QB#9468

FASTQ Processing Tools for Data Analysis

Zhongfeng Wang
TAS | 03/26/2020

Outline of the webinar

Overview of Illumina FASTQ generation

- Data analysis pipeline overview
- FASTQ format
- Demultiplexing and FASTQ generation

FASTQ tools

- Adapter trimming
- Quality trimming
- Read merging

Does my run look good vs does my FASTQ look good?

Illumina metrics vs third party QC tools

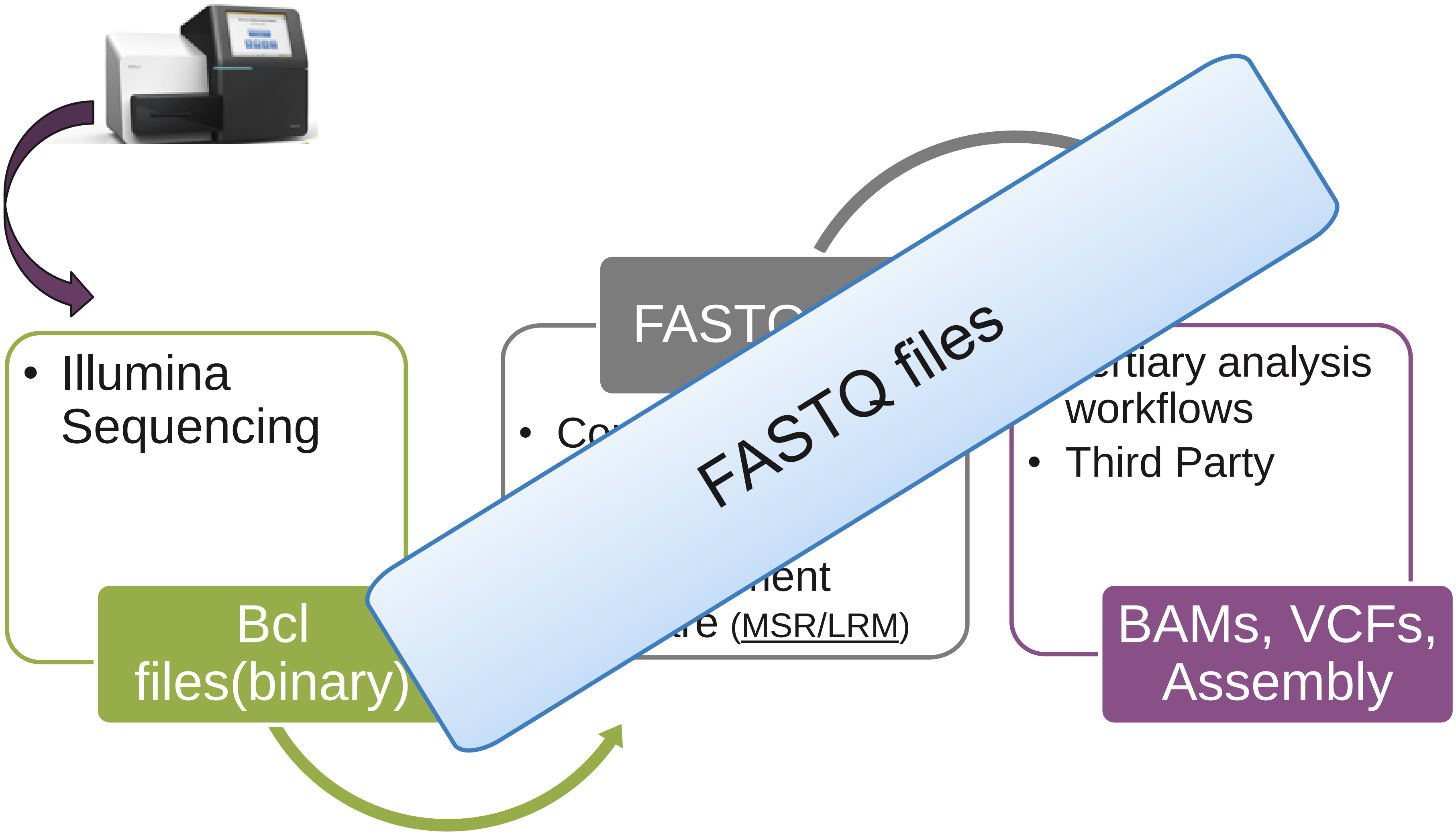
Intended audience

- **Users who are new to Illumina sequencing platforms**
  - End users who use a core facility to sequence their data and need some guidance on data optimization

- **Not intended for bioinformaticians**

Overview of Illumina FASTQ Generation

Pipeline for data analysis



- Illumina Sequencing

Bcl files(binary)

FASTC

- Co

ment

re (MSR/LRM)

rtiary analysis workflows
- Third Party

BAMs, VCFs, Assembly

FASTQ files

For Research Use Only. Not for use in diagnostic procedures.

BCL Files

- **BCL = Base CalL**
  - The “raw” data
  - Binary file containing base calls and quality scores for each tile for each cycle
  - Produced by on-instrument Real-Time Analysis Software (RTA)

FASTQ Files-format

- **Text file that contains the sequence data from clusters**

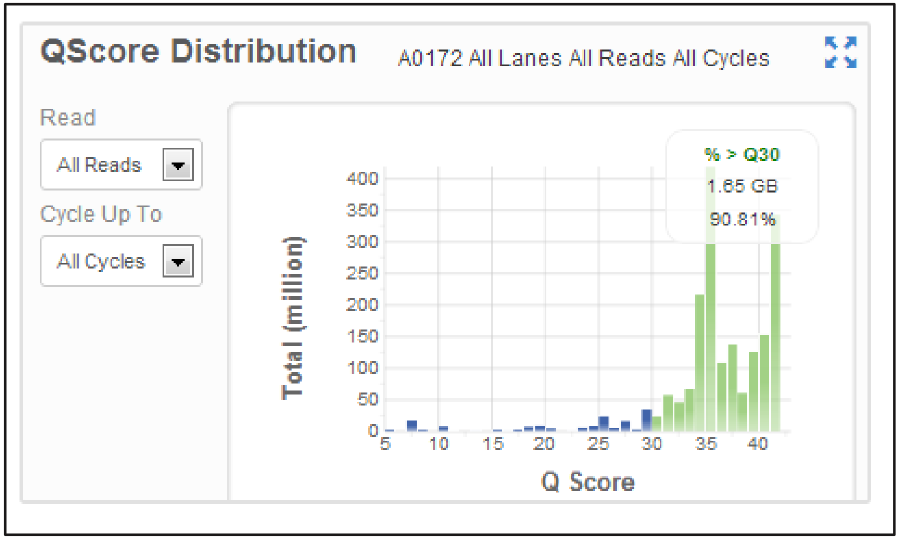
- **Each entry consists of 4 lines:**

1. Sequence identifier with information about the sequencing run and the cluster.
   i. Header format = @<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<xpos>:<y-pos>:<UMI> <read>:<is filtered>:<control number>:<index>
2. The sequence (the base calls; A, C, T, G and N).
3. A separator, which is simply a plus (+) sign.
4. The base call quality scores. Phred +33 encoded (ASCII)

```
@NB987655:1:ABCD12345:1:11401:6329:1045 1:N:0:ATCACG
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

Qscore – Quality Score

- The QScore is based on the Phred scale – Chance that a base call is incorrectly called
  - Q10: 10% chance of incorrect base call
  - Q20: 1% chance of incorrect base call
  - Q30: 0.1% chance of incorrect base call
  - Q40: 0.01% chance of incorrect base call
- Quality Scores for Next-Generation Sequencing



QScore Distribution A0172 All Lanes All Reads All Cycles

Read: All Reads

Cycle Up To: All Cycles

% > Q30
1.65 GB
90.81%

Symbol	ASCII Code	Q-Score
<	60	27
=	61	28
>	62	29
?	63	30
@	64	31
A	65	32
B	66	33
C	67	34
D	68	35
E	69	36
F	70	37

FASTQ Files- How to view

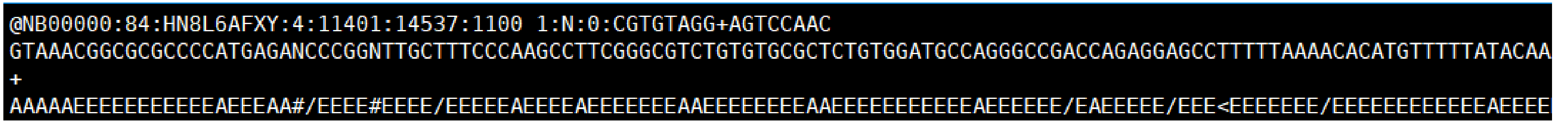
FASTQ files are large (file size) – best to not open these

How to view contents without opening?

- Use a terminal window on Linux or Mac
- zcat *.fastq.gz | less
- zcat *.fastq.gz | head

How to open a FASTQ file?

- gzipped
- 7-zip for windows to unzip
- Open using a notepad



