Universidade de São Paulo
Instituto de Química

USP

Análise de Microbiomas
aula 1

João Carlos Setubal

2024

Os micro-organismos 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

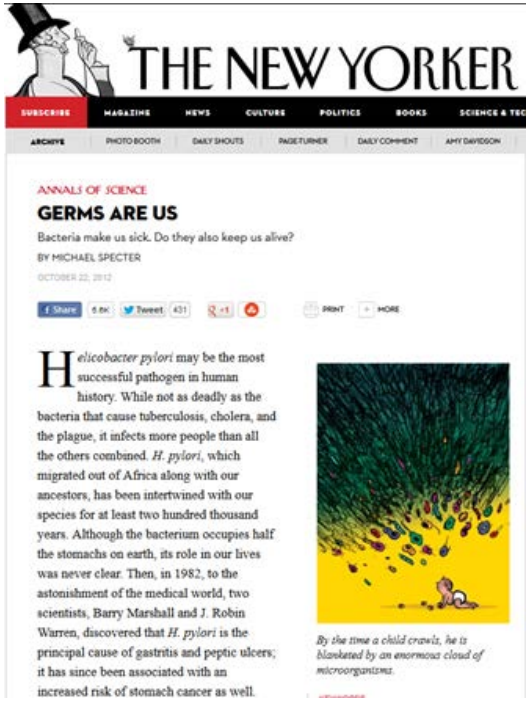
Projeto Microbioma Humano

junho 2012

Science
AAAS

My Microbiome and Me

Science 8 June 2012:


THE NEW YORKER

SUBSCRIBE | MAGAZINE | NEWS | CULTURE | POLITICS | BOOKS | SCIENCE & TECH

ARCHIVE | PHOTO BOOTH | DAILY SHOUTS | PAGE TURNER | DAILY COMMENT | AMY DAVIDSON

ANNALS OF SCIENCE

GERMS ARE US

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

BY MICHAEL SPECTER

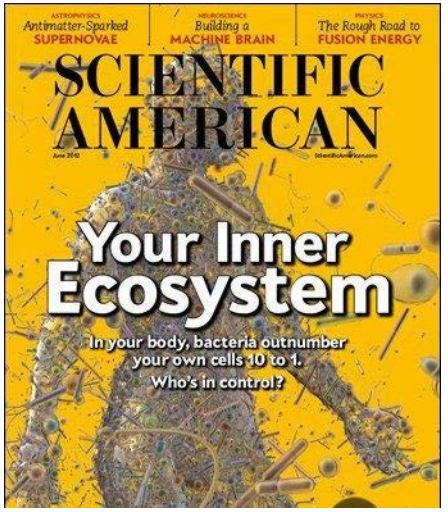
OCTOBER 22, 2012

Share 6.6K | Tweet 431 | +1 | PRINT | MORE

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

tem por objetivo sequenciar amostras dos mais variados ambientes do planeta

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.

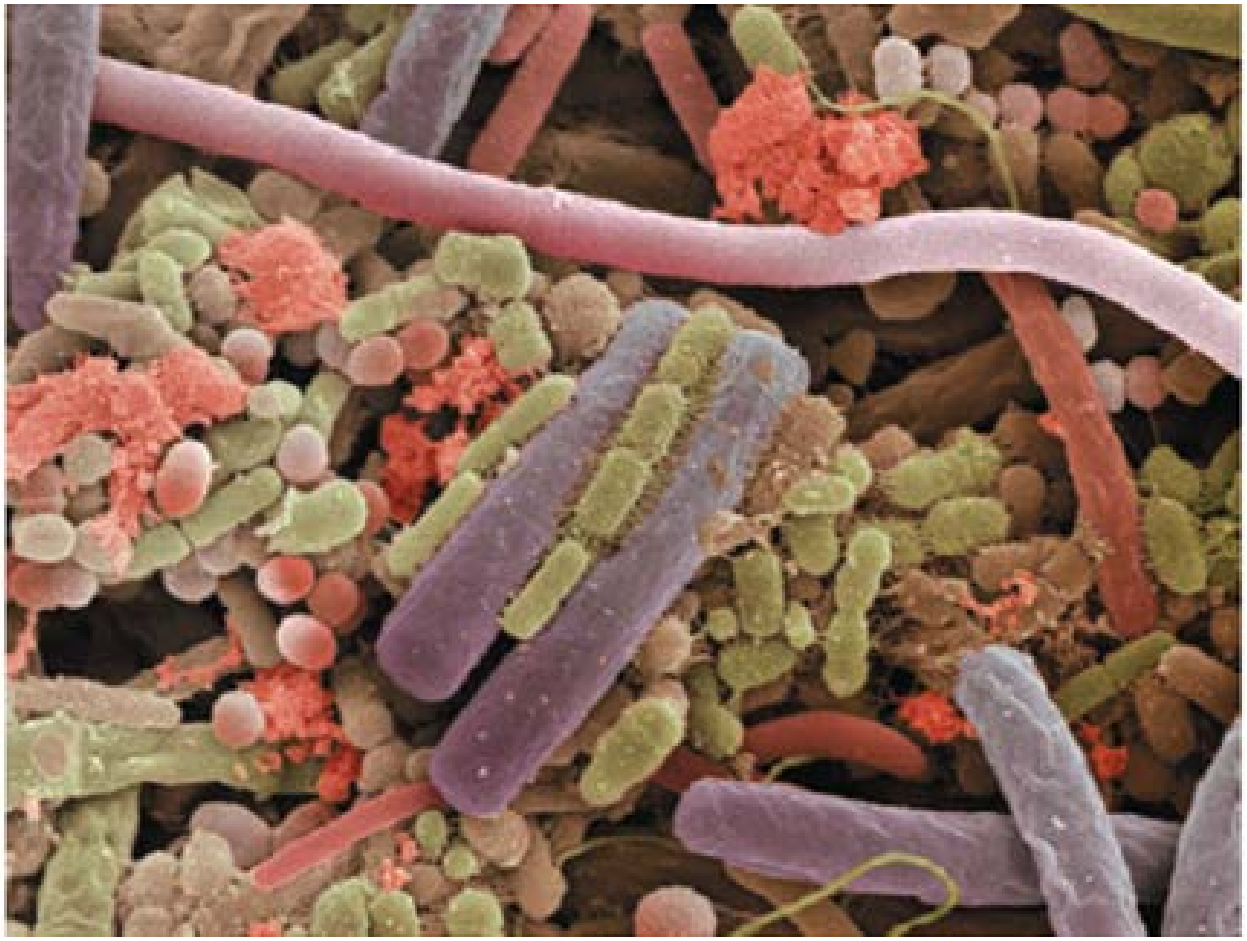
News

Earth Microbiome Project: Rick Stevens at TEDxNaperville

Não confundir...

