Better lives through livestock

Sequence Data Quality Control

Kennedy Mwangi

International Livestock Research Institute (ILRI)

Viral Pathogen Genome Sequencing and Bioinformatics Analysis Training Workshop
6th – 17th May, 2024

ILRI
INTERNATIONAL
LIVESTOCK RESEARCH
INSTITUTE

CGIAR

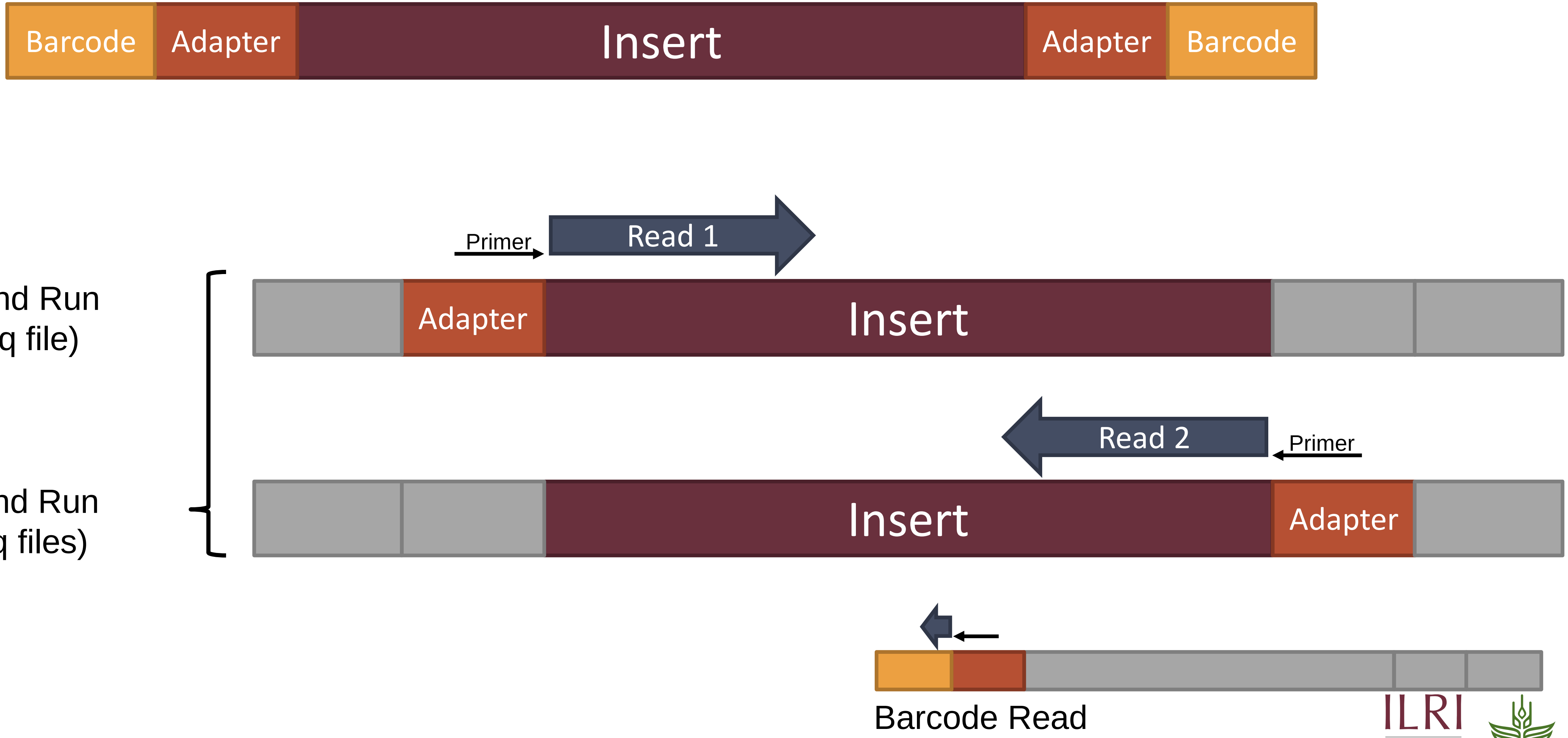
AFRICA CDC
Centres for Disease Control and Prevention
Safeguarding Africa's Health

ASLM
AFRICAN SOCIETY FOR LABORATORY MEDICINE

Overview

- Should be first step!
  - What your data look like
  - Uses tools such as FastQC & MultiQC
- Removes:
  - Low quality bases
  - Low complexity sequences
  - Adaptor sequences

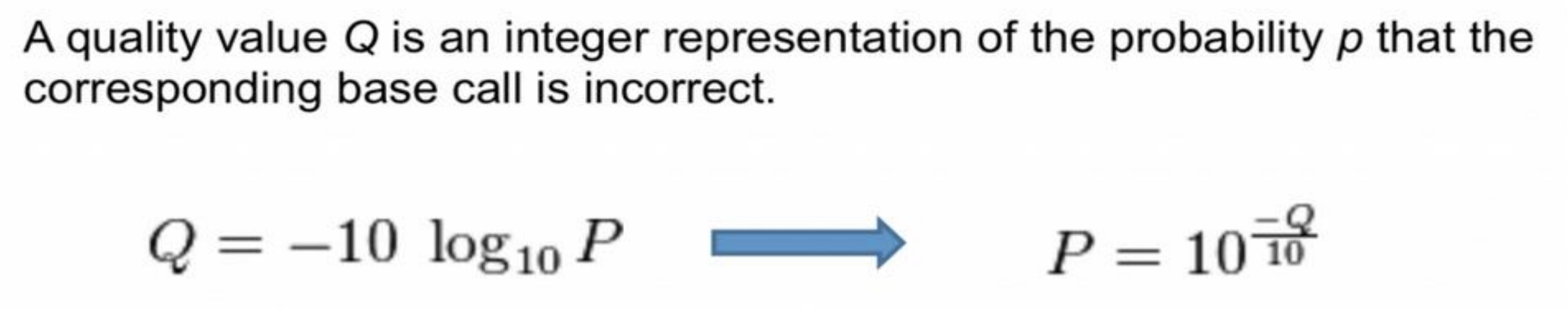
What Reads Do You Get



FastQ Format Data

```
@HWUSI-EAS611:34:6669YAAXX:1:1:5069:1159 1:N:0:
TCGATAATACCGTTTTTTTCCGTTTGATGTTGATACCATT
+
DF=DBD<BBFGGGGGGGBD@GGGD4@CA3CGG>DDD:D,B
@HWUSI-EAS611:34:6669YAAXX:1:1:5243:1158 1:N:0:
TATCTGTAGATTTCACAGACTCAAATGTAAATATGCAGAG
+
IIHIIHIIIIIIIIIIIIIIIIIIIIIIIHIIIIHIIIII
@HWUSI-EAS611:34:6669YAAXX:1:1:5266:1162 1:N:0:
GGAGGAAGTATCACTTCCTTGCCTGCCTCCTCTGGGGCCT
+
:GBGGGGGGGGGDGGDEDGGDGGGGDHHDHGHHGBGG:GG
```

Read Quality – Phred Score



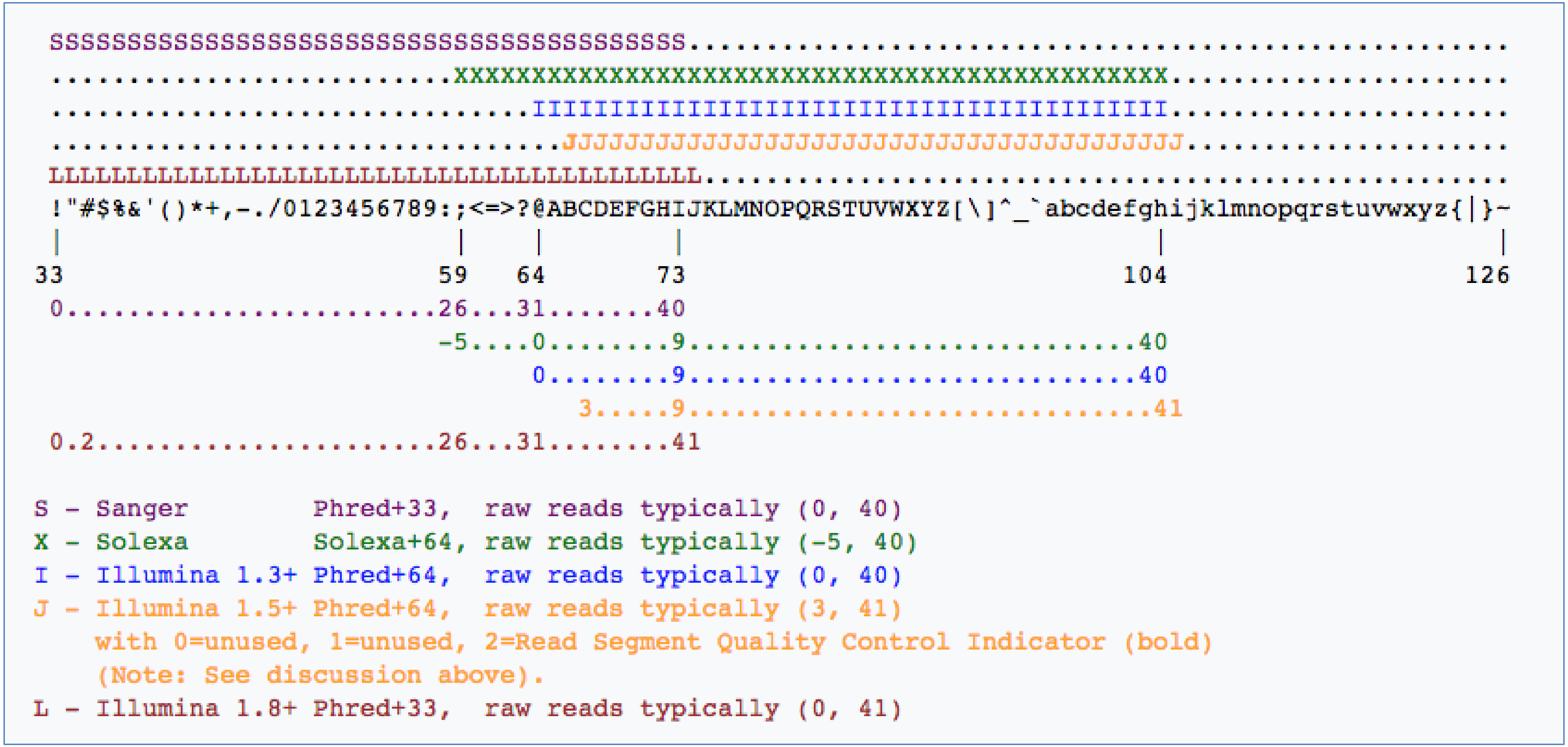
A quality value Q is an integer representation of the probability *p* that the corresponding base call is incorrect.

