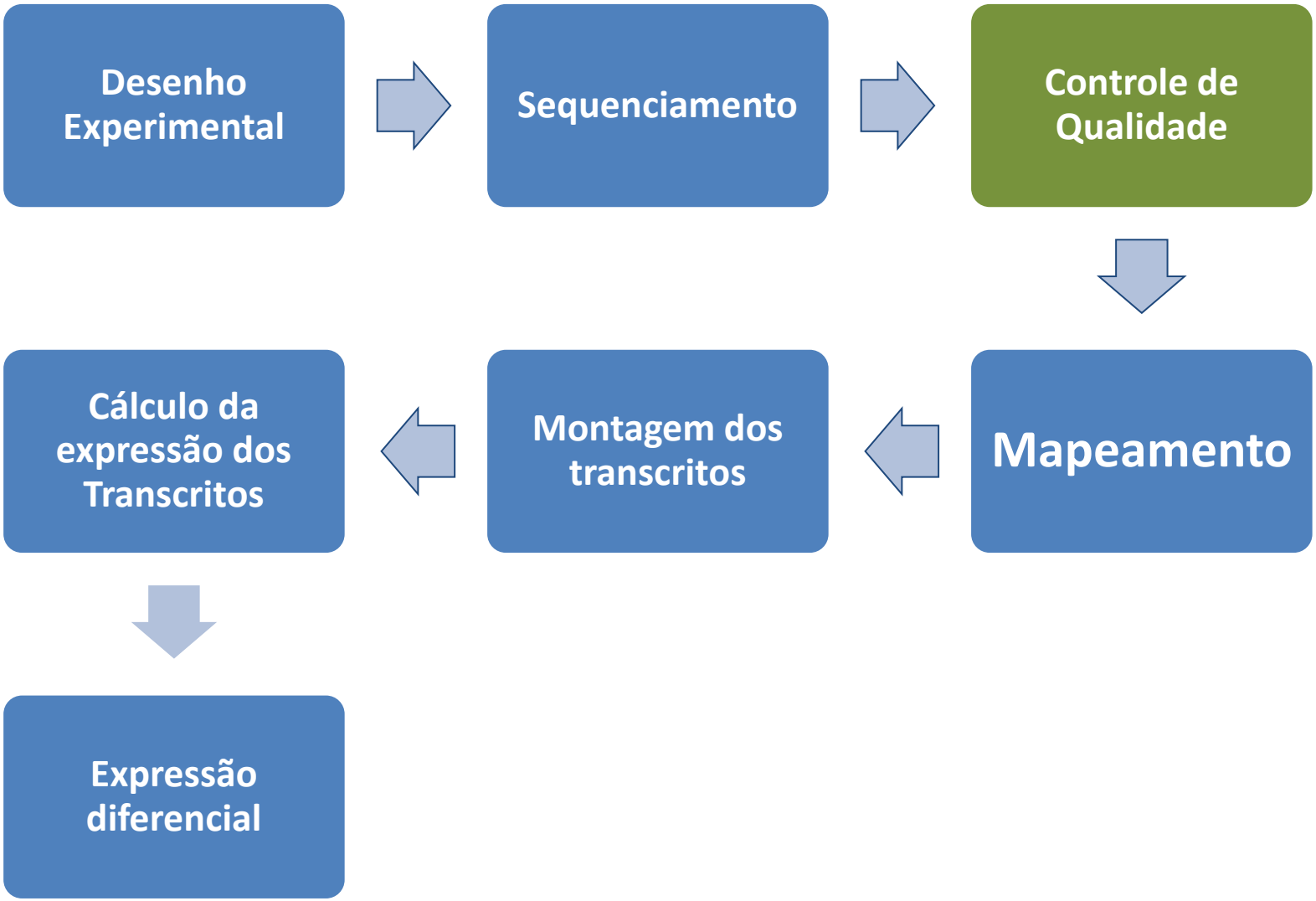
RNA-Seq
Controle de Qualidade

Vagner Okura

LaCTAD

vagnerko@unicamp.br

RNA-Seq


Desenho Experimental
Sequenciamento
Controle de Qualidade
Mapeamento
Montagem dos transcritos
Cálculo da expressão dos Transcritos
Expressão diferencial

Controle de Qualidade

- ✓ Motivação
  - ✓ Reads contém erros;
  - ✓ Sequências com qualidade ruim afetam os passos posteriores - mapeamento, montagem de transcritos;
  - ✓ Correção de erros aumenta a taxa de reads mapeados e diminui o uso de memória RAM em montagens.

