MG-RAST Tutorial

Moon Kim and Nick Falkowski

Agenda

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

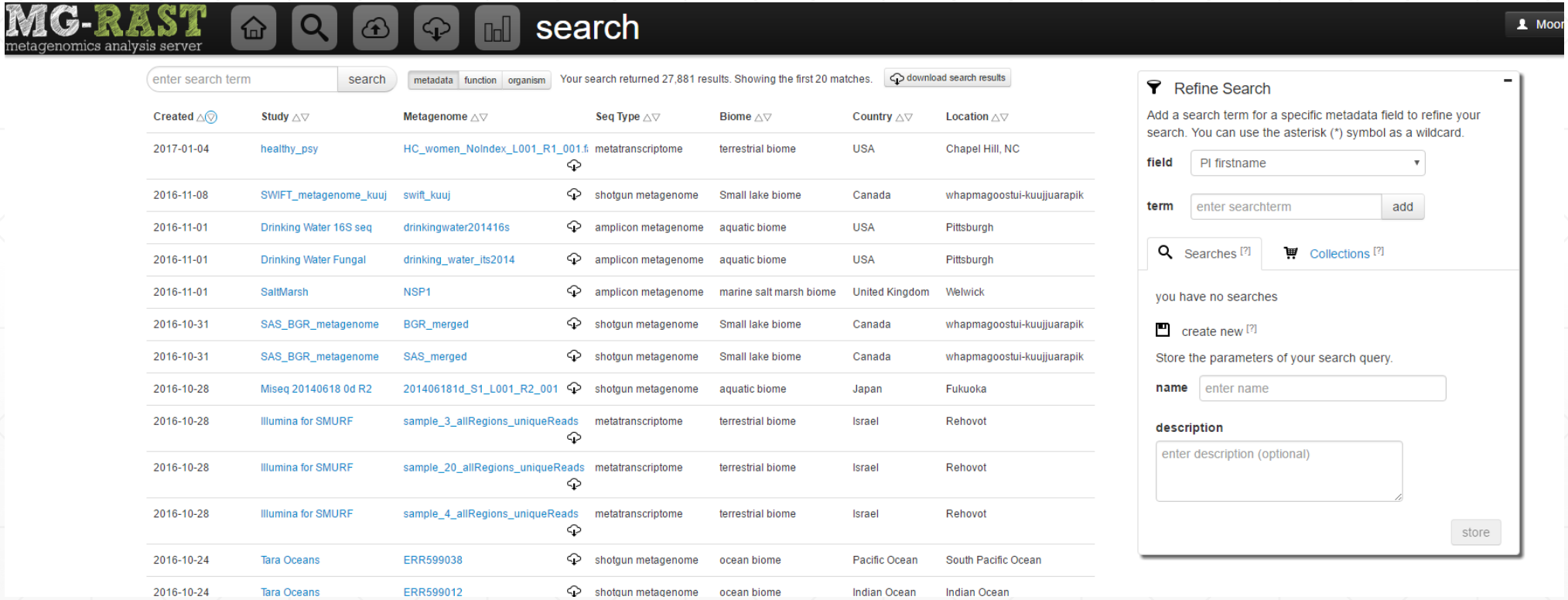
Introduction

- 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_NoIndex_L001_R1_001.fa 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 201406181d_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

-

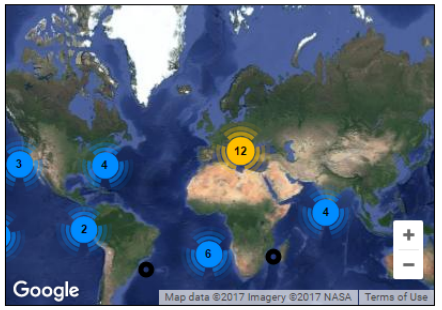
funding source

-

contact

Administrative
Coordinators Tara Oceans Consortium (kandels@embl.de)
Tara Oceans Consortium (-)
Heidelberg, Germany, Germany

