IX International Course of Massive Data Analysis FOR GENOMICS

Francisco García
fgarcia@cipf.es

Quality Control for NGS Raw Data

Contents

- Data formats
  - Sequence capture
  - Fasta and fastq formats
  - Sequence quality encoding
- Quality Control
  - Evaluation of sequence quality
  - Quality control tools
  - Identification of artifacts & filtering
- Practical session

Sequence capture

illumina

Roche

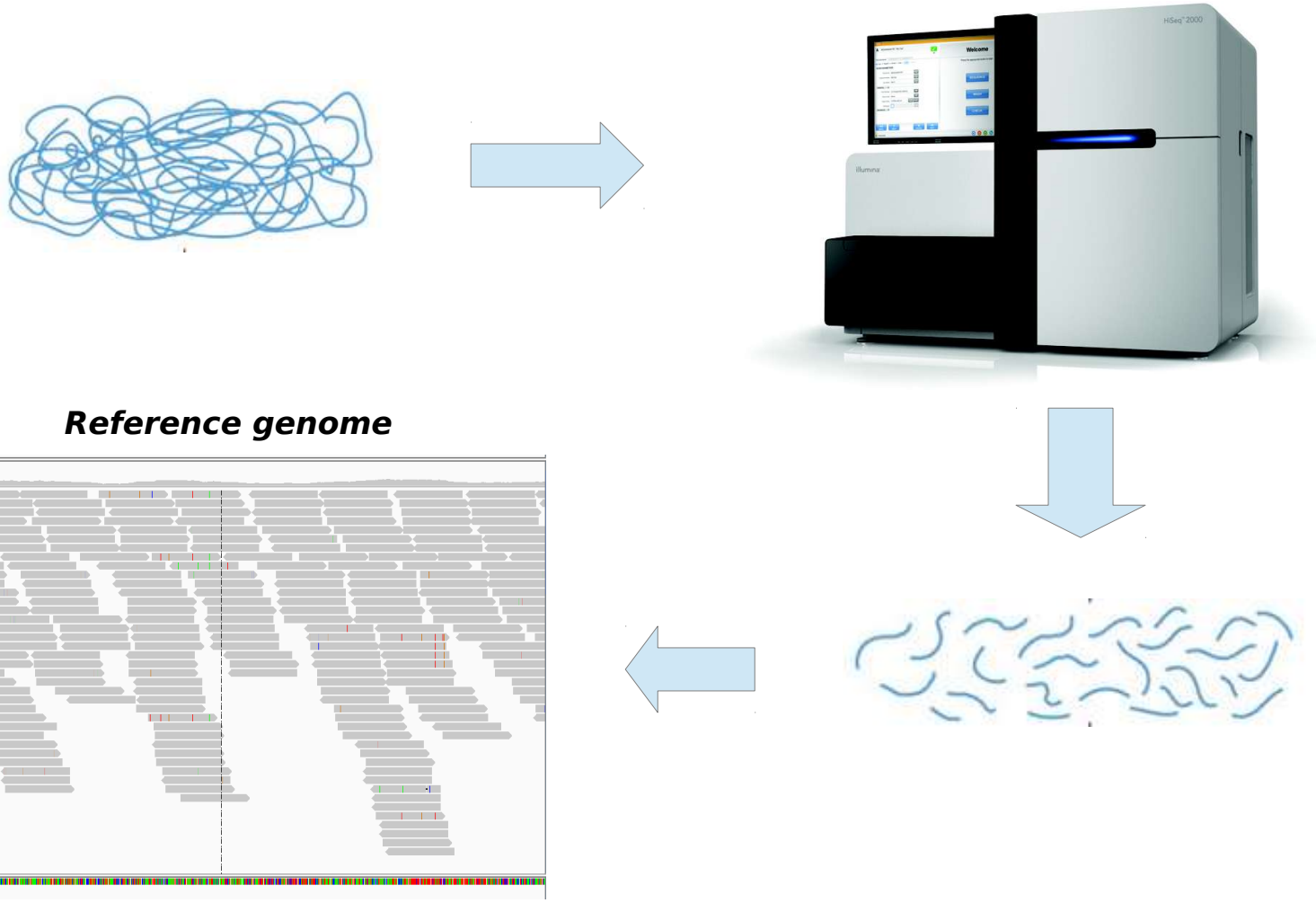
applied biosystems

RAW data
Propietary format

FastQ

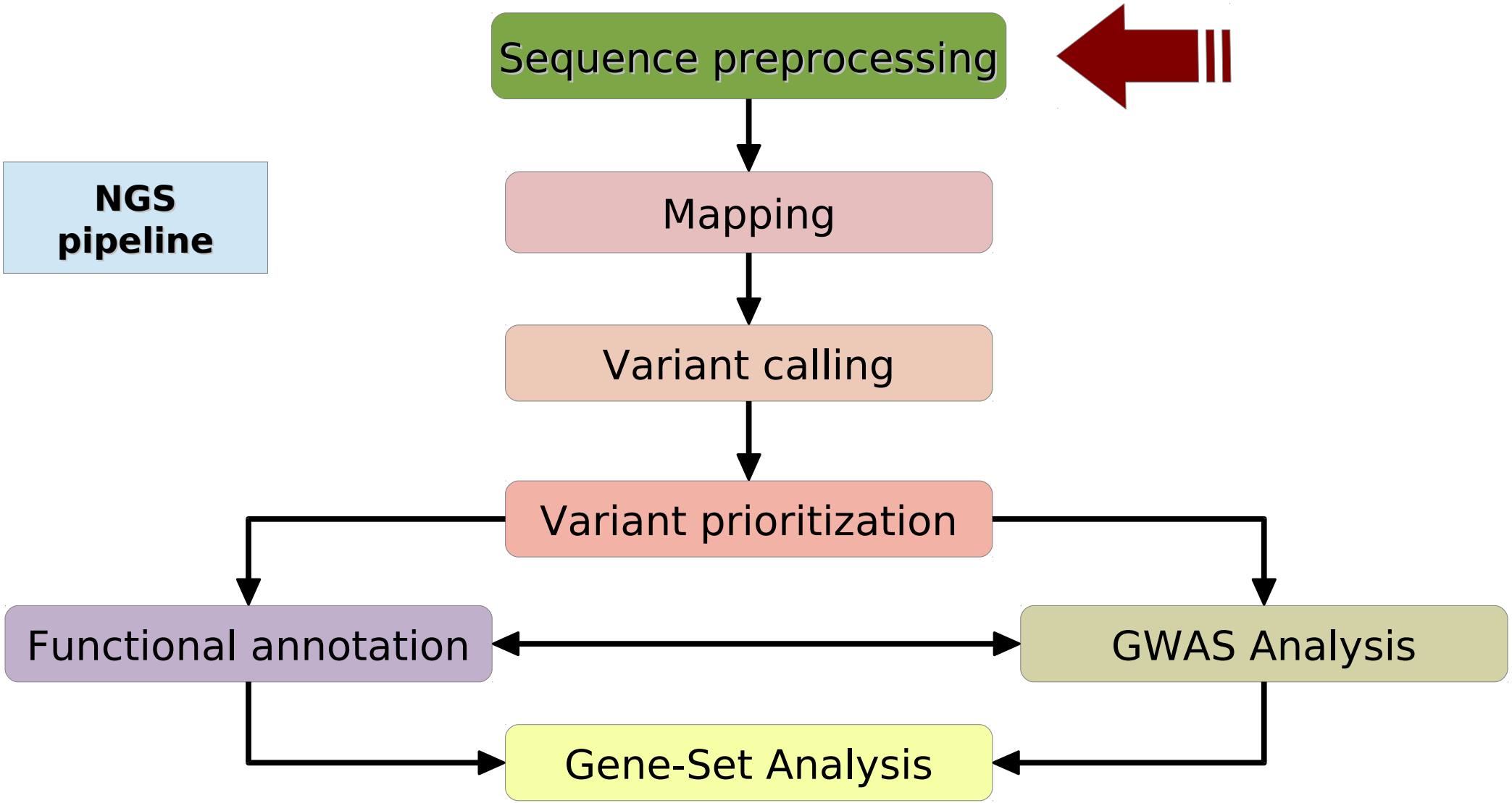
Francisco García
fgarcia@cipf.es

Genome sequencing



Francisco García
fgarcia@cipf.es

Where are we?



Francisco García
fgarcia@cipf.es

From sequencers to digital data

- **What structure does the data have?**
  - Text-based formats (easy to use!)
  - If not compressed, it can be huge

- **Different data formats:**
  - Different sequencers output different files (sff, fasta, csfasta, qual file, fastq...)
  - There are some data formats widely accepted (e.g. FastQ format)

Fasta format

- Two lines per sequence:
  - 1. Header lines starts with “>” followed by a sequence ID
  - 2. Sequence (string of nt or peptides)

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

```
>BBTBSCRYR
tgcaccaaacatgtctaaagctggaaccaaaattactttctttgaagacaaaaactttca
aggccgccactatgacagcgattgcgactgtgcagatttccacatgtacctgagccgctg
caactccatcagagtggaaggaggcacctgggctgtgtatgaaaggcccaattttgctgg
gtacatgtacatcctaccccggggcgagtatcctgagtaccagcactggatgggcctcaa
```

- Typical file extensions (.fasta, .fa, .fna, .fnn, .faa, ...)

