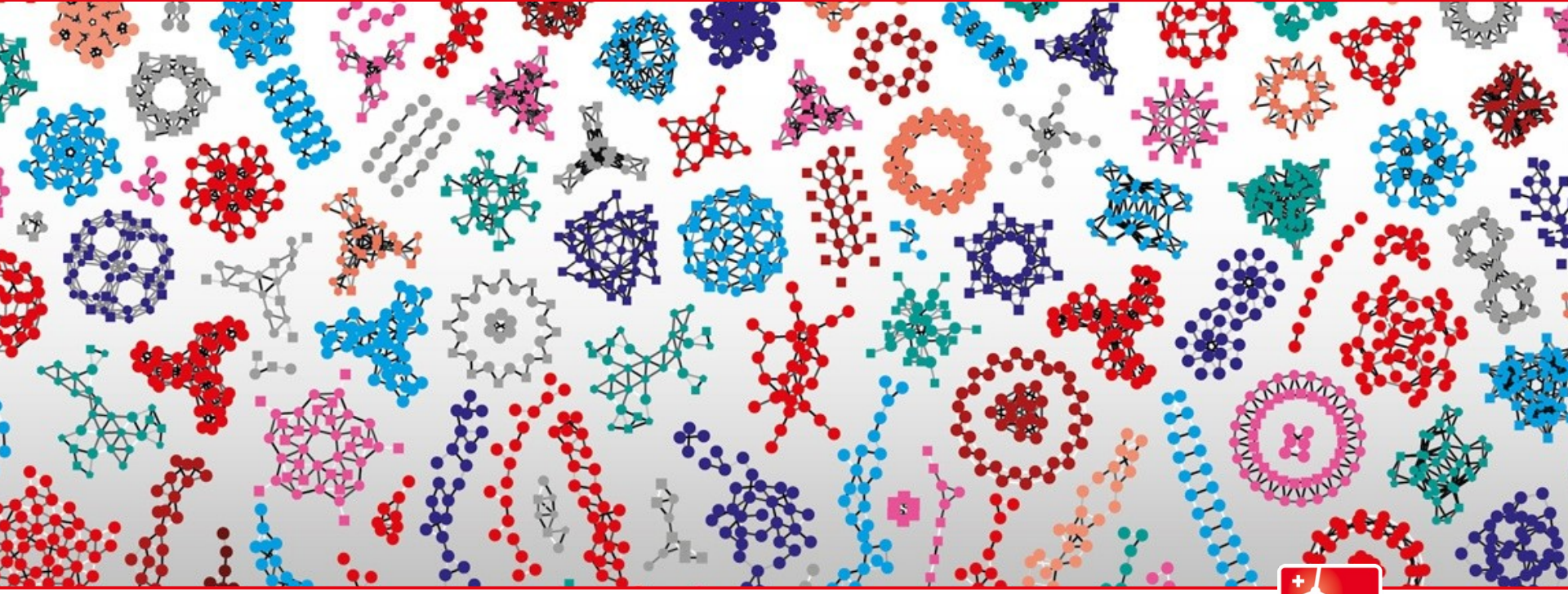


Swiss Institute of Bioinformatics

Introduction to RNA-Seq – Quality Control

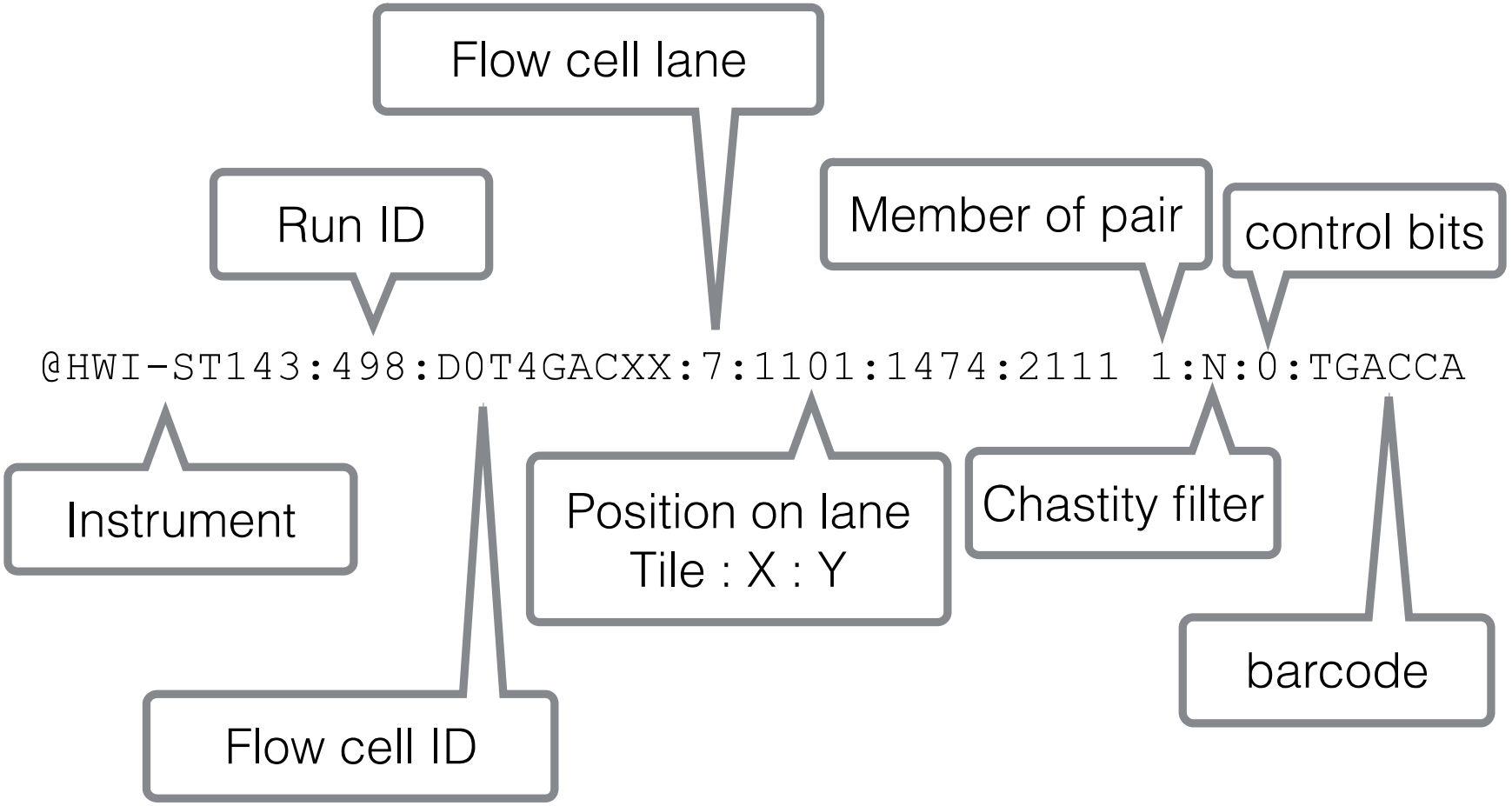
Wandrille Duchemin

”Raw data”: FASTQ format

- **FASTQ format stores sequences and “*Phred*” quality score in a single file**
  - Sequence header starts with @
  - Quality header usually starts with +
- **Paired data: two separate files for forward and reverse with same ordering**

```
@HWI-ST143:498:D0T4GACXX:7:1101:1474:2111 1:N:0:TGACCA
TCTCNAGATAAAATCAAAACCAACAGAGAGTCTAGAATAAAAGTGAATAG
+
@@BF#2ADHHHHHHJJJJIJJJJJJIJGJIJHIIGIHIIIIIJJHIHIJJJ
@HWI-ST143:498:D0T4GACXX:7:1101:1424:2149 1:N:0:TGACCA
CAGATGTTTTGGTAAGAAGAATGCCATTAGCAATTGGTGGAGGAGCAGG
A
+
BBCFFFEFHHHHHJIJIJJHIIJIIJJJJJJJJJJIJHIJIJJIJIGIIG
```

FastQ header



FASTQ Header

- **Depends on sequencing technology used, changed several times by Illumina and others**
  - Chastity filter switched (now N is good)
  - Representation of paired reads modified

@HWI-ST143:498:D0T4GACXX:7:1101:1474:2111 2:N:0:TGACCA

@HWI-ST143:498:D0T4GACXX:7:1101:1474:2111#TGACCA/2

Phred Quality Scores

- **ASCII encoded**
- **Represents probability (p) that base call is incorrect**
  - Quality (Q) = $-10 \log_{10}(p)$

P-value	Phred	Probability of incorrect base call	Base call accuracy
1E-01	10	1 in 10	90%
1E-02	20	1 in 100	99%
1E-03	30	1 in 1000	99.9%
1E-04	40	1 in 10,000	99.99%

Phred Quality Scores

Sanger, Illumina v1.3 to 1.7 (ASCII_BASE=64)

Q	ASCII	P	Q	ASCII	P	Q	ASCII	P	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	Q	ASCII	P	Q	ASCII	P	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			

