METAREP

JCVI Metagenomics Reports

website www.jcvi.org/metarep
source code http://github.com/jcvi/METAREP
blog http://blogs.jcvi.org/tag/metarep
contact metarep-support@jcvi.org

About

What METAREP Is

JCVI Metagenomics Reports (METAREP) is a new **open source tool for high-performance comparative metagenomics**. It provides a suite of web based tools to help scientists to **view, query, browse, and compare** metagenomics annotation data derived from ORFs called on metagenomics reads or assemblies.

METAREP supports browsing of functional and taxonomic assignments. Users can either specify fields, or logical combinations of fields to **filter and refine datasets**. Users can compare multiple datasets at various functional and taxonomic levels applying statistical tests as well as hierarchical clustering, multidimensional scaling ,and heatmaps.

For each of these features **tab delimited files can be exported** for downstream analyses. The web site is optimized to be user friendly and fast.

Highlights

- **Handle extremely large datasets**. Uses scalable high-performance search engine (we have indexed 300 million annotation entries, but much larger volumes can be handled).
- **Compare 20+ datasets at the same time**. Use various compare options including statistical tests and plot options to visualize dataset differences at various taxonomic and functional levels.
- **Use statistical tests** such as METASTATS (White *et al.*, 2009), a modified non-parametric t-test, to compare two sample populations (e.g. metagenomics samples from healthy and diseased individuals).
- **Export publication-ready graphics**. Export heatmaps, hierarchical clustering, and multi-dimensional scaling plots in PDF format.
- **Analyze KEGG metabolic pathways**. Summaries include enzyme highlights on KEGG maps, pathway enzyme distributions and statistics about pathway coverage at various pathway levels.
- **Search using a SQL-like query syntax**. Build your query using 14 different fields that can be combined logically.
- **Drill down into data** using METAREP's NCBI Taxonomy, Gene Ontology, Enzyme Classification or KEGG Pathway browser.
- **Install your own METAREP version**. Flexible central configuration, METAREP and 3rd party code base is completely open source.
- **Cross-link function with phylogeny**. Slice your data at various taxonomic and/or functional levels. E.g. search for all bacteria or exclude eukaryotes or search for a certain (GO/EC ID)/taxonomic combination.
- **Generic data format**. Data types that can be populated include a free text functional description, best BLAST hit information as well as GO ID, EC ID, and HMMs.

Under the hood - the METAREP search engine

METAREP uses the open source enterprise search platform Solr/Lucene for extremely fast querying of large metagenomics datatsets. Currently, we have indexed 68 million documents distributed over 330 index files. Much larger index volumes can be handled as shown by HAthi Trust, a digital library, which currently indexes 227 terabytes of data. Our current Solr server setup peaks at a query response performance of 3,000 search requests per second (benchmark was carried out for a 8 million document index). To improve query response time, METAREP can be configured to run on two Solr servers in a index replication and load balancing set-up (see section 1.3).

©J. Craig Venter Institute

Contents

©J. Craig Venter Institute

List of Figures

List of Tables

Chapter 1

How to set up METAREP

1.1 Manual Installation

For questions/feedback please send an email to metarep-support@jcvi.org

Note, that although NETAREP components can run on many operating systems, METAREP has only been successfully deployed and run on a linux operating system.

Step 1. Install required 3rd party tools METAREP uses the following open source tools that need to be installed prior to use of METAREP. Please refer to the individual tool pages for installation instructions.

- Java Virtual Machine - if suitable, a HotSpot 64-Bit Server VM (build 1.5.0) is recommended [download]
- Apache Solr/Lucene 1.4.1 Search Server [download tgz] [download zip] [installation guide]
- MySQL Database Server - version 5.1.42 is recommended [download] [installation guide]
- R Statistical Software - version 2.8.0 is recommended [download]
- Apache Web Server - version 2.2.14 is recommended [download]
- PHP Version - version 5.2.11 is recommended [download]

CentOs 5.5 installation

- install minimal CentOs 5.5 version [download]
- use the CentOs yum package manager to install the following packages
  - mysql.i386 5.0.77-4.el5_5.3
  - R.i386 2.10.0-2.e15
  - perl.i386 4:5.8.8-32.el5_5.1
  - httpd.i386 2.2.3-43.el5.centos.3
  - php.i386 5.2.10-1
- Java Virtual Machine - if suitable, a HotSpot 64-Bit Server VM (build 1.5.0) is recommended [download]
- Apache Solr/Lucene 1.4.1 Search Server [download tgz] [download zip] [installation guide]

Step 2. Download and install METAREP source code

- download the latest stable/tagged METAREP source code [download]
- untar/unzip source code
- create a metarep folder under your Apache Web Server directory
  `mkdir /<APACHE_WEB_HOME>/htdocs/metarep`
- move source code content to the new metarep folder
- check that the directory structure under `/<APACHE_WEB_HOME>/htdocs/metarep` contains an `app` and a `cake` directory
- the path to this metarep directory will subsequently be referred to as `<METAREP_HOME>`

Step 3. Set up METAREP MySQL database

- start up MySQL server if not already running
- if your have installed MySQL as root and have not yet assigned a password, create a password by executing
  `mysqladmin -u root password <password>`
- create the metarep MySQL database by typing
  `mysql -u root -p -e "create database metarep"` in the command line or
  `create database metarep` using the MySQL client download
- import the metarep MySQL dump file using
  `mysql -u root -p metarep < /<METAREP_HOME>/mysql/metarep_mysql_dump.sql`

Step 4. Set up Gene Ontology MySQL database

- create a gene_ontology MySQL database by typing
  `mysql -u root -p -e "create database gene_ontology"` in the command line or
  `create database gene_ontology` using the MySQL client
- import gene ontology MySQL dump file using
  `mysql -u root -p gene_ontology < /<METAREP_HOME>/mysql/gene_ontology_mysql_dump.sql`.

  Alternatively, if you wish to import the latest Gene Ontology version, download the latest term db dump file from the gene ontology website [download] and import the `graph_path` and `term` database tables into the gene_ontology database. The other database tables are not needed.

Step 5. Set up METAREP Apache Solr/Lucene instance

- cd into your Apache Solr installation directory, e.g. apache-solr-1.4.1 (subsequently referred to as `<SOLR_HOME>` directory)
  `cd /<SOLR_HOME>`
- create metarep-solr directory within your Solr home directory
  `mkdir /<SOLR_HOME>/metarep-solr`
- copy the METAREP Solr configuration files into the new metarep-solr directory
  `cp -a /<METAREP_HOME>/solr/* /<SOLR_HOME>/metarep-solr`

©J. Craig Venter Institute

- cd into the example folder of your Solr home directory
  `cd /<SOLR_HOME>/example`
- start jetty web server. Adjust port (-Djetty.port) and java maximum heap size (-Xmx) if needed.
  `java -Djetty.port=1234 -Dsolr.solr.home=/<SOL_HOME>/metarep-solr -jar -Xms156m -Xmx14000m start.jar`
- test it at `http://localhost:1234/solr`. If the page shows **Welcome to Solr!** everything has been correctly set up.

Step 6. Import example datasets

- tab delimited example files can be found in `/<METAREP_HOME>/data/tab`
- the import script `metarep_loader.pl` can be found in the `/<METAREP_HOME>/scripts/perl` directory.
- the script takes several arguments. Adjust the arguments to match your MySQL and Solr connection parameters. The specified tmp directory will be used to store intermediate XML files. Execute the loading script
  `perl /<METAREP_HOME>/scripts/perl/metarep_loader.pl -project_id=1 -project_dir=/<METAREP_HOME>/data/tab -tmp_dir=/<tmp-dir> -mysql_host=<host>:<port> -metarep_db=metarep -metarep_username=<username> -metarep_password=<password> -solr_url=http://localhost:1234 -solr_home_dir=/<SOLR_HOME> -solr_instance_dir=/<SOLR_HOME>/metarep-solr`
- check if load was successful. Go to your Solr default admin page at `http://localhost:1234/solr` (adjust host/port if needed). You should see 7 links that represent the newly loaded datasets. Click on a dataset. Enter `protein` into the query text field and hit the search button. If the load was successful, you should see XML formatted search results. In addition, you can check the number of loaded documents per dataset by using the statistics link at the top menu for each dataset admin page. It displays the number of documents in the stats section. If the document number is greater than zero, your load was successful.

Step 7. Adjust METAREP configuration files

- edit the METAREP configuration file to match your web server and Solr configuration */<METAREP_HOME>/app/config/metarep.php*. Detailed information about each of the variables can be found here
- edit the METAREP database configuration file to connect to your metarep and gene_ontology MySQL databases */<METAREP_HOME>/app/config/database.php*. (see also section 1.3).

Step 8. Configure and start Apache web server

- modify the `httpd.conf` webserver configuration file to work with mod_rewrite module
  Uncomment the LoadModule rewrite_module line so that it looks like this

```
# LoadModule foo_module modules/mod_foo.so
LoadModule php5_module modules/libphp5.so
LoadModule rewrite_module libexec/apache2/mod_rewrite.so
```

- add the following metarep directory configuration

```
<Directory /metarep>
Options FollowSymLinks
AllowOverride All
</Directory>
```

- start Apache web server
  `/<APACHE_WEB_HOME>/bin/apachectl start`
- go to `http://localhost:80/metarep` to see the METAREP welcome page

Step 9. Create New Project

- login with username:admin and password:admin
- click on New Project in the top menu
- enter project name: *HOTS vertical ocean depth profile*
- enter project description: 'Planktonic microbial communities in the North Pacific Subtropical Gyre, from the ocean's surface to near-sea floor depths'
- hit the submit button
- go to `http://localhost:80/metarep/projects/view/1` to analyze the project's datasets

Step 10. Change METAREP admin password

- login with username:admin and password:admin
- click on change password on the dashboard page and enter your new admin password

1.2 Virtual Machine Installation

We have created an open virtualization format (OVF) package that bundles all 3rd party tools and is configured to run out of the box in a virtual machine. To run a virtual version of METAREP on your machine, follow these steps

1. download the METAREP OVF package from our FTP site at `ftp://ftp.jcvi.org/pub/software/metarep/vm`
2. unzip the OVF package
3. download and install Oracle's Virtual Box, a OVF compatible virtualization software [download]
4. start Virtual Box
5. click File/Import Appliance and select the OVF file.
6. adjust RAM/CPU usage using the Appliance Import Wizard (see image)
7. start VM
8. double-Click on the METAREP firefox link on the VM desktop
9. log into METAREP with username=admin and password=admin

1.3 Configuration

METAREP Configuration Files

METAREP Configuration File You can configure METAREP by editing the METAREP configuration file which is located in `<installation-dir>/METAREP/app/conf/metarep.php`. The following example configuration file shown in Listing 1.1 describes the variables that can be defined.