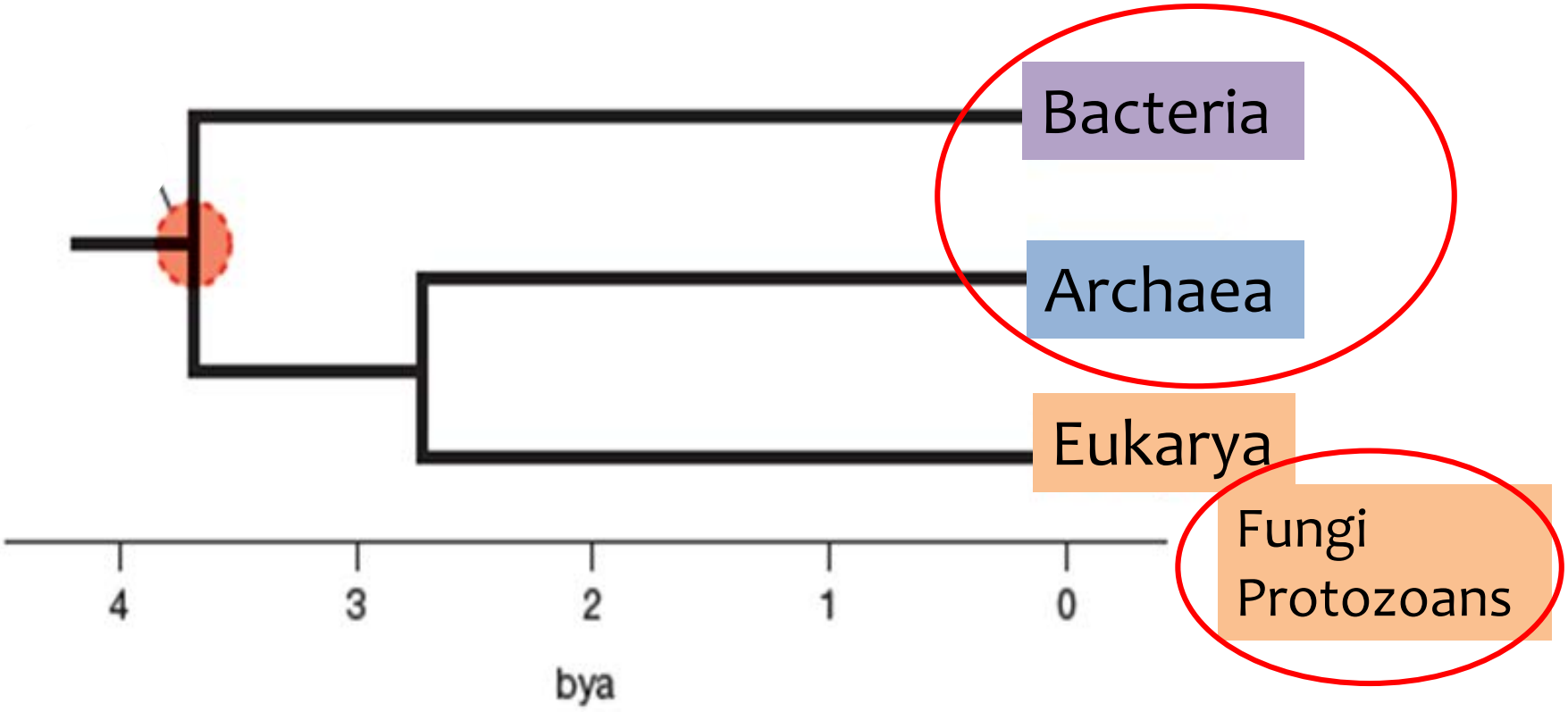
- Earth Microbiome Project
- com
- Earth Biogenome Project (EBP)
  - Este é um projeto lançado em 2017 que pretende sequenciar “all life on Earth”
  - voltado para eucariotos

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



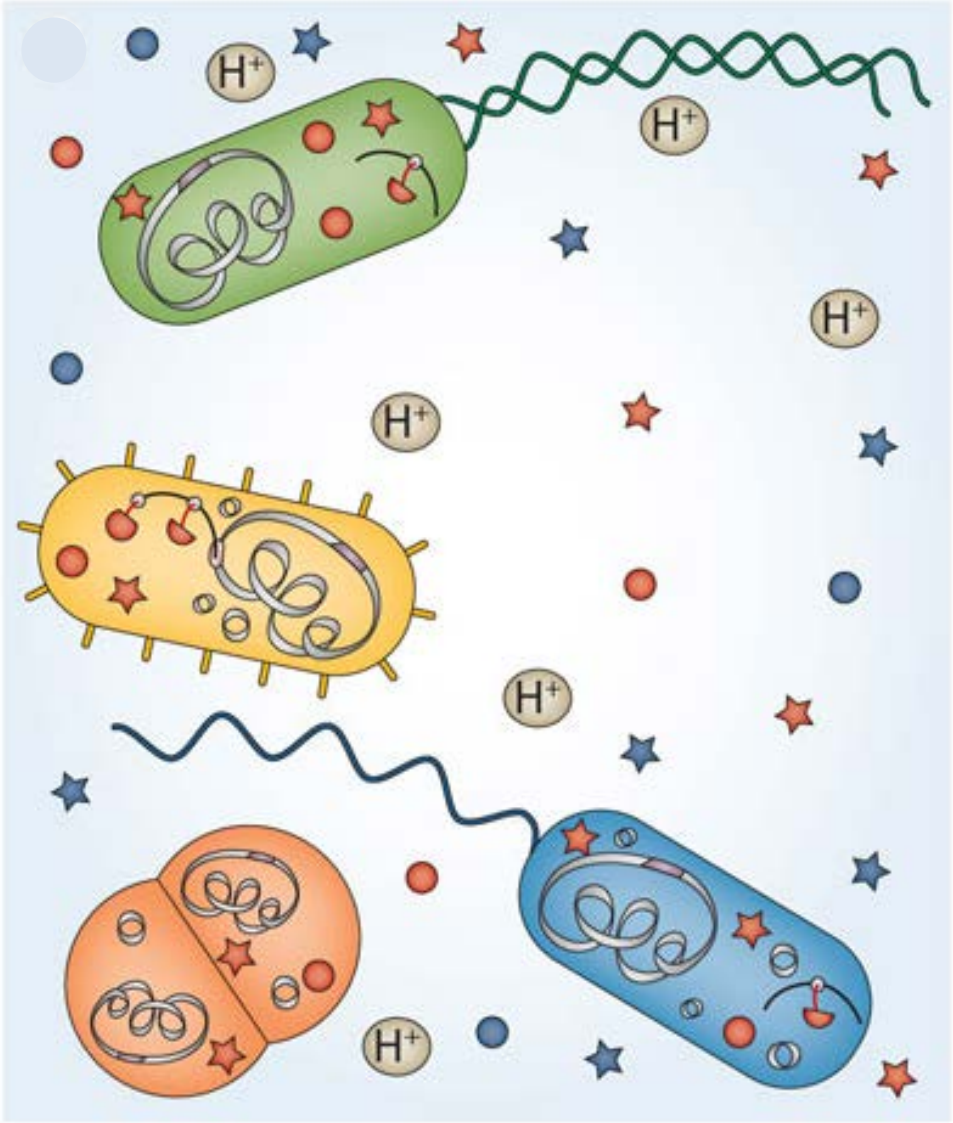
ecossistema microbiano

Microbiotas contêm variedade de microrganismos



+ Vírus e Bacteriófagos

Definição de Microbioma



Genes, Genomas,

Proteínas e Metabólitos da Microbiota

+

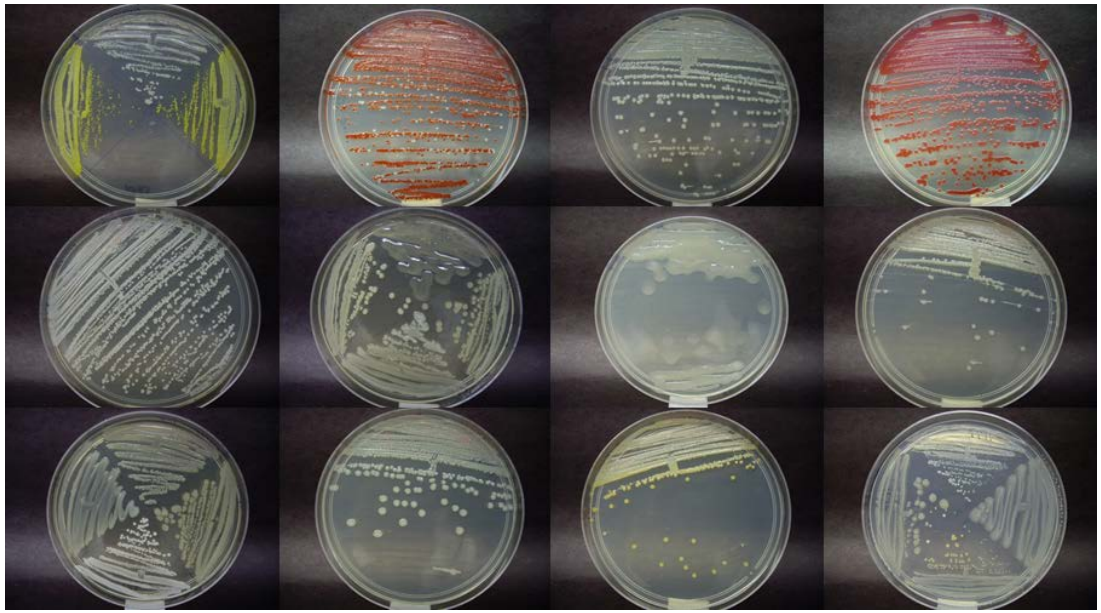
Proteínas e Metabólitos da resposta do Hospedeiro à interação com a microbiota

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?