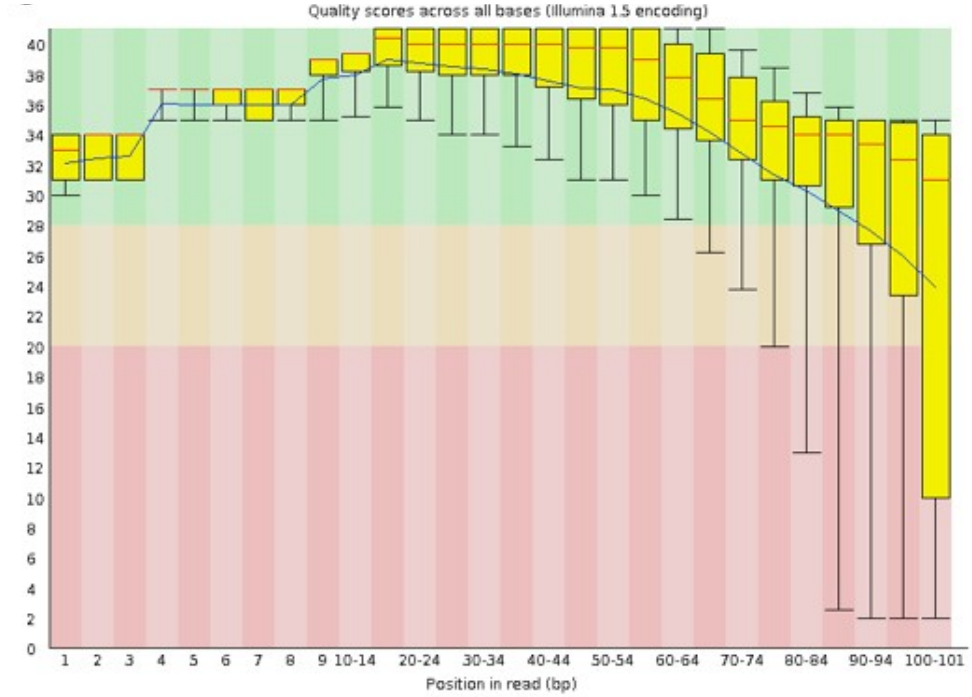
Quality Control using FastQC

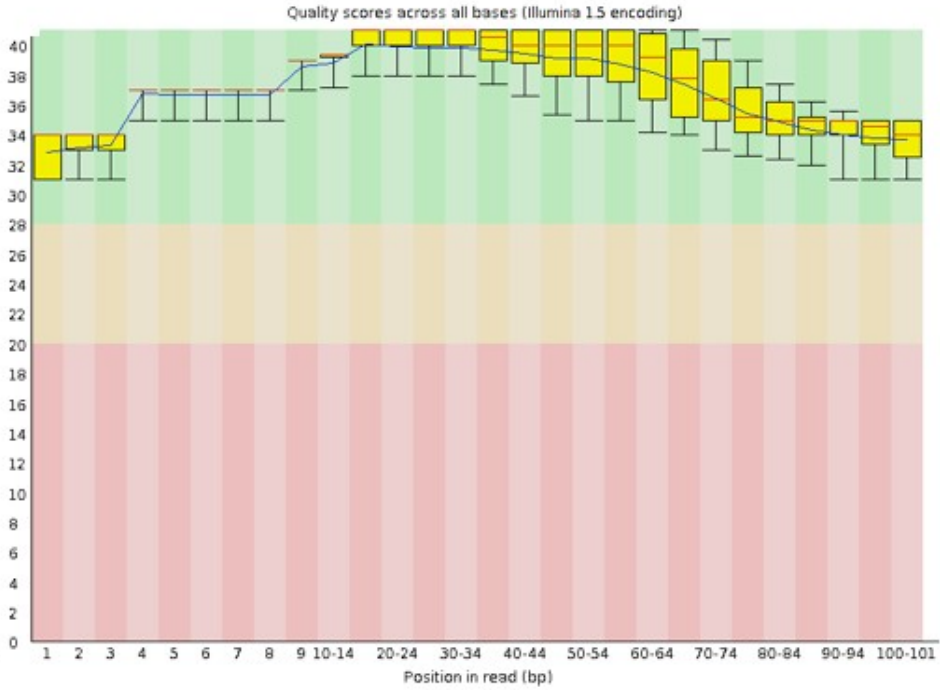
- **FastQC aims to provide a QC report which can spot problems which originate either in the sequencer or in the starting library material**
- **Generates an HTML report with various metrics**
- **Supported file formats:**
  - FASTQ (can be gzip compressed), SAM, BAM

