

Analysis of Oral Microbiome Sequence Data



Part I: DOE bioinformatics resources for metagenomics analysis

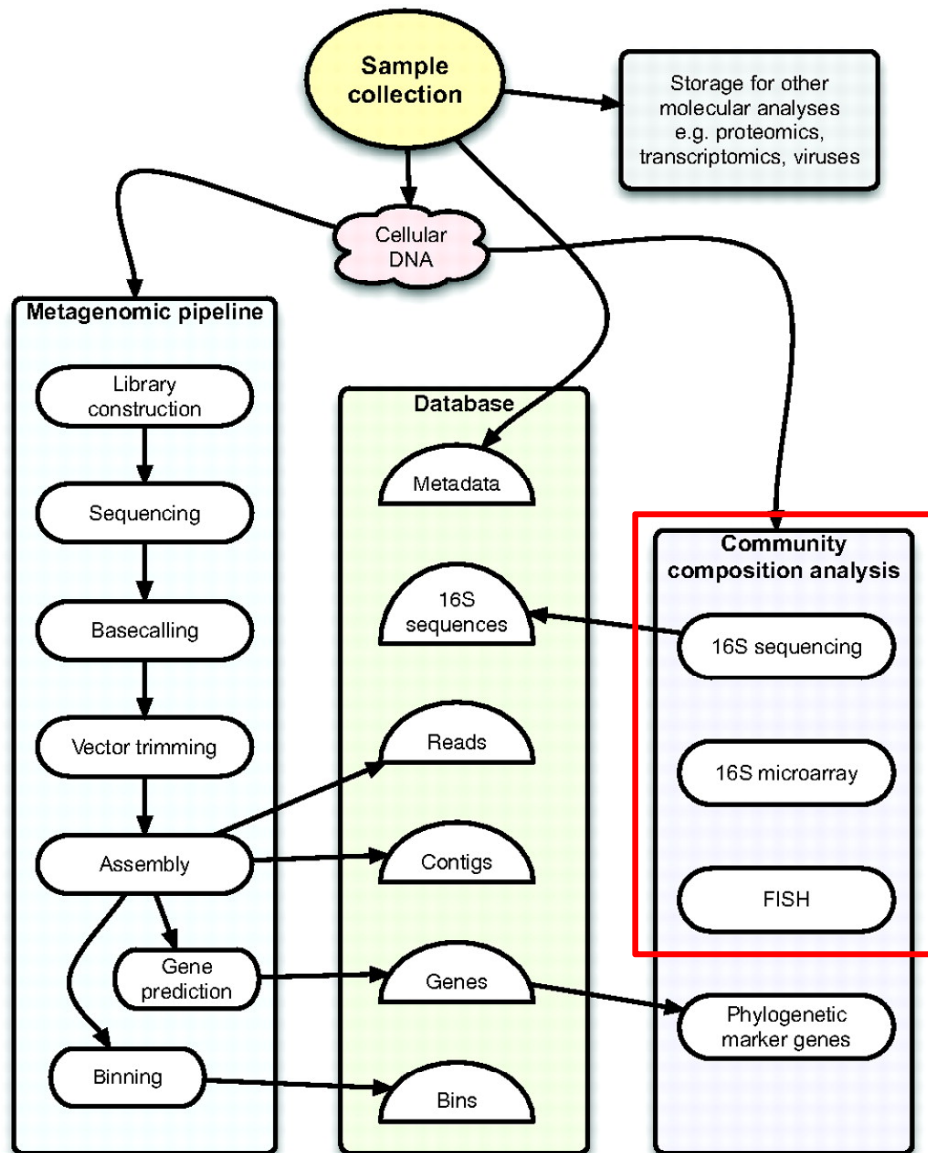
Gary Xie

April 3rd, 2009

Outline

- Premetagenome Community Composition Profiling
- Sequence Processing
- **Data Analysis**
 - **Postsequencing Community Composition Estimates**
 - **Binning**
 - **Analyzing Dominant Populations**
 - **Gene-Centric Analysis**

Typical workflow for Sanger-based metagenomic projects at the JGI



•Premetagenome Community Composition Profiling

- 4 approaches

- 3 resources in 16S rRNA sequence analysis

- Statistical Toolbox

- Sequence Processing

- Data Analysis

Premetagenome Community Composition Profiling

Limitations:

- 16s rRNA
 - 16s copy number
 - PCR induced bias
 - JGI: ribosomal panel-384-well plate
- Phylochip
 - Limited by probes on the microarray
 - Relative abundance estimation
- Pyrotags (Pyrosequencing PCR amplified 16s rRNA)
 - Reduced phylogenetic resolution (100-200bp 454 reads)
 - Dependent on reference 16s database
- FISH (using group-specific 16S rRNA gene-targeted probes)
 - Number of fluorochromes, availability, and cost
 - Target broader groups such as domains or phyla

Resources in 16S rRNA sequence analysis

- Various 16S rRNA databases and resources available:
 - RDPII Project
 - Greengenes
 - SILVA rRNA database project

Available 16S rRNA databases (January 2008):

Database	B ¹	A ²	E ³	Alignment positions	Number of sequences
RDPII (http://rdp.cme.msu.edu/)	+	-	-	~53,000	~472,000 ⁴
ARB SILVA (http://www.arb-silva.de/)	+	+	+	~46,000	~504,000 ⁵
greengenes (http://greengenes.lbl.gov/)	+	+	(+)	7,682	~180,000 ⁶

¹ Bacteria; ² Archaea; ³ Eukarya; ⁴ including partial sequences; ⁵ >300 nt; ⁶ >1250 nt

Current database status

	SSU Parc	SSU Ref	LSU Parc	LSU Ref
Minimal length	300	1200/900	300	1900
Quality filtering	basic	strong	basic	strong
Guide Tree	no	yes	yes	yes
Release version	93	93	93	93
Release date	01.02.08	01.02.08	01.02.08	01.02.08
Aligned rRNA sequences	566,047	224,967	102,021	10,128

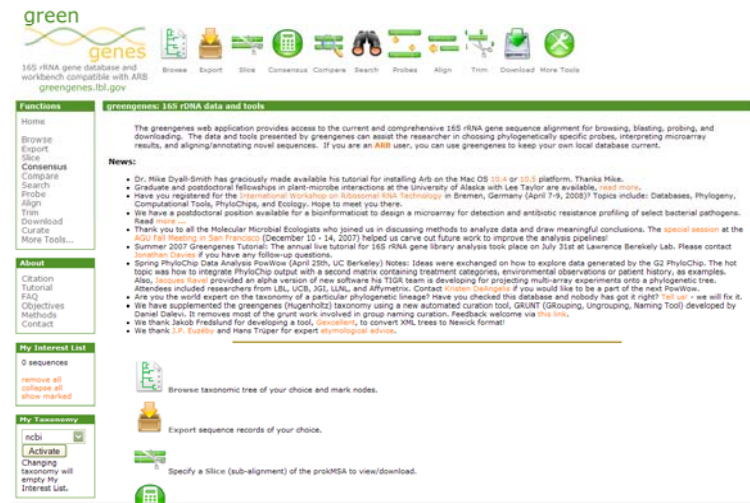
<http://www.arb-silva.de/>

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

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

The Green Genes

- Run by LLNL
- Hugenholtz alignment
- Not limited to size
- Includes phyla with no cultivated representatives
- SSU rRNA gene sequences from Bacteria, Archaea, & Eukarya
- Artificial “group names” are made
- Allows:
 - Detection of chimeric sequences
 - Trimming of chimeric sequences
 - Visualization of 16S sequence alignments
 - Browsing of sequences
 - Searching of sequences
 - Aligning of sequences
 - Probe searches



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

SILVA rRNA database project

- 16S/23S or 18S/28S rRNA databases for (Eukarya, Archaea, and Bacteria)
- Developed by German consortium (MPI, TUM, and Ribocon)
- Fully compatible with popular software (ARB) in 16S rRNA analysis
- Able to download an Arb database with reference sequences alignment & reference helix
- Cultivated & non-cultivated representatives
- Limited to 50 sequences; have to pay Ribocon for >50
- Allows:
 - Browsing of sequences
 - Searching of sequences
 - Aligning of sequences
 - Designing of FISH probes

silva 

Home Browser Search Aligner List Download Background Living Tree FAQs Contact FISH & Probes Shop rRNA workshop

SILVA

Welcome to the SILVA rRNA database project
A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data, free for academic use. SILVA provides comprehensive, quality checked and regularly updated databases 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). The SILVA databases can be used for rapid rRNA sequence retrieval, phylogenetic analysis, and the design of probes and primers for diversity analysis. They are fully compatible to the software package ARB. For more background information → [Click here](#)

ARB
The software package ARB represents a graphically-oriented, fully-integrated package of cooperating software tools for handling and analysis of sequence information. It was originally developed for the analysis of ribosomal RNA sequences (16S/18S & 23S/28S) but can easily be used for any kind of sequence information. The ARB project has been started more than 12 years ago by Wolfgang Ludwig at the Technical University in Munich, Germany, see [/J](#) www.arb-home.de

News

Latest News
18.02.08
rRNA technology Workshop Handout ready for download [\[more\]](#)
07.02.08
Registration for rRNA Workshop closed
Many thanks for the overwhelming feedback! With more than a 100% of over-subscription and... [\[more\]](#)
30.01.08
SILVA 93 now online
Have a look at the latest statistics... [\[more\]](#)
[go to Archive](#) →

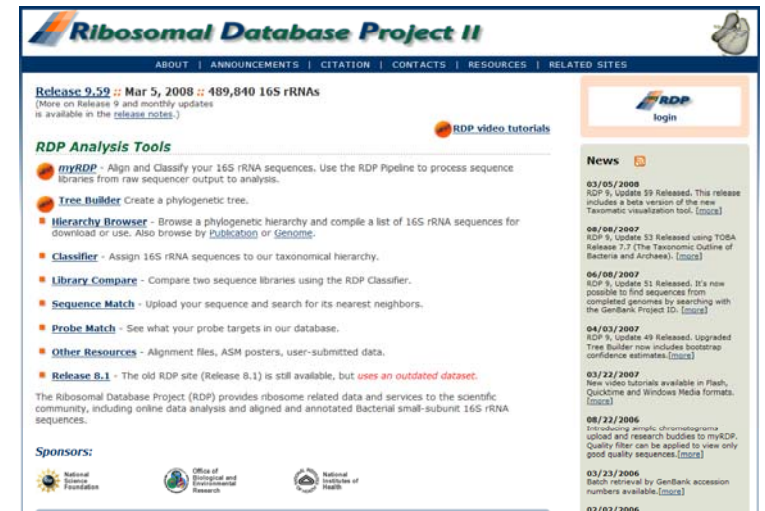
Current database status

	SSU Parc	SSU Ref	LSU Parc	LSU Ref
Minimal length	300	1200/900	300	1900
Quality filtering	basic	strong	basic	strong
Guide Tree	no	yes	yes	yes
Release version	93	93	93	93
Release date	01.02.08	01.02.08	01.02.08	01.02.08
Aligned rRNA sequences	566,047	224,967	102,021	10,128

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

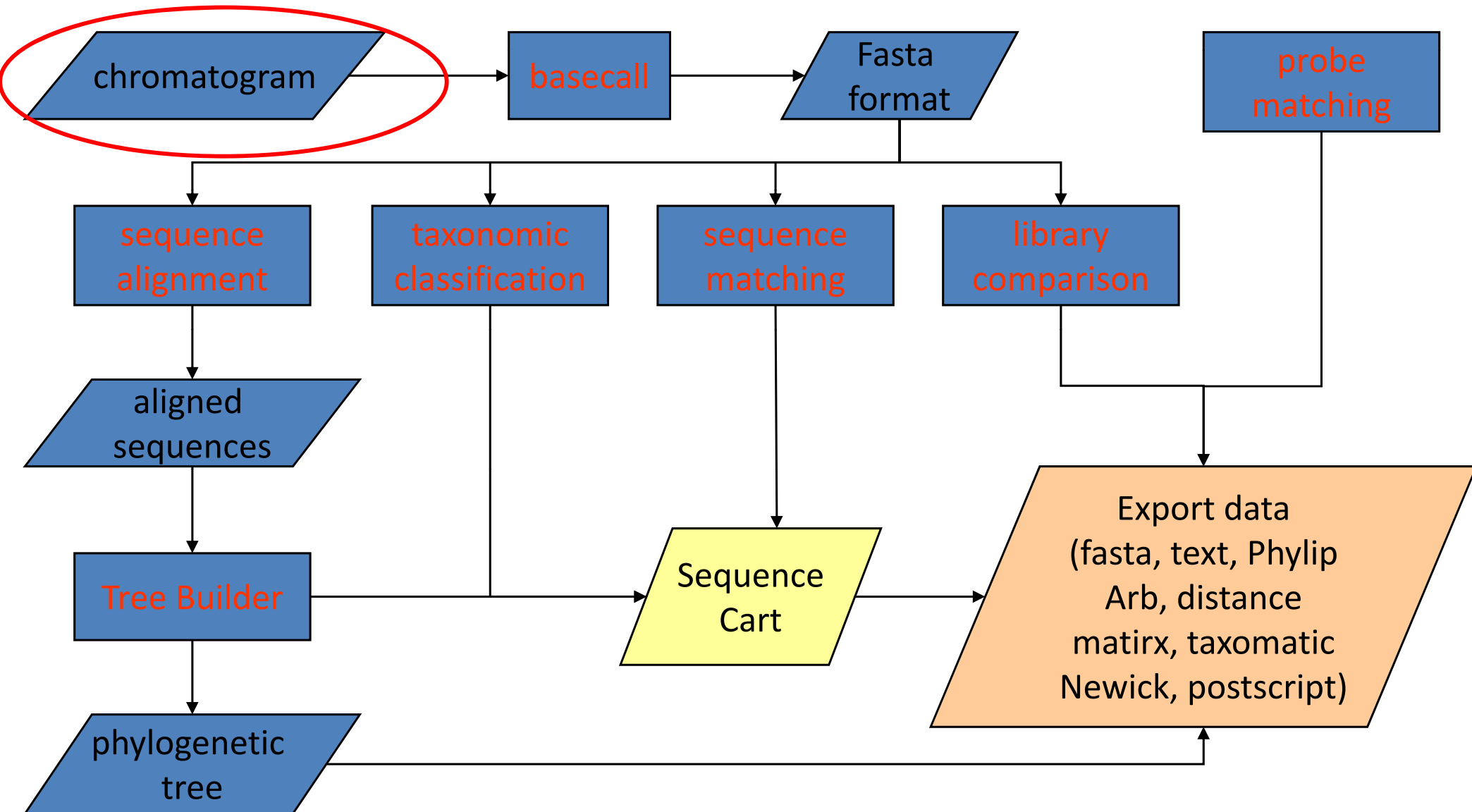
RDPII Project

- Run by MSU
- Not limited to size
- RDP taxonomy is based primarily on cultivated representatives
- No fungal sequences available
- Able to align sequences & obtain a distance matrix, however there is no seed or reference helix information
- Alignment can vary among batches of sequences
- Allows:
 - Users to create accounts and upload sequences for analysis
 - Creation of phylogenetic trees quickly
 - Raw chromatograms to be processed
 - Comparison of different libraries
 - Browsing of sequences
 - Searching of sequences
 - Aligning of sequences
 - Probe searches



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

RDPII sequence analysis flowchart



Uploading chromatograms

Create a New Library Run

Logged in as [jsaw@lanl.gov]

summary **dotur/estimateS** help

Return to myRDP

Fill in the Mandatory Fields Below

* indicates mandatory field

* Gene: Bacteria 16S rRNA

NOTE: Library Name Abbreviation cannot exceed 5 characters

* Library Name Abbrev: test2

* Library Name: test2

NOTE: If your vector is not in the list, please send an email to rdpstaff@msu.edu.

* Vector: Invitrogen pCR2.1

Account Number:

Project:

Folder:

Primer Name:

Note(s): second testing

Define well mapping Upload data without well mapping

Return to RDP »

To run dotur/estimateS programs on samples (to determine species richness)

Choose gene (16S or unknown)

Fill in required information

Choose cloning/sequencing vector

To define reads according to well position

To upload without well position

Uploading chromatograms

Data Loader Status

Logged in as [jsaw@lanl.gov]

Start uncompressing file....Done
Start checking vector....Done
Start running Phred....Done
Start running Lucy....Done
Start loading data to database....

[refresh](#)

Loader Complete

Logged in as [jsaw@lanl.gov]

[summary](#)

[dotur/estimateS](#)

[help](#)

[Return to myRDP](#)

Your data has been loaded into the database.

View Completed Library Runs

(Clicking the following link will open a new window)

1. [LR00002171](#)

[Return to RDP »](#)

Downloading data

myRDP

RDp HOME | BROWSER | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | TREE | myRDP | seqCART

Welcome, Jimmy Saw! [\[account info\]](#) [\[logout\]](#) [overview](#) | [upload](#) | [download](#) | [pipeline](#) | [help](#)

Download Format Options

Your seq cart contains 189 sequences:

- 9 myRDP sequences, 9 are aligned *
- 180 public sequences

If you do not wish to include public sequences in your download, please deselect those sequences from the [hierarchy browser](#).