Tradicionalmente com Abordagens dependentes de cultivo



Cultivo de bactérias em meio sólido

Imagem: Julio Oliveira

Porém...

cultivo significa ser capaz de fazer o micro-organismo crescer em laboratório; ou seja, "acertar" o meio de cultura do qual o micro-organismo precisa, assim como demais condições de sobrevivência

Dado empírico: a **fração cultivável** da vasta riqueza microbiana da biosfera é muito pequena (estimada em apenas 1%). Ou seja, para 99% dos procariotos que se estima que existam, não sabemos como cultivá-los

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)

Esta são as Meta-ômicas

MetaGenômica e MetaTranscritômica

Amostra ambiental

Extrair o DNA
(ou RNA)

Sequenciar

Analisar as sequências de
DNA: metagenômica
cDNA: metatranscritômica
Bioinformática!

Sequenciamento de DNA
alto-desempenho

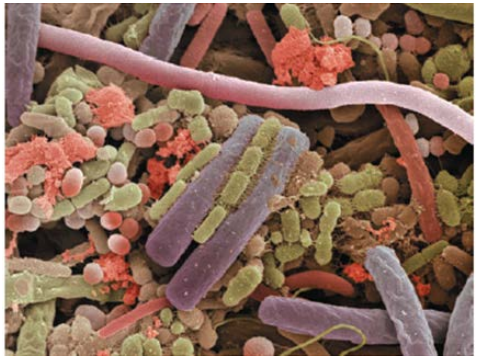
Tecnologias de sequenciamento

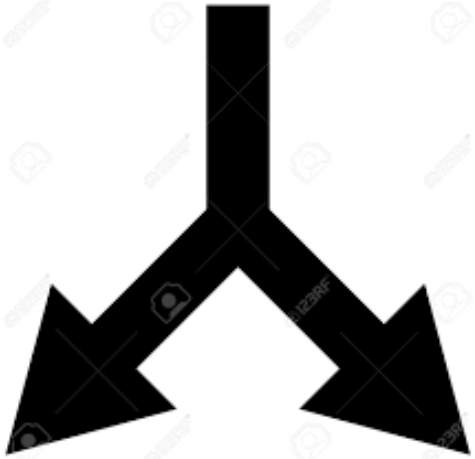
- NGS – next generation sequencing
  - Illumina
    - 90% do mercado
    - Em metagenômica talvez seja perto de 100%
  - PacBio
    - Long reads
  - Nanopore
    - Long reads

Metagenômica é Big Data

- Uma corrida de sequenciamento (Illumina) de uma amostra ambiental resulta em milhões de reads
- 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 em uma única corrida do sequenciador
- A bioinformática é essencial

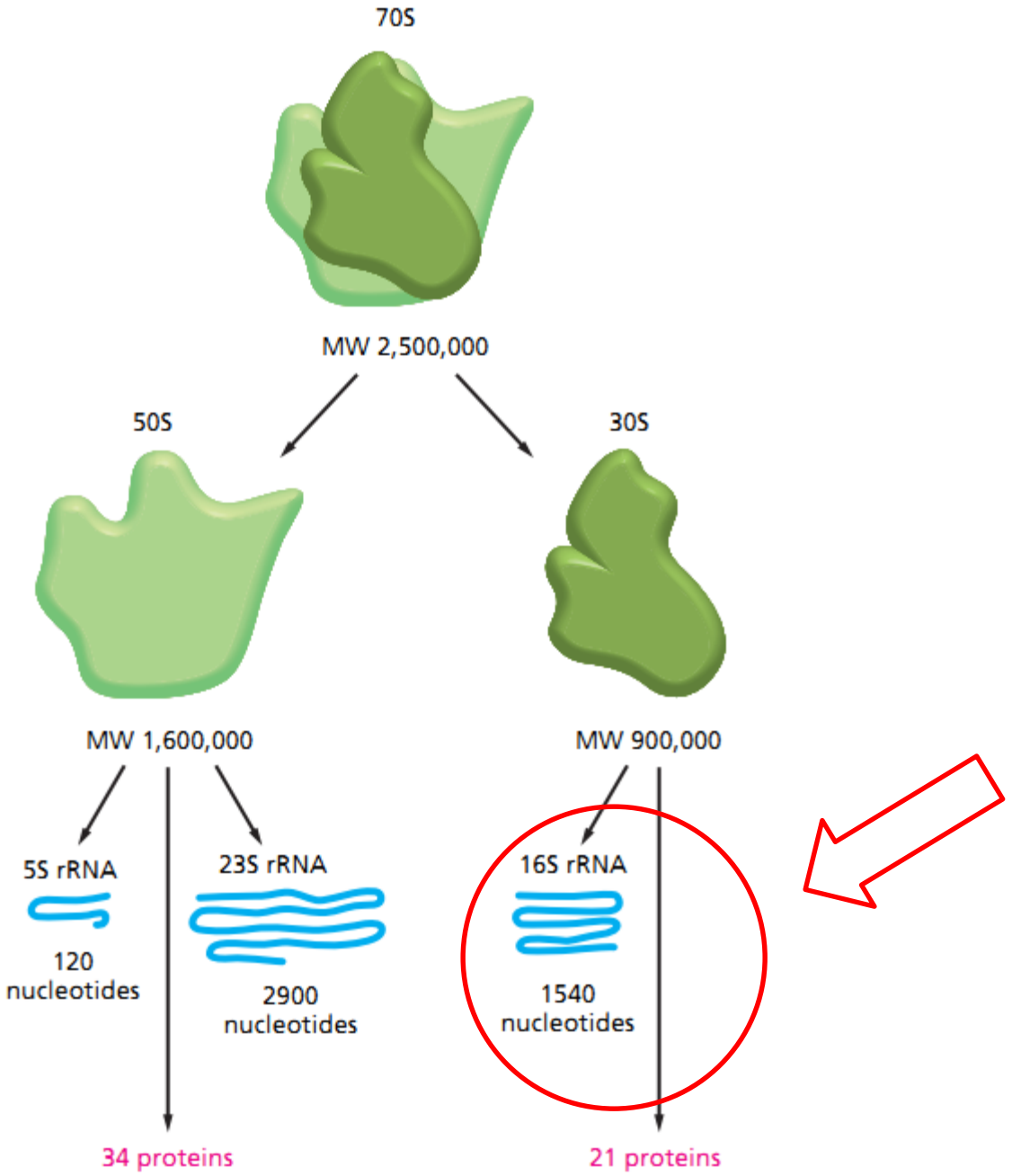
Metagenômica: tipos de Dados





16S / 18S / ITS

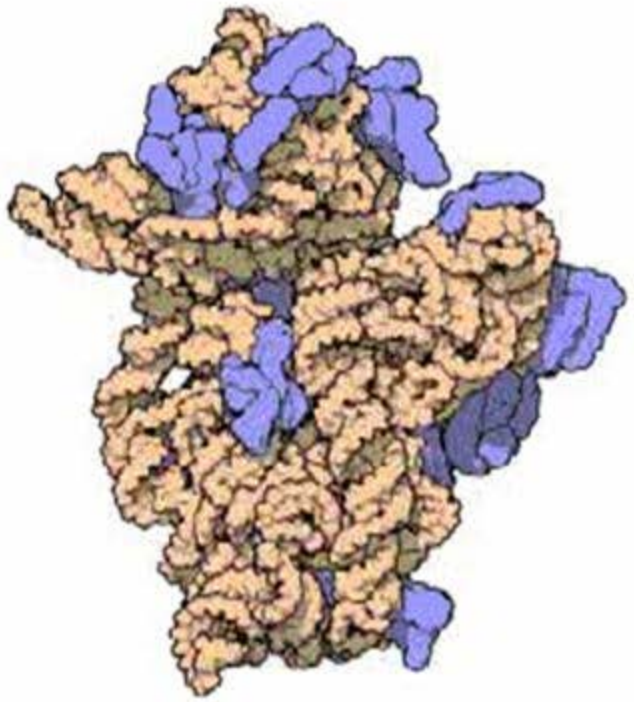
DNA total ou shotgun



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/

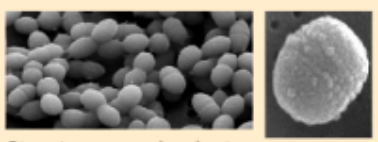
Sequenciamento da unidade 16S do RNA ribosomal

- 16S é um marcador
- a ideia é “pescar” um trecho do 16S de “todos” os procariotos presentes na amostra, e sequenciar esses trechos
- fazendo a classificação taxonômica desses trechos, teremos um perfil da população de procariotos presentes na amostra