Fastq format

- We could say “it is a fasta with **qualities**”:
  - 1. Header (like the fasta but starting with “@”)
  - 2. Sequence (string of nt)
  - 3. “+” and sequence ID (optional)
  - 4. Encoded quality of the sequence

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

Quality codification

- Phred quality score
  - Error probability
  - ASCII encoded

- Phred +33
  - Sanger [0,40]
  - Illumina 1.8 [0,41]
  - Illumina 1.9 [0,41]

- Phred +64
  - Illumina 1.3 [0,40]
  - Illumina 1.5 [3,40]

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

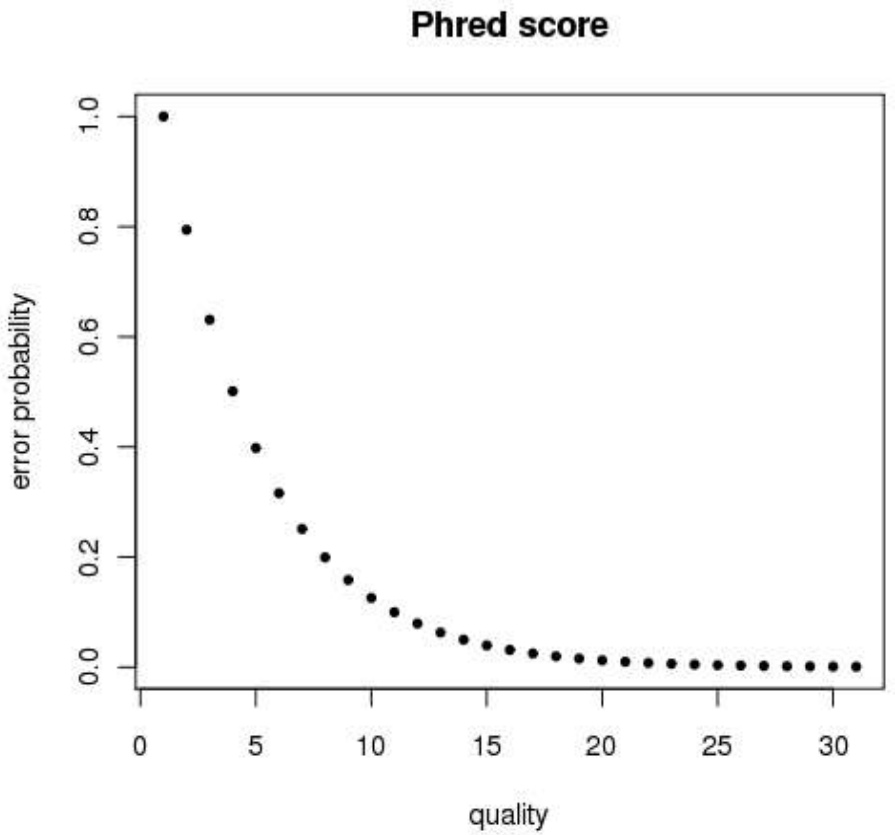
Francisco García
fgarcia@cipf.es

Quality codification

- Phred scores

$$Q = -10 \log_{10} P \quad \Longleftrightarrow \quad P = 10^{\frac{-Q}{10}}$$

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 10000	99.99%
50	1 in 100000	99.999%



Francisco García
fgarcia@cipf.es

Sequence quality evaluation

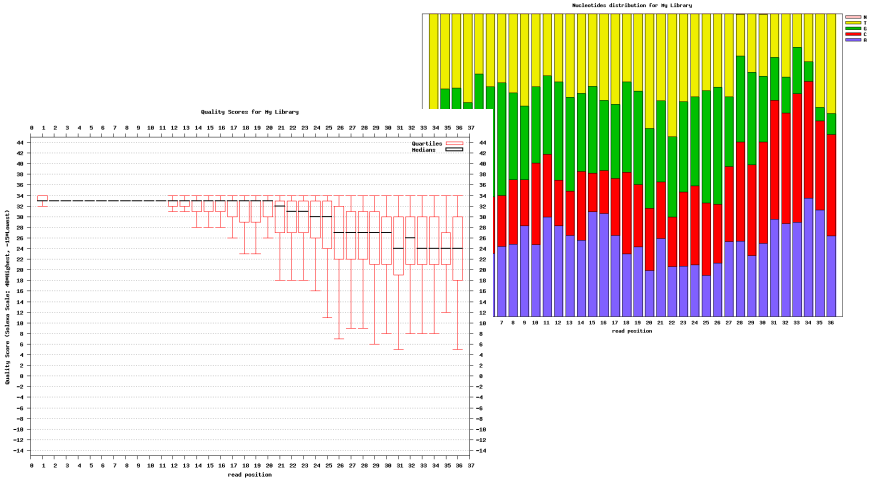
- If we evaluate our sequence in depth ...

  ... we will know how reliable our results are

- **Problem:**
  - **Huge files** → Need of a tool to do it

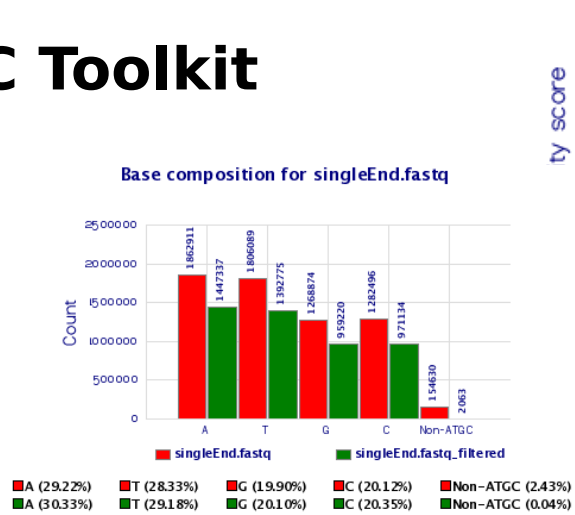
Sequence quality evaluation

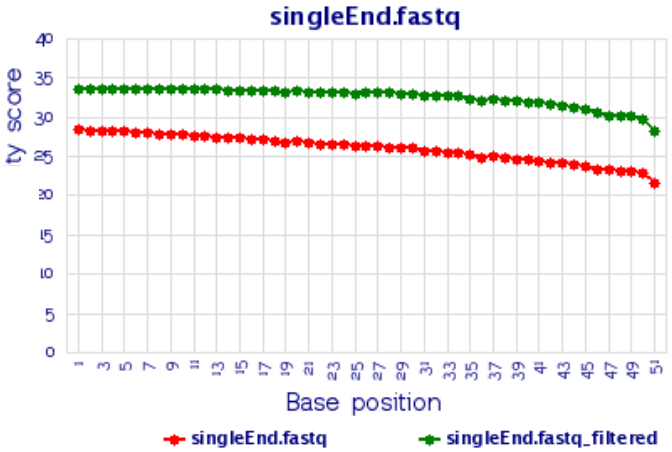
- Quality control tools:
  - **Fastx-toolkit**



