
DTU

22126: Next Generation Sequencing Analysis
DTU - January 2026
Mick Westbury

Mick Westbury
Associate Professor
Section of Bioinformatics
Technical University of Denmark
micwe@dtu.dk

DATA BASICS

NGS Analysis workflow

Question

Raw data

Pre-processing

Assembly – mapping or de novo

Variant calling

Post-processing

Comparison

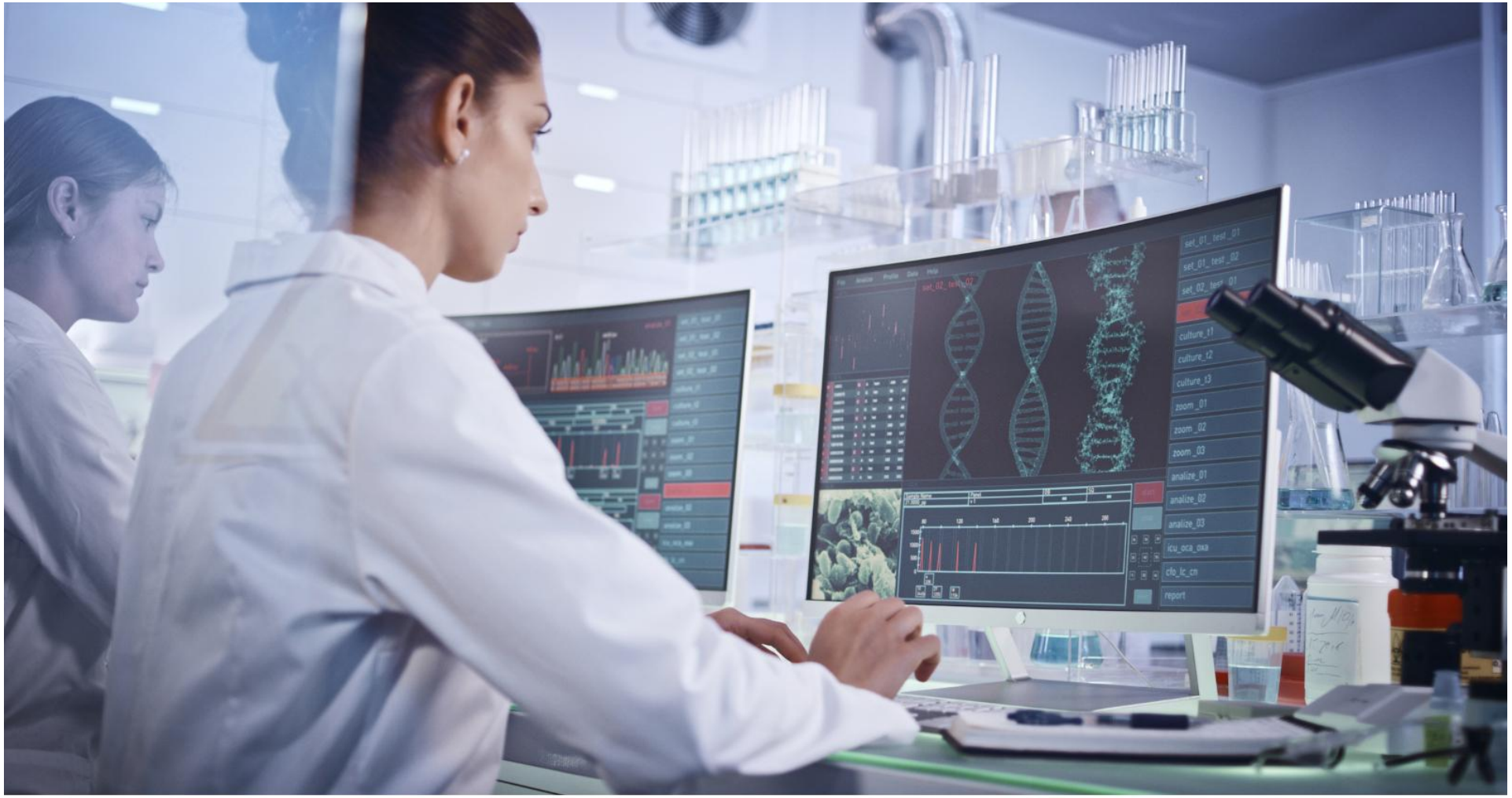
Answer

Why Raw Data Matters

- NGS analysis begins with raw sequencing reads
- Understanding data quality prevents downstream errors
- Poor input leads
  - Poor variant calling
  - Poor assembly
  - Poor quantification


