Metagenomic Analysis

Analyzing multibacterial (multispecies) samples from the environment

November 22, 2016

Metagenomics

- Metagenomics deals with samples taken directly from the environment
  - Soil, water, hot spring, oil sands, human gut, stool
  - Also called environmental genomics

- Necessarily more complex than genomics
  - Mixture of multiple organisms
  - Many have never been looked at on the molecular level at all

Metagenomic analysis

- Who’s there?
  - Taxonomy analysis

- How many/much of them are there?
  - Relative abundance of organisms

- What do they do?
  - Functional analysis

Basic principles

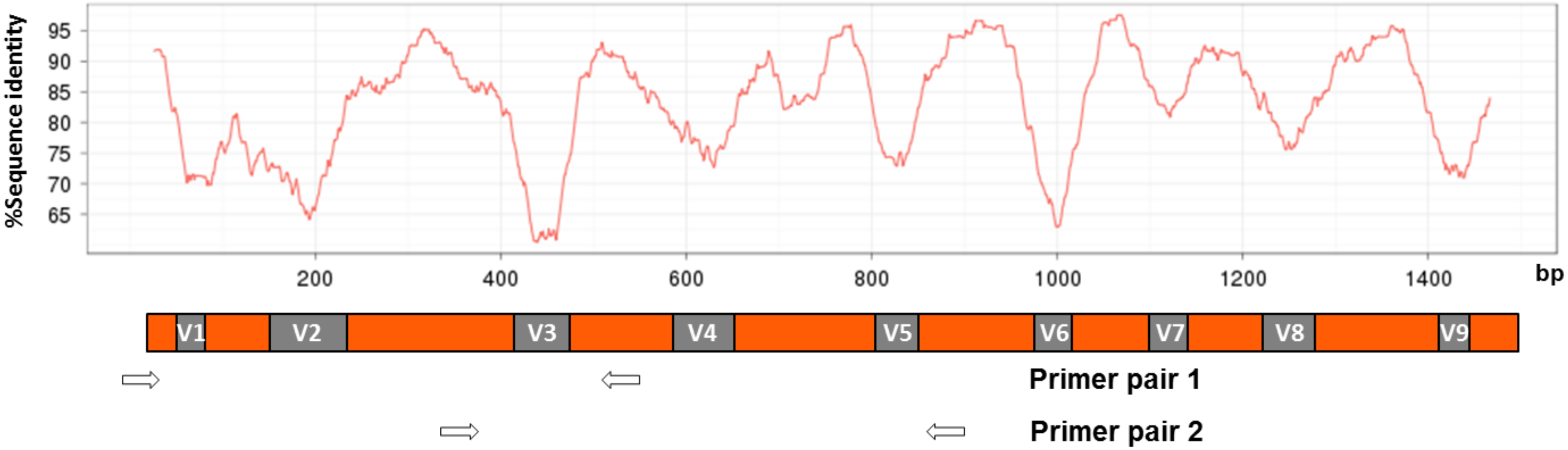
- Simple analysis using a reference gene 16SrRNA
  - Amplification of reference gene with general primers
  - Overall microbial community composition (presence-absence)
- Whole genome sequencing (WGS)
  - Sequencing + mapping to known genomes (or to specific marker database)
  - Overall microbial community composition (quantitative)
  - Dominant functions
  - It can also be done with assembly

Comparative metagenomic analysis

- Across different environments
  - Individual taxon abundance
  - Overall microbial community composition
  - Dominant functions
  - Environment-specific functions of same taxon
- Within same environment
  - Changes over time

16S rRNA sequencing

- Early and still common method
- Highly conserved yet unique to individual (mostly bacterial) species
  - Consists of variable and conserved regions
- Targeted sequencing using primer pairs



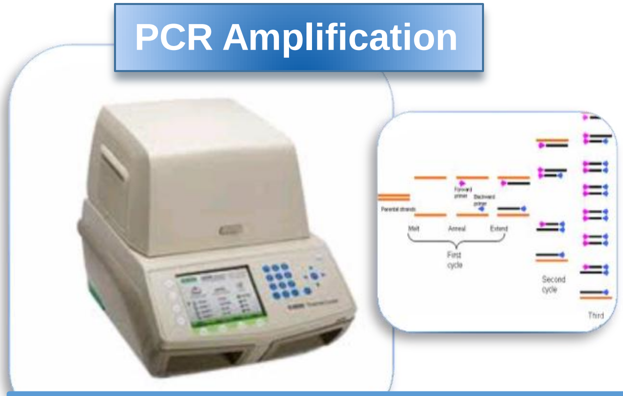
http://www.gatc-biotech.com/index.php?id=1025&L=1

16s rRNA based community analysis procedures

DNA Extraction

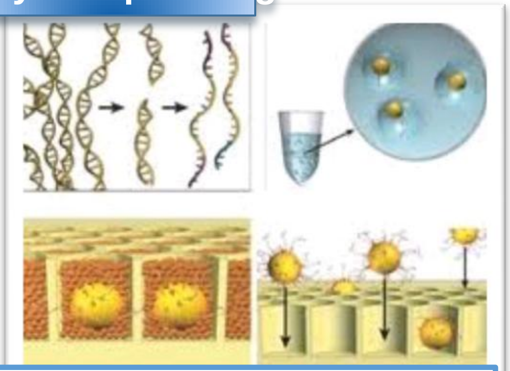
From Environment Samples
From Enrichment Cultures

PCR Amplification



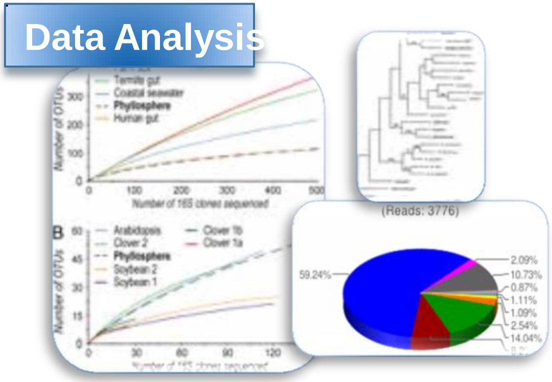
Gene: 16s rRNA
Primer sets: 1392R454A, 926F454B

Pyrosequencing



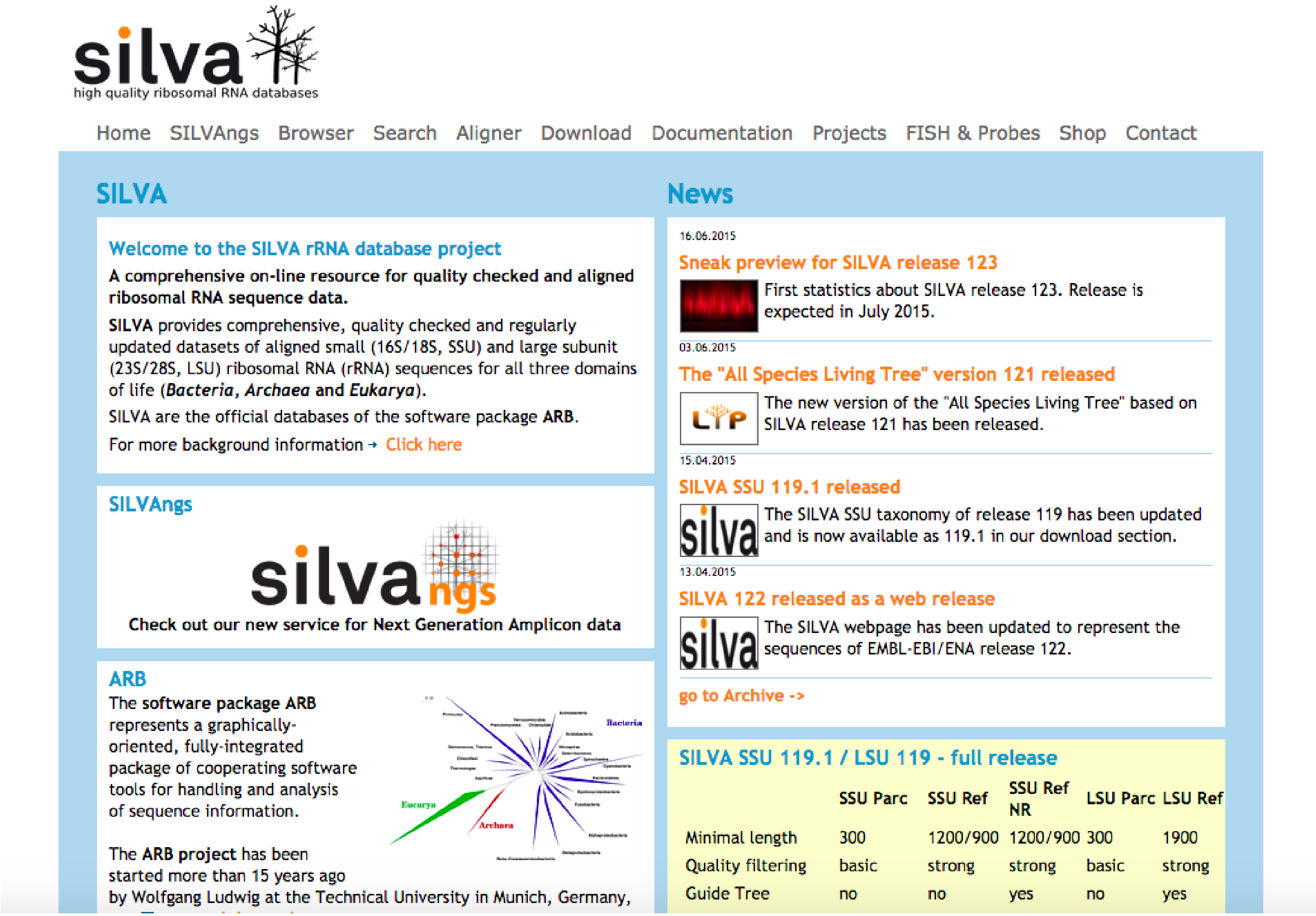
454 Titanium Pyrosequencing

Data Analysis



Microbial Community Analysis

SILVA Database



SILVA

Welcome to the SILVA rRNA database project

A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data.

SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (***Bacteria***, ***Archaea*** and ***Eukarya***).

SILVA are the official databases of the software package **ARB**.

For more background information → Click here

SILVAngs

silva ngs

Check out our new service for Next Generation Amplicon data

ARB

The **software package ARB** represents a graphically-oriented, fully-integrated package of cooperating software tools for handling and analysis of sequence information.

The **ARB project** has been started more than 15 years ago by Wolfgang Ludwig at the Technical University in Munich, Germany,

News

16.06.2015

Sneak preview for SILVA release 123

First statistics about SILVA release 123. Release is expected in July 2015.

03.06.2015

The "All Species Living Tree" version 121 released

The new version of the "All Species Living Tree" based on SILVA release 121 has been released.

15.04.2015

SILVA SSU 119.1 released

The SILVA SSU taxonomy of release 119 has been updated and is now available as 119.1 in our download section.

13.04.2015

SILVA 122 released as a web release

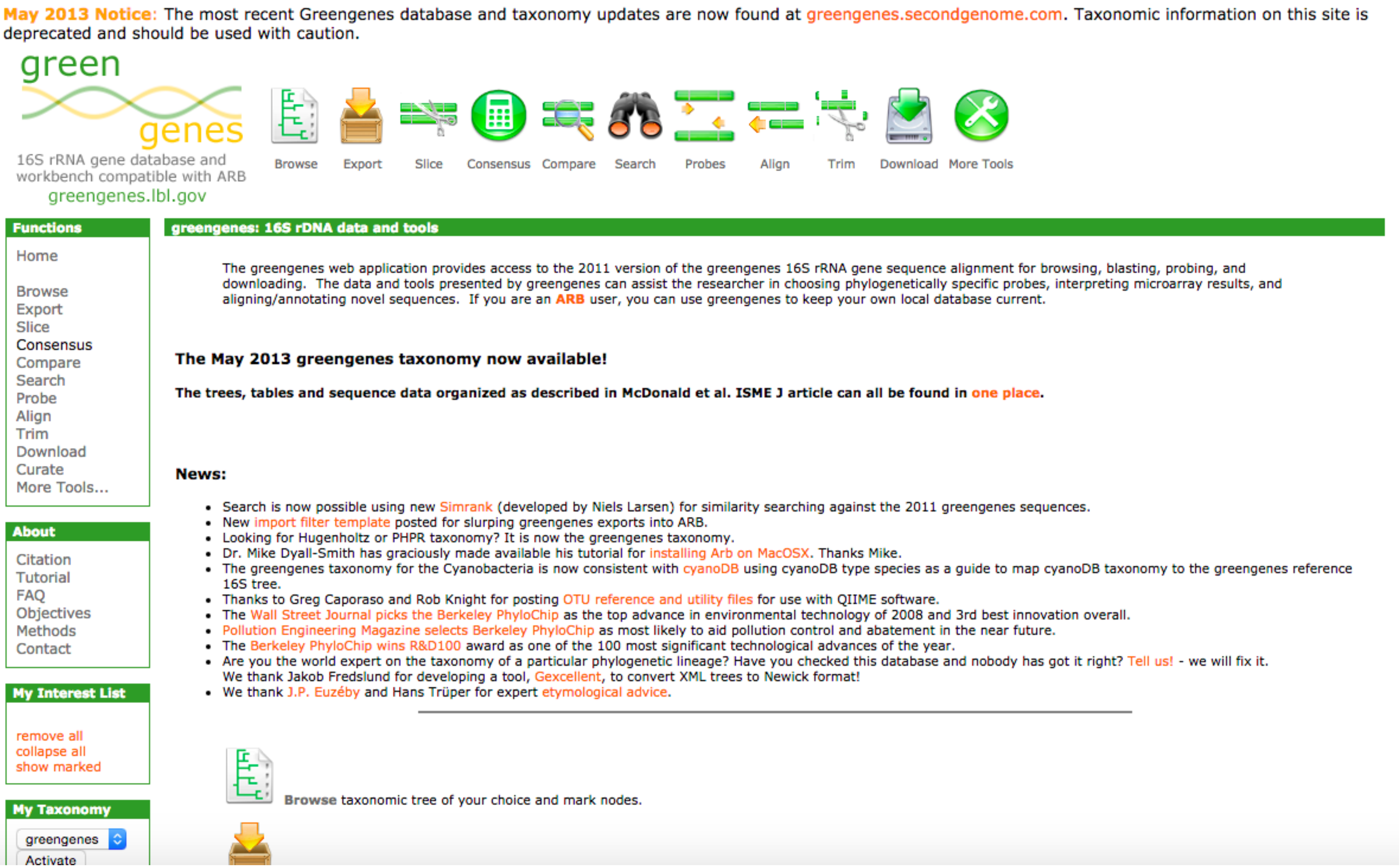
The SILVA webpage has been updated to represent the sequences of EMBL-EBI/ENA release 122.

go to Archive ->

SILVA SSU 119.1 / LSU 119 - full release

	SSU Parc	SSU Ref	SSU Ref NR	LSU Parc	LSU Ref
Minimal length	300	1200/900	1200/900	300	1900
Quality filtering	basic	strong	strong	basic	strong
Guide Tree	no	no	yes	no	yes

GreenGenes Database



May 2013 Notice: The most recent Greengenes database and taxonomy updates are now found at greengenes.secondgenome.com. Taxonomic information on this site is deprecated and should be used with caution.

green genes
16S rRNA gene database and workbench compatible with ARB
greengenes.lbl.gov

greengenes: 16S rDNA data and tools

The greengenes web application provides access to the 2011 version of the greengenes 16S rRNA gene sequence alignment for browsing, blasting, probing, and downloading. The data and tools presented by greengenes can assist the researcher in choosing phylogenetically specific probes, interpreting microarray results, and aligning/annotating novel sequences. If you are an **ARB** user, you can use greengenes to keep your own local database current.

