Tutorial Bioinformatica

Marcel Ferreira - Bolsista/CAPES

2024-10-12

Table of contents

Sobre esse curso

Realização

Figure 1: Realização UnB, UNESP e USP.

Apoio

Figure 2: CAPES-PROCAD Edital n° 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

João Victor Hammamura Toloi

Matheus de Souza Ferrari

Pedro Mendes Laprega

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

! Confira os arquivos baixados

Ao realizar o download, confira se os arquivos foram baixados corretamente.

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 informção da posição baseda no cromossomo (**coluna 1**) e seu inicio (**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.

Figure 3: Estrutura de um arquivo BED.

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

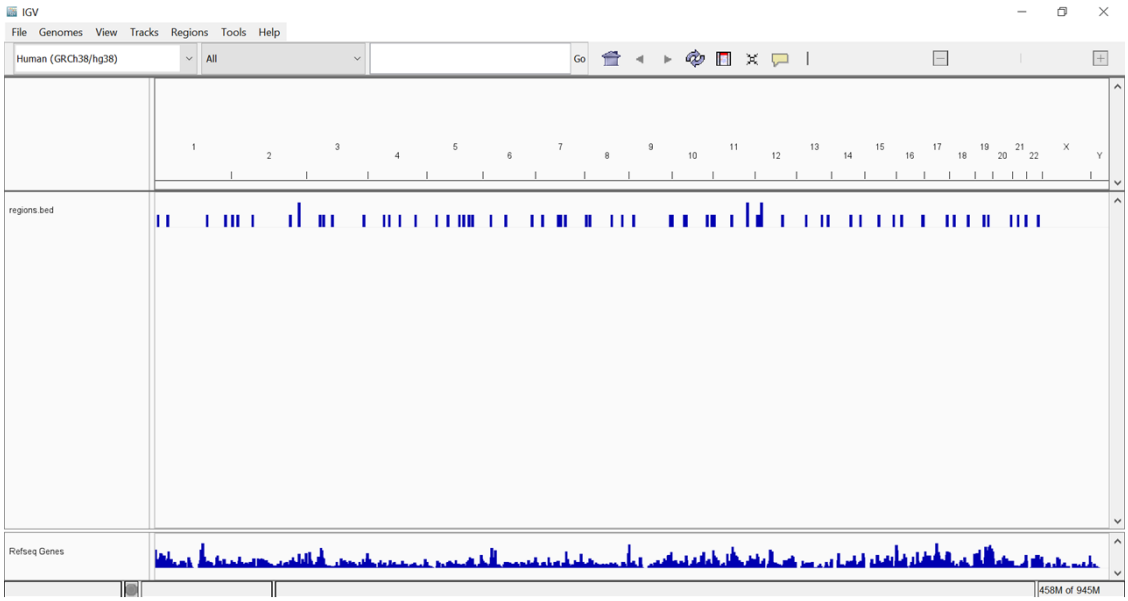

Figure 4: Abrindo o arquivo regions.bed no IGV

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

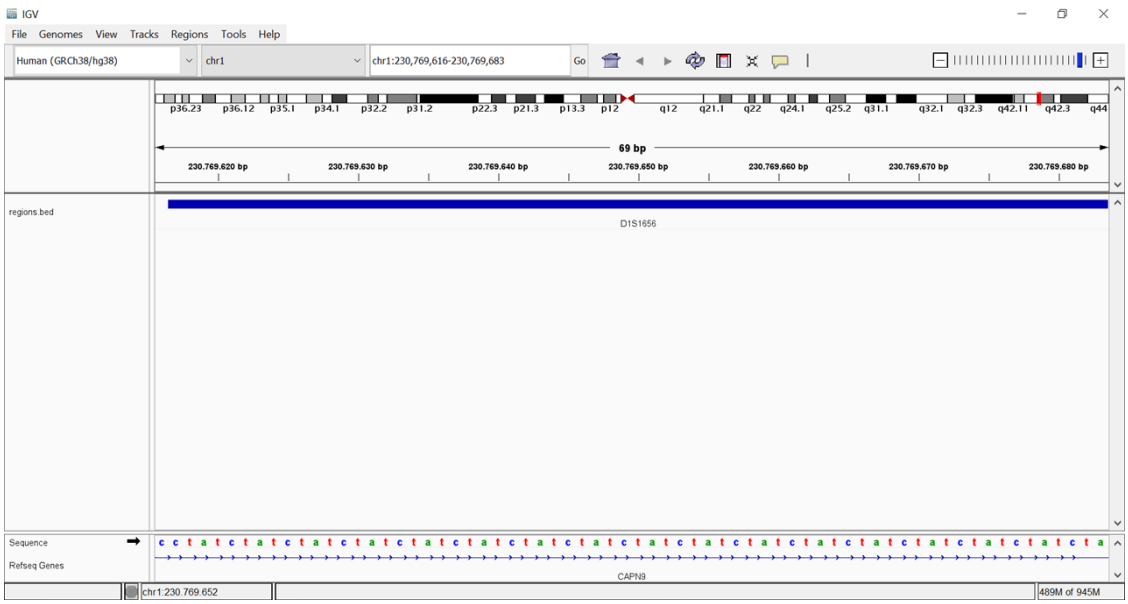

Figure 5: Região do D1S1656.

