Universidade de São Paulo
Instituto de Química

USP

Análise de Microbiomas
aula 2

João Carlos Setubal

2021

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”
- no caso de OTUs, 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

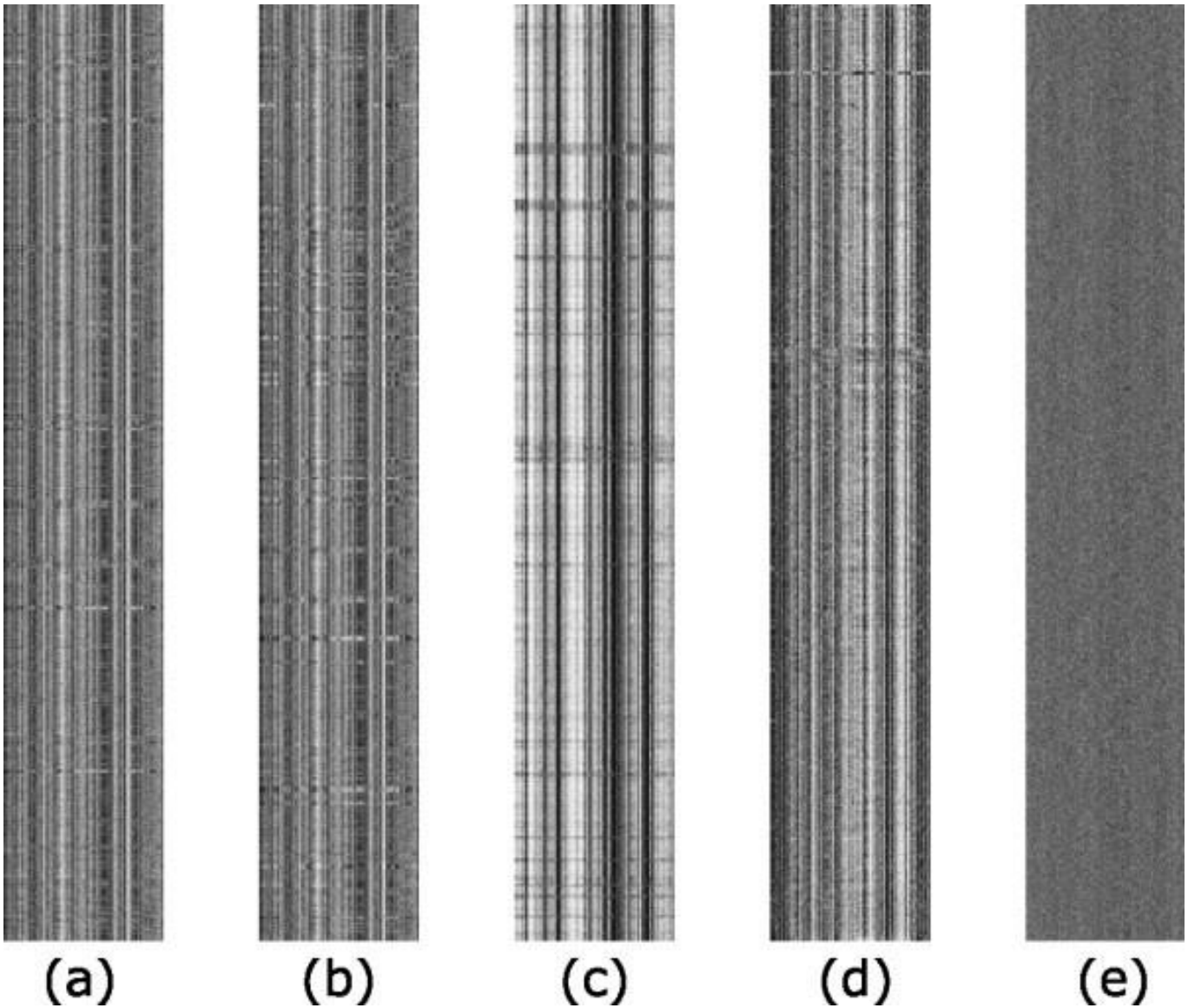
janela	AAAA	AAAC	AAAG	AAAT	ACAA	ACAC	ACAG	ACAT
1	15	2	9	...	0			
2	16	3						
3	14	0						
4	13	2						
5	15	4						
6	12	0						
7	18	1						
8	17	3						
9	16	1						

Exercício

- S_1 = TTCTACTACT
- S_2 = TTGTACTAGG
- S_3 = ACTTCTACTA
- Montar as matrizes de frequências para essas 3 sequências, supondo palavras de tamanho 2
- Quais duas sequências são mais similares entre si em termos das frequências dessas palavras?

Genome “barcodes”

Burkholderia pseudomallei



E. coli K12

E. coli O157

Pyrococcus furiosus

random

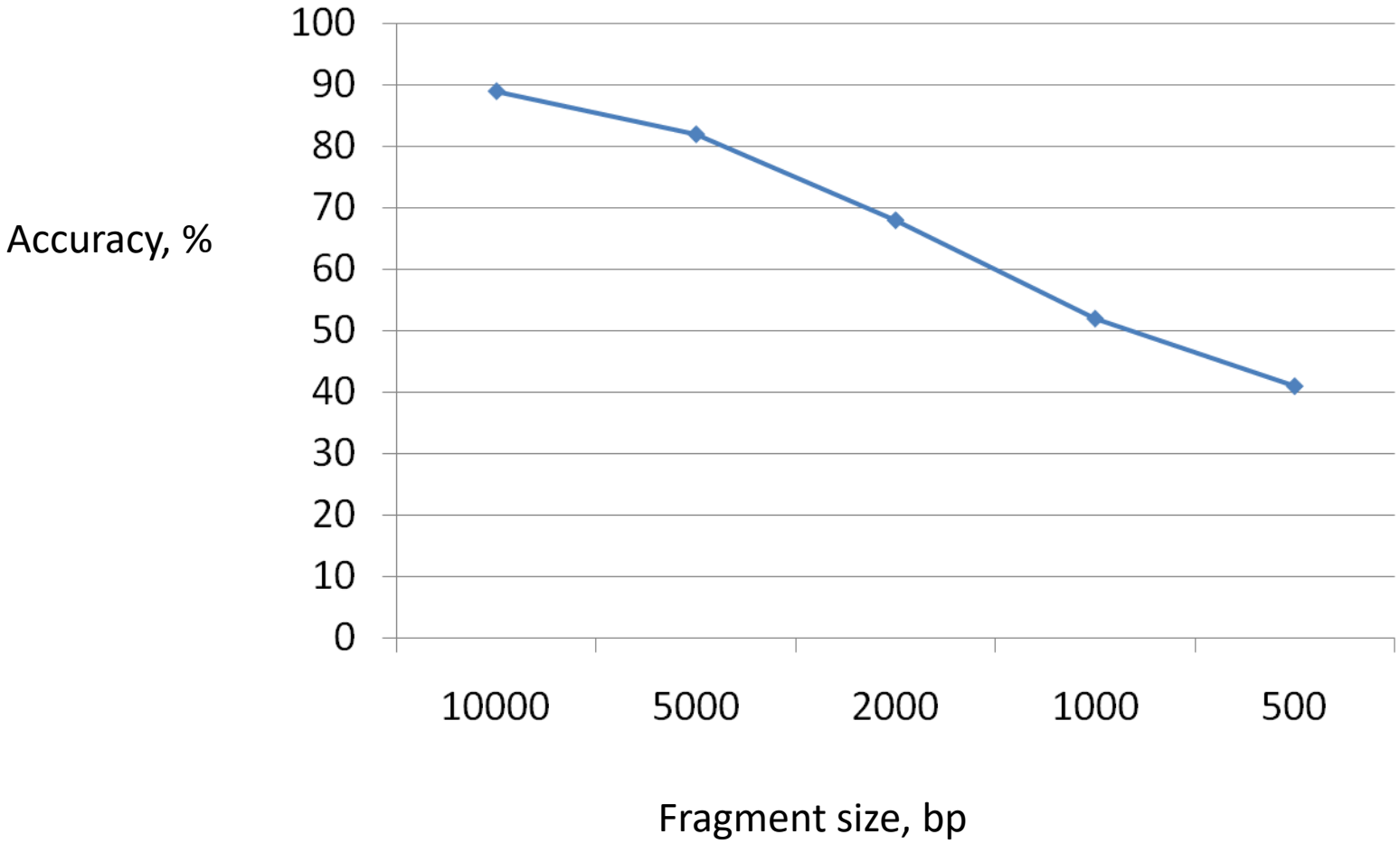
Zhou, Olman, Xu, *BMC Bioinformatics*, 2009

