Data Formats
Data Quality
Read Mapping

Riddhiman Dhar
Department of Biotechnology
IIT Kharagpur

Data formats

- Roche 454

  - FASTA format

- Illumina

  - FASTQ format

Data formats

- PacBio

  - basecall HDF5 format earlier

  - currently BAM format (convertible to fastq format)

- Ion-torrent

  - DAT files (Raw data)

  - BAM, FASTQ or VCF format

- Nanopore

  - HDF5 format

  - FAST5 format

Fastq format

- Fasta format with Quality scores

Fastq format

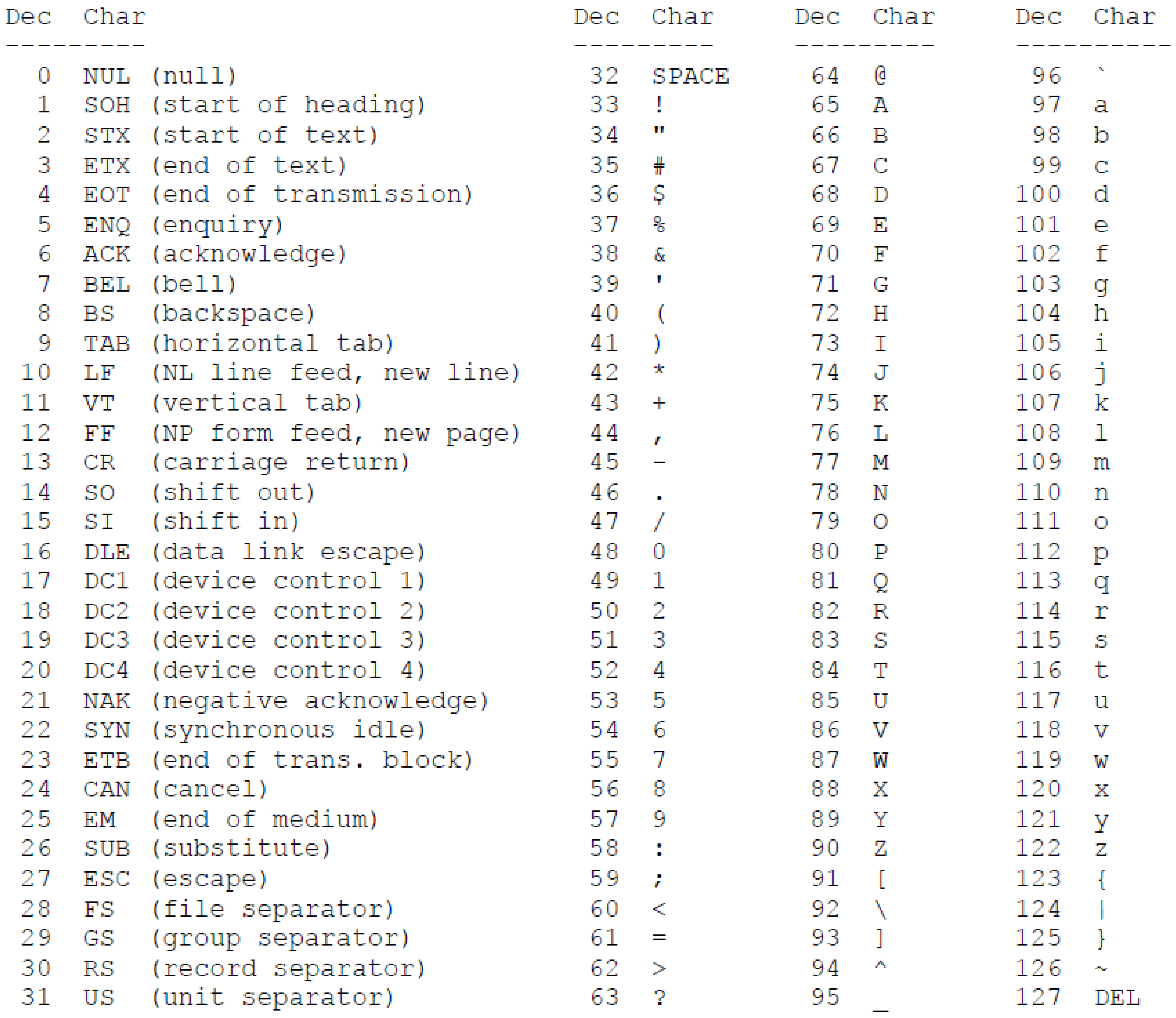
- Fasta format with Quality scores

- @D00733:181:CAH6EANXX:8:2210:1499:2056 1:N:0:GTAGAG
- NCTTTGTACTATGACCGATACACTCAACCGGCGAAAAGTGGAACTTGAGAATTGATGTCTTCATCTTATTATCTGTCTCTTATACACATCTCCGAGCCCACGAGACGTAGAGGAATCTCGTATGC
- +
- #<=BBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG@GGGGGGGGGGGGGEGGEGG8CEGGGGGGG
- @D00733:181:CAH6EANXX:8:2210:1668:2177 1:N:0:GTAGAG
- GCTTTTTTTACGGCAGCATTTTTTTTTCAACTCTGATCGCCCCTTTACTGCTCCCTCCGCCCAAATTCCATTGCAGTTCAAATGTATACTGAAAAAAACCCCATTGCTATTGTTAAACAGTGAAC
- +
- BBCCCGGGGBFGGGGGGEGGGGGGGGGGCGFECCGGGGGGGGGGCDGGGGGGGGED>BGGGGBD@CGGEG>CCFGDCCGCDGG8FCGG=FEGGGGGGGDGDDEGG/6D@/DGGCB//6C9CD/

Header: @D00733:181:CAH6EANXX:8:2210:1668:2177 1:N:0:GTAGAG

Element	Requirements	Description
@	@	Each sequence identifier line starts with @
<instrument>	Characters allowed: a–z, A–Z, 0–9 and underscore	Instrument ID
<run number>	Numerical	Run number on instrument
<flowcell ID>	Characters allowed: a–z, A–Z, 0–9	
<lane>	Numerical	Lane number
<tile>	Numerical	Tile number
<x_pos>	Numerical	X coordinate of cluster
<y_pos>	Numerical	Y coordinate of cluster
<read>	Numerical	Read number. 1 can be single read or Read 2 of paired-end
<is filtered>	Y or N	Y if the read is filtered (did not pass), N otherwise
<control number>	Numerical	0 when none of the control bits are on, otherwise it is an even number. On HiSeq X and NextSeq systems, control specification is not performed and this number is always 0.
<sample number>	Numerical	Sample number from sample sheet

