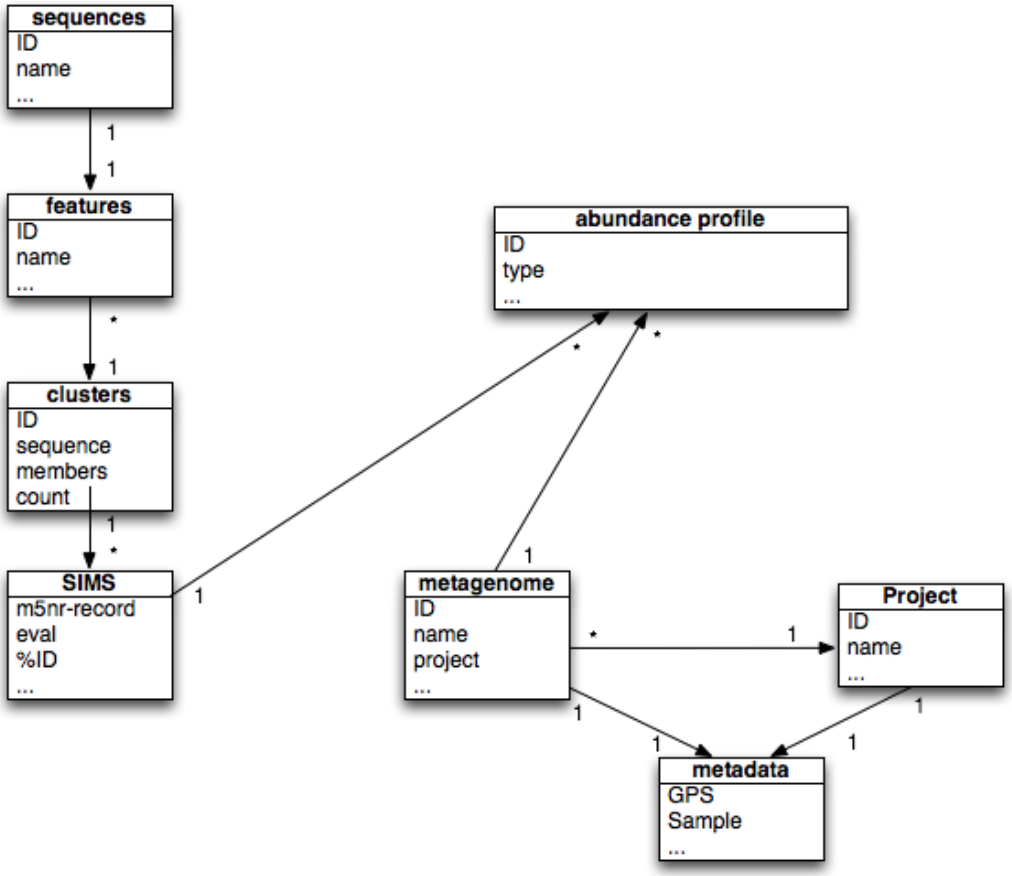


Figure 2.1: The MG-RAST v3 data model.

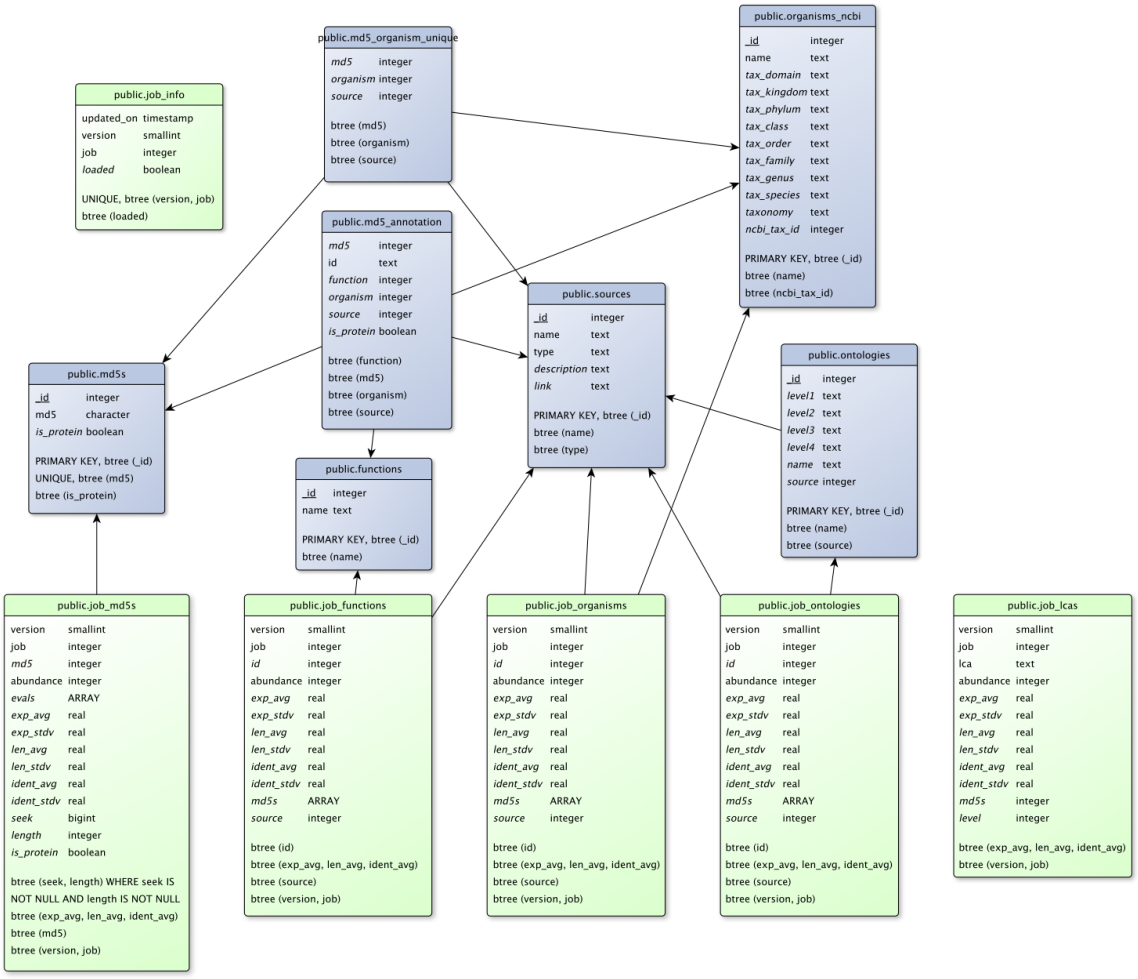


Figure 2.2: The analysis database schema, shows the static objects in blue and the per metagenome (variable) objects in green.

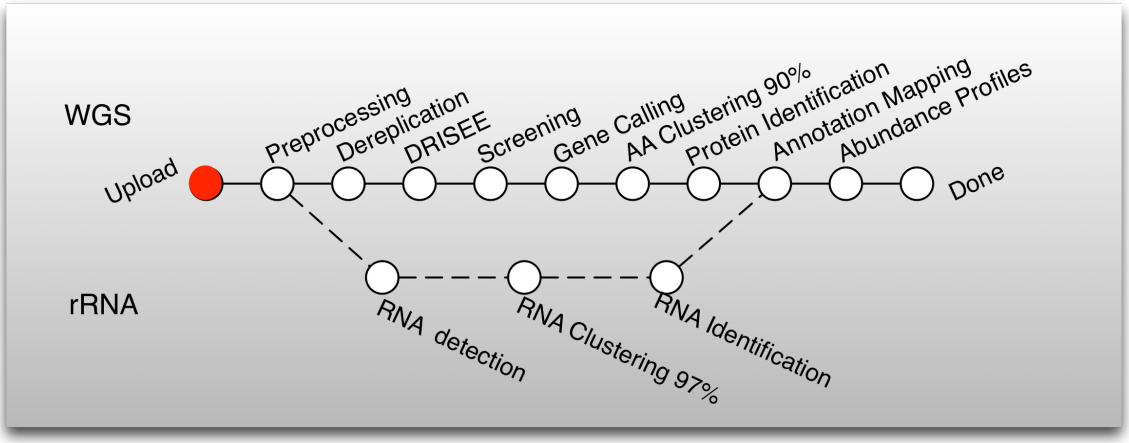


Figure 2.3: Details of the analysis pipeline for MG-RAST version 3.x.

2.3 Details on the new MG-RAST pipeline

The pipeline shown in Figure 2.3) contains a significant number of improvements over version 3.0. Using the M5NR [41] the new pipeline computes results against many reference databases instead of only SEED. Several key algorithmic improvements were needed to support the flood of user-generated data (see Figure 2.4). Using a dedicated software to perform gene prediction instead of using a similarity based approach reduces runtime requirements, the additional clustering of proteins at 90% identity reduces data while preserving biological signal.

Due to the amount of sequence data submitted to MG-RAST (see Figure 2.4) only protein coding genes and ribosomal RNA (rRNA) genes will be annotated by the pipeline.

Below we describe each step of the pipeline in some detail, all data sets generated by the individual stages of the processing pipeline are made available as downloads. Appendix A lists the available files for each data set.

2.3.1 Preprocessing

After upload, data is pre-processed using SolexaQA [7] to trim low quality regions from FASTQ data. Platform specific approaches are used for 454 data submitted in FASTA format: reads more than than two standard deviations away from the mean read length are discarded following [13]. All sequences submitted to the system are available but discarded reads will not be analyzed further.

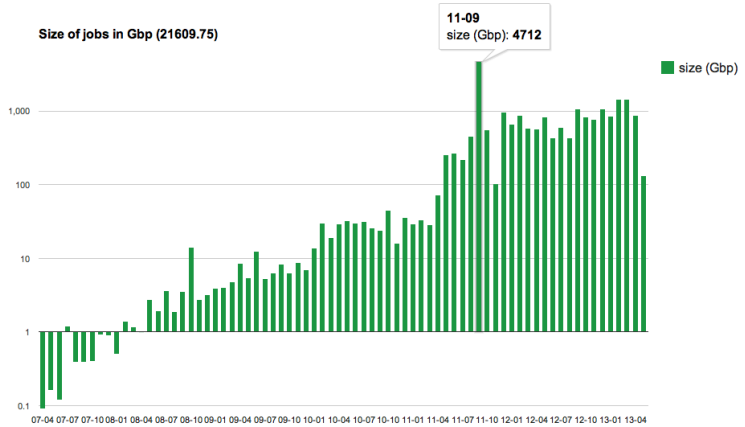


Figure 2.4: The sizes of MG-RAST jobs per month in gigabasepairs.

2.3.2 Dereplication

For shotgun metagenome and shotgun metatranscriptome data sets we perform a de-replication step. We use a simple k-mer based approach to rapidly identify all 20 character prefix identical sequences. This step is required to remove to remove artificial duplicate reads (ADRs) [12]. Instead of simply discarding the ADRs, we set them aside and use them later.

We note that de-replication is not suitable for amplicon data sets that are likely to share common pre-fixes.

2.3.3 DRISEE

MG-RAST version 3 uses DRISEE [17] to analyze the sets of Artificial Duplicate Reads [12] and determine the degree of variation among prefix identical sequences derived from the same template. See below for details.

2.3.4 Screening

The pipeline provides the option to remove 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 [20] 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 to avoid storing any human DNA on MG-RAST.

2.3.5 Gene Calling

While the previous version of MG-RAST used similarity-based gene predictions, this approach is significantly more expensive computationally than de-novo gene prediction. After an in depth investigation of tool performance [38], we have moved to a machine learning approach: FragGeneScan [32]. 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.

It is important to note that FragGeneScan is trained for prokaryotes only. While it will identify proteins for eukaryotic sequences, the results should be viewed as more or less random.

2.3.6 AA Clustering

MG-RAST builds clusters of proteins at the 90% identity level using the uclust [9] implementation in QIIME [5] 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 signal.

2.3.7 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 [18], which we parallelized using OpenMPI [11] 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.

2.3.8 Annotation Mapping

Sequence similarity searches are computed against a protein database derived from the M5NR [41], which provides a non-redundant integration of many databases (GenBank, [3], SEED [28], IMG [22], UniProt [21], KEGG [16] and eggNOGs [15]). Unlike MG-RAST 2, which relied solely on SEED, MG-RAST now supports many complementary views into the data with one similarity search, including different functional hierarchies: SEED Subsystems, IMG terms, COG [36], eggNOGs [15] and ontologies such as GO (Gene Ontology Consortium, 2013). 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 dataset of four cow rumen metagenomes.

2.3.9 Abundance Profiles

Anbundance profiles are the primary data product that MG-RAST's user interface uses to display information on the data sets. Abundance profiles functional and taxonomic information.

Using the abundance profiles the MG-RAST systems deffers making a decision on when to transfer annotations. As 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 cut-off values.

The threshold for annotation transfer can be set using the following parameters:

- e-value
- percent identity
- minimal alignment length

The taxonomic profiles use the NCBI taxnomy, all taxonomic information is projected against this data. The functional profile are available for data sources that provide hierarchical information. These are currently:

- SEED Subsystems

  The SEED Subsystems [28] represent an independent re-annotation effort that powers e.g. the RAST [2] effort. Manual curation of subsystems makes them an extremely valuable data source.

  Subsystems represent a hierarchy:

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

  The page at `http://pubseed.theseed.org//SubsysEditor.cgi` allows browsing the Subsystems.

- KEGG Orthologues

  We use the KEGG [16] enzyme number hierarchy to implement a four level hierarchy

  1. KEGG level 1 – first digit of the EC number (EC:X.*.*.*)
  2. KEGG level 2 – the first two digits of the EC number (EC:X.Y.*.*)

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

We note that KEGG data is no longer available for free download and we thus have to rely on using the latest freely downloadable version of the data.

The high level KEGG categories are:

1. Cellular Processes
2. Environmental Information Processing
3. Genetic Information Processing
4. Human Diseases
5. Metabolism
6. Organisational Systems

- COG and EGGNOG categories The high level COG and EGGNOG categories are:

  1. Cellular Processes
  2. Information Storage and Processing
  3. Metabolism
  4. Poorly Characterized

  We note that for most metagenomes the coverage of each of the four namespaces is quite different. The ”Source hit Distribution” (see 3.6.1.2) provides information on how many sequences per data set were found for each database.

2.4 The rRNA pipeline

rRNA reads are identified using a simple rRNA detection pipeline and are searched in a separate flow in the pipeline.

2.4.1 rRNA detection

An initial BLAT [18] search against a reduced RNA database efficiently identifies RNA.

The reduced database is a 90% identity clustered version of the Silva database and is merely used to rapidly identify sequences with similarities to ribosomal RNA.

2.4.2 rRNA CLUSTERING

The rRNA-similar reads are then clustered at 97% identity and the longest sequence is picked as the cluster representative.

2.4.3 rRNA IDENTIFICATION

A BLAT similarity search is performed for the longest cluster representative against the M5rna database, integrating SILVA [29], Greengenes [8] and RDP [6].

2.5 Quality Assessment

The MG-RAST pipeline offers a variety of summaries of technical aspects of the sequence quality to enable sequence data triage. These tools include DRISEE for estimating sequence error, summaries of the spectra of long kmers, and visualizations of the base caller output.

2.5.1 DRISEE, (Duplicate Read Inferred Sequencing Error Estimation)

DRISEE [17] is a method to provide a measure for sequencing error for whole genome shotgun metagenomic sequence data that is independent of sequencing technology, and accounts for many of the shortcomings of Phred. It utilizes ADRs (artifactual/artificial duplicate reads) to generate internal sequence standards from which an overall assessment of sequencing error in a sample is derived. DRISEE values are normally reported as percent error.

DRISEE values can be used to assess the overall quality of sequence samples. DRISEE data are presented on the Overview page for each MG-RAST sample for which a DRISEE profile can be determined. Total DRISEE Error presents the overall DRISEE based assessment of the sample as a percent error:

$TotalDRISEEError = base\_errors/total\_bases * 100$

where base_errors refers to the sum of DRISEE detected errors and total_bases refers to the sum of all bases considered by DRISEE.

The current implementation of DRISEE is not suitable for amplicon sequencing data, or other samples that may contain natural duplicated sequences (e.g. eukaryotic DNA where gene duplication and other forms of highly repetitive sequences are common) in high abundance.

2.5.2 Kmer profiles

k-mer digests are an annotation-independent method to describe sequence datasets that can support inferences about genome size and coverage. Here the overview page presents several visualizations of the kmer spectrum of each dataset, evaluated at k=15.

Three visualizations provided of the kmer spectrum are the kmer spectrum, kmer rank abundance, and ranked kmer consumed. All three graphs represent the same spectrum, but in different ways. The kmer spectrum plots the number of distinct kmers against kmer coverage. The kmer coverage is equivalent to number of observations of each kmer. The kmer rank abundance plots the relationship between kmer coverage and the kmer rankanswering the quesiton what is the coverage of the nth most-abundant kmer. Ranked kmer consumed plots the largest fraction of the data explained by the nth most abundant kmers only.

2.5.3 Nucleotide histograms

These graphs show the fraction of base pairs of each type (A, C, G, T, or ambiguous base N) at each position starting from the beginning of each read. Amplicon datasets (see Figure 2.5) should show biased distributions of bases at each position, reflecting both conservation and variability in the recovered sequences:

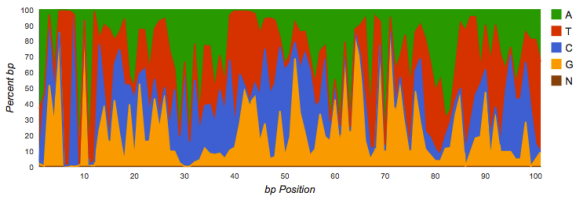


Figure 2.5: Nucleotide histogram with biased distributions typical for an amplicon data set

Shotgun datasets should have roughly equal proportions of A, T, G and C basecalls, independent of position in the read as shown in Figure 2.6.

Vertical bars at the beginning of the read indicate untrimmed (see Figure 2.7), contiguous barcodes. Gene calling via FragGeneScan [32] and RNA similarity searches are not impacted by the presence of barcodes. However if a significant fraction of the reads is consumed by barcodes it reduces the biological information contained in the reads.

If a shotgun dataset has clear patterns in the data (see Figure 2.8, this indicates likely contamination with artificial sequences. This dataset had a large fraction of adapter dimers:

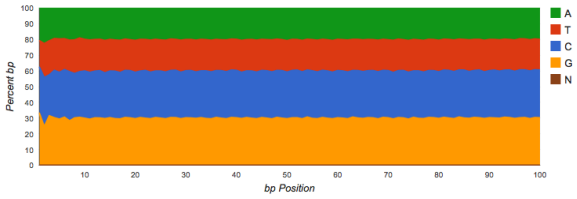


Figure 2.6: Nucleotide histogram showing ideal distributions typical for a shotgun metagenome.

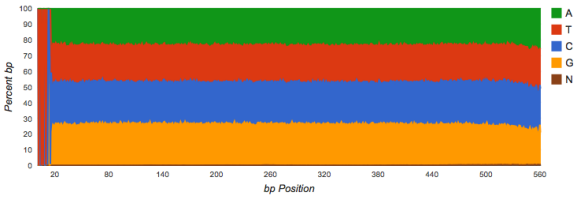


Figure 2.7: Nucleotide histogram with untrimmed barcodes.

2.6 Representative hit, best hit, and Lowest Common Ancestor interpretation

MG-RAST searches the non-redundant 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 different ways of finding the right function or taxon information. This impacts the end-user experience as they have three different methods to choose the number of hits reported for a given sequence in their data set. The details on the three different classification functions implemented are below:

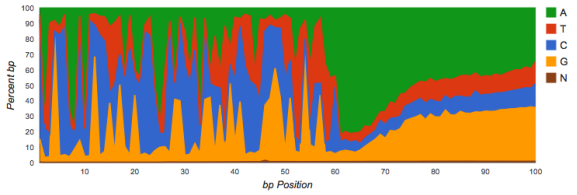


Figure 2.8: Nucleotide histogram with contamination.

2.6.1 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 we have decided to double count all annotations with identical match properties and leave 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. If you are looking for a specific species or function in your results, the best hit function is likely what you are looking for.

2.6.2 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 our database. This makes counts additive across functional and taxonomic levels and thus allows for example to compare functional and taxonomic profiles of different metagenomes.

2.6.3 Lowest Common Ancestor (LCA)

To avoid the problem of multiple taxonomic annotations for a single feature we provide taxonomic annotations based on the widely used LCA-method (lowest common ancestor) introduced by MEGAN [14]. In this method all hits that have a bit score close to the bit score of the best hit are collected. 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.

The number of hits (occurrences of the input sequence in the database) may be inflated if the best hit filter is used, or your favorite species might be missing despite a very similar sequence similarity result if using the representative hit classifier function (in fact 100% identical match to your favorite species exists).

One way to consider both representative and best hit is that they over-interpret the available evidence, with the LCA classifier function any input sequence is only classified down to a trustworthy taxonomic level. While naively this seems to be the best function to choose in all cases as it classifies sequences to varying depths, this causes problems for downstream analysis tools that might rely on everything being classified to the same level.

2.7 Why dont the numbers of annotations add up to the number of reads ?

The MG-RAST v3 annotation pipeline does not usually provide a single annotation for each submitted fragment of DNA. There are steps in the pipeline that map one read to multiple annotations and one annotation to multiple reads. These steps are a consequence of genome structure, pipeline engineering, and the character of the sequence databases that MG-RAST uses for annotation.

The first step that is not one-to-one is gene prediction. Long reads (¿400bp) and contigs can contain pieces of two or more microbial genes; when the gene caller makes this prediction, the multiple predicted protein sequences (called fragments) are annotated separately.

There is an intermediate clustering step that identifies sequences at 90% amino acid identity and performs one search for each cluster. Sequences that do not fall into clusters are searched separately. The abundance column in the MG-RAST tables presents the estimate of the number of sequences that contain a given annotation, found by multiplying each selected database match (hit) by the number of representatives in each cluster. The final step that is not one-to-one is the annotation process itself. Sequences can exist in the underlying data sources many times with different labels. When those sequence are the best hit similarity, we do not have a principled way to choosing the correct label. For this reason we have decided to double count these annotations and leave determination of truth to our users. Note: Even when considering a single data source, double-counting can occur depending on the consistency of annotations. Also note, Hits refers to the number of unique database sequences that were found in the similarity search, NOT the number of reads. The hit count can be smaller than the number of reads because of clustering or larger due to double counting.

2.8 Metadata, Publishing and Sharing

As mentioned above MG-RAST is both an analytical platform and a data integration. To enable data re-use for e.g. metaanalyses 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 Genomics Standards Consortium (GSC) [10].

2.8.1 Metadata

While the GSC provide a GCDML [19] encoding this XML based format is more useful to programmers than to end users submitting data.

We have therefore elected to use spreadsheets to transport metadata. Specifically we use MIxS (Minimum information about any (x) sequence (MIxS) and MIMARKS (Minimum Information about a MARKer gene Survey) to encode minimal metadata [44].

The metadata describe both origins of samples and provide details on the generation of the sequence data. While the GSC checklist aims at capturing a minimum of information, MG-RAST can handle additional metadata if supplied by the user. The metadata is stored in a simple key value format and is displayed on the Metagenome Overview page.

Once uploaded metadata spreadsheets are validated automatically and users informed of any problems.

The presence of metadata enables discovery by end-users using contextual metadata. Users can perform searches like retrieve soil samples from the continental U.S.. If the users have added additional metadata (domain specific extension) additional queries are enabled e.g. restrict the results to soils with a specific pH.

2.8.2 Publishing

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.

2.8.3 Sharing

In addition to publishing, data and analysis can also be shared with specific users. To share data with a user simply enter their email address via clicking sharing on the overview page. The dialogue shown in Figure 2.9 will allow entering email addresses.

As shown in Figure 2.10 we tend to see data sets sharing between small groups of users.