Explicação da imagem anterior

- Cada barra (ou bar code vertical) corresponde a 1 Mbp de um genoma de um procarioto (com exceção da última)
- cada barra pode ser entendida como a matriz de frequência desse fragmento genômico, com valores de frequência convertidos para tons de cinza
- as faixas verticais de cada barra significam k-meros de alta frequência (mais escuro) ou de baixa frequência (mais claro) ao longo desse trecho do genoma
- o fato de que existem essas faixas verticais mostra que diferentes k-meros tem diferentes frequências ao longo do genoma, e essas frequências são razoavelmente constantes ao longo do genoma

- as faixas horizontais indicam prováveis regiões de transferência horizontal, um fenômeno comum em bactérias. Essas regiões tem frequências de k-meros diferentes das frequências que caracterizam o genoma
- a última barra é uma sequência artificial, em que as bases A,T,C,G foram escolhidas aleatoriamente
  - note sua falta de estrutura
- Conclusão: podemos usar as frequências de k-meros como assinaturas genômicas
  - crie uma biblioteca com esses “códigos de barras”
  - compare as frequências de k-meros de um novo fragmento com os códigos da biblioteca; se houver “similaridade suficiente”, teremos uma identificação

Esta técnica não funciona bem com fragmentos curtos

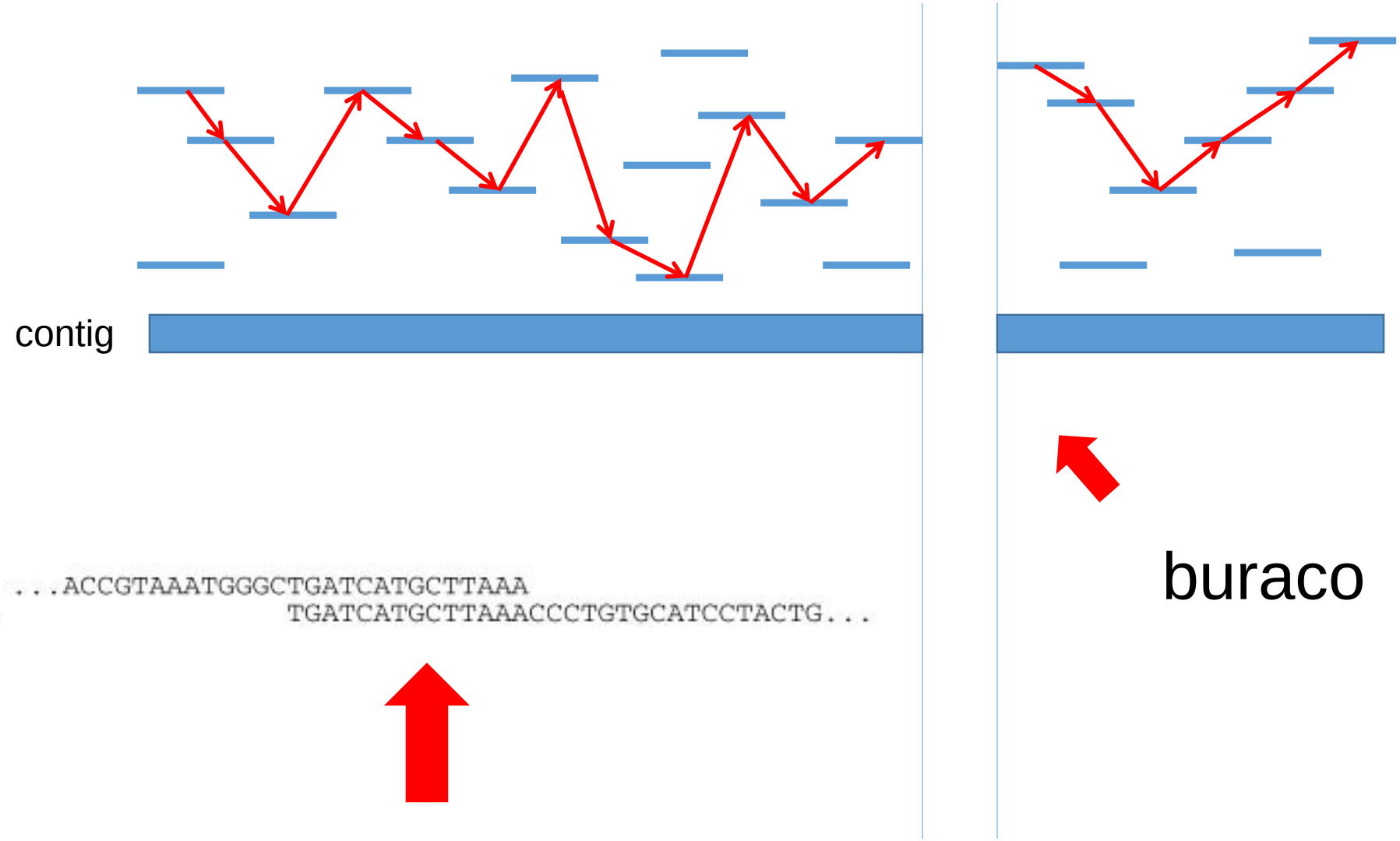


Zhou et al, 2009 simulated data

Exercício

- Que explicação você daria para o decréscimo de acurácia com decréscimo de comprimento?