The May 2013 greengenes taxonomy now available!

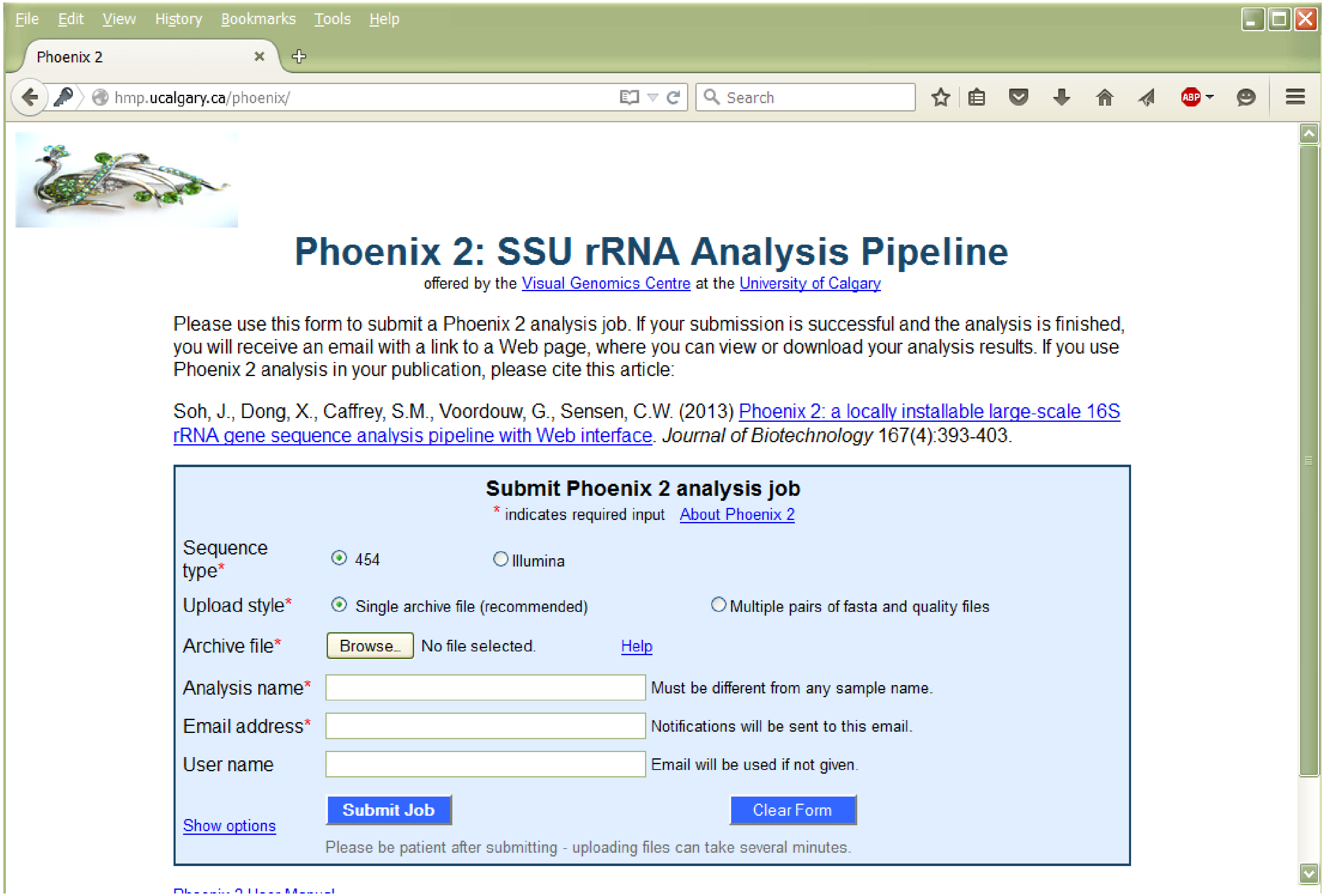
The trees, tables and sequence data organized as described in McDonald et al. ISME J article can all be found in one place.

News:

- Search is now possible using new Simrank (developed by Niels Larsen) for similarity searching against the 2011 greengenes sequences.
- New import filter template posted for slurping greengenes exports into ARB.
- Looking for Hugenholtz or PHPR taxonomy? It is now the greengenes taxonomy.
- Dr. Mike Dyall-Smith has graciously made available his tutorial for installing Arb on MacOSX. Thanks Mike.
- The greengenes taxonomy for the Cyanobacteria is now consistent with cyanoDB using cyanoDB type species as a guide to map cyanoDB taxonomy to the greengenes reference 16S tree.
- Thanks to Greg Caporaso and Rob Knight for posting OTU reference and utility files for use with QIIME software.
- The Wall Street Journal picks the Berkeley PhyloChip as the top advance in environmental technology of 2008 and 3rd best innovation overall.
- Pollution Engineering Magazine selects Berkeley PhyloChip as most likely to aid pollution control and abatement in the near future.
- The Berkeley PhyloChip wins R&D100 award as one of the 100 most significant technological advances of the year.
- Are you the world expert on the taxonomy of a particular phylogenetic lineage? Have you checked this database and nobody has got it right? Tell us! - we will fix it.
  We thank Jakob Fredslund for developing a tool, Gexcellent, to convert XML trees to Newick format!
- We thank J.P. Euzéby and Hans Trüper for expert etymological advice.

Browse taxonomic tree of your choice and mark nodes.

A 16S rRNA analysis pipeline



Phoenix 2: SSU rRNA Analysis Pipeline

offered by the Visual Genomics Centre at the University of Calgary

Please use this form to submit a Phoenix 2 analysis job. If your submission is successful and the analysis is finished, you will receive an email with a link to a Web page, where you can view or download your analysis results. If you use Phoenix 2 analysis in your publication, please cite this article:

Soh, J., Dong, X., Caffrey, S.M., Voordouw, G., Sensen, C.W. (2013) Phoenix 2: a locally installable large-scale 16S rRNA gene sequence analysis pipeline with Web interface. *Journal of Biotechnology* 167(4):393-403.

Submit Phoenix 2 analysis job

* indicates required input About Phoenix 2

Sequence type*	● 454	○ Illumina
Upload style*	● Single archive file (recommended)	○ Multiple pairs of fasta and quality files
Archive file*	Browse... No file selected.	Help
Analysis name*		Must be different from any sample name.
Email address*		Notifications will be sent to this email.
User name		Email will be used if not given.
Show options	Submit Job	Clear Form

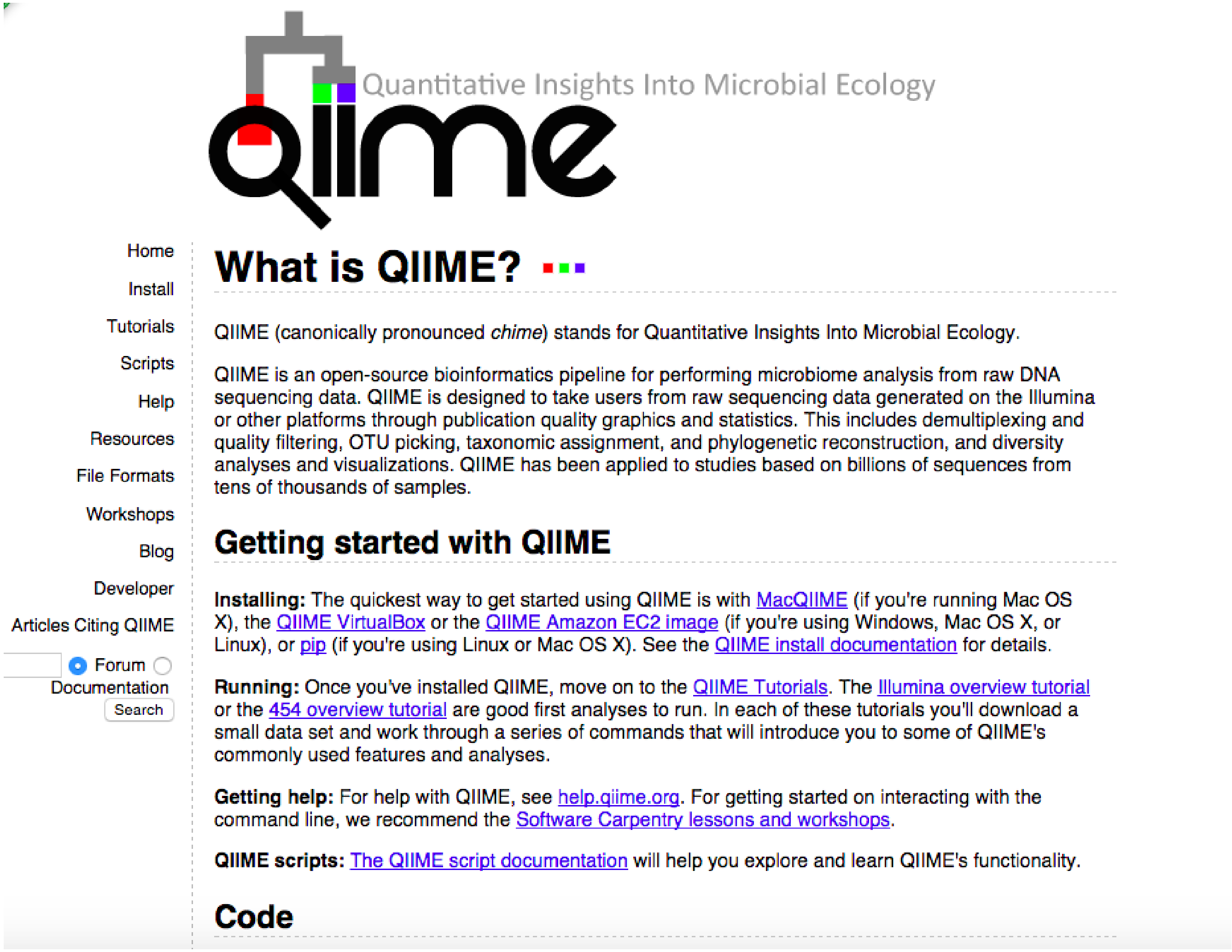
Please be patient after submitting - uploading files can take several minutes.

Phoenix 2 User Manual

Metagenomic Tools (Local)

- Qiime (Quantitative Insights Into Microbial Ecology)
  - Consists of Python scripts
  - Taxonomy and diversity statistics/visualization
- Mothur
  - Commands written in C++ programs
  - Taxonomy and diversity statics/visualization
- MEGAN (Metagenome Analyzer)
  - Provides functional as well as taxonomy analysis
  - GUI with tree-based visualization

QUIIME (Local Installation)



What is QIIME? 🟥🟩🟪

QIIME (canonically pronounced *chime*) stands for Quantitative Insights Into Microbial Ecology.

QIIME is an open-source bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data. QIIME is designed to take users from raw sequencing data generated on the Illumina or other platforms through publication quality graphics and statistics. This includes demultiplexing and quality filtering, OTU picking, taxonomic assignment, and phylogenetic reconstruction, and diversity analyses and visualizations. QIIME has been applied to studies based on billions of sequences from tens of thousands of samples.

Getting started with QIIME

Installing: The quickest way to get started using QIIME is with MacQIIME (if you're running Mac OS X), the QIIME VirtualBox or the QIIME Amazon EC2 image (if you're using Windows, Mac OS X, or Linux), or pip (if you're using Linux or Mac OS X). See the QIIME install documentation for details.

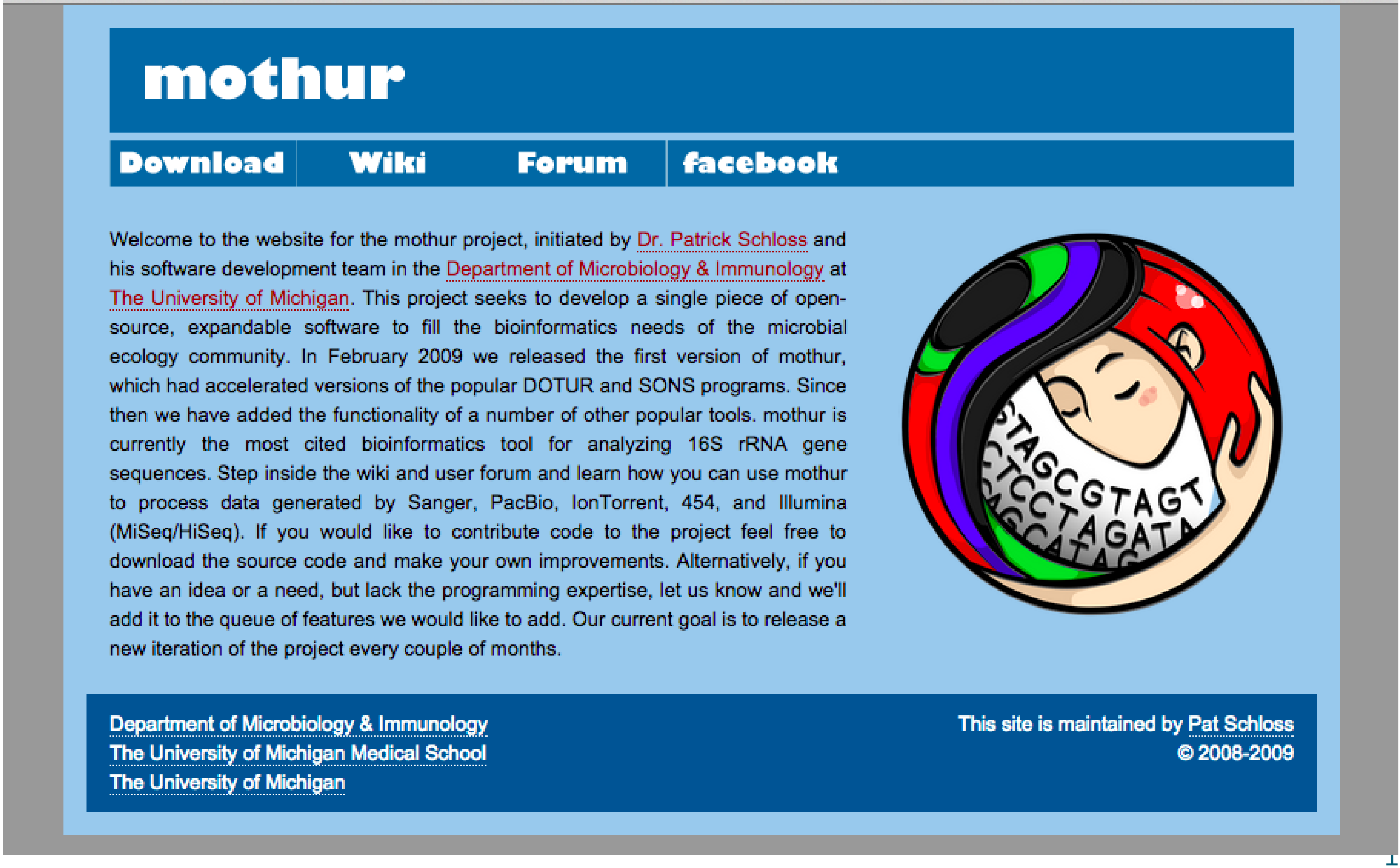
Running: Once you've installed QIIME, move on to the QIIME Tutorials. The Illumina overview tutorial or the 454 overview tutorial are good first analyses to run. In each of these tutorials you'll download a small data set and work through a series of commands that will introduce you to some of QIIME's commonly used features and analyses.

