Manual - Metabarcoding com QIIME 2

Prof: Guilherme Baião - l2026/05

Objetivo

Este manual tem como objetivo guiar uma análise prática de dados de metabarcoding utilizando o QIIME 2. Ao final, o aluno deverá ser capaz de:

- Entender o fluxo básico de análise de dados de sequenciamento de amplicons
- Avaliar qualidade de dados
- Processar dados com DADA2
- Gerar e interpretar perfis taxonômicos

Contexto Biológico

Metabarcoding é uma abordagem que utiliza sequenciamento de fragmentos específicos de DNA (amplicons) para identificar comunidades biológicas em uma amostra (ex: microbioma). Neste exercício, utilizamos dados de sequenciamento da região 16S rRNA para caracterizar comunidades bacterianas.

São utilizadas amostras de solo e de raízes de plantas.

Primers utilizados:
Forward: CCTACGGGNGGCWGCAG
Reverse: GACTACHVGGGTATCTAATCC

Plataformas de software utilizadas

Micromamba
Micromamba é um gerenciador de ambientes computacionais que permite instalar e organizar programas e suas dependências de forma reprodutível, evitando conflitos entre diferentes ferramentas bioinformáticas.

QIIME 2
QIIME 2 é uma plataforma de bioinformática voltada para análise de dados de microbioma e metabarcoding, permitindo processar sequências, controlar qualidade, inferir ASVs, atribuir taxonomia e gerar visualizações interpretáveis.

---

0. Preparando o ambiente de análise

Ativar ambiente micromamba

```
micromamba activate barcode2
```

Preparando os arquivos necessários

De dentro da sua pasta pessoal, crie links para o fodler de dados e outros aquivos que vamos utilizar:

```
ln -s /home/bioufmg/baiao/Data
ln -s /home/bioufmg/baiao/metadata.tsv
ln -s /home/bioufmg/baiao/manifest.tsv
ln -s /home/bioufmg/baiao/classifier.qza
```

Inspecione o conteúdo da pasta Data
O que são esses arquivos?
Inspecione o arquivo metadata.tsv. Que informações são essas?
Falaremos sobre o arquivo manifest em breve.

---

1. Importação dos dados para o qiime

O QIIME 2 trabalha com arquivos em um formato próprio chamado "artifact" (.qza). Por isso, precisamos importar os arquivos FASTQ para esse formato.

Comando

```
qiime tools import \
 --type 'SampleData[PairedEndSequencesWithQuality]' \
 --input-path manifest.tsv \
 --output-path fastq-data.qza \
 --input-format PairedEndFastqManifestPhred33V2
```

Perguntas

- O que é o arquivo manifest? Que informações ele contém? (Use o comando “less” para visualizar o arquivo)
- Por que precisamos importar os dados para o formato QIIME?

---

2. Avaliação da qualidade dos dados

Antes de processar os dados, é importante avaliar a qualidade das sequências. Essa etapa nos ajuda a decidir como tratar os dados depois.

Comando

```
qiime demux summarize \
  --i-data fastq-data.qza \
  --o-visualization fastq-data.qzv
```

Importante: Os arquivos .qzv são arquivos de visualização. Para ver o seu conteudo, utilize o site: https://view.qiime2.org/. Talvez você tenha que transferir o arquivo para o seu computador antes de passá-lo ao site.

Perguntas

- Quantas reads contém cada amostra?
- Qual o tamanho delas?
- Como a qualidade muda ao longo da read?
- Há diferença de qualidade entre reads forward e reverse?
- Em que posição a qualidade começa a cair mais fortemente?
- Qual a qualidade (Phred score) das bases ruins? Essa qualidade é muito ruim?

3. Remoção de primers (Cutadapt)

Primers e adaptadores não são sequências biológicas e devem ser removidos para evitar interferência nas análises. Para isso vamos utilizar o software cutadapt através de um plugin do qiime

Comando

- **Importante: Insira a sequencia dos primers nos local adequado no comando abaixo!!**

```
qiime cutadapt trim-paired \
  --i-demultiplexed-sequences fastq-data.qza \
  --p-front-f PRIMER_FORWARD \
  --p-front-r PRIMER_REVERSO \
  --p-error-rate 0.1 \
  --p-overlap 3 \
  --p-quality-cutoff-3end 30 \
  --p-match-adapter-wildcards \
  --o-trimmed-sequences fastq-data-trimmed.qza \
  --verbose
```

Gerando uma visualização:

```
qiime demux summarize \
  --i-data fastq-data-trimmed.qza \
  --o-visualization fastq-data-trimmed.qzv
```

Perguntas

- O tamanho das reads mudou?
- A qualidade melhorou?
- Por que remover primers pode melhorar resultados?

Importante:
O gráfico mantém o eixo X original (tamanho das reads original). Portanto, ignore regiões que estão além do tamanho real das reads sem adaptadores.

4. DADA2 (Filtragem, Denoising, Merging)

DADA2 corrige erros de sequenciamento, gera ASVs (variantes de sequência exatas) e remove sequências quiméricas.

