Next Generation Sequencing (NGS) Data Analysis for the Oral Microbiome

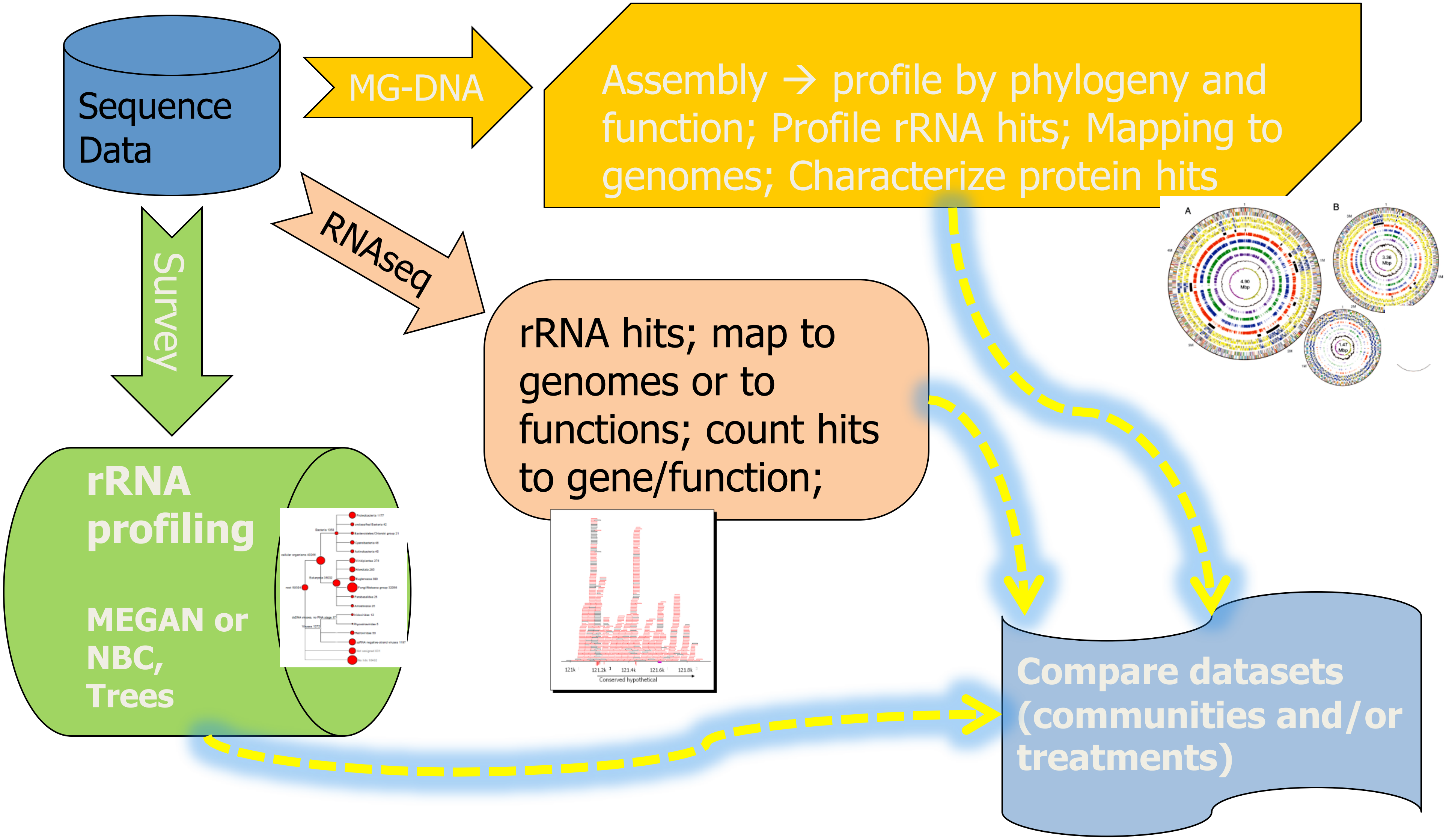
Part 2: oral metagenome annotation and analysis

Gary Xie
xie@lanl.gov

NIDCR National Institute of Dental and Craniofacial Research National Institutes of Health

Los Alamos NATIONAL LABORATORY EST. 1943

Toward a comprehensive bioinformatic workbench for the genomics community



Sequence Data

MG-DNA

Assembly → profile by phylogeny and function; Profile rRNA hits; Mapping to genomes; Characterize protein hits

RNAseq

rRNA hits; map to genomes or to functions; count hits to gene/function;

Survey

rRNA profiling

MEGAN or NBC, Trees

Compare datasets (communities and/or treatments)

Community Composition Profiling

- 16s based Community Composition Profiling
  - Pyrotags
  - RDP pyro-pipeline
  - Mothur
  - Naïve Bayesian classifier
- Random shotgun metagenome based Community Composition Profiling
  - Read based
    - MEGAN
    - Sequence Recruitment
  - Contig/protein based
    - AMPHORA

16s Based Community Composition Profiling

Limitations:

PCR introduced bias
Access to the DNA (preparation work)

- Phylochip
  - Limited by probes on the microarray
  - Relative abundance estimation

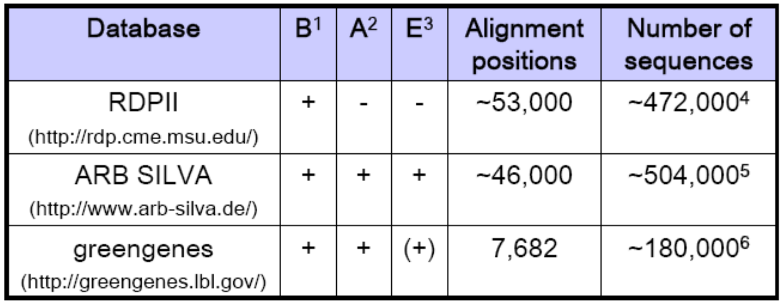
- Pyrotags (Pyrosequencing PCR amplified 16s rRNA)
  - Homopolymer quality issues
  - Shorter reads: targeted variable regions

- i-tags (Illumina sequencing PCR amplified 16s rRNA)
  - hard to find universally conserved flanking region immediately adjacent to phylogenetically informative variable region
  - region of high-quality sequence is very short (<76bp)
    - reduced phylogenetic resolution
    - more dependent on reference 16s database

Resources in 16S rRNA sequence analysis

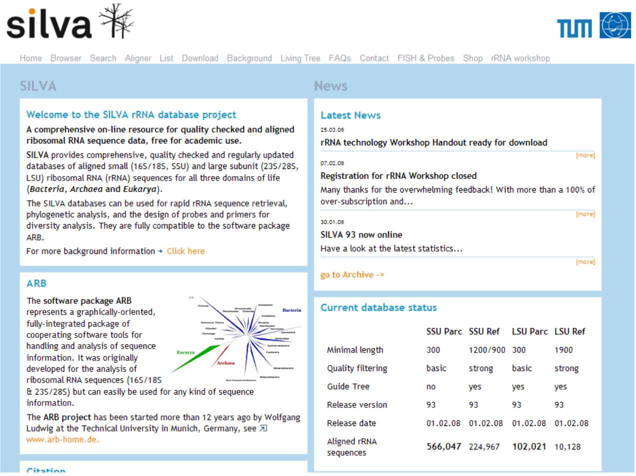
- Various 16S rRNA databases and resources available:
  - RDPII Project
  - Greengenes
  - SILVA rRNA database project

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


silva
TUM
Home Browser Search Aligner List Download Background Living Tree FAQs Contact FISH & Probes Shop rRNA workshop
Bacteria
Eucarya
Archaea

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 www.arb-home.de.

Citation

News

Latest News

25.03.08
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]

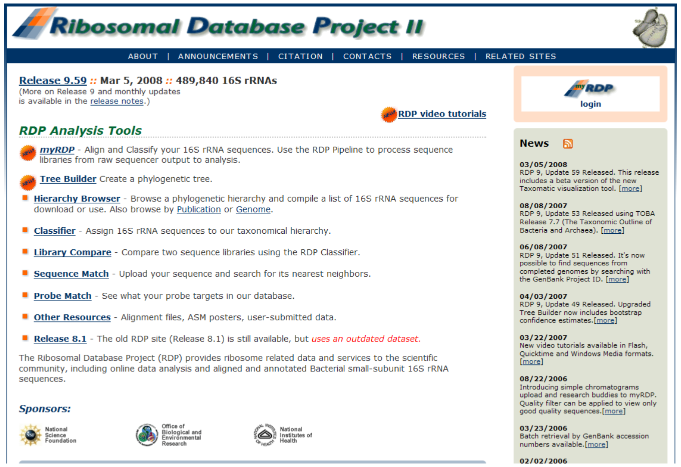
30.01.08
SILVA 93 now online
Have a look at the latest statistics...
[more]

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/


Ribosomal Database Project II
ABOUT | ANNOUNCEMENTS | CITATION | CONTACTS | RESOURCES | RELATED SITES
myRDP
login
NEW!
National Science Foundation
Office of Biological and Environmental Research
National Institutes of Health

Release 9.59 :: Mar 5, 2008 :: 489,840 16S rRNAs
(More on Release 9 and monthly updates is available in the release notes.)

RDP video tutorials

RDP Analysis Tools

- *myRDP* - Align and Classify your 16S rRNA sequences. Use the RDP Pipeline to process sequence libraries from raw sequencer output to analysis.
- Tree Builder Create a phylogenetic tree.
- Hierarchy Browser - Browse a phylogenetic hierarchy and compile a list of 16S rRNA sequences for download or use. Also browse by Publication or Genome.
- Classifier - Assign 16S rRNA sequences to our taxonomical hierarchy.
- Library Compare - Compare two sequence libraries using the RDP Classifier.
- Sequence Match - Upload your sequence and search for its nearest neighbors.
- Probe Match - See what your probe targets in our database.
- Other Resources - Alignment files, ASM posters, user-submitted data.
- Release 8.1 - The old RDP site (Release 8.1) is still available, but *uses an outdated dataset.*

The Ribosomal Database Project (RDP) provides ribosome related data and services to the scientific community, including online data analysis and aligned and annotated Bacterial small-subunit 16S rRNA sequences.

