Universidade de São Paulo
Instituto de Química

USP

Análise de Microbiomas
aula 3

João Carlos Setubal

2021

Análise de dados 16S - etapas

- pré-processamento dos reads
- separação dos reads “em caixinhas”
  - OTUs (ver aula anterior)
  - ASV – amplicon sequence variant (ver adiante)
- classificação taxonômica
- análise de rarefação
- análise de diversidade alfa e comparação entre os índices de cada amostra
- comparação entre amostras
  - diversidade beta
- Análise de redes (networks)

Amplicon Sequence Variant

- Uma ASV é uma sequência única entre todas que estão representadas nos reads obtidos
- Esta unicidade se manifesta
  - na composição de bases
  - no tamanho das sequências
- Em termos de composição, isto quer dizer que duas ASVs de mesmo tamanho necessariamente tem que ter pelo menos uma base de diferença entre si (e basta 1 base diferente para termos 2 ASVs distintas)
- Mas 2 ASVs podem ser distintas também porque tem tamanhos diferentes, mesmo que uma delas seja subcadeia da outra
- é um conceito mais preciso do que OTU
- mas nada impede que agrupemos ASVs em OTUs

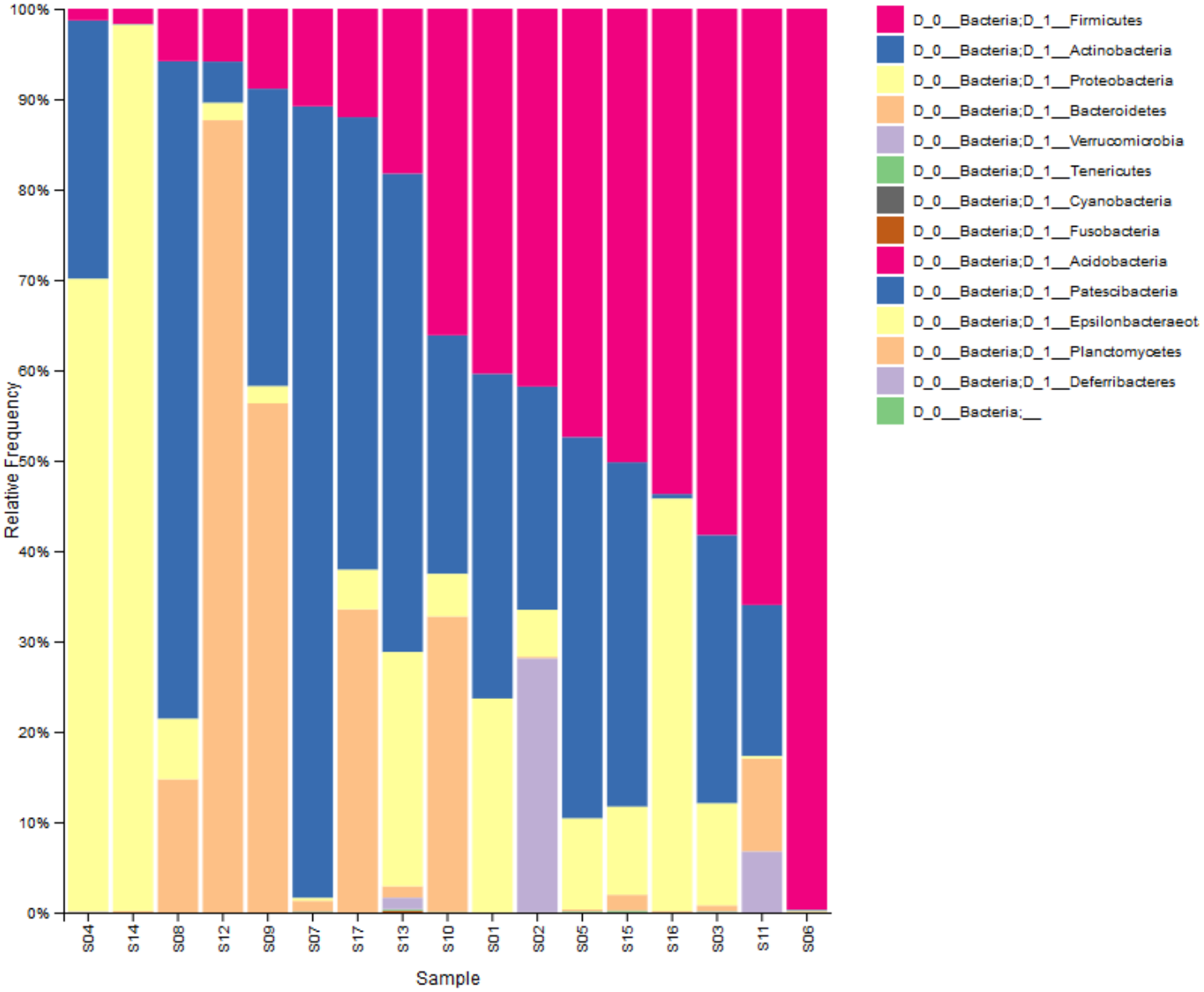