Getting help: For help with QIIME, see help.qiime.org. For getting started on interacting with the command line, we recommend the Software Carpentry lessons and workshops.

QIIME scripts: The QIIME script documentation will help you explore and learn QIIME's functionality.

Code

MOTHUR (Local Installation)

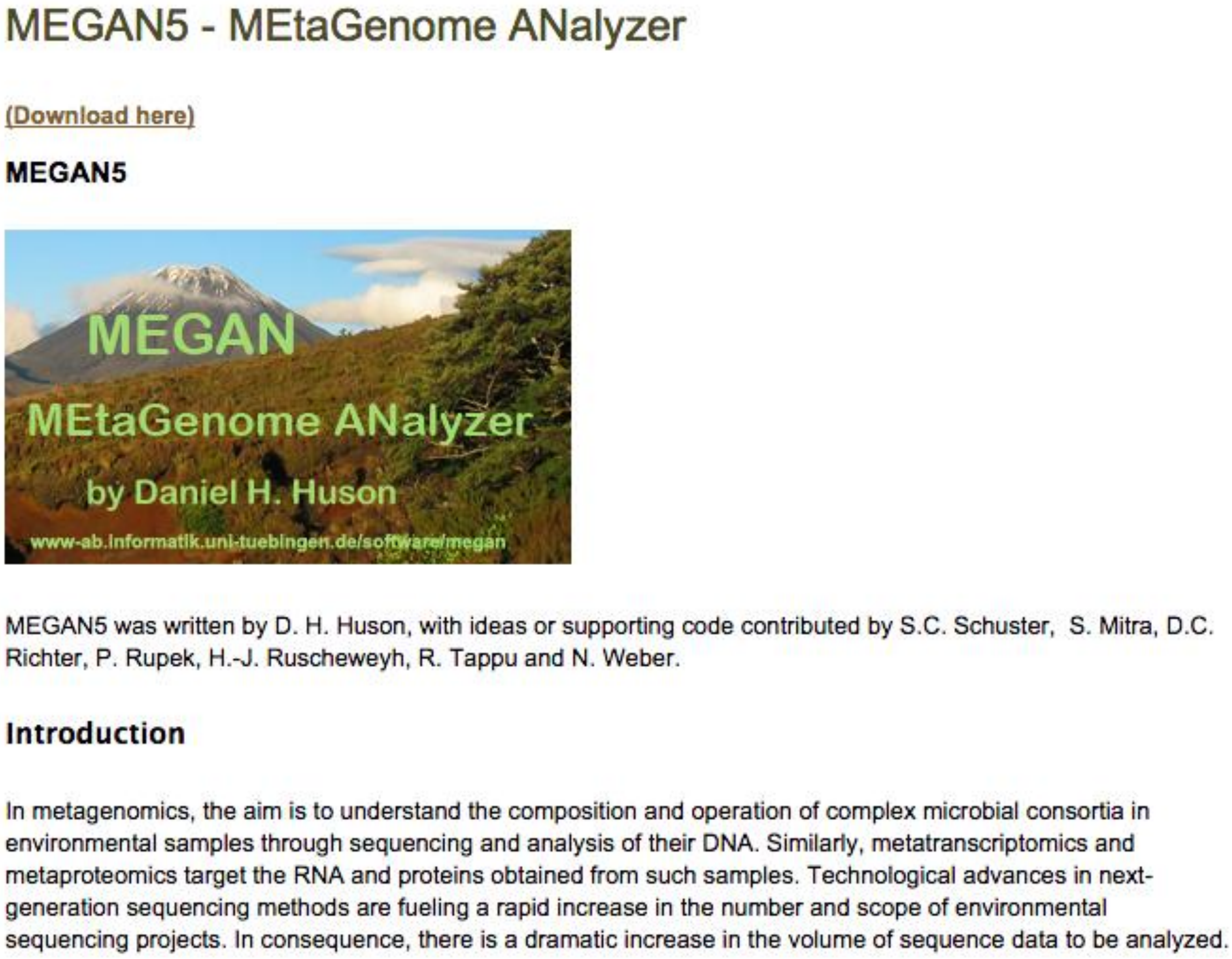


Welcome to the website for the mothur project, initiated by Dr. Patrick Schloss and his software development team in the Department of Microbiology & Immunology at The University of Michigan. This project seeks to develop a single piece of open-source, expandable software to fill the bioinformatics needs of the microbial ecology community. In February 2009 we released the first version of mothur, which had accelerated versions of the popular DOTUR and SONS programs. Since then we have added the functionality of a number of other popular tools. mothur is currently the most cited bioinformatics tool for analyzing 16S rRNA gene sequences. Step inside the wiki and user forum and learn how you can use mothur to process data generated by Sanger, PacBio, IonTorrent, 454, and Illumina (MiSeq/HiSeq). If you would like to contribute code to the project feel free to download the source code and make your own improvements. Alternatively, if you have an idea or a need, but lack the programming expertise, let us know and we'll add it to the queue of features we would like to add. Our current goal is to release a new iteration of the project every couple of months.

Department of Microbiology & Immunology
The University of Michigan Medical School
The University of Michigan

This site is maintained by Pat Schloss
© 2008-2009

Megan 5 (Local Installation)



MEGAN5 - MEtaGenome ANalyzer

(Download here)

MEGAN5

MEGAN5 was written by D. H. Huson, with ideas or supporting code contributed by S.C. Schuster, S. Mitra, D.C. Richter, P. Rupek, H.-J. Ruscheweyh, R. Tappu and N. Weber.

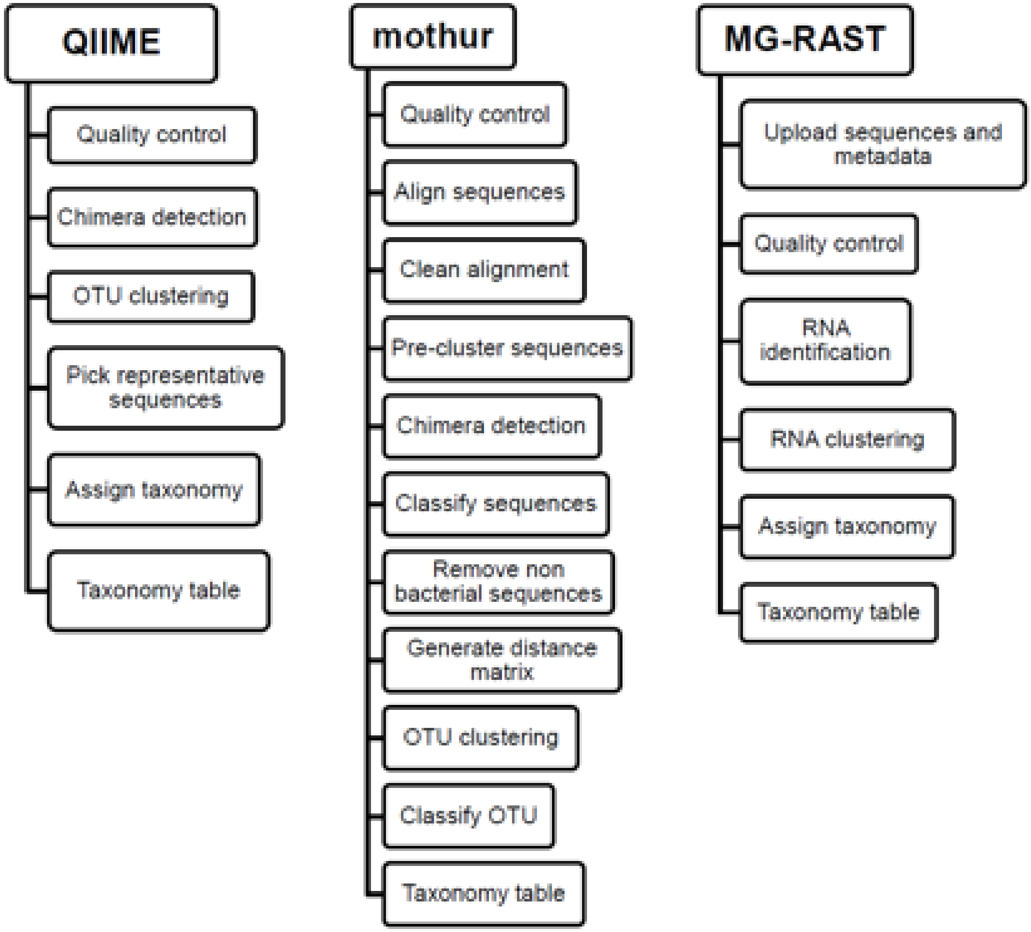
Introduction

In metagenomics, the aim is to understand the composition and operation of complex microbial consortia in environmental samples through sequencing and analysis of their DNA. Similarly, metatranscriptomics and metaproteomics target the RNA and proteins obtained from such samples. Technological advances in next-generation sequencing methods are fueling a rapid increase in the number and scope of environmental sequencing projects. In consequence, there is a dramatic increase in the volume of sequence data to be analyzed.

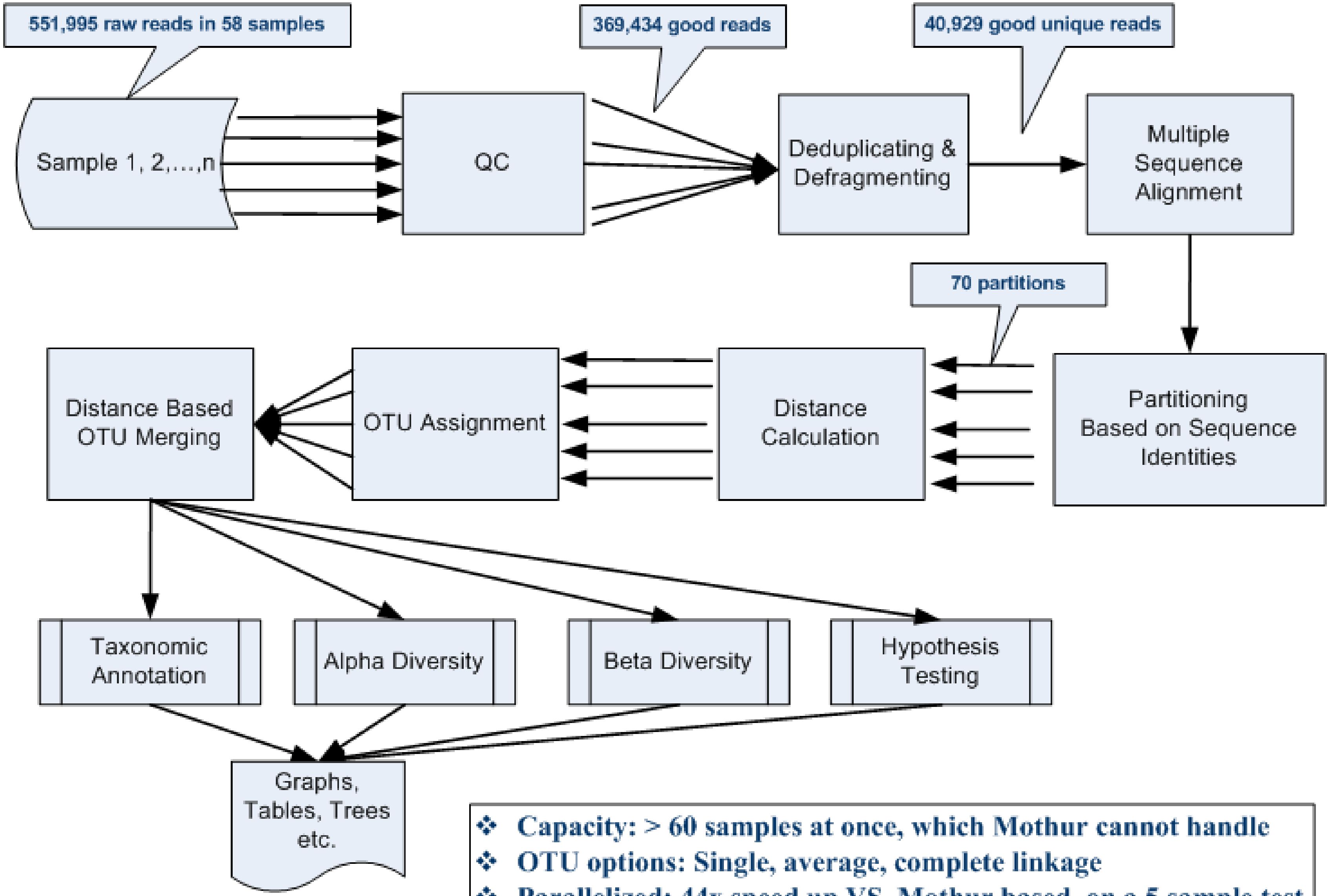
Operational Taxonomic Units

- OTUs are basic units of taxonomy analysis
- 16S rRNA reads are clustered into OTUs
  - At 97% or 95% sequence identity (usually)
- OTUs are mapped to a taxon

What do the analysis pipelines do?