To display Genbank ID or RDP ID

- ☒ display Genbank ACCNO / seq name
- ☐ display RDP id

Fasta or Phylip formats (aligned or unaligned)

- ☒ fasta
- ☐ phylip

Arb compatible format

- ☐ navigation tree (ARB compatible) [more ARB help](#)

To download distance matrix

- ☐ distance matrix (DNADist format)

To display sequences in Taxomatic format

- ☐ display on Taxomatic

☐ aligned

- ☐ show mask/structure
- ☐ unaligned

☐ uncorrected

- ☐ jukes cantor

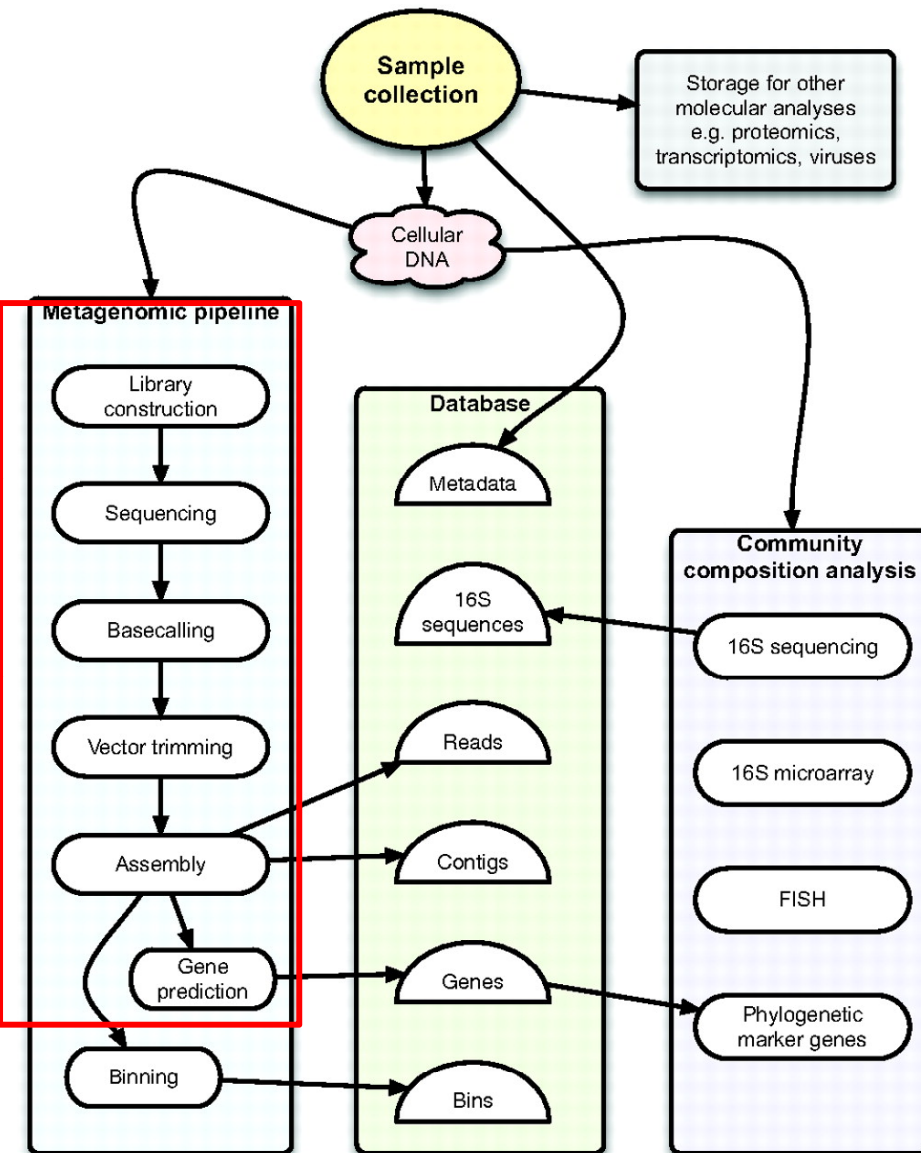
DOWNLOAD

[Note: Phylip, Distance Matrix, Navigation Tree and Taxomatic are limited to 2000 sequences.]

Statistical Toolbox

- OTU-based approach
 - DOTUR: define & count OTU
(distance matrix as input)
 - SONS: measuring overlap between communities
(OTU designation as input)
 - Estimate S - beta diversity index
(Need abundance file for each OTU)
- Community-based approach
 - S-LIBSHUFF (distance matrix as input)
 - Treeclimber (phylogenetic tree as input)
 - Unifrac (phylogenetic tree as input)

Typical workflow for Sanger-based metagenomic projects at the JGI



- Premetagenome Community Composition Profiling

- **Sequence Processing**

 - **Metagenomic assembly**

 - **Gene Prediction**

 - **Most popular CDS-finding tools**

 - **2 Full-service servers (IMG-M/ER, MG-RAST)**

- Data Analysis

Typical workflow for Sanger-based metagenomic projects of bacterial and archaeal communities at the JGI

Basecalling:

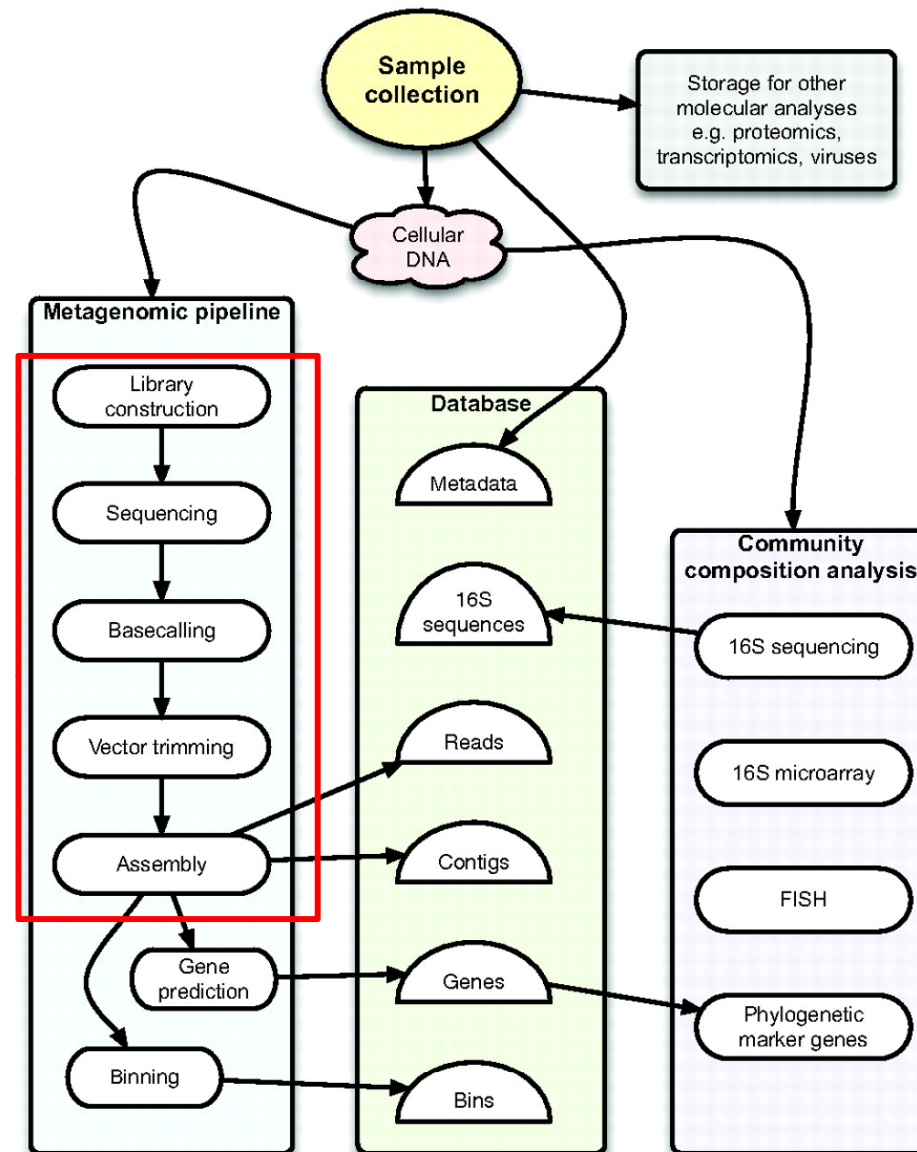
- Phred
- Paracel's Trace Tuner
- ABI's KB

Vector Screening:

- Cross_match
- LUCY
- Vector_clip

Assembly:

- PGA
- Arachne
- Phrap
- Celera



Metagenomic Assembly (Sanger) at JGI

-Use Trimmer (Lucy etc) to treat reads PRIOR to assembly

-PGA work better with paired reads information and produce better assemblies

-Not using phrap for metagenomic projects

Gene Prediction

evidence-based (ORFs with translations homologous to the known proteins are CDSs)

Blast, CRITICA, Orpheus

- Advantages: finds “unusual” genes (e. g. horizontally transferred); relatively low rate of false positive predictions
- Limitations: cannot find “unique” genes; low sensitivity towards short genes; prone to propagation of false positive results of *ab initio* annotation tools

ab initio (ORFs with nucleotide composition similar to CDSs are also CDSs)

fgenesB, Genemark, GLIMMER, MetaGene

- Advantages: finds “unique” genes; high sensitivity
- Limitations: often misses “unusual” genes; high rate of false positives

Most popular CDS-finding tools

- CRITICA
<http://www.es.embnnet.org/Doc/critica/index.html>
- Glimmer family (Glimmer2, Glimmer3, RBS finder)
<http://glimmer.sourceforge.net/>
- GeneMark family (GeneMark-hmm, GeneMarkS)
<http://exon.gatech.edu/GeneMark/>
- EasyGene
- AMIGENE
- **PRODIGAL (default JGI gene finder)**
<http://compbio.ornl.gov/prodigal/>

Combinations and variations of the above

- REGANOR (CRITICA + Glimmer3 + pre-processing)
- mORFind (Orpheus + CRITICA + GLIMMER)
- RAST (Glimmer2 + pre- and post-processing)

Metagenome Annotation

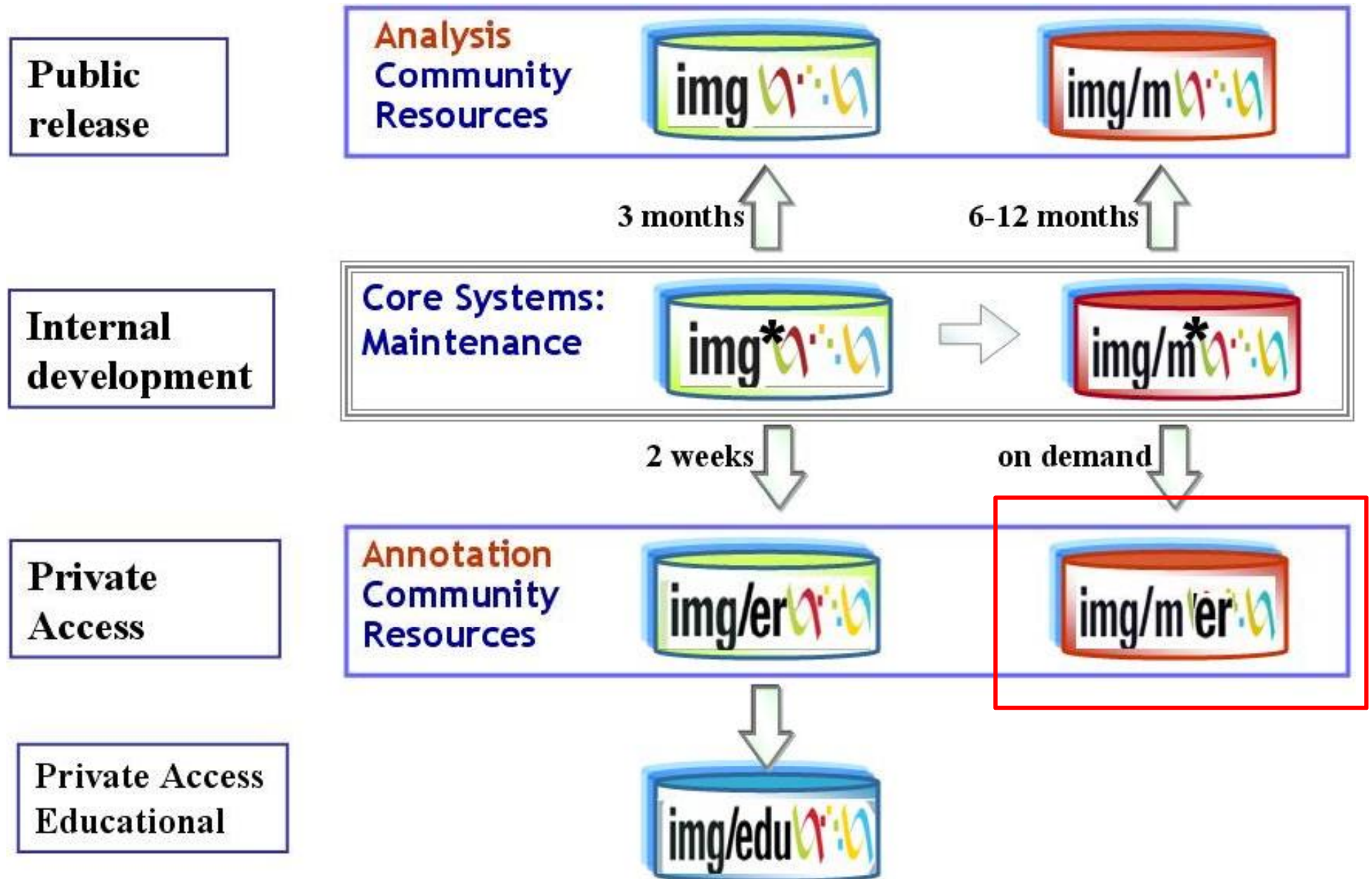
- BLASTx
- Fgenesb
<http://linux1.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb>
- GeneMark (GeneMark-hmm for reads, GeneMarkS for longer contigs)
<http://exon.gatech.edu/GeneMark/>
- MetaGene
<http://metagene.cb.k.u-tokyo.ac.jp/metagene/>
- GISMO
<http://www.cebitec.uni-bielefeld.de/groups/brf/software/gismo/>

Full-service servers

- IMG/M-ER – uses GeneMark for Sanger, proxygenes for 454
<http://img.jgi.doe.gov/submit>
- MG-RAST
<http://metagenomics.nmpdr.org/>

The Integrated Microbial Genomes (IMG) system

<http://img.jgi.doe.gov>





IMG ER & IMG/M ER dataset submission

Project Search Result

If you cannot find a project for your submission, go to [IMG-GOLD](#) to define a new project

Show Samples

Selection	ER Project ID	Project Display Name	GOLD
☐	10736	Extreme microbial communities from Yellowstone National Park	Gm00
☒	10739	Aquatic microbial communities from Yellowstone Bison Hot Spring Pool	Gm00
☐		Aquatic microbial communities from Yellowstone Bath	

Select Sample to Submit

Project 10739: Aquatic microbial communities from Yellowstone

If you cannot find a sample for your submission, either create a default sample or

Selection	ER Sample ID	Sample Display Name	Sample Site
☐	10022	2_050719S	Bison Pool, 6 m downstream of
☐	10023	5_050719P	Bison Pool, downstream photos
☐	10024	4_050719Q	Bison pool, downstream edge of
☐	10025	1_050719N	Bison Pool boiling source pool
☐	10026	3_050719R	Bison Pool, 11 m downstream of

(Number of rows displayed: 5)

Submit Sample

Create Default Sample

New Metagenome Dataset Submission

Submit isolate genome or metagenome for inclusion into IMG/M ER system.

Submission	
Target ER System (*)	IMG/M ER
ER Submission Sample ID	10022
Genbank/Fasta File (*)	
Or, file name in JGI file system (**)	
JGI Project ID	
EC computation by PRIAM needed? (*)	Yes
Gene calling needed? (*)	No
Product name computation needed? (*)	GeneMark Proxygenes
Is public in IMG database? (*)	No Unknown
Unique species code or locus tag prefix (* required if gene calling is needed)	
Topology	
Comments	
Additional Metagenome Files	
Contigs File	
Singlets File	
Binning File	
READS Files	



MG-RAST

Meta Genome Rapid Annotation
using Subsystem Technology

version 1.2

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

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. Although we do not guarantee a maximum turnover time, the current average processing time is about 24 hours. Currently the server handles 454 and Sanger sequence data. Data sets supplied by 454 can be uploaded directly.

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](#).

To be able to contact you once the computation is finished and in case user intervention is required, we request that you register with email address.

Login

Password

Login

[Forgot your password?](#)
[Register a new account](#)

Current Step **Upload Summary**

Please enter a project name and metagenome names for all uploaded files:

Project information:

Project Name:

mg-rast-test_87739d27295c4735ffa42383d2cbf440_jSC5POj.fasta

Metagenome Name:

Description:

Job ▲▼	Owner	ID	Name	Size (bp) ▲▼	Creation Date	Annotation Progress
2023	McNeil, Leslie		SID2303	0	2008-12-19	 [view details]
2022	McNeil, Leslie	4442110.3	SID2302	6954286	2008-12-19	 [view details]

Jobs Details #2023

» [Browse annotated metagenome in SEED Viewer](#)

» Available downloads for this job: Genbank export ▼ Download

» [Share this metagenome with selected users](#)

» [Make this metagenome publicly accessible](#)

» [Back to the Jobs Overview](#)



Genome Upload has been successfully completed.

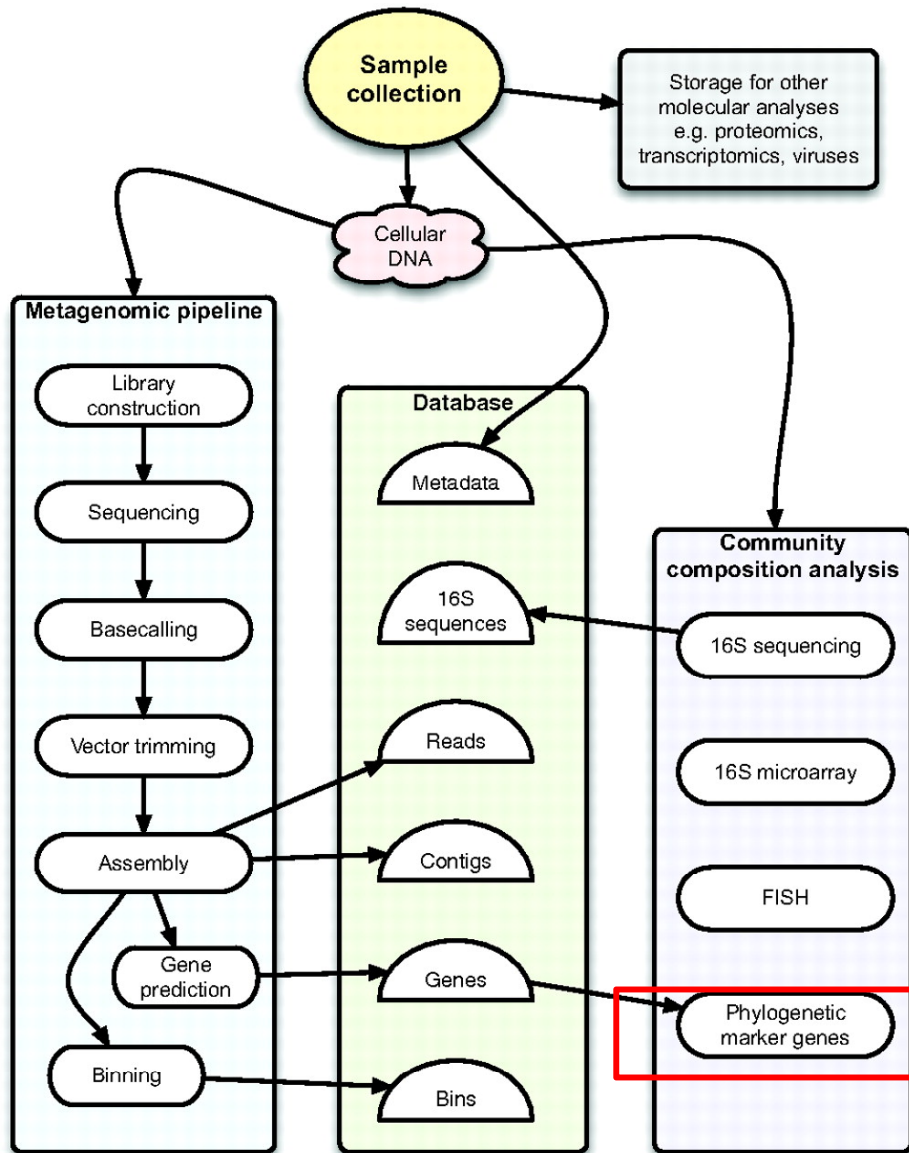


Enter an email address

Enter an email address:

