



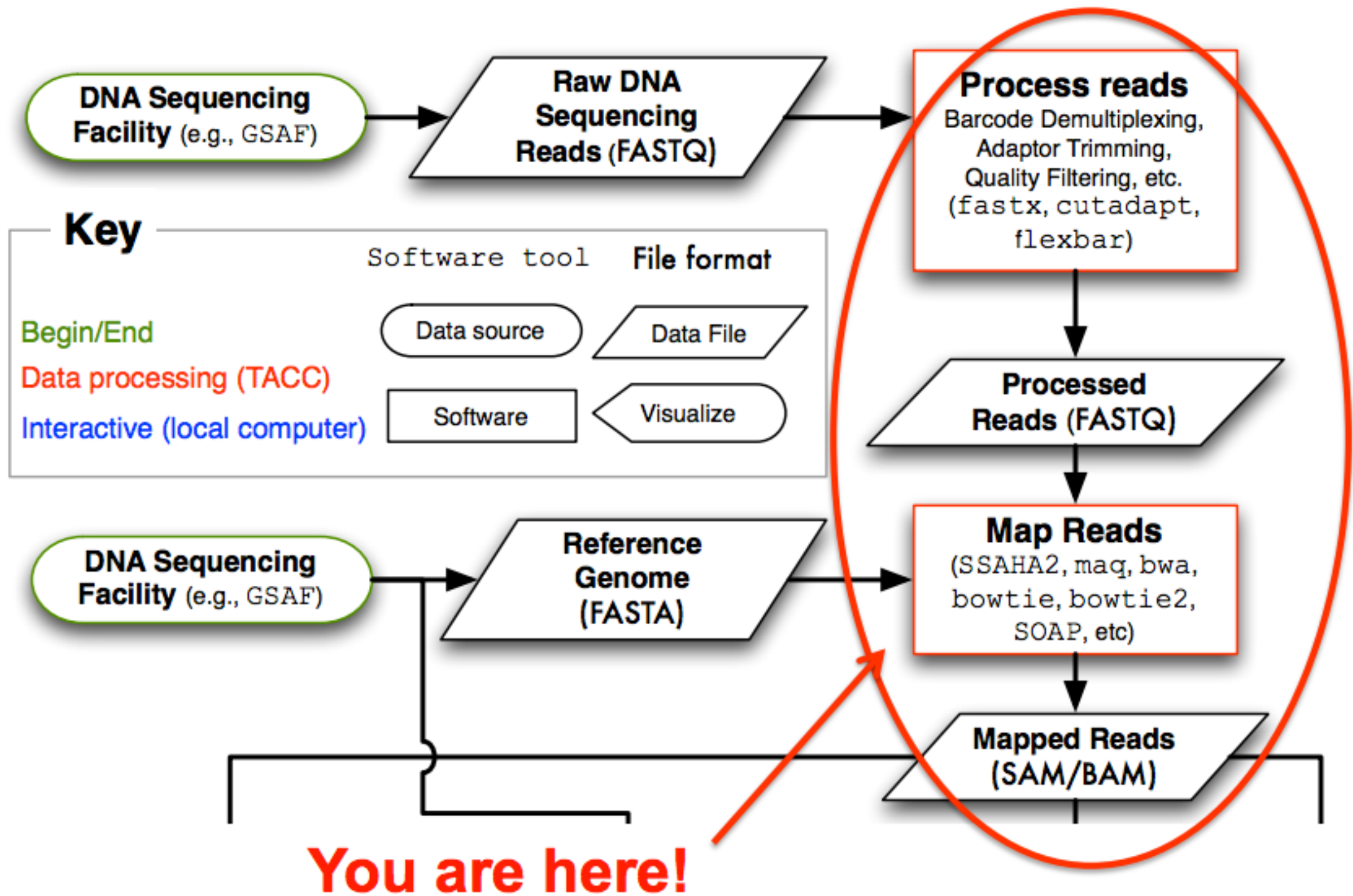
**GENOMIKA**

---

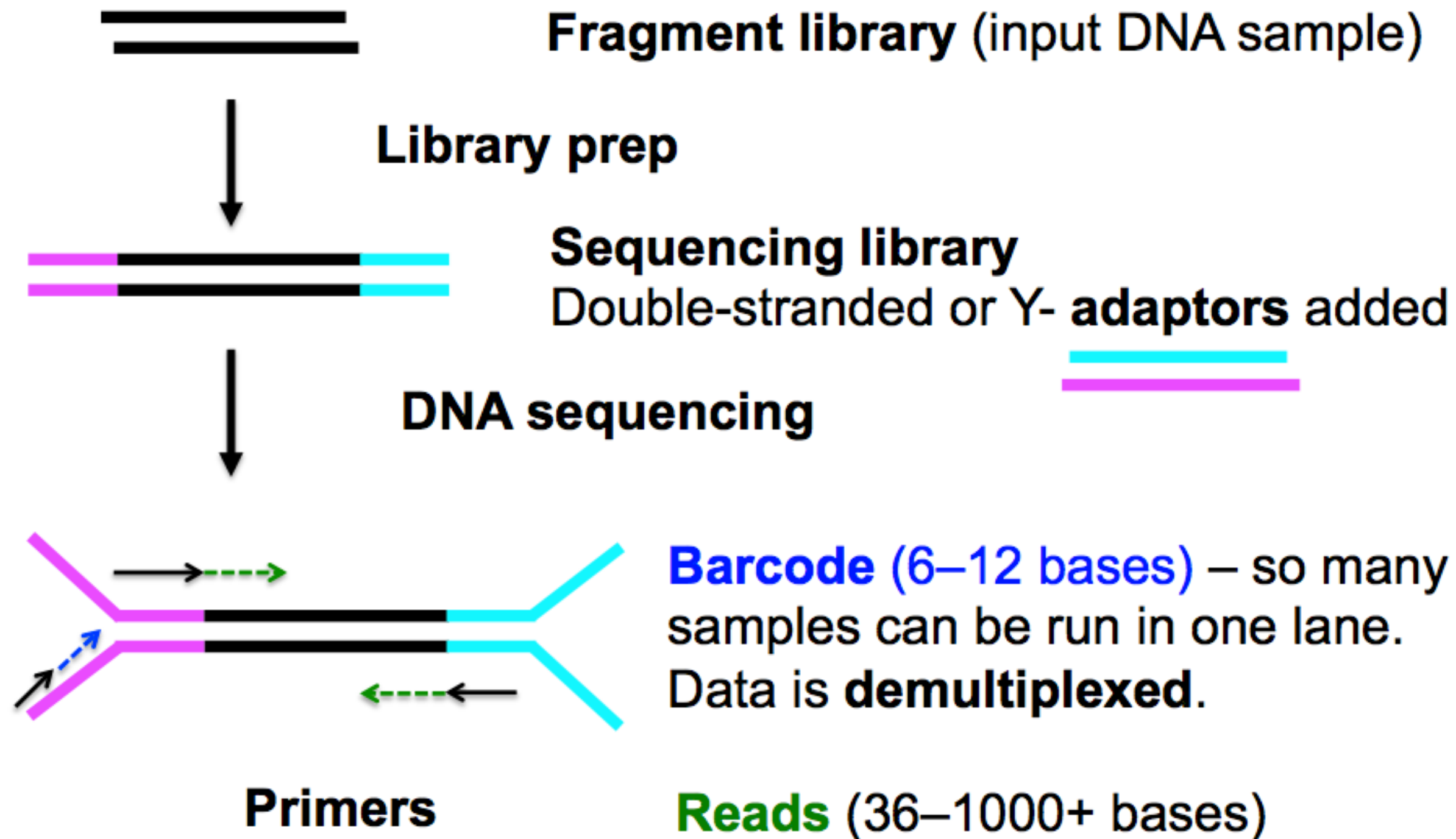
## **Introdução ao NGS**

Rodrigo Bertollo  
[rodrigo@genomika.com.br](mailto:rodrigo@genomika.com.br)

