HKU SPACE and Edinburgh Napier University
Msc Biomedical Science
Scientific Skills Module

Data handling on Bioinformatics II

GU Haogao
The University of Hong Kong
guhaogao@hku.hk

2024-11-08

The Learning Objective

Data handling on Bioinformatics

- Analyse and interpret DNA and protein sequence data using bioinformatics.
  - Knowing how DNA and protein sequences are stored as data
  - Knowing how to create and manipulate such data
  - How to use existing tools to boost your analysis
  - Standard practice and advices for bioinformatics projects
- In these two lessons, we will learn:
  - **General data skills** that applicable to a wide range of (bio)informatics projects.

Rundown

Lesson 2

- Storing sequencing data (50 mins):
  - Recap on sequencing
  - fasta/fastq format
  - Nucleotide and amino acid codes
- Break (10 mins)
- Working with sequence data (50 mins)
  - Example: reading sequences data and counting nucleotides
  - Example: trimming low-quality bases
- Break (10 mins)
- Phylogenetic analysis and application (30 mins)

Sequencing

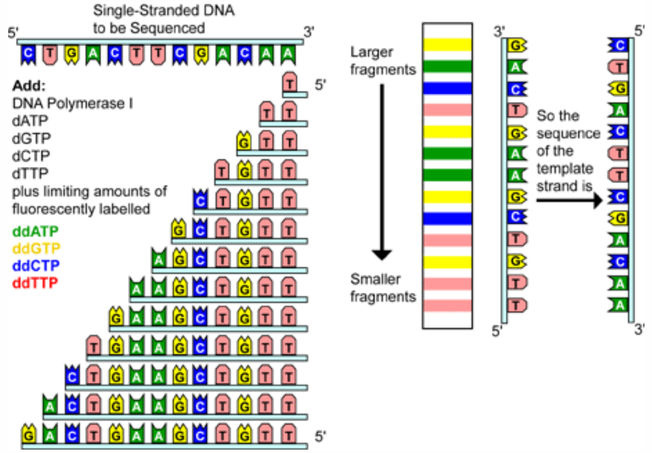
DNA Sequencing

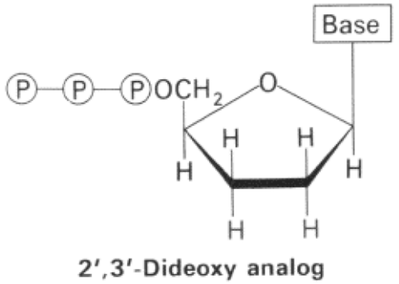
- Sanger sequencing
  - Chain termination method
  - Also known as Sequencing **after** elongation
  - Requires the construction of millions of sequences with different sizes

