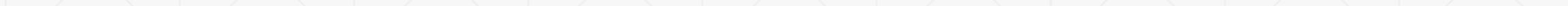


MG-RAST Tutorial

Moon Kim and Nick Falkowski

Agenda

- Getting Started
 - Introduction
- Uploading Data
 - Data formats
 - Metadata creation
- Data Products
 - Pipeline
 - Analytic tools
- API



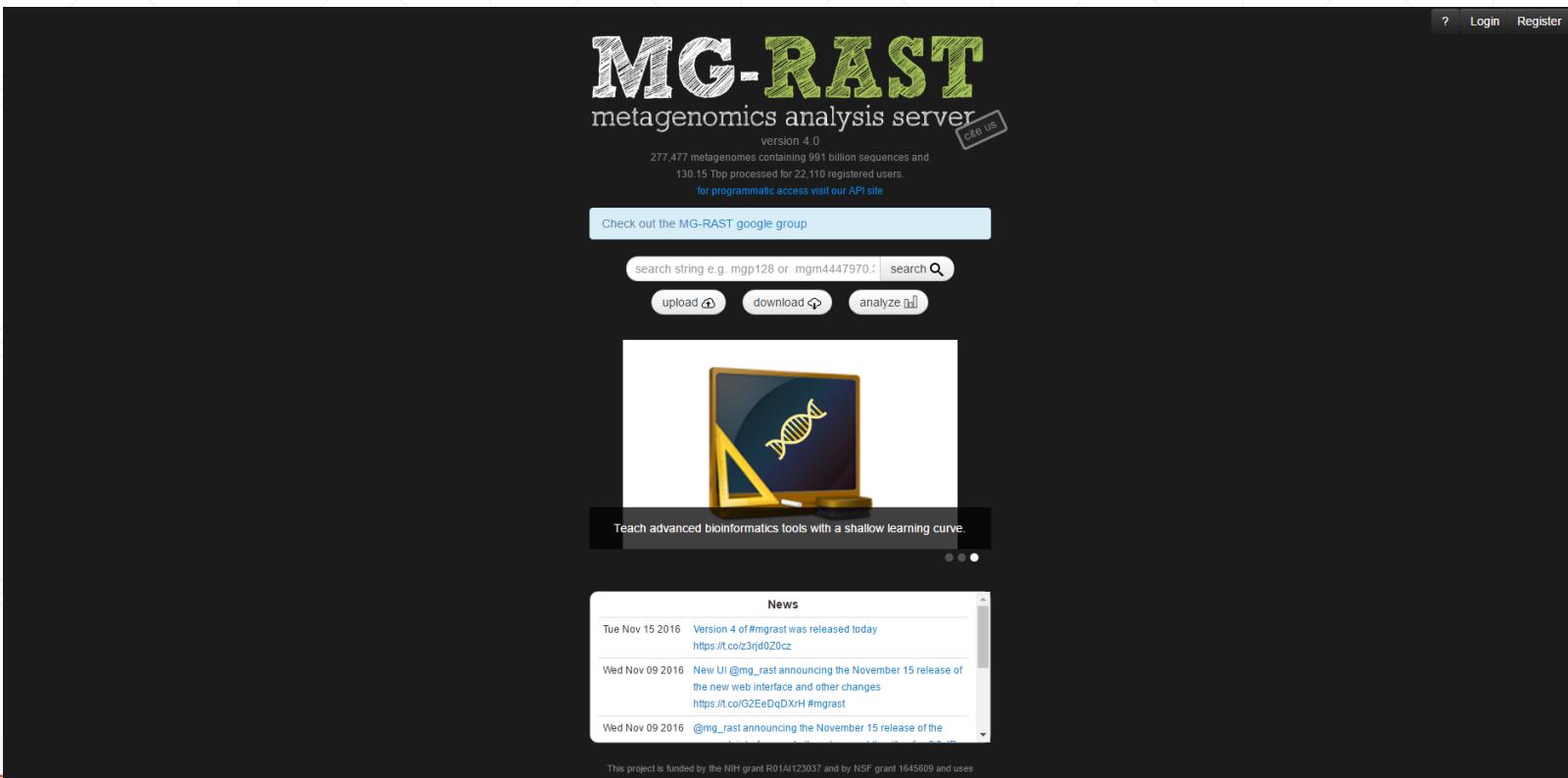
Introduction

- Metagenomic Rapid Annotations using Subsystems Technology
- Open source web application server that suggests automatic phylogenetic and functional analysis of metagenomes
- One of the biggest repositories for metagenomic data
- Also supports amplicon sequences and metatranscritome sequence processing, but cannot perform eukaryotic metagenomes analysis



