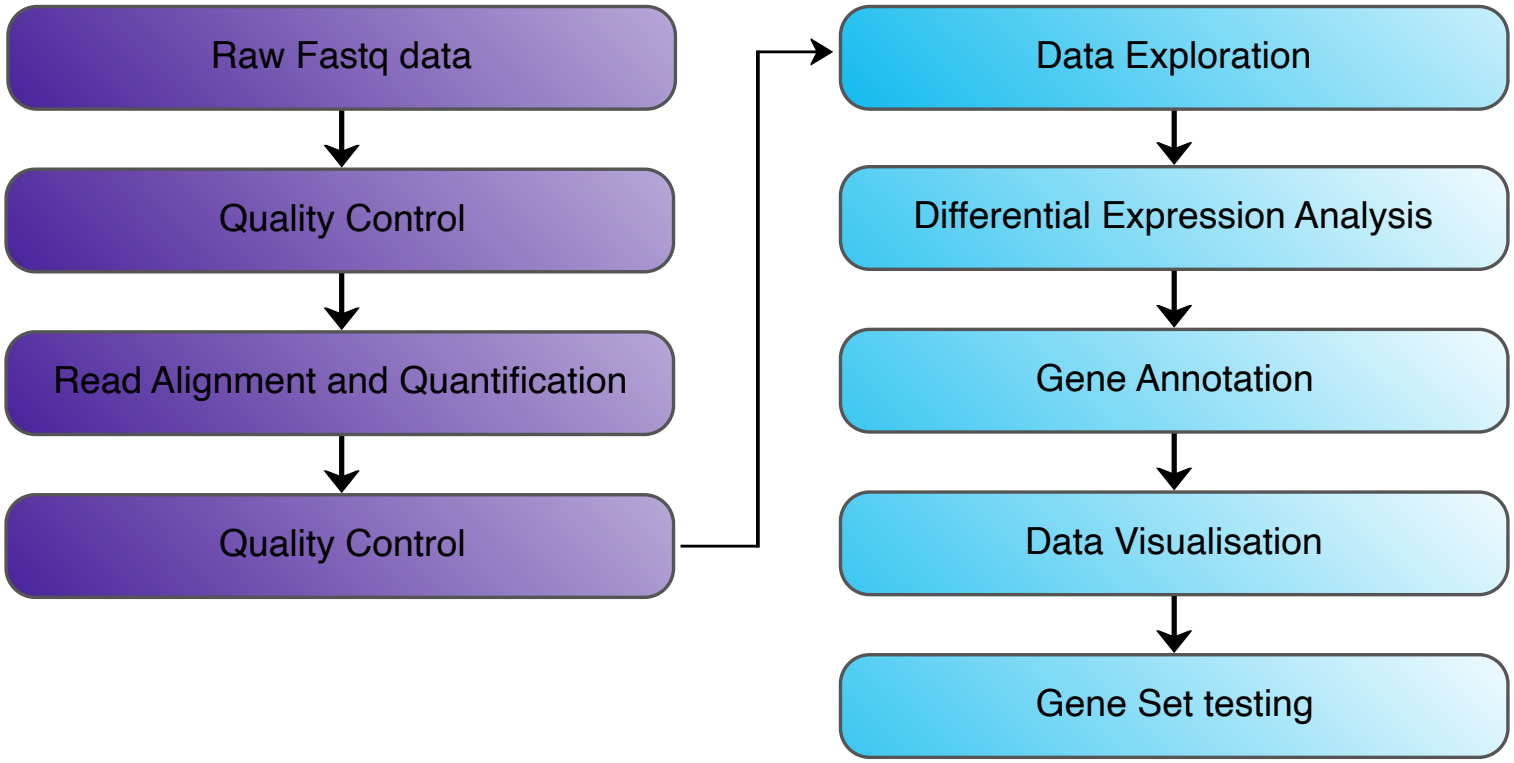
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Basic quality control with FastQC

