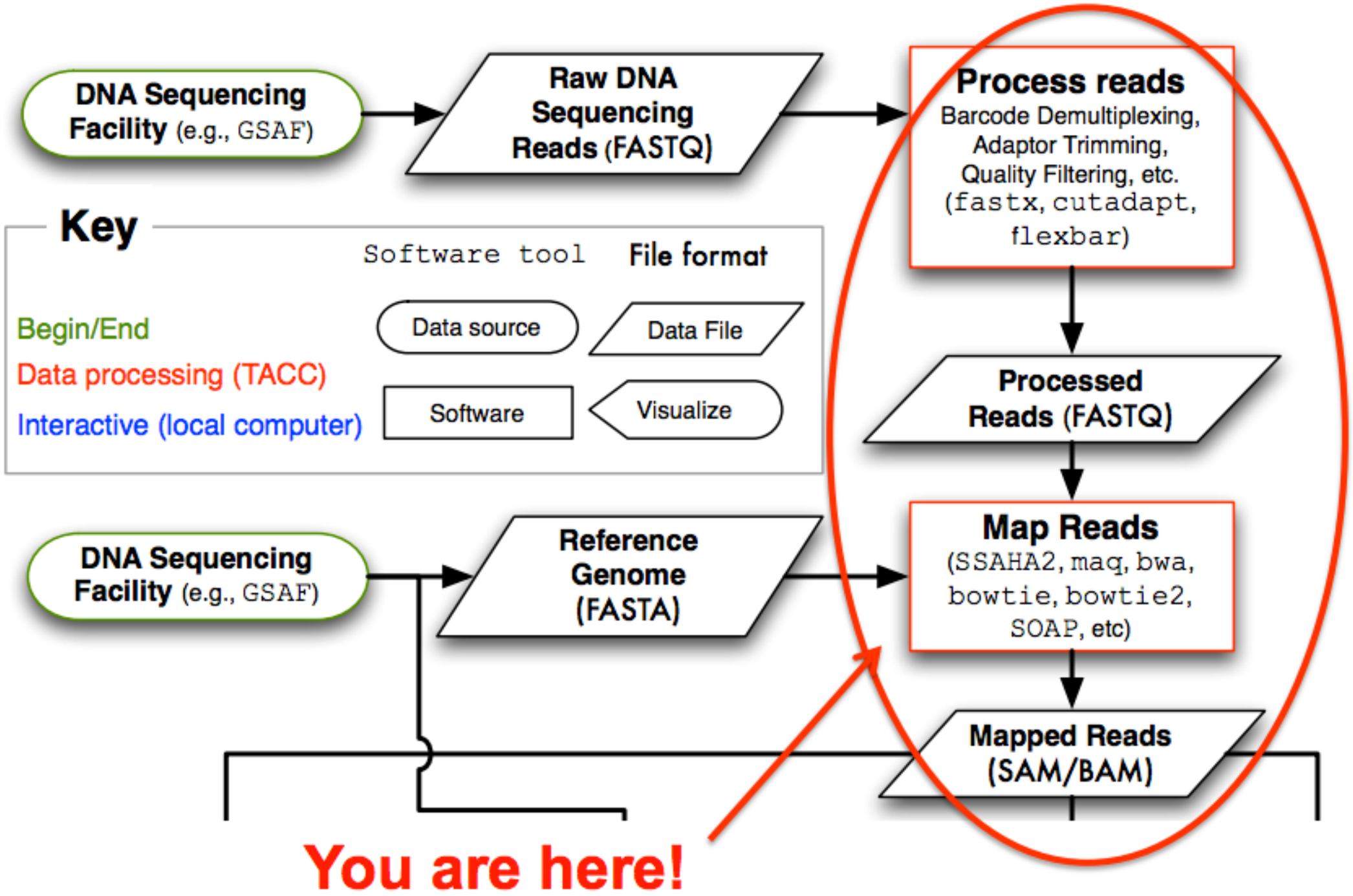
GENOMIKA

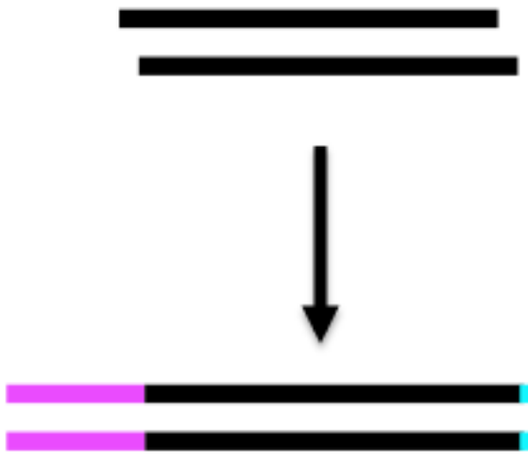
Introdução ao NGS

Rodrigo Bertollo
rodrigo@genomika.com.br

Pipeline


DNA Sequencing Facility (e.g., GSAF)
Raw DNA Sequencing Reads (FASTQ)
Process reads
Barcode Demultiplexing, Adaptor Trimming, Quality Filtering, etc. (fastx, cutadapt, flexbar)
Key
Software tool
File format
Begin/End
Data processing (TACC)
Interactive (local computer)
Data source
Data File
Software
Visualize
Processed Reads (FASTQ)
DNA Sequencing Facility (e.g., GSAF)
Reference Genome (FASTA)
Map Reads
(SSAHA2, maq, bwa, bowtie, bowtie2, SOAP, etc)
Mapped Reads (SAM/BAM)
You are here!

Terminology

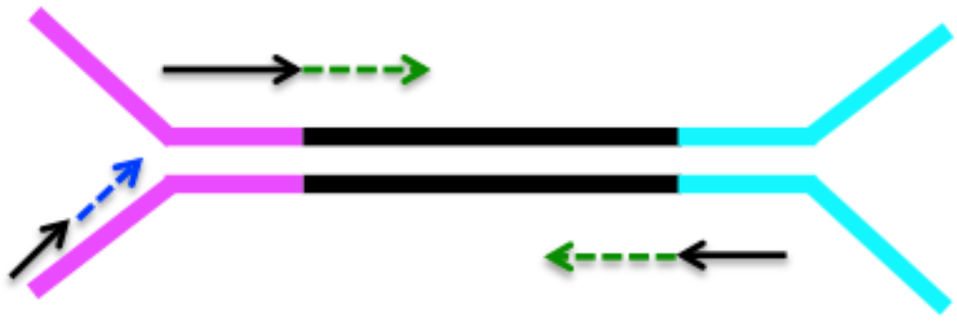


Fragment library (input DNA sample)

Library prep

Sequencing library
Double-stranded or Y- **adaptors** added

DNA sequencing



Barcode (6–12 bases) – so many samples can be run in one lane.
Data is **demultiplexed**.