http://hannonlab.cshl.edu/fastx_toolkit/download.html

  - **NGS QC Toolkit**



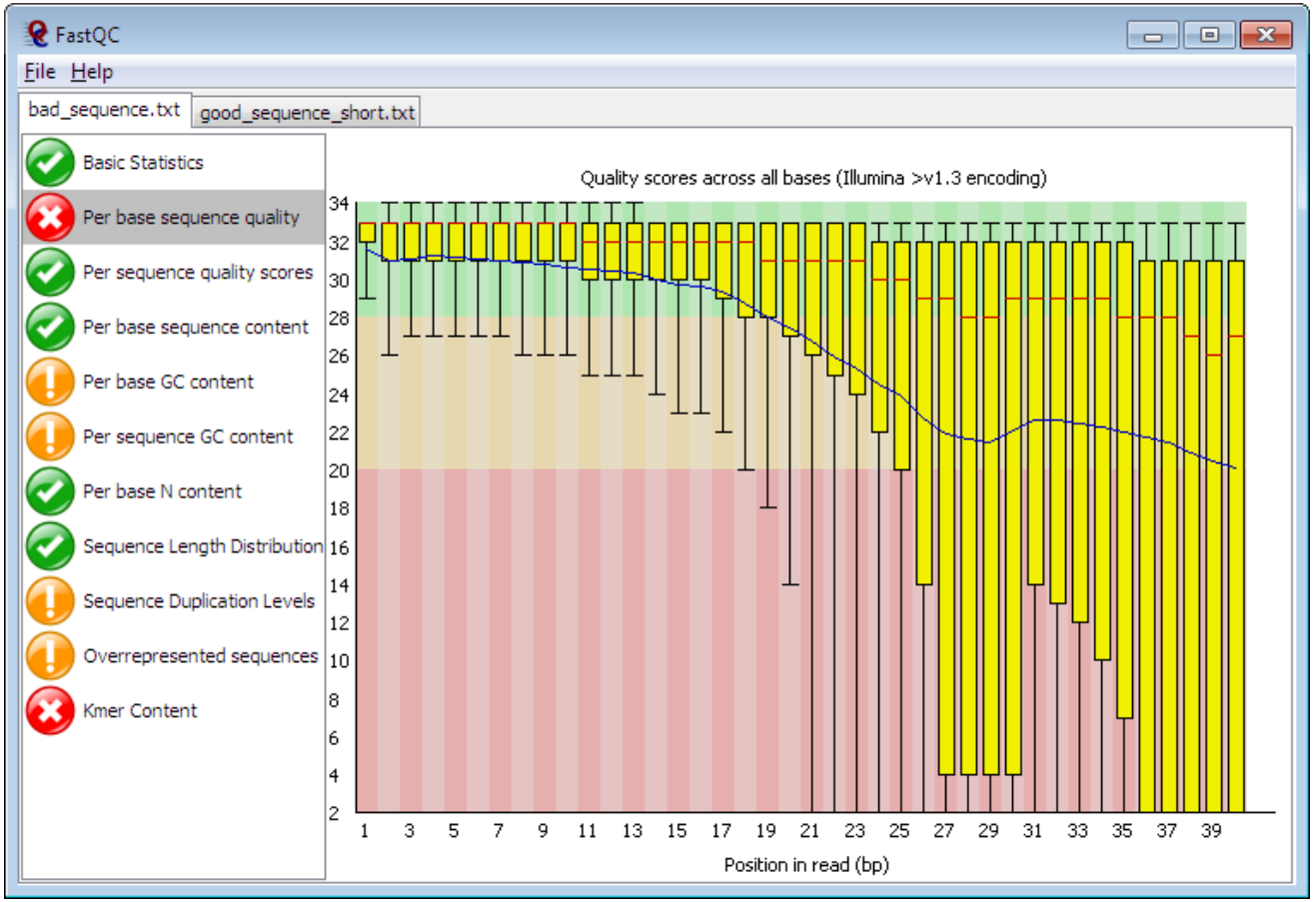


http://www.nipgr.res.in/ngsqctoolkit.html

Francisco García
fgarcia@cipf.es

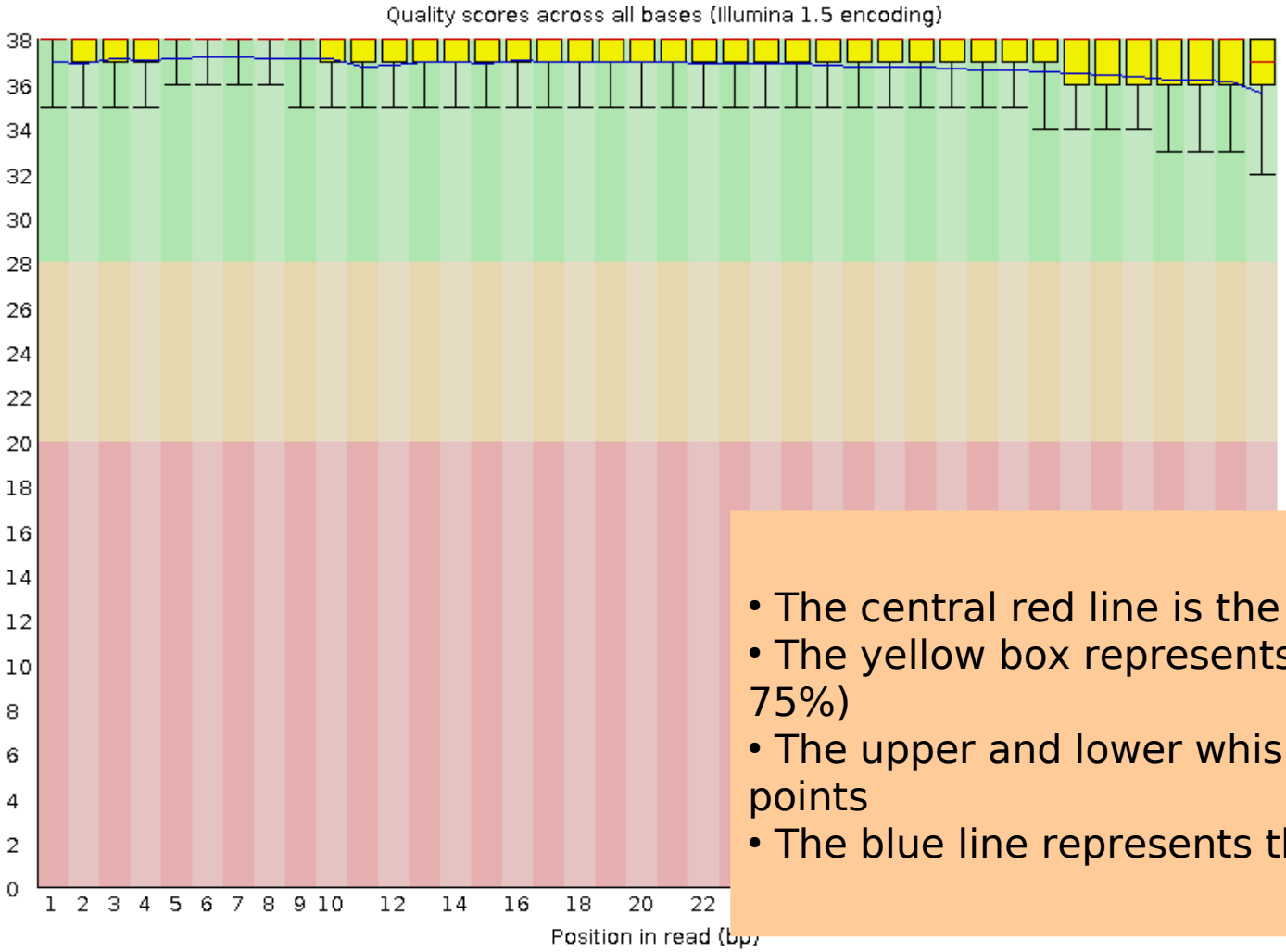
Sequence quality evaluation

- Other quality control tool: **FastQC**



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

Sequence quality per base position



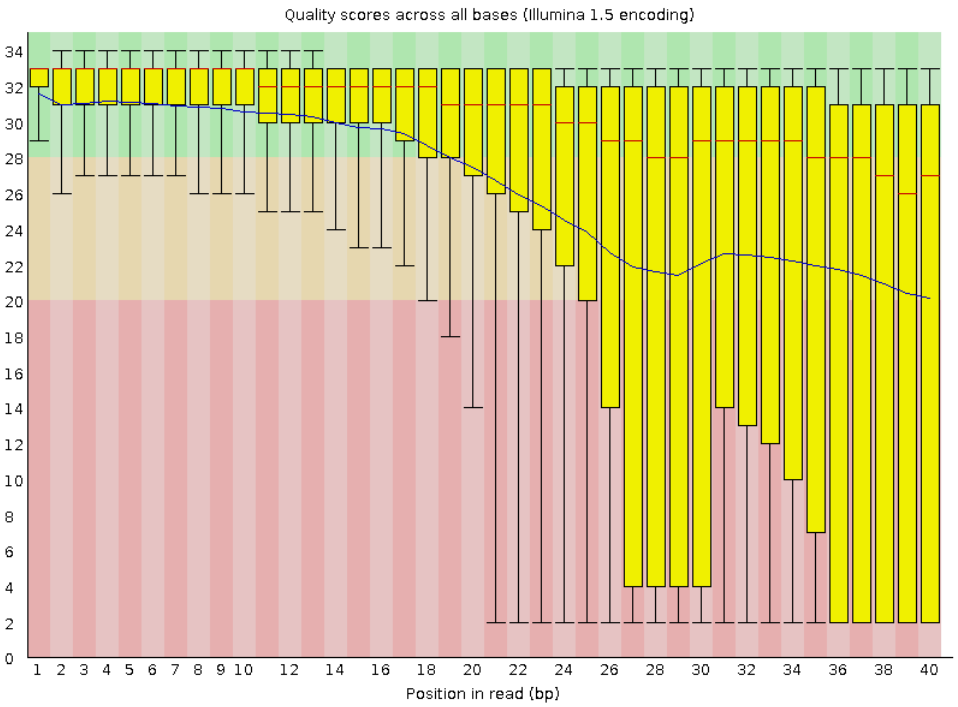
- Good data
  - Consistent
  - High quality along the read

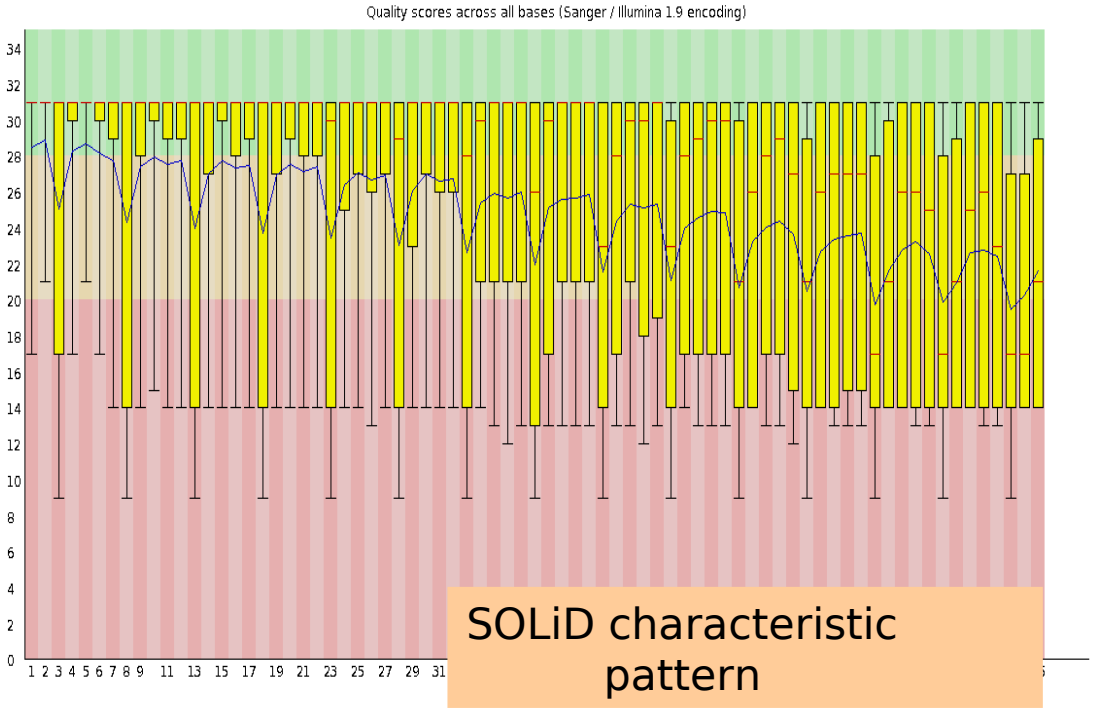
- The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality

Francisco García
fgarcia@cipf.es

Sequence quality per base position

- Bad data
  - High variance
  - Quality decrease with length

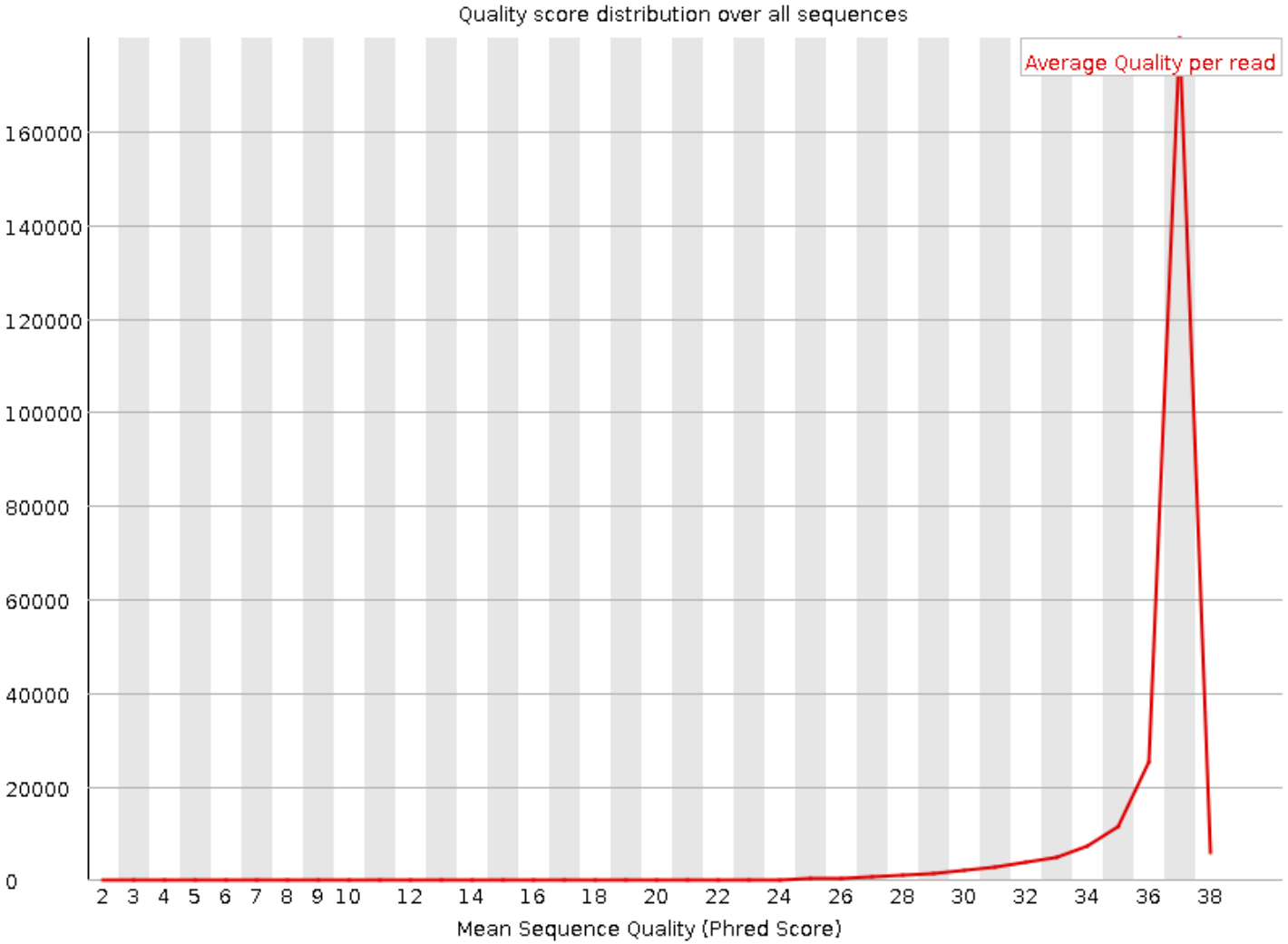




SOLiD characteristic pattern

Francisco García
fgarcia@cipf.es

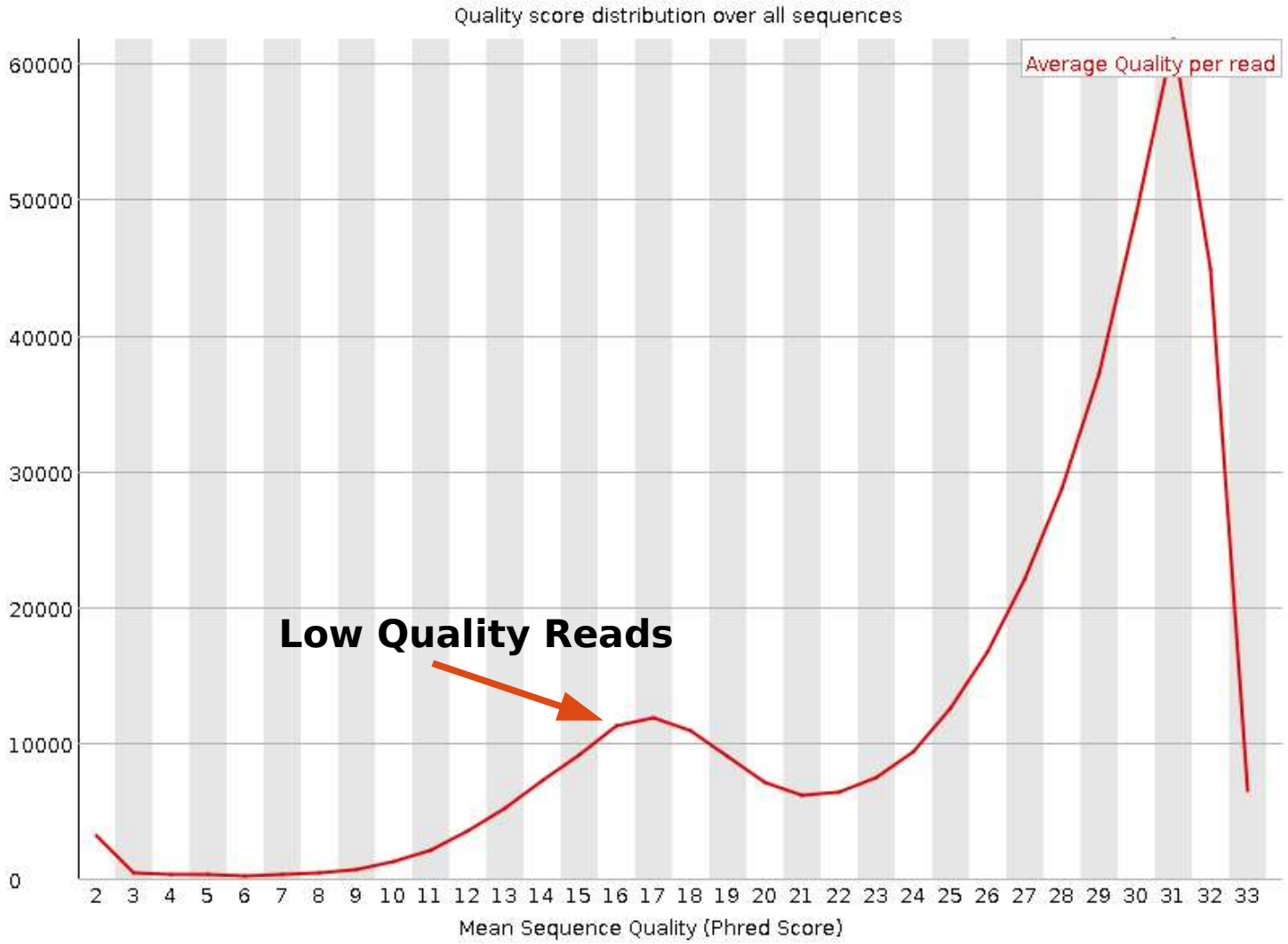
Per sequence quality distribution



- Good data
  - Most are high-quality sequences

Francisco García
fgarcia@cipf.es

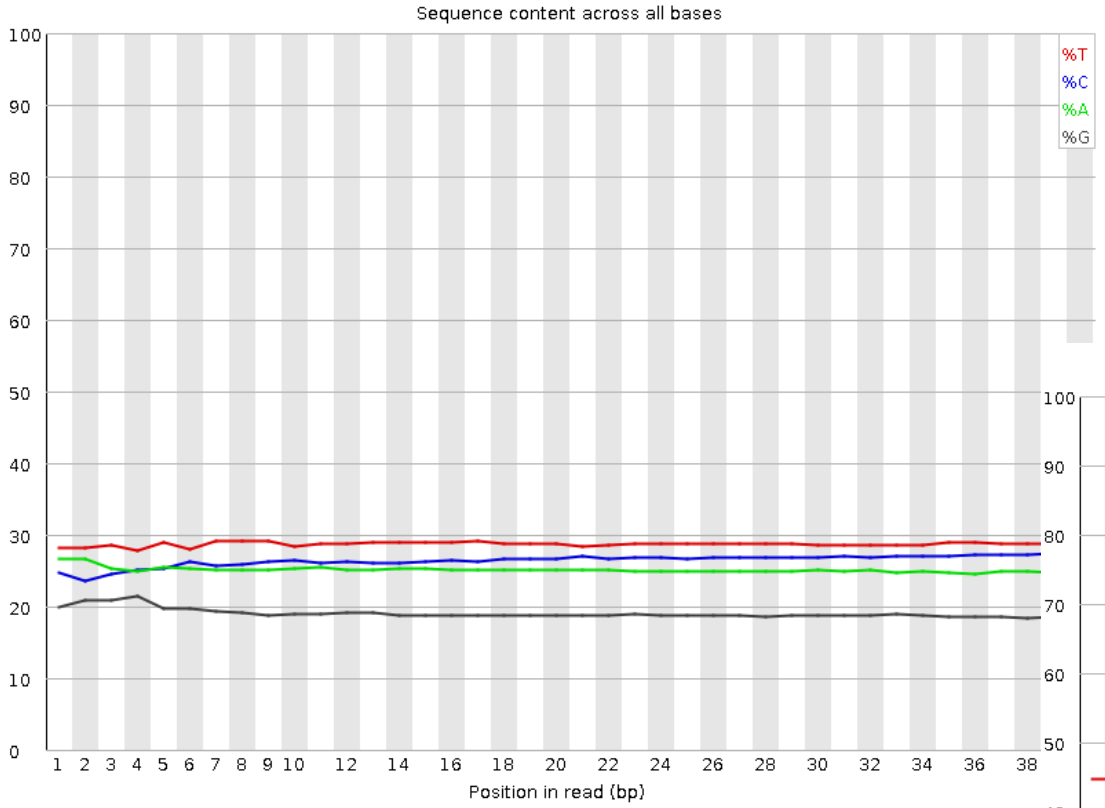
Per sequence quality distribution



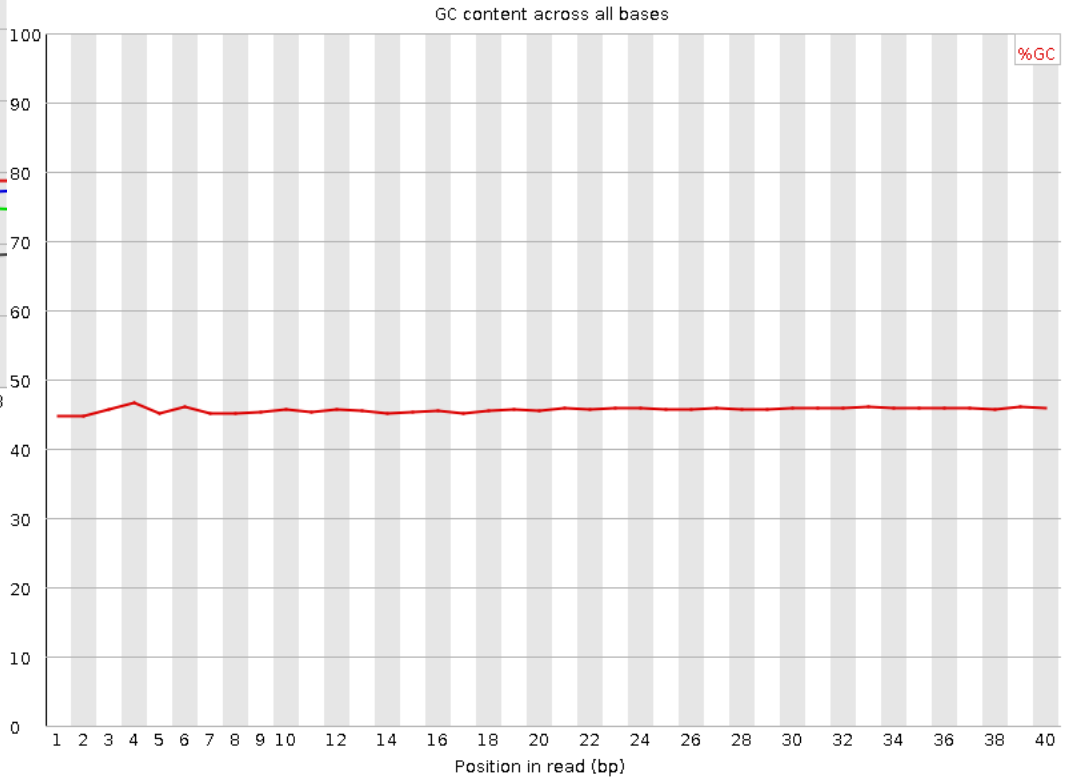
- Bad data
  - Non-uniform distribution

Francisco García
fgarcia@cipf.es

Per base sequence content

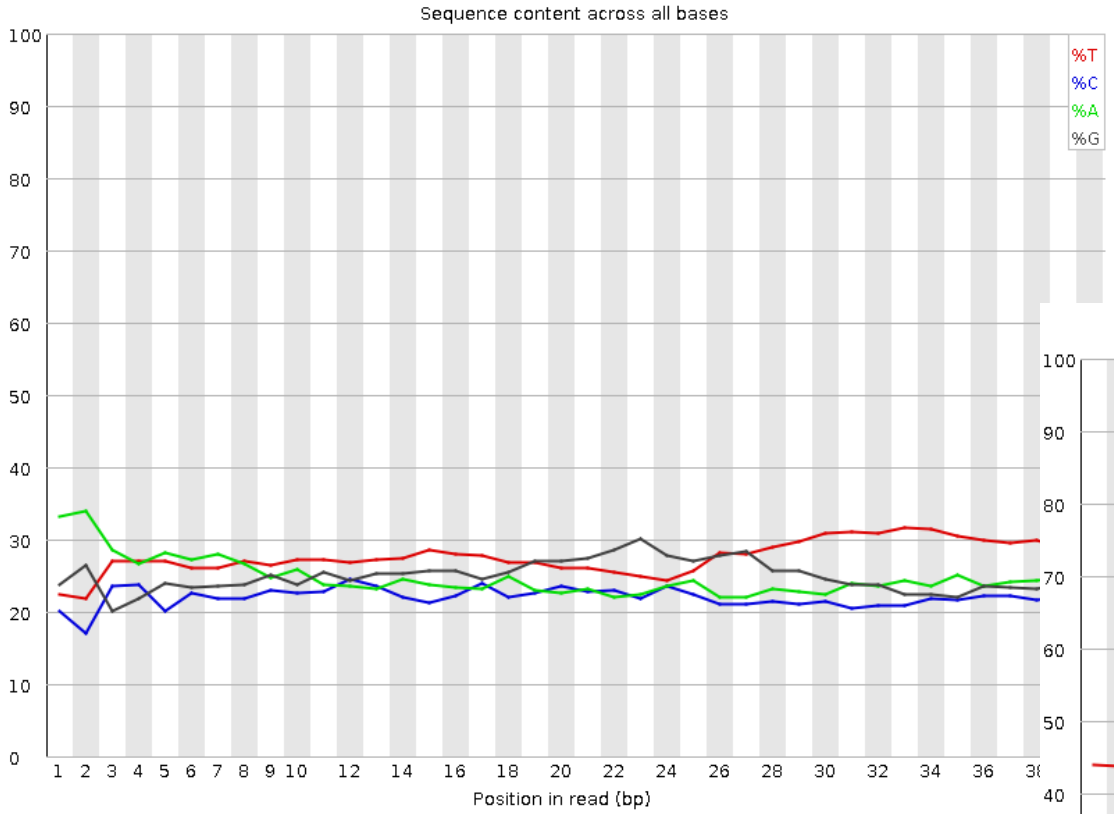


- Good data
  - Smooth over length
  - Organism dependent (GC)

