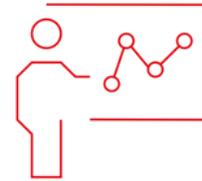




Swiss Institute of  
Bioinformatics

# Introduction to RNA-Seq: Quality Control

Wandrille Duchemin



# "Raw data": FASTQ format

```
@HWI-ST143:498:D0T4GACXX:7:1101:1474:2111 1:N:0:TGACCA  
TCTCNAGATAAAATCAAACCAACAGAGAGTCTAGAATAAAAGTGAATAG  
+  
@@BF#2ADHHHHJJJJJJJJJJJJGJJHIIIGIHIIIIJJHHIJJJ  
@HWI-ST143:498:D0T4GACXX:7:1101:1424:2149 1:N:0:TGACCA  
CAGATGTTTTGGTAAGAAGAATGCCATTAGCAATTGGTGGAGGAGCAGGA  
+  
BBCFFFEFHHHHHJJJJJJHIIJJJJJJJJJJJJJJJJHJJJJJJGIIIG
```

Paired data: two separate files for forward and reverse with same ordering

# "Raw data": FASTQ format

Read ID

line starts with @

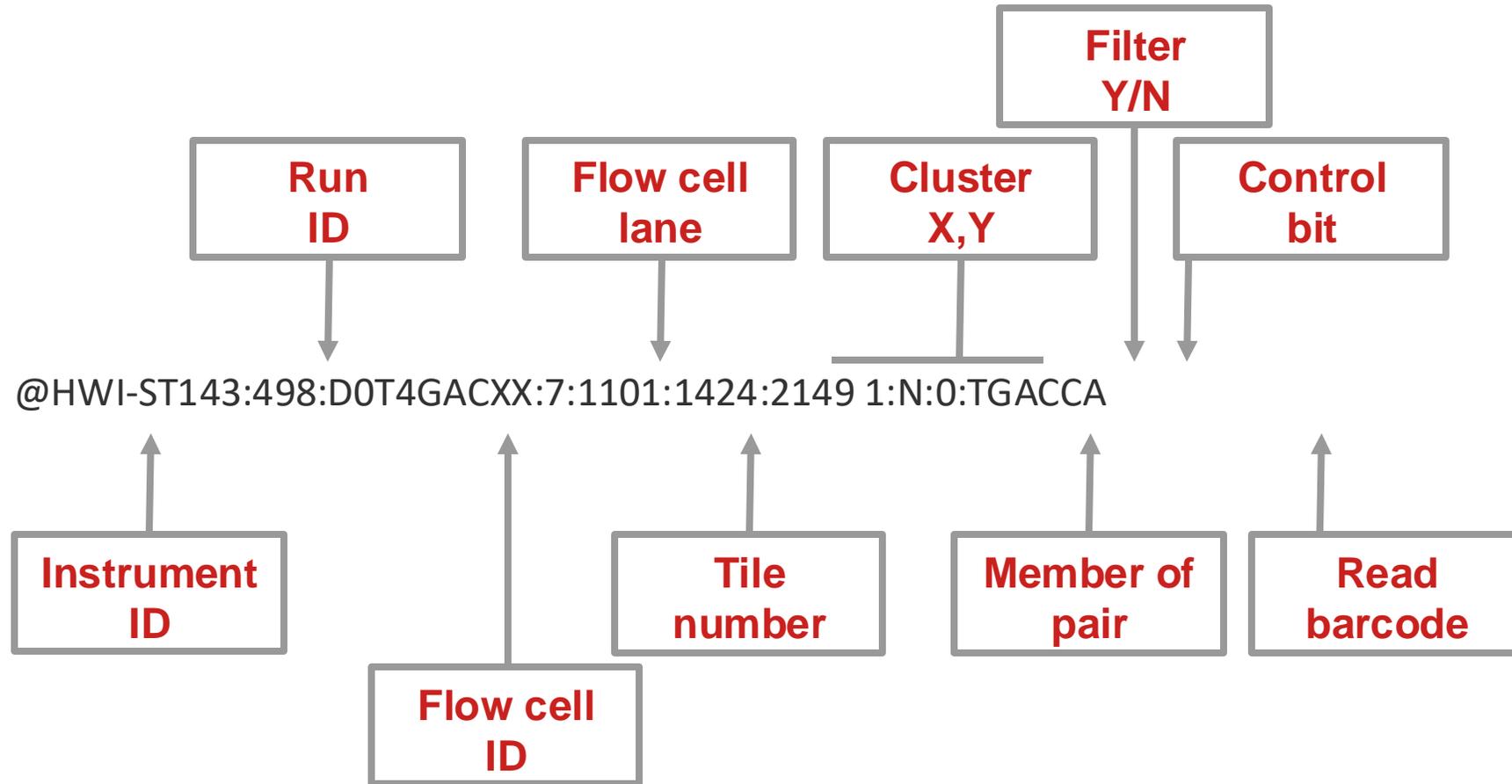
```
@HWI-ST143:498:D0T4GACXX:7:1101:1474:2111 1:N:0:TGACCA  
TCTCNAGATAAAATCAAACCAACAGAGAGTCTAGAATAAAAGTGAATAG  
+  
@@BF#2ADHHHHHJJJJJJJJJJJJGJIJHIIGIHIIIIJJHIIJJ  
@HWI-ST143:498:D0T4GACXX:7:1101:1424:2149 1:N:0:TGACCA  
CAGATGTTTTGGTAAGAAGAATGCCATTAGCAATTGGTGGAGGAGCAGGA  
+  
BBCFFFEFHHHHHHJJJJJJHIIJJJJJJJJJJJJJJHIIJJJJJJIGIIG
```