Phoenix 2: Pyrotag Analysis Pipeline



551,995 raw reads in 58 samples

369,434 good reads

40,929 good unique reads

Sample 1, 2,…,n

QC

Deduplicating & Defragmenting

Multiple Sequence Alignment

70 partitions

Partitioning Based on Sequence Identities

Distance Calculation

OTU Assignment

Distance Based OTU Merging

Taxonomic Annotation

Alpha Diversity

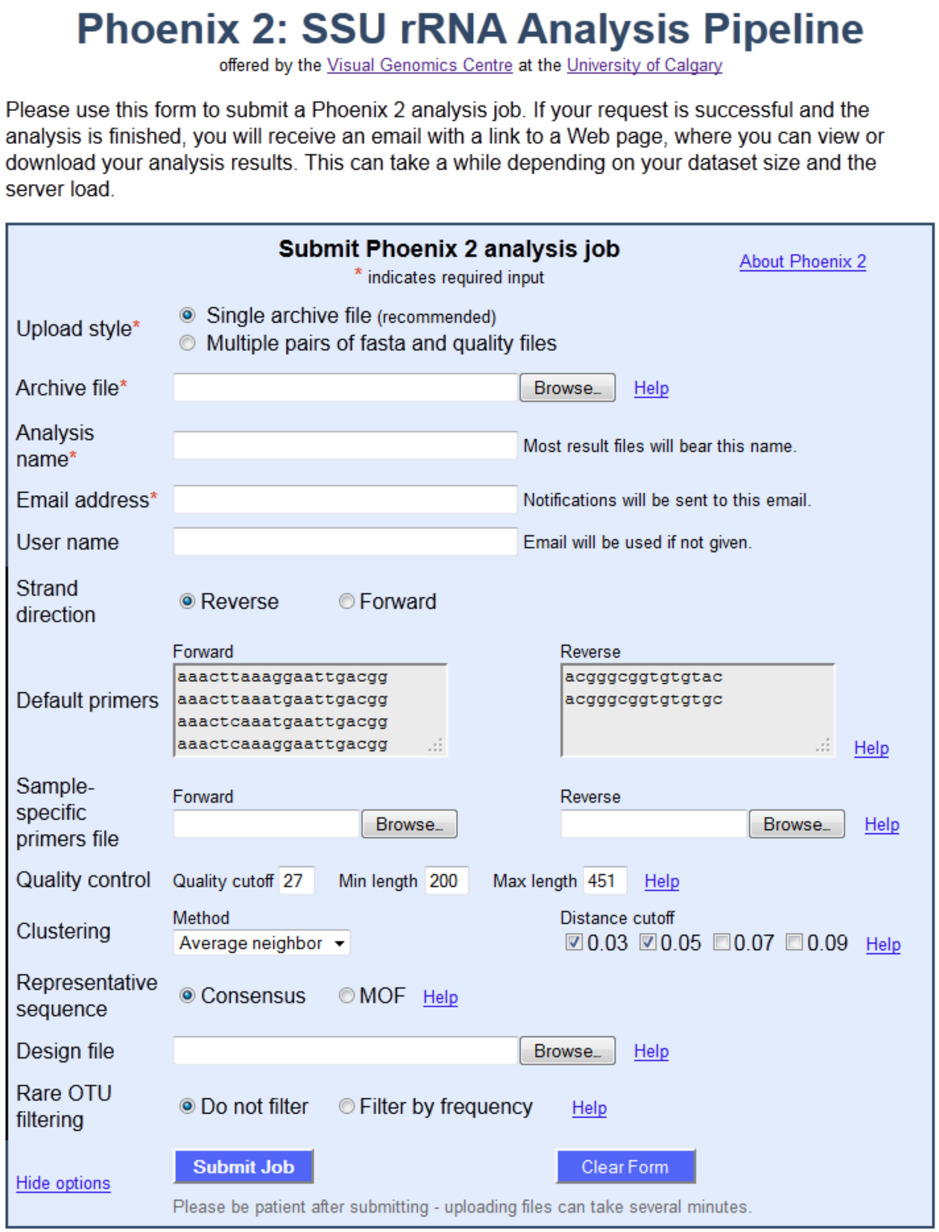
Beta Diversity

Hypothesis Testing

Graphs, Tables, Trees etc.

- ❖ Capacity: > 60 samples at once, which Mothur cannot handle
- ❖ OTU options: Single, average, complete linkage
- ❖ Parallelized: 44x speed up VS. Mothur based on a 5 sample test
- ❖ The results are almost the same as the results from Mothur

Phoenix 2 Web Interface



Phoenix 2: SSU rRNA Analysis Pipeline

offered by the Visual Genomics Centre at the University of Calgary

Please use this form to submit a Phoenix 2 analysis job. If your request is successful and the analysis is finished, you will receive an email with a link to a Web page, where you can view or download your analysis results. This can take a while depending on your dataset size and the server load.

Submit Phoenix 2 analysis job

About Phoenix 2

* indicates required input

Upload style*: ● Single archive file (recommended) ○ Multiple pairs of fasta and quality files

Archive file*: [ ] Browse... Help

Analysis name*: [ ] Most result files will bear this name.

Email address*: [ ] Notifications will be sent to this email.

User name: [ ] Email will be used if not given.

Strand direction: ● Reverse ○ Forward

Default primers:

Forward

```
aaacttaaaggaattgacgg
aaacttaaatgaattgacgg
aaactcaaatgaattgacgg
aaactcaaaggaattgacgg
```

Reverse

```
acgggcggtgtgtac
acgggcggtgtgtgc
```

Help

Sample-specific primers file: Forward [ ] Browse... Reverse [ ] Browse... Help

Quality control: Quality cutoff 27 Min length 200 Max length 451 Help

Clustering: Method: Average neighbor ▾ Distance cutoff: ☑ 0.03 ☑ 0.05 □ 0.07 □ 0.09 Help

Representative sequence: ● Consensus ○ MOF Help

Design file: [ ] Browse... Help

Rare OTU filtering: ● Do not filter ○ Filter by frequency Help

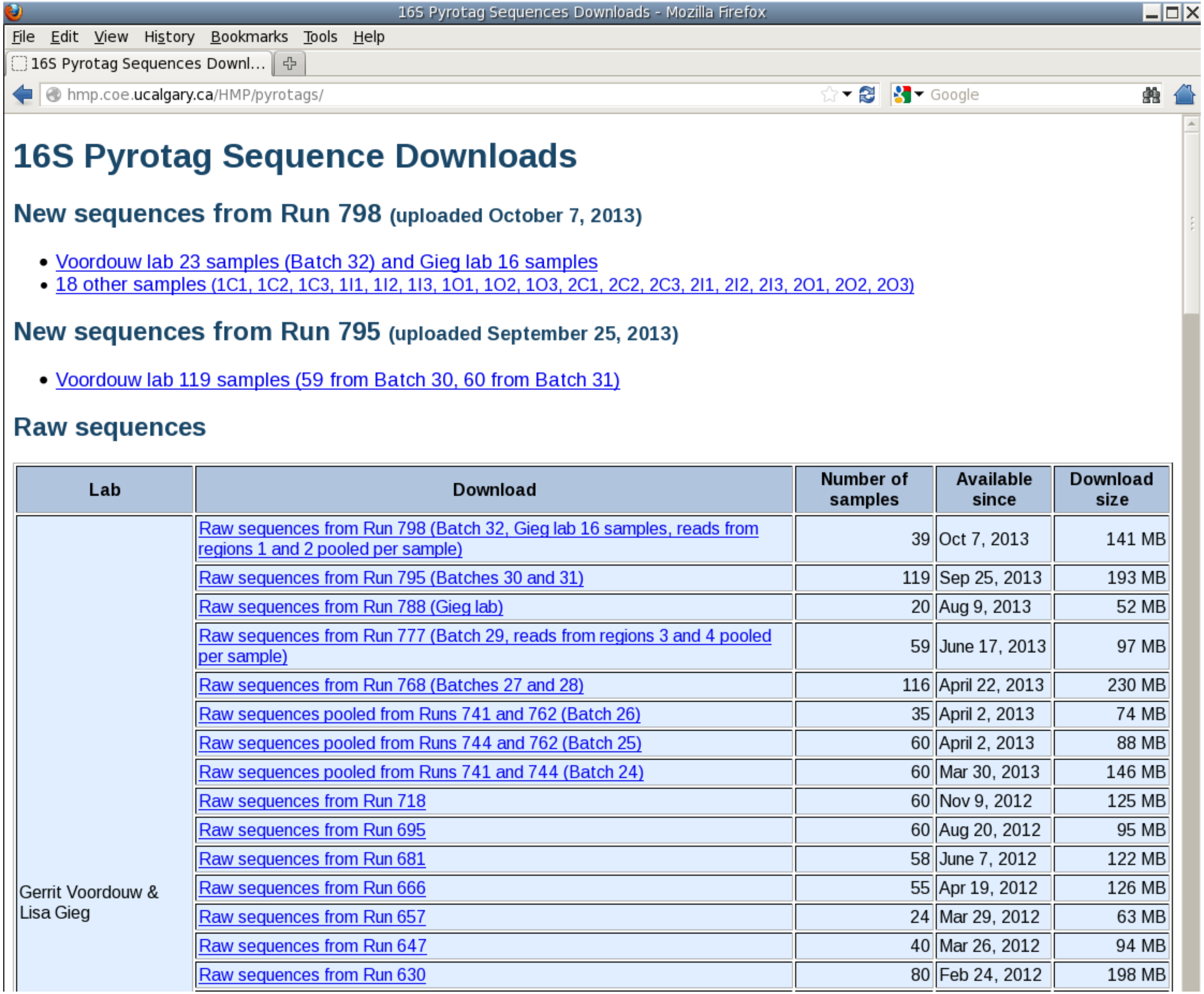
Hide options

Submit Job

Clear Form

Please be patient after submitting - uploading files can take several minutes.

Pyrotag Download Table


16S Pyrotag Sequences Downloads - Mozilla Firefox
File Edit View History Bookmarks Tools Help
16S Pyrotag Sequences Downl...
hmp.coe.ucalgary.ca/HMP/pyrotags/
Google

16S Pyrotag Sequence Downloads

New sequences from Run 798 (uploaded October 7, 2013)

- Voordouw lab 23 samples (Batch 32) and Gieg lab 16 samples
- 18 other samples (1C1, 1C2, 1C3, 1I1, 1I2, 1I3, 1O1, 1O2, 1O3, 2C1, 2C2, 2C3, 2I1, 2I2, 2I3, 2O1, 2O2, 2O3)

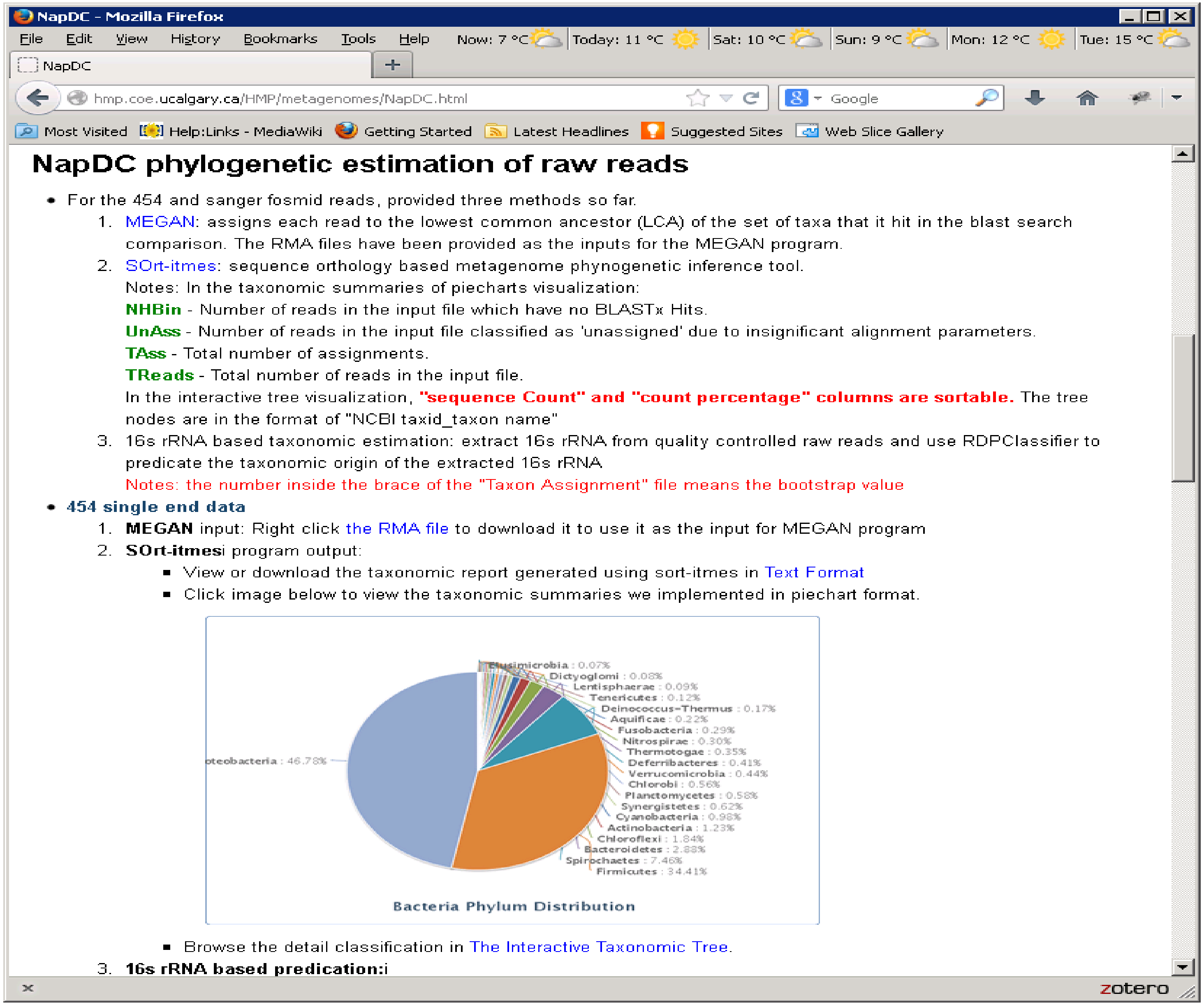
New sequences from Run 795 (uploaded September 25, 2013)

- Voordouw lab 119 samples (59 from Batch 30, 60 from Batch 31)

Raw sequences

Lab	Download	Number of samples	Available since	Download size
Gerrit Voordouw & Lisa Gieg	Raw sequences from Run 798 (Batch 32, Gieg lab 16 samples, reads from regions 1 and 2 pooled per sample)	39	Oct 7, 2013	141 MB
	Raw sequences from Run 795 (Batches 30 and 31)	119	Sep 25, 2013	193 MB
	Raw sequences from Run 788 (Gieg lab)	20	Aug 9, 2013	52 MB
	Raw sequences from Run 777 (Batch 29, reads from regions 3 and 4 pooled per sample)	59	June 17, 2013	97 MB
	Raw sequences from Run 768 (Batches 27 and 28)	116	April 22, 2013	230 MB
	Raw sequences pooled from Runs 741 and 762 (Batch 26)	35	April 2, 2013	74 MB
	Raw sequences pooled from Runs 744 and 762 (Batch 25)	60	April 2, 2013	88 MB
	Raw sequences pooled from Runs 741 and 744 (Batch 24)	60	Mar 30, 2013	146 MB
	Raw sequences from Run 718	60	Nov 9, 2012	125 MB
	Raw sequences from Run 695	60	Aug 20, 2012	95 MB
	Raw sequences from Run 681	58	June 7, 2012	122 MB
	Raw sequences from Run 666	55	Apr 19, 2012	126 MB
	Raw sequences from Run 657	24	Mar 29, 2012	63 MB
	Raw sequences from Run 647	40	Mar 26, 2012	94 MB
	Raw sequences from Run 630	80	Feb 24, 2012	198 MB

Data Access: Phylogeny



NapDC - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Now: 7 °C Today: 11 °C Sat: 10 °C Sun: 9 °C Mon: 12 °C Tue: 15 °C
NapDC
hmp.coe.ucalgary.ca/HMP/metagenomes/NapDC.html
Google
Most Visited Help:Links - MediaWiki Getting Started Latest Headlines Suggested Sites Web Slice Gallery

NapDC phylogenetic estimation of raw reads

- For the 454 and sanger fosmid reads, provided three methods so far.
  1. MEGAN: assigns each read to the lowest common ancestor (LCA) of the set of taxa that it hit in the blast search comparison. The RMA files have been provided as the inputs for the MEGAN program.
  2. SOrt-itmes: sequence orthology based metagenome phynogenetic inference tool.
     Notes: In the taxonomic summaries of piecharts visualization:
     NHBin - Number of reads in the input file which have no BLASTx Hits.
     UnAss - Number of reads in the input file classified as 'unassigned' due to insignificant alignment parameters.
     TAss - Total number of assignments.
     TReads - Total number of reads in the input file.
     In the interactive tree visualization, **"sequence Count" and "count percentage" columns are sortable.** The tree nodes are in the format of "NCBI taxid_taxon name"
  3. 16s rRNA based taxonomic estimation: extract 16s rRNA from quality controlled raw reads and use RDPClassifier to predicate the taxonomic origin of the extracted 16s rRNA
     Notes: the number inside the brace of the "Taxon Assignment" file means the bootstrap value
- **454 single end data**
  1. **MEGAN** input: Right click the RMA file to download it to use it as the input for MEGAN program
  2. **SOrt-itmes**i program output:
     - View or download the taxonomic report generated using sort-itmes in Text Format
     - Click image below to view the taxonomic summaries we implemented in piechart format.

