CURSO DE CURTA DURAÇÃO - 2017

BIOINFORMÁTICA

BioME – CENTRO MULTIUSUÁRIO DE BIOINFORMÁTICA - UFRN

NEXT GENERATION SEQUENCING

Análise de Dados de Sequenciadores de Segunda Geração

Prof. Dr. JORGE ESTEFANO SANTANA DE SOUZA

E-mail: jorge@imd.ufrn.br

Bioinformatics Multidisciplinary Environment | Centro Multiusuário de Bioinformática

2Bio Instituto de Bioinformática e Biotecnologia

IMD Instituto Metrópole Digital | PPG BioInfo UFRN | UFRN Universidade Federal do Rio Grande do Norte

ISO-BIOINFO

Análise de Qualidade

Objetivo:

Utilizar as ferramentas básicas de analise de qualidade para obter um perfil inicial da qualidade do sequenciamento.

Comandos Básicos de Linux:

Para trabalhar com nossos dados, vamos precisar saber alguns comandos básicos do Linux. Podem procurar mais informação no site:

http://wiki.ubuntubr.org/ComandosBasicos

Ferramentas:

1- Linux.
2- WebServer.

3- fastq_screen
4- fastqc
5- samstat
6- DynamicTrim.pl
7- trim_galore
8- cutadapt

Inicial:

Login maquina local:

Login:

Senha:

Login no server:

ssh -p 4422 bif@10.7.5.38

Senha: bif0003

Inicial:

Pasta com dados iniciais:

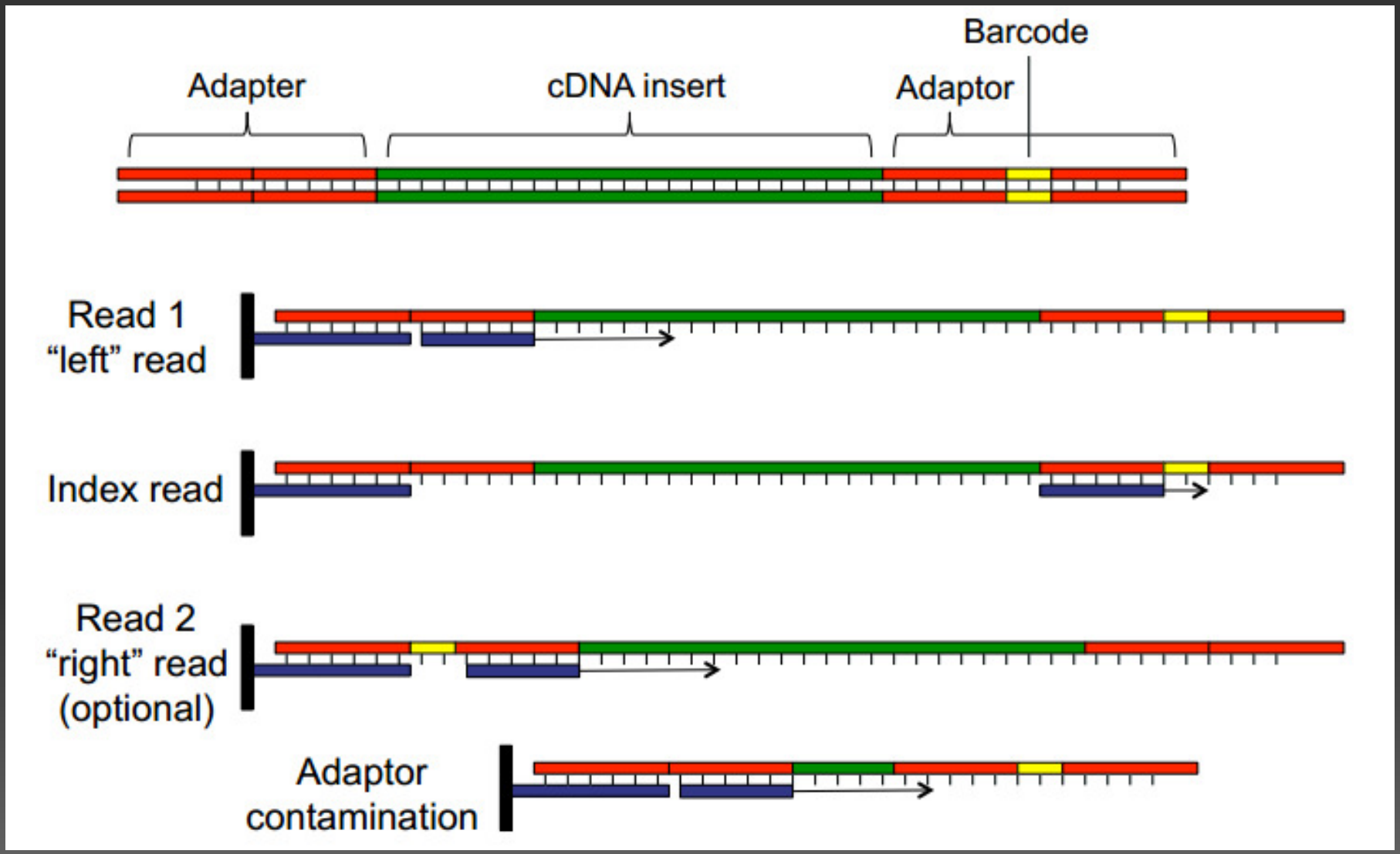
/home/treinamento/NGS/

Pasta servidor WEB:

/home/bif/public_html/

Algumas nomenclaturas e formatos basicos

Sequencing


Barcode
Adapter
cDNA insert
Adaptor
Read 1
“left” read
Index read
Read 2
“right” read
(optional)
Adaptor
contamination

Experimental Design

Biological comparisons.

Paired-end, Single-end Reads.

Read Length.

Deep sequencing.

Biological and experimental replicates.