PRACTICAL

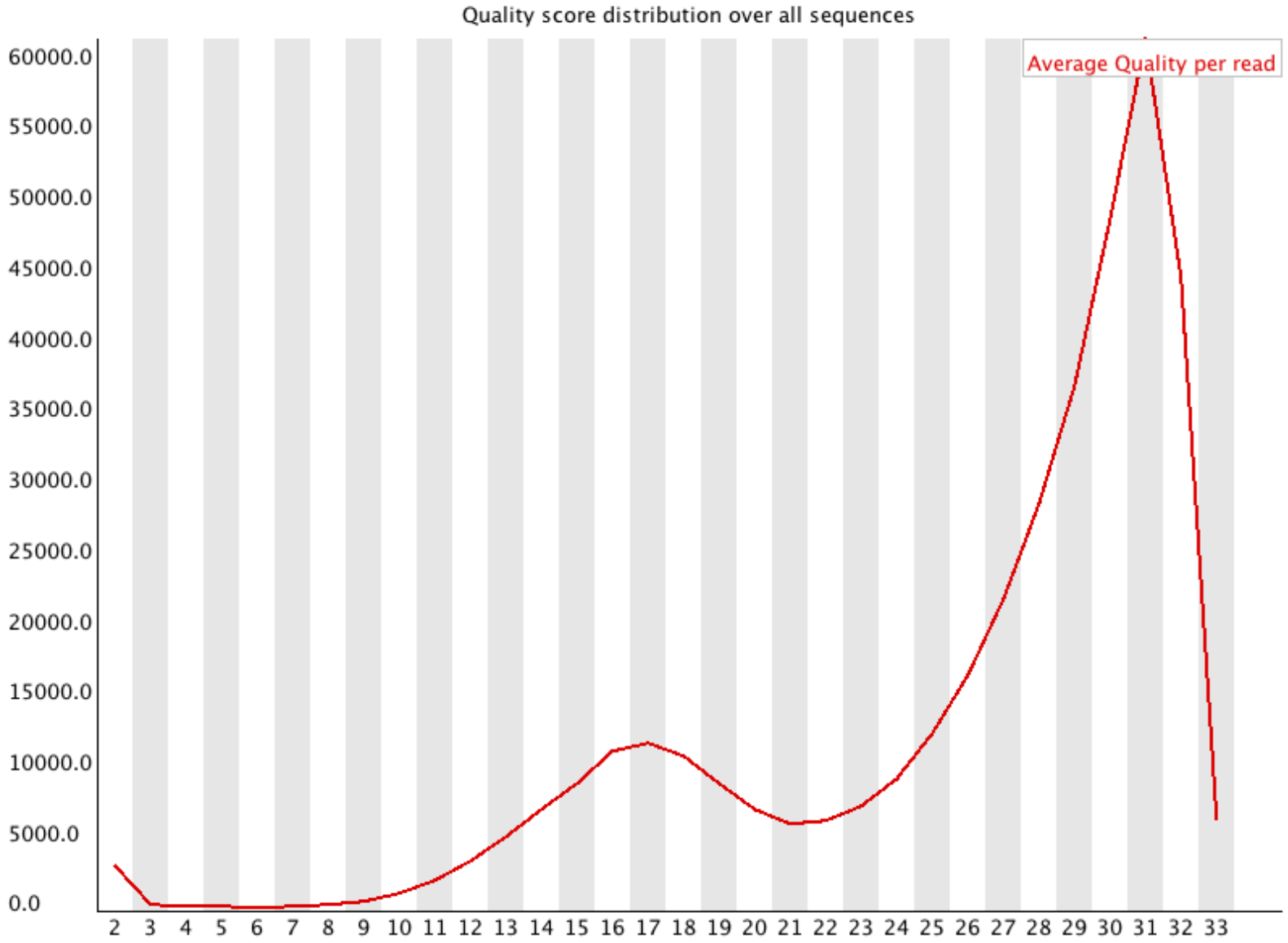
- **Go to the website and do the fastQC practical**