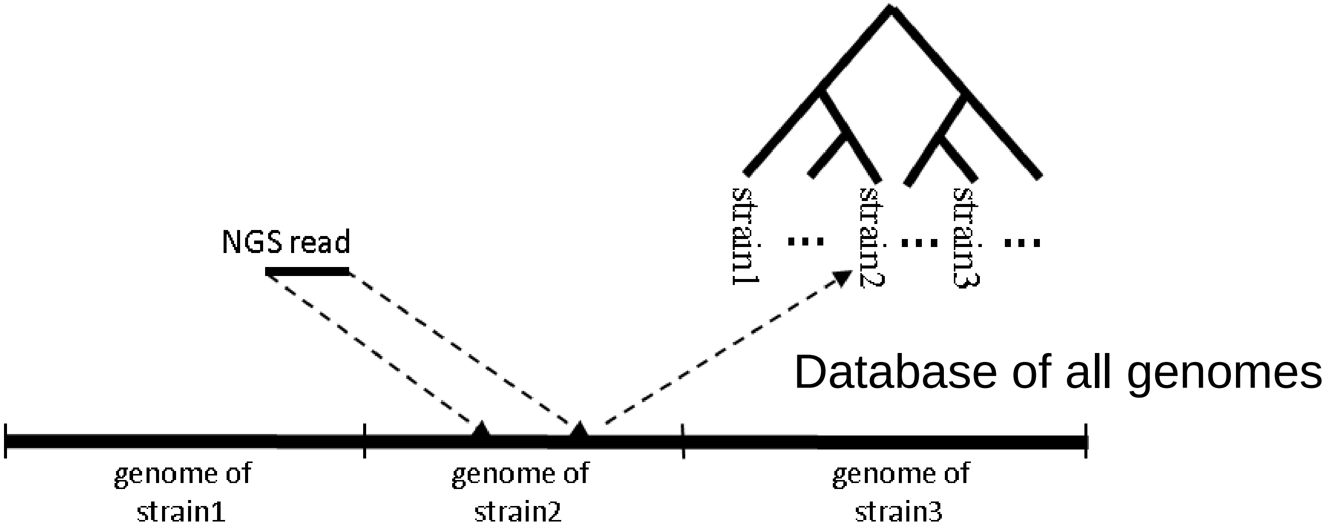
     - Browse the detail classification in The Interactive Taxonomic Tree.
  3. **16s rRNA based predication:**i

Metagenomic sequencing

- Whole metagenome with „binning”
  - Map genes to (annotated) genomic sequences
  - Count hits by taxa („bins”) – gives taxonomic composition incl. Quantitation of taxa.
  - You can use marker database instead of full genomes. Faster but less sensitive.

- Whole metagenome with assembly
  - Assemble reads just like in genome assembly
  - Complicated because of multiple unknown sources of metagenomic reads, lower coverage on individual genomes.
  - Requires large computers.

1) The input reads are aligned against a dbase of genomes
2) The alignments are processed using the lowest common ancestor algorithm
3) Each read is assigned to a taxonomic level
4) The result is a list of reads with the assigned taxonomy

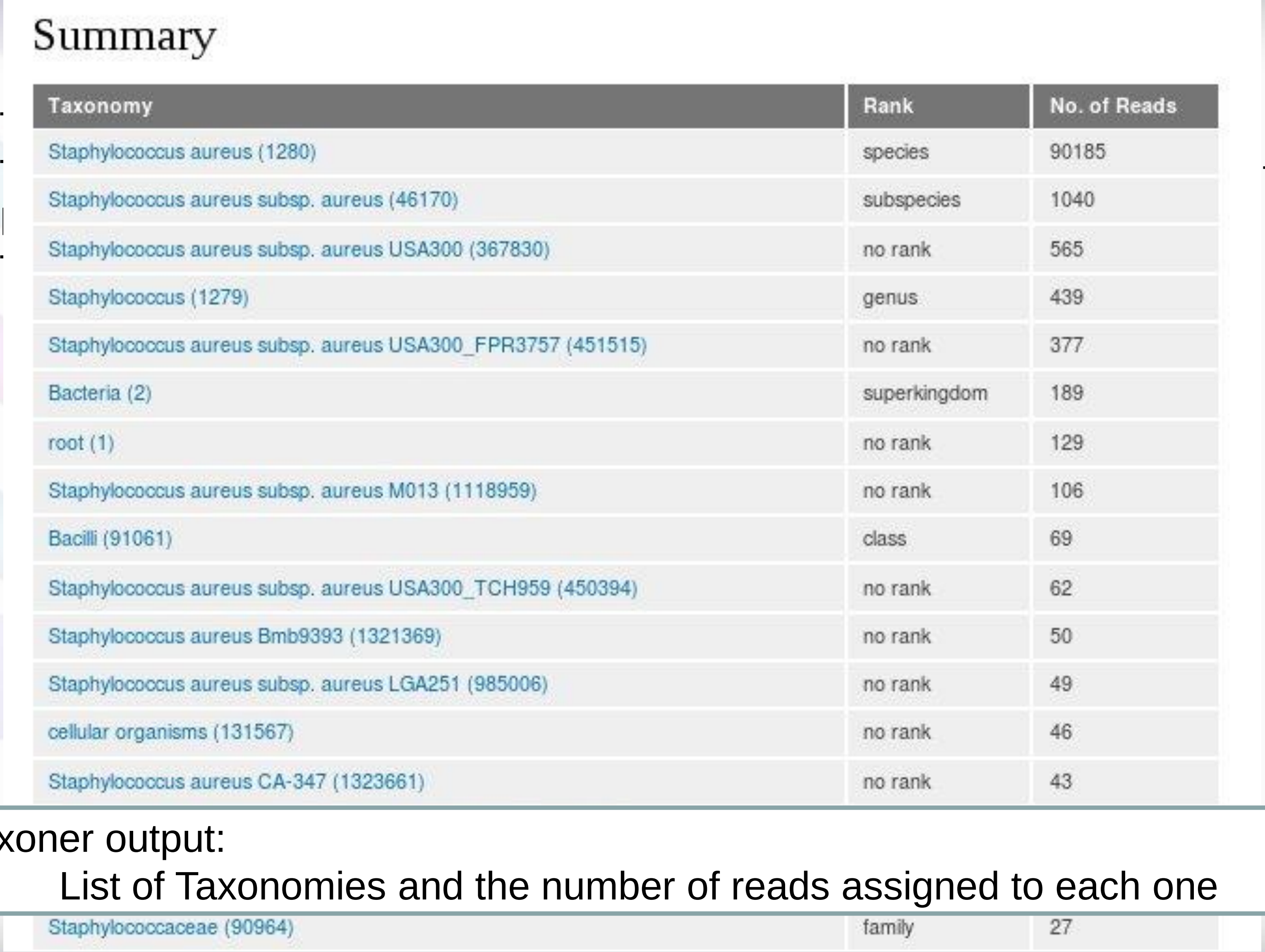


Taxoner output:

List of Taxonomies and the number of reads assigned to each one

Taxoner: Principle

1)
2) thm
3)
4)



Summary

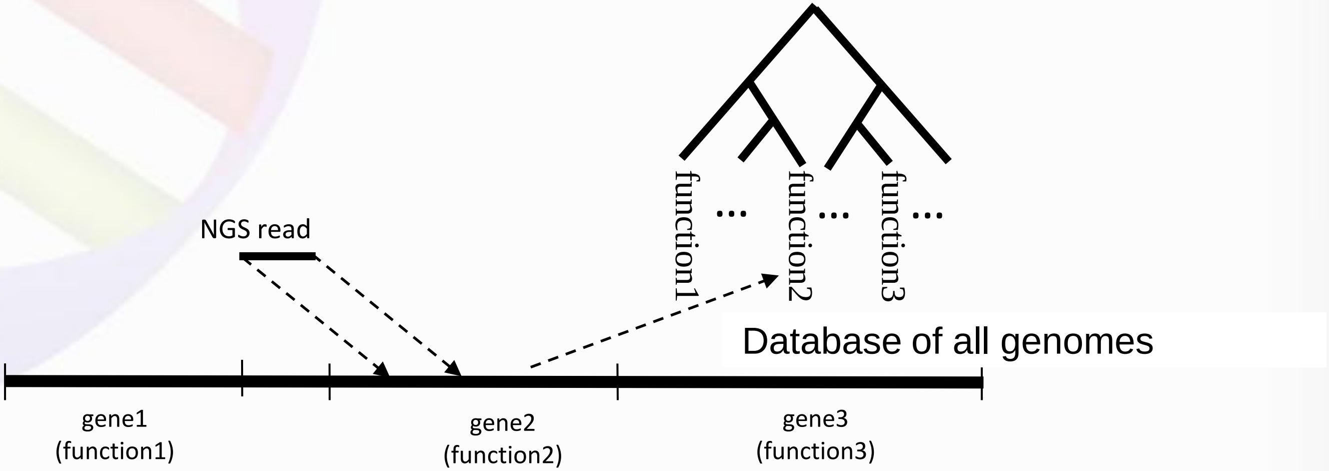
Taxonomy	Rank	No. of Reads
Staphylococcus aureus (1280)	species	90185
Staphylococcus aureus subsp. aureus (46170)	subspecies	1040
Staphylococcus aureus subsp. aureus USA300 (367830)	no rank	565
Staphylococcus (1279)	genus	439
Staphylococcus aureus subsp. aureus USA300_FPR3757 (451515)	no rank	377
Bacteria (2)	superkingdom	189
root (1)	no rank	129
Staphylococcus aureus subsp. aureus M013 (1118959)	no rank	106
Bacilli (91061)	class	69
Staphylococcus aureus subsp. aureus USA300_TCH959 (450394)	no rank	62
Staphylococcus aureus Bmb9393 (1321369)	no rank	50
Staphylococcus aureus subsp. aureus LGA251 (985006)	no rank	49
cellular organisms (131567)	no rank	46
Staphylococcus aureus CA-347 (1323661)	no rank	43
Staphylococcaceae (90964)	family	27

Taxoner output:

List of Taxonomies and the number of reads assigned to each one

Taxoner: Gene Assignment

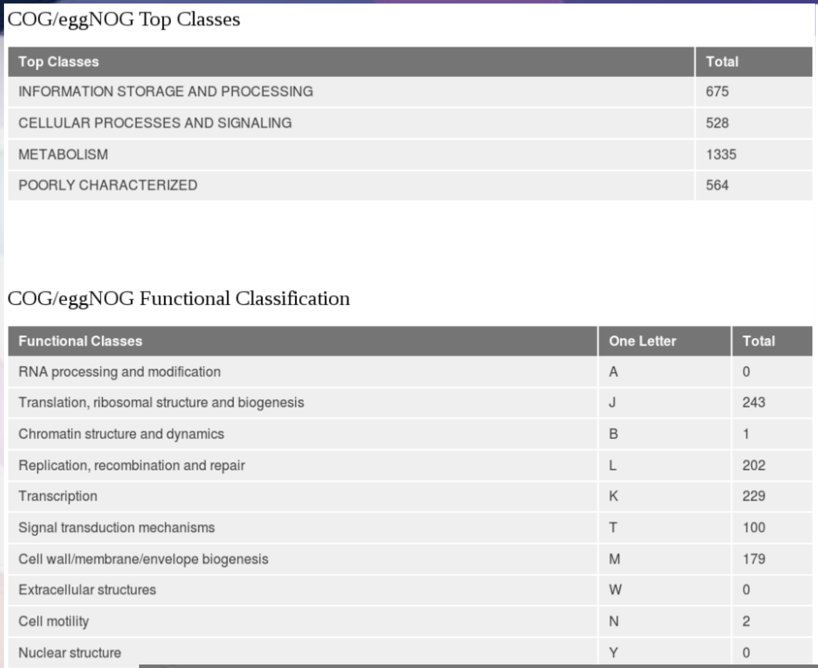
1) Genes and functions are assigned to the lower taxonomic levels (species, subspecies and strain)
2) The algorithm is based on an integrated database created with JBioWH using Gene, PTT, COG and eggNOG databases
3) The result is a list of genes per taxonomy and COG-eggNOG functional classification



Gene and function assignment output:

List of COG groups and number of genes per taxonomies

Taxoner: Gene Assignment

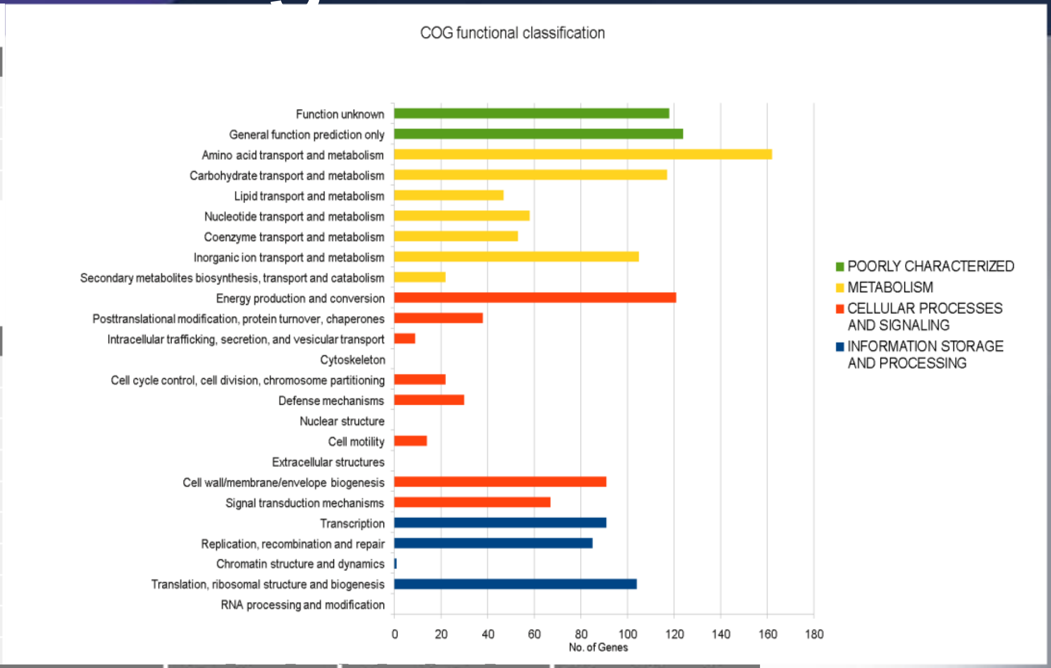


COG/eggNOG Top Classes

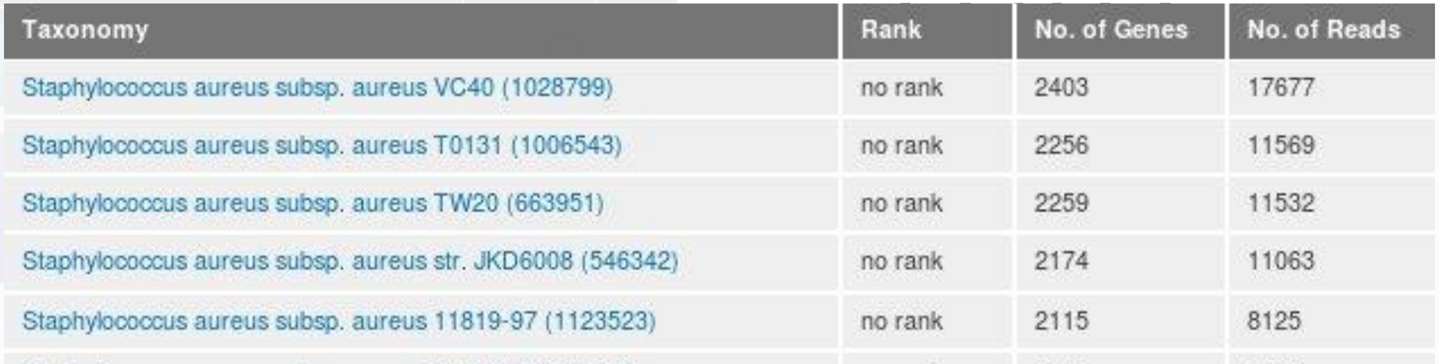
Top Classes	Total
INFORMATION STORAGE AND PROCESSING	675
CELLULAR PROCESSES AND SIGNALING	528
METABOLISM	1335
POORLY CHARACTERIZED	564

COG/eggNOG Functional Classification

Functional Classes	One Letter	Total
RNA processing and modification	A	0
Translation, ribosomal structure and biogenesis	J	243
Chromatin structure and dynamics	B	1
Replication, recombination and repair	L	202
Transcription	K	229
Signal transduction mechanisms	T	100
Cell wall/membrane/envelope biogenesis	M	179
Extracellular structures	W	0
Cell motility	N	2
Nuclear structure	Y	0



COG functional classification



Taxonomy	Rank	No. of Genes	No. of Reads
Staphylococcus aureus subsp. aureus VC40 (1028799)	no rank	2403	17677
Staphylococcus aureus subsp. aureus T0131 (1006543)	no rank	2256	11569
Staphylococcus aureus subsp. aureus TW20 (663951)	no rank	2259	11532
Staphylococcus aureus subsp. aureus str. JKD6008 (546342)	no rank	2174	11063
Staphylococcus aureus subsp. aureus 11819-97 (1123523)	no rank	2115	8125

Gene and function assignment output:

List of COG groups and number of genes per taxonomies

Taxoner: Run times and examples

		Running time[1]		
	Dbase	1 thread	4 threads	12 threads
MetaPhlAn	own bacterial marker dbase[2]	14 sec	7 sec	6 sec
Taxoner	NCBI nt Bacteria[3]	**165 sec**	**105 sec**	**90 sec**
Taxoner	NCBI nt full dbase[4]	2446 sec	2031 sec	1866 sec
MEGABLAST	NCBI nt bacteria[3]	8.3 h	n/a	3.9 h
MEGABLAST	NCBI nt full dbase[4]	37.6 h	n/a	9.4h

- MetaPhlAn was selected because of its speed and accuracy in estimating taxon composition
- Megablast was selected because of its reputation in alignments

Processor: Intel Xeon CPU: E5-2640

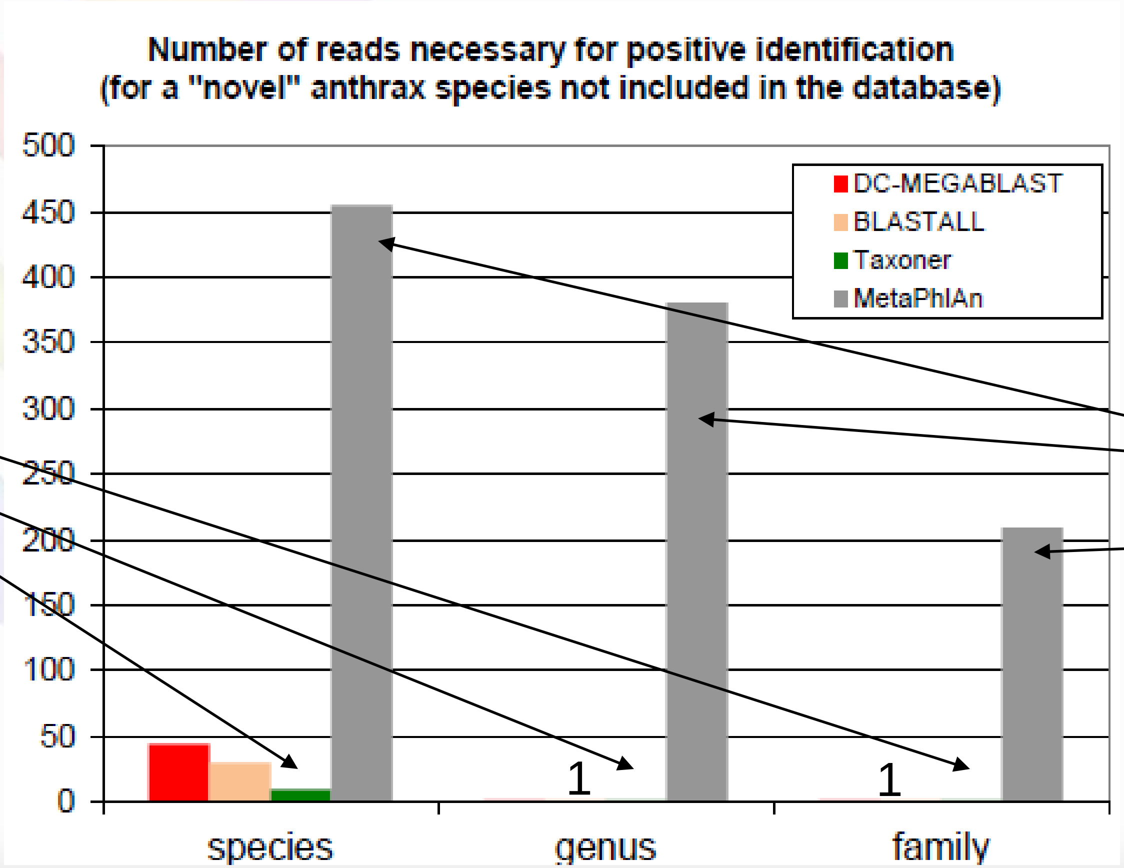
[1] Dataset: SRR292150, Reads 183 203 (27 MB)

[2] Own database with 366 988 039 nucleotides (367 MB)

[3] 15 400 949 699 nucleotides (15 GB)

[4] 52 380 339 934 nucleotides (54 GB)

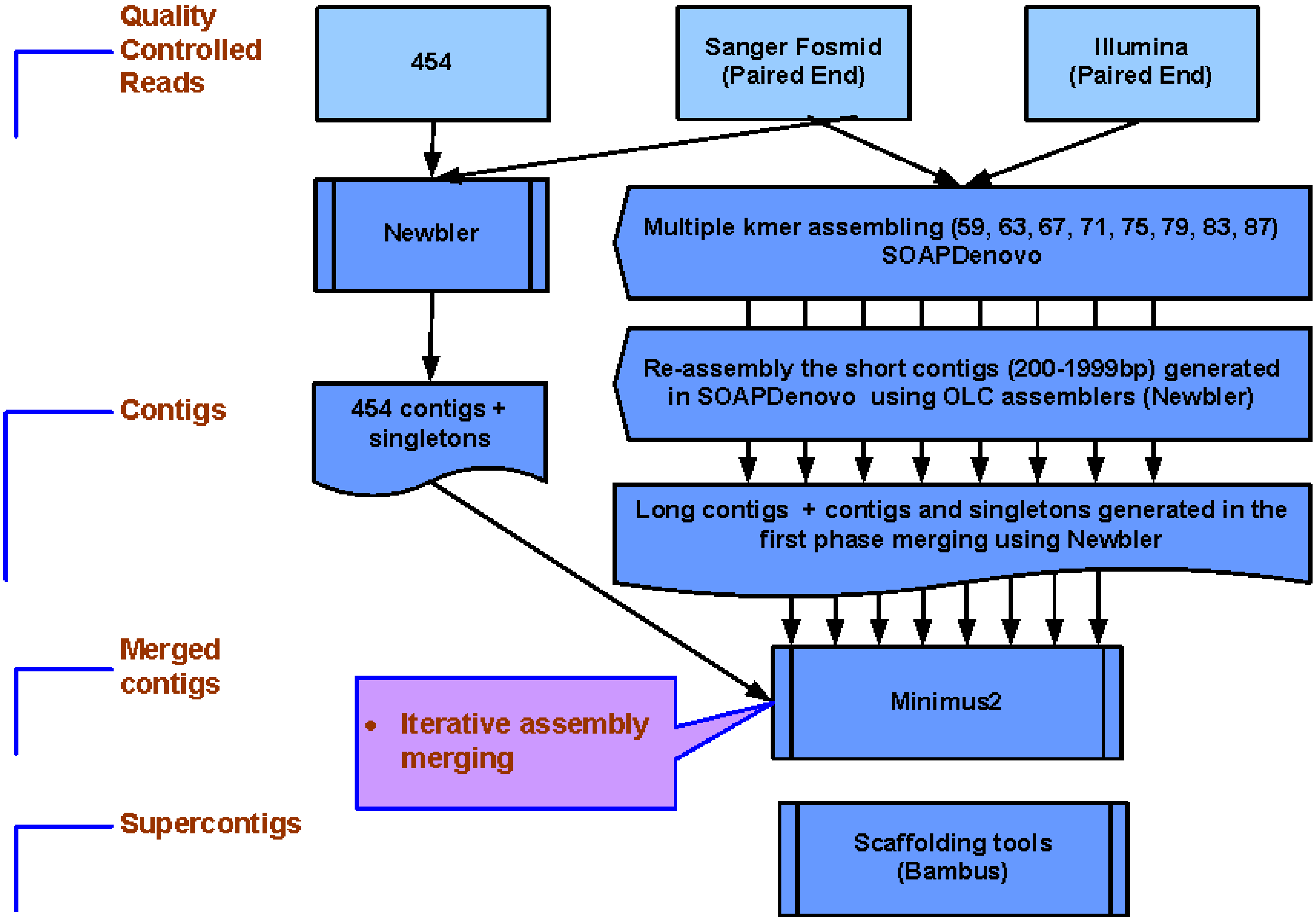
Marker databases make analysis fast but much less sensitive..



Full dbase
165 sec
(Metaphlan)

Marker dbase
14 sec
(Metaphlan)

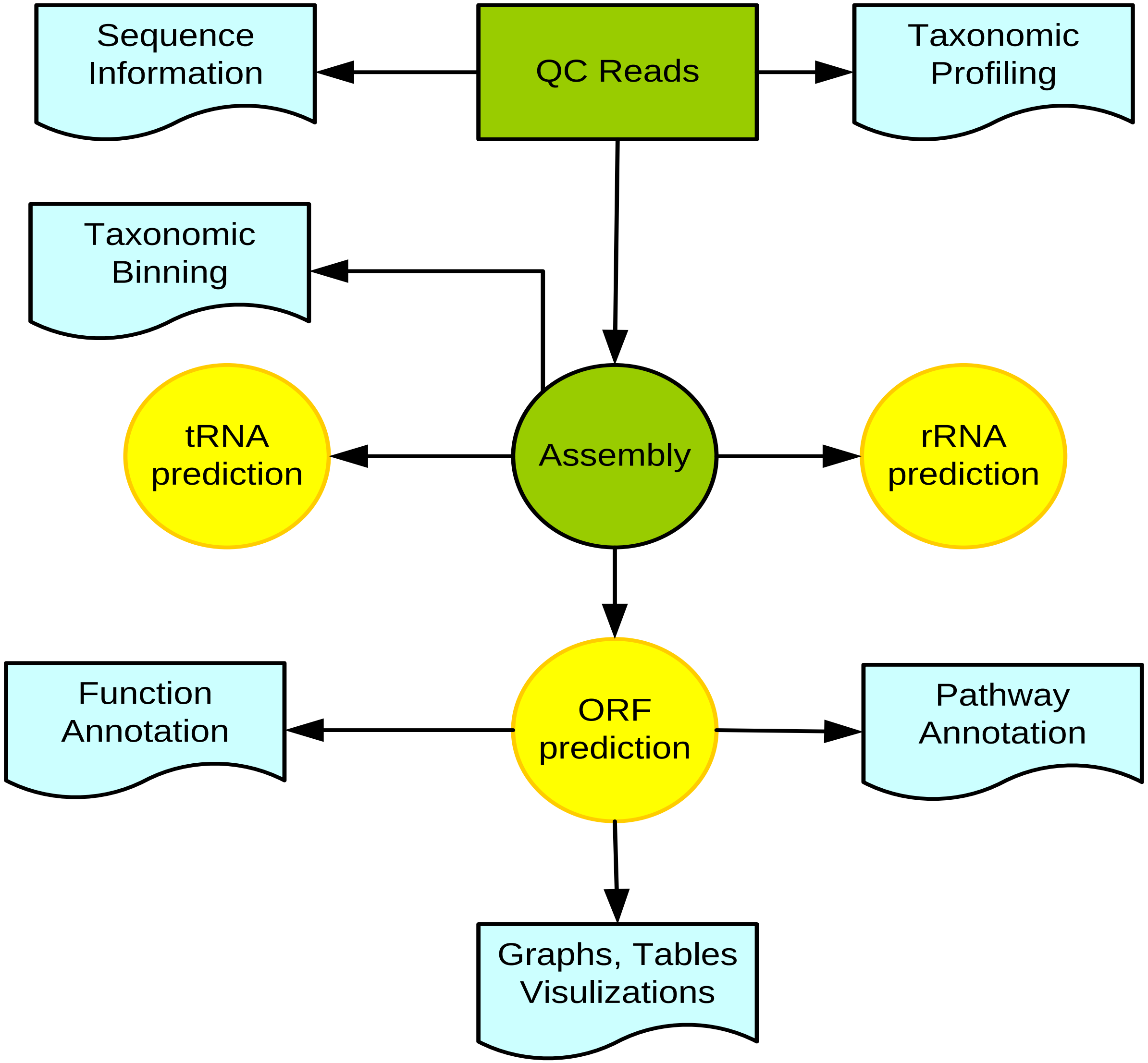
HMP Metagenome Assembly Pipeline



Quality Controlled Reads
454
Sanger Fosmid (Paired End)
Illumina (Paired End)
Newbler
Multiple kmer assembling (59, 63, 67, 71, 75, 79, 83, 87) SOAPDenovo
Re-assembly the short contigs (200-1999bp) generated in SOAPDenovo using OLC assemblers (Newbler)
Contigs
454 contigs + singletons
Long contigs + contigs and singletons generated in the first phase merging using Newbler
Merged contigs
Minimus2
Supercontigs
Scaffolding tools (Bambus)

- Iterative assembly merging

HMP Metagenome Analysis Flowchart



Example: Tailings Pond Composition

Sand, clay, fines (< 44 μm)
Bitumen
Alkaline water
Hydrocarbon diluent
Naphthenic acids
Metals (V, Ni, Sr)

Microorganisms!