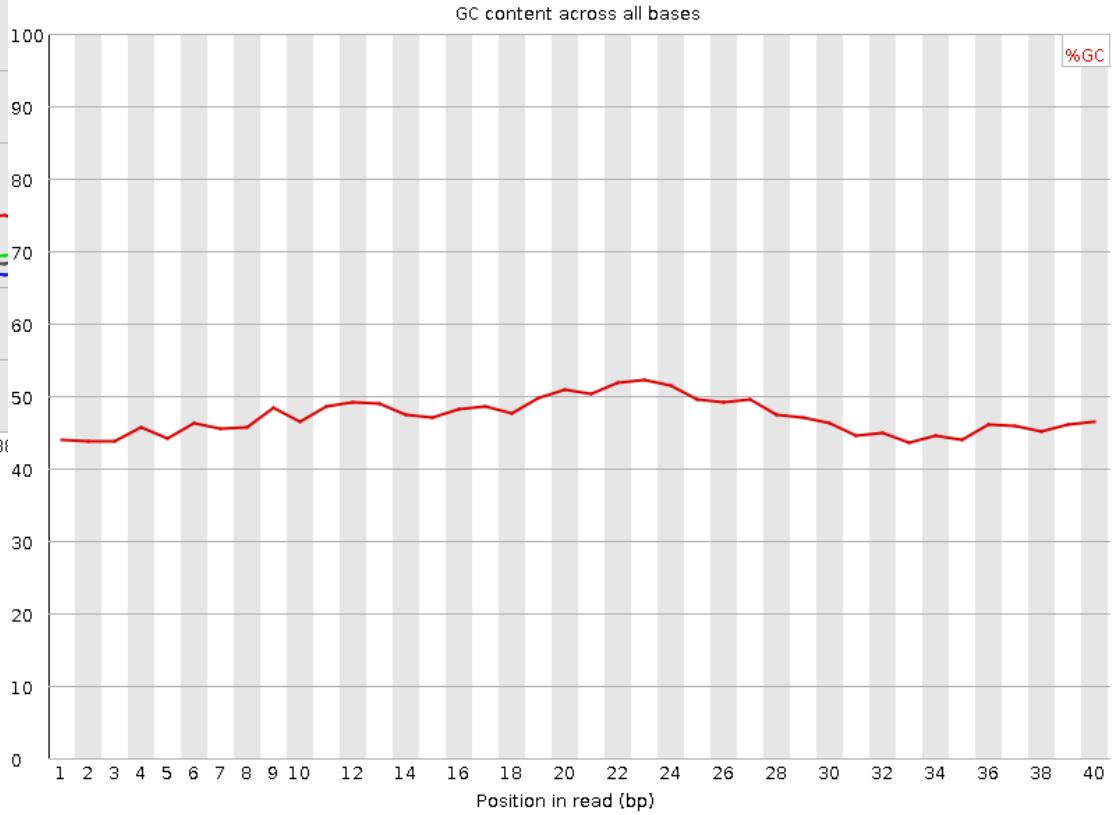


Francisco García
fgarcia@cipf.es

Per base sequence content



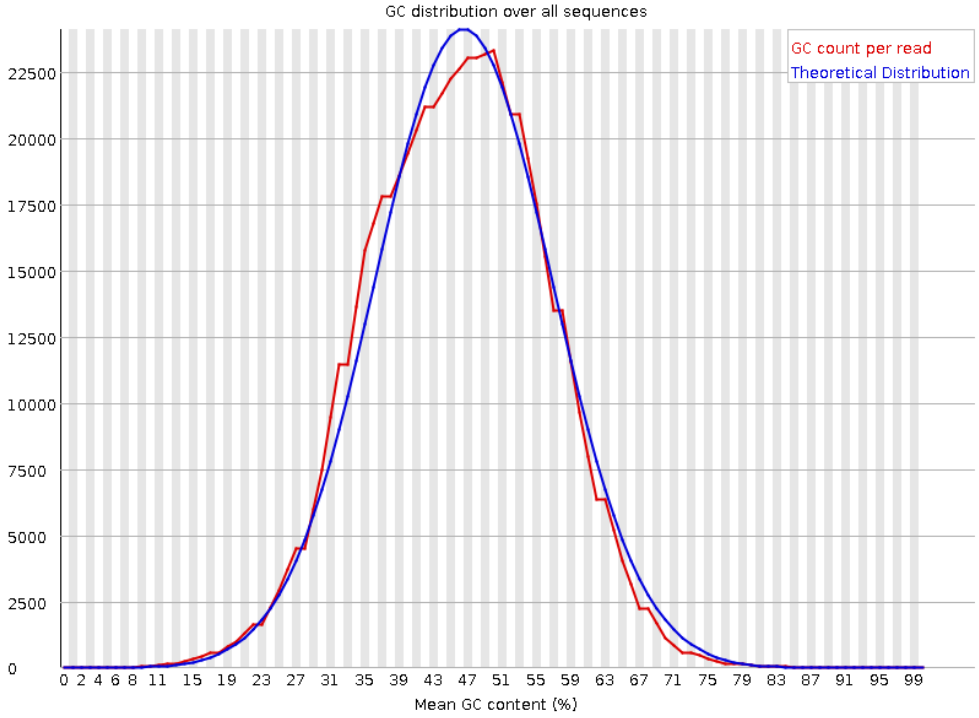
- Bad data
  - Sequence position bias



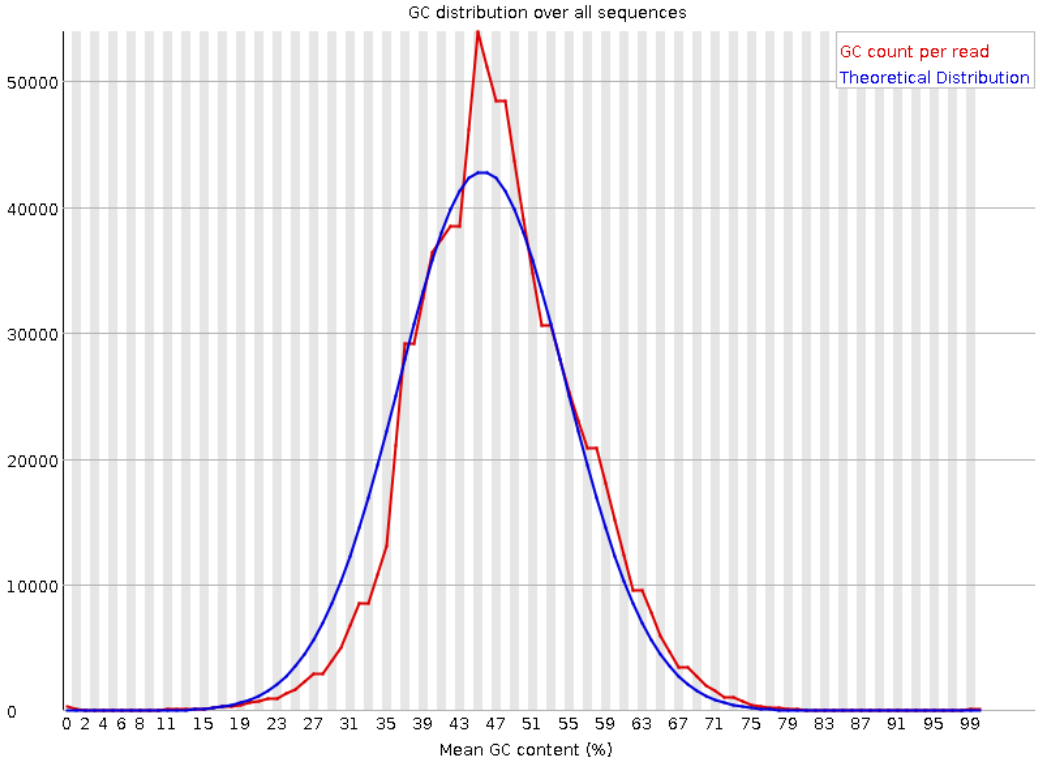
Francisco García
fgarcia@cipf.es

Per sequence GC content

- Good data
  - Fits with expected
  - Organism dependent



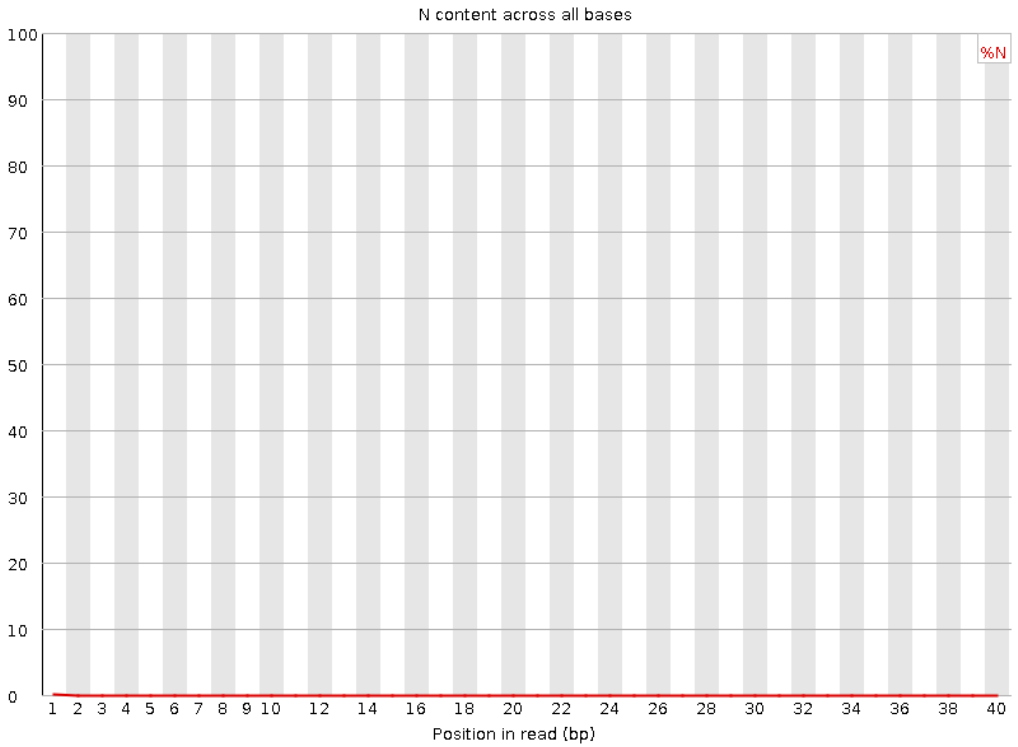
