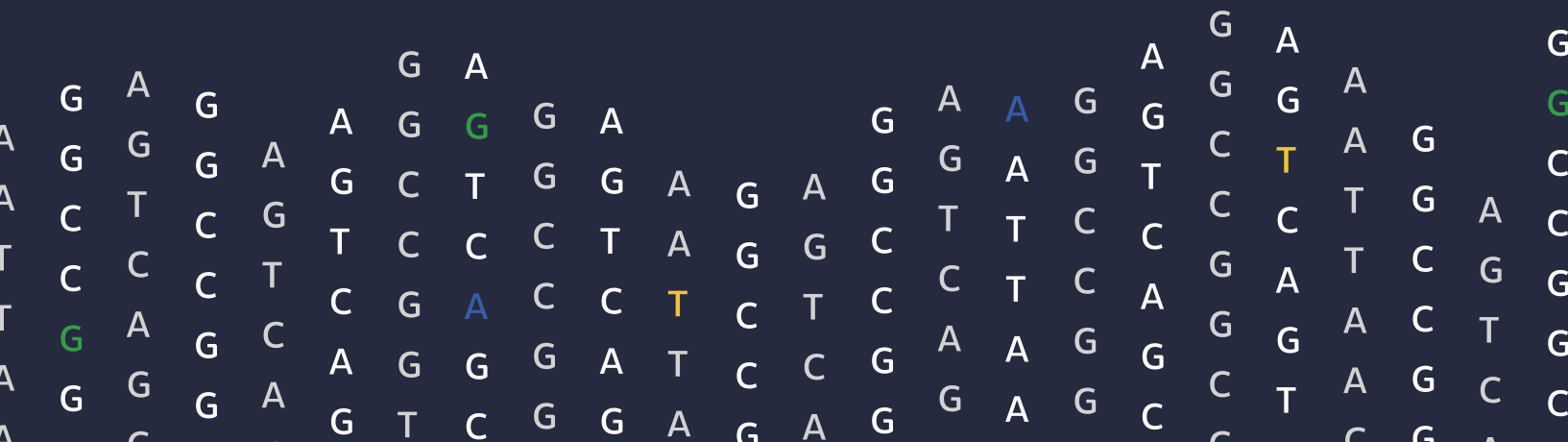


# All-In-One Service

Report for Whole-Genome Sequencing



ichrogene

시 료 명 EUS001

의뢰기관 아이크로진 생명정보 연구소

의뢰자 홍0동

분석일 2022-02-07



# All-In-One Service

Report for Whole-Genome Sequencing

ichrogene

시 료 명 EUS001

의뢰기관 아이크로진 생명정보 연구소

의 료 자 홍0동

분 석 일 2022-02-07





# Contents

All-In-One Service

## 1. Flow chart and Tools for Analysis

1-1 Flow chart

1-2 Tools and parameters

## 2. Raw Data QC (Read 1, 2)

2-1 Per base quality score, Per sequence quality score

2-2 Per sequence content, Per sequence GC content

2-3 Per base N content, Sequence duplication level

2-4 Adapter content

## 3. Mapped Data QC

3-1 Per sequence content, Base quality score

3-2 Mapping quality score, Coverage histogram

3-3 GC content, Distribution of Read length

3-4 InDel size, Insert Size

## 4. Sequence Variant QC

Description of all columns

### 리포트 형식

PDF : Fastq 파일과 bam 파일에 대한 quality control 결과 리포트입니다. [본 보고서](#)

CSV : 분석 완료 후 vcf의 QC 결과 리포트입니다. [PDF 파일과 별도 제공](#)



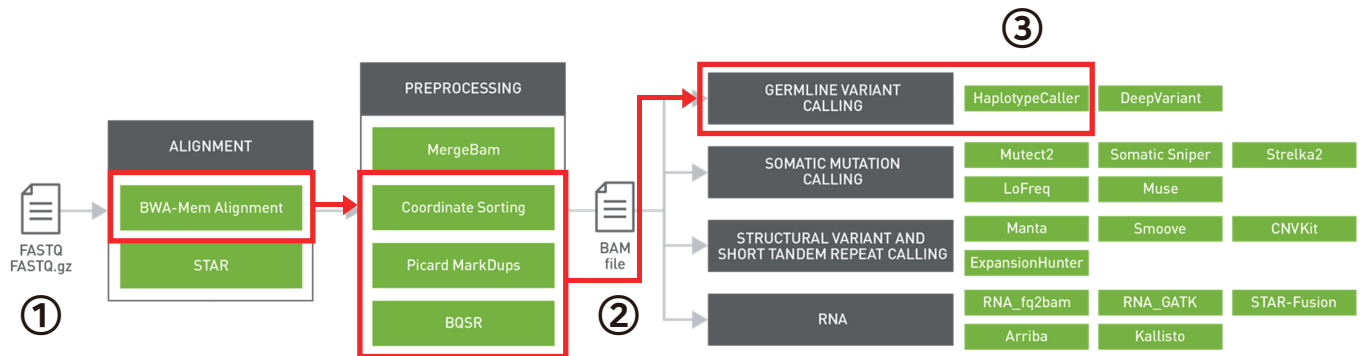
## **Flow chart and Tools for Analysis**



# Flow chart

All-In-One Service

## NVIDIA 社の Clara Parabricks 소프트웨어의 파이프라인을 활용하여 분석 수행



[ NVIDIA Clara Parabricks 파이프라인 (빨간 박스 : WGS 파이프라인) ]

## 본 QC 리포트에는 아래와 같은 3단계 데이터의 QC가 포함 (번호는 상단 그림 참조)

- ① RAW data QC (fastq 파일; Paired-end의 경우 Read1와 Read2, Single-end의 경우 Read1)
- ② Mapped Data QC (bam 파일)
- ③ Sequence variant QC (vcf 파일)

# Tools and parameters

All-In-One Service

## 1. Tools

- ① NVIDIA Clara Parabricks
- ② GATK
- ③ Fastqc
- ④ Alfred
- ⑤ rtg-tool

## 2. Parameters

- ① Sequencing library type : Paired End sequencing
- ② NGS platform : Whole Genome Sequencing
- ③ Input files import : /home/test/XXXX
- ④ Target import : Agilent SureSelect All Exon V.5
- ⑤ Genome build : hg19
- ⑥ Bam out : On
- ⑦ Output format : uncompressed VCF

## Raw Data QC

(Read 1)