Sponsors:

News

03/05/2008
RDP 9, Update 59 Released. This release includes a beta version of the new Taxomatic visualization tool. [more]

08/08/2007
RDP 9, Update 53 Released using TOBA Release 7.7 (The Taxonomic Outline of Bacteria and Archaea). [more]

06/08/2007
RDP 9, Update 51 Released. It's now possible to find sequences from completed genomes by searching with the GenBank Project ID. [more]

04/03/2007
RDP 9, Update 49 Released. Upgraded Tree Builder now includes bootstrap confidence estimates.[more]

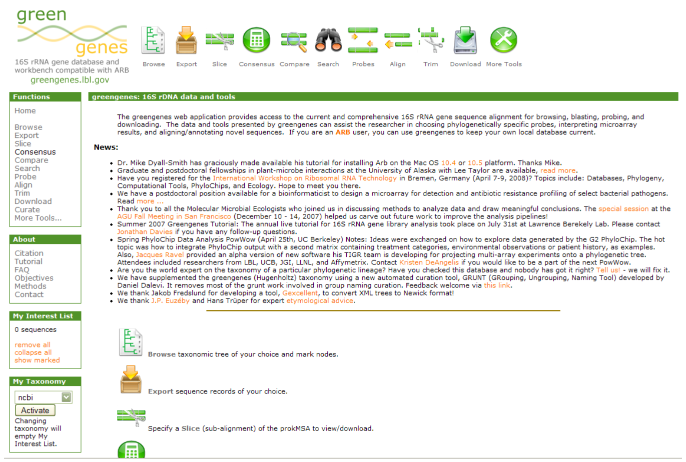
03/22/2007
New video tutorials available in Flash, Quicktime and Windows Media formats. [more]

08/22/2006
Introducing simple chromatograms upload and research buddies to myRDP. Quality filter can be applied to view only good quality sequences.[more]

03/23/2006
Batch retrieval by GenBank accession numbers available.[more]

02/02/2006

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



0 sequences

Changing taxonomy will empty My Interest List.

greengenes: 16S rDNA data and tools

The greengenes web application provides access to the current and comprehensive 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.

News:

- Dr. Mike Dyall-Smith has graciously made available his tutorial for installing Arb on the Mac OS 10.4 or 10.5 platform. Thanks Mike.
- Graduate and postdoctoral fellowships in plant-microbe interactions at the University of Alaska with Lee Taylor are available, read more.
- Have you registered for the International Workshop on Ribosomal RNA Technology in Bremen, Germany (April 7-9, 2008)? Topics include: Databases, Phylogeny, Computational Tools, PhyloChips, and Ecology. Hope to meet you there.
- We have a postdoctoral position available for a bioinformaticist to design a microarray for detection and antibiotic resistance profiling of select bacterial pathogens. Read more ...
- Thank you to all the Molecular Microbial Ecologists who joined us in discussing methods to analyze data and draw meaningful conclusions. The special session at the AGU Fall Meeting in San Francisco (December 10 - 14, 2007) helped us carve out future work to improve the analysis pipelines!
- Summer 2007 Greengenes Tutorial: The annual live tutorial for 16S rRNA gene library analysis took place on July 31st at Lawrence Berekely Lab. Please contact Jonathan Davies if you have any follow-up questions.
- Spring PhyloChip Data Analysis PowWow (April 25th, UC Berkeley) Notes: Ideas were exchanged on how to explore data generated by the G2 PhyloChip. The hot topic was how to integrate PhyloChip output with a second matrix containing treatment categories, environmental observations or patient history, as examples. Also, Jacques Ravel provided an alpha version of new software his TIGR team is developing for projecting multi-array experiments onto a phylogenetic tree. Attendees included researchers from LBL, UCB, JGI, LLNL, and Affymetrix. Contact Kristen DeAngelis if you would like to be a part of the next PowWow.
- 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 have supplemented the greengenes (Hugenholtz) taxonomy using a new automated curation tool, GRUNT (GRouping, Ungrouping, Naming Tool) developed by Daniel Dalevi. It removes most of the grunt work involved in group naming curation. Feedback welcome via this link.
- 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.

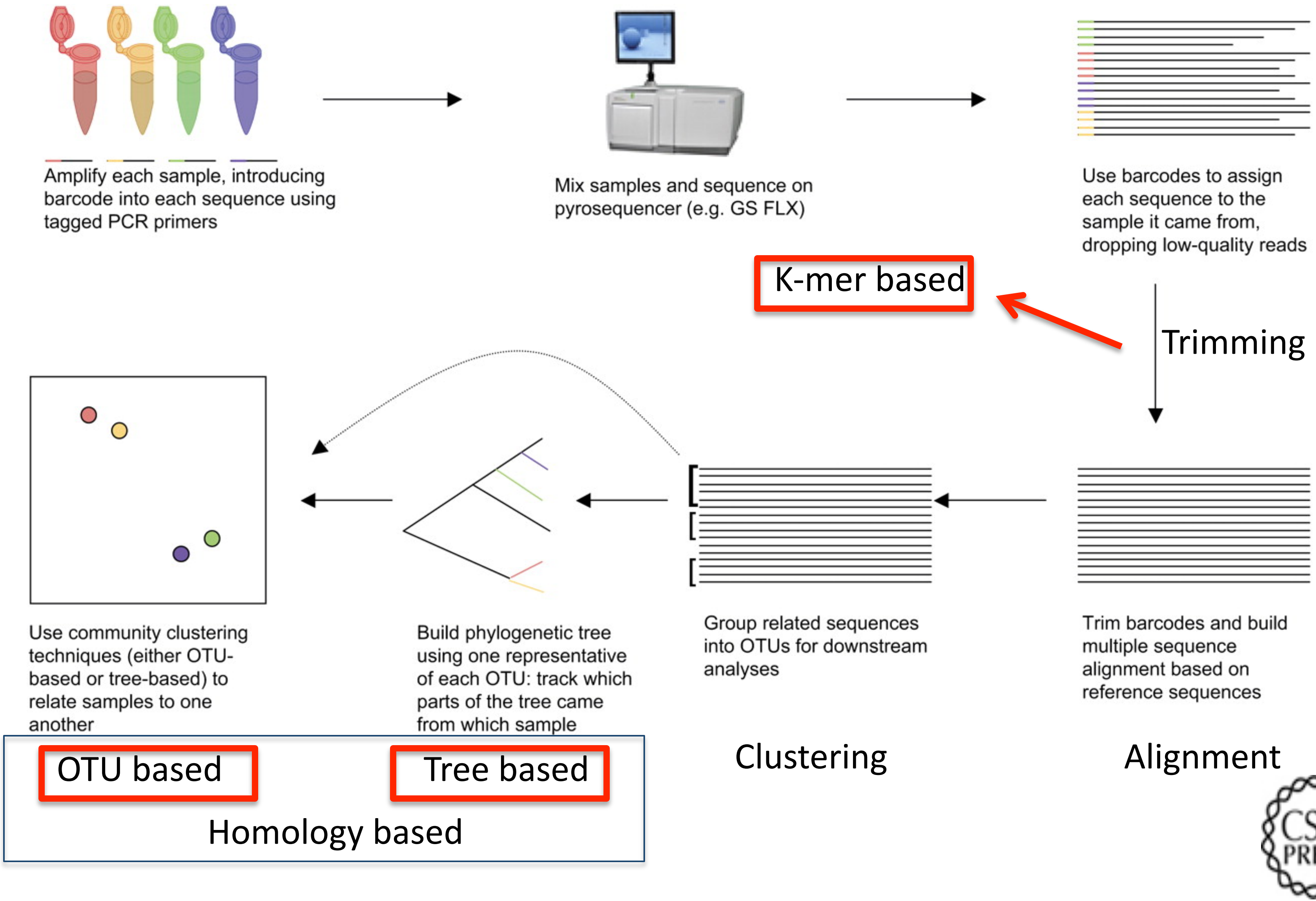
Browse taxonomic tree of your choice and mark nodes.

Export sequence records of your choice.

Specify a **Slice** (sub-alignment) of the prokMSA to view/download.

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

Overview of barcoded pyrosequencing workflow for rRNA-based community composition profiling



Amplify each sample, introducing barcode into each sequence using tagged PCR primers

Mix samples and sequence on pyrosequencer (e.g. GS FLX)

Use barcodes to assign each sequence to the sample it came from, dropping low-quality reads

K-mer based

Trimming

Trim barcodes and build multiple sequence alignment based on reference sequences

Alignment

Group related sequences into OTUs for downstream analyses

Clustering

Build phylogenetic tree using one representative of each OTU: track which parts of the tree came from which sample

Use community clustering techniques (either OTU-based or tree-based) to relate samples to one another

OTU based

Tree based

Homology based

Hamady M, Knight R Genome Res. 2009;19:1141-1152

Analysis Toolbox

K-mer based approach

- NBC
- Markov

OTU-based approach

- DOTUR: define & count OUT (distance matrix as input)
- SONS: measuring overlap between communities (OTU designation as input)
- Estimate S - beta diversity index (Need abundance file for each OTU)

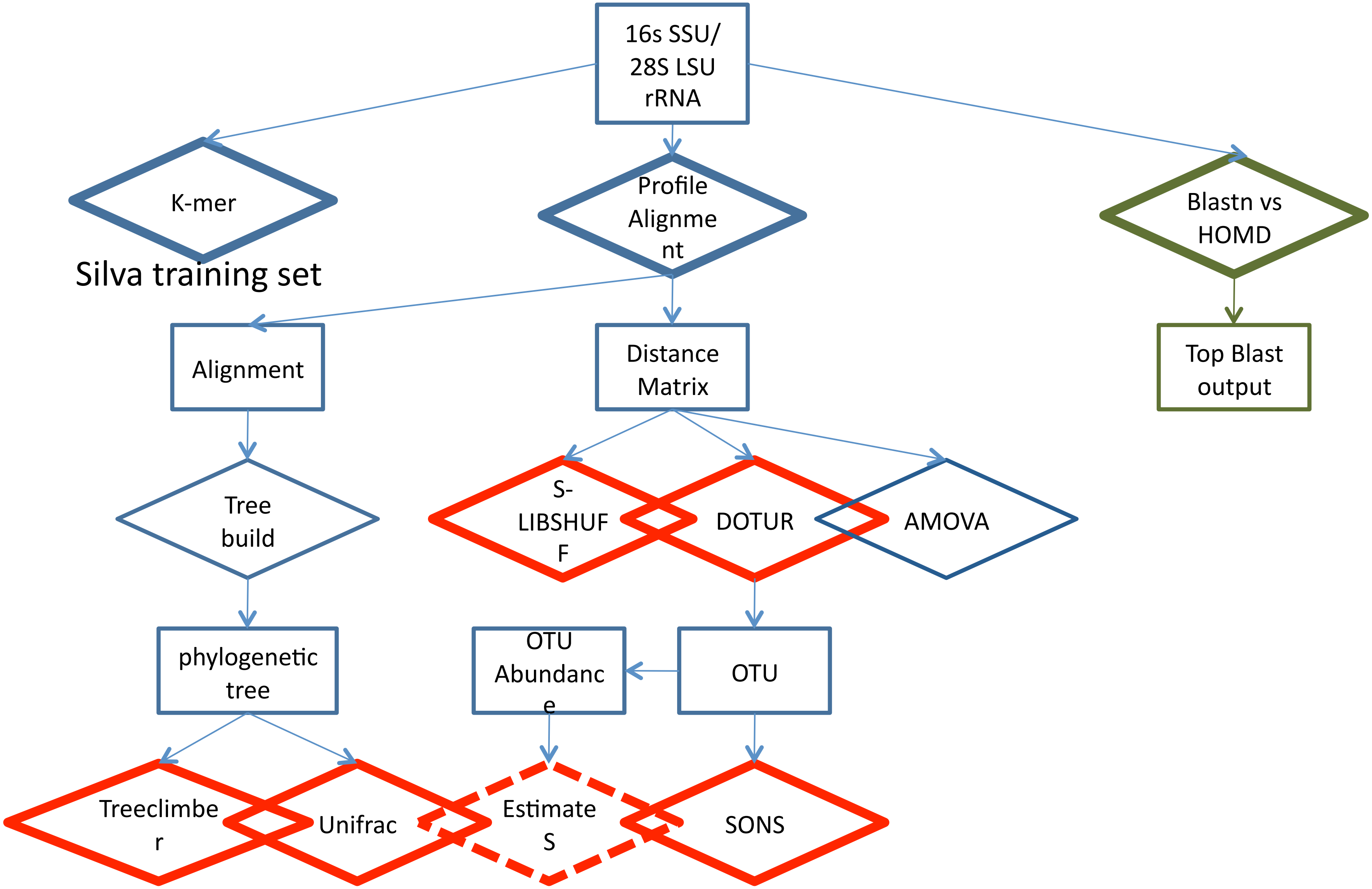
Tree-based approach

- S-LIBSHUFF (distance matrix as input)
- Treeclimber (phylogenetic tree as input)
- Unifrac (phylogenetic tree as input)

