

## **Ferramentas em Bioinformática 4** **Filtragem de Sequências**

Introdução à Bioinformática 1/2020  
Professor João Victor de Araujo Oliveira

Trabalharemos com dados públicos do genoma da bactéria *Escherichia coli* disponíveis no SRA (Sequence Read Archive: <https://www.ncbi.nlm.nih.gov/sra/SRX023780>). Por questões de tempo e flexibilização devido à pandemia, não será necessário baixar este conjunto de reads. Além disso o professor irá fornecer a saída dos programas utilizados durante a atividade.

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*?
- b) Execute o programa FastQC sobre o arquivo .fastq (**já realizado pelo professor**) e analise a saída .html do programa ([Saída\\_fastqcSRR060738\\_fastqc.zip](#)):
  - i) Tire uma print da seção *basic statistics*.
  - ii) Qual a variação do tamanho das *reads*?
  - 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.
  - iv) A maior parte das *reads* possui aproximadamente que tamanho? justifique sua resposta usando algum teste realizado pelo FastQC.

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

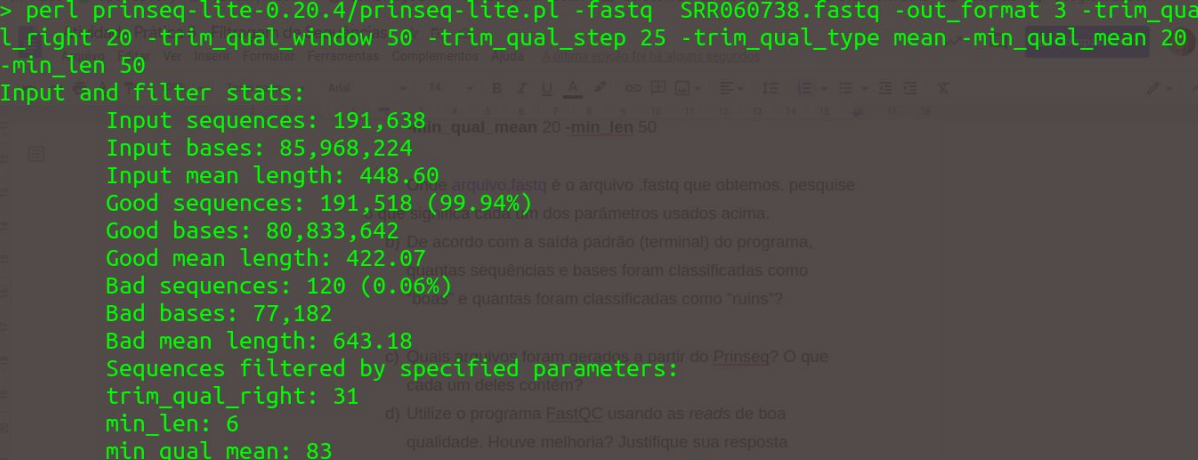
2) Usaremos o programa *Prinseq* para filtrar *reads* de baixa qualidade:

- 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.

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



```
> perl prinseq-lite-0.20.4/prinseq-lite.pl -fastq SRR060738.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  
Input and filter stats:  
Input sequences: 191,638  
Input bases: 85,968,224  
Input mean length: 448.60  
Good sequences: 191,518 (99.94%)  
Good bases: 80,833,642  
Good mean length: 422.07  
Bad sequences: 120 (0.06%)  
Bad bases: 77,182  
Bad mean length: 643.18  
Sequences filtered by specified parameters:  
trim_qual_right: 31  
min_len: 6  
min_qual_mean: 83
```

- 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](#)).

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

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