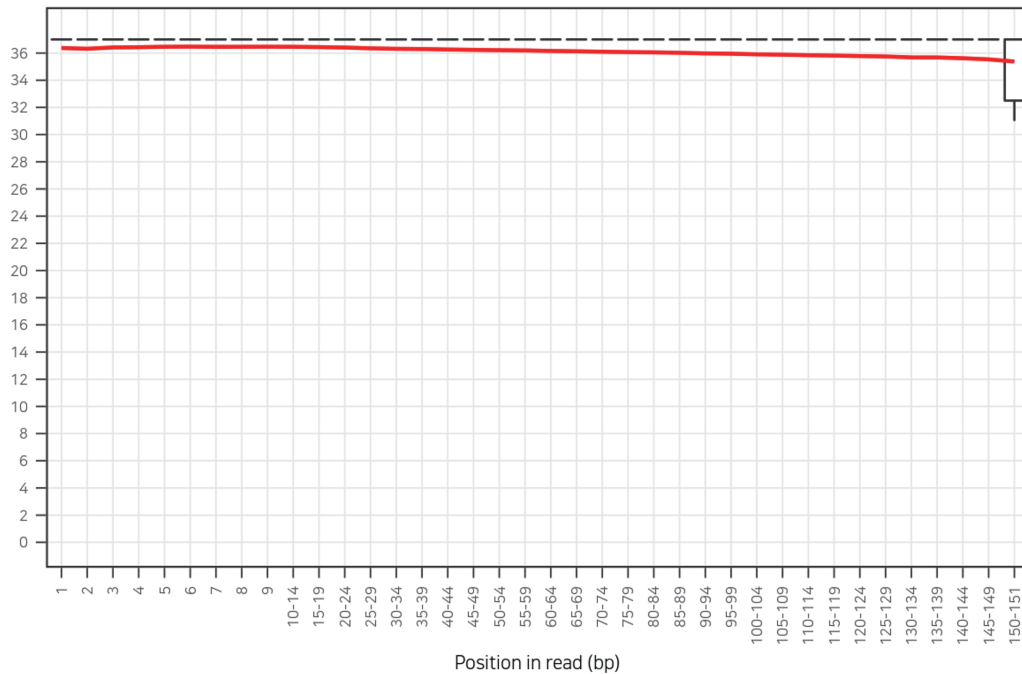




## Per base quality score (base pair 당 품질 점수)

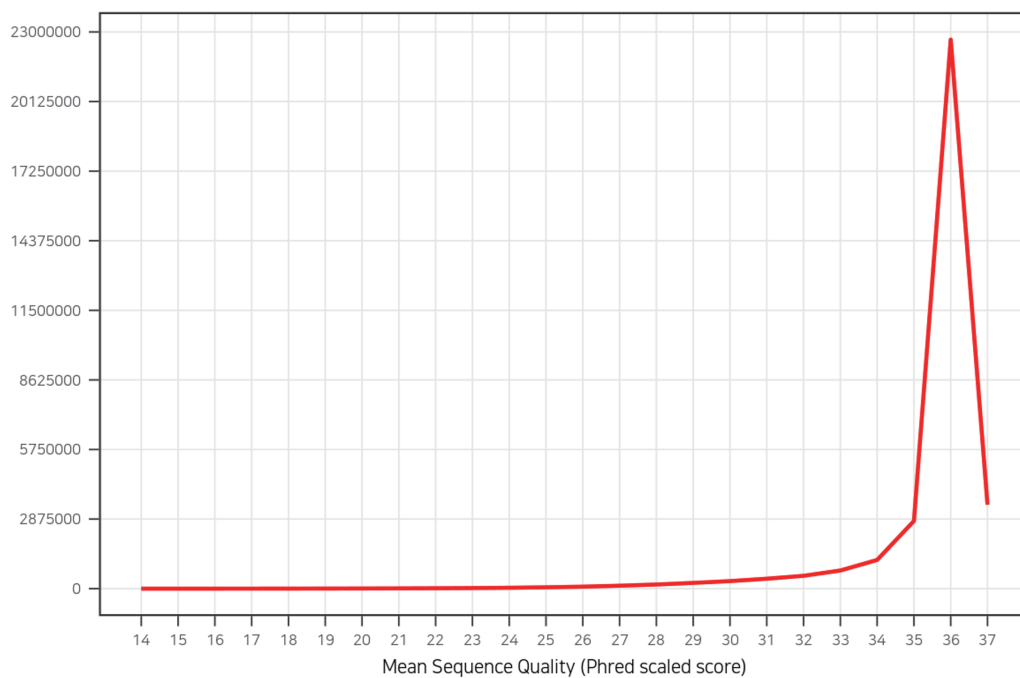
FASTQ 파일의 각 위치 (position) 에서 모든 베이스의 품질 값의 범위에 대한 개요이며, 빨간 선은 중앙값을 의미합니다. 시퀀싱이 진행됨에 따라 대부분의 플랫폼에서 call의 품질이 저하되기 때문에 우측으로 갈 수록 품질이 저하되는 것이 일반적입니다.

· 정상 범위 : 20 이상



## Per sequence quality score (Read의 품질 점수 분포)

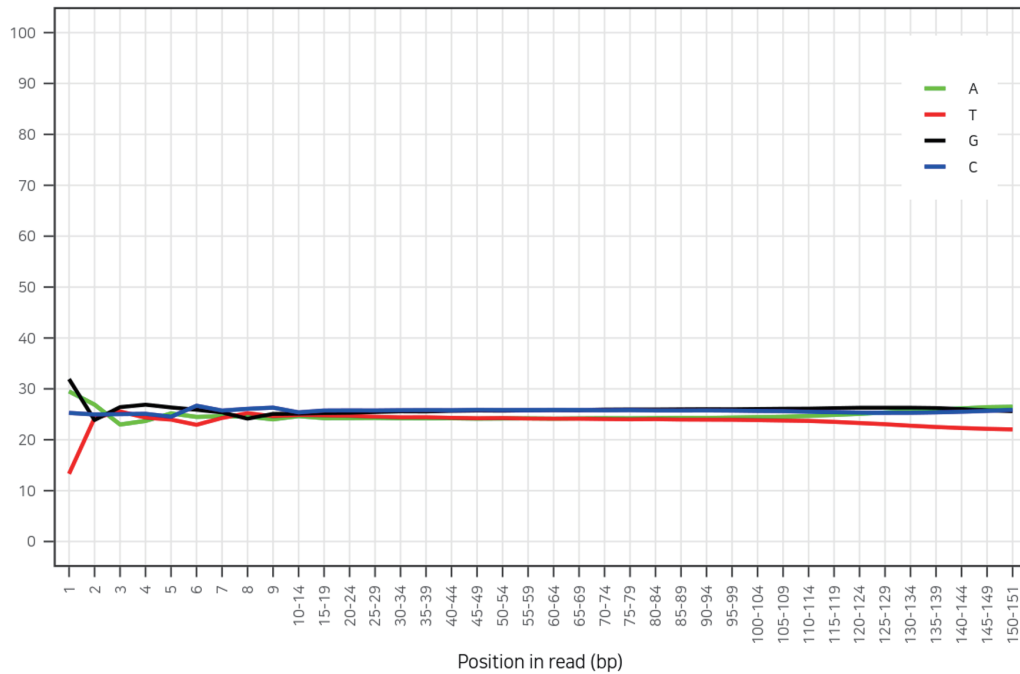
특정 시퀀스 품질값을 가지는 read의 수를 나타냅니다. Y축은 시퀀스 품질 점수(x축)에 해당하는 read의 개수이며, 그래프의 우측으로 갈 수록 (시퀀스 품질이 높을 수록) read의 개수가 많은 것이 일반적입니다.



## Per sequence content (Nucleotide 비율)

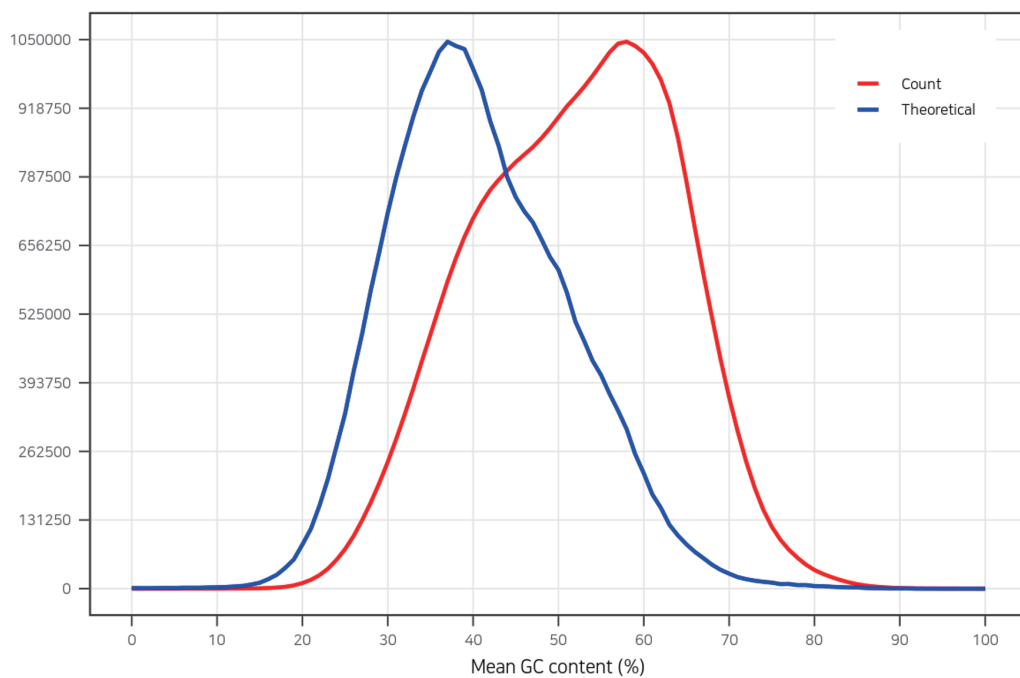
전체 read의 특정 position에 존재하는 ATGC의 비율을 나타냅니다. Nucleotide는 A-T, G-C 쌍을 이루기 때문에 각 쌍의 비율이 일정해야 합니다.