# Pipeline

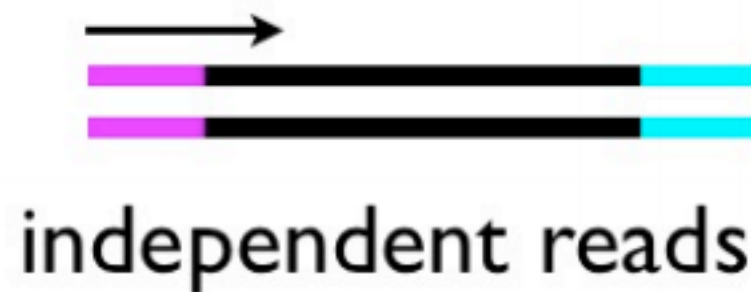


# Terminology

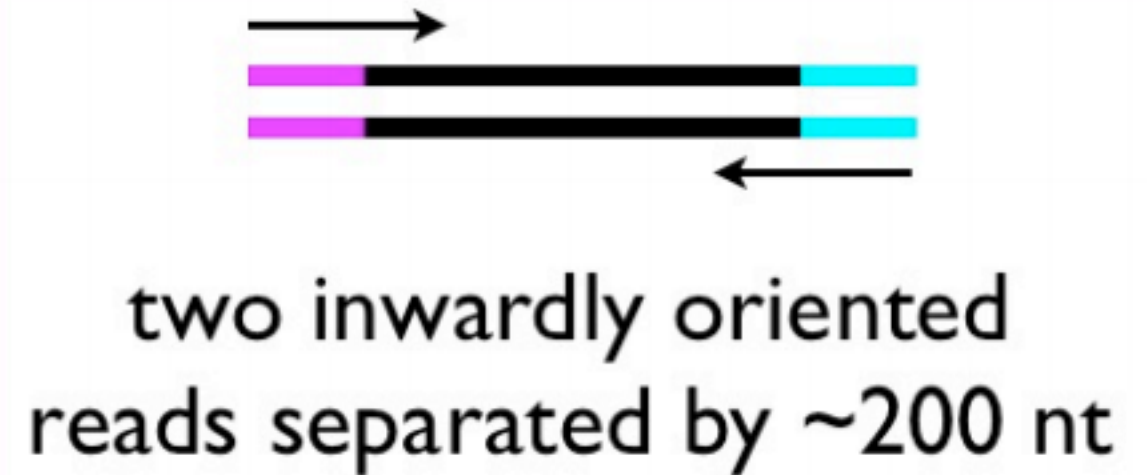


# Types of Illumina fragment libraries

## single-end



## paired-end



## mate-paired



# Read Sequences Quality Control

## Garbage in = garbage out

- Contaminated with other samples?
- Sample barcodes removed?
- Adaptor sequences trimmed?
  - RNAseq, MiSeq data
- Trim ends of reads with poor quality?
  - *de novo* Assembly
- Know your data
  - Paired reads? Relative orientations?
  - Technology specific concerns?
    - Indels with 454





# Read Sequences

## FASTQ Format

```
@HWI-EAS216_91209:1:2:454:192#0/1
GTTGATGAATTTCTCCAGCGCGAATTTGTGGGCT
+HWI-EAS216_91209:1:2:454:192#0/1
B@BBBBBBB@BBBBAAAA>@AABA?BBBAAB??>A?
```

Line 1: @read name

Line 2: called base sequence

Line 3: +read name (optional after +)

Line 4: base quality scores

# FASTQ format

Standard Format for NGS data

Conversion can be done from sff, fasta + qual, . . .

Extension of the Fasta format

Text-based formats (easy to use!)

If not compressed, it can be huge

[http://en.wikipedia.org/wiki/FASTQ\\_format](http://en.wikipedia.org/wiki/FASTQ_format)

# Decipher base quality scores

<http://www.asciitable.com/>

|                   |  |
|-------------------|--|
| Quality character | !"#\$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI |
| ASCII Value       | 33 43 53 63 73                             |
| Base Quality (Q)  | 0 10 20 30 40                              |

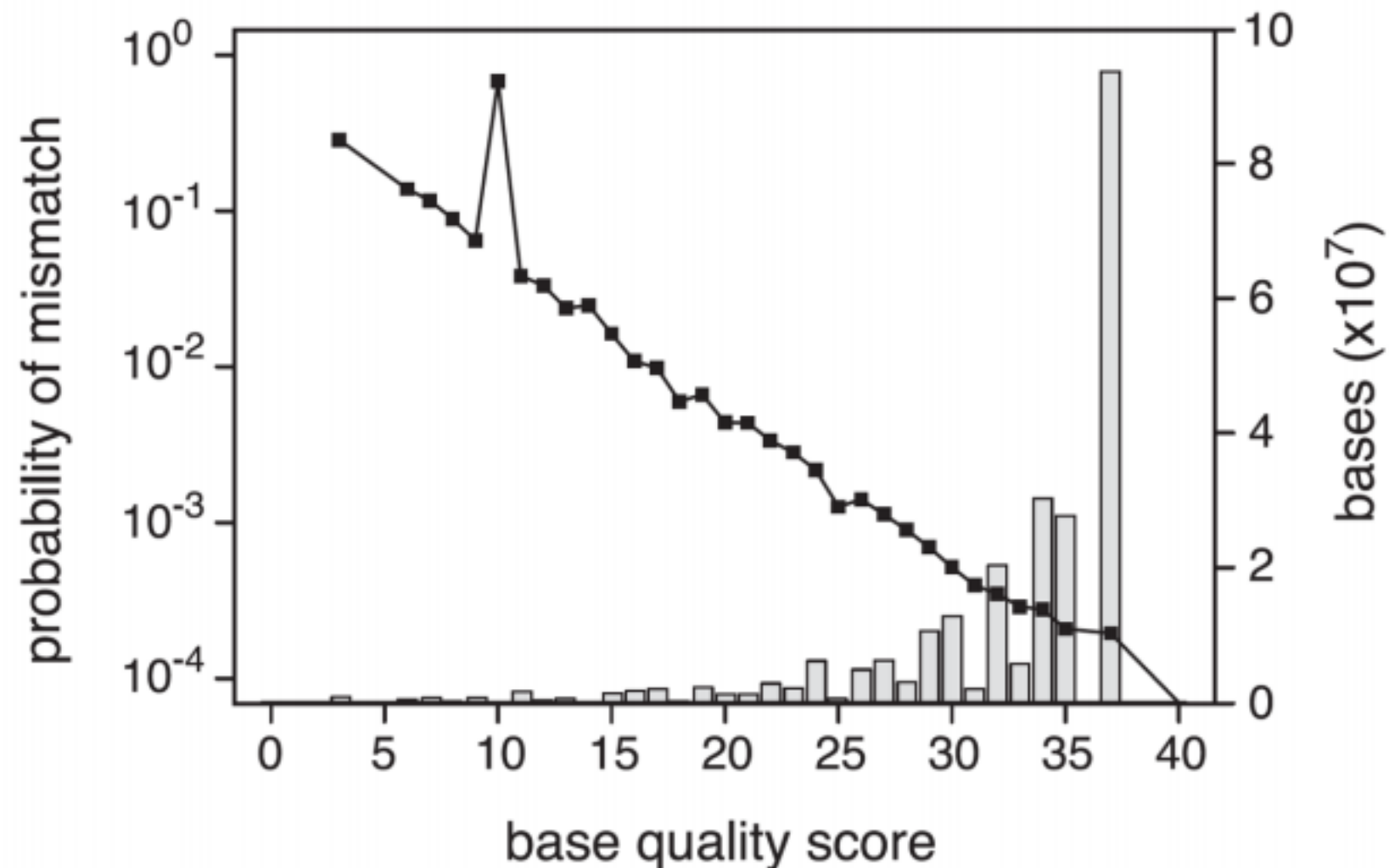
$$\text{Probability of Error} = 10^{-Q/10}$$

(This is a **Phred** score, also used for other types of qualities.)

- \* Very low quality scores can mean something special –  
Illumina  $Q \leq 3$  means something like: "I'm lost, you might want to stop believing sequencing cycles from here on out."
- \* In older FASTQ files, the formula and ASCII offset might differ.  
Consult: [http://en.wikipedia.org/wiki/FASTQ\\_format](http://en.wikipedia.org/wiki/FASTQ_format)