Share job with this user or group

Typical workflow for Sanger-based metagenomic projects at the JGI



- Premetagenome Community Composition Profiling
- Sequence Processing
- Data Analysis

– Postsequencing Community Composition Estimates

- Binning
- Analyzing Dominant Populations
- Gene-Centric Analysis

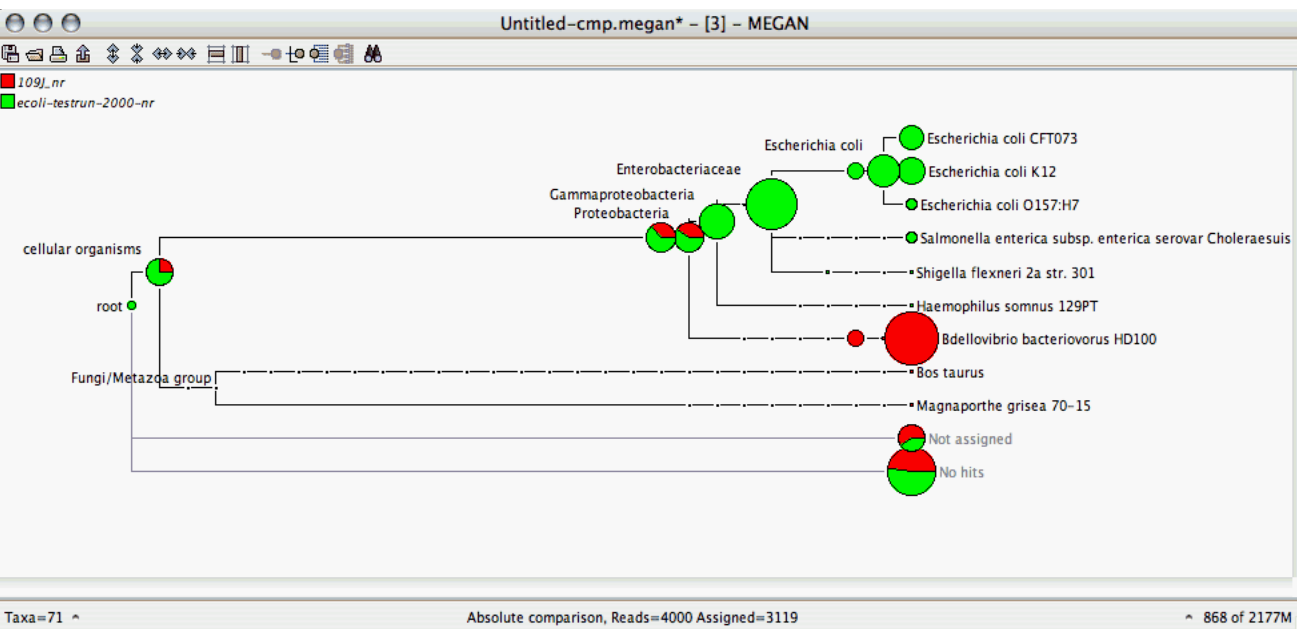
Problem of the BLAST-based composition estimation

- The poor representation of microbial diversity by sequenced isolates , often resulting in remote matches to phylogenetically distant organisms or the absence of any hits

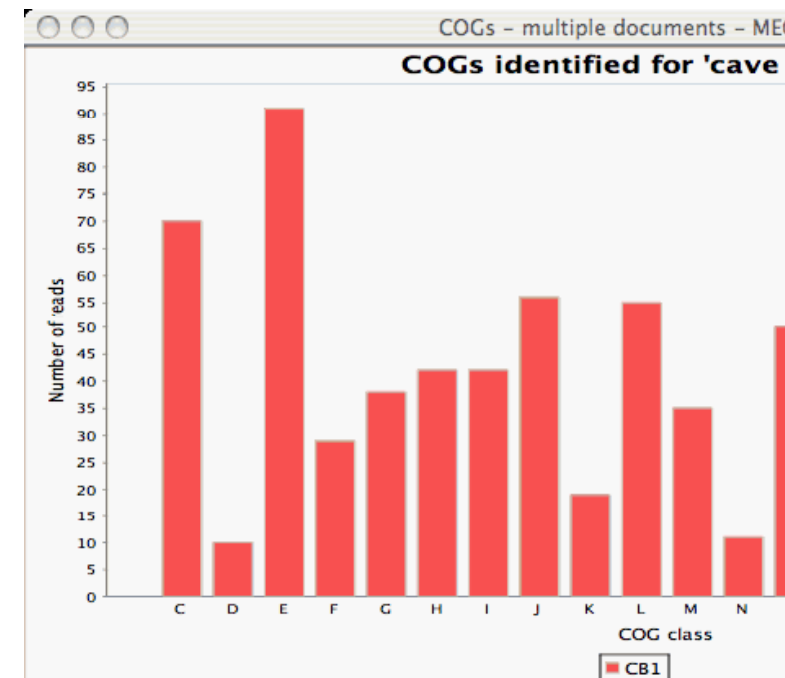
<http://www-ab.informatik.uni-tuebingen.de/software/megan>

MEGAN addresses this problem by assigning sequence fragments to the lowest common ancestor of the set of taxa that it hit in the comparison

Comparison and Analysis of Multiple Datasets

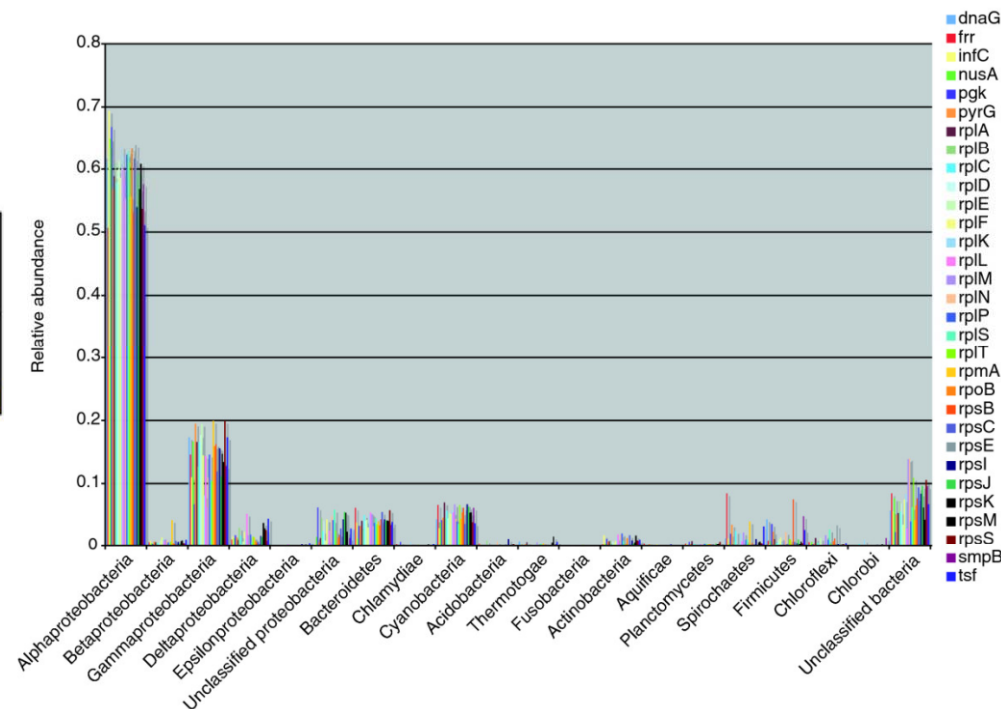
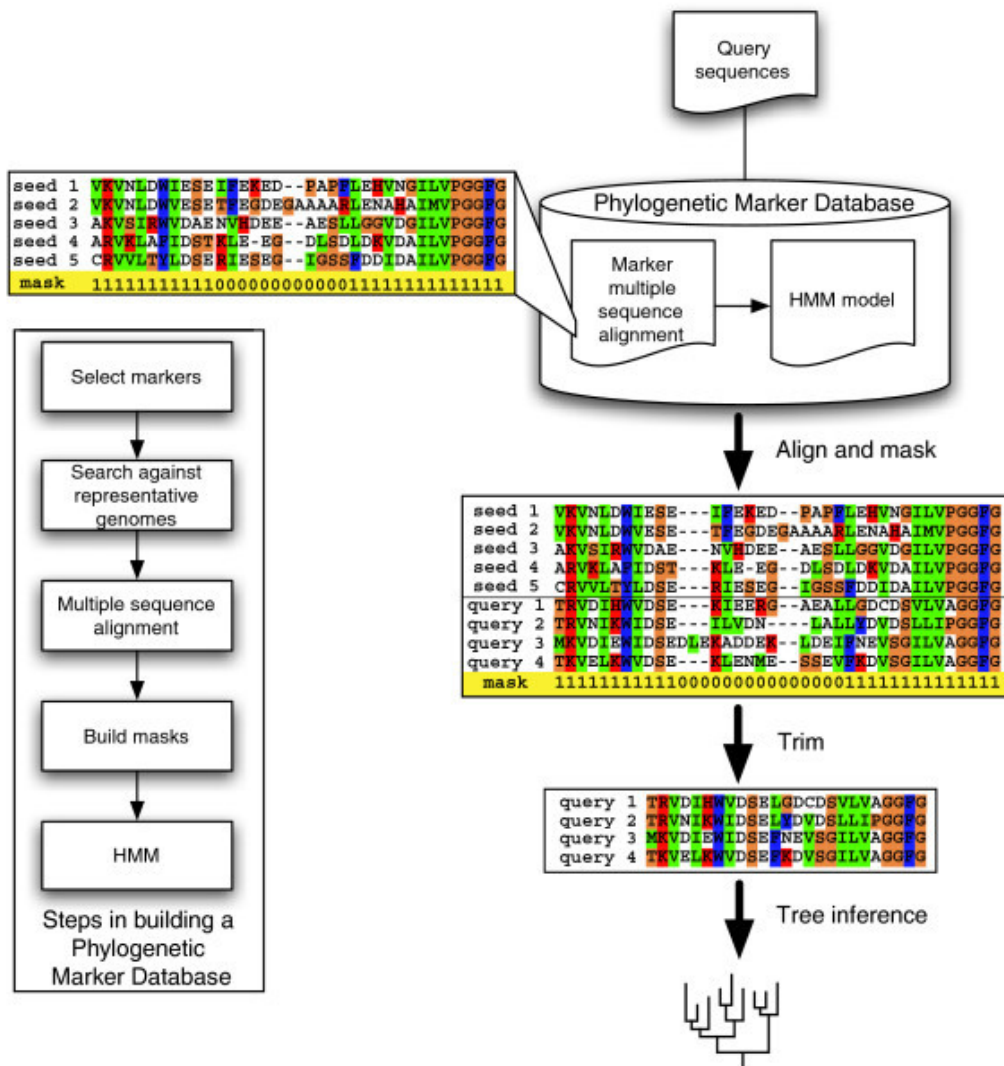


COG analysis



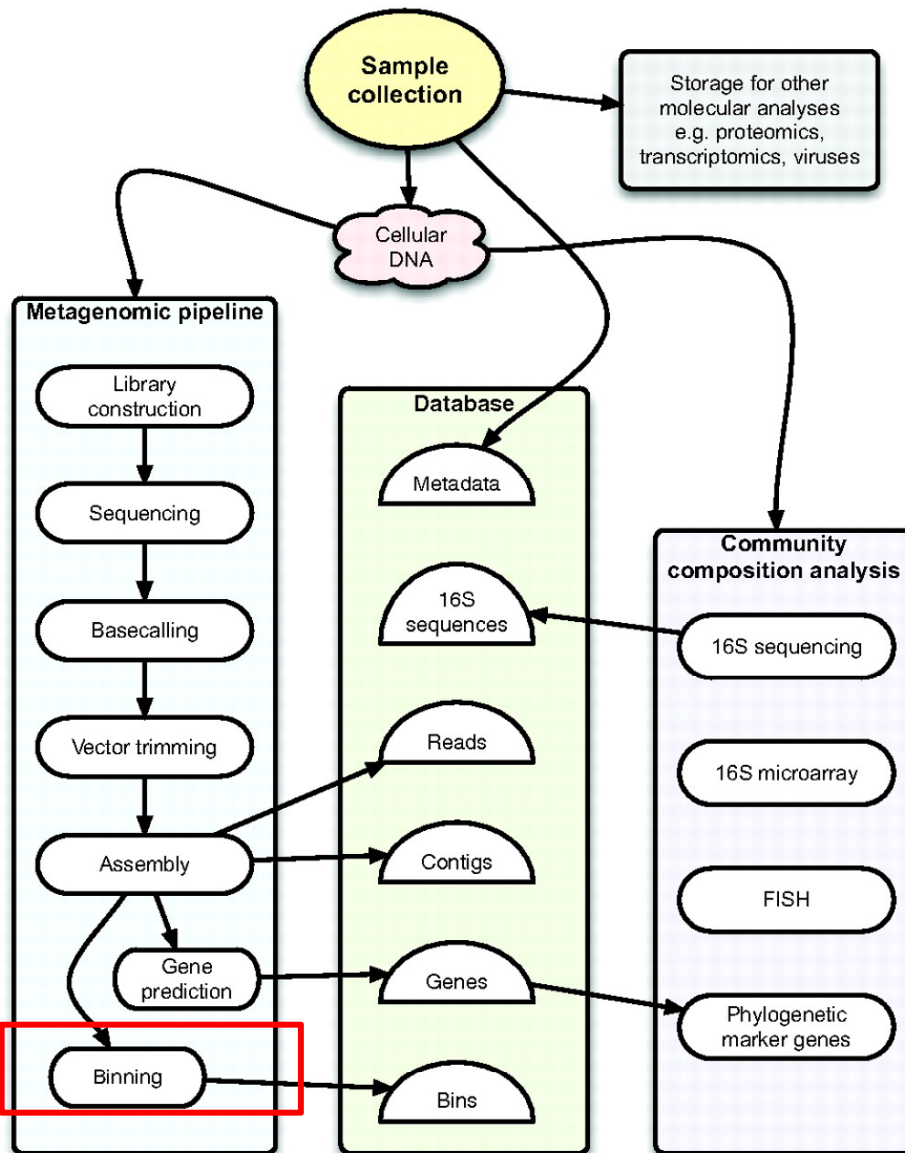
Automated pipeline for phylogenomic analysis (AMPHORA)

<http://bobcat.genomecenter.ucdavis.edu/AMPHORA/>



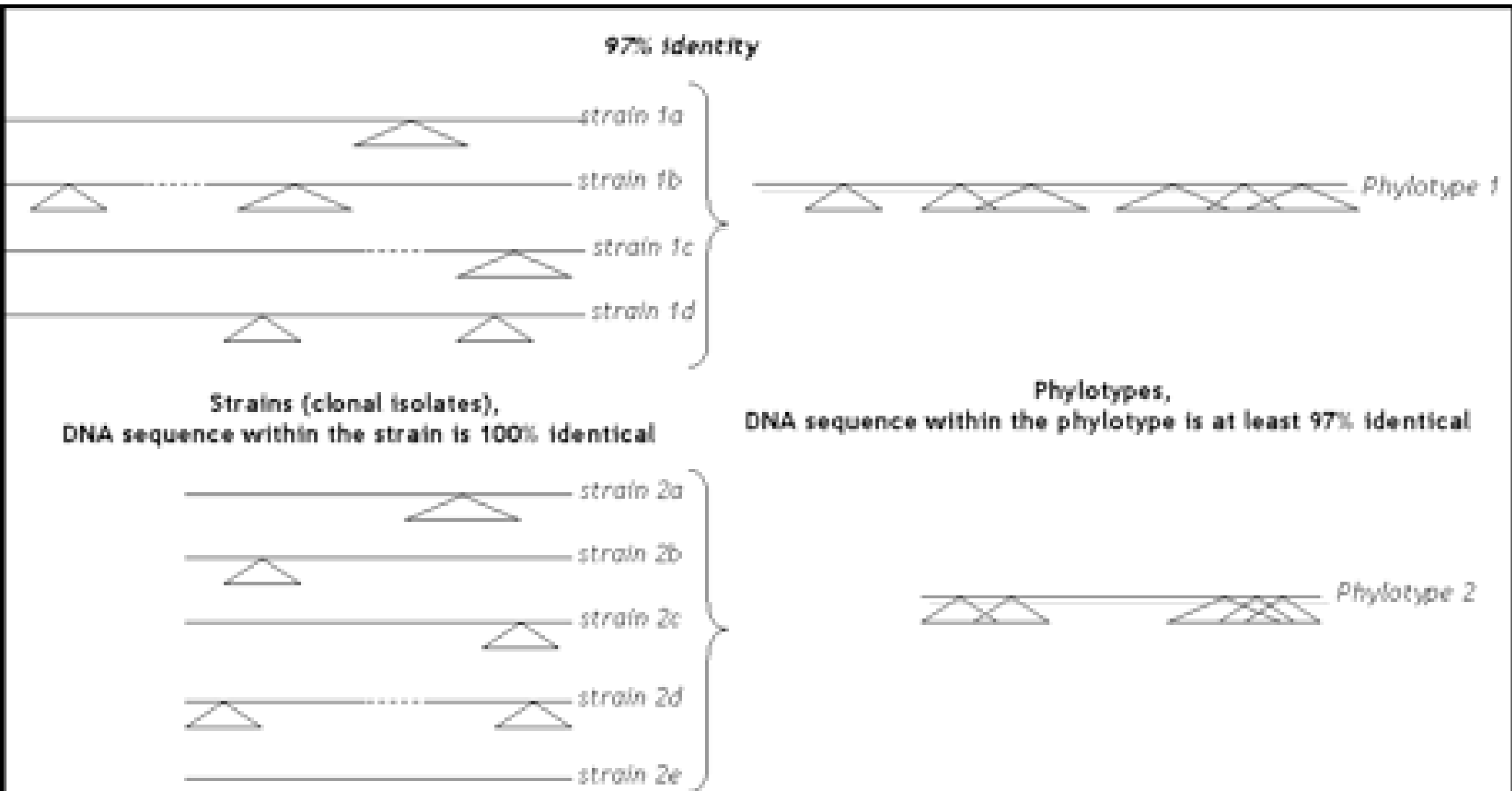
Wu M and Eisen JA. A simple, fast, and accurate method of phylogenomic inference [Genome Biology 2008, 9:R151](https://doi.org/10.1186/gb-2008-9-r151)