Signing Up

- <http://metagenomics.anl.gov/>



Public Data

MG-RAST metagenomics analysis server

search

enter search term search metadata function organism Your search returned 27,881 results. Showing the first 20 matches. [download search results](#)

Created ▲▼	Study ▲▼	Metagenome ▲▼	Seq Type ▲▼	Biome ▲▼	Country ▲▼	Location ▲▼
2017-01-04	healthy_psy	HC_women_Nolndex_L001_R1_001.f	metatranscriptome	terrestrial biome	USA	Chapel Hill, NC
2016-11-08	SWIFT_metagenome_kuuj	swift_kuuj	shotgun metagenome	Small lake biome	Canada	whapmagoostui-kuujjuarapik
2016-11-01	Drinking Water 16S seq	drinkingwater201416s	amplicon metagenome	aquatic biome	USA	Pittsburgh
2016-11-01	Drinking Water Fungal	drinking_water_its2014	amplicon metagenome	aquatic biome	USA	Pittsburgh
2016-11-01	SaltMarsh	NSP1	amplicon metagenome	marine salt marsh biome	United Kingdom	Welwick
2016-10-31	SAS_BGR_metagenome	BGR_merged	shotgun metagenome	Small lake biome	Canada	whapmagoostui-kuujjuarapik
2016-10-31	SAS_BGR_metagenome	SAS_merged	shotgun metagenome	Small lake biome	Canada	whapmagoostui-kuujjuarapik
2016-10-28	Miseq 20140618 0d R2	20140618d_S1_L001_R2_001	shotgun metagenome	aquatic biome	Japan	Fukuoka
2016-10-28	Illumina for SMURF	sample_3_allRegions_uniqueReads	metatranscriptome	terrestrial biome	Israel	Rehovot
2016-10-28	Illumina for SMURF	sample_20_allRegions_uniqueReads	metatranscriptome	terrestrial biome	Israel	Rehovot
2016-10-28	Illumina for SMURF	sample_4_allRegions_uniqueReads	metatranscriptome	terrestrial biome	Israel	Rehovot
2016-10-24	Tara Oceans	ERR599038	shotgun metagenome	ocean biome	Pacific Ocean	South Pacific Ocean
2016-10-24	Tara Oceans	ERR599012	shotgun metagenome	ocean biome	Indian Ocean	Indian Ocean

Refine Search

Add a search term for a specific metadata field to refine your search. You can use the asterisk (*) symbol as a wildcard.

field PI firstname

term enter searchterm add

Searches [?] Collections [?]

you have no searches

create new [?]

Store the parameters of your search query.

name enter name

description

enter description (optional)

store

Public Data

Tara Oceans (mgp20413)

principle investigator Coordinators Tara Oceans Consortium, Tara Oceans Consortium

visibility public

static link <http://metagenomics.anl.gov/linkin.cgi?project=mgp20413>

description

-

funding source

-

contact

Administrative

Coordinators Tara Oceans Consortium (kandels@embl.de)

Tara Oceans Consortium (-)

Heidelberg, Germany, Germany

Technical

-- (-)

- (-)