GIGO
POWERFUL COMPUTER MODEL
GARBAGE IN
GARBAGE OUT

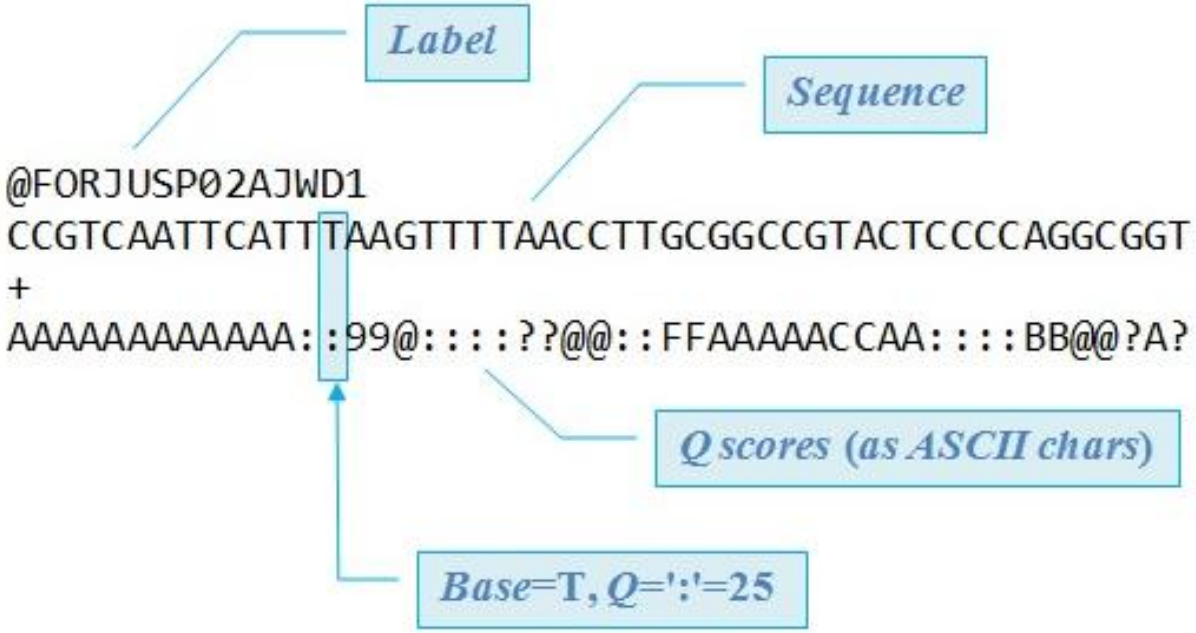
What NGS Data Looks Like



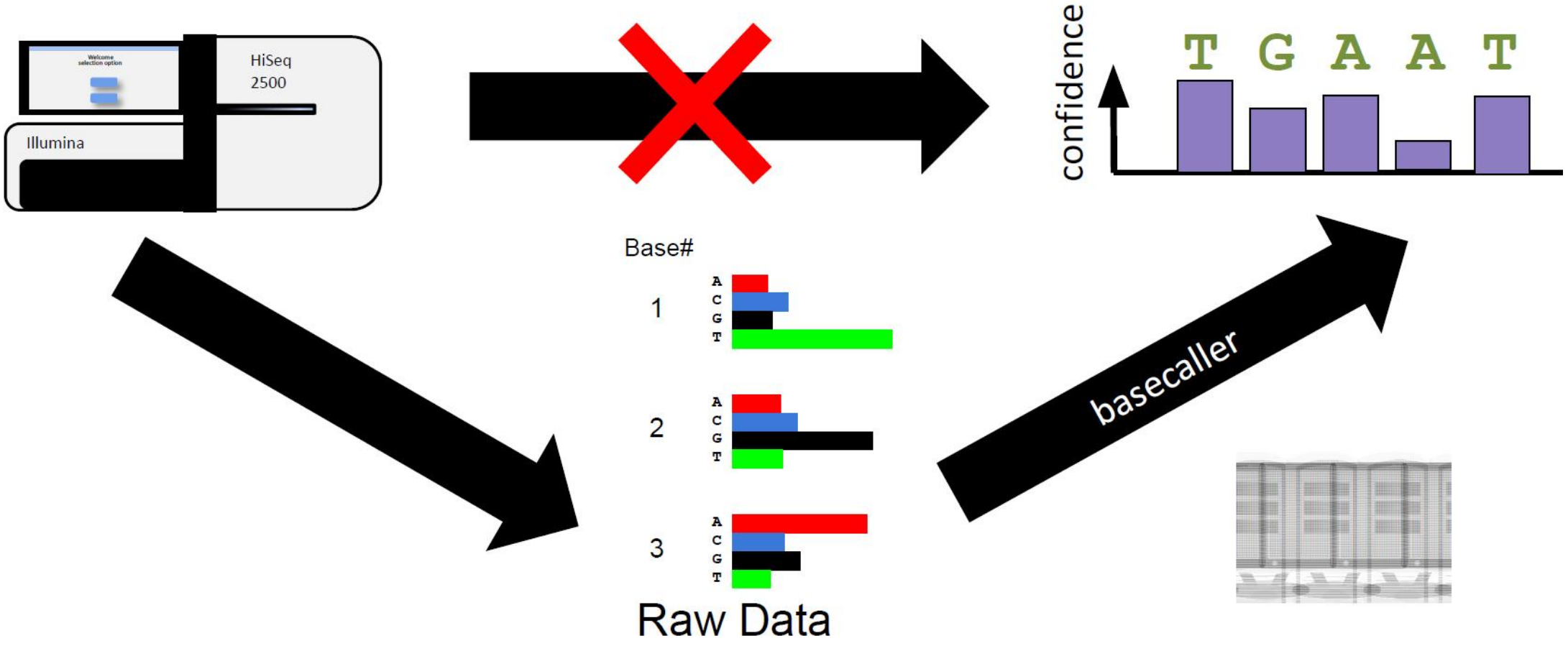
What NGS Data (Actually) Looks Like