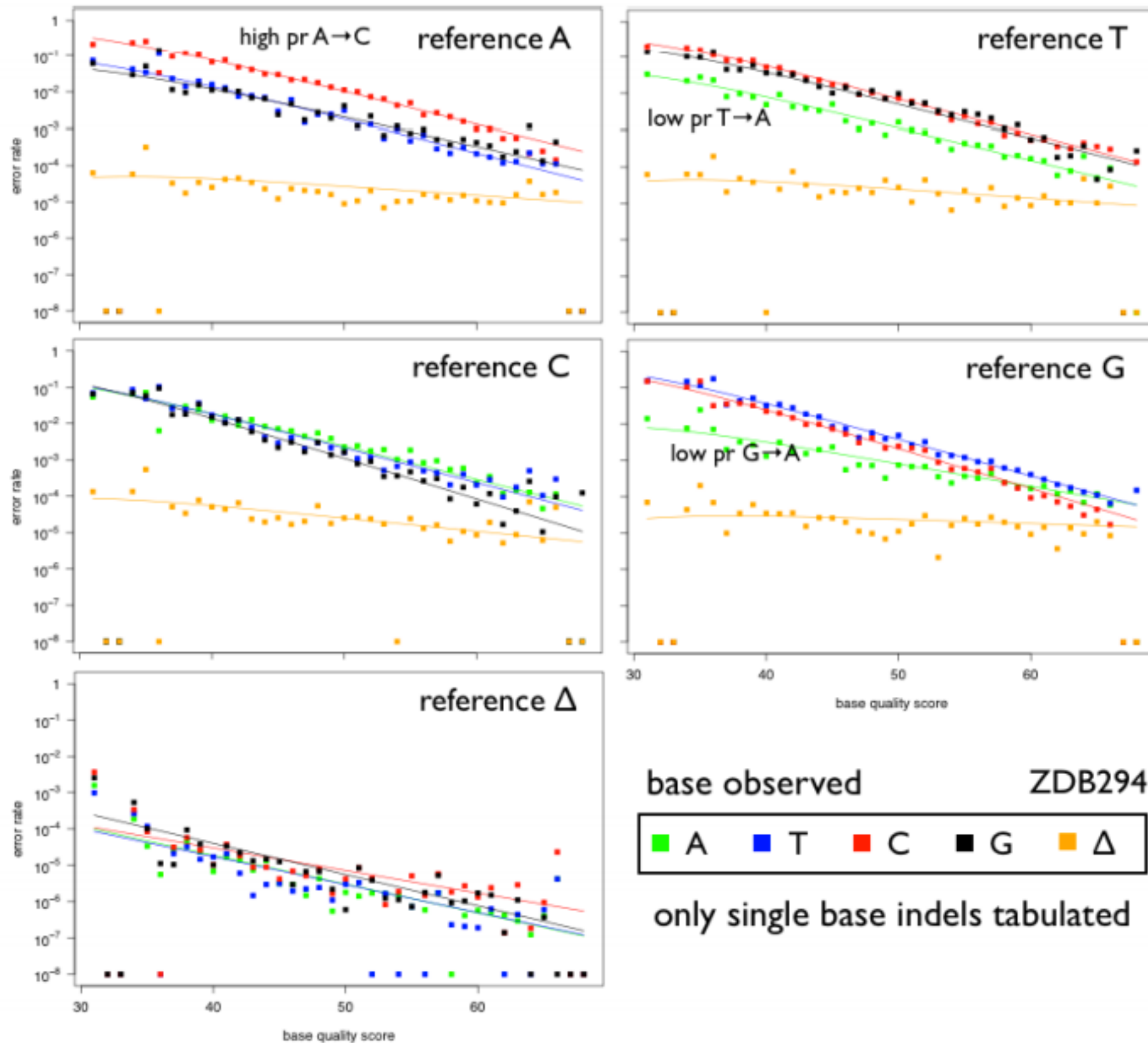


# Example of Illumina Data



- Most bases have high qualities ( $Q > 30$ ).
- Overall qualities are well calibrated\*.

# Example of Illumina Data



# FASTQC

## **Quality Assurance tool for FASTQ sequences**

FastQC website:

<http://www.bioinformatics.babraham.ac.uk>

FastQC report documentation:

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/>

Good Illumina dataset:

[http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good\\_sequence\\_short\\_fastqc/fastqc\\_report.html](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc/fastqc_report.html)

Bad Illumina dataset:

[http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad\\_sequence\\_fastqc/fastqc\\_report.html](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc/fastqc_report.html)

# FASTQC












- Report basic statistics on your data
- Identify issues with your data

## Basic Statistics

| Measure            | Value                   |
|--------------------|-------------------------|
| Filename           | tmp.fastq               |
| File type          | Conventional base calls |
| Encoding           | Illumina 1.5            |
| Total Sequences    | 250000                  |
| Filtered Sequences | 0                       |
| Sequence length    | 101                     |
| %GC                | 51                      |

## FastQC Report

### Summary

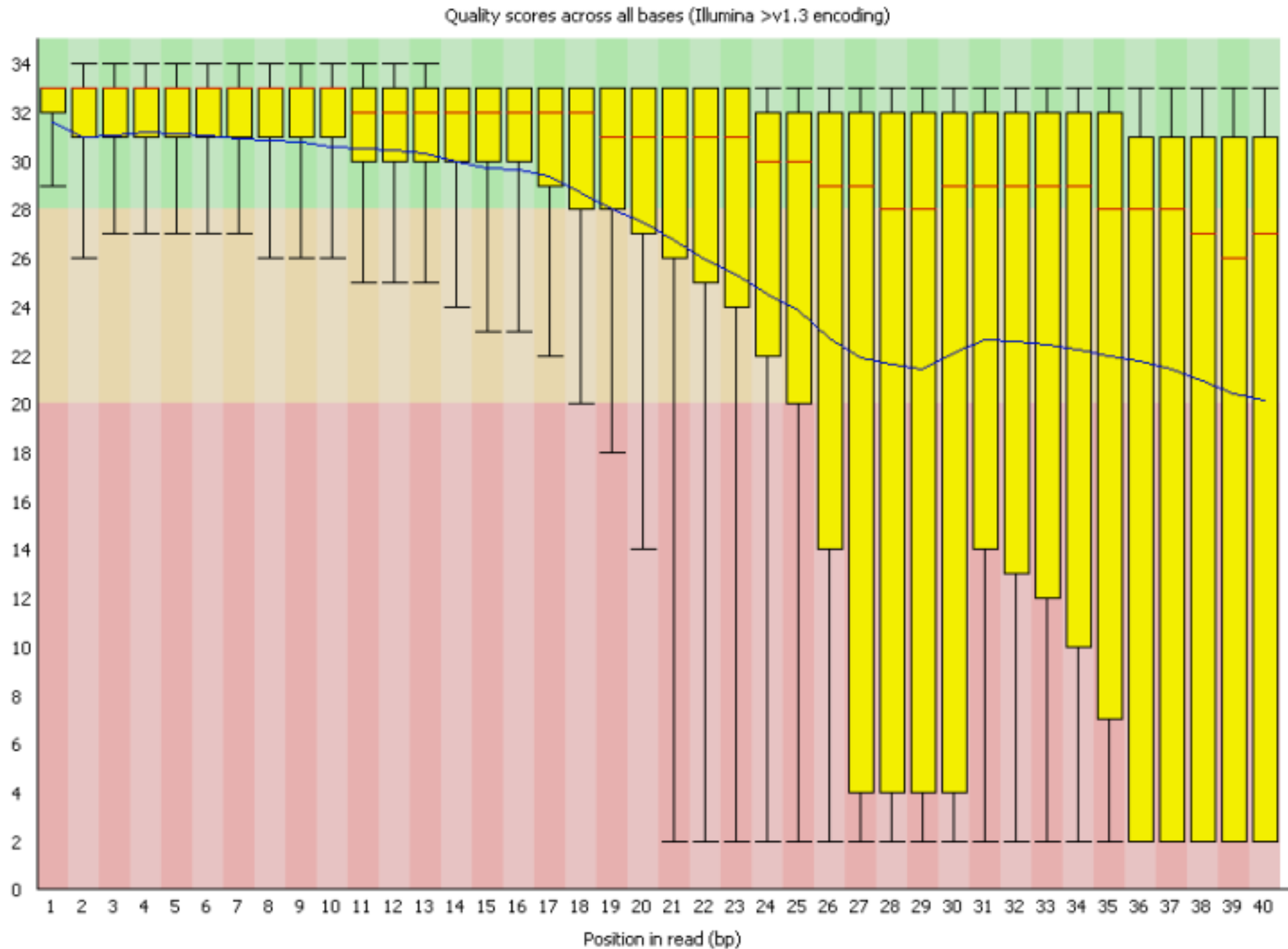
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Kmer Content](#)