Listing 1.1: METAREP Configuration File

```
<?php
/*************************************************************
* File: metarep.php
* Description: METAREP configuration file
*
* PHP versions 4 and 5
*
* METAREP : High-Performance Comparative Metagenomics Framework (http://www.jcvi.org/metarep)
* Copyright(c)  J. Craig Venter Institute (http://www.jcvi.org)
*
* Licensed under The MIT License
* Redistributions of files must retain the above copyright notice.
*
* @link http://www.jcvi.org/metarep METAREP Project
* @package metarep
* @version METAREP v 1.2.0
* @author Johannes Goll
* @lastmodified 2010-09-16
* @license http://www.opensource.org/licenses/mit-license.php The MIT License
**/

/**
 * METAREP Version
 *
 */

define('METAREP_VERSION','1.2.0-beta');

/**
 * METAREP Running Title
 *
 * customize your METAREP application title. It is used by Browser as the window title,
 * the default web layout uses it besides the METAEP logo. The title
 * is used at various other places throughout the application.
 */

define('METAREP_RUNNING_TITLE','JCVI Metagenomics Reports');

/**
 * METAREP Web Root
 *
 * Point this variable to your Apache METAREP webroot directory
 * Default: /<your-installation-dir>/apache-2.2.14/htdocs/metarep
 */

define('METAREP_WEB_ROOT','/<your-installation-dir>/apache-2.2.14/htodocs/metarep');

/**
 * METAREP Url Root
 * Default: http://localhost:80/metarep
 */
```

```
define('METAREP_URL_ROOT','http://localhost:80/metarep');

/**
 * Directory to store temporary files
 *
 * Temporary files include CAKEPHP cache/application and R files
 * Default: /tmp
 */

define('METAREP_TMP_DIR','/tmp');

/**
 * Solr instance dir
 *
 * Contains Solr configuration files in conf/  subdirectory
 * Default: /your-installation-dir>/apache-solr-1.4.0/metarep-solr
 */

define('SOLR_INSTANCE_DIR','/<your-installation-dir>/apache-solr-1.4.0/metarep-solr');

/**
 * Solr port
 *
 * Defines the Solr port
 * Default: 1234
 */

define('SOLR_PORT','1234');

/**
 * Solr data dir
 *
 * Defines location of Solr index files
 * Default: /<your-installation-dir>/apache-solr-1.4.0/metarep-solr/data
 */

define('SOLR_DATA_DIR','/<your-installation-dir>/apache-solr-1.4.0/metarep-solr/data');

/**
 * Solr master server host
 *
 * Takes on role of the Solr master server in a
 * load balanced/replication set-up.
 * Default: localhost
 */

define('SOLR_MASTER_HOST','localhost');

/**
 * Solr slave server host
 *
 * Define the Solr slave host if you use METAREP
 * in a load balanced/replication set-up
 */

//define('SOLR_SLAVE_HOST','');

/**
 * Solr big ip; define if you use a
 * Define Solr BIG-IP if you use METAREP in a
 * load balanced/replication set-up
 */

//define('SOLR_BIG_IP_HOST','');
```

```
/**
 * FTP host
 *
 * Specify FTP host if you like to provide
 * additional data for your METAREP dataset
 */

//define('FTP_HOST','');

/**
 * FTP suser name
 */

//define('FTP_USERNAME','');

/**
 * FTP password
 */

//define('FTP_PASSWORD','');

/**
 * Email to send bug reports and feature requests.
 *
 * Email is displayed if METAREP can not access the Solr or
 * MySQL servers. It is also used to provide users an Email
 * address send bug reports and feature requests.
 * Default: metarep-support@jcvi.org
 */

define('METAREP_SUPPORT_EMAIL','metarep-support@jcvi.org');

/**
 * Internal Email Extension
 *
 * METAREP distinguishes between four types of users:
 * ADMIN, INTERNAL, EXTERNAL, and PUBLIC.
 *
 * ADMIN and INTERNAL users can access all METAREP datasets, while
 * EXTERNAL and PUBLIC have restricted access. The variable defines
 * the Email extension that is used to identify INTERNAL users. This
 * is especially helpful if you like to grant dataset access to all
 * users of your institution - just specify your institute's email
 * extension, e.g. jcvi.org for the J. Craig Venter Institute.
 */

//define('INTERNAL_EMAIL_EXTENSION','');

/**
 * Number of Top Facet Counts
 *
 * The METAREP search and browse pages summarize annotation data
 * types in the form of sorted top ten lists. Change this variable
 * to increase/decrease the number of top hits shown for each data type.
 * Default: 10
 */

define('NUM_TOP_FACET_COUNTS',10);

/**
 * Number of Search Results
 *
 * The METAREP search page displays pages of found annotation results.
 * By default, ten hits are shown per page. Change this variable to
 * increase/decrease the number of results that are shown for each
 * result page.
```

```
 * Default: 10
 */

define('NUM_SEARCH_RESULTS',10);

/**
 * Number of METASTATS bootstrap permutations
 * Used for estimating null distribution of the
 * METASTATS t statistic.
 */

define('NUM_METASTATS_BOOTSTRAP_PERMUTATIONS',1000);

/**
 * Path to R Executable
 *
 * Define the path to your R executable
 * Default: /usr/local/bin/R
 */

define('R_PATH','/usr/local/bin/R');

/**
 * Path to Rscript Executable
 *
 * Define the path to your Rscript executable
 * Default: /usr/local/bin/Rscript
 */

define('RSCRIPT_PATH','/usr/local/bin/Rscript');

/**
 * Activate/Deactivate JCVI-only features
 *
 * Sett this variable to 1, activates JCVI-only
 * features that access JCVI resources that are
 * not included in this distribution.
 * Default: 0
 */

define('JCVI_INSTALLATION',0);
?>
```

Database Configuration File You can configure MySQL database connection parameters and the Blog RSS feed by editing the database configuration file which is located in `<installation-dir>/METAREP/app/conf/database.php`. The following Listing describes the variables that can be defined:

Listing 1.2: METAREP Database Configuration File

```
<?php
/*********************************************************
* File: database.php
* Description: configuration file for METAREP datasources
*
* PHP versions 4 and 5
*
* METAREP : High-Performance Comparative Metagenomics Framework (http://www.jcvi.org/metarep)
* Copyright(c)  J. Craig Venter Institute (http://www.jcvi.org)
*
* Licensed under The MIT License
```

```
* Redistributions of files must retain the above copyright notice.
*
* @link http://www.jcvi.org/metarep METAREP Project
* @package metarep
* @version METAREP v 1.2.0
* @author Johannes Goll
* @lastmodified 2010-07-09
* @license http://www.opensource.org/licenses/mit-license.php The MIT License
**/

class DATABASE_CONFIG {

  //METAREP MySQL database connection parameters
  var $default = array(
    'driver' => 'mysqli',
    'persistent' => true,
    'host' => 'localhost',
    'login' => '<your-login>',
    'password' => '<your-password>',
    'database' => 'metarep',
  );

  //GO MySQL database connection parameters
  var $go = array(
    'driver' => 'mysqli',
    'persistent' => true,
    'host' => 'localhost',
    'login' =>  '<your-login>',
    'password' =>  '<your-password>',
    'database' => 'gene_ontology',
  );

  //METAREP Blog connection parameters
  var $blog = array(
    'datasource' => 'rss',
    'feedUrl' => 'http://blogs.jcvi.org/tag/metarep/feed/',
    'encoding' => 'UTF-8',
    'cacheTime' => '+1 day',
  );
}
?>
```

Solr Index Replication and Load Balancing

METAREP works well with a single Solr server. However, if you expect to have a lot of traffic (many concurrent users) you may want configure METAREP to use two load-balanced Solr servers to improve query response time). In such a setup, two Solr servers replicate each others index files and user traffic is balanced between the two using a load balancer. This is achieved by defining one Solr server as a master server and another Solr server as a slave server (in theory more than two slave servers can be defined but METAREP currently supports only one slave server). New Index files that are submitted to the master server will automatically be replicated by the Solr slave server using Solr's inbuilt replication functionality. To configure METAREP for index replication and load balancing:

- repeat Step 4 (except server start up) of the Installation Guide (see section 1.1) to create two separate Solr server installations (master & slave) .
- uncomment the replication requestHandler (lines 505-520) in the master solrconfig.xml file. The file can be found at `/<master-installation-dir>/apache-solr-1.4.1/metarep-solr/conf/solrconfig.xml`
- adjust the masterUrl (line 496) in the master solrconfig_slave.xml file. The file can be found at `/<master-installation-dir>/apache-solr-1.4.1/metarep-solr/conf/solrconfig_slave.xml`

- copy your edited solrconfig_slave.xml file located on the master server to the slave server. Rename the copy on the slave to solrconfig.xml (override existing solrconfig.xml).
  `cp <master-installation-dir>/apache-solr-1.4.1/metarep-solr/conf/solrconfig_slave.xml <slave-installation-dir>/apache-solr-1.4.1/metarep-solr/conf/solrconfig.xml`
- start the master server
  `java -Djetty.port=1234 -Dsolr.solr.home= /<SOLR_HOME_MASTER>/metarep-solr -jar -Xms156m -Xmx14000m start.jar`
- start the slave server
  `java -Djetty.port=1234 -Dsolr.solr.home= /SOLR_HOME_SLAVE>/metarep-solr -jar -Xms156m -Xmx14000m start.jar`
- update the Solr Slave (SOLR_SLAVE_HOST) and Solr Master host (SOLR_MASTER_HOST), ports (SOLR_PORT) and load balancer IP (SOLR_BIG_IP_HOST) in the METAREP configuration file located at `<installation-dir>/METAREP/app/conf/metarep.php` (see also Listing 1.1).

Chapter 2

How to annotate and import data

METAREP does not allow the importation of sequence data. For users to analyze their own raw metagenomics sequences or assemblies in METAREP, users need to annote their sequences first. However, if users have already run blast or HMM homology searches, they do not have to rerun the searches and can parse the existing results to load their data into METAREP.

Step 1 Annotation The METAREP tab-delimited data format contains the most common metagenomics annotation data types (Table 2.1). For each sequence, you can assign a sequence ID, a library ID and functional description (fields 1-4), GO, HMM and EC data types (fields 5 -9) and best Blast Hit information (fields 10-13). In addition, a categorical filter tag can be assigned to a sequence. This field is useful, for example, to tag sequences as duplicates and artifacts to filter them out later on the fly. Some of the fields may contain several values. E.g. a sequence may have several HMMs, GO IDs or functional descriptions. If a field can contain multiple values is indicated in the last column of Table 2.1.

Table 2.1: METAREP Data Format.

Column	Field ID	Description	Type/Range	Example	Multi-Valued
01	peptide_id	Unique peptide ID	text	1120333534885	no
02	library_id	Library ID	alphanumeric	GS-00a-01-01-2P5KB	yes
03	com_name	Functional Description	text	periplasmic sugar binding protein	yes
04	com_name_src	Common Name Source	text	RF|YP_167057.1	yes
05	go_id	Gene Ontology ID	text	GO:0009265	yes
06	go_id_src	Source of Gene Ontology assignment	text	PF02511	yes
07	ec_id	Enzyme Commission ID	text	2.1.1.148	yes
08	ec_id_src	Source of Enzyme Commission ID	text	PRIAM	yes
09	hmm_id	Hidden Markov Model hits (e.g. PFAM/TIGRFAM)	text	246194	yes
10	blast_taxon	NCBI Taxon ID	integer	246194	yes
11	blast_evalue	BLAST E-Value	double	1.78E-20	no
12	blast_pid	BLAST Percent Identity	double	0.93	yes
13	blast_cov	BLAST sequence coverage of shortest sequence	double	0.82	yes
14	filter	Any filter tag (categorical variable)	text	duplicate	yes

Users can either use existing annotation pipelines (Li, 2009; Meyer *et al.*, 2008; Tanenbaum *et al.*, 2010) or create/use their own pipeline.

To produce all METAREP data types, a basic pipeline could consist of the following steps:

1. run metagene (Noguchi *et al.*, 2008, 2006) to identify prokaryotic and phage coding genes; translate coding genes
2. blast peptide sequences (blastp) against a comprehensive protein database (set the BLAST output format to tab format for easier parsing).
3. in parallel, run HMM searches against the Pfam and TIGRFAM HMM collections
4. parse BLAST results to assign a common name (fields 3-4), taxonomy, E-value, percent identity and sequence coverage (fields 10-13).
5. parse HMM results to assign hmm ids (field 9).
6. mapp HMMs to GO terms and EC numbers using hmm to go and hmm to ec mappings (fields 5-9).
7. summarize all assignments in one tab delimited file. The name of the file is used to assign the dataset name in METAREP.

