

ctctcatgcacattggattcaggtttgattagaaatccagatctggaagaggcaggatgcaactatcagtgggcaatttgagtttgc  
ggggcattcaatcaaggagcggcactattgggggttggctgcttgcactgacgagtggttgatggaatctgocaggcgggcaaggctg  
cgttcagacagcggggcggcgtctctgaaactcggcggcagccacggcggcagcaacggttgacggcggtagcgggtgaggatgctc  
cggggcattccacctaacgtgtagggcgacgagtgagtgcttaaacctgcgggaaagagctacactcaggttccaaatgcagcggc  
yatecgggtctctccacgttcacaattgggacagttatcggcgtttgtctgggtgggtgacgggctcagcaaccogaaactgac  
yactgctcgggtgcccgaactcggcctcagagcagggtcacgggaggcttctgccccgaggtaatcggaggcggctgaaagt  
ntccgggcaaccgacgtaaatctgcttgcctatgntgtacgggtatctgtaaaacggcattaaagcacacgcatgatacggcgggat  
ctgtctcgggtgtttcactagcatcaagttcaccattgaccaaccacccattcctttgacgattcgggtcacctccaatcgcaac  
aaccccgggcggggcgttctgctgctgtggcgtggccatcttaccgctaccagctctcggaaacgggaagtgttggctgactcg  
ctctctatogagacgattcctcgggttagatccogtgcgggcagcggcttgcgggacacagagcgttccgggtagagtaaccgta  
acatcagctgctcctgctgtagatcttagaggatoccccccctgcaattggcgttgggttcttgatgcaattatgagacaaccag  
gtacgtggccgaaacagagacgaaacgggtagactcggcgtcaatttagtgaagggactcagctcaccactgccaaccgggttggc  
gtaagcctggacggggcattccacggttccggcgggagcgggtgcccggcgtgtctctcagggccggagacggctggcgtagcgg  
tgaggaaatttgcgactgcccgtagatgtaccaggatgcccgtgctgagtagcttcatcttgataattggtaacggctgcttga  
ygtccggggcactccacttatacttccatgaggatgcagcaacttctctctaaatgaaattacatggggcgaatgtaccgggtt  
ggggggggagggggggtgctataatacagctgctcggctcggctgactgataaggatccacataagctctgtaacgtaacga  
ggcacaacagctagatagccgtctccggctaaagccctctcgtctgtgcccggctgtctcggcactgtacggctggctggctg  
pacatccaccgggtgaggtagtaagtaaacgtaacgtagaacggcgcaaccacactatgcccggcgtgataccctccctacg  
ggcgtctggcggtagatctctgataatcttggctatgggtggatttgggattaattctgagaaaattcgtgttgacttccag  
caactgtctgctgctcactetaaggatccacgtacgacgaggtagggaaattnaacacataccttactgggtatacgggtgta  
cggctcatggtnctagtgagtc

# Metagenomics and the NGS technology

## Francisco Rodriguez-Valera



**1: Introduction to Metagenomics (The era of cloning)**

**2: The advent of NGS, is there a role for clone libraries?**

**4 : Screening for genes: activity vs sequence**

**5: What the future may bring**

# The pure culture



Louis Pasteur



Robert Koch

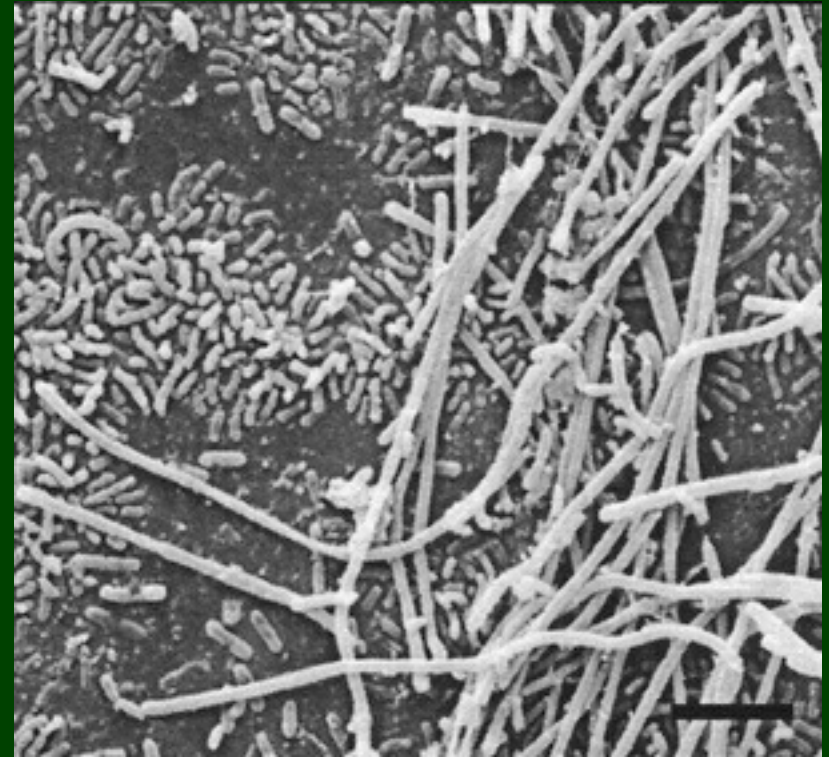
Study of large populations of microorganisms that reproduce clonally from a single cell

# Culture is not enough

- Most prokaryotes are extremely difficult to retrieve in pure culture
- Even if you get them in pure culture you might not be able to perform experiments with them (physiology)
- The genome of one strain may not represent the genetic repertoire of the “species” in the community PAN-GENOME
- Many microbes require close interaction with others to show their abilities

# Genomics and Metagenomics

**isolate** **community**



**Genomics**

**Sequencing  
analysis**

Not only  
**16S rRNA!**

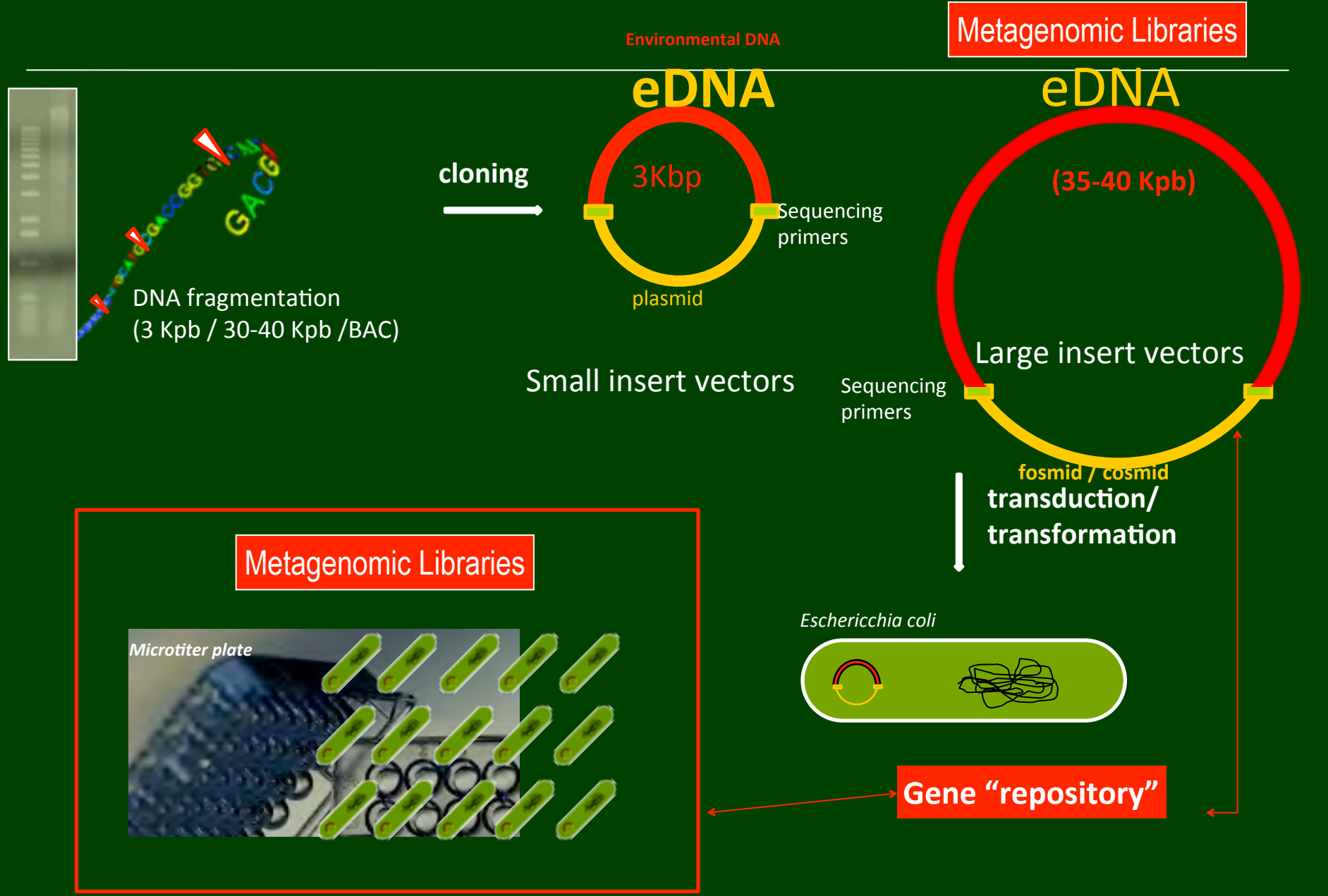


**Metagenomics**

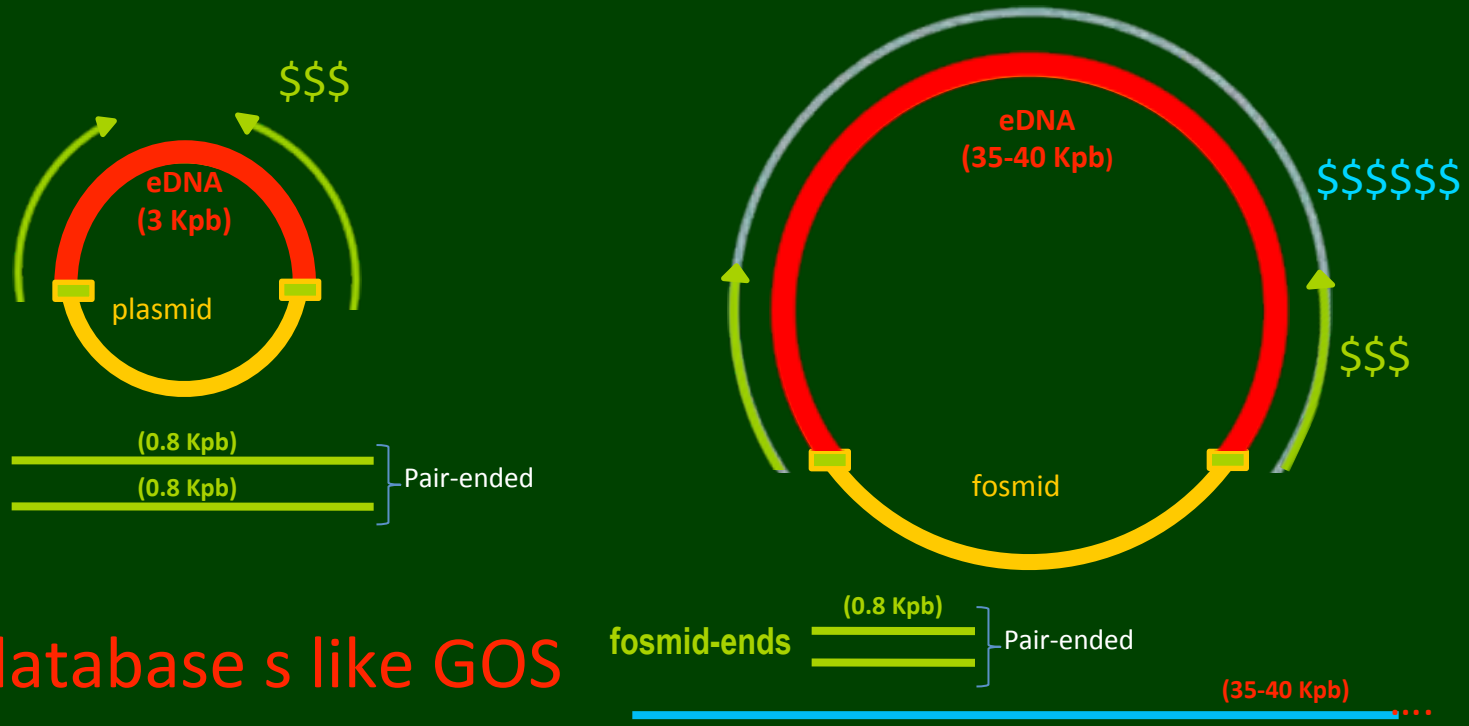
# Applications

- Exploration and conservation
- Metabiogeochemistry
- Systematics
- Population genomics and evolution
- Biotechnology

