SciLifeLab

NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

Quality Assessment of sequencing data

elixir

excelerate

Outline

- Why QC?
  - Bad data = bad assembly
    - Partial / missing data?
    - Is there enough data?
    - Did I get what I expect?
    - Can I assemble it?
    - Do I need to change my analysis workflow?
    - Is it the correct type?
    - Are there biases?
    - Is there contamination?
  - Most checks can be made before assembly and assembly validation

- Focus:
  - Illumina data (PE + MP)
  - Pacbio data
  - Nanopore data

First steps

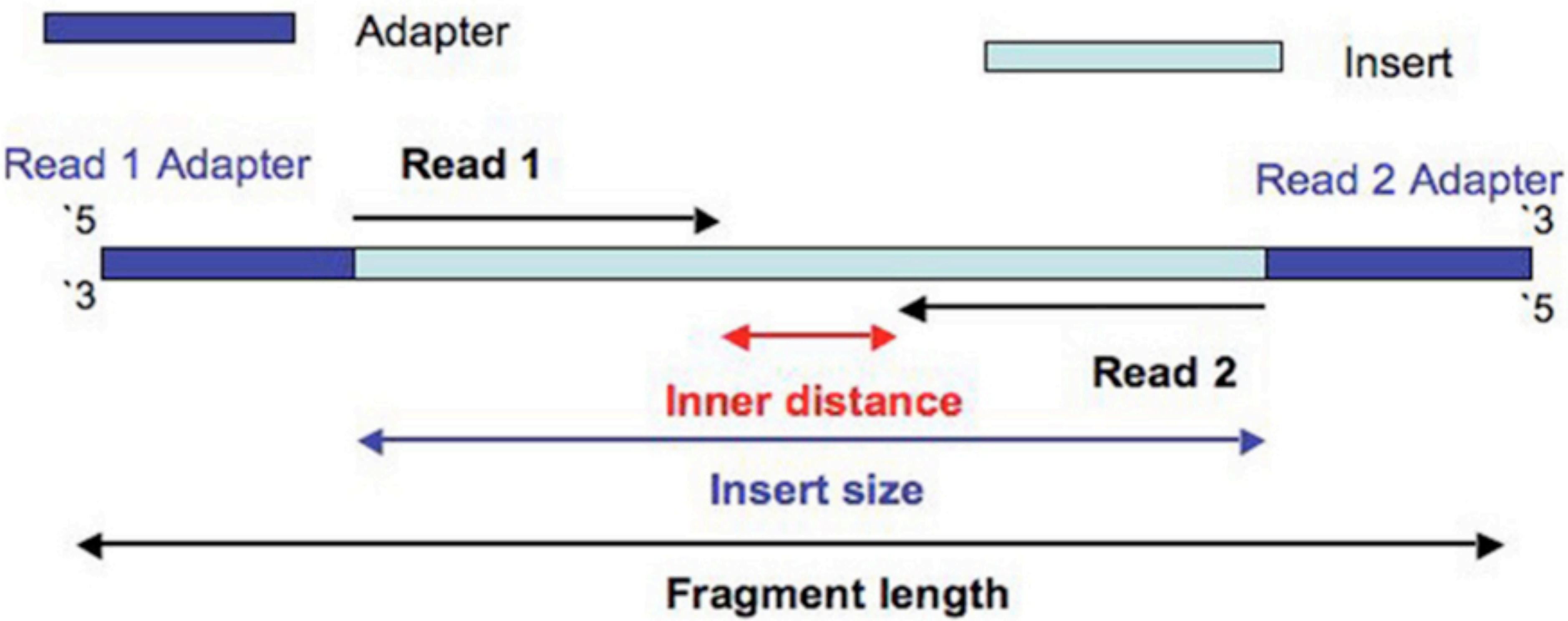
- Make sure your data is whole.
  - File checksums ensure data integrity
    - MD5

      -
        ```
        $ md5sum file1.fastq.gz # before
        823fc8b0ca72c6e9bd8c5dcb0a66ce9b   file1.fastq.gz
        ```

      -
        ```
        $ md5sum -c checksums.md5 # after
        file1.fastq.gz: OK
        file2.fastq.gz: OK
        file3.fastq.gz: FAILED
        md5sum: WARNING: 1 of 3 computed checksums did NOT match
        ```
  - Calculate file checksums before transfer.
  - Validate checksums against the transferred files after the transfer.

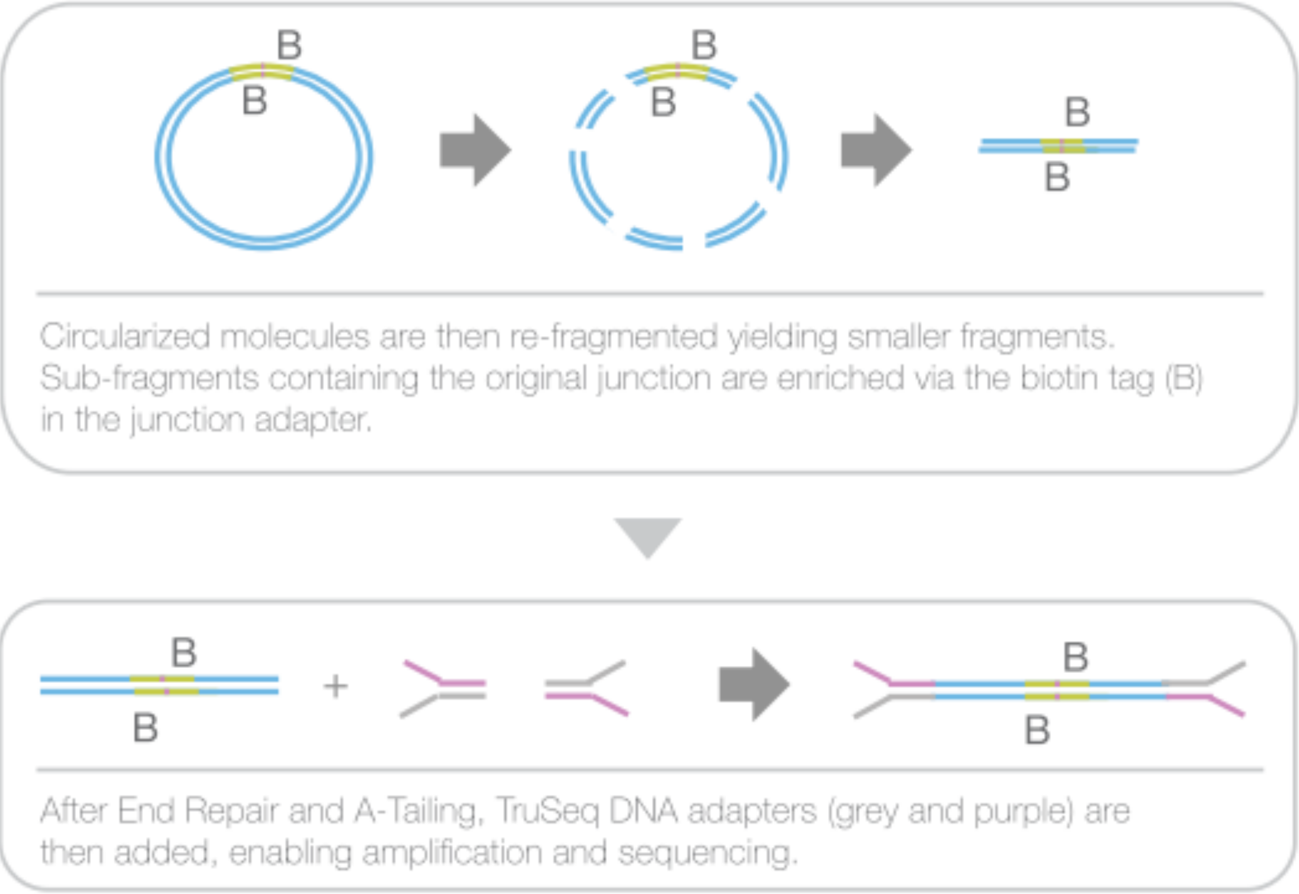
Illumina Sequencing

- Paired end Illumina library



Illumina Sequencing

- Mate pair Illumina library



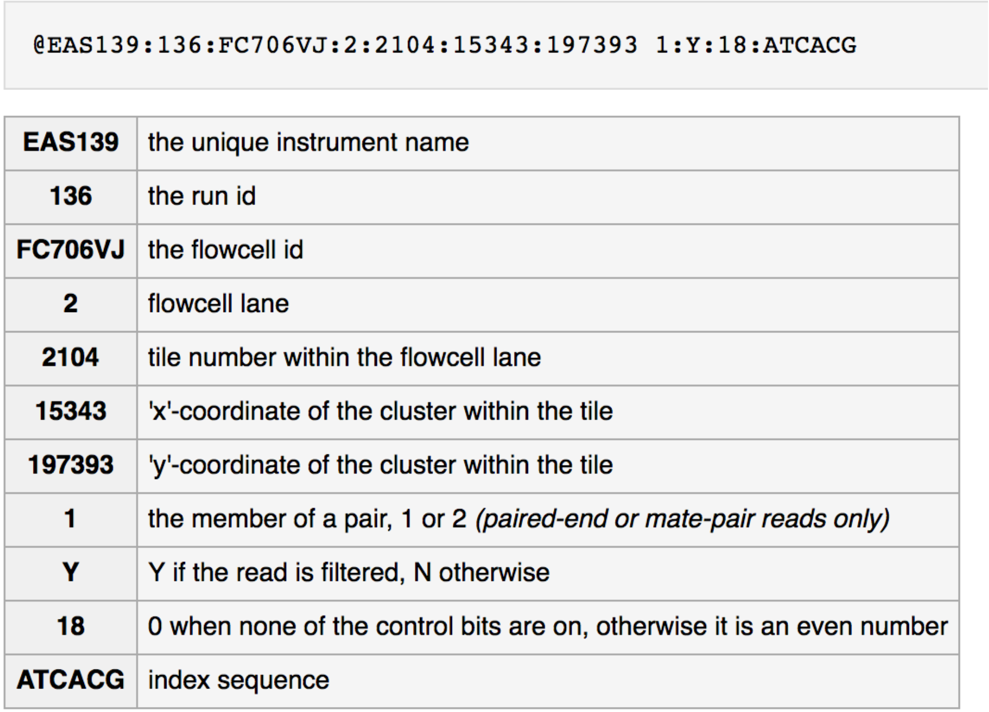
Circularized molecules are then re-fragmented yielding smaller fragments. Sub-fragments containing the original junction are enriched via the biotin tag (B) in the junction adapter.

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
+
CCCFFFFFHHHHHJIIIJHIJJHIJJJJJJJJJJIJJJIJJJJJIIIJJIJJJJGIJJJJH
```



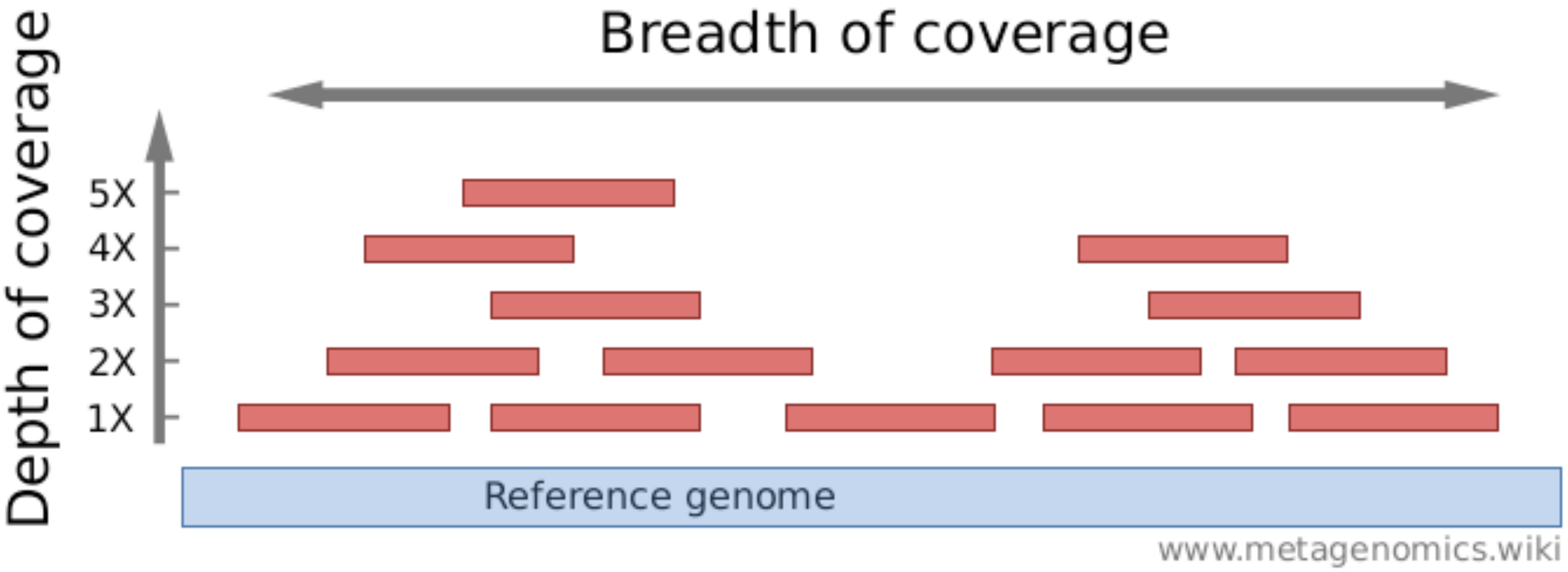
```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

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

Do I have enough data?

- What is my expected genome size?
- What depth of coverage should I expect?
  - Illumina:
    - 100x coverage in total

- Coverage = Number of bases sequenced / Estimated genome size



Calculating data quantity

- FastQC / MultiQC summary reports

- Third party scripts

- Command line calculation (my favourite way)
  - Can use Seqtk to convert files to fasta

  -
    ```
    zcat *.fastq.gz | seqtk seq -A - | 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)

  -
    ```
    parallel ‘seqtk seq -A {} | grep -v “^>” | tr -dc
    “ACGTNacgtn” | wc -m’ ::: *.fastq.gz | paste -sd+ |
    bc -l
    ```

Calculating data quantity

- 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.

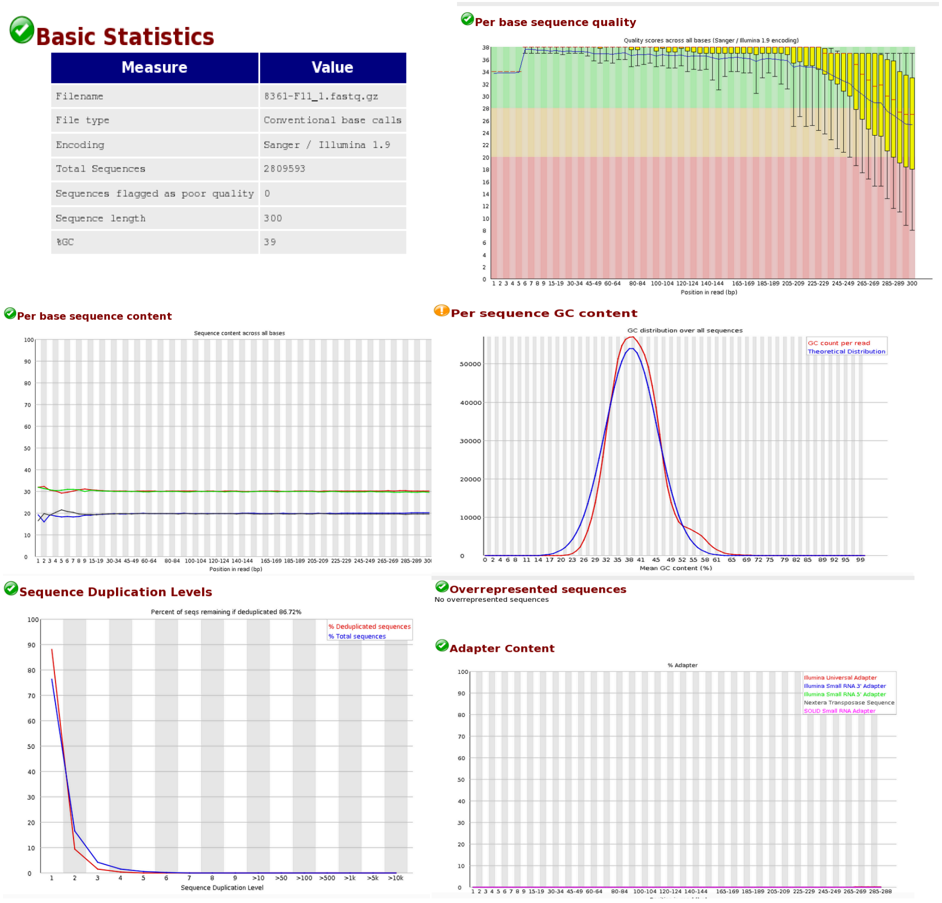
- How should I subsample?
  - Use a random fraction of the reads maintaining read pairing.
    - E.g. Use the same seed (-s) and give the fraction (0.1) in Seqtk.

      ```
      seqtk sample -s100 read1.fq 0.1 > sub1.fq
      seqtk sample -s100 read2.fq 0.1 > sub2.fq
      ```
  - Normalize uneven coverage (e.g. bbnorm)

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

- Look at MultiQC for multiple samples



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

Per base sequence quality

Per base sequence content

Per sequence GC content

Sequence Duplication Levels

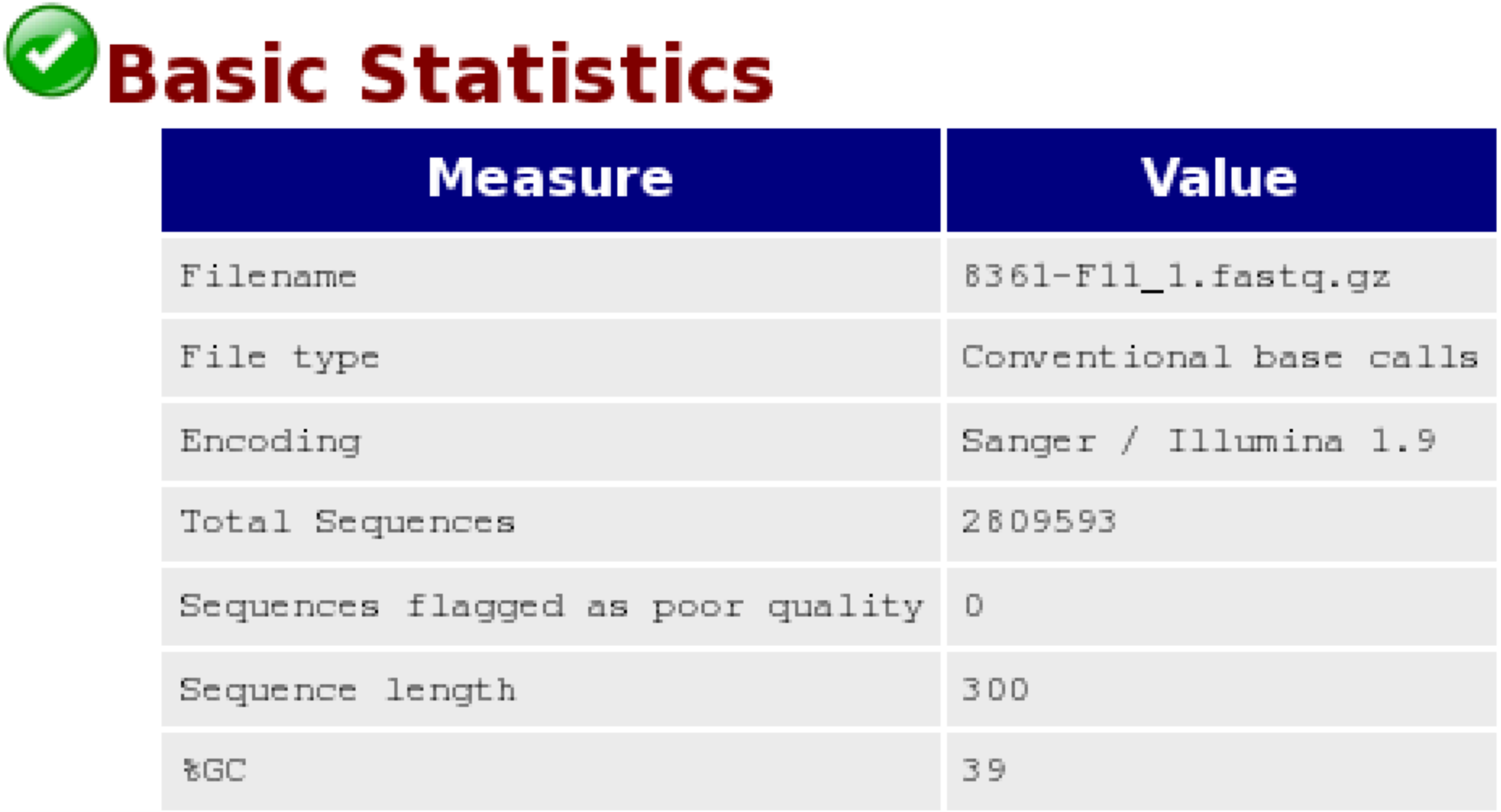
Overrepresented sequences

No overrepresented sequences

Adapter Content

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)
Sequence content across all bases
Position in read (bp)
GC distribution over all sequences
GC count per read
Theoretical Distribution
Mean GC content (%)
Percent of seqs remaining if deduplicated 86.72%
% Deduplicated sequences
% Total sequences
Sequence Duplication Level
% Adapter
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter

FastQC

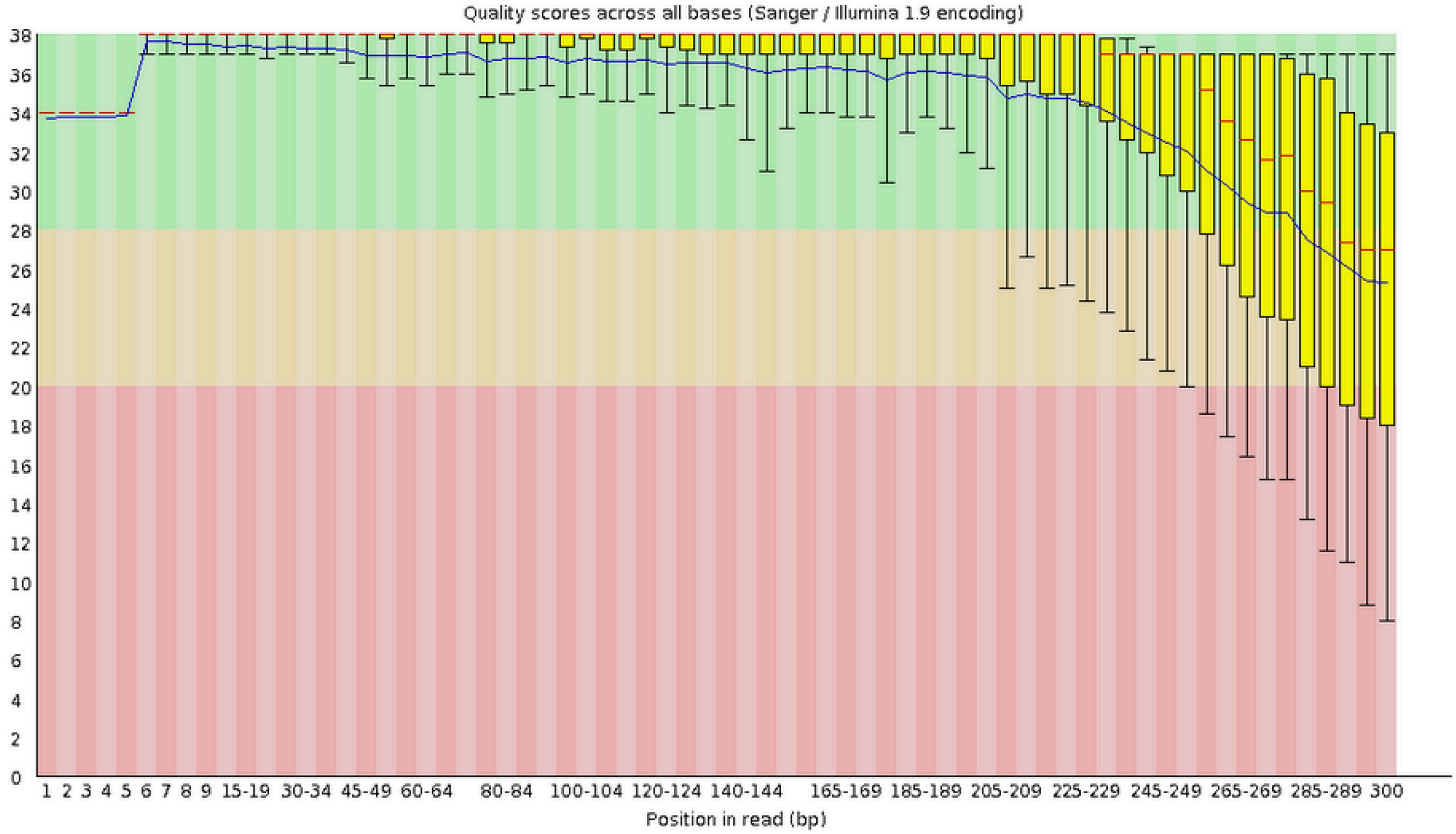


✓ Basic Statistics

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



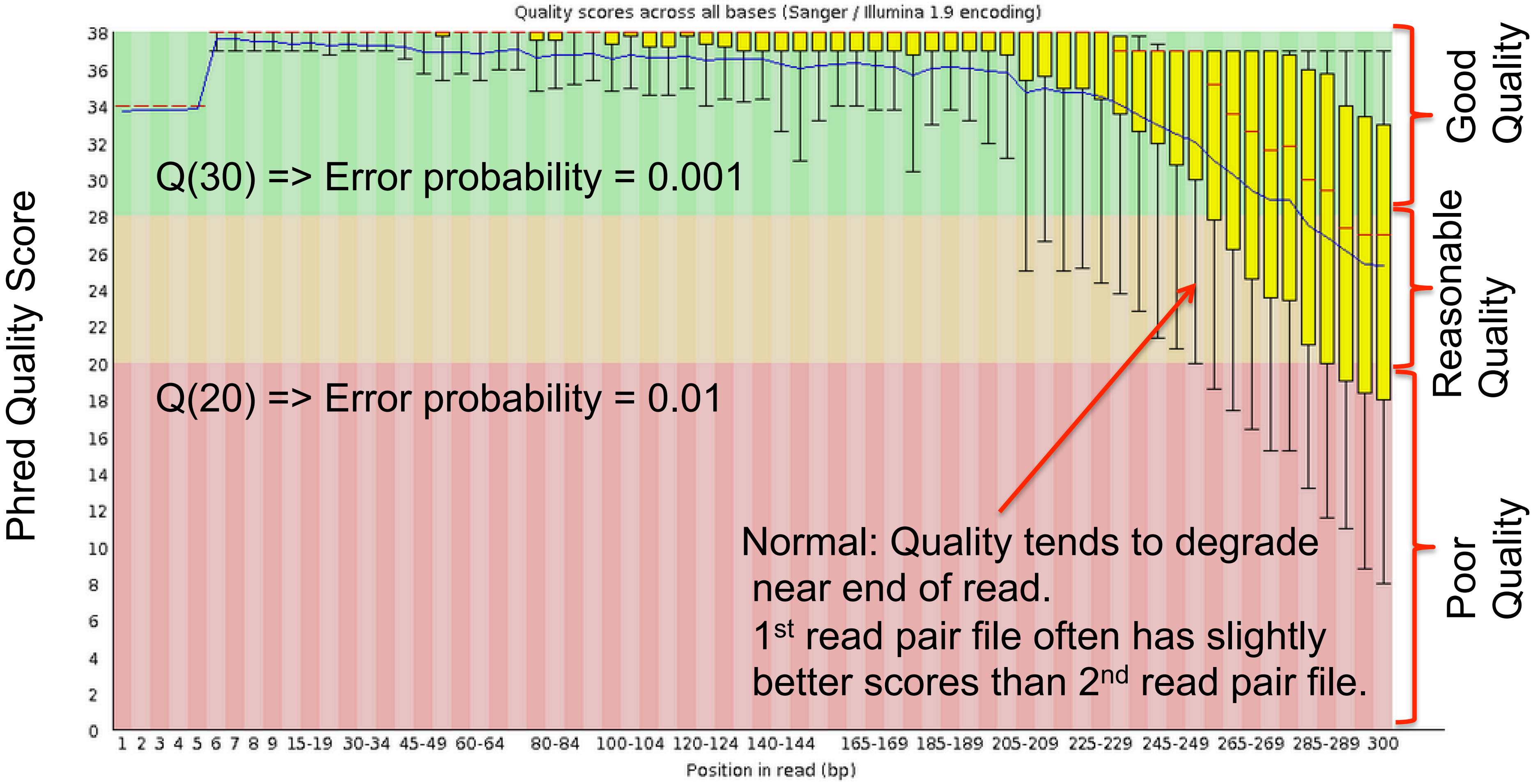
FastQC

SciLifeLab

Q(40) => Error probability = 0.0001

Per base sequence quality

Blue line = mean quality



Q(30) => Error probability = 0.001

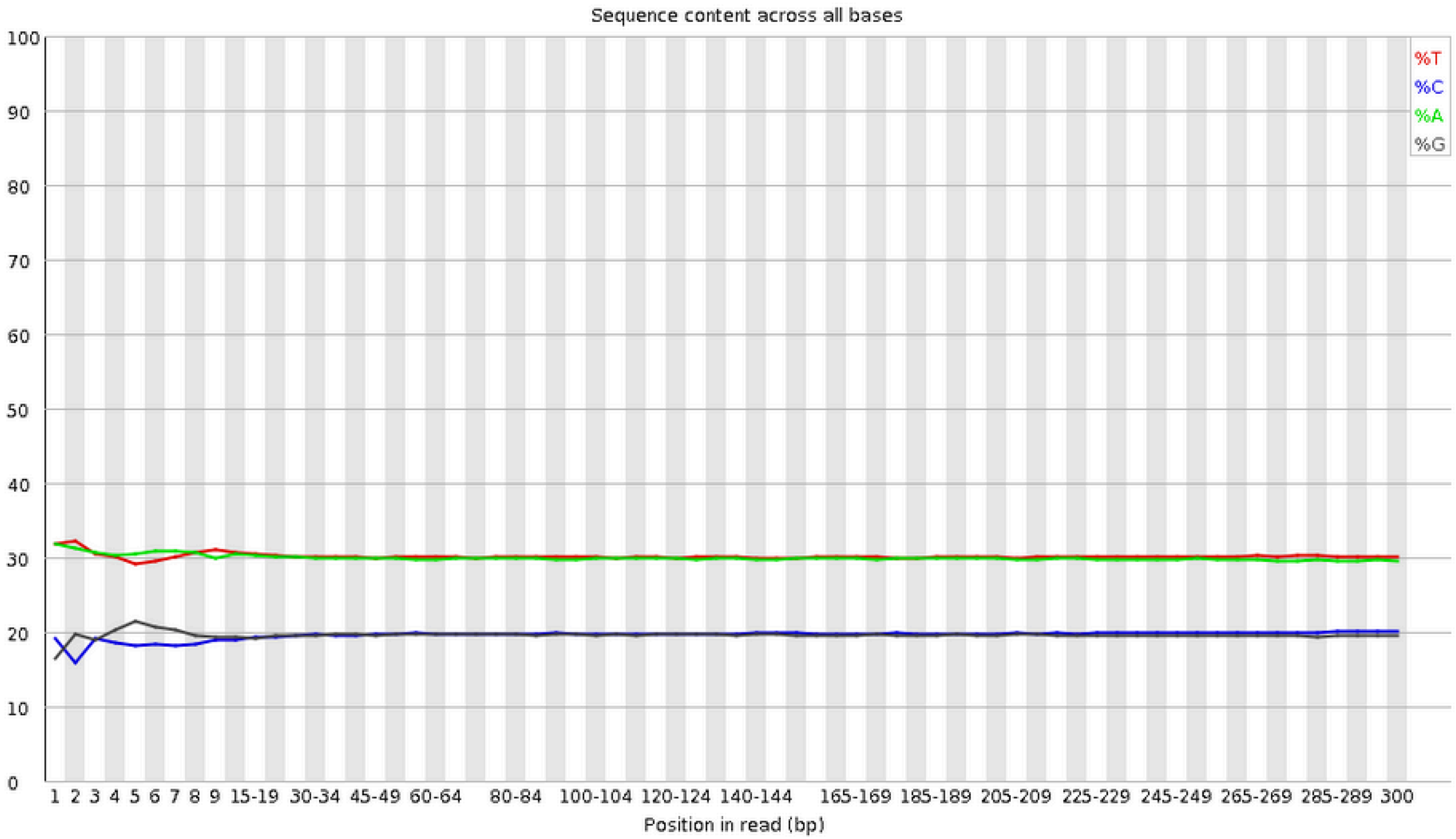
Q(20) => Error probability = 0.01

Normal: Quality tends to degrade near end of read.

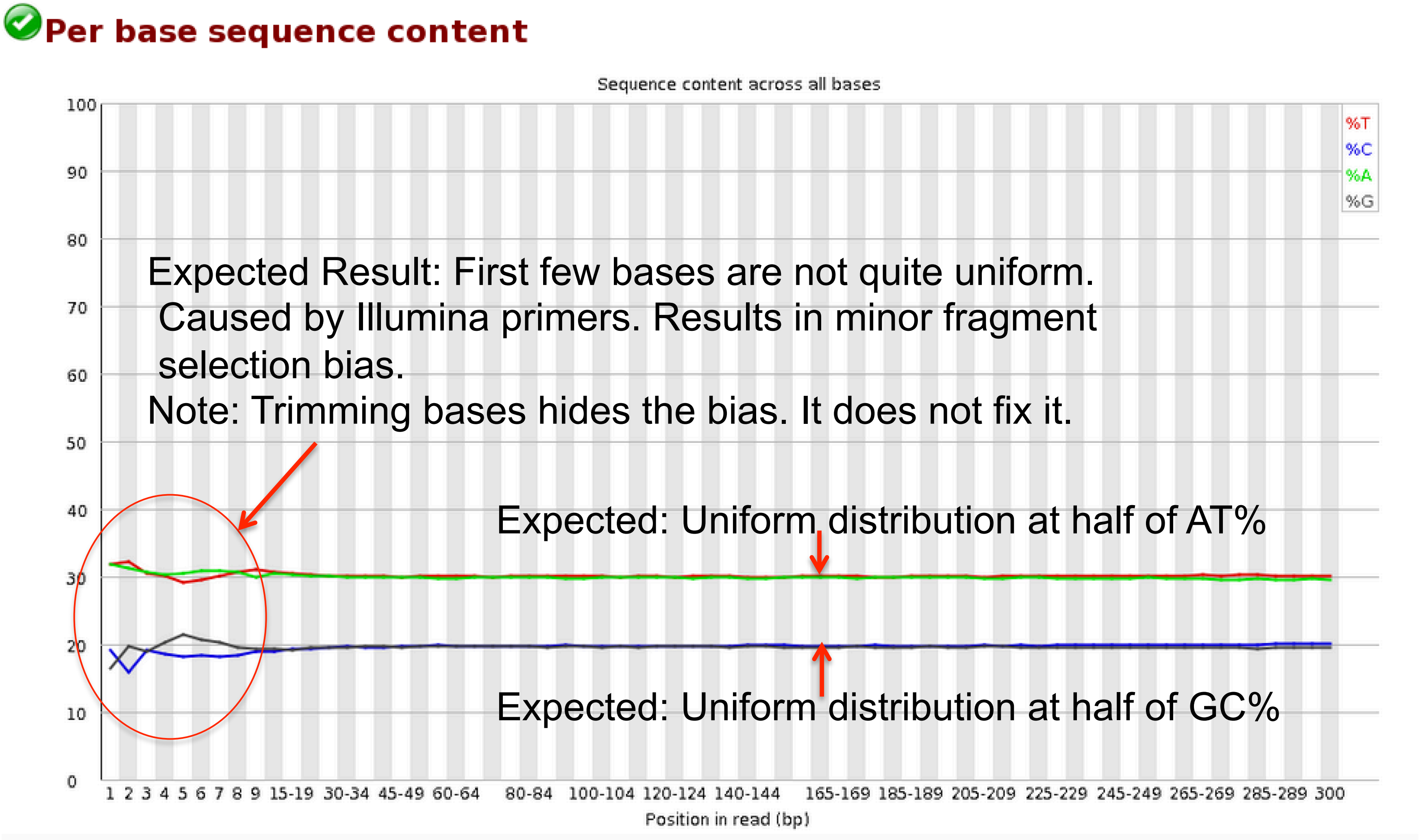
1st read pair file often has slightly better scores than 2nd read pair file.

FastQC

✓ Per base sequence content



FastQC



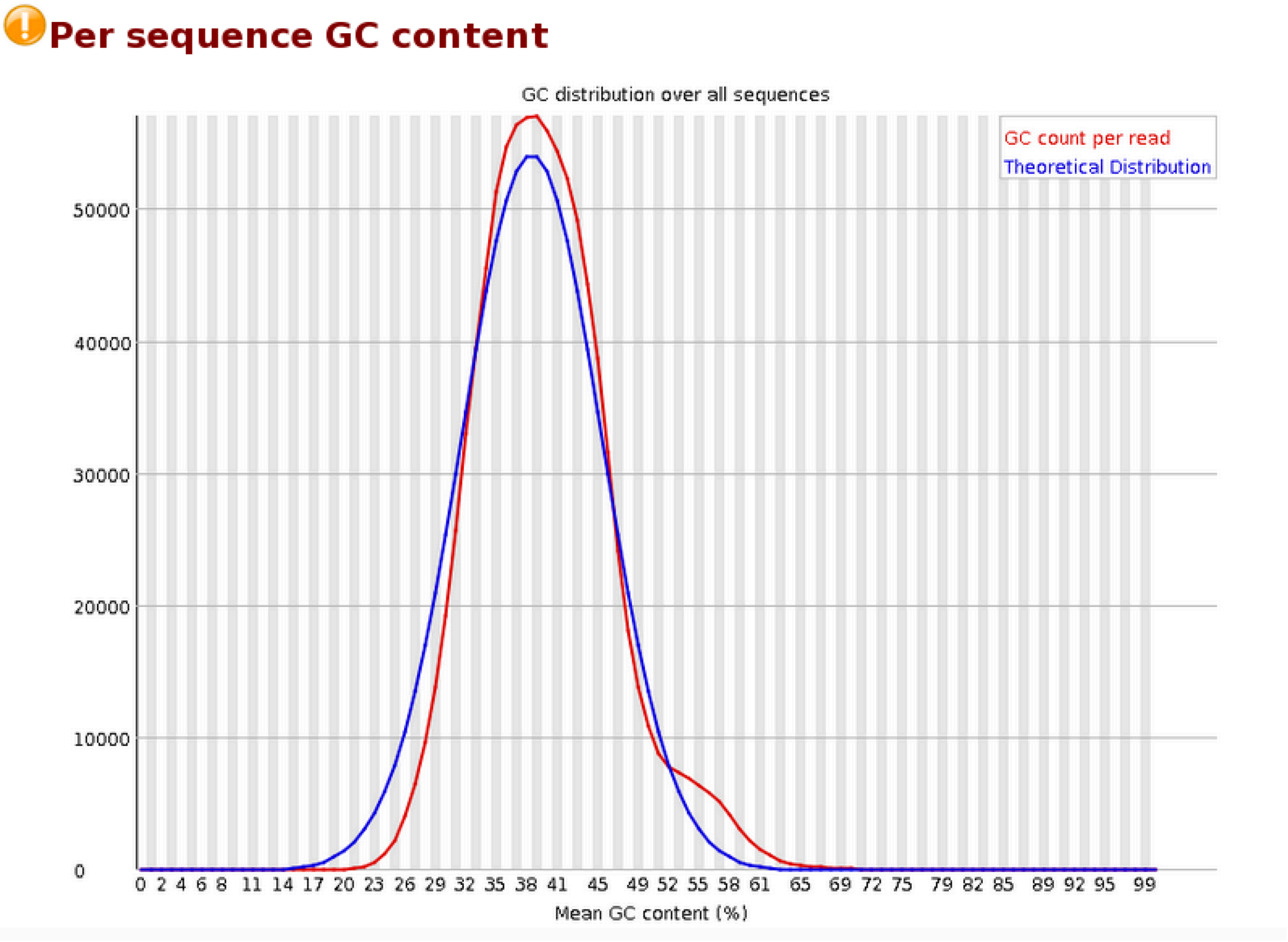
Expected Result: First few bases are not quite uniform. Caused by Illumina primers. Results in minor fragment selection bias.

Note: Trimming bases hides the bias. It does not fix it.

Expected: Uniform distribution at half of AT%

Expected: Uniform distribution at half of GC%

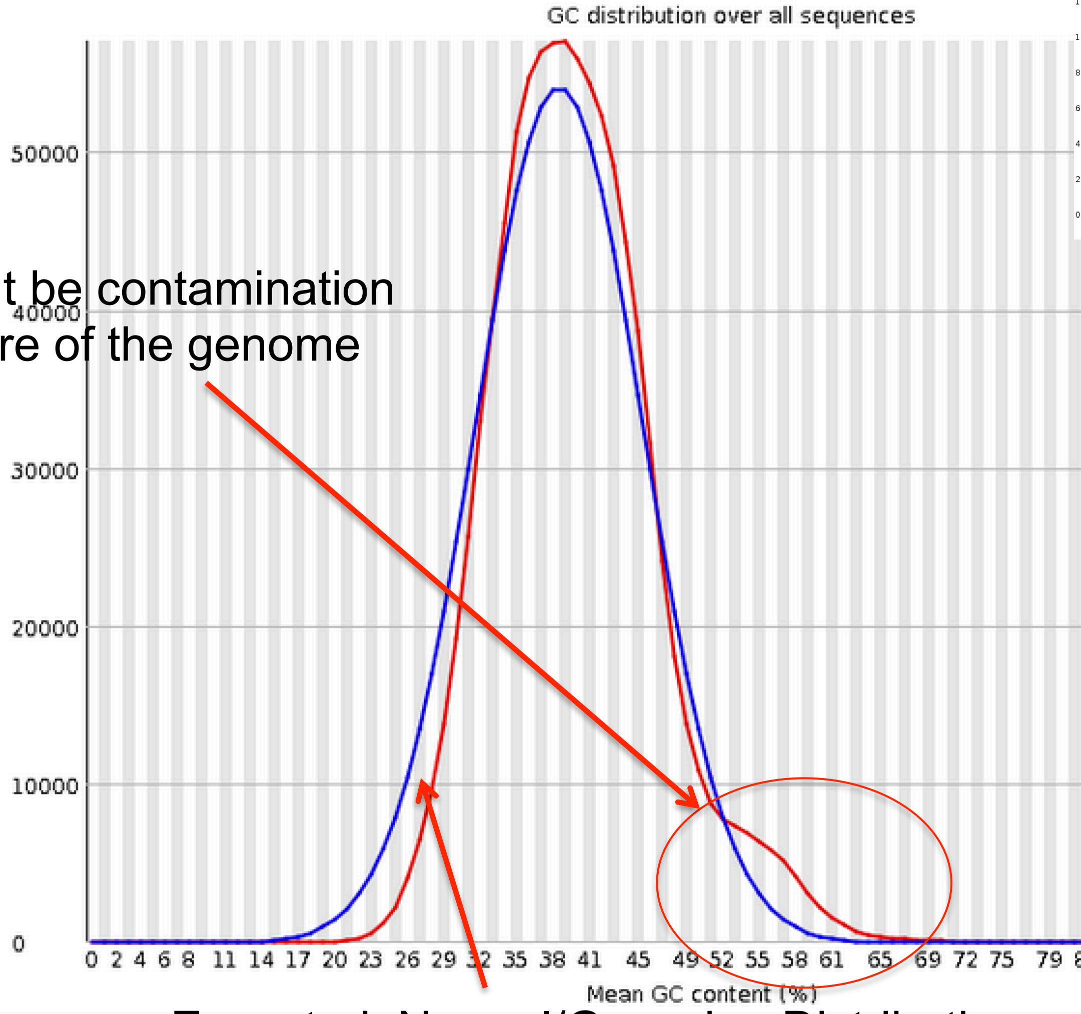
FastQC



Per sequence GC content

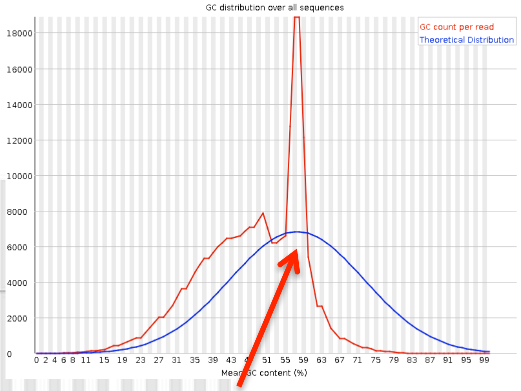
FastQC

Per sequence GC content

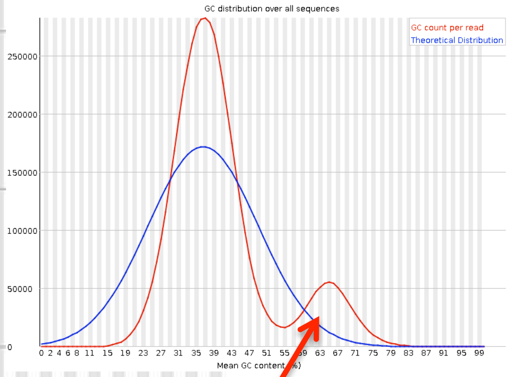


This might be contamination or a feature of the genome

Expected: Normal/Gaussian Distribution



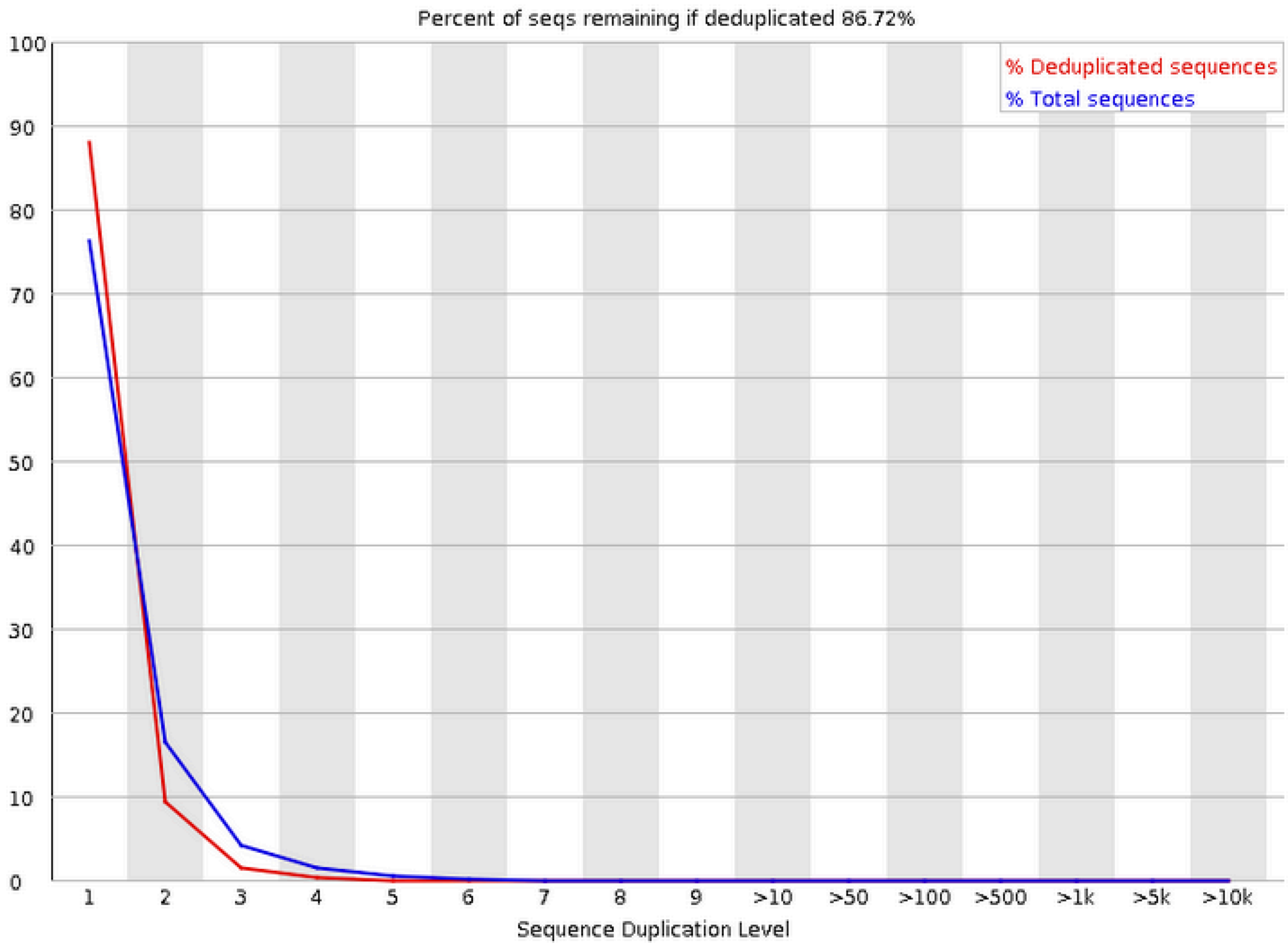
Sharp peak indicates specific motif. Adapters are the usual suspect.



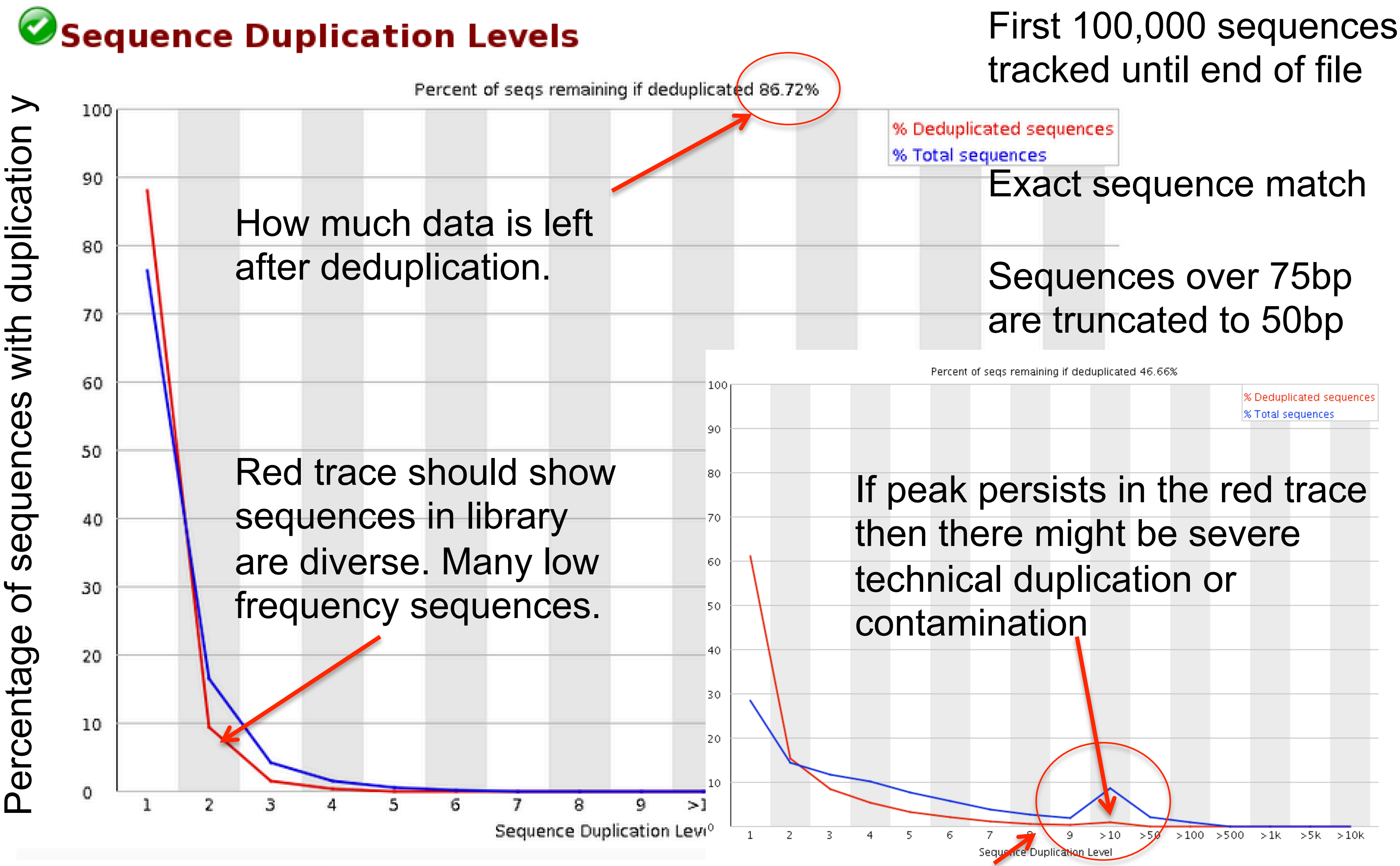
Wider or multiple distributions suggest contamination.

FastQC

✓ Sequence Duplication Levels



FastQC



First 100,000 sequences tracked until end of file

Exact sequence match

Sequences over 75bp are truncated to 50bp

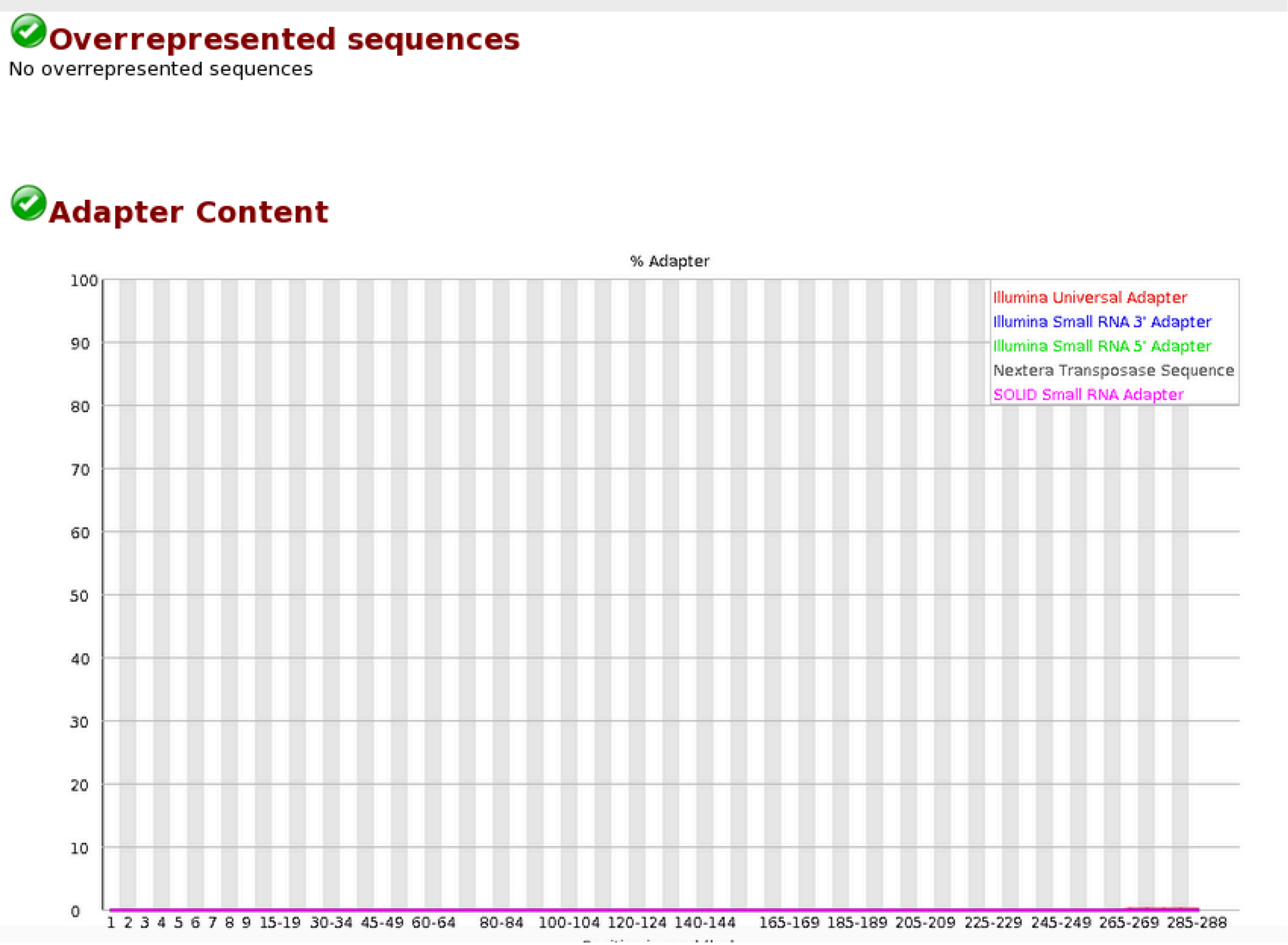
How much data is left after deduplication.

Red trace should show sequences in library are diverse. Many low frequency sequences.

If peak persists in the red trace then there might be severe technical duplication or contamination

Peak shows 10%+ sequences with high duplication levels

FastQC



✓ Overrepresented sequences

No overrepresented sequences

✓ Adapter Content

FastQC

✓ Overrepresented sequences

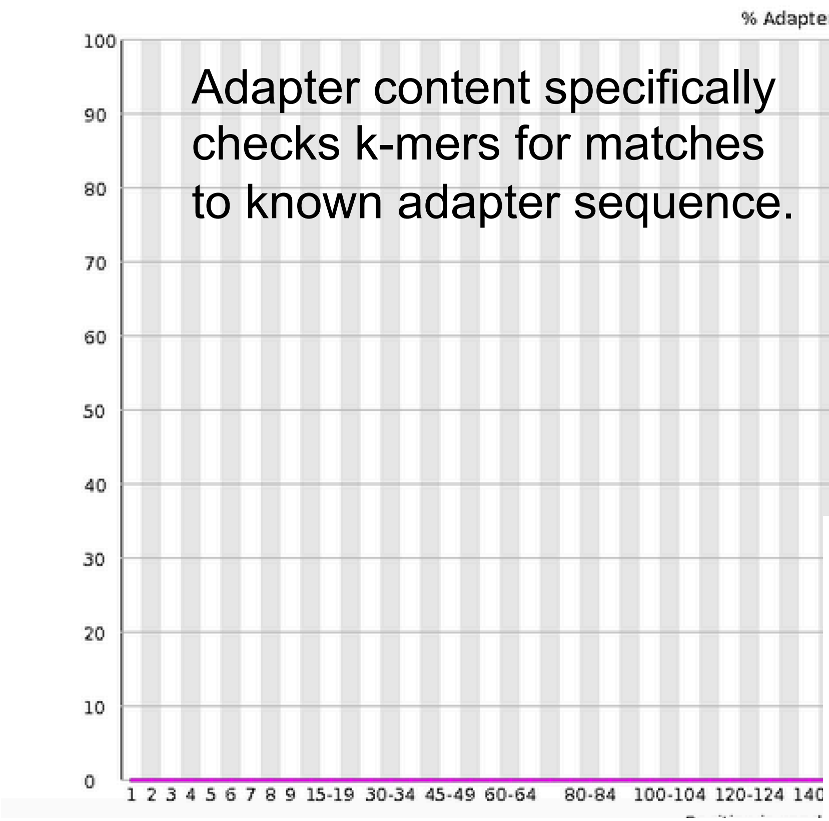
No overrepresented sequences

Lists sequence that is more than 0.1%

First 100,000 sequences tracked until end of file

Overrepresented sequences are matched against known contaminants.

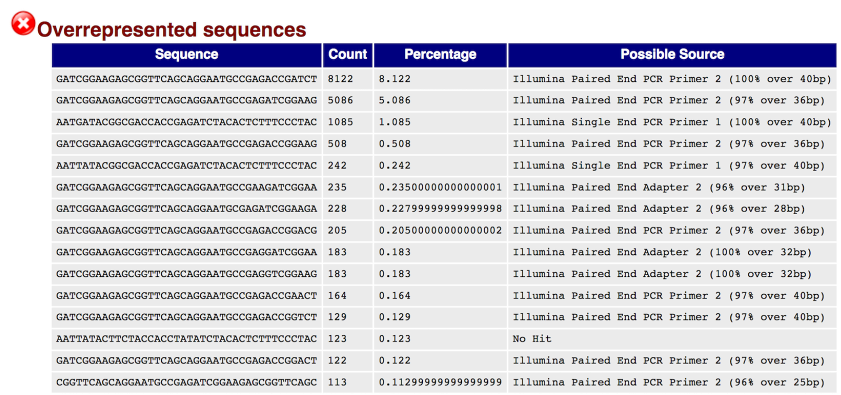
✓ Adapter Content



Adapter content specifically checks k-mers for matches to known adapter sequence.

Match hits are not conclusive, but indicative.

Matches must be >20bp and only 1 mismatch.

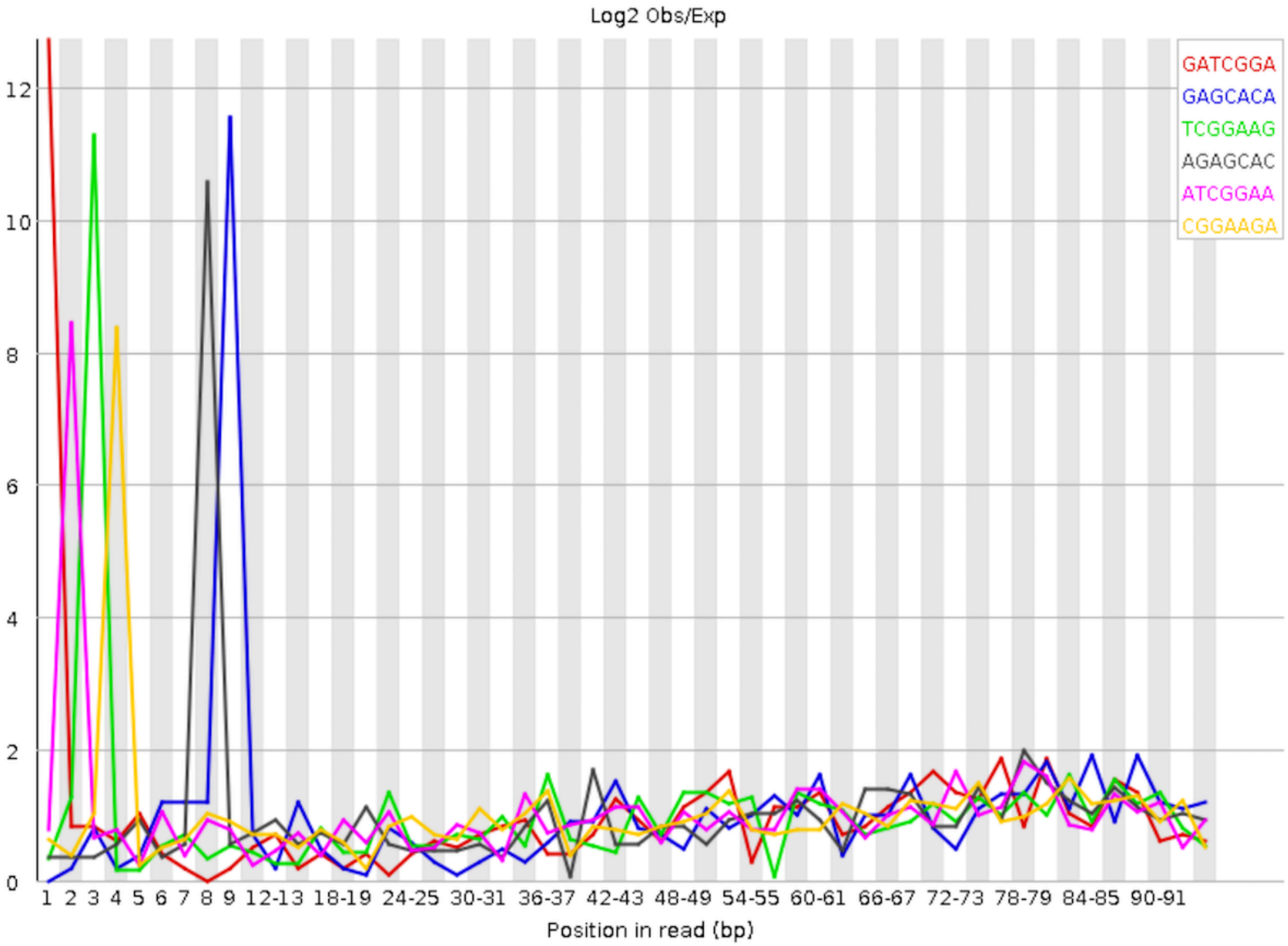


☒ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	5086	5.086	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	1085	1.085	Illumina Single End PCR Primer 1 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGAAG	508	0.508	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATTATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	242	0.242	Illumina Single End PCR Primer 1 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAAGATCGGAA	235	0.23500000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACG	205	0.20500000000000002	Illumina Paired End PCR Primer 2 (97% over 36bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGATCGGAA	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGTCGGAAG	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAACT	164	0.164	Illumina Paired End PCR Primer 2 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGTCT	129	0.129	Illumina Paired End PCR Primer 2 (97% over 40bp)
AATTATACTTCTACCACCTATATCTACACTCTTTCCCTAC	123	0.123	No Hit
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACT	122	0.122	Illumina Paired End PCR Primer 2 (97% over 36bp)
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	113	0.11299999999999999	Illumina Paired End PCR Primer 2 (96% over 25bp)

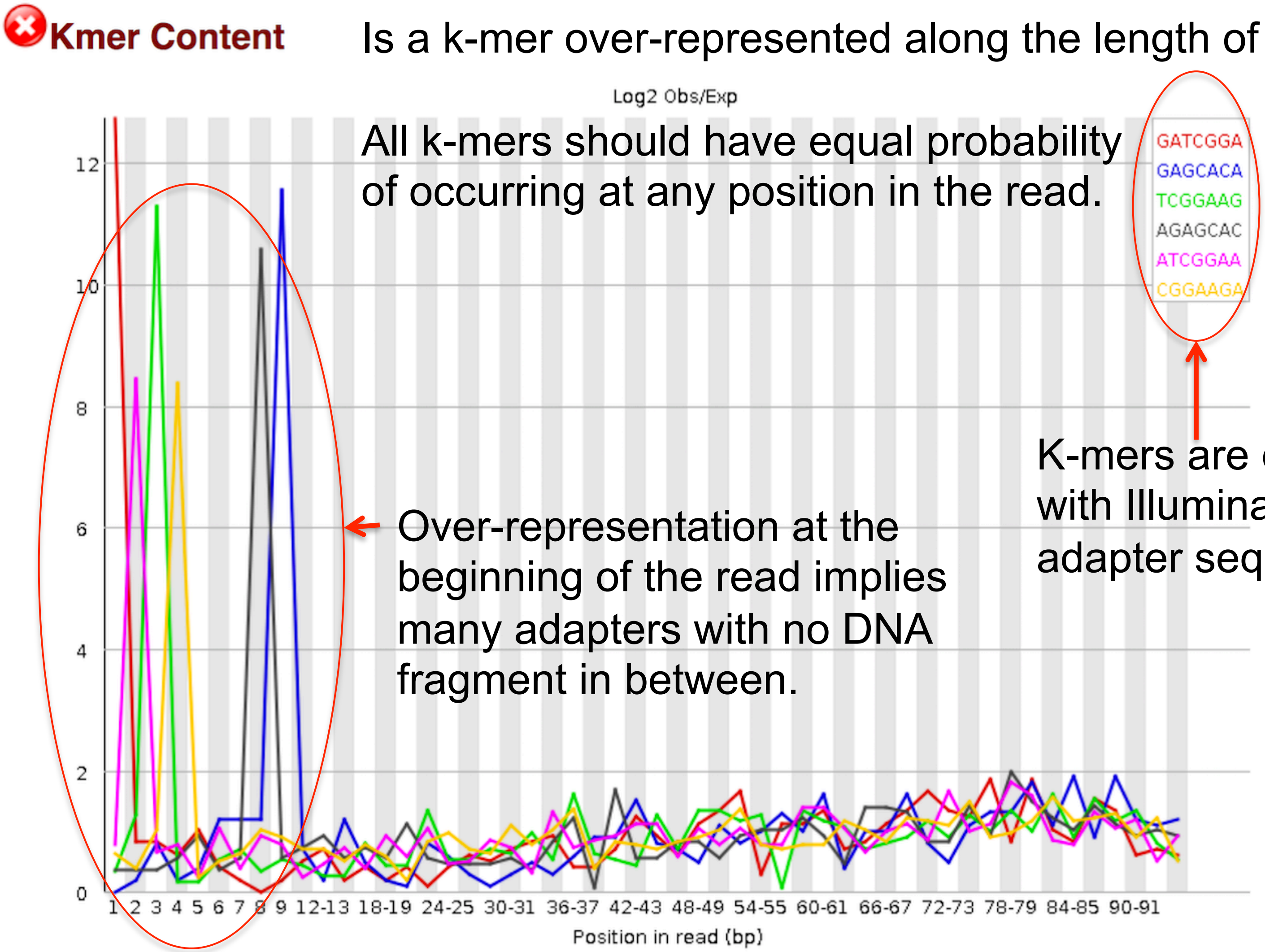
FastQC

☒ Kmer Content



FastQC

Is a k-mer over-represented along the length of a read?



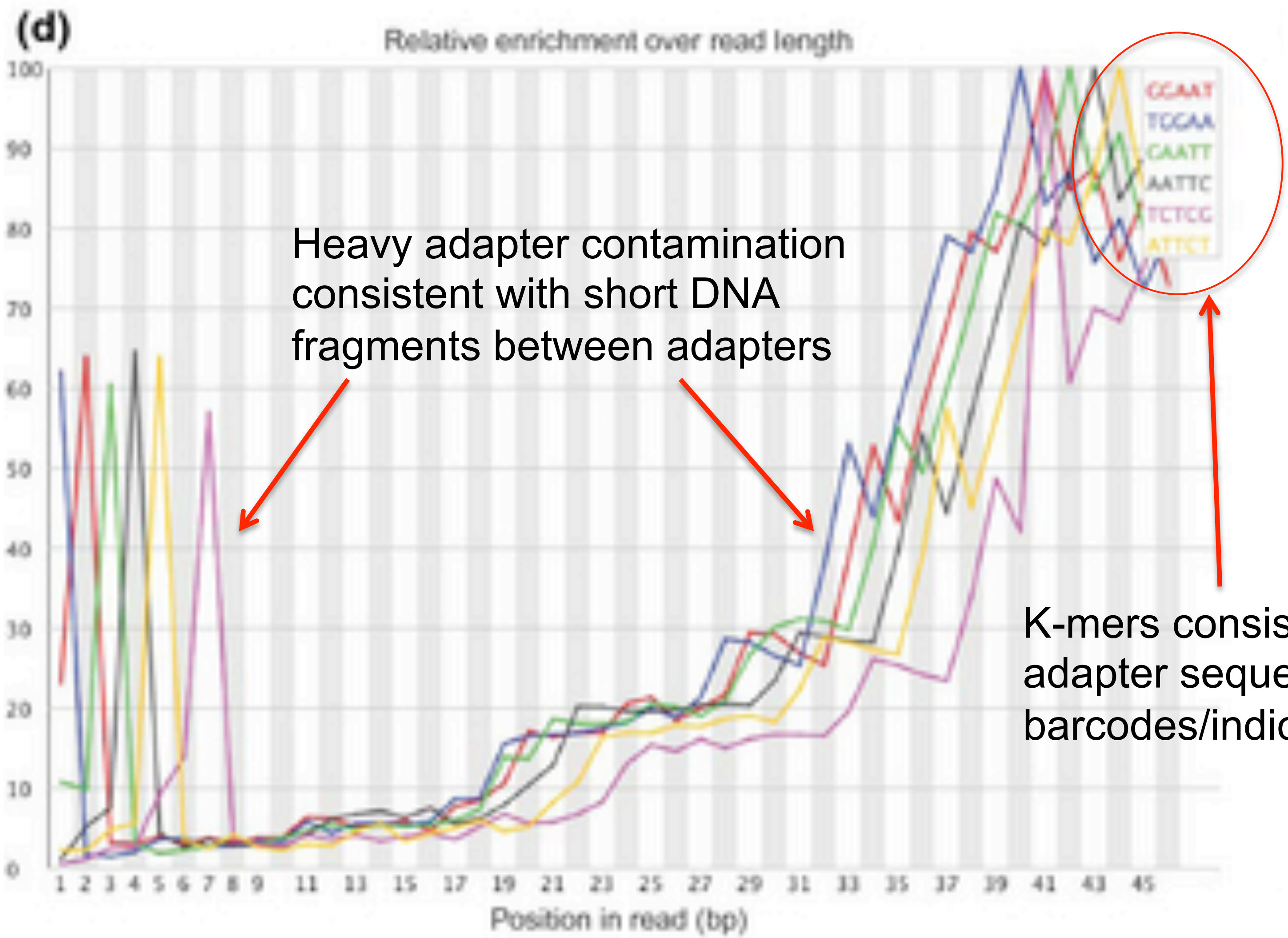
All k-mers should have equal probability of occurring at any position in the read.

Over-representation at the beginning of the read implies many adapters with no DNA fragment in between.

K-mers are consistent with Illumina TruSeq adapter sequence.

Default k is 7. K-mer size can be increased with option -k

FastQC

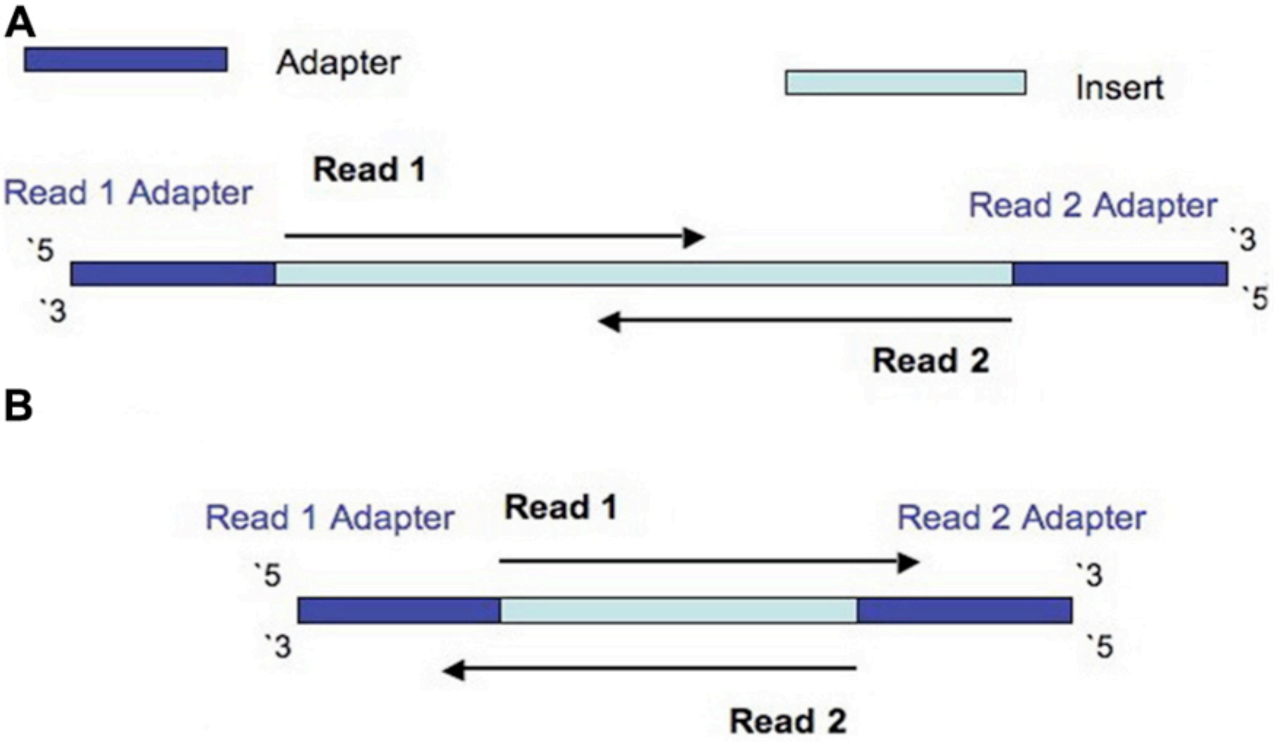


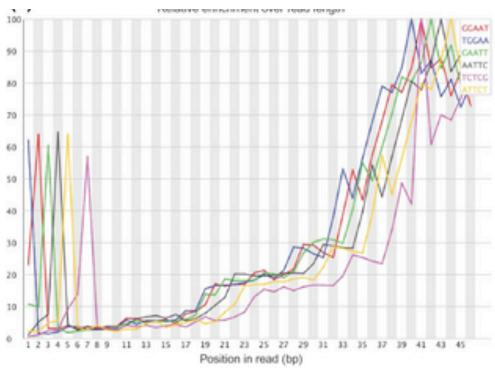
Heavy adapter contamination consistent with short DNA fragments between adapters

K-mers consistent with adapter sequences and barcodes/indices

Trimming reads

- Why trim reads?
  - Remove adapter read through.

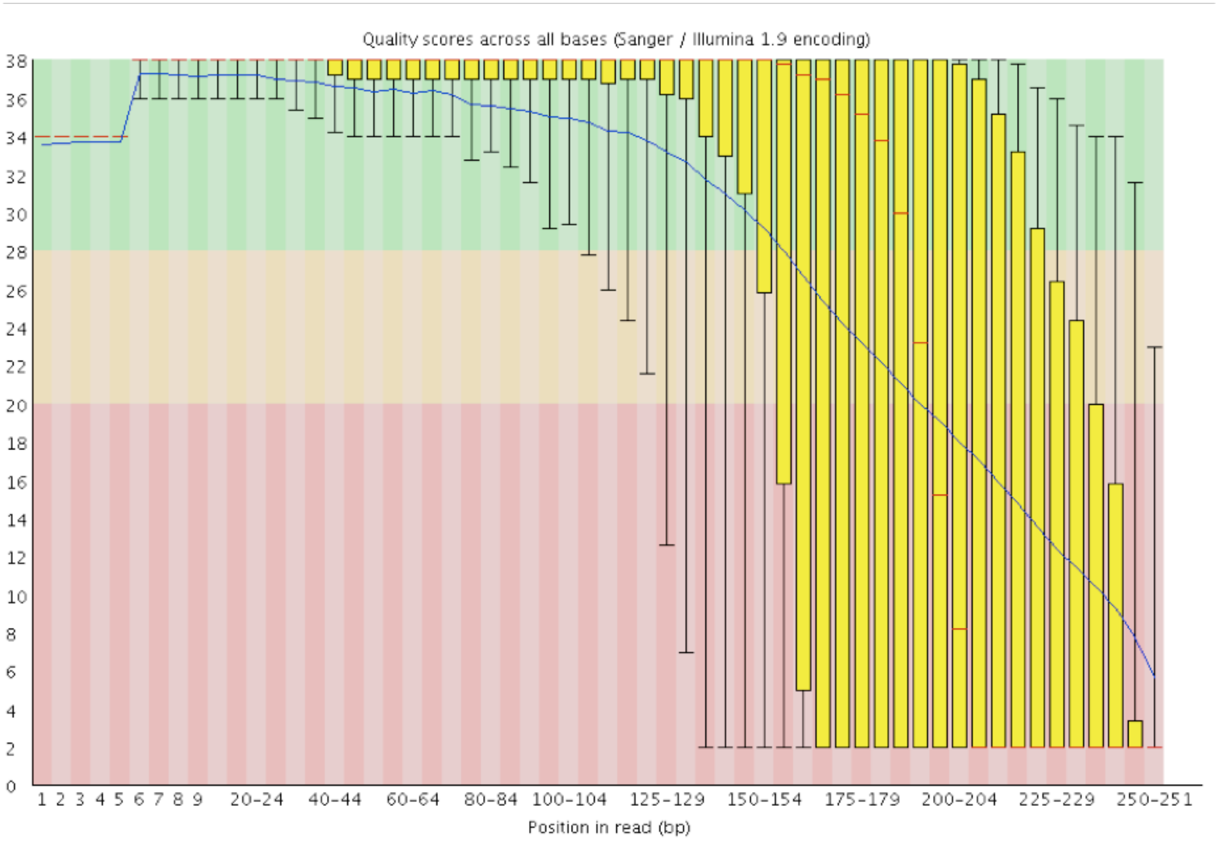




  - Update to date list of Illumina adapters:
    https://support.illumina.com/downloads/illumina-customer-sequence-letter.html

Trimming reads

- Why trim reads?
  - Remove poor quality reads



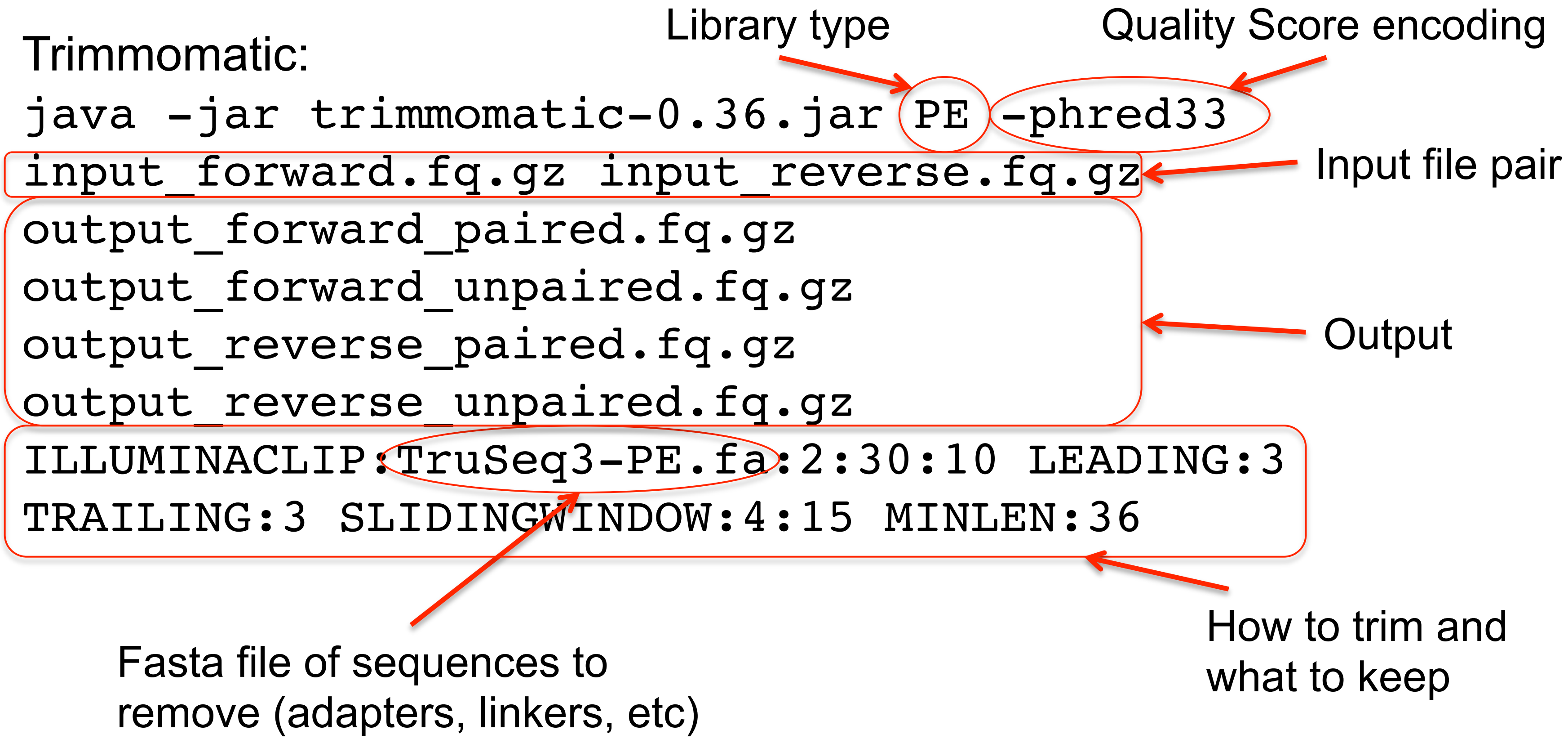
Trimming reads

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

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

Trimming reads

- Trimmomatic:



```
java -jar trimmomatic-0.36.jar PE -phred33
input_forward.fq.gz input_reverse.fq.gz
output_forward_paired.fq.gz
output_forward_unpaired.fq.gz
output_reverse_paired.fq.gz
output_reverse_unpaired.fq.gz
ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3
TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

Library type

Quality Score encoding

Input file pair

Output

Fasta file of sequences to remove (adapters, linkers, etc)

How to trim and what to keep

- BBMerge can be used to discover adapters:

```
bbmerge.sh in=reads.fq outa=adapters.fa
```

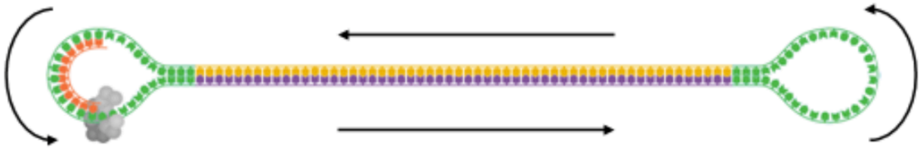
Duplication Removal

- Why do duplicates arise?
  - Optical duplicates (amplified cluster mistaken for multiple clusters)
  - PCR duplicates

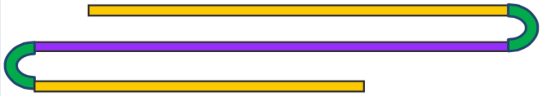
- Why are duplicates bad?
  - Poor overlap information
  - Increased variance of coverage
  - Increased computation time and resources

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

PacBio Sequencing



SMRTbell™ Template



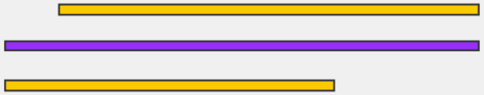
Polymerase Read

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

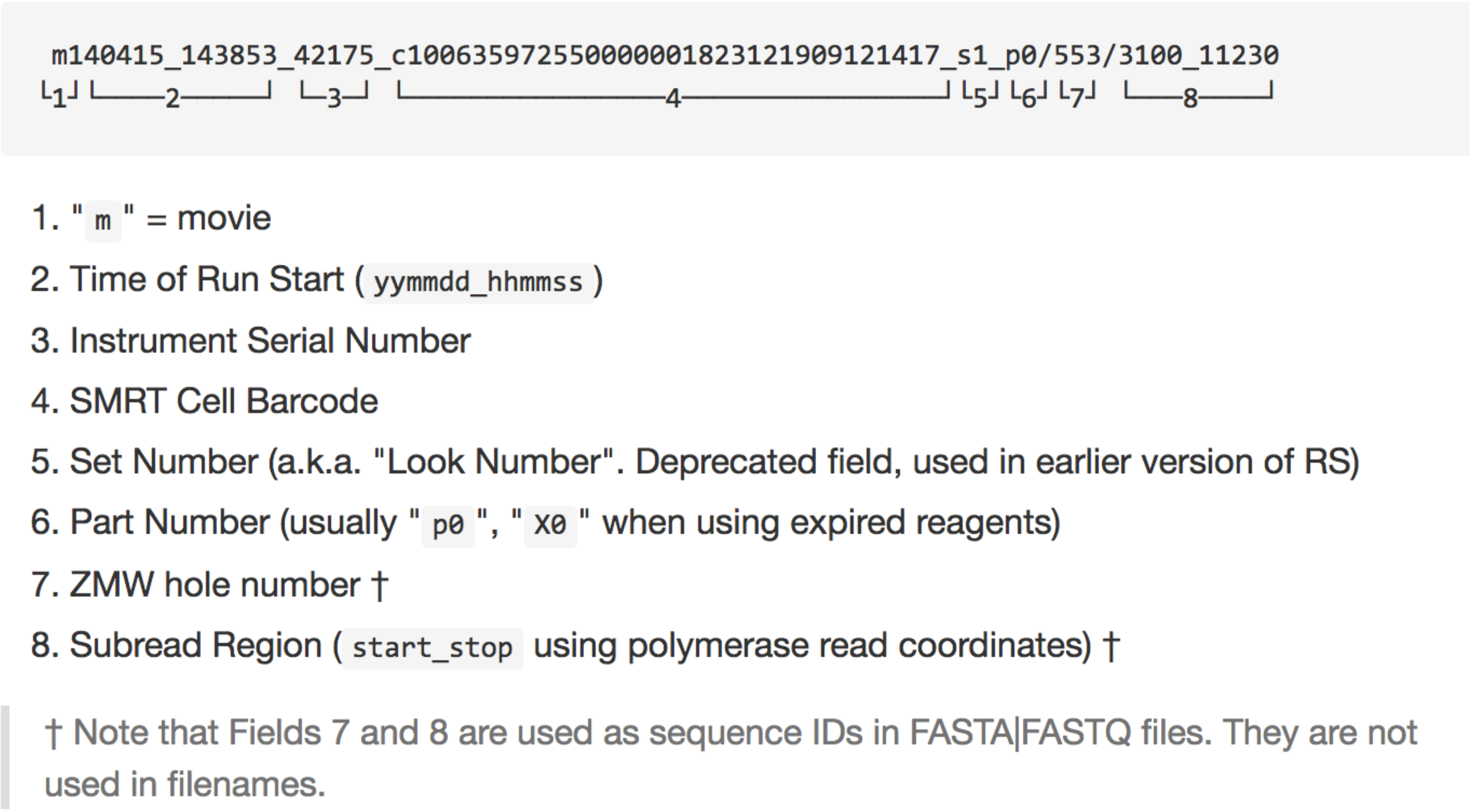
Definition:

- 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

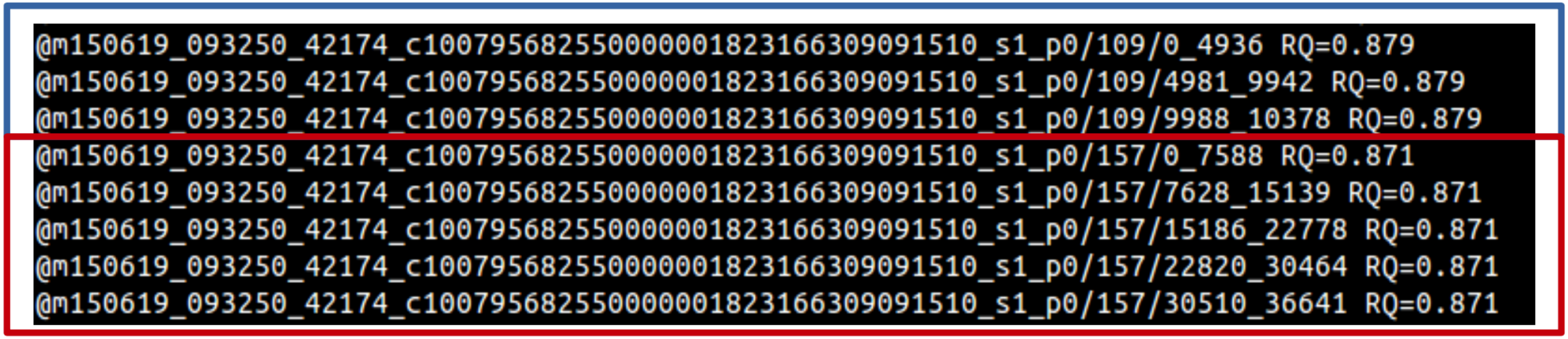


```
 m140415_143853_42175_c100635972550000001823121909121417_s1_p0/553/3100_11230
└1┘ └────2────┘ └─3─┘ └────────────────4───────────────┘ └5┘└6┘└7┘ └───8────┘
```

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

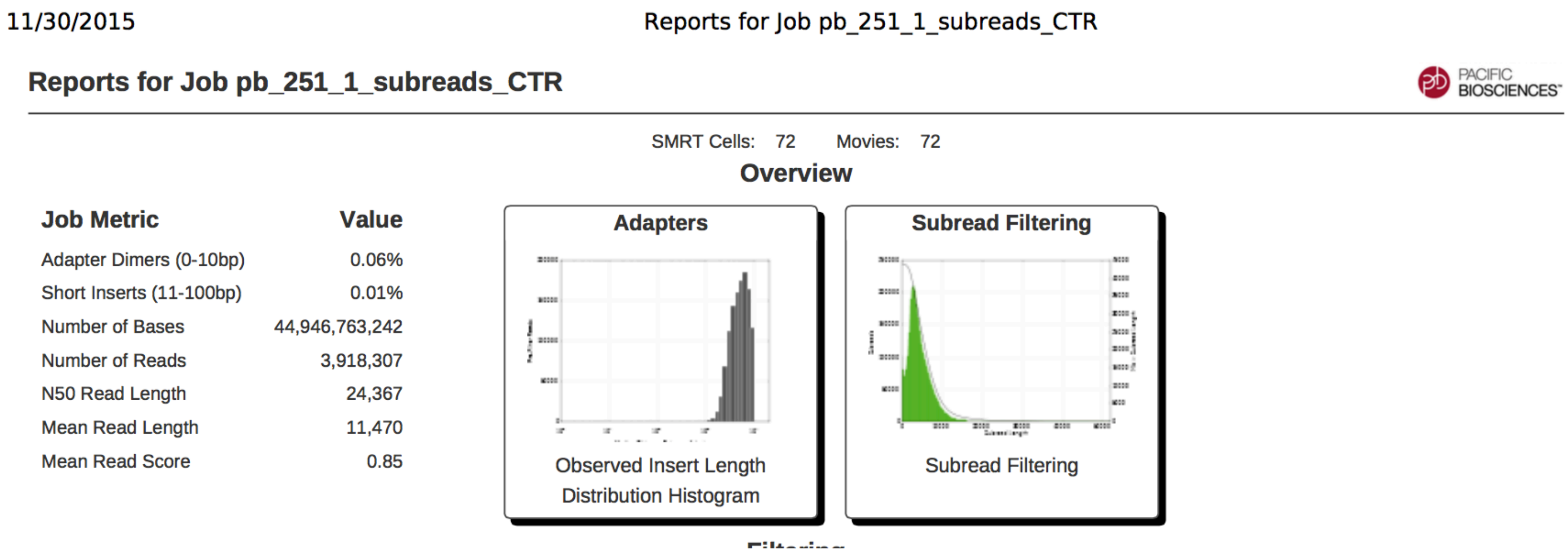


```
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/109/0_4936 RQ=0.879
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/109/4981_9942 RQ=0.879
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/109/9988_10378 RQ=0.879
@m150619_093250_42174_c100795682550000001823166309091510_s1_p0/157/0_7588 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).

Do I have enough data?

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



11/30/2015 Reports for Job pb_251_1_subreads_CTR

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

Observed Insert Length Distribution Histogram

Subread Filtering

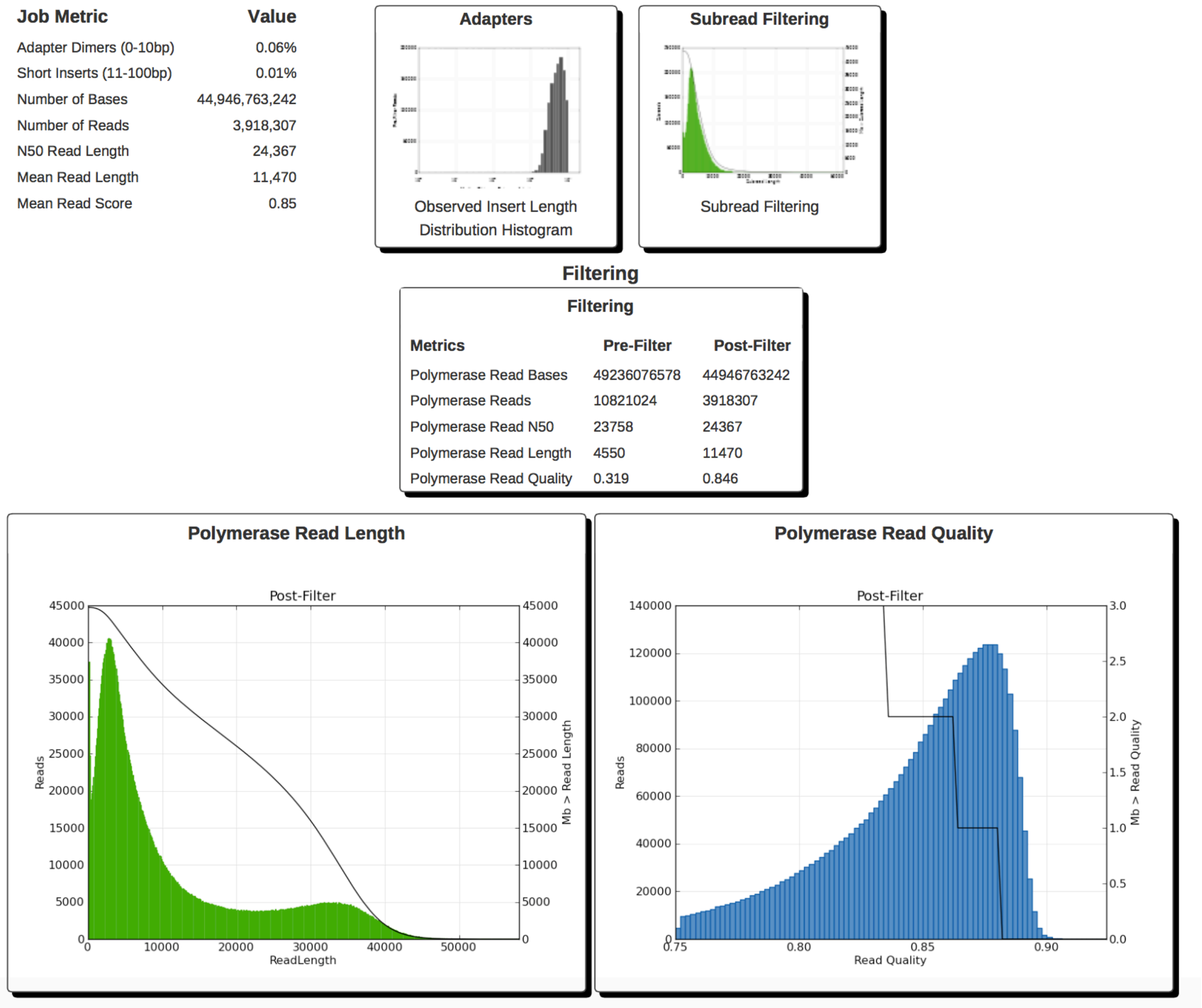
Filtering

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)
  - **`parallel ‘seqtk seq -A -L 10000 {} | grep -v “^>” | tr -dc “ACGTNacgtn” | wc -m’ ::: *.fastq.gz | paste -sd+ | bc -l`**

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

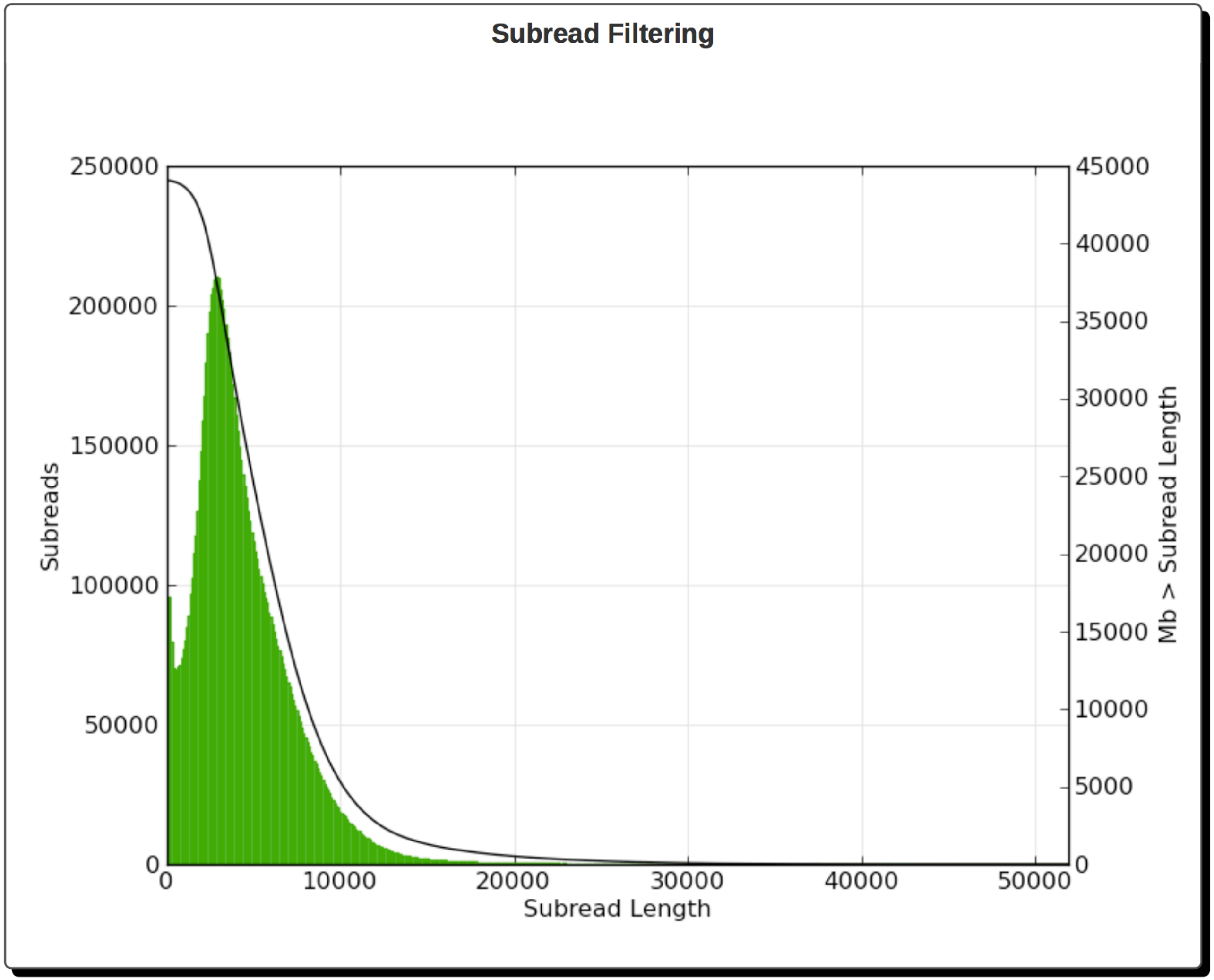
Filtering

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

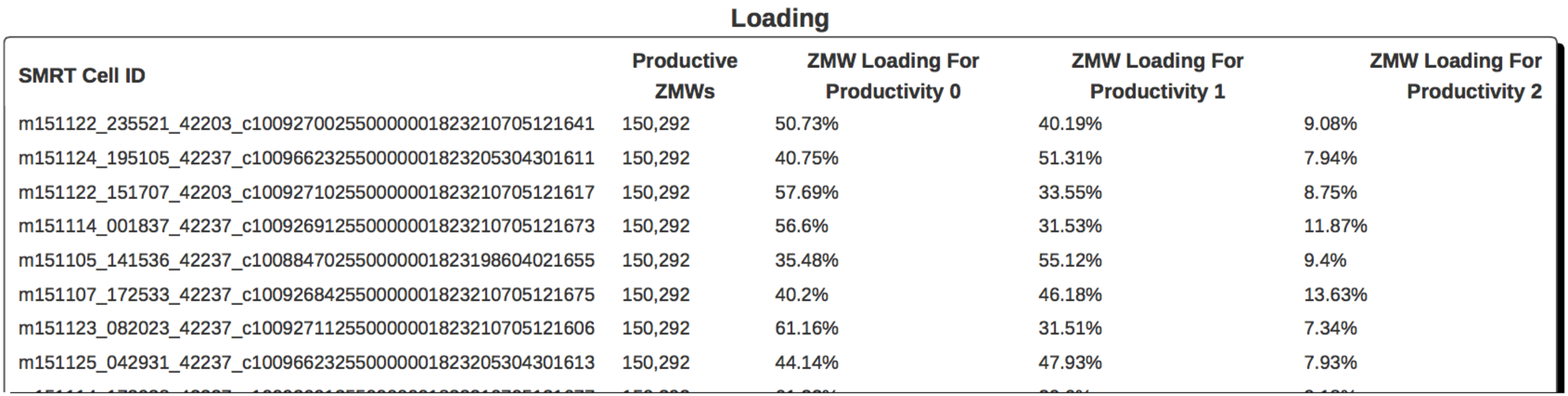
SMRT Portal Report



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)).

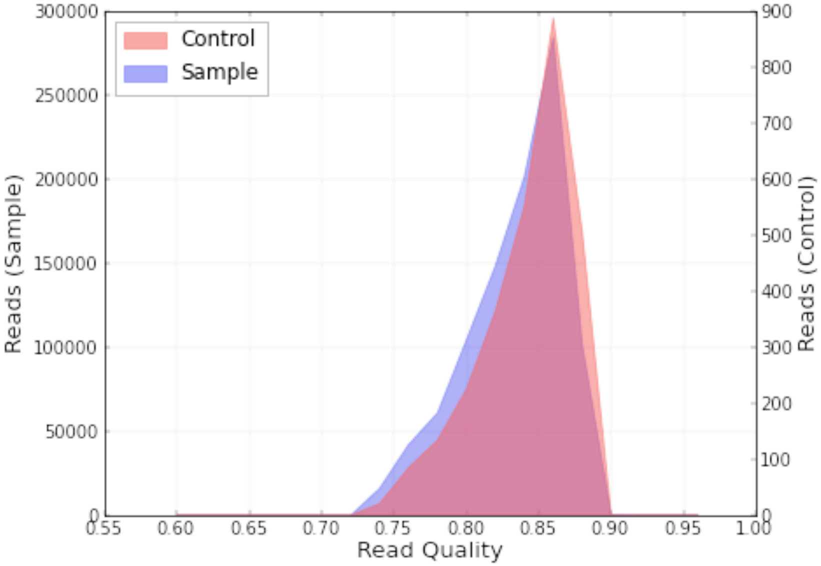
  - Maximize P1 and minimize P0 + P2.
  - High P0 indicates underloading (too low concentration of molecules)
  - High P2 indicates overloading (too high concentration) or poor prep.

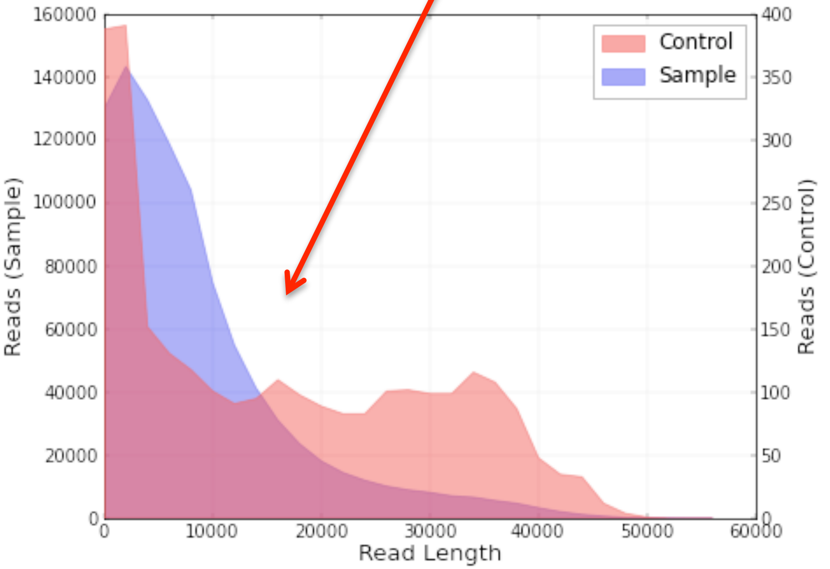
Control reads

- SMRT portal does not filter out control reads unless the protocol is included.

Control Sequence	2kb_Control	Number of Control Reads	2775
Fraction Control Reads	0.0028828862522167057	Control Subread Accuracy	0.861804283567377
Control Polymerase Read Length N50	29699.0	Control Polymerase Read Length 95%	39372
Control Polymerase Read Length	14950		

Control reads are longer than Sample reads indicating good sequencing but bad Sample DNA quality

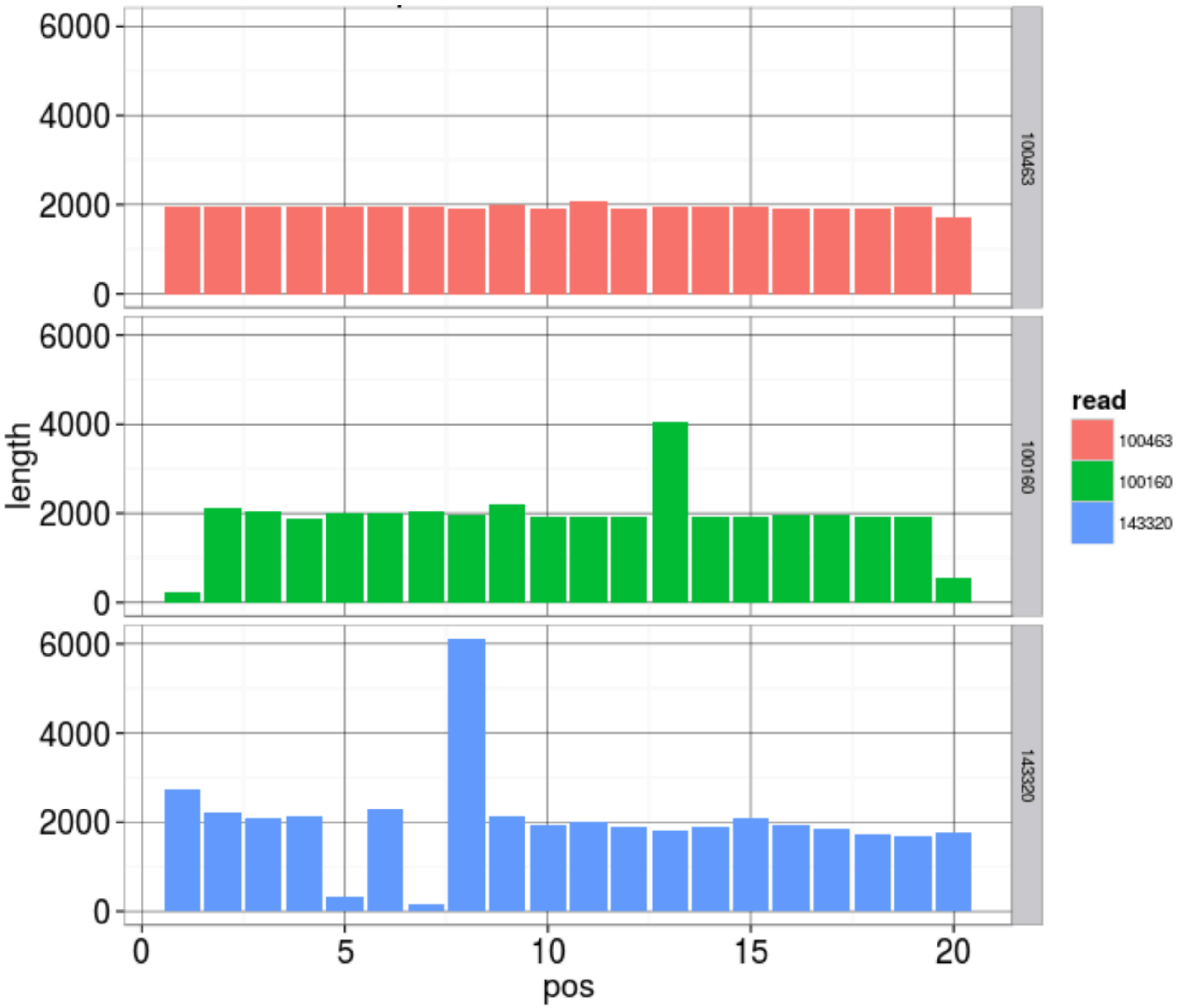




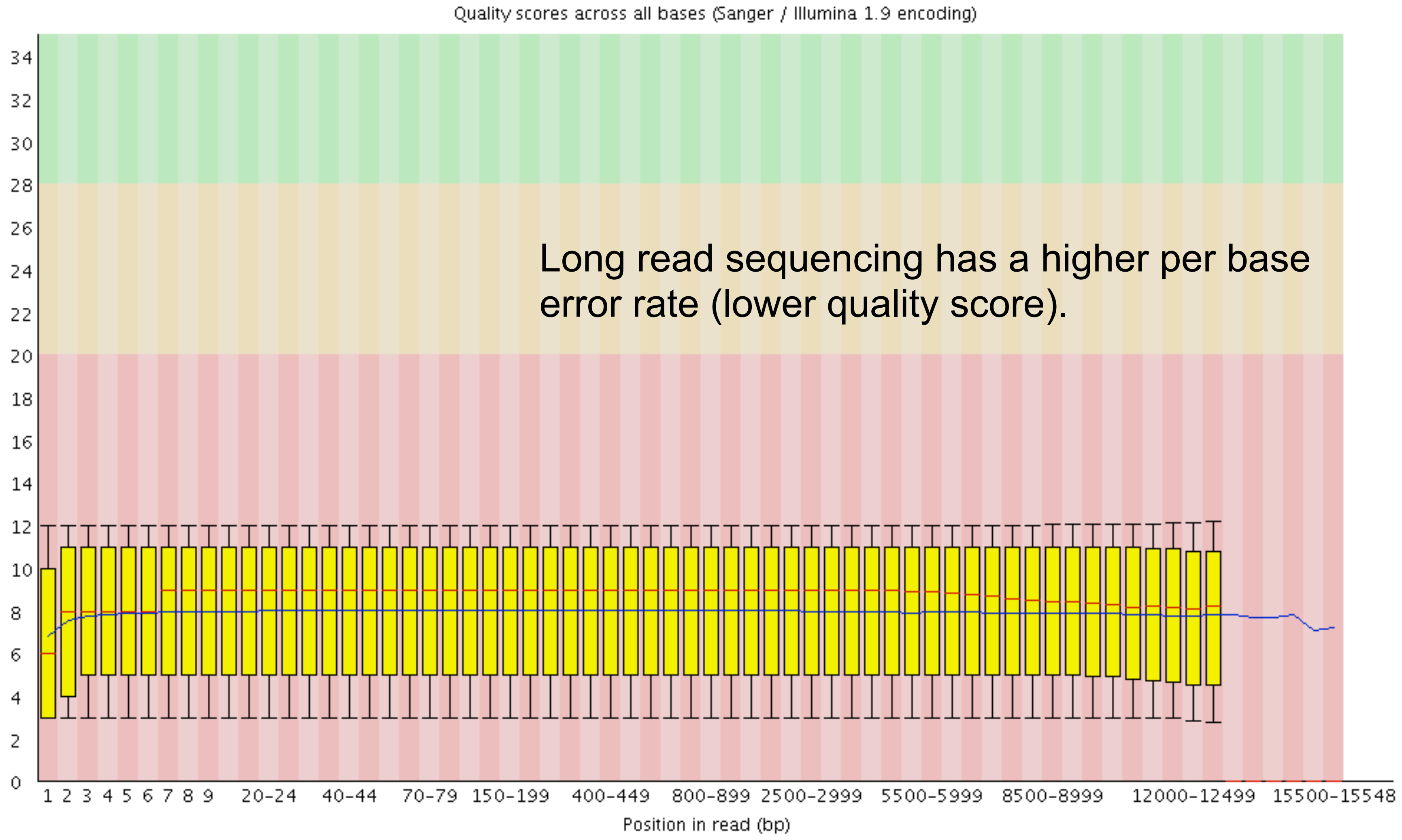
Adapter Trimming

SciLifeLab

SMRTbell adapter:
ATCTCTCTCTTTTCCTCCTCCTCCGTTGTTGTTGTTGAGAGAGAT

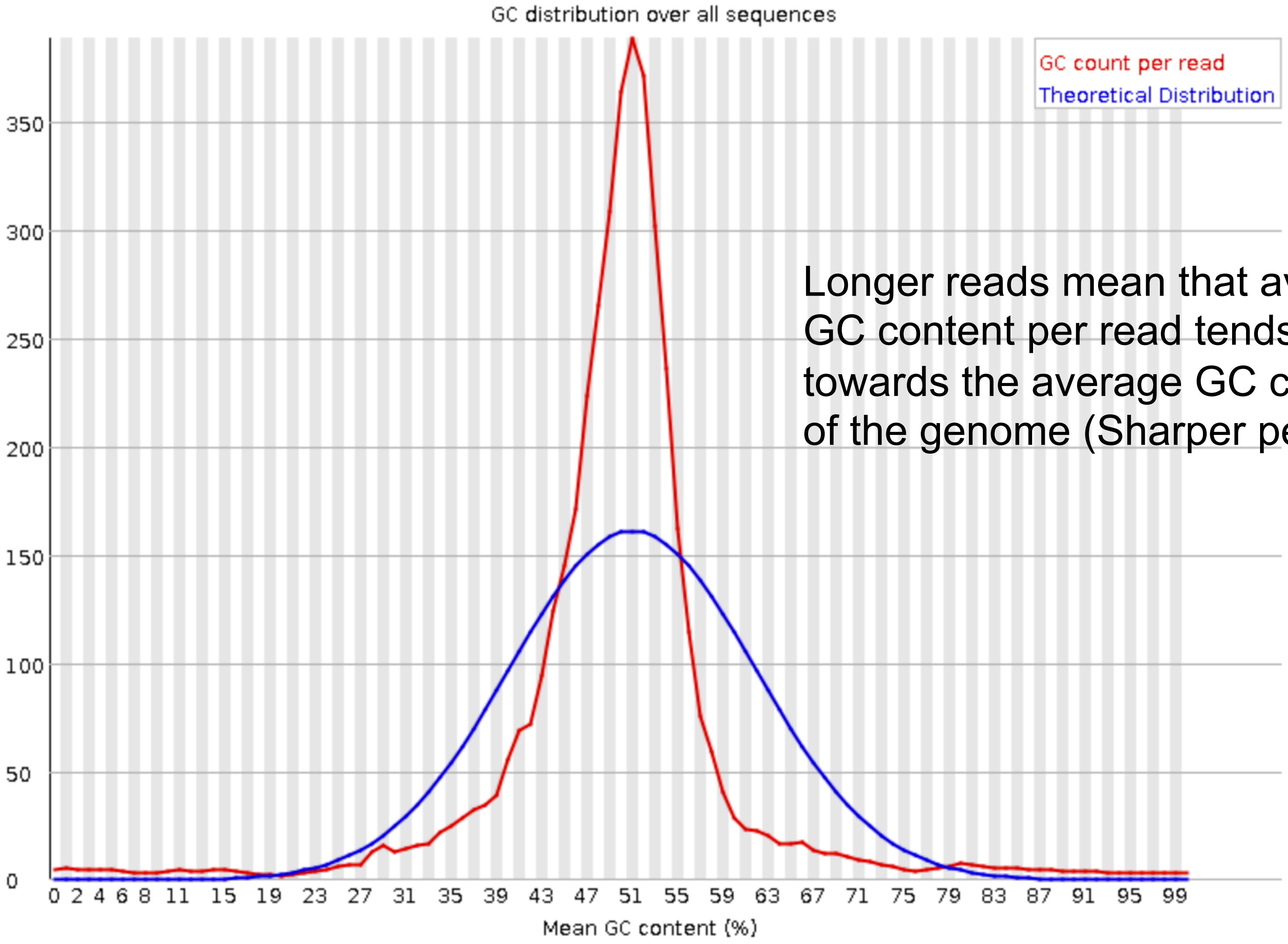


FastQC



Long read sequencing has a higher per base error rate (lower quality score).

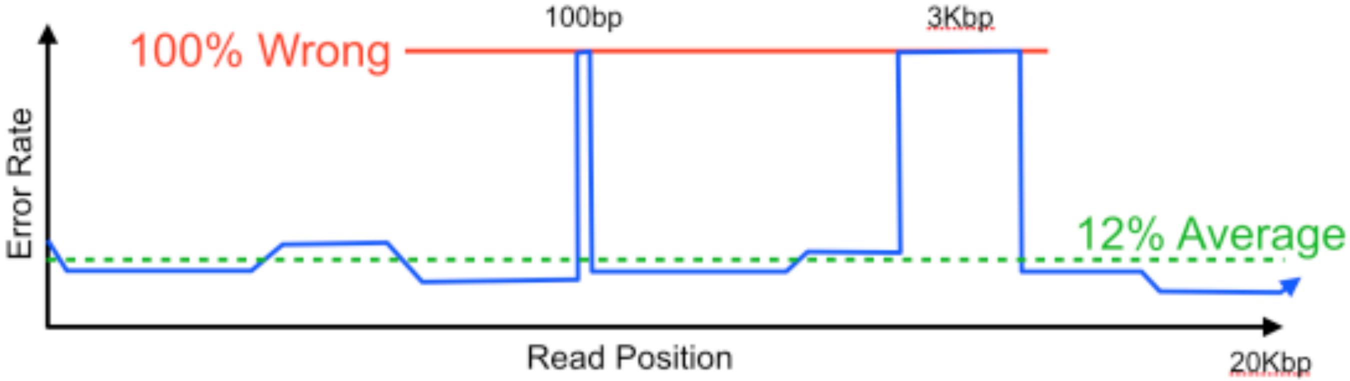
FastQC

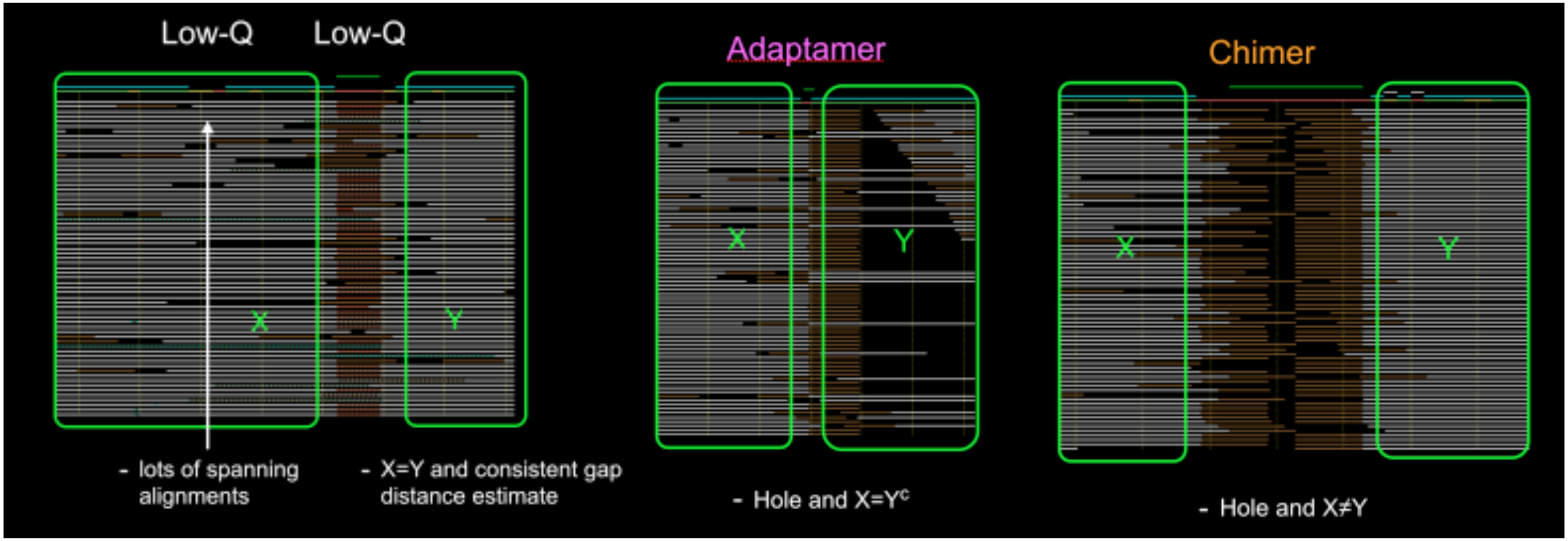


Longer reads mean that average GC content per read tends more towards the average GC content of the genome (Sharper peak).

Read Errors

Quality score is inferred from the light signal.
Strong light signal does not imply correct base calls.



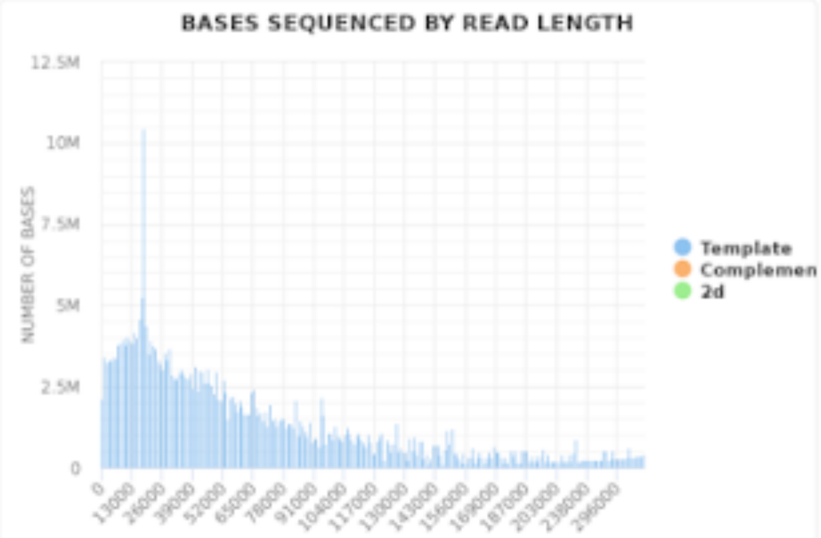


- lots of spanning alignments
- X=Y and consistent gap distance estimate
- Hole and X=Y^c
- Hole and X≠Y

Nanopore sequencing

- Single Molecule sequencing
- No fragmentation required, therefore read length is theoretically unlimited. (Pacbio has size selected library)
- Support is limited to the ONT community portal.
- Limited QC tools.
  - Poretools (output summary)
  - Porechop (adapters)

GGGCAATAACGTTTATGTTGGTTTCATGGTTTGGTCTAACTTTACC
DATA FOR ILLUSTRATIVE PURPOSES ONLY



Summary

- Illumina
  - Md5sum
  - Format
  - Data Quantity
  - Quality Scores
  - GC Content
  - Nucleotide bias
  - Duplication
  - Adapter content
  - Fragment distribution

- PacBio / Nanopore
  - Md5sum
  - Format
  - Data Quantity
  - GC Content
  - Adapter & Control check
  - Subread distribution

Information on sequence QC issues

- Sequencing Fail
  https://sequencing.qcfail.com

- SEQanswers
  http://seqanswers.com

- BioStar
  https://www.biostars.org

- Bioinformatics StackExchange
  https://bioinformatics.stackexchange.com