-,-

metagenomes

	<input type="checkbox"/>	<input type="button" value="create collection"/>																						
name	Q	bp count	Q	seq. count	Q	material	Q	sample	Q	library	Q	location	Q	country	Q	coordinates	Q	type	Q	method	Q	download	Q	
ERR599135		6,938,884,806		52,709,744		water		mgs556807		mgI556809		South Atlantic Ocean		Atlantic Ocean		-20.4091, -3.1759		WGS		illumina		<input type="button" value="metadata"/>	<input type="button" value="submitted"/>	<input type="button" value="results"/>
ERR599010		5,129,921,923		36,449,856		water		mgs556759		mgI556761		South Atlantic Ocean		Atlantic Ocean		-20.9354, -35.1803		WGS		illumina		<input type="button" value="metadata"/>	<input type="button" value="submitted"/>	<input type="button" value="results"/>
ERR599129		6,127,630,616		43,744,550		water		mgs556801		mgI556803		South Atlantic Ocean		Atlantic Ocean		-31.0266, 4.665		WGS		illumina		<input type="button" value="metadata"/>	<input type="button" value="submitted"/>	<input type="button" value="results"/>
ERR598968		28,412,269,137		178,227,284		water		mgs556729		mgI556731		North Atlantic Ocean		Atlantic Ocean		34.6712, -71.3093		WGS		illumina		<input type="button" value="metadata"/>	<input type="button" value="submitted"/>	<input type="button" value="results"/>
ERR599052		46,461,017,602		286,911,517		water		mgs556795		mgI556797		North Pacific Ocean		Pacific Ocean		35.3671, -127.7422		WGS		illumina		<input type="button" value="metadata"/>	<input type="button" value="submitted"/>	<input type="button" value="results"/>
ERR315861		6,785,017,005		42,781,232		water		mgs556705		mgI556707		Mediterranean Sea		Mediterranean Sea		42.2038, 17.715		WGS		illumina		<input type="button" value="metadata"/>	<input type="button" value="submitted"/>	<input type="button" value="results"/>



Getting sequence data

- Couple of options we tried
 - Downloading ‘fasta’ files directly from NCIB website
 - Using R to download ‘sra’ files and converting to ‘fastq’



Download from NCIB

Guerrero Negro Hypersaline Microbial Mat Metagenome encompasses the following 10 sub-projects:

Project Type			Number of Projects
Genome sequencing Highest level of assembly : Scaffolds or contigs			10
BioProject accession	Assembly level	Organism	Title
PRJNA29605	Scaffolds or contigs	microbial mat metagenome	Metagenome from Guerrero Negro hypersaline microbial mat 01 (DOE Joint Genome Institute)
PRJNA29611	Scaffolds or contigs	microbial mat metagenome	Metagenome from Guerrero Negro hypersaline microbial mat 02 (DOE Joint Genome Institute)
PRJNA29613	Scaffolds or contigs	microbial mat metagenome	Metagenome from Guerrero Negro hypersaline microbial mat 03 (DOE Joint Genome Institute)
PRJNA29615	Scaffolds or contigs	microbial mat metagenome	Metagenome from Guerrero Negro hypersaline microbial mat 04 (DOE Joint Genome Institute)
PRJNA29617	Scaffolds or contigs	microbial mat metagenome	Metagenome from Guerrero Negro hypersaline microbial mat 05 (DOE Joint Genome Institute)
PRJNA29619	Scaffolds or contigs	microbial mat metagenome	Metagenome from Guerrero Negro hypersaline microbial mat 06 (DOE Joint Genome Institute)
PRJNA29621	Scaffolds or contigs	microbial mat metagenome	Metagenome from Guerrero Negro hypersaline microbial mat 07 (DOE Joint Genome Institute)
PRJNA29623	Scaffolds or contigs	microbial mat metagenome	Metagenome from Guerrero Negro hypersaline microbial mat 08 (DOE Joint Genome Institute)
PRJNA29625	Scaffolds or contigs	microbial mat metagenome	Metagenome from Guerrero Negro hypersaline microbial mat 09 (DOE Joint Genome Institute)
PRJNA29627	Scaffolds or contigs	microbial mat metagenome	Metagenome from Guerrero Negro hypersaline microbial mat 10 (DOE Joint Genome Institute)

Less...