- Bad data
  - Does not fit with expected
  - Library contamination?



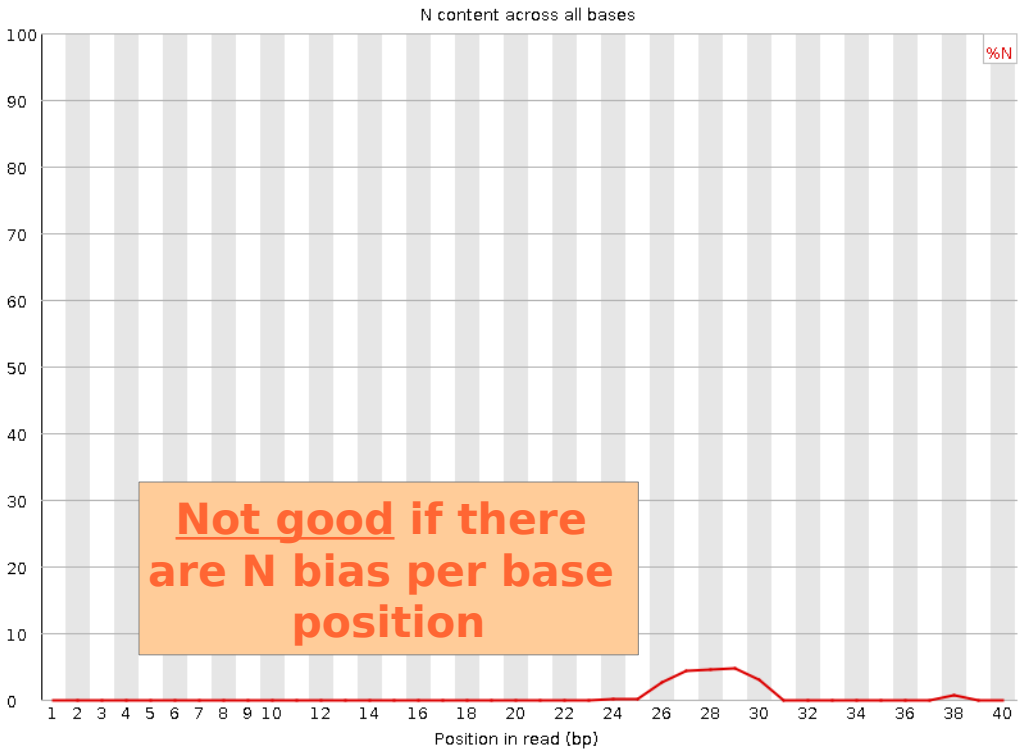
Francisco García
fgarcia@cipf.es

Per base N content

- Good data

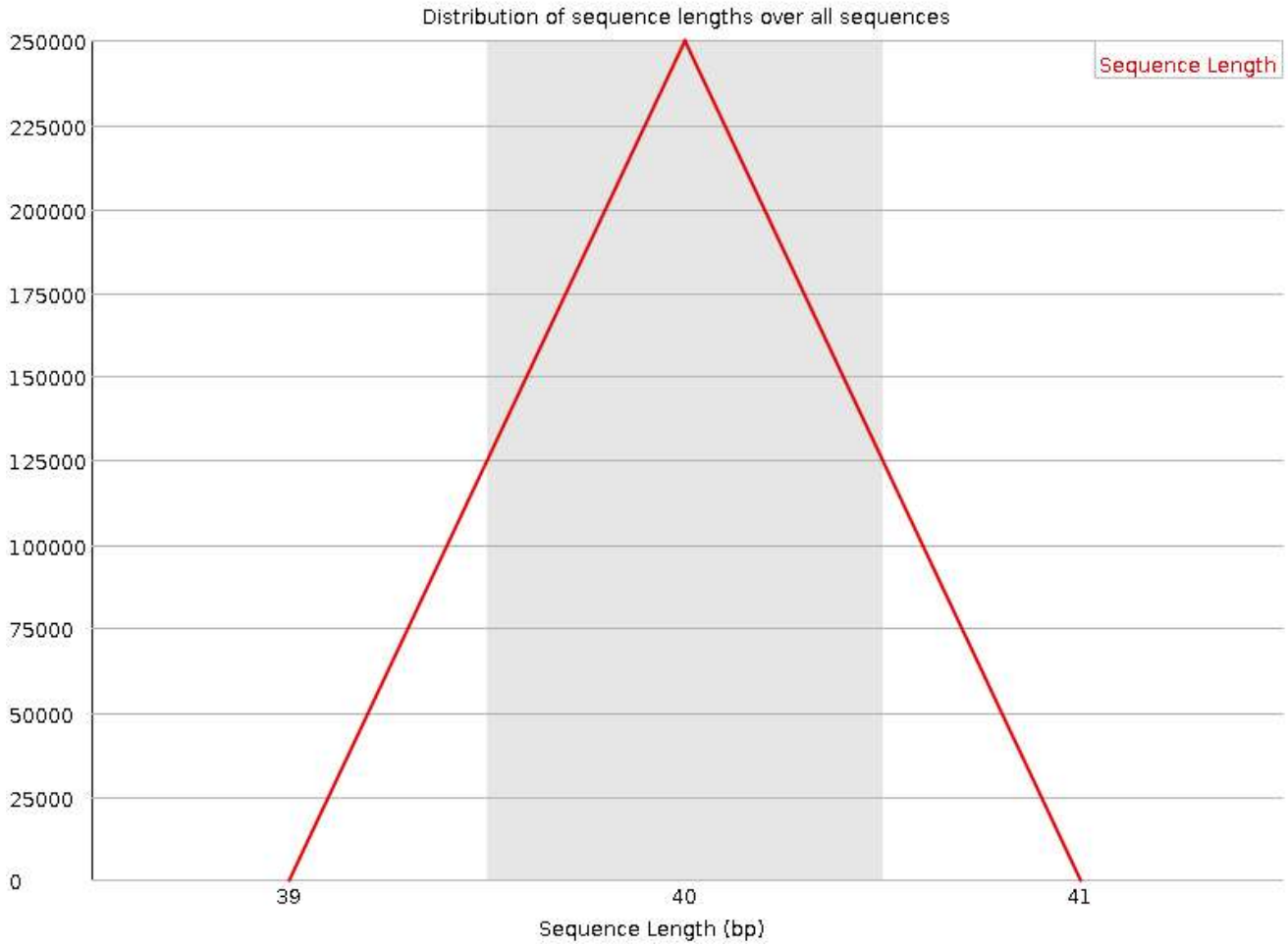


- Bad data



Sequence length distribution

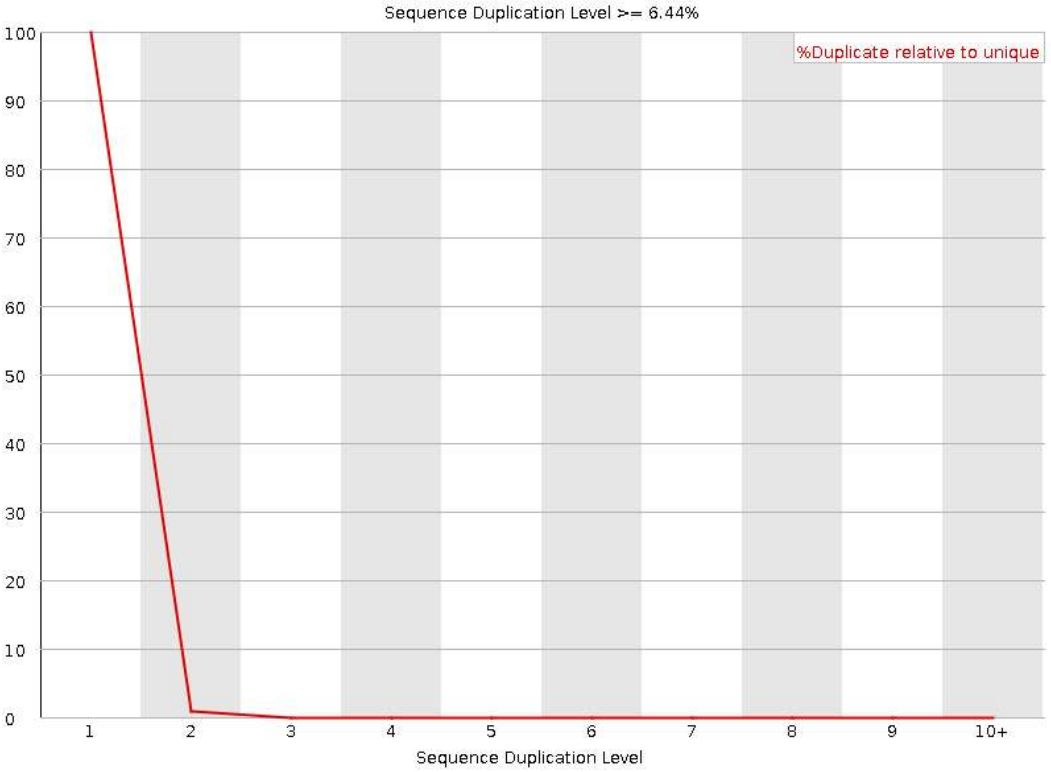
- Just descriptive:
  - Some sequencers output sequences of different length (e.g. 454)