· 정상 범위 : 20~30% 사이에서 일정한 비율 분포 (이론상 모든 base는 25%씩 동일한 비율로 존재)



## Per sequence GC content (Read의 GC 비율)

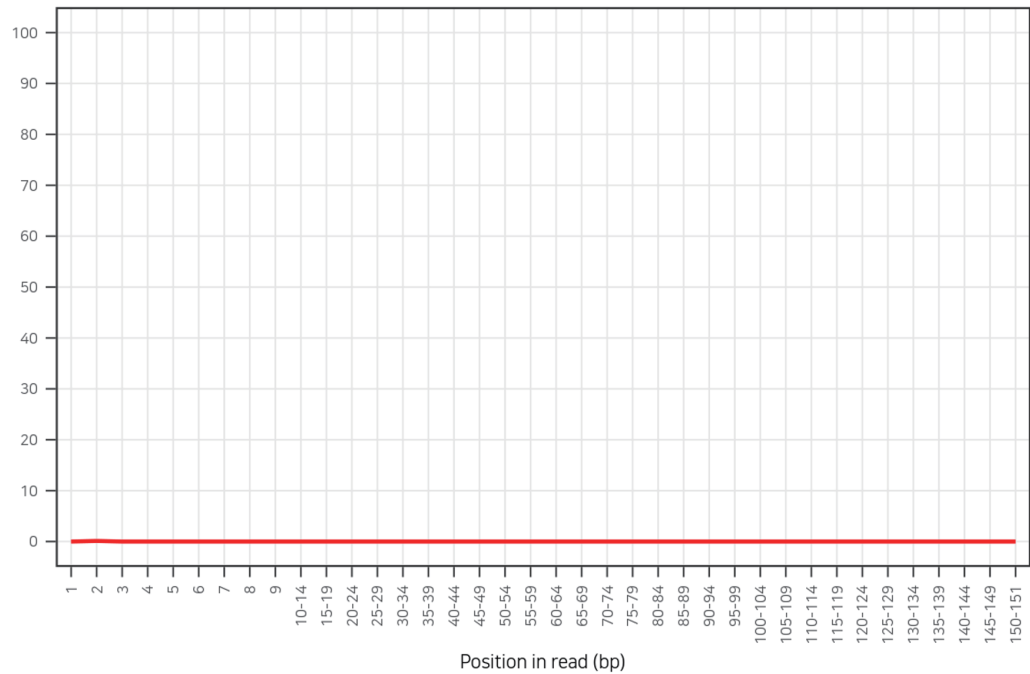
각 read에 포함된 GC의 비율을 나타낸 그래프로, 파란색이 이론상의 GC content의 분포이며, 빨간색과 파란색의 그래프가 비슷한 양상으로 보이는 것이 일반적입니다.



Per base N content (Read의 N 비율)

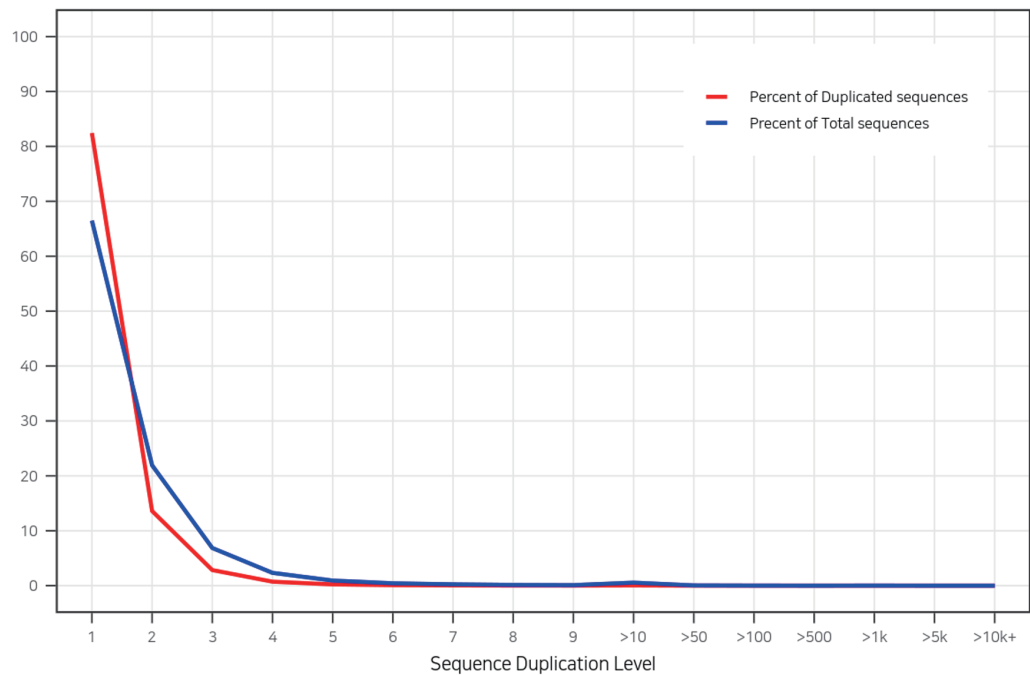
전체 read의 특정 위치에서 N (Any nucleotide)으로 시퀀싱된 비율을 나타냅니다. N은 어떤 nucleotide인지 판별이 되지 않은 상태를 의미합니다.

· 정상 : 빨간 선이 y축 값 0 (그래프의 하단)에 위치



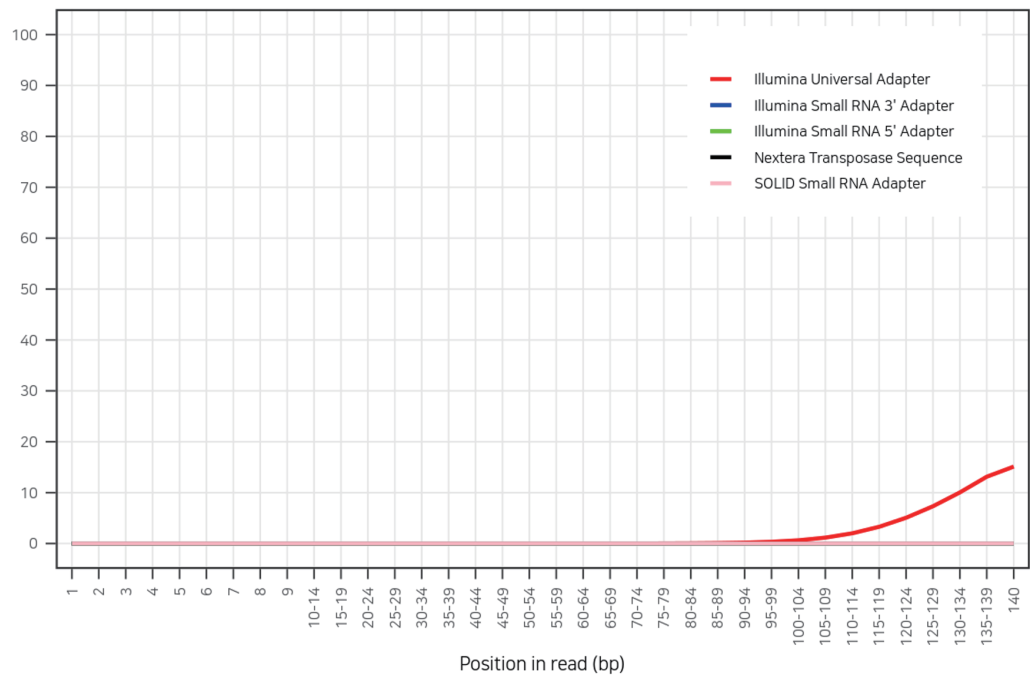
Sequence duplication level (중복된 read의 비율)

전체 read 중에서 중복된 read의 수를 나타낸 그래프입니다. 좌측으로 갈 수록 중복이 없다는 의미로 우측으로 갈수록 그래프가 0으로 낮아지는 것이 일반적입니다. x축 기준 2이상의 read는 중복된 read라는 의미이며, PCR 중복일 가능성이 높습니다.



### Adapter content (Read내의 adapter 비율)

시퀀싱 라이브러리 준비 과정에서 삽입되는 외부 유래의 시퀀스가 사라지지 않고 남아있는 비율을 나타내는 그래프로, 측정한 5개의 외부 유래 시퀀스가 모두 y축 기준 하단에 위치하는 것이 정상입니다. x축 기준 우측에 0 이상의 구간이 존재하는 것은 외부 유래 시퀀스가 아직 제거되지 않았다는 의미이며, 이를 제거하는 과정인 trimming 과정이 필요할 수 있습니다. 하지만, 현재의 대부분의 mapping tool은 외부유래 시퀀스를 감지하여 자동으로 mapping 보정을 하는 기능이 내재되어 있습니다.