# CLASSIC METAGENOMICS: AMPLIFICATION BY CLONING



# Sanger Sequencing



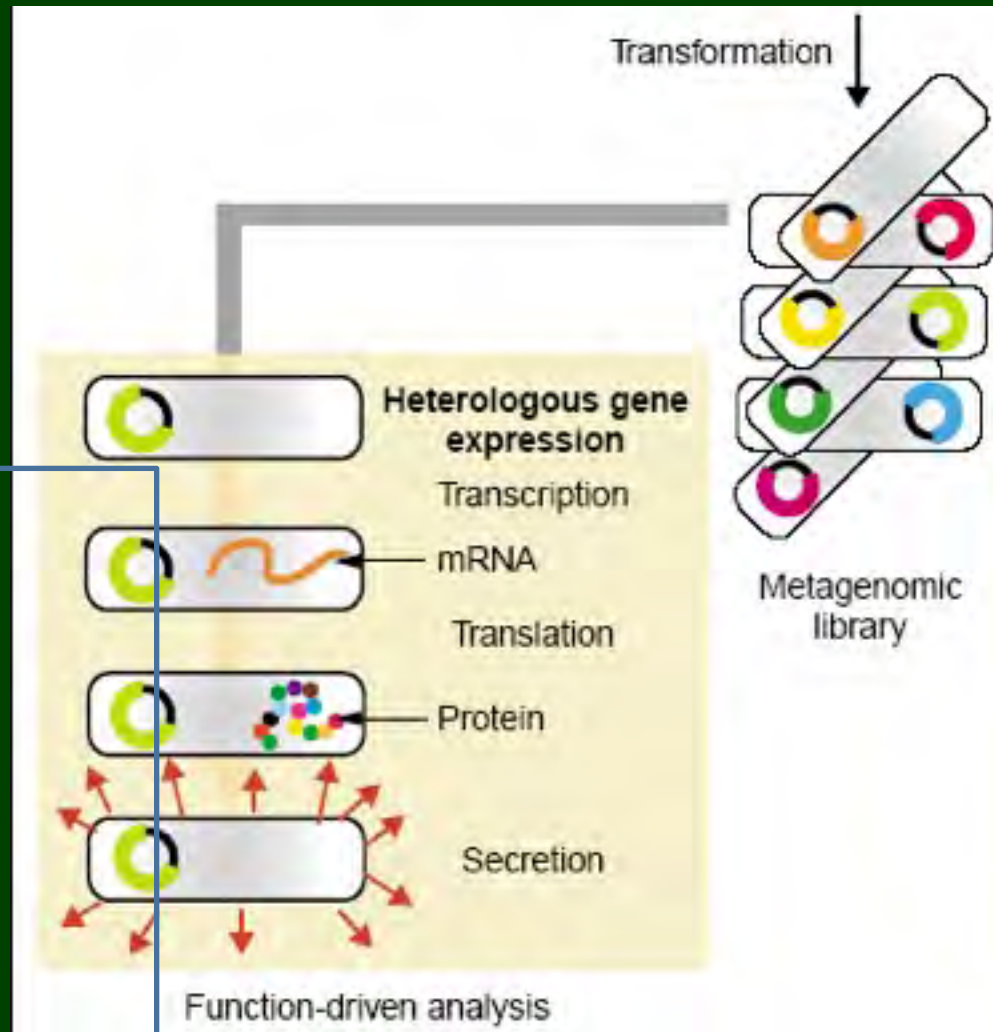
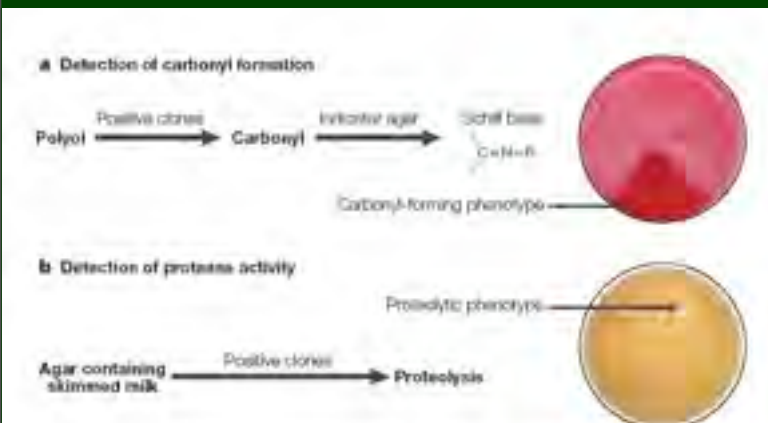
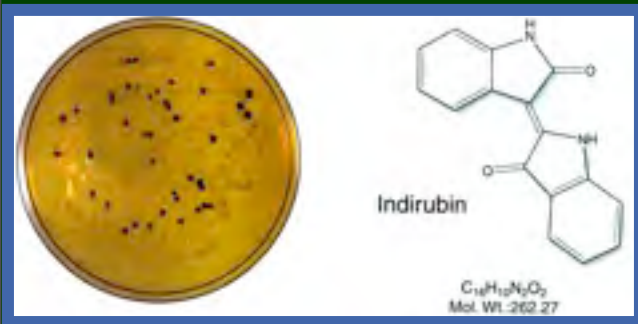
Large databases like GOS

Interesting fosmids can be fully sequenced!! but

Interesting genes are at the end → incomplete operons sometimes

# eDNA can be screen for interesting phenotypes

## Function-driven analysis



# Biological activity screening

Gillespie *et al* (2002). **Applied Environmental Microbiology**, 68: 4301-4306.

Isolation of **Antibiotics** Turbomycin A and B from a Metagenomic Library of Soil Microbial DNA

Kneietsch *et al* (2004) **AEM** 69:3048. **Dehydratases**

Kneietsch *et al* (2003) **J Mol Micro Biotech.** 5:46. **Alcohol dehydrogeases**

Yun *et al* (2004) **AEM** 70:7229. **Amylases**

Rhee *et al* (2005) **AEM** 71:817. **Thermophilic esterase**

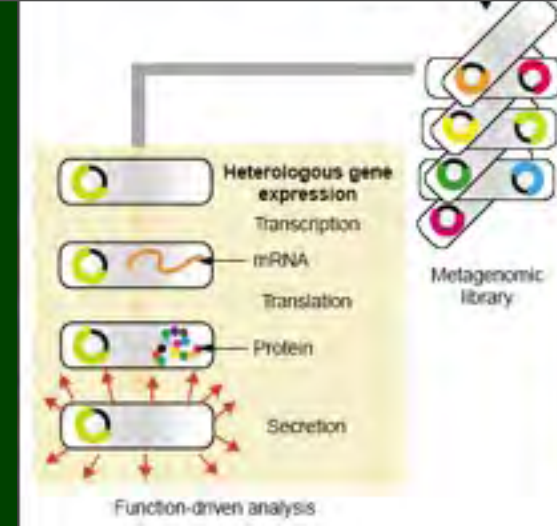
Song *et al* (2005) **J Microbiol** 43:172. **b-lactamase** in deep mud off New Guinea.

.....

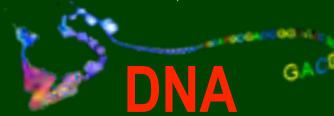
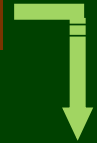
Walter *et al* (2005) **AEM** 71:2347. **b-glucanase** from mouse guts

Bunterngsook B. *et al* 2010. **Lipolytic enzymes** from a peat-swamp forest soil. **Biosci Biotechnol Biochem.**

González-Pastor JE *et al* (2010). **Metal resistance genes.** **Methods Mol Biol.**



# METAGENOMICS : Sanger Sequencing



**Cloning: Small insert vector (ca 3 Kbp)**

**Cloning: long insert vector (e.g. fosmids, ca. 35-40 Kbp eDNA)**

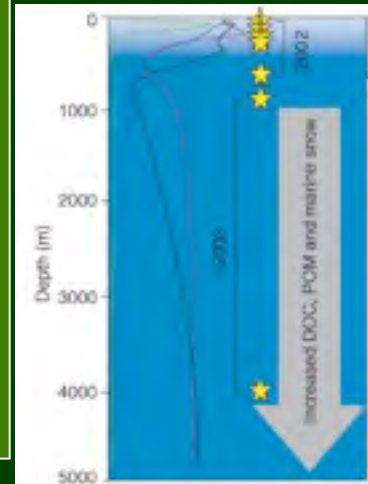


e.g. HOTs  
(Hawaii Ocean Time-Series)

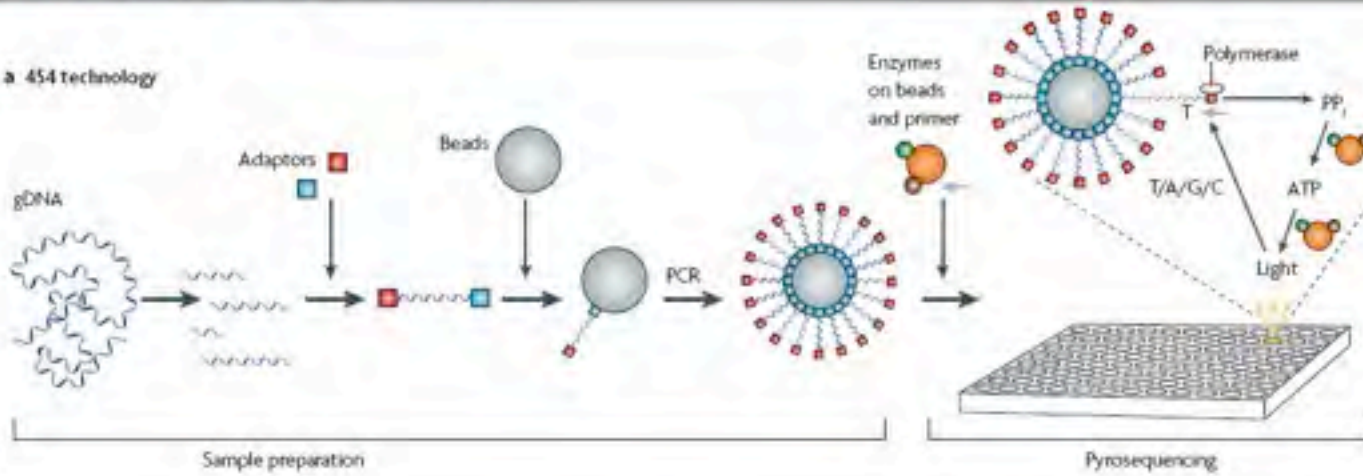
e.g. GOS  
(Global Ocean Sampling)

- Large libraries easy to generate
- Natural contigs of ca: 3Kbp (pair ended)
- Annotation of single genes (unreliable)
- Phenotype can be detected (very unlikely)

- Large libraries more difficult
- Natural contigs of ca: 35 Kbp (pair ended)
- Complete sequence allows annotation of clusters of genes (very reliable)
- PCR or fosmid end screening
- Phenotype can be detected (unlikely)

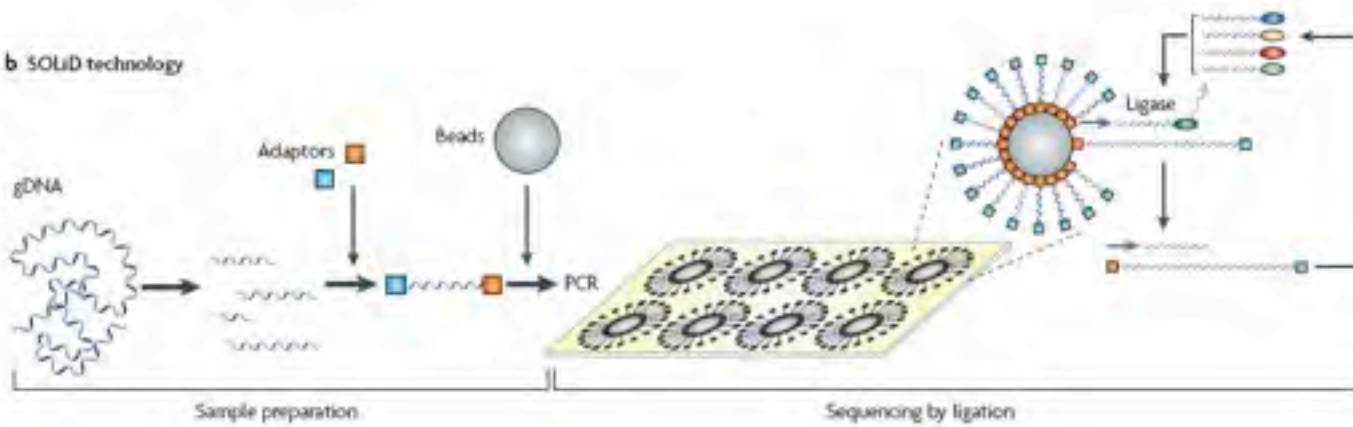


a 454 technology



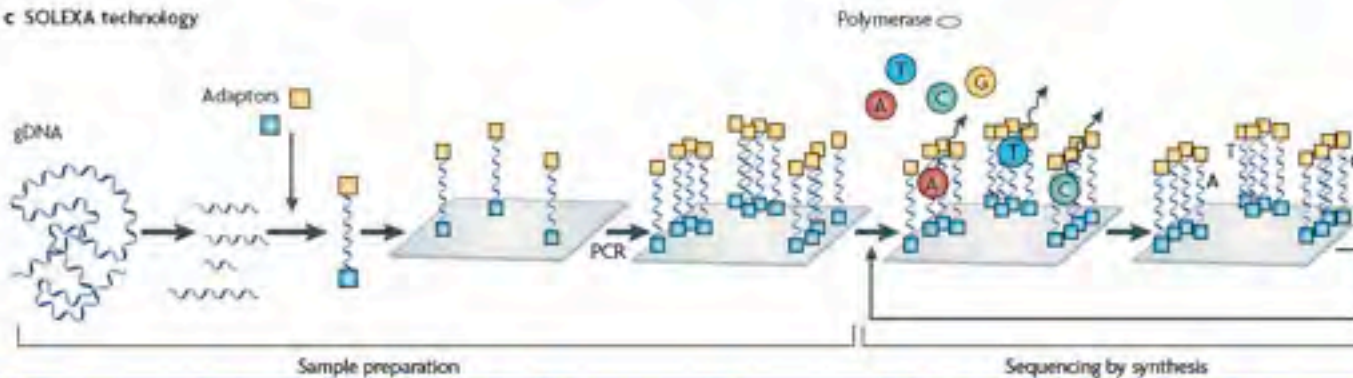
AMPLIFICATON  
emPCR

b SOLiD technology



**NGS (New Generation Sequencing)**  
Originally, Next Generation Sequencing but actually 2nd generation, we are now on the brink of the 3rd

