Introduction to Applied Bioinformatics - day2

KAUST Bioinformatics platform | KAUST

Recap of day1

Tools: Linux commands, seqkit

Bioinformatics is a multidisciplinary field (bio, cs, stat)

It covers various domains and applications

Linux command line (pwd, mkdir, ls, cd, nano, wc, etc)

File formats (fasta, fastq)

Seqkit for sequence file overview

Agenda for day2

Tools: fastqc, fastp, star, samtools

Day 2: Data Access, Quality Control & Processing

Objectives: Access and download data from public repositories, perform quality evaluation and control, and understand and execute alignment

Time	Session
09:00-09:30	**S4: Sequencing Technologies** Journey through the evolution of DNA sequencing – from Sanger to Illumina short-reads to the long-read revolution of PacBio and Nanopore. Understand which technology to choose for your research question.
09:30-10:00	**S2: Analysis Pipeline Steps** See the big picture. Walk through a complete bioinformatics workflow – from raw reads to biological insight – and understand how each step connects to the next.
10:00-10:15	Coffee Break
10:15-12:15	**S1: Public Data Repositories** Tap into the world's largest collections of biological data. Learn to navigate NCBI SRA, GEO, Ensembl, and other repositories that give you free access to millions of experiments.
	L3: Data Retrieval Hands-on Download real sequencing data and reference genomes using professional command-line tools – the same workflow used by researchers worldwide. View Lab Instructions →
12:15-13:30	Lunch Break
13:30-15:00	**S3: Data Quality Control** Great analysis starts with great data. Learn to spot problems early by understanding Phred scores, adapter contamination, GC bias, and duplication – before they derail your results.
	L4: FastQC & Trimming Hands-on Run FastQC and fastp on real data, trim away low-quality reads, and see the dramatic difference quality control makes. View Lab Instructions →
15:00-15:15	Coffee Break
15:15-17:00	**S4: Sequence Alignment Fundamentals** Understand how millions of short reads find their home on a reference genome. Dive into the algorithms behind splice-aware alignment and learn what mapping quality really means.
	L5: Reference Genome Alignment & BAM files - Hands-on Align reads to a reference genome and evaluate the results with samtools – a core skill for any genomics project. View Lab Instructions →

A Typical Bioinformatics Workflow

1

Question

Define a clear, testable biological question

2

Data Acquisition

Obtain raw data (sequencing, public databases, experiments)

3

Quality Control

Assess data quality, trim adapters, filter low-quality reads

4

Processing

Alignment, assembly, quantification, variant calling

5

Analysis

Statistical testing, clustering, differential expression, enrichment

6

Interpretation

Biological context, pathway analysis, literature integration

7

Visualization

Plots, heatmaps, genome browsers, interactive dashboards

8

Reporting

Reproducible documentation, data deposition, publication

Sequencing Technologies

Session Objectives

- Understand the principles of DNA sequencing
- Compare different sequencing technologies
- Learn about common sequencing experiments
- Match technologies to appropriate use cases

Part 1: Sequencing Fundamentals

What is DNA Sequencing?

Definition

DNA Sequencing is the process of determining the precise order of nucleotides (A, T, G, C) within a DNA molecule.

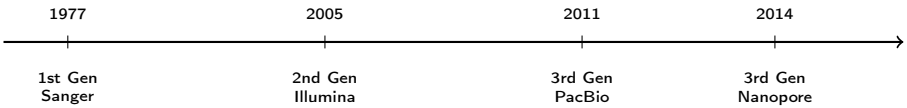
Why Sequence DNA?

- Identify genetic variants
- Discover new genes
- Understand disease mechanisms
- Study evolution
- Develop personalized medicine

Key Metrics

- **Read length**: How long each sequence
- **Throughput**: How much data
- **Accuracy**: Error rate
- **Cost**: Per base or per genome
- **Speed**: Time to results

Generations of Sequencing



1st Generation

- Sanger sequencing
- Long reads, low throughput
- Still gold standard for validation

2nd Generation

- Short reads
- Massive parallelization
- High throughput, low cost

3rd Generation

- Long reads
- Single molecule
- Real-time sequencing

Part 2: Sanger Sequencing

Sanger Sequencing: The Pioneer

Principle: Chain Termination

1. DNA template + primer
2. Add DNA polymerase
3. Normal dNTPs + fluorescent ddNTPs
4. ddNTPs terminate chain extension
5. Separate fragments by size
6. Read fluorescent labels

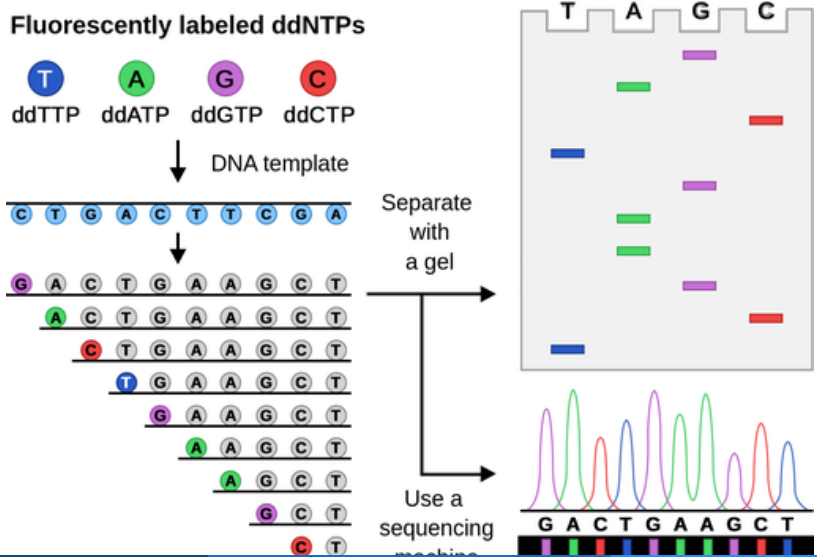
Characteristics

- Read length: 700–1000 bp
- Very high accuracy (99.99%)
- Low throughput
- Cost: $1–5 per reaction

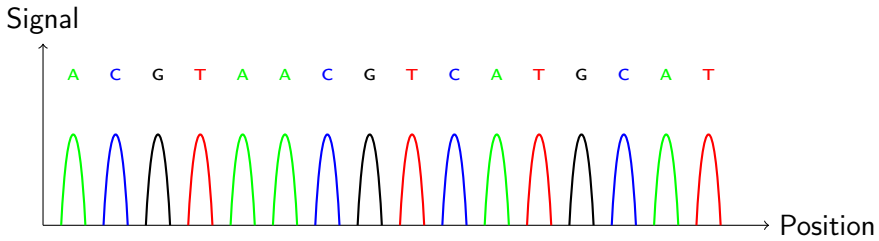
Still Used For

- Validating variants
- Small-scale projects
- Plasmid sequencing

Sanger Sequencing: cnt.



Sanger Sequencing Output



Key Concept

A **chromatogram** shows the fluorescent signal for each base. Peak height indicates confidence. Overlapping or weak peaks suggest low quality or heterozygosity.

Part 3: Short-Read Sequencing (Illumina)

Illumina Sequencing: Overview

Key Features

- Sequencing by synthesis (SBS)
- Massive parallel sequencing
- Read length: 50–300 bp
- Very high accuracy (>99.9%)
- Dominant platform (>80% market)

Throughput

- MiSeq: 15 Gb
- NextSeq: 120 Gb
- NovaSeq: 6000 Gb (6 Tb!)

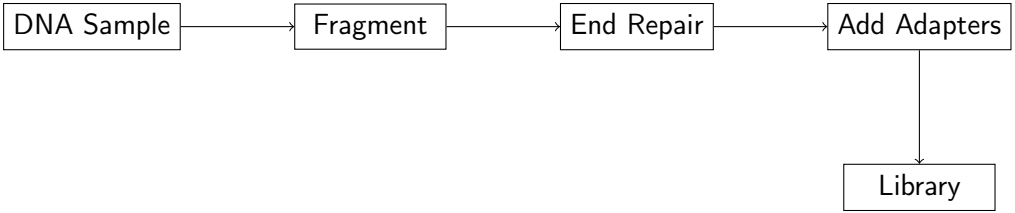
Workflow Steps

1. Library preparation
2. Cluster generation
3. Sequencing by synthesis
4. Data analysis

Cost Revolution

Human genome sequencing dropped from $3 billion (2003) to under $200 (2024)!

Illumina: Library Preparation



Adapters Include:

- Sequences for flow cell binding
- Primer binding sites
- Index/barcode for multiplexing

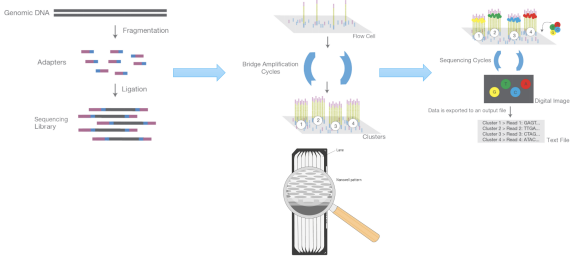
Multiplexing

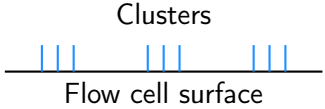
- Pool multiple samples
- Each has unique barcode
- Demultiplex after sequencing
- Cost-effective!