JCVI's current annotation pipeline workflow is given in Figure 2.1. To load several annotation files at a time, store all files in one project directory.

Step 2. Import Annotations

- login as admin
- click on New Project in the top menu to create a new project
- click on List Projects in the top menu
- copy the project ID of your new project (shown in the first column)
- the import script metarep_loader.pl can be found in `/<installation-dir>/METAREP/scripts/perl` directory
- import tab delimited files stored in your project directory. Specify your MySQL and Solr server connection parameters. Use the `-tmp` argument to specify a temporary directory to store intermediate XML files.

```
perl /<installation-dir>/METAREP/scripts/perl/metarep_loader.pl
-project_id=<your-project-id>
-project_dir=<path to the root directory that contains your tab-delimited annotation
files>
-tmp_dir=/<tmp-directory>
-mysql_url=<host>:<port>
-metarep_db=metarep -metarep_username=<username>
-metarep_password=<password>
-solr_url=http://localhost:1234
-solr_home_dir=/<installation-dir>/apache-solr-1.4.0
-solr_instance_dir=/<installation-dir>/apache-solr-1.4.0/metarep-solr
```

- click on List Projects in the top menu. Your project should now have one library.
- click on the View link to start analyzing your dataset.

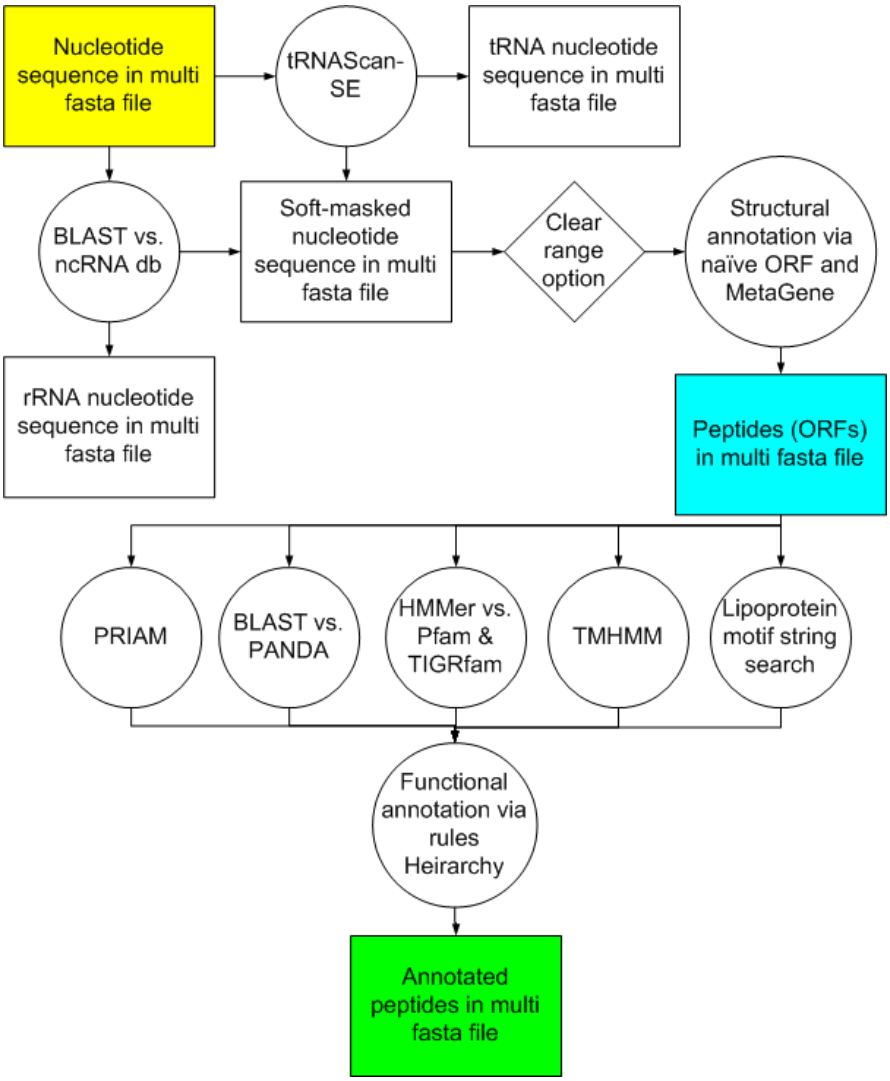


Figure 2.1: JCVI Metagenomics Annotation Pipeline Workflow
taken from (Tanenbaum *et al.*, 2010)

Chapter 3

How to use METAREP

3.1 Glossary

METAREP distinguishes between *libraries*, *populations* and *datasets* that can be analyzed and organized into *projects*. There is one METAREP *admin user* and multiple *project admin users* who can specify how project datasets are shared among registered users.

Library In METAREP the term library stands for a basic sequence collection whose annotation is loaded into METAREP. A library could be, for example, a set of annotations derived from a 454 half plate, a Sanger based library, or an assembly. All analysis functionality is available for libraries except for the non-parametric t-test (METASTATS) that can only be applied to compare two populations (section 3.4).

Population Annotation data from multiple libraries can be merged by creating populations of libraries. During the population creation process, library annotations are first copied and then merged. Thus if a population is deleted, annotation data of the underlying libraries is still available. All analysis functionality that is available for libraries can be applied to populations. Population are especially helpful to aggregate annotation data across libraries and thus provide access to higher level summaries. A non-parametric t-test (METASTATS) can be used to compare two populations to identify significant differences taking variation among the individual libraries into account (section 3.4).

Dataset Either a library or a population.

Project Multiple libraries and populations can be organized into projects. Each project may have one project admin user who can modify the project's user permissions. Depending on the settings a user either has access to the project and its datasets or not. Projects can also be published and made accessible to any (outside) user.

Admin User The METAREP Admin User has access and can edit all projects and datasets. The admin can specify a project admin user for each project, edit project and library descriptions, and create/delete populations.

Project Admin User A METAREP project admin user has access to a project (of which she/he is project admin) and its datasets and can edit project and library descriptions, modify user permissions and create/delete populations.

Internal User Any user who has a registered and created a METAREP account.

Guest Users Users who browse public datasets in METAREP without creating a user account.

3.2 User Management

Dashboard Page The Dashboard page allows users to login, register, and reset their password if they have forgotten it. It also features general information about METAREP and links to the METAREP JCVI blog.

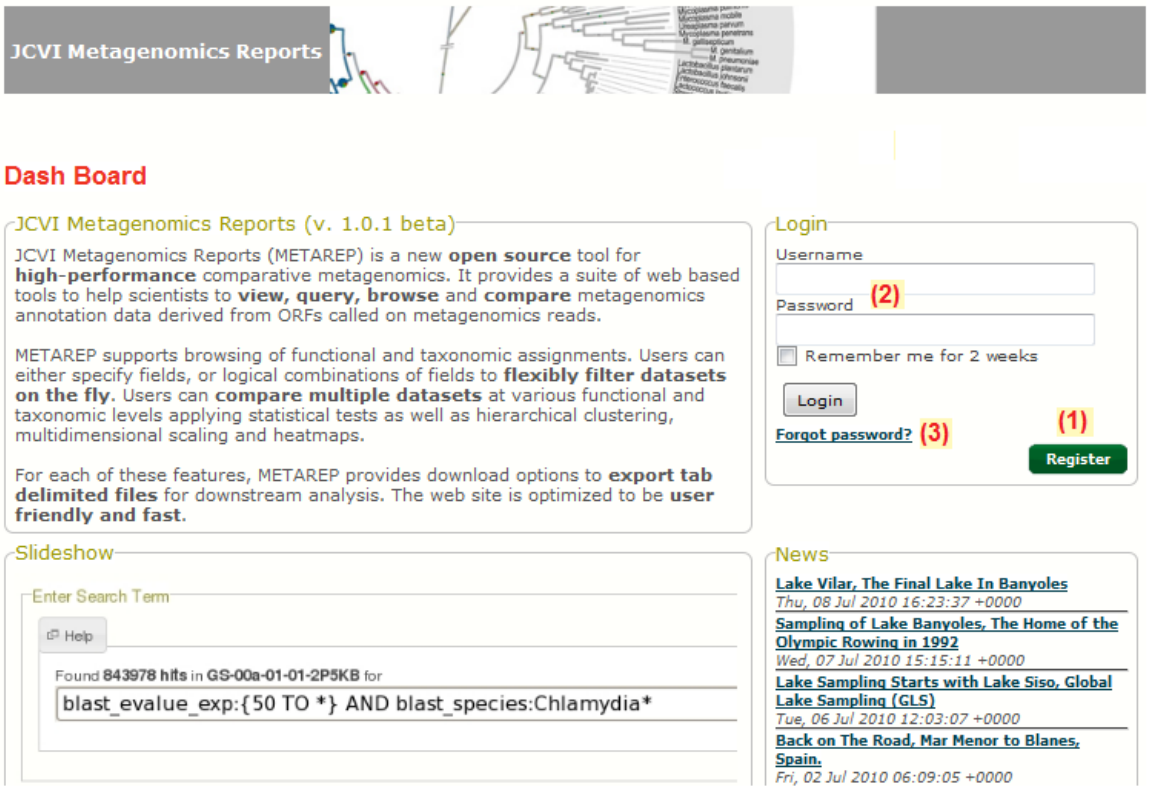


Figure 3.1: Dashboard Page

Options (Figure 3.1):

1. **register** - opens up a new page that allows users to create a METAREP account.
2. **login** - enter your username and password and click the login button. Click 'Remember me for two weeks' to staty logged in for two weeks.
3. **forgot password** - opens up a new page that asks for your email address. After confirmation a password-reactivation link is sent to the specified address. The link opens up a new page to enter and confirm a new password.

Configuration (Listing 1.1):

1. **change institute-wide permissions** - define `INTERNAL_EMAIL_EXTENSION`. Users that register with an email address that ends with the specified extension are registered as Internal METAREP Users with read access to all projects and datasets.
2. **change title** - define `METAREP_RUNNING_TITLE` to change page heading and browse window title.
3. **change blog** - change the blog RSS feed in the database.php configuration file (Listing 1.2) to see your own blog entries on the dashboard page.

User Dashboard Page The User Dashboard page allows users to modify their account information and change their password. It also lists the projects for which the logged in user is a project admin. Project admin users can edit the project description and share the project's datasets with other registered users (Figure 3.3). The METAREP admin user can open a page with all registered users, can edit all project descriptions and can grant project admin user permissions (one user per project).

Figure 3.2: Dashboard Index Page

Options (Figure 3.2):

1. **manage users** - opens a new page that shows a list of all registered users (option is only available if you are logged in as the METAREP admin user).
2. **change account information** - opens up a new page that lets you update your user information.
3. **change password** - lets you change your existing password.
4. **logout** - logs you out of the system and brings you back to the Dashboard page (Figure 3.1).
5. **edit project information** - opens up new page that allows you to edit project information details (option is only available if you are the project admin). The METAREP admin user can assign a project admin user by selecting a user from the drop down list.
6. **manage project permissions** - opens up new page that allows you to grant other registered users access to your project (option is only available if you are the project admin) (Figure 3.3).
7. **feedback** - select type of feedback from drop down (feature request, bug report, data load, other). Enter a message in the textbox and click submit. An email is sent to your email address as well as to the METAREP support email.

