



Análise e interpretação de métricas de qualidade

Luís Vieira

Objetivos

- Avaliar qualitativamente as *reads* de NGS utilizando o software *FastQC*



FastQC

- Software desenvolvido pelo Grupo de Bioinformática do *Babraham Institute (Cambridge)*
- Permite verificar de forma simples um conjunto de parâmetros de qualidade das amostras (análise primária) antes de se progredir com a análise secundária
- As funções principais são:
 - Fornecer uma visão geral das áreas em que poderão ter ocorrido problemas ou *biases (analysis modules)*
 - Apresentar tabelas e gráficos para visualizar rapidamente os dados
 - Exportar os resultados para um relatório *.html*

FastQC

- O FastQC pode utilizar ficheiros *fastq*. (comprimidos ou não), *bam*. ou *sam*. (por defeito, o FastQC adivinha o tipo de ficheiro)
- Pode abrir vários ficheiros em simultâneo e, dependendo do tamanho dos ficheiros, pode levar vários minutos a abrir
- Os resultados da análise podem ser guardados num relatório *html*, que pode ser distribuído. Por defeito, o nome do relatório contém a designação do ficheiro *fastq* e uma terminação *_fastqc.html*
- O relatório pode ser aberto num browser (e.g., *Google Chrome*)

Ficheiros *fastq*

- Os ficheiros *fastq* são ficheiros em formato de texto que armazenam informação sobre a sequência de nucleótidos e os correspondentes valores de qualidade de cada *read*
- Por questões de economia de espaço, os valores de qualidade estão codificados com um único carácter ASCII

Códigos ASCII

Table 4 ASCII Codes for Q-Scores 0–40

Symbol	ASCII Code	Q-score	Symbol	ASCII Code	Q-score
!	33	0	6	54	21
"	34	1	7	55	22
#	35	2	8	56	23
\$	36	3	9	57	24
%	37	4	:	58	25
&	38	5	;	59	26
'	39	6	<	60	27
(40	7	=	61	28
)	41	8	>	62	29
*	42	9	?	63	30
+	43	10	@	64	31
,	44	11	A	65	32
-	45	12	B	66	33
.	46	13	C	67	34
/	47	14	D	68	35
0	48	15	E	69	36
1	49	16	F	70	37
2	50	17	G	71	38
3	51	18	H	72	39
4	52	19	I	73	40
5	53	20			

Ficheiros *fastq*

- O ficheiro *fastq* contém normalmente 4 linhas:
 - Linha 1: Identificador (após o símbolo @)
@Instrument:RunID:FlowCellID:Lane:Tile:X:Y ReadNum:FilterFlag:0:SampleNumber
 - Linha 2: Sequência
 - Linha 3: Um sinal + (seguido de uma descrição – opcional)
 - Linha 4: Valores de qualidade (o número de caracteres tem de ser o mesmo que o da sequência)

Exemplo:

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#<=
```

http://en.wikipedia.org/wiki/FASTQ_format, acedido 21/05/2015; MiSeq Reporter User Guide (Illumina, Feb 2014)

Valores de qualidade

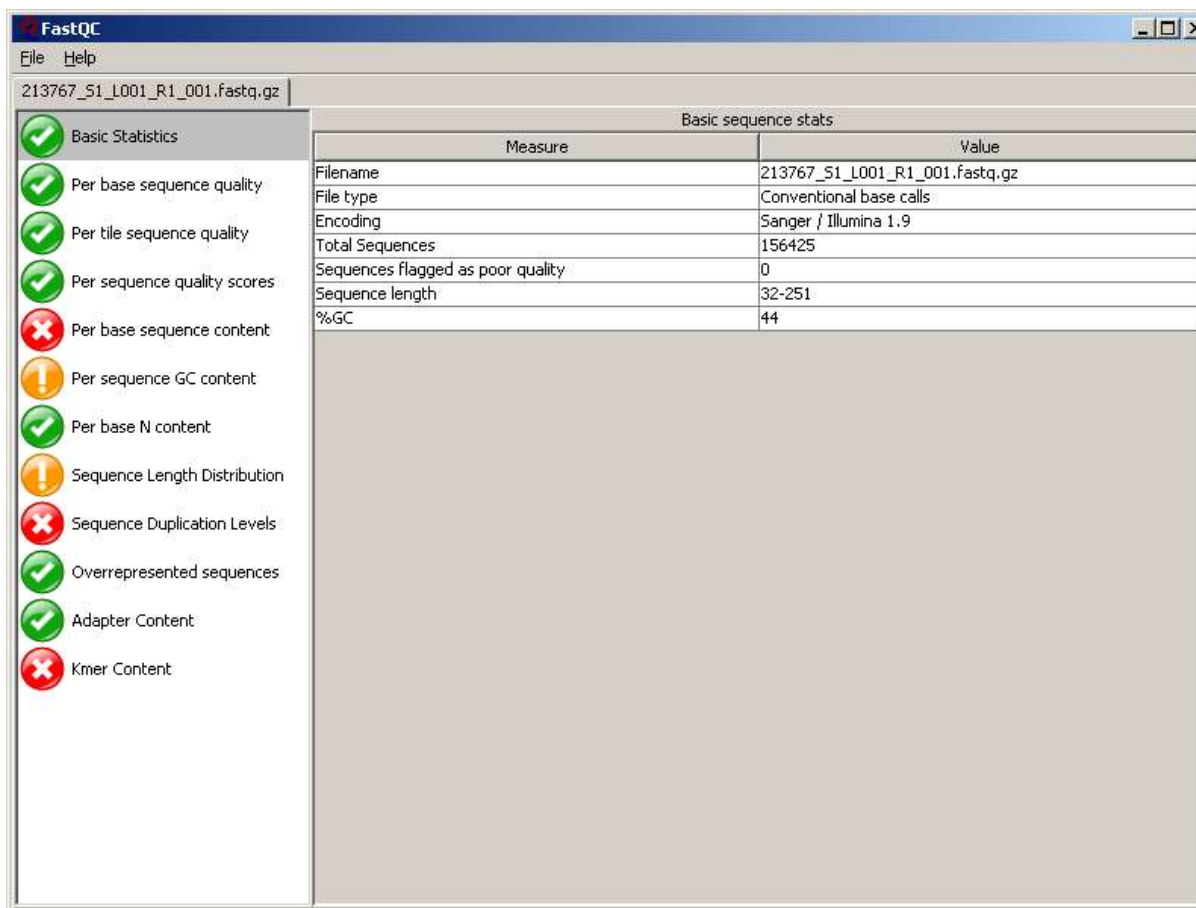
- Os valores de qualidade estão associados logaritmicamente às probabilidades de erro:

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

p – estimativa da probabilidade de erro para um *base-call*

Valores de qualidade	Probabilidade da base atribuída estar errada	Precisão da base atribuída
Q10	1 em 10	90%
Q20	1 em 100	99%
Q30	1 em 1.000	99.9%
Q40	1 em 10.000	99.99%

Analysis modules



The image shows the FastQC software interface. On the left, a list of analysis modules is displayed, each with a status icon (green checkmark for passed, red X for failed, orange exclamation mark for warning). The 'Basic Statistics' module is selected and expanded, showing a table of basic sequence statistics.