sequence

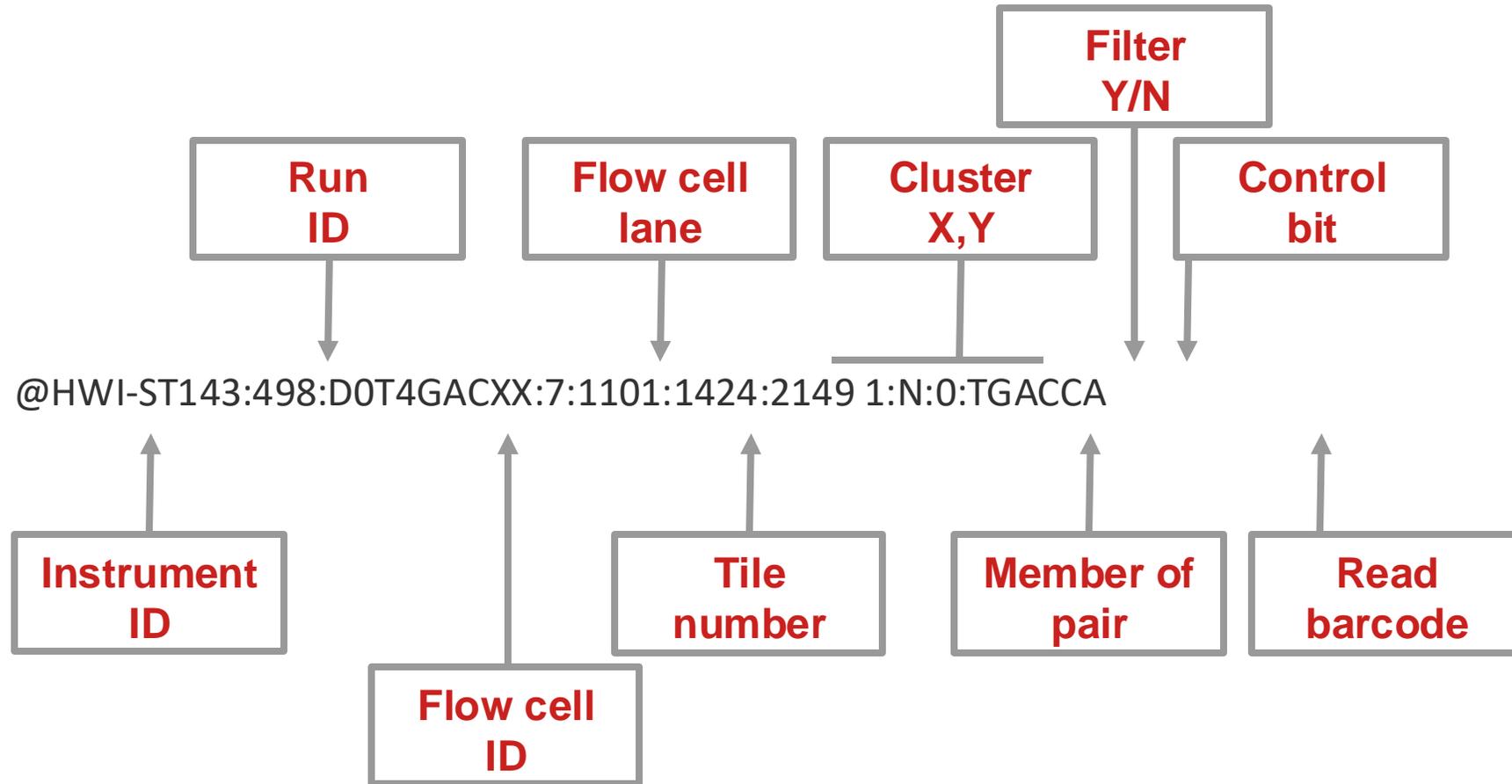
quality line

Paired data: two separate files for forward and reverse with same ordering

# "Raw data": FASTQ format - header



# "Raw data": FASTQ format - header



Depends on the sequencing technology.  
It was changed several times by illumina and others

# "Raw data": FASTQ format - quality line

TCTCNAGATAAAATCAAACCAACAGAGAGTCTAGAATAAAAGTGAATAG

@@BF#2ADHHHHJJJJJJJJJJGJJHHIIGIHIIIIJJHHIJJ

Each nucleotide is associated to a quality line

# "Raw data": FASTQ format - PHRED score

Probability that a base is incorrect (p)

- Quality (Q) =  $-10 \log_{10}(p)$

ASCII encoded

P-value	PHRED	Probability of incorrect base call	Base call accuracy
$10^{-1}$	10	1/10	90%
$10^{-2}$	20	1/100	99%
$10^{-3}$	30	1/1000	99.9%
$10^{-4}$	40	1/10'000	99.99%

# "Raw data": FASTQ format - quality line

TCTCNAGATAAAATCAAACCAACAGAGAGTCTAGAATAAAAGTGAATAG

@@BF#2ADHHHHJJJJJJJJJJJJGJIJHIIGIHIIIIJJHIHIJJ

Illumina v1.8 and later (ASCII\_BASE=33)

Q	ASCII	P									
1	"	0.79433	12	-	0.06310	23	8	0.00501	34	C	0.00040
2	#	0.63096	13	.	0.05012	24	9	0.00398	35	D	0.00032
3	\$	0.50119	14	/	0.03981	25	:	0.00316	36	E	0.00025
4	%	0.39811	15	0	0.03162	26	;	0.00251	37	F	0.00020
5	&	0.31623	16	1	0.02512	27	<	0.00200	38	G	0.00016
6	'	0.25119	17	2	0.01995	28	=	0.00158	39	H	0.00013
7	(	0.19953	18	3	0.01585	29	>	0.00126	40	I	0.00010
8	)	0.15849	19	4	0.01259	30	?	0.00100	41	J	0.00008
9	*	0.12589	20	5	0.01000	31	@	0.00079			
10	+	0.10000	21	6	0.00794	32	A	0.00063			
11	,	0.07943	22	7	0.00631	33	B	0.00050			

# "Raw data": FASTQ format - quality line

TCTCNAGATAAAATCAAACCAACAGAGAGTCTAGAATAAAAGTGAATAG

@@BF#2ADHHHHJJJJJJJJJJJJGJIJHIIGIHIIIIJJHIHIJJ

Illumina v1.8 and later (ASCII\_BASE=33)

Q	ASCII	P									
1	"	0.79433	12	-	0.06310	23	8	0.00501	34	C	0.00040
2	#	0.63096	13	.	0.05012	24	9	0.00398	35	D	0.00032
3	\$	0.50119	14	/	0.03981	25	:	0.00316	36	E	0.00025
4	%	0.39811	15	0	0.03162	26	;	0.00251	37	F	0.00020
5	&	0.31623	16	1	0.02512	27	<	0.00200	38	G	0.00016
6	'	0.25119	17	2	0.01995	28	=	0.00158	39	H	0.00013
7	(	0.19953	18	3	0.01585	29	>	0.00126	40	I	0.00010
8	)	0.15849	19	4	0.01259	30	?	0.00100	41	J	0.00008
9	*	0.12589	20	5	0.01000	31	@	0.00079			
10	+	0.10000	21	6	0.00794	32	A	0.00063			
11	,	0.07943	22	7	0.00631	33	B	0.00050			

# "Raw data": FASTQ format - PHRED +33/+64

Sanger, Illumina v1.3 to 1.7 (ASCII\_BASE=64)

Q	ASCII	P									
1	A	0.79433	12	L	0.06310	23	W	0.00501	34	b	0.00040
2	B	0.63096	13	M	0.05012	24	X	0.00398	35	c	0.00032
3	C	0.50119	14	N	0.03981	25	Y	0.00316	36	d	0.00025
4	D	0.39811	15	O	0.03162	26	Z	0.00251	37	e	0.00020
5	E	0.31623	16	P	0.02512	27	[	0.00200	38	f	0.00016
6	F	0.25119	17	Q	0.01995	28	\	0.00158	39	g	0.00013
7	G	0.19953	18	R	0.01585	29	]	0.00126	40	h	0.00010
8	H	0.15849	19	S	0.01259	30	^	0.00100			
9	I	0.12589	20	T	0.01000	31	_	0.00079			
10	J	0.10000	21	U	0.00794	32	`	0.00063			
11	K	0.07943	22	V	0.00631	33	a	0.00050			

Illumina v1.8 and later (ASCII\_BASE=33)

Q	ASCII	P									
1	"	0.79433	12	-	0.06310	23	8	0.00501	34	C	0.00040
2	#	0.63096	13	.	0.05012	24	9	0.00398	35	D	0.00032
3	\$	0.50119	14	/	0.03981	25	:	0.00316	36	E	0.00025
4	%	0.39811	15	0	0.03162	26	;	0.00251	37	F	0.00020
5	&	0.31623	16	1	0.02512	27	<	0.00200	38	G	0.00016
6	'	0.25119	17	2	0.01995	28	=	0.00158	39	H	0.00013
7	(	0.19953	18	3	0.01585	29	>	0.00126	40	I	0.00010
8	)	0.15849	19	4	0.01259	30	?	0.00100	41	J	0.00008
9	*	0.12589	20	5	0.01000	31	@	0.00079			
10	+	0.10000	21	6	0.00794	32	A	0.00063			
11	,	0.07943	22	7	0.00631	33	B	0.00050			

# Quality Control of FASTQ files with fastqQC

Helps spot problems in the sequencer or in starting library material

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

fastqc generates an html report :

- Average quality per position
- GC% profile
- Adapter presence
- ...