Per Base Sequence Quality

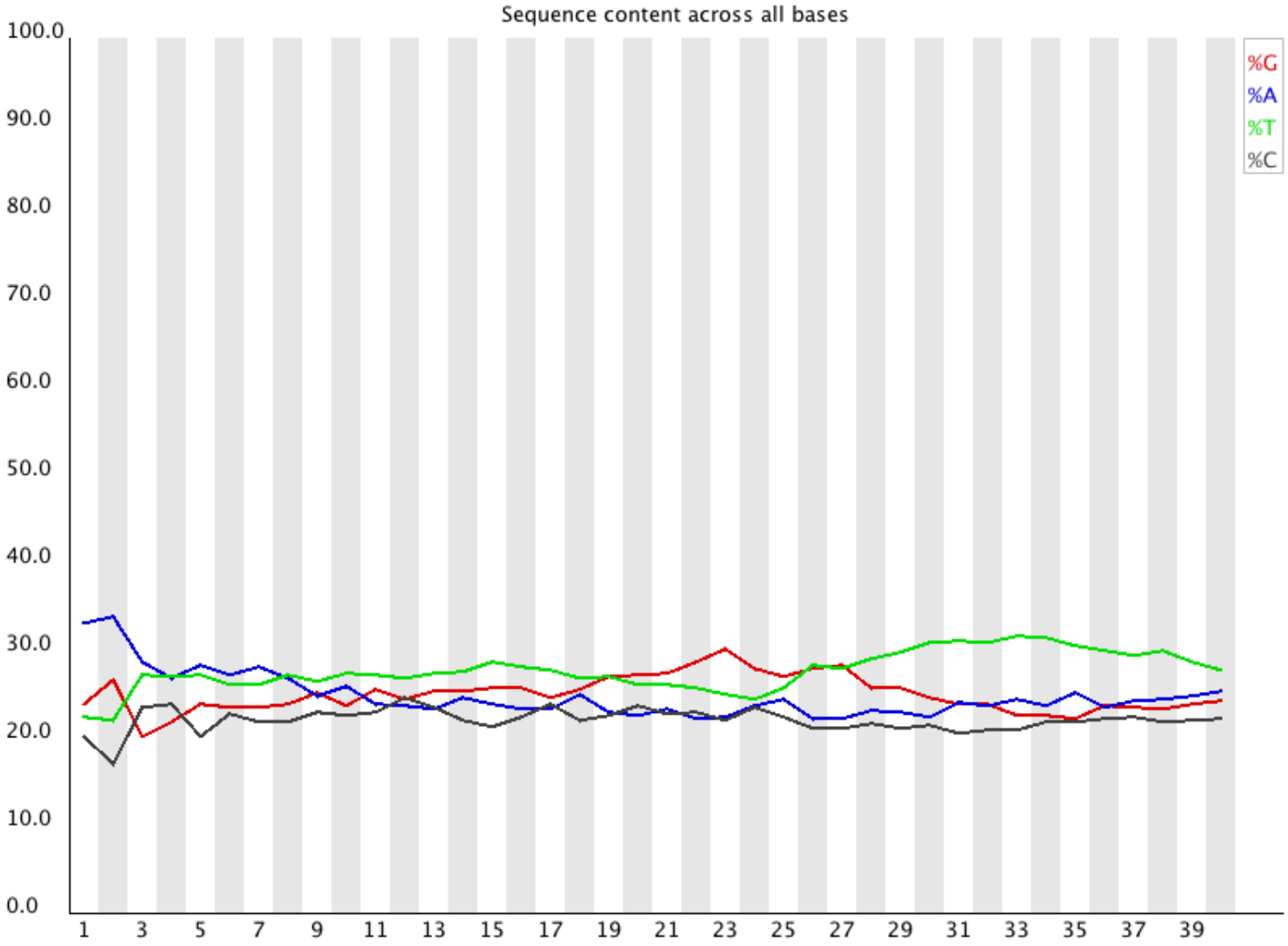




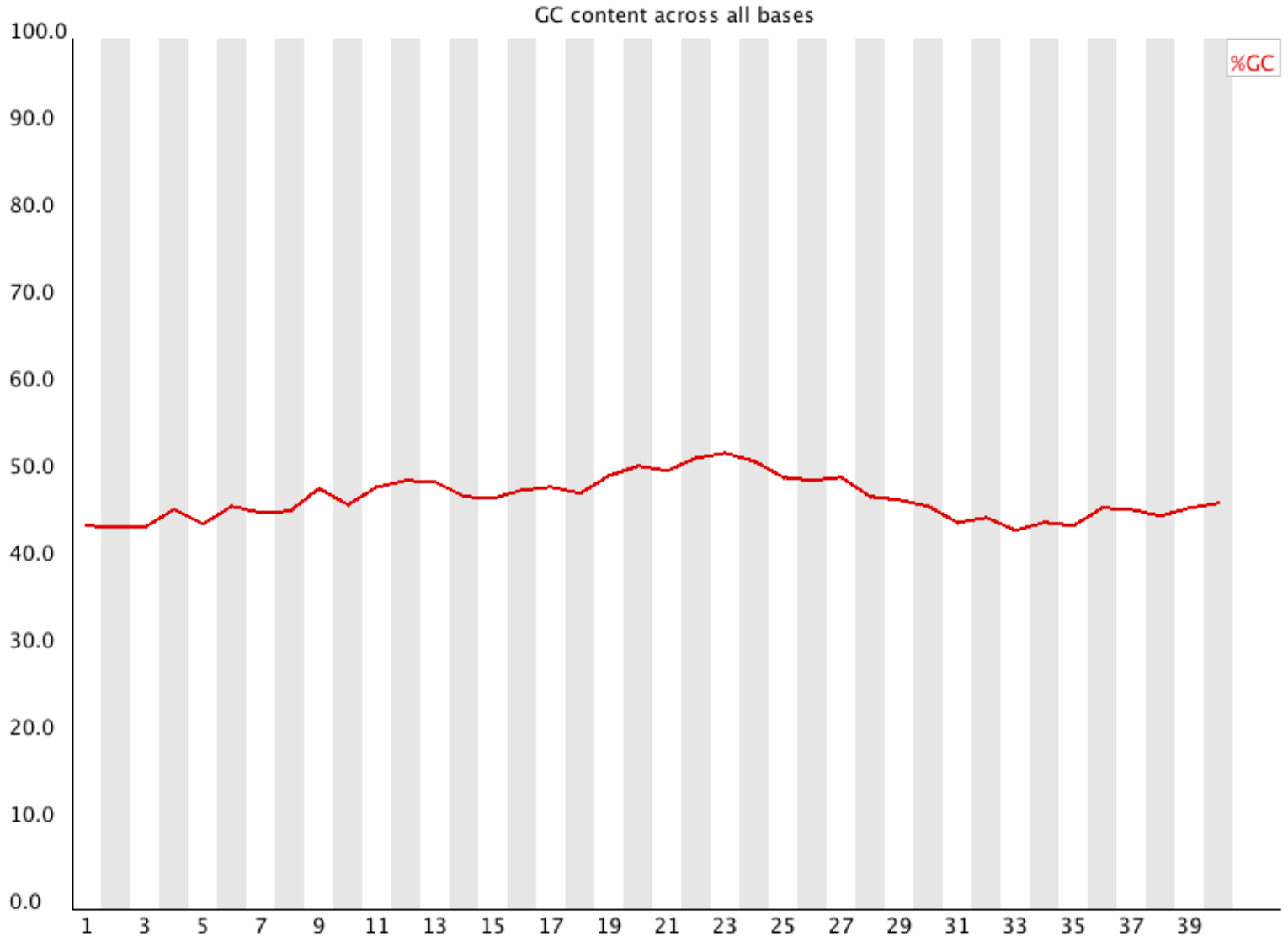
Per Sequence Quality Scores



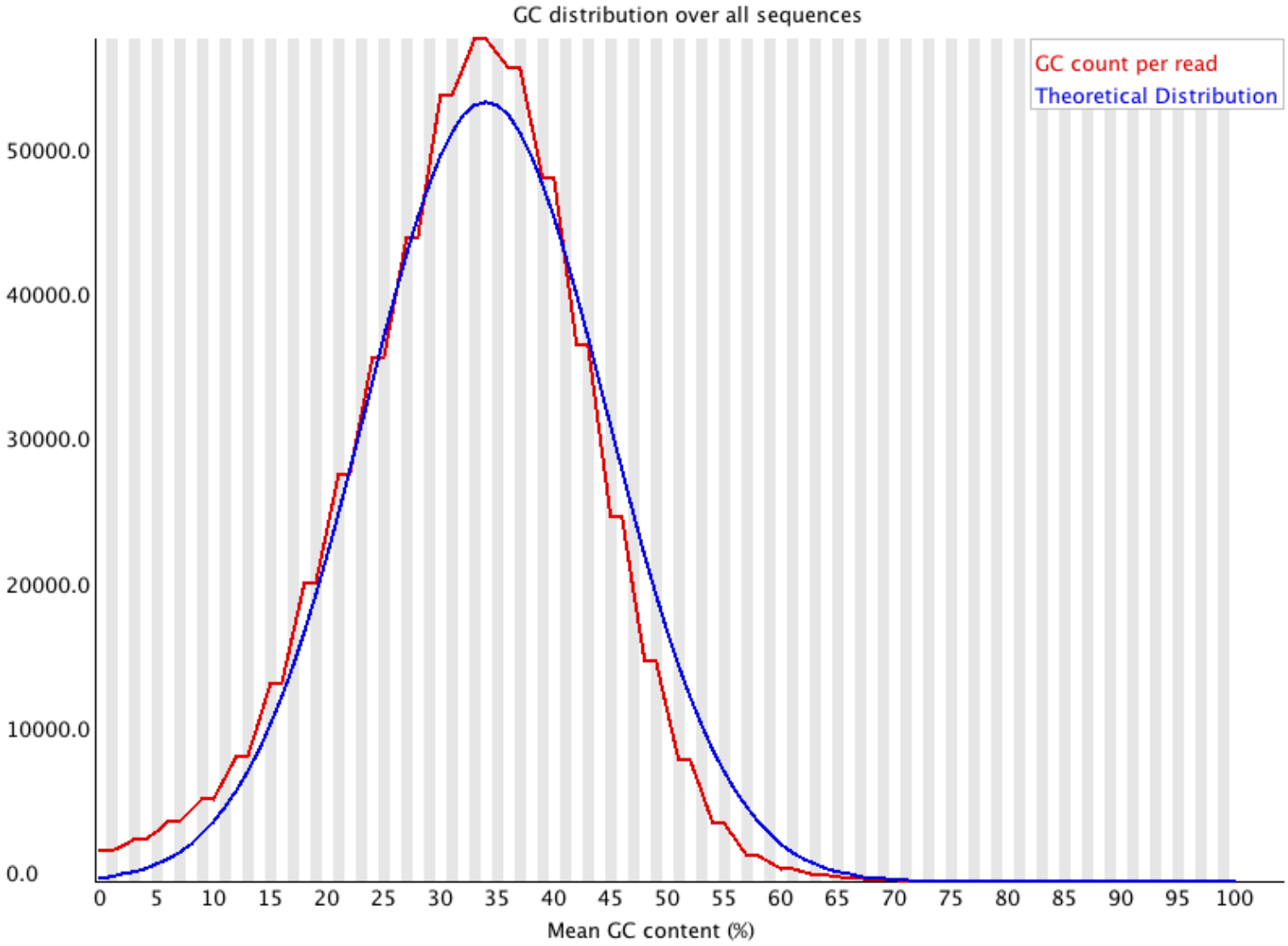
Per Base Sequence Content



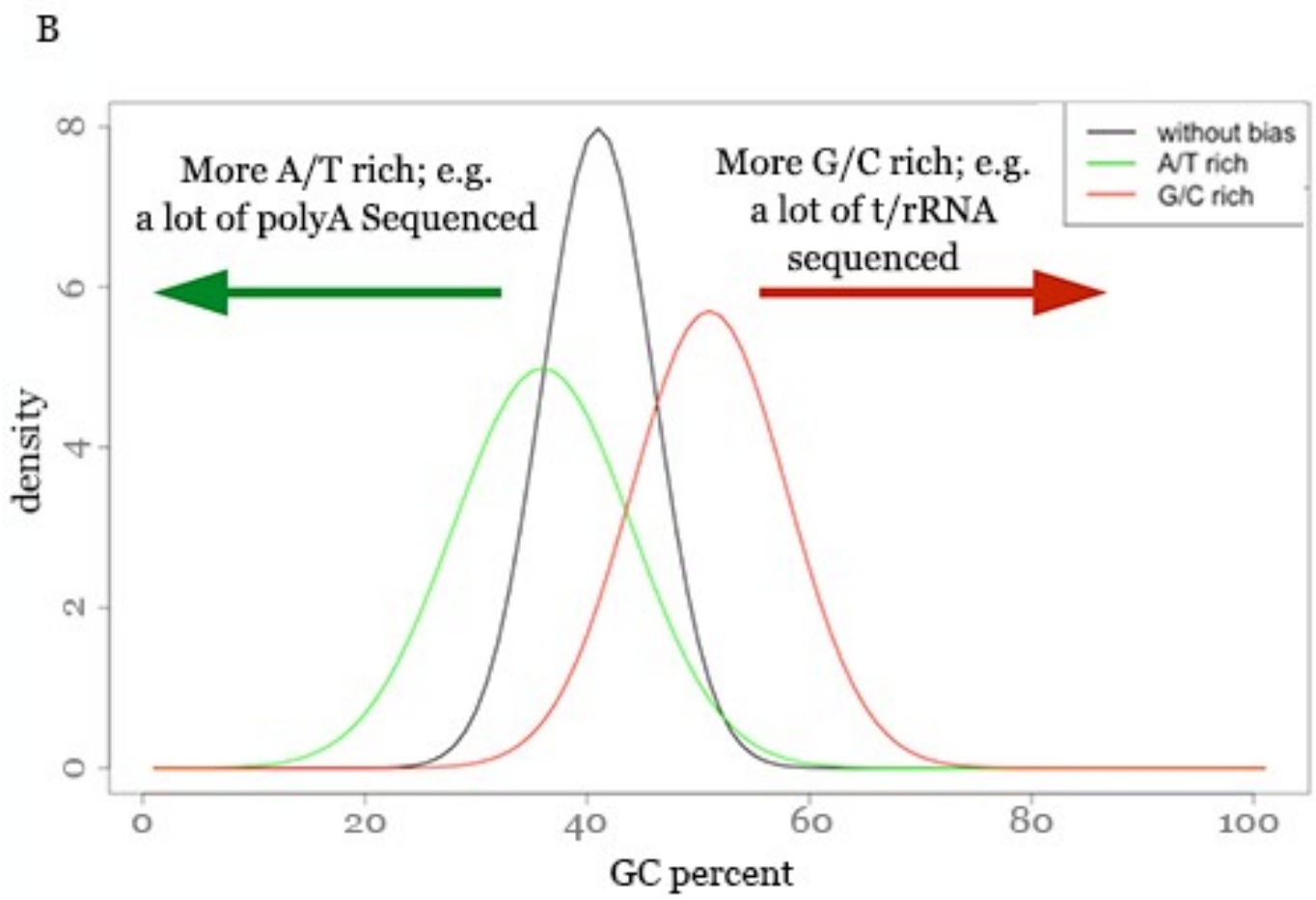
Per Base GC Content



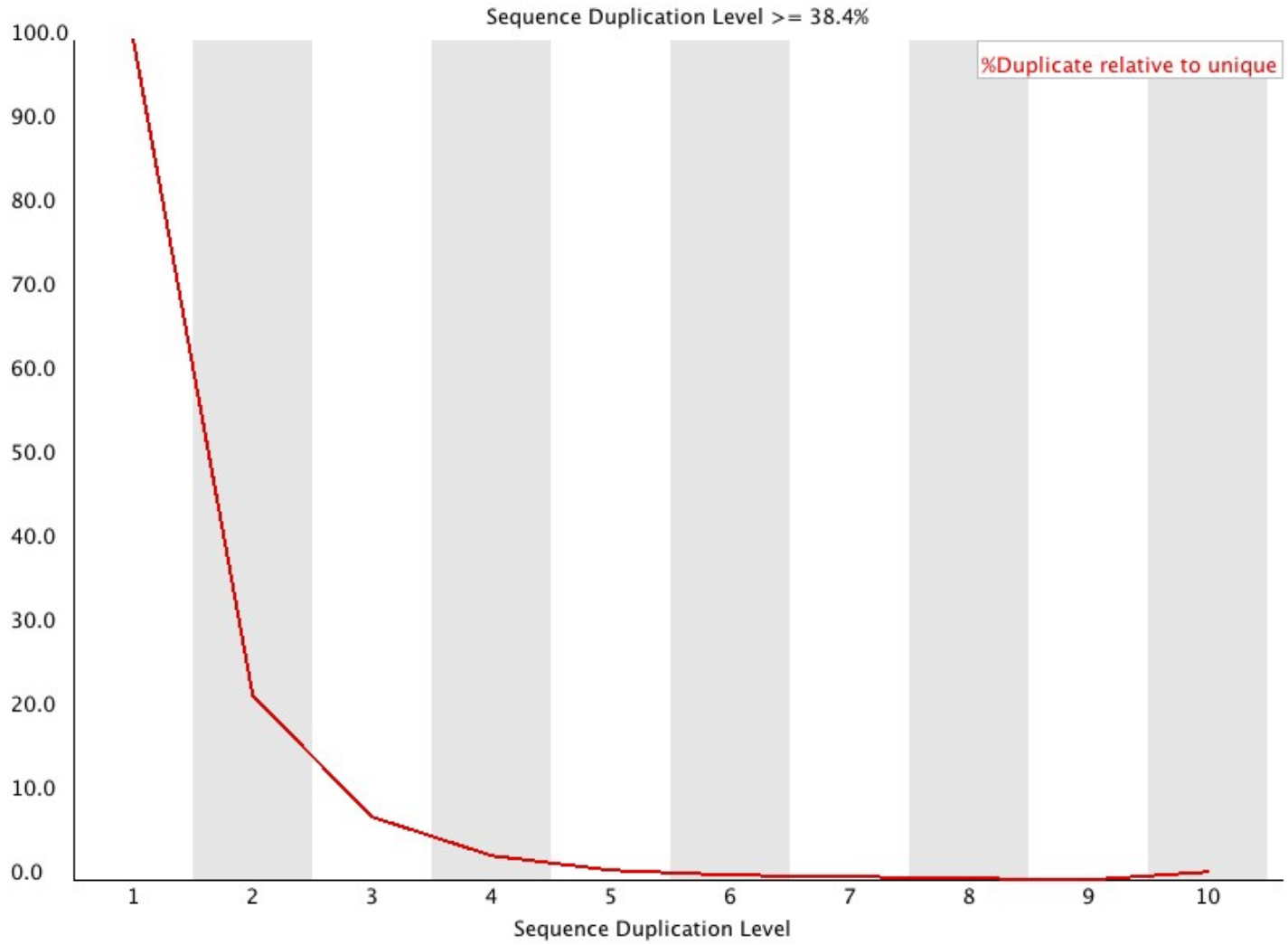
Per Sequence GC Content



Per Sequence GC Content



Duplicate Sequences



Overrepresented Sequences