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

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 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

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

Different Phred Scores

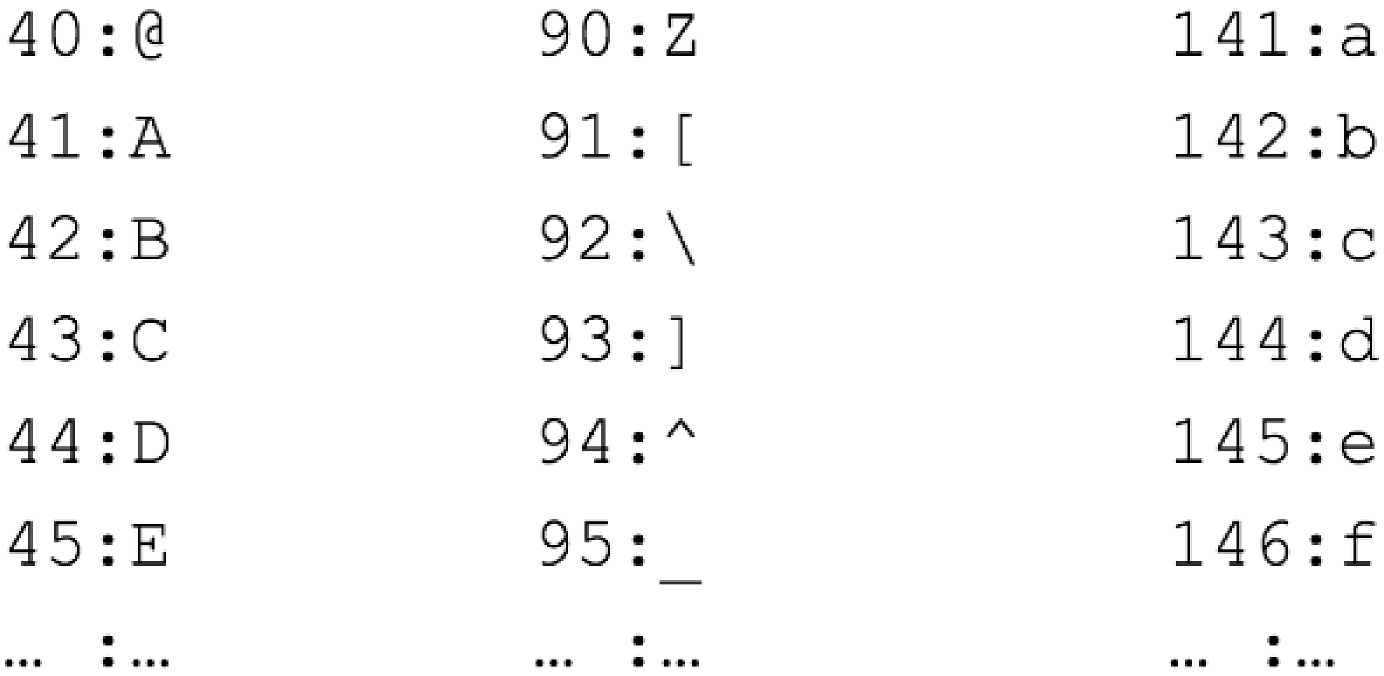


```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126
  0........................26...31.......40
                           -5....0........9.............................40
                                 0........9.............................40
                                    3.....9..............................41
  0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

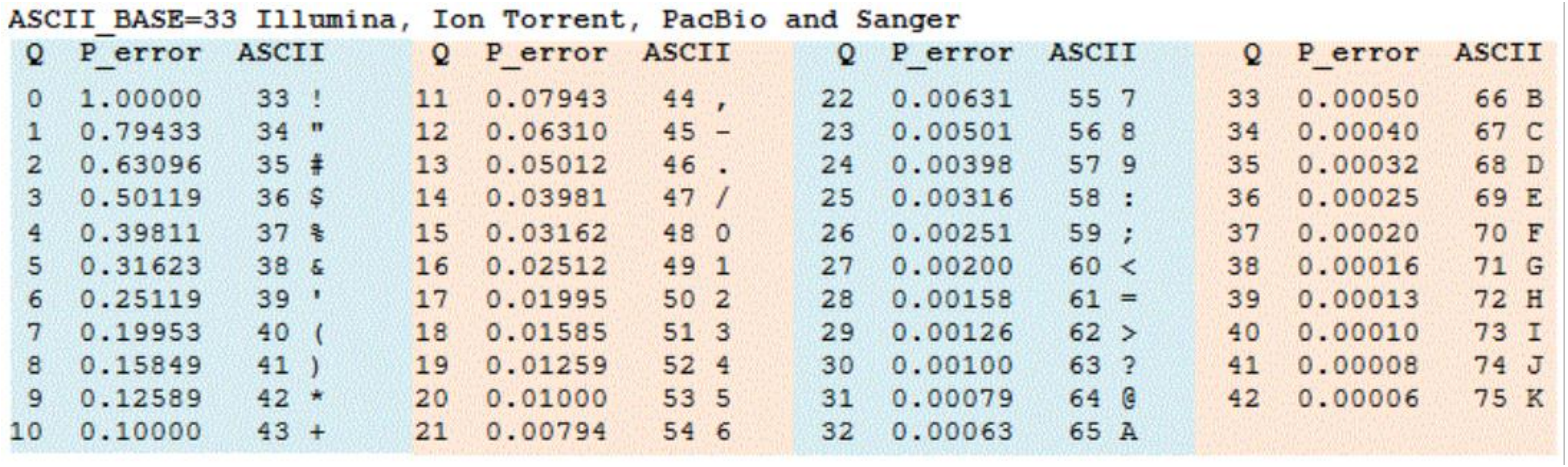
ASCII Encoding

- Each number is converted to one symbol:



```
40:@            90:Z              141:a
41:A            91:[              142:b
42:B            92:\              143:c
43:C            93:]              144:d
44:D            94:^              145:e
45:E            95:_              146:f
…  :…           …  :…             …  :…
```

ASCII Encoding: cont…

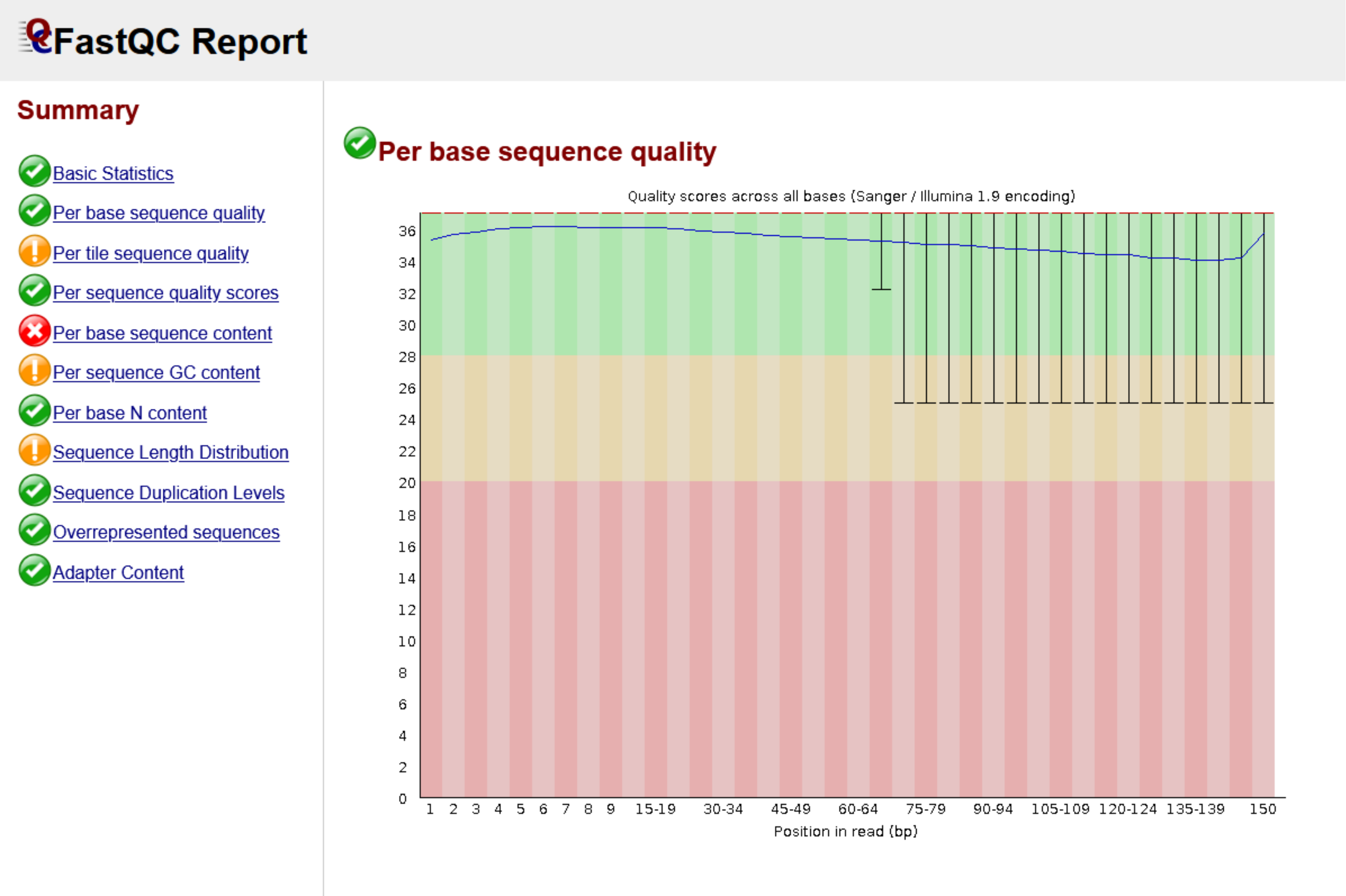


ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

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

Read Quality: FastQC



FastQC Report

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ! Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content

✓ **Per base sequence quality**

Reads raw fastq files

Performs multiple checks

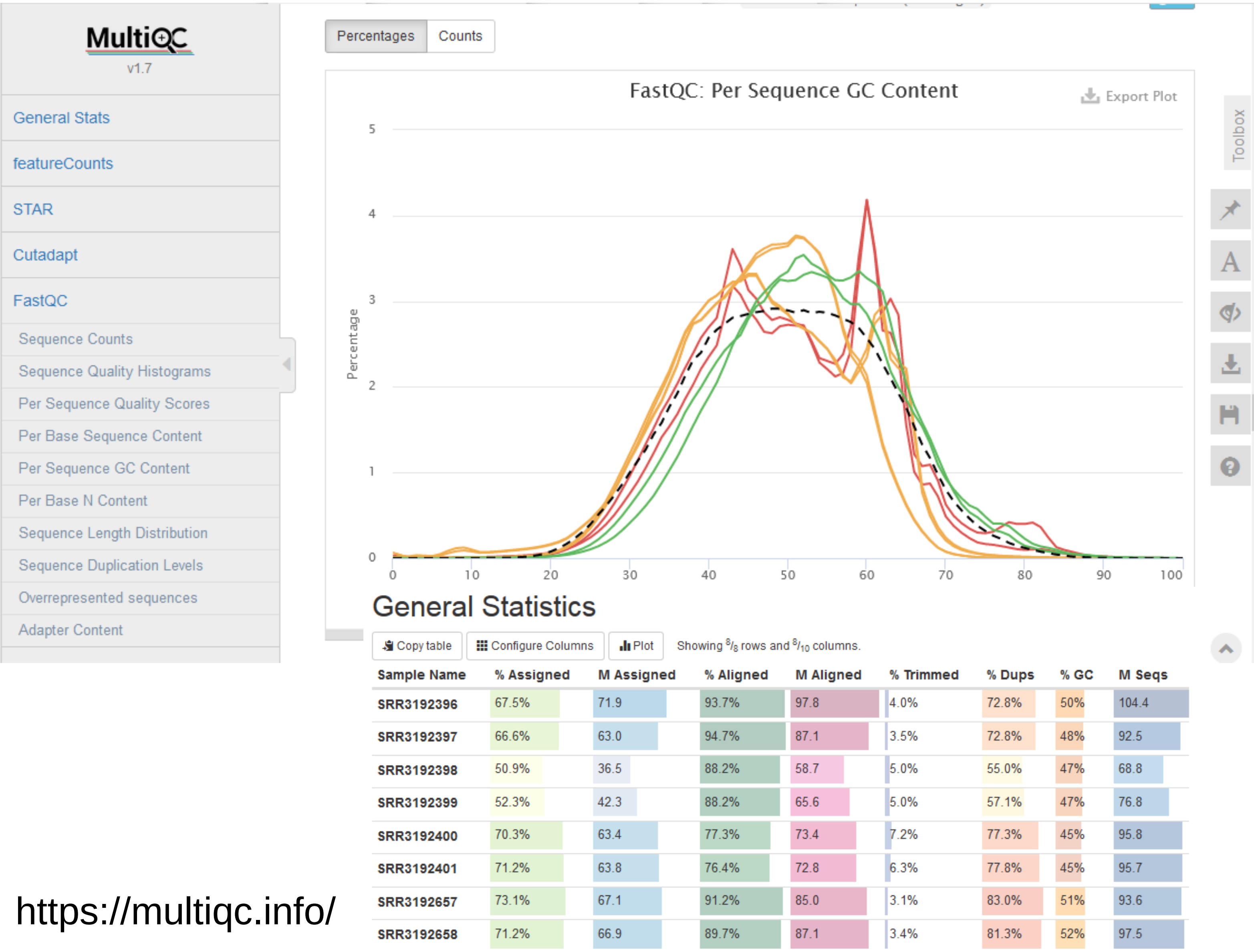
– Pass/warn/fail

– Compares to genomic library

HTML Report

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

Read Quality: MultiQC



General Statistics