Primers

Reads (36–1000+ bases)

Types of Illumina fragment libraries

single-end

independent reads

paired-end

two inwardly oriented reads separated by ~200 nt

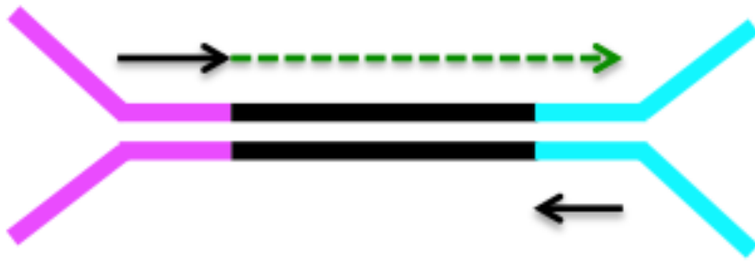
mate-paired

two outwardly oriented reads separated by ~3000 nt

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@BBBBBB@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

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 ≤ 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

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

Bad Illumina dataset:
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

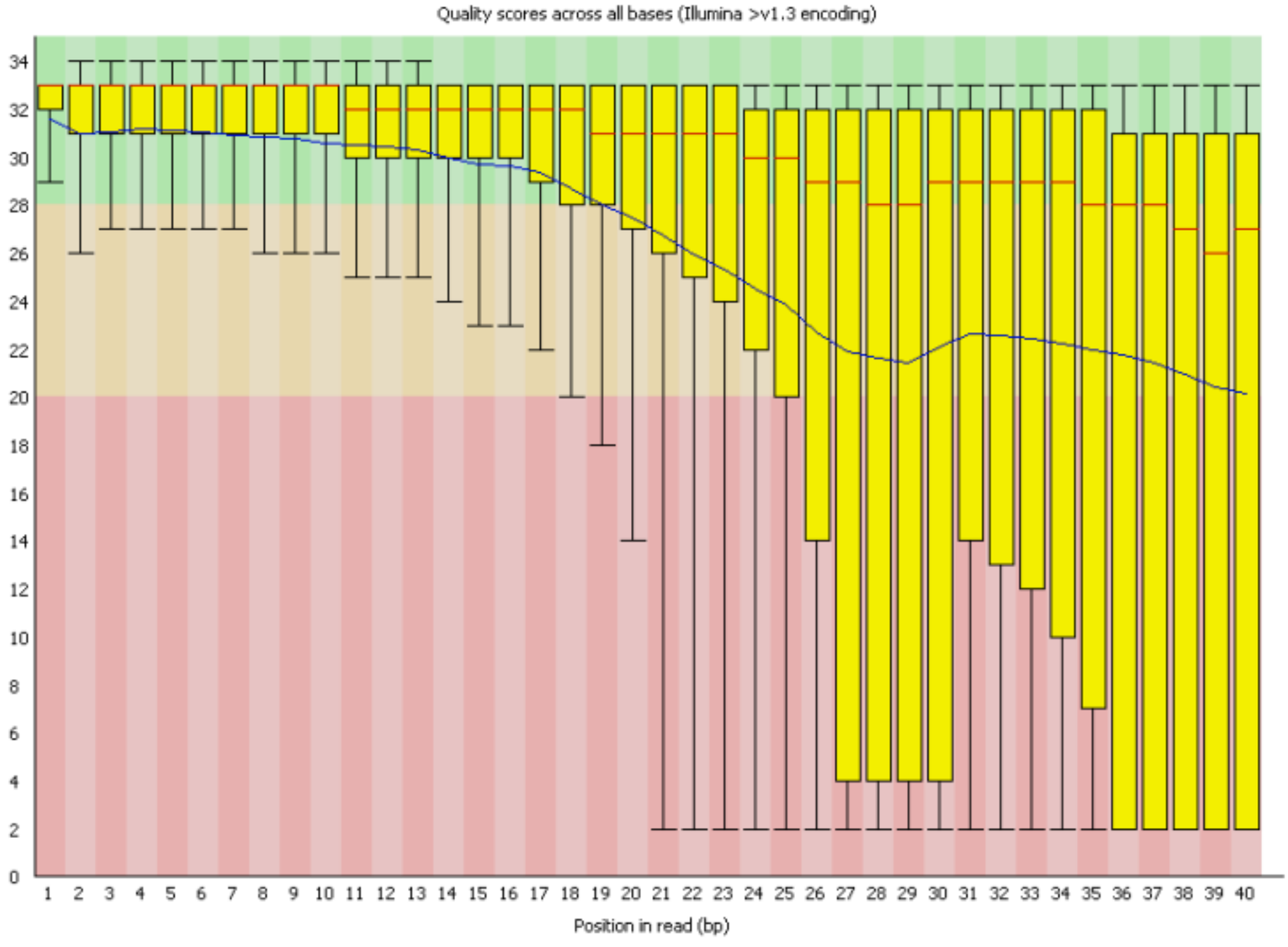
Summary

- 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

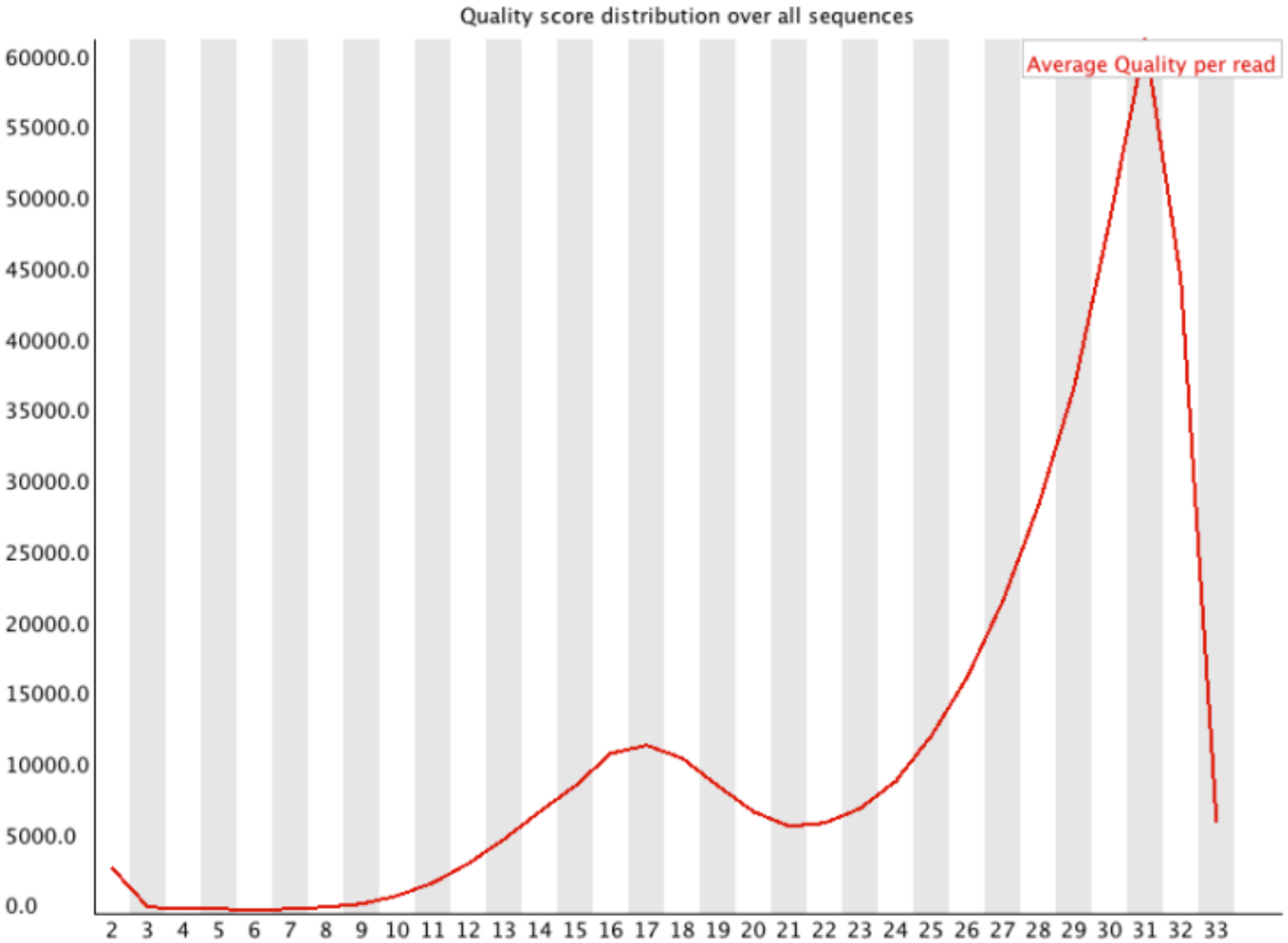
- Should I trim low quality bases?
  - Per-base sequence quality Report
    - based on *all* sequences
- Do I need to remove adapter sequences?
  - Overrepresented sequences Report
    - based on *1st 200,000* sequences
- How complex is my library?
  - Sequence duplication levels Report
    - estimate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

