

Doutoranda pelo Programa Interunidades de Pós-Graduação em Bioinformática da USP

Orientador: Prof. Dr. João Carlos Setubal Monitor: Robson Pontes

Organização da aula de hoje

Apresentação

Explicações

Links

Prática

Aspectos técnicos

Materiais

Visualização e Interpretação



Tutorial "Moving Pictures" (microbioma humano)



https://docs.qiime2.org

Version: 2022.8 -

Table of Contents

- Getting started
- What is QIIME 2?
- Core concepts
- Installing QIIME 2
- Jupyter Book Tutorials
- Tutorials
- Interfaces
- Plugins
- Semantic types
- Community
- Data resources
- Supplementary resources
- User Glossary
- Citing QIIME 2

Quick search



QIIME 2 user documentation

This site is the official user documentation for QIIME[™] 2, including installation instructions, tutorials, and other important information. Visit http://qiime.org for information on QIIME[™] 1.

Getting started

Check out the getting started guide to begin using QIIME 2.

Table of contents

- Getting started
- What is QIIME 2?
- Core concepts
 - Data files: QIIME 2 artifacts
 - Data files: visualizations
 - Semantic types
 - Plugins
 - Methods and visualizers
 - Next steps
- Installing QIIME 2
 - Nativelv installing QIIME 2

Parte Prática: https://bit.ly/16scolab

https://bit.ly/16scolab

III Faça uma cópia do Python Notebook para o seu Google Drive e abra-o para executar o código.

≣

Q

 $\{x\}$

!!! Cuidado para não fechar a guia ou a janela.

Deixe executando a instalação do QIIME2 e dependências

Arquivo Editar Ver Inserir Ambiente de ex	cecução l	Ferramentas
Localizar no Drive Abrir no modo Plavaround		
Novo notebook		ciamo
Abrir notebook	Ctrl+O	clame
Fazer upload de notebook		seado no l
Rename		
Mover		
Mover para a lixeira		ento no se
Salvar uma cópia no Drive		
Salvar uma cópia como Giet do GitHub		utorial fut
Salvar uma cópia no GitHub		. O Google
Salvar	Ctrl+S	los desde
Salvar e fixar revisão	Ctrl+M S	

Microbioma

Coleção de microrganismos e seus componentes biológicos em um ambiente





Microrganismos + Genes + Genomas + Proteínas + Metabólitos



LibreTexts Biology. Book Microbiology (Boundless).

Por que estudar um microbioma?

Participação de microrganismos no ciclo biogeoquímico.



Por que estudar um microbioma?

Fecal Microbiota Transplant (FMT)

Clostridium difficile Toxinas \rightarrow intestino

BORODY, Thomas J.; KHORUTS, Alexander. Fecal microbiota transplantation and emerging applications. **Nature reviews Gastroenterology & hepatology**, v. 9, n. 2, p. 88-96, 2012.

Existem microrganismos não cultiváveis na bancada (laboratório) (~99%)



Como podemos estudar um microbioma?

Sequenciamento de DNA / RNA



https://earthmicrobiome.org/

"... a massively collaborative effort to characterize microbial life on this planet."

"We set out to analyze 200,000 samples from these communities..."

NIH Human Microbiome Project



Characterization of the microbiomes of healthy human subjects at five major body sites, using 16S and metagenomic shotgun sequencing.



Characterization of microbiome and human host from three cohorts of microbiome-associated conditions, using multiple 'omics technologies.

Sequenciamento de DNA / RNA

https://www.hmpdacc.org/

~US\$170 milhões

STUDIES

PRIMARY SITE



SAMPLES 31,596

FILES 161,265



BIKEL, Shirley et al. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. Computational and structural biotechnology journal, v. 13, p. 390-401, 2015.



BIKEL, Shirley et al. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. Computational and structural biotechnology journal, v. 13, p. 390-401, 2015.