Quality score – ASCII table



Dec	Char	Dec	Char	Dec	Char	Dec	Char
0	NUL (null)	32	SPACE	64	@	96	`
1	SOH (start of heading)	33	!	65	A	97	a
2	STX (start of text)	34	"	66	B	98	b
3	ETX (end of text)	35	#	67	C	99	c
4	EOT (end of transmission)	36	$	68	D	100	d
5	ENQ (enquiry)	37	%	69	E	101	e
6	ACK (acknowledge)	38	&	70	F	102	f
7	BEL (bell)	39	'	71	G	103	g
8	BS (backspace)	40	(	72	H	104	h
9	TAB (horizontal tab)	41	)	73	I	105	i
10	LF (NL line feed, new line)	42	*	74	J	106	j
11	VT (vertical tab)	43	+	75	K	107	k
12	FF (NP form feed, new page)	44	,	76	L	108	l
13	CR (carriage return)	45	-	77	M	109	m
14	SO (shift out)	46	.	78	N	110	n
15	SI (shift in)	47	/	79	O	111	o
16	DLE (data link escape)	48	0	80	P	112	p
17	DC1 (device control 1)	49	1	81	Q	113	q
18	DC2 (device control 2)	50	2	82	R	114	r
19	DC3 (device control 3)	51	3	83	S	115	s
20	DC4 (device control 4)	52	4	84	T	116	t
21	NAK (negative acknowledge)	53	5	85	U	117	u
22	SYN (synchronous idle)	54	6	86	V	118	v
23	ETB (end of trans. block)	55	7	87	W	119	w
24	CAN (cancel)	56	8	88	X	120	x
25	EM (end of medium)	57	9	89	Y	121	y
26	SUB (substitute)	58	:	90	Z	122	z
27	ESC (escape)	59	;	91	[	123	{
28	FS (file separator)	60	<	92	\	124	|
29	GS (group separator)	61	=	93	]	125	}
30	RS (record separator)	62	>	94	^	126	~
31	US (unit separator)	63	?	95	_	127	DEL

HDF5 format

- Hierarchical Data Format (HDF)
- Nested data

- Groups

- Datasets

- Attributes

FAST5 format

- Specifically structured HDF5 data
- “Raw” data and “Analysis” data
- Raw data – values of pico-amp currents in nanopore
- Analysis data – basecall data