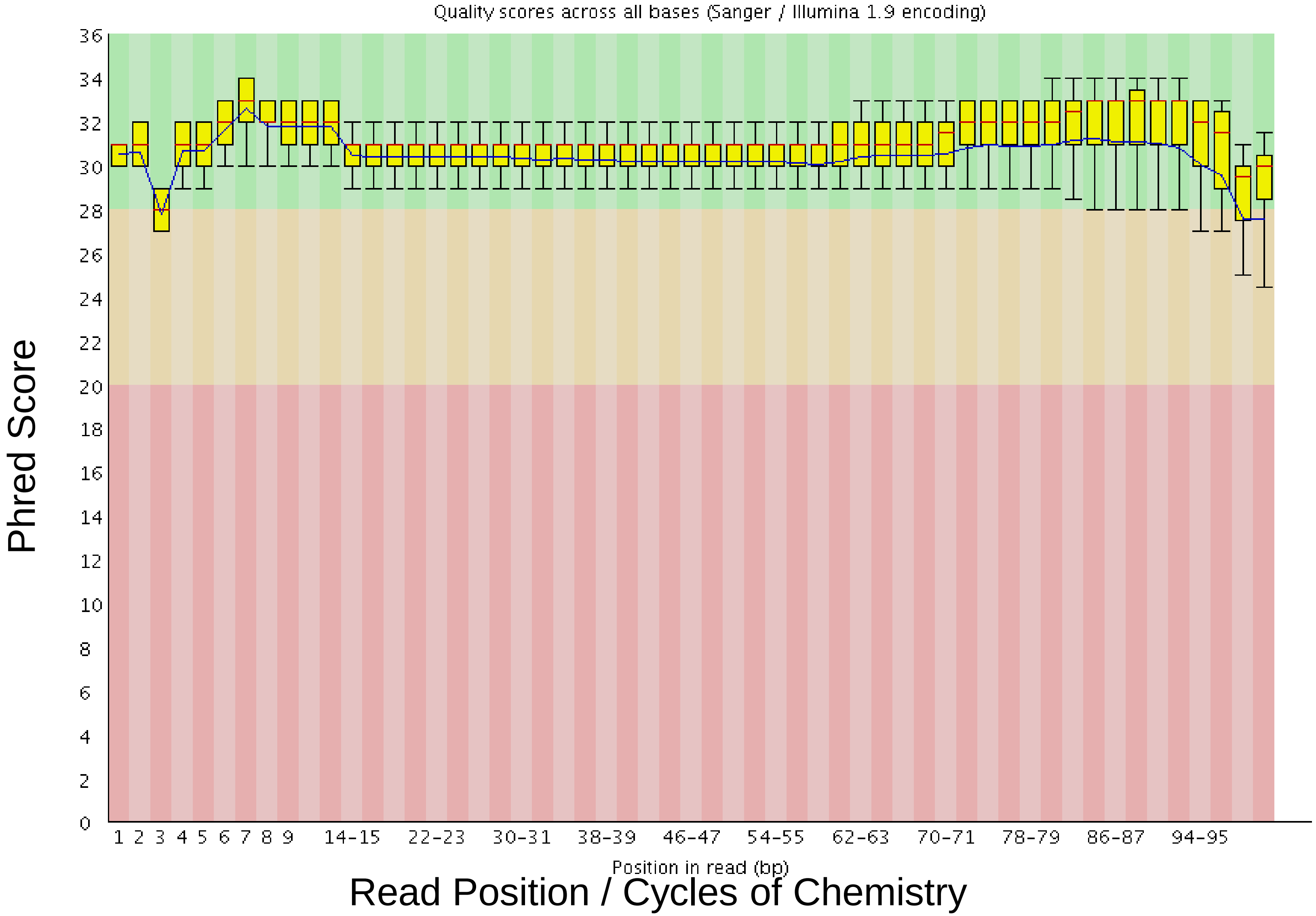
Showing $^8/_8$ rows and $^8/_{10}$ columns.

Sample Name	% Assigned	M Assigned	% Aligned	M Aligned	% Trimmed	% Dups	% GC	M Seqs
SRR3192396	67.5%	71.9	93.7%	97.8	4.0%	72.8%	50%	104.4
SRR3192397	66.6%	63.0	94.7%	87.1	3.5%	72.8%	48%	92.5
SRR3192398	50.9%	36.5	88.2%	58.7	5.0%	55.0%	47%	68.8
SRR3192399	52.3%	42.3	88.2%	65.6	5.0%	57.1%	47%	76.8
SRR3192400	70.3%	63.4	77.3%	73.4	7.2%	77.3%	45%	95.8
SRR3192401	71.2%	63.8	76.4%	72.8	6.3%	77.8%	45%	95.7
SRR3192657	73.1%	67.1	91.2%	85.0	3.1%	83.0%	51%	93.6
SRR3192658	71.2%	66.9	89.7%	87.1	3.4%	81.3%	52%	97.5

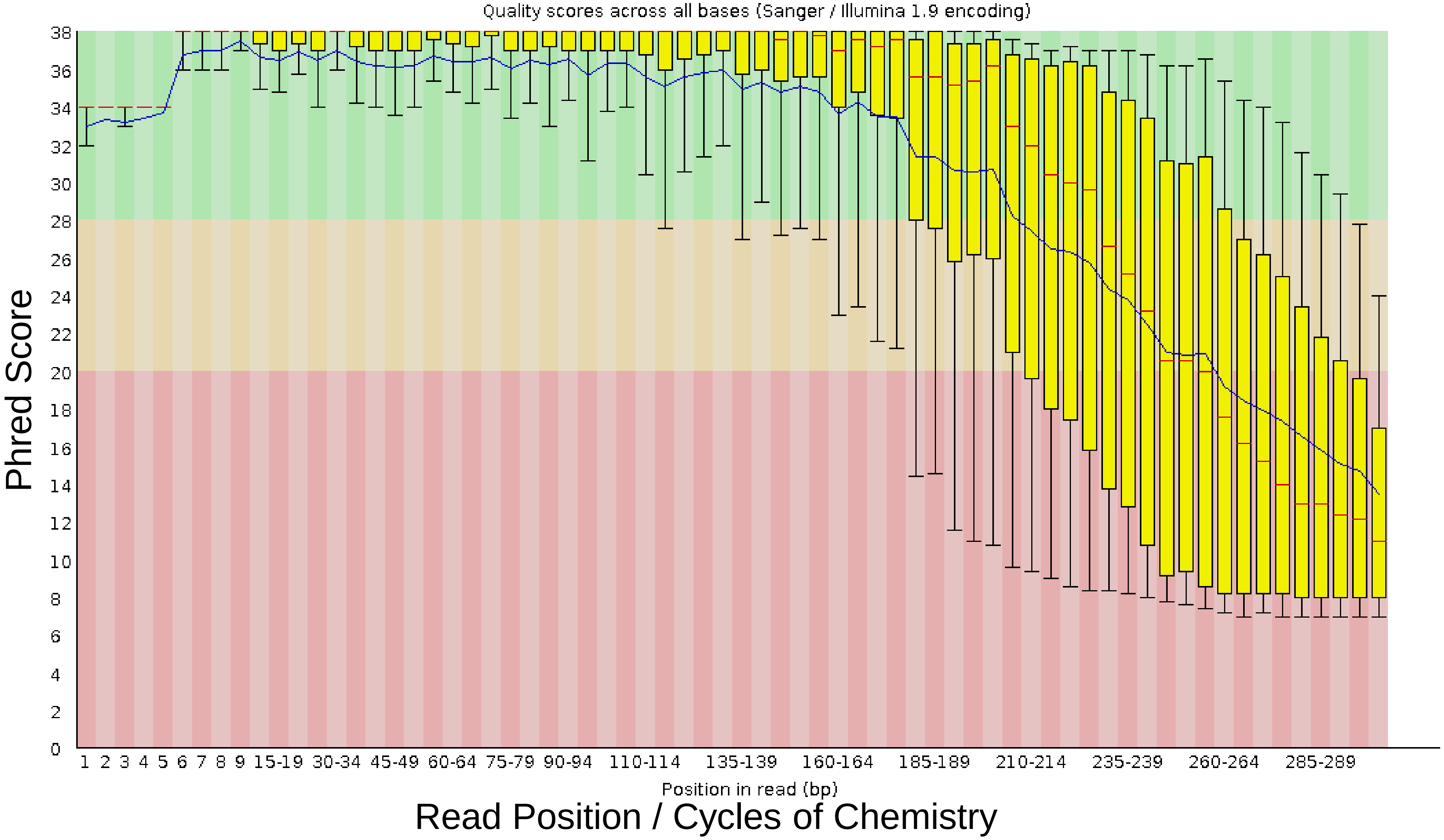
- Aggregates QC information from multiple samples
- Large number of programs supported
- Combined HTML report

https://multiqc.info/

Base Call Qualities – Per Cycle



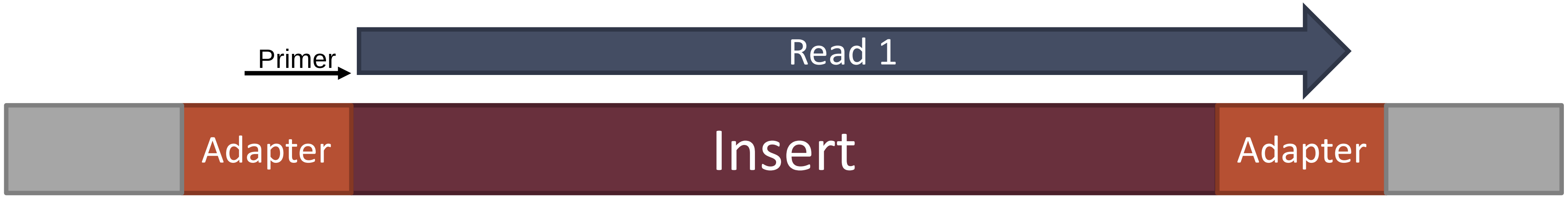
Base Call Qualities – Per Cycle



