Avaliação e processamento de dados brutos de sequenciamento de genomas e transcriptomas

Desirrê Petters-Vandresen

Módulo I – Genômica no Estudo de Microrganismos

Por que devemos avaliar os dados brutos com atenção?

- Muitas perguntas são respondidas e muitas hipóteses são testadas com base na informação existente nas sequências de um genoma ou transcriptoma

- Diversas ferramentas de anotação utilizam características presentes nas sequências para realizar predições

- Sequências erradas podem ter um grande impacto negativo nos resultados e conclusões de um estudo

Por que devemos avaliar os dados brutos com atenção?

- Muitas perguntas são respondidas e muitas hipóteses são testadas com base na informação existente nas sequências de um genoma ou transcriptoma

- [illegible]

- [illegible] vo nos resultados e conclusões de um estudo

Montar todos os reads brutos que saem do equipamento em um genoma ou transcriptoma sem qualquer tipo de controle de qualidade não é uma boa ideia!

O que devemos avaliar nos dados brutos antes de prosseguir com montagens?

- Qualidade das bases
- Presença de adaptadores e contaminantes
- Comprimento dos reads
- Quantidade de reads

Indicador de qualidade Q (*Phred quality score*)

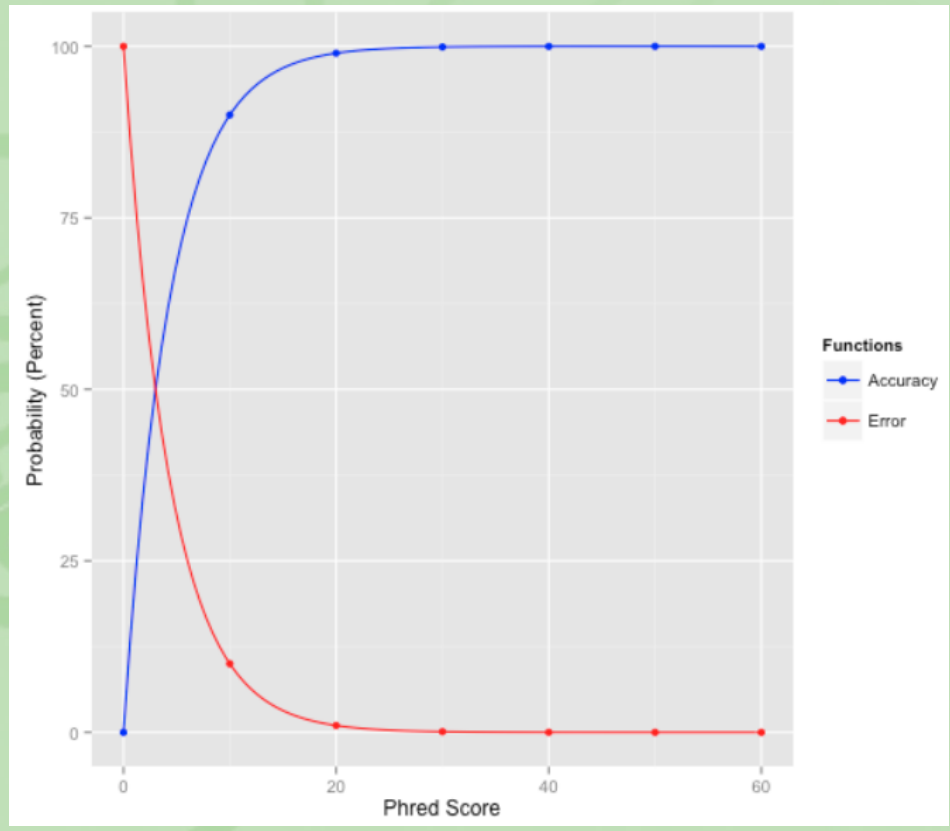
- Baseado na probabilidade de erro (*E*) na identificação de uma base em determinada posição do read

$$Q = -10 \log E$$

- Define a acurácia de uma base
  - 90%: um erro em cada 10 leituras (0.1), Q = 10
  - 99%: um erro em cada 100 leituras (0.01), Q = 20
  - 99,9%: um erro em cada 1.000 leituras (0.001), Q = 30
  - 99,99%: um erro em cada 10.000 (0.0001), Q = 40

- Q < 20, a perda de confiabilidade é muito alta e rápida
- Q > 20, o aumento na confiabilidade não é tão significativo

- **20 ou 25 como valores de corte em muitos casos**



Referência da Imagem

Formato FASTQ

- Formato de armazenamento de sequências biológicas e scores de qualidade correspondentes às bases

