December | 2008

RAST and MG-RAST Workshop II

Agenda

Thursday, December 4th
8:30 ~ Coffee
9:00 ~ Welcome
9:10 ~ The subsystems based approach to annotation
9:30 ~ Overview of RAST
9:45 ~ What you can submit and what you get back
10:00 ~ Break
10:15 ~ RAST interface and tools /Hands-on tutorial
12:00 ~ Breakout sessions
1:00 ~ Metabolic Reconstructions
2:00 ~ Break
2:15 ~ Metagenomics RAST Overview
2:30 ~ Formats and results provided
2:40 ~ MG-RAST Interface and Tools /Hands-on tutorial
4:30 ~ Adjourn

Friday, December 5th
9am-2pm, Open working groups

Table of Contents

Reflections on Accurate Annotations: The Basic Cycle and Its Significance

by Ross Overbeek

I have recently been reflecting on the status of the Project to Annotate 1000 Genomes, and in this short essay I will argue that it has been an overwhelming success due to issues that became apparent only as the project progressed. A thousand more-or-less complete genomes now exist, a framework for rapidly annotating new genomes with remarkable accuracy is now functioning, and we are on the verge of another major shift in the world of annotations. This reflection is based on an informal note that I sent to friends on the last day of 2007, but my thoughts have clarified somewhat since then.

The Production of Accurate Annotations

The efforts required to establish a framework for high-volume, accurate annotation are substantial. I believe that it is important that we reflect on what we have learned about the factors that determine productivity. So, what have we learned from the project?

First, subsystem-based annotation is the key to accuracy. While there are certainly numerous efforts still focusing on annotation of a single genome, the recognition that comparative analysis is the key to everything, and that focusing on the variations of a single component of cellular machinery as they are manifested over the entire collection of existing genomes is the key to accuracy, are both widely accepted principles at this stage. Manually-based subsystem creation and maintenance is the rate-limiting component of successful annotation efforts, and the factors that constrain this process are at the heart of the matter. We have understood this for some time now.

However, I am going to argue a new position in this short essay:

1. There are three distinct components that make up our strategy for rapid accurate annotation: subsystems-based annotation, FIGfams as a framework for propagating the subsystems annotations, and RAST as a technology for using FIGfams and subsystems to consistently propagate annotations to newly-sequenced genomes.

2. These three components form a cycle (subsystems => FIGfams => RAST technology => subsystems). This cycle creates a feedback that rapidly accelerates the productivity achievable in all three components. Further, failure in any of these components impairs productivity dramatically in the others. Understanding this cycle will be the key to supporting higher productivity in subsystem maintenance and creation.

3. To understand the dependencies, we need to consider each of the components:

* The key to accurate FIGfam creation and maintenance is to couple it directly to subsystem maintenance. Once the initial release of the FIGfams was created, updating them occurs automatically based on changes in the subsystem collection. Thus, FIGfams are automatically split, merged and added as the subsystem collection is maintained. There remains one area of substantial cost in FIGfam development -- creation of family-dependent decision procedures that are occasionally required to achieve the required accuracy. At this point we have approximately 10,000 subsystem-based FIGfams, although the overall collection contains over 100,000 families (the majority containing only 2-3 members).

* RAST has a central dependency on FIGfams for assertion of function to newly-recognized genes. In this sense, the main dependency of RAST is on the FIGfam collection. The more accurate the FIGfams and their associated decision procedures, the more accurate the assignments of function made to genes in genomes processed by RAST.

* Finally, the central costs of maintenance of subsystems are cleaning up errors in existing subsystems (often indicated by multiple genes having the same function) and by adding new genomes to existing subsystems. Once a subsystem has reached an acceptable level of accuracy (and many are not there yet), the central cost is integration of new genomes after annotation by RAST. The speed with which new genomes can be added depends on how well RAST assigns gene function (and, secondarily, on how accurately these RAST-based annotations can be used to infer operational variants of subsystems).

4. The main costs of increasing the speed and accuracy of annotations split into two categories: those relating to maintenance of existing subsystems, and those relating to generation of new subsystems. The maintenance costs are containable, if the cycle is established and functions smoothly. Otherwise, I suspect they inevitably grow rapidly.

Let me begin by depicting the cycle:
I have argued that the costs in achieving rapid, accurate annotations is limited by the rate at which subsystems can be maintained and created. I place the maintenance ahead of creation at this stage. As the collection grows (it now contains over 600 subsystems with over 6800 distinct functional roles), costs of maintenance will tend to dominate. The creation of new subsystems will always be a critical activity, but each new subsystem will impact smaller sets of genomes as we "move into the tail of the distribution".

The costs relating to subsystem maintenance, which will quickly dominate, depend critically on how smoothly the cycle I described functions. We have just established the complete cycle.

The two central costs that cannot be avoided will be creation of FIGfam-dependent decision procedures and the creation of new subsystems. The manual work on FIGfams will be necessary to achieve near-100% accuracy on annotation of seriously ambiguous paralogs. However, in the vast majority of cases, this effort will be restricted to specific curators who are willing to spend massive effort to get things perfect. The more central cost relates to manual curation of the subsystems.

More Effective Integration of Existing Annotation Efforts

In the section above, I reflected on the cycle that we shall depend upon for supporting increased volume and accuracy of our own efforts. Other groups are certainly experimenting with their own solutions, and in some cases with clear successes. I have no desire to rate these competing efforts. I sincerely believe that cooperative activity is the key to enhanced achievements by everyone. However, effective cooperation is often elusive. I think that we have put in place an extremely important mechanism for making cooperation much easier, and the benefits more compelling.

Anyone working for one of the main annotation efforts realizes that it is not easy to really benefit from access to the annotation efforts of other groups. The efforts required characterizing discrepancies between local annotations and those produced externally often outweigh any benefits that result.

Two events of major importance have occurred:

1. Both PIR and the SEED Project decided to build correspondences between IDs used by different annotation projects. The PIR effort produced BioThesaurus and the SEED effort produced the Annotation Clearing House. The fact that it will become trivial to reconcile IDs between the different annotation efforts will undoubtedly support rapid increases in cross-linking entries. The SEED is working with UniProt to cross-link proteins from all of our complete genomes, and I am sure similar efforts are happening between the other major annotation efforts.
2. Within the Annotation Clearing House, a project to allow experts to assert that specific annotations are reliable (using whatever IDs they wish) has been initiated. This has led to many tens of thousands of assertions that specific annotations are highly reliable. PIR is preparing a list of assertions that they consider highly reliable, and both institutions are making these lists openly available.

To see the utility of exchanging expert assertions in a framework in which it is easy to compare the results, let me describe how we intend to use these assertions:

1. We begin with a 3-column table of reliable annotations containing [ProteinID,AssertedFunction,IDofExpert]
2. We then take our IDs and construct a 2-column table [FIG-function,AssertedFunction]. This table gives a correspondence between each of our functional roles and the functional roles used by the expert making the assertion of reliability.

3. Then, we go through this correspondence table (using both tools and manual inspection) and split it into one set in which we believe both columns are essentially identical and a second set that we believe represent errors (either our own or those of the expert asserting reliability). We anticipate that in most cases the expert assertion will be accurate, which is what makes this exercise so beneficial to ourselves.

4. We take the table of "essentially the same" assertions and distribute it as a table of synonyms (which we consider to be a very useful resource).

We are strongly motivated to resolve differences between our annotations and high-reliability assertions made by experts. The production of the table of synonyms both reduces the effort to redo such a comparison in the future, but is also a major asset by itself. I am confident that any serious annotation group that participates will benefit, and I believe that these exchanges will accelerate in 2008 and 2009.

Summary

I have tried to express the significance of the cycle depicted above, but I think that I failed to really convey the epiphany, so let me end by expressing it somewhat more emphatically. I believe that there will be a very rapid acceleration in the sequencing of new, complete genomes (although frequently the quality of the sequence wil be far from perfect, and I am willing to say that a genome in 100 contigs is "essentially complete"). Groups that now try to provide accurate integrations of all (or most) complete genomes will be strained heavily. The tendency will be to go one of two directions:

1. Some will swing to completely automated approaches. This will result in rapid propagation of errors (for those portions of the cellular mechanisms that are not yet accurately characterized -- which is quite a bit).
2. Others will give up any attempt at comprehensive annotation and focus on accurate annotation of a slowly growing subset.

The problem with the second approach is that accurate annotation of new cellular mechanisms (i.e., the introduction of new subsystems) will increasingly depend on a comprehensive set of genomes (comparative analysis is central to working out any of the serious difficulties, and the larger the set of accurately annotated genomes, the better framework for careful correction.

The cycle depicted above is the only viable strategy that I know of to handle the deluge of genomes accurately. I claim that as time goes by, the SEED effort to implement the above cycle will emerge in a continuously strengthening position. Other groups will be forced to rapidly copy it, but it really was not that easy to establish, and I believe the odds are that the SEED effort will be the only group standing in 2-3 years (i.e., it will be the only group claiming both accuracy and comprehensive integration).

How to Annotate a Genome

by Ross Overbeek

We at the Fellowship for Interpretation of Genomes (FIG) have actively led the Project to Annotate a 1000 Genomes since its inception in 2003. In that effort we pioneered what we called the subsystems approach to annotation in which experts annotated a single subsystem across the entire set of genomes. This was a radically different approach than the more usual of attempting to annotate all of the genes in a single genome. The effort to develop well-curated sets of subsystems has led to a collection of 400-600 subsystems (depending on where you choose to impose a threshold of acceptable quality). We believe that the number will continue to grow for reasons that will become apparent in this short note.

It is time to revisit the issue of how to annotate a specific genome of interest, since numerous biologists are now faced with that opportunity. For what it is worth, here is our advice.

Begin by Identifying the Recognizable Instances of Subsystems

When you are able to annotate a complete subsystem, the individual assignments are all somewhat more reliable. Most of the common machinery can easily be identified, and this establishes a starting point for the more difficult remaining tasks. The easiest way to perform this initial stage of analysis is to proceed through two tasks:

1. Submit the genome sequence to the RAST server maintained at Argonne National Laboratory. This can be done by going to the RAST server, registering yourself as a user (anyone is welcome to use the site), uploading your sequence, and getting an initial annotation back in about 12 hours. You can then download the initial annotation to your site and work on it using any tools you prefer. The initial annotation from RAST gives you three things:
    * protein-encoding genes (CDSs),
    * RNA-encoding genes (tRNAs and rRNAs)
    * identified subsystems

2. Once you have an initial set of identified subsystems, you should manually go through and see where RAST missed identifying active variants. It is fairly conservative in its calls, so if there were a mis-called gene (e.g., due to a frameshift) or an unusual form of a gene (e.g., an unknown form of an enzyme) you would see almost all of the genes in a subsystem accounted for, but not enough for RAST to say that the subsystem is really there. If you do this analysis within RAST, you can compare the metabolic reconstruction for your genome against related genomes, focusing on the specific differences.

If your genome is close to a previously annotated and studied genome, we suggest that you do a detailed analysis of what genes distinguish the new genome from the previously annotated genome (or genomes). The SEED provides a tool for easily doing such a comparison, and similar tools are either available or becoming available from a number of sources.

Note that this initial step can be done very rapidly -- in a few days.

Fix Frameshifts, Annotate Insertion Sequences, and Process Pseudo-genes

RAST often fails to identify the functional role of a particular gene due to frameshifts. This is very common in low-quality sequence or sequence produced by 454 technology. It is not particularly serious, but we do recommend that you post-process the gene calls to clean up the frameshifts. Biologists are justifiably reluctant to change sequence data without resequencing; hence, we recommend that the actual DNA sequence remain unchanged, that the correction be embodied in the proposed translation of the feature, and that the discrepancy between the actual DNA sequence and the translation be recorded with the feature. We note that you can automatically correct obvious frameshifts using tools within the SEED environment, and we anticipate that these will become increasingly important as larger volumes of low-quality sequence data becomes available.

The issue of detecting insertion sequences, mobile elements, prophages and so forth is important for a number of reasons. Determining the set of impacted genes (often pseudo-genes) is extremely time-consuming. We would guess that tools to support this type of analysis will appear soon, but for now you will need to determine how much effort you are willing to expend on the task. So, this part of the effort can take from a few days (to automatically detect and correct frameshifts) to man-years (to characterize insertion sequences, pseudo-genes, and prophages).

Look at Identified Functions that are Not in Subsystems

As you scan through the genes not yet placed in subsystems that were identified by RAST, some correspond to FIGfams, and some do not. Some are closely similar to well-annotated proteins (e.g., to Swiss Prot entries), and some are not.

We recommend that you scan through these focusing on those that correspond to functional roles that should be encoded into subsystems. It is particularly important to examine those for which "functional coupling" information exists (RAST will give you this information). When strong functional coupling data exists, and when the functional role can be identified with reasonable certainty, you have a particularly good candidate for a new subsystem. If you can connect any of the genes in the cluster (in, say, genomes that are "close" and have been actively studied) to literature, you need to get the relevant papers before deciding how to proceed. We suggest making a rapid pass through the set of genes that have not been assigned to subsystems, prioritizing these genes for possible use in starting new subsystems.

We urge you to develop new subsystems when possible and to publish these subsystems (which makes them accessible to users working on other versions of the SEED).

Summary
So, our approximate approach to annotating a new genome would be:

1. Run the genome through RAST.

2. Do a detailed metabolic comparison (within RAST) between your new genome and one or more of its closest relatives. Follow this by a general comparison of what genes distinguish it from its closest relatives.

3. Correct obvious frameshifts.

4. Decide whether or not you are willing to spend the effort needed to identify IS elements, prophages and other mobile elements. Similarly, decide whether or not you wish to expend the effort to carfully identify pseudo-genes.

5. If you have substantially changed the gene calls, rerun your genome through RAST again (keeping the gene calls that you have now established).

6. Go through the genes that have not yet been placed into subsystems, determine whether or not it makes sense to construct a limited set of new subsystems (especially if they capture aspects of the genome which may have motivated the sequencing effort in the first place).

This can be done either very rapidly or more time can be taken. It all depends on the anticipated role of the genome. In many cases, these tasks can be performed in a few weeks, and we believe that the overall time will continue to drop as the quality of the RAST analysis (due to an expanded library of subsystems) improves.

NMPDR

NMPDR is a Bioinformatics Resource Center that provides the advanced bioinformatics environment needed to identify genetic polymorphisms correlated with pathogenicity, drug resistance, morbidity, and infectivity. NMPDR is funded by the National Institute of Allergy and Infectious Disease (NIAID) to support research in biodefense, emerging infectious diseases, and re-emerging pathogens. NMPDR is both a central repository for a wide variety of scientific data on these pathogenic microorganisms and a platform for software tools that support investigator-driven data analysis, such as the identification of potential targets for the development of vaccines, therapeutics, and diagnostics.

NMPDR's goal is to support existing and newly developed techniques for comparative analysis to obtain a deeper understanding of the fundamental biology of a specific set of pathogenic organisms, and to promote efforts to counter the threats posed by these pathogens. The resource will integrate hundreds (and, eventually, thousands) of prokaryotic genomes with existing and newly generated functional data to support characterization and analysis of Category B food and water-borne diarrheagenic bacteria, *Campylobacter jejuni, Vibrio cholerae, Vibrio parahaemolyticus, Vibrio vulnnificus*, and *Listeria monocytogenes*. Also included are the enterotoxin B producing *Staphylococcus aureus*, as well as the re-emerging, antibiotic resistant pathogens *Streptococcus pneumoniae* and *Streptococcus pyogenes* (Group A Strep).

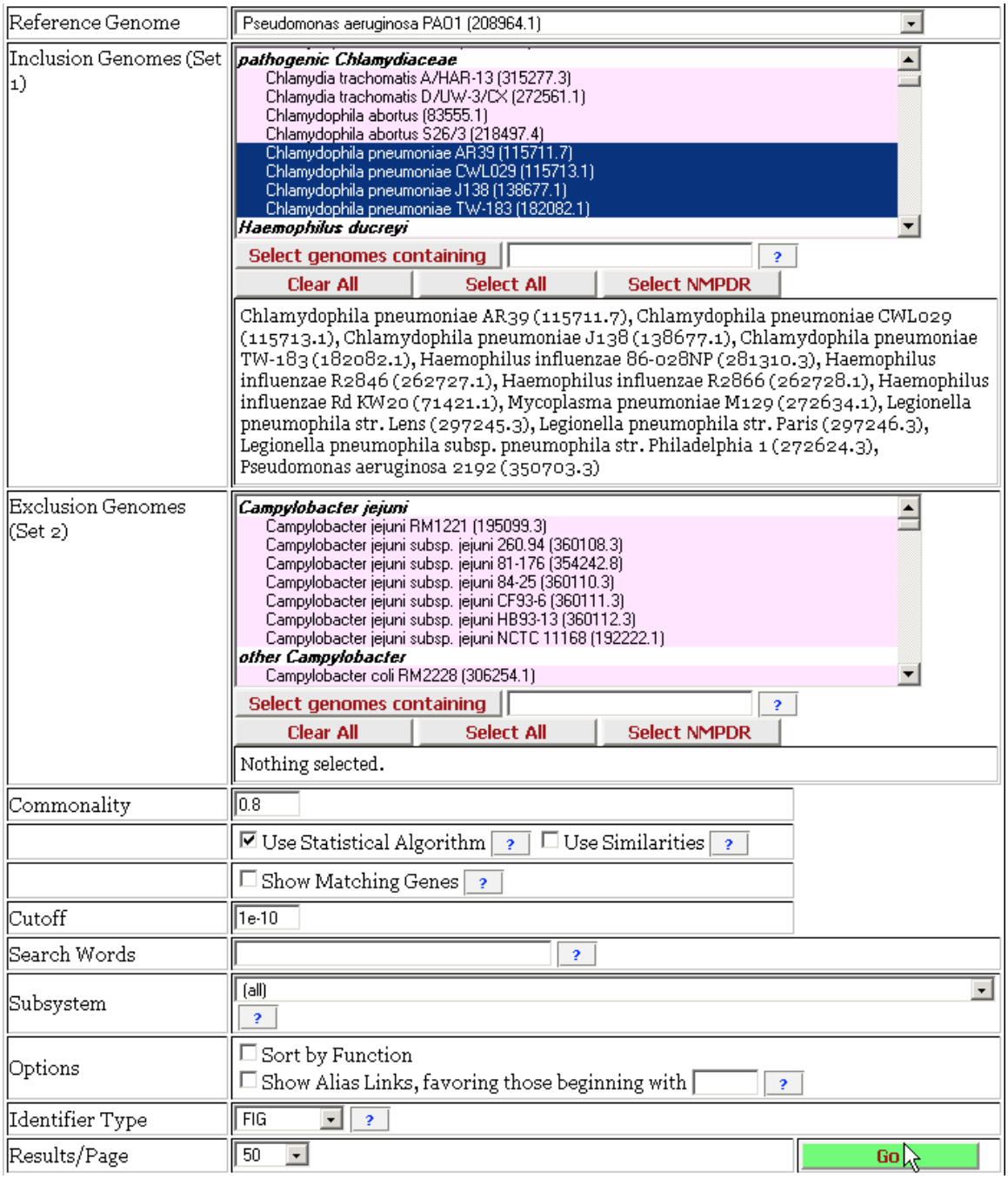
NMPDR employs a strategy of subsystems annotation to provide researchers with corrected functional annotations in a structured biological context. More than 600 distinct subsystems have been developed using the SEED by professional curators and community experts to describe central and secondary metabolism, complex structures, and virulence or other phenotypes. NMPDR includes all essentially complete, public genomes that provide a rich context for comparative analysis with tools such as Signature Genes, Homolog Spreadsheet, and Compare Regions.

