Data structure

Illumina fastq

```
@D44TDFP1_1:1:1101:1320:1948/1
NGGAGGCAGAGGCAGGTGGATTTCTGAGTTCAAGGCCAGCCTGGTCTACAAAGTGAGTNCCAGGACGGCCAGGGCTATACAGAGAAACAGAGAAACCCTGT
+
#1=DDDDDHFHHHIIIAEHGHIIGIIGHGHHIIIIGIIGHIIIIFHIIIIIIIFHIIG#-5@EHHHECCBBBBBBBCECECCCCCCCCCCCCCCCABBCC
@D44TDFP1_1:1:1101:1817:1955/1
NGGGTTGGGGAGGAGAAGATGACGACATTTTTAACAGATTAGTTCATAAAGGCATGTCNATATCACGTCCAAATGCTGTAGTAGGGAGGTGTCGAATGATC
+
#1=DBDDFHHHHHGIIJJJJJIJHGIJJJJJJJJJJJIJJJJIIJJJJJJJJJJJHHH#-;BFAEEDEDDDDDDDCDDDDDDEEDDCBD<BCDDDDDDDDD
@D44TDFP1_1:1:1101:1790:1968/1
GAGGCCAGGTTGAGGATTTTGGAGGACAGAGGGATAAGAAAAATAAGTGGAACAGGAANGGCATTAGCAAAAGCAGAAAAGTATGAACACAAAAGTGAAGT
+
CCCFFFFFHFFHHJJIJHJJJJGHIIJJIGHJJHIJJJJJJJJJJJGIJJJJJJJJJ#-;EHHHHFFFFFFEEECEDDDDACDEDEDDDDDDDDDDDEDC
@D44TDFP1_1:1:1101:1870:1994/1
AGGGGCTGAGTGACTCGGGGCCACATAGGCAGCAAGGAGCAAGGGGCCTGAGCAAGAGNTACCATATTTACCTCAGTGTGTGAAGATCATTTGCCCAGGCT
+
CCCFFFFFHHHHHJJJJJJJJJIJJIJJJIIIIJIIJIGIIIIIJJFGHHHFFFFFEE#,5=BDDDDEFEEDDDDDDDEDDDDEDDDDDDEEEDDDDDDDD
@D44TDFP1_1:1:1101:2070:1923/1
NGCAGNCCNAGGTCTGAGTTCCAAGGACANGTATGTGAAAGGCCTGATTGAGGGCAAANCGGATCCCTACGCGCTCGTCCGTGTGGGCACCCAGACGTTCT
+
#0;@@#2@#2=?=@@@?@?@@@@@@@@@@#1:???>??????????????????????#-;????????==<<<<<:<<:<<<<<:<<<<<<<<<:<<<<
```

fastq format

- Each read is represented by four lines:
- '@', followed by read ID
- Sequence
- '+', optionally followed by repeated read ID
- quality string:
  - same length as sequence, each character encodes the base-call quality of one base

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

FASTQ: Phred base-call qualities

quality score Q_{phred}	error prob. p	characters
0 .. 9	1 .. 0.13	!"#$%&'()*
10 .. 19	0.1 .. 0.013	+,-./01234
20 .. 29	0.01 .. 0.0013	56789:;<=>
30 .. 39	0.001 .. 0.00013	?@ABCDEFGH
40	0.0001	I

- If p is the probability that the base call is wrong, the Phred score is:
- $Q = -10 \log_{10} p$
- The score is written with the character whose ASCII code is Q +33 (Sanger Institute standard).

Fastq format – fasta with qualities

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- p = the probability that the corresponding base call is wrong
- Qualities $Q_{\text{sanger}} = -10 \log_{10} p$
  - p = 0.1 ➔ Q = 10
  - p = 0.01 ➔ Q = 20
  - P = 0.001 ➔ Q = 30
- Encoding: Sanger/Phred format can encode a quality score from 0 to 93 using ASCII 33 to 126: Q + 33 ➔ ASCII code

Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
32	20	040		Space	64	40	100	@	@
33	21	041	!	!	65	41	101	A	A
34	22	042	"	"	66	42	102	B	B
35	23	043	#	#	67	43	103	C	C
36	24	044	$	$	68	44	104	D	D
37	25	045	%	%	69	45	105	E	E
38	26	046	&	&	70	46	106	F	F
39	27	047	'	'	71	47	107	G	G
40	28	050	(	(	72	48	110	H	H
41	29	051	)	)	73	49	111	I	I
42	2A	052	*	*	74	4A	112	J	J
43	2B	053	+	+	75	4B	113	K	K
44	2C	054	,	,	76	4C	114	L	L
45	2D	055	-	-	77	4D	115	M	M
46	2E	056	.	.	78	4E	116	N	N
47	2F	057	/	/	79	4F	117	O	O
48	30	060	0	0	80	50	120	P	P
49	31	061	1	1	81	51	121	Q	Q
50	32	062	2	2	82	52	122	R	R
51	33	063	3	3	83	53	123	S	S
52	34	064	4	4	84	54	124	T	T
53	35	065	5	5	85	55	125	U	U
54	36	066	6	6	86	56	126	V	V
55	37	067	7	7	87	57	127	W	W
56	38	070	8	8	88	58	130	X	X
57	39	071	9	9	89	59	131	Y	Y
58	3A	072	:	:	90	5A	132	Z	Z
59	3B	073	;	;	91	5B	133	[	[
60	3C	074	<	<	92	5C	134	\	\
61	3D	075	=	=	93	5D	135	]	]
62	3E	076	>	>	94	5E	136	^	^
63	3F	077	?	?	95	5F	137	_	_

fastq format

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ................................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |        |                              |                     |
33                         59   64       73                            104                   126

S - Sanger, Illumina 1.8+    Phred+33,  raw reads typically (0, 41)
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)
```

Fastq QC

- Before starting a RNA-seq analysis it is better to have a look at the overall quality of raw data.
- oneChannelGUI has an interface to FastQC a java tool that allows quality controls at the level of various type of sequencing files.