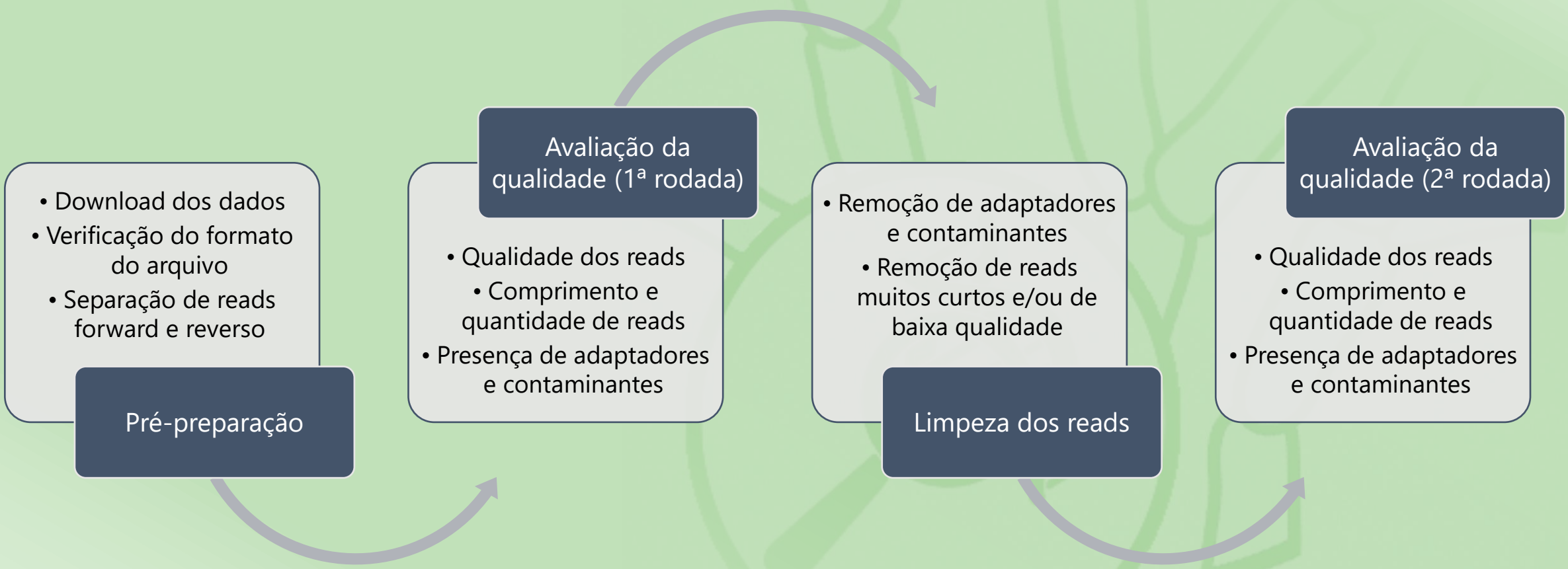
```
@A00178:149:H7K7YDSXY:4:1101:1506:1000 1:N:0:GCACTCAT+ATGAGTGC
CNGCTCTGGTCATCCGTCTCGGCTCGCGAGATTCAAGCGTTGCCGTCAACCTTGGCAATGTAGACAAGGA
GGTCGAGGACACGGCGGGAGTAGCCCCACTCGTTGTCGTACCAGGAGACGAGCTTGACGAAGTTCTCGTT
GAGCGAGATAC
+
F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFF:
FFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFF:FFFFFFFFFFFFF:FF,FFFFFFFFFF
FFFF:FF:FF:
```

Linha 01: começa com um @ e contém o identificador da sequência (similar à primeira linha do formato FASTA)
Linha 02: sequência em nucleotídeos
Linha 03: começa com um + e pode conter o identificar da sequência novamente
Linha 04: contém os valores de qualidade para a sequência na linha 02. Mesmo número de caracteres que a linha 02 (cada símbolo é correspondente à uma letra)

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

https://drive5.com/usearch/manual/quality_score.html


• Download dos dados
• Verificação do formato do arquivo
• Separação de reads forward e reverso
Pré-preparação
Avaliação da qualidade (1ª rodada)
• Qualidade dos reads
• Comprimento e quantidade de reads
• Presença de adaptadores e contaminantes
• Remoção de adaptadores e contaminantes
• Remoção de reads muitos curtos e/ou de baixa qualidade
Limpeza dos reads
Avaliação da qualidade (2ª rodada)
• Qualidade dos reads
• Comprimento e quantidade de reads
• Presença de adaptadores e contaminantes

Pré-preparação - Download

- Organizar os arquivos para que possam ser avaliados e processados:
  - Servidores online como o Galaxy
  - Servidor interno (e. g. cluster de um instituto de pesquisa)
  - Computador pessoal

- Obtenção de dados em bases de dados públicas
  - NCBI SRA
  - JGI Mycocosm

Pré-preparação – Verificação do formato do arquivo

- **Formato FASTQ (compactado ou descompactado):** formato compatível com os softwares de avaliação e processamento

- **Formato SRA:** exige conversão para o formato FASTQ para que possa ser avaliado e processador

Avaliação de qualidade (1ª rodada)

- Qualidade dos reads
- Comprimento e quantidade de reads
- Presença de adaptadores e contaminantes

- Diferentes módulos de análises e resultados no FastQC

Limpeza dos reads

- Remoção de adaptadores e contaminantes
- Remoção de reads muitos curtos e/ou de baixa qualidade

- Uso do software Trimmomatic e arquivo de sequências de adaptadores

Avaliação de qualidade (2ª rodada)

- Qualidade dos reads
- Comprimento e quantidade de reads
- Presença de adaptadores e contaminantes

- Diferentes módulos de análises e resultados no FastQC
- Determinar se a limpeza realizada pelo Trimmomatic foi suficiente e se os dados estão adequados para as análises posteriores