Per Base Sequencing Quality



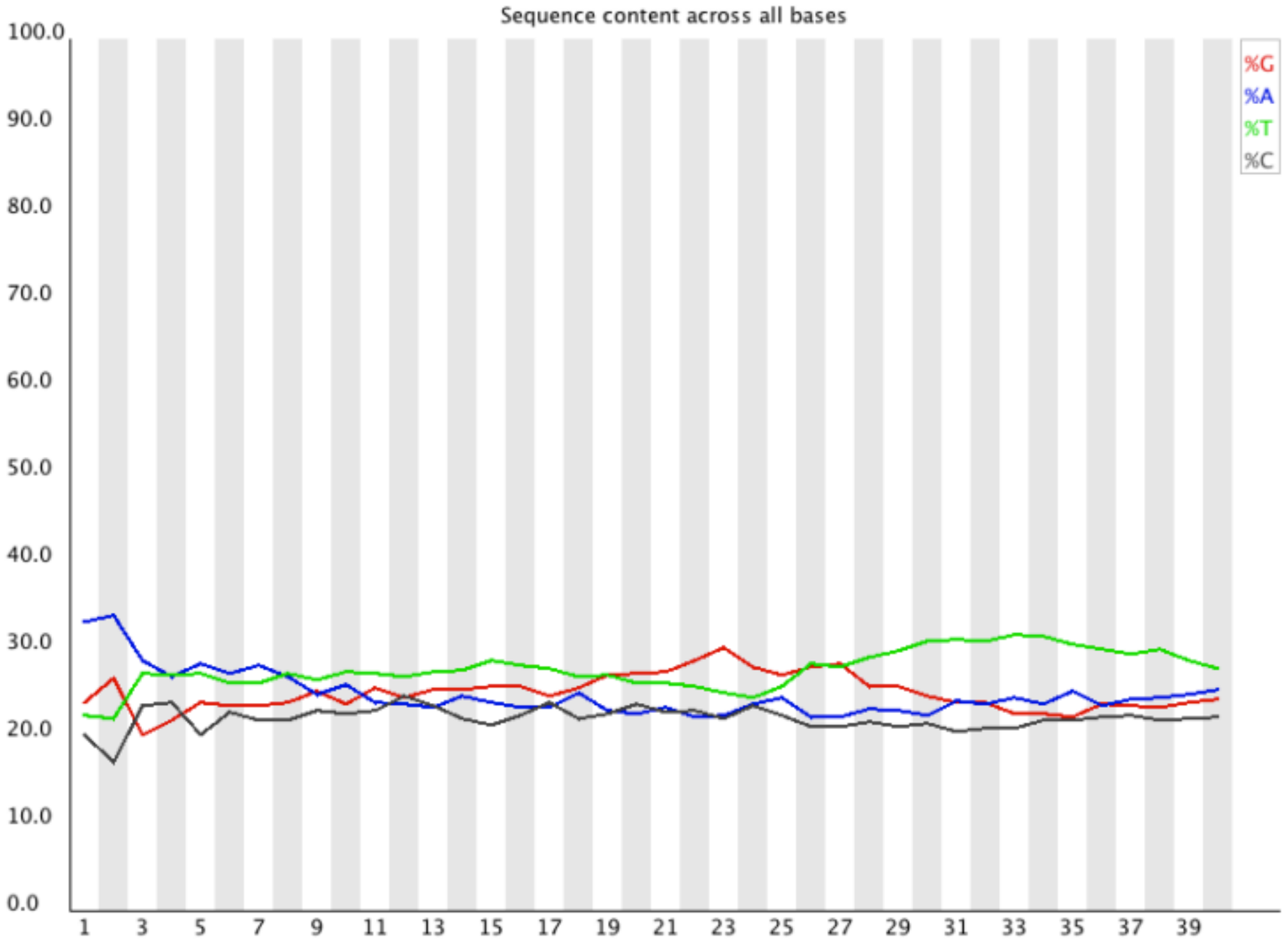
GENOMIKA

Per Sequencing Quality

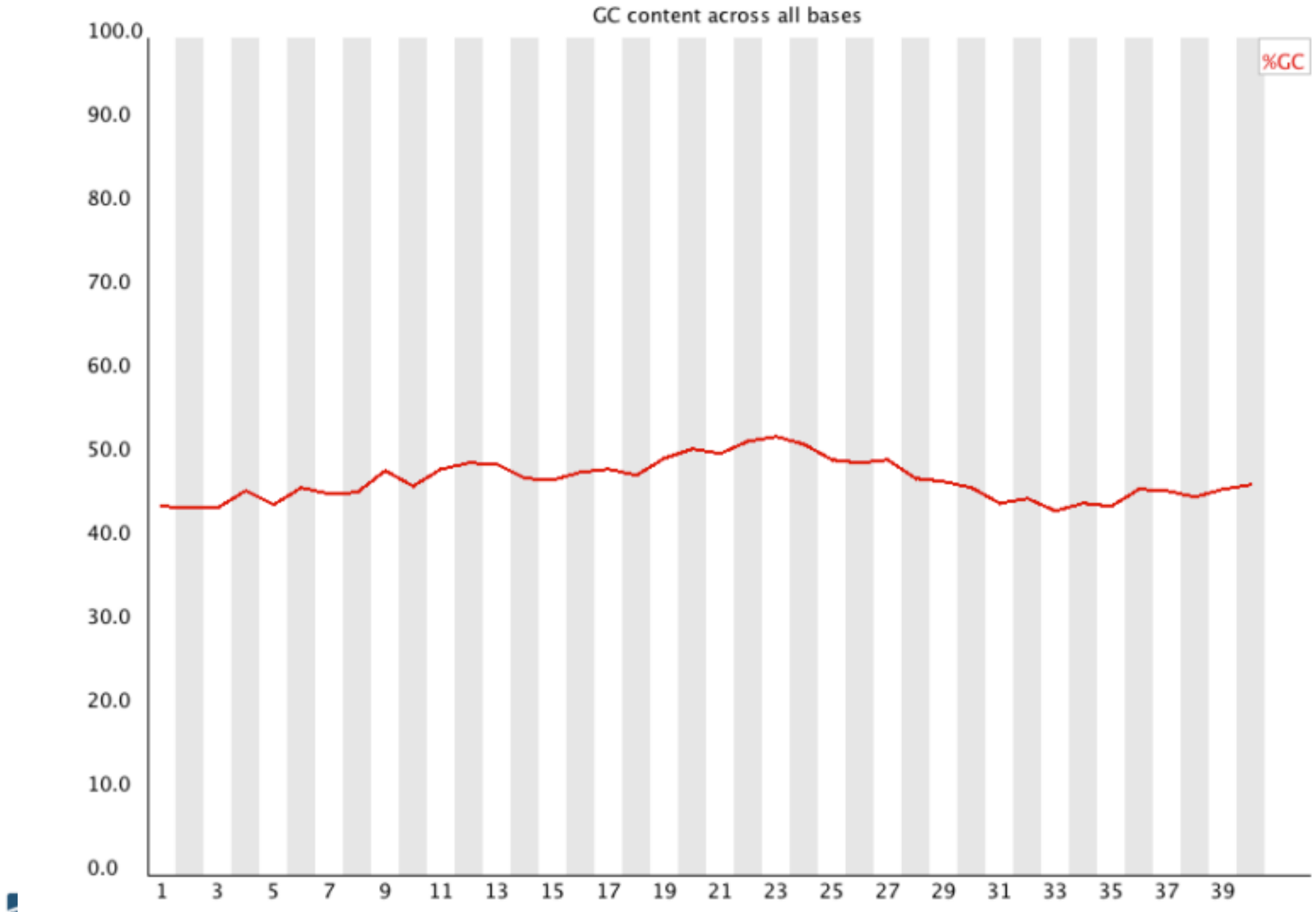