March 2023

Differential Gene Expression Analysis Workflow



Fastq file format

```
@SRR7657883.sra.1 1 length=150
CTTTGGGTGTTCTCTTGTTTTGCCTCAAGGATGGTCCTTGACTTCCTCACAAGGGTTGTGTCGAATGTTTTGACAGTTACTAATTCTACCACTGTGTTGCCACCATAGAGGCTGTAGATGTCATCTGCGAACTTTTGGAGAGCATTAGTA
+SRR7657883.sra.1 1 length=150
AAFAFJJJJJJJJJJJJAFJJJFJFJFFFJJJJJJJJFFJJFJAAAJJFJFFFJFJJ--AAJ--AA-AF-<-AF-7AAJ7J7-7--7--<---7FA-AAA7F-A<A<A-A-7-7A-A--7-<--<--77-))-)-<-7-7--7)7-------
@SRR7657883.sra.2 2 length=150
CCTGGTCATCGGGCACCTGGACACTTTCACGGCCCAAGGCGGGCCGAGACCCTTGGGGTTCTGTTGAGGCAGCTGGCTTTTCCATAGGGAGCGTGGGGAGGCCCGGGGCGCTGGTCTTAGTCCGACTCTCCTGGGTCCTCCGCGGAGGCT
+SRR7657883.sra.2 2 length=150
AAFFFJFJJJJJ<JJJJJJFJJJJJJ<JJJJJJJJJJJJJJJJJJJJJJJJJFAJJJJJJJJJJFFJJAJJJJJJJAFJAFAAFAAJFAAAJFAJFJ-AJF<AFJJJ7-7FJJJ7-7-<-A<AA-7)A7<F<<A7))A)-<<F---7-)
@SRR7657883.sra.3 3 length=150
GTCCACTCTGCACTCCGCAGCTTTCTTCTCCCTGTCCAGGTCAGCACCTTTTCGAGACGTGAAGATGTTAGAACGCCGCTTGGACTGCTTGCGGATGGGCGTGGGGGTGGAGGAGGTAGCGATGGTCTCGATGCGCAGCTCCCGCTGGCT
+SRR7657883.sra.3 3 length=150
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJFJJJJJJJJJJA7AFJFFJJFJFJJFJJFJJFJJJFFFJ-AAAFFJAAJJ77A-7AJFAJFJJJF7-FJAJF-A<<FFFJJ<FA<<AAJFAA7FF<)<-<)7A<F7
@SRR7657883.sra.4 4 length=150
CGGGGGATTGCAACACAATACCGTGCTGGTGGGCAGCAGCATGAGCATGCCCCCGTCGTGCACAATCCACCACACGTGAATGCTGCCCTCCGAGAAGCGCTCAGGGTTCCCGGGAATCATGGAAATATTCTTGGTGACCAGCAGGGCGAG
+SRR7657883.sra.4 4 length=150
AAAFFFAJJ<AFJJJJJJFJAJJFAJFF<JAJJJJAFJJJJJF<FJFFJJJJJAJ7AJFJFFJJJFJ7J7JFFAA<J-7-AFJ<F<-JFJ7JJF-F<-7FF7-AFAAF---77<-----77<F77-7---A7-A-77)7--))))-)--)
@SRR7657883.sra.5 5 length=150
CGGGAGTGCTCACTGTCACCCAAATCCACATCCCCCCCACCGTCTCCAGAGGTGTGGGCGGAGAGCCGGACTCTTGGAATCATCTCTCAGGCTTCAACCCCAAGGACCACAGGGACTCCCAGGACTAGCCTGCCCCACTTCCTCAGATCG
+SRR7657883.sra.5 5 length=150
AAAFFAFFFJJJJJJJJJFJJJJJJJJJJJ7FJJJJJJJFFJFJJFJJFJFJJAAAAJ<JF<JAJJJAJ<FJJJFFF-AF7AJ-<J7JAA-AAA<JF--A<JJFJJFFA<FFJJFF-J-<<A<AJ-A7A)F<FF7F--F<<AA--)-7--
@SRR7657883.sra.6 6 length=150
CCCATTTCAGTATGATCCCTCTTACCGCTCAGTCCGAGAAATTCGCGAGCATCTTAAGGCCAGGGAAAGTGCACAGCCAGAGAATTGGTCTTGCAACTGCATCCAGTGTGAGCTTAGAATTGGTTCTGAGGAGTTTGAGGAGCTTATTTT
+SRR7657883.sra.6 6 length=150
AAFFFFJJJJJJJJJJJJJJJJFJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJFJJJFJFJFJJJJJJJJJJJJJJJJJ<JJ<JAAFJJJJJJJJFFJJJJJJJJJJJJJJ<FJFJFFF7J<FFFFFJ-FF<AA7JJ<AF7F77FF)7--77-
@SRR7657883.sra.7 7 length=150
CTTTGTAGGTATCACAGGAGTTGAAGAAGGGGCGGATGCAGACCTCGTATTTGTCCAGGTTGATGGCAGCCAGTTCTGTCTGGTCCAGGAAGAGGTCTCCACTAGTGTCCAGCTTAGAGAACATCCAACCGATGGAGTCCTCGCAACTGG
+SRR7657883.sra.7 7 length=150
AAFFFFJJJJJJJJJJJJJJFJFJJJJJFJJAJJJJJJFJJJJJFJFFFFJFJJJJFAJJFFFFJJJJ7FFJJJJJJJAJJAJJJFJ-FFFJFJFFFJ--AAFFJJJJJF-<AJJJA-A7---AAF-7<-AF<JJAFA7<)--77)-<<A-
@SRR7657883.sra.8 8 length=150
GGGGGCTGCTGACCATCCGATACCCCATGGAACACGGTGTGGTACGGGATTGGAATGACATGGAACGCGAGGGCGGCAGGGGAGAGGTTTGTGTTCCTGTCATCCGCTTTGATCATGAGTGCAGTGGTGGGGGGGGGGTTACATGGTATT
+SRR7657883.sra.8 8 length=150
AAFFFFJJJJJJJJJJJJJJJJJJJJFFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJFFJJJJJJAFFJA-AJA--7-----7--A-A-----A7--<FF---77--7-FAF-------A----A-))7)-)))-)))----------
@SRR7657883.sra.9 9 length=150
GTCATGCACTTCTTCGTCATAACTATCGTCTTCATTATCAGACACCGGCTCAGGATCTGGGTGACTTACAGGGATCTGAACTTTCTTCTTCCTTCGGATAGTAATTTTTGCGTTTTTCCCACTCTTCCTTAGCTGCTGAACAGCAATAGT
+SRR7657883.sra.9 9 length=150
AAFFFFJJJJJJJJJJJJJJJJJJJJFJJJJJFJJJJJJJJJJJJJJJJFJJJJJJFJJJJFAFFFFJJ7AAJFAFJFJFFF-FFJFJJJFJFJJJ7JF-7AJJA-AAAJA77FJFFF-7-7-<FAF-FA<F77FJ7AAA7---<-----7-
```