## Raw Data QC

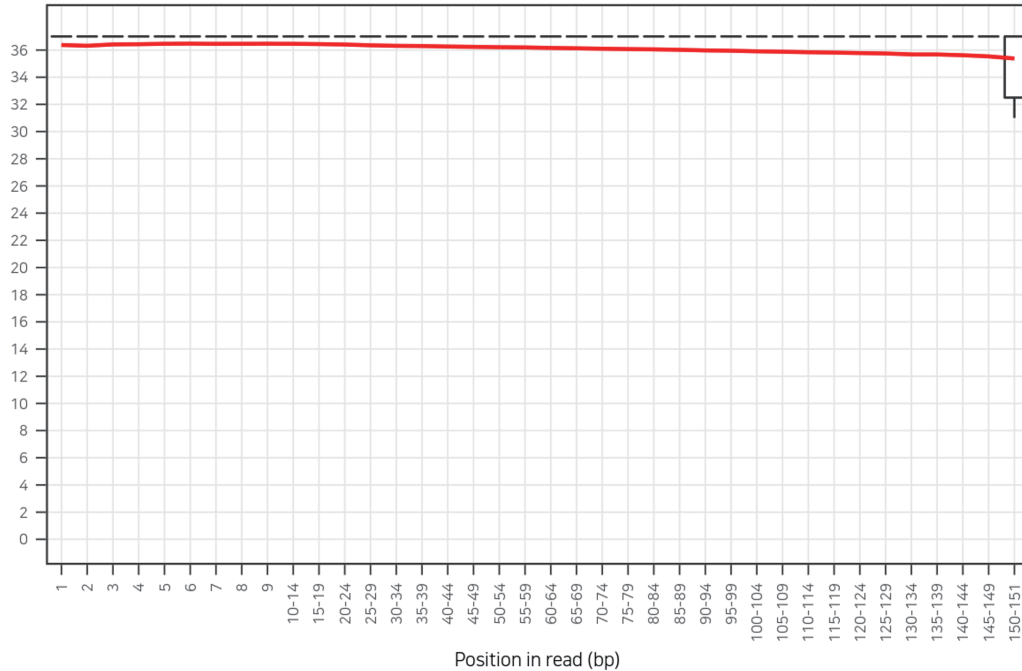
(Read 2)



## Per base quality score (base pair 당 품질 점수)

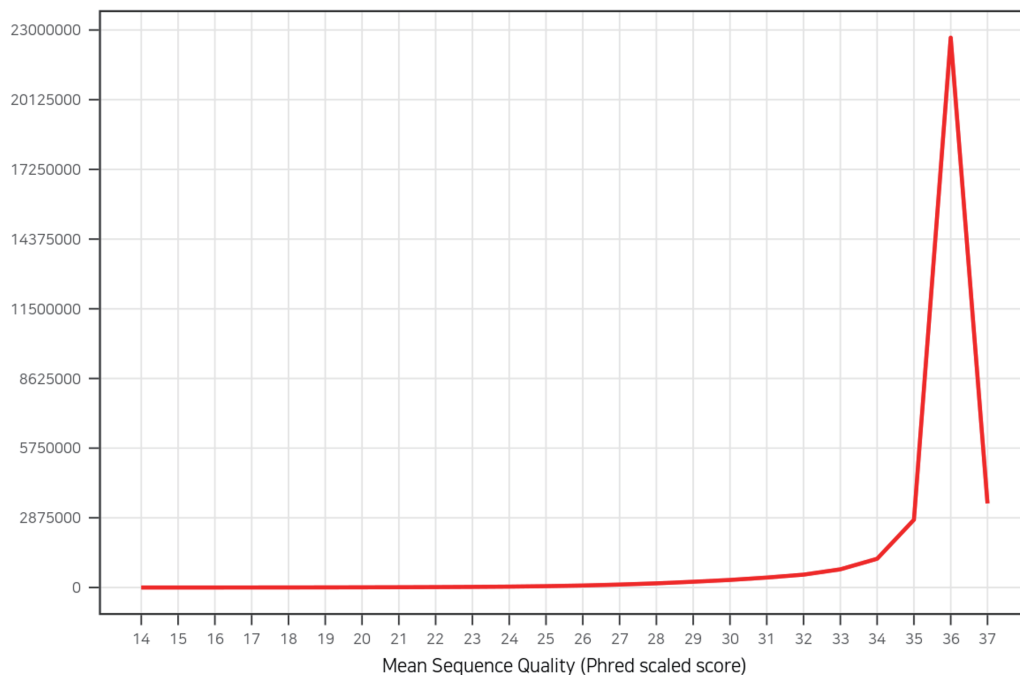
FASTQ 파일의 각 위치 (position) 에서 모든 베이스의 품질 값의 범위에 대한 개요이며, 빨간 선은 중앙값을 의미합니다. 시퀀싱이 진행됨에 따라 대부분의 플랫폼에서 call의 품질이 저하되기 때문에 우측으로 갈 수록 품질이 저하되는 것이 일반적입니다.

· 정상 범위 : 20 이상



## Per sequence quality score (Read의 품질 점수 분포)

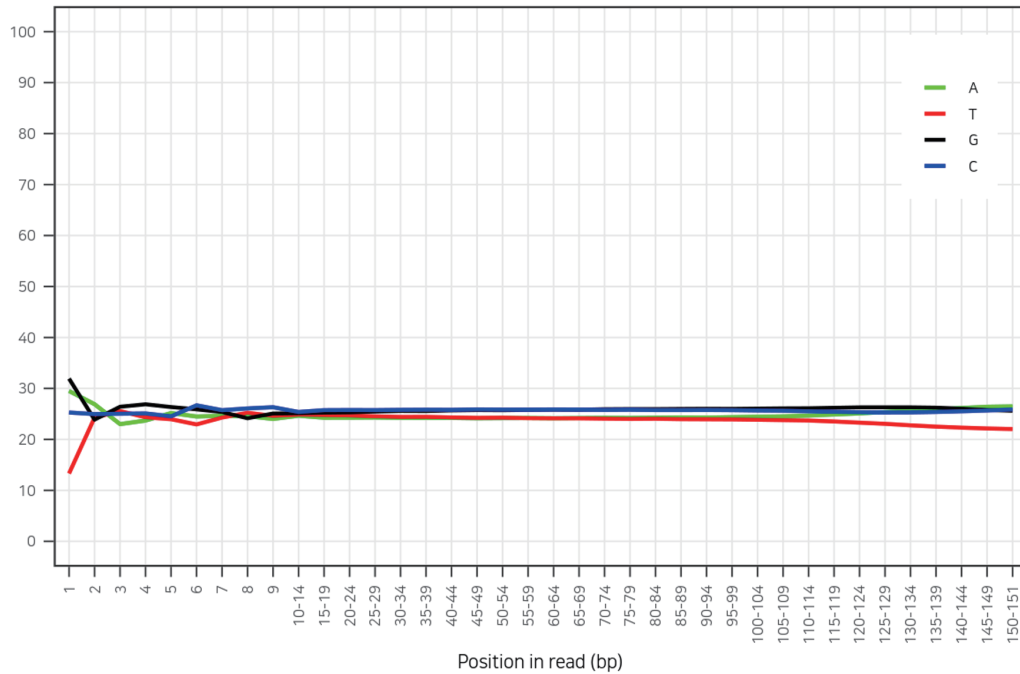
특정 시퀀스 품질값을 가지는 read의 수를 나타냅니다. Y축은 시퀀스 품질 점수(x축)에 해당하는 read의 개수이며, 그래프의 우측으로 갈 수록 (시퀀스 품질이 높을 수록) read의 개수가 많은 것이 일반적입니다.



