

# **Tutorial Bioinformatica**

Marcel Ferreira

2023-11-27

# Table of contents

<b>Sobre esse tutorial</b>	<b>3</b>
Autores . . . . .	3
Softwares necessários . . . . .	3
Opcionais . . . . .	3
Dados utilizados . . . . .	4
<b>Introdução</b>	<b>5</b>
Primeiros passos . . . . .	6
<b>Dia 1 - Sequenciamento de DNA</b>	<b>7</b>
Arquivos . . . . .	7
Métricas . . . . .	7
Atividades . . . . .	7
<b>Dia 2 - Alinhamento de sequências de DNA</b>	<b>9</b>
Arquivos . . . . .	9
Atividades . . . . .	9
<b>Dia 3 - Genotipagem</b>	<b>10</b>
<b>Dia 4 - Análise de sequenciamento Oxford Nanopore</b>	<b>11</b>
Atividades . . . . .	11
<b>Dia 5 - Genotipagem de STRs a partir de dados de NGS</b>	<b>13</b>
<b>Referências</b>	<b>14</b>

# Sobre esse tutorial

## Autores

Marcel Rodrigues Ferreira ()

DEMAIS

## Softwares necessários

- WSL (Windows Subsystem for Linux)
- [IGV](#)(Robinson et al. 2011)
- [fastqc](#)
- [bwa](#)(Li 2013)
- [minimap2](#)(Li 2018, 2021)
- [samtools](#)(Danecek et al. 2021)
- [freebayes](#)(Garrison and Marth 2012)
- [gatk](#)
- [vcftools](#)(Danecek et al. 2011)
- [bcftools](#)(Danecek et al. 2021)
- [whasthap](#)(Martin et al. 2016)
- [NanoPlot](#)(De Coster and Rademakers 2023)
- [chopper](#)(De Coster and Rademakers 2023)

## Opcionais

- [notepad++](#)
- [gzip](#)
- [HTSlib](#)

## Dados utilizados

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

# Introdução

Bem-vindos ao Workshop de Bioinformática Aplicada à Genética Forense: Análise de Dados de Sequenciamento de Segunda e Terceira Geração. Este curso abrangente foi projetado para fornecer a vocês, participantes entusiasmados, uma imersão prática nas técnicas avançadas de análise de dados genômicos, com foco especial na aplicação forense.

A genética forense tornou-se uma ferramenta essencial na resolução de casos criminais, identificação de indivíduos e estabelecimento de relações familiares. Neste workshop de cinco dias, exploraremos os fundamentos e as aplicações práticas do sequenciamento de DNA, abordando desde os conceitos básicos até as técnicas avançadas de genotipagem de STRs (Short Tandem Repeats) a partir de dados de Next-Generation Sequencing (NGS).

**Dia 1 - Sequenciamento de DNA:** Iniciaremos nossa jornada explorando os princípios fundamentais do sequenciamento de DNA de segunda e terceira geração. Compreenderemos as tecnologias por trás desses métodos e sua importância na geração de dados genômicos de alta qualidade.

**Dia 2 - Alinhamento de Sequências de DNA:** No segundo dia, mergulharemos na etapa crucial de alinhamento de sequências de DNA. A precisão dessa fase é vital para extrair informações significativas dos dados brutos e identificar variações genéticas relevantes.

**Dia 3 - Identificação de Variantes:** Aprofundando-nos ainda mais, dedicaremos o terceiro dia à identificação de variantes genéticas. Exploraremos ferramentas e estratégias para detectar mutações, SNPs (Single Nucleotide Polymorphisms) e outras alterações que desempenham um papel crucial na individualidade genômica.

**Dia 4 - Análise de Sequenciamento Oxford Nanopore:** No quarto dia, abordaremos uma tecnologia revolucionária: o sequenciamento Oxford Nanopore. Compreenderemos suas vantagens, desafios e exploraremos casos de uso específicos na genética forense.

**Dia 5 - Genotipagem de STRs a partir de dados de NGS:** Encerraremos o workshop com uma exploração prática da genotipagem de STRs, uma ferramenta valiosa para estabelecer perfis genéticos únicos. Aprenderemos a interpretar e analisar esses dados, fornecendo insights fundamentais para investigações forenses.

Ao longo desta semana, vocês serão desafiados a aplicar os conhecimentos adquiridos em exercícios práticos e estudos de caso, preparando-os para enfrentar os desafios reais da genética forense na era da bioinformática avançada. Esteja preparado para uma jornada intensiva de aprendizado e descoberta!

## Primeiros passos

- Instale os Softwares necessários (?@sec-softwares-necessários);

### Usuários windows

Usuários windows precisam instalar o Subsistema Windows para Linux (WSL).  
Os softwares FASTQC e IGV precisam ser instalados no windows e **não no WSL**.

- Baixe os dados que serão utilizados neste workshop (?@sec-dados-utilizados);

### Utilize o email correto

Para ter acesso aos dados utilize o email que foi fornecido durante a inscrição no evento.  
Em caso de erro entre em contato com a organização.