Configuration:

- **change feedback email contact** - change the `METAREP_SUPPORT_EMAIL` to redirect user feedback (Listing 1.1).

Manage Project Permissions Page Project admin users can share their project's datasets with other registered users for collaborative data analysis. The page shows all registered users on the right. Users that currently have read access to the project are shown on the left.

Options (Figure 3.3):

Figure 3.3: Manage Project Permissions Page

1. **search users** - enter search term into text box to filter the list of registered users (right panel)
2. **select user** - select a user from the list by clicking the plus symbol or by dragging and dropping the user into the left panel.
3. **select all users** - select all users from the right list.
4. **remove all users** - remove all users from the left list.
5. **remove user** - remove a user from the left list by clicking on the minus symbol.

©J. Craig Venter Institute

3.3 Navigation

Figure 3.4: Main Menu

Main Menu Options (Figure 3.4):

1. **quick navigation** - opens up a menu that allows you to navigate from projects to project datasets to dataset analysis pages (Figure 3.5).
2. **new project** - open up a new page that allows you to create a new project (this option is only visible to the METAREP admin user).
3. **list projects** - opens up a new page that lists all projects for which you have read permissions.
4. **list populations** - opens up a new page that lists all populations that belong to projects for which you have read permissions.
5. **dashboard** - brings you back to your user dashboard page (Figure 3.2).
6. **logout** - logs you out of the system and brings you back to the Dashboard page (Figure 3.1).

Quick Navigation Menu Allows you to quickly navigate from projects to project datasets to a dataset's analysis pages. The first level shows all projects ordered alphabetically. The second level shows all of a project's datasets (populations and libraries). The third level shows analysis actions that are available for the respective dataset.

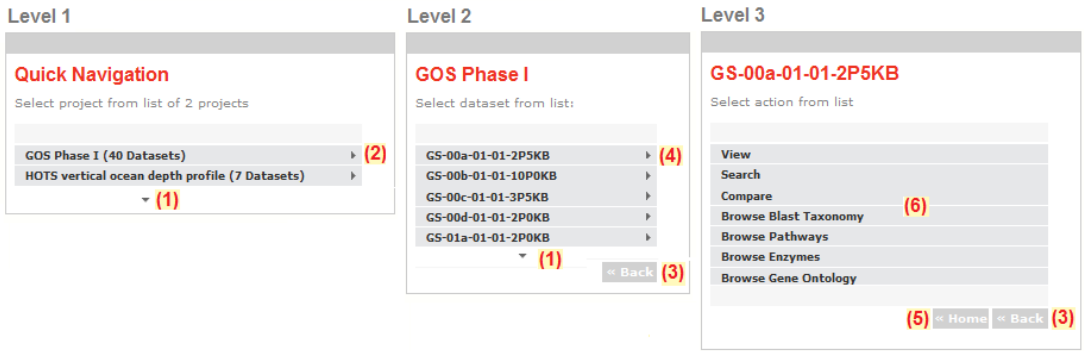


Figure 3.5: Quick Navigation Menu

Options (Figure 3.5):

1. **scroll down** - scroll down list. A scroll up arrow shows up if you have scrolled passed more than 20 items.
2. **select project** - brings you to the next level which shows the project's datasets (populations and libraries).
3. **go back** - brings you back to the previous level.

4. **select dataset** - brings you to the next level which shows analysis actions that are available for the respective dataset.
5. **go home** - brings you back to the project level.
6. **analyze dataset** - select from the set of analysis actions that are available for the respective dataset (library or population).

View Project Page Shows basic project information and lists project populations and libraries along with the number of annotation entries and sample meta information. If you are the project's admin user you are able to update library information and create populations of libraries.

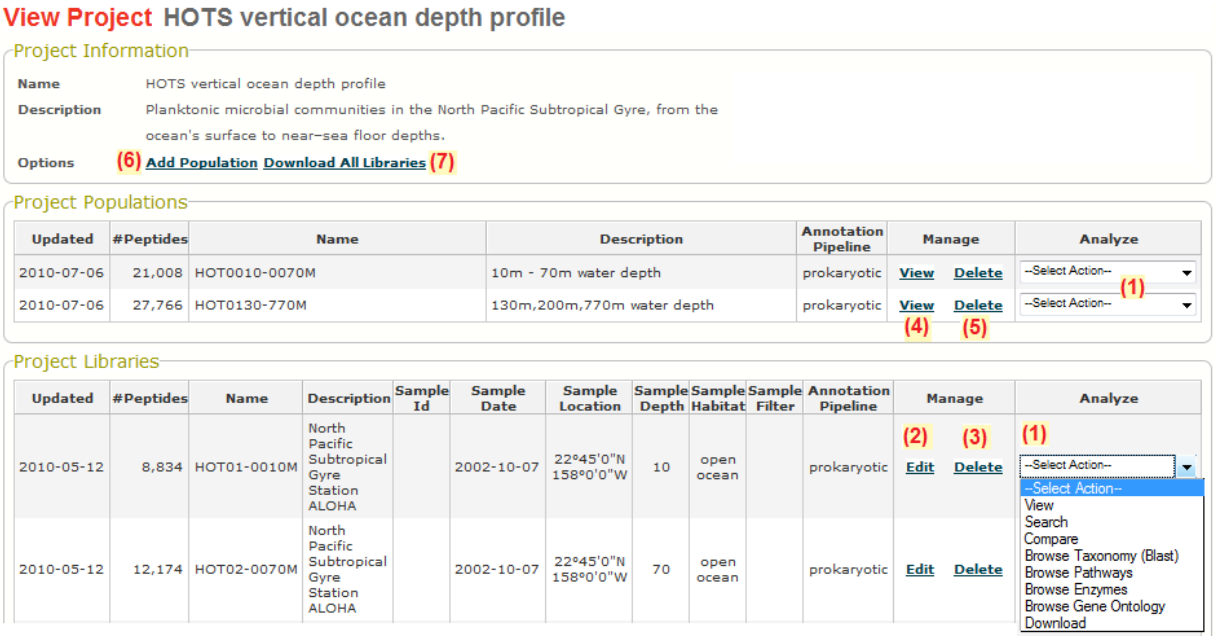


Figure 3.6: View Project Page

Options (Figure 3.6):

1. **analyze dataset** - select from the set of analysis actions that are available for the respective dataset (library or population).
2. **edit library** - opens up a new page that lets you edit the library's description field (this option is only visible if you are the project's admin user).
3. **delete library** - deletes the respective library (this option is only visible if you are the project's admin user)
4. **view population** - opens up new page that shows basic population information and lists the population's libraries along with the number of annotation entries and sample meta information (this option is only visible if you are the project's admin user).
5. **delete population** - deletes the respective library (this option is only visible if you are the project's admin user)
6. **add population** - opens up a new page that lets you choose from the project libraries to create a new population (this option is only visible if you are the project's admin user).
7. **download all libraries** - opens up a download dialog that lets you export all project data from the specified METAREP FTP server (this option only visible if data has been uploaded to the FTP server).

©J. Craig Venter Institute

3.4 Analysis

View Dataset The View Dataset page shows the underlying metagenomics annotation data (Figure 3.7) as well as provide high level summaries for each of the metagenomics annotation data types (Figure 3.8). Summaries include top species, KEGG metabolic pathways , Gene Ontology terms, Enzyme Classification IDs, HMMs (such as TIGRFAM and Pfam HMMs) and common names. For each of these data types, a tab with a ranked list and a bar chart with the relative frequencies for the respective data type category is displayed. Users can adjust the number of ranks displayed (up to 1,000 ranks) and download the data in tab delimited format.

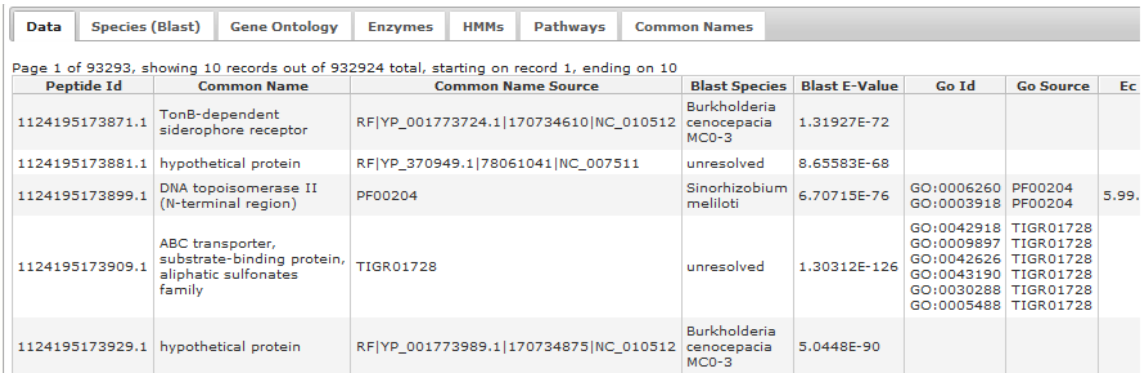


Figure 3.7: View Dataset Page - Data Tab

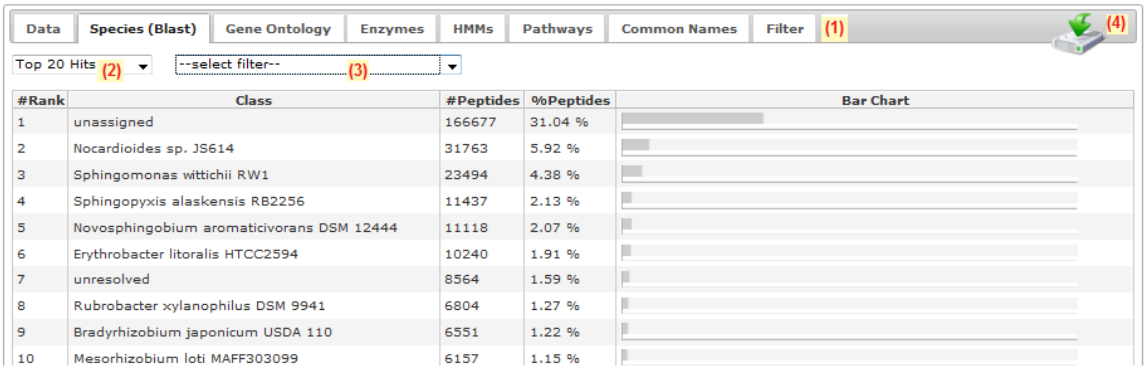


Figure 3.8: View Dataset Page - Species (Blast) Tab

Options (Figure 3.8):

1. **select annotation data type** - click on a tab label to show summary statistics for the respective annotation data type (here Species (Blast)).

2. **adjust top ranks** - select top 10, 20, 50, 100 or 1000 hits from the left upper drop down to adjust the number of ranks shown for the selected annotation data type. Once a level has been selected it is remembered and used for the other tabs as well.

3. **adjust filter** - select filter category from filter drop down (only shown if dataset contains filter tags). After selecting the filter, the view pages are updated and show statistics for only the subset of the data that is tagged with the respective filter type. Once a filter has been selected it is remembered and used for the other tabs as well. To unselect a filter, select [–select filter–].

4. **download** - opens download dialog to export absolute and relative counts of top ranks for the respective data type.