FastQC in basespace

Output

App Session Name: FastQC 03/14/2015 5:33:55

Output Project: Select Project(s):

splicing_dataset mouse ×

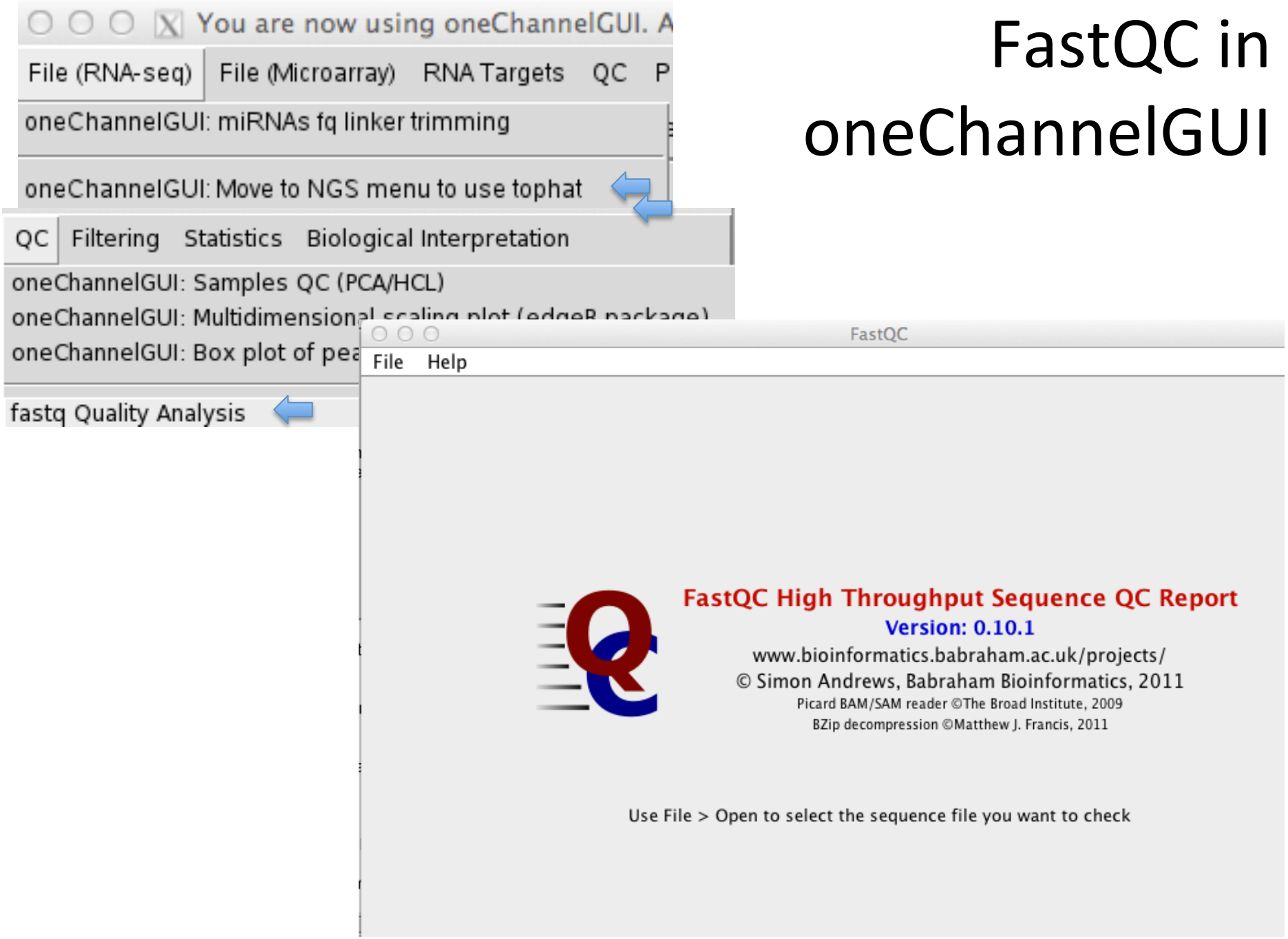
Input

Input Sample: Select Sample(s):

SRR1300868_GSM1399376-C-80-TS-Mus-musculus-RNA-Seq ×

This app is free.

FastQC in oneChannelGUI



You are now using oneChannelGUI. A

File (RNA-seq) | File (Microarray) | RNA Targets | QC | P

oneChannelGUI: miRNAs fq linker trimming

oneChannelGUI: Move to NGS menu to use tophat

QC | Filtering | Statistics | Biological Interpretation

oneChannelGUI: Samples QC (PCA/HCL)