GENOMIKA

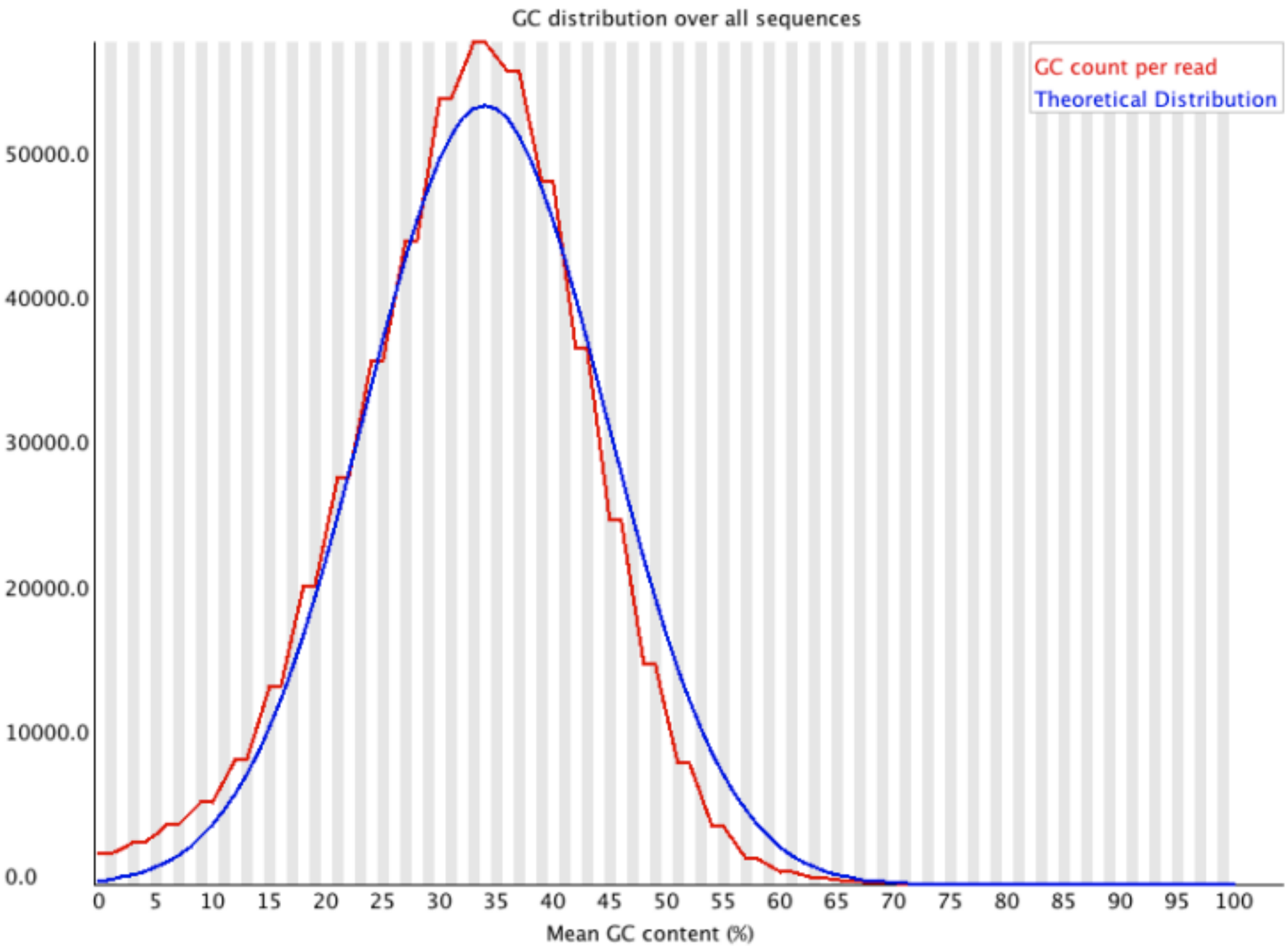
Per Base Sequencing content



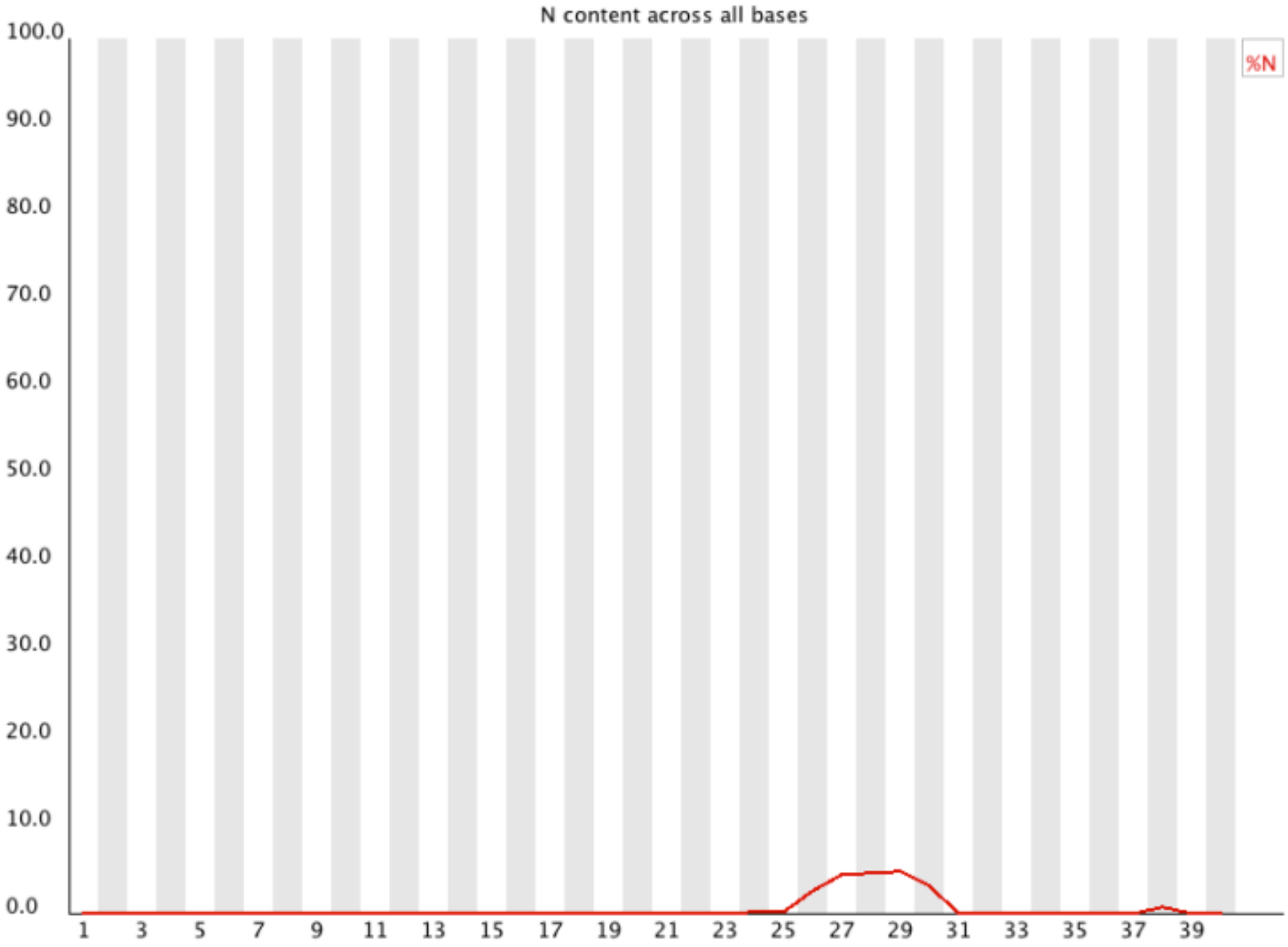