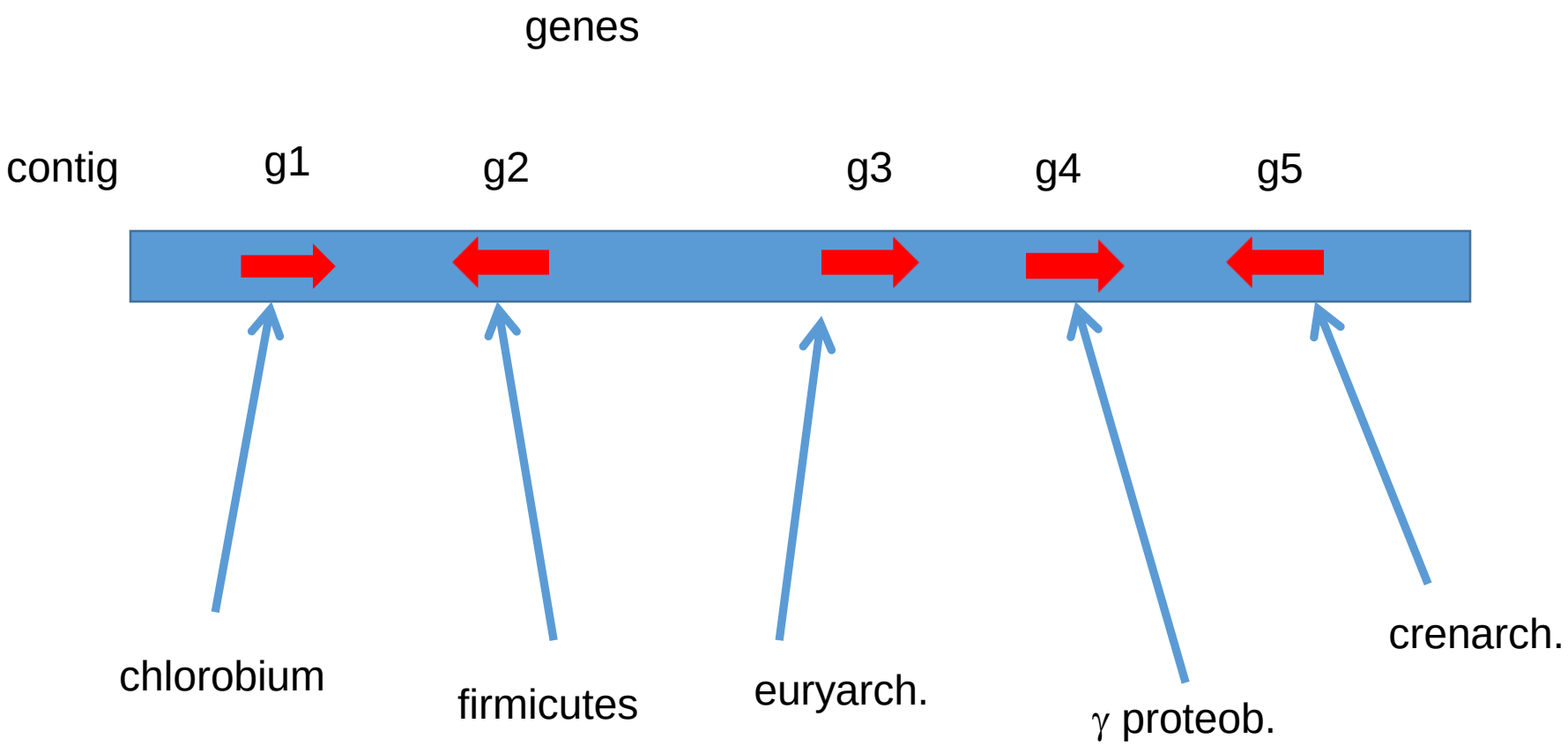
Montagem de genomas



Montagem

- Em genomas bacterianos isolados, é um processo razoavelmente bem compreendido
- Em metagenomas há velhas e novas dificuldades
  - Mistura de organismos
    - pode causar quimeras
    - Transferência lateral pode causar erros
  - Repetições
    - sempre um problema, especialmente se forem longas
  - Tamanho dos conjuntos de dados
    - Chegando a bilhões de reads

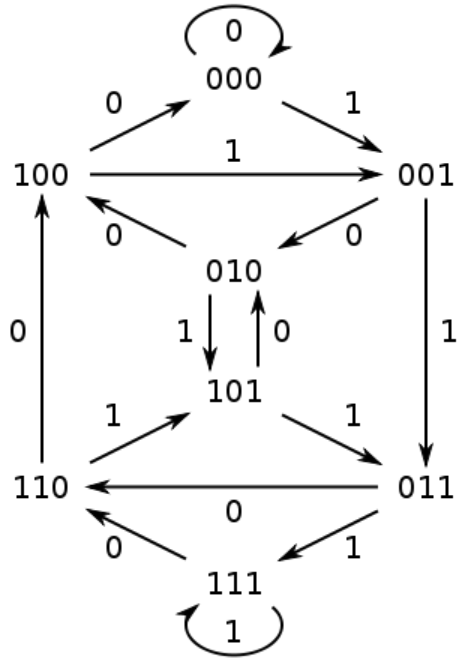
Exemplo de quimerismo: os organismos identificados são muito discrepantes entre si



