Instituto Federal de Brasília

Campus Taguatinga

Superior em Computação

Filtragem de sequências

1/2020

Danyelle da Silva Oliveira Angelo

1. Avalie o conjunto das reads antes da filtragem:
    a. De acordo com as informações disponíveis no banco SRA, qual sequenciador foi utilizado para se obter as reads? As reads são single-end ou paired-end?
    Resposta: Layout -> single-end, sequenciador-> 454 GS FLX

    b. Execute o programa FastQC sobre o arquivo .fastq (já realizado pelo professor) e analise a saída .html do programa (Saida_fastqcSRR060738_fastqc.zip):
        i. Tire uma print da seção basic statistics.
        Resposta:

Basic Statistics

Measure	Value
Filename	SRR060738.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	191638
Sequences flagged as poor quality	0
Sequence length	125-1891
%GC	48

        ii. Qual a variação do tamanho das reads?
        Resposta: Varia entre 125 e 1891

iii. Quais testes o arquivo .fastq não passou (símbolo de cruz vermelha)? Para cada teste tire um print do gráfico gerado pela ferramenta e justifique o motivo de não ter passado.

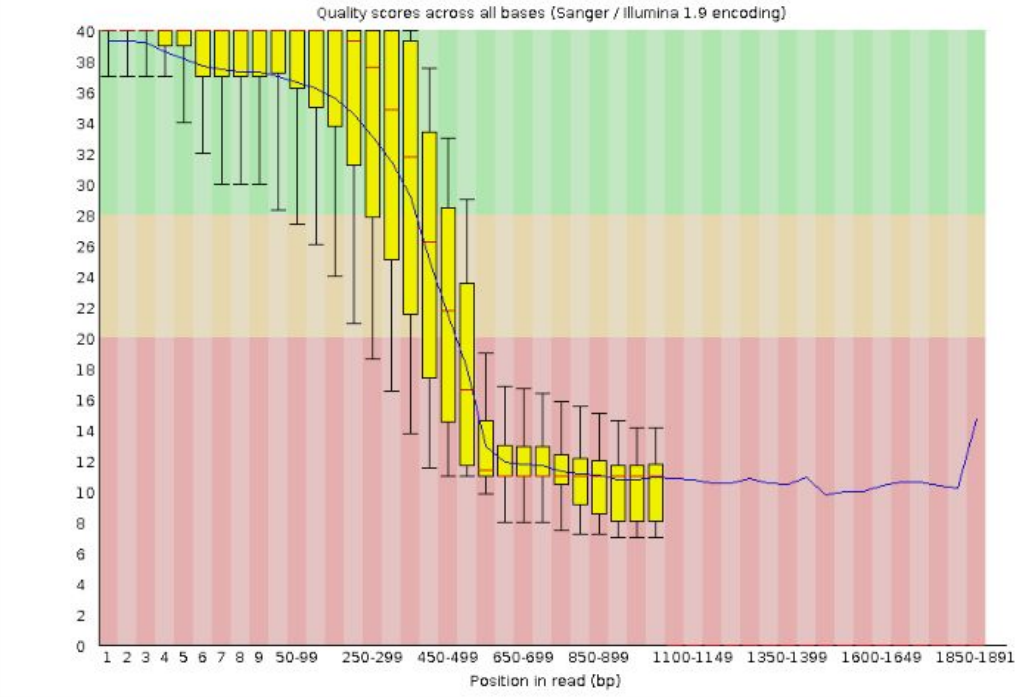
Resposta:

1. Per Base Sequence quality

   Primeiro é preciso entender o que nos diz o gráfico abaixo:

   - qualidade dos escores por regiões: em verde (boa qualidade), em amarelo (qualidade razoável), e em vermelho (baixa qualidade).
   - as barrinhas em amarelo representam o intervalo interquartil (25-75%),
   - os bigodes superior e inferior representam os pontos de 10% e 90%,
   - os traços vermelhos nos dão a mediana dos intervalos interquartis,
   - a linha em azul nos mostra a qualidade média (eixo y), quanto maior a pontuação, melhor a leitura da base,