ATGAAAGTTCAACCCAATGCGCTCTGGGAGCAGTGTCTTCAACTTATCAGAGATAATGTAACCGAGCAGC AGTTTACAACATGGTTCAAGCCAATTACCTTCGAAGCGTTCGATGCGGCTACCAATATCCTATTGGTGCA GGTGCCCAGTCCGTTCGTTTACGAGTACTTGGAGGAAAACTATGTTGATTTGAGCTAAAGTGCTTACC CGTATATATGGCAAAGGCGTTCAGCTGAAGTACAGAATTGTTACCGACAAGGCGCACAATCTTACGCAGG ACCATTAGACGTTGCTCTTCAAGAAATAGACCCACAGTTAGACTTGCACAAATCGTTTTCAAACTATATT GAAGGCGATAGCAACAAGCTGCCACGCTCTATTGGCTTGTCAATAGCAGAACACCCGAACACCACGCAGT TCAATCCGATGTTCATTTATGGTCCATCGGGTTGTGGAAAAACCCCACCTTATCAATGCGATAGGTCTGCG CATAAAGCAGCTTTACCCACAAAAGCGTGTGTTGTATATATCTGCTCGTCTGTTCCAAGTACAGTACACA AATTCTGTTTTGAATAATACGACCAACGATTTCATCAACCTTCTATCAAACAATAGATGTTTTAATCGTCG

Quais microrganismos estão presentes? (Taxonomia)

Gene ribossomal da subunidade 16S



16S: Bactérias e arqueias 18S ou ITS: Fungos

Maior cobertura e menor custo (comparado ao shotgun)

Limitações:

- Identificação taxonômica menos precisa (Confiabilidade até nível de gênero!)
- Não é possível realizar análise funcional de genes (embora há programas que fazem inferências a partir da taxonomia)

YANG, Bo; WANG, Yong; QIAN, Pei-Yuan. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. **BMC bioinformatics**, v. 17, n. 1, p. 1-8, 2016.

Gene ribossomal da subunidade 16S



JOHNSON, Jethro S. et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. **Nature communications**, v. 10, n. 1, p. 1-11, 2019.



• Quantitative insights into Microbial Ecology

"...is a powerful, extensible, and decentralized microbiome analysis package with a focus on data and analysis transparency."

- Pacote de ferramentas para análise de microbioma
- Criado em ~2010 durante o Projeto de Microbioma Humano (2007-2016)
- Liderado por Greg Caporaso and Rob Knight.

Tutorial "Moving Pictures" (microbioma humano)



https://docs.qiime2.org

Version: 2022.8 -

Table of Contents

- Getting started
- What is QIIME 2?
- Core concepts
- Installing QIIME 2
- Jupyter Book Tutorials
- Tutorials
- Interfaces
- Plugins
- Semantic types
- Community
- Data resources
- Supplementary resources
- User Glossary
- Citing QIIME 2

Quick search



QIIME 2 user documentation

This site is the official user documentation for QIIME[™] 2, including installation instructions, tutorials, and other important information. Visit http://qiime.org for information on QIIME[™] 1.

Getting started

Check out the getting started guide to begin using QIIME 2.

Table of contents

- Getting started
- What is QIIME 2?
- Core concepts
 - Data files: QIIME 2 artifacts
 - Data files: visualizations
 - Semantic types
 - Plugins
 - Methods and visualizers
 - Next steps
- Installing QIIME 2
 - Natively installing QIIME 2

Comandos: <u>https://bit.ly/16scolab</u>



Documentação e tutorial: https://docs.giime2.org/2022.8/tutorials/moving-pictures/

Análise de rRNA 16S

- 1. Obter dados (sequências e metadados)
- 2. Importar dados para QIIME2
- 3. Filtragem e agrupamento dos reads em ASVs
- 4. Construção da tabela de ASVs (contagem de sequências por ASVs em cada amostra)
- 5. Análise da diversidade:
 - a. Alfa-diversidade
 - b. Beta-diversidade
- 6. Classificação taxonômica

ASVs (Amplicon Sequence Variants): sequências únicas de DNA que representam um organismo.

Antes de iniciar o processamento dos dados

1. Instalação do QIIME2 e dependências

- 2. Criação de diretórios:
 - a. dados_mp
 - b. dados_mp/seqs

Sobre os dados



Amostras (versão simplificada do artigo original)

- Dados sequenciados na plataforma Illumina HiSeq
- Protocolo de sequenciamento do Earth Microbiome Project para a região hipervariável 4 (V4) do gene de rRNA 16S

(No artigo original, foram 396 amostras coletadas em tempos diferentes) CAPORASO, J. Gregory et al. Moving pictures of the human microbiome. **Genome biology**, v. 12, n. 5, p. 1-8, 2011.

3) Obter os dados: Metadados

3.1 - Baixe a tabela de metadados e dê uma olhada.

!wget -O "sample-metadata.tsv"
"https://data.qiime2.org/2022.8/tutorials/moving-pi
ctures/sample_metadata.tsv"

!cat sample-metadata.tsv

wget = obter arquivo

cat = visualizar arquivo

Há quantas amostras?