Input formats: fastq (gzip), sam, bam

# Combining multiple reports: multiQC

- fastQC: 1 report for each fastq file
- MultiQC: combines individual reports in a single file

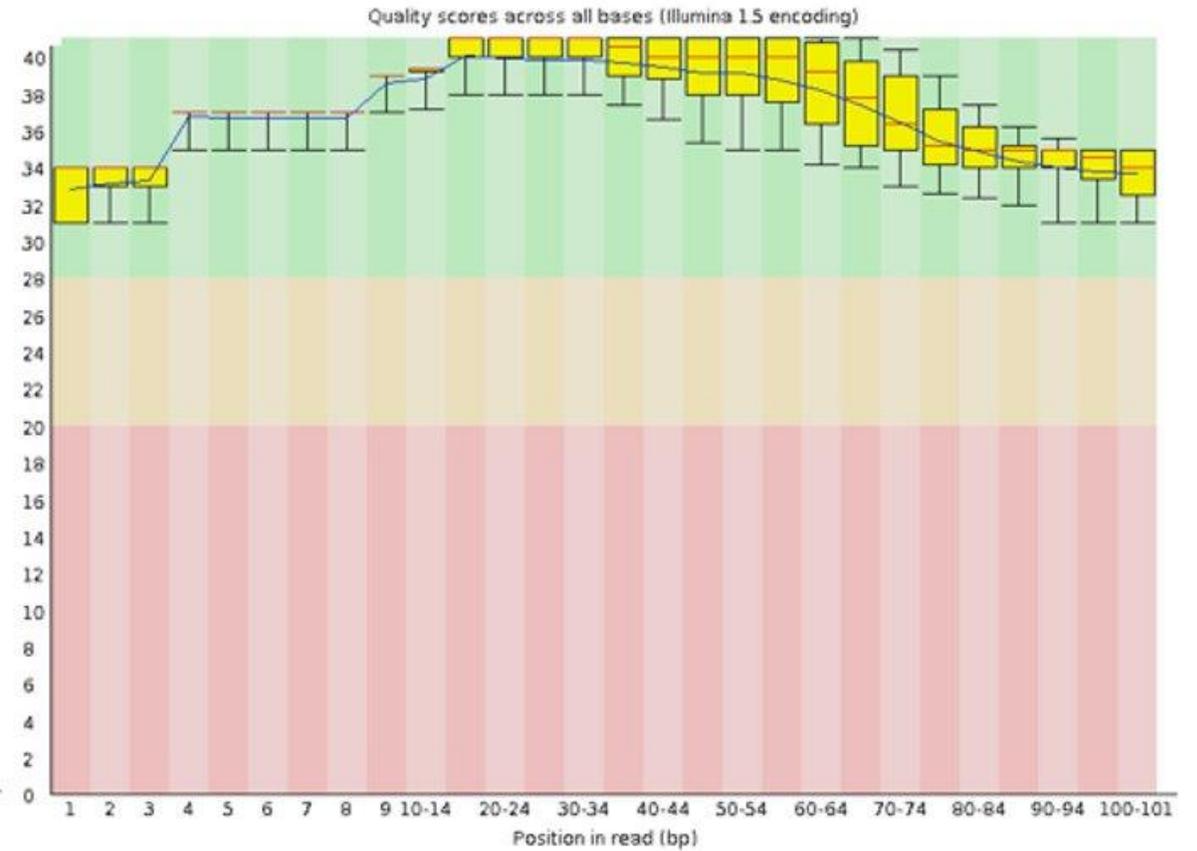
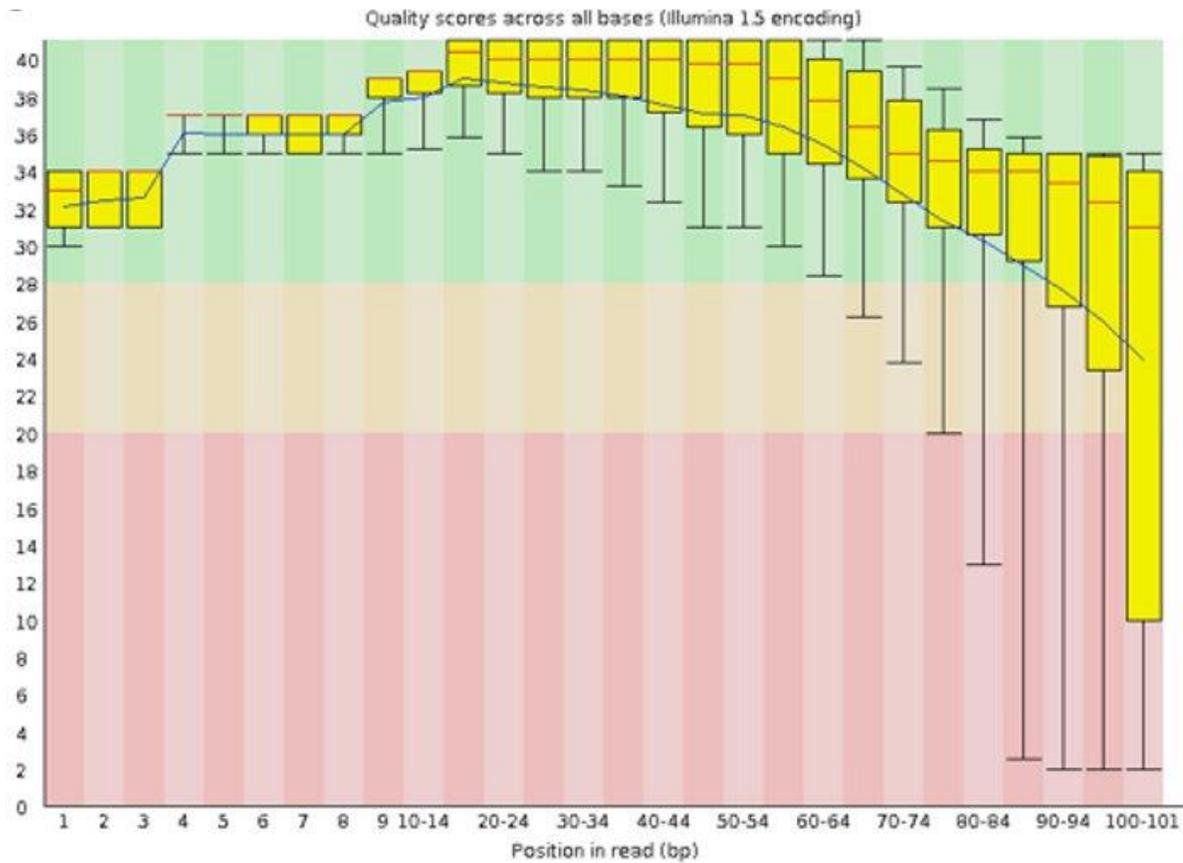
<https://www.multiqc.info>

MultiQC also works with other tools outputs:

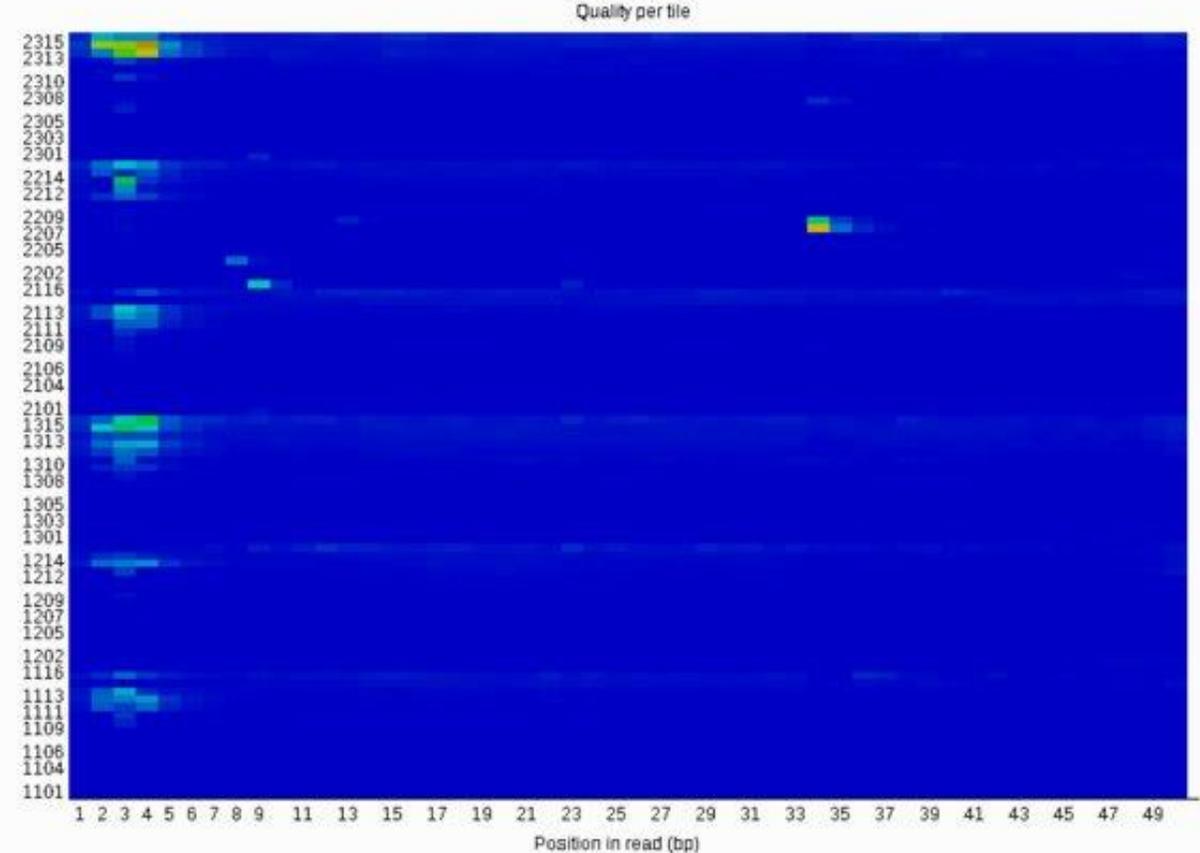
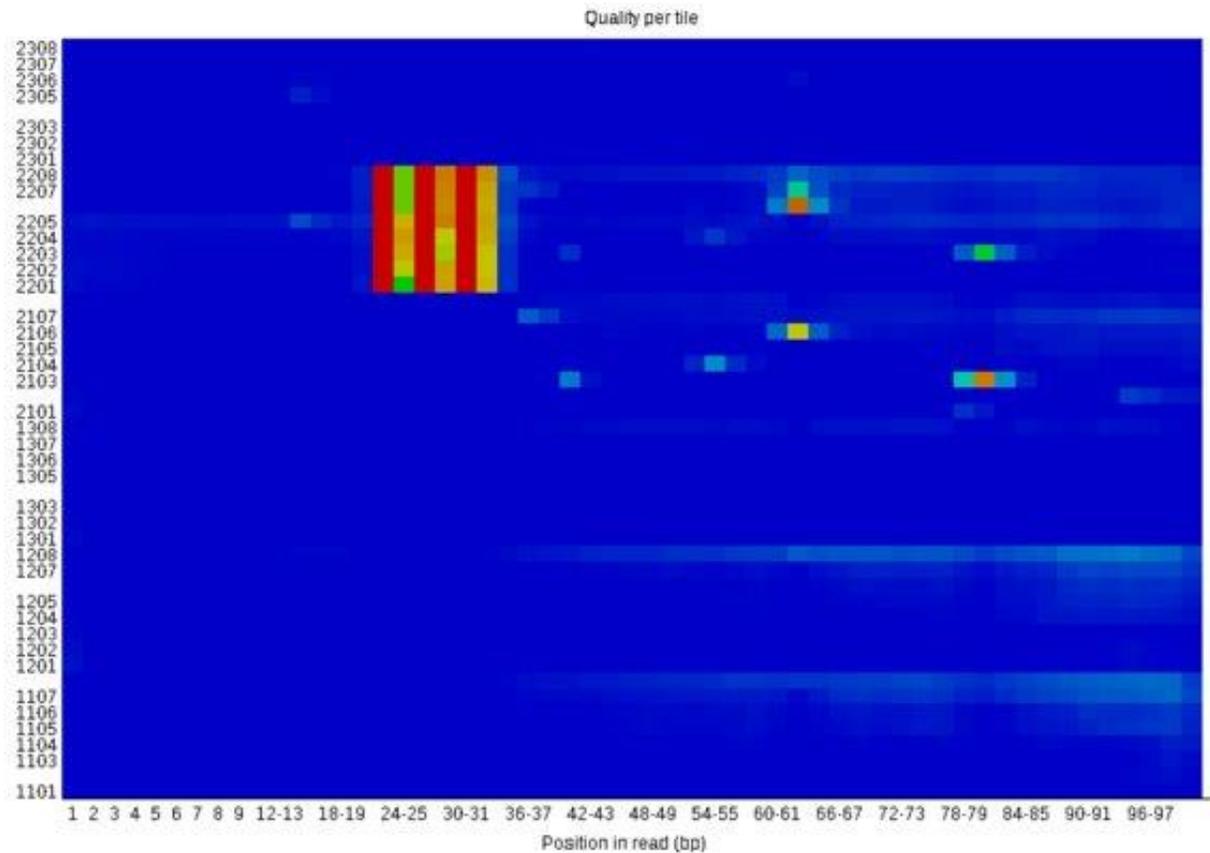
- Trimming outputs
- Mapping outputs
- ...