# Useful reports

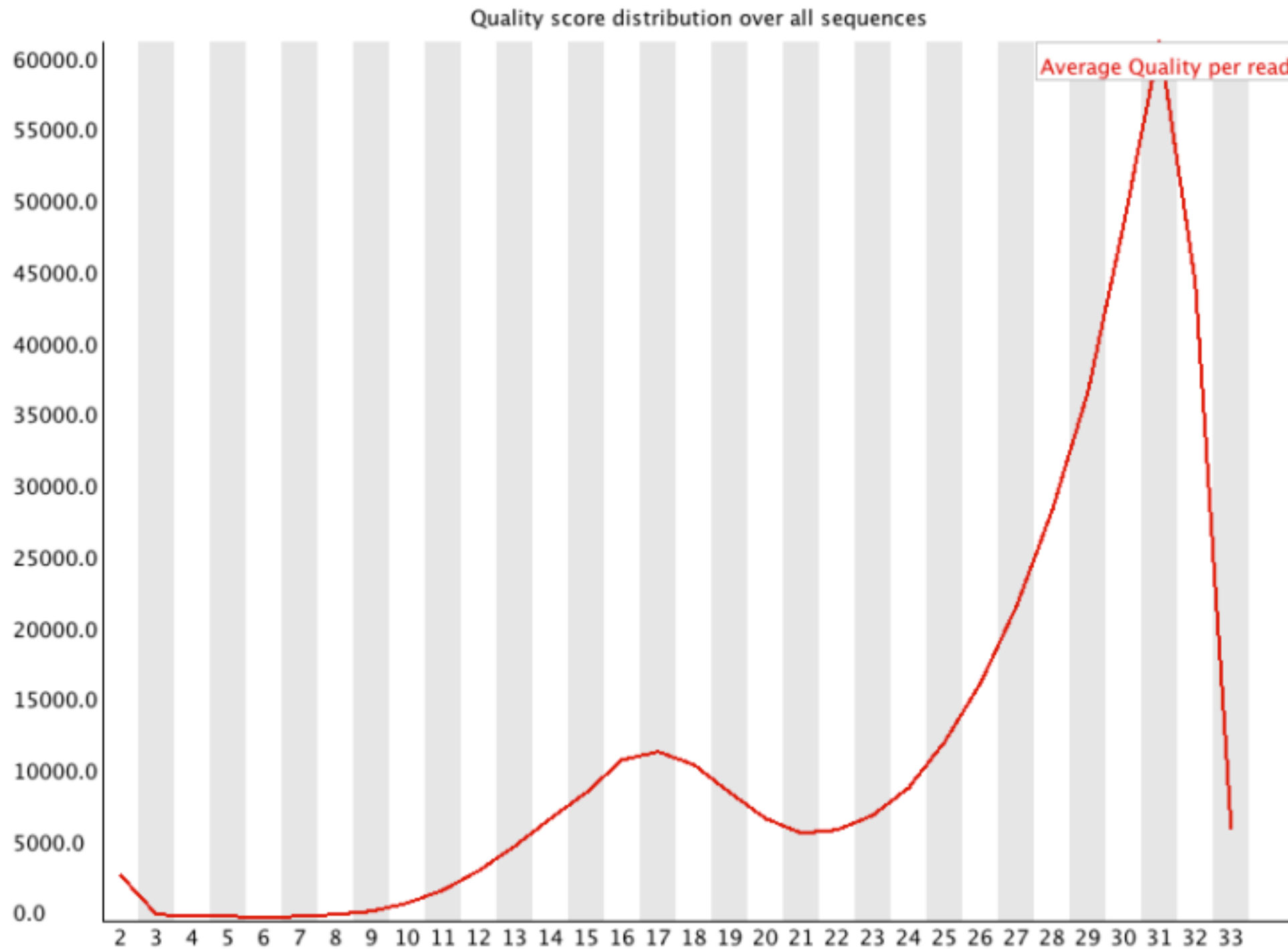
- Should I trim low quality bases?
  - Per-base sequence quality Report
    - based on *all* sequences
- Do I need to remove adapter sequences?
  - Overrepresented sequences Report
    - based on *1<sup>st</sup> 200,000* sequences
- How complex is my library?
  - Sequence duplication levels Report
    - estimate based on *1<sup>st</sup> 200,000* sequences



