Universidade de São Paulo
Instituto de Química

USP

Análise de Microbiomas

João Carlos Setubal

Os microorganismos estão por toda parte

- São responsáveis por muitos processos fundamentais para a vida do planeta em geral e para a vida dos seres humanos em particular

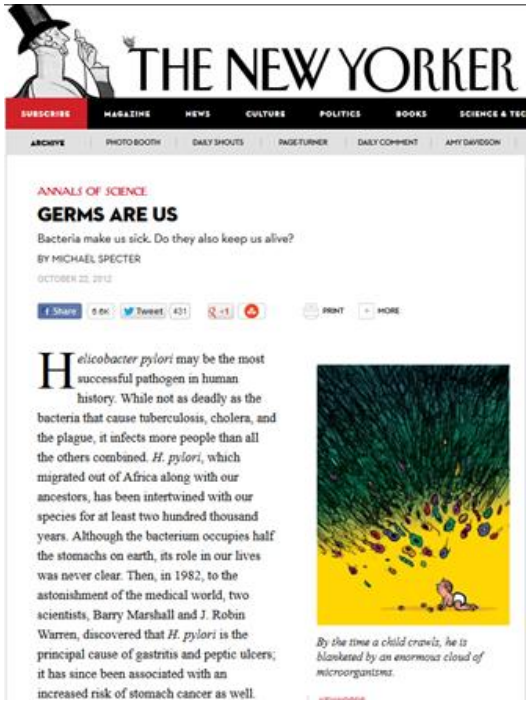
Projeto Microbioma Humano

junho 2012

Science
AAAS

My Microbiome and Me

Science 8 June 2012:


THE NEW YORKER

ANNALS OF SCIENCE

GERMS ARE US

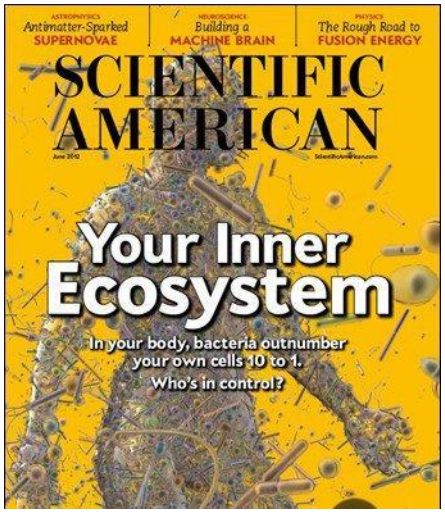
Bacteria make us sick. Do they also keep us alive?

BY MICHAEL SPECTER

Helicobacter pylori may be the most successful pathogen in human history. While not as deadly as the bacteria that cause tuberculosis, cholera, and the plague, it infects more people than all the others combined. *H. pylori*, which migrated out of Africa along with our ancestors, has been intertwined with our species for at least two hundred thousand years. Although the bacterium occupies half the stomachs on earth, its role in our lives was never clear. Then, in 1982, to the astonishment of the medical world, two scientists, Barry Marshall and J. Robin Warren, discovered that *H. pylori* is the principal cause of gastritis and peptic ulcers; it has since been associated with an increased risk of stomach cancer as well.

By the time a child crawls, he is blanketed by an enormous cloud of microorganisms.

outubro 2012



June 2012 Issue

maio 2013

www.earthmicrobiome.org

earth microbiome project

Home | Defining the Tasks | Getting Involved | EMP Protocols and Standards | Affiliations | Publications | Meetings | EMP Logo | No categories

The Earth Microbiome Project is a systematic attempt to characterize the global microbial taxonomic and functional diversity for the benefit of the planet and mankind

Constructing the Microbial Biomap for Planet Earth

The Earth Microbiome Project is a proposed massively multidisciplinary effort to analyze microbial communities across the globe. The general premise is to examine microbial communities from their own perspective. Hence we propose to characterize the Earth by environmental parameter space into different biomes and then explore these using samples currently available from researchers across the globe. We will analyze 200,000 samples from these communities using metagenomics, metatranscriptomics and amplicon sequencing to produce a global Gene Atlas describing protein space, environmental metabolic models for each biome

SEARCH

Meetings

There are currently no EMP centric meetings planned, however we will update this space as soon as the next meeting is organized.

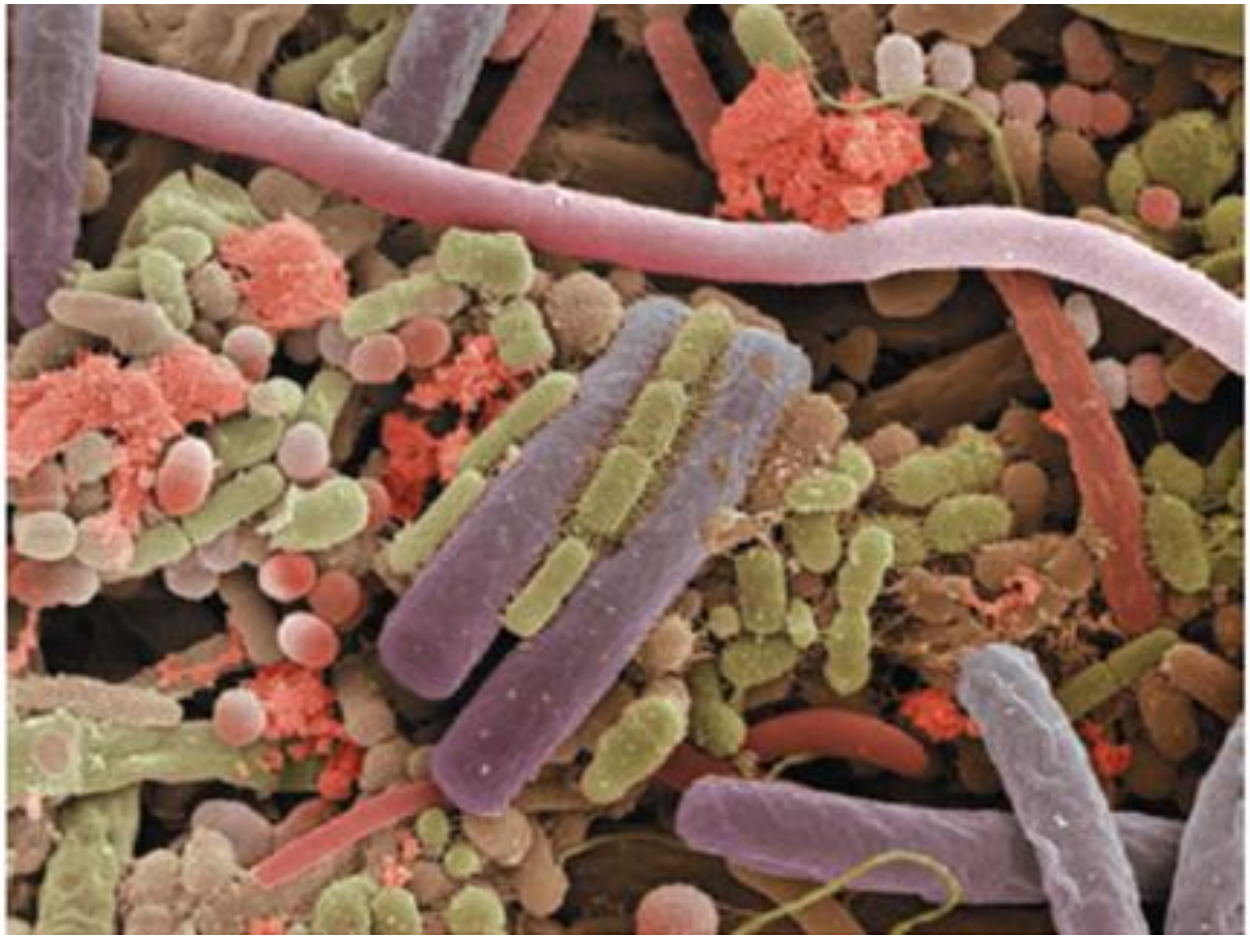
News

Earth Microbiome Project: Rick Stevens at TEDxNaperville

Há uma certa confusão

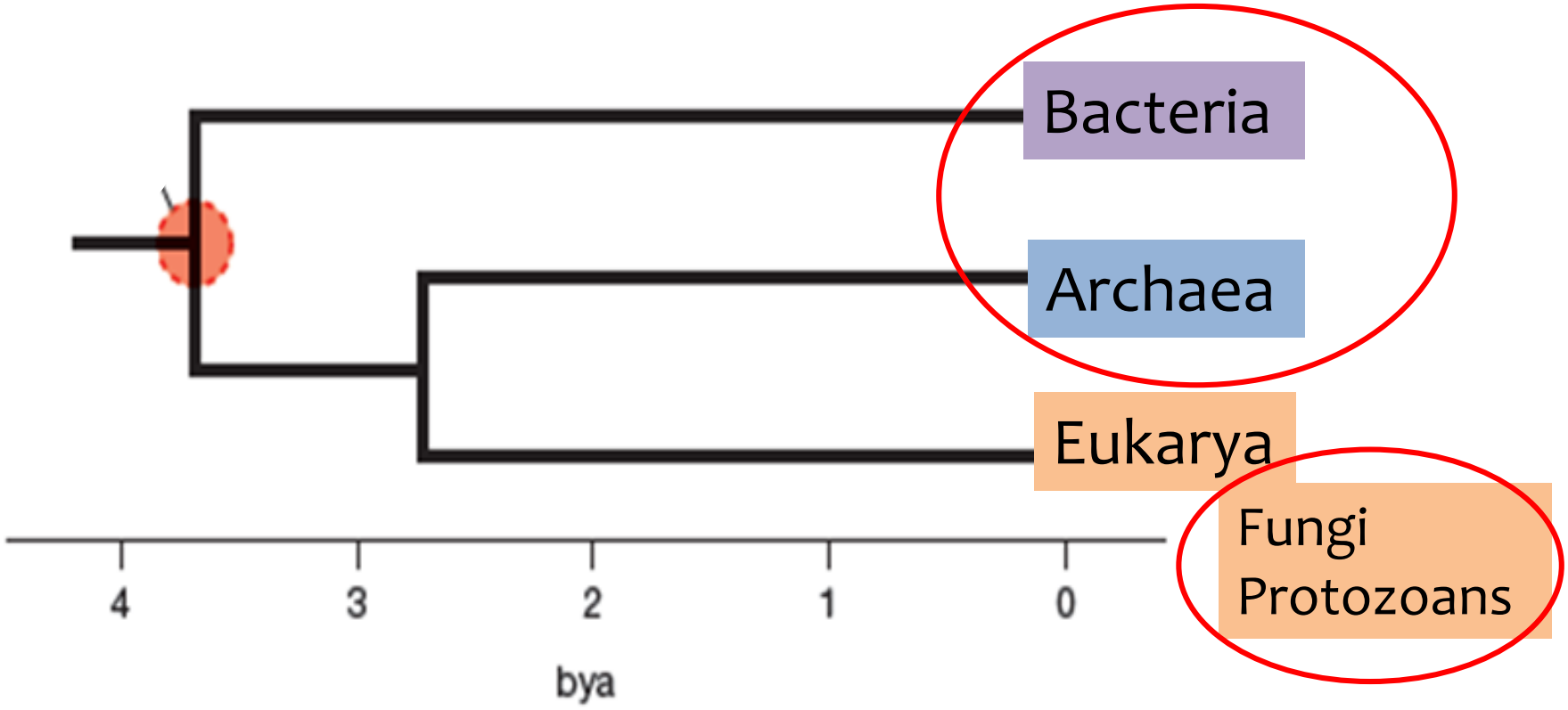
- Earth Biogenome Project (EBP)
- Projeto lançado em 2017 que pretende sequenciar "all life on Earth"
  - voltado para eucariotos