Clean-up options

Trimming 3’ end:

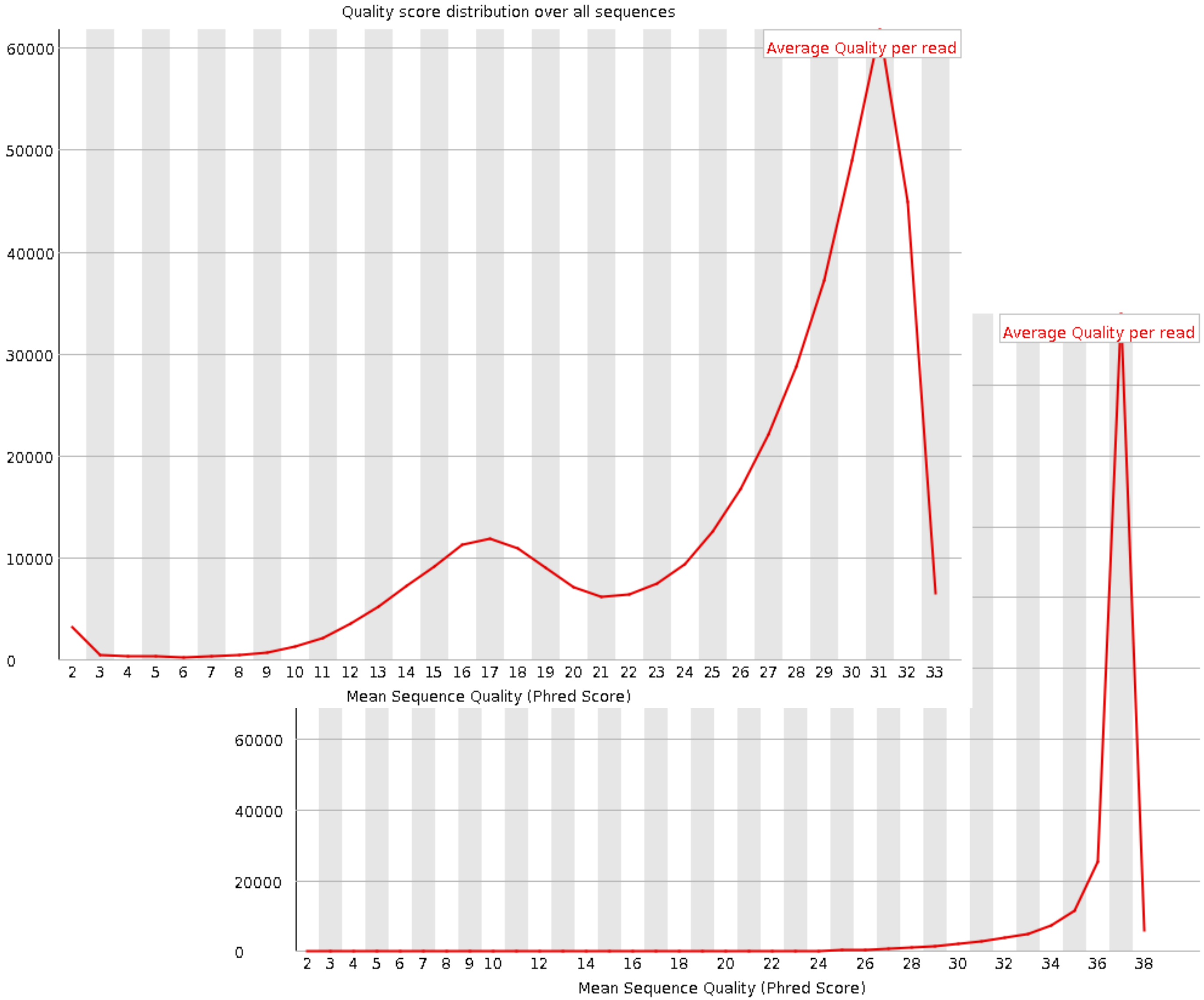
- Remove adapter read through
- Remove poor quality bases



Some quality issues may need to also remove specific reads

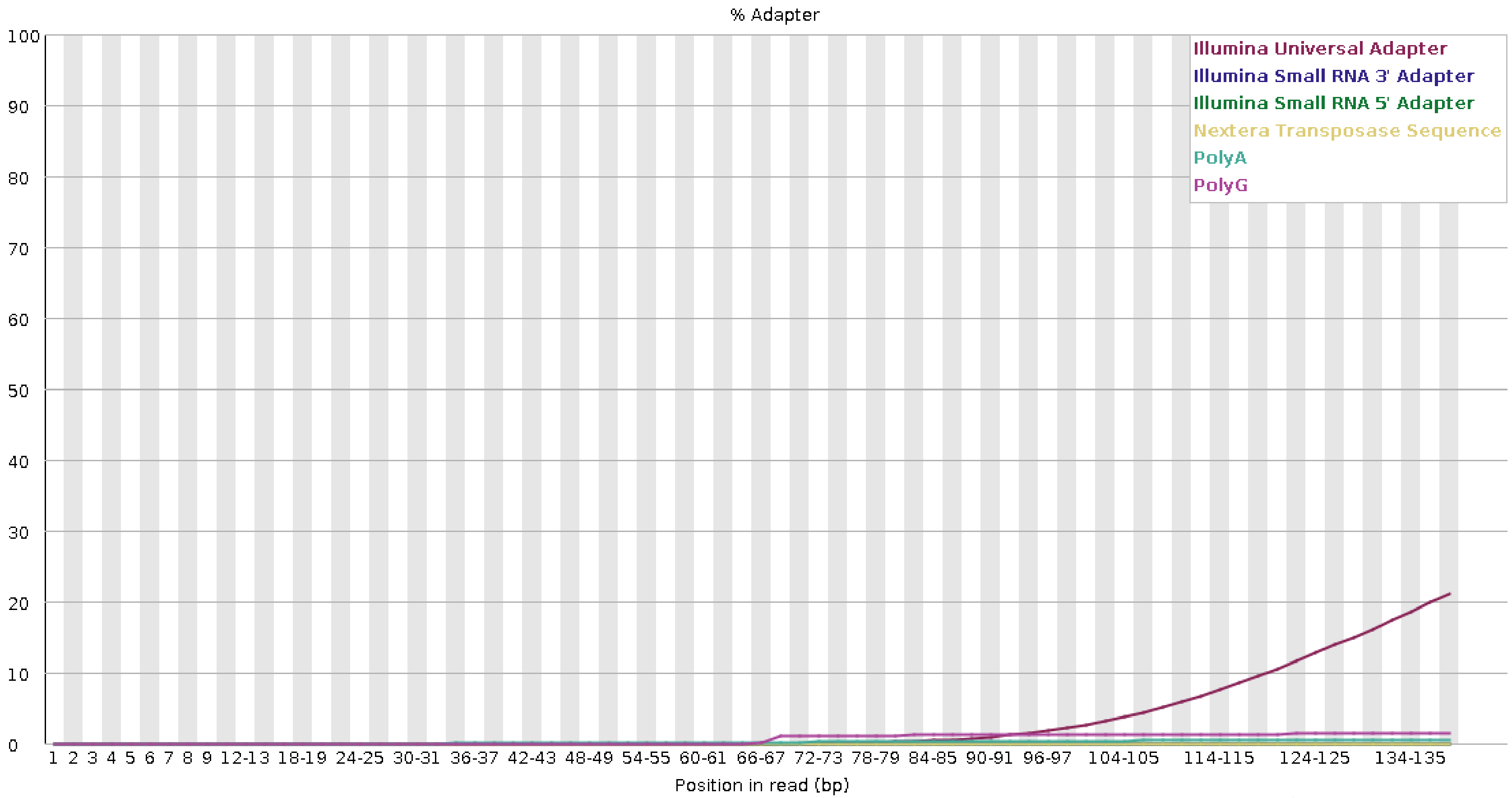
Despite issues may still be good enough for what is needed e.g. mapping

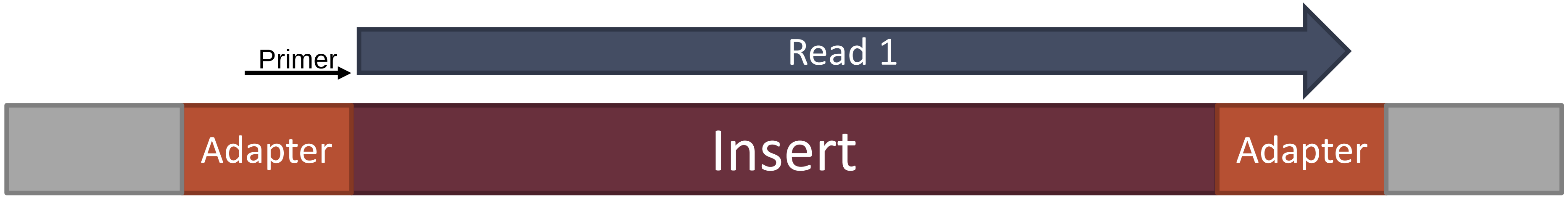
Per-Read Quality



- Are all reads equally affected?
- Is there a subset of reads which are always poor whilst others are good?

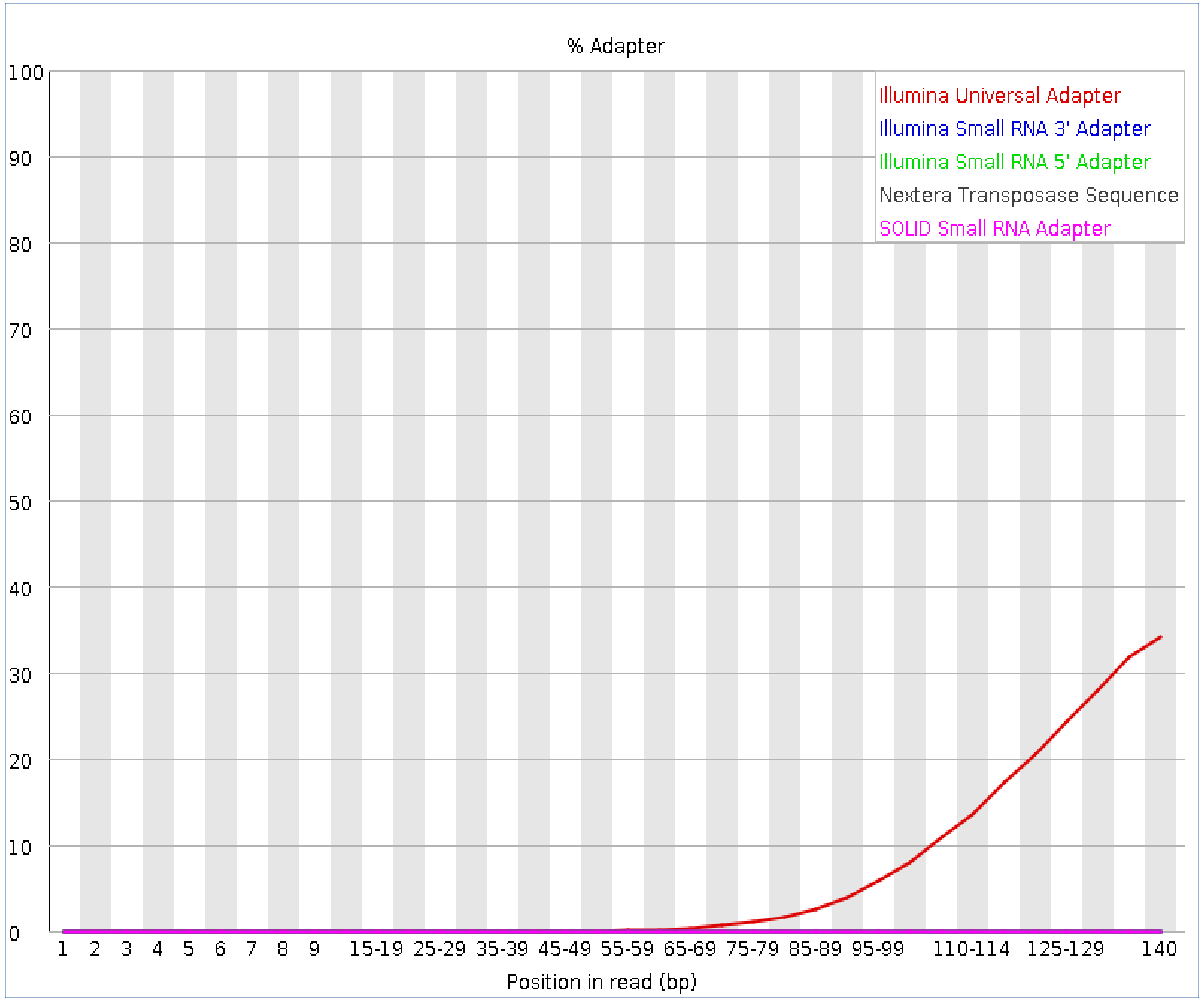
Measuring Read-though Adapters



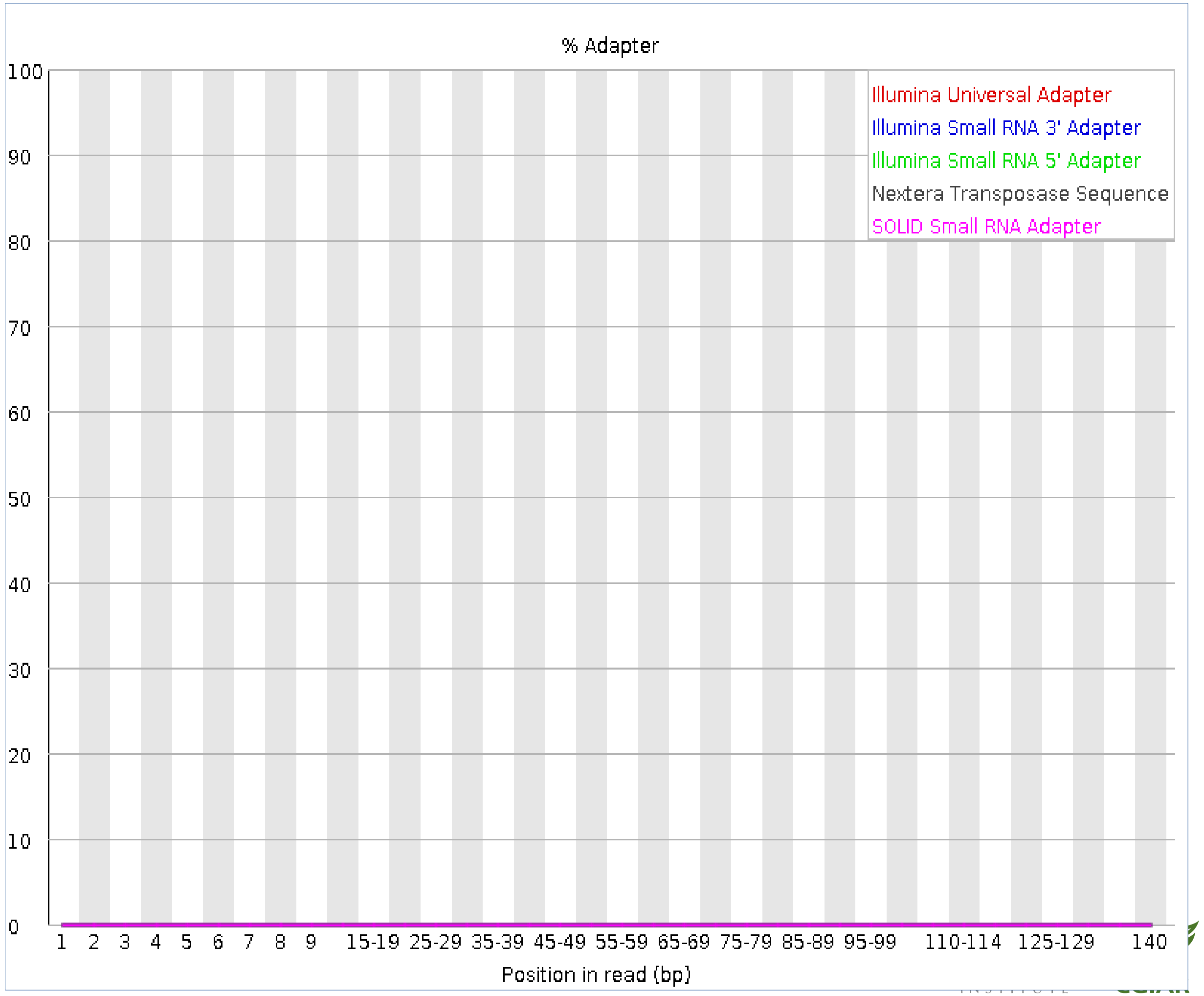


Adapter removal

Before Trimming



After Trimming



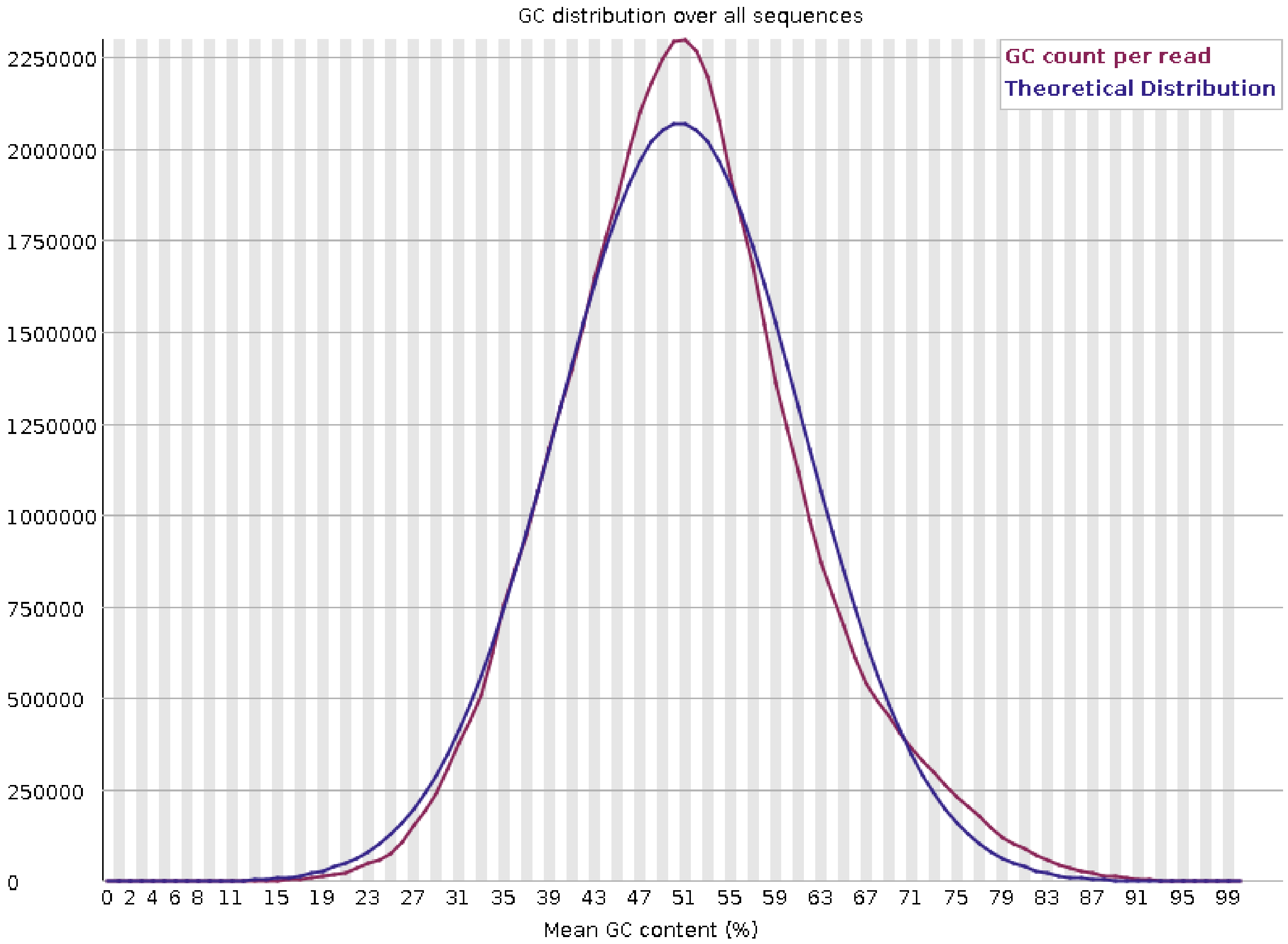
Library Dependent QC Metrics

Some QC metrics will be influenced by what you are sequencing

✓ ! ✗ Concern or Expected?

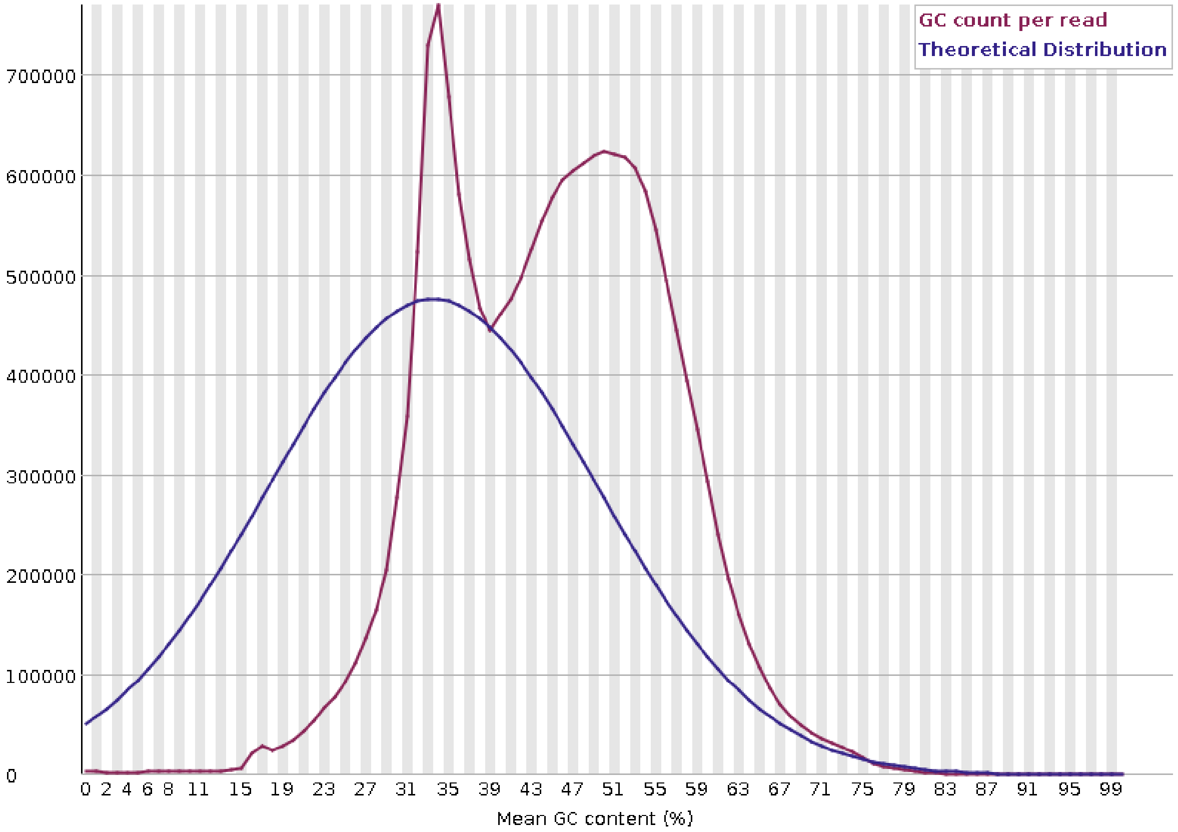
- GC Content
- Base Composition
- Duplication

