

Tutorial Bioinformatica

Marcel Ferreira - Bolsista/CAPES

2024-10-06

Table of contents

Sobre esse curso	5
Realização	5
Apoio	5
Autores	5
Colaboradores	6
Introdução	7
Primeiros passos	8
Configurações de sistema	8
Softwares necessários	8
Usuários Windows	8
No Ubuntu	9
Instalando via CONDA	9
Instalando o CONDA via Miniconda	9
Crie e ative o ambiente	10
Instalação via linha de comando	10
Dados utilizados	10
Dia 1 - Sequenciamento de DNA	15
Arquivos	15
Atividades da manhã	15
Atividades da tarde	16
Dia 2 - Alinhamento de sequências de DNA	17
Arquivos	17
Atividades da manhã	17
Atividades da tarde	18
Dia 3 - Genotipagem	20
Arquivos	20
Atividades da manhã	20
Atividades da tarde	21
Dia 4 - Análise de sequenciamento Oxford Nanopore	23
Atividades da manhã	23

Atividades da tarde	26
Dia 5 - Marcadores STRs	27
PRIMER PASO - DESCARGAS	27
HIPSTR	28
INSTALAÇÃO	28
CÓMO CRIAR OS ÍNDICES DAS AMOSTRAS?	29
COMANDOS	29
INTERPRETAÇÃO DO ARQUIVO DE SAÍDA	30
Referências	31
ANEXO: Dicas para uso do Ubuntu/WSL	32
Navegação e Diretórios	32
Listar Conteúdo do Diretório	32
Mudar de Diretório	32
Diretório Atual	32
Criar Diretório	32
Manipulação de Arquivos	32
Copiar Arquivo	32
Mover/Renomear Arquivo	33
Remover Arquivo	33
Visualização de Conteúdo	33
Visualizar Conteúdo do Arquivo	33
Visualizar Conteúdo do Arquivo (página por página)	33
Pacotes e Atualizações	33
Atualizar Lista de Pacotes	33
Atualizar Pacotes Instalados	33
Instalar Novo Pacote	34
Gerenciamento de Usuários	34
Adicionar Usuário ao Grupo	34
Mudar Senha do Usuário	34
Processos	34
Listar Processos	34
Matar um Processo por ID	34
Monitorar Recursos do Sistema (htop)	34
Medir Tempo de Execução de um Comando (time)	35
Rede	35
Verificar Configurações de Rede	35
Testar Conexão com um Endereço IP	35
Entrada e Saída Padrão (stdin/stdout)	35
Outros Comandos Úteis	35
Ajuda sobre um Comando	35

Sair do Terminal	36
-----------------------------------	-----------

Sobre esse curso

Realização



Figure 1: Realização UnB, UNESP e USP.

Apoio



Figure 2: CAPES-PROCAD Edital nº 16/2020. Processos 88887.516236/2020-00 e 88881.516238/2020-01.

Autores

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

Colaboradores

Gabriela Sato Paes

Ícaro Scalisse de Freitas Santos

Matheus de Souza Ferrari

Thássia Mayra Telles Carratto

Vítor Matheus Soares Moraes

Viviane Aparecida de Oliveira Ciriaco

Introdução

Bem-vindos ao II 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 geração. Compreenderemos as tecnologias por trás desses métodos e sua importância na geração de dados genômicos de alta qualidade. Também serão analisados dados brutos de sequenciamento e seu controle 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.

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*), InDels, e outras variantes 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. 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, 40 GB de armazenamento, 8 GB de memória RAM e um processador i5/i7 ou compatível. No entanto, para uma performance ideal e considerando o potencial de expansão das aplicações, encorajamos a utilização de um sistema com mais de 60 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>

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**. Anote a **senha** que você configurou. Ela será fundamental durante o uso do WSL!!!!

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)
- [vcftools](#) (Danecek et al. 2011)
- [bcftools](#) (Danecek et al. 2021)
- [NanoPlot](#) (De Coster and Rademakers 2023)
- [chopper](#) (De Coster and Rademakers 2023)
- [HipSTR](#) (Willems et al. 2017)
- [gzip](#)

Instalando via CONDA

Usar o CONDA para criar um ambiente garante que todos os participantes estejam utilizando exatamente os mesmos programas e versões, evitando problemas que podem surgir por diferenças entre os computadores. Assim, você terá todas as ferramentas necessárias instaladas de forma organizada e padronizada, facilitando o acompanhamento durante o workshop, mesmo se você não tiver muita experiência com a instalação de programas. Isso ajuda a garantir que todos possam focar no conteúdo sem se preocupar com configurações complicadas.

Instalando o CONDA via Miniconda

Utilize os comandos a baixo

```
cd ~

mkdir -p ~/miniconda3

wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh -O ~/miniconda3/m

bash ~/miniconda3/miniconda.sh -b -u -p ~/miniconda3

rm -rf ~/miniconda3/miniconda.sh

~/miniconda3/bin/conda init bash
```

```
~/miniconda3/bin/conda init zsh
```

Reinicie o sistema. No **WSL** é só fechar a abrir novamente.

Baixe o arquivo de configuração do CONDA para o Workshop clicando [aqui](#).

Crie e ative o ambiente

```
conda env create -f workshopbioinfo.yml  
  
conda activate workshopbioinfo
```

Instalação via linha de comando

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 no linux (diretamente ou no WSL) utilize o comando **apt install** da seguinte forma:

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

Dados utilizados

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

! 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 ~20 GB. **Atente-se para isso!!!**

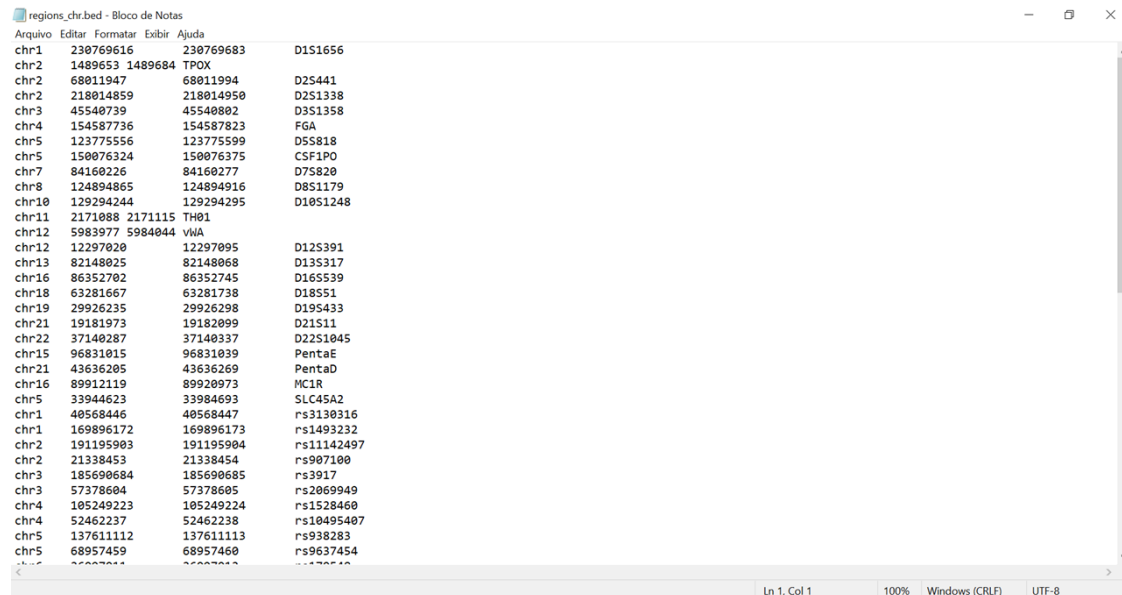
As amostras utilizadas neste curso foram sequenciadas pelo projeto 1000 genomas.

! Regiões com reads

Os arquivos **fastq** fornecidos para este curso foram preparados contendo apenas as regiões que estão nos arquivos **bed** fornecidos na pasta **Genome/**.

Exemplo:

Abrindo o arquivo BED podemos notar que ele contém a informação da posição baseada no cromossomo (**coluna 1**) e seu início (**coluna 2**) e fim (**coluna 3**). De fato, estas são as únicas colunas obrigatórias de um BED. A coluna 4 neste arquivo contém anotação da região que iremos trabalhar.



Chromosome	Start	End	Annotation
chr1	230769616	230769683	D1S1656
chr2	1489653	1489684	TPOX
chr2	68011947	68011994	D2S441
chr2	218014859	218014950	D2S1338
chr3	45540739	45540802	D3S1358
chr4	154587736	154587823	FGA
chr5	123775556	123775599	D5S818
chr5	150076324	150076375	CSF1PO
chr7	84160226	84160277	D7S820
chr8	124894865	124894916	D8S1179
chr10	129294244	129294295	D10S1248
chr11	2171088	2171115	TH01
chr12	5983977	5984044	vWA
chr12	12297020	12297095	D12S391
chr13	82148025	82148068	D13S317
chr16	86352702	86352745	D16S539
chr18	63281667	63281738	D18S51
chr19	29926235	29926298	D19S433
chr21	19181973	19182099	D21S11
chr22	37140287	37140337	D22S1045
chr15	96831015	96831039	PentaE
chr21	43636205	43636269	PentaD
chr16	89912119	89920973	MC1R
chr5	33944623	33984693	SLC45A2
chr1	40568446	40568447	rs3130316
chr1	169896172	169896173	rs1493232
chr2	191195903	191195904	rs11142497
chr2	21338453	21338454	rs907100
chr3	185690684	185690685	rs3917
chr3	57378604	57378605	rs2069949
chr4	105249223	105249224	rs1528460
chr4	52462237	52462238	rs10495407
chr5	137611112	137611113	rs938283
chr5	68957459	68957460	rs9637454

Figure 3: Estrutura de um arquivo BED.

Quando abrimos o arquivo BED no IGV ele marca em azul as posições das coordenadas genômicas:

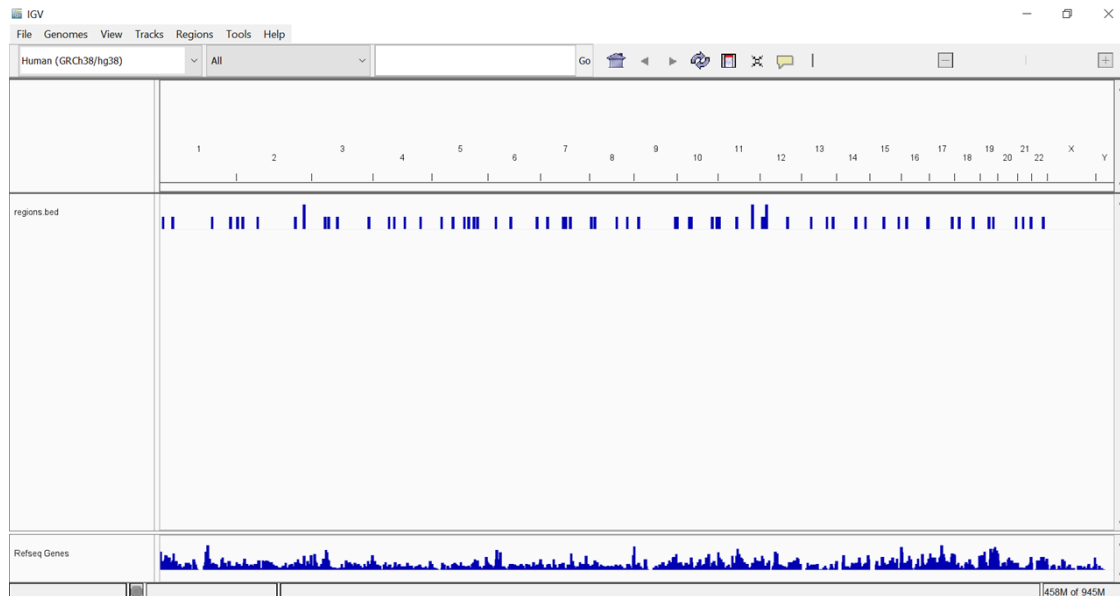


Figure 4: Abrindo o arquivo regions.bed no IGV

Para visualizar a região no IGV devemos utilizar a seguinte notação: {cromossomo}:{start}-{end}. Para o D1S1656 fica chr1:230769616-230769683.

Exome_ShortReads/

fastq/

bam/

WGS_ShortReads/

fastq/

bam/

SimulatedReads/

Dia 1 - Sequenciamento de DNA

! Importante

Verifique se o **FASTQC** está instalado.

Arquivos

Serão utilizados os arquivos contidos na pastas:

- `WorkshopBioinfo2024/WGS_ShortReads/fastq/`
- `WorkshopBioinfo2024/Exome_ShortReads/fastq/`

Estão dados vieram do projeto [1000 genomas](#).

Atividades da manhã

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 Adaptadores e Trimagem:

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

- Adapte o comando abaixo para a amostra que esta analisando, substituindo HG00097 pelo código dela:

```
trimmomatic PE -phred33 \
WGS_ShortReads/fastq/HG00097_r1.fastq WGS_ShortReads/fastq/HG00097_r2.fastq \
trimmed_r1.fastq trimmed_r1_unpaired.fastq trimmed_r2.fastq trimmed_r2_unpaired.fastq
ILLUMINACLIP:~/miniconda3/pkgs/trimmomatic-0.39-hdfd78af_2/share/trimmomatic-0.39-2
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

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. Utilize o comando a baixo para manter leituras com qualidade acima de 27 e tamanho mínimo de 100.

```
trimmomatic PE -phred33 \
WGS_ShortReads/fastq/HG00097_r1.fastq WGS_ShortReads/fastq/HG00097_r2.fastq \
trimmed_r1.fastq trimmed_r1_unpaired.fastq trimmed_r2.fastq trimmed_r2_unpaired.fastq
ILLUMINACLIP:~/miniconda3/pkgs/trimmomatic-0.39-hdfd78af_2/share/trimmomatic-0.39-2
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:27 MINLEN:100
```

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.

Atividades da tarde

De 1 a 3, compare dos dados de cada amostra para WGS e Exoma.

1. Qual é a pontuação geral de qualidade em todas as bases de sua amostra?
2. Há alguma contaminação de adaptador nas leituras de sequenciamento?
3. As sequências têm algum conteúdo incomum de GC?
4. Execute FASTQC para as leituras simuladas ([WorkshopBioinfo2024/SimulatedReads](#)). Quais as principais diferenças entre as amostras?
5. O que acontece se você realizar a trimagem utilizando `trimmomatic` com os parâmetros qualidade acima de 30 e tamanho mínimo de 100 com ambas as amostras?

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`. Eles estão nas pastas `Exome_ShortReads/fastq` e `WGS_ShortReads/fastq`.

Atividades da manhã

1. Indexação do Genoma de Referência:

- Para o alinhamento de sequências utilizaremos o programa `bwa`. Há outros, como `bowtie2`, `minimap2`, que podem ser utilizados caso seja conveniente. Os comandos abaixo preparam o genoma de referência para o `bwa`.
- 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 hg38.fa
```

Indexação do genoma

O processo para criar o índice do genoma via `bwa index` demora bastante tempo para ser realizado. Mesmo em máquinas com grandes capacidades de memória. Devido a isso colocamos na pasta `WorkshopDados/genome/` os arquivos resultantes desta etapa, que são os arquivos com extensão `.amb`, `.ann`, `.bwt`, `.fai`, `.pac` e `.sa`. O genoma de referência humano está no arquivo `hg38.fa`. Esse é a última versão do genoma de referência e a mais utilizada no mundo.

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}" hg38.fa {SAMPLE}_r1.fastq {SAMPLE}_r2.fastq
```

Substitua **{SAMPLE}** pelos nomes dos arquivos correspondentes.

💡 Automatizando o processo (Opcional)

Caso tenha experiência em programação, você pode utilizar um **loop**, como **for** ou **while**, para rodar todas as amostras ao mesmo tempo. Você pode utilizar o próprio **bash** (ubuntu) ou sua linguagem de programação favorita, como **python**, **perl**, **R**, etc.

Você poderá incluir as etapas seguintes no mesmo **loop**.

3. Converter Formato SAM para BAM e ordenar os reads:

- O arquivo de saída do **bwa mem** é no formato SAM. Converta-o para o formato BAM, mais compacto e eficiente.

```
samtools sort {SAMPLE}.sam > {SAMPLE}.bam
```

4. Indexar o Arquivo BAM:

- Ordene o arquivo BAM para facilitar a busca e indexe-o para melhorar o desempenho de ferramentas subsequentes.

```
samtools index {SAMPLE}.bam
```

5. Visualização do Alinhamento:

- Use o IGV (Integrative Genomics Viewer) para visualizar o alinhamento (**{SAMPLE}.bam**) e verificar sua qualidade.

! Regiões com reads

Os arquivos **fastq** fornecidos para este curso foram preparados contendo apenas as regiões que estão nos arquivos **bed** fornecidos na pasta **Genome/**.

6. Use o IGV para observar várias amostras ao mesmo tempo

Atividades da tarde

Para resolver os exercícios, utilizem todas as amostras do conjunto de dados, com exceção da amostra HG00097.

1. Qual a cobertura das amostras para os marcadores PentaE, PentaD, vWA [PODEM SER OUTROS].
2. Qual é a profundidade média de cobertura das leituras alinhadas?
3. Há alguma região com uma alta porcentagem de *mismatches* ou *indels* no genes MC1R e SLC45A2?
4. Quão uniformemente as leituras estão distribuídas no genoma de referência?

Dia 3 - Genotipagem

Arquivos

Serão utilizados os arquivos BAM que foram gerados no dia 2.

Atividades da manhã

1. Preparação do Ambiente:

- Certifique-se de que o **freebayes** está instalado no seu ambiente. Você pode instalar com:

```
sudo apt install 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 hg38.fa
```

- Caso utilize um genoma menor, atualize o nome.

3. Crie um arquivo com os nomes das amostras BAM:

- Entre na pasta onde estão os arquivos BAM gerados e utilize o comando **ls** como abaixo

```
ls *.bam > bam_list.txt
```

4. 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 -L bam_list.txt -t bed_file.bed > variantes.vcf
```

Otimizando o uso do **freebayes**

Caso o computador que esteja utilizando aborte o processo por falta de memória, você pode optar por reduzir o número de amostras em **bam_list.txt**. É importante destacar que a análise correta de genotipagem via **freebayes** requer que todas as amostras sejam analisadas simultaneamente, mas para fim de aprendizado esta é uma estratégia.

5. 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. Abaixo estão alguns exemplos:

```
bcftools view --exclude 'QUAL<1' variantes.vcf > variantes_filtradas.vcf
```

```
bcftools view --trim-alt-alleles variantes_filtradas.vcf > variantes_filtradas_trim.vcf
```

```
bcftools view --min-ac 1 variantes_filtradas_trim.vcf > variantes_filtradas_trim_minac.vcf
```

```
bcftools norm -f hg38.fa variantes_filtradas_trim_minac.vcf > variantes_filtradas_trim_minac_norm.vcf
```

- Adapte os critérios de filtragem conforme necessário.

6. Análise e Interpretação de Variantes:

- Utilize ferramentas como **VCFtools** para realizar análises adicionais no arquivo VCF, como filtragem específica e anotações.

7. Visualização de Variantes:

- Use o IGV para visualizar as variantes em relação ao genoma de referência e avaliar sua qualidade.
- Você pode abrir ao mesmo tempo o VCF e o alinhamento (BAM) de uma mesma amostra. Caso deseje.

Atividades da tarde

1. Quantos SNPs e *indels* foram identificados em suas amostras?
2. Qual é a qualidade das variantes identificadas?
3. As variantes identificadas estão localizadas em regiões codificadoras ou não codificadoras do genomas?
4. Como é a relação heterozigosidade/homozigosidade em todas as amostras?

5. Há algum gene ou região genômica específica com uma densidade excepcionalmente alta de variantes?

Dia 4 - Análise de sequenciamento Oxford Nanopore

! Importante

O software guppy só esta disponível para download via comunidade da Oxford Nanopore. Para este tutorial fornecemos um arquivo .tar para instalação em sua máquina. O arquivo esta na pasta **WorkshopDados/guppy_installer/**

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 no tutorial do Tim Kahlke¹ e em nossas experiências durante os trabalhos com ONT.

Atividades da manhã

1. Realize uma chamada de base utilizando guppy (Opicional):

- Verifique os workflows disponíveis para esta versão de guppy;

```
./Downloads/ont-guppy-cpu/bin/guppy_basecaller --print_workflows
```

- Sabendo que este sequenciamento foi realizado utilizando o kit SQK-LSK108 e a flowcell MIN106, qual a configuração a ser utilizada?

¹ https://timkahlke.github.io/LongRead_tutorials/

- Realize a chamada de base para todos os arquivos `fast5` contidos na pasta **WorkshopDados/fast5/**;

```
./Downloads/ont-guppy-cpu/bin/guppy_basecaller -i [PASTAFAST5] -s /
./guppy_out -c [CONFIG].cfg --num_callers 2 --cpu_threads_per_caller 1
```

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 **WorkshopDados/LongReadsFastq/**

2. Avaliação da qualidade do arquivo resultado `.fastq`:

- Utilize o `NanoPlot` para gerar gráficos de qualidade da amostra;

```
NanoPlot --fastq [AMOSTRA].fastq -o [OUTDIR] --N50 --verbose
```

3. Filtre as leituras baseado no tamanho e qualidade:

- Utilize `chopper` para isso.

```
chopper < [AMOSTRA].fastq -q [QUALIDADE] -l [TAMANHO_MIN] > [AMOSTRA]_filtrada.fastq
```

Por que usar `<` no comando `chopper`?

O comando `<` garante que o arquivo `fastq` da amostra seja direcionado para entrada (**stdin**) do comando no `chopper`. Em nossas experiências já tivemos a necessidade de utilizar algumas vezes e outras não.

4. Alinhamento das Sequências:

- Use o `minimap2` para alinhar suas sequências de DNA ao genoma de referência.

```
minimap2 -ax map-ont [REFERENCE_GENOME].fa [AMOSTRA].fastq /
-R "@RG\tID:{SAMPLE}\tSM:{SAMPLE}" -t [THREADS] > [AMOSTRA].sam
```

Consumo de memória durante a etapa

O alinhamento via `minimap2` tem pico de consumo de memória de ~ 13 GB.

- Realize novoamento utilizando `bwa mem` desta vez (Opicional);

⚠ Tempo de execução do `bwa`

Atenção, o `bwa` demora quase 10x mais que o `minimap2`

```
bwa mem -x ont2d [REFERENCE_GENOME].fa [AMOSTRA].fastq /
-R "@RG\tID:{SAMPLE}\tSM:{SAMPLE}" -t [THREADS] > [AMOSTRA].sam
```

5. Gerar o arquivo BAM indexado:

- Repita os passos 3 e 4 do dia 2 utilizando `samtools`;

6. Visualização do Alinhamento:

- Importe o BAM para o IGV e avalie sua qualidade;
- Compare as amostras com suas respectivas amostras de short reads;

7. Realize a Genotipagem das Amostras de long reads:

- Utilize o `freebayes` para realizar a genotipagem e modo similar ao passo 3 do dia 3;

⚠ Tempo de execução da genotipagem em Long Reads

O tempo de execução do `freebayes` para arquivos de ONT pode ser bem demorado! Você pode optar por modificar o arquivo `bed` fornecido para trabalhar com menos regiões.

Para os formulários utilize o `vcf` resultante em **WorkshopDados/pre_run/Dia4/**

- Repita as métricas de filtragem utilizadas no passo 4 do dia 3;
- Repita os passos 5 e 6 do dia 3, comparando aos resultados de short reads.

8. Análise de modificações de bases

- Use `modkit` para avaliar as modificações de base.

```
modkit pileup {sample}.bam {sample}_pileup.bed --log-filepath {sample}.log
--ref path/to/reference.fasta
--preset traditional
```

- Avalie as modificações no IGV.

Atividades da tarde

1. Avalie a qualidade das leituras das amostras com **NanoPlot**. Qual a qualidade média das sequencias?
2. Quais o intervalo de tamanhos das leituras de cada amostra?
3. Compare a genotipagem dos genes MC1R e SLC45A2 obtidas com ONT com as obtidas com short reads.
4. s
5. Avalie as modificações 5mC e 5hmC nos genes MC1R e SLC45A2.

Dia 5 - Marcadores STRs

Para esse curso, utilizaremos um computador virtual, devido à incompatibilidades com a nova versão do compilador C++. Portanto, teremos que baixar o computador virtual com os arquivos necessários para o curso, e virtual box que permitirá executar o computador virtual.

PRIMER PASO - DESCARGAS

1. Baixar o computador virtual Ubuntu (baixar e descomprimir o arquivo zip).

Link para baixar o computador virtual: https://drive.google.com/file/d/1Pnw3KGixyB6ur2h9zDA3zSAeHj8Oz/view?usp=drive_link

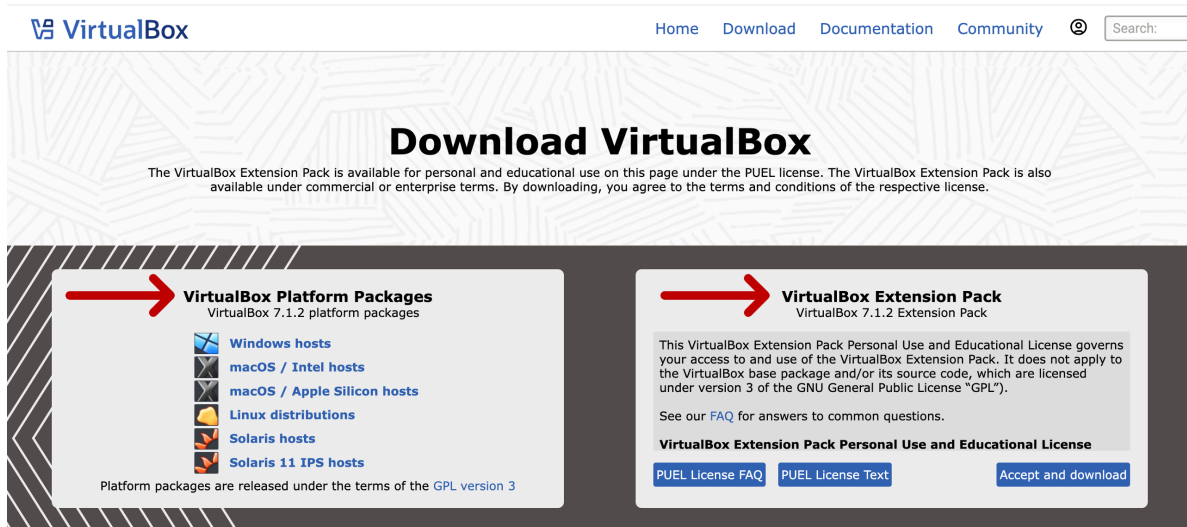
1. Baixar virtual box e o pacote de instalação.

Link para baixar virtual box que mostrará o computador virtual Linux: <https://www.virtualbox.org/wiki/Downloads>

Baixar DUAS coisas deste link virtual box:

1. **VirtualBox Platform Packages**
2. **VirtualBox Extension Pack**

Escolher o sistema operativo do computador (windows, mac ou linux) e baixar.



Logo de instalar, escolher o computador virtual Ubuntu que baixamos no primer link.

Senha do computador virtual instalado: bioinformatica

Link tutorial de ajuda: https://drive.google.com/file/d/1bPgrcmzdD3b_ELsGPKUMBllZ7G0SJR18/view?usp=drive_link

HIPSTR

Haplotype inference and **p**hasing for **S**hort **T**andem **R**epeats ou HipSTR é um programa que foi desenvolvido por Thomas Willems para capturar marcadores STRs a partir de dados de sequenciamento de nova geração.

Link do Github:

<https://github.com/tfwillems/HipSTR>

INSTALAÇÃO

Infelizmente a última versão do compilador C++ apresenta uma incompatibilidade com HipSTR. Portanto, iremos utilizar um computador virtual que já contem o programa instalado.

O computador virtual tem uma pasta chamada “curso” no desktop, onde se encontra:

- O genoma de referência (`arquivo.fa`)
- O índice do genoma de referência (`arquivo.fai`)

- Amostras em formato bam (HG00118.bam, etc)
- O arquivo bed com as regiões que serão analisadas.

Com isso, iremos correr o comando do HipSTR para genotipar essas amostras

CÓMO CRIAR OS ÍNDICES DAS AMOSTRAS?

- Utilizaremos `samtools index`

```
cd rota_da_pasta
samtools index nome_da_amostra.bam
```

Fazer esse comando para todas as amostras.

Link tutorial de ajuda: https://drive.google.com/file/d/17vv41Hf0ezyEHZAHne1e4iPbzCemDgXN/view?usp=drive_link

COMANDOS

Arquivos necessários

```
./HipSTR --bams      run1.bam, run2.bam, run3.bam, run4.bam
          --fasta     genome.fa
          --regions   str_regions.bed
          --str-vcf    str_calls.vcf.gz
```

! IMPORTANTE

- Antes de correr o comando do HipSTR, verifiquem que todas as amostras que serão incluídas no comando, já tenham o seu índice (Arquivo fai, gerado com samtools index)
- Preparar o comando previamente em algum arquivo de texto, substituindo `nome_da_amostra1` pelo nome real da amostra.
- Podem rodar varias amostras no mesmo comando, separando elas por virgulas SEM espaço, como no exemplo embaixo.

```
cd rota_da_pasta
--bams nome_da_amostra1.bam,nome_da_amostra2.bam --fasta GRCh38_full_analysis_set_plus_decoy
```

Link de ajuda: https://drive.google.com/file/d/1Vzw8K7-Jh-ud7ul_0bNhKpJ4gw0zxqvo/view?usp=drive_link

INTERPRETAÇÃO DO ARQUIVO DE SAÍDA

HipSTR gera como arquivo de saída um arquivo vcf.gz. Logo de descomprimir o arquivo teremos um vcf.

O vcf podemos abrir com Excel e realizar a interpretação dos dados para genotipar as amostras.

Durante a aula aprenderemos a realizar essa interpretação.

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>.
- 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>.
- Willems, Thomas, Dina Zielinski, Jie Yuan, Assaf Gordon, Melissa Gymrek, and Yaniv Erlich. 2017. “Genome-Wide Profiling of Heritable and de Novo STR Variations.” *Nature Methods* 14 (6): 590–92. <https://doi.org/10.1038/nmeth.4267>.

ANEXO: Dicas para uso do Ubuntu/WSL

Navegação e Diretórios

Listar Conteúdo do Diretório

```
ls
```

Mudar de Diretório

```
cd nome_do_diretorio
```

Diretório Atual

```
pwd
```

Criar Diretório

```
mkdir nome_do_novo_diretorio
```

Manipulação de Arquivos

Copiar Arquivo

```
cp arquivo_origem destino
```


Mover/Renomear Arquivo

```
mv arquivo_origem novo_nome_ou_destino
```

Remover Arquivo

```
rm nome_do_arquivo
```

Visualização de Conteúdo

Visualizar Conteúdo do Arquivo

```
cat nome_do_arquivo
```

Visualizar Conteúdo do Arquivo (página por página)

```
less nome_do_arquivo
```

Pacotes e Atualizações

Atualizar Lista de Pacotes

```
sudo apt update
```

Atualizar Pacotes Instalados

```
sudo apt upgrade
```

Instalar Novo Pacote

```
sudo apt install nome_do_pacote
```

Gerenciamento de Usuários

Adicionar Usuário ao Grupo

```
sudo usermod -aG nome_do_grupo nome_do_usuario
```

Mudar Senha do Usuário

```
passwd nome_do_usuario
```

Processos

Listar Processos

```
ps aux
```

Matar um Processo por ID

```
kill -9 processo_id
```

Monitorar Recursos do Sistema (htop)

```
htop
```

Medir Tempo de Execução de um Comando (time)

```
time comando_a_ser_medido
```

Rede

Verificar Configurações de Rede

```
ifconfig
```

Testar Conexão com um Endereço IP

```
ping endereco_ip
```

Entrada e Saída Padrão (stdin/stdout)

- **stdin (Standard Input):** É a entrada padrão de dados. Um programa pode ler dados a partir do stdin. Exemplo:

```
cat < nome_do_arquivo
```

- **stdout (Standard Output):** É a saída padrão de dados. Um programa geralmente imprime resultados no stdout. Exemplo:

```
ls > lista_de_arquivos.txt
```

- **stderr (Standard Error):** É a saída padrão para mensagens de erro. Exemplo:

```
comando_inexistente 2> erro.log
```

Outros Comandos Úteis

Ajuda sobre um Comando

```
man nome_do_comando
```

Sair do Terminal

```
exit
```