# Practical

# Per base sequence quality

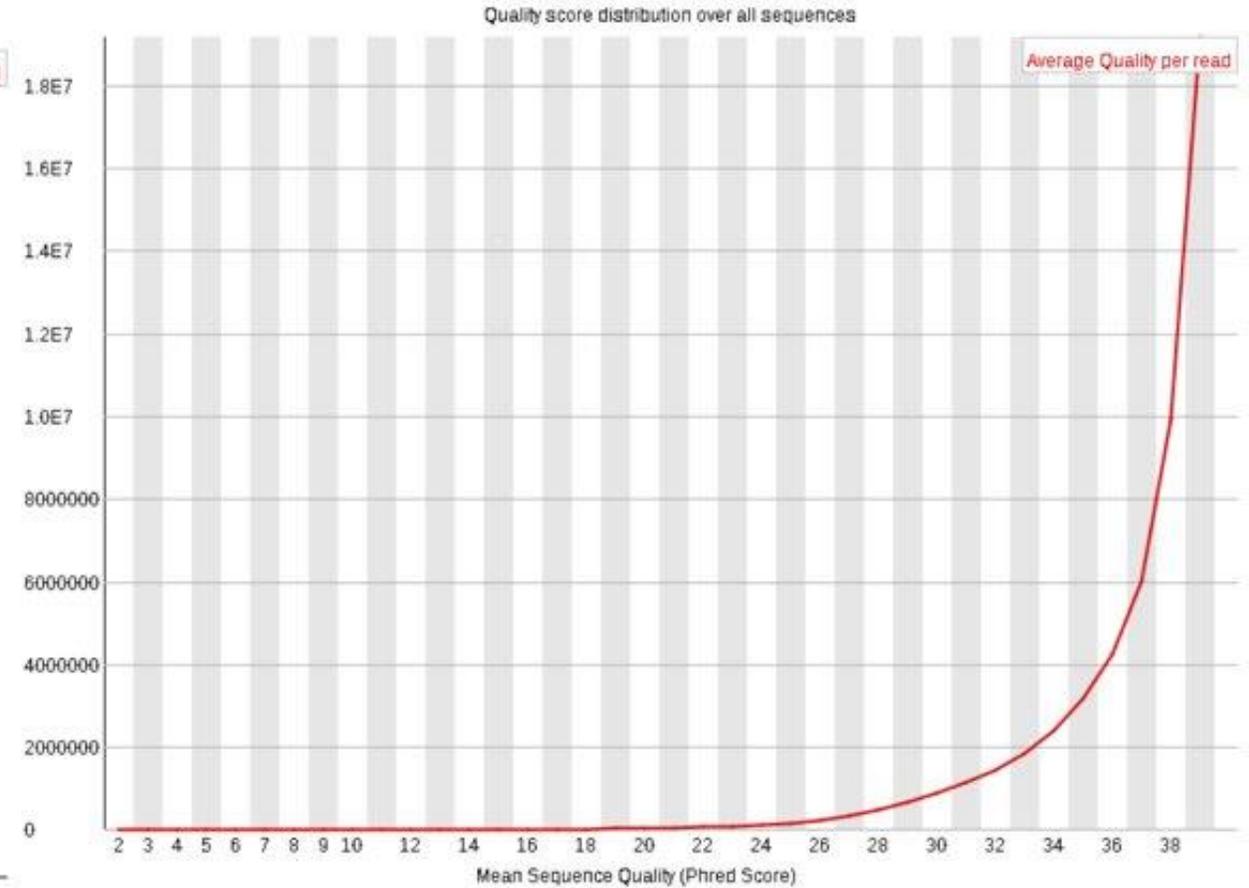
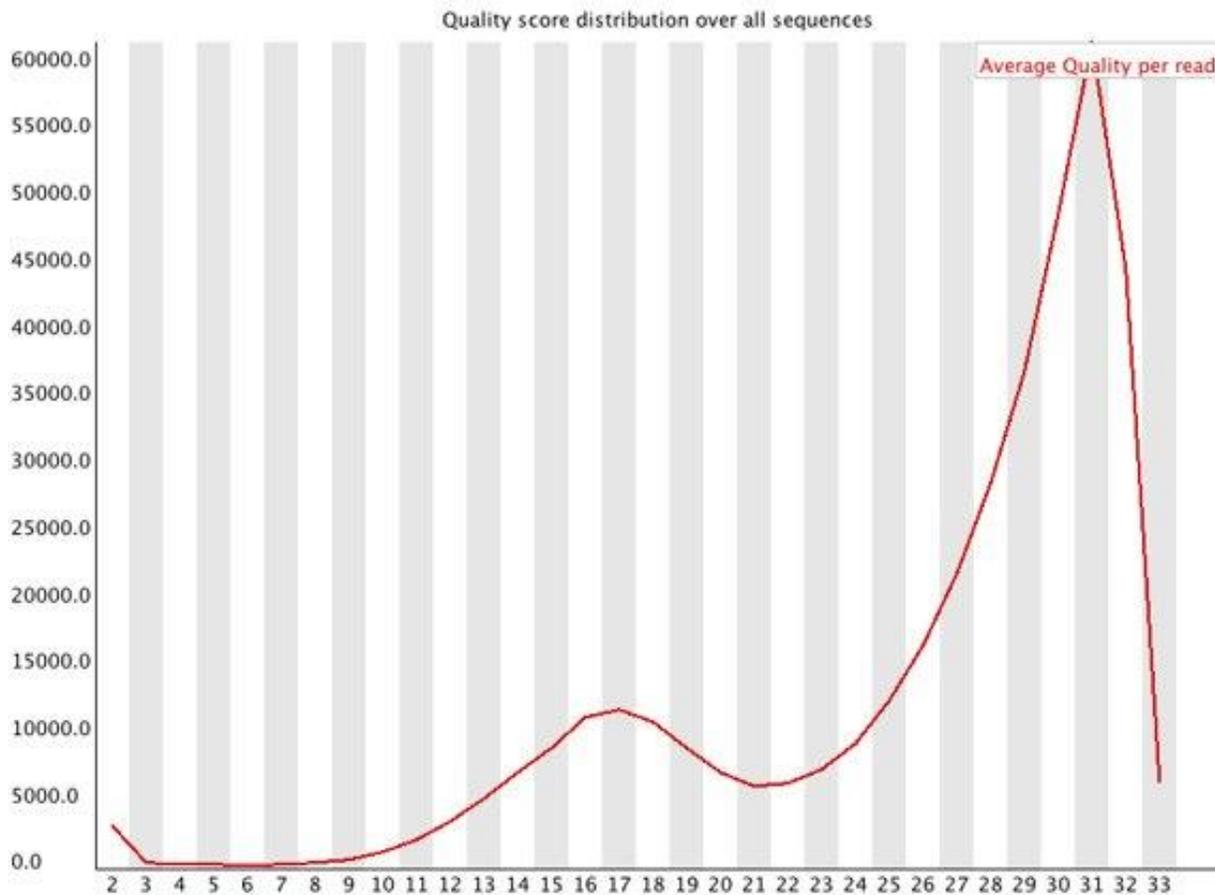