Programas para determinar ASVs

- DADA2
  - B.J. Callahan et al. Dada2: high-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7):581, 2016. doi:10.1038/nmeth.3869.
- Deblur
  - A. Amir et al. Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems*, 2(2):e00191–16, 2017.
- Esses programas são complexos; determinação de ASVs não é um processo simples!
- Em geral, DADA2 resulta em bem mais ASVs do que Deblur

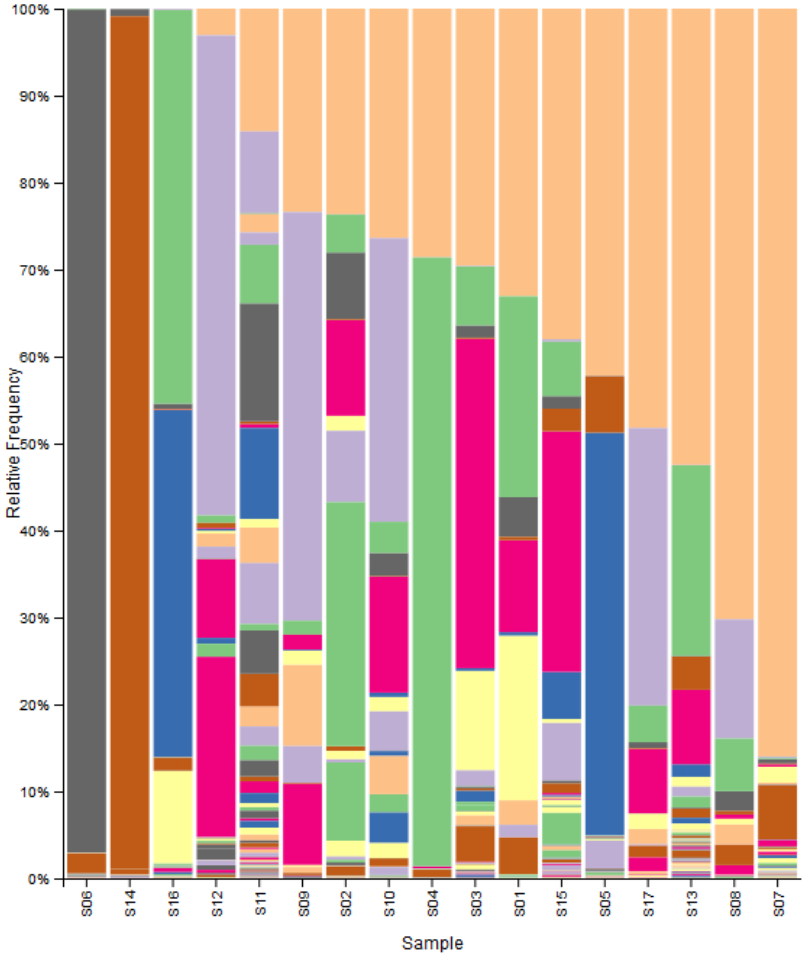
Classificação taxonômica

- Já visto em aulas anteriores
- É importante lembrar que fragmentos de 16S (ou mesmo o gene 16S completo) em geral não nos permite chegar ao nível de espécie
  - pois é comum a situação em que diferentes espécies tem genes 16S que são idênticos!
- Este fato é ainda mais pronunciado quando o dado em análise é apenas um pedaço do gene 16S (por exemplo, a região V3/V4)

Exemplo de visualização da classificação taxonômica de ASVs de amostras (colunas) ao nível de Filo, com suas abundâncias relativas (eixo Y)



Exemplo de visualização da classificação taxonômica de ASVs de amostras ao nível de gênero e suas abundâncias relativas