oneChannelGUI: Multidimensional scaling plot (edgeR package)

oneChannelGUI: Box plot of pea

fastq Quality Analysis

FastQC

File | Help

Open

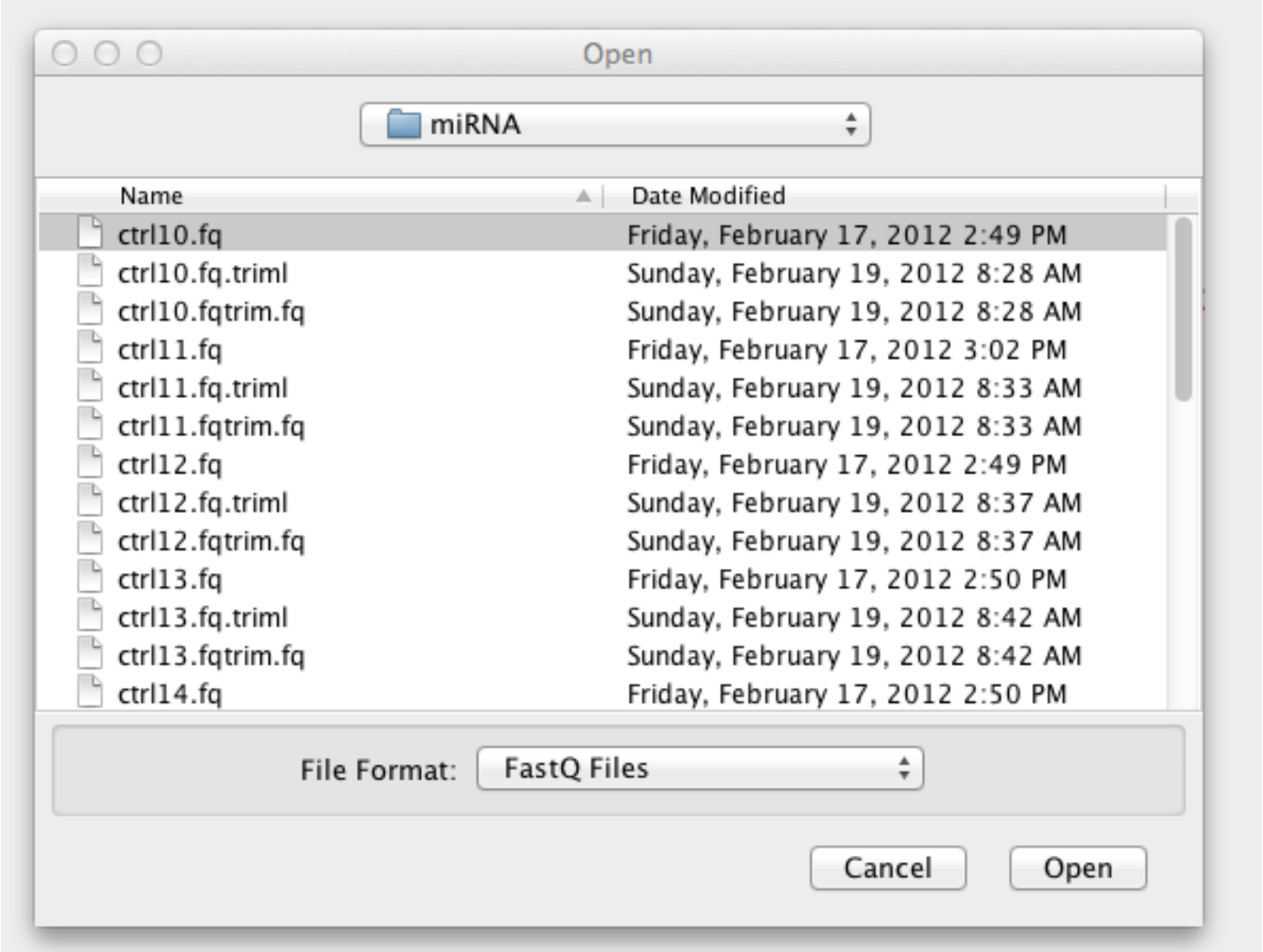
raffaelecalogero

Name	Date Modified
A1190_control_summary.pdf	Wednesday, March 14, 2012 4:40 PM
A1190_Imputed_and_excluded.txt	Wednesday, March 14, 2012 4:40 PM
A1190_TableControl.txt	Wednesday, March 14, 2012 4:40 PM
Applications	Wednesday, July 25, 2012 11:01 AM
Applications (Parallels)	Wednesday, July 25, 2012 11:01 AM
AT.postflight.287	Tuesday, January 17, 2012 7:56 AM
barabino.zip	Wednesday, July 11, 2012 2:33 PM
bin	Friday, September 7, 2012 5:57 PM
