Universidade de São Paulo
Instituto de Química

USP

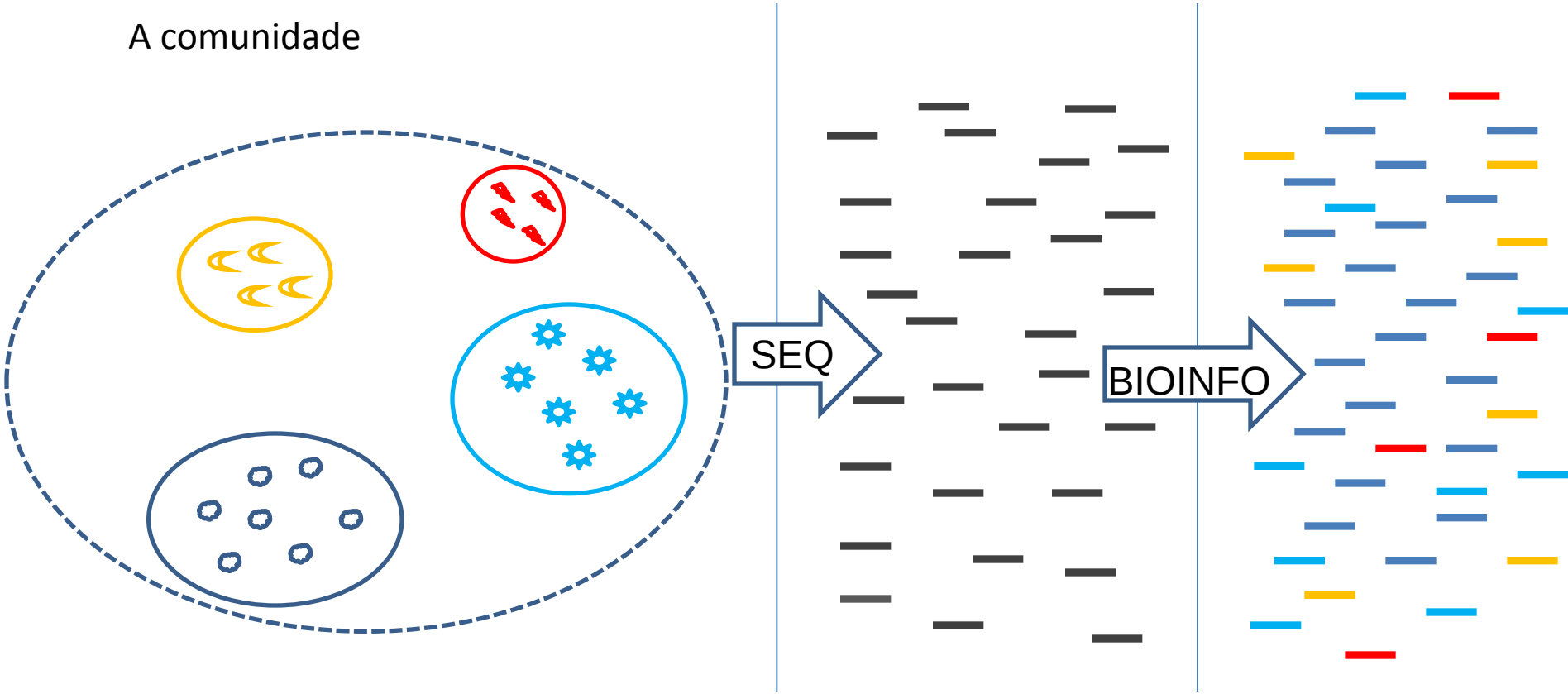
Metagenômica

João C. Setubal

2016

DNA

A comunidade



Taxonomia

- Filo: proteobacteria
  - Classe: proteobacteria gama
    - Ordem: xanthomonadales
      - Família: xanthomonadácea
        - Gênero: xanthomonas
          - Espécie: citri

OTU

- Unidade taxonômica operacional
- Se for conhecida, leva um rótulo padronizado
- Mas pode ser desconhecida

Identificação taxonômica

- Estrutura da comunidade microbiana
  - Descoberta de quais OTUs conhecidos estão presentes na amostra
  - Descoberta de novos OTUs
  - Os “conhecidos” podem ser na verdade novas OTUs parentes próximas de reais conhecidos
  - Quem são os agentes principais?
    - Nem sempre são os mais abundantes
- Dinâmica temporal da comunidade

Dados (sequências)

- 16S
  - Primers específicos
- Dados de DNA total (WMS)
  - Maior parte das sequências vem de genes codificadores de proteínas, mas também tem 16S
- Qual escolher depende dos objetivos do projeto

Muitas variáveis

Variável	valores
Data type	WMS, 16S amplicon
Sequencing technology	Illumina, 454, Ion, pacBio
WMS Reference database:	NR, NT, M5NR
16S Reference database	Greengenes, RDP, Silva
WMS identification program	Many programs available
Taxonomy level	Phylum, class, order, family, genus, species

Muitos programas disponíveis!

Bazinet & Cummings, 2012

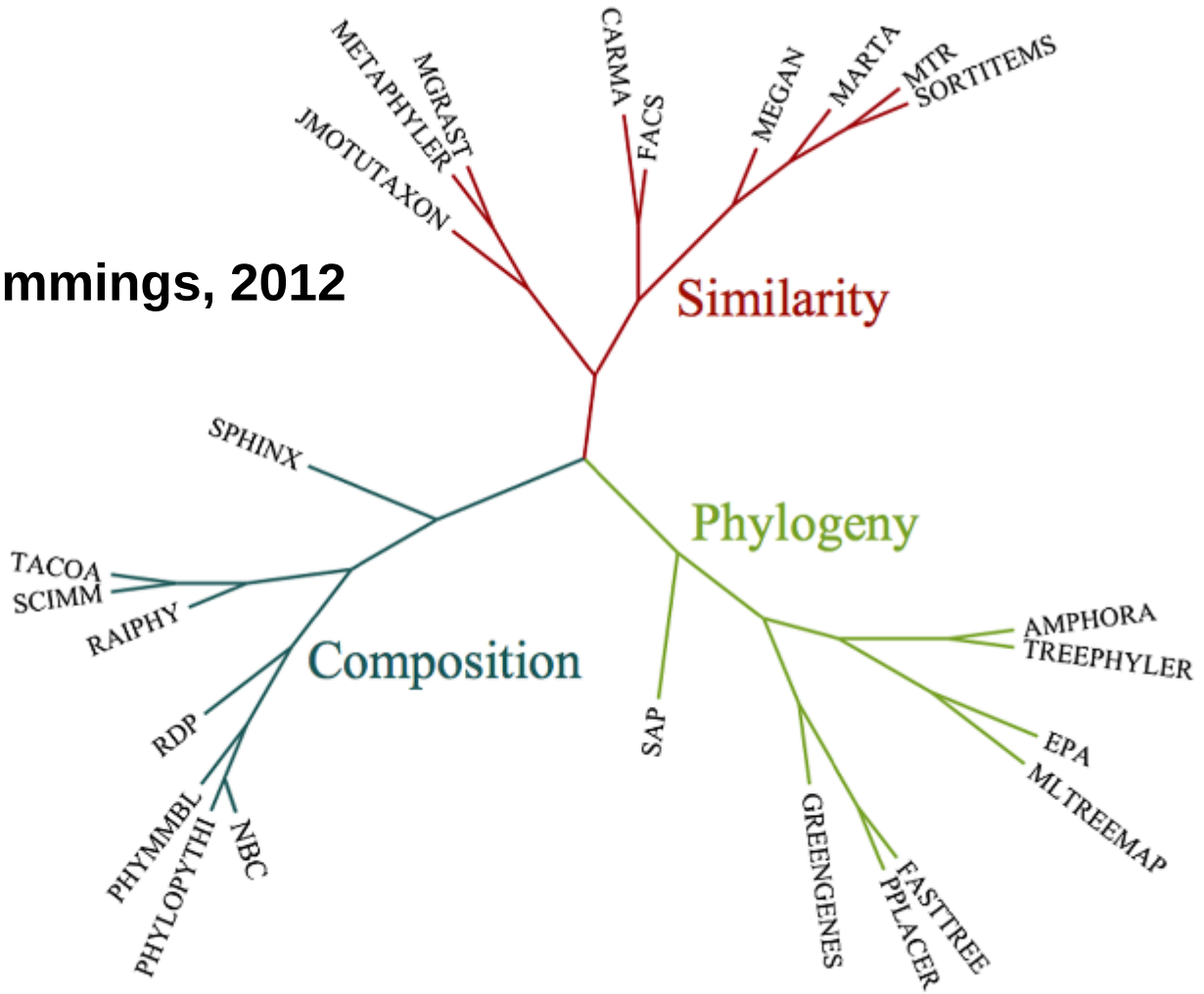


Figure 1 Program clustering. A neighbor-joining tree that clusters the classification programs based on their similar attributes.

Muitas fontes de erro

- Amostragem
- Preparação da biblioteca
- Sequenciamento
- Tamanho da sequência (pode ser curta demais)
- Algoritmo de identificação
- Viéses dos bancos de dados

Binning e classificação

- Binning
  - Juntamos em diferentes caixinhas as sequências que são parecidas entre si
    - Não sabemos o que contém cada caixinha
    - OTU1, OTU2, OTU3, etc
- Classificação
  - Procuramos associar um rótulo taxonômico a cada caixinha (ou a cada read ou fragmento)

Análise de abundância

- Que organismos ou funções são mais abundantes num nicho?
- Usar contagem de reads como indicador de abundância

Classificação de reads de DNA total

- Similaridade com sequências de origem conhecida
  - BLAST
- Propriedades intrínsicas de cada sequência
  - Assinaturas genômicas
    - Apropriado para binning

Classificação com base na frequência de palavras de *k* bases

k = 4: AAAA, AAAC, AAAG, AAAT, CAAA, etc...

Dada uma janela de *x* kb, podemos contar as ocorrências de cada uma dessas palavras dentro da janela

Exemplo:

AGATTAGCGACTATTATAGCCTAGATCGATCATTACC

AGAT ocorre 2 vezes

ATTA ocorre 3 vezes

etc

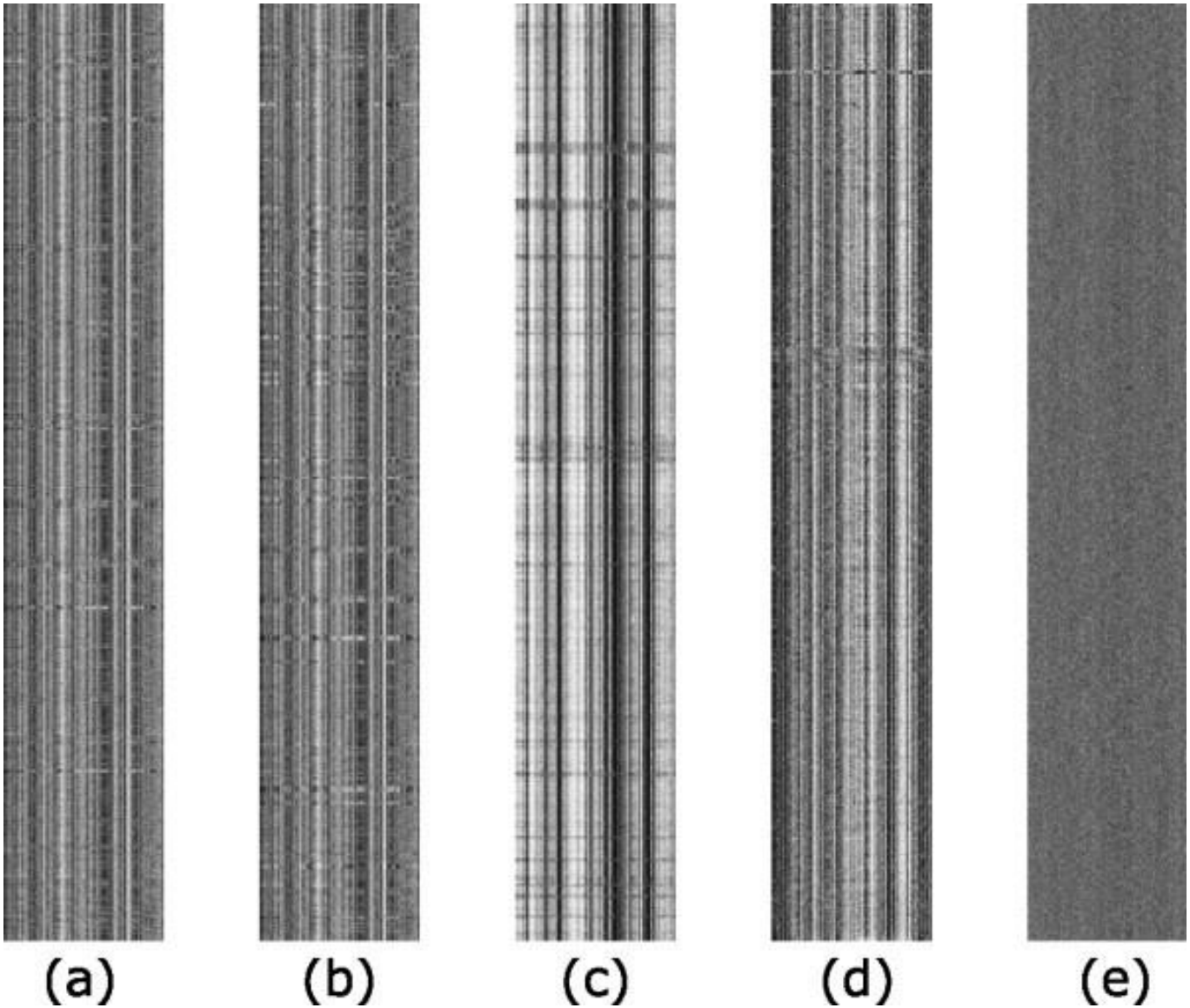
Palavras de k bases: k-mers (kâmeros)

Matriz de frequências

janela	AAAA	AAAC	AAAG	AAAT	ACAA	ACAC	ACAG	ACAT
1	15	2						
2	16	3						
3	14	0						
4	13	2						
5	15	4						
6	12	0						
7	18	1						
8	17	3						
9	16	1						

Genome “barcodes”

Burkholderia pseudomallei



E. coli K12

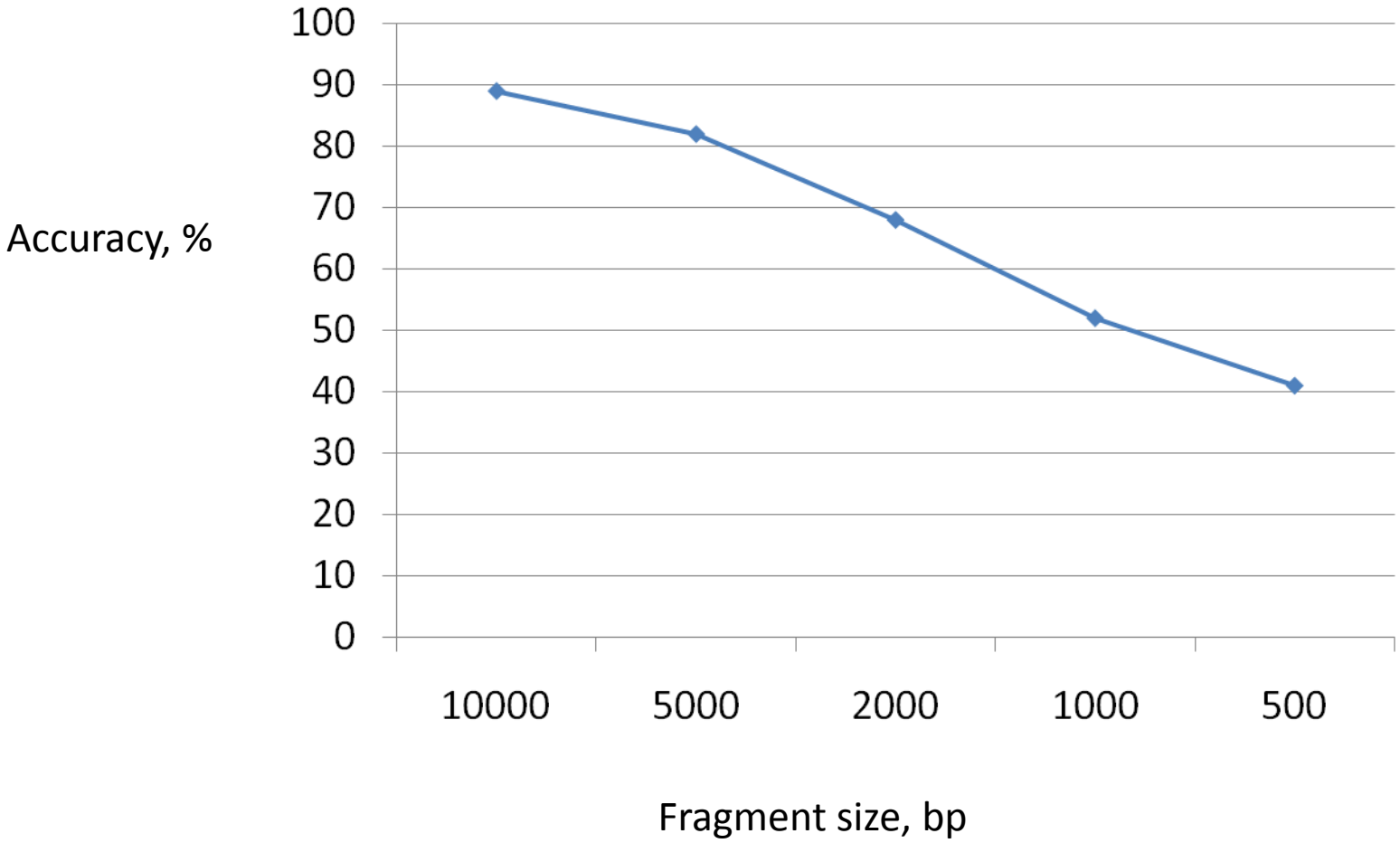
E. coli O157

Pyrococcus furiosus

random

Zhou, Olman, Xu, *BMC Bioinformatics*, 2009

Não funciona bem com fragmentos curtos



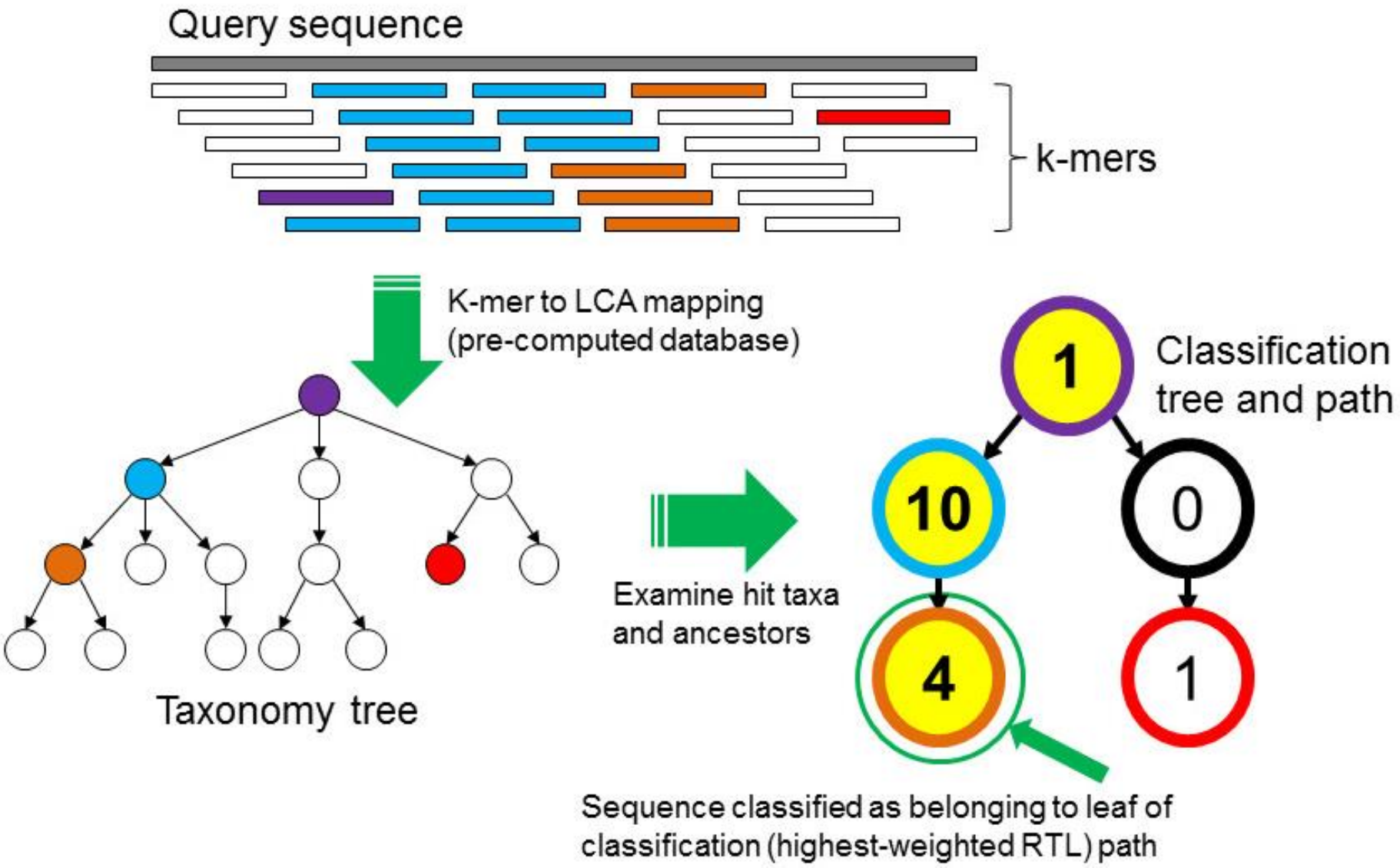
Zhou et al, 2009 simulated data

Exercício

- S_1 = TTCTACTACT
- S_2 = TTGTACTAGG
- S_3 = ACTTCTACTA
- Contar palavras de tamanho 2