Technical
- - (-)
- (-)
-, -



metagenomes

create collection

name	bp count	seq. count	material	sample	library	location	country	coordinates	type	method	download
ERR599135	6,938,884,806	52,709,744	water	mgs556807	mgl556809	South Atlantic Ocean	Atlantic Ocean	-20.4091, -3.1759	WGS	illumina	metadata submitted results
ERR599010	5,129,921,923	36,449,856	water	mgs556759	mgl556761	South Atlantic Ocean	Atlantic Ocean	-20.9354, -35.1803	WGS	illumina	metadata submitted results
ERR599129	6,127,630,616	43,744,550	water	mgs556801	mgl556803	South Atlantic Ocean	Atlantic Ocean	-31.0266, 4.665	WGS	illumina	metadata submitted results
ERR598968	28,412,269,137	178,227,284	water	mgs556729	mgl556731	North Atlantic Ocean	Atlantic Ocean	34.6712, -71.3093	WGS	illumina	metadata submitted results
ERR599052	46,461,017,602	286,911,517	water	mgs556795	mgl556797	North Pacific Ocean	Pacific Ocean	35.3671, -127.7422	WGS	illumina	metadata submitted results
ERR315861	6,785,017,005	42,781,232	water	mgs556705	mgl556707	Mediterranean Sea	Mediterranean Sea	42.2038, 17.715	WGS	illumina	metadata submitted 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 ▾ Send: ▾ **Filters:** Manage Filters

Items: 1 to 20 of 10530

<< First < Prev Pag

1. Uncultured organism clone SBXZ_6525 16S ribosomal RNA gene, partial
1,364 bp linear DNA
Accession: JN437545.1 GI: 364589549
GenBank FASTA Graphics

2. Uncultured organism clone SBXZ_6524 16S ribosomal RNA gene, partial
1,261 bp linear DNA
Accession: JN437544.1 GI: 364589548
GenBank FASTA Graphics

3. Uncultured organism clone SBXZ_6523 16S ribosomal RNA gene, partial
1,132 bp linear DNA
Accession: JN437543.1 GI: 364589547
GenBank FASTA Graphics

Complete Record
Coding Sequences
Gene Features

Choose Destination

File Clipboard
Collections

Download 10530 items.