# Per tile sequence quality

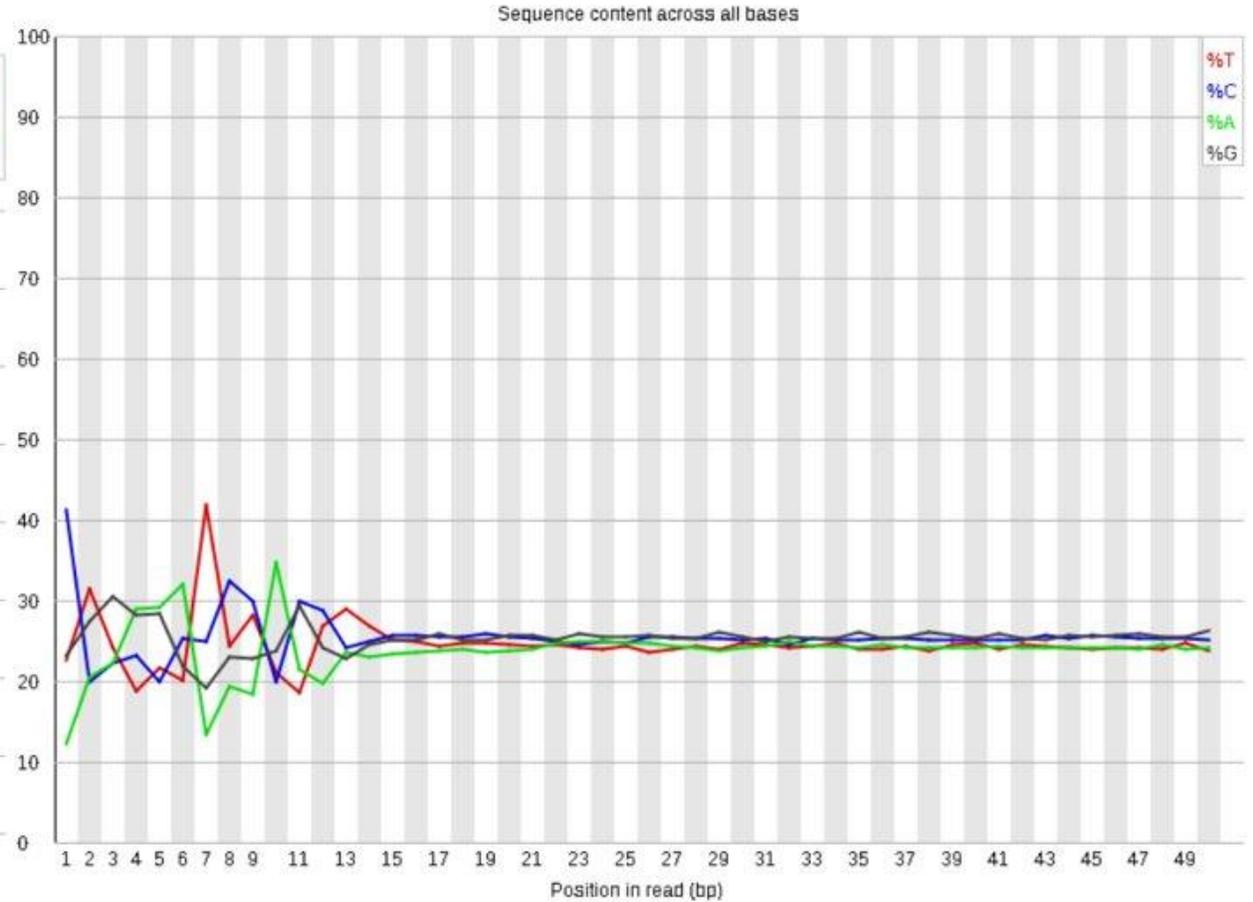
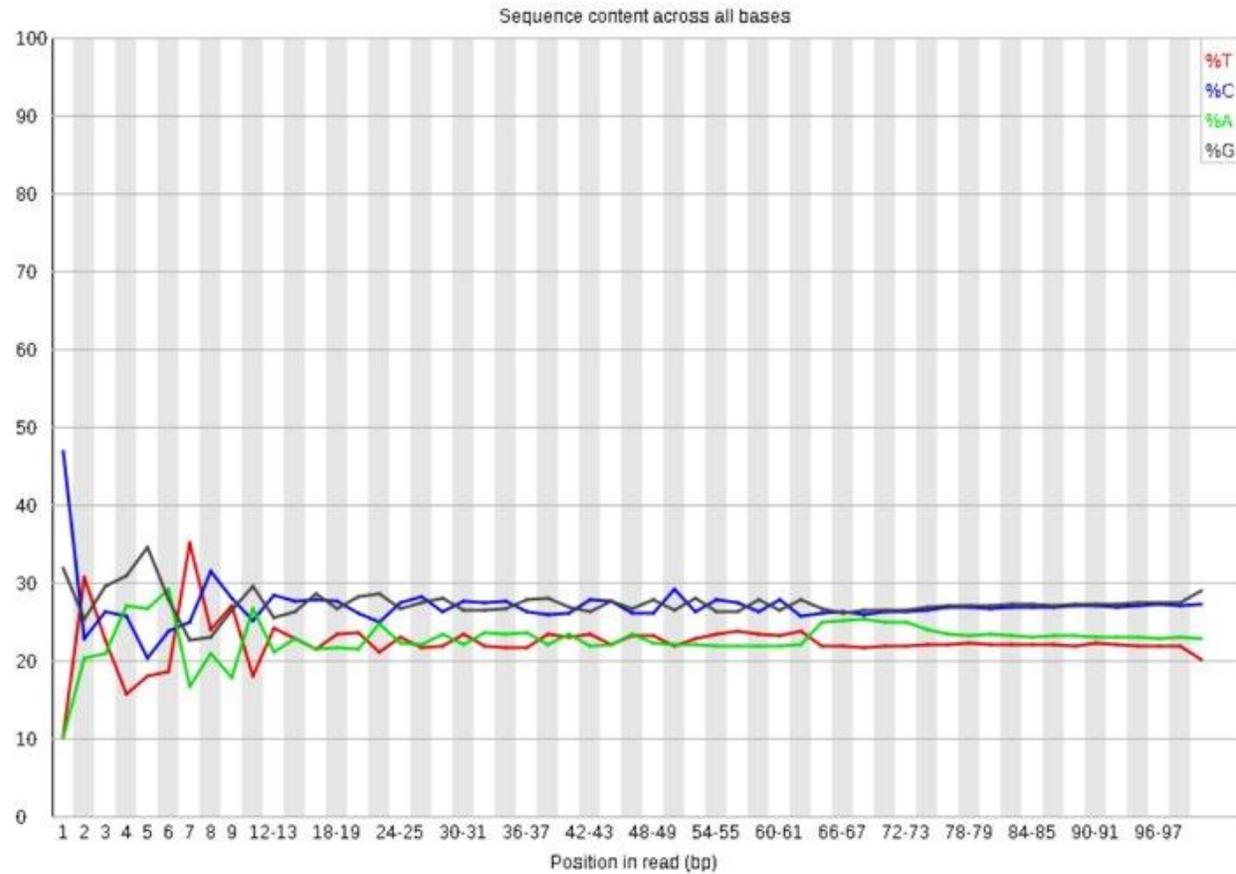