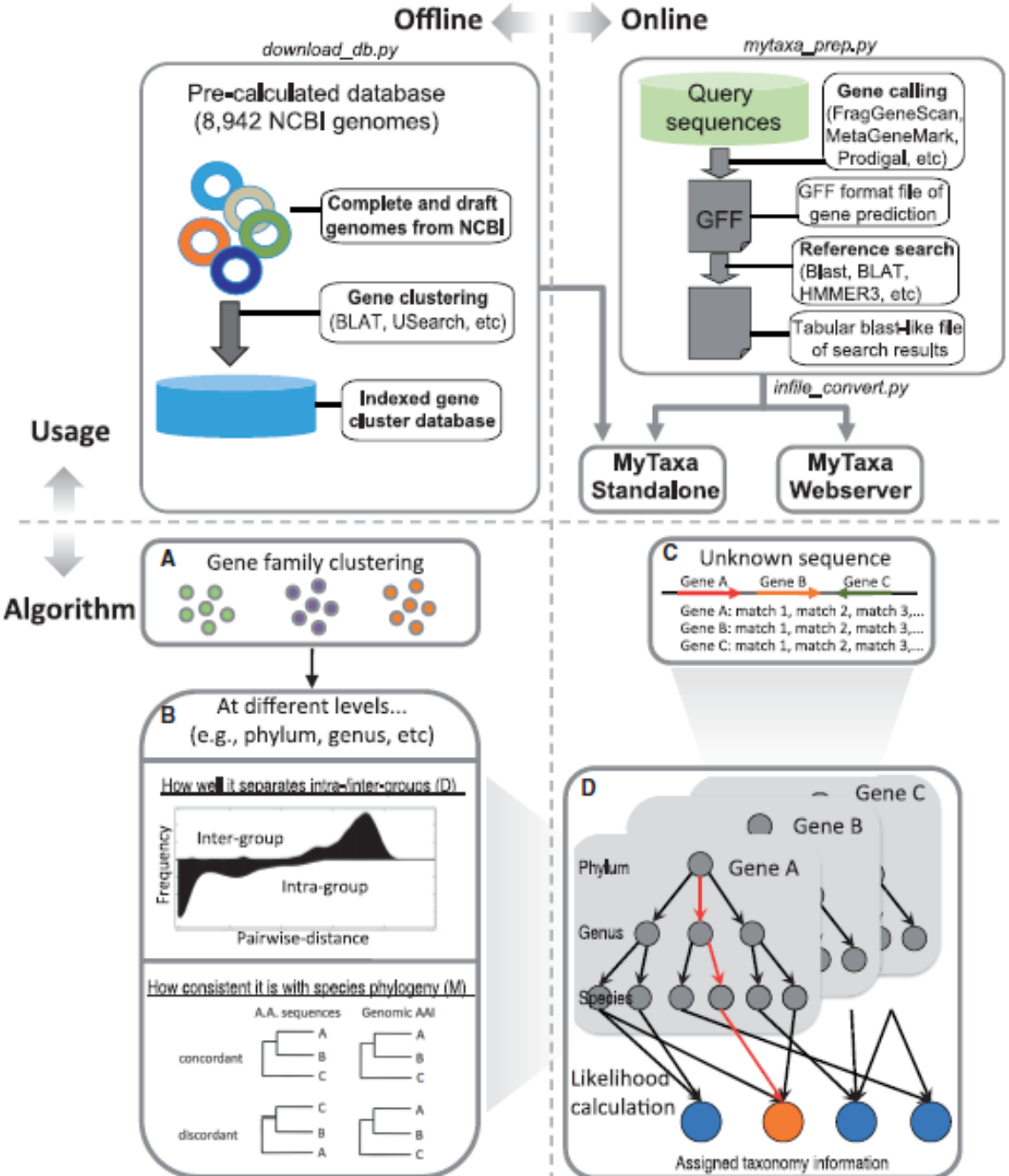
Kraken

- Wood & Salzberg: *Genome Biology*, 2014
- Ideia: um banco de dados com k-mers e o LCA (ancestral comum mais baixo/próximo) de todos os organismos que contém aquele k-mer


Query sequence
k-mers
K-mer to LCA mapping
(pre-computed database)
Taxonomy tree
Examine hit taxa
and ancestors
1
10
0
4
1
Classification
tree and path
Sequence classified as belonging to leaf of
classification (highest-weighted RTL) path

myTaxa

- Luo, Rodriguez-R, Konstantinidis: *Nucleic Acids Research*, 2014
- The distinguishing aspect of MyTaxa is that it employs all genes present in an unknown sequence as classifiers, weighting each gene based on its (predetermined) classifying power at a given taxonomic level and frequency of horizontal gene transfer

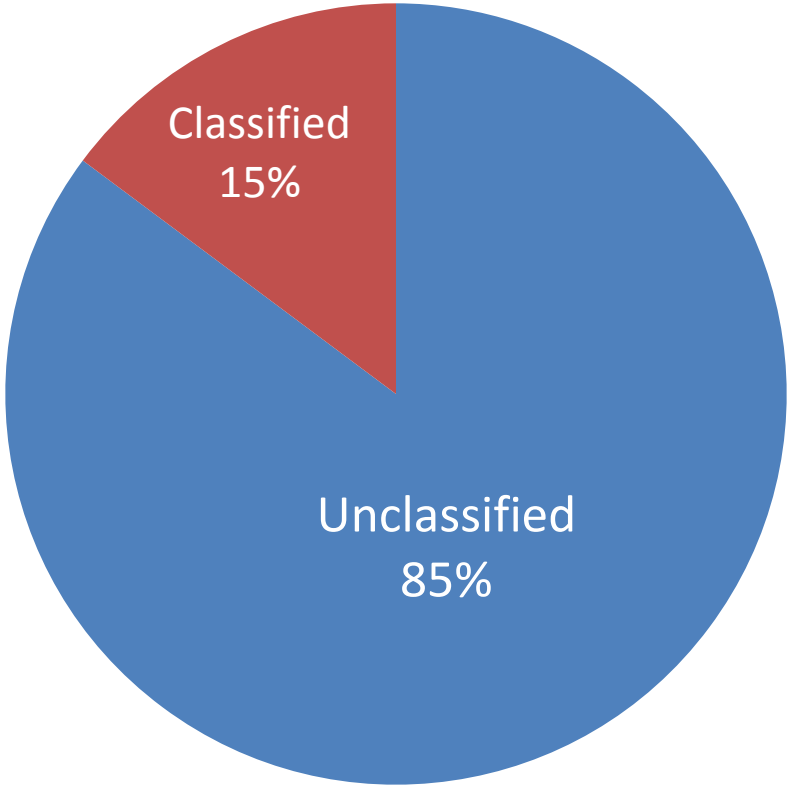

Offline
Online
download_db.py
Pre-calculated database
(8,942 NCBI genomes)
Complete and draft genomes from NCBI
Gene clustering
(BLAT, USearch, etc)
Indexed gene cluster database
mytaxa_prep.py
Query sequences
Gene calling
(FragGeneScan, MetaGeneMark, Prodigal, etc)
GFF
GFF format file of gene prediction
Reference search
(Blast, BLAT, HMMER3, etc)
Tabular blast-like file of search results
infile_convert.py
MyTaxa Standalone
MyTaxa Webserver
Usage
Algorithm
A
Gene family clustering
B
At different levels...
(e.g., phylum, genus, etc)
How well it separates intra-/inter-groups (D)
Inter-group
Intra-group
Frequency
Pairwise-distance
How consistent it is with species phylogeny (M)
A.A. sequences
Genomic AAI
concordant
discordant
C
Unknown sequence
Gene A
Gene B
Gene C
Gene A: match 1, match 2, match 3,...
Gene B: match 1, match 2, match 3,...
Gene C: match 1, match 2, match 3,...
D
Gene C
Gene B
Gene A
Phylum
Genus
Species
Likelihood calculation
Assigned taxonomy information

Uma comparação usando dados reais

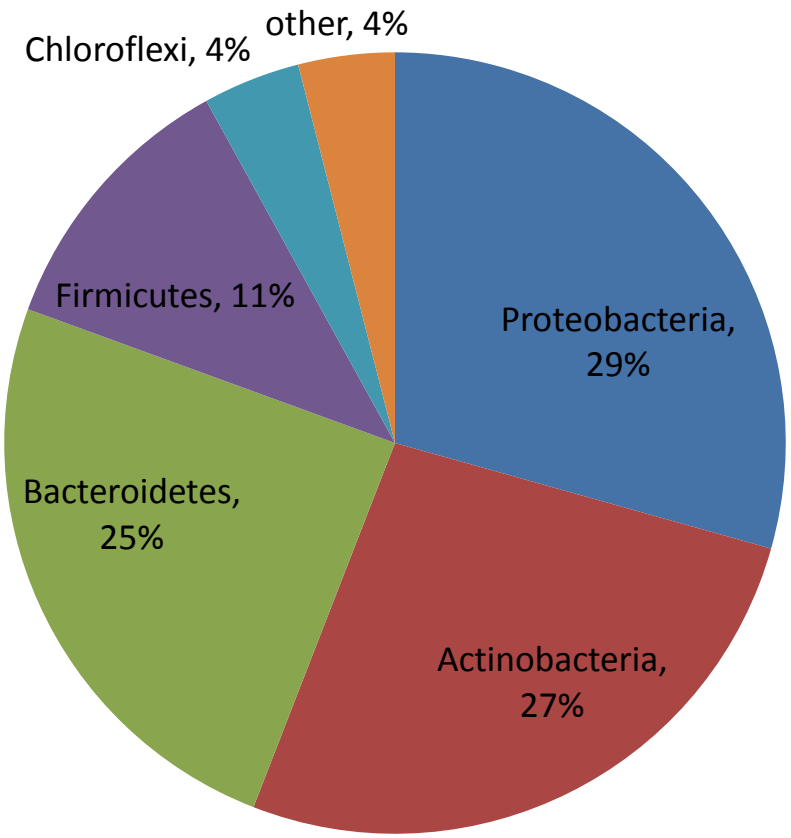
Dados de entrada

Variable	value
dataset	ZC4, day 15 (metazoo)
Data type	WMS
Sequencing technology	Illumina miSeq
WMS program	Kraken [Wood & Salzberg, 2014]
WMS Reference database:	Kraken database
Taxonomy level	Phylum

Resultado

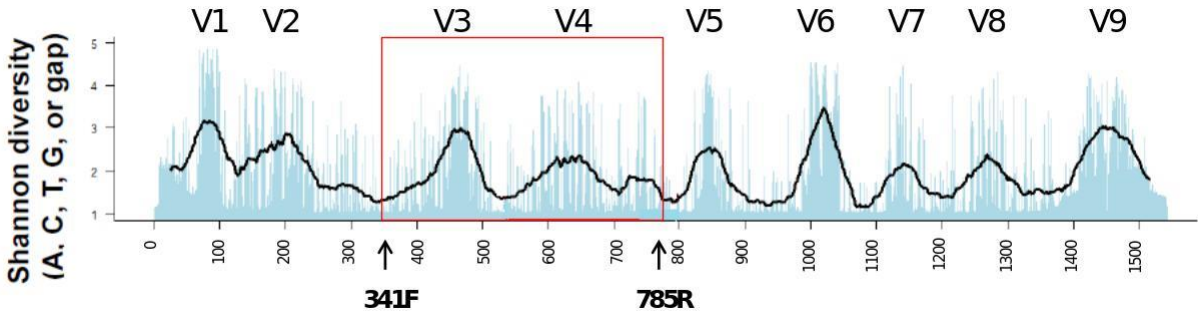


Dentre os reads classificados como bactérias (99%)



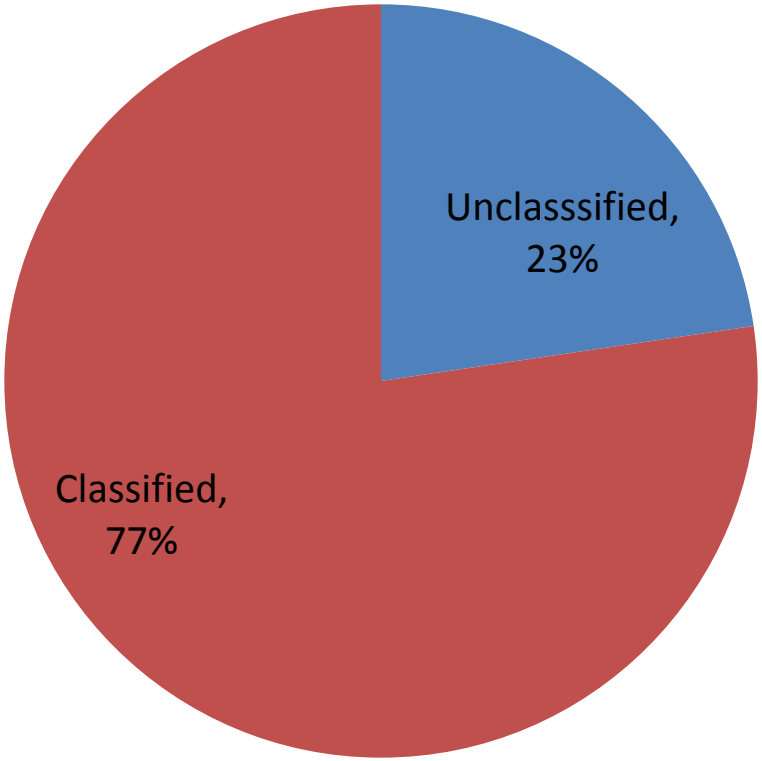
Outro tipo de análise

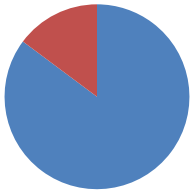
Variable	value
dataset	ZC4, day 15
Data type	16S
Data type details	V3 and V4, read size ~416 bp
Sequencing technology	Illumina (miSeq)
16S analysis pipeline	Qiime (RDP classifier + Greengenes)
Taxonomy level	Phylum



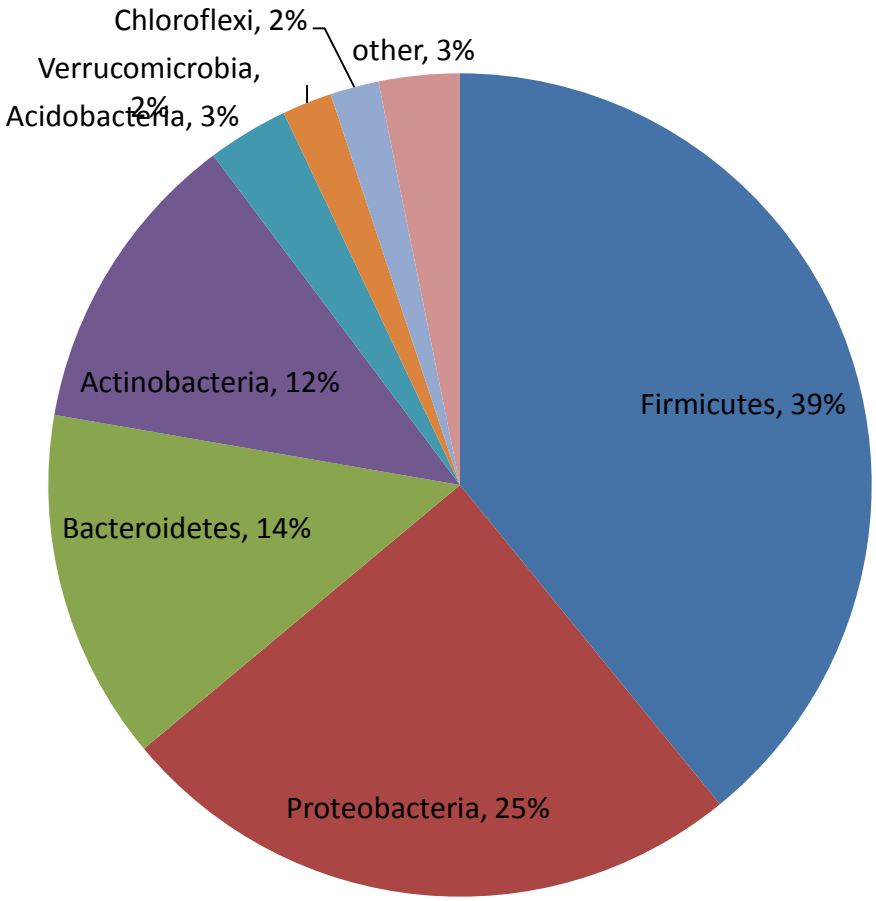
Qiime (Caporaso *et al.*, 2010)

Resultados

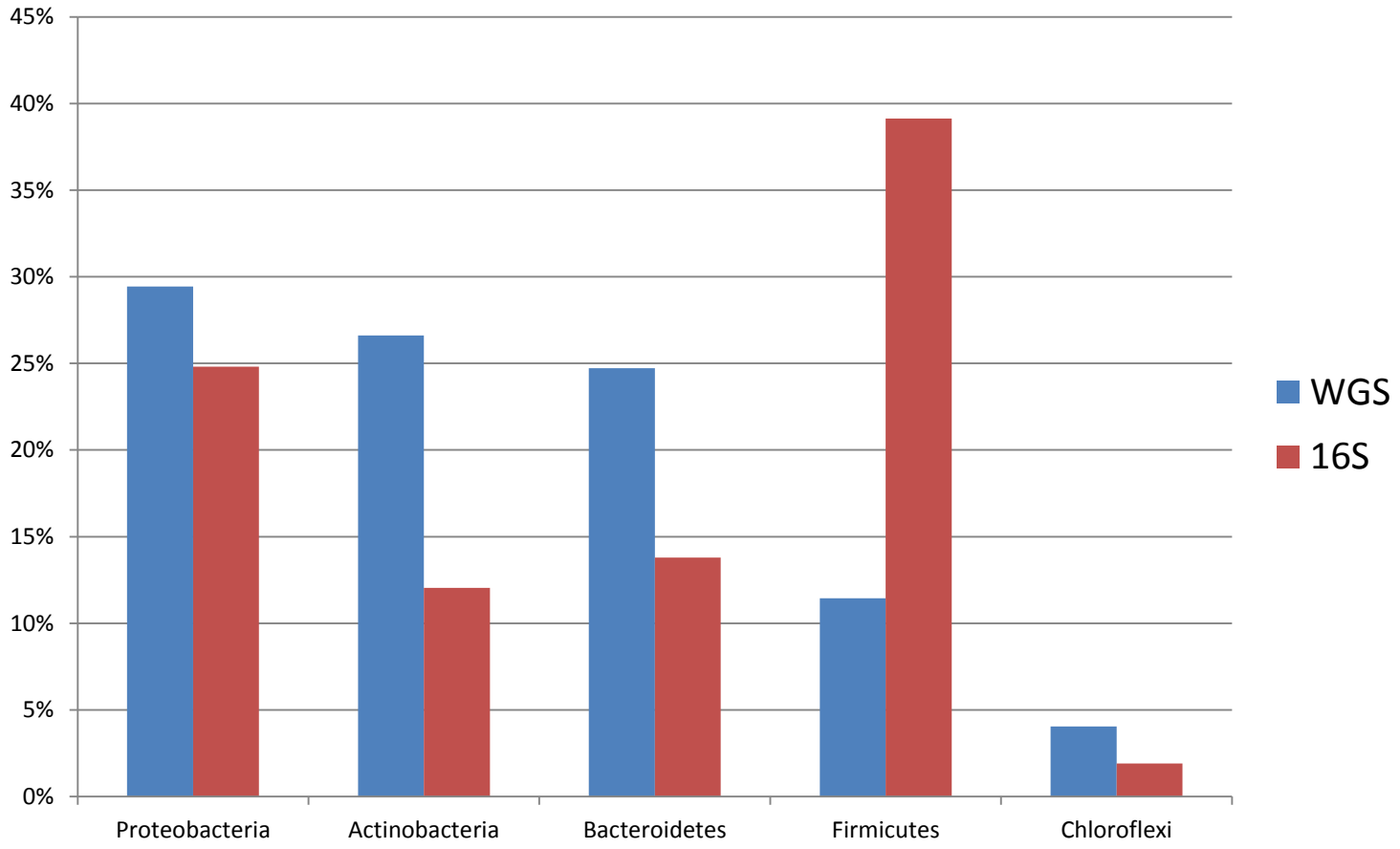




Dentre os classificados



WMS e 16S


45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Proteobacteria
Actinobacteria
Bacteroidetes
Firmicutes
Chloroflexi
WGS
16S