Format
FASTA

Sort by
Default order

Show GI

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

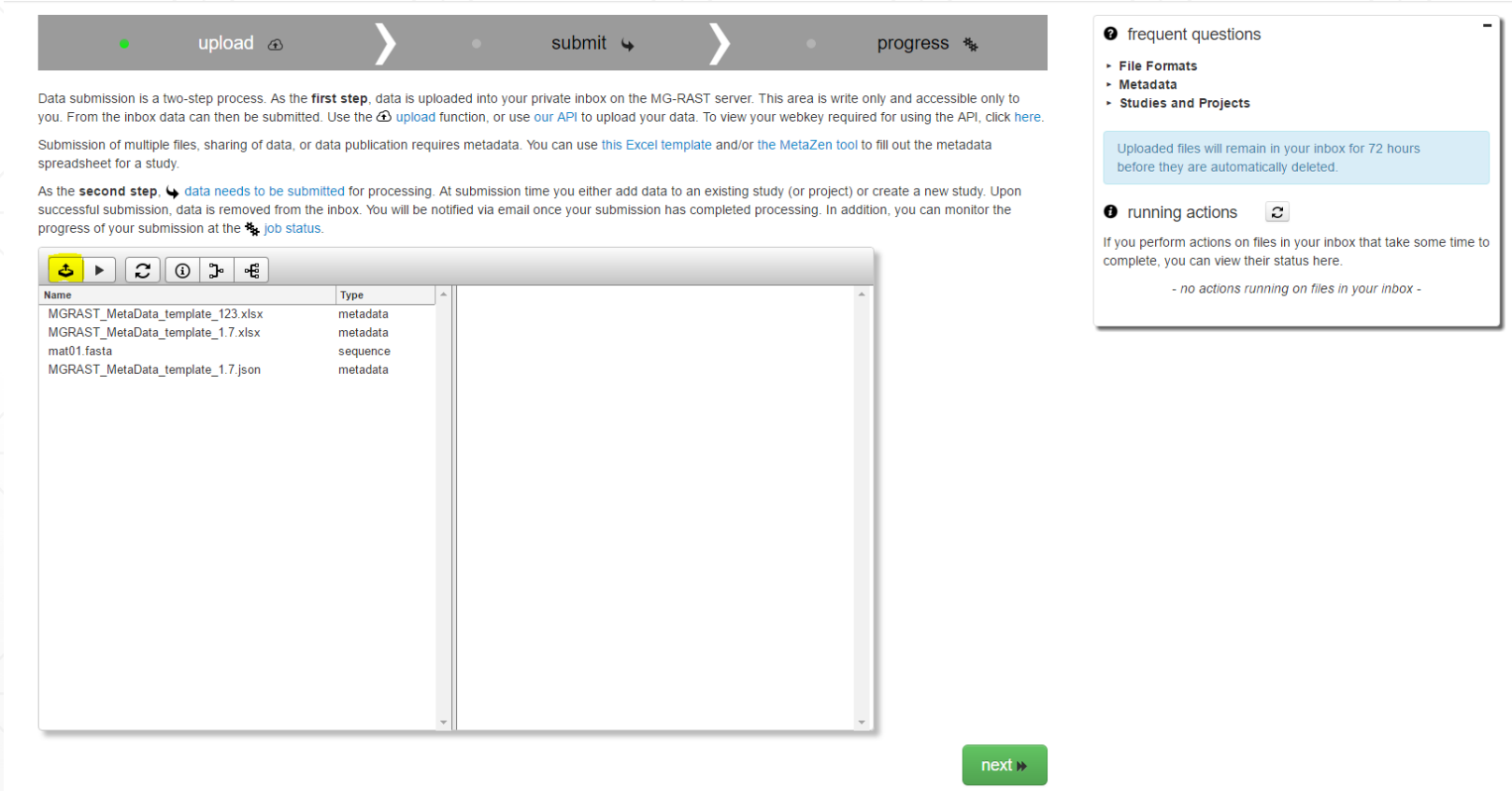
Formatting data for MG-RAST

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

Uploading MetaData

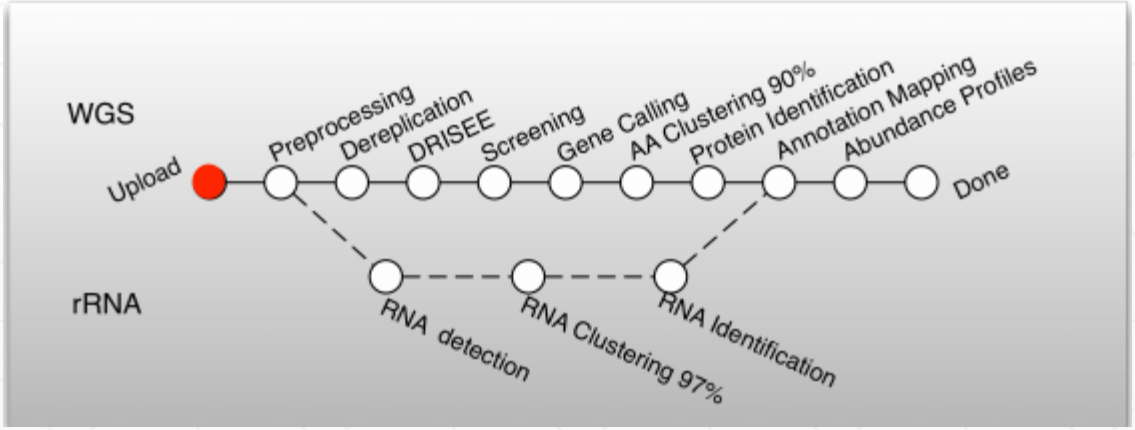
- MetaData not necessary but strongly encouraged
- Datasets with metadata given priority in the pipeline
- MetaData in the form of GSC standard compliant checklists (http://gensc.org)