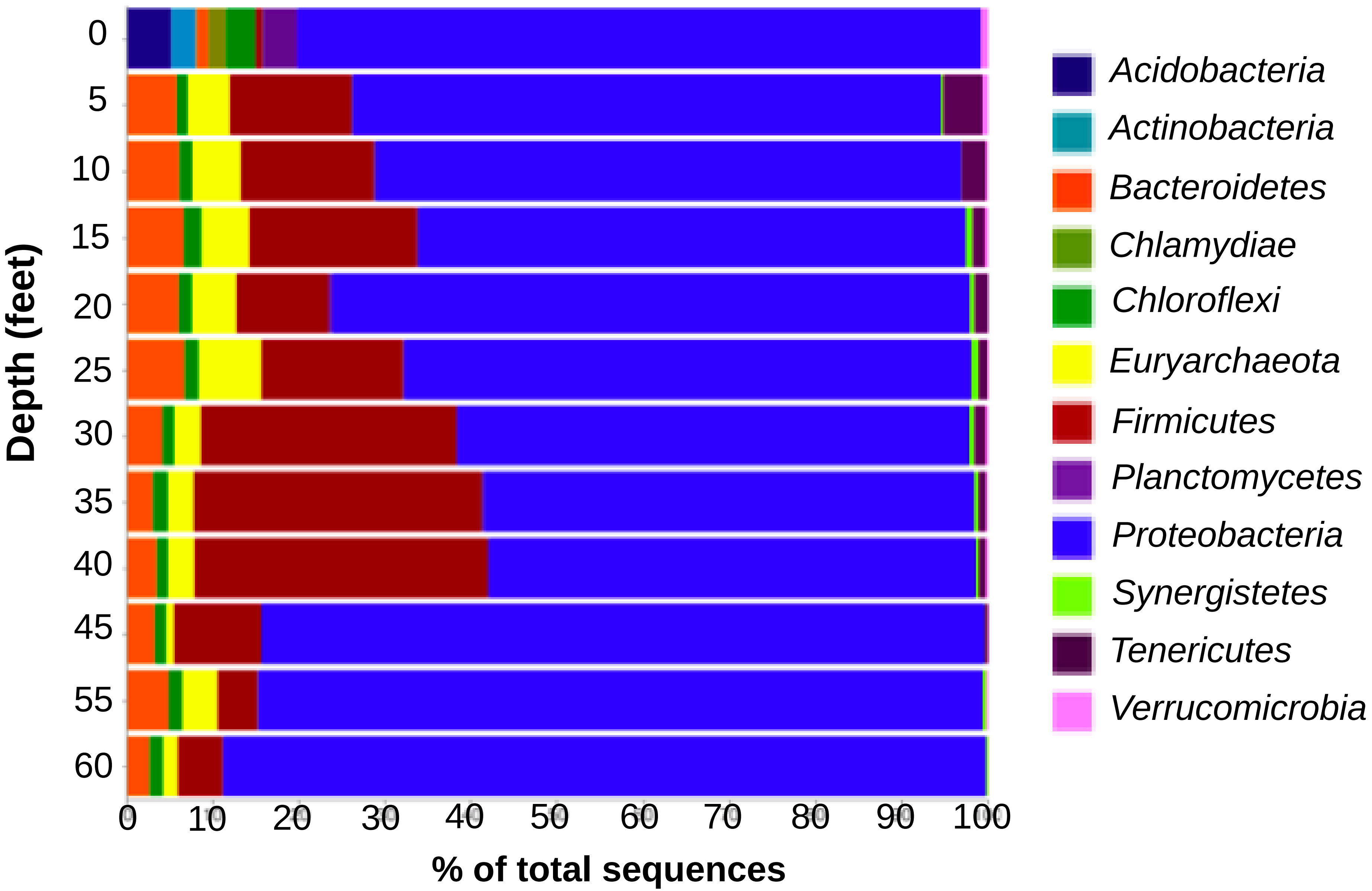
Aerobes at surface
Anaerobes at depth

Depth Profile of an Active Tailings Pond

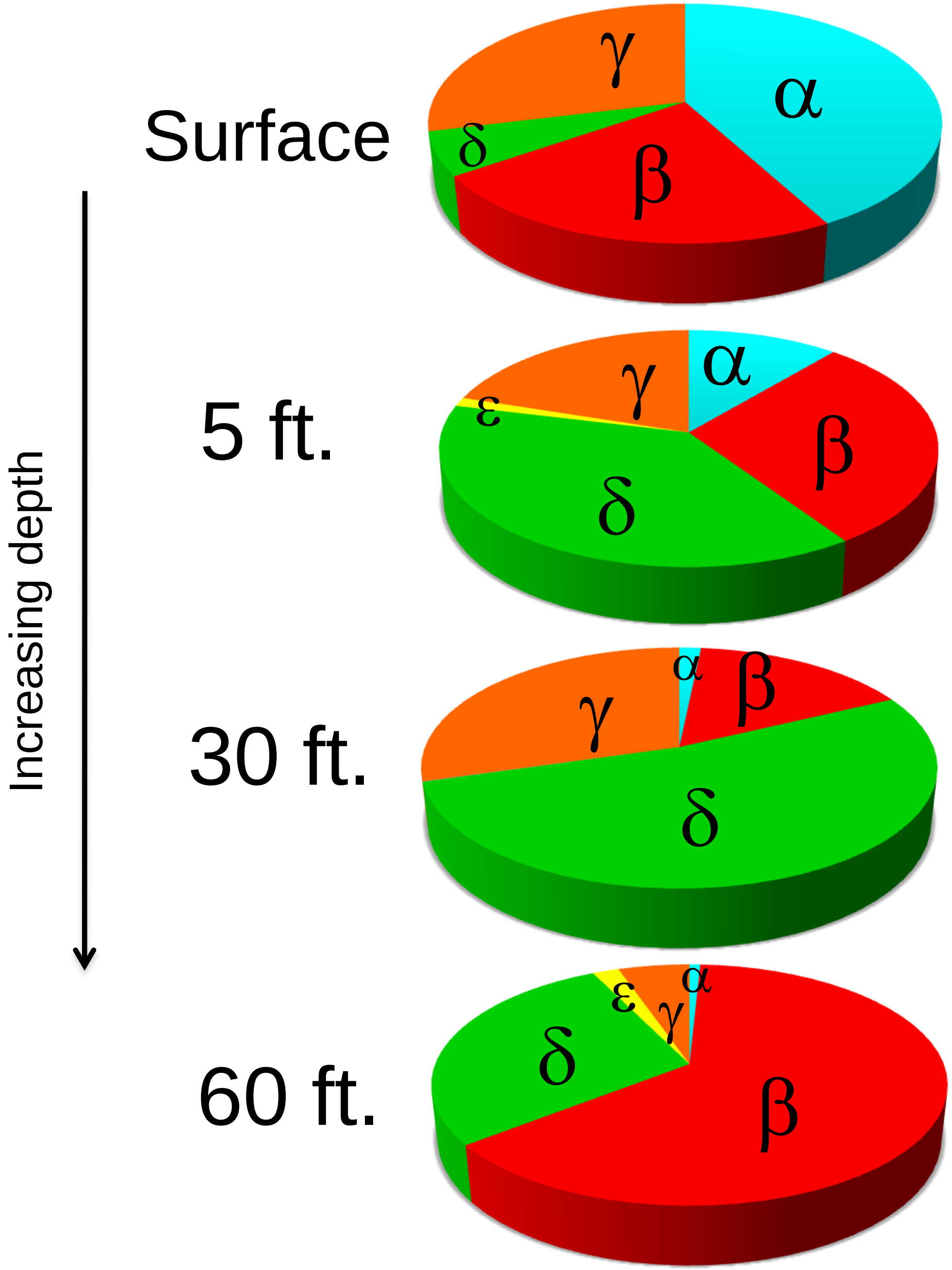
- The pond in operation since 2004
- Sampled in October 2008
- Samples collected from surface → 60 ft. deep
- “Soft” or pre-consolidated (pre-CT) tailings
- Pond is routinely treated with gypsum ($CaSO_4$)

Meddőhányó

Overview of Sequences Found at Phylum Level as a Function of Depth



Proteobacteria



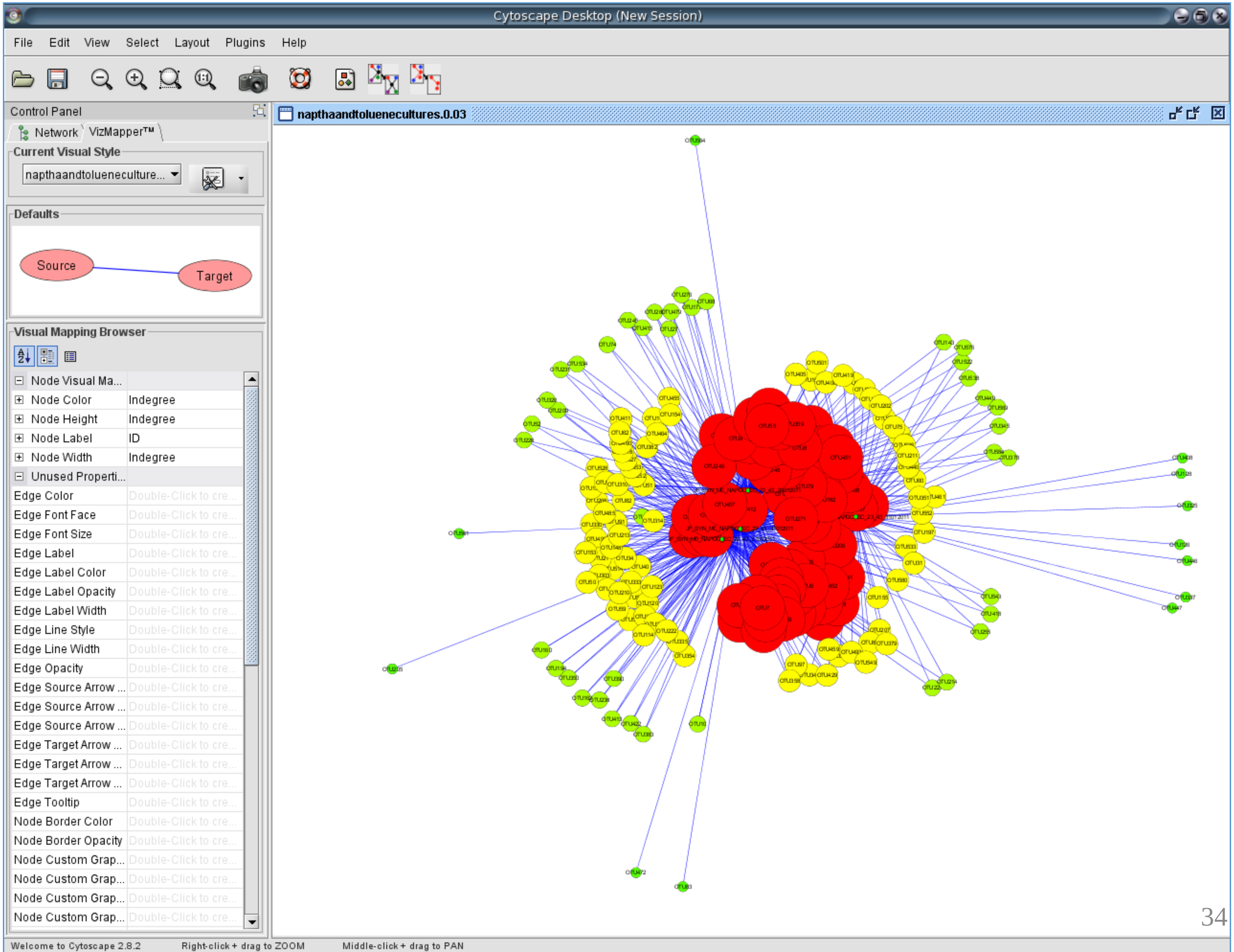
α: *Hypomicrobium*
β: *Methyloversatilis*
δ: *Pelobacter*
γ: *Hydrocarboniphaga*

α: *Porphyrobacter*
β: *Rhodoferax, Hydrogenophaga*
δ: *Syntrophus, Desulfocapsa*
γ: *Coxiella*

β: *Rhodoferax, Azoarcus*
δ: *Syntrophus, Geobacter, Desulfatibacillum*
γ: *Coxiella*

β: *Brachymonas*
δ: *Syntrophus, Desulfuromonas*
γ: *Coxiella*

Degrees of OTU sharing among samples



Current Visual Style: napthaandtolueneculture...