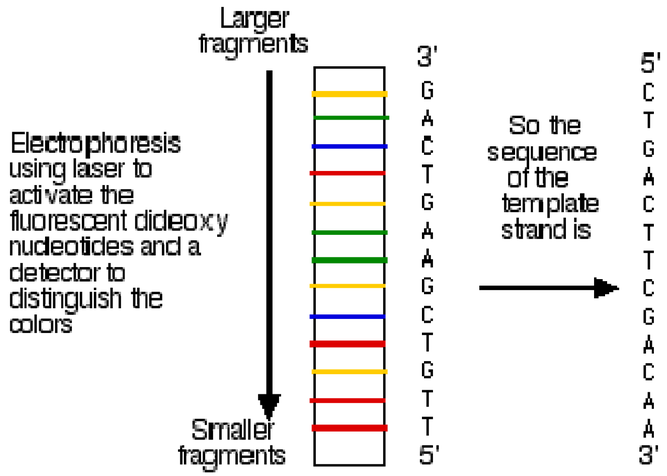
Sanger Method

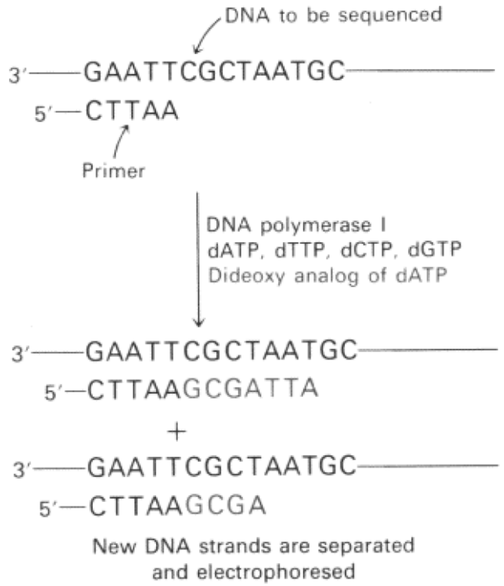
- Uses dideoxy-nucleotides (ddNTP) that inhibits DNA elongation after integration
- Requires primer, DNA polymerase, template, mixture of deoxyribonucleic acids (dNTP) and ddNTP
- Incorporation of ddNTP into growing strand to terminate DNA synthesis during **cycle sequencing**
- The synthesized strand sizes are determined using capillary electrophoresis

Sanger sequencing





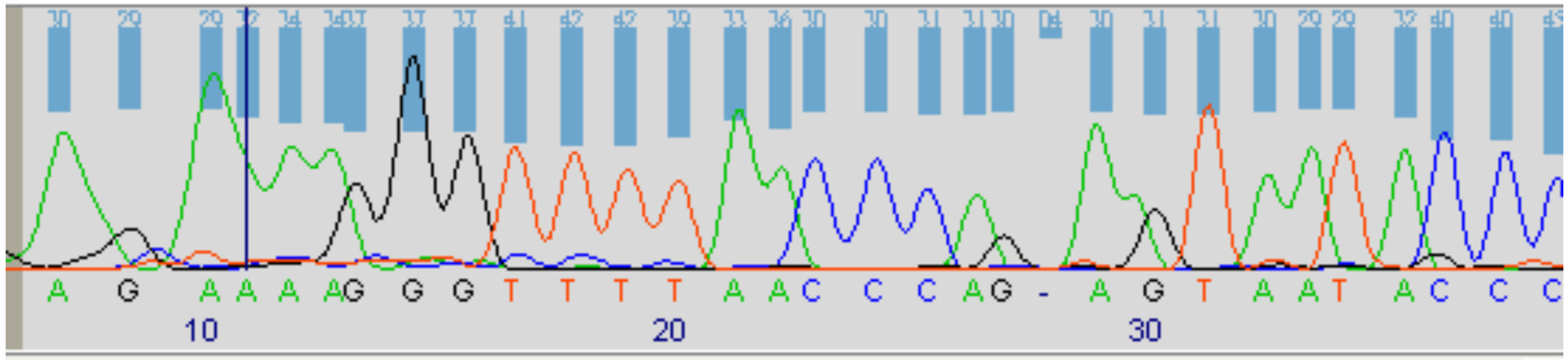




Previous note by Kingsley Tam

Why do we have to do manual sequencing proof-reading and editing?

What is the right sequence at this particular position????????