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

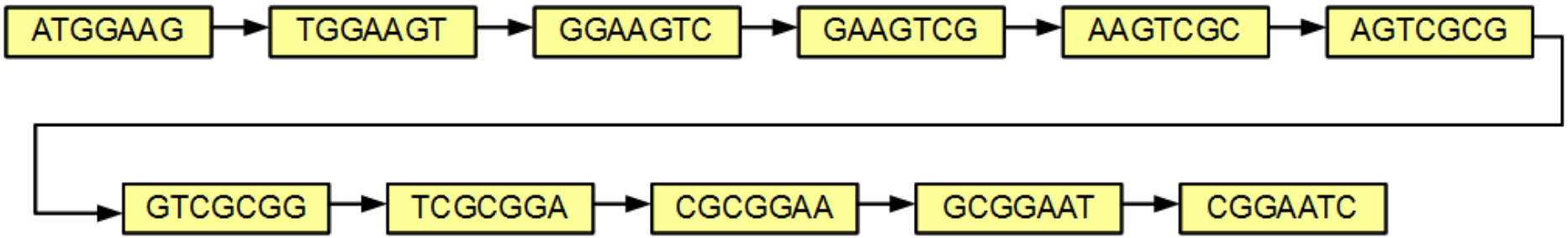
Grafo de de Bruijn em montagem

sequence **ATGGAAGTCGCGGAATC**

7mers

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

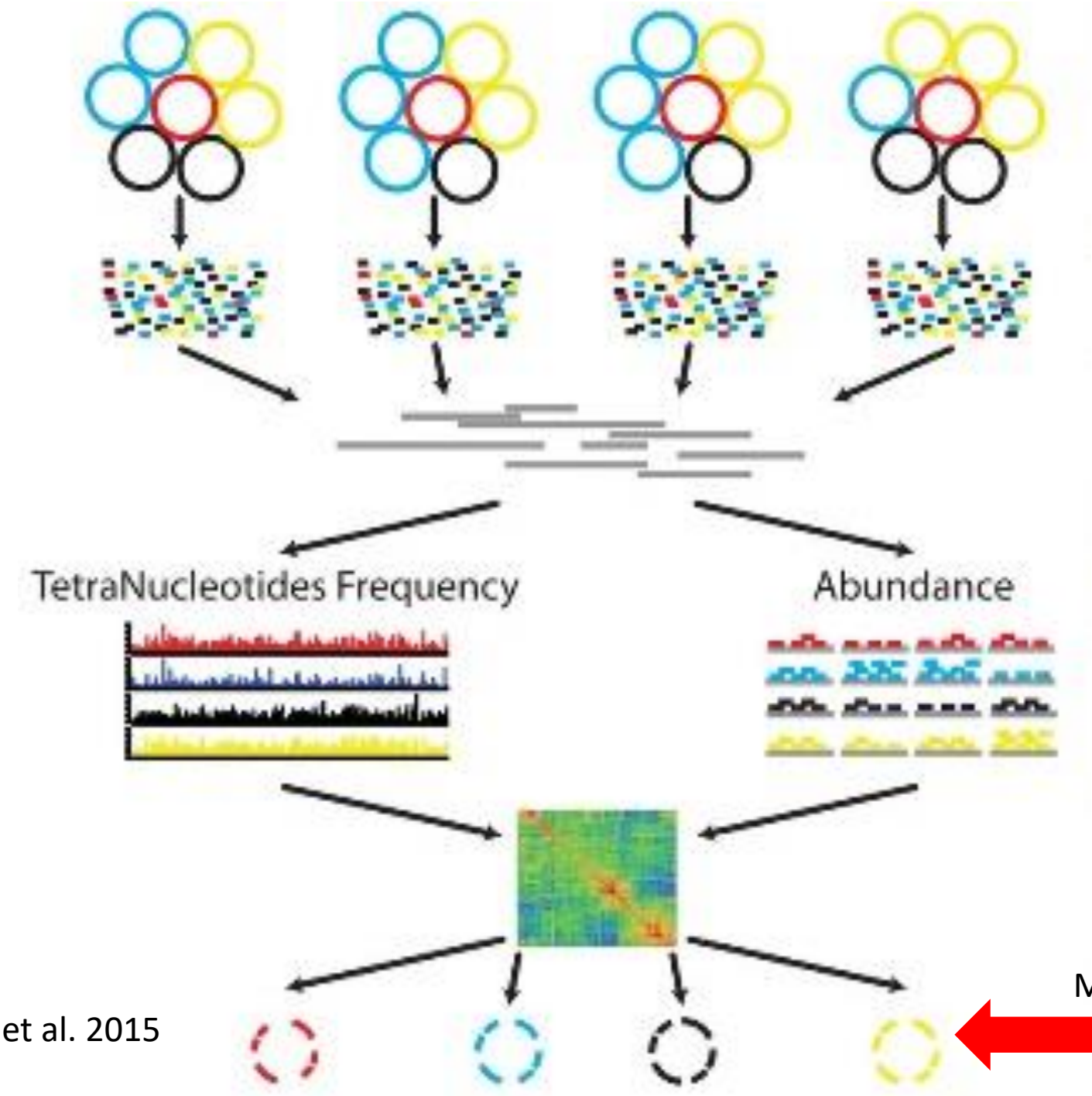
de Bruijn graph



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

Metagenome-assembled genomes ou MAGs

- Genomas de micro-organismos obtidos a partir de sequenciamento metagenômico shotgun e posterior agrupamento/montagem
- Ou seja, genomas que NÃO são obtidos a partir de sequenciamento de isolados
- Tornaram-se ferramenta importante no estudo das microbiotas



Kang et al. 2015

explicação do diagrama anterior

- É para ler de cima para baixo
- no topo: representação dos genomas nas amostras
- sequenciamento
- montagem
- análise de contigs por tetranucleotídeos e abundância relativa
- recuperação dos genomas

MetaWRAP é um ótimo pipeline para recuperar genomas

Uritskiy et al. Microbiome (2018) 6:158
https://doi.org/10.1186/s40168-018-0541-1

Microbiome

SOFTWARE Open Access

MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis

Gherman V. Uritskiy, Jocelyne DiRuggiero* and James Taylor*

fluxo de processamento no metaWRAP

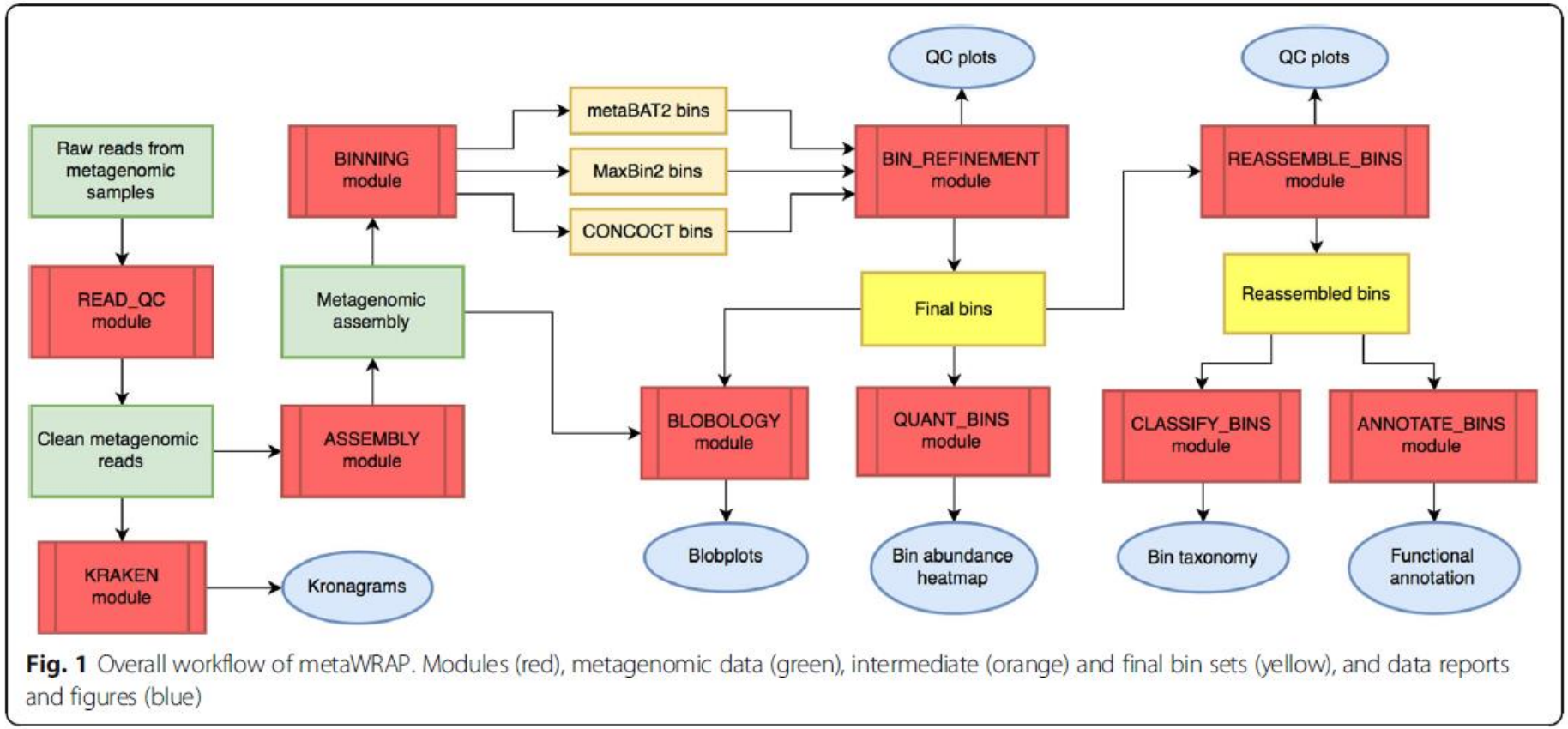


Fig. 1 Overall workflow of metaWRAP. Modules (red), metagenomic data (green), intermediate (orange) and final bin sets (yellow), and data reports and figures (blue)