Patrick Schloss' Mothur has most of these packages

SA Eichorst

Comparing community memberships and structures

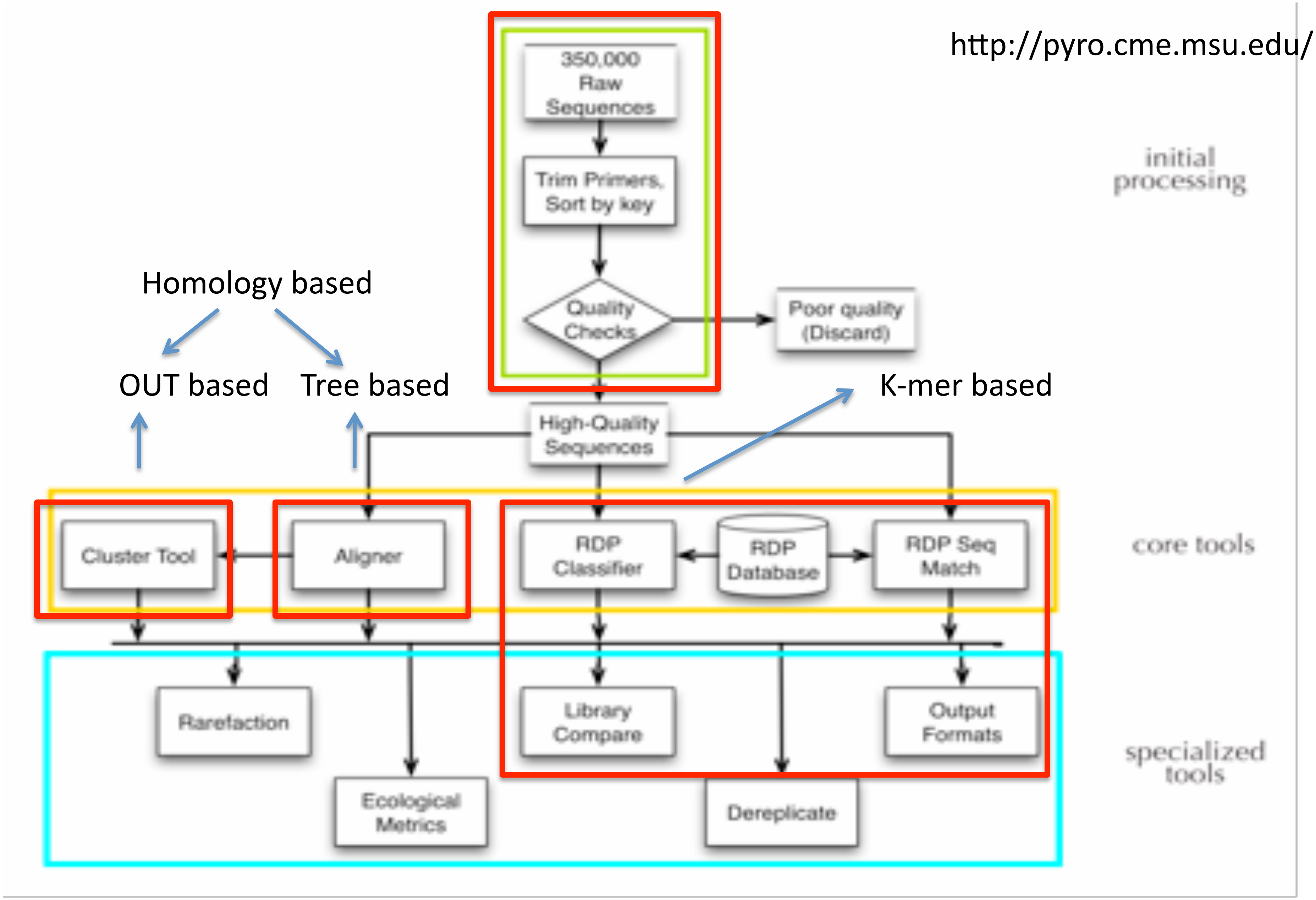


Patrick Schloss' Mothur

http://www.mothur.org/

RDP has better GUI and plug-in

RDP Pyrosequencing Pipeline



http://pyro.cme.msu.edu/

Homology based

OUT based

Tree based

K-mer based

RDP Pyrosequencing Pipeline

Pro:

- Including the initial process: sorting, trimming tag and primers, removing low quality
- Aligner based on the 2nd structure (Nawrocki & Eddy, 2007)
- Including complete-linkage clustering, formatted for downstream analysis
- Including package for study community richness & diversity:
    - Generating Shannon/Chao1
    - Jaccard/sorensen index
    - Rarefaction curve
- Including the K-mer analysis: NBC and library comparison

Con:

- Only includes "Bacteria" and "Archaea" sequences, no Eukaryote
- No oral specific training set
- Library comparison only handle two samples at a time

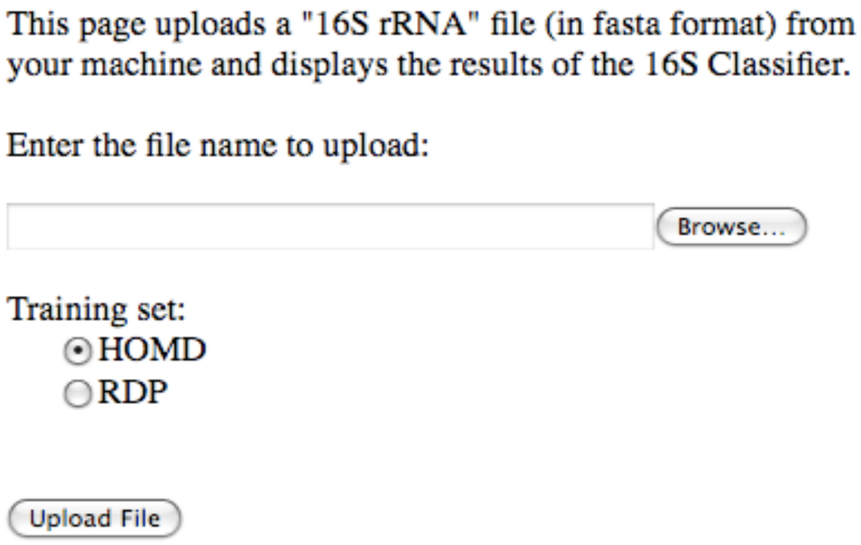
K-mer based NBC and LANL-Markov classifier

* **No multiple sequence alignment** required

* Provide confidence estimates for each assignment.

* Tested on the leave-one-out cross validation (LOOCV)

* Different from RDP classifier, our Markov classifier assumes a dependency on sequence reads



This page uploads a "16S rRNA" file (in fasta format) from your machine and displays the results of the 16S Classifier.

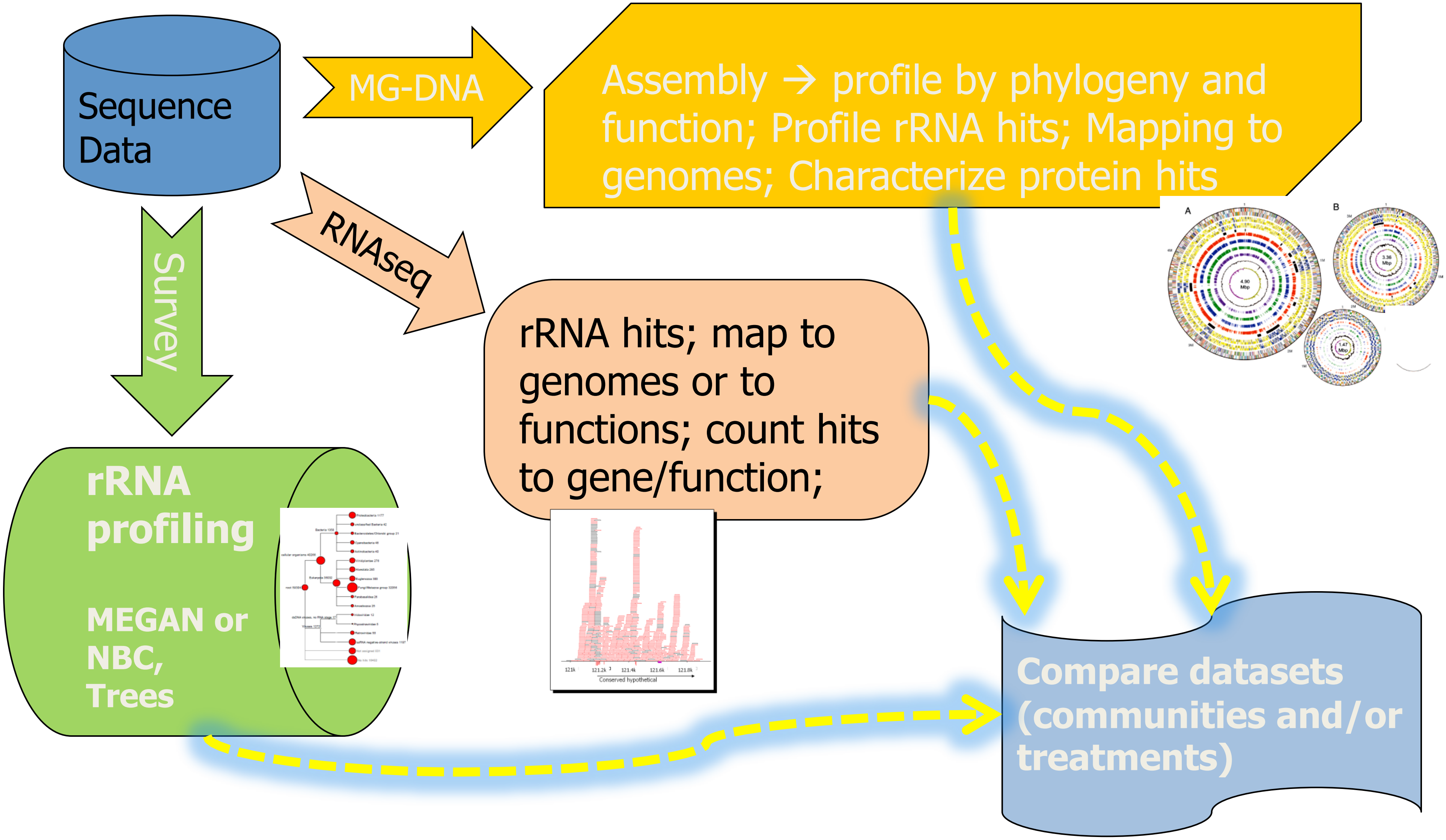
Enter the file name to upload:

Training set:

● HOMD
○ RDP

http://oralgen.lanl.gov/oralgen-tng/missimp.html

Toward a comprehensive bioinformatic workbench for the genomics community



Sequence Data

MG-DNA

Assembly → profile by phylogeny and function; Profile rRNA hits; Mapping to genomes; Characterize protein hits

RNAseq

rRNA hits; map to genomes or to functions; count hits to gene/function;

Survey

rRNA profiling

MEGAN or NBC, Trees

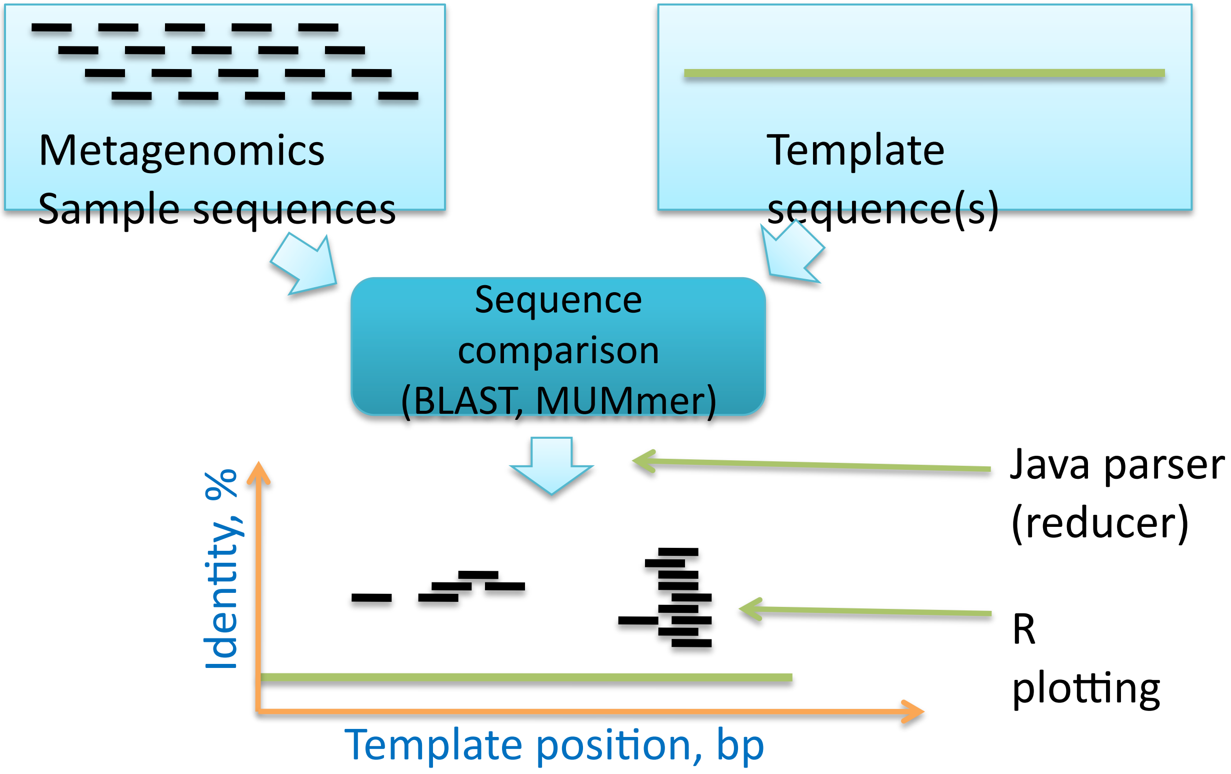
Compare datasets (communities and/or treatments)

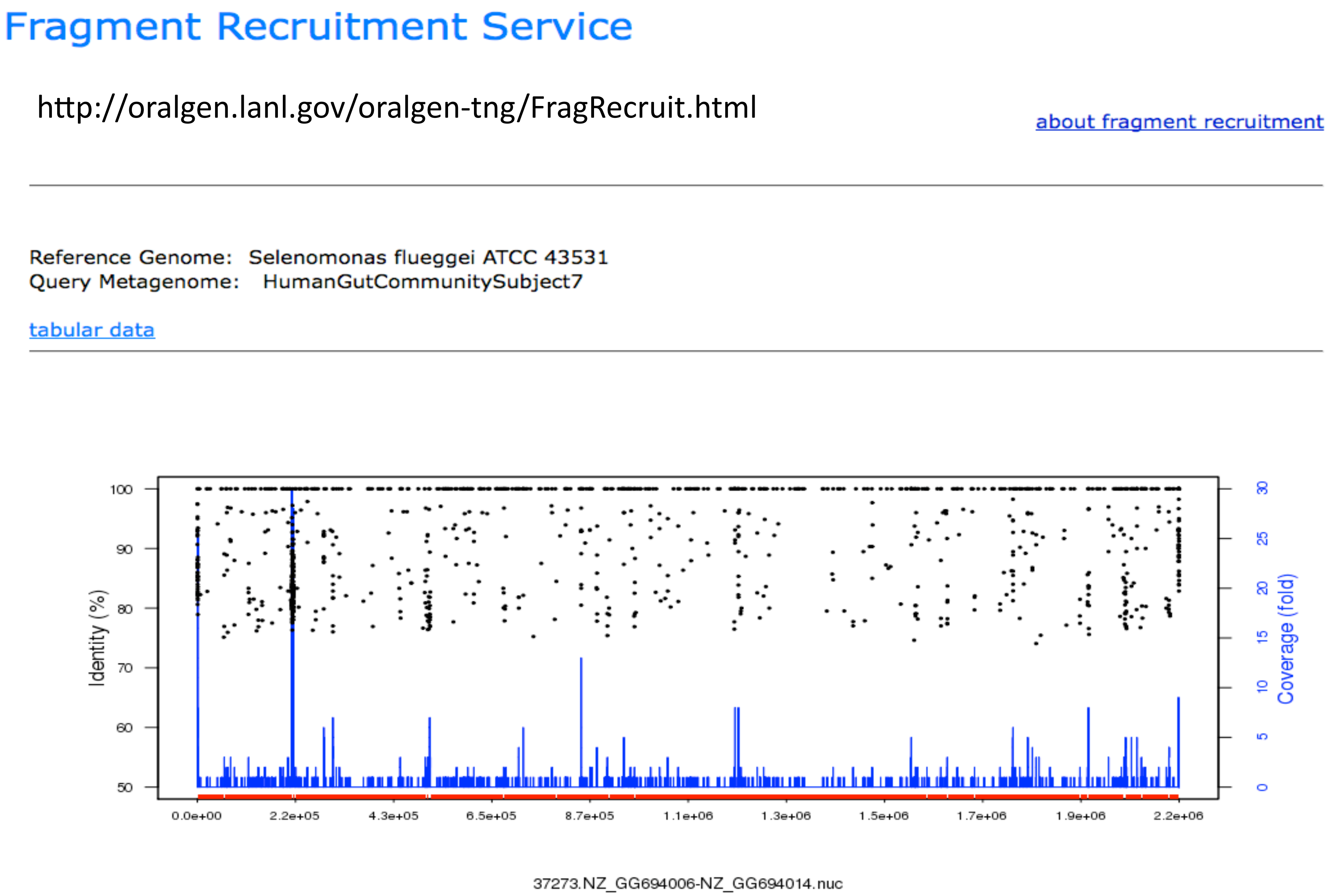
Community Composition Profiling

- 16s based Community Composition Profiling
  - Pyrotags
  - RDP pyro-pipeline
  - Mothur
  - Naïve Bayesian classifier
- Random shotgun based Community Composition Profiling
  - Read based
    - Sequence Recruitment
    - MEGAN
  - Contig/protein based
    - AMPHORA