## Per sequence content (Nucleotide 비율)

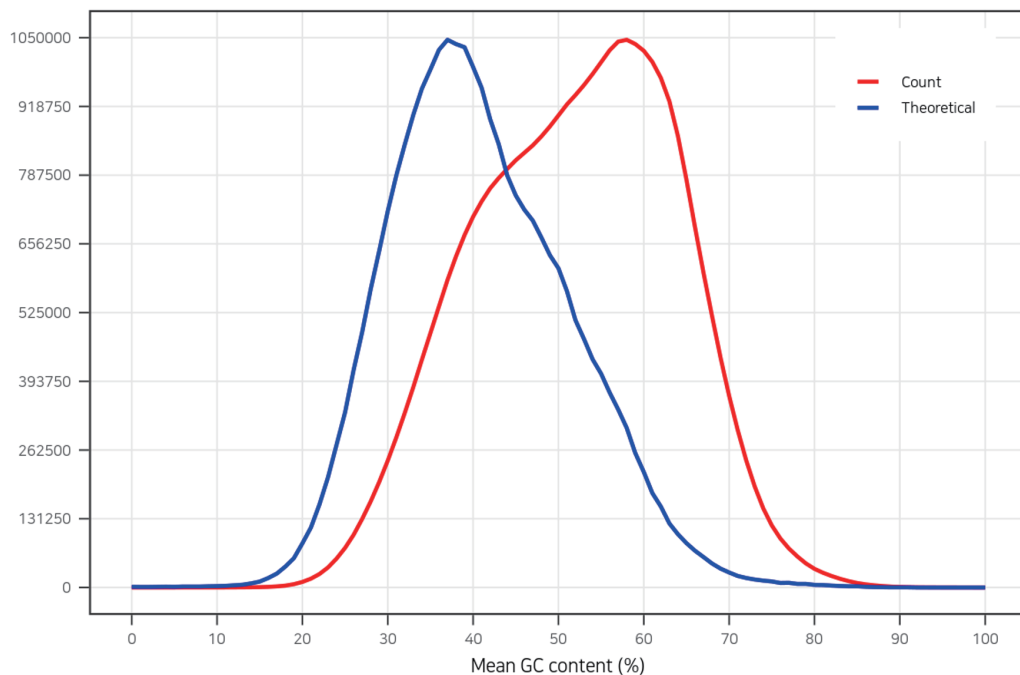
전체 read의 특정 position에 존재하는 ATGC의 비율을 나타냅니다. Nucleotide는 A-T, G-C 쌍을 이루기 때문에 각 쌍의 비율이 일정해야 합니다.

· 정상 범위 : 20~30% 사이에서 일정한 비율 분포 (이론상 모든 base는 25%씩 동일한 비율로 존재)



## Per sequence GC content (Read의 GC 비율)

각 read에 포함된 GC의 비율을 나타낸 그래프로, 파란색이 이론상의 GC content의 분포이며, 빨간색과 파란색의 그래프가 비슷한 양상으로 보이는 것이 일반적입니다.

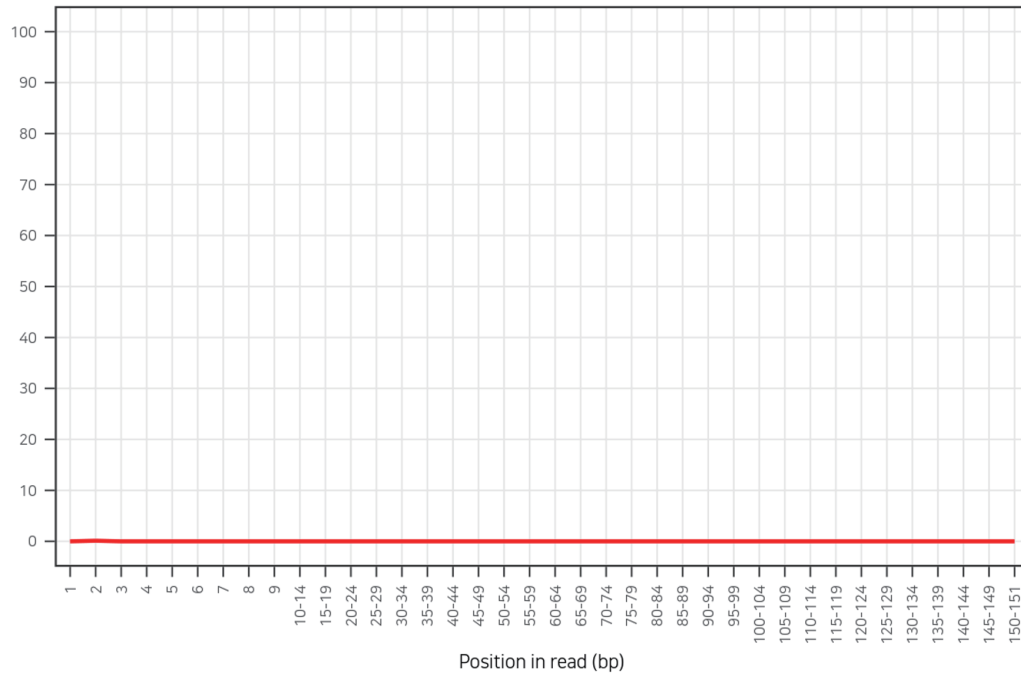




## Per base N content (Read의 N 비율)

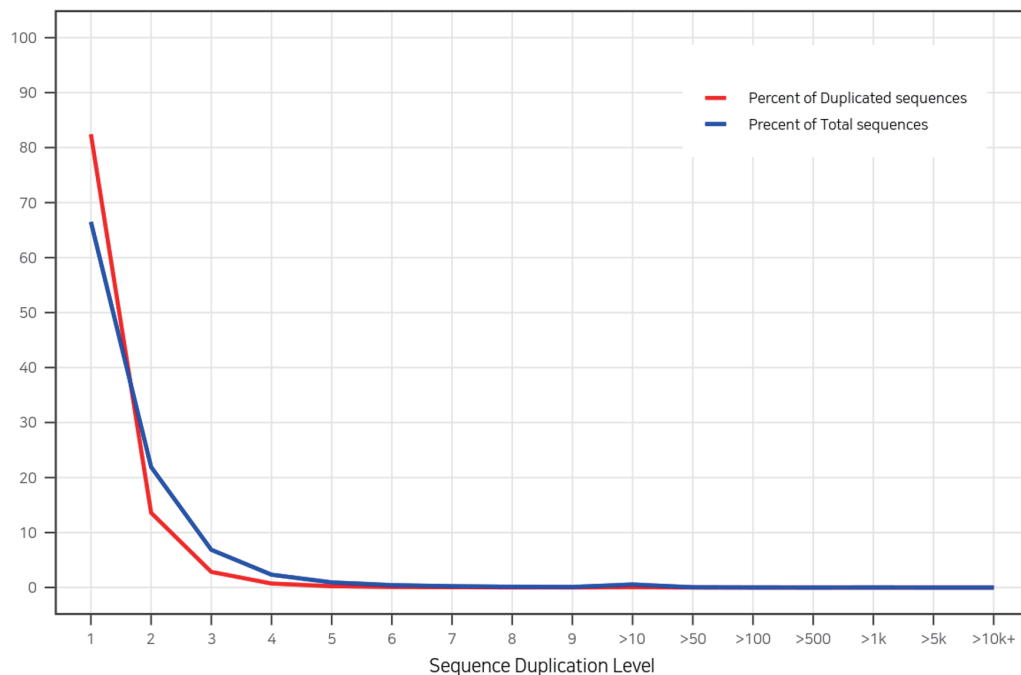
전체 read의 특정 위치에서 N (Any nucleotide)으로 시퀀싱된 비율을 나타냅니다. N은 어떤 nucleotide인지 판별이 되지 않은 상태를 의미합니다.