Browse Dataset Dataset annotations can be browsed using four distinct hierarchies: NCBI Taxonomy (Figure 3.9), Gene Ontology , Enzyme Classification and KEGG metabolic pathways (Figure 3.10). The number of hits are displayed for each node in the tree, and a user can click on a tree node and expand further. After clicking a node, a summary of that node is shown in the right panel featuring a pie chart calculated from its sub-nodes and top lists of functional and taxonomic assignments. Here we describe the Browse Taxonomy and Browse Pathway pages. The Browse Enzyme Classification and Browse Gene Ontology pages provide the same options that are available for the Browse Taxonomy pages (Figure 3.9).

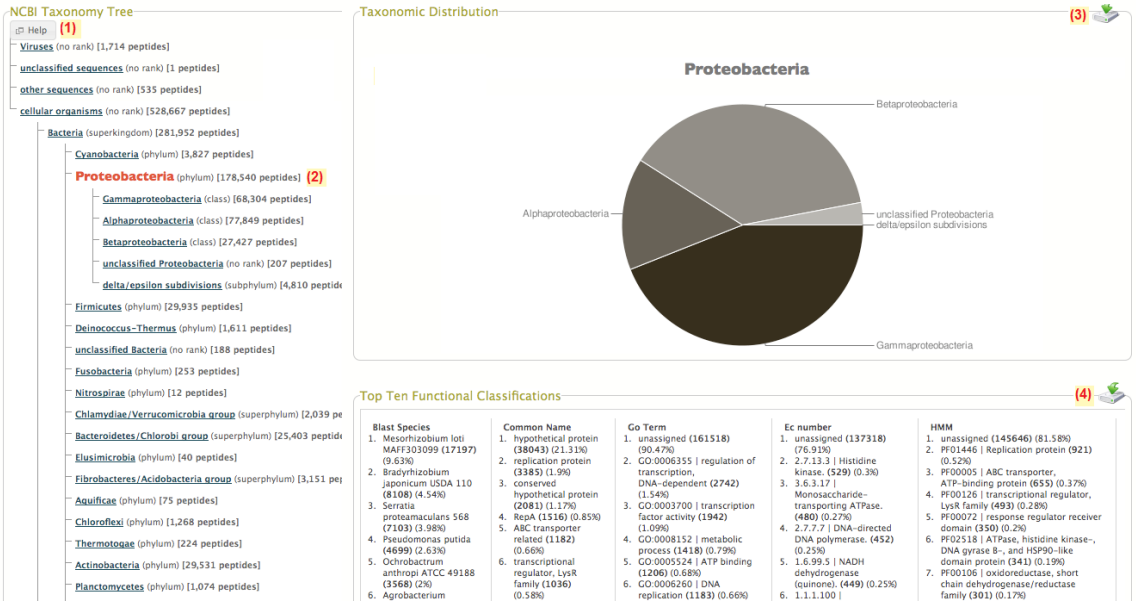


Figure 3.9: Browse NCBI Taxonomy Page

Options (Figure 3.9):

1. **open help dialog** - opens a help dialog that explains browse options.
2. **expand node** - click on a node to expand the tree and see a node's sub-nodes. The right panel is updated and a summary of that node is shown in the right including a pie chart summarizing sub-nodes as well as top lists of functional and taxonomic assignments.
3. **download sub-node counts** - opens download dialog to export absolute and relative counts of sub-nodes of the selected node.
4. **download top lists of functional and taxonomic assignments** - opens download dialog to export absolute and relative counts for top 10 functional and taxonomic assignments for that node.

Options (Figure 3.10):

1. **open help dialog** - opens a help dialog that explains browse options.
2. **expand node** - click on a node to expand the tree and see a node's sub-nodes. The right panel is updated and a summary of that node is shown on the right including a pie chart summarizing sub-nodes we well as top lists of functional and taxonomic assignments.
3. **download sub-node counts** - opens download dialog to export absolute and relative counts of sub-nodes of the selected node.
4. **look-up KEGG enzyme information** - clicking on an Enzyme within the KEGG pathway map updates the pathway panel with a KEGG page that shows additional enzyme information. Click your web browser's back button to get back to the original pathway map.

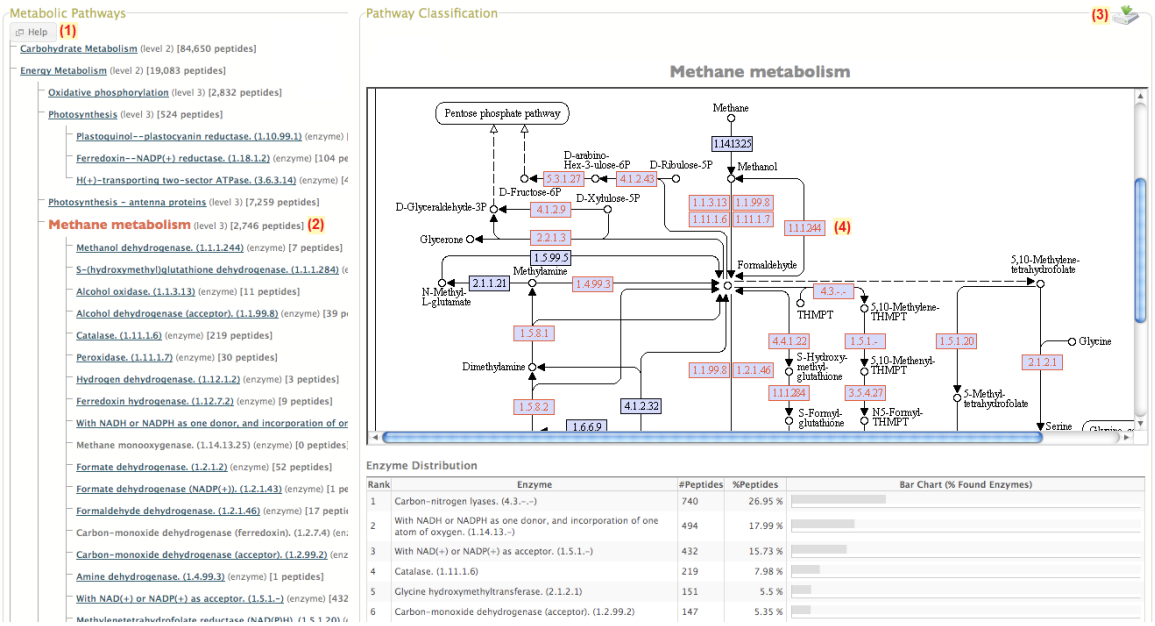


Figure 3.10: Browse KEGG Pathways Page

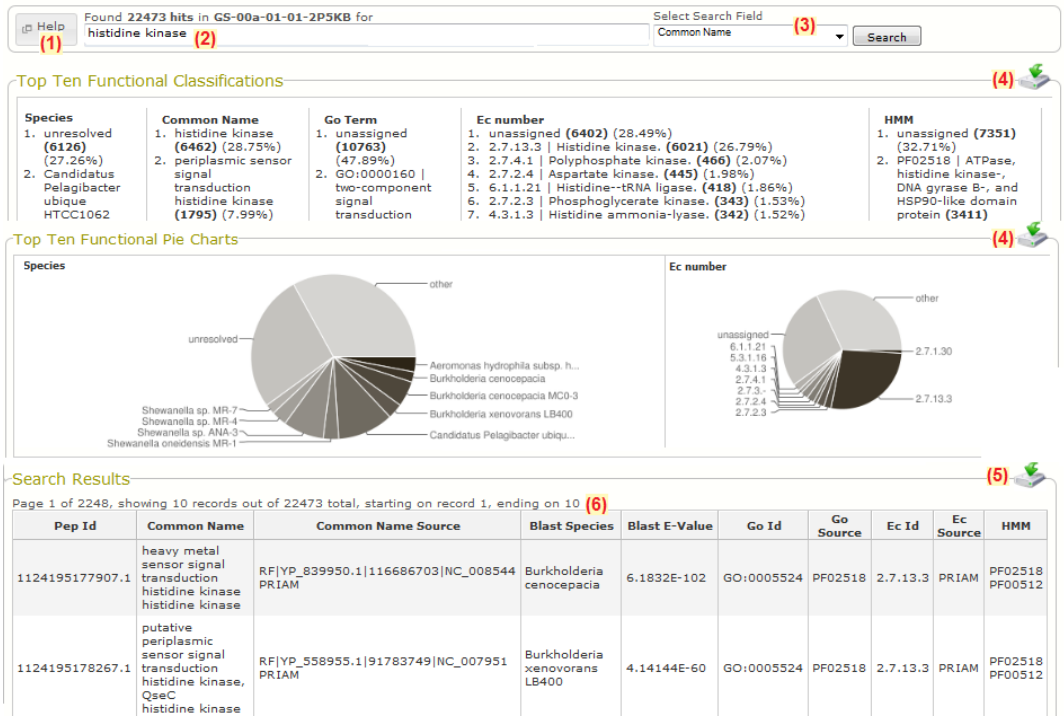
Search Dataset Users can use a SQL like query syntax including logical combinations of annotation fields to filter datasets (Figure 3.11). For example a user may filter results based on the BLAST E-Value or combination of BLAST E-Value and percent identity, search for only bacterial species, or choose to exclude results that have BLAST hits to eukaryotes (Table 3.1). The search returns results as well as frequency count lists and pie charts, that summarize the top functional and taxonomic categories for the identified subset. Counts and identifiers can be exported as tab delimited files.

Options (Figure 3.11):

1. **open help dialog** - opens a help dialog that explains search fields.
2. **enter search term** - enter a search term and select the corresponding search field (option 3) or enter a Lucene query by typing the field name followed by a colon and then the term you are looking for `<field-name>:<search-term>`. Supported fields and example queries are listed in Table 3.1. If no field has been specified, the default field com_name_txt (common name) is being searched.
3. **select search field** - select a search field from the drop down that matches your search term. If no field has been selected, the default field com_name_txt ((common name)is being searched. After clicking the search button, search results are returned and your entered search term is automatically translated into the Lucene query language (`<selected-field-name>:<search-term>`) for you to refine.
4. **download top lists of functional and taxonomic assignments** - opens download dialog to export absolute and relative counts for top 10 functional and taxonomic assignments of matching annotation entries.
5. **download identifiers** - opens download dialog to export matching identifiers (unique entry IDs).
6. **page through search results** - click on the result page navigation links to page through the results.

Example Queries (see also Table 3.1):

- **blast_tree:1224 AND -blast_tree:28211**
  slice taxonomy to fetch all entries that are *Proteobacteria* (NCBI Taxon ID 1224) but not *Alphaproteobacteria* (NCBI Taxon ID 28211). To search for a certain NCBI Taxon ID go to http://www.ncbi.nlm.nih.gov/Taxonomy.



Pep Id	Common Name	Common Name Source	Blast Species	Blast E-Value	Go Id	Go Source	Ec Id	Ec Source	HMM
1124195177907.1	heavy metal sensor signal transduction histidine kinase histidine kinase	RF|YP_839950.1|116686703|NC_008544 PRIAM	Burkholderia cenocepacia	6.1832E-102	GO:0005524	PF02518	2.7.13.3	PRIAM	PF02518 PF00512
1124195178267.1	putative periplasmic sensor signal transduction histidine kinase, QseC histidine kinase	RF|YP_558955.1|91783749|NC_007951 PRIAM	Burkholderia xenovorans LB400	4.14144E-60	GO:0005524	PF02518	2.7.13.3	PRIAM	PF02518 PF00512

Figure 3.11: Search Page

- **blast_species:unassigned AND hmm_id:unassigned AND ec_id:unassigned**
  to fetch hypothetical proteins, i.e. all entries without any homology hits (no blast hits, HMMs or EC IDs).
- **blast_tree:2 AND blast_evalue_exp:{50 TO *}** AND blast_pid:{0.9 TO *}
  to fetch highly conserved bacterial proteins with a Blast E-value $\leq 10^{-50}$ and percent identity $\geq 90\%$.