2.8.4 Identifiers

MG-RAST automatically assigns a unique identifier to every data set submitted. Upon completion of the automated pipeline data sets can be viewed via the web interface using the identifiers.

The data set identifiers are of the form `integer prefix.revision`, an example is `4440283.3`.

In addition to data sets MG-RAST supports projects (groups of data sets) that can be addressed with simple numerical project identifiers.

An example: `http://metagenomics.anl.gov/linkin.cgi?project=128`

» Back to the Metagenome Select

Job Information

Name - ID:	4447970.3 - CA_05_4.6
Job:	#1
User:	Pedro.Belda

share multiple metagenomes

To share the above job and its data with another user, please enter the email address of the user. Please note that you have to enter the email address which that person used to register at the MG-RAST service. The user will receive an email that notifies him how to access the data. Once you have granted the right to view one of your MG-RAST jobs to another user or group, the name will appear at the bottom of the page with the option to revoke it.

Enter an email address

Enter an email address:

Share job with this user or group

This job is currently available to:

Figure 2.9: The sharing mechanisms requires a valid email address for the user the data is to be shared with. A list of users with access to the data is displayed at the bottom on the page.

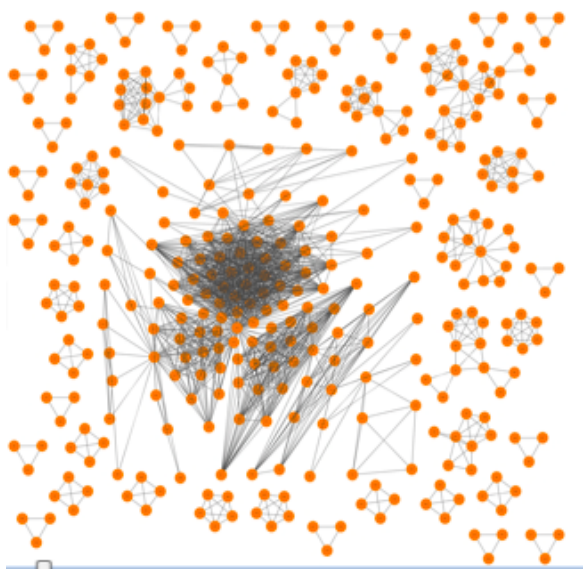


Figure 2.10: Data sets shared in MG-RAST by users (orange dots) are shown as connecting edges.

2.8.5 Linking to MG-RAST

As future version of MG-RAST may change, we provide a link-in mechanism as a stable way of linking to MG-RAST. If you need to link out to data sets in MG-RAST the `linkin.cgi` is the right way to do it.

```
http://metagenomics.anl.gov/linkin.cgi?metagenome=
http://metagenomics.anl.gov/linkin.cgi?project=
```

Figure 2.11: The `linkin.cgi` mechanism provides stable URLs for linking to MG-RAST.

For example, for the metagenome ID `4440283.3` the URL is: `http://metagenomics.anl.gov/linkin.cgi?metagenome=4440283.3`

This URL provides a stable method of linking to your data which does not require the viewer to have an MG-RAST account. Please do not use the URL you see when you are browsing the site.

By default your data is not visible to others, you will need to explicitly grant permission for it to be visible to anyone on the internet by making it public through the MG-RAST website.

Chapter 3

The MG-RAST v3 web 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 components 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 (Yilmaz, 2011) (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 datasets. An example query using the metadata browser, enabling the user to interrogate the existing pool of public data sets for a Biome of interest (e.g. Hot springs) and performing comparisons and a search for organisms encoding a specific gene function (e.g. Beta-lactamase or Aldo/keto reductase; see Figure 3.1.

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 e.g. a comparison between samples. This is done without the need for re-computation of results.

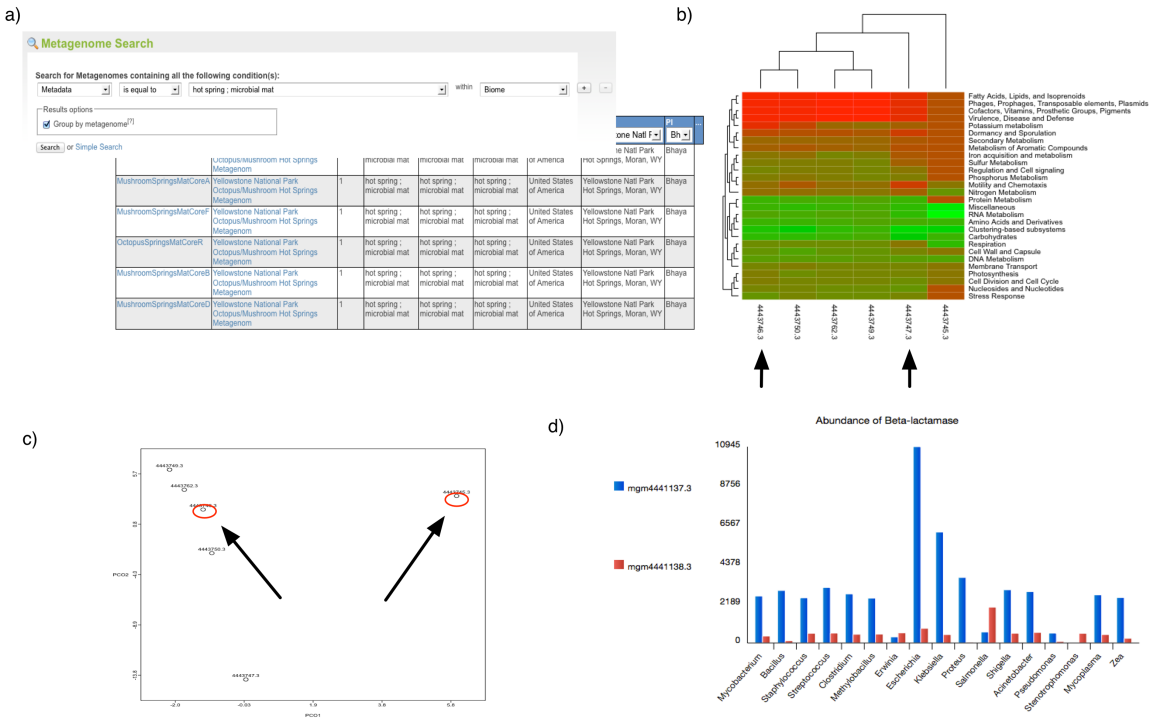


Figure 3.1: a) Using the web interface for a search of metagenomes for microbial mats in hot-springs (GSC-MIMS-Keywords Biome=hotspring; microbial mat) we find 6 metagenomes (refs: 4443745.3, 4443746.3, 4443747.3, 4443749.3, 4443750.3, 4443762.3). b) Initial comparison reveals some differences in protein functional class abundance (using SEED subsystens level 1). C) From the PCoA plot using normalized counts of functional SEED Subsystem based functional annotations (level 2) and Bray-Curtis as metric, we attempt to find differences between two similar datasets (MG-RAST-IDs: 444749.3, 4443762.3). d) Using exported tables with functional annotations and taxonomic mapping we analyze the distribution of organisms observed to contain Beta-lactamase and plot the abundance per species for two distinct samples.

3.1 Technical details

3.1.1 Browser requirements

The current web interface for MG-RAST is being developed for recent versions of Firefox. If you are using another browser please understand that some or all of the web site will not function.

We understand that this may not be your favorite (or institutionally prescribed) browser, but writing interactive web sites for many browsers is hard. While we are aiming to create a multi-browser version of the web interface, the current version is limited to Firefox.

3.1.2 Downloading figures

Almost all figures and tables are downloadable into either graphics or spreadsheets. Please look for a download chart data link next to the graphic.

3.2 Sitemap for MG-RAST

The MG-RAST web site (as shown in Figure 3.2) is rather complex and offers a lot of different options.

In addition to the home page, the site at `http://metagenomics.anl.gov` has 5 main pages (shown in blue in Figure 3.2).

- Download page

  This page lists all publicly available data for download. The data is structured into projects.

- the Browse page

  This page allows interactive browsing of all data sets and is powered by metadata.

- the Search page

  The search page allows identifier, taxonomy and function driven searches against all public data.

- the Analysis page

  The Analysis page enables comparisons between data sets and in depth analyses.

- the Upload page

  This is allowing users to provide their samples and metadata to MG-RAST. More details on uploading are below.

- the Metagenome Overview page

  For each indivual data set this page provides an overview.

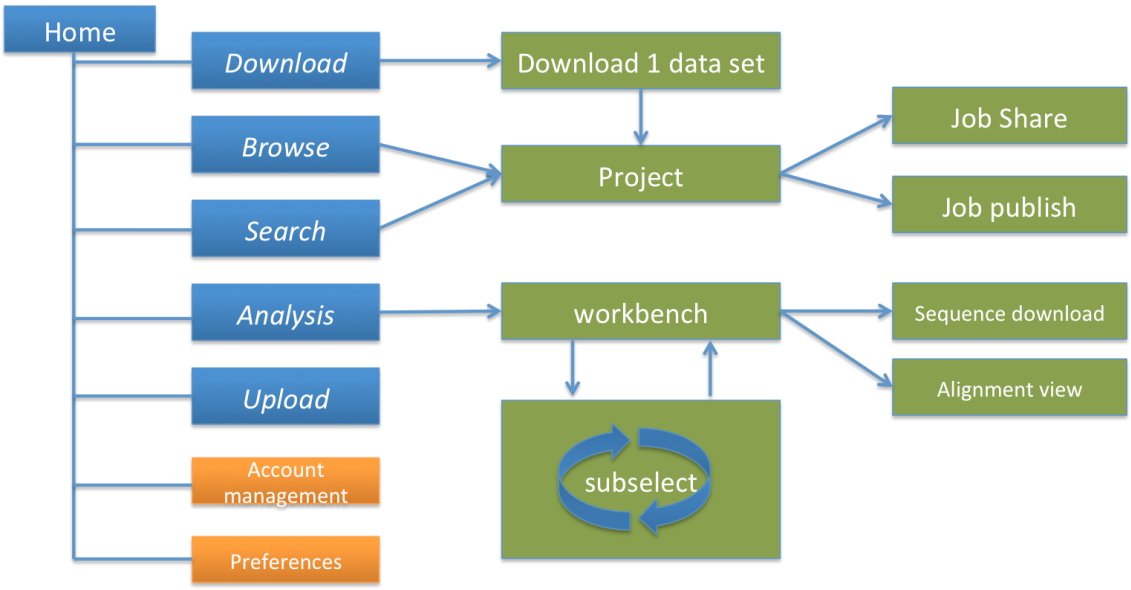


Figure 3.2: The sitemap for the MG-RAST version 3 web site. On the site map the main pages are shown in blue, management pages in orange. The green boxes represent pages that are not directly accessible from the home page.

It is important to mention that if you want to create links to the MG-RAST web site you should use the *linkin* mechanism instead of linking to any web page directly. All pages intended for users to create external links to provide the linkin feature.

3.3 The Upload and Metadata pages

Data and Metadata can be uploaded in the form of spreadsheets along with the sequence data 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 data sets of over 100GBp (gigabasepairs) often resulting in file sizes in excess of 150 GB.

3.4 The Browse page – Metadata enabled data discovery

The Metagenome Browse page lists all data sets visible to the user.[1] This page also provides an overview of the non-public data sets submitted by the user or shared with them.

Figure 3.3 shows the interactive metagenome browse table, which 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).

3.5 Project Page

Shown in Figure 3.4 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 via 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 labelled arrows.

For the data set owners the project page provides editing capability using a number of menu entries at the top of the page. Figure 3.5 shows all available options.

- Share Project

  Make the data in this project avaialble to third parties via sending them access tokens.

- Add Jobs

  Add additional data sets to this project.

- Edit Project data

  Edit the contents of this page.

- Upload info

  Upload information to be displayed on this page.

[1] Datasets in MG-RAST are private by default, but the submitting user has the option to share datasets with specific users or to make datasets public.

ALL METAGENOMES

group by project

Current table counts public (12535) private (0) shared (0)

metagenomes	projects	biomes	features	materials	altitudes	depths	locations	ph's	countries	temperatures	pi's
12535	373	102	101	102	115	296	477	116	71	1005	144

clear table filters add selected to a collection

display 10 items per page

displaying 1 - 10 of 12535 next» last»

project	name	bps	sequences	biome	feature	material	sequencing type		select all	...
		<	<	all	all	all	all			
The oral metagenome in health and disease	CA_05_4.6	27669924	70503	human-associated habitat	human-associated habitat	human-associated habitat	WGS	public		
The oral metagenome in health and disease	CA_06_1.6	37519874	97722	human-associated habitat	human-associated habitat	human-associated habitat	WGS	public		
cDNA - Plymouth Marine Lab Coastal Waters project	1-19-DNA-flx	59316369	344216	marine habitat	marine habitat	marine habitat	WGS	public		
cDNA - Plymouth Marine Lab Coastal Waters project	6-19-DNA-flx	68187679	304020	marine habitat	marine habitat	marine habitat	WGS	public		
Northern Line Islands	FannLIMic20050811	30909241	290844	marine habitat	marine habitat	marine habitat	WGS	public		
Northern Line Islands	FannLIVir20050811	39607682	380355	marine habitat	marine habitat	marine habitat	WGS	public		
Soudan Mine Metagenome	RedSoudMineMic20050331	35439683	334386	mine drainage	mine drainage	mine drainage	WGS	public		
Soudan Mine Metagenome	BlackSoudMineMic20050331	38502057	388627	mine drainage	mine drainage	mine drainage	WGS	public		
Chicken Cecum Microbiome	Chicken Cecum A	32296796	310801	animal-associated habitat	animal-associated habitat	animal-associated habitat	WGS	public		
Chicken Cecum Microbiome	Chicken_Cecum_B	26378422	254712	animal-associated habitat	animal-associated habitat	animal-associated habitat	WGS	public		

displaying 1 - 10 of 12535 next» last»

Figure 3.3: The Metagenome Browser page enables sorting and data search. Users can select the metadata they wish to view and search. Some of the metadata is hidden by default and can be viewed by clicking on the header on the right side of the table and selecting the desired columns, this can also be used to hide unwanted columns.

THE ORAL METAGENOME IN HEALTH AND DISEASE (ID 128)

metagenomes | project metadata

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

Share Project | Add Jobs | Edit Project Data | Upload Info | Upload MetaData | Export MetaData

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.

Export Jobs Table

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	

Figure 3.4: The project page provides a summary of all data in the project and provides an interface for downloads.

Figure 3.5: If you are the data set owner, the project page will display these buttons.

Figure 3.6: Top of the metagenome overview page.

- Upload metadata

  Upload a metadata spreadsheet for the project.

- Export metadata[2]

  Export the metadata spreadsheet for this project.

3.6 The 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 data set. The page is made available by the automated pipeline once the computation is finished.

This page is a good starting for looking at a particular data set and provides a signifcant 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 data sets with default parameters. Further analysis are available on the Analysis page.

3.6.1 The technical part of the overview page – Details on sequencing and analysis

The Overview page provides the MG-RAST ID for a data set, a unique identifier that is usable as accession number for publications. Additional information like the Name of the submitting PI and

[2]This option is available to non data set owners

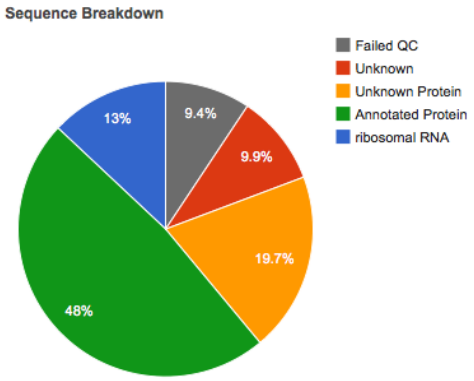


Figure 3.7: Sequences to the pipeline are classified into one of 5 categories. grey = failed the QC, red = unknown sequences, yellow = unknown function but protein coding, green = protein coding with known function and blue = ribosomal RNA. For this example over 50% of sequences were either filtered by QC or failed to be recognized as either protein coding or ribosomal.

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 as additional information (Figure 3.6).

We provide an automatically generated paragraph of text describing the submitted data and the results computed by the pipeline. Via 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 (Figure 3.7) 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.

It should be noted that for performance reasons no other sequence features are annotated by the default pipeline. Other feature types e.g. small RNAs or regulatory motifs (e.g. CRISPRS [4]) will not only require significantly higher computational resources but also are frequently not supported by the unassembled short reads that comprise the vast majority of todays 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 of the data supports.

The overview page also provides metadata (data describing data) for each data set to the extent that data 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.

GSC MIxS INFO

Investigation Type	**metagenome**
Project Name	**The oral metagenome in health and disease**
Latitude and Longitude	**39.481448, 0.353066**
Country and/or Sea, Location	**Spain Valencia**
Collection Date	**2010-03-01 10:00:00 UTC**
Environment (Biome)	**human-associated habitat**
Environment (Feature)	**human-associated habitat**
Environment (Material)	**human-associated habitat**
Environmental Package	**human-oral**
Sequencing Method	**454**
More Metadata	

Figure 3.8: The information from the GSC MIxS checklist providing minimal metadata on the sample.

This is implemented using the standards developed by the Genomics Standards Consortium. Figure 3.8 shows the metadata summary for a data set.

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 are the environmental packages are explicitly encouraged, but not required when using MG-RAST.

3.6.1.1 Metagenome QC

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

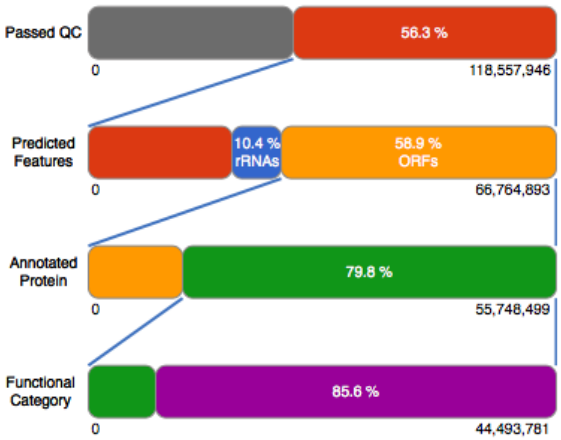


Figure 3.9: The analysis flowchart provides an overview of the fractions of sequences surviving the various steps of the automated analysis. In this case about 20% of sequences were filtered during quality control. From the remaining 37,122,128 sequences, 53.5% were predicted to be protein coding, 5.5% hit ribosomal RNA. From the predicted proteins, 76.8% could be annotated with a putative protein function. Out of 32 million annotated proteins, 24 million have been assigned to a functional classification (SEED, COG, EggNOG, KEEG), representing 84% of the reads. labelfig:analysis-flowchart

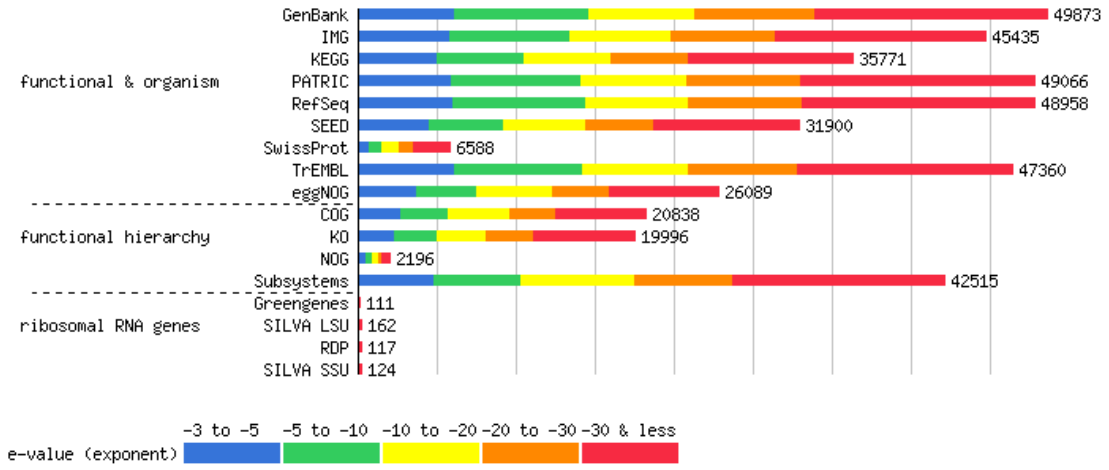


Figure 3.10: The graph shows the number of features in this dataset that were annotated by the different databases. The bars representing annotated reads are colored by e-value range. Different databases have different numbers of hits, but can also have different types of annotation data.

3.6.1.2 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 also mapped to the rRNA databases.

The graph 3.10 shows the number of features in this dataset that were annotated by the different databases. These include protein databases, protein databases with functional hierarchy information, and ribosomal RNA databases.

In addition this display will print the number of records in the M5NR protein database and in the M5RNA ribosomal databases.

3.6.1.3 yet more statistics in Technical Data

This part provides a quick links to a general statistical overview of the different analysis steps performed (see Analysis flowchart), a comprehensive list of all metadata for the data set, sequence length and GC distributions and a breakdown of blat hits per data source (e.g. hits to RefSeq [30], UniProt [21] or SEED [28]).

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.

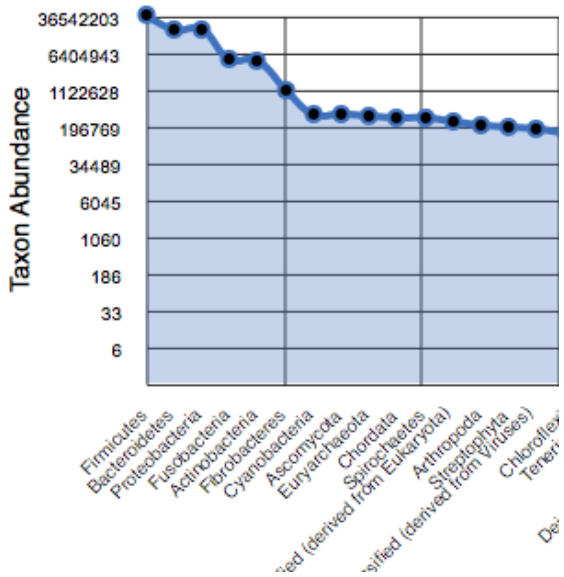


Figure 3.11: Organism breakdown: Sample rank abundance plot by phylum.

Metadata is presented in a searchable table which contains contextual metadata describing sample location, acquisition, library construction and sequencing using GSC compliant metadata. All metadata can be downloaded from the table.

3.6.2 The biological part of the Overview page – Organism Breakdown

The taxonomic hit distribution display breaks down 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, e.g. those hitting a specific taxonomic unit.

3.6.2.1 Rank abundance

The rank abundance plot ((Figure 3.11) provides a rank-ordered list of taxonomic units at a user-defined taxonomic level, ordered by their abundance in the annotations.

3.6.2.2 Rarefaction

The rarefaction curve of annotated species richness is a plot (see 3.12 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. When the rarefaction

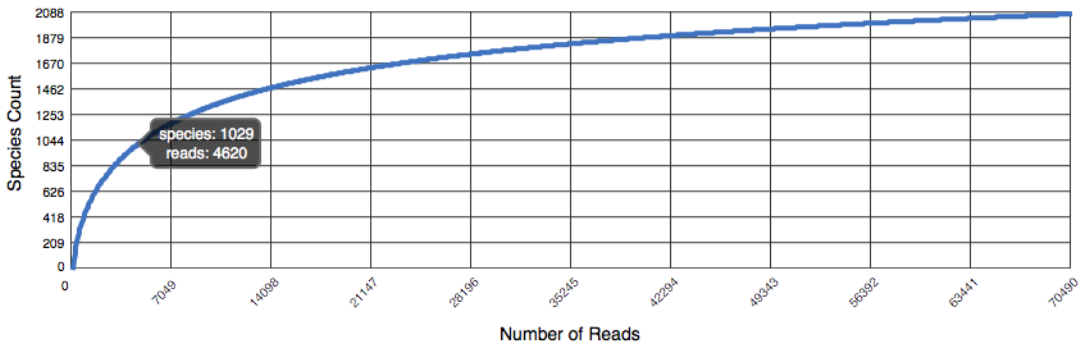


Figure 3.12: The rarefaction plot shows a curve of annotated species richness. This curve is a plot of the total number of distinct species annotations as a function of the number of sequences sampled.

curve is flat, more intensive sampling is likely to yield only 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 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 very quickly at first and then level off towards 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 the complete dataset.

