

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

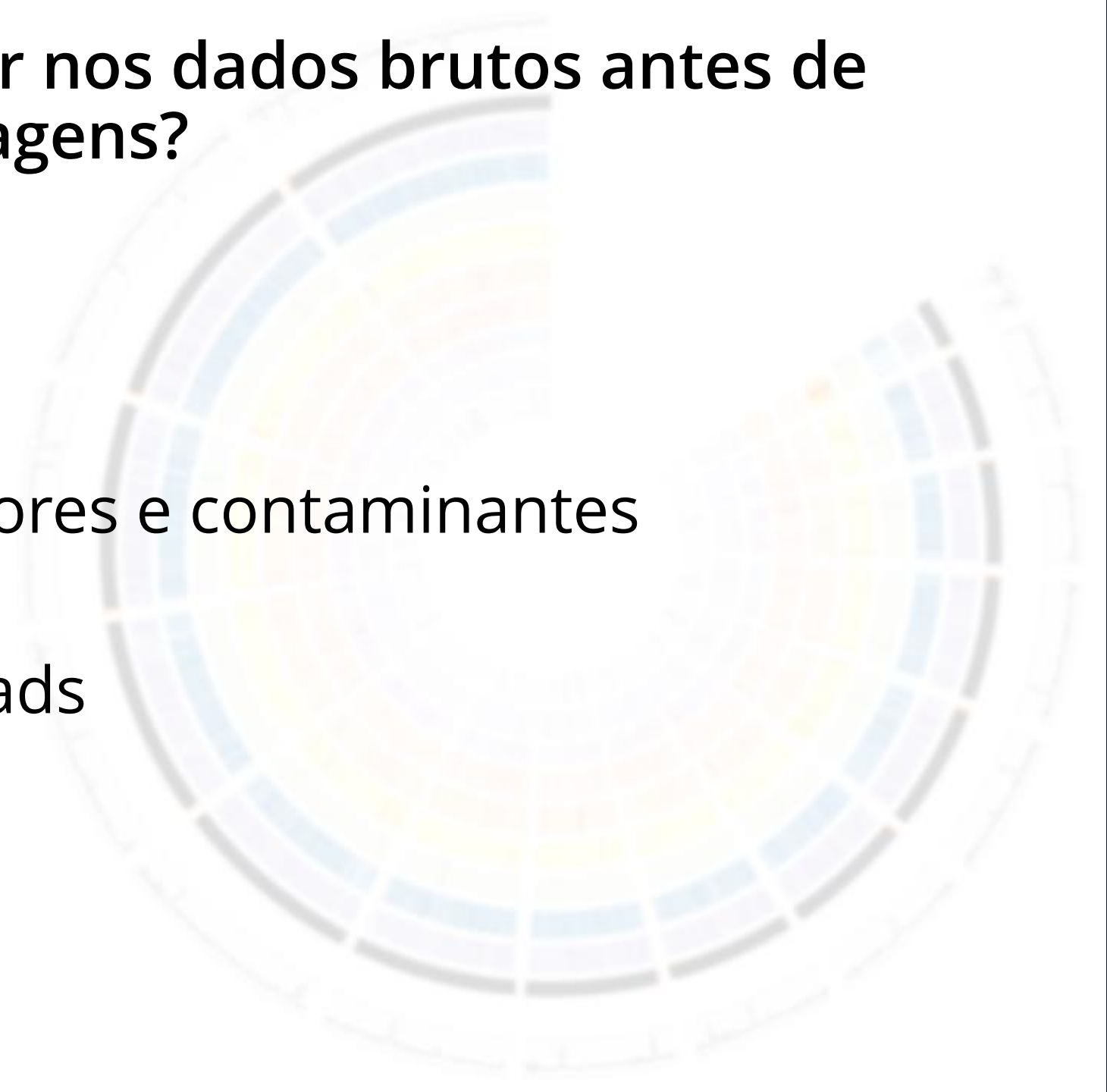
Dr<sup>a</sup> Desirrê Petters-Vandresen