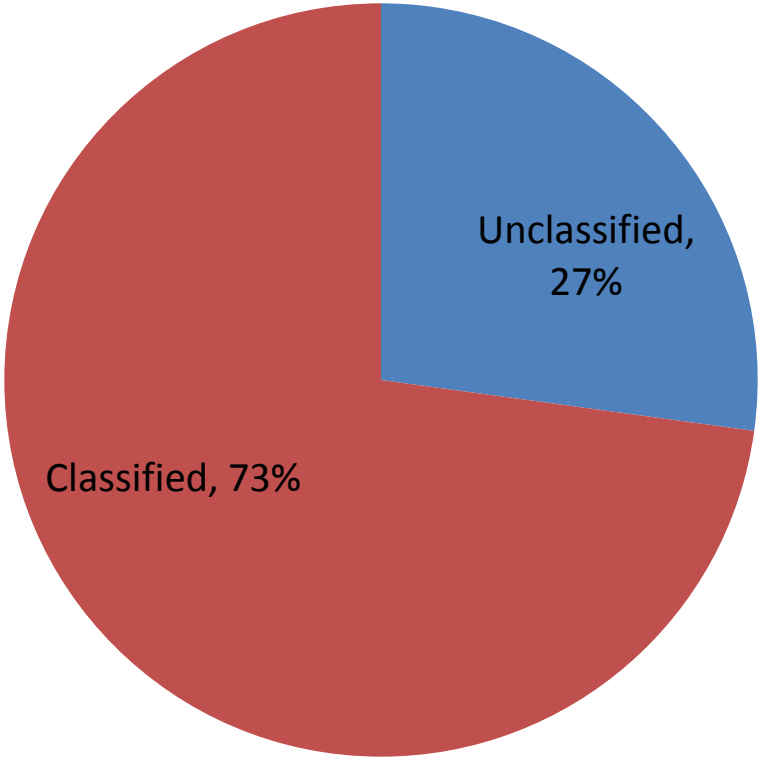
Comentários

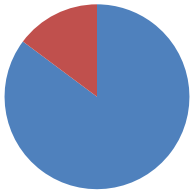
- Concordância qualitativa
  - boa
- Concordância quantitativa (abundância relativa)
  - fraca
    - esp. Firmicutes
- Kraken deixa muitos reads não classificados

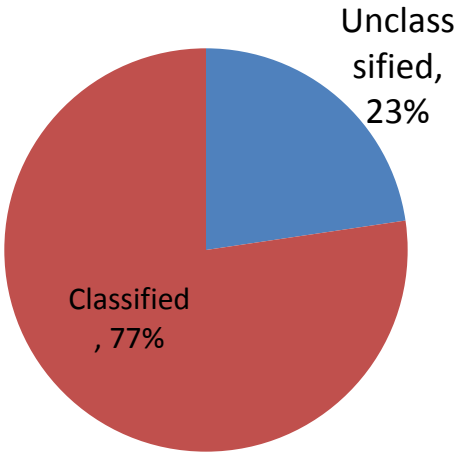
Vamos mudar o programa identificador

- myTaxa

Resultados



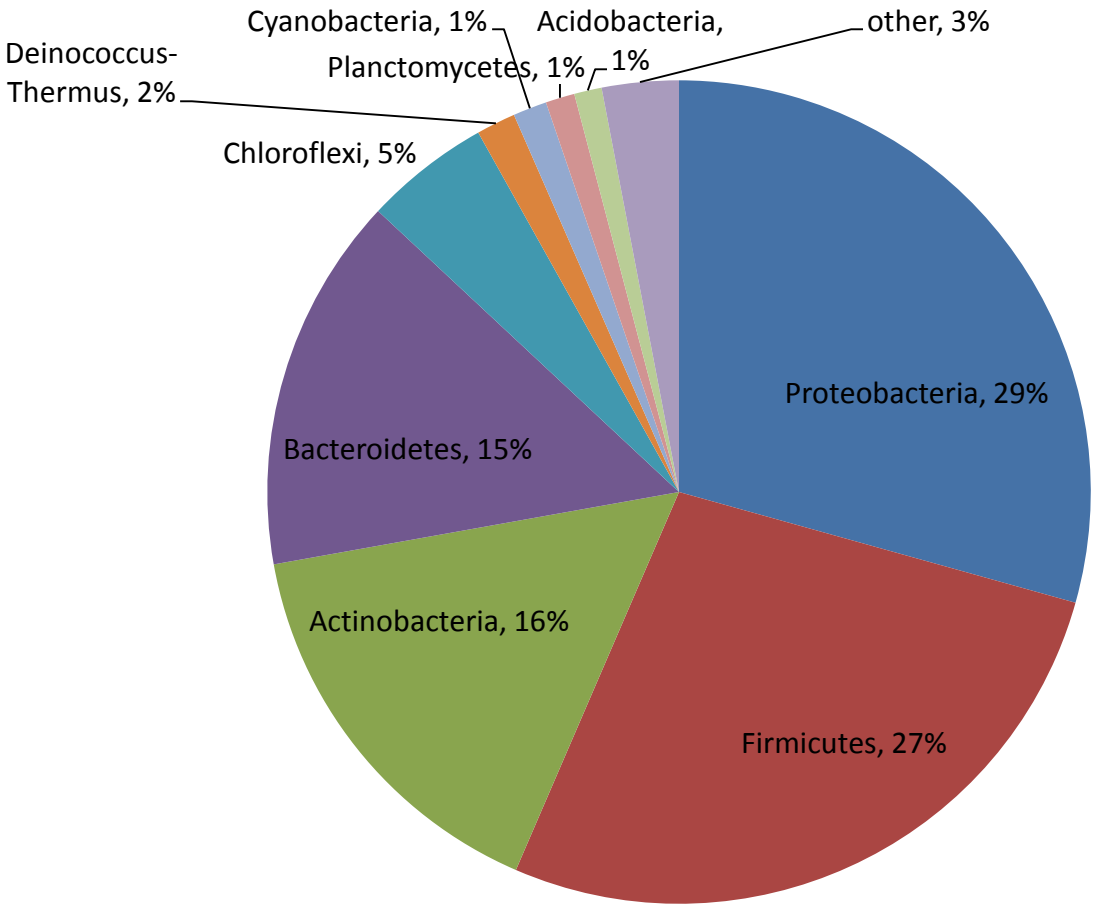




Kraken

16S

Dentre os classificados como bactéria (99%)

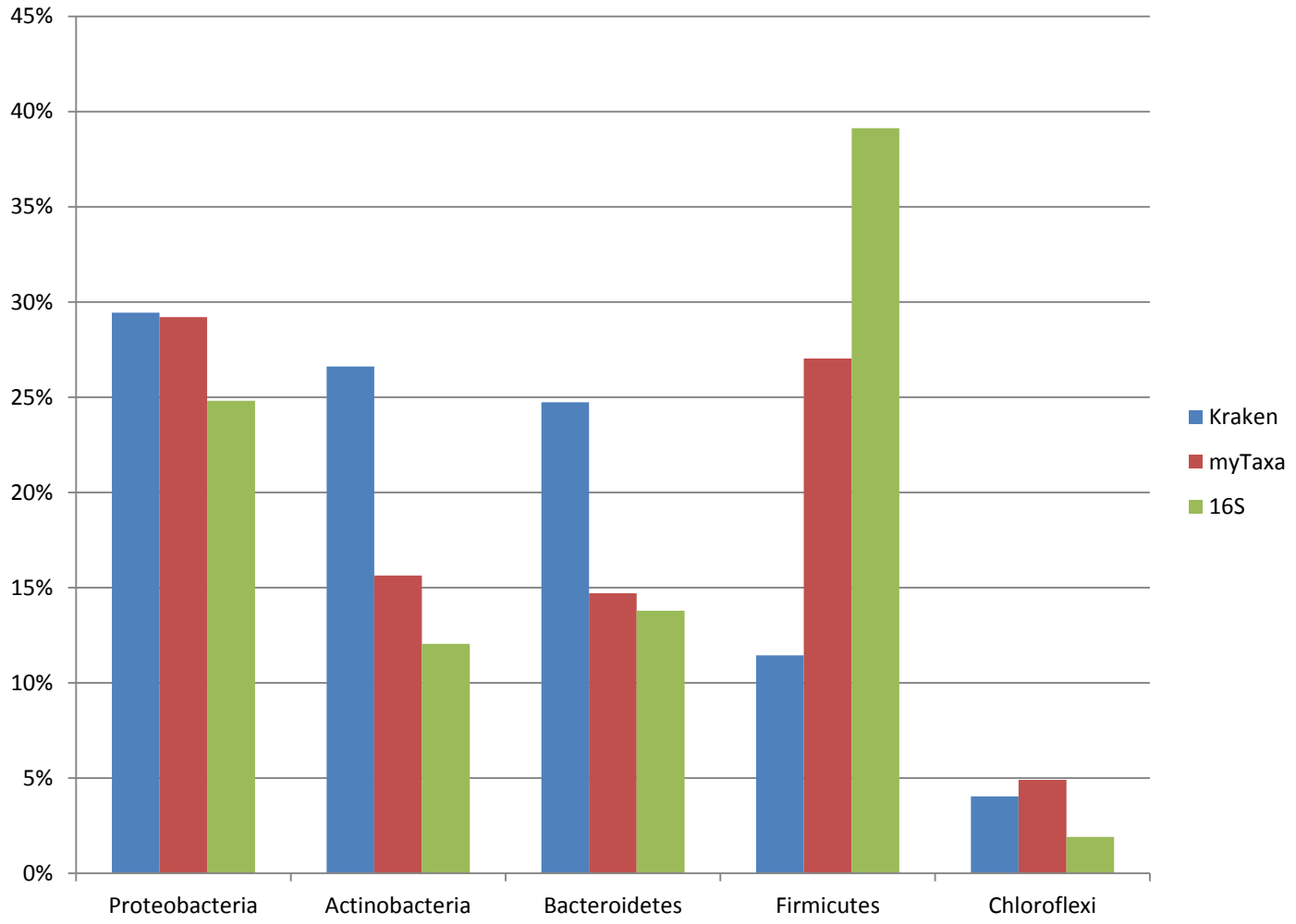


Prokaryotes in Genbank

filo	# genomas	%
Actinobacteria	4059	13
Bacteroidetes/chlorobi	932	3
Cyanobacteria	340	1
Firmicutes	9628	31
Proteobacteria	14268	46
Spirochaetes	525	2
Others	1500	5

Source: Land et al. 2015

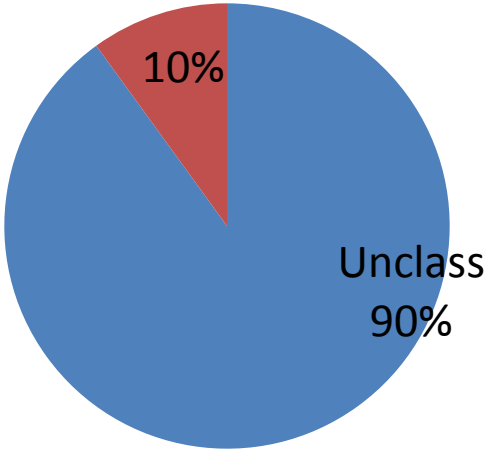
Comparação de todos os resultados



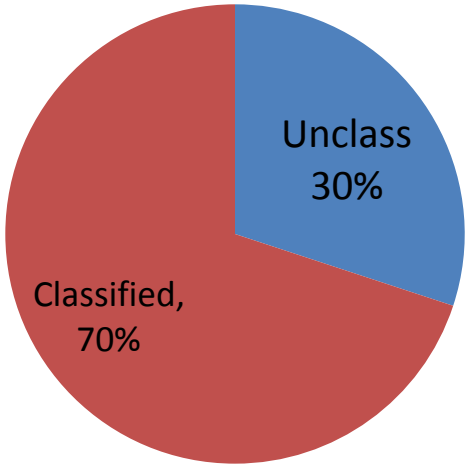
Réplica

- Outro conjunto de dados: ZC4, day 99
- Mesmas metodologias
  - Kraken
  - myTaxa
  - 16S

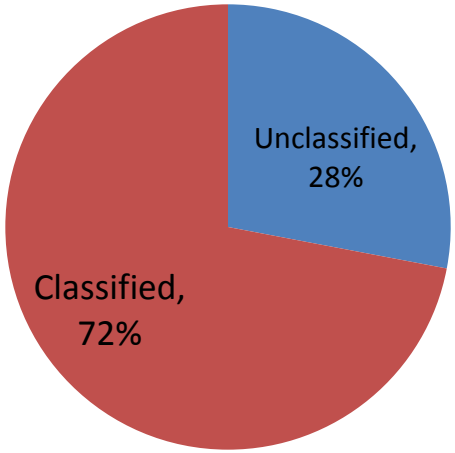
Resultados



kraken

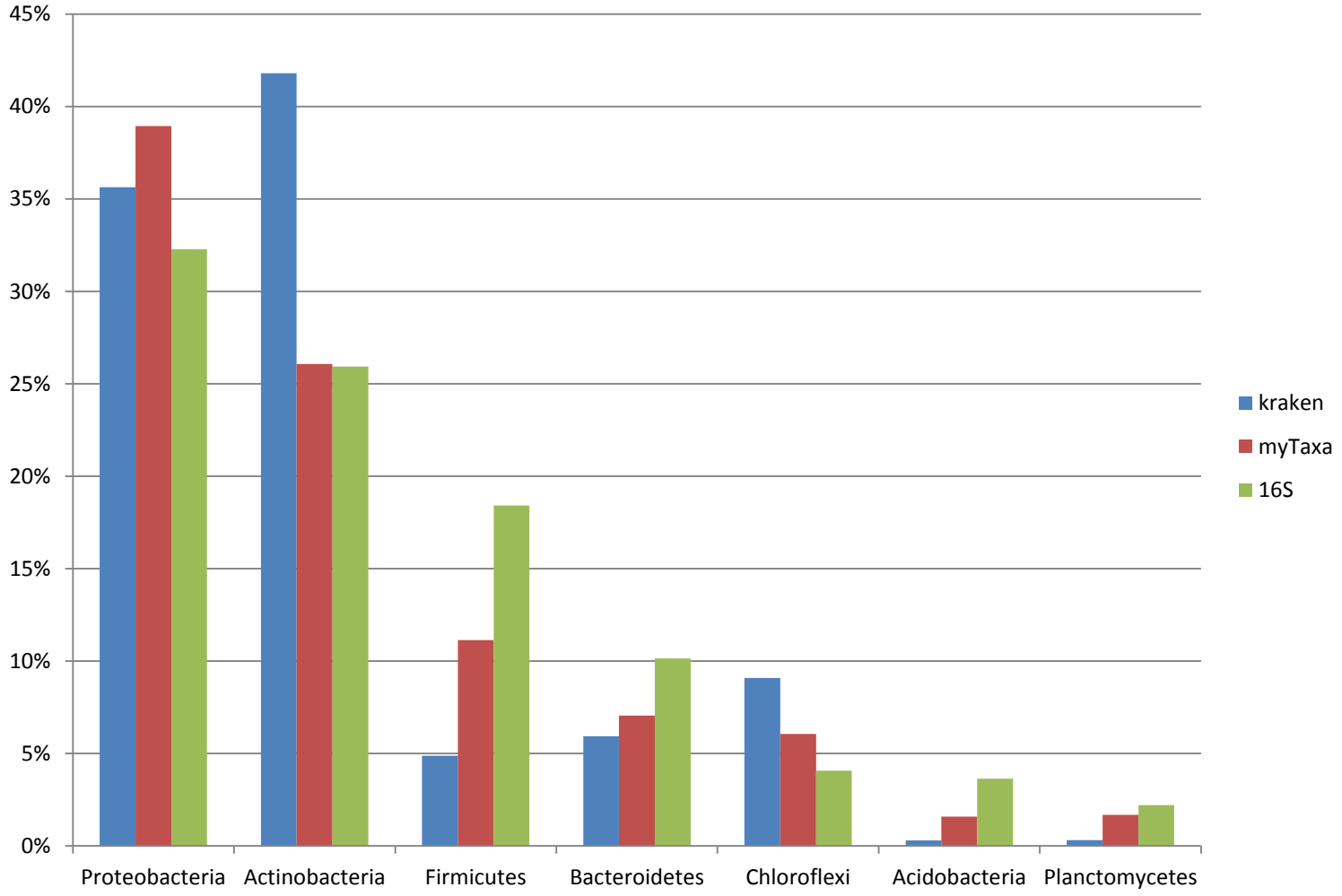


16S



myTaxa

comparação


45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Proteobacteria
Actinobacteria
Firmicutes
Bacteroidetes
Chloroflexi
Acidobacteria
Planctomycetes
kraken
myTaxa
16S

Conclusões

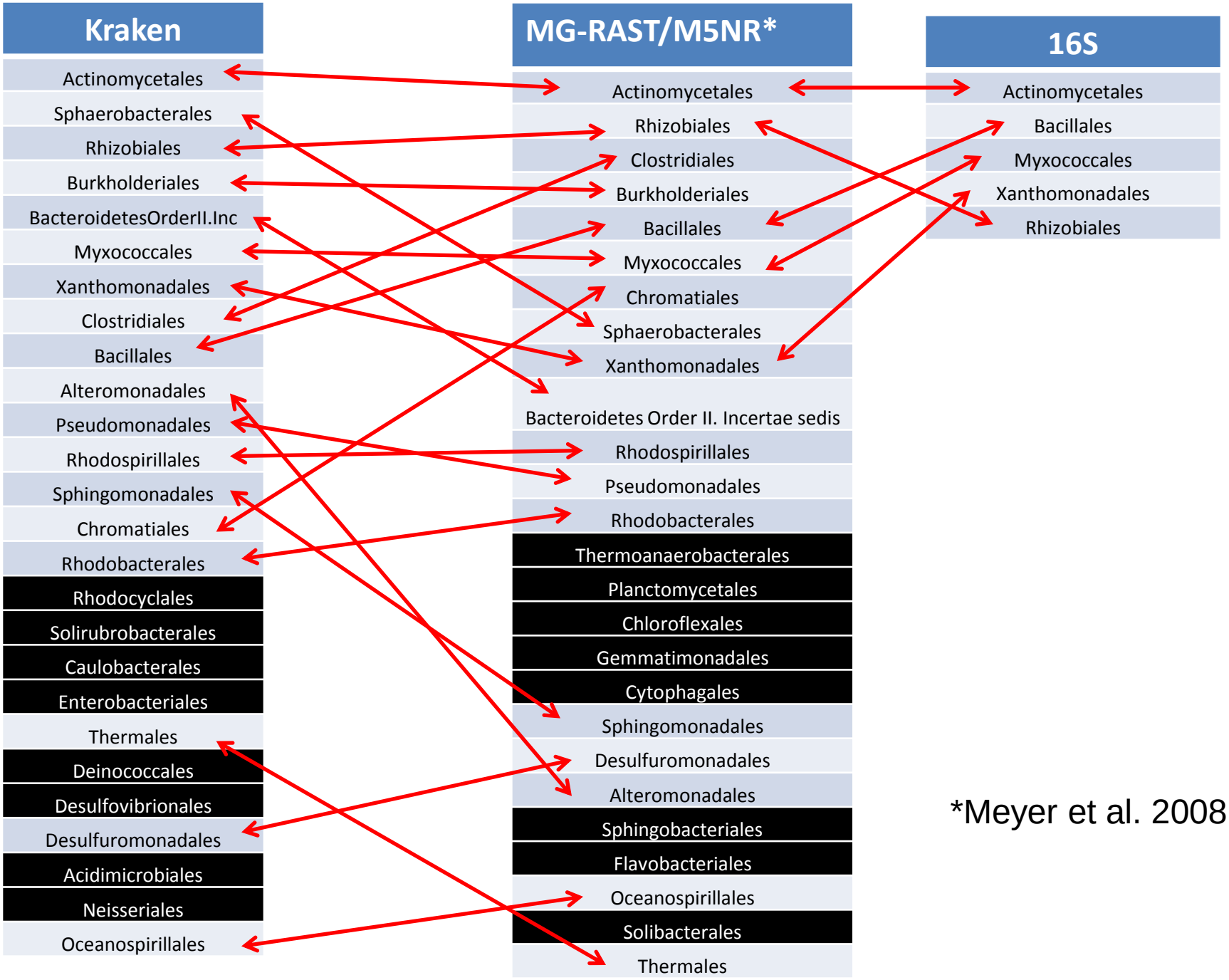
- WMS vs. 16S
  - 16S pode perder OTUs pela especificidade dos primers
    - Menos sensibilidade
  - É mais difícil chegar ao nível de espécie
  - Identificações positivas são confiáveis
    - Especificidade boa
  - WMS tem melhor sensibilidade (pega tudo)
    - Mais sensível aos vieses dos bancos
- myTaxa é um programa melhor do que kraken

