

Tutorial Bioinformatica

Marcel Ferreira - Bolsista/CAPES

2023-11-29

Table of contents

| | |
|---|-----------|
| Sobre esse curso | 3 |
| Realização | 3 |
| Apoio | 3 |
| Autores | 3 |
| Colaboradores | 3 |
| Introdução | 4 |
| Primeiros passos | 5 |
| Configurações de sistema | 5 |
| Softwares necessários | 5 |
| Usuários Windows | 5 |
| No Ubuntu | 5 |
| Instalação | 6 |
| Dados utilizados | 7 |
| Dia 1 - Sequenciamento de DNA | 9 |
| Arquivos | 9 |
| Métricas | 9 |
| Atividades | 9 |
| Dia 2 - Alinhamento de sequências de DNA | 11 |
| Arquivos | 11 |
| Atividades | 11 |
| Dia 3 - Genotipagem | 13 |
| Arquivos | 13 |
| Atividades | 13 |
| Dia 4 - Análise de sequenciamento Oxford Nanopore | 15 |
| Atividades | 15 |
| Dia 5 - Genotipagem de STRs a partir de dados de NGS | 17 |
| Referências | 18 |

Sobre esse curso

Realização



Figure 1: Realização UnB, UNESP e USP.

Apoio



Figure 2: CAPES-PROCAD Edital n° 16/2020

Autores

Celso Teixeira Mendes Junior
Erick da Cruz Castelli
Marcel Rodrigues Ferreira
Tamara Soledad Frontanilla Recalde

Colaboradores

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 uma imersão prática nas técnicas 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

Configurações de sistema

Antes de iniciarmos o tutorial, é imperativo garantir que o sistema atenda às configurações mínimas para uma experiência estável. Utilizaremos sistema Linux. Recomenda-se que a máquina disponha de, no mínimo, 30 GB de armazenamento e 8 GB de memória RAM. No entanto, para uma performance ideal e considerando o potencial de expansão das aplicações, encorajamos a utilização de um sistema com mais de 50 GB de armazenamento e, no mínimo, 16 GB de memória RAM. Essas configurações mais robustas assegurarão não apenas a instalação suave do software, mas também a capacidade de executar múltiplas aplicações de forma eficiente, proporcionando uma experiência mais fluida e responsiva ao usuário.

Softwares necessários

Usuários Windows

- WSL (Windows Subsystem for Linux)
- [IGV](#) (Robinson et al. 2011)
- [FASTQC](#)
- [notepad++](#)

💡 Tutorial para instalar o WSL

Siga o tutorial da microsoft para instalar o WSL.

<https://learn.microsoft.com/pt-br/windows/wsl/install>

No Ubuntu

- [IGV](#) (Robinson et al. 2011)
- [FASTQC](#)
- [Trimmomatic](#) (Bolger, Lohse, and Usadel 2014)
- [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)
- [WhatsHap](#) (Martin et al. 2016)
- [NanoPlot](#) (De Coster and Rademakers 2023)
- [chopper](#) (De Coster and Rademakers 2023)
- [cramino](#) (De Coster and Rademakers 2023)

Opcionais

- [gzip](#)
- [HTSLib](#)

Instalação

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**.

Ao terminar a instalação do WSL e de configurar seu usuário no linux utilize os seguintes comandos:

```
sudo apt-get update
```

```
sudo apt-get upgrade
```

Estes comandos irão garantir que o seu sistema esteja atualizado.

Sobre o comando `sudo`

O comando `sudo` permite ao usuário executar comandos com permissão superior. Para isso você precisará da sua **senha** (ou do administrador)!

Para instalar softwares utilize o comando `apt install` da seguinte forma:

```
sudo apt install [SOFTWARE]
```

Dados utilizados

Baixe os dados que serão utilizados neste workshop via [Google Drive](#);

! 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.
Os dados totalizam ~13 GB, se atente para isso.

Os dados estão contidos nesta estrutura de pastas descritas a baixo:

WorkshopDados/

```
--genome/
|  |--chr17.fas1
|  °--hg38.fa
|--fast5/
|  °--*arquivos.fast52
|--guppy__installer/
|  °--ont-guppy-cpu__6.5.7__linux64.tar.gz3
|--LongReadsFastq/
|  |--HG00096.hg38.fastq
|  °--HG00099.hg38.fastq
°--ShortReadsFastq/
|  |--HG00096__r1.fastq
|  |--HG00096__r2.fastq
|  |--HG00097__r1.fastq
|  |--HG00097__r2.fastq
|  |--HG00099__r1.fastq
|  |--HG00099__r2.fastq
|  |--HG00100__r1.fastq
|  |--HG00100__r2.fastq
|  |--NA18486__r1.fastq
|  |--NA18486__r2.fastq
|  |--NA18487__r1.fastq
|  |--NA18487__r2.fastq
|  |--NA18488__r1.fastq
|  |--NA18488__r2.fastq
|  |--NA19648__r1.fastq
|  |--NA19648__r2.fastq
```