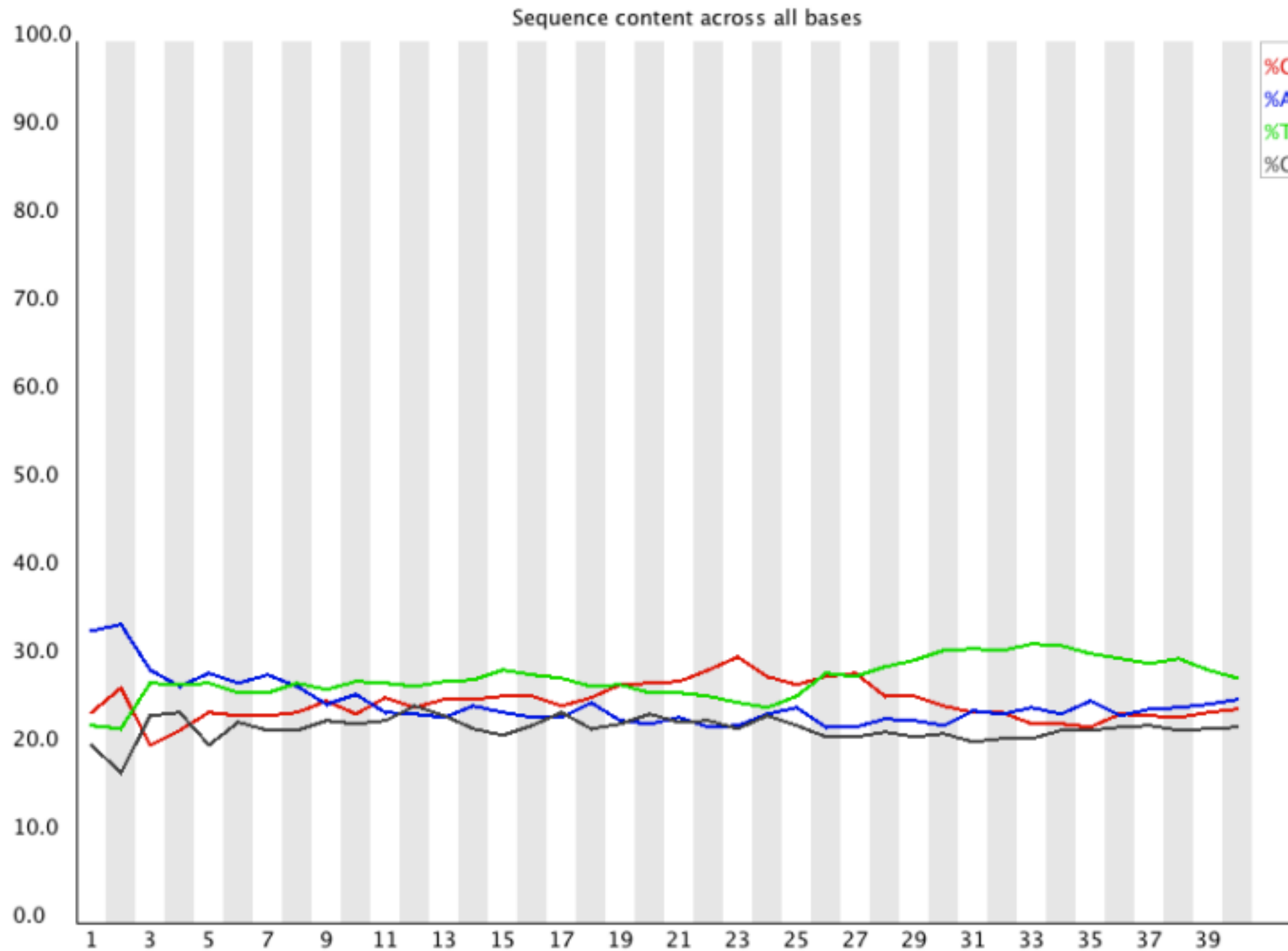
# Per Base Sequencing Quality



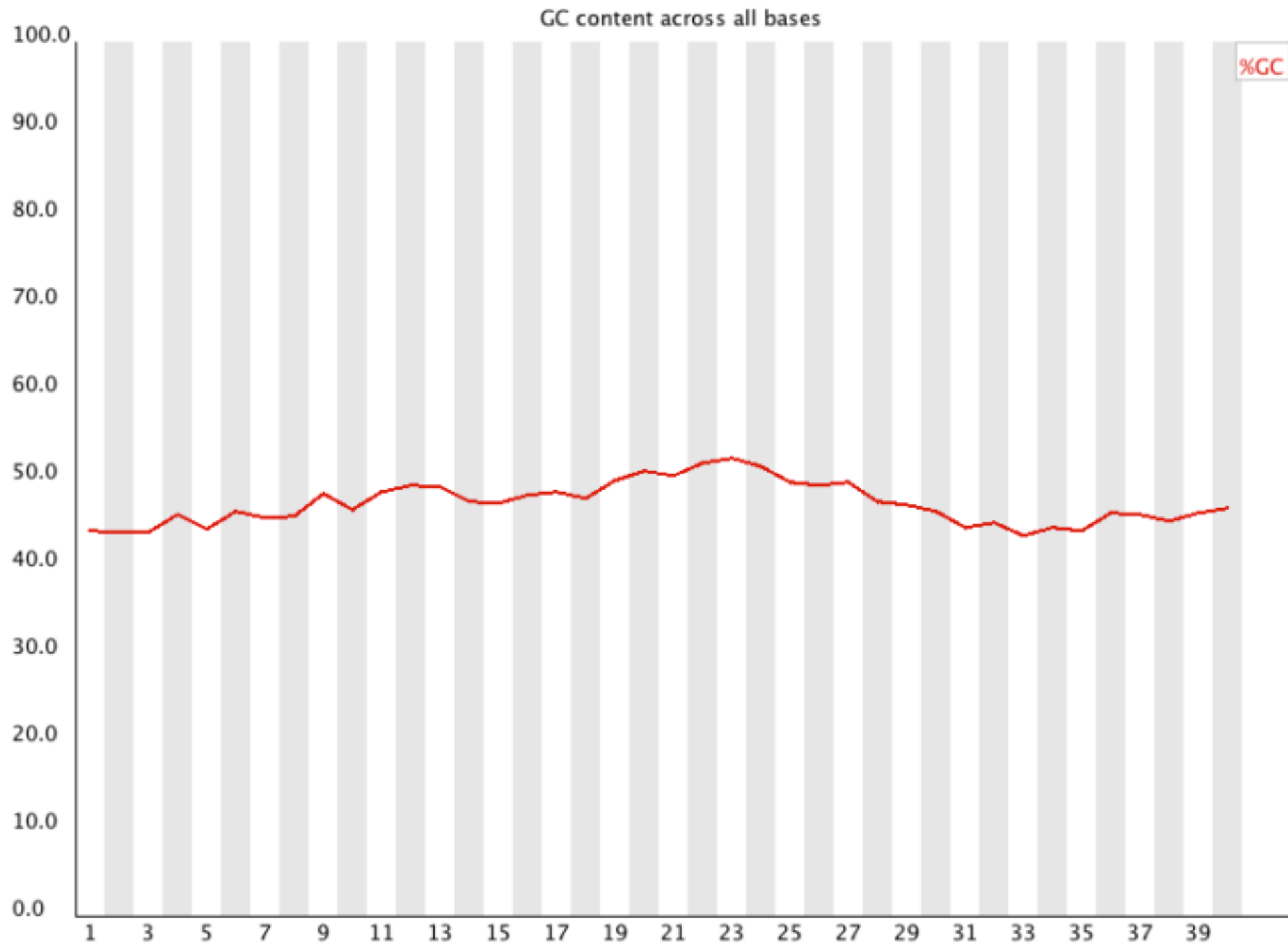
# Per Sequencing Quality



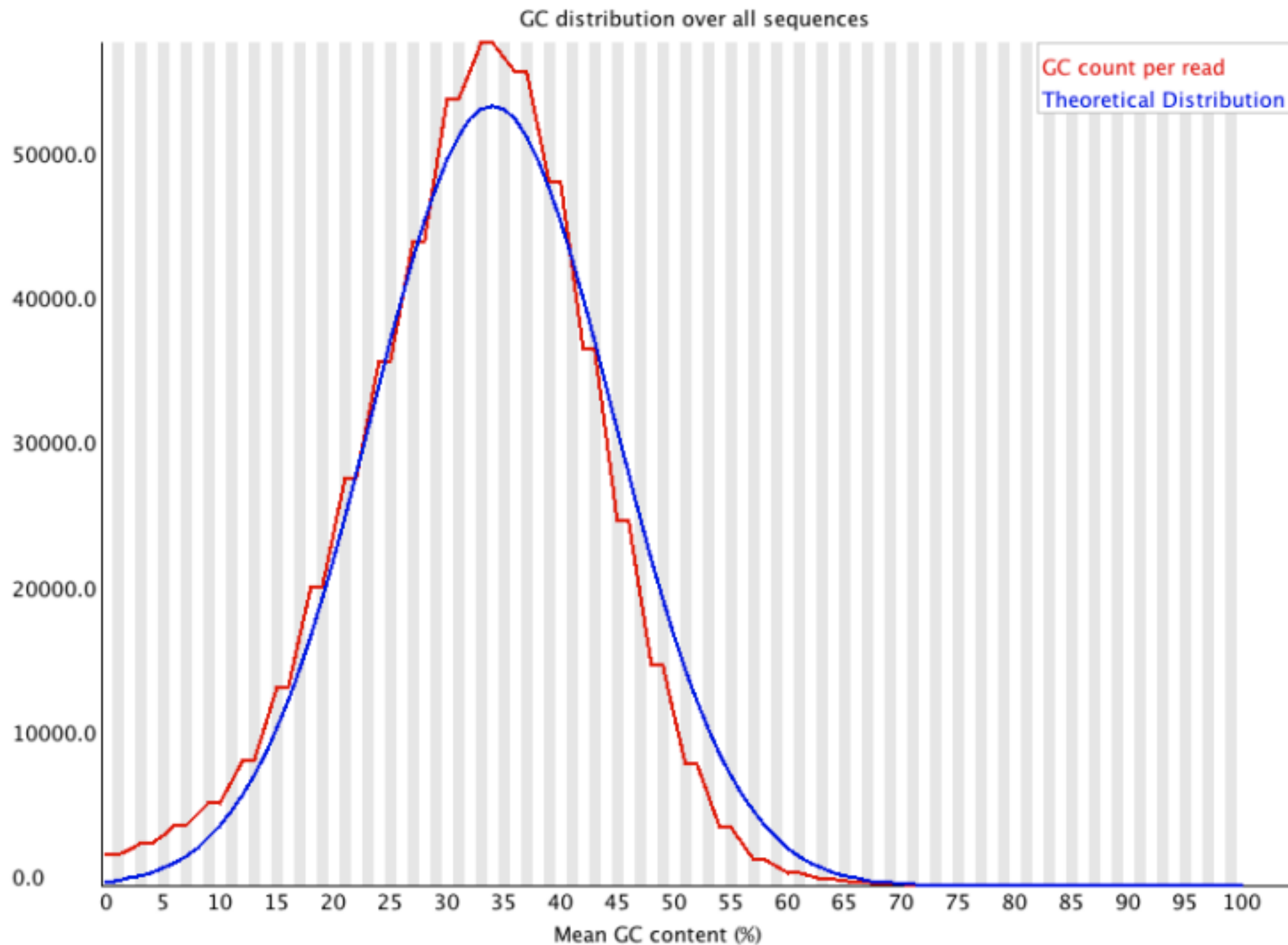
# Per Base Sequencing content



# Per Base GC Content

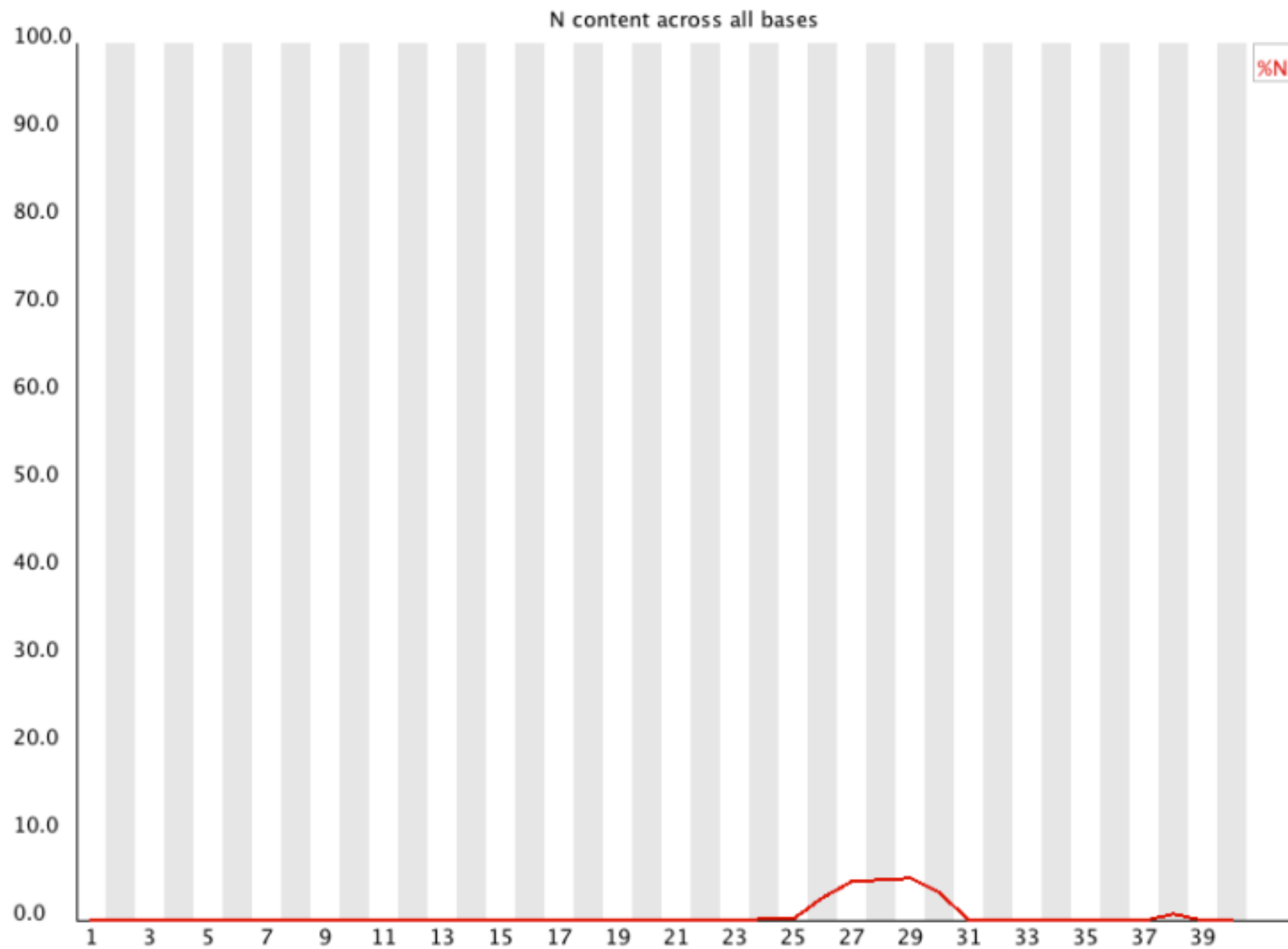


# Per Sequencing Nucleotide Content

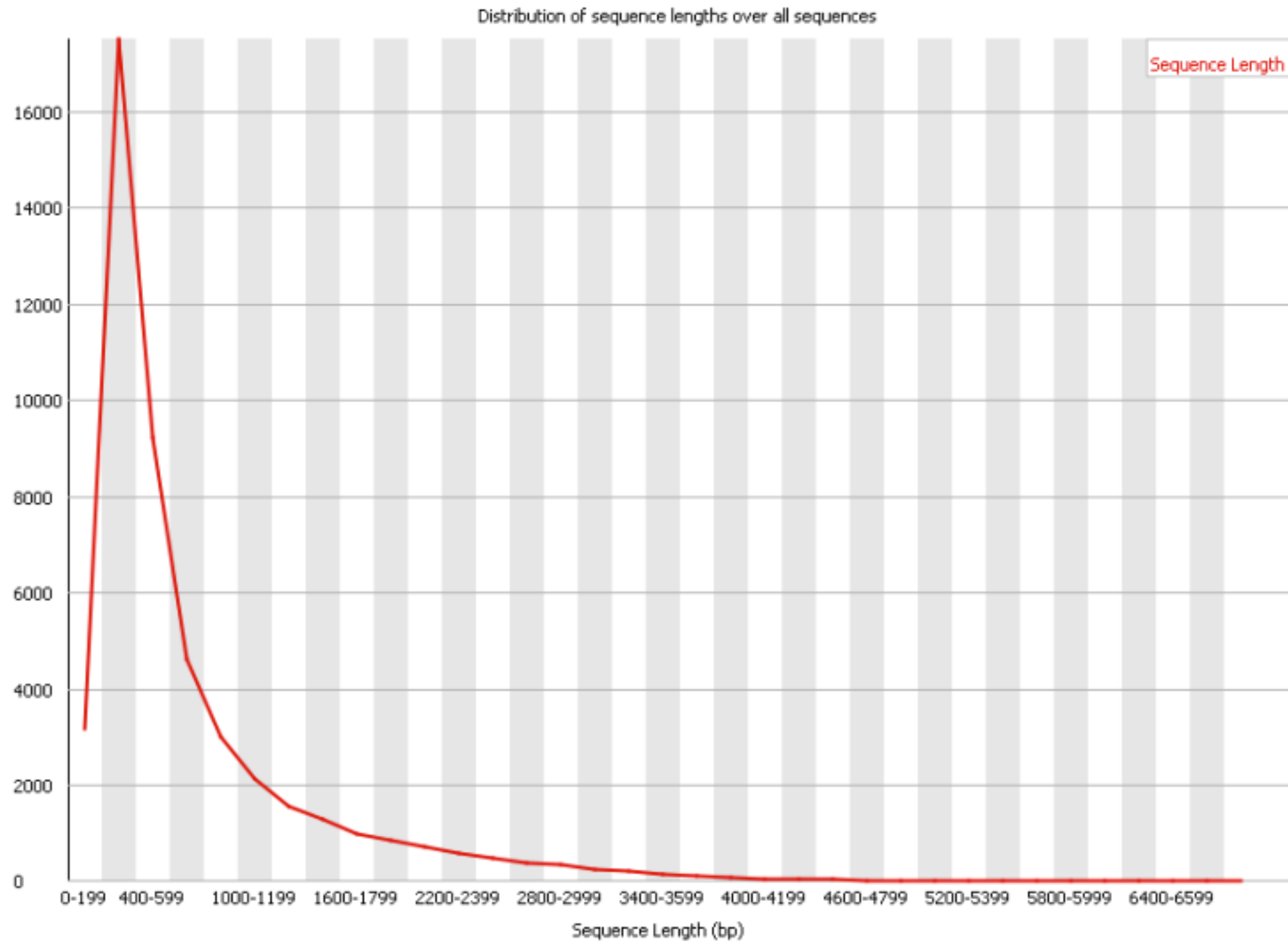




