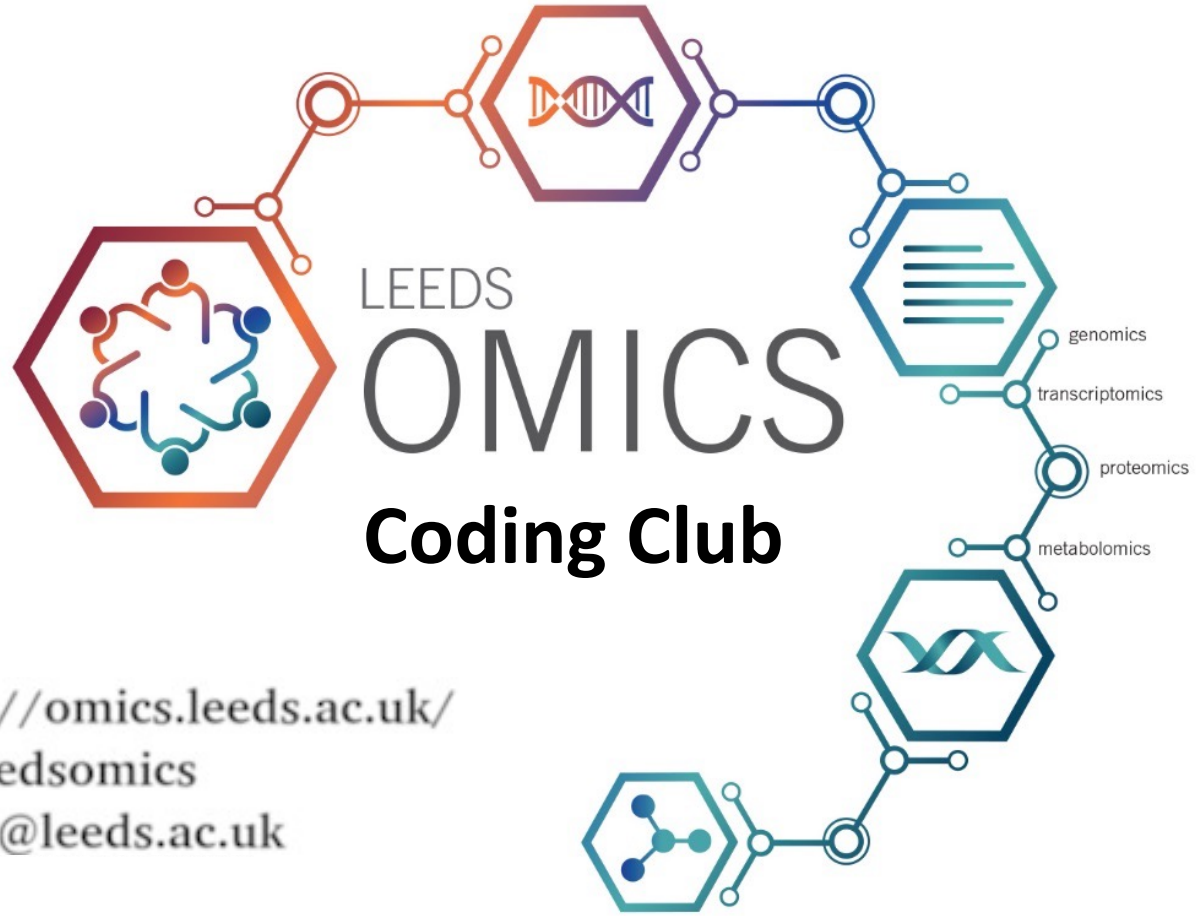


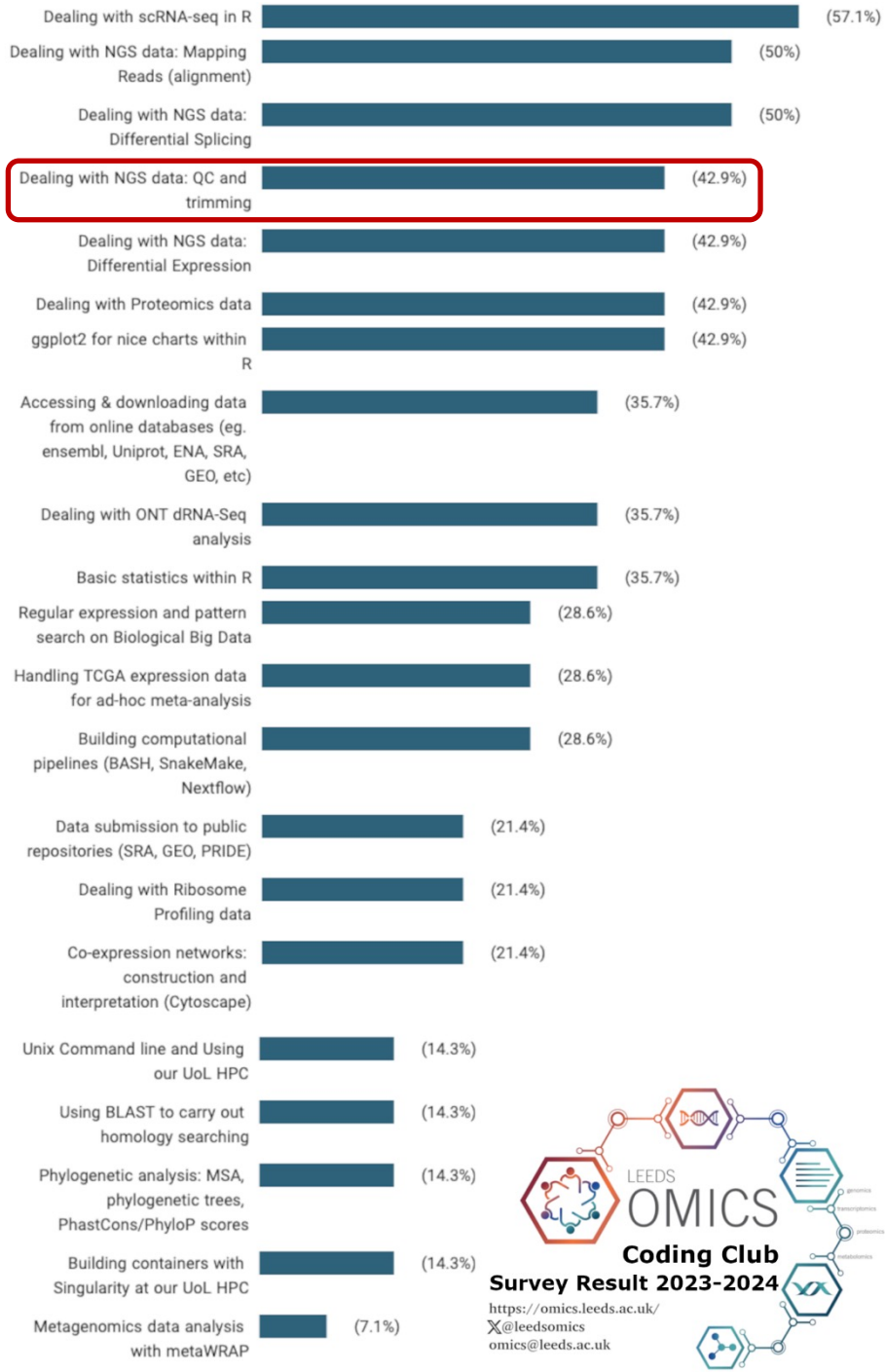
Coding Club

https://omics.leeds.ac.uk/
@leedsomics
omics@leeds.ac.uk

Dealing with NGS data: QC and trimming

Club Moderator(s): Elton Vasconcelos and Eilidh Ward

Topics to be addressed on the **2023-24** season - **Survey Result**



LEEDS OMICS

Coding Club

Survey Result 2023-2024

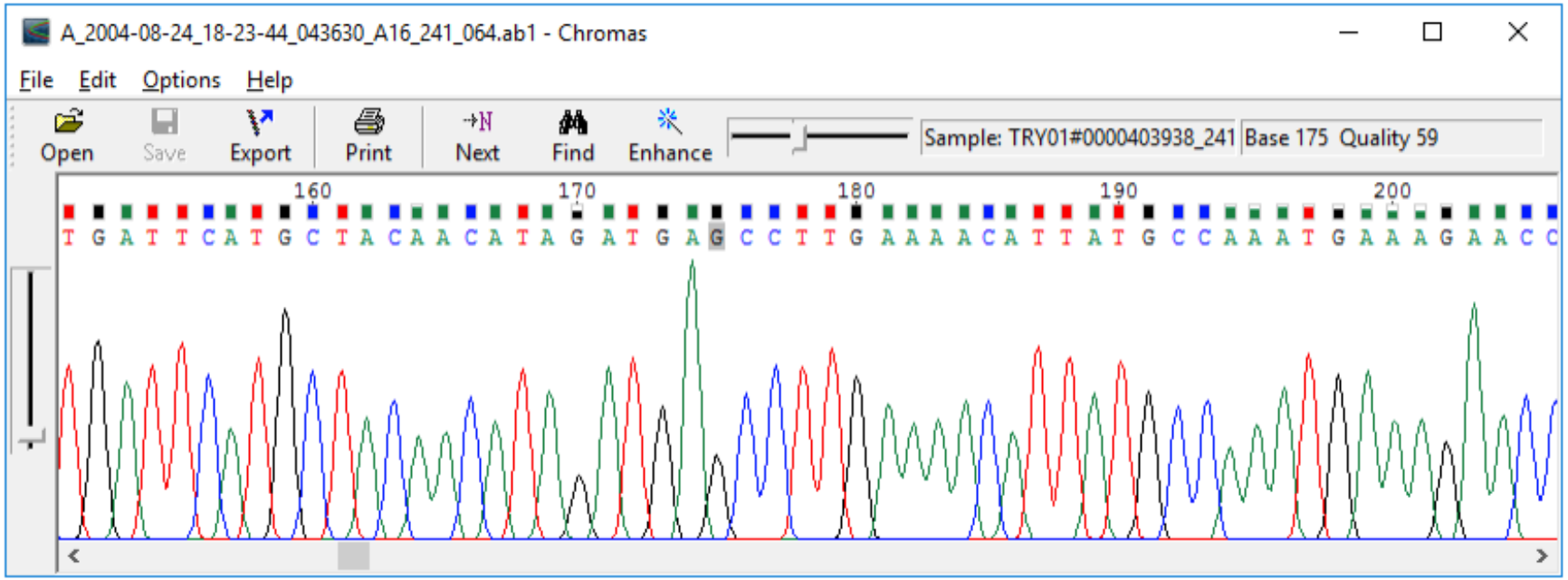
https://omics.leeds.ac.uk/

@leedsomics

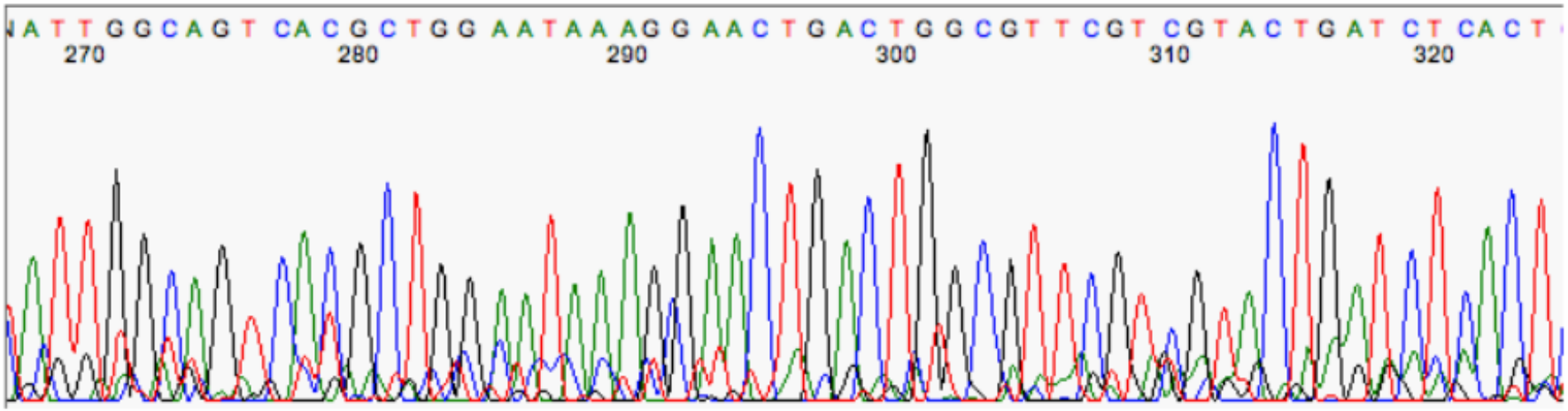
omics@leeds.ac.uk

Going back in time: **Sanger Sequencing Strategy**

Good quality sequencing



Not as good



https://www.phrap.com/phred/

Phred - Quality Base Calling

Phred is a base-calling program for DNA sequence traces. Phred reads DNA sequence chromatogram files and analyzes the peaks to call bases, assigning quality scores ("Phred scores") to each base call.

Phred was developed by Drs. Phil Green and Brent Ewing, and is distributed by CodonCode Corporation under license from the University of Washington. Phred is widely used by the largest academic and commercial DNA sequencing laboratories. This page gives a brief description of Phred. For information about Phrap, Cross_match, and Consed, please visit www.phrap.com.

Ewing et al.1998a, Genome Research 8:175-85

$$QV = -10 \log_{10} EP$$

EP = error probability

Phred quality score	Probability that the base is called wrong	Accuracy of the base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Mid 2000's: **Next-Generation Sequencing (NGS)**

Fastq format

R1 fastq file

```
1  @K00193:38:H3MYFBBXX:4:1101:10003:44458/1
2  TTCCTTATGAAACAGGAAGAGTCCCTGGGCCCAGGCCTGGCCCACGGTTGTCAAGGCACATCATTGCCAGCAAGCTGAAG
3  +
4  AAFFFKKKKKKKKKKKKKKKKKKKKKKKKKFKKFKKKKF<AAKKKKKKKKKKKKKKKKFKKKFKKKKKKKKKKKKFKAFKKKKKKKI
5  @K00193:38:H3MYFBBXX:4:1101:10003:46427/1
6  CGCCTGGCGACCTTGGTGTCCGCGATGTGGATCATGTCCTTATCAATGTAGAAGTCAACCTCTGAGGCTGACTGAAATTCA
7  +
8  ,AFFFFKKKKFKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKAKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKI
9  @K00193:38:H3MYFBBXX:4:1101:10003:5079/1
10 CTTTGACCTTGATGAGGATCTGCAGGTCCTCCAGGTCCTTGCCATCCCACTTGTCAAAAGGCTCCAGGAGCTGCAGGCGCTC
11 +
12 AAFFFKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKK
```

R2 fastq file

```
1  @K00193:38:H3MYFBBXX:4:1101:10003:44458/2
2  ATGGAACTCCCTACTGACCTCTGAGAACTGGAAACGAGTTTGTACAGAAGTCAGAACTTTGGGTTGGGAATGAGATCT
3  +
4  A<FFFKKKKKKKKKKKKKKKKKKKKKKKFKAKKKKKKK7FFKKKKKKKKK7FKKFKKKKKKKKKKKKKKKFKKFKKKKKKK
5  @K00193:38:H3MYFBBXX:4:1101:10003:46427/2
6  CTGTGGCCAAGCTGGCTGGCTTCCTGGACCTCACAGAGCAGGAGTTCGGATCCAGCTTCTTGTCTTCAAACACAAGAT
7  +
8  AAFAFKKKKK7FFFFFKKK<KKKKKKKKKFK7AA<KKKKKKKKKKKAKKKKFAKKKKKKKKKKKKKKKKKKKKKKAFKKKI
9  @K00193:38:H3MYFBBXX:4:1101:10003:5079/2
10 GAGCTTCCCAAAGGGGAGTTTGACCCAGGGCAGGACACCTACCAGCACCCACCCAAGGACAGCAGCGGGCAGCATG
11 +
12 AAFFFKKKKKKKKKKKKFKKKKKKKKKKKKKKFKKKKKKKKKKKFKKKKKKFKKKFKKKKKKKKKKKKKKKKKKKFKKK
```

Line 1: This line contains the character '@' and followed by the sequences.

Line 2: The reads sequence.

Line 3: To save space, this line is now redundant and just has the '+' character.

Line 4: The quality string for the linked sequence. The quality string is encoded as an ASCII character.

American Standard Code for Information Interchange (ASCII)

Code	Char	Code	Char	Code	Char	Code	Char	Code	Char	Code	Char
32	[space]	48	0	64	@	80	P	96	`	112	p
33	!	49	1	65	A	81	Q	97	a	113	q
34	"	50	2	66	B	82	R	98	b	114	r
35	#	51	3	67	C	83	S	99	c	115	s
36	$	52	4	68	D	84	T	100	d	116	t
37	%	53	5	69	E	85	U	101	e	117	u
38	&	54	6	70	F	86	V	102	f	118	v
39	'	55	7	71	G	87	W	103	g	119	w
40	(	56	8	72	H	88	X	104	h	120	x
41	)	57	9	73	I	89	Y	105	i	121	y
42	*	58	:	74	J	90	Z	106	j	122	z
43	+	59	;	75	K	91	[	107	k	123	{
44	,	60	<	76	L	92	\	108	l	124	|
45	-	61	=	77	M	93	]	109	m	125	}
46	.	62	>	78	N	94	^	110	n	126	~
47	/	63	?	79	O	95	_	111	o	127	[backspace]

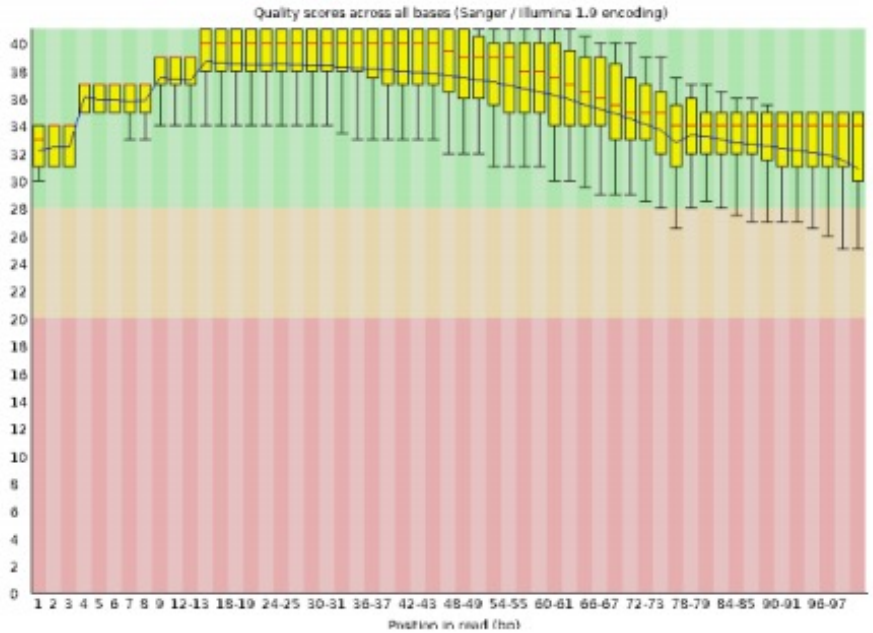
ASCII table for character and encoding

-> Fastq format Phred Quality Score-decoding

https://en.wikipedia.org/wiki/FASTQ_format

FastQC for Quality Control (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

Per base sequence quality

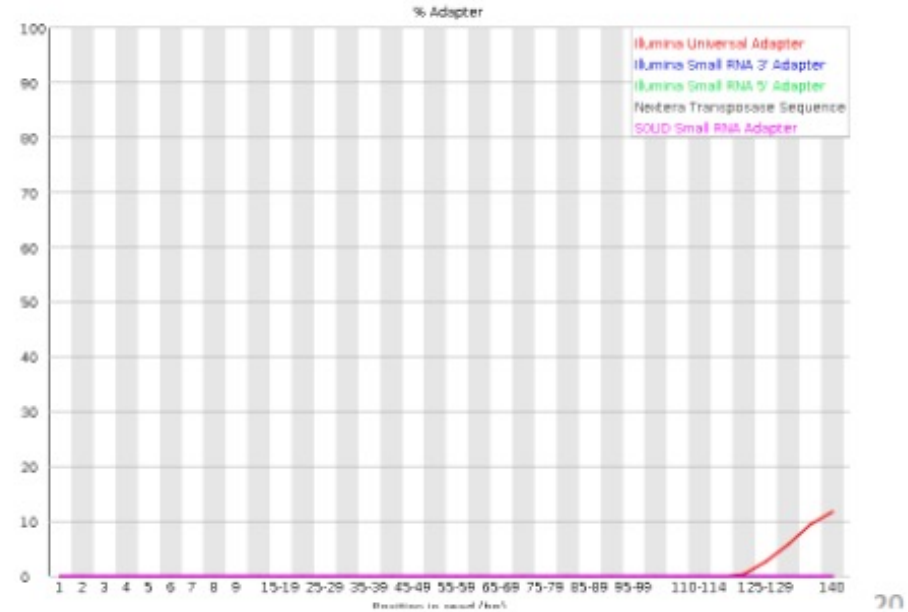



```
$ fastqc -t 4 --nogroup *.fastq.gz
```

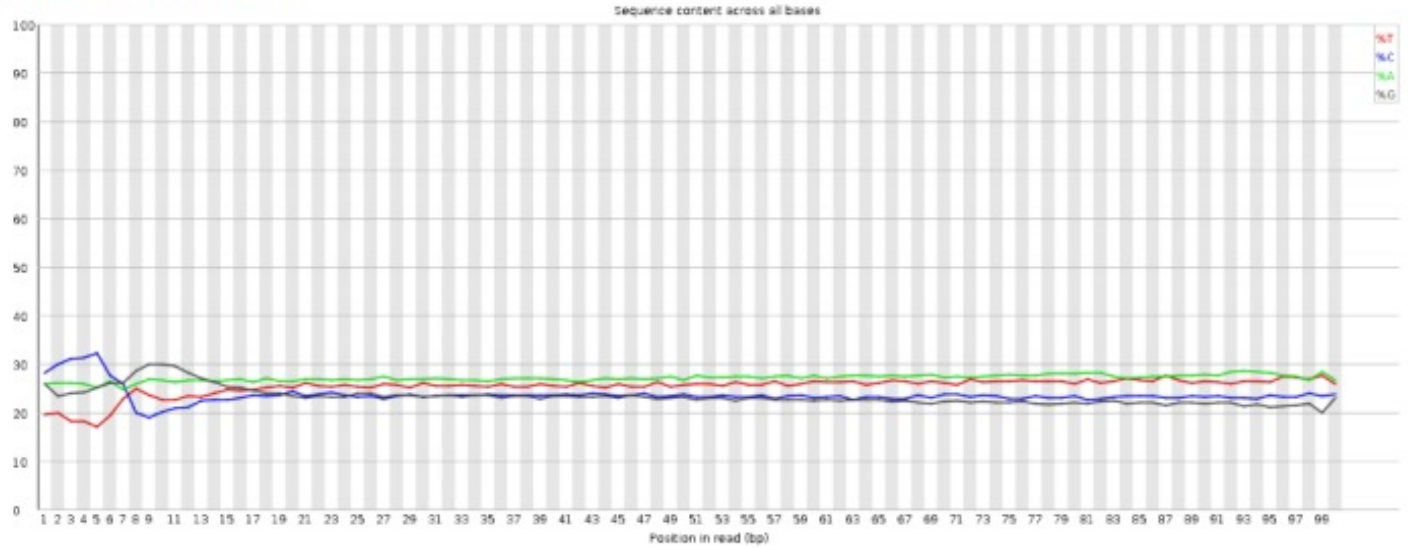
- **MultiQC** (https://multiqc.info/)

```
$ multiqc –o multiQC_out/ .
```

Adapter Content



Per base sequence content



Trimming adapters and low quality bases at reads' extremities

- Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic)
- Cutadapt (https://cutadapt.readthedocs.io/en/stable/)
- Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)

Example of trimmomatic command line:

```
$ java -jar /your/trimmomatic/folder/path/trimmomatic-0.36.jar PE -threads 24 -phred33
input_R1.fastq.gz input_R2.fastq.gz output_R1-trimmed_paired.fq.gz output_R1-
trimmed_unpaired.fq.gz output_R2-trimmed_paired.fq.gz output_R2-trimmed_unpaired.fq.gz
HEADCROP:15 ILLUMINACLIP:/your/trimmomatic/folder/path/adapters/user-
defined_adapters_file.fasta:2:30:7 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:30
```

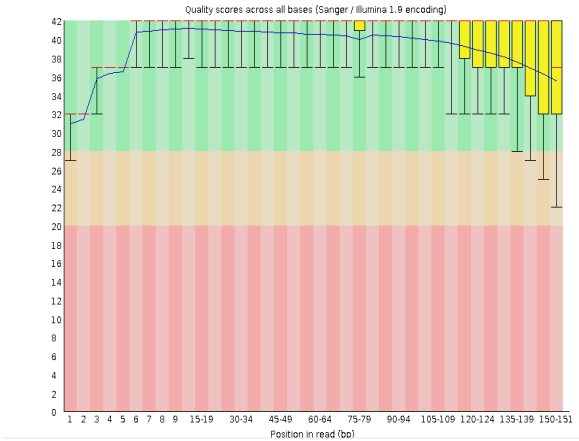
Re-analyse the trimmed data using FastQC

Before trimming

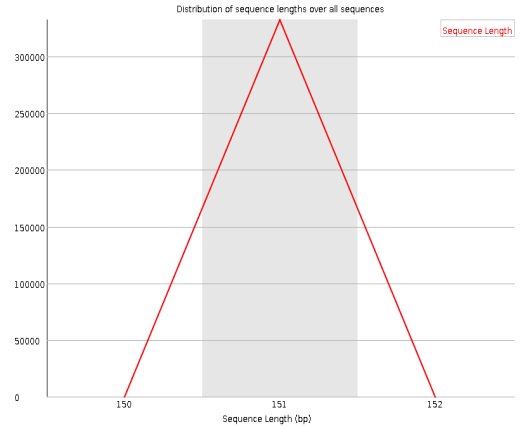
Basic Statistics

Measure	Value
Filename	normal_rep1_r1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	331958
Sequences flagged as poor quality	0
Sequence length	151
%GC	54

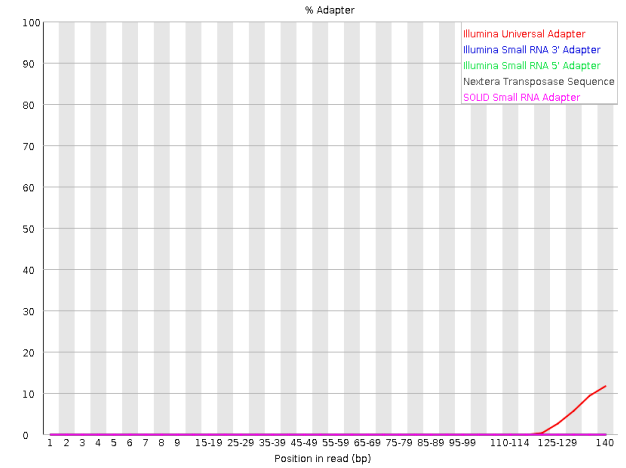
Per base sequence quality



Sequence Length Distribution



Adapter Content

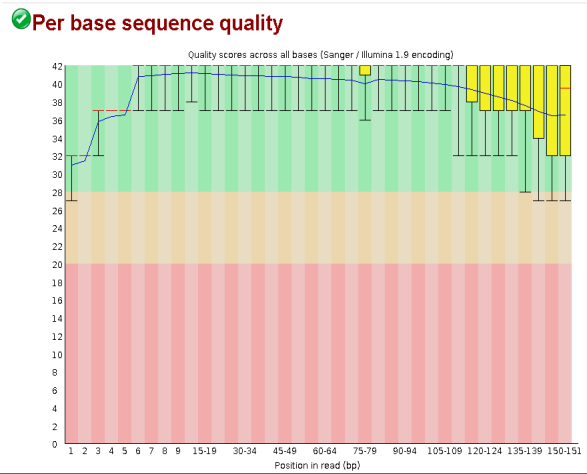


After trimming

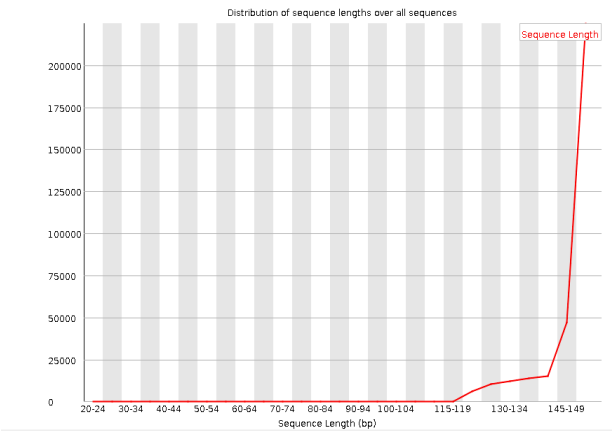
Basic Statistics

Measure	Value
Filename	normal_rep1_r1_val_1.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	331389
Sequences flagged as poor quality	0
Sequence length	21-151
%GC	54

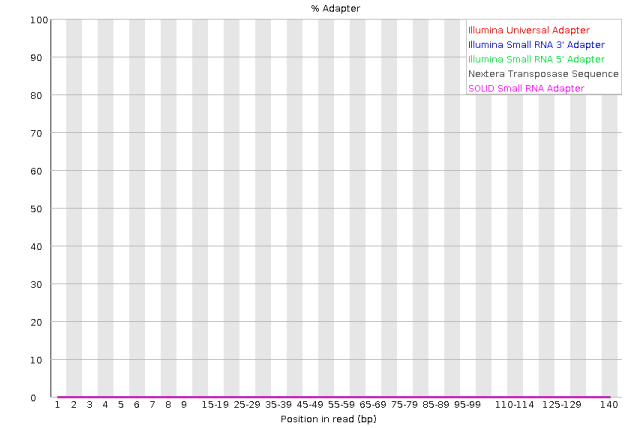
Per base sequence quality



Sequence Length Distribution



Adapter Content



Bring your issues on!