```
@NB00000:84:HN8L6AFXY:4:11401:14537:1100 1:N:0:CGTGTAGG+AGTCCAAC
GTAAACGGCGCGCCCCATGAGANCCCGGNTTGCTTTCCCAAGCCTTCGGGCGTCTGTGTGCGCTCTGTGGATGCCAGGGCCGACCAGAGGAGCCTTTTTAAAACACATGTTTTTATACAA
+
AAAAAEEEEEEEEEEEAEEEAA#/EEEE#EEEE/EEEEEAEEEEAEEEEEEEAAEEEEEEEEAAEEEEEEEEEEEAEEEEEE/EAEEEEE/EEE<EEEEEEE/EEEEEEEEEEEEAEEEE
```

Anatomy of a FASTQ file

- **Naming**

  SampleName_SampleNumber_Lane_Read_FlowCellIndex.fastq.gz

  SampleName_S1_L001_R1_001.fastq.gz

  SampleName_S1_L001_R2_001.fastq.gz

- **e.g. NextSeq500/550, PE sequencing. 8 fastq files for each sample**

  SampleName_S1_L001_R1_001.fastq.gz

  SampleName_S1_L002_R1_001.fastq.gz

  SampleName_S1_L003_R1_001.fastq.gz

  SampleName_S1_L004_R1_001.fastq.gz

  SampleName_S1_L001_R2_001.fastq.gz

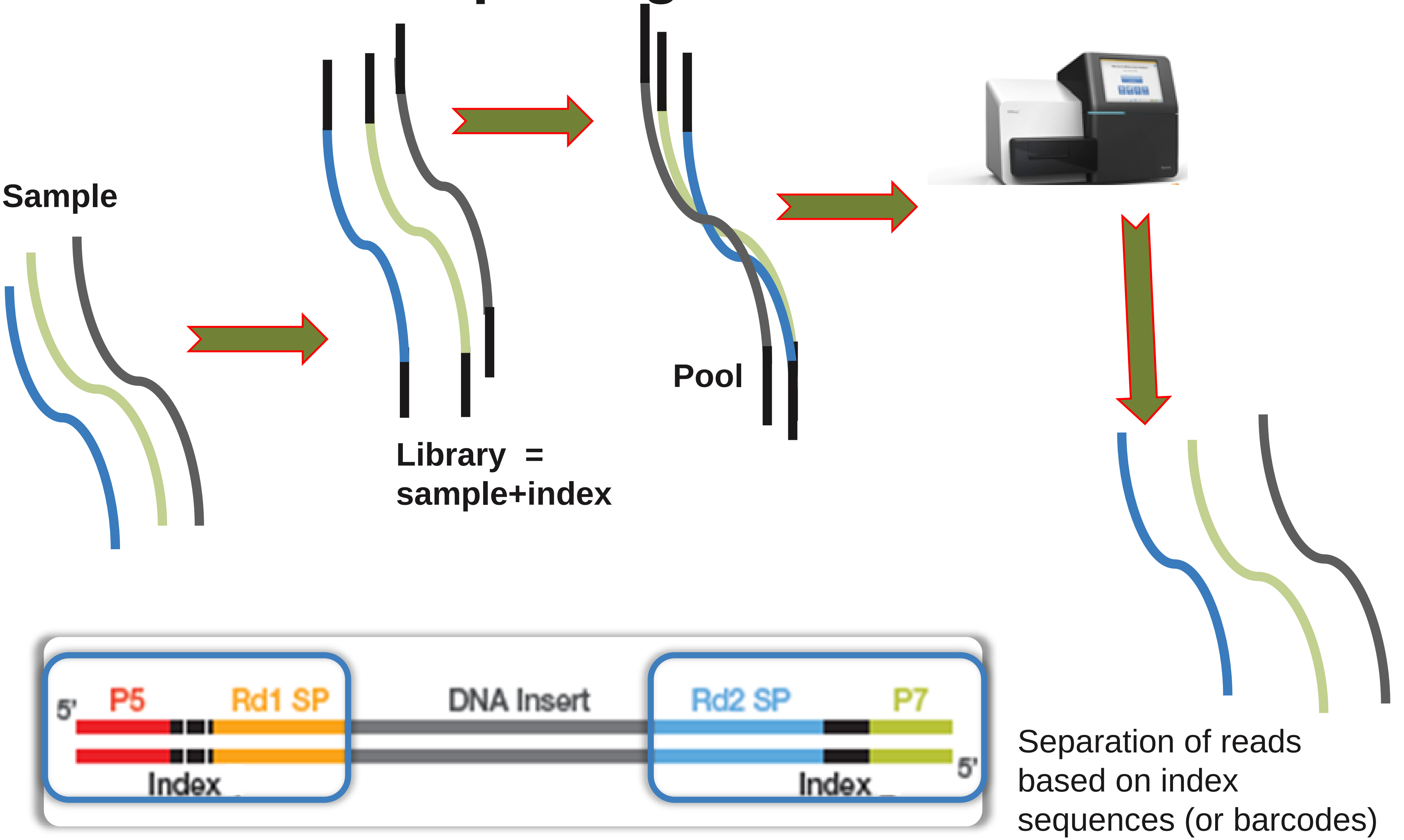
  SampleName_S1_L002_R2_001.fastq.gz

  SampleName_S1_L003_R2_001.fastq.gz

  SampleName_S1_L004_R2_001.fastq.gz

Demultiplexing and FASTQ generation

What is demultiplexing?



Sample

Library = sample+index

Pool

Separation of reads based on index sequences (or barcodes) using bioinformatics

For Research Use Only. Not for use in diagnostic procedures.

Illumina Tools for Demultiplexing and Fastq Generation

On-Board

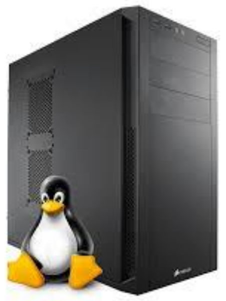
MiSeq™ Reporter

Local Run Manager

Cloud based

BaseSpace™ Sequence Hub (BSSH)

Linux based



Bcl2fastq

DRAGEN ™ bclconvert

Supports All Illumina Sequencing Systems

Bcl2fastq v2 is required for platforms using RTA 2 & 3.

Note: For sequencing systems running an earlier software version of Real-Time Analysis (RTA) than v1.18.54, use bcl2fastq v1.8.4.

For Research Use Only. Not for use in diagnostic procedures.

What is bcl2fastq v2?

- **Linux-based software to convert BCLs → FASTQs**
- **Optionally performs demultiplexing per sample sheet (created with IEM) instructions**
- **Command line example**