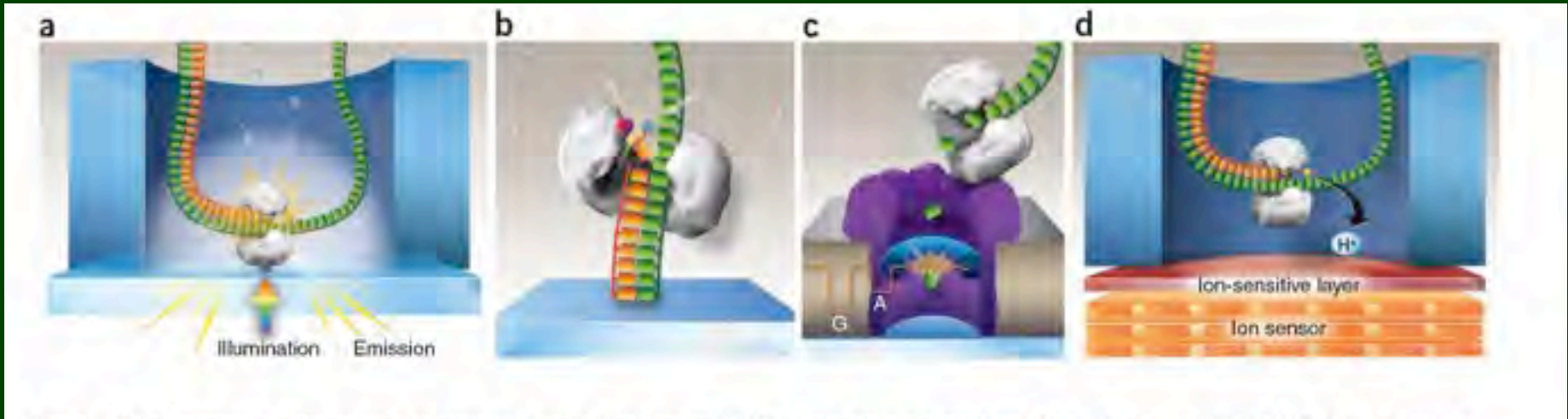
c SOLEXA technology



Array PCR

# Single Molecule Sequencing

Munroe and Harris, Nature Biotechnology, 28: 226 (2010)



Pac Bio

Helicos

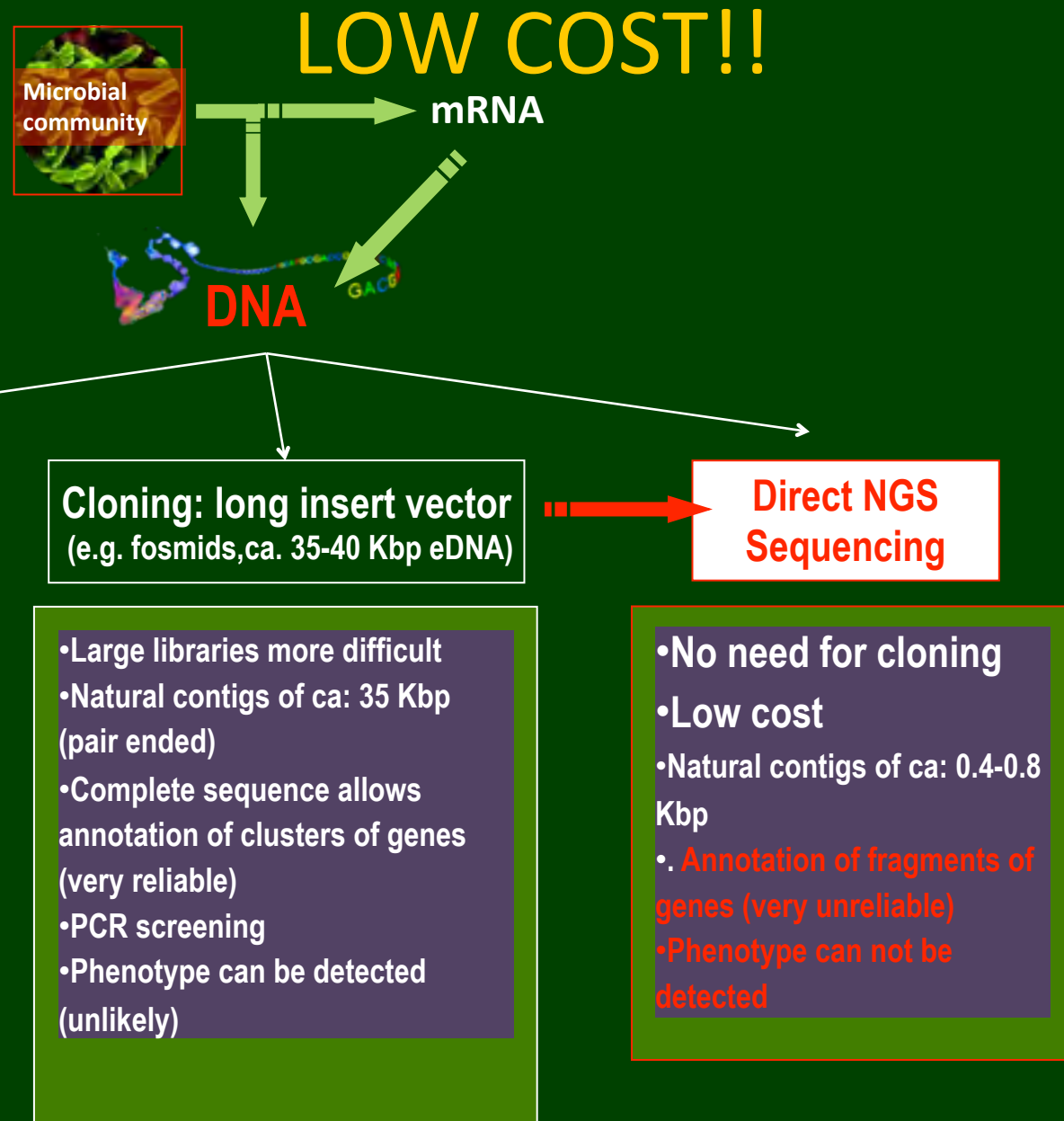
Nanopore

Ion Torrent

**NO AMPLIFICATION**

Gigantic technological drive for the individual human genome

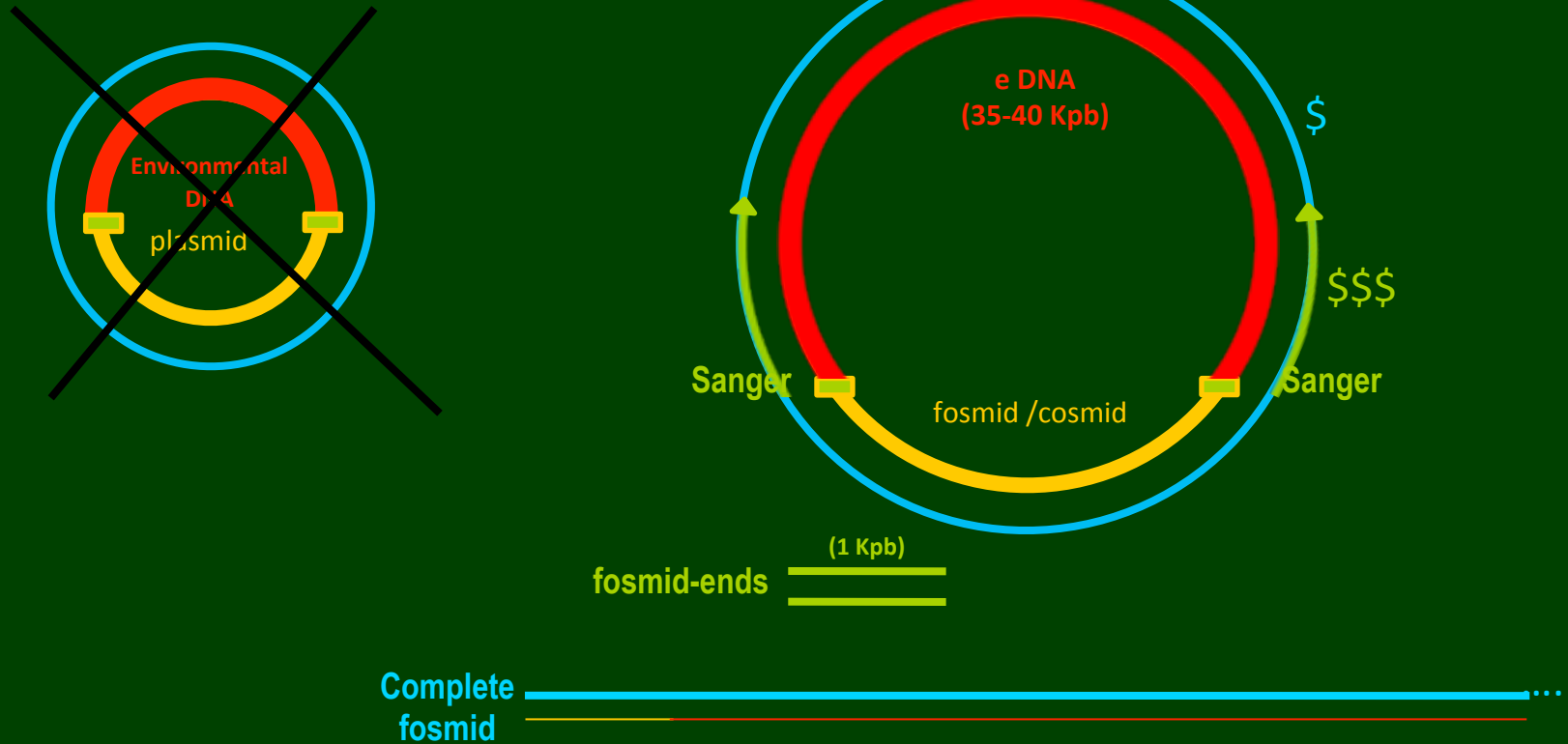
# METAGENOMICS BY NGS: NO NEED FOR CLONING



# NGS METAGENOMICS

- Straight forward simple and cheap
- Large volume of sequence (400 Mbp by Titanium 454 pyrosequencing), 10 Gb Solexa
- Thanks to the high coverage assembly of large fragments is feasible → Annotation reliable
- Large insert libraries can be sequenced by NGS
- Sequence driven search for activities

# Metagenomic Libraries NGS sequencing



Interesting fosmids can be fully sequenced!!