3.6.2.3 Alpha Diversity

Finally 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 Figure 3.13).

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:

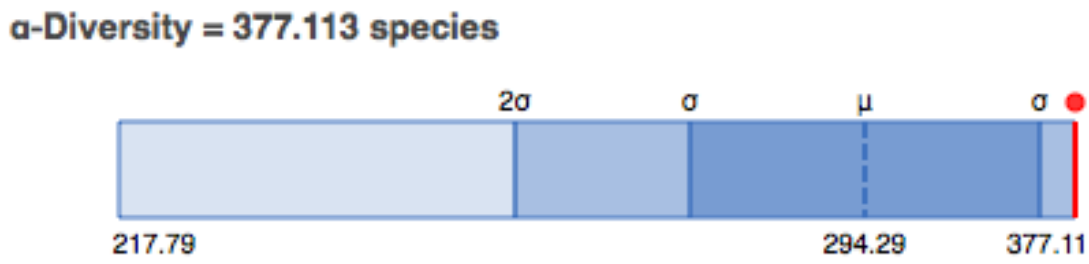


Figure 3.13: The alpha diversity plot shows the range of -diversity values in the project the data set belongs to. The min, max, and mean values are shown, with the standard deviation ranges (and 2) in different shades. The -diversity of this metagenome is shown in red.

$$\text{Richness} = 10^{-\sum_i p_i \log(p_i)}$$

where p_i are the proportions of annotations in each of the species categories.

Shannon species richess is the antilog of the Shannon index It has units of effective number of species. Each p is a ratio of the number of annotations for each species to the total number of annotations and m is the total number of different species annotations.

The species-level annotations are from all the annotation source databases used by MG-RAST. The table of species and number of observations used to calculate this diversity estimate can be downloaded under download source data on the overview page.

3.6.2.4 Functional Breakdown

This section contains four pie charts providing a breakdown of the functional categories for KEGG [16], COG [36], SEED Subsystems [28] and EggNOGs [15]. 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 via the Krona tool [27]integrated in the page.

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

We note that users can explore subsystems abundance on various levels using the analysis page.

QUESTION:: What are the default parameters used to display these graphs?? BEST HIT REP HIT?

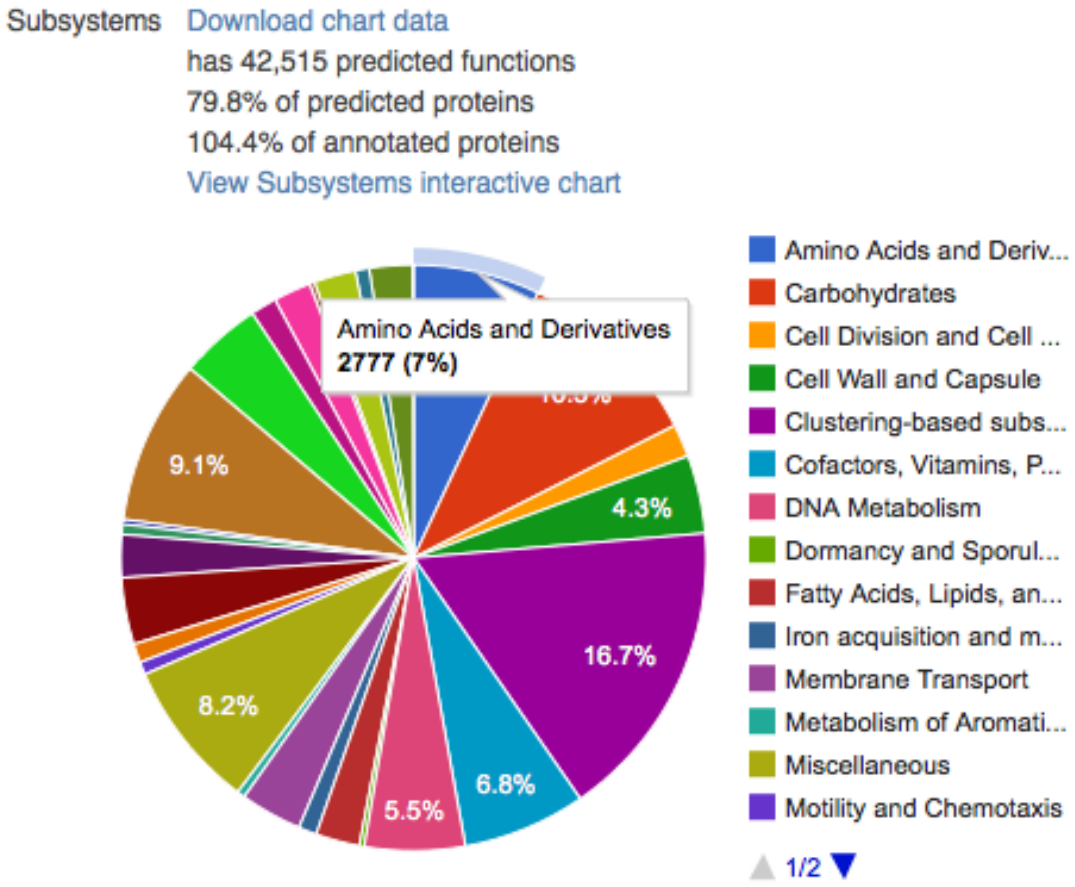


Figure 3.14: The glssubsystems function piechart classifies reads into the Subsystem level one functions. In contrast to the COG, EGGNOG and KEGG classification schemes there are over 20 top level subsystem categories creating a more highly resolved ”fingerprint” for the metagenome.

3.6.2.5 Taxonomic breakdown

This section contains 6 pie charts representing the break

3.7 Download page

The download page provides all publicly available data sets for download. Three types of data are available for download:

- metadata

  Data describing data in GSC compliant format.

- submitted data

  The original user submission

- analysis results

  The results of running the MG-RAST pipeline, the list includes all intermediate data products and is intended to serve as a basis for further analysis outside the MG-RAST pipeline.

  Details on the individual files are in A.

3.8 The Search Page

The search page is an additional way to find data sets in MG-RAST. (The other being the Metagenome Browse page.

The basic function of the Search page is to find data sets that 1) contain a search string in the metadata (data set name, project name, project description, GSC metadata), or 2) contain specific functions (e.g. SEED functional roles, SEED subsystems or GenBank annotations or 3) contain specific organisms. The default search uses all 3 kinds of data. In addition to a google like search that searches all data fields, we also provide specialized searches in one of the 3 data types. Figure 3.15 shows the result of a metadata search for ”oral health”.

Figure 3.16 shows the results from Figure 3.15 after sorting by metagenome ID.

3.9 The 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

Found 11 metagenomes containing 3 metadata matches.

map metagenomes | create collection | download results table | clear all filters

display 25 items per page

displaying 1 - 11 of 11

Metagenome	MG-RAST ID	Project	Public	Match Counts	Biome	Feature	Material	Country	Location	PI	...
		all	al		all	all	all	all	all	al	
NOCA_03P	4447102.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
NOCA_01P	4447192.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
February coral	4445755.3	Project for: February coral	1	1	animal-associated habitat	animal-associated habitat	animal-associated habitat	Australia	Nelly Bay Magnetic Island		
CA_06P	4447903.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
CA_06_1.6	4447971.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
CA_05_4.6	4447970.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
CA_04P	4447943.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
CA1_02P	4447101.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
CA1_01P	4447103.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
October coral	4445829.3	Project for: October coral	0	2	animal-associated habitat	animal-associated habitat	animal-associated habitat	Australia	Nelly Bay Magnetic Island		
October coral	4445756.3	Project for: October coral	1	2	organism-associated habitat	organism-associated habitat	organism-associated habitat	Australia	Nelly Bay Magnetic Island		

displaying 1 - 11 of 11

Figure 3.15: Searching for ”oral health” returns 11 data sets for two projects.

Found 11 metagenomes containing 3 metadata matches.

map metagenomes | create collection | download results table | clear all filters

display 25 items per page

displaying 1 - 11 of 11

Metagenome	MG-RAST ID	Project	Public	Match Counts	Biome	Feature	Material	Country	Location	PI	...
		all	al		all	all	all	all	all	al	
October coral	4445829.3	Project for: October coral	0	2	animal-associated habitat	animal-associated habitat	animal-associated habitat	Australia	Nelly Bay Magnetic Island		
October coral	4445756.3	Project for: October coral	1	2	organism-associated habitat	organism-associated habitat	organism-associated habitat	Australia	Nelly Bay Magnetic Island		
NOCA_03P	4447102.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
NOCA_01P	4447192.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
February coral	4445755.3	Project for: February coral	1	1	animal-associated habitat	animal-associated habitat	animal-associated habitat	Australia	Nelly Bay Magnetic Island		
CA_06P	4447903.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
CA_06_1.6	4447971.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
CA_05_4.6	4447970.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
CA_04P	4447943.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
CA1_02P	4447101.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	
CA1_01P	4447103.3	The oral metagenome in health and disease	1	1	human-associated habitat	human-associated habitat	human-associated habitat	Spain	Valencia	Mira	

displaying 1 - 11 of 11

Figure 3.16: The search results from the previous search sorted by projects

can be used to view these profiles for a single metagenome, or compare profiles from multiple metagenomes using various visualizations (e.g. heatmap) and statistics (e.g. PCoA , normalization).

The page breaks down in three parts following a typical workflow (Figure 3.17):

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 as explained in 2.6.

   We note that when choosing the LCA annotations, not all downstream tools are available. This is due to the fact that for the LCA annotations not all sequences will be annotated to the same level, it returns classifications on different taxonomic levels.

   Functional annotations can either be grouped into mappings to functional hierarchies or 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 which can be compared undividually, 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 using barcharts, trees, spreadsheet like tables , heatmaps, PCoA, rarefaction plots, Circular recruitment plot and KEGG maps.

The analysis page offers several hit classification schemes that are explained in 2.6.

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

Once a category is selected, the data browser underneath available metagenomes will display data of the selected category. The text field under available metagenomes displays the available data sets or group identifiers.

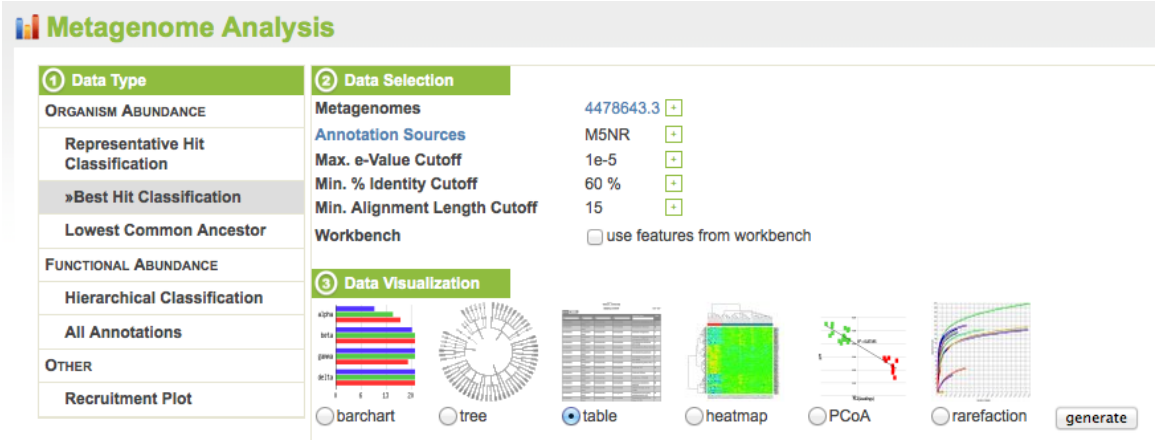


Figure 3.17: Using the analysis page is a three step process. First select a profile and hit (see below) type. Second select a list of metagenomes and set annotation source and similarity parameters. Third chose a comparison.

The use of MG-RAST identifiers (e.g. 4447971.3) is possible in the text field underneath ”available metagenomes”.

- the **private** data

  This will display a list of private or shared data sets for browsing under available metagenomes.

- the **collections**

  Collections are used defined sets of metagenomes grouped for easier analysis. This is the recommended way of working with the analysus page.

- the **projects**

  Projects are global groups of data sets grouped by the submitting user. The project name will be displayed

- the **public** data

  All public data sets will be displayed.

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

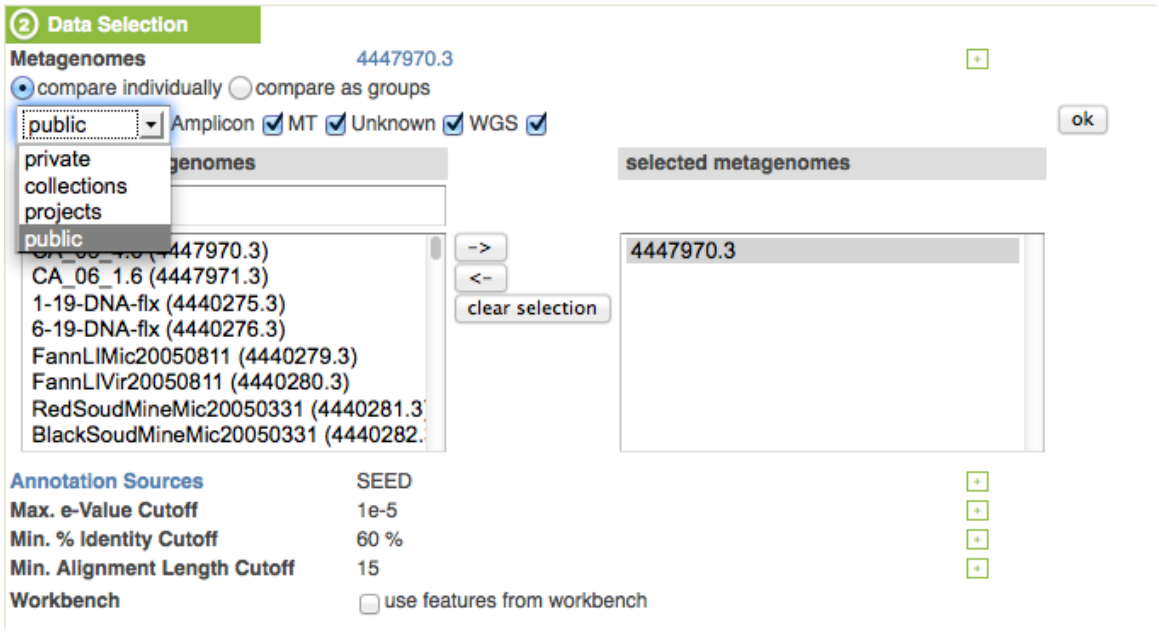


Figure 3.18: A view of the data selection dialogue, with the list of four data categories expanded.

3.9.1 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 you transform the data 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 additional downstream tests like 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.

The analysis page calculates the ordination visualizations with either raw or normalized counts, at the users option. The normalization procedure is to take

$normalized\_value\_i = log2(raw\_counts\_i + 1)$

And then the standardized values are calculated from the normalized values by subtracting the mean of each samples normalized values and dividing by the standard deviation of each samples normalized values.

$standardized\_i = (normalized\_i - mean(normalized\_i))/stddev(normalized\_i)$ You can read

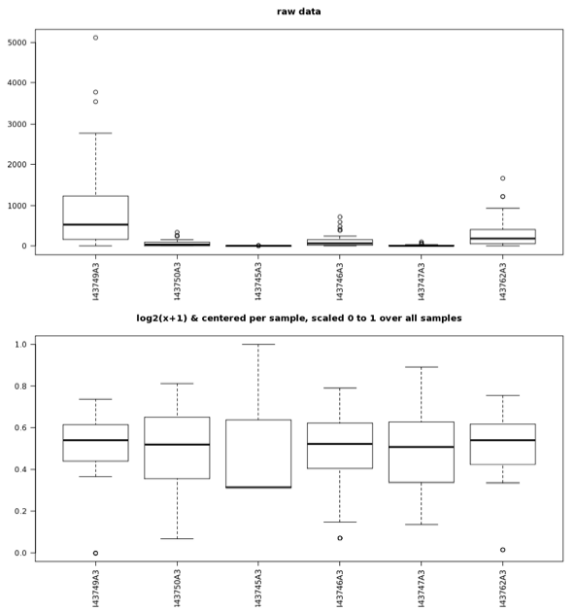


Figure 3.19: Boxplots of the abundance data for raw values (top) as well as values that have undergone the normalization and standardization procedure described above (bottom). It is clear that after normalization and standardization, samples exhibit value distributions that are much more comparable, and that exhibit a normal distribution; the normalized and standardized data are suitable for analysis with parametric tests, the raw data are not.

more about these procedures in a number of texts - We recommend Terry Speeds Statistical Analysis of Gene Expression in Microarray Data [35].

When data exhibit a non-normal, normal or unknown distribution, non-parametric tests (e.g. Man-Whitney or Kurskal-Wallis) should be used. Boxplots are an easy way to check and the MG-RAST analysis page provides boxplots of the standardized abundance values for checking the comparability of samples (Figure 3.19).

3.9.2 Rarefaction

The rarefaction view is only available for taxonomic data. The rarefaction curve of annotated species richness is a plot (see 3.20 of the total number of distinct species annotations as a function of the number of sequences sampled. As shown in 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 few additional species. The rarefaction curve is derived from the protein taxonomic annotations and is

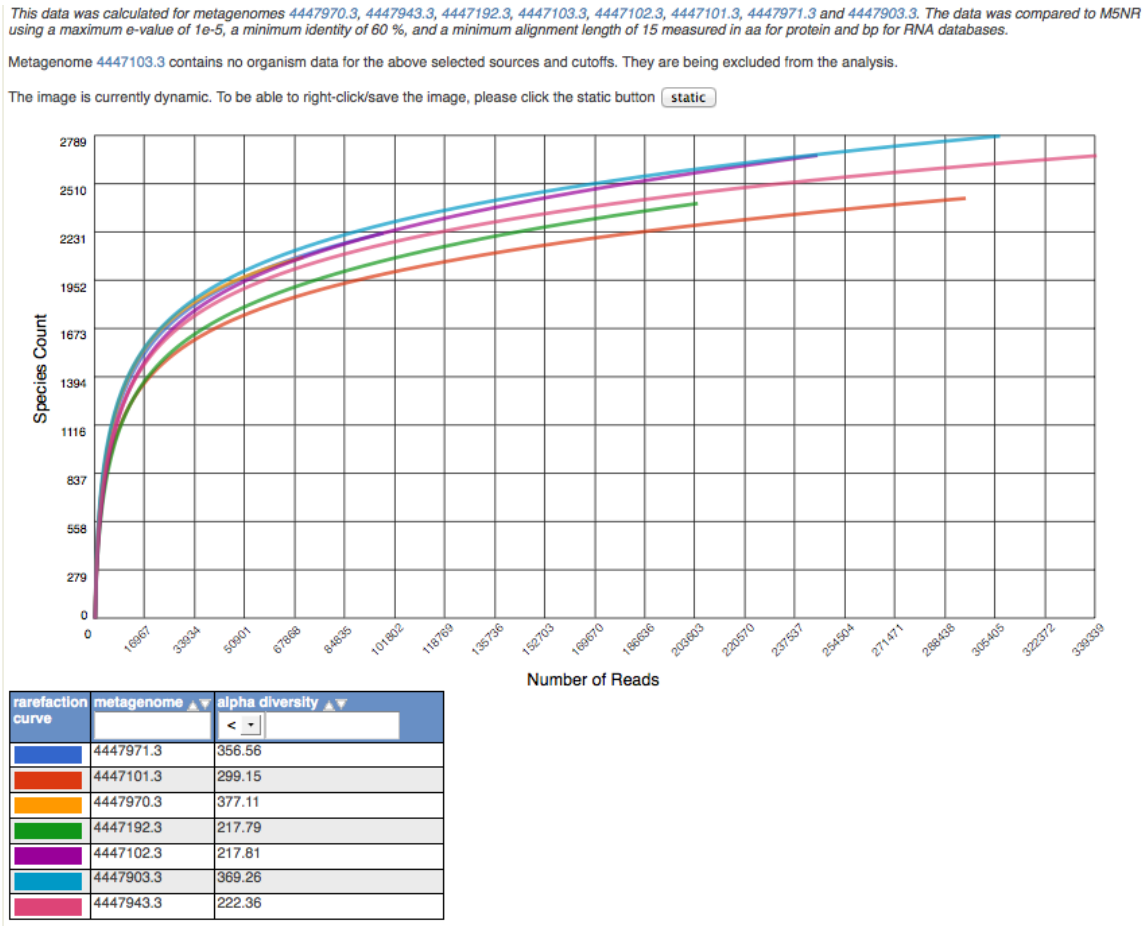


Figure 3.20: The rarefaction plot shows a curve of annotated species richness. This curve is a plot of the total number of distinct species annotations as a function of the number of sequences sampled.

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 sampling depth independent way.

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 very quickly at first and then level off towards an asymptote as fewer new species are found per unit of individuals collected. These rarefaction curves are calcu-

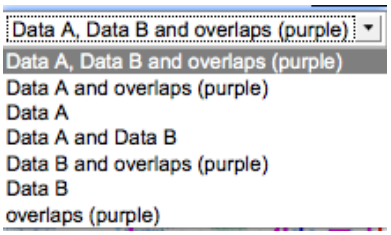


Figure 3.21: The options available for coloring the KEGG maps

lated from the table of species abundance. The curves represent the average number of different species annotations for subsamples of the the complete dataset.

3.9.3 The KEGG Mapper

The KEGG mapper is only available for functional data, it provides the ability to compare data sets visually on the basis of them mapped onto a KEGG pathway map. Users can select from any available KEGG pathway map.

Different colors indicate different metagenomic data sets.

The KEGG mapper works by providing two buffers that users can assign data sets to. After loading the buffers with the intended data sets, the KEGG mapper can highlight parts of the KEGG map that are present in the data set.

Several combinations of the two data sets can be displayed, those are shown in Figure 3.21.

The KEGG map tool allows the visual comparison of predicted metabolic pathways in metagenomic samples. It maps the abundance of identified enzymes onto a KEGG [16] map of functional pathways. Metagenomes can be assigned into one of two groups and those groups can be visually compared.

3.9.4 Recruitment plots

The recruitment page allows mapping of protein sequences in a single metagenome onto the complete genome sequences that are represented in the M5NR.

Once the metagenome is selected, the page will provide a list of genomes sorted by the amount of hits per genome is displayed for the user to choose from (see Figure 3.23).