Illumina: Cluster Generation

Bridge Amplification

1. Library binds to flow cell
2. DNA bends to form bridge
3. PCR creates copies
4. Clusters of identical sequences
5. Each cluster = one read





Key Concept

Millions of clusters form on a flow cell, enabling massive parallel sequencing.

Illumina: Sequencing by Synthesis

How It Works

1. Add fluorescently labeled nucleotides
2. Only one nucleotide incorporates (blocked)
3. Image the flow cell (detect fluorescence)
4. Cleave blocker and fluorophore
5. Repeat for next base

Four Colors = Four Bases

- Each base has unique fluorescent label
- Camera captures all clusters simultaneously
- Software calls bases from images

Paired-End Sequencing

- Sequence both ends of fragment
- 2×150 bp common
- Better mapping
- Detect structural variants

Illumina: Strengths & Limitations

Strengths

- Very high accuracy (>99.9%)
- Massive throughput
- Low cost per base
- Well-established pipelines
- Extensive tool support
- Good for quantification

Limitations

- Short read length (300 bp)
- Difficult with repetitive regions
- GC bias in some protocols
- Requires PCR amplification
- Library prep takes time
- Cannot detect some modifications

Key Concept

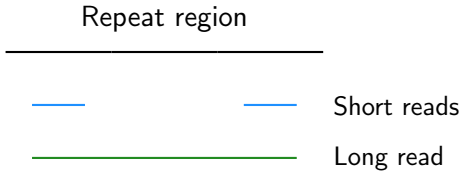
Illumina is ideal for: variant calling, RNA-seq, ChIP-seq, metagenomics, and any application needing high accuracy and depth.

Part 4: Long-Read Sequencing

Why Long Reads?

Short Read Challenges

- Repetitive regions
- Structural variants
- Haplotype phasing
- Full-length transcripts
- De novo assembly



Solution

Long reads span difficult regions!

PacBio: Single Molecule Real-Time (SMRT)

Technology

- Zero-mode waveguides (ZMWs)
- Single polymerase at bottom
- Watch incorporation in real-time
- Circular consensus sequencing (HiFi)

Specifications

- Read length: 10–25 kb (up to 100 kb)
- HiFi accuracy: >99.9%
- CLR accuracy: 90%
- Throughput: 25 Gb (Revio)

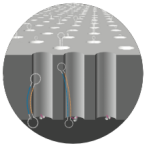
HiFi Reads

- Circular DNA template
- Polymerase reads multiple times
- Consensus from passes
- High accuracy + long length!

Best For

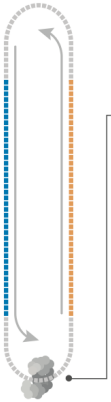
- Genome assembly
- Full-length transcripts
- Structural variants
- Methylation detection

How SMRT sequencing works

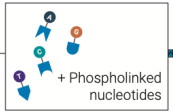


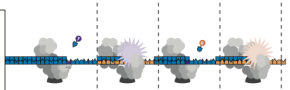
SMRT® Cells contain millions of zero-mode waveguides (ZMWs)

SMRTbell templates enable repeated sequencing of circular template with real-time detection of base incorporation



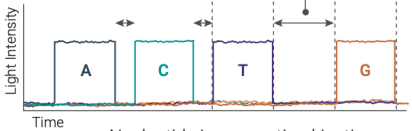
A single molecule of DNA is immobilized in each ZMW





As anchored polymerases incorporate labeled bases, light is emitted

Directly detect DNA modifications during sequencing



Nucleotide incorporation kinetics are measured in real time

Oxford Nanopore: Sequencing Through a Pore

Technology

- Protein nanopore in membrane
- DNA passes through pore
- Each base disrupts current
- Measure current changes
- Real-time base calling

Specifications

- Read length: No limit! (record >4 Mb)
- Accuracy: 95–99% (improving)
- Throughput: varies by device

Unique Features

- Portable devices (MinION)
- Real-time analysis
- Direct RNA sequencing
- Native DNA modifications
- No amplification needed

Device Options

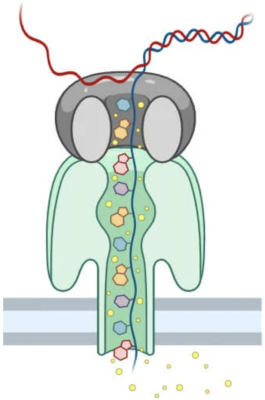
- MinION: Portable, USB
- GridION: Benchtop
- PromethION: High throughput

MinION

GridION

Flongle

PromethION



Nanopore Sequencing

Long-Read Comparison

Feature	PacBio HiFi	Oxford Nanopore
Read length	10–25 kb	Unlimited (avg 10–30 kb)
Accuracy	>99.9%	95–99%
Throughput	High (Revio)	Variable
Speed	Hours to days	Real-time
Portability	Lab-based	Portable (MinION)
Cost per Gb	Medium	Low–Medium
DNA modifications	Yes (kinetics)	Yes (direct)
Direct RNA	No	Yes

Key Concept

Both platforms are improving rapidly. Choice depends on application, accuracy needs, and resources.

Part 5: Common Sequencing Experiments

DNA Sequencing Experiments

Whole Genome Sequencing (WGS)

- Sequence entire genome
- Reference: 30× coverage
- Detects all variant types
- Most comprehensive

Whole Exome Sequencing (WES)

- Only protein-coding regions (1.5%)
- More cost-effective
- Good for clinical genetics
- Misses non-coding variants

Targeted Sequencing

- Specific genes/regions
- Very high depth possible
- Gene panels (cancer, inherited)
- Amplicon sequencing

Coverage Depth

- WGS: 30–60×
- WES: 100–200×
- Targeted: 500–1000×

RNA Sequencing Experiments

Bulk RNA-seq

- Gene expression profiling
- Differential expression
- Alternative splicing
- Typically 20–50M reads/sample

Single-cell RNA-seq

- Expression per cell
- Cell type identification
- Developmental trajectories
- Thousands of cells

Total RNA vs mRNA

- Poly-A selection: mRNA only
- Ribosomal depletion: all RNA
- Small RNA-seq: miRNA, etc.

Long-read RNA-seq

- Full-length transcripts
- Isoform identification
- No assembly needed

Epigenomics & Chromatin

ChIP-seq

- Chromatin immunoprecipitation
- Protein-DNA interactions
- Histone modifications
- Transcription factor binding

ATAC-seq

- Chromatin accessibility
- Open chromatin regions
- Regulatory elements
- Low input requirement

Methylation Sequencing

- WGBS: Whole genome bisulfite
- RRBS: Reduced representation
- Long-read: Direct detection

Why Epigenomics?

- Gene regulation
- Disease mechanisms
- Development
- Environmental effects

Choosing the Right Technology

Application	Recommended	Why
Variant calling (SNPs)	Illumina	High accuracy
Structural variants	Long-read	Span breakpoints
De novo assembly	Long-read + Illumina	Length + polish
RNA-seq (bulk)	Illumina	Cost-effective
Full-length transcripts	PacBio/ONT	No assembly
Metagenomics	Illumina or ONT	Throughput or portability
Clinical diagnostics	Illumina	Validated pipelines
Field work	Nanopore	Portable
Methylation	Long-read or WGBS	Direct or comprehensive

Key Concept

Often the best approach combines technologies — e.g., long reads for assembly, short reads for polishing and variant calling.

Session Summary

Key Concept

Sequencing Technologies

- **Sanger**: Gold standard, low throughput, validation
- **Illumina**: Short reads, high accuracy, dominant platform
- **PacBio**: Long reads, HiFi accuracy, assembly
- **Nanopore**: Ultra-long reads, portable, real-time

Common Experiments

- DNA: WGS, WES, targeted panels
- RNA: Bulk RNA-seq, single-cell, isoforms
- Epigenomics: ChIP-seq, ATAC-seq, methylation

Questions?