# NG-Sequencing

Examples  
Fosmids by Pyrosequencing

Metagenomic Libraries

ISME J, 2010

The metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing

Rohit Ghai<sup>1</sup>, Ana-Belen Martin-Cuadrado<sup>1</sup>, Aitor Gonzaga Molin<sup>1</sup>, Inmaculada Garcia-Heras<sup>1</sup>, Raul Cabrera<sup>1</sup>, Javier Martin<sup>1</sup>, Miguel Virdo<sup>1</sup>, Philipp Duschamps<sup>1</sup>, David Moreira<sup>2</sup>, Purificación López-García<sup>1</sup>, Alex Mira<sup>3</sup> and Francisco Rodriguez-Valera<sup>1</sup>

Comparison of “bulk DNA” (1 “run”) vs fosmid library (12 Elisa plates)



Clonation bias

PLOS one, 2007

Metagenomics of the Deep Mediterranean, a Warm Bathypelagic Habitat

Ana-Belen Martin-Cuadrado<sup>1</sup>\*, Purificación López-García<sup>2</sup>\*, Juan-Carlos Alba<sup>1</sup>, David Moreira<sup>3</sup>, Luis Montecchi<sup>4</sup>, Axel Strömstedt<sup>5</sup>, Gerhard Gomschall<sup>6</sup>, Francisco Rodriguez-Valera<sup>1</sup>\*

Fosmid-end random sequencing of KM3 metagenomic library Sanger

AEM, 2009

CODH

CO Dehydrogenase Genes Found in Metagenomic Fosmid Clones from the Deep Mediterranean Sea<sup>†</sup>

Ana-Belen Martin-Cuadrado, Rohit Ghai, Aitor Gonzaga, and Francisco Rodriguez-Valera\*



Interesting genes at the end, incomplete operons sometimes

ISME J, 2008

Hindsight in the relative abundance, metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions

Ana-Belen Martin-Cuadrado<sup>1</sup>, Francisco Rodriguez-Valera<sup>1</sup>, David Moreira<sup>2</sup>, José C. Alba<sup>1</sup>, Gloria Ivorra-Martinez<sup>3</sup>, Matthew R. Heupel<sup>4</sup>, Emmanuel Talla<sup>5</sup> and Purificación López-García<sup>1</sup>



Big genomic fragments from unknown species

Env. Micro, 2008

Comparative analysis of genome fragments of Acidobacteria from deep Mediterranean plankton.

Quaiser A, López-García P, Zivanovic Y, Henn MR, Rodriguez-Valera F, Moreira D.

On-going: Around 600 fosmids of marine Archaea

MAGYK

The ISME Journal (2010), 1–13  
© 2010 International Society for Microbial Ecology All rights reserved 1751-7362/10 \$32.00  
[www.nature.com/ismej](http://www.nature.com/ismej)



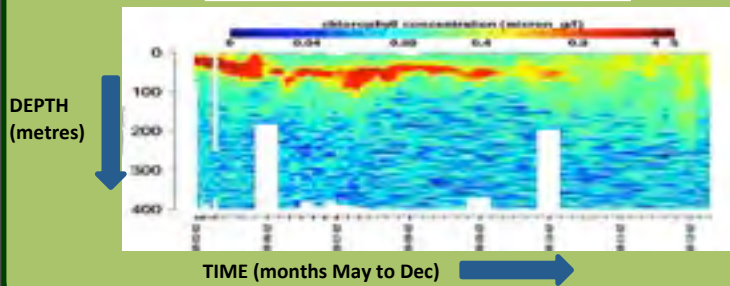
## ORIGINAL ARTICLE

# Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing

Rohit Ghai<sup>1</sup>, Ana-Belén Martín-Cuadrado<sup>1</sup>, Aitor Gonzaga Molto<sup>1</sup>, Inmaculada García Heredia<sup>1</sup>, Raúl Cabrera<sup>2</sup>, Javier Martín<sup>3</sup>, Miguel Verdú<sup>3</sup>, Philippe Deschamps<sup>4</sup>, David Moreira<sup>4</sup>, Purificación López-García<sup>4</sup>, Alex Mira<sup>2</sup> and Francisco Rodríguez-Valera<sup>1</sup>

<sup>1</sup>*Departamento de Producción Vegetal y Microbiología, Evolutionary Genomics Group, Universidad Miguel Hernández, Apartado 18 San Juan de Alicante, Alicante, Spain;* <sup>2</sup>*Department of Genomics and Health, Centro Superior de Investigación en Salud Pública, Avda. Cataluña 21, Valencia, Spain;* <sup>3</sup>*Mediterraneo Servicios Marinos S.L. Nueva Dársena Pesquera s/n, Alicante, Spain;* <sup>4</sup>*Unité d'Ecologie, Systématique et Evolution, CNRS UMR8079, Université Paris-Sud 11, Orsay, France*

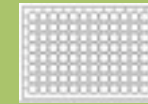
## DEEP CHLOROPHYLL MAXIMUM



FOSMIDS  
(30-40kb)

FOSMID vs DIRECT  
PYROSEQUENCING

BULK  
DNA

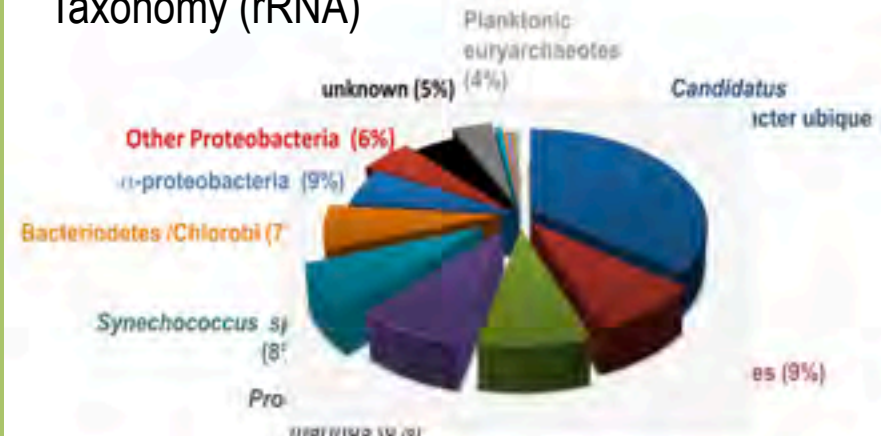


454 plate  
~**1000** fosmids  
(12 collections)

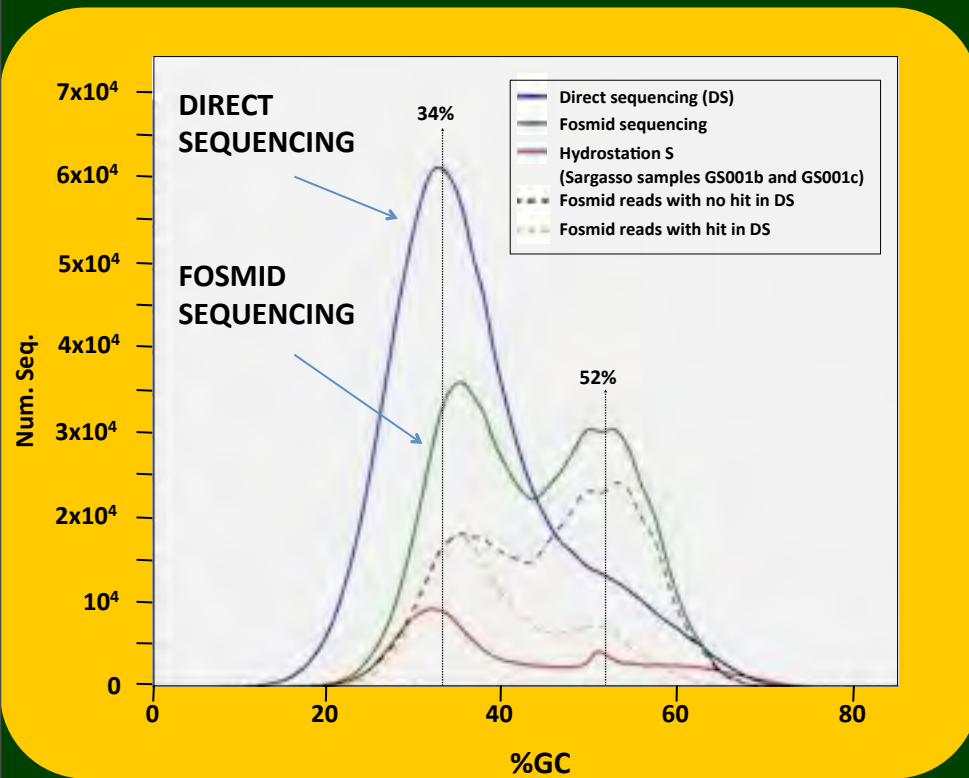
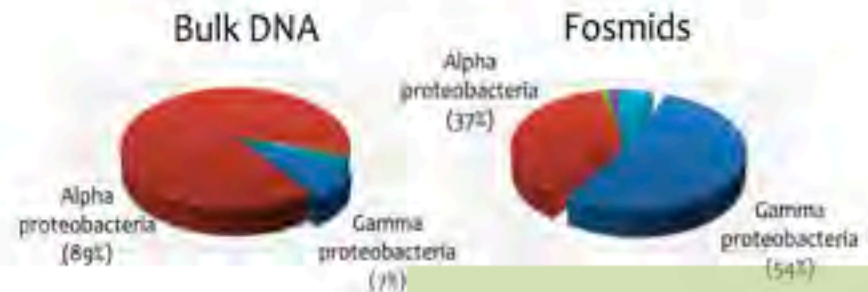


454 plate

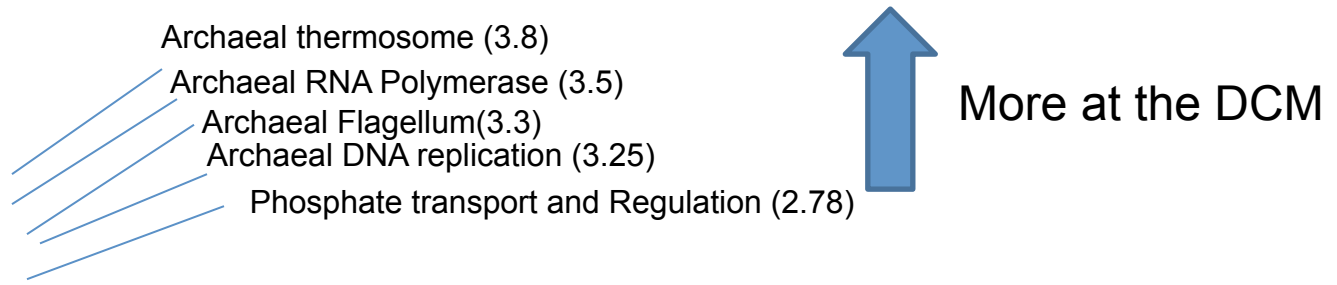
## Taxonomy (rRNA)



## Bias against alpha proteobacteria in fosmids



Relative abundance ( $\log_2$  scale)