A circular genome plot or a table will be printed. See Figure 3.24 for an example. The following elements are contained in the figure:

- outmost circle: forward strand genes (red: protein, black: RNA)

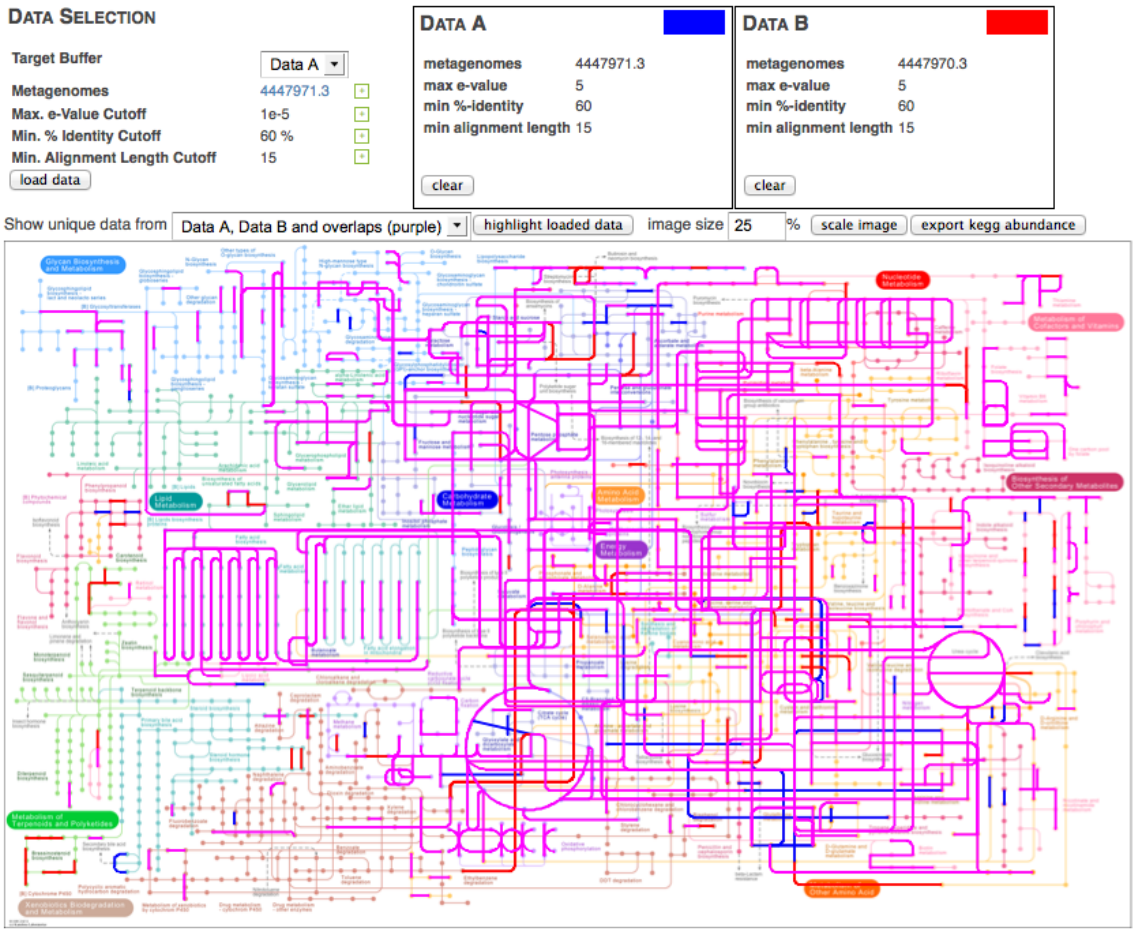


Figure 3.22: A comparison of two data sets using the KEGG mapper. Parts of metabolism common are shown in purple, unique to A are in blue, unique to B in red.

- 2nd circle: contigs for the reference genome
- 2rd circle: reverse strand genes (red: protein, black: RNA)
- innermost circle: abundance information (color coded for evalue)

The table view has the same information as the circular view and can easily be downloaded into a local spreadsheet. We use RefSeq [30] identifiers for the table as well as RefSEQ functions because the underlying contig information is present in the GenBank [3] downloads.

The recruitment plot uses the best hit approach.

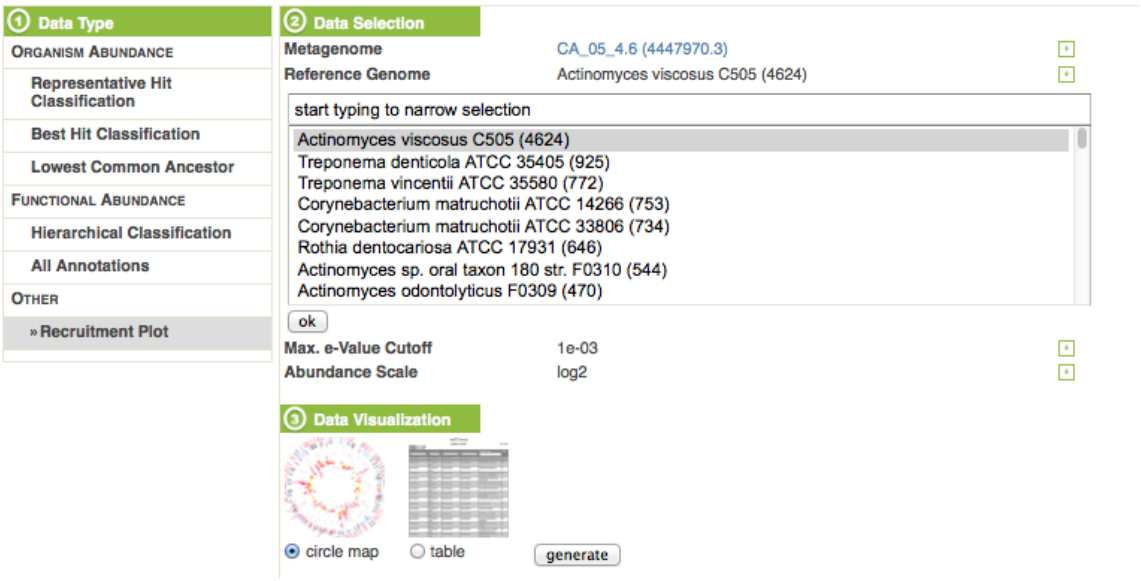


Figure 3.23: Selection of a genome to display sorted by number of hits per genome.

3.9.5 The Bar charts

Figure 3.25 shows the bar chart visualization option on the analysis page. One important property of the page is the built-in ability to drill down by clicking on a specific category. In this example we have expanded the domain Bacteria to show the normalized abundance (adjusted for sample sizes) of bacterial phyla. 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.

3.9.6 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; see Figure 3.27.

The tree offers another of capabilities via the options shown in Figure 3.28.

- export of a high resolution image

  For publication purposes we provide a SVG version of the image.

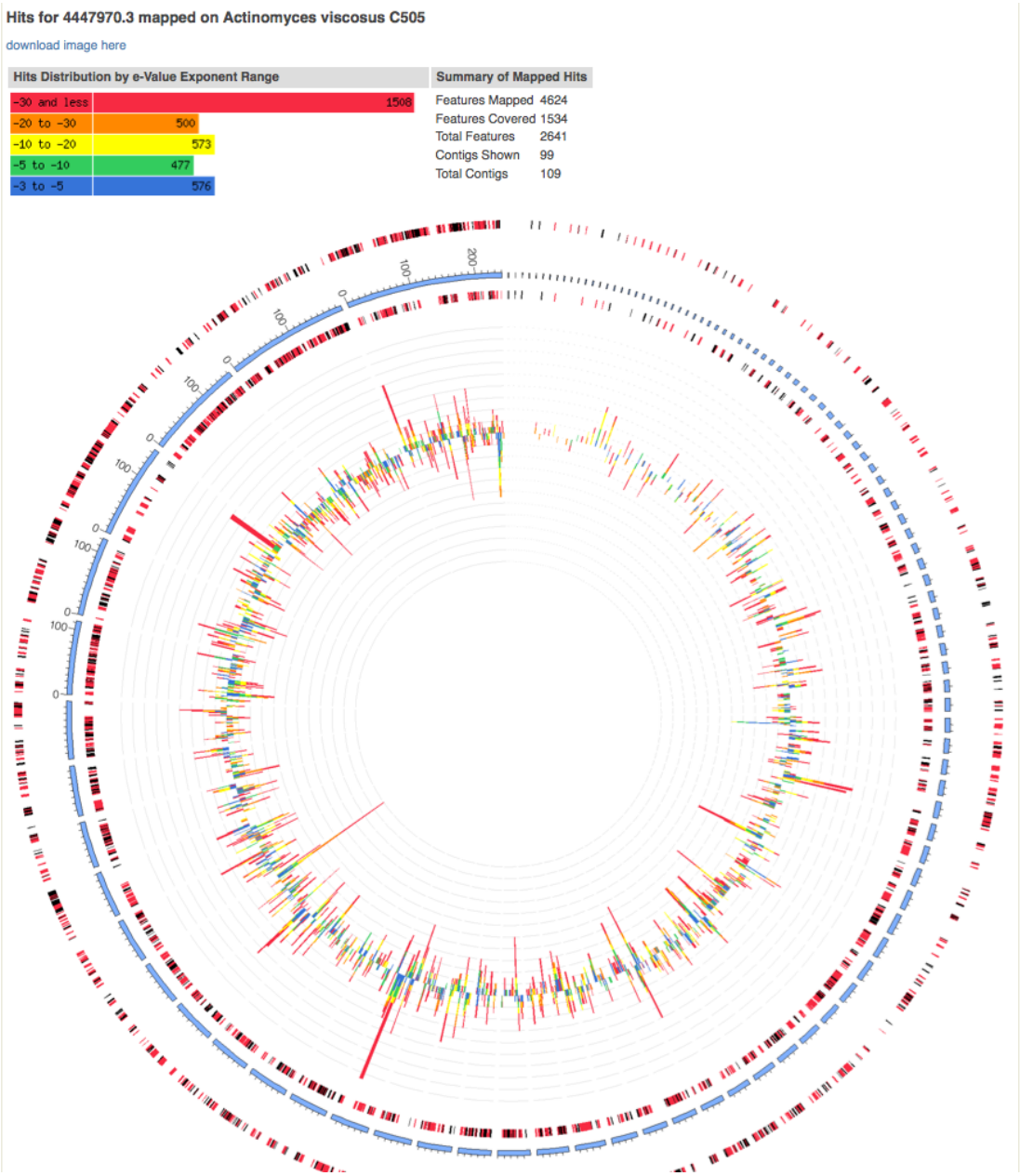


Figure 3.24: An example recruitment plot with the parameters from the previous Figure for *Actinomyces viscosus* C505.

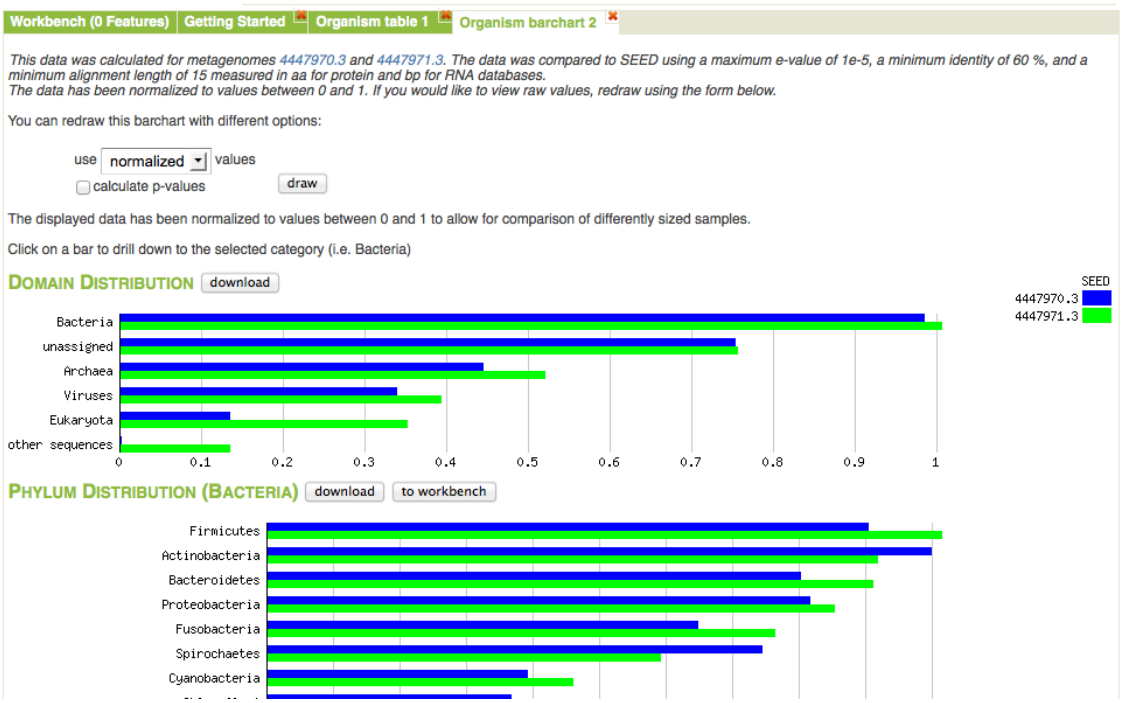


Figure 3.25: The bar chart view comparing normalized abundance of taxa. We have expanded the Bacteria domain to display the next level of the hierarchy.

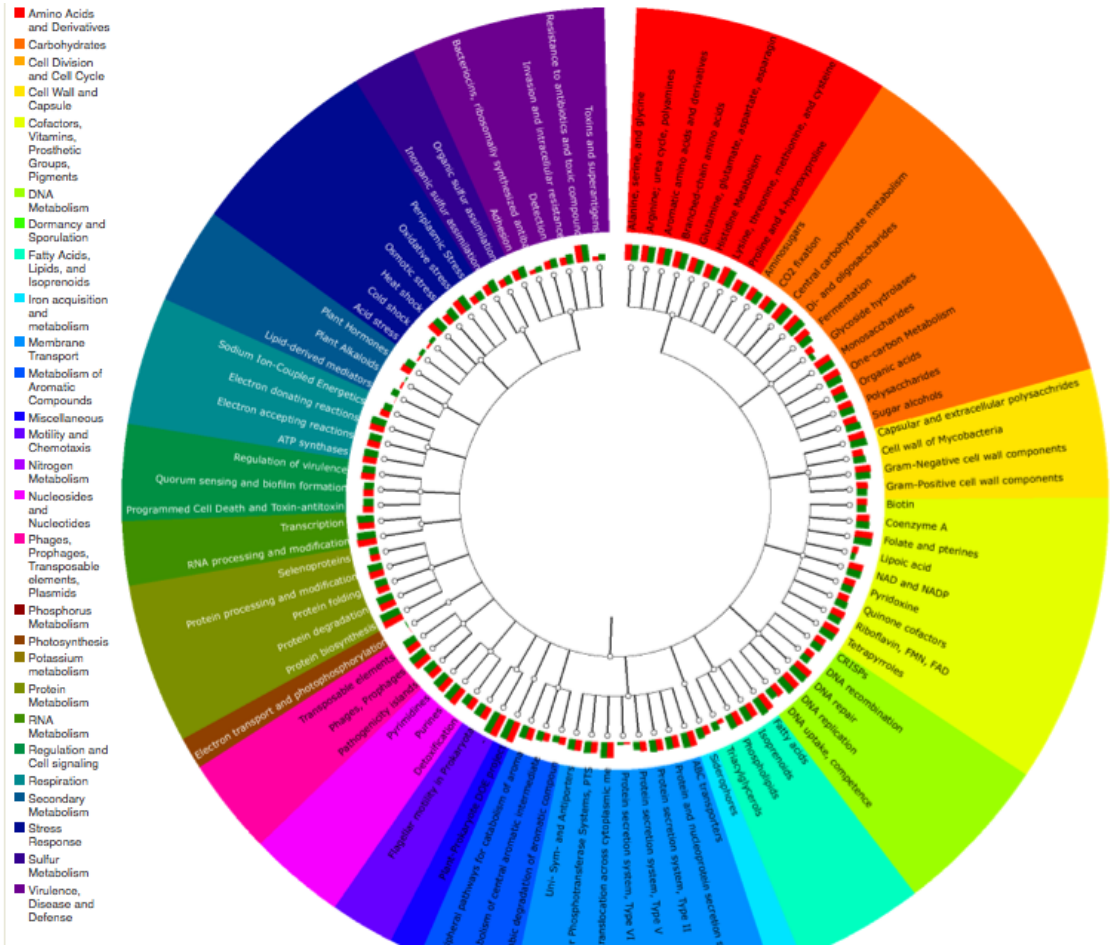
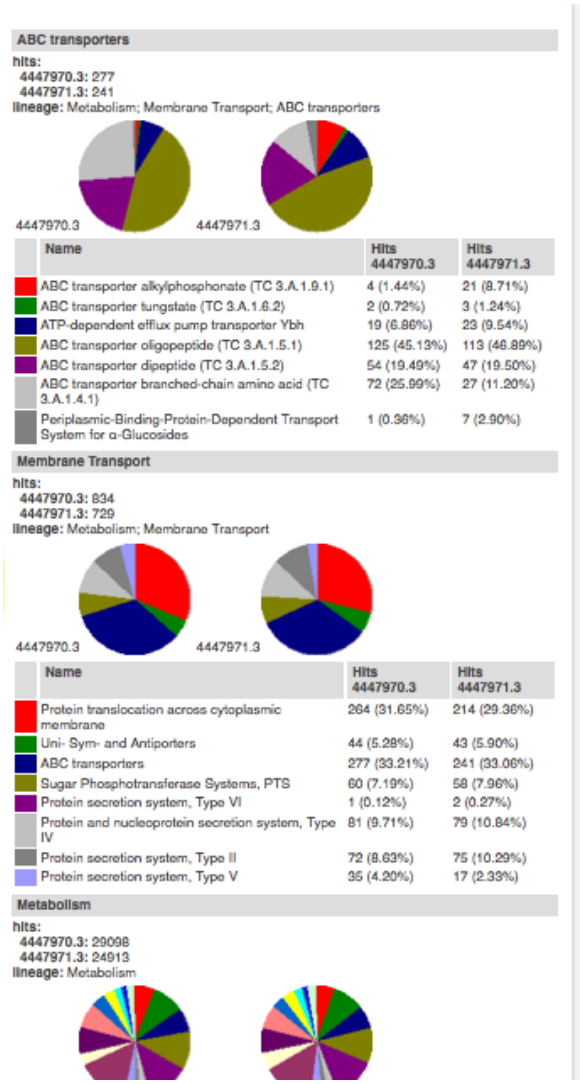


Figure 3.26: The tree diagram is a visualization method on the analysis page.



ABC transporters

hits:
4447970.3: 277
4447971.3: 241
lineage: Metabolism; Membrane Transport; ABC transporters

4447970.3 4447971.3

	Name	Hits 4447970.3	Hits 4447971.3
	ABC transporter alkylphosphonate (TC 3.A.1.9.1)	4 (1.44%)	21 (8.71%)
	ABC transporter tungstate (TC 3.A.1.6.2)	2 (0.72%)	3 (1.24%)
	ATP-dependent efflux pump transporter Ybh	19 (6.86%)	23 (9.54%)
	ABC transporter oligopeptide (TC 3.A.1.5.1)	125 (45.13%)	113 (46.89%)
	ABC transporter dipeptide (TC 3.A.1.5.2)	54 (19.49%)	47 (19.50%)
	ABC transporter branched-chain amino acid (TC 3.A.1.4.1)	72 (25.99%)	27 (11.20%)
	Periplasmic-Binding-Protein-Dependent Transport System for α-Glucosides	1 (0.36%)	7 (2.90%)

Membrane Transport

hits:
4447970.3: 834
4447971.3: 729
lineage: Metabolism; Membrane Transport

4447970.3 4447971.3

	Name	Hits 4447970.3	Hits 4447971.3
	Protein translocation across cytoplasmic membrane	264 (31.65%)	214 (29.36%)
	Uni- Sym- and Antiporters	44 (5.28%)	43 (5.90%)
	ABC transporters	277 (33.21%)	241 (33.06%)
	Sugar Phosphotransferase Systems, PTS	60 (7.19%)	58 (7.96%)
	Protein secretion system, Type VI	1 (0.12%)	2 (0.27%)
	Protein and nucleoprotein secretion system, Type IV	81 (9.71%)	79 (10.84%)
	Protein secretion system, Type II	72 (8.63%)	75 (10.29%)
	Protein secretion system, Type V	35 (4.20%)	17 (2.33%)

Metabolism

hits:
4447970.3: 29098
4447971.3: 24913
lineage: Metabolism

Figure 3.27: Clicking on a node in the tree diagram will display addition information to the right of the tree display.

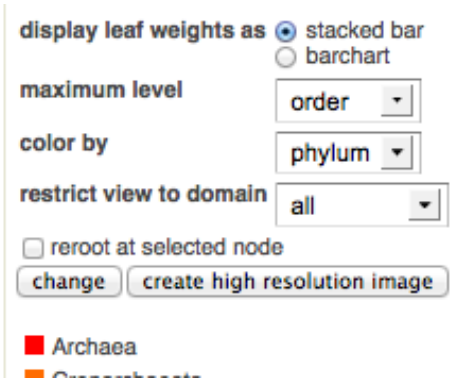


Figure 3.28: The options for the tree view.

- re-rooting

  The tree display allows zooming in by re-rooting the tree display. For this select an node inside the tree (turning it red) and select reoot at the selected node. See Figure **??**).

- bar chart or stacked chart

  The abundances of the hierarchy entries can be displayed as bar charts per node or as a stacked graph.

- restrict to domain

  Is identical to re-rooting the tree for a specific domain.

- maximum level

  This setting determines the depth of the tree being displayed.

- color by

  Determines the color (if any) used for the out circle of the display.

  Figure 3.29 shows the result of changing the display depth and coloring options. The color is used to group organism visually into order level groups.

3.9.7 Heatmap/Dendrogram

The heatmap/dendrogram (Figure 3.30) is a tool that 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

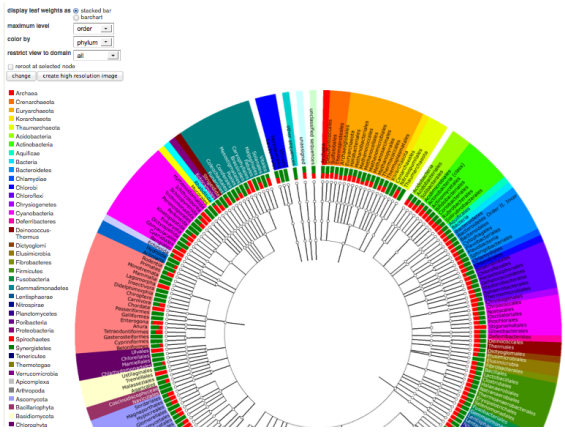


Figure 3.29: A tree view at order level with coloring set to phylum level.

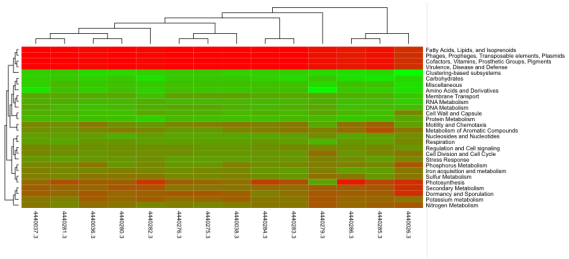


Figure 3.30: Heatmap/dendogram example in MG-RAST. The MG-RAST heatmap/dendrogram has two dendrograms, one indicating the similarity/dissimilarity among metagenomic samples (x axis dendrogram) and another to indicate the similarity/dissimilarity among annotation categories (e.g., functional roles; the y-axis dendrogram).

two dendrograms, one indicating the similarity/dissimilarity among metagenomic samples (x axis dendrogram) and another to indicate 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.

The barchart and tree tools map raw or normalized abundances onto functional or taxonomic hierarchies. The barchart tool presents mapping onto the highest category of a hierarchy (e.g. Domain) and allows a drill down into the hierarchy. In addition reads from a specific level can be added into the workbench.

3.9.8 Ordination

MG-RAST uses Principle Coordinate Analysis (PCoA) to reduce the dimensionality of comparisons of multiple samples that consider functional or taxonomic annotations.

PCoA is a well known method for dimensionality reduction of large data sets. 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 2 or 3 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 further 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 etc.) 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 to identify individual sample features (taxa or functions) that drive correlations observed in PCoA visualizations. This can be accomplished with permutation based statistics applied directly to the data before calculation of distance measures used to produce PCoAs, or by applying conventional statistical approaches (e.g. ANOVA or Kruskal-Wallis test) to groups observed in PCoA based visualizations.

3.9.9 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 using Krona [27], can be exported in BIOM [23] format to be used in other tools, e.g. QIIME [5] or the tables 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.