- ✓ O controle de qualidade envolve dois passos:
  - ✓ Análise e identificação - FASTQC
  - ✓ Limpeza - Trimmomatic
  - ✓ Dica: FASTQC – Trimmomatic - FASTQC

Formato FASTA

```
>SEQ1_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
>SEQ2_ID
TTAATTGGTAAATATAGTTTCTTGAGATTTGTTGGGGGAGACATTTTTGTGATTGCCTTGA
>SEQ3_ID
TAGCGATGCGTCACGACTCGTCAGCTCAGCTCGCCTTCGAGACCGCCTACGCATCGCCT
>SEQ4_ID
CCGCTAGCCATCAGCGCAGTCGCTCGACATCGATGCGCGGGAAAGAGAGACATCGCAG
```

Extensões: fasta, fa, fna.

Formato FASTQ

```
@HWUSI:155:C6013ACXX:6:73:941:1973 1:Y:0:CGATGT
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
@HWUSI:155:C6013ACXX:6:73:941:1988 1:Y:0:CGATGT
TTAATTGGTAAATATAGTTTCTTGAGATTTGTTGGGGGAGACATTTTTGTGATTGCCTTGA
+
efcfffffcfeefffcfddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`d
```

Extensões: fastq, fq.

Formato FASTQ

@HWUSI:155:C6013ACXX:6:73:941:1973 1:Y:0:CGATGT

Versão Illumina 1.8

HWUSI	Identificador único da máquina
155	ID da corrida
C6013ACXX	ID da flowcell
6	Lane
73	Número do bloco da lane
941	Coordenada X no bloco
1973	Coordenada Y no bloco
1	Reads pareados: 1 ou 2 *(paired-end ou mate-pair)*
Y	Y para read filtrado, ou N, caso contrário.
0	Usado para bits de controle. 0=desligado
CGATGT	Seqüência do index (multiplexing)

@HWUSI:6:73:941:1973#0/1

Versões anteriores

#0: index para amostras multiplexadas (ou 0); /1 reads pareados

Formato FASTQ

Codificação para valores de qualidade

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ...............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |        |                              |                     |
33                         59   64       73                            104                   126
 0........................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31.........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Formato FASTQ

O que significa esses valores de qualidade?

É um valor inteiro associado à probabilidade de que a base correspondente está incorreta.

- ✓ Usualmente, valor de qualidade >= 20;
- ✓ Quando a qualidade é 0, a base é substituida pela letra N.

$$Q_{phred} = -10 \log_{10} (p)$$

$$Q_{solexa} = -10 \log_{10} \frac{p}{1-p}$$

Score	p_{error}
10	0.01
20	0.001
30	0.0001

Ferramenta: FASTQC

- ✓ Os sequenciadores geram relatórios de controle de qualidade direcionados a para identificação de problemas com a corrida (números gerais, por lane);
- ✓ FastQC fornece um relatório que proporciona analisar base a base problemas relacionados ao sequenciamento das bibliotecas (amostras).
- ✓ Possui interface gráfica, ou pode ser executado por linha de comando, permitindo processamento de um grande número de arquivos.

Ferramenta: FASTQC

Illumina, 454 e PacBio

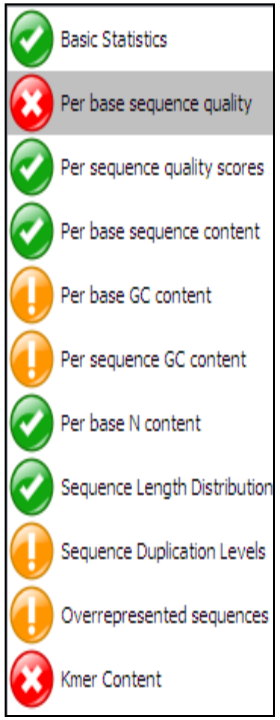
FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution

Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45



http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

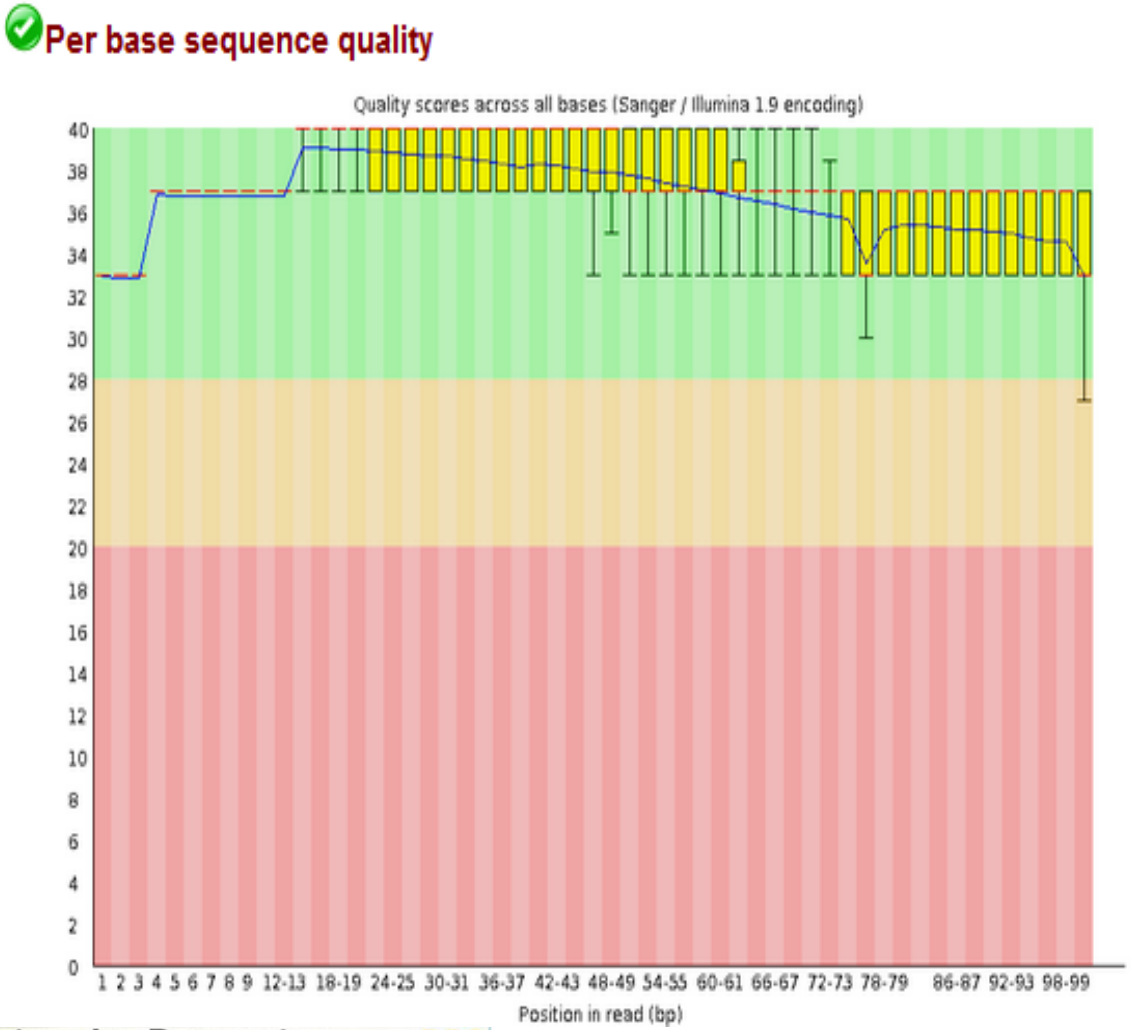
Ferramenta: FASTQC

Basic Statistics

Measure	Value
Filename	s_1_sample_1A
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	35446174
Sequences flagged as poor quality	0
Sequence length	100
%GC	49