- Illumina sequencing by synthesis
  - GAIIx
    - replaced by HiSeq
  - HiSeq2000
  - MiSeq
    - low throughput, fast turnaround
- SOLiD (not available at BMGC)
  - "Color-space" reads (require special mapping software)
  - Low error rate
- 454 pyrosequencing
  - Longer reads, lower throughput

What are the goals?
- Somatic alterations.
- SNPs.
- Structural variations.

What are the characteristics of my system?
- Complex genome, much?
- Well annotated?

Fragment size?

Barcode or Lane?

Samples per Lane?

Fastq:

```
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)
```

- **Fastq format** (Illumina Casava 1.8.0)— ⚠ Formats vary
  - 4 lines per read

Machine ID

QC Filter flag
Y=bad
N=good

barcode

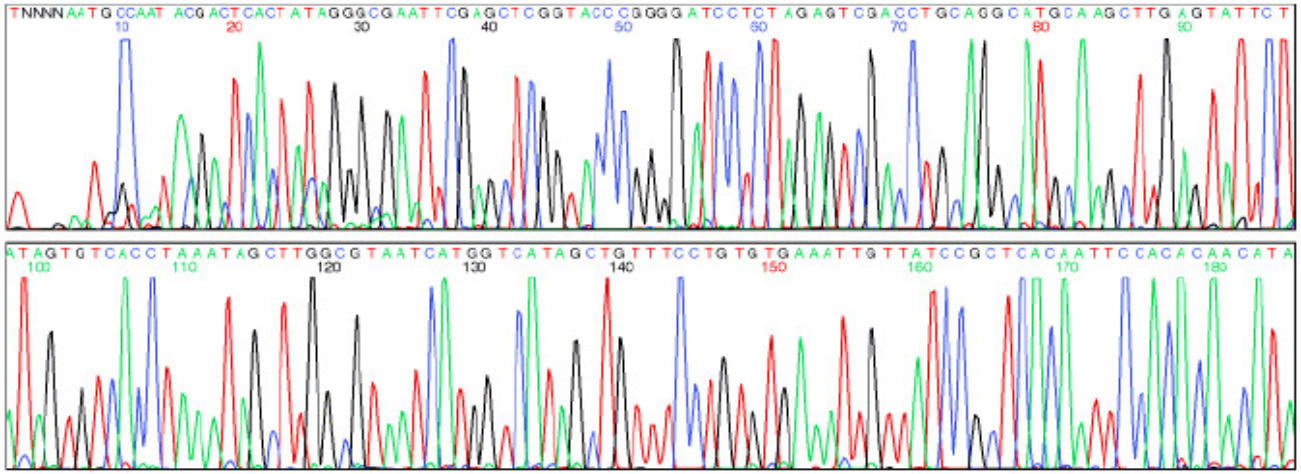
Read ID → @HWI-M00262:4:000000000-A0ABC:1:1:18376:2027 1:N:0:AGATC
Sequence → TTCAGAGAGAATGAATTGTACGTGCTTTTTTTGT
\+ → +
Quality score Phred+33 → =1:?7A7+?77+<<@AC<3<,33@A;<A?A=:4=

Read pair #

Dec	Hx	Oct	Char		Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL	(null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH	(start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX	(start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX	(end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT	(end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ	(enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK	(acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL	(bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS	(backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB	(horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF	(NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT	(vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF	(NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR	(carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO	(shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI	(shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE	(data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1	(device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2	(device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3	(device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4	(device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK	(negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN	(synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB	(end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN	(cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM	(end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB	(substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC	(escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS	(file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS	(group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS	(record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US	(unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

Interrupção da cadeia / Sanger



Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

RAW DATA

Fasta

```
>seq1
CTAGCTGAGCATCGGTAGCTAGCTGAGGTAGCTAG
>seq2
CTAGCTGAGCATCGTAGCTACGTAGTAGCTAGCTG
>seq3
AGCTACGTAGCTAGCTGAGCATCGTAGCTACGTAT
>seq4
ATGTCACGACGAGCATCGTAGCTACGTAGCTAGCT
```

csFasta

```
>186_2041_1641_F3
T122233110.3012011122133012030.1110.31220022220.120
>186_2041_1706_F3
T111321213122013212201032301 23.2113.31201112230.031
>186_2041_1709_F3
T2103022220322301123212223030330323320201102233.123
```

			2nd base		
		A	C	G	T
	A	0	1	2	3
1st	C	1	0	3	2
base	G	2	3	0	1
	T	3	2	1	0

Fastq

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;;7;;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;;9;7;;.7;393333
```

Map Reads

1 - Download Genome. (NC_012967.1)

2 - Sorting the Chromosomes.

```
>chr1
GAATTCTCTGTAACACTAAGCTCTCTTCCTCAAAACCAGAGGTAGATAGA
ATGTGTAATAATTTACAGAATTTCTAGACTTCAACGATCTGatttttaa
atttatttttatttttCAGGTTGAGACTGAGCTAAAGTTAATCTGTGGC
GACGTTCTGGATGTACTGGACAAACACCTCATTCCAGCAGCTACAACTGG
CAAGTCCAAGGTTTTCTATCATGAAATGTAGGTTCTATACTAACAATTAA
CAAGTGTACTTCAATAAATTTAAACATTCTCAGGAATAGTTGACTTTGTT
TCTTTTTTTCTTAGACATTTCATATTATTTTCCTTATTAAATATAACCAA
```

3 - Index the Genome.

SAM Format Specification

Map Reads

1.1 An example

Suppose we have the following alignment with bases in lower cases clipped from the alignment
r001/1 and r001/2 constitute a read pair; r003 is a chimeric read; r004 represents a split a

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                         CAGCGCCAT
```

Col	Field	Brief description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string
7	RNEXT	Ref. name of the mate/next segment
8	PNEXT	Position of the mate/next segment
9	TLEN	observed Template LENgth
10	SEQ	segment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33

The corresponding SAM format is:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

https://broadinstitute.github.io/picard/explain-flags.html

Quality control

Assess the quality of reads.

- Identify contaminants.

- Identify samples with low performance sequencing.