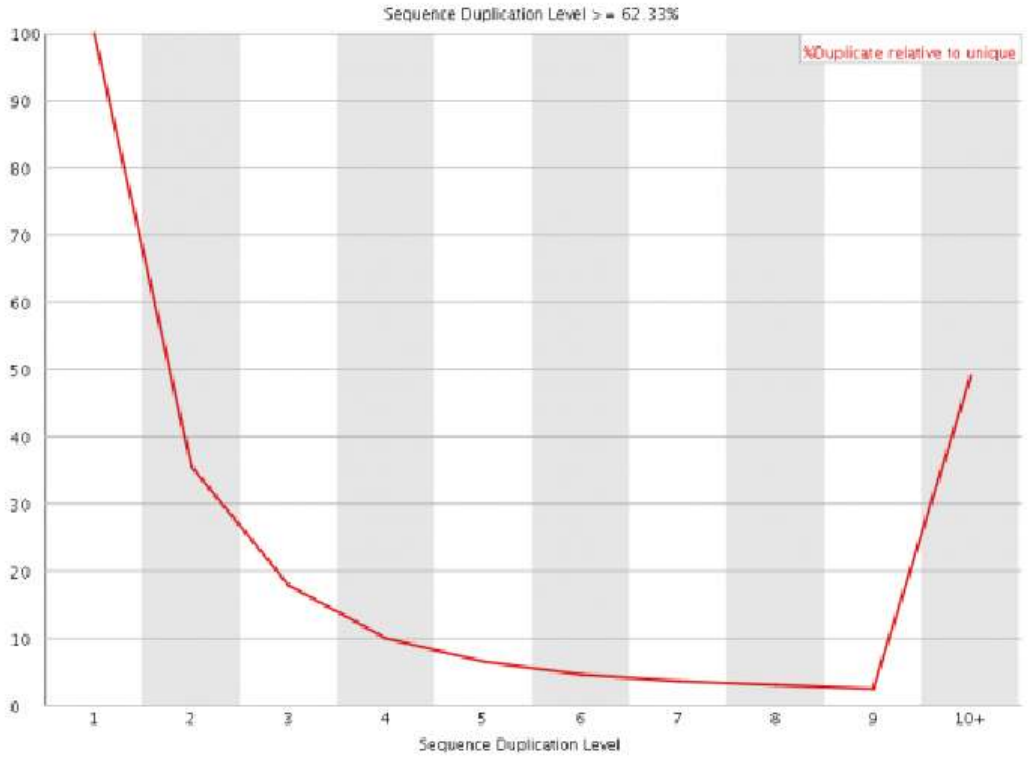


Francisco García
fgarcia@cipf.es

Sequence duplication levels

- In **transcriptomics**, you expect higher number of duplicated sequences.
- In **genomics** you should be worried if this happens → PCR artifact?





Francisco García
fgarcia@cipf.es

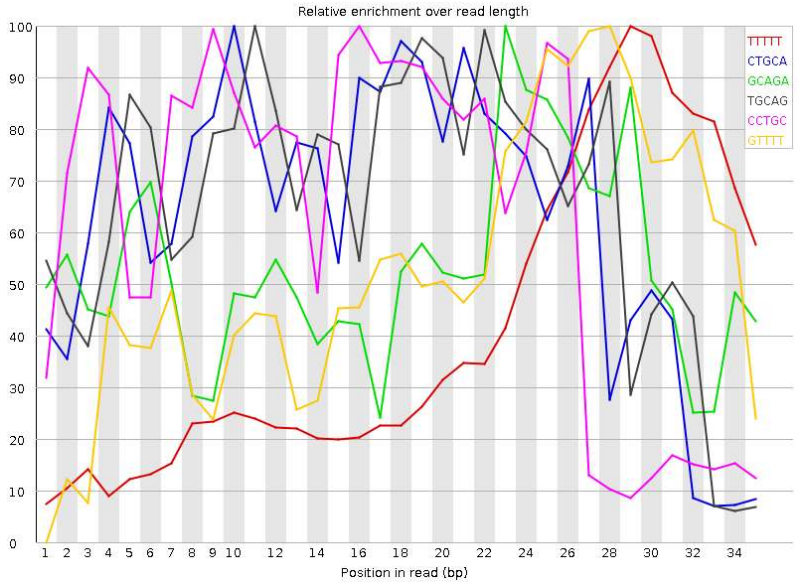
Overrepresented sequences & Kmer content

- Question:
  - If we obtain the exact same sequences too many times
    → **Do we have a problem?**

- Answer:
  - **Sometimes !**

- Examples:
  - PCR primers, adapters ...

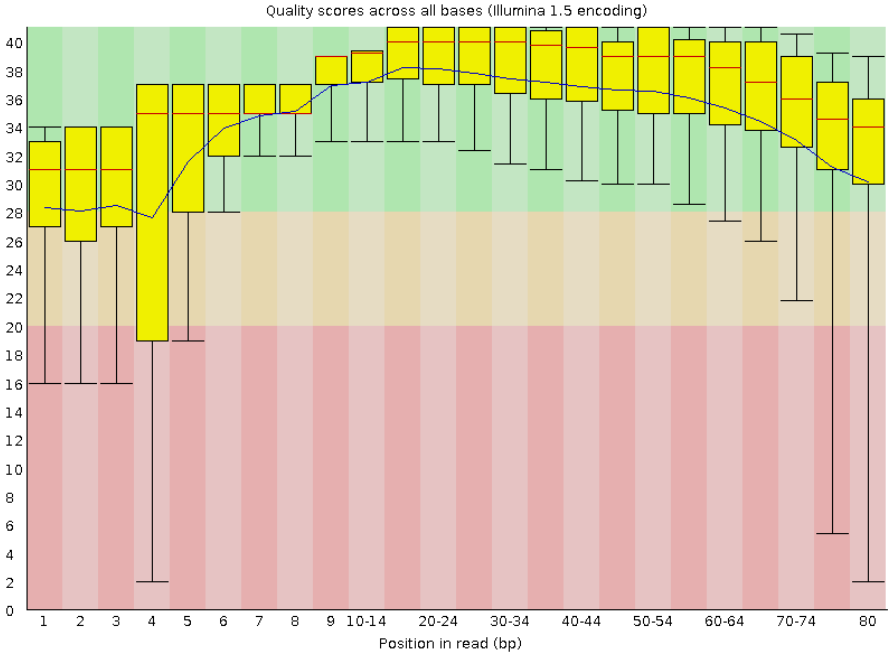
Sequence	Count	Percentage
AGAGTTTTATCGCTTCCATGAC GCAGAAGTTAACACTTTC	2065	0.5224039181558763
GATTGGCGTATCCAACCTGCA GAGTTTTATCGCTTCCATG	2047	0.5178502762542754

