Introduction to MetaGenomics

Olivier Croce *- croce@unice.fr - bioinfomed.fr*

IRCAN

INSTITUT DE RECHERCHE SUR LE CANCER ET LE VIEILLISSEMENT, NICE

INSTITUTE FOR RESEARCH ON CANCER AND AGING, NICE

Inserm · cnrs · UNIVERSITÉ CÔTE D'AZUR · chu de Nice · Antoine Lacassagne

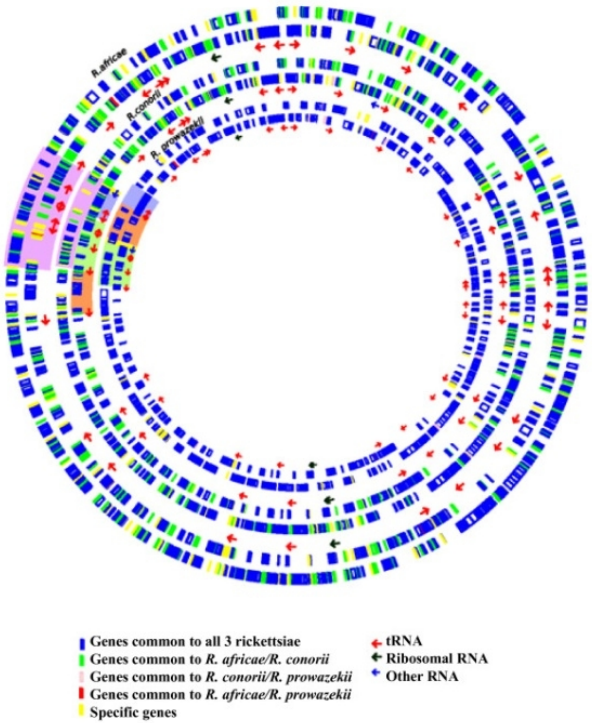
IRCAN BIOINFORMATICS & INFORMATICS

Definitions

Genomics:

=> **Analysis** of the genome from a **single organism**, phenotype prediction

- Presence/absence of genes
- Chromosomal gene order (synteny)
- Comparison of presence/absence of orthologous genes
- Presence of indels or SNPs in conserved genes
- Repetitive motifs
- …



Applications for procaryotes => optimization or design of culture media, resistance to antibiotics, detection of virulence

MetaGenomics:

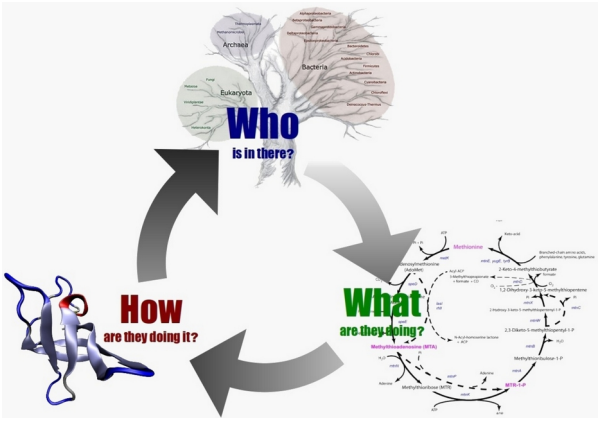
=> **Analysis** of the genetic material recovered from an environmental sample. The sample contains **>1 species**

Metagenome: the collective genome of all the microorganisms in a given environment.

Metagenomics

Characterization of the diversity :

- **Who** is there ?
- **What** are they doing ?
- **How** are they doing ?



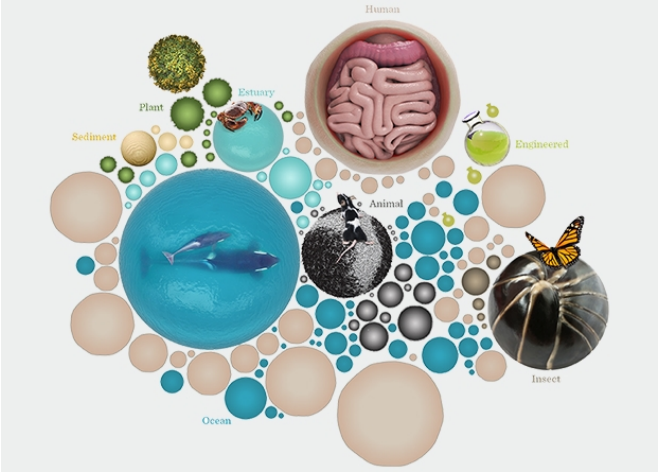
Applications :

Fundamental research, industrial, clinical applications

(growth, proteins, vaccine, secondary metabolites, resistance, etc.)

Localisation:

Almost everywhere ! microbiome, waters, soils and deep ground, space, and more



Microbiome

= microbiota = microbiote

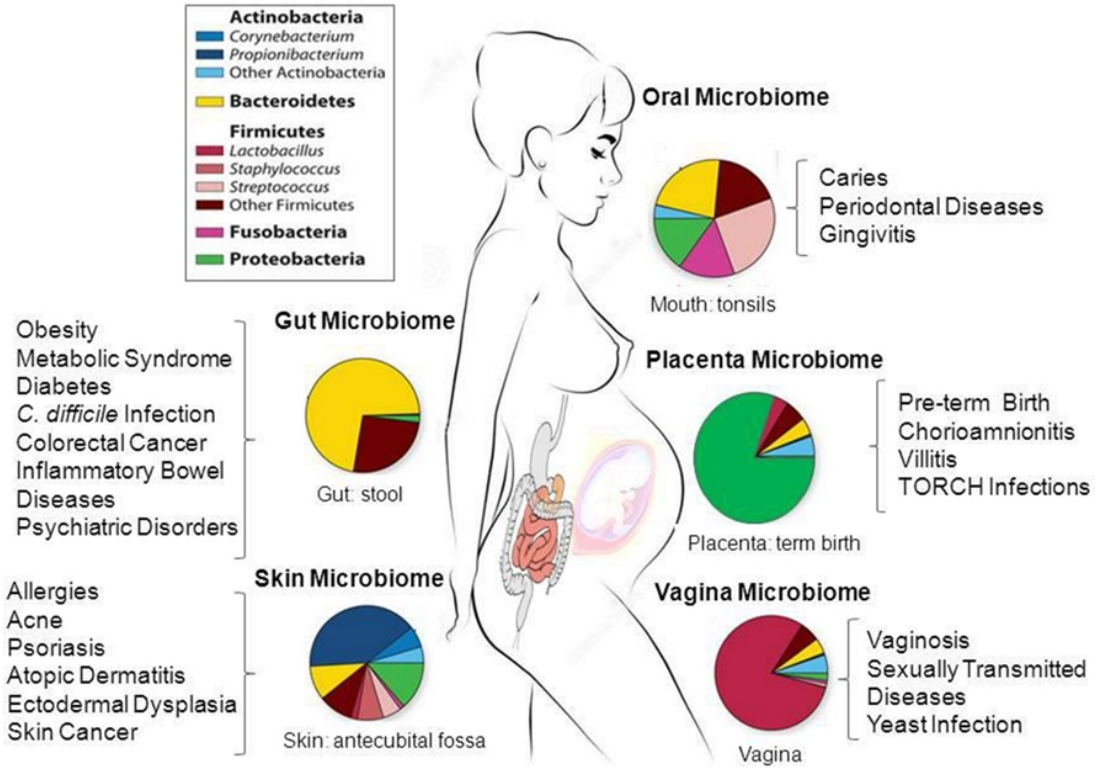
- All the micro-organisms (bacteria, archaea, fungi, viruses), colonizing and interact with the host in various organs
- The microbiota is involved in many essential functions

- Altered in many pathologies :
  - Metabolic diseases (Gurung *et al.* 2020, Liu *et al.* 2021)
  - Auto-immune diseases (De Luca *et al.* 2019)
  - Liver diseases (Qin *et al*. 2014, Loomba *et al.* 2017)
  - Chronic intestinal diseases (Nishida *et al.* 2017)
  - Cancer (Garrett *et al.* 2015)

- Potential applications:
  - Non-invasive diagnostic or prognostic marker
  - Interventional levers (diet, probiotics, fecal transplantation, etc.)

Microbiome

- More cell in guts than in the rest of the body ! (>10^{14})
- Be considered as a real organ
- The most studied microbiome : gut microbiome



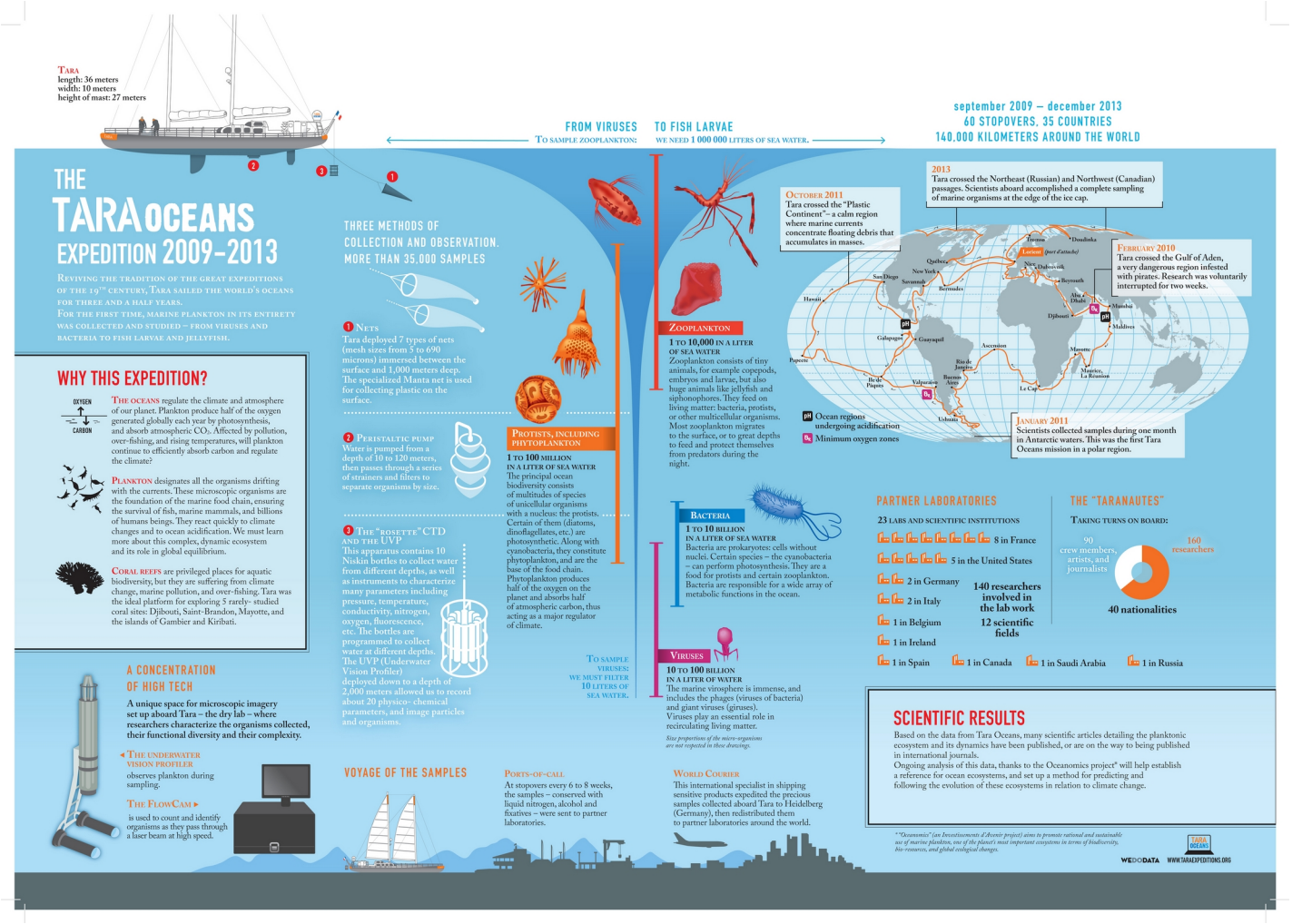
Colonization of the intestine:

- Initial acquisition during delivery by vagina, stool and skin
- Breastfeeding
- Breast milk: Bifidobacterium (more than 90% of the flora)
- Industrial milk: more Bacteroides and Clostridium
- Stool begins to become hard = adult flora

Waters

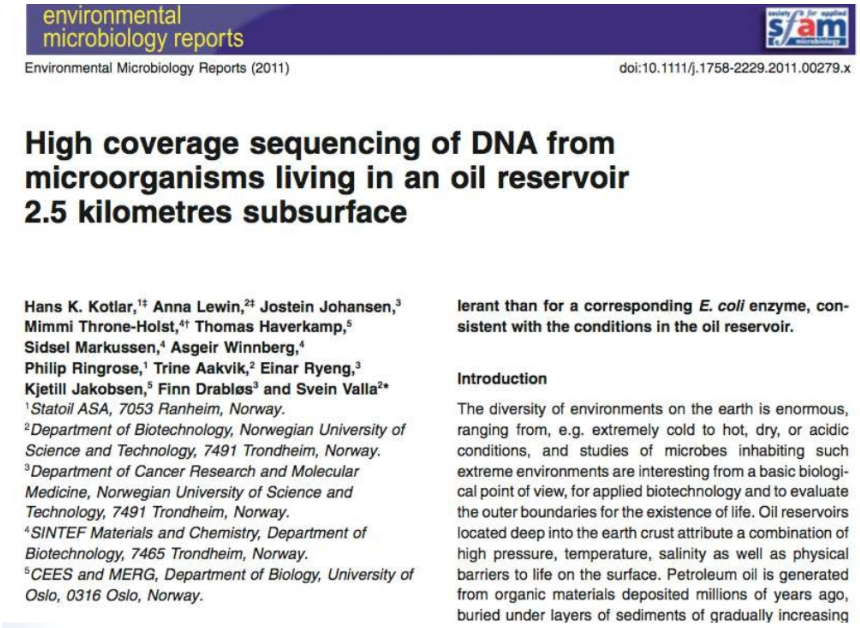
- Oceans, huge microbial diversity

Tara Oceans project. 117 millions of oceanic genes founded !



Soils and deep ground

- Microbes are also present at great depths



environmental microbiology reports

Environmental Microbiology Reports (2011) doi:10.1111/j.1758-2229.2011.00279.x

High coverage sequencing of DNA from microorganisms living in an oil reservoir 2.5 kilometres subsurface

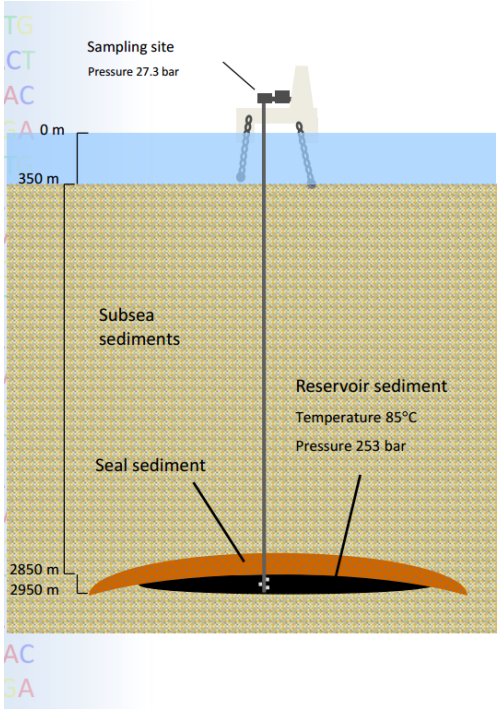
Hans K. Kotlar,[1‡] Anna Lewin,[2‡] Jostein Johansen,[3] Mimmi Throne-Holst,[4†] Thomas Haverkamp,[5] Sidsel Markussen,[4] Asgeir Winnberg,[4] Philip Ringrose,[1] Trine Aakvik,[2] Einar Ryeng,[3] Kjetill Jakobsen,[5] Finn Drabløs[3] and Svein Valla[2*]

[1]*Statoil ASA, 7053 Ranheim, Norway.*
[2]*Department of Biotechnology, Norwegian University of Science and Technology, 7491 Trondheim, Norway.*
[3]*Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, 7491 Trondheim, Norway.*
[4]*SINTEF Materials and Chemistry, Department of Biotechnology, 7465 Trondheim, Norway.*
[5]*CEES and MERG, Department of Biology, University of Oslo, 0316 Oslo, Norway.*

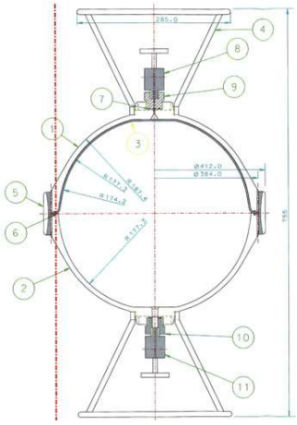
lerant than for a corresponding *E. coli* enzyme, consistent with the conditions in the oil reservoir.

Introduction

The diversity of environments on the earth is enormous, ranging from, e.g. extremely cold to hot, dry, or acidic conditions, and studies of microbes inhabiting such extreme environments are interesting from a basic biological point of view, for applied biotechnology and to evaluate the outer boundaries for the existence of life. Oil reservoirs located deep into the earth crust attribute a combination of high pressure, temperature, salinity as well as physical barriers to life on the surface. Petroleum oil is generated from organic materials deposited millions of years ago, buried under layers of sediments of gradually increasing



Sampling an oil field



Xpand Pressure Flask

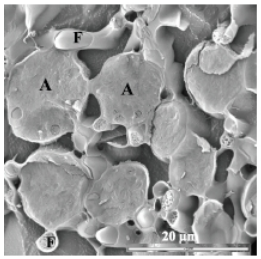
Space

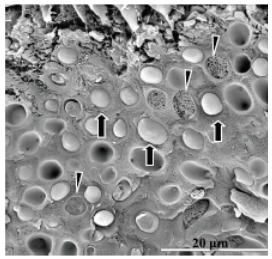
- Microbes can survive in extreme conditions such as outer space

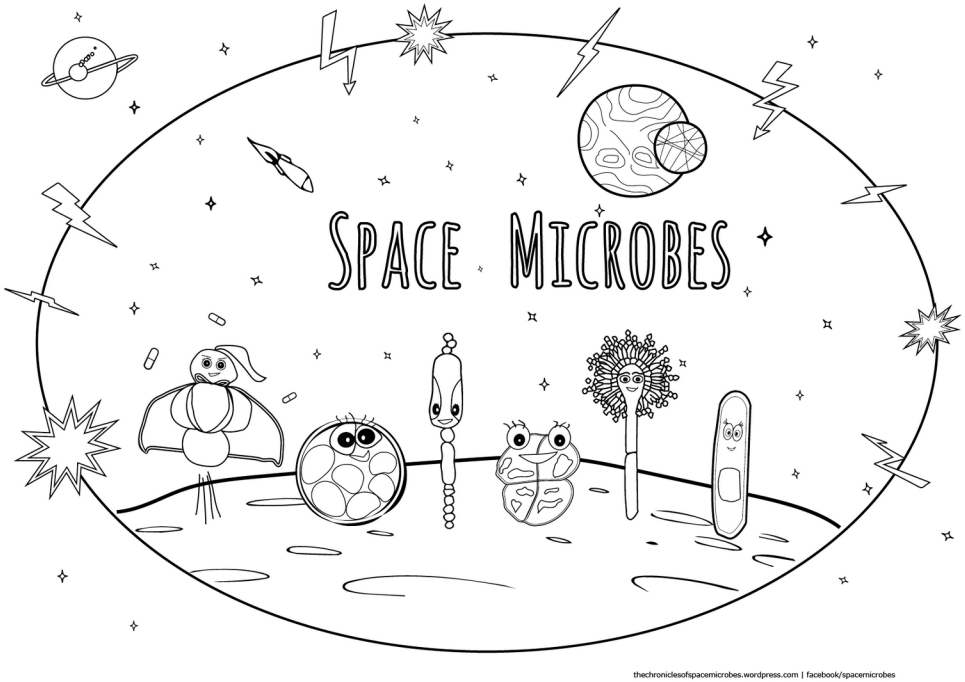
Survival of *B.subtilis* spores :
Unprotected : several seconds
Protected : more than 6 years

Others microorganisms :
Phage T1, *Synechococcus*
Haloarcula, Deinococcus

Recently : surprising **survival of lichen**

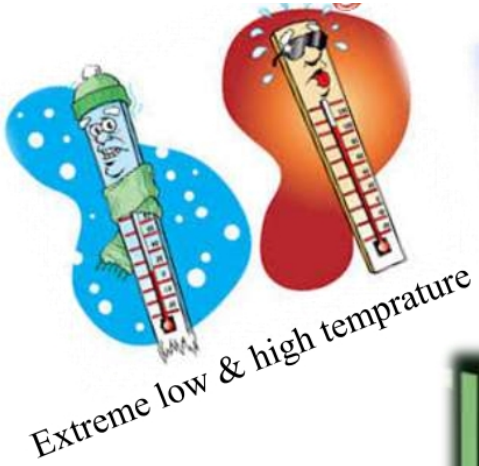






And more !

- Microbes are present in any other metagenomics hot spots
- Almost everywhere with extremophiles



Havey metal composition

Methods

- **Microarrays**

Requires knowledge of the community in advance :PhyloChip (taxonomic), Geochip (metabolic)

- **(Meta) *Barcoding sequencing***

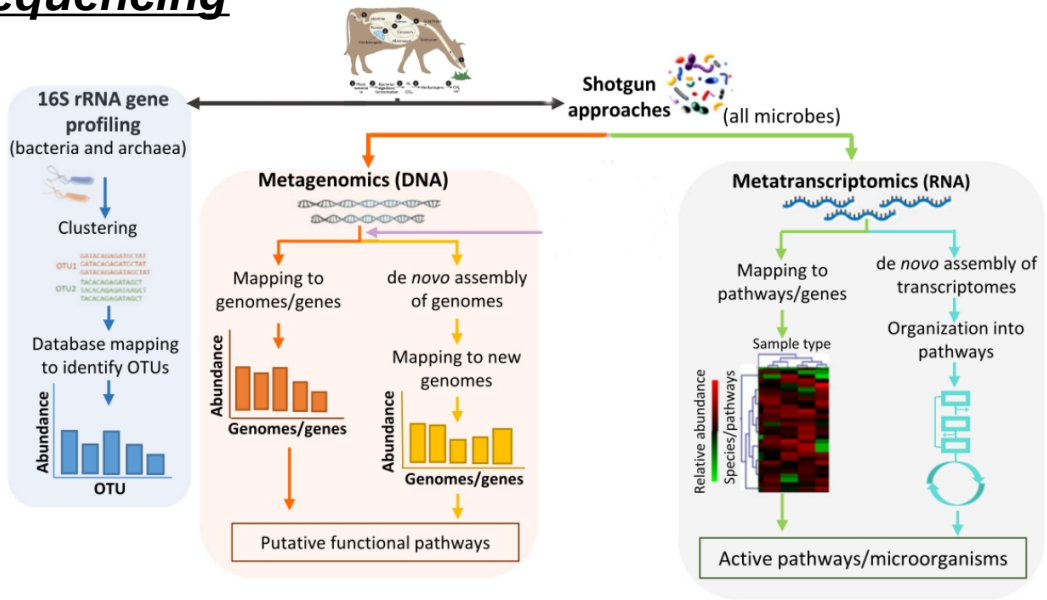
Amplicon based analysis through High throughput sequencing of a given gene (or part of) after amplification

- *SSU rRNA* (=16S/18S)
- Protein coding genes: rpoB, nifH, IRS, cytC, RecA,…
- ITS (internal transcribed spacer)

- ***(WGS / Total / Shotgun) MetaGenomics sequencing***

Sequencing of the whole DNA in a sample.

Complete community analysis, characterization of the pangenome



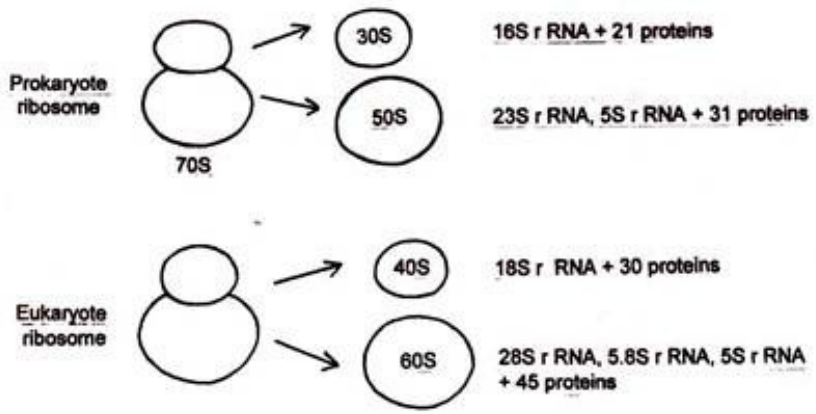
*** *MetaTranscriptomics***

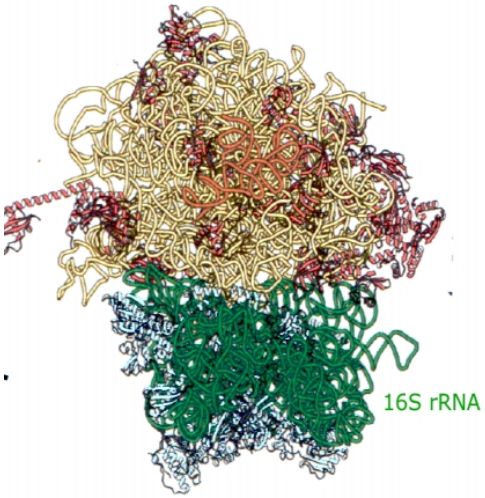
Idem TotalMetagenomics, but from RNA

(Meta) Barcoding sequencing

- Usually based on rRNA **16S for bacteria** (~1540nt), rRNA **18S for eucaryotes**

  => current taxonomic classification for prokaryotes & eukaryotes. Species definition !



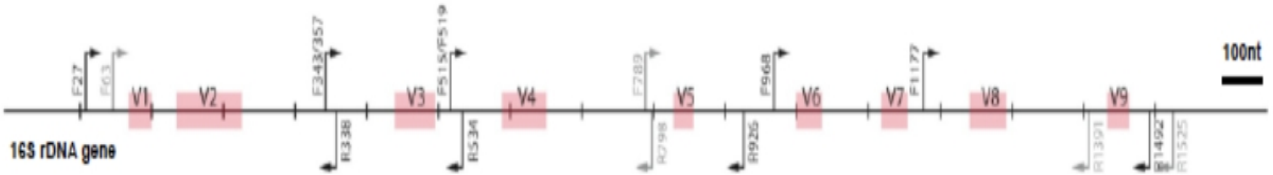


- Most commonly used molecular marker: essential function, ubiquity, evolutionary properties
- Highly conserved gene, 9 hypervariable domains interspersed with conserved fragments
- Rapid and cost-effective approaches for assessing diversity and abundance.
- OTU (operational taxonomic units) definition based on 16S rRNA gene

  => organisms displaying 97 to 98% identity in this gene to be part of the same OTU

(Meta) Barcoding sequencing

- Design primers for PCR (polymerase chain reaction) for specific regions of the 16s rRNA, not the whole molecule



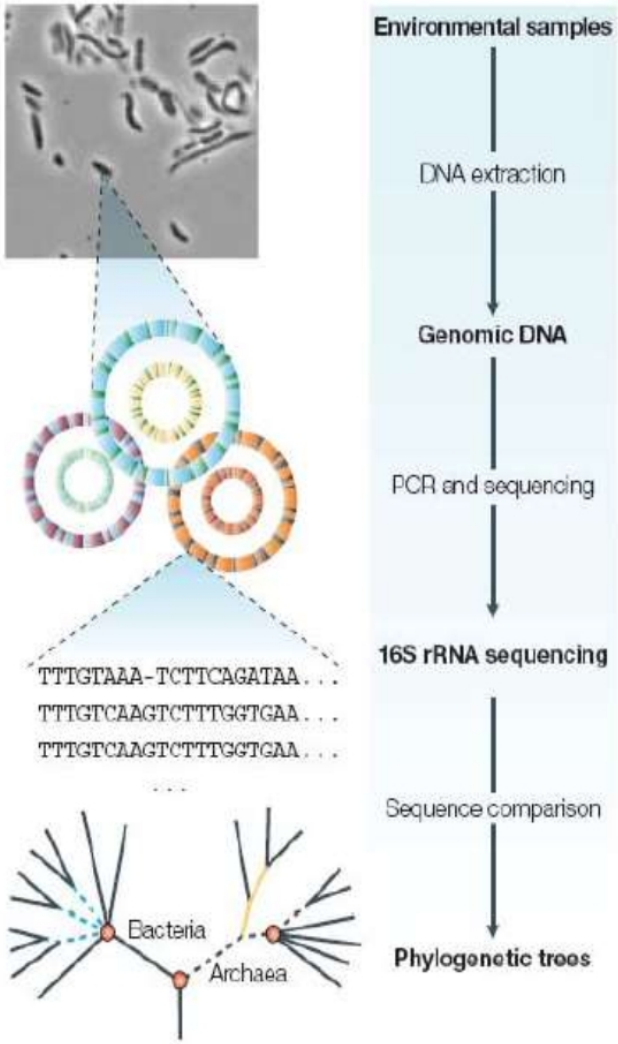
- Use of universal and specific primers that hybridize to highly conserved regions

- Usually, amplification of the V4 and V6 region of the rRNA 16S

- Sequencing amplicons (generally using Illumina paired-end)

- Comparison with other closed sequences (public databases) Same species => homology >97.8% , genera >95%

- Phylogenetic analysis of 16S rRNA helps to reveal the species diversity in a community



(Meta) Barcoding sequencing

Limits of rRNA use

- Sampling challenges : rich species Vs sparse species. Rare species could not be sequenced
- Do not tell much about the functional abilities of a community
- Based on the assumption that the level of interspecies rRNA variation is homogeneous among genera
- PCR bias: not all rRNA genes amplify equally well with the same “universal” primers
- Multiple copies of rRNA genes in some species (which may artificially lead to the over-representation of some species)
- Speed of evolution of rRNA genes may vary according to the phylum
- Homology cutoff not applicable to all genera => may underestimate, or overestimate the diversity
- The discriminatory power may be insufficient at the species level, especially for closely related species

Examples:
- Pantoea agglomerans strains may exhibit up to 27 bp differences, which does not validate the 98.7% cutoff
- Clostridium tetani and C. innocuum exhibit 104 bp differences => < 95% homology => classification in distinct genera?

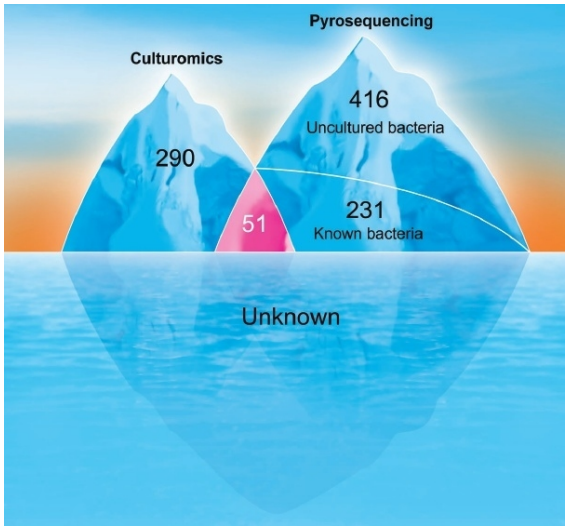
(Meta) Barcoding sequencing

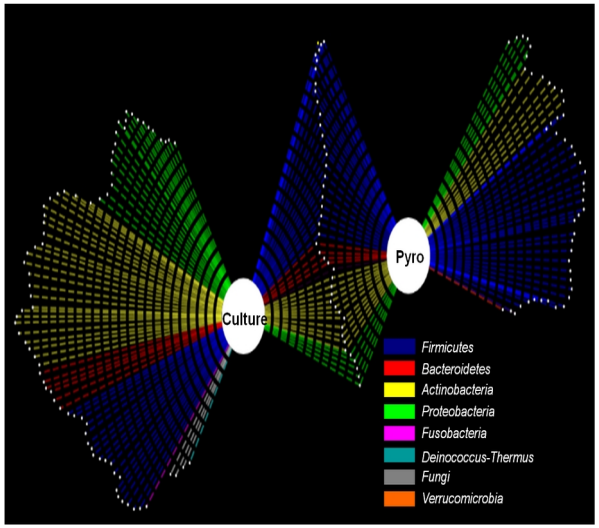
Limits of rRNA use

- Possibility of 16S rRNA genes acquired by HGT

  [Jain et al. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA 1999;96:3801-6]

- Missing sequences or variable quality of available sequences, including from validly published species

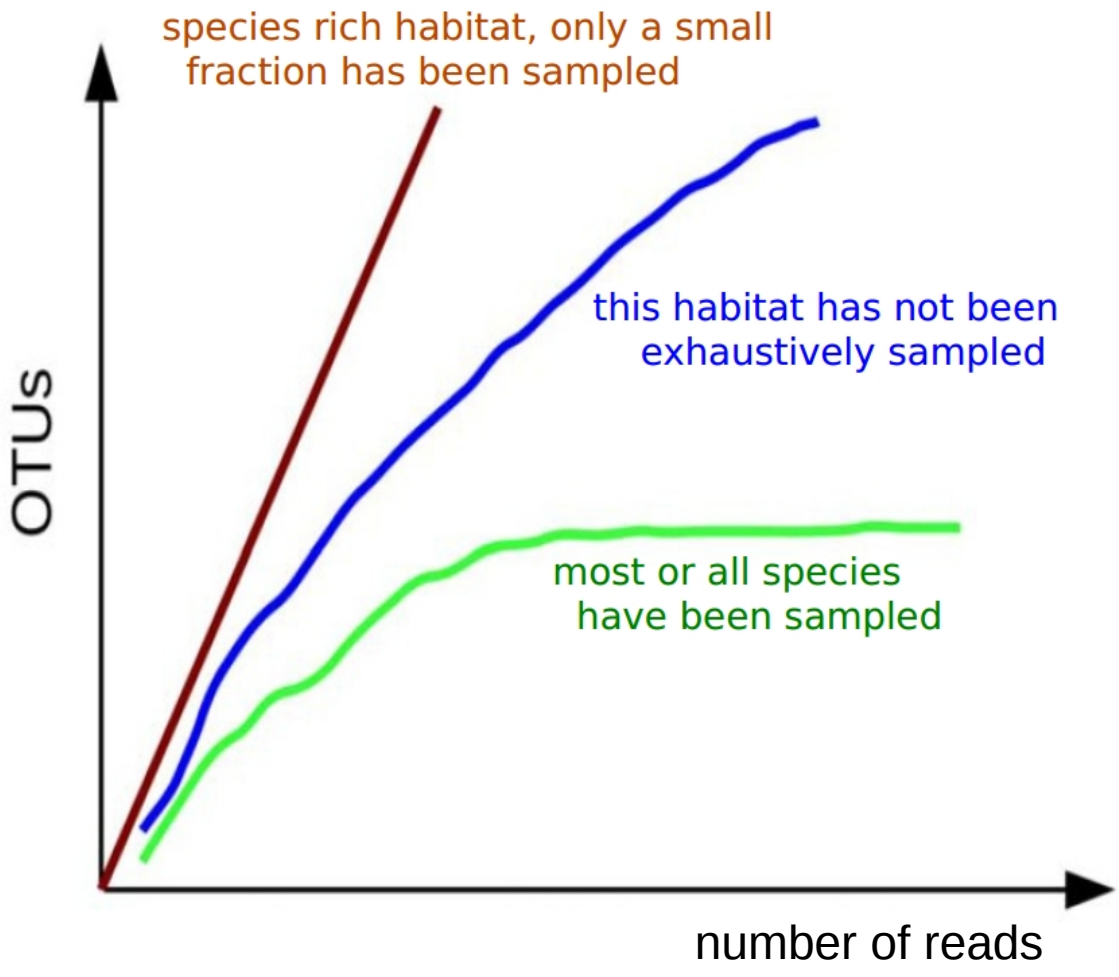




- **For procaryotes** : NGS could be complementary to culturomics. 90-95% microorganisms remain uncultivable in laboratory.

- **For eucaryotes** : cultures is not even possible (plankton, fungi..), direct observation (microscope) is exhausting => NGS is widely uses (rRNA 18S, or other genes)

(Meta) Barcoding sequencing



Rarefaction allows the calculation of species richness for a given number of individual samples, based on the construction of so-called rarefaction curves. This curve is a plot of the number of species as a function of the number of samples.

(Total) MetaGenomics sequencing

- Environmental sample : all DNA sequencing, no specific amplification
  - Who is here ? => Biodiversity caracterization
  - Who does what ? => Physiological caracterization

RESEARCH ARTICLE

Environmental Genome Shotgun Sequencing of the Sargasso Sea

J. Craig Venter,[1*] Karin Remington,[1] John F. Heidelberg,[3] Aaron L. Halpern,[2] Doug Rusch,[2] Jonathan A. Eisen,[3] Dongying Wu,[3] Ian Paulsen,[3] Karen E. Nelson,[3] William Nelson,[3] Derrick E. Fouts,[3] Samuel Levy,[2] Anthony H. Knap,[6] Michael W. Lomas,[6] Ken Nealson,[5] Owen White,[3] Jeremy Peterson,[3] Jeff Hoffman,[1] Rachel Parsons,[6] Holly Baden-Tillson,[1] Cynthia Pfannkoch,[1] Yu-Hui Rogers,[4] Hamilton O. Smith[1]

We have applied "whole-genome shotgun sequencing" to microbial populations collected en masse on tangential flow and impact filters from seawater samples collected from the Sargasso Sea near Bermuda. A total of 1.045 billion base pairs of nonredundant sequence was generated, annotated, and analyzed to elucidate the gene content, diversity, and relative abundance of the organisms within these environmental samples. These data are estimated to derive from at least 1800 genomic species based on sequence relatedness, including 148 previously unknown bacterial phylotypes. We have identified over 1.2 million previously unknown genes represented in these samples, including more than 782 new rhodopsin-like photoreceptors. Variation in species present and stoichiometry suggests substantial oceanic microbial diversity.

1.2 million unknown genes
(Venter et al., 2004)

INTRODUCTION TO SPECIAL ISSUE

Tara Oceans studies plankton at planetary scale

P. Bork[1], C. Bowler[2], C. de Vargas[3,4], G. Gorsky[5,6], E. Karsenti[2,7], P. Wincker[8]
+ See all authors and affiliations

Science 22 May 2015:
Vol. 348, Issue 6237, pp. 873
DOI: 10.1126/science.aac5605

Article | Figures & Data | Info & Metrics | eLetters | PDF

The ocean is the largest ecosystem on Earth, and yet we know very little about it. This is particularly true for the plankton that inhabit the ocean. Although these organisms are at least as important for the Earth system as the rainforests and form the base of marine food webs, most plankton are invisible to the naked eye and thus are largely uncharacterized. To study this invisible world, the multinational *Tara* Oceans consortium, with use of the 110-foot research schooner *Tara*, sampled microscopic plankton at 210 sites and depths up to 2000 m in all the major oceanic regions during expeditions from 2009 through 2013 (*1*).

Success depended on collaboration between scientists and the *Tara* Expeditions logistics team. The journey involved not only science but also outreach and education as well as negotiation through the shoals of legal and political regulations, funding uncertainties, threats from pirates, and unpredictable weather (*2*). At various times, journalists, artists, and teachers were also on board. Visitors included Ban Ki-moon (Secretary-General of the United Nations) and numerous youngsters, including

Tara Oceans : 117 millions of oceanic genes
(Bork and al., 2015)

(Total) MetaGenomics sequencing

Principe of metagenomics

Sample collection

↓

Whole DNA extraction

↓

Whole DNA amplification

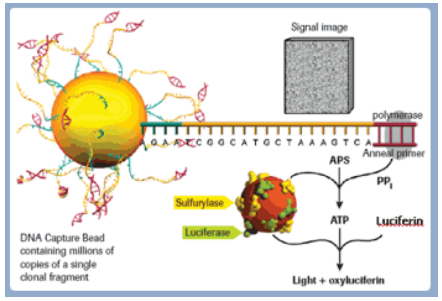
↓

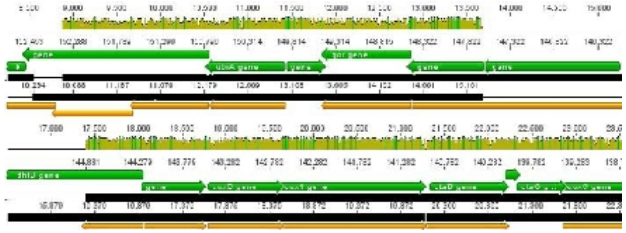
Whole DNA sequencing

↓

Data analysis - Bioinformatics

(alignement, comparaison of sequences,gene prediction, metabolic pathway reconstruction)

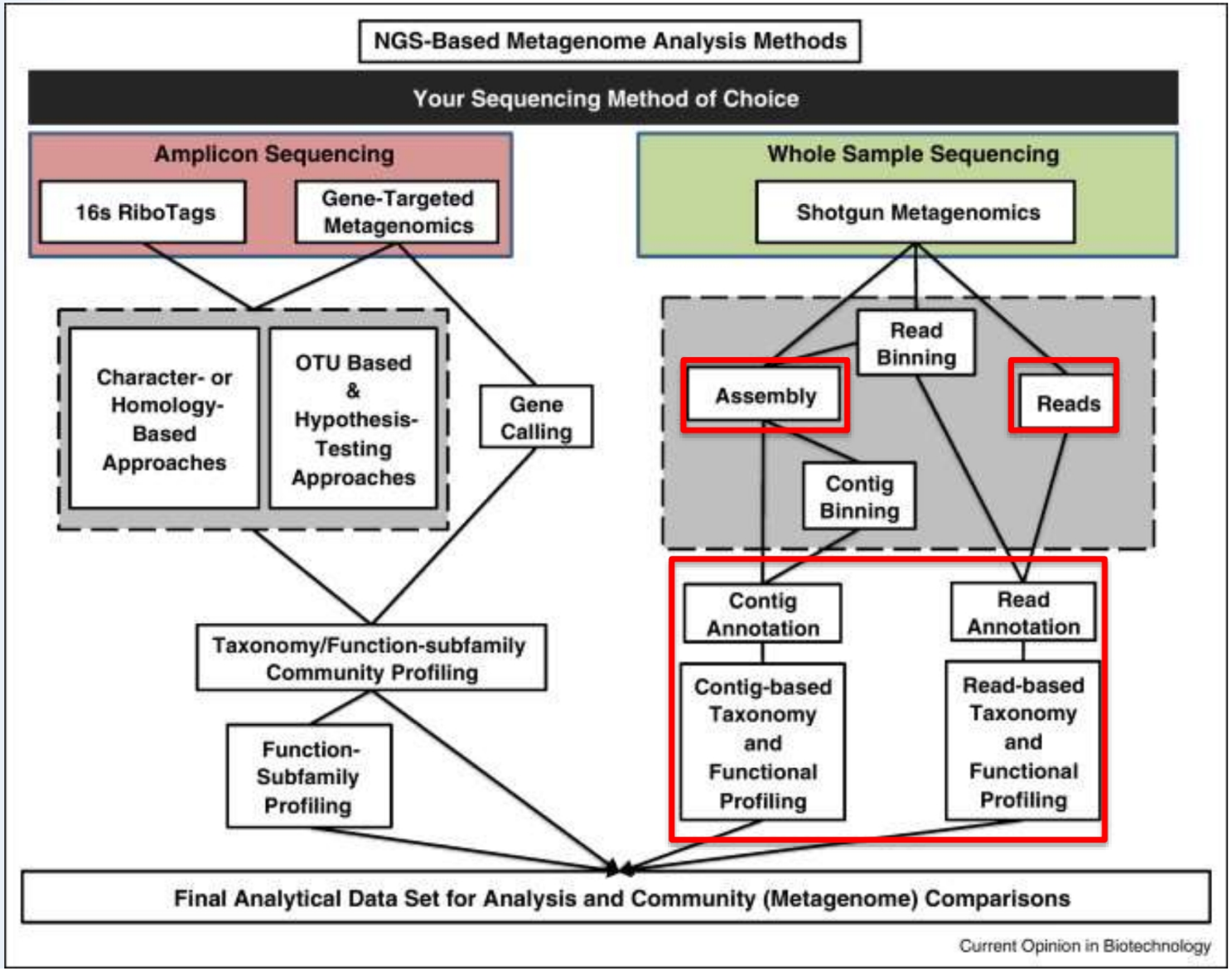




- Information about : biodiversity, but also physiology, metabolic pathways…

(Total) MetaGenomics sequencing

- Methods differences between barcoding and shotgun sequencing



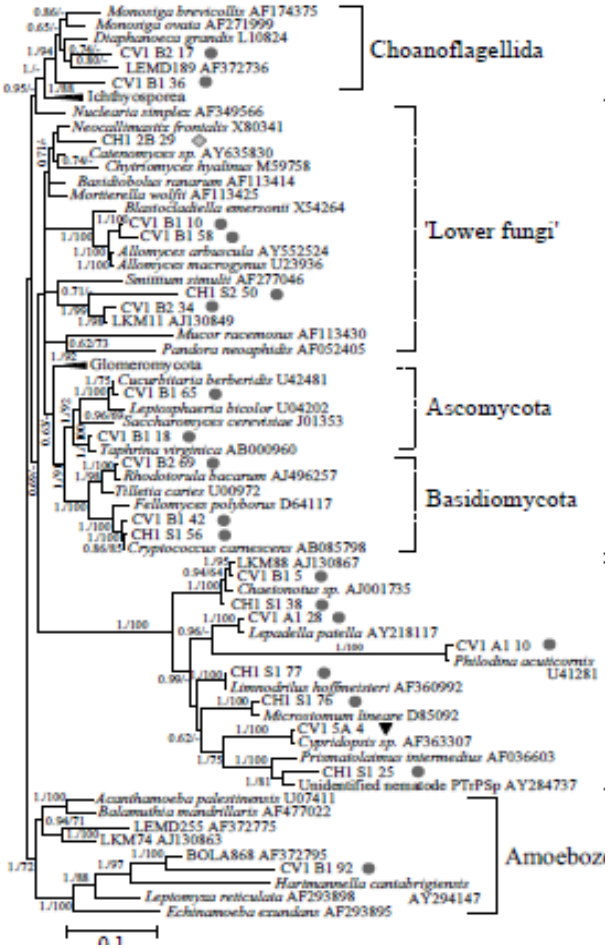
(Total) MetaGenomics sequencing

(1) Biodiversity analysis

- Building of **phylogenetic trees** :

  Positioning of known species and unknown microorganisms (to who are they closest to ?)

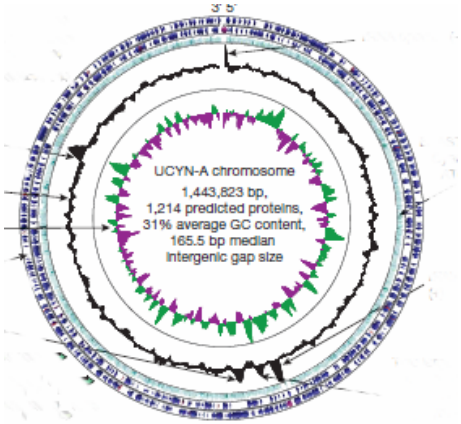
  Discovery of : new species, new phyla, etc



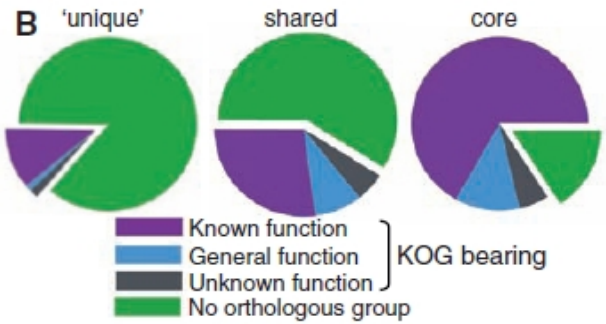
(Total) MetaGenomics sequencing

(2) Physiology and genomic

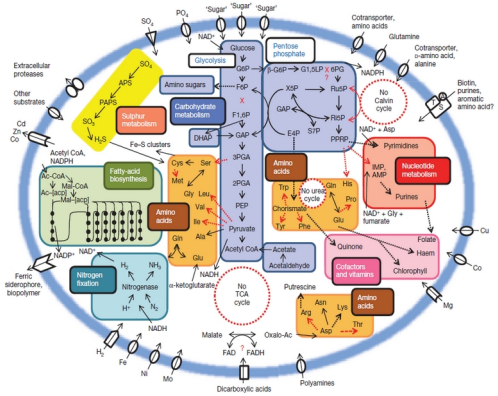
Genomes



Genes
functional groups



Pathways
metabolism, physiology



Informations as :

- G+C contents
- Genome sizes
- DNA repair mechanisms
- Pathways of excretion, polysaccharides secretion
- …

(Total) MetaGenomics sequencing

- **Initial bioinformatics step** :

- Direct use of reads (after trimming and quality tests)

- Assemble reads into contigs

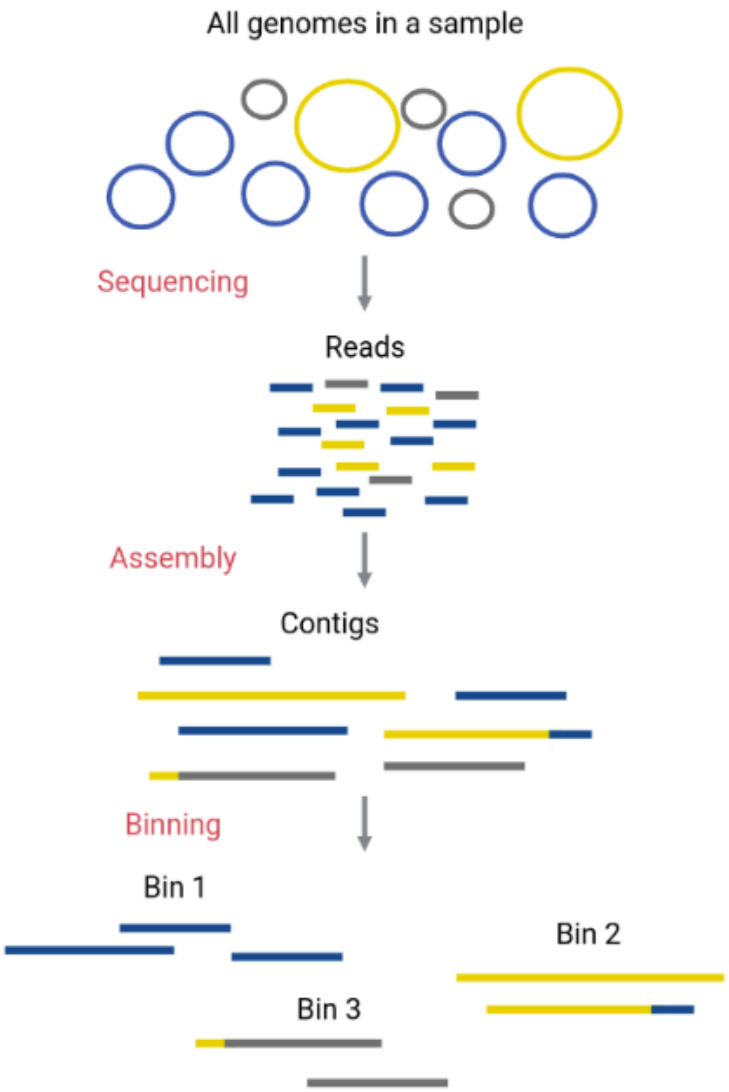
- Both

- **Sequence classification (=clusterisation)**

More difficult than for barcoding sequencing => need to create “bins”

Sequence classification (binning) is the process of separating sequence data using specific information.

Sequence classification by: sequences composition or sequences similarity

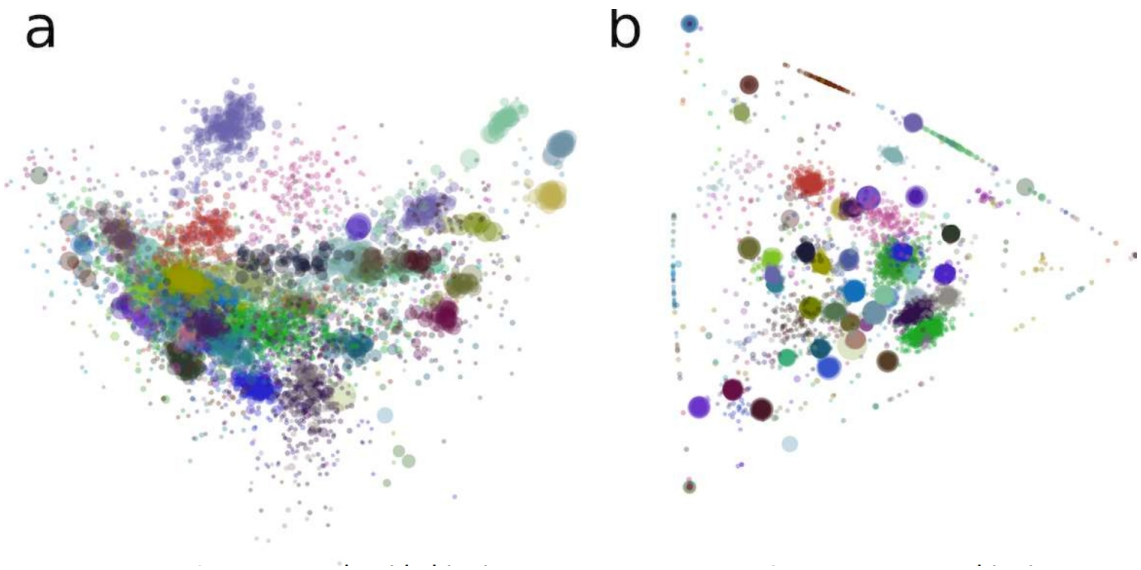


(Total) MetaGenomics sequencing

(1) Sequences composition

- Tetranucleotide frequency (kmer counting)
- Clustering of reads. (e.g. swarm, cd-hit)
- Sequence (co-) assembly (e.g. MetaHit, Metavelvet)
- Differential coverage of contigs (e.g. GroopM, Concoct)

Advantage : read with unknown origin can be classified into a bin
Disadvantage: impossible to determine taxonomy or function of the reads



Input data: 1159 (Imelfort et al., PeerJ, 2014)

a) PCA - Tetranucleotide binning
b) GroopM coverage binning

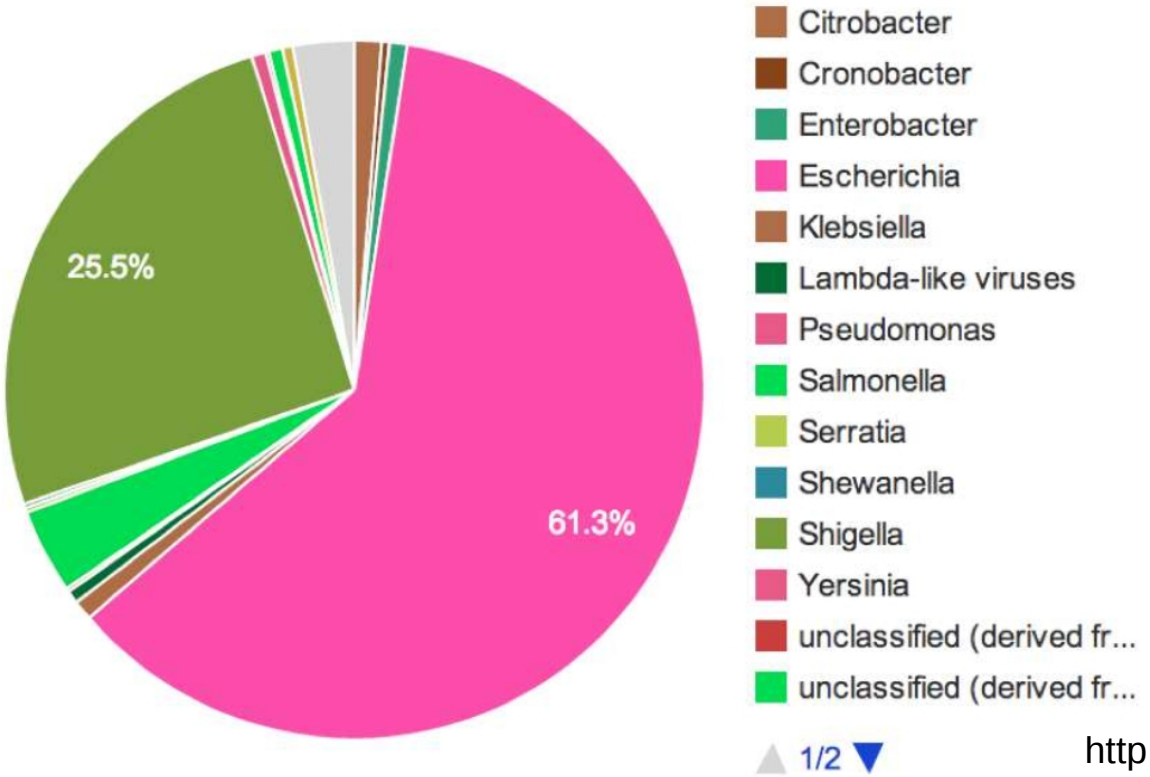
(Total) MetaGenomics sequencing

(2) Sequences similarity

- Compare sequences to reference database (e.g. Blast, bwa, bowtie)
- Use phylogenetics to classify sequences.

Advantage: One can determine taxonomy and function of reads.
Disadvantage: reads with no similarity to databases sequences, can not be classified

Usual example : Using the best blast hit



http://nickloman.github.io/

(Total) MetaGenomics sequencing - tools

Tools for sequences classification

Nucleotide composition:

CompostBin , PCA-analysis of k-mer, frequencies, Self-Organizing Maps (different variants), MetaCluster, PhyloPythia, Naïve Bayes classifier (NBC), etc

Sequence similarity:

MEGAN, SorT-Items, Threephyler, COMET,Metaphlan, PhyloSift, Kraken, etc

Both:

Phymm / PhymmBL, Phylophytia, RAIphy, Metaxa2 (rRNA), PhyloOTU (rRNA), MLTreeMap, RITA, STAMP, WGSQuikr.

Differential Coverage:

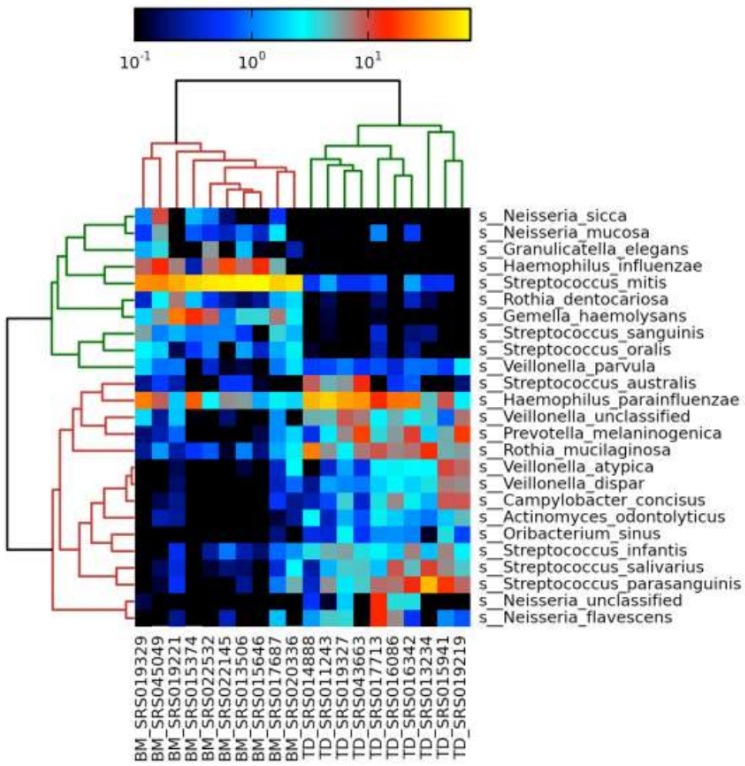
GroopM, Concoct, Blobology

(Total) MetaGenomics sequencing - tools

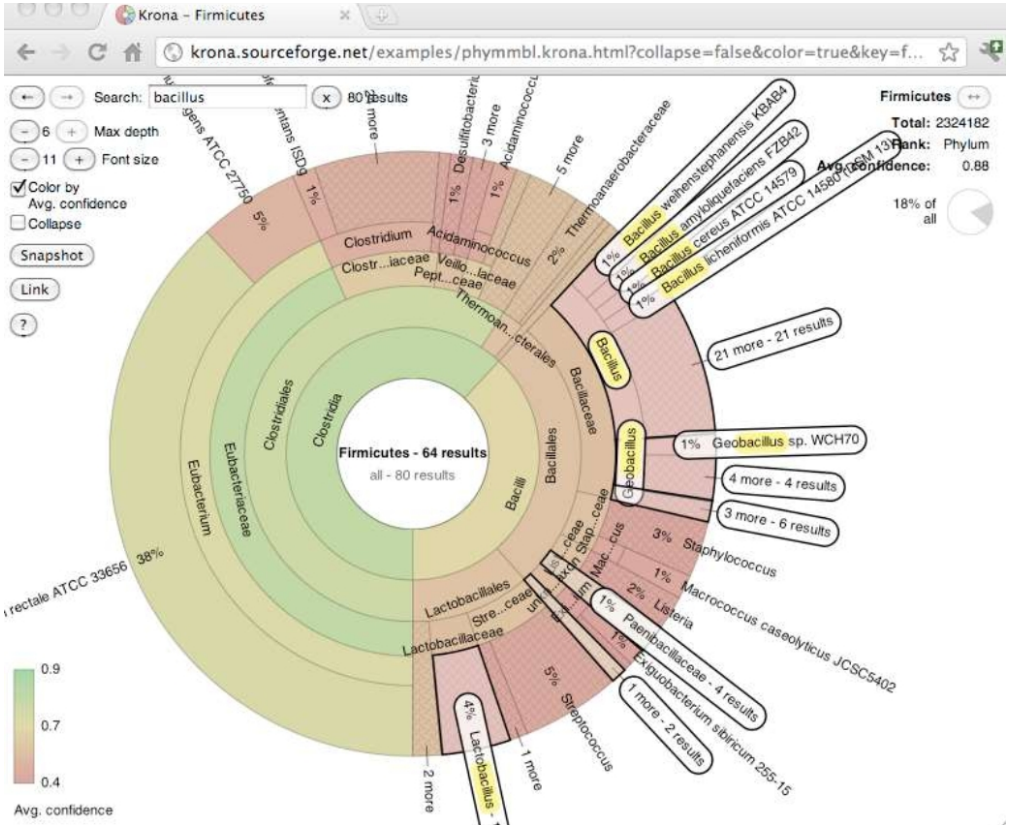
MetaPhlAn: Metagenomic Phylogenetic Analysis
* Uses a database of taxon specific marker genes
* Works well with known ecosystems: e.g. gut communities

Phylosift:
* Uses a database of 37 universal proteins & rRNA genes.
* Designed to classify using phylogenies

Both databases are smaller than NCBI NR, depending on your ecosystem, one will work better



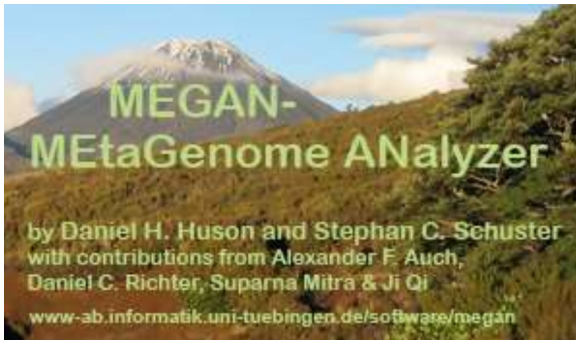
https://bitbucket.org/nsegata/metaphlan/



http://sourceforge.net/p/krona

(Total) MetaGenomics sequencing - tools

MEGAN (Huson et al., Genome Research, 2007)



- Developed for characterization of metagenomic shotgun reads
- LCA assignment based on BLAST hit scores
- Support for paired-end reads and comparison of datasets.
- Latest version can analyze RDP files / QIIME OTU files
- Analysis of metabolism via SEED, KEGG or COG maps
- Comparison of multiple metagenomes (> 2)

https://github.com/husonlab/megan-ce
http://megan.informatik.uni-tuebingen.de/

Why to use MEGAN ?

Easy to work with on a desktop / laptop computer:
Extra things needed: Java, a BLAST server

MEGAN gives a visualization of BLAST results
- Study diversity
- Compare samples
- Contamination filtering
- Special gene of interest
- Extraction of sequences based on taxonomic / metabolic information.

(Total) MetaGenomics sequencing - tools

The basics of MEGAN

MEGAN uses BLAST, a database and a taxonomy file

- BLAST N : nucleotides against a nucleotide database.
- BLAST X : Translated nucleotide against a protein database.

• Which database?

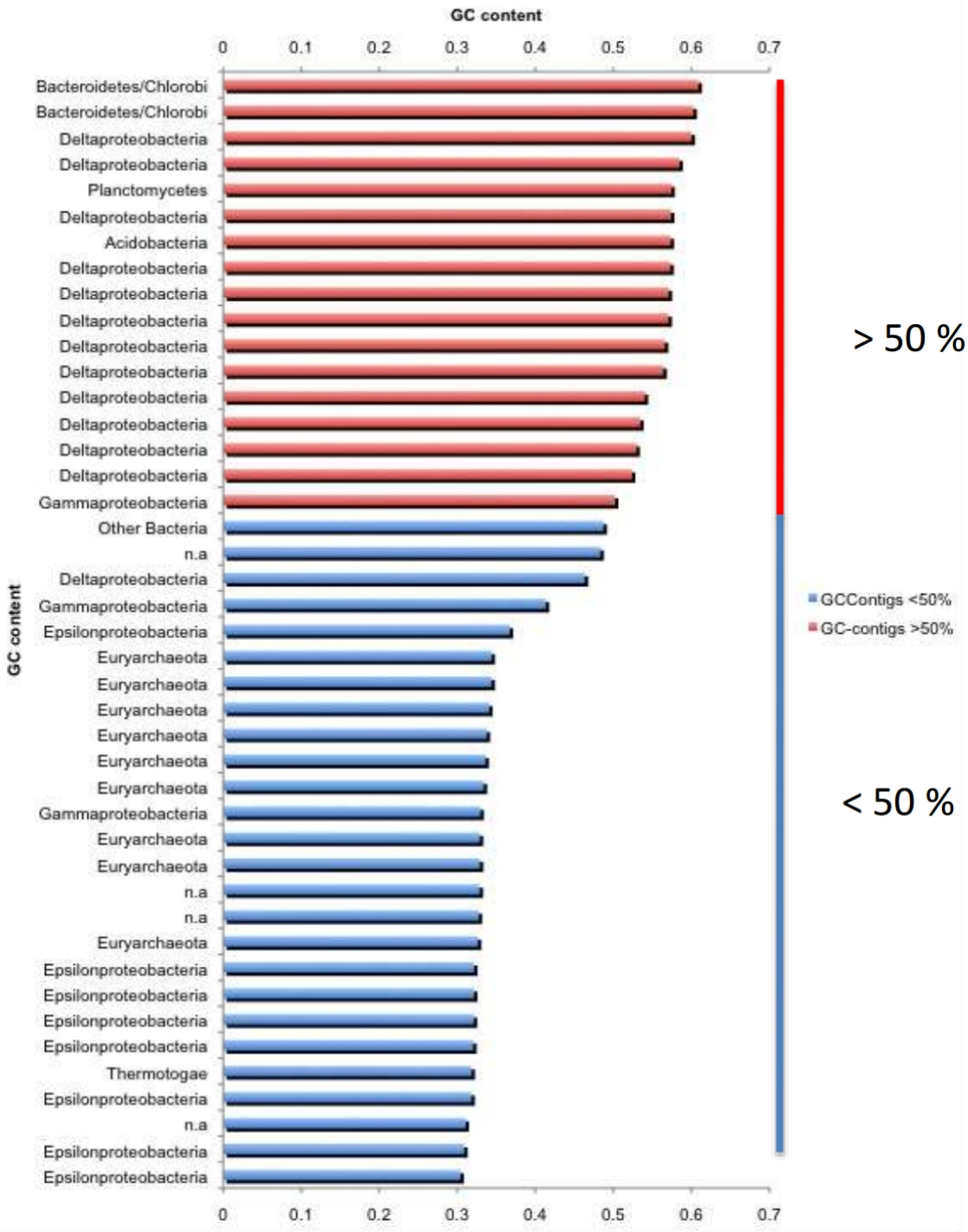
one of the many available database like the NCBI-nonredundant database (nr), or a your own custom database.

• Taxonomy: NCBI taxonomy, or your own custom taxonomy

• LCA clustering
BLAST output file is used to bin sequences using the LCA (“Lowest Common Ancestor” **) assignment algorithm into specific taxons

“ In this approach, every read is assigned to some taxon. If the read aligns very specifically only to a single taxon, then it is assigned to that taxon. The less specifically a read hits taxa, the higher up in the taxonomy it is placed. Reads that hit ubiquitously may even be assigned to the root node of the NCBI taxonomy” (MEGAN manual)

(Total) MetaGenomics sequencing - tools



MEGAN classifications

Major taxa are separated by different GC content

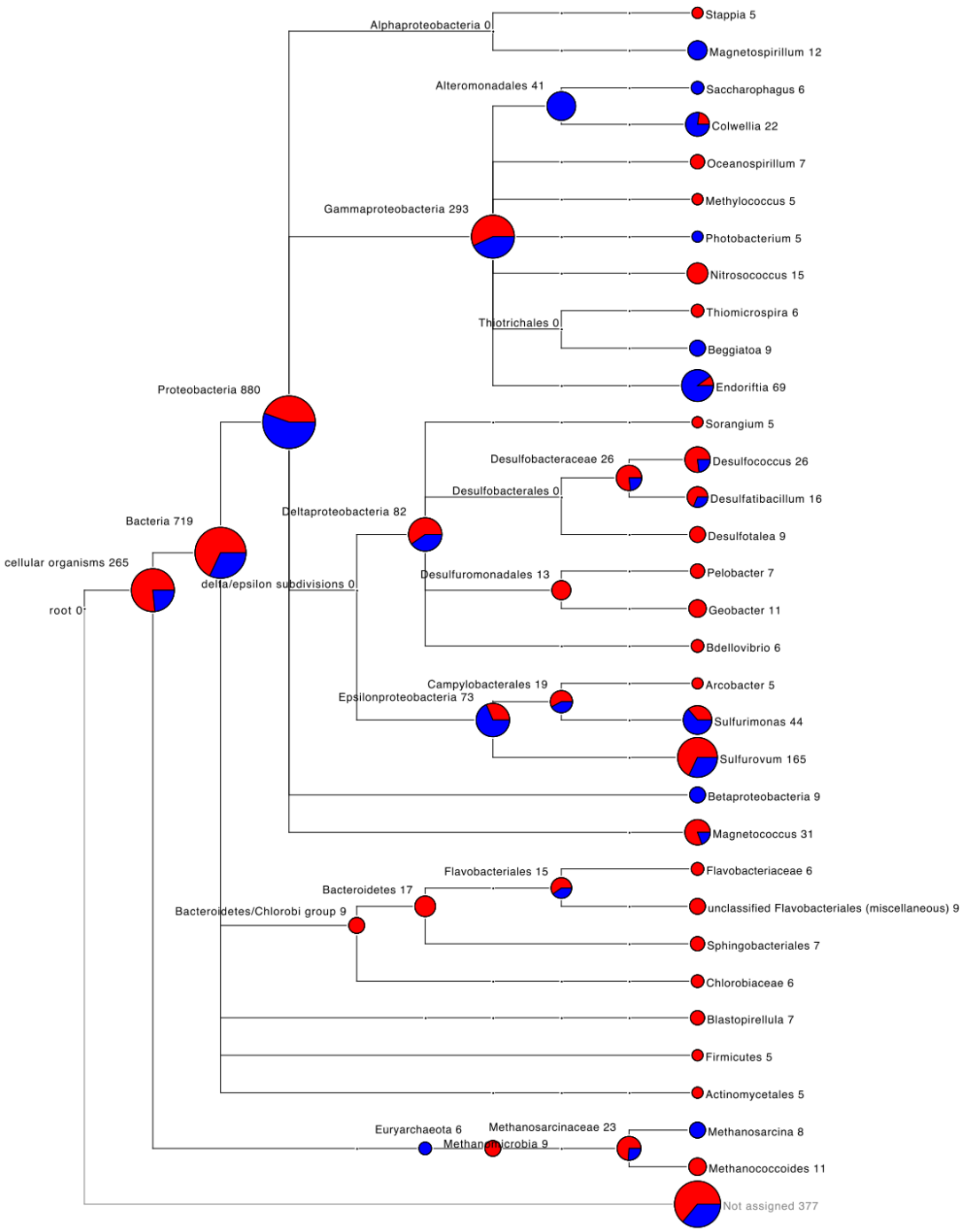
Adaptation to environment

(Kotlar et al., 2011)

(Total) MetaGenomics sequencing - tools

multiple samples

Comparison between reads assigned to **Phosphorus** metabolism and **Nitrogen** metabolism



(Total) MetaGenomics sequencing - tools

Other tools

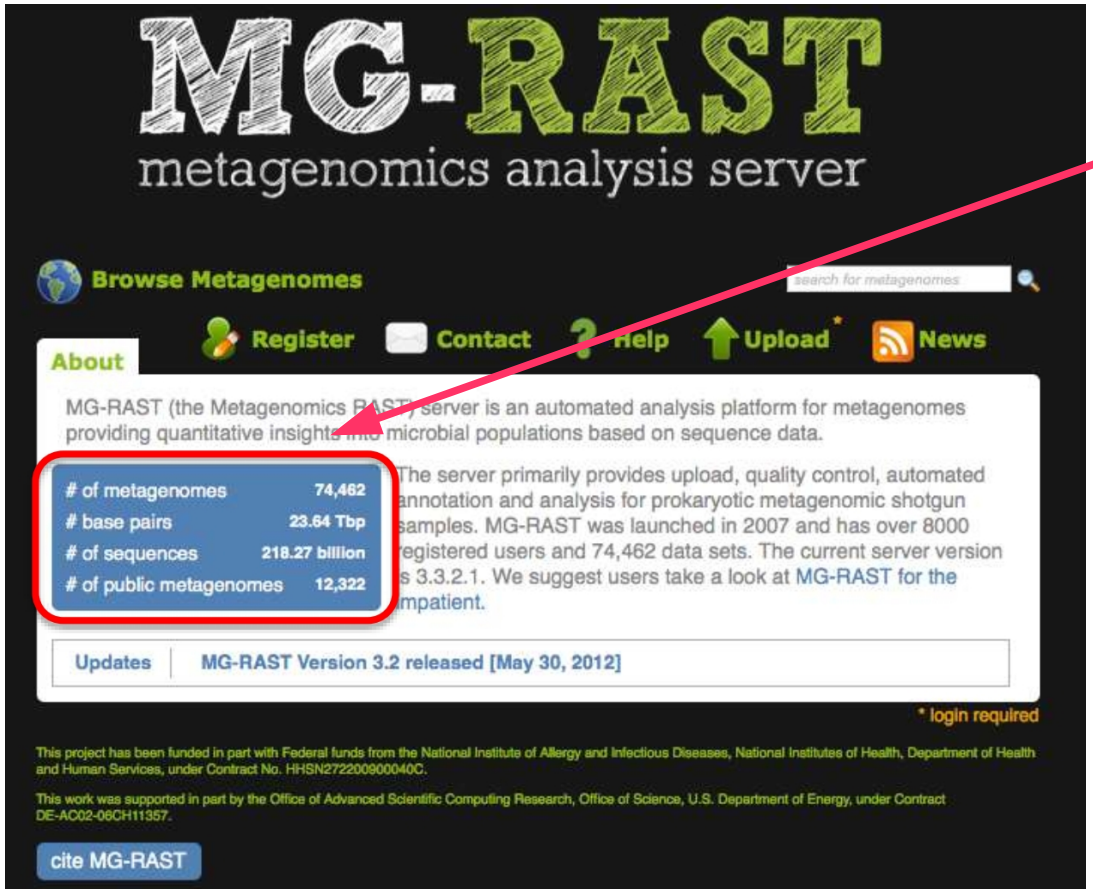
- MG-RAST* (http://metagenomics.anl.gov/) (1,2)
- IMG/M (http://img.jgi.doe.gov/) (1)
- WebMGA (http://weizhong-lab.ucsd.edu/metagenomic-analysis/) (1)
- METAgen assist* (http://www.metagenassist.ca/METAGENassist/faces/Home.jsp) (1,2)
- Real-Time metagenomics (https://edwards.sdsu.edu/RTMg/) (1)
- Ribosomal Database Project (RDP) (rdp.cme.msu.edu) (2)
- Qiime (Quantitative Insights Into Microbial Ecology) (www.qiime.org) (1)
- Mega (and not “Megane”), more focus on phylogeny (https://www.megasoftware.net/) (2)
- Mothur (https://www.mothur.org/) (1,2)

(1) MetaG
(2) Barcoding / amplicon sequences

(Total) MetaGenomics sequencing - tools

MG -RAST (online tool)