Basic sequence stats	
Measure	Value
Filename	213767_S1_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	156425
Sequences flagged as poor quality	0
Sequence length	32-251
%GC	44

Analysis Modules and Status:

- Basic Statistics (Passed)
- Per base sequence quality (Passed)
- Per tile sequence quality (Passed)
- Per sequence quality scores (Passed)
- Per base sequence content (Failed)
- Per sequence GC content (Warning)
- Per base N content (Passed)
- Sequence Length Distribution (Warning)
- Sequence Duplication Levels (Failed)
- Overrepresented sequences (Passed)
- Adapter Content (Passed)
- Kmer Content (Failed)

Analysis modules

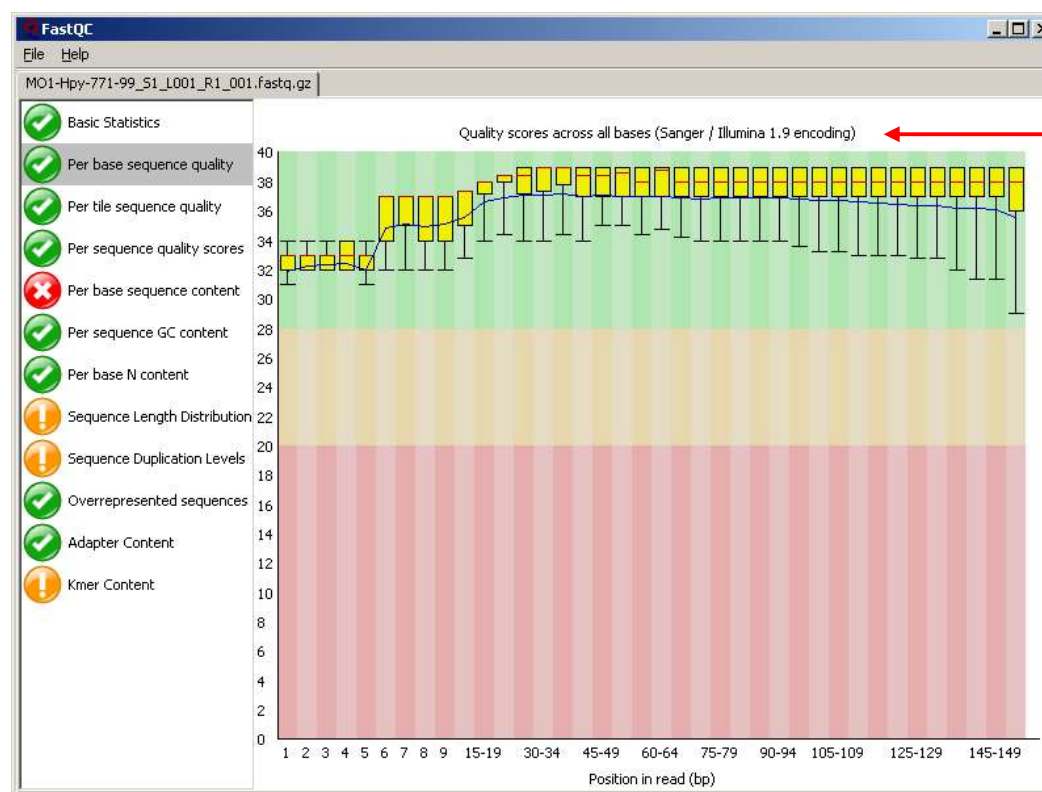