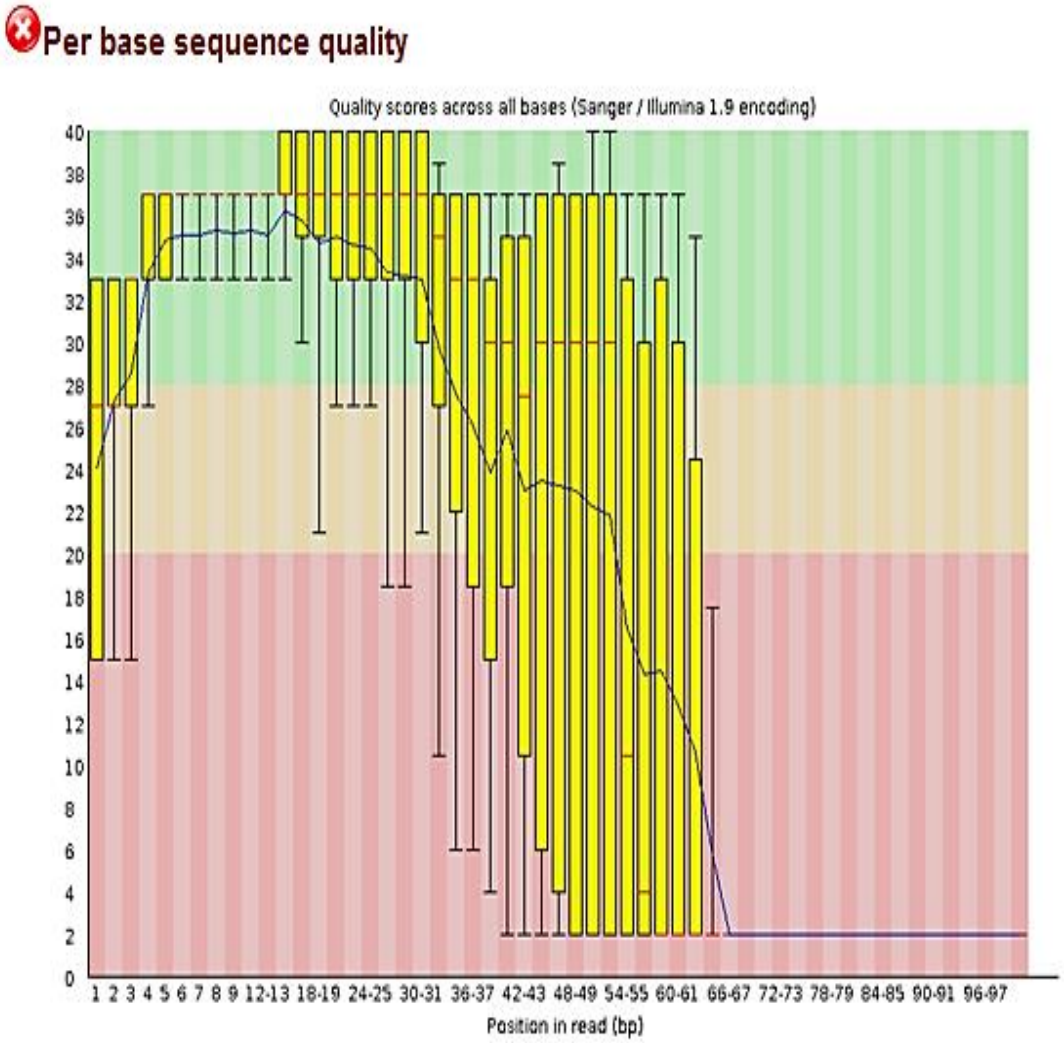
Ferramenta: FASTQC

Qualidade boa



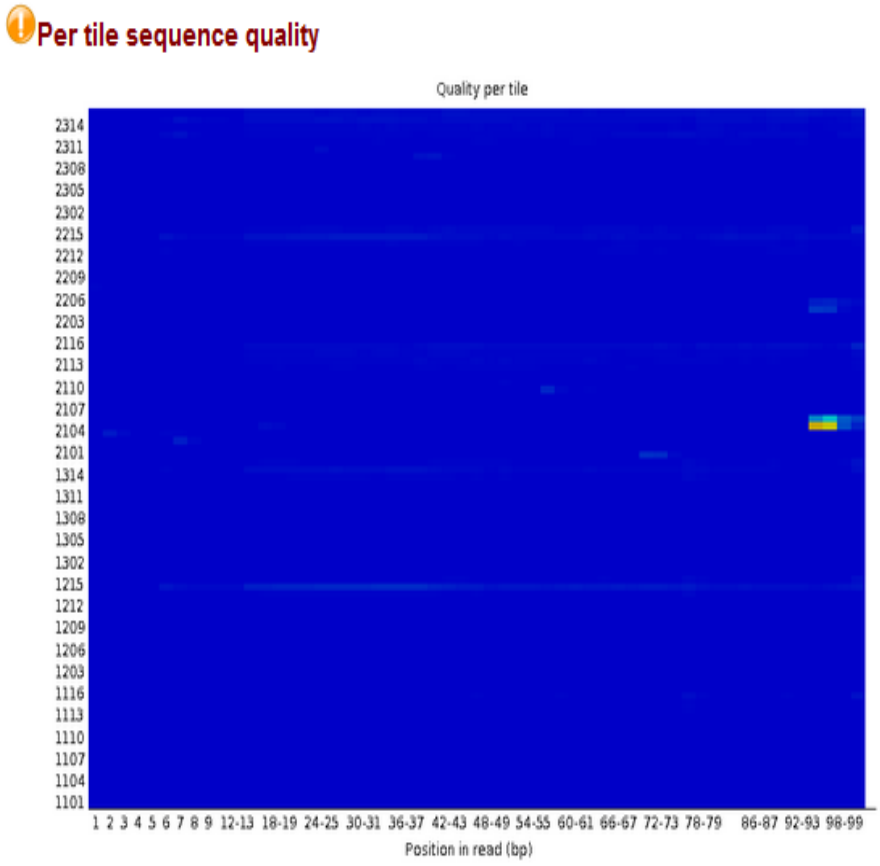
Ferramenta: FASTQC

Qualidade ruim

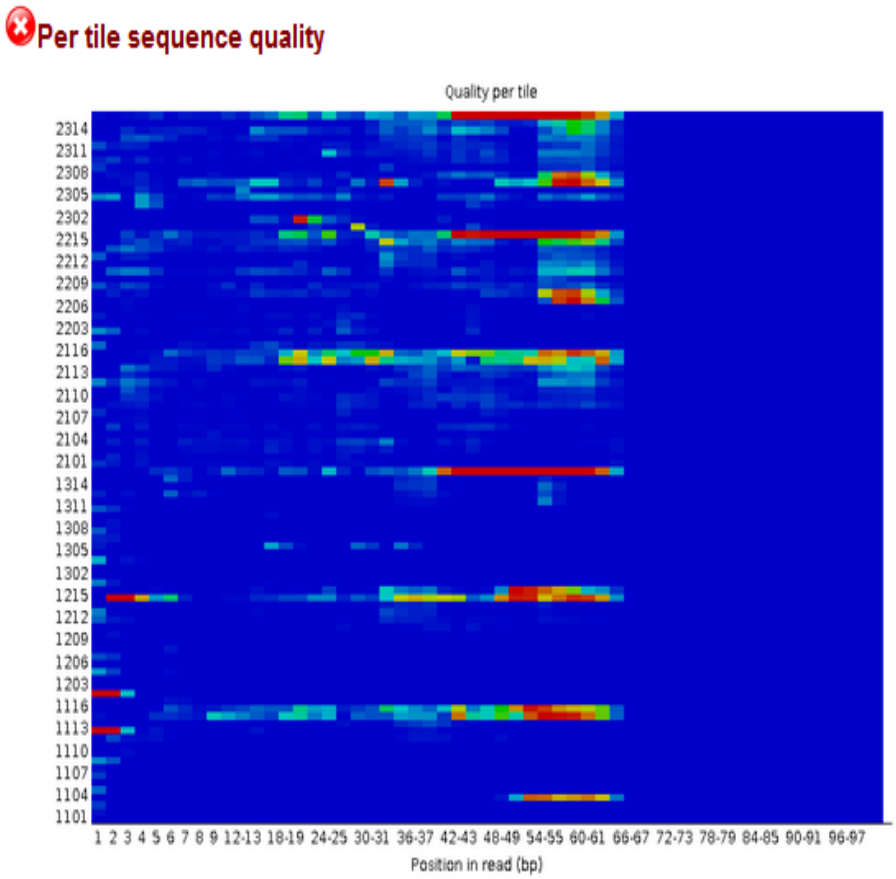


Ferramenta: FASTQC