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (total)	10531
WGS master	1
Genomic DNA	10530
PUBLICATIONS	
PubMed	2
PMC	2
OTHER DATASETS	
BioSample	1
Assembly	1

Summary ▾ 20 per page ▾ Sort by Default order ▾

Items: 1 to 20 of 10530

- [Uncultured organism clone SBXZ_6525 16S ribosomal RNA gene, partial](#)
1. 1,364 bp linear DNA
Accession: JN437545.1 GI: 364589549
[GenBank](#) [FASTA](#) [Graphics](#)
- [Uncultured organism clone SBXZ_6524 16S ribosomal RNA gene, partial](#)
2. 1,261 bp linear DNA
Accession: JN437544.1 GI: 364589548
[GenBank](#) [FASTA](#) [Graphics](#)
- [Uncultured organism clone SBXZ_6523 16S ribosomal RNA gene, partial](#)
3. 1,132 bp linear DNA
Accession: JN437543.1 GI: 364589547
[GenBank](#) [FASTA](#) [Graphics](#)

Send: ▾ Filters: [Manage Filters](#)

Complete Record Coding Sequences Gene Features

Choose Destination

File Clipboard Collections

Download 10530 items.

Format [FASTA](#) ▾

Sort by [Default order](#) ▾

Show CI

[Create File](#)

Using R and SRA database

Basic method

1. Download SRAdb
2. Search database for SR* and filter or convert to what you need
3. Download SRA files
4. Convert to 'fastq' with SRA-Toolkit
5. Gzip 'fastq' files so you can upload to MG-RAST

Sample R code

```
# configuration
sqlfile <- 'SRAmetadb.sqlite'
sra_download = 'F:\\\\sra'
fastq_output = 'F:\\\\fastq_split'
fastq_bin = 'F:\\\\sra\\\\fastq-dump.exe'

library(stringr)
library(SRAdb)
library(Biostrings)

# if you don't have the sql file (24GB download)
if (! file.exists(sqlfile)) sqlfile <- getSRAdbFile()
sra_con <- dbConnect(SQLite(),sqlfile)

# Microbial communities within an urban mass transit system == SRP073814
sra_info <- getSRA(search_terms='SRP073814', out_types=c('sra'), sra_con)

# filter only WGS runs
sra_subset <- subset(sra_info, library_strategy=='WGS')

# download the sra files (55.2GB download)
for (r in sra_subset$run) getSRAfile(r, sra_con, fileType = 'sra',destDir=sra_download)

# extract fastq from sra files (~200GB)
for (f in sra_subset$run){
  print(sprintf('run %d/%d',which(sra_subset$run == f),length(sra_subset$run)))
  system2(fastq_bin, args = c('-O',fastq_output,'--split-files',f))
}

# at this point you probably want to gzip all the fastq files (~71GB)
```

Creating metadata for MG-RAST

Finding metadata

- Collect from NCIB project descriptions/other sources
- Use NCIB Run Selector (<https://www.ncbi.nlm.nih.gov/Traces/study/?go=home>)
- Data from R queries

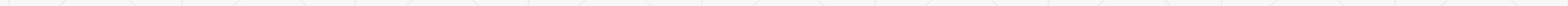
Formatting data for MG-RAST

- Download Excel template and fill in, upload with sequence data
- Use “MetaZen” tool on MG-RAST



Uploading MetaData

- MetaData not necessary but strongly encouraged
- Datasets with metadata given priority in the pipeline
- MetaData in the form of GSC standard compliant checklists (<http://gensc.org>)



Uploading Data

upload ➞ submit ➞ progress *

Data submission is a two-step process. As the **first step**, data is uploaded into your private inbox on the MG-RAST server. This area is write only and accessible only to you. From the inbox data can then be submitted. Use the [upload](#) function, or use [our API](#) to upload your data. To view your webkey required for using the API, click [here](#).

Submission of multiple files, sharing of data, or data publication requires metadata. You can use [this Excel template](#) and/or the [MetaZen tool](#) to fill out the metadata spreadsheet for a study.

As the **second step**, data needs to be submitted for processing. At submission time you either add data to an existing study (or project) or create a new study. Upon successful submission, data is removed from the inbox. You will be notified via email once your submission has completed processing. In addition, you can monitor the progress of your submission at the [job status](#).