Per Base GC Content



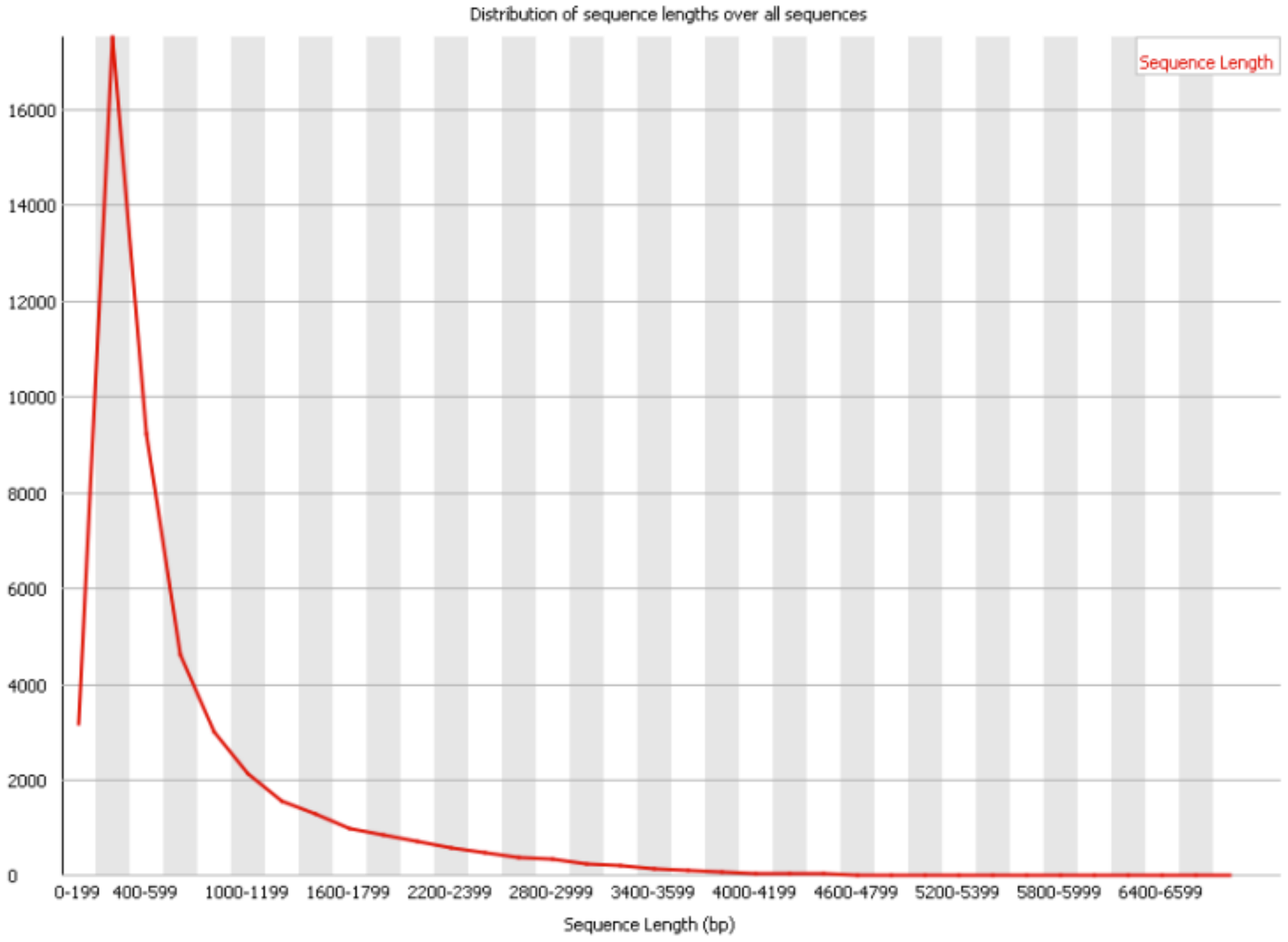
Per Sequencing Nucleotide Content



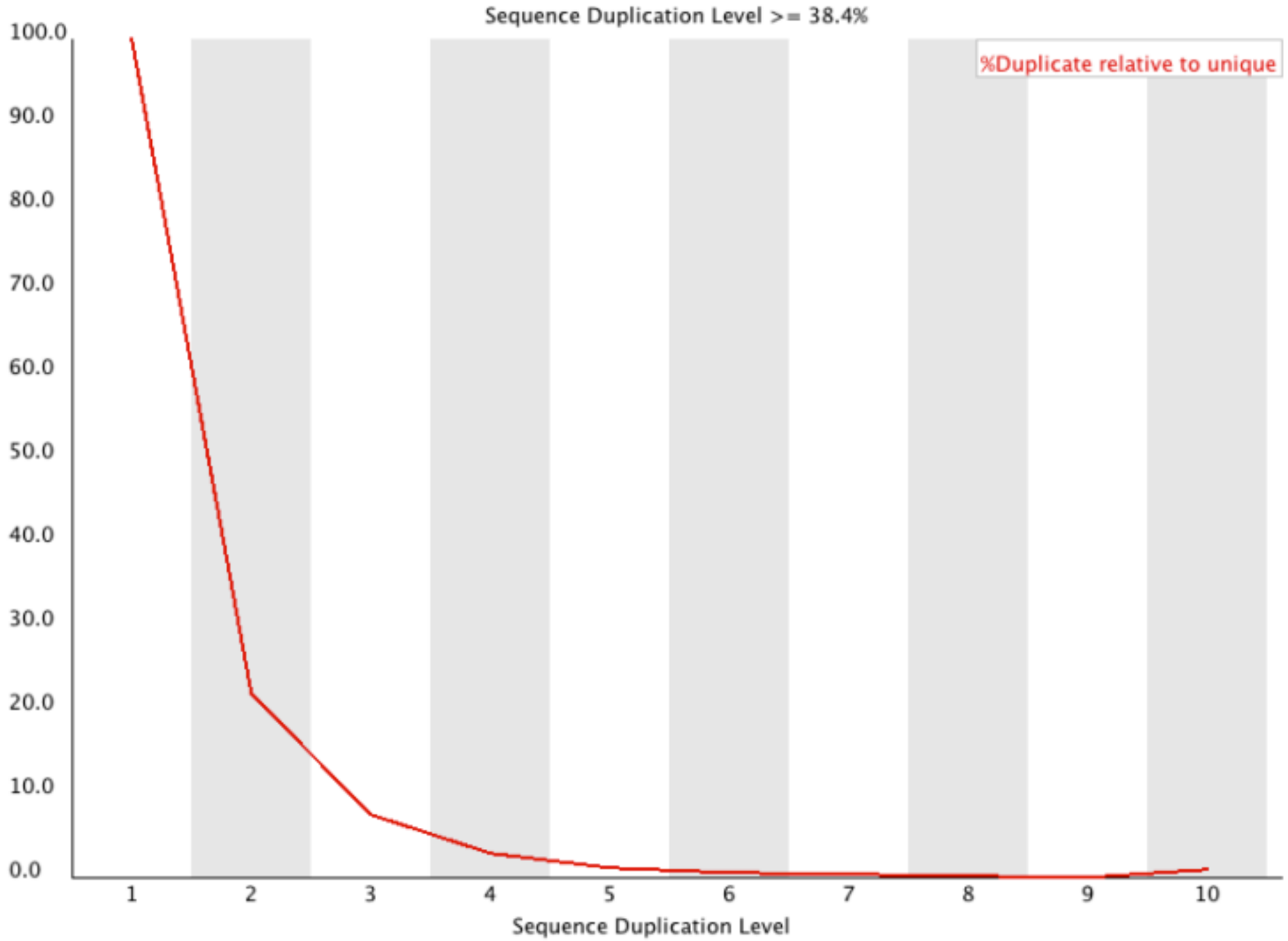