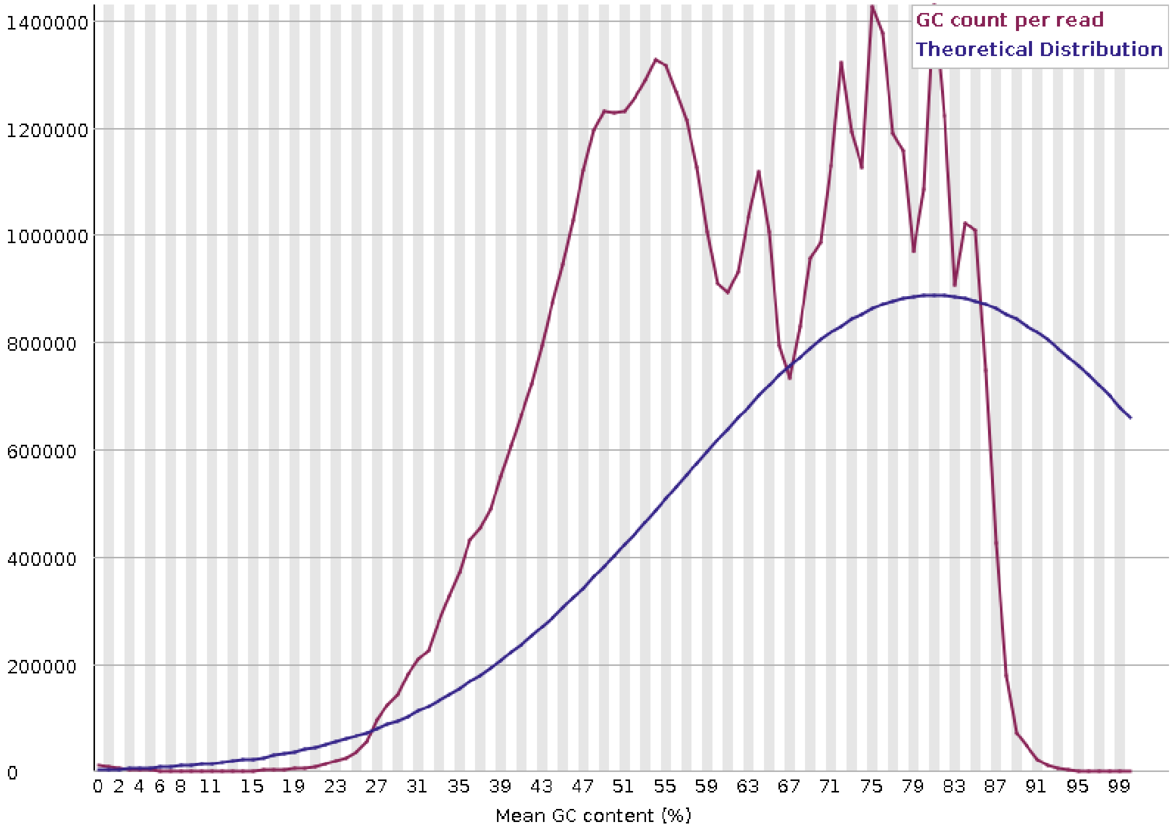
Library GC Content



- Generic summary of library composition at a read level
- Expect a normally distributed set of values centred on the overall GC content

GC Content: Cont…

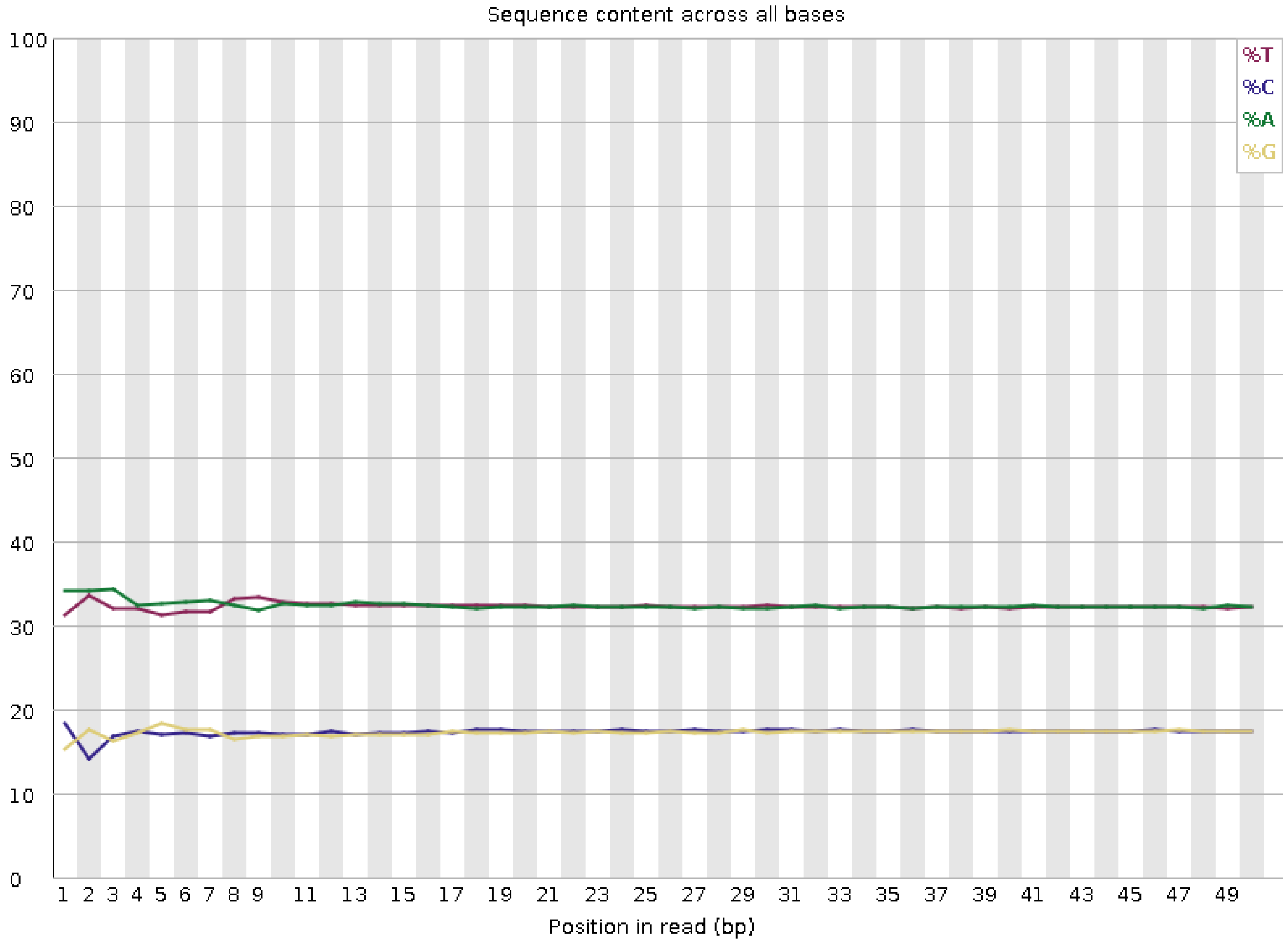




Specific Contamination with single sequence or closely related sequences

Artificial sequences, ribosomal RNA, contaminants

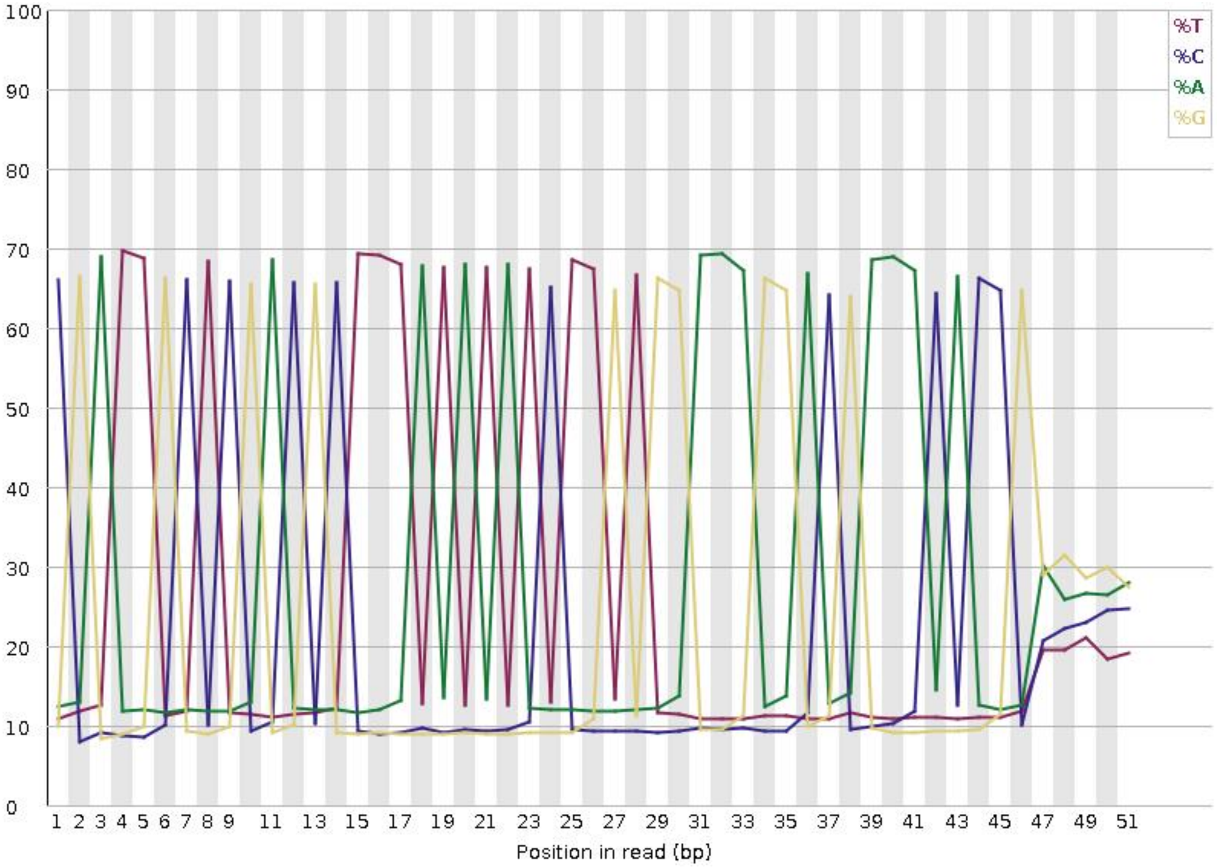
Library Base Composition



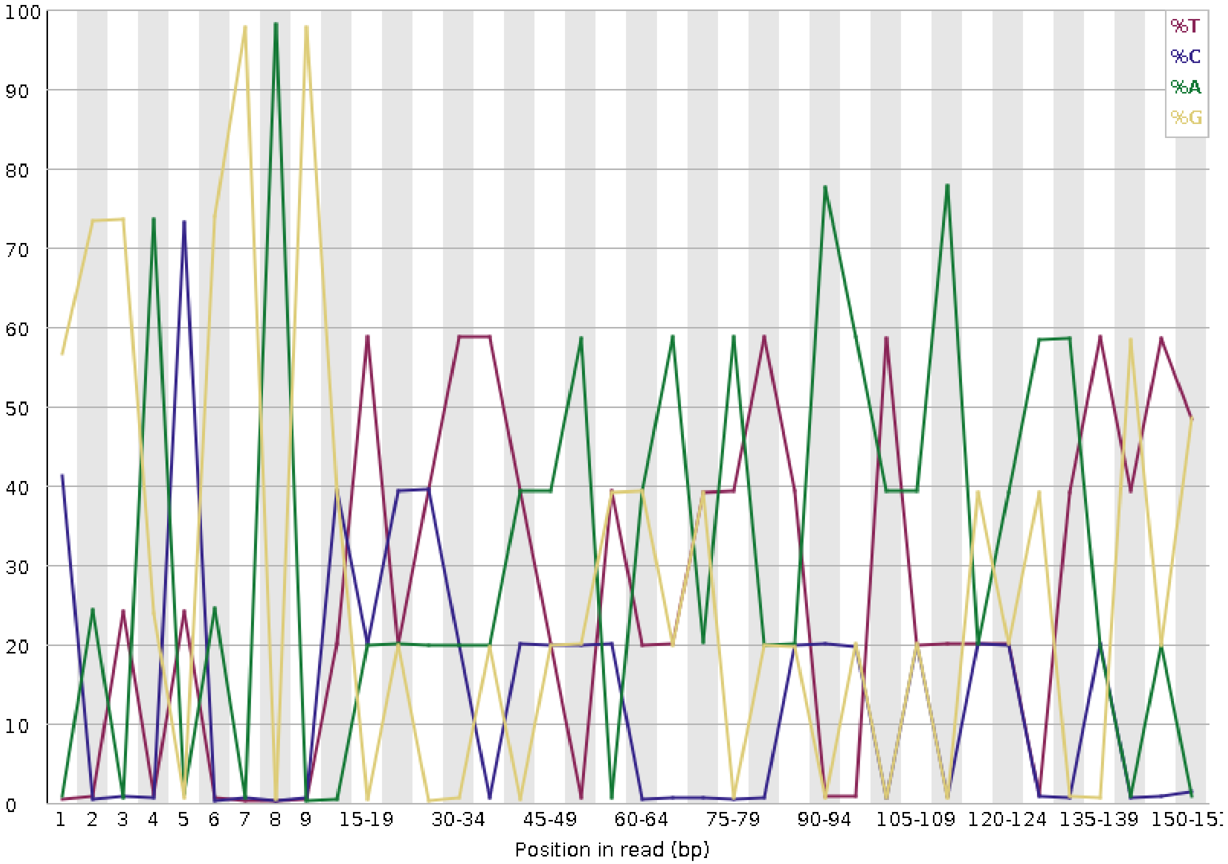
- For every chemistry cycle we can look at the number of ATGC we call
- For Libraries with random start positions the composition should be the same for all cycles

Bias Composition Throughout

Wrong Sequence

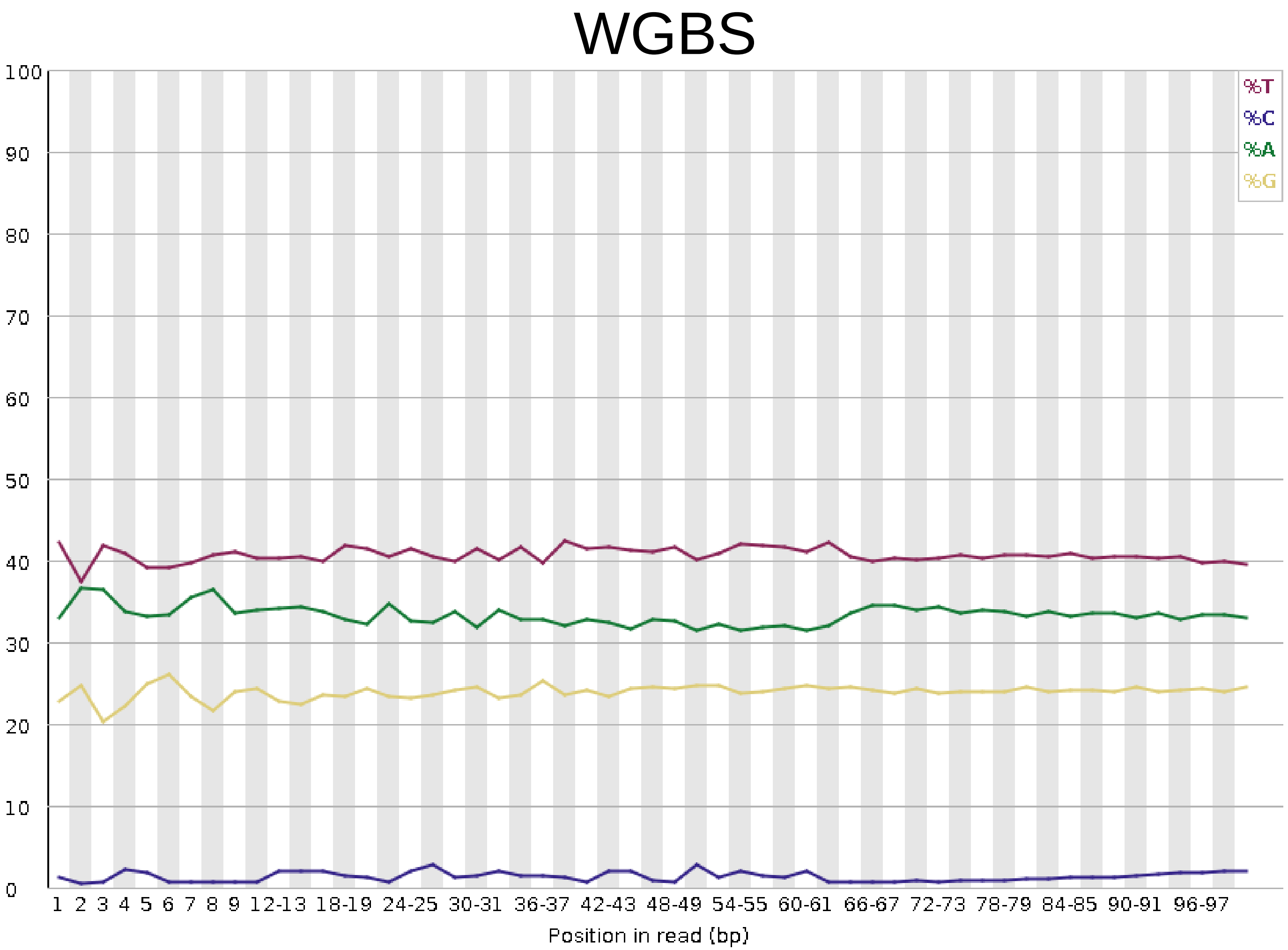


Amplicon



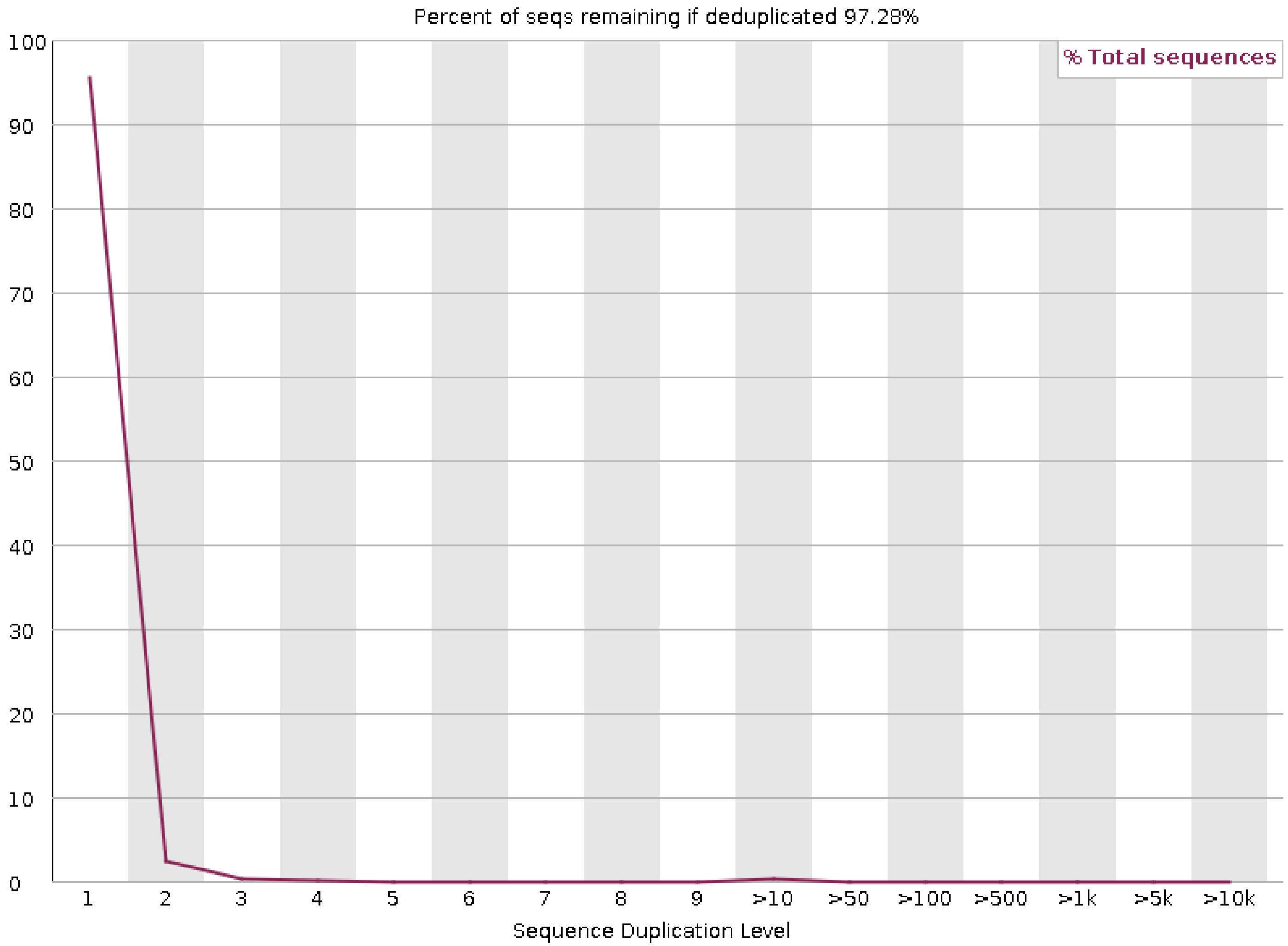
Proportional biases of bases at specific positions: Very low diversity

Bias Composition Throughout Cont…