Uploading Data



MG-Rast Pipeline

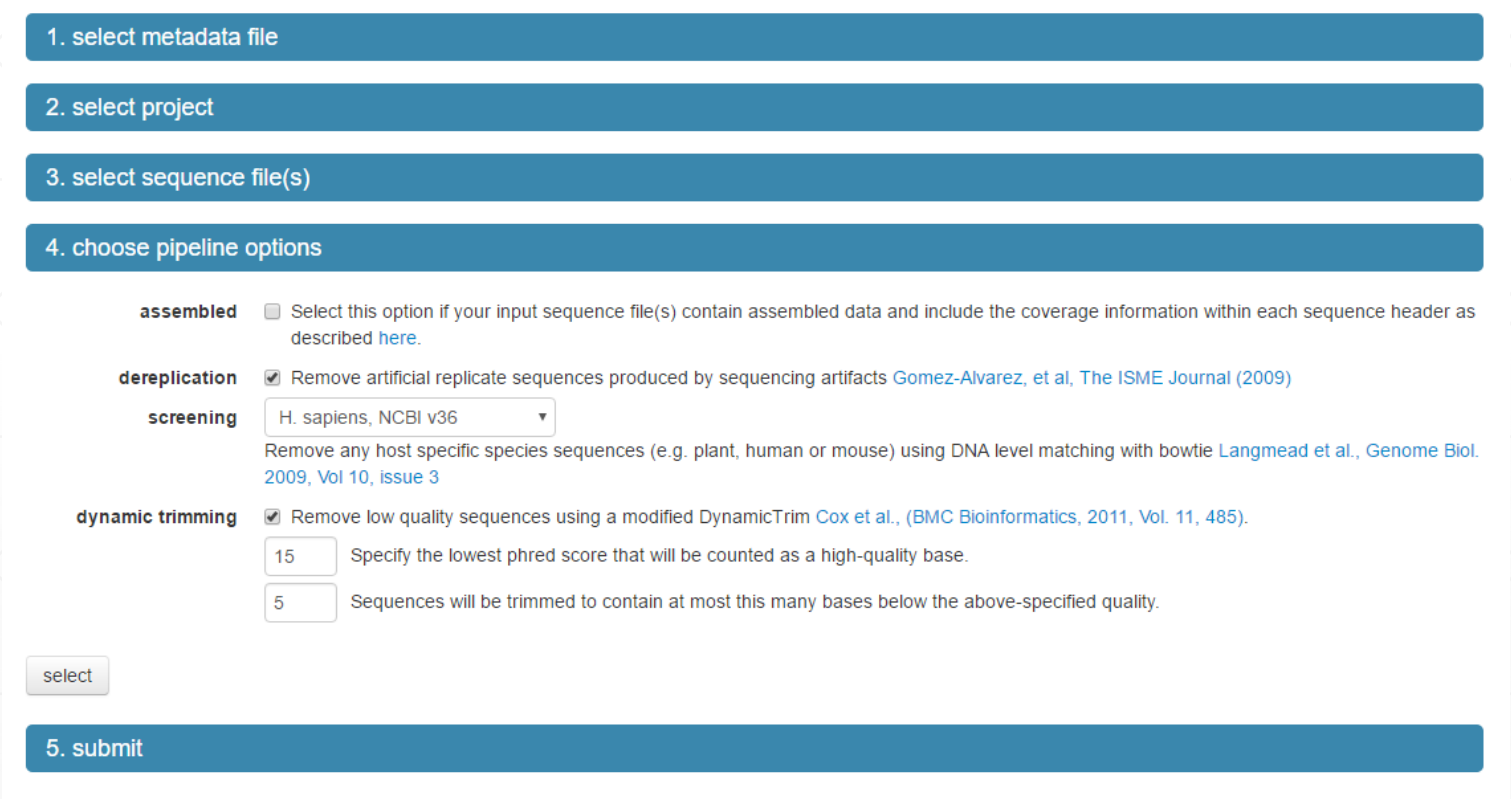
- Performs quality control, protein prediction, clustering and similarity based annotation on nucleic acid sequence datasets using a number of bioinformatics tools
- Data on MG-RAST is private to the submitting user unless shared with other users or made public by the user. Public data given highest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