- O *FastQC* assume que as amostras “normais” são constituídas por bibliotecas de fragmentos aleatórios e diversas; logo, a escala de cores (verde, amarelo, vermelho) pode não refletir os resultados para o tipo de experiência efetuada, i.e., os resultados de cada módulo de análise devem ser analisados no contexto daquilo que esperamos para as nossas bibliotecas
- Para alguns módulos, o *FastQC* analisa apenas as primeiras 100.000-200.000 sequências de forma a utilizar pouca memória e processamento



Exemplos

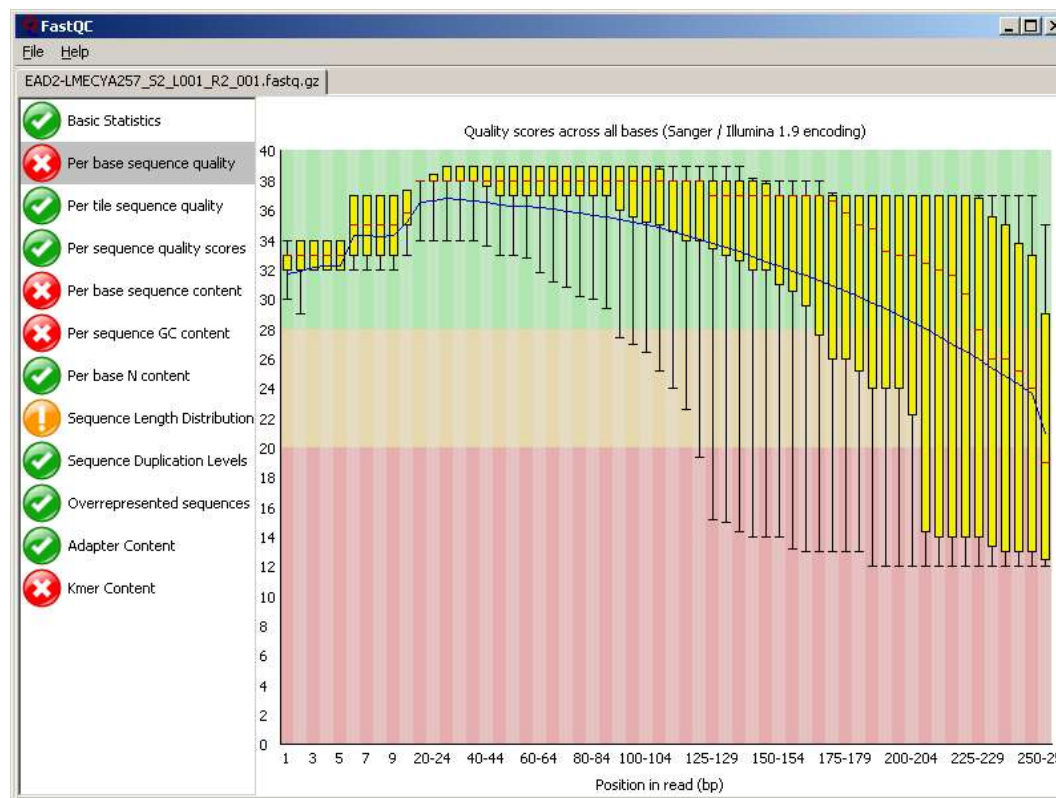
Per base sequence quality

- Sequenciação de genoma –qualidade elevada por base

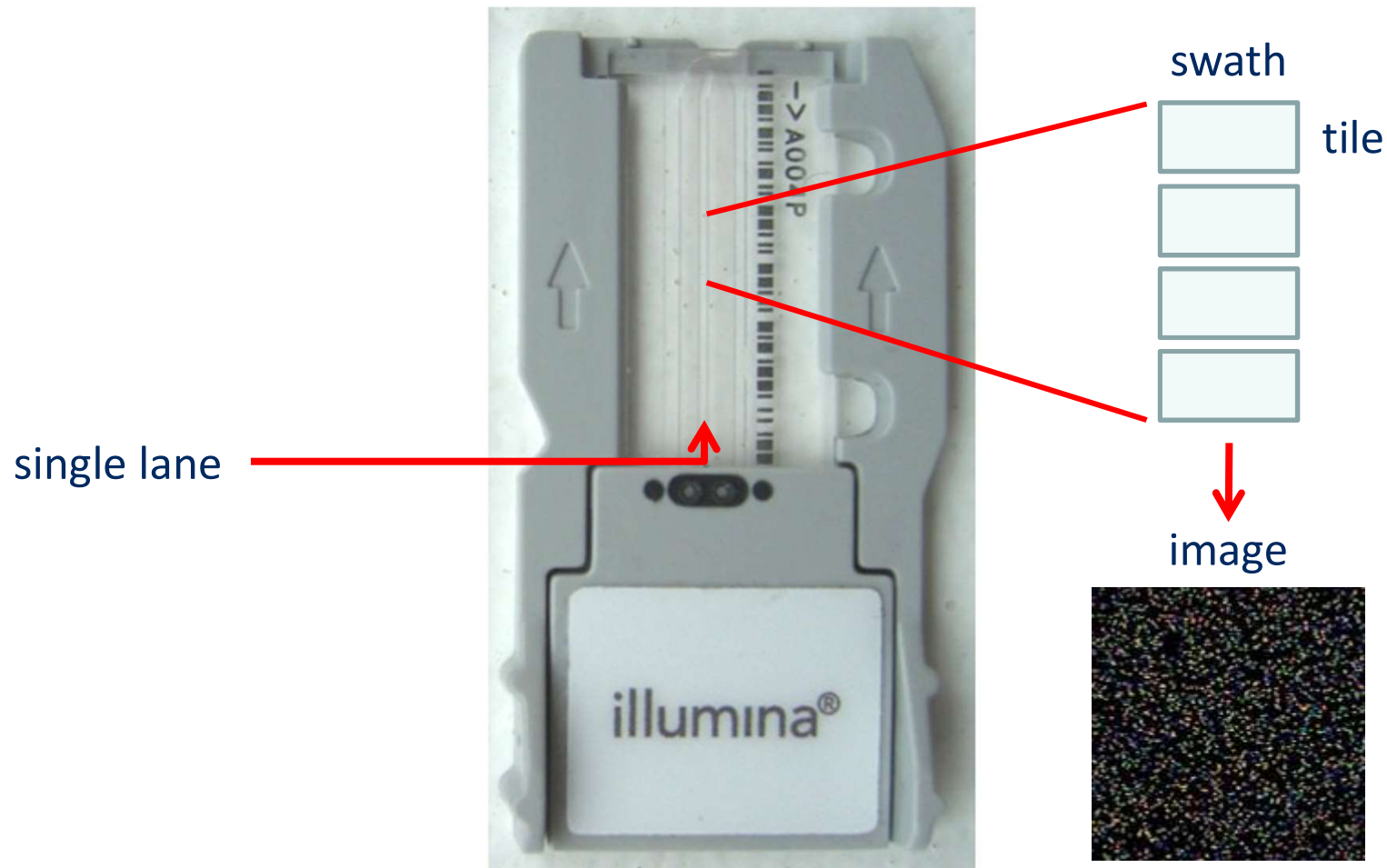


Per base sequence quality

- Sequenciação de genoma – baixa qualidade > base 170

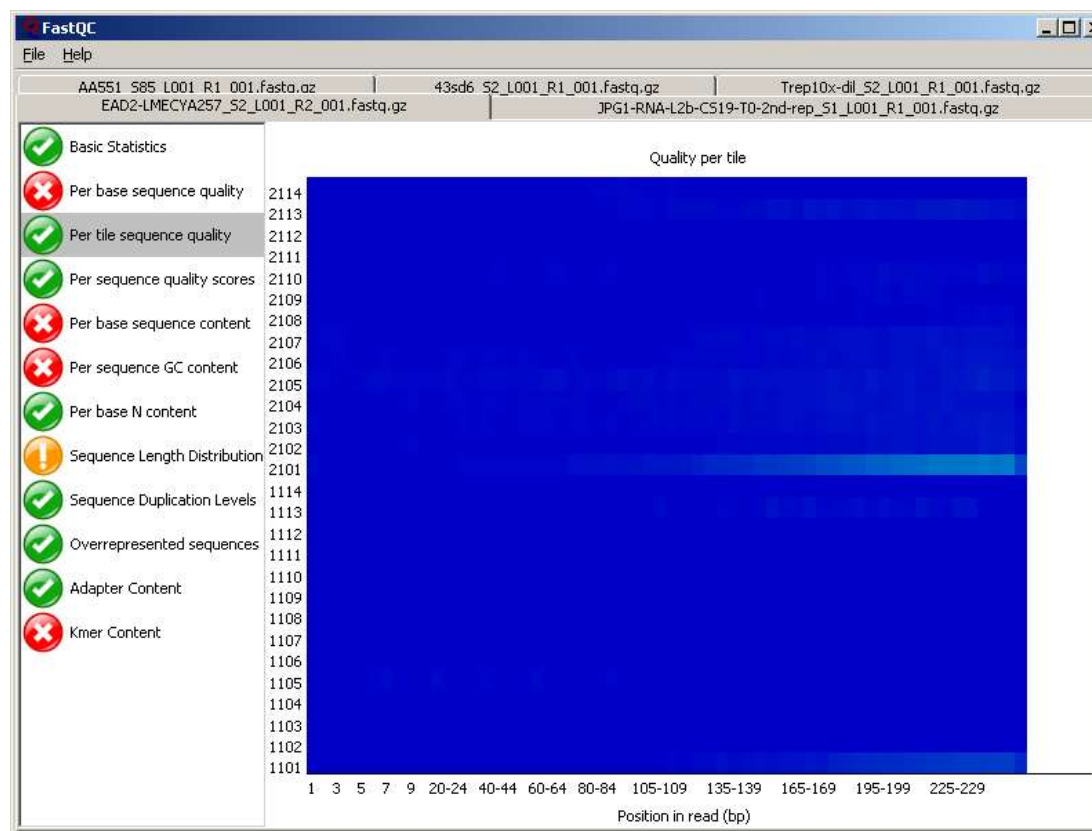


MiSeq flow cell



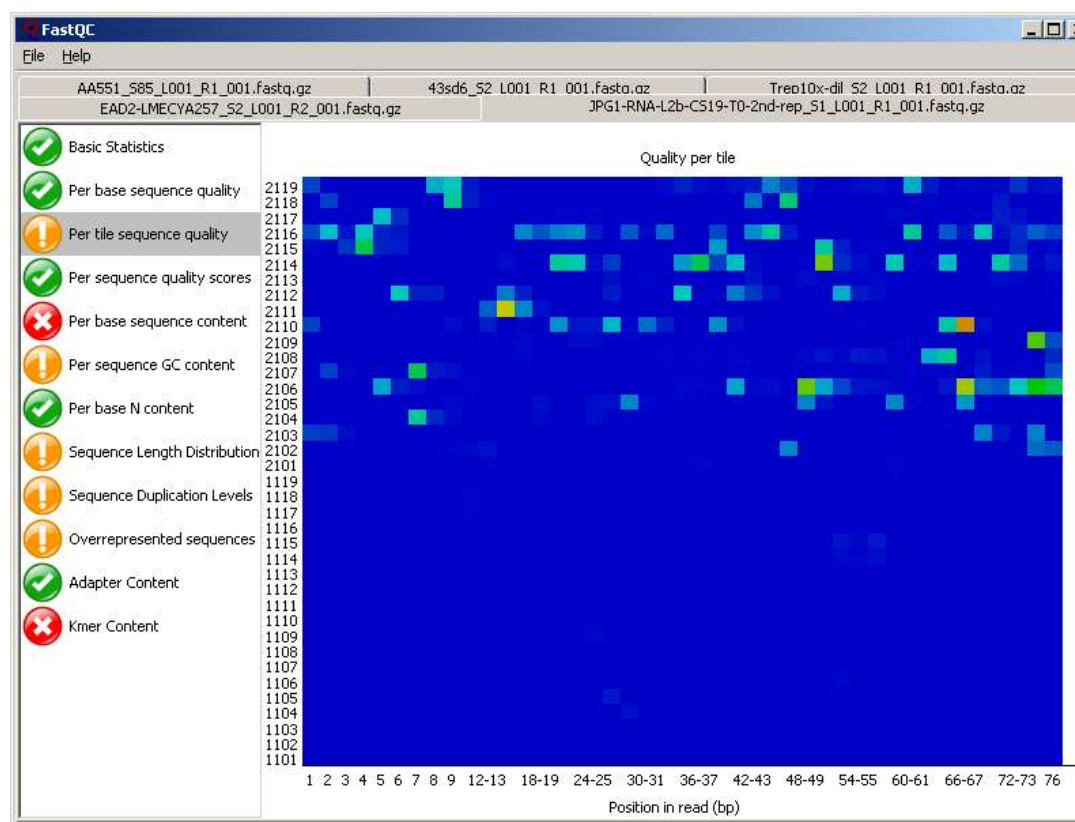
Per tile sequence quality

- Boa qualidade por *tile*/base



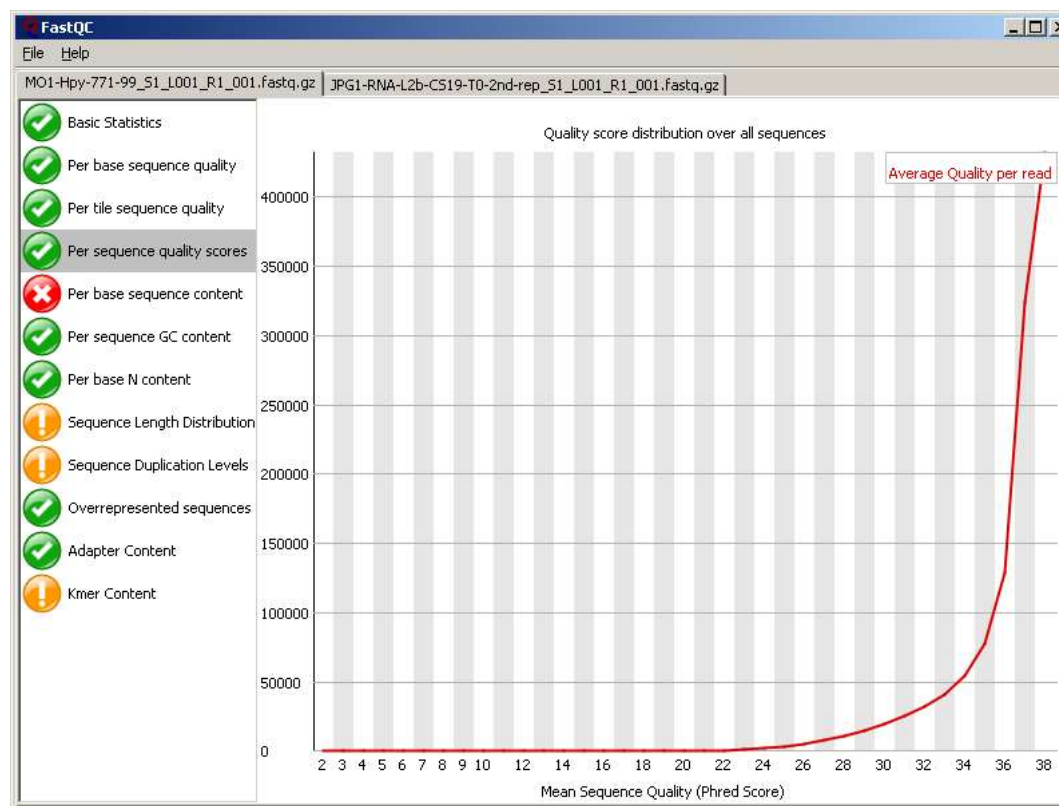