Read based approach: Sequence Recruitment

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





Fragment Recruitment Service

http://oralgen.lanl.gov/oralgen-tng/FragRecruit.html

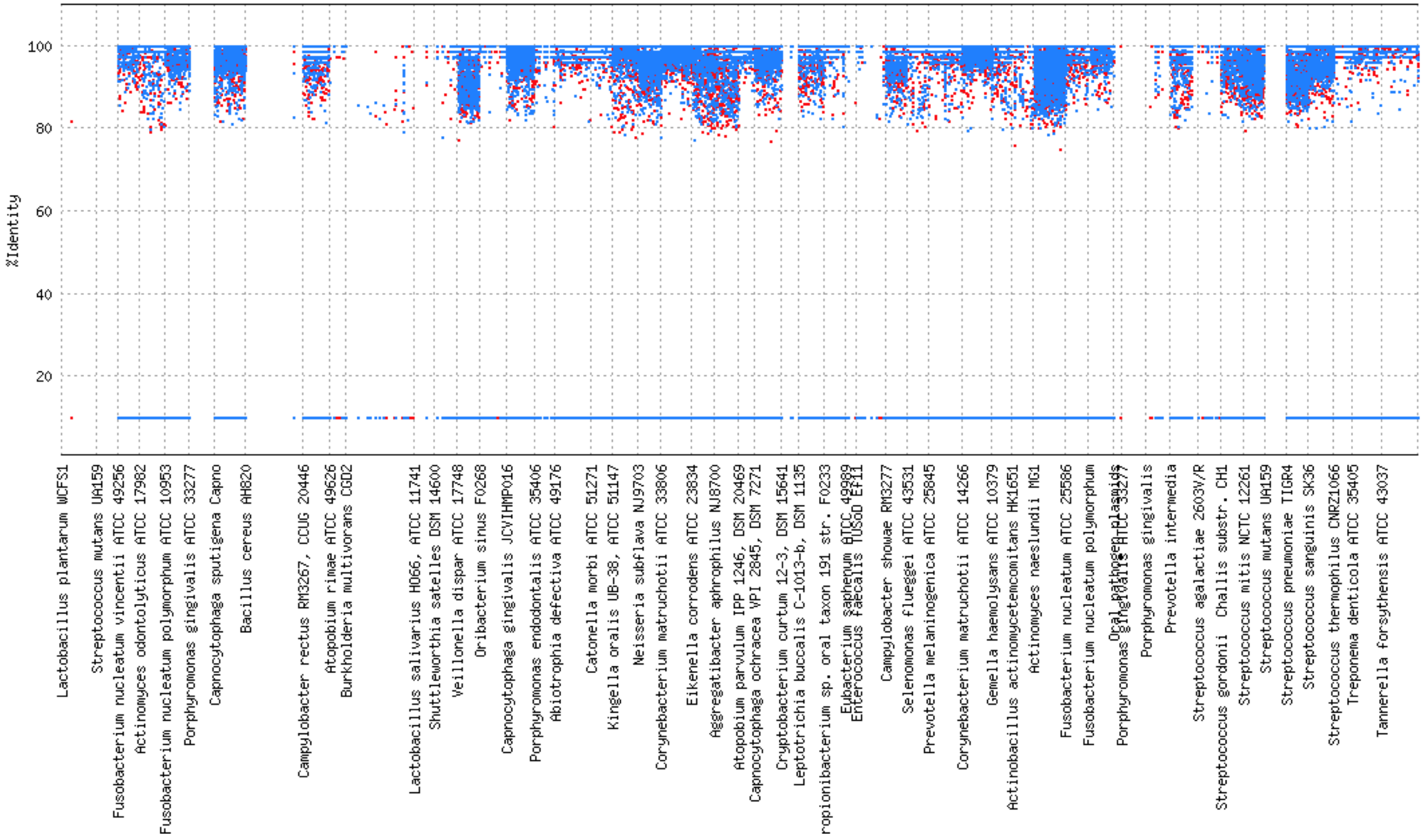
about fragment recruitment

Reference Genome: Selenomonas flueggei ATCC 43531
Query Metagenome: HumanGutCommunitySubject7

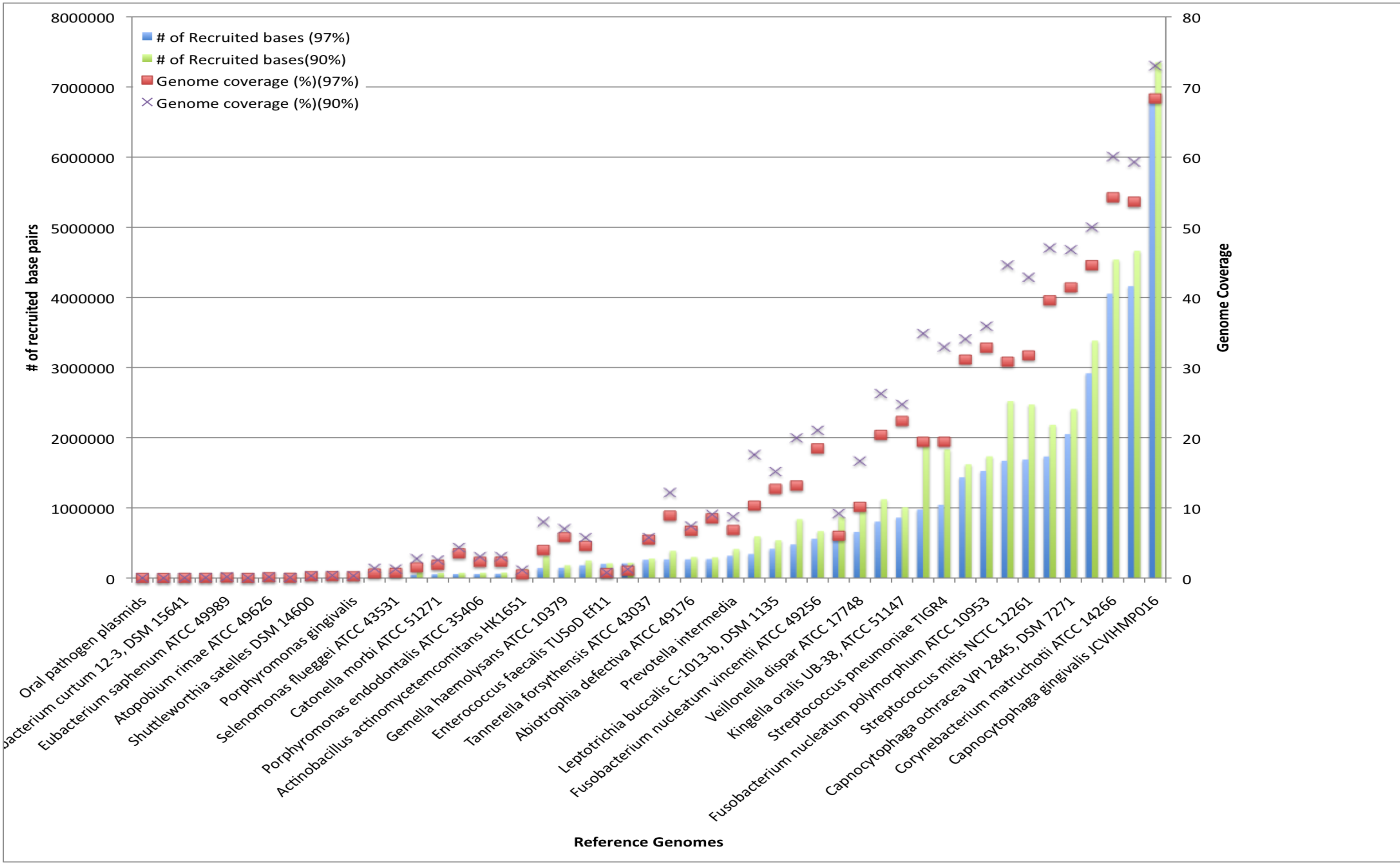
tabular data

37273.NZ_GG694006-NZ_GG694014.nuc

Mapping the metagenome reads to reference genomes



HMP reference genomes: a starting point for reconstruction of microbial genomes from metagenomic sequences



Read based: lowest common ancestor vs top blast approach

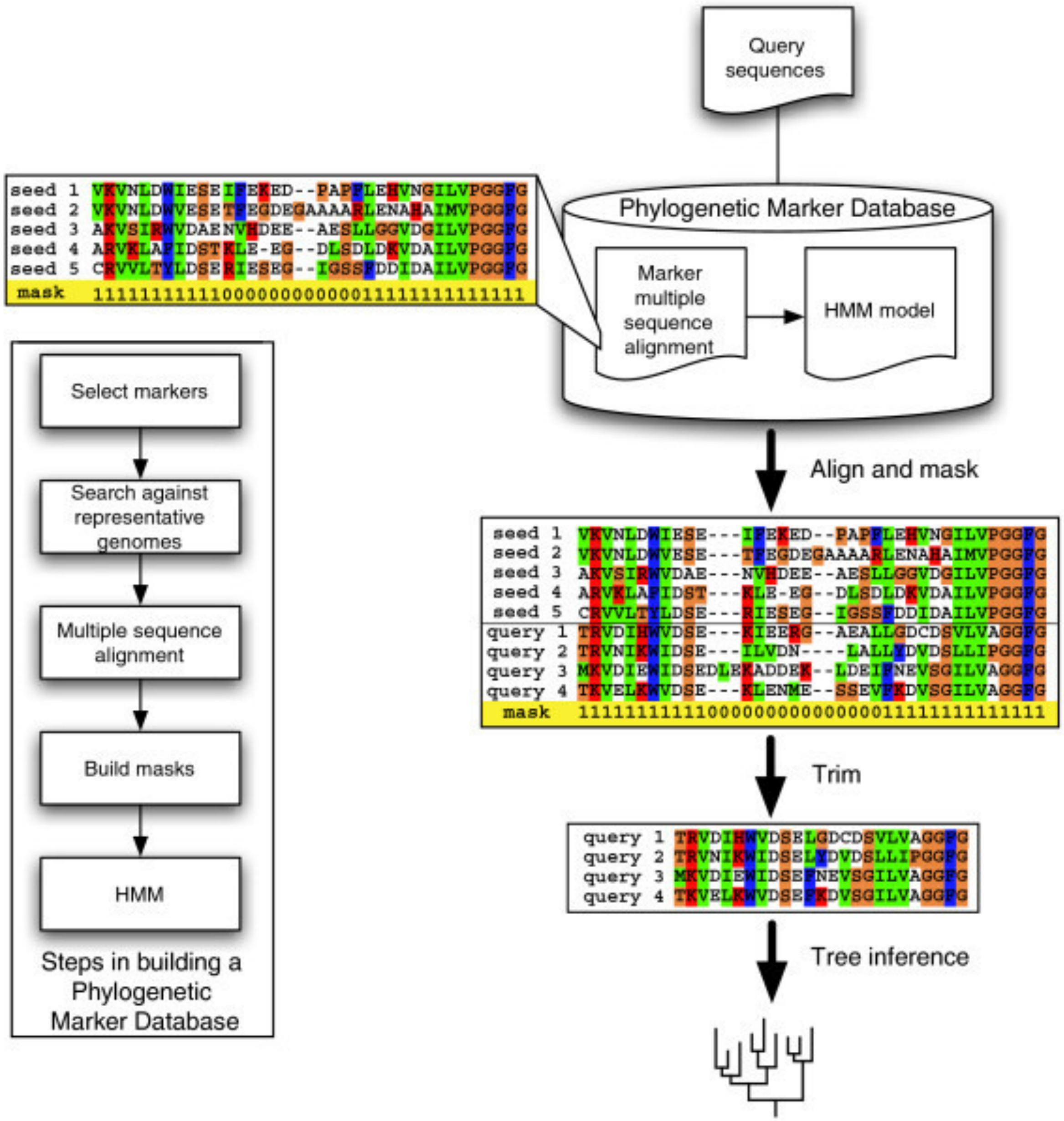
Top Blast hit:
–The poor representation of microbial diversity by sequenced isolates
–remote matching to phylogenetically distant organisms or the absence of any hits.

Such as MG-RAST, HOMD

Lowest common ancestor:
–All reads count
–Trace back to the lowest common ancestor of the set of taxa.