The screenshot shows a file list in a central window. The files listed are:

Name	Type
MGRAST_MetaData_template_123.xlsx	metadata
MGRAST_MetaData_template_1.7.xlsx	metadata
mat01.fasta	sequence
MGRAST_MetaData_template_1.7.json	metadata

To the right, a sidebar contains:

- frequent questions**
 - File Formats
 - Metadata
 - Studies and Projects
- running actions**

Uploaded files will remain in your inbox for 72 hours before they are automatically deleted.

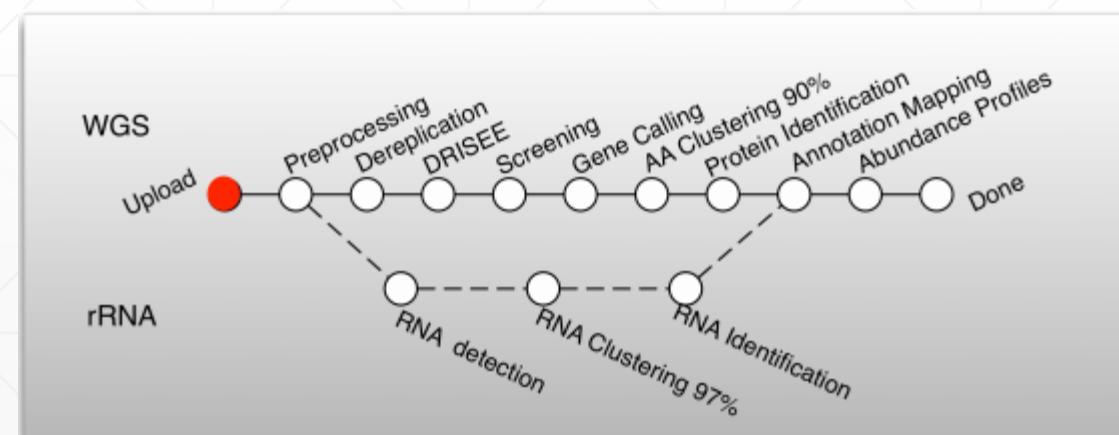
If you perform actions on files in your inbox that take some time to complete, you can view their status here.

- no actions running on files in your inbox -

next ➞

MG-Rast Pipeline

- Performs quality control, protein prediction, clustering and similarity based annotation on nucleic acid sequence datasets using a number of bioinformatics tools
- Data on MG-RAST is private to the submitting user unless shared with other users or made public by the user. Public data given highest priorities.



MG-Rast Pipeline

- Data hygiene
 - Quality control and removal of artifacts.
- Feature extraction
 - Identification of protein coding and rRNA features (aka “genes”)
- Feature annotation
 - Identification of putative functions and taxonomic origins for each of the features
- Profile generation
 - Creation of multiple on disk representations of the information obtained above.
- Data loading
 - Loading the representations into the appropriate databases.



MG-Rast Pipeline

1. select metadata file

2. select project

3. select sequence file(s)

4. choose pipeline options

assembled Select this option if your input sequence file(s) contain assembled data and include the coverage information within each sequence header as described [here](#).

dereplication Remove artificial replicate sequences produced by sequencing artifacts [Gomez-Alvarez, et al, The ISME Journal \(2009\)](#)

screening H. sapiens, NCBI v36 ▾

Remove any host specific species sequences (e.g. plant, human or mouse) using DNA level matching with bowtie [Langmead et al., Genome Biol. 2009, Vol 10, issue 3](#)

dynamic trimming Remove low quality sequences using a modified DynamicTrim [Cox et al., \(BMC Bioinformatics, 2011, Vol. 11, 485\)](#).

15 Specify the lowest phred score that will be counted as a high-quality base.

5 Sequences will be trimmed to contain at most this many bases below the above-specified quality.

5. submit

MG-Rast Pipeline

• upload ➞ • submit ➞ • progress *

The submission process is complete and you can now safely close your browser. You can come back to this page or reload it at any time to monitor the progress of your submission.