Comunidades microbianas **–Microbiotas–** são típicas de cada ambiente



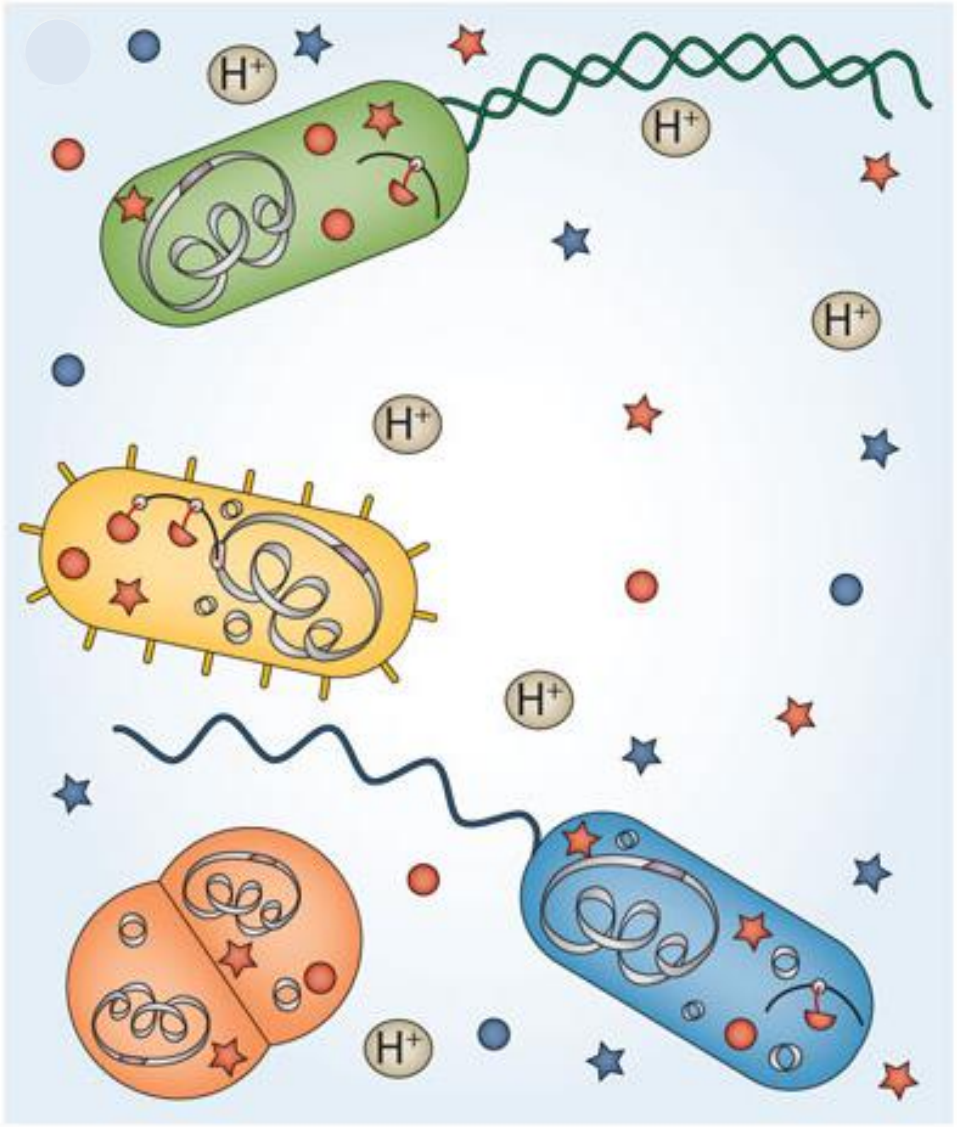
ecossistema microbiano

Microbiotas contêm variedade de microrganismos



\+ Vírus e Bacteriófagos

Microbioma



Genes, Genomas,

Proteínas e Metabólitos da Microbiota

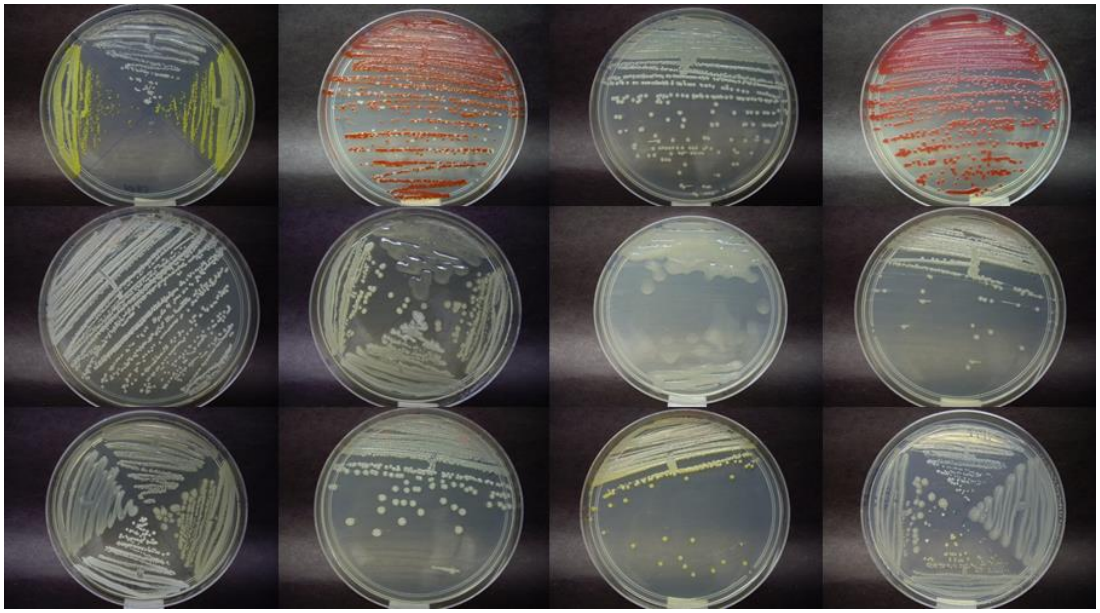
+

Proteínas e Metabólitos da resposta do Hospedeiro à interação com a microbiota

Metabólito da microbiota
Proteína da microbiota
Metabólitos do hospedeiro
Proteínas do hospedeiro

Imagem adaptada de: Whiteside, S. A. *et al.* (2015) The microbiome of the urinary tract—a role beyond infection *Nat. Rev. Urol.* doi:10.1038/nrurol.2014.361

Como acessar essa extraordinária riqueza microbiológica?



Abordagens dependentes de cultivo

Cultivo de bactérias em meio sólido

Imagem: Julio Oliveira

Porém...

A **fração cultivável** da vasta riqueza microbiana da biosfera é muito pequena (estimada em 1%)

Como acessar a extraordinária maioria invisível?

→ Abordagens independentes do cultivo

MetaGenômica

revela as espécies, os genes e genomas de comunidades microbianas

---

MetaTranscritômica

revela os genes expressos (microbiota ativa)

MetaProteômica

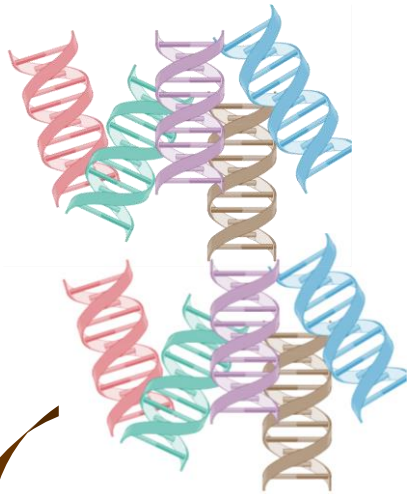
revela as proteínas expressas (microbiota ativa)

Meta-ômicas

MetaGenômica e MetaTranscritômica

Amostra ambiental

Extrair o DNA
(ou RNA)



Sequenciar

Sequenciamento de DNA
alto-desempenho

Analisar as sequências de
DNA: metagenômica
cDNA: metatranscritômica

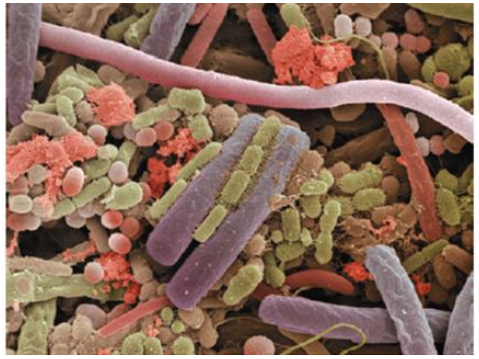
Tecnologias de sequenciamento

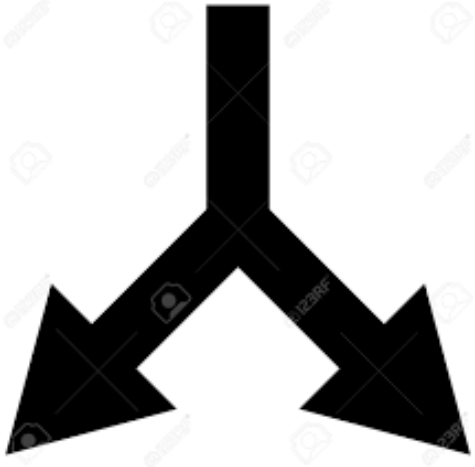
- NGS – next generation sequencing
  - Illumina
    - 90% do mercado
    - Em metagenômica talvez seja perto de 100%
  - PacBio
    - Long reads
  - Nanopore
    - Long reads

Big Data

- Milhões de reads
- Que significa isto?
- Supondo
  - cada read com 300 bp
  - 10 milhões de reads para uma amostra
  - $10 \times 10^6 \times 300 = 3 \times 10^9$ bp
  - Um genoma bacteriano: 5×10^6 bp
  - Equivalente a 600 genomas bacterianos
- A bioinformática é essencial