Per Base N content



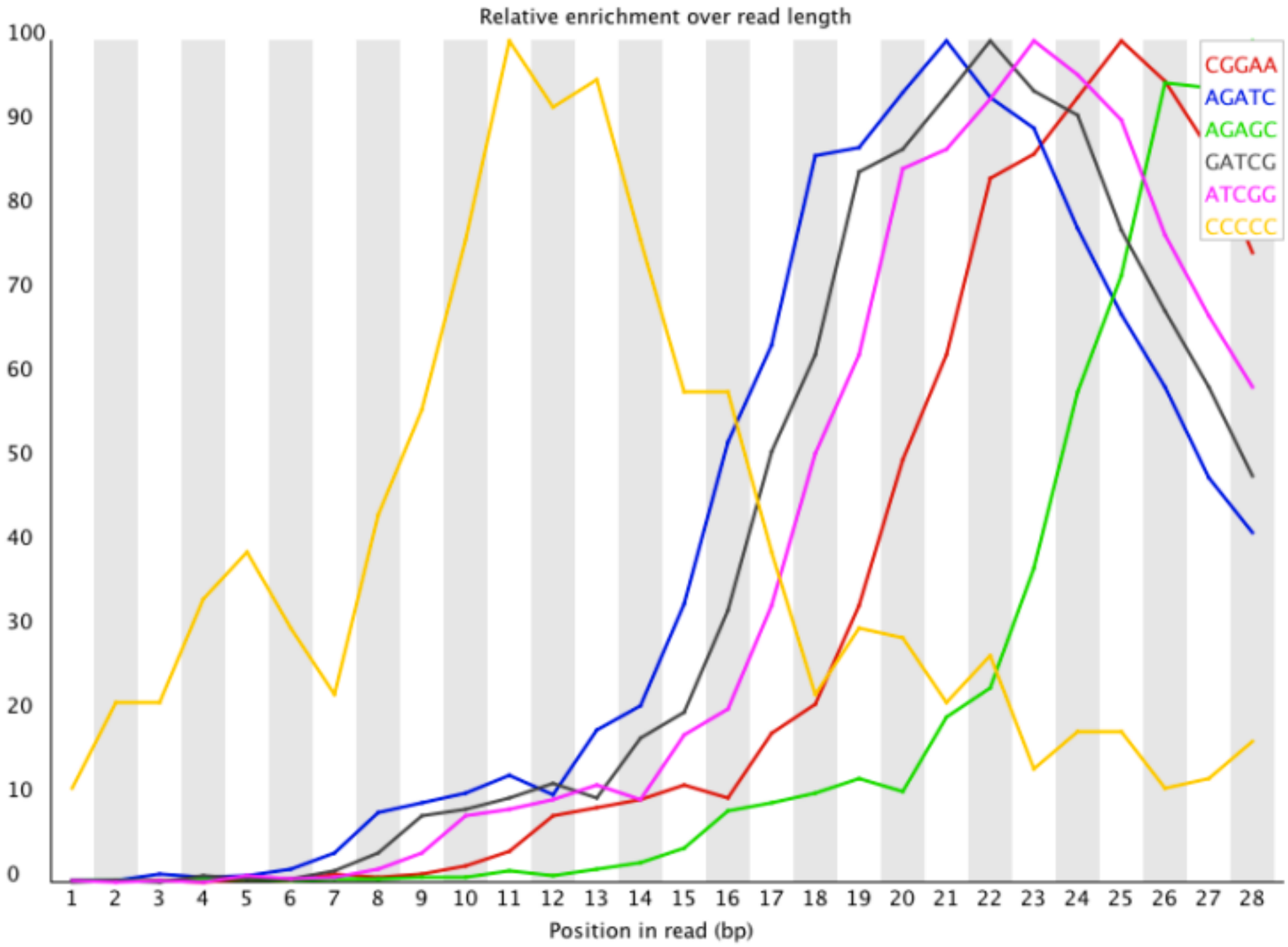
Sequence Length Distribution



Duplicate Sequences Distribution