The table below shows the status of your submissions as they progress through our pipeline. Click on the job number in the first column to view details of that specific job. The status lights in the last column show an overview of which tasks have been completed and which still need to run.

job	stage	status	tasks
300611	dereplication	* in-progress	● ● ● ○ ○ ● ● ● ● ● ● ● ● ● ● ●
300610	dereplication	* in-progress	● ● ○ ○ ● ● ● ● ● ● ● ● ● ● ● ●

showing rows 1-2 of 2

mat01 (300611)

Status Pipeline Settings

Below are all tasks that are part of the MG-RAST pipeline for your submission.

- Green bars, indicating completed tasks, can be expanded via mouseclick
- Blue bars indicate tasks currently being computed on
- Orange bars represent the next tasks to be queued
- Gray tasks are waiting for completion of another task they depend on
- Red bars indicate an error

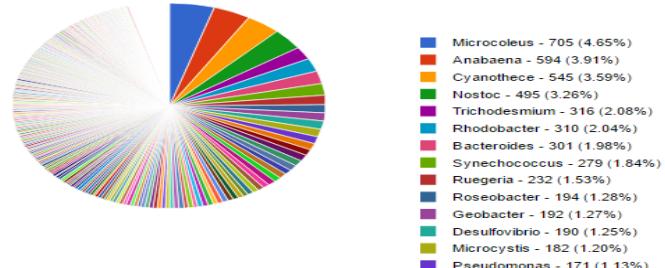
qc_stats	(in queue)
✓ preprocess	1/28/2017, 4:40:40 PM
✓ dereplication	1/28/2017, 4:41:20 PM
○ screen	(in queue)
○ rna detection	(in queue)
○ rna clustering	(not started)
○ rna sims blat	(not started)
○ genecalling	(not started)
○ aa filtering	(not started)
○ aa clustering	(not started)
○ aa sims blat	(not started)
○ aa sims annotation	(not started)
○ rna sims annotation	(not started)
○ index sim seq	(not started)
○ md5 abundance	(not started)
○ ica abundance	(not started)
○ source abundance	(not started)

MG-Rast Data Products

- Sequence Breakdown
 - Predicted Features
 - Analysis Statistics
 - DRISEE
 - K-mer Profile
 - Nucleotide Histogram
 - Source Hits Distribution
 - Functional Category Hits Distribution
 - Taxonomic Hits Distribution
 - Rank Abudnace Plot
 - Rarefaction Curve
 - Alpha Diversity
 - Sequence Length Histogram
 - Sequence GC Distribution
 - Metadata
-

MG-Rast Data Products

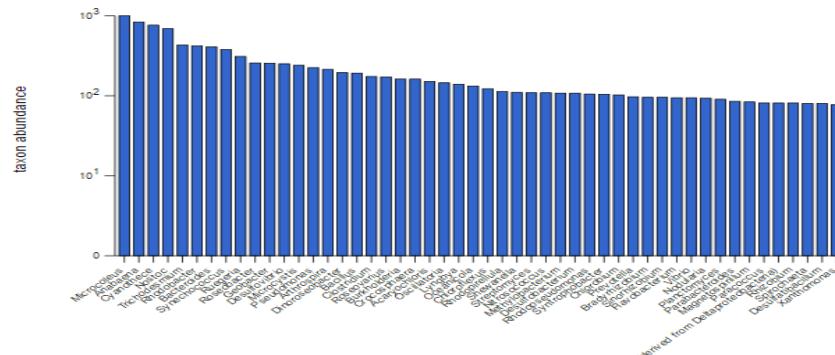
Genus



Rank Abundance Plot

The plots below show the taxonomic abundances ordered from the most abundant to least abundant. Only the top 50 most abundant are shown. The y-axis plots the abundances of annotations on a log scale. The rank abundance chart is a tool for visually representing taxonomic richness and evenness.

Domain Phylum Class Order Family Genus



Custom Analysis

Create a new Analysis

To perform an analysis, you must first load the metagenomic profiles to analyze. A profile holds the abundance values and cutoffs for a list of database sources for a specific dataset. You can select the databases and datasets, as well as a name for your analysis below. Click the ✓-button to load the data from our server.

Profiles are generated on demand. Depending on profile size the initial calculation may take some time. Once computed they will be cached and subsequent requests will download immediately. You can use the -icon in the top menu bar to store profiles on your harddrive and upload them back into your browser cache (without requiring interaction with our server).

