SciLifeLab

NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

Quality Assessment of sequencing data

elixir

excelerate

Outline

- General Principles
  - Why QC?
  - Data Integrity
- Illumina
  - Data Format
  - FastQC
- PacBio
  - Data Format
  - FastQC
  - SMRT Portal

Quality Assessment

- Why check your data?
  - Data quality affects the final assembly
  - Contamination
  - Preparation biases and errors
  - Missing data
  - Difficulty assessment

Data Integrity

- Ensure all your data is there.
  - Many tools cannot tell if data is complete
  - File checksums ensure data integrity
    - MD5

      -
        ```
        823fc8b0ca72c6e9bd8c5dcb0a66ce9b    file1.fastq.gz
        ```

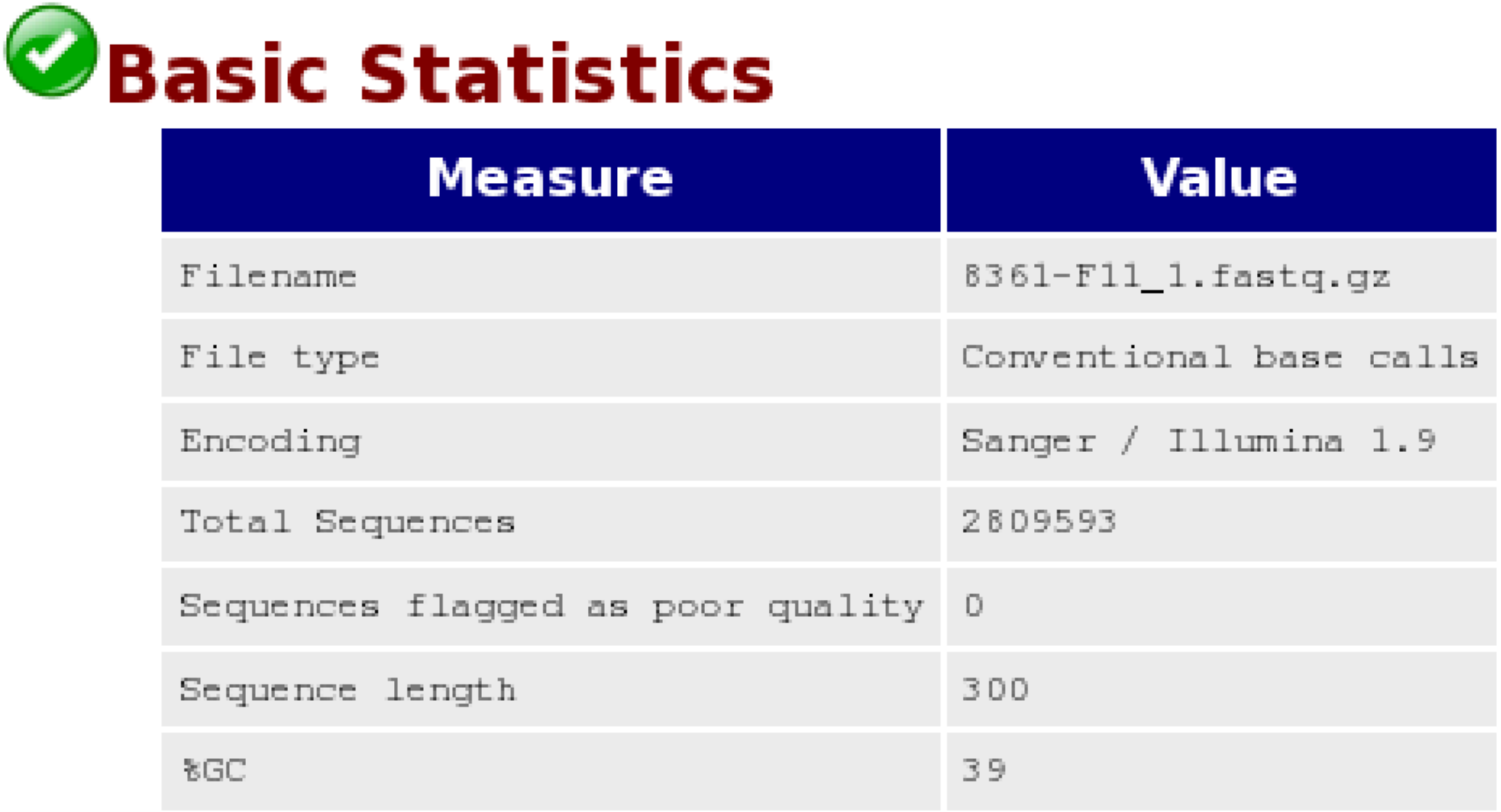
      -
        ```
        $ md5sum -c md5.txt
        file1.fastq.gz: OK
        file2.fastq.gz: OK
        file3.fastq.gz: FAILED
        md5sum: WARNING: 1 of 3 computed checksums did NOT match
        ```
  - Calculate checksum before transfer, check after.

Do I have enough data?

- What is my expected genome size?
- What depth of coverage should I expect?
  - Illumina:
    - 100x coverage in total
  - PacBio:
    - 70x coverage in total from subreads
    - At least 30x coverage of reads >10kb
- Coverage = Number of bases/Genome Size

- Check your reports from the sequencing provider
  - Illumina: FastQC / MultiQC / Sissyphus
  - PacBio: SMRT portal report

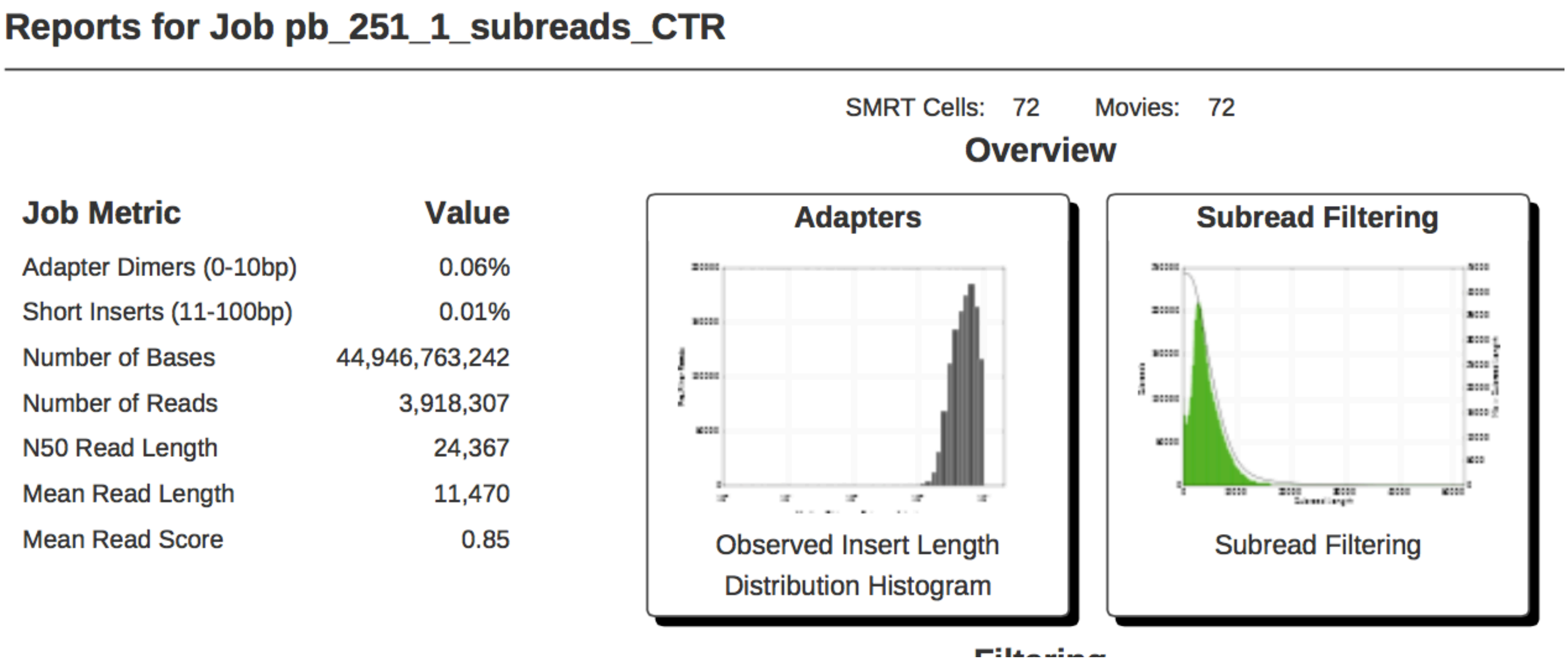
FastQC



✓ Basic Statistics

Measure	Value
Filename	8361-F11_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2809593
Sequences flagged as poor quality	0
Sequence length	300
%GC	39

SMRT Portal



Reports for Job pb_251_1_subreads_CTR

SMRT Cells: 72 Movies: 72

Overview

Job Metric	Value
Adapter Dimers (0-10bp)	0.06%
Short Inserts (11-100bp)	0.01%
Number of Bases	44,946,763,242
Number of Reads	3,918,307
N50 Read Length	24,367
Mean Read Length	11,470
Mean Read Score	0.85