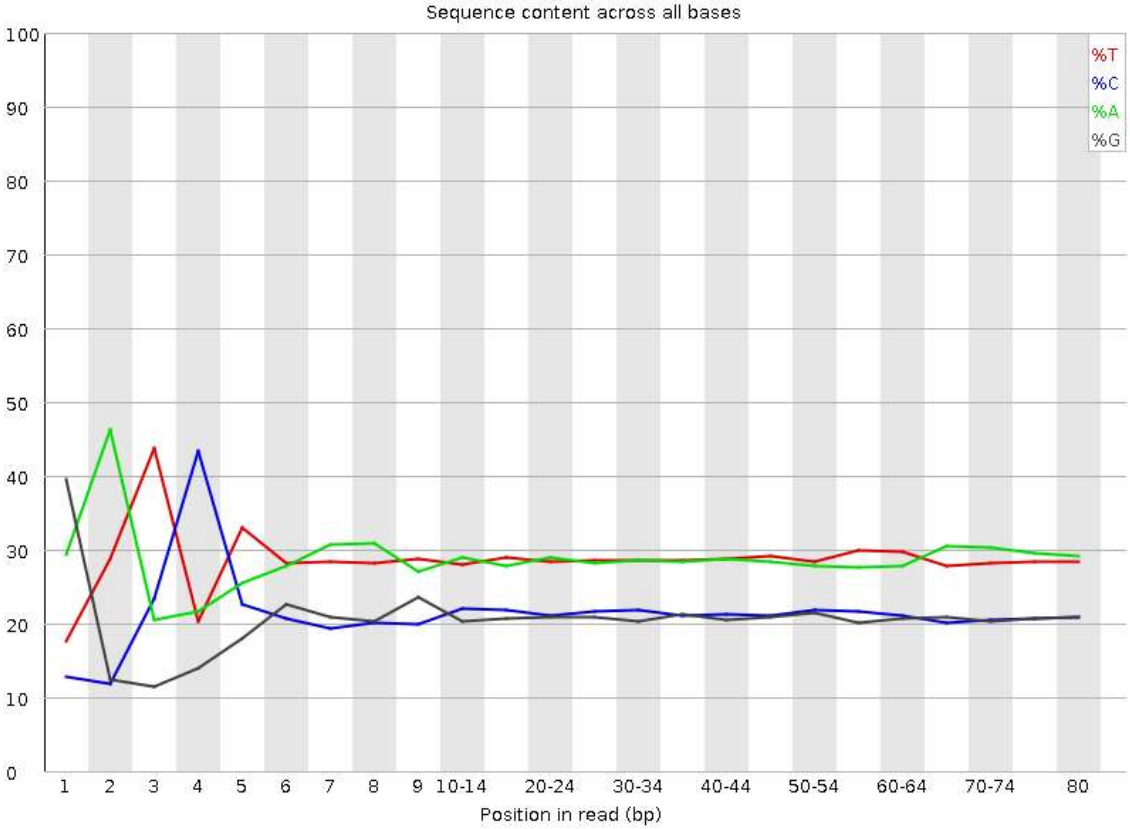


Francisco García
fgarcia@cipf.es

Typical artifacts

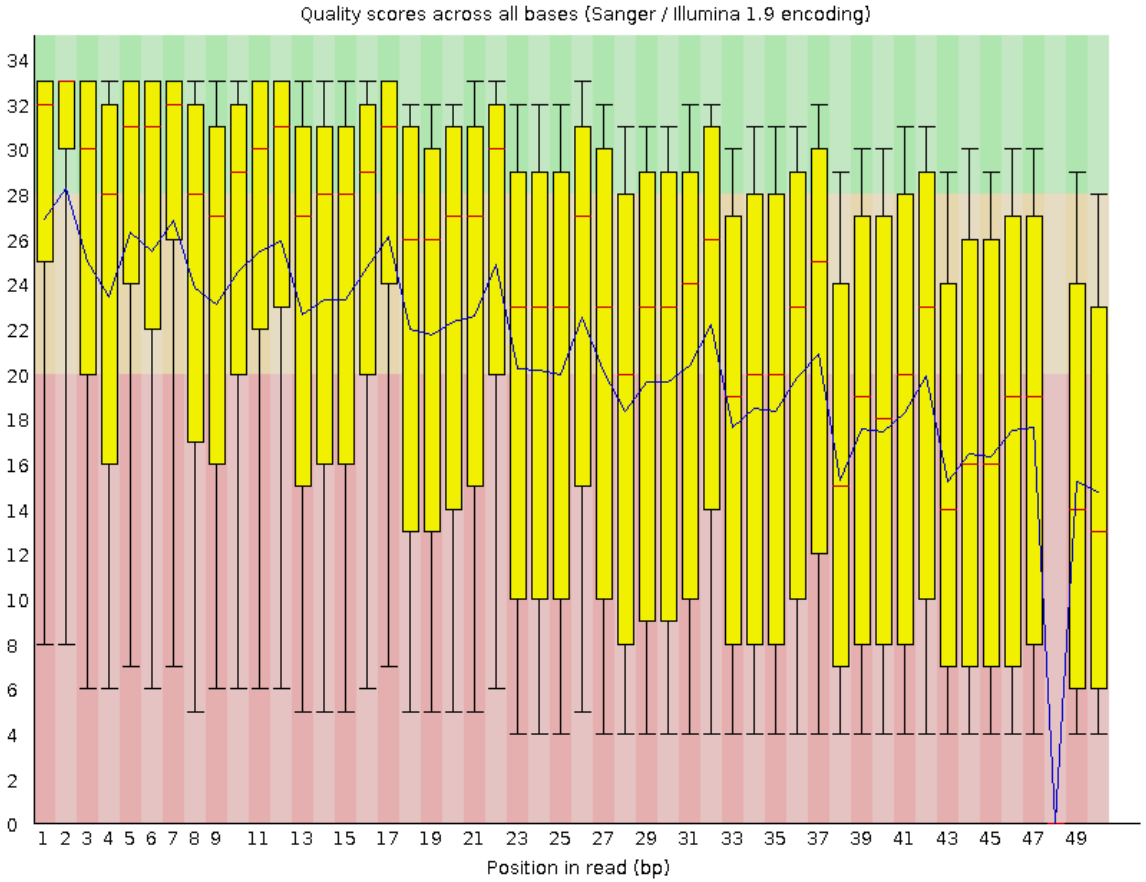
- Sequence adapters



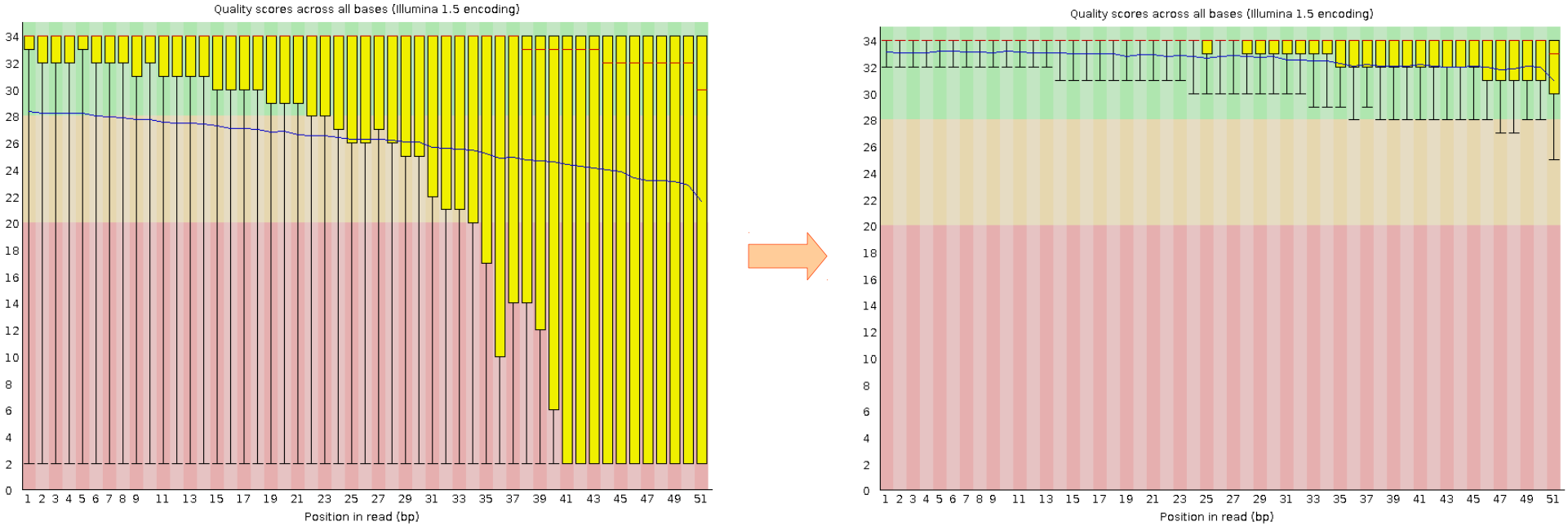


Typical artifacts

- Platform dependent



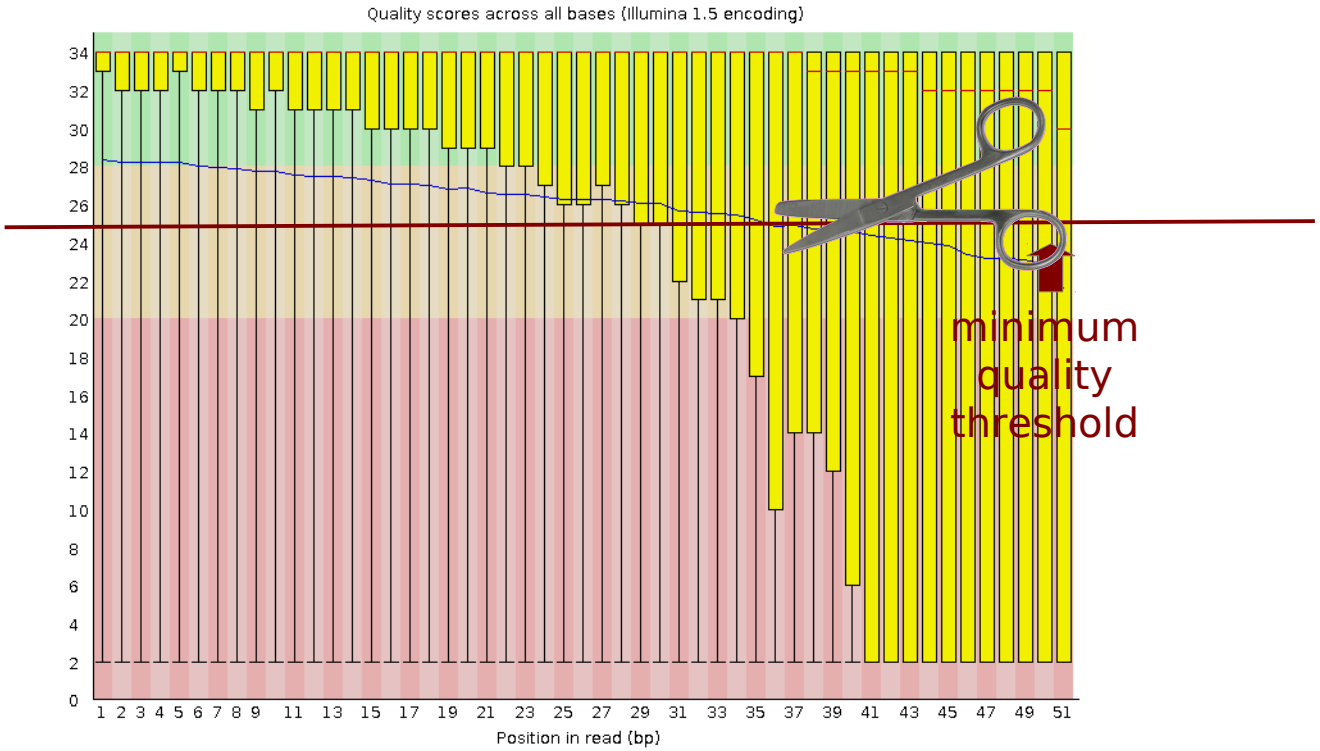
Filtering & trimming



- Removing bad quality data will improve our confidence on downstream analysis

Filtering & trimming

- Sequence filtering
  - Mean quality
  - Read length
  - Read length after trimming
  - Percentage of bases above Q
  - Adapter trimming
  - Adapter reads



Francisco García
fgarcia@cipf.es

Filtering & trimming

- Sequence filtering tools
  - Fastx-toolkit
  - Galaxy (https://main.g2.bx.psu.edu/)
  - SeqTK (https://github.com/lh3/seqtk)
  - Cutadapt (http://code.google.com/p/cutadapt/)
  - And more....

Practical: FastQC & Fastx-toolkit

- Use **FastQC** to see your starting state.
- Use **Fastx-toolkit** to optimize different datasets and then visualize the result with FastQC to prove your success!

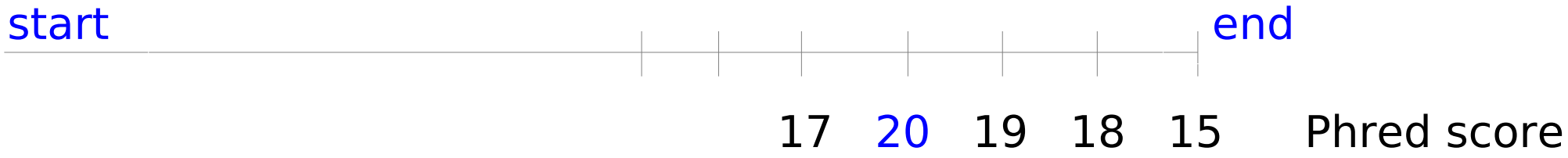
Hints: Try trimming, clipping and quality filtering.

Go to the tutorial and try the exercises...

Trimming

Trimming the sequence with a minimum quality threshold of 20:

A



B