Qualidade boa



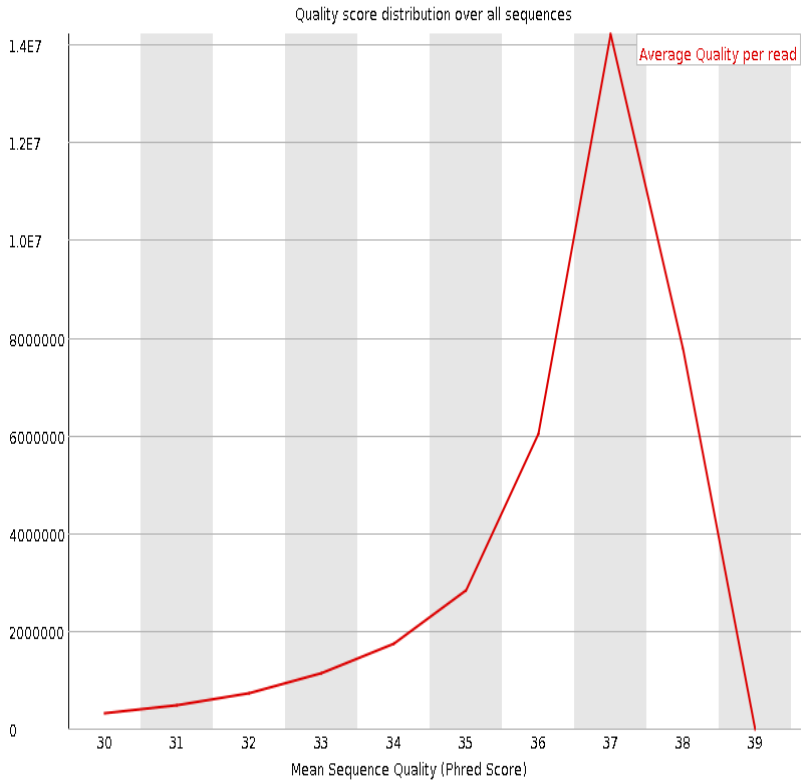
Qualidade ruim



Ferramenta: FASTQC

Per sequence quality scores

Qualidade boa



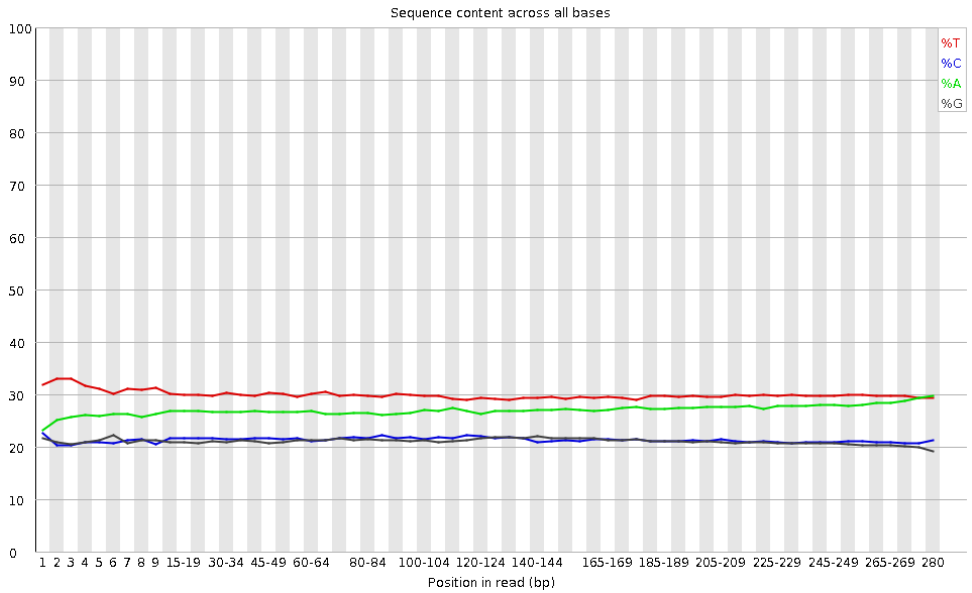
Qualidade ruim