Put names of genes in the picture

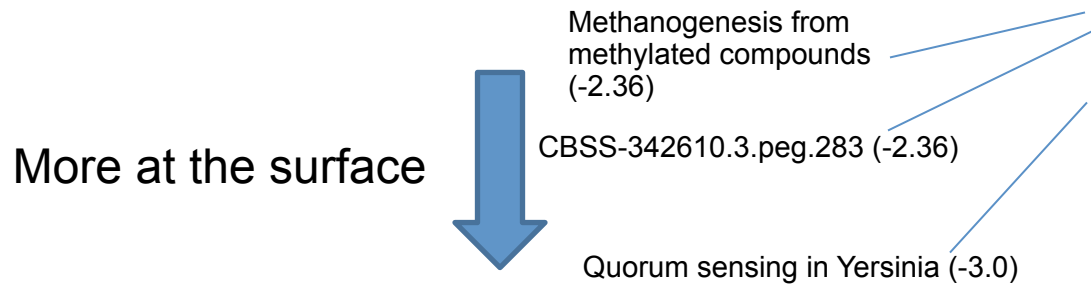


Figure 3

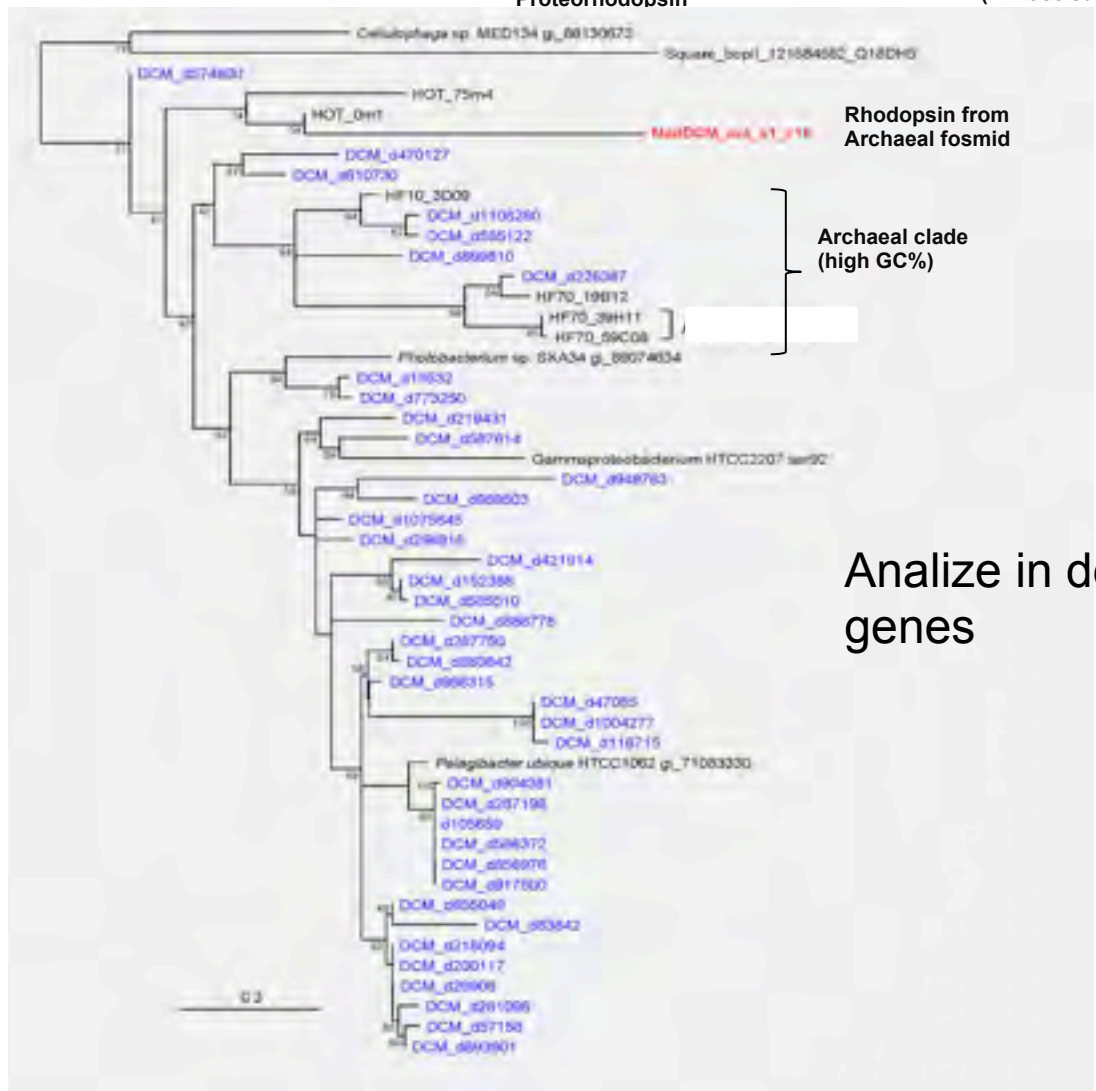
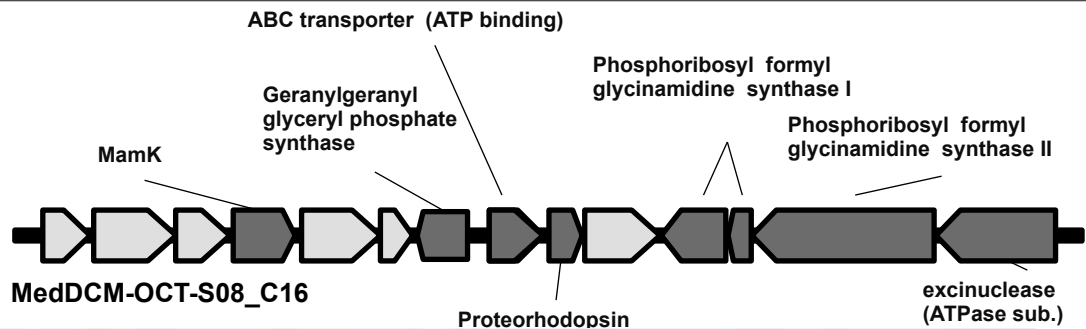
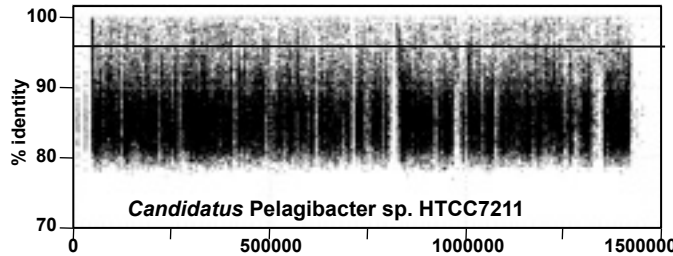
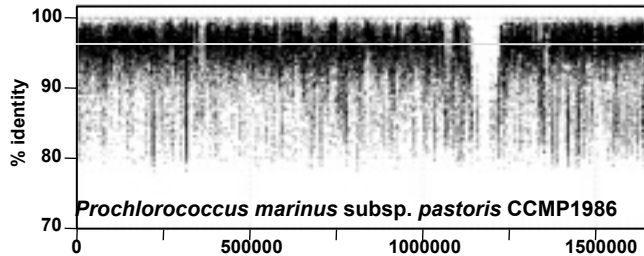


Figure 4

# Put genomes in the picture Genome recruitment

Number of hits



# Metagenomics of the Water Column in the Pristine Upper Course of the Amazon River

Rohit Ghai<sup>1,3</sup>, Francisco Rodriguez-Valera<sup>1,3</sup>, Katherine D. McMahon<sup>1</sup>, Danyelle Toyama<sup>2</sup>, Raquel Rinke<sup>2</sup>, Tereza Cristina Souza de Oliveira<sup>3</sup>, José Wagner Garcia<sup>4</sup>, Fernando Pellon de Miranda<sup>5</sup>, Flavio Henrique-Silva<sup>2\*</sup>

**1** Evolutionary Genomics Group, Departamento de Producción Vegetal y Microbiología, Universidad Miguel Hernández, San Juan de Alicante, Alicante, Spain, **2** Laboratory of Molecular Biology, Departamento de Genética e Evolução, Universidade Federal de São Carlos, São Carlos, SP, Brasil, **3** Universidade Federal do Amazonas, Setor Sul do Campus Universitário, Manaus, AM, Brasil, **4** Noosfera Projetos Especiais Ltda, São Paulo, SP, Brasil, **5** Petróleo Brasileiro S.A. – Petrobras, Centro de Pesquisas e Desenvolvimento Leopoldo Américo Miguez de Melo, Rio de Janeiro, RJ, Brasil

## Abstract

River water is a small percentage of the total freshwater on Earth but represents an essential resource for mankind. Microbes in rivers perform essential ecosystem roles including the mineralization of significant quantities of organic matter originating from terrestrial habitats. The Amazon river in particular is famous for its size and importance in the mobilization of both water and carbon out of its enormous basin. Here we present the first metagenomic study on the microbiota of this river. It presents many features in common with the other freshwater metagenome available (Lake Gatun in Panama) and much less similarity with marine samples. Among the microbial taxa found, the cosmopolitan freshwater *act* lineage of the actinobacteria was clearly dominant. Group I Crenarchaea and the freshwater sister group of the marine SAR11 clade, LD12, were found alongside more exclusive and well known freshwater taxa such as *Polynucleobacter*. A metabolism-centric analysis revealed a disproportionate representation of pathways involved in heterotrophic carbon processing, as compared to those found in marine samples. In particular, these river microbes appear to be specialized in taking up and mineralizing allochthonous carbon derived from plant material.

# The Amazon River Basin



Largest river in the world

6280 km long

15% of total runoff of all rivers in the world

Size of the South America: 17,840,000 sq. Km

Amazon Rainforest Size: 8,235,430 sq. Km

- Lowland moist forest
- Mangrove and coastal swamp forest
- Sub-montane forest
- Montane forest
- Fragmented forest
- Converted forest
- Inland water
- No data

- Savannah woodlands
- Grasslands
- Subdesertic vegetation
- Montane mosaics
- Seasonally flooded grasslands
- Agricultural mosaics
- Subdesertic vegetation

Based on the Vegetation Map of Tropical South America, H.D. Eva et al. (1999) TREES Publications Series, European Commission, with modifications by Rhett A. Butler / mongabay.com

# Sample Location

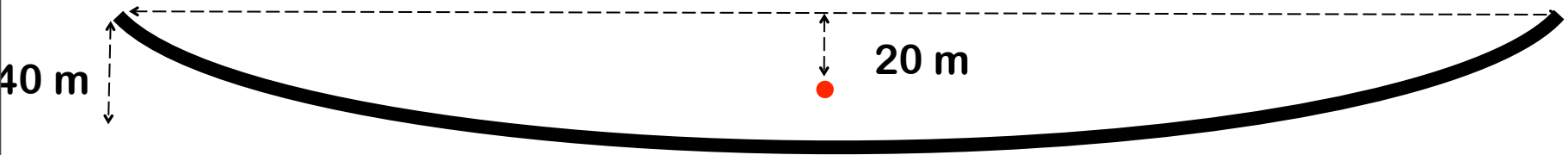


Sample Collection Depth (m)	20
Total Depth (m)	40
Temperature (°C)	29.8
pH	7.09
Oxygen (mg L <sup>-1</sup> )	5.05
Turbidity (NTU)	18.20
M susp(mg L <sup>-1</sup> )	45.67
DBP (mg L <sup>-1</sup> )	-2.90
Transparency (m)	2
Conductivity (μS cm <sup>-1</sup> )	109.7



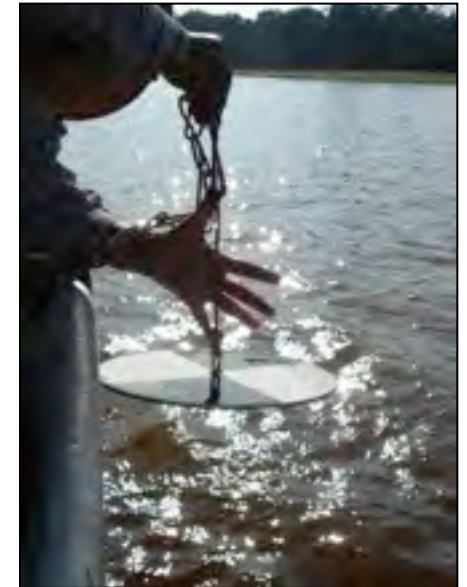
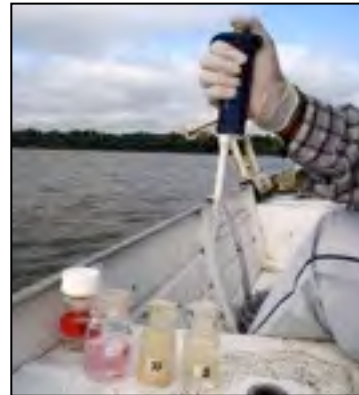
GPS Coordinates  
03° 93'151" S  
63° 18'56" W

2379 metres





## COLETA

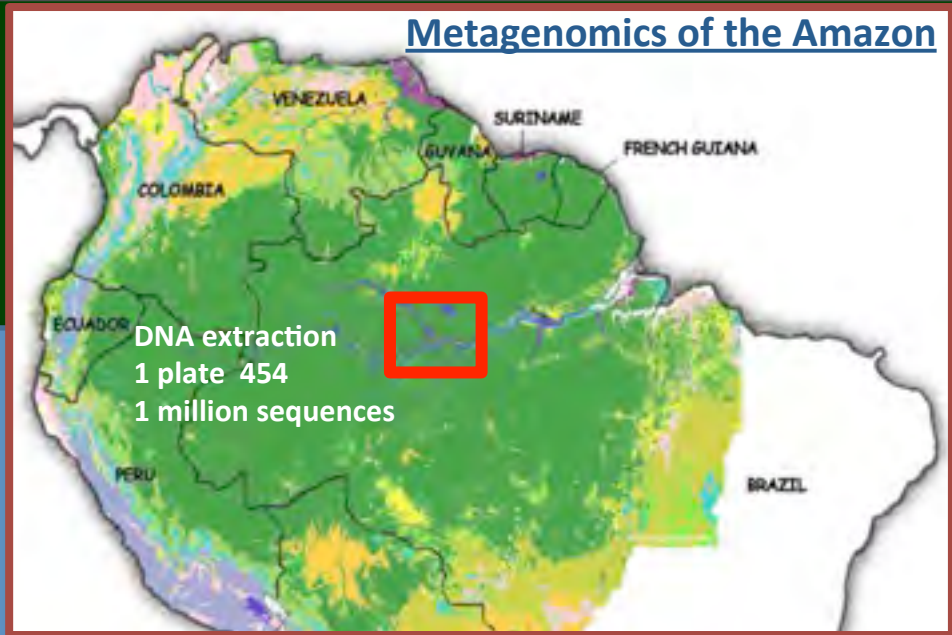




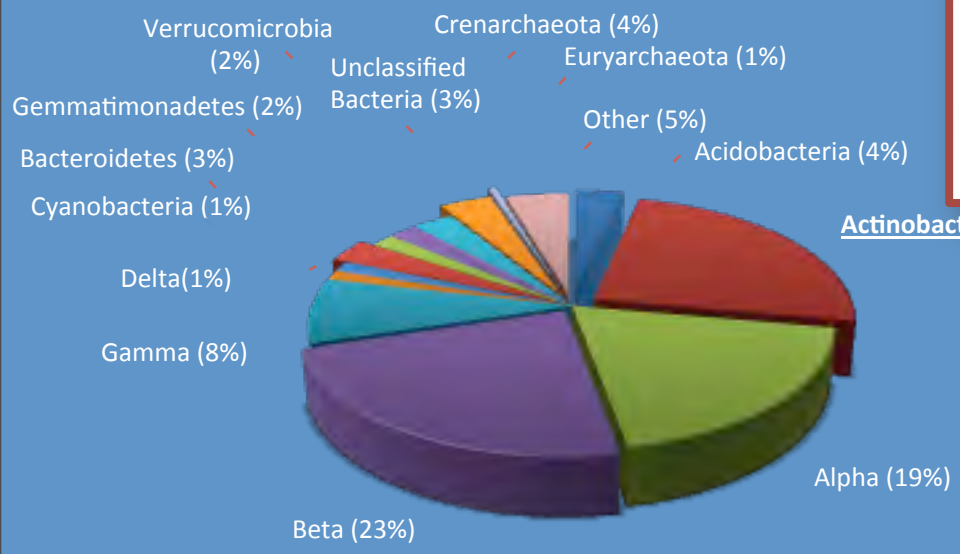
## FILTRAÇÃO

# Metagenomics of the Amazon

- Largest river in the world, 6280 km long
- 15% of total runoff of all rivers in the world
- Size of South America: 17,840,000 sq. Km
- Amazon Rainforest Size: 8,235,430 sq. Km