Vamos mudar o nível taxonômico

- Ordem
- Dados: ZC4, day 99

Fica mais complicado…

- RDP muitas vezes diz que reads pertencem a uma ordem desconhecida
- Idem myTaxa
- Estimativas de abundância ficam ainda menos confiáveis do que ao nível de filo
- Existe maior sensibilidade ao contéudo dos bancos utilizados

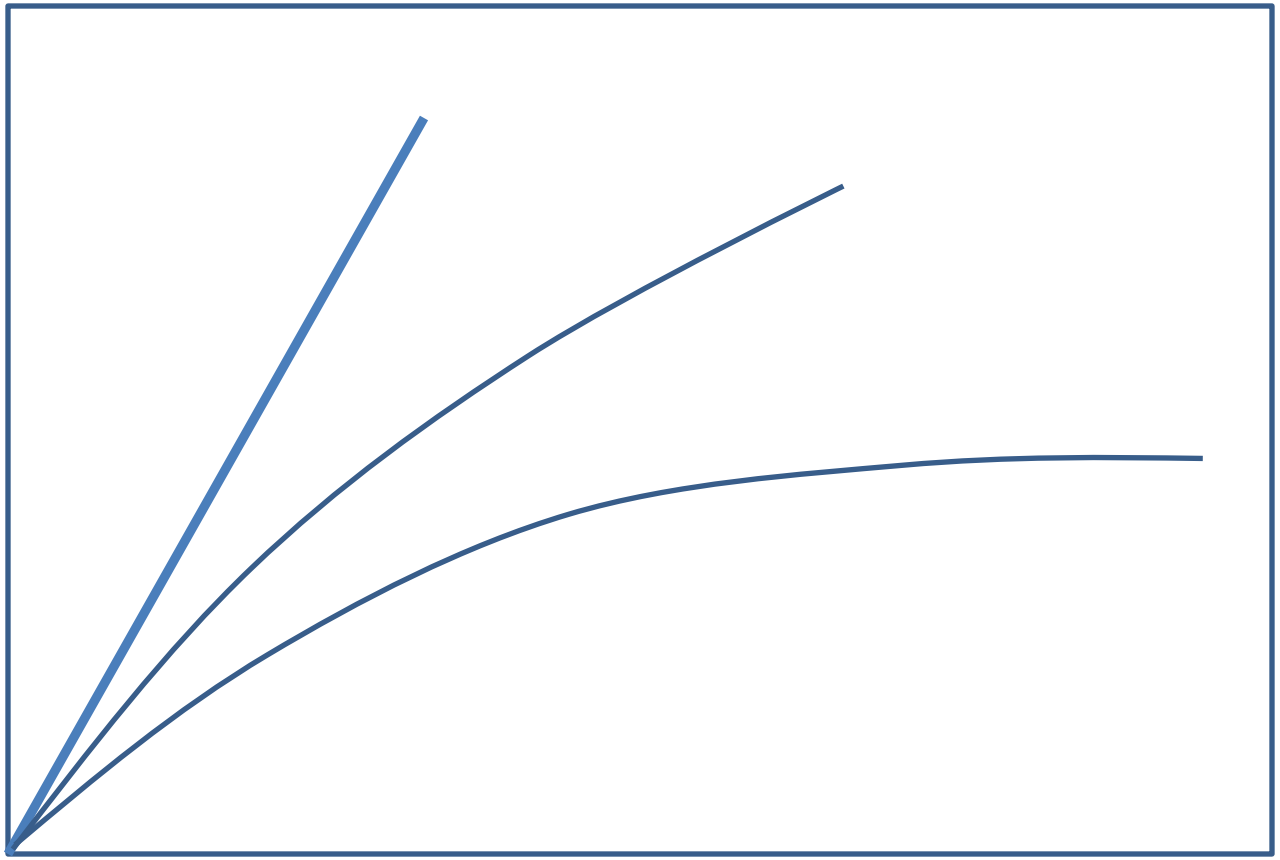


*Meyer et al. 2008

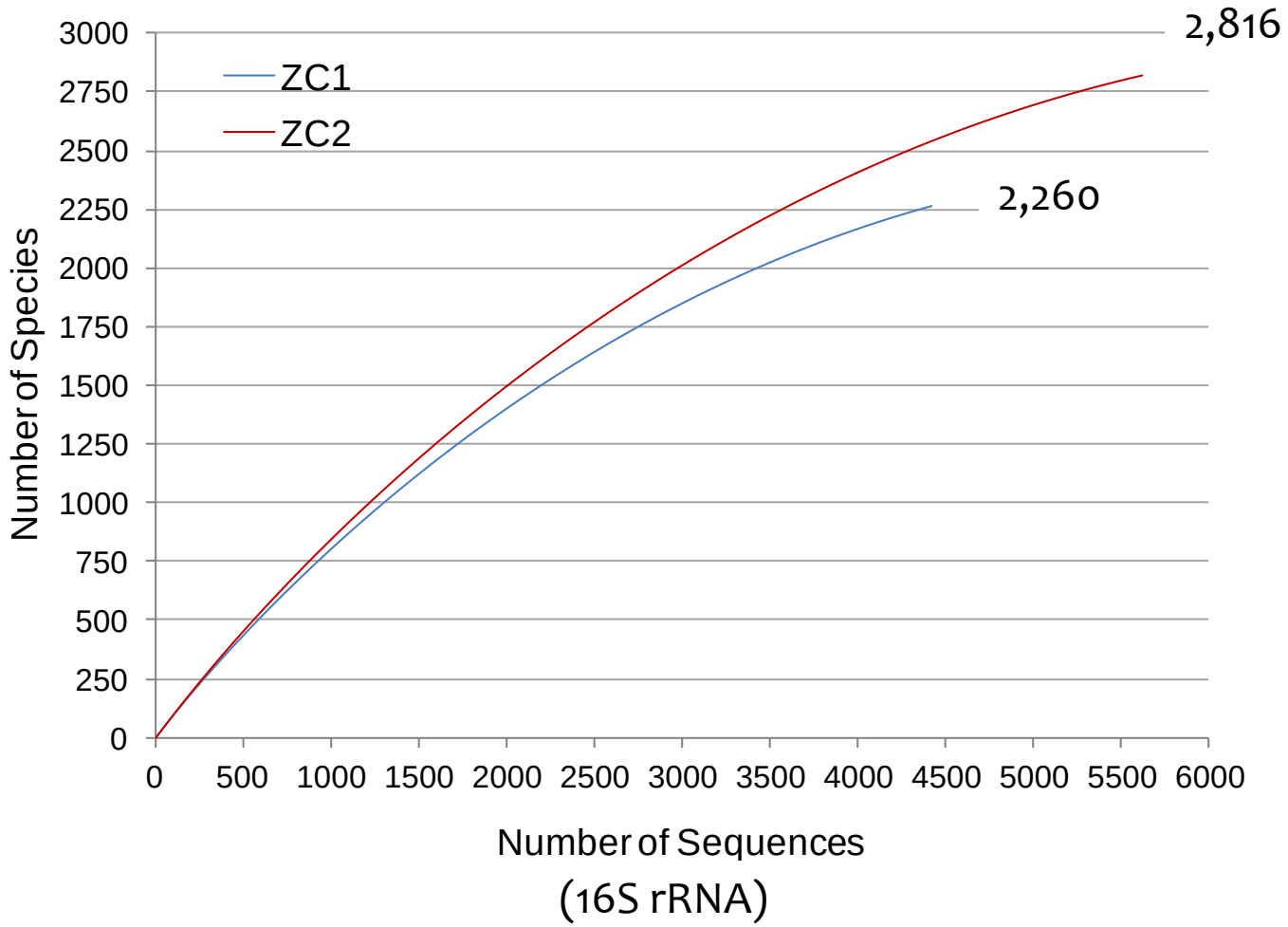
Análise de abundância

- Abundância relativa
  - Curvas de rarefação
- Variação
  - temporal
  - espacial
  - Entre diferentes condições

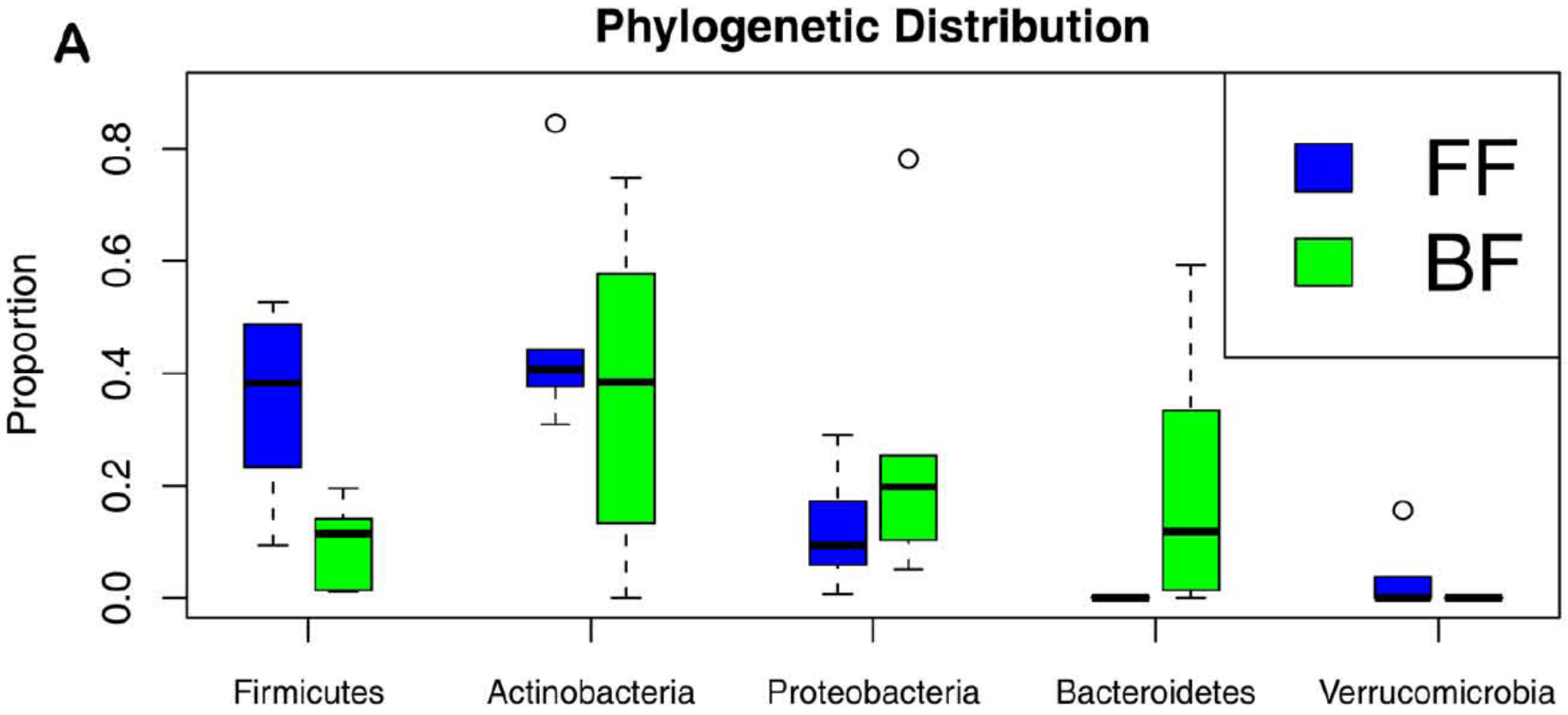
Curvas de rarefação (ou saturamento)



Amostras metazoo

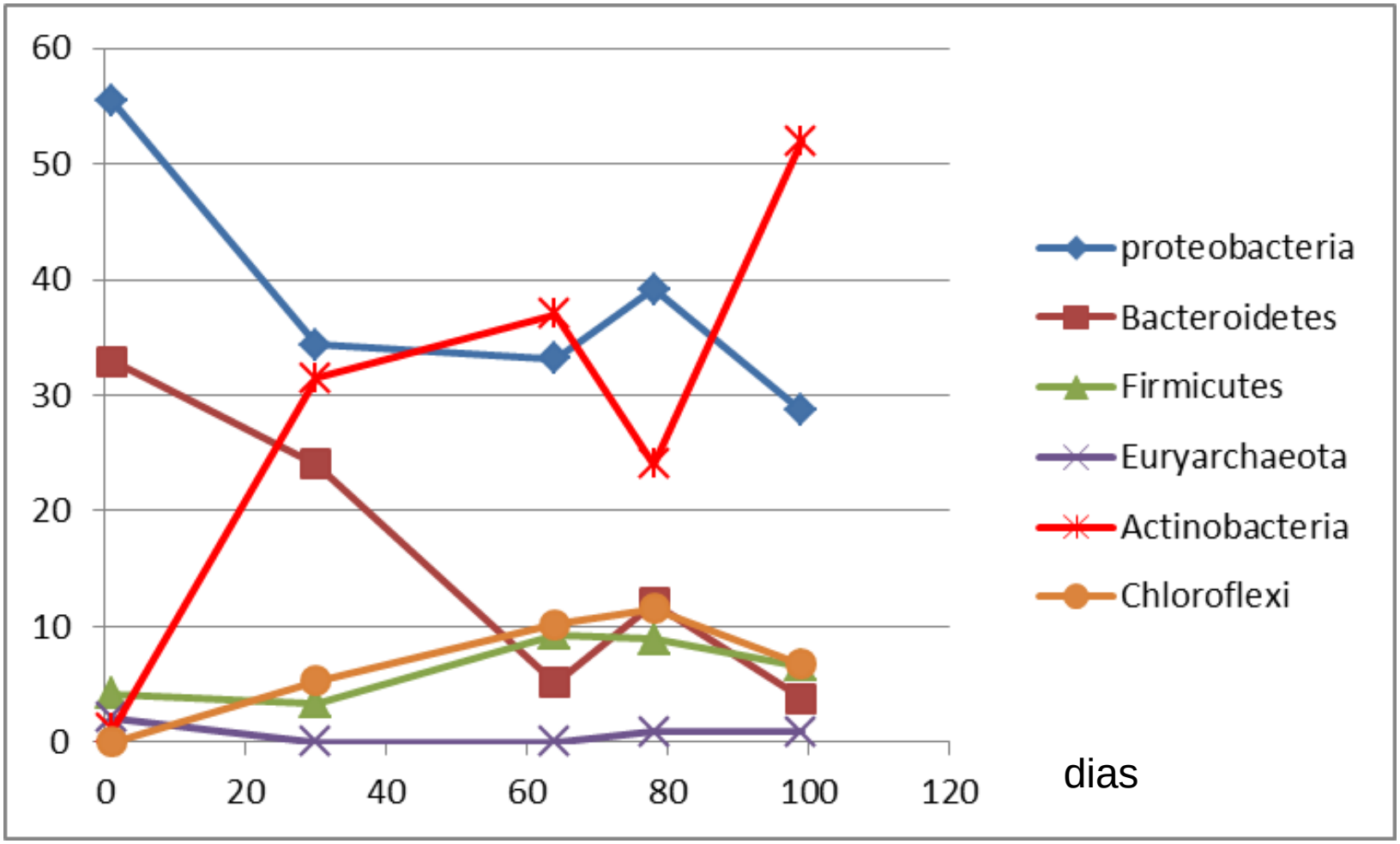


Variação da microbiota intestinal entre bebês amamentados no seio e com mamadeira



Schwartz *et al. Genome Biology* 2012 **13**:r32 doi:10.1186/gb-2012-13-4-r32

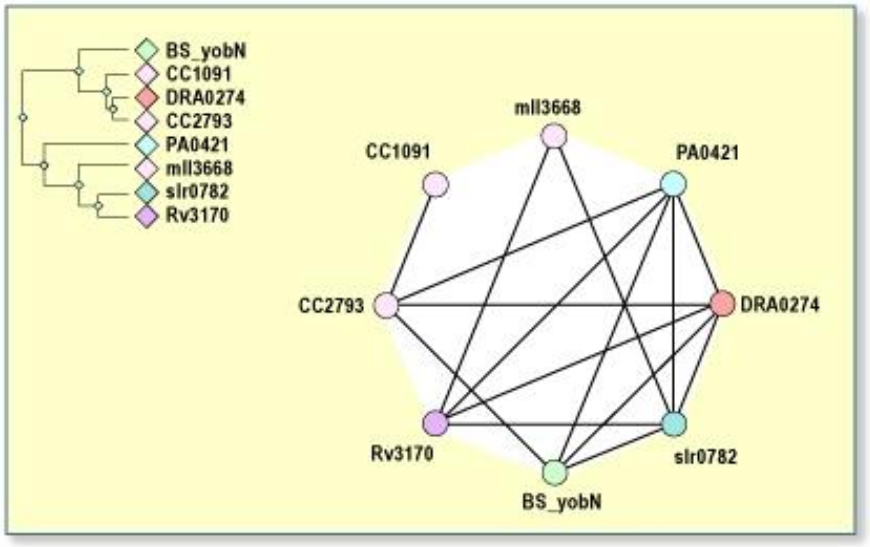
Variação da abundância relativa no tempo



Prejeto metagenoma zoológico - compostagem

Abundância de funções

- BLASTX de reads contra um banco de COGs Cluster of Orthologous Groups



Example of a COG: monoamine oxidase

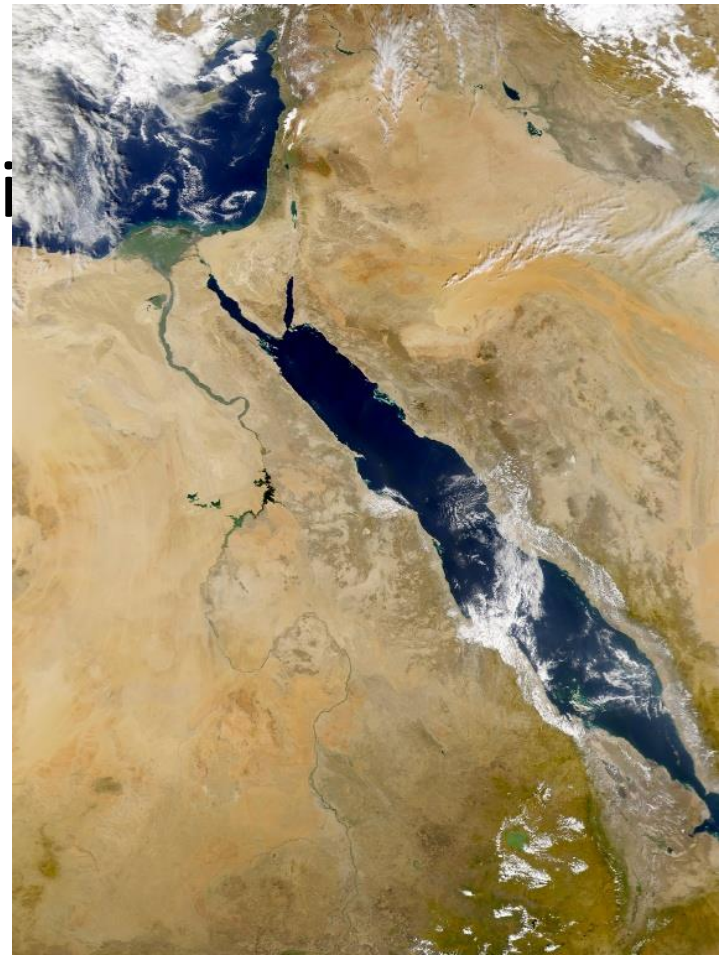
http://www.ncbi.nlm.nih.gov/books/NBK21090/ eggNOG [Jensen et al. 2008]