Análise de diversidade alfa

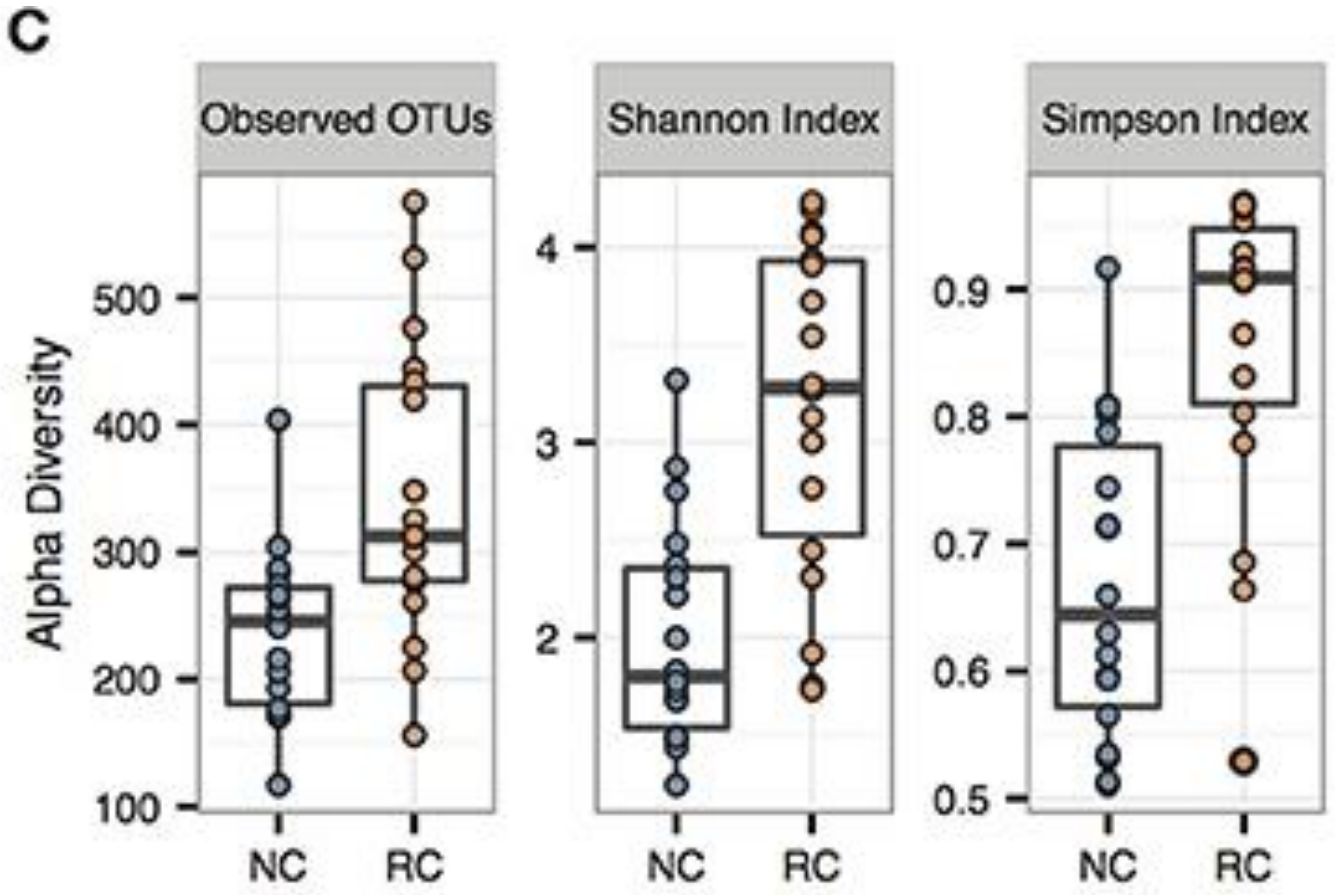
- É um conceito de ecologia
- para uma determinada amostra, queremos quantificar
  - número de diferentes espécies (caracterizadas por OTUs ou ASVs) – isto é chamado de riqueza (richness) ou *observed OTUs* ou *observed ASVs*
  - distribuição das diferentes espécies em termos de suas abundâncias
    - a distribuição é mais ou menos uniforme, ou poucas espécies são muito abundantes e muitas espécies são pouco abundantes?
    - em inglês: evenness
  - distribuição filogenética das espécies