```
/usr/local/bin/bcl2fastq --runfolder-dir . --output-directory
./Data/Intensities/BaseCalls/  --input-dir
./Data/Intesities/Basecalls  --sample-sheet ./SampleSheet.csv
```

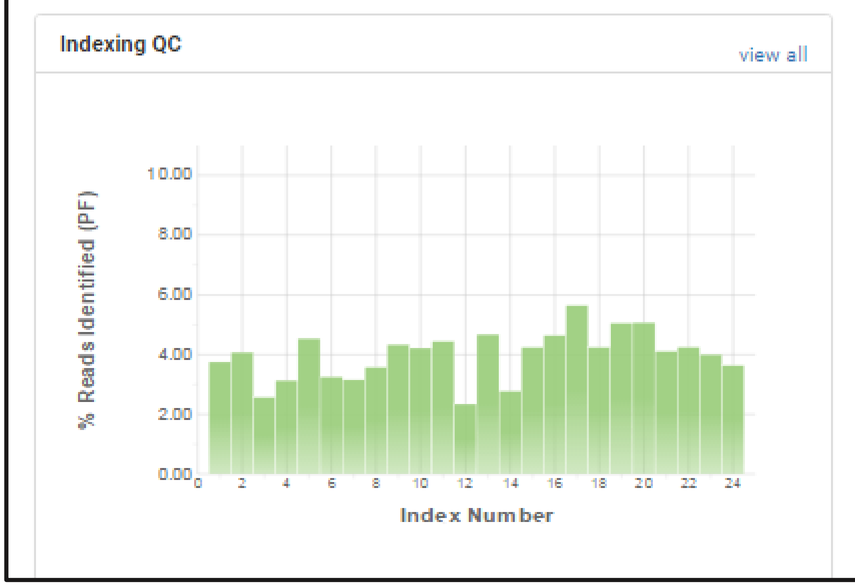
- **Command line configurable options**

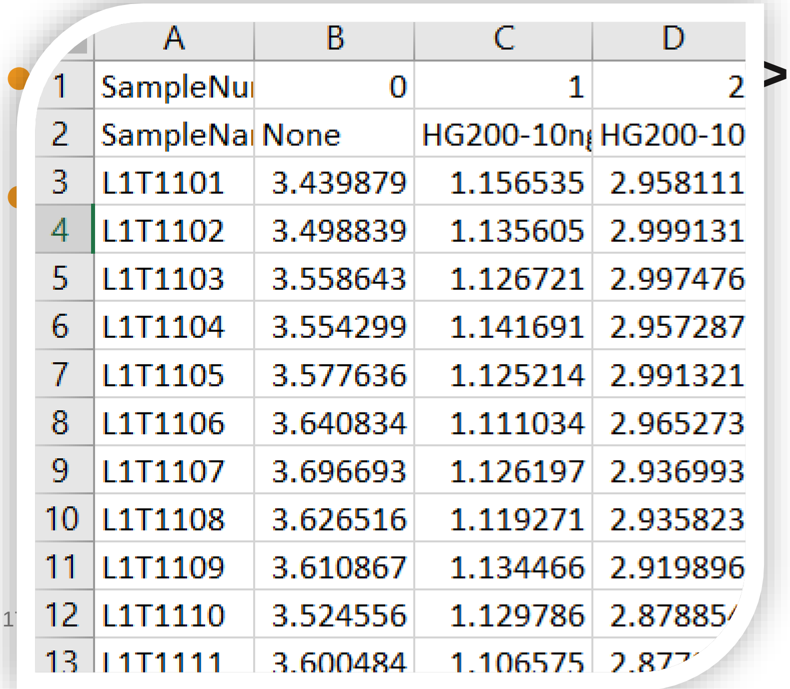
  E.g. `--no-lane-splitting`

  - SampleName_S1_R1_001.fastq.gz
  - SampleName_S1_R1_001.fastq.gz

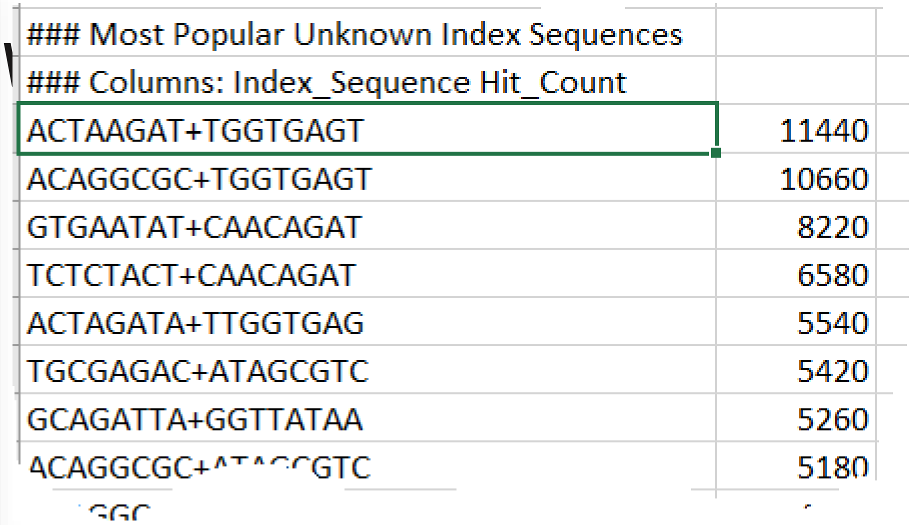
Demultiplexing reports

- **IndexQC on SAV**
- **DemultiplexSummaryF1L#.txt**
  - Helps troubleshoot sample dropouts
- **Depending on what software is used for demultiplexing, DemultiplexSummaryF1L1.txt location will vary**





	A	B	C	D
1	SampleNu	0	1	2
2	SampleNa	None	HG200-10n	HG200-10
3	L1T1101	3.439879	1.156535	2.958111
4	L1T1102	3.498839	1.135605	2.999131
5	L1T1103	3.558643	1.126721	2.997476
6	L1T1104	3.554299	1.141691	2.957287
7	L1T1105	3.577636	1.125214	2.991321
8	L1T1106	3.640834	1.111034	2.965273
9	L1T1107	3.696693	1.126197	2.936993
10	L1T1108	3.626516	1.119271	2.935823
11	L1T1109	3.610867	1.134466	2.919896
12	L1T1110	3.524556	1.129786	2.87885
13	L1T1111	3.600484	1.106575	2.87



### Most Popular Unknown Index Sequences	
### Columns: Index_Sequence Hit_Count	
ACTAAGAT+TGGTGAGT	11440
ACAGGCGC+TGGTGAGT	10660
GTGAATAT+CAACAGAT	8220
TCTCTACT+CAACAGAT	6580
ACTAGATA+TTGGTGAG	5540
TGCGAGAC+ATAGCGTC	5420
GCAGATTA+GGTTATAA	5260
ACAGGCGC+ATAGCGTC	5180

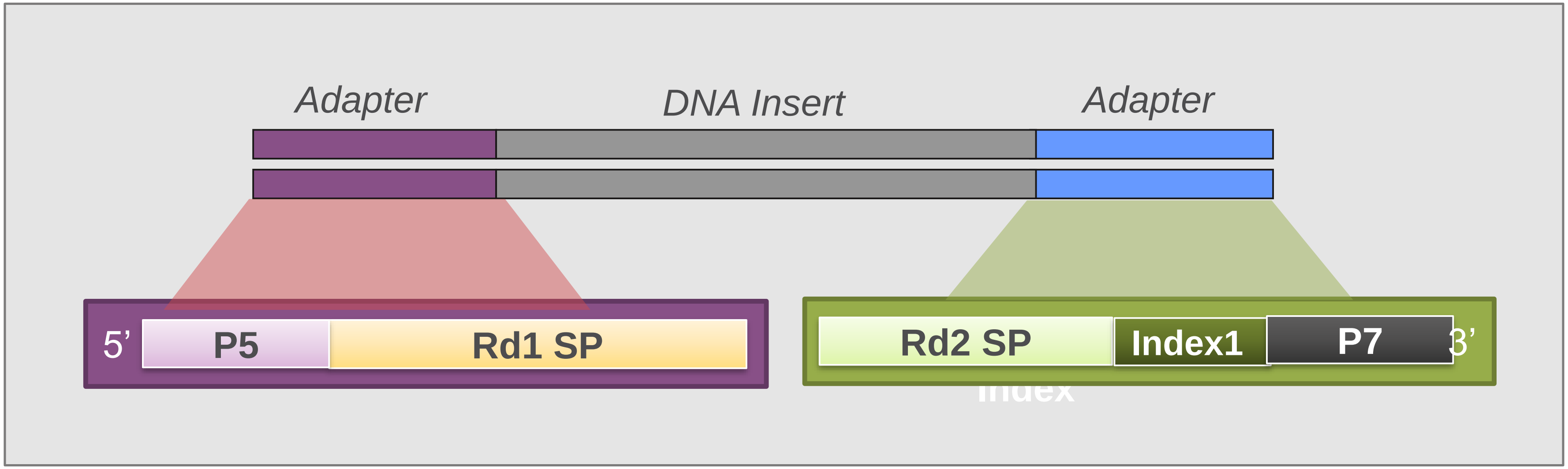
Tools: Adapter Trimming

Illumina library design

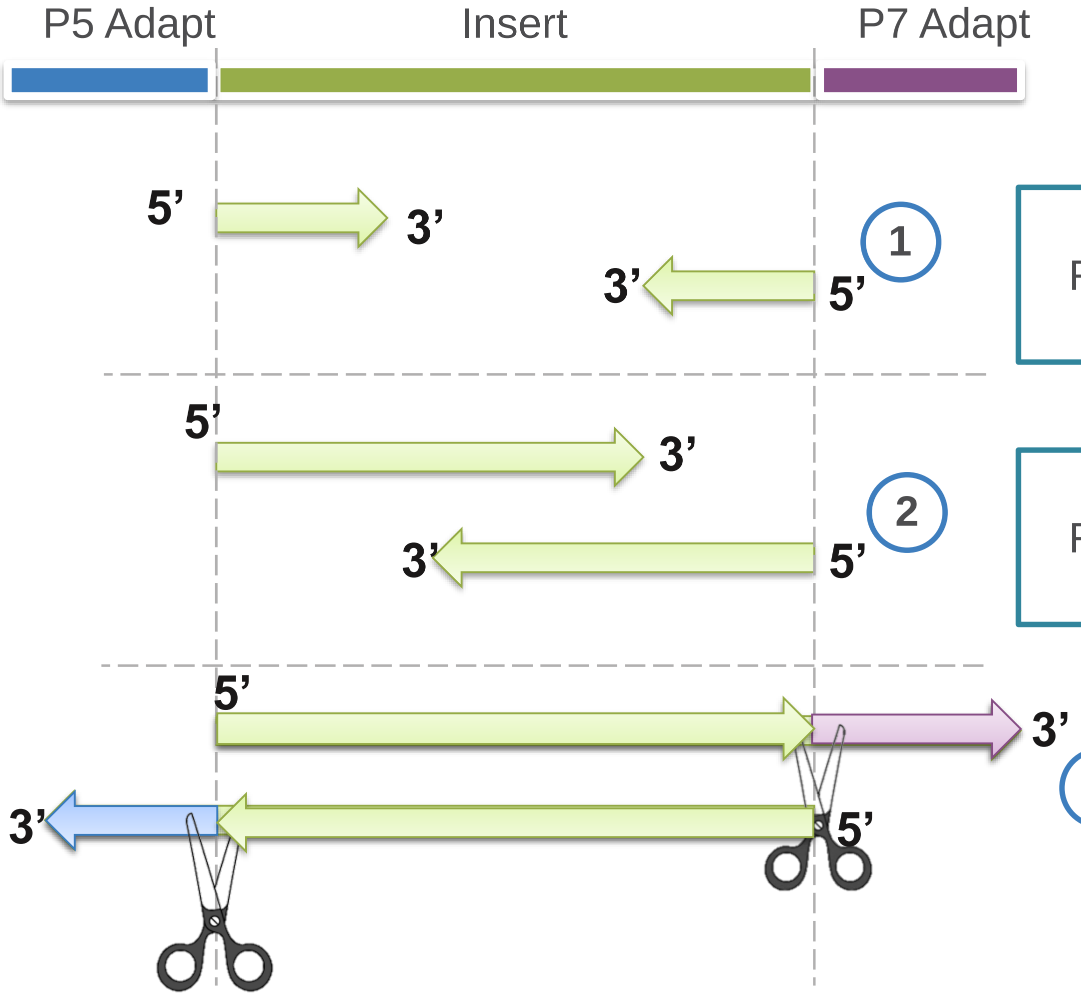
○ Library – DNA insert plus full adapter

○ Read 1 Sequencing Primer (Rd1 SP)

○ Read 2 Sequencing Primer (Rd2 SP)



Adapter trimming overview
PE 151



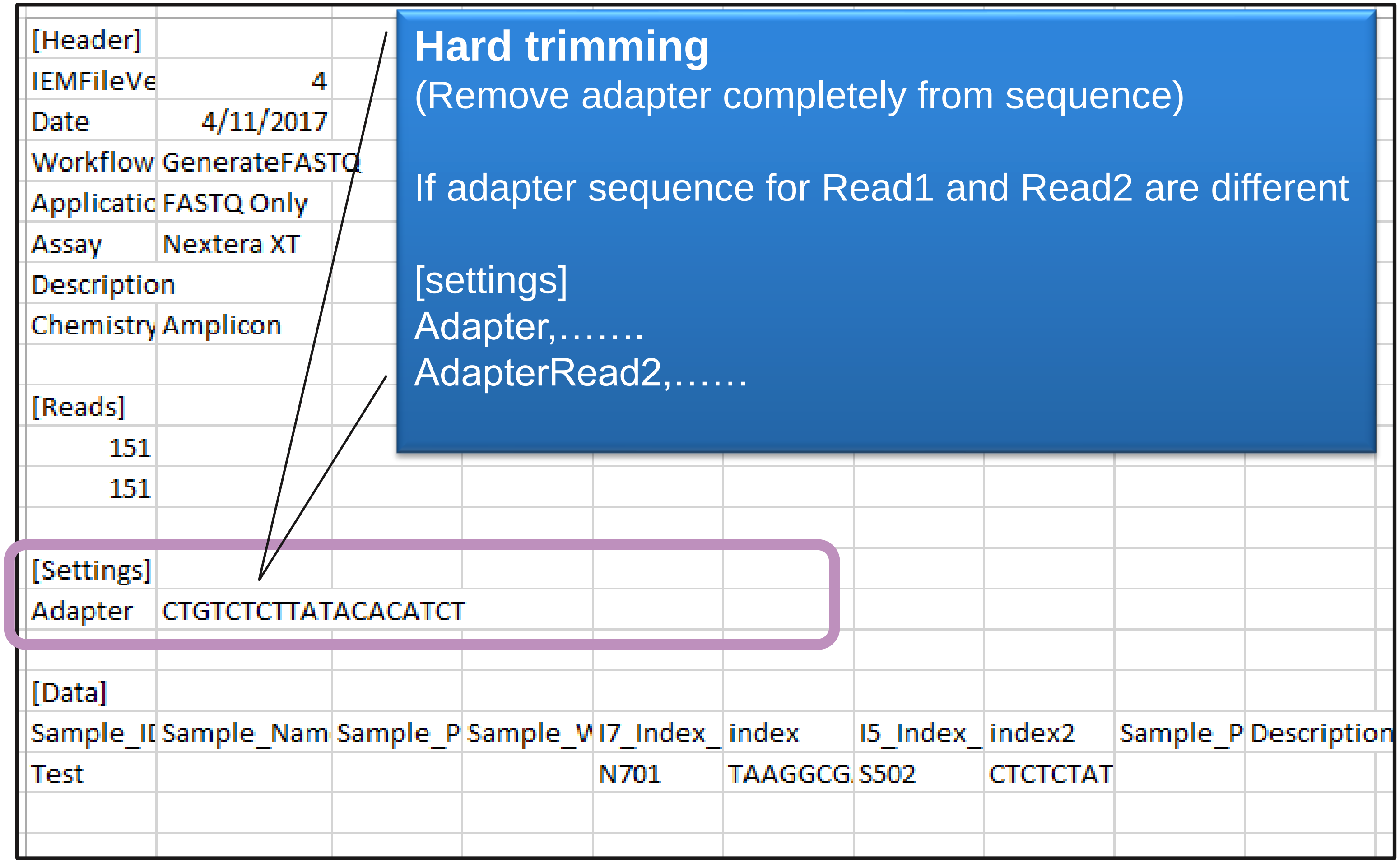
Insert size = 1000bp
Read length shorter than insert

Insert size =200bp
Read length shorter than insert (pairs overlap)

Insert size =100bp
Read length longer than insert

Adapter trimming options

MiSeq™ Reporter, Local Run Manager, BaseSpace™ Fastq Generation and Bcl2fastq



[Header]									
IEMFileVe	4								
Date	4/11/2017								
Workflow	GenerateFASTQ								
Applicatic	FASTQ Only								
Assay	Nextera XT								
Description									
Chemistry	Amplicon								
[Reads]									
151									
151									
[Settings]									
Adapter	CTGTCTCTTATACACATCT								
[Data]									
Sample_II	Sample_Nam	Sample_P	Sample_W	I7_Index_	index	I5_Index_	index2	Sample_P	Description
Test				N701	TAAGGCG.	S502	CTCTCTAT		

Hard trimming

(Remove adapter completely from sequence)

If adapter sequence for Read1 and Read2 are different

[settings]
Adapter,…….
AdapterRead2,……

If runs are setup using PrepTab for NextSeq / MiniSeq, adapter trimming is on by default.

https://support.Illumina.com/bulletins/2016/12/what-sequences-do-i-use-for-adapter-trimming.html

For Research Use Only. Not for use in diagnostic procedures.

Adapter trimming example (hard trimming)

Adapter match > 90% (default)

```
@M00000:71:000000000-D00LW:1:1101:16265:1658 1:N:0:1
ACTCTGCGTTGCGCTTCTGCTCGGCCTCCAGCTCACCCTCCCTGTCTCTTATACACATCTCCGAGCCCA
+
BCCCCFFCCBCCGGGGGGGGGGGGGGGHHHHHHHHHHHHHGGHHHHHHHHHHHHHHHHHHHHHGGGGGH
```

```
@M00000:71:000000000-D00LW:1:1101:16265:1658 1:N:0:1
ACTCTGCGTTGCGCTTCTGCTCGGCCTCCAGCTCACCCTCC
+
BCCCCFFCCBCCGGGGGGGGGGGGGGGHHHHHHHHHHHHHG
```

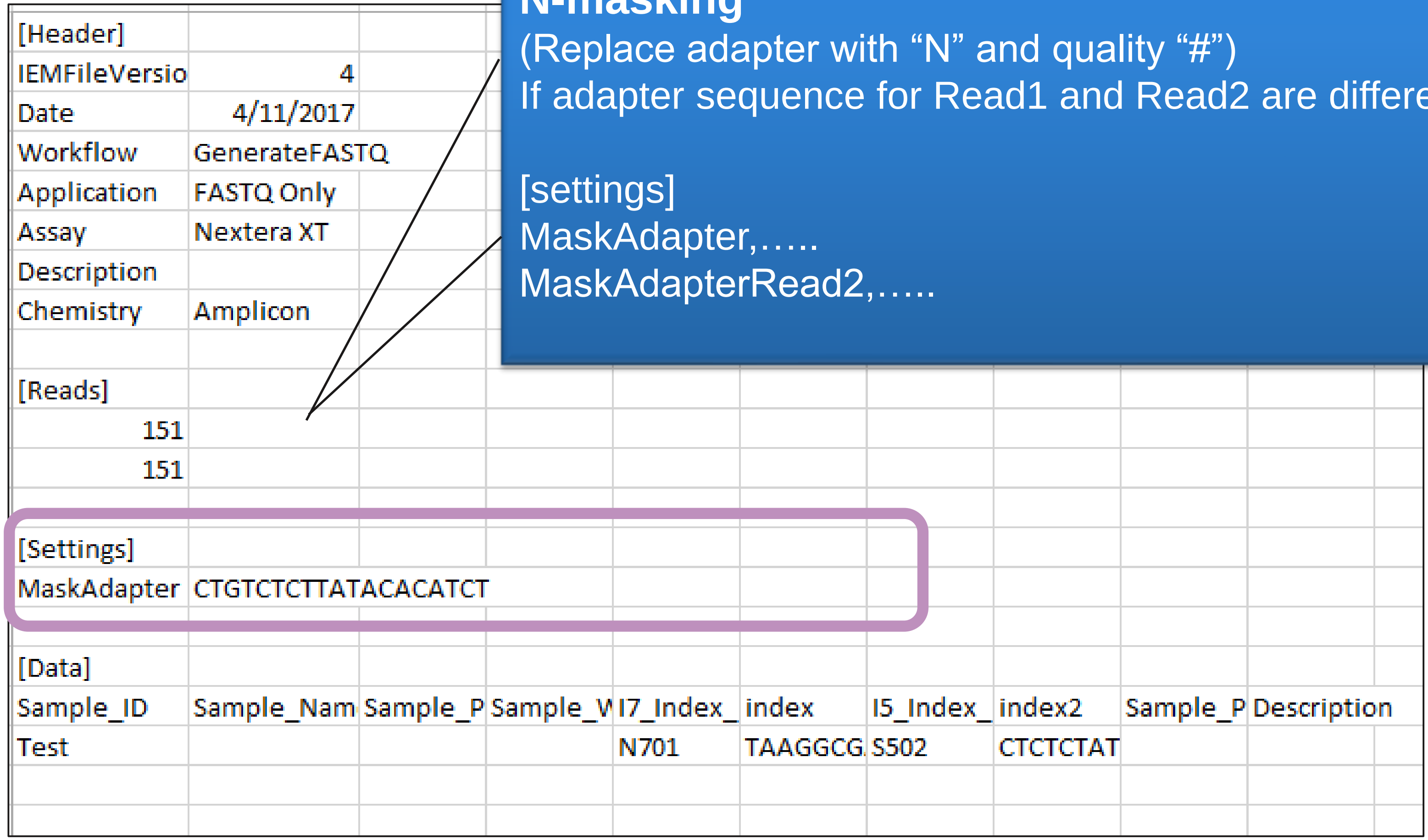
Bases trimmed (removed) from the start of the Adapter

Adapter masking options

Bcl2fastq

N-masking
(Replace adapter with “N” and quality “#”)
If adapter sequence for Read1 and Read2 are different

[settings]
MaskAdapter,…..
MaskAdapterRead2,…..



[Header]									
IEMFileVersio	4								
Date	4/11/2017								
Workflow	GenerateFASTQ								
Application	FASTQ Only								
Assay	Nextera XT								
Description									
Chemistry	Amplicon								
[Reads]									
151									
151									
[Settings]									
MaskAdapter	CTGTCTCTTATACACATCT								
[Data]									
Sample_ID	Sample_Nam	Sample_P	Sample_W	I7_Index_	index	I5_Index_	index2	Sample_P	Description
Test				N701	TAAGGCG	S502	CTCTCTAT		

Adapter masking example