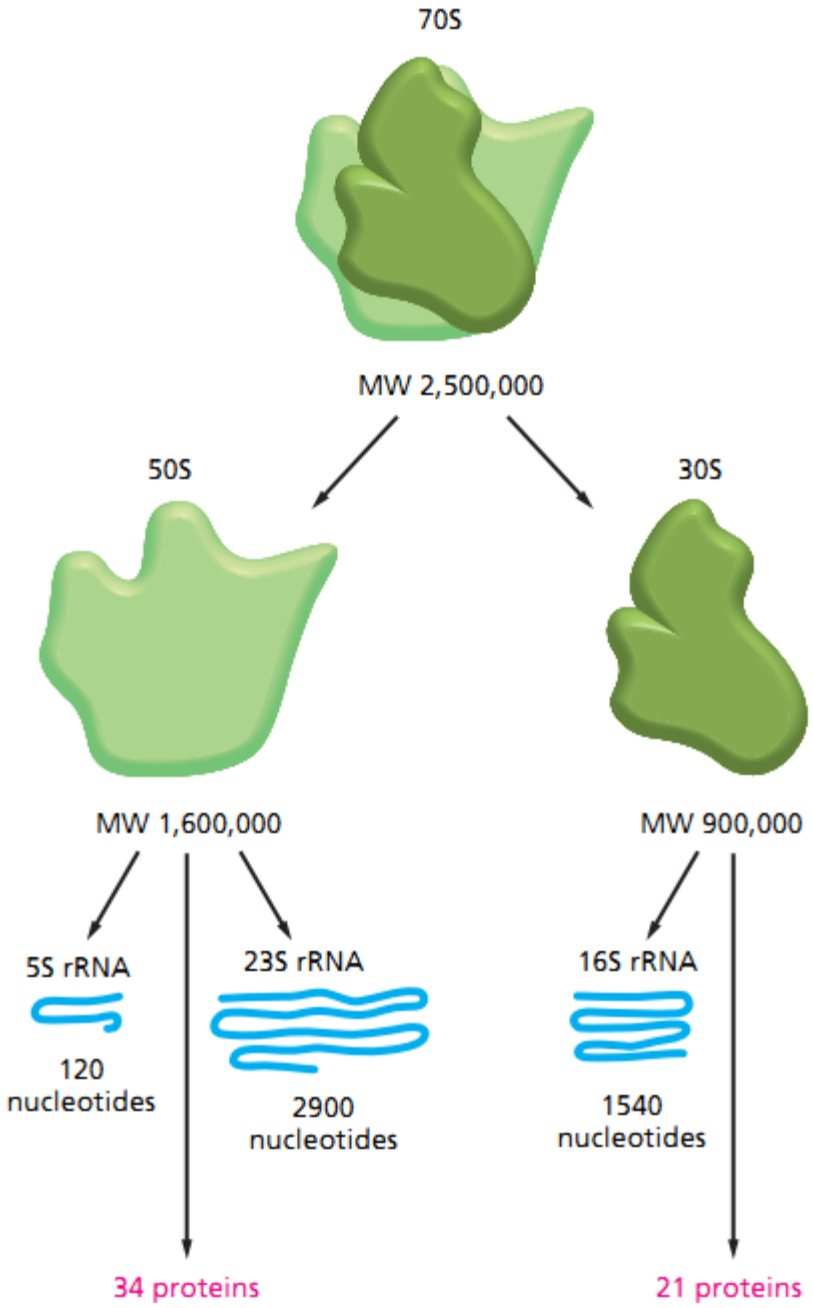
Metagenômica: tipos de Dados





16S / 18S / ITS

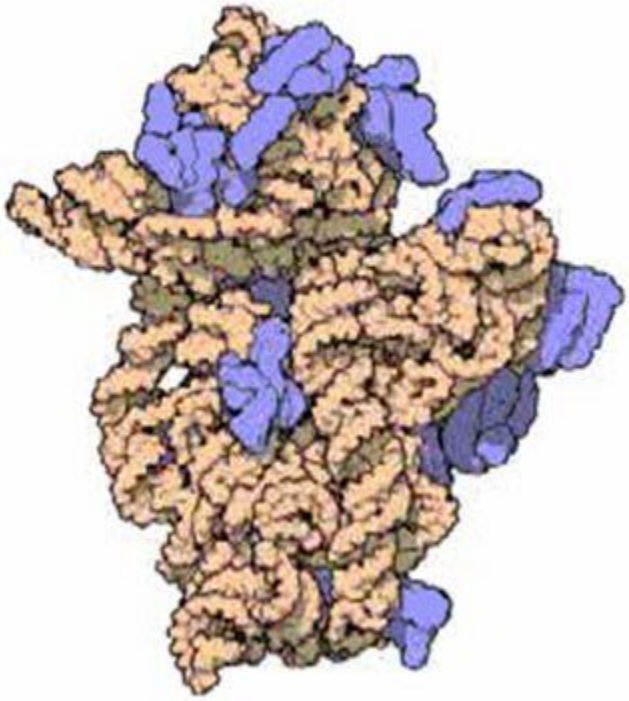
shotgun



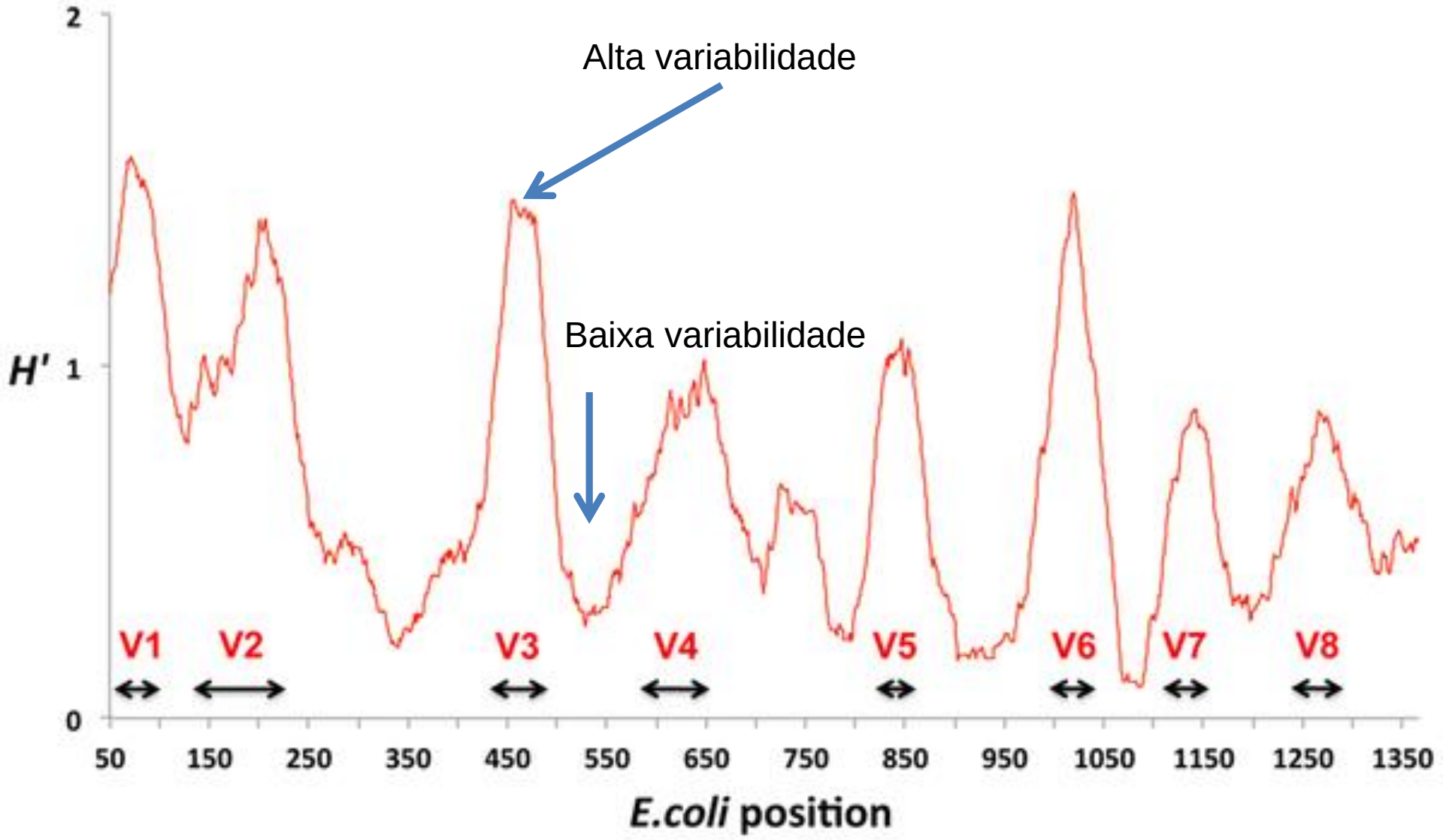
Alberts et al. 2008

16S rRNA

- 16S
  - Ribosomal RNA
  - Large RNA component of the small subunit of the ribosome
  - Phylogenetic Markers
    - Species Identification
  - 1542 bp



Crédito: Christel Chehoud, http://slideplayer.com/slide/4641762/

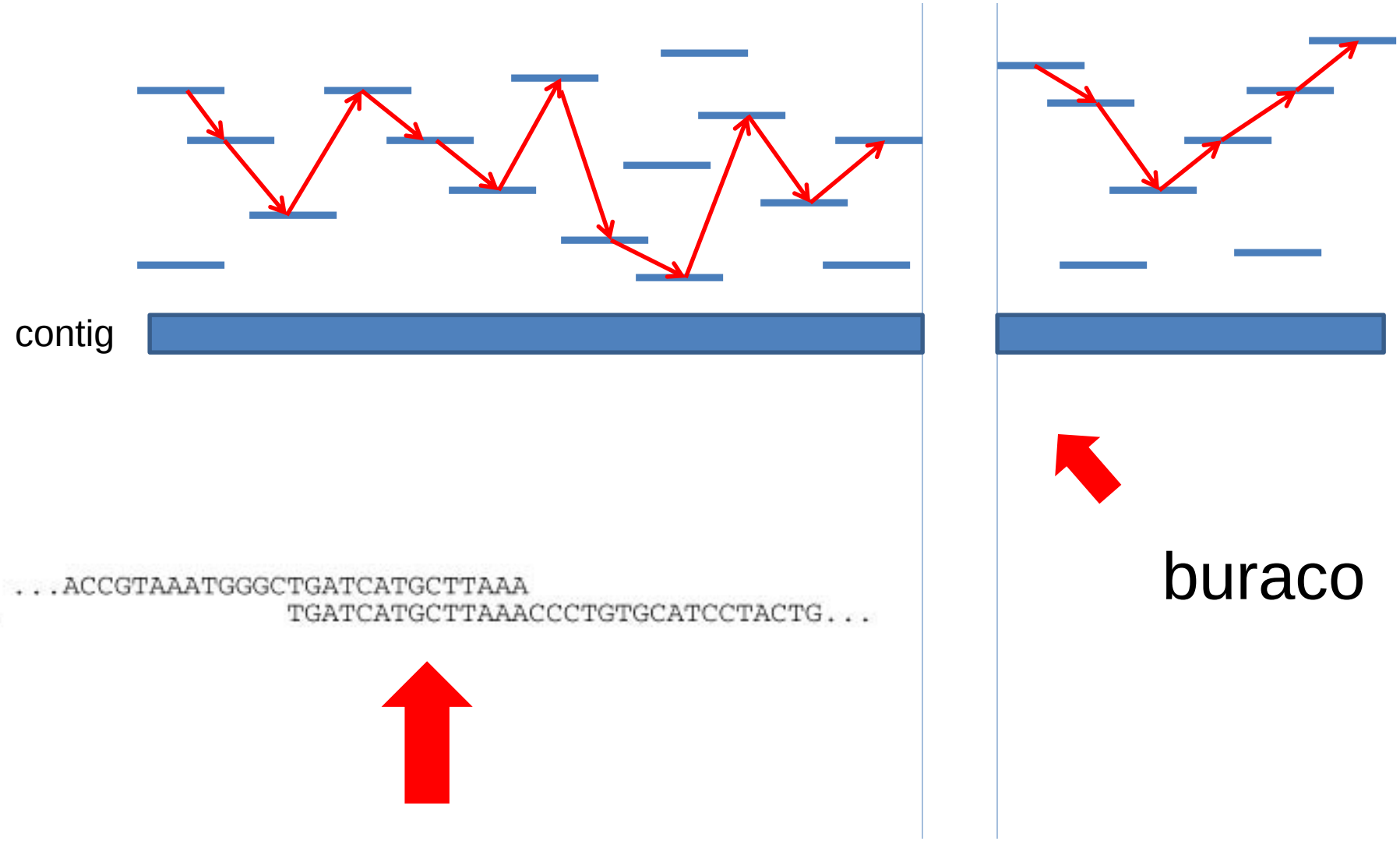


⇨ Primers “universais”

DNA shotgun

- Sequenciar o DNA total da amostra
- Resultado
  - Milhões de fragmento
  - Mistura dos DNAs dos diversos organismos presents
  - fragmentos devem ser montados

Montagem de genomas



Montagem

- Montagem é essencial para
  - Análise funcional
  - Recuperação de genomas
- Objeto principal resultante
  - contigs
  - genomas draft
- Em raros casos
  - genomas completos

16S vs. shotgun: objetivos

- 16S
  - Composição e estrutura da microbiota
    - “perfil taxonômico”
- Shotgun
  - Resultados mais detalhados
    - Perfil taxonômico
    - Funções gênicas
    - genomas

16S e shotgun: positivos e negativos

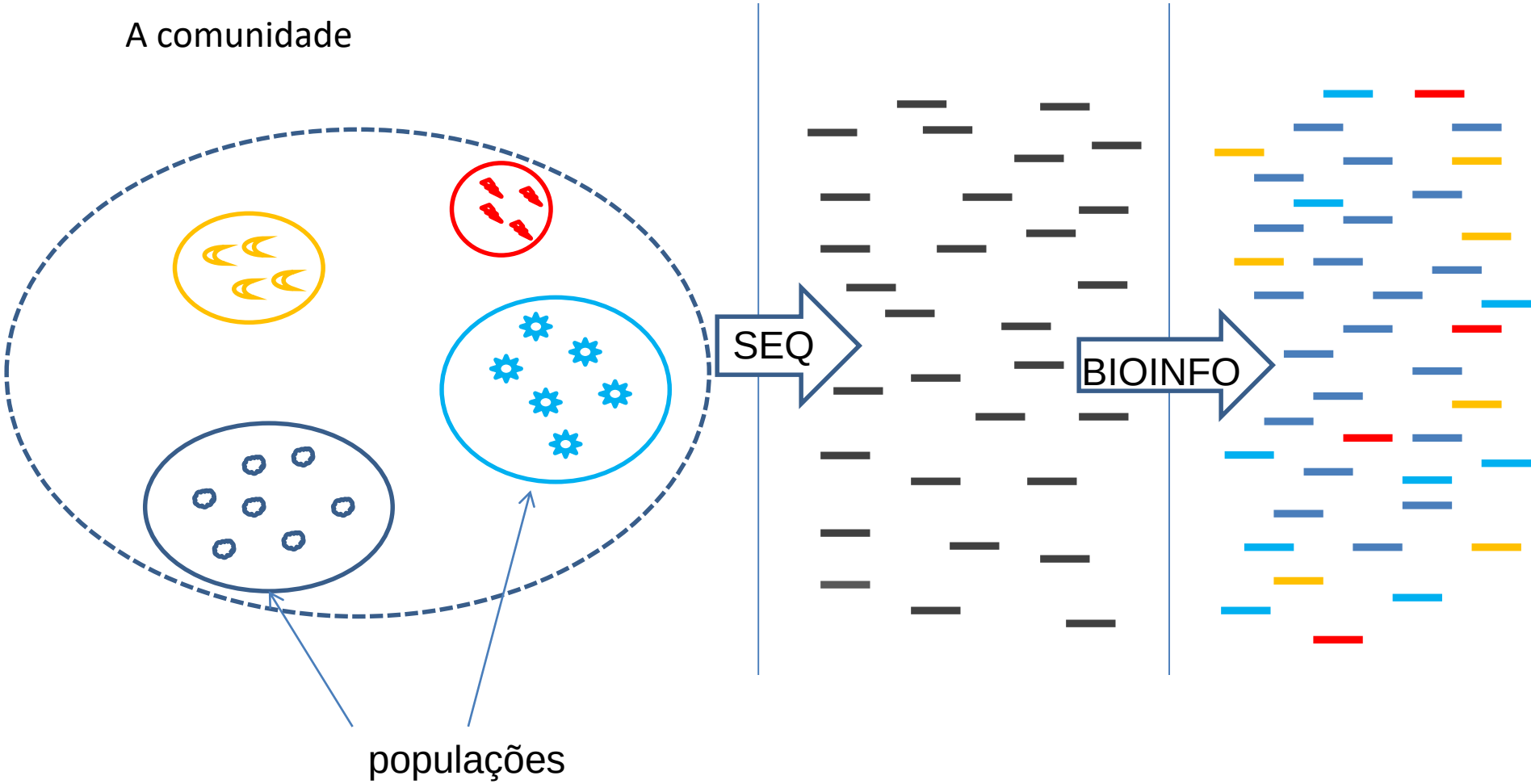
	16S	shotgun
custo	Mais baixo	Mais alto
Vieses (biases)	Menor chance de ser representativo	Maior chance de “pegar tudo”
Bancos de dados	Maior cobertura	Menor cobertura
Identificação taxonômica	Menos precisa (em geral, não mais do que gênero)	Mais precisa, podendo chegar a especie, e talvez cepas

Que perguntas queremos fazer?