FASTQ Format

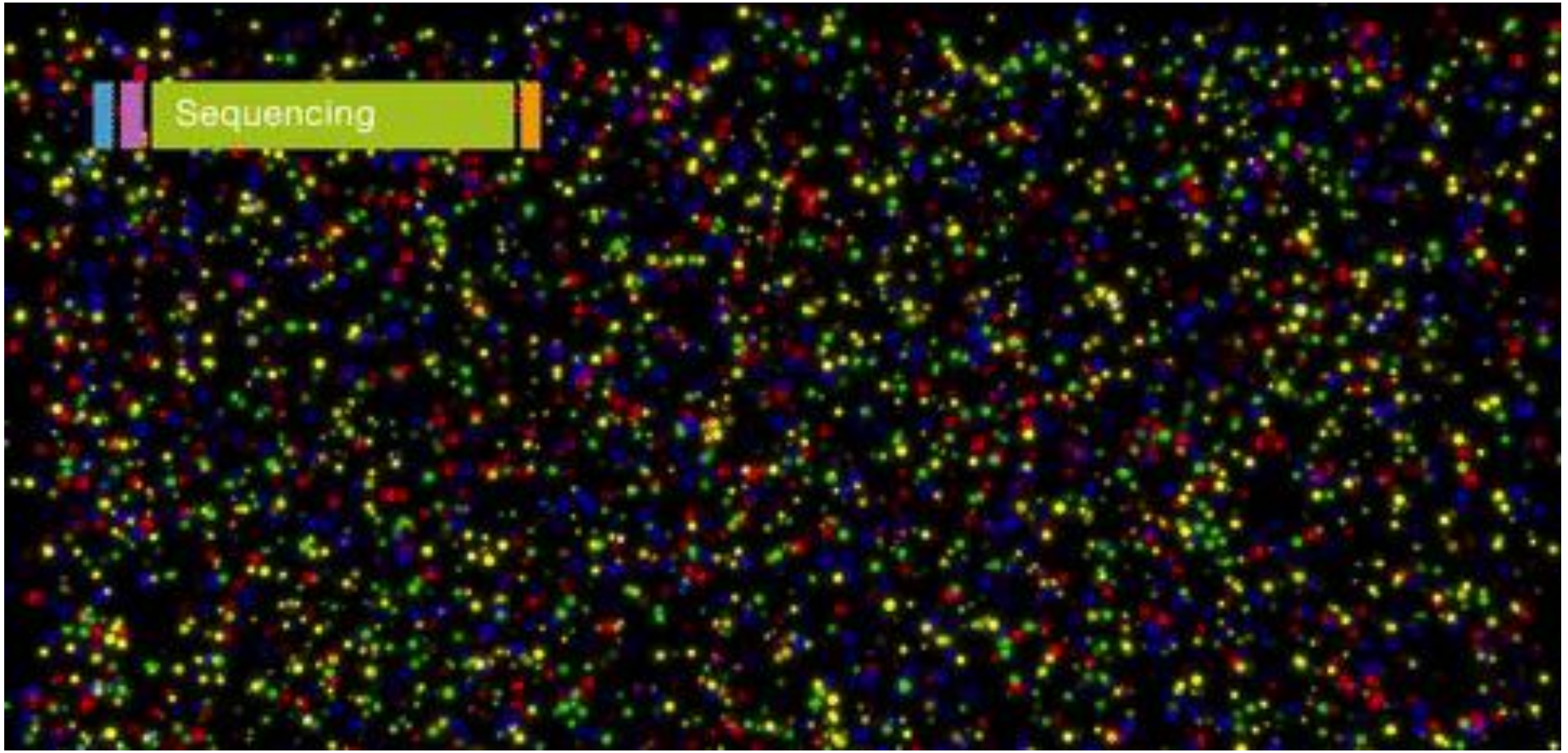
- Header line starts with “@”
- Sequence line (A, C, G, T, N)
- “+” line
- Quality line with ASCII-encoded PHRED values