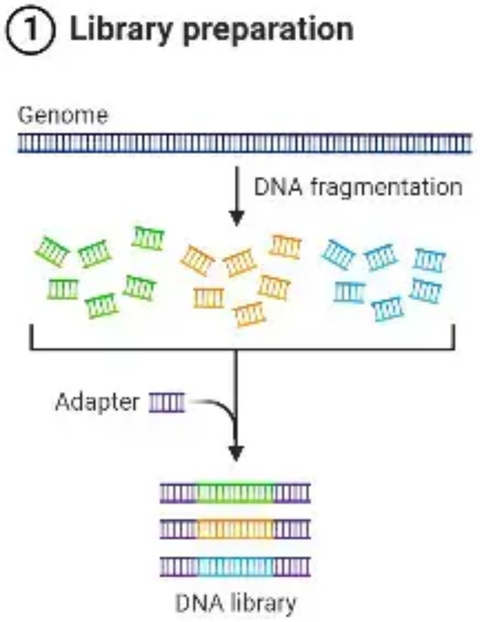


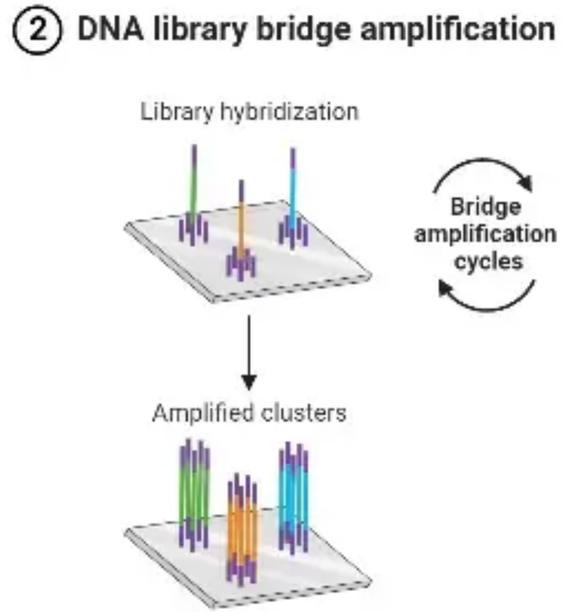
Previous note by Kingsley Tam

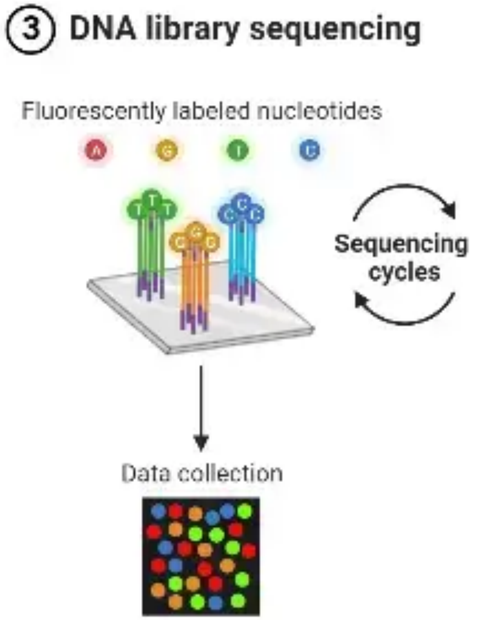
NGS

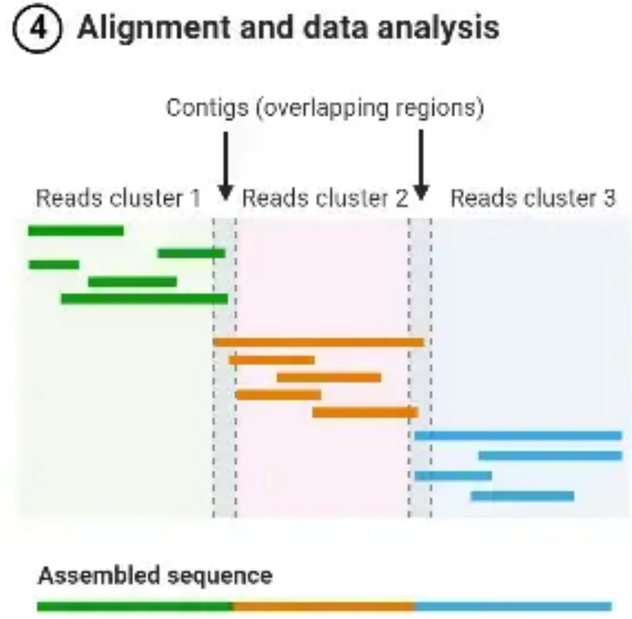
High-throughput sequencing

- Sequencing Process:
  - NGS: Massively parallel, short reads, needs assembly.
  - Sanger: Chain termination, longer reads, direct sequence.
- Read Lengths:
  - NGS: Short reads (50–300 bp), high volume.
  - Sanger: Long reads (up to 1000 bp), single fragments.