Abundância relativa espacial

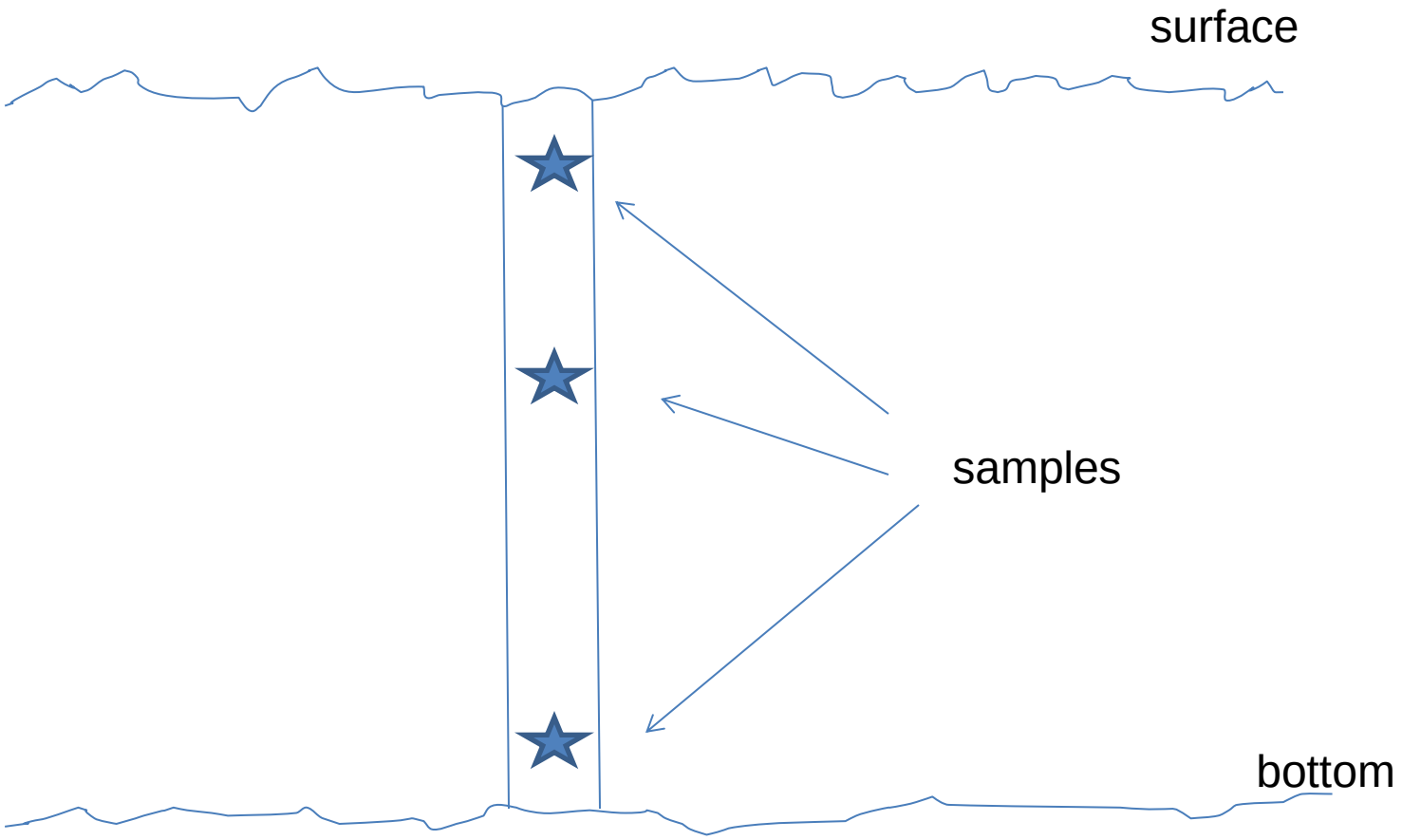
- **COGs diferencialmente representados**
- Semelhante a genes diferencialmente expressos
- Heat maps, clusterização hierárquica

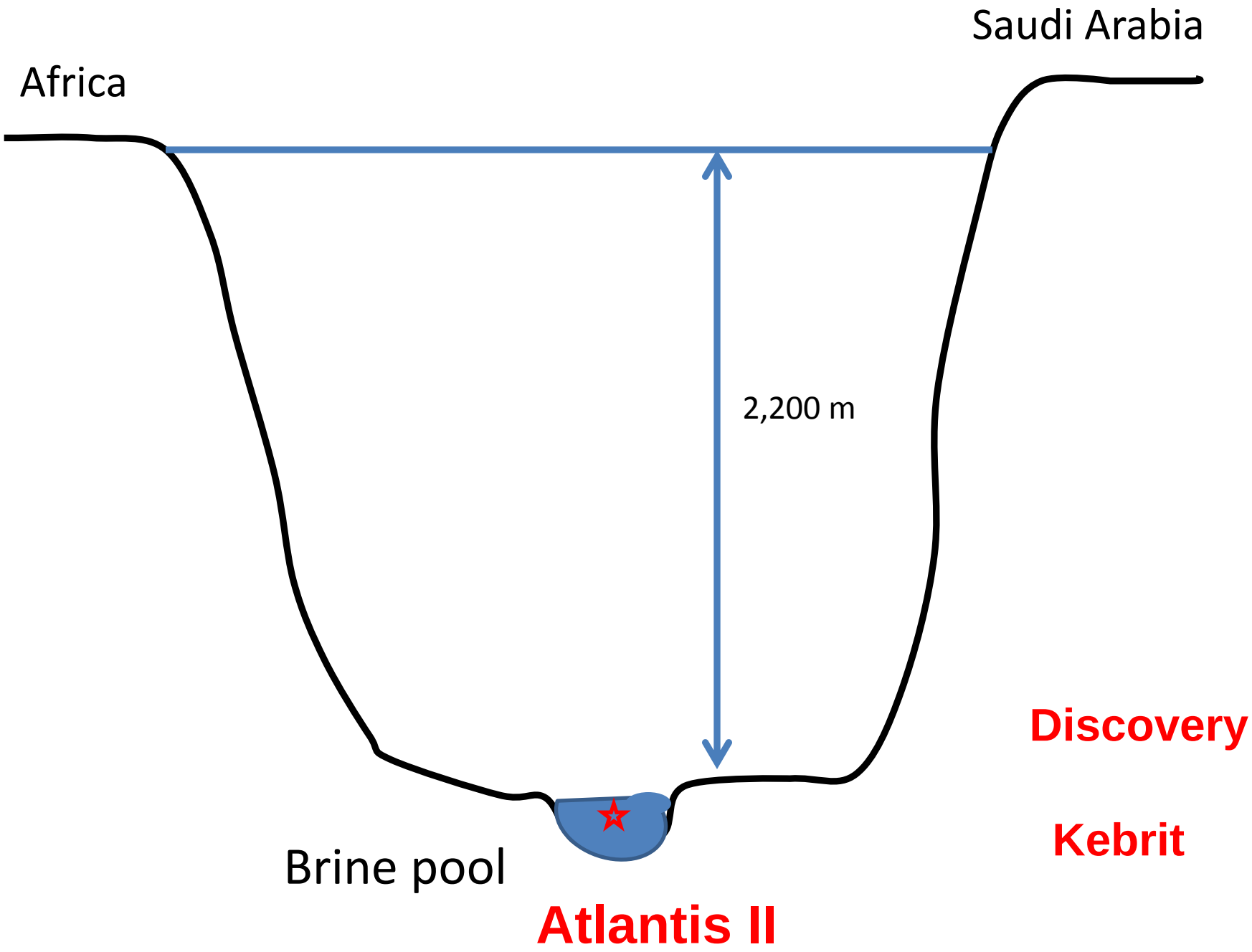
Motivation

- Red Sea
- American University in Cai
- KAUST funding



Ocean water columns



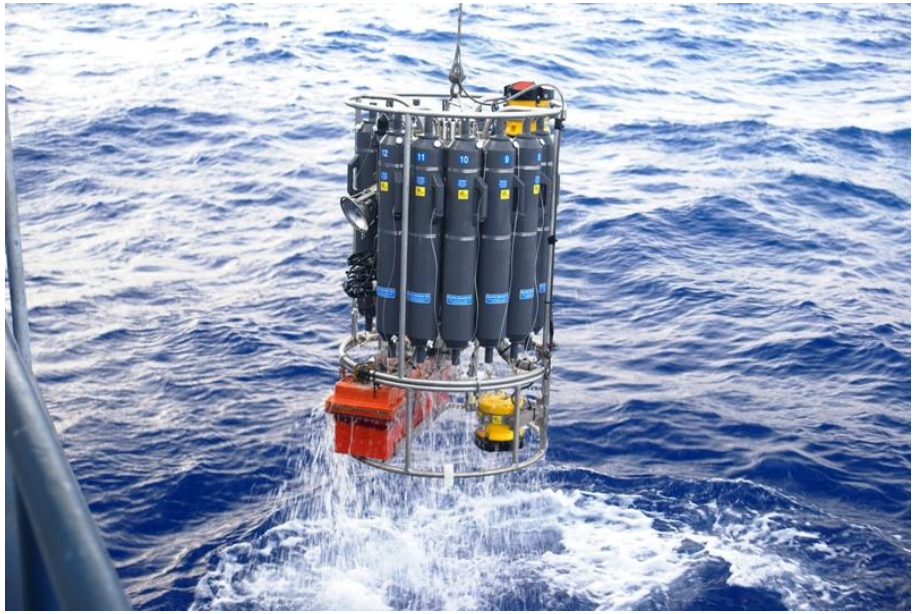

Saudi Arabia
Africa
2,200 m
Discovery
Kebrit
Brine pool
Atlantis II

The brine pools are a special niche

- High salinity (10X more than surface water)
- Enriched with heavy metals: iron, manganese, copper, zinc (1000X more concentrated than normal water)
- High temperatures (70 °C)
- High pressure
- No light

The Oceanus research ship belongs to the Woods Hole Oceanographic Institute.

Source: Hamza El Dorry



A CTD: Conductivity, Temperature, and Depth (CTD) sensors

A CTD determines the essential physical properties of ocean water. It gives scientists a precise and comprehensive charting of the distribution and variation of water temperature, salinity, and density that helps to understand how the oceans affect life. (Media Relations, Woods Hole Oceanographic Institution)

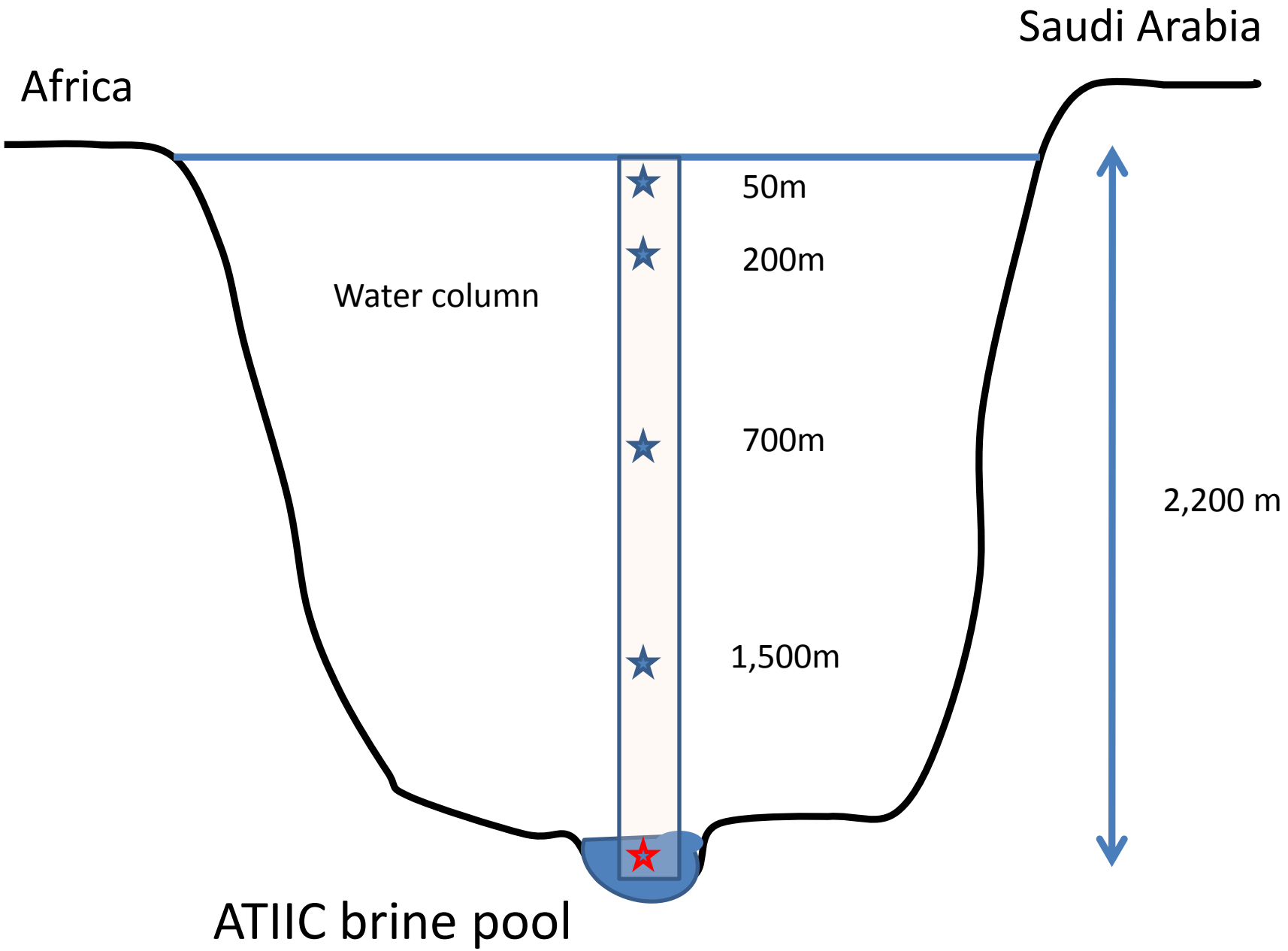
As the ship steams slowly ahead, scientists in a lab on the ship will guide the CTD up and down in the water column, occasionally sending the instrument an electronic signal to collect a water sample in a bottle mounted on the instrument's cage. (© C.A. Linder, WHOI)

One of the labs on **OCEANUS**

Source: Hamza El Dorry

Data

- Pyrosequencing with Roche 454 (AUC)
- 2 columns (above two different brine pools)
  - 5 samples in each column
- **Esta aula**
  - data for the column above the brine pool **Atlantis II** (**ATIIC**)


Saudi Arabia
Africa
50m
200m
Water column
700m
2,200 m
1,500m
ATIIC brine pool

First results

- Comparison among several ocean water columns
  - 11 locations worldwide
  - 24 samples at different depths

Location of the 11sites and number of sequences of the 24 data sets

Figure S1

Depth (m)	**North Pacific Gyre** ALOHA Station, Hawaii, USA	**Sargasso Sea** BATS Station, Bermudas	**Red Sea** Atlantis II Basin	**Southern Pacific** Station 3, Coast of Iquique, Chile	**Mediterranean Sea** Coast of Alicante, Spain	**Sea of Marmara** Central Basin, Turkey	**Atlantic Ocean, Puerto Rico Trench** North of Puerto Rico
20		334					
25	581						
50		410	**1100**	292	1204		
75	645						
85				529			
100		484					
110	454			334			
200			**655**	453			
500	954	158					
700			**654**			253	
1000							
1500			**1138**				
4000	116						
6000							545

Sequences, x10³

GOS (surface water):	**Caribbean Sea** (GS018)	**Eastern Tropical Pacific** (GS023)	**Galapagos Islands** (GS034)	**Indian Ocean** (GS114)
Sequences x10³	143	133	134	349

Comparison: gene functions

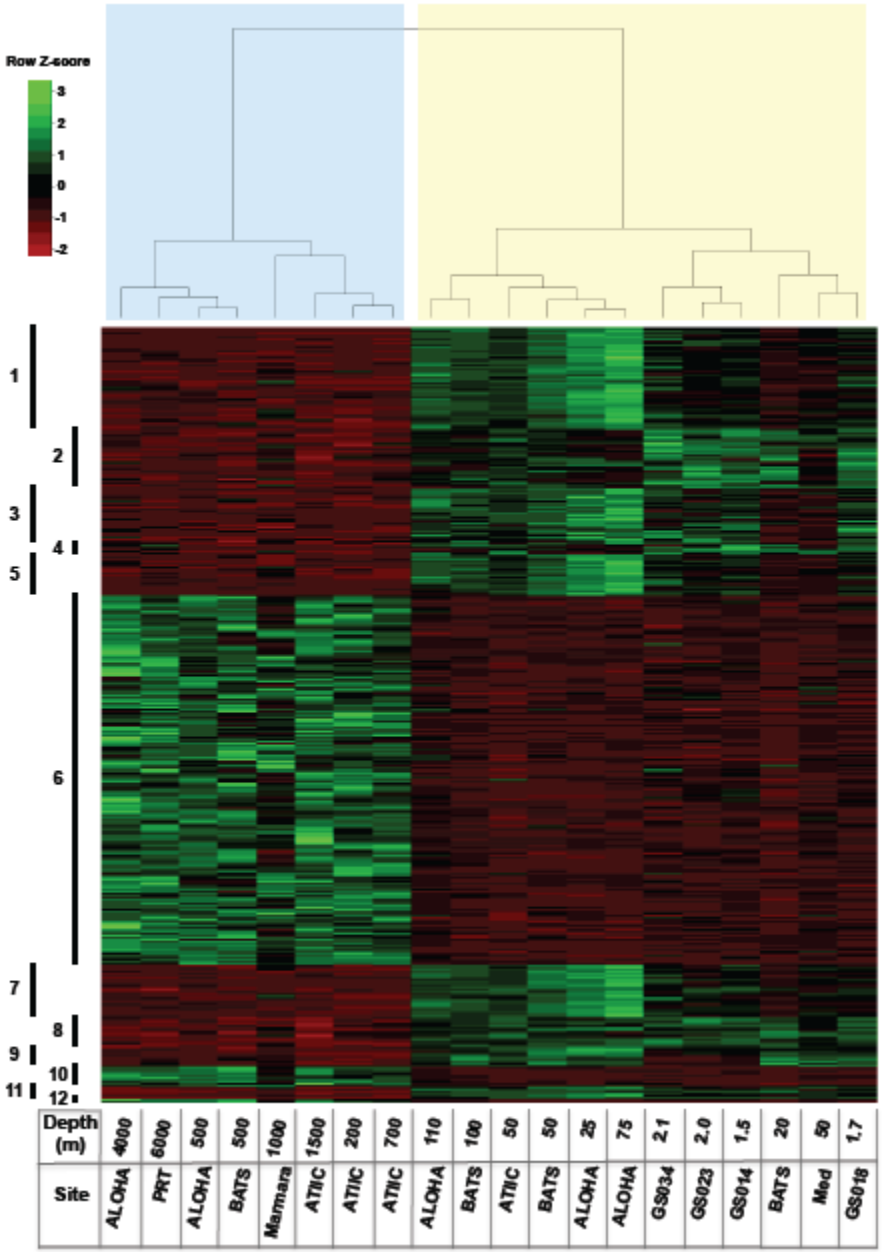
- Based on COG assignments
  - BLASTX against eggNOG [Jensen et al. 2008]
- What functions were present in one site but not in others
- What functions were present at a certain depth but not in others
- Rather than presence/absence
  - **Differentially represented COGs**
  - Similar to differentially expressed genes
  - Heat maps, hierarchical clustering

Methodological issues

- Comparative metagenomics
- How to determine whether an assigned COG is differentially represented
  - Normalization, statistics