· 정상 : 빨간 선이 y축 값 0 (그래프의 하단)에 위치



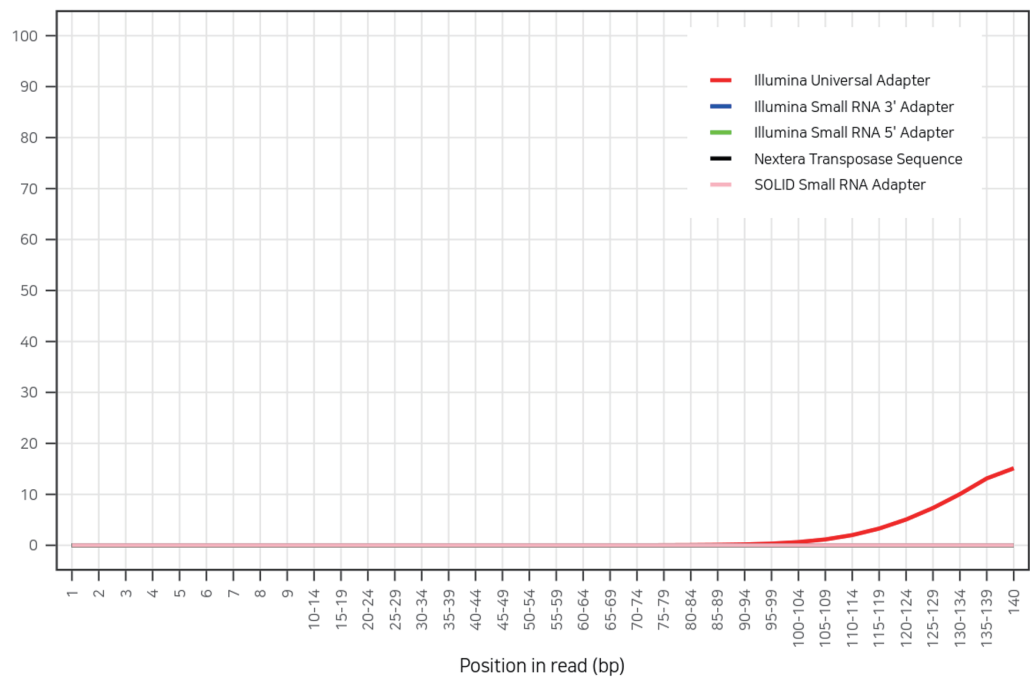
## Sequence duplication level (중복된 read의 비율)

전체 read 중에서 중복된 read의 수를 나타낸 그래프입니다. 좌측으로 갈 수록 중복이 없다는 의미로 우측으로 갈수록 그래프가 0으로 낮아지는 것이 일반적입니다. x축 기준 2이상의 read는 중복된 read라는 의미이며, PCR 중복일 가능성이 높습니다.



## Adapter content (Read내의 adapter 비율)

시퀀싱 라이브러리 준비 과정에서 삽입되는 외부 유래의 시퀀스가 사라지지 않고 남아있는 비율을 나타내는 그래프로, 측정한 5개의 외부 유래 시퀀스가 모두 y축 기준 하단에 위치하는 것이 정상입니다. x축 기준 우측에 0 이상의 구간이 존재하는 것은 외부 유래 시퀀스가 아직 제거되지 않았다는 의미이며, 이를 제거하는 과정인 trimming 과정이 필요할 수 있습니다. 하지만, 현재의 대부분의 mapping tool은 외부유래 시퀀스를 감지하여 자동으로 mapping 보정을 하는 기능이 내재되어 있습니다.



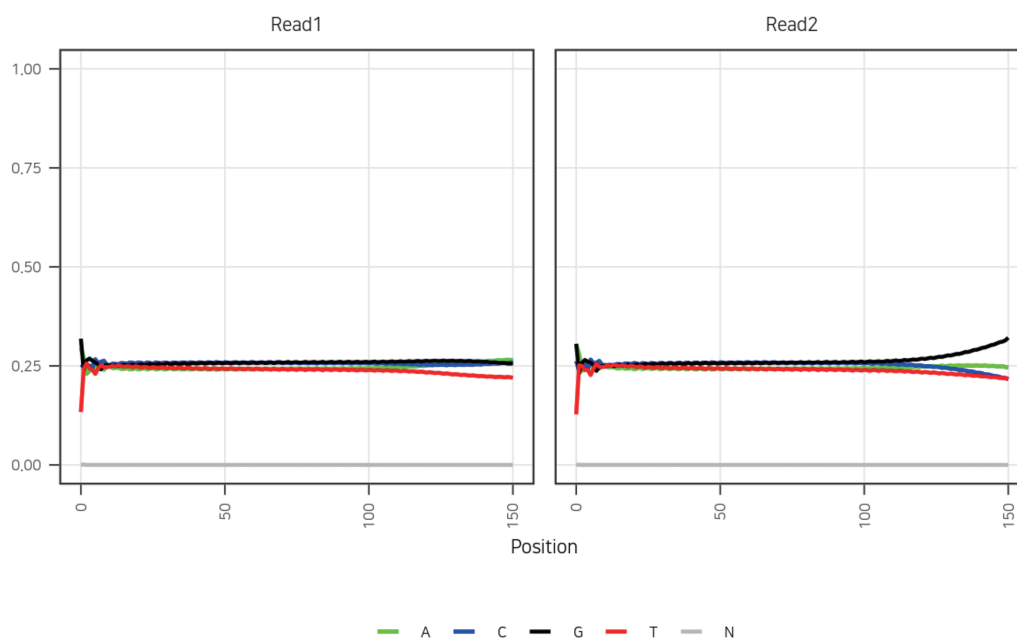
## **Mapped Data QC**



## Per base quality score (base pair 당 품질 점수)

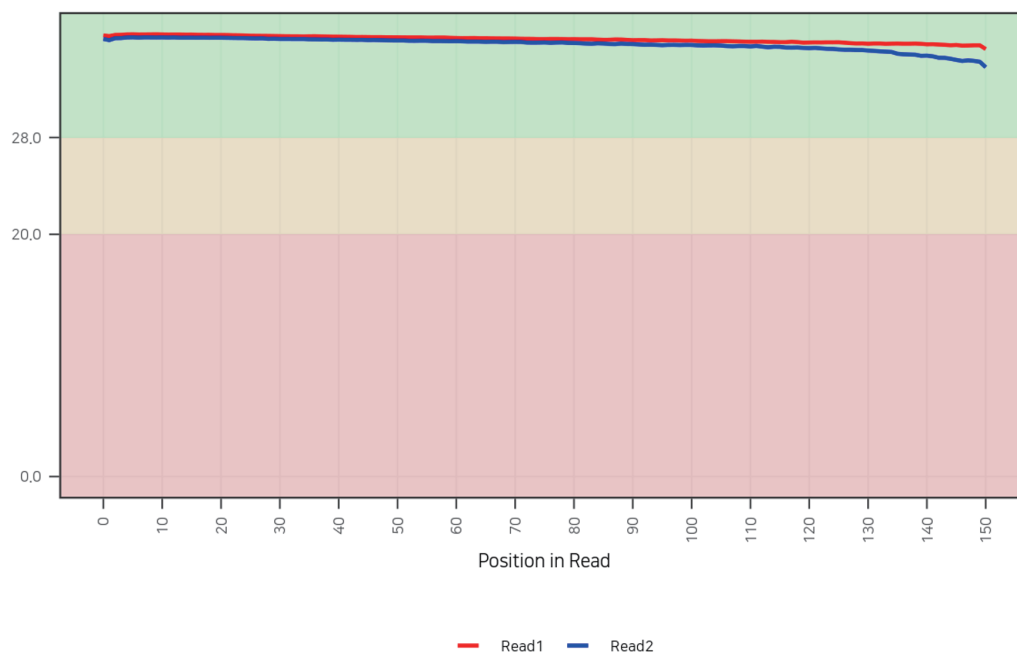
매핑 후 각 read의 특정 위치에 존재하는 ATGC와 N (Any nucleotide)의 비율입니다. 매핑 후에도 nucleotide의 비율이 일정하게 유지되는 것이 일반적입니다.

· 정상 범위 : ATGC는 20~30% 사이에서 일정한 비율 분포, N의 비율이 0%



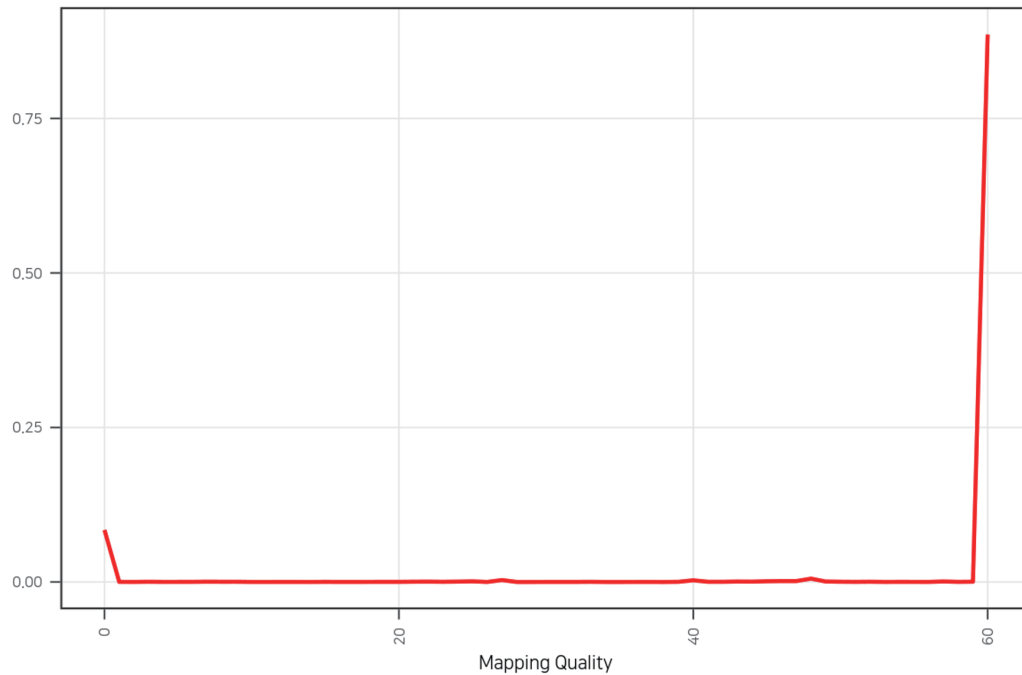
## Base quality score (평균 mapping score)

전체 read에서 각 위치의 평균 매핑 점수입니다. 매핑 점수가 높다는 것은 참조유전체와 비교하여 정확한 위치에 매핑되어, 분석 시 에러 발생확률이 낮음을 의미합니다. 빨간선과 파란선은 paired-end의 각 read를 의미합니다. 일반적으로 초록색 영역에 그래프가 존재하는 경우 높은 품질 수준을 의미하며, 빨간색 영역의 경우 나쁜 품질을 의미합니다.



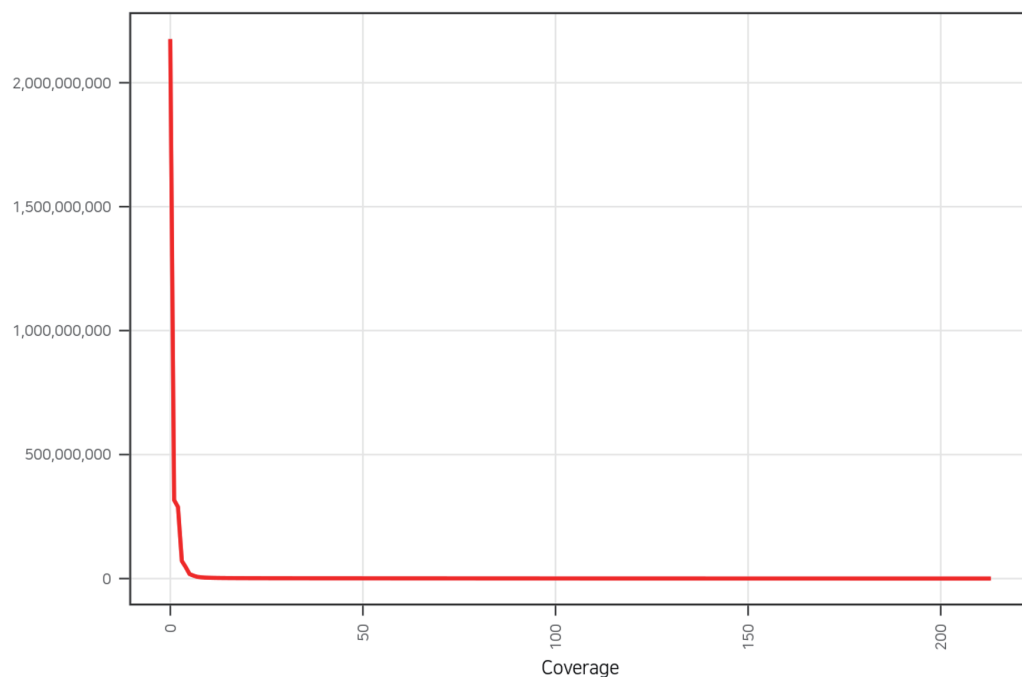
## Mapping quality score (매핑 품질 점수별 read 분포)

Mapping된 전체 read의 품질 정보를 나타내는 그래프입니다. 전체 read의 수에서 특정 매핑 점수를 가지는 read의 분포입니다. x축 기준 우측이 높은 mapping 점수를 갖는 read이며 그래프가 오른쪽으로 치우칠수록 정확하게 매핑된 read의 비율이 높다는 것을 의미합니다.



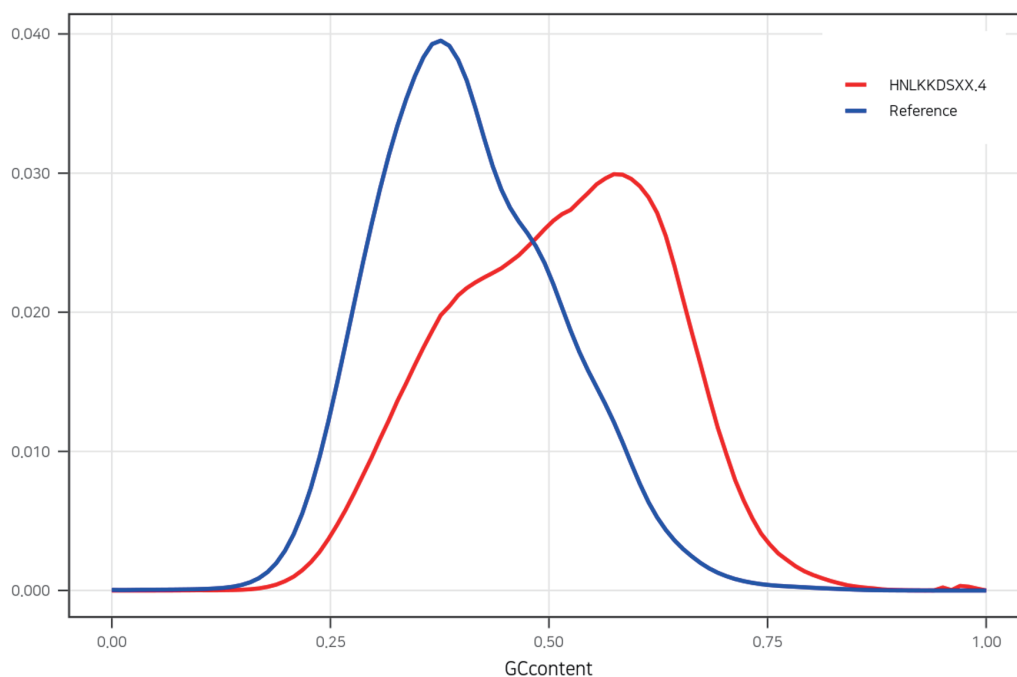
## Coverage histogram (Coverage 분포)

주어진 coverage값을 가지는 base의 수를 나타냅니다. 시퀀싱 전에 목표로 한 coverage 값과 동일한 값의 coverage를 가지는 base가 많을수록 좋습니다. 모든 base에서 coverage가 모두 동일하지는 않기 때문에, coverage가 낮은 부분이 일부 존재할 가능성이 있습니다.



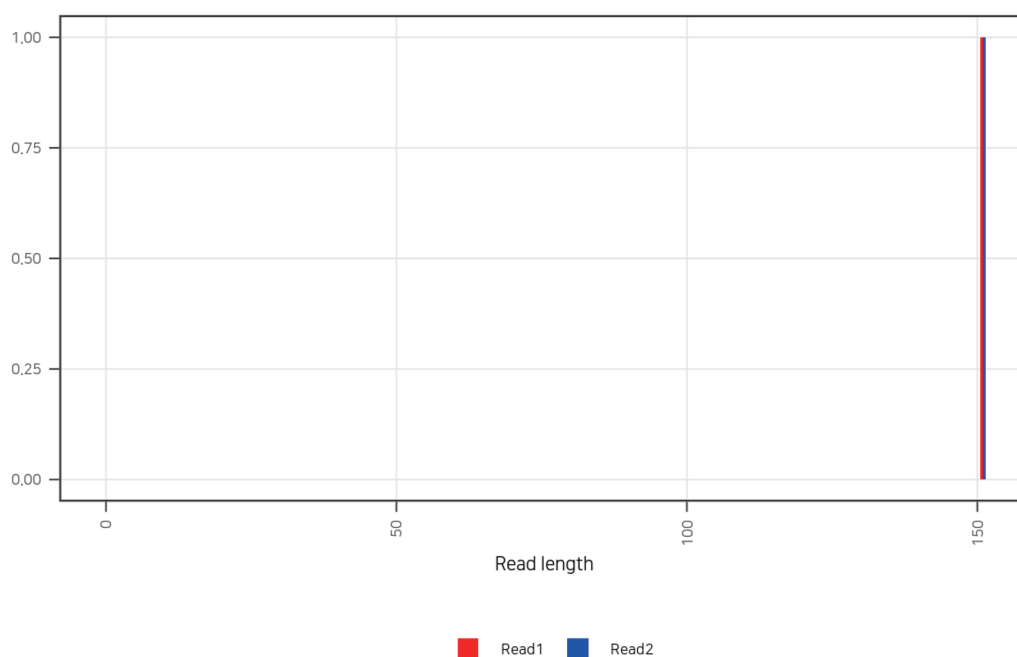
## GC content (Mapping된 read의 GC 비율)

각 read에 포함된 GC의 비율을 나타낸 그래프로, 파란색이 이론상의 GC content의 분포이며, 빨간색과 파란색의 그래프가 비슷한 양상으로 보이는 것이 일반적입니다.



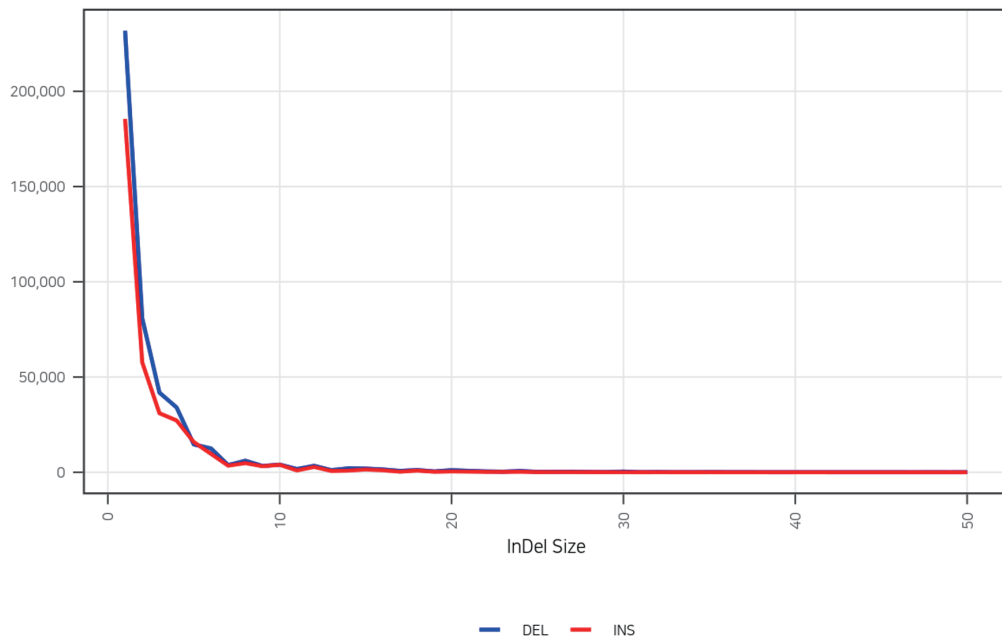
## Distribution of Read length (Read 길이 분포)

Mapping된 read의 길이 분포입니다. 일정한 길이의 read비율이 100% (1.0)에 가까울수록 좋습니다. Read의 길이는 시퀀싱 환경 및 trimming 여부에 따라 차이가 있을 수 있습니다.



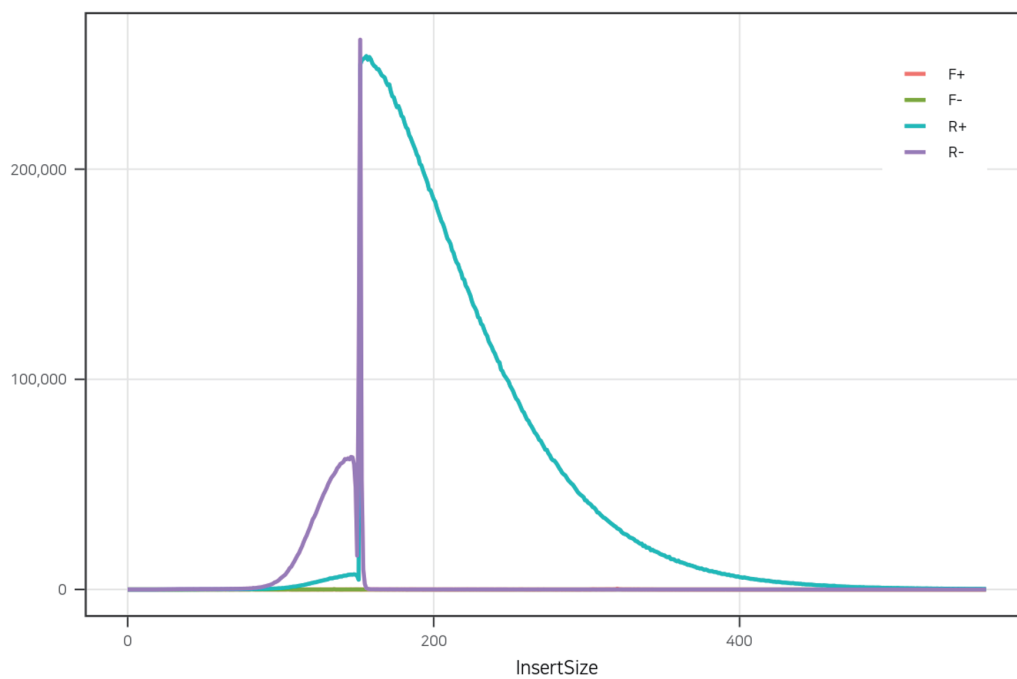
## InDel size (Insertion, deletion의 크기 분포)

참조유전체에 대한 매핑 실행 후, 확인되는 insertion과 deletion의 길이 분포도입니다.



## Insert size (Insert sequence의 크기 분포)

Adapter 사이에 삽입되어 있던 DNA의 길이 분포도입니다.





**Sequence Variant QC**  
(Column Description)



컬럼 상세

| 데이터 열 이름                      | 예시                     | 의미                                |
|-------------------------------|------------------------|-----------------------------------|
| Sample name                   | WES_Sample1            | Sample 이름                         |
| Failed Filters                | 0                      | 필터링을 통과하지 못한 변이의 수                |
| Passed Filters                | 65713885               | 필터링을 통과한 변이의 수                    |
| SNPs                          | 1401810                | SNP의 수                            |
| MNPs                          | 0                      | MNP의 수                            |
| Insertions                    | 73087                  | 삽입(insertion) 수                   |
| Deletions                     | 107667                 | 결실(deletion)의 수                   |
| Indels                        | 971                    | Indel의 수                          |
| Symbolic SVs                  | 70                     | 구조변이의 수                           |
| Same as reference             | 64124443               | 참조유전체와 동일한 변이의 수                  |
| Missing Genotype              | 2390                   | Missing 유전형인 변이의 수                |
| Partial Genotype              | 3447                   | 부분적으로 유전형이 missing인 변이의 수         |
| Phased Genotypes              | 0.3% (225323/65711425) | Haplotype phasing이 된 변이의 수        |
| SNP Transitions/Transversions | 1.62 (1424530/878570)  | Transition과 Transversion의 비율      |
| Total Het/Hom ratio           | 0.54 (559181/1027871)  | Heterozygous/Homozygous 비율        |
| SNP Het/Hom ratio             | 0.56 (501029/900781)   | SNP의 Heterozygous/Homozygous 비율   |
| MNP Het/Hom ratio             | - (0/0)                | MNP의 Heterozygous/Homozygous 비율   |
| Insertion Het/Hom ratio       | 0.38 (19970/53117)     | 삽입의 Heterozygous/Homozygous 비율    |
| Deletion Het/Hom ratio        | 0.48 (34761/72906)     | 결실의 Heterozygous/Homozygous 비율    |
| Indel Het/Hom ratio           | - (971/0)              | Indel의 Heterozygous/Homozygous 비율 |
| Symbolic SV Het/Hom ratio     | 0.01 (1/69)            | 구조변이의 Heterozygous/Homozygous 비율  |
| Insertion/Deletion ratio      | 0.68 (73087/107667)    | Insertion/Deletion 비율             |
| Indel/SNP+MNP ratio           | 0.13 (181725/1401810)  | Indel과 SNP+MNP의 비율                |

리포트는 CSV 파일 형태로 따로 제공됩니다.





ichrogene   
All-In-One Service