Quais as principais variáveis que podemos analisar?

e2

Sobre os metadados

Há quantas amostras?

Quais as principais variáveis que podemos analisar?

- Dois indivíduos
 - Quatro regiões do corpo
 - Cinco amostras temporais (Day $0 \rightarrow$ uso de antibiótico)
- Total: 34 amostras (faltam algumas amostras)
- A comunidade microbiana difere entre dois indivíduos?
- A comunidade microbiana difere ao longo do tempo após o uso de antibiótico?



Obter os dados: Sequências e Barcode

3.2 - Baixe as sequências de 16S rRNA para o diretório "seqs"

!wget -0 "seqs/sequences.fastq.gz" \
"https://data.qiime2.org/2022.8/tutorials/moving-pictures/emp-sing
le-end-sequences/sequences.fastq.gz"

3.3 - Baixe os barcodes (sequências identificadoras das amostras)

!wget -0 "seqs/barcodes.fastq.gz" \
"https://data.qiime2.org/2022.8/tutorials/moving-pictures/emp-sing
le-end-sequences/barcodes.fastq.gz"

Importar arquivos para QIIME2

!qiime tools import \

--type EMPSingleEndSequences \

--input-path seqs \

--output-path seqs.qza

EMPSingleEndSequences = Sequências multiplexadas e únicas (não pareadas)

Para importar outros formatos: https://docs.qiime2.org/2022.8/tutorials/importing/

4) Demultiplexar reads

Demultiplexar sequências

```
!qiime demux emp-single \
    --i-seqs seqs.qza \
    --m-barcodes-file sample-metadata.tsv \
    --m-barcodes-column barcode-sequence \
    --o-per-sample-sequences demux.qza \
    --o-error-correction-details demux-details.qza
```

"Demultiplex" \rightarrow separar sequências por amostras de acordo com a sequência identificadora (barcode)

4) Demultiplexar reads

Visualizar quantidade de sequências por amostra

```
!qiime demux summarize \
```

```
--i-data demux.qza \
```

```
--o-visualization qualities.qzv
```

Arquivos que terminam em .qzv podem ser visualizados na plataforma online do QIIME2: <u>https://view.qiime2.org/</u>

"Arrastar" ou fazer upload do arquivo demux.qzv na plataforma online.

O que você observa ao longo dos reads? A partir de qual posição dos reads você "trimaria" (cortaria)?



DADA2 (Divisive Amplicon Denoising Algorithm) Detecta e corrige (quando possível) dados de sequência de amplicon Illumina. Filtra sequências quiméricas.

Esse passo é o mais demorado (então vamos deixar executando durante a explicação)

5.1) Executar DADA2

!qiime dada2 denoise-single \

- --i-demultiplexed-seqs demux.qza \
- --p-trim-left ? \
- --p-trunc-len ? \
- --o-representative-sequences rep-seqs-dada2.qza \

--o-table table-dada2.qza \

--o-denoising-stats stats-dada2.qza

--p-trim-left $\,\text{m} \rightarrow \text{corta}$ as primeiras m bases de cada sequência

--p-trim-length $n \rightarrow corta$ a sequência da posição n até o final

Baseado nos gráficos que você visualizou no demux.qzv, que valores você escolheria para --p-trunc-len e --p-trim-left neste caso?



- Filtrar e 'trimar' os reads
- Encontrar o conjunto mais provável de sequências únicas na amostra (ASVs)
- 3. Remover quimeras
- 4. Contar as abundâncias de cada ASV



Schematic of OTU and DADA2 approaches towards amplicon sequencing errors.

https://igcbioinformatics.github.io/biomeshinycourse/pages/dada2/Biodata.ptCrashCourses.html



ASV	A1	A2	Sequência
ASV1	300	350	ATGGCATGCAAG
ASV2	156	250	AAGGCATCCTAG
ASVn	367	210	ATGCCTTCCAAG

5.2 - Visualizar quantidade de ASVs resultantes de cada passo do DADA2

!qiime metadata tabulate \

--m-input-file stats-dada2.qza \

--o-visualization stats-dada2.qzv

Qual a proporção de reads que foi mantido após o processamento em ASVs? Observe o número final de reads (non-chimeric). O que você observa quando compara esses valores entre as amostras e como isso pode afetar a métrica de diversidade?

