UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

PPGBCM
CBiot/UFRGS

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

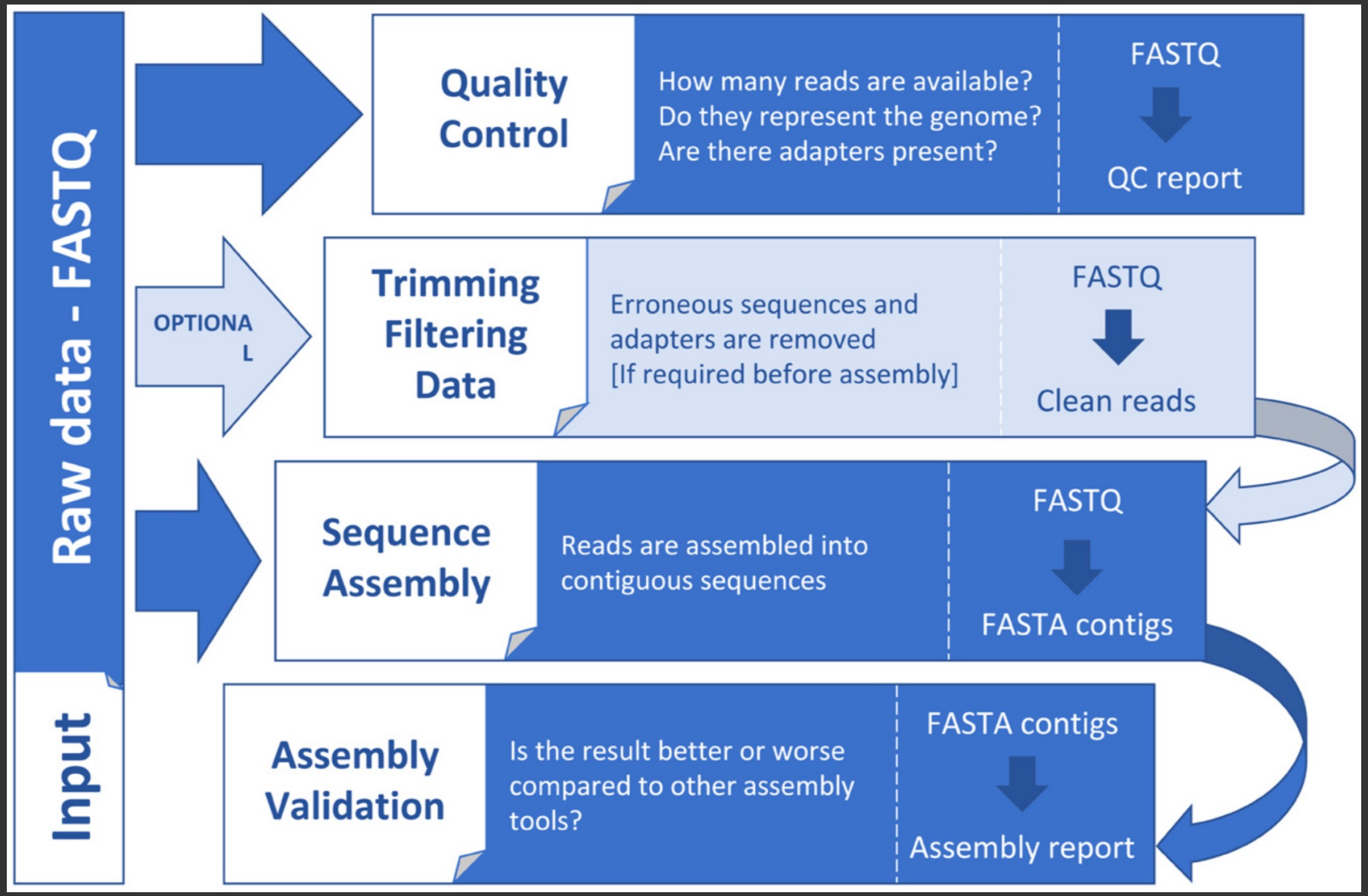
CSC

Introdução à metagenômica: curso prático

Controle de qualidade de sequências

[igor_pessi@UniversityOfHelsinki]$ _

Outubro/2022



Del Angel et al. (2018)