Example: Use NMPDR to find genes important in respiratory pathogens

Use NMPDR's Signature Genes Tool to compare whole genomes with the goal of defining a set of protein-encoding genes that are shared among selected genomes. For example, what genes are common to Gram-negative respiratory pathogens?

1. As the reference genome, select a commonly used strain of a relevant species., For example, choose *Pseudomonas aeruginosa* strain PAO1, which is a well characterized lab strain.

2. In the inclusion set, select any number of genomes that share a phenotype with the reference genome. For example, include all available strains of *Chlamydophila pneumoniae, Haemophilus influenzae, Mycoplasma pneumoniae, Legionella pneumophila*, and *Pseudomonas aeruginosa* to determine the core set of genes shared by these strains.
3. Leave the exclusion set empty.
4. Leave the remainder of the settings at their default values and click the Go button, as illustrated below.



5. The tool searches the database to find every protein in the reference genome that has a bidirectional, best BLASTP hit (BBH) with the genomes in the inclusion set, but not the exclusion set. Results with a score of 1.000 have a bidirectional best hit in every genome from Set 1 (and no bidirectional best hit against any genome in Set 2 when that set has members). When sets 1 and/or 2 are large, proteins with less than perfect scores will be returned.
    - For example, 22 proteins in Pseudomonas aeruginosa strain PA01 have BBH with the all of the other genomes selected (which are of various quality), and another 163 have BBH in most of the Gram-negative respiratory pathogens selected. These are a starting point for finding genes that may be targets for therapeutics. Results may be saved as protein or DNA sequences, or as a tab-delimited file suitable for opening and resorting in a spreadsheet application such as Excel.
6. To refine the list of potential drug targets, examine the subsystems these candidates play roles in, and examine their closest sequence matches using the Similarities on the respective evidence pages.
7. Because we have annotated the results of a genome-wide essentiality screen for the reference genome, Pseudomonas aeruginosa strain PA01, this list may be narrowed to those genes demonstrated to be essential in the reference strain. Type "essential" in the Search Words field of the Signature Genes form and repeat the search. This reduces the number of candidates to 107.

Another avenue for investigating essential genes is via the Essential Genes page, which lists genes demonstrated to be essential (or potentially essential when there is severe growth attenuation or the possibility of polar effects) in genome-wide screens of 10 different pathogens.

A list of candidate drug targets defined as the set of proteins that have been determined to be essential in at least one of the NMPDR pathogens, have been included in subsystems by our curators, have orthologs in the Protein Data Bank, and have orthologs in a substantial number of the pathogenic bacterial species curated in the BRC system is available for browsing on the Drug Targets page.

New in NMPDR --- The Protein Target Search Tool.

The NMPDR team has developed a new discovery tool to allow users to identify and characterize proteins and genes based on various attributes and features in the NMPDR and SEED databases. This tool provides users with flexible searching (Boolean) (Figure 1 & 2) and returns a results table in which the user can add or delete information from it (Figure 3).

All results are downloadable and in our next release, will be connected to a set of tools to perform operations on sets of sequences (e.g. get FASTA sequences, multiple sequence alignment). To complement this search capability we have also included a

'batch search tool' which takes a list of IDS (NCBI, UniProt , Locus) and returns information for each protein-encoding gene in tabular form.
The Target Search search parameters include:

* Various Sequence IDs * Pfam Name or ID * Taxonomy ID * Organism Name * Lineage * Subsystem * In Conserved Neighborhood * EC Number or Function * PatScan Sequence, AA * PatScan Sequence, DNA * Transmembrane Domains * Signal Peptide * Subcellular Location * Similar to Human * Molecular Weight * Isoelectric Point * Sequence Length	* Amino Acid Content * General Organism Phenotype: o Gram Stain o GC Content o Shape o Arrangement o Endospores o Motility o Pathogenic o Salinity o Oxygen Requirements o Temperature Range /Optimum o Habitat o Pathogenic In o Pathogenic/Non-pathogenic o Disease

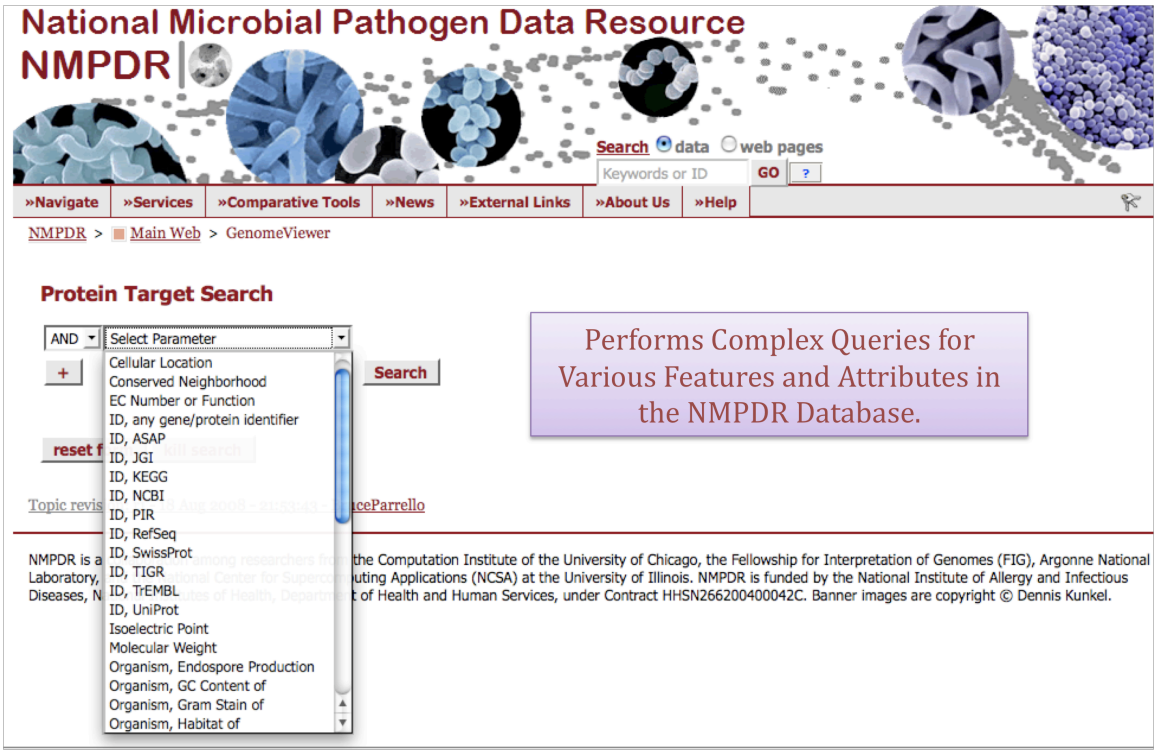


Figure 1. The NMPDR Protein Target Search Tool. Users can construct complex Boolean searches for various features and sequence and organism attributes to identify proteins of interest.

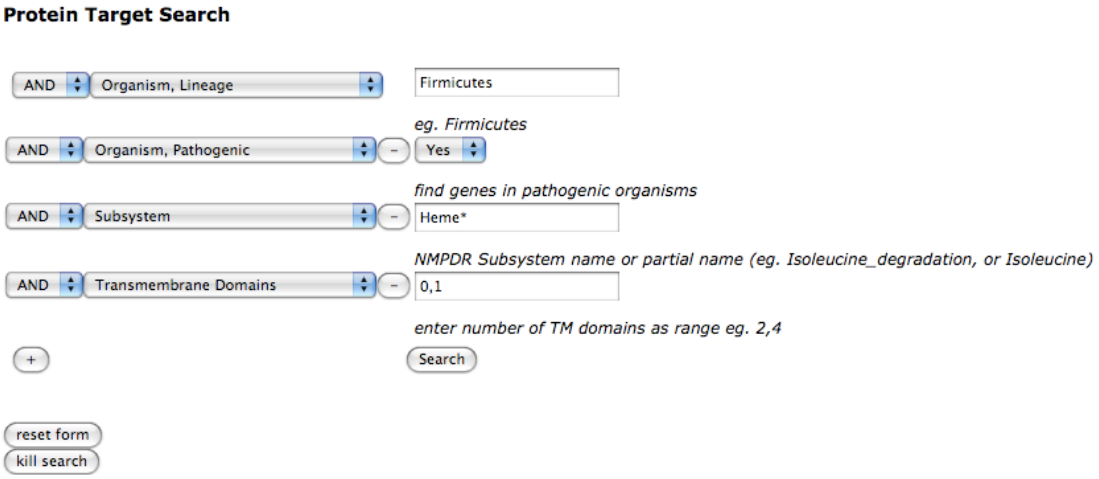
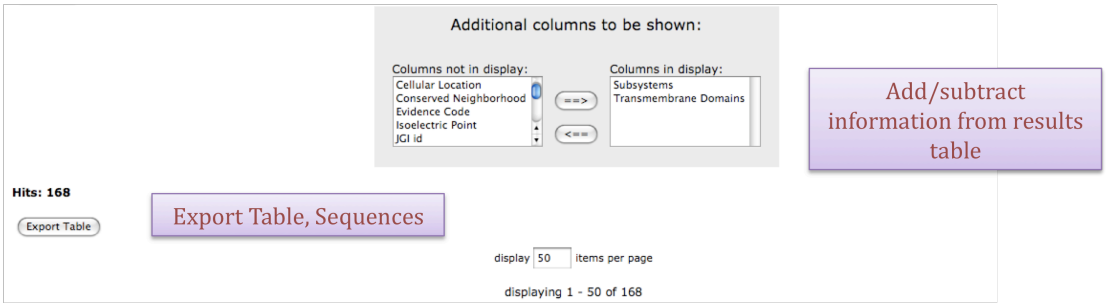


Figure 2. An example search using the Protein Target Search. Here a user has constructed a query to identify soluble proteins annotated as being involved in subsystems involved with Heme (biosynthesis, utilization, transport, etc) in Firmicutes that are pathogenic.



Parameters Hit	FIG ID	Organism	Aliases	Function	Subsystems	Transmembrane Domains
Transmembrane Domains:0,1 Subsystem:Heme* Pathogenic:Yes Lineage:'Firmicutes'	fig|401650.3.peg.2001	Listeria monocytogenes Aureli 1997		Cell surface protein IsdA, transfers heme from hemoglobin to apo-IsdC	1. Heme,_hemin_uptake_and_utilization_systems_in_GramPositives 2. Sortase	545-563
Transmembrane Domains:0,1 Subsystem:Heme* Pathogenic:Yes Lineage:'Firmicutes'	fig|267409.1.peg.1075	Listeria monocytogenes str. 1/2a F6854	LMOf6854_2249 ZP_00233365.1 gi|47095759	Cell surface protein IsdA, transfers heme from hemoglobin to apo-IsdC	1. Heme,_hemin_uptake_and_utilization_systems_in_GramPositives 2. Sortase	545-563
Transmembrane Domains:0,1 Subsystem:Heme* Pathogenic:Yes Lineage:'Firmicutes'	fig|260799.1.peg.4398	Bacillus anthracis str. Sterne	BAS4442 YP_030689.1 gi|49187437	Cell surface protein IsdA, transfers heme from hemoglobin to apo-IsdC	1. Heme,_hemin_uptake_and_utilization_systems_in_GramPositives 2. Sortase	859-877
Transmembrane Domains:0,1 Subsystem:Heme* Pathogenic:Yes Lineage:'Firmicutes'	fig|66692.3.peg.3435	Bacillus clausii KSM-K16	GeneID:3202633 LocusTag:ABC3423 YP_176917.1 gi|56965185	Cell surface protein IsdA, transfers heme from hemoglobin to apo-IsdC	1. Heme,_hemin_uptake_and_utilization_systems_in_GramPositives 2. Sortase	769-788
Transmembrane Domains:0,1 Subsystem:Heme* Pathogenic:Yes	fig|158878.1.peg.1130	Staphylococcus aureus subsp. aureus Mu50	GeneID:1121107 NP_371654.1 SAV1130 gi|15924120	Cell surface protein IsdA, transfers heme from	1. Heme,_hemin_uptake_and_utilization_systems_in_GramPositives 2. Sortase	324-344

Figure 3. An example results table for Protein Target Search Tool. Users are given results in tabular form in which they can search, sort or add additional information. All results are downloadable and in the next release, will allow for selection of proteins to be submitted to analysis tools (multiple sequence alignment, find consensus, etc).

Suggestions for Constructing Queries

1. Faster, more successful searches will likely include at least one parameter that will reduce the search space such as 'EC Number or Function', 'Subsystem', 'Organism Name', 'Taxon ID', 'Lineage', 'PFAM ID', 'PFAM Name', or any of the ID parameters with AND assigned as the logical operator.

2. The order of the parameters selected does not matter. The query is rearranged automatically for optimized performance.

3. All proteins returned by the search must match all parameters assigned the logical AND operator, at least one of the parameters assigned a logical OR operator, and none of the parameters assigned the logical NOT operator.

4. If a single parameter is assigned OR as the logical operator, it is the equivalent of an AND.

5. Including multiple OR operators can *significantly increase search time*.

6. The NOT logical operator should be used in conjunction with parameters assigned AND/OR logical operators.

7. If more than 10,000 hits are found, the search will need to be resubmitted with a refined set of parameters.

8. The following cellular locations apply only to gram negative bacteria: Periplasmic, OuterMembrane.

9. The following cellular locations apply only to gram positive bacteria: CellWall, CytoplasmicMembrane.

10. Prefix and/or append your 'EC Number or Function' parameter with a * as a wildcard. Otherwise, only exact matches will be returned.

Using the RAST prokaryotic genome annotation server

RAST is designed to rapidly call and annotate the genes of a complete or essentially complete prokaryotic genome. RAST uses a "Highest Confidence First" assignment propagation strategy based on manually curated subsystems and subsystem-based protein families that automatically guarantees a high degree of assignment consistency. RAST returns an analysis of the genes and subsystems in your genome, as supported by comparative and other forms of evidence.

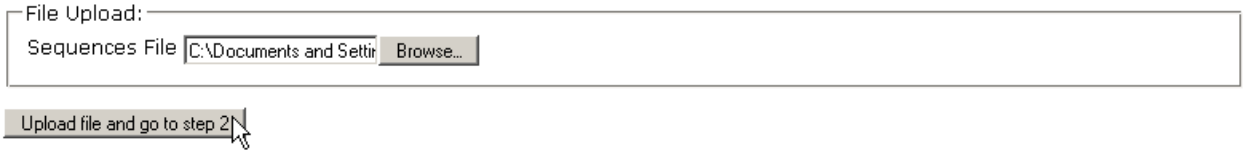
Because the SEED and NMPDR provide access to all essentially complete, public genomes without a user account, the use of RAST without an account makes no sense—you must have a free account in order for access to your data to be kept under your control. The tools available in RAST for comparing your new private data to public genomes are mostly the same as those available for analyzing public genomes at The SEED (http://seed-viewer.theseed.org/) and NMPDR (www.nmpdr.org).

The tour of the site will follow this workflow:

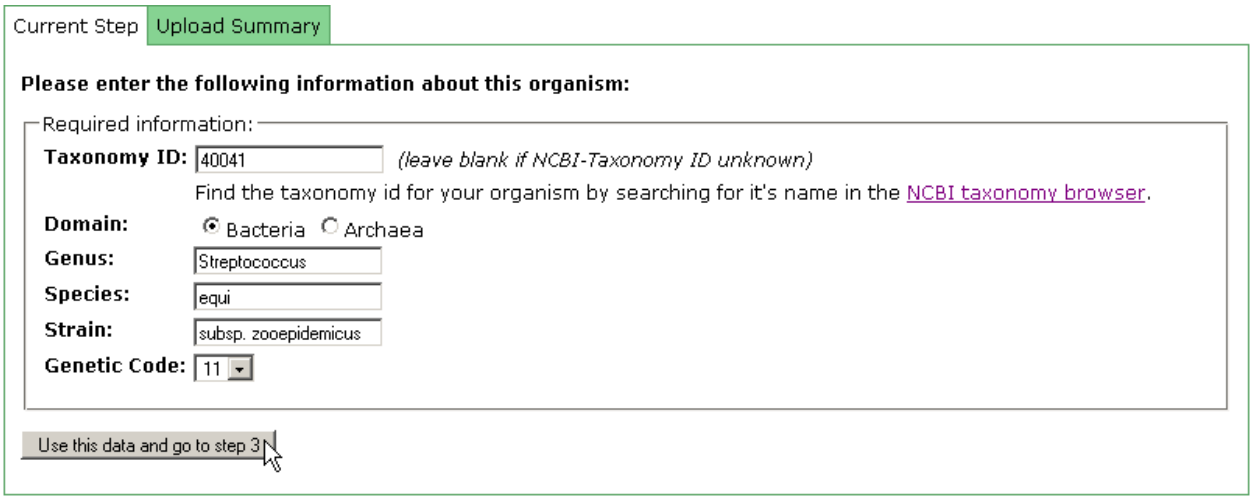
- upload data
    - file formats, tax id, translation table, check status, final upload
    - manage/share job
- view annotated genome
    - genome overview—basic stats, features in vs. not in subsystems
    - browse genome—walk chromosome, color by subsystem
    - metabolic reconstruction—expand categories, view table, select subsystem to view
    - explore subsystem—private genome highlighted
    - explore one protein page—compare regions with SEED genomes
- compare new data to another genome
    - compare proteomes (sequence-based comparison)
    - compare metabolic reconstructions (annotation-based comparison)

Upload your genome sequence to RAST

- Begin at http://rast.nmpdr.org (Or from www.nmpdr.org click on the link to the RAST server)
- Log in and select "Upload New Job" from the "Your Jobs" menu.
- **Step 1:** Browse for the sequence file which must be a plain text file in either FASTA or GenBank format only. Click the button to "Upload and go to step 2."



- **Step 2:** Provide the name of the organism. If you know or find the taxonomy ID from NCBI, paste it into the text box. Then, when you select either Bacteria or Archaea with the radio button, the corresponding genus, species, and strain will autofill accurately. If you do not know or cannot find an ID in the NCBI Taxonomy database, then fill in the genus, species, and strain. RAST will provide a dummy ID number corresponding to nothing in the taxonomy database.
- Choose a translation table. Most bacteria use version 11 of the genetic code, but mycoplasmas and spiroplasmas use version 4.