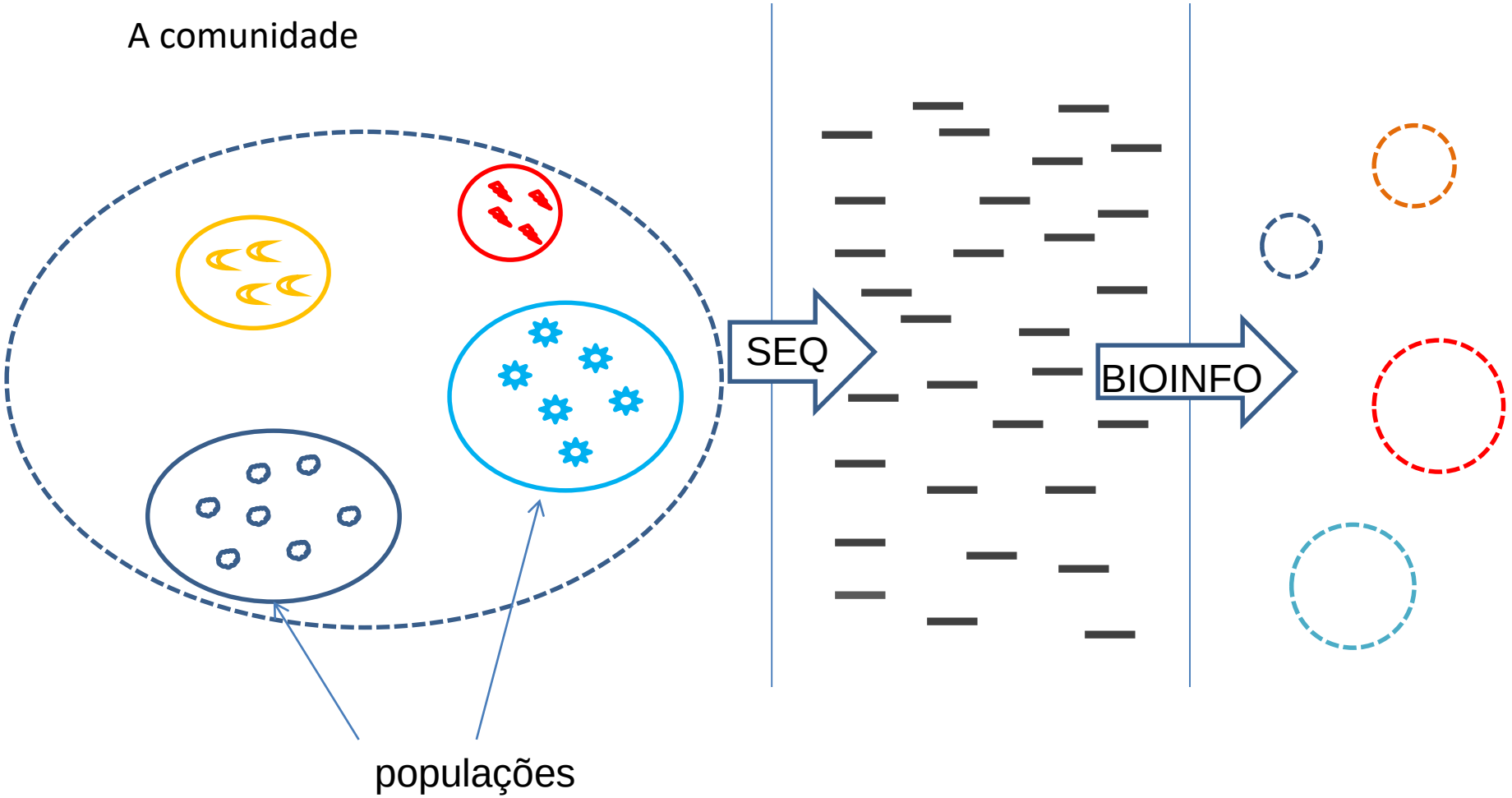
Quem está na amostra?

- Identificação taxonômica (16S, shotgun)
- Recuperação de genomas (shotgun)

16S / DNA shotgun

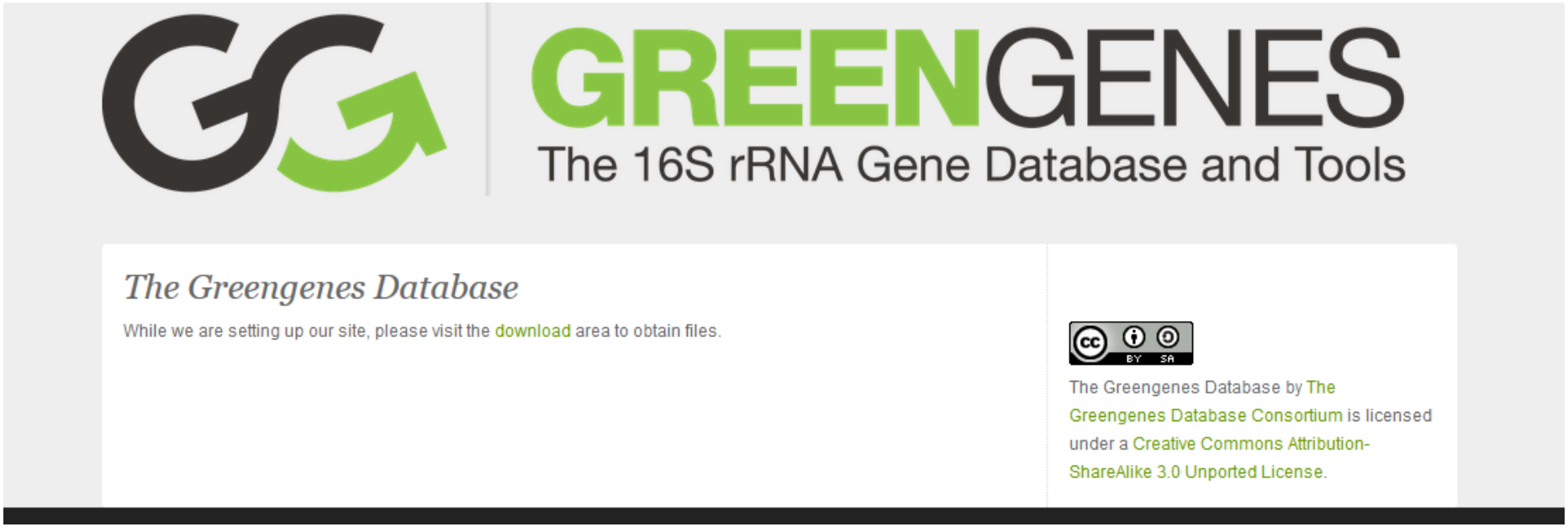


Recuperação de genomas



Identificação taxonômica depende de bancos de dados

Bancos de dados de 16S



silva
high quality ribosomal RNA databases

SILVA

Welcome to the SILVA rRNA database project

A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data.

SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (***Bacteria***, ***Archaea*** and ***Eukarya***).

SILVA are the official databases of the software package **ARB**.

For more background information → Click here

SILVAngs

silva ngs

Check out our new service for Next Generation Amplicon data

SILVA Tree Viewer

The **SILVA Tree Viewer** is a web application to browse and query the SILVA guide trees.

A technical preview is available at ↗www.arb-silva.de/treeviewer

ARB

The **software package ARB** represents a graphically-oriented, fully-integrated package of cooperating software tools for handling and analysis of sequence information.

The **ARB project** has been started more than 15 years ago by Wolfgang Ludwig at the Technical University in Munich, Germany, see ↗www.arb-home.de.

News

23.11.2017

The 10th de.NBI Quaterly Newsletter published

Good news for de.NBI, the German Network for Bioinformatics Infrastructure: In September, de.NBI has passed successfully the midterm evaluation in Berlin. The international evaluation panel stated that de.NBI is working successfully from the beginning and that it should be continued.

09.11.2017

Call for Action - We need your Help!

UniEuk

The UniEuk project needs your help to launch EukBank 1.0.

28.10.2017

de.NBI Handbook ready for Download

The Handbook is the first comprehensive document that lists the work and effort of all de.NBI partners. Content: How de.NBI is structured, Presentations of all Partners, Index of Persons/Contact Details.

05.10.2017

SILVA TreeViewer published

SILVA TreeViewer: interactive web browsing of the SILVA phylogenetic guide trees now published in BMC Bioinformatics.

go to Archive ->

User satisfaction survey

SILVA is now part of the German Network for Bioinformatics Infrastructure **de.NBI**.

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

To evaluate and improve our quality of service we need your feedback. **Please help us** by participating in this short ↗survey.

SILVA SSU / LSU 128 - full release

	SSU Parc	SSU Ref	SSU Ref NR 99	LSU Parc	LSU Ref

myrdp login

ANNOUNCEMENTS

RDP News

11/10/2017 myRDP login problem fixed!

11/09/2017 Apologize for the problem with myRDP login.
Our team is working to fix it as soon as possible.

05/16/2017 Apology for slow/NO connection to RDP tools today
Thanks to Alex/Brian, etc. for working things out in the server room

05/10/2017 RDP Director at GSC 19, May 14-17
Genomic Standards Consortium Meeting, Brisbane, Queensland, Australia

05/10/2017 Possible Friday, May 12, morning interruptions
Emergency Generator Testing from 9-10 A.M.

12/13/2016 Most Highly Cited Researchers
Congratulations to RDP Director James Cole

09/30/2016 RDP Release 11.5 available
Updated 16S rRNA training set to training set No. 16.

08/16/2016 Possible Friday morning interruptions
Building electrical testing/maintenance

06/30/2016 RDP Classifier Updates
The Classifier 16S training set and Fungal ITS Warcup set have been updated

06/03/2016 RDP staff on the road!
Teaching in China, Genomic Standards Consortium meeting in Crete, special ASM Microbe events in Boston

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.

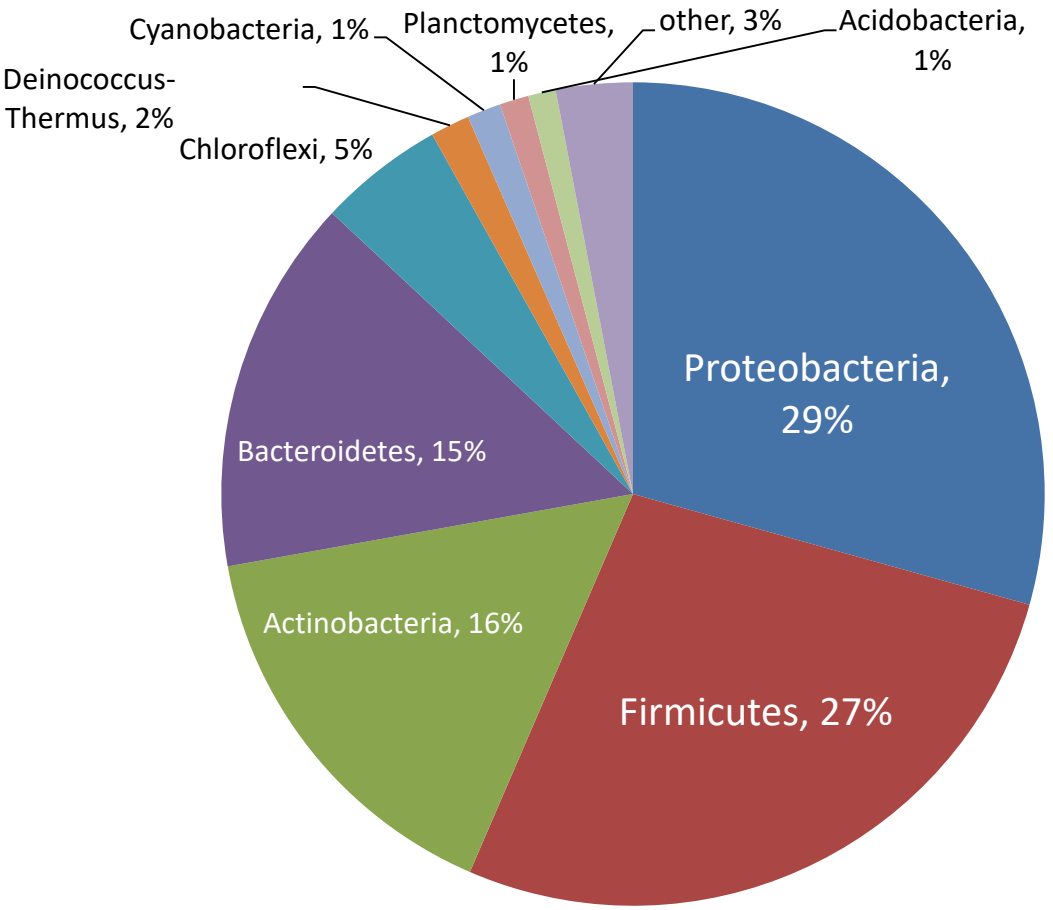
RDP's mission and funding:

Part of RDP's mission is to provide support to our users. Email and phone contacts are available on the *contacts page*.

Bancos de dados para DNA total

- GenBank
  - nt
  - nr
  - env_nr
  - refSeq
  - WGS

Classificação taxonômica e abundância relativa



Genomas de procariotos no GenBank

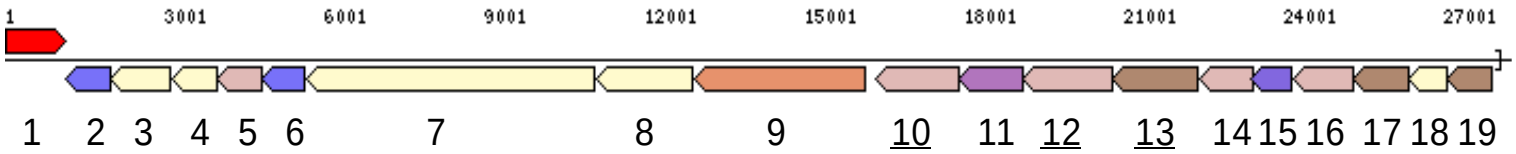
filo	# genomas	%
Actinobacteria	4059	13
Bacteroidetes/chlorobi	932	3
Cyanobacteria	340	1
Firmicutes	9628	31
Proteobacteria	14268	46
Spirochaetes	525	2
Others	1500	5

Source: Land et al. 2015

Quais funções estão presentes?