Search All Datasets Users can use the same SQL like query syntax to query all datasets and get a breakdown of counts for several sample meta-information categories (Figure 3.12). Summaries include project, sample habitat, sample filter, sample depth and sample location. The search returns a list of datasets ordered by the number of hits per dataset. Top meta-information category counts be downloaded and datasets can be further analyzed by selecting an action from the dataset's analyze drop down

Options (Figure 3.12):

1. **open help dialog** - opens a help dialog that explains search fields.
2. **enter search term** - enter a search term and select the corresponding search field (option 3) or enter a Lucene query by typing the field name followed by a colon and then the term you are looking for `<field-name>:<search-term>`. Supported fields and example queries are listed in Table 3.1. If no field has been specified, the default field com_name_txt (common name) is being searched.
3. **select search field** - select a search field from the drop down that matches your search term. If no field has been selected, the default field com_name_txt ((common name)is being searched. After clicking the search button, search results are returned and your entered search term is automatically translated into the Lucene query language (`<selected-field-name>:<search-term>`) for you to refine.

Table 3.1: METAREP Search Fields

Field Name	Description	Type/Range	Example
Core Annotation Fields			
peptide_id	Peptide ID	text	**peptide_id:1120333534885** *retrieve hit with the specified peptide id*
com_name_txt	Common Name (default field)	text	**com_name_txt:phage** *all hits containing the word phage*
com_name_src	Common Name Source	text	**com_name_src:PRIAM** *all hits having names assigned based on PRIAM*
go_id	Gene Ontology ID	text	**go_id:GO:0000160** *all hits with GO:0000160*
go_tree	Gene Ontology Tree	integer portion of ID	**go_tree:160** *all hits with GO:0000160 or lower (including all hits with GO IDs that are lower (more specific) in the GO hierarchy)*
go_src	Gene Ontology Source	text	**go_src:PF00204** *all hits that have GO terms assigned based on PF00204*
ec_id	Enzyme ID	text	**ec_id:5.99.1.3** *all hits with Enzyme ID 5.99.1.3*
ec_src	Enzyme Source	text	**ec_src:PRIAM** *all hits that have EC IDs assigned based on PRIAM*
hmm_id	HMM ID	text	**hmm_id:PF00204** *all hits that have a PF00204 HMM assignment*
library_id	Library ID	text	**library_id:GS-00a-01-01-2P5KB** *all hits that belong to library GS-00a-01-01-2P5KB (helpful to search for library entries within populations)*
filter	any filter tag (e.g. sequence duplicates)	text	**filter:duplicate** *all hits with filter tagged with duplicate* **-filter:duplicate** *exclude entries with filter tag duplicate*
Best BLAST Hit Fields			
blast_species	Species	text	**blast_species:Chlamydia*** *all Chlamydia species*
blast_tree	Taxonomy	integer (NCBI Taxon ID)	**blast_tree:2** *all bacteria* **-blast_tree:2** *exclude all bacteria*
blast_evalue_exp	Negative E-Value Exponent	positive integer	**blast_evalue_exp:{20 TO *}** *all hits with Blast E-value* $\leq 10^{-20}$ **blast_evalue_exp:{10 TO 20}** *all hits with* $10^{-20} \leq E-value \leq 10^{-10}$
blast_pid	Percent Identity	float between 0-1	**blast_pid:{0.9 TO *}** *all hits with Blast percent identity* $\geq 90\%$ **blast_pid:{0.6 TO 0.8}** *all hits with* $60\% \leq percent\ identity \leq 80\%$
blast_cov	Percent Sequence Coverage	float between 0-1	**blast_cov:{0.8 TO *}** *all hits with Blast percent sequence coverage* $\geq 80\%$ **blast_cov:{0.2 TO 0.3}** *all hits with* $20\% \leq sequence\ coverage \leq 30\%$
Optional Fields			
apis_tree	Taxonomy	integer (NCBI Taxon ID)	**apis_tree:2** *all bacteria* **-apis_tree:2** *exclude all bacteria*
env_lib	Environmental Library	text	**env_lib:Whale*** *all hits that match the whale fall metagenome*

©J. Craig Venter Institute

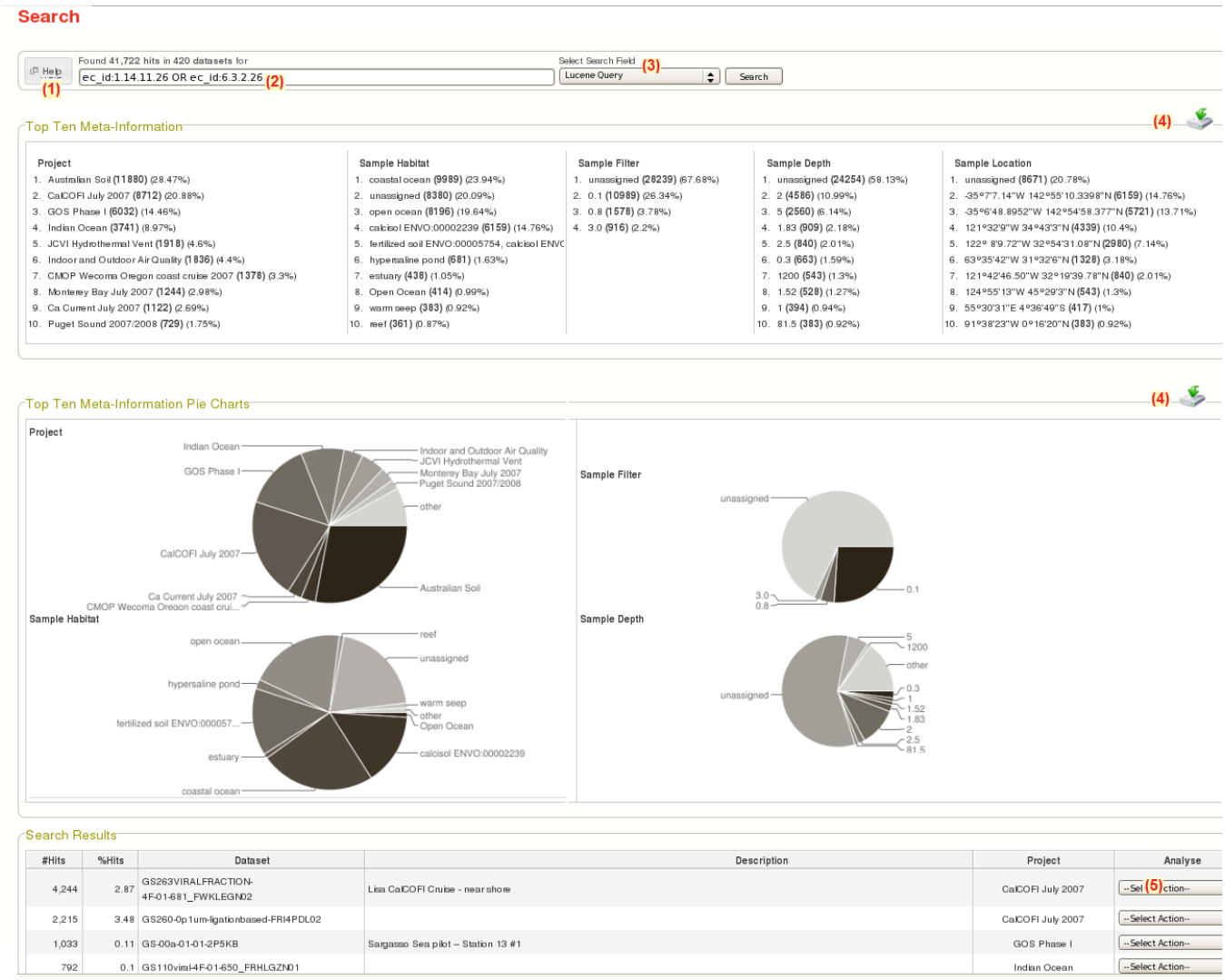


Figure 3.12: Search All Page

4. **download meta-information top lists** - opens download dialog to export absolute and relative counts for top 10 meta-information categories.

5. **analyze dataset** - select from the set of analysis actions that are available for the respective dataset (library or population).

Compare Page The compare page allows each user to compare their own datasets, or compare them with datasets shared by other users (Figure 3.13). The page features a select box to choose datasets. Users can also filter the shown datasets by entering a search keyword. Furthermore, users can specify a query using the Lucene query language (Table 3.1) to filter all selected datasets before the comparison. For example, users may want to only compare hits below a certain E-Value cut off. A minimum absolute count can be entered by users to filter out categories with only a few hits. Various comparison options can be selected such as absolute and relative counts, statistical tests multidimensional scaling, heatmap and hierarchical cluster plots. Similar to the View pages (Figure 3.8), users can choose from several tabs to indicate the annotation data type they wish to compare. Choices are NCBI Taxonomy, Gene Ontology terms, KEGG metabolic pathways, Enzyme Classification, HMMs, and common names. Absolute and relative counts as well as statistics

can be exported in tab delimited format. Graphics can be exported in PDF format.

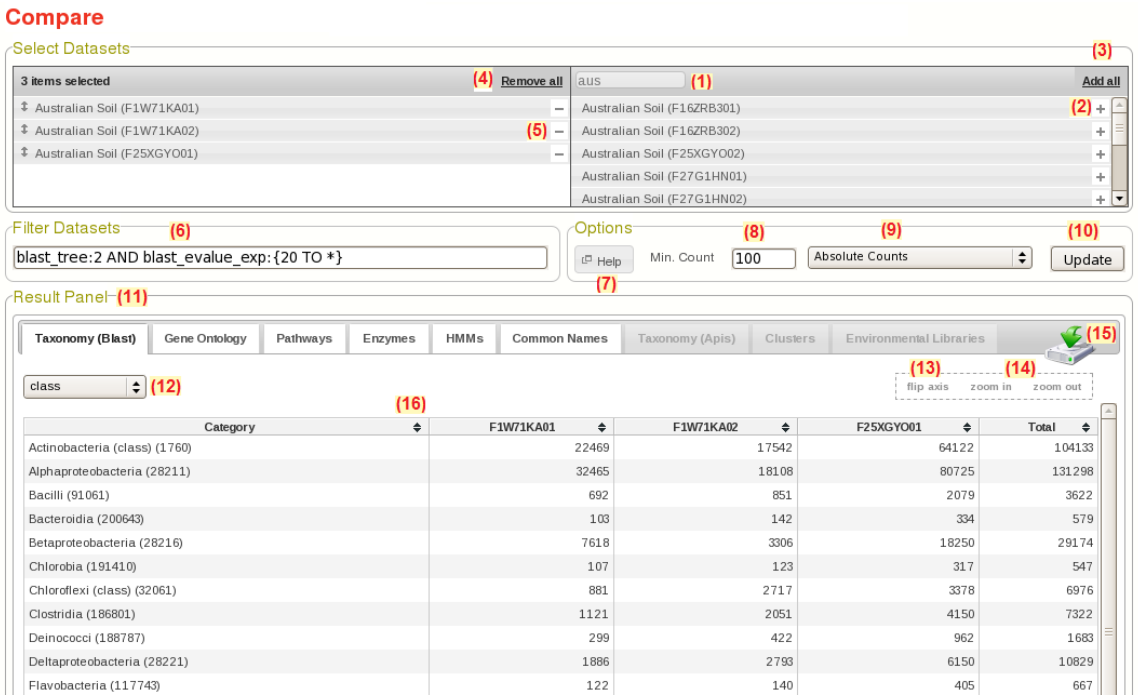


Figure 3.13: Compare Page

Options (Figure 3.13):

1. **search datasets** - enter search term into text box to filter available datasets (shown in the upper right panel).
2. **select dataset** - select a dataset from the list by clicking the plus symbol or by dragging and dropping it into the upper left panel.
3. **select all datasets** - select all datasets from right list.
4. **remove all datasets** - remove all datasets from the left list.
5. **remove dataset** - remove dataset from the left list by clicking on the minus symbol.
6. **filter datasets** - specify a query using the Lucene query language to filter all selected datasets before the comparison. See Table 3.1 for supported fields and example queries.
7. **open help dialog** - opens a help dialog that explains compare options.
8. **set minimum count** - specify the minimum count to filter out spurious categories. If any of the selected dataset has less than the specified count for a certain category that category is excluded from the result table. Default is 0. If the chi-square test is selected, it is automatically set to 5.
9. **select compare option** - select compare option from drop down. See next paragraph for detailed description of available compare options.
10. **update results** - click the update button to reflect changes of the dataset selection, filter query, minimum count or compare option.