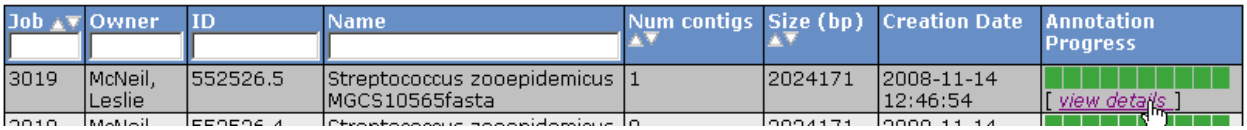


- **Step 3:** Provide information about the sequence data and select settings. If you uploaded a GenBank file, you may elect to preserve gene calls. Since there are no gene calls in a FASTA file, this choice would be unavailable.
- Select whether the translated proteins should reflect corrected frame shifts if you have low-quality sequence data.
- Look at the information in the "Upload summary" tab to confirm that the system detected the sequence data you intended to upload.

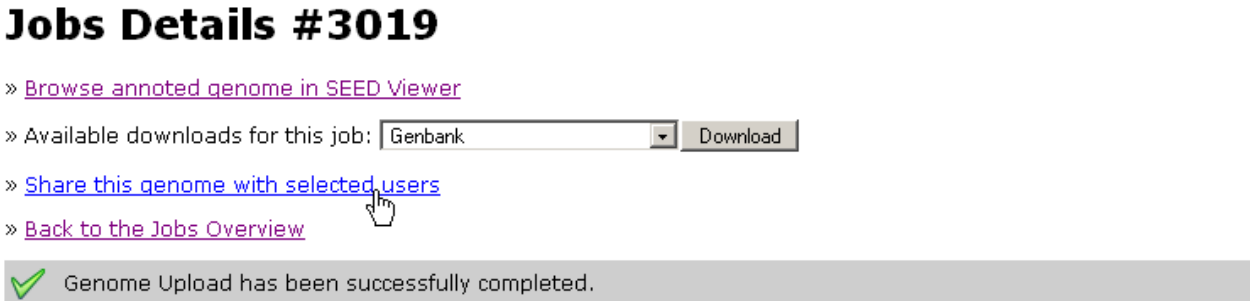
- Click the button to "Finish Upload."

Manage your job in RAST

- From the "Your Jobs" menu, select "Jobs Overview." If you have logged out, you will be directed to your jobs overview upon logging back in.
- The status and progress of the new job can be tracked in the table. Select the link to view details of one job.



- From here you may download the completed job, or you may delete it from the system using the green menu bar.



- You can share this job with others by clicking the link and adding the email addresses of those to whom you would like to grant access to your otherwise private data.

- Request a new group by emailing rast@mcs.anl.gov. Groups can be managed from the account management page, which is accessed by clicking on the pair of people at the far right of the green menu bar.
- Your annotation could be complete in as few as 8 hours.

View your annotated genome results in RAST

- Begin at http://rast.nmpdr.org (Or from www.nmpdr.org click on the link to the RAST server)

- Log in and find your new genome in your jobs overview
- Click on "view details" and then "Browse annotated genome in SEED Viewer"

Organism Overview for Streptococcus zooepidemicus MGCS10565fasta (552526.5)

Genome	Streptococcus zooepidemicus MGCS10565fasta (Taxonomy ID: 552526)
Domain	Bacteria
Size	2,024,171 bp
Number of Contigs	1
Number of Subsystems	217
Number of Coding Sequences	2014
Number of RNAs	72

For each genome we offer a wide set of information to browse,

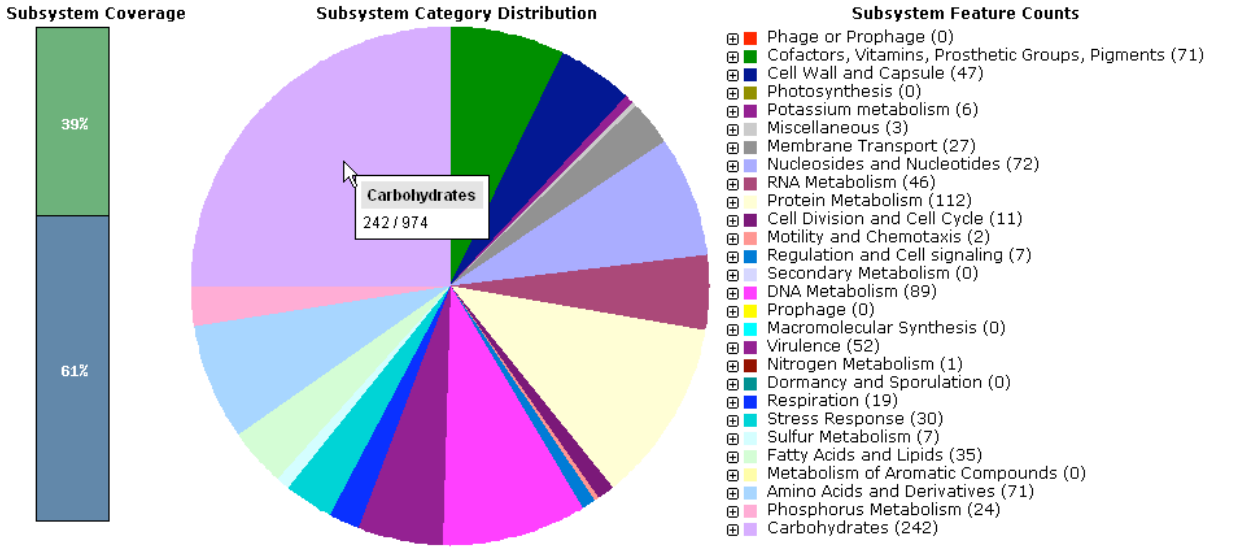
Browse | Compare | Download

Browse through the features of Streptococcus zooepidemicus MGCS10565fasta both graphically and through a table. Both allow quick navigation and filtering for features of your interest. Each feature is linked to its own detail page.

Click here to get to the Genome Browser

Subsystem Information

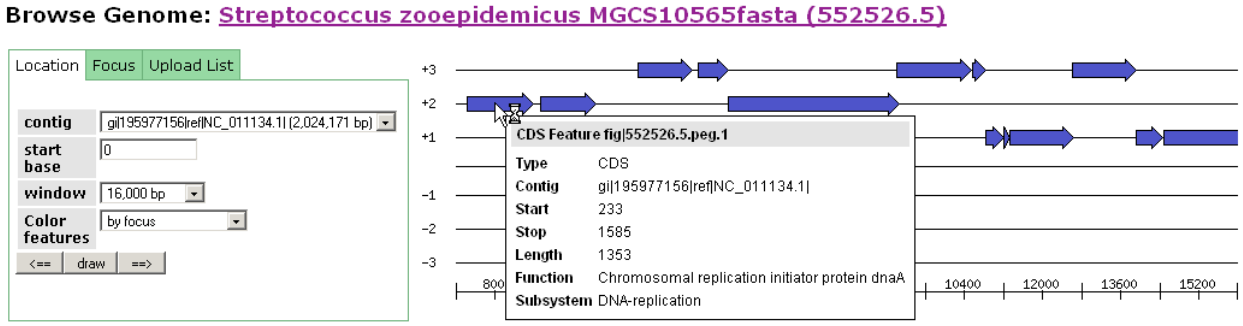
Subsystem Statistics | Features in Subsystems



- Notice that the genome summary page shows how many contigs, how many genes, the number of genes that are assigned to complete subsystems, and the subsystem categories represented.
- The green "Features in Subsystems" tab displays all genes and other features that are automatically included in subsystems because similar sequences are found for all roles in a functional variant of the subsystem. The table is resortable and downloadable.
- In the menu bar, under Organism, the feature table will display all annotated features in your genome—both those in and not in complete subsystems.

Walk the chromosome or contigs of your new genome:

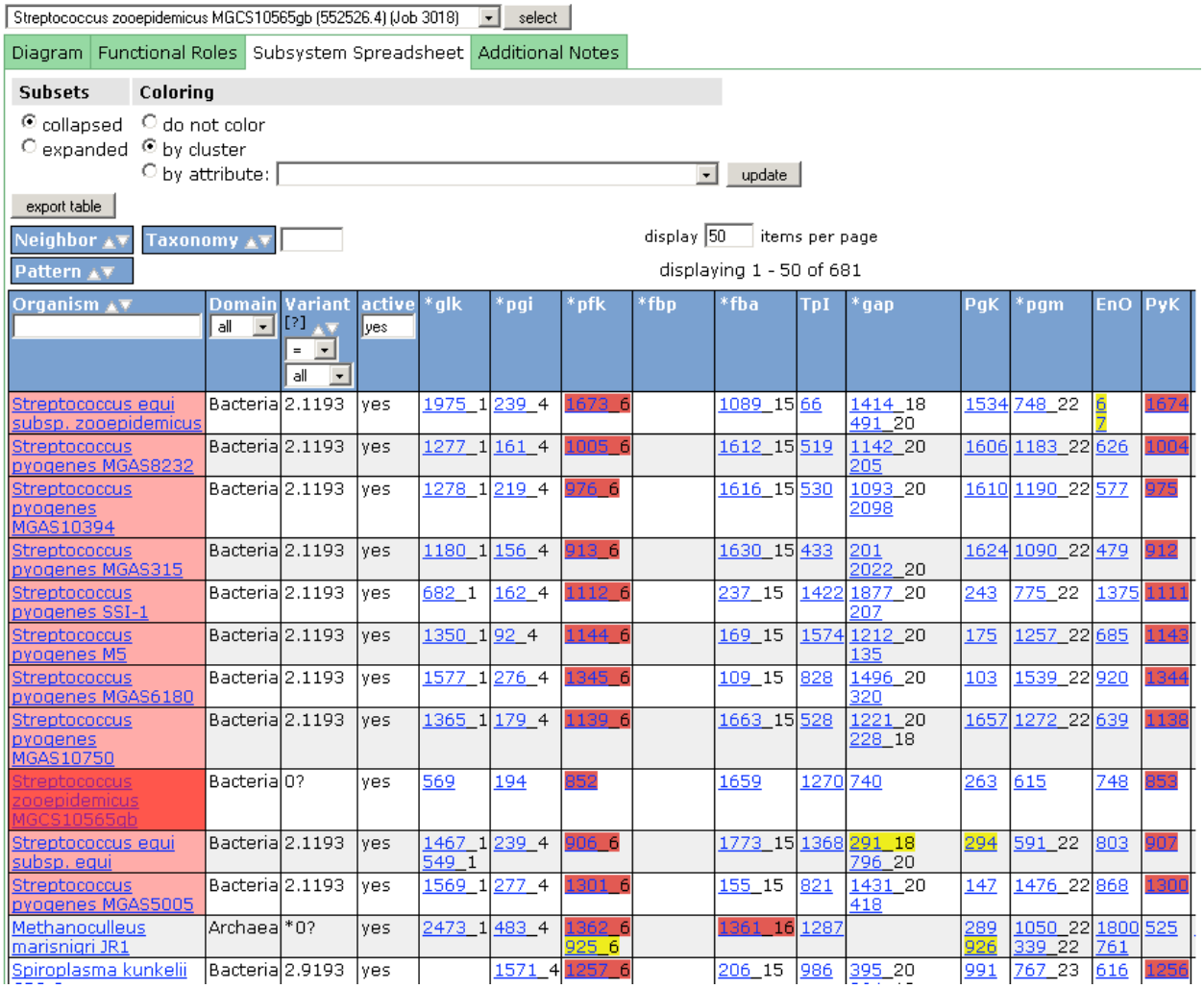
- **Browse genome** Access to the genome browser is available from the menu bar, under Organism -> Browse Genome, and in the hint box at the top of the overview page.
- The genome browser provides a visual tour of the annotated features. You may choose a larger window, and you may color the features by subsystem or clustering.



- The table beneath the graphic allows you to scan or search for a feature of interest. Click the "Region" button in the last column to focus the browser graphic on the selected feature.

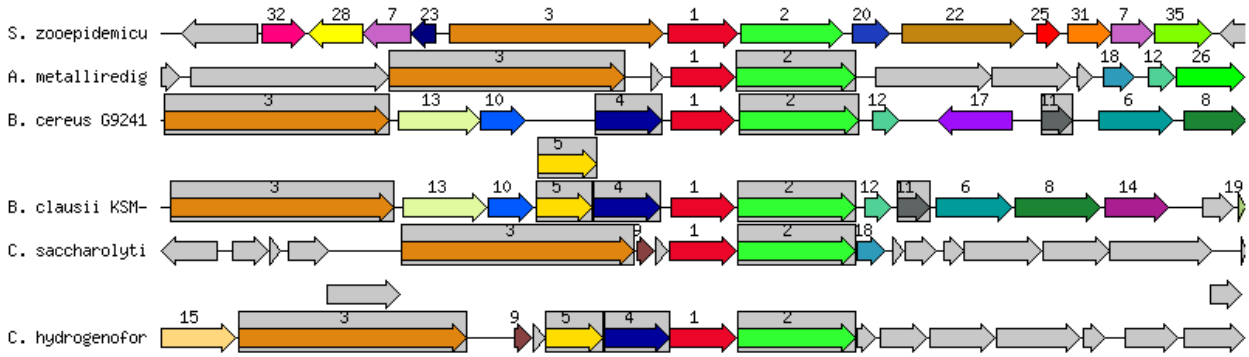
Metabolic reconstruction of your new genome--subsystems:

- Return to the genome overview page to view the pie chart of complete subsystems identified in the genome.
- Expand the categories to see subcategories and subsystem names.
- The table provides similar access; you can select the "carbohydrates" category from the column header.
- From the Carbohydrates category either in the chart or table, click to open the Glycolysis and Gluconeogenesis subsystem
- The new genome is highlighted and displayed in the context of closely related public genomes. The spreadsheet is arranged with functional roles in columns, genomes in rows, and genes annotated to those roles in the respective cells. Within one row, genes that are clustered on the genome are shown in the same color.



Organism	Domain	Variant [?]	active	*glk	*pgi	*pfk	*fbp	*fba	TpI	*gap	PgK	*pgm	EnO	PyK
	all	= all	yes											
Streptococcus equi subsp. zooepidemicus	Bacteria	2.1193	yes	1975_1	239_4	1673_6		1089_15	66	1414_18 491_20	1534	748_22	6 7	1674
Streptococcus pyogenes MGAS8232	Bacteria	2.1193	yes	1277_1	161_4	1005_6		1612_15	519	1142_20 205	1606	1183_22	626	1004
Streptococcus pyogenes MGAS10394	Bacteria	2.1193	yes	1278_1	219_4	976_6		1616_15	530	1093_20 2098	1610	1190_22	577	975
Streptococcus pyogenes MGAS315	Bacteria	2.1193	yes	1180_1	156_4	913_6		1630_15	433	201 2022_20	1624	1090_22	479	912
Streptococcus pyogenes SSI-1	Bacteria	2.1193	yes	682_1	162_4	1112_6		237_15	1422	1877_20 207	243	775_22	1375	1111
Streptococcus pyogenes M5	Bacteria	2.1193	yes	1350_1	92_4	1144_6		169_15	1574	1212_20 135	175	1257_22	685	1143
Streptococcus pyogenes MGAS6180	Bacteria	2.1193	yes	1577_1	276_4	1345_6		109_15	828	1496_20 320	103	1539_22	920	1344
Streptococcus pyogenes MGAS10750	Bacteria	2.1193	yes	1365_1	179_4	1139_6		1663_15	528	1221_20 228_18	1657	1272_22	639	1138
Streptococcus zooepidemicus MGCS10565gb	Bacteria	0?	yes	569	194	852		1659	1270	740	263	615	748	853
Streptococcus equi subsp. equi	Bacteria	2.1193	yes	1467_1 549_1	239_4	906_6		1773_15	1368	291_18 796_20	294	591_22	803	907
Streptococcus pyogenes MGAS5005	Bacteria	2.1193	yes	1569_1	277_4	1301_6		155_15	821	1431_20 418	147	1476_22	868	1300
Methanoculleus marisnigri JR1	Archaea	*0?	yes	2473_1	483_4	1362_6 925_6		1361_16	1287		289 926	1050_22 339_22	1800 761	525
Spiroplasma kunkelii	Bacteria	2.9193	yes		1571_4	1257_6		206_15	986	395_20	991	767_23	616	1256

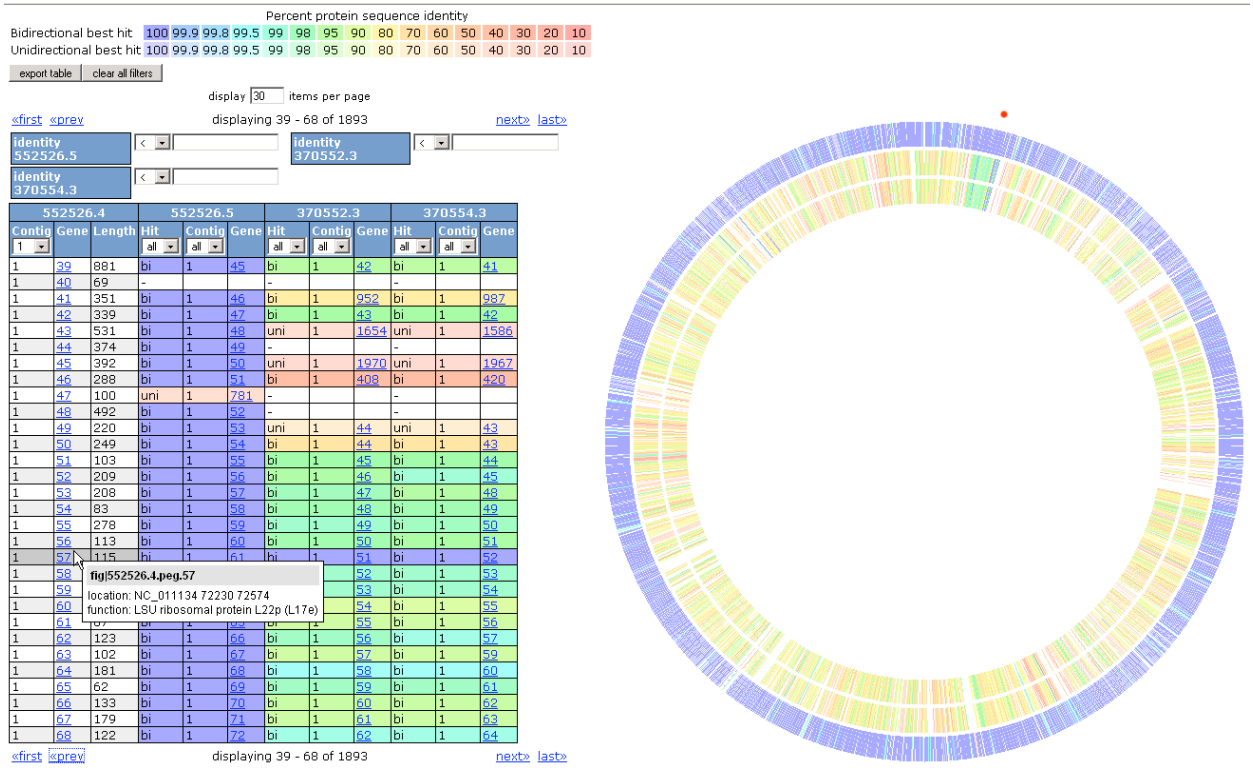
- Click on any of the genes in the newly annotated genome. The annotation overview will open in a new window or tab.
- The compare regions view displays the new genome at the top in comparison with 4 other closely related genomes.
- To expand the graphic comparison to more genomes, click the advanced button, select the option to collapse close genomes, type in a larger number of genomes (20), select PCH pin for clustered genes, or leave at similarity for isolated genes, then click the button to redraw the graphic.