We show how to use the taxonomic information derived from an analysis of protein similarities found for the data set 4447970.3. We are using best hit classifcation, SEED, $10\hat{5}$, 60% identity and minimal alignment length of 15 amino acids and selecting table output. The resulting table output is shown in Figure 3.31.

The following control elements are connected to the table:

- group by

  This allows summarizing entries below the level chosen here to be subsumed.

- download table

  This will download the entire table as a spreadsheet.

- Krona

  This will invoke KRONA [27] with the table data.

- QIIME

  Creates a BIOM [23] format file with the data being displayed in the table.

- table size

  Changing the number of elements to display for the web page.

Below we explain the columns of the table and the functions available for them. For each column we allow sorting the table via clicking on the upwards and downwards pointing triangles.

- metagenome

  In the case of multiple data sets being displayed, this column allows sorting by metagenome ID or selecting a single metagenome.

- source

  Displays the annotation source for the data being displayed.

- domain

  The domain column allows sub selecting from Archaea, Bacteria, Eukarya and Viruses.

This data was calculated for metagenome 4447970.3. The data was compared to SEED 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.

group table by class change

available plugins

krona graph QIIME report

download this table

display 15 items per page

displaying 1 - 15 of 54 next» last»

metagenome	source	domain	phylum	class	abundance	avg eValue	avg % ident	avg align len	# hits	to workbench
4447970.3	SEED	all			<	<	<	<	<	
4447970.3	SEED	Archaea	Crenarchaeota	Thermoprotei	8	-12.00	68.08	51.60	7	
4447970.3	SEED	Archaea	Euryarchaeota	Archaeoglobi	2	-5.00	66.67	36.00	2	
4447970.3	SEED	Archaea	Euryarchaeota	Halobacteria	3	-8.50	68.16	43.25	3	
4447970.3	SEED	Archaea	Euryarchaeota	Methanobacteria	31	-14.75	67.90	55.97	16	
4447970.3	SEED	Archaea	Euryarchaeota	Methanococci	9	-10.39	68.21	49.39	8	
4447970.3	SEED	Archaea	Euryarchaeota	Methanomicrobia	59	-10.01	68.94	47.00	59	
4447970.3	SEED	Archaea	Euryarchaeota	Thermococci	15	-17.69	66.70	65.76	15	
4447970.3	SEED	Archaea	Euryarchaeota	Thermoplasmata	4	-5.00	68.10	33.40	4	
4447970.3	SEED	Archaea	Thaumarchaeota	unclassified (derived from Thaumarchaeota)	3	-10.00	60.04	50.67	3	
4447970.3	SEED	Bacteria	Acidobacteria	Solibacteres	19	-15.24	71.74	57.14	19	
4447970.3	SEED	Bacteria	Acidobacteria	unclassified (derived from Acidobacteria)	13	-21.64	62.83	78.64	13	
4447970.3	SEED	Bacteria	Actinobacteria	Actinobacteria (class)	13339	-20.36	70.30	69.27	8111	
4447970.3	SEED	Bacteria	Aquificae	Aquificae (class)	23	-19.74	67.38	69.09	23	
4447970.3	SEED	Bacteria	Bacteroidetes	Bacteroidia	2350	-27.79	77.40	75.62	1718	
4447970.3	SEED	Bacteria	Bacteroidetes	Cytophagia	102	-13.50	68.32	53.87	102	

displaying 1 - 15 of 54 next» last»

Figure 3.31: A view of the analysis page table.

- phylum, class

  Since we have selected to group results at the class level only phylum and class are being displayed. The text fields in the colum headers allow subsection (e.g. by entering Acidobacteria or Actinobacteria in the phylum field). The searches are performed inside the web browser and are very efficient.

  Any subselection will narrow down all data sets being displayed in the table.

  Users can elect to have the results grouped by other taxonomy levels, e.g. genus, creating more colums in the table view.

- abundance

  The number of sequences found with the parameters selected matching this taxonomic unit. (Not the parameters chosen are displayed on top of the table). Clicking on the abundance displays another page displaying the BLAT alignments underlying the assignments.

  The abundance is calculated by multiplying the acutal number of database hits found for the clusters by the number of cluster members.

- avg. evalue, avg percent identity, average alignment length

  The average values for evalue, percent identity and alignment length

- hits

  The number of clusters found for this entity (function or taxon) in the metagenome.

- ...

  Allows extending the table to add additional columns.

3.9.10 The workbench

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 the fact that data sent to the workbench only exist until you close your current session.

The workbench was designed to allow users to select subsets of the data for comparison or export.

Chapter 4

User Manual

4.1 Privacy, identifiers, sharing and publication

Data in MG-RAST is private unless published to everyone or shared with specific users by the submitter.

Once data is submitted to the pipeline a unique identifier is assigned (see 2.8.4 for details).

The web interface allows sharing and publication of data, requiring the presence of minimal metadata (see 2.8.1). Data can only be shared once the computation has finished.

4.2 Uploading to MG-RAST

MG-RAST was designed to allow users to upload sequence data directly from next generation sequencing machines. Data can be in FASTA, FASTQ or SFF format.

All uploaded sequence files must have one of the following extensions:

- .fasta
- .fna
- .fastq
- .fq
- .sff

Compressing large files will reduce the upload time and the chances of a failed upload, you can use Zip (.zip) and gzip (.gz) as well as tarred gzipped files (.tgz).

We suggest you upload raw data (in FASTQ or SFF format) and let MG-RAST perform the quality control step as this will allow us to identify any issues with the sequencing run. Frequently local quality control will identify some issues but mask others.

It is not neccesary to assemble data prior to upload to MG-RAST, the system has been optimized for short reads and can handle uploads of many hundred gigabytes.

4.2.1 Assembled data with read abundance info

For assembled data (in FASTA format) uploaded to MG-RAST read abundance information for contigs can be imported as well. The assembled option for the pipeline will attempt to retrieve read abundance information from the sequence files using the following simple format:

```
>sequence_number_1_[cov=2]
CTAGCGCACATAGCATTCAGCGTAGCAGTCACTAGTACGTAGTACGTACC
>sequence_number_2_[cov=4]
ACGTAGCTCACTCCAGTAGCAGGTACGTCGAGAAGACGTCTAGTCATCAT
....
```

The abundance information must be appended without spaces to the end of the sequence name (also without whitespace) in the format _[cov=n] where n is the coverage or abundance of each contig.

4.2.2 Steps for submission via the web interface

To start uploading data to MG-RAST through the website, click on the up arrow, this opens the Upload page.

On this page you can upload files, modify the files where needed, add metadata and finally submit for analysis.

The page is split into two sections, Prepare Data to upload, manipulate and assemble all the files required for a submission and Submission to create the MG-RAST job(s), set analysis parameters and start the analysis. Each Section contains subsections which you can click to expand.

4.2.2.1 Prepare Data

4.2.2.1.1 Download metadata spreadsheet template The best time to enter metadata for a dataset is at this stage and we provide a spreadsheet template which can be filled out with all the available information. The metadata can be modified at a later date to add information as it

Technology	Rate (bit/s)	Time for 1GB Upload
Modem 14.4 (2400 baud)	14.4 kbit/s	154 hours
ADSL Lite	1.5 Mbit/s	1.5 hours
Ethernet	10 Mbit/s	13.33 minutes
T3	44.736 Mbit/s	3 minutes
Fast Ethernet	100 Mbit/s	1.33 minutes

Table 4.1: Summary of upload times

becomes available or to correct errors. While the number of fields in the template are large, the number of required fields, labelled in red in the template, is small. The template file can be used to upload metadata for one or multiple samples and submit them to MG-RAST as a single project.

4.2.2.2 Upload files

All files uploaded to MG-RAST should be named using alphanumeric and .-_ characters without spaces. Files larger than 50MB should be compressed before upload using gzip (preferable) or Zip, this will reduce the time taken for the upload of the file which in turn reduces the chance that the upload will fail.

Types of files: Sequence files FASTA, FASTQ or SFF formats Metadata files filled out spreadsheet Barcode files The barcode file should be plain text ASCII containing lines with a barcode sequence followed by a unique filename separated by a tab, with as many lines as necessary for the barcodes in the sequence file you are submitting.

Click on the Browse button to select the file or files and the upload will begin automatically after the files are selected.

4.2.2.2.1 Uploading For the actual uploading we use an HTML5 feature [40] that will automatically break up the files into chunks on the client side and send them. Note: This is one of the reasons we request that you use a recent version of Firefox as older versions might be slower.

Below we provide a summary of observed upload times that might help adjust exectations on how long the upload should take.

Based on observed values, upload times per 1GB (10**9 bytes) vary from 2 minutes to over an hour with typical times being 10 to 15 minutes. Your experience will vary depending on the speed of your connection to the internet and the quality of service in your region.

In practice the time taken will be more than the figure above.

4.2.2.2.2 Verifying the integrity of the uploaded files. When the upload of your files has completed, you will be prompted with the MD5-sum of each file. You should generate an md5 sum for each uploaded file on your machine, paste it into the appropriate box in the prompt and click the ”check” button. A popup will show you whether there is a match or not. Additionally the check button will turn green upon success and red upon failure. Click the ”Close” button if you have completed the checks or if you wish to skip this step.

Checking the integrity especially of large files is important, because it will give you immediate feedback about whether your upload was successful or not. If not detected at upload time, a damaged file will lead to errors later in the pipeline, wasting both valuable compute cycles and even more importantly, your time. To generate an MD5 sum of your file you can use the ”md5” shell command on a Mac, the ”md5sum” shell command on Unix systems or use freely available md5-sum tools on windows, e.g. from http://www.winmd5.com/.

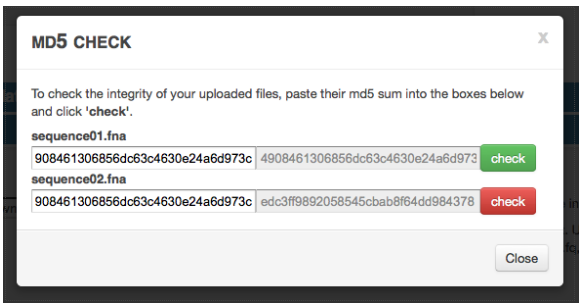


Figure 4.1: A dialogue will request the user put in locally generated MD5 checksum for the files to identify any data curruption during the upload.

4.2.2.2.3 File filters in place for uploading As MG-RAST has been designed to work with metagenomic and metatranscriptomic data sets, there is a filter in place trying to identify data sets not suitable for MG-RAST. Those data sets will be painted in red and cannot be submitted.

Below we list the criteria for rejection

- protein sequences

  MG-RAST is optimized to perform translation from DNA to proteins.

- reads shorter than 75 basepairs

  The gene prediction stage performance deteriorates significantly with shorter reads.

- genomes

Submissions with complete genomes or a small number of contigs are rejected as well. Here our sister service RAST at `http://rast.nmpdr.org` should be used instead of MG-RAST.

- files that are too small (sequence data less than 1Mbp)

  Files that are too small for MG-RAST to properly function are rejected at the submission stage. There is a minimal size requirement of 1 metagebase.

- corrupted files

  If he number of unique identifiers is not matching the number of sequence records in a file.

- alignments

  We cannot identify proteins from sequences containing alignment information.

- Colorspace

  The tool chain does not function for ABIsolid sequences in colorspace, please translate to standard FASTA.

- rar compressed files and Zip files over 4GB

  We cannot decompress these files.

In addition we will filter at the upload stage any Word documents, Rich Text Format files and all files without the extension .fna, .fasta, .fq, .fastq or .sff in their name.

Please note: We recommend computing an MD5 checksum and verifying that the checksum computed by MG-RAST is identical to the locally computed checksum. This is the best way to ensure data integrity.

4.2.2.2.3.1 De-Multiplexing for 454 and similar data sets MG-RAST performs the de-multiplexing based on the presence of the barcode sequence at the beginning of the reads.

Assuming you have a sequence file testseq.fasta and your barcode file has tab-separated lines like:

```
AAAAAAAA      fileA
CCCCCCCC      fileC
```

The demultiplexing step will split your sequence file into three files: fileA.fasta containing all the reads that begin with AAAAAAAA, fileC.fasta containing all the reads that begin with CCCCCCCC, and testseq_no_MID_tag.fasta containing reads which do not match either of the two.

We note that demultiplexing for Illumina needs to be done outside the MG-RAST system. Illumina barcodes work differently from 454 barcodes.

4.2.2.2.4 Managing the inbox All files uploaded to MG-RAST will be displayed in your inbox and you can perform certain functions operations on them. Compressed and/or archived files can be unpacked, SFF files can be converted to FASTQ and sequence data can be demultiplexed using barcodes contained in uploaded files, files can also be deleted. When a sequence file is selected some information about the sequence data is displayed. It is a good idea to check that the uploaded files in your inbox match your local copies. The file MD5 checksum, file size, sequence count, and basepair count can be used for this purpose.

4.2.2.3 Submission

The submission process allows you to create MG-RAST jobs using files in your inbox and is designed to facilitate the creation of a large number of jobs easily.

4.2.2.3.1 Select metadata file We recommend you supply metadata for all your samples and will assign a higher processing priority to samples with metadata and which will be made public. Metadata files which will have successfully passed out validation step will be displayed for selection. READ more on how to create a valid metadata file.

4.2.2.3.2 Select project All jobs created in MG-RAST will need to be placed in a project which can be an existing project or a new one created by you during the submission. The project can be specified one of 3 ways: in the metadata file if supplied, selected from the existing projects you have access to, or a new project name can be entered into the text box.

4.2.2.3.3 Select sequence file(s) All the sequence files in your inbox will be displayed for selection and job submission, each sequence file can be used to create a single MG-RAST job.

4.2.2.3.4 Choose pipeline options The MG-RAST analysis can be influenced by the options selected here which affect dereplication, screening and quality filtering of the reads. The options selected are applied to all the sequence files selected.

4.2.2.3.5 Submit job This is the final step after which the analysis pipeline takes over and the processing begins. Once the job or jobs have been submitted all the files required to create them will be removed from your inbox.

You can monitor the progress of your jobs in ”My Data Summary” on the Browse Metagenomes page.

4.2.3 The cmd-line uploader

The following syntax will allow uploading to MG-RAST from the command line.

```
curl -H "auth: webkey" -X POST -F "upload=@/path_to_file/metagenome.fasta" '
```

where you need to substitute webkey with the unique string of text generated by MG-RAST for your account. Your webkey is valid for a limited time period and ensures that the uploads you perform from the command line are recognized as belonging to your MG-RAST account and placed in the correct inbox.

4.2.4 Managing the Inbox

The Inbox is a temporary storage location for sequence and metadata files prior to submission to the pipeline. To protect us from any misuse of the facility, we have limited the Inbox to metadata spreadsheets and sequence files.

Files are visible only to the uploading user and will automatically be deleted after 72 hours.

4.2.4.1 File processing options in the Inbox

- unpack selected Unpacks selected zip, gzip, or tar files.
- convert sff to fastq Converts selected sff files to fastq format. Only FASTQ and FASTA files can be submitted to the system.
- demultiplex Demultiplexes selected files.

  Note that this is only suitable for 454 type barcodes that are actual pre-fixes of the reads. This approach does not work for the Illumina barcode approach (basically a third read for each paired end read).
- join paired ends Joins overlapping paired-end reads.

Please note: After the actual upload is complete the system will compute the statistics shown in Figure 4.3. Computing this information takes some time, so your data will not immediately be visible after you uploaded it.

You can unpack, delete, convert and demultiplex files from your inbox below. Metadata files will automatically appear in the '*select metadata file*' section below. Sequence files will automatically appear in the '*select sequence file(s)*' section below after sequence statistics are calculated (may take anywhere from seconds to hours depending on file size).

Filenames in gray in your Inbox are undergoing analysis and cannot be moved or submitted to a different process until analysis is complete. Filenames in red have encountered an error.

File Processing Operations

unpack selected	Unpacks selected zip, gzip, or tar files.	demultiplex	Demultiplexes selected files.
convert sff to fastq	Converts selected sff files to fastq format.	join paired-ends	Joins overlapping paired-end reads.

200_2M_09Nov09_lane_8_1.txt
200_2M_09Nov09_lane_8_1.txt.zip
E4GC80A02.fastq
E4GC80A02.sff
E4GC80A02.sff.fasta
E4GC80A02.sff.qual
E4GC80A02.xml
(uploading) E8N68KH02.sff
metadata_spreadsheet_biogas_reads.xls
MGRAST_MetaData_template_1.0.xlsx

Directory Management Operations

update inbox	Refreshes the contents of your inbox.
move selected	Moves the selected files into or out of a directory.
delete selected	Deletes the selected files.
create directory	Creates a new directory in your inbox.
delete directory	Allows you to select and delete an empty directory.

Figure 4.2: The Inbox provides temporary storage before submitting data and limited editing features.

File Information

sequence content	DNA
unique id count	118196
sequence type	WGS
standard deviation gc content	4.387
standard deviation length	23.360
standard deviation gc ratio	0.101
sequencing method guess	454
bp count	29996553
ambig sequence count	10432
length max	509
suffix	1
file size	59.5 MB
ambig char count	31973
sequence count	118196
length min	50
average gc content	69.145
average gc ratio	0.452
average ambig chars	0.271
file checksum	d5f9cdd37554e6a858c84154aa0d2047
average length	253.787
type	ASCII text
creation date	2013 May 09 08:44:33
file type	fastq

Figure 4.3: The information displayed by the inbox for one file (once selected).

4.2.4.2 Directory management operations for the Inbox

- update inbox

  Refreshes the contents of your inbox.

- move selected

  Moves the selected files into or out of a directory.

- delete selected

  Deletes the selected files.

- create directory

  Creates a new directory in your inbox.

- delete directory

  Allows you to select and delete an empty directory.

Users should always double check the MD5 checksum for files that are uploaded to the system to verify the integrity. Figure 4.3 shows the MD5 finger print that is computed upon upload for each file.

4.2.5 Generating metadata for the submission

MG-RAST uses questionnaires to capture metadata for each project with one or more samples. Users download and fill out the questionnaire, then submit it. Questionnaires are validated for completeness and compliance with the controlled vocabularies for certain fields automatically by MG-RAST.

MG-RAST has implemented the use of Minimum Information about any (X) Sequence (MIxS) [44] developed by the Genomic Standards Consortium (GSC). In addition to the minimal checklists, more detailed data can be captured in optional environmental packages.

We use simple spreadsheets to capture metadata, with a minimal number of required fields (in red in the spreadsheets) and a number of optional fields. The spreadsheet is separated into multiple tabs representing the different metadata categories. The MG-RAST metadata spreadsheet template is available on the MG-RAST upload page or here `ftp://ftp.metagenomics.anl.gov/data/misc/metadata/MGRAST_MetaData_template_1.3.xlsx`.

A filled out version of the spreadsheat is available here: `ftp://ftp.metagenomics.anl.gov/data/misc/metadata/MGRAST_MetaData_template_example.xlsx`.

In Figure 4.4 we show the template tab for project and the required field labels (in red); in essence your contact information.]

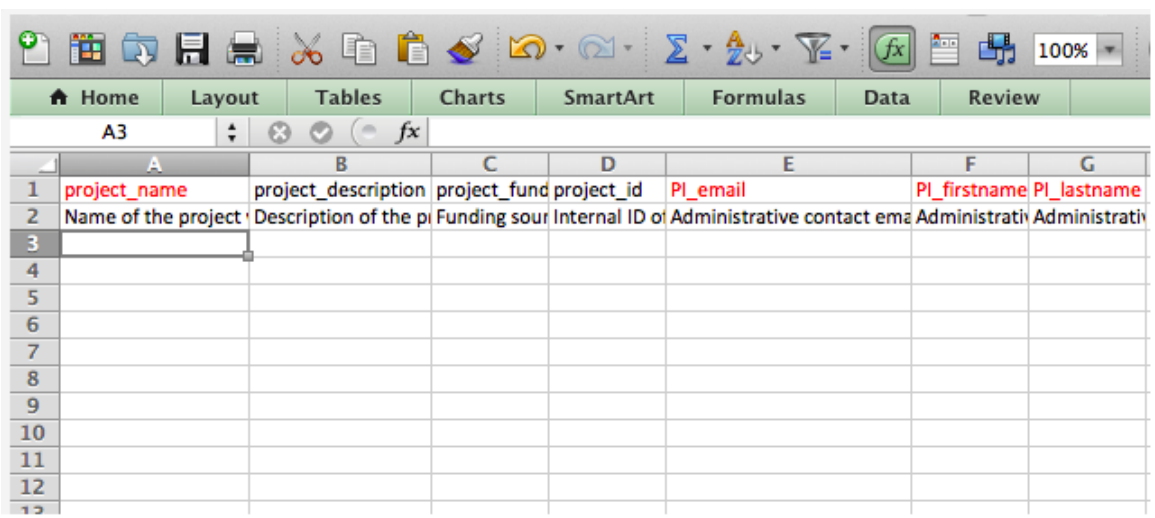


Figure 4.4: The project spreadsheet. In red are required fields. Note that the 2nd row contains information on how to fill out the form.

Note: Use the third line in the spreadsheet and below as shown in Figure 4.6 to enter your data. Do not attempt to alter the first two lines or delete them, they are read only. The first line contains

Figure 4.5: The various tabs in the spreadsheet. Project, sample and one of library metagenome or library mimarks survey are required.

the field labels and second line contains descriptions that can help explain how to fill out the fields, along with what unit to use, e.g. temperature in Celsius and distance in meters.

	A	B	C	D	E	F	G	H
1	sample_nam	sample_id	latitude	longitude	continent	country	location	depth
2	Unique nam	Internal ID o	The geograp	The geograp	The geograp	The geograp	The geograp	Depth is
3	sample1							
4	sample2							
5	sample3							
6								
7								

Figure 4.6: The sample tab with 3 new samples (sample1, sample2 and sample3) added. Again red text in the first row indicates required fields. Rows 1 and 2 cannot be altered.

Required sheets You need to fill out four sheets to describe your metadata:

1. project

   This sheet has only one row, and describes a set of samples uploaded together; the other sheets have one row per sample

2. sample

   This sheet includes either the filename or metagenome name used for matching

3. library

   either metagenome (for WGS and WXS) or mimarks-survey (for 16s amplicon)

4. environmental package

   at least one of the environmental packages of suggested standard metadata. Choose the package that best describes your dataset (e.g. water, human-skin, soil)

The sample section (below) requires minimal information (including the sample name) about where and when the sample was taken. Note that some fields in the spreadsheet must be filled out with terms from a controlled vocabulary or in a certain way. Country and environment (biome, feature, material) fields require entries from curated ontologies, gazetteer and environmental ontology respectively.

The sample tab with 3 new samples (sample1, sample2 and sample3) added. Again red text in the first row indicates required fields.

Mandatory fields

- country United States of America, Netherlands, Australia, Uruguay
- latitude and longitude 106.84517, -104.60667, 28 42.306N, 88 24.099W, 45.30 N, 73.35 W
- biome Small lake biome, Tropical humid forests, Mangrove biome This term must be one of the terms from the bioportal ontology. Terms that are not on this list are not valid.
- feature city, fish farm, livestock-associated habitat, marine habitat, ocean basin, microbial mat This term must be one of the terms from the bioportal ontology. Terms that are not on this list are not valid.
- material air, dust, volcanic soil, saliva, blood, dairy product, surface water, piece of gravelThis term must be one of the terms from the bioportal ontology. Terms that are not on this list are not valid.

4.2.5.0.0.1 The library section captures technical data on the preparation and sequencing done. Chose the library tab to fill out (metagenome for shotgun sequencing or mimarks-survey for amplicon) based on the type of sequencing done. These are separated as different sequencing techniques involved different metadata fields. Each row describes one library for one sample, the samples need to have the identical sample name you used in the sample tab before.

The library_metagenome tab, required fields in red. The **file_name** field holds the filename of the sequence file uploaded, or the filename to use for creating the demultiplexed file if you uploaded a multiplexed sequence file and have barcode sequences in the spreadsheet. This is used for mapping sequence files to metadata.