- Em genes (shotgun)
- Em genes expressos (metaTranscritômica)

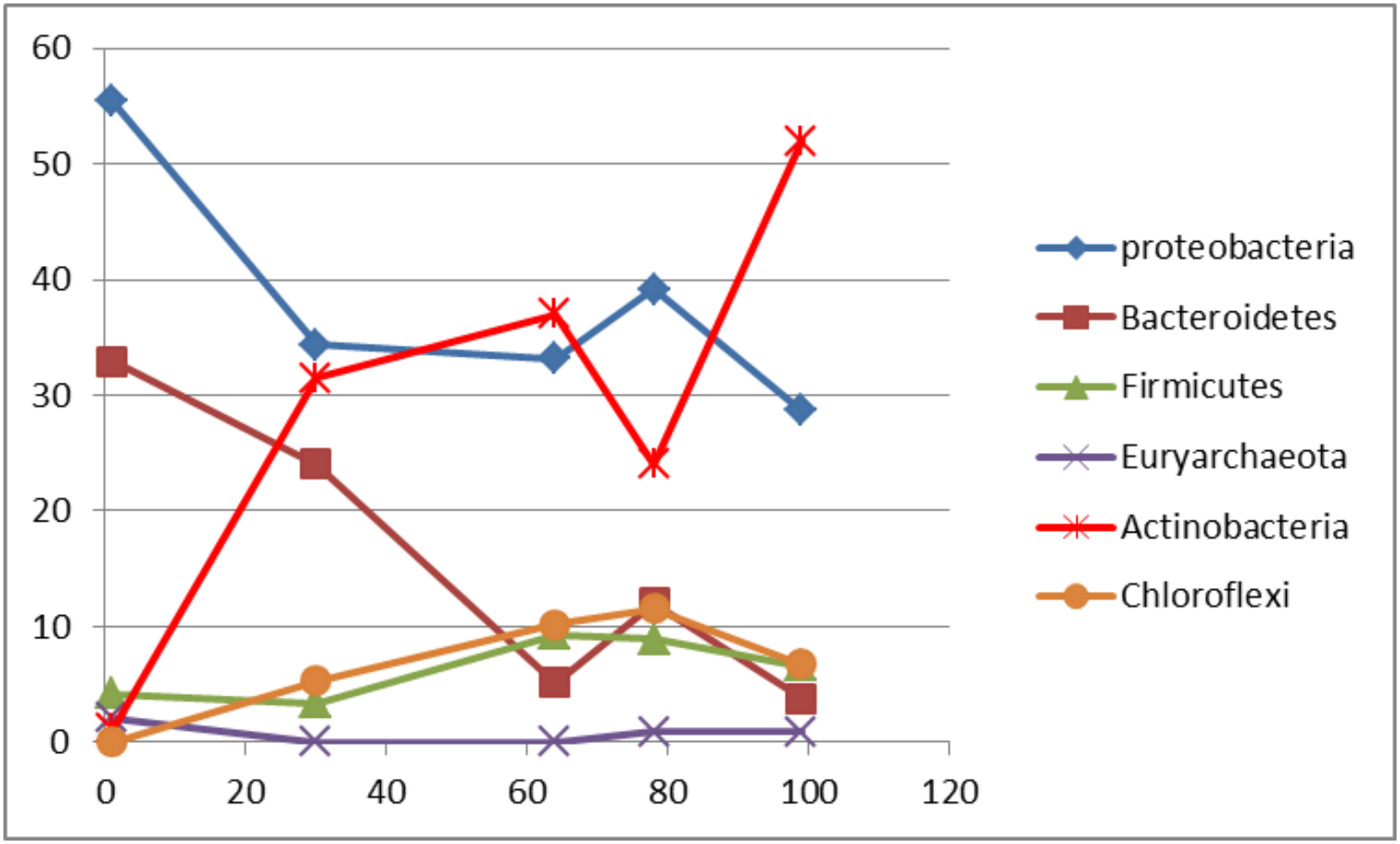
ZC1 contig00009.9 (27,919 bp)



1. Beta-xylosidase (376aa, COG3507)
2. Dehydrogenases (280aa, COG1028)
3. hypothetical protein (379aa);
4. hypothetical protein (283aa)
5. 5-keto 4-deoxyuronate isomerase (280aa, COG3717)
6. Dehydrogenases (267aa, COG1028)
7. hypothetical protein (1799aa)
8. SusD family protein (606aa, pfam07980)
9. TonB-linked outer membrane protein (1068aa, COG4771);
10. Pectate lyase (518aa, COG3866)
11. Predicted unsaturated glucuronyl hydrolase
12. Pectin methylesterase (568aa, COG4677)
13. Endopolygalacturonase (523aa, COG5434)
14. Nucleoside-diphosphate-sugar epimerase (326aa, COG0451)
15. Nucleoside-diphosphate-sugar pyrophosphorylase (249aa, pfam00483)
16. Galactokinase (377aa, COG0153)
17.Soluble lytic murein transglycosylase (347aa, COG0741)
18. hypothetical protein (235aa)
19. Predicted UDP-glucose 6-dehydrogenase (283aa, COG1004).

Metagenômica comparativa

Mesmo local, variação no tempo



Mesmo local, variação de indivíduos

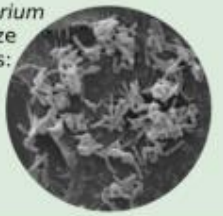
- Amostras da boca
  - Indivíduos que fumam
  - Indivíduos que não fumam

A map of diversity in the human microbiome

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

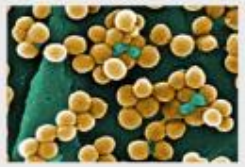
Propionibacterium acnes lives on the skin and nose of most people

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

Staphylococcus epidermidis colonizes external body sites



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

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

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

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

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

Taxonomia

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

GTDB

2018

RESOURCE

nature biotechnology

A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life

Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil & Philip Hugenholtz

Taxonomy is an organizing principle of biology and is ideally based on evolutionary relationships among organisms. Development of a robust bacterial taxonomy has been hindered by an inability to obtain most bacteria in pure culture and, to a lesser extent, by the historical use of phenotypes to guide classification. Culture-independent sequencing technologies have matured sufficiently that a comprehensive genome-based taxonomy is now possible. We used a concatenated protein phylogeny as the basis for a bacterial taxonomy that conservatively removes polyphyletic groups and normalizes taxonomic ranks on the basis of relative evolutionary divergence. Under this approach, 58% of the 94,759 genomes comprising the Genome Taxonomy Database had changes to their existing taxonomy. This result includes the description of 99 phyla, including six major monophyletic units from the subdivision of the Proteobacteria, and amalgamation of the Candidate Phyla Radiation into a single phylum. Our taxonomy should enable improved classification of uncultured bacteria and provide a sound basis for ecological and evolutionary studies.

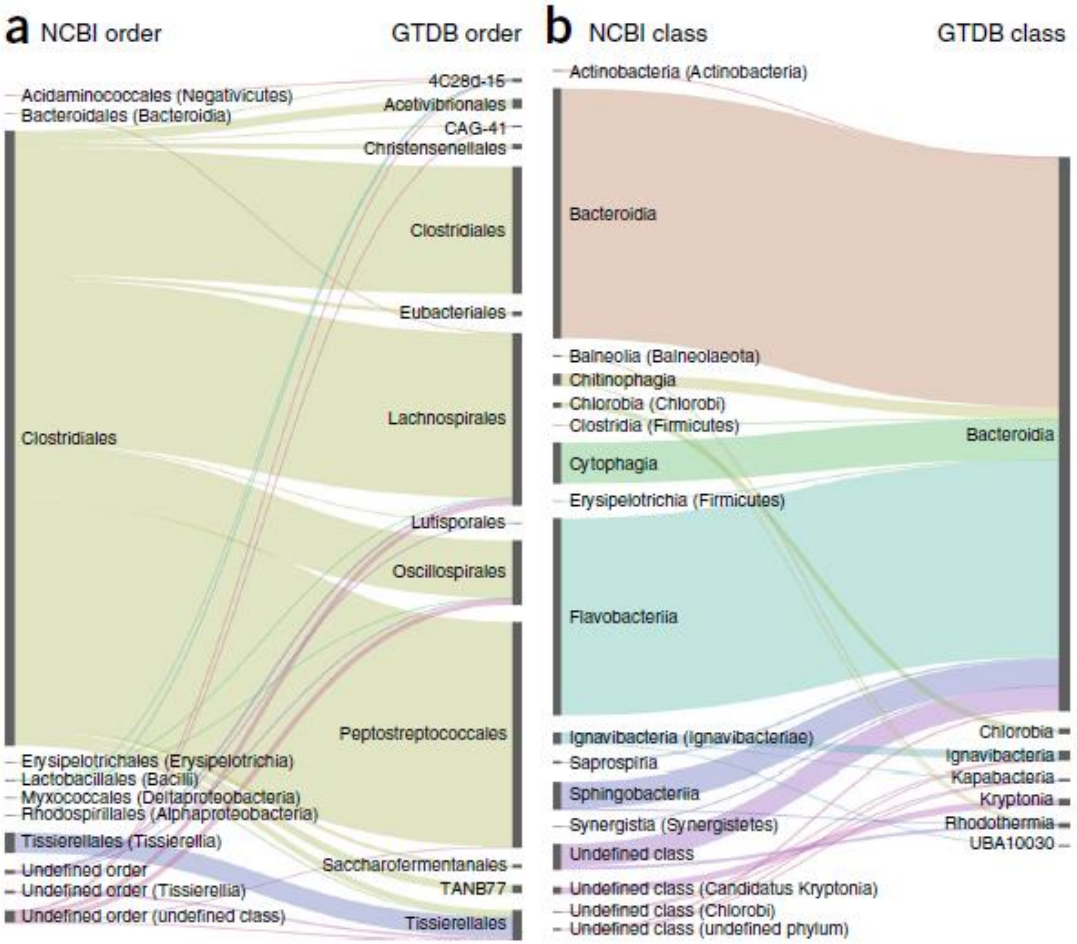


Figure 5 Comparisons of NCBI and GTDB classifications of genomes designated as Clostridia or Bacteroidetes in the GTDB taxonomy. (**a**) Comparison of NCBI (left) and GTDB (right) order-level classifications of the 2,368 bacterial genomes assigned to the class Clostridia in the GTDB taxonomy. Genomes classified in a class other than Clostridia by NCBI are indicated in parentheses. (**b**) Comparison of NCBI and GTDB class-level classifications of the 2,058 bacterial genomes assigned to the phylum Bacteroidetes in the GTDB taxonomy. Genomes classified in a phylum other than the Bacteroidetes by NCBI are indicated in parentheses.

OTU

- Unidade taxonômica operacional
- Se for conhecida, leva um rótulo padronizado
  - *Xanthomonas citri*
- Mas pode ser desconhecida
  - Nesse caso, recebe um número, que varia de análise para análise
- Conceito comum em análise de dados de 16S