Once all required data is loaded you can start the analysis.

selected databases

RefSeq Subsystems

metagenomes

add collection ▾

Enter filter name

sequence type shotgun amplicon metatranscriptome status public private

0.2-um-passable microorganisms in deep-sea hydrothermal fl
00000N1_S1_L001_R2_001
00000N3_S2_L001_R1_001
00000N3_S2_L001_R2_001
00000N5_S3_L001_R2_001
00000N6_S4_L001_R2_001
00000N7_S5_L001_R1_001
00000N8_S6_L001_R1_001
0000N11_S7_L001_R1_001
0000N13_S8_L001_R2_001

available databases

IMG add

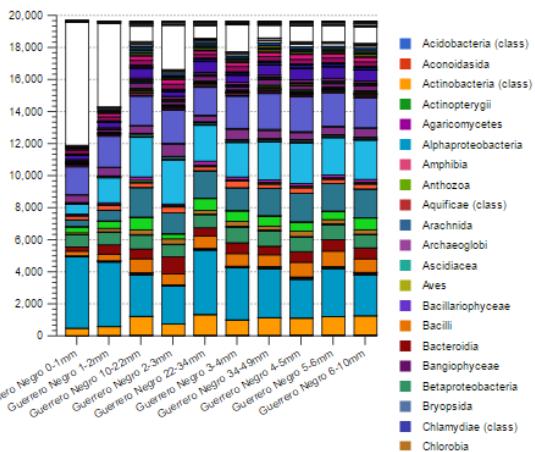
Custom Analysis

stacked bar-chart of analysis 1 [?]

adjust graph data

metadatum	metagenome name	metadatum to name the datasets by
perform normalization	yes	normalize the datasets

layout



Analysis

analysis 1

e-value 5 %-ident 60 length 15 min.abundance 1

source RefSeq type taxonomy level class

- no filter -

name	hits
Guerrero Negro 0-1mm	19,689
Guerrero Negro 1-2mm	16,350
Guerrero Negro 10-22mm	13,208
Guerrero Negro 2-3mm	14,734
Guerrero Negro 22-34mm	13,458
Guerrero Negro 3-4mm	14,708
Guerrero Negro 34-49mm	11,421
Guerrero Negro 4-5mm	16,128
Guerrero Negro 5-6mm	13,275
Guerrero Negro 6-10mm	15,873

View Metadata Plugins Export

table matrix pie-chart donut-chart rarefaction barchart stacked bar PCoA

heatmap differential

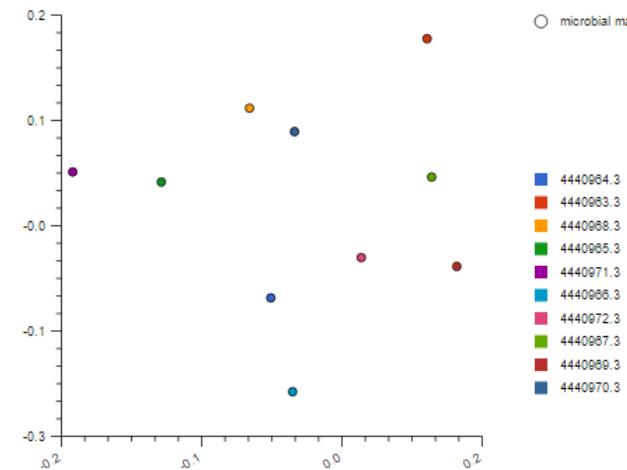
Custom Analysis

PCoA of analysis 1 [?]

adjust graph data

metadatum	metagenome name	metadatum to name the datasets by
color attribute	metagenome id	metadatum to color by
shape attribute	biome	metadatum to determine shape
PC X	6	set the principle x-component
PC Y	4	set the principle y-component
distance method	braycurtis	method for distance matrix
perform normalization	yes	normalize the datasets

layout



Custom Analysis

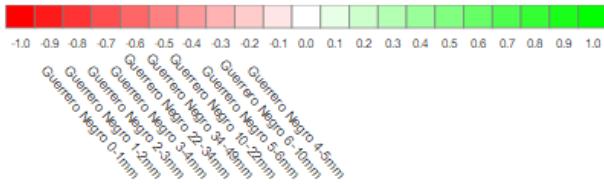
heatmap of analysis 1 [?]