Erros e vieses de sequenciamento

Extração de DNA
    # Envelope celular
    # Biofilmes
    # Contaminação (manipulação, hospedeiro)
Preparação da biblioteca
    # PCR
    # Regiões rica em GC
Sequenciamento
    # Mudanças de bases
    # Inserções/deleções
    # Presença de adaptadores

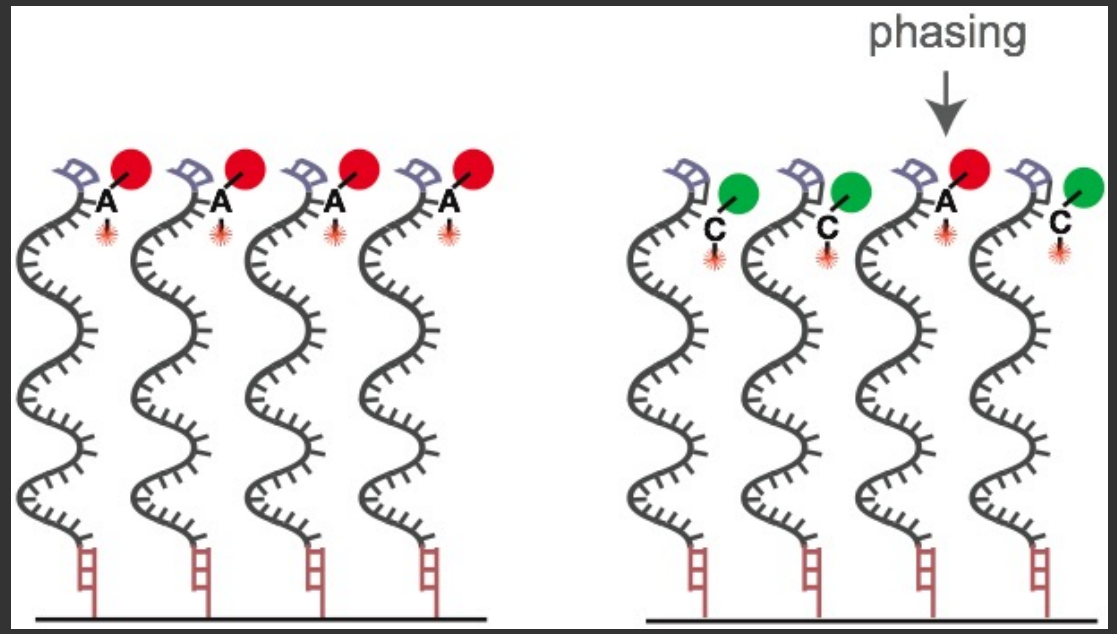
Erros e vieses de sequenciamento

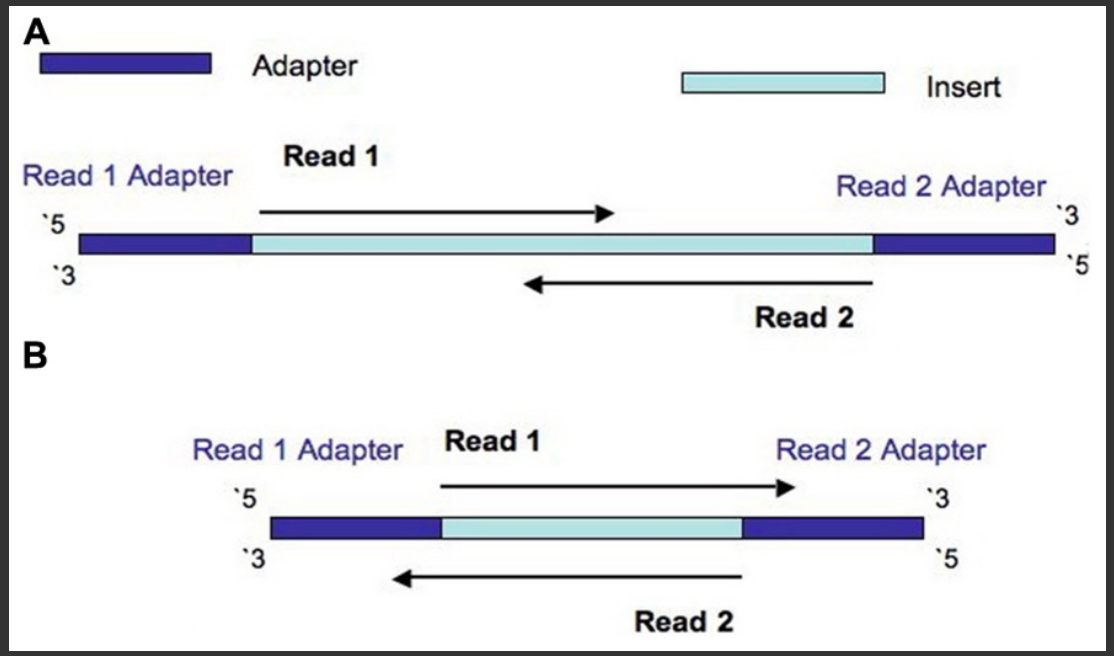
```
# Extração de DNA
    # Envelope celular
    # Biofilmes
    # Contaminação (manipulação, hospedeiro)    // CORRIGÍVEL //
# Preparação da biblioteca
    # PCR
    # Regiões rica em GC
# Sequenciamento
    # Mudanças de bases                         // CORRIGÍVEL //
    # Inserções/deleções                        // CORRIGÍVEL //
    # Presença de adaptadores                   // CORRIGÍVEL //
```