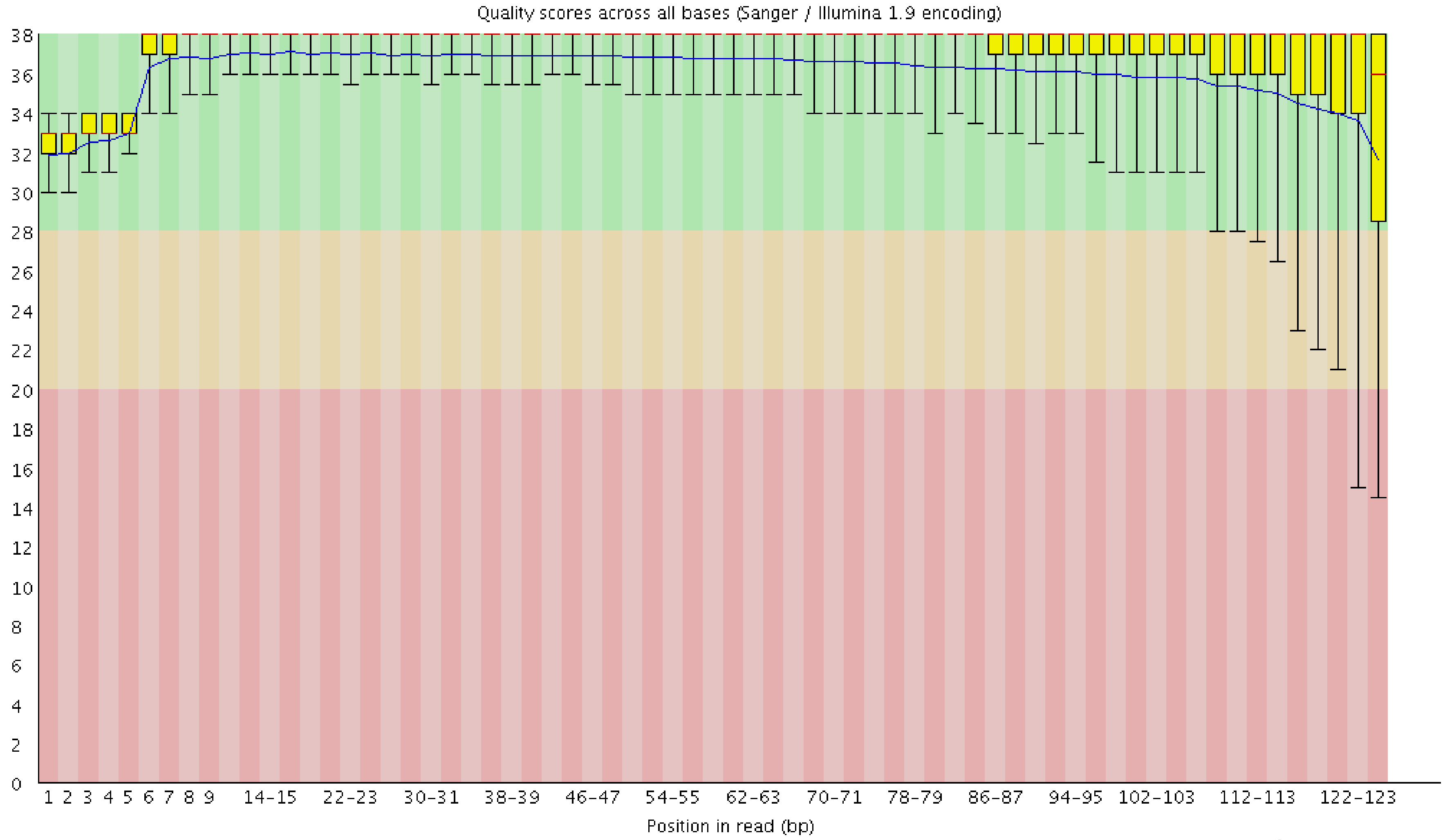
Checking NGS Data Quality

Checking NGS Data Quality

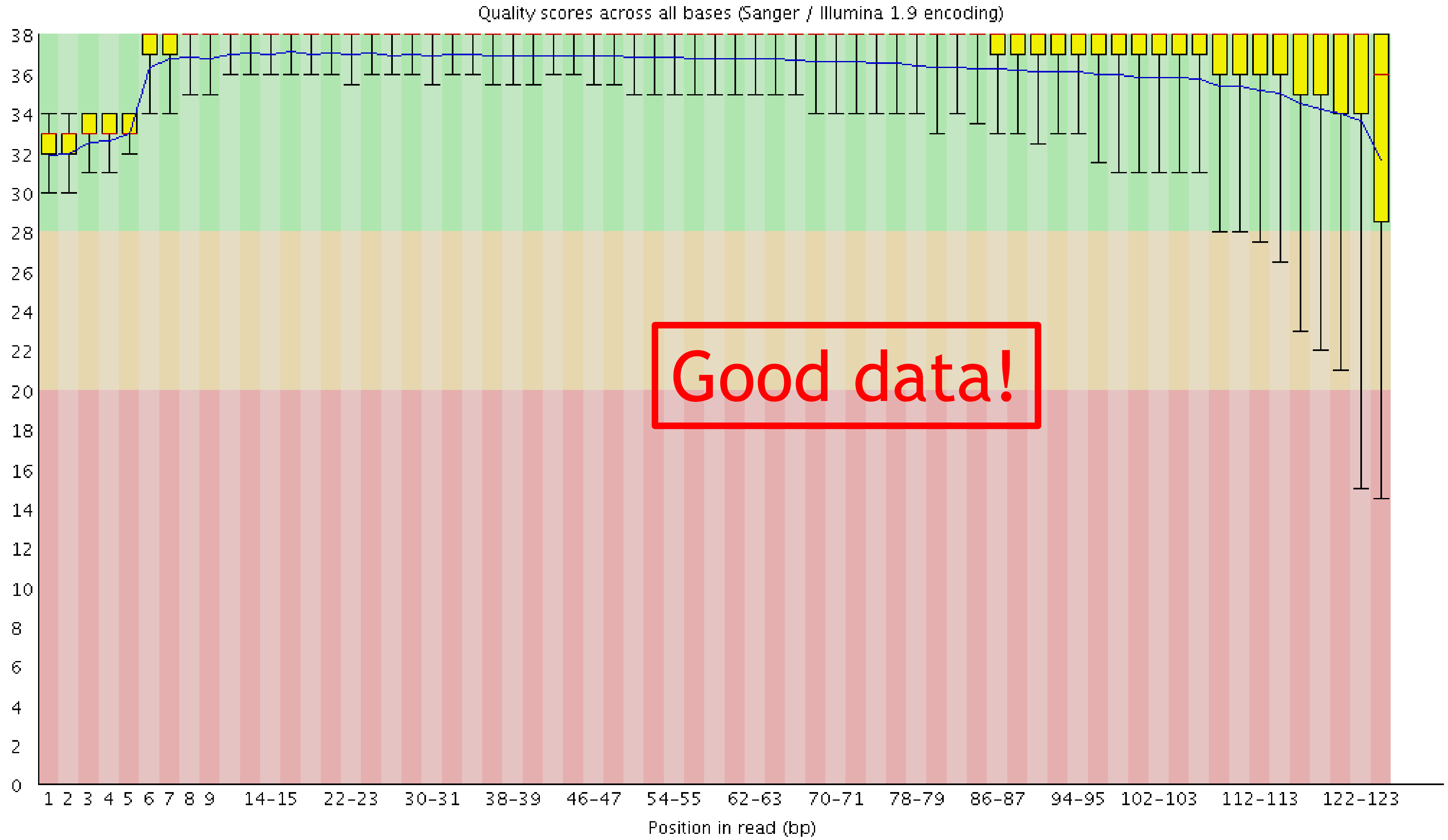
FastQC

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

Per base quality score