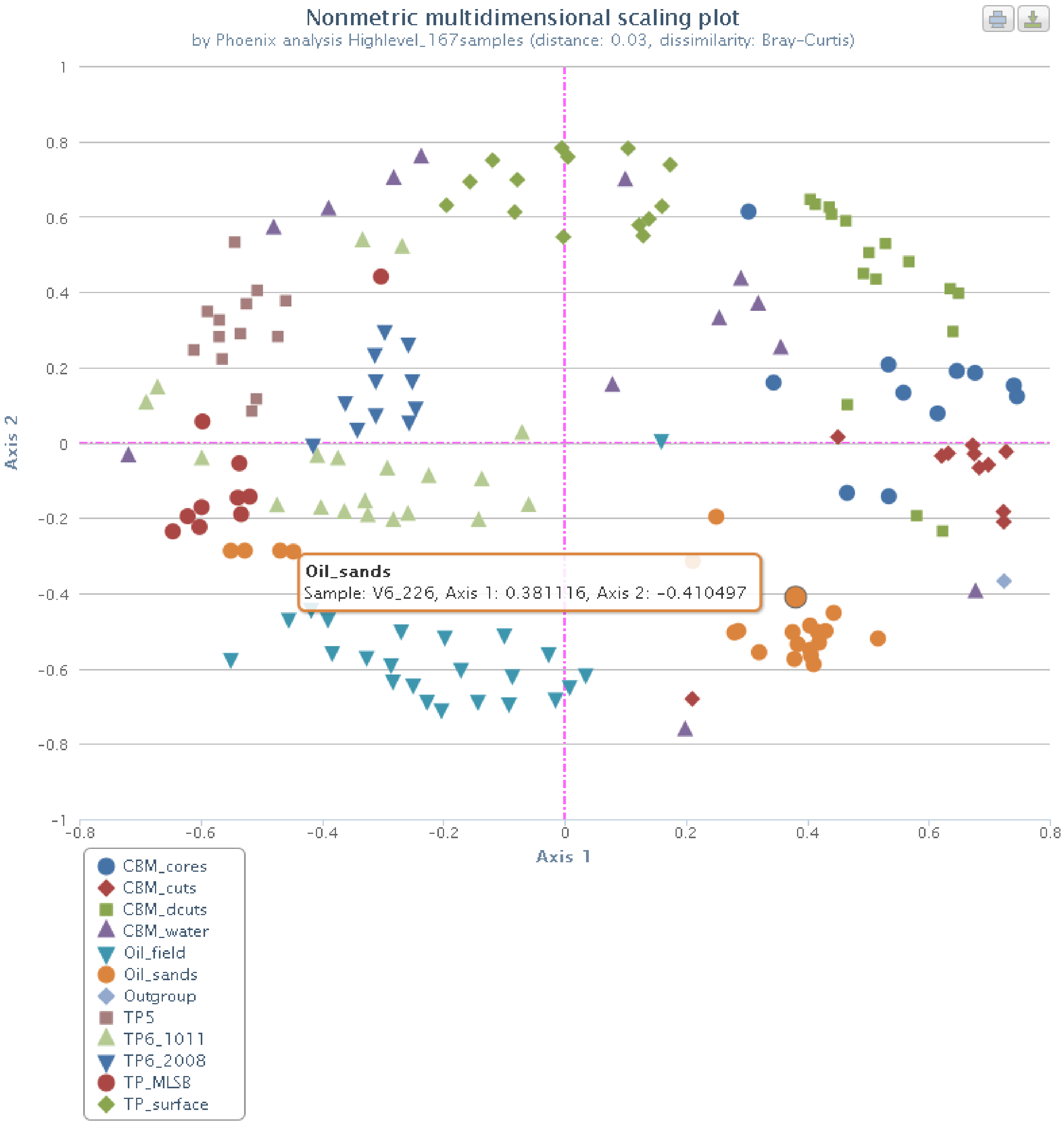
Defaults

Visual Mapping Browser

⊟ Node Visual Ma...	
⊞ Node Color	Indegree
⊞ Node Height	Indegree
⊞ Node Label	ID
⊞ Node Width	Indegree
⊟ Unused Properti...	
Edge Color	Double-Click to cre...
Edge Font Face	Double-Click to cre...
Edge Font Size	Double-Click to cre...
Edge Label	Double-Click to cre...
Edge Label Color	Double-Click to cre...
Edge Label Opacity	Double-Click to cre...
Edge Label Width	Double-Click to cre...
Edge Line Style	Double-Click to cre...
Edge Line Width	Double-Click to cre...
Edge Opacity	Double-Click to cre...
Edge Source Arrow ...	Double-Click to cre...
Edge Source Arrow ...	Double-Click to cre...
Edge Source Arrow ...	Double-Click to cre...
Edge Target Arrow ...	Double-Click to cre...
Edge Target Arrow ...	Double-Click to cre...
Edge Target Arrow ...	Double-Click to cre...
Edge Tooltip	Double-Click to cre...
Node Border Color	Double-Click to cre...
Node Border Opacity	Double-Click to cre...
Node Custom Grap...	Double-Click to cre...
Node Custom Grap...	Double-Click to cre...
Node Custom Grap...	Double-Click to cre...
Node Custom Grap...	Double-Click to cre...

Welcome to Cytoscape 2.8.2 Right-click + drag to ZOOM Middle-click + drag to PAN

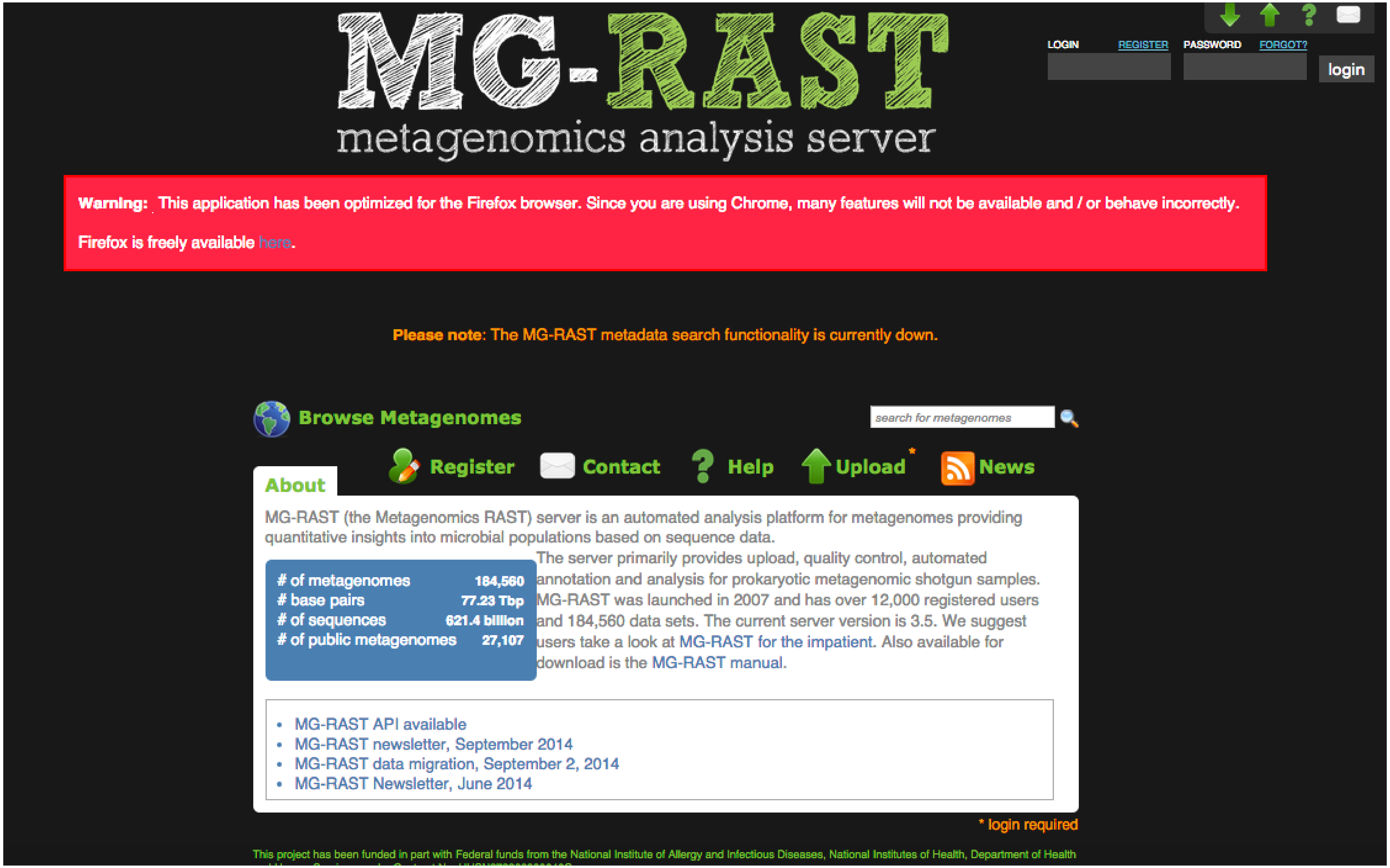
Phoenix 2 NMDS



Metagenomic Analysis Related Sites

- MG-RAST (Metagenomic Analysis Server)
  - https://metagenomics.anl.gov
  - 27,003 public metagenomes
- IMG (Integrated Microbial Genomes)
  - https://img.jgi.doe.gov
- SILVA and GreenGenes (Ribosomal RNA Collection)
  - http://www.arb-silva.de
  - http://greengenes.lbl.gov/cgi-bin/nph-index.cgi
  - Ribosomal RNA databases and tools

MG-RAST Server



Warning: This application has been optimized for the Firefox browser. Since you are using Chrome, many features will not be available and / or behave incorrectly.

Firefox is freely available here.

Please note: The MG-RAST metadata search functionality is currently down.

MG-RAST (the Metagenomics RAST) server is an automated analysis platform for metagenomes providing quantitative insights into microbial populations based on sequence data.

# of metagenomes	184,560
# base pairs	77.23 Tbp
# of sequences	621.4 billion
# of public metagenomes	27,107

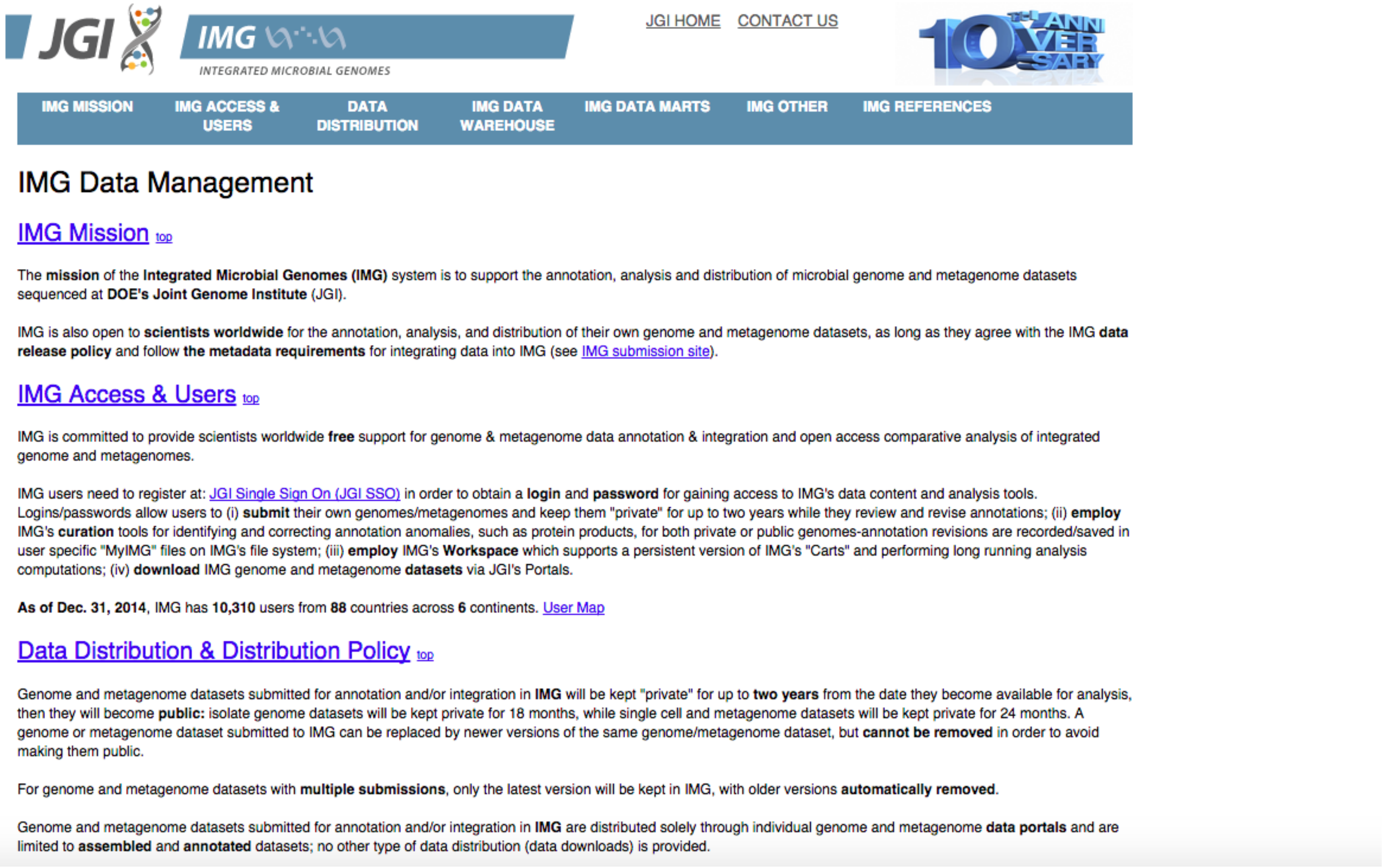
The server primarily provides upload, quality control, automated annotation and analysis for prokaryotic metagenomic shotgun samples. MG-RAST was launched in 2007 and has over 12,000 registered users and 184,560 data sets. The current server version is 3.5. We suggest users take a look at MG-RAST for the impatient. Also available for download is the MG-RAST manual.

- MG-RAST API available
- MG-RAST newsletter, September 2014
- MG-RAST data migration, September 2, 2014
- MG-RAST Newsletter, June 2014

* login required

This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health

Integrated Microbial Genomes



IMG Data Management

IMG Mission top

The **mission** of the **Integrated Microbial Genomes (IMG)** system is to support the annotation, analysis and distribution of microbial genome and metagenome datasets sequenced at **DOE's Joint Genome Institute** (JGI).

IMG is also open to **scientists worldwide** for the annotation, analysis, and distribution of their own genome and metagenome datasets, as long as they agree with the IMG **data release policy** and follow **the metadata requirements** for integrating data into IMG (see IMG submission site).

IMG Access & Users top

IMG is committed to provide scientists worldwide **free** support for genome & metagenome data annotation & integration and open access comparative analysis of integrated genome and metagenomes.

IMG users need to register at: JGI Single Sign On (JGI SSO) in order to obtain a **login** and **password** for gaining access to IMG's data content and analysis tools. Logins/passwords allow users to (i) **submit** their own genomes/metagenomes and keep them "private" for up to two years while they review and revise annotations; (ii) **employ** IMG's **curation** tools for identifying and correcting annotation anomalies, such as protein products, for both private or public genomes-annotation revisions are recorded/saved in user specific "MyIMG" files on IMG's file system; (iii) **employ** IMG's **Workspace** which supports a persistent version of IMG's "Carts" and performing long running analysis computations; (iv) **download** IMG genome and metagenome **datasets** via JGI's Portals.

As of Dec. 31, 2014, IMG has **10,310** users from **88** countries across **6** continents. User Map

Data Distribution & Distribution Policy top

Genome and metagenome datasets submitted for annotation and/or integration in **IMG** will be kept "private" for up to **two years** from the date they become available for analysis, then they will become **public:** isolate genome datasets will be kept private for 18 months, while single cell and metagenome datasets will be kept private for 24 months. A genome or metagenome dataset submitted to IMG can be replaced by newer versions of the same genome/metagenome dataset, but **cannot be removed** in order to avoid making them public.

For genome and metagenome datasets with **multiple submissions**, only the latest version will be kept in IMG, with older versions **automatically removed**.

Genome and metagenome datasets submitted for annotation and/or integration in **IMG** are distributed solely through individual genome and metagenome **data portals** and are limited to **assembled** and **annotated** datasets; no other type of data distribution (data downloads) is provided.

What have we learnt?

- Metagenomics is the analysis of samples - usually environmental or gut microflora samples - with many thousand species
- Traditional approach uses one reference gene, 16S rRNA, amplified by PCR and NGS sequenced. Bacterial composition is obtained..
- In whole genome sequencing (WGS) the reads can be mapped to annotated genome sequences, bacterial composition and biological functions are obtained.
- Alternatively, WGS reads can be assembled at large computational overheads which makes analysis more accurate.
- Main programs: MG RAST, Megan, Mothur