Such as MEGAN, AMPHORA

Contig/protein based: Automated pipeline for phylogenomic analysis (AMPHORA)


Query sequences
Phylogenetic Marker Database
Marker multiple sequence alignment
HMM model
Align and mask
Trim
Tree inference
Select markers
Search against representative genomes
Multiple sequence alignment
Build masks
HMM
Steps in building a Phylogenetic Marker Database

```
seed 1  VKVNLDWIESEIFEKED--PAPFLEHVNGILVPGGFG
seed 2  VKVNLDWVESETFEGDEGAAAARLENAHAIMVPGGFG
seed 3  AKVSIRWVDAENVHDEE--AESLLGGVDGILVPGGFG
seed 4  ARVKLAFIDSTKLE-EG--DLSDLDKVDAILVPGGFG
seed 5  CRVVLTYLDSERIESEG--IGSSFDDIDAILVPGGFG
 mask   1111111111100000000000011111111111111
```

```
 seed 1  VKVNLDWIESE---IFEKED--PAPFLEHVNGILVPGGFG
 seed 2  VKVNLDWVESE---TFEGDEGAAAARLENAHAIMVPGGFG
 seed 3  AKVSIRWVDAE---NVHDEE--AESLLGGVDGILVPGGFG
 seed 4  ARVKLAFIDST---KLE-EG--DLSDLDKVDAILVPGGFG
 seed 5  CRVVLTYLDSE---RIESEG--IGSSFDDIDAILVPGGFG
query 1  TRVDIHWVDSE---KIEERG--AEALLGDCDSVLVAGGFG
query 2  TRVNIKWIDSE---ILVDN-----LALLYDVDSLLIPGGFG
query 3  MKVDIEWIDSEDLEKADDEK--LDEIFNEVSGILVAGGFG
query 4  TKVELKWVDSE---KLENME--SSEVFKDVSGILVAGGFG
  mask   1111111111100000000000000011111111111111
```

```
query 1  TRVDIHWVDSELGDCDSVLVAGGFG
query 2  TRVNIKWIDSELYDVDSLLIPGGFG
query 3  MKVDIEWIDSEFNEVSGILVAGGFG
query 4  TKVELKWVDSEFKDVSGILVAGGFG
```

Two applications:

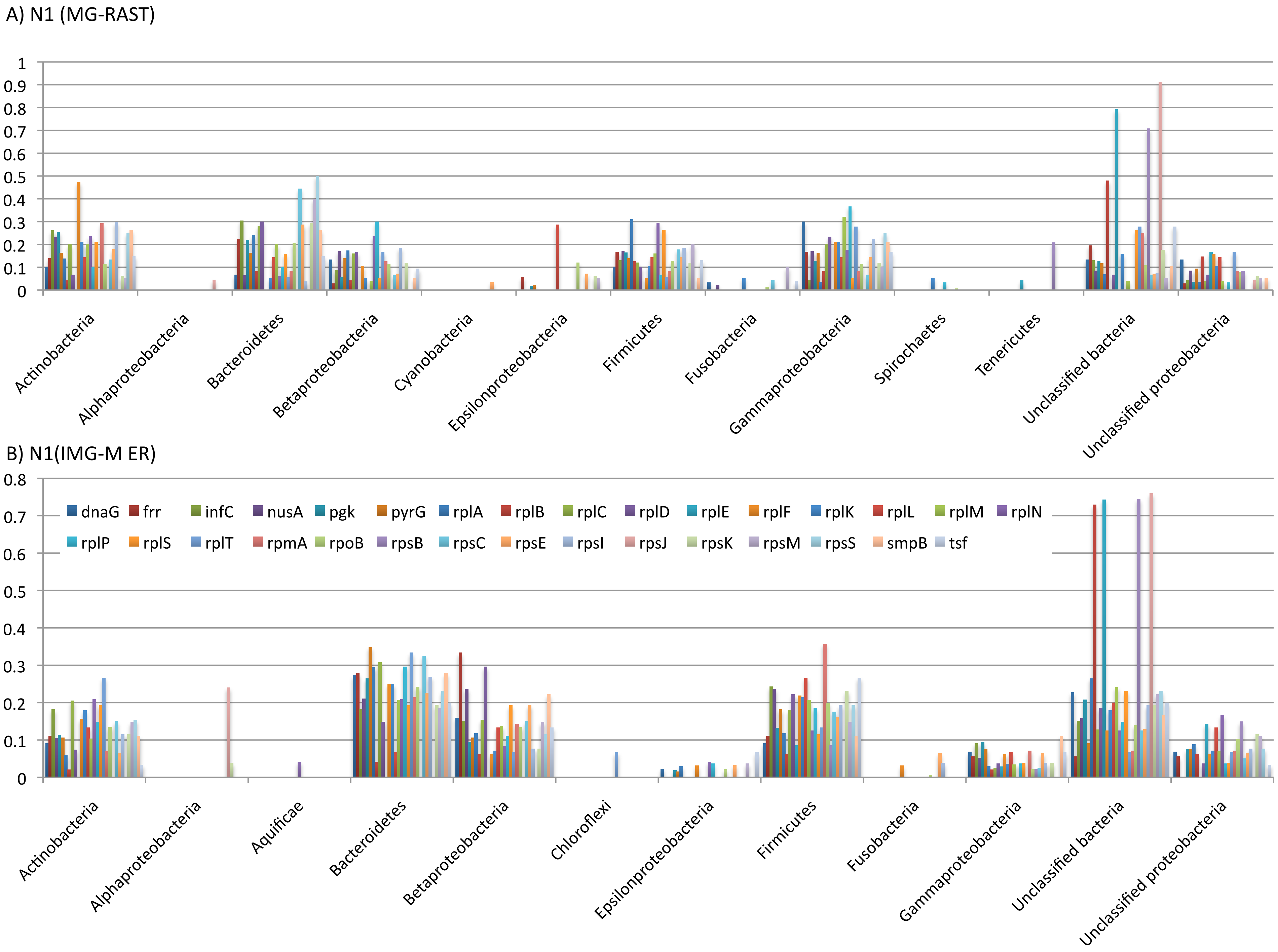
– Build a genome tree from 578 complete bacterial genomes

– Identify bacterial phylotypes from metagenomic data

Wu et al 2008

AMPHORA Features

- Fully Automated: in a pipeline
- Can be used for phylogenetic analyses of single gene or whole microbiomes.
- 31 pre-build phylogenetic marker genes
  - Most are single copy genes within each genome
  - Housekeeping genes
- Profile HMM-based multiple sequence alignment
  - High quality alignment according to seed alignment
  - Reproducible
  - Speed


A) N1 (MG-RAST)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Actinobacteria
Alphaproteobacteria
Bacteroidetes
Betaproteobacteria
Cyanobacteria
Epsilonproteobacteria
Firmicutes
Fusobacteria
Gammaproteobacteria
Spirochaetes
Tenericutes
Unclassified bacteria
Unclassified proteobacteria
B) N1(IMG-M ER)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
dnaG
frr
infC
nusA
pgk
pyrG
rplA
rplB
rplC
rplD
rplE
rplF
rplK
rplL
rplM
rplN
rplP
rplS
rplT
rpmA
rpoB
rpsB
rpsC
rpsE
rpsI
rpsJ
rpsK
rpsM
rpsS
smpB
tsf
Actinobacteria
Alphaproteobacteria
Aquificae
Bacteroidetes
Betaproteobacteria
Chloroflexi
Epsilonproteobacteria
Firmicutes
Fusobacteria
Gammaproteobacteria
Unclassified bacteria
Unclassified proteobacteria

Summary: Community Composition Profiling

Approaches for studying the community composition

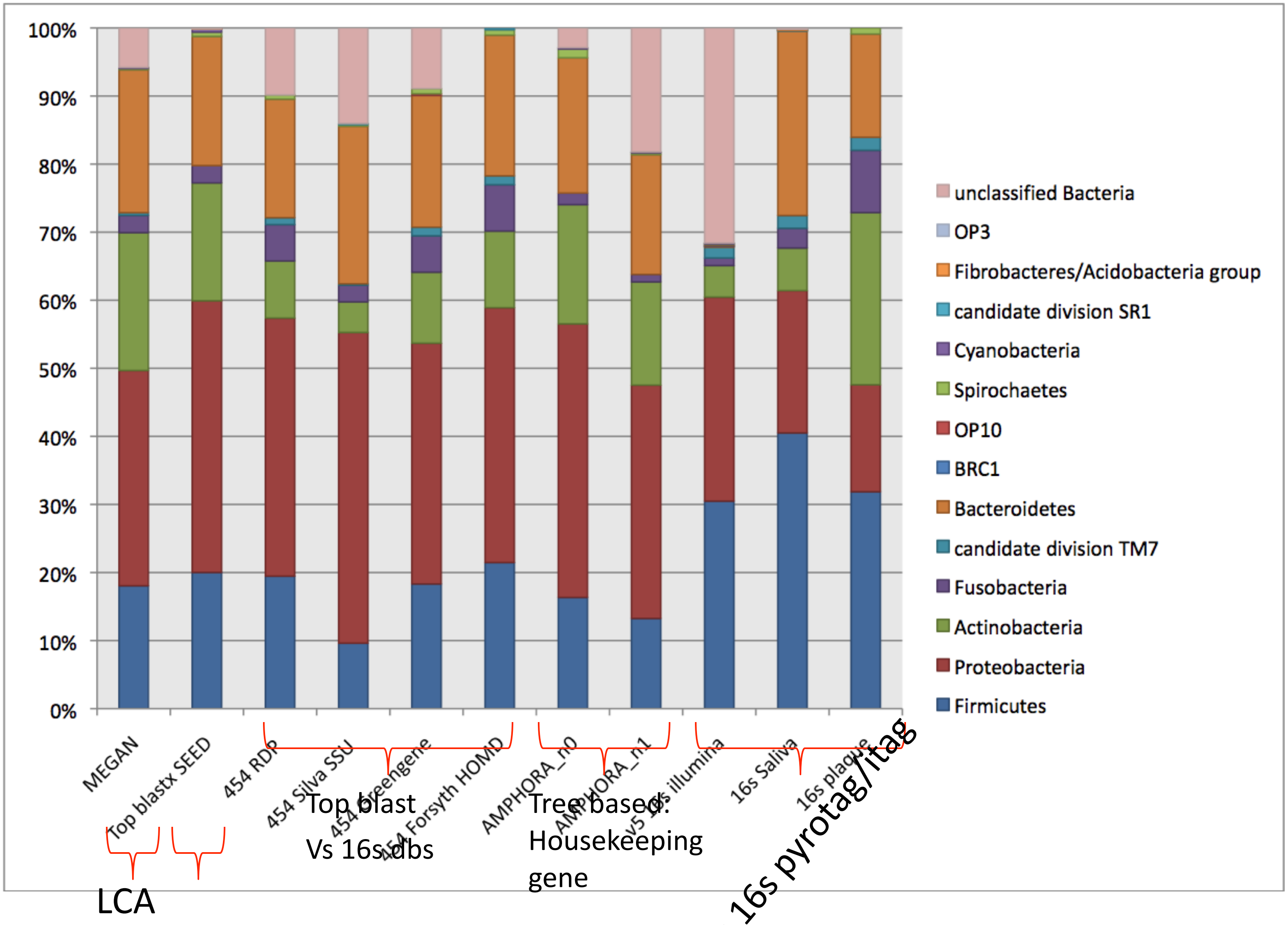
- Homology based
  - Lowest common ancestor (MEGAN)
  - Top blast hits (shotgun reads, 16s)
  - Phylogenetic tree (31 house keeping gene)