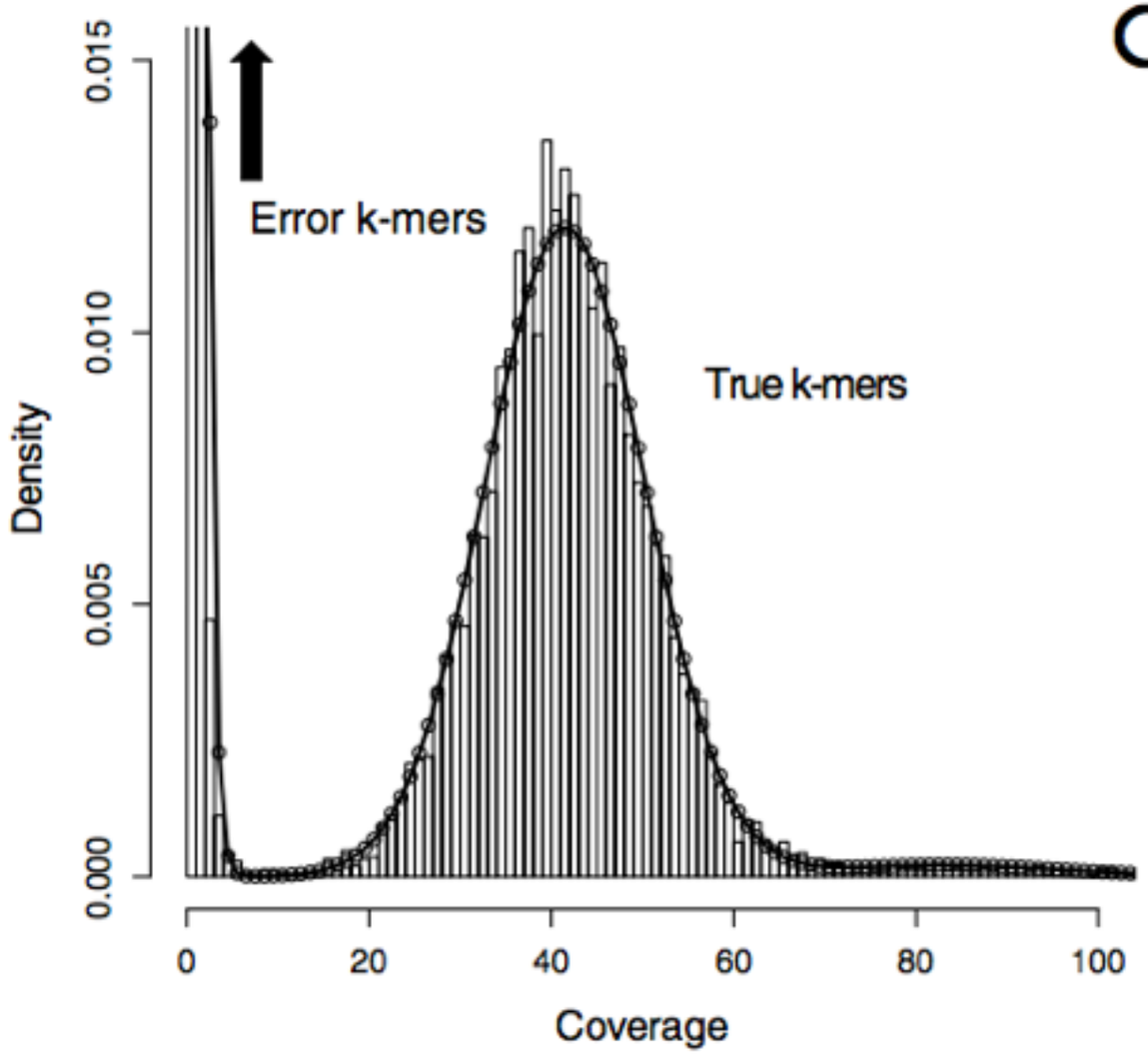
Overrepresented K-mers



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!