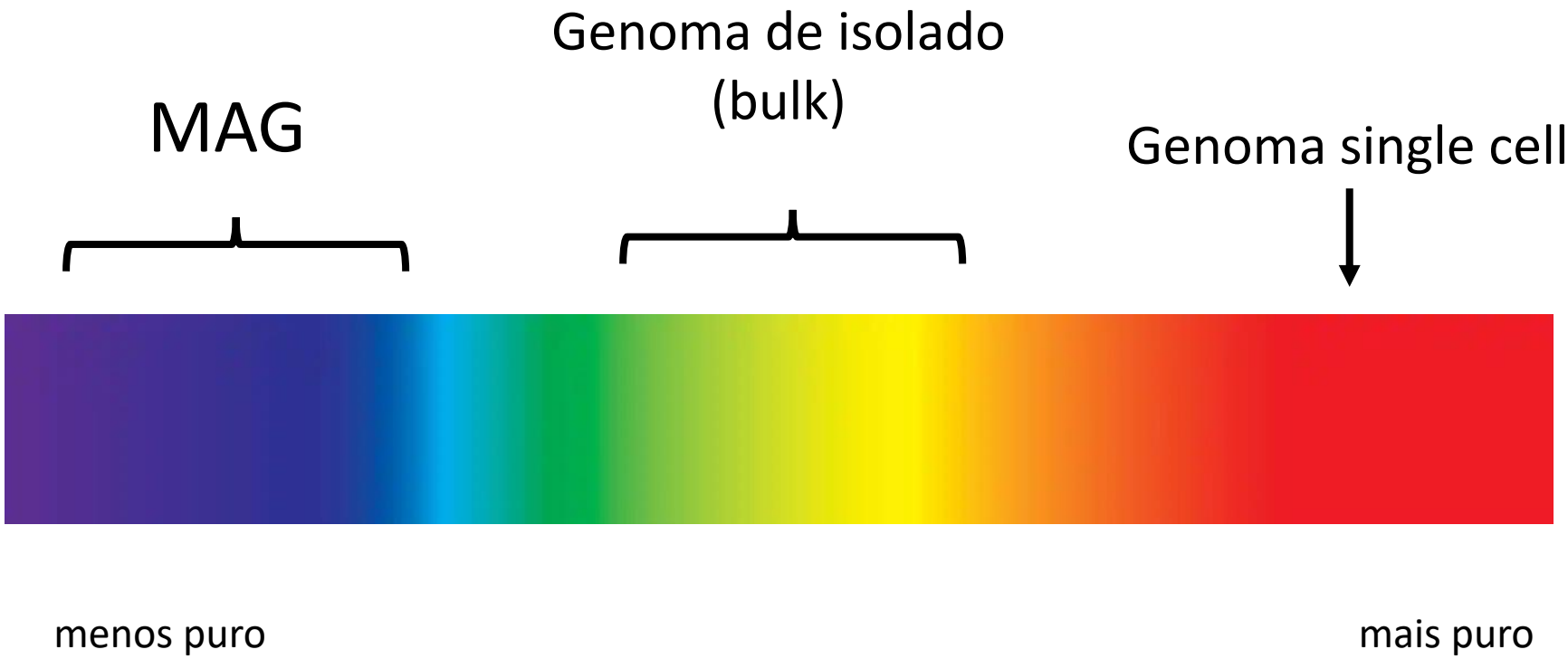
MAGs podem nos dar uma visão detalhada do ambiente de interesse

- Classificação pode chegar ao nível de espécie
  - Em alguns casos, de cepas
- abundância relativa
- Informações sobre genes e suas funções
- Redes de interações
  - Positivas e negativas
  - **Ecologia microbiana**
- Muitas oportunidades para análise computacional

MAGs são reais?

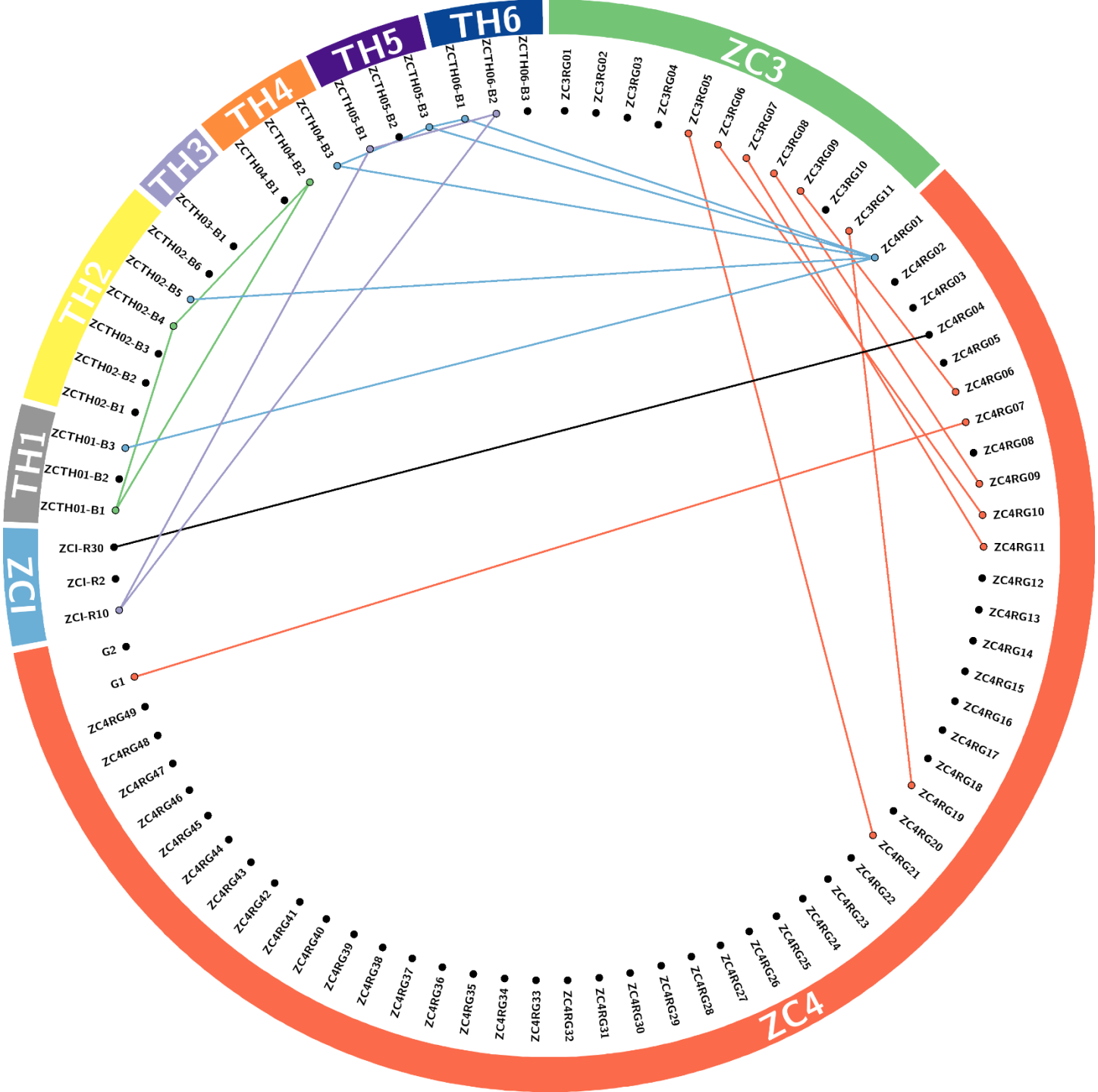
- Precisam passar por um controle de qualidade
  - Completude
  - Contaminação
  - o programa CheckM (Parks et al. 2015) faz essas estimativas
- MAGs em geral são mosaicos de cepas

O espectro de “pureza genômica”



Confirmação adicional

- Um MAG deveria poder ser recuperado de diferentes amostras, que sejam totalmente independentes uma da outra
- Na próxima imagem, os pontos em volta do círculo são MAGs, recuperados de diferentes amostras
- cada amostra é um arco do círculo em sua própria cor
- as linhas ligam MAGs que foram considerados “os mesmos” entre amostras