Consistent disproportional expression of bases

Duplication

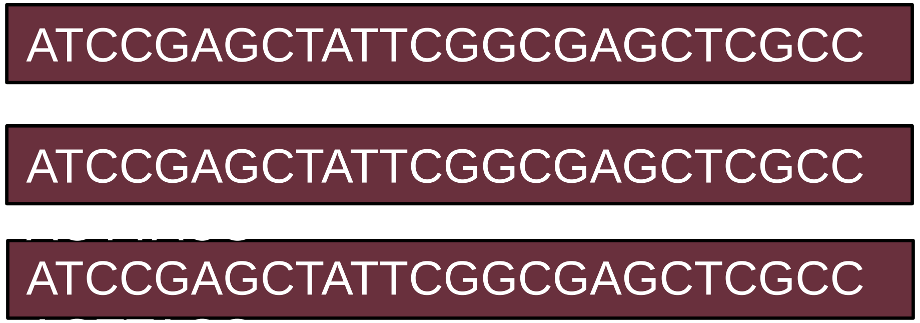


- How frequently the exact same sequence appears in your library
- For WGS expect most sequences to be unique

Duplication: Cont…

If the exact same sequence appears more than once it could be…

Technical:



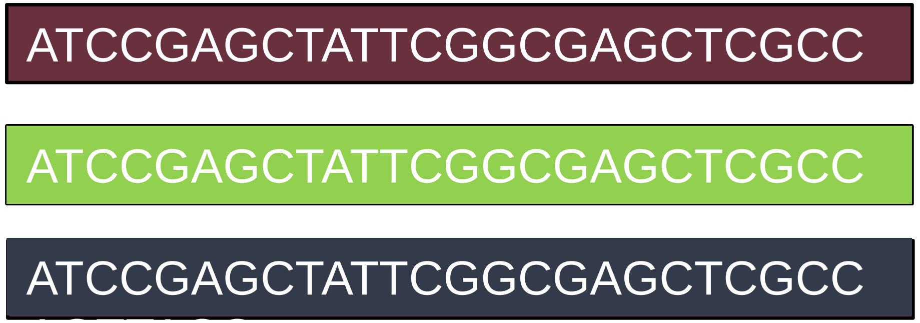
ATCCGAGCTATTCGGCGAGCTCGCC

ATCCGAGCTATTCGGCGAGCTCGCC

ATCCGAGCTATTCGGCGAGCTCGCC

- PCR duplicates

Coincidental:



ATCCGAGCTATTCGGCGAGCTCGCC

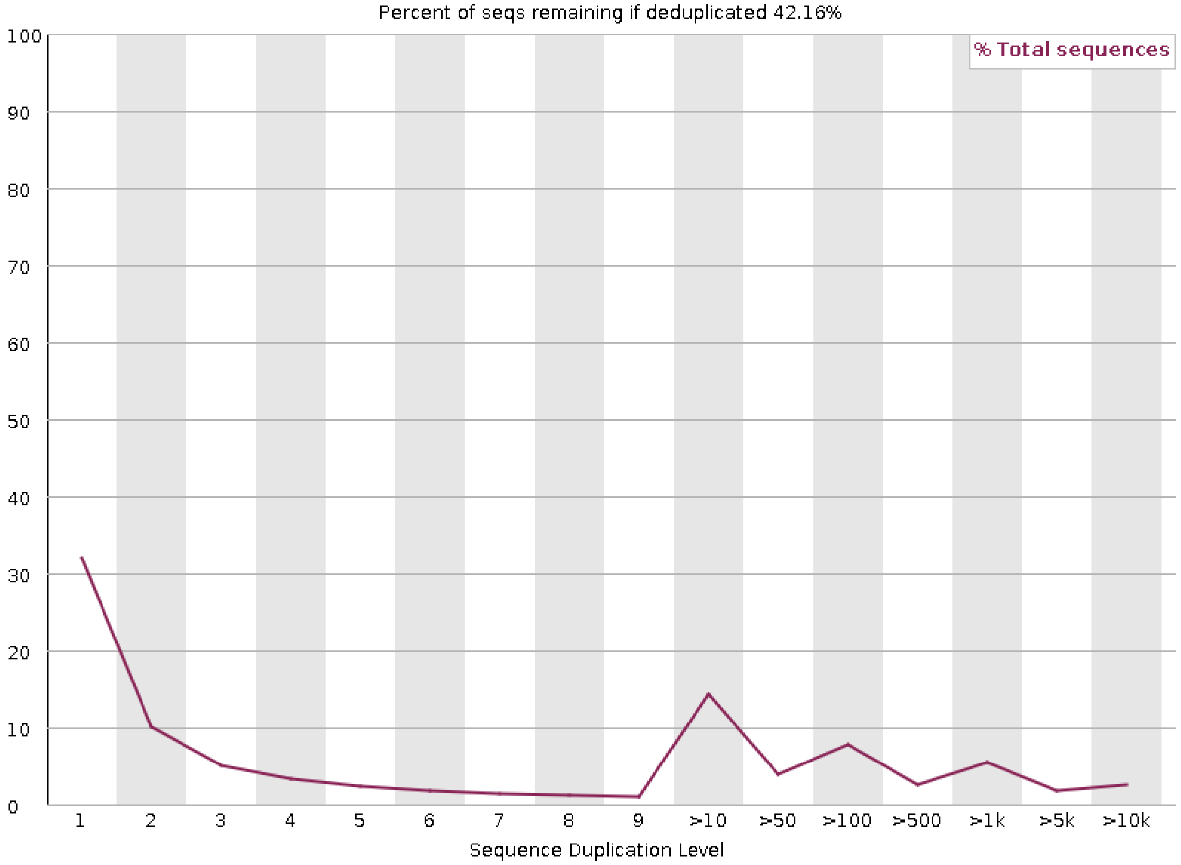
ATCCGAGCTATTCGGCGAGCTCGCC

ATCCGAGCTATTCGGCGAGCTCGCC

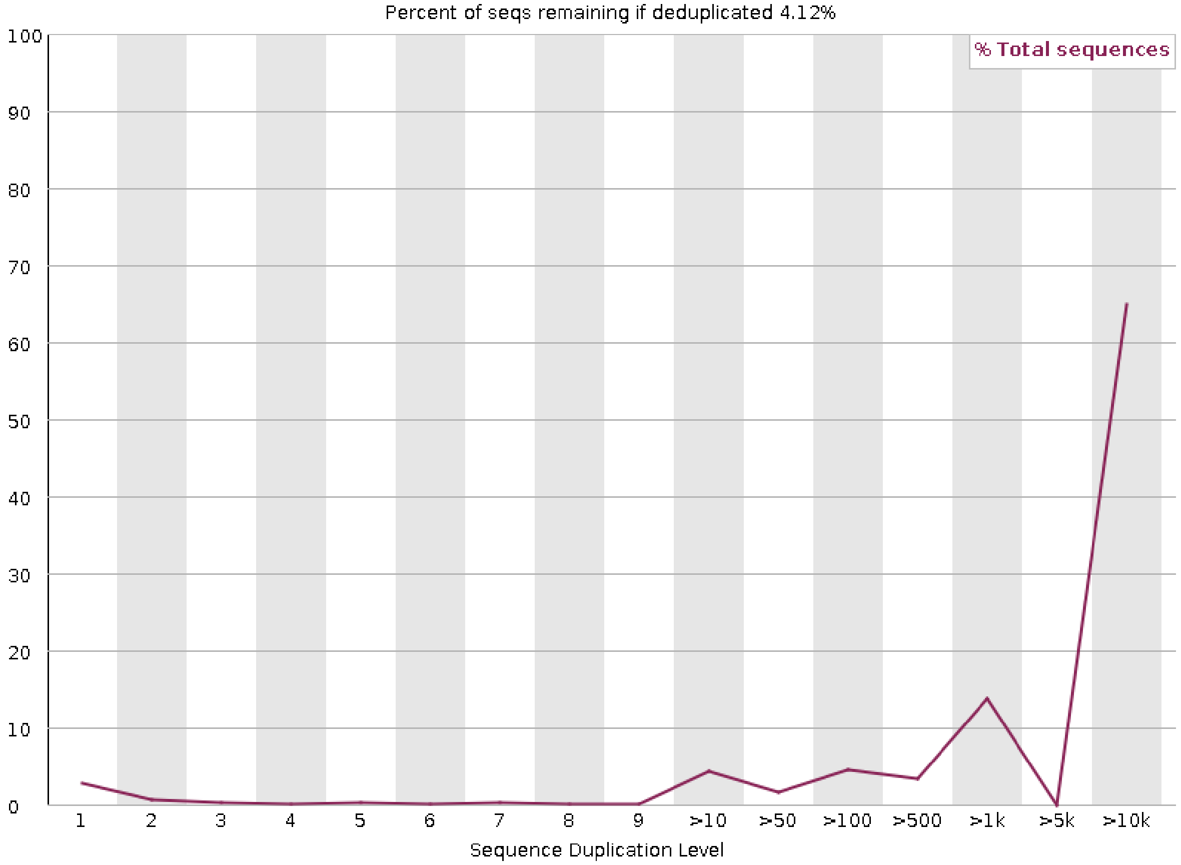
- Deep sequencing
- Highly present sequences
- Restricted diversity libraries

Duplication: Cont…

RNA-Seq



Amplicon

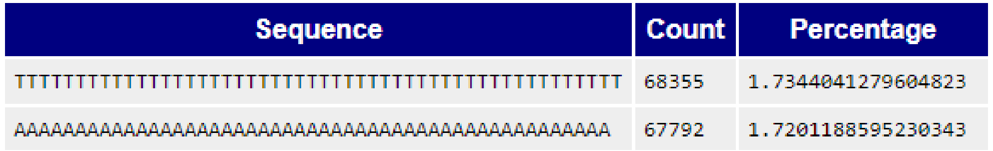


Overrepresented Sequences

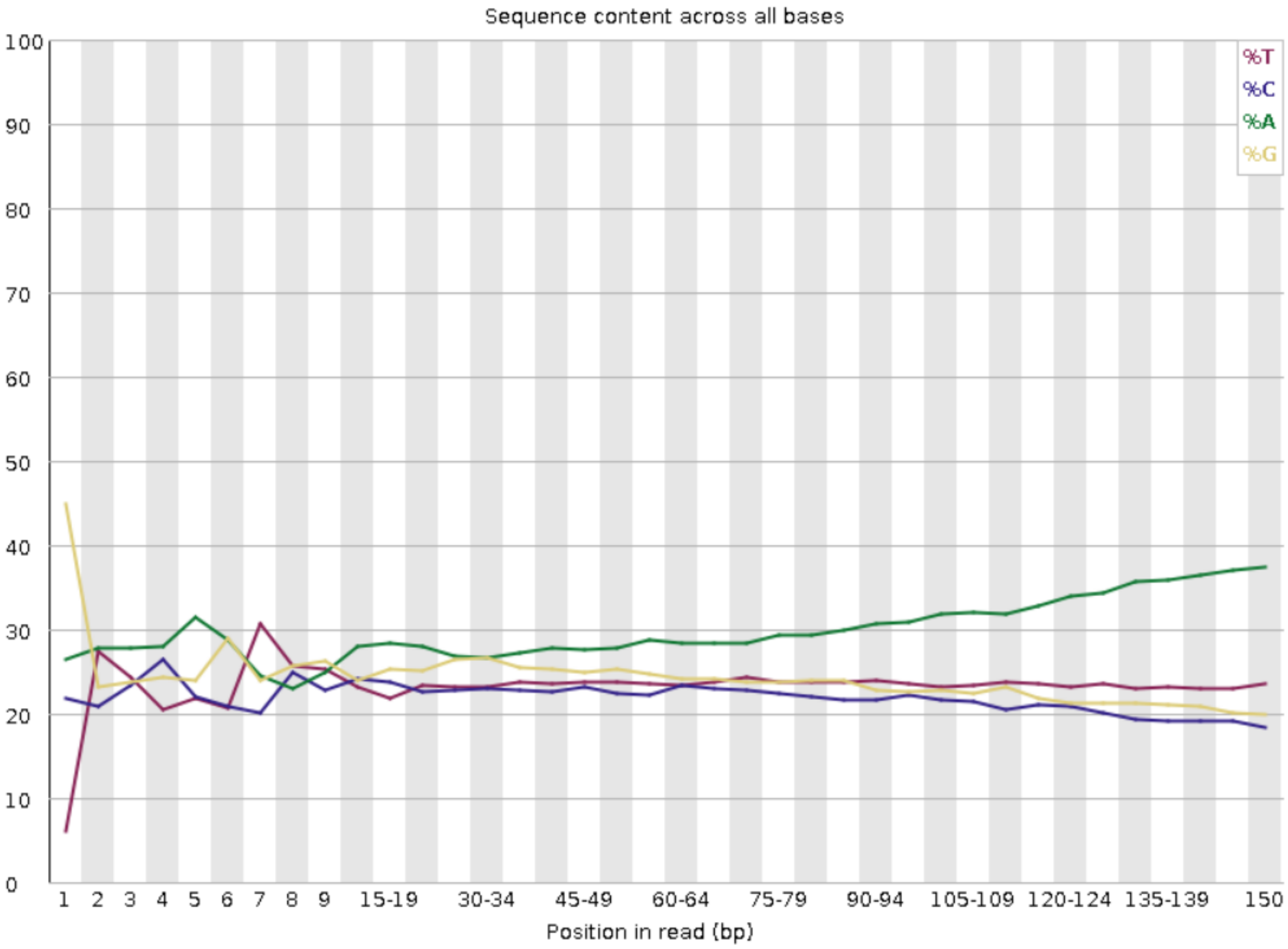
- Extreme duplication