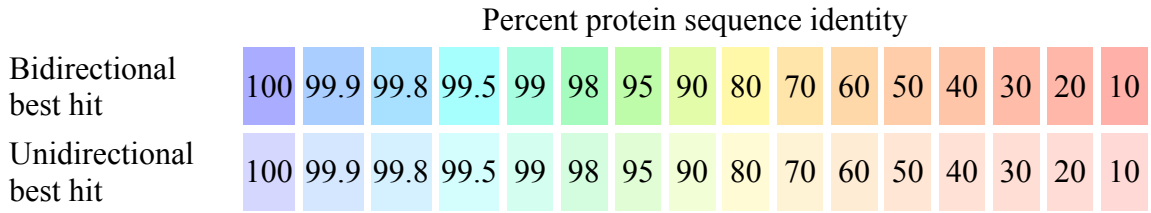
How does your genome <u>sequence</u> compare with others in the database?

- From the page showing the subsystem, go back on your browser to return to the overview, then select "Sequence-based comparison" from the Comparative Tools menu.
- **Step 1: Select a reference genome.** If your private sequence is in many contigs, the best selection may be a known, high-quality genome.
- **Step 2:** Select up to three comparison genomes.
- **Step 3**: Click button to compute

Result:



- This example uses the GenBank and FASTA versions of the same *Streptococcus equi subsp. zooepidemicus* MGCS10565 genome in order to compare the published gene calls to the RAST-computed gene calls. The second two genomes are nephritogenic strains of *Streptococcus pyogenes*.
- Results are computed using BLASTP to compare every protein in the reference genome to every protein in the comparison genomes.
- Results are presented in a color-coded table and in a circular map, in order of the contigs/genes in the reference genome.
- Comparison proteins are listed with their contig number, gene number, and length. The gene numbers are linked to pop-up boxes that list the annotation (name) as well as the proportion of identical amino acids.
- The amino acid identity of the comparison genomes relative to the reference is color-coded on a scale that is not linear, but logarithmic, and that follows the order of the visible spectrum.



How does your genome <u>content</u> compare with others in the database?

- From the Comparative Tools menu of the organism summary page, select "Function-based comparison"
- Your newly annotated genome is already input as the reference.
- Select one comparison genome.
- Click the button

Result:

- The table opens with all features in your genome (A) and the comparison genome (B) that are associated with a complete subsystem.
- The table is sortable and searchable by subsystem category or name
- The first column may be reset to show features in genome A, but not B; in both A and B; or in B but not A. When genome B is very closely related to the new genome, A, this comparison of what functions are annotated in B but not automatically associated with subsystems in A indicates a place to begin looking at the annotations to evaluate accuracy or find missing functions. It is important to keep in mind that automated subsystem assignments are only

made if all roles required for one functional variant of a subsystem can initially be recognized by sequence similarity. If only some of the roles are found, the subsystem is not declared to be present.

- The table is downloadable.

Compare Metabolic Reconstruction of Streptococcus zooepidemicus MGCS10565gb (A) and Streptococcus pyogenes MGAS10750 (B) select other

save to file
clear all filters

display 15 items per page

displaying 1 - 15 of 1361 next» last»

Presence (all / all / A / A and B / B)	Category (all)	Subcategory (all)	Subsystem	Role	Organism A	SS active A (all)	Organism B	SS active B (all)
	Amino Acids and Derivatives	Arginine; urea cycle, polyamines	Arginine and Ornithine Degradation	NADP-specific glutamate dehydrogenase (EC 1.4.1.4)	fig|552526.4.peg.1055	yes	find	no
A	Amino Acids and Derivatives	Aromatic amino acids and derivatives	Common Pathway For Synthesis of Aromatic Compounds (DAHP synthase to chorismate)	3-dehydroquinate dehydratase (EC 4.2.1.10)	fig|552526.4.peg.836	yes	find	no
A	Amino Acids and Derivatives	Aromatic amino acids and derivatives	Common Pathway For Synthesis of Aromatic Compounds (DAHP synthase to chorismate)	3-phosphoshikimate 1-carboxyvinyltransferase (EC 2.5.1.19)	fig|552526.4.peg.676	yes	find	no
A	Amino Acids and	Aromatic amino acids	Common Pathway For	Shikimate kinase (EC 2.7.1.71)	fig|552526.4.peg.677	yes	find	no

MG-RAST Overview

The metagenomics RAST server (http://metagenomics.nmpdr.org) is a SEED-based environment that allows users to upload metagenomes for automated analyses. The server is built as a modified version of the RAST server. The RAST (Rapid Annotation using Subsystem Technology) technology was originally implemented to allow automated high-quality annotation of complete or draft microbial genomes using SEED data, and has been adapted for metagenome analysis.

Our freely available server provides the annotation of sequence fragments, their phylogenetic classification, functional classification of samples, and comparison between multiple metagenomes. The server also computes an initial metabolic reconstruction for the metagenome and allows comparison of metabolic reconstructions of metagenomes and genomes.

User submission and analysis are confidential. Currently the server handles 454 and Sanger sequence data. Data sets supplied by 454 can be uploaded directly. In either case, the data needs to be in valid FASTA format. For more information, please see Which Sequences Should I Upload. For the metagenomics service please also read this explanation of metagenomics sequence formats.

The server relies on the technology and data established by FIG and the NMPDR team at Argonne National Laboratory and the University of Chicago.

In addition to SEED data we use the following ribosomal RNA databases for our analyses: greengenes, RDP-II and European ribosomal RNA database.

What's new in v2.0? **(Changes from MG-RAST v1.2)**

We have gone through several rounds of feedback with users of version 1.2 (Many thanks to all who send suggestions!) and have included the following new capabilities in version 2.0:

- Significant improvements of user interface responsiveness and overall performance.
- The ability to "publish" metagenomes on the MG-RAST server for public use.
- The ability to download subsets of fragments as FASTA (e.g. all fragments matching a given e.g. a subsystem or a functional role).
- The ability to modify parameters for sequence comparison underlying both metabolic reconstruction and phylogenetic reconstruction on the fly.
- The same capability for the heat map style comparisons of both metabolisms

and phylogenetic reconstructions.

- We have added a recruitment plot feature, plotting fragments against microbial genomes.
- We have added the ability to view all BLAST hits for a fragment and show the individual BLAST alignments.
- The ability to use KEGG maps map explore and compare metabolic reconstructions on several hierarchy levels (e.g. the high level metabolism overview).
- We have changed the pipeline that computes the underlying data so all numbers/percentages/comparisons/etc. will have changed if you look at your data in v2.0
- We have also updated the underlying databases. Most notably the SEED NR no longer from represents the status from 2006, we have added the Silva RNA database)
- Added the Invite a friend feature to share data that you submitted with other users by just entering their email addresses.
- Support for user driven creation and maintenance of groups.
- The ability to support arbitrary sets and versions of databases.
- Many small detailed fixes and improvements.

Start Page

The start page of MG-RAST (Figure 1) provides users with access to registration, data submission and management tools for uploaded data. Access to public genomes is also available once you login. Only once you have logged in and have selected a metagenome, can you gain access to your jobs, view your results and use comparative tools.

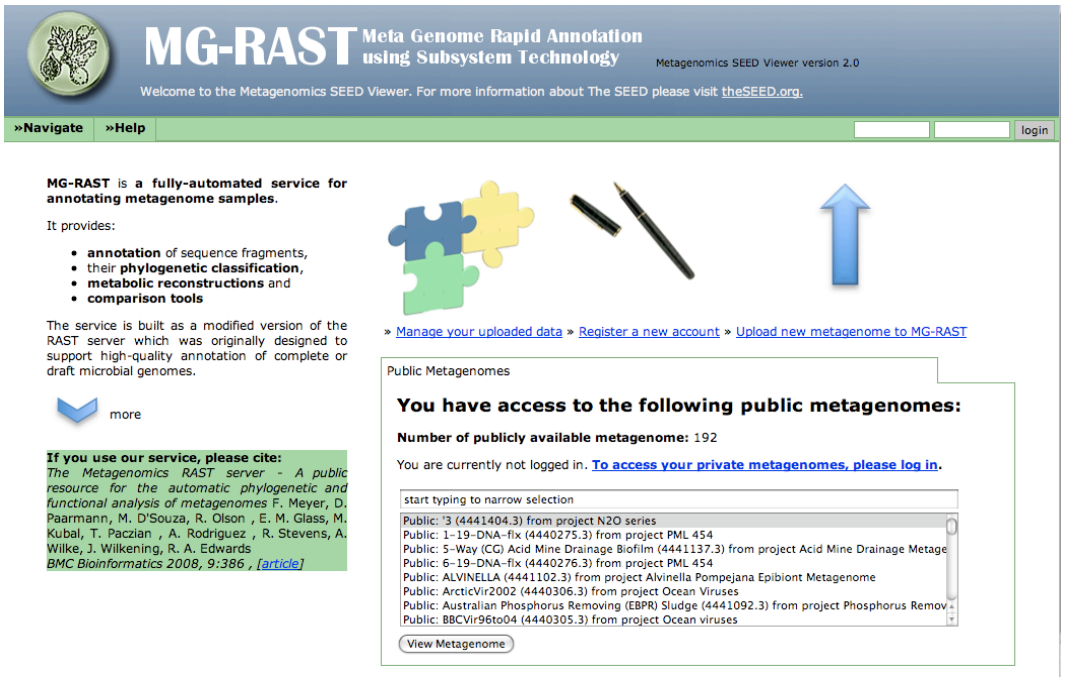


Figure 1. MG-RAST start page. Users must login to view their private metagenomes. Public metagenomes are available for browsing with or without an account.

Registration

Registering for the first time? → Choose Register A New Account/New Account. Please enter your first and last name as well as your email address into the registration form. Then please select your country and choose a login name. *It's recommended to use only letters and digits for your login name, without spaces.* You will then shortly receive an email with more information about your new account.

Already have an account for one of our other services? → Choose Register A New Account/Existing Account. Please enter your login and email of that account. If your group administrator has given you a group name, please enter it in the group name field, otherwise leave this field blank.

Upload a Genome/Creating a Job

Uploading your metagenome has a few steps. The first is uploading your FASTA file. File requirements and suggestions:

- The FASTA file name must end in .fa, .fasta, .fas, .fsa, or .fna.
- Files larger than 30 MB should compress (.tgz) their file or contact us for other options.
- Quality files (.qual) may also be included along with the sequence file for submission to MG-RAST. The quality file should be combines into a single archive (that ends in .tgz) and then uploaded to the server.

```
How to I create a .tgz file?
Create the file metagenome.tar.gz from two fna files.
tar -cvzf metagenome.tar.gz  seqfile1.fna seqfile2.fna
```

The second step requires that you provide a project name, a name for your metagenome and a brief description of the sample. The second tab shows an Upload Summary of the number of files uploaded for submission.

The third and last step asks for users to supply information about the metagenome sample. These description fields were adopted from MIGs (Minimum Information about a Genome Sequence) specification. You can also elect to make your metagenome publicly available; this option is also available if you wish to do so at a later date. During this step you also have the option of removing duplicate sequences from the analysis.

Which Sequences Should I Upload?

The Metagenomics RAST Server is designed to annotate **nucleotide** sequences from metagenome projects. You can supply either assembled or unassembled data, and reads can be as short as 100 bp and as long as you would like. There are some caveats to the system. Please also read this explanation of appropriate metagenomics sequence formats.

Unassembled Data -- If you want to do statistical comparisons between metagenomes, you most likely need unassembled sequences. The frequency that any gene is found is an approximation of the abundance of that gene in the environment. Thus if you two different samples you can compare gene frequencies between them to figure out which are the important environments. In this case, just upload the unassembled nucleotide sequences in valid FASTA format

Assembled Data -- If you want to look for complete genes or pieces of a genome, then you can use assembled sequences. These are typically longer, and the ORF caller we use on the short fragments and sequences may have problems with longer sequences. On the to do list is to add specific ORF callers for different sequence sets.

For sequences over about 1,000,000 bp (1 Mbp) you should consider pulling out those sequences individually and running them through the RAST Server for complete genomes. This server uses far superior gene identification and analysis algorithms that are only applicable once you have longer sequences. However, the algorithm will not work very well with sequences under about

1 Mbp. If you assemble sequences you will loose the frequency information, and cannot easily do statistical comparisons between metagenomes.

Common Errors and Submission Questions

- Sequence data
  - MG-RAST does not take protein sequences
  - Watch out for illegal characters in your FASTA file
  - Submission of entire genome, MG-RAST designed for short reads, use RAST for complete genomes
- Sequence data format
  - Sequences not in proper FASTA format
  - Do not use non-unique ids,
  - Watch out for sequence ids without sequences, file may be truncated.
- Sequence filename
  - Files do not have proper extensions. They must be .fa, .fasta, .fas, .fsa or .fna
  - Archive not labeled .tgz
- Submission:
  - I do not see my newly uploaded job in job list! -- your job will show up after the first step of processing is started.
- Updating analysis -- currently only way to do this is to resubmit.
- Kill a job -- email mg-rast
- Delete a job -- email mg-rast
- Deletion policy -- we will keep the data for 120 days, after that no guarantees
- Make a job public -- do this from the overview page

Jobs Overview

The overall status of your metagenome analyses can be viewed from the main portion of the Jobs Overview page. This contains information regarding a user's personal jobs and all that are public. Information includes each job/metagenome

and its status and contains information including job number, name of the user who started the job, metagenome name, and annotation progress.

The table of jobs can be sorted on Job ID or searchable (text boxes in header row). This is especially useful when the user has numerous metagenomes to select from.

Clicking on the bars for a given job in the annotation progress column directs the user to the "Job Details" page where the detailed job status and access to the metagenome analysis can be found. Job Details

Here you are able to:

- Share with selected users by providing their email addresses.
- Make the metagenome publicly accessible
- View detailed information on the processing of your job.
- View your results!!
- Download your results!!

Metagenome Overview

The overview page has several sections, all outlining various statistics of the sample and the results of the analysis (Figure 2). The overview also provides a table that allows quick navigation and information about the tools in MG-RAST which can be used to further analyze your sample.

Please note: The navigation bar has new options not previously seen on the start page or job management pages. Now you have access to tools that will allow you to compare your metagenome to other metagenomes in regard to metabolism and phylogeny (Profile). Also available is metabolic comparisons against bacterial genomes (also known as a recruitment plot).

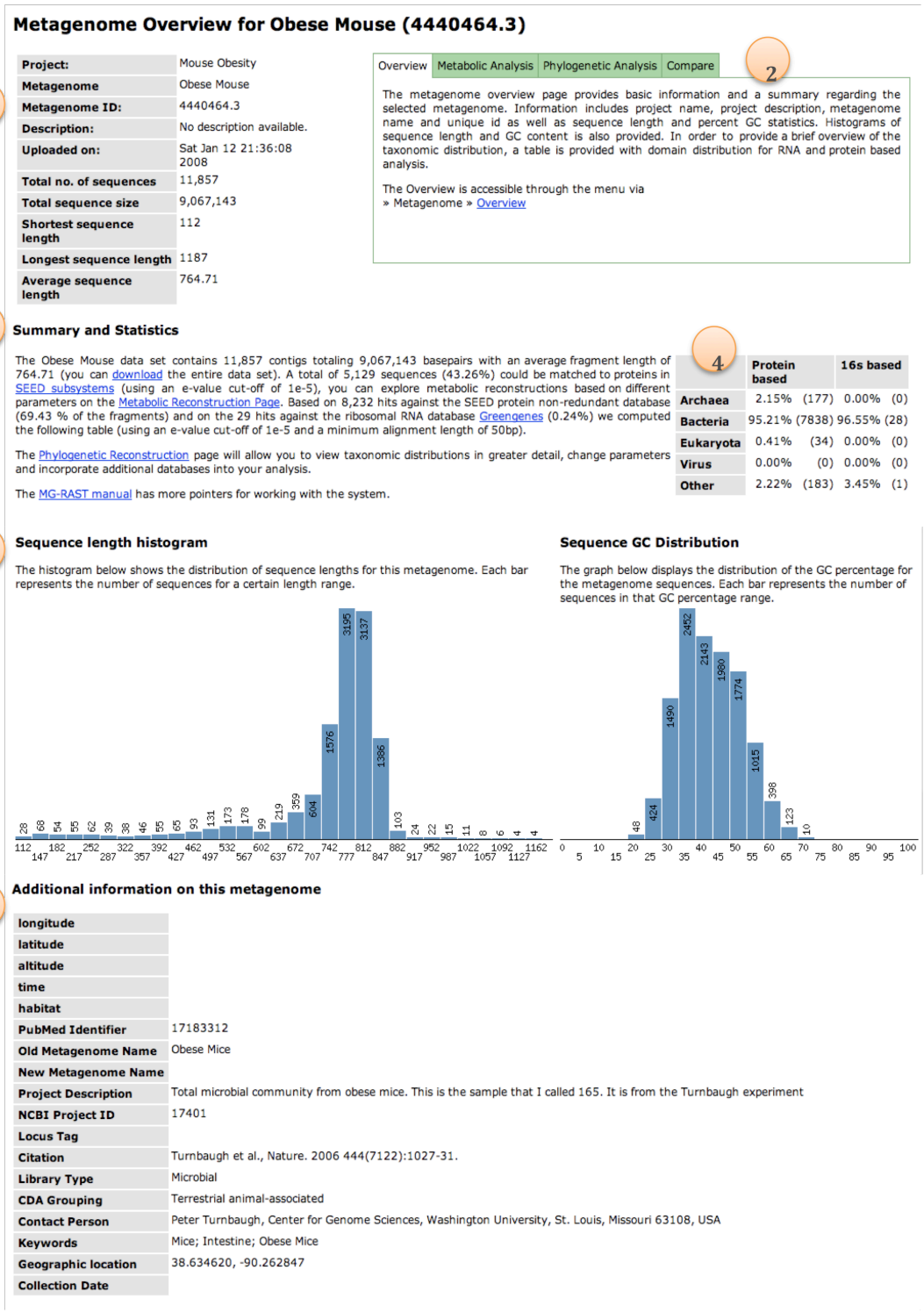


Figure 2. Metagenome Overview. 1. The overall statistics of your sample are provided (see below as to how these are calculated). 2. Links to Profile and Comparative Analysis tools. 3. A statistical summary in paragraph form. 4. A summary table of taxonomic distribution based on best protein similarity to SEED and 16S based similarity to RDP. 5. Graphical representations of sequence length and GC distributions. 6. An outline of the metagenome description and MIGS data you submitted along with your sequence file.

How are Overview statistics calculated?

* Total number of sequences

  This is the total number of sequences submitted by the user for this metagenome. Not all of these will produce results later on. It is possible and very probable that some sequences can not be matched to anything in our database.

* Total sequence size

  This is the sum of the lengths (bp) of all submitted sequences.

* Average sequence length

  This is the Total sequence length divided by the Number of sequences

* Longest sequence length

  This is the length (bp) of the longest sequence submitted.