Books	Saturday, August 18, 2012 6:42 PM
dataframeR1.plasmid_R2.cho.xlsx	Tuesday, May 15, 2012 8:36 AM
Desktop	Monday, September 10, 2012 4:01 PM
Documents	Sunday, September 9, 2012 9:40 AM
Downloads	Monday, September 10, 2012 3:56 PM

File Format: FastQ Files

Cancel | Open

ort


Open
miRNA
Name
Date Modified
ctrl10.fq
Friday, February 17, 2012 2:49 PM
ctrl10.fq.triml
Sunday, February 19, 2012 8:28 AM
ctrl10.fqtrim.fq
Sunday, February 19, 2012 8:28 AM
ctrl11.fq
Friday, February 17, 2012 3:02 PM
ctrl11.fq.triml
Sunday, February 19, 2012 8:33 AM
ctrl11.fqtrim.fq
Sunday, February 19, 2012 8:33 AM
ctrl12.fq
Friday, February 17, 2012 2:49 PM
ctrl12.fq.triml
Sunday, February 19, 2012 8:37 AM
ctrl12.fqtrim.fq
Sunday, February 19, 2012 8:37 AM
ctrl13.fq
Friday, February 17, 2012 2:50 PM
ctrl13.fq.triml
Sunday, February 19, 2012 8:42 AM
ctrl13.fqtrim.fq
Sunday, February 19, 2012 8:42 AM
ctrl14.fq
Friday, February 17, 2012 2:50 PM
File Format:
FastQ Files
Cancel
Open

ctrl10.fq

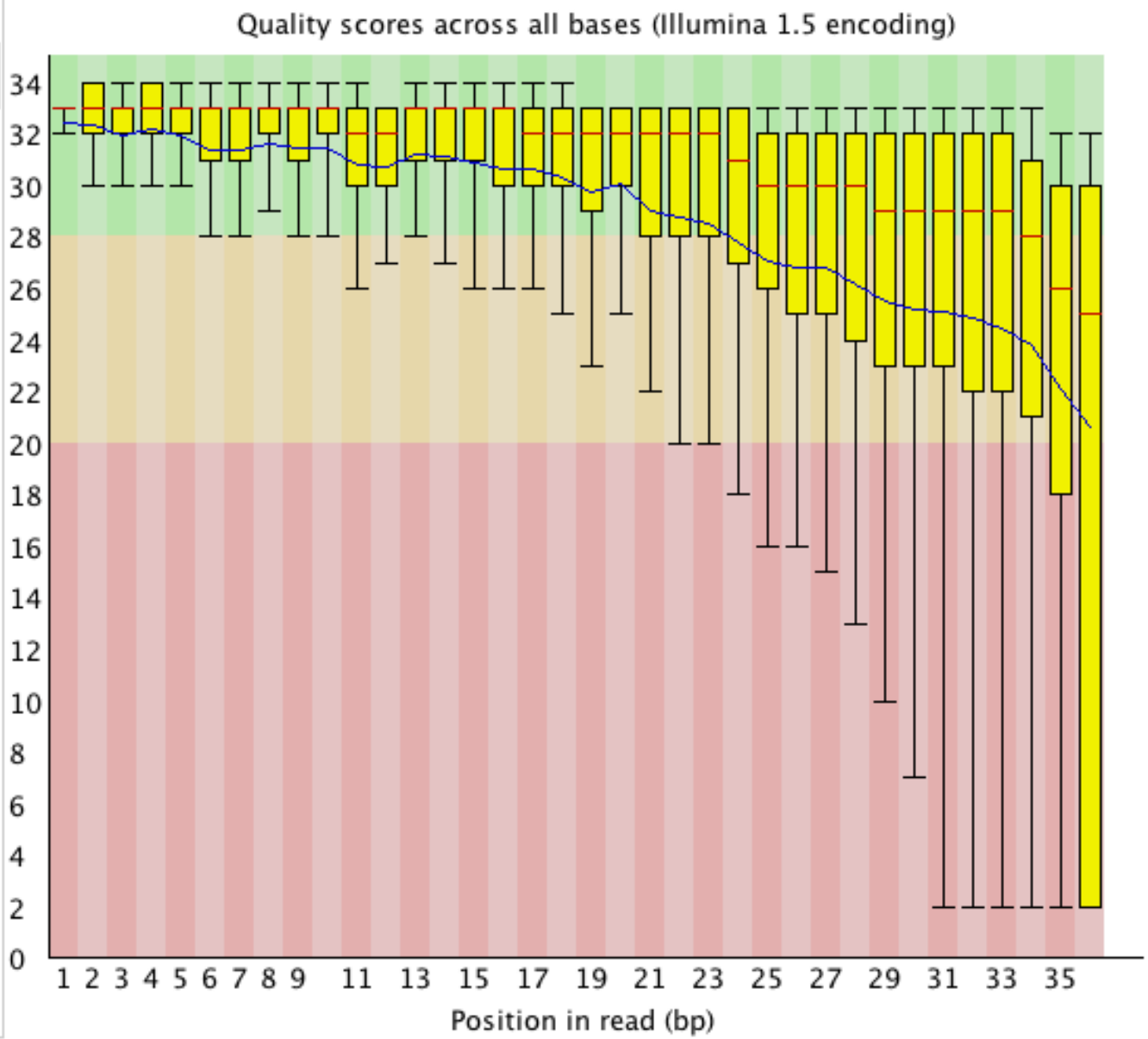
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic sequence stats

Measure	Value
Filename	ctrl10.fq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	1000000
Filtered Sequences	0
Sequence length	36
%GC	47