```
@M00000:71:000000000-D00LW:1:1101:16265:1658 1:N:0:1
ACTCTGCGTTGCGCTTCTGCTCGGCCTCCAGCTCACCCTCCCTGTCTCTTATACACATCTCCGAGCCCA
+
BCCCCFFCCBCCGGGGGGGGGGGGGGGHHHHHHHHHHHHHGGHHHHHHHHHHHHHHHHHHHHHGGGGGH
```

```
@M00000:72:000000000-D00LW:1:1101:16265:1658 1:N:0:1
ACTCTGCGTTGCGCTTCTGCTCGGCCTCCAGCTCACCCTCCNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+
BCCCCFFCCBCCGGGGGGGGGGGGGGGHHHHHHHHHHHHHG############################
```

Bases masked to “N” and quality score to 2 “#” from the start of the adapter
(Useful if analysis program requires reads to be same read length)

illumina®

Masking short adapters – bcl2fastq

```
@M00000:71:000000000-D00LW:1:1101:16265:1658 1:N:0:1
ACCTGTCTCTTATACACATCTCCGAGTCTGCGTTGCGCTTCTGCTCGGCCTCCAGCTCACCCTCCCCCA
+
BCCCCFFCCBCCGGGGGGGGGGGGGGGHHHHHHHHHHHHHGGHHHHHHHHHHHHHHHHHHHHGGGGGH
```

```
@M00000:72:000000000-D00LW:1:1101:16265:1658 1:N:0:1
NNNNNNNNNNNNNNNNNNNNNNNNNNNN
+
############################
```

--minimum-trimmed-read-length 35
--mask-short-adapter-reads 22

Entire read is converted to 35bp Ns and quality score to "#" from the start of the sequence

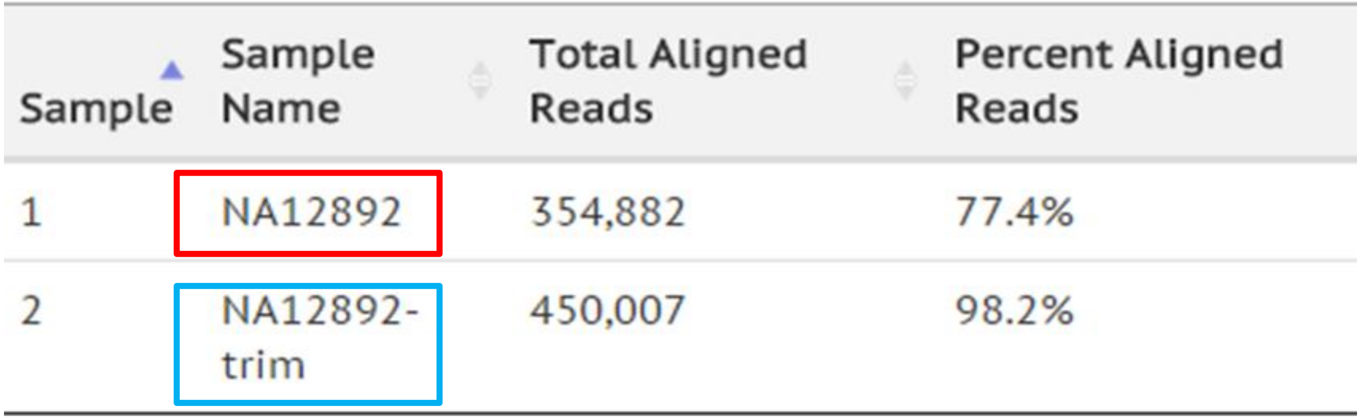
- Useful if adapter dimers in the sample
- Prevents generation of empty reads

Why adapter trim?

BWA
Enrichment
V2.1

1 Higher alignment %

BWA
(backtrace)



Sample	Sample Name	Total Aligned Reads	Percent Aligned Reads
1	NA12892	354,882	77.4%
2	NA12892-trim	450,007	98.2%

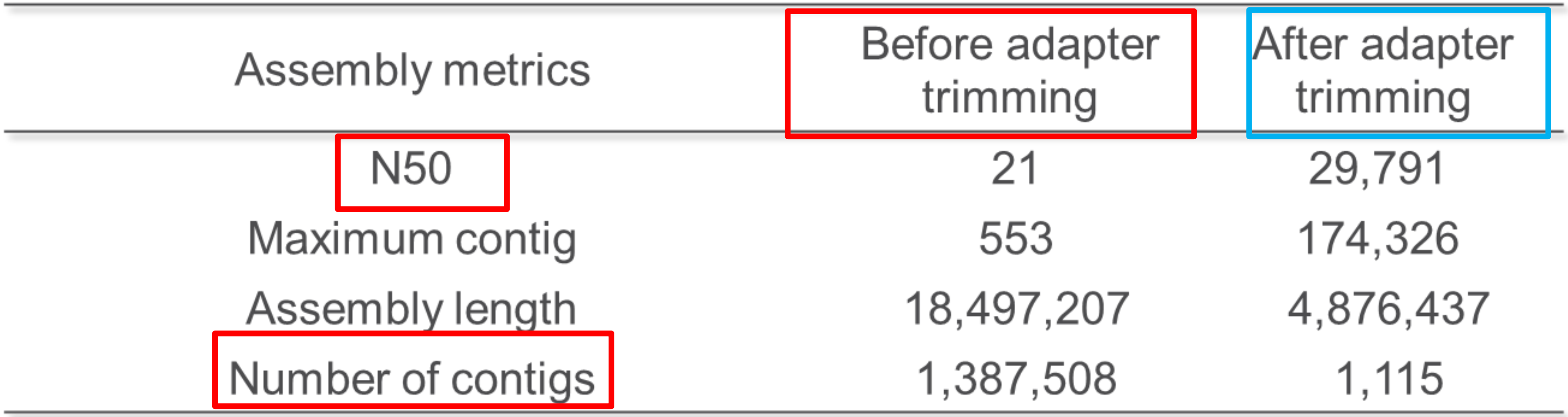
Why adapter trim?

Velvet de novo Assembly
BASESPACE LABS

2

Improved assemblies

Data: 2 x 250bp, *E.coli* (Nextera™ XT)



Assembly metrics	Before adapter trimming	After adapter trimming
N50	21	29,791
Maximum contig	553	174,326
Assembly length	18,497,207	4,876,437
Number of contigs	1,387,508	1,115

For Research Use Only. Not for use in diagnostic procedures.

Tools: Quality trimming

Quality trimming

- **Filter the end of reads based on read end average quality**
- **When to trim?**
  - Where algorithms sensitive to quality – De Novo assembly, merging reads, Metagenomics (using 16S rRNA gene for classification)
- **When not to trim?**
  - Resequencing. Most aligners take quality scores into account (i.e. BWA, Isaac) will soft clip the ends of reads if low quality
- **FASTQ toolkit app on BSSH (BaseSpace Sequence Hub) can be used to perform quality trimming**

Quality score trimming example

QualityScoreTrim,20

```
@M00000:72:000000000-D00LW:1:1101:22420:18334 1:N:0:1
CACCAAGGGCCTGGGGTGTCAATGGCGGGGCTTGTGACTGCACAAAAGGGGCCTCCCGCAGGGGCTCCCGCC
+
BBBBBBFBBBBBGGGGEEFGGGHHHHGGG00>A0B355@BB3@3BGB?E1///1/11//////////?////
```