adjust graph data

metadatum metagenome name

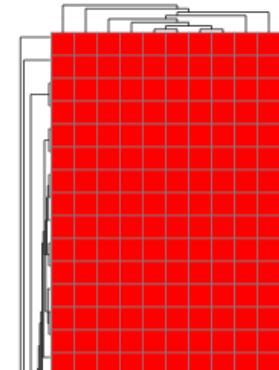
metadatum to name the datasets by

layout

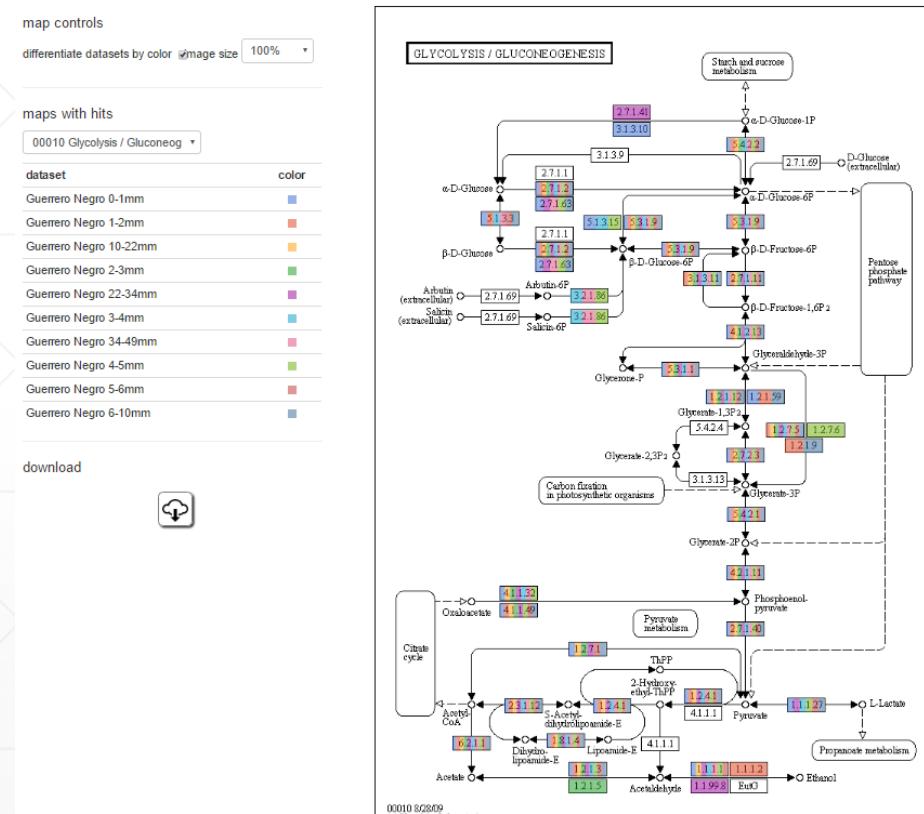


Guerrero Negro 4.5mm
Guerrero Negro 6-10mm
Guerrero Negro 10-22mm
Guerrero Negro 22-34mm
Guerrero Negro 34-45mm
Guerrero Negro 1.2mm
Guerrero Negro 2.5mm
Guerrero Negro 0.1mm

Microsporidia
Phaeophyceae
Platyhelminthes
(her sequences)
Xanthophyceae
Chytridiomycota
Placozoa
Hemichordata
Nanoarchaeota
Echinodermata
Thaumarchaeota
Korarchaeota
Apicomplexa
Cnidaria
Bacillariophyta



Custom Analysis



API

- <https://github.com/MG-RAST/MG-RAST-Tools>
- <http://api.metagenomics.anl.gov/api.html>
- <ftp://ftp.metagenomics.anl.gov/manual.pdf>