Typical workflow for Sanger-based metagenomic projects at the JGI



- Premetagenome Community Composition Profiling
- Sequence Processing
- Data Analysis
 - Postsequencing Community Composition Estimates
 - **Binning**
 - Analyzing Dominant Populations
 - Gene-Centric Analysis

Bins are sets of metagenomic sequence fragments originating from one phylogenetic group, preferably from the same species



binning methods

- Binning methods developed can fall into two categories (biology perspective):
 - similarity-based
 - composition-based
- From machine learning perspective, binning methods can also be divided into
 - supervised learning
 - unsupervised learning

Biology perspective

- similarity-based
 - Assign metagenomic fragments to their closest phylogenetic neighbor based on **coding-sequence identity**.
 - ex. BLAST, dotter
- composition-based
 - Distinguish genomes from one another by **intrinsic features of the sequence**.
 - ex. oligonucleotide frequencies, GC content

machine learning perspective

- Supervised learning
 - Methods that build a classifier using the knowledge of completed genomes
 - (Chen et al., 2005)Current amount of known genomes is insufficient to represent the almost limitless microbial.
- unsupervised learning
 - Do not have dependence on training data
 - Directly clustering metagenomic samples
 - Focuses on the long fosmid-sized fragments

Binning methods matrix

Composition-based

Similarity-based

supervised

PhyloPythia (IBM), 2007

BLAST

MEGAN, 2007

Naïve Basin
Classifier, 2008

S-GSOM, 2008

TETRA, 2005

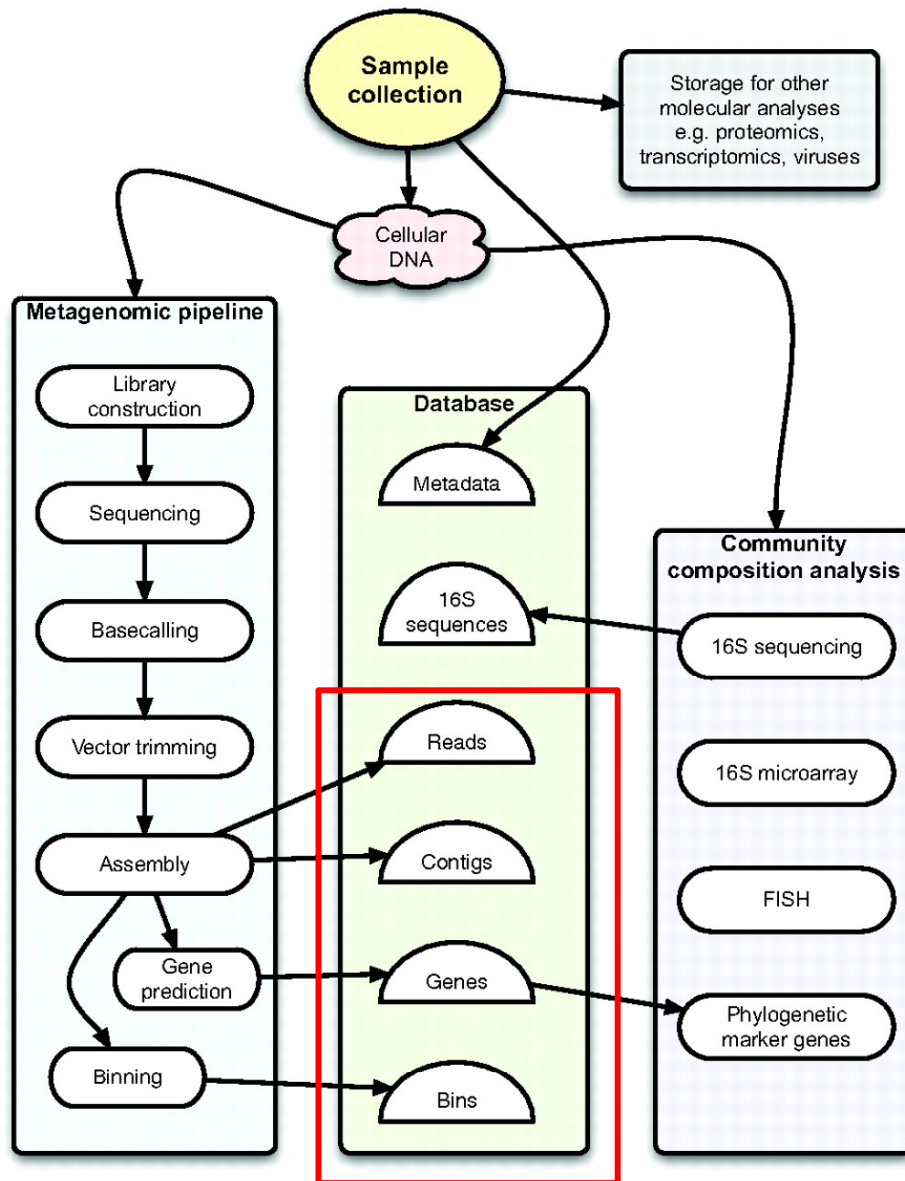
unsupervised

Fuzzy k-means
classifier, 2008

CompostBin, 2008

GSOM, 2008

Typical workflow for Sanger-based metagenomic projects at the JGI



- Premetagenome Community Composition Profiling
- Sequence Processing
- Data Analysis
 - Postsequencing Community Composition Estimates
 - Binning
 - **Analyzing Dominant Populations**
 - Gene-Centric Analysis

Aim 1: Determine the SNP frequency profile of a dominant population

- The ratio of nonsynonymous to synonymous SNPs in protein-coding genes within a population provides an estimate of the fraction of genes under selective pressure.
- The ratio of haplotypes for individual SNPs (site frequency spectra) can be used to estimate important parameters in population genetics such as the scaled mutation rate and scaled exponential growth rate. (Johnson, P. L., and M. Slatkin. 2006. *Genome Res.* 16:1320-1327)
- SNPs highlight junctions of homologous recombination between strains, allowing the degree of sexuality within a population to be estimated. (Whitaker, R. J., and J. F. Banfield. 2006. *Trends Ecol. Evol.* 21:508-516)

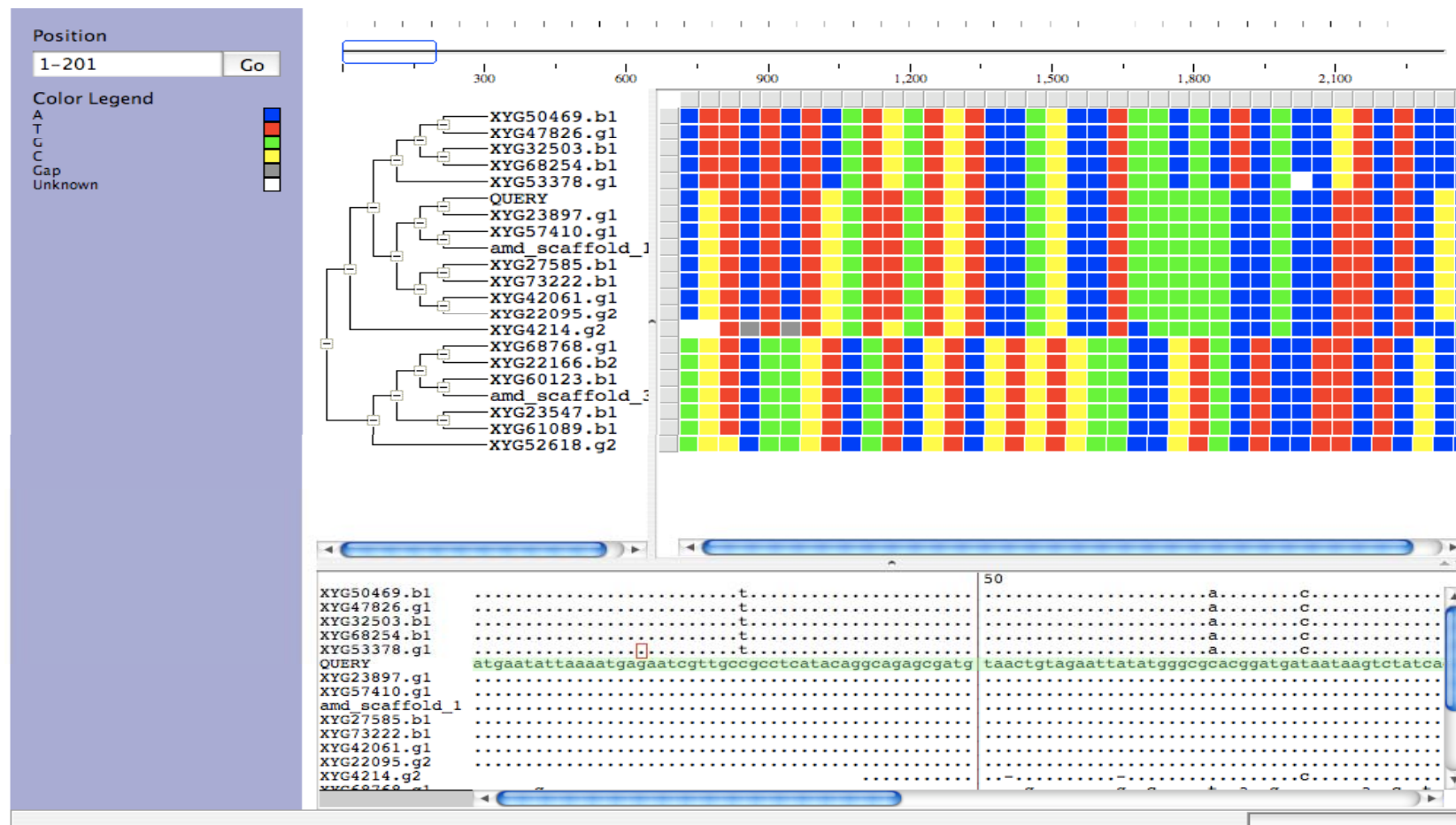
Fine-scale population structure

The visualization and analysis of polymorphisms in composite population assemblies.

- Consed (<http://www.phrap.org/>)
- SNP-VISTA (<http://genome.lbl.gov/vista/snpvista/>)
- Strainer (<http://www.bioinformatics.org/strainer/wiki/>)
- Maq: Mapping and Assembly with Qualities (<http://maq.sourceforge.net/maq-man.shtml>)
- UCSC Genome Browser (<http://genome.ucsc.edu/>)

SNP VISTA

Please be patient for applet to load.



<http://genome.ucsc.edu/>

Use UCSC Genome Browser to host SNP data

Home Genomes Blat Tables DNA PDF/PS Session Help

UCSC Genome Browser on *P. blakesleeanus* Sep. 2006 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search scaffold_1:36,001-70,000 jump clear size 40,000 bp configure

scaffold_1: 40000 45000 50000 55000 60000 65000 70000 75000

SNPs

User Supplied Track

Best Gene Predictions

ewise1Plus.C_10016
gw1.1.194.1
gw1.1.1073.1
ewise1Plus.C_10023
heshPB_pg.C_10017
e_gw1.1.301.1
geneshPB_pg.1__20
geneshPB_pg.1__21
e_gw1.1.307.1
heshPB_pg.C_10023
fgeneshPB_pm.1__2
heshPB_pg.C_10025
Genewise1.C_10038
heshPB_pg.C_10027

move start < 2.0 >

Click on a feature for details. Click on base position to zoom in around cursor. Click gray/blue bars on left for track options and descriptions.

default tracks hide all manage custom tracks configure refresh

Use drop down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

- Custom Tracks

SNPs
dense

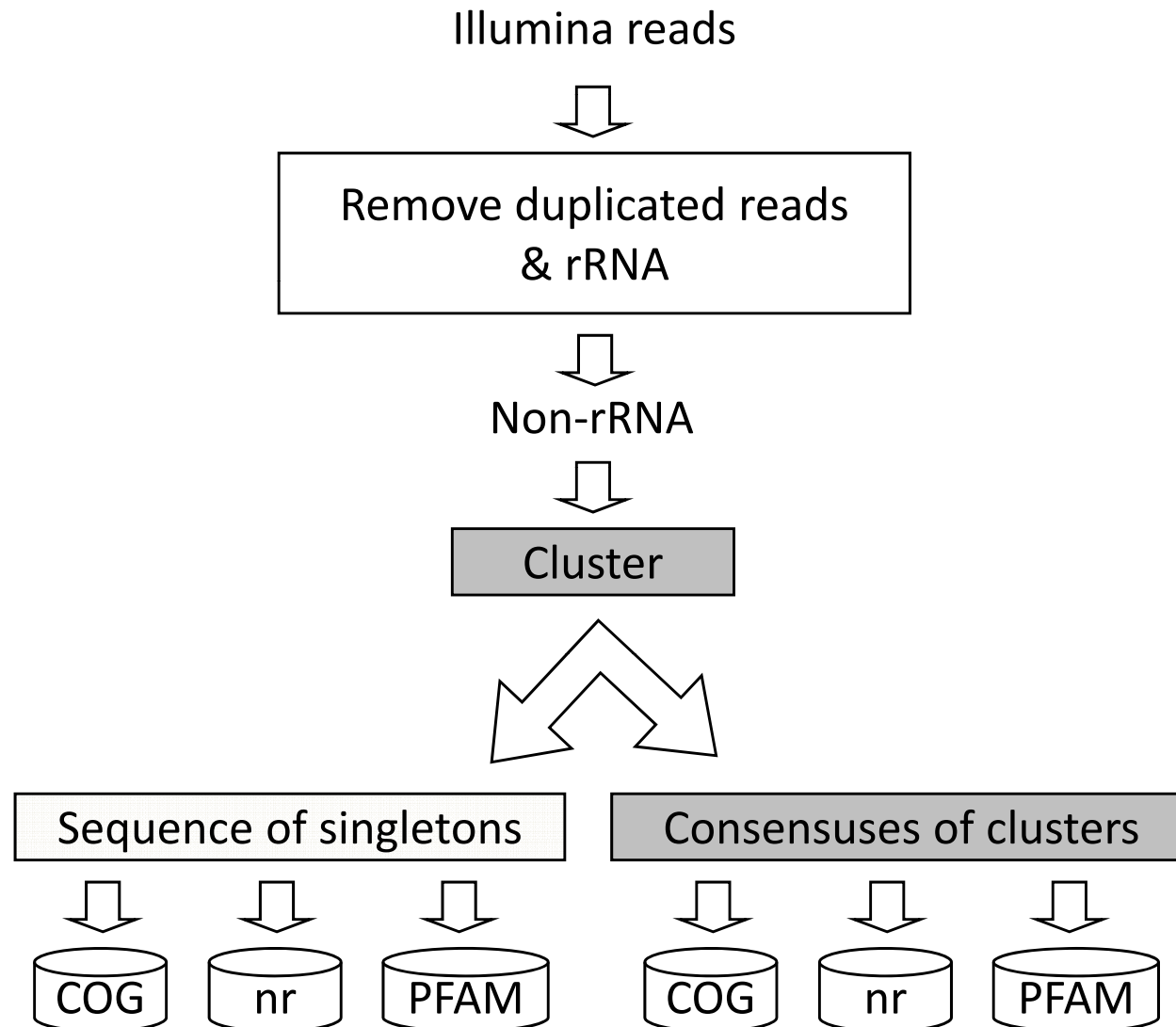
- Mapping and Sequencing Tracks

Base Position Short Match
dense hide

- Genes and Gene Prediction Tracks

Best Genes
full

Metatranscriptome analysis



Short-Read Alignment Tools

Eland (<http://www.fejes.ca/2008/01/aligning-dna-eland.html>)

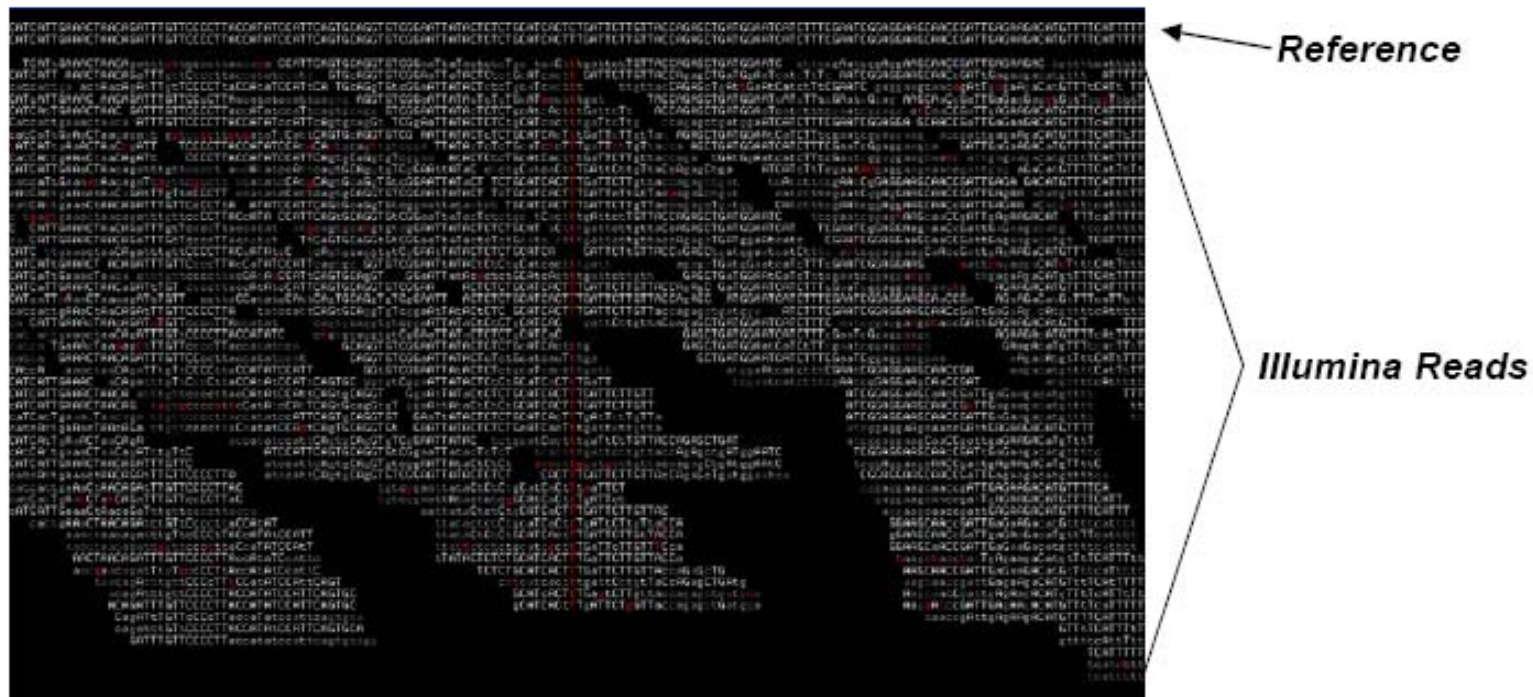
RMAP (<http://rulai.cshl.edu/rmap/>)