First result: COGs per water column

site	#COGs
ATIIC	483
ALOHA	337
BATS	360
Iquique	174
Total unique	**790**



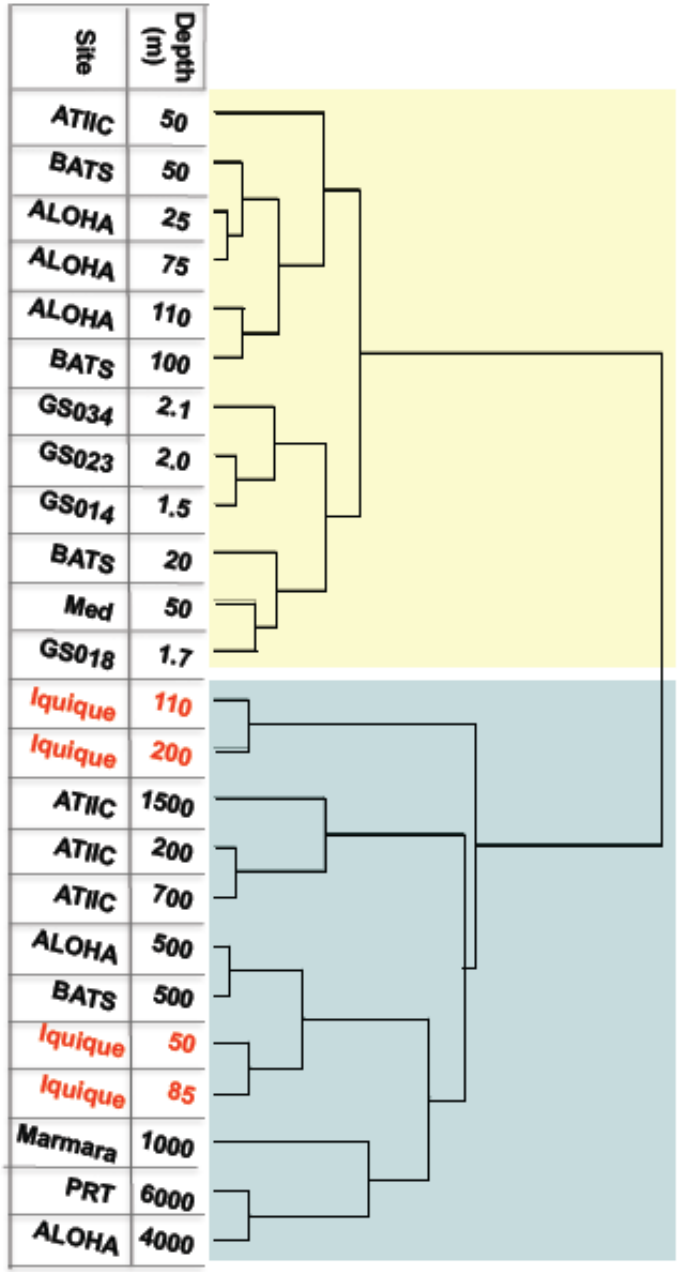
Based on 386 COGs shared by ATIIC, Aloha, BATS with differential representation

COGs

Iquique not included

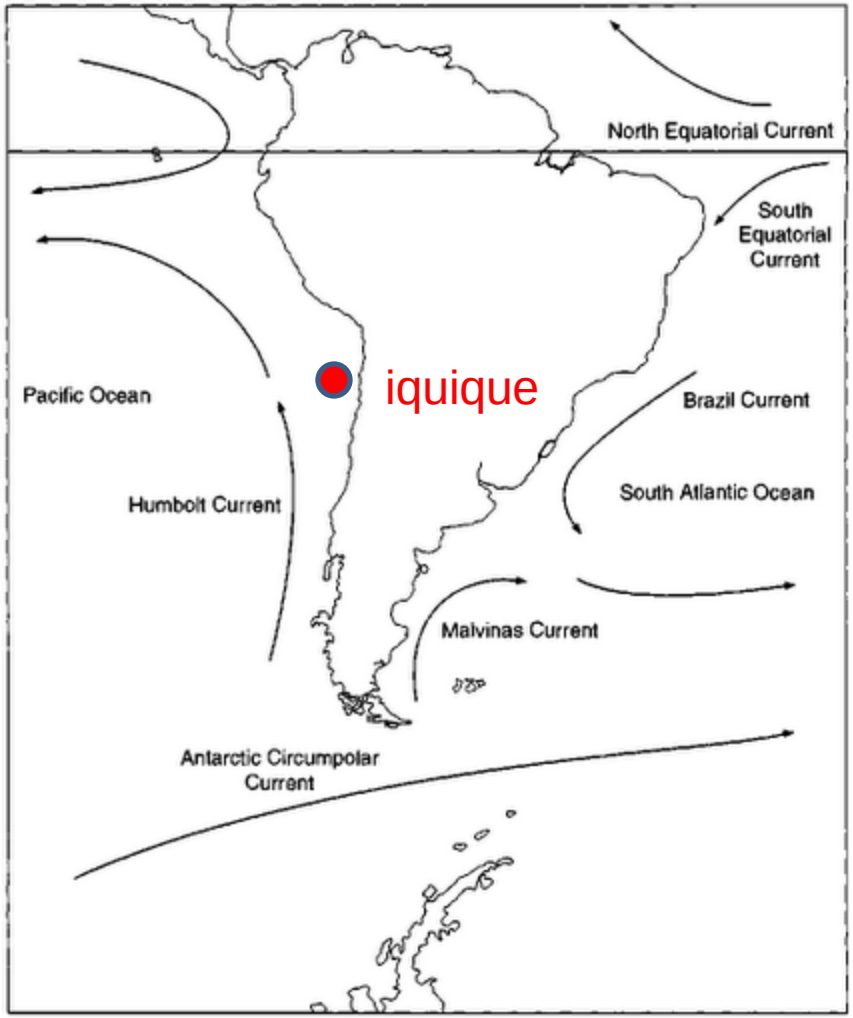
And Iquique?

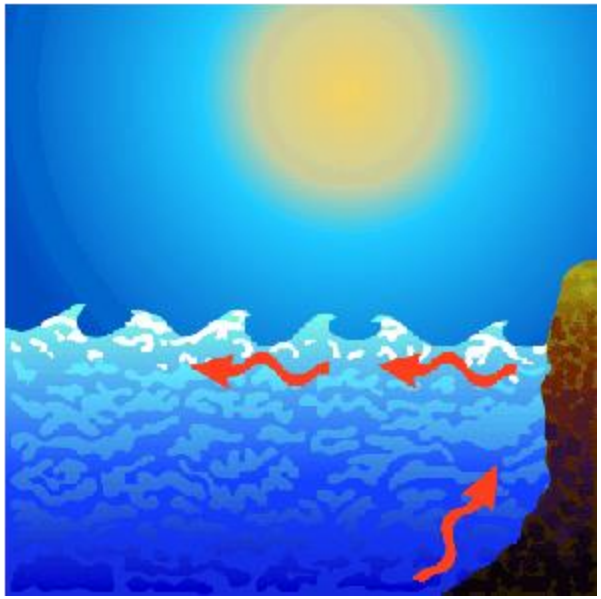
Figure 2A


Site
Depth (m)
ATlIC 50
BATS 50
ALOHA 25
ALOHA 75
ALOHA 110
BATS 100
GS034 2.1
GS023 2.0
GS014 1.5
BATS 20
Med 50
GS018 1.7
Iquique 110
Iquique 200
ATlIC 1500
ATlIC 200
ATlIC 700
ALOHA 500
BATS 500
Iquique 50
Iquique 85
Marmara 1000
PRT 6000
ALOHA 4000

Why is Iquique different?

- Humboldt current
- Upwelling phenomenon
  - Cold nutrient-rich water is taken to the surface





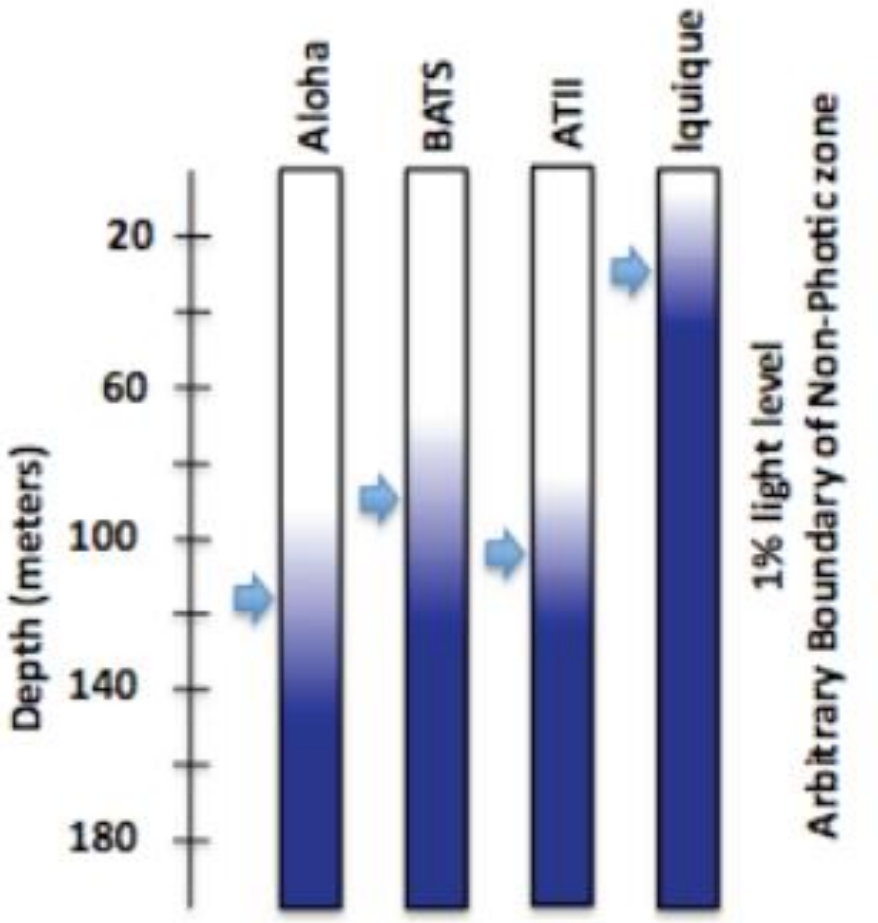
upwelling

http://mynasadata.larc.nasa.gov/glossary.php?&word=upwelling

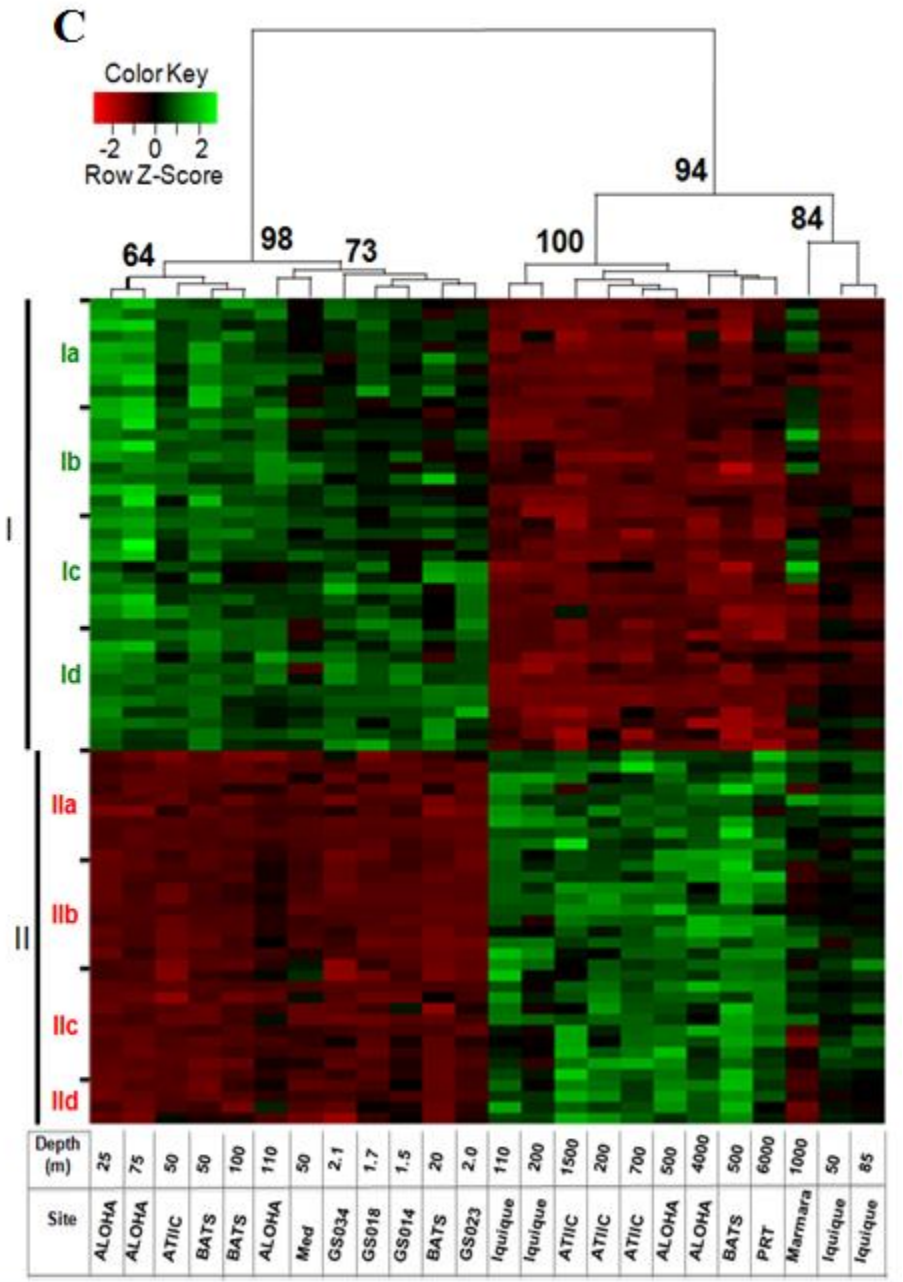
http://what-when-how.com/marine-mammals/south-american-aquatic-mammals

Figure 2B

PAR values (Photosynthetically active radiation)



“Photic and aphotic COGs”



41 COGs with higher abundance in photic zones

34 COGs with higher abundance in aphotic zones

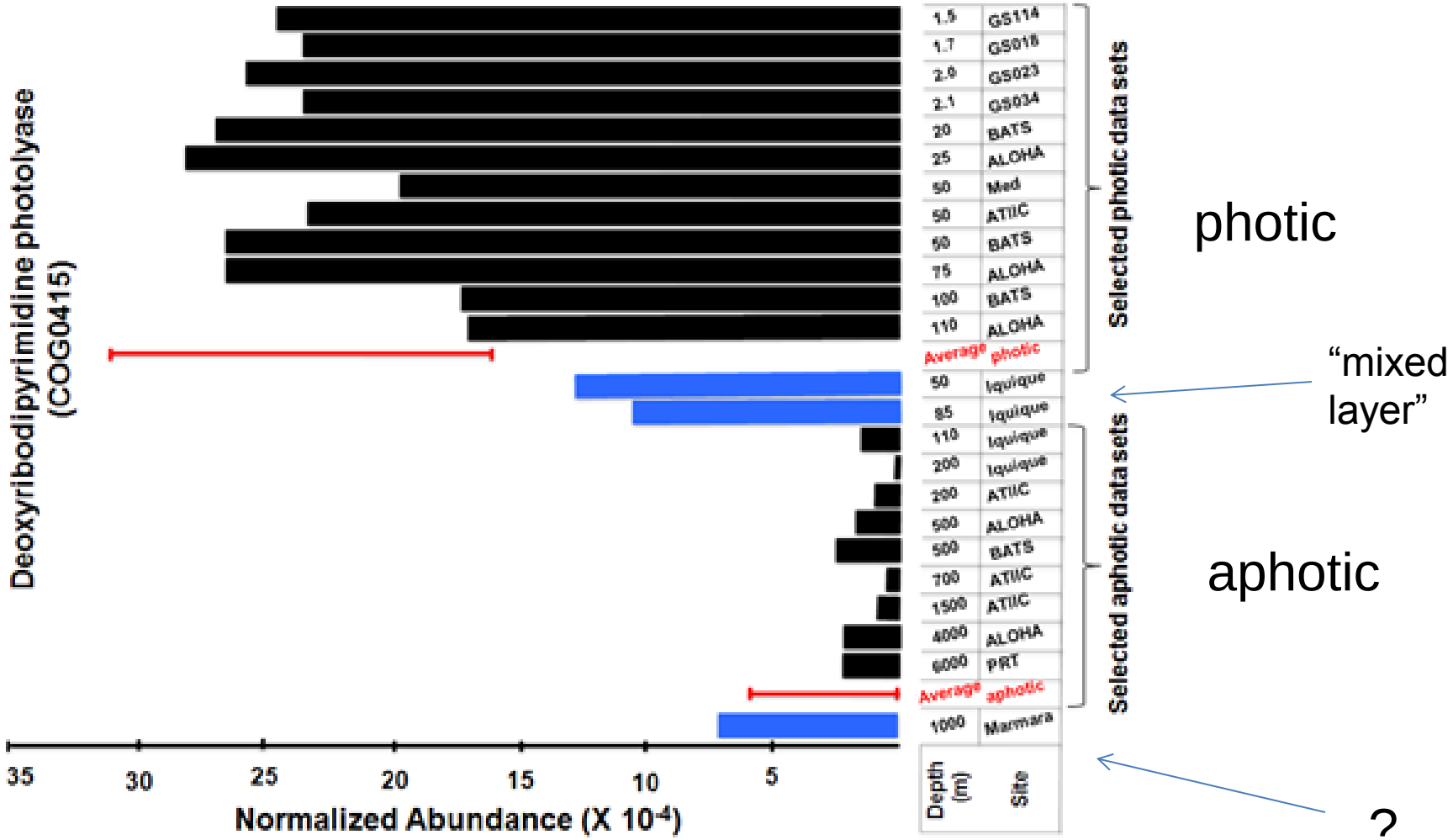
Photic COGs

- Photosyntesis
- biosynthesis of light-harvesting pigments
- assimilation of CO_2 by photosynthetic bacteria
- Light-induced DNA repair
- oxidative stress response
- N_2 fixation
- phosphate metabolism

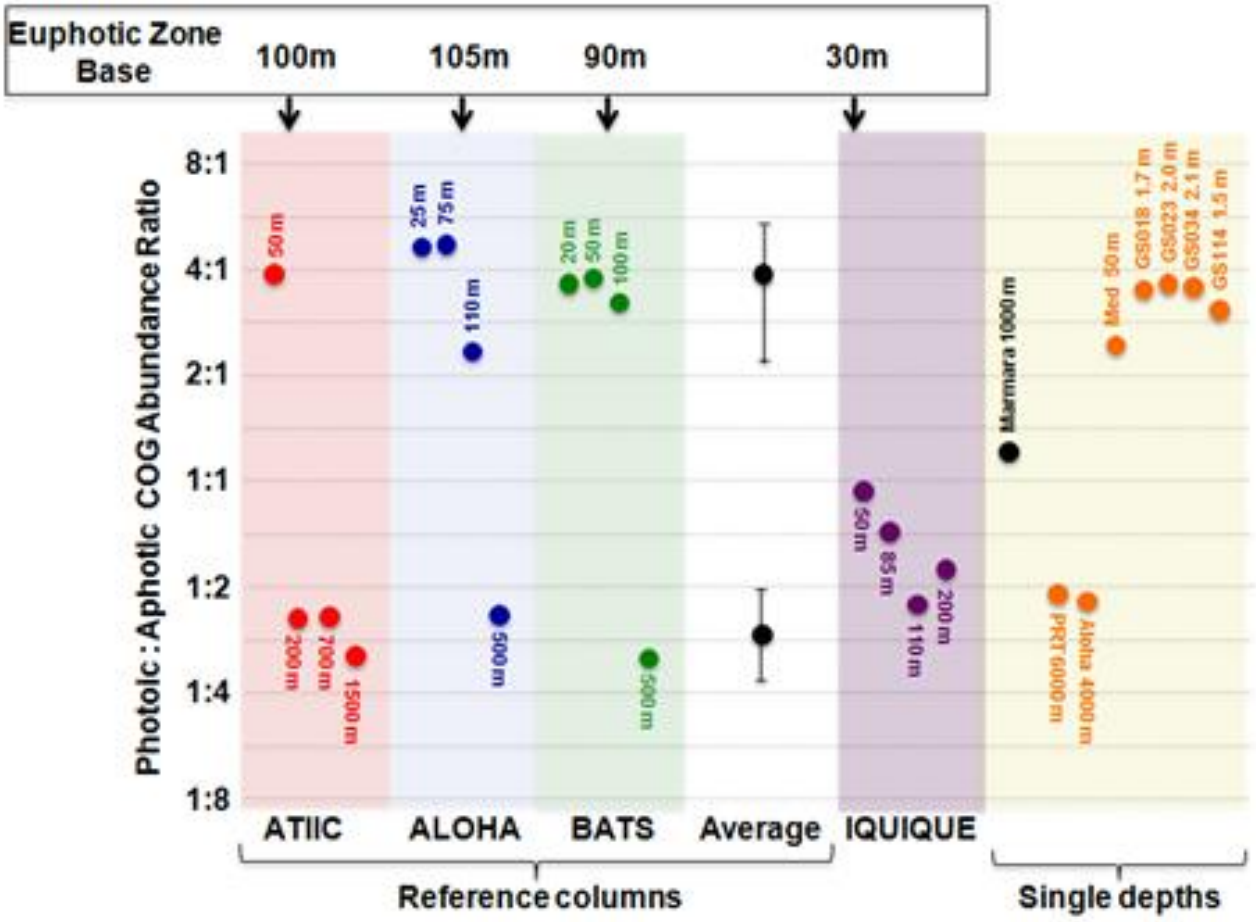
Aphotic COGs

- Catabolism of proteins and aminoacids
- Methane oxidation
- sulfate assimilation and metabolism
- selenocysteine metabolism
- terpenoid biosynthesis