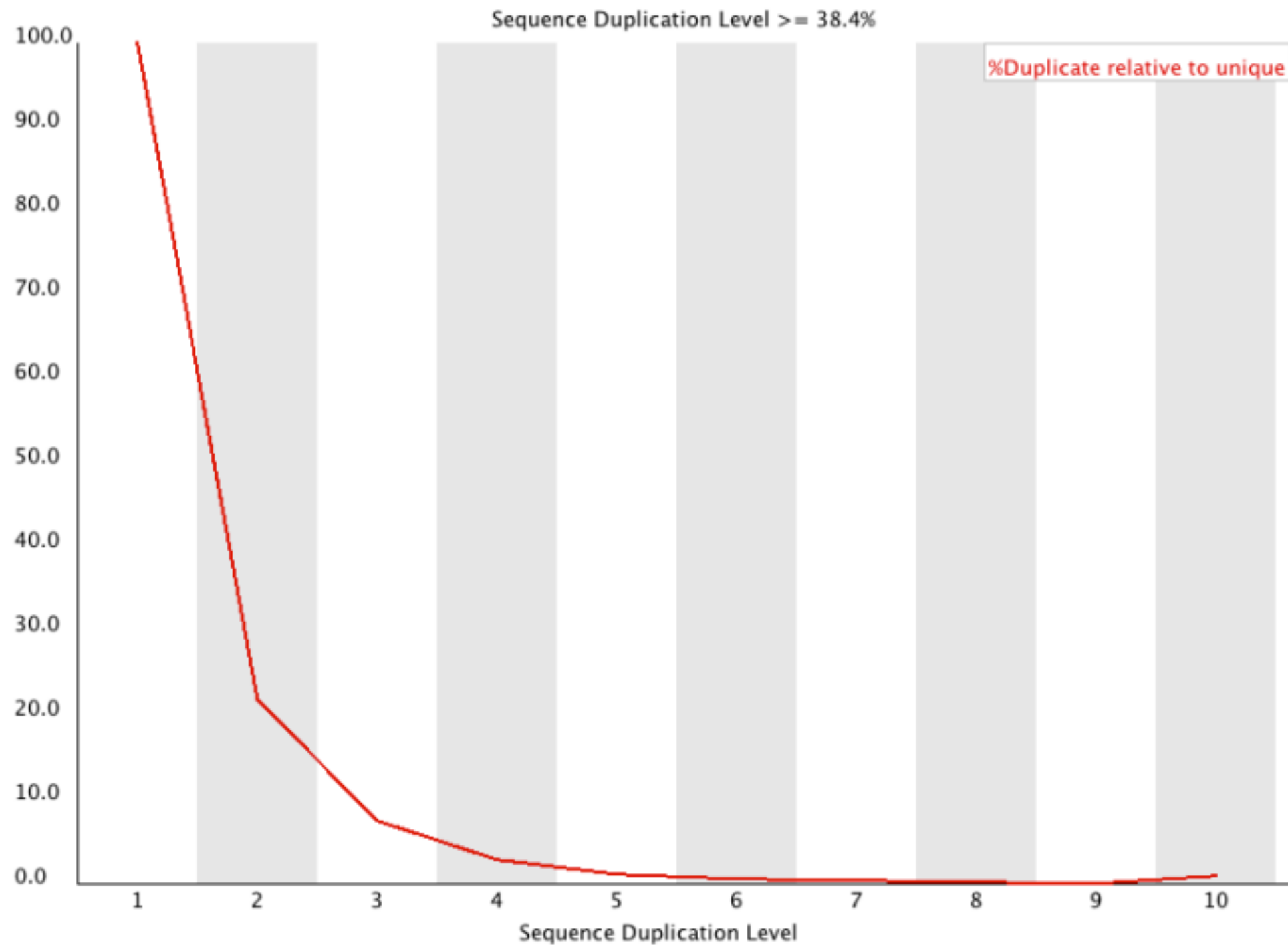
# Per Base N content



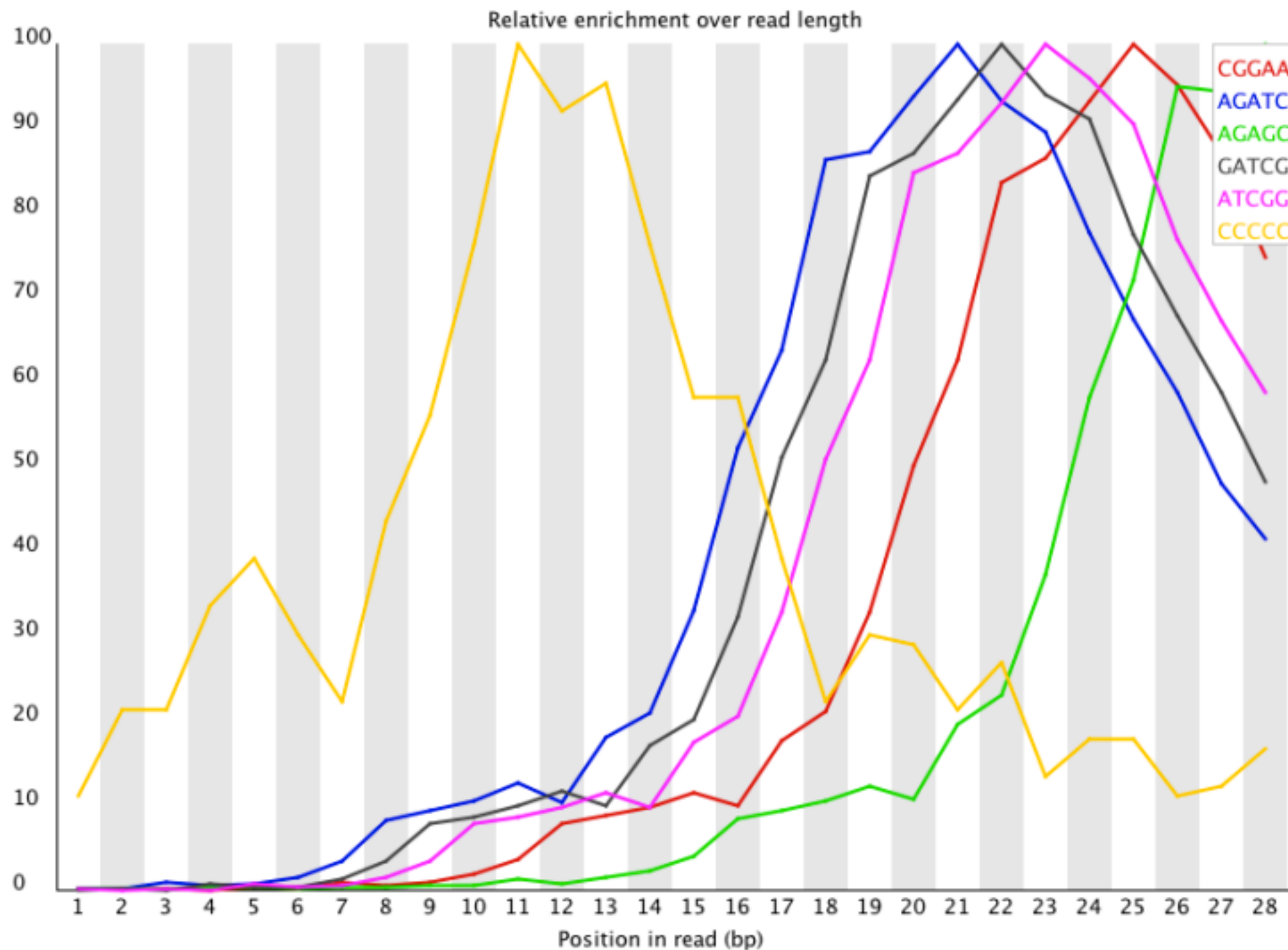
# Sequence Length Distribution



# Duplicate Sequences Distribution



# Overrepresented K-mers



# Overrepresented K-mers

- What is a  $k$ -mer?
- Create a sliding window of size  $k$ , move it over all your reads and count occurrence of  $k$ -mers