FASTQ PHRED Scores (Quality)



FASTQ PHRED Scores (Quality)



FASTQ PHRED Scores (Quality)

S - Sanger	Phred+33,	raw reads typically (0, 40)	
X - Solexa	Solexa+64,	raw reads typically (-5, 40)	
I - Illumina 1.3+	Phred+64,	raw reads typically (0, 40)	
J - Illumina 1.5+	Phred+64,	raw reads typically (3, 41)	
	with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator		
	(Note: See discussion above).		
L - Illumina 1.8+	Phred+33,	raw reads typically (0, 41)	
N - Nanopore	Phred+33,	Duplex reads typically (0, 50)	
E - ElemBio AVITI	Phred+33,	raw reads typically (0, 55)	
P - PacBio	Phred+33,	HiFi reads typically (0, 93)	

FASTQ PHRED Scores (Quality)

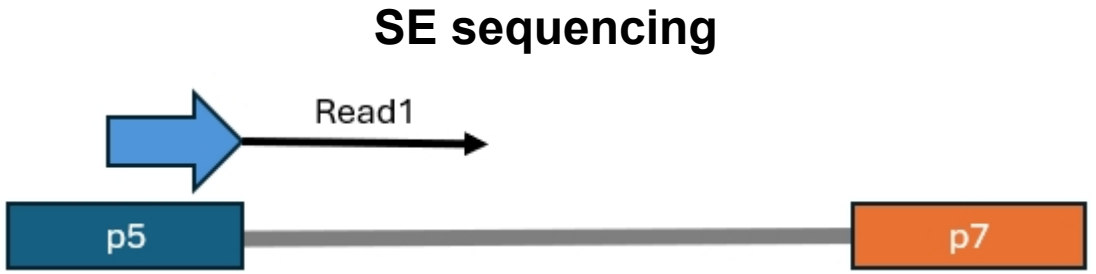
- Measure probability of a sequencing error
- PHRED = –10 × log10(error probability)
- Higher score means more confidence
- Quality typically decreases toward end of read

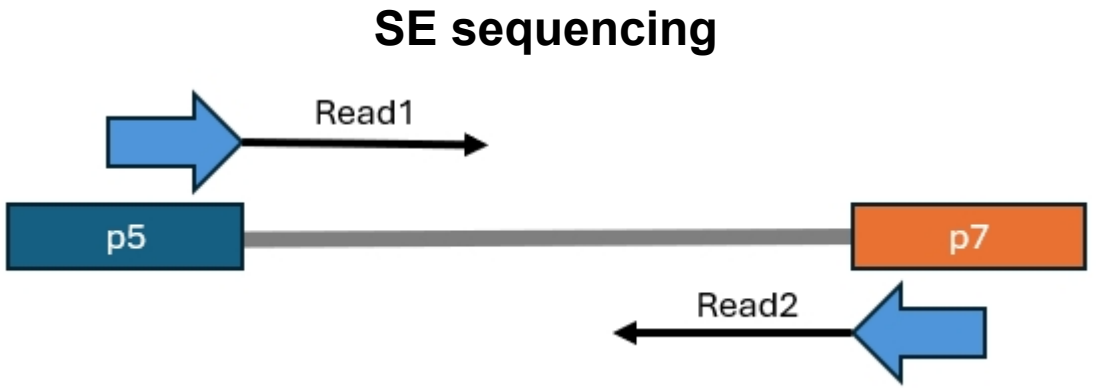
Sequencing Errors

- Substitutions (most common in Illumina)
- Insertions or deletions
- Systematic errors:
  - –GC-rich motifs, repeats, homopolymers
- Random errors increase with cycle number

Read Layouts

- Single-end reads (SE)
- Paired-end reads (PE)
- Insert size
- Fragment size





Library Preparation Biases

- PCR amplification bias
- GC bias
- Adapter sequences
- Primer dimers
- Overrepresented sequences

Basic QC Concepts

- Per-base sequence quality
- Sequence composition
- GC content distribution
- Read length distribution
- N (missing data) content
- Duplicate sequences

Coverage Concepts

- **Depth**: average reads per base
- **Breadth**: proportion of genome covered
- Depends on
  - –read count
  - –read length
  - –genome size
- Uneven coverage affects variant calling