The **metagenome_name** field holds the name of the metagenome you are submitting. If the file_name field is empty it will be used for mapping metadata to sequence files, in this case it would need to match the uploaded sequence filename (not including file extension).

The **investigation_type** field is required to be metagenome for shotgun metagenome samples and mimarks-survey for amplicon studies (reflecting what tab was filled out).

The type of sequencing instrument used is another required field values are e.g. Illumina, 454, Ion Torrent, Sanger or assembled.

Again, only a limited number of fields are required. However, the more info you provide the easier it is for you and others to understand any potential uses of your data and to understand why results appear in a particular way. It might, for example, allow understanding of specific biases caused by technology choices or sampled environments.

You can fill out one or more environmental metadata packages. Currently we provide support for the following GSC environmental packages:

- Air
- Built Environment
- Host-associated
- Human-associated
- Human-oral
- Human-skin
- Human-vaginal
- Microbial mat/biofilm
- Miscellaneous natural or artificial environment
- Plant-associated
- Sediment
- Soil
- Wastewater sludge
- Water

We strongly encourage users to submit rich metadata but understand the effort required in providing it. Using the environmental packages (which were designed and are used by practitioners in the respective field) should make it reasonably simple to report the essential metadata required to analyze the data. If there is no environmental package to report metadata for your specific sample, please contact MG-RAST staff, we will work with the GSC [10] to create the required questionnaire.

4.3 How to work with projects and collections

Collections provide an efficient way to create multiple sets of metagenomes for analysis. If you wanted e.g. to compare human gut to cow rumen samples, you probably want to see a dialogue like this:

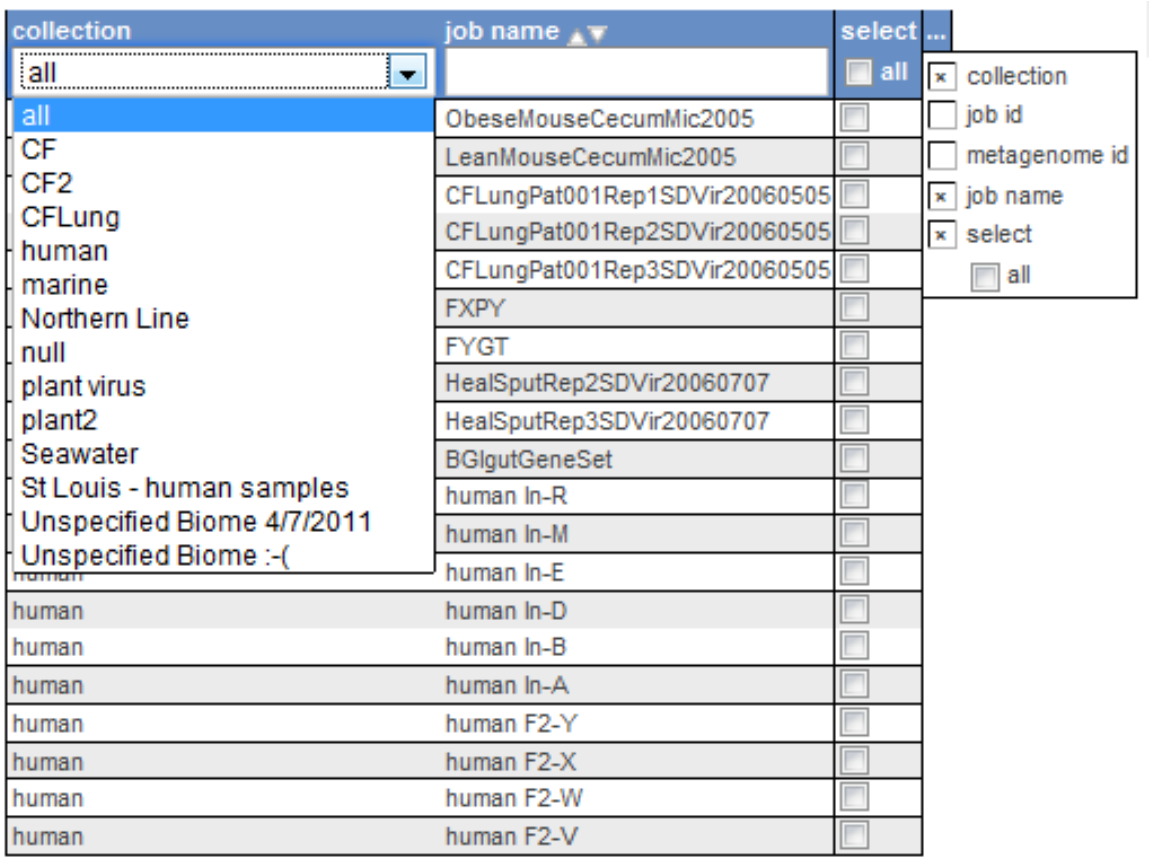


Figure 4.7: A view of the browse table with the collection column enabled. Clicking on the "..." at the right end of the table allows expanding the table columns.

Luckily MG-RAST v3 provides a mechanism to make this happen. Users can create collections that are persistent across multiple sessions

Below we show how to define a collection that allows comparison of multiple data sets. Please note that collections are just short-cuts to the actual samples, they can not be shared at this time.

Step 1: Start with the metadata browser (either on the front page or in the little menu block in the top right hand corner) and click on the globe symbol. See Figure 4.8 for the symbol.

Step 2: This will take you to the Browser dialogue, showing a large number of metagenomes.

Figure 4.8: The symbol for the MG-RAST metagenome browser

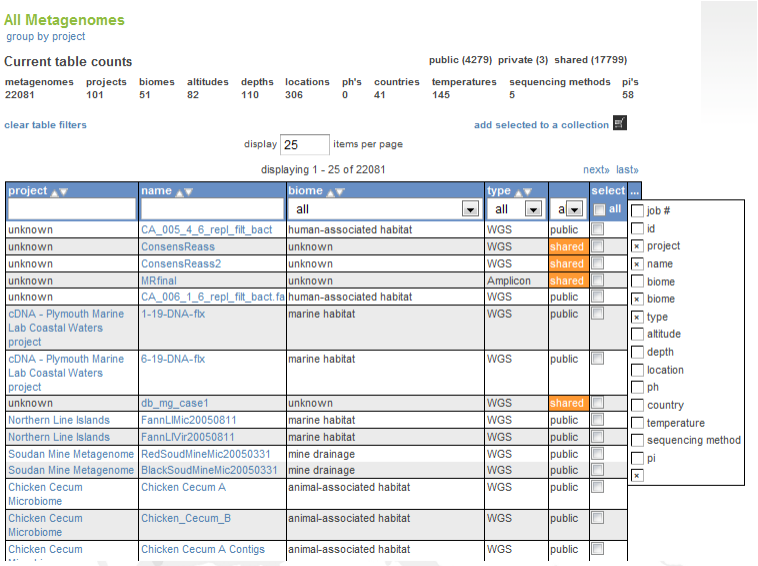


Figure 4.9: The MG-RAST metagenome browser

Step 3: Clicking on biome will allow selecting a specific Biome (here we pick Animal associated)

This results in the following list of metagenomes to be shown:

The list of samples (show above) still shows too many samples when restricted to just animal associated metagenomes.

Step 4: To downselect, search for Twin to further restrict to samples from Peter Turnbaugh and Jeffrey Gordons Human Twin study.

Step 5: Clicking on the black shopping cart symbol in the top right hand corner will allow the creation of a new collection entry. The next step is naming the collection.

Here we name the collection Twin Study and hit OK.

Step 6: Once the collection is added, the new collection will appear in the list of collections (in Your Data Summary).

Step 7: Use collection in the Metagenome selection on the Analysis page. It is possible to do analysis on the metagenomes where they are compared individually, or, alternatively, you may compare whole groups of metagenomes.

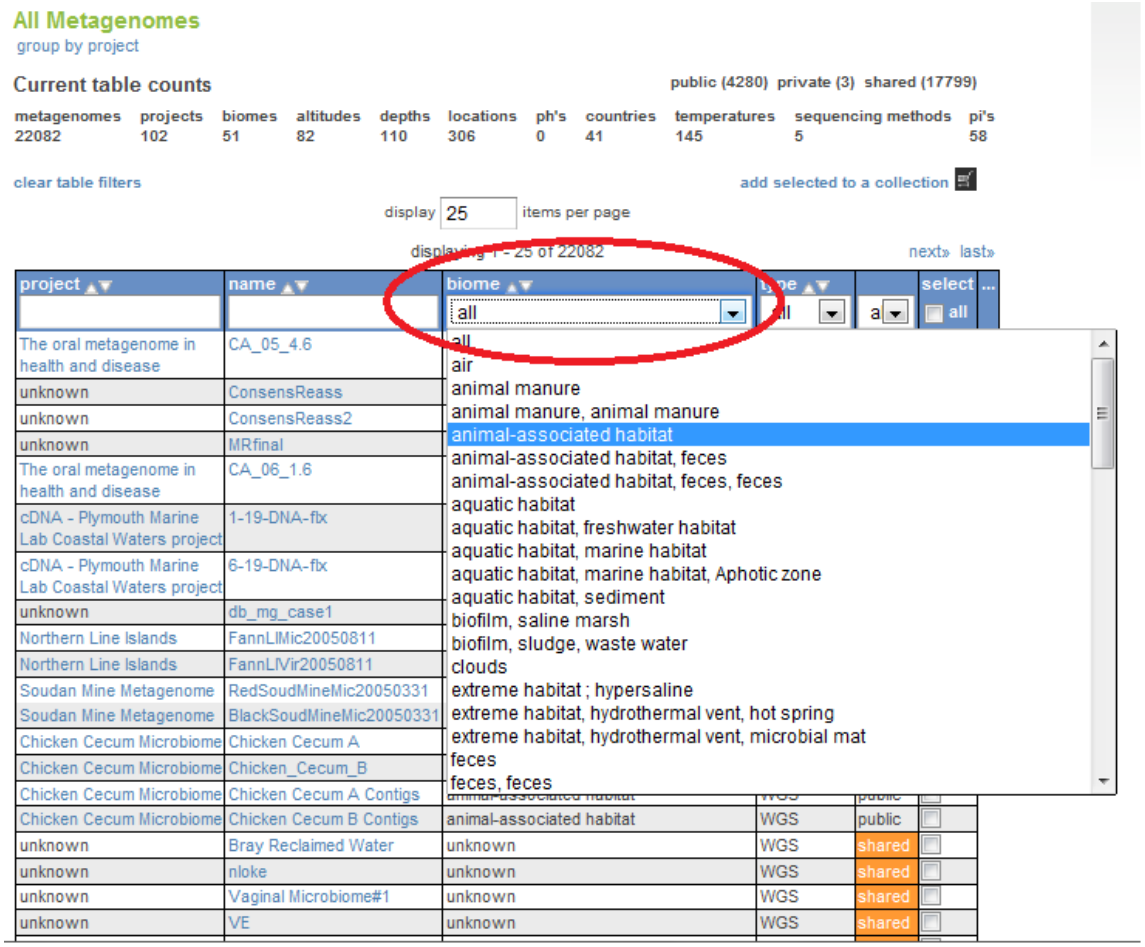


Figure 4.10: The browser allows filtering by e.g. specific BIOME information.

All Metagenomes

group by project

Current table counts public (114) private (0) shared (19)

metagenomes	projects	biomes	altitudes	depths	locations	ph's	countries	temperatures	sequencing methods	pi's
133	18	1	9	7	35	0	10	10	5	15

clear table filters add selected to a collection

display 25 items per page

displaying 1 - 25 of 133 next» last»

project	name	biome	type		select all
		animal-associate	all	a	
Chicken Cecum Microbiome	Chicken Cecum A	animal-associated habitat	WGS	public	
Chicken Cecum Microbiome	Chicken_Cecum_B	animal-associated habitat	WGS	public	
Chicken Cecum Microbiome	Chicken Cecum A Contigs	animal-associated habitat	WGS	public	
Chicken Cecum Microbiome	Chicken Cecum B Contigs	animal-associated habitat	WGS	public	
Human Lung Healthy vs Cystic Fibrosis Metagenome	CFLungPat001Rep1SDVir20060505	animal-associated habitat	WGS	public	
Mosquito Metagenome	Mosq1SDVir20060125	animal-associated habitat	WGS	public	
Human Lung Healthy vs Cystic Fibrosis Metagenome	CFLungPat001Rep2SDVir20060505	animal-associated habitat	WGS	public	
Mosquito Metagenome	MosqDigSDVir20060606	animal-associated habitat	WGS	public	
Mosquito Metagenome	Mosq2SDVir2060606	animal-associated habitat	WGS	public	
Aquacultured Fish (Kent State)	FishHealGutKentSTMic20060504	animal-associated habitat	WGS	public	
Aquacultured Fish (Kent State)	FishMorGutKentSTMic20060504	animal-associated habitat	WGS	public	
Aquacultured Fish (Kent State)	FishHealSlimKentSTMic20060504	animal-associated habitat	WGS	public	
Aquacultured Fish (Kent State)	FishMorSlimKentSTMic20060504	animal-associated habitat	WGS	public	
Aquacultured Fish (Kent State)	FishHealGutKentSTVir20060504	animal-associated habitat	WGS	public	
Aquacultured Fish (Kent State)	FishMorGutKentSTVir20060504	animal-associated habitat	WGS	public	
Aquacultured Fish (Kent State)	FishHealSlimKentSTVir20060504	animal-associated habitat	WGS	public	
Aquacultured Fish (Kent State)	FishMorSlimKentSTVir20060504	animal-associated habitat	WGS	public	
Stressed Coral Holobionts	BocasPAMic20050921	animal-associated habitat	WGS	public	
Stressed Coral Holobionts	DOCPorCompHawMic200602	animal-associated habitat	WGS	public	
Stressed Coral Holobionts	pHPorCompHawVir200602	animal-associated habitat	WGS	public	
Stressed Coral Holobionts	ConPorCompHawVir200602	animal-associated habitat	WGS	public	
Stressed Coral Holobionts	DOCPorCompVirHaw200602	animal-associated habitat	WGS	public	

Figure 4.11: A reduced list of metagenomes for one BIOME.

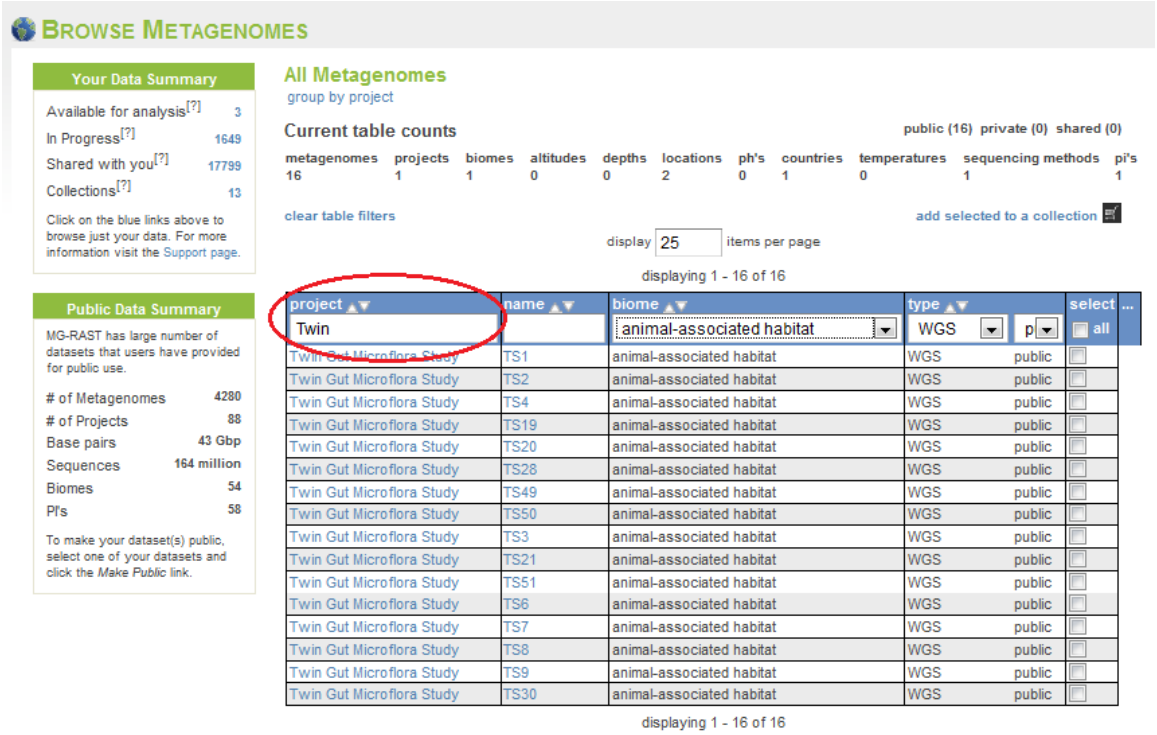


Figure 4.12: Selecting a specific project further reduces the number of data sets being displayed.

4.4 Understanding data sets – Has my sequencing worked?

Unfortunately not every sequencing run works equally well. Users of MG-RAST have provided data with many different sources of error allowing us to provide a number of tools to identify the most common errors.

The quality assessment tools described in 2.5 provide a good tool set for a first data quality analysis. While there are many potential sources of error, some common problems can be easily identified with just the nucleotide histograms, if your data exhibits patterns like the ones described in 2.5 it is likely there were problems with sequencing.

4.4.1 Why are so many reads failing QC?

We frequently find data sets with high numbers of reads filtered out by the quality control. Below we list the major reasons for filtered reads:

1. Artificial duplicate reads (ADRs)

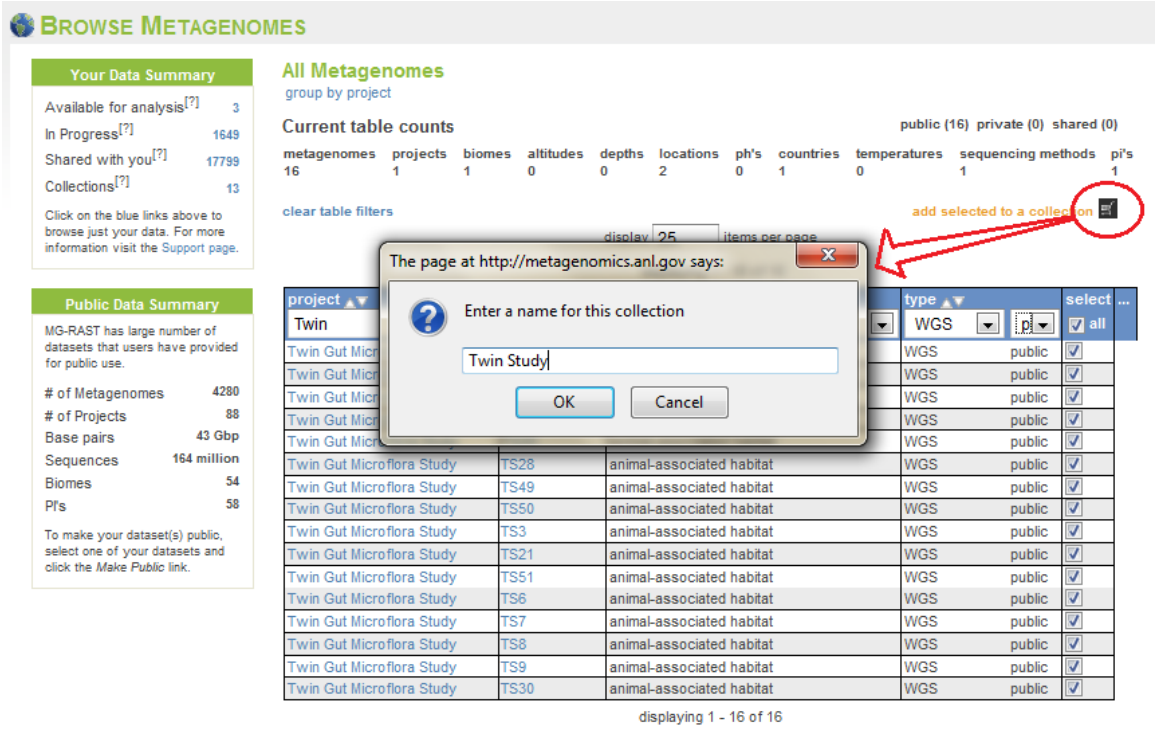


Figure 4.13: Saving a collection

The presence of significant amounts of ADRs hints at problems with a PCR step. Frequently a problem with 454 runs, but also seen with other platforms, a set of DNA templates was copied many times during a PCR step, often consuming up to 80% of the entire data set.

2. filtered reads

   If you have selected screening against a host organism, e.g. the human genome, reads matching that particular genome will be removed from the data set. This frequently consumes a large fraction of the sequence run. There are *in vitro* techniques to minimize the amount of host DNA in your sample, see [37].

3. filter rRNA reads in metatranscriptomes

   If you are trying to analyze a metatranscriptome, reducing the amount of ribosomal is essential. Unless an *in vitro* rRNA knockdown method is applied, up to 97% of all reads will be ribosomal.

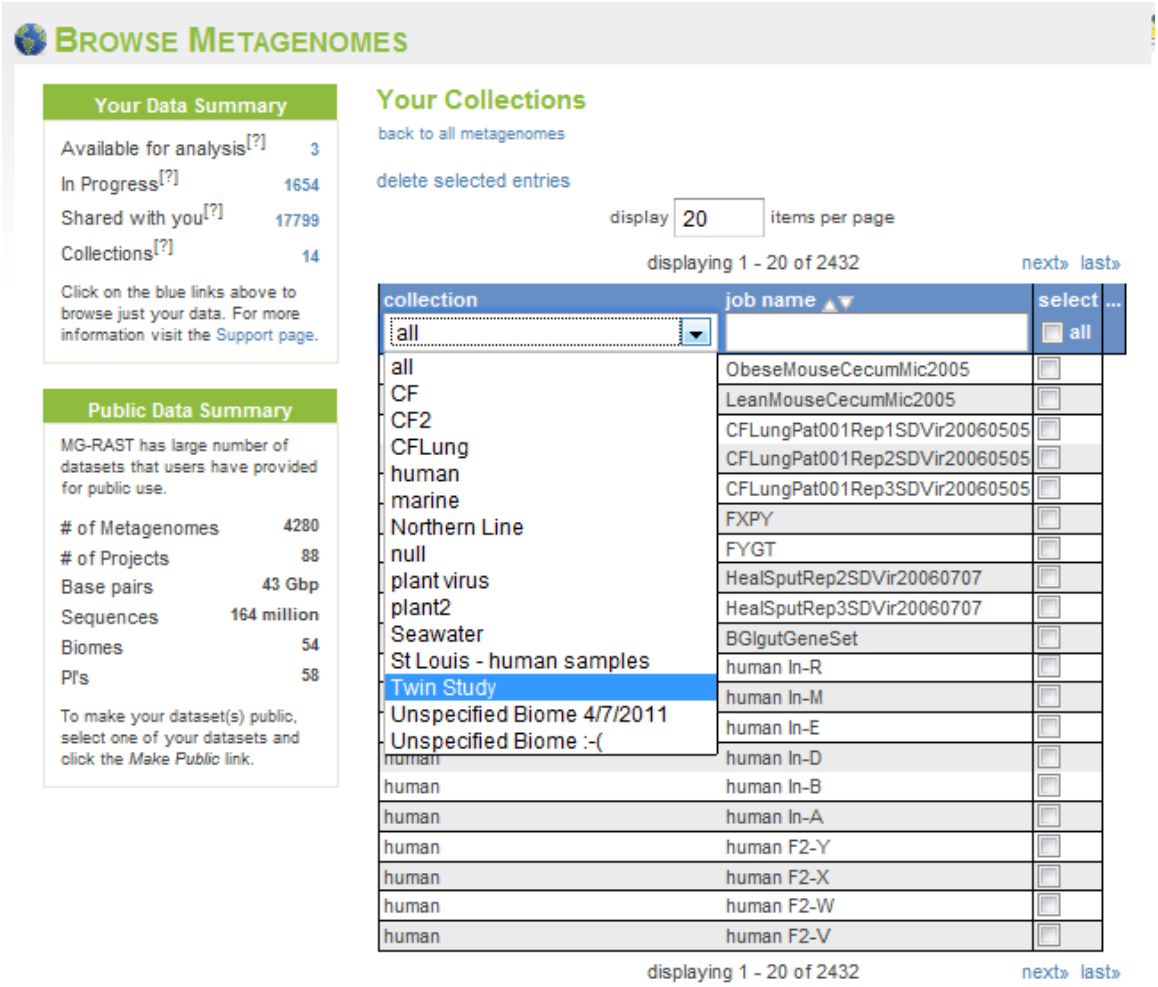


Figure 4.14: A list of collections

4.5 How to drill down using the workbench

One of the new features of MGRAST v3 is the workbench. It is the main mechanism for exchanging subsets of data between analysis views. It also allows you to download the FASTA files of a selection of proteins.