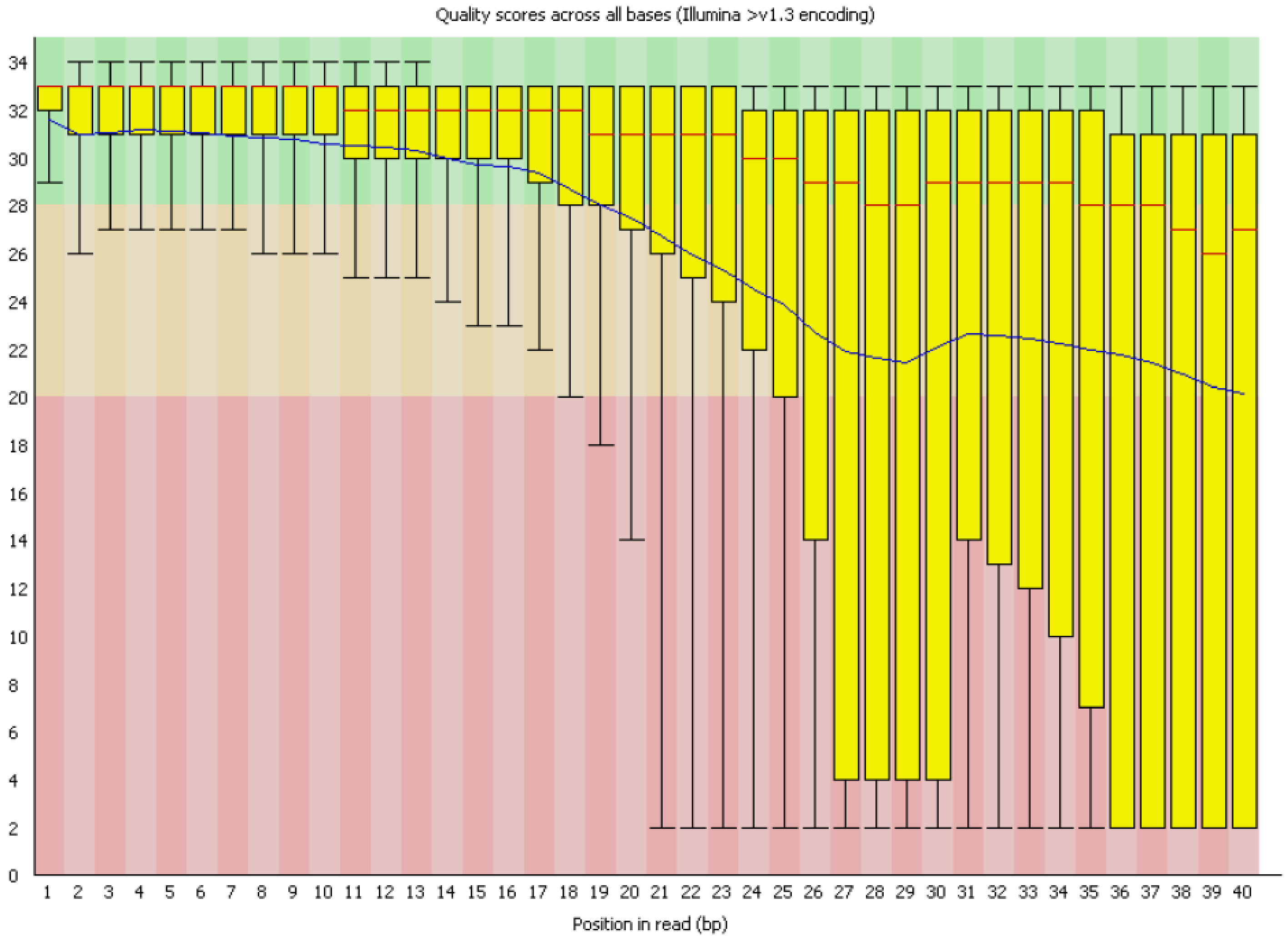


Per base quality score



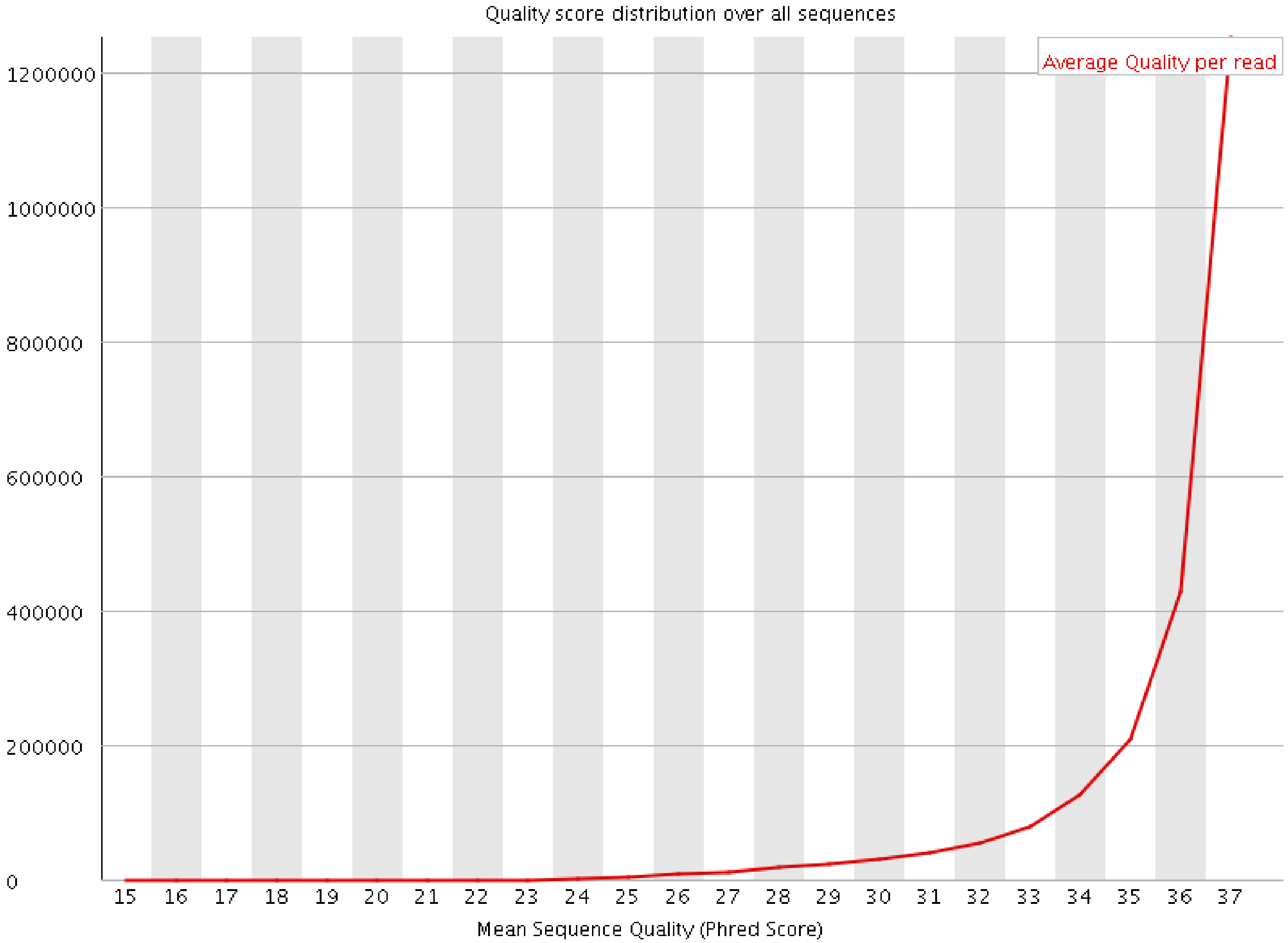
Good data!

Per base quality score

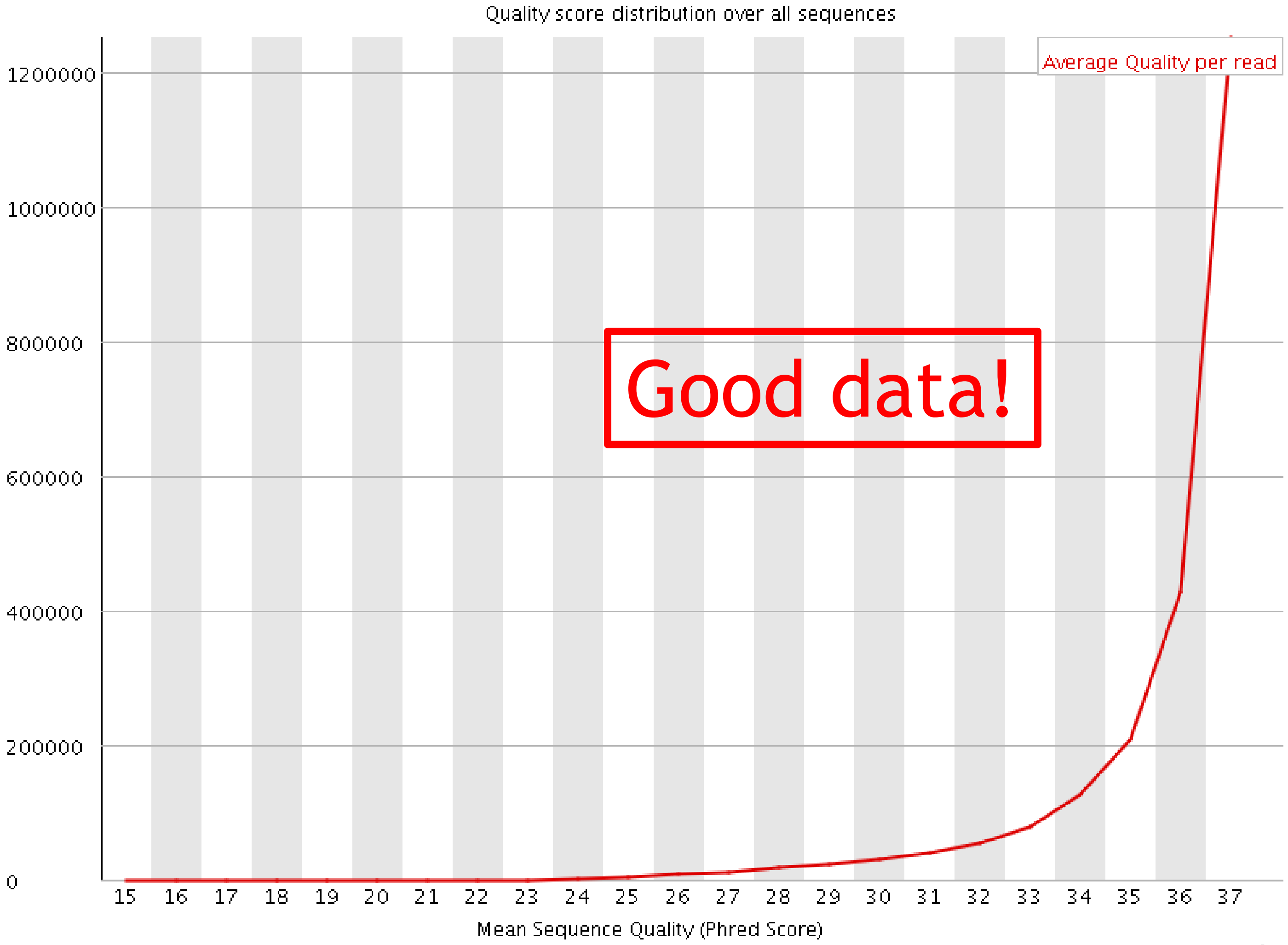


Bad data!