Índices para a quantificação da diversidade alfa

- Riqueza (OTUs ou ASVs observadas)
  - é o mais simples
- Índice (ou entropia) de Shannon
  - Fórmula: $H' = - \sum p_i \ln p_i$
    - onde i é uma espécie (OTU ou ASV) distinta, e p_i é a frequência observada dessa espécie na amostra (abundância relativa)
    - Este índice procura quantificar o "grau de surpresa" das espécies da amostra. Numa amostra com apenas uma espécie, esse índice resultaria zero. Numa amostra com uma espécie dominante, e muitas outras pouco abundantes, o índice resultaria próximo de zero
- Índice de Simpson
  - Fórmula: $\lambda = \sum p_i^2$
  - modela a probabilidade de se obter a mesma espécie na amostra olhando para 2 pontos ao acaso
  - Às vezes aparece como inverse Simpson index ($1/\lambda$) ou Gini–Simpson index ($1 - \lambda$)

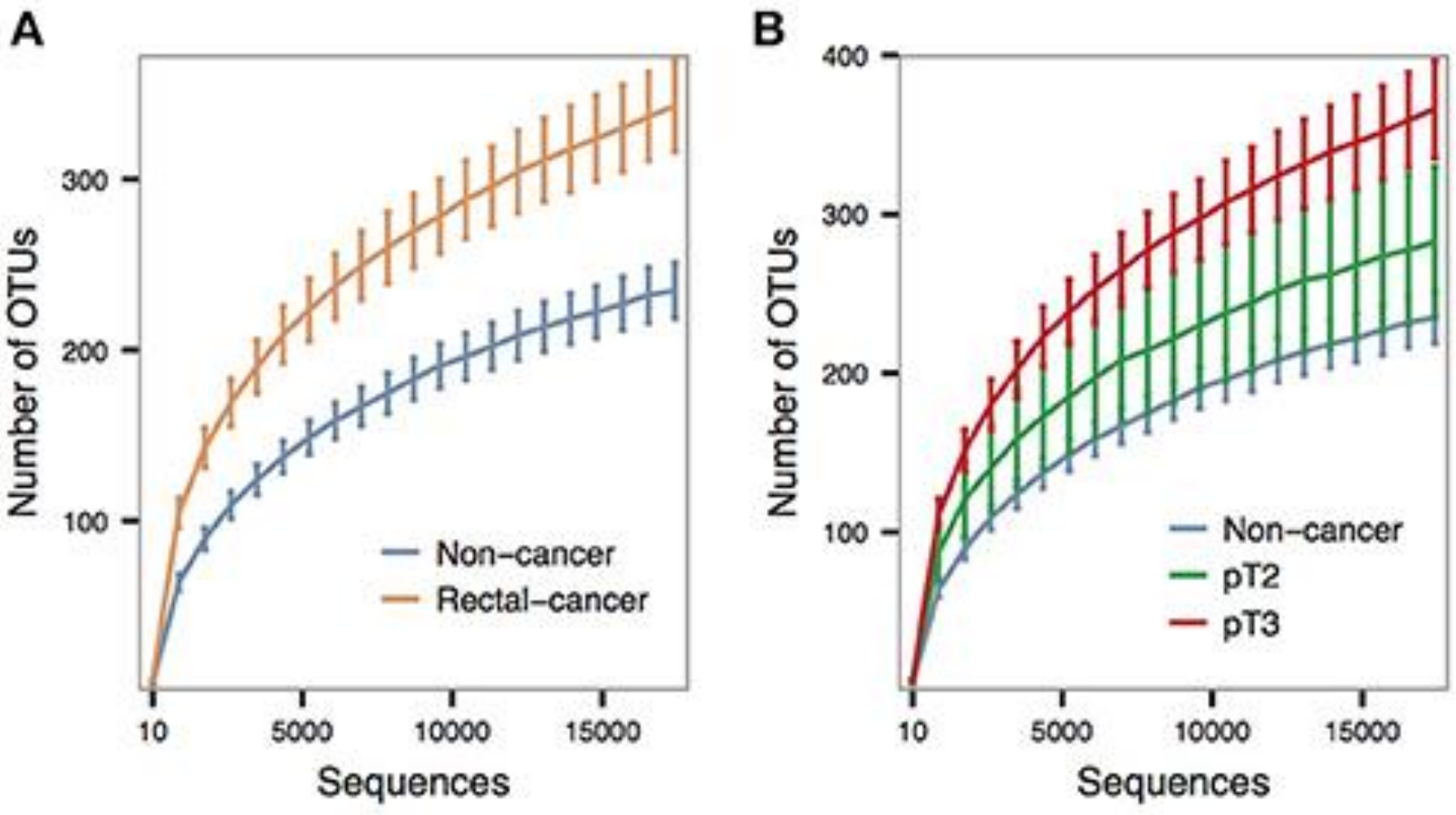
Exemplo

- Nos próximos slides, usarei figuras do seguinte artigo
- A.M. Thomas et al. Tissue-associated bacterial alterations in rectal carcinoma patients revealed by 16S rRNA community profiling. *Frontiers in Cellular and Infection Microbiology*, 6:179, 2016
- Foi um estudo dos microbiomas de pacientes com cancer retal por meio do gene 16S efetuado no AC Camargo Cancer Center (em São Paulo, também conhecido com Hospital AC Camargo)



