

# Lecture 2: Sequencing & k-mers

## Student Handout & In-Class Exercises

**Course:** BINF301 — Computational Biology

**Instructor:** Tom Michoel

**Date:** 19.01.2026

### 1. Overview

This handout summarizes the central ideas from Lecture 2, including:

- genome sequencing technologies,
- FASTA & FASTQ formats,
- quality scores and read trimming,
- definition and use of  $k$ -mers,
- genome size estimation using  $k$ -mer distributions.

### 2. Sequencing Technologies

Below is a comparison of the four major sequencing platforms:

Technology	Read Length	Accuracy	Throughput	Notes
Sanger	500–900 bp	Very high	Low	Chain termination; used for small fragments.
Illumina	100–300 bp	Very high	Very high	Short reads; quality decreases at 3' end.
PacBio HiFi	10–25 kb	Very high	Medium	Long reads; extremely accurate HiFi mode.
Nanopore	10 kb–1 Mb	Moderate	High	Electrical-signal based; ultra-long reads; portable devices.

**Discussion prompts:**

- Which technology is best for assembly, variant calling, or metagenomics?
- How do read length and accuracy interact?

### 3. FASTA & FASTQ

#### FASTA

- Stores only sequences (DNA, RNA, protein).
- Simple two-line format: header + sequence.

#### FASTQ

- Stores nucleotide sequences *and* per-base quality scores.
- Uses Phred encoding:  $Q = -10 \log_{10}(P_{\text{error}})$ .
- ASCII + 33 encoding for quality symbols.

### 4. Quality Control & Trimming

- Read quality often drops toward the 3' end of short reads.

- Adapters may be present and must be trimmed.
- Trimming improves downstream assembly and  $k$ -mer profiling.

## 5. k-mers

A  $k$ -mer is a substring of length  $k$  extracted from a sequence.

- Unique  $k$ -mer count approximates genome size.
- Sequencing errors introduce low-frequency unique  $k$ -mers.
- Repeats create high-frequency peaks.
- Tools like **GenomeScope** model errors, repeats, heterozygosity.

## 6. Discussion Starters

- Differences among sequencing generations.
- FASTA vs FASTQ usage.
- How trimming affects  $k$ -mer spectra.
- Why repeats cause high-frequency  $k$ -mers.
- How the Poisson distribution relates to coverage.

## In-Class Exercises

### Exercise 1 — Compare Sequencing Technologies

Using the table in Section 2, decide which platform you would use for:

- (a) genome assembly,
- (b) variant calling,
- (c) metagenomics.

Discuss trade-offs in read length, accuracy, speed, throughput, and biases.

**Solution. (a) Assembly: PacBio HiFi or Nanopore.** Long reads resolve repeats and SVs. HiFi pairs long length with very high accuracy; Nanopore provides ultra-long reads when needed (telomere-to-telomere).

**(b) Variant calling: Illumina or PacBio HiFi.** Illumina excels for SNPs/indels due to low error rates; HiFi adds long-range context and strong SV detection.

**(c) Metagenomics: Illumina** (deep, accurate profiling) or **Nanopore** (rapid, long reads; helpful for assembly in complex communities).

### Exercise 2 — Quality Score Interpretation

Given this FASTQ quality string:

```
@@@DDDDFFFFFGHIJ
```

- Convert several characters to Phred quality values.
- Identify low-quality read regions.
- Discuss how trimming would affect downstream  $k$ -mer counting.

**Solution. ASCII+33 mapping (examples):** @ (64)  $\rightarrow Q = 31$ ; D (68)  $\rightarrow Q = 35$ ; F (70)  $\rightarrow Q = 37$ ; G (71)  $\rightarrow Q = 38$ ; H (72)  $\rightarrow Q = 39$ ; I (73)  $\rightarrow Q = 40$ ; J (74)  $\rightarrow Q = 41$ .

**Interpretation:** All values are high ( $Q_{31-41}$ ), so no clear low-quality tail in this toy string.

**Trimming rationale:** In real data, low-quality 3' tails create spurious unique  $k$ -mers (singletons), distorting histograms and inflating genome-size estimates. Trimming reduces this noise.

### Exercise 3 — k-mer Counting Thought Experiment

Sequence:

```
ATGATGCT
```

Tasks:

- List all 4-mers.
- Count their frequencies.
- Predict how a sequencing error or repeat affects the histogram.

**Solution. 4-mers (sliding window):** ATGA, TGAT, GATG, ATGC, TGCT. Each occurs once.

**Effect of one sequencing error:** Typically creates novel singletons (low-frequency  $k$ -mers), adding a left-tail to the histogram.

**Effect of a repeat:** Increases counts proportionally (e.g., doubling the region gives each 4-mer count 2), creating higher-frequency peaks or secondary peaks.

**Exercise 4 — Mini Case: Genome Size Estimation**

Toy dataset:

ATG	20
TGA	19
GAT	22
ATC	1
TCA	1

Tasks:

- Identify likely sequencing errors.
- Estimate approximate coverage from the main peak.
- Explain how GenomeScope models errors, repeats, heterozygosity.

**Solution. Likely errors:** ATC and TCA (singletons).

**Coverage estimate:** Main cluster  $\approx 20\times$ .

**GenomeScope intuition:** Fits a mixture to the k-mer spectrum.

Errors  $\Rightarrow$  leftmost tail (low counts).

Repeats  $\Rightarrow$  peaks at multiples of coverage ( $2C, 3C, \dots$ ).

Heterozygosity  $\Rightarrow$  sub-peak near  $C/2$  (e.g.,  $\sim 10$  here).