SOAP (<http://soap.genomics.org.cn/>)

SHRiMP (<http://compbio.cs.toronto.edu/shrimp/>)

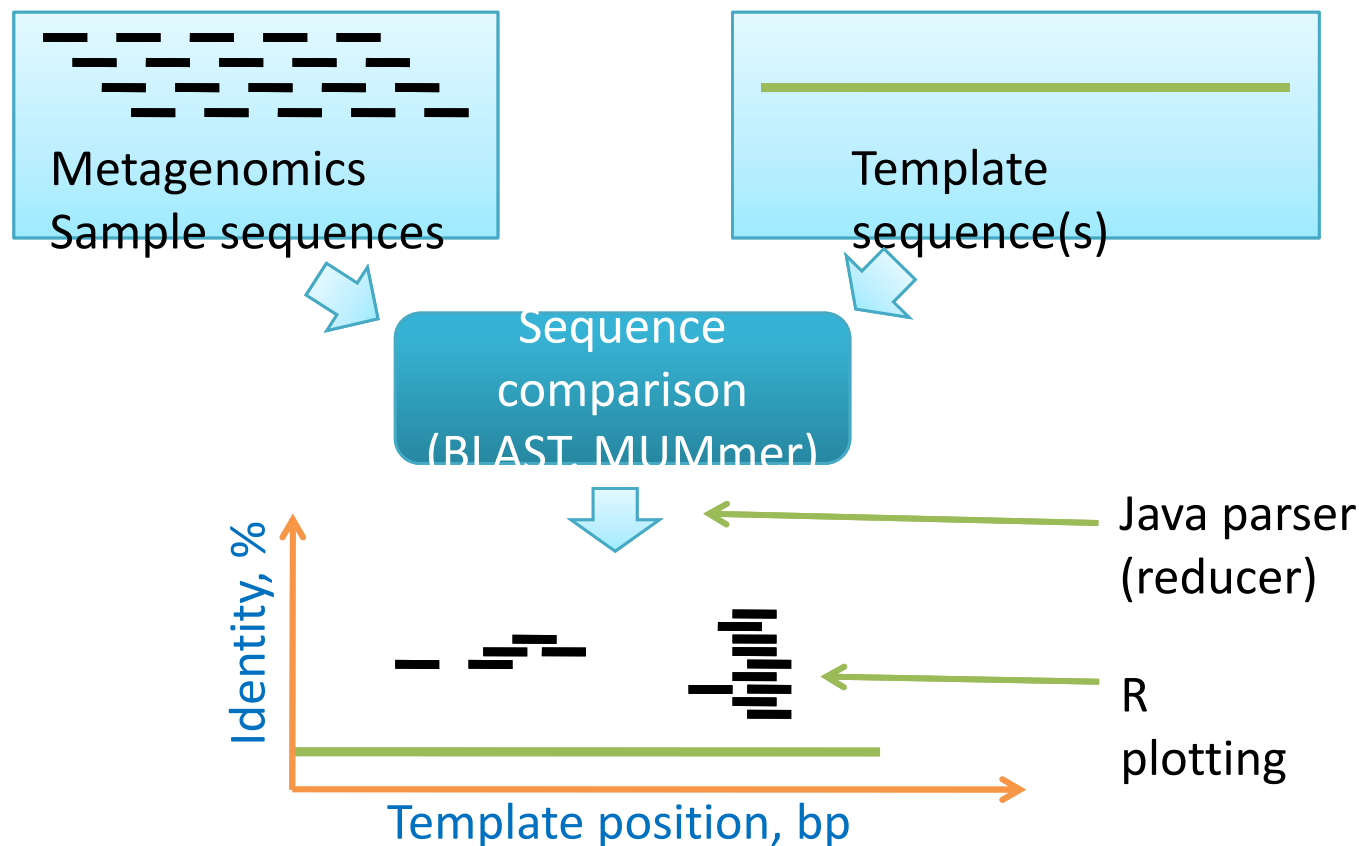
MAQ: Mapping and Assembly with Qualities (<http://maq.sourceforge.net/>)

Align Illumina reads to Reference using MAQ

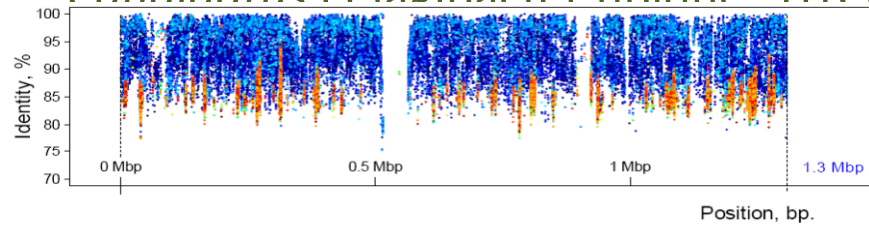


Aim 2: Determine the Gross-scale population structure

Fragment recruitment---align metagenomic reads against reference genomes or genome fragments

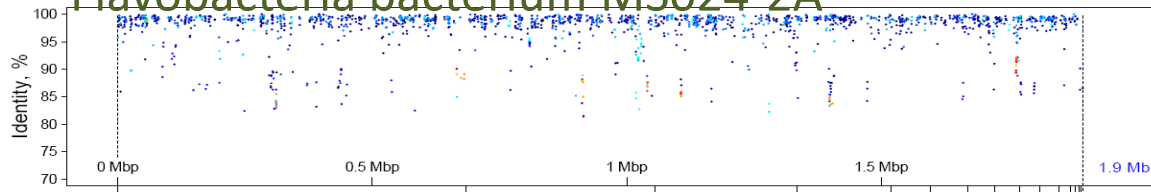


"Candidatus Pelagibacter ubique" HTCC1062



Fragment recruitment application

Flavobacteria bacterium MS024-2A



Flavobacteria bacterium MS024-3C

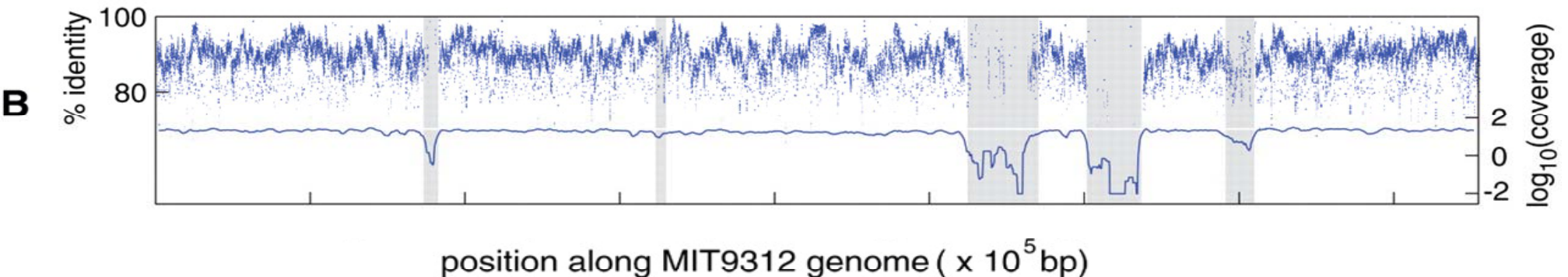
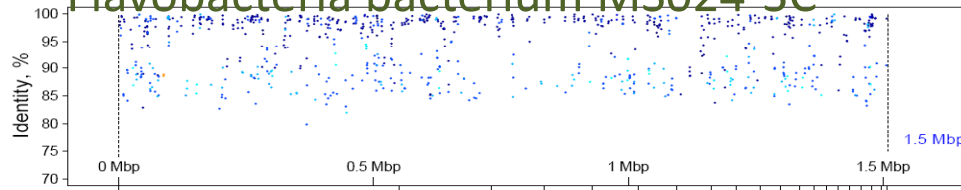
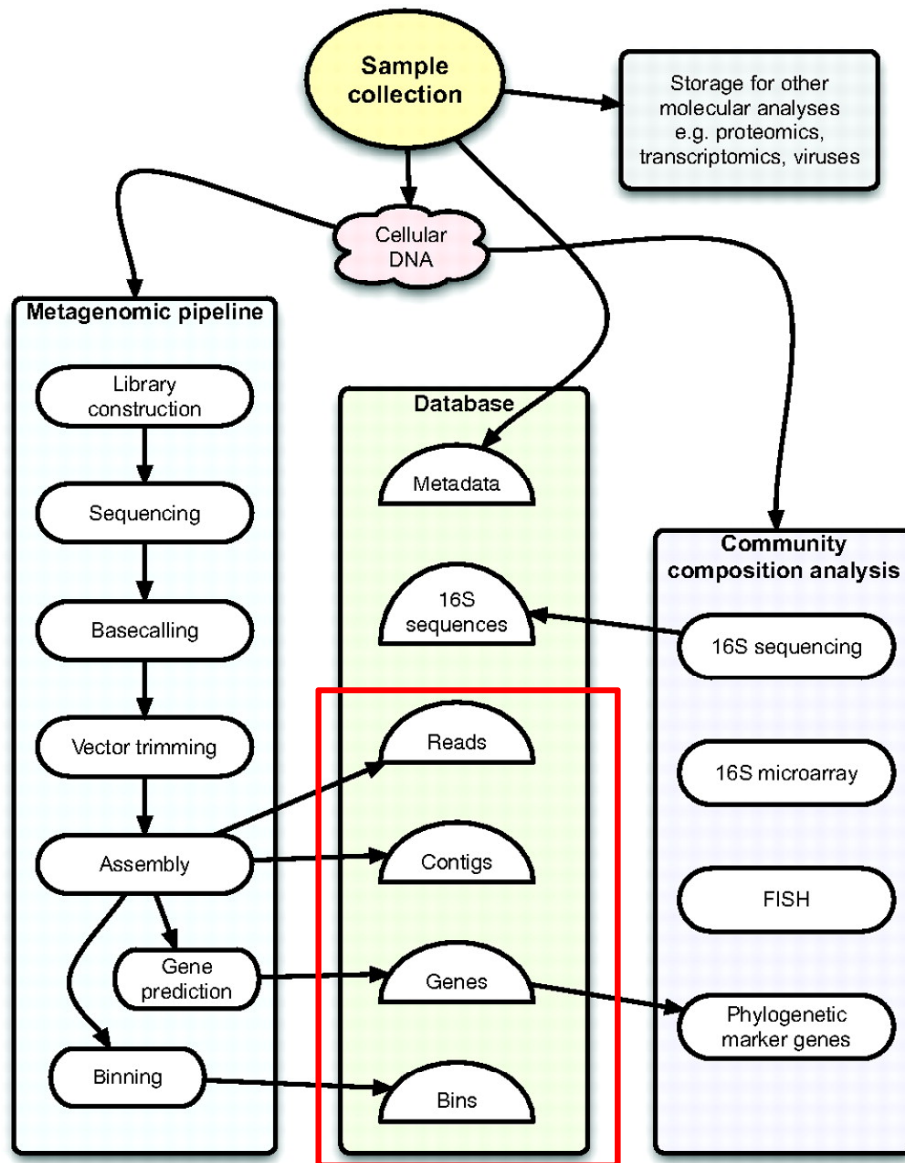


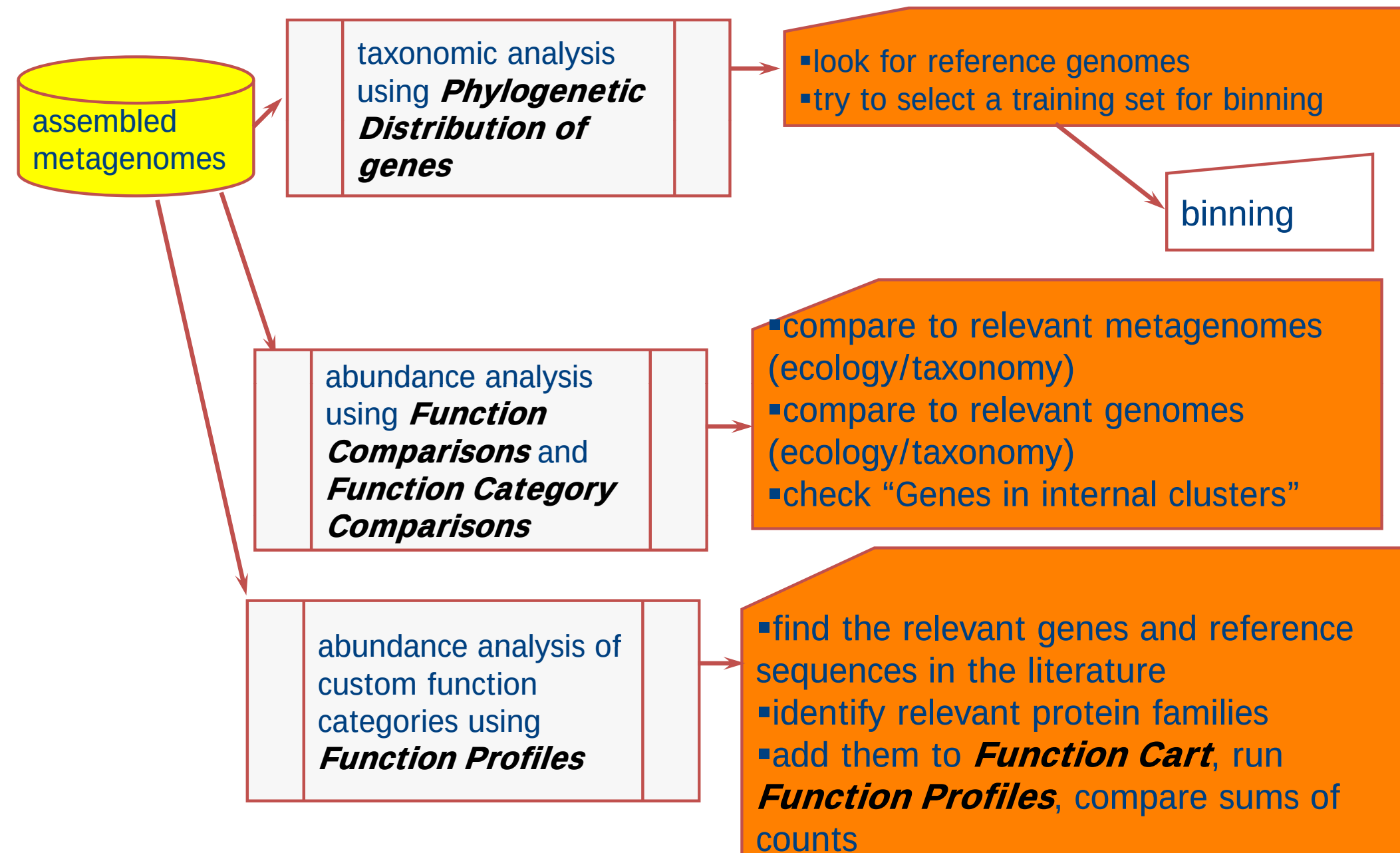
Fig. 2. Features of genomic islands (shaded) in the *Prochlorococcus* strain MIT9312 genome compared with wild sequences from the Atlantic and Pacific Oceans. Coleman et al. , 2006

Typical workflow for Sanger-based metagenomic projects at the JGI

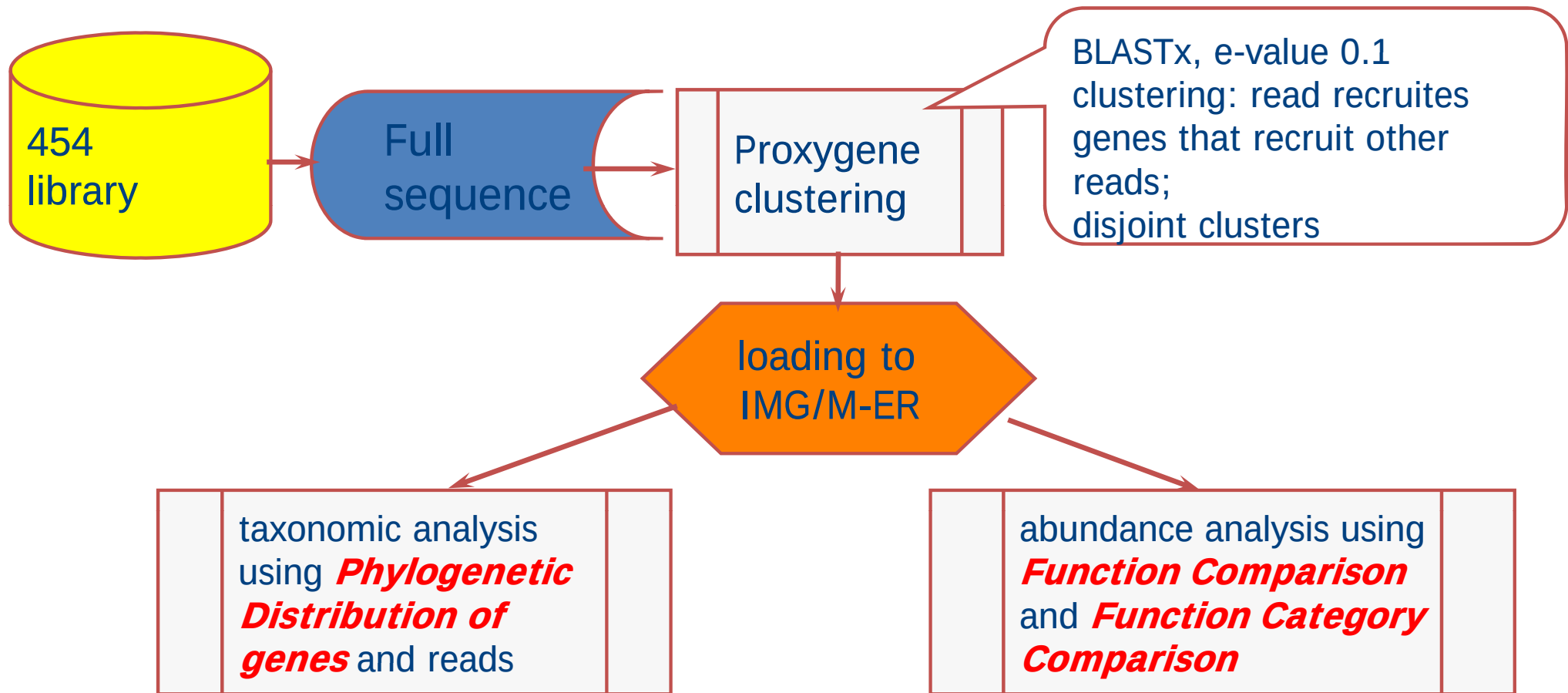


- Premetagenome Community Composition Profiling
- Sequence Processing
- Data Analysis
 - Postsequencing Community Composition Estimates
 - Binning
 - Analyzing Dominant Populations
 - Gene-Centric Analysis

Sanger/Titanium metagenomes: assembled data



454 metagenomes



Analysis of Community Diversity & Abundance

- **IMG: Phylogenetic Distribution of Genes**
- **Purpose.** Assess phylogenetic composition of a metagenome sample based on the distribution of the best BLAST hits of the protein-coding genes found in the dataset.