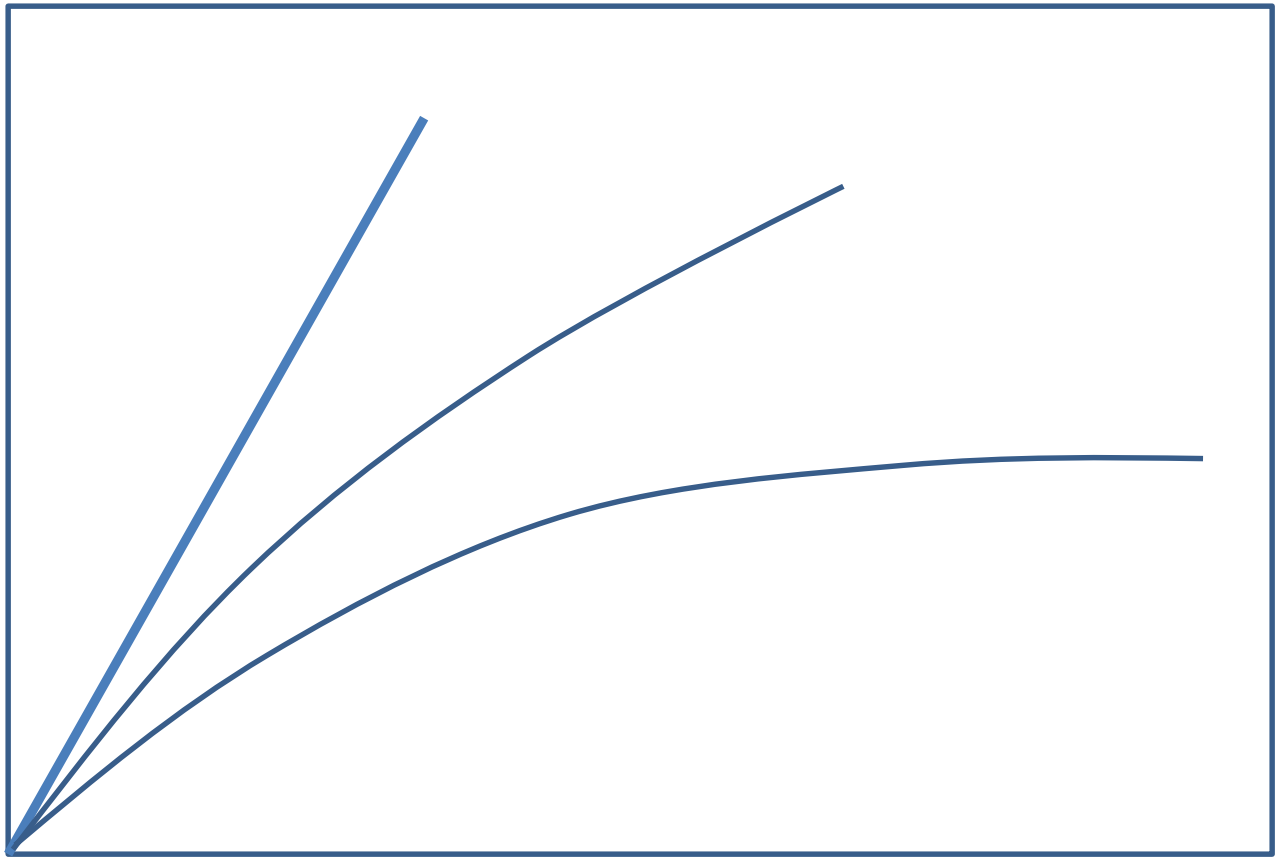
ASV

- Amplicon Sequence Variant
- A diferença entre duas ASVs pode ser apenas uma base
- é um conceito mais preciso do que OTU
- Deve ser usado em lugar de OTU
  - Mas podem ser usados em conjunto
    - diferentes ASVs podem corresponder a uma OTU
      - por exemplo: um gênero

A amostra é representativa?

- Curvas de rarefação

Curvas de rarefação (ou saturamento)



Muitas fontes de erro

- Amostragem
- Preparação da biblioteca
- Sequenciamento
- Tamanho da sequência (pode ser curta demais)
- Programas (montadores, classificadores)
- Viéses dos bancos de dados

Classificação de reads de DNA total

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

Por analogia com classificação de reads em dados de 16S (OTUs)

- Separar reads em “caixinhas”
- cada caixinha tem os reads que mutuamente se parecem num nível de 97 ou 98% de identidade
- qual seria o análogo para DNA total?

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

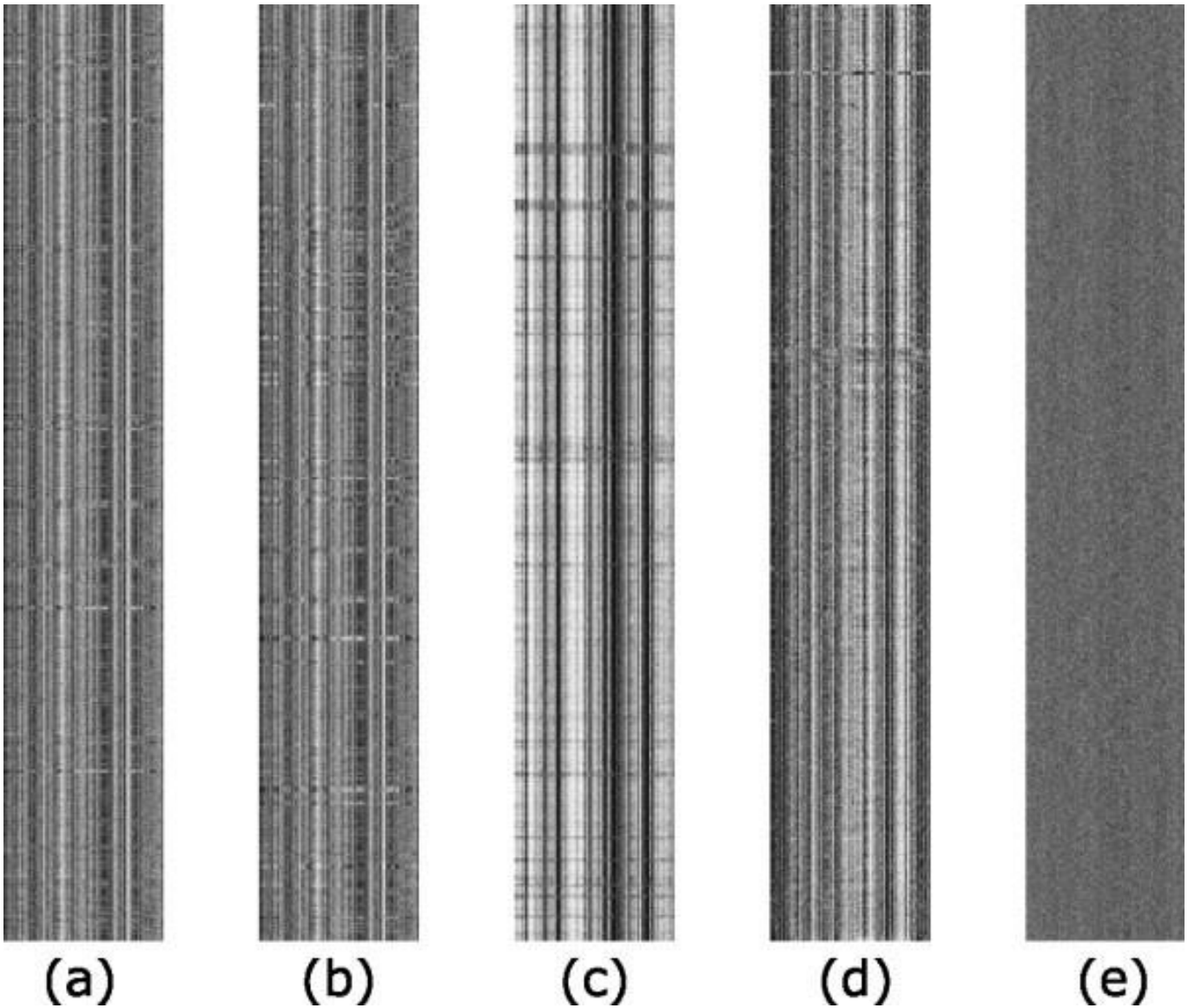
Palavras de k bases: k-mers (kâmeros)

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

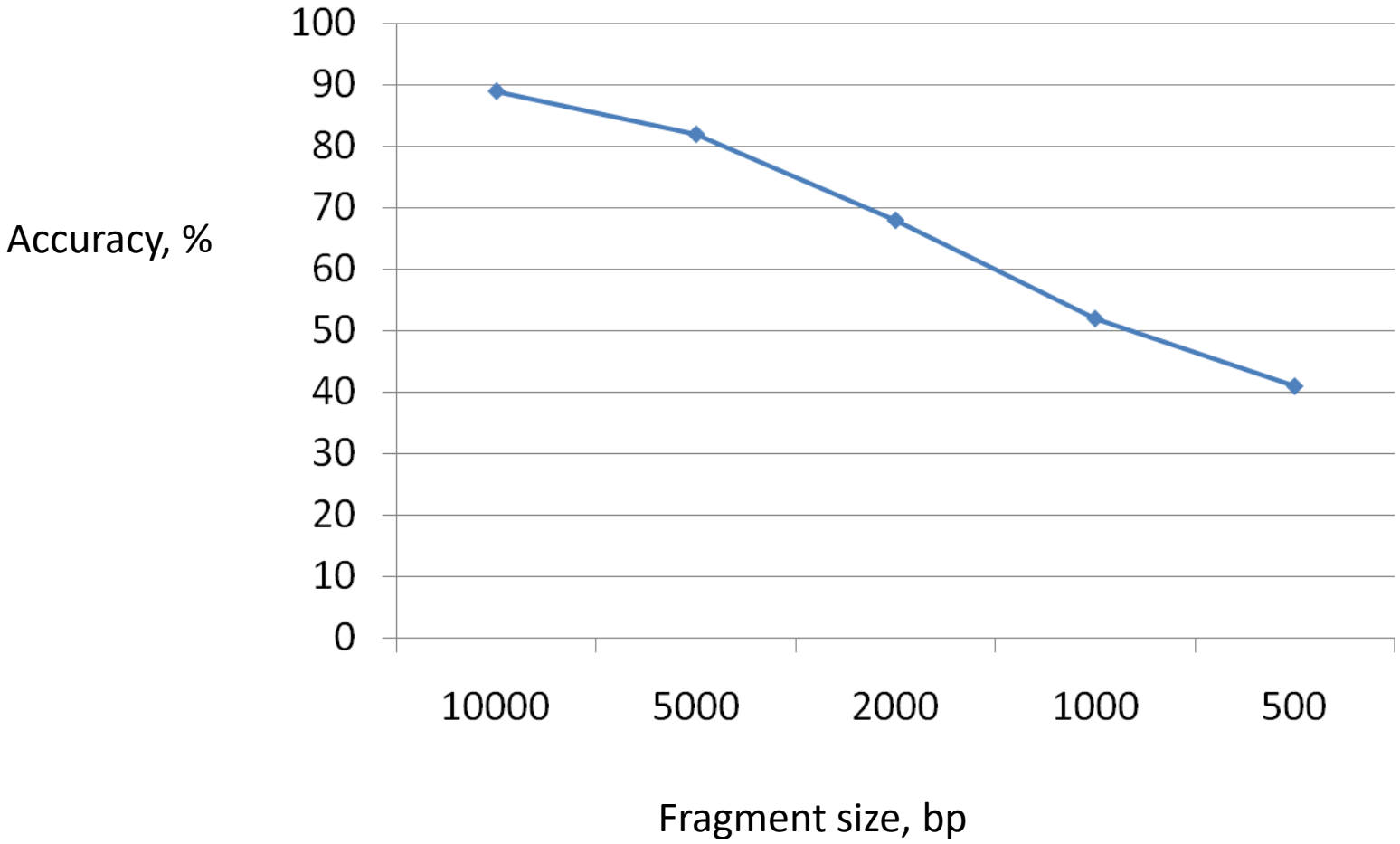
E. coli O157

Pyrococcus furiosus

random

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

Não funciona bem com fragmentos curtos

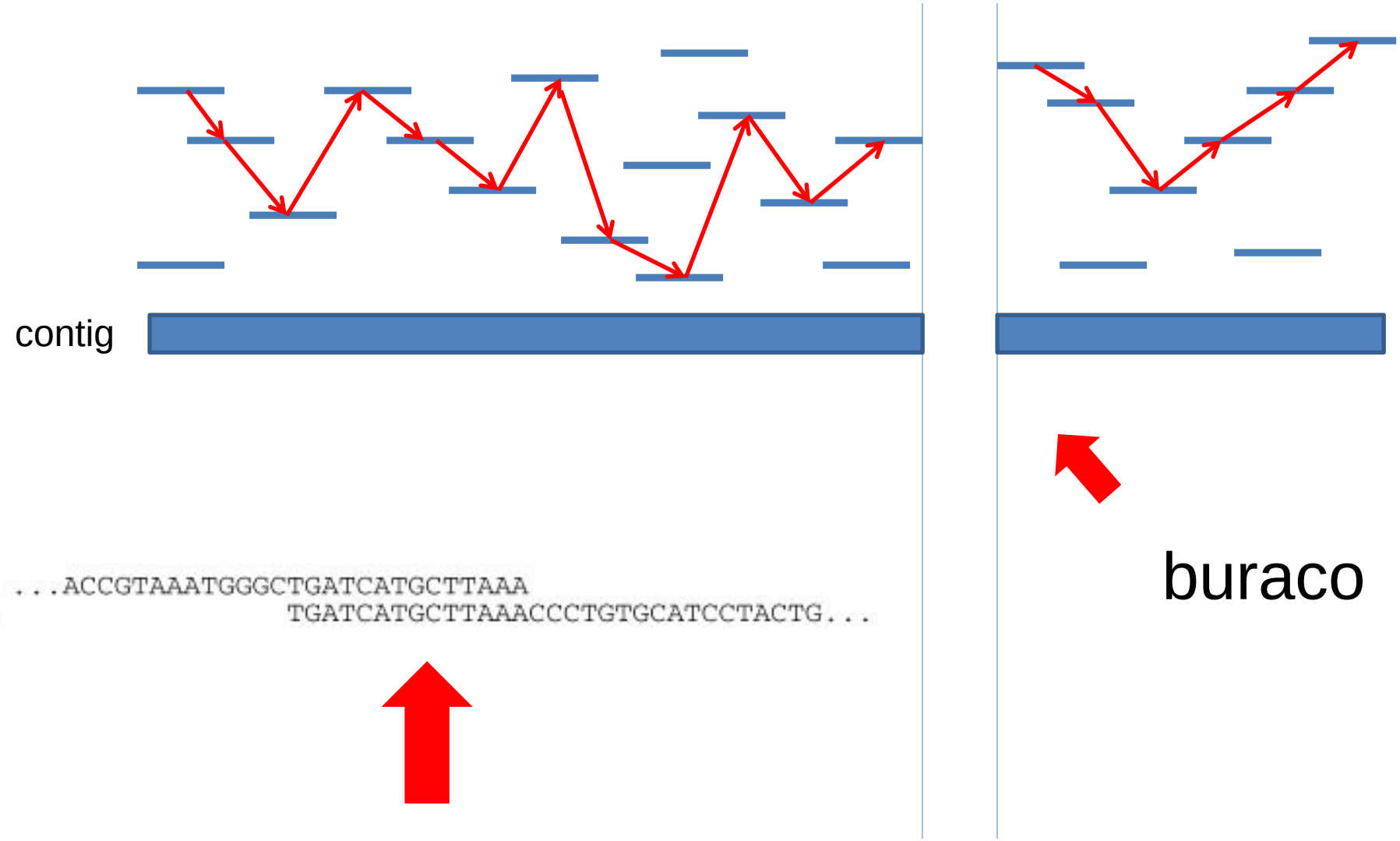


Zhou et al, 2009 simulated data

Exercício

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

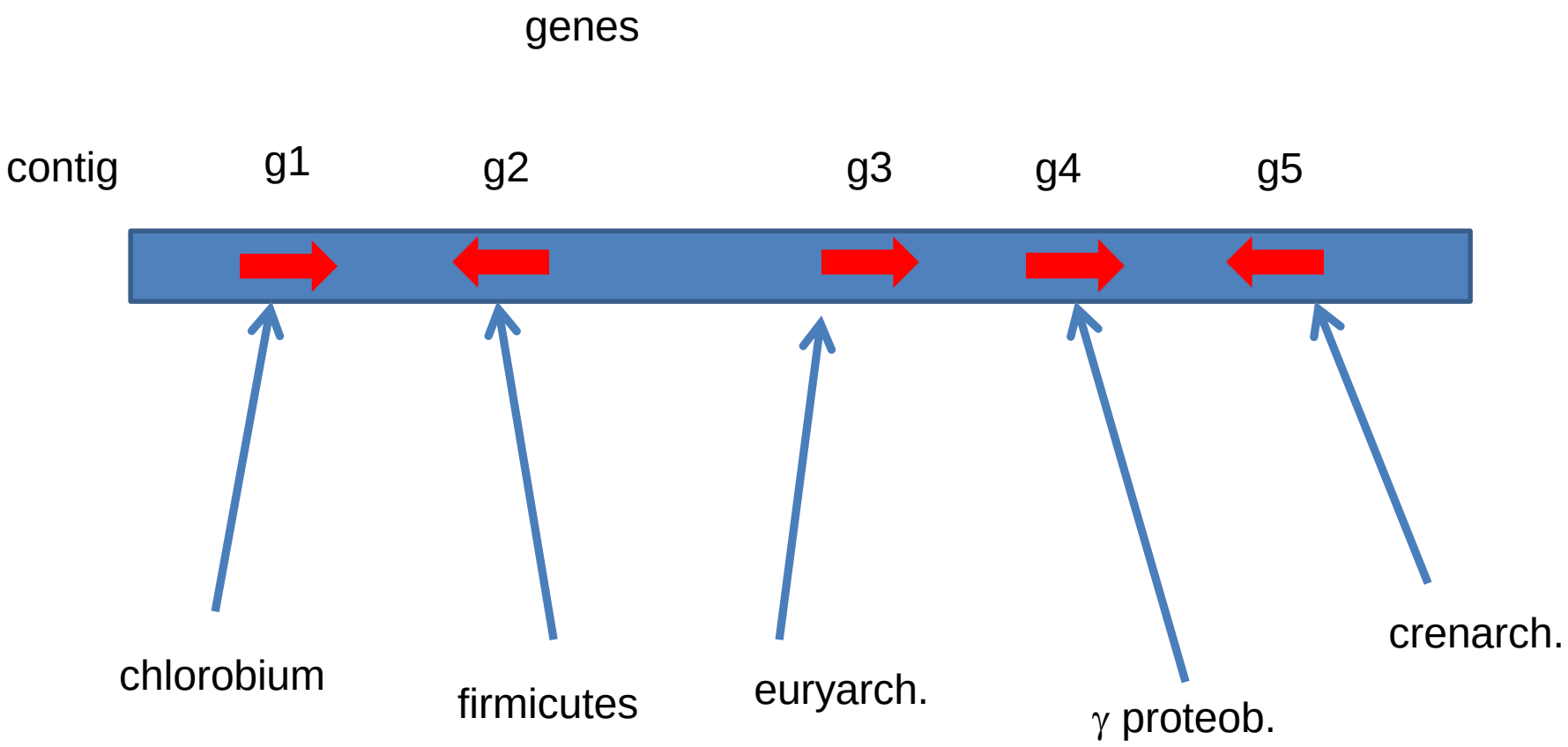
Montagem de genomas



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

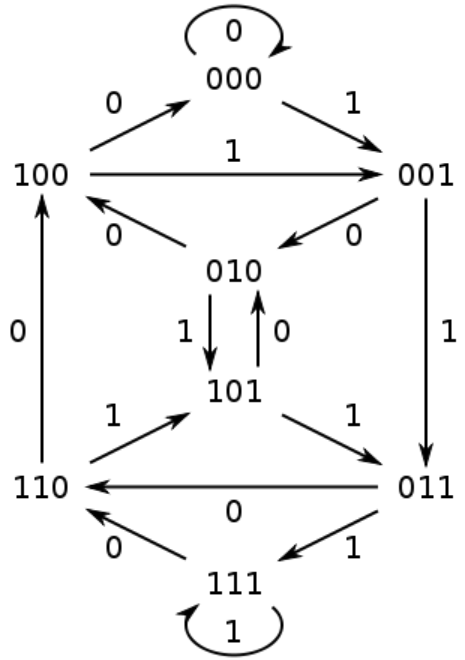
Exemplo de quimerismo



Paradigmas de montagem

- OLC
  - overlap, layout, consensus
  - mais rigoroso, mas mais lento
- k-meros + grafos de de Bruijn
  - menos rigoroso, mas muito mais rápido
  - mais apropriado para metagenômica

grafos de de Bruijn



Sobreposição de *k*-mers

alfabeto binário

$k = 1$

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

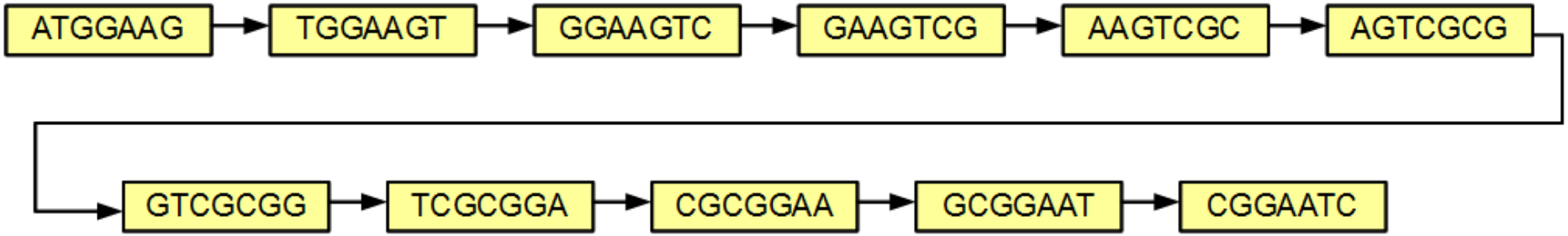
Grafo de de Bruijn em montagem

sequence **ATGGAAGTCGCGGAATC**

7mers

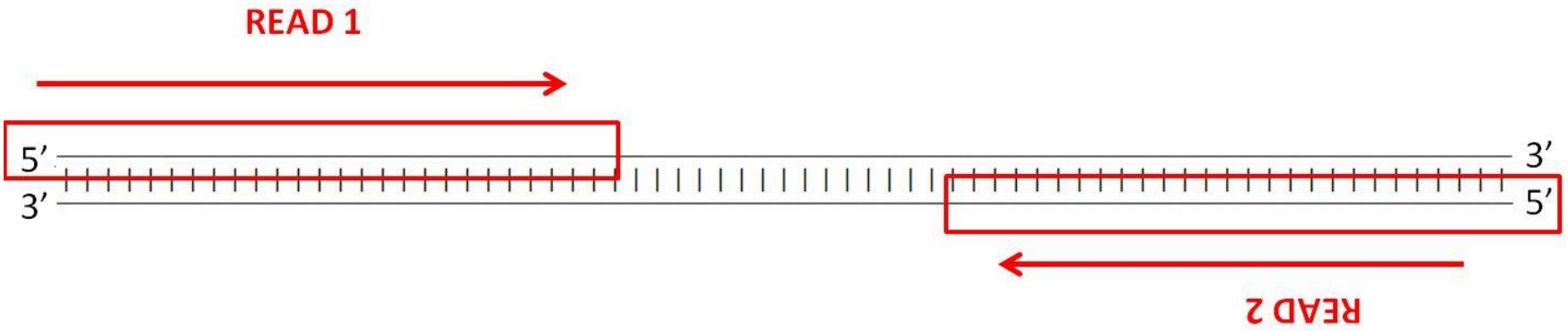
ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

de Bruijn graph



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

Single-end and Paired-end reads



Crédito: http://www.cureffi.org/2012/12/19/forward-and-reverse-reads-in-paired-end-sequencing/

Anotação funcional

- Pipeline para genomas completos pode ser usado
  - Exemplo: IMG/M
- Revejam aula sobre anotação de genomas

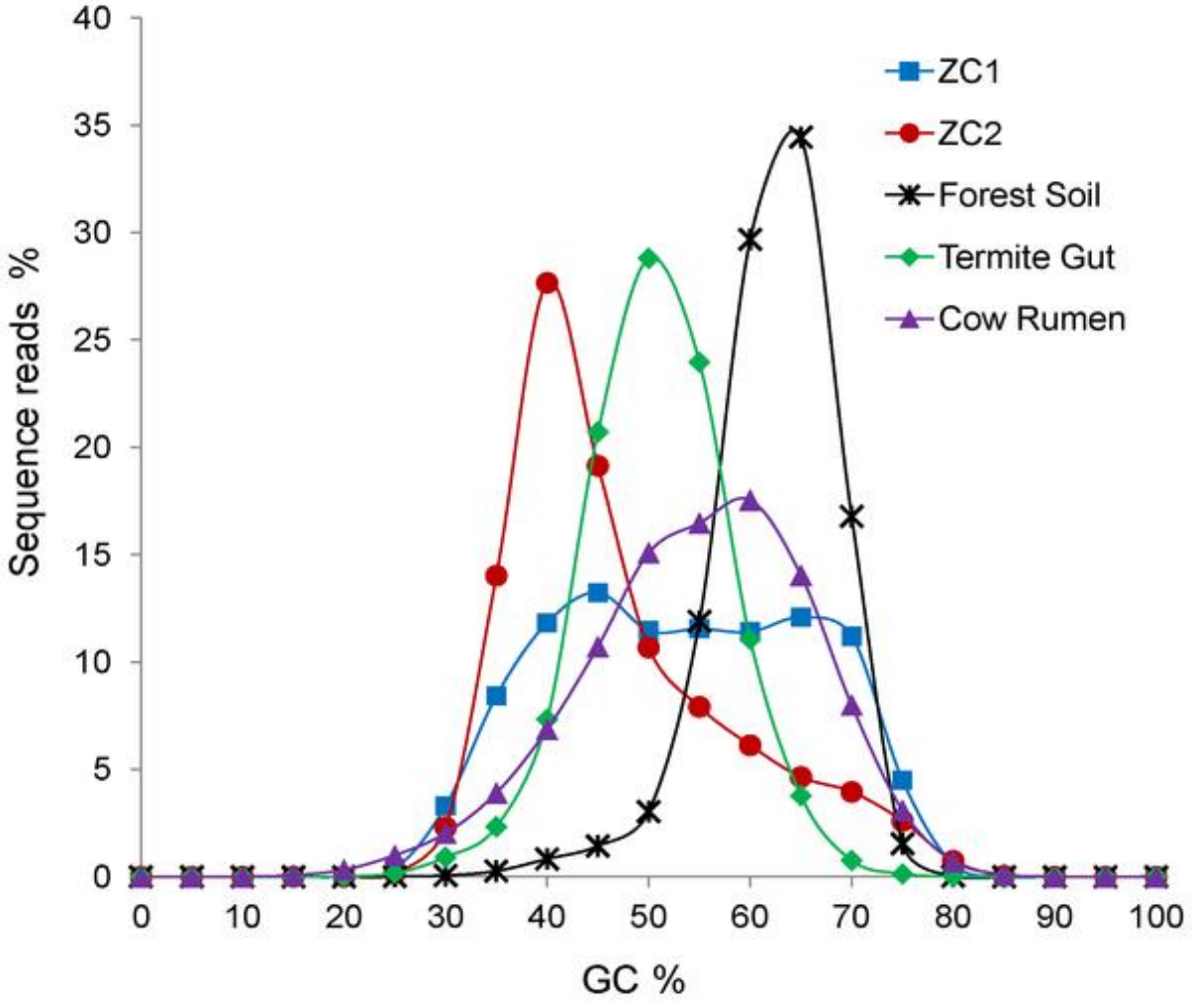
Cobertura

- Quanto cada genoma é coberto pelos reads obtidos
- Ambientes de grande riqueza: cobertura baixa
- Cobertura baixa cria contigs pequenos
  - 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:

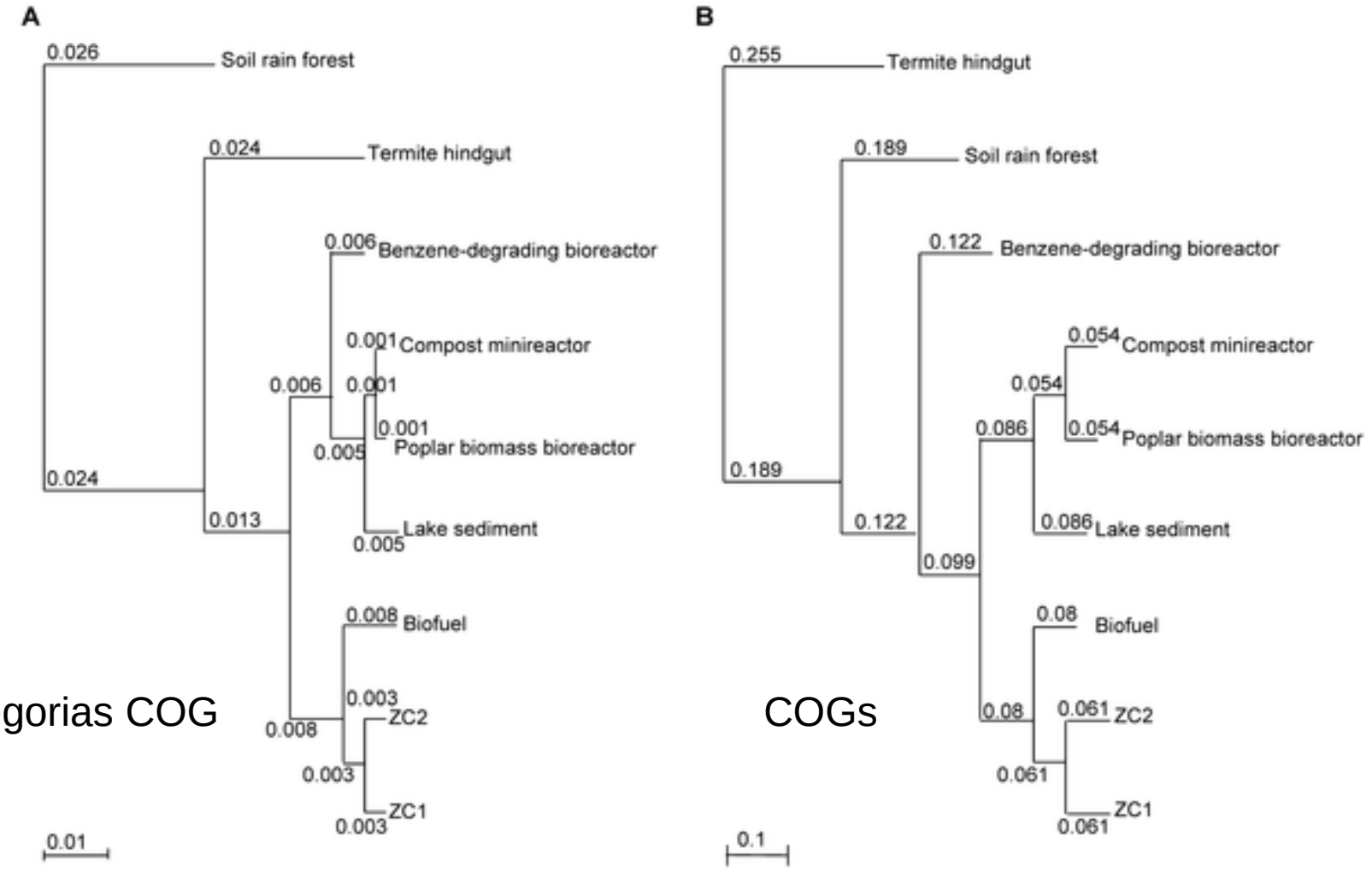
- ( ) COG Categories
- ( ) COG Pathways
- ( ) KEGG Pathway Categories (KO)
- ( ) KEGG Pathway Categories (EC)
- ( ) KEGG Pathways (KO)
- ( ) KEGG Pathways (EC)
- ( ) Pfam Categories

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

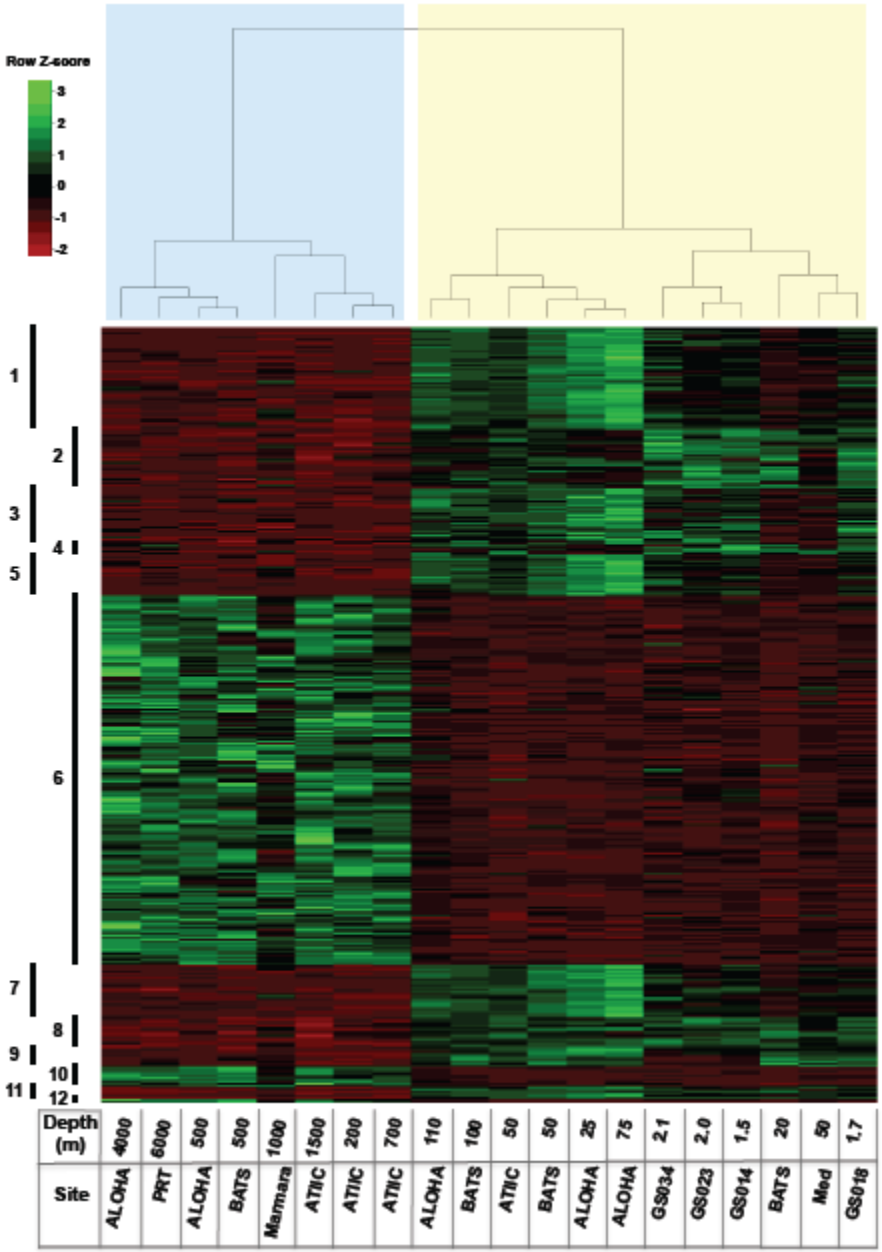
Martins LF, Antunes LP, Pascon RC, de Oliveira JCF, Digiampietri LA, et al. (2013) Metagenomic Analysis of a Tropical Composting Operation at the São Paulo Zoo Park Reveals Diversity of Biomass Degradation Functions and Organisms. PLoS ONE 8(4): e61928. doi:10.1371/journal.pone.0061928
http://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.006192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

Abundância de funções

- mapeamento de reads em ORFs anotadas

Abundância relativa espacial

- é necessário o conceito de família gênica
- COG
  - Clusters of Orthologous Groups
- **COGs diferencialmente representados**
- Semelhante a genes diferencialmente expressos
- Heat maps, clusterização hierárquica



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

COGs

Iquique not included

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

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:

Sugestões 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

International Journal of
Molecular Sciences

MDPI

Review

Human Microbiome Acquisition and Bioinformatic Challenges in Metagenomic Studies

Valeria D'Argenio [1,2,3]

[1] CEINGE-Biotecnologie Avanzate, via G. Salvatore 486, 80145 Naples, Italy; dargenio@ceinge.unina.it; Tel.: +39-081-373-7909
[2] Department of Molecular Medicine and Medical Biotechnologies, University of Naples Federico II, via Pansini 5, 80131 Naples, Italy
[3] Task Force on Microbiome Studies, University of Naples Federico II, 80131 Naples, Italy

Received: 14 December 2017; Accepted: 24 January 2018; Published: 27 January 2018

Abstract: The study of the human microbiome has become a very popular topic. Our microbial counterpart, in fact, appears to play an important role in human physiology and health maintenance. Accordingly, microbiome alterations have been reported in an increasing number of human diseases. Despite the huge amount of data produced to date, less is known on how a microbial dysbiosis effectively contributes to a specific pathology. To fill in this gap, other approaches for microbiome study, more comprehensive than 16S rRNA gene sequencing, i.e., shotgun metagenomics and metatranscriptomics, are becoming more widely used. Methods standardization and the development of specific pipelines for data analysis are required to contribute to and increase our understanding of the human microbiome relationship with health and disease status.

Keywords: human microbiome; 16S rRNA analysis; metagenomics; metrascriptomics; data analysis; bioinformatics

Nature Reviews Microbiology 2018

MICROBIOME

Best practices for analysing microbiomes

Rob Knight[1,4,6,12], Alison Vrbanac[2,12], Bryn C. Taylor[2,12], Alexander Aksenov[3], Chris Callewaert[4,5], Justine Debelius[4], Antonio Gonzalez[4], Tomasz Kosciolek[4], Laura-Isobel McCall[3], Daniel McDonald[4], Alexey V. Melnik[3], James T. Morton[4,6], Jose Navas[6], Robert A. Quinn[3], Jon G. Sanders[4], Austin D. Swafford[1], Luke R. Thompson[7,8], Anupriya Tripathi[9], Zhenjiang Z. Xu[4], Jesse R. Zaneveld[10], Qiyun Zhu[4], J. Gregory Caporaso[11] and Pieter C. Dorrestein[1,3,4]*

Abstract | Complex microbial communities shape the dynamics of various environments, ranging from the mammalian gastrointestinal tract to the soil. Advances in DNA sequencing technologies and data analysis have provided drastic improvements in microbiome analyses, for example, in taxonomic resolution, false discovery rate control and other properties, over earlier methods. In this Review, we discuss the best practices for performing a microbiome study, including experimental design, choice of molecular analysis technology, methods for data analysis and the integration of multiple omics data sets. We focus on recent findings that suggest that operational taxonomic unit-based analyses should be replaced with new methods that are based on exact sequence variants, methods for integrating metagenomic and metabolomic data, and issues surrounding compositional data analysis, where advances have been particularly rapid. We note that although some of these approaches are new, it is important to keep sight of the classic issues that arise during experimental design and relate to research reproducibility. We describe how keeping these issues in mind allows researchers to obtain more insight from their microbiome data sets.

Exact sequence variants
For marker gene sequencing, the exact DNA sequence for each read is used instead of operational taxonomic unit clustering.

Operational taxonomic units
(OTUs). A group of closely related individuals or sequences (often 97% sequence similarity threshold).

Machine learning
The use of algorithms to learn from and make predictions about data.

Advances in DNA sequencing technologies have transformed our capacity to investigate the composition and dynamics of complex microbial communities that inhabit diverse environments, from mammalian gastrointestinal tracts to deep ocean sediments. These developments have led to vast increases in the number of microbiome studies being performed in many fields of science, from clinical research to biotechnology. With this transformation, researchers are often left holding massive amounts of data and are confronted with a bewildering array of computational tools and methods for analysing their data. Conducting a robust experiment is not trivial in microbiome research, and as with any study, experimental methods, environmental factors and analysis methods can affect results. Standards for data collection and analysis are still emerging in the field, yet many compelling results can be achieved with current practices.

and functional assignment; integration of data sets from multiple sequencing runs; and further improvement in machine learning, compositional data analysis and multi-omics analyses. However, many of the most fundamental issues that concern microbiome studies arise from statistical and experimental design issues. The most important challenge for the field is to integrate new approaches that are unique to microbiome studies, while remembering standard practices that are broadly applicable to all scientific studies.

Although it is impossible to be fully comprehensive in one article, this Review aims to provide straightforward guidelines for designing and executing a microbiome experiment and analysing the resulting data, with a particular focus on human, model organism and environmental microbiomes. We direct the reader to more specialized reviews on specific topics where these exist.