11. **change compare data type** - change the annotation type by clicking on the respective tab label. Some datasets my not have all annotation data types which is indicated by inactive tabs.
12. **change compare level** - specify the level at which you like to compare the selected datasets (e.g. rank for in the Taxonomy Tab, distance from root node for the Gene Ontology Tab, etc.).
13. **flip axis** - flip the axis of the result table (the column heading will be replaced by category names). Click again to reset the orientation.
14. **zoom** - click on zoom in/out to enlarge/shrink the font size of the result table.
15. **download** - download results shown in the table in tab delimited format.
16. **sort compare results** - change the sorting order of compare results by clicking on the column header.

Compare Options (Figure 3.13 option (9)):

- **absolute counts** - shows absolute number of entries that fall into a certain category.
- **relative counts** - shows absolute number of peptides that fall into a certain category divided by the total number of entries. If a filter query has been applied (Figure 3.13 option (6)) it shows the absolute number of entries that match the filter query and fall into a certain category divided by the number of total entries found for the filter query.
- **heatmap counts** - shows the relative counts per dataset divided by the sum of all relative counts per row. Table cells are colored according to their relative row-wise count according to the color legend shown at the top. The heatmap color scheme can be changed using the heatmap color drop down.
- **chi-square test of independence** - tests the association between two categorical variables, here the variable *category* (e.g. a certain taxon, enzyme, or pathway) and *dataset*. The null hypothesis states that the two variables are independent. The alternative hypothesis states that they are not independent. Here, for each category a 2 (rows) x datasets (columns) contingency table is constructed with the first row containing the counts for entries per dataset that fall into the respective category and the second the counts for entries per dataset that do not fall into that category. The approximation to the chi-square distribution does not hold well if the counts are too low (usually < 5). To avoid this, a minimum count cut off of 5 (Figure 3.13 option 8) is applied before generating the contingency table. The result table is sorted by ascending p-value. Often, significance levels α equal to 0.01 or 0.05 are chosen. But any value between 0 and 1 can be used depending on your Type I error threshold (the probability of falsely rejecting the null hypothesis, i.e. wrongly stating that the category and the samples are not independent). As multiple categories are simultaneously tested, bonferroni corrected p-values are listed as well (and are recommended to be used instead of the p-values). You can state that with (α there is evidence in the data that the respective category is not independent from the datasets selected. Note that this suggests that the category and the datasets are related but it may not be a causal relationship.
- **wilcoxon rank sum test** - option performs multiple two sample non-parametric Wilcoxon rank sum tests (also known as *Mann-Whitney test*). Each category (e.g. a certain taxon, enzyme, or pathway) is compared across *two* populations to see whether differences in the normalized population means for that category are due to chance or not. The null hypothesis states that there is no difference between the normalized population medians. The alternative hypothesis states that there is a significant difference. Often, significance levels α equal to 0.01 or 0.05 are chosen. But any value between 0 and 1 can be used depending on your Type I error threshold (the probability of falsely rejecting the null hypothesis, i.e. wrongly stating that there is a difference in the normalized population medians). As multiple categories are simultaneously tested, bonferroni corrected p-values are listed as well (and are recommended to be used instead of the p-values). You can state that for a certain category with α there is evidence in the data that there is a significant difference between the normalized population medians.

  Column Descriptions (**??**):

  1. category
  2. %median (population 1)

3. %median absolute deviation (MAD) (population 1)
4. %median (population 2)
5. %median absolute deviation (MAD) (population 2)
6. median ratio (median population 1/median population 2)
7. p-value
8. bonferroni corrected p-value

- **METASTATS - a modified non parametric t-test** for detecting differentially abundant features in metagenomics samples (White *et al.*, 2009). The test can be used to compare a category (e.g. a certain taxon, enzyme, or pathway) across *two* populations to see whether differences in the normalized population means for that category are due to chance or not. The null hypothesis states that there is no difference between the normalized population means. The alternative hypothesis states that there is a significant difference. The null distribution is modeled by randomization and a t-statistic is being applied (for low counts < 8 a Fisher's exact test is used instead of the non-parametric t-test). The METASTATS result table is sorted by default by ascending p-value (Figure 3.14). Often, significance levels α equal to 0.01 or 0.05 are chosen. But any value between 0 and 1 can be used depending on your Type I error threshold (the probability of falsely rejecting the null hypothesis, i.e. wrongly stating that there is a difference in the normalized population means). As multiple categories are simultaneously tested, bonferroni corrected p-values are listed as well (and are recommended to be used instead of the p-values). You can state that for a certain category with α there is evidence in the data that there is a significant difference between the normalized population means.

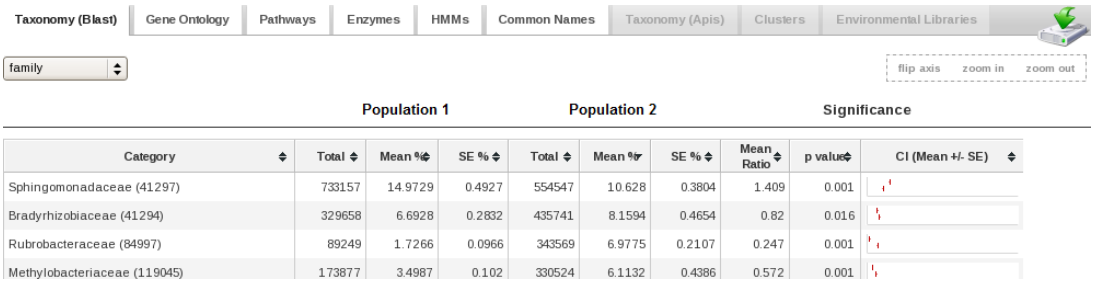


Figure 3.14: METASTATS Result Panel

Column Descriptions (Figure 3.14):

1. category
2. total (population 1)
3. %mean (population 1)
4. %standard error (population 1)
5. total (population 2)
6. %mean (population 2)
7. %standard error (population 2)
8. mean ratio (mean population 1/mean population 2)
9. p-value
10. bonferroni corrected p-value (not shown in Figure 3.14)
11. graphical representation of two confidence intervals (CIs). Each CI is calculated as follows: %mean $\pm$ 1x standard error. The CI of the first population is shown above the CI of the second population.

- **Hierarchical Cluster Plots** hierarchical cluster analysis is based on a dissimilarity matrix of samples based on differences in their category counts. Figure 3.15 shows 11 samples that were clustered based on taxonomic assignments on the family level. The matrix is generated by calculating the euclidean distances among samples using normalized category counts (count divided by the number of overall counts per sample). If a filter query has been applied (Figure 3.13) option 6), hits for a certain category are divided by the overall hits per sample. Click on the download button to download the euclidean dataset distances between. After euclidean distances have been computed, datasets are clustered. Initially, each dataset belongs to its own cluster. During each iteration of the algorithm, the two most similar clusters are joined, until there is single cluster. After each iteration, distances among clusters are recomputed for the next iteration using various clustering methods. According to Milligan *et al.*, the method with the best overall performance has been either average linkage or Ward's minimum variance method (Milligan, 1980)

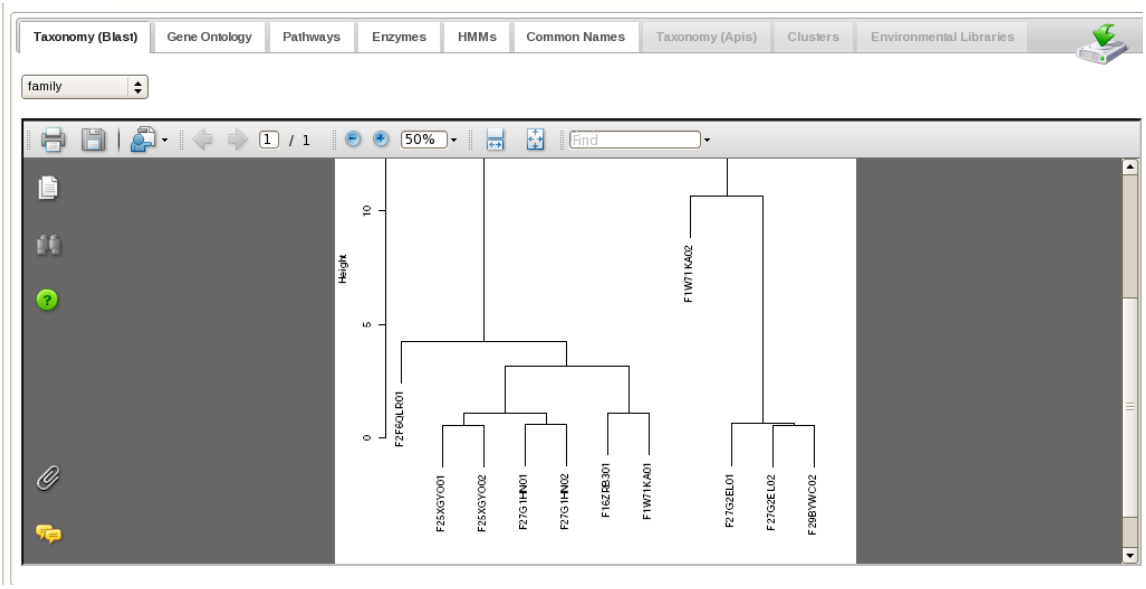


Figure 3.15: Hierarchical Clustering Plot

Clustering Methods:

  - **Complete Linkage Cluster Plot** uses the maximum distance of two merged clusters for next clustering iteration.
  - **Average Linkage Cluster Plot** uses average distance of two merged clusters for the next clustering iteration *(tends to find spherical clusters with equal variance)*
  - **Single Linkage Cluster Plot** uses minimum distance of two merged clusters for the next clustering iteration (*tends to have less bias for detecting highly elongated or irregular shaped clusters*).
  - **Ward's Minimum Variance Cluster Plot** *tends to find spherical clusters with approximately the same number of observations in each* cluster.
  - **Median Cluster Plot** *uses median distance of two merged clusters for next clustering iteration.*
  - **McQuitty Cluster Plot**
  - **Centroid Cluster Plot**

- **Multidimensional Scaling Plot** applies non-metric multidimensional scaling to project differences between samples onto a two dimensional space where samples that are close are more similar than those that a farther apart. Like for hierarchical clustering, a dissimilarity matrix is used as input for the algorithm. The matrix is generated by calculating the euclidean distances among all normalized counts (matrix cell count divided by the number of overall hits). If a filter query has been applied (Figure 3.13 option (6)), hits for a certain matrix cell are divided by the overall hits. Click on the download button to download the euclidean dataset distances.