```
@M00000:72:000000000-D00LW:1:1101:22420:18334 1:N:0:1
CACCAAGGGCCTGGGGTGTCAATGGCGGGGCTTGTGACTGCACAAAAGG
+
BBBBBBFBBBBBGGGGEEFGGGHHHHGGG00>A0B355@BB3@3BGB?E
```

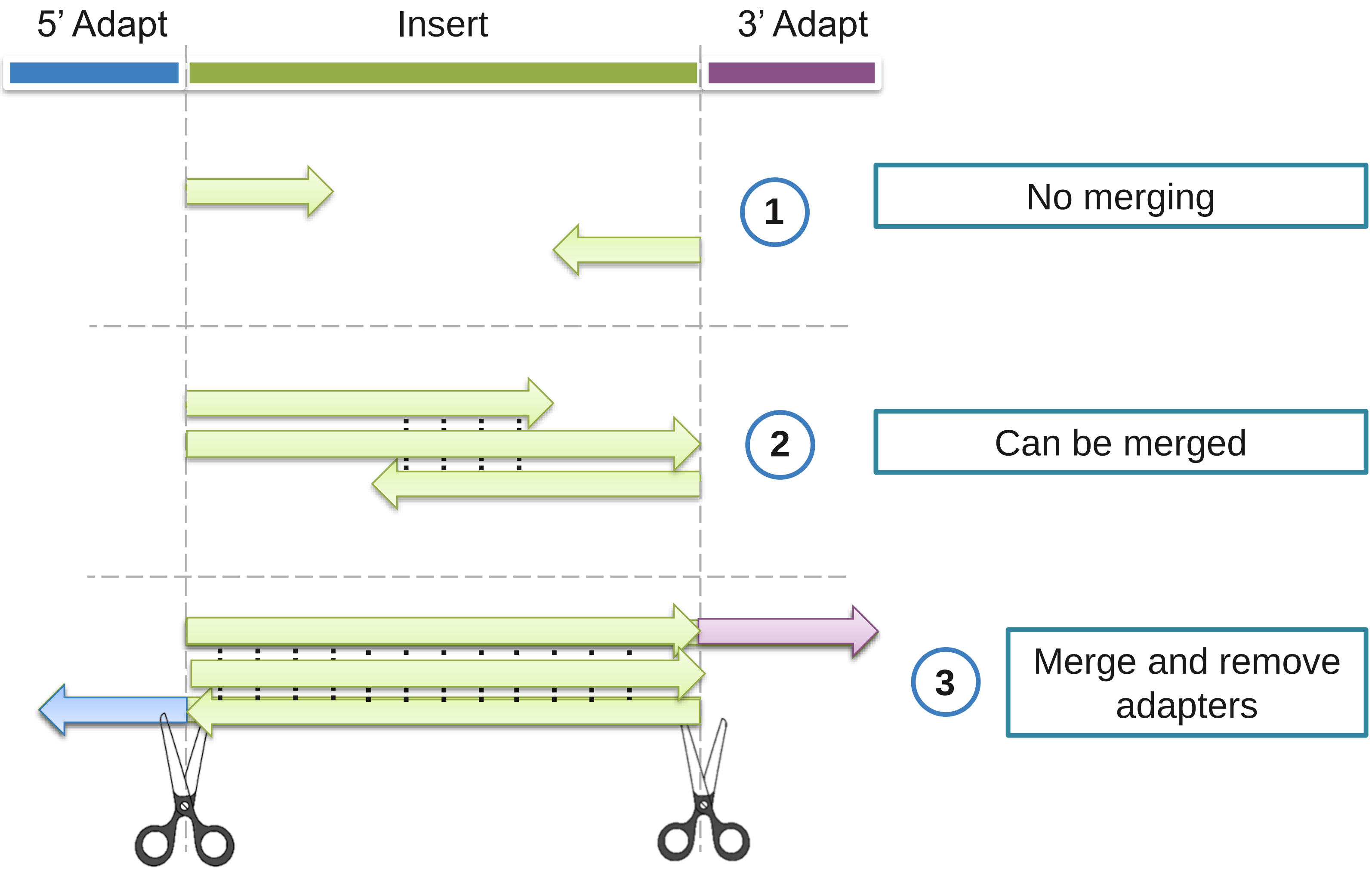
Q	ASC
13	.
14	/
15	0
16	1
18	3
20	5
22	7
25	9
30	?
31	@
32	A
33	B

Tools: Read merging

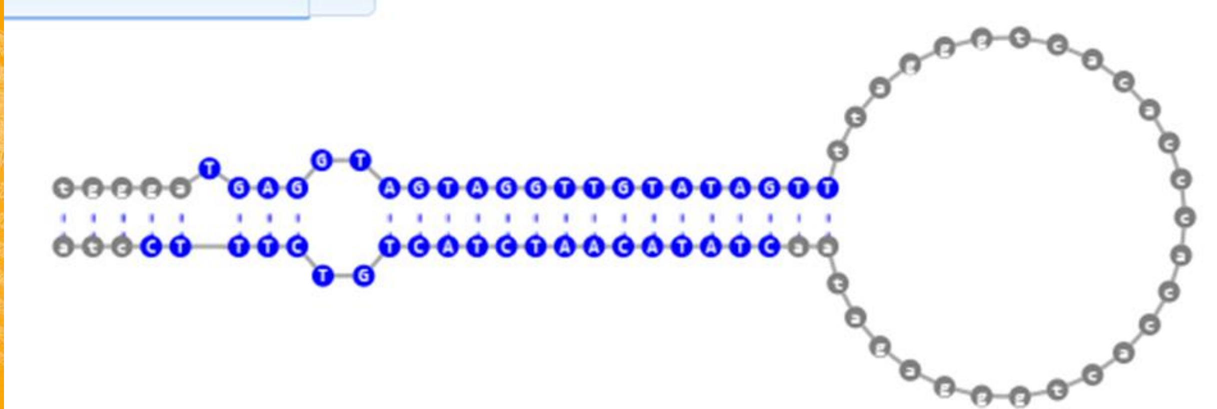
Read merging

- **Combines paired end read into a single read**
- **Generally requires substantially high overlap**
  - StitchReads,1 option in MiSeq™ Reporter
  - Workflows = Amplicon – DS, GenerateFASTQ workflow, TruSeq™ Amplicon
  - 10 bp overlap
  - Not supported in any other software
- **When to merge?**
  - Where read continuity is important
  - Improve indel detection in some cases (i.e. where Indel is in overlap region)
  - To use tools that only take single end reads (for example some metagenomics software)
  - Where majority of reads overlap
- **When not to merge?**
  - When some proportion of reads do not overlap
  - Where there are simple repeats in overlap region (i.e. sequencing amplicons with expansion repeats)
  - Most downstream aligners will handle unmerged reads

Read merging



Data optimization with an example:
Using FASTQ toolkit app for Small RNA workflow



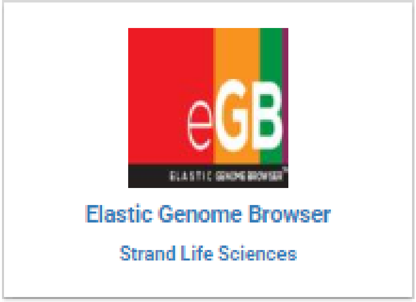
BaseSpace™

- BaseSpace™ is the Illumina cloud-based genomics computing environment for next-generation sequencing (NGS) data management and analysis.
- Is used to analyze or manipulate samples – results stored in the project. There are currently about 80+ apps

Core App

BaseSpace Lab App

Third party App



Please Read:
This is a BaseSpace™ Labs App. BaseSpace Labs Apps are developed using an accelerated development process in order to make them available to BaseSpace users faster than conventional Illumina Apps. Illumina may provide support for BaseSpace Labs Apps at its sole discretion. BaseSpace Labs Apps are provided AS-IS without any warranty of any kind. BaseSpace Labs Apps are used at the user's sole risk. Illumina is not responsible for any loss of data, incorrect results, or any costs, liabilities, or damages that may result from the use of a BaseSpace Labs App.

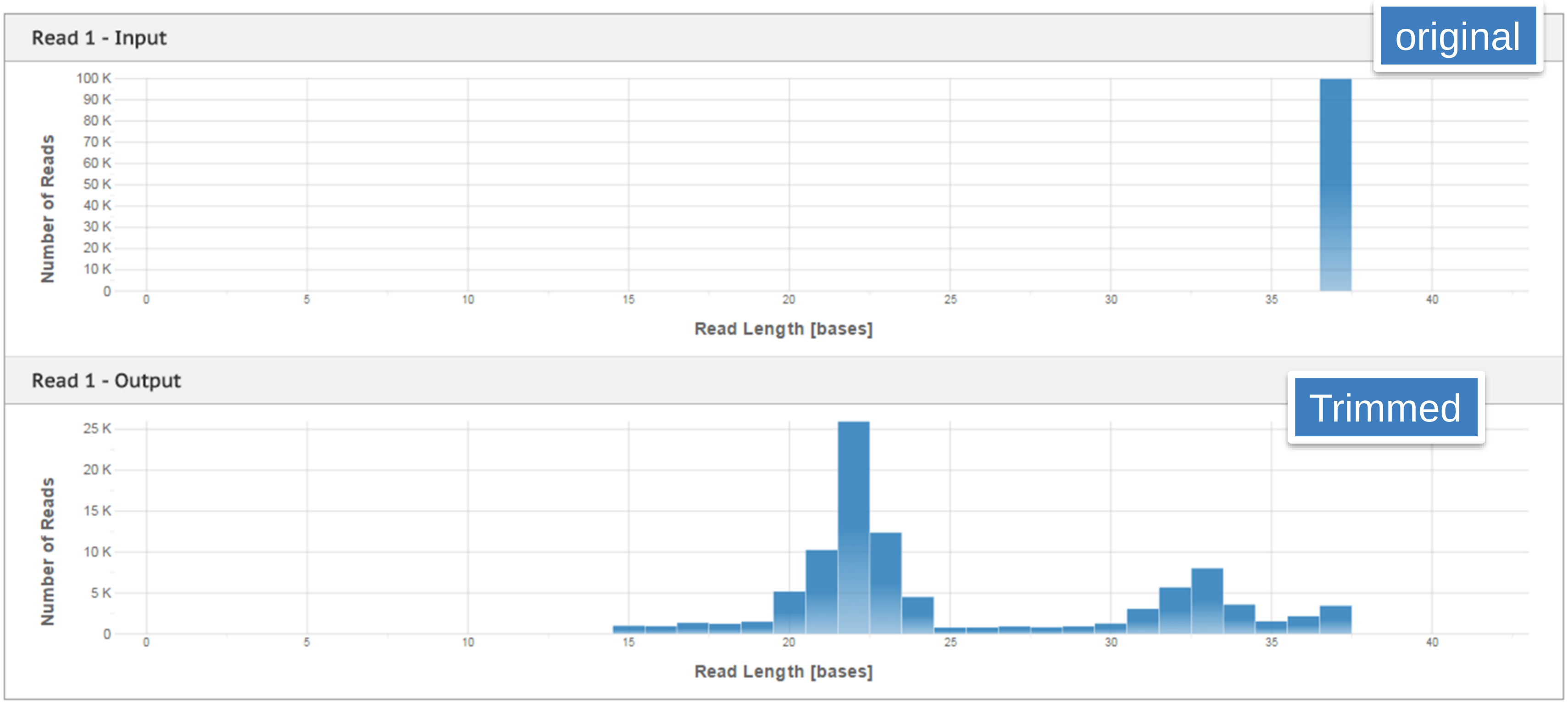
Small RNA sample prep kit

Small RNA

Illumina, Inc.

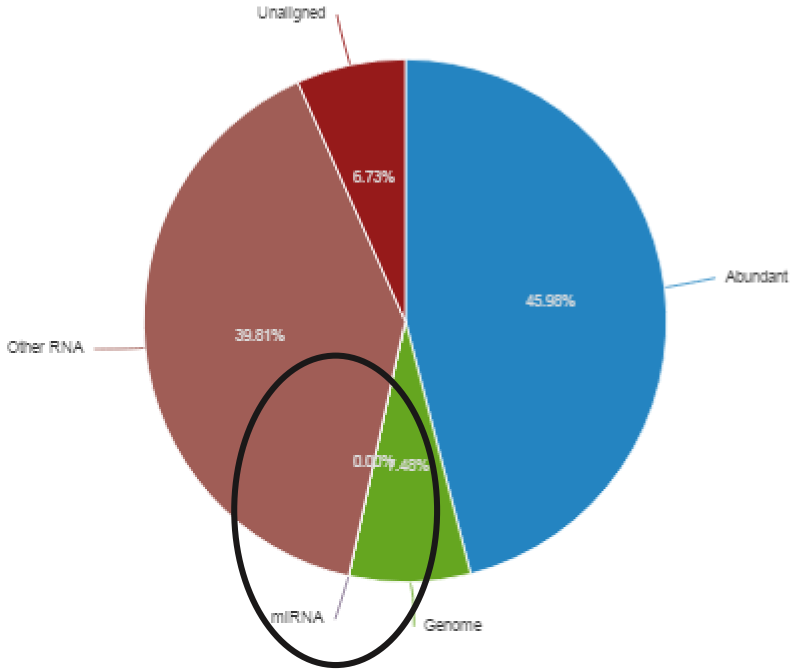
- **Human small RNA peaks at ~22bp**
- **Typical run setup is Single End = 36 to 50cycles. So reads have to be adapter trimmed**
- **Bcl2fastq**: --minimum-trimmed-read-length 20 and --mask-short-adapter-reads 20
- **Recommendation: For workflow GenerateFASTQ – do not turn on adapter trimming**
- **Generate FASTQs with 50bp and trim downstream using 3rd party tools or BSSH FASTQ toolkit**

FASTQ toolkit output- Results

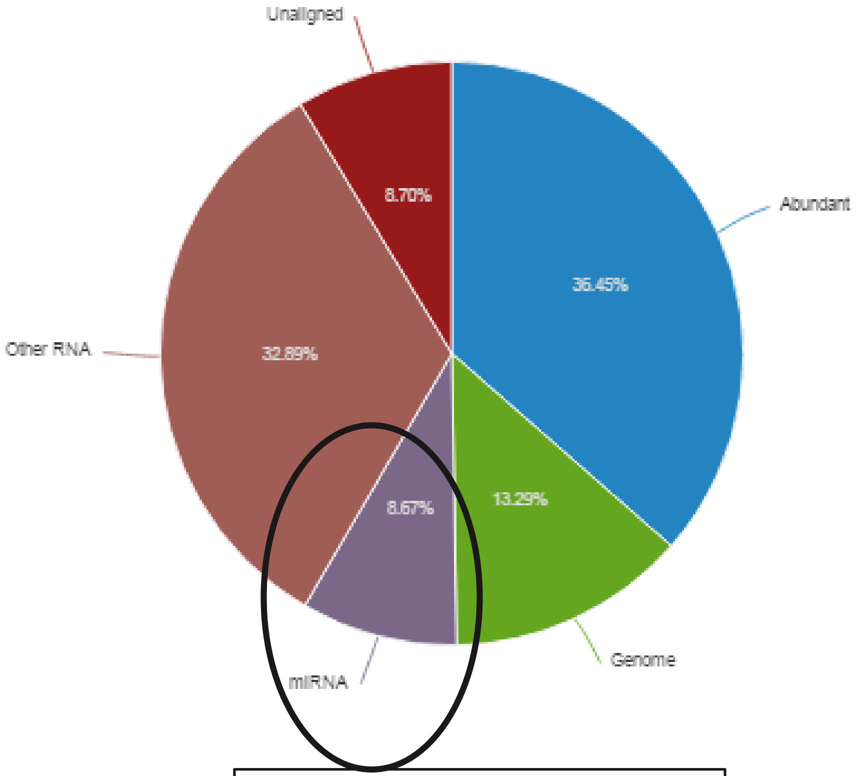


Small RNA app output

Small RNA
Illumina, Inc.



Untrimmed FASTQ

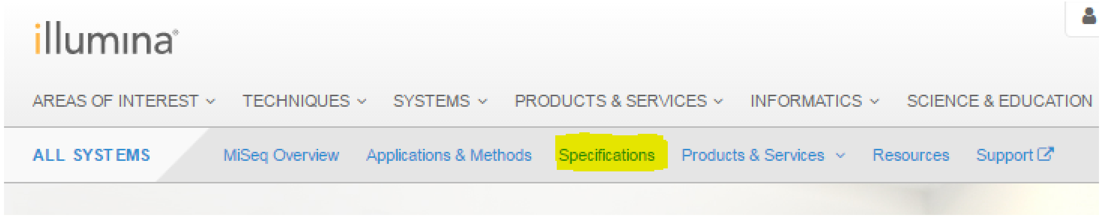


Trimmed FASTQ

Does my run look good?
vs
Does my FASTQ look good?

Previous Webinar -Sequence Analysis Viewer (SAV): A Beginner's Guide

Metrics



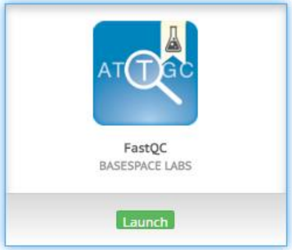
Illumina

- Based on Sequencing Analysis Viewer – taking a whole run into consideration
- Yield, Q30, Reads passing filter