FastQC Report

Mon 24 Feb 2014
SRR576933.fastq

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACA	1060621	29.432719567181643	TruSeq Adapter, Index 5 (100% over 36bp)
GCTAACAAATACCCGACTAAATCAGTCAAGTAAATA	13630	0.37823875606902535	No Hit
NATCGGAAGAGCACACGTCTGAACTCCAGTCACACA	11728	0.3254573830651159	TruSeq Adapter, Index 5 (97% over 36bp)
GTTAGCTATTTACTTGACTGATTTAGTCGGGTATTT	10983	0.304783291115635	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACC	3658	0.10151117899490057	TruSeq Adapter, Index 1 (97% over 36bp)

Kmer Content

Produced by FastQC (version 0.10.1)

MultiQC Reports

- **Generally, you will be interested in how the QC metrics of all your samples compare with each other**
- **The software MultiQC will combine your per sample FastQC reports into a single report**

PRACTICAL

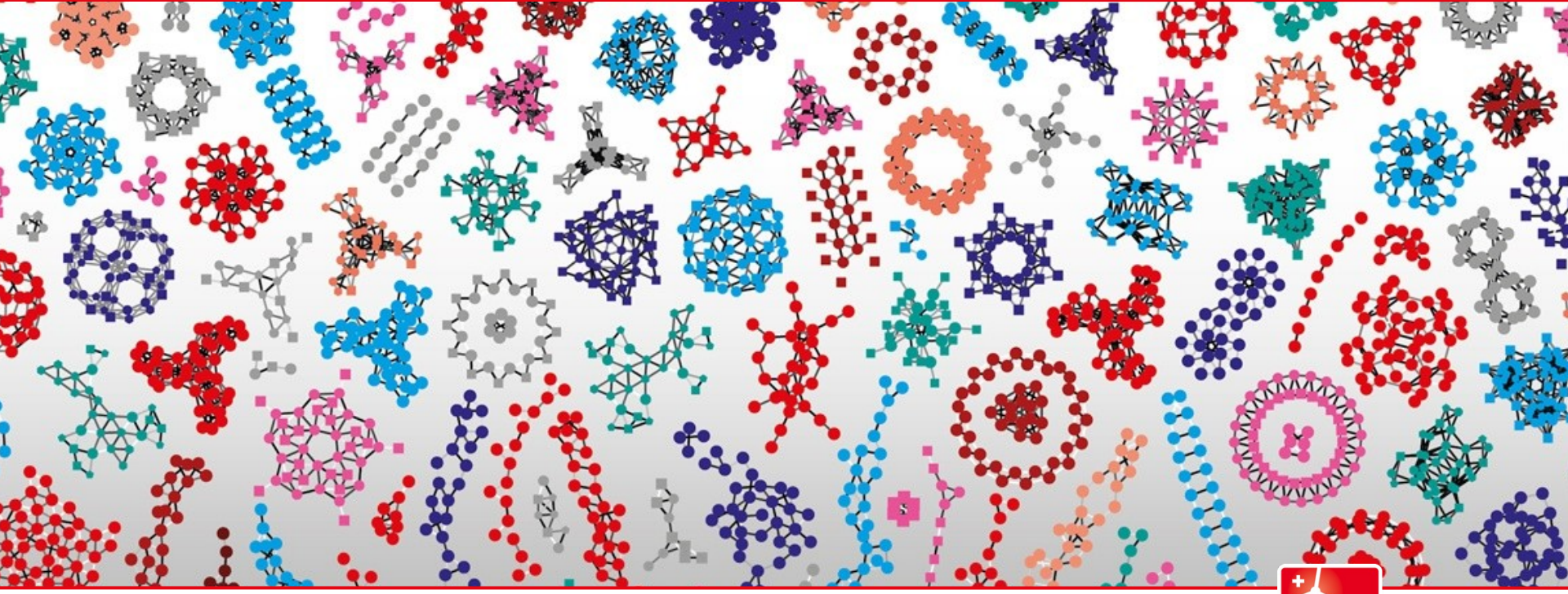
- **Go to the website and do the multiQC practical**

REFERENCES

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

P. Ewels, et al. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics (2016). doi: 10.1093/bioinformatics/btw354

http://multiqc.info



Swiss Institute of Bioinformatics

Contributors:

Wandrille Duchemin
Geoffrey Fucile
Walid Gharib
Pablo Escobar Lopez
Mihaela Zavolan