Comando

- **Importante: Insira a posição de truncagem desejada. Veja sugestões abaixo!!**

```
qiime dada2 denoise-paired \
  --i-demultiplexed-seqs fastq-data-trimmed.qza \
  --p-trunc-len-f XXX \
  --p-trunc-len-r YYY \
  --o-representative-sequences asv-seqs.qza \
  --o-table asv-table.qza \
  --o-denoising-stats denoising-stats.qza \
  --o-base-transition-stats base-transition-stats.qza
```

Trunc-len é a posição onde a sequência será cortada (truncada). Todas as bases após essas posição serão descartadas em todas as reads. Temos que escolher um valor para as reads forward (f) e reverso (r). Um bom valor de trunc-len:

- Remove o máximo possível de regiões de baixa qualidade
- Mantém um tamanho que garanta sobreposição entre forward/reverse. Precisamos de **≥20 bp de overlap**
- **O valor de trun-len é aplicado na read já sem o adaptador!! Então faça a conta com base no que vê no gráfico.**
- O amplicon tem ~440 bp. Precisamos de **≥20 bp de overlap entre forward e reverso.**

Algumas sugestões de valores:
150, 150
270, 270
280, 210

Visualização

```
qiime metadata tabulate \
  --m-input-file denoising-stats.qza \
  --o-visualization denoising-stats.qzv
```

Perguntas

- Quantas reads foram removidas de cada amostra?
- Qual etapa remove mais dados?
- Por que precisamos de sobreposição entre reads?
- O que acontece se mudar trunc-len? Repita essa etapa com valores diferentes e veja se alguma coisa muda.
- Qual foi o melhor valor dentre os que você testou?

---

5. Tabela de Features

A **feature table** do QIIME 2 é uma tabela que registra quantas vezes cada sequência/ASV foi observada em cada amostra. Em termos práticos, ela é a matriz principal de abundância usada nas etapas seguintes da análise, como diversidade alfa/beta, comparação entre grupos e visualização da composição das comunidades.

Veja a que foi gerada pela sua análise:

Comando

```
qiime feature-table summarize \
  --i-table asv-table.qza \
  --m-metadata-file metadata.tsv \
  --o-summary asv-table.qzv \
  --o-sample-frequencies sample-frequencies.qza \
  --o-feature-frequencies asv-frequencies.qza
```

Investigue o arquivo **asv-table.qzv**

Perguntas

- Quantas ASVs existem?
- Algumas amostras têm poucas reads?
- Isso pode afetar análise?

6. Sequências representativas

As **rep-seqs** do QIIME 2 são as sequências representativas de cada feature/ASV presente na análise. Enquanto a *feature table* diz **quantas vezes** cada ASV aparece em cada amostra, o arquivo de *rep-seqs* contém a **sequência de DNA** correspondente a cada ASV, sendo usado para etapas como atribuição taxonômica, alinhamento e construção de árvores filogenéticas.

Comando

```
qiime feature-table tabulate-seqs \
  --i-data asv-seqs.qza \
  --o-visualization asv-seqs.qzv
```

Investigue o arquivo **asv-seqs.qzv**

Perguntas

- Qual sequência é mais abundante?
- O que faz uma sequencia ser mais abundante na amostra?
- Isso significa que é biologicamente mais importante?

7. Filtragem

As opções de filtragem no QIIME 2 permitem remover amostras ou sequências com baixa qualidade, baixa abundância ou critérios específicos definidos pelo usuário, tornando a análise mais confiável e adequada ao objetivo do estudo.

Vamos filtrar com base em quantas amostras cada ASV foi identificada.

Comando

Para filtrar a feature-table:

```
qiime feature-table filter-features \
  --i-table asv-table.qza \
  --p-min-samples 2 \
  --o-filtered-table asv-table-ms2.qza
```

Para filtrar as sequências representativas:

```
qiime feature-table filter-seqs \
  --i-data asv-seqs.qza \
  --i-table asv-table-ms2.qza \
  --o-filtered-data asv-seqs-ms2.qza
```

Gere visualizações da tabela filtrada. Analise.

Perguntas

- Quais ASVs estamos removendo da análise?
- Isso pode afetar diversidade?
- Por que faríamos esse tipo de filtragem?

---

8. Classificação taxonômica

O **classificador taxonômico** atribui nomes taxonômicos às ASVs comparando suas sequências representativas com um banco de referência. No QIIME 2, isso é frequentemente feito com o comando classify-sklearn, que usa um classificador **Naive Bayes** treinado com sequências do banco **SILVA**, uma base de dados curada e amplamente usada para genes ribossomais, como o 16S rRNA.

Comando

```
qiime feature-classifier classify-sklearn \
  --i-classifier classifier.qza \
  --i-reads asv-seqs-ms2.qza \
  --o-classification taxonomy.qza
```

Perguntas

- Como funciona classificação taxonômica?
- O que significa baixa confiança?

---

9. Barplot taxonômico

O **barplot taxonômico** do QIIME 2 é uma visualização interativa que mostra a composição taxonômica das amostras em diferentes níveis, como filo, classe, ordem, família ou gênero. Ele permite comparar visualmente quais grupos microbianos são mais abundantes em cada amostra ou grupo experimental, a partir da combinação entre a *feature table* e a taxonomia atribuída às ASVs.

Comando

```
qiime taxa barplot \
  --i-table asv-table-ms2.qza \
  --i-taxonomy taxonomy.qza \
  --m-metadata-file sample-metadata.tsv \
  --o-visualization taxa-bar-plots.qzv
```

Visualize o gráfico.

Perguntas

- Explore o gráfico e as opções interativas
- As amostras são similares?
- Existem padrões entre grupos?