2023.06.23

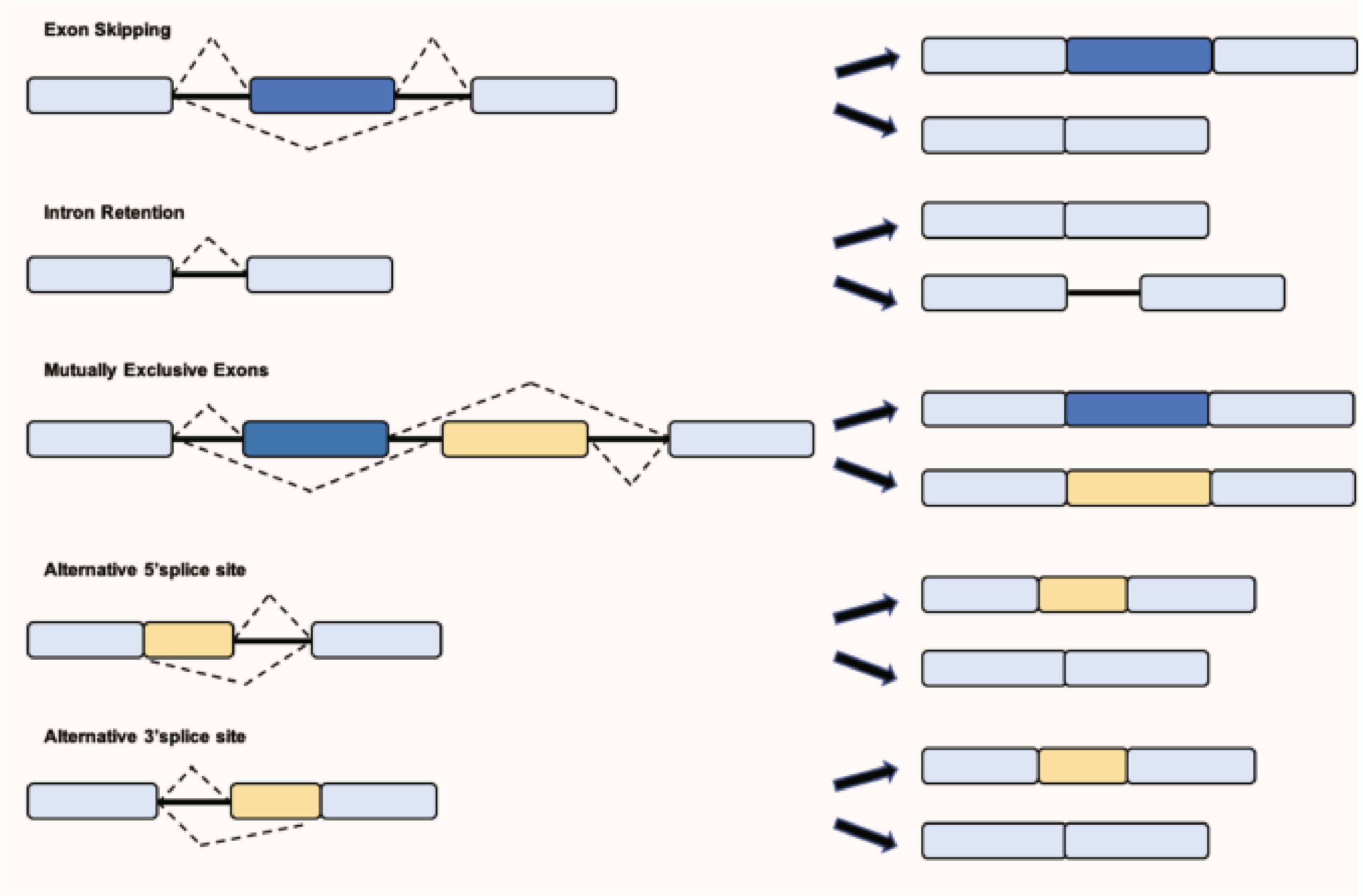
RNA Sequencing Quality Control

FastQC

Seoungbeom Jin

RNA sequencing – alternative splicing

A cellular process in which **exons** from the **same gene** are **joined in different combinations**, leading to different, but related, **mRNA transcripts**. These **mRNAs** can be **translated to produce different proteins** with distinct **structures and functions** all from a single gene.

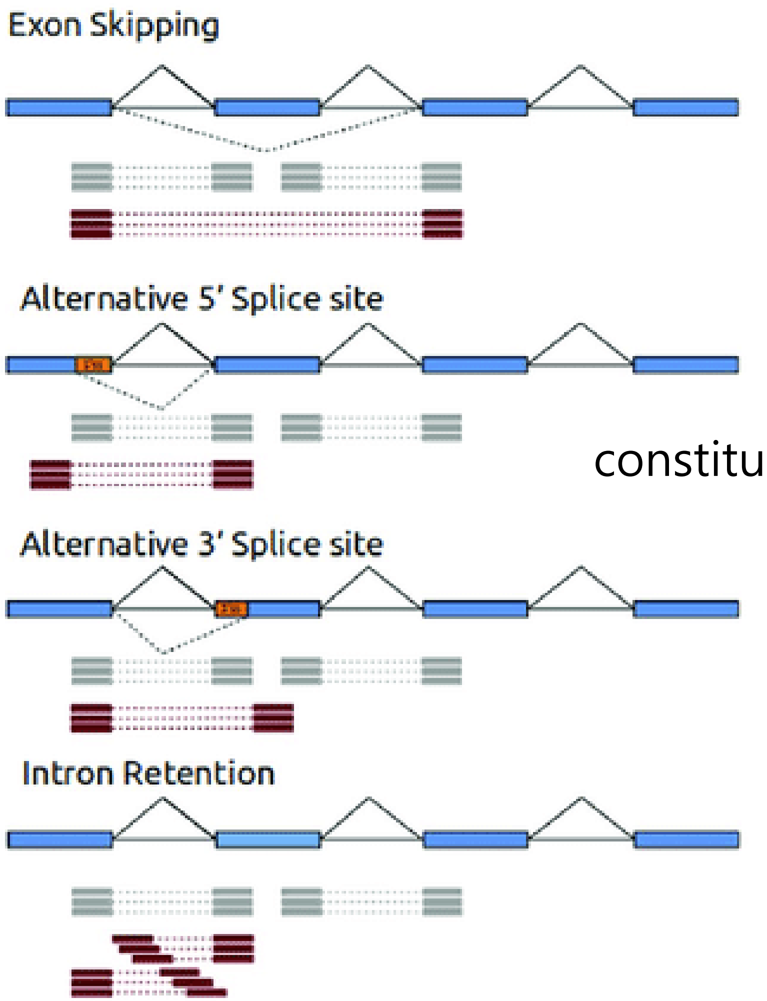


- Exon skipping : intron이 제거될 때 exon과 함께 제거
- Intron retention : intron 부분이 제거되지 않음
- Mutually exclusive exons : exons의 선택적인 제거
- Alternative 5′ SS : exon의 5’부분과 함께 제거
- Alternative 3′ SS : exon의 3’부분과 함께 제거

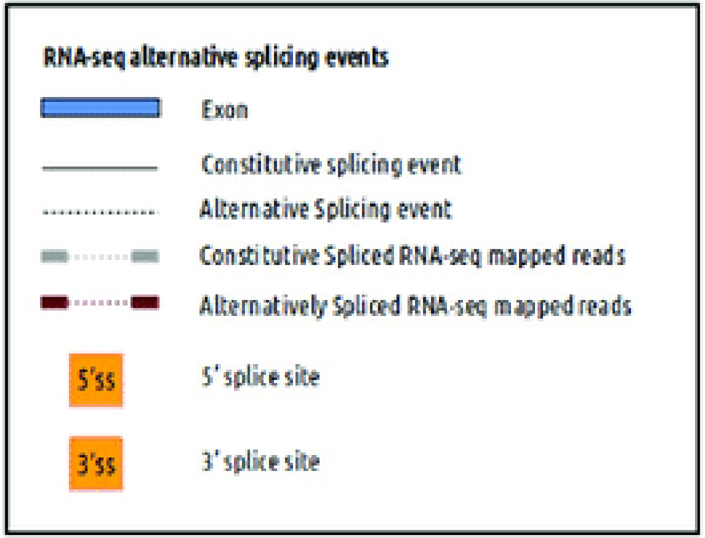
https://www.genome.gov/genetics-glossary/Alternative-Splicing, 2023-06-22
Yuanjiao Zhang et al., *Signal Transduction and Targeted Therapy*, 2021

RNA sequencing – alternative splicing

Paired-end sequencing가 alternative splicing alignment에 유리한 이유



constitutive

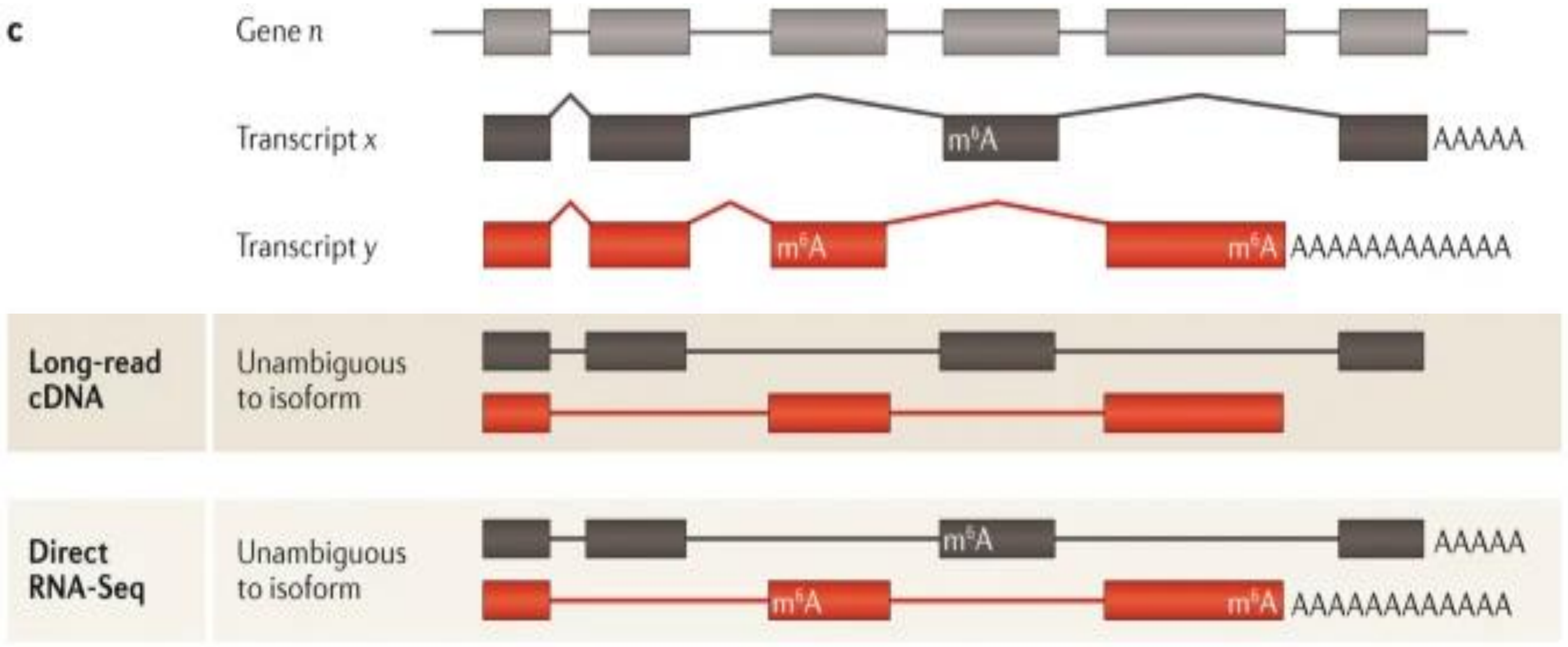

RNA-seq alternative splicing events
Exon
Constitutive splicing event
Alternative Splicing event
Constitutive Spliced RNA-seq mapped reads
Alternatively Spliced RNA-seq mapped reads
5'ss
5' splice site
3'ss
3' splice site

- Paired-end sequencing은 fragment의 3'과 5'에서 pair로 sequencing이 일어나는데 이때 read pair의 사이의 거리를 알 수 있기 때문에 이 정보를 활용할 수 있다
- No alternative splicing : read pair의 거리 유지
- Alternative splicing : read pair의 거리가 늘어남

RNA sequencing – alternative splicing

Paired-end sequencing가 alternative splicing alignment에 대한 한계

- Other **biases and limitations** can result from the myriad computational methods that can be applied to RNA-seq data, such as differences in how **ambiguous** or **multi-mapped** reads are handled.
- the greatest potential for fundamentally addressing the inherent **limitations of short-read cDNA sequencing** lies with **long-read cDNA(e.g. full-length isoform reads)** sequencing and **dRNA-seq(direct RNA sequencing)** methods.



Rory Stark et al., *Nature Reviews Genetics*, 2019

RNA sequencing – Quality Control

Pre trimming read QC

- Check pre trimming read quality
- Using FastQC

Trimming

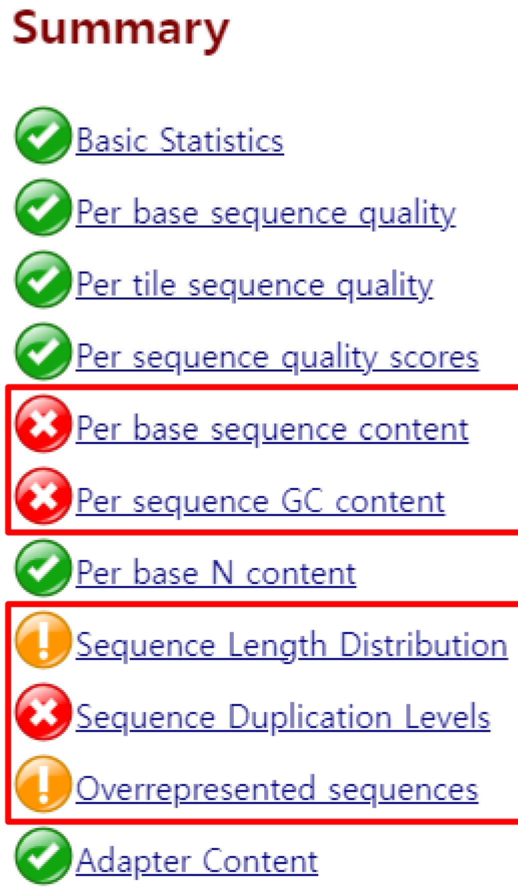
- Remove low read quality (quality score < 20)
- Remove short read (read length < 20)
- Remove adapter sequence

Post trimming read QC

- Check post trimming read quality
- Using FastQC

RNA sequencing – Quality Control

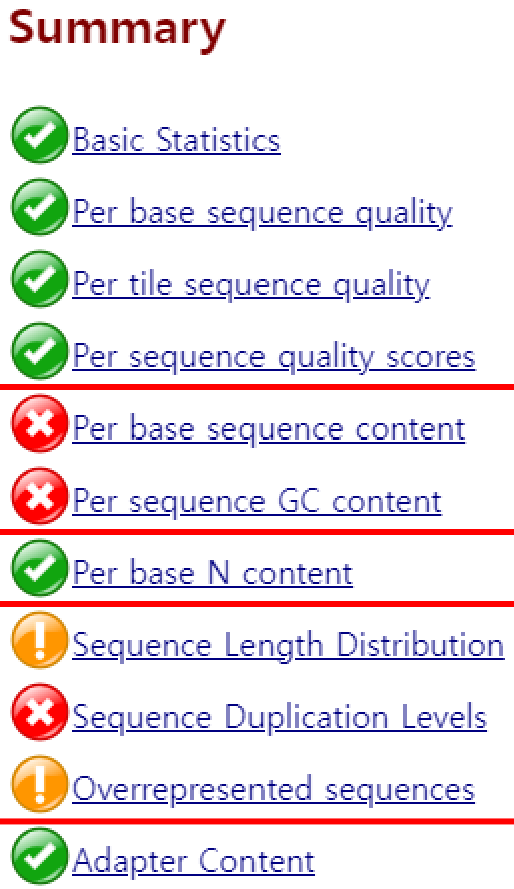
Pre trimming



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

Post trimming

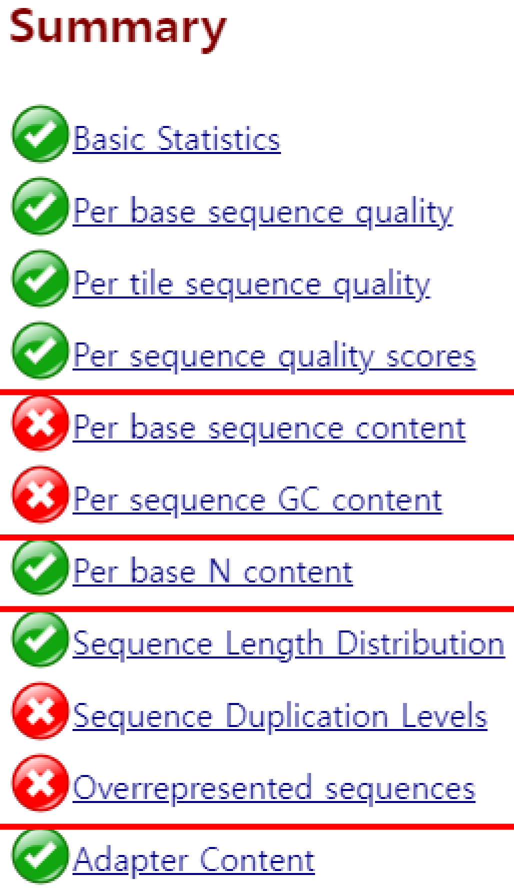


Summary

- ✓ 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

BID01_1

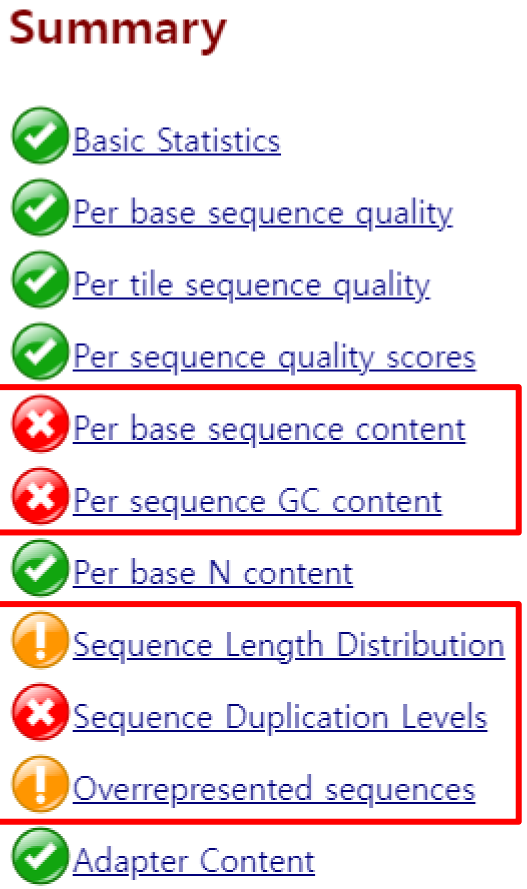
Pre trimming



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

Post trimming



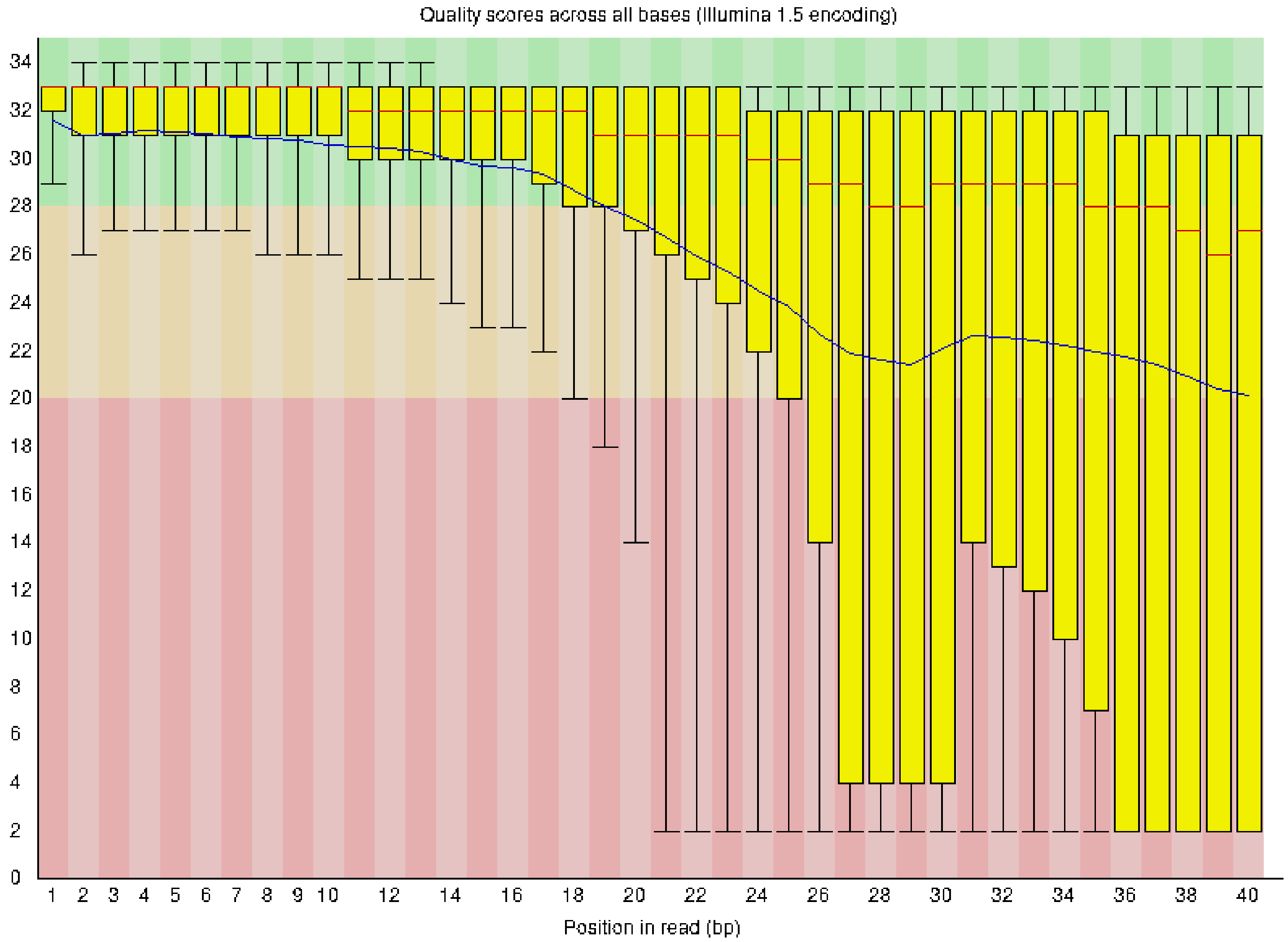
Summary

- ✓ 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

BID01_2

RNA sequencing – Quality Control

Pre base sequence quality



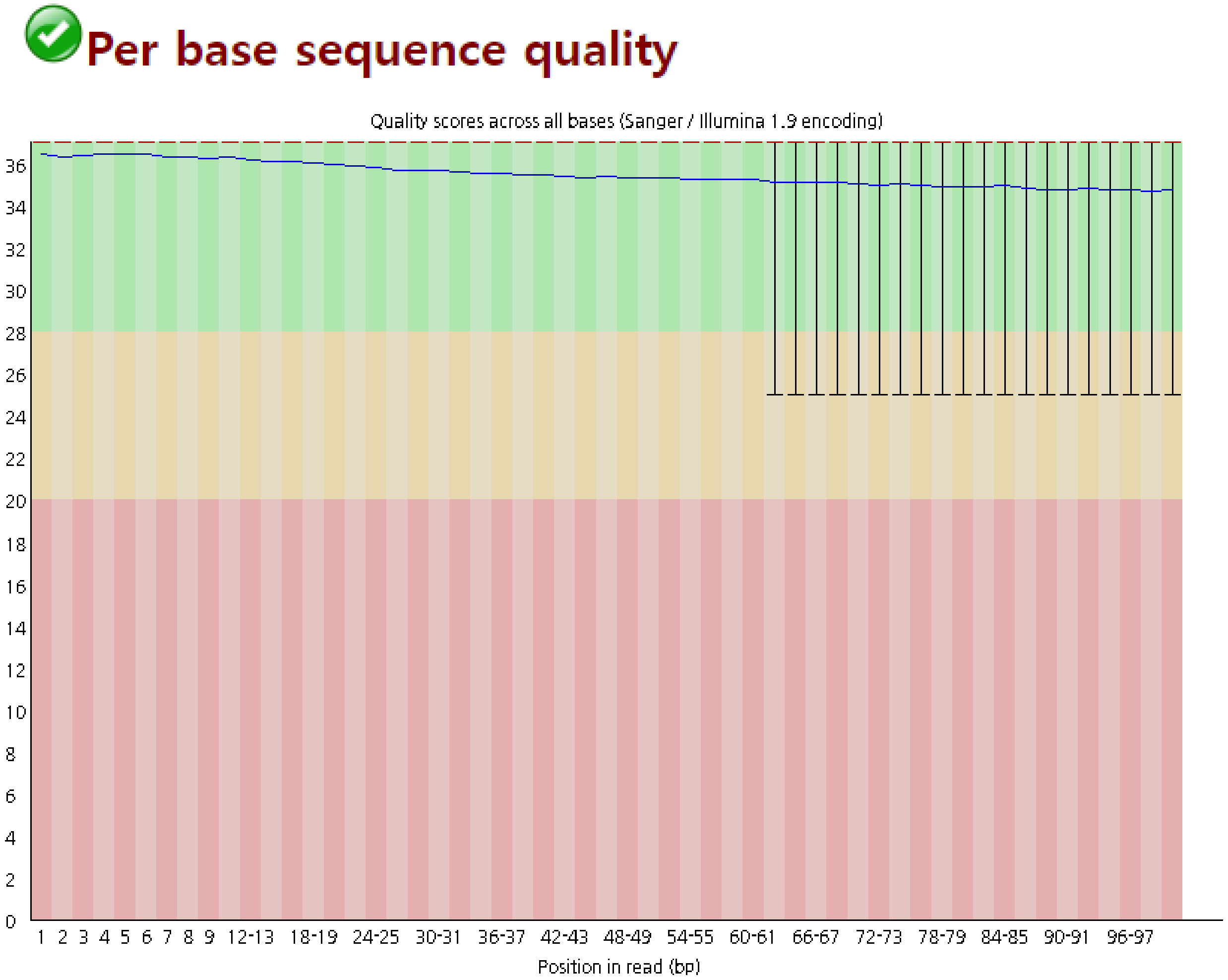
Bad Illumina Data

- 모든 read의 각 position에서 quality score의 분포를 나타낸다.