* Shortest sequence length

  This is the length of the shortest sequence submitted

Profiles

To view your metabolic or phylogenetic profiles (Figure 3), first select the category. Once a category is selected you can then choose your dataset in which to based you profile. For metabolic reconstructions the Subsystem dataset is available. For phylogeny, RDP, Silva, European Ribosomal and GREENGENES are all options. Parameters are also changeable; users can change e-value, p-value, percent identity, and minimum alignment length. This will allow you to refine the analysis to suit the sequence characteristics of your sample. We recommend a minimal alignment length of 50bp be used with all RNA databases.

* Note: Metabolic reconstructions are based on SEED functional roles and Subsystems. (There is also a tool to view this via KEGG maps and do comparisons by going to the "Compare Metagenomes" link in the navigation bar.)

Profile results are presented in two ways: Pie chart and table. Phylogeny and Metabolism are hierarchical and the pie charts reflect that notion. By clicking on a section of the pie chart, an additional chart appears detailing the breakdown of that group. This is possible down to a third level. All selections made to the chart are reflected in the accompanying table (second tab). The numbers shown in the chart and table are actual counts.

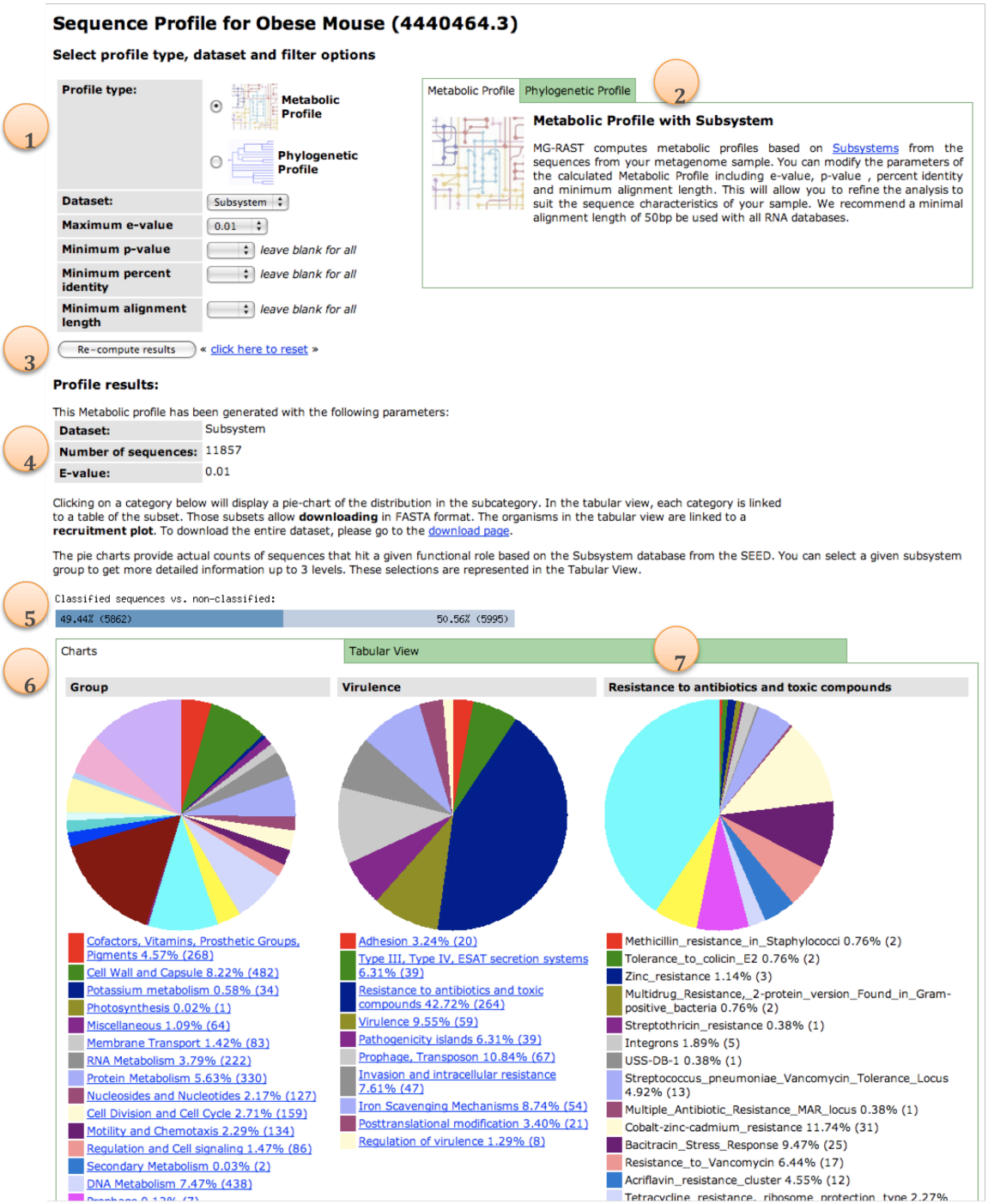


Figure 3. Sequence Profiles. This figure shows a Metabolic Profile for on of the public metagenomes in MG-RAST. Profiles have parameters that can be modified by the user as well as interactive displays. 1. Changeable parameters: select the type of profile you would like to view as well as change the parameters to identify similarity between your sample and that of the proteins in subsystems or the RNA databases. 2. Information about both profile types and recommendations. 3. Once you have modified the type of profile you want to view or any parameters, click on re-compute results. This will show the new profile based on your selections. 4. A summary of your profile restrictions. 5. A summary graph of the number of sequences that were classified given the parameters chosen. 6. This is the results pie chart. Clicking on a group will create a second pie chart

breaking down the distribution in that group. This can be iterated until you reach the final sets of subsystems or organism names. 7. A tabular view of your results are also provided. Each column is searchable and sortable and the table can be downloaded.

Compare Metagenome to other Metagenomes - Heat Maps

You can compare the metabolism or phylogeny of your metagenome with one more other metagenomes (Figure 4). Just as was seen looking at the Fragment Profile, you can select your database and modify your parameters. For metabolic reconstructions the Subsystem dataset is available. For phylogeny, RDP, Silva, European Ribosomal and GREENGENES are all options. Parameters are also changeable; users can change e-value, p-value, percent identity, and minimum alignment length. This will allow you to refine the analysis to suit the sequence characteristics of your sample. We recommend a minimal alignment length of 50bp be used with all RNA databases. The Heat Maps show the relative abundance, which is calculated using the number of sequences in a subsystem/tax class as a fraction of the total number of sequences in a subsystem/dataset. This allows for correction based on the sample size.

Compare Metagenome to Organism - Recruitment Plot

You can compare metabolism of your sample with the metabolic reconstructions from bacterial genomes (Figure 5). Choosing an organism predicted in your sample, you can compare the metabolic coverage (Figure 6). Like most of the comparative tools in MG-RAST you can modify the parameters of the calculated Metabolic Reconstruction including e-value, p-value , percent identity and minimum alignment length.

Compare Metagenome – KEGG Map

MG-RAST also enables users to view their sample on KEGG maps and compare with others (Figure 7). Mapping of functional roles to KEGG maps was done using functional assignments from analysis against the SEED. Absolute counts are provided for each KEGG map. These maps are hierarchical, just like the Subsystems, which allow you to browse the sample on various levels or compare it with other metagenomes.

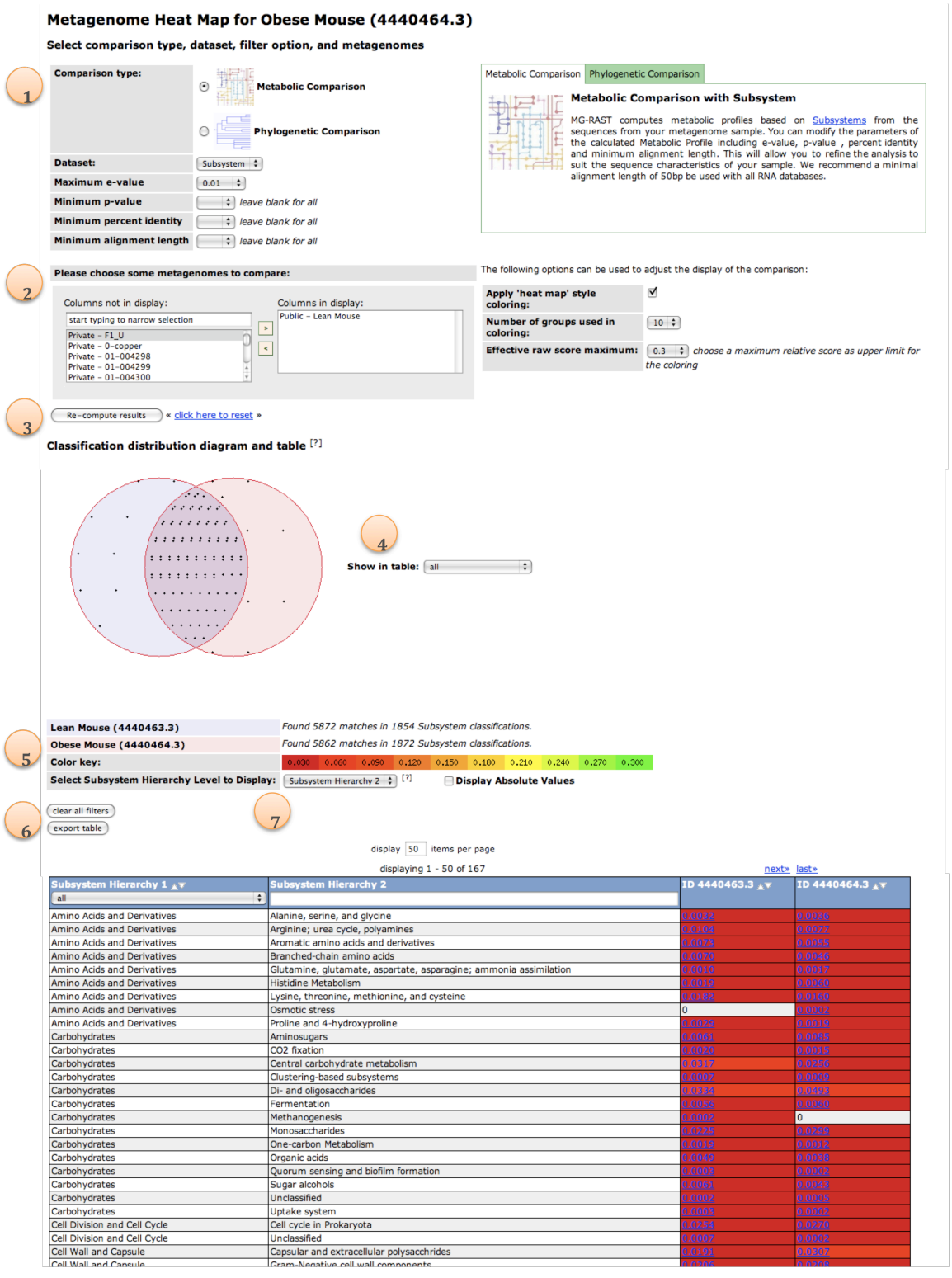


Figure 4. Heat Map Comparison. Shown is a simple example of two metagnome samples being compared with regard to their metabolic profiles. 1. Changeable parameters: select the type of profile comparison you would like to view as well as change the parameters to identify similarity between your sample and that of the proteins in subsystems or the RNA databases. 2. Here you can browse or search for metagenomes to compare with your metagenome. Make sure to add them to the

comparison by using the left and right arrow keys. 3. Once you have modified the type of comparison you want to view or any parameters, click on re-compute results. This will show the new comparison based on your selections. 4. A Venn diagram of your comparisons. Mousing over the dots provides information of what is in the union or intersections. You can also chose what group of organisms or subsystems are unique or similar to one another in the table below by using the drop down menu next to the Venn diagram. 5. A summary of the metagenomes chosen. 6. You can download the table by export the results. 7. Taxonomy and subsystems are hierarchical. Users can select what hierarchical level in which to view the results. All results tables are searchable and sortable.

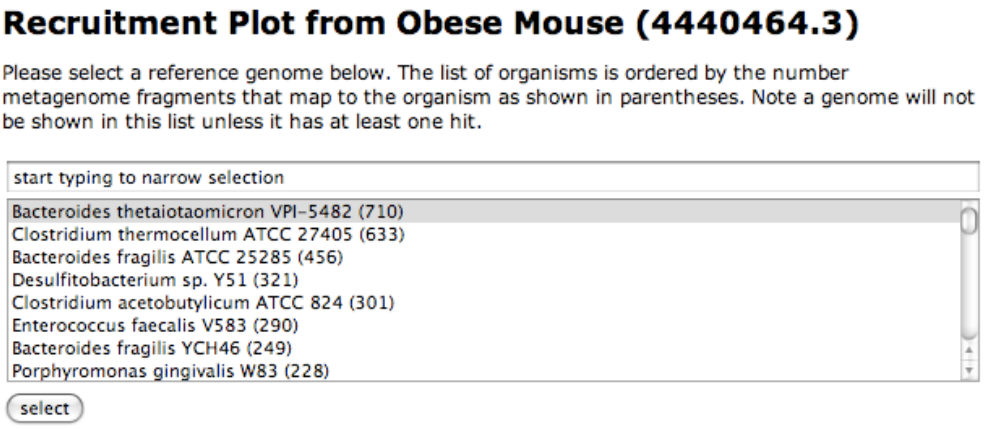


Figure 5. Choosing an organism to compare with a metagenome. The Recruitment Plot allows comparison of hits against individual bacterial genomes. The selection box shows the organisms that have hits against the metagenome as well as the number. The organisms are in order of greatest number of hits to lowest.

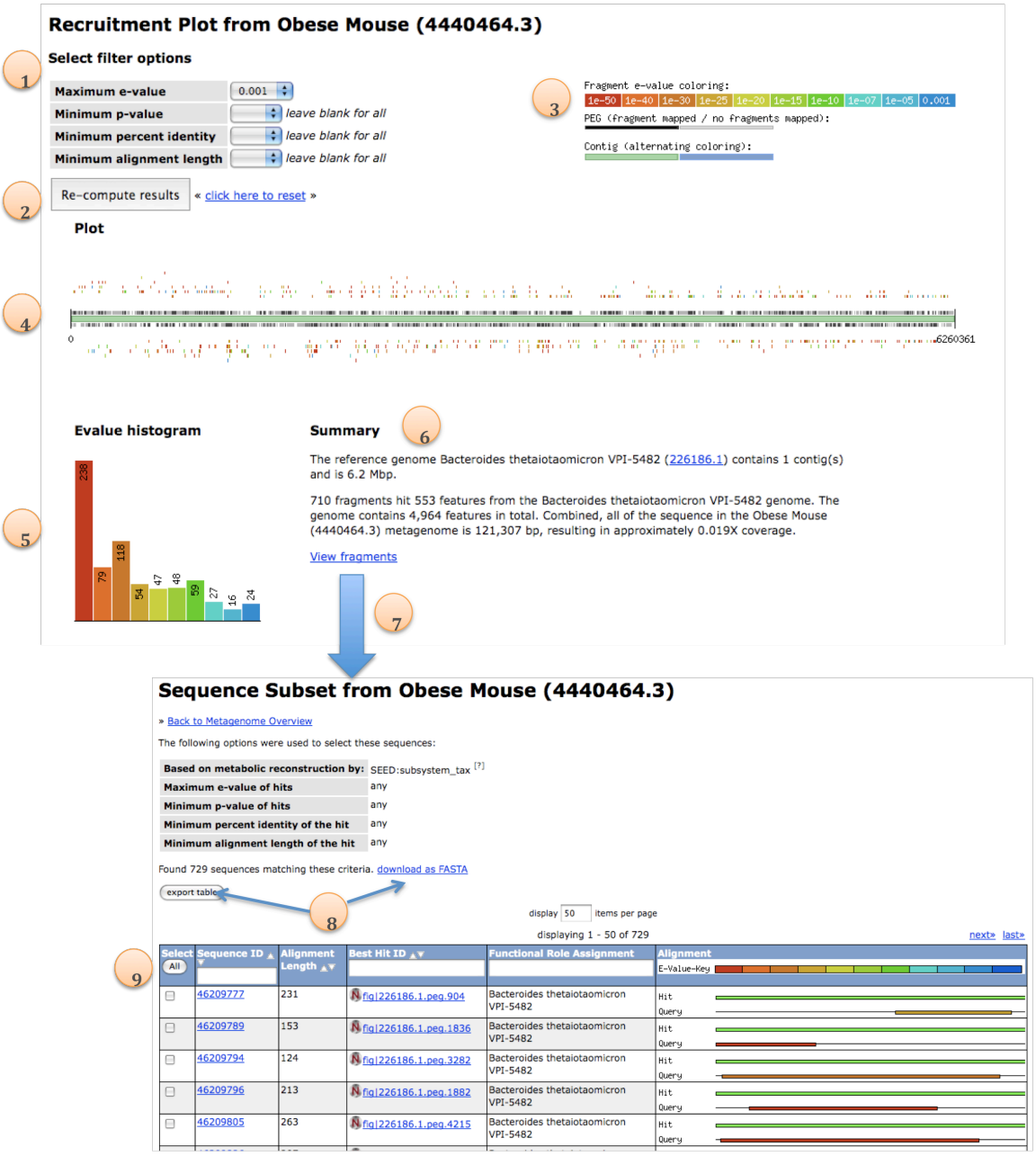


Figure 6. Recruitment Plot and fragment details. You can compare metabolism of your sample with the metabolic reconstructions from bacterial genomes. Using the organisms predicted to be in you sample, you can see the metagenome coverage of a given bacterium. 1. Changeable parameters: e-value, p-value, percent identity and minimum alignment length will allow you to modify the confidence level of your hits against the bacterial genome. 2. After you change parameters, you will need to re-computed the results. 3. E-value color legend. 4. Linear view of the bacterial chromosome with metagenome fragment hits above and below, colored by e-value. 5. E-value histogram provides the distribution of hits to the genome. Best hits are shown and used in the count. 6. Summary of the results, 7. Viewing the BLAST hits against the genome. 8. Downloadable results table or FASTA sequences. 9. BLAST results and alignments. Tables are searchable and sortable.