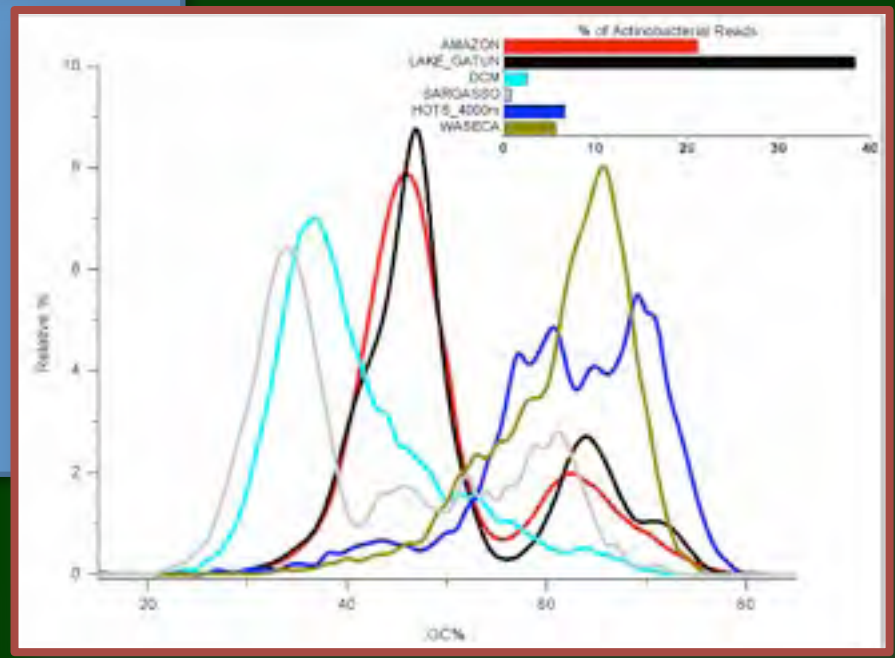
## rRNA Taxonomy



Actinobacteria (24%)

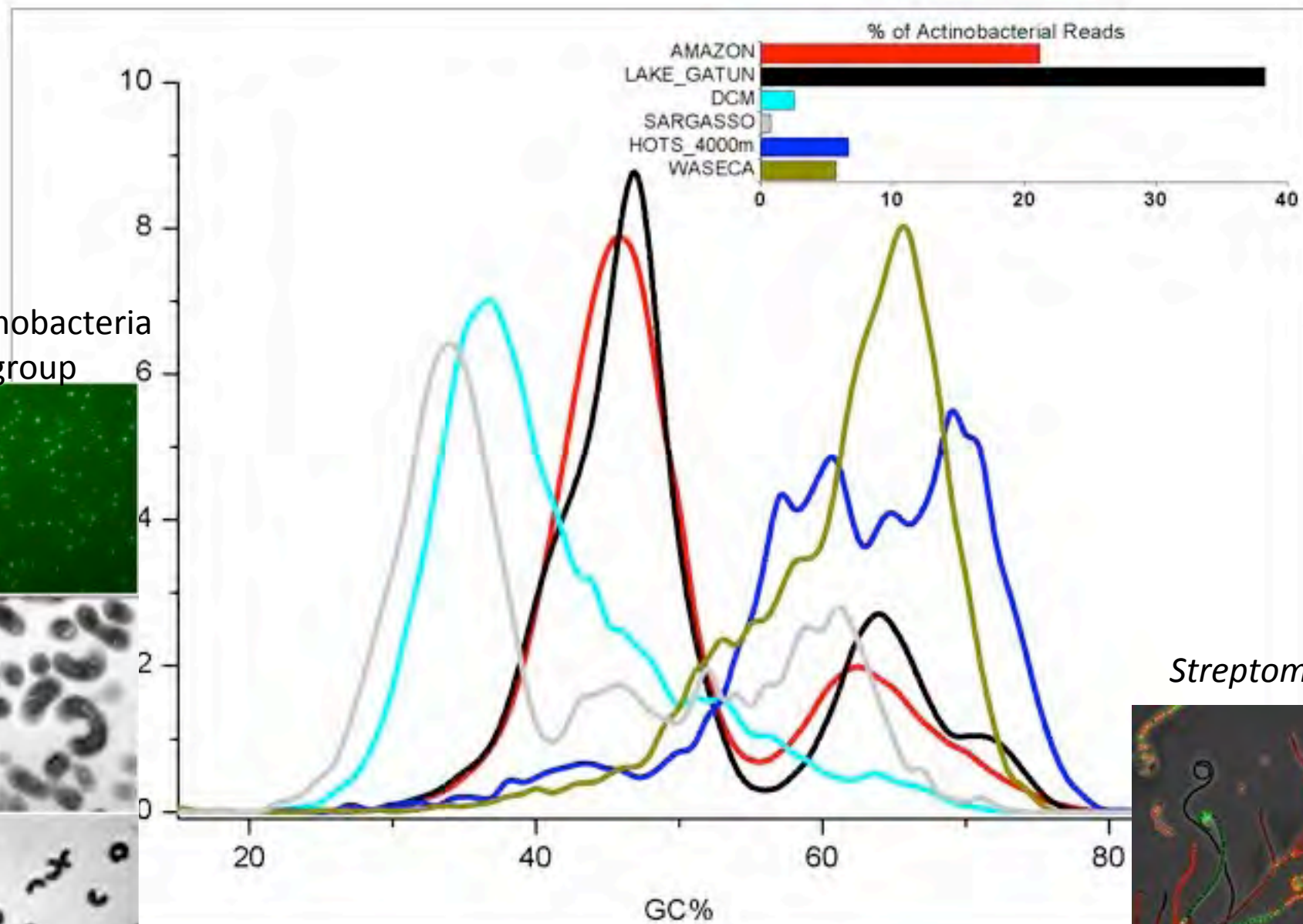
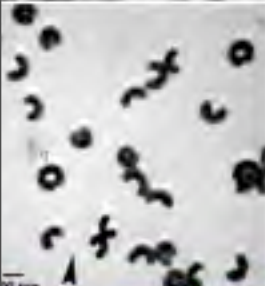
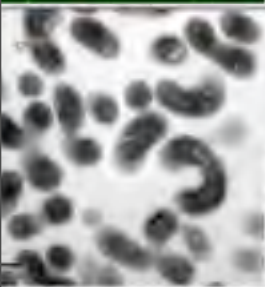
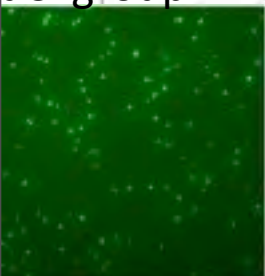
Dominance of Betaproteobacteria(*Polynucleobacter*), Freshwater Alphaproteobacteria (LD12, SAR11 -related) and Freshwater Actinobacteria (acl lineage)

Freshwater relative of ammonia-oxidizing crenarchaeon *Nitrosopumilus maritimus* ( synteny in amoABC gene cluster)

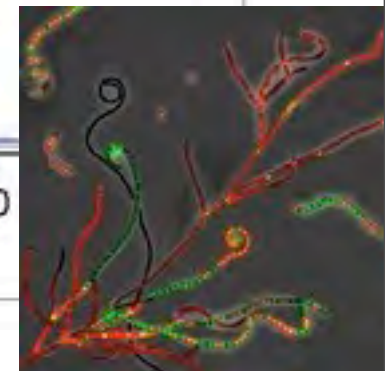


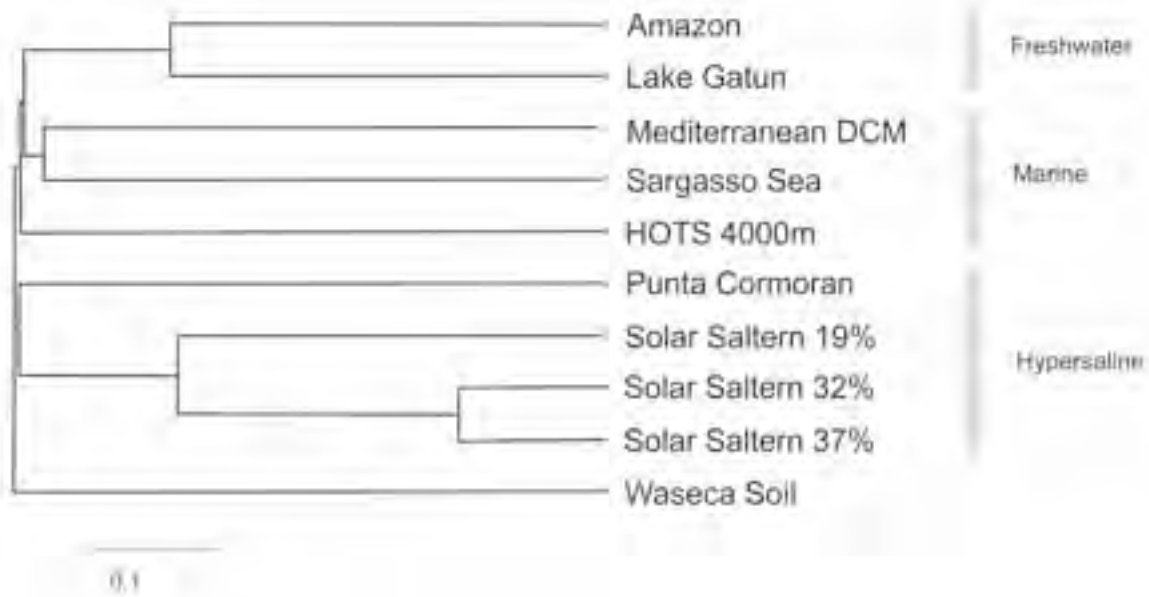
Low GC Actinobacteria

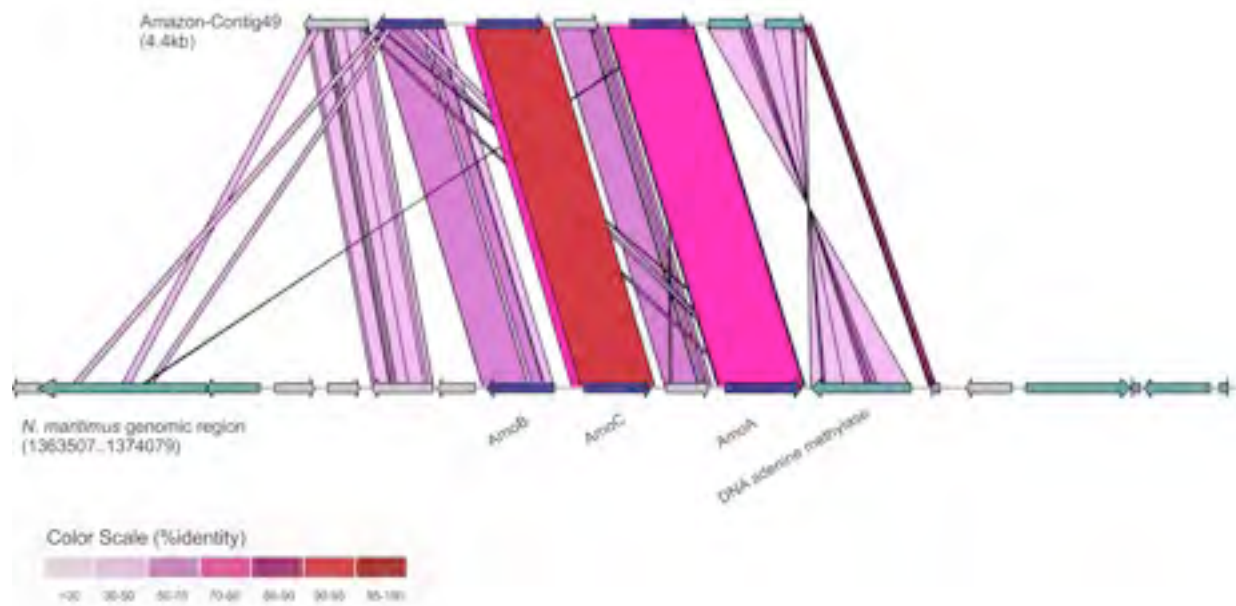
Actinobacteria  
aCl group



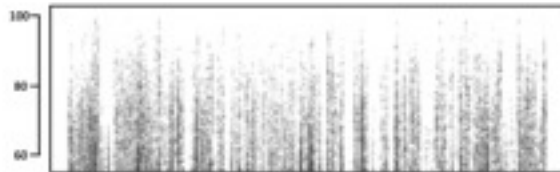
*Streptomyces*



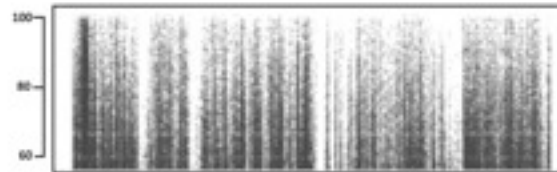




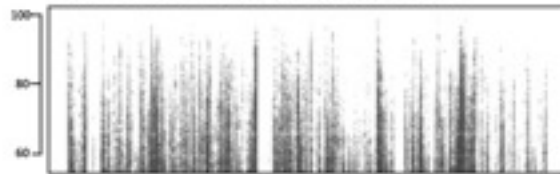
*Nitrosopumilus maritimus*



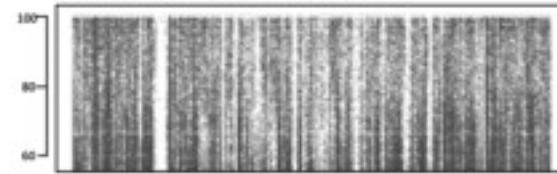
*Polynucleobacter necessarius*QLW-P1DMWA