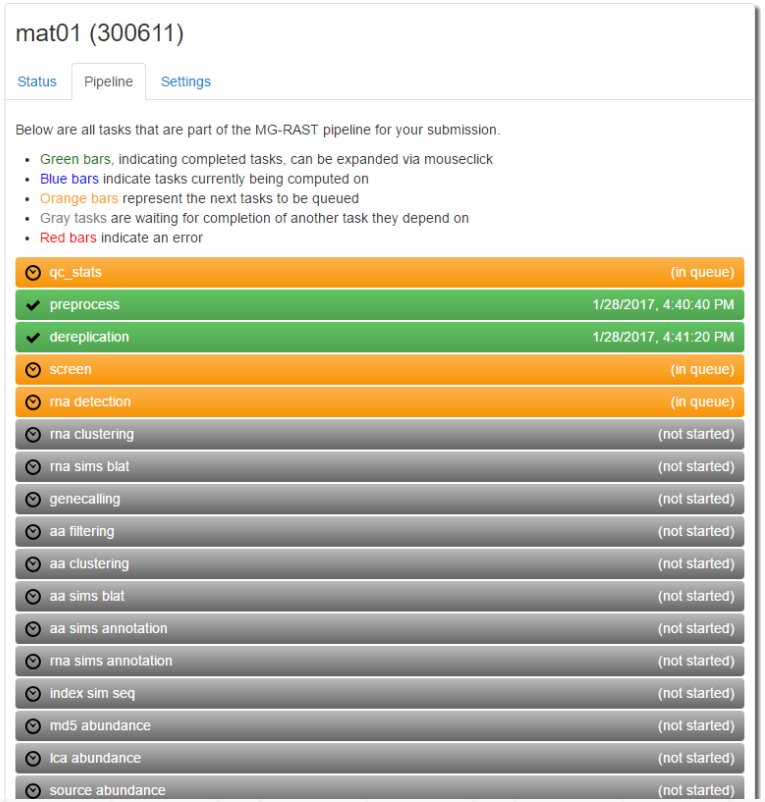
MG-Rast Pipeline

- 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



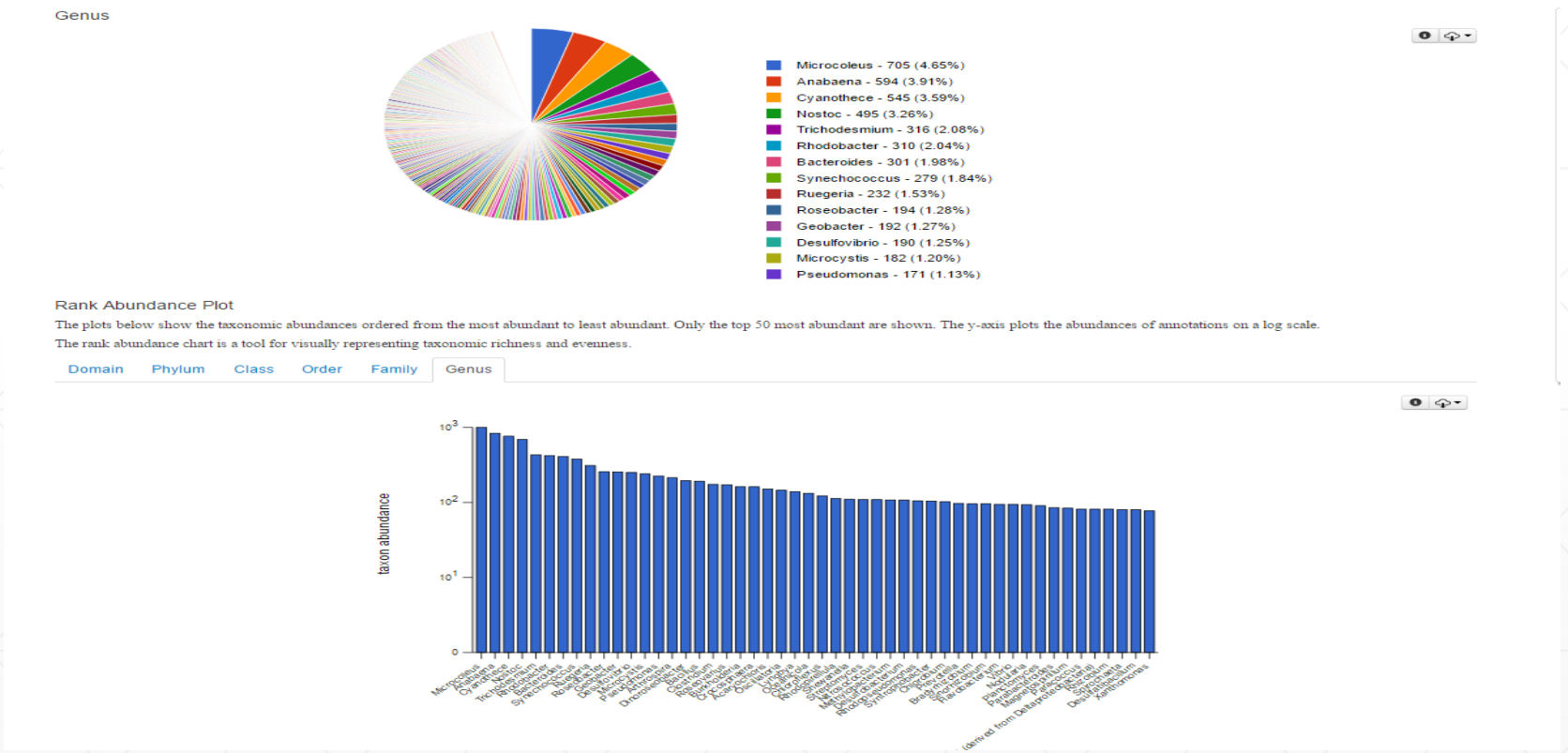