Fastq file format - Headers

```
@SRR7657883.sra.1 1 length=150
CTTTGGGTGTTCTCTTGTTTTGCCTCAAGGATGGTCCTTGACTTCCTCACAAGGGTTGTGTCGAATGTTTTGACAGTTACTAATTC
+SRR7657883.sra.1 1 length=150
AAFAFJJJJJJJJJJJAFJJJFJFJFFFJJJJJJJFFJJFJAAAJJFJFFFJFJJ--AAJ--AA-AF-<-AF-7AAJ7J7-7--7-
@SRR7657883.sra.2 2 length=150
CCTGGTCATCGGGCACCTGGACACTTTCACGGCCCAAGGCGGGCCGAGACCCTTGGGGTTCTGTTGAGGCAGCTGGCTTTTCCATA
+SRR7657883.sra.2 2 length=150
AAFFFJFJJJJJ<JJJJJJFJJJJJJ<JJJJJJJJJJJJJJJJJJJJJJJJJJFAJJJJJJJJJJFFJJAJJJJJJJAFJAFAAFA
@SRR7657883.sra.3 3 length=150
GTCCACTCTGCACTCCGCAGCTTTCTTCTCCCTGTCCAGGTCAGCACCTTTTCGAGACGTGAAGATGTTAGAACGCCGCTTGGACT
+SRR7657883.sra.3 3 length=150
AAFFFJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJFJJJJJJJJJA7AFJFFJJFJFJJFJJFJJFJJJFFFJ-AA
@SRR7657883.sra.4 4 length=150
CGGGGGATTGCAACACAATACCGTGCTGGTGGGCAGCAGCATGAGCATGCCCCCGTCGTGCACAATCCACCACACGTGAATGCTGC
+SRR7657883.sra.4 4 length=150
AAAFFFAJJ<AFJJJJJJFJAJJFAJFF<JAJJJJAFJJJJJF<FJFFJJJJJAJ7AJFJFFJJJFJ7J7JFFAA<J-7-AFJ<F<
@SRR7657883.sra.5 5 length=150
CGGGAGTGCTCACTGTCACCCAAATCCACATCCCCCCCACCGTCTCCAGAGGTGTGGGCGGAGAGCCGGACTCTTGGAATCATCTC
+SRR7657883.sra.5 5 length=150
AAAFFAFFFJJJJJJJJJFJJJJJJJJJJ7FJJJJJJJFFJFJJFJJFJFJJAAAAJ<JF<JAJJJAJ<FJJJFFF-AF7AJ-<J
```

Fastq file format - Sequences

```
@SRR7657883.sra.1 1 length=150
CTTTGGGTGTTCTCTTGTTTTGCCTCAAGGATGGTCCTTGACTTCCTCACAAGGGTTGTGTCGAATGTTTTGACAGTTACTAATTC
+SRR7657883.sra.1 1 length=150
AAFAFJJJJJJJJJJJAFJJJFJFJFFFJJJJJJJFFJJFJAAAJJFJFFFJFJJ--AAJ--AA-AF-<-AF-7AAJ7J7-7--7-
@SRR7657883.sra.2 2 length=150
CCTGGTCATCGGGCACCTGGACACTTTCACGGCCCAAGGCGGGCCGAGACCCTTGGGGTTCTGTTGAGGCAGCTGGCTTTTCCATA
+SRR7657883.sra.2 2 length=150
AAFFFJFJJJJJ<JJJJJJFJJJJJJ<JJJJJJJJJJJJJJJJJJJJJJJJJJFAJJJJJJJJJJFFJJAJJJJJJJAFJAFAAFA
@SRR7657883.sra.3 3 length=150
GTCCACTCTGCACTCCGCAGCTTTCTTCTCCCTGTCCAGGTCAGCACCTTTTCGAGACGTGAAGATGTTAGAACGCCGCTTGGACT
+SRR7657883.sra.3 3 length=150
AAFFFJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJFJJJJJJJJJA7AFJFFJJFJFJJFJJFJJFJJJFFFJ-AA
@SRR7657883.sra.4 4 length=150
CGGGGGATTGCAACACAATACCGTGCTGGTGGGCAGCAGCATGAGCATGCCCCCGTCGTGCACAATCCACCACACGTGAATGCTGC
+SRR7657883.sra.4 4 length=150
AAAFFFAJJ<AFJJJJJJFJAJJFAJFF<JAJJJJAFJJJJJF<FJFFJJJJJAJ7AJFJFFJJJFJ7J7JFFAA<J-7-AFJ<F<
@SRR7657883.sra.5 5 length=150
CGGGAGTGCTCACTGTCACCCAAATCCACATCCCCCCCACCGTCTCCAGAGGTGTGGGCGGAGAGCCGGACTCTTGGAATCATCTC
+SRR7657883.sra.5 5 length=150
AAAFFAFFFJJJJJJJJJFJJJJJJJJJJ7FJJJJJJJFFJFJJFJJFJFJJAAAAJ<JF<JAJJJAJ<FJJJFFF-AF7AJ-<J
```