- Used for assessing run quality – instrument performance
- Illumina.com > Systems > MiSeq™ > Specifications

3rd party tools *Eg.,* FASTQC

- Based on individual FASTQ file
- Sample dependent and hence library kit
- Used for assessing sample performance – primarily for troubleshooting individual sample

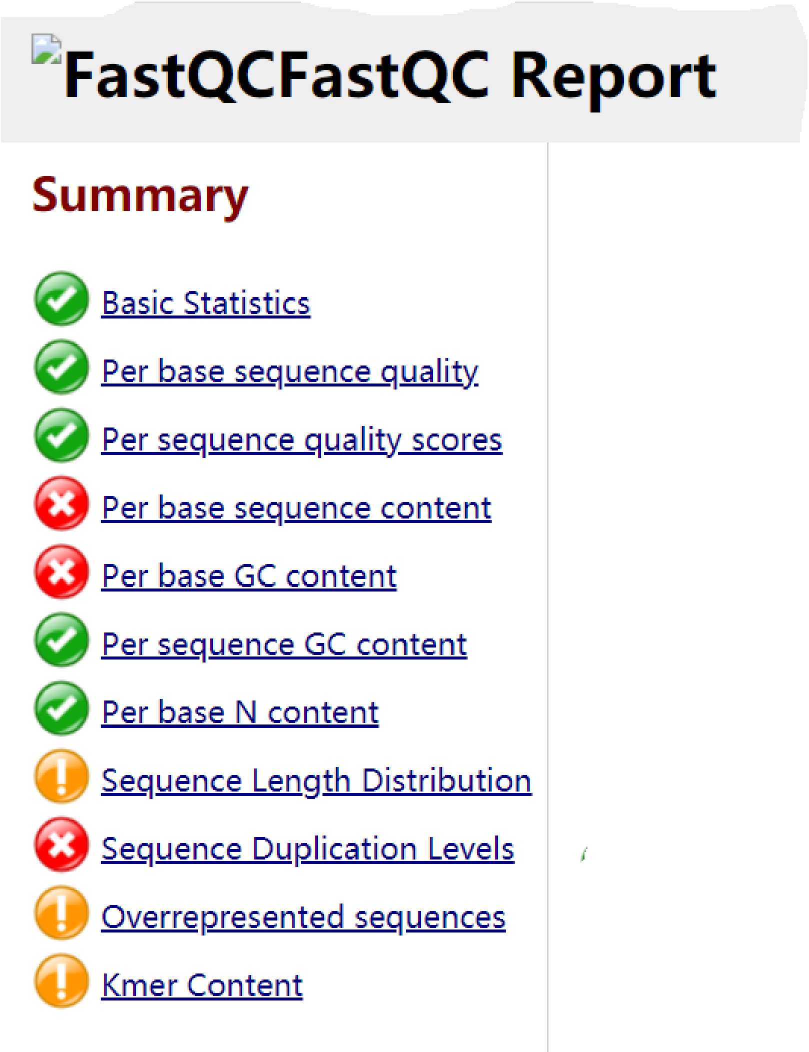
Does my FASTQ look good?



- FASTQC app quick QC step to assess sample quality - subsamples from read pool
- FASTQC tool helps understand the nature of your library
- Not all graphs are meaningful - Understand documentation
- Failure of QC plots in the app does not translate to poor data quality – understand your data
- Sometimes it needs more in-depth downstream data analysis

✓ ! ✗

What FastQC report looks like

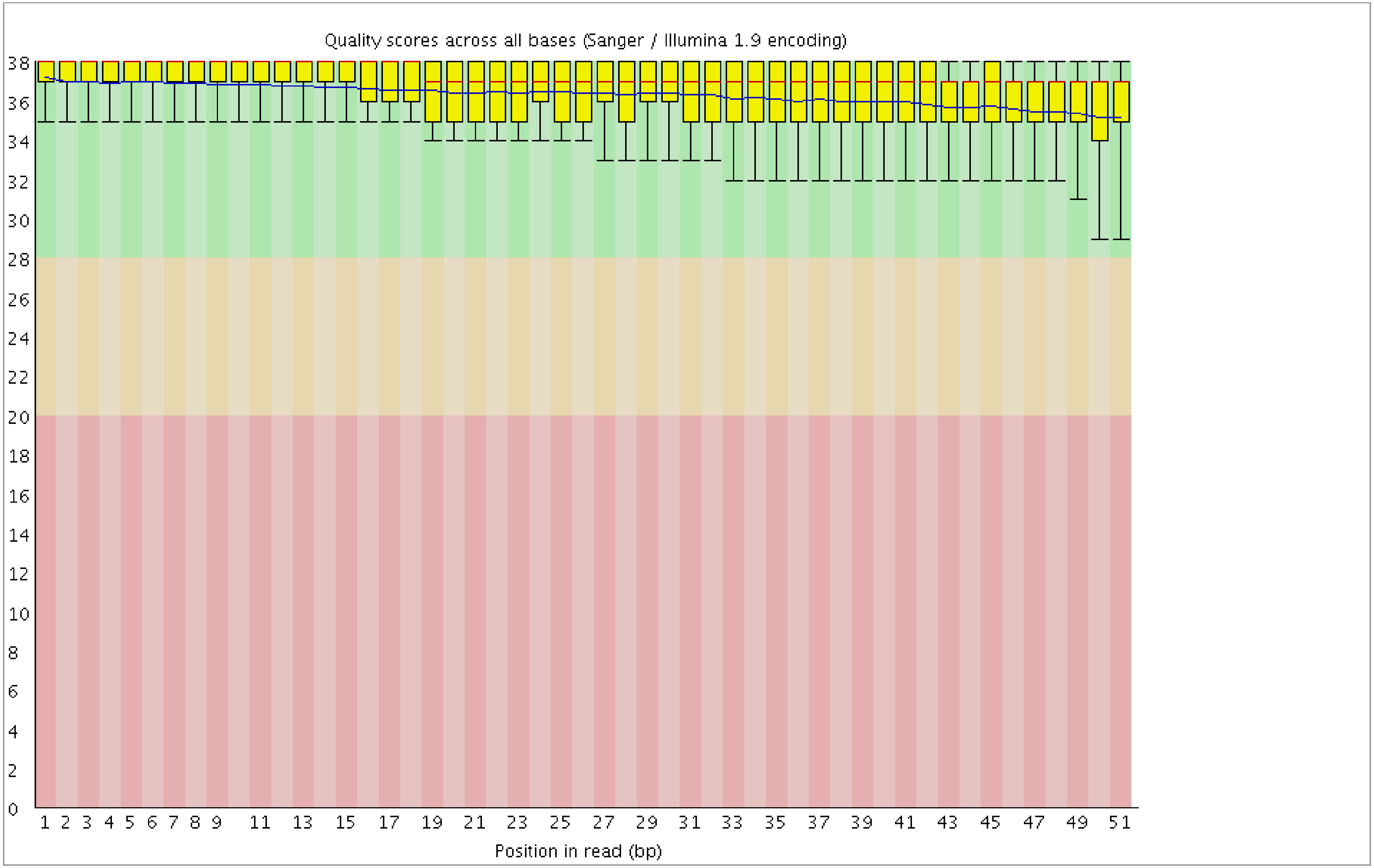


FastQCFastQC Report

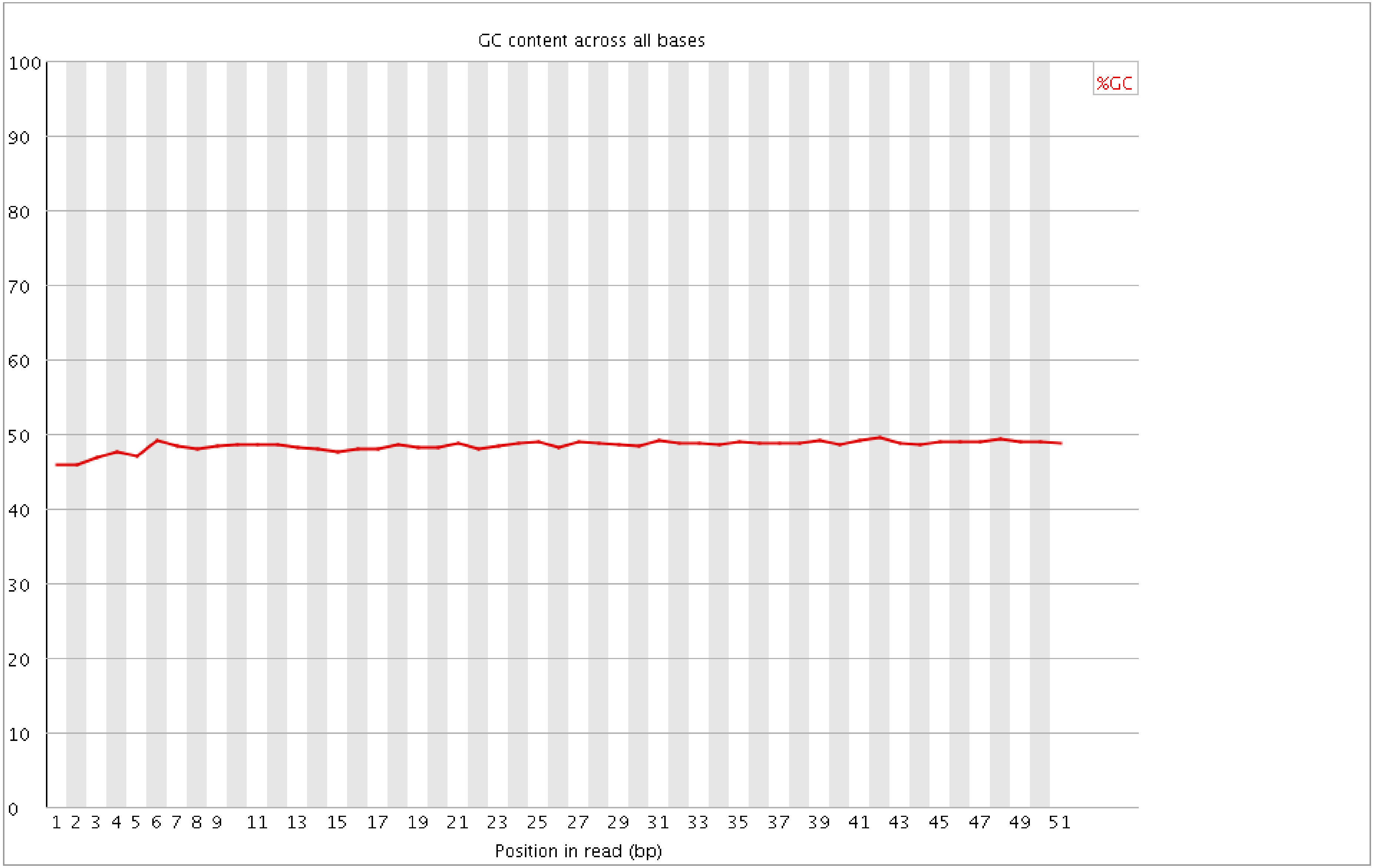
Summary

- ✓ 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

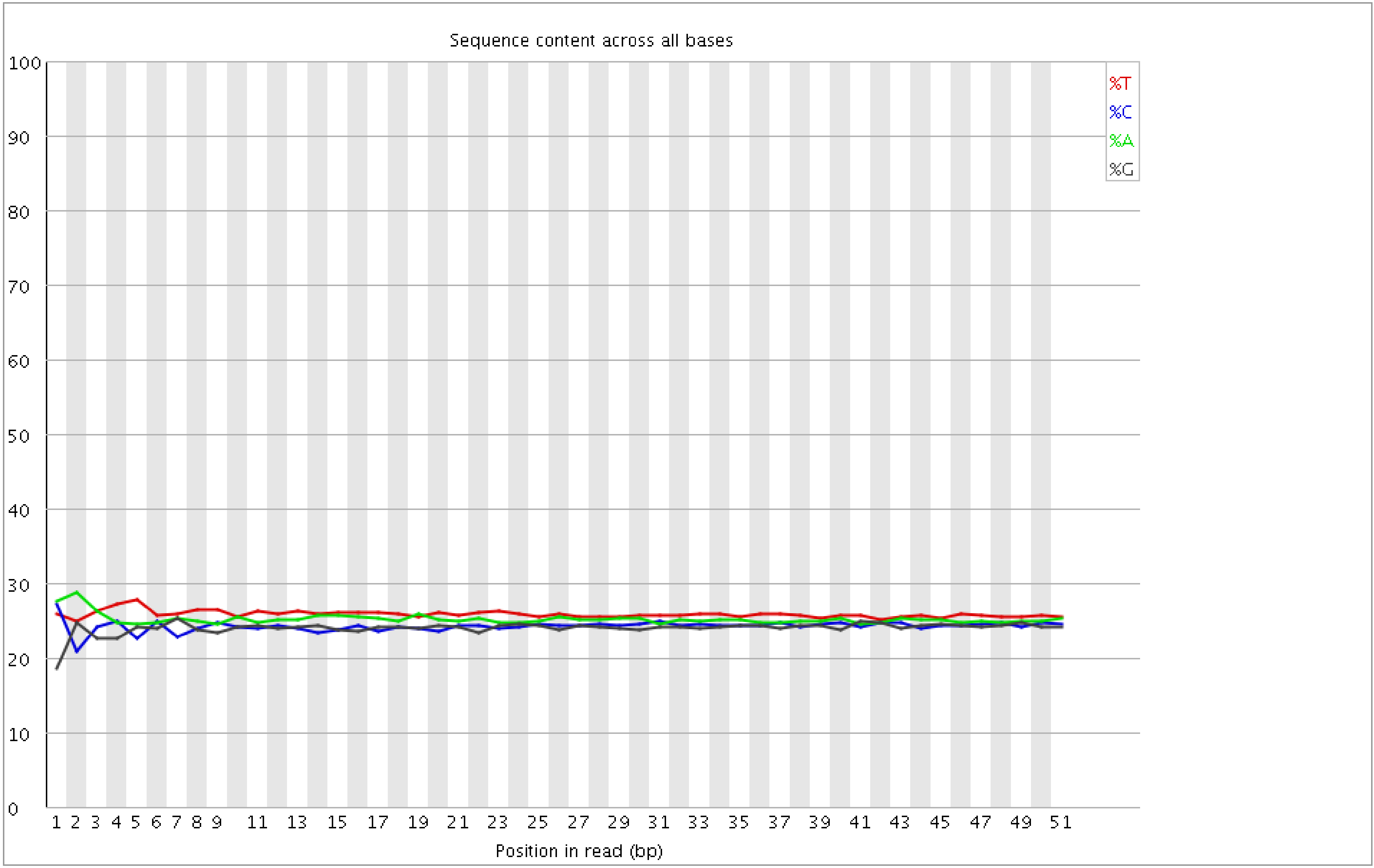
Per Base Sequence Quality



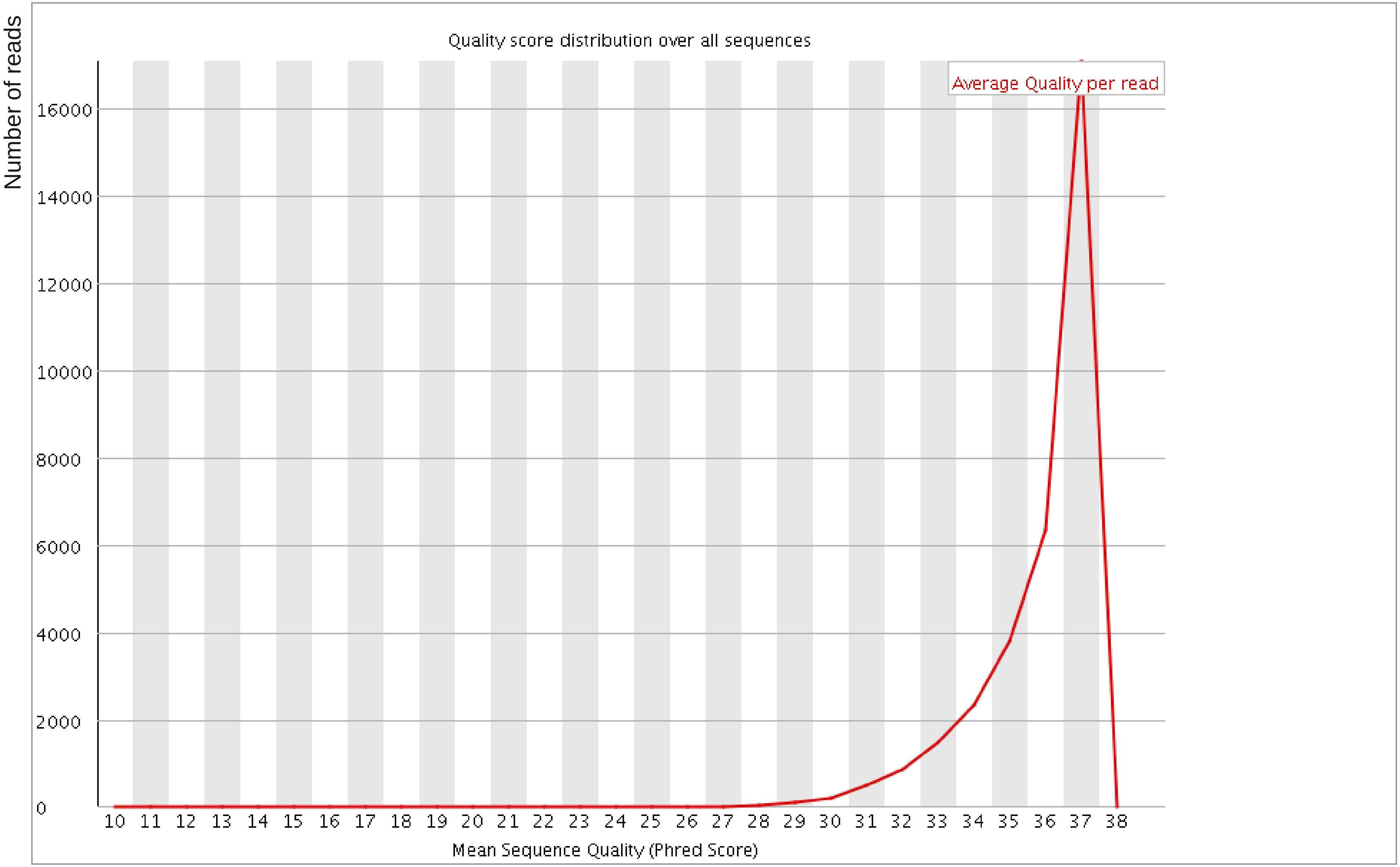
✓ Per Base GC Content



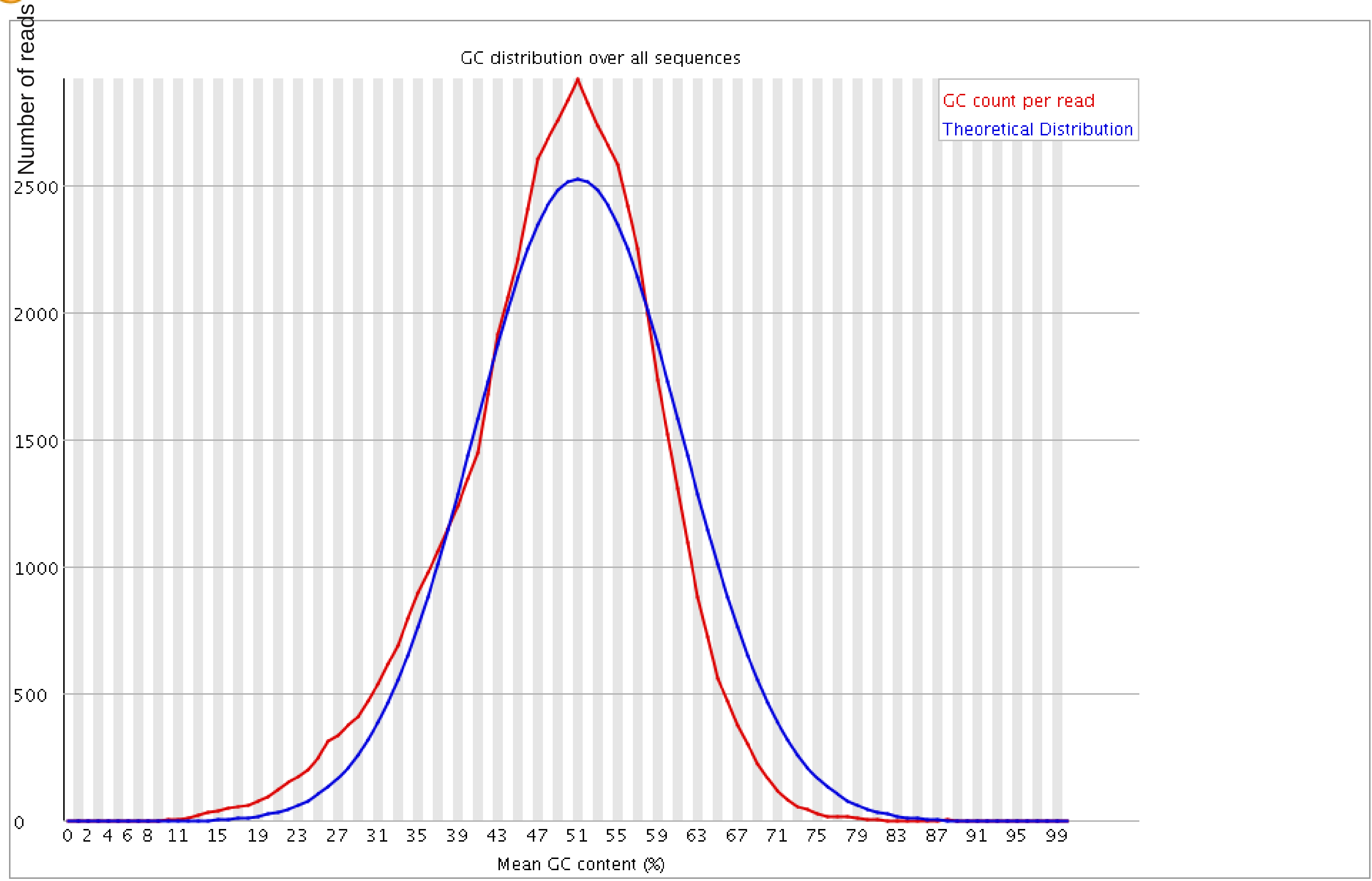
Per Base Sequence Content



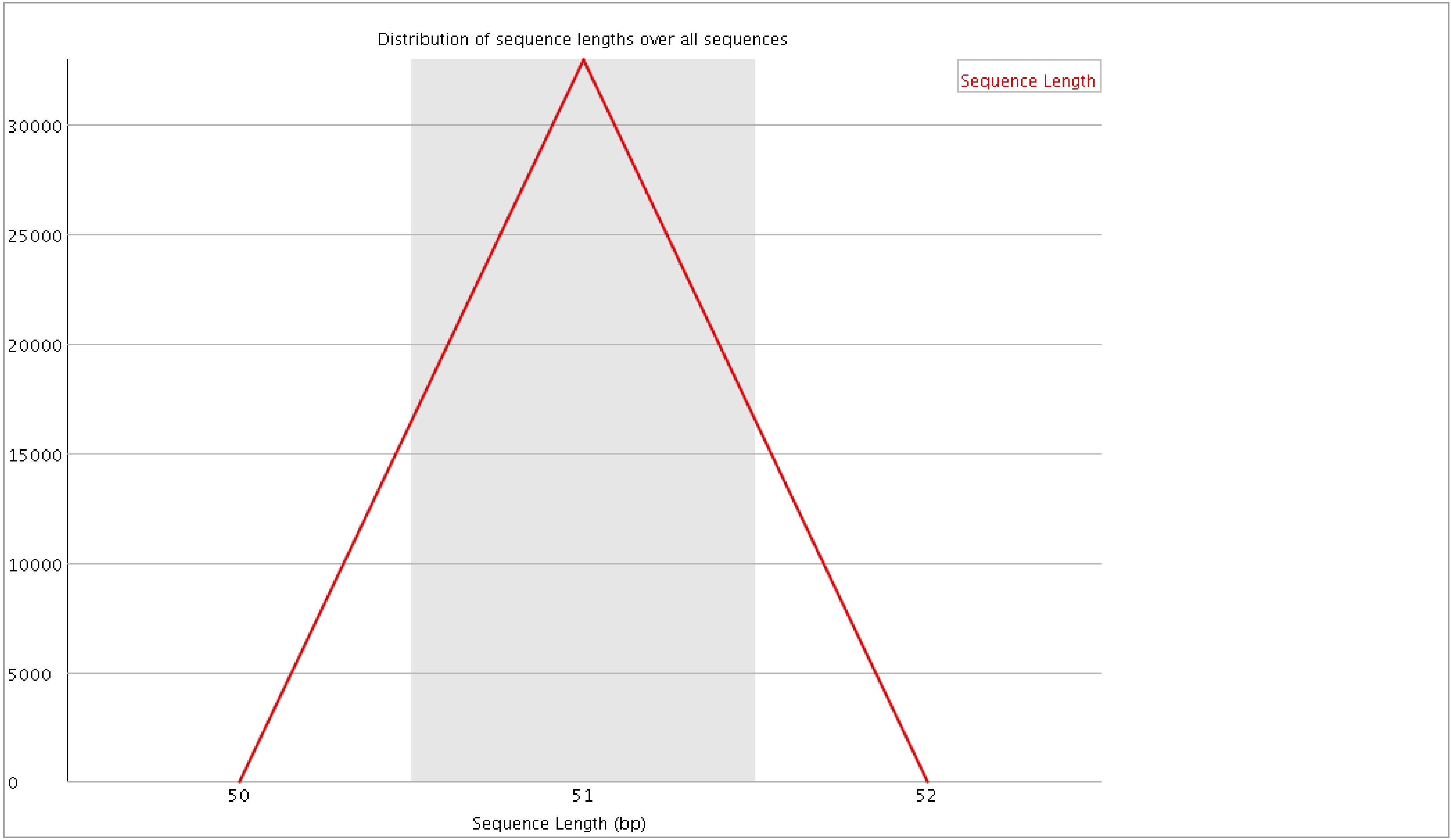
✓ Per Sequence Quality Scores



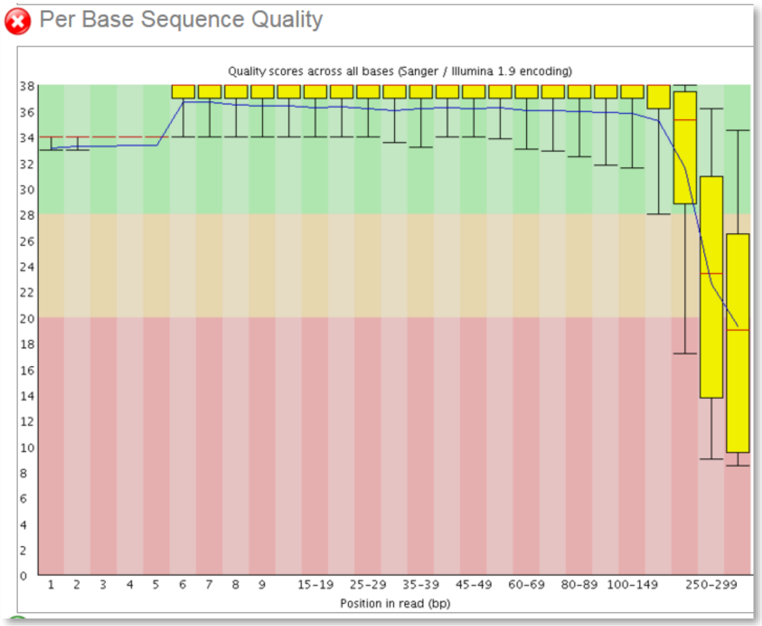
Per Sequence GC Content

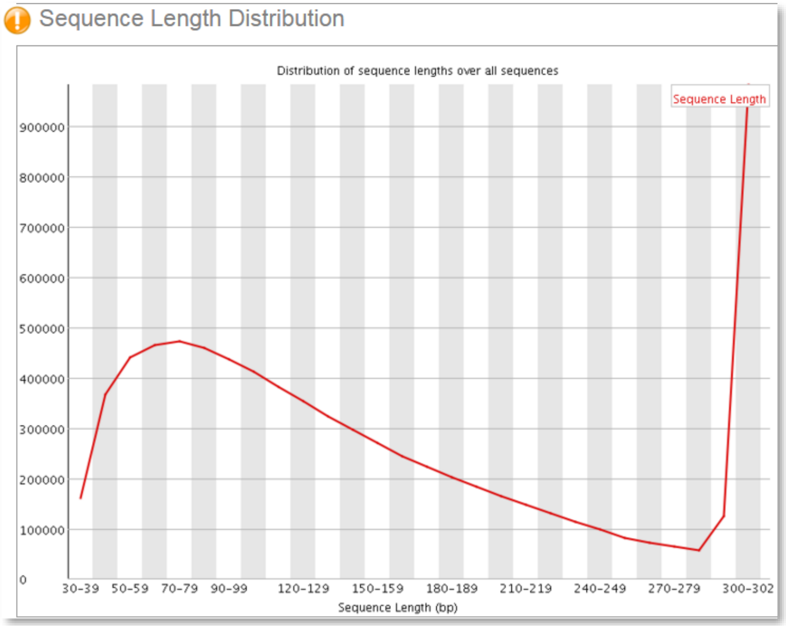


Sequence Length Distribution



Example of less than ideal data

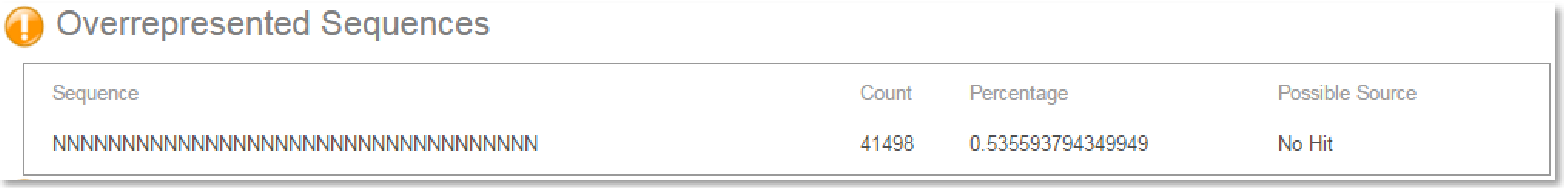




- Indication of potential run problem, check SAV data

- Indication of mixed library types or potential library problem?

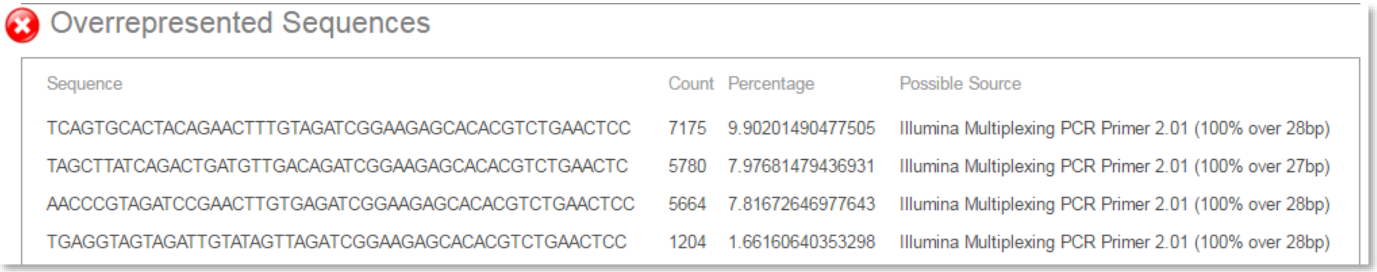