Sequenciamento Illumina: dois tipos principais de erros/artefatos

Erros de fase resultam na queda da qualidade da sequência

Adapter read-through resulta na presença de adaptadores na região 3’ da sequência





Anatomia de arquivos .fastq

```
@M02764:119:000000000-C5R9K:1:1101:15139:1363 1:N:0:24
```

Cabeçalho

Anatomia de arquivos .fastq

```
@M02764:119:000000000-C5R9K:1:1101:15139:1363 1:N:0:24
GCTAACCCCATTGCAACGTGGTAACTTGTTAGACCGTTTTTAAAAGTCGCTGAAGCAGCCACGATAAAC
```

Cabeçalho

Sequência de DNA

Anatomia de arquivos .fastq

```
@M02764:119:000000000-C5R9K:1:1101:15139:1363 1:N:0:24
GCTAACCCCATTGCAACGTGGTAACTTGTTAGACCGTTTTTAAAAGTCGCTGAAGCAGCCACGATAAAC
+
```

Cabeçalho

Sequência de DNA

Separador

Anatomia de arquivos .fastq

```
@M02764:119:000000000-C5R9K:1:1101:15139:1363 1:N:0:24
GCTAACCCCATTGCAACGTGGTAACTTGTTAGACCGTTTTTAAAAGTCGCTGAAGCAGCCACGATAAAC
+
CCCCCEFGGGGGFCGGGGGGGGGGGGFCCDFGGFFCCCFCC8EFF@FFCCECC@F<FGFGGGGGG7B7<C
```

Cabeçalho

Sequência de DNA

Separador

Valor Phred

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			

Índice de qualidade Phred

Medida de qualidade da identificação das bases

Definido logaritmicamente à probabilidade de erro

$$Q = \log_{10}(p) \times -10$$

Valor Phred	Prob. de base incorreta	Precisão	ASCII
10	1/10 (0.1)	90%	+
20	1/100 (0.01)	99%	5
30	1/1000 (0.001)	99.9%	?
40	1/10000 (0.0001)	99.99%	I

O que fazer?

Table 2. A selection of quality control software tools for metagenomics data

Tool	Synopsis	Reference	Web site
FastQC	Quality control tool showing statics such as quality values, sequence length distribution and GC content distribution	[33]	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
FastQ Screen	Screen a library against sequence databases to see if composition of library matches expectations	[37]	http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen
BBtools	BBDuk trims and filters reads using *k*-mers and entropy information. BBNorm normalizes coverage by down-sampling reads (digital normalization)	[35]	http://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/
Trimmomatic	Flexible read trimming tool for Illumina data	[36]	http://www.usadellab.org/cms/?page=trimmomatic
Cutadapt	Find and remove adapter sequences, primers, poly-A tails and other types of unwanted sequence	[34]	https://cutadapt.readthedocs.io
khmer/diginorm	Tools for *k*-mer error trimming of reads and digital normalization of samples	[38, 39]	http://khmer.readthedocs.io
MultiQC	Summarize results from different analysis (such as FastQC) into one report	[40]	http://multiqc.info

Note: Most of these tools can also be used for other types of genome sequence data, e.g. whole-genome or RNA-seq data.

Breitwieser et al. (2019)