Tutorial Bioinformatica

Marcel Ferreira

2023-11-20

Table of contents

Prefacio **3**
Autores . . . 3
Softwares necessários . . . 3
Opcionais . . . 3
Dados utilizados . . . 3

1 Introduction **4**

2 Dia 1 - Sequenciamento de DNA **5**
2.1 Arquivos . . . 5
2.2 Métricas . . . 5
2.3 Atividades . . . 5

3 Dia 2 - Alinhamento de sequências de DNA **7**
3.1 Arquivos . . . 7

4 Dia 3 - Genotipagem **8**

5 Dia 4 - Análise de sequenciamento Oxford Nanopore **9**

6 Dia 5 - Genotipagem de STRs a partir de dados de NGS **10**

References **11**

Prefacio

Autores

Marcel Rodrigues Ferreira ()

Softwares necessários

- WSL (Windows Subsystem for Linux)
- IGV (site)
- fastqc (github)
- bwa (github)
- minimap2 (github)
- samtools (github)
- freebayes (github)
- gatk (github)
- vcftools (github)
- bcftools (site) (github)
- WhatsHap (github)

Opcionais

- notepad++
- gzip
- HTSlib

Dados utilizados

- fast5/
- fastq/
- genome/
- bam/
- vcf/

1 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

```
1 + 1
```

```
[1] 2
```

2 Dia 1 - Sequenciamento de DNA

2.1 Arquivos

> Dica
>
> Preste atenção nos seus arquivos

2.2 Métricas

$$Read\ Accuracy = \frac{N_{match}}{N_{match} + N_{mis} + N_{del} + N_{ins}} \tag{2.1}$$

$$Mis/Ins/Del = \frac{N_{mis/ins/del}}{N_{match} + N_{mis} + N_{del} + N_{ins}} \tag{2.2}$$

$$P = 10^{\frac{-Qscore}{10}} \tag{2.3}$$

$$Read\ Qscore = -10 \log_{10} [\frac{1}{N} \sum 10^{\frac{-q_1}{10}}] \tag{2.4}$$

2.3 Atividades

O controle de qualidade (QC) dos dados é uma etapa crítica na análise de sequenciamento de nova geração (NGS) para garantir a confiabilidade dos resultados. Abaixo estão as etapas típicas do controle de qualidade:

1. **Análise Inicial com FASTQC:**
    - Execute o FASTQC nas suas leituras brutas para avaliar a qualidade geral. Isso inclui gráficos e estatísticas que indicam a distribuição da qualidade das bases ao longo das reads, a presença de adaptadores, a presença de sequências overrepresented, entre outros.

2. **Identificação de Adaptação (Adapter) e Trimagem:**

    - Com base nos resultados do FASTQC, identifique a presença de adaptadores e sequências indesejadas nas extremidades das reads. Utilize ferramentas como Trimmomatic, Cutadapt ou similar para remover essas sequências, garantindo que apenas dados de alta qualidade sejam mantidos.

3. **Remoção de Leituras de Baixa Qualidade:**

    - Algumas leituras podem conter regiões de baixa qualidade. Considere a remoção dessas leituras ou a trimagem de regiões específicas usando ferramentas adequadas, dependendo da natureza do problema.

4. **Filtragem de Leituras Curtas ou Longas:**

    - Dependendo do seu experimento, você pode querer filtrar leituras muito curtas ou muito longas que possam representar artefatos ou problemas experimentais.

5. **Avaliação de Qualidade Pós-Trimagem:**

    - Após a trimagem e filtragem, execute novamente o FASTQC para avaliar como essas etapas afetaram a qualidade dos dados. Isso ajudará a garantir que você atingiu os padrões de qualidade desejados.

6. **Relatório Final de Controle de Qualidade:**

    - Compile todos os resultados de QC em um relatório final que destaque os principais aspectos da qualidade dos dados. Isso é útil para comunicação interna, bem como para garantir a transparência na publicação de resultados.

3 Dia 2 - Alinhamento de sequências de DNA

3.1 Arquivos

4 Dia 3 - Genotipagem

5 Dia 4 - Análise de sequenciamento Oxford Nanopore

6 Dia 5 - Genotipagem de STRs a partir de dados de NGS

References

Knuth, Donald E. 1984. "Literate Programming." *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.