- **Heatmap Plot** This heatmap plot is helpful to get a quick visual impression of differences between datasets and categories Figure 3.16 . Differences are highlighted by a color gradient and dendrograms (tree like structures) that are added to the left and to the top axis (Figure 3.16). The heatmap is generated by using normalized category counts (count divided by the number of overall counts per dataset). If a filter query has been applied (Figure 3.13) option 6), hits for a certain category are divided by the overall hits per dataset. Next, two sets of euclidean distances are computed. One set for columns (datasets) differences and one set for row (category) differences. Next, columns and rows are reordered to optimize the two dendograms drawings on each axis. Click on the download button to download both sets of euclidean distances.

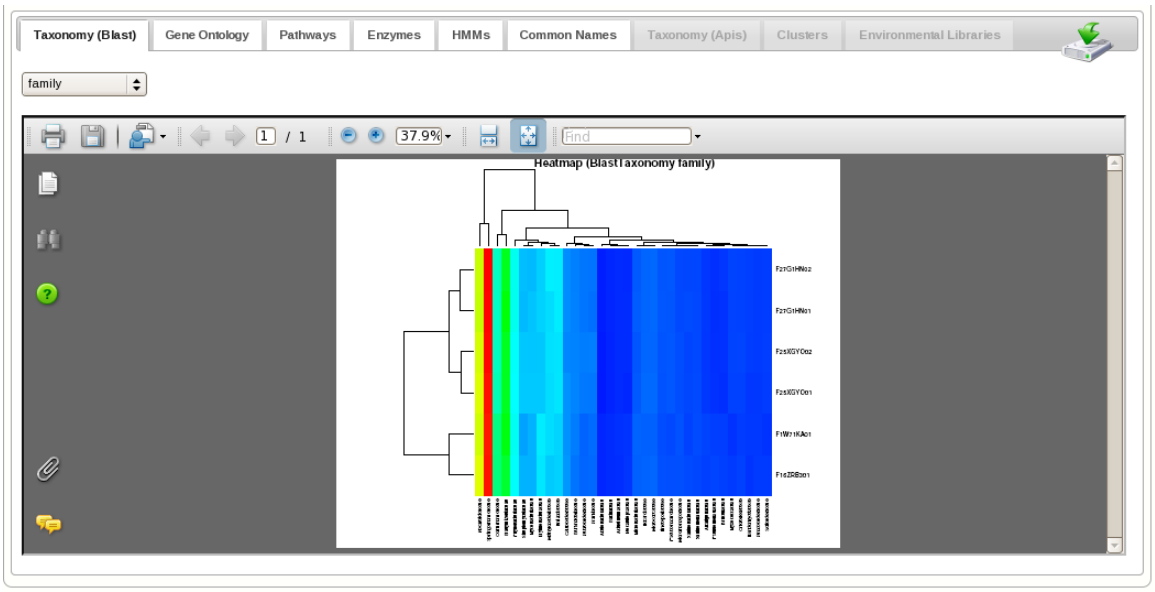


Figure 3.16: Heatmap Plot

3.5 Example Workflow

Project Description To investigate how land-use and agricultural management practices impact microbial soil communities, soil from two distinct sites have been investigated using metagenomics approaches. Both soils are calcarosols (high pH, low C and N). One site has been continually farmed (MANAGED) while a second site close by is currently under native vegetation (REMNANT). Within each site, soil samples have been collected from each location from which whole genome shotgun sequencing was carried out using Titanium 454 pyrosequencing producing 20 half plate runs for both the REMNANT and MANAGED sites. In total approximately 20 million reads were generated. Below are examples of two biological questions that can be explored from this data.

In the first example, a subset of signal transduction genes will be studied and in the second, genes involved in sphingolipid metabolism will be studied.

Workflow

Step 1. Annotate Reads Annotate reads using a metagenomics annotation pipeline (see chapter 2). For the 20 millions reads, 16 million peptide annotation have been identified.

Step 2. Create Project Create a METAREP project via the METAREP web interface (Figure 3.4).

Step 3. Load Annotations Load annotation results for each of the 454 half plates into METAREP using the loading script that is part of the METAREP installation (see chapter 2).

Step 4. Create Populations To compare the two sites, create two library populations. One based on the 20 REMNANT libraries another one based on the 20 MANAGED libraries each comprising 8 million annotation entries (Figure 3.6 option 6).

Step 5. Analysis

Case Study 1. C-di-GMP signaling EAL and GGDEF domain proteins are signal transduction proteins ubiquitous in Bacteria (Taxon ID 2) and play roles in hydrolysis of the novel second messenger cyclic dimeric GMP (c-di-GMP). The GGDEF domain protein is thought to stimulate c-di-GMP production while the EAL domain protein promotes c-di-GMP degradation (Seshasayee *et al.*, 2010). From each site, MANAGED and REMNANT, the researcher would like to identify all EAL and GGDEF domain proteins, proteins that contain both domains as well as to compare their taxonomic distribution at the genus level.

- **identify EAL and GGDEF domain related accessions and descriptions** including GO IDs, common names and HMM IDs by querying public databases such as UniProt, Pfam, GO and TIGRFAM. Here, PF00563 corresponds to the EAL domain and PF00990 or TIGR00254 correspond to the GGDEF domain.

- **search for EAL domain proteins** for each population using the populations's search page (Figure 3.11 option 2)
  query 1 (less specific): [com_name_txt:EAL AND com_name_txt:domain AND blast_tree:2]
  query 2 (more specific): [hmm_id:PF00563 AND blast_tree:2]

- **search for GGDEF domain proteins** for each population using the populations's search page (Figure 3.11 option 2)
  query 1 (less specific) [com_name_txt:GGDEF AND com_name_txt:domain AND blast_tree:2]
  query 2 (more specific) [hmm_id:PF00990 OR hmm_id:TIGR00254 AND blast_tree:2]

©J. Craig Venter Institute

- **search for proteins that contain both the GGDEF and EAL domain** for each population using the populations's search page (Figure 3.11 option 2)
  query 1 (less specific) [com_name_txt:GGDEF AND com_name_txt:EAL com_name_txt:domain AND blast_tree:2]
  query 2 (more specific) [(hmm_id:PF00990 OR hmm_id:TIGR00254) AND hmm_id:PF00563 AND blast_tree:2]

- **compare the taxonomic distribution of GGDEF domain proteins on the genus level** across the two populations using the compare page (Figure 3.13).
  1. select the two population from the multi-select box (option 2).
  2. enter filter query [hmm_id:PF00990 OR hmm_id:TIGR00254 AND blast_tree:2] (option 6).
  3. select relative counts from the compare option drop down. Alternatively, statistical tests such as METASTATS or Wilcoxon Rank Sum test can be selected to identify significant differences in the relative GGFDEF counts between the two sites (option 9).
  4. click update (option 10).
  5. select level genus in the taxonomy tab (option 12).
  6. export relative counts (option 15).
- **compare the taxonomic distribution of EAL domain proteins on the genus level** across the two populations using the compare page (Figure 3.13).
  - repeat steps 1-6 above with filter query [hmm_id:PF00563 AND blast_tree:2] (option 6).
- **compare the relative abundance of GGDEF and EAL domain proteins** across the two populations using the compare page (Figure 3.13).
  1. select the two population from the multi-select box (option 2).
  2. enter filter query [hmm_id:PF00990 OR hmm_id:PF00563 AND blast_tree:2] (option 6).
  3. select relative counts from the compare option drop down. Alternatively, heatmap counts can be selected to visualize sample differences (option 9).
  4. click update (option 10).
  5. select the HMM tab in the result tabbed panel (option 11).
  6. select level 'Pfam' (option 12).
  7. export relative counts (option 15).

Case Study 2. Sphingolipid metabolism Sphingolipids comprise a variety of lipids in which fatty acids are linked via amide bonds to a long-chain base or sphingoid. While they constitute critical components of mammalian cell membranes, they are not present in all lineages of bacteria and fungi, Therefore, the presence of genes linked to sphingolipid biosynthesis in a microbial community would constitute interesting biomarkers of a particular biological function (Olsen and Jantzen, 2001). Based on a previous study of taxonomic diversity of 16S rRNA gene sequences from these sites, it is expected that a significant proportion of community membership will be comprised of alpha-proteobacteria that are known to synthesize these compounds. Therefore, sphingolipid biosynthesis pathways may be present in the metagenomic data. One approach to examining this question is to first identify at each site the presence of genes involved in sphingolipid biosynthesis using the appropriate KEGG pathway map. Next, one can compare the relative abundance of these genes at each site and finally their taxonomic affiliations. Of particular importance in this investigation is to determine the presence, relative abundance and taxonomic affiliations of serine palmitoyltransferase(SPT) which catalyzes the first step in the *de novo* biosynthesis of sphingolipids.

- **compare relative abundance of sphingolipid metabolism enzymes** for each population using the compare page (Figure 3.13)
  1. select the two population from the multi-select box (option 2).

2. select relative counts from the compare option drop down. Alternatively, a statistical test can be selected to identify significant differences in the relative sphingolipid metabolism counts of the two sites (option 9).
3. click update (option 10).
4. select the pathway tabs in the result tabbed panel (option 11).
5. select 'Metabolic Pathways (level 3)' (option 12).
6. export results (option 15).
7. open exported results in text editor or Excel and search for 'sphingolipid metabolism'.

- **analyze enzyme distribution and top ten species of the sphingolipid metabolism** for each population using the browse pathway page (Figure 3.10).
  - open browse page for MANAGED population (Figure 3.6 option 1)
  - expand node 'Lipid Metabolism (level 2)'
  - expand node 'Sphingolipid Metabolism (level 3)'. The Sphingolipid metabolic pathway is shown on the right with found enzymes highlighted in red. Check if *Serine C-palmitoyltransferase (2.3.1.50)* is highlighted in red.
  - click on download icon to export individual enzyme counts (option 3)
  - click on download icon to export top ten species summaries shown below the pathway distribution
  - open browse page for REMNANT population and repeat the previous steps to download enzyme counts and top ten species for the REMNANT site.

©J. Craig Venter Institute

Bibliography

Li, W. (2009). Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics*, **10**, 359. 11

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., and Edwards, R. A. (2008). The metagenomics rast server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386. 11

Milligan, G. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, **45**(3), 325–342. 29

Noguchi, H., Park, J., and Takagi, T. (2006). Metagene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res*, **34**(19), 5623–5630. 12

Noguchi, H., Taniguchi, T., and Itoh, T. (2008). Metageneannotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res*, **15**(6), 387–396. 12

Olsen, I. and Jantzen, E. (2001). Sphingolipids in bacteria and fungi. *Anaerobe*, **7**(2), 103 – 112. 32

Seshasayee, A. S. N., Fraser, G. M., and Luscombe, N. M. (2010). Comparative genomics of cyclic-di-gmp signalling in bacteria: post-translational regulation and catalytic activity. *Nucleic Acids Res*. 31

Tanenbaum, D. M., Goll, J., Murphy, S., Kumar, P., Zafar, N., Thiagarajan, M., Madupu, R., Davidsen, T., Kagan, L., Kravitz, S., Rusch, D. B., and Yoosep, S. (2010). The jcvi standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data. *Stand. Genomic Sci.*, **2**, 2. 11, 13

White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol*, **5**(4), e1000352. 1, 28

Index