¹Genoma do cromossomo 17 obtido em https://timkahlke.github.io/LongRead_tutorials/

²Arquivos **fast5** obtidos em https://timkahlke.github.io/LongRead_tutorials/

³Instalador obtido em <https://community.nanoporetech.com/>

```
|--NA19649_r1.fastq  
|--NA19649_r2.fastq  
|--NA19651_r1.fastq  
°--NA19651_r2.fastq
```

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** esta instalado.

Arquivos

Serão utilizados os arquivos contidos na pasta **WorkshopDados/shortReads/**.

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

1. Indexação do Genoma de Referência:

- Antes de realizar o alinhamento, é necessário indexar o genoma de referência usando o comando **bwa index**. Isso cria arquivos que aceleram o processo de alinhamento.

```
bwa index reference_genome.fa
```

2. Alinhamento de Sequências:

- Use o **bwa mem** para alinhar suas sequências de DNA ao genoma de referência.
- ```
bwa mem -R "@RG\tID:{SAMPLE}\tSM:``{SAMPLE}``" reference_genome.fa``{SAMPLE}_r1``.f
> alinhamento.sam
```

Substitua `reference_genome.fa`, `read1.fq` e `read2.fq` pelos nomes dos arquivos correspondentes.

### 3. Converter Formato SAM para BAM:

- O arquivo de saída do **bwa mem** é no formato SAM. Converta-o para o formato BAM, mais compacto e eficiente.
- ```
samtools sort alinhamento.sam > alinhamento_sorted.bam
```

4. Ordenar e Indexar o Arquivo BAM:

- Ordene o arquivo BAM para facilitar a busca e indexe-o para melhorar o desempenho de ferramentas subsequentes.
- `samtools index alinhamento_sorted.bam`

5. Remoção de Duplicatas (opcional):

- Dependendo da aplicação, você pode querer remover duplicatas do seu arquivo BAM para evitar viés em análises subsequentes.
- `samtools rmdup alinhamento_sorted.bam alinhamento_nodups.bam`

6. Visualização do Alinhamento:

- Use o IGV (Integrative Genomics Viewer) para visualizar o alinhamento e verificar sua qualidade.

7. Avaliação de Cobertura:

- Utilize ferramentas como `samtools depth` para calcular a cobertura do genoma e avaliar a profundidade de sequenciamento em diferentes regiões.
- `samtools depth alinhamento_sorted.bam > cobertura.txt`

Dia 3 - Genotipagem

Arquivos

Atividades

1. Preparação do Ambiente:

- Certifique-se de que o FreeBayes está instalado no seu ambiente. Você pode instalar o FreeBayes usando gerenciadores de pacotes, como o **conda** ou o **pip**, ou compilando-o a partir do código-fonte.
- `conda install -c bioconda freebayes`

2. Indexação do Genoma de Referência (se ainda não estiver indexado):

- Assim como na etapa de alinhamento, o genoma de referência deve ser indexado.
- `samtools faidx reference_genome.fa`

3. Chamada de Variantes com FreeBayes:

- Execute o FreeBayes para chamar variantes a partir do arquivo BAM gerado após o alinhamento.
- `freebayes -f reference_genome.fa alinhamento_sorted.bam > variantes.vcf`

4. Filtragem de Variantes (opcional):

- Dependendo dos seus critérios e do tipo de análise, pode ser necessário filtrar as variantes chamadas pelo FreeBayes para reduzir o número de falsos positivos.
- `bcftools view --exclude 'QUAL>5' variantes.vcf > variantes_filtradas.vcf`

Adapte os critérios de filtragem conforme necessário.

5. Análise e Interpretação de Variantes:

- Utilize ferramentas como **VCFtools**, **bcftools** ou **GATK** para realizar análises adicionais no arquivo VCF, como filtragem específica, anotações e interpretação biológica.

6. Visualização de Variantes:

- Use o IGV para visualizar as variantes em relação ao genoma de referência e avaliar sua qualidade.

7. Anotação de Variantes:

- Anote as variantes chamadas usando ferramentas como ANNOVAR, Variant Effect Predictor (VEP) ou AnnovarR.

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 `NanoPlot`;
- 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

- Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20. <https://doi.org/10.1093/bioinformatics/btu170>.
- 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>.