Candidatus *Pelagibacter ubique* HTCC1062



*Acinetobacter baumannii*ATCC 17978



# Assembly

Metagenomic Libraries



Seqman – DNASTAR (Sanger, 454 )



CLC genomics (454, Solexa, both), Java \* Trial Version

*“Sequences –Group 5”*



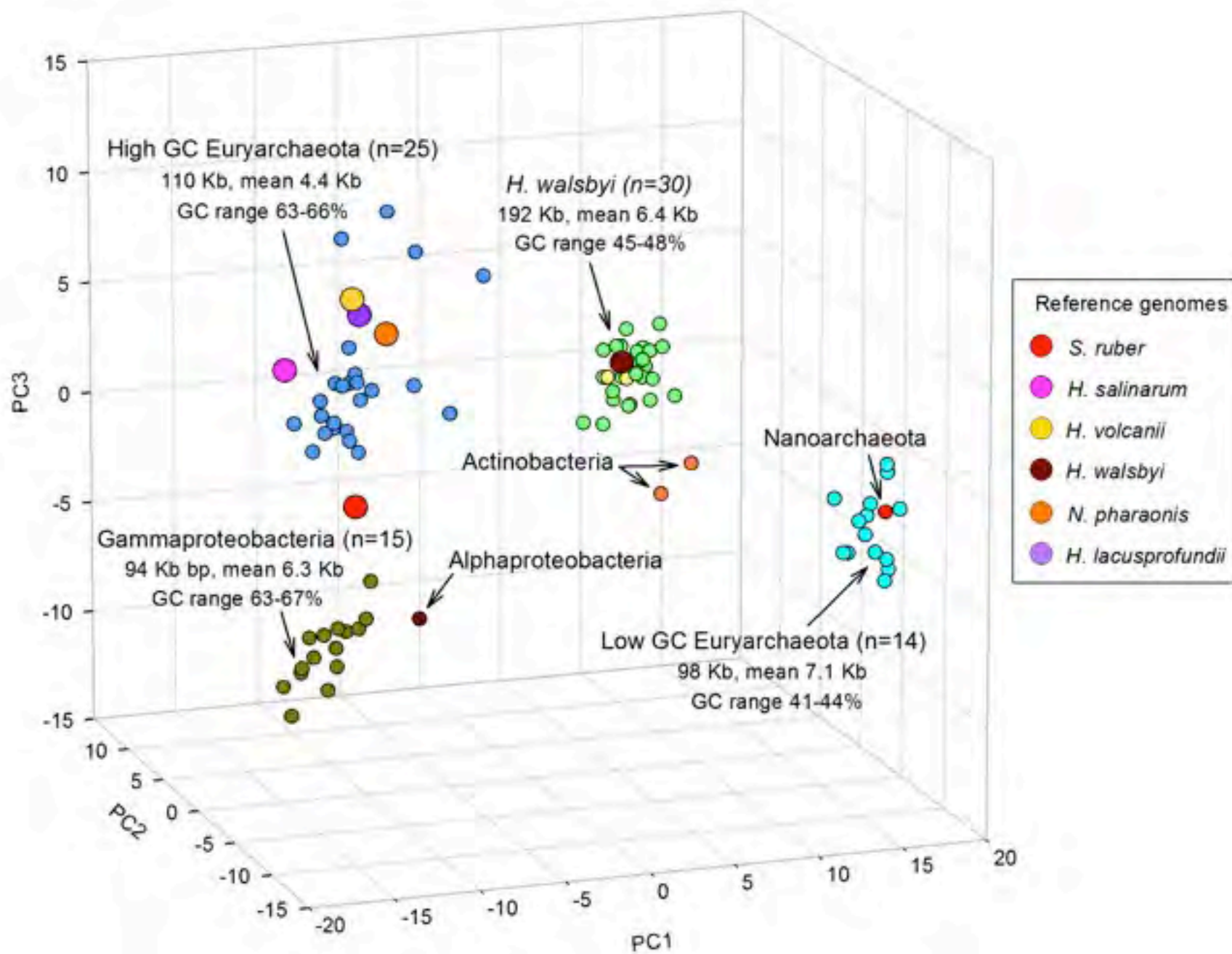
Genious (454, Solexa, both), Java

MIRA (Sanger, 454, Solexa, PacBIO ) Linux\*

... much more

Practical Exercise

Figure 5

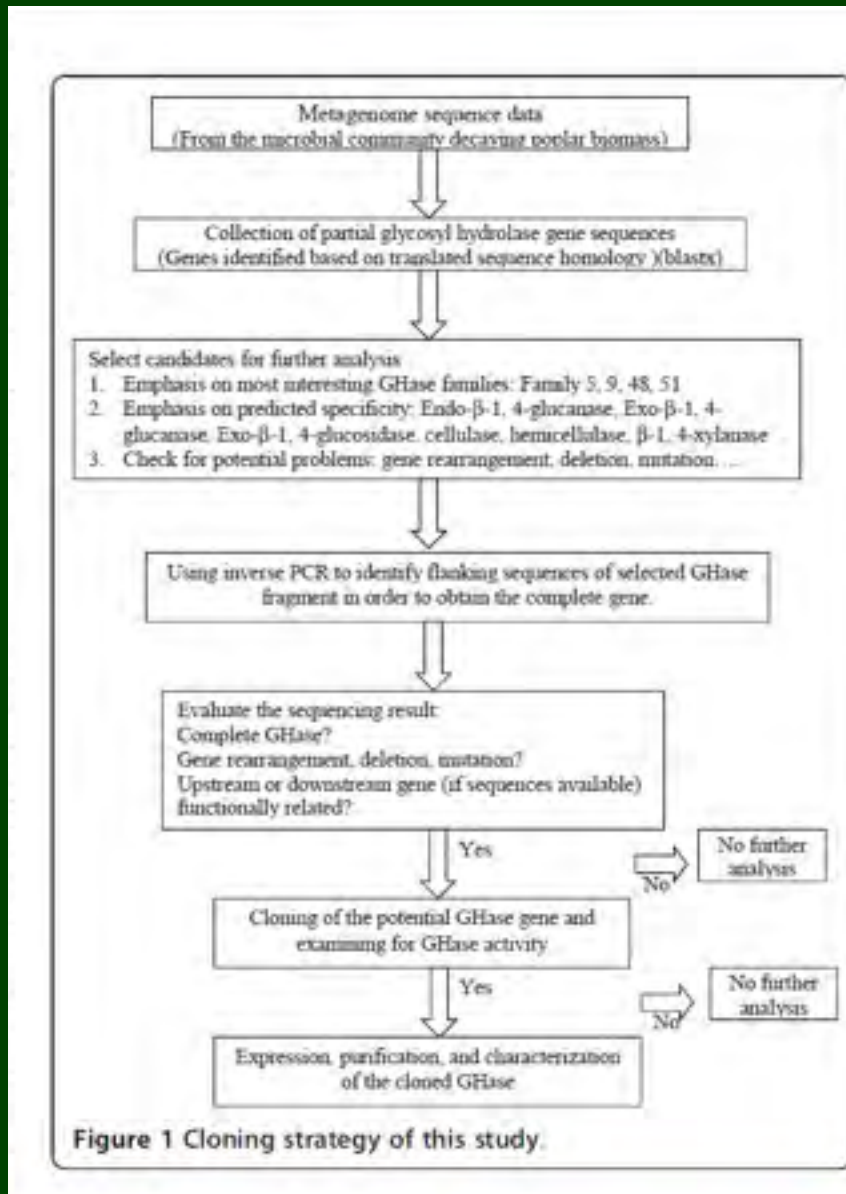


# What about screening for useful genes ?

- From sequence to function
  - Screen bulk sequences for tell-tale domains
  - Synthetic DNA from eDNA seq
  - Clone in adequate host



# Bioprospecting metagenomes for glycoside hydrolases



**DCM** 50m, Mediterranean sea  
12,192 clones

**AD** 1000m, Adriatic sea  
14,491 fosmid-end

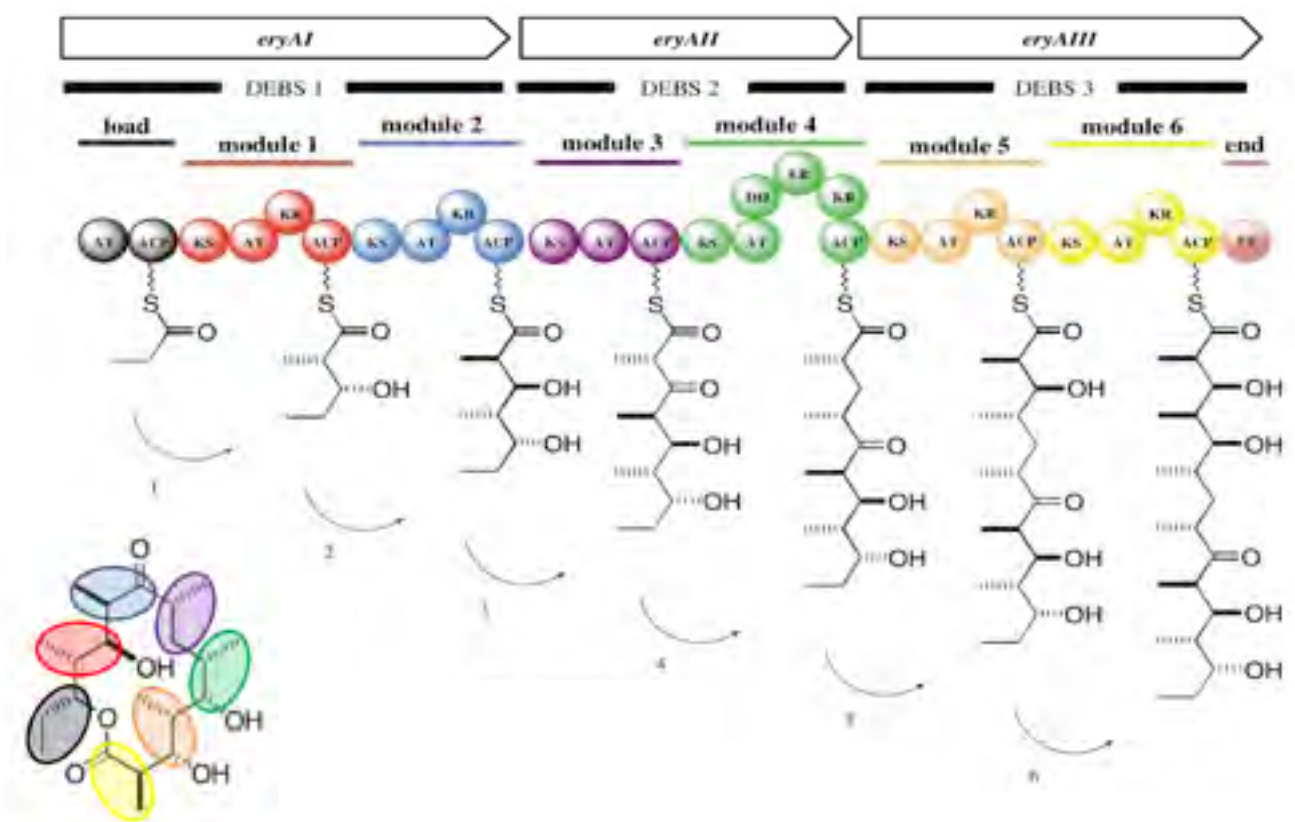
**KM3** 3000m, Ionian Sea  
15,468 fosmid-end

**IN SILICO  
SCREENING**

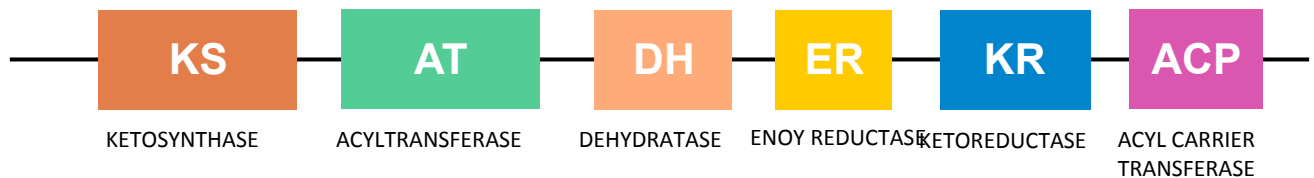
**706 PKS-related  
domains identified**

**110 FOSMIDS  
SELECTED FOR  
SEQUENCING**

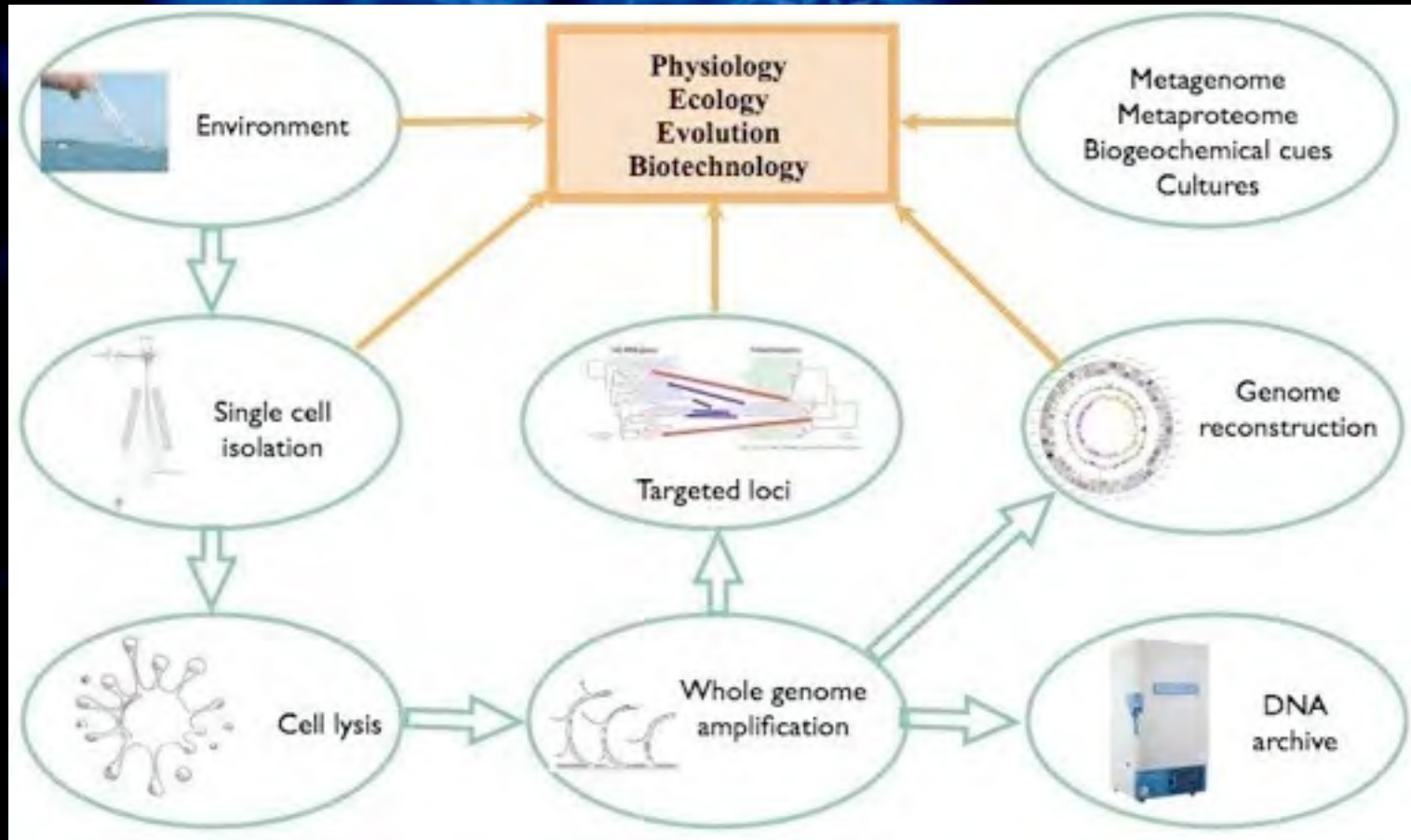
## ERYTHROMYCIN BIOSYNTHESIS – PKS GENE CLUSTER AND DOMAINS



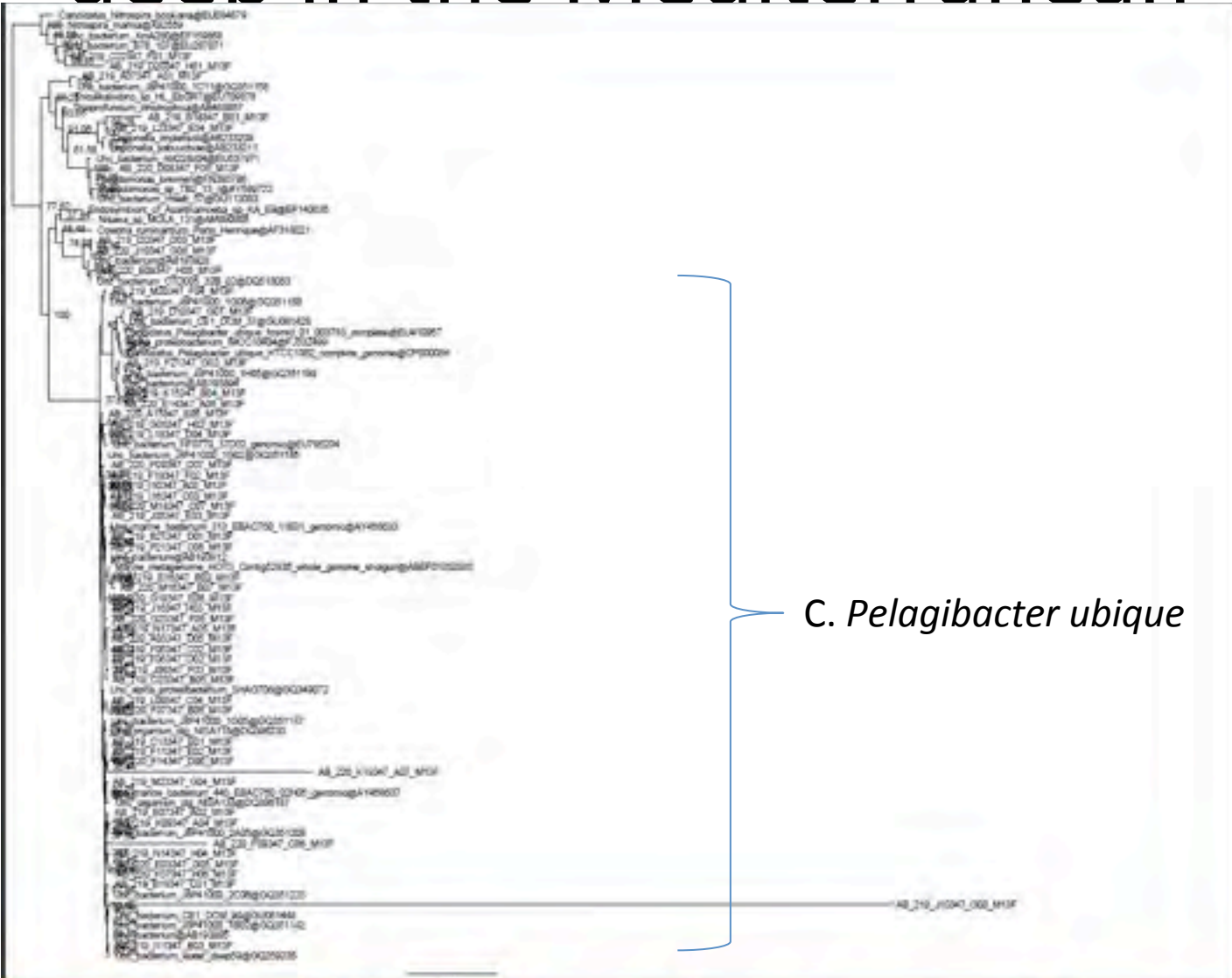
## PKS RELATED DOMAINS



# Single-cell genomics



# Single-cell 16S rRNA seqs. from 3000 m deep in the Mediterranean

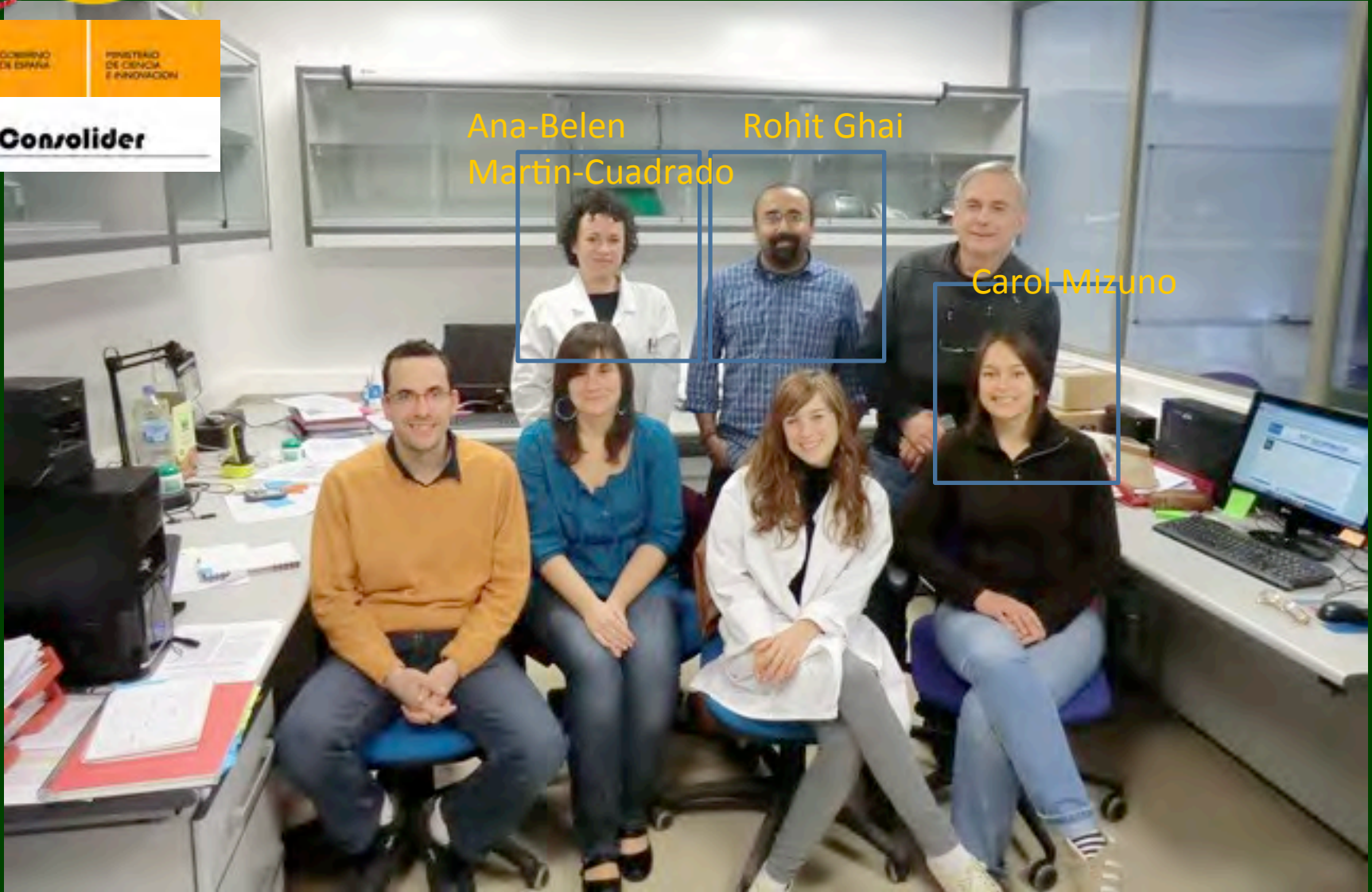


# Outlook for the next 10 years

- Human, farm animals and Earth microbiomes catalogued → In-depth exploitation of microbial diversity
- Sequence analysis (annotation) limiting step
- Sequence driven screening for useful metabolic pathways (PKs, NRP etc), enzymes, new antimicrobials, new probiotics → Huge opportunities for biotech → Better health



# Acknowledgments



Ana-Belen

Rohit Ghai

Martin-Cuadrado

Carol Mizuno