Hands-on Session: Retrieving Reference Genomes

Tools: datasets & seqkit

Biological resources

NCBI

Genbank: https://www.ncbi.nlm.nih.gov/genbank/

an annotated collection of all publicly available DNA sequences

What It Contains

- Nucleotide sequences
- Annotations and features
- Literature references
- Part of INSDC collaboration

Key Features

- Comprehensive (all organisms)
- Links to other NCBI resources
- Multiple format downloads

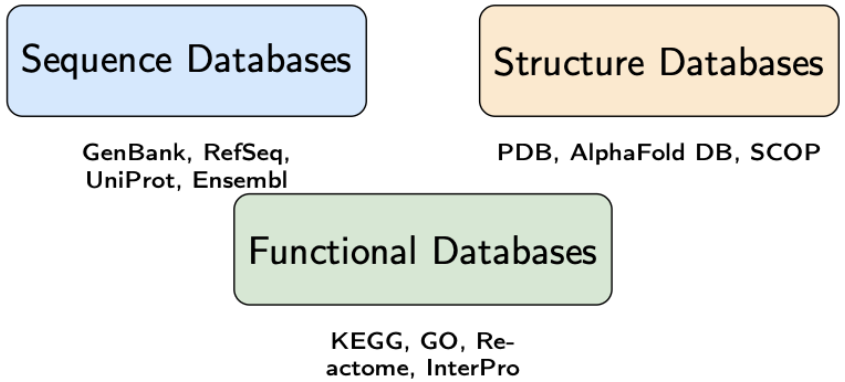
refseq: https://www.ncbi.nlm.nih.gov/refseq/

What Makes It Different

- Curated, non-redundant
- Reference standard sequences
- Consistent annotation
- Quality controlled

Accession Prefixes

- NM_: mRNA
- NR_: non-coding RNA
- NP_: Protein
- NC_: Chromosome
- NG_: Genomic region



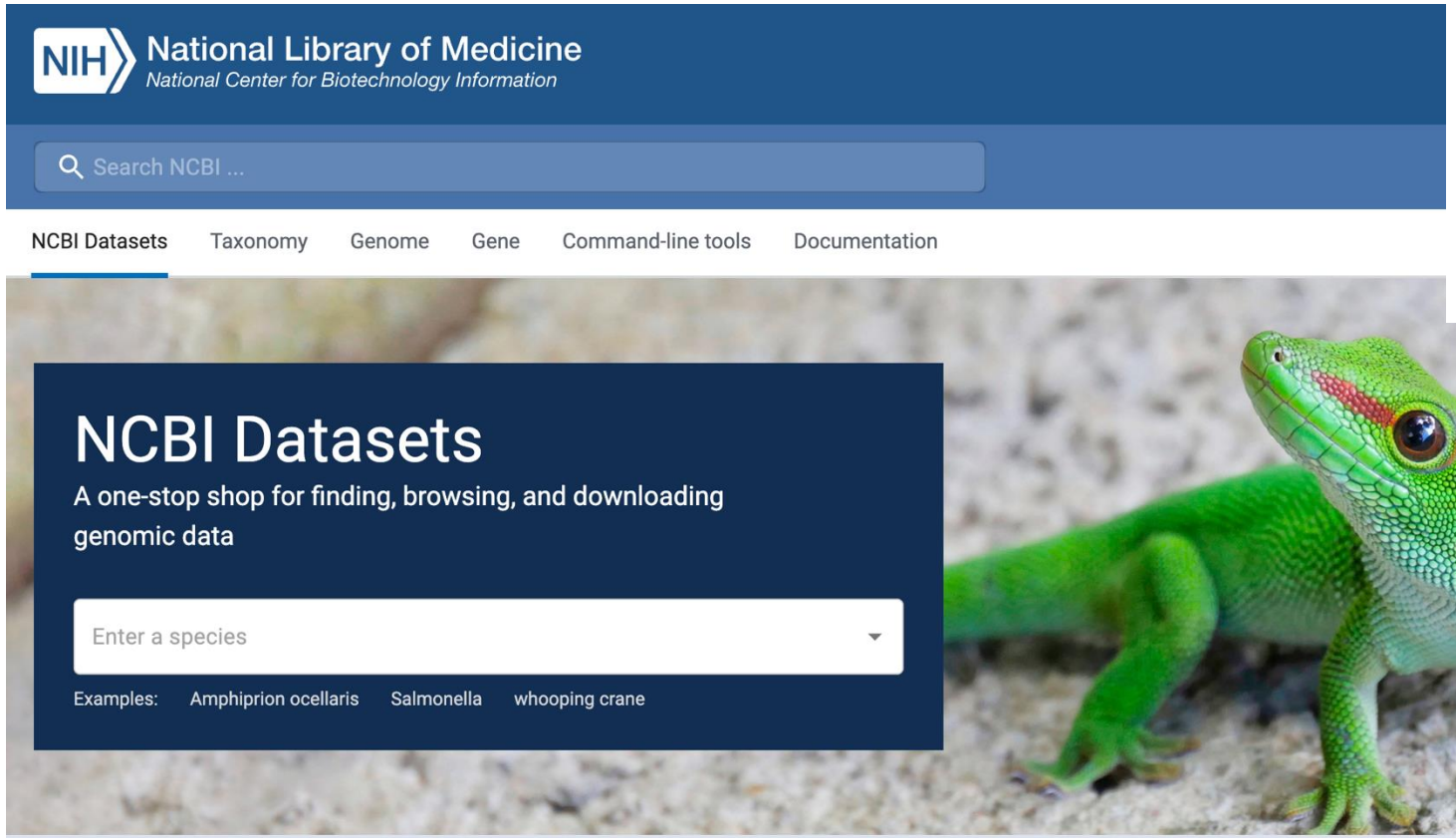
Ensembl: https://www.ensembl.org

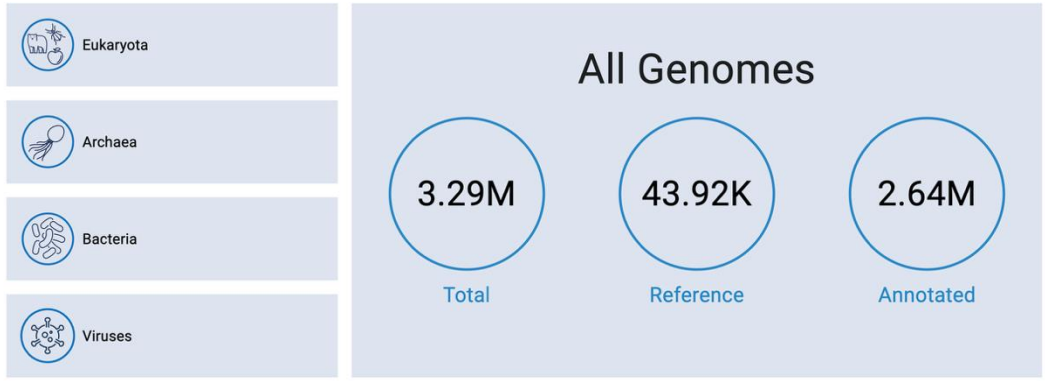
What It Contains
- Genome assemblies
- Gene annotations
- Comparative genomics
- Variation data
- Regulatory features

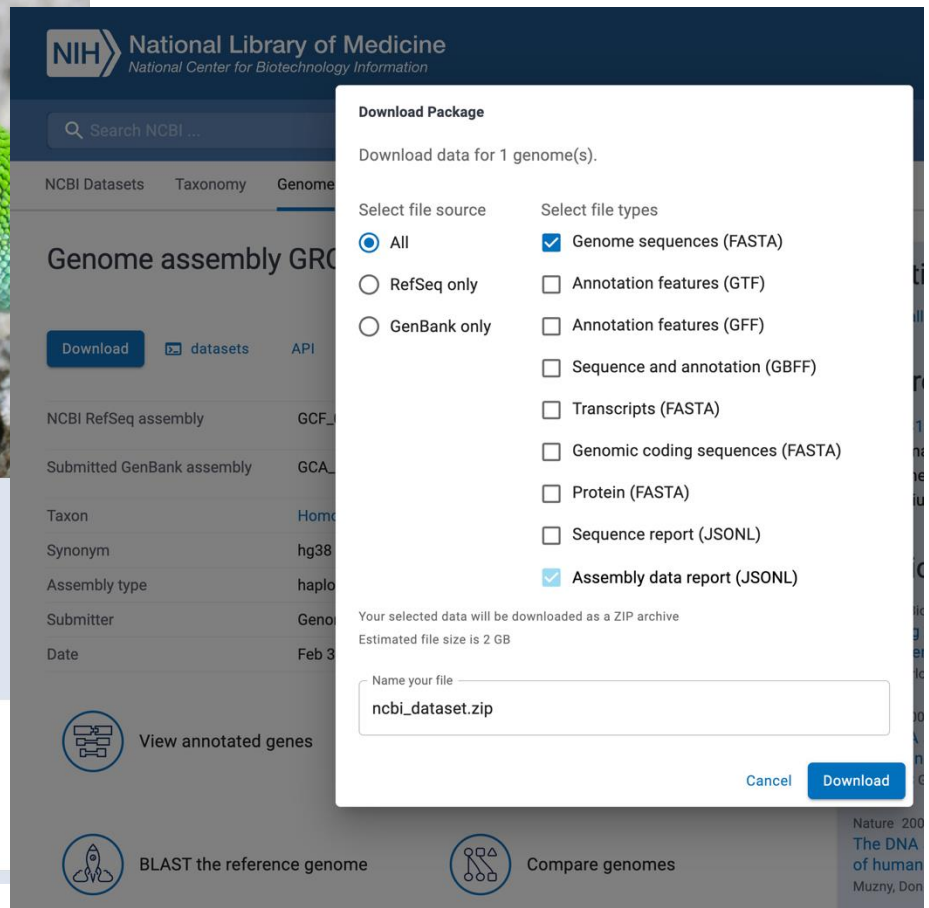
Useful For
- Downloading reference genomes
- Gene/transcript information
- Ortholog identification
- Variant annotation (VEP)

