Tutorial Bioinformatica

Marcel Ferreira

2023-11-20

Table of contents

Sobre esse tutorial

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 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!

1.1 Primeiros passos

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

Referências