Fastq file format - Third line

```
@SRR7657883.sra.1 1 length=150
CTTTGGGTGTTCTCTTGTTTTGCCTCAAGGATGGTCCTTGACTTCCTCACAAGGGTTGTGTCGAATGTTTTGACAGTTACTAATTC
+SRR7657883.sra.1 1 length=150
AAFAFJJJJJJJJJJJAFJJJFJFJFFFJJJJJJJFFJJFJAAAJJFJFFFJFJJ--AAJ--AA-AF-<-AF-7AAJ7J7-7--7-
@SRR7657883.sra.2 2 length=150
CCTGGTCATCGGGCACCTGGACACTTTCACGGCCCAAGGCGGGCCGAGACCCTTGGGGTTCTGTTGAGGCAGCTGGCTTTTCCATA
+SRR7657883.sra.2 2 length=150
AAFFFJFJJJJJ<JJJJJJFJJJJJJ<JJJJJJJJJJJJJJJJJJJJJJJJJFAJJJJJJJJJJFFJJAJJJJJJJAFJAFAAFA
@SRR7657883.sra.3 3 length=150
GTCCACTCTGCACTCCGCAGCTTTCTTCTCCCTGTCCAGGTCAGCACCTTTTCGAGACGTGAAGATGTTAGAACGCCGCTTGGACT
+SRR7657883.sra.3 3 length=150
AAFFFJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJFJJJJJJJJJA7AFJFFJJFJFJJFJJFJJFJJJFFFJ-AA
@SRR7657883.sra.4 4 length=150
CGGGGGATTGCAACACAATACCGTGCTGGTGGGCAGCAGCATGAGCATGCCCCCGTCGTGCACAATCCACCACACGTGAATGCTGC
+SRR7657883.sra.4 4 length=150
AAAFFFAJJ<AFJJJJJJFJAJJFAJFF<JAJJJJAFJJJJJF<FJFFJJJJJAJ7AJFJFFJJJFJ7J7JFFAA<J-7-AFJ<F<
@SRR7657883.sra.5 5 length=150
CGGGAGTGCTCACTGTCACCCAAATCCACATCCCCCCCACCGTCTCCAGAGGTGTGGGCGGAGAGCCGGACTCTTGGAATCATCTC
+SRR7657883.sra.5 5 length=150
AAAFFAFFFJJJJJJJJJFJJJJJJJJJJ7FJJJJJJJFFJFJJFJJFJFJJAAAAJ<JF<JAJJJAJ<FJJJFFF-AF7AJ-<J
```

Fastq file format - Quality Scores

```
@SRR7657883.sra.1 1 length=150
CTTTGGGTGTTCTCTTGTTTTGCCTCAAGGATGGTCCTTGACTTCCTCACAAGGGTTGTGTCGAATGTTTTGACAGTTACTAATTC
+SRR7657883.sra.1 1 length=150
AAFAFJJJJJJJJJJJAFJJJFJFJFFFJJJJJJJFFJJFJAAAJJFJFFFJFJJ--AAJ--AA-AF-<-AF-7AAJ7J7-7--7-
@SRR7657883.sra.2 2 length=150
CCTGGTCATCGGGCACCTGGACACTTTCACGGCCCAAGGCGGGCCGAGACCCTTGGGGTTCTGTTGAGGCAGCTGGCTTTTCCATA
+SRR7657883.sra.2 2 length=150
AAFFFJFJJJJJ<JJJJJJFJJJJJJ<JJJJJJJJJJJJJJJJJJJJJJJJJFAJJJJJJJJJJFFJJAJJJJJJJAFJAFAAFA
@SRR7657883.sra.3 3 length=150
GTCCACTCTGCACTCCGCAGCTTTCTTCTCCCTGTCCAGGTCAGCACCTTTTCGAGACGTGAAGATGTTAGAACGCCGCTTGGACT
+SRR7657883.sra.3 3 length=150
AAFFFJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJFJJJJJJJJJA7AFJFFJJFJFJJFJJFJJFJJJFFFJ-AA
@SRR7657883.sra.4 4 length=150
CGGGGGATTGCAACACAATACCGTGCTGGTGGGCAGCAGCATGAGCATGCCCCCGTCGTGCACAATCCACCACACGTGAATGCTGC
+SRR7657883.sra.4 4 length=150
AAAFFFAJJ<AFJJJJJJFJAJJFAJFF<JAJJJJAFJJJJJF<FJFFJJJJJAJ7AJFJFFJJJFJ7J7JFFAA<J-7-AFJ<F<
@SRR7657883.sra.5 5 length=150
CGGGAGTGCTCACTGTCACCCAAATCCACATCCCCCCCACCGTCTCCAGAGGTGTGGGCGGAGAGCCGGACTCTTGGAATCATCTC
+SRR7657883.sra.5 5 length=150
AAAFFAFFFJJJJJJJJJFJJJJJJJJJJ7FJJJJJJJFFJFJJFJJFJFJJAAAAJ<JF<JAJJJAJ<FJJJFFF-AF7AJ-<J
```