Adapters

Observed Insert Length Distribution Histogram

Subread Filtering

Subread Filtering

Filtering

Calculating data quantity

- Third party scripts

- Command line calculation (my favourite way)
  - Can use Seqtk to convert and filter on read length

  -
    ```
    zcat *.fastq.gz | seqtk seq -A -L 10000 - | grep -v
    “^>” | tr -dc “ACGTNacgtn” | wc -m
    ```
    - zcat (concatenates the compressed fastq files into one stream)
    - seqtk (converts to fasta format and drops reads less than 10k)
    - grep (-v excludes lines starting with “>”, i.e. fasta headers)
    - tr (-dc removes any characters not in set “ACGTNacgtn”)
    - wc (-m counts characters)

Calculating data quantity

SciLifeLab

- How much data is too much data?
  - Greater than 200X coverage is considered extreme.

- Why is too much data bad?
  - Increased computation time and resources
  - Errors begin to compound and start to look like real data.
  - Assemblies become more fragmented and inaccurate.

- How should I subsample?
  - Illumina: Use a random fraction of the reads maintaining read pairing.
    - E.g. Use the same seed (-s) and give the fraction (0.1) in Seqtk.

      ```
      seqtk sample -s100 read1.fq 0.1 > sub1.fq
      seqtk sample -s100 read2.fq 0.1 > sub2.fq
      ```
  - PacBio: Filter out shorter length reads
    - E.g. Keep reads greater than 5kb:

      ```
      seqtk seq -L 5000 reads.fq.gz > reads_5kbplus.fq
      ```

Sidebar - Unix notes

- Sequence files are best kept compressed.

- **`zcat`** prints **`gzip`** compressed files to the screen.
- **`bzcat`** prints **`bzip2`** compressed files to the screen.

- **`file`** tests the type of file.

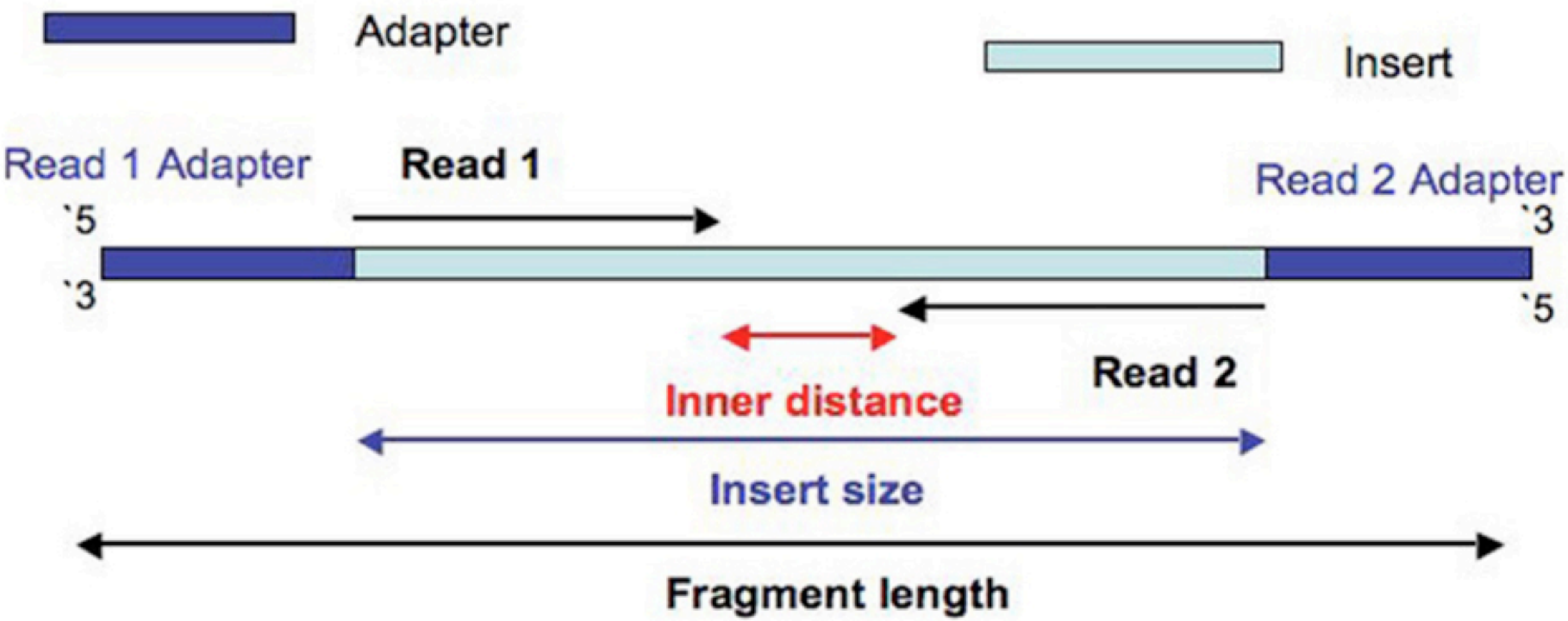
```
$ file bacteria_R1.fastq.gz
bacteria_R1.fastq.gz: gzip compressed data, from NTFS
filesystem (NT), max speed
```

- Try **`man <command>`** or **`<command> -h/--help`** to understand how unix commands work
  - Press **`q`** to exit the **`man`** page

Illumina Specific Quality Checks And Clean Up

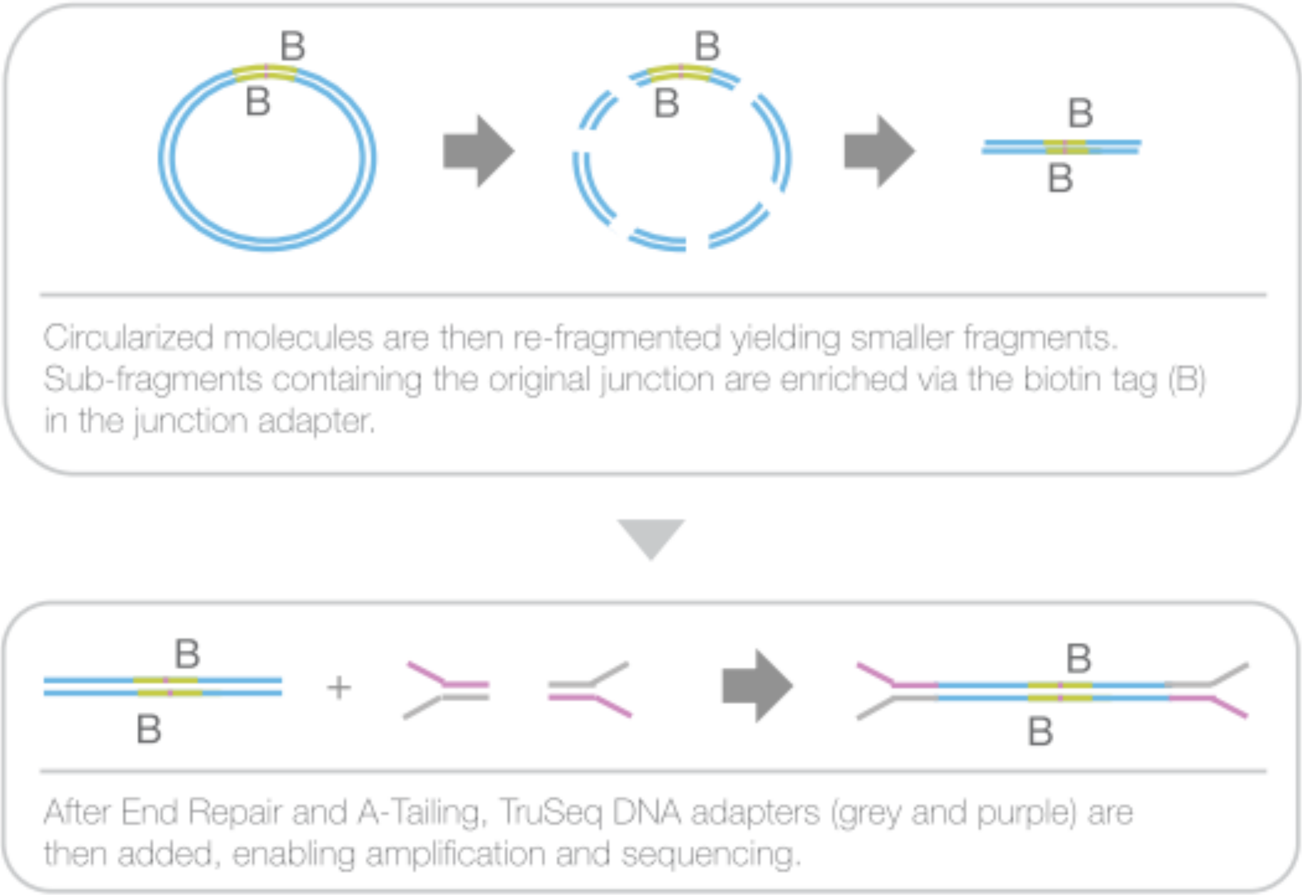
Data Recap - Illumina

- Paired end Illumina library



Data Recap - Illumina

- Mate pair Illumina library



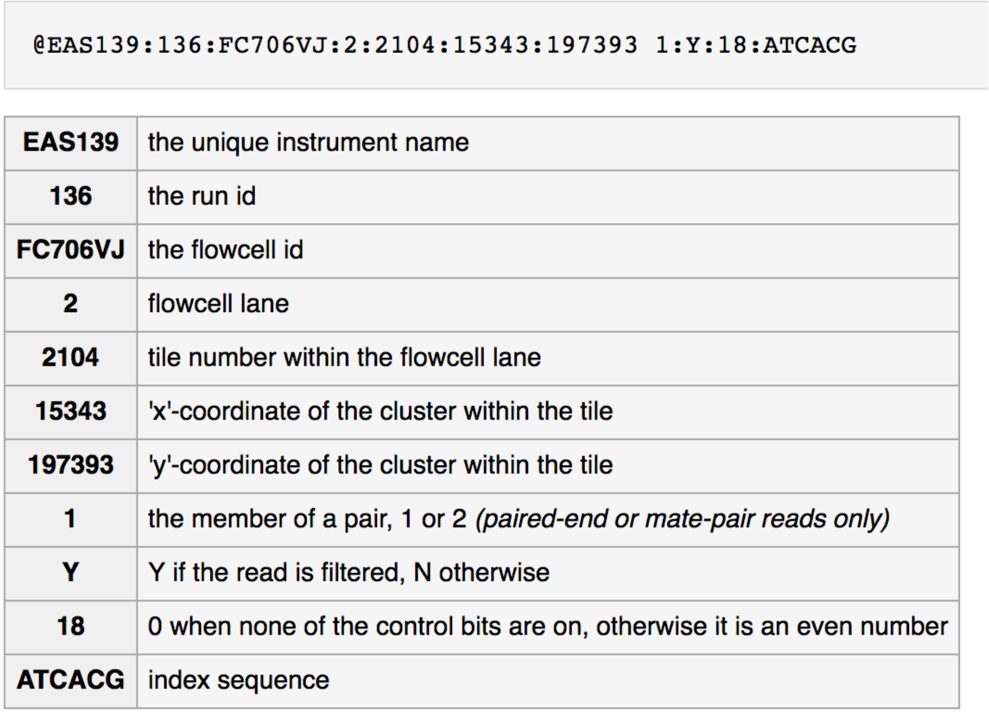
Circularized molecules are then re-fragmented yielding smaller fragments. Sub-fragments containing the original junction are enriched via the biotin tag (B) in the junction adapter.

After End Repair and A-Tailing, TruSeq DNA adapters (grey and purple) are then added, enabling amplification and sequencing.

Format Check

- Check the format
  - **$ zcat file1.fastq.gz | head**