Exemplo de perfil taxonômico obtido com 16S -- microbioma humano (próximo slide)

- separado por região do corpo
- Exercício:
  - indique quais são os principais grupos bacterianos em cada região do corpo que foi amostrada
    - fezes, bochecha, placa dentária, língua, nariz, vagina, pele

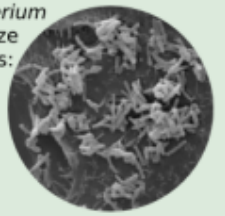
A map of diversity in the human microbiome



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

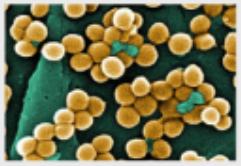
Propionibacterium acnes lives on the skin and nose of most people

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

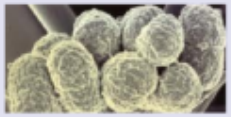


Lactobacillus species (*L. gasseri*, *L. jensenii*, *L. crispatus*, *L. iners*) are predominant but mutually exclusive in the vagina


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

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

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

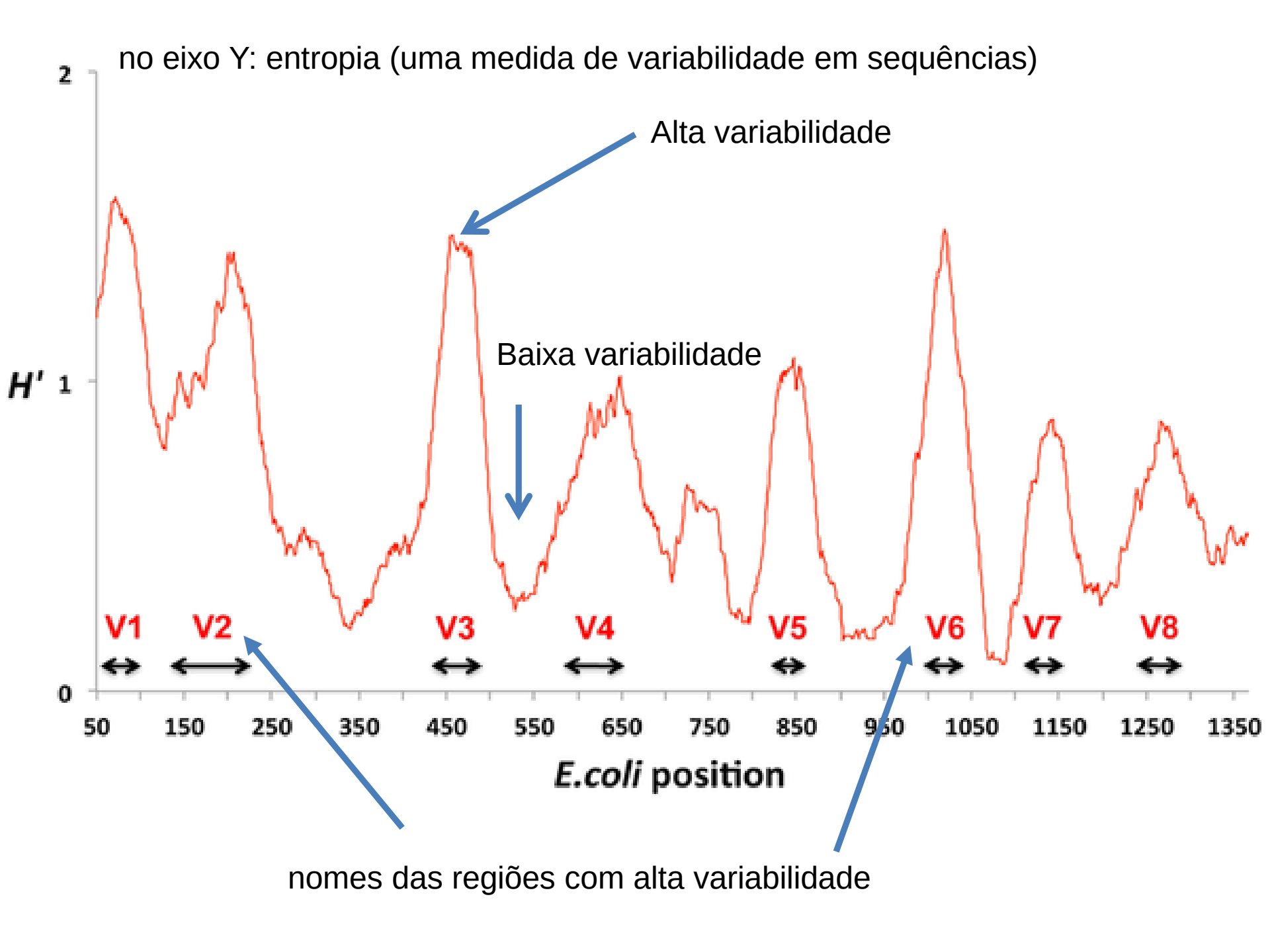
National Institutes of Health Human Microbiome Project

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

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

o 16S rRNA é um bom marcador, por que...

- tem regiões altamente conservadas entre diferentes espécies de bactérias e de arquéias
  - o que permite "primers universais"
- tem também regiões de alta variabilidade, o que permite distinguir o 16S entre diferentes organismos (geralmente apenas até o nível de gênero)

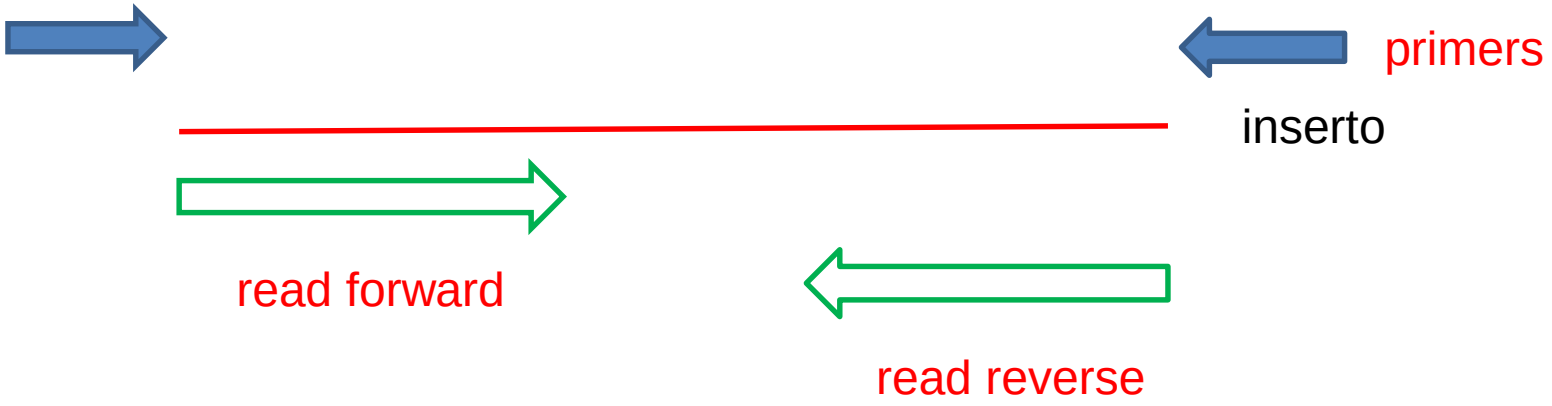

no eixo Y: entropia (uma medida de variabilidade em sequências)
Alta variabilidade
Baixa variabilidade
H'
2
1
0
V1
V2
V3
V4
V5
V6
V7
V8
50
150
250
350
450
550
650
750
850
950
1050
1150
1250
1350
E.coli position
nomes das regiões com alta variabilidade

tamanho esperado do inserto para V3/V4

- 550 bp

O que é o inserto?