Comparação das microbiotas de pacientes sem cancer (NC) e pacientes com cancer retal (RC) em termos de diversidade alfa, com 3 métricas diferentes. Em todos os casos a diferença observada é estatisticamente significativa ($p < 0.002$). Análise com base em OTUs.

Curvas de rarefação



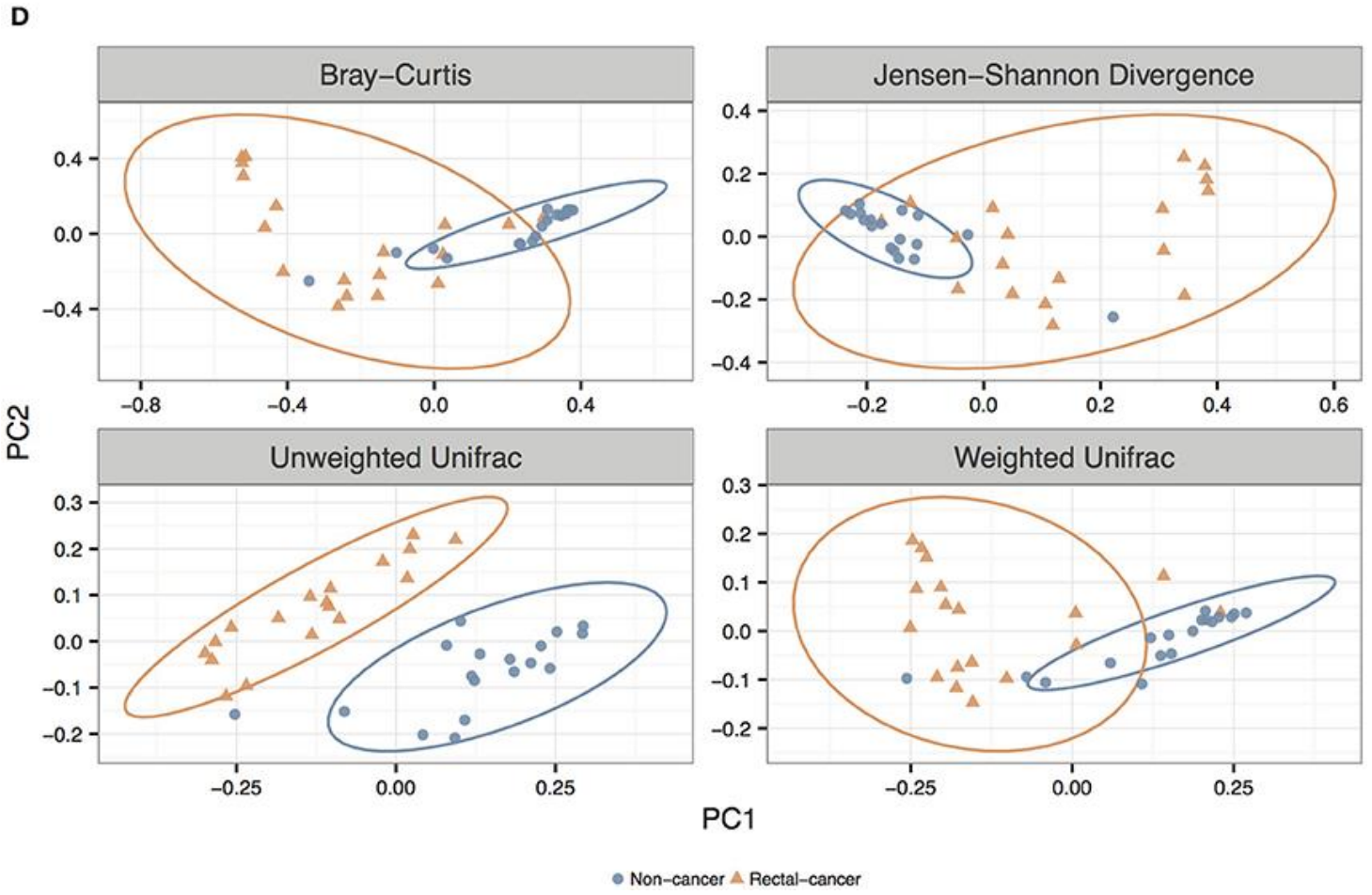
Esta análise mostra que as diferenças vistas em diversidade alfa existem mesmo quando se particionam as sequências para simular um processo de amostragem