Amostra	Sexo	População
HG00097	F	British
HG00142	M	British
HG00143	M	British
HG00145	M	British
HG00263	F	British
HG00277	M	Finnish
HG00372	M	Finnish
HG00463	M	Han Chinese
HG01063	M	Puerto Rican

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

WorkshopBioinfo2024/

Genome/

R10/

fastq/

bam/

Exome_ShortReads/

fastq/

bam/

WGS_ShortReads/

fastq/

bam/

SimulatedReads/

Dia 1 - Sequenciamento de DNA

> **!** Importante
>
> Verifique se o `FASTQC` esta instalado.

Arquivos

Serão utilizados os arquivos contidos na pastas:

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

Estes 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.fast
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.fast
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

1. No relatório do FASTQC, o que significa uma bandeira vermelha (*Fail*) no gráfico de '*Per base sequence content*'?
2. Qual métrica do FASTQC pode indicar a presença de contaminantes ou adaptadores remanescentes nas leituras?
3. Quando o gráfico de '*Per base sequence quality*' mostra uma queda significativa nas últimas bases de uma leitura, o que isso pode indicar?
4. O gráfico de '*Per sequence GC content*' avalia a distribuição do conteúdo GC nas leituras. O que um pico inesperado nesse gráfico pode sugerir?

5. Execute `FASTQC` para as leituras simuladas (**`WorkshopBioinfo2024/SimulatedReads`**). Quais as principais diferenças entre as amostras?
6. O que acontece se você realizar a trimagem utilizando `trimmomatic` com os paramêtros 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 analises 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 indice do genoma via **`bwa index`** demora bastante tempo para ser realizado. Mesmo em máquimas com grandes capacidades de memória. Devido a isso colocamos na pasta **`Genome/index`** os arquivo resultantes desta etapa, que são os arquivos com extensão `.amb`, `.ann`, `.bwt`, `.fai`, `.pac` e `.sa`.
   > O genoma de referência humano esta 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.fas
```

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

> **Automatizando o processo (Opicional)**
>
> 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 própio `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 dos seguintes programas é amplamente utilizado para alinhar sequências curtas de DNA em genomas de referência?
2. Qual formato de arquivo não é gerado por `bwa index`.
3. Qual é o principal formato de saída gerado por ferramentas como BWA, Bowtie e STAR após o alinhamento de sequências de DNA?
4. O que o programa SAMtools realiza no fluxo de análise de dados de alinhamento de DNA?
5. Qual é a profundidade média de cobertura das leituras alinhadas para o gene MC1R?

   > Dica
   >
   > Use o comando a baixo para amostras Exoma e WGS:
   > `samtools coverage -b bam_list.txt -r {chr}:{start}-{end} > cov.txt`

6. Qual o número de leituras alinhadas em para os genes MC1R e SLC45A2 de exoma da amostra HG00372?

   > Dica
   >
   > Use o comando abaixo e olhe no arquivo resultante na linha 10 “SN sequences: VALOR”.
   > `samtools stats HG00372.bam {chr}:{start}{end} > {stat}.txt`

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 da 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 analise 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

bcftools view --min-ac 1 variantes_filtradas_trim.vcf > variantes_filtradas_trim_mi

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

   - 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. Olhando no VCF do Exome, no gene “SLC45A2”, na posição “33984321” (chr5), qual é a qualidade no IGV?
2. Qual o genótipo da amostra HG00463 no SNP localizado na posição “89914679” (chr16) (WGS)? O que se conclui com esse tipo de genótipo em questão de ref. e alt.?
3. Qual a qualidade do SNP que está na posição “89914789” (chr16) para Exome e WGS?

4. Sobre a qualidade do SNP da questão anterior, qual a sua interpretação do resultado?
5. Na posição "89914936" de WGS (chr16) (Excel), qual o tipo de polimorfismo observado no IGV? Qual o genótipo das amostras HG00142, HG00145 e HG00263?
6. No IGV, na posição "89914473" (chr16), qual a cobertura? (WGS)

Dia 4 - Análise de sequenciamento Oxford Nanopore

Atividades da manhã

1. **Chamada de bases via Dorado:**
   - Qual o modelo deve ser usado?
   - Realize a chamada de base para todos os arquivos `pod5` contidos na pasta **WorkshopDados/pod5/**;