Per tile sequence quality

- Má qualidade em alguns *tiles*



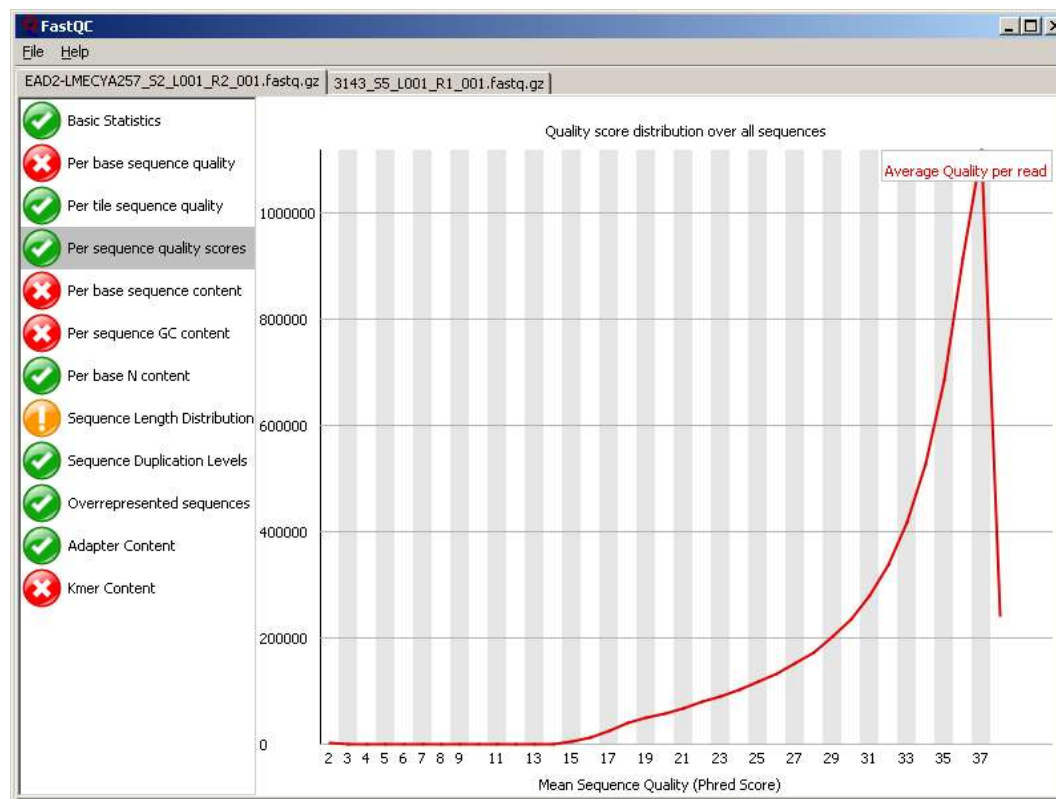
Per sequence quality scores

- Sequenciação de genoma – *reads* de qualidade elevada



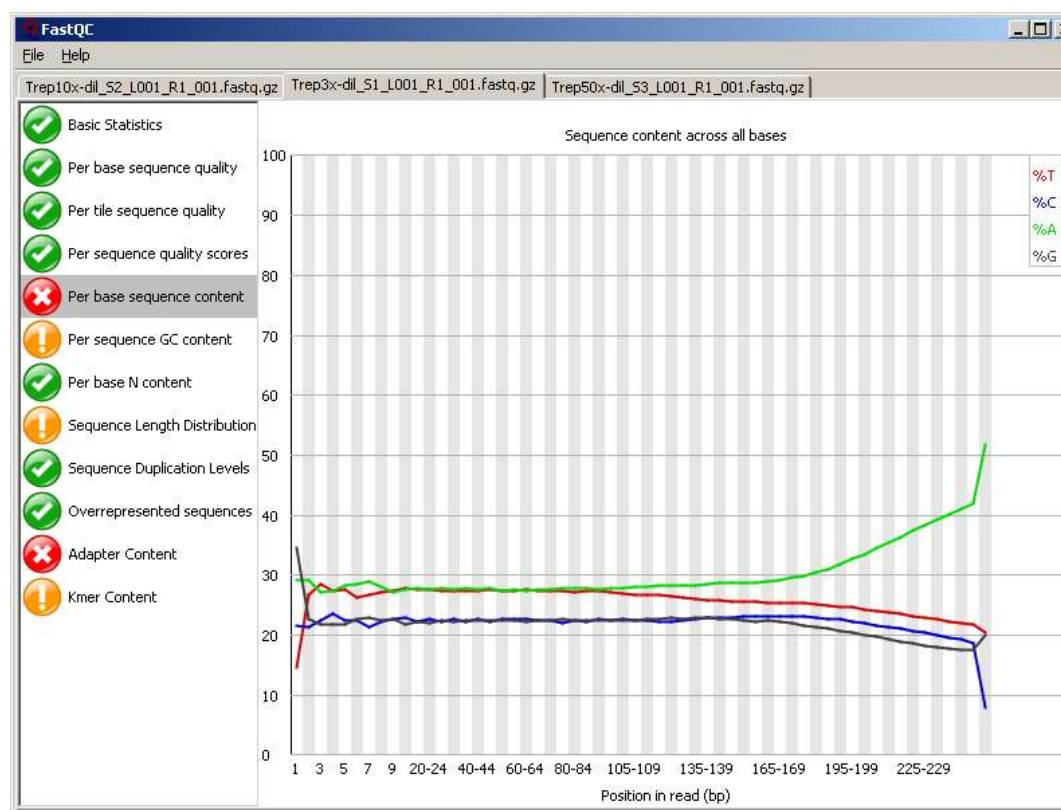
Per sequence quality scores

- Sequenciação de genoma – *reads* de qualidade baixa



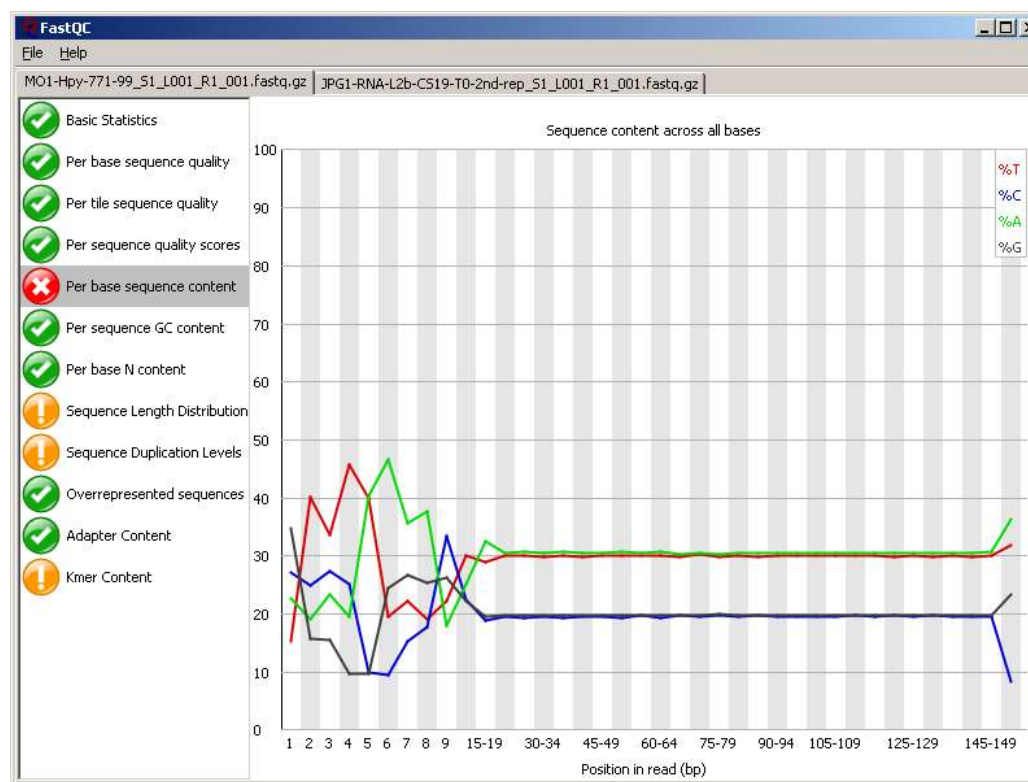
Per base sequence content

- Sequenciação de genoma fragmentado por sonicação



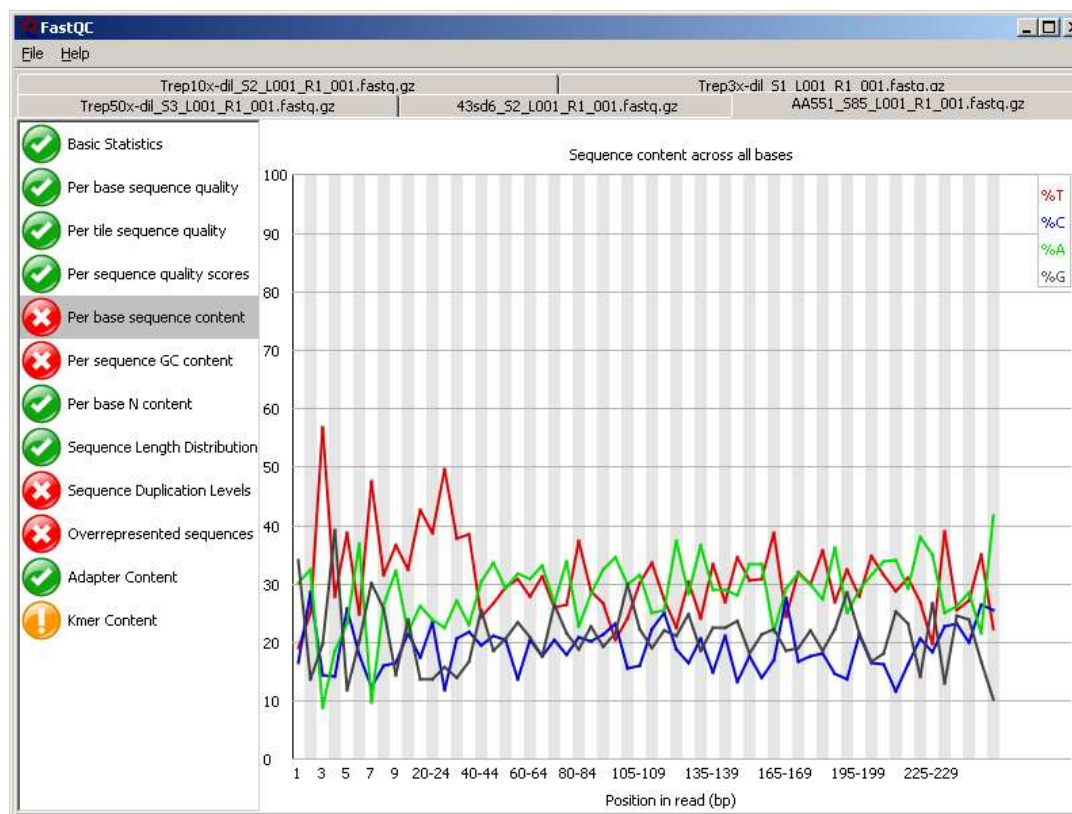
Per base sequence content

- Sequenciação de genoma fragmentado com transposases



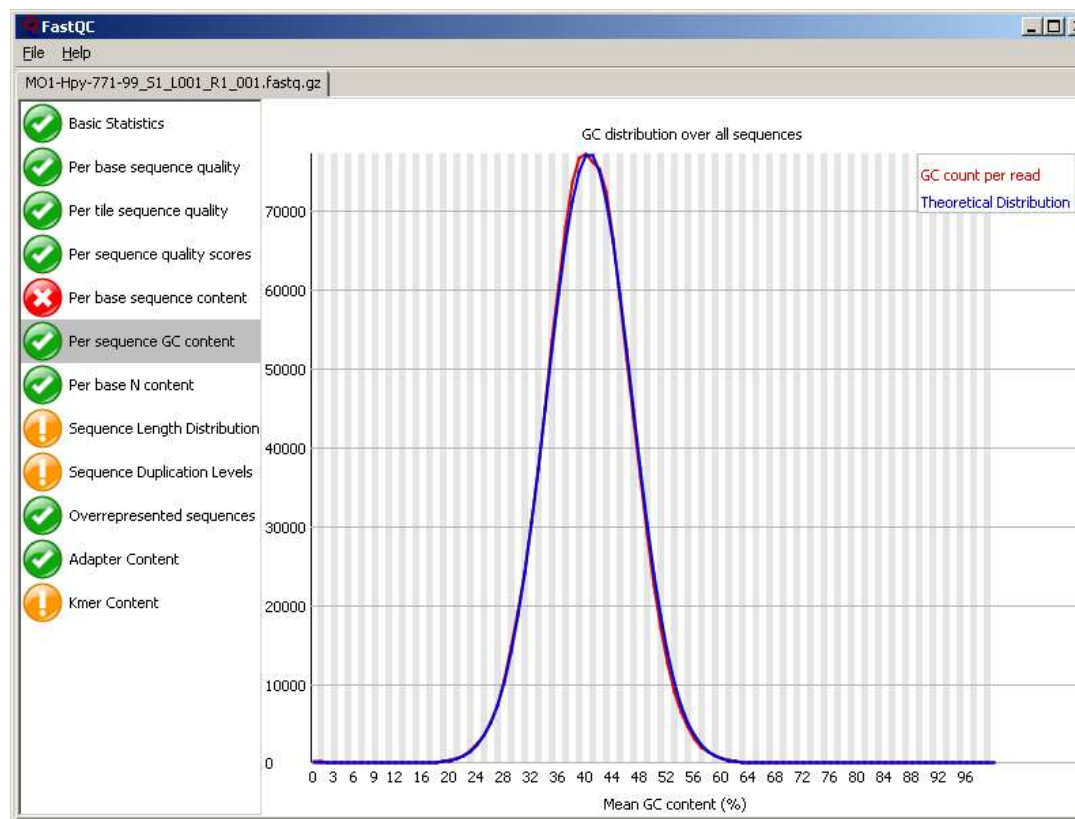
Per base sequence content

- Sequenciação de amplicões



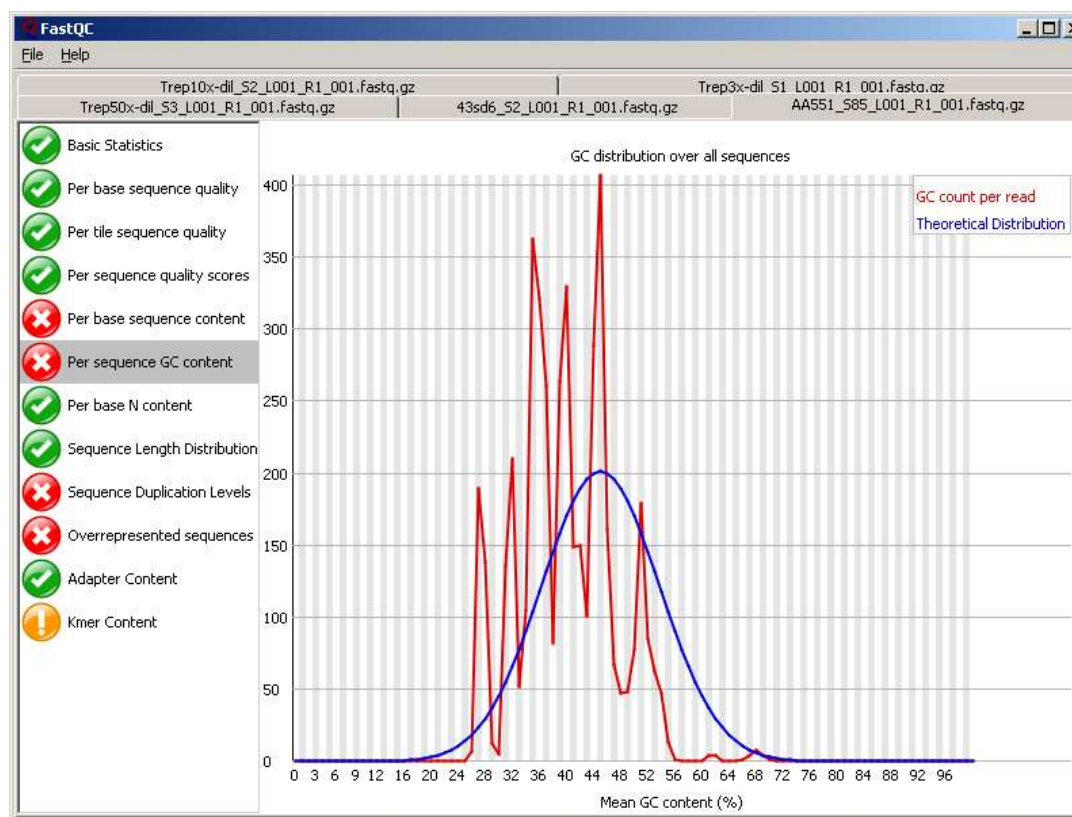
Per sequence GC content

- Sequenciação de genoma – biblioteca equilibrada



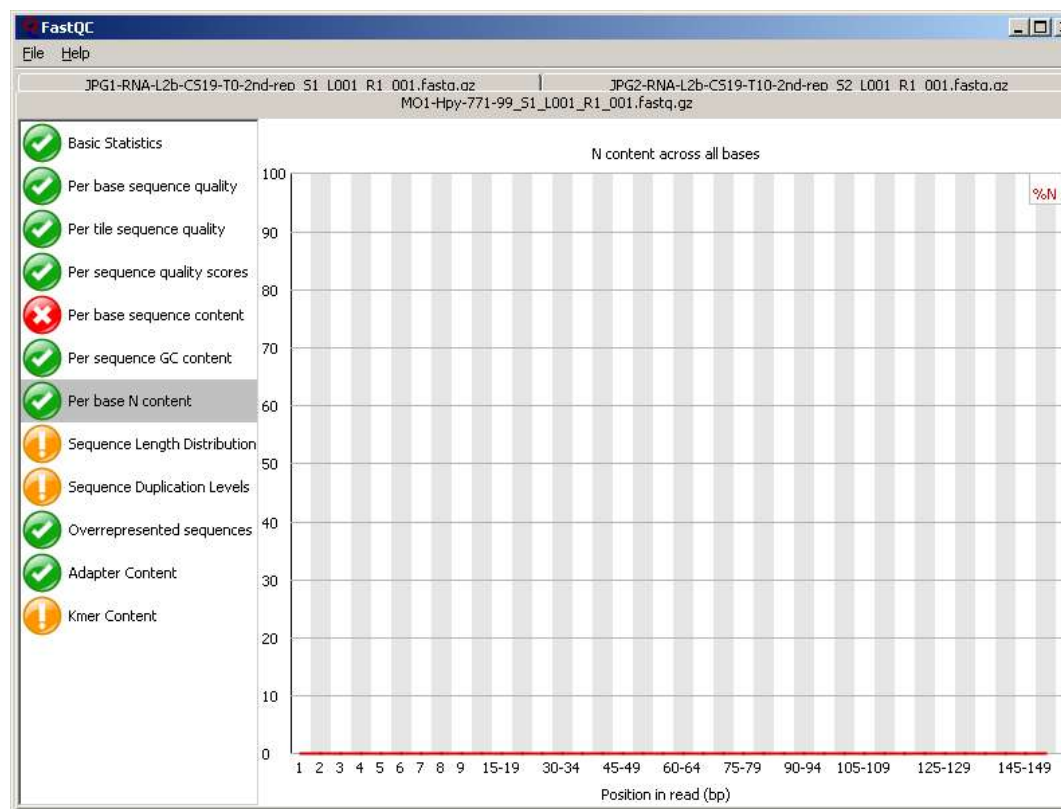
Per sequence GC content