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

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

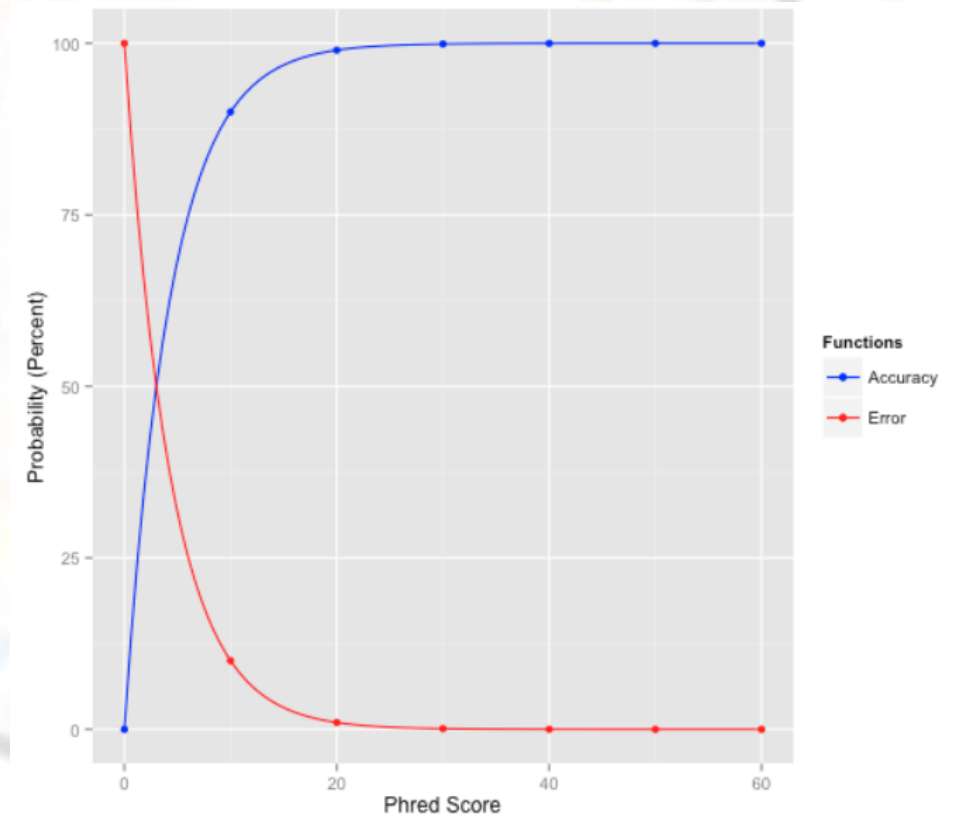
- Qualidade das bases
- Presença de adaptadores e contaminantes
- Comprimento dos reads
- Quantidade de reads



# Indicador de qualidade Q (*Phred quality score*)

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

$$Q = -10 \log E$$



# Formato FASTQ

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

```
1 @A00178:149:H7K7YDSXY:4:1101:1506:1000 1:N:0:GCACTCAT+ATGAGTGC
2 CNGCTCTGGTCATCCGTCTCGGCTCGCGAGATTCAAGCGTTGCCGTCAACCTTGGCAATGTAGACAAGGA
  GGTCGAGGACACGGCGGGAGTAGCCCCACCTCGTTGTCGTACCAGGAGACGAGCTTGACGAAGTTCTCGTT
  GAGCGAGATAC
3 +
4 F#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:
  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:
  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:
  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:
  FFFF:FF:FF:
```