TH6
TH5
TH4
TH3
TH2
TH1
ZCI
ZC3
ZC4
ZCTH06-B3
ZCTH06-B2
ZCTH06-B1
ZCTH05-B3
ZCTH05-B2
ZCTH05-B1
ZCTH04-B3
ZCTH04-B2
ZCTH04-B1
ZCTH03-B1
ZCTH02-B6
ZCTH02-B5
ZCTH02-B4
ZCTH02-B3
ZCTH02-B2
ZCTH02-B1
ZCTH01-B3
ZCTH01-B2
ZCTH01-B1
ZCI-R30
ZCI-R2
ZCI-R10
G2
G1
ZC3RG01
ZC3RG02
ZC3RG03
ZC3RG04
ZC3RG05
ZC3RG06
ZC3RG07
ZC3RG08
ZC3RG09
ZC3RG10
ZC3RG11
ZC4RG01
ZC4RG02
ZC4RG03
ZC4RG04
ZC4RG05
ZC4RG06
ZC4RG07
ZC4RG08
ZC4RG09
ZC4RG10
ZC4RG11
ZC4RG12
ZC4RG13
ZC4RG14
ZC4RG15
ZC4RG16
ZC4RG17
ZC4RG18
ZC4RG19
ZC4RG20
ZC4RG21
ZC4RG22
ZC4RG23
ZC4RG24
ZC4RG25
ZC4RG26
ZC4RG27
ZC4RG28
ZC4RG29
ZC4RG30
ZC4RG31
ZC4RG32
ZC4RG33
ZC4RG34
ZC4RG35
ZC4RG36
ZC4RG37
ZC4RG38
ZC4RG39
ZC4RG40
ZC4RG41
ZC4RG42
ZC4RG43
ZC4RG44
ZC4RG45
ZC4RG46
ZC4RG47
ZC4RG48
ZC4RG49

Classificação taxonômica de MAGs

Existe um descompasso entre taxonomia (o processo de dar rótulos válidos e universalmente aceitos a organismos) e a genômica

- genômica avança muito mais rápido do que taxonomia
- então hoje em dia existem milhares de MAGs que estão sem rótulo!
- Para lidar com esta situação, um grupo na Austrália criou um banco chamado GTDB
- Junto com o banco, eles disponibilizam uma ferramenta chamada GTDB-tk, que permite classificar um novo MAG de acordo com as informações do banco

RESOURCE

nature biotechnology

GTDB

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.

August 2018

- há discrepâncias entre a taxonomia tradicional do NCBI e a taxonomia proposta pelo GTDB
- veja figura no próximo slide

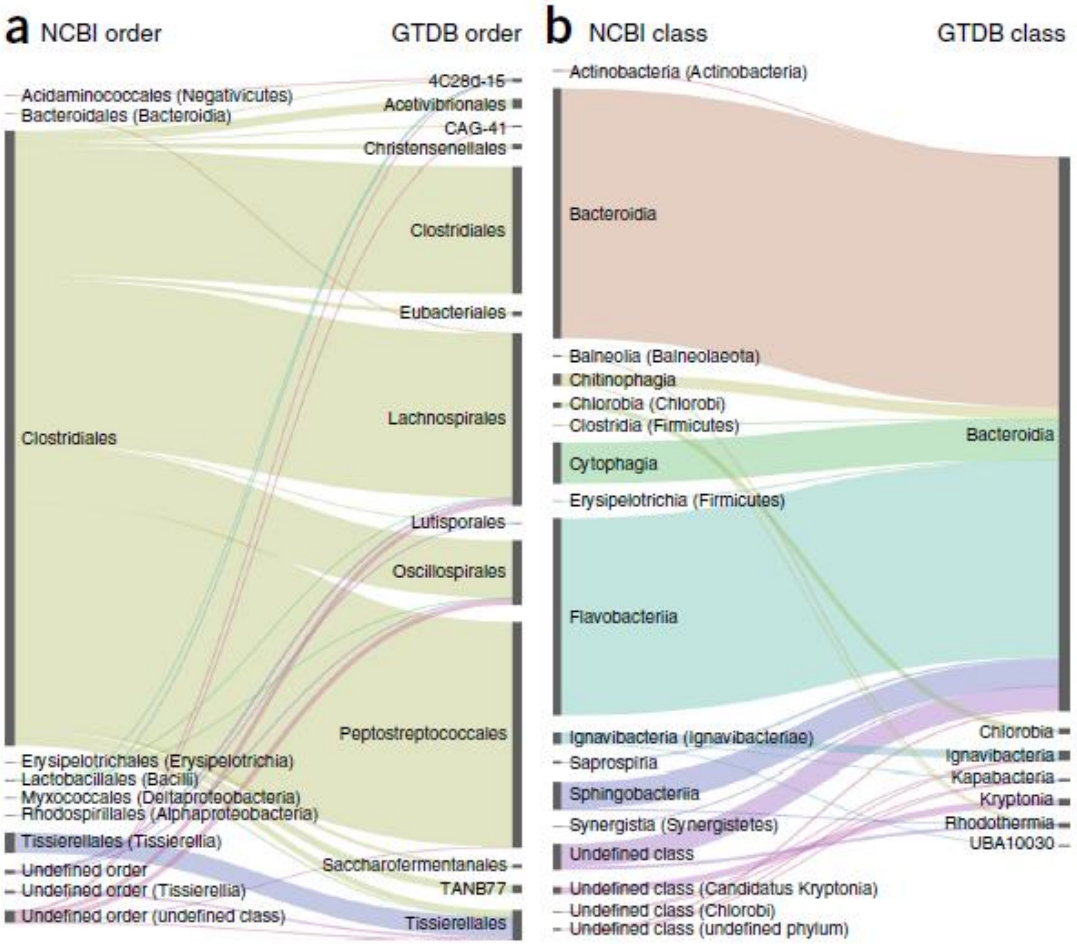


Figure 5 Comparisons of NCBI and GTDB classifications of genomes designated as Clostridia or Bacteroidetes in the GTDB taxonomy. (**a**) Comparison of NCBI (left) and GTDB (right) order-level classifications of the 2,368 bacterial genomes assigned to the class Clostridia in the GTDB taxonomy. Genomes 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.

Anotação funcional

- Pipelines para genomas completos podem ser usados em MAGs
  - IMG/M
  - RAST
  - PGAP
- 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

Sumário de MAGs

- MAGs são “reais”
- Contribuem para lançar luz na “materia escura microbiana”
- Permitem melhor compreensão dos seus ambientes
- Cuidado com a representatividade dos MAGs
  - Complementar com análise de todos os reads/contigs de sua amostra, pois os MAGs representam apenas uma fração dessa massa de dados