When you initially go to the analysis page (see 3.9) , your workbench will be empty. It is displayed as the leftmost tab in the data tabular view. So how do you get data into the workbench? There are two simple ways to select data subsets from any generated table or from the drilldown of a barchart.

Try this example: Start by selecting the lean and obese mouse cecum samples (MG-RAST

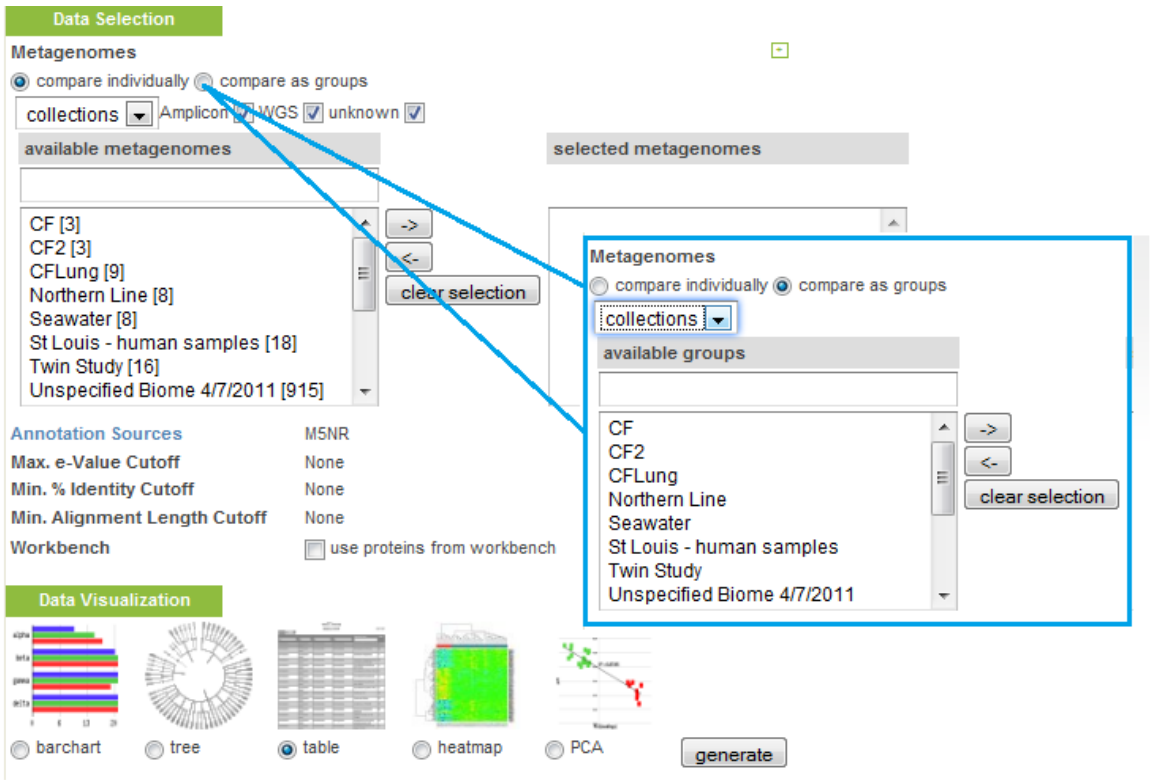


Figure 4.15: Collections can be compared as groups.

IDs 4440463.3 and 4440464.3) [39] in the data selection and creating a table. To do this go to the analysis page and select the analysis view Organism Classification. Expand the metagenome selection by clicking the plus symbol next to metagenomes. Select public from the dropdown-box (to view only public data sets) and type mouse into the filter box. Select the two samples and click the button with the right arrow, then the ok button. The default data visualization is table, so you can click the generate button (Figure 4.16).

After a short wait, a new tab will appear in the tabview below (see Figure 20), showing the data table with organism classifications for the two samples. The last column of this table will have a button labeled to workbench as the column header. Each cell in that column will have a checkbox. Checking a checkbox and clicking the to workbench-button will send the proteins identified by that row to the workbench (Figure 19). Note that you only have one workbench and putting a new set of proteins into it will replace the current content. So what if I want to select all Bacteria, do I really need to click through all those checkboxes? No you can use the grouping feature of the table, so you only have to click one checkbox per metagenome.

Above the table you will find a dropdown-box labeled group table by (Figure 4.17). Select

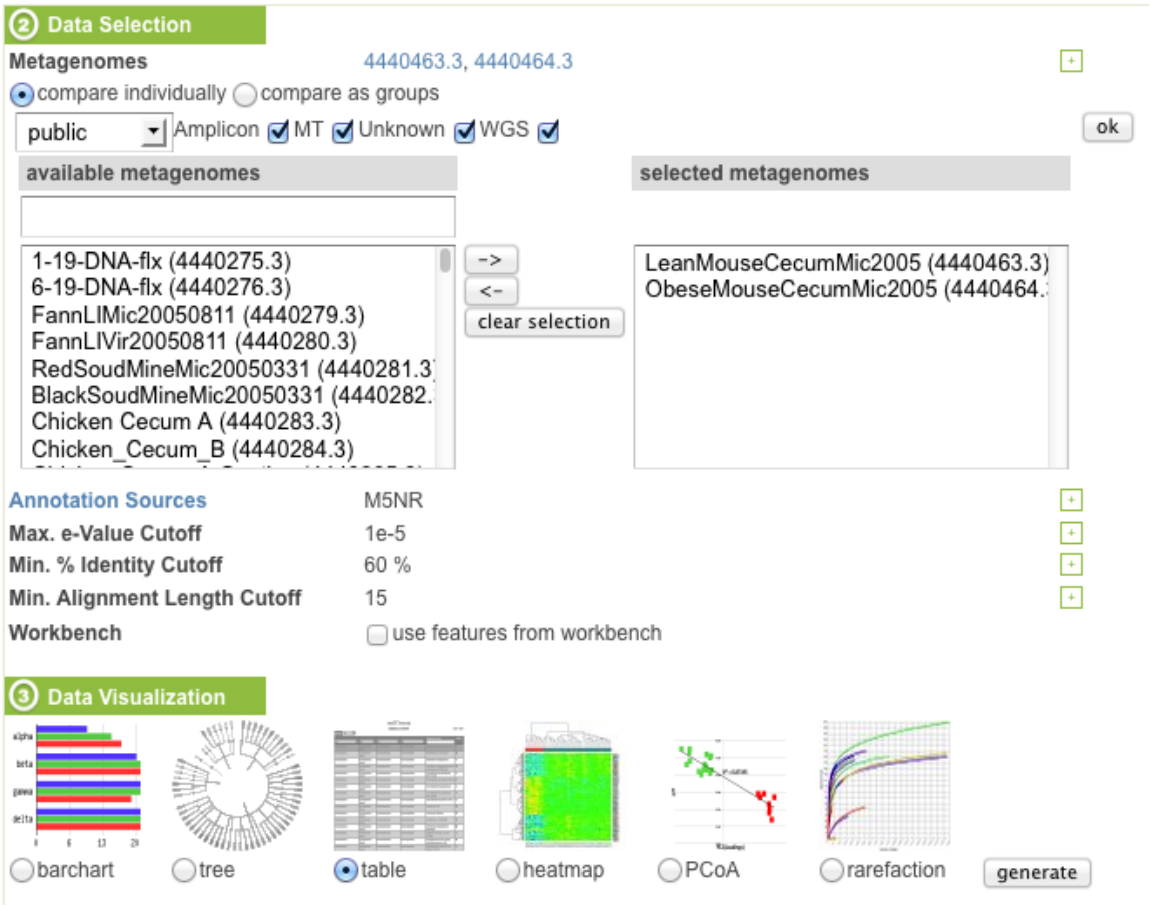


Figure 4.16: Screenshot of the Analysis Page and Workbench tab. Note that users can search and select metagenomes to analyze, the annotation cources and par ameters to set, along with the analysis and visualization they want to perform.

domain and the table will be grouped, so there is only one row per metagenome and domain.

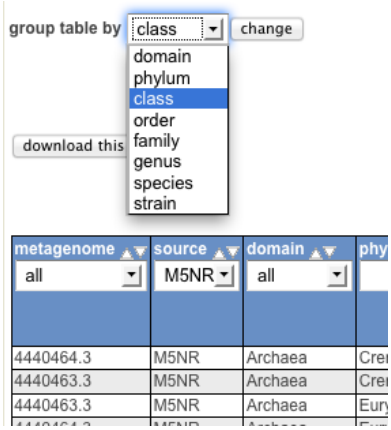


Figure 4.17: Using the tables to group results.

Now check the two boxes in the Bacteria rows and click the to workbench button (see Figure 4.18.

A pop-up message will appear, telling you how many proteins have been sent to the workbench. If you take a look at the tabular view now, you will notice that the workbench tab shows the number of proteins it currently contains (see Figure 4.19). If you click on that tab, you will get information about what the workbench contains. On this tab you will also find a download as FASTA button,

Aside from being able to download the sequences of your selected proteins, you can also use them to generate other visualizations. This includes switching from organism to functional classification. To to this, simply check the use proteins from workbench checkbox in the data selection when generating a new visualization, e.g. a circular tree using the proteins we just buffered.

The table is not the only visualization that allows to put a subselection into the workbench. You can also use the barchart to do this (Figure 4.20). Simply click on the to workbench button next to the headline of a drilldown. Note that you cannot put the topmost barchart into the workbench, as it is not yet a subselection of proteins.

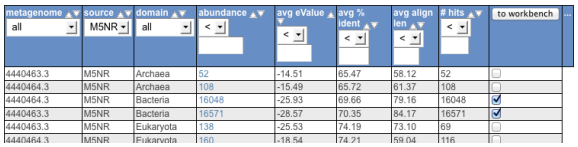


Figure 4.18: Use the table to select results you want to add to your workbench for further analyses.

Workbench (31058 Features) Getting Started Organism table 1

workbench buffers
31058 features

The workbench contains 31058 unique features.

They were selected from the following table lines:

metagenome	source	domain	abundance	avg eValue	avg % ident	avg align len	# hits
4440464.3	M5NR	Bacteria	16048	-25.93	69.66	79.16	16048
4440463.3	M5NR	Bacteria	16571	-28.57	70.35	84.17	16571

download metagenome dna FASTA annotated by GenBank

display annotated hits below

Figure 4.19: View of the workbench with the summary of the proteins that have been added.

Figure 4.20: In addition to the results table, users can download results or add to their workbench from barcharts.

4.6 Downloads from the workbench

The workbench feature stores sub-selections of data and allows those to be used as input for further selection or displays, e.g. select all E. coli reads and then display the functional categories present just in E. coli reads across multiple data sets. In addition the workbench allows downloading the annotated reads for the sub-selection stored in the workbench as fasta (Figure 4.21).

Once processing data sets in MG-RAST is finished a download page is created for the project. On this page all data products created during the computation are made available as files. In addition, datasets which have been published in MG-RAST have links to an ftp site at the top of this page where you can download additional information.

4.7 Viewing Evidence

For individual proteins, the MG-RAST page allows users to retrieve the sequence alignments underlying the annotation transfers (see Figure 4.22). Using the M5NR [41] technology users can retrieve alignments against the database of interest with no additional overhead.

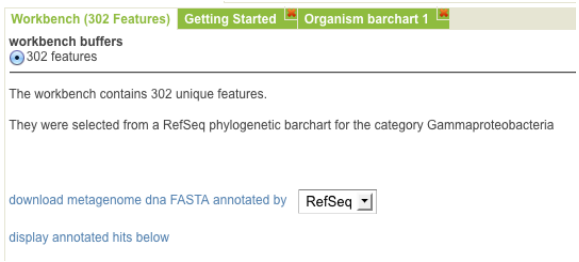


Figure 4.21: The workbench facilitates the download of selected reads using the name space of the selection.

service	Web Interface	API	FTP server	comment
public data access	Y	Y	Y	
private data access	Y	Y	N	
upload	Y	T	N	Unless specifically arranged by help desk

Table 4.2: Differences between the various access modalities

4.8 MG-RAST Outputs

There are three ways to access the data in MG-RAST:

- through the website (which is authenticated)
- through the MG-RAST API (which is also authenticated)
- through the ftp site (which is not authenticated, and for which we only put data in public projects)

Access to private or shared data requires either password access via the web interface or a web-key generated for access via the API.

If you data is private or merely shared the web site and the API are the only two ways to get to it. A way around this is making the data public.

All of the data files available on the website should be available via the urls returned by a query to http://api.metagenomics.anl.gov/analysisset/mgm4450212.3 This returns a JSON data structure with urls for all the data files, for instance 900.abundance.organism.gz can be retrieved from http://api.metagenomics.anl.gov/analysisset/mgm4450212.3-900-5

(This is a public job, the private jobs can be accessed by providing the webkey in the auth field in the GET request.)

BLAT alignments

The sequence alignments underlying functional and organismal classification are stored in MG-RAST in an abbreviated format. This page allows re-creation of these alignments using the original parameters and tools.

NOTE: Since different annotation providers have different interpretation of the sequences, you can switch between **name spaces** when performing this query.

select annotation name space RefSeq

fragment	organism	name space	name space ID	function	e-value[?]	score[?]	identity
4443749.3 CYOBK37TF	Calditerrivibrio nitroreducens DSM 19672	RefSeq	YP_004051066.1	alpha-glucan phosphorylase	3e-40	163 bits (420)	77/128 (60%)
download all sequences download predicted coding sequences			download database sequences				

```
>RefSeq: YP_004051066.1 alpha-glucan phosphorylase [Calditerrivibrio nitroreducens DSM 19672]

          Length = 850

 Score = 163 bits (420), Expect = 3e-40
 Identities = 77/128 (60%), Positives = 88/128 (69%), Gaps = 0/128 (0%)

Query: 7   VAGVDVWLNNPLRPREASGTSGMKAAANGGQNLSILDGWWDEADYYQTGWPIGRGEEYED 66
           V GVDVWLNNP RP EASGTSGMKAA NG  N SILDGWW E      GW IG GEEY D
Sbjct: 585 VRGVDVWLNNPRRPMEASGTSGMKAAINGALNFSILDGWWVEGYKNNNGWSIGAGEEYSD 644

Query: 67  RAYQDEVESNALYDLLEQEVAPLFYQRGSDGLPHQWIQRMKQAIRLNCPQFSTQRMVLEY 126
             YQD VE   LYD LE E+ PLFY +   GLP +W++ MK +I + C +FST RMV+EY
Sbjct: 645 PKYQDFVEGQELYDKLENEIVPLFYAKDRSGLPREWLKMMKNSIFIGCSEFSTSRMVMEY 704

Query: 127 VQRAYIPL 134
            ++ Y PL
Sbjct: 705 HEKYYTPL 712
```

Figure 4.22: BLAT hit details with alignment.

In general we preserve the inputs and outputs for every stage of the pipeline for download and to ensure reproducibility. As an example of why this is useful is the use of dereplicated reads for error estimation by DRISEE (see 2.3.3) or the Lowest Common Ancestor (LCA) (see 2.6.3) algorithm being used to re-interpret the similarities for a given cluster.

4.8.1 Data products on the web site

4.8.1.1 Spreadsheets available on the web pages

Many of the web pages provide spreadsheets for download with the information rendered into tables or graphical displays.

While most of the graphics can be downloaded directly, some require creating a static version for download (which can be achieve through a button next to the graphic.)

Note that the option to use a screenshot will provide images at screen resolution.

4.8.1.2 BIOM file format exports

From the table on the analysis page, users can download BIOM [23] formatted streams, reflecting the paramter choices made. Using this approach users can download abundance profiles in BIOM format. This enables downstream processing with BIOM compliant tools e.g. QIIME [5].

4.8.1.3 Sequence files via the workbench

The workbench allows download of small subsets of sequences with annotations.

4.8.2 The FTP server

All public data is made available on the FTP server.

The FTP server provides a number of services:

- projects

  This is the are where we make public data avaialable for download, sorted into projects.

- data

  Data created to enable MG-RAST e.g. the M5NR is made available here

- tools

  Tools developed by the MG-RAST team will be made available here in addition to github.

- private

  This is a private upload area. MG-RAST help desk staff will provide a private upload location for you in certain situations.

A project directory for `ftp://ftp.metagenomics.anl.gov/projects/128/` is shown in Figure 4.23.

4.8.3 Downloads

One of the critical insights when developing MG-RAST version 3 was the need to make a maximum number of data products available for download for downstream analysis. For this purpose we have created the download page that contains all automatically created data products in a single location for each metagenome. In addition a global download page provides access to all public data sets grouped by projects.

Index of ftp://ftp.metagenomics.anl.gov/projects/128/

Up to higher level directory

Name	Size	Last Modified
4447101.3		5/29/12 12:00:00 AM
4447102.3		5/29/12 12:00:00 AM
4447103.3		5/29/12 12:00:00 AM
4447192.3		5/29/12 12:00:00 AM
4447903.3		5/29/12 12:00:00 AM
4447943.3		5/29/12 12:00:00 AM
4447970.3		5/29/12 12:00:00 AM
4447971.3		5/29/12 12:00:00 AM
metadata.project-128.json	100 KB	6/11/12 12:00:00 AM
metadata.project-128.xls	15 KB	5/29/12 12:00:00 AM
metadata.project-128.xlsx	15 KB	6/11/12 12:00:00 AM

Figure 4.23: Listing of the project directory for `ftp://ftp.metagenomics.anl.gov/projects/128/`

We list the data products available on the download page for each metagenome using a specific example (MG-RAST ID: 4465825.3) in the Appendix (see Appendix A.

The general paradigm is to make all files available that are generated during the automated analysis, in addition the user submitted data and metadata are made available.

Chapter 5

Putting it all in perspective

5.1 Discussion

We have described MG-RAST, a community resource for the analysis of metagenomic sequence data. We have developed a new pipeline and environment for automated analysis of shotgun metagenomic data, as well as a series of interactive tools for comparative analysis. The pipeline is also being used for the analysis of metatranscriptome data as well as amplicon data of various kinds. This service is being used by thousands of users worldwide, many contributing their data and analysis results to the community. We believe that community resources, such as MG-RAST, will fill a vital role in the bioinformatics ecosystem in the years to come.

MG-RAST has become a community clearinghouse for metagenomic data and analysis, with over 12,000 public data sets that can be freely used. Because analysis was performed in a uniform way, these data sets can be used as building blocks for new comparative analysis; so long as new data sets are analyzed similarly, results are robustly comparable between new and old data set analysis. These data sets (and the resulting analysis data products) are made available for download and reuse as well.

Community resources like MG-RAST provide an interesting value proposition to the metagenomics community: First, it enables low-cost meta-analysis. Users utilize the data products in MG-RAST as a basis for comparison without the need to re-analyze every data set used in their studies. The high computational cost of analysis [43] makes pre-computation a prerequisite for large scale meta-analyses. In 2001, Angiuli et al. [1], determined the real currency cost of re-analysis for the over 12,000 data sets openly available on MG-RAST to be in excess of 30 million US-dollars if Amazons EC2 platform is used. This figure doesnt consider the 66,000 private data sets that have been analyzed with MG-RAST.

Second, it provides incentives to the community to adopt standards, both in terms of metadata

and analysis approaches. Without this standardization, data products arent readily reusable, and computational costs quickly become unsustainable. We are not arguing that a single analysis is necessarily suitable for all users, rather, we are pointing out that if one particular type of analysis is run for all data sets, the results can be efficiently reused, amortizing costs. Open access to data and analyses foster community interactions that make it easier for researchers efforts to achieve consensus with respect to establishing best practises as well as identifying methods and analyses that could provide misleading results.

Third, community resources drive increased efficiency and computational performance. Community resources consolidate the demand for analysis resources sufficiently to drive innovation in algorithms and approaches. Due to this demand, the MG-RAST team has needed to scale the efficiency of their pipeline by a factor of nearly 1000 over the last four years. This drive has caused improvements in gene calling, clustering, sequence quality analysis, as well as many other areas. In less specialized groups with less extreme computational needs, this sort of efficiency gain would be difficult to achieve. Moreover, the large quantities of data sets that flow through the system have forced the hardening of the pipeline against a large variety of sequence pathology types that wouldnt be readily observed in smaller systems.

We believe that our experiences in the design and operation of MG-RAST are representative of bioinformatics as a whole. The community resource model is critical if we are to benefit from the exponential growth in sequence data. This data has the potential to enable new insights into the world around us, but only if we can analyze it effectively. It is only due to this approach that we have been able to scale to the demands of our users effectively, analyzing over 200 billion sequences thus far.

We note that scaling to the required throughput by adding hardware to the system or simply renting time using an unoptimized pipeline on e.g. Amazons EC2 machine would not be economically feasible. The real currency cost on EC2 for the data currently analyzed in MG-RAST (26 Terabasepairs) would be in excess of 100 million US dollars using an unoptimized workflow like CLOVR [1].

All of MG-RAST is open source and available on `https://github.com/MG-RAST`.

5.2 Future Work

While MG-RAST v3 is a substantial improvement over prior systems, much work remains to be done. Data set sizes continue to increase at an exponential pace. Keeping up with this change remains a top priority, as metagenomics users continue to benefit from increased resolution of microbial communities. Upcoming versions of MG-RAST will include: (1) mechanisms for speeding

pipeline up using data reduction strategies that are biologically motivated; (2) opening up the data ecosystem via an API that will enable third-party development and enhancements; (3) providing distributed compute capabilities using user-provided resources; as well as (4) providing virtual integration of local data sets to allow comparison between local data and shared data without requiring full integration.

5.2.1 Roadmap

We maintain a rough roadmap for future version of MG-RAST.

version 3.4

- web services API
- R client
- provide alpha version of MG-RAST remote compute client (using VMs)

3.5

- provide reviewer access tokens
- consolidate all SQL onto PostGRES
- provide beta version of MG-RAST remote compute client (using VMs)
- include IPython based notebooks for analysis
- use AWE for all computations and SHOCK for all pipeline storage
- multi-metagenome recruitment plot

4.0

- re-write web interface to support many browsers
- BAM upload support
- BAM download support
- variation study support
- convert all file access to SHOCK

5.0

- provide federated SHOCK system
- provide an assembly based pipeline

5.3 Acknowledgments

This work used the Magellan machine (Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, Contract grant DE-AC02-06CH11357) at Argonne National Laboratory, and the PADS resource (National Science Foundation grant OCI-0821678) at the Argonne National Laboratory/University of Chicago Computation Institute.This work was supported in part by the U.S. Dept. of Energy under Contract DE-AC02-06CH11357, the Sloan Foundation (SLOAN #2010-12), NIH NIAID (HHSN272200900040C) and the NIH Roadmap HMP program (1UH2DK083993-01).

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

Bibliography

[1] S. V. Angiuoli, M. Matalka, A. Gussman, K. Galens, M. Vangala, D. R. Riley, C. Arze, J. R. White, O. White, and W. F. Fricke. Clovr: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC Bioinformatics*, 12:356, 2011.