**Linha 01:** começa com um @ e contém o identificador da sequência (similar à primeira linha do formato FASTA)

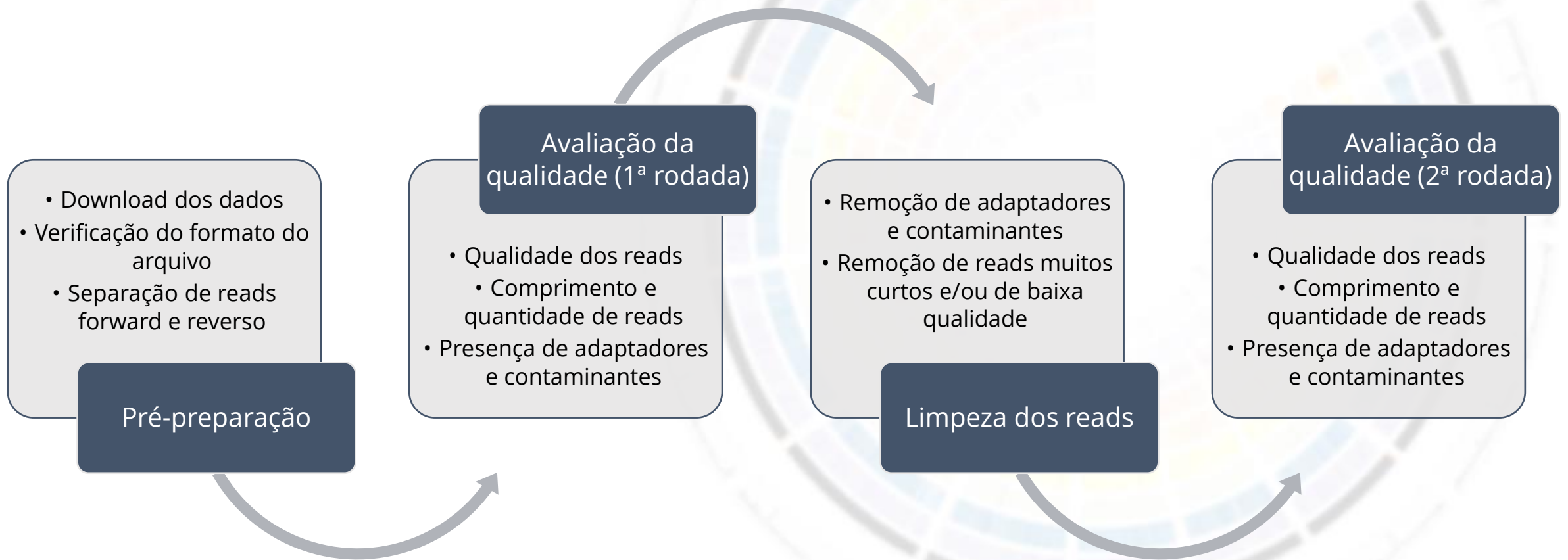
**Linha 02:** sequência em nucleotídeos

**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			





# 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 muito 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

# E se o genoma/transcriptoma já estiver pronto?

- Observar seção de materiais e métodos na publicação associada ao genoma ou transcriptoma e confirmar se o controle de qualidade foi realizado antes da montagem

For *P. capitalensis* LGMF01 and *P. citricarpa* LGMF06, libraries of the paired-end reads were processed with NxTrim (O'Connell et al., 2015) to remove Nextera adapters and generate mate-pair, paired-end, single-end and unknown libraries. The script “deinterleave\_fastq.sh” (<https://gist.github.com/nathanhaigh/3521724>) was used to separate the reads from the mate-pair and paired-end libraries in “forward” and “reverse” files. For quality filtering, Trimmomatic v 0.38 (Bolger et al., 2014) was used to (1) trim bases at the start and end of the reads below a quality threshold of 25, (2) trim low quality segments using a 4 bp sliding window and a quality threshold of 15, and (3) discard reads shorter than 50 bp. For *P. citribraziliensis* LGMF08, the reads were filtered in Trimmomatic to remove Illumina adapters and trim for quality using the same parameters described before but discarding reads shorter than 90 bp.

FastQC v 0.11.8 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) was employed to check the quality of reads after processing in Trimmomatic. De novo genome assemblies were generated with SPAdes v 3.13 (Bankevich et al., 2012), using default parameters. Contigs smaller than 500 bp were filtered and removed from the final assemblies, which were then evaluated with QUAST v 4.6.3 (Gurevich et al., 2013). Library and assembly statistics for the new assemblies are summarized in Table S1 and assemblies are available at Zenodo (<https://doi.org/10.5281/zenodo.3750350>).