BioMart tool for advanced queries

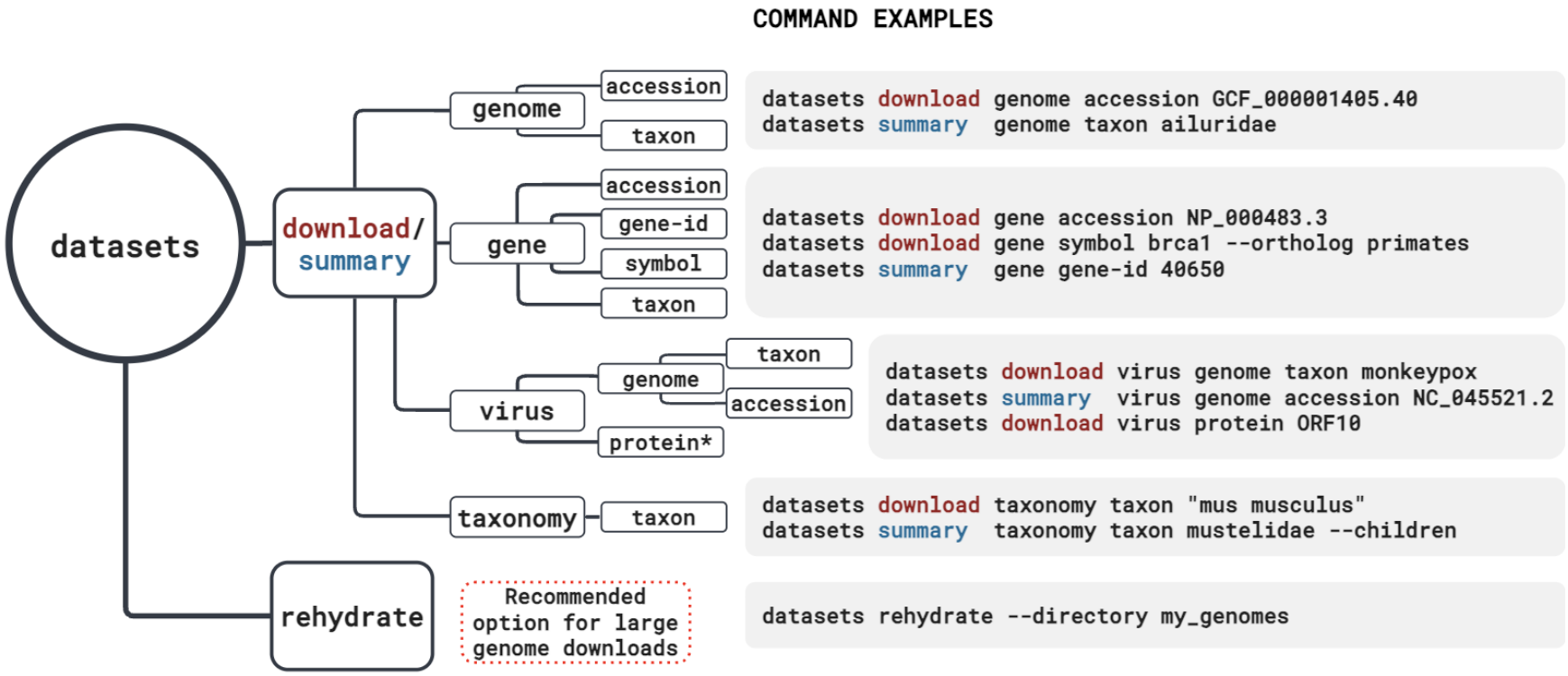
Open NCBI: https://www.ncbi.nlm.nih.gov/datasets/







Obtaining data from NCBI



* virus protein: restricted to download of SARS-Cov-2 proteins

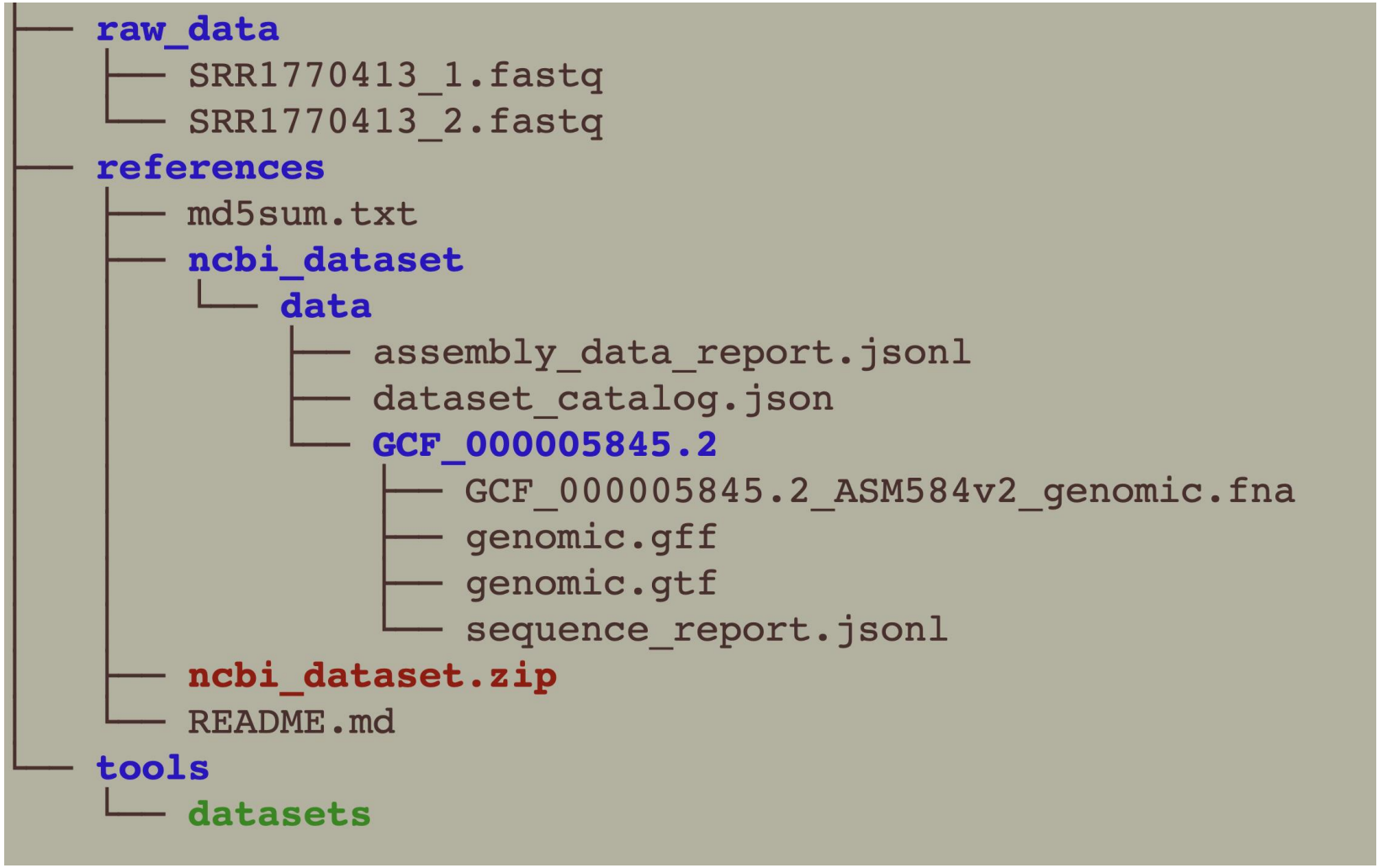
Instructionsto use datasets: Lab 3

Hands-on Session:

Retrieving Raw Sequencing Data From Public Repositories

Tools: sratoolkits: prefetch & fasterq-dump

Output


raw_data
SRR1770413_1.fastq
SRR1770413_2.fastq
references
md5sum.txt
ncbi_dataset
data
assembly_data_report.jsonl
dataset_catalog.json
GCF_000005845.2
GCF_000005845.2_ASM584v2_genomic.fna
genomic.gff
genomic.gtf
sequence_report.jsonl
ncbi_dataset.zip
README.md
tools
datasets

Sequence Quality Evaluation & Quality Control

Interactive including Hands-on sessions using: **fastqc**, **fastp**

Dataset inspection

Before getting into the data ensure:

- Review the metadata or sample sheet
- You have the right dataset (files from the correct project and study)
- Match number of samples and replicates with number of FastQ files
- Check the size of all files and look for any big difference (**du -sh**)
- Use **md5** checksum if it is copied from another source
- Print the last few lines of each file to make sure they end correctly
- Count number of lines and they should be divisible by 4, why ? (**wc -l**)
- Make sure the files are named properly (don't start with numbers, no spaces or special characters in the file name, etc.
- If data is paired-end, the FastQ files should usually end with as *_1.fastq* and *_2.fastq*

Why Quality Control Matters

Poor Quality Bases

Low-quality base calls at read ends can lead to incorrect alignments and false variant calls.