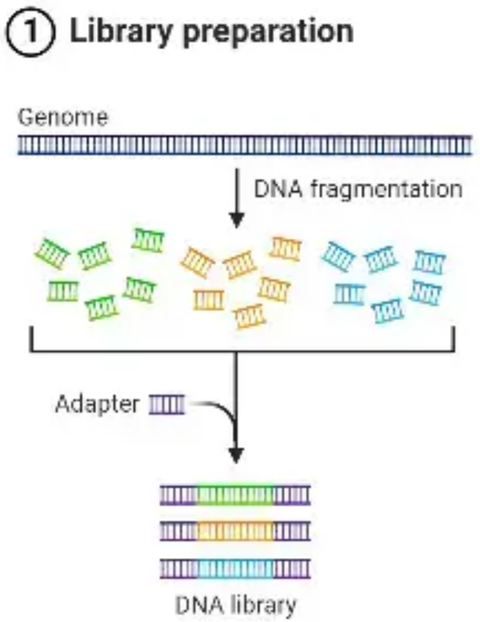


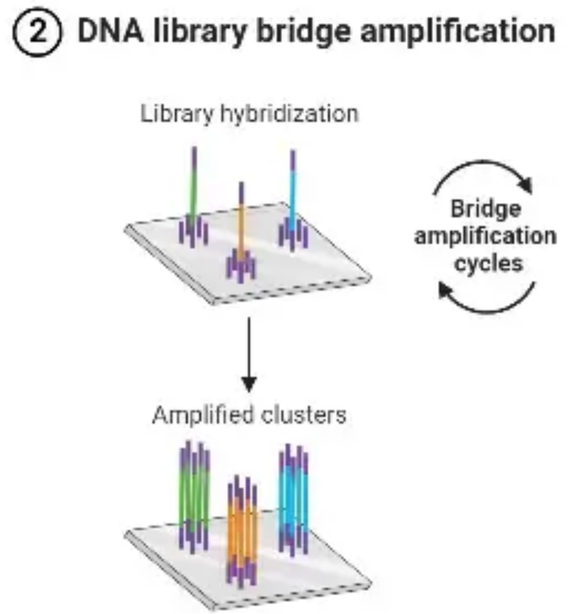


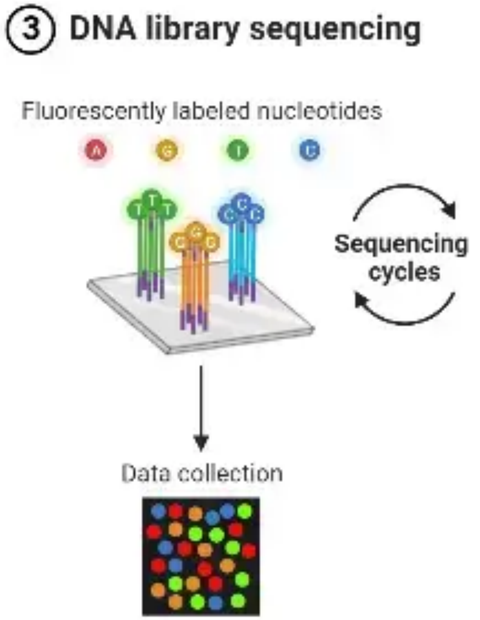
NGS

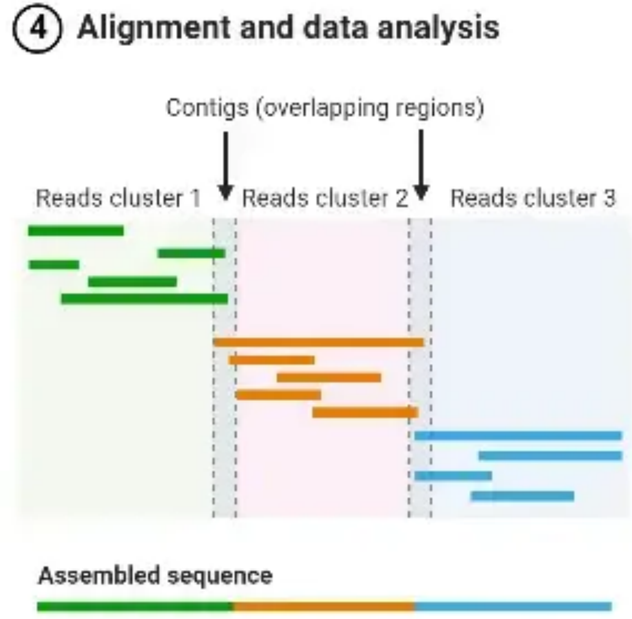
High-throughput sequencing

- Data Detection:
  - NGS: Real-time base detection, needs bioinformatics for assembly.
  - Sanger: Fluorescent chain-termination, minimal assembly.
- Amplification:
  - NGS: Cluster amplification (bridge or emulsion PCR).
  - Sanger: Traditional PCR, no clustering needed.