- The exact same sequence is a significant proportion of the whole library (which might not be duplicated overall)
  - Poly Sequences
  - Specific Sequences

Poly Sequences

PolyA (or PolyT) – Common in RNA-Seq

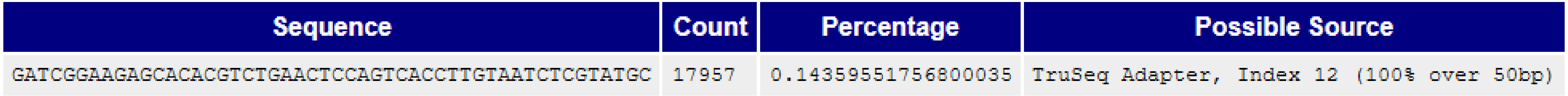


Sequence	Count	Percentage
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	68355	1.734404127960482 3
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	67792	1.720118859523034 3



Overrepresented Specific Sequences

- Normally artificial sequences (primers, adapters, vectors etc)
- Can search a database of known sequences to find matches



Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGC	17957	0.14359551756800035	TruSeq Adapter, Index 12 (100% over 50bp)

Acknowledgments

ILRI | CGIAR

ASLM
AFRICAN SOCIETY FOR LABORATORY MEDICINE

AFRICA CDC
Centres for Disease Control and Prevention
Safeguarding Africa's Health