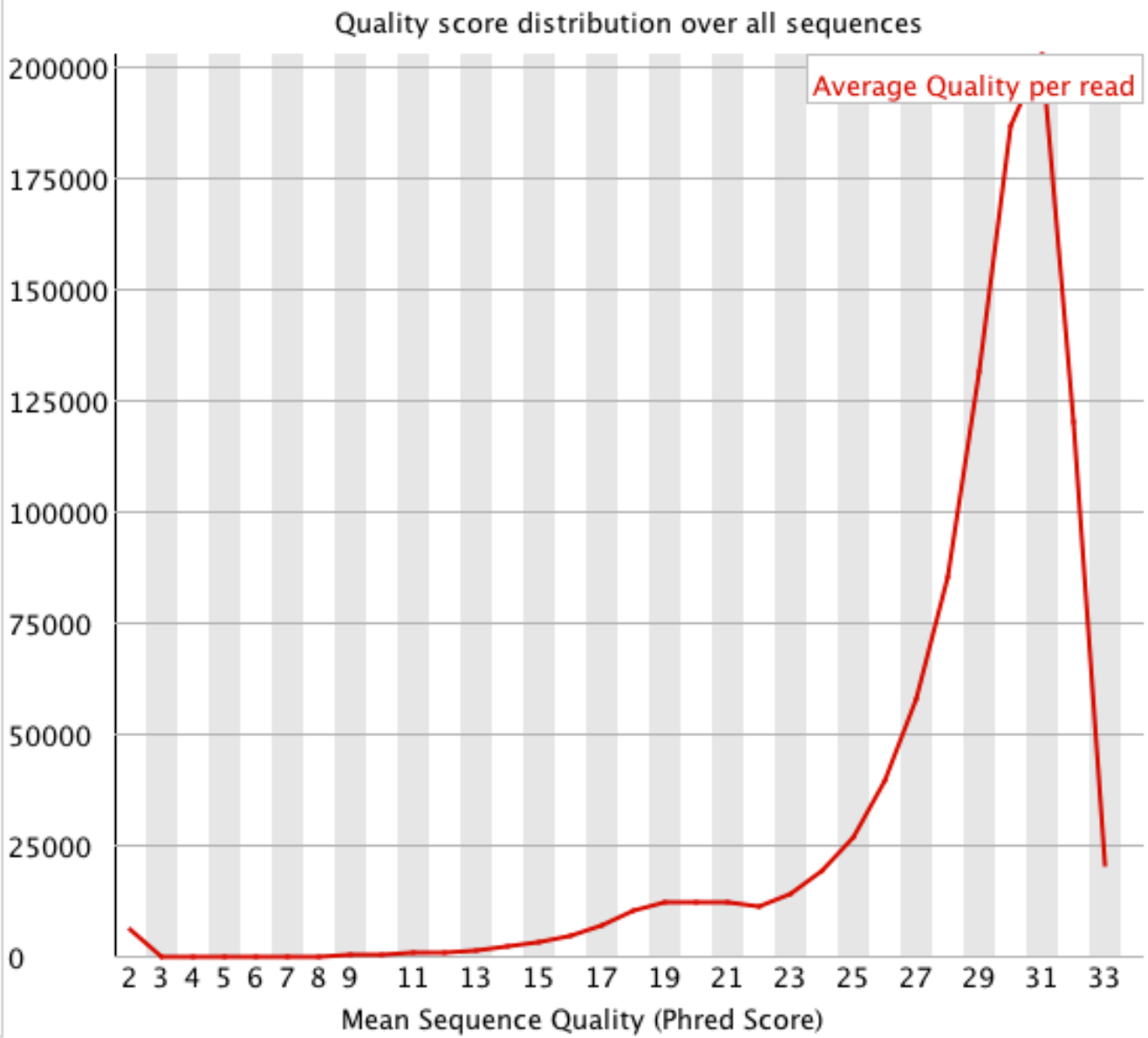
ctrl10.fq

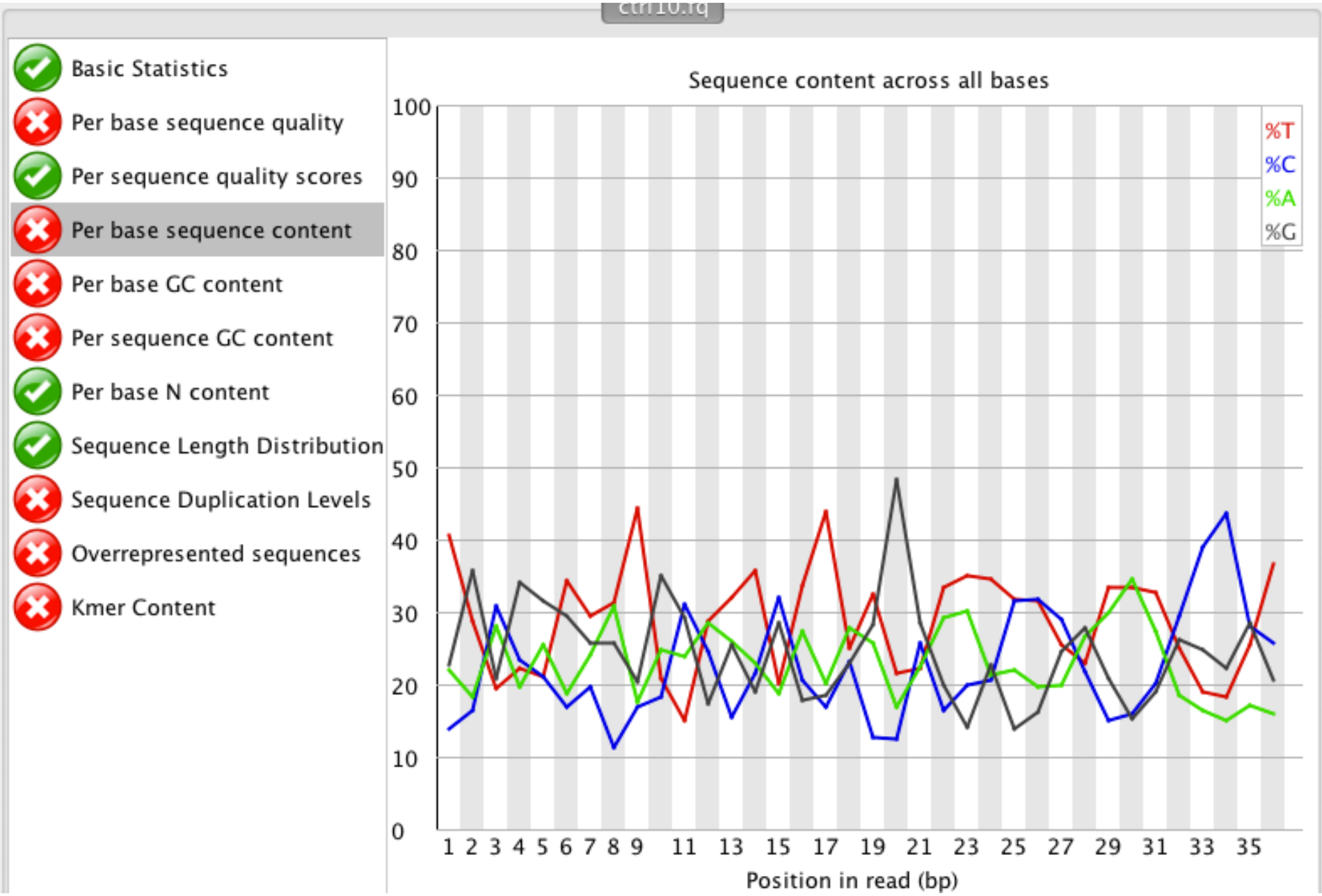
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content



ctrl10.fq

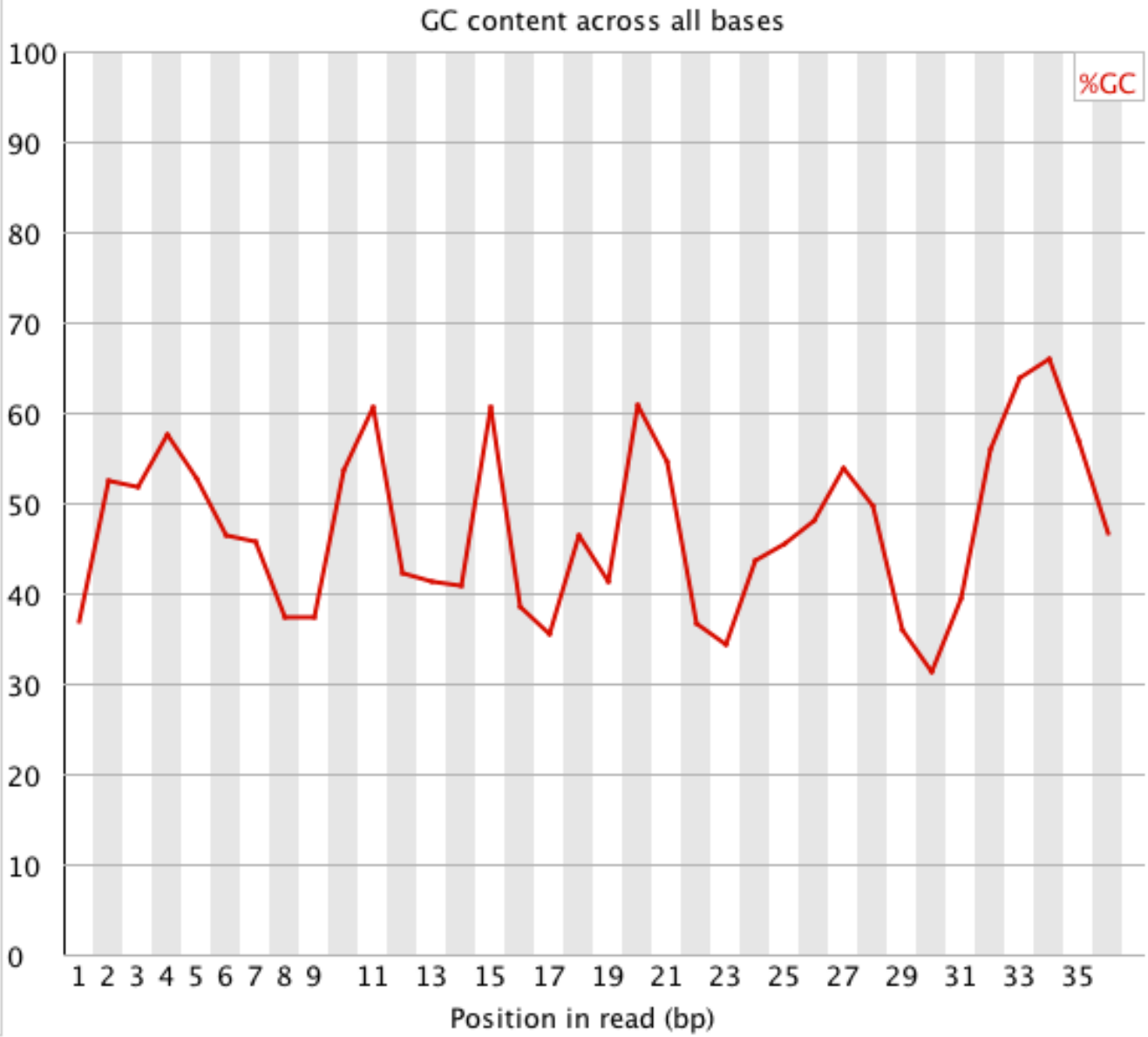
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content




ctrl10.fq
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
Sequence content across all bases
%T
%C
%A
%G
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 35
Position in read (bp)

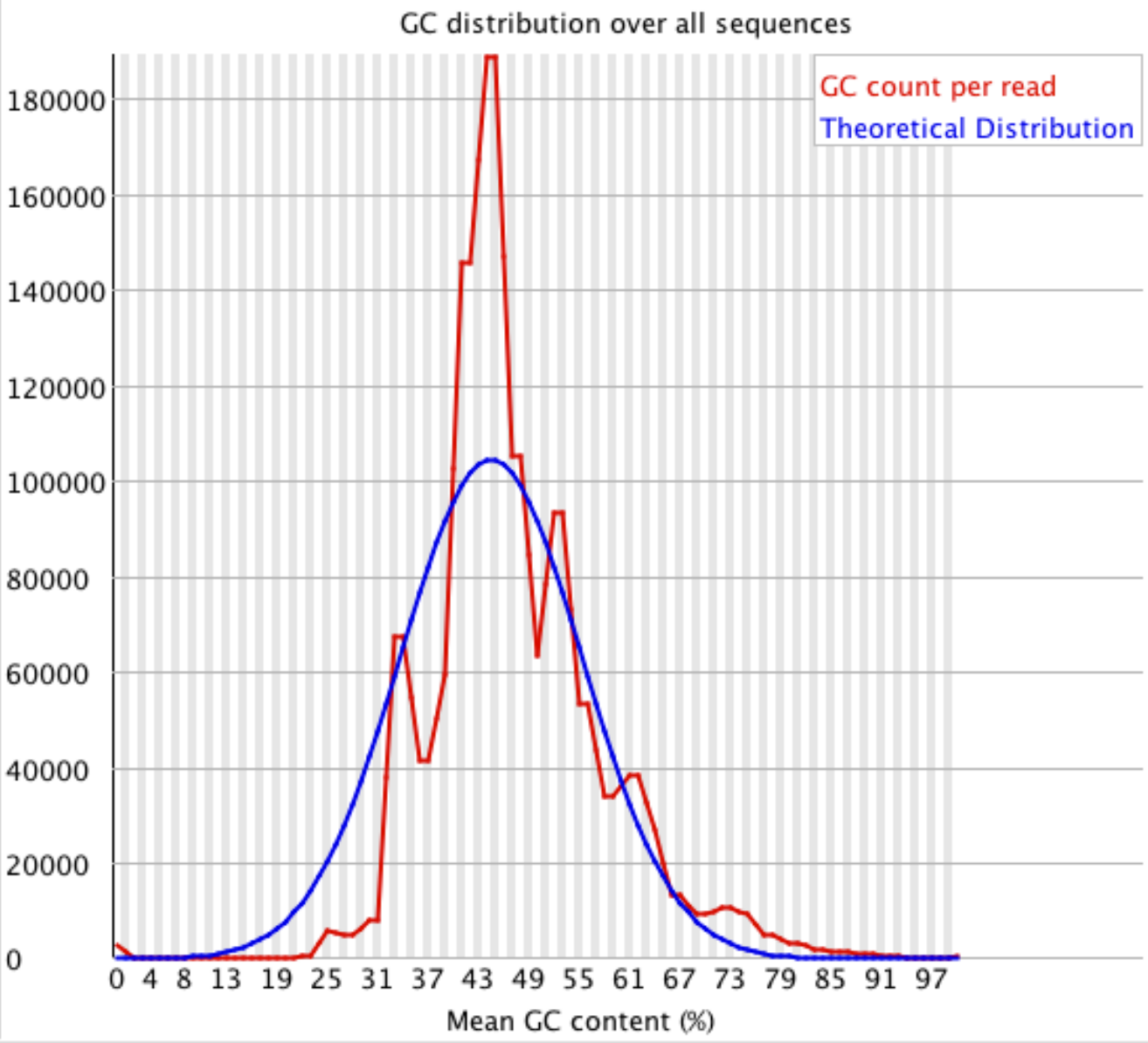
ctrl10.fq

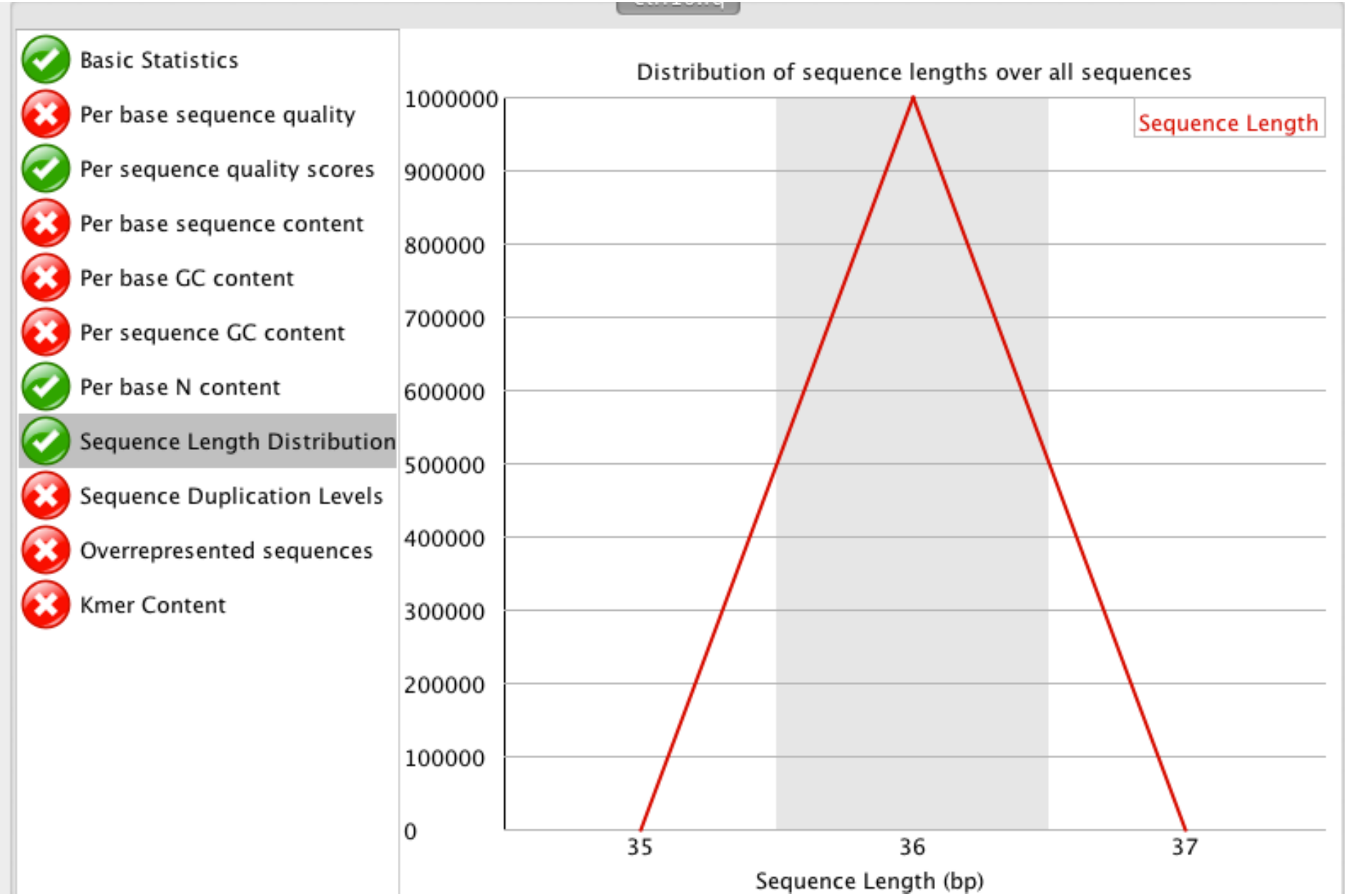
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content



ctrl10.fq