- Sequenciação de amplicões – biblioteca desequilibrada



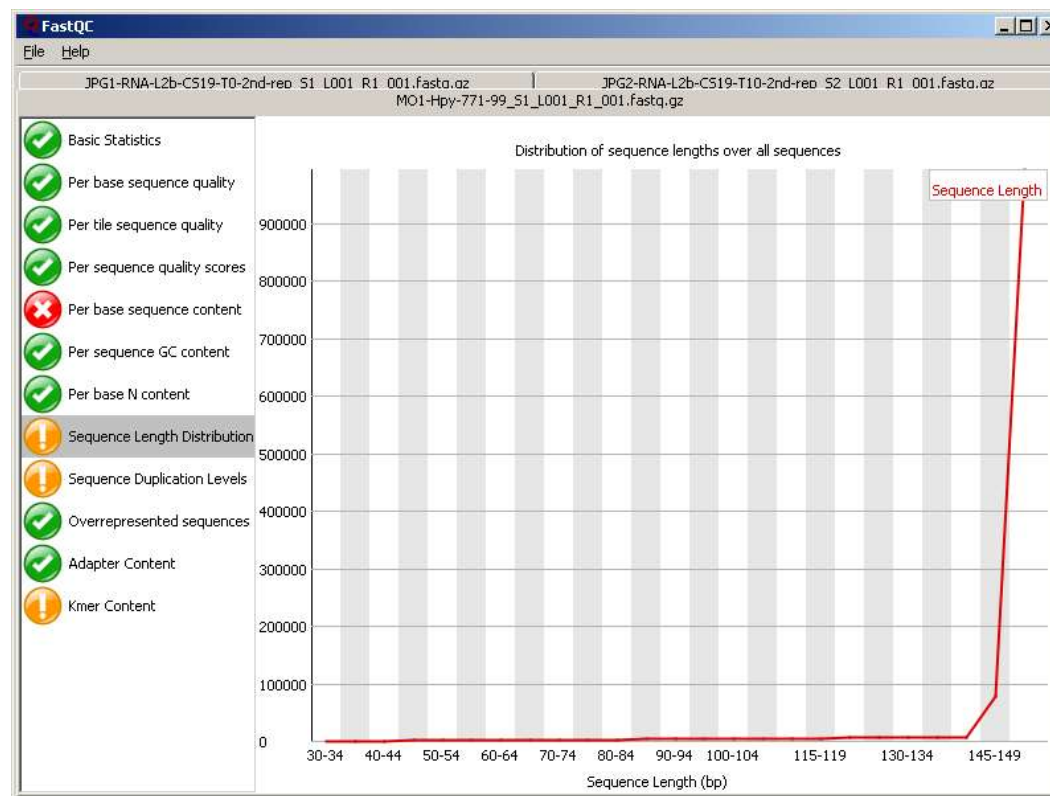
Per base N content

- Ausência de bases N ao longo da *read*



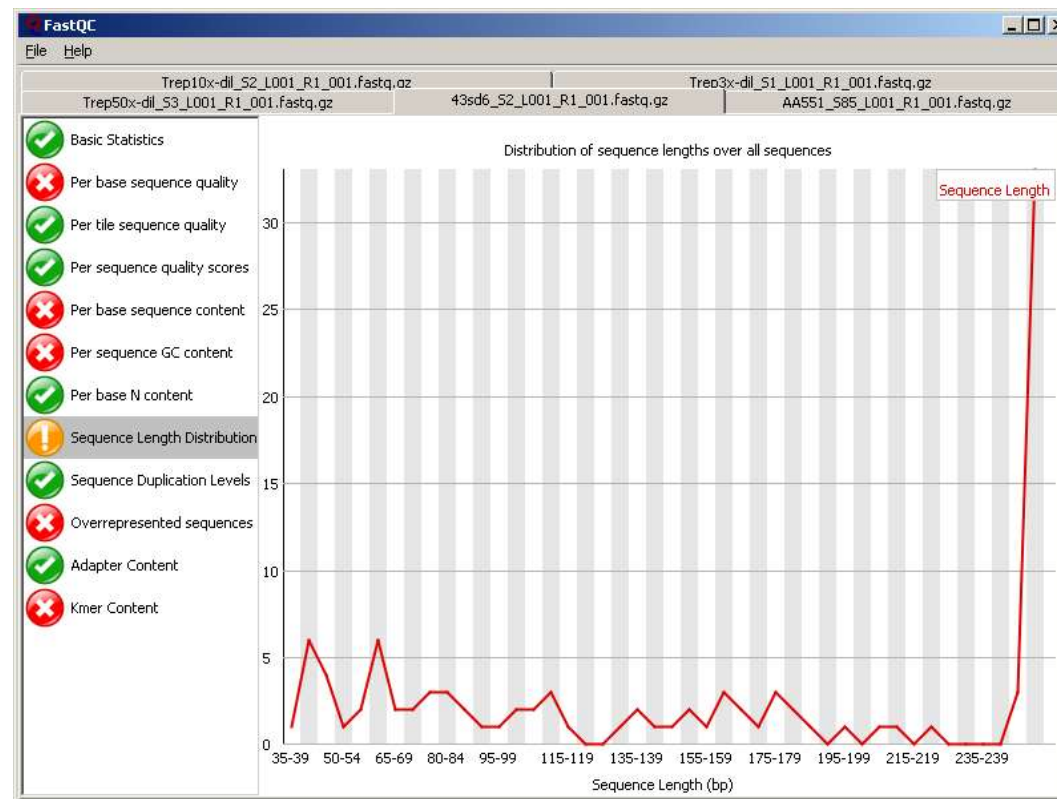
Sequence length distribution

- Reads com comprimento uniforme



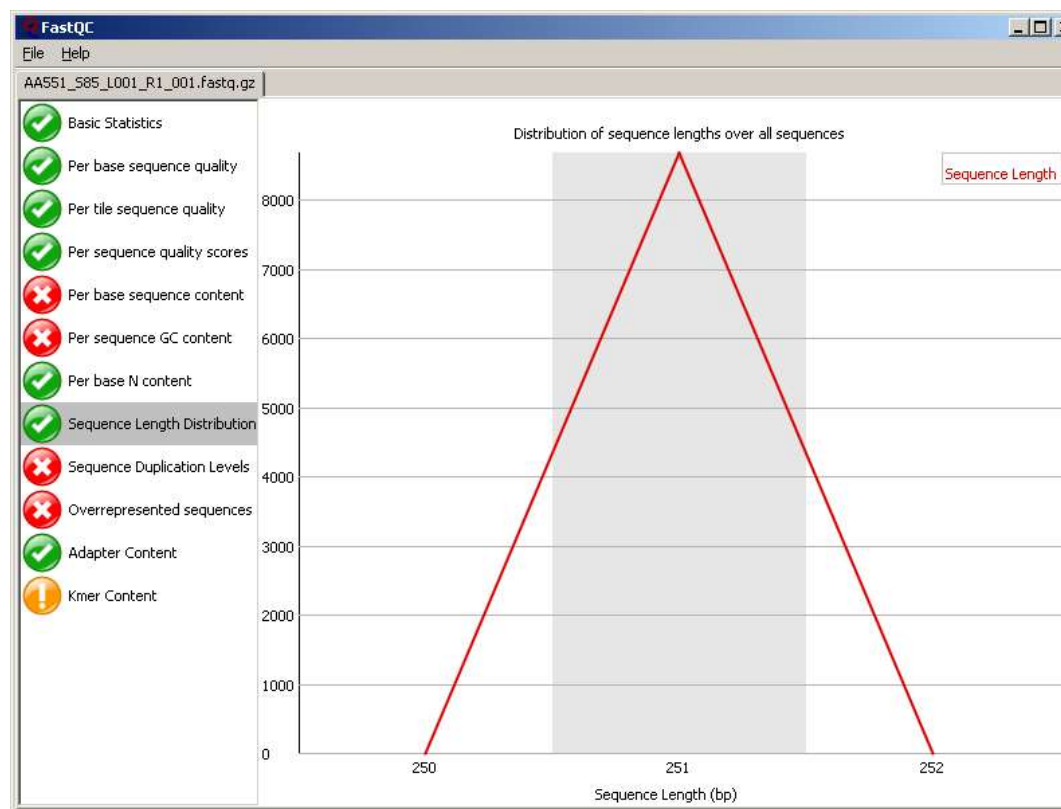
Sequence length distribution

- Reads com comprimentos diferentes



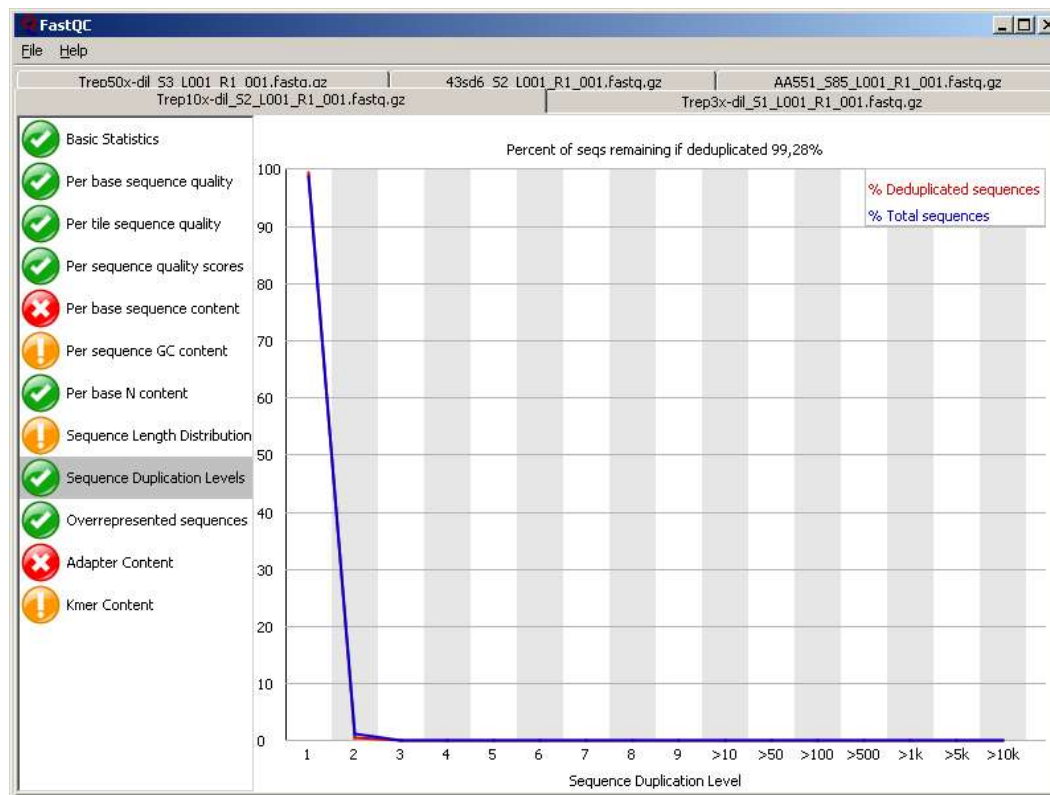
Sequence length distribution

- Reads com comprimento único



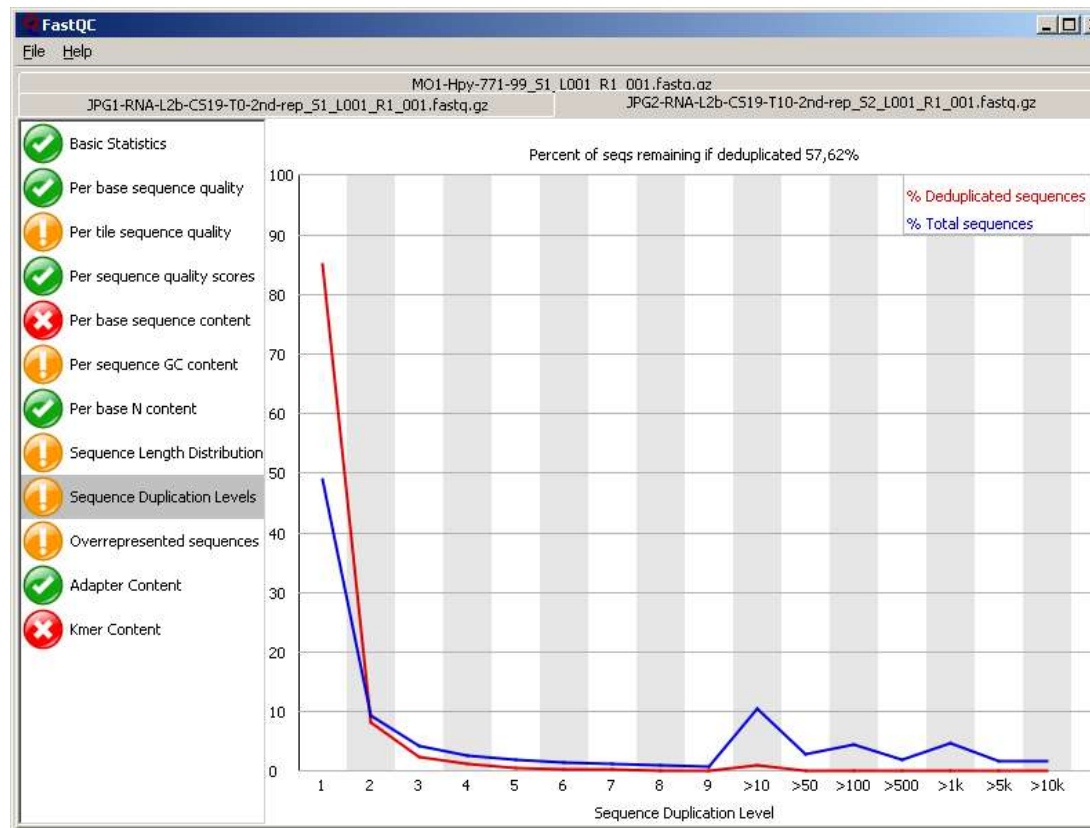
Sequence duplication levels

- Sequenciação de genoma



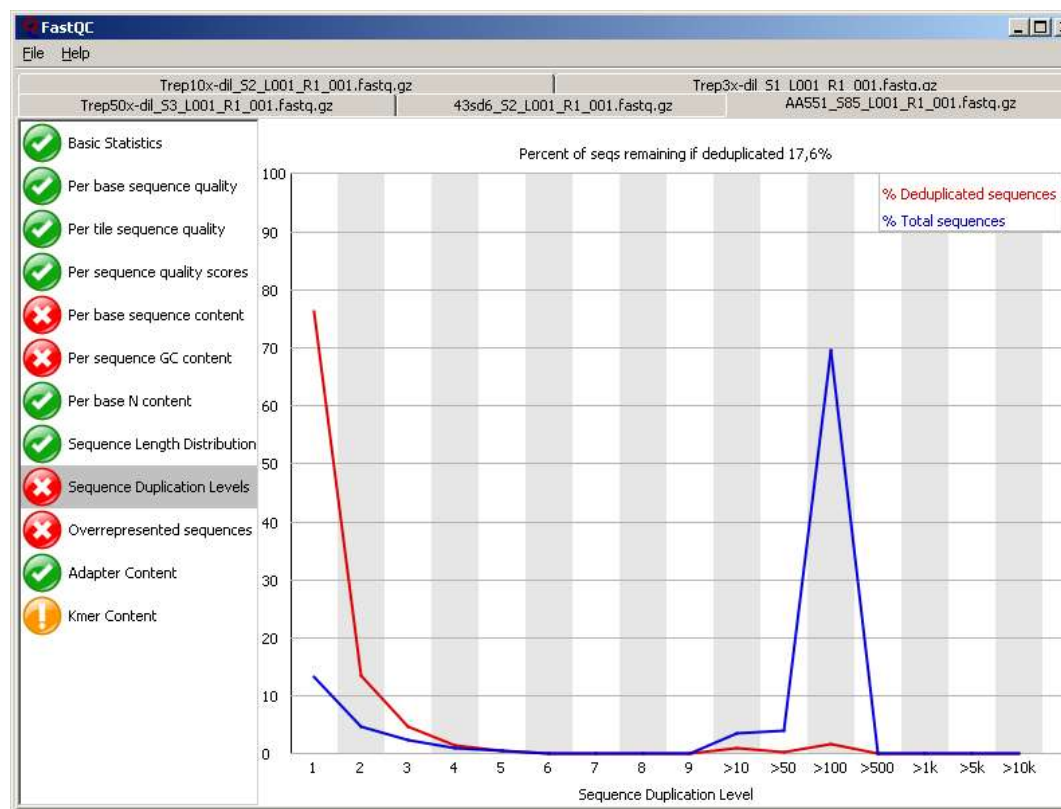
Sequence duplication levels

- Sequenciação de RNA



Sequence duplication levels

- Sequenciação de amoções



Overrepresented sequences

FastQC

File Help

MO1-Hpy-771-99_S1_L001_R1_001.fastq.gz JPG1-RNA-L2b-CS19-T0-2nd-rep_S1_L001_R1_001.fastq.gz