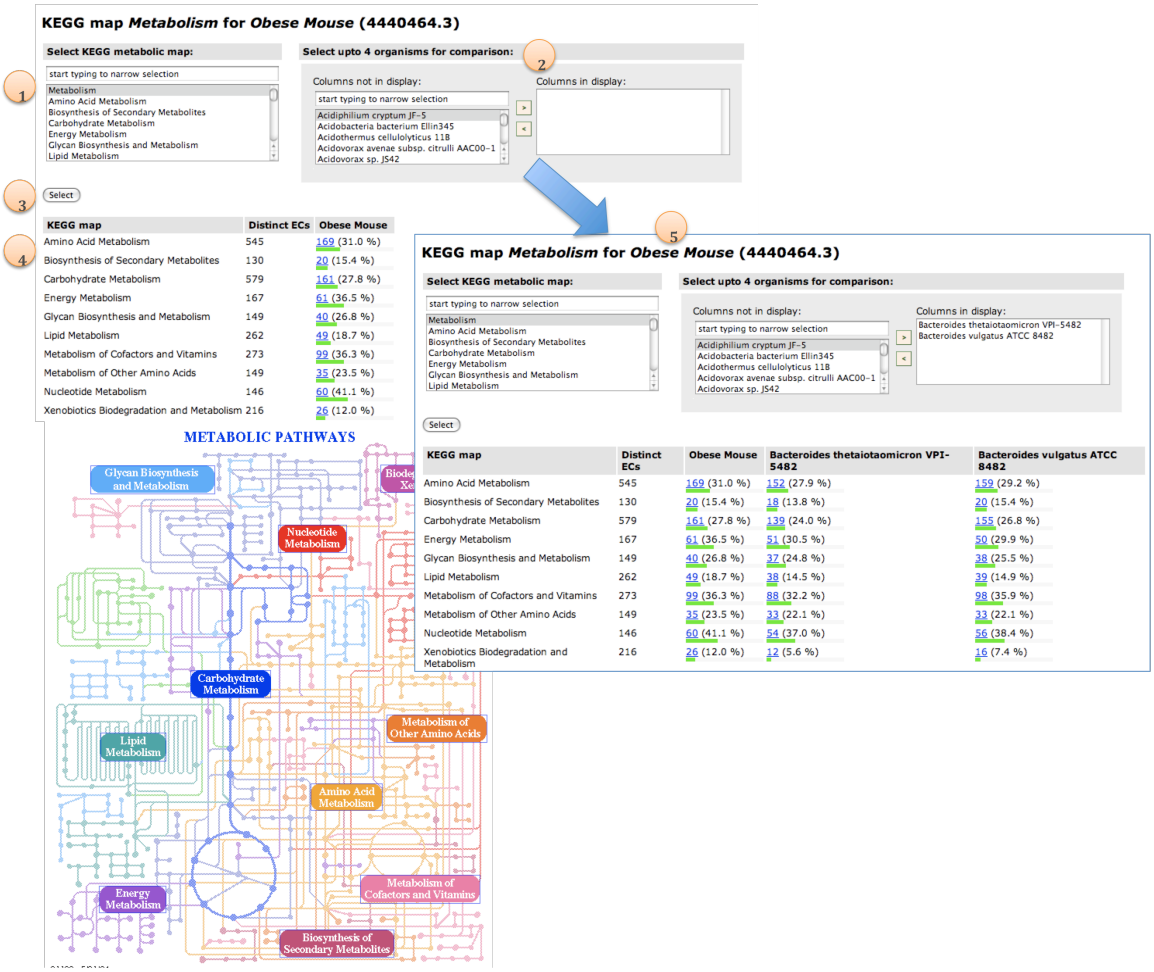


Figure 7. KEGG Maps and Comparisons. View the metabolic distribution using KEGG maps and hierarchy. User can also compare their metagenome with up to four genomes or metagenomes. Shown is the highest metabolic level in KEGG for this metagenome and two organisms that have a large number of fragments similar to their genomes. 1. Select a metabolic category. 2. Choose genome or metagenomes to compare with. 3. Click on select to view selections. 4. Pathway distribution for the metagenome. 5. After selecting two genomes to compare with, a table of results is shown. Each category is selectable to get a more refined view of the pathway or process.

MG-RAST FAQ's

What are projects?

Projects are related sets of metagenomes. If you for example plan on studying a set of samples from a chrono-series, it might me useful to group them into a project.

What level of privacy does MG-RAST v2.0 provide?

We provide password control and the ability for the submitting entity to control access on a username/password basis to the submitted data sets.

Note that we currently do not provide industry standard encryption as this would put additional load on our server infrastructure and is not strictly required for scientific purposes.

Do you support BLASTing against my private database XYZ?

We currently do not explicitly support this, however the underlying software design end system architecture support this.

How frequently do you update the underlying NR for MG-RAST?

With version 2.0 we have added the support for multiple concurrent sets of sequence similarity results to be stored per metagenome. We can add results for newer NRs. However once you start comparing results for metagenomes (say you are interested in the phylogenetic reconstruction) different versions of the NR used for the underlying data will lead to incorrect comparison results as older versions of NR will miss certain organisms and or annotations.

How long does it take to analyze my Metagenome?

The answer depends on two factors a) the size of your data set and b) the current server load. Under optimal conditions, it takes about 18 hours to run a 100 million basepair 454 metagenome through the pipeline.

How many metagenomes can I submit?

We do not restrict user submission of samples. However the computation required is massive and samples are processed on a first come first serve basis.

What parameters should I use to analyze my data?

The answer depends on your sample. In any case we recommend that you modify e-value, minimal alignment length and percent identity requirements for the BLAST results underlying the results. The effects of this are different for each sample. Depending on sample complexity, sample size, number of species an diversity of species present your results will vary dramatically when modifying these parameters. For RNA based phylogenetic reconstruction, we recommend requiring a minimum alignment length of 50bp for exact matches.

Where can people access my "published" metagenomes?

The MG-RAST v2.0 Homepage has a list of publicly accessible metagenomes.

Future versions will continue to support this feature, also we will provide a metadata based selection tool, that will allow the user to focus on metagenome data sets from the environment or condition etc they are interested in.

What about HIPAA relevant data?

MG-RAST is provided under the assumption that all data is anonymized, no HIPAA relevant data should be stored on MG-RAST.

How can I download a subset of fragments in FASTA format?

Many pages support downloading the data into a spreadsheet format (e.g. MS Excel). One the Metabolic Reconstruction page or the Phylogenetic reconstruction, you can download a subset of the fragments contained in the sample matching a specific group of organisms or matching a specific part of metabolism via clicking on the tab for Tabular view. There you click on a given subset.

Glossary

Accession number1 (GenBank). The accession number is the unique identifier assigned to the entire sequence record when the record is submitted to GenBank. The GenBank accession number is a combination of letters and numbers that are usually in the format of one letter followed by five digits (e.g., M12345) or two letters followed by six digits (e.g., AC123456). The accession number for a particular record will not change even if the author submits a request to change some of the information in the record. Take note that an accession number is a unique identifier for a complete sequence record, while a Sequence Identifier, such as a Version, GI, or ProteinID, is an identification number assigned just to the sequence data. The NCBI Entrez System is searchable by accession number using the Accession [ACCN] search field.

Accession number2 (RefSeq). This accession number is the unique identification number for a complete RefSeq sequence record. RefSeq accession numbers are written in the following format: two letters followed by an underscore and six digits (e.g., NT_123456). The first two letters of the RefSeq accession number indicate the type of sequence included in the record as described below:

- NT_123456 constructed genomic contigs
- NM_123456 mRNAs (actually the cDNA sequences constructed from mRNA)
- NP_123456 proteins
- NC_123456 chromosomes

Annotation. Please see Assigning a gene function and annotation.

Assigning a gene function and annotation. Annotators assign gene functions to genes, and we call this process annotation. In most contexts, people use the term annotation to refer to assignments of function to the genes within a single organism. We certainly use the term in this sense, but we also use it to describe the process of assigning functions to corresponding genes from numerous genomes. Our basic approach to annotation is to ask our annotators to annotate the genes included in a Subsystem (e.g., glycolysis) across all genomes. This process of annotation of the genes within a subsystem across a set of genomes, rather than annotation of genes within a single genome, allows our annotators to focus on a constrained set of functional roles and attempt to accurately identify exactly what variant, if any, of a subsystem exists in each of the genomes.

We use the term annotation to refer to assigning functions to genes (either within a single organism or to a constrained set of gene/protein families across a set of organisms). This activity certainly is closely related to the construction of subsystems and protein families (which we call FIGfams), but we will describe those activities elsewhere.

Assignment. Please see Assigning a gene function and annotation

Bidirectional Best Hit (BBH). The paper "The use of gene clusters to infer functional coupling" defines a Bidirectional Best Hit or BBH as follows:

Given two genes Xa and Xb from two genomes Ga and Gb, Xa and Xb are called a "bidirectional best hit (BBH)" if and only if recognizable similarity exists between them (in our case, we required fasta3 scores lower than 1.0 × 10-5), there is no gene Zb in Gb that is more similar than Xb is to Xa, and there is no gene Za in Ga that is more similar than Xa is to Xb. Genes (Xa, Ya) from Ga and (Xb, Yb) from Gb form a "pair of close bidirectional best hits (PCBBH)" if and only if Xa and Ya are close, Xb and Yb are close, Xa and Xb are a BBH, and Ya and Yb are a BBH.

Bit score. The value S' is derived from the raw alignment score S in which the statistical properties of the scoring system used have been taken into account. Because bit scores have been normalized with respect to the scoring system, they can be used to compare alignment scores from different searches.

BLAST. The Basic Local Alignment Search Tool finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

	Default	Special Cases		
		Short Query	Large Sequence Family	Ungapped BLAST
Filter	on	off	on	on
Scoring Matrix	BLOSUM62	PAM30 for 35 and under	BLOSUM62	BLOSUM62
Word Size	3	3, or reduce to 2	3	3
E value	10	1000 or more	10	10
Gap costs	11,1	11,1	11,1	4
Alignments	50	50	2000	50

CDD Conserved Domain Database. This database is a collection of sequence alignments and profiles representing protein domains conserved during molecular evolution.

CDS/coding sequence. CDS refers to the portion of a genomic DNA sequence that is translated, from the start codon to the stop codon, inclusively, if complete. A partial CDS lacks part of the complete CDS (it may lack either or both the start and stop codons). Successful translation of a CDS results in the synthesis of a protein.

E-value/Expect value. The E-value is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. It decreases exponentially with the score (S) that is assigned to a match between two sequences. It is important to note that searches with short sequences can be virtually identical and have relatively high E-value. This is because the calculation of the E-value takes into account the length of the query sequence (besides database size). This is because shorter sequences have a high probability of occurring in the database purely by chance.

EC number. A number assigned to a type of enzyme according to a scheme of standardized enzyme nomenclature developed by the Enzyme Commission of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB). Every enzyme code consists of the letters "EC" followed by four numbers separated by periods. Those numbers represent a progressively finer classification of the enzyme. The main classification can be broken into 6 groups: 1. oxidoreductases, 2. Transferases, 3. Hydrolases, 4. Lyases, 5. Isomerases, and 6. Ligases.

European ribosomal RNA database. A database containing all complete or nearly complete SSU (small subunit) and LSU (large subunit) ribosomal RNA sequences. http://bioinformatics.psb.ugent.be/webtools/rRNA/

FastGroupII. It is a web-based bioinformatics platform to de-replicate large 16S rDNA libraries. FastGroupII provides users with the option of four different de-replication methods, performs rarefaction analysis, and automatically calculates the Shannon-Wiener Index and Chao1. It is recommended that one use FastGroupII for clustering and primary analysis of 16S libraries, and then the data from that can be fed into RDP Classifier and other programs.

FASTA. FASTA format is a text-based format for representing either nucleic acid sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences.

A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. The word following the ">" symbol is the identifier of the sequence, and the rest of the line is the description (both are optional). There should be no space between the ">" and the first letter of the identifier. It is recommended that all lines of text be shorter than 80 characters. The sequence ends if another line starting with a ">" appears; this indicates the start of another sequence.

Feature. A feature is a defined region in the DNA. A PEG is the most prevalent feature type in the SEED. Some other feature types include RNA, prophage and pathogenicity islands. The format for a feature ID is fig|genome_id.feature_abbreviation.feature_number (ie fig|83333.1.peg.100).

FIGfam. FIGfams are protein families generated by the Fellowship for Interpretation of Genomes (FIG). These families are based on the collection of subsystems, as well as correspondences between genes in closely related strains (we describe the construction of FIGfams in a separate SOP). The important properties of these families are as follows:

1. Two PEGs which both occur within a single FIGfam are believed to have the same function.

2. There is a decision procedure associated with the family which can be invoked to determine whether or not a gene can be "safely" assigned the function associated with the family.

FIG Identifier / FIG-IDs. We provide identifiers for genome sequences and features in the following form:

Entity type	key	identifier
Genome	genome	fig|83331.1
PEG	id	fig|83331.peg.123
RNA feature	id	fig|83331.rna.1

Functional role. The concept of functional role is both basic and primitive in the sense that we will not pretend to offer a precise definition. It corresponds roughly to a single logical role that a gene or gene product may play in the operation of a cell.

GBK. ".gbk" GenBank flat file format for complete bacterial genomes. They are available at the NCBI FTP site along with a variety of formats (GenBank summary file (*.gbs), FASTA Nucleic Acid file (*.fna), FASTA Amino Acid file (*.faa), Protein Table (*.ptt), etc). ftp://ftp.ncbi.nih.gov/genomes/Bacteria/.

Gene function. The function of a protein-encoding gene (PEG) is the functional role played by the product of the gene or an expression describing a set of roles played by the encoded protein. The operators used to construct expressions and the meanings associated with the operators are described in

http://www.nmpdr.org/FIG/Html/SEED_functions.html

Genes other than PEGs can also be assigned functions (e.g., SSU rRNA). However, in most cases the functions assigned to genes other than PEGs tend not to be problematic.

Gene Ontology. A controlled vocabulary of terms relating to molecular function, biological process, or cellular components developed by the Gene Ontology Consortium. A controlled vocabulary allows scientists to use consistent terminology when describing the roles of genes and proteins in cells.

Greengenes. A database and web application that provides access to the current and comprehensive 16S rRNA gene sequence alignment for browsing, blasting, probing, and downloading. http://greengenes.lbl.gov/cgi-bin/nph-index.cgi

gzip. A software application used for file compression. gzip is short for GNU zip; the program is freely available at http://www.gnu.org/software/gzip/.

```
gzip -c metagenome.tar  > metagenome.tar.gz
```

Homologous. The term refers to similarity attributable to descent from a common ancestor.

InterPro. A searchable database providing information on sequence function and annotation. Sequences are grouped based on protein signatures or 'methods'. These groups represent superfamilies, families or sub-families of sequences. The groups may be defined as FAMILIES, DOMAINS, REPEATS OR SITES. The function of sequences within any group may be confined to a single biological process or it may be diverse range of functions (as in a superfamily) or the group may be functionally uncharacterized but without exception every entry has an abstract and references are provided where possible.

Metabolic Reconstruction. When we use the term metabolic reconstruction of a given genome we will simply mean the set of populated subsystems that contain the genome, the PEGs (and their assigned functions) that are connected to functional roles in these populated subsystems, and the specific variant code associated with the genome in each of the populated subsystems.

NCBI Accession number. An Accession number is a unique identifier given to a sequence when it is submitted to one of the DNA repositories (GenBank, EMBL, DDBJ). The initial deposition of a sequence record is referred to as version 1. If the sequence is updated, the version number is incremented, but the Accession number will remain constant.

Ortholog. Orthologs are genes in different species that derive from a common ancestor, i.e., they are direct evolutionary counterparts.

Pair of Close Homologs (PCH). The paper "The use of gene clusters to infer functional coupling" defines a Pair of Close Homologs as follows:

We can also define the concept of "pairs of close homologs" (PCHs) as follows: genes (X′a, Y′a) from Ga and (X′b, Y′b) from Gb form a PCH if and only if X′a and Y′a are close, X′b and Y′b are close, X′a and X′b are recognizably similar, and Y′a and Y′b are recognizably similar. Here, we will consider two genes to be recognizably similar if their gene products produce fasta3 scores lower than 1.0 × 10-5. We use a scoring scheme analogous to the one described for PCBBHs to evaluate the connections between PCHs, except that if Ga and Gb are the same genome, we assign an arbitrary "same-genome score" ("same-genome" pairs cannot occur for PCBBHs by definition, but for PCHs they are possible). Unlike PCBBHs from two very close genomes for which contiguity is completely uninformative in the vast majority of cases, PCHs allow recognition of gene clusters that play similar (but usually not identical) roles (such as two transport cassettes

containing pairs of homologs) in the same or similar organisms. The arbitrary "same-genome score" should, we believe, have a value that is high enough to rank such instances as significant.

Paralog. A paralog is one of a set of homologous genes that have diverged from each other as a consequence of gene duplication.

PEG. A Protein Encoding Gene (PEG) is equivalent to a CDS (Coding Sequence).

Pfam . Pfam is a database of multiple sequence alignments and hidden Markov models covering many common protein domains.

Phenotype. Phenotype is any observable traits or characteristics of an organism, e.g., gram stain, shape, or the presence or absence of a disease. Phenotypic traits are not necessarily genetic.

Populated Subsystem. Please see Subsystem

PSORT. A program for analysis of protein sorting signals and prediction of subcellular localization. PSORT receives the information of an amino acid sequence and its source orgin, as inputs. Then, it analyzes the input sequence by applying the stored rules for various sequence features of known protein sorting signals. Finally, it reports the possibility for the input protein to be localized at each candidate site with additional information.

Radar. **R**apid **A**utomatic **D**etection and **A**lignment of **R**epeats in protein sequences. Many large proteins have evolved by internal duplication and many internal sequence repeats correspond to functional and structural units. Radar is uses an automatic algorithm, for segmenting your query sequence into repeats, it identifies short composition biased as well as gapped approximate repeats and complex repeat architectures involving many different types of repeats in your query sequence.

RAST. RAST or Rapid Annotation using Subsystem Technology is a rapid and very accurate annotation technology. We make a RAST server available for public use at: http://rast.nmpdr.org

RDP. The Ribosomal Database Project (RDP) complies ribosomal sequences and related data, and redistributes them in aligned and phylogenetically ordered form to the scientific community. RDP also has a variety of software tools for handling, analyzing and displaying sequences. http://rdp.cme.msu.edu/index.jsp

SEED-Viewer. The SEED Viewer is a web-based application that allows browsing of SEED data structures. We use the SEED-Viewer to provide a public read-only version of the latest SEED data at: http://seed-viewer.theseed.org

Silva. A database that provides comprehensive, quality checked and regularly updated databases of aligned small and large subunit ribosomal RNA sequences for all three domains of life. http://www.arb-silva.de/

Subsystem. A subsystem is a set of functional roles that an annotator has decided should be thought of as related. Frequently, subsystems represent the collection of functional roles that make up a metabolic pathway, a complex (e.g., the ribosome), or a class of proteins (e.g., two-component signal-transduction proteins within Staphylococcus aureus). A populated subsystem is a subsystem with an attached spreadsheet. The rows of the spreadsheet represent genomes and the columns represent the functional roles of the spreadsheet. Each cell contains the identifiers of genes from the corresponding genome the implement the specific functional role. That is, a populated subsystem specifies which genes implement instances of the subsystem in each of the genomes. The rows of a populated genome are assigned variant codes which describe which of a set of possible variants of the subsystem exist within each genome (special codes expressing a total absence of the subsystem or remaining uncertainty exist). Construction of a large set of curated populated subsystems is at the center of the NMPDR and SEED annotation efforts.

Tar. Is a (Unix) a single specialist program designed to store and extract files from an archive file known as a tarfile. a tape archive. A ".tar" file is not a compressed files, it is actually a collection of files within a single file uncompressed. If the file is a .tar.gz ("tarball") or ".tgz" file it is a collection of files that is compressed. If you are looking to compress a file you would create the tar file then **gzip** the file.