```
dorado basecaller pod5/dna_r10.4.1_e8.2_260bps_fast@v4.1.0 pod5/ --modified-bases 5m
```

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 novamento 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/R10/vcf/variantes.vcf**

   - 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 pat
     ```

   - Avalie as modificações no IGV.

Atividades da tarde

1. Qual dos seguintes programas é especializado no alinhamento de leituras longas, como as geradas por sequenciadores Oxford Nanopore e PacBio?
2. Avalie a qualidade das leituras das amostras com `NanoPlot`. Qual das 9 amostras apresentou uma qualidade média mais baixa?
3. Sobre o intervaldo de tamanhos das leituras de cada amostra, qual apresenta a maior média de tamanho?
4. Para a amostra HG00372, corte a amostra para qualidades acima de 15 e tamanho mínimo de 2000 bases. Qual o número de bases original e após corte?
5. Compare a genotipagem dos genes MC1R e SLC45A2 obtidas com ONT com as obtidas com short reads.
6. Avalie as modificações 5mC e 5hmC na região promotora dos genes MC1R e SLC45A2.

Dia 5 - Marcadores STRs

HipSTR

Haplotype**i**nference 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

Atualmente, HipSTR apresenta uma incompatibilidade com a nova versão do compilador do C++. Portanto, iremos utilizar um repositório transitório que já contém uma mudança no código fonte, até o Dr. Thomas Willems modificar diretamente no repositório original.

Link do Github para o repositório alternativo:

https://github.com/Tfronta/HipSTR2

REQUERIMENTOS

Para compilar HipSTR, são necessários os seguintes pacotes:

- make
- g++
- zlib
- libhts
- libbz2
- liblzma

Se vocês estiverem executando o Ubuntu 16+, os pacotes podem ser facilmente instalados executando:

```
sudo apt install make g++ zlib1g-dev libhts-dev libbz2-dev liblzma-dev
```

INSTALAÇÃO DO HIPSTR

HipSTR requer um compilador c++ padrão. Para obter HipSTR, use:

```
git clone https://github.com/Tfronta/HipSTR2 HipSTR
```

Para construir, usem Make:

```
cd HipSTR
make
```

Para que HipSTR esteja disponível no computador, sem necessidade de copiar a rota donde ele está instalado, por exemplo "downloads", criaremos un symlink. Um **symlink** (ou link simbólico) é um "atalho" que aponta para outro arquivo ou diretório no sistema. Isso permite acessar um arquivo de diferentes locais como se estivesse em várias pastas, mas ele só existe em uma.

```
sudo ln -s $(pwd)/HipSTR /usr/local/bin/HipSTR
```

Para saber se o HipSTR foi instalado corretamente ou solicitar ajuda use:

```
HipSTR --help
```

Tutorial de ajuda

GENOTIPAGEM

ARQUIVOS NECESSÁRIOS

- O genoma de referência HG38 (arquivo.fa)
- O índice do genoma de referência HG38 (arquivo.fai)
- Amostras em formato bam (HG00118.bam, etc)
- O índice das amostras em formato bai (HG00118.bai, etc)
- O arquivo bed com as regiões que serão analizadas em este curso

Estes arquivos se encontram dentro de uma pasta de google drive. O genoma de referência já possui o índice, por questões de tempo. Entretanto, vamos a criar os índices para as amostras

Acesso aos arquivos

! IMPORTANTE

- Deverão baixar todos estes arquivos e coloca-los dentro DA MESMA PASTA
- Se todos estes arquivos não se encontrarem dentro da MESMA pasta, o comando não funcionará.

RECOMENDAÇÕES

- Não usar espaços ou maiúsculas no nome da pasta criada. Exemplos sugeridos: curso, curso_bioinfo, cursobioinfo, etc.

COMO CRIAR OS ÍNDICES DAS AMOSTRAS?

- Utilizaremos samtools index

```
cd rota_da_pasta_criada
samtools index nome_da_amostra.bam
```

Fazer esse comando para todas as amostras.

Tutorial de ajuda

COMANDO DO HIPSTR

- Antes de correr o comando do HipSTR, verifiquem que todas as amostras que serão incluidas no comando, já tenham o 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
HipSTR --bams nome_da_amostra1.bam,nome_da_amostra2.bam --fasta GRCh38_full_analysis_set_plus
```

Tutoria de ajuda

INTERPRETAÇÃO DO ARQUIVO DE SAÍDA

- HipSTR gera como arquivo de saída um arquivo vcf.gz. Logo de descomprimir o arquivo teremos um vcf.
- Podem abrir com Excel e realizar a interpretação dos dados para genotipar as amostras. Durante a aula aprenderemos a realizar essa interpretação.

Atividades da manhã

- Extrair os valores GB e DP de todas as amostras
- Genotipar todas as amostras, incluir valores de cobertura

Atividades da tarde

1. Quais são os genótipos do marcador CSF1PO na amostra HG00239?
2. Quais são os genótipos do marcador TPOX na amostra NA18602?
3. Qual é o comando utilizado para criar os índices das amostras?
4. O valor DP indica a diferença de número de bases com o alelo de referênça?
5. Alguma das amostras não foi possível genotipar?

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
```