Softwares:

- **FastQ_Screen.**

- **FastQC.**

- **SAMStat.**

Fastq:

jorge@jorge-virtual: ~/aula02/fastq

```
@HWUSI-EAS1881:8:RNASeqAline:1:1:14230:1008 1:N:0:NGATGTA
NCAAGAGATAGAAAGACCAGTCCTTGCTGAAAGACAAGTCNNAANNNNCNNNNNNNTNNNNNNNNNNNTNNNN
+
#////66556C@@CC@CC@@CC@@@@CC@@CC@@C##########################################
@HWUSI-EAS1881:8:RNASeqAline:1:1:16075:1009 1:N:0:NGATGTA
NGAGGATCTGCTTGAGAACTACGACAACGTGTGCACGTTGNNGTNNNNGNNNNNNNANNNNNNNNNNNTNNNN
+
#.,,,66666@CC@@@@@C@@@C@C@@@C@@C@CC##########################################
@HWUSI-EAS1881:8:RNASeqAline:1:1:17360:1009 1:N:0:NGATTAT
NCAGGATGGATTTTAGATCTTGTTGAAAGCAGCCACATCCNNGGNNNNCNNNNNNNCNNNNNNNNNNNGNNNN
+
#2.2.66666CCCC@@@@@CC@CC@@CC@C@CC@C##########################################
@HWUSI-EAS1881:8:RNASeqAline:1:1:18765:1009 1:N:0:NGATGTA
NCACACCATATATTTACAGTAGGAATAGACGTAGACACACNNGCNNNNTNNNNNNNCNNNNCNNNNNNTNNNN
+
#,+))22232@@C@@@@@@@@@@@@@@@C@@@@@@##########################################
@HWUSI-EAS1881:8:RNASeqAline:1:1:11153:1009 1:N:0:NGATGTA
NCCACATCTACAAAATGCCAGTATCAGGCGGCGGCTTCGANNGCNNNNGNNNNNNNTNNNNNNNNNNNTNNNN
+
#---+66656C@@@22@@CCCC@@@@@CC@CCC@@##########################################
```

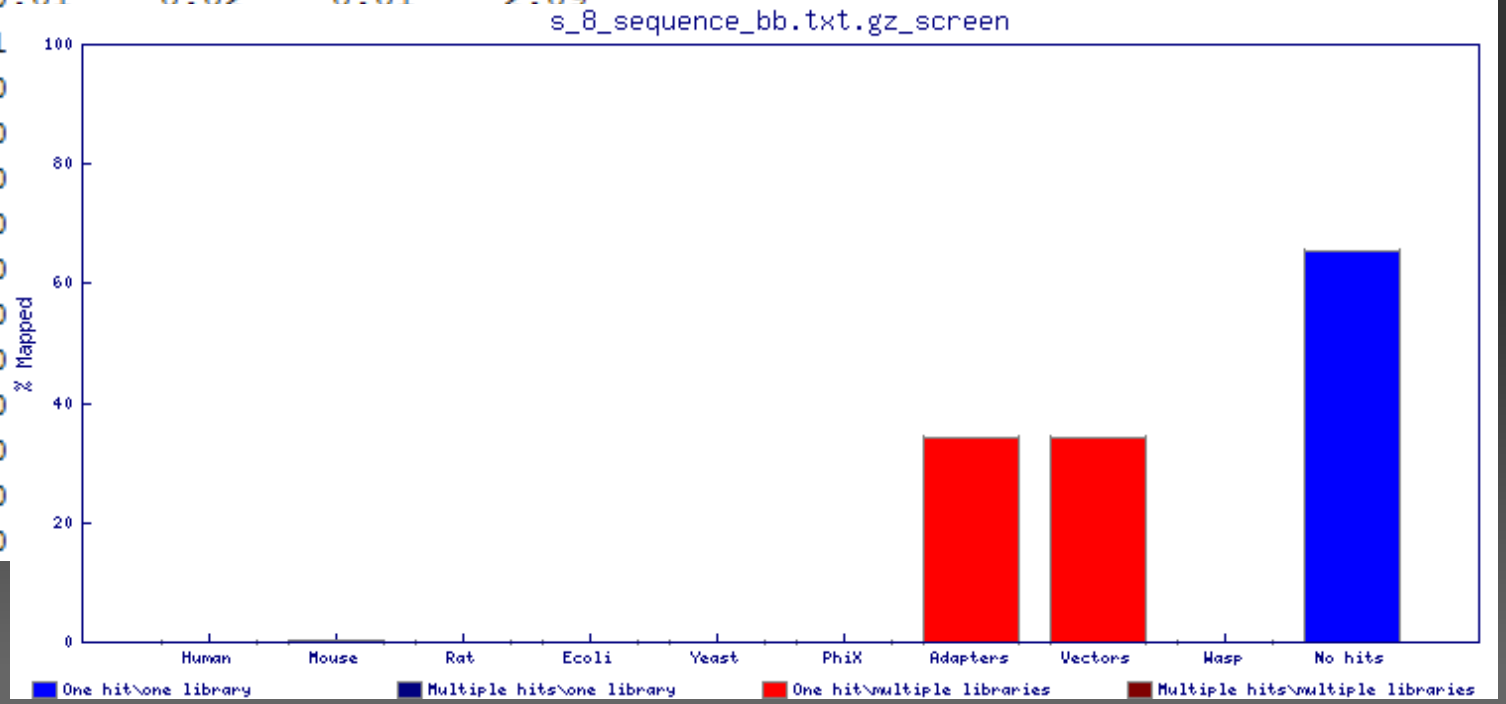
NEXT

Run FastQ_Screen:

Quality control

```
fastq_screen –nohits –subset 0  /home/treinamento/NGS/ERR844339.fastq --outdir .
```

```
#Fastq_screen version: 0.4
Library          %Unmapped   %One_hit    %Multiple_hits  %One_hit_multiple   %Multiple_hits_mult
18S              99.77   0.00    0.00    0.00    0.23
28S              99.53   0.00    0.00    0.02    0.45
45S              99.27   0.00    0.00    0.73    0.00
5S               100.00  0.00    0.00    0.00    0.00
Contaminants     97.87   0.01    0.02    0.01    2.09
Vectors          95.80   1
Ribosomal        99.27   0
C_elegans        97.49   0
Ciona_intest     97.61   0
E.coli           100.00  0
Phix             100.00  0
S_cerevisiae     97.64   0
SalmonellaTyphi  100.00  0
Shigellaflex     97.98   0
Yersinia_pest    100.00  0
ALU              97.62   0
Virus            97.66   0
```



Run FastQ_Screen:

Quality control

```
fastq_screen –nohits –subset 0  /home/treinamento/NGS/ERR844339.fastq --outdir .
```

```
# This is a configuration file for fastq_screen

############
## Bowtie #
############
## If the bowtie binary is not in your PATH then you can
## set this value to tell the program where to find it.
## Uncomment the line below and set the appropriate location
##

#BOWTIE /usr/local/bin/bowtie2
BOWTIE2 /usr/local/bin/bowtie2
```

```
##virus
DATABASE        virus       /home/databases/virus/virus     BOWTIE2

##ribossomal
DATABASE        ribosomal   /home/databases/ribosomal/ribosomal BOWTIE2

##Ecoli- sequence available from EMBL accession U00096.2
DATABASE        E.coli      /home/databases/E.coli/E.coli  BOWTIE2
```

Run FastQ_Screen:

Quality control

```
fastqutils stats  ERR844339_no_hits.fastq   | more

Total:   ???
```

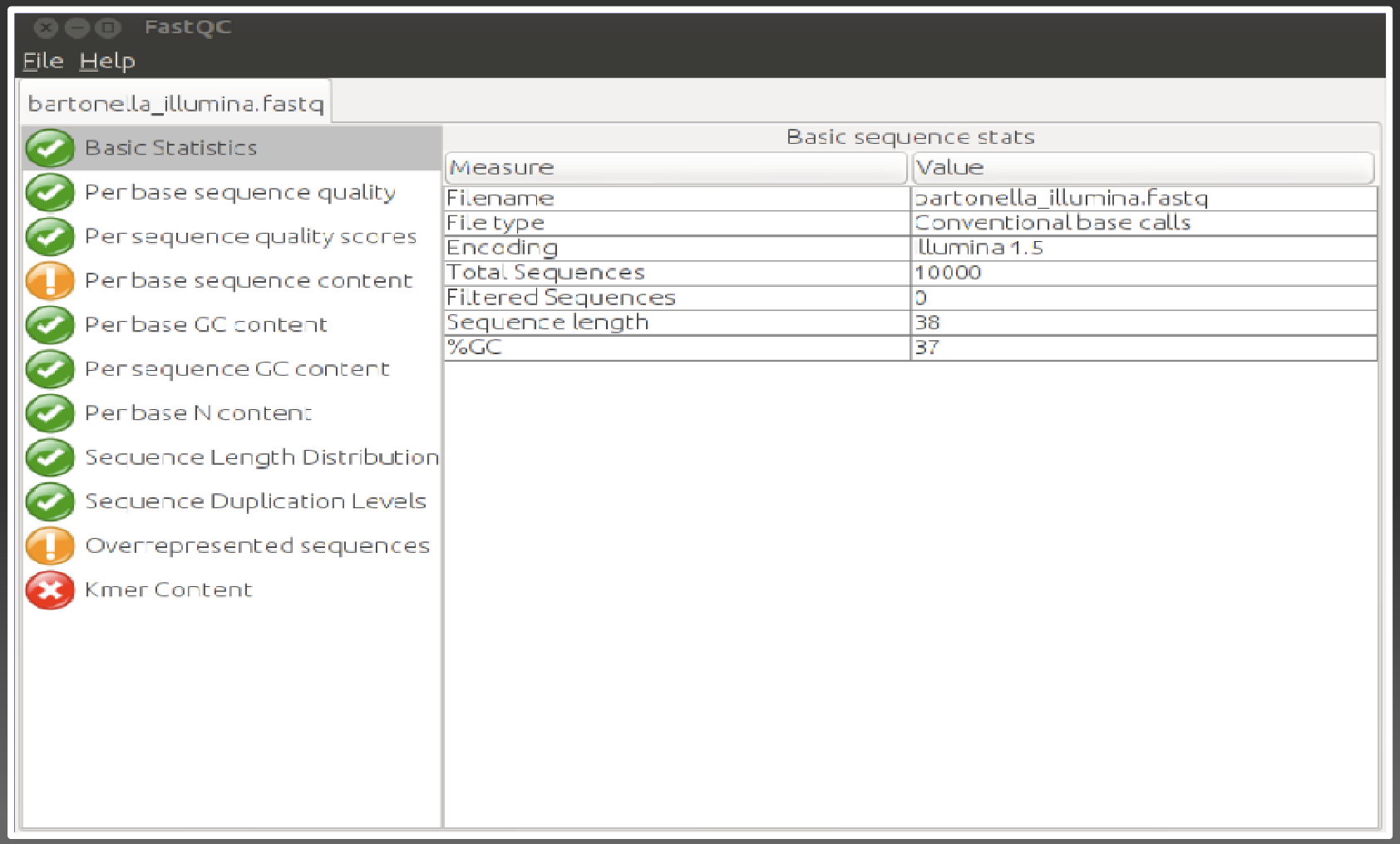
```
fastqutils stats  /home/treinamento/NGS/ERR844339.fastq  | more

Total:   ???
```

NEXT

Run FastQC:

Quality control

```
fastqc /home/treinamento/NGS/ERR844339.fastq    -o  .
```



FastQC:

Quality control

Good

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

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

Bad

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

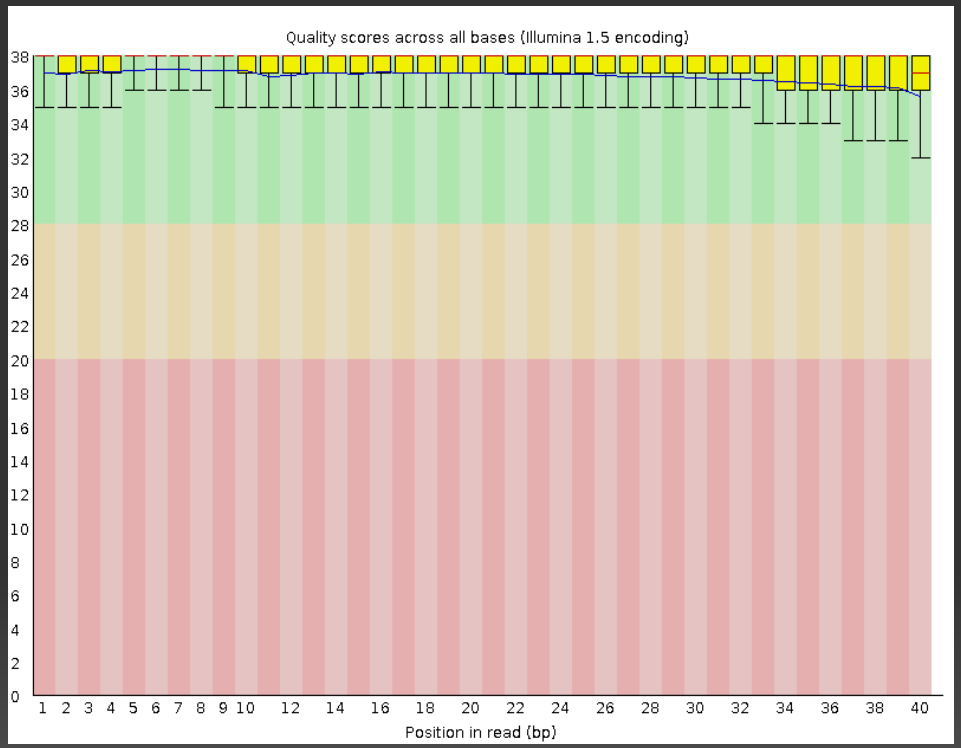
Basic Statistics

Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequence length	40
%GC	47

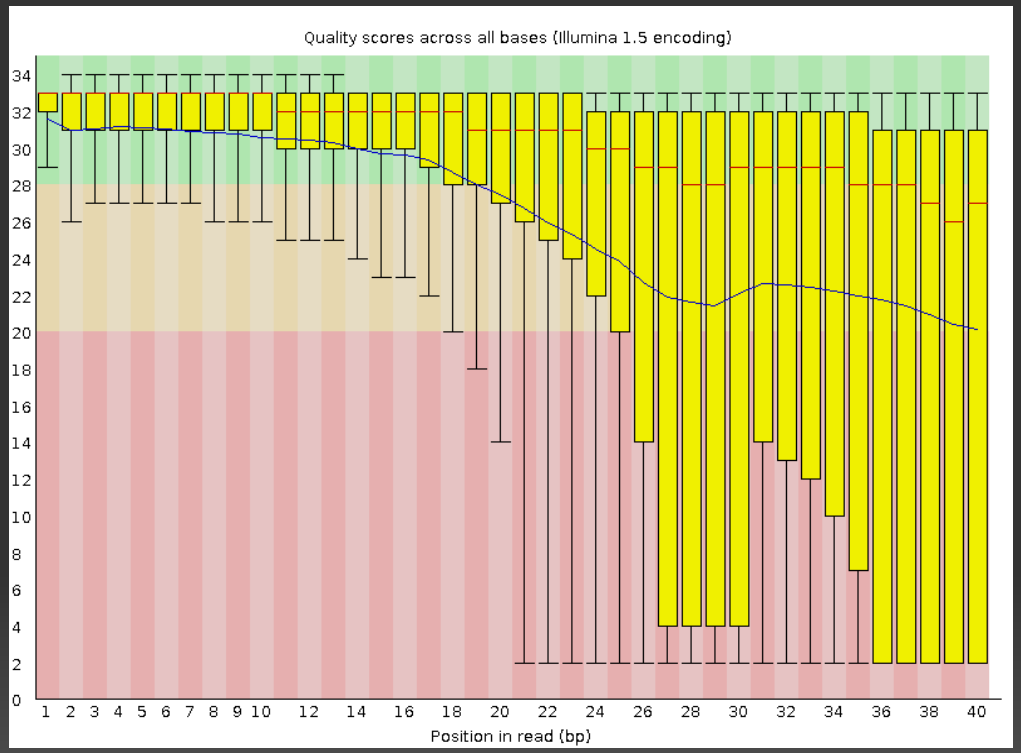
FastQC:

Quality control

Good



Bad



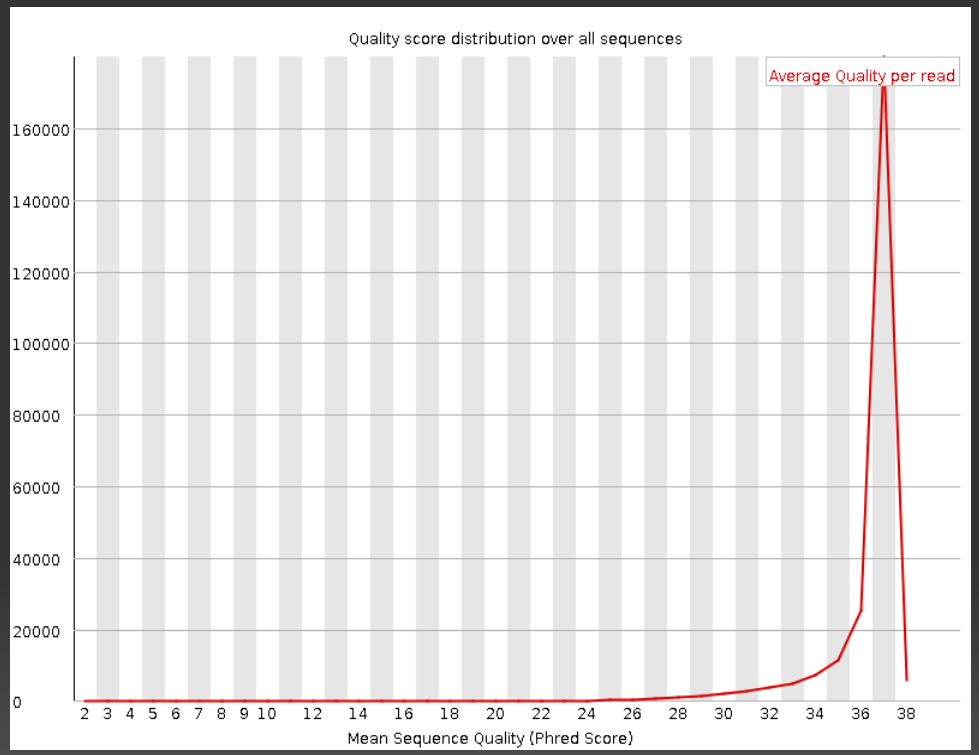
Per base sequence quality

Per base sequence quality

FastQC:

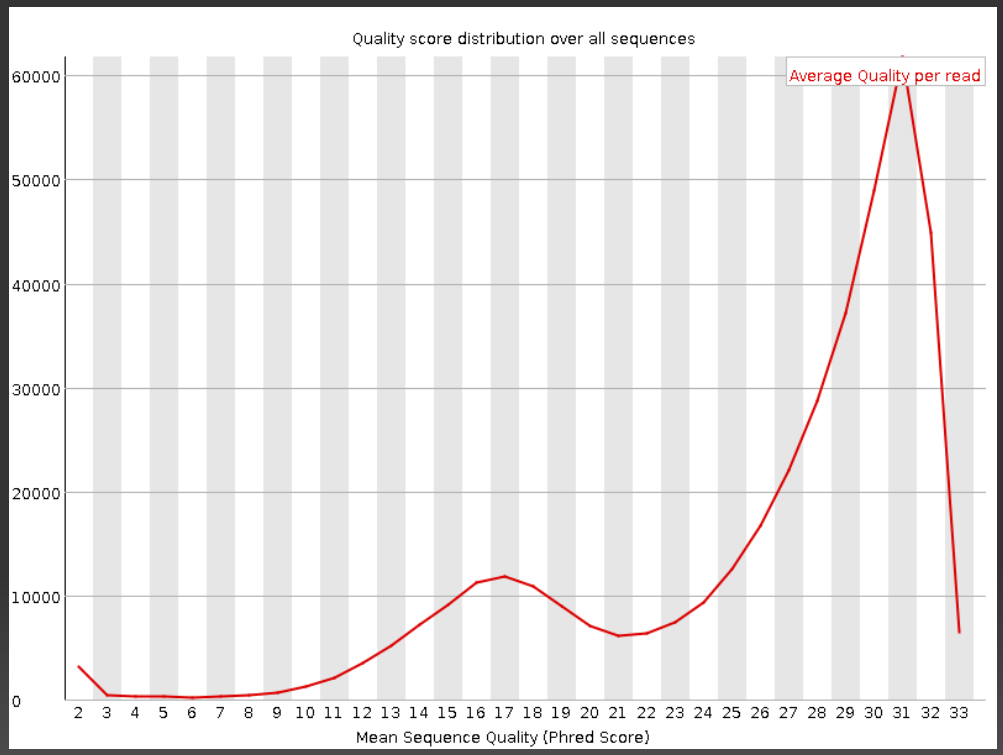
Quality control

Good



Per sequence quality scores

Bad

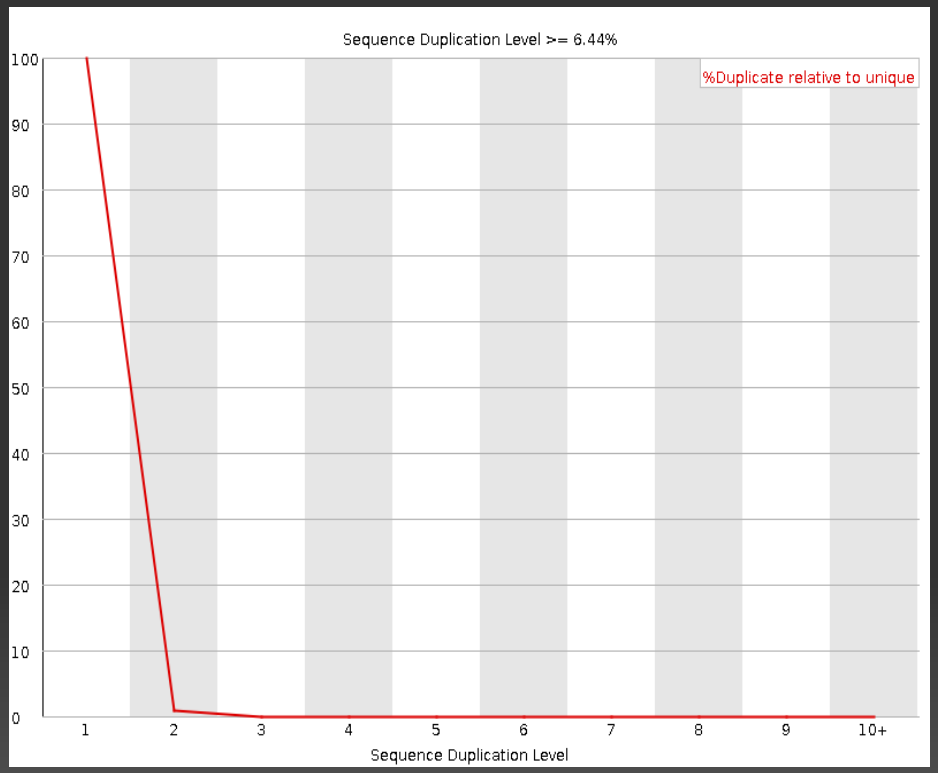


Per sequence quality scores

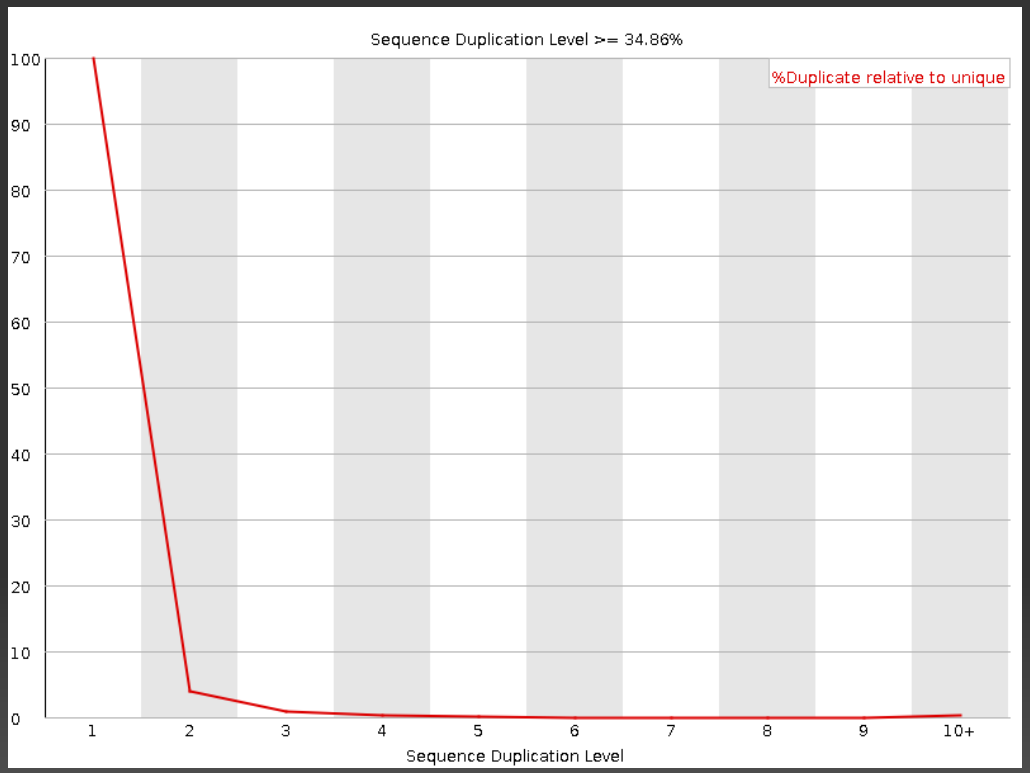
FastQC:

Quality control

Good



Bad



Sequence Duplication Levels

Sequence Duplication Levels

FastQC:

Quality control

Good

No overrepresented sequences

Bad

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAA	1836	0.46447147396328753	No Hit
GATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1831	0.4632065734350651	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.45005160794155147	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	1779	0.45005160794155147	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCC	1760	0.4452449859343061	No Hit
AAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTT	1729	0.4374026026593269	No Hit
CGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAG	1713	0.43335492096901496	No Hit
ATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAG	1708	0.43209002044079253	No Hit

Overrepresented sequences

Overrepresented sequences

Run FastQC:

Quality control

```
fastqc /home/treinamento/NGS/ERR844339.fastq   -o  .
```

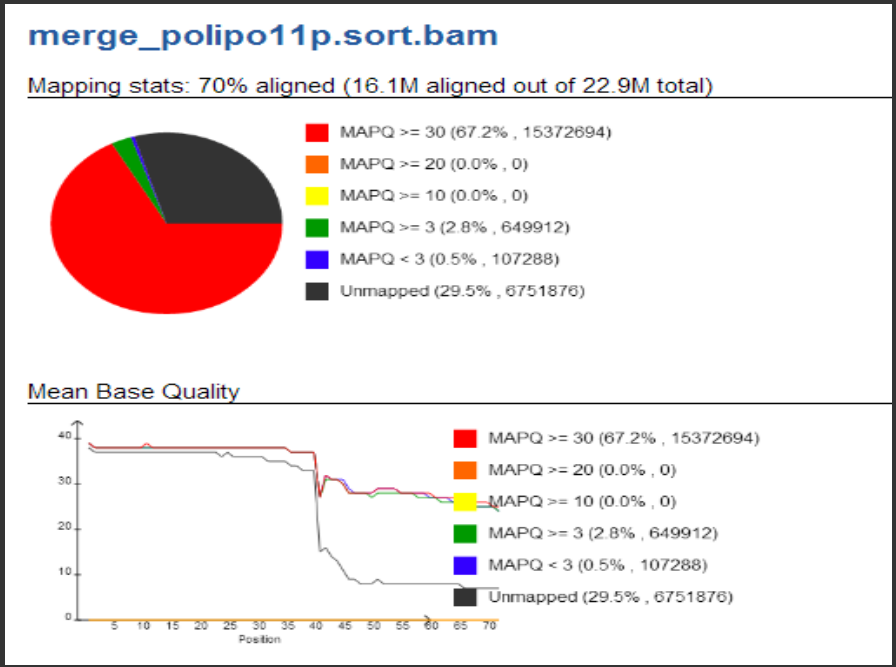
```
fastqc /home/treinamento/NGS/10_S5_R1_001.fastq   -o .
```