- To create a Tar file

  Creates a GZIP-compressed Tar file of the name metagenome.tar.gz of twofiles.

  tar -cvzf metagenome.tar.gz seqfile1.fna seqfile2.fna

- To list files in a compressed Tar file

  tar -tzf metagenome.tar.gz

- To extract files from a Tar file

  Extracts all files from a compressed Tar file of the name metagenome.tar.gz.

  tar -xvf metagenome.tar.gz

TaxID/Taxonomy Identifier. The TaxID is a stable unique identifier for each taxon (for a species, a family, an order, or any other group in the taxonomy database). The taxID is seen in the GenBank records as a "source" feature table entry; for example, db_xref="taxon:83333" is the TaxID for E. coli K12.

Variant Code. Please see Subsystems

Acknowledgments

National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract HHSN266200400042C.

APPENDIX A. Publications that cite out tools and work.

Over 100 publications have cited our systems or work!!

Describing our tools

1. **The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.** Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. BMC Bioinformatics. 2008 Sep 19;9:386. PMID: 18803844
2. **The RAST Server: Rapid annotations using subsystems technology.** Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. BMC Genomics. 2008 Feb 8;9:75. PMID: 18261238
3. **The National Microbial Pathogen Database Resource (NMPDR): A genomics platform based on subsystem annotation.** McNeil LK, Reich C, Aziz RK, Bartels D, Cohoon M, Disz T, Edwards RA, Gerdes SY, Hwang K, Kubal M, Margaryan GR, Meyer F, Mihalo W, Olsen GJ, Olson R, Osterman AL, Paarmann D, Paczian T, Parrello B, Pusch GD, Rodionov DA, Shi X, Vassieva O, Vonstein V, Zagnitko OP, Xia F, Zinner J, Overbeek R, Stevens R. Nucleic Acids Res. 2007 Jan;35(Database issue):D347-53. [Epub 2006 Dec 1] PMID: 17145713
4. **National Institute of Allergy and Infectious Diseases bioinformatics resource centers: New assets for pathogen informatics.** Greene JM, Collins F, Lefkowitz EJ, Roos D, Scheuermann RH, Sobral B, Stevens R, White O, Di Francesco V. Infect Immun. 2007 Jul;75(7):3212-9. [Epub 2007 Apr 9] PMID: 17420237
5. **Annotation of bacterial and archaeal genomes: improving accuracy and consistency.** Overbeek R, Bartels D, Vonstein V, Meyer F. Chem Rev. 2007 Aug;107(8):3431-47. [Epub 2007 Jul 21] PMID: 17658903
6. **The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.** Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O,

Ye Y, Zagnitko O, Vonstein V. Nucleic Acids Res. 2005 Oct 7;33(17):5691-702. PMID: 16214803

Using our tools

1. **Identification of genes encoding the folate- and thiamine-binding membrane proteins in firmicutes.** Eudes A, Erkens GB, Slotboom DJ, Rodionov DA, Naponelli V, Hanson AD. J Bacteriol. 2008 Nov;190(22):7591-4. [Epub 2008 Sep 5] PMID: 18776013
2. **The dual transcriptional regulator CysR in *Corynebacterium glutamicum* ATCC 13032 controls a subset of genes of the McbR regulon in response to the availability of sulphide acceptor molecules.** Rückert C, Milse J, Albersmeier A, Koch DJ, Pühler A, Kalinowski J. BMC Genomics. 2008 Oct 14;9:483. PMID: 18854009
    - Despite a lack of functional clustering, Cg0156 was shown to activate transcription of the genes in the pathway for assimilatory reduction of sulphate, *fpr2 cysIXHDNYZ,* in the Cysteine Biosynthesis subsystem.
3. **RNomics and Modomics in the halophilic archaea *Haloferax volcanii:* identification of RNA modification genes.** Grosjean H, Marck C, Gaspin C, Decatur WA, de Crecy-Lagard V. BMC Genomics. 2008 Oct 9;9(1):470. PMID: 18844986
4. **Identification and characterization of genes underlying chitinolysis in *Collimonas fungivorans* Ter331.** Fritsche K, de Boer W, Gerards S, van den Berg M, van Veen JA, Leveau JH. FEMS Microbiol Ecol. 2008 Oct;66(1):123-35. [Epub 2008 Jul 30] PMID: 18671744
5. **Rise and persistence of global M1T1 clone of *Streptococcus pyogenes.*** Aziz RK, Kotb M. Emerg Infect Dis. 2008 Oct;14(10):1511-7. PMID: 18826812
6. ***Vibrio cholerae* VciB promotes iron uptake via ferrous iron transporters.** Mey AR, Wyckoff EE, Hoover LA, Fisher CR, Payne SM. J Bacteriol. 2008 Sep;190(17):5953-62. [Epub 2008 Jun 27] PMID: 18586940
7. **Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities.** Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P, Joint I. PLoS ONE. 2008 Aug 22;3(8):e3042. PMID: 18725995
    - MG-RAST and our subsystems were used to reconstruct and compare the taxonomy and metabolism of metagenomic and metatranscriptomic data sets collected in replicate from two communities at two time points.
8. **Genome sequence of a Lancefield group C *Streptococcus zooepidemicus* strain causing epidemic nephritis: New information about an old disease.** Beres SB, Sesso R, Pinto SW, Hoe NP, Porcella SF, Deleo FR, Musser JM. PLoS ONE. 2008 Aug 21;3(8):e3026. PMID: 18716664
    - The RAST server was used as one tool to develop a comprehensive annotation of the features in this genome.
9. **The type III pantothenate kinase encoded by coaX is essential for growth of *Bacillus anthracis.*** Paige C, Reid SD, Hanna PC, Claiborne A. J Bacteriol. 2008 Sep;190(18):6271-5. [Epub 2008 Jul 18] PMID: 18641144

10. **Comparative metagenomics reveals host specific metavirulomes and horizontal gene transfer elements in the chicken cecum microbiome.** Qu A, Brulc JM, Wilson MK, Law BF, Theoret JR, Joens LA, Konkel ME, Angly F, Dinsdale EA, Edwards RA, Nelson KE, White BA. PLoS ONE. 2008 Aug 13;3(8):e2945. PMID: 18698407
    - MG-RAST and our subsystems were used to characterize the microbial community structure and functional gene content of the chicken cecal microbiome from a pathogen-free chicken and one that had been challenged with *Campylobacter jejuni.*
11. **Towards environmental systems biology of *Shewanella.*** Fredrickson JK, Romine MF, Beliaev AS, Auchtung JM, Driscoll ME, Gardner TS, Nealson KH, Osterman AL, Pinchuk G, Reed JL, Rodionov DA, Rodrigues JL, Saffarini DA, Serres MH, Spormann AM, Zhulin IB, Tiedje JM. Nat Rev Microbiol. 2008 Aug;6(8):592-603. [Epub 2008 Jul 7] PMID: 18604222
12. ***Streptococcus iniae* M-Like protein contributes to virulence in fish and is a target for live attenuated vaccine development.** Locke JB, Aziz RK, Vicknair MR, Nizet V, Buchanan JT. PLoS ONE. 2008 3(7): e2824. doi:10.1371/journal.pone.0002824. PMID: 18665241
13. **Identification of a cellobiose utilization gene cluster with cryptic beta-galactosidase activity in *Vibrio fischeri.*** Adin DM, Visick KL, Stabb EV. Appl Environ Microbiol. 2008 Jul;74(13):4059-69. [Epub 2008 May 16] PMID: 18487409
14. **Large-scale transposon mutagenesis of *Mycoplasma pulmonis.*** French CT, Lao P, Loraine AE, Matthews BT, Yu H, Dybvig K. Mol Microbiol. 2008 Jul;69(1):67-76. [Epub 2008 Apr 28] PMID: 18452587
15. **Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome.** Urich T, Lanzén A, Qi J, Huson DH, Schleper C, Schuster SC. PLoS ONE. 2008 Jun 25;3(6):e2527. PMID: 18575584
    - MG-RAST and our subsystems were used to reconstruct and compare the taxonomy and metabolism of a soil community from metatranscriptomic data.
16. **Biochemical and phylogenetic characterization of a novel diaminopimelate biosynthesis pathway in prokaryotes identifies a diverged form of LL-diaminopimelate aminotransferase.** Hudson AO, Gilvarg C, Leustek T. J Bacteriol. 2008 May;190(9):3256-63. [Epub 2008 Feb 29] PMID: 18310350
17. **Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes.** Frank JA, Reich CI, Sharma S, Weisbaum JS, Wilson BA, Olsen GJ. Appl Environ Microbiol. 2008 Apr;74(8):2461-70. [Epub 2008 Feb 22] PMID: 18296538
    - A phylogenetically representative set of full-length bacterial SSU rRNA sequences was extracted from The SEED.
18. **Phylogenomic and functional analysis of pterin-4a-carbinolamine dehydratase family (COG2154) proteins in plants and microorganisms.** Naponelli V, Noiriel A, Ziemak MJ, Beverley SM, Lye LF, Plume AM, Botella

JR, Loizeau K, Ravanel S, Rébeillé F, de Crécy-Lagard V, Hanson AD. Plant Physiol. 2008 Apr;146(4):1515-27. [Epub 2008 Feb 1] PMID: 18245455

- The COG2154 family of proteins was discovered to have members in genomes with or without aromatic amino acid hydroxylases (AAHs), which generate oxidized pterin cofactors recycled by Pterin-4a-carbinolamine dehydratases (PCDs). Partnerless PCDs are hypothesized to support the function of presently unrecognized pterin-dependent enzymes. A signature motif for PCD activity, which may be used as a query term in protSCAN (available from our Sequence Search page), was discovered to be: any(EDKH) 3...3 H any(HN) any(PCS) 5...6 any(YWF) 9...9 any(HW) 8...15 D

19. **Transcriptional regulation of NAD metabolism in bacteria: NrtR family of Nudix-related regulators.** Rodionov DA, De Ingeniis J, Mancini C, Cimadamore F, Zhang H, Osterman AL, Raffaelli N. Nucleic Acids Res. 2008 Apr;36(6):2047-59. [Epub 2008 Feb 14] PMID: 18276643
20. **Transcriptional regulation of NAD metabolism in bacteria: Genomic reconstruction of NiaR (YrxA) regulon.** Rodionov DA, Li X, Rodionova IA, Yang C, Sorci L, Dervyn E, Martynowski D, Zhang H, Gelfand MS, Osterman AL. Nucleic Acids Res. 2008 Apr;36(6):2032-46. [Epub 2008 Feb 14] PMID: 18276644
21. **Cohesion Group Approach for Evolutionary Analysis of TyrA, a Protein Family with Wide-Ranging Substrate Specificities.** Bonner CA, Disz T, Hwang K, Song J, Vonstein V, Overbeek R, Jensen RA. Microbiol Mol Biol Rev. 2008 Mar;72(1):13-53. PMID: 18322033
    - The TyrA dehydrogenases are used as a prototype example of how a credible picture of evolutionary events can be deduced within the vertical trace of inheritance in combination with intervening events of lateral gene transfer (LGT). Figures and tables supplemental to the paper are found on the TyrA page.
22. **Functional metagenomic profiling of nine biomes.** Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, Mc Daniel L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F. Nature 2008 Mar 12. PMID: 18337718
    - MG-RAST and our subsystems were used to reconstruct and compare the metabolism of different microbial environments.
23. **ComPath: Comparative enzyme analysis and annotation in pathway/subsystem contexts.** Choi K, Kim S. BMC Bioinformatics. 2008 Mar 6;9:145. PMID: 18325116
    - ComPath provides a total of 327 model pathways combining 205 pathways from KEGG database and 122 subsystems from The SEED and NMPDR.
24. **Glycerate 2-kinase of *Thermotoga maritima* and genomic reconstruction of related metabolic pathways.** Yang C, Rodionov DA, Rodionova IA, Li X, Osterman AL. J Bacteriol. 2008 Mar;190(5):1773-82. [Epub 2007 Dec 21] PMID: 18156253

25. **Microbial ecology of four coral atolls in the Northern Line Islands.** Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, Wegley L, Hatay M, Hall D, Brown E, Haynes M, Krause L, Sala E, Sandin SA, Thurber RV, Willis BL, Azam F, Knowlton N, Rohwer F. PLoS ONE. 2008 Feb 27;3(2):e1584. PMID: 18301735
    - MG-RAST and our subsystems were used to characterize differences in microbial communities across atolls that could reflect natural environmental variation or human impacts.
26. **Sialic acid mutarotation is catalyzed by the *Escherichia coli* beta-propeller protein YjhT.** Severi E, Müller A, Potts JR, Leech A, Williamson D, Wilson KS, Thomas GH. J Biol Chem. 2008 Feb 22;283(8):4841-9. [Epub 2007 Dec 5] PMID: 18063573
    - A previously uncharacterized protein present in many sialic acid-utilizing pathogens, YjhT, was proven to accelerate the equilibration of the alpha- and beta-anomers of N-acetylneuraminic acid, thus describing a novel sialic acid mutarotase activity. The conservation of its genomic position near sialometabolic and sialic acid-inducible genes was explored using Compare Regions.
27. **Bacterial carbon processing by generalist species in the coastal ocean.** Mou X, Sun S, Edwards RA, Hodson RE, Moran MA. Nature. 2008 Feb 7;451(7179):708-11.[Epub 2008 Jan 27] PMID: 18223640
    - MG-RAST and our subsystems were used to directly measure niche breadth for bacterial functional assemblages.
28. **Structural basis for substrate binding and the catalytic mechanism of type III pantothenate kinase.** Yang K, Strauss E, Huerta C, Zhang H. Biochemistry. 2008 Feb 5;47(5):1369-80. [Epub 2008 Jan 11] PMID: 18186650
    - A comprehensive analysis of the PanK-encoding genes in the Coenzyme A Biosynthesis subsystem revealed that PanK-III enzymes have a much wider phylogenetic distribution than the better known PanK-I, being present in 12 of the 13 major bacterial groups, and in many pathogens.
29. **Bifunctional NMN adenylyltransferase/ADP-ribose pyrophosphatase: Structure and function in bacterial NAD metabolism.** Huang N, Sorci L, Zhang X, Brautigam CA, Li X, Raffaelli N, Magni G, Grishin NV, Osterman AL, Zhang H. Structure. 2008 Feb;16(2):196-209. PMID: 18275811
30. **An in vivo expression technology screen for *Vibrio cholerae* genes expressed in human volunteers.** Lombardo MJ, Michalski J, Martinez-Wilson H, Morin C, Hilton T, Osorio CG, Nataro JP, Tacket CO, Camilli A, Kaper JB. Proc Natl Acad Sci U S A. 2007 Nov 13;104(46):18229-34. [Epub 2007 Nov 6] PMID: 17986616
31. **Comparative RNomics and modomics in Mollicutes: Prediction of gene function and evolutionary implications.** de Crécy-Lagard V, Marck C, Brochier-Armanet C, Grosjean H. IUBMB Life. 2007 Oct;59(10):634-58. PMID: 17852564
32. **The biological role of death and lysis in biofilm development.** Bayles KW. Nat Rev Microbiol. 2007 Sep;5(9):721-6. PMID: 17694072

- The functional roles, phylogenetic distribution, and biological importance of the Murein Hydrolase Regulation and Cell Death subsystem, which was developed in conjunction with an NMPDR curator, are described.

33. **Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation.** Gupta N, Tanner S, Jaitly N, Adkins JN, Lipton M, Edwards R, Romine M, Osterman A, Bafna V, Smith RD, Pevzner PA. Genome Res. 2007 Sep;17(9):1362-77. [Epub 2007 Aug 9] PMID: 17690205
    - Subsystems were used to identify and categorize the functions of expressed proteins detected in the proteome of Shewanella oneidensis MR-1.
34. **Identification of genes encoding tRNA modification enzymes by comparative genomics.** de Crécy-Lagard V. Methods Enzymol. 2007;425:153-83. PMID: 17673083
35. **Comparative genomics of bacterial and plant folate synthesis and salvage: Predictions and validations.** de Crécy-Lagard V, El Yacoubi B, de la Garza RD, Noiriel A, Hanson AD. BMC Genomics. 2007 Jul 23;8:245. PMID: 17645794
    - Subsystem construction tools in the SEED and the SignatureGenesTool at NMPDR were used to predict the pathways and to identify cases of missing genes for almost every step of the Folate Biosynthesis subsystem. Candidates for such missing genes in bacteria and plants were then predicted using our compare regions view, and representative candidates were verified experimentally.
36. **Free methionine-(R)-sulfoxide reductase from *Escherichia coli* reveals a new GAF domain function.** Lin Z, Johnson LC, Weissbach H, Brot N, Lively MO, Lowther WT. Proc Natl Acad Sci U S A. 2007 Jun 5;104(23):9597-602. [Epub 2007 May 29]
    - The *yebR* gene of E.coli was proven to function as a free methionine-(R)-sulfoxide reductase, and the conservation of its genomic position adjacent to ProQ was explored using Compare Regions
37. **Characterization of a TIR-like protein from *Paracoccus denitrificans.*** Low LY, Mukasa T, Reed JC, Pascual J. Biochem Biophys Res Commun. 2007 May 4;356(2):481-6. [Epub 2007 Mar 7] PMID: 17362878
38. **Toward the automated generation of genome-scale metabolic networks in the SEED.** DeJongh M, Formsma K, Boillot P, Gould J, Rycenga M, Best A.BMC Bioinformatics. 2007 Apr 26;8:139. PMID: 17462086
    - Subsystems were used to guide the automated generation of substantially complete metabolic networks from a collection of modular components called "scenarios."
39. **The IclR-type transcriptional repressor LtbR regulates the expression of leucine and tryptophan biosynthesis genes in the amino acid producer *Corynebacterium glutamicum.*** Brune I, Jochmann N, Brinkrolf K, Huser AT, Gerstmeir R, Eikmanns BJ, Kalinowski J, Puhler A, Tauch A. J Bacteriol. 2007 Apr;189(7):2720-33. [Epub 2007 Jan 26] PMID: 17259312
    - The Cg1486 gene was proven to function as a repressor of genes in the leucine and tryptophan biosynthetic pathways, and the conservation of its

genomic position upstream of *leuCD* was explored using Compare Regions.

40. **Genomic identification and in vitro reconstitution of a complete biosynthetic pathway for the osmolyte di-myo-inositol-phosphate.** Rodionov DA, Kurnasov OV, Stec B, Wang Y, Roberts MF, Osterman AL. Proc Natl Acad Sci U S A. 2007 Mar 13;104(11):4279-84. [Epub 2007 Mar 2] PMID: 17360515
    - Comparative genomic analyses predicted two genes that were previously missing, which were included in a new subsystem that accurately describes the Di-Inositol-Phosphate biosynthesis pathway.
41. **Structure of the type III pantothenate kinase from *Bacillus anthracis* at 2.0 A resolution: Implications for coenzyme A-dependent redox biology.** Nicely NI, Parsonage D, Paige C, Newton GL, Fahey RC, Leonardi R, Jackowski S, Mallett TC, Claiborne A. Biochemistry. 2007 Mar 20;46(11):3234-45.[Epub 2007 Feb 27] PMID: 17323930
42. **Biotin uptake in prokaryotes by solute transporters with an optional ATP-binding cassette-containing module.** Hebbeln P, Rodionov DA, Alfandega A, Eitinger T. Proc Natl Acad Sci U S A. 2007 Feb 20;104(8):2909-14. [Epub 2007 Feb 14] PMID: 17301237
    - The ECF class transporters subsystem describes a group of bacterial and archaeal transporters containing typical ABC proteins that seem to be independent of solute-binding proteins.
43. **GISMO--gene identification using a support vector machine for ORF classification.** Krause L, McHardy AC, Nattkemper TW, Pühler A, Stoye J, Meyer F. Nucleic Acids Res. 2007 January; 35(2): 540–549. [Epub 2006 Dec 14] PMID: 17175534
    - ThiS in the Thiamin Biosynthesis subsystem was given as one example of how our subsystems were used to validate the automatic identification of very small genes.
44. **Computational reconstruction of iron- and manganese-responsive transcriptional networks in alpha-proteobacteria.** Rodionov DA, Gelfand MS, Todd JD, Curson AR, Johnston AW. PLoS Comput Biol. 2006 Dec 15;2(12):e163. [Epub 2006 Oct 18] PMID: 17173478
45. **Discovery of a new prokaryotic type I GTP cyclohydrolase family.** El Yacoubi B, Bonnett S, Anderson JN, Swairjo MA, Iwata-Reuyl D, de Crecy-Lagard V. J Biol Chem. 2006 Dec 8;281(49):37586-93. [Epub 2006 Oct 10] PMID: 17032654
    - The SignatureGenesTool was used in conjunction with Compare Regions and the Folate Biosynthesis subsystem to discover the function of COG1469 family proteins.
46. **Experimental and computational assessment of conditionally essential genes in *Escherichia coli.*** Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, Mori H, Lesely SA, Palsson BO, Agarwalla S. J Bacteriol. 2006 Dec;188(23):8259-71. [Epub 2006 Sep 29] PMID: 17012394
    - EssentialGenes of *E. coli* were analyzed in terms of metabolic subsystems across multiple genomes.
47. **The thioredoxin domain of *Neisseria gonorrhoeae* PilB can use electrons from DsbD to reduce downstream methionine sulfoxide reductases.** Brot N, Collet

JF, Johnson LC, Jonsson TJ, Weissbach H, Lowther WT. J Biol Chem. 2006 Oct 27;281(43):32668-75. [Epub 2006 Aug 22] PMID: 16926157
    - The domains that are fused in this Neisseria protein were identified as separate but clustered peptides using the Find Clusters function.
48. **Essential genes on metabolic maps.** Gerdes S, Edwards R, Kubal M, Fonstein M, Stevens R, Osterman A. Curr Opin Biotechnol. 2006 Oct;17(5):448-56. [Epub 2006 Sep 15] PMID: 16978855
    - The genomic scale screens for essential genes that are displayed on the EssentialGenes page are reviewed and discussed in the context of comparative analysis of subsystems.
49. **Comparative genomics and experimental characterization of N-acetylglucosamine utilization pathway of *Shewanella oneidensis.*** Yang C, Rodionov DA, Li X, Laikova ON, Gelfand MS, Zagnitko OP, Romine MF, Obraztsova AY, Nealson KH, Osterman AL. J Biol Chem. 2006 Oct 6;281(40):29872-85. [Epub 2006 Jul 20] PMID: 16857666
    - A novel variant of the classical three-step biochemical conversion of GlcNAc to fructose 6-phosphate was described in the Chitin and N-acetylglucosamine Utilization subsystem. The functional roles GlcN-6-P deaminase and GlcNAc kinase were assigned to two genes of previously unknown function based on Compare Regions and experimental verification.
50. **Characterization of the *Staphylococcus aureus* heat shock, cold shock, stringent, and SOS responses and their effects on log-phase mRNA turnover.** Anderson KL, Roberts C, Disz T, Vonstein V, Hwang K, Overbeek R, Olson PD, Projan SJ, Dunman PM. J Bacteriol. 2006 Oct;188(19):6739-56 PMID: 16980476
51. **Genome sequence of the bioplastic-producing "Knallgas" bacterium *Ralstonia eutropha H16.*** Pohlmann A, Fricke WF, Reinecke F, Kusian B, Liesegang H, Cramm R, Eitinger T, Ewering C, Pötter M, Schwartz E, Strittmatter A, Voss I, Gottschalk G, Steinbüchel A, Friedrich B, Bowien B. Nat Biotechnol. 2006 Oct;24(10):1257-62. [Epub 2006 Sep 10] PMID: 16964242
52. **Crystal structure of a type III pantothenate kinase: insight into the mechanism of an essential coenzyme A biosynthetic enzyme universally distributed in bacteria.** Yang K, Eyobo Y, Brand LA, Martynowski D, Tomchick D, Strauss E, Zhang H. J Bacteriol. 2006 Aug;188(15):5532-40. PMID: 16855243
    - A comprehensive survey of the phylogenetic distribution of type I, II and III PanKs in more than 300 complete or nearly complete genomes from the Archaea, Eukarya, and 13 major groups of Bacteria was performed.
53. **Random mutagenesis in *Corynebacterium glutamicum* ATCC 13032 using an IS6100-based transposon vector identified the last unknown gene in the histidine biosynthesis pathway.** Mormann S, Lömker A, Rückert C, Gaigalat L, Tauch A, Pühler A, Kalinowski J. BMC Genomics. 2006 Aug 10;7:205. PMID: 16901339
54. **Study of an alternate glyoxylate cycle for acetate assimilation by *Rhodobacter sphaeroides.*** Alber BE, Spanheimer R, Ebenau-Jehle C, Fuchs G. Mol Microbiol. 2006 Jul;61(2):297-309. PMID: 16856937

55. **Comparative genomics of NAD biosynthesis in cyanobacteria.** Gerdes SY, Kurnasov OV, Shatalin K, Polanuyer B, Sloutsky R, Vonstein V, Overbeek R, Osterman AL. J Bacteriol. 2006 Apr;188(8):3012-23. PMID: 16585762
56. **A hidden metabolic pathway exposed.** Osterman A. Proc Natl Acad Sci U S A. 2006 Apr 11;103(15):5637-8. [Epub 2006 Apr 4] PMID: 16595627
57. **Using pyrosequencing to shed light on deep mine microbial ecology.** Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander EC Jr, Rohwer F. BMC Genomics. 2006 Mar 20;7:57. PMID: 16549033
    - Subsystems were used to characterize two distinctly different microbial communities from short pyrosequence reads during the development of the MG-RAST server.
58. **An application of statistics to comparative metagenomics.** Rodriguez-Brito B, Rohwer F, Edwards RA. BMC Bioinformatics. 2006 Mar 20;7:162. PMID: 16549025
    - Subsystems that were overrepresented in the Sargasso Sea and Acid Mine Drainage metagenomes when compared to non-redundant databases were identified using statistical methods later incorporated into the MG-RAST server.
59. **Community genomics among stratified microbial assemblages in the ocean's interior.** DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM. Science. 2006 Jan 27;311(5760):496-503. PMID: 16439655
60. **Comparative and functional genomic analysis of prokaryotic nickel and cobalt uptake transporters: Evidence for a novel group of ATP-binding cassette transporters.** Rodionov DA, Hebbeln P, Gelfand MS, Eitinger T. J Bacteriol. 2006 Jan;188(1):317-27. PMID: 16352848
    - Positional gene clustering of identified candidate nickel/cobalt transporters with known Ni- and Co-containing enzymes was analyzed with Compare Regions and captured in the Transport of Nickel and Cobalt subsystem.
61. **Functional genomics and expression analysis of the *Corynebacterium glutamicum fpr2-cysIXHDNYZ* gene cluster involved in assimilatory sulphate reduction.** Ruckert C, Koch DJ, Rey DA, Albersmeier A, Mormann S, Puhler A, Kalinowski J. BMC Genomics. 2005 Sep 13;6:121. PMID: 16159395
    - The conservation of this cluster among the Actinomycetales was explored using Compare Regions.
62. **Low-molecular-weight protein tyrosine phosphatases of *Bacillus subtilis.*** Musumeci L, Bongiorni C, Tautz L, Edwards RA, Osterman A, Perego M, Mustelin T, Bottini N. J Bacteriol. 2005 Jul;187(14):4945-56. PMID: 15995210

Citing our work

63. **Strepto-DB, a database for comparative genomics of group A (GAS) and B (GBS) streptococci, implemented with the novel database platform 'Open Genome Resource' (OGeR).** Klein J, Münch R, Biegler I, Haddad I, Retter I, Jahn D. Nucleic Acids Res. 2008 Oct 14. [Epub ahead of print] PMID: 18854354

64. **MetaSim: A sequencing simulator for genomics and metagenomics.** Richter DC, Ott F, Auch AF, Schmid R, Huson DH. PLoS ONE. 2008 Oct 8;3(10):e3373. PMID: 18841204
65. **Incorporating genome-wide DNA sequence information into a dynamic whole-cell model of Escherichia coli: Application to DNA replication.** Atlas JC, Nikolaev EV, Browning ST, Shuler ML. IET Syst. Biol. 2008 Sept;2(5):369-82. DOI:10.1049/iet-syb:20070079
66. **Annotation of metagenome short reads using proxygenes.** Dalevi D, Ivanova NN, Mavromatis K, Hooper SD, Szeto E, Hugenholtz P, Kyrpides NC, Markowitz VM. Bioinformatics. 2008 Aug 15;24(16):i7-i13. PMID: 18689842
67. **Hyperbolic SOM-based clustering of DNA fragment features for taxonomic visualization and classification.** Martin C, Diaz NN, Ontrup J, Nattkemper TW. Bioinformatics. 2008 Jul 15;24(14):1568-74. [Epub 2008 Jun 5] PMID: 18535082
68. **GeConT 2: Gene context analysis for orthologous proteins, conserved domains and metabolic pathways.** Martinez-Guerrero CE, Ciria R, Abreu-Goodger C, Moreno-Hagelsieb G, Merino E. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W176-80. [Epub 2008 May 29] PMID: 18511460
69. **On application of directons to functional classification of genes in prokaryotes.** Wu H, Mao F, Olman V, Xu Y. Comput Biol Chem. 2008 Jun;32(3):176-84. [Epub 2008 Mar 2] PMID: 18440870
70. **Laying the foundation for a Genomic Rosetta Stone: Creating information hubs through the use of consensus identifiers.** Van Brabant B, Gray T, Verslyppe B, Kyrpides N, Dietrich K, Glöckner FO, Cole J, Farris R, Schriml LM, De Vos P, De Baets B, Field D, Dawyndt P; Genomic Standards Consortium. OMICS. 2008 Jun;12(2):123-7. PMID: 18479205
71. **An instant cell recognition system using a microfabricated coordinate standard chip useful for combinable cell observation with multiple microscopic apparatuses.** Yamada Y, Yamaguchi N, Ozaki M, Shinozaki Y, Saito M, Matsuoka H. Microsc Microanal. 2008 Jun;14(3):236-42. [Epub 2008 Mar 3] PMID: 18312725
72. **The minimum information about a genome sequence (MIGS) specification.** Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, DePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kravitz S, Kyrpides N, Leebens-Mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, Methe B, Mizrachi I, Moxon R, Nelson K, Parkhill J, Proctor L, White O, Sansone SA, Spiers A, Stevens R, Swift P, Taylor C, Tateno Y, Tett A, Turner S, Ussery D, Vaughan B, Ward N, Whetzel T, San Gil I, Wilson G, Wipat A. Nat Biotechnol. 2008 May;26(5):541-7. PMID: 18464787
73. **Genome-enabled approaches shed new light on plant metabolism.** DellaPenna D, Last RL. Science. 2008 Apr 25;320(5875):479-81. PMID: 18436775

74. **Large-scale prediction of drug-target relationships.** Kuhn M, Campillos M, González P, Jensen LJ, Bork P. FEBS Lett. 2008 Apr 9;582(8):1283-90. [Epub 2008 Feb 20] PMID: 18291108
75. **The multiple facets of homology and their use in comparative genomics to study the evolution of genes, genomes, and species.** Descorps-Declère S, Lemoine F, Sculo Q, Lespinet O, Labedan B. Biochimie. 2008 Apr;90(4):595-608. [Epub 2007 Sep 22] PMID: 17961904
76. **Comparative genomics-based investigation of resequencing targets in *Vibrio fischeri:* Focus on point miscalls and artefactual expansions.** Mandel MJ, Stabb EV, Ruby EG. BMC Genomics. 2008 Mar 25;9:138. PMID: 18366731
77. **Comparative genomics and functional annotation of bacterial transporters.** Gelfand MS, Rodionov DA. Physics of Life Reviews. 2008 March;5(1):22-49. [Epub 2007 Oct 24] doi:10.1016/j.plrev.2007.10.003
78. **Mining the genomes of plant pathogenic bacteria: How not to drown in gigabases of sequence.** Vinatzer BA, Yan S. Mol Plant Pathol. 2008 Jan;9(1):105-18. PMID: 18705888
79. **Finding novel metabolic genes through plant-prokaryote phylogenomics.** de Crécy-Lagard V, Hanson AD. Trends Microbiol. 2007 Dec;15(12):563-70. [Epub 2007 Nov 9] PMID: 17997099
80. **Annotation, comparison and databases for hundreds of bacterial genomes.** Médigue C, Moszer I. Res Microbiol. 2007 Dec;158(10):724-36. [Epub 2007 Oct 6] PMID: 18031997
81. **Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data.** Lemoine F, Lespinet O, Labedan B. BMC Evol Biol. 2007 Nov 29;7:237. PMID: 18047665
82. **Metabolic reconstruction and analysis for parasite genomes.** Pinney JW, Papp B, Hyland C, Wambua L, Westhead DR, McConkey GA. Trends Parasitol. 2007 Nov;23(11):548-54. [Epub 2007 Oct 22] PMID: 17950669
83. **GOing from functional genomics to biological significance.**!McCarthy FM, Bridges SM, Burgess SC. Cytogenet Genome Res. 2007;117(1-4):278-87. PMID: 17675869
84. **Current approaches to gene regulatory network modelling.** Schlitt T, Brazma A. BMC Bioinformatics. 2007 Sep 27;8 Suppl 6:S9. PMID: 17903290
85. **Genome-wide analysis of intergenic regions of *Mycobacterium tuberculosis* H37Rv using affymetrix GeneChips .** Fu LM, Shinnick TM. EURASIP J Bioinform Syst Biol. 2007 September 16; 2007:23054-10. PMID: 18253472
86. **Sequence-based analysis of pQBR103; a representative of a unique, transfer-proficient mega plasmid resident in the microbial community of sugar beet.** Tett A, Spiers AJ, Crossman LC, Ager D, Ciric L, Dow JM, Fry JC, Harris D, Lilley A, Oliver A, Parkhill J, Quail MA, Rainey PB, Saunders NJ, Seeger K, Snyder LA, Squares R, Thomas CM, Turner SL, Zhang XX, Field D, Bailey MJ. ISME J. 2007 Aug;1(4):331-40. [Epub 2007 Jul 5] PMID: 18043644
87. **Sensitivity and control analysis of periodically forced reaction networks using the Green's function method.** Nikolaev EV, Atlas JC, Shuler ML. J Theor Biol. 2007 Aug 7;247(3):442-61. [Epub 2007 Feb 28] PMID: 17481665

88. **Comparative genomic reconstruction of transcriptional regulatory networks in bacteria.** Rodionov DA. Chem Rev. 2007 Aug;107(8):3467-97. [Epub 2007 Jul 18] PMID: 17636889
89. **Protein annotation at genomic scale: The current status.** Frishman D. Chem Rev. 2007 Aug;107(8):3448-66. [Epub 2007 Jul 21] PMID: 17658902
90. **Mining enzymes from extreme environments.** Ferrer M, Golyshina O, Beloqui A, Golyshin PN. Curr Opin Microbiol. 2007 Jun;10(3):207-14. [Epub 2007 Jun 4] PMID: 17548239
91. **Microbial genome data resources.** Markowitz VM. Curr Opin Biotechnol. 2007 Jun;18(3):267-72. [Epub 2007 Apr 30] PMID: 17467973
92. **The pan-genome: Towards a knowledge-based discovery of novel targets for vaccines and antibacterials.** Muzzi A, Masignani V, Rappuoli R. Drug Discov Today. 2007 Jun;12(11-12):429-39. [Epub 2007 May 7] PMID: 17532526
93. **The positive role of the ecological community in the genomic revolution.** Field D, Kyrpides N. Microb Ecol. 2007 Apr;53(3):507-11. [Epub 2007 Apr 12] PMID: 17436031
94. **Accurate phylogenetic classification of variable-length DNA fragments.** McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Nat Methods. 2007 Jan;4(1):63-72. [Epub 2006 Dec 10] PMID: 17179938
95. **CutDB: A proteolytic event database.** Igarashi Y, Eroshkin A, Gramatikova S, Gramatikoff K, Zhang Y, Smith JW, Osterman AL, Godzik A. Nucleic Acids Res. 2007 Jan;35(Database issue):D546-9. [Epub 2006 Nov 16] PMID: 17142225
96. **How do we compare hundreds of bacterial genomes?** Field D, Wilson G, van der Gast C. Curr Opin Microbiol. 2006 Oct;9(5):499-504. [Epub 2006 Aug 30] PMID: 16942900
97. **Automated bacterial genome analysis and annotation.** Stothard P, Wishart DS. Curr Opin Microbiol. 2006 Oct;9(5):505-10. [Epub 2006 Aug 22] PMID: 16931121
98. **New metrics for comparative genomics.** Galperin MY, Kolker E. Curr Opin Biotechnol. 2006 Oct;17(5):440-7. [Epub 2006 Sep 15] PMID: 16978854
99. **Systems biology as a foundation for genome-scale synthetic biology** Barrett CL, Kim TY, Kim HU, Palsson BØ, Lee SY. Curr Opin Biotechnol. 2006 Oct;17(5):488-92. [Epub 2006 Aug 23] PMID: 16934450
100. **AgBase: A functional genomics resource for agriculture.** McCarthy FM, Wang N, Magee GB, Nanduri B, Lawrence ML, Camon EB, Barrell DG, Hill DP, Dolan ME, Williams WP, Luthe DS, Bridges SM, Burgess SC. BMC Genomics. 2006 Sep 8;7:229. PMID: 16961921
101. **A phosphatidic acid-binding protein of the chloroplast inner envelope membrane involved in lipid trafficking.** Awai K, Xu C, Tamot B, Benning C. Proc Natl Acad Sci U S A. 2006 Jul 11;103(28):10817-22. [Epub 2006 Jul 3] PMID: 16818883
102. **MaGe: A microbial genome annotation system supported by synteny results.** Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S, Lajus A, Pascal G, Scarpelli C, Médigue C. Nucleic Acids Res. 2006 Jan 10;34(1):53-65. PMID: 16407324

103. **Automatic detection of subsystem/pathway variants in genome analysis.** Ye Y, Osterman A, Overbeek R, Godzik A. Bioinformatics. 2005 Jun;21 Suppl 1:i478-86. PMID: 15961494