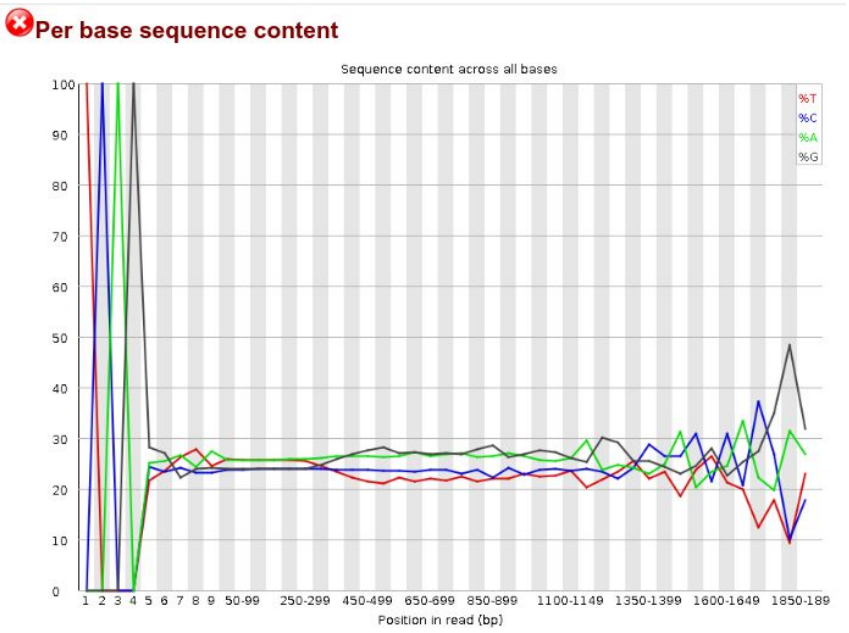
   Dito isso, o FastQC retorna FALHA se a mediana para qualquer base for menor que 20, que é o que vemos mais a direita do nosso gráfico, por esse motivo o arquivo não passou no teste.



2. Per Base Sequence Content

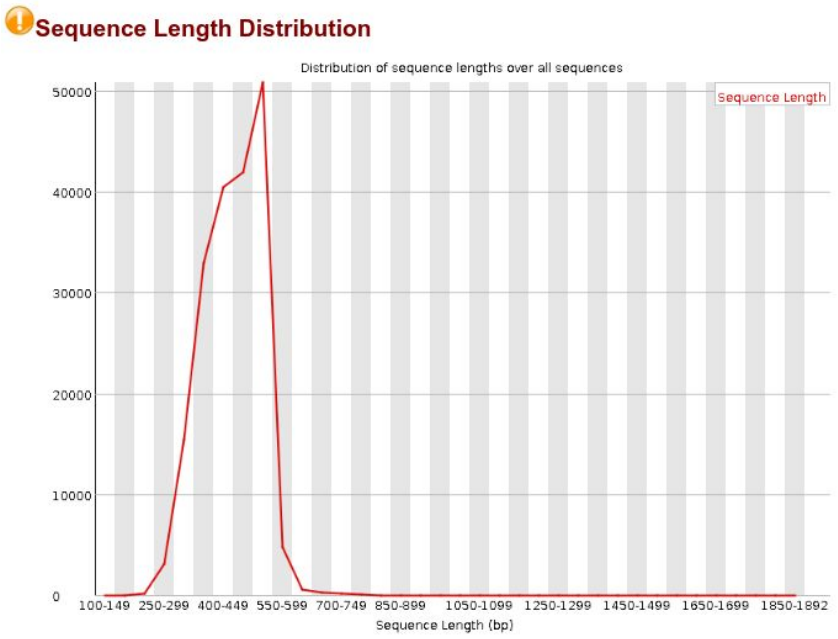
   Este método mostra a frequência de cada um dos tipo de base encontrado nas diferentes posições da sequência. No gráfico abaixo, o eixo X indica a posição dentro do read e o eixo do Y a frequência de A,C,T,G.

