Neoprospecta - Desafio 4

Participante: Thayana Vieira Tavares

- Os reads provenientes de um sequenciamento podem não representar fielmente a amostra coletada, isto pode acontecer por diversos motivos: contaminação na bancada, adição de adaptadores, erro do sequenciador, entre outros. Visando diminuir esses erros para que seja possível utilizar os arquivos em análises posteriores, faz-se o processo de trimagem e controle de qualidade dos reads.

- Para gerar um relatório inicial, foi utilizado o software **FastQC** <*disponível em: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/*>.

  No diretório *'fqs'*, onde estão todos os arquivos fastqs, o seguinte comando foi digitado no terminal linux:

  ```
  fastqc *.fastq -o ../resultados/.
  ```

  Assim, o software conseguiu analisar todos os arquivos .fastq presentes no diretório e retornou arquivos zip e html contendo as análises.

- Visando gerar um relatório mais compacto, utilizou-se o software **MultiQC** <*disponível em: https://multiqc.info/*> para gerar um único relatório com todos os resultados analisados pelo FastQC.

  O seguinte comando foi digitado no linux, estando no diretório *'resultados'*, onde estão todos os arquivos gerados pelo fastQC:

  ```
  multiqc .
  ```

  Assim, o software retorna um arquivo html com todas as análises do fastQC em um único local.

- É possível observar no relatório que a qualidade de alguns reads não está tão boa, por isso foi realizado o processo de trimagem com o Trimmomatic <*disponível em: http://www.usadellab.org/cms/?page=trimmomatic*> , com o seguinte comando, no diretório *'fqs'*:

  ```
  for file in *.fastq; do trimmomatic SE $file ../trimados/$file HEADCROP:20
  SLIDINGWINDOW:5:28; done
  ```

  - HEADCROP:20 foi utilizado para cortar as 20 primeiras bases de todos os reads, visando retirar adaptadores ou primers.
  - SLIDINGWINDOW:5:28 foi utilizado para selecionar as bases com qualidade acima de 28 em uma janela de 5 em 5, visando eliminar bases de baixa qualidade.

- Por fim, o relatório final foi gerado utilizando os novos arquivos fastq trimados. No diretório '*trimados*', com o comando:

```
fastqc *.fastq -o ../resultados_trimados/.
```

Os relatórios gerados com fastQC mostram informações importantes acerca dos reads:

1. **Basic Statistics**: contém um resumo, onde se pode encontrar o nome do arquivo, o total de reads, o tamanho do menor read e do maior, conteúdo GC, entre outros;
2. **Per base sequence quality:** apresenta a média das qualidades das bases (em boxplots) nos intervalos indicados no eixo x. É apresentado um gráfico com 3 faixas horizontais: vermelha (qualidade péssima), amarela (qualidade intermediária) e verde (qualidade ótima);
3. **Per tile sequence quality:** essa seção indica a qualidade do sequenciamento por Tile dentro do sequenciador. A boa qualidade é indicada com a cor azul escuro, cores diferentes podem indicar uma má qualidade;
4. **Per sequence quality scores:** exibe um gráfico que indica a qualidade média por reads no x, essa qualidade é medida pelo Phred score. Já em y, estão as frequências que essas qualidades ocorrem;
5. **Per base sequence content:** mostra o percentual das bases nucleotídicas T, C, A, G encontrado nas diferentes posições da sequência. Teoricamente, o esperado é que esse percentual seja uniforme, porém em muitos casos não é o que acontece, devido a sequências repetitivas, má qualidade dos reads, adaptadores, entre outros;
6. **Per sequence GC content:** exibe em azul, a distribuição normal do conteúdo GC que seria esperado, já em vermelho, indica a distribuição observada por reads;
7. **Per base N content:** o gráfico mostra a quantidade de bases indicadas como N (não distinguidas) ao longo das posições dos reads;
8. **Sequence Length Distribution:** o gráfico desta seção indica a distribuição dos tamanhos de reads observados na amostra. No eixo x, encontra-se a escala de tamanho e no y a frequência com que ocorre.
9. **Sequence Duplication Levels:** mostra a quantidade de sequências duplicadas encontradas nos reads. Em azul, é indicado a porcentagem total de sequências, enquanto a linha vermelha mostra a porcentagem das sequências duplicadas.
10. **Overrepresented sequences:** exibe as sequências mais encontradas nos reads, suas quantidades e porcentagens;
11. **Adapter Content:** mostra quais adaptadores illumina foram encontrados.