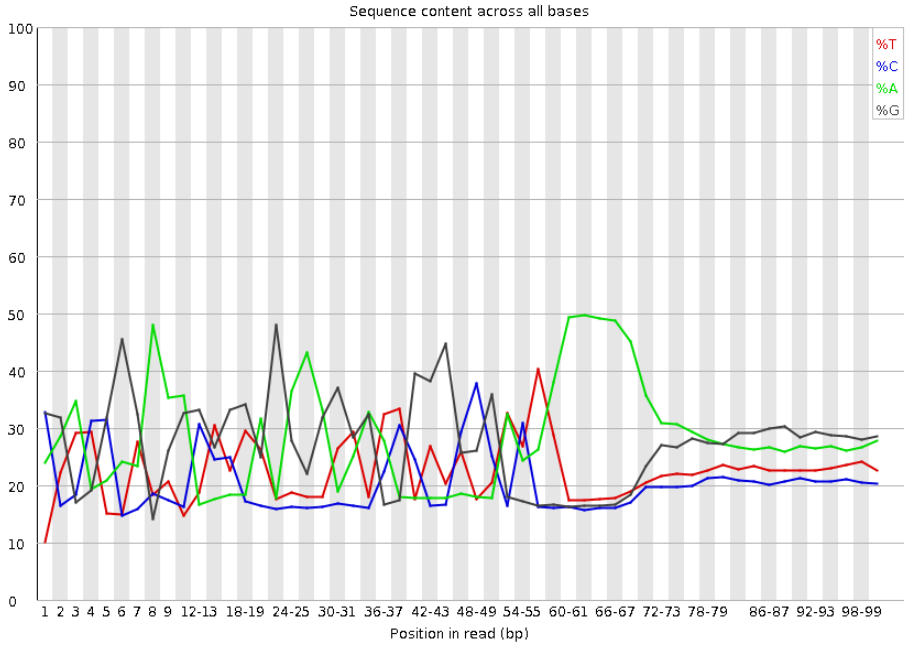
Ferramenta: FASTQC

Per Base Sequence Content

Proporção esperada entre bases

Proporção não balanceada pode indicar contaminação ou bias da biblioteca, ou problema no sequenciamento.