   Aqui é esperado que a diferença entre as frequências de cada base seja pequena (*se ela for maior que 20% em uma posição o FastQC já indica um erro*),mas não é o que acontece no gráfico abaixo, por exemplo, *temos uma diferença de cerca de de 40% entre a timina e guanina na posição 1850.*



iv. A maior parte das reads possui aproximadamente que tamanho? justifique sua resposta usando algum teste realizado pelo FastQC.

Resposta: O módulo Sequence Length Distribution mostra a distribuição de tamanho de reads observada na amostra. O mesmo retorna um aviso de erro caso encontre um read de tamanho 0, e emite um Warning quando encontra sequências de tamanho variado (que é o nosso caso). Encontraremos a resposta para esta pergunta olhando para o pico mais alto do eixo Y, que nos diz que a *maioria das sequências possui um tamanho aproximado a 550.*



v. Analisando a saída do FastQC, as reads são de qualidade ou requerem atenção? Caso precisem de atenção, o que fazer?

Resposta: Elas são de baixa qualidade, portanto requerem atenção. Pode-se: remover os nucleotídeos de baixa qualidade (trimming) ou melhorar a qualidade das reads sobrepondo os padrões de baixa frequência pelos de alta frequência.

2. Usaremos o programa Prinseq para filtrar reads de baixa qualidade:
   a. Execute **(o professor já executou :-))** o programa Prinseq da seguinte forma:

```
perl prinseq-lite.pl -fastq arquivo.fastq -out_format 3
-trim_qual_right 20 -trim_qual_window 50
-trim_qual_step 25 -trim_qual_type mean
-min_qual_mean 20 -min_len 50
```

Onde arquivo.fastq é o arquivo .fastq que obtemos. pesquise o que significa cada um dos parâmetros usados acima.

Resposta:

Parâmetro	Significado
-fastq	indica que o arquivo a ser lido (arquivo.fastq) está no formato fastq
-out_format 3	indica o formato de saída, nesse caso é fastq (use 1 para fasta, 2 para fasta e qual, 4 para fastq e fasta e 5 para fastq, fasta e qual)

-trim_qual_right 20	elimina as sequências com escore de qualidade inferior a 20.
-trim_qual_window 50	tamanho da janela (50) utilizada para calcular o índice de qualidade
-trim_qual_step 25	tamanho do passo (25) utilizado para mover a janela
-trim_qual_type mean	tipo de método (mean) para o cálculo de qualidade
-min_qual_mean 20	sequência de filtro com média de pontuação abaixo de 20
-min_len 50	elimina as sequências com menos de 50 nucleotídeos

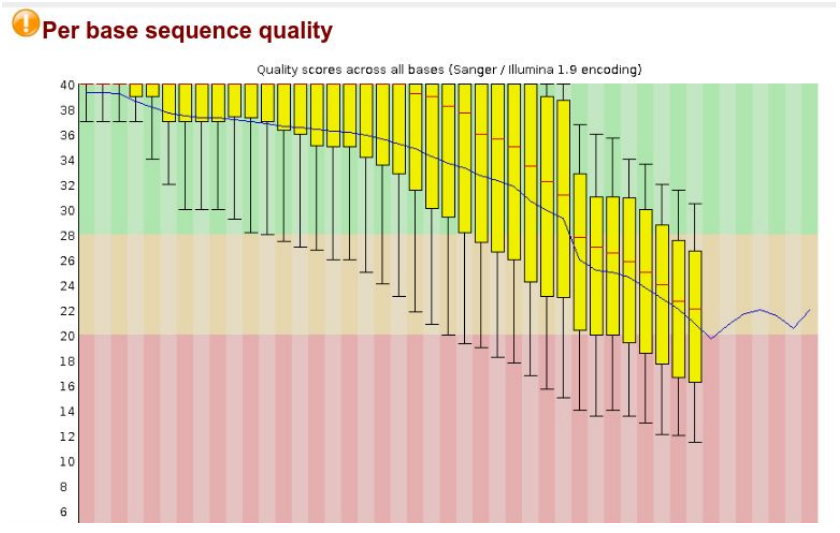
b. De acordo com a saída padrão (terminal) do programa, quantas sequências e bases foram classificadas como “boas” e quantas foram classificadas como “ruins”?