MG-Rast Pipeline



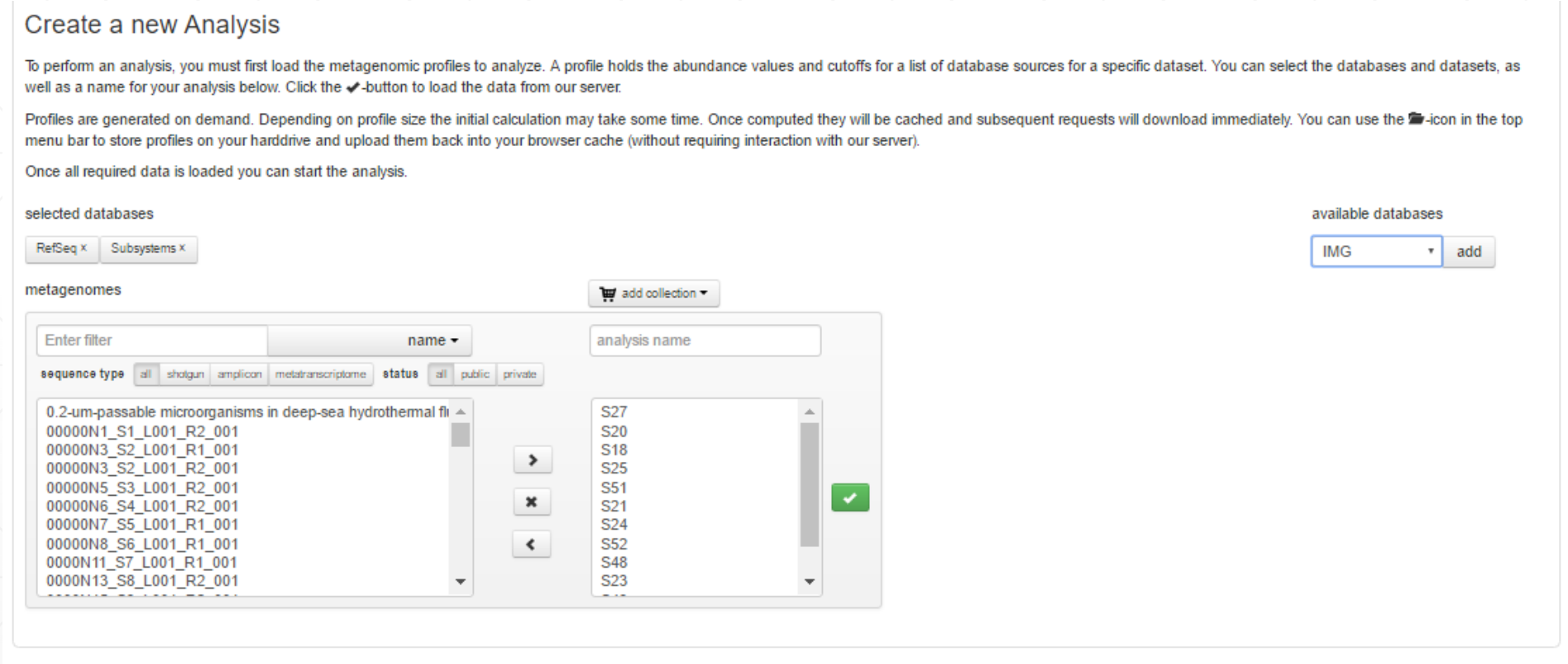
MG-Rast Data Products

- 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

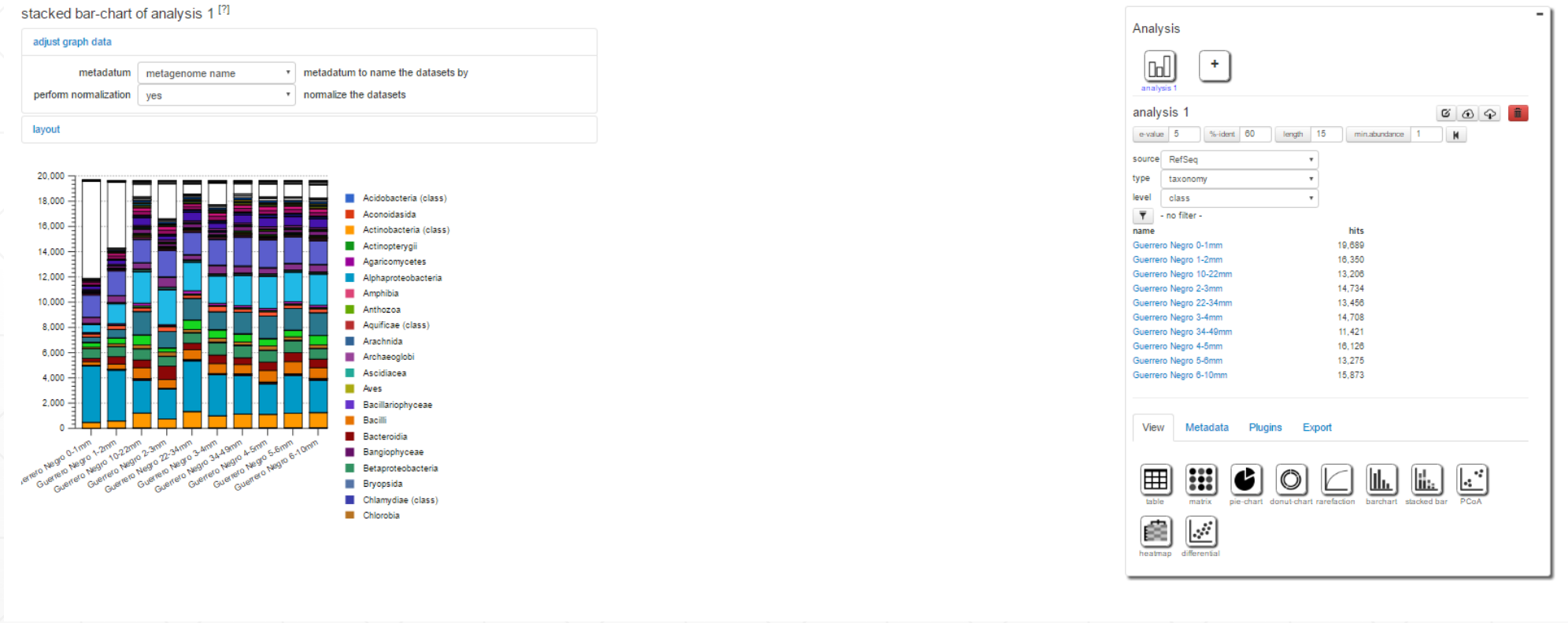
MG-Rast Data Products



Custom Analysis



Custom Analysis



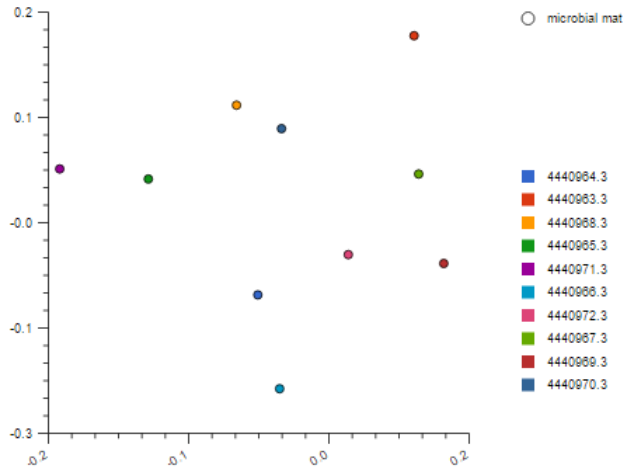
Custom Analysis

PCoA of analysis 1 [?]