```
@HWI-ST486:212:D0C8BACXX:6:1101:2365:1998 1:N:0:ATTCCT
CTTATCGGATCGATCCCAGTTTGGGCTTGTAAACGGTGAATCCTCAAAGACCACCAATGTTG
+
CCCFFFFFHHHHHJJJJJHIJIIJGGJGFEGIGHIBFGHJIJIICHIIIDHGGIGIGHEFG
@HWI-ST486:212:D0C8BACXX:6:1101:2365:1998 2:N:0:ATTCCT
TAACCGAGCAAACAAAAGTTGGTTGTCACAAATTGTAATGACCTGATTAAACTTGATTTTTT
+
CCCFFFFFHHHHHJIIIJHIJJHIJJJJJJJJJJIJJJIJJJJIIIJJIJJJJGIJJJJH
```



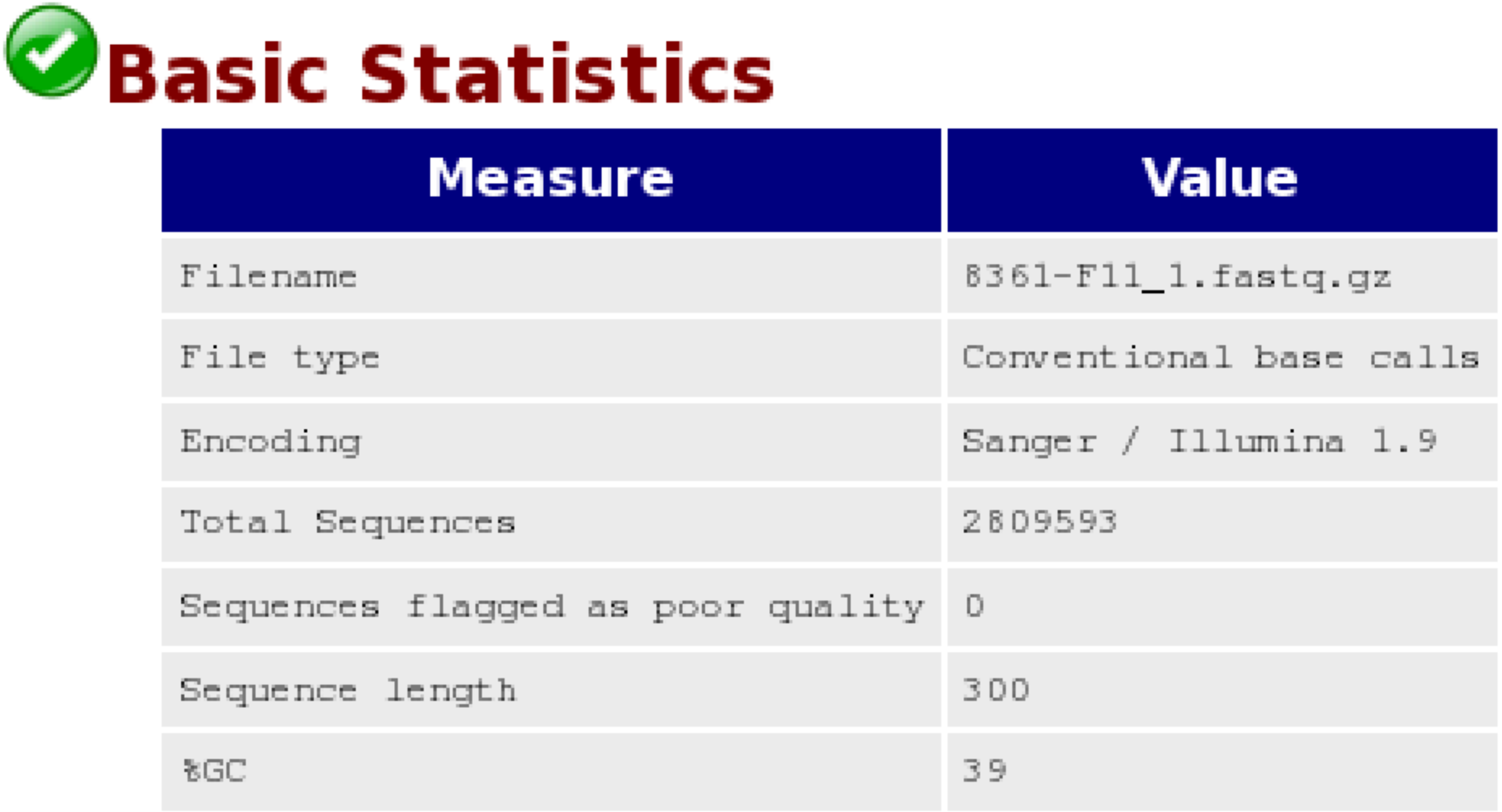
```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

FastQC

- What does it tell you?
  - Total read pairs
  - Sequence length
  - Quality Score Encoding
  - Average GC%
  - Base quality along the read
  - Nucleotide % along the read
  - Sequence GC content
  - Duplication %
  - Adapter content

FastQC

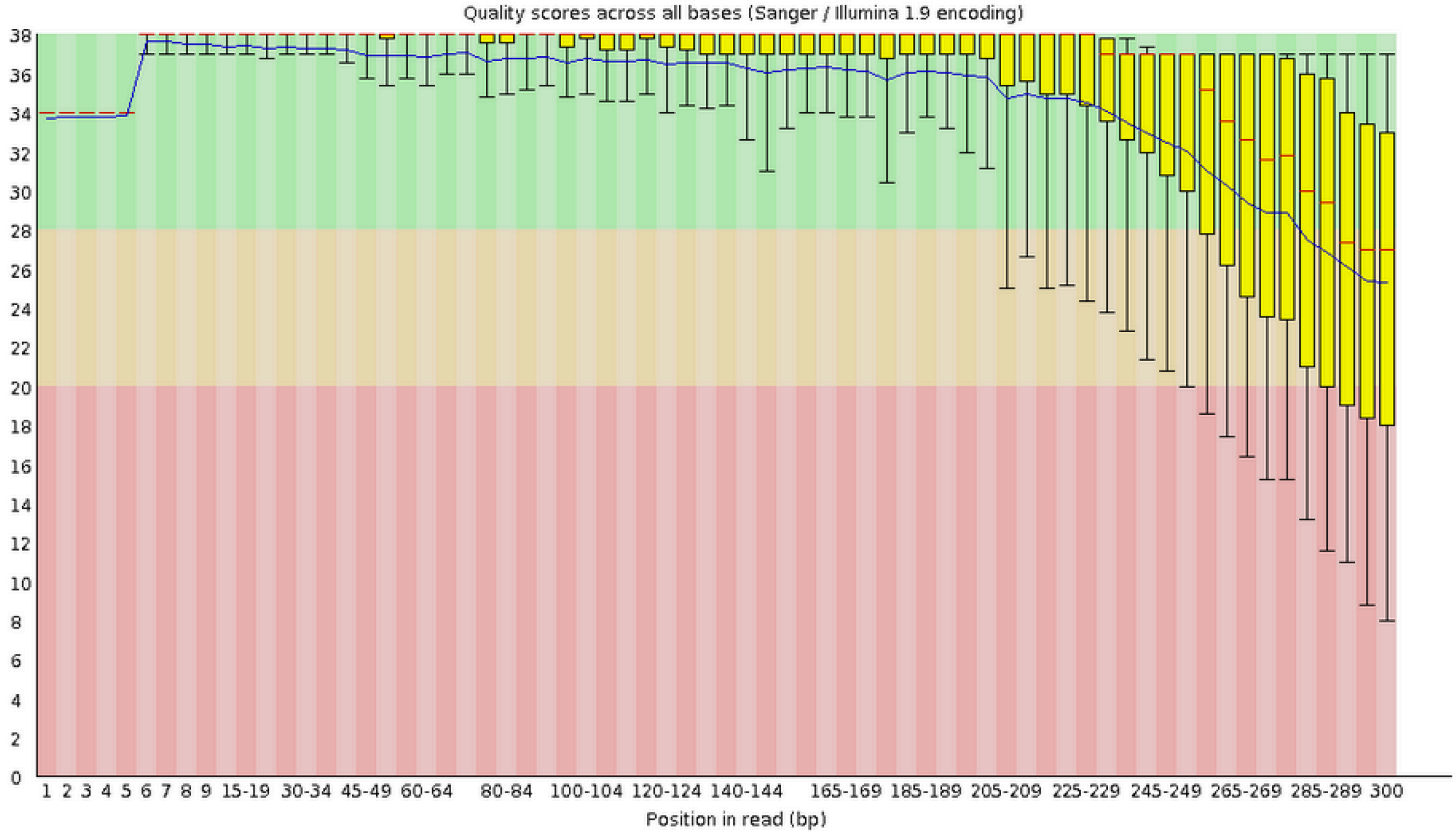


Basic Statistics

Measure	Value
Filename	8361-F11_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2809593
Sequences flagged as poor quality	0
Sequence length	300
%GC	39