(Phred) Quality Scores

Sequence quality scores are transformed and translated p-values

- Sequence bases are called after image processing (base calling)
  - Each base in a sequence has a *p-value* associated with it
  - p-values range from 0-1 (e.g.: 0.05, 0.01, 1e-30)
  - p-value of 0.01 inferred as 1 in 100 chance that called base is wrong

(Phred) Quality Scores …

How do we assign p-values to bases in the fastq file?

- P-vales can be many characters long (e.g.:0.000005)
- Transform to Phred quality scores *Q*
- $Q = -10(log_{10}P)$ (e.g.: 0.01 = Q value of 20, 0.001 = Q value of 30)
- Translate *Q* values to ASCII characters (adding 33) (Q value of 30 = ?, Q value of 40 = I)

Dec	Hex	Chr
0	00	NUL
1	01	SOH
2	02	STX
3	03	ETX
4	04	EOT
5	05	ENQ
6	06	ACK
7	07	BEL
8	08	BS
9	09	HT
10	0A	LF
11	0B	VT
12	0C	FF
13	0D	CR
14	0E	SO
15	0F	SI
16	10	DLE
17	11	DC1
18	12	DC2
19	13	DC3
20	14	DC4
21	15	NAK
22	16	SYN
23	17	ETB
24	18	CAN
25	19	EM
26	1A	SUB
27	1B	ESC
28	1C	FS
29	1D	GS
30	1E	RS
31	1F	US

Dec	Hex	Chr
32	20	Space
33	21	!
34	22	"
35	23	#
36	24	$
37	25	%
38	26	&
39	27	'
40	28	(
41	29	)
42	2A	*
43	2B	+
44	2C	,
45	2D	-
46	2E	.
47	2F	/
48	30	0
49	31	1
50	32	2
51	33	3
52	34	4
53	35	5
54	36	6
55	37	7
56	38	8
57	39	9
58	3A	:
59	3B	;
60	3C	<
61	3D	=
62	3E	>
63	3F	?

Dec	Hex	Chr
64	40	@
65	41	A
66	42	B
67	43	C
68	44	D
69	45	E
70	46	F
71	47	G
72	48	H
73	49	I
74	4A	J
75	4B	K
76	4C	L
77	4D	M
78	4E	N
79	4F	O
80	50	P
81	51	Q
82	52	R
83	53	S
84	54	T
85	55	U
86	56	V
87	57	W
88	58	X
89	59	Y
90	5A	Z
91	5B	[
92	5C	\
93	5D	]
94	5E	^
95	5F	_

Dec	Hex	Chr
96	60	`
97	61	a
98	62	b
99	63	c
100	64	d
101	65	e
102	66	f
103	67	g
104	68	h
105	69	i
106	6A	j
107	6B	k
108	6C	l
109	6D	m
110	6E	n
111	6F	o
112	70	p
113	71	q
114	72	r
115	73	s
116	74	t
117	75	u
118	76	v
119	77	w
120	78	x
121	79	y
122	7A	z
123	7B	{
124	7C	|
125	7D	}
126	7E	~
127	7F	DEL

QC is important

Check for any problems before we put time and effort into analysing potentially bad data