Adapter Contamination

Residual adapter sequences cause mis-alignments or complete removal of reads during mapping.

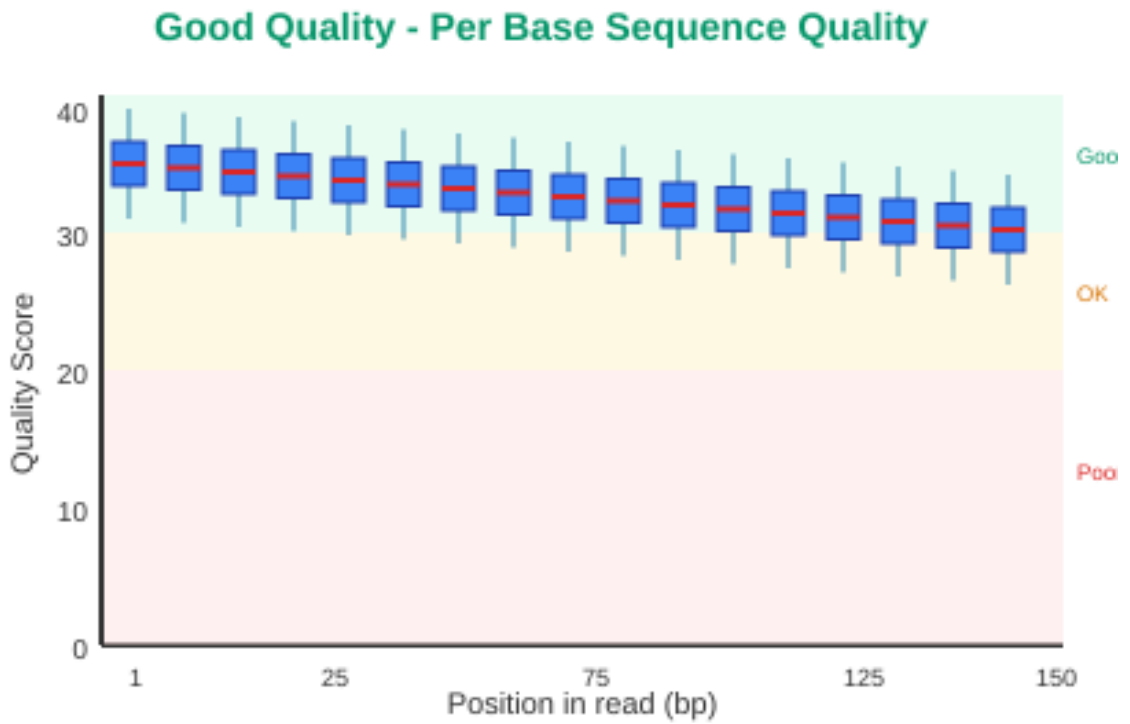
Too-Short Reads

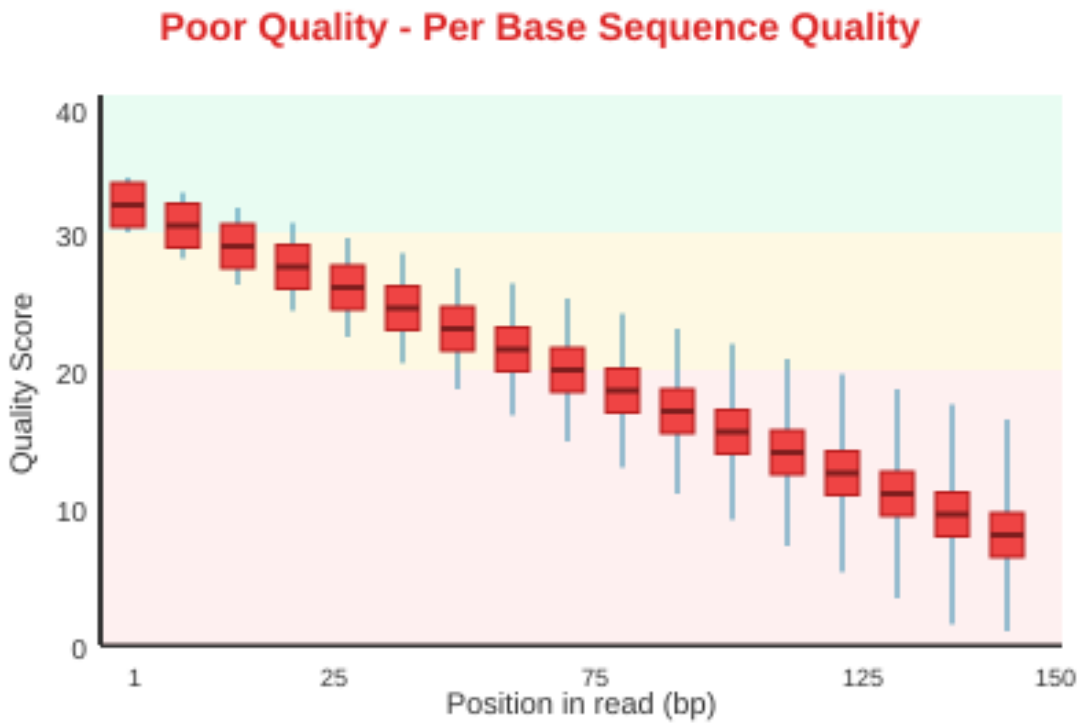
Very short reads after trimming cannot be reliably aligned and may crash alignment software.

Method: *Trim Galore* a wrapper for *FastQC* and *Cutadapt* or *fastp* that has many features

1) Per-Base Sequence Quality

Phred quality scores at each position along the read





Interpretation:

• Green zone (Q30+): Excellent quality, 99.9% accuracy • Yellow zone (Q20-30): Acceptable • Red zone (<Q20): Poor quality
• Quality drop at 3' end is normal — trim if severe • Box plots show median (red line), IQR (box), and range (whiskers)

Understanding Phred Quality Scores

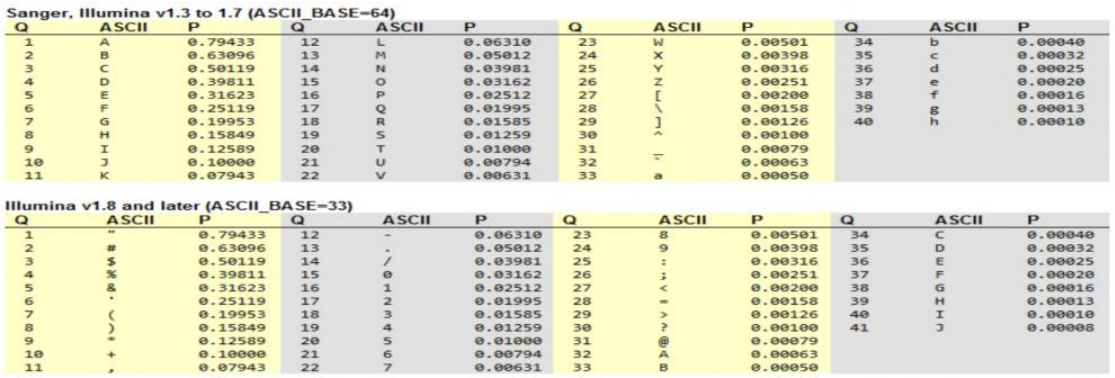
Phred quality scores (Q) indicate the probability of an incorrect base call:

$Q = -10 \times \log_{10}(P)$ where P = probability of error

Phred Score	Error Probability	Base Call Accuracy	Interpretation
Q10	1 in 10 (10%)	**90%**	Poor - not usable
Q20	1 in 100 (1%)	**99%**	Acceptable minimum
Q30	1 in 1,000 (0.1%)	**99.9%**	Good - standard target
Q40	1 in 10,000 (0.01%)	**99.99%**	Excellent