- Os reads podem ser paired-end ou single-end

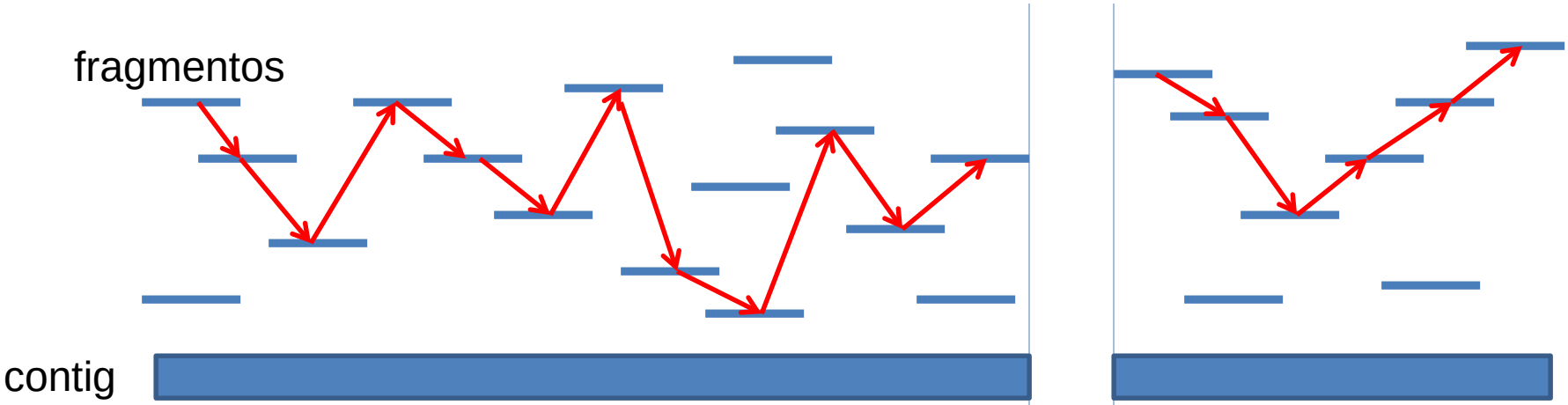


Nesta ilustração, temos paired-end, e o “miolo” do inserto não será sequenciado

DNA shotgun

- Sequenciar o DNA total da amostra
- Resultado
  - Milhões de fragmentos (reads)
  - Mistura dos DNAs dos diversos organismos presentes
  - fragmentos precisam ser montados

Montagem de genomas



...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...

a montagem é possível por causa de sobreposição entre os reads

buraco

buracos ocorrem quando trechos do genoma não são cobertos por reads

Montagem

- Montagem é essencial para
  - Análise funcional (genes)
  - Recuperação de genomas (falaremos disto mais tarde)
- Objeto principal resultante
  - contigs
  - um contig é uma sequência que foi montada
  - presume-se que um contig se refere a uma região *contígua* de um genoma de um organismo presente na amostra

Comparação entre 16S e shotgun

- 16S
  - Composição e estrutura da microbiota
    - “perfil taxonômico”
- DNA total ou 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 (não mais do que gênero)	Mais precisa, podendo chegar a especie, e talvez cepas

Em dados de 16S é comum os reads serem agrupados em OTUs

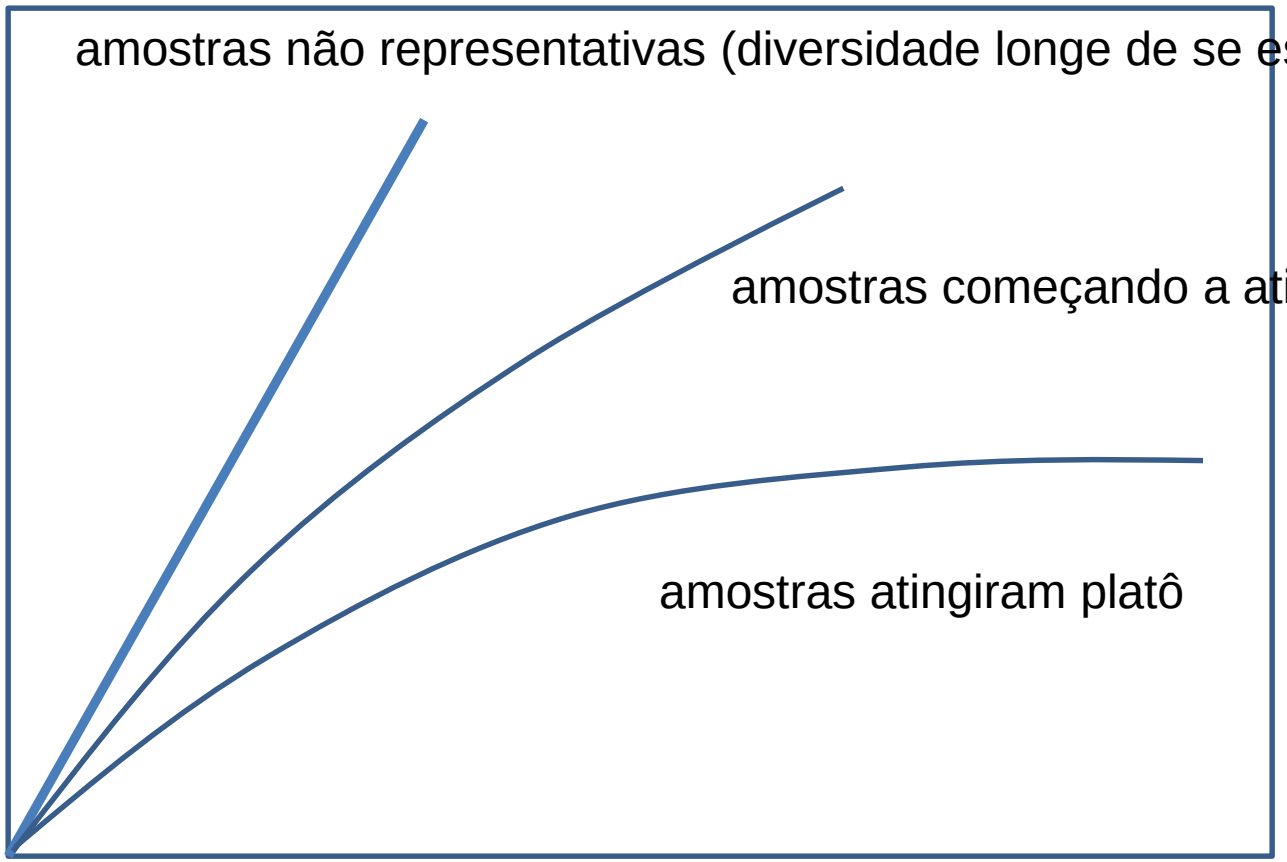
- *Operational Taxonomic Unit* ou *Unidade taxonômica operacional*
- Ideia básica: agrupar os reads em caixinhas por meio de similaridade de modo que
  - numa dada caixinha, todos os reads se parecem entre si com pelo menos 97% de identidade
  - não existe read em nenhuma outra caixinha que seja pelo menos 97% similar a reads desta caixinha
- Pega-se uma sequência representativa de uma caixinha, e faz-se uma busca num banco de 16S
- Se houver similaridade de pelo menos 97%, podemos rotular a OTU com o mesmo rótulo da sequência do banco
- Caso contrário, a OTU fica sem classificação

Que perguntas básicas queremos fazer com dados metagenômicos?

pergunta 1: A amostra é representativa?

- Curvas de rarefação

Curvas de rarefação (ou saturamento)



número de especies ou OTUs

número de amostras

Pergunta 2: Quem está na amostra?