Example of less than ideal data



Overrepresented Sequences

Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	41498	0.535593794349949	No Hit

- Indication of poor sequence quality, using non-PF data?
- Using adapter masks, short reads?



Overrepresented Sequences

Sequence	Count	Percentage	Possible Source
TCAGTGCACTACAGAACTTTGTAGATCGGAAGAGCACACGTCTGAACTCC	7175	9.90201490477505	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TAGCTTATCAGACTGATGTTGACAGATCGGAAGAGCACACGTCTGAACTC	5780	7.97681479436931	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)
AACCCGTAGATCCGAACTTGTGAGATCGGAAGAGCACACGTCTGAACTCC	5664	7.81672646977643	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TGAGGTAGTAGATTGTATAGTTAGATCGGAAGAGCACACGTCTGAACTCC	1204	1.66160640353298	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)

- Indication of adapter presence
- adapter-dimer or adapter contamination

Summary

- **FASTQ format, demultiplexing and FASTQ generation using Illumina tools**
- **FASTQ processing tools for adapter trimming, quality trimming, read merging** *et.al.*
- **Illumina metrics are used for the overall run performance whereas third party tools such as FASTQC are used to look at sample performance**

Resources

- **FASTQC detailed documentation**
- What sequences do I use for adapter trimming?
- **Demultiplexing resource bulletin**
- FASTQ files explained
- Adapter trimming: Why are adapter sequences trimmed from only the 3' ends of reads?
- **Illumina在线技术培训研讨会 — Assessing Run Quality With SAV And FastQC**
- **Illumina在线技术培训研讨会 — Introduction To Bcl2fastq V2+**

QUESTIONS?