[2] Ramy Aziz, Daniela Bartels, Aaron Best, Matthew DeJongh, Terrence Disz, Robert Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth Glass, Michael Kubal, Folker Meyer, Gary Olsen, Robert Olson, Andrei Osterman, Ross Overbeek, Leslie McNeil, Daniel Paarmann, Tobias Paczian, Bruce Parrello, Gordon Pusch, Claudia Reich, Rick Stevens, Olga Vassieva, Veronika Vonstein, Andreas Wilke, and Olga Zagnitko. The rast server: Rapid annotations using subsystems technology. *BMC Genomics*, 9(1):75, 2008.

[3] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Res*, 41(Database issue):D36–42, 2013.

[4] A. Bolotin, B. Quinquis, A. Sorokin, and S. D. Ehrlich. Clustered regularly interspaced short palindrome repeats (crisprs) have spacers of extrachromosomal origin. *Microbiology*, 151(Pt 8):2551–61, 2005.

[5] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. Qiime allows analysis of high-throughput community sequencing data. *Nat Methods*, 7(5):335–6, 2010.

[6] J. R. Cole, B. Chai, T. L. Marsh, R. J. Farris, Q. Wang, S. A. Kulam, S. Chandra, D. M. McGarrell, T. M. Schmidt, G. M. Garrity, J. M. Tiedje, and Ribosomal Database Project. The ribosomal database project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic acids research*, 31(1):442–443, January 2003.

[7] M. P. Cox, D. A. Peterson, and P. J. Biggs. Solexaqa: At-a-glance quality assessment

of illumina second-generation sequencing data. *BMC Bioinformatics*, 11:485, 2010.

[8] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. Greengenes, a Chimera-Checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, 72(7):5069–5072, July 2006.

[9] R. C. Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–1, 2010.

[10] D. Field, L. Amaral-Zettler, G. Cochrane, J.R. Cole, P. Dawyndt, G.M. Garrity, J. Gilbert, F.O. Glöckner, L. Hirschman, and I. Karsch-Mizrachi. The genomic standards consortium. *Plos Biology*, 9(6):e1001088, 2011.

[11] Edgar Gabriel, Graham E. Fagg, George Bosilca, Thara Angskun, Jack J. Dongarra, Jeffrey M. Squyres, Vishal Sahay, Prabhanjan Kambadur, Brian Barrett, Andrew Lumsdaine, Ralph H. Castain, David J. Daniel, Richard L. Graham, and Timothy S. Woodall. Open MPI: Goals, concept, and design of a next generation MPI implementation. In *Proceedings, 11th European PVM/MPI Users' Group Meeting*, pages 97–104, Budapest, Hungary, September 2004.

[12] V. Gomez-Alvarez, T. K. Teal, and T. M. Schmidt. Systematic artifacts in metagenomes from complex microbial communities. *ISME J*, 3(11):1314–7, 2009.

[13] S. M. Huse, J. A. Huber, H. G. Morrison, M. L. Sogin, and D. M. Welch. Accuracy and quality of massively parallel dna pyrosequencing. *Genome Biol*, 8(7):R143, 2007.

[14] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster. Megan analysis of metagenomic data. *Genome Res*, 17(3):377–86, 2007.

[15] L. J. Jensen, P. Julien, M. Kuhn, C. von Mering, J. Muller, T. Doerks, and P. Bork. eggnog: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res*, 36(Database issue):D250–4, 2008.

[16] M. Kanehisa. The kegg database. *Novartis Found Symp*, 247:91–101; discussion 101–3, 119–28, 244–52, 2002.

[17] K. P. Keegan, W. L. Trimble, J. Wilkening, A. Wilke, T. Harrison, M. D'Souza, and F. Meyer. A platform-independent method for detecting errors in metagenomic sequencing data: Drisee. *PLoS Comput Biol*, 8(6):e1002541, 2012.

[18] W. J. Kent. Blat–the blast-like alignment tool. *Genome Res*, 12(4):656–64, 2002.

[19] Murphy S Kagan L Kravitz S Lombardot T Field D Glckner FO; Genomic Standards Consortium Kottmann R, Gray T. A standard migs/mims compliant xml schema: toward the development of the

genomic contextual data markup language (gcdml). *OMICS*, 12(2):115–21, 2008.

[20] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[21] Michele Magrane and Uniprot Consortium. UniProt knowledgebase: a hub of integrated protein data. *Database : the journal of biological databases and curation*, 2011, January 2011.

[22] V. M. Markowitz, N. N. Ivanova, E. Szeto, K. Palaniappan, K. Chu, D. Dalevi, I. M. Chen, Y. Grechkin, I. Dubchak, I. Anderson, A. Lykidis, K. Mavromatis, P. Hugenholtz, and N. C. Kyrpides. Img/m: a data management and analysis system for metagenomes. *Nucleic Acids Res*, 36(Database issue):D534–8, 2008.

[23] D. McDonald, J.C. Clemente, J. Kuczynski, J. Rideout, J. Stombaugh, D. Wendel, A. Wilke, S. Huse, J. Hufnagle, and F. Meyer. The biological observation matrix (biom) format or: how i learned to stop worrying and love the ome-ome. *Gigascience*, 2012.

[24] F Meyer, D Paarmann, M D'Souza, R Olson, EM Glass, M Kubal, T Paczian, A Rodriguez, R Stevens, A Wilke, J Wilkening, and RA Edwards. The metagenomics rast server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(1):386, 2008.

[25] NHGRI. Cost per raw megabase of dna sequence, 2012.

[26] Timothy J. Dallman Chrystala Constantinidou Saheer E Gharbia John Wain Mark J. Pallen Nicholas J. Loman, Raju V Misra. Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, (5):434439, 2012.

[27] B. D. Ondov, N. H. Bergman, and A. M. Phillippy. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics*, 12:385, 2011.

[28] R. Overbeek, T. Begley, R.M. Butler, J.V. Choudhuri, N. Diaz, H.-Y. Chuang, M. Cohoon, V. de Crécy-Lagard, T. Disz, R Edwards, M Fonstein, E.D. Frank, S. Gerdes, E.M. Glass, A. Goesmann, L. Krause, B. Linke, A.C. McHardy, F. Meyer, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, M. Kubal, N. Larsen, H. Neuweger, C. Rückert, G. J. Olsen, R. Olson, A. Osterman, V. Portnoy, G.D. Pusch, D.A. Rodionov, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*, 33(17), 2005.

[29] Elmar Pruesse, Christian Quast, Katrin Knittel, Bernhard M. Fuchs, Wolfgang

Ludwig, Jörg Peplies, and Frank Oliver O. Glöckner. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research*, 35(21):7188–7196, December 2007.

[30] K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 35(Database issue), January 2007.

[31] J. Reeder and R. Knight. The 'rare biosphere': a reality check. *Nat Methods*, 6(9):636–7, 2009.

[32] Mina Rho, Haixu Tang, and Yuzhen Ye. Fraggenescan: Predicting genes in short and error-prone reads,. *NAR*, (in print), 2009.

[33] C. S. Riesenfeld, P. D. Schloss, and J. Handelsman. Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet*, 38:525–52, 2004.

[34] E. E. Snyder, N. Kampanya, J. Lu, E. K. Nordberg, H. R. Karur, M. Shukla, J. Soneja, Y. Tian, T. Xue, H. Yoo, F. Zhang, C. Dharmanolla, N. V. Dongre, J. J. Gillespie, J. Hamelius, M. Hance, K. I. Huntington, D. Jukneliene, J. Koziski, L. Mackasmiel, S. P. Mane, V. Nguyen, A. Purkayastha, J. Shallom, G. Yu, Y. Guo, J. Gabbard, D. Hix, A. F. Azad, S. C. Baker, S. M. Boyle, Y. Khudyakov, X. J. Meng, C. Rupprecht, J. Vinje, O. R. Crasta, M. J. Czar, A. Dickerman, J. D. Eckart, R. Kenyon, R. Will, J. C. Setubal, and B. W. Sobral. PATRIC: the VBI PathoSystems resource integration center. *Nucleic Acids Res*, 35(Database issue), January 2007.

[35] Terry Speed. *Statistical Analysis of Gene Expression Microarray Data*. Chapman and Hall/CRC, 2003.

[36] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, 2003.

[37] Torsten Thomas, Jack Gilbert, and Folker Meyer. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, 2(1):3, 2012.

[38] W. L. Trimble, K. P. Keegan, M. D'Souza, A. Wilke, J. Wilkening, J. Gilbert, and F. Meyer. Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. *BMC Bioinformatics*, 13(1):183, 2012.

[39] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon. An obesity-associated gut microbiome with increased capacity for en-

ergy harvest. *Nature*, 444(7122):1027–31, 2006.

[40] W3C. File api; w3c working draft 25 october 2012, 2012.

[41] A. Wilke, T. Harrison, J. Wilkening, D. Field, E. M. Glass, N. Kyrpides, K. Mavrommatis, and F. Meyer. The m5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics*, 13:141, 2012.

[42] A. Wilke, J. Wilkening, E.M. Glass, N. Desai, and F. Meyer. An experience report: porting the mg-rast rapid metagenomics analysis pipeline to the cloud. *Concurrency and Computation: Practice and Experience*, 23(17):22502257, 2011.

[43] J. Wilkening, A. Wilke, Narayan Desai, and Folker Meyer. Using clouds for metagenomics: A case study. In *IEEE Cluster 2009*, 2009.

[44] Pelin Yilmaz, Renzo Kottmann, Dawn Field, Rob Knight, James R. Cole, Linda Amaral-Zettler, Jack A. Gilbert, Ilene Karsch-Mizrachi, Anjanette Johnston, Guy Cochrane, Robert Vaughan, Christopher Hunter, Joonhong Park, Norman Morrison, Phillipe Rocca-Serra, Peter Sterk, Mani Arumugam, Laura Baumgartner, Bruce W. Birren, Martin J. Blaser, Vivien Bonazzi, Tim Booth, Peer Bork, Frederic D. Bushman, Pier Luigi Buttigieg, Patrick Chain , Elizabeth K. Costello, Heather Huot-Creasy, Peter Dawyndt, Todd DeSantis , Noah Fierer, Jed Fuhrman, Rachel E. Gallery , Dirk Gevers , Richard A. Gibbs , Michelle Gwinn Giglio , Inigo San Gil , Antonio Gonzalez3 , Jeffrey I. Gordon, Robert Guralnick , Wolfgang Hankeln , Sarah Highlander , Philip Hugenholtz, Janet Jansson , Scott T. Kelley , Jerry Kennedy , Dan Knights , Omry Koren , Justin Kuczynski , Nikos Kyrpides , Robert Larsen , Christian L. Lauber , Teresa Legg , Ruth E. Ley , Catherine A. Lozupone , Wolfgang Ludwig , Donna Lyons , Eamonn Maguire , Barbara A. Methé , Folker Meyer , Sara Nakielny , Karen E. Nelson , Diana Nemergut , Lindsay K. Neubold , Josh D. Neufeld , Anna E. Oliver , Norman R. Pace , Giriprakash Palanisamy , Jörg Peplies , Jane Peterson , Joseph Petrosino , Lita Proctor , Elmar Pruesse , Christian Quast , Jeroen Raes , Sujeevan Ratnasingham , Jacques Ravel , David A. Relman , Susanna Assunta-Sansone , Patrick D. Schloss , Lynn Schriml , Erica Sodergren , Aymé Spor , Jesse Stombaugh , James M. Tiedje , Doyle V. Ward , George M. Weinstock , Doug Wendel , Owen White , Andrew Whitely , Andreas Wilke , Jennifer Wortmann , and Frank Oliver Glöckner . The "minimum information about an environmental sequence" (miens) specification. *Nature*

Biotechnology, 2010.

Appendix A

The downloadable files for each data set

Uploaded File(s) DNA (4465825.3.25422.fna)

Uploaded nucleotide sequence data in FASTA format. Preprocessing

Depending on the options chosen, the preprocessing step filters sequences based on length, number of ambiguous bases and quality values if available.

passed, DNA (4465825.3.100.preprocess.passed.fna)

A FASTA formatted file containing the sequences which were accepted and will be passed on to the next stage of the analysis pipeline.

removed, DNA (4465825.3.100.preprocess.removed.fna)

A FASTA formatted file containing the sequences which were rejected and will not be passed on to the next stage of the analysis pipeline. Dereplication

The optional dereplication step removes redundant technical replicate sequences from the metagenomic sample. Technical replicates are identified by binning reads with identical first 50 base-pairs. One copy of each 50-base-pair identical bin is retained.

passed, DNA (4465825.3.150.dereplication.passed.fna)

A FASTA formatted file containing one sequence from each bin which will be passed on to the next stage of the analysis pipeline.

removed, DNA (4465825.3.150.dereplication.removed.fna)

A FASTA formatted file containing the sequences which were identified as technical replicates and will not be passed on to the next stage of the analysis pipeline. Screening

The optional screening step screens reads against model organisms using bowtie to remove reads which are similar to the genome of the selected species.

passed, DNA (4465825.3.299.screen.passed.fna)

A FASTA formatted file containing the reads which which had no similarity to the selected genome and will be passed on to the next stage of the analysis pipeline. Prediction of protein

coding sequences

Coding regions within the sequences are predicted using FragGeneScan, an ab-initio prokaryotic gene calling algorithm. Using a hidden Markov model for coding regions and non-coding regions, this step identifies the most likely reading frame and translates nucleotide sequences into amino acids sequences. The predicted coding regions, possibly more than one per fragment, are called features.

coding, Protein (4465825.3.350.genecalling.coding.faa)

A amino-acid sequence FASTA formatted file containing the translations of the predicted coding regions.

coding, DNA (4465825.3.350.genecalling.coding.fna)

A nucleotide sequence FASTA formatted file containing the predicted coding regions. RNA Clustering

Sequences from step 2 (before dereplication) are pre-screened for at least 60% identity to ribosomal sequences and then clustered at 97% identity using UCLUST. These clusters are checked for similarity against the ribosomal RNA databases (Greengenes [8], LSU and SSU from [29], and RDP [6]).

rna97, DNA (4465825.3.440.cluster.rna97.fna)

A FASTA formatted file containing sequences that have at least 60% identity to ribosomal sequences and are checked for RNA similarity.

rna97, Cluster (4465825.3.440.cluster.rna97.mapping)

A tab-delimited file that identifies the sequence clusters and the sequences that comprise them.

The columns making up each line in this file are:

Cluster ID, e.g. rna97_998

Representative read ID, e.g. 11909294

List of IDs for other reads in the cluster, e.g. 11898451,11944918

List of percentage identities to the representative read sequence, e.g. 97.5%,100.0%

RNA similarities

The two files labelled expand are comma- and semicolon- delimited files that provide the mappings from md5s to function and md5s to taxonomy:

annotated, Sims (4465825.3.450.rna.expand.lca)

annotated, Sims (4465825.3.450.rna.expand.rna)

Packaged results of the blat search against all the DNA databases with MD5 value of the database sequence hit followed by sequence or cluster ID, similarity information, annotation, organism, database name.

raw, Sims (4465825.3.450.rna.sims)

This is the similarity output from BLAT. This includes the identifier for the query which is either the FASTA id or the cluster ID, and the internal identifier for the sequence that it hits.

The fields are in BLAST m8 format:

Query id (either fasta ID or cluster ID), e.g. 11847922

Hit id, e.g. lcl—501336051b4d5d412fb84afe8b7fdd87

percentage identity, e.g. 100.00

alignment length, e.g. 107

number of mismatches, e.g. 0

number of gap openings, e.g. 0

q.start, e.g. 1

q.end, e.g. 107

s.start, e.g. 1262

s.end, e.g. 1156

e-value, e.g. 1.7e-54

score in bits, e.g. 210.0

filtered, Sims (15:04 4465825.3.450.rna.sims.filter)

This is a filtered version of the raw Sims file above that removes all but the best hit for each data source. Gene Clustering

Protein coding sequences are clustered at 80% identity with UCLUST. This process does not remove any sequences but instead makes the similarity search step easier. Following the search, the original reads are loaded into MG-RAST for retrieval on-demand.

aa90, Protein (4465825.3.550.cluster.aa90.faa)

An amino acid sequence FASTA formatted file containing the translations of one sequence from each cluster (by cluster ids starting with aa90_) and all the unclustered (singleton) sequences with the original sequence ID.

aa90, Cluster (4465825.3.550.cluster.aa90.mapping)

A tab-separated file in which each line describes a single cluster.

The fields are:

Cluster ID, e.g. aa90_3270

protein coding sequence ID including hit location and strand, e.g. 11954908_1_121_+

additional sequence ids including hit location and strand, e.g. 11898451_1_119_+,11944918_19_121_+

sequence % identities, e.g. 94.9%,97.0%

Protein similarities

annotated, Sims (4465825.3.650.superblat.expand.lca)

The expand.lca file decodes the MD5 to the taxonomic classification it is annotated with.

The format is:

md5(s), e.g. cf036dfa9cdde3a8a4c09d7fabfd9ba5;1e538305b8319dab322b8f28da82e0a1

feature id (for singletons) or cluster id of hit including hit location and strand, e.g. 11857921_1_101_-

alignment %, e.g. 70.97;70.97

alignment length, e.g. 31;31

E-value, e.g. 7.5e-05;7.5e-05

Taxonomic string, e.g. Bacteria;Actinobacteria;Actinobacteria (class);Coriobacteriales;Coriobacteriaceae;Slackia;Slackia exigua;-

annotated, Sims (4465825.3.650.superblat.expand.protein)

Packaged results of the blat search against all the protein databases with MD5 value of the database sequence hit followed by sequence or cluster ID, similarity information, functional annotation, organism, database name.

Format is:

md5 (identifier for the database hit), e.g. 88848aa7224ca2f3ac117e7953edd2d9

feature id (for singletons) or cluster ID for the query, e.g. aa90_22837

alignment % identity, e.g. 76.47

alignment length, e.g. 34

E-value, e.g. 1.3e-06

protein functional label, e.g. SsrA-binding protein

Species name associated with best protein hit, e.g. Prevotella bergensis DSM 17361 RefSeq 585502

raw, Sims (4465825.3.650.superblat.sims)

Blat output with sequence or cluster ID, md5 value for the sequence in the database and similarity information.

filtered, Sims (4465825.3.650.superblat.sims.filter)

Blat output filtered to take only the best hit from each data source.

Appendix B

Terms of Service

- MG-RAST is a web-based computational metagenome analysis service provided on a best-effort basis. We strive to provide correct analysis, privacy, but can not guarantee correctness of results, integrity of data or privacy. That being said, we are not responsible for any HIPPA regulations regarding human samples uploaded by users. We will try to provide as much speed as possible and will try to inform users about wait times. We will inform users about changes to the system and the underlying data.
- We reserve the right to delete non public data sets after 120 days.
- We reserve the right to reject data set that are not complying with the purpose of MG-RAST.
- We reserve the right to perform additional data analysis (e.g. search for novel sequence errors to improve our sequence quality detection, clustering to improve sequence similarity searches etc.) AND in certain cases utilize the results. We will NOT release user provided data without consent and or publish on user data before the user.
- User acknowledges the restrictions stated about and will cite MG-RAST when reporting on their work.
- User acknowledges the fact that data sharing on MG-RAST is meant as a pre-publication mechanism and we strongly encourage users to make data publicly accessible in MG-RAST once published in a journal (or after 120 days).
- User acknowledges that data (including metadata) provided is a) correct and b) user either owns the data or has the permission of the owner to upload data and or publish data on MG-RAST.

- We reserve the right to curate and update public meta data.
- We reserve the right at any time to modify this agreement. Such modifications and additional terms and conditions will be effective immediately and incorporated into this agreement. MG-RAST will make a reasonable effort to contact users via email of any changes and your continued use of MG-RAST will be deemed acceptance thereof.

Appendix C

Tools and data used by MG-RAST

The MG-RAST team is happy to acknowledge the use of the following great software and data products: Databases

MG-RAST uses a number of protein and ribosomal RNA databases integrated into the M5NR [41] (Wilke et al, BMC Bioinformatics 2012. Vol 13, No. 151) non-redundant database using the M5NR tools.

C.1 Databases

C.1.1 Protein databases

- The SEED [28] (Overbeek et al., NAR, 2005, Vol. 33, Issue 17)
- GenBank [3] (Benson et al., NAR, 2011, Vol. 39, Database issue)
- RefSeq [30] (Pruitt et al., NAR, 2009, Vol. 37, Database issue)
- IMG/M (Markowitz et al., NAR, 2008, Vol. 36, Database issue)
- UniProt [21] (Apweiler et al., NAR, 2011, Vol. 39, Database issue)
- eggNOGG [15] (Muller et al., NAR, 2010, Vol. 38, Database issue)
- KEGG [16] (Kanehisa et al., NAR, 2008, Vol. 36, Database issue)
- PATRIC [34] (Gillespie et al., Infect. Immun., 2011, Vol. 79, no. 11)

C.1.2 Ribosomal RNA databases:

- greengenes [8] (DeSantis et al., Appl Environ Microbiol., 2006, Vol. 72, no. 7)
- SILVA [29] (Pruesse et al., NAR, 2007, Vol. 35, issue 21)
- RDP [6] (Cole et al., NAR, 2009, Vol. 37, Database issue)

C.2 Software

C.2.1 Bioinformatics codes:

- FragGeneScan [32] (Rho et al, NAR, 2010, Vol. 38, issue 20)
- BLAT [18] (J. Kent, Genome Res, 2002, Vol. 12, No. 4)
- QIIME [5] (Caporaso et al, Nature Methods, 2010, Vol. 7, No. 5) (we also use uclust that is part of QIIME)
- Biopython
- Bowtie [20] (Langmead et al., Genome Biol. 2009, Vol 10, issue 3)
- sff_extract, Jose Blanca and Joaquin Caizares
- Dynamic Trim, part of SolexaQA, [7] (Cox et al., BMC Bioinformatics, 2011, Vol. 11, 485)
- FastqJoin

C.2.2 Web/UI tools:

- Krona [27] (Ondov et. al. BMC Bioinformatics, 2011, Vol. 12, 385)
- Raphal JavaScript Library (Dmitry Baranovskiy)
- jQuery
- Circos (Krzywinski et al., Genome Res. 2009, Vol. 19)
- cURL

C.2.3 Behind the scenes:

- Perl
- Python
- R
- Googles V8 JavaScript engine
- Node.js
- nginx
- OpenStack

List of Figures

Glossary

16s 16S ribosomal RNA (or 16S rRNA) is a component of the 30S small subunit of prokaryotic ribosomes. 4, 5, 74, 114

ADR Artificial duplicate read. 4, 16, 20, 114

DNA Deoxyribonucleic acid. 4, 114

EC2 Amazon Elastic Compute Cloud. 4, 5, 114

MD5 The MD5 message-digest algorithm is a widely used cryptographic hash function that produces a 128-bit (16-byte) hash value. Specified in RFC 1321, MD5 has been utilized in a wide variety of security applications, and is also commonly used to check data integrity.. 4, 32, 66–68, 71, 101–103, 112, 114

RNA Ribonucleic acid. 4, 21, 114

rRNA ribosomal ribonucleic acid. 4, 12, 15, 19, 114

SEED The SEED effort led by Ross Overbeek is a systematic annotation effort for prokaryotic genomes using Subsystems.. 4, 7, 15, 17, 18, 29, 38, 39, 42, 60, 106, 110, 111, 114

Subsystem A subsystem is a set of functional roles that an annotator has decided should be thought of as related. Frequently, subsystems represent the collection of functional roles that make up a metabolic pathway, a complex (e.g., the ribosome), or a class of proteins (e.g., two-component signal-transduction proteins within Staphylococcus aureus). Construction of a large set of curated populated subsystems is at the center of the SEED annotation efforts.. 4, 17, 18, 29, 42, 52, 110, 114

Mathematics and Computer Science Division
Argonne National Laboratory
9700 South Cass Avenue, Bldg. 240
Argonne, IL 60439-4847

www.anl.gov

Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC