

Bioinformatic tools for metagenomic data analysis

MEGAN

- blast-based tool for exploring taxonomic content

MG-RAST (SEED, FIG)

- rapid annotation of metagenomic data, phylogenetic classification and metabolic reconstruction

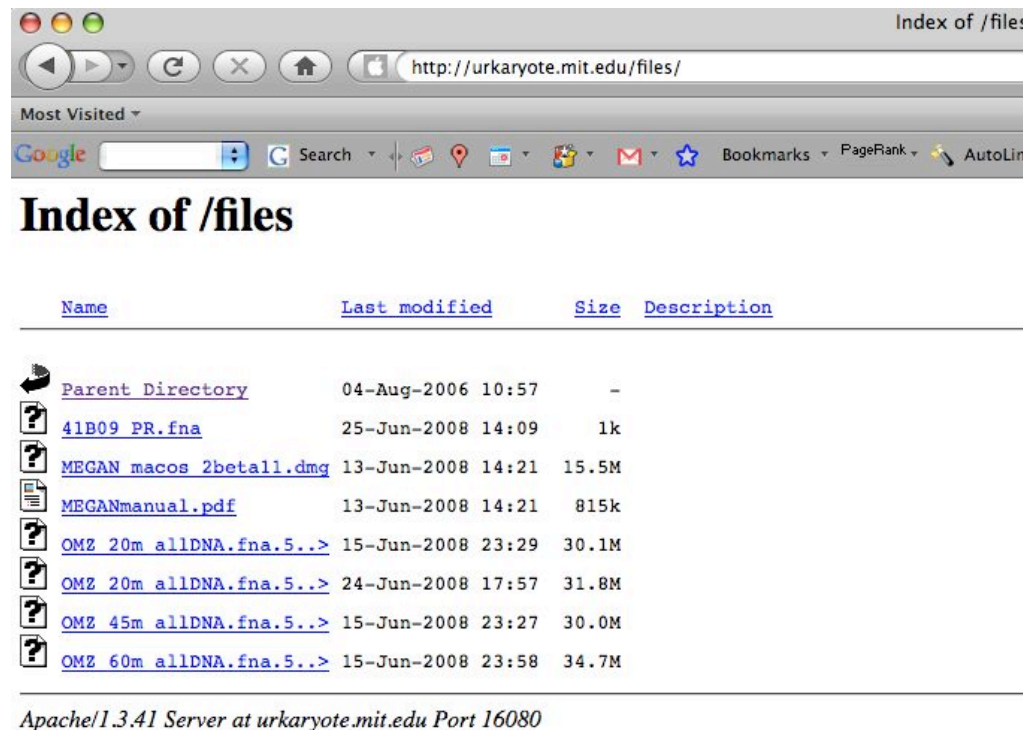
CAMERA (JCVI, Calit2, UCSD)

- metagenomic data repository and blast server

Getting started

<http://urkaryote.mit.edu/files/>

Download MEGAN and metagenomic datasets BLAST files:



The screenshot shows a web browser window with the address bar displaying <http://urkaryote.mit.edu/files/>. The page title is "Index of /files". Below the browser window, the "Index of /files" page is displayed, featuring a table with columns for Name, Last modified, Size, and Description. The table lists several files, including a Parent Directory, 41B09 PR.fna, MEGAN macos 2beta11.dmg, MEGANmanual.pdf, and several OMZ allDNA.fna.5..> files. The footer of the page reads "Apache/1.3.41 Server at urkaryote.mit.edu Port 16080".

Name	Last modified	Size	Description
Parent Directory	04-Aug-2006 10:57	-	
41B09 PR.fna	25-Jun-2008 14:09	1k	
MEGAN macos 2beta11.dmg	13-Jun-2008 14:21	15.5M	
MEGANmanual.pdf	13-Jun-2008 14:21	815k	
OMZ 20m allDNA.fna.5..>	15-Jun-2008 23:29	30.1M	
OMZ 20m allDNA.fna.5..>	24-Jun-2008 17:57	31.8M	
OMZ 45m allDNA.fna.5..>	15-Jun-2008 23:27	30.0M	
OMZ 60m allDNA.fna.5..>	15-Jun-2008 23:58	34.7M	

Apache/1.3.41 Server at urkaryote.mit.edu Port 16080

MEGAN (Metagenome analyzer)

D.H. Huson et al., MEGAN analysis of metagenome data. **Genome Research**, 17:377-386 (2007).

<http://www-ab.informatik.uni-tuebingen.de/software/megan/welcome.html>

Software designed to analyze metagenomic datasets and
assign reads to a known taxon.

What organisms are present and what are their relative proportions?

What types of types of genes are present?

- Input files:
- BLAST files; BLASTX, BLASTN or BLASTZ
 - NCBI-NR, NCBI-NT, NCBI-ENV-NR, NCBI-ENV-NT or genome specific databases

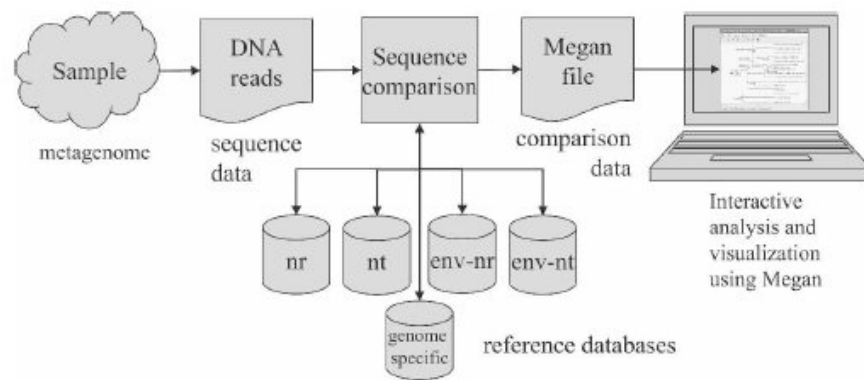


Image from Huson et al., 2007



Taxonomy Nodes (all dates)

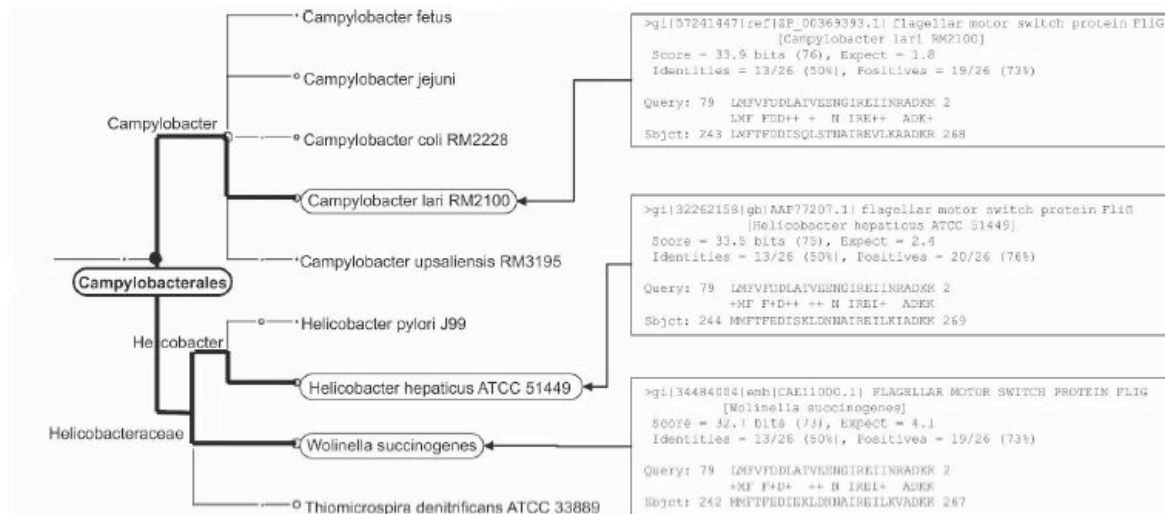
Ranks:	higher taxa	genus	species	lower taxa	total
All taxa	18384	45663	327292	48065	439404
Archaea	193	107	3368	99	3767
Bacteria	2378	1824	102732	4453	111387
Eukaryota	15345	43449	211648	12403	282845
Viruses	444	276	4866	31080	36666
Dates: 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 all dates					

MEGAN uses NCBI taxonomy to summarize and order results.

<http://www.ncbi.nlm.nih.gov/Taxonomy/Selector/taxse.cgi>

MEGAN (Metagenomics analyzer)

Lowest common ancestor (LCA) assignment algorithm:

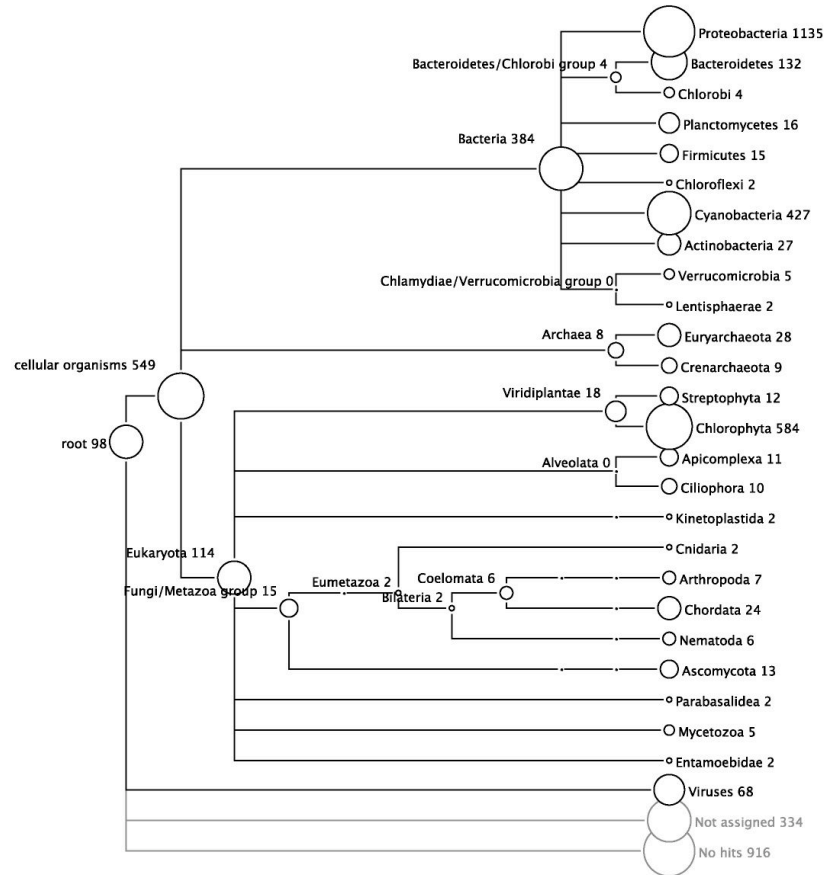


Filters set by user:

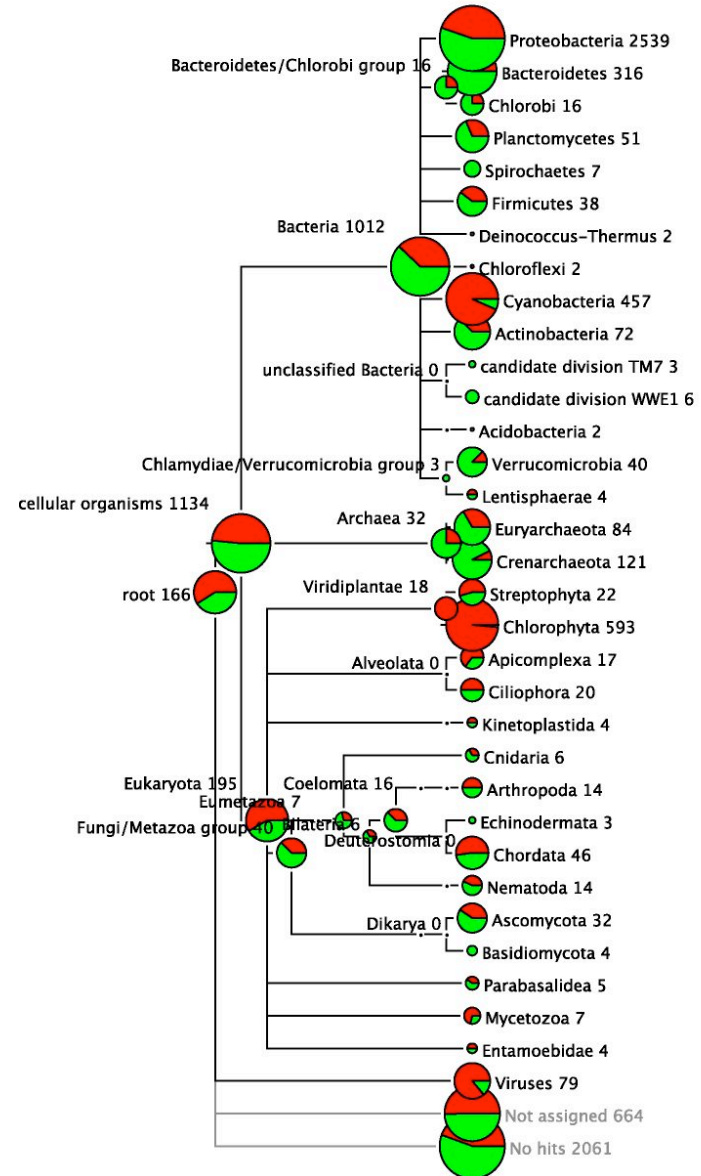
- Min. Support for Taxa (default = 2)
- Min. Score (default = 30.0)
- Top Percentage (default = 10.0)
- Win ("Winner") Score (default = 0)

Comparison of metagenomic samples

Single metagenomic sample



■ OMZ_20m_allDNA.fna.5000.keepVSnr_20080515.FF.b10.v10.c250
■ OMZ_45m_allDNA.fna.5000.keepVSnr_20080515.FF.b10.v10.c250



MEGAN (Metagenomics analyzer)

Let's start playing with MEGAN!

MEGAN on startup screen:

- will take a minute or so to load NCBI tree and taxonomy

The screenshot displays the MEGAN software interface on a Mac OS desktop. The main window, titled "Untitled - MEGAN (version 2beta11, built 5 Jun 2008)", shows a hierarchical tree structure representing the NCBI taxonomy. The tree is rooted at "root" and branches into several categories: "cellular organisms", "other sequences", "Viruses", "Viroids", and "unclassified sequences". Each category lists specific taxonomic groups, such as "Bacteria", "Archaea", "Eukaryota", "transposons", "plasmids", "artificial sequences", "insertion sequences", "midvariant sequence", "dsDNA viruses, no RNA stage", "Retro-transcribing viruses", "Deltavirus", "ssDNA viruses", "dsRNA viruses", "ssRNA viruses", "unclassified viruses", "unclassified phages", "Satellites", "unclassified archaeal viruses", "environmental samples", "Pospiviroidae", "Avsunviroidae", "unclassified viroids", "unidentified", "hybrid", "environmental samples", "putative agent of rhinosporidiosis", "metagenomes", and "unidentified Cotton leaf curl Rajasthan virus-associated DNA".

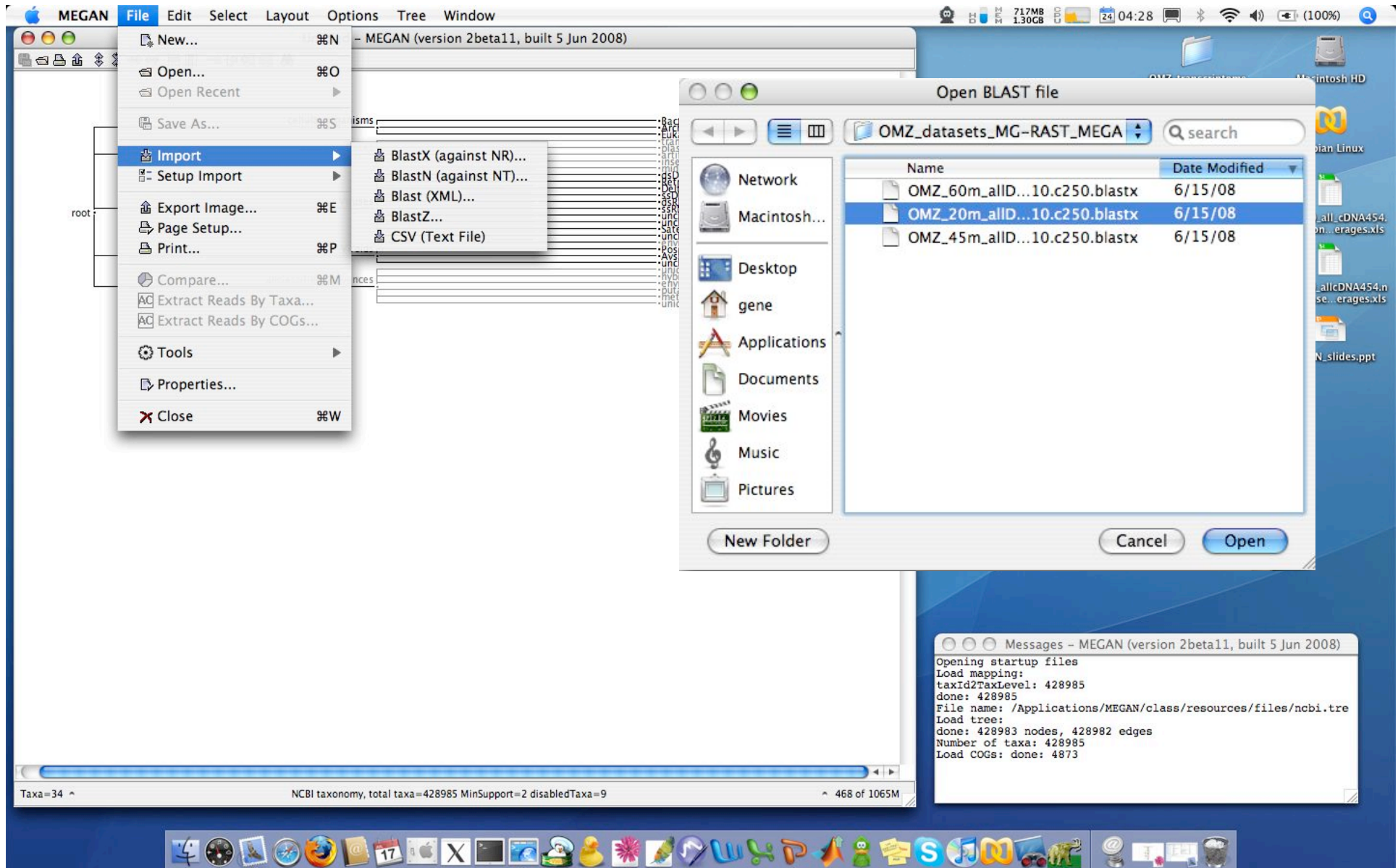
Below the tree, a status bar indicates "Taxa=34" and "NCBI taxonomy, total taxa=428985 MinSupport=2 disabledTaxa=9". A "Messages" window in the bottom right corner provides details about the startup process:

```
Opening startup files
Load mapping:
taxId2TaxLevel: 428985
done: 428985
File name: /Applications/MEGAN/class/resources/files/ncbi.tre
Load tree:
done: 428983 nodes, 428982 edges
Number of taxa: 428985
Load COGs: done: 4873
```

The desktop background is blue with various icons, including "OMZ_transcriptome", "Macintosh HD", "ESOM", "Debian Linux", and several files related to "HOT179_all_cDNA454". The dock at the bottom contains various application icons.

Step 1: Importing a BLAST file:

- File→Import. Choose import file type (e.g. BLASTX, BLASTN)



Step 2: Choosing MEGAN output format:

- Full Dataset OR Summary (choose Full Dataset - greater control over filters)

The screenshot shows the MEGAN software interface (version 2beta11, built 5 Jun 2008) on a Mac OS X desktop. The main window displays a taxonomic tree with categories like 'cellular organisms', 'other sequences', 'Viruses', and 'unclassified sequences'. A 'Choose' dialog box is open, asking 'Generate summary or full dataset?'. The 'Full Dataset' option is selected with a checkmark, and the 'Summary' option is also visible. The dialog has 'Cancel' and 'OK' buttons. Below the dialog, a red-bordered box contains the following text:

- If **full dataset** is requested, then the parser generates a large file that contains all necessary information for interactive exploration of the given data set.
- If **summary** is requested, then the parser generates a taxonomical summary of the dataset on the fly while parsing the data. If this option is used, then a **Parameters** Dialog is presented to provide control over the parameters used by the LCA-assignment algorithm to generate the summary.

The bottom status bar shows 'Taxa=34', 'NCBI taxonomy, total taxa=428985 MinSupport=2 disabledTaxa=9', and '468 of 1065M'. A 'Messages' window in the bottom right corner displays the following log:

```
Messages - MEGAN (version 2beta11, built 5 Jun 2008)
Opening startup files
Load mapping:
taxId2TaxLevel: 428985
done: 428985
File name: /Applications/MEGAN/class/resources/files/ncbi.tre
Load tree:
done: 428983 nodes, 428982 edges
Number of taxa: 428985
Load COGs: done: 4873
```

Step 3: Importing and parsing the BLAST file:

- Note: some taxa will remain unidentified!

The screenshot displays the MEGAN (version 2beta11, built 5 Jun 2008) application window. The main interface shows a taxonomic tree with categories: cellular organisms, other sequences, and Viruses. A progress dialog box titled "Importing BlastX:" is visible, showing the file "OMZ_20m_allDNA.fna.5000.keepVSnr_20080515.FF.b10". A message dialog box with an exclamation mark icon states: "Unable to identify taxa for 411 matches (out of 34270)".

Summary of blast results

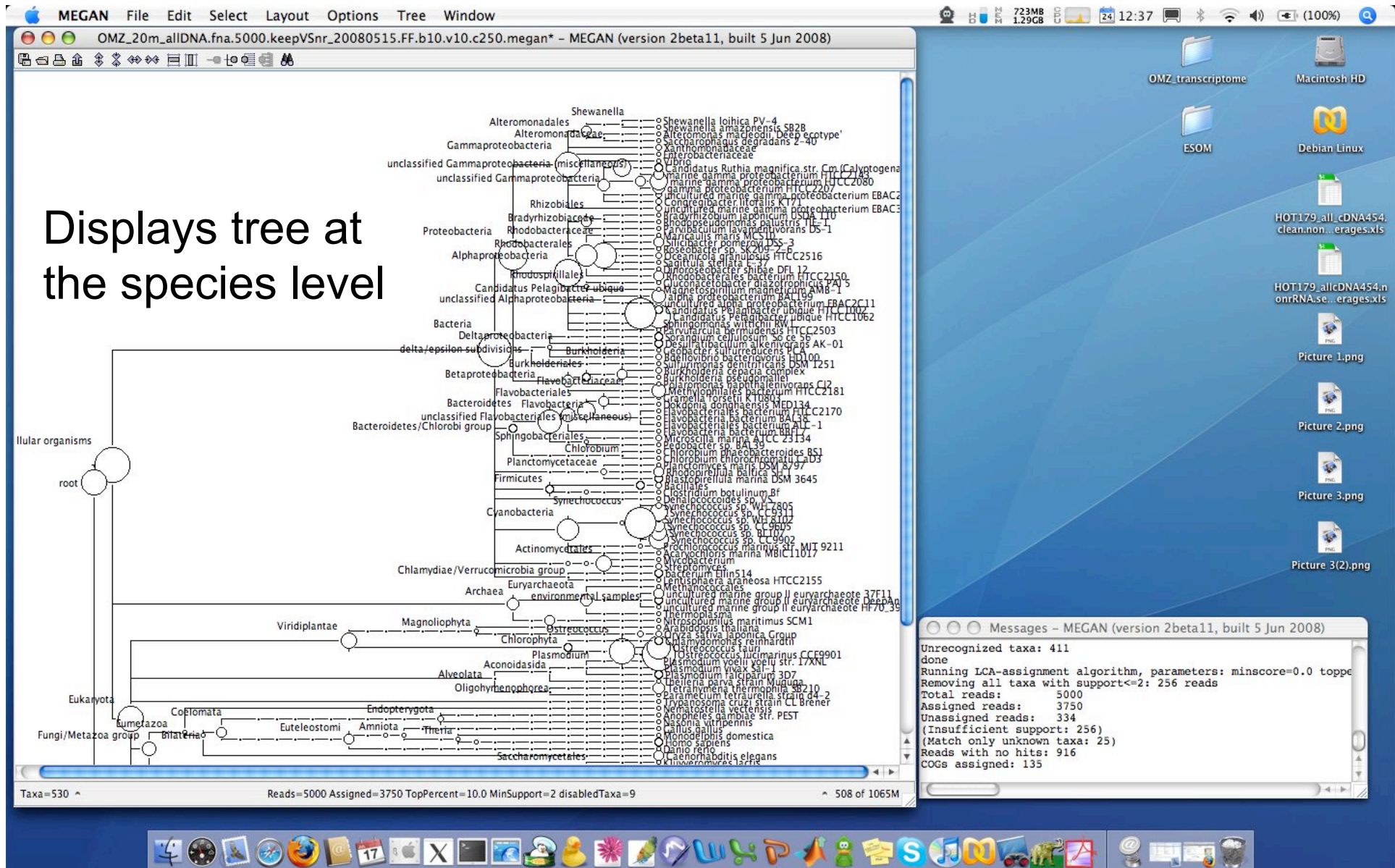
Messages - MEGAN (version 2beta11, built 5 Jun 2008)

```
Line 668051: unidentified taxon: 'acceptor'
Line 756264: unidentified taxon: '(5-phosphoribosylamino)methyl
#30000
Line 808058: unidentified taxon: 'Bacteriophage RB49'
Line 808109: unidentified taxon: 'Bacteriophage KVP20'
Line 890870: unidentified taxon: 'swine, fetal liver, Peptide Pe
Total matches: 34270
Reads with hits: 4084
Reads with no hits: 916
Unrecognized taxa: 411
```


Step 4: Generating tree showing taxonomic distribution of reads:

- Note: NCBI tree replaced by taxonomy of assigned reads

Displays tree at the species level



Step 5: Setting filtering criteria

- Options → Min Support, Min Score, Top Percentage and Win Score

The screenshot displays the MEGAN (version 2beta11, built 5 Jun 2008) software interface. The main window shows a phylogenetic tree with various taxa labeled, including Shewanella, Burkholderia, and Eukaryota. The 'Options' menu is open, showing settings for Min Support For Taxa (2), Min Score (40.03), Top Percentage (10.0), and Win Score (0.0). Three input dialog boxes are overlaid on the right side of the screen, each showing a specific setting: 'Set min support' (2), 'Set min score' (40.03), and 'Set percentage of best score' (10.0). The status bar at the bottom indicates 'Taxa=530', 'Reads=5000', 'Assigned=3750', 'TopPercent=10.0', 'MinSupport=2', 'disabledTaxa=9', and '537 of 1065M'.

MEGAN File Edit Select Layout Options Tree Window

OMZ_20m_allDNA.fna.5000.keepVSr - MEGAN (version 2beta11, built 5 Jun 2008)

Options

- Min Support For Taxa...
- Min Score...
- Top Percentage...
- Win Score...
- Taxon Disabling
- Set Number Of Reads...
- List Summary
- List Assignments
- Add Taxa To Chart
- Remove Taxa From Chart
- Inspect...

Input

Set min support

2

Cancel OK

Set min score

40.03 (23004 of 31502)

Cancel Apply

Input

Set percentage of best score

10.0

Cancel OK

Input

Set win score

0.0

Cancel OK

Total reads: 5000
Assigned reads: 3750
Unassigned reads: 334
(Insufficient support: 256)
(Match only unknown taxa: 25)
Reads with no hits: 916
COGs assigned: 135

Taxa=530 Reads=5000 Assigned=3750 TopPercent=10.0 MinSupport=2 disabledTaxa=9 537 of 1065M

Step 6: Show the number of reads assigned to each taxa/node

- Tree → Show number of Reads Assigned

The screenshot displays the MEGAN (version 2beta11, built 5 Jun 2008) software interface. The main window shows a taxonomic tree with read counts assigned to each node. The tree is rooted at 'cellular organisms' and branches into various taxonomic groups, including Bacteroidetes/Chlorobi group, Firmicutes, Cyanobacteria, Actinomycetales, Archaea, Euryarchaeota, Eukaryota, and Fungi/Metazoa group. The tree is labeled with 'Taxa=338' and 'Reads=5000 Assigned=2912 MinScore=40.0 TopPercent=10.0 MinSupport=2 disabledTaxa=9'. A context menu is open over the tree, showing options like 'Collapse', 'Uncollapse', 'Show Taxon Names', 'Show Taxon Ids', 'Show Number of Reads Assigned' (which is selected), 'Display the number of reads assigned to a taxon', 'Node Labels On', 'Node Labels Off', and 'Show Intermediate Labels'. A messages window in the bottom right corner displays the following information:

```
Messages - MEGAN (version 2beta11, built 5 Jun 2008)
Removing all taxa with support<=2: 179 reads
Total reads: 5000
Assigned reads: 2912
Unassigned reads: 1172
(Insufficient support: 179)
(Match only unknown taxa: 25)
Reads with no hits: 916
COGs assigned: 104
executing nodelabels names=true ids=false assigned=true sum
executing nodelabels names=true ids=false assigned=false sum
```

Step 7: Collapsing/expanding the tree to different taxonomic levels

-Tree → Collapse at Taxonomical Level

- choose higher level to collapse, or lower level to expand

The screenshot displays the MEGAN 2 software interface. The main window shows a taxonomic tree with various levels of collapse and expansion. A context menu is open, showing options for collapsing and expanding the tree at different taxonomic levels. The menu options include:

- Kingdom
- Phylum
- Class
- Order
- Family
- Varietas
- Genus
- Species group
- Subspecies
- Species

The tree itself shows various taxonomic groups, including Bacteroidetes/Chlorobi group 3, Bacteroidetes (miscellaneous), unclassified, Sphingobacteriales, Planctomycetaceae, Firmicutes, Clostridia, Cyanobacteria, Proteobacteria, Actinomycetales, Euryarchaeota, Archaea, Eukaryota, Fungi/Metazoa group, and dsDNA viruses, no RNA stage.

A summary window titled "Messages - MEGAN (version 2beta11, built 5 Jun 2008)" is open in the bottom right corner, displaying the following statistics:

```
Total reads: 5000
Assigned reads: 2912
Unassigned reads: 1172
(Insufficient support: 179)
(Match only unknown taxa: 25)
Reads with no hits: 916
COGs assigned: 104
executing nodelabels names=true ids=false assigned=true sum
executing nodelabels names=true ids=false assigned=false sum
executing nodelabels names=true ids=false assigned=true sum
```

The status bar at the bottom of the MEGAN window shows: Taxa=338, Reads=5000, Assigned=2912, MinScore=40.0, TopPercent=10.0, MinSupport=2, disabledTaxa=9, 513 of 1065M.

Step 8: Changing the vertical and horizontal view of the tree

The screenshot displays the MEGAN (version 2beta11, built 5 Jun 2008) software interface. The main window shows a phylogenetic tree with various taxonomic groups. A toolbar at the top contains icons for file operations and view controls. A red box highlights the view control icons, which include a vertical double-headed arrow, a horizontal double-headed arrow, and a square icon. Arrows point from these icons to labels: 'Expand vertically', 'Contract vertically', 'Expand horizontally', and 'Contract horizontally'. The status bar at the bottom indicates 'Taxa=49', 'Reads=5000', 'Assigned=3750', 'TopPercent=10.0', 'MinSupport=2', 'disabledTaxa=9', and '1004 of 1065M'.

MEGAN File Edit Select Layout Options Tree Window

OMZ_20m_allDNA.fna.5000.keepVSnr_20080515.FF.b10.v10.c250.megan* - MEGAN (version 2beta11, built 5 Jun 2008)

Expand vertically

Contract vertically

Expand horizontally

Contract horizontally

cellular organisms 549

root 98

Eukaryota 114

Fungi/Metazoa group 15

Viridiplantae 18

Alveolata 0

Chlorophyta 584

Apicomplexa 11

Ciliophora 10

Kinetoplastida 2

Cnidaria 2

Eumetazoa 2

Coelomata 6

Arthropoda 7

Chordata 24

Nematoda 6

Ascomycota 13

Parabasalidea 2

Mycetozoa 5

Entamoebidae 2

Viruses 68

Not assigned 334

No hits 916

Proteobacteria 1135

Bacteroidetes/Chlorobi group 4

Bacteroidetes 132

Chlorobi 4

Planctomycetes 16

Firmicutes 15

Chloroflexi 2

Cyanobacteria 427

Actinobacteria 27

Verrucomicrobia group 0

Verrucomicrobia 5

Lentisphaerae 2

Archaea 8

Euryarchaeota 28

Crenarchaeota 9

Assigned reads: 7275

Unassigned reads: 664

Reads with no hits: 2061

Executing collapse level='Phylum';

Executing node labels names=true ids=false assigned=true summarized=false

Executing collapse level='Class';

Executing collapse level='Order';

Executing collapse level='Phylum';

Written to file: /Users/gene/Desktop/Untitled-cmp.eps

Executing node labels names=true ids=false assigned=true summarized=false

Executing collapse level='Phylum';

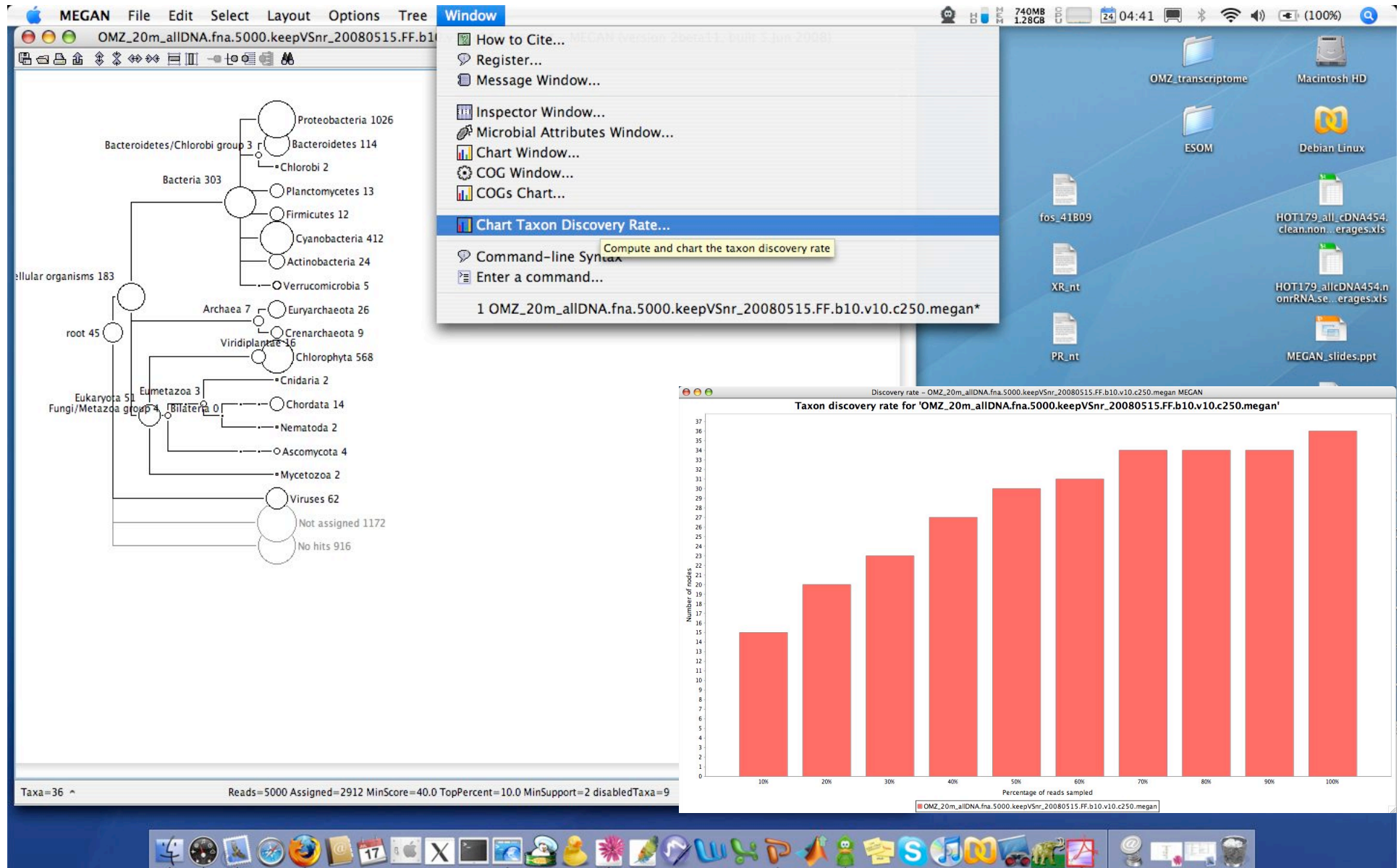
Executing uncollapse all;

Executing collapse level='Phylum';

Taxa=49 Reads=5000 Assigned=3750 TopPercent=10.0 MinSupport=2 disabledTaxa=9 1004 of 1065M

Step 9: Taxon discovery rate

- Window→Chart Taxon Discovery Rate



Step 9: Exporting and saving MEGAN output

- File→Save As
- File→Export Image

The screenshot displays the MEGAN (version 2beta11, built 5 Jun 2008) application window. The 'File' menu is open, with 'Save As...' and 'Export Image...' highlighted by red boxes. The main window shows a phylogenetic tree with various taxonomic groups and their associated read counts. The status bar at the bottom indicates 'Taxa=36', 'Reads=5000', 'Assigned=2912', 'MinScore=40.0', 'TopPercent=10.0', 'MinSupport=2', 'disabledTaxa=9', and '550 of 1065M'.

Phylogenetic Tree Data:

Group	Count
Proteobacteria	1026
Bacteroidetes/Chlorobi group 3	3
Bacteroidetes	114
Chlorobi	2
Bacteria	303
Planctomycetes	13
Firmicutes	12
Cyanobacteria	412
Actinobacteria	24
Verrucomicrobia	5
Archaea	7
Euryarchaeota	26
Crenarchaeota	9
Viridiplantae	16
Chlorophyta	568
Cnidaria	2
Chordata	14
Nematoda	2
Ascomycota	4
Mycetozoa	2
Viruses	62
Not assigned	1172
No hits	916

File Menu Options:

- New... (%N)
- Open... (%O)
- Open Recent
- Save As... (%S)
- Import
- Setup Import
- Export Image... (%E)
- Page Setup...
- Print... (%P)
- Compare... (%M)
- Extract Reads By Taxa...
- Extract Reads By COGs...
- Tools
- Properties...
- Close (%W)

Status Bar:

Taxa=36 Reads=5000 Assigned=2912 MinScore=40.0 TopPercent=10.0 MinSupport=2 disabledTaxa=9 550 of 1065M

Messages - MEGAN (version 2beta11, built 5 Jun 2008)

```
40 27
50 30
60 31
70 34
80 34
90 34
100 36
executing chart cogs;
executing select intermediate;
executing select level='Phylum';
```

Exercise: Comparing metagenomic samples using MEGAN

- 1) Import the BLASTX against NR file for the OMZ 20m dataset.
Note: this should be done already
(File name: OMZ_20m_allDNA.fna.5000VSnr.FF.b10.v10.c250.blastx)
- 2) Import the BLASTX against NR files for the OMZ 45m OR 60m dataset.
(File name: OMZ_45m_allDNA.fna.5000VSnr.FF.b10.v10.c250.blastx OR
OMZ_60m_allDNA.fna.5000VSnr.FF.b10.v10.c250.blastx)
- 3) Change the Min Score filter to a bitscore of 40 for each dataset.
- 4) Using the Compare tool (found under File--->Compare), select both datasets and choose the normalize over all reads option.
- 5) Collapse the resulting tree to the phylum level. Note any major differences in the representation of difference taxonomic groups.

MG-RAST

MetaGenome Rapid Annotation using Subsystems Technology

<http://metagenomics.nmpdr.org/mg-rast/FIG/rast.cgi>


Login: CMORE

Password: CMORE

- Accurate and consistent annotations of metagenomic data (12-24 hours)
- Currently accepts 454 and Sanger sequences
- Automatic metabolic reconstruction
- Phylogenetic classification of rRNAs and proteins
- Comparison tools

MG-RAST

Subsystem – a generalization of “pathway” as a collection of functional roles jointly involved in a biological process or complex



MG-RAST Meta Genome Rapid Annotation
using Subsystem Technology version 2.0

The NMPDR, SEED-based, metagenome annotation service.
For more information about The SEED please visit theSEED.org.

Home Your Jobs CMORE class

Upload a metagenome

The upload page will guide you through the upload process.
The *Current Step* tab is used to enter all necessary information step by step.
The *Upload Summary* tab provides an overview of the current upload.

Upload Summary

To upload sequence data to the metagenomics RAST server:

(1) You can upload a fasta file containing just the nucleotide sequences.
In this case the file name should end in .fa, .fasta, .fas, .fsa or .fna.
If your sequence file is larger than 30 MB please use the compressed format (below) or contact us for other options.

(2) You can compress the sequence file containing just the nucleotide sequences with tar and gzip a popular compression tool.
In this case the compressed file name should end in .tgz and the fasta file should end in .fa, .fasta, .fas, .fsa or .fna.

(3) Optionally, you can include a quality file along with the sequence file in a single compressed file. To do this, compress both files into a single archive and then upload the archive file.
In this case the sequence file name should end in .fa, .fasta, .fas, .fsa or .fa, the quality file name should end in .qual, and the archive name should end in .tgz.

File Upload:

Sequences File

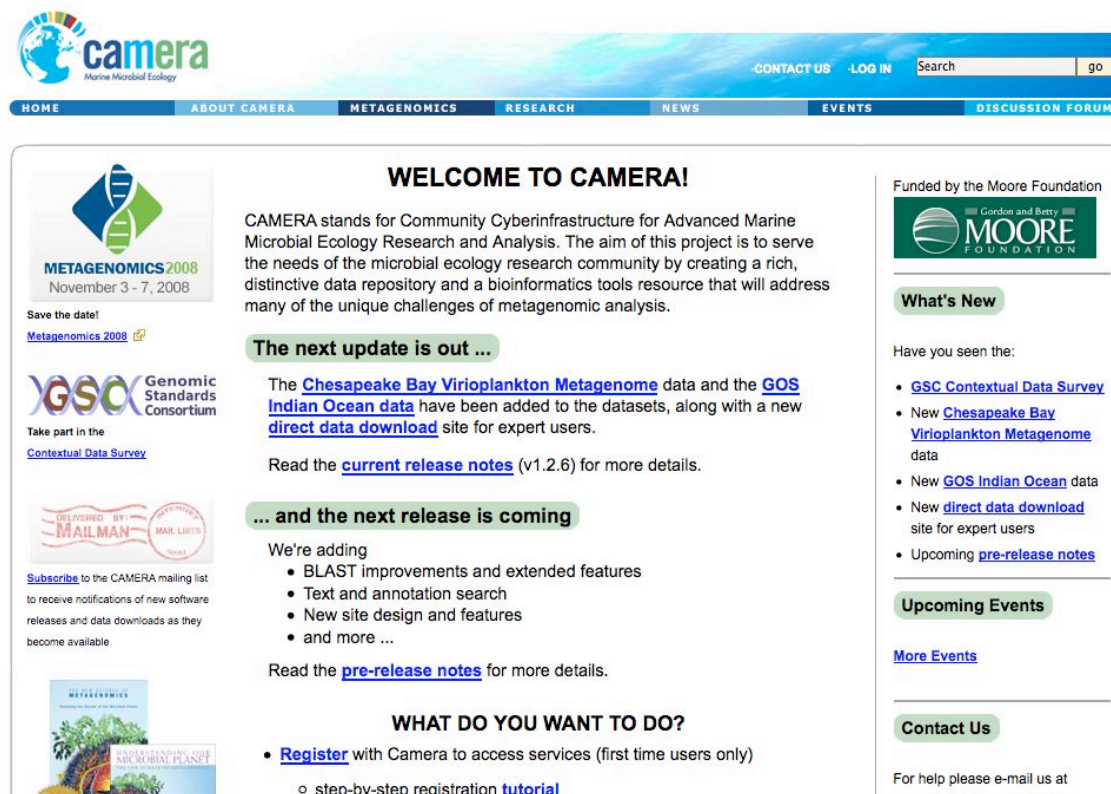
- * Submit fasta nucleotide sequences
- * Project name and description
- * Metadata (location, depth, altitude, time of collection, habitat)

CAMERA

Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis

“The aim of this project is to serve the needs of the microbial ecology research community by creating a rich, distinctive data repository and a bioinformatics tools resource that will address many of the unique challenges of metagenomic analysis.”

<http://camera.calit2.net/index.php>



The screenshot shows the CAMERA website homepage. At the top is a navigation bar with links: HOME, ABOUT CAMERA, METAGENOMICS, RESEARCH, NEWS, EVENTS, and DISCUSSION FORUMS. A search bar is also present. The main content area is titled "WELCOME TO CAMERA!" and describes the project's goal: to serve the microbial ecology research community by creating a rich, distinctive data repository and a bioinformatics tools resource. It mentions the addition of the Chesapeake Bay Virioplankton Metagenome data and the GOS Indian Ocean data, along with a new direct data download site for expert users. A section titled "The next update is out ..." lists updates: BLAST improvements, text and annotation search, new site design, and more. Another section titled "... and the next release is coming" lists upcoming features: BLAST improvements, text and annotation search, new site design, and more. A section titled "WHAT DO YOU WANT TO DO?" lists options: Register with Camera to access services (first time users only) and step-by-step registration tutorial. The right sidebar features the Moore Foundation logo and a "What's New" section listing updates: GSC Contextual Data Survey, New Chesapeake Bay Virioplankton Metagenome data, New GOS Indian Ocean data, New direct data download site for expert users, and Upcoming pre-release notes. There is also an "Upcoming Events" section and a "Contact Us" section.

WELCOME TO CAMERA!

CAMERA stands for Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis. The aim of this project is to serve the needs of the microbial ecology research community by creating a rich, distinctive data repository and a bioinformatics tools resource that will address many of the unique challenges of metagenomic analysis.

The next update is out ...

The [Chesapeake Bay Virioplankton Metagenome](#) data and the [GOS Indian Ocean data](#) have been added to the datasets, along with a new [direct data download](#) site for expert users.

Read the [current release notes](#) (v1.2.6) for more details.

... and the next release is coming

We're adding

- BLAST improvements and extended features
- Text and annotation search
- New site design and features
- and more ...

Read the [pre-release notes](#) for more details.

WHAT DO YOU WANT TO DO?

- [Register](#) with Camera to access services (first time users only)
 - step-by-step registration [tutorial](#)

What's New

Have you seen the:

- [GSC Contextual Data Survey](#)
- New [Chesapeake Bay Virioplankton Metagenome](#) data
- New [GOS Indian Ocean](#) data
- New [direct data download](#) site for expert users
- Upcoming [pre-release notes](#)

Upcoming Events

[More Events](#)

Contact Us

For help please e-mail us at [camera.help@calit2.net](#)

CAMERA

Community Cyberinfrastructure for Advanced Marine Microbial
Ecology Research and Analysis

CAMERA requires registration:



Please enter your username and password to login:

username ← CMORE_2008

password ← CMORE2008

Not a registered CAMERA user? [Request a CAMERA account](#)

Forgot your password? [Reset it here](#)

USEFUL LINKS
MARINE MICRO SPECIFIC

<http://camera.calit2.net/>

<http://www.moore.org/microgenome/>

<http://egg.umh.es/micromar/>

<http://www.megx.net/>

GENERIC TOOLS AND MICROBIAL GENOME EXPLORATION

http://genome.igi-psf.org/mic_home.htm

<http://www.softberry.com/all.htm>

<http://www.ncbi.nih.gov/>

<http://img.igi.doe.gov/cgi-bin/pub/main.cgi>

RIBOSOMAL RNA DATABASE AND PROBE RESOURCES AND TOOLS

<http://greengenes.lbl.gov/cgi-bin/nph-index.cgi>

<http://www.microbial-ecology.net/probebase/>

<http://www.arb-home.de/>

<http://rdp.cme.msu.edu/>

TARGETED PROTEIN DATABASES AND SEARCH TOOLS ONLINE

http://nar.oxfordjournals.org/content/vol34/suppl_1/index.dtl

<http://www.genome.jp/kegg/>

<http://www.ncbi.nlm.nih.gov/COG/>

<http://pfam.sanger.ac.uk/>

<http://www.ebi.ac.uk/interpro/>

<http://string.embl.de>