- K-mer character based
  - NBC
  - Markov

Input data used:

- Targeted:
  - 16s rRNA reads
  - Other housekeeping genes
- Random shotgun reads

Combination of different strategies



Top blast
Vs 16s dbs

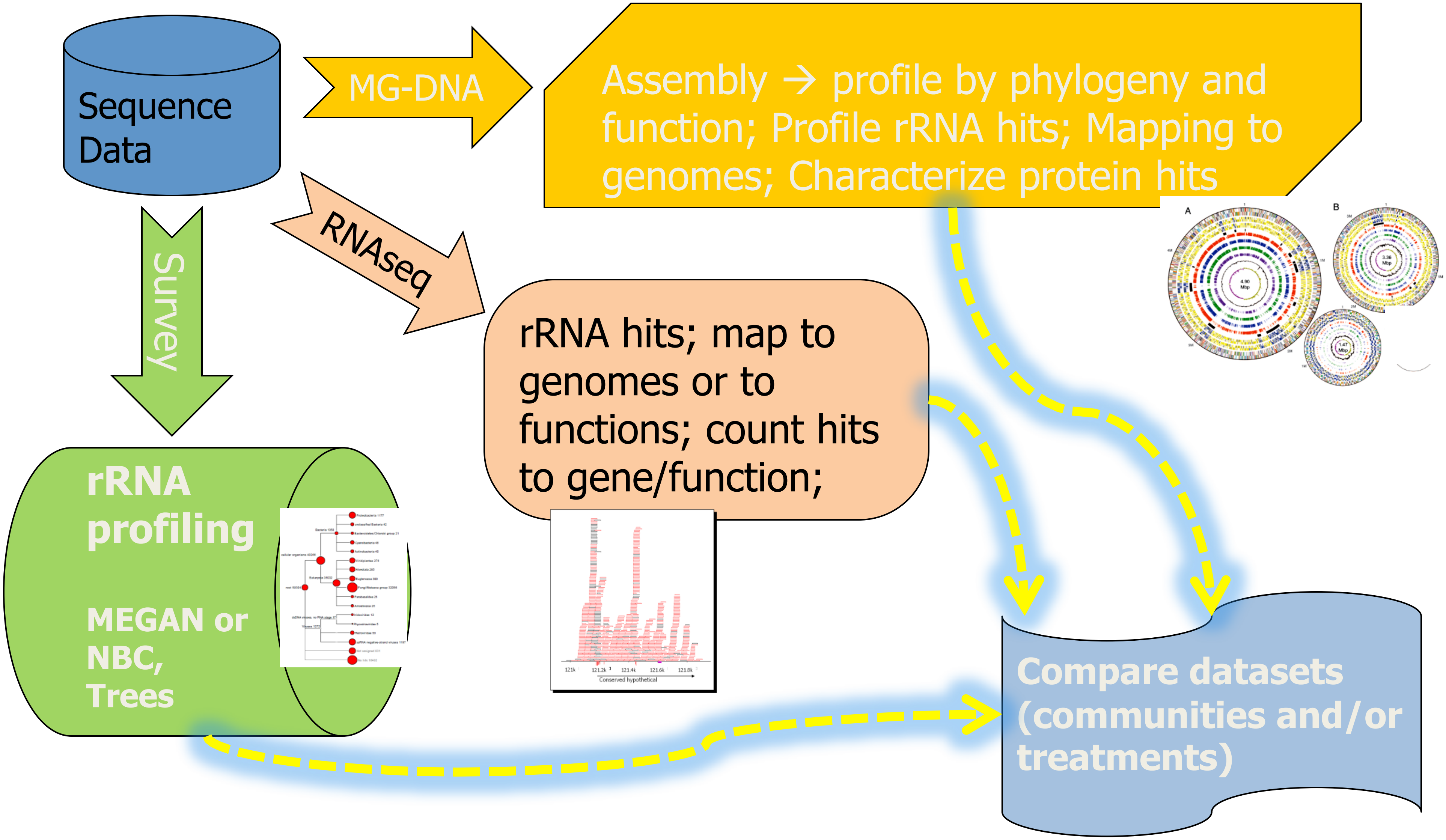
Tree based:
Housekeeping
gene

LCA

Top Blast

Previous 16s pyrotag/itag

Toward a comprehensive bioinformatic workbench for the genomics community



Sequence Data

MG-DNA

Assembly → profile by phylogeny and function; Profile rRNA hits; Mapping to genomes; Characterize protein hits

RNAseq

rRNA hits; map to genomes or to functions; count hits to gene/function;

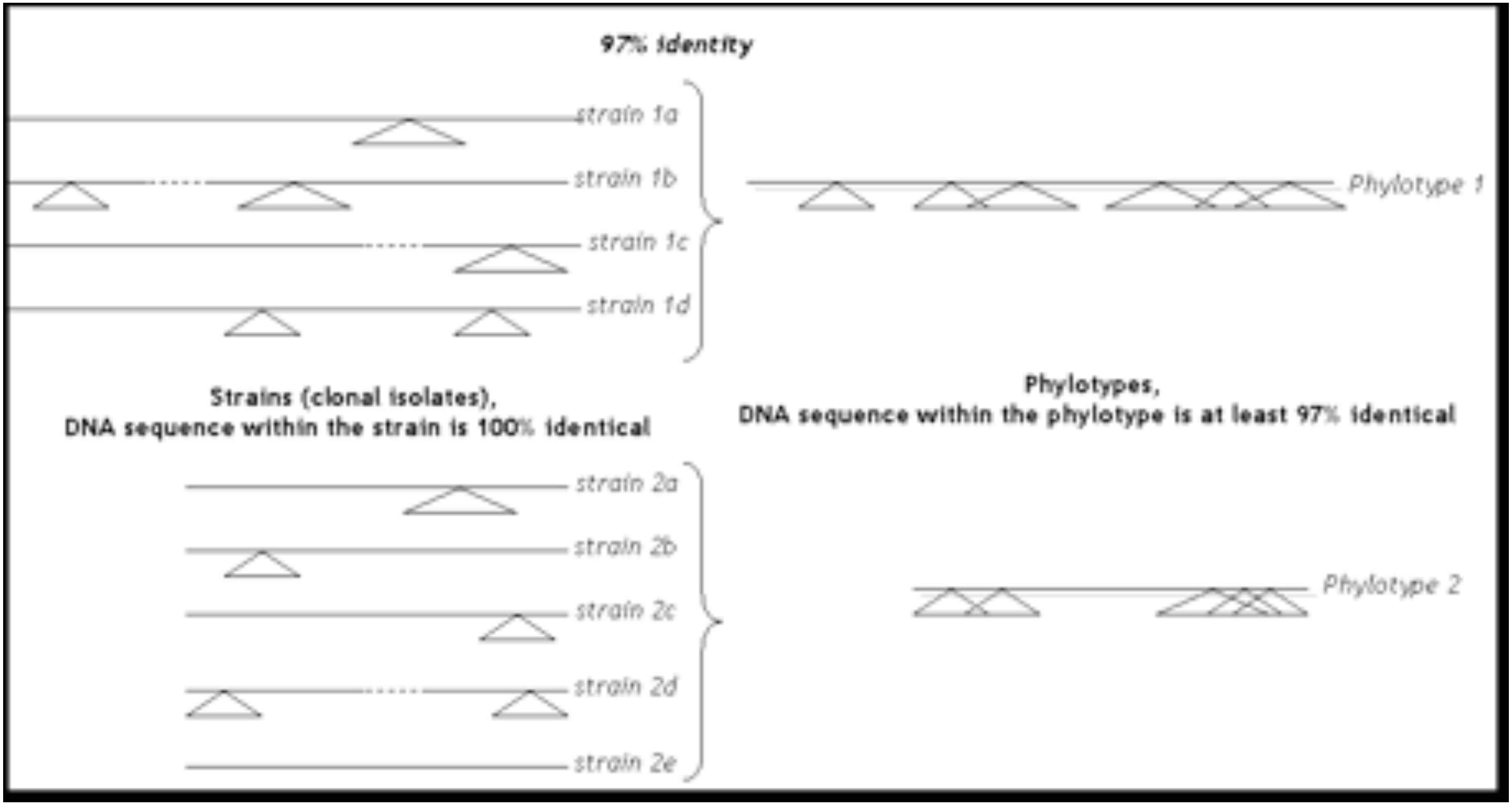
Survey

rRNA profiling

MEGAN or NBC, Trees

Compare datasets (communities and/or treatments)

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


97% identity
strain 1a
strain 1b
strain 1c
strain 1d
Phylotype 1
Strains (clonal isolates),
DNA sequence within the strain is 100% identical
Phylotypes,
DNA sequence within the phylotype is at least 97% identical
strain 2a
strain 2b
strain 2c
strain 2d
strain 2e
Phylotype 2