5.3 - Visualizar contagem de reads por ASV em cada amostra (histograma)

!qiime feature-table summarize \

--i-table table-dada2.qza \

--o-visualization table.qzv \

--m-sample-metadata-file sample-metadata.tsv

Análise de Diversidade

Métricas de Diversidade

Alpha diversidade: diversidade ecológica dentro de uma amostra

Beta diversidade: diferenças ecológicas (distâncias) entre amostras

Análise de Diversidade: Alfa-diversidade

Quão diversa é uma única amostra?



Riqueza (Richness): Quantos taxa estamos observado? \rightarrow #taxa observados, Índice Simpson, Faith's PD (Phylogenetic Diversity)

Uniformidade (Evenness): Quão homogêneas são as abundâncias entre os taxa? → Índice Evenness

"Mistura": Métricas que combinam ambos riqueza e uniformidade \rightarrow Índice Shannon

Análise de Diversidade: Alfa-diversidade

Testes estatísticos para alfa diversidade

- Alfa diversidade pode fornecer um único valor para cada amostra.
- Pode ser tratado como qualquer outra medida de amostra e pode ser usado em testes univariados clássicos (t-teste, Teste de Mann-Whitney U)

Análise de Diversidade: Beta-diversidade

Qual a diferença entre duas ou mais amostras/indivíduos/locais?



- Unweighted (não ponderado = presença/ausência): Quantos taxa estão compartilhados entre as amostras?
 - Índice Jaccard, unweighted Unifrac
- Weighted (ponderado = considera abundância como peso): Os taxa compartilhados possuem abundâncias similares?
 - Distância Bray-Curtis, weighted Unifrac



As amostras compartilham taxa similar geneticamente?



• Weighted Unifrac escala os ramos pela abundância

Análise de Diversidade: Beta-diversidade

Podemos comparar a dissimilaridade (ou distância) das amostras entre si. Para isso, usamos técnicas de redução de dimensionalidade como:

PCoA - Análise de Coordenadas Principais

Análise de Coordenadas Principais (PCoA)

- Análise multidimensional para visualizar quais as **fontes principais de variabilidade dos dados**.
- Duas (ou três) dimensões (as que mais contribuem para a variabilidade dos dados)
- Amostras são vistas como pontos num espaço e podemos calcular a distância entre elas.

Análise de Coordenadas Principais (PCoA)



https://gibbons-lab.github.io/isb_course_2020/16S/

Análise de Coordenadas Principais (PCoA)

Teste estatístico para beta diversidade



- Mais complicado.
- Geralmente não é normal e é heterogêneo.
- PERMANOVA pode lidar com isso.

6.1 - Calcular distância filogenética entre ASVs

!qiime phylogeny align-to-tree-mafft-fasttree \

- --i-sequences rep-seqs-dada2.qza \
- --o-alignment aligned-rep-seqs.qza \
- --o-masked-alignment masked-aligned-rep-seqs.qza \
- --o-tree unrooted-tree.qza \
- --o-rooted-tree rooted-tree.qza

6.2 - Alpha e Beta diversidade

Executar a análise de diversidade (alfa e beta)

!qiime diversity core-metrics-phylogenetic \

- --i-phylogeny rooted-tree.qza \
- --i-table table-dada2.qza \
- --p-sampling-depth ? \
- --m-metadata-file sample-metadata.tsv \
- --output-dir core-metrics-results

Visualize o arquivo table.qzv e, em particular, a guia "Interactive Sample Detail" nessa visualização. Que valor você escolheria para --p-sampling-depth? Quantas amostras serão excluídas de sua análise com base nessa escolha? Quantas sequências totais você analisará no comando core-metrics-phylogenetic?

--p-sampling-depth M Subamostragem dos reads de cada amostra com a quantidade total de M

6.2 - Alpha e Beta diversidade

Visualize os arquivos .qzv gerados no diretório core-metrics-results na plataforma online do QIIME2 (<u>https://view.qiime2.org/</u>)

Arquivos terminados em ... emperor. qzv são gráficos de PCoA.

Mude as cores para visualizar o agrupamento no PCoA.



Beta-diversidade (PCoA)

Axis 2 (22.71 %) Distância Bray-curtis Região do corpo gut left palm right palm tongue Tempo depois do antibiótico ▼ Day 0 • After two months Axis 1 (32.51 %),7

6.3 - Análises estatísticas

Teste para alfa diversidade entre grupos

!qiime diversity alpha-group-significance \

- --i-alpha-diversity core-metrics-results/faith_pd_vector.qza \
- --m-metadata-file sample-metadata.tsv \
- --o-visualization core-metrics-results/faith-pd-group-significance.qzv

!qiime diversity alpha-group-significance \
 --i-alpha-diversity core-metrics-results/evenness_vector.qza \
 --m-metadata-file sample-metadata.tsv \
 --o-visualization core-metrics-results/evenness-group-significance.qzv

Vemos uma diferença na diversidade por amostra entre subject (indivíduos)? E entre body-site (regiões do corpo)? Quais colunas dos metadados apresentam diferenças na **riqueza** da comunidade microbiana? E na **uniformidade**?