- We can use this to correct sequencing errors!

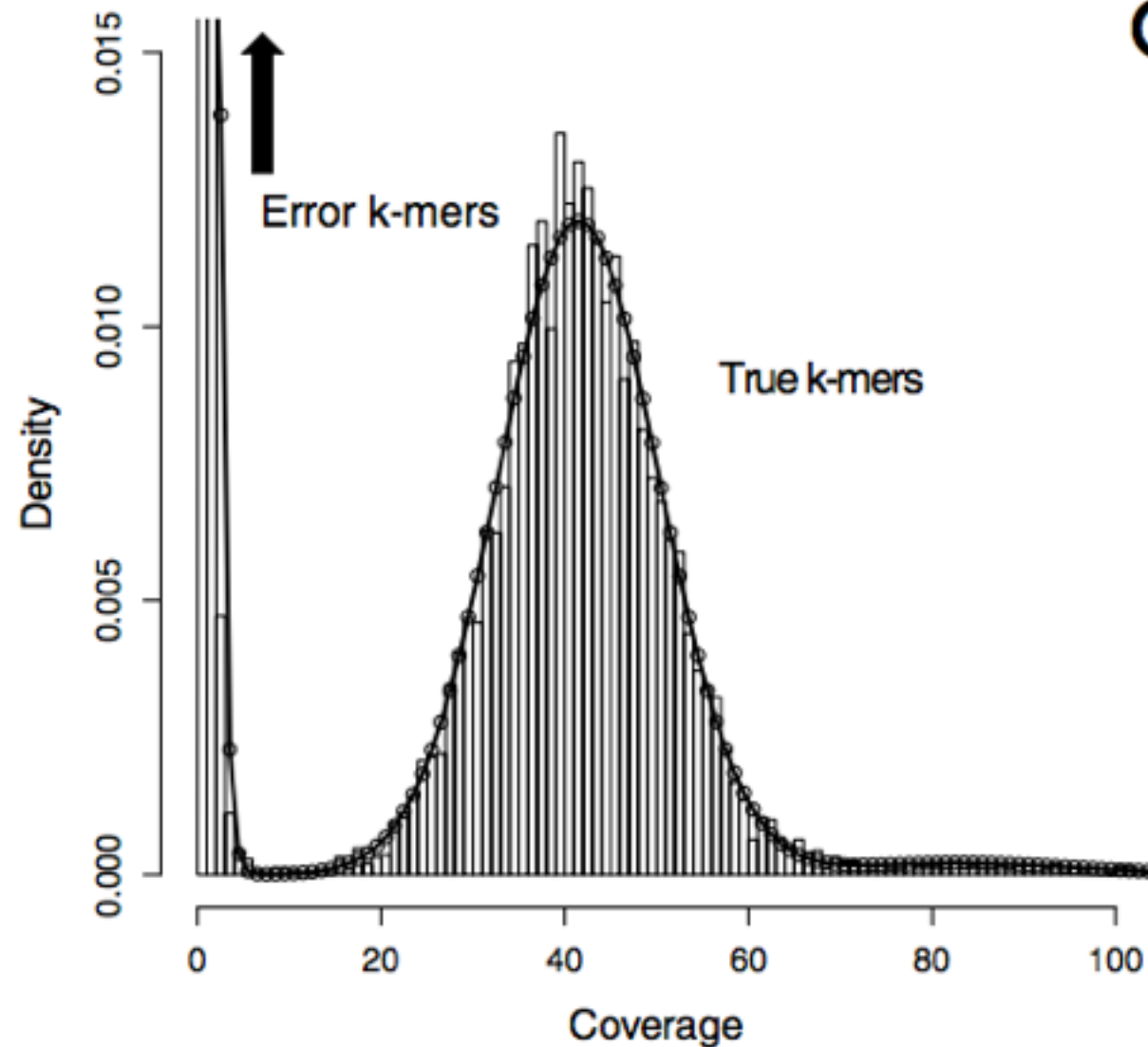
→  
DNA: ACGTGTAACGTGACGTTGGA  
ACGTG  
CGTGT  
GTGTA

$k=5$



# Overrepresented K-mers

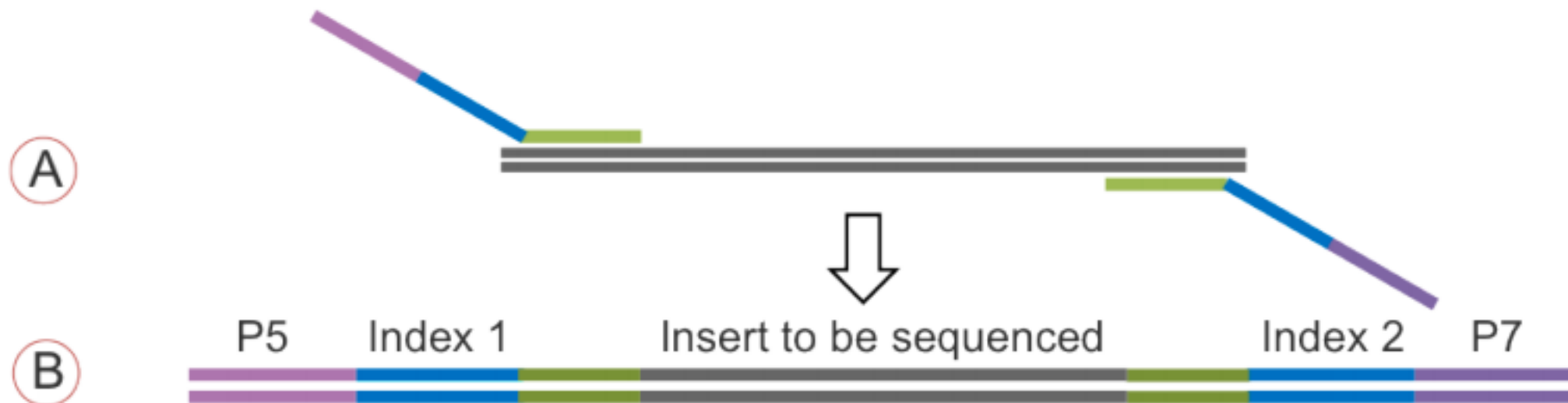
Concept: Rare *k*-mers are seq. errors  
Need >15X coverage



```
ACGTGGTTGCCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
```