- Start with FastQC
  - Quick
  - Outputs an easy to read html report

We run fastQC from the terminal with the command

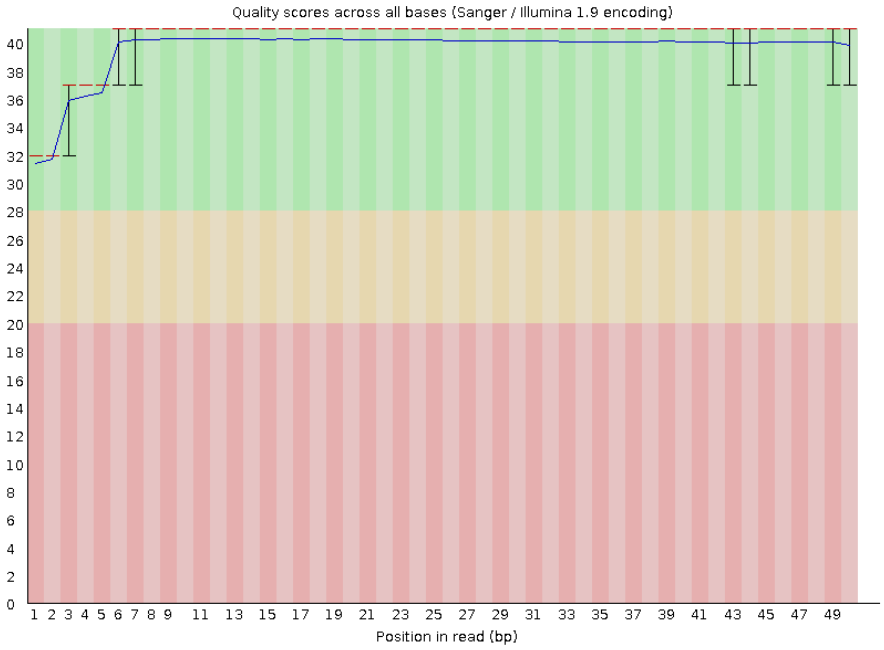
fastqc `<fastq>`

but there are lots of other parameters which you can find to tailor your QC by typing

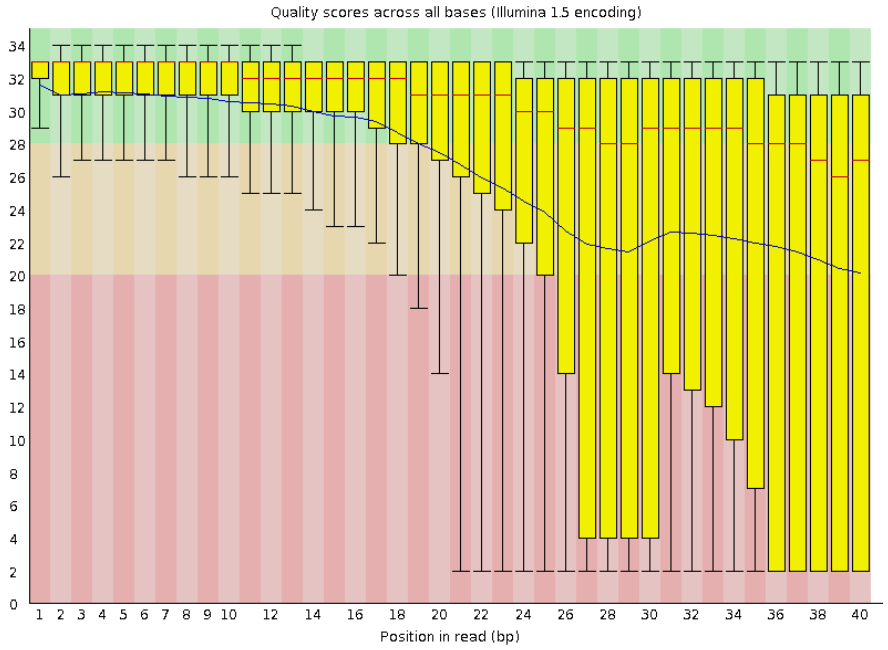
fastqc -h

Per base sequence quality

Good Data

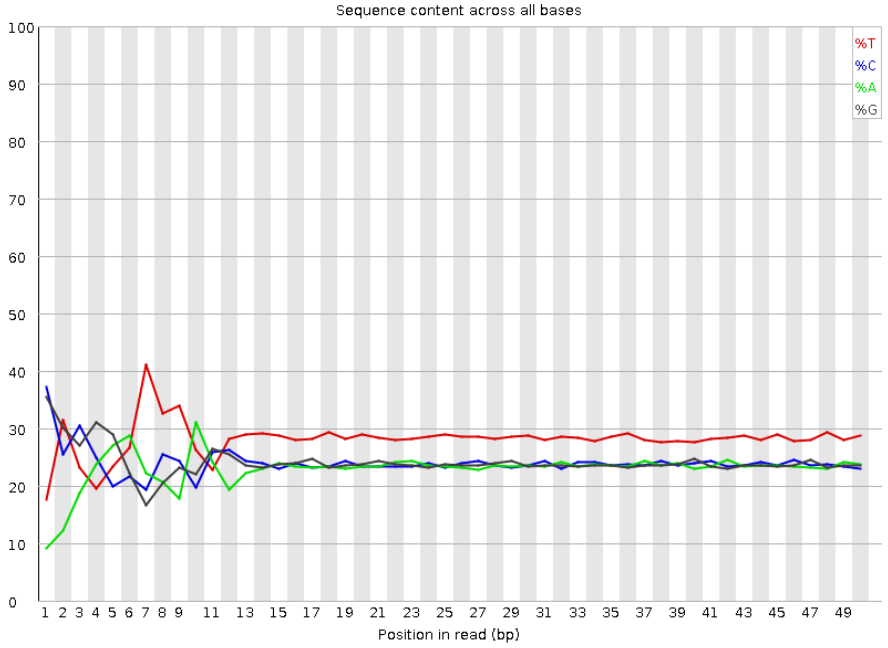


Bad Data

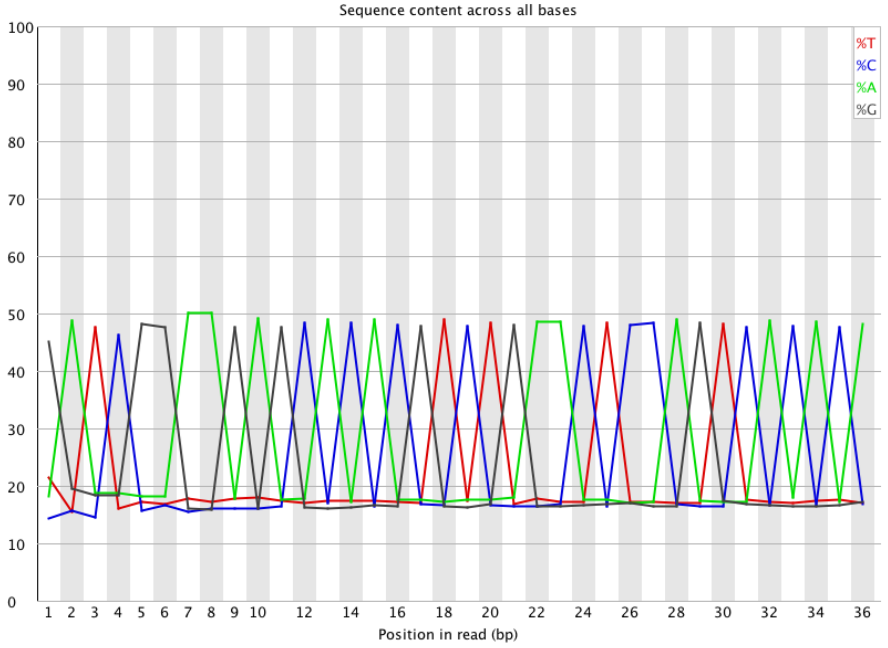


Per base sequence content

Good Data

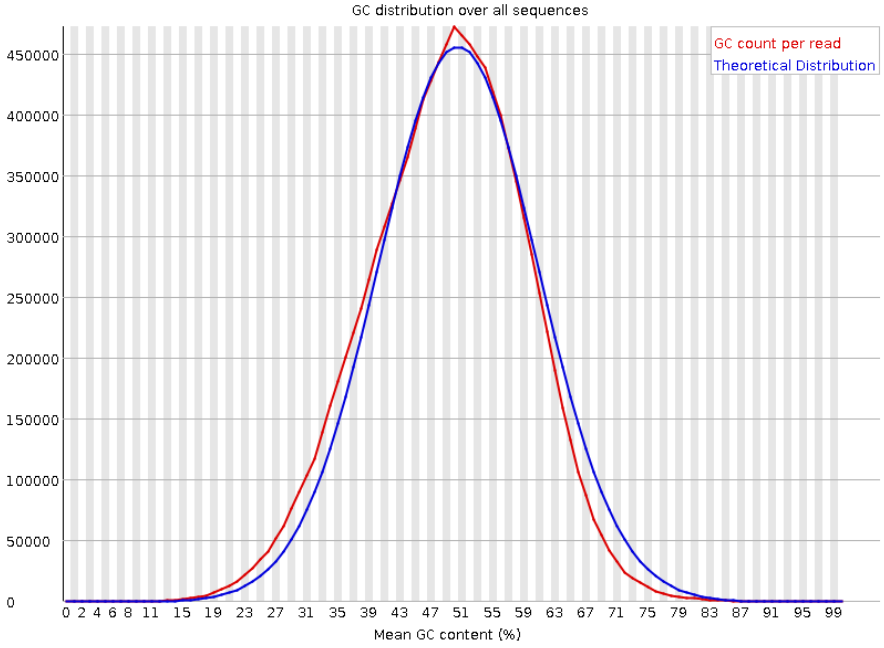


Bad Data

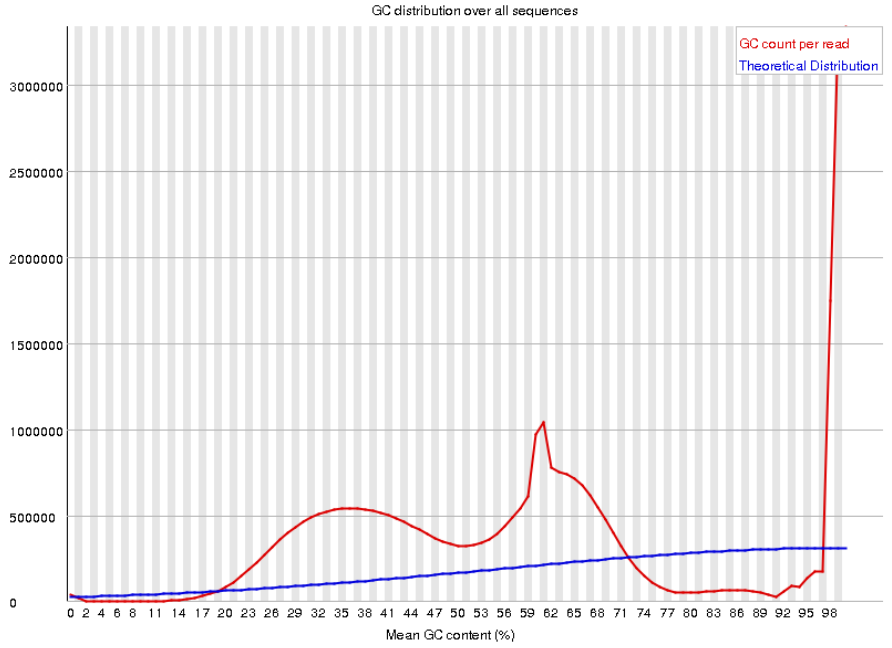


Per sequence GC content

Good Data

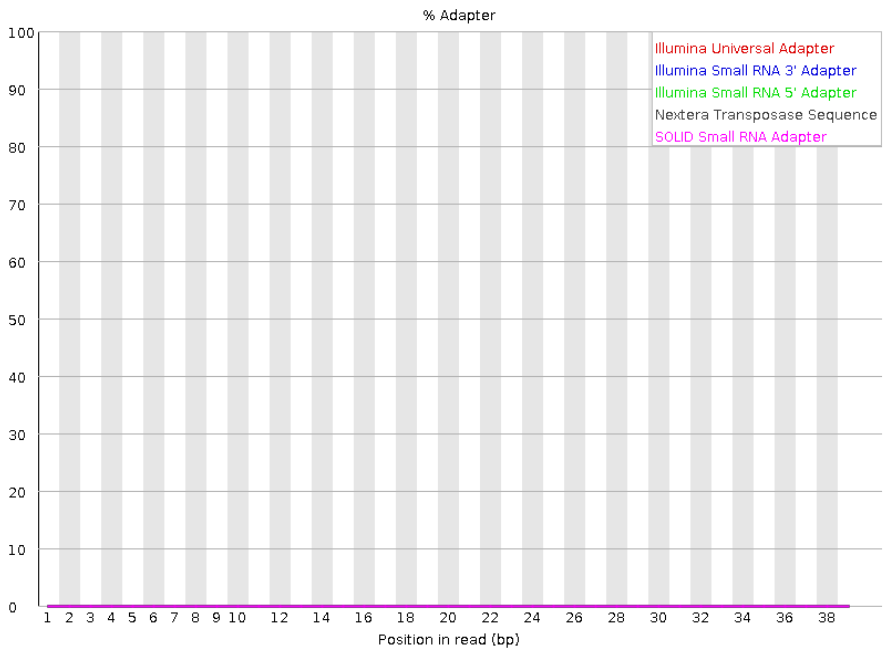


Bad Data

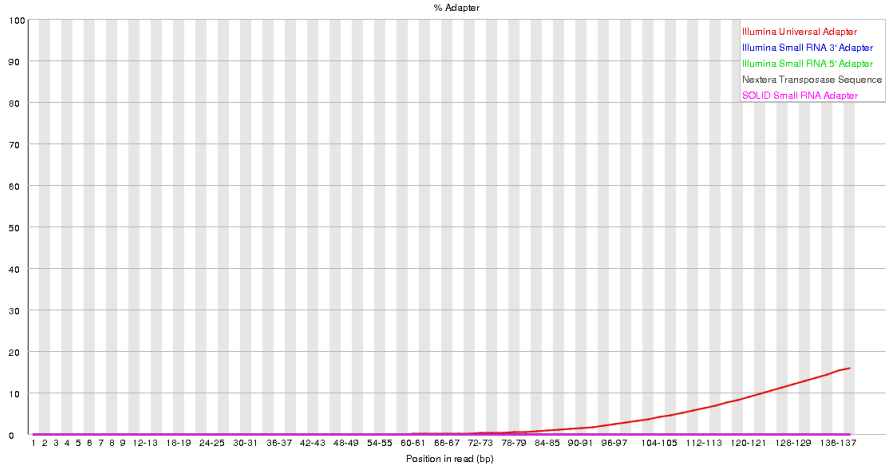


Adaptor content

Good Data



Bad Data



And now onto the exercise…

- Go to: https://ushers.bio.cam.ac.uk/guacamole2
- Log on with YOUR credentials that were emailed to you

A quick intro to the environment

- The terminal is just a text based version of the operating system
- We will look at an example with side by side GUI and text file system…
- You use commands instead of mouse clicks - commands are case-senstitve and can be followed by arguments with spaces
  - cd
  - pwd
  - ls
  - flags - e.g. ls -a
  - the directory structure is like a tree, you can go back with cd ..
  - Up arrows to get through history
  - tab complete to avoid errors
  - More to look at the files and q to exit
  - ctrl-c