Only present when the fastq id contains the tile id

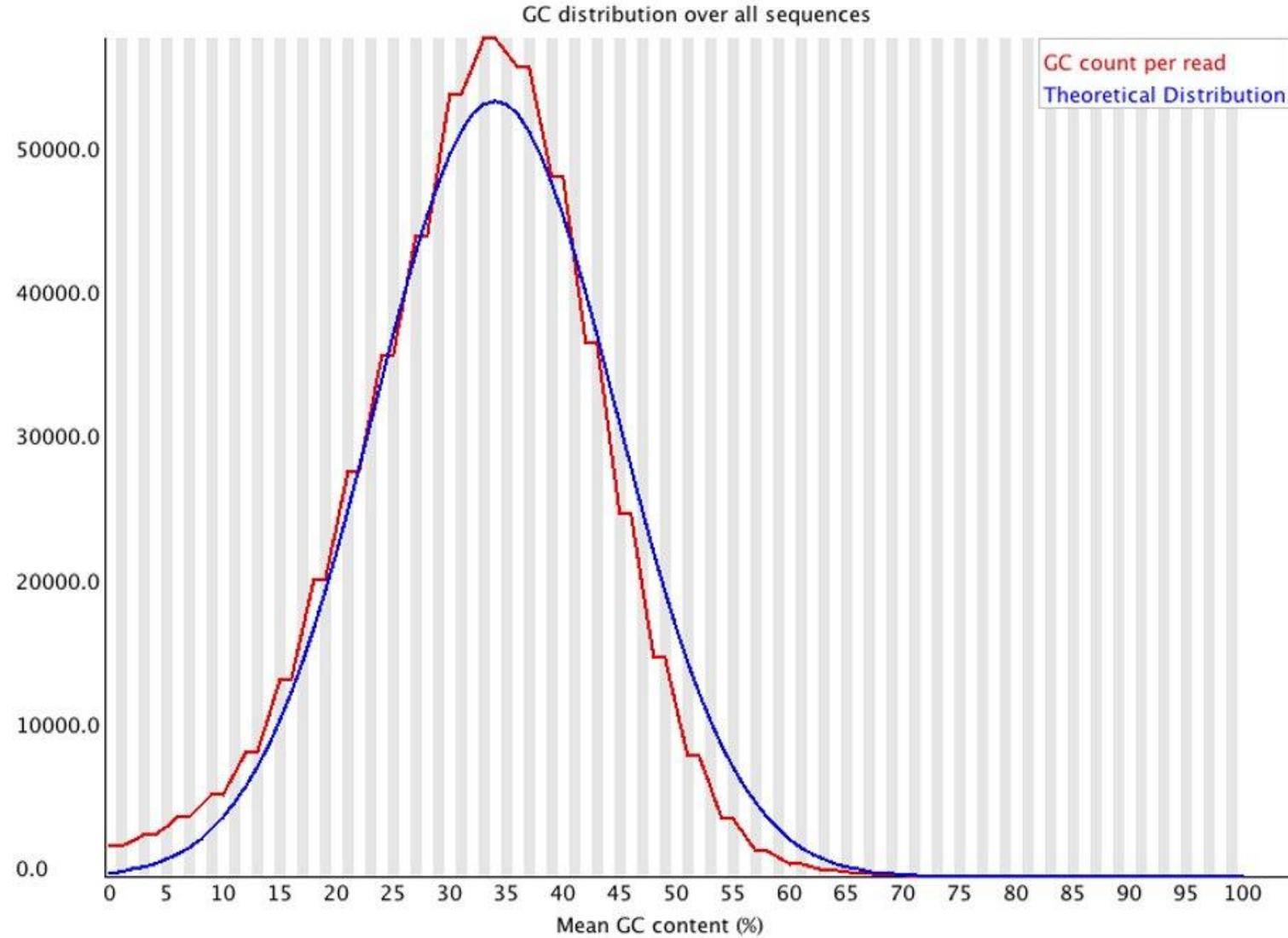
# Per sequence quality score



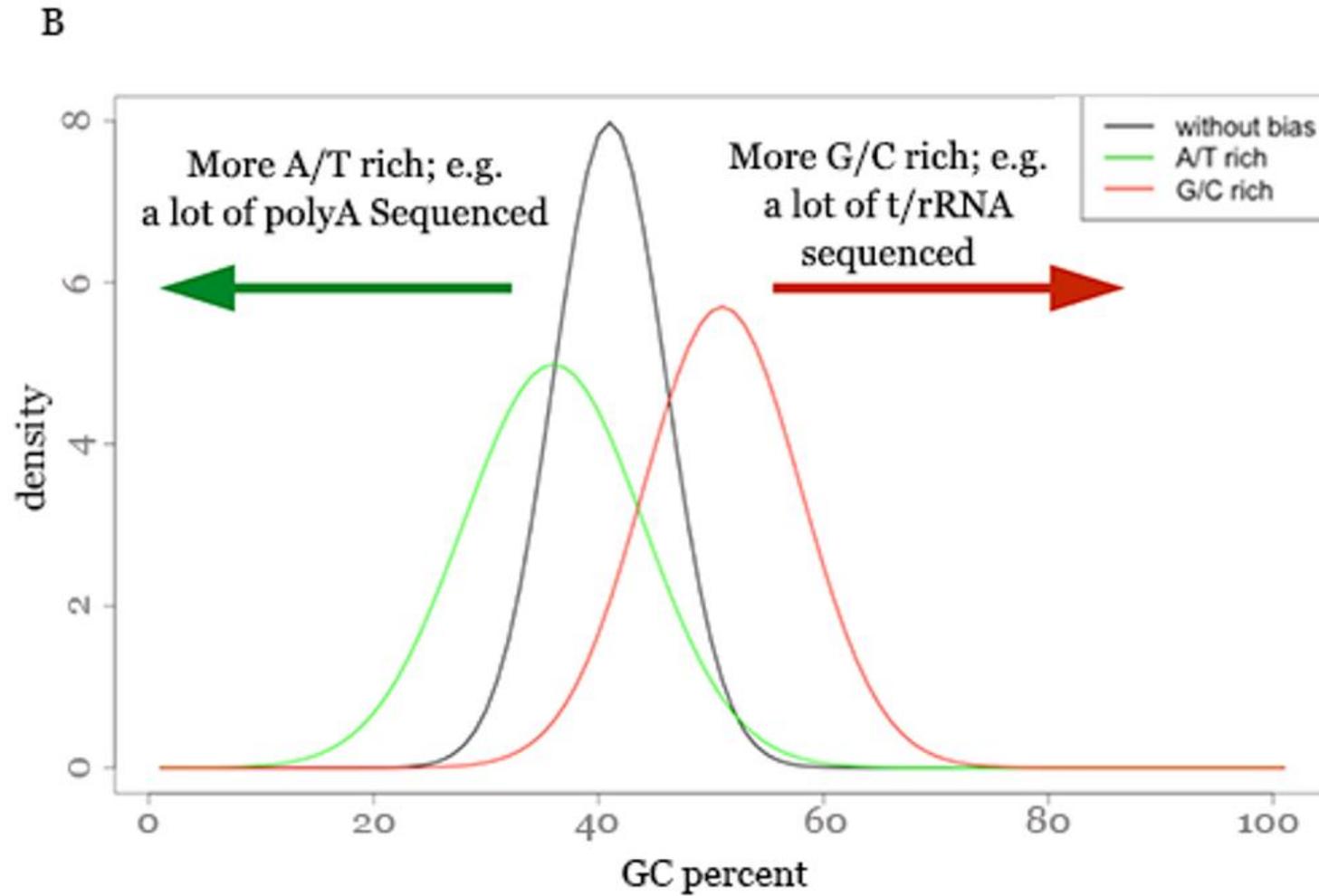
# Per base sequence content



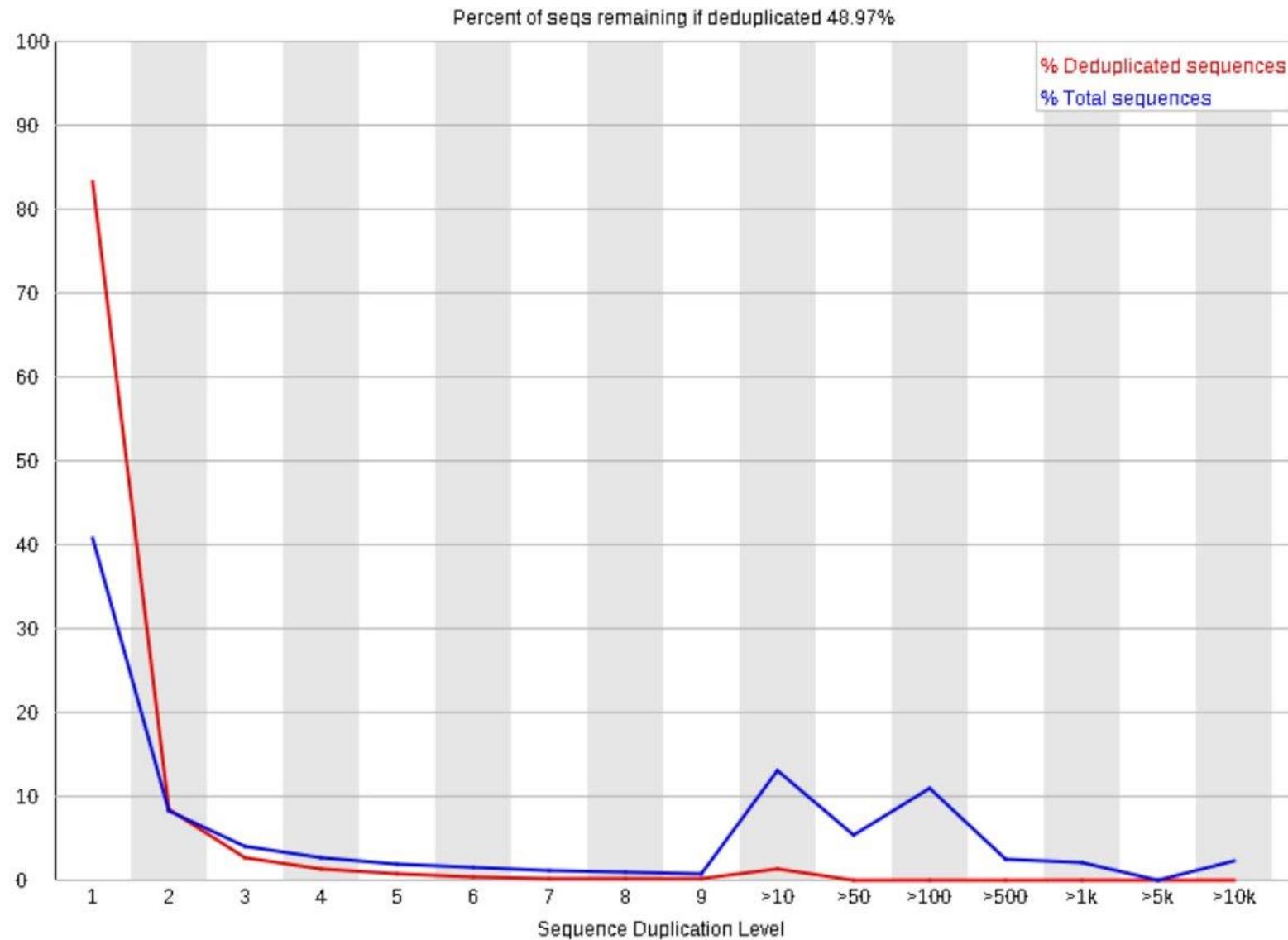
# Per sequence GC content



# Per sequence GC content

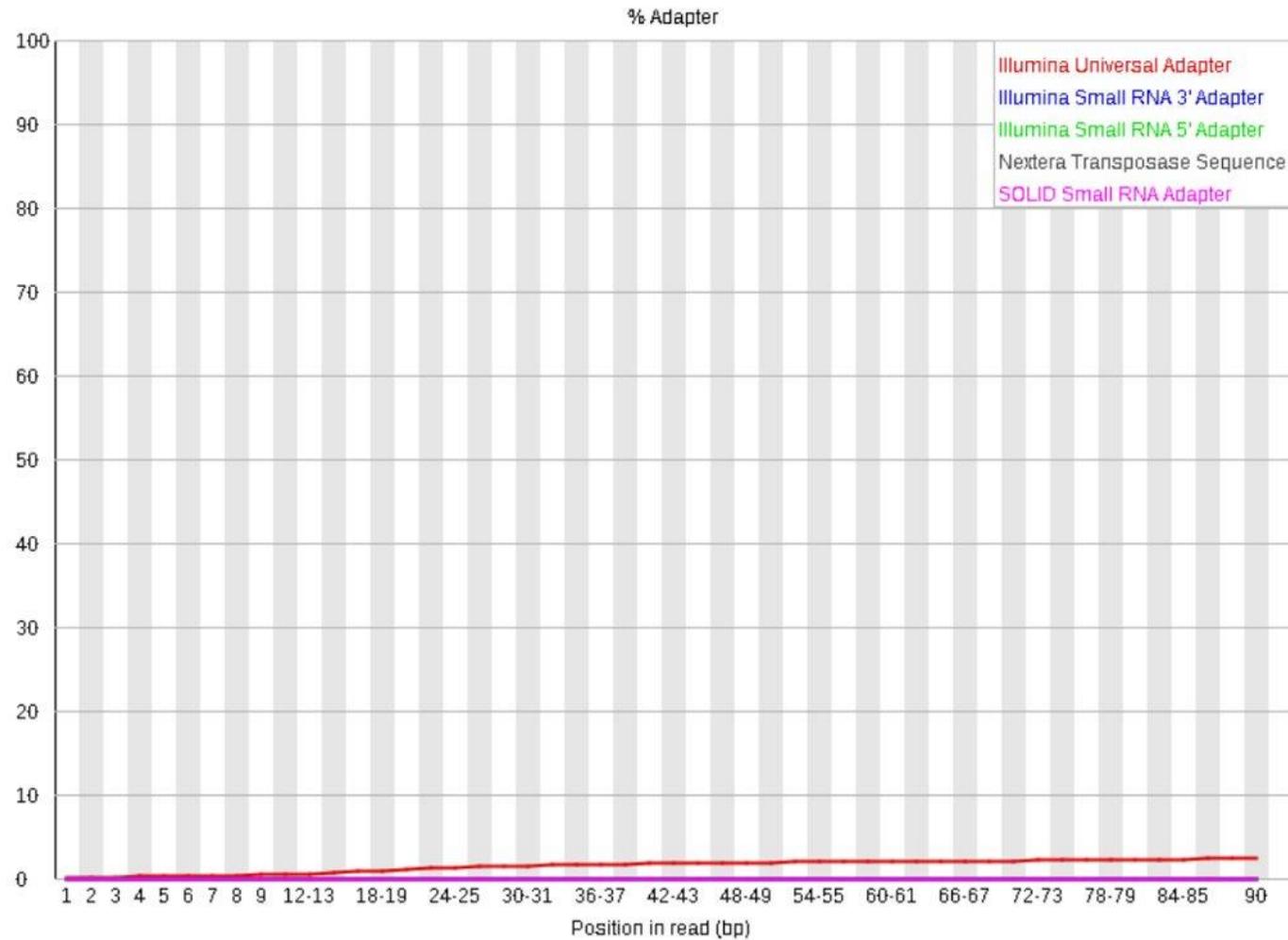


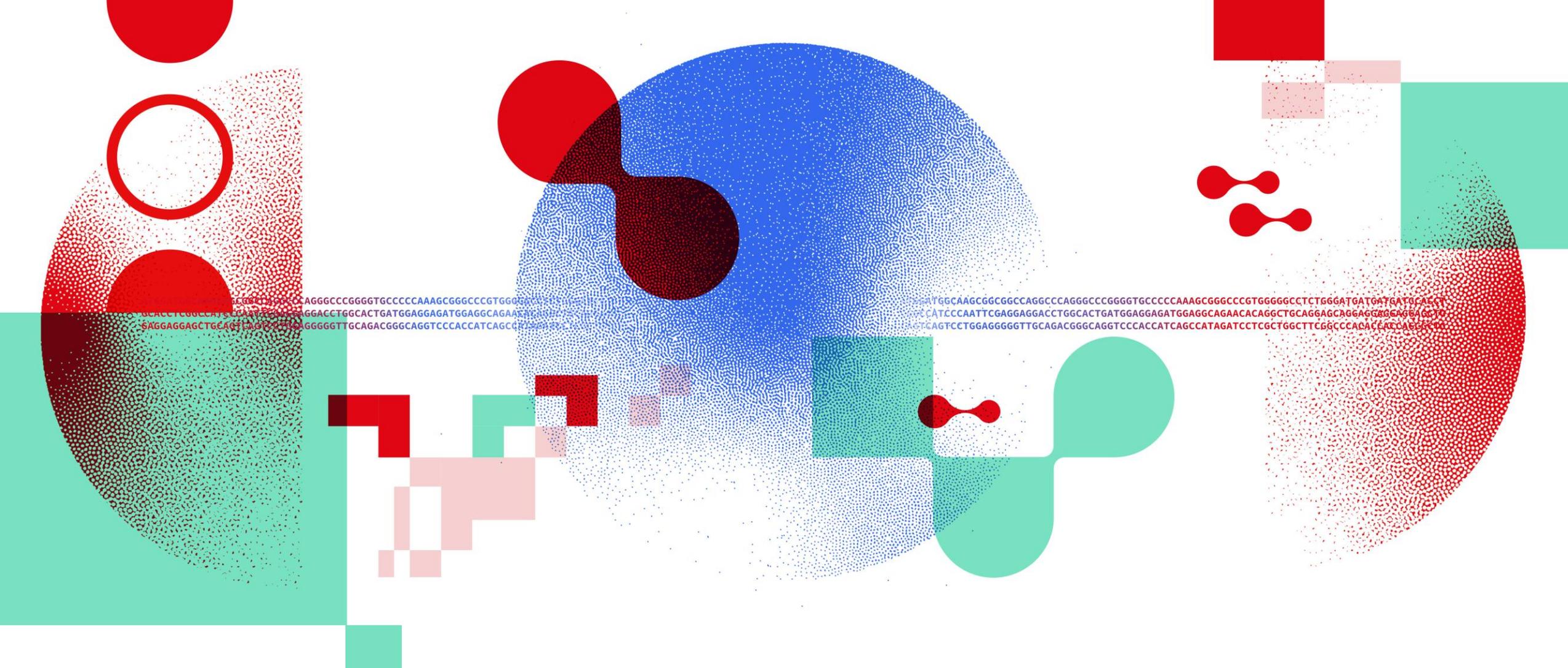
# Duplicate sequences



# Over-represented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGC	355643	2.113348167370486	TruSeq Adapter, Index 5 (100% over 50bp)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATG	42318	0.2514675327414971	TruSeq Adapter, Index 5 (100% over 49bp)





...AGGGCCCGGGTGCCCCAAAGCGGGCCGTGGG...  
...GACCTCGCCATGCAATG...GGACCTGGCACTGATGGAGGAGATGGAGGCAGAA...  
...SAGGAGGAGCTGCAGT...AGGGGGTTGCAGACGGGCAGGTCCACCATCAGCC...  
...TGGCAAGCGGGCGGCAGGCCAGGGCCCGGGTGCCCCAAAGCGGGCCGTGGGGCCTCTGGGATGATGATGATGCACT...  
...CATCCAATTCGAGGAGGACCTGGCACTGATGGAGGAGATGGAGGCAGAACACAGGCTGCAGGAGCAGGAGGAGGAGGAGCT...  
...TCAGTCCTGGAGGGGGTTGCAGACGGGCAGGTCCACCATCAGCCATAGATCCTCGCTGGCTTCGGCCACACATCAGGCT...

Thank you