Muitos MAGs estão sendo recuperados de amostras humanas

Resource

Cell

Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle

Edoardo Pasolli,[1] Francesco Asnicar,[1,8] Serena Manara,[1,8] Moreno Zolfo,[1,8] Nicolai Karcher,[1] Federica Armanini,[1] Francesco Beghini,[1] Paolo Manghi,[1] Adrian Tett,[1] Paolo Ghensi,[1] Maria Carmen Collado,[2] Benjamin L. Rice,[3] Casey DuLong,[4] Xochitl C. Morgan,[5] Christopher D. Golden,[4] Christopher Quince,[6] Curtis Huttenhower,[4,7] and Nicola Segata[1,9,*]

janeiro 2019

nature biotechnology

RESOURCE

https://doi.org/10.1038/s41587-020-0603-3

Check for updates

OPEN

A unified catalog of 204,938 reference genomes from the human gut microbiome

Alexandre Almeida[1,2], Stephen Nayfach[3,4], Miguel Boland[1], Francesco Strozzi[5], Martin Beracochea[1], Zhou Jason Shi[6,7], Katherine S. Pollard[6,7,8,9,10,11], Ekaterina Sakharova[1], Donovan H. Parks[12], Philip Hugenholtz[12], Nicola Segata[13], Nikos C. Kyrpides[3,4] and Robert D. Finn[1]

julho 2020

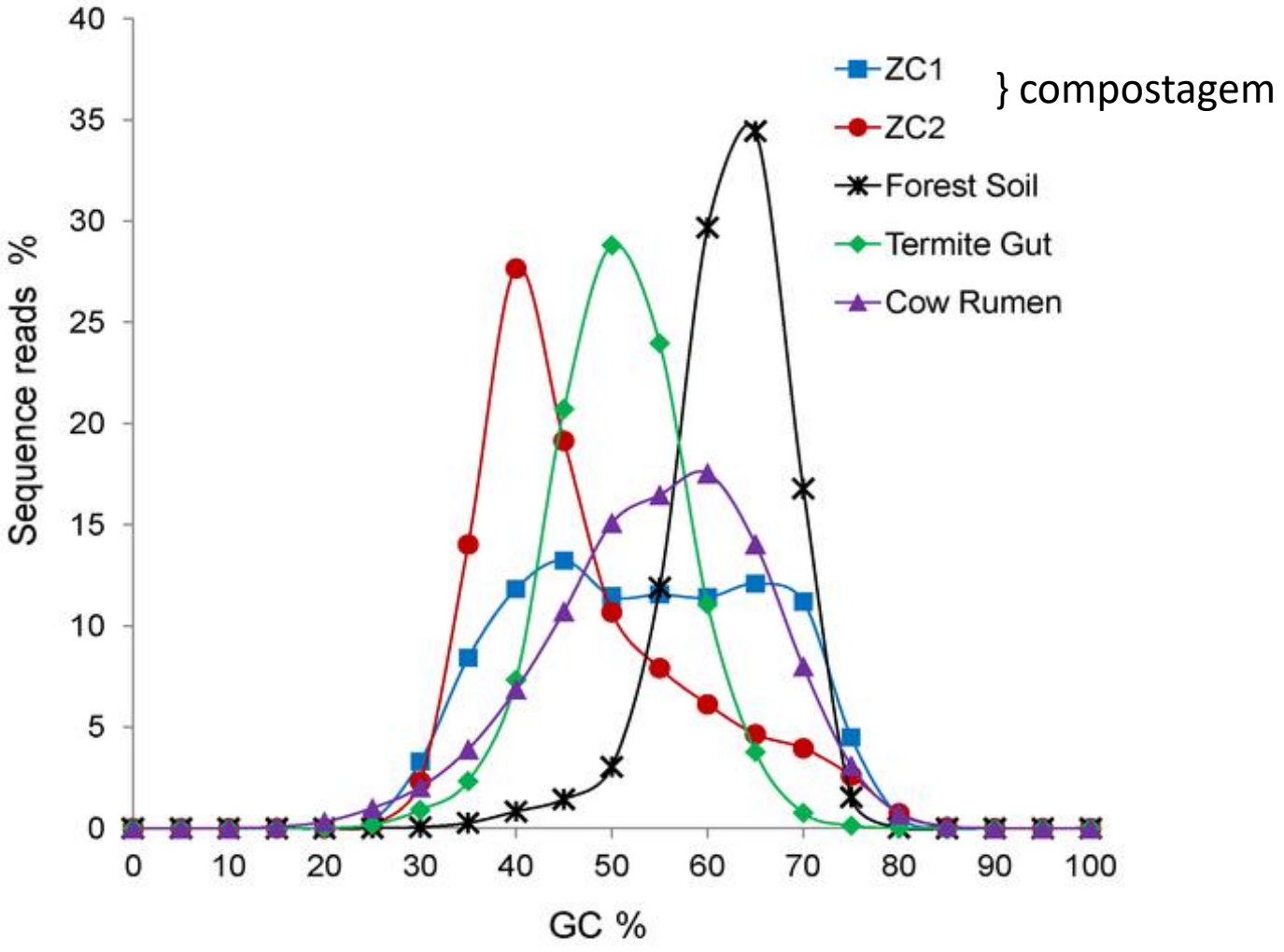
Comparação de metagenomas

- Aqui estamos falando não mais em termos de MAGs, mas em termos de coleção de reads
- Genomicamente
- Taxonomicamente
- Funcionalmente
- Recursos oferecidos pelo IMG/M

Uma comparação muito simples de se fazer é %GC

- verificamos como %GC nos reads varia para um dado metagenoma (coleção de reads que veio de um determinado ambiente – pode ser junção de mais de uma amostra)
- comparamos essa variação entre vários metagenomas
- Veja próxima figura

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

Outras comparações mais complexas são oferecidas pelo mecanismo de Genome clustering do IMG/M

Clustering Type:

By Function:

- (•) COG
- ( ) Pfam
- ( ) KO

By Taxonomy:

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

By Function Category:

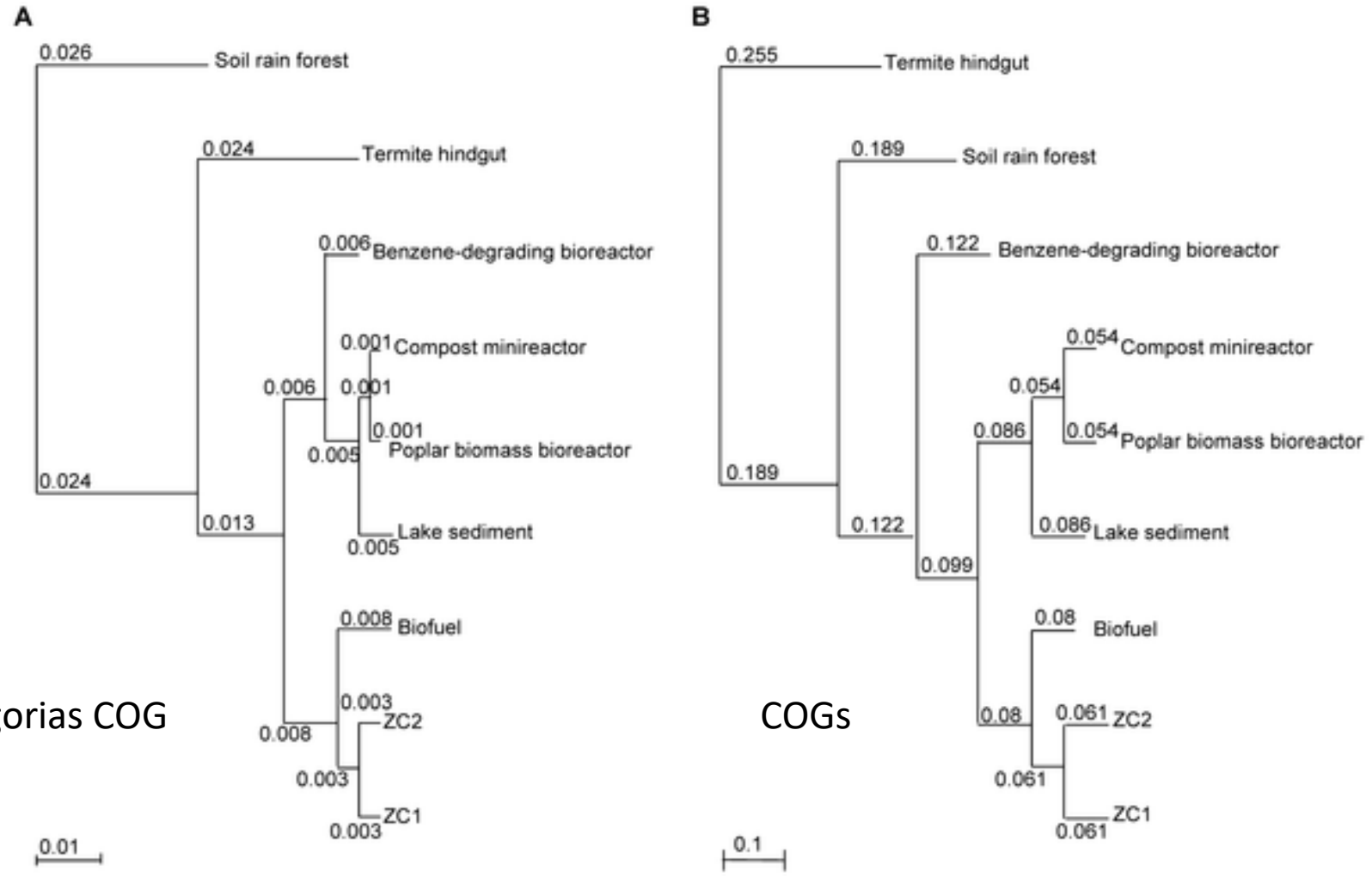
- ( ) COG Categories
- ( ) COG Pathways
- ( ) KEGG Pathway Categories (KO)
- ( ) KEGG Pathway Categories (EC)
- ( ) KEGG Pathways (KO)
- ( ) KEGG Pathways (EC)
- ( ) Pfam Categories