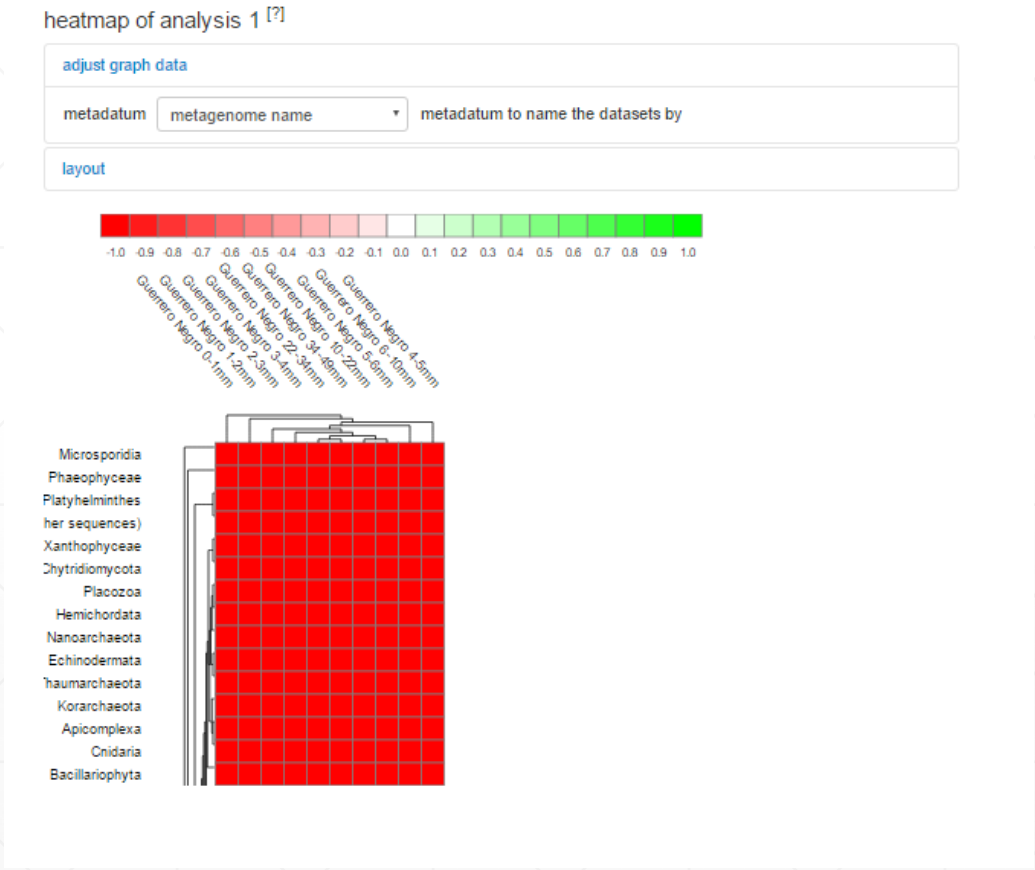
adjust graph data

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



Custom Analysis

map controls

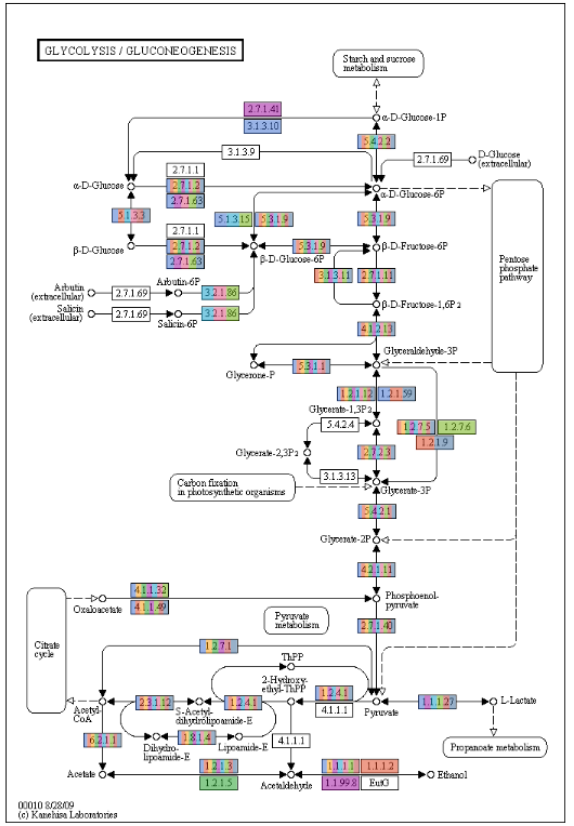
differentiate datasets by color ☑mage size 100%

maps with hits

00010 Glycolysis / Gluconeog

dataset	color
Guerrero Negro 0-1mm	
Guerrero Negro 1-2mm	
Guerrero Negro 10-22mm	
Guerrero Negro 2-3mm	
Guerrero Negro 22-34mm	
Guerrero Negro 3-4mm	
Guerrero Negro 34-49mm	
Guerrero Negro 4-5mm	
Guerrero Negro 5-6mm	
Guerrero Negro 6-10mm	

download



API

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