Resposta: Boas = 191.518, ruins = 120

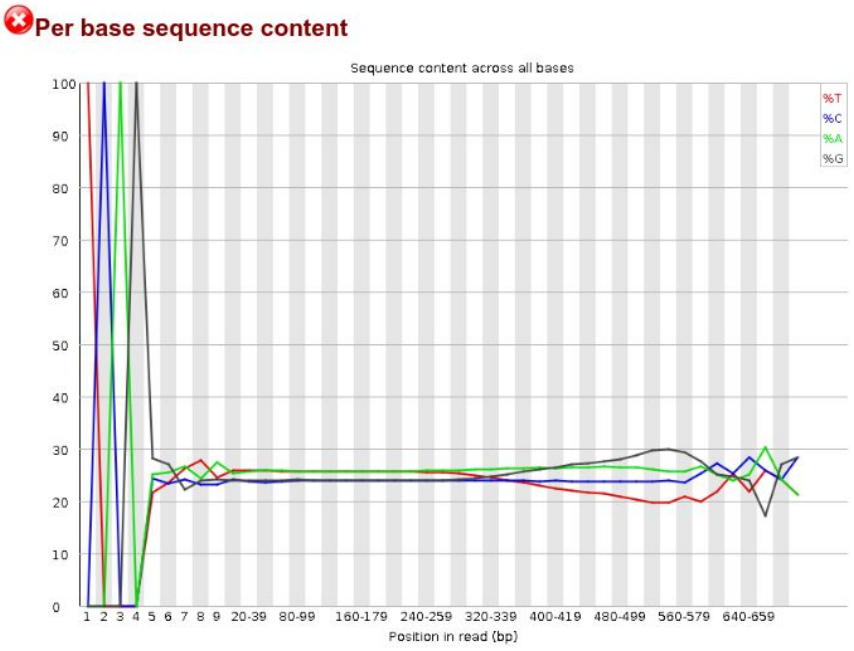
c. Utilize o programa FastQC usando as reads de boa qualidade (O prinseq retorna dois conjuntos de reads, as de qualidade boa e as ruins). Houve melhoria? Justifique sua resposta apresentando os gráficos gerados pelos testes que tiveram sua classificação alterada (Arquivo de saída: prinseq_fastqc.zip).

Resposta: Sim, houve melhorias. Veja os gráficos abaixo,

1. Per Base Sequence quality: veja que dessa vez a mediana para todas as bases é maior do que 20, para algumas ela é menor do que 25, por isso o FastQC retorna apenas uma mensagem de aviso.



2. Per Base Sequence Content: verifique que aqui a diferença entre os nucleotídeos também diminuiu.



d. Houve algum teste que permaneceu com classificação negativa? Quais foram?

Resposta: Teve o “Per base sequence content” apesar da qualidade dele ter melhorado ainda existe uma diferença entre as frequências de cada base em cada posição muito grande.

3. De acordo com os resultados, qual decisão você tomaria: prosseguir para a etapa de montagem, ou solicitar um novo sequenciamento? Justifique sua resposta.

Resposta: Depois da separação com o printseq eu solicitaria um novo sequenciamento apenas para tentar diminuir a diferença entre os nucleotídeos das posições 1 a 4.

Referências:

- VERIFICAÇÃO de qualidade FastQ. **Blast2Go**. Disponível em: <http://docs.blast2go.com/user-manual/tools-(pro-feature)/fastq-quality-check/#FASTQQualityCheck-Results>. Acesso em: 13 de outubro de 2020.
- REIS, Clovis F. Testando a qualidade de um sequenciamento. **Duna Bioinformática**. Disponível em: <http://dunabioinfo.com/pt/blog/mc01/testeQual.pdf>. Acesso em: 13 de outubro de 2020.