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

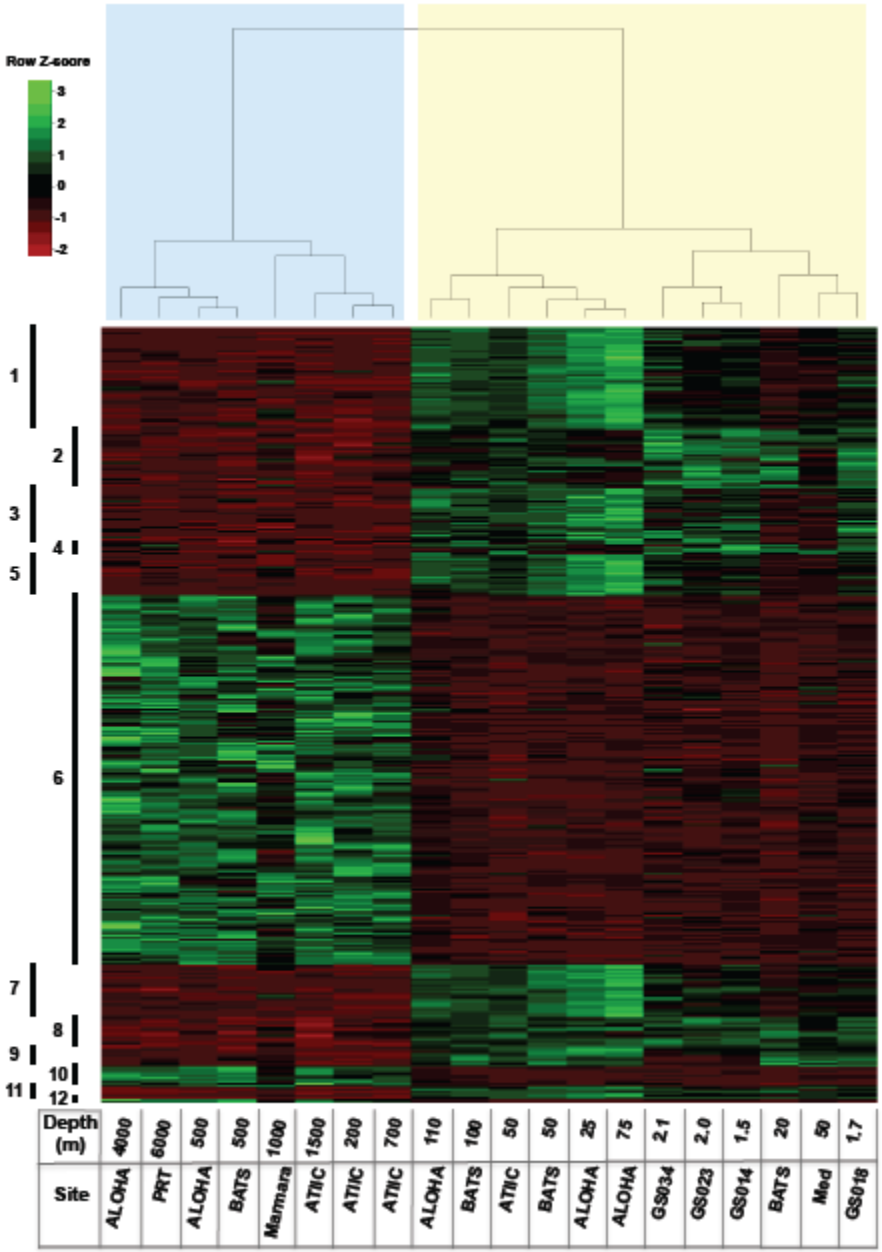
Martins LF, Antunes LP, Pascon RC, de Oliveira JCF, Digiampietri LA, et al. (2013) Metagenomic Analysis of a Tropical Composting Operation at the São Paulo Zoo Park Reveals Diversity of Biomass Degradation Functions and Organisms. PLoS ONE 8(4): e61928. doi:10.1371/journal.pone.0061928
http://127.0.0.1:8081/plosone/article?id=info:doi/10.1371/journal.pone.006192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

Abundância de funções

- mapeamento de reads em ORFs anotadas

Abundância relativa espacial

- é necessário o conceito de família gênica
- COG: Clusters of Orthologous Groups
- É um jeito de agrupar genes em famílias
- Se temos os genes dos metagenomas classificados por COGs, podemos computar
  - representação diferencial dos COGs
- Semelhante a genes diferencialmente expressos
- Ou seja, há COGs que estão mais representados (mais abundantes) em certas amostras comparadas com outras?
- Podemos representar o resultado por heat maps com clusterização hierárquica



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

COGs

Iquique not included

Exercício

- Na imagem anterior, as linhas representam COGs e as colunas representam amostras
- a cor vermelha indica sub-representação e a cor verde indica super-representação; preto (ou escuro) significa estar próximo da media (nem sub-, nem super-)
- Quais resultados este diagrama nos mostra?

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.

MICROBIAL GENOMICS

REVIEW

Pérez-Cobas *et al.*, *Microbial Genomics*

DOI 10.1099/mgen.0.000409

Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses

Ana Elena Pérez-Cobas, Laura Gomez-Valero and Carmen Buchrieser*

Abstract

Metagenomics and marker gene approaches, coupled with high-throughput sequencing technologies, have revolutionized the field of microbial ecology. Metagenomics is a culture-independent method that allows the identification and characterization of organisms from all kinds of samples. Whole-genome shotgun sequencing analyses the total DNA of a chosen sample to determine the presence of micro-organisms from all domains of life and their genomic content. Importantly, the whole-genome shotgun sequencing approach reveals the genomic diversity present, but can also give insights into the functional potential of the micro-organisms identified. The marker gene approach is based on the sequencing of a specific gene region. It allows one to describe the microbial composition based on the taxonomic groups present in the sample. It is frequently used to analyse the biodiversity of microbial ecosystems. Despite its importance, the analysis of metagenomic sequencing and marker gene data is quite a challenge. Here we review the primary workflows and software used for both approaches and discuss the current challenges in the field.

julho de 2020