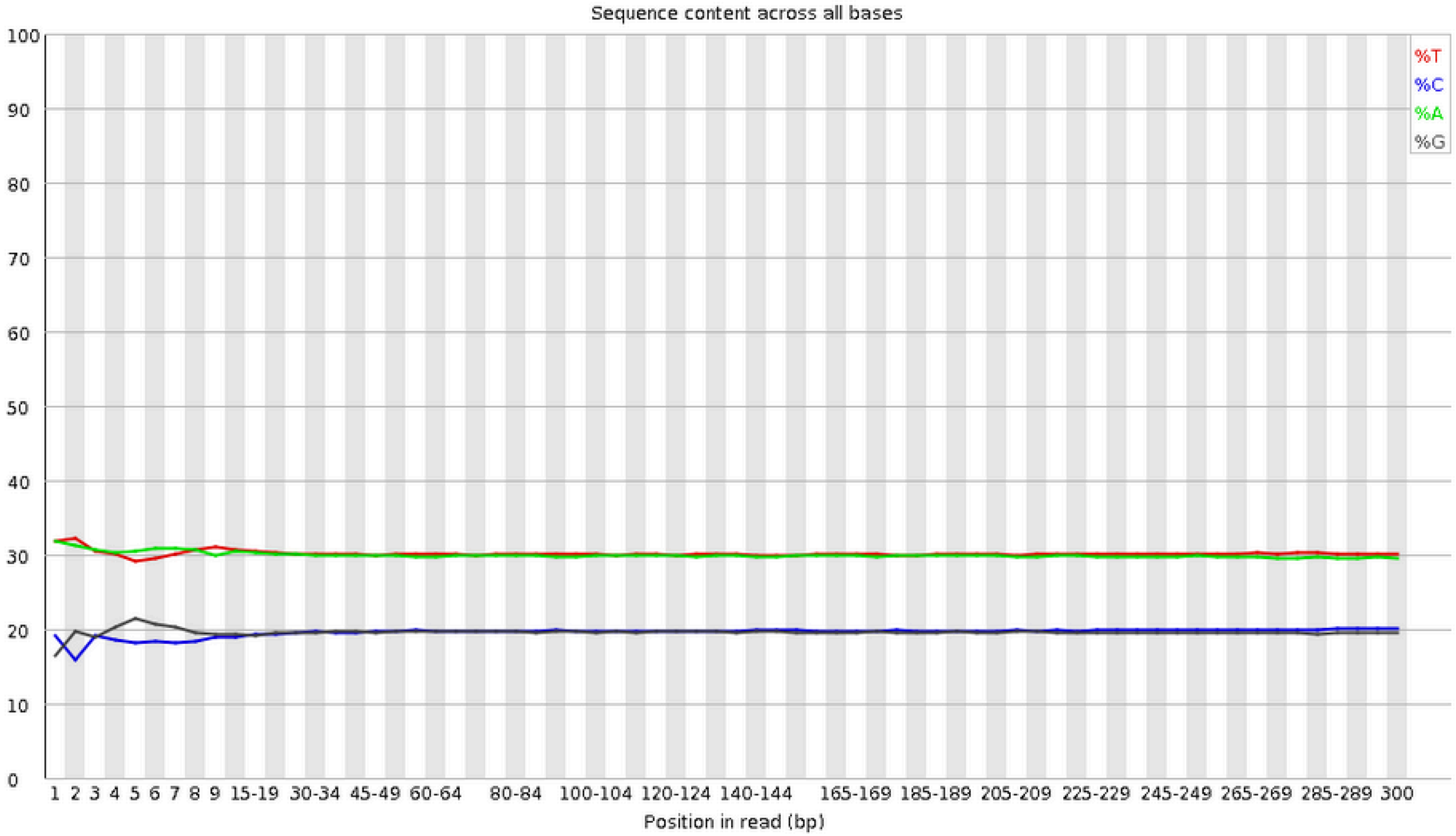
FastQC

✓ Per base sequence quality

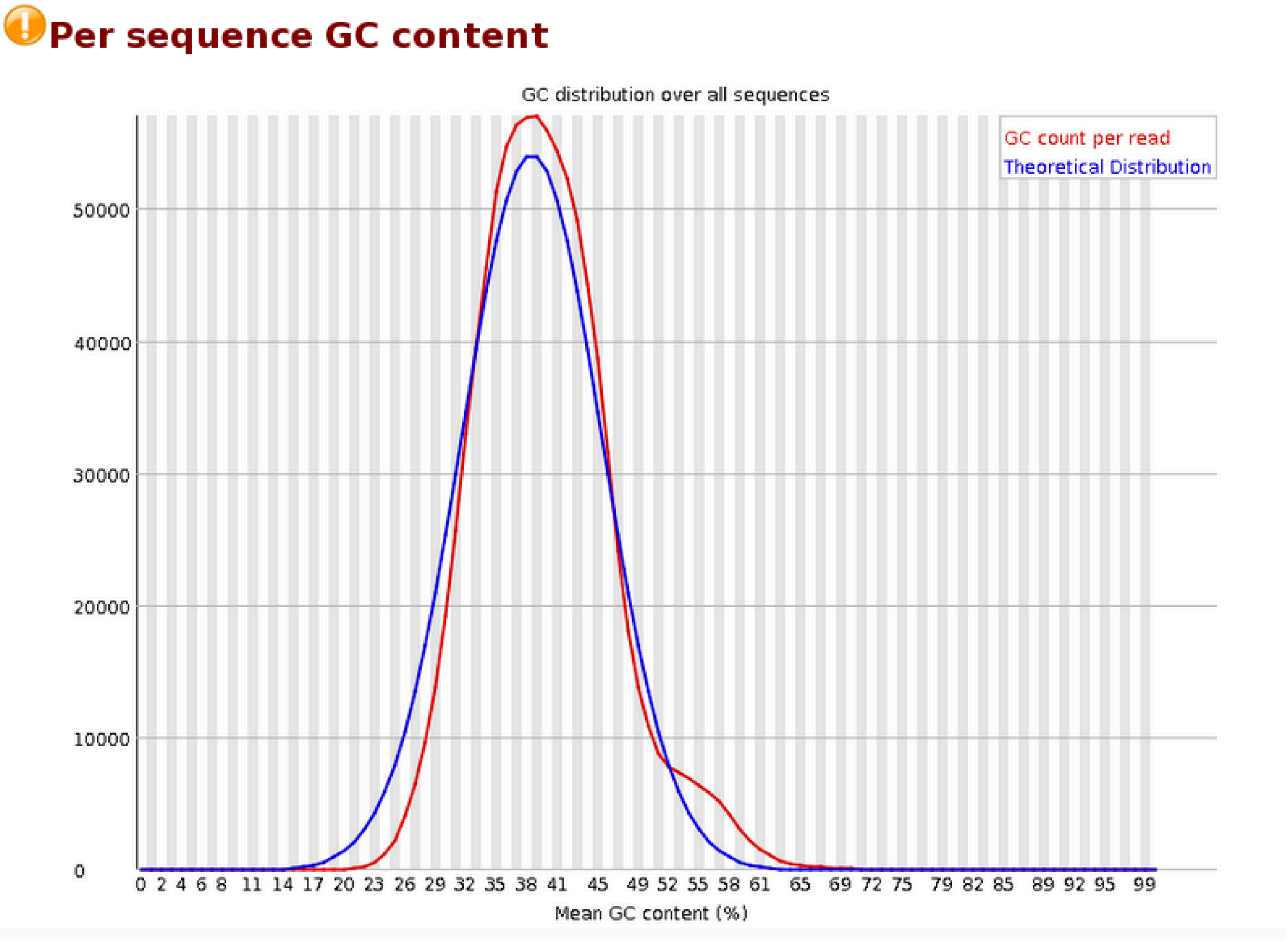


FastQC

✓ Per base sequence content

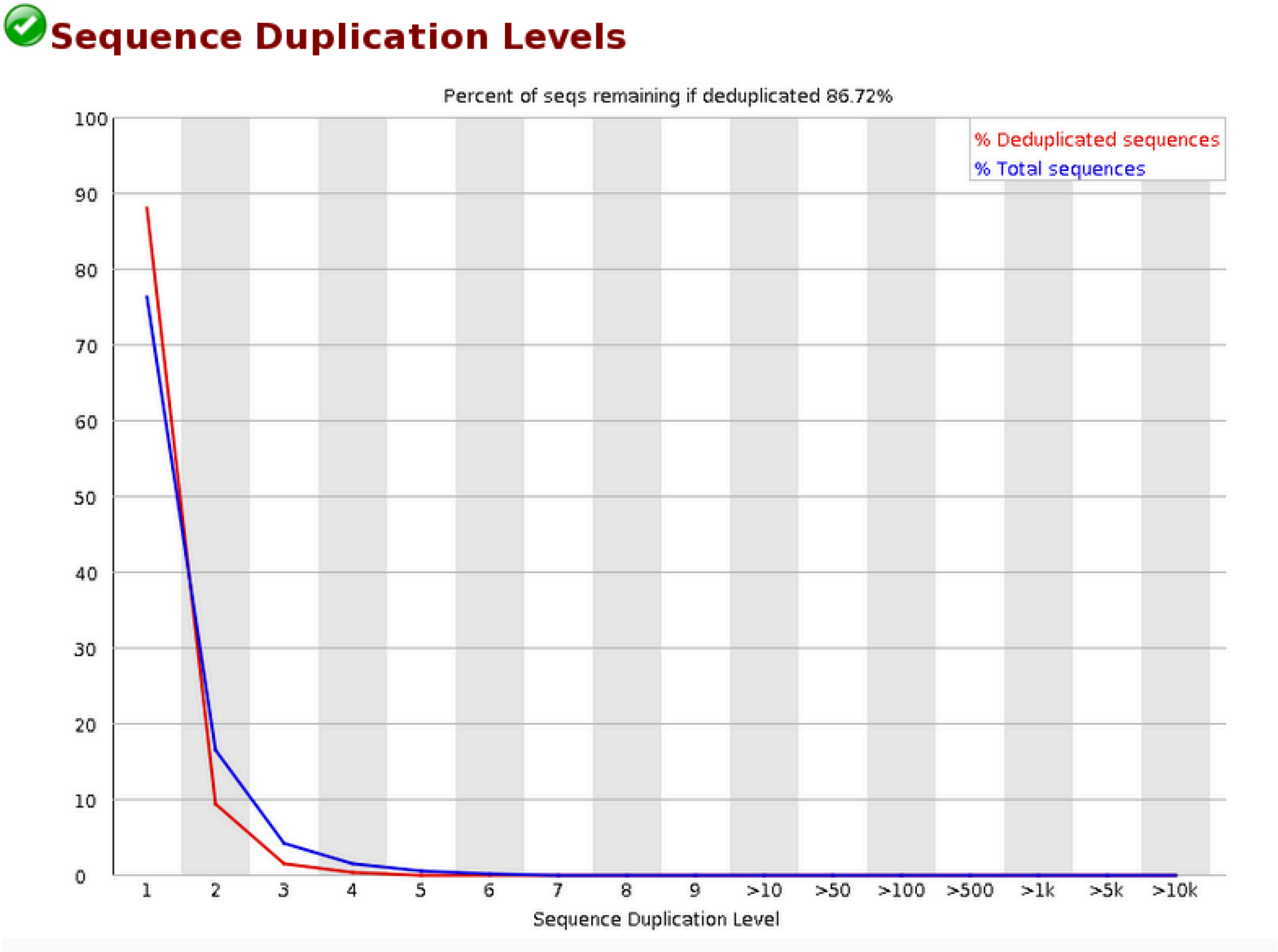


FastQC



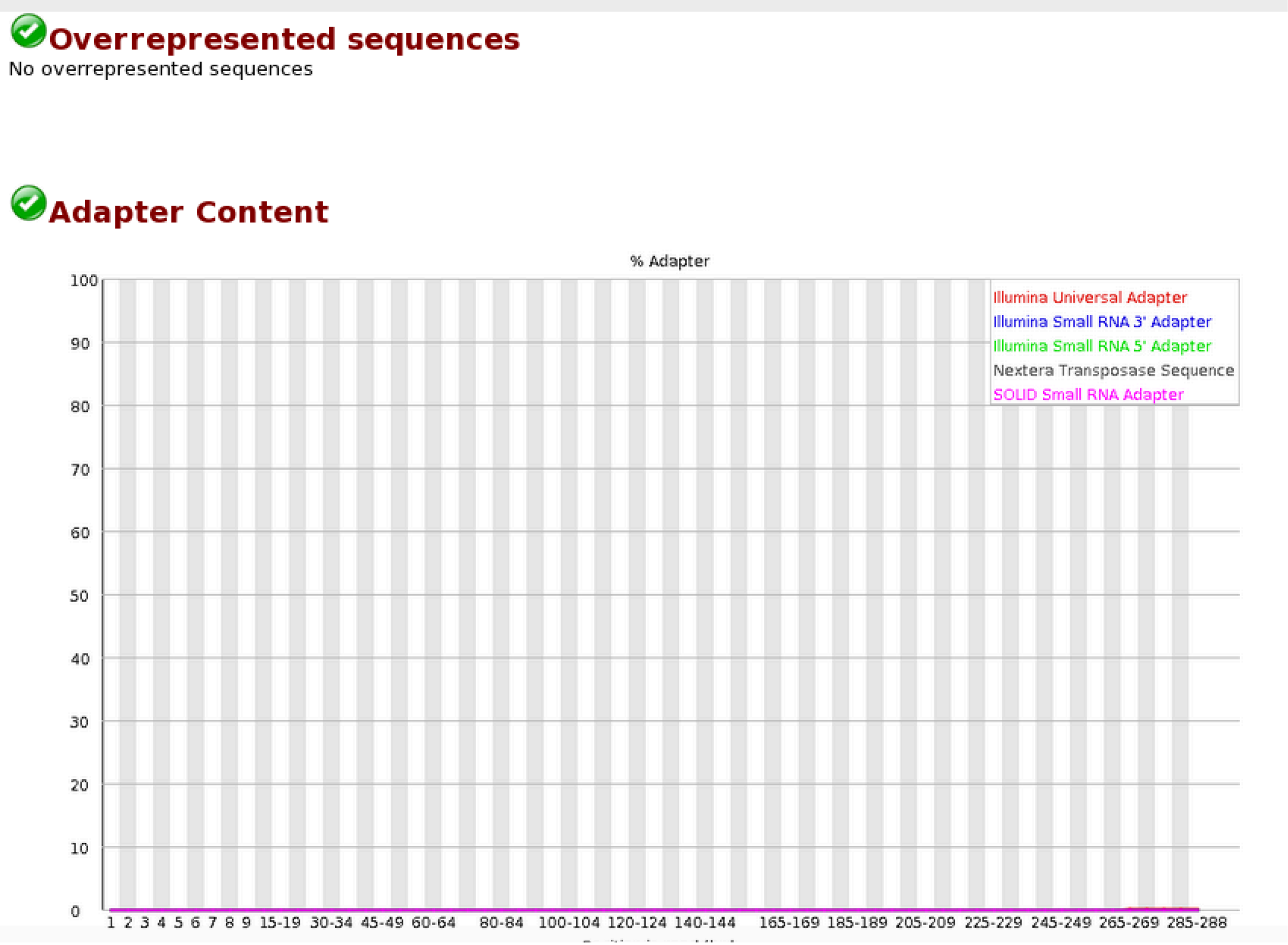
Per sequence GC content

FastQC



Sequence Duplication Levels

FastQC



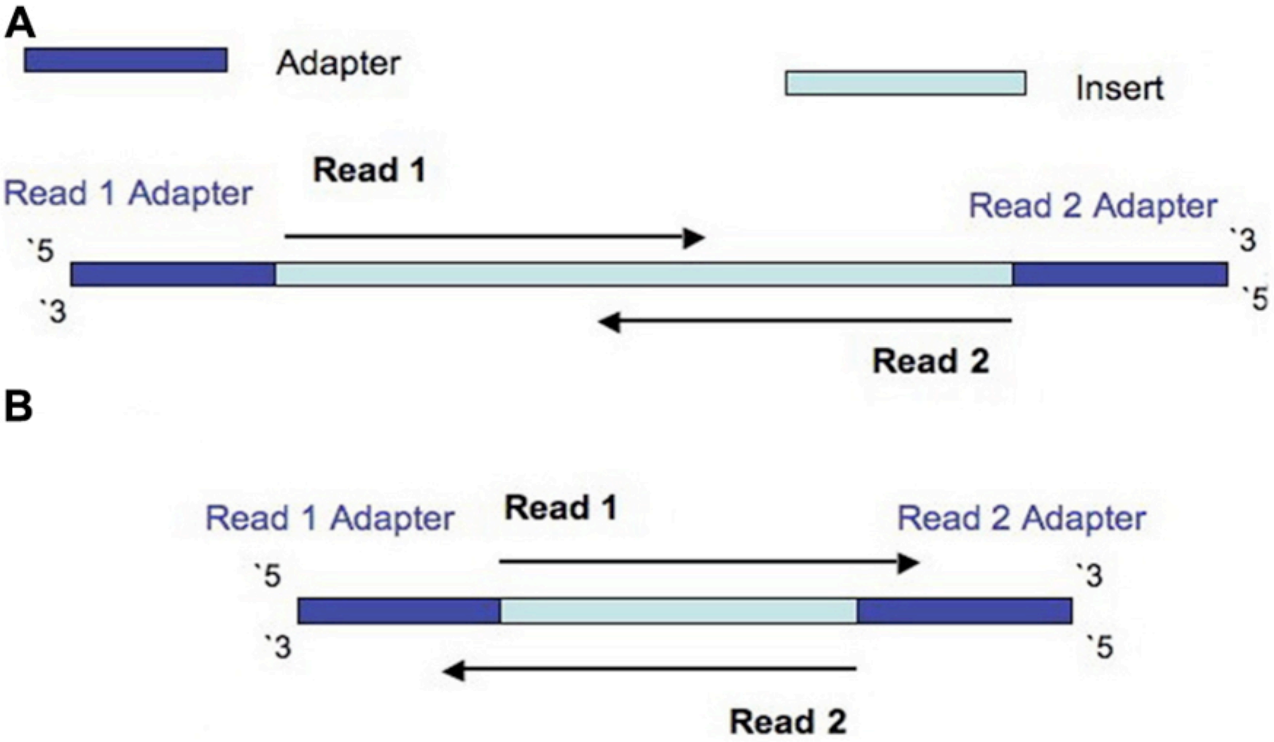
✓ Overrepresented sequences

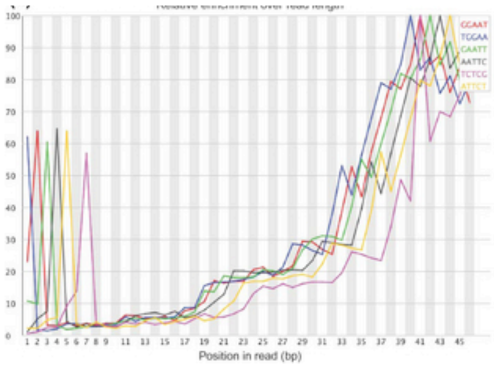
No overrepresented sequences

✓ Adapter Content

Trimming reads

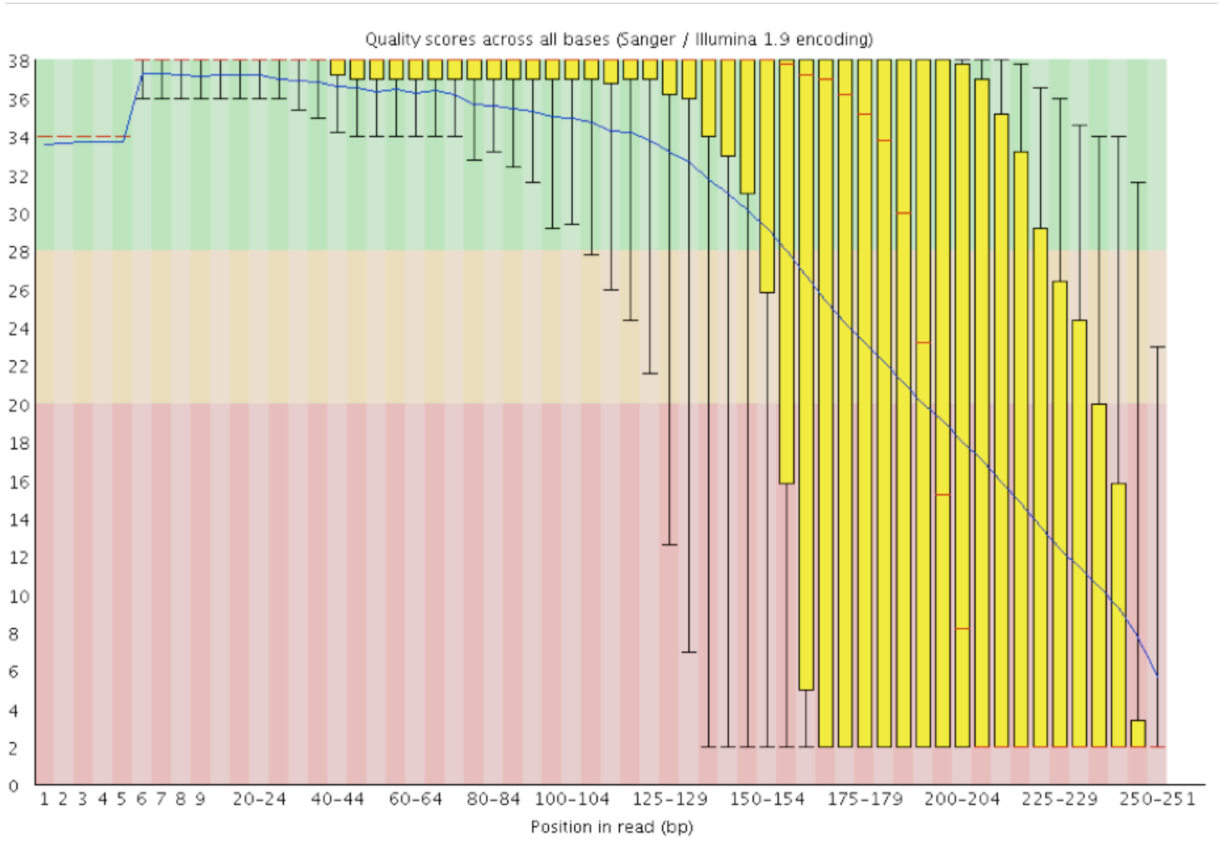
- Why trim reads?
  - Remove adapter read through.





Trimming reads

- Why trim reads?
  - Remove poor quality reads