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

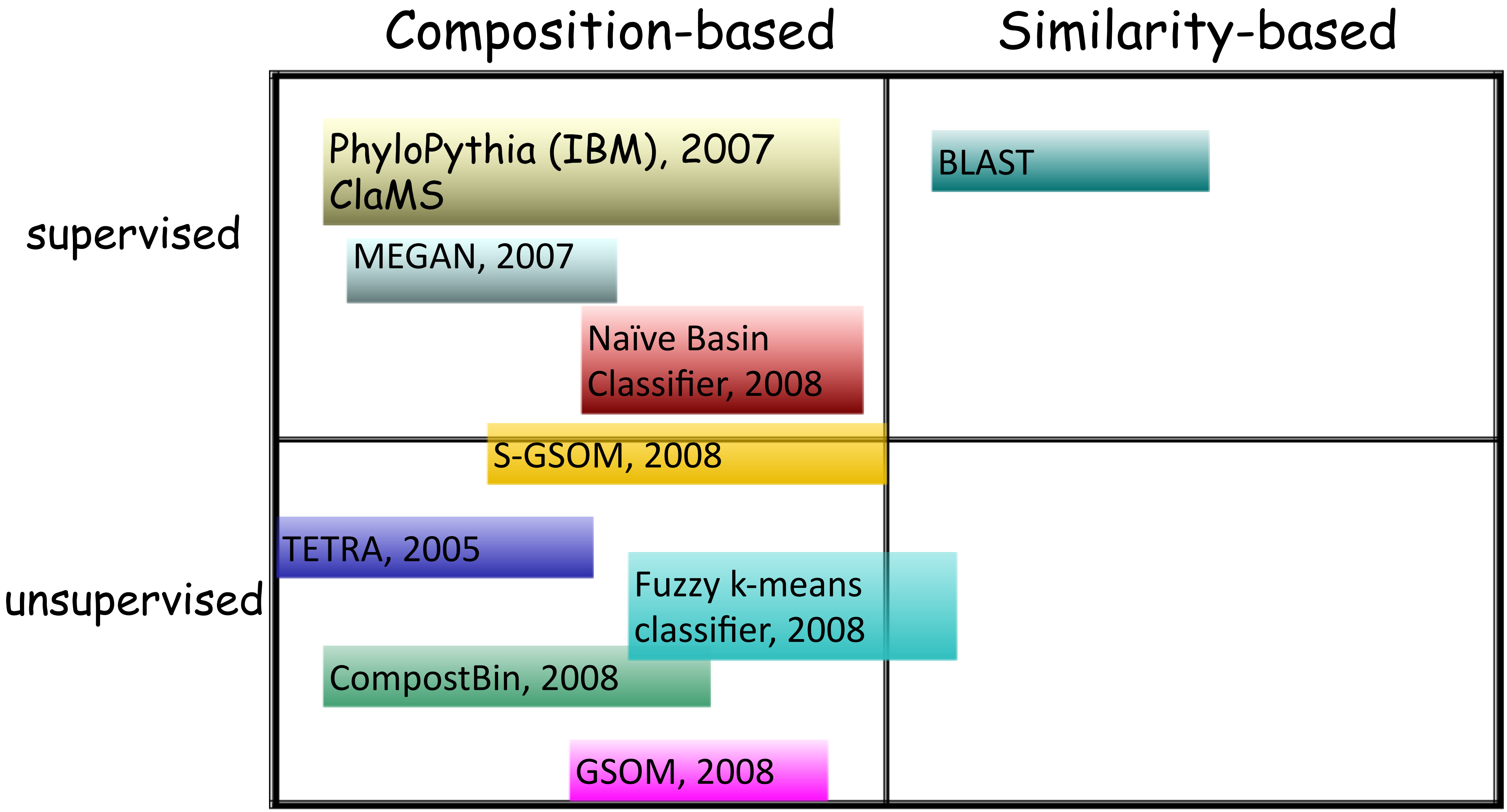
Biology perspective

- 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. olionucleotide 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

unsupervised

PhyloPythia (IBM), 2007
ClaMS

MEGAN, 2007

Naïve Basin Classifier, 2008

S-GSOM, 2008

TETRA, 2005

Fuzzy k-means classifier, 2008

CompostBin, 2008

GSOM, 2008

BLAST

KL Liu

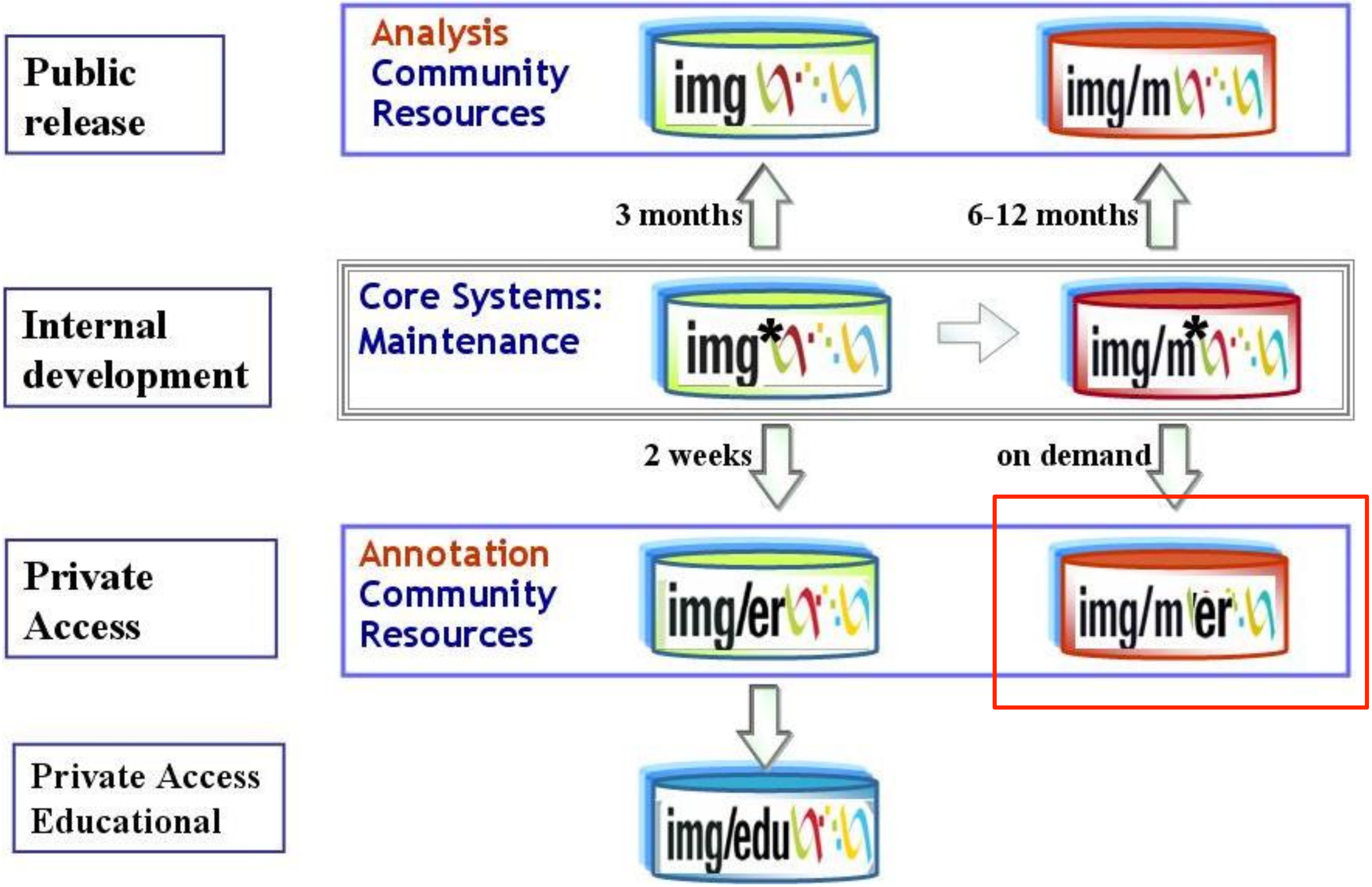
Metagenome Annotation

IMG/M-ER – the combined approach of BLASTX similarity search and de novo gene prediction.

- 80-299 bp:
  - MultiBLASTx against IMG-NR
  - All frameshift fragments are joined afterwards.
- 300-699 bp:
  - MultiBLASTx > GeneMark> Metagene.
- >700 bp:
  - GeneMark>Metagene

MG-RAST –read based only
Top blastx hit to SEED (reference db) only

The Integrated Microbial Genomes (IMG) system
http://img.jgi.doe.gov



Public release

Analysis Community Resources

img

img/m

3 months

6-12 months

Internal development

Core Systems: Maintenance

img*

img/m*

2 weeks

on demand

Private Access

Annotation Community Resources

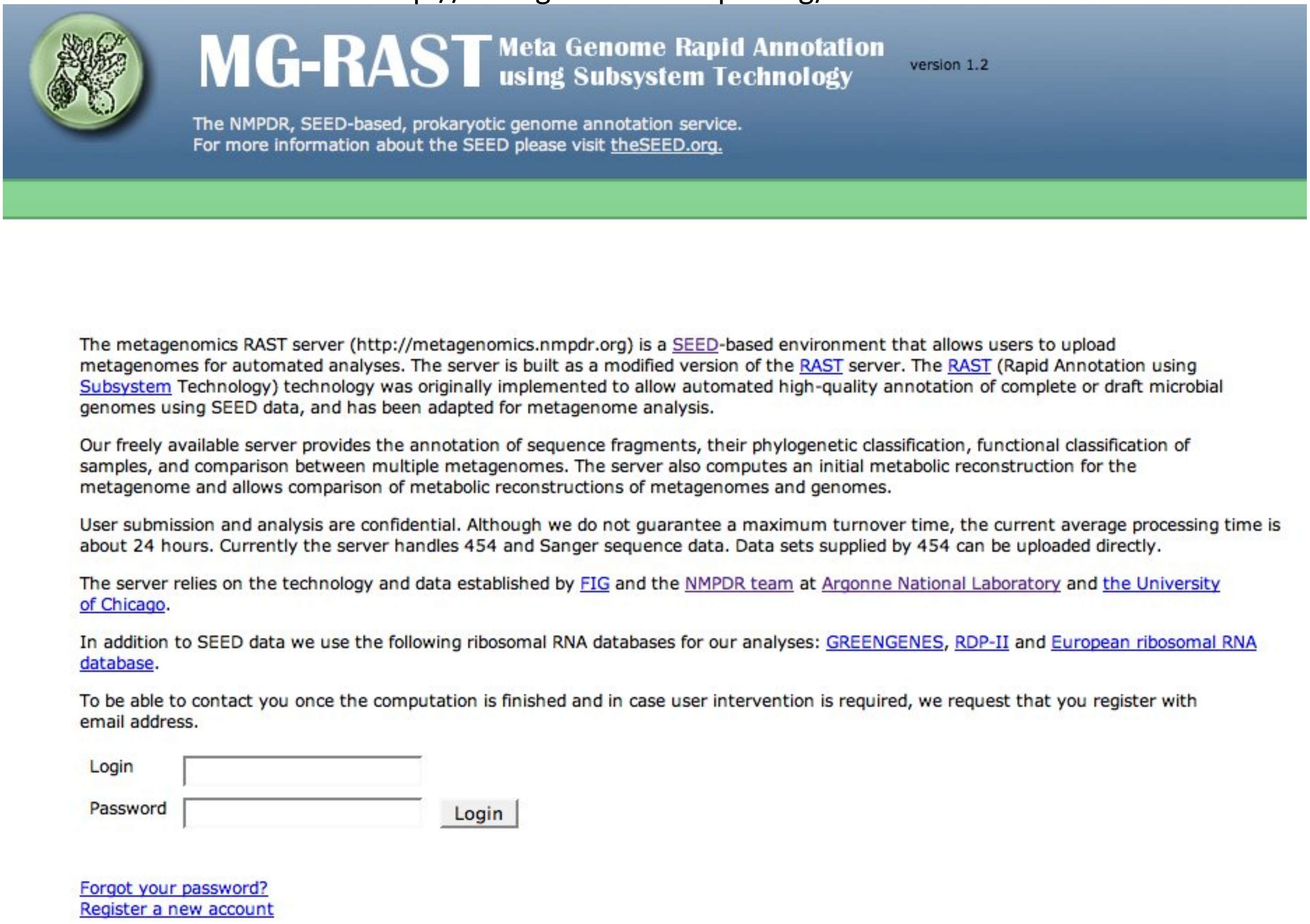
img/er

img/mer

Private Access Educational

img/edu

http://metagenomics.nmpdr.org/



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.

Credits

- NIH-NIDCR
- LANL-ORALGEN
  - Chienchi Lo
  - Pavel Senin
  - Andrew Liu
  - Patrick Chain