Deoxyribodipyrimidine photolyase (repairs DNA damage caused by exposure to ultraviolet light, COG0415)



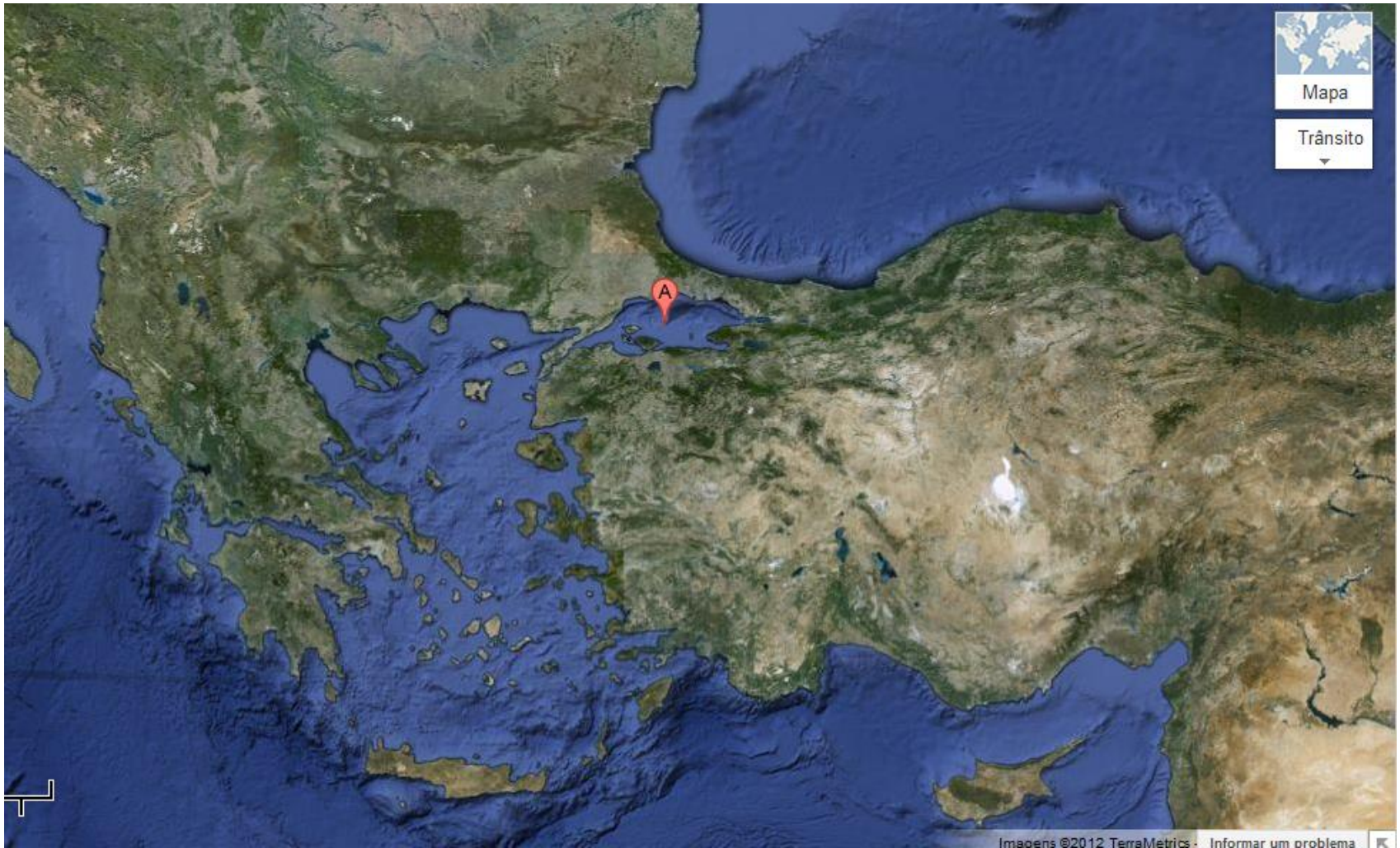
Abundance ratios



Another puzzle: Marmara

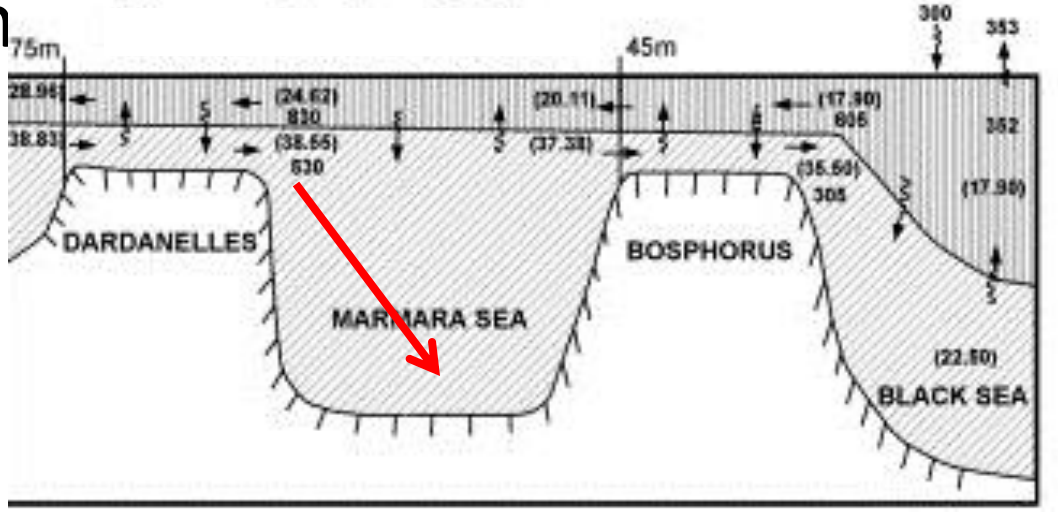
- Relatively high ratio photic:aphotic COGs for the depth
- Overrepresentation of the photolyase COG

Sea of Marmara



Conjectured explanation

- The Marmara sea receives saltier water inflow from the Mediterranean through the Dardan



- This water is denser therefore it sinks

http://www.sciencedirect.com/science/article/pii/S0034666703001179 Mudie et al., 2004

Oxygen limitation

- Hypoxic regions in Iquique
  - 85, 110, 200m
  - Lowest O_2 values among all samples
- Are there COGs significantly associated with lack of oxygen?
- We removed 383 photic/aphotic COGs and verified the rest
- Result: 22 differentially represented COGs (11 up, 11 down)

Row Z-Score
4
2
0
-2
COG ID
COG0397
COG1054
COG0366
COG2124
COG0376
COG2109
COG3670
COG2802
COG1201
COG1236
COG1463
COG2181
COG1014
COG1924
COG1775
COG1251
COG3243
COG2223
COG1140
COG2963
COG5013
COG2801
Normalized Abundance
Depth (m)
50 500 500 20 50 110 100 2.0 1.5 1.7 2.1 75 50 50 25 1500 6000 4000 1000 200 700 85 110 200
Oxygen (µmol.kg-1)
184 157 184 203 304 216 219 >200 >200 >200 >200 226 238 213 206 115 240 148 30 73 81 11 4.0 3.2
Site
Iquique ALOHA BATS BATS Med ALOHA BATS GS023 GS014 GS018 GS034 ALOHA ATIIC BATS ALOHA ATIIC PRT ALOHA Marmara ATIIC ATIIC Iquique Iquique Iquique

Gene favored by low oxygen levels: *narG*, involved in denitrification

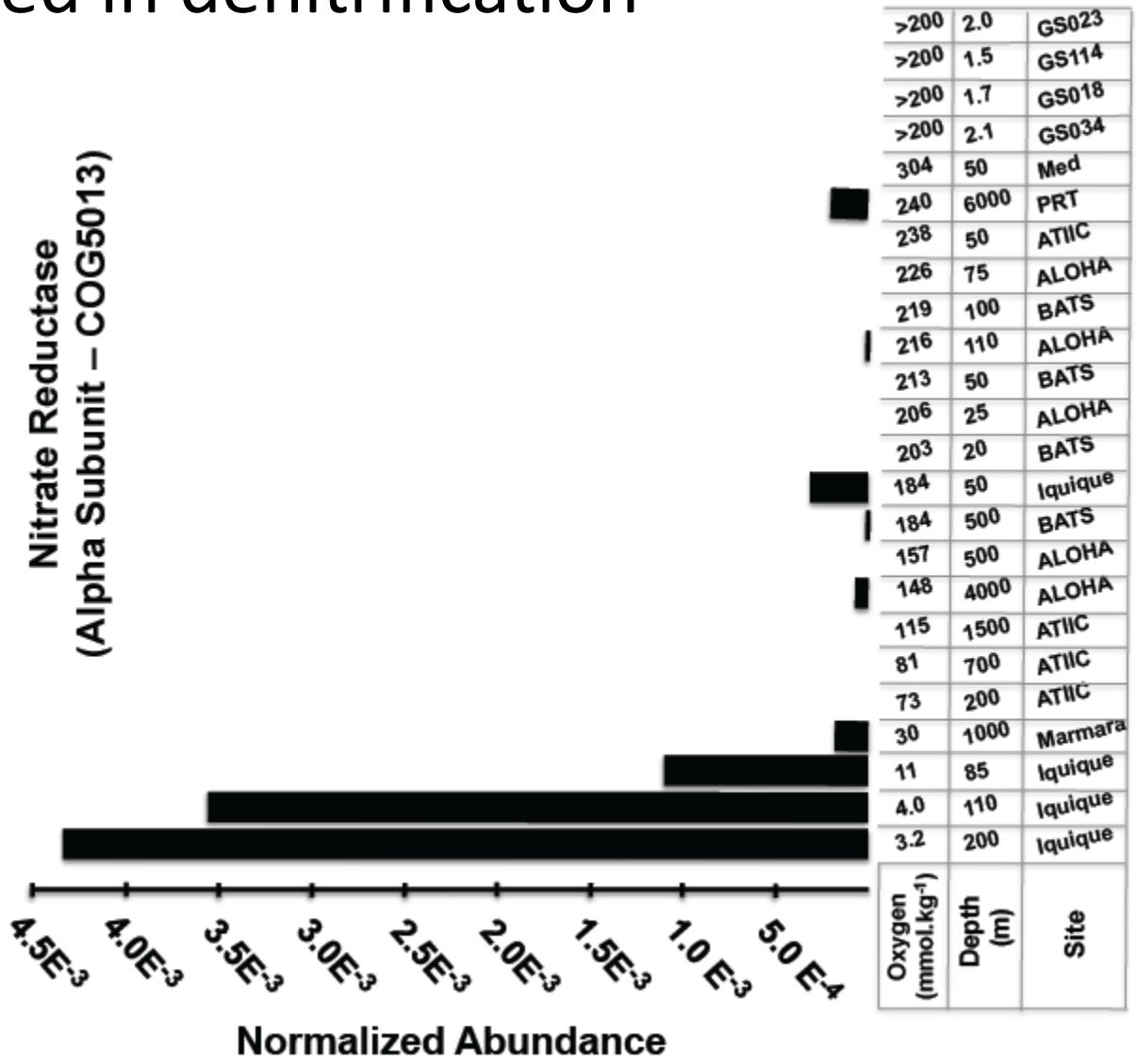


Figure 4B

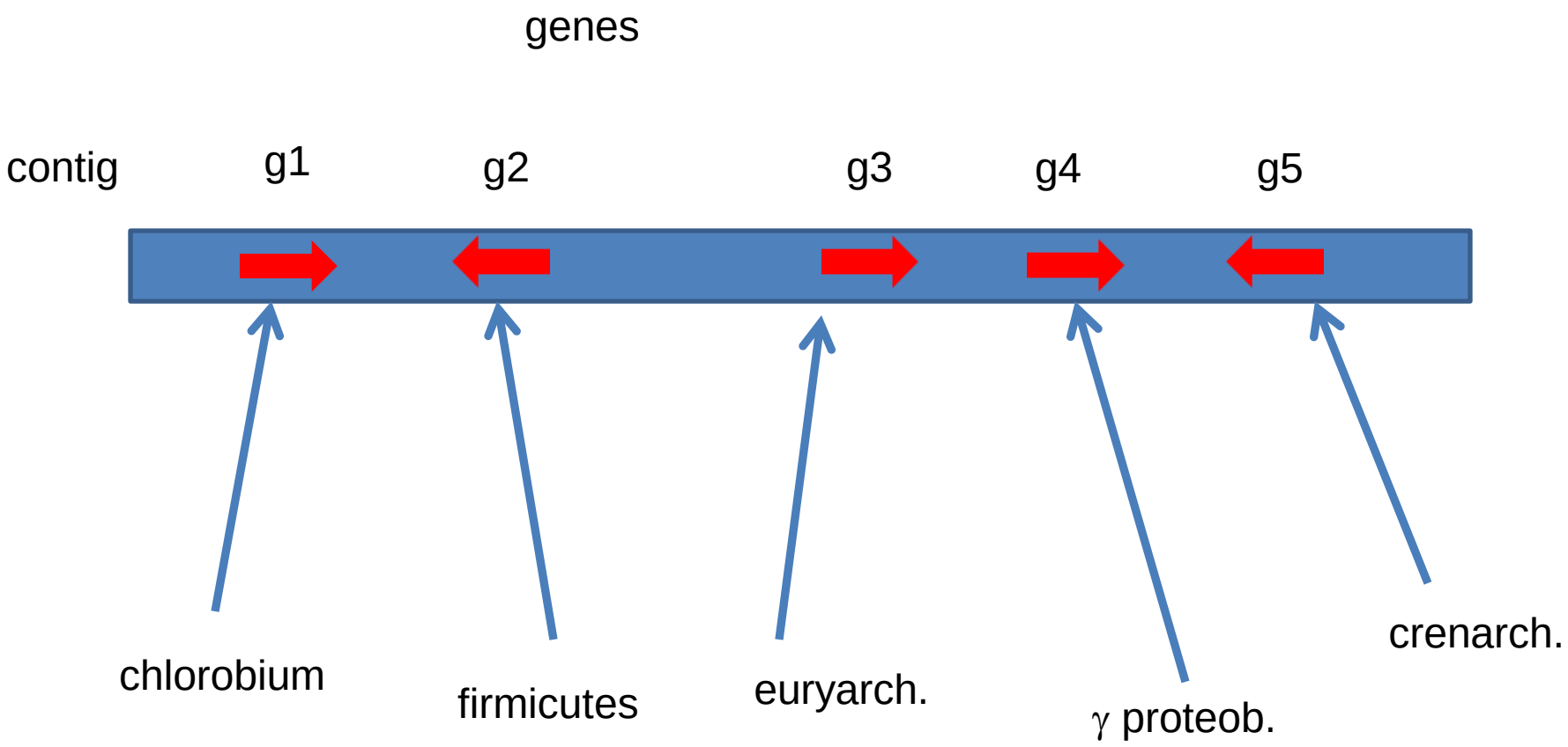
Summary

- A metagenomics-enabled functional profiling analysis
  - Light
  - Oxygen
- Reference sets of functional biological activities
  - diagnosis of the physiological and biochemical capabilities of marine microorganisms
- May help monitor “the oceans’ health”

Montagem

- Em genomas bacterianos isolados, é um processo razoavelmente bem compreendido
- Em metagenomas há velhas e novas dificuldades
  - Mistura de organismos
    - Quimeras
    - Transferência lateral
  - Repetições
  - Tamanho dos conjuntos de dados
    - Chegando a bilhões de reads

Exemplo de quimerismo



Reads vs. contigs

- Reads: o que sai da máquina sequenciadora
- Contigs: resultado da montagem

Reads

- Essencial usar reads para análises de abundância
- Também é melhor usar reads para identificação taxonômica por causa do possível quimerismo

Contigs

- Bons para achar genes
  - Mais provável achar ORFs completas em contigs
- Para simples presença de genes quimerismo não é um problema sério

Montagem de Metagenomas

Algoritmos de montagem especializados para metagenômica

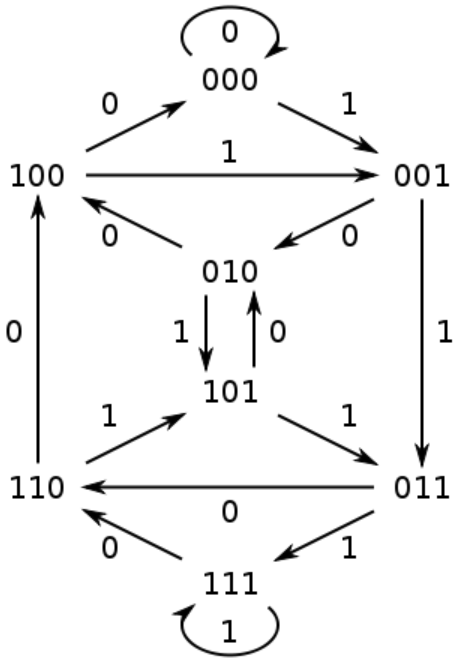
- Genovo [Laserson, Jojic, Koller 2011]
- Metavelvet [Namiki et al. 2012]
- Differential-coverage binning [Albertsen et al. 2013]

genovo

- Algoritmo probabilístico
- Procura a montagem mais provável para um dado conjunto de reads
- Melhor que newbler
- Muito lento! (3 dias x 2 horas [newbler])

metavelvet

- Baseado em velvet
- Velvet é baseado em grafos de de Bruijn



Sobreposição de *k*-mers

$k = 1$

http://chessprogramming.wikispaces.com/De+Bruijn+sequence

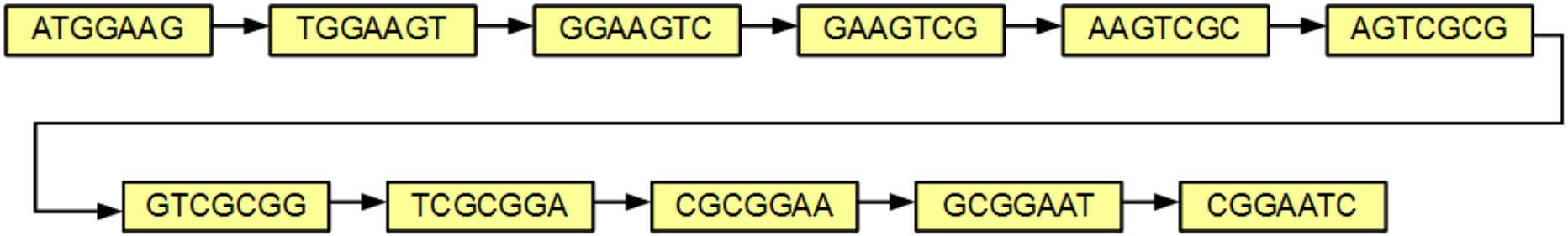
Grafo de de Bruijn em montagem

sequence **ATGGAAGTCGCGGAATC**

7mers

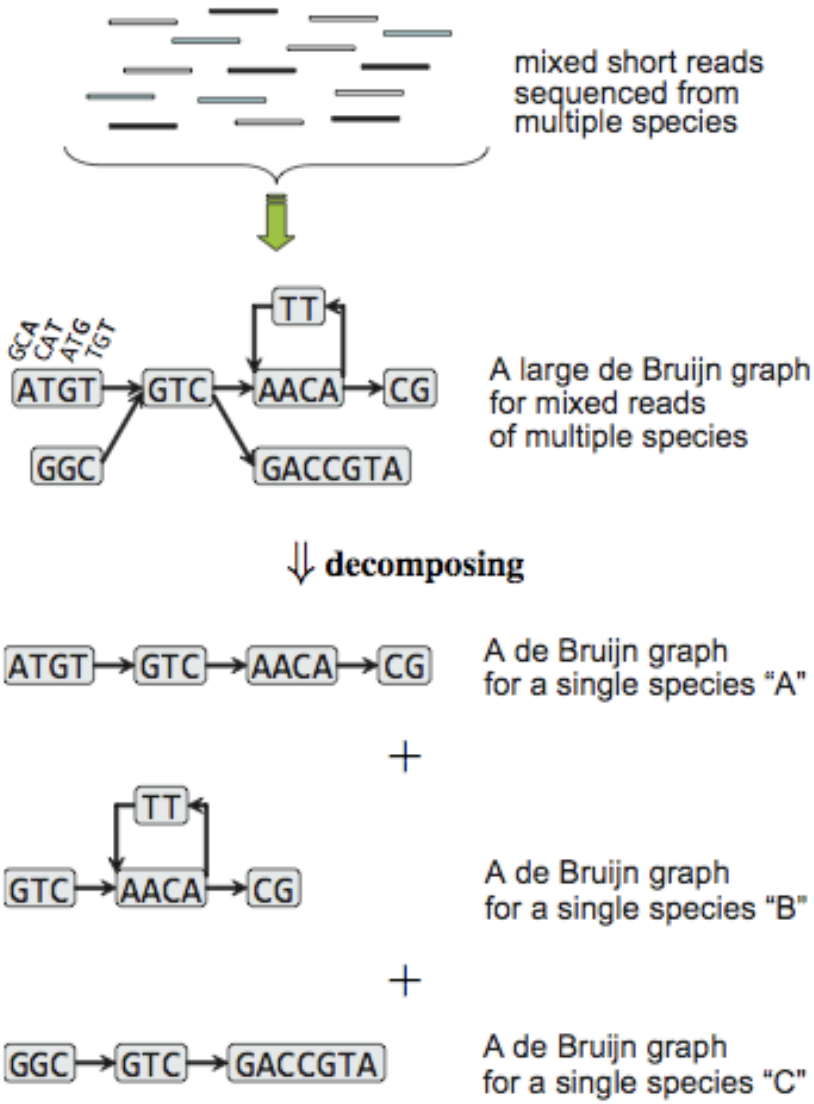
ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

de Bruijn graph



http://www.homolog.us/blogs/wp-content/uploads/2011/07/i6.png

metavelvet



Permite escolha de mais de um valor para *k*

http://metavelvet.dna.bio.keio.ac.jp/

Na prática, no meu lab

- SoapDeNovo
  - Mais rápido, consegue dar conta de conjuntos grandes de dados, resultados suficientemente bons
  - Luo et al.: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 2012 1:18.

