

Análise de dados de Metagenômica

Dalmolin Systems Biology Group

Table of contents

Processamento e Análise de Dados de Metagenoma	3
1 Organizando Ambiente para Análise	4
2 Controle de Qualidade e Pré-processamento	5
2.1 Baixar Amostras	5
2.2 Controle de Qualidade com FastQC e MultiQC	5
2.2.1 Estrutura de Arquivos .fastq	5
2.2.2 Ferramentas de Controle de Qualidade	6
2.2.3 Gerando Relatórios de Qualidade	6
3 Montagem	8
3.1 Contexto	8
3.2 Realizando a montagem	8
4 Classificação Taxonômica	9
5 Anotação Funcional	10
6 Análises Downstream	11
6.1 Cálculos de Diversidade	11
6.2 Extraindo informação funcional	11

Processamento e Análise de Dados de Metagenoma

Neste repositório está o material para o curso **Análise de dados de Metagenômica**, organizado pelo Prof. Rodrigo Dalmolin, do Centro Multiusuário de Bioinformática da UFRN.

O curso é dividido em 5 módulos:

- Controle de qualidade e pré-processamento
- Montagem
- Classificação taxonômica
- Anotação funcional
- Análises downstream:
 - Cálculos de diversidade e visualizações
 - Visualizações da anotação funcional

1 Organizando Ambiente para Análise

Primeiro, é necessário criar e ativar um ambiente Conda que contenha todas as ferramentas necessárias para os próximos passos. Caso ainda não tenha o Conda instalado, siga as instruções [neste link](#).

```
$ conda env create -f medusaPipeline.yml
$ conda activate medusaPipeline
```

A estrutura de diretórios a seguir ajudará na organização dos arquivos baixados, arquivos intermediários e seus outputs:

```
$ mkdir -p ./Pipeline/{result,data/{merged,assembled,collapsed,removal/{index,reference},raw
```

2 Controle de Qualidade e Pré-processamento

 WORK IN PROGRESS

2.1 Baixar Amostras

Mude para o diretório `Pipeline/data` e baixe os arquivos de exemplo em pares (pair-end):

```
$ cd Pipeline/data
$ fasterq-dump SRR579292 -e 8
```

 Nota

O argumento `-e` no comando `fasterq-dump` especifica o número de threads que serão utilizadas. Adapte este argumento conforme o desempenho do seu sistema.

2.2 Controle de Qualidade com FastQC e MultiQC

2.2.1 Estrutura de Arquivos `.fastq`

O formato `.fastq` contém as sequências de DNA geradas no sequenciamento, junto com informações sobre a qualidade de cada nucleotídeo. A estrutura de um arquivo `.fastq` segue um padrão repetitivo a cada fragmento sequenciado:

1. **Linha 1:** Contém o identificador único da leitura, que pode incluir informações sobre o sequenciador e a amostra (começa com o símbolo `@`).
2. **Linha 2:** A sequência de nucleotídeos resultante do sequenciamento.
3. **Linha 3:** Opcionalmente usada para anotações ou descrições adicionais (começa com um símbolo `+`).
4. **Linha 4:** A qualidade de cada base na sequência, representada pelo Phred Score, que reflete a confiabilidade da leitura.

A imagem abaixo exemplifica essa estrutura:

```
@Sample Label
AATGATACGGCGTATGTATTCCTCCTATTTGCGAAGTCCAATGGTGCGTATATGCCTTCTGCTTGTC
+
{:FFFFFFF,FFHHHF4FFFF<FED@FFFF:F,,FFFF9C;=?FFF?2AADDFF:FF:FF?GBB?<F99FPFF
```

- **Em Roxo:** Identificador do sequenciamento.
- **Em Laranja:** Sequências adaptadoras (caso ainda não tenham sido removidas), que podem variar dependendo da plataforma de sequenciamento.
- **Em Azul:** O fragmento de DNA sequenciado, conhecido como “*Insert Size*”.
- **Em Verde:** Phred Score, que indica a qualidade de cada base na sequência.

2.2.2 Ferramentas de Controle de Qualidade

Para garantir que as sequências obtidas sejam de boa qualidade e adequadas para análise, utilizamos ferramentas como o **FastQC** e o **MultiQC**.

- **FastQC:** Avalia a qualidade de cada arquivo de sequenciamento individualmente, gerando relatórios com informações sobre a qualidade das bases, conteúdo GC, presença de adaptadores e outros aspectos que podem impactar a análise.
- **MultiQC:** Agrega os relatórios gerados pelo FastQC (ou outras ferramentas), criando um único relatório consolidado que facilita a visualização e interpretação dos dados de múltiplas amostras.

2.2.3 Gerando Relatórios de Qualidade

Para gerar relatórios de controle de qualidade com o FastQC para cada uma das amostras baixadas, use o seguinte comando:

```
# Gerando relatórios de qualidade para cada amostra
for sample in $(ls Pipeline/data/*.fastq.gz);
do
fastqc $sample -o Pipeline/data
done
```

Gere o relatório consolidado contendo todas as amostras a partir do output do FastQC:

```
# Consolidando relatórios com MultiQC
$ multiqc Pipeline/data
```

Interpretando os resultados do MultiQC

Verifique os gráficos de qualidade, presença de contaminantes e distribuições de qualidade das bases. Ajustes podem ser necessários para garantir a integridade dos dados para as etapas seguintes.

3 Montagem

 WORK IN PROGRESS

3.1 Contexto

As *reads*, ou leituras, fragmentos de sequências gerados pelo processo de sequenciamento, podem ser re-organizadas e mescladas em sequências mais longas e contíguas, ou *contigs*. Esse processo é denominado de **Montagem**.

Para se realizar montagem, você pode utilizar uma referência, como o genoma referência de um organismo, que servirá como base para organizar as *contigs*. No entanto, no contexto metagenômico, a modalidade de montagem geralmente realizada é a montagem livre de referência, ou montagem *de novo*.

Como montador *de novo*, utilizaremos o [MEGAHIT](#).

3.2 Realizando a montagem

Vamos montar as leituras pós-descontaminação usando o MEGAHIT:

```
megahit -1 ../removal/unaligned_1.fastq -2  
../removal/unaligned_2.fastq -o SRR579292 -t 8
```

O arquivo de saída `SRR579292.contigs.fa` contém as sequências contíguas correspondentes às leituras usadas.

Nota

Os passos seguintes, como a classificação taxonômica, podem se utilizar tanto das leituras quanto dos contigs. Iremos utilizar as leituras por motivos de didática, mas a maioria das ferramentas utilizadas para a análise de metagenômica podem fazer uso tanto de *reads* quanto de *contigs*.

4 Classificação Taxonômica

 WORK IN PROGRESS

5 Anotação Funcional



WORK IN PROGRESS

6 Análises Downstream

 WORK IN PROGRESS

6.1 Cálculos de Diversidade

6.2 Extraíndo informação funcional