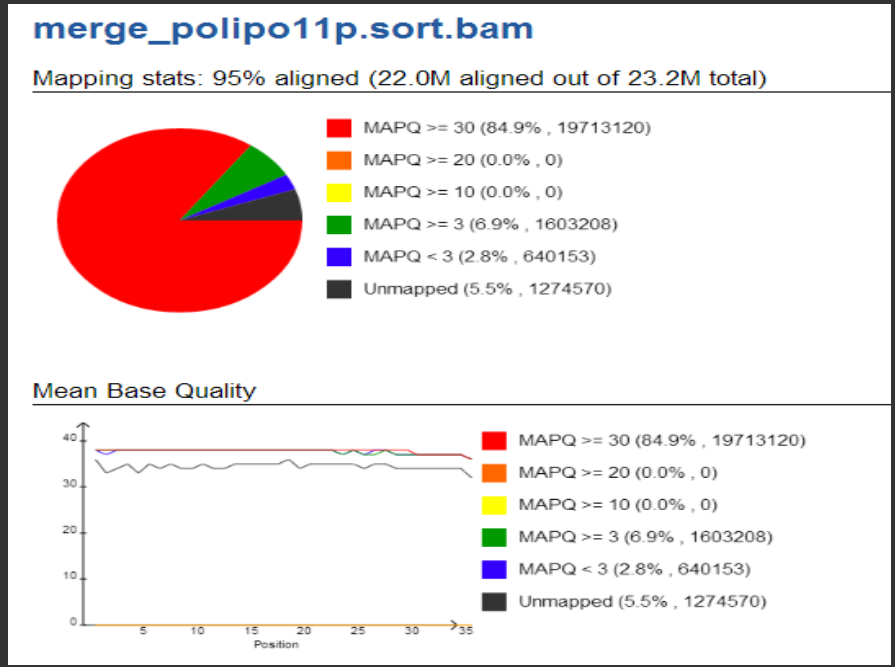
```
fastqc /home/treinamento/NGS/polipo.fastq   -o .
```

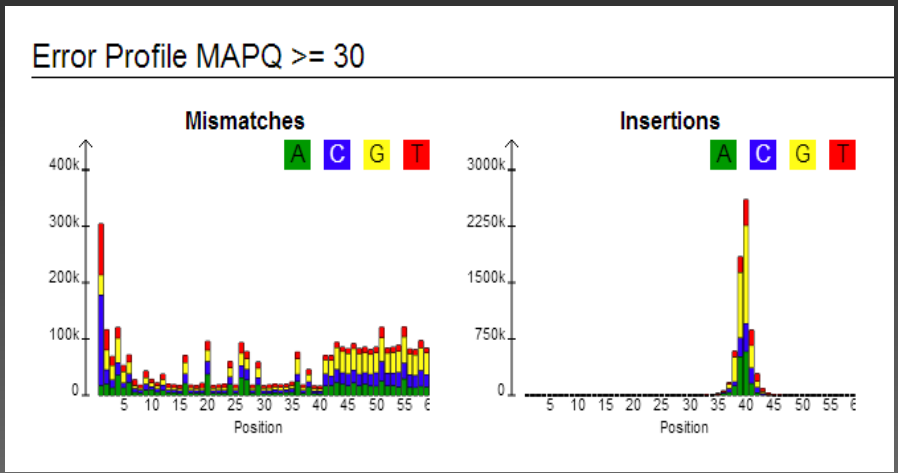
NEXT

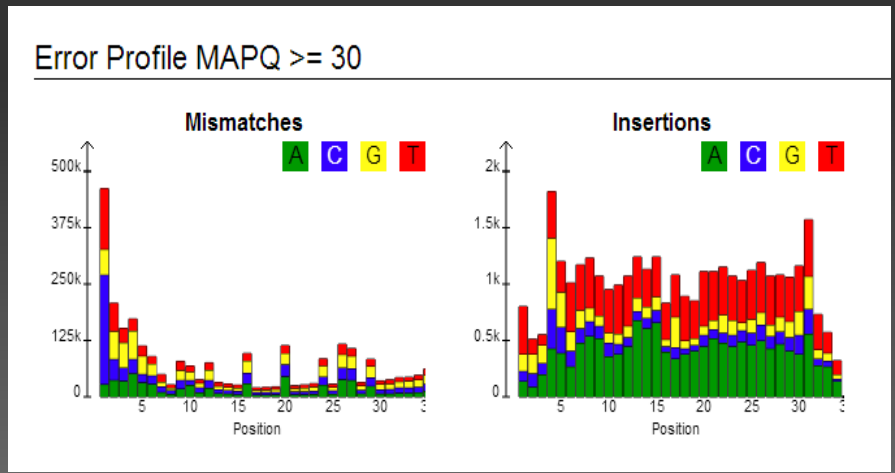
SAMstat :

Quality control









Run FastQC:

Quality control

```
ln -s  /home/treinamento/NGS/ERR844339.bam  .
```

```
samstat  ERR844339.bam
```

```
samstat  polipo.bam
```

NEXT

Run DynamicTrim.pl:

Quality control

DynamicTrim.pl -h 20 /home/treinamento/NGS/10_S5_R1_001.fastq

Base	A	C	A	C	C	T	G	C	C	G
Quality score	30	30	30	30	30	30	30	30	30	10

Base	A	C	A	C	C	T	G	C	C	G
Quality score	30	10	30	30	30	30	30	30	30	10

Base	A	C	A	C	C	T	G	C	C
Quality score	30	30	30	30	30	30	30	30	30

Base	A	C	C	T	G	C	C
Quality score	30	30	30	30	30	30	30

```
-rwxr-xr-x 1 jorge jorge 287M 2012-11-16 19:26 L001_R1_001.fastq
-rw-rw-r-- 1 jorge jorge 186M 2012-11-16 18:43 L001_R1_001.fastq.trimmed
-rwxr-xr-x 1 jorge jorge  46M 2012-11-16 18:15 L002_R1_001.fastq
-rw-rw-r-- 1 jorge jorge  31M 2012-11-16 18:09 L002_R1_001.fastq.trimmed
```

Run FastQC:

Quality control

```
trim_galore  /home/treinamento/NGS/10_S5_R1_001.fastq
```

```
cutadapt -a TGGAATTCTCGG /home/treinamento/NGS/10_S5_R1_001.fastq
```

Obrigado.

E-mail: jorge@imd.ufrn.br