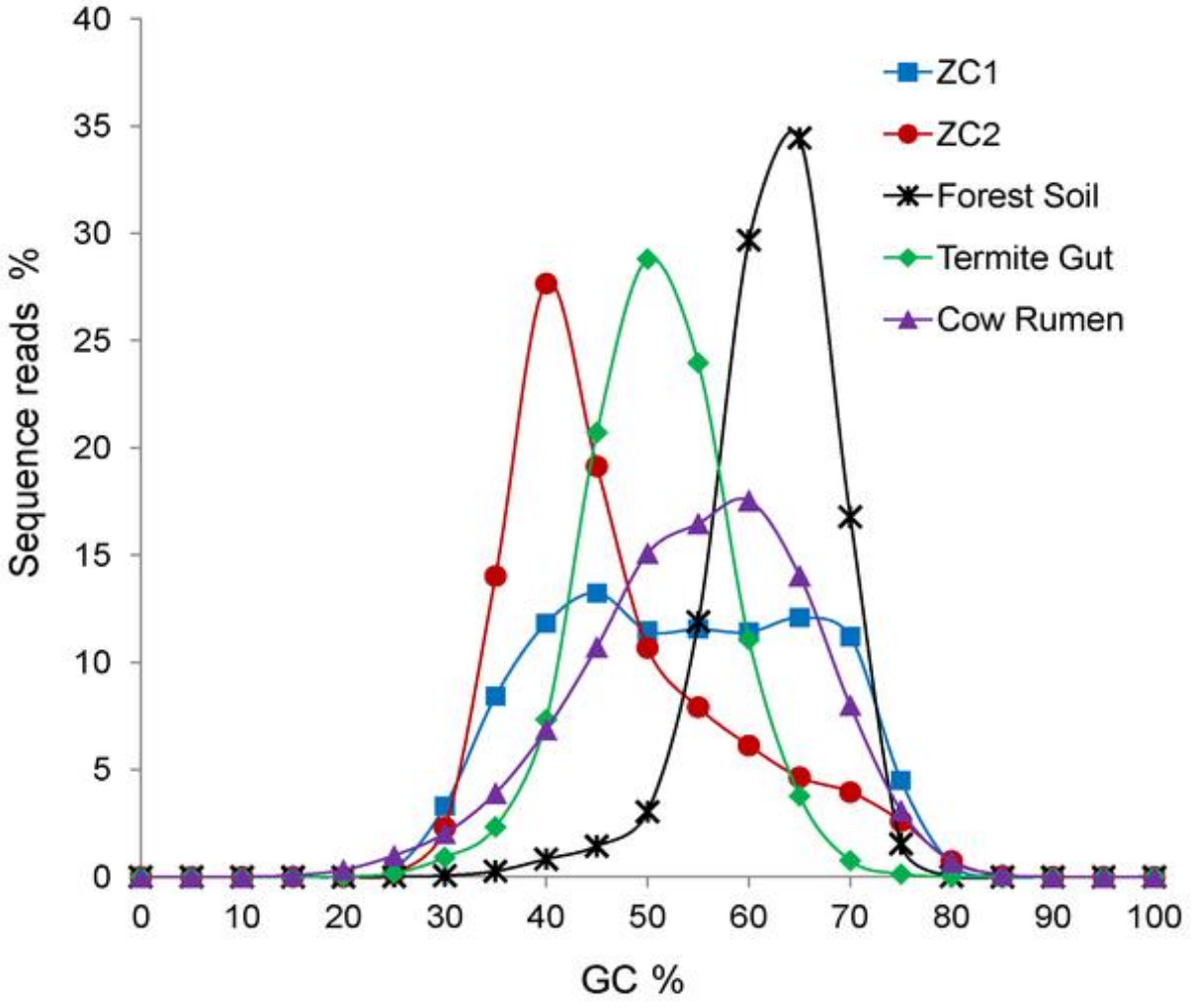
Anotação funcional

- Pipeline para genomas completos pode ser usado
  - Exemplo: IMG/M
- Problema: maioria das ORFs são parciais
  - Dificulta atribuição de função
    - Potencial gerador de erros

Comparação de metagenomas

- Genomicamente
- Taxonomicamente
- Funcionalmente
- Recursos oferecidos pelo IMG/M

Figure 1. Distribution of the GC content percentage for ZC1 and ZC2 compared with selected metagenomes.



Martins LF, Antunes LP, Pascon RC, de Oliveira JCF, Digiampietri LA, et al. (2013) Metagenomic Analysis of a Tropical Composting Operation at the São Paulo Zoo Park Reveals Diversity of Biomass Degradation Functions and Organisms. PLoS ONE 8(4): e61928. doi:10.1371/journal.pone.0061928
http://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.0061928

Genome clustering (IMG/M)

Clustering Type:

By Function:

- (•) COG
- ( ) Pfam
- ( ) KO

By Taxonomy:

- ( ) Class
- ( ) Family
- ( ) Genus

By Function Category:

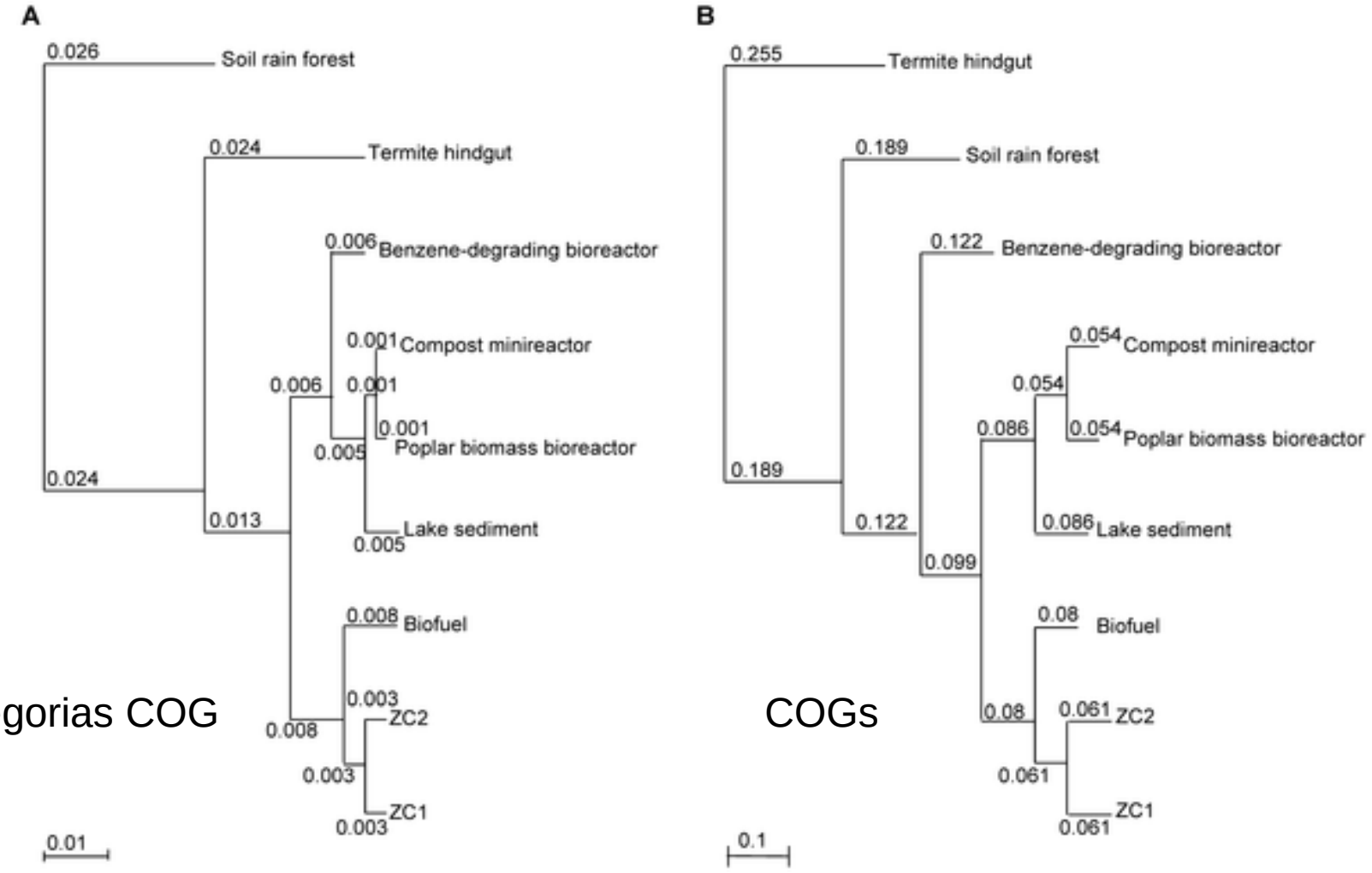
- ( ) COG Categories
- ( ) COG Pathways
- ( ) KEGG Pathway Categories (KO)
- ( ) KEGG Pathway Categories (EC)
- ( ) KEGG Pathways (KO)
- ( ) KEGG Pathways (EC)
- ( ) Pfam Categories

Clustering Method:

- (•) Hierarchical Clustering
- ( ) Principal Components Analysis (PCA)
- ( ) Principal Coordinates Analysis (PCoA)
- ( ) Non-metric MultiDimensional Scaling (NMDS)
- ( ) Correlation Matrix

Go Reset

Figure 8. Hierarchical clustering of functional gene groups of ZC1 and ZC2 and seven public metagenomes.



Categorias COG

COGs

Martins LF, Antunes LP, Pascon RC, de Oliveira JCF, Digiampietri LA, et al. (2013) Metagenomic Analysis of a Tropical Composting Operation at the São Paulo Zoo Park Reveals Diversity of Biomass Degradation Functions and Organisms. PLoS ONE 8(4): e61928. doi:10.1371/journal.pone.0061928
http://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.0061928

Normalização

- É um assunto que requer competência em estatística
- A seguir são apresentadas duas regras práticas
- Motivação: Metagenômica comparativa requer normalização do número de reads das amostras
- Variação pode ser grande; por exemplo:
  - Amostra 1: 200 mil reads
  - Amostra 2: 2 milhões de reads

Método 1

- Determinar a amostra com menor número de reads (suponha n_{min})
- Para cada outra amostra
  - Selecionar aleatoriamente n_{min} reads
- Desvantagem
  - Pode jogar muito dado fora

Método 2

- Seja n_A o numero de reads da amostra *A*
- Seja σ o número médio de reads por amostra
- Seja *x* a contagem da característica de interesse em *A* (ou seja, *x* reads tem essa característica)
- Então normalizamos *x* pela fórmula

$$-\log_{10}\left[\left(\frac{x}{n_A}\right).\sigma + 1\right]$$

A map of diversity in the human microbiome


Streptococcus dominates the oral cavity with *S. mitis* > 75% in the cheek

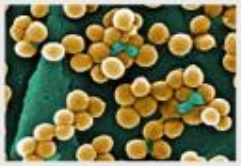
Propionibacterium acnes lives on the skin and nose of most people

Many *Corynebacterium* species characterize different body sites:
C. matruchoti the plaque
C. accolens the nose
C. croppenstedtii the skin

Lactobacillus species (*L. gasseri*, *L. jensenii*, *L. crispatus*, *L. iners*) are predominant but mutually exclusive in the vagina


Several *Prevotella* species are present in the gastrointestinal tract. *P. copri* is present in 19% of the subjects and dominates the intestinal flora when present

Staphylococcus epidermidis colonizes external body sites



Bacteroides is the most abundant genus in the gut of almost all healthy subjects

Campylobacter includes opportunistic pathogens, but members live in the oral cavities of most healthy people in the cohort


E. coli is present in the gut of the majority of healthy subjects but at very low abundance

Phylogenetic map of the microbial species detected in the healthy human microbiome

Commensal microbes

Potential pathogens

The four most abundant phyla

- Actinobacteria
- Bacteroidetes
- Firmicutes
- Proteobacteria

Low abundance phyla

- Chloroflexi
- Cyanobacteria
- Euryarchaeota
- Fusobacteria
- Lentisphaerae
- Spirochaetes
- Synergistetes
- Tenericutes
- Thermi
- Verrucomicrobia

National Institutes of Health Human Microbiome Project

N. Segata & C. Huttenhower
http://huttenhower.sph.harvard.edu

http://huttenhower.sph.harvard.edu/metaphlan

Platformas web de processamento

- Laboratórios governamentais
- Serviços padronizados de processamento

MG-RAST

metagenomics analysis server

LOGIN

Browse Metagenomes

search for 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	77,307
# base pairs	25.81 Tbp
# of sequences	236.94 billion
# of public metagenomes	12,527

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 77,307 data sets. The current server version is 3.3.3.3. We suggest users take a look at MG-RAST for the impatient.

Updates | MG-RAST 3.2.4 release notes [October 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

loaded.

Microbiome Details (Assembled Data)

Add to Genome Cart | Browse Genome | BLAST Genome

About Genome

- Overview
- Statistics
- Genes

Overview

Proposal Name	Sao Paulo Zoo Compost
Sample Name	Sample C4
Taxon Object ID	2156126000
IMG Submission ID	2671
GOLD ID in IMG Database	Project Id: Gm0002180
External Links	
Genome type	metagenome
Sequencing Status	Draft
IMG Release	
Comment	
Sample Information	
Sample Site	Sao Paulo Zoo composting operation
Sample Collection Date	January 26, 2011
Isolation Country	Brazil
Sampling Strategy	8 days after composting started
Sample Isolation	done 8 days after composting started
Temperature Range	Thermophile
Sample Assembly Method	newbler
Sample Geographic Location	Sao Pulo Zoo
Longitude	-46.62
Latitude	-23.65

EBI Metagenomics

Easy submission

Manually supported submission process, with help available for meta-data provision. Accepted data formats include SFF (454) and FASTQ (Illumina and IonTorrent).

Find out more

Powerful analysis

Functional analysis of metagenomic sequences using InterPro - a powerful and sophisticated alternative to BLAST-based analyses. Taxonomy diversity analysis is performed using Qiime.

Find out more

Data archiving

Data automatically archived at the Sequence Read Archive (SRA), ensuring accession numbers are supplied - a prerequisite for publication in many journals.

Find out more

Projects

Latest public projects (Total: 37)

Metatranscriptomics of the marine sponge Geodia barretti: Tackling phylogeny and function of its microbial community.
Geodia barretti is a marine cold-water sponge harbouring high numbers of microorganisms. ...
View more - 1 sample

A metagenomics transect into the deepest point of the Baltic Sea reveals clear stratification of microbial functional capacities
The Baltic Sea is characterized by hyposaline surface waters, hypoxic and anoxic deep waters and ...
View more - 6 samples

Gut metagenome in European women with normal, impaired and diabetic glucose control
Type 2 diabetes (T2D) is a result of complex gene–environment interactions, and several risk ...
View more - 147 samples

Samples

Latest public samples (Total: 1053)

Fecal sample from Crohn's patient 1
Fecal sample from Crohn's patient 1 ...
View more - Taxonomy | Function results |

Fecal sample from Crohn's patient 10
Fecal sample from Crohn's patient 10 ...
View more - Taxonomy | Function results |

Fecal sample from Crohn's patient 2
Fecal sample from Crohn's patient 2 ...
View more - Taxonomy | Function results |

Fecal sample from Crohn's patient 3
Fecal sample from Crohn's patient 3 ...
View more - Taxonomy | Function results |

Fecal sample from Crohn's patient 4
Fecal sample from Crohn's patient 4 ...
View more - Taxonomy | Function results |

Data content

1053 public samples (37 public projects)
191 private samples (13 private projects)

News & events

Tweets

Follow @EBImetagenomics

EBI Metagenomics @EBImetagenomics 30 Sep
Check out our new analysis page, using improved data visualisation (Google & Krona charts), and with taxonomic info: ebi.ac.uk/metagenomics/
Expand

EBI Metagenomics @EBImetagenomics 8 Aug
The poster we presented at #ISMBECCB is now available at F1000 posters and describes the EBI metagenomics pipeline:

Sugestão de leitura

G&I Genomics & Informatics

pISSN 1598-866X eISSN 2234-0742
Genomics Inform 2013;11(3):102-113
http://dx.doi.org/10.5808/GI.2013.11.3.102

REVIEW ARTICLE

Analytical Tools and Databases for Metagenomics in the Next-Generation Sequencing Era

Mincheol Kim[1], Ki-Hyun Lee[1], Seok-Whan Yoon[1], Bong-Soo Kim[2], Jongsik Chun[1,2], Hana Yi[3,4,5*]

[1]School of Biological Sciences & Institute of Bioinformatics (BIOMAX), Seoul National University, Seoul 151-742, Korea, [2]Chunlab Inc., Seoul National University, Seoul 151-742, Korea, [3]Department of Environmental Health, Korea University, Seoul 136-703, Korea, [4]Department of Public Health Sciences, Graduate School, Korea University, Seoul 136-703, Korea, [5]Korea University Guro Hospital, Korea University College of Medicine, Seoul 136-703, Korea

Metagenomics has become one of the indispensable tools in microbial ecology for the last few decades, and a new revolution in metagenomic studies is now about to begin, with the help of recent advances of sequencing techniques. The massive data production and substantial cost reduction in next-generation sequencing have led to the rapid growth of metagenomic research both quantitatively and qualitatively. It is evident that metagenomics will be a standard tool for studying the diversity and function of microbes in the near future, as fingerprinting methods did previously. As the speed of data accumulation is accelerating, bioinformatic tools and associated databases for handling those datasets have become more urgent and necessary. To facilitate the bioinformatics analysis of metagenomic data, we review some recent tools and databases that are used widely in this field and give insights into the current challenges and future of metagenomics from a bioinformatics perspective.

Keywords: computational biology, high-throughput nucleotide sequencing, metagenomics