Per sequence quality scores

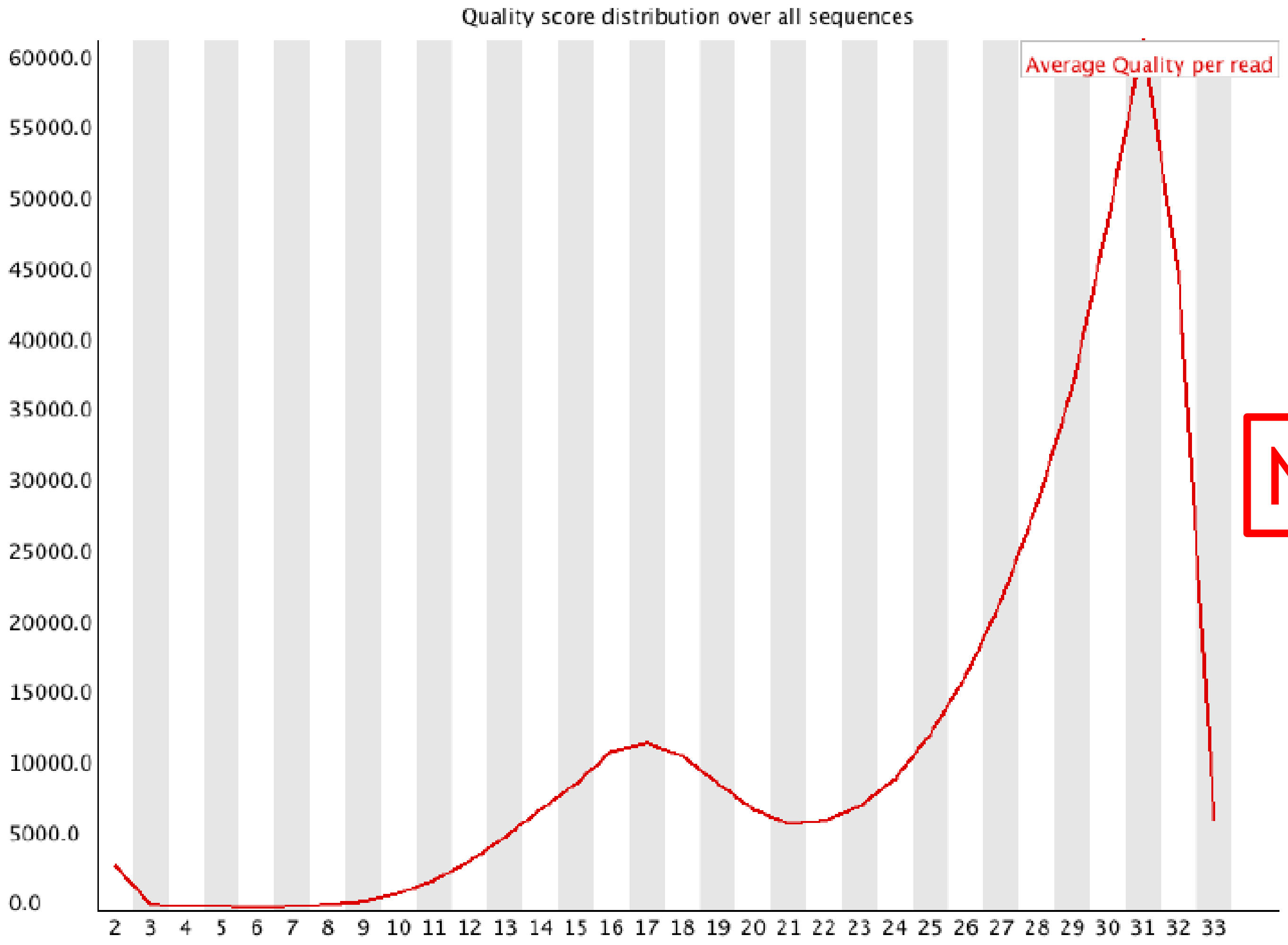


Per sequence quality scores



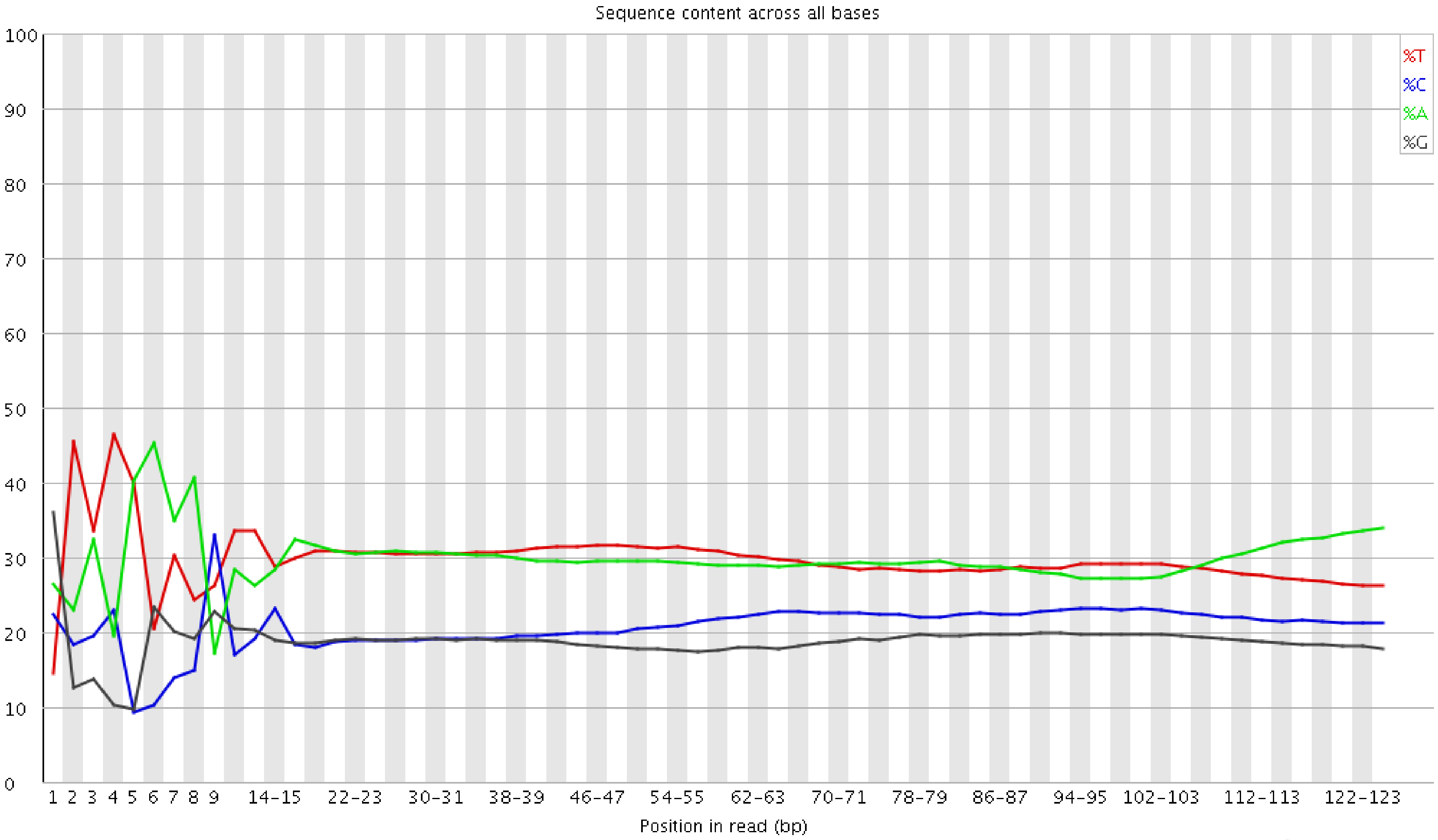
Good data!