Navigation: IMG/M Microbiomes or Find Genomes/Genome Browser → Microbiome Details → Phylogenetic Distribution of Genes.

Phylogenetic Distribution of Genes

img/m

INTEGRATED MICROBIAL GENOME with MICROBIOME SAMPLES

IMG/M Home Find Genes Find Genomes Find Functions Compare Genomes MyIMG/M Analysis Carts About IMG/M Using IMG/M

Genome Browser

Save Selections Select All

hint: Selections do not take effect until you click on [Preferences](#) to show or hide genomes. Go to home page statistics under [IMG Genomes](#).

Domains(D): * = Microbiome, B = Bacteria, A = Archaea, E = Eukarya, V = Viruses. Genome Completion(C): F = Finished, D = Draft. Click on column name to sort.

Select ☒ D ☒ C ☐ F ☐ G ☐ H ☐ I ☐ J ☐ K ☐ L ☐ M ☐ N ☐ O ☐ P ☐ Q ☐ R ☐ S ☐ T ☐ U ☐ V ☐ W ☐ X ☐ Y ☐ Z ☐ AA ☐ AB ☐ AC ☐ AD ☐ AE ☐ AF ☐ AG ☐ AH ☐ AI ☐ AJ ☐ AK ☐ AL ☐ AM ☐ AN ☐ AO ☐ AP ☐ AQ ☐ AR ☐ AS ☐ AT ☐ AU ☐ AV ☐ AW ☐ AX ☐ AY ☐ AZ ☐ BA ☐ BB ☐ BC ☐ BD ☐ BE ☐ BF ☐ BG ☐ BH ☐ BI ☐ BJ ☐ BK ☐ BL ☐ BM ☐ BN ☐ BO ☐ BP ☐ BQ ☐ BR ☐ BS ☐ BT ☐ BU ☐ BV ☐ BW ☐ BX ☐ BY ☐ BZ ☐ CA ☐ CB ☐ CC ☐ CD ☐ CE ☐ CF ☐ CG ☐ CH ☐ CI ☐ CJ ☐ CK ☐ CL ☐ CM ☐ CN ☐ CO ☐ CP ☐ CQ ☐ CR ☐ CS ☐ CT ☐ CU ☐ CV ☐ CW ☐ CX ☐ CY ☐ CZ ☐ DA ☐ DB ☐ DC ☐ DD ☐ DE ☐ DF ☐ DG ☐ DH ☐ DI ☐ DJ ☐ DK ☐ DL ☐ DM ☐ DN ☐ DO ☐ DP ☐ DQ ☐ DR ☐ DS ☐ DT ☐ DU ☐ DV ☐ DW ☐ DX ☐ DY ☐ DZ ☐ EA ☐ EB ☐ EC ☐ ED ☐ EE ☐ EF ☐ EG ☐ EH ☐ EI ☐ EJ ☐ EK ☐ EL ☐ EM ☐ EN ☐ EO ☐ EP ☐ EQ ☐ ER ☐ ES ☐ ET ☐ EU ☐ EV ☐ EW ☐ EX ☐ EY ☐ EZ ☐ FA ☐ FB ☐ FC ☐ FD ☐ FE ☐ FF ☐ FG ☐ FH ☐ FI ☐ FJ ☐ FK ☐ FL ☐ FM ☐ FN ☐ FO ☐ FP ☐ FQ ☐ FR ☐ FS ☐ FT ☐ FU ☐ FV ☐ FW ☐ FX ☐ FY ☐ FZ ☐ GA ☐ GB ☐ GC ☐ GD ☐ GE ☐ GF ☐ GG ☐ GH ☐ GI ☐ GJ ☐ GK ☐ GL ☐ GM ☐ GN ☐ GO ☐ GP ☐ GQ ☐ GR ☐ GS ☐ GT ☐ GU ☐ GV ☐ GW ☐ GX ☐ GY ☐ GZ ☐ HA ☐ HB ☐ HC ☐ HD ☐ HE ☐ HF ☐ HG ☐ HH ☐ HI ☐ HJ ☐ HK ☐ HL ☐ HM ☐ HN ☐ HO ☐ HP ☐ HQ ☐ HR ☐ HS ☐ HT ☐ HU ☐ HV ☐ HW ☐ HX ☐ HY ☐ HZ ☐ IA ☐ IB ☐ IC ☐ ID ☐ IE ☐ IF ☐ IG ☐ IH ☐ II ☐ IJ ☐ IK ☐ IL ☐ IM ☐ IN ☐ IO ☐ IP ☐ IQ ☐ IR ☐ IS ☐ IT ☐ IU ☐ IV ☐ IW ☐ IX ☐ IY ☐ IZ ☐ JA ☐ JB ☐ JC ☐ JD ☐ JE ☐ JF ☐ JG ☐ JH ☐ JI ☐ JJ ☐ JK ☐ JL ☐ JM ☐ JN ☐ JO ☐ JP ☐ JQ ☐ JR ☐ JS ☐ JT ☐ JU ☐ JV ☐ JW ☐ JX ☐ JY ☐ JZ ☐ KA ☐ KB ☐ KC ☐ KD ☐ KE ☐ KF ☐ KG ☐ KH ☐ KI ☐ KJ ☐ KK ☐ KL ☐ KM ☐ KN ☐ KO ☐ KP ☐ KQ ☐ KR ☐ KS ☐ KT ☐ KU ☐ KV ☐ KW ☐ KX ☐ KY ☐ KZ ☐ LA ☐ LB ☐ LC ☐ LD ☐ LE ☐ LF ☐ LG ☐ LH ☐ LI ☐ LJ ☐ LK ☐ LL ☐ LM ☐ LN ☐ LO ☐ LP ☐ LQ ☐ LR ☐ LS ☐ LT ☐ LU ☐ LV ☐ LW ☐ LX ☐ LY ☐ LZ ☐ MA ☐ MB ☐ MC ☐ MD ☐ ME ☐ MF ☐ MG ☐ MH ☐ MI ☐ MJ ☐ MK ☐ ML ☐ MN ☐ MO ☐ MP ☐ MQ ☐ MR ☐ MS ☐ MT ☐ MU ☐ MV ☐ MW ☐ MX ☐ MY ☐ MZ ☐ NA ☐ NB ☐ NC ☐ ND ☐ NE ☐ NF ☐ NG ☐ NH ☐ NI ☐ NJ ☐ NK ☐ NL ☐ NM ☐ NO ☐ NP ☐ NQ ☐ NR ☐ NS ☐ NT ☐ NU ☐ NV ☐ NW ☐ NX ☐ NY ☐ NZ ☐ OA ☐ OB ☐ OC ☐ OD ☐ OE ☐ OF ☐ OG ☐ OH ☐ OI ☐ OJ ☐ OK ☐ OL ☐ OM ☐ ON ☐ OO ☐ OP ☐ OQ ☐ OR ☐ OS ☐ OT ☐ OU ☐ OV ☐ OW ☐ OX ☐ OY ☐ OZ ☐ PA ☐ PB ☐ PC ☐ PD ☐ PE ☐ PF ☐ PG ☐ PH ☐ PI ☐ PJ ☐ PK ☐ PL ☐ PM ☐ PN ☐ PO ☐ PP ☐ PQ ☐ PR ☐ PS ☐ PT ☐ PU ☐ PV ☐ PW ☐ PX ☐ PY ☐ PZ ☐ QA ☐ QB ☐ QC ☐ QD ☐ QE ☐ QF ☐ QG ☐ QH ☐ QI ☐ QJ ☐ QK ☐ QL ☐ QM ☐ QN ☐ QO ☐ QP ☐ QQ ☐ QR ☐ QS ☐ QT ☐ QU ☐ QV ☐ QW ☐ QX ☐ QY ☐ QZ ☐ RA ☐ RB ☐ RC ☐ RD ☐ RE ☐ RF ☐ RG ☐ RH ☐ RI ☐ RJ ☐ RK ☐ RL ☐ RM ☐ RN ☐ RO ☐ RP ☐ RQ ☐ RR ☐ RS ☐ RT ☐ RU ☐ RV ☐ RW ☐ RX ☐ RY ☐ RZ ☐ SA ☐ SB ☐ SC ☐ SD ☐ SE ☐ SF ☐ SG ☐ SH ☐ SI ☐ SJ ☐ SK ☐ SL ☐ SM ☐ SN ☐ SO ☐ SP ☐ SQ ☐ SR ☐ SS ☐ ST ☐ SU ☐ SV ☐ SW ☐ SX ☐ SY ☐ SZ ☐ TA ☐ TB ☐ TC ☐ TD ☐ TE ☐ TF ☐ TG ☐ TH ☐ TI ☐ TJ ☐ TK ☐ TL ☐ TM ☐ TN ☐ TO ☐ TP ☐ TQ ☐ TR ☐ TS ☐ TT ☐ TU ☐ TV ☐ TW ☐ TX ☐ TY ☐ TZ ☐ UA ☐ UB ☐ UC ☐ UD ☐ UE ☐ UF ☐ UG ☐ UH ☐ UI ☐ UJ ☐ UK ☐ UL ☐ UM ☐ UN ☐ UO ☐ UP ☐ UQ ☐ UR ☐ US ☐ UT ☐ UU ☐ UV ☐ UW ☐ UX ☐ UY ☐ UZ ☐ VA ☐ VB ☐ VC ☐ VD ☐ VE ☐ VF ☐ VG ☐ VH ☐ VI ☐ VJ ☐ VK ☐ VL ☐ VM ☐ VN ☐ VO ☐ VP ☐ VQ ☐ VR ☐ VS ☐ VT ☐ VU ☐ VV ☐ VW ☐ VX ☐ VY ☐ VZ ☐ WA ☐ WB ☐ WC ☐ WD ☐ WE ☐ WF ☐ WG ☐ WH ☐ WI ☐ WJ ☐ WK ☐ WL ☐ WM ☐ WN ☐ WO ☐ WP ☐ WQ ☐ WR ☐ WS ☐ WT ☐ WU ☐ WV ☐ WW ☐ WX ☐ WY ☐ WZ ☐ XA ☐ XB ☐ XC ☐ XD ☐ XE ☐ XF ☐ XG ☐ XH ☐ XI ☐ XJ ☐ XK ☐ XL ☐ XM ☐ XN ☐ XO ☐ XP ☐ XQ ☐ XR ☐ XS ☐ XT ☐ XU ☐ XV ☐ XW ☐ XX ☐ XY ☐ XZ ☐ YA ☐ YB ☐ YC ☐ YD ☐ YE ☐ YF ☐ YG ☐ YH ☐ YI ☐ YJ ☐ YK ☐ YL ☐ YM ☐ YN ☐ YO ☐ YP ☐ YQ ☐ YR ☐ YS ☐ YT ☐ YU ☐ YV ☐ YW ☐ YX ☐ YY ☐ YZ ☐ ZA ☐ ZB ☐ ZC ☐ ZD ☐ ZE ☐ ZF ☐ ZG ☐ ZH ☐ ZI ☐ ZJ ☐ ZK ☐ ZL ☐ ZM ☐ ZN ☐ ZO ☐ ZP ☐ ZQ ☐ ZR ☐ ZS ☐ ZT ☐ ZU ☐ ZV ☐ ZW ☐ ZX ☐ ZY ☐ ZZ

Microbiome Details

Microbiome Information

Microbiome Name	Acid Mine Drainage
Taxon Object ID	2001200000
NCBI Taxon ID	256318
NCBI Project ID	13696
GOLD ID	Gm00001
External Links	NCBI/GenBank: AADL00000000; PUM
Sequencing Status	Draft
Sequencing Center	JGI
Funding Agency	DOE
Phenotype	

Phylogenetic Distribution of Genes

Distribution of genes binned by BLAST percent identities

Acid Mine Drainage Phylogenetic Distribution of Genes

Distribution of Best Blast Hits
View COG Functional Category Statistics
View COG Pathway Statistics

The Phylogenetic Distribution of Genes allows to assess the phylogenetic composition of a metagenome sample based on the distribution of best BLAST hits of protein-coding genes in the dataset. The phylogenetic distribution can be projected onto the families in a phylum/class (click on phylum/class name), and further onto species in a family. For a reference genome within a species, the metagenome genes are viewed using the Protein Recruitment Plot or the Reference Genome Context Viewer.

Distribution of Best Blast Hits

Domains(D): * = Microbiome, B = Bacteria, A = Archaea, E = Eukarya, P = Plasmids, V = Viruses.

Compare	D	Phylum/Class	No. Of Genomes	No. Of Hits	Histogram 30%	No. Of Hits	Histogram 60%
☐	A	Crenarchaeota	15	181		6	
☐	A	Euryarchaeota	34	1319		1946	
☐	B	Acidobacteria	2	78		12	
☐	D	Actinobacteria					
☐	R	Aquificae					
☐	B	Bacteroidetes					
☐	B	Chlamydiae					
☐	B	Chlorobi					
☐	B	Chloroflexi					
☐	B	Cyanobacteria					
☐	B	Deinococcus-Thermobacteria					
☐	B	Bacilli					
☐	B	Clostridia					
☐	B	Mollicutes					
☐	B	Fusobacteria					
☐	B	Lentisphaerae					
☐	B	Planctomycetes					
☐	B	Alphaproteobacteria					
☐	B	Betaproteobacteria					
☐	B	Deltaproteobacteria					
☐	B	Epsilonproteobacteria					

Best Hits at 90% Identity

Euryarchaeota

Table View COG Functional Cat COG Pathways Hide genes unassociated to cog

Cog Function View

Amino acid transport and metabolism (110) Select Clear

- ☐ 2001200362 100% 1096 3-dehydroquinate synthetase
- ☐ 2001200501 100% 883 Glycine cleavage system protein P(pyridoxal-binding), C-terminal domain
- ☐ 2001200479 100% 1494 Shikimate 5-dehydrogenase
- ☐ 2001200187 100% 4137 Amino acid transporters
- ☐ 2001200584 100% 4137 Amino acid transporters
- ☐ 2001200522 100% 2614 Xaa-Pro aminopeptidase
- ☐ 2001200417 100% 1499 Amino acid transporters
- ☐ 2001200334 100% 1312 Ornithine carbamoyltransferase
- ☐ 2001200359 100% 1494 Shikimate 5-dehydrogenase

Best Hits at 90% Identity

Euryarchaeota

Table View COG Functional Cat COG Pathways Hide genes unassociated to cog

Cog Pathway View

Aminoacyl-tRNA synthetases and alternate systems for amino acid activation (21) Select Clear

- ☐ 2001200813 100% 1406 Threonyl-tRNA synthetase
- ☐ 2001200311 100% 1403 Valyl-tRNA synthetase
- ☐ 2001200283 100% 897 Aspartyl-tRNA synthetase
- ☐ 2001200312 100% 1403 Valyl-tRNA synthetase
- ☐ 2001200364 100% 1358 Arginyl-tRNA synthetase
- ☐ 2001200386 99% 1461 Alaninyl-tRNA synthetase
- ☐ 2001200324 99% 1553 Isoleucyl-tRNA synthetase
- ☐ 2001200314 99% 1199 Phenylalanyl-tRNA synthetase alpha subunit

Best Hits at 90% Identity

Euryarchaeota

Table View COG Functional Cat COG Pathways Hide genes unassociated to cog

Select	Gene ID	Percent	Name	Cog ID	Cog Name	Cog Function	Cog Gene Count
☐	2001200003	100	Prothymine ferro-lyase (ferrochelatase) [Acid Mine Drainage]	COG0276	Coenzyme transport and metabolism	Coenzyme transport and metabolism	937
☐	2001200004	98	Selenocysteine-specific translation elongation factor [Acid Mine Drainage]	COG3276	Translation, ribosomal structure and biogenesis	Translation, ribosomal structure and biogenesis	355
☐	2001200006	100	Phosphoribosylamine-glycine ligase [Acid Mine Drainage] EC 6.3.4.13	COG0151	Nucleotide transport and metabolism	Nucleotide transport and metabolism	1221
☐	2001200007	100	Predicted transcriptional regulator containing the HTH domain [Acid Mine Drainage]	COG1386	Transcription	Transcription	753
☐	2001200008	100	Purine-cytosine permease and related proteins [Acid Mine Drainage]	COG1457	Nucleotide transport and metabolism	Nucleotide transport and metabolism	576
☐	2001200009	100	Deoxycytidine deaminase [Acid Mine Drainage]	COG0717	Nucleotide transport and metabolism	Nucleotide transport and metabolism	800

Navigation (cont.): Phylogenetic Distribution of Genes → Compare COG Functions; View all COG Functional Category/Pathways

Phylogenetic Distribution of Genes

Acid Mine Drainage Phylogenetic Mapping Of Genomic Fragments

Domains(D): *=Microbiome,
B=Bacteria, A=Archaea, E=Eukarya, V=Viruses.