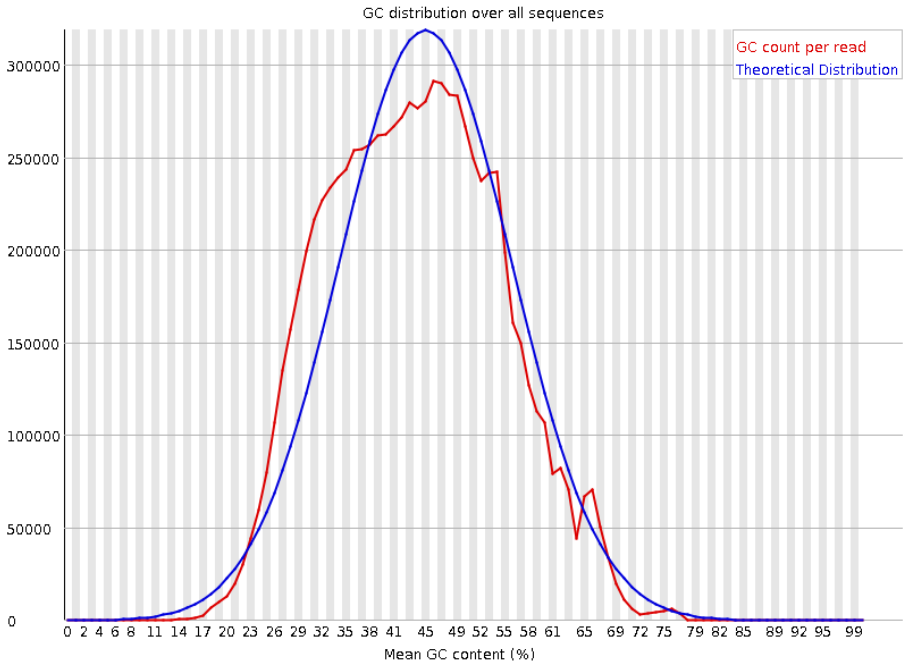




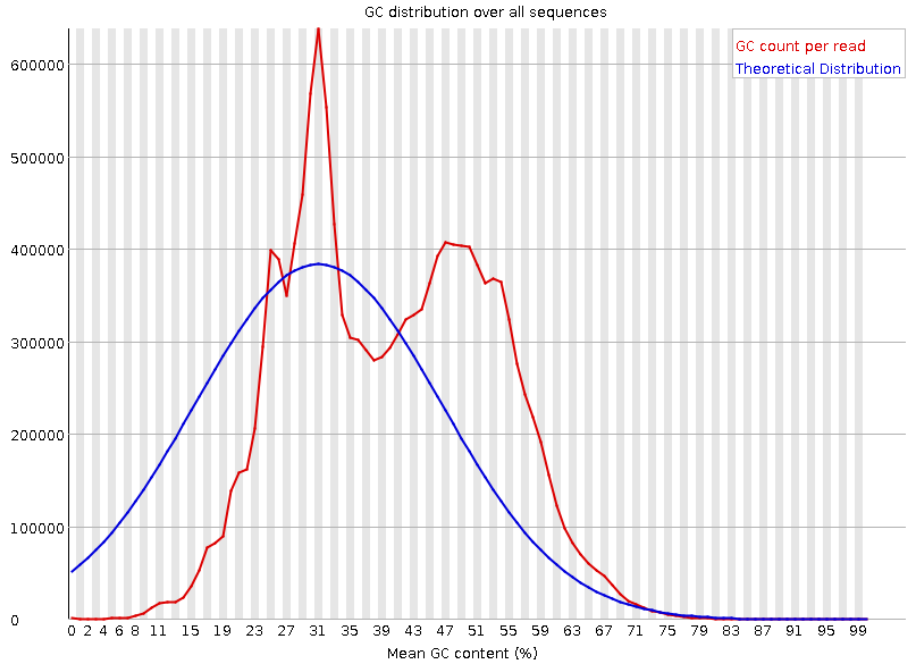
Ferramenta: FASTQC

Per Sequence GC Content

Distribuição normal observada é similar a distribuição modelada.



Distribuição diferente pode indicar contaminação da biblioteca, ou outro bias.



Ferramenta: FASTQC

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC	2295719	47.036867813501395	TruSeq Adapter, Index 2 (100% over 50bp)
CATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC	115042	2.3570895858773775	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTTTGC	70115	1.4365826073416	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGCATCTCGTATGC	65650	1.3450994533548604	TruSeq Adapter, Index 2 (98% over 50bp)
NATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC	63095	1.2927501905472187	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATTTCGTATGC	50134	1.0271929321324078	TruSeq Adapter, Index 2 (98% over 50bp)
AATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC	49997	1.0243859462206086	TruSeq Adapter, Index 2 (98% over 50bp)
GACCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC	36977	0.7576198398583804	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGATGTATCTCGTATGC	29030	0.5947941680257669	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTATGTATCTCGTATGC	26354	0.5399657424785071	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGCTGTATCTCGTATGC	24414	0.5002171828515698	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTTGTATGC	24074	0.4932509404427252	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGATGTATCTCGTATGC	21536	0.44124998975555259	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGACGTATCTCGTATGC	20529	0.4206176188563888	TruSeq Adapter, Index 2 (98% over 50bp)
TATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC	13555	0.27772769368202793	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGT	13502	0.27664177942417856	TruSeq Adapter, Index 2 (97% over 49bp)
GTTCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC	13333	0.27317914716801756	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCATCGATGTATCTCGTATGC	10130	0.20755304588704854	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATCTATCTCGTATGC	10034	0.20558610685396297	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGTTGTATCTCGTATGC	9798	0.20075071506429434	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATTC	9550	0.19566945589548998	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCTTATGC	8668	0.17759820352901645	TruSeq Adapter, Index 2 (98% over 50bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACGTATCTCGTATGC	8378	0.17165640853323716	TruSeq Adapter, Index 4 (97% over 37bp)
GAACGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGC	8266	0.16936164632797068	TruSeq Adapter, Index 2 (98% over 50bp)

Busca por possíveis fontes: contaminações ou adaptadores.

Ferramenta: TRIMMOMATIC

http://www.usadellab.org/cms/?page=trimmomatic

Ferramenta: TRIMMOMATIC

ILLUMINACLIP: corta os adaptadores ou outra sequências específicas dos *reads*.

- *Passa o arquivo contendo adaptadores; número máximo de mismatches; simples ou palindrômico.*

Ferramenta: TRIMMOMATIC

SLIDINGWINDOW: corte usando janela deslizante;

- *Passa o tamanho da janela e a qualidade média*

ACTTAGCTAGCTAGAGATCGTATAGCTTCTACGCATAGA
TCAGACGTAAG

LEADING: corta bases com qualidade abaixo de x (parâmetro) no começo da sequência.
TRAILING: corta bases com qualidade abaixo de x (parâmetro) no final da sequência.

Ferramenta: TRIMMOMATIC

CROP: corta o read em um tamanho específico (corta no final)

- *Passa o tamanho que a sequência deve ficar*

HEADCROP: corta o read em um tamanho específico (corta no começo)

- *Passa o tamanho que deve ser cortado da sequência*

MINLEN: descarta reads com tamanho menos que x (parâmetro)

- *Passa o tamanho mínimo para manter o read*