Per sequence quality scores

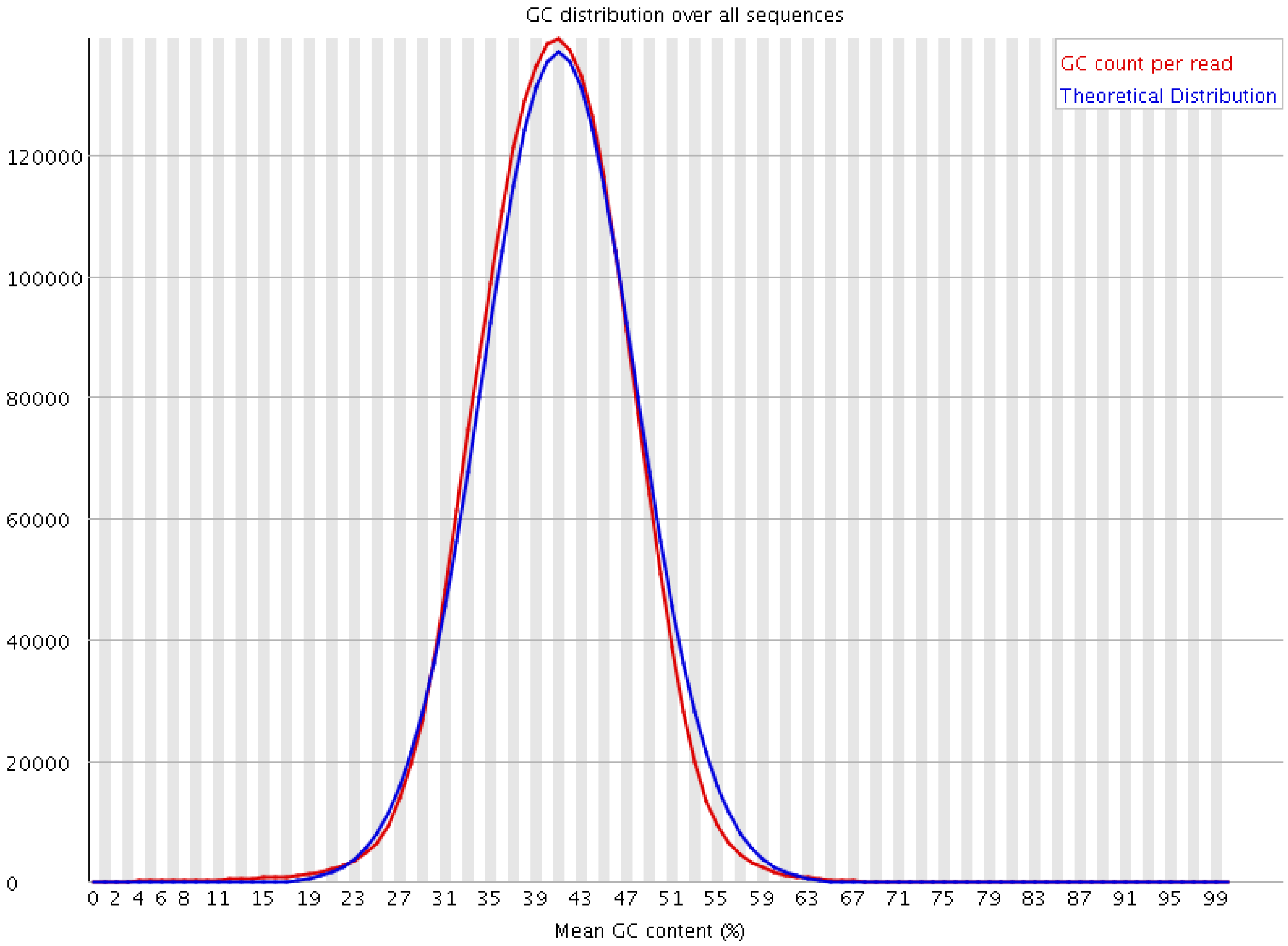


Not so good!

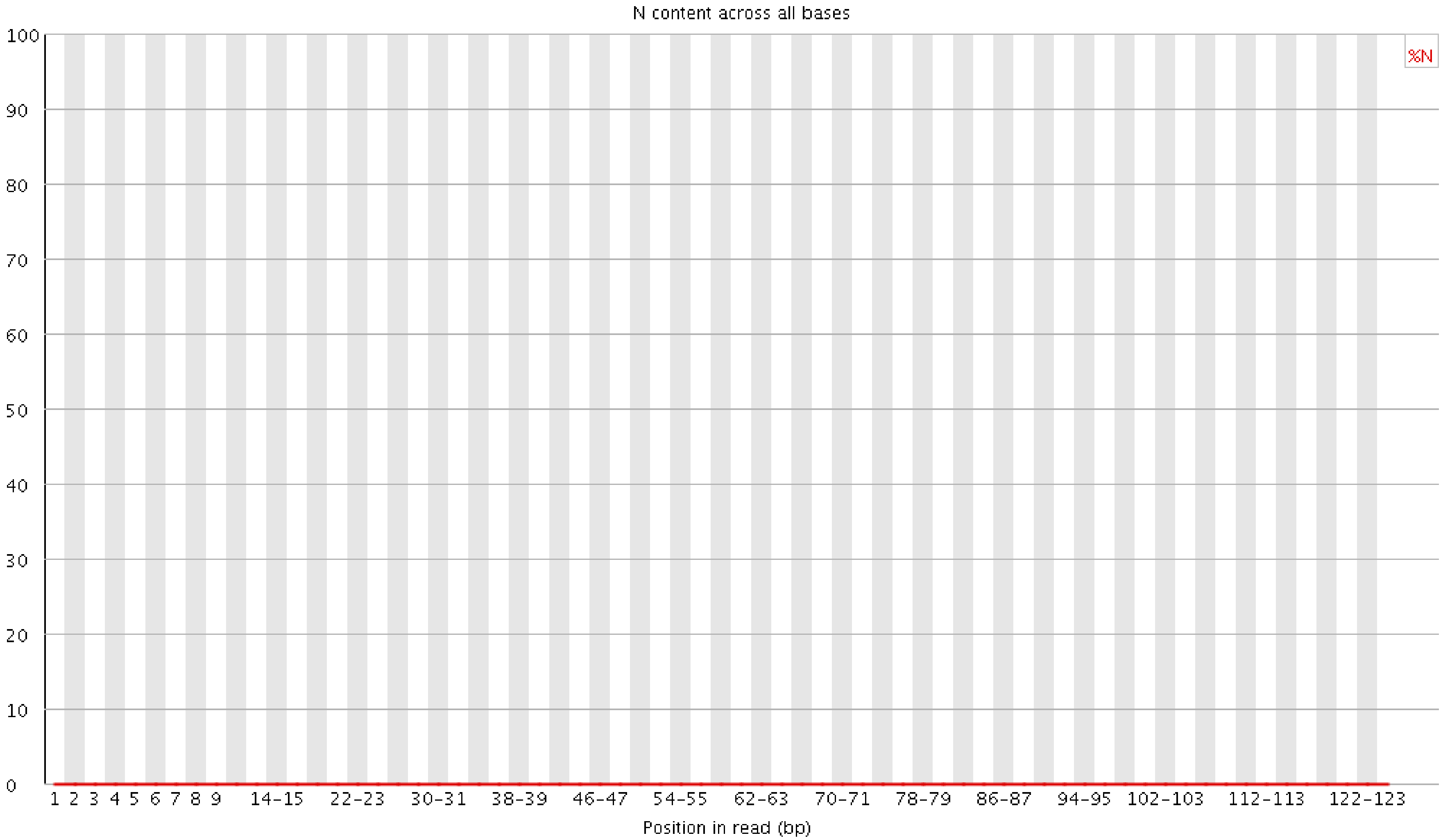
Per base sequence content



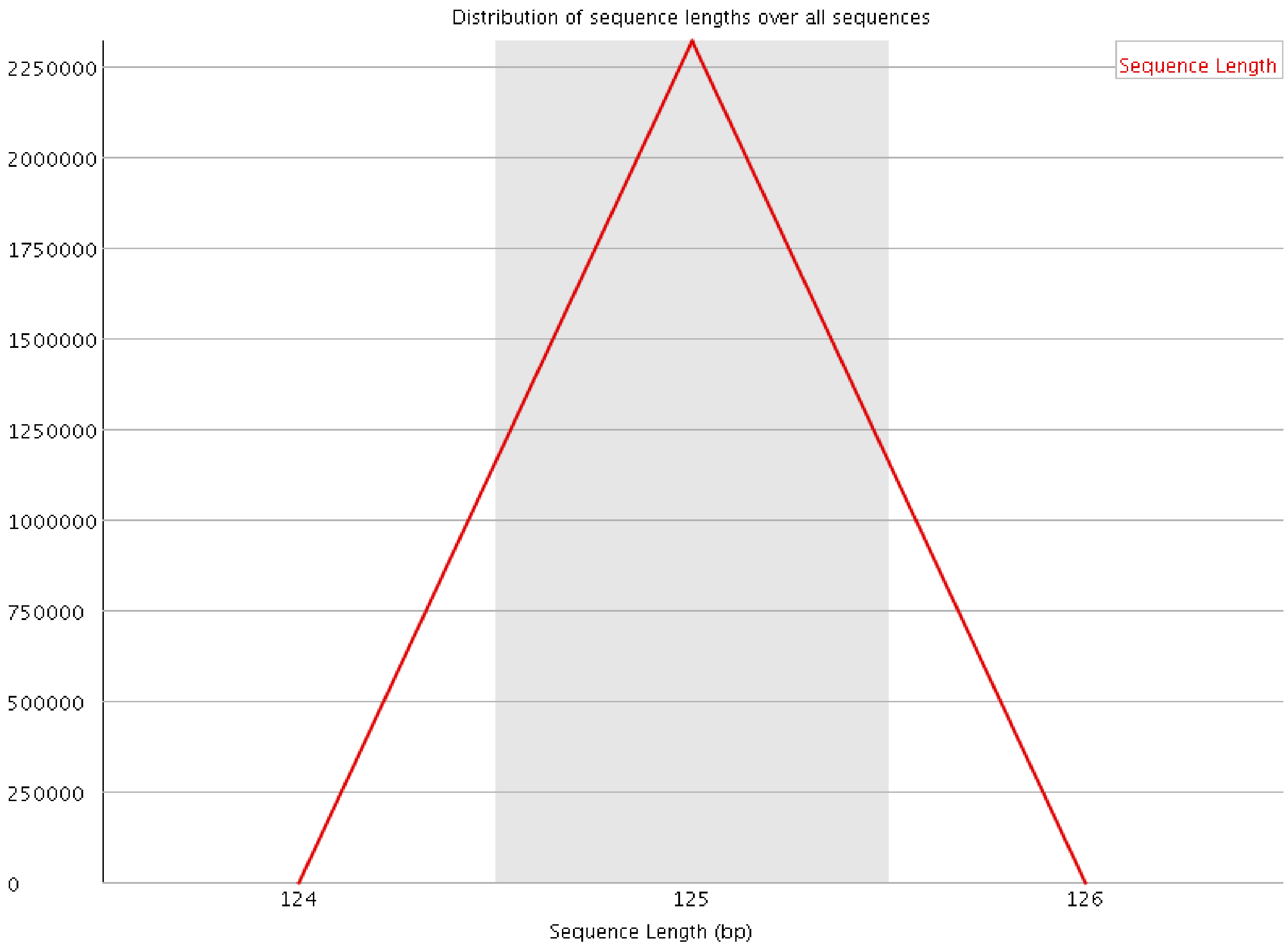
Per sequence GC content



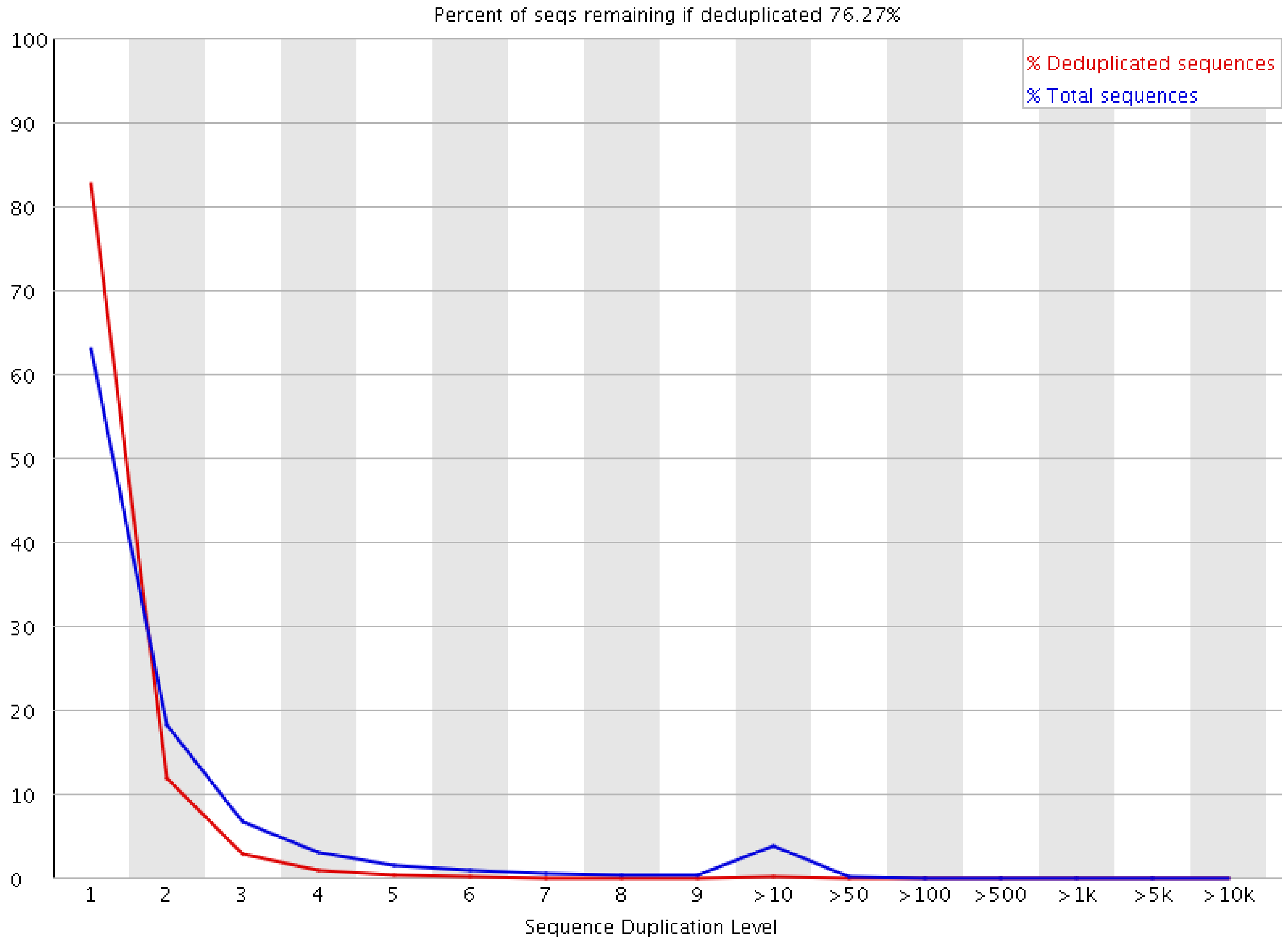
Per Base N content



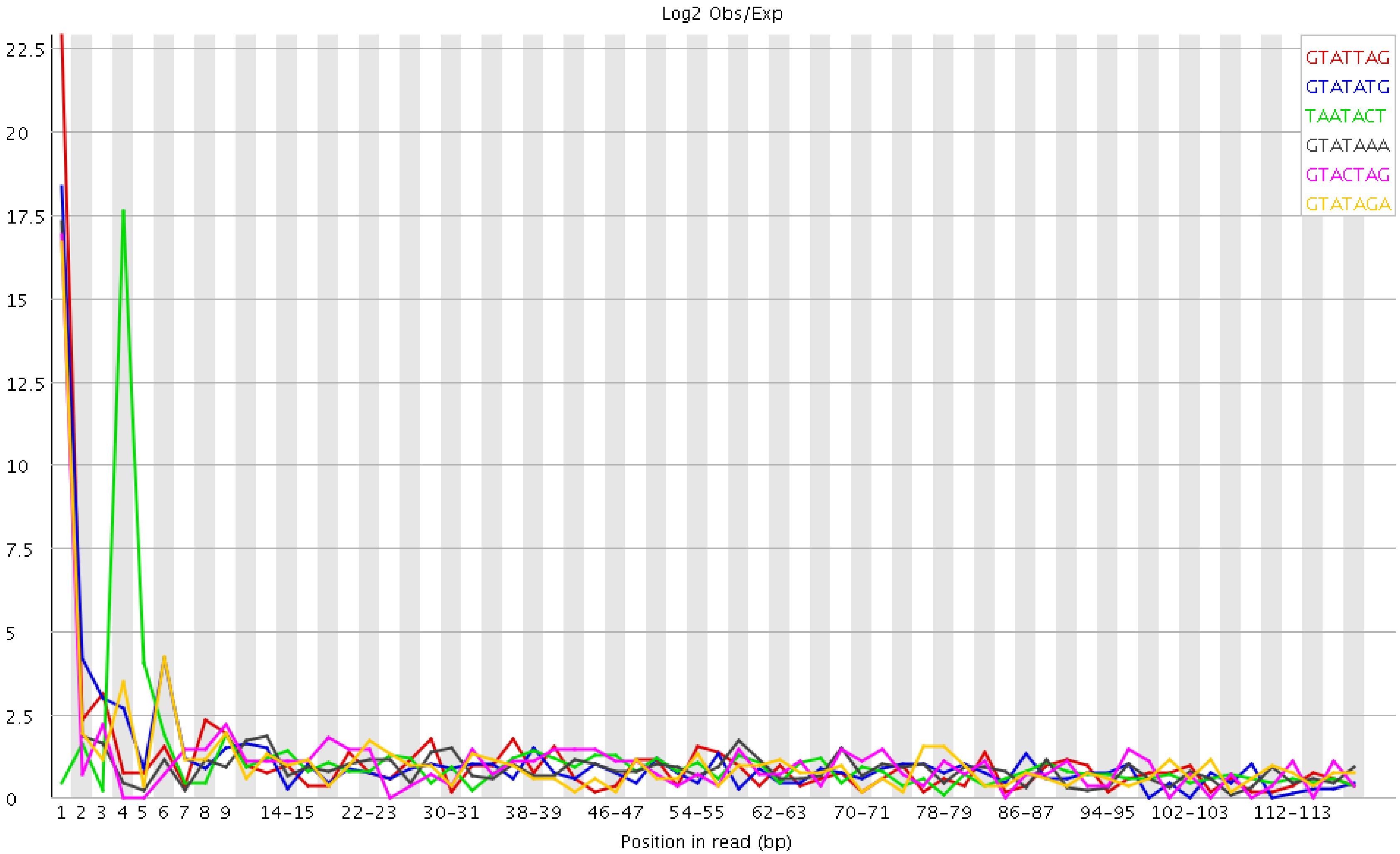
Sequence length distribution



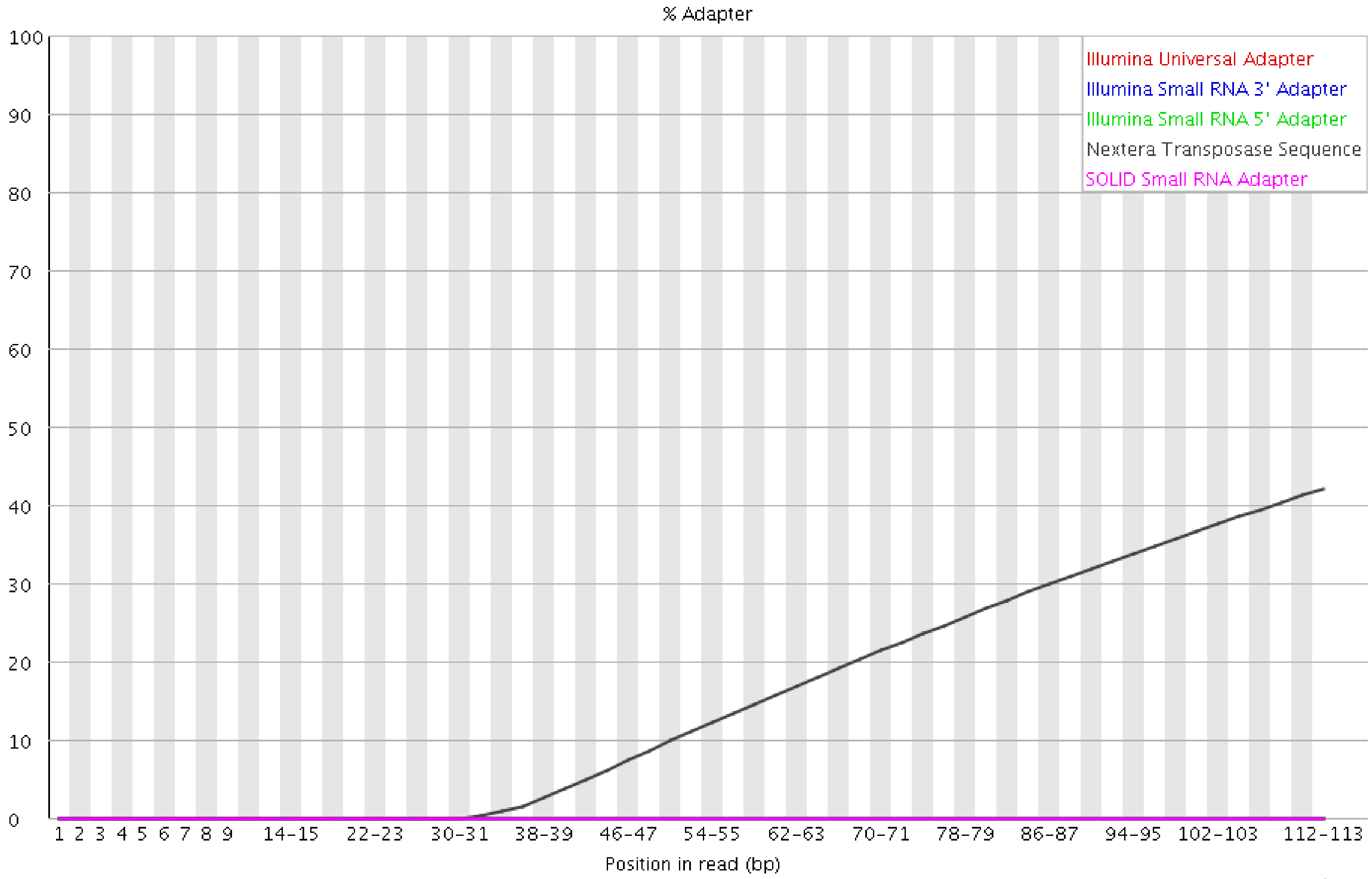
Duplicate Sequences



Over-represented k-mers



Adapter content



Read trimming tools

- Trimmomatic

  http://www.usadellab.org/cms/?page=trimmomatic

- bbduk

  https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/

- Cutadapt

  https://pypi.org/project/cutadapt/1.3/