Key takeaway: Aim for >80% of bases at Q30 or above.
Modern Illumina sequencers typically achieve >85% Q30

```
@K00235:171:H2VKHBBXY:3:1101:19380:2475 1:N:0:CGATGT
CAGAGCATTCNGCTACTCCATTACAACGCGTGTGGAATGGG
+
AAAAAEEEEE#EEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```



Sanger, Illumina v1.3 to 1.7 (ASCII_BASE=64)

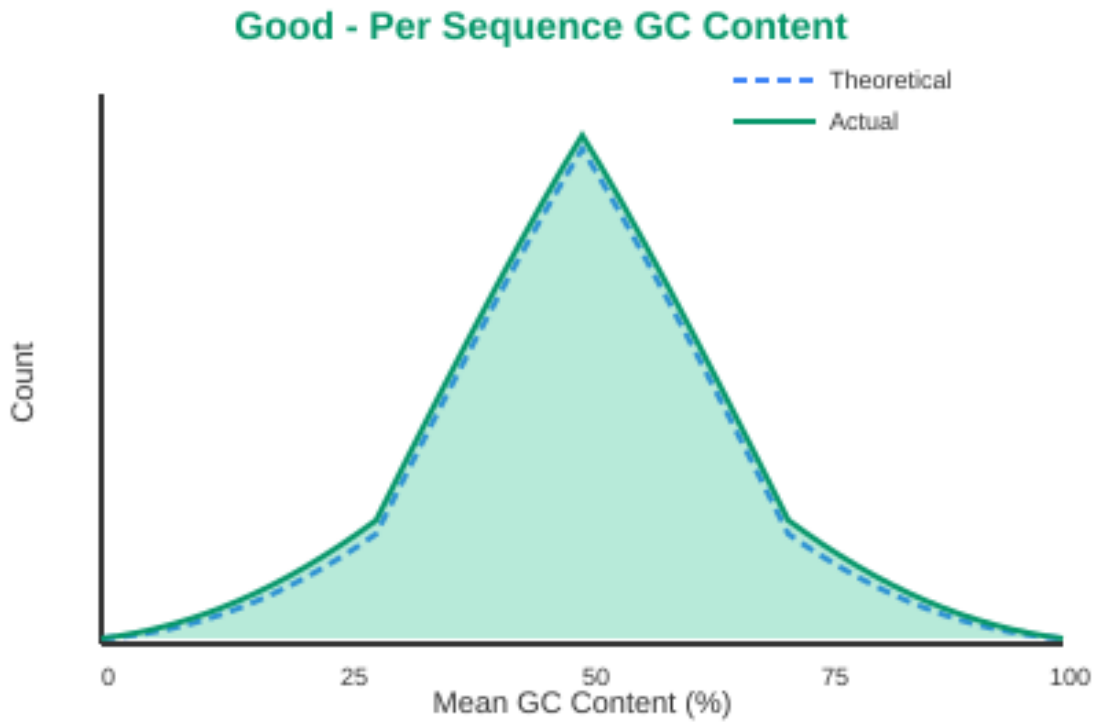
Q	ASCII	P	Q	ASCII	P	Q	ASCII	P	Q	ASCII	P
1	A	0.79433	12	L	0.06310	23	W	0.00501	34	b	0.00040
2	B	0.63096	13	M	0.05012	24	X	0.00398	35	c	0.00032
3	C	0.50119	14	N	0.03981	25	Y	0.00316	36	d	0.00025
4	D	0.39811	15	O	0.03162	26	Z	0.00251	37	e	0.00020
5	E	0.31623	16	P	0.02512	27	[	0.00200	38	f	0.00016
6	F	0.25119	17	Q	0.01995	28	\	0.00158	39	g	0.00013
7	G	0.19953	18	R	0.01585	29	]	0.00126	40	h	0.00010
8	H	0.15849	19	S	0.01259	30	^	0.00100			
9	I	0.12589	20	T	0.01000	31	_	0.00079			
10	J	0.10000	21	U	0.00794	32	`	0.00063			
11	K	0.07943	22	V	0.00631	33	a	0.00050			

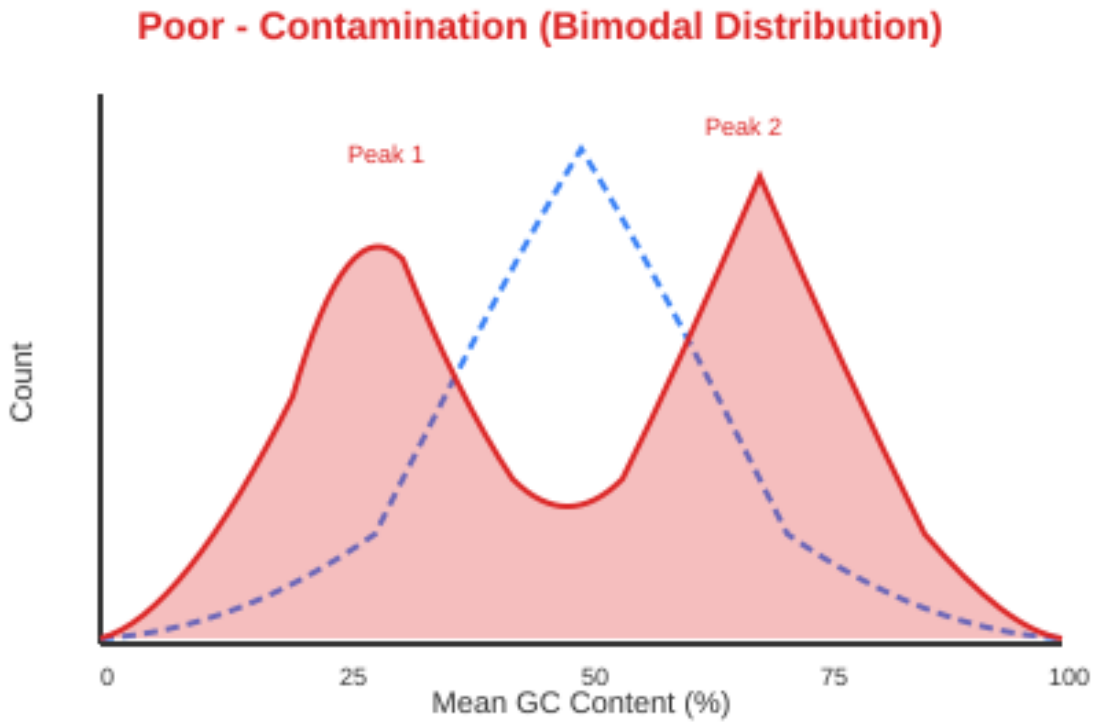
Illumina v1.8 and later (ASCII_BASE=33)

Q	ASCII	P	Q	ASCII	P	Q	ASCII	P	Q	ASCII	P
1	"	0.79433	12	-	0.06310	23	8	0.00501	34	C	0.00040
2	#	0.63096	13	.	0.05012	24	9	0.00398	35	D	0.00032
3	$	0.50119	14	/	0.03981	25	:	0.00316	36	E	0.00025
4	%	0.39811	15	0	0.03162	26	;	0.00251	37	F	0.00020
5	&	0.31623	16	1	0.02512	27	<	0.00200	38	G	0.00016
6	'	0.25119	17	2	0.01995	28	=	0.00158	39	H	0.00013
7	(	0.19953	18	3	0.01585	29	>	0.00126	40	I	0.00010
8	)	0.15849	19	4	0.01259	30	?	0.00100	41	J	0.00008
9	*	0.12589	20	5	0.01000	31	@	0.00079			
10	+	0.10000	21	6	0.00794	32	A	0.00063			
11	,	0.07943	22	7	0.00631	33	B	0.00050			

2) GC Content Distribution

Distribution of GC content across all reads — should match expected genome GC%



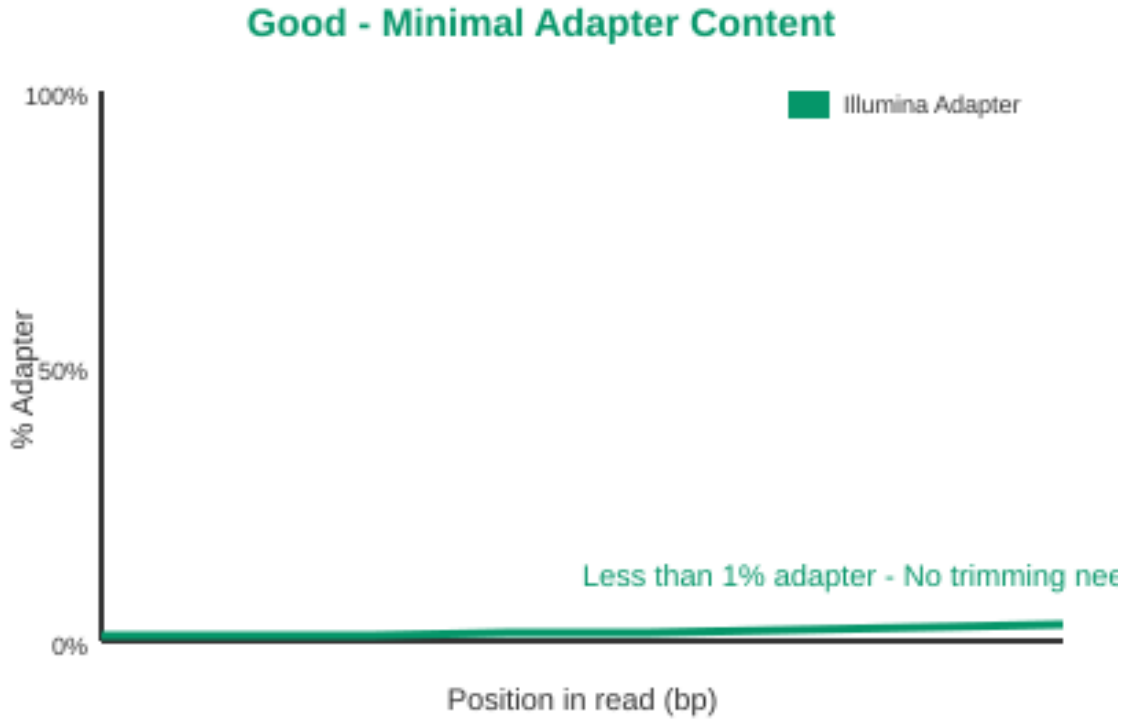


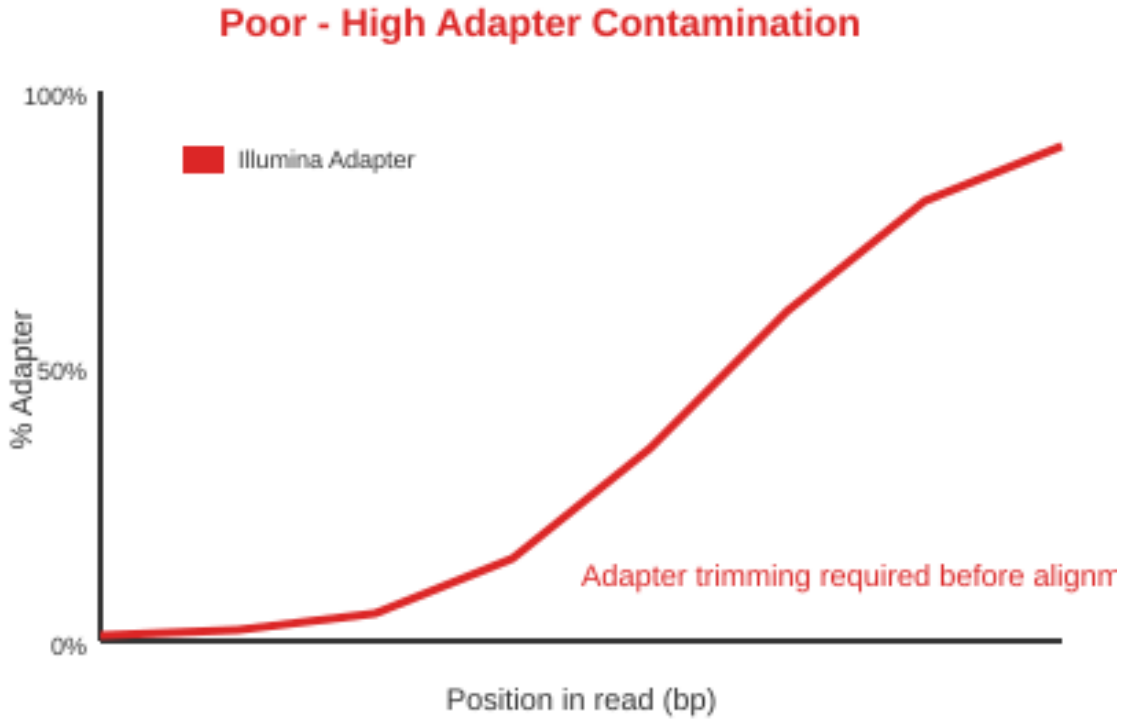
Interpretation:

- Good: Single peak matching theoretical distribution (blue dashed line) — indicates clean, uncontaminated sample
- Bad: Multiple peaks suggest contamination (bacteria, adapter dimers) or significant library bias — investigate source

3) Adapter Conten

Cumulative percentage of reads with adapter sequence at each position





Interpretation:

- Good: Flat line near 0% — adapters were properly removed or insert size is appropriate
- Bad: Rising curve at 3' end — short inserts or read-through into adapter; use Cutadapt or Trimmomatic to remove

Adapter Trimming

Detecting and removing sequencing adapter sequences

What It Does

- Removes adapter sequences from 3' end of reads
- Auto-detects adapter type (Illumina, Nextera, sRNA) (fastp uses PE read overlap analysis)
- Uses stringent overlap detection (even 1 bp match)
- Searches first 1m reads to find adapters

Why It's Critical

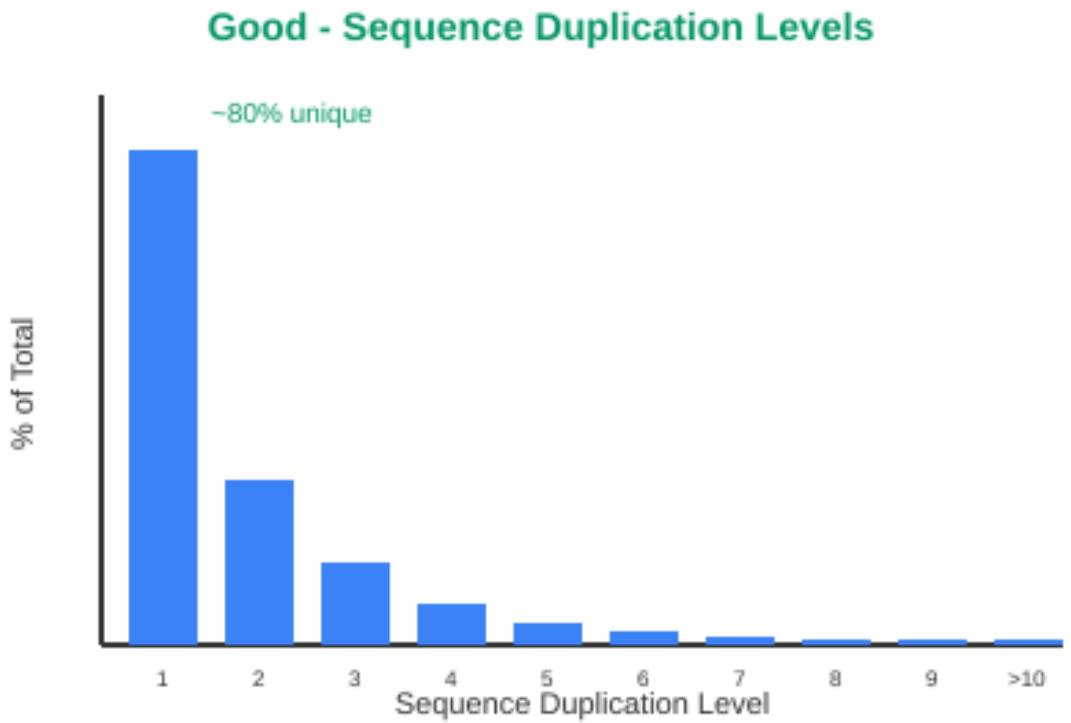
- Adapter sequences are not part of the biological sample
- Contamination causes incorrect alignments
- In bisulfite-seq: leads to wrong methylation calls

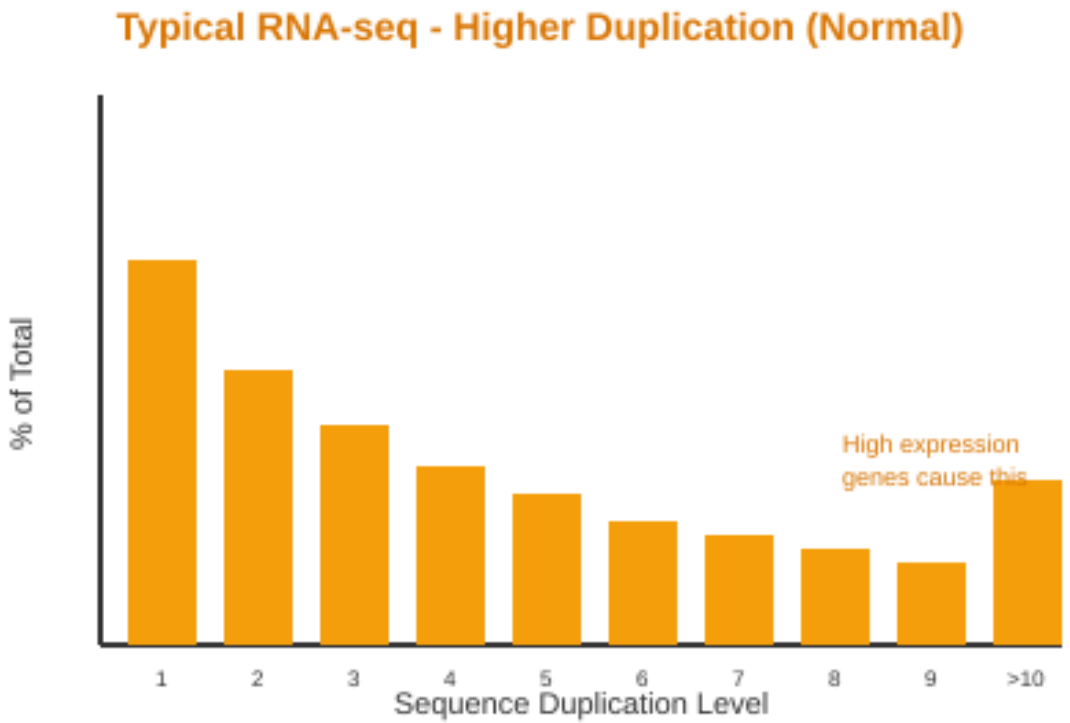
Example:

- Illumina paired-end adapters ('AGATCGGAAGAGC') are automatically recognized and removed. For reads shorter than the insert size, adapter read-through is detected and trimmed.
- For paired-end data, fastp analyzes where read pairs overlap. If Read 1 ends with adapter sequence and Read 2 begins with it (palindrome pattern), this confirms adapter contamination, and both are trimmed.

4) Sequence Duplication Levels

Percentage of sequences appearing at different duplication levels





Interpretation:

- DNA-seq: Most sequences should be unique (left plot) — high duplication indicates low library complexity or PCR artifacts
- RNA-seq: Higher duplication is NORMAL (right plot) — highly expressed genes naturally produce many identical reads

Duplication Analysis

Evaluating and optionally removing PCR duplicates (tool: fastp)

What It Does

- Calculates duplication rate across the dataset
- Optional deduplication to remove identical reads

Source:

- Very deep sequencing?
- Specific genes?
- PCR issues?
- Ribosomal RNAs sequenced?

Example:
fastp reports duplication levels with GC content correlation. High-duplication reads often have biased GC, indicating PCR artifacts. This helps diagnose library prep problems early.

Subsampling / Read Limiting

Processing only a subset of reads

What It Does

- --reads_to_process: Limit total reads processed
- Default 0 means process all reads
- Creates filtered subset of original data

Use Cases

- Quick quality preview of large datasets
- Create test subsets for pipeline development
- Downsample for resource-limited analysis

Example:
Using --reads_to_process 1000000 processes only the first 1M reads. Ideal for quickly checking data quality before committing to full processing.

Decontaminating reads

Why is Decontamination Critical?

Contaminated reads lead to unreliable results and wasted resources

Common Contamination Sources

- **Ribosomal RNA (rRNA)**

Can comprise 40-80% of total RNA-seq reads

- **Host genomic DNA**

Human/mouse contamination in samples

- **Microbial contamination**

Bacteria, fungi, or viral sequences

- **Technical artifacts**

Adapters, primers, PhiX spike-in

Impact on Analysis

- **Reduced mapping rates**

Fewer reads align to target genome

- **Skewed gene expression**

False differential expression results

- **Wasted sequencing depth**

Paying to sequence contaminants

- **Assembly artifacts**

Chimeric contigs, misassemblies

Up to 80% of RNA-seq reads can be rRNA without depletion

Tools for Decontamination

Tool	Target	Best For
Kraken2 + KrakenTools	Bacteria, viral, human, fungi	General-purpose decontamination
BBSplit (BBTools)	Host genomes, multiple refs	Host removal, PDX experiments
SortMeRNA	rRNA only	RNA-seq rRNA depletion
Bowtie2 + SAMtools	Any reference genome	Precise removal, custom refs

Tip: Combine tools for comprehensive decontamination

Use SortMeRNA for rRNA, then Kraken2/BBSplit for remaining contaminants

Best Practices & Quality Control

Best Practices

- **Run FastQC before AND after**

Compare read counts, quality distributions

- **Use appropriate databases**

Match to your expected contaminants

- **Keep logs and statistics**

Track % reads removed at each step

- **Validate with known samples**

Test pipeline on positive controls

- **Document tool versions**

Ensure reproducibility

Warning Signs

- **>30% reads removed**

May indicate sample or library issue

- **Very low mapping after cleaning**

Check if correct reference was used

- **Unexpected species in Kraken**

Cross-contamination or mislabeling

- **Inconsistent results across replicates**

Batch effects or technical issues

- **Quality drop after filtering**

Review filtering parameters

Hands-on: Quality Control

Lab 3: Quality Control Hands-on

Learning Objectives

- Run FastQC to assess raw read quality
- Interpret FastQC reports and identify quality issues
- Use fastp for quality trimming and adapter removal
- Compare pre- and post-trimming quality with MultiQC

Step 1: Initial QC with FastQC

```
fastqc *.fastq.gz -o fastqc_raw/ -t 8
```

Step 2: Trim adapters and low-quality bases

```
fastp -i R1.fq.gz -I R2.fq.gz -o trim_R1.fq.gz -O
trim_R2.fq.gz  --detect_adapter_for_pe -q 20 -l 36 --
html fastp.html
```

Step 3: Post-trim QC

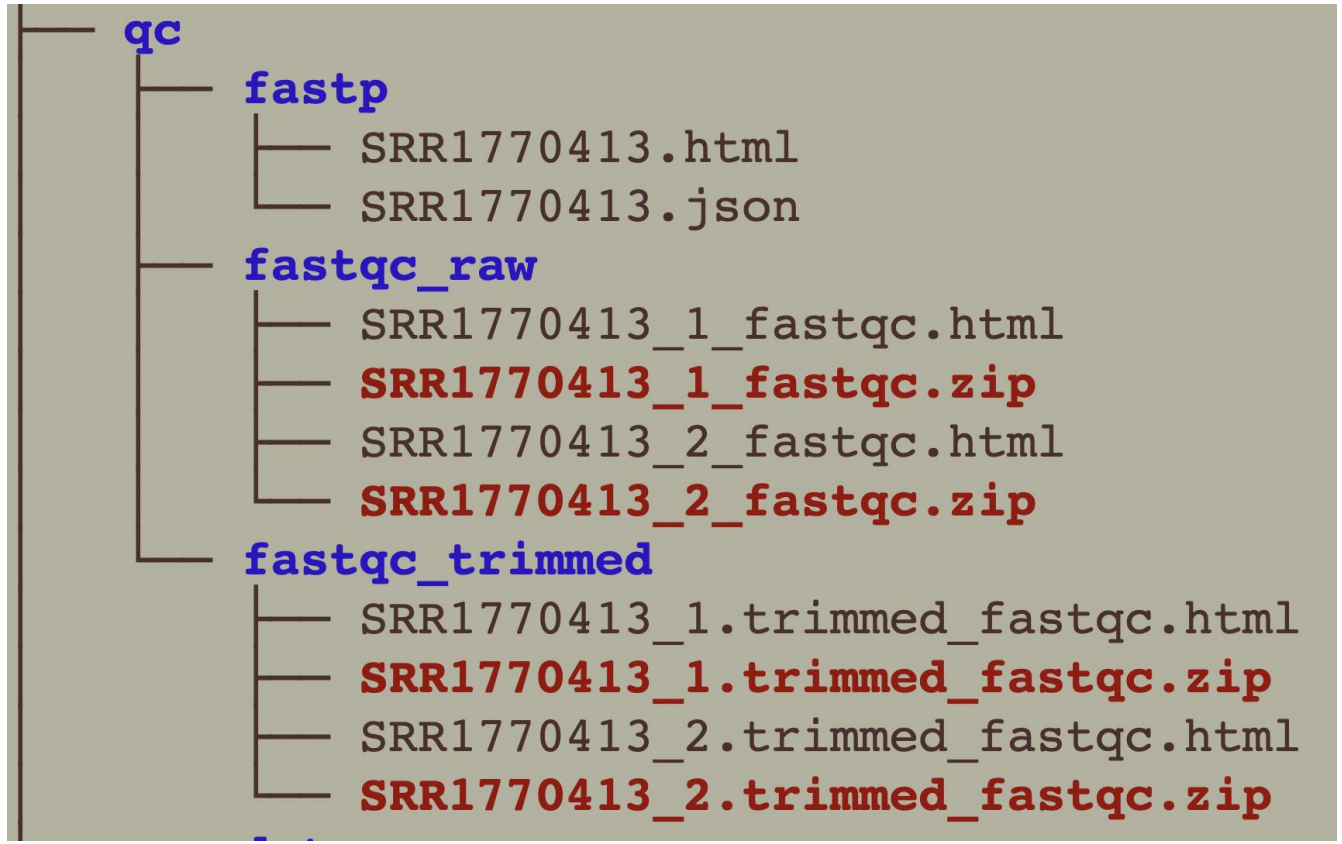
```
fastqc *trim*.fq.gz -o fastqc_trimmed/ -t 8
```

Step 6: Aggregate all QC reports

```
multiqc . -o multiqc_report/
```

Output

QC Results



Processed Data (cleaned)