Análise de diversidade beta

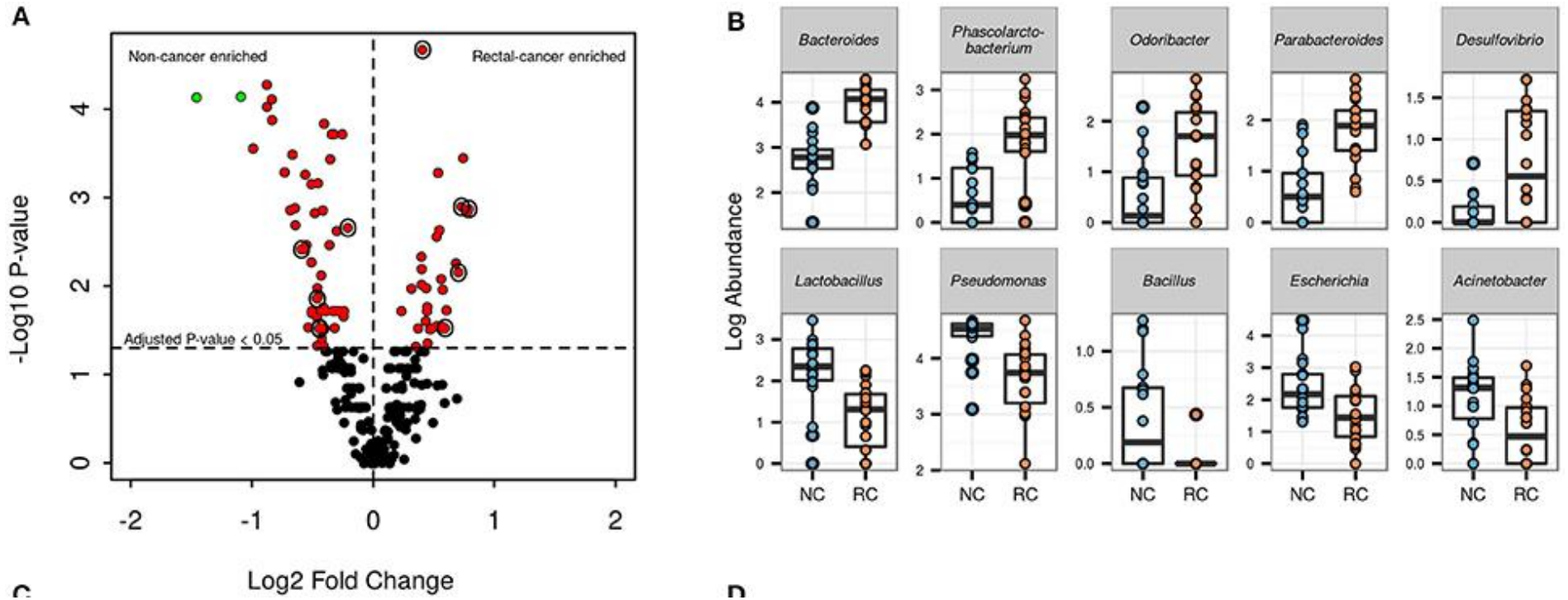
- Neste caso queremos comparar a dissimilaridade (ou distância) das amostras entre si (ao invés de calcular um índice específico de cada amostra, como no caso de diversidade alfa)
- Há várias métricas para calcular dissimilaridade; as mais comuns são
  - Bray-Curtis: compara a composição em OTUs ou ASVs das amostras
  - Unifrac: similar a Bray-Curtis, mas leva em conta a distância filogenética entre as OTUs ou ASVs
    - não-ponderado (leva em conta apenas presença/ausência de OTUs ou ASVs)
    - ponderado (leva em conta os taxons e sua abundância)