Compare max. 2	D	Phylum/Class	No. Of Genomes	No. Of HITS 30%	Histogram 30%
☐	A	Crenarchaeota	6	155	
☐	A	Euryarchaeota	25	1352	
☐	B	Acidobacteria	2	99	
☐	D	Actinobacteria	42	194	
☐	B	Aquificae	1	42	
☐	B	Bacteroidetes	17	33	
☐	B	Chlamydiae	11	6	
☐	B	Chlorobi	10	77	
☐	B	Chloroflexi	6	79	
☐	B	Cyanobacteria	25	139	
☐	B	Deinococcus-Thermus	4	33	
☐	B	Bacilli	38	116	
☐	B	Clostridia	21	235	
☐	B	Mollicutes	17	2	
☐	V	unclassified	27	1	
-	-	Unassigned	-	3678	

Histogram is a count of best hits within the phylum/class at
Unassigned are the remainder of genes less than the percent

Percent Difference Factor

Compare COG Functions

[View all COG Functional Category 30% Statistics](#)
[View all COG Functional Category 60% Statistics](#)
[View all COG Functional Category 90% Statistics](#)

[View all COG Pathways 30% Statistics](#)
[View all COG Pathways 60% Statistics](#)
[View all COG Pathways 90% Statistics](#)

Comparison Summary Statistics for Crenarchaeota and Euryarchaeota COG Functional Categories 30%

COG Functional Category	Crenarchaeota	Euryarchaeota
Amino acid transport and metabolism	19 (16.8)	107 (11.1)
Carbohydrate transport and metabolism	9 (8.0)	71 (7.4)
Cell cycle control, cell division, chromosome partitioning	0	10 (1.0)
Cell wall/membrane/envelope biogenesis	3 (2.7)	22 (2.3)
Coenzyme transport and metabolism	3 (2.7)	53 (5.5)
Defense mechanisms	3 (2.7)	13 (1.3)
Energy production and conversion	10 (8.8)	71 (7.4)
Function unknown	14 (12.4)	92 (9.5)
General function prediction only	22 (19.5)	125 (13.0)
Inorganic ion transport and metabolism	4 (3.5)	37 (3.8)

Summary Statistics of COG Functional Categories 30%

COG Functional Category	Crenarchaeota	Euryarchaeota	Acidobacteria	Actinobacteria	Aquificae	Bacteroidetes	Chlamydiae
Amino acid transport and metabolism	19 (16.8)	107 (11.1)	3 (3.7)	9 (6.8)	8 (19.5)	3 (11.1)	0
Carbohydrate transport and metabolism	9 (8.0)	71 (7.4)	5 (6.2)	5 (3.8)	3 (7.3)	7 (25.9)	1 (20.0)
Cell cycle control, cell division, chromosome partitioning	0	10 (1.0)	0	0	0	0	0
Cell motility	0	0	1 (1.2)	1 (0.8)	0	0	0
Cell wall/membrane/envelope biogenesis	3 (2.7)	22 (2.3)	8 (9.9)	8 (6.1)	1 (2.4)	0	1 (20.0)
Chromatin structure and dynamics	0	0	0	0	0	0	0
Coenzyme transport and metabolism	3 (2.7)	53 (5.5)	4 (4.9)	6 (4.5)	5 (12.2)	0	0
Defense mechanisms	3 (2.7)	13 (1.3)	7 (8.6)	2 (1.5)	0	1 (3.7)	0

Navigation (cont.): Phylogenetic Distribution of Genes → (Project on Family) →
 (Project on Species) → Protein Recruitment Plot / Reference Genome Context Viewer

Phylogenetic Distribution of Genes

Acid Mine Drainage Family Statistics

Archaea Euryarchaeota

Loaded.

Family	No. Of Hits 30%	Histogram 30%	No. Of Hits 60%	Histogram 60%	No. Of Hits 90%	Histogram 90%
Archaeoglobaceae	26		-		-	
Ferropasmaceae	362		1507		1920	
Halobacteriaceae	12		-		-	
Methanobacteriaceae	8		1		-	
Methanocaldococcaceae	7		-		-	
Methanococcaceae	4		-		-	
Methanopyraceae	2		-		-	
Methanosactaceae	9		1		-	
Methanosarcinaceae	92		2		-	
Methanospirillaceae	17		3		-	
Picrophilaceae	273		131		1	
Thermococcaceae	36		7		-	
Thermoplasmataceae	490		287		17	

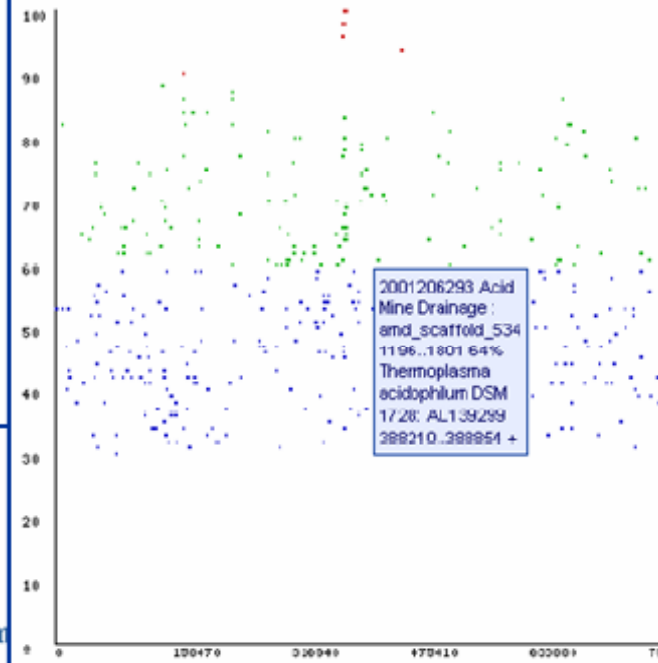
Acid Mine Drainage Protein Recruitment Plot

Thermoplasmataceae

Red 90%
 Green 60%
 Blue 30%

All Strands Plot

0 .. 1584691



Acid Mine Drainage Species Statistics

Archaea Euryarchaeota

Family Thermoplasmataceae

Protein Recruitment Plot

- Percent identity plot for all scaffolds
- Percent identity plot for all positive strand scaffolds
- Percent identity plot for all negative strand scaffolds

Genus Species	No. Of Hits 30%	Histogram 30%	No. Of Hits 60%
Thermoplasma acidophilum	112		112
Thermoplasma volcanium	273		172

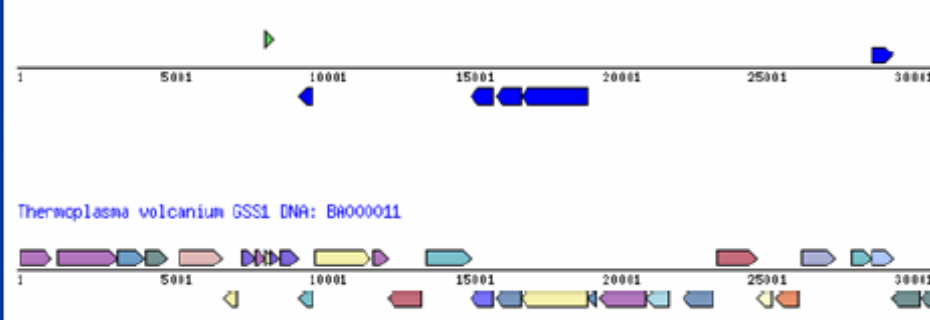
Reference Genome Context Viewer for Fragments

Archaea Euryarchaeota

Thermoplasmataceae Thermoplasma volcanium

Blue - 30%
 Green - 60%
 Red - 90%

Acid Mine Drainage 1 .. 40001



Analysis Protein Family Relative Abundance

- **IMG: Abundance Profile Overview**
- **Purpose.** Examine relative abundance of all protein families (COGs and Pfams) and functional families (Enzymes) in metagenomes and isolate genomes.

Navigation: Compare Genomes → Abundance Profiles → Abundance Profile Overview

Abundance Profile Overview



IMG/M Home Find Genomes Find Genes Find Functions **Compare Genomes** Analysis Carts

Genome Statistics | VISTA **Abundance Profiles** Genome Clustering

Abundance Profiles Tools

The following tools operate on functional profiles of multiple genomes.

Tool

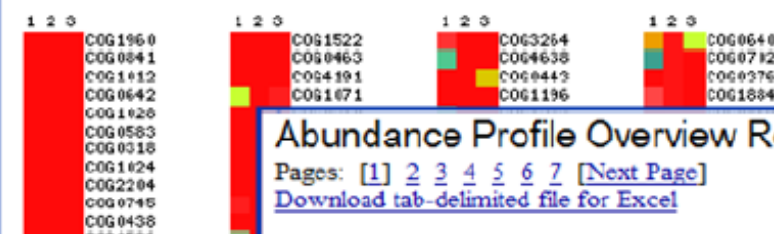
[Abundance Profile Overview\(All Functions\)](#)

[Abundance Profile Search](#)

[Function Comparisons](#)

Abundance Profile Overview Results

- 1 - [Whalefall Sample #1](#)
- 2 - [Whalefall Sample #2](#)
- 3 - [Whalefall Sample #3](#)



Abundance Profile Overview Results

Pages: [\[1\]](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [\[Next Page\]](#)

[Download tab-delimited file for Excel](#)

Add Selected to Function Cart

Select All

Clear All

Select	Row No.	ID	Name	Wha Sam #11	Wha Sam #22	Wha Sam #33
☐	1	COG0001	Glutamate-1-semialdehyde aminotransferase	19	5	11
☐	2	COG0002	Acetylglutamate semialdehyde dehydrogenase	9	10	10
☐	3	COG0003	Oxvanion-translocating ATPase	1	1	4

Abundance Profile Overview

Display Options:

Output Type

☒ Heap Map

or

☐ Matrix

Normalization Method

☒ None

☐ Scale for genome size

☒ Gene count

☐ Estimated gene copies² ☐ Include

Enter matching text for highlighting clusters/rows. (E.g.,

Function:

☒ COG

☐ Pfam

☐ Enzyme

☐ TIGRfam

Genomes¹:

Please select 1 to 100 genomes.

[Whalefall Sample #1 \(*\)](#)[D]
[Whalefall Sample #2 \(*\)](#)[D]
[Whalefall Sample #3 \(*\)](#)[D]

Analysis Protein Family Relative Abundance

- **Abundance Profile Search**
- **Purpose.** Selection of protein families (COGs and Pfams) in metagenomes and isolate genomes based on their relative abundance; similar to Phylogenetic Profiler for gene selection, but operates on protein families rather than individual genes.

Navigation. Compare Genomes → Abundance Profiles → Abundance Profile Search

Abundance Profile Search



IMG/M Home | Find Genomes | Find Genes | Find Functions | **Compare Genomes**

Genome Statistics | VISTA | **Abundance Profiles** | Genome Clustering

Abundance Profiles Tools

The following tools operate on functional profiles of multiple genomes.

Tool
Abundance Profile Overview(All Functions)
Abundance Profile Search
Function Comparisons
Function Category Comparisons

Abundance Profile Search

Find genes in genome (bin) of interest qualified by similarity to sequences in other genomes (based on BLASTP alignments). Only user-selected genomes appear in the profiler.

Genome Completion: [F]inished, [D]raft.

Select functional classification

☐ COG
☒ Pfam

Normalization Method

☐ None
☒ Frequency
☐ z-score

Show Results As:

☐ Gene Counts
☐ Normalized Values
☒ Both

More Abundant Cut-Off Less Abundant Cut-Off

Enter matching text for highlighting clusters/rows (E.g., 'kinase')

Find functions in	More Abundant Than In	Less Abundant Than In	Ignoring	TaxonName
☐	☐	☐	☐	*Microbiome
☐	☐	☐	☐	Sample
☐	☐	☐	☐	AMD
☐	☐	☐	☒	[Acid Mine Drainage]
☒	☐	☐	☐	Ferroplasma acidarmanus Type I
☐	☒	☐	☐	Ferroplasma acidarmanus Type II
☐	☐	☐	☒	Leptospirillum sp. Group II
☐	☐	☐	☒	Leptospirillum sp. Group III
☐	☐	☒	☐	Thermoplasmatales archaeon Gpl

Abundance Profile Search Results

Add Selected to Function Cart Select All Clear All

Red - Less abundant.
Green - More abundant.

Selection	Pfam Id	Pfam Name	Ferroplasma acidarmanus Type I	Ferroplasma acidarmanus Type II	Thermoplasmatales archaeon Gpl
☐	pfam00488	MutS_V	1 (369.959)	0 (0.000)	5 (1230.921)
☐	pfam00654	Voltage_CLC	1 (369.959)	0 (0.000)	3 (738.552)
☐	pfam00903	Glyoxalase	1 (369.959)	0 (0.000)	2 (492.368)
☐	pfam00925	GTP_cyclohydro2	1 (369.959)	0 (0.000)	3 (738.552)
☐	pfam01112	Asparaginase_2	1 (369.959)	0 (0.000)	4 (984.737)

Analysis Protein Family Relative Abundance

- **IMG: Function Comparison**
- **Purpose.** Comparison of a query metagenome/ isolate genome with one or several reference metagenomes/ isolate genomes, in terms of their relative abundance of protein families (COGs, Pfams, TIGRfams) and functional families (Enzymes), with estimates of the *statistical significance* of the observed differences

Navigation. Compare Genomes → Abundance Profiles → Function Comparisons

Function Comparisons



IMG/M Home Find Genomes Find Genes Find Functions **Compare Genomes** Analysis Carts MyIMG/M

Genome Statistics VISTA **Abundance Profiles** Genome Clustering

Abundance Profiles Tools

The following tools operate on functional profiles of multiple genomes.

Tool
Abundance Profile Overview(All Functions)
Abundance Profile Search
Function Comparisons
Function Category Comparisons

Function Comparisons

Page Options:

500 functions per page.

Select One Query Genome¹:

Whalefall Sample #1 (*)[D]
Whalefall Sample #2 (*)[D]
Whalefall Sample #3 (*)[D]

Select Multiple Reference Genomes¹:

Whalefall Sample #1 (*)[D]
Whalefall Sample #2 (*)[D]
Whalefall Sample #3 (*)[D]

Function:

- ☒ COG
- ☐ Pfam
- ☐ Enzyme
- ☐ TIGRfam

Measurement:

- ☒ Gene count
- ☐ Estimated gene copies²

5 Minimum function gene count to c

Output:

- ☐ Show all rows.
- ☒ Show only rows with at least one non-
- ☐ Show only rows with significant hits.

Function Comparisons

Loaded.

Add Selected to Function Cart

Select All

Clear All

Select	Row No.	ID	Name	Whalefa Sample #1 Gene Count (Q)	Whalefa Sample #2 (R)	Whalefa Sample #2 p-value (R)	Whalefa Sample #2 Gene Count (R)			
☐	1	COG0001	Glutamate-1-semialdehyde aminotransferase	19	2.69	3.54e-03	5			
☐	2	COG0002	Acetylglutamate semialdehyde dehydrogenase	0	-0.38	3.53e-01	10			
☐	21	COG0022	Pyruvate/2-oxoglutarate dehydrogenase complex, dehydrogenase (E1) component, eukaryotic type, beta subunit	9	-4.47	3.93e-06	38	-1.56	5.95e-02	15

Analysis Protein Family Relative Abundance

- **IMG: Functional Category Comparison**
- **Purpose.** Comparison of a query metagenome/ isolate genome with one or several reference metagenomes/ isolate genomes, in terms of the relative abundance of the genes assigned to different functional categories (COG Pathway, KEGG Pathway, KEGG Pathway Category, Pfam Category, TIGRfam sub-roles), with estimates of the *statistical significance* of the observed differences

Function Category Comparisons

 **INTEGRATED MICROBIAL GENOMICS**
with MICROBIOLOGY

[IMG/M Home](#) | [Find Genomes](#) | [Find Genes](#) | [Find Functions](#) | **[Compare Genomes](#)** | [Analysis Carts](#) | [My Lists](#)

[Genome Statistics](#) | [VISTA](#) | **[Abundance Profiles](#)** | [Genome Clustering](#)

Abundance Profiles Tools

The following tools operate on functional profiles of multiple genomes.

Tool
Abundance Profile Overview(All Functions)
Abundance Profile Search
Function Comparisons
Function Category Comparisons

Function Category Comparisons

Page Options:

500 functions per page.

Select One Query Genome¹:

Whalefall Sample #1 (*)
Whalefall Sample #2 (*)
Whalefall Sample #3 (*)

Select Multiple Reference Genomes¹:

Whalefall Sample #1 (*)
Whalefall Sample #2 (*)
Whalefall Sample #3 (*)

Function Category:

☒ COG Pathways
☐ Kegg Pathways
☐ Kegg Pathways Category
☐ Pfam Categories (formed from COG categories / pathways)
☐ TIGRfam Category Roles

Measurement:

☒ Gene count
☐ Estimated gene copies²

5 Minimum function gene count to calculate d-score³

Output:

☒ D-rank⁴
☐ D-rank (unsigned)⁵

☐ Include all rows, including those without hits.

Function Category Comparisons

Pages: [1]
[Download tab-delimited file for Excel](#)

Row No.	ID	Name	Whalefa Sample #1 Gene Count (Q)	Whalefa Sample #2 (R)	Whalefa Sample #2 p-value (R)	Whalefa Sample #2 Gene Count (R)	Whalefa Sample #3 (R)		
1	1	Archael/Vacuolar-type H+ ATPase subunits	13	0.00	5.00e-01	1	0.00		
2	10	Isoleucine biosynthesis	123	-0.41	3.40e-01	115	-1.77	3.86e-02	136
3	11	Leucine biosyntehsis	183	0.26	3.99e-01	155	-1.07	1.42e-01	185
4	12	Methionine biosynthesis	108	-2.65	4.02e-03	137	-2.11	1.75e-02	129

Search Terms and Pathways

Find functions in selected genomes by keyword.

Keyword: lipid

Filters: COG

Select from gene product names, EC numbers, EC numbers or enzyme names, Pfam domains, TIGRfam domains, InterPro domains, KEGG Pathway Enzymes, IMG Term and Synonyms, IMG Pathways, IMG Parts List

Select from genome names, assume current version, Domains: (B)acteria, (V)iruses, Genome Compaction

Aeromonas sp. ATCC 35061 (B)[D]
Aeromonas sp. ATCC 35061 (B)[D]
Agrobacterium sp. ATCC 29420 (B)[D]
Agrobacterium sp. ATCC 29420 (B)[D]
Alcanivorax borkumensis ORZ2 (B)[D]
Algoriphagus sp. PR1 (B)[D]
Alkalilimnicola ehrlichi MLHE-1 (B)[F]
Alkaliphilus metalliredigens QYMF (B)[D]
Alteromonadales bacterium TV-7 (B)[D]
Alteromonas macleodii Deep ecotype (B)[D]
Anaerobaculum sp. ATCC 29420 (B)[D]

Go Reset

Examples

- "kinase" as a Function Name may get "Shikimate kinase", etc.
- "EC:1.2.3.3" or "1.2.3.3" as an EC Number (exact) gets only genes associated with EC 1.2.3.3.
- "1.2.3" as an EC Number (inexact) returns genes associated with EC 1.2.3.3 and EC 1.1.2.3.
- "IPR004462" (exact) or "kinase" or "Acetylglutamate" (inexact) gets "IPR004462 Acetylglutamate kinase".
- "pfam00698" will retrieve that Pfam for Acl trans 1.

Function Search Results

(ii)

The number of genes is shown in parentheses.

☐ COG0416 - Fatty acid/phospholipid biosynthesis enzyme (1)
☐ COG3307 - Lipid A core - O-antigen ligase and related enzymes (1)
☐ COG0671 - Membrane-associated phospholipid phosphatase (1)
☐ COG3963 - Phospholipid N-methyltransferase (1)

IMG Home Find Genomes Find Genes

Search **COG** Pfam KEGG Enzyme

Add Selected to Function Cart

Select All

Clear All

COG Browser

[COG list](#)

[Amino acid transport and metabolism](#) [E]

[Arginine biosynthesis](#)
[Histidine biosynthesis](#)
[Isoleucine biosynthesis](#)
[Leucine biosynthesis](#)
[Methionine biosynthesis](#)
[Phenylalanine/tyrosine biosynthesis](#)
[Proline biosynthesis](#)
[Threonine biosynthesis](#)
[Tryptophan biosynthesis](#)
[Valine biosynthesis](#)

Details for COG Category *Amino acid transport and metabolism*.