# Sequencing Process: PCR primers

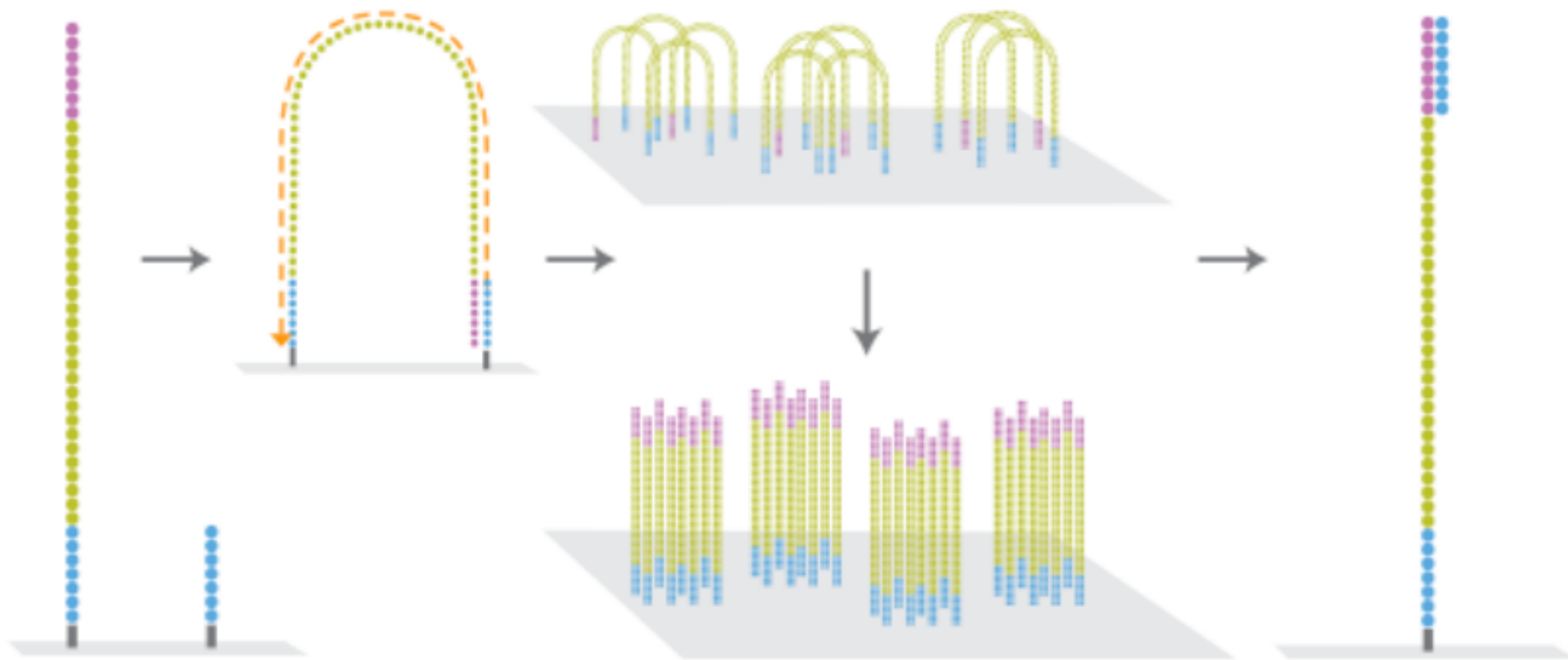
## One-step PCR Method



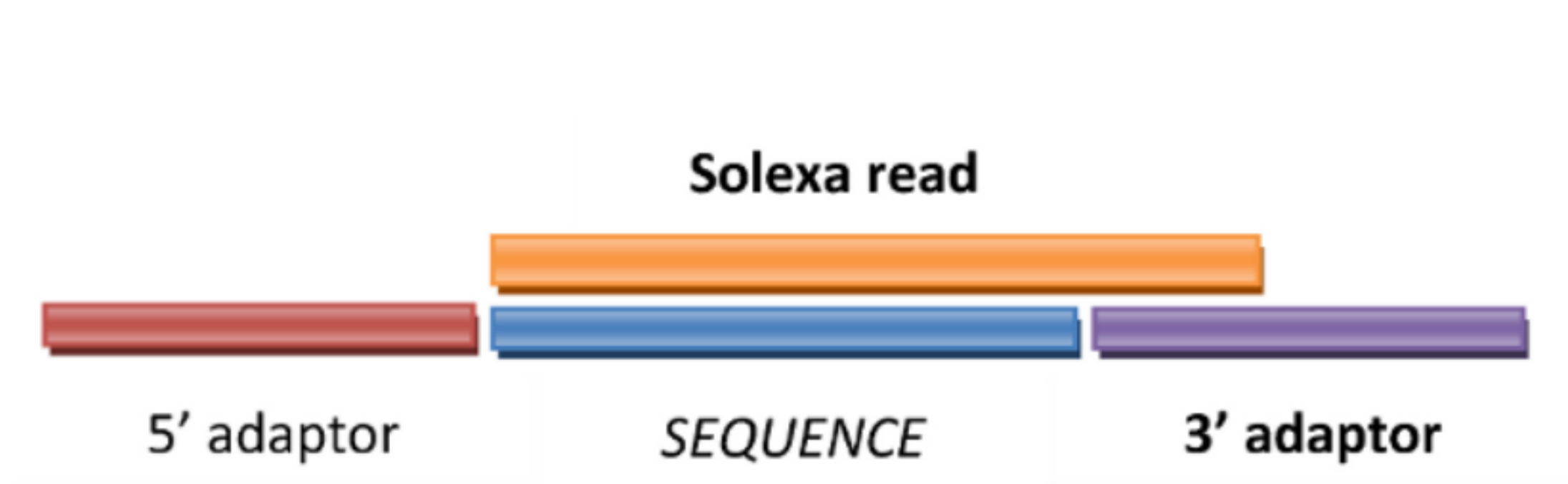
A Target-specific PCR with indices and sequencing adaptors

B Final amplicon ready to be sequenced

# PCR primers



# NGS adaptors & Cutadapt



# NGS adaptors & Cutadapt

## Overrepresented sequences

| Sequence  | Count   | Percentage          | Possible Source                             |
|---|---------|---------------------|---|
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATATCGTATGC  | 1547768 | 38.192098035156306  | TruSeq Adapter, Index 1<br>(98% over 50bp)  |
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGC  | 146635  | 3.61830603513262    | TruSeq Adapter, Index 1<br>(100% over 50bp) |
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCAAGATATCGTATGC  | 6639    | 0.16382128255358863 | TruSeq Adapter, Index 1<br>(97% over 41bp)  |
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATTTTCGTATGC | 6462    | 0.15945370204267054 | TruSeq Adapter, Index 1<br>(98% over 50bp)  |
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACATTACGATATCGTATGC  | 5433    | 0.1340625136486891  | TruSeq Adapter, Index 1<br>(97% over 41bp)  |
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACATAACGATATCGTATGC  | 5147    | 0.1270052931621209  | TruSeq Adapter, Index 1<br>(97% over 41bp)  |
| GATCGGAAGAGCACACGTCTGAACTCCAGTCACACCACGATATCGTATGC  | 4703    | 0.11604932849066535 | TruSeq Adapter, Index 1<br>(97% over 41bp)  |

Very important if your DNA fragment is shorter than read length



# Coverage

- Coverage/depth is how many times that your data covers the genome (on average)
- Example:
  - N: Number of reads: 5 mill
  - L: Read length: 100
  - G: Genome size: 5 Mbases
  - $C = 5 * 100 / 5 = 100X$
  - On average there are 100 reads covering each position in the genome

$$C = N \times \frac{L}{G}$$



---

## **Pré-processando os FASTQ's**

Marcel Caraciolo, CTO  
[marcel@genomika.com.br](mailto:marcel@genomika.com.br)