Visualização de diversidade beta

- Dados de 16S podem ser entendidos como pontos num espaço hiperdimensional (o número de dimensões está relacionado com a riqueza das amostras)
- Uma vez que as amostras são vistas como pontos num espaço, podemos calcular a distância entre elas
- As possíveis métricas de distância foram mencionadas no slide anterior
- Para visualização, somos obrigados a reduzir o número de dimensões
- Isto pode ser feito por técnicas de redução de dimensionalidade
  - análise de componentes principais (PCA)
  - análise de coordenadas principais (PCoA)
  - Non-metric multidimensional scaling (NMDS)



Aqui estamos vendo a diversidade beta das mesmas amostras vistas anteriormente. Claramente existe uma diferença da microbiota de pacientes sem cancer e da microbiota de pacientes com cancer, visto por diversas métricas. A técnica de redução de dimensionalidade neste caso foi PCA



Esta análise procura identificar quem são os taxons responsáveis pelas diferenças vistas anteriormente

À esquerda temos o assim-chamado gráfico de vulcão (volcano plot). Ele mostra uma comparação das OTUs (cada ponto) presentes em amostras de pacientes com cancer e pacientes sem cancer. Se uma OTU é bem mais abundante em pacientes com cancer, vai aparecer como um ponto à direita do centro. O eixo X dá a significância estatística dessa diferença. No painel B estão mostradas as OTUs (em termos de gênero) que mais se destacam no gráfico de vulcão (tanto para um lado quanto para outro)

O pacote qiime2

- é um ambiente que facilita a execução dos diversos programas para análise de dados de 16S

Nature Biotechnology, 2019 Rob Knight et al.

Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2

Há extensa documentação para qiime2

- https://docs.qiime2.org/2020.6/
- https://docs.qiime2.org/2020.6/tutorials/pd-mice/
  - passo a passo de análise de dados de microbioma fecal de camundongos e seu papel no desenvolvimento da doença de parkinson

Testes estatísticos

- Análise de dados de 16S depende muito de testes estatísticos
- Em geral queremos verificar se o perfil do microbioma de grupo de amostras X é diferente do grupo de amostras Y, com diferentes métricas
  - precisamos dos testes para determinar se as diferenças observadas são estatisticamente signficativas
- Diferentes análises requerem diferentes tipos de teste

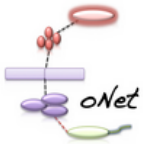
Testes disponíveis no qiime2

- Diversidade alfa
  - Kruskal-Wallis
- Diversidade beta
  - Permanova
  - Anosim

Redes de co-ocorrência

- O objetivo aqui é verificar se gêneros, OTUs ou ASVs co-ocorrem ou tem correlação negativa em diferentes amostras
- Isto pode nos dar pistas sobre as interações entre os microrganismos presentes nas amostras
- É um tema complexo
- Existem diferentes programas que usam diferentes técnicas para montar tais redes
- Exemplo: coNet

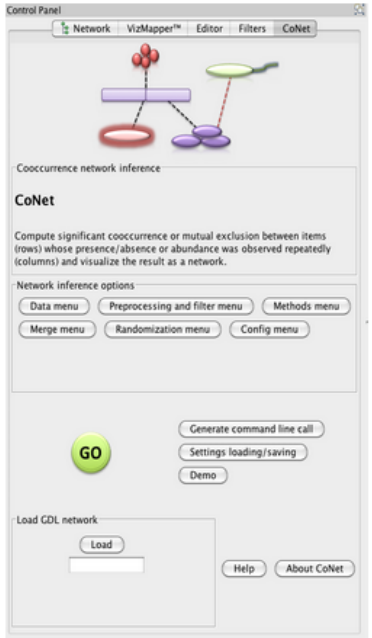
http://psbweb05.psb.ugent.be/conet/



CoNet - Co-occurrence Network inference

CoNet - Home

CoNet is a tool that detects significant non-random patterns of co-occurrence (copresence and mutual exclusion) in incidence and abundance data. It has been designed with (microbial) ecological data in mind, but can be applied in general to infer relationships between objects observed in different samples (for example between genes present or absent across organisms). CoNet runs on command line and as a Cytoscape plugin.



Features

- new: can parse biom files
- support for lagged similarity computation in time series
- automatic assignment of higher-level taxa from lineages
- large choice of correlation, distance and similarity measures
- measures can be combined in multiple ways
- implements the ReBoot procedure (published in PLoS Comp Bio), which is also available through ccrepe
- significance can be tested with various randomization routines and multiple testing corrections
- supports row groups and combination of 2 input matrices

Articles featuring CoNet

Para saber mais sobre análise de 16S

- no próximo slide segue 1a página de review recentemente publicado que faz um apanhado geral de técnicas de análise de microbiomas, incluindo 16S
- No outro slide, reproduzo figura desse artigo com a sequência de etapas de análise de dados de 16S
  - em cada etapa, estão indicados os programas recomendados ou existentes para aquela etapa

julho de 2020

MICROBIAL GENOMICS

REVIEW

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

DOI 10.1099/mgen.0.000409

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

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.

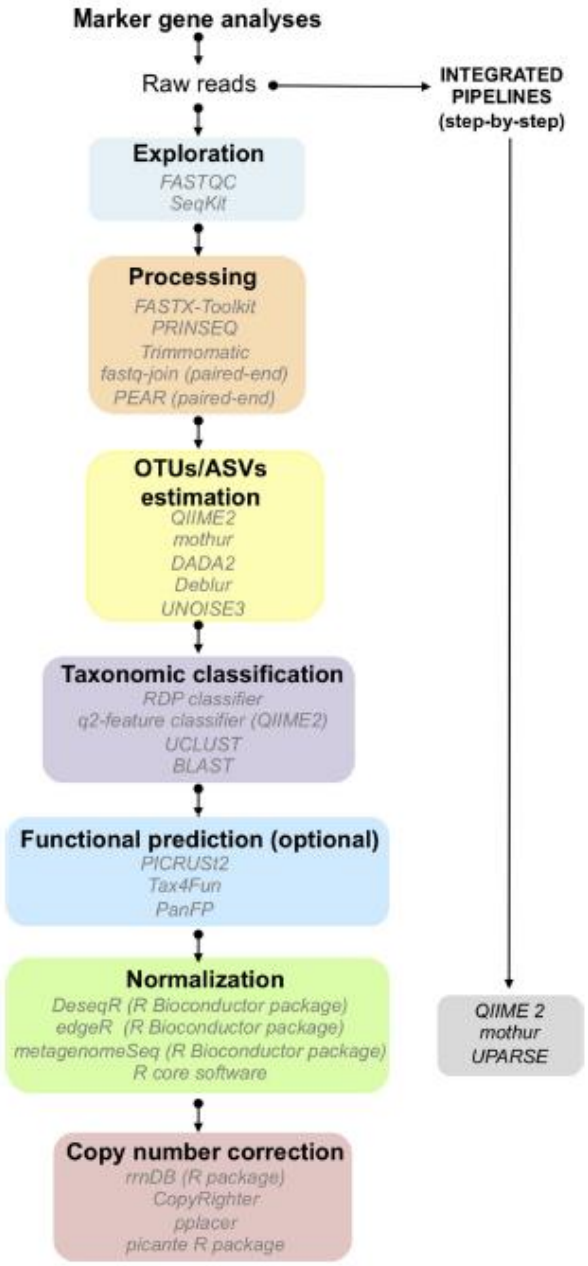


Fig. 2. Schematic representation of the main steps necessary for the analysis of marker gene-derived data. The software related to each step is shown in italics.

Esta figura, retirada do paper do slide anterior, mostra uma possível sequência de etapas na análise de dados de 16S, indicando em cada etapa os principais programas que existem para aquela etapa

Sugiro também consultar o paper mencionado no próximo slide

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing

Andrei Prodan, Valentina Tremaroli, Harald Brolin, Aeilko H. Zwinderman, Max Nieuwdorp, Evgeni Levin

Published: January 16, 2020 • https://doi.org/10.1371/journal.pone.0227434

Abstract

Microbial amplicon sequencing studies are an important tool in biological and biomedical research. Widespread 16S rRNA gene microbial surveys have shed light on the structure of many ecosystems inhabited by bacteria, including the human body. However, specialized software and algorithms are needed to convert raw sequencing data into biologically meaningful information (i.e. tables of bacterial counts). While different bioinformatic pipelines are available in a rapidly changing and improving field, users are often unaware of limitations and biases associated with individual pipelines and there is a lack of agreement regarding best practices. Here, we compared six bioinformatic pipelines for the analysis of amplicon sequence

DADA2 offered the best sensitivity, at the expense of decreased specificity compared to USEARCH-UNOISE3 and Qiime2-Deblur. USEARCH-UNOISE3 showed the best balance between resolution and specificity. OTU-level USEARCH-UPARSE and MOTHUR performed well, but with lower specificity than ASV-level pipelines. QIIME-uclust produced large number of spurious OTUs as well as inflated alpha-diversity measures and should be avoided in future studies.