Overrepresented K-mers

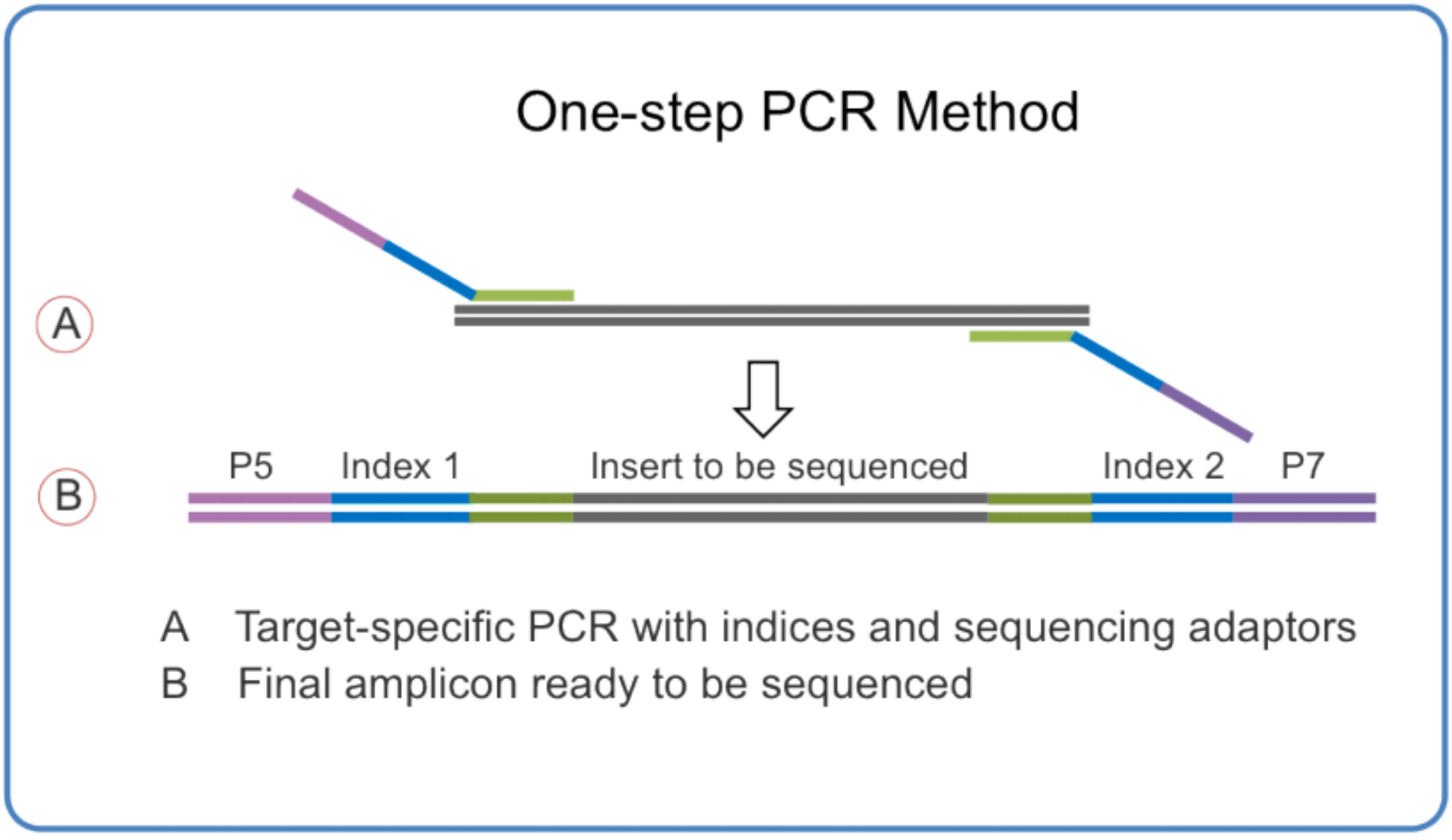


Concept: Rare *k*-mers are seq. errors
Need >15X coverage

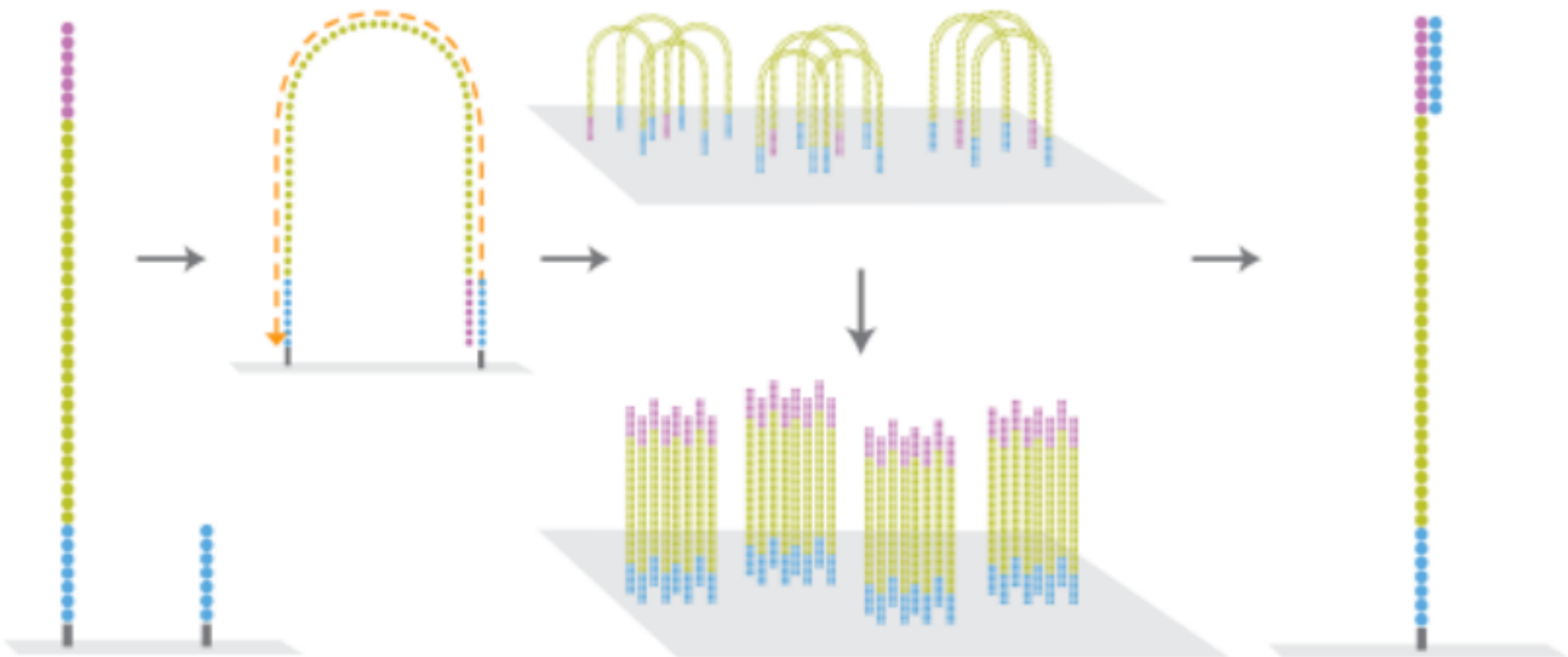
ACGTGGTTGCCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA
ACGTGGTTACCCTTAAA

GENOMIKA

Sequencing Process: PCR primers

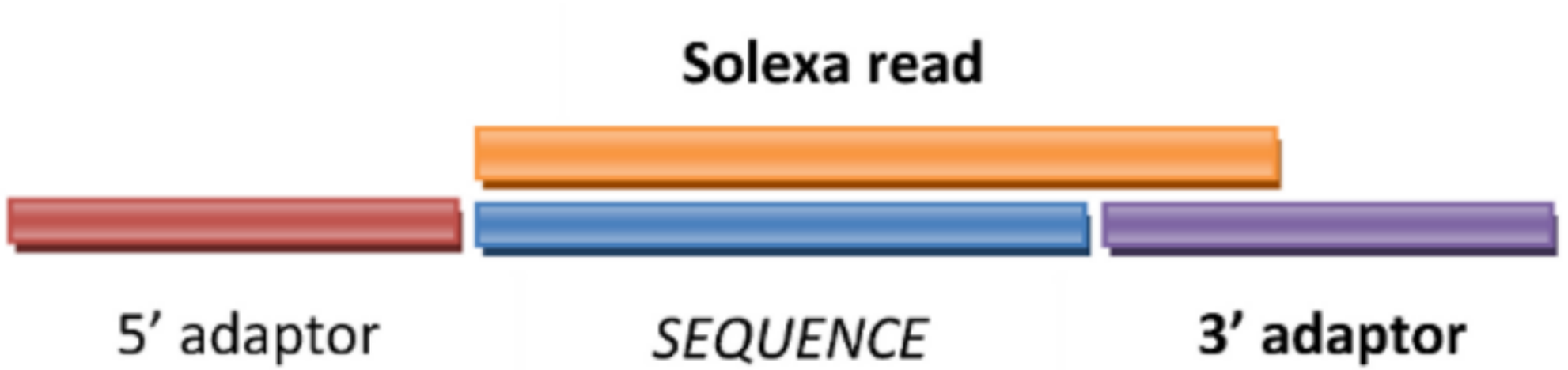


PCR primers



GENOMIKA

NGS adaptors & Cutadpt



GENOMIKA

NGS adaptors & Cutadpt

Overrepresented sequences

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

Very important if your DNA fragment is shorter than read length

GENOMIKA

Coverage

- Coverage/depth is how many times that your data covers the genome (on average)

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

- 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

GENOMIKA

Pré-processando os FASTQ's

Marcel Caraciolo, CTO
marcel@genomika.com.br