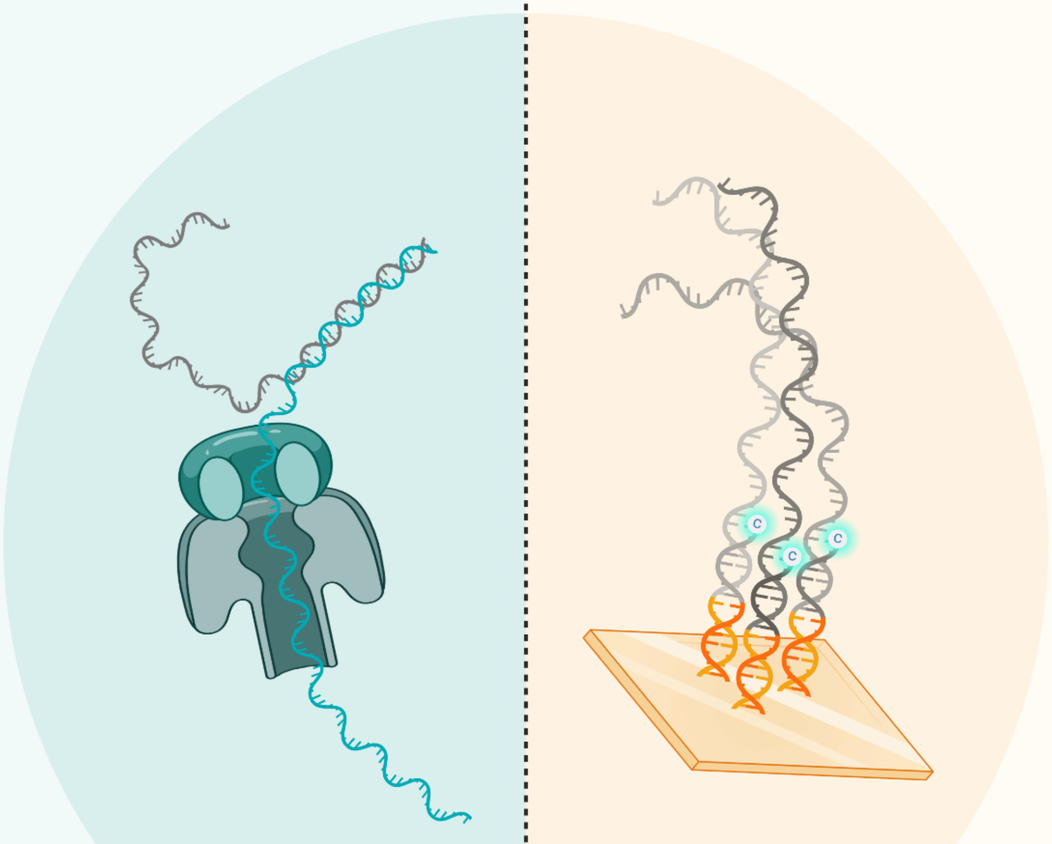






Oxford NANOPORE Technologies

illumina



Fasta and Fastq

Storing sequences

In Fasta format

- When we don't need to record the quality information.
- Only need to record the sequence.
- Two fields:
  - Header
  - Sequence

Storing sequences

In Fasta format

- Pros:
  - Simple
- Cons:
  - No universal specification for the format of an header
- Use a strict naming convention and be consistent.