Click on column name to sort.

Select	COG ID	COG Name	Genome Count
☐	COG0002	Acetylglutamate semialdehyde dehydrogenase	624
			802

Genomes with COG0002 (iv)

Genomes with *Acetylglutamate semialdehyde dehydrogenase*.

Phylogenetic Distribution

D	C	Genome	Gene Count
A	D	Caldivirga maculigenensis IC-167	1
A	D	Ferroplasma acidarmanus Fer1	

COG Pathway Details (iii)

Details for COG Pathway *Arginine biosynthesis*.

Add Selected to Function Cart

Select All

Clear All

Click on column name to sort.

Select	COG ID	COG Name	Genome Count
☐	COG0002	Acetylglutamate semialdehyde dehydrogenase	624

Phylogenetic Distribution for COG0002 (v)

(Hits are shown in red.)

A .01 Archaea (40)
A . .02 Crenarchaeota (10)
A . . .03 Thermoprotei (10)
A04 Desulfurococcales (1)
A05 Desulfurococcaceae (1)
A06 Aeropyrum (1)
A08 Aeropyrum pernix K1 [F] (1)
A05 unclassified
A06 Hyperthermus

Compare Genomes – Function Profile.

IMG Home | Find Genomes | Find Genes | Find Functions | Compare Genomes

Genome Statistics | VISTA | Abundance Profiles | **Function Profile** | Genome Clustering

Function List

Remove Selected | Select All | Clear All

Selection	Function ID	Name
☒	EC:1.2.4.1	Pyruvate dehydrogenase (acetyl-transferring).
☒	EC:1.4.3.16	L-aspartate oxidase.
☒	EC:1.8.1.4	Dihydrolipoyl dehydrogenase.

Genome List

Reset | Show All Genomes

Select	Ignore	Taxon Name
☐	☐	Bacteria
☐	☐	Chloroflexi
☒	☐	Chloroflexus
☒	☐	Chloroflexus aggregans DSM 9485 [D]
☒	☐	Chloroflexus aurantiacus J-10-fl [F]
☒	☐	Chloroflexus sp. Y-400-fl [D]
☐	☐	Roseiflexus
☐	☒	Roseiflexus castenholzii DSM 13941 [F]
☐	☒	Roseiflexus sp. RS-1 [F]

View Functions vs. Genomes | **View Genomes vs. Functions** →

Function Profile

Genome		EC:	EC:	EC:
		1	1	1
	D	2	4	8
		4	3	1
		1	16	4
Chloroflexus aggregans DSM 9485	B	7	1	1
Chloroflexus aurantiacus J-10-fl	B	8	1	1
Chloroflexus sp. Y-400-fl	B	4	1	0



MG-RAST Meta Genome Rapid Annotation using Subsystem Technology

Metagenomics SEED Viewer version 2.0

Welcome to the Metagenomics SEED Viewer. For more information about The SEED please visit theSEED.org.

Profiles

<http://metagenomics.nmpdr.org/>

Sequence Profile for Obese Mouse (4440464.3)

Select profile type, dataset and filter options

1

Profile type: ☒ Metabolic Profile ☐ Phylogenetic Profile

Dataset:

Maximum e-value:

Minimum p-value:

Minimum percent identity:

Minimum alignment length:

3 Re-compute results * [click here to reset](#) *

Profile results:

This Metabolic profile has been generated with the following parameters:

4 Dataset: Subsystem
Number of sequences: 11857
E-value: 0.01

Clicking on a category below will display a pie-chart of the distribution in the subcategory. In the tabular view, each category is linked to a table of the subset. Those subsets allow **downloading** in FASTA format. The organisms in the tabular view are linked to a **recruitment plot**. To download the entire dataset, please go to the [download page](#).

The pie charts provide actual counts of sequences that hit a given functional role based on the Subsystem database from the SEED. You can select a given subsystem group to get more detailed information up to 3 levels. These selections are represented in the Tabular View.

5 Classified sequences vs. non-classified:

49,442 (5662) 50,562 (5995)

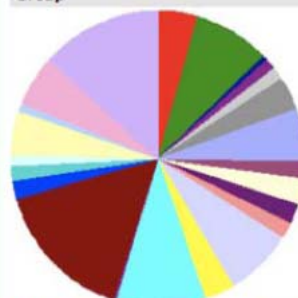
6 Charts

7 Tabular View

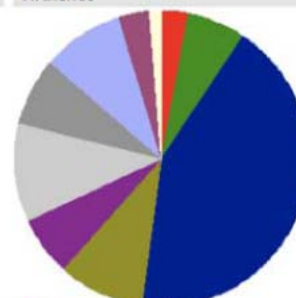
Group

Virulence

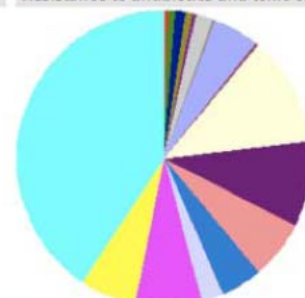
Resistance to antibiotics and toxic compounds



Cofactors, Vitamins, Prosthetic Groups, Pigments 4.57% (268)



Adhesion 3.24% (20)
Type III, Type IV, ESAT secretion systems



Methicillin resistance in Staphylococci 0.76% (2)
Tolerance to colicin E2 0.76% (2)

Compare Metagenome to other Metagenomes Heat Maps

Metagenome Heat Map for Obese Mouse (4440464.3)

Select comparison type, dataset, filter option, and metagenomes

1

Comparison type:

☒ Metabolic Comparison

☐ Phylogenetic Comparison

Dataset: Subsystem

Maximum e-value: 0.01

Minimum p-value: leave blank for all

Minimum percent identity: leave blank for all

Minimum alignment length: leave blank for all

2

Please choose some metagenomes to compare:

Columns not in display:
start typing to narrow selection
Private - F1_U
Private - 0-copper
Private - 01-004298
Private - 01-004299
Private - 01-004300

Columns in display:
Public - Lean Mouse

Re-compute results [click here to reset](#)

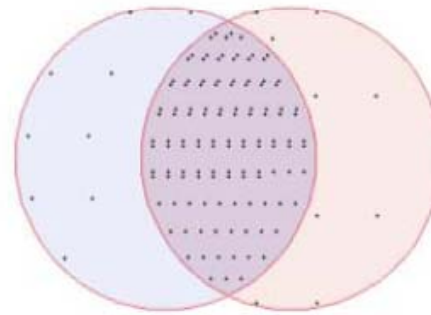
3

Classification distribution diagram and table ^[?]

Apply 'heat map' style coloring: ☒

Number of groups used in coloring: 10

Effective raw score maximum: 0.3 choose a maximum relative score as upper limit for the coloring



4
Show in table: all

5

Lean Mouse (4440463.3) Found 5872 matches in 1854 Subsystem classifications.

Obese Mouse (4440464.3) Found 5862 matches in 1872 Subsystem classifications.

Color key: 0.000 0.050 0.090 0.120 0.150 0.180 0.210 0.240 0.270 0.300

Select Subsystem Hierarchy Level to Display: Subsystem Hierarchy 2 ^[?] ☐ Display Absolute Values

6

clear all filters

export table

7

display 58 items per page

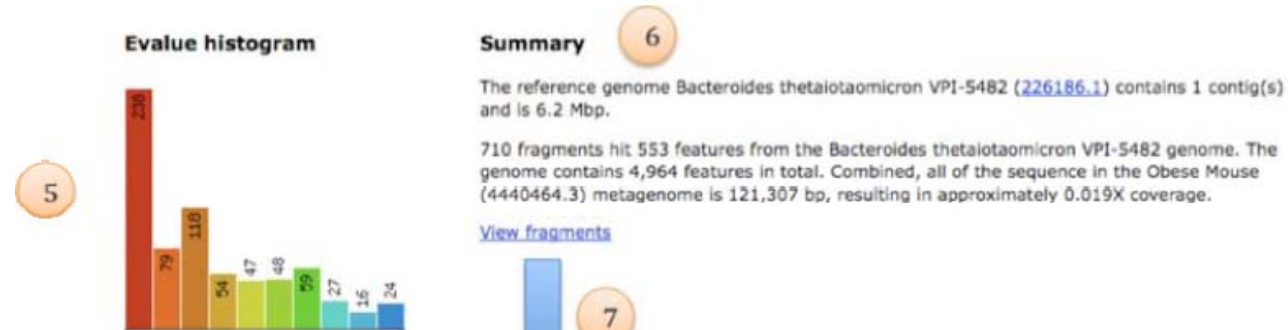
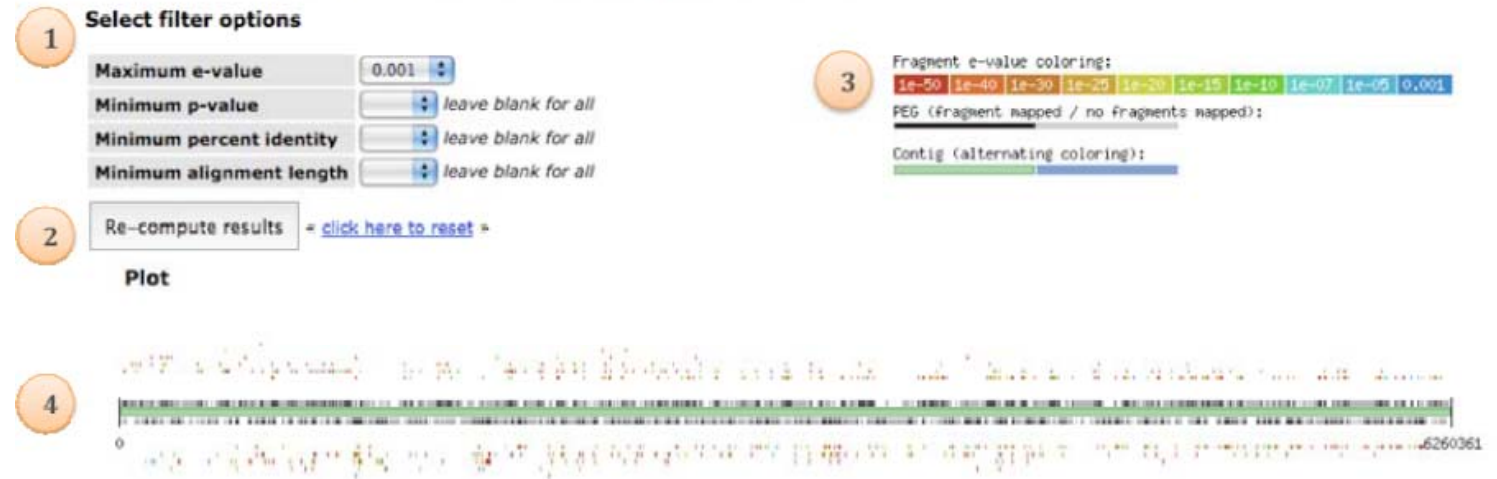
displaying 1 - 50 of 167

[next](#) [last](#)

Subsystem Hierarchy 1 ^{▲▼}	Subsystem Hierarchy 2	ID 4440463.3 ^{▲▼}	ID 4440464.3 ^{▲▼}
all			
Amino Acids and Derivatives	Alanine, serine, and glycine	0.0000	0.0000
Amino Acids and Derivatives	Arginine; urea cycle, polyamines	0.0108	0.0077
Amino Acids and Derivatives	Aromatic amino acids and derivatives	0.0000	0.0000

Compare Metagenome to Organism Recruitment Plot

Recruitment Plot from Obese Mouse (4440464.3)



Sequence Subset from Obese Mouse (4440464.3)

= [Back to Metagenome Overview](#)

The following options were used to select these sequences:

Based on metabolic reconstruction by: SEED:subsystem_tax^(?)
 Maximum e-value of hits: any
 Minimum p-value of hits: any
 Minimum percent identity of the hit: any
 Minimum alignment length of the hit: any

Found 729 sequences matching these criteria. [download as FASTA](#)

[export table](#)

display 50 items per page

displaying 1 - 50 of 729

[next](#) [last](#)

9

Select	Sequence ID	Alignment Length	Best Hit ID	Functional Role Assignment	Alignment
All					E-Value-Key
☐	46209777	231	fig 226186.1.pep.904	<i>Bacteroides thetaiotaomicron</i> VPI-5482	Hit
☐	46209789	153	fig 226186.1.pep.1036	<i>Bacteroides thetaiotaomicron</i> VPI-5482	Hit
☐	46209794	124	fig 226186.1.pep.3282	<i>Bacteroides thetaiotaomicron</i>	Hit

KEGG map *Metabolism for Obese Mouse* (4440464.3)

Select KEGG metabolic map:

start typing to narrow selection

Metabolism
Amino Acid Metabolism
Biosynthesis of Secondary Metabolites
Carbohydrate Metabolism
Energy Metabolism
Glycan Biosynthesis and Metabolism
Lipid Metabolism

1

Select upto 4 organisms for comparison:

2

Columns not in display:

start typing to narrow selection

Acidiphilium cryptum JF-5
Acidobacteria bacterium Elin345
Acidothermus cellulolyticus 118
Acidovorax avenae subsp. citrulli AAC00-1
Acidovorax sp. J542

Columns in display:

3

Select

KEGG map	Distinct ECs	Obese Mouse
Amino Acid Metabolism	545	169 (31.0 %)
Biosynthesis of Secondary Metabolites	130	20 (15.4 %)
Carbohydrate Metabolism	579	161 (27.8 %)
Energy Metabolism	167	61 (36.5 %)
Glycan Biosynthesis and Metabolism	149	40 (26.8 %)
Lipid Metabolism	262	49 (18.7 %)
Metabolism of Cofactors and Vitamins	273	99 (36.3 %)
Metabolism of Other Amino Acids	149	35 (23.5 %)
Nucleotide Metabolism	146	60 (41.1 %)
Xenobiotics Biodegradation and Metabolism	216	26 (12.0 %)

4

KEGG map *Metabolism for Obese Mouse* (4440464.3)

Select KEGG metabolic map:

start typing to narrow selection

Metabolism
Amino Acid Metabolism
Biosynthesis of Secondary Metabolites
Carbohydrate Metabolism
Energy Metabolism
Glycan Biosynthesis and Metabolism
Lipid Metabolism

Select

Select upto 4 organisms for comparison:

Columns not in display:

start typing to narrow selection

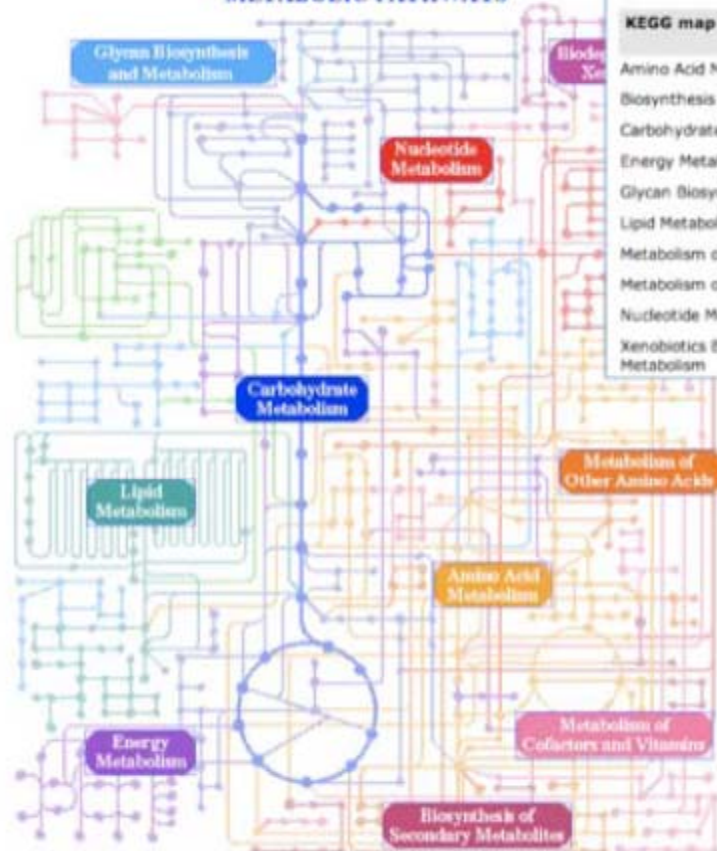
Acidiphilium cryptum JF-5
Acidobacteria bacterium Elin345
Acidothermus cellulolyticus 118
Acidovorax avenae subsp. citrulli AAC00-1
Acidovorax sp. J542

Columns in display:

Bacteroides thetaiotaomicron VPI-5482
Bacteroides vulgatus ATCC 8482

5

METABOLIC PATHWAYS



KEGG map	Distinct ECs	Obese Mouse	Bacteroides thetaiotaomicron VPI-5482	Bacteroides vulgatus ATCC 8482
Amino Acid Metabolism	545	169 (31.0 %)	152 (27.9 %)	159 (29.2 %)
Biosynthesis of Secondary Metabolites	130	20 (15.4 %)	18 (13.8 %)	20 (15.4 %)
Carbohydrate Metabolism	579	161 (27.8 %)	139 (24.0 %)	155 (26.8 %)
Energy Metabolism	167	61 (36.5 %)	51 (30.5 %)	50 (29.9 %)
Glycan Biosynthesis and Metabolism	149	40 (26.8 %)	37 (24.8 %)	38 (25.5 %)
Lipid Metabolism	262	49 (18.7 %)	38 (14.5 %)	39 (14.9 %)
Metabolism of Cofactors and Vitamins	273	99 (36.3 %)	88 (32.2 %)	98 (35.9 %)
Metabolism of Other Amino Acids	149	35 (23.5 %)	33 (22.1 %)	33 (22.1 %)
Nucleotide Metabolism	146	60 (41.1 %)	54 (37.0 %)	56 (38.4 %)
Xenobiotics Biodegradation and Metabolism	216	26 (12.0 %)	12 (5.6 %)	16 (7.4 %)

DOE Joint Genome Institute

Microbial Genomics & Metagenomics (MGM) Workshop

- Agenda
- Registration
- Travel Info

Walnut Creek, California



http://img.jgi.doe.gov/m/doc/using_index.html

<http://www.jgi.doe.gov/meetings/mgm/>



RAST User Workshop II - 2008

The National Microbial Pathogen Database Resource (NMPDR) will be hosting a 1-day workshop on **December 4th** to introduce the RAST and MG-RAST analysis servers.

<http://www.nmpdr.org/FIG/wiki/view.cgi/Main/MG-RAST>

<http://www.nmpdr.org/FIG/wiki/view.cgi/Main/HowToUseNMPDR>

Credits

- NIH-NIDCR
- JGI-Genome Biology Program (GBP)
- LBL-Biological Data Management & Technology Center (BDMTC)
- ANL-MG/RAST
- LANL-ORALGEN
 - Jimmy Saw
 - Pavel Senin
 - Andrew Liu