### Dicas

Utilize o comando `htop` para monitorar o consumo de memória em seu computador durante todas as atividades deste tutorial.

Utilize os argumento `-h` ou `--help` para visualizar ajuda sobre o uso dos softwares via terminal.

`SOFTWARE -h`

Adicionar o comando `time` antes de rodar os códigos para saber o tempo que demorou irá te ajudar a se programar para atividades futuras.

`time [COMANDOS...]`

# Dia 1 - Sequenciamento de DNA

## ! Importante

Verifique se o FASTQC está instalado.

## Arquivos

## Métricas

$$Read\ Accuracy = \frac{N_{match}}{N_{match} + N_{mis} + N_{del} + N_{ins}} \quad (0.1)$$

$$Mis/Ins/Del = \frac{N_{mis/ins/del}}{N_{match} + N_{mis} + N_{del} + N_{ins}} \quad (0.2)$$

$$P = 10^{\frac{-Q_{score}}{10}} \quad (0.3)$$

$$Read\ Q_{score} = -10 \log_{10} \left[ \frac{1}{N} \sum 10^{\frac{-q_1}{10}} \right] \quad (0.4)$$

## 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.



# Dia 2 - Alinhamento de sequências de DNA

## Arquivos

Os arquivos utilizados para estas análises serão os `.fastq` analisados no primeiro dia.

### Arquivos `.fastq`

Preste atenção para o caminho da pasta aonde estão os `.fastq`.

## Atividades

- Realize o alinhamento das amostras utilizando `BWA` e o genoma de referência `hg38.fa`;
- Converta os arquivos `.sam` para o formato `.bam`;
- Crie o índice dos arquivos `.bam`;
- Alinhe novamente, desta vez utilizando `minimap2` e repita as demais etapas;
- Utilize o `IGV` para visualizar os alinhamentos;

## **Dia 3 - Genotipagem**

## Dia 4 - Análise de sequenciamento Oxford Nanopore

### ! Importante

O software `guppy` só está disponível para download via comunidade da Oxford Nanopore. Para este tutorial fornecemos um arquivo `.tar` para instalação em sua máquina. Para instalar siga os seguintes passos:

Acesse a pasta que contem o arquivo `.tar` e descompacte;

```
tar -xf ont-guppy-cpu_6.5.7_linux64.tar.gz
```

Verifique o caminho completo para a pasta

```
pwd
```

Executando `guppy` via caminho completo (Exemplo pedindo ajuda)

```
./guppy_basecaller --help
```

Este tutorial foi inspirado...

### Atividades

- Verifique os workflows disponíveis para esta versão de `guppy`;
- Realize uma chamada de base para todos os arquivos `fast5` contidos na pasta ...;

### ⚠ Aviso

O tempo de execução da chama de base (para este conjunto de dados) é superior a **12 horas** em máquinas de uso pessoal, e pode acabar inutilizando seu uso. Caso deseje praticar, recomendamos que seja feito em um momento onde não precise da máquina para outras atividades. Para este tutorial utilize o resultado da chamada de base contido na pasta ...

- Avalie a qualidade do arquivo resultado `.fastq` utilizando `FASTQC` e `nanoQC`;
- Filtre a read baseado no tamanho e na qualidade utilizando `chopper`;
- Monte o genoma utilizando `BWA` e `minimap2`;

### Opcional

Realize uma montagem *de novo* utilizando o pipeline minimap2-miniasm. Este processo pode levar várias horas/dias. Tome cuidado!

## **Dia 5 - Genotipagem de STRs a partir de dados de NGS**

# Referências

- Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics* 27 (15): 2156–58. <https://doi.org/10.1093/bioinformatics/btr330>.
- Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. 2021. “Twelve Years of SAMtools and BCFtools.” *GigaScience* 10 (2). <https://doi.org/10.1093/gigascience/giab008>.
- De Coster, Wouter, and Rosa Rademakers. 2023. “NanoPack2: Population-Scale Evaluation of Long-Read Sequencing Data.” Edited by Can Alkan. *Bioinformatics* 39 (5). <https://doi.org/10.1093/bioinformatics/btad311>.
- Garrison, Erik, and Gabor Marth. 2012. “Haplotype-Based Variant Detection from Short-Read Sequencing.” <https://doi.org/10.48550/ARXIV.1207.3907>.
- Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” <https://doi.org/10.48550/ARXIV.1303.3997>.
- . 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” Edited by Inanc Birol. *Bioinformatics* 34 (18): 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>.
- . 2021. “New Strategies to Improve Minimap2 Alignment Accuracy.” Edited by Can Alkan. *Bioinformatics* 37 (23): 4572–74. <https://doi.org/10.1093/bioinformatics/btab705>.
- Martin, Marcel, Murray Patterson, Shilpa Garg, Sarah O Fischer, Nadia Pisanti, Gunnar W Klau, Alexander Schöenhuth, and Tobias Marschall. 2016. “WhatsHap: Fast and Accurate Read-Based Phasing.” <http://dx.doi.org/10.1101/085050>.
- Robinson, James T, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1): 24–26. <https://doi.org/10.1038/nbt.1754>.