Trimming reads

- Many tools available
  - Trimmomatic
  - CutAdapt
  - AlienTrimmer
  - Sickle
  - Trim Galore
  - Scythe
  - Prinseq
  - …

- **Warning:** Some assemblers expect untrimmed input
  - Allpaths-LG
  - Mira

Duplication Removal

- Why do duplicates arise?
  - Optical duplicates
  - PCR duplicates

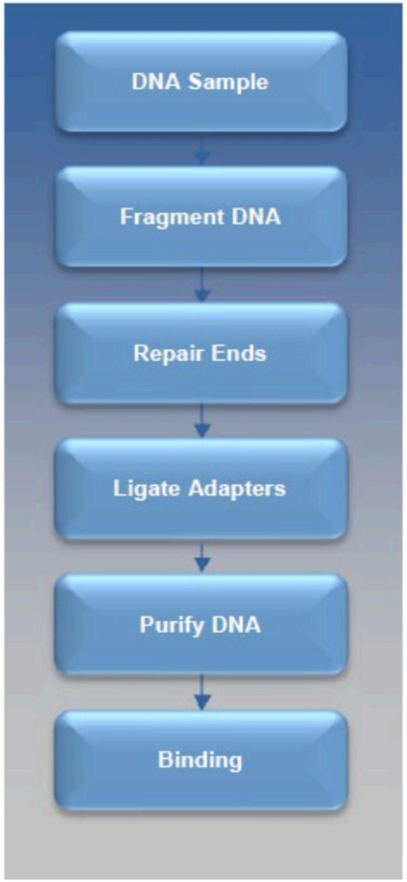
- Why are duplicates bad?
  - Poor overlap information
  - Increased variance of coverage
  - Increased computation time and resources

- How to remove duplicates:
  - Prinseq
  - FastUniq
  - ParDRe
  - …

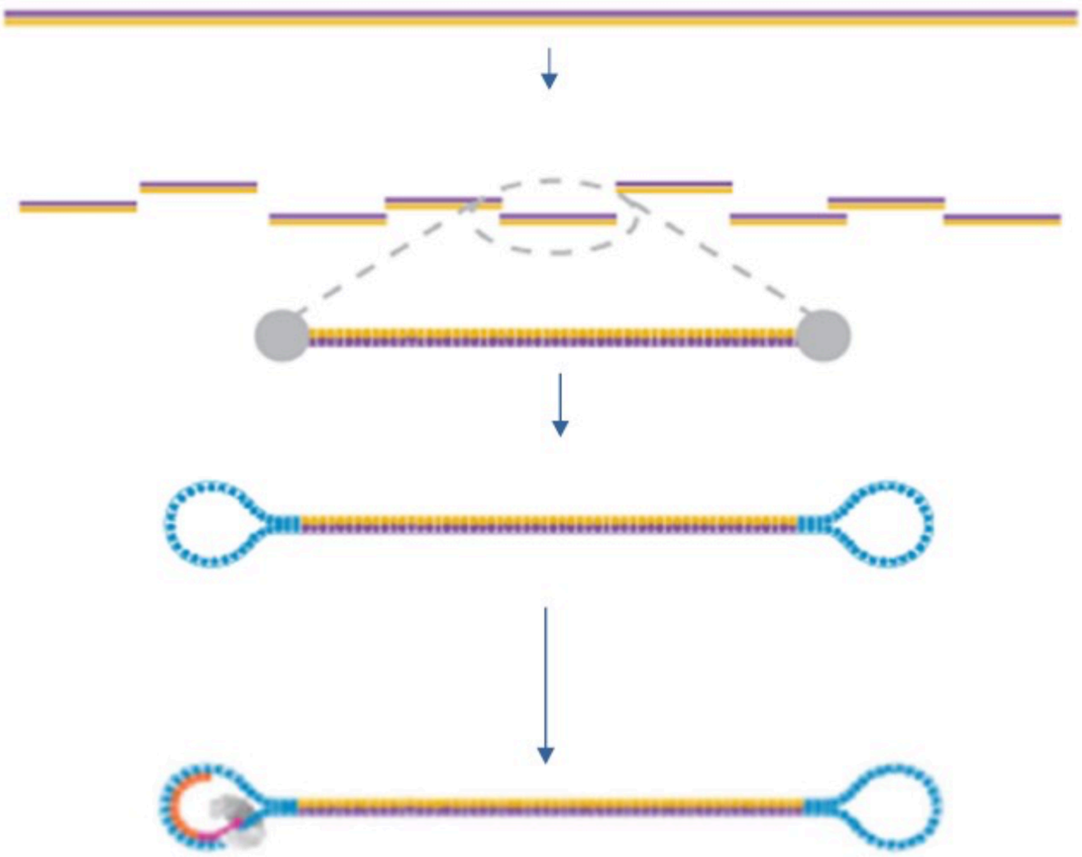
PacBio Specific Quality Checks And Clean Up

PacBio Sequencing

Sample Preparation



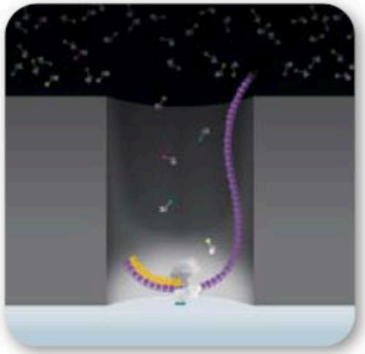
Building of the SMRTbell™



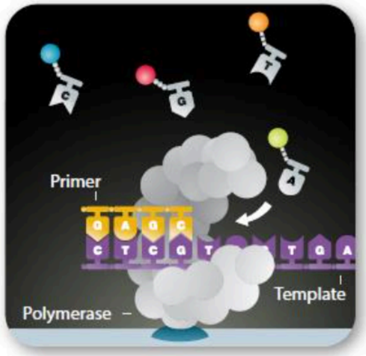
PacBio Sequencing

SMRT® Cells

Zero-Mode Waveguides

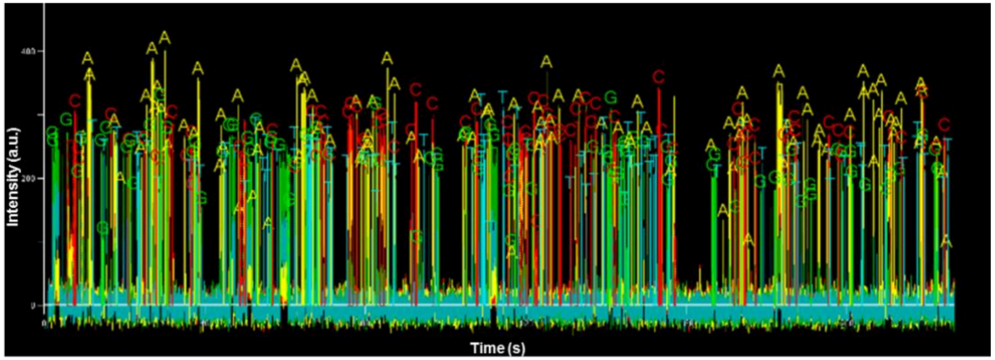


Phospholinked Nucleotides

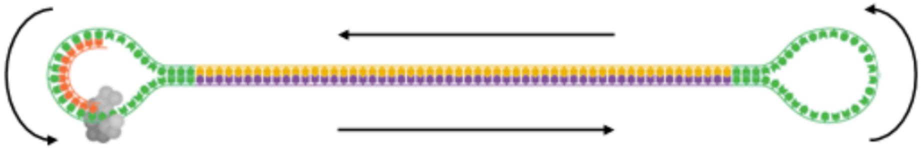


PacBio® RS II

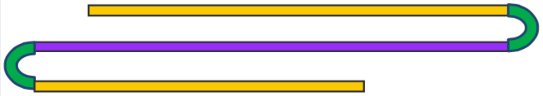
Trace



PacBio Sequencing



SMRTbell™ Template



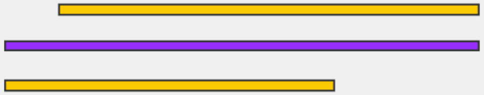
Polymerase Read

Definition:

- Sequence of nucleotides incorporated by polymerase while reading a template
- Includes adapters
- Often called “read”
- Includes adapters
- 1 molecule, 1 pol. read

Purpose:

- QC of instrument run
- Benchmarking



Subread

Definition:

- Single pass of template
- Adapters removed
- 1 molecule, ≥1 subreads

Unique data:

- Kinetic measurements
- Rich QVs

Purpose:

- For subsequent analysis

Read of Insert

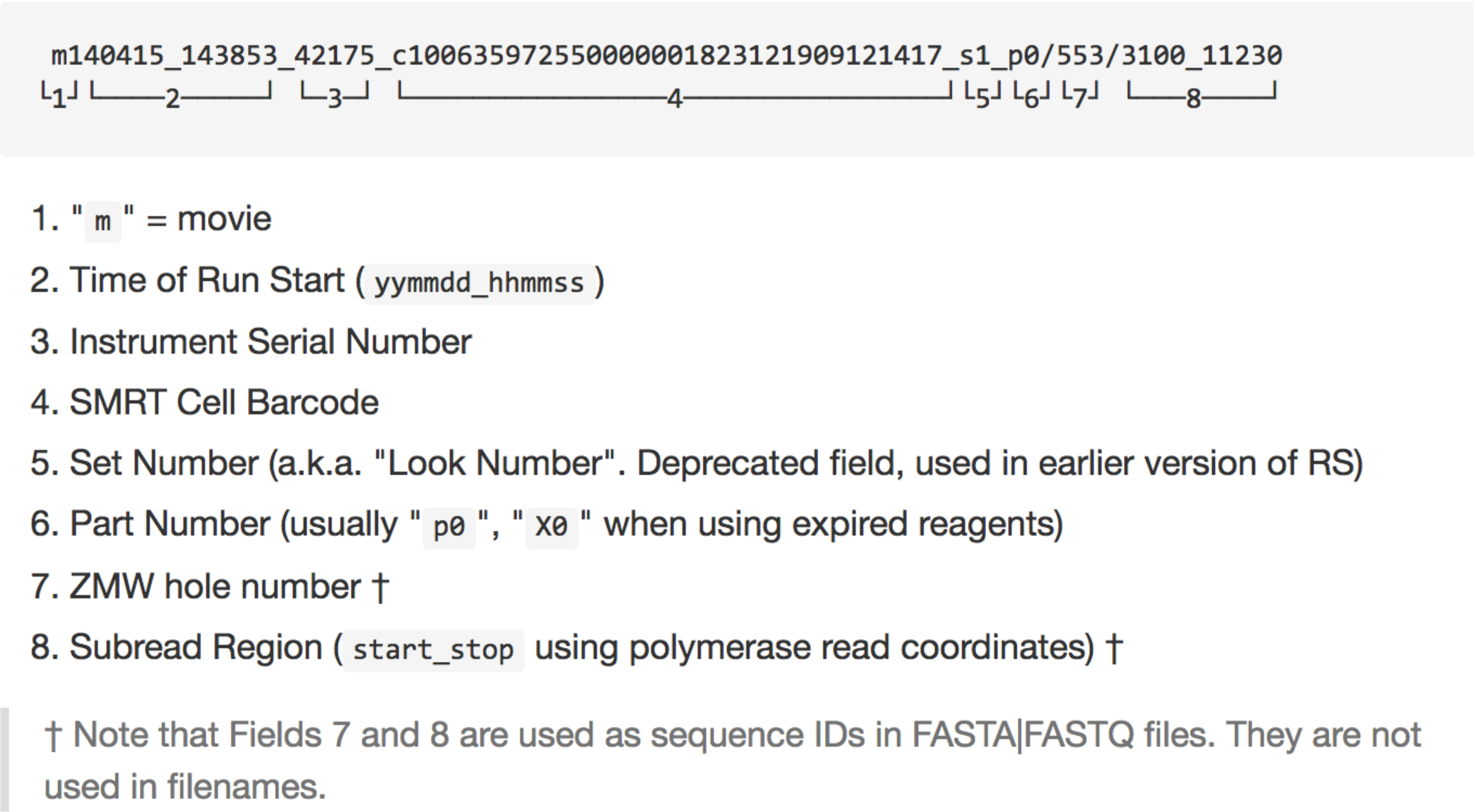
Definition:

- Represents highest-quality single-sequence for an insert, regardless of number of passes
- Generalizes CCS for <2 passes and RQ <0.9
- 1 or more passes
- 1 molecule, 1 read

Purpose:

- For Library QC
- For subsequent analysis

PacBio Sequencing

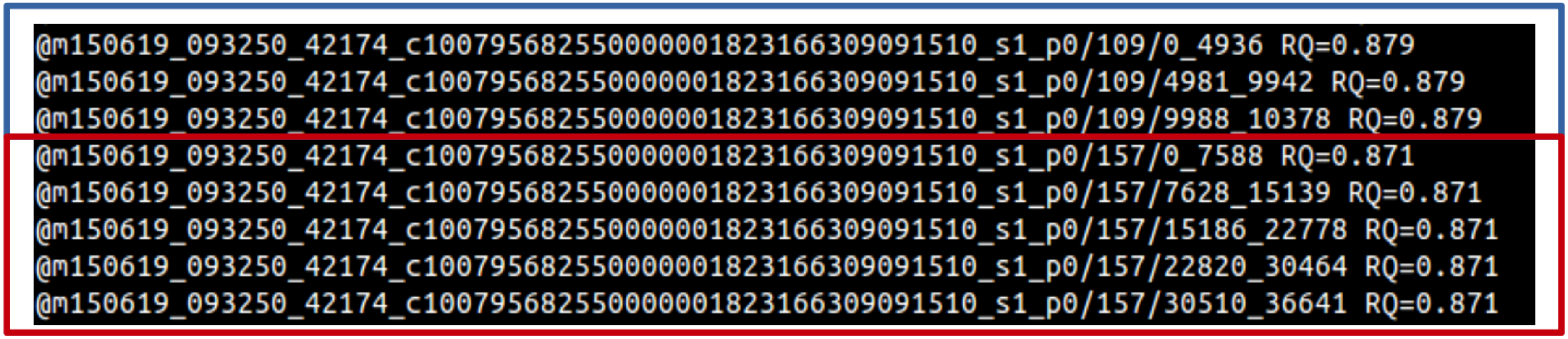


```
 m140415_143853_42175_c100635972550000001823121909121417_s1_p0/553/3100_11230
└1┘└─────2─────┘ └─3─┘ └───────────────4────────────────┘ └5┘└6┘└7┘ └───8────┘
```

1. "`m`" = movie
2. Time of Run Start (`yymmdd_hhmmss`)
3. Instrument Serial Number
4. SMRT Cell Barcode
5. Set Number (a.k.a. "Look Number". Deprecated field, used in earlier version of RS)
6. Part Number (usually "`p0`", "`X0`" when using expired reagents)
7. ZMW hole number †
8. Subread Region (`start_stop` using polymerase read coordinates) †

> † Note that Fields 7 and 8 are used as sequence IDs in FASTA|FASTQ files. They are not used in filenames.

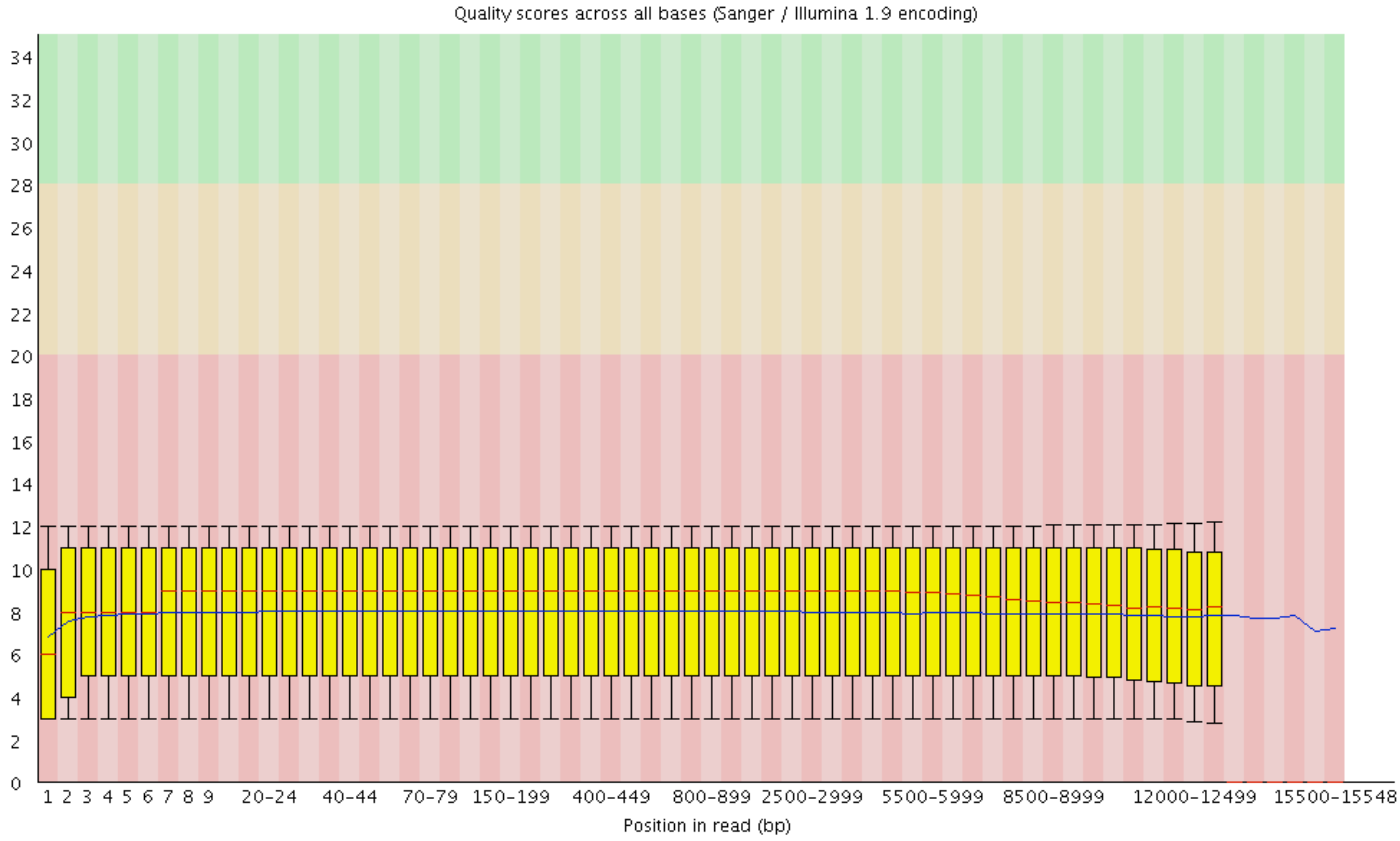
PacBio Sequencing



```
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/109/0_4936 RQ=0.879
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/109/4981_9942 RQ=0.879
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/109/9988_10378 RQ=0.879
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/157/0_7588 RQ=0.871
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/157/7628_15139 RQ=0.871
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/157/15186_22778 RQ=0.871
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/157/22820_30464 RQ=0.871
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/157/30510_36641 RQ=0.871
```

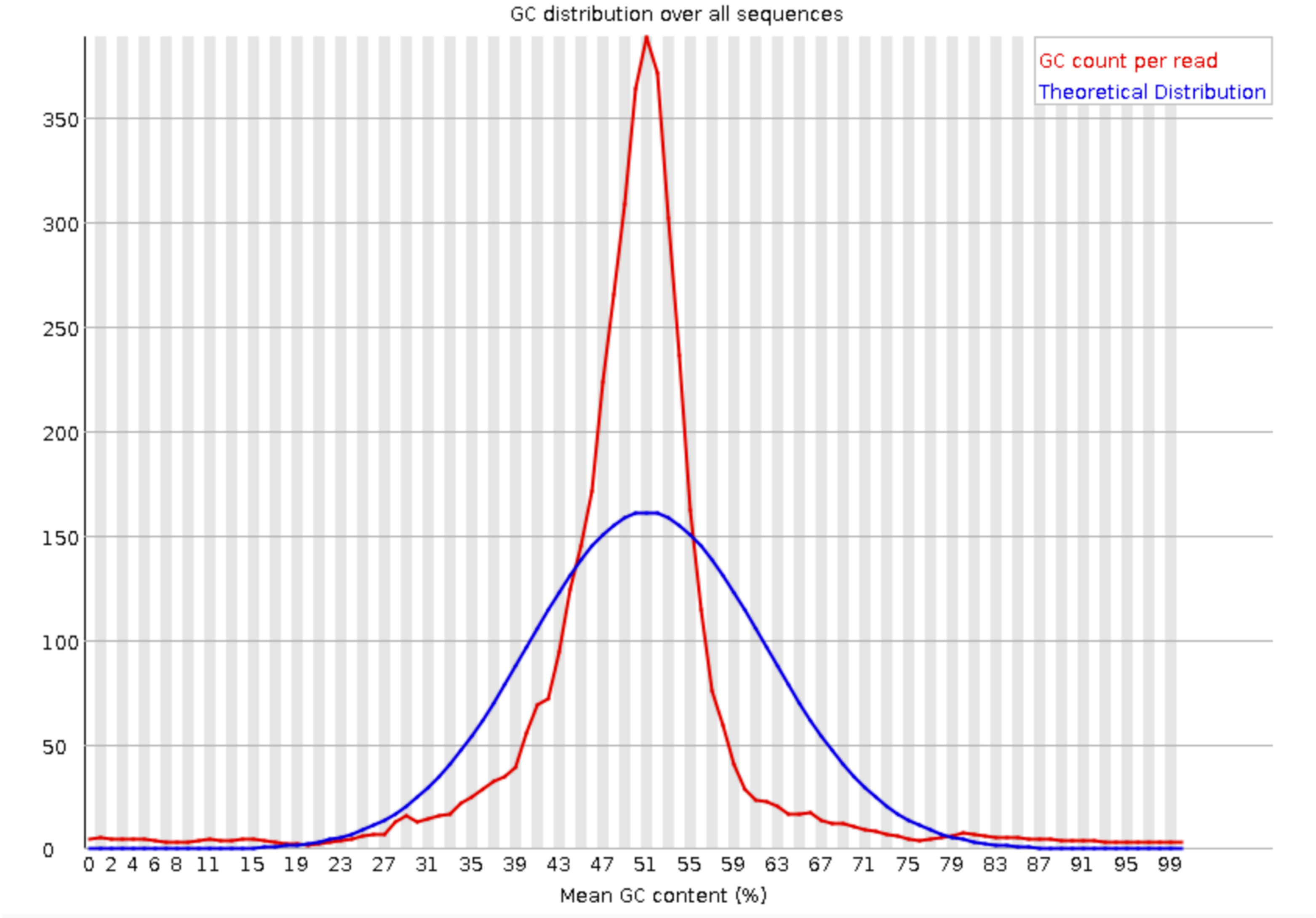
- The subreads fastq file contains all the subreads from a SMRT movie.
- The reads from a ZMW after adapter removal are oriented in the direction forward, reverse, forward, and so on.
- Read Quality (RQ) Assignment: A trained prediction of a read’s mapped accuracy based on its pulse and base file characteristics (peak signal-to-noise ratio, average base QV, interpulse distance, and so on).
- Quality Value (QV): The total probability that the basecall is an insertion or substitution or is preceded by a deletion. QV = -10 * log10(p).

FastQC

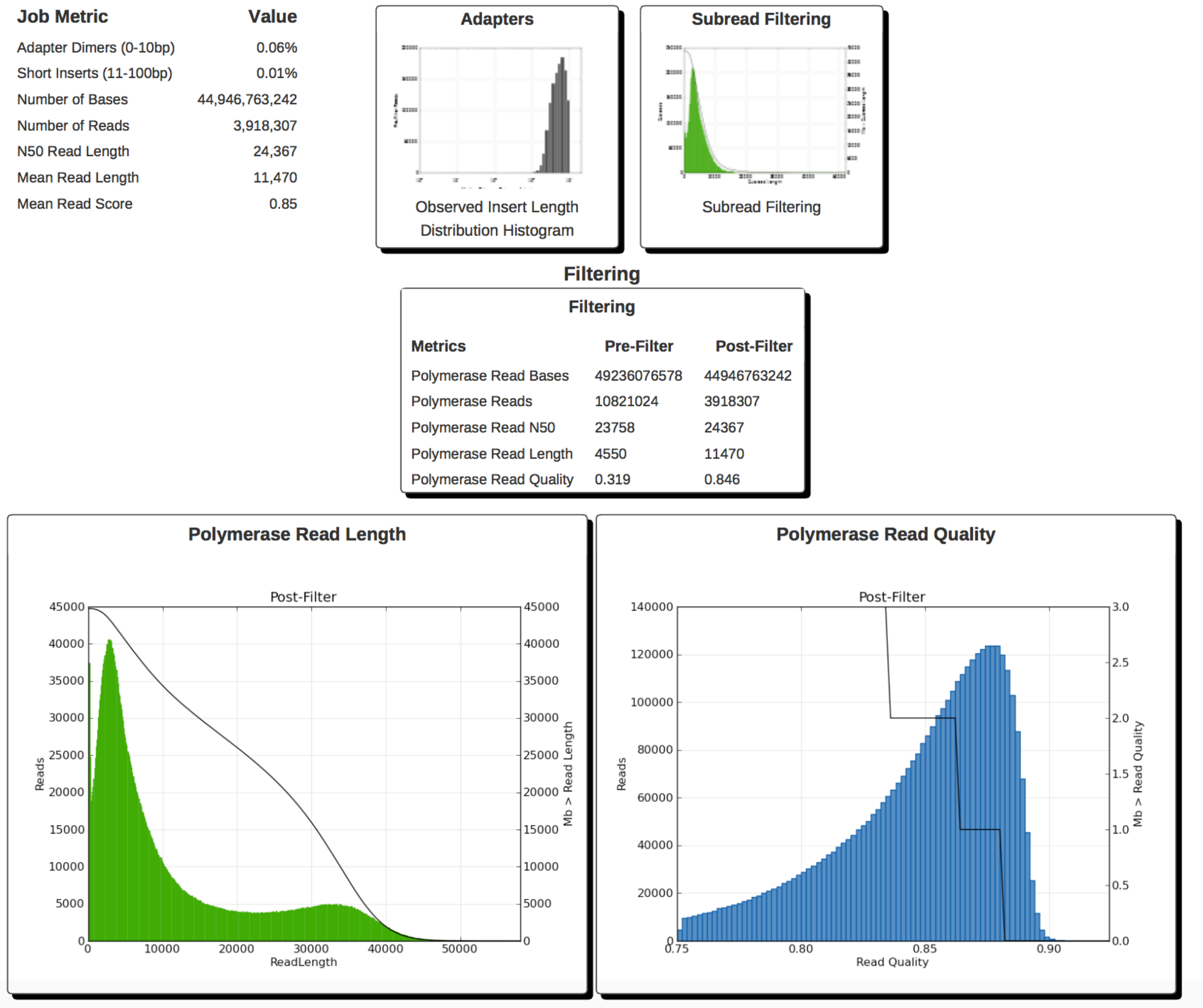


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9
20-24
40-44
70-79
150-199
400-449
800-899
2500-2999
5500-5999
8500-8999
12000-12499
15500-15548
Position in read (bp)

FastQC


GC distribution over all sequences
GC count per read
Theoretical Distribution
350
300
250
200
150
100
50
0
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)

SMRT Portal Report



Job Metric	Value
Adapter Dimers (0-10bp)	0.06%
Short Inserts (11-100bp)	0.01%
Number of Bases	44,946,763,242
Number of Reads	3,918,307
N50 Read Length	24,367
Mean Read Length	11,470
Mean Read Score	0.85

Adapters

Observed Insert Length Distribution Histogram

Subread Filtering

Subread Filtering

Filtering

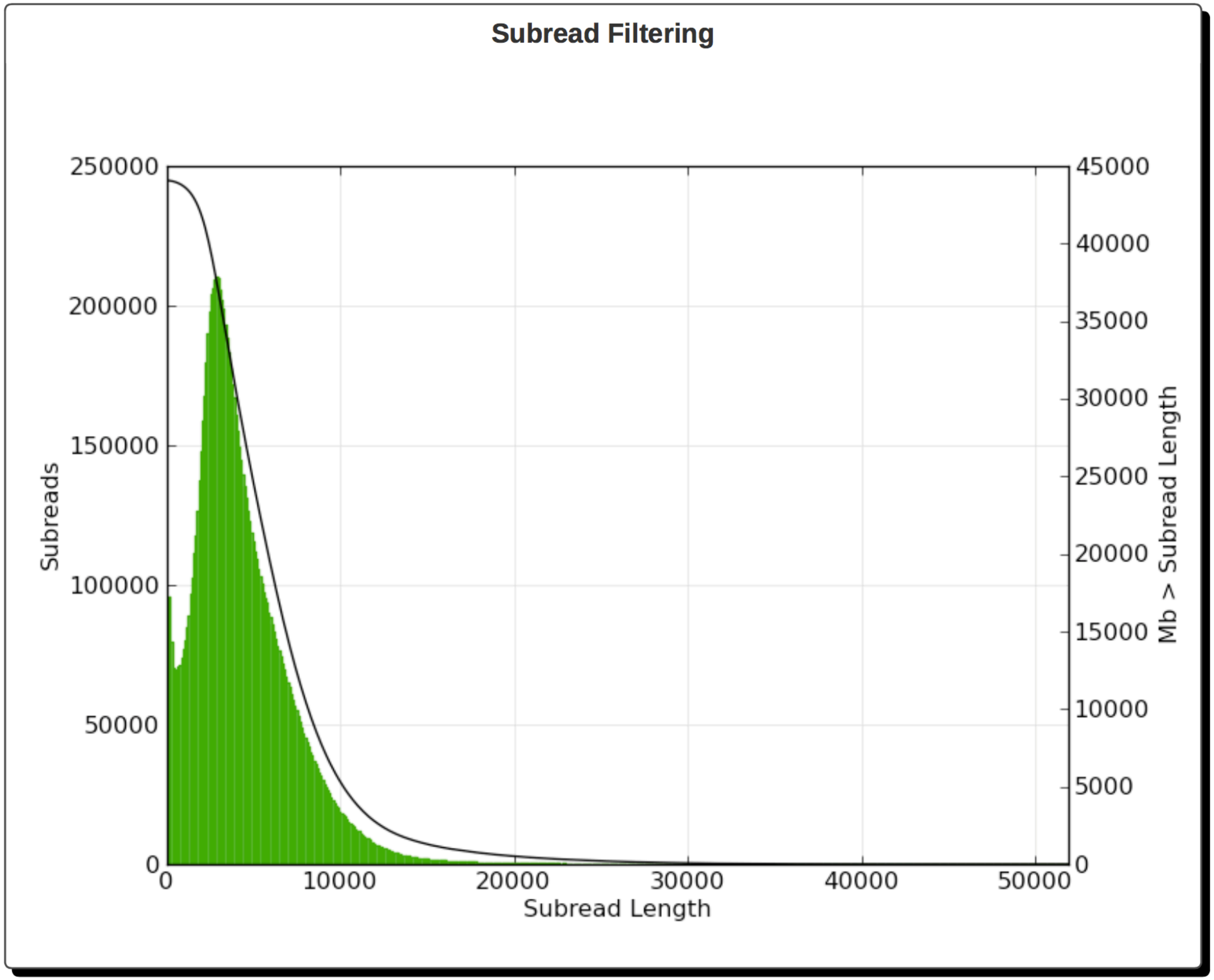
Filtering

Metrics	Pre-Filter	Post-Filter
Polymerase Read Bases	49236076578	44946763242
Polymerase Reads	10821024	3918307
Polymerase Read N50	23758	24367
Polymerase Read Length	4550	11470
Polymerase Read Quality	0.319	0.846

Polymerase Read Length

Polymerase Read Quality

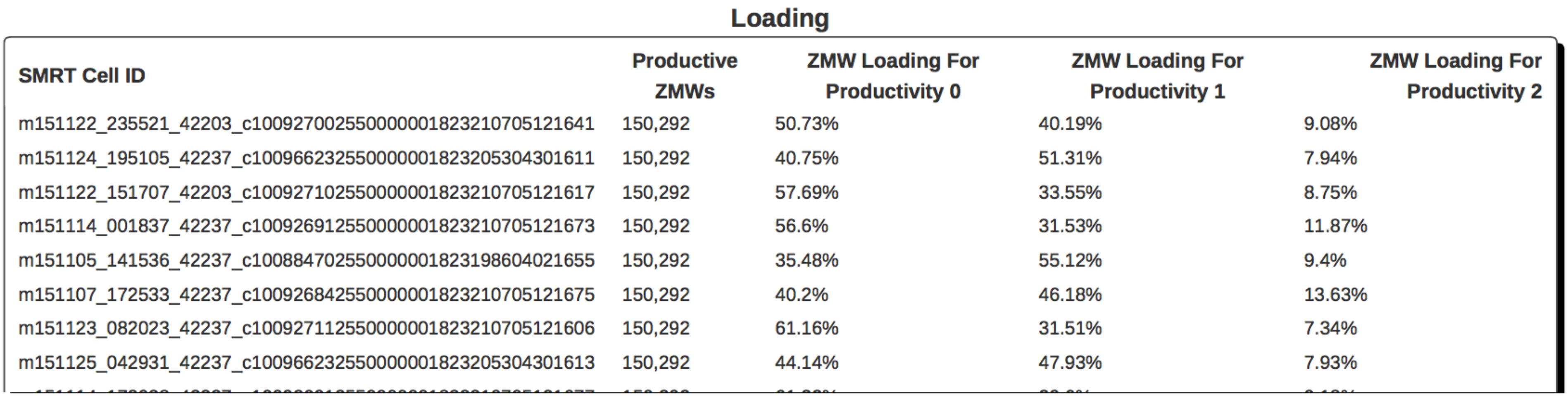
SMRT Portal Report



Adapters

Adapter Dimers (0-10bp)	0.06%
Short Inserts (11-100bp)	0.01%

SMRT Portal Report



Loading

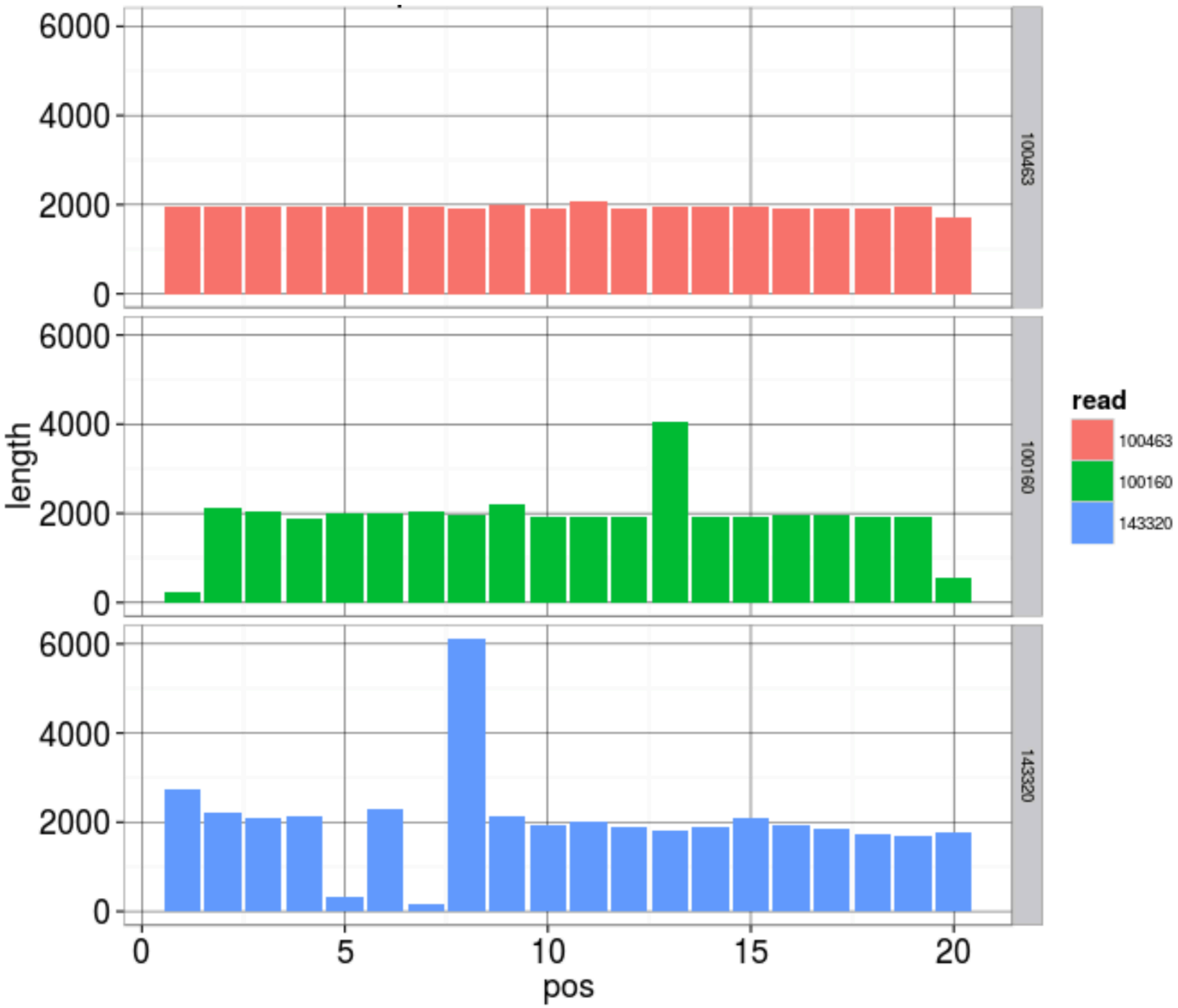
SMRT Cell ID	Productive ZMWs	ZMW Loading For Productivity 0	ZMW Loading For Productivity 1	ZMW Loading For Productivity 2
m151122_235521_42203_c100927002550000001823210705121641	150,292	50.73%	40.19%	9.08%
m151124_195105_42237_c100966232550000001823205304301611	150,292	40.75%	51.31%	7.94%
m151122_151707_42203_c100927102550000001823210705121617	150,292	57.69%	33.55%	8.75%
m151114_001837_42237_c100926912550000001823210705121673	150,292	56.6%	31.53%	11.87%
m151105_141536_42237_c100884702550000001823198604021655	150,292	35.48%	55.12%	9.4%
m151107_172533_42237_c100926842550000001823210705121675	150,292	40.2%	46.18%	13.63%
m151123_082023_42237_c100927112550000001823210705121606	150,292	61.16%	31.51%	7.34%
m151125_042931_42237_c100966232550000001823205304301613	150,292	44.14%	47.93%	7.93%

- SMRT cell loading
  - P0: % of ZMWs that are empty with no polymerase
  - P1: % of ZMWs that are productive and sequencing
  - P2: % of ZMWs that are not P0 or P1 (e.g. unbound polymerase, more than one molecule in a well (overloaded cell).

  - Maximize P1 and minimize P0 + P2.
  - High P0 indicates underloading (too low concentration of molecules)
  - High P2 indicates overloading (too high concentration) or poor prep.

Adapter Misidentification

SMRTbell adapter:
ATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGAT



Up Next

- Sequence quality assessment
  - K-mer analyses
    - Histograms
    - genome size estimation
    - GC plots
    - data set comparision
  - Contamination analyses
  - Mapping based analysis