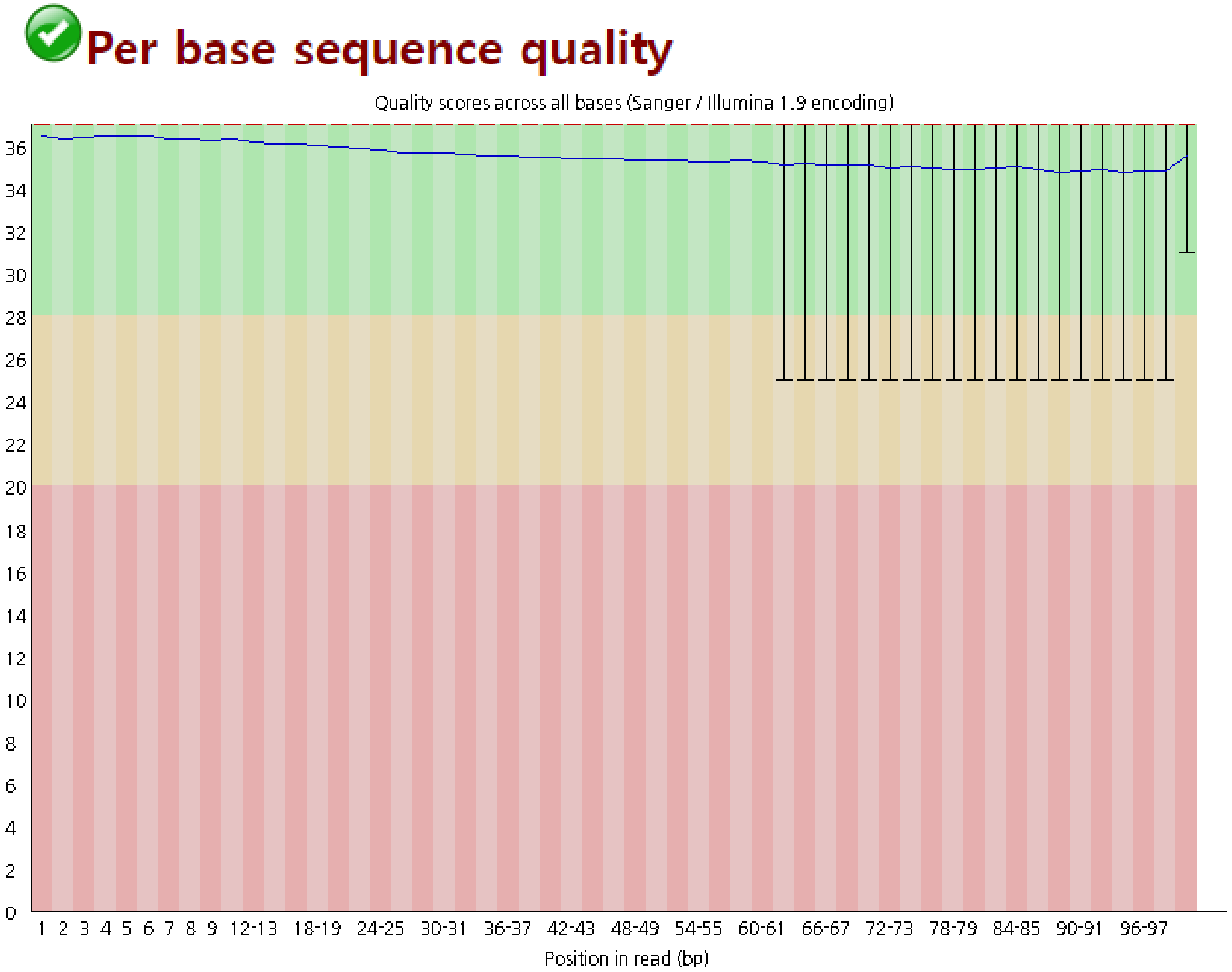
- X axis : 모든 read의 base position
- Y axis : quality score
- Yellow box : represents 25 ~ 75 Precentiles
- Whiskers line : represents 10 ~ 90 Precentiles
- Red line : median
- Blue line : average quality score

RNA sequencing – Quality Control

Pre base sequence quality(BID01_1)



Pre trimming

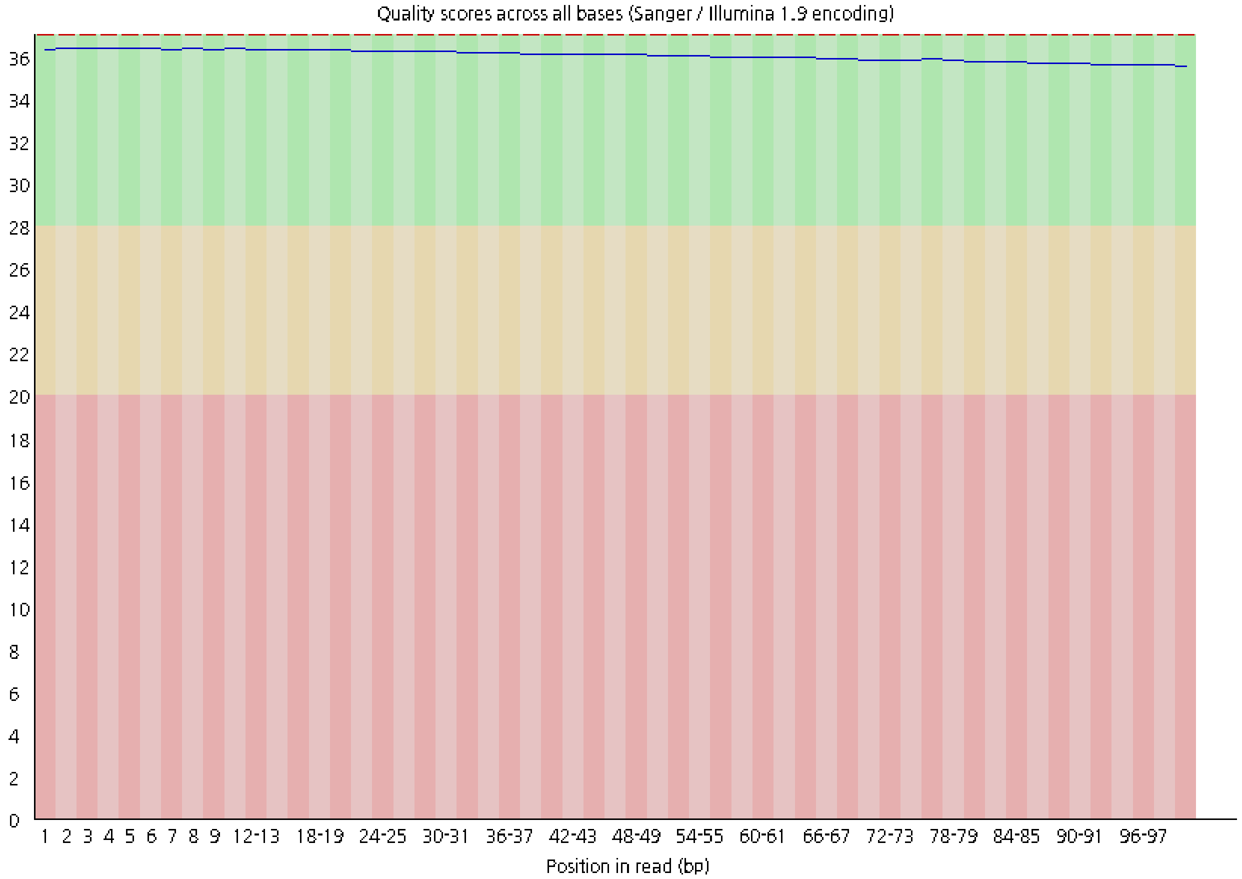


Post trimming

RNA sequencing – Quality Control

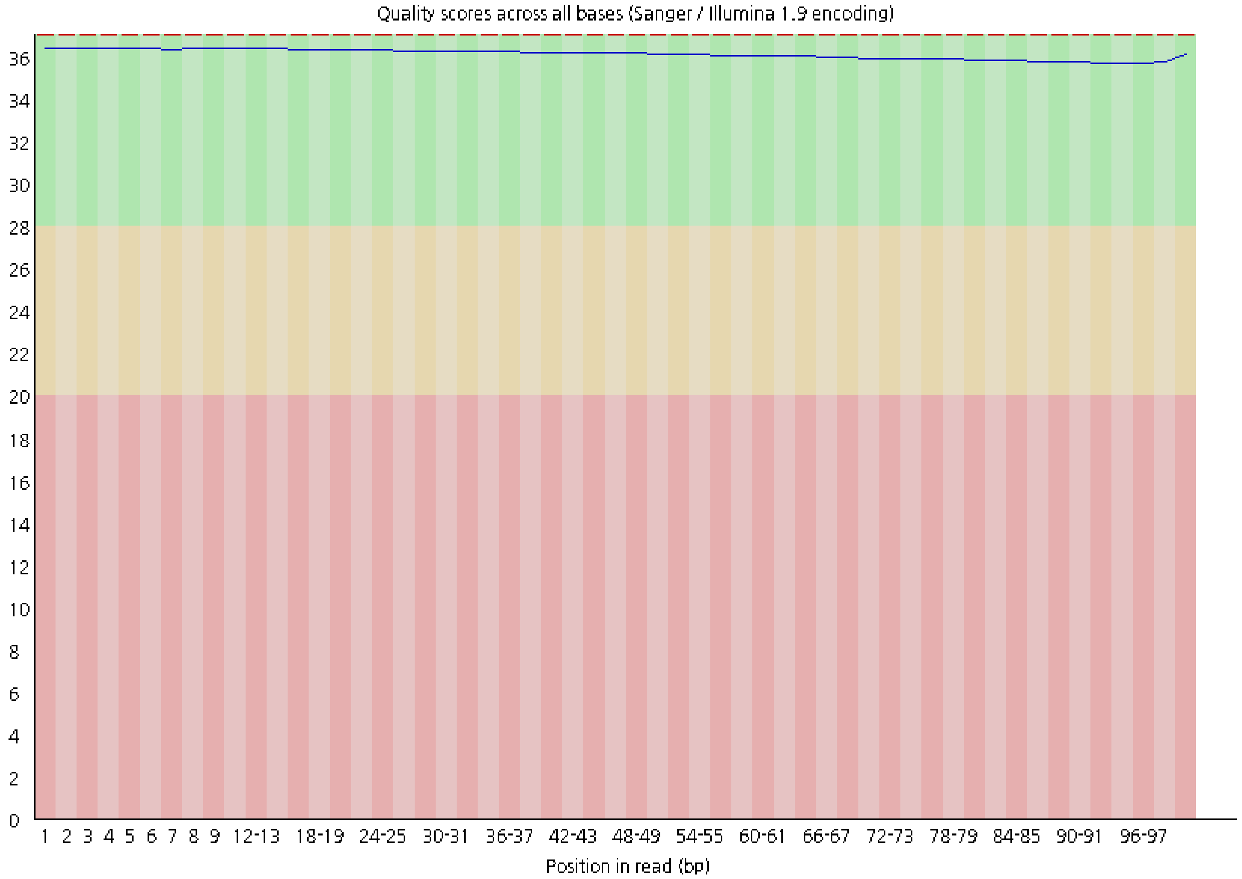
Pre base sequence quality(BID01_2)

✓ Per base sequence quality



Pre trimming

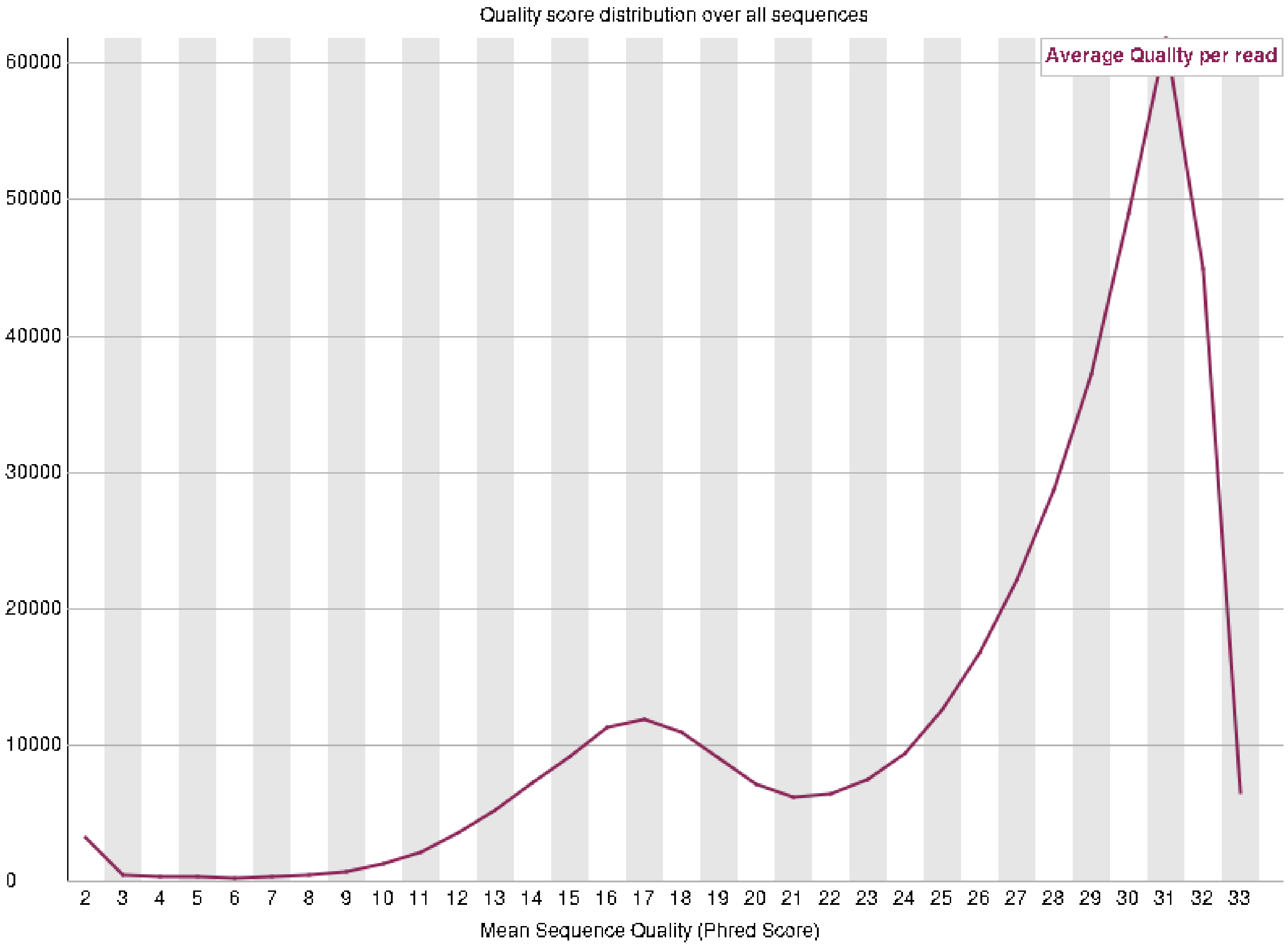
✓ Per base sequence quality



Post trimming

RNA sequencing – Quality Control

Pre sequence quality scores



Bad Illumina Data

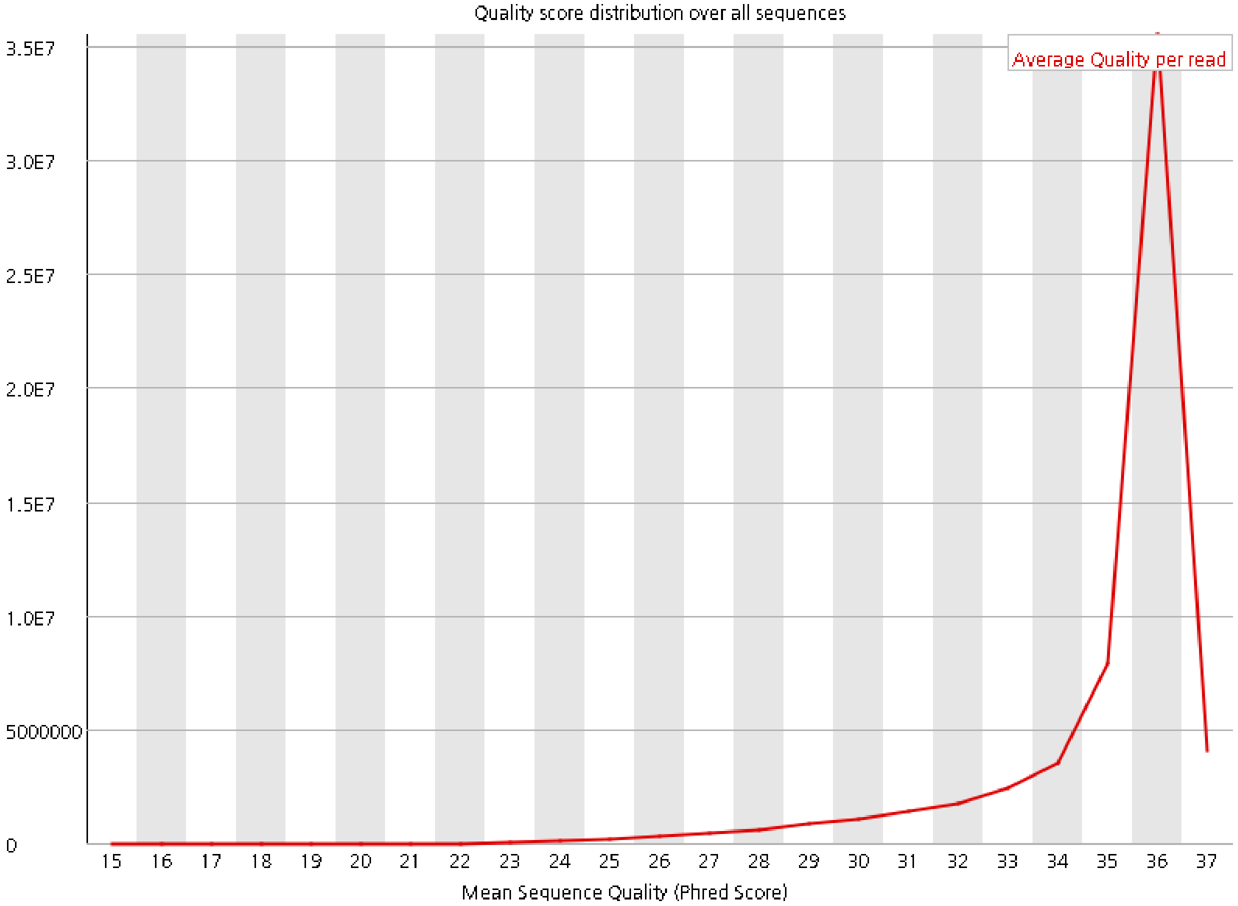
- 모든 read의 평균 quality score를 나타낸다.

- X axis : Mean sequence quality(Phred score)
- Y axis : Number of read

RNA sequencing – Quality Control

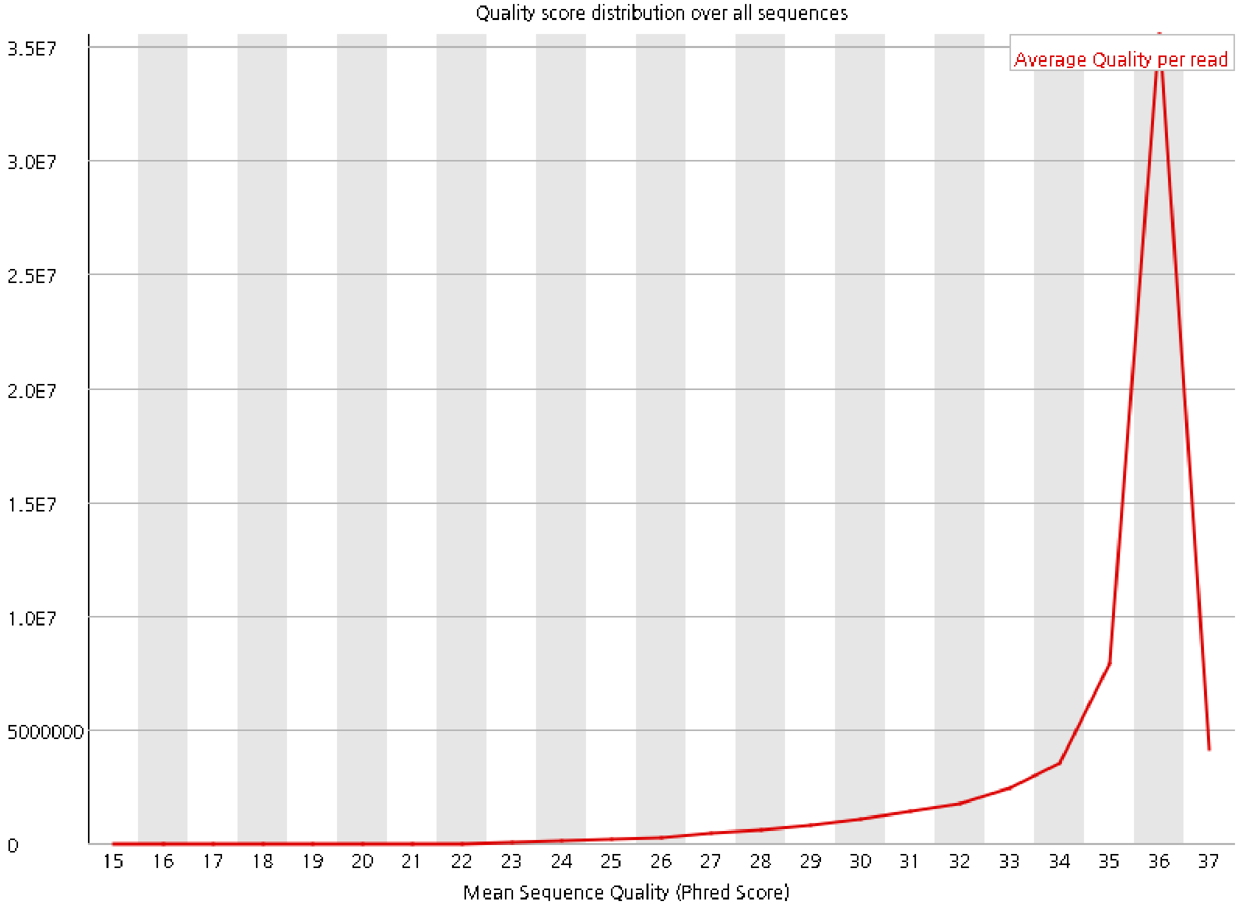
Pre sequence quality scores(BID01_1)

✓ Per sequence quality scores



Pre trimming

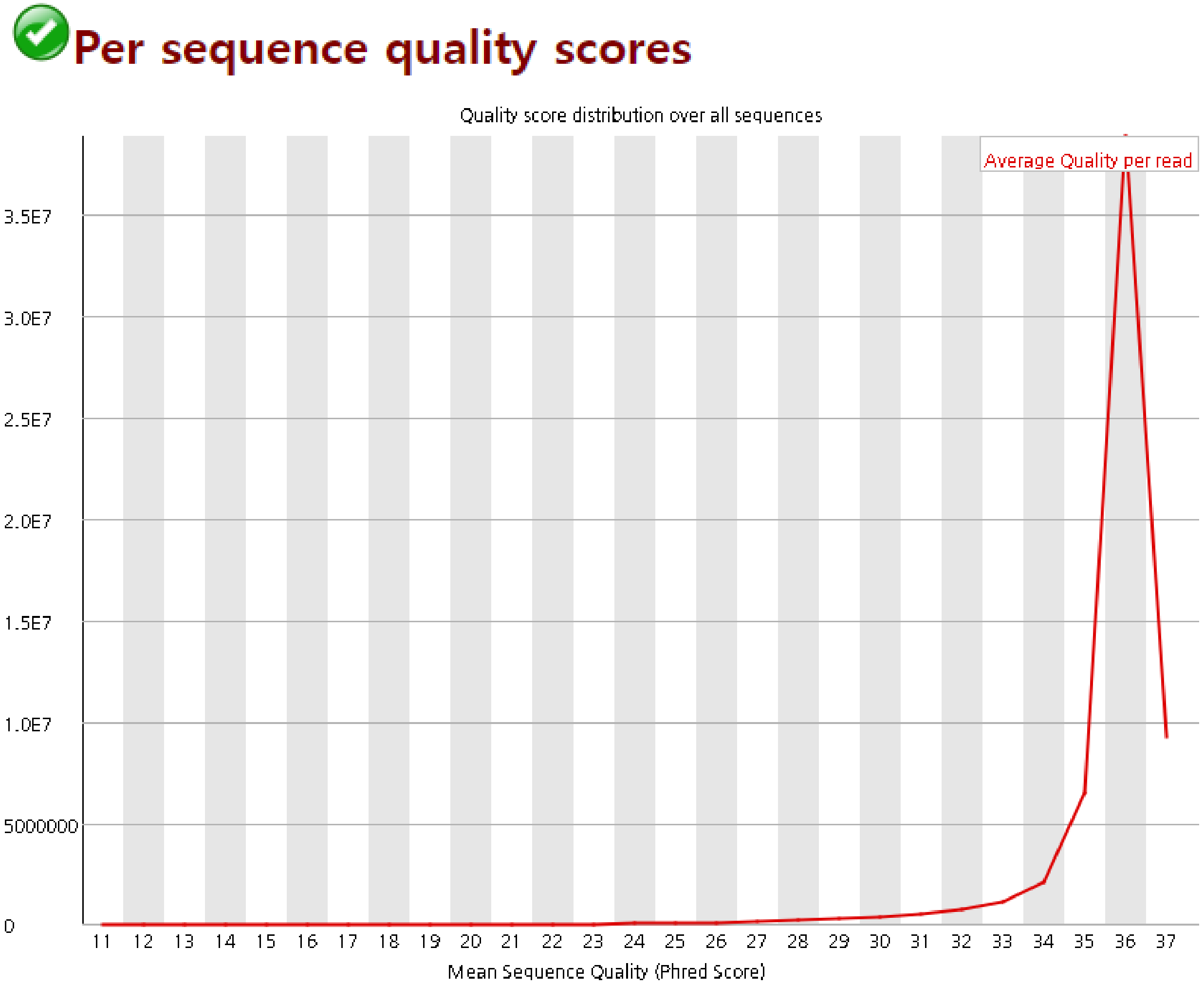
✓ Per sequence quality scores



Post trimming

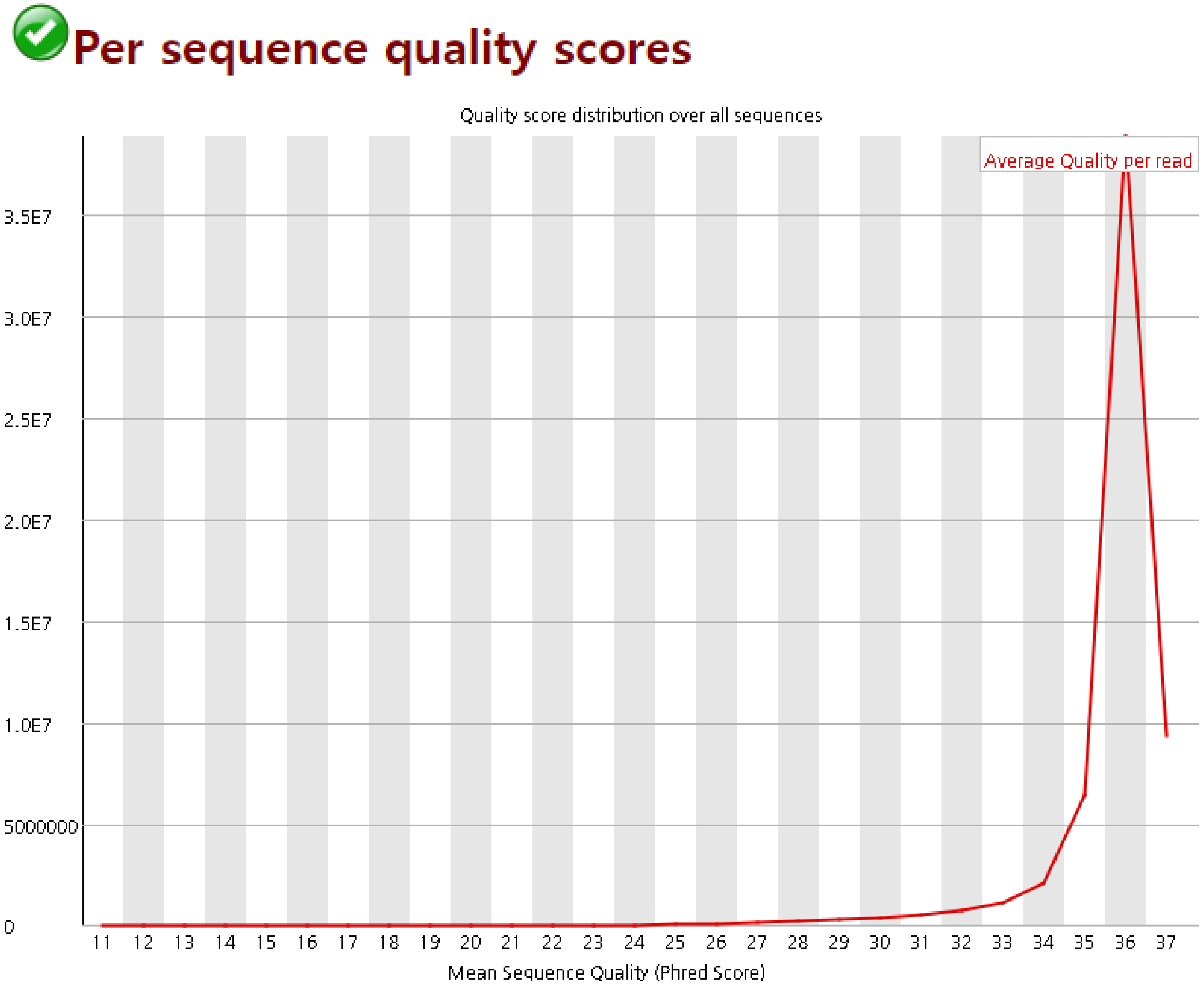
RNA sequencing – Quality Control

Pre sequence quality scores(BID01_2)



✓ Per sequence quality scores

Pre trimming

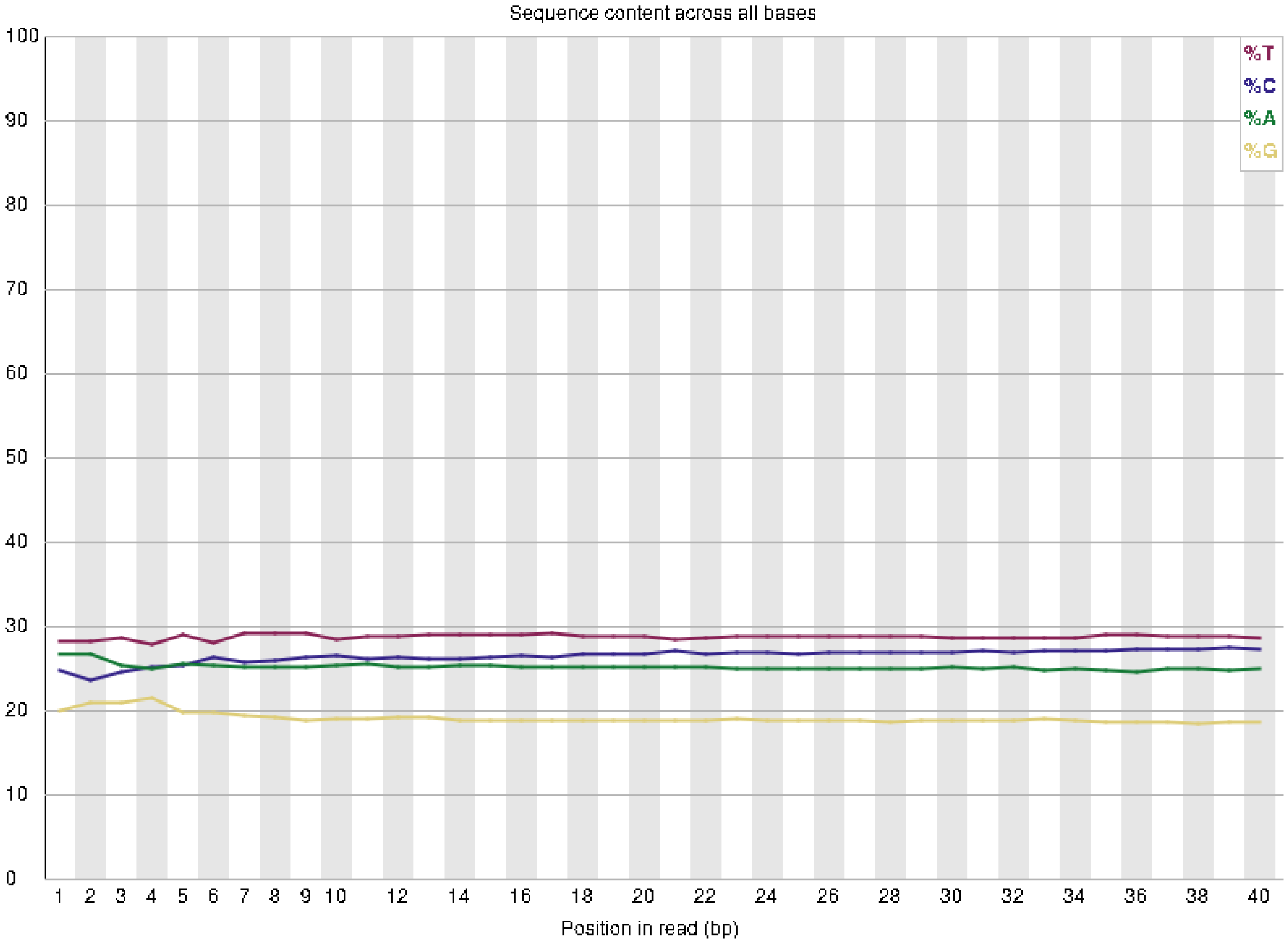


✓ Per sequence quality scores

Post trimming

RNA sequencing – Quality Control

Pre base sequence content



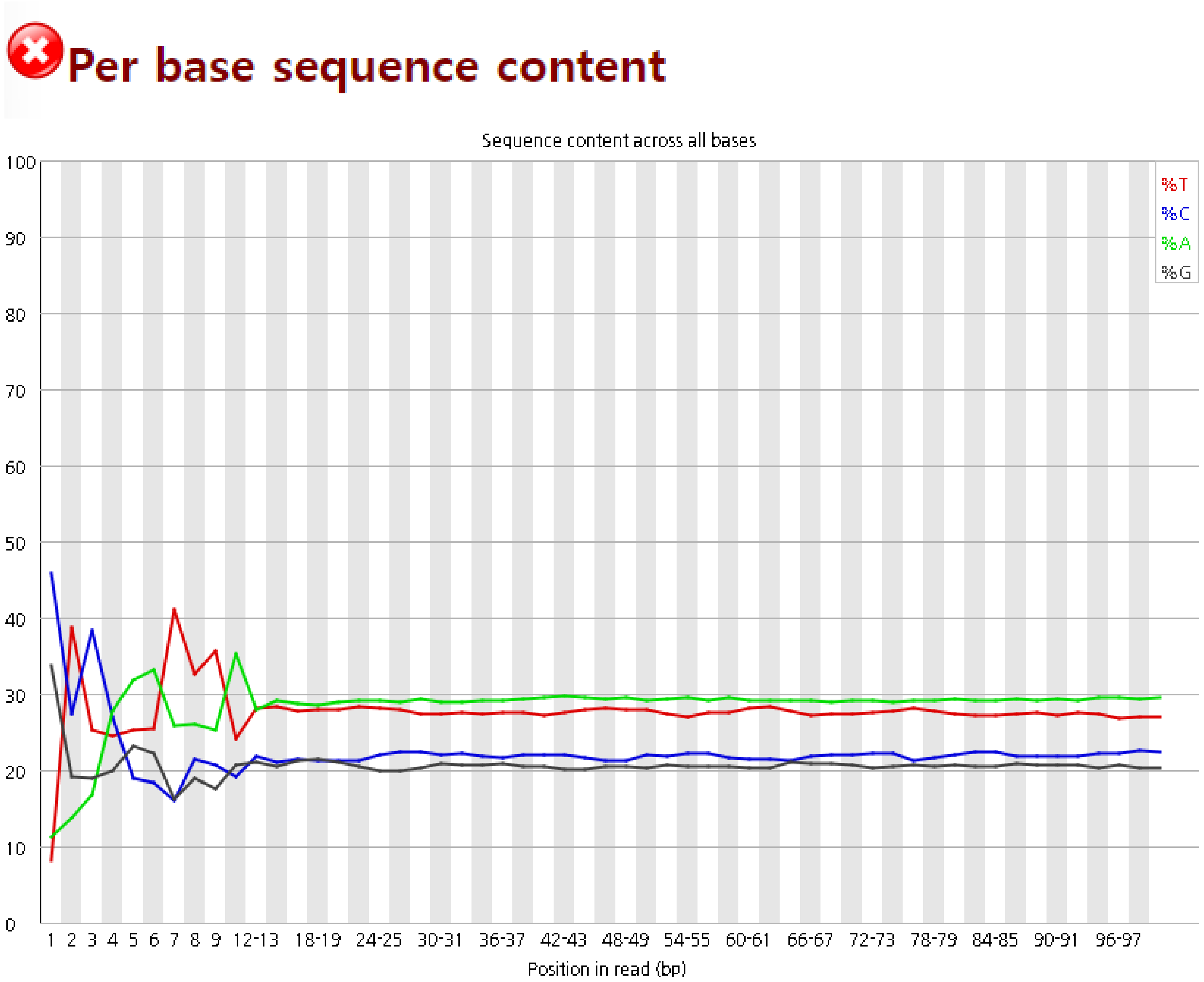
Good Illumina Data

- 각 base position에 대해 base의 각각의 비율을 나타냄
- Human의 base의 수는 균등하므로 plot의 line들은 서로 평행하게 나타나야 한다

- X axis : 모든 read의 base position
- Y axis : Base Precentage
- Red line : T%
- Blue line : C%
- Green line : A%
- Yellow(black) line : G%

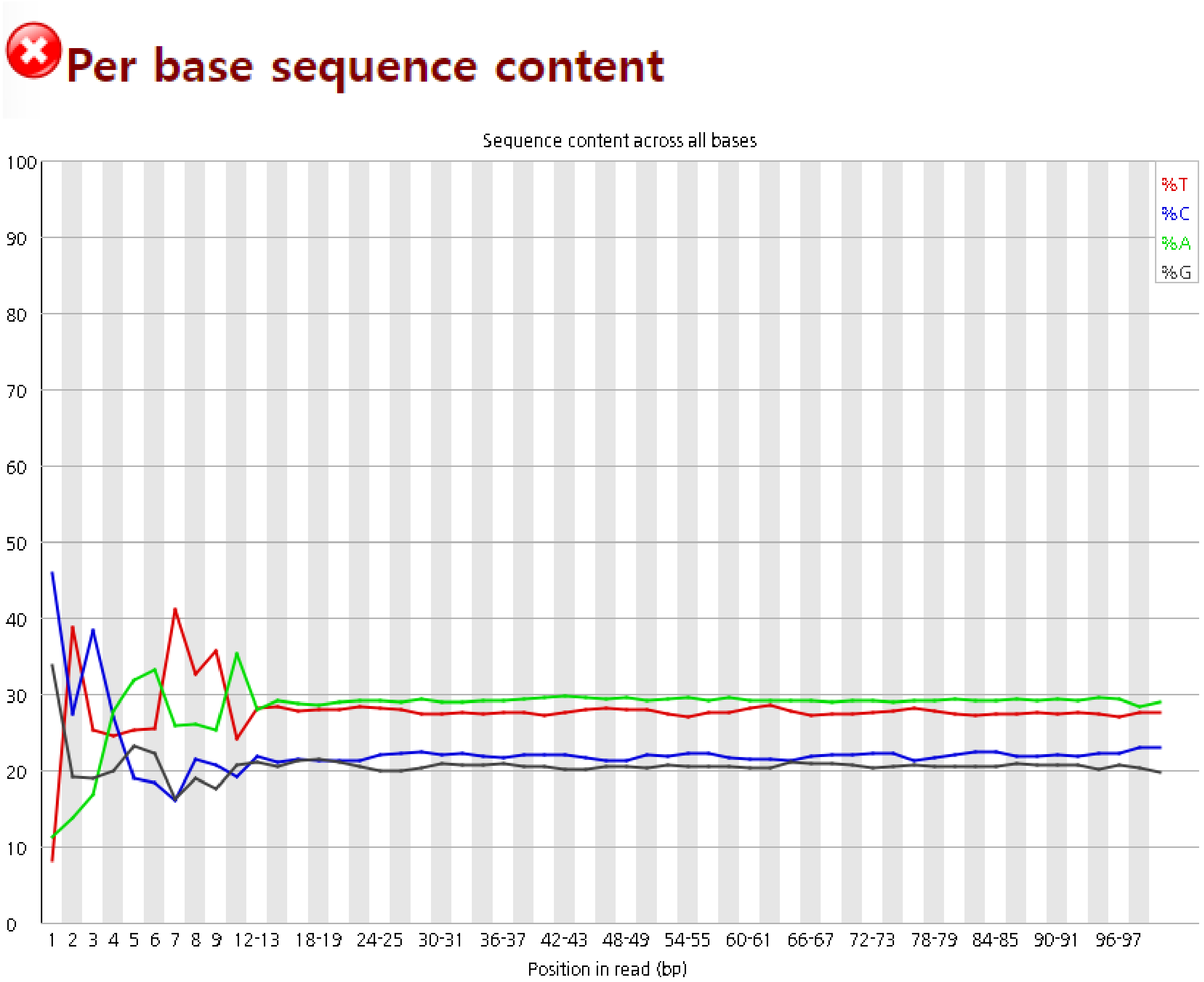
RNA sequencing – Quality Control

Pre base sequence content(BID01_1)



Per base sequence content

Pre trimming

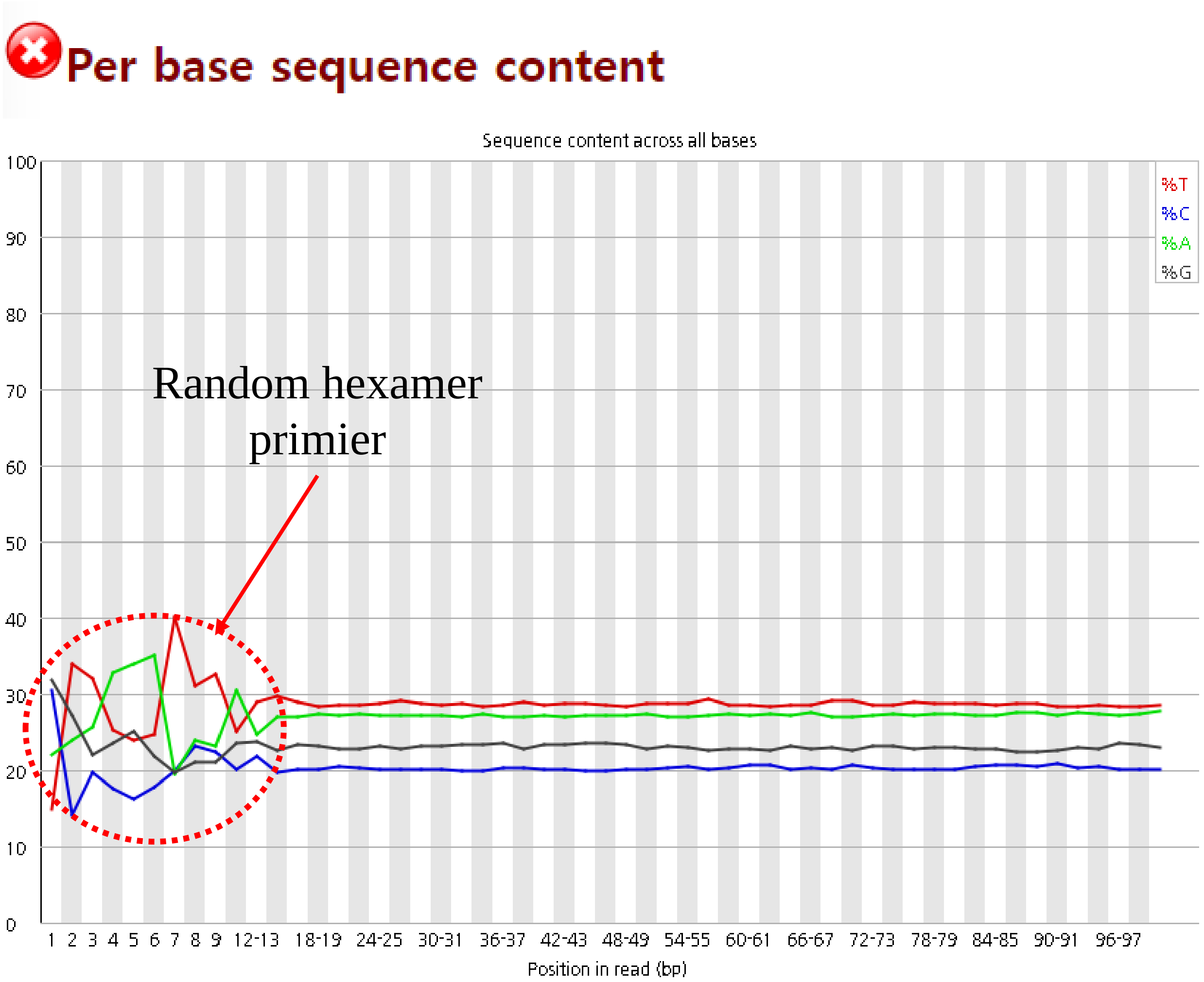


Per base sequence content

Post trimming

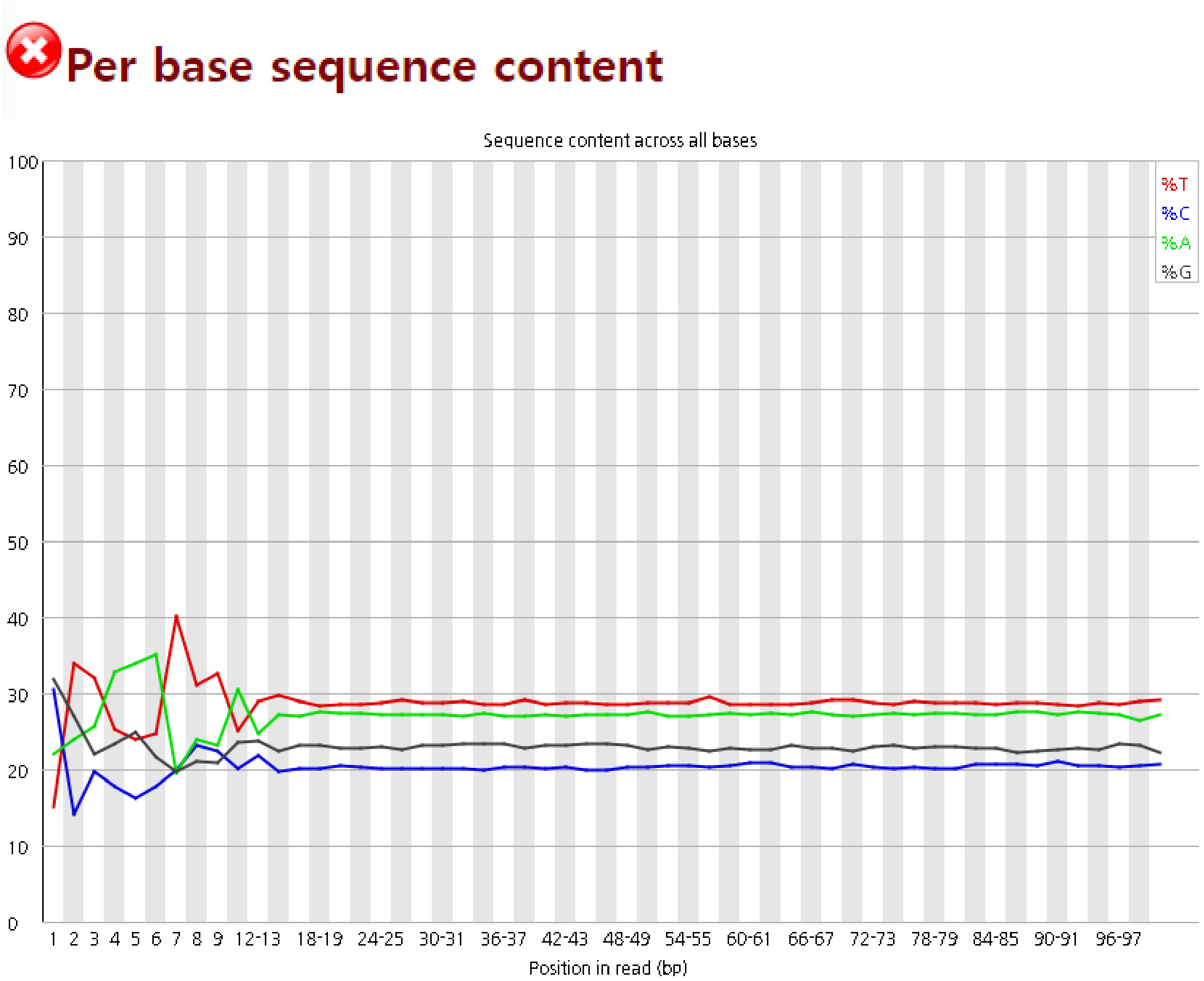
RNA sequencing – Quality Control

Pre base sequence content(BID01_2)



Random hexamer primier

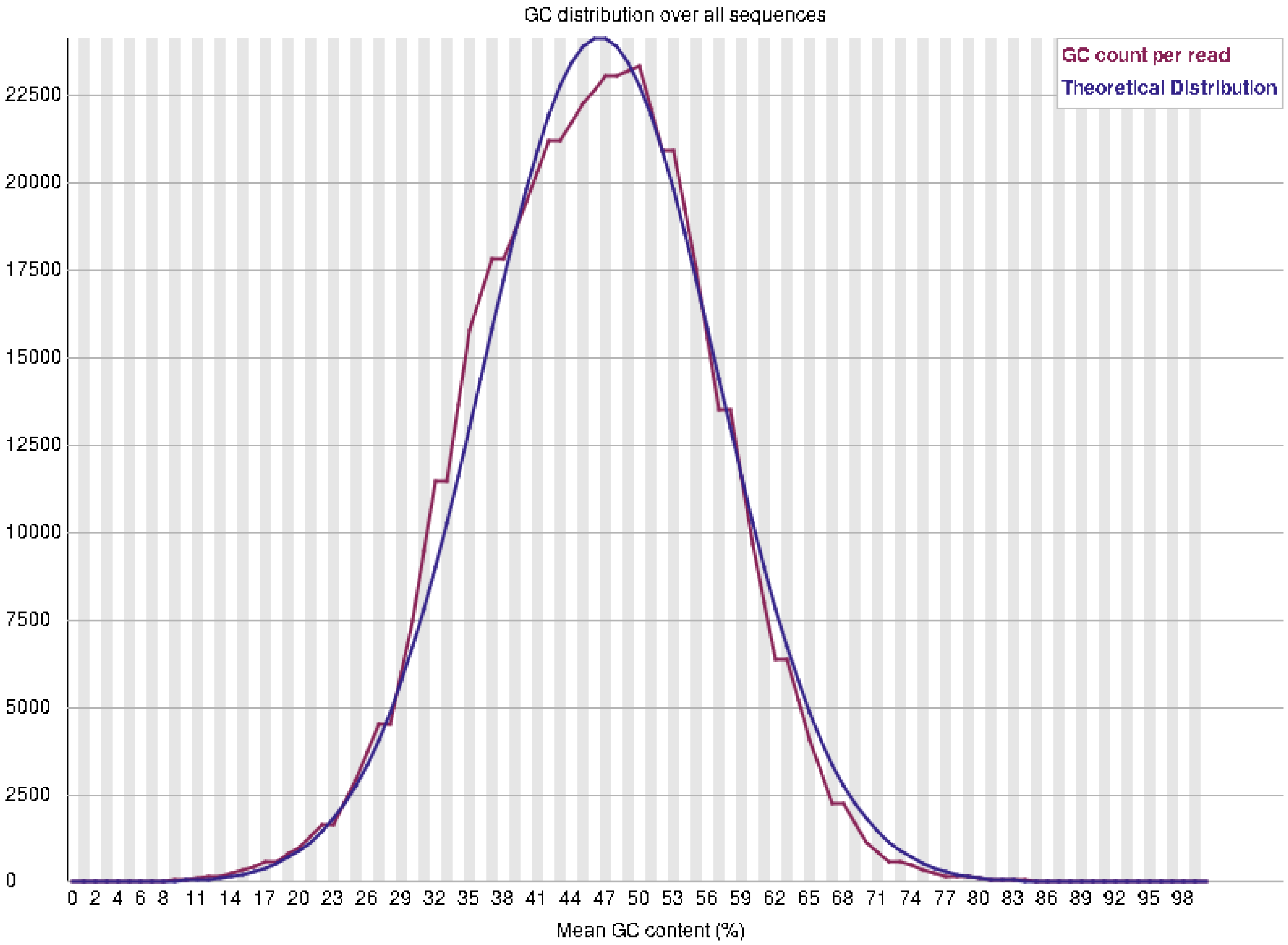
Pre trimming



Post trimming

RNA sequencing – Quality Control

Pre sequence GC content



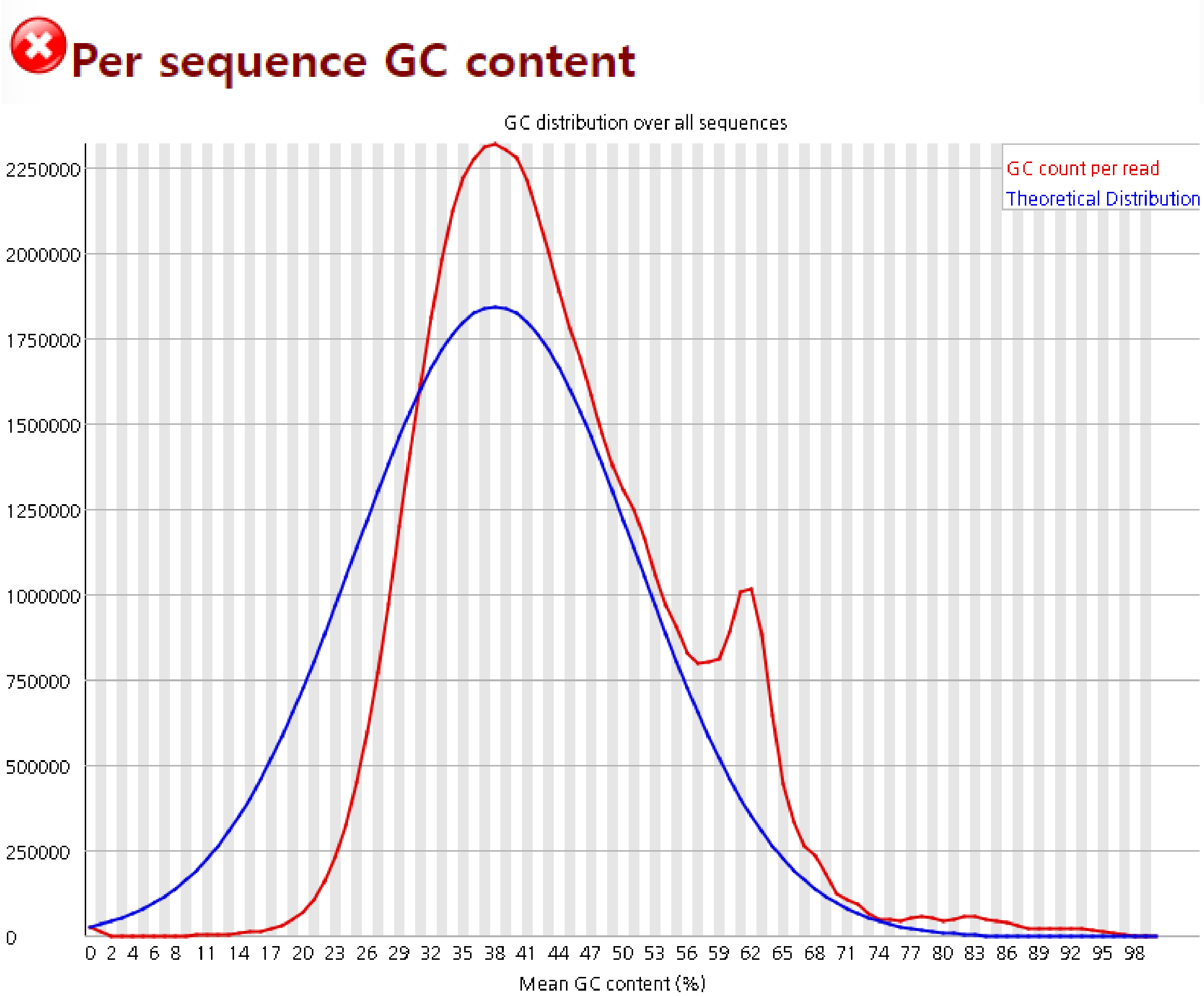
Good Illumina Data

- 각 read의 GC content 분포를 나타낸다.
- Human DNA에는 GC content의 분포가 정규분포로 나타난다.
- 때문에 예상되는 GC content의 분포와 관측되는 GC content의 분포를 비교한다.

- X axis : Mean GC content
- Y axis : Number of read
- Blue line : expected GC content 분포
- Read line : observed GC content 분포

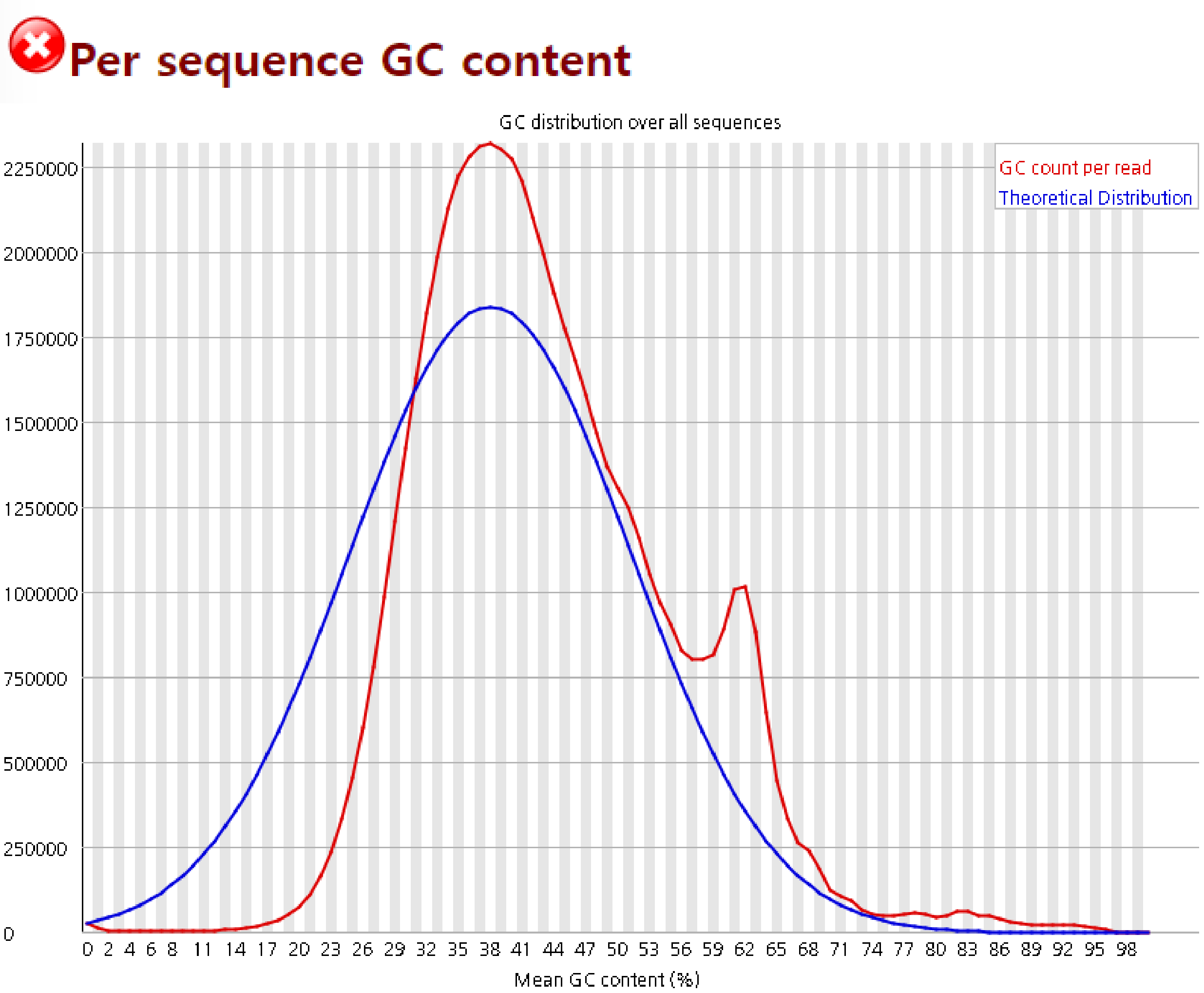
RNA sequencing – Quality Control

Pre sequence GC content(BID01_1)



Per sequence GC content

Pre trimming

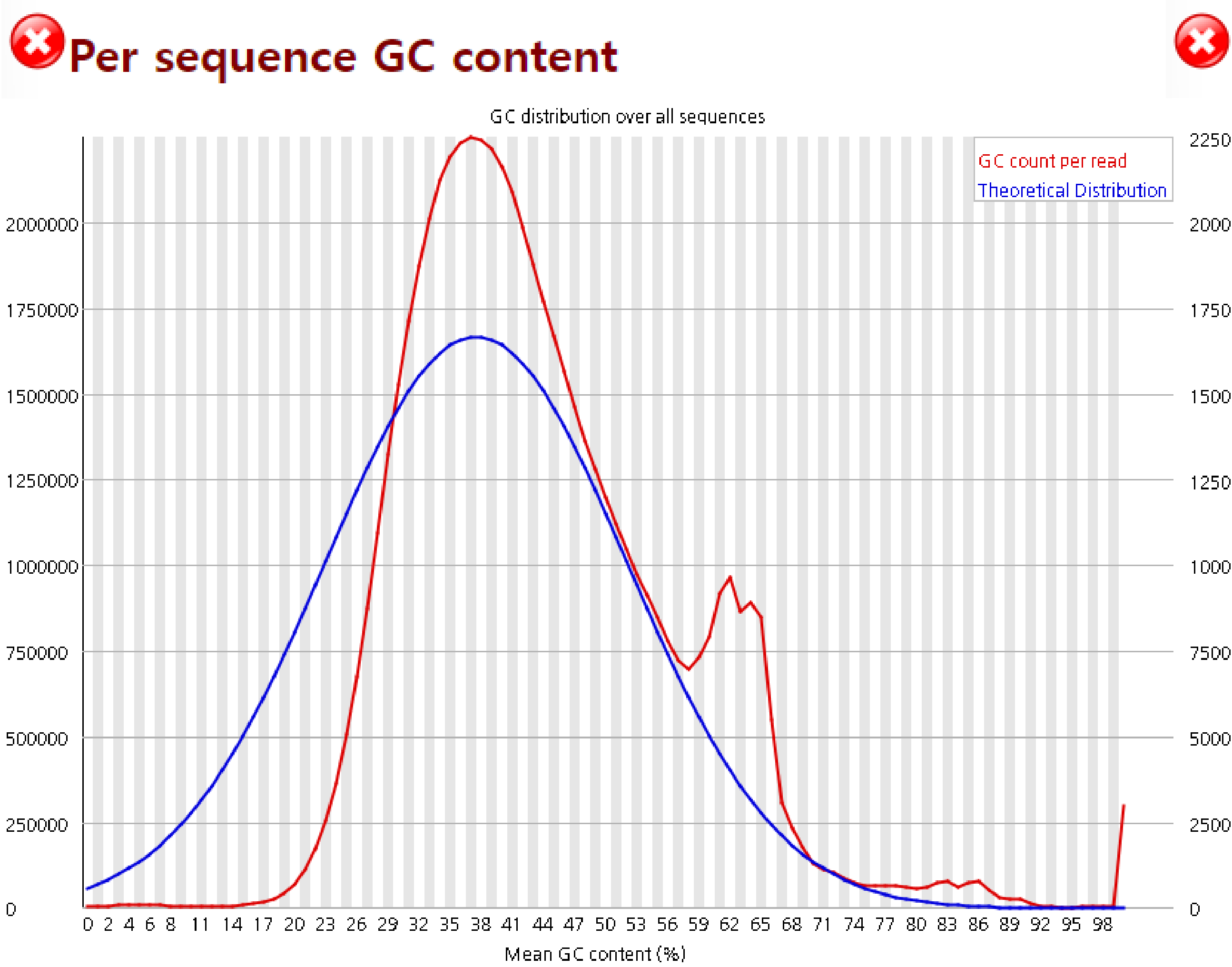


Per sequence GC content

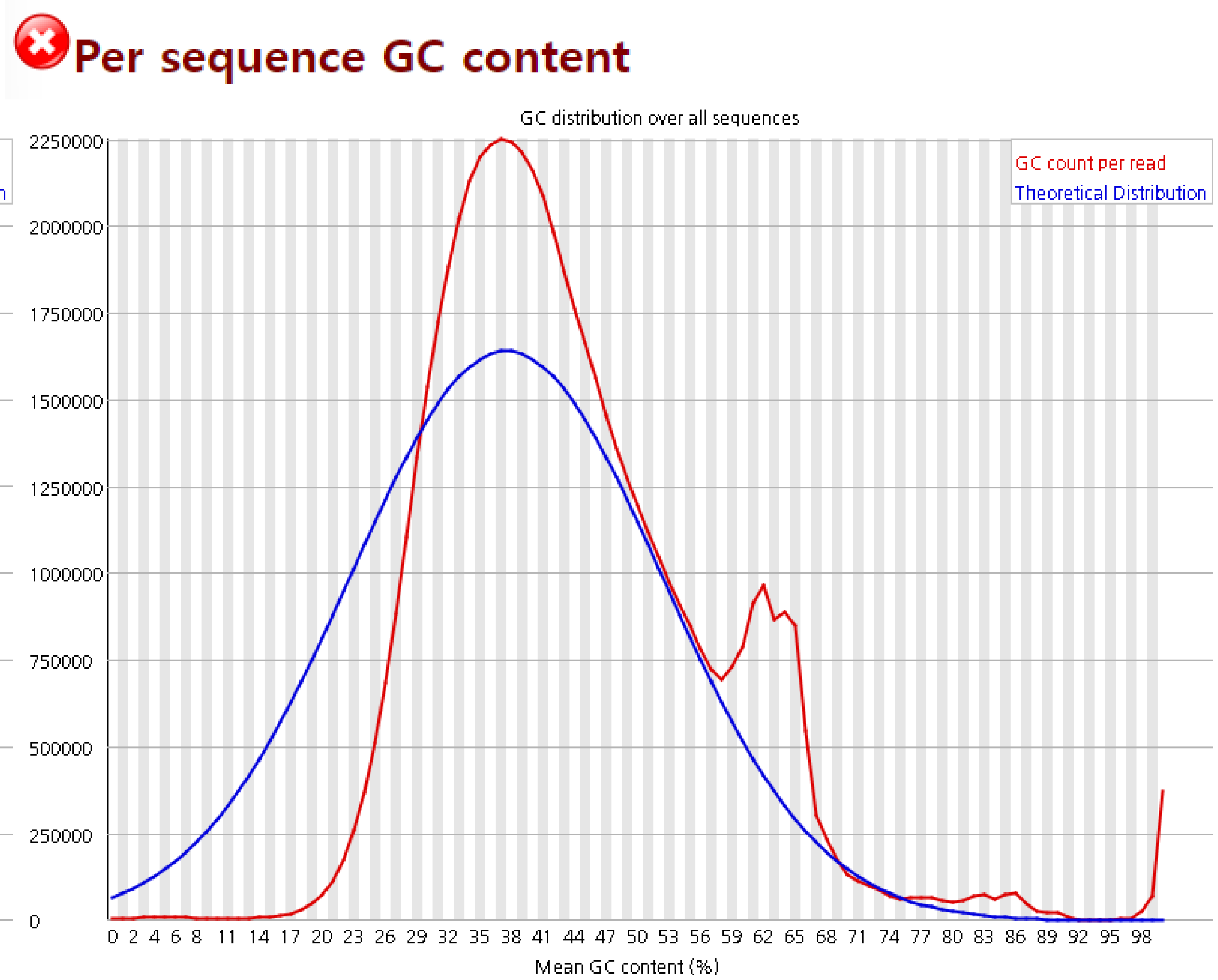
Post trimming

RNA sequencing – Quality Control

Pre sequence GC content(BID01_2)



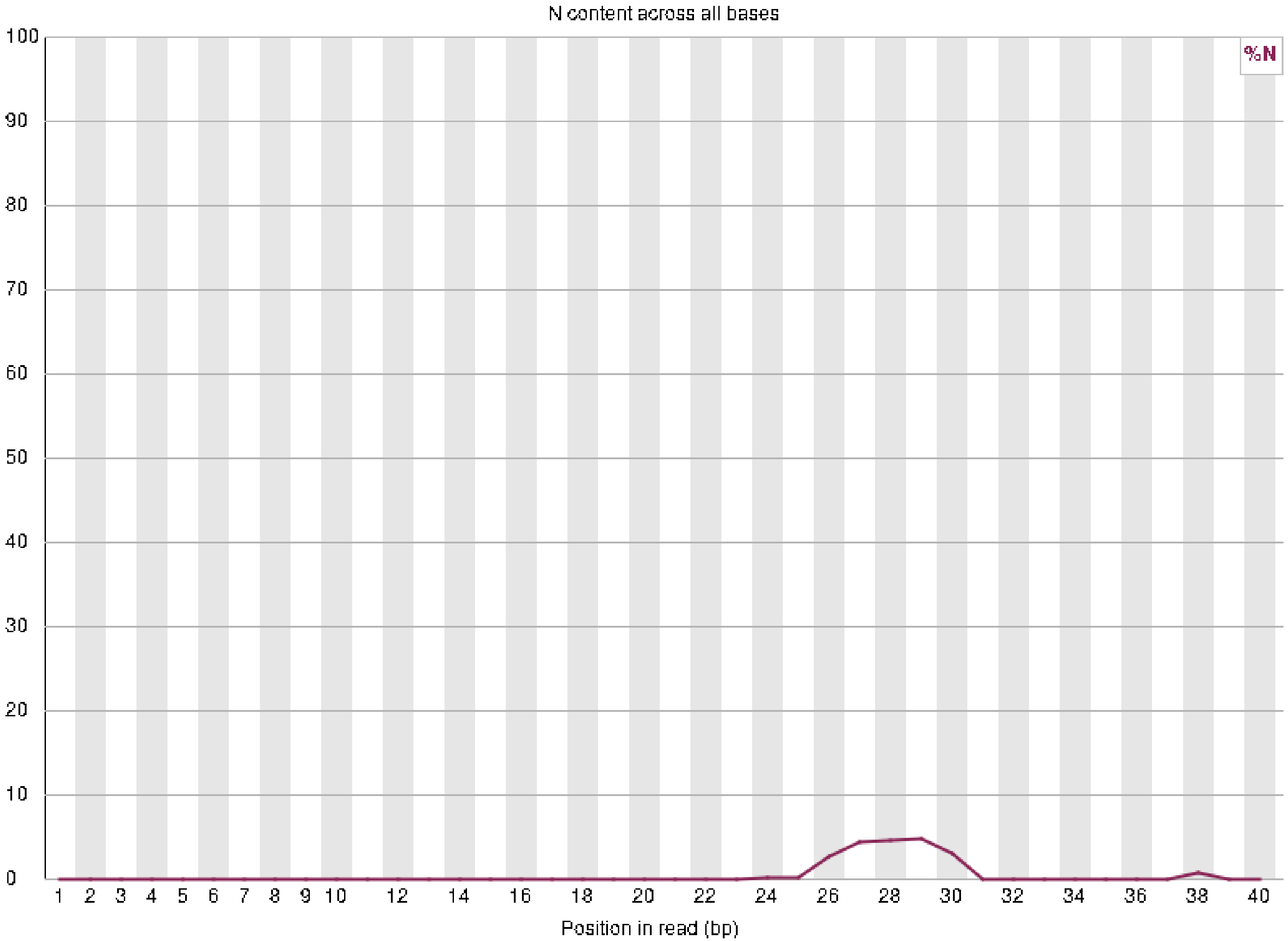
Pre trimming



Post trimming

RNA sequencing – Quality Control

Pre Base N Content



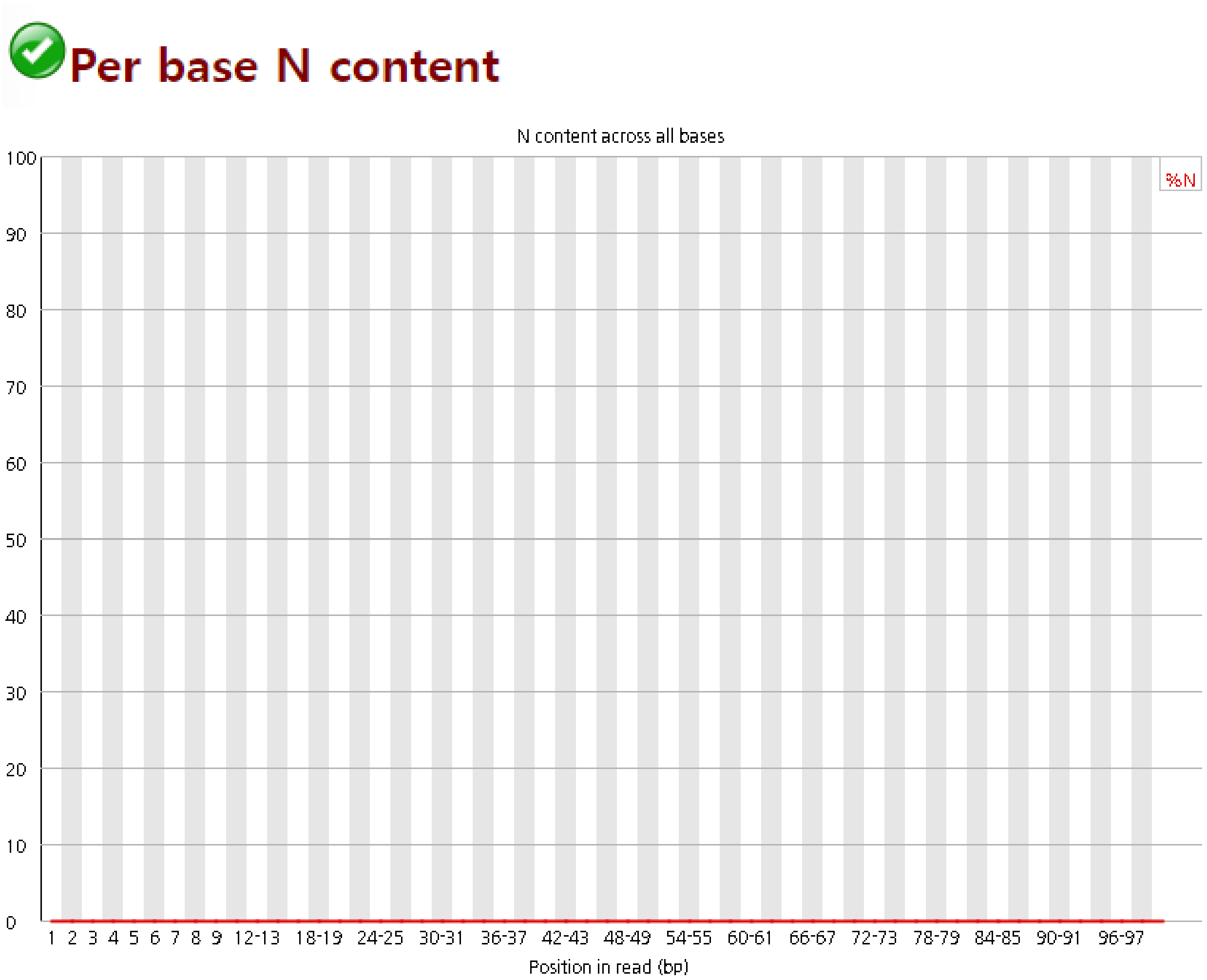
Bad Illumina Data

- 각 position의 base에서 N이 call된 Precentage를 나타낸다
- Sequencing 과정에서 염기서열을 정확히 알 수 없을 때 N을 call 한다
- N content가 높으면 read의 quality가 떨어져 mapping 과정에서 문제가 발생한다

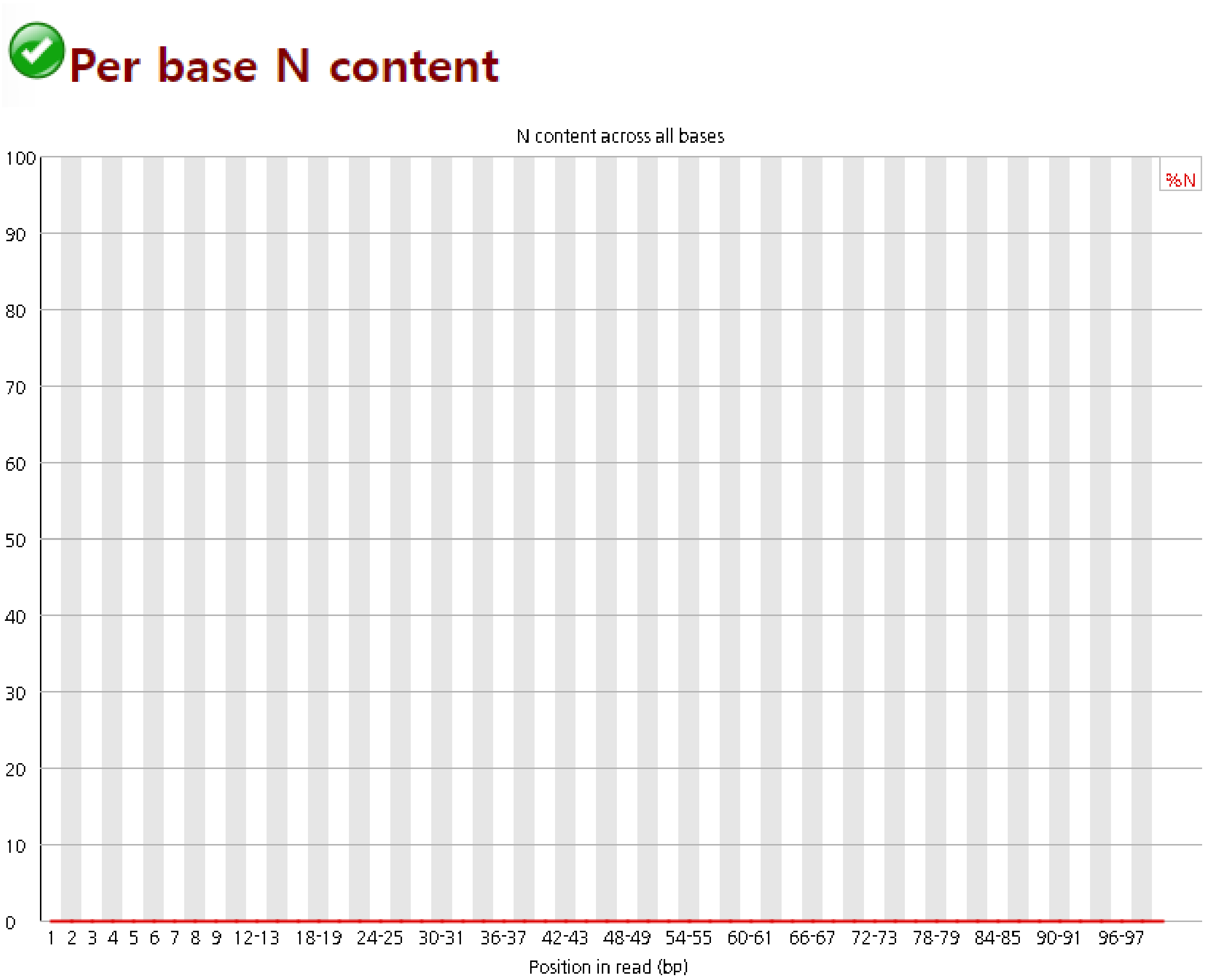
- X axis : 모든 read의 base position
- Y axis : Precentage of N content
- Red line : 해당 base position에 N content의 Precentage

RNA sequencing – Quality Control

Pre Base N Content(BID01_1)



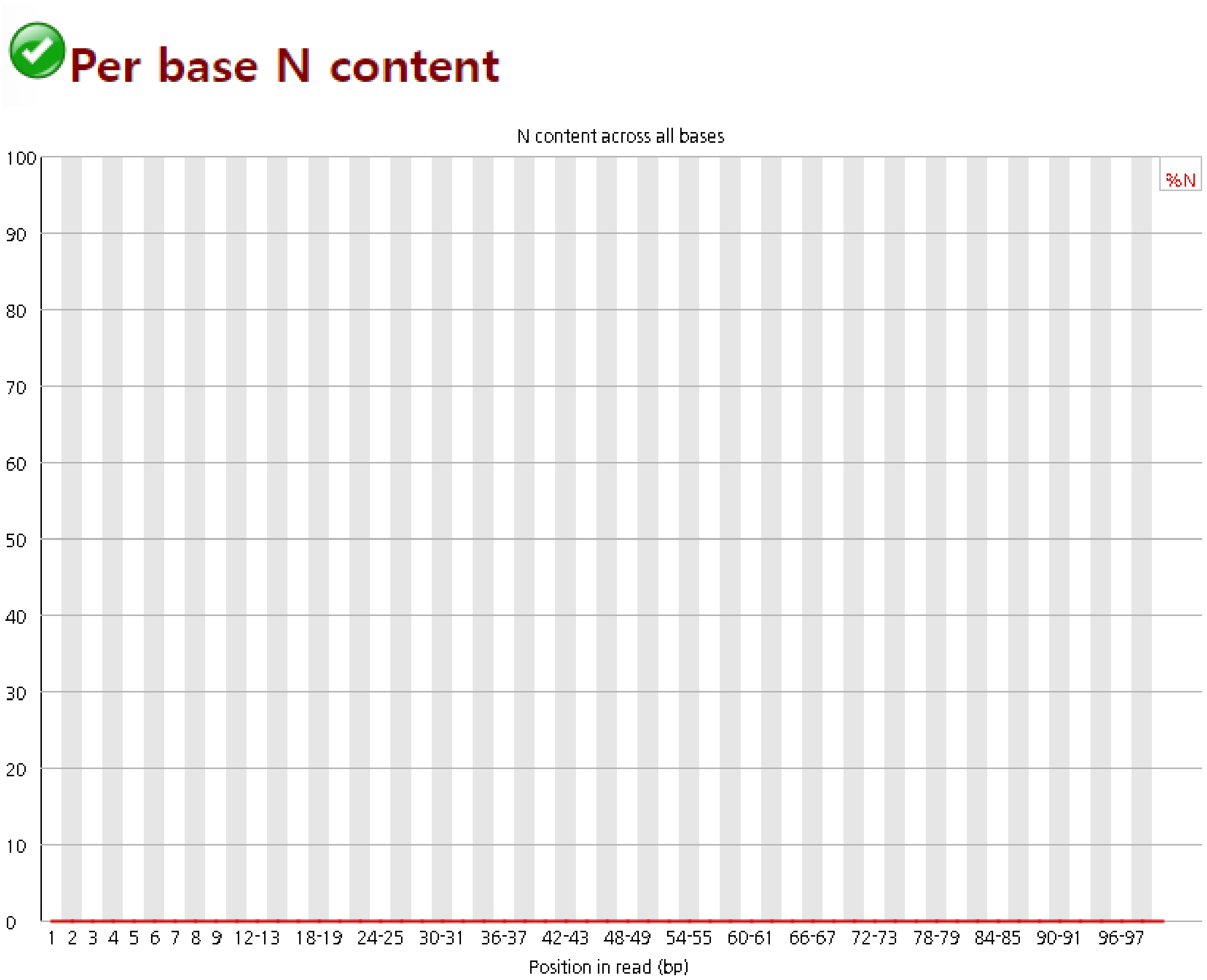
Pre trimming



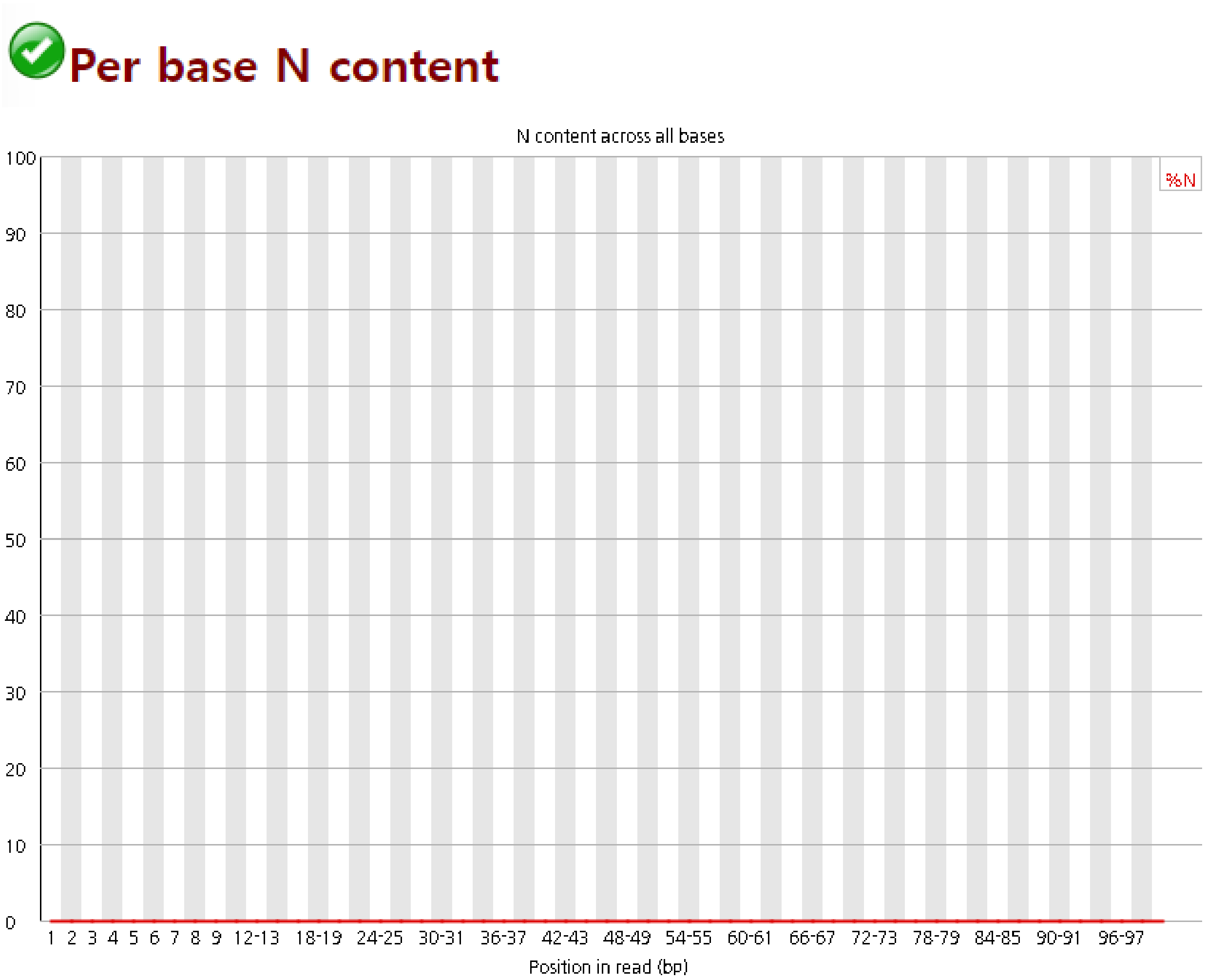
Post trimming

RNA sequencing – Quality Control

Pre Base N Content(BID01_2)



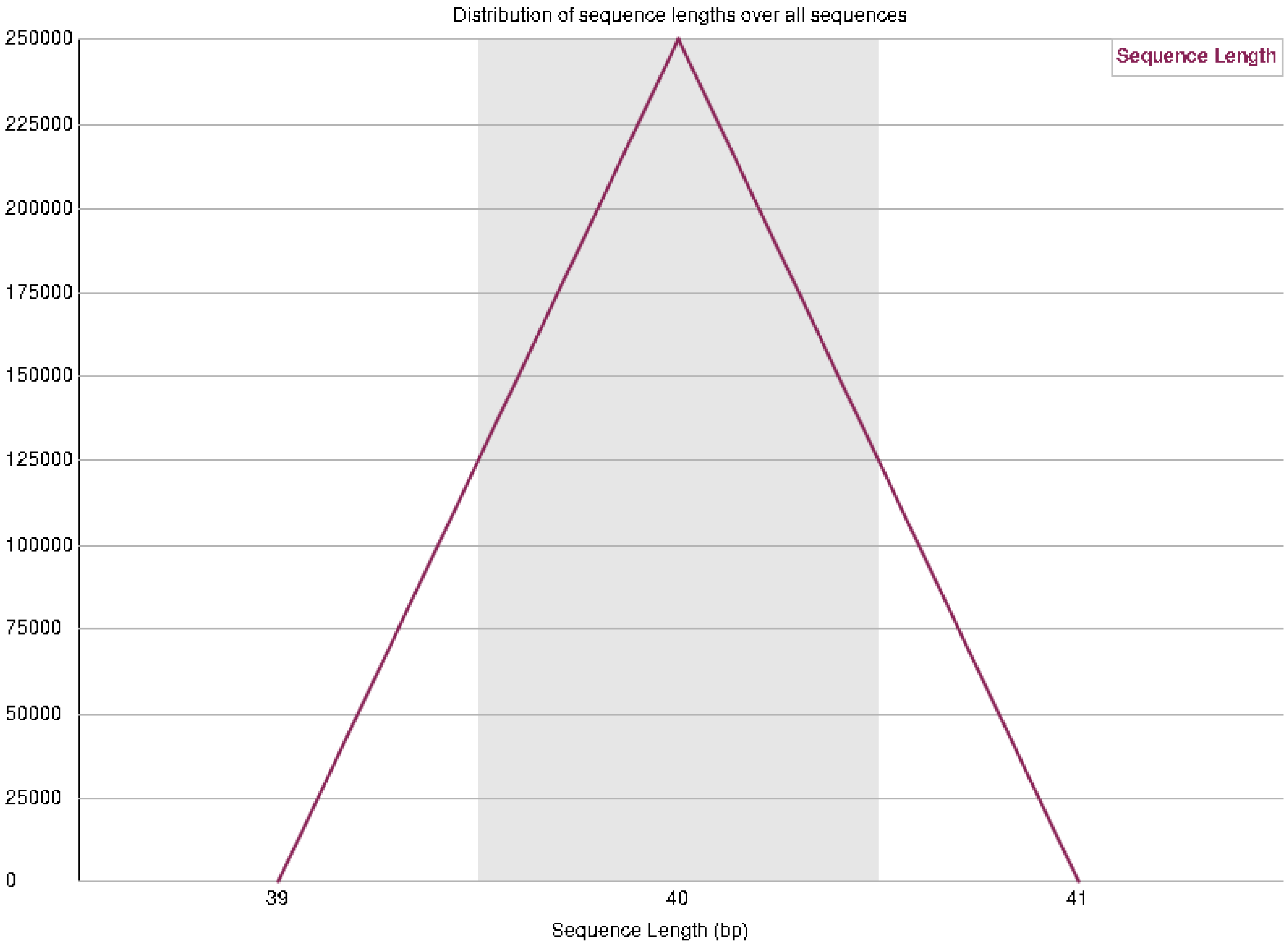
Pre trimming



Post trimming

RNA sequencing – Quality Control

Sequence Length Distribution



Good Illumina Data

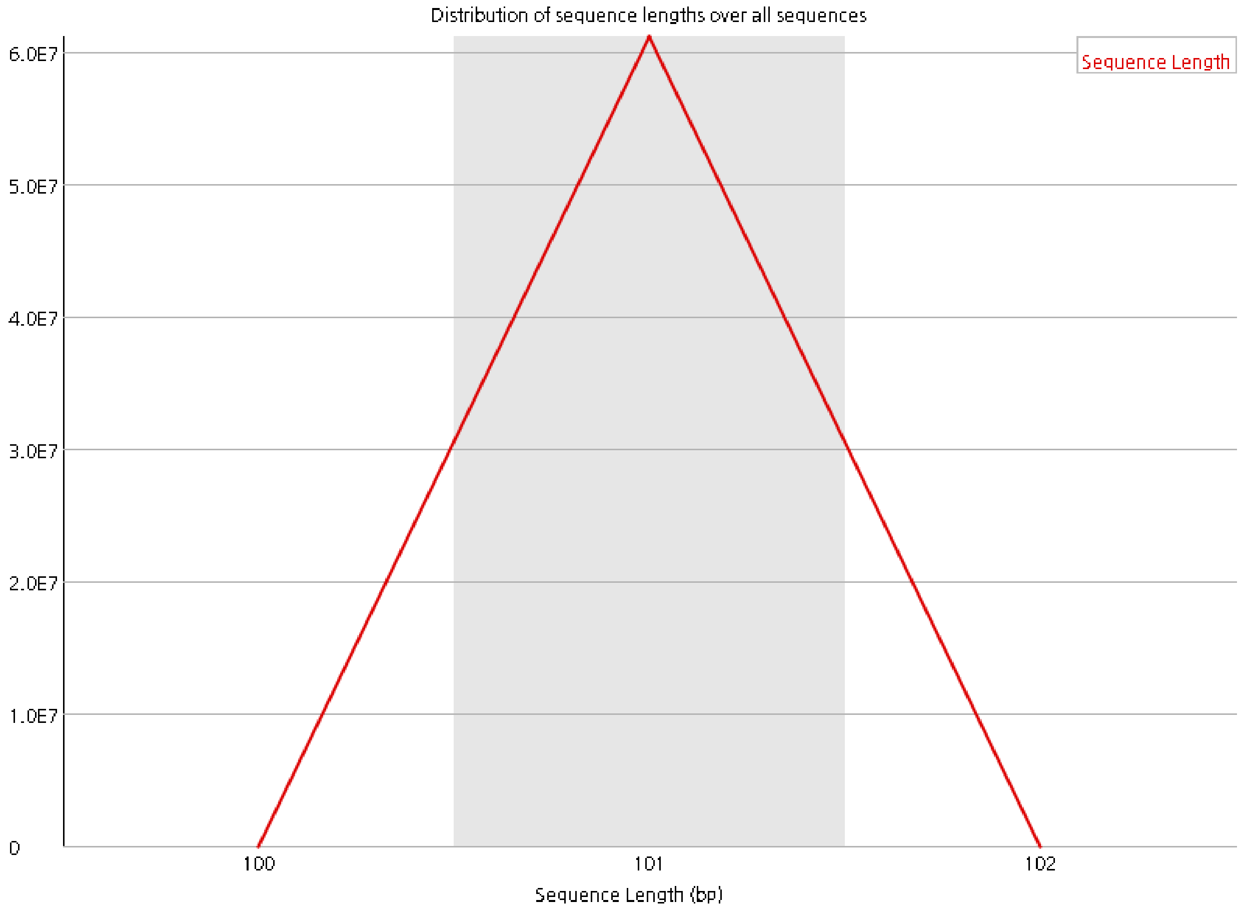
- Read들의 sequence length를 나타낸다.

- X axis : sequence length
- Y axis : Number of read
- Read line : 해당 sequence length의 read의 개수

RNA sequencing – Quality Control

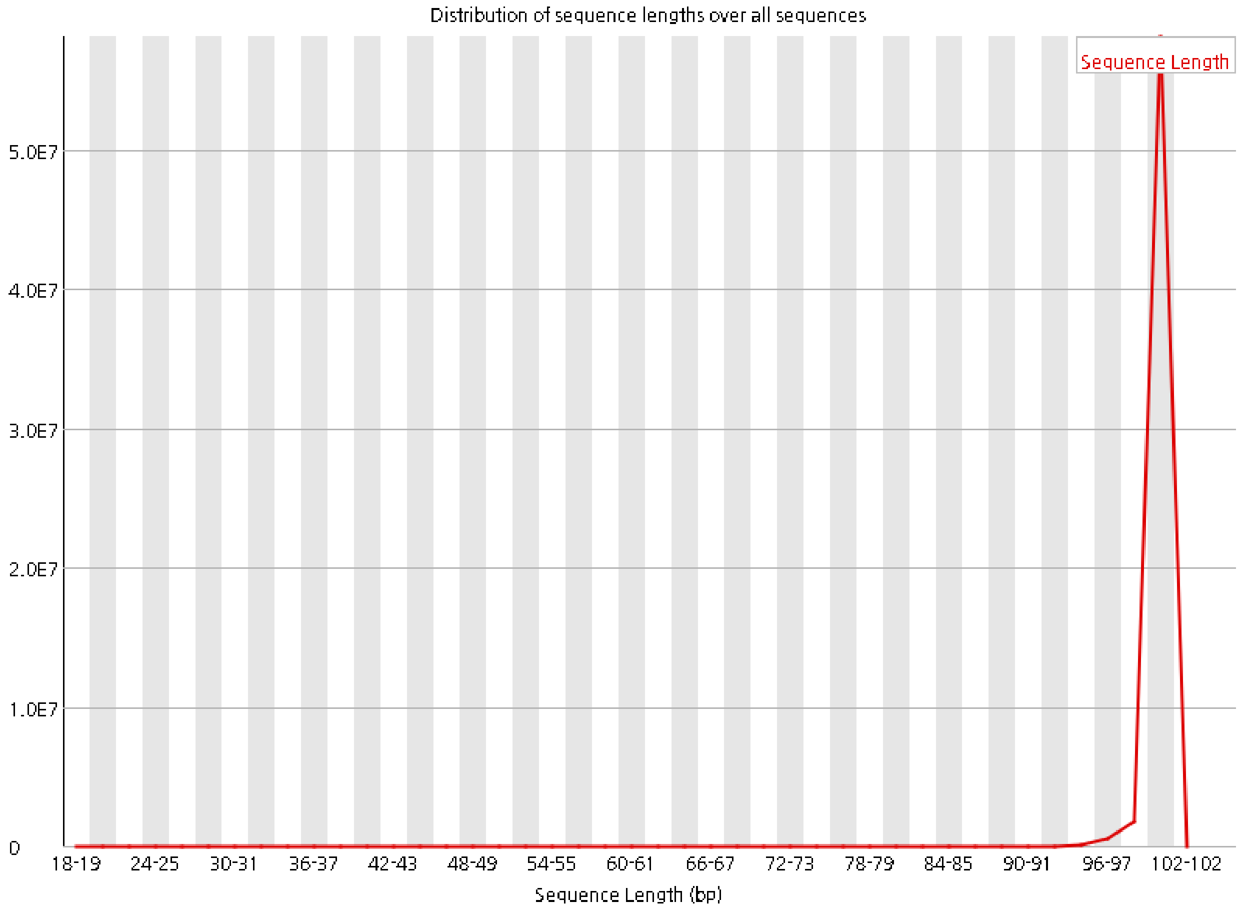
Sequence Length Distribution(BID01_1)

✓ **Sequence Length Distribution**



Pre trimming

! **Sequence Length Distribution**

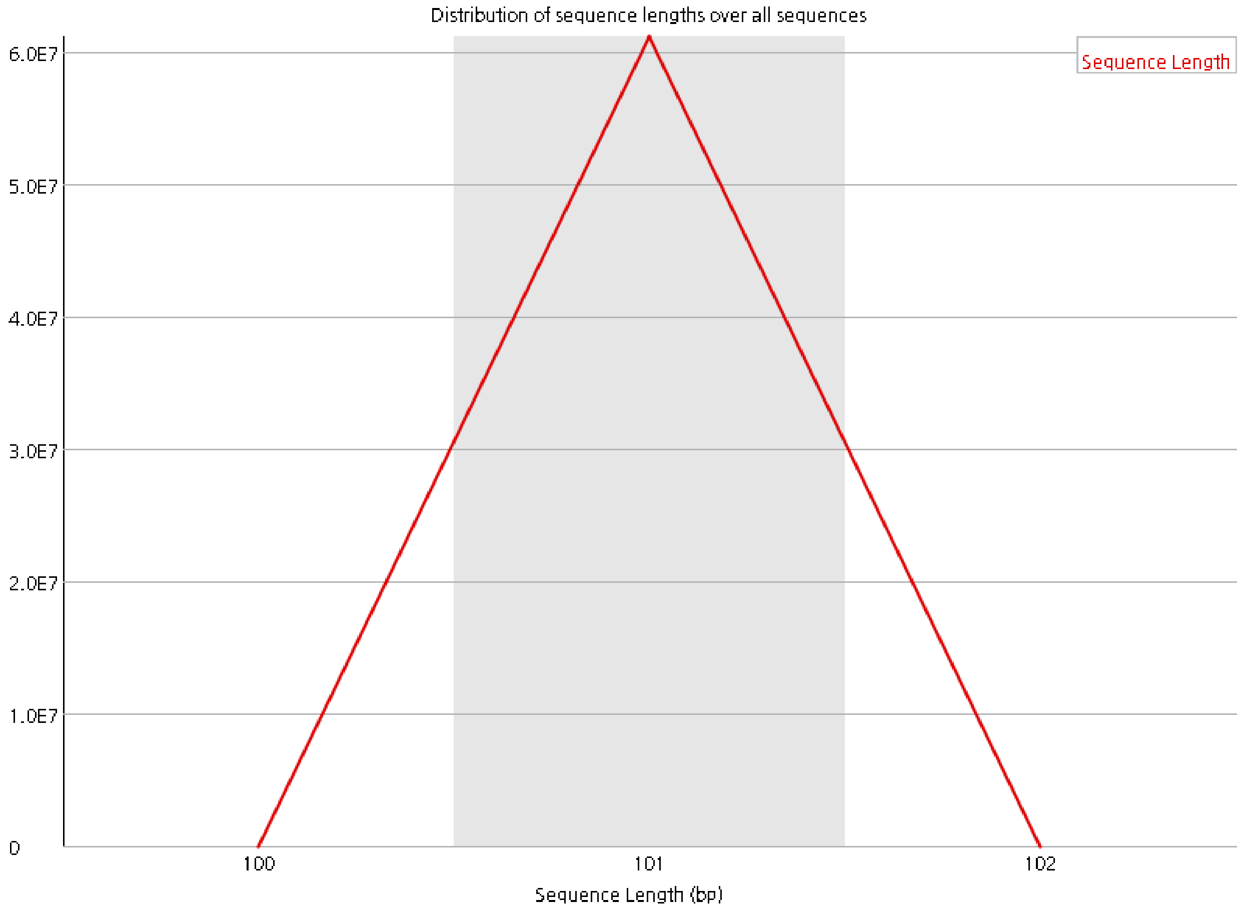


Post trimming

RNA sequencing – Quality Control

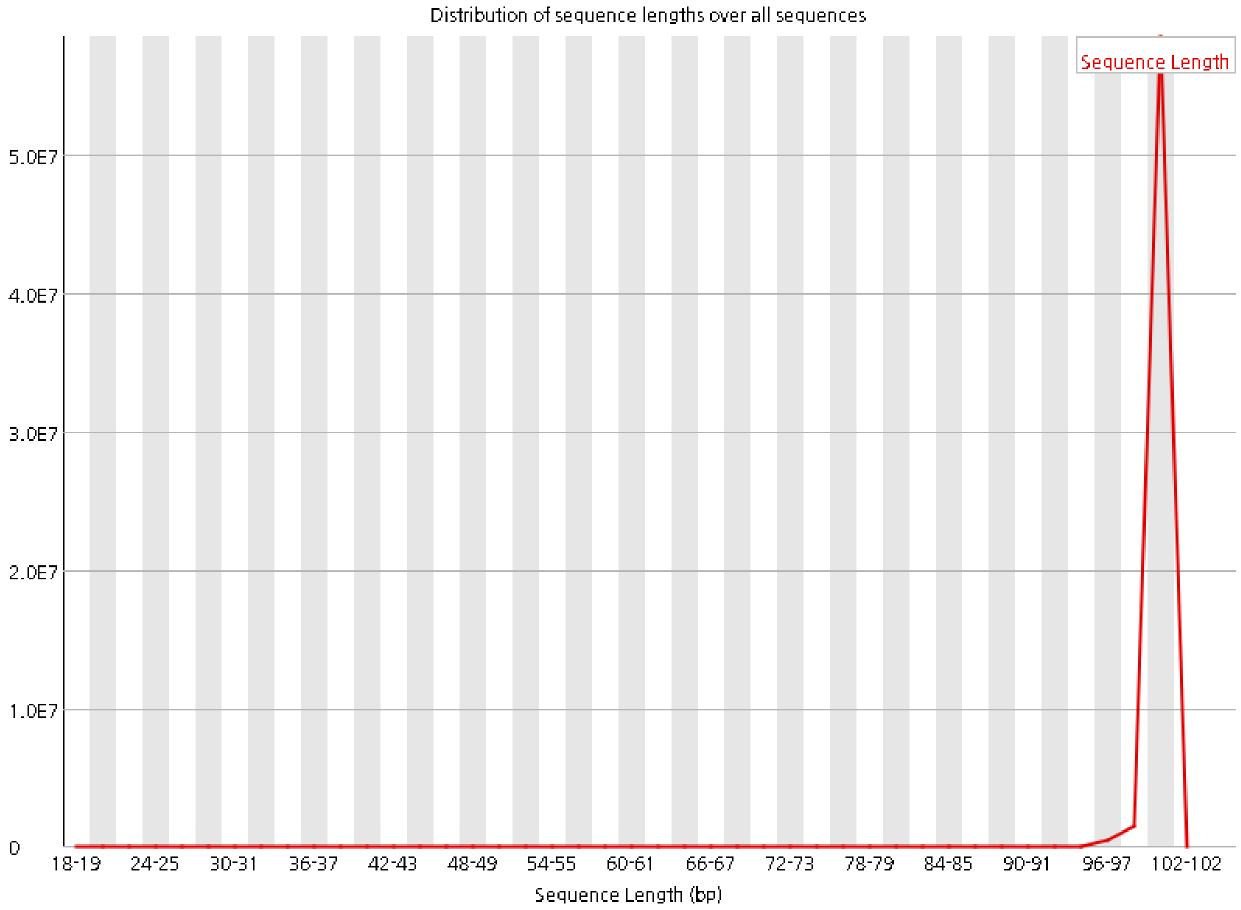
Sequence Length Distribution(BID01_2)

Sequence Length Distribution



Pre trimming

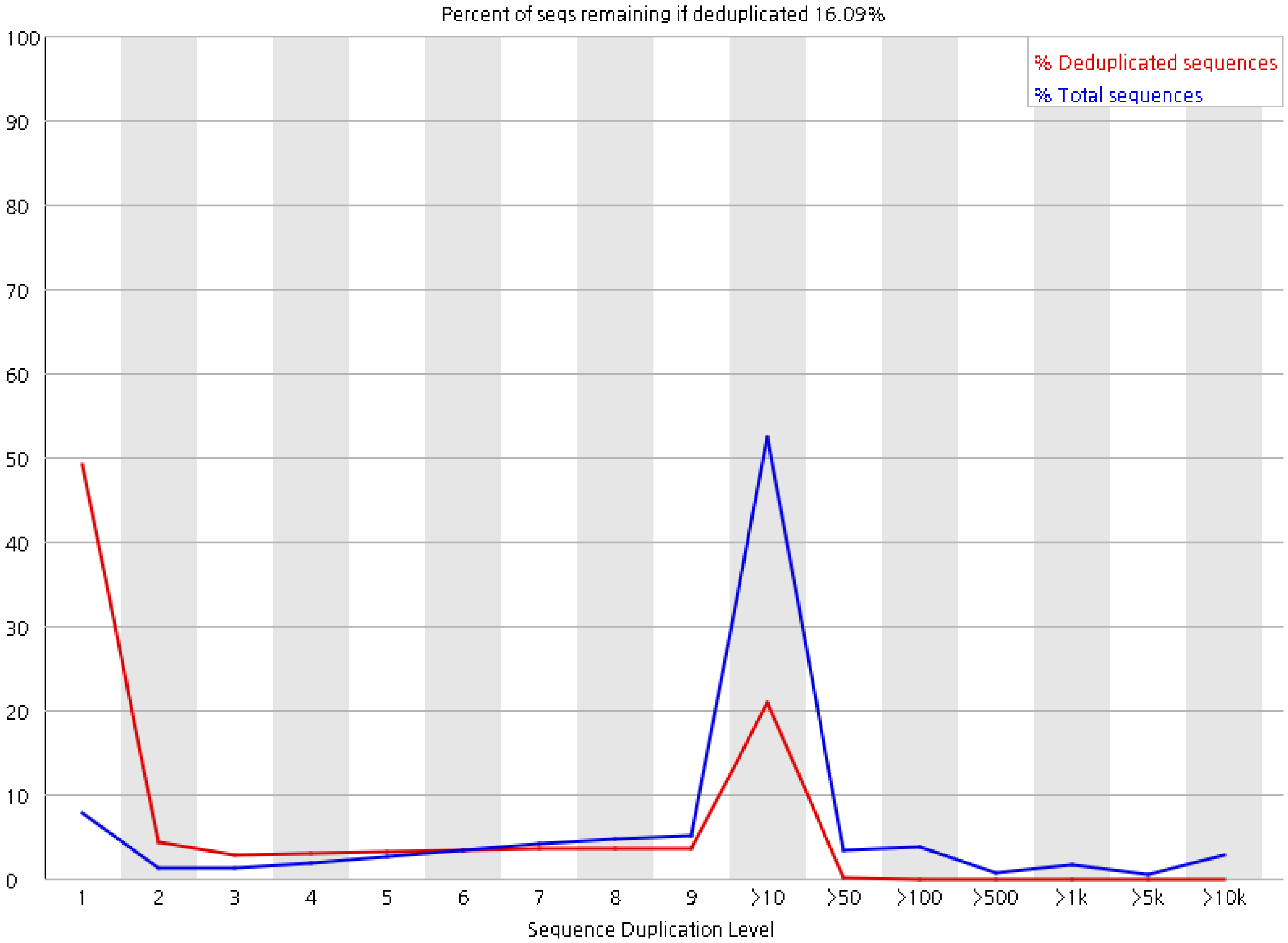
Sequence Length Distribution



Post trimming

RNA sequencing – Quality Control

Sequence Duplication Levels



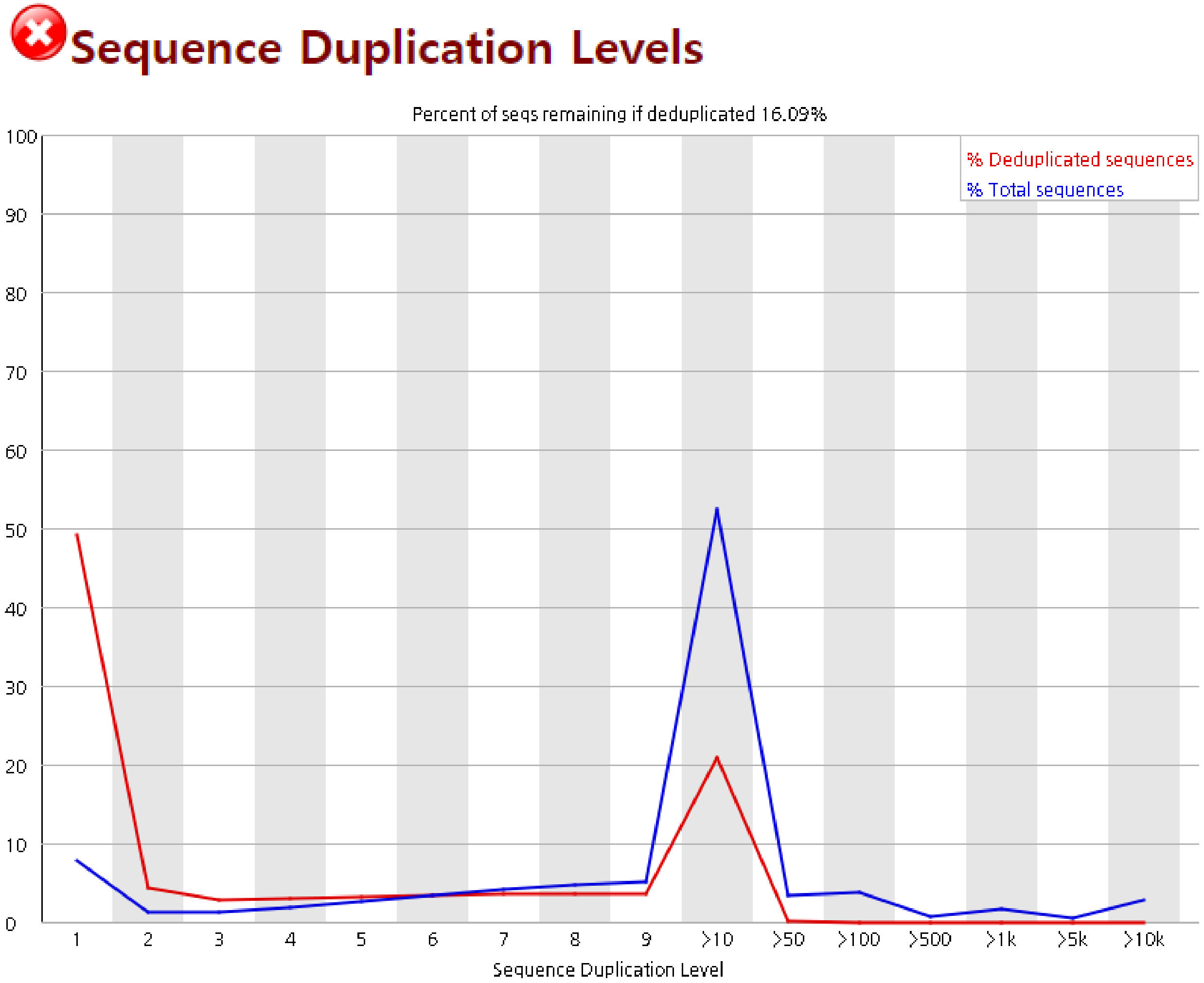
BID01_1 Pre trimming

- Duplicated sequence를 가진 read들의 수를 나타낸다.

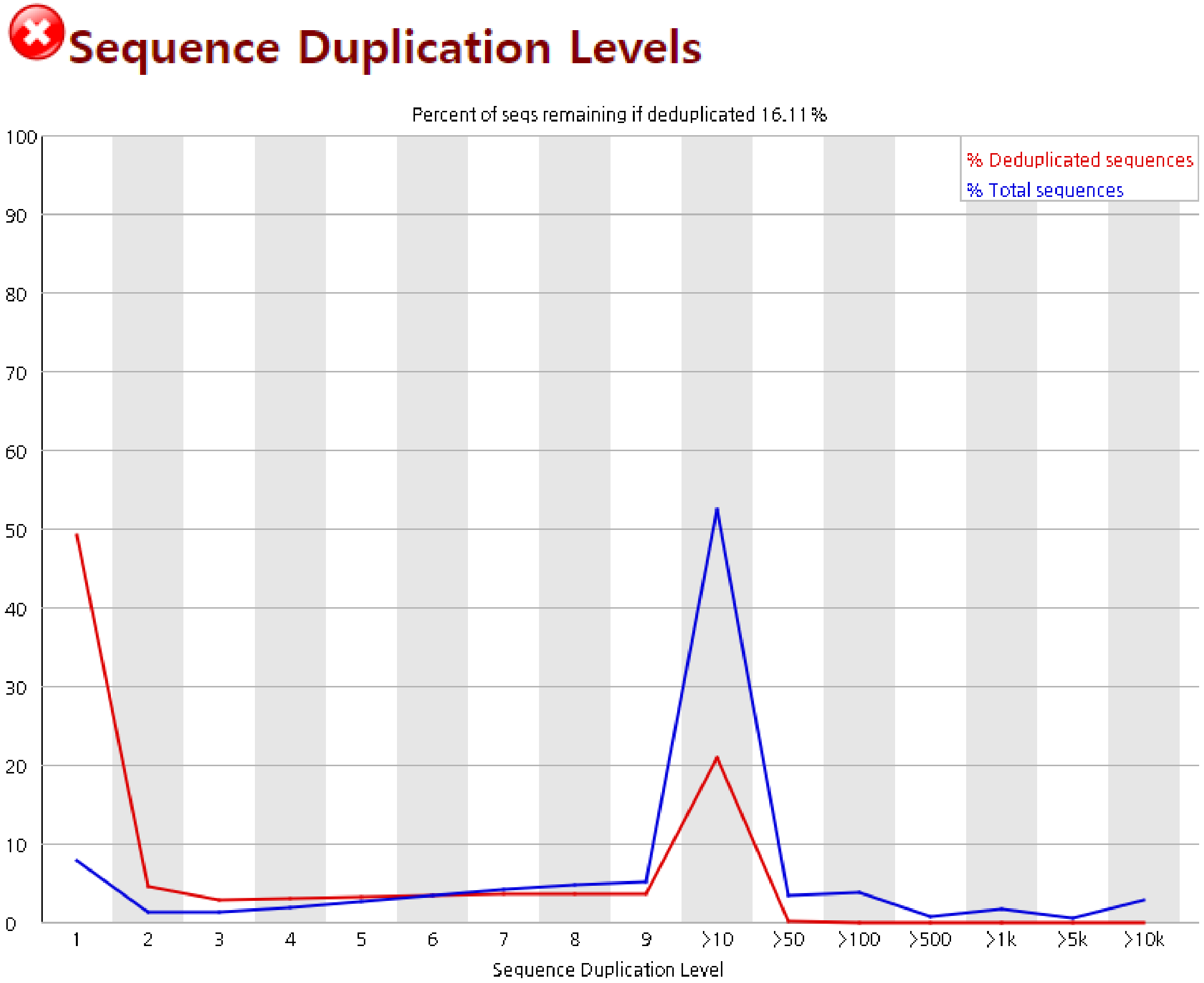
- X axis : sequence duplication level
- Y axis : Precentage
- Blue line : 해당 sequence duplication level에서 duplication된 Precentage
- Read line : 해당 sequence duplication level에서 duplication를 제거한 후 duplication Precentage

RNA sequencing – Quality Control

Sequence Duplication Levels(BID01_1)



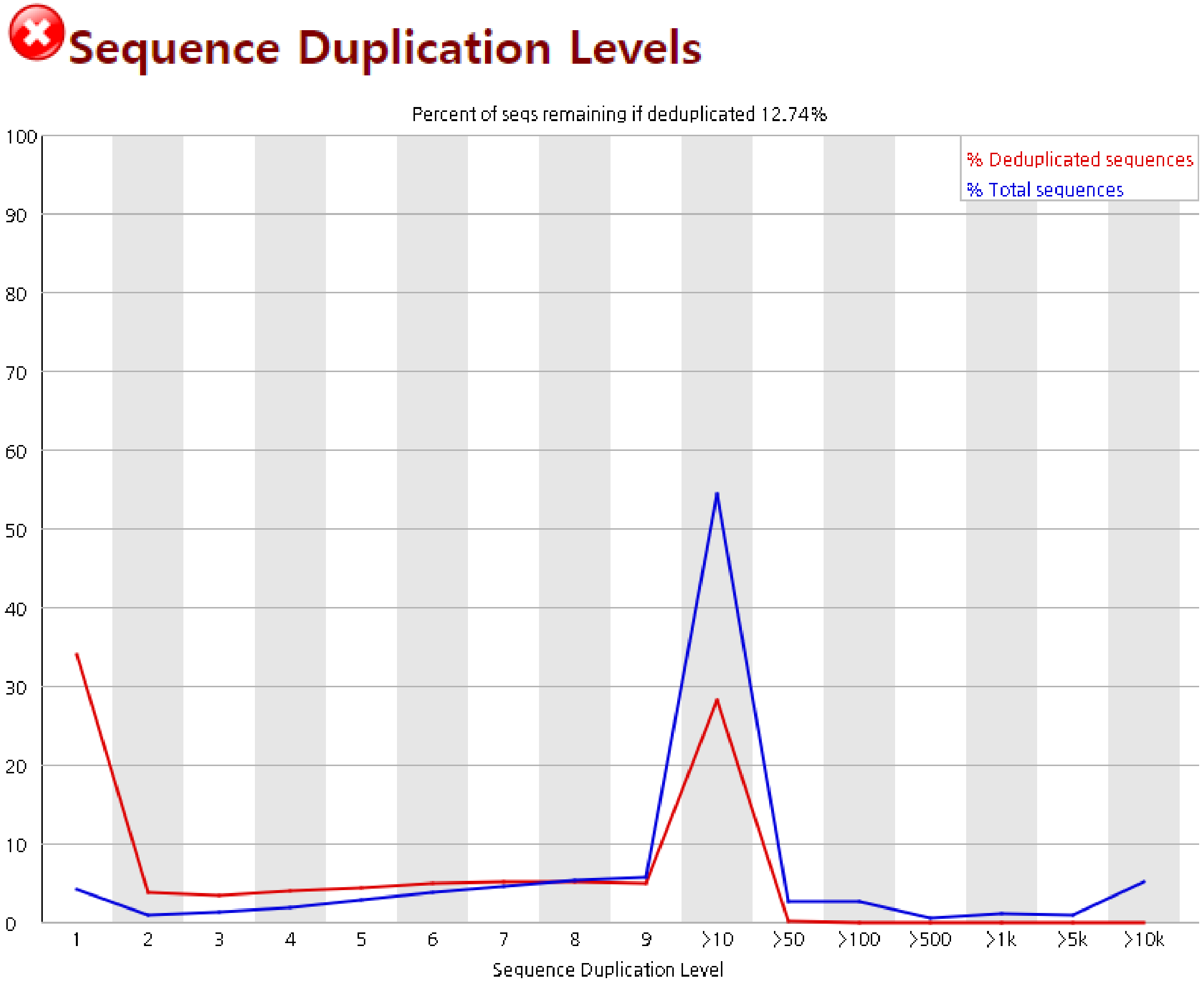
Pre trimming



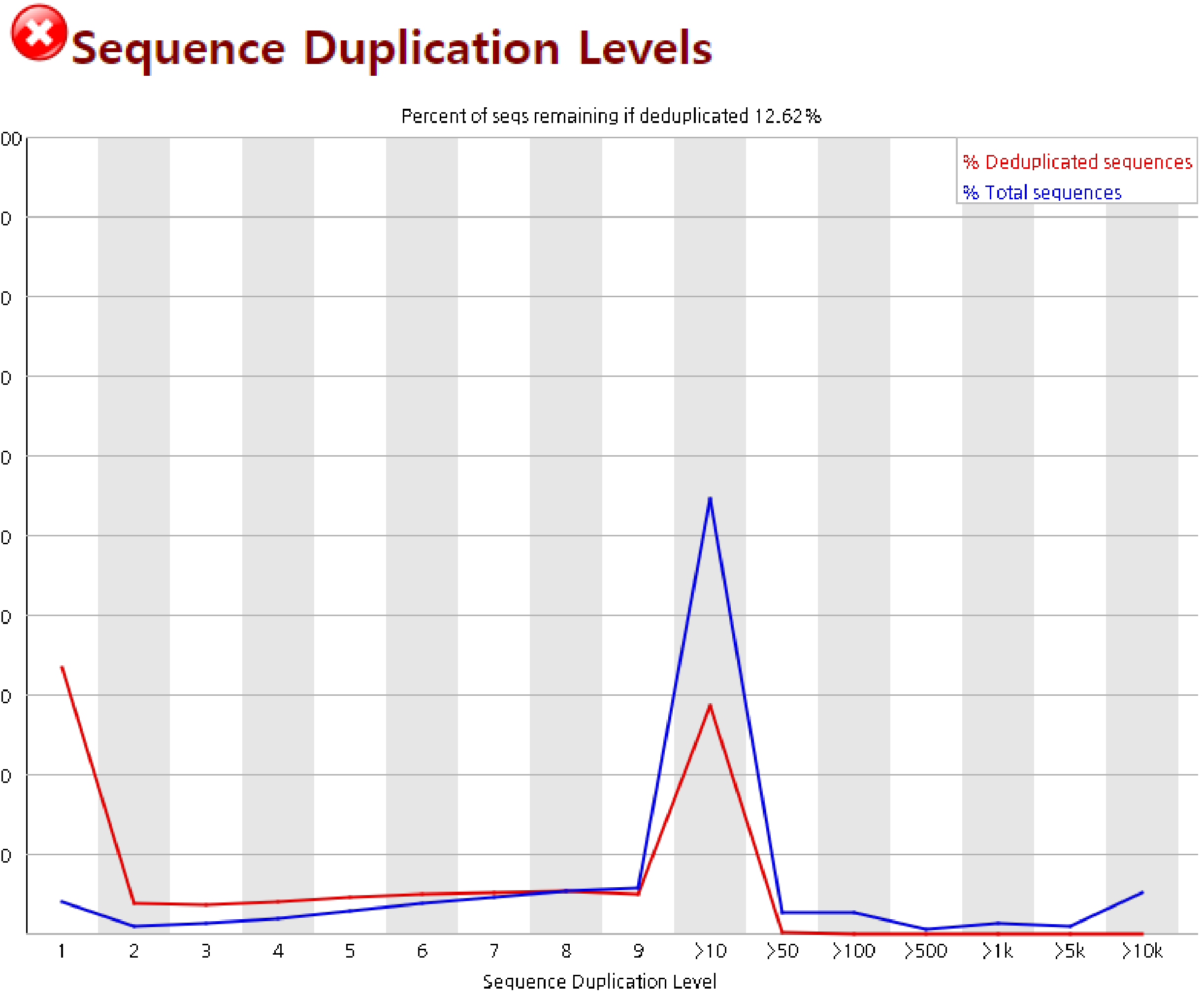
Post trimming

RNA sequencing – Quality Control

Sequence Duplication Levels(BID01_2)



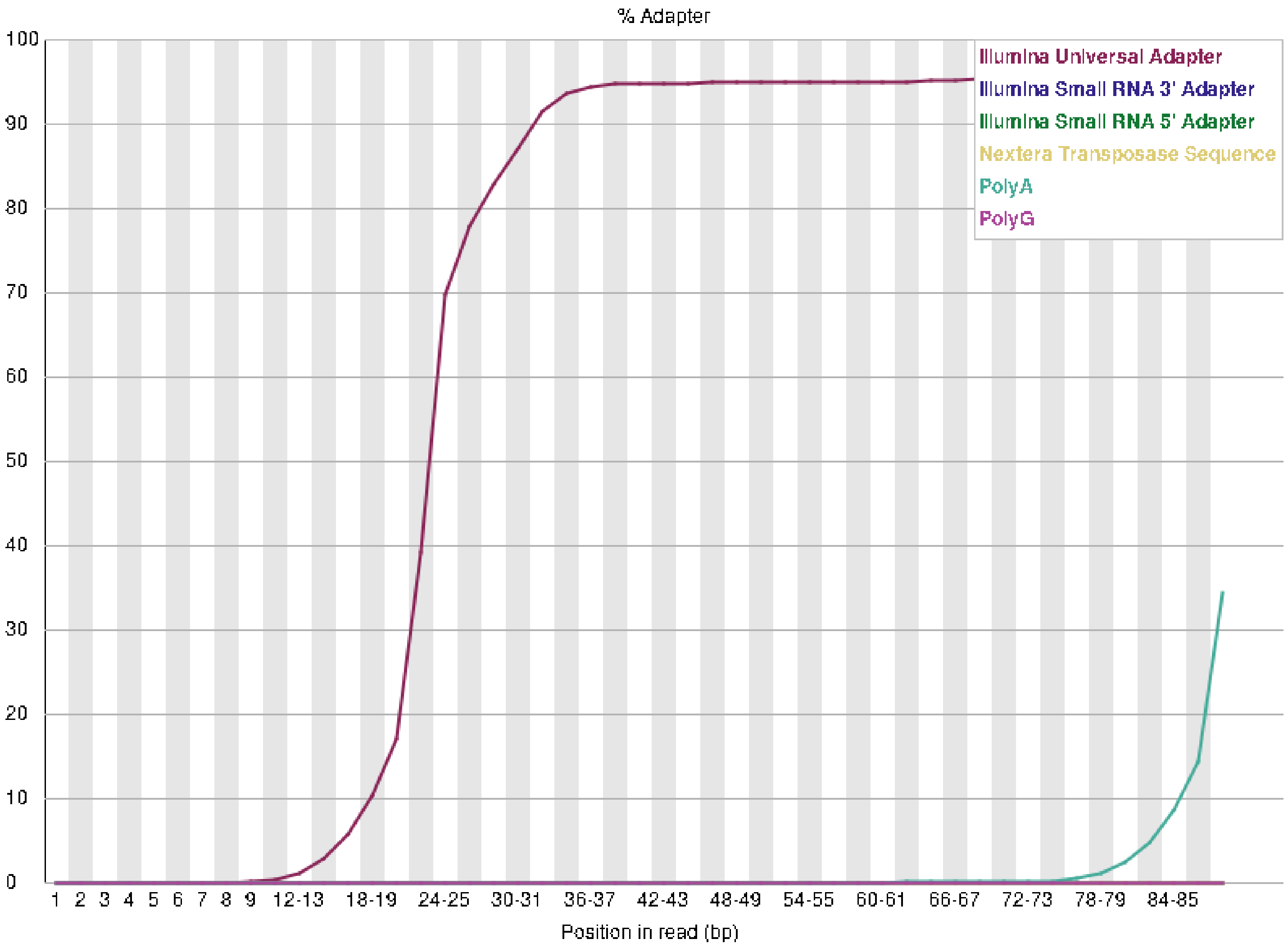
Pre trimming



Post trimming

RNA sequencing – Quality Control

Adapter Content



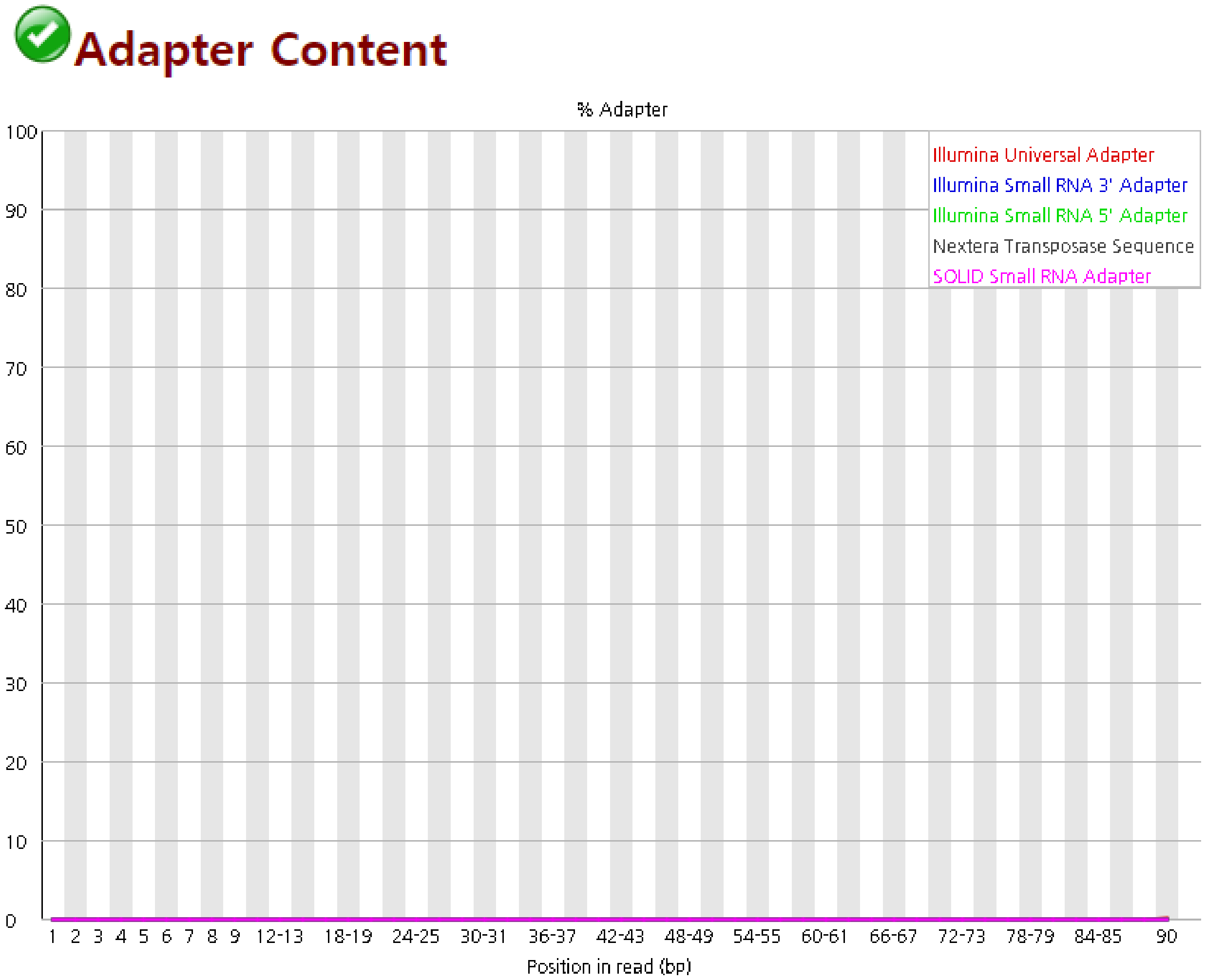
Small RNA with read-through adapter

- 각 read의 position 마다 adapter sequence가 얼마나 포함되어 있는지 Precentage로 나타낸다.

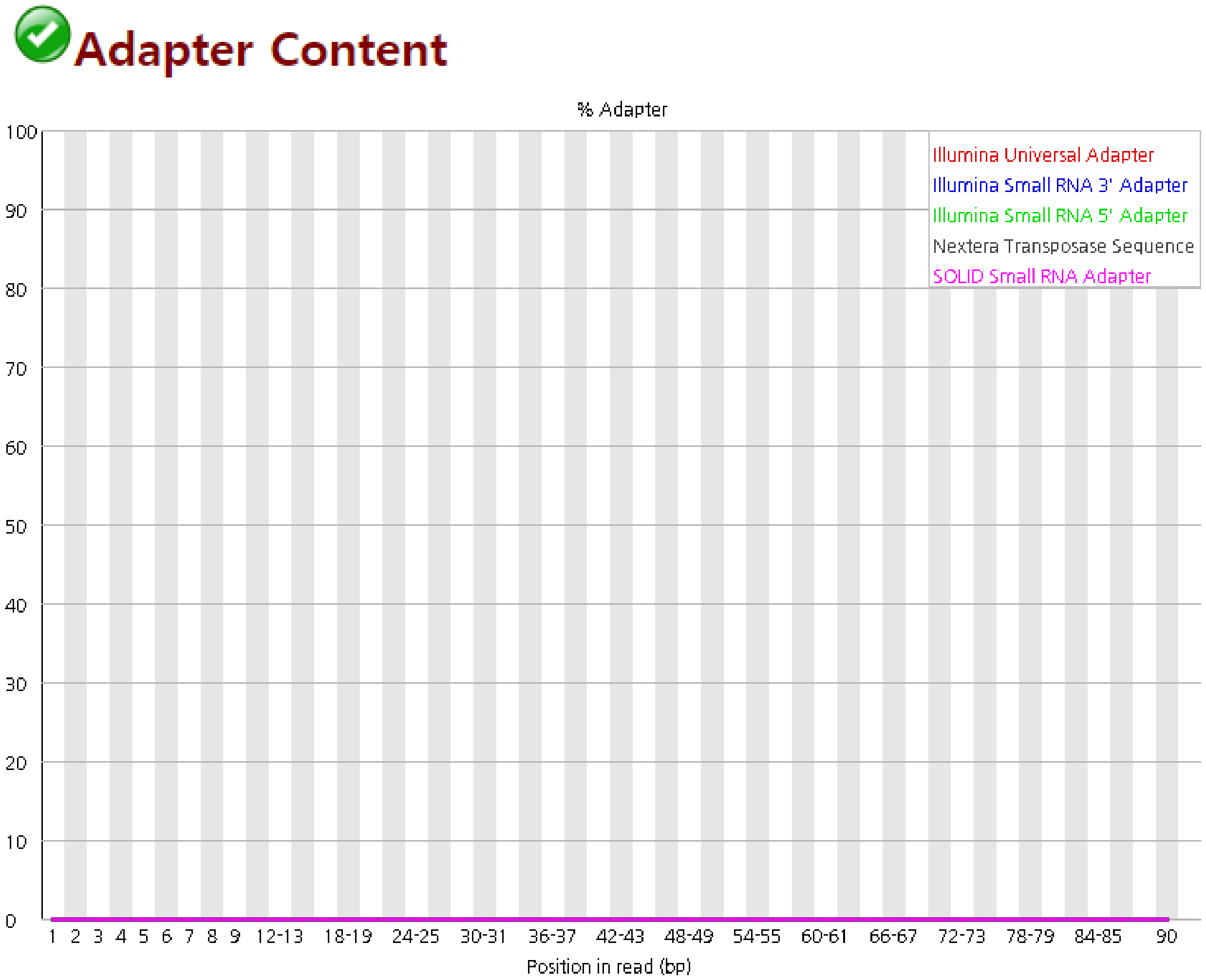
- X axis : 모든 read의 base position
- Y axis : adapter Precentage
- Color line : read에 존재하는 다른 종류의 adapter와 poly A, poly G를 의미

RNA sequencing – Quality Control

Adapter Content(BID01_1)



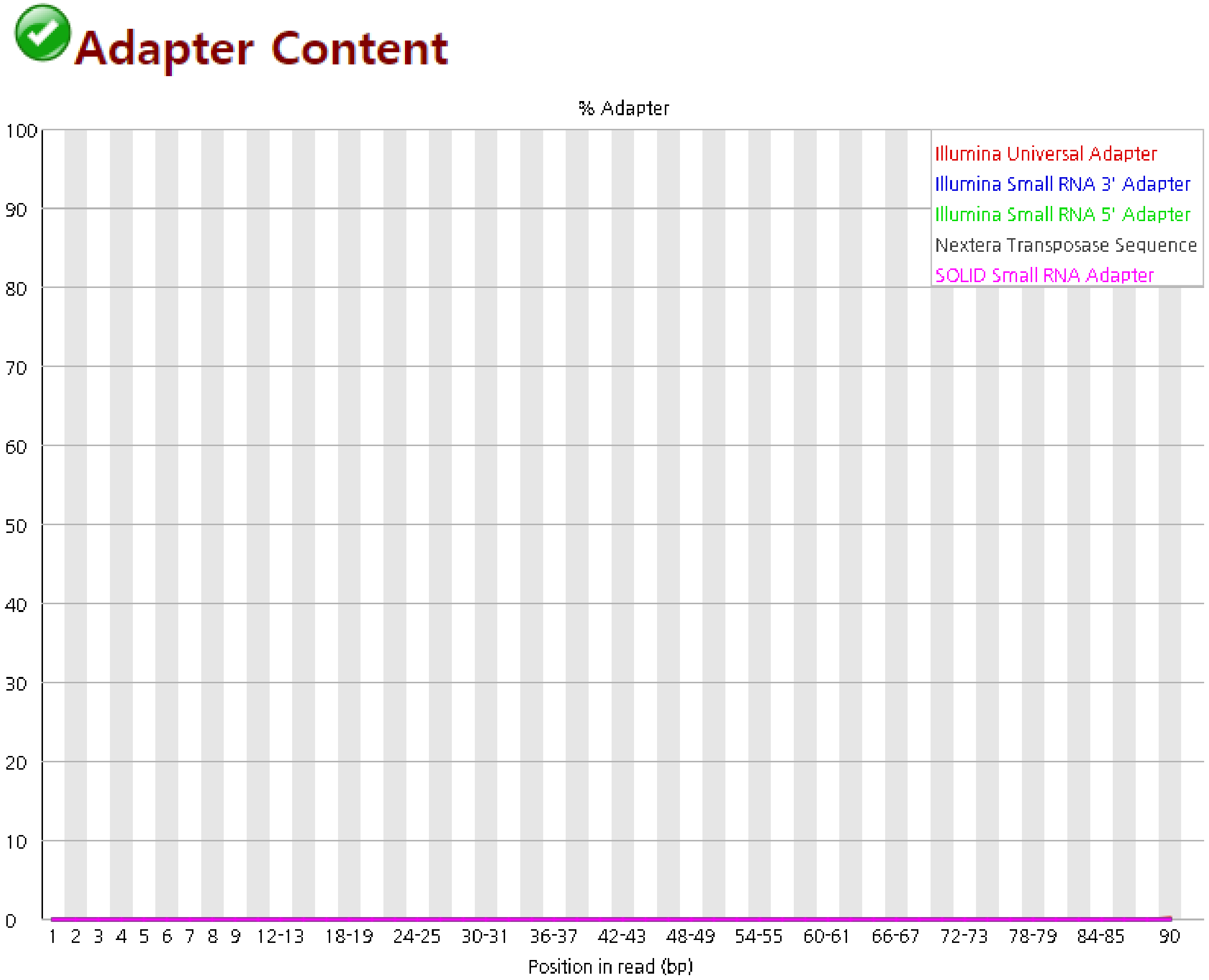
Pre trimming



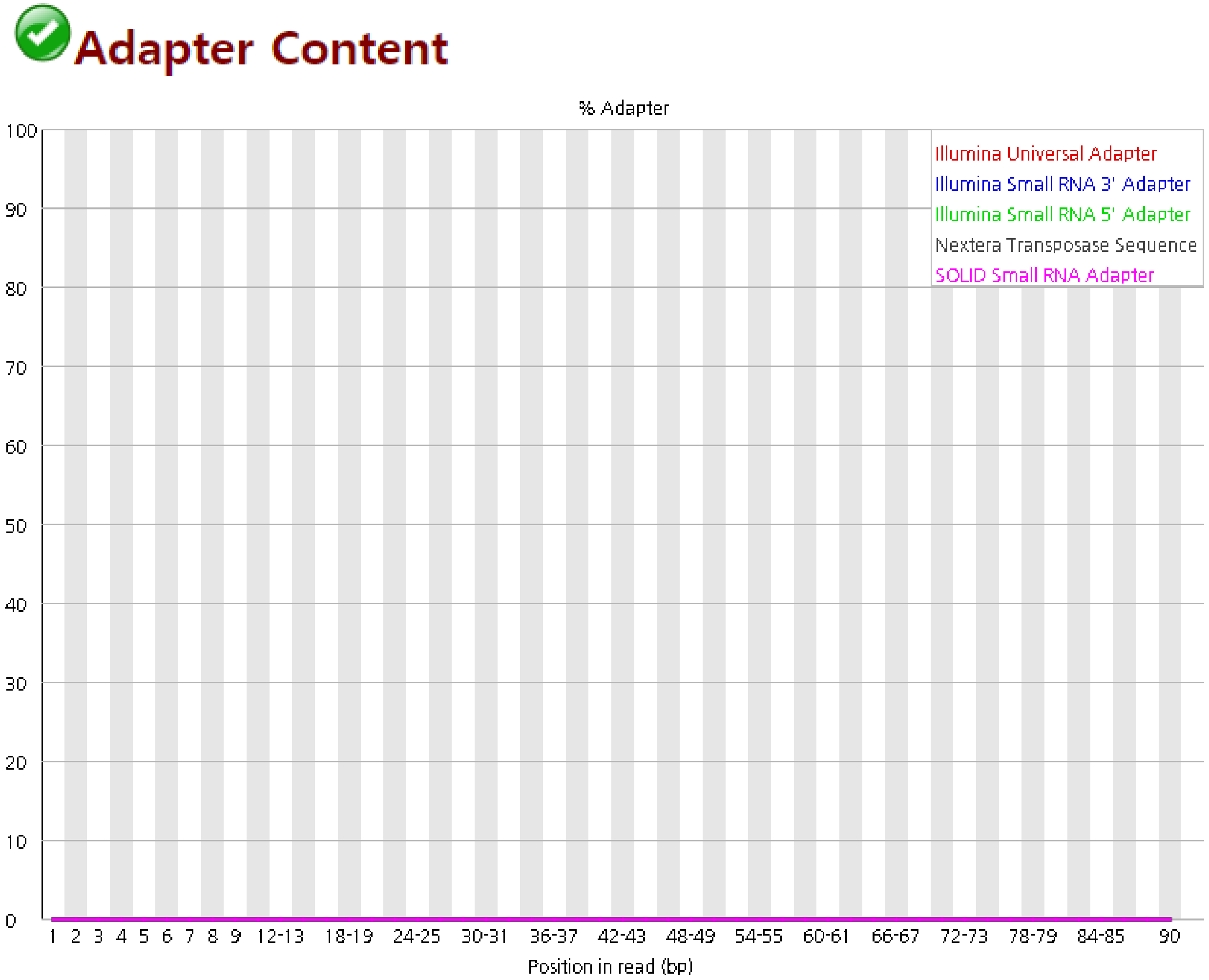
Post trimming

RNA sequencing – Quality Control

Adapter Content(BID01_2)



Pre trimming

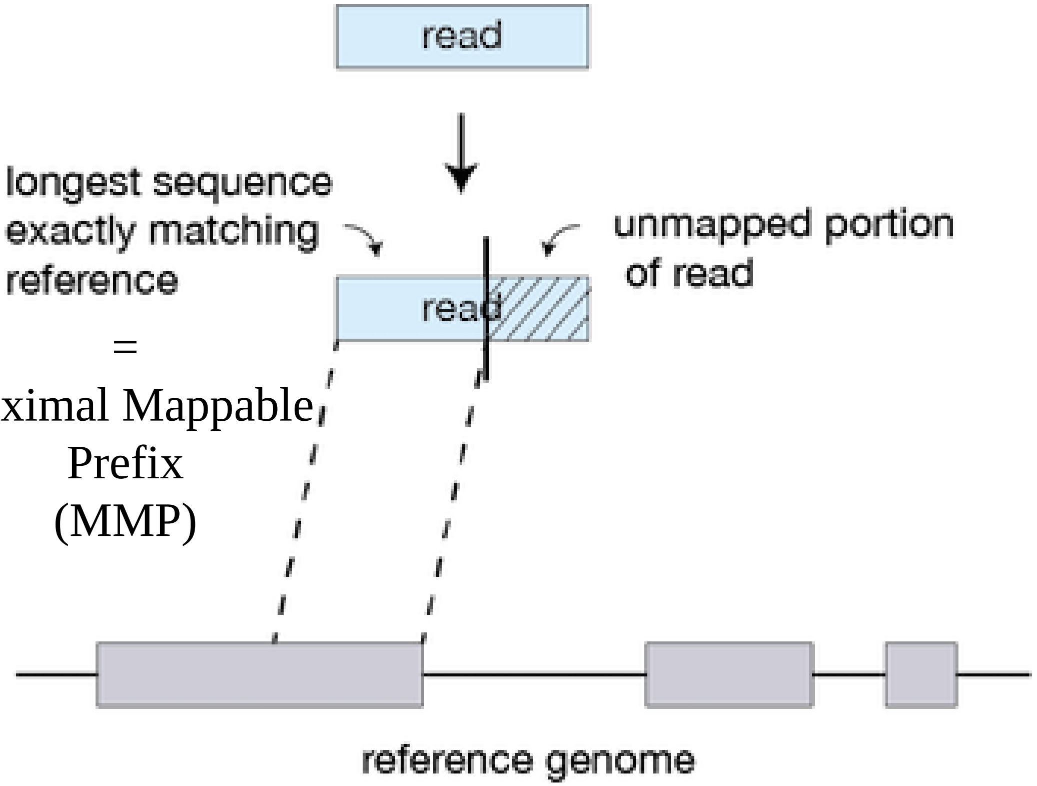


Post trimming

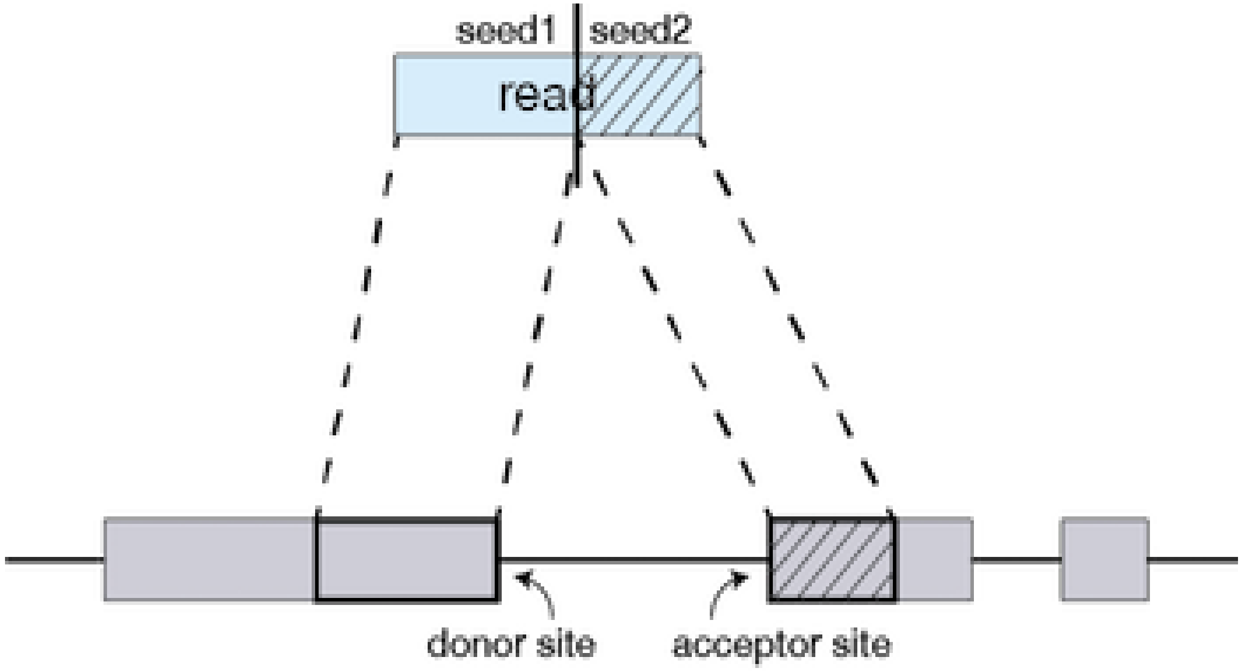
RNA sequencing – Alignment

- Alignment of the large sets of sequenced reads to a reference genome.
- Using STAR tool

➢ Seed search 과정과 clustering, stitching and scoring 과정으로 mapping 진행

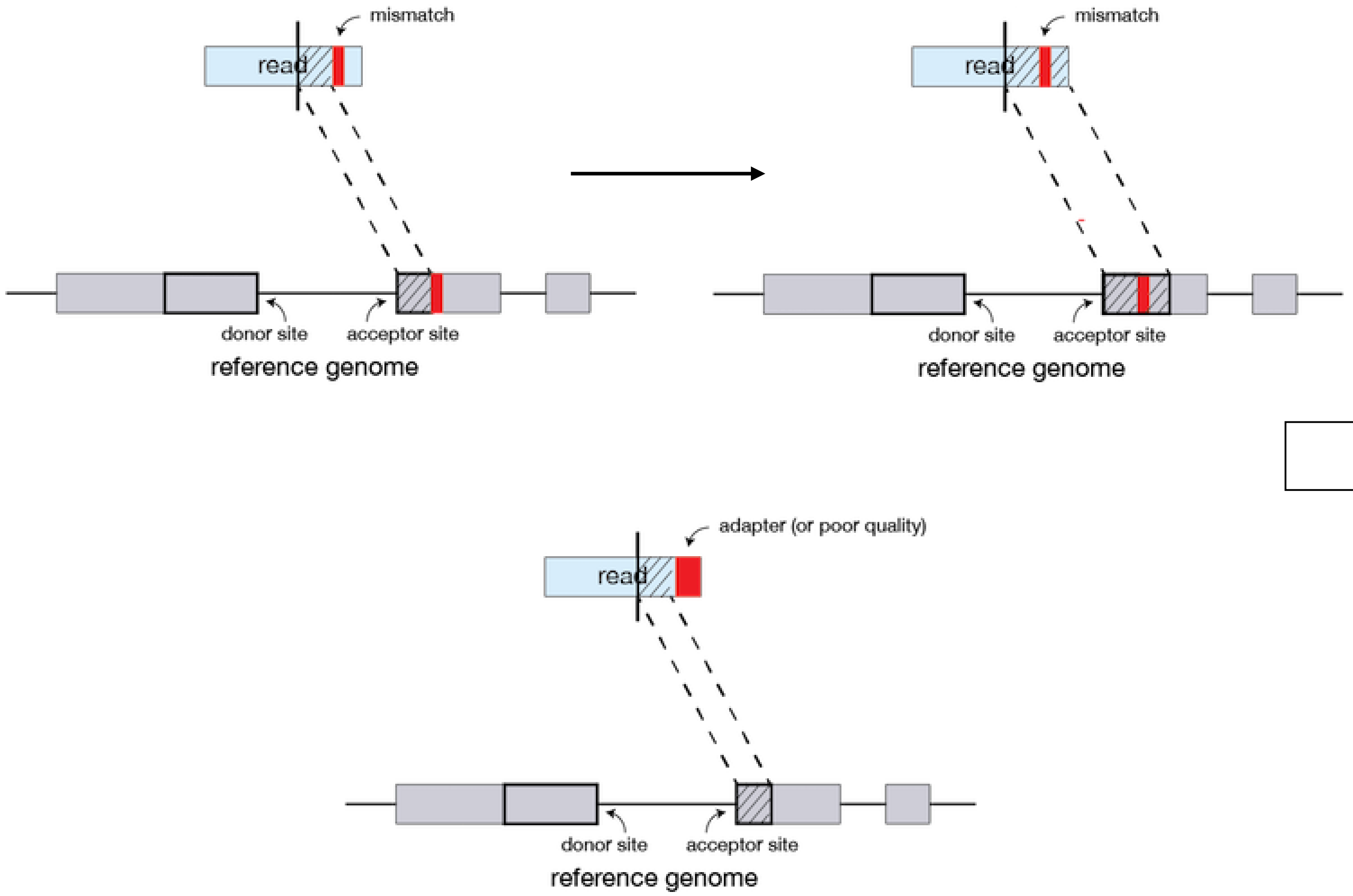


=
Maximal Mappable Prefix
(MMP)



https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html,2023-06-07

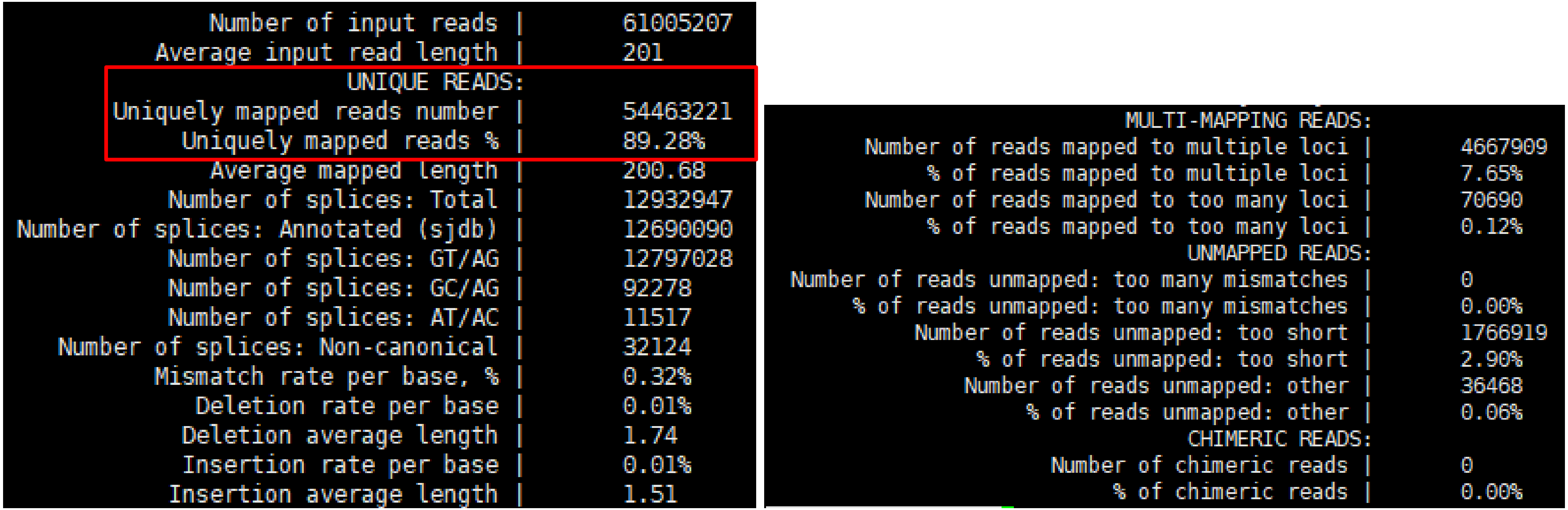
RNA sequencing – Alignment



Clustering, stitching and scoring : Alignments of the entire read sequence by stitching together all the seeds that were aligned to the genome in the first phase.

https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html,2023-06-07

RNA sequencing – Alignment

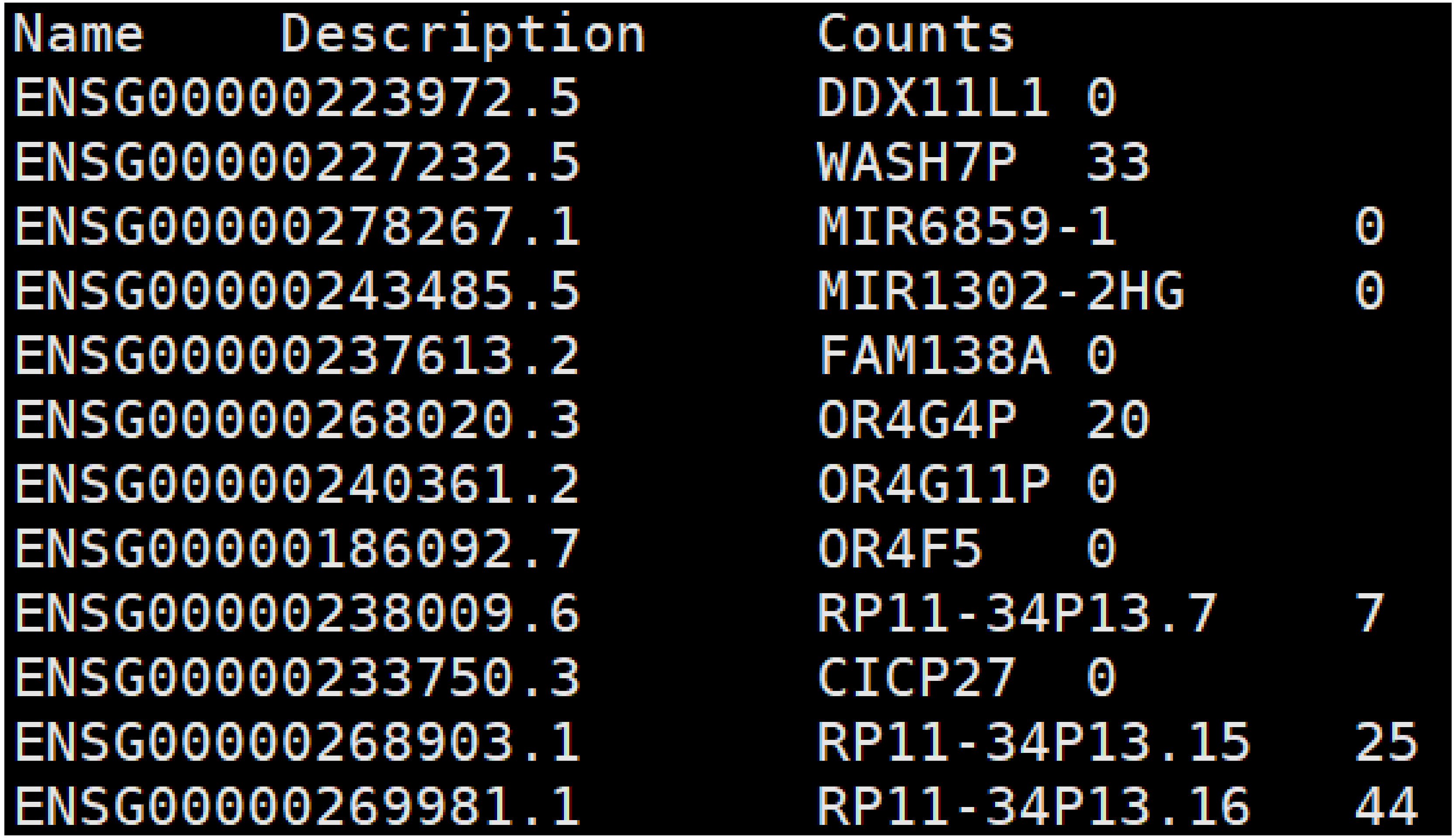


```
                      Number of input reads |       61005207
                  Average input read length |       201
                                UNIQUE READS:
               Uniquely mapped reads number |       54463221
                    Uniquely mapped reads % |       89.28%
                      Average mapped length |       200.68
                   Number of splices: Total |       12932947
        Number of splices: Annotated (sjdb) |       12690090
                   Number of splices: GT/AG |       12797028
                   Number of splices: GC/AG |       92278
                   Number of splices: AT/AC |       11517
           Number of splices: Non-canonical |       32124
                  Mismatch rate per base, % |       0.32%
                     Deletion rate per base |       0.01%
                    Deletion average length |       1.74
                    Insertion rate per base |       0.01%
                   Insertion average length |       1.51
```

```
                             MULTI-MAPPING READS:
       Number of reads mapped to multiple loci |       4667909
            % of reads mapped to multiple loci |       7.65%
       Number of reads mapped to too many loci |       70690
            % of reads mapped to too many loci |       0.12%
                                  UNMAPPED READS:
  Number of reads unmapped: too many mismatches |       0
       % of reads unmapped: too many mismatches |       0.00%
           Number of reads unmapped: too short |       1766919
                % of reads unmapped: too short |       2.90%
               Number of reads unmapped: other |       36468
                    % of reads unmapped: other |       0.06%
                                  CHIMERIC READS:
                      Number of chimeric reads |       0
                           % of chimeric reads |       0.00%
```

- **STAR mapping result**

RNA sequencing – Quantify



```
Name    Description     Counts
ENSG00000223972.5       DDX11L1 0
ENSG00000227232.5       WASH7P  33
ENSG00000278267.1       MIR6859-1       0
ENSG00000243485.5       MIR1302-2HG     0
ENSG00000237613.2       FAM138A 0
ENSG00000268020.3       OR4G4P  20
ENSG00000240361.2       OR4G11P 0
ENSG00000186092.7       OR4F5   0
ENSG00000238009.6       RP11-34P13.7    7
ENSG00000233750.3       CICP27  0
ENSG00000268903.1       RP11-34P13.15   25
ENSG00000269981.1       RP11-34P13.16   44
```

- ➢ Quantify results

RNA sequencing – 101 sample mapping reads percentage

CURRENT PROTOCOLS A Wiley Brand in Bioinformatics

UNIT | Full Access

Mapping RNA-seq Reads with STAR

Alexander Dobin, Thomas R. Gingeras

First published: 03 September 2015 | https://doi.org/10.1002/0471250953.bi1114s51 | Citations: 578

- Very good mapping rate : Uniquely mapping 90% 초과
- Good mapping rate : Uniquely mapping 80% 이상
- **Low mapping rate : Uniquely mapping 50% 미만**
  - ➢ **Indicative of a problem with library preparations or data processing**

- Insufficient depletion of ribosomal RNA(rRNA)
- Poor sequencing quality
- Exogenous RNA/DNA contamination
- Computational processing problems

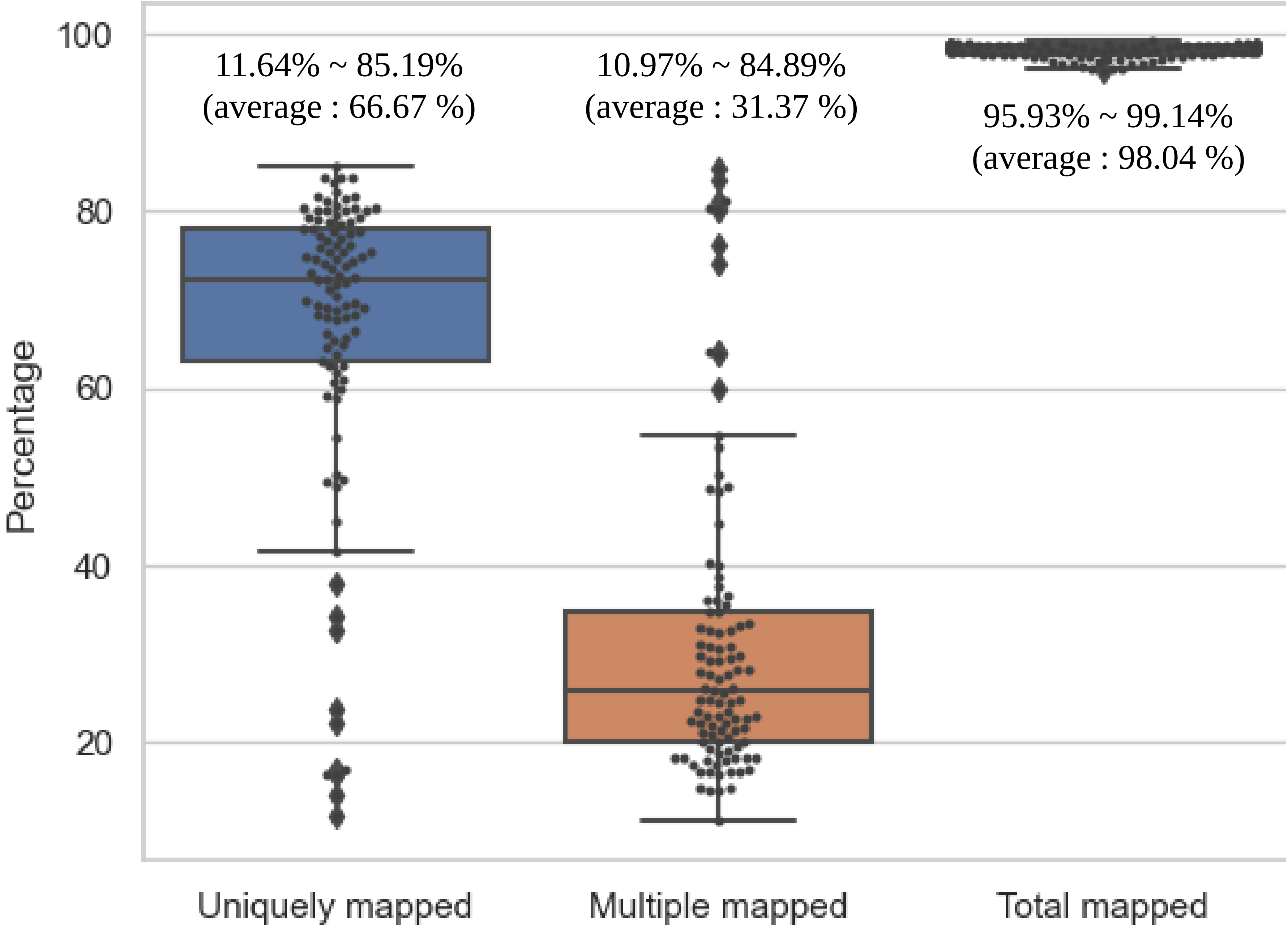
Alexander Dobin et al., *CURRENT PROTOCOLS*, 2015

RNA sequencing – 101 sample mapped reads %

Sample ID	Uniquely mapped reads %	% of reads mapped to multiple loci	Uniquely + multiple mapped reads %
ID1	68.94	29.12	98.06
ID2	77.75	20.46	98.21
ID3	83.83	14.52	98.35
ID4	79.29	19.04	98.33
ID5	74.55	22.93	97.48
ID6	74.83	22.88	97.71
ID7	75.4	23.37	98.77
ID8	74.41	21.52	95.93
ID9	16.21	80.2	96.41
ID11	74.86	22.72	97.58
…	…	…	…
ID100	49.53	48.86	98.39
ID101	54.3	44.62	98.92
ID102	16.22	81.06	97.28
ID103	74.13	24.72	98.85
ID104	37.89	59.85	97.74
ID105	64.51	34.63	99.14
ID106	62.56	36.54	99.1
ID107	48.81	50.19	99

- Uniquely mapped reads % < 86% (11.64% ~ 85.19%)
- % of reads mapped to multiple loci < 85% (10.97% ~ 84.89%)
- Uniquely + multiple mapped reads % : 95.93% ~ 99.14%

RNA sequencing – 101 sample mapped reads percentage



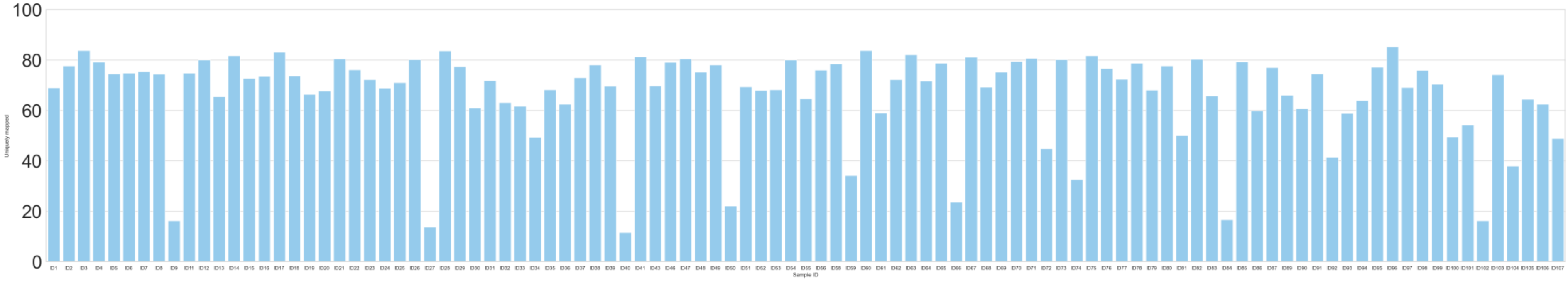
11.64% ~ 85.19%
(average : 66.67 %)

10.97% ~ 84.89%
(average : 31.37 %)

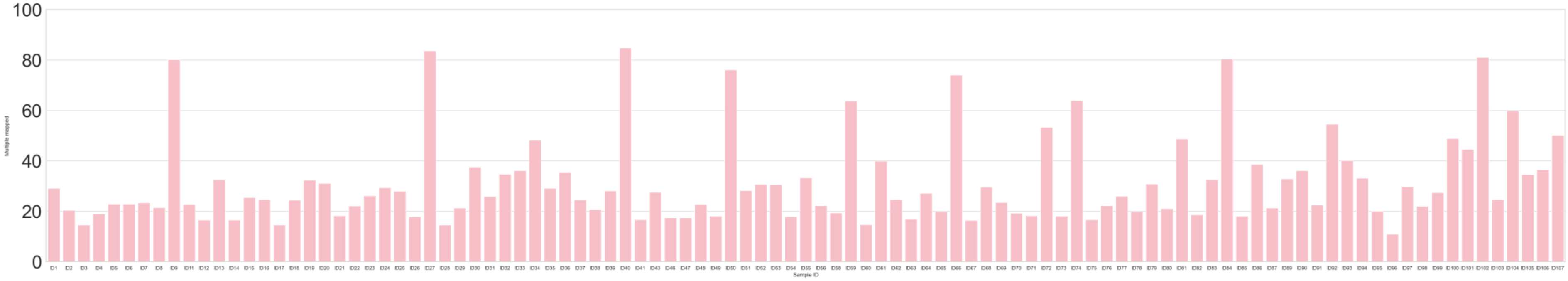
95.93% ~ 99.14%
(average : 98.04 %)

RNA sequencing – 101 sample mapped reads percentage

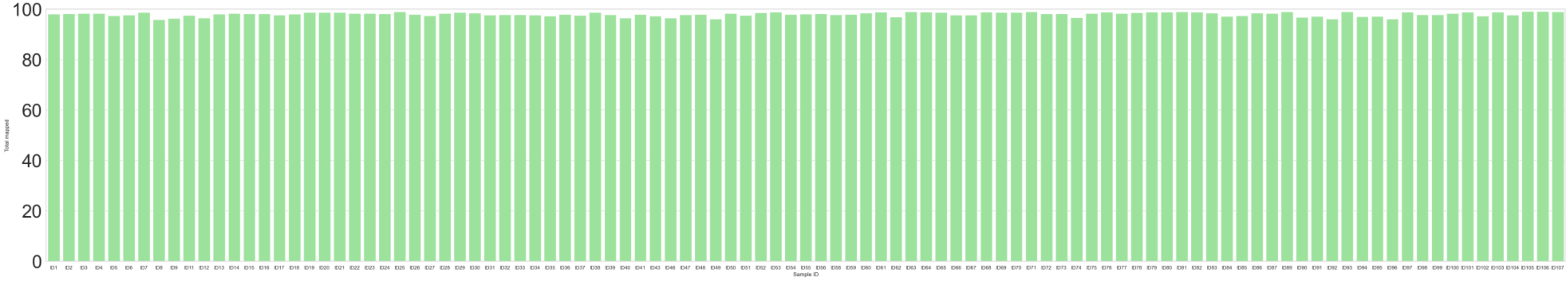
- Uniquely mapped reads : 11.64% ~ 85.19% (average : 66.67 %)



- Multiple loci mapped reads : 10.97% ~ 84.89% (average : 31.37 %)

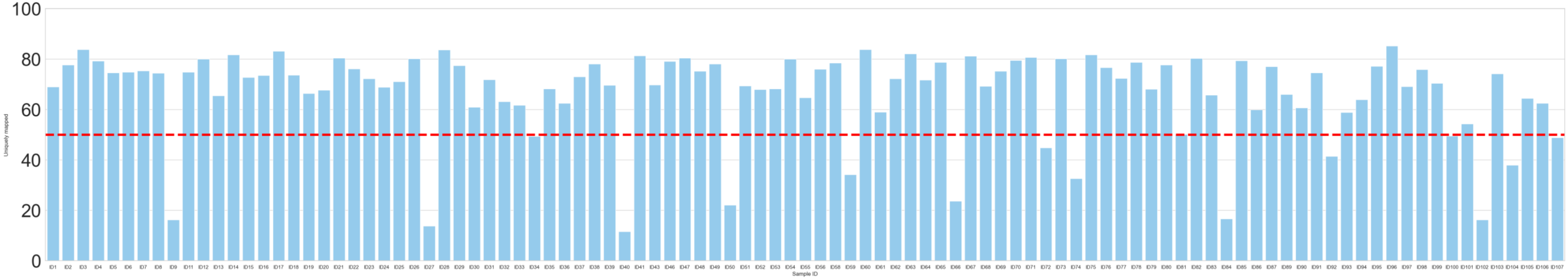


- Total mapped reads : 95.93% ~ 99.14% (average : 98.04 %)



RNA sequencing – 101 sample mapping reads percentage

- Uniquely mapping reads



11.64% ~ 85.19% (average : 66.67 %)

Low mapping rate (< 50%) : 15 sample

(ID9, ID27, ID34, ID40, ID50, ID59, ID66, ID72, ID74, ID84, ID92, ID100, ID102, ID104, ID107)

As 15 CD patients showed high rRNA ratio (>40%) in sample QC, the RNA sequencing of these 15 samples were repeated and aligned to the reference genome

86 sample

101 sample

RNA sequencing – normalization

RPKM & FPKM(reads or fragments Pre kilobase of transcript Pre million mapped reads)

- It is suitable to **compare gene expression levels** within a **single sample**
- **Rescaled** to correct for both **library size** and **gene length**

➢ RPKM : read 기준으로 gene or transcript length를 보정 → paired-end의 경우 2의 리드로 간주

➢ FPKM **:** fragment 기준으로 gene or transcript length를 보정 → paired-end의 경우 두 개의 read가 한 fragment를 이루므로 한 개로 간주

$$RPKM_i \text{ or } FPKM_i = \frac{q_i}{\frac{l_i}{10^3} * \frac{\sum_j q_j}{10^6}} = \frac{q_i}{l_i * \sum_j q_j} * 10^9$$

- q_i : read(RPKM) or fragment(FPKM) counts
- l_i : gene or transcript length
- $\sum_j q_j$: total read(RPKM) or fragment(FPKM) counts

Yingdong Zhao et al., *BMC*, 2021

RNA sequencing – normalization

TPM(transcripts Pre million)

- TPM was introduced in an attempt to facilitate **comparisons across samples**
- The sum of all TPM values is the same in all samples

$$TPM_i = \frac{q_i/l_i}{\sum_j (q_j/l_j)} * 10^6 \;\; = \left(\frac{FPKM_i}{\sum_j FPKM_j} \right) * 10^6$$

- q_i : reads mapped to transcript
- l_i : transcript length
- $\sum_j (q_j/l_i)$: the sum of mapped reads to transcript normalized by transcript length.

Yingdong Zhao et al., *BMC*, 2021

RNA sequencing – normalization

RPKM & FPKM vs TPM example(× 10^9 생략)

Gene	Length	Sample 1 read count	Sample 2 read count	Sample 1 RPKM (FPKM)	Sample 2 RPKM (FPKM)	Sample 1 TPM	Sample 2 TPM
A	10	10	10	10 / (10 × 25) = **0.04**	10 / (10 × 40) = **0.025**	0.04 / 0.1 = **0.4**	0.025 / 0.0625 = **0.4**
B	5	5	0	5 / (5 × 25) = **0.04**	5 / (0 × 40) = **0**	0.04 / 0.1 = **0.4**	0 / 0.0625 = **0**
C	20	10	30	10 / (20 × 25) = **0.02**	30 / (20 × 40) = **0.0375**	0.02 / 0.1 = **0.2**	0.0375 / 0.0625 = **0.6**
Total				**0.1**	**0.0625**	**1.0**	**1.0**