



# Introduction to MetaGenomics



Institut de Recherche sur le Cancer et le Vieillessement, Nice  
Institute for Research on Cancer and Aging, Nice  
CNRS UMR 7284 - INSERM U 1081 - UNS



**Olivier Croce** – [croce@unice.fr](mailto:croce@unice.fr)  
Bioinformatics service - IRCAN

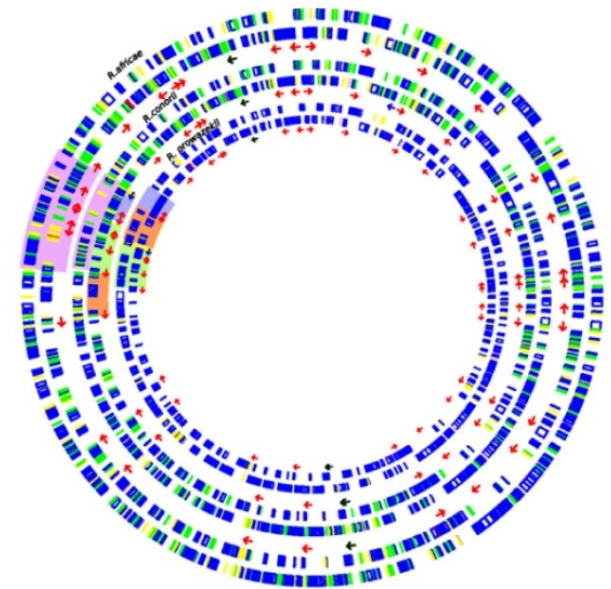
# Definitions

## 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



■ Genes common to all 3 rickettsiae  
■ Genes common to *R. africae*/*R. conorii*  
■ Genes common to *R. conorii*/*R. prowazekii*  
■ Genes common to *R. africae*/*R. prowazekii*  
■ Specific genes  
♦ tRNA  
♦ Ribosomal RNA  
♦ Other RNA

## MetaGenomics:

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

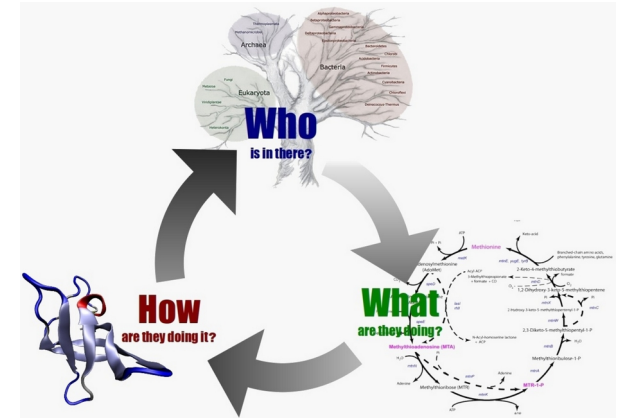
**Metagenome:** the collective genome of all the microorganisms in a given environment.



# 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

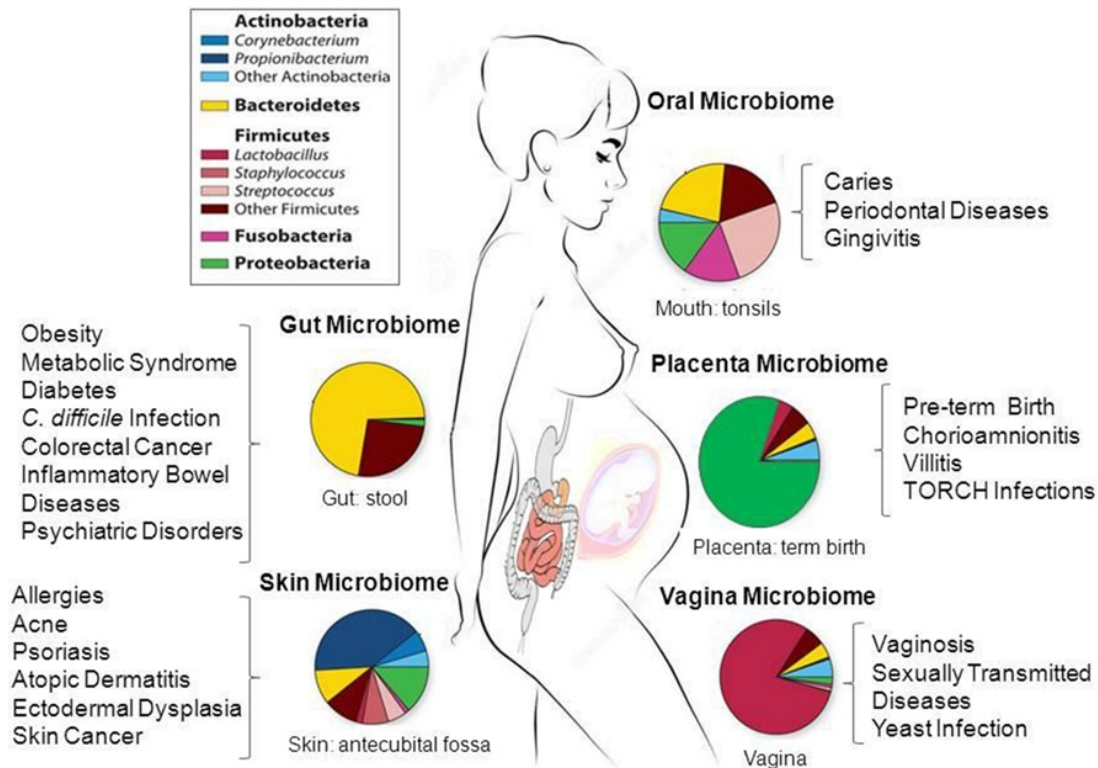
- All the micro-organisms (bacteria, archaea, fungi, viruses), colonizing and interact with the host in various organs
- The microbiota is involved in many essential functions
- Altered in many pathologies :
  - Metabolic diseases (Gurung *et al.* 2020, Liu *et al.* 2021)
  - Auto-immune diseases (De 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 !



TARA  
length: 36 meters  
width: 10 meters  
height of mast: 27 meters

## THE TARA OCEANS EXPEDITION 2009-2013

REVIVING THE TRADITION OF THE GREAT EXPEDITIONS OF THE 19<sup>TH</sup> CENTURY, TARA SAILED THE WORLD'S OCEANS FOR THREE AND A HALF YEARS. FOR THE FIRST TIME, MARINE PLANKTON IN ITS ENTIRETY WAS COLLECTED AND STUDIED - FROM VIRUSES AND BACTERIA TO FISH LARVAE AND JELLYFISH.

**WHY THIS EXPEDITION?**

**OXYGEN**  
**CARBON**

THE OCEANS regulate the climate and atmosphere of our planet. Plankton produce half of the oxygen generated globally each year by photosynthesis, and absorb atmospheric CO<sub>2</sub>. Affected by pollution, over-fishing, and rising temperatures, will plankton continue to efficiently absorb carbon and regulate the climate?

PLANKTON designates all the organisms drifting with the currents. These microscopic organisms are the foundation of the marine food chain, ensuring the survival of fish, marine mammals, and billions of humans beings. They react quickly to climate changes and to ocean acidification. We must learn more about this complex, dynamic ecosystem and its role in global equilibrium.

CORAL REEFS are privileged places for aquatic biodiversity, but they are suffering from climate change, marine pollution, and over-fishing. Tara was the ideal platform for exploring 5 rarely-studied coral sites: Djibouti, Saint-Brandon, Mayotte, and the islands of Gambier and Kiribati.

**A CONCENTRATION OF HIGH TECH**

A unique space for microscopic imagery set up aboard Tara - the dry lab - where researchers characterize the organisms collected, their functional diversity and their complexity.

THE UNDERWATER VISION PROFILER observes plankton during sampling.

THE FLOW CAM is used to count and identify organisms as they pass through a laser beam at high speed.

**VOYAGE OF THE SAMPLES**

PORTS-OF-CALL  
At stopovers every 6 to 8 weeks, the samples - conserved with liquid nitrogen, alcohol and fixatives - were sent to partner laboratories.

WORLD COLLABOR  
This international specialist in shipping sensitive products expedited the precious samples collected aboard Tara to Heidelberg (Germany), then redistributed them to partner laboratories around the world.

**FROM VIRUSES TO FISH LARVAE**  
TO SAMPLE ZOOPLANKTON, WE NEED 1 000 000 LITERS OF SEA WATER.

**THREE METHODS OF COLLECTION AND OBSERVATION. MORE THAN 35,000 SAMPLES**

- 1 Nets**  
Tara deployed 7 types of nets (mesh sizes from 5 to 690 microns) immersed between the surface and 3,000 meters deep. The specialized Manta net is used for collecting plastic on the surface.
- 2 PERISTALTIC PUMP**  
Water is pumped from a depth of 10 to 120 meters, then passes through a series of strainers and filters to separate organisms by size.
- 3 THE "BOUTEILLE" CTD AND THE MVP**  
This apparatus contains 10 Niskin bottles to collect water from different depths, as well as instruments to characterize many parameters including pressure, temperature, conductivity, nitrogen, oxygen, fluorescence, etc. The bottles are programmed to collect water at different depths. The MVP (Underwater Vision Profiler) deployed down to a depth of 2,000 meters allowed us to record about 20 physico-chemical parameters, and image particles and organisms.

**ZOOPLANKTON**  
1 TO 10,000 IN A LITER OF SEA WATER.  
Zooplankton consists of tiny animals, for example copepods, embryos and larvae, but also huge animals like jellyfish and siphonophores. They feed on living matter: bacteria, protists, or other multicellular organisms. Most zooplankton migrates to the surface, or to great depths to feed and protect themselves from predators during the night.

**PROTISTS, INCLUDING PHYTOPLANKTON**  
1 TO 100 MILLION IN A LITER OF SEA WATER.  
The principal ocean biodiversity consists of multitudes of species of unicellular organisms with a nucleus: the protists. Certain of them (diatoms, dinoflagellates, etc.) are photosynthetic. Along with cyanobacteria, they constitute phytoplankton, and are the base of the food chain. Phytoplankton produces half of the oxygen on the planet and absorbs half of atmospheric carbon, thus acting as a major regulator of climate.

**BACTERIA**  
1 TO 10 BILLION IN A LITER OF SEA WATER.  
Bacteria are prokaryotes: cells without nuclei. Certain species - the cyanobacteria - can perform photosynthesis. They are a food for protists and certain zooplankton. Bacteria are responsible for a wide array of metabolic functions in the ocean.

**VIRUSES**  
10 TO 100 BILLION IN A LITER OF WATER.  
The marine virosphere is immense, and includes the phages (viruses of bacteria) and giant viruses (giruses). Viruses play an essential role in recirculating living matter.

**SCIENTIFIC RESULTS**

Based on the data from Tara Oceans, many scientific articles detailing the planktonic ecosystem and its dynamics have been published, or are on the way to being published in international journals. Ongoing analysis of this data, thanks to the Oceanomics project\* will help establish a reference for ocean ecosystems, and set up a method for predicting and following the evolution of these ecosystems in relation to climate change.

\*Oceanomics (an Investissements d'Avenir project) aims to promote national and worldwide use of marine plankton, one of the planet's most important organisms in terms of biodiversity, bio-resources, and global ecological changes.

september 2009 - december 2013  
60 STOPOVERS, 35 COUNTRIES  
140,000 KILOMETERS AROUND THE WORLD

2013  
Tara crossed the Northeast (Russian) and Northwest (Canadian) passages. Scientists aboard accomplished a complete sampling of marine organisms at the edge of the ice cap.

OCTOBER 2011  
Tara crossed the "Plastic Continent" - a calm region where marine currents concentrate floating debris that accumulates in masses.

FEBRUARY 2010  
Tara crossed the Gulf of Aden, a very dangerous region infested with pirates. Research was voluntarily interrupted for two weeks.

JANUARY 2011  
Scientists collected samples during one month in Antarctic waters. This was the first Tara Oceans mission in a polar region.

2 Ocean regions undergoing acidification  
3 Minimum oxygen zones

**PARTNER LABORATORIES**  
23 LABS AND SCIENTIFIC INSTITUTIONS

- 8 in France
- 5 in the United States
- 2 in Germany
- 2 in Italy
- 1 in Belgium
- 1 in Ireland
- 1 in Spain
- 1 in Canada
- 1 in Saudi Arabia
- 1 in Russia

**THE "TARANAUTES"**  
TAKING TURNS ON BOARD:

- 90 crew members, artists, and journalists
- 160 researchers
- 140 researchers involved in the lab work
- 12 scientific fields
- 40 nationalities

WE DO DATA WWW.TARA OCEANS.ORG

# 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,<sup>1\*</sup> Anna Lewin,<sup>2\*</sup> Jostein Johansen,<sup>3</sup> Mimmi Throne-Holst,<sup>4\*</sup> Thomas Haverkamp,<sup>5</sup> Sidsel Markussen,<sup>4</sup> Asgeir Winnberg,<sup>4</sup> Philip Ringrose,<sup>1</sup> Trine Aakvik,<sup>2</sup> Einar Ryeng,<sup>3</sup> Kjetill Jakobsen,<sup>5</sup> Finn Drablos<sup>3</sup> and Svein Valla<sup>2\*</sup>

<sup>1</sup> Statoil ASA, 7053 Ranheim, Norway.

<sup>2</sup> Department of Biotechnology, Norwegian University of Science and Technology, 7491 Trondheim, Norway.

<sup>3</sup> Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, 7491 Trondheim, Norway.

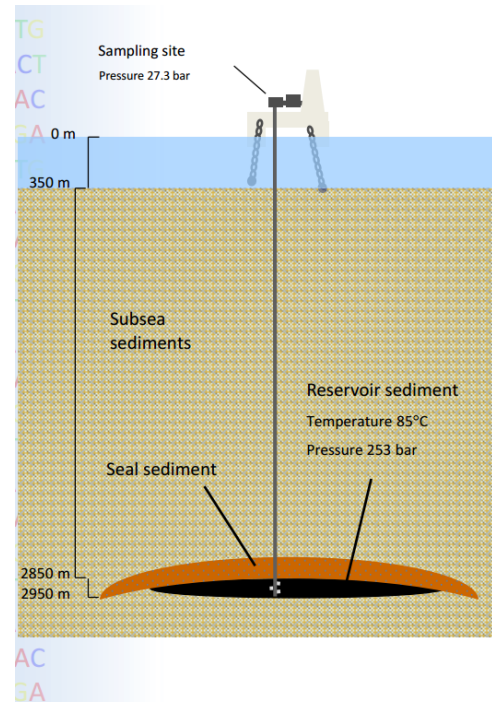
<sup>4</sup> SINTEF Materials and Chemistry, Department of Biotechnology, 7465 Trondheim, Norway.

<sup>5</sup> CEES and MERG, Department of Biology, University of Oslo, 0316 Oslo, Norway.

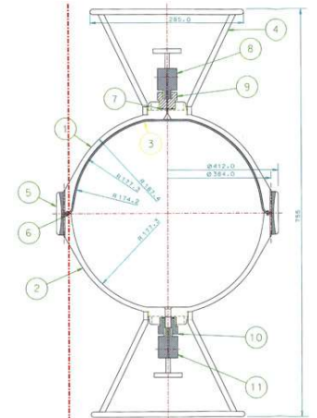
erant than for a corresponding *E. coli* enzyme, consistent with the conditions in the oil reservoir.

### 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



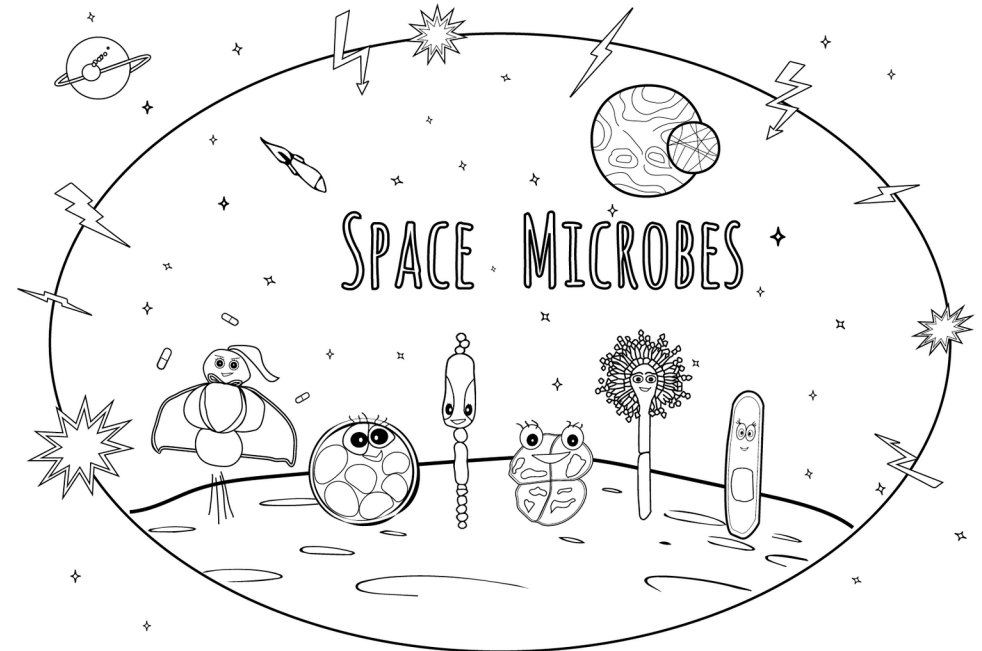
# Space

- Microbes can survive in extreme conditions such as outer space

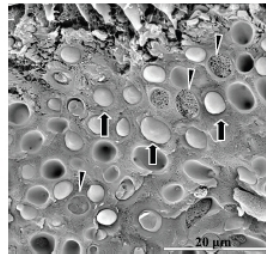
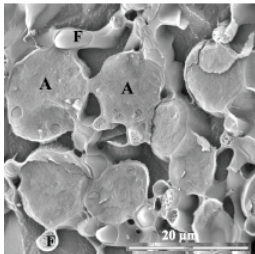
**Survival of *B.subtilis* spores :**  
Unprotected : several seconds  
Protected : more than 6 years

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

Recently : surprising **survival of lichen**

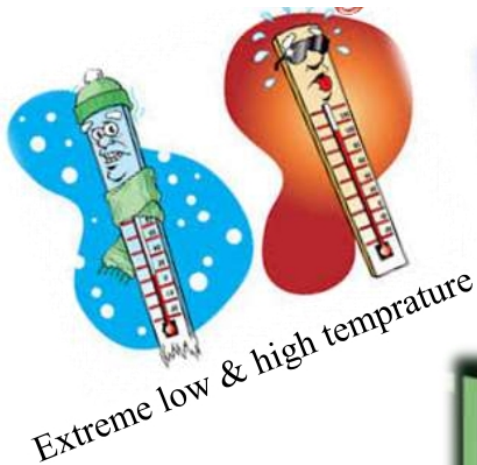


thechroniclesofspacemicrobes.wordpress.com | facebook/spacemicrobes



# And more !

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



volcano



Soil



Waste water



Acidic



alkaline



# 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)

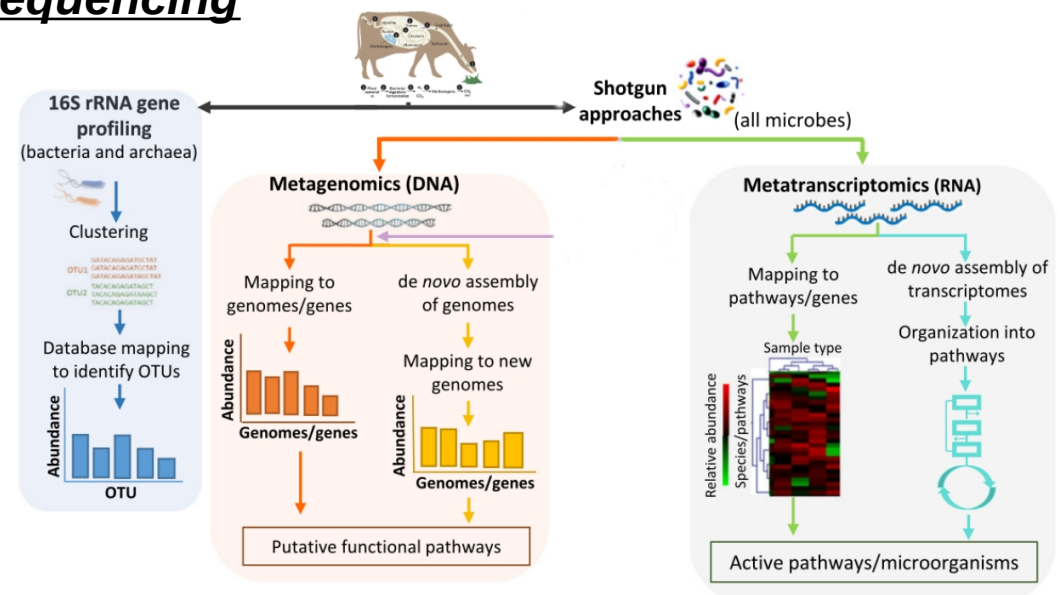
- (WGS / Total / Shotgun) MetaGenomics sequencing

Sequencing of the whole DNA in a sample.

Complete community analysis, characterization of the pangenome

- \* MetaTranscriptomics

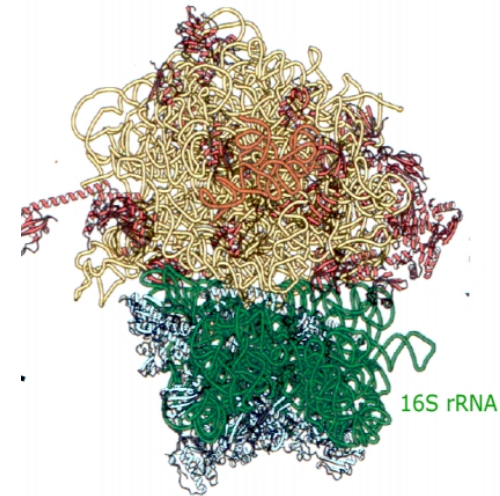
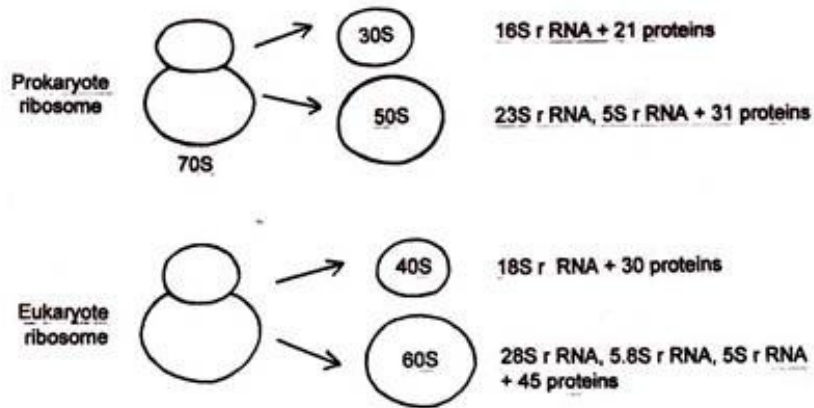
Idem TotalMetagenomics, but from RNA



# (Meta) Barcoding sequencing

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

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

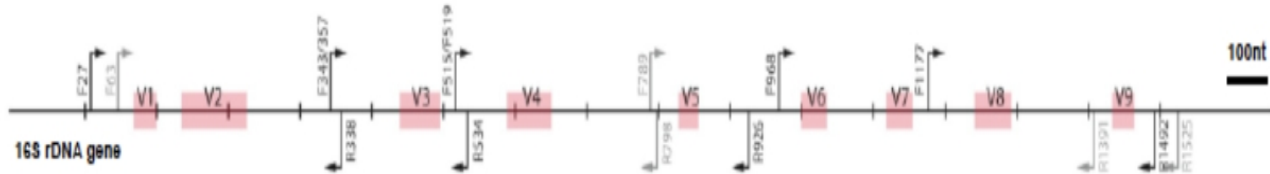


- Most commonly used molecular marker: essential function, ubiquity, evolutionary properties
- Highly conserved gene, 9 hypervariable domains interspersed with conserved fragments
- Rapid and cost-effective approaches for assessing diversity and abundance.
- OTU (operational taxonomic units) definition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

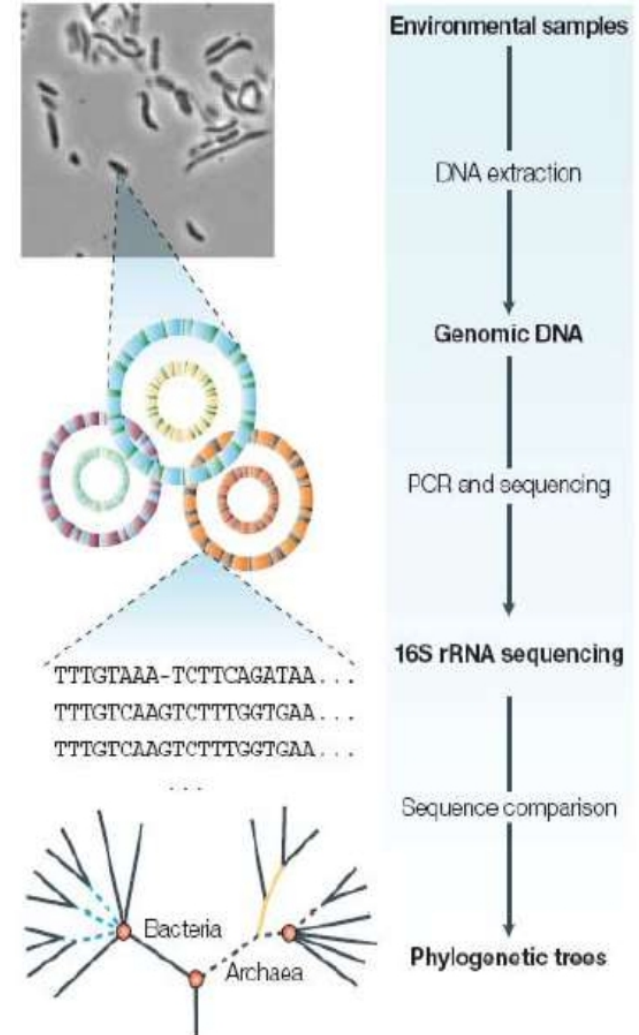
=> organisms displaying 97 to 98% identity in this gene to be part of the same OTU

# (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?

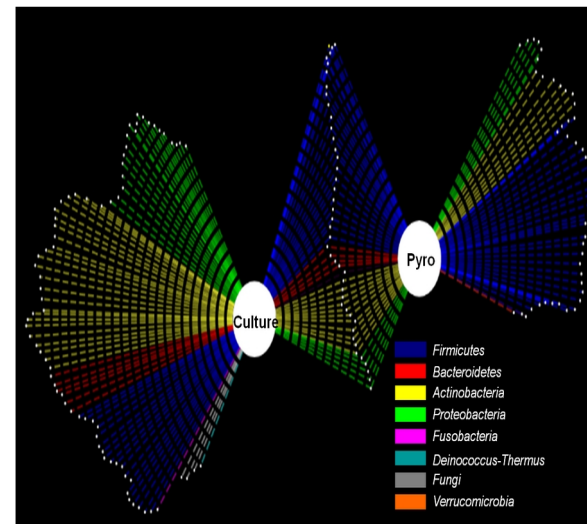
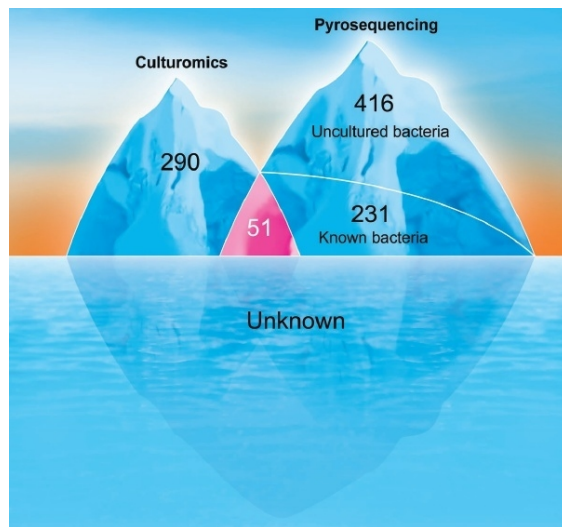
# (Meta) Barcoding sequencing

## 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]

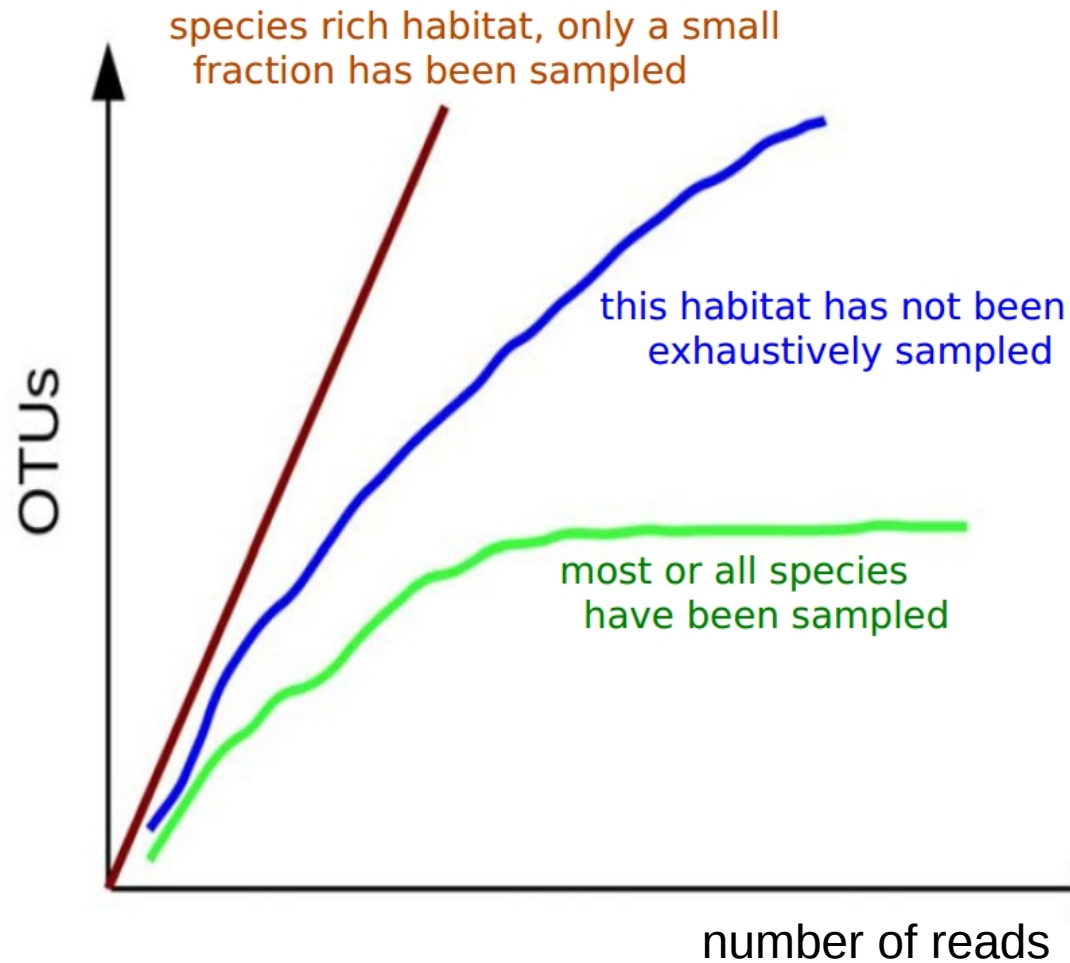
- Missing sequences or variable quality of available sequences, including from validly published species



- **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 characterization
  - Who does what ? => Physiological characterization

## RESEARCH ARTICLE

### Environmental Genome Shotgun Sequencing of the Sargasso Sea

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

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<sup>1</sup>, C. Bowler<sup>2</sup>, C. de Vargas<sup>3,4</sup>, G. Gorsky<sup>5,6</sup>, E. Karsenti<sup>2,7</sup>, P. Wincker<sup>8</sup>  
+ 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).

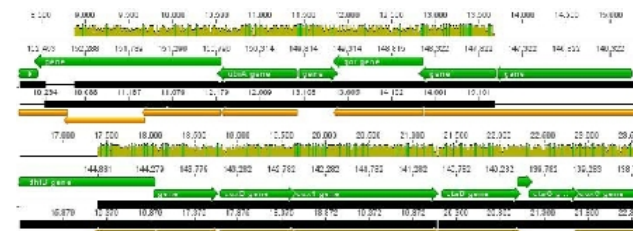
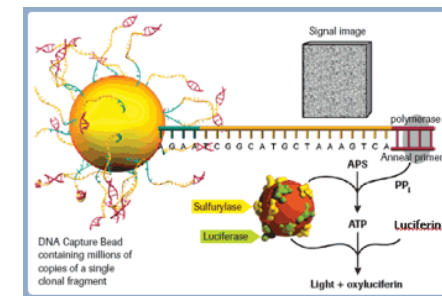
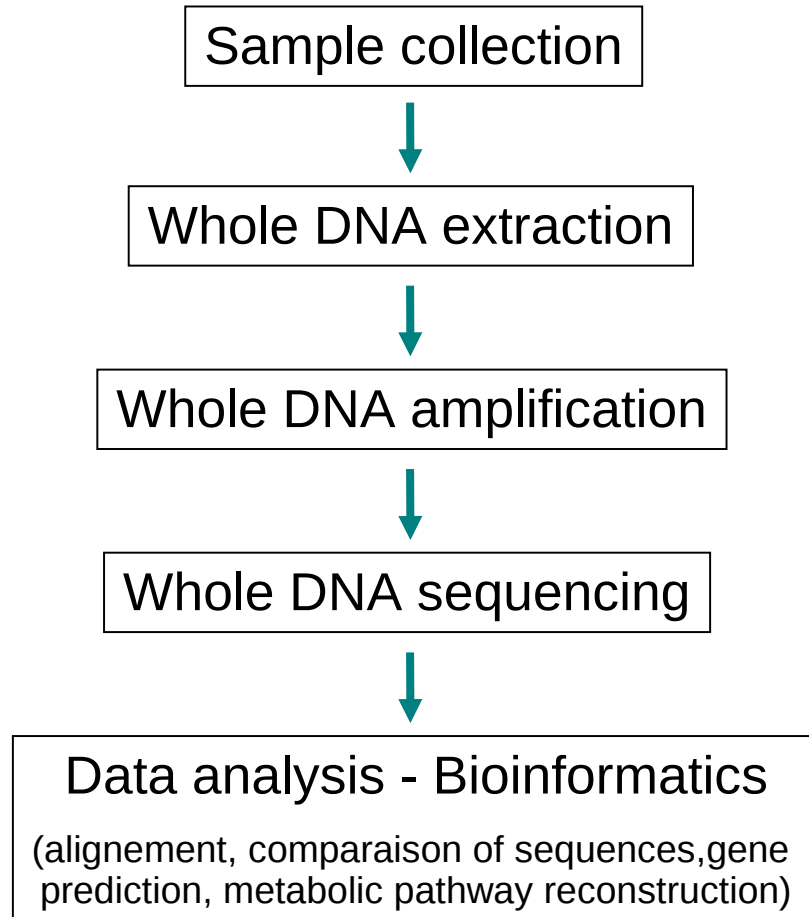
Success depended on collaboration between scientists and the *Tara* Expeditions logistics team. The journey involved not only science but also outreach and education as well as negotiation through the shoals of legal and political regulations, funding uncertainties, threats from pirates, and unpredictable 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

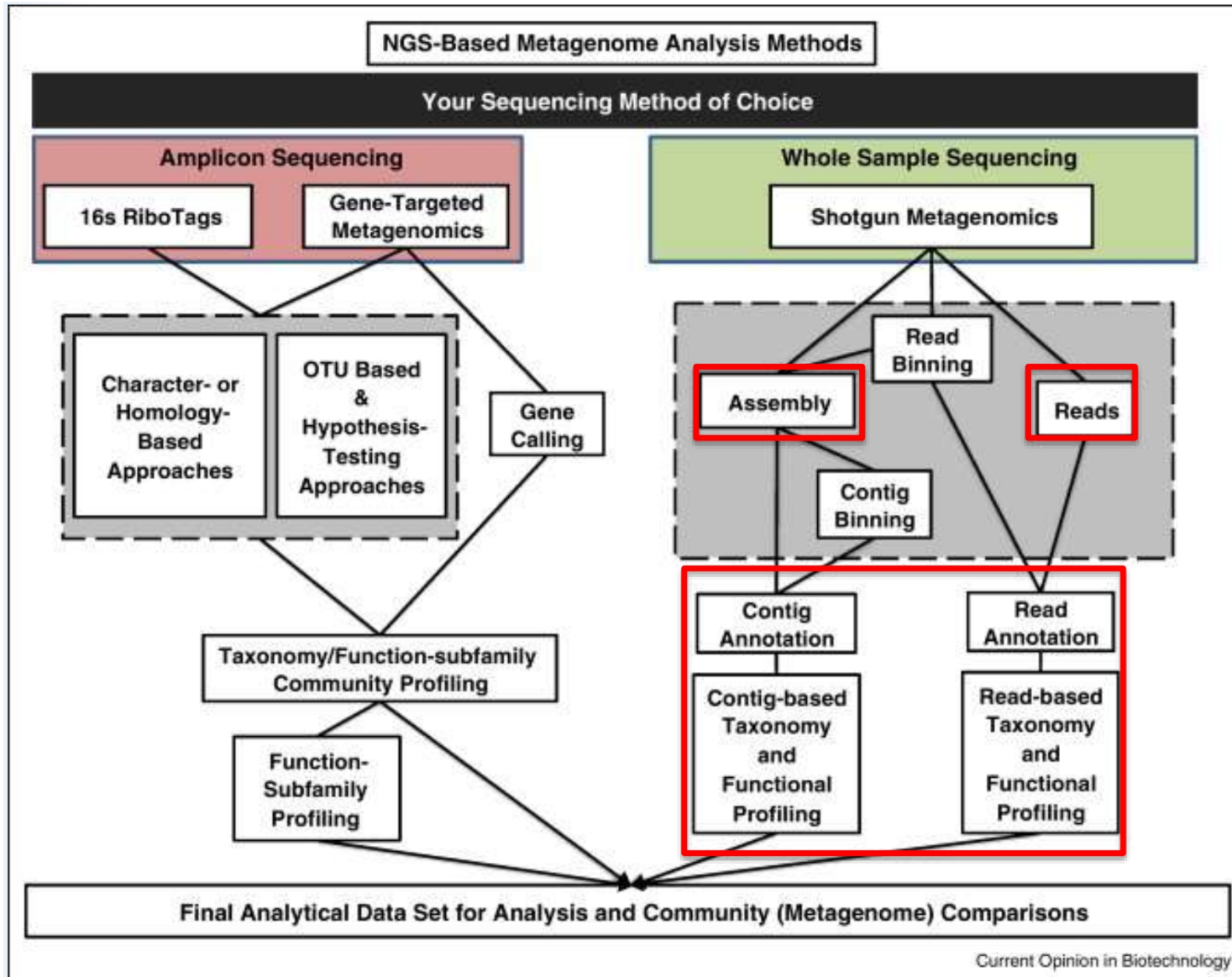
## Principle of metagenomics



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

# (Total) MetaGenomics sequencing

- Methods differences between barcoding and shotgun sequencing



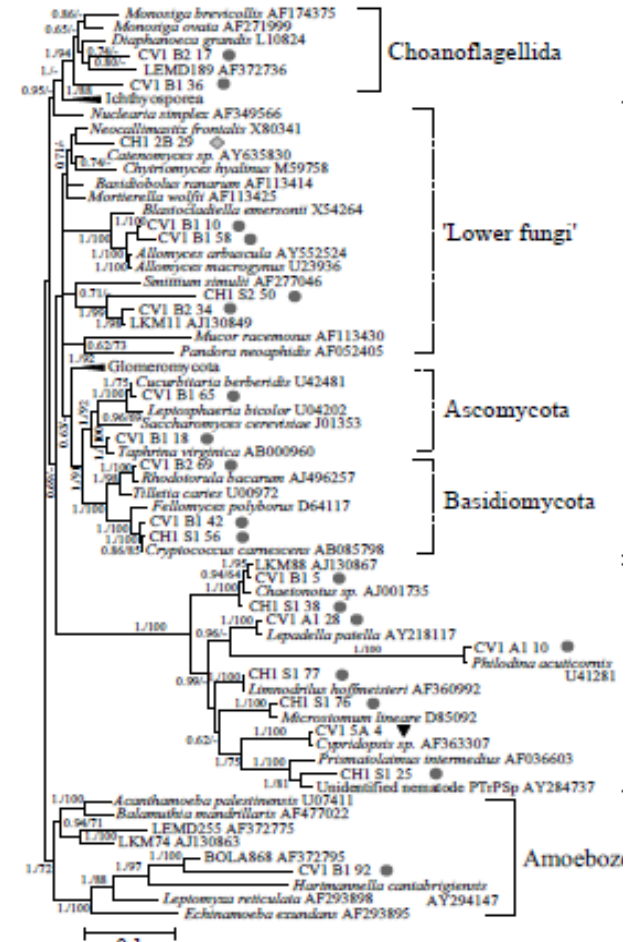
# (Total) MetaGenomics sequencing

## (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

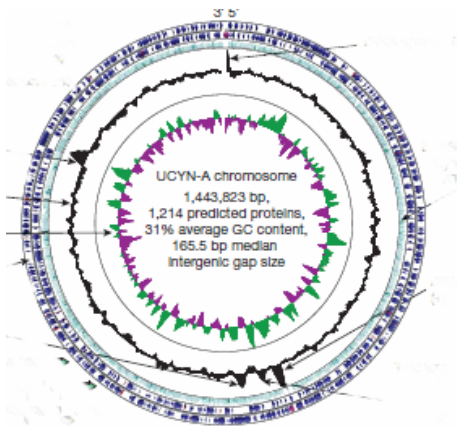




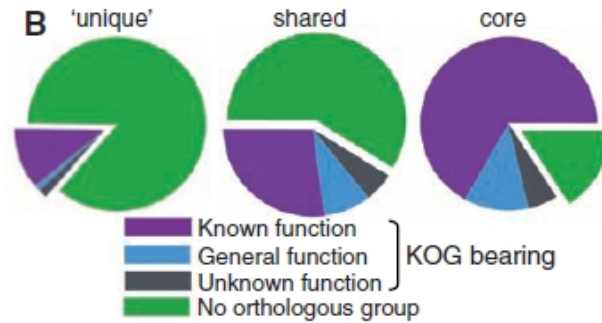
# (Total) MetaGenomics sequencing

## (2) Physiology and genomic

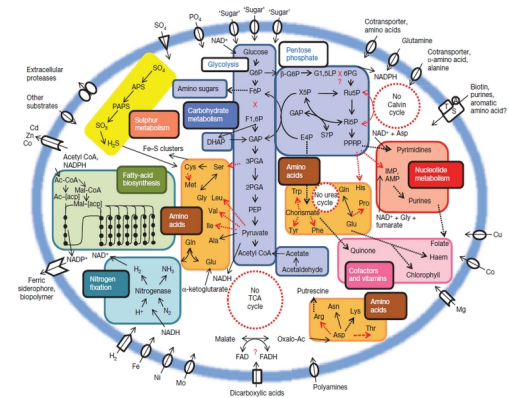
### Genomes



### 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 :**

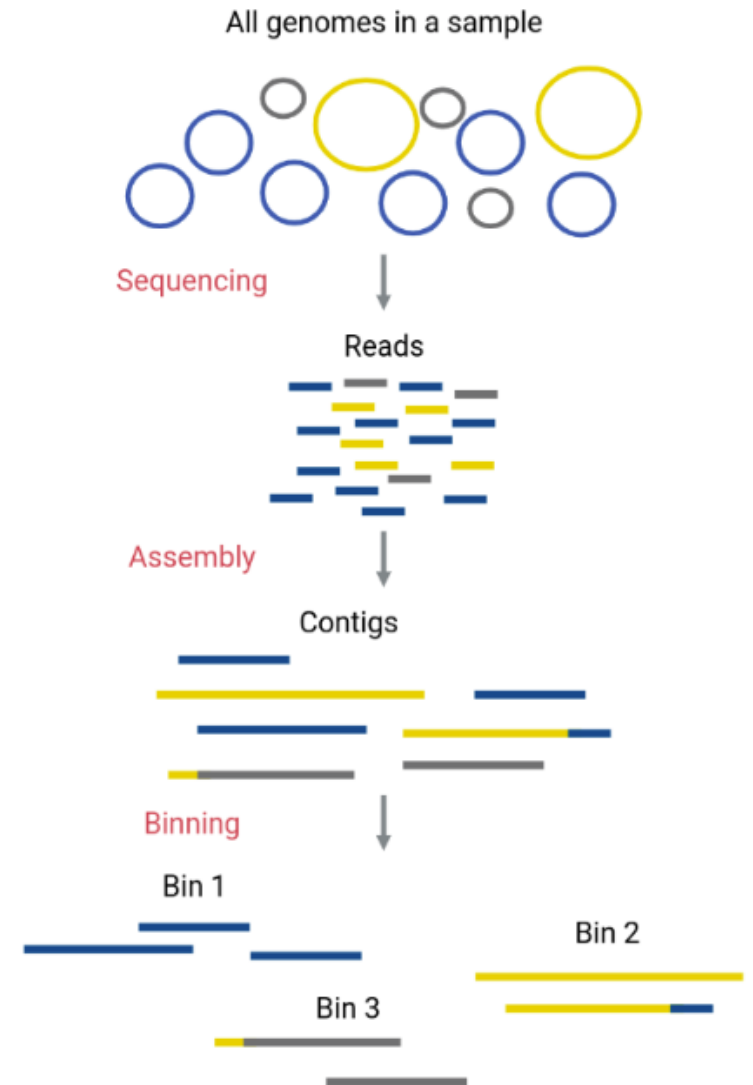
- Direct use of reads (after trimming and quality tests)
- Assemble reads into contigs
- Both

- **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



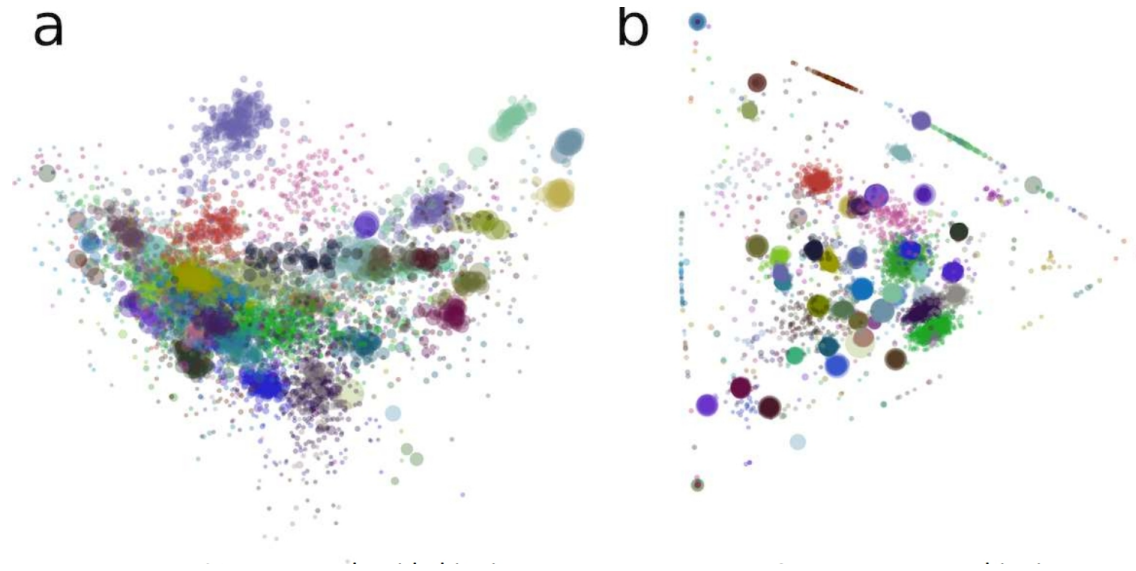
# (Total) MetaGenomics sequencing

## (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

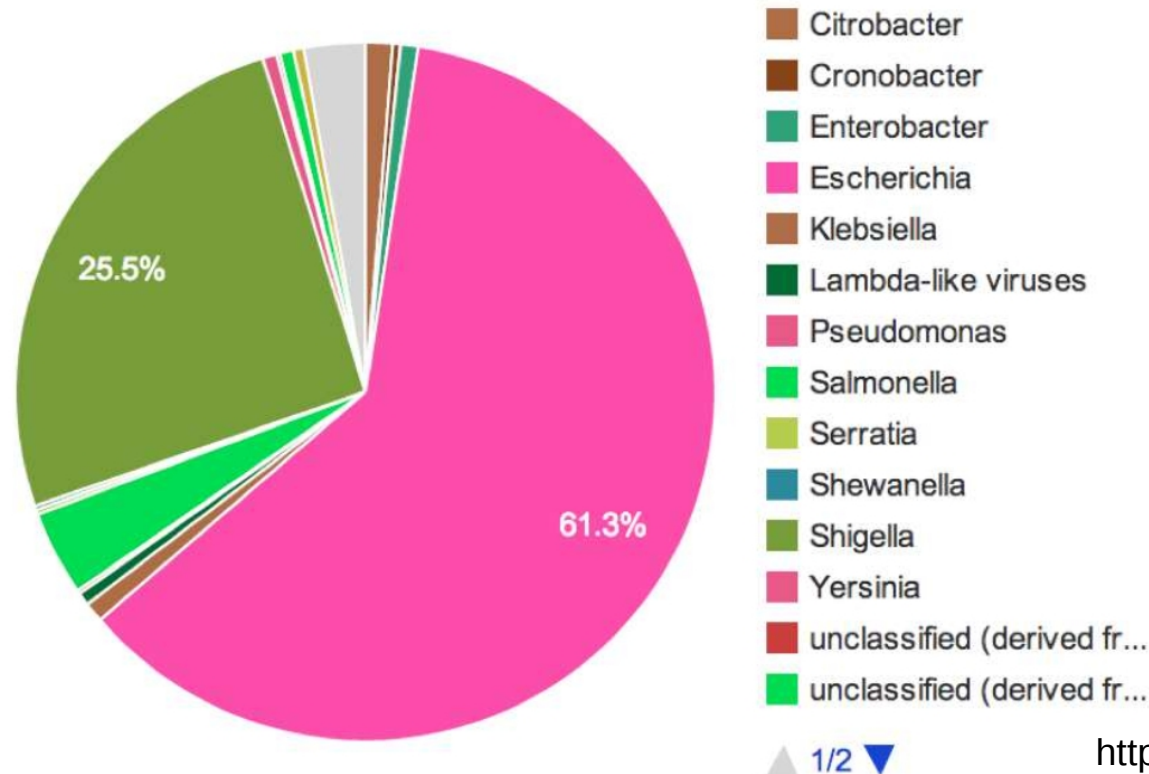
## (2) Sequences similarity

- 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



# (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, Phylopythia, RAphy, Metaxa2 (rRNA), PhyloOTU (rRNA), MLTreeMap, RITA, STAMP, WGSQuikr.

### **Differential Coverage:**

GroopM, Concoct, Blobology



# (Total) MetaGenomics sequencing - tools

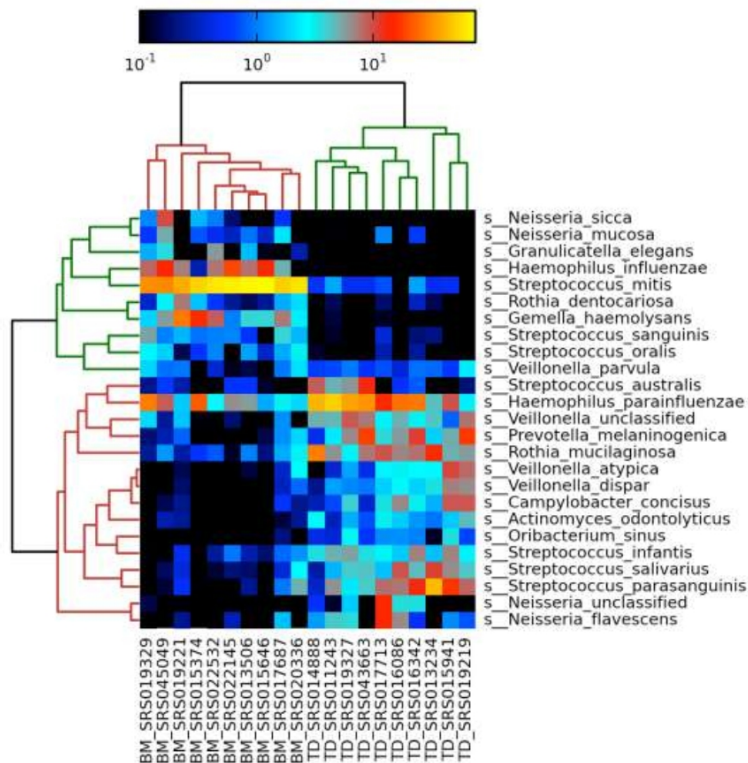
## MetaPhlan: Metagenomic Phylogenetic Analysis

- \* 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)

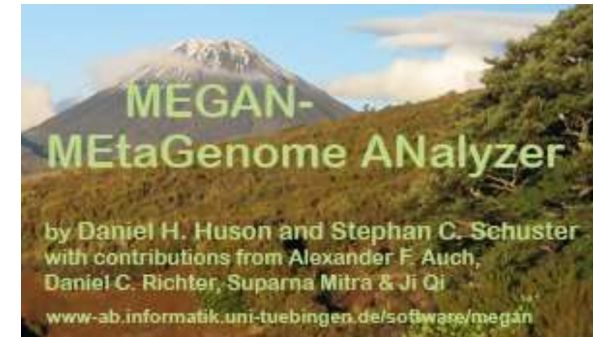
<https://github.com/husonlab/megan-ce>  
<http://megan.informatik.uni-tuebingen.de/>

*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

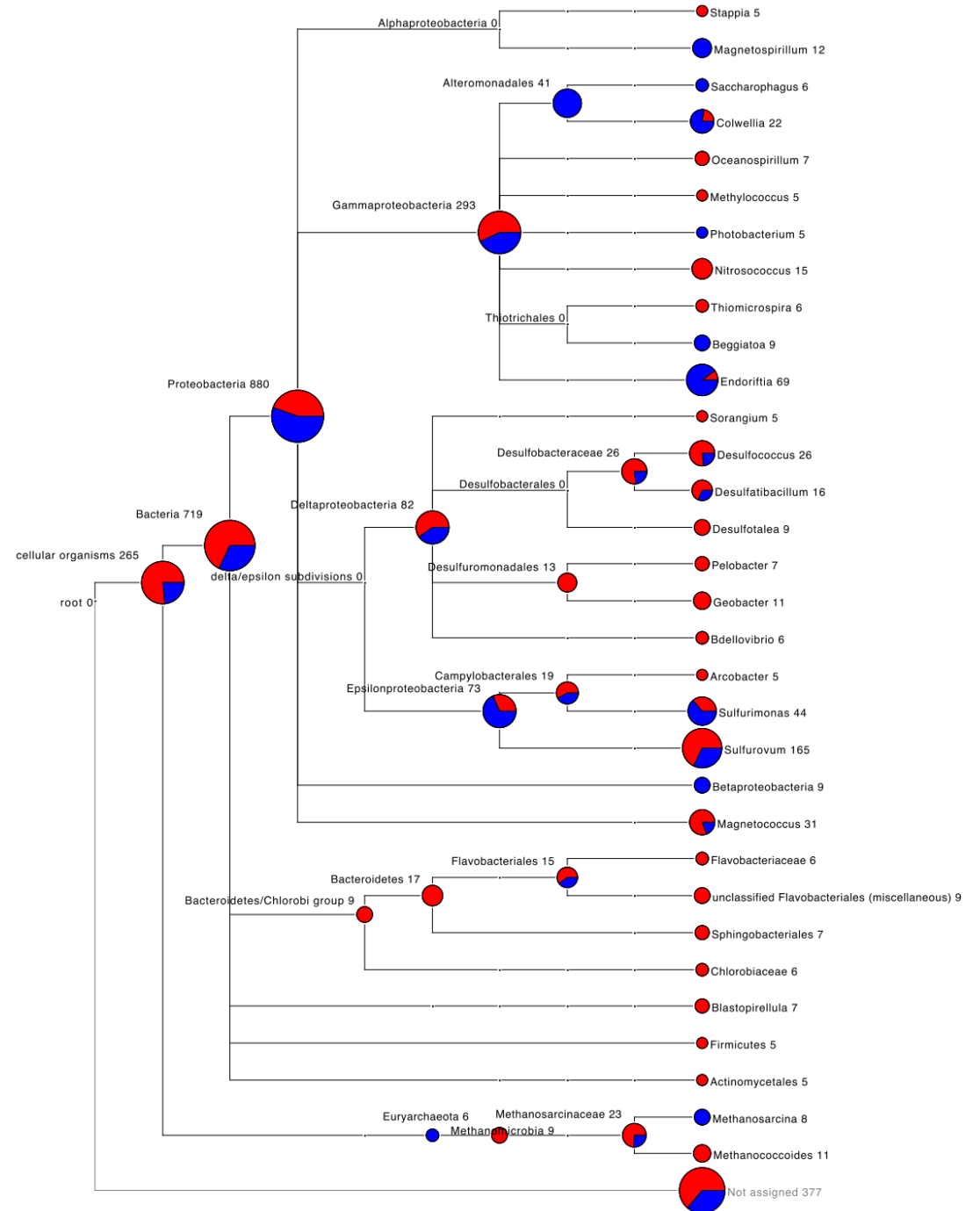
*“ In this approach, every read is assigned to some taxon. If the read aligns very specifically only to a single taxon, then it is assigned to that taxon. The less specifically a read hits taxa, the higher up in the taxonomy it is placed. Reads that hit ubiquitously may even be assigned to the root node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



# (Total) MetaGenomics sequencing - tools

## 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](http://rdp.cme.msu.edu)) (2)
- Qiime (Quantitative Insights Into Microbial Ecology) ([www.qiime.org](http://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/>

**MG-RAST**  
metagenomics analysis server

Browse Metagenomes

Register Contact Help Upload News

About

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

# of metagenomes	74,462
# base pairs	23.64 Tbp
# of sequences	218.27 billion
# of public metagenomes	12,322

The server primarily provides upload, quality control, automated annotation and analysis for prokaryotic metagenomic shotgun samples. MG-RAST was launched in 2007 and has over 8000 registered users and 74,462 data sets. The current server version is 3.3.2.1. We suggest users take a look at **MG-RAST** for the impatient.

Updates | MG-RAST Version 3.2 released [May 30, 2012]

\* login required

This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900040C.

This work was supported in part by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract DE-AC02-06CH11357.

cite MG-RAST

2012

**MG-RAST**  
metagenomics analysis server

version 4.0.3

509,065 metagenomes containing 2,251 billion sequences and 336.87 Tbp processed for 38,448 registered users.

for programmatic access visit our API site

Please use your institutional email address for account requests.

search string e.g. mgp128 or mgm4447970.3 search

upload download analyze

Report

Turn your raw sequence into analyzed data.

2023

# (Total) MetaGenomics sequencing - tools

## MG -RAST

### Metagenome Analysis

**1 Data Type**

ORGANISM ABUNDANCE

Representative Hit Classification

» Best Hit Classification

Lowest Common Ancestor

FUNCTIONAL ABUNDANCE

Hierarchical Classification

All Annotations

OTHER

Recruitment Plot

**2 Data Selection**

Metagenomes 4441147.3

Annotation Sources M5NR

Max. e-Value Cutoff 1e-5

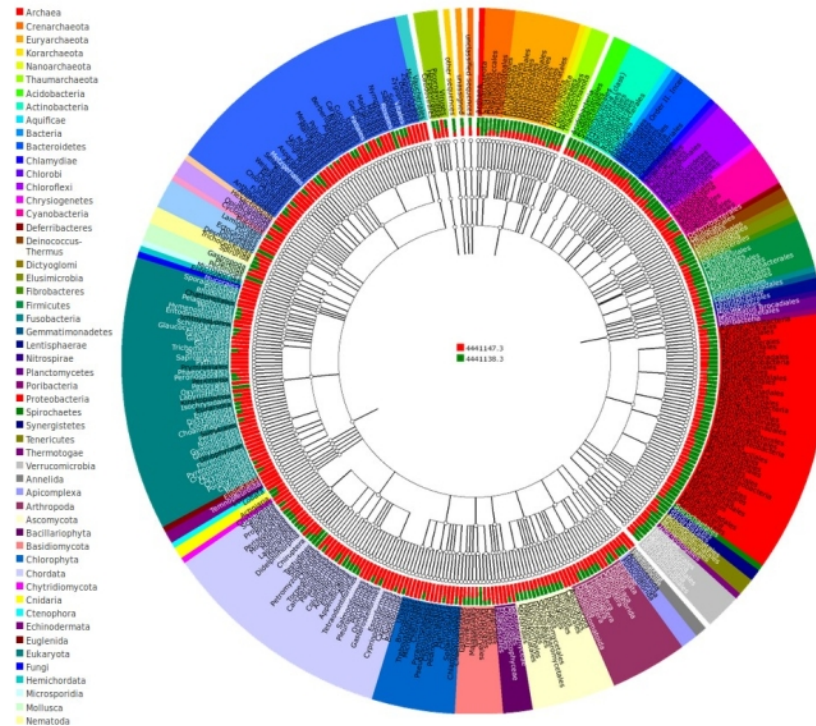
Min. % Identity Cutoff 60 %

Min. Alignment Length Cutoff 15

Workbench  use features from workbench

**3 Data Visualization**

barchart  tree  table  heatmap  PCoA  rarefaction



# (Meta) Barcoding sequencing - tools

## Ribosomal Database Project (RDP)

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



RDP Release 11, Update 5 :: September 30, 2016

3,356,809 16S rRNAs :: 125,525 Fungal 28S rRNAs  
Find out what's new in RDP Release 11.5 [here](#).



[Cite RDP's latest tool articles.](#)

RDP provides quality-controlled, aligned and annotated Bacterial and Archaeal 16S rRNA sequences, and Fungal 28S rRNA sequences, and a suite of analysis tools to the scientific community. New to RDP release 11:

- RDP tools have been updated to work with the new fungal 28S rRNA sequence collection.
- A new Fungal 28S Aligner and updated Bacterial and Archaeal 16S Aligner. We optimized the parameters for these secondary-structure based Infernal aligners to provide improved handling for partial sequences.
- Updated RDPipeline offers extended processing and analysis tools to process high-throughput sequencing data, including single-strand and paired-end reads.
- Most of the RDP tools are now available as open source packages for users to incorporate in their local workflow.

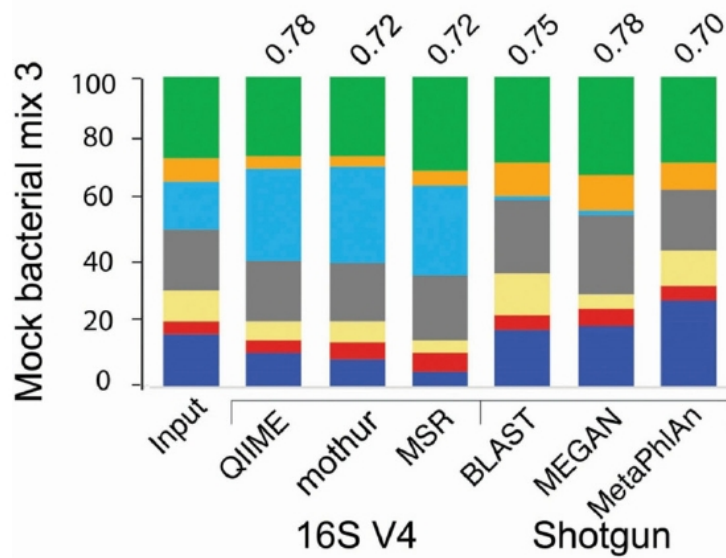
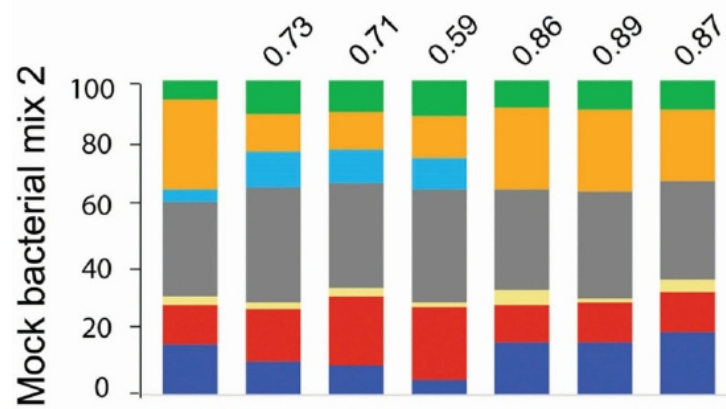
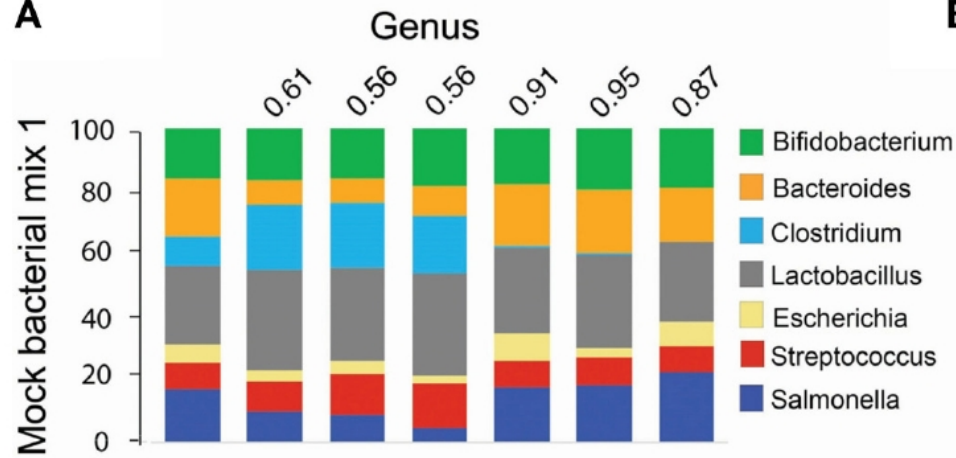


<https://rdp.cme.msu.edu/help/tutorial.jsp>

[http://rdp.cme.msu.edu/tutorials/init\\_process/RDPtutorial\\_INITIAL-PROCESS.html](http://rdp.cme.msu.edu/tutorials/init_process/RDPtutorial_INITIAL-PROCESS.html)

[https://rdp.cme.msu.edu/tutorials/classifier/classifer\\_cover\\_page.html](https://rdp.cme.msu.edu/tutorials/classifier/classifer_cover_page.html)



**A****B**