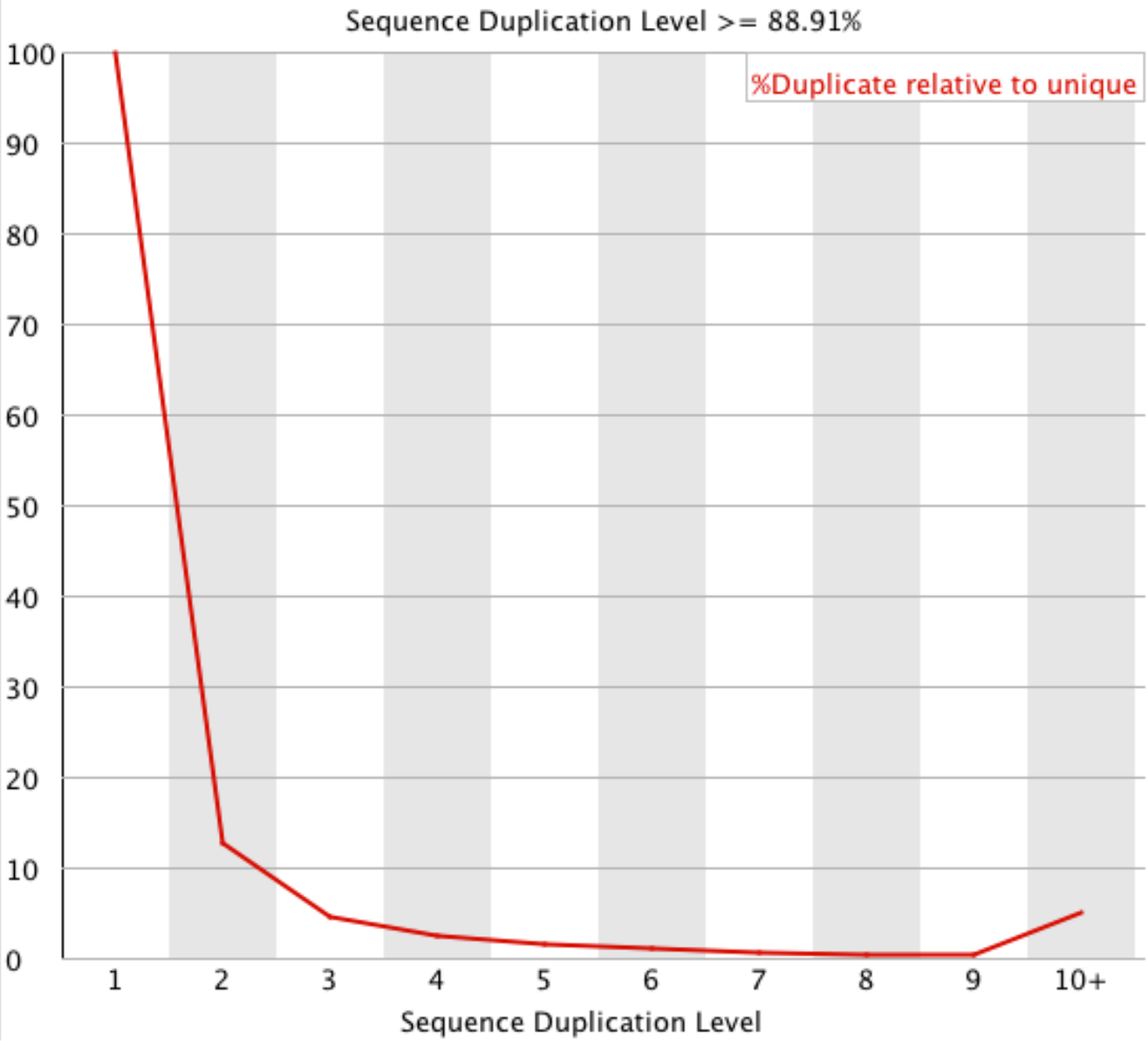
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content




Basic Statistics
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
Distribution of sequence lengths over all sequences
Sequence Length
1000000
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
35
36
37
Sequence Length (bp)

ctrl10.fq

- 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



ctrl10.fq

- 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

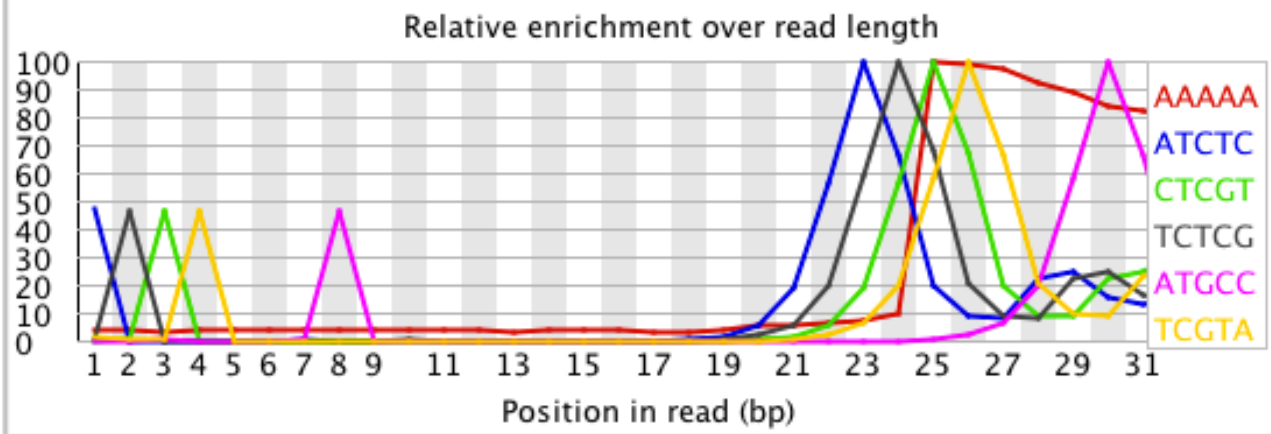
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
ATCTCGTATGCCGT...	54458	5.446	Illumina Single End...
GTCTGTGATGAATT...	25188	2.519	No Hit
GTAGTGTTTCCTAC...	15986	1.599	No Hit
TGAGAACTGAATTC...	15884	1.588	No Hit
TAGCTTATCAGACT...	13178	1.318	No Hit
TGAGAACTGAATTC...	13039	1.304	No Hit
TAGCTTATCAGACT...	12679	1.268	No Hit
TGAGGTAGTAGATT...	12223	1.222	No Hit
AGTCTGTGATGAAT...	12120	1.212	No Hit
TGAGGTAGTAGTTT...	11836	1.184	No Hit
GTAGTGTTTCCTAC...	10634	1.063	No Hit
TGTAGTGTTTCCTA...	10451	1.045	No Hit
TGTAGTGTTTCCTA...	9442	0.944	No Hit
GCGGGTGATGCGAA...	9417	0.942	No Hit
TAGCTTATCAGACT...	7950	0.795	No Hit
TGAGGTAGTAGATT...	6885	0.688	RNA PCR Primer, In...
TGGCTCAGTTCAGC...	6681	0.668	No Hit
ACTGCTGACGCGGG...	6384	0.638	No Hit
TGAGAACTGAATTC...	6349	0.635	RNA PCR Primer, In...
TGAGGTAGTAGTTT...	5833	0.583	No Hit
TCTCACACAGAAAT...	5786	0.579	No Hit
TTCAAGTAATCCAG...	5416	0.542	No Hit
TACCACAGGGTAGA...	5266	0.527	No Hit
GTAGTGTTTCCTAC...	5131	0.513	No Hit
AATGTGTGACTGAA...	5124	0.512	No Hit
TCAAGTGATGTCAT...	5110	0.511	No Hit
TCTCCCAACCCTTG...	4876	0.488	No Hit
TGAGAACTGAATTC...	4539	0.454	No Hit
TGTAAACATCCTTG...	4385	0.438	No Hit
ATGTGTGACTGAAA...	4187	0.419	No Hit

ctrl10.fq

- 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

Overrepresented Kmers



Sequence	Count	Obs/Exp Overall	Obs/Exp Max	Max Obs/Exp Po...
AAAAA	694695	31.094	117.404	25
ATCTC	705450	21.392	155.654	23
CTCGT	707985	20.332	152.508	25
TCTCG	700230	20.109	148.83	24
ATGCC	542740	19.454	186.069	30
TCGTA	659300	18.43	146.109	26
CGTAT	614585	17.18	146.022	27
TGCCG	505625	17.164	176.528	31

SAM File format

Headers

Alignments

```
1 @HD VN:1.0  SO:unsorted
2 @SQ SN:gi|110640213|ref|NC_008253.1|    LN:4938920
3 @PG ID:bowtie2  PN:bowtie2  VN:2.1.0
4 gi|110640213|ref|NC_008253.1|_418_952_1:0:0_1:0:0_0/1   0   gi|110640213|ref|NC_
  008253.1|   418 42  70M *   0   0   CCAGGCAGTGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAA
  ATCACCAACCATCTGGTAGCGATGAT  22222222222222222222222222222222222222222222222222
  222222222222222222  AS:i:-3 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:8G61
  YT:Z:UU
5 gi|110640213|ref|NC_008253.1|_31_476_0:0:0_0:0:0_1/1    16  gi|110640213|ref|NC_
  008253.1|   407 42  70M *   0   0   GGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTG
  CCCCCGCCAAAATCACCAACCATCTG  22222222222222222222222222222222222222222222222222
  222222222222222222  AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:70 YT:Z
  :UU
6 gi|110640213|ref|NC_008253.1|_210_743_2:0:0_1:1:0_2/1   0   gi|110640213|ref|NC_
  008253.1|   210 42  70M *   0   0   CATTACCACCACCATCACCATTACCACAGGAAACGGTGCGGGCT
  GACGCGTACAGGAAACACCGAAAAAA  22222222222222222222222222222222222222222222222222
  222222222222222222  AS:i:-6 XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:30T31A7
      YT:Z:UU
```

Each row describes a single alignment of a raw read against the reference genome. Each alignment has 11 mandatory fields, followed by any number of optional fields.

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	`[0,`2^{16}`-1]`	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	`[0,`2^{31}`-1]`	1-based leftmost mapping POSition
5	MAPQ	Int	`[0,`2^{8}`-1]`	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	`[0,`2^{31}`-1]`	Position of the mate/next read
9	TLEN	Int	`[-`2^{31}`+1,`2^{31}`-1]`	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

FLAG: bitwise FLAG. Each bit is explained in the following table:

Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate
0x800	supplementary alignment

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

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

- H can only be present as the first and/or last operation.
- S may only have H operations between them and the ends of the CIGAR string.
- For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- Sum of lengths of the M/I/S/=/X operations shall equal the length of SEQ.

Tag[4]	Type	Description
X?	?	Reserved fields for end users (together with Y? and Z?)
AM	i	The smallest template-independent mapping quality of segments in the rest
AS	i	Alignment score generated by aligner
BC	Z	Barcode sequence, with any quality scores stored in the QT tag.
BQ	Z	Offset to base alignment quality (BAQ), of the same length as the read sequence. At the i-th read base, $\mathrm{BAQ}_i = Q_i - (\mathrm{BQ}_i - 64)$ where Q_i is the i-th base quality.
CC	Z	Reference name of the next hit; '=' for the same chromosome
CM	i	Edit distance between the color sequence and the color reference (see also NM)
CO	Z	Free-text comments
CP	i	Leftmost coordinate of the next hit
CQ	Z	Color read quality on the original strand of the read. Same encoding as QUAL; same length as CS.
CS	Z	Color read sequence on the original strand of the read. The primer base must be included.
CT	Z	Complete read annotation tag, used for consensus annotation dummy features.[5]
E2	Z	The 2nd most likely base calls. Same encoding and same length as QUAL.
FI	i	The index of segment in the template.
FS	Z	Segment suffix.
FZ	B,S	Flow signal intensities on the original strand of the read, stored as (uint16_t) round(value * 100.0).
LB	Z	Library. Value to be consistent with the header RG-LB tag if @RG is present.
H0	i	Number of perfect hits
H1	i	Number of 1-difference hits (see also NM)
H2	i	Number of 2-difference hits
HI	i	Query hit index, indicating the alignment record is the i-th one stored in SAM
IH	i	Number of stored alignments in SAM that contains the query in the current record
MC	Z	CIGAR string for mate/next segment
MD	Z	String for mismatching positions. *Regex*: [0-9]+(([A-Z]|\^[A-Z]+)[0-9]+)*[6]
MQ	i	Mapping quality of the mate/next segment
NH	i	Number of reported alignments that contains the query in the current record
NM	i	Edit distance to the reference, including ambiguous bases but excluding clipping