6.3 - Análises estatísticas

Teste PERMANOVA para beta diversidade

```
!qiime diversity beta-group-significance \
```

- --i-distance-matrix core-metrics-results/unweighted_unifrac_distance_matrix.qza \
- --m-metadata-file sample-metadata.tsv \
- --m-metadata-column body-site \
- --o-visualization core-metrics-results/unweighted-unifrac-body-site-significance.qzv \

--p-pairwise

```
!qiime diversity beta-group-significance \
```

- --i-distance-matrix core-metrics-results/unweighted_unifrac_distance_matrix.qza \
- --m-metadata-file sample-metadata.tsv \

```
--m-metadata-column subject \
```

--o-visualization core-metrics-results/unweighted-unifrac-subject-group-significance.qzv \

--p-pairwise

Há diferenças na composição microbiana estatisticamente significantes entre os indivíduos ("subjects")? E em relação aos "body-sites"? Quais pares de "body-sites" são significativamente diferentes um do outro?

6.4 - Classificação Taxonômica

Obter classificador taxonômico

Classificador treinado com Base de Dados Greengenes v.13_8 99% OTUs para V4

!wget -0 "gg-13-8-99-515-806-nb-classifier.qza"
"https://data.qiime2.org/2022.8/common/gg-13-8-99-515-806-nb-class
ifier.qza"

6.4 - Classificação Taxonômica

Atribuir taxonomia às sequências representativas (ASVs) usando o classificador

!qiime feature-classifier classify-sklearn \
 --i-classifier gg-13-8-99-515-806-nb-classifier.qza \
 --i-reads rep-seqs-dada2.qza \
 --o-classification taxonomy.qza

(opcional) Para visualizar as taxonomias correspondentes à cada ASV: **!qiime metadata tabulate **

```
--m-input-file taxonomy.qza \
```

```
--o-visualization taxonomy.qzv
```

6.4 - Classificação Taxonômica

Visualize a abundância relativa

!qiime taxa barplot \

- --i-table table.qza \
- --i-taxonomy taxonomy.qza \
- --m-metadata-file sample-metadata.tsv \
- --o-visualization taxa-bar-plots.qzv

Visualize as amostras no "Level 2" (filo) e ordene as amostras por "body-site".

Quais são os phyla dominantes em cada body-site? Você observa alguma mudança consistente dos dois indivíduos (subject) entre os diferentes tempos?





Abundância Relativa

a nível de filo





Abundância Relativa

a nível de filo

Beta-diversidade (PCoA)

Axis 2 (22.71 %) Distância Bray-curtis Região do corpo gut left palm right palm tongue Tempo depois do antibiótico ▼ Day 0 • After two months Axis 1 (32.51 %)



Abundância Relativa

a nível de classe



Outras análises

- Abundância diferencial
 - ANCOM
 - Gneiss
- Análise de associação
 - mantel
 - bioenv
- Rede de co-ocorrência
 - CoNet
 - Cytoscape / Gephy
- Diagramas de Venn
- Análise de Diversidade no R: phyloseq, vegan, microbiome, etc.

Tutorial e documentação disponível



Instruções para instalação: https://docs.qiime2.org/2022.8/install/

Tutorial: https://docs.qiime2.org/2022.8/tutorials/moving-pictures/

Obrigada!

Contato: suzy.eiko@gmail.com

Formato fastq

1 @HWI-EAS440_0386:1:23:17547:1423#0/1
2 TACGNAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATC
3 +
4 IIIE)EEEEEEEGFIIGIIIHIHHGIIIGIIHHHGIIHGHEGDGIFIGEHGIH

1) Identificador da sequência (Inicia com @)

- 2) Sequência nucleotídica
- 3) Identificador da sequência (opcional) (Inicia com +)
- 4) Qualidade em Phred score caracteres ASCII

Phred quality scores are logarithmically linked to error probabilities

Formato	fastq
---------	-------

 $P = 10^{\frac{-Q}{10}}$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

ASCII BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	Perror	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 \$	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

https://www.drive5.com/usearch/manual/guality_score.html