http://metagenomics.anl.gov/

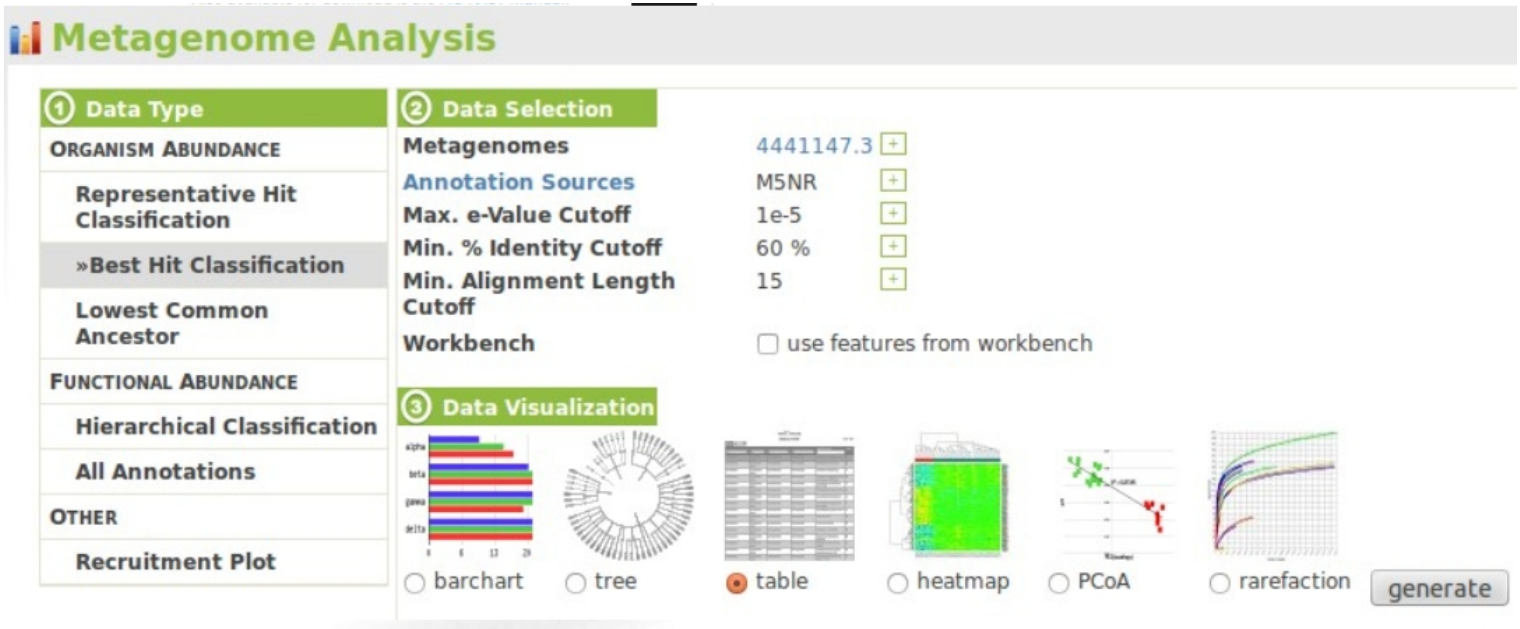


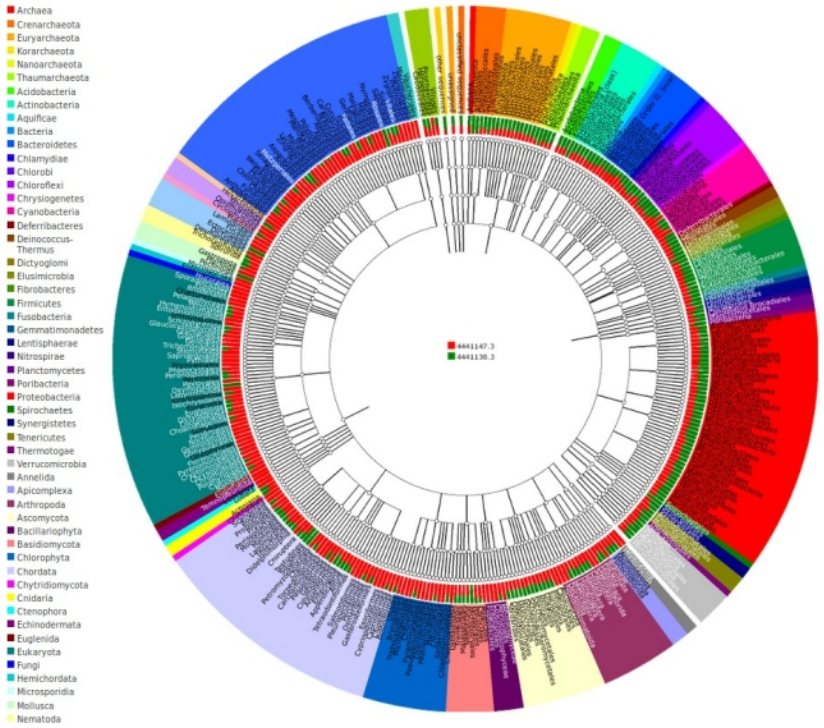
2012

2023

(Total) MetaGenomics sequencing - tools

MG -RAST

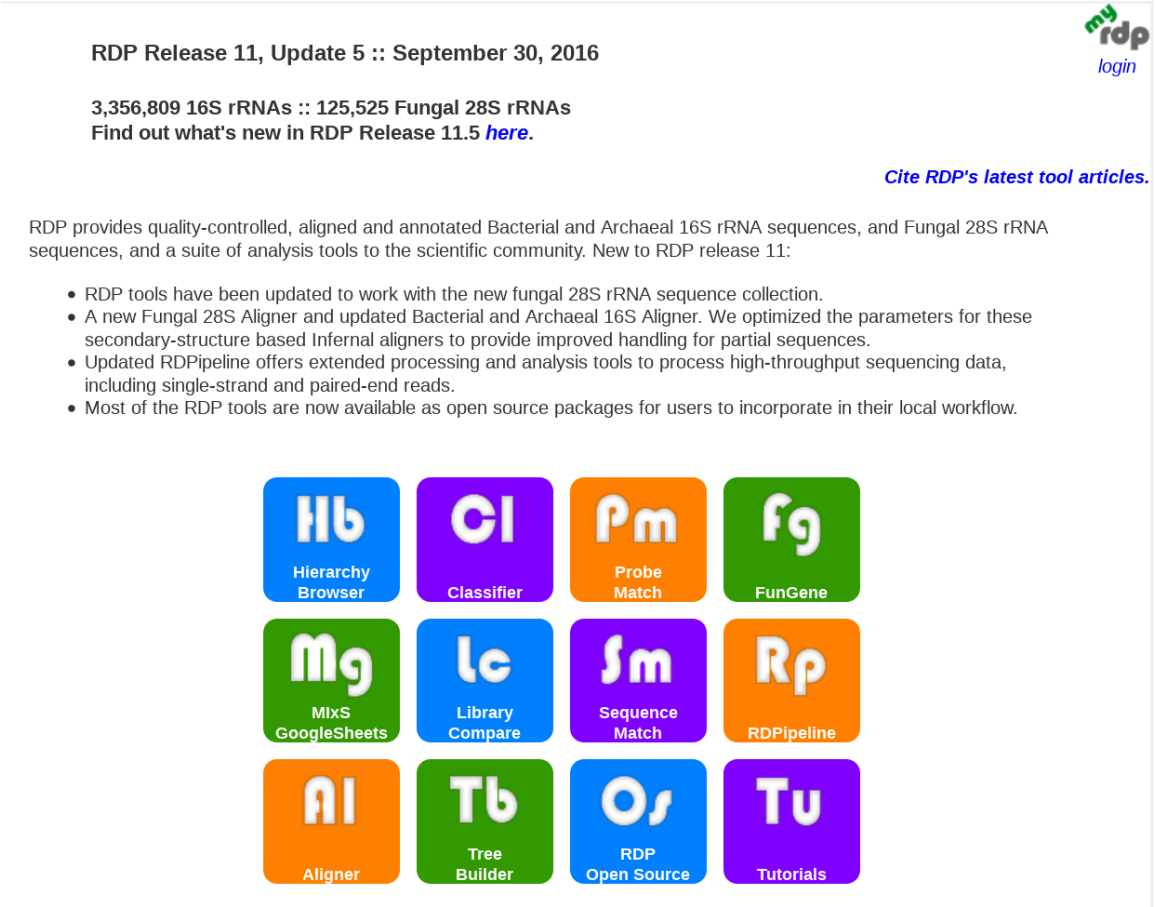




(Meta) Barcoding sequencing - tools

Ribosomal Database Project (RDP)

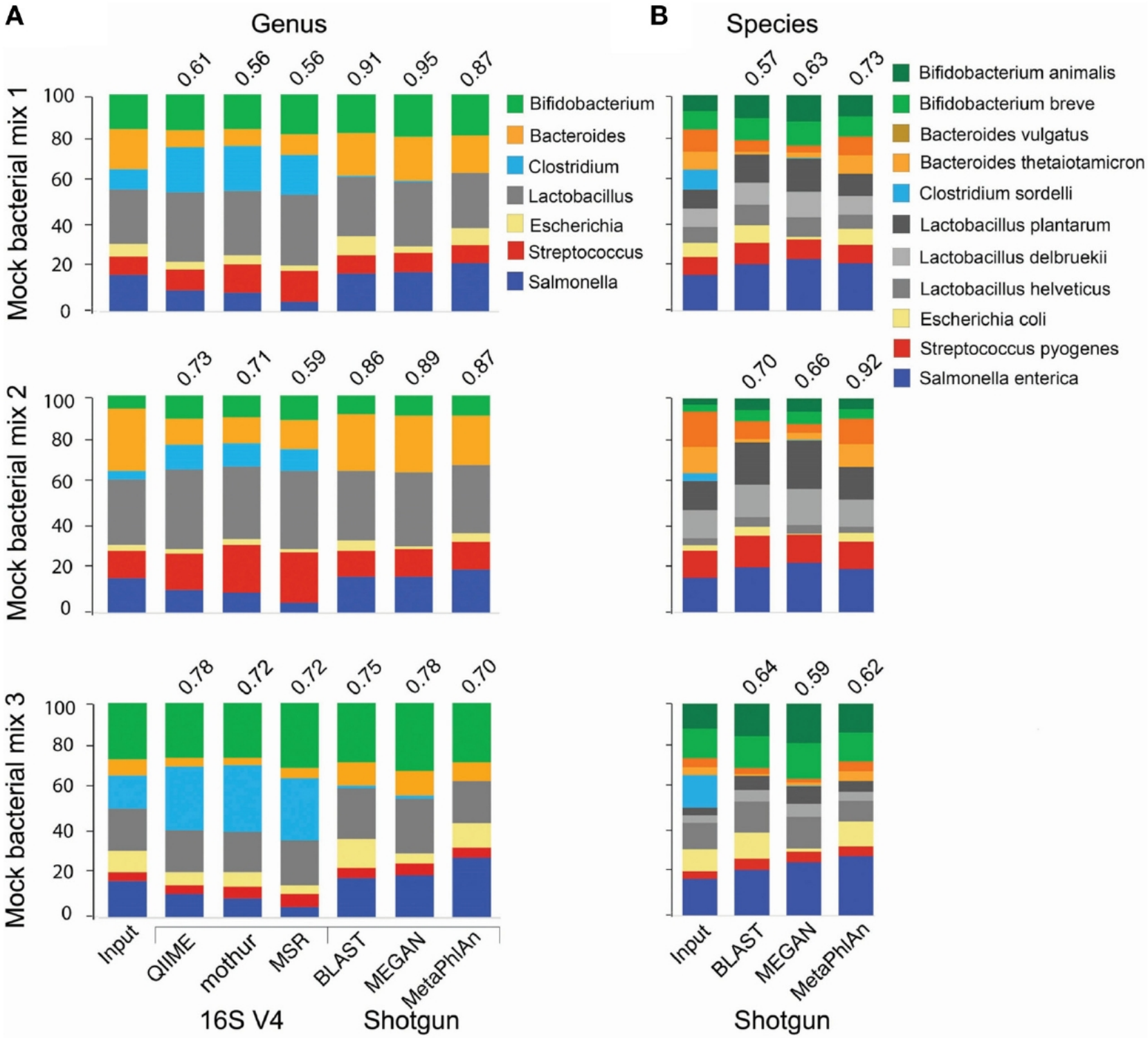
https://rdp.cme.msu.edu/



https://rdp.cme.msu.edu/help/tutorial.jsp

http://rdp.cme.msu.edu/tutorials/init_process/RDPtutorial_INITIAL-PROCESS.html

https://rdp.cme.msu.edu/tutorials/classifier/classifer_cover_page.html


A
Genus
B
Species
Mock bacterial mix 1
Mock bacterial mix 2
Mock bacterial mix 3
0.61
0.56
0.56
0.91
0.95
0.87
0.73
0.71
0.59
0.86
0.89
0.87
0.78
0.72
0.72
0.75
0.78
0.70
0.57
0.63
0.73
0.70
0.66
0.92
0.64
0.59
0.62
Bifidobacterium
Bacteroides
Clostridium
Lactobacillus
Escherichia
Streptococcus
Salmonella
Bifidobacterium animalis
Bifidobacterium breve
Bacteroides vulgatus
Bacteroides thetaiotamicron
Clostridium sordelli
Lactobacillus plantarum
Lactobacillus delbruekii
Lactobacillus helveticus
Escherichia coli
Streptococcus pyogenes
Salmonella enterica
Input
QIIME
mothur
MSR
BLAST
MEGAN
MetaPhlAn
16S V4
Shotgun
Input
BLAST
MEGAN
MetaPhlAn
Shotgun