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GGGGTCTCGCTATGTTGCCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCG...	13545	0,164	No Hit
GTCTGGAGTCTTGGAAGCTTGACTACCCTACGTTCTCCTACAAATGGACC	13071	0,159	No Hit
GTGGCTATTCACAGGCGCGATCCCACTACTGATCAGCACGGGAGTTTTGA	12952	0,157	No Hit
CCCCCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCATAT	12896	0,156	No Hit
CTCCGTTTCCGACCTGGGCCGTTACCCCTCCTTAGGCAACCTGGTGGT	12758	0,155	No Hit
GCTCCGTTTCCGACCTGGGCCGTTACCCCTCCTTAGGCAACCTGGTGGT...	12356	0,15	No Hit
CCCTCCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCATATT	12250	0,149	No Hit
GGGGTCTCGCTATGTTGCTCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCG...	11658	0,141	No Hit
GGATGTGTCTGGAGTCTTGGAAGCTTGACTACCCTACGTTCTCCTACAAA	10009	0,121	No Hit
CCTCCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCATATTGAT...	8871	0,108	No Hit
CCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCATATTGATG	8548	0,104	No Hit

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

Sequence Duplication Levels

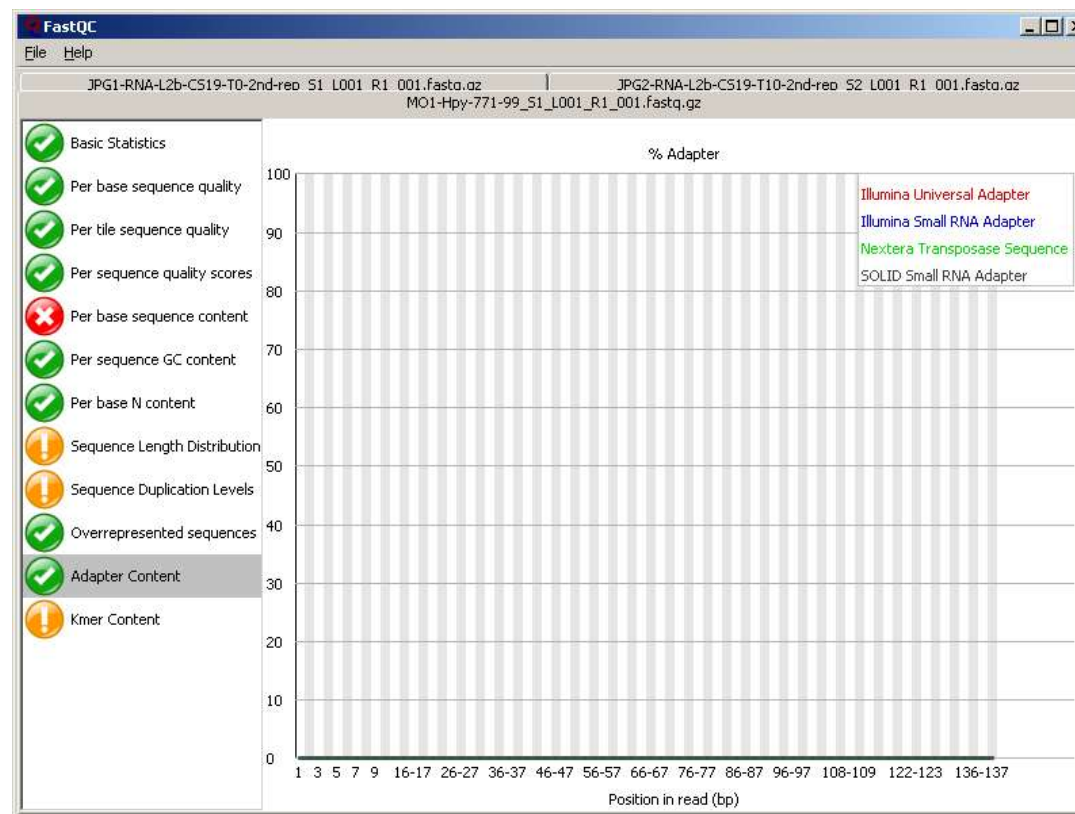
Overrepresented sequences

Adapter Content

Kmer Content

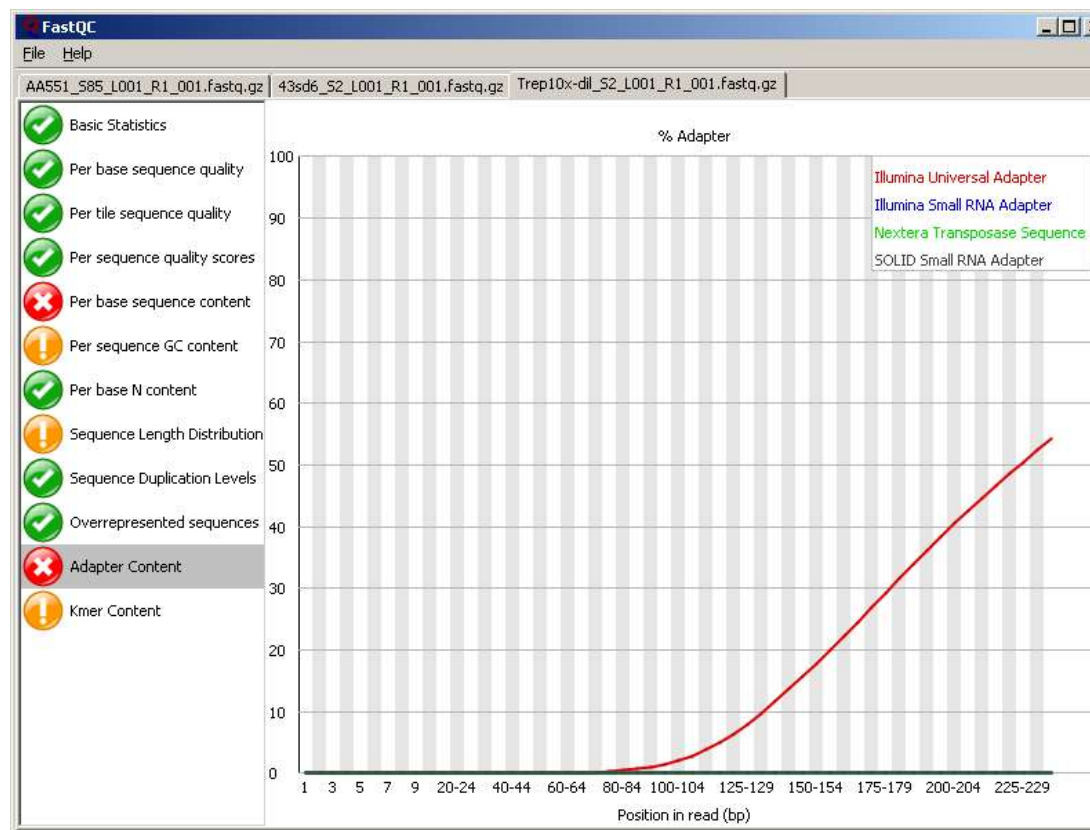
Adapter content

- Ausência de sequências de adaptadores



Adapter content

- Presença de sequências de adaptadores Illumina



Bibliografia

- **Download**

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

- **Manual do FastQC**

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/>

- **Videos online**

- “Using FastQC to check the quality of high throughput sequence”:

<https://www.youtube.com/watch?v=bz93ReOv87Y>

- “Fastqc Linux Install and Usage (Commandline & GUI)”:

<https://www.youtube.com/watch?v=5nth7o-f0Q>