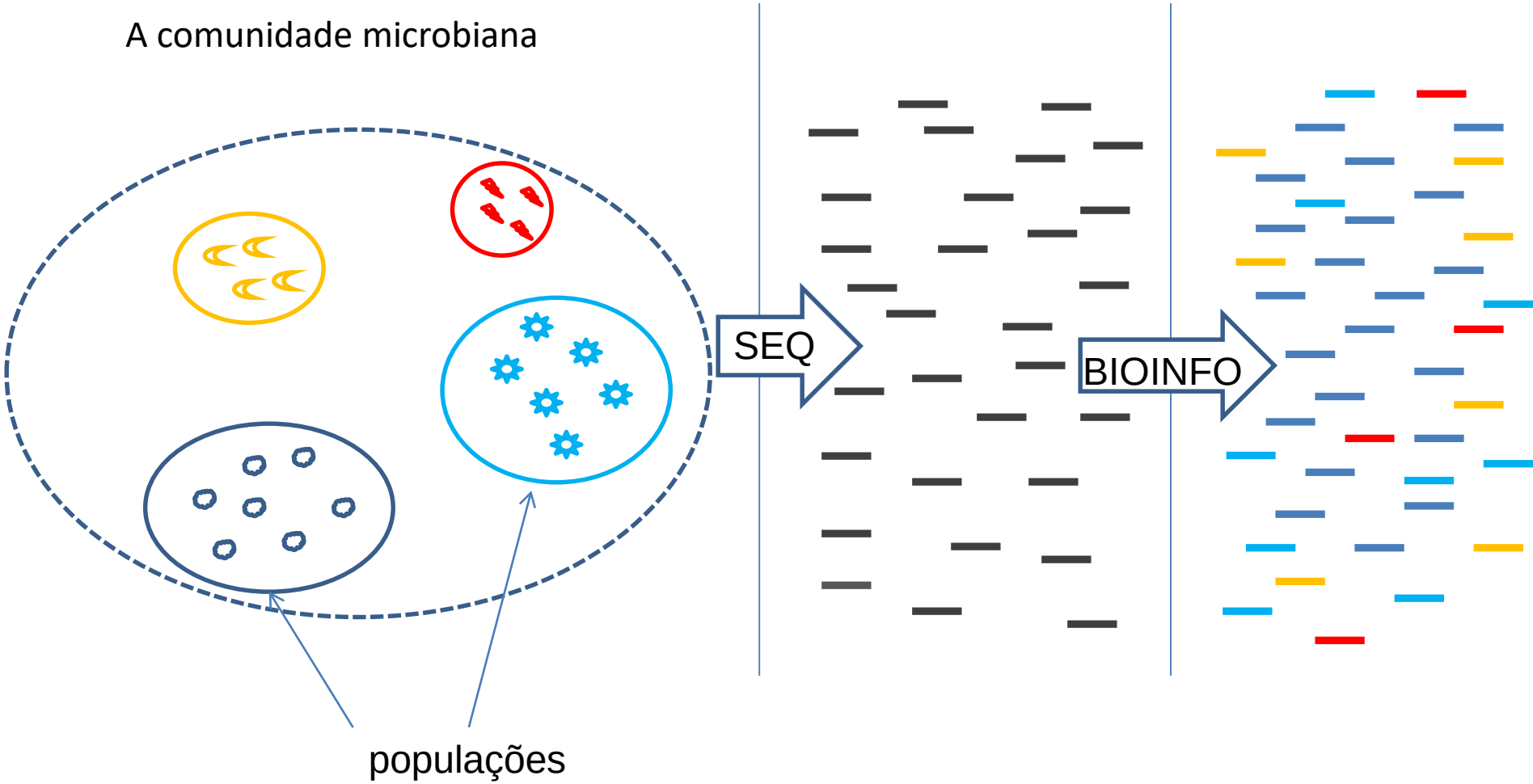
- Identificação taxonômica (16S, shotgun)
- Recuperação de genomas (shotgun)

Taxonomia

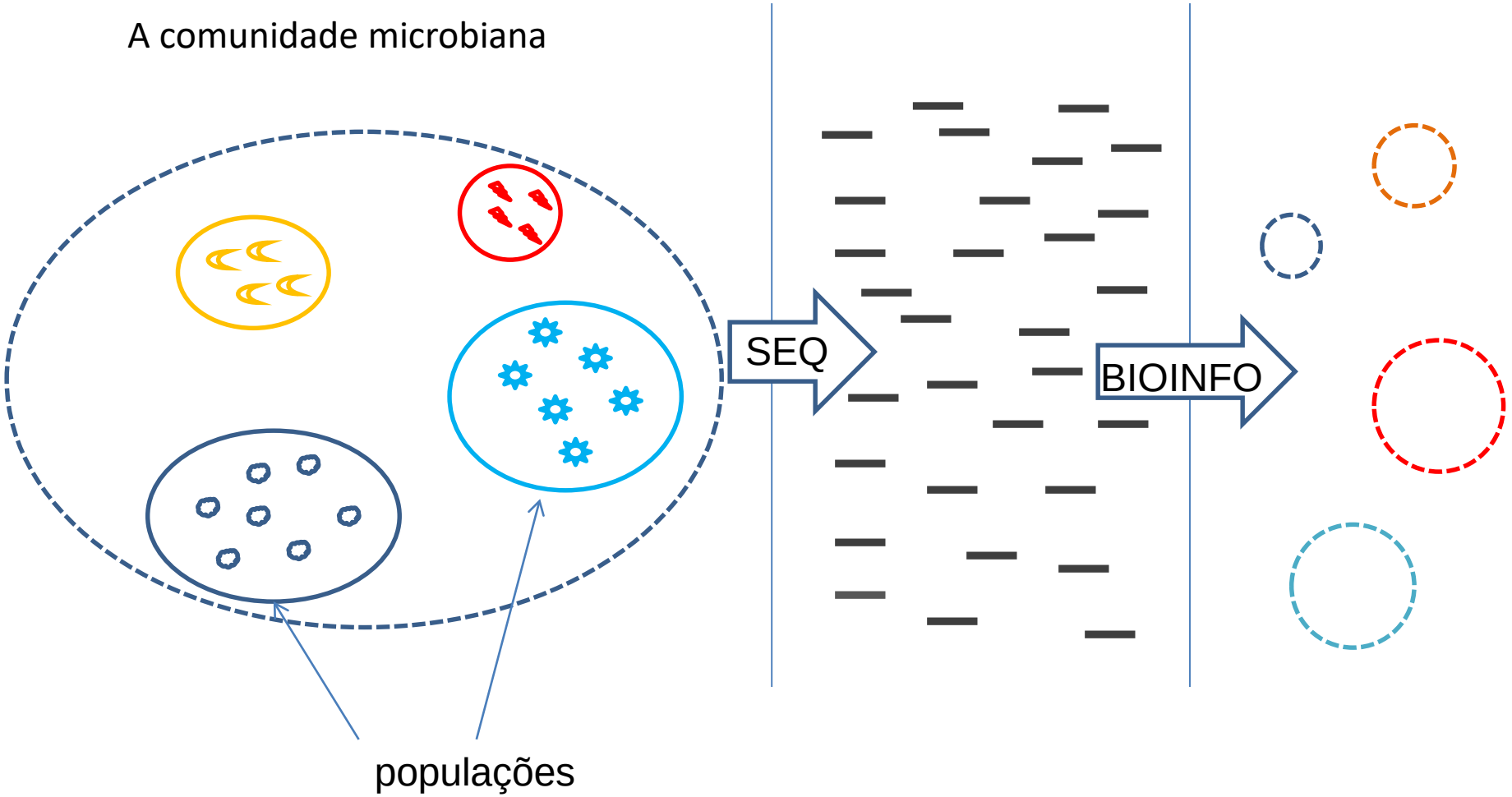
- *Xanthomonas citri*
- Filo: proteobacteria
  - Classe: proteobacteria gama
    - Ordem: xanthomonadales
      - Família: xanthomonadácea
        - Gênero: xanthomonas
          - Espécie: citri

16S / DNA shotgun

A comunidade microbiana

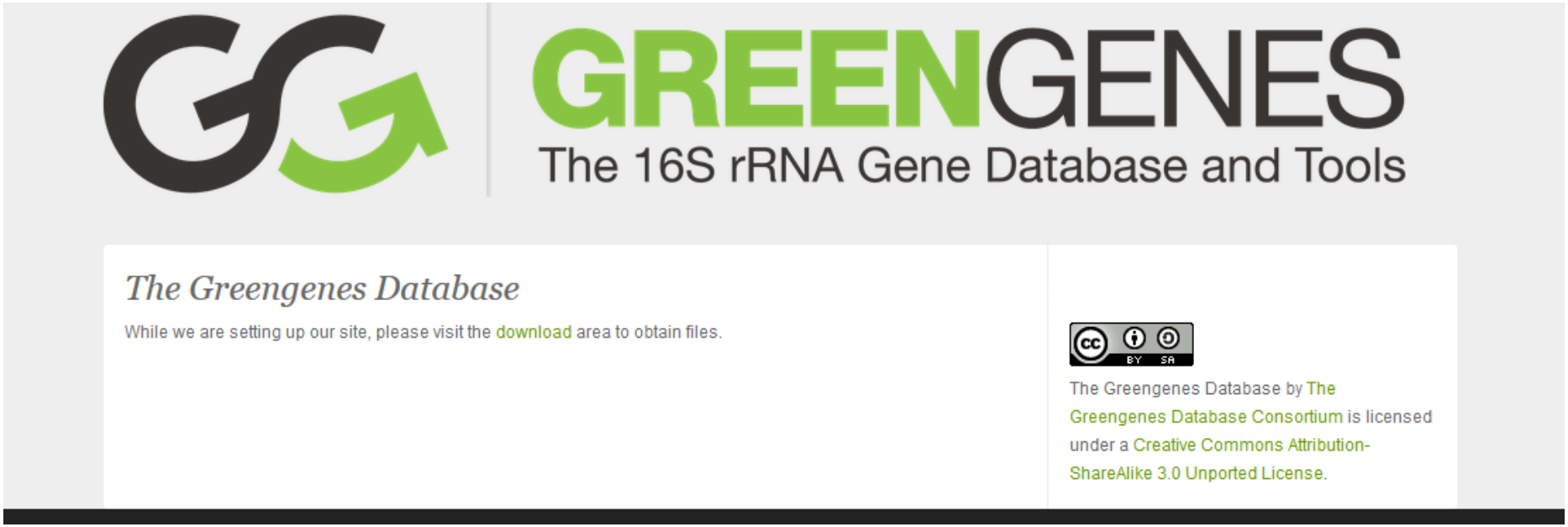


Recuperação de genomas



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

silvangs

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

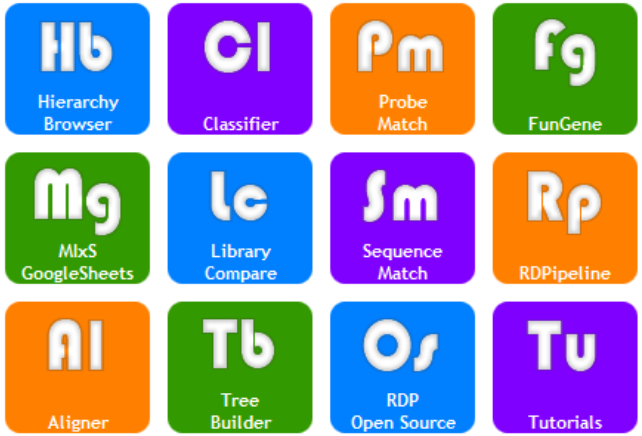
RDP Release 11, Update 5 :: September 30, 2016

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.



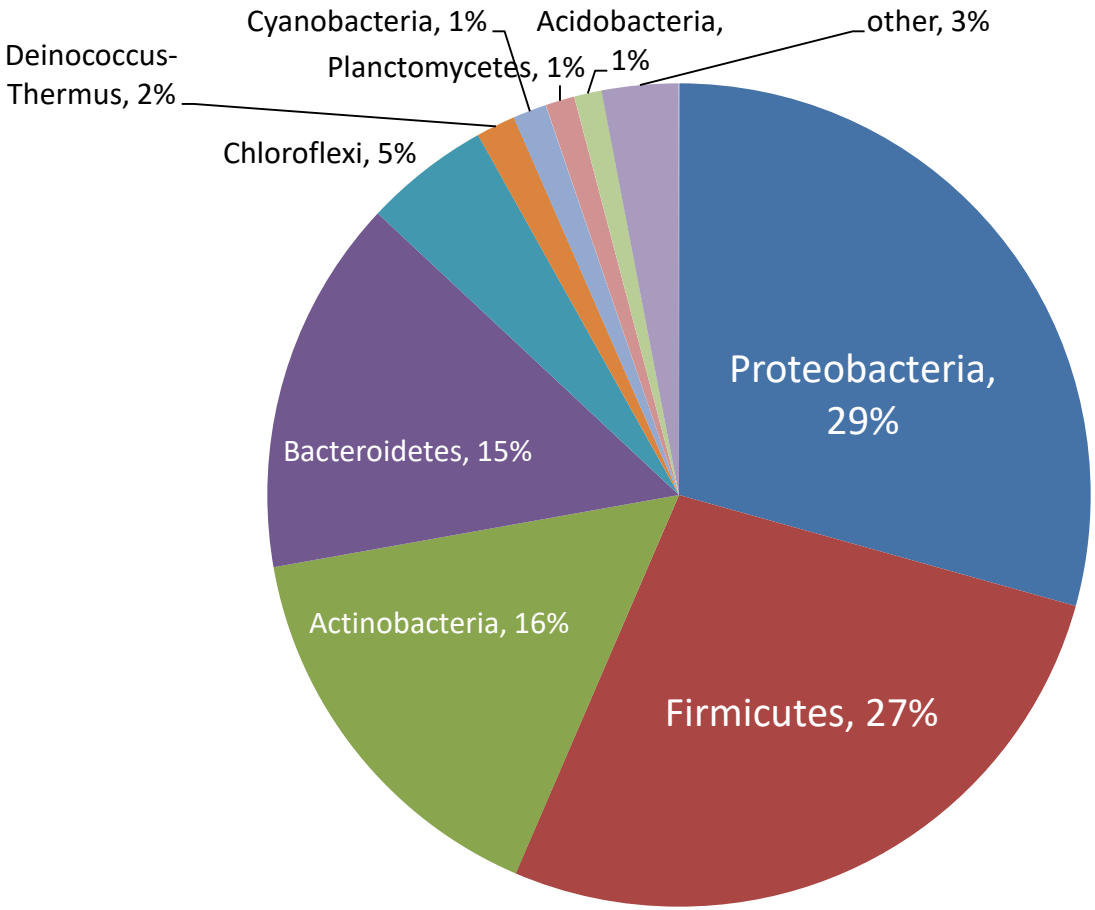
RDP's mission and funding:

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: nucleotídeos
  - nr: proteínas
  - env_nr: proteínas inferidas de dados metagenômicos
  - refSeq: genomas de referência
  - WGS: whole genome shotgun
    - aqui estão dados de genomas draft de isolados

Classificação taxonômica e abundância relativa: tipicamente expressas por um gráfico de pizza



É preciso cuidado com viéses

- A abundância relativa “observada” pode ser apenas um reflexo das abundâncias relativas de sequências em bancos de dados
- principalmente quando de omitem das tabelas ou gráficos as sequências sem classificação

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

Exercício

- Compare as abundâncias relativas do gráfico de pizza de slide anterior com as abundâncias relativas da tabela do slide anterior (que mostra como eram as abundâncias no GenBank em 2015)
- Os números são muito parecidos!
- Duas hipóteses
  - 1) essa é a abundância relativa na natureza
  - 2) a abundância da amostra é enviesada; apenas reflete o que se tem no banco de dados

O problema do viés dos bancos tem diminuído com o passar do tempo

- Esforços de muitos grupos de pesquisa ao redor do mundo tem gerado sequências de novos grupos taxonômicos, até agora desconhecidos

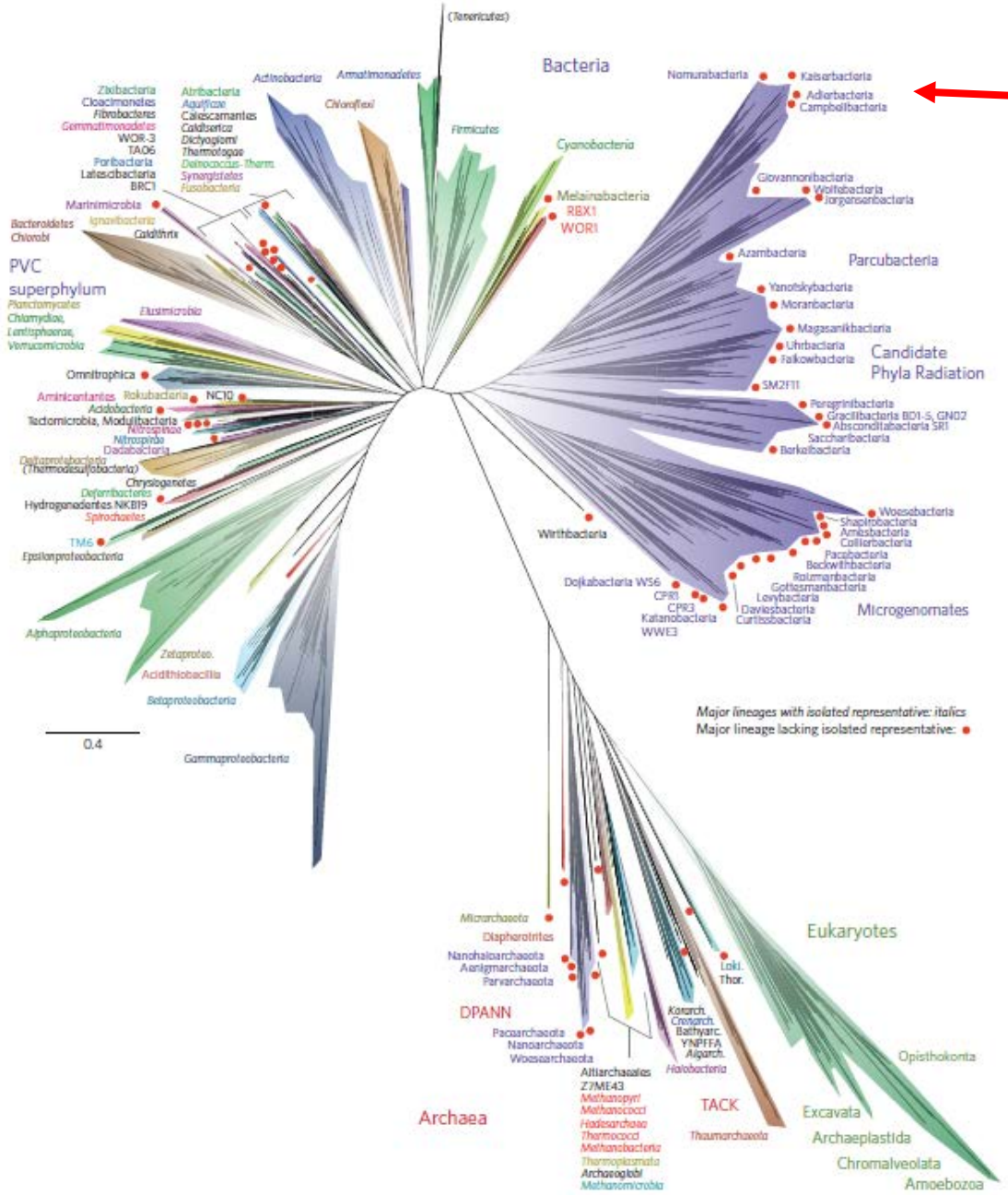


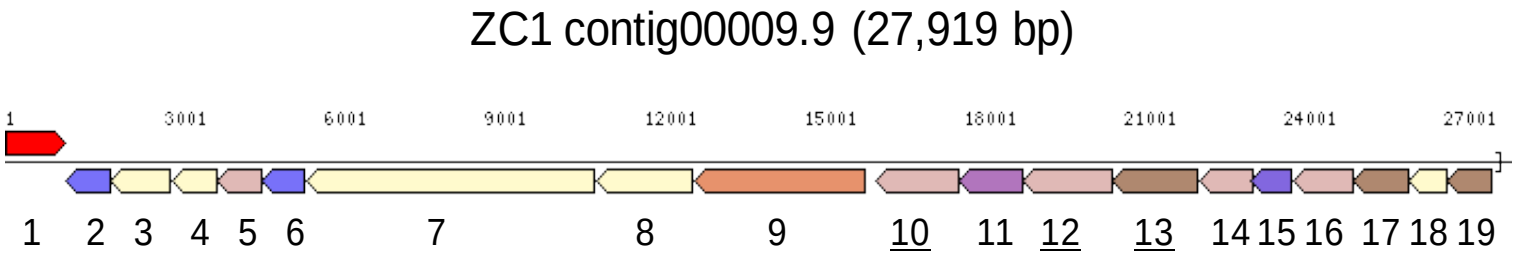
Figure 1 | A current view of the tree of life, encompassing the total diversity represented by sequenced genomes. The tree includes 92 named bacterial

pontos vermelhos indicam novas categorias taxonômicas para as quais não havia isolados quando este paper foi publicado (2016)

pergunta 3: Quais funções estão presentes?

- Em genes (shotgun)
- Em genes expressos (metaTranscritômica)
- precisamos anotar contigs

Exemplo de anotação de um contig

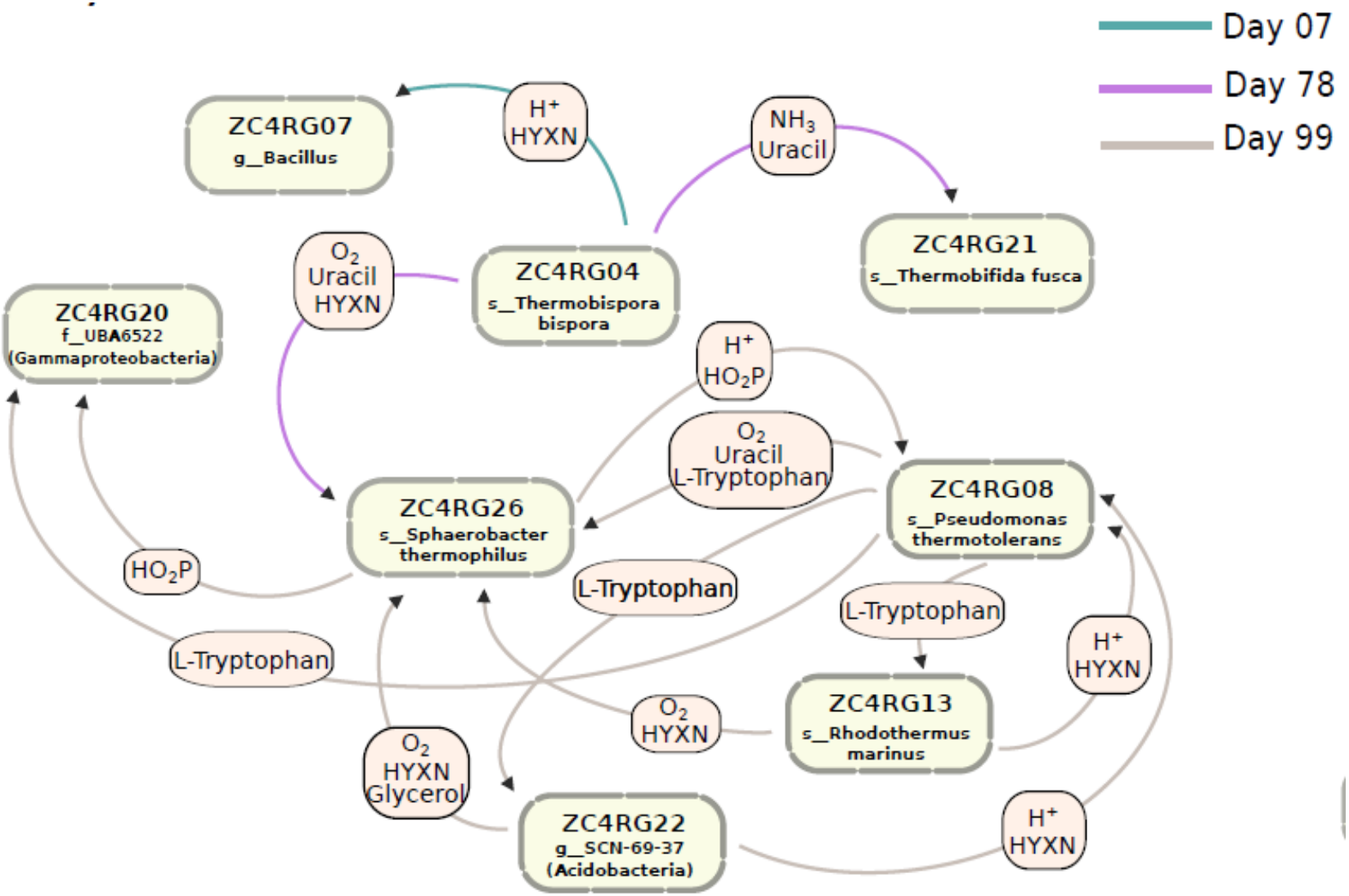


1. Beta-xylosidase (376aa, COG3507)
2. Dehydrogenases (280aa, COG1028)
3. hypothetical protein (379aa);
4. hypothetical protein (283aa)
5. 5-keto 4-deoxyuronate isomerase (280aa, COG3717)
6. Dehydrogenases (267aa, COG1028)
7. hypothetical protein (1799aa)
8. SusD family protein (606aa, pfam07980)
9. 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).

pergunta 4: como são as interações entre os organismos presentes na amostra?

- Responder a esta pergunta exige a inferência de redes de interação
  - geralmente aproximadas por redes de co-ocorrência
    - em diferentes locais
    - em diferentes pontos do tempo
  - co-ocorrência negativa (sempre que *A* está presente, *B* está ausente, ou vice versa) também é importante

Exemplo de rede de interações para amostras seriadas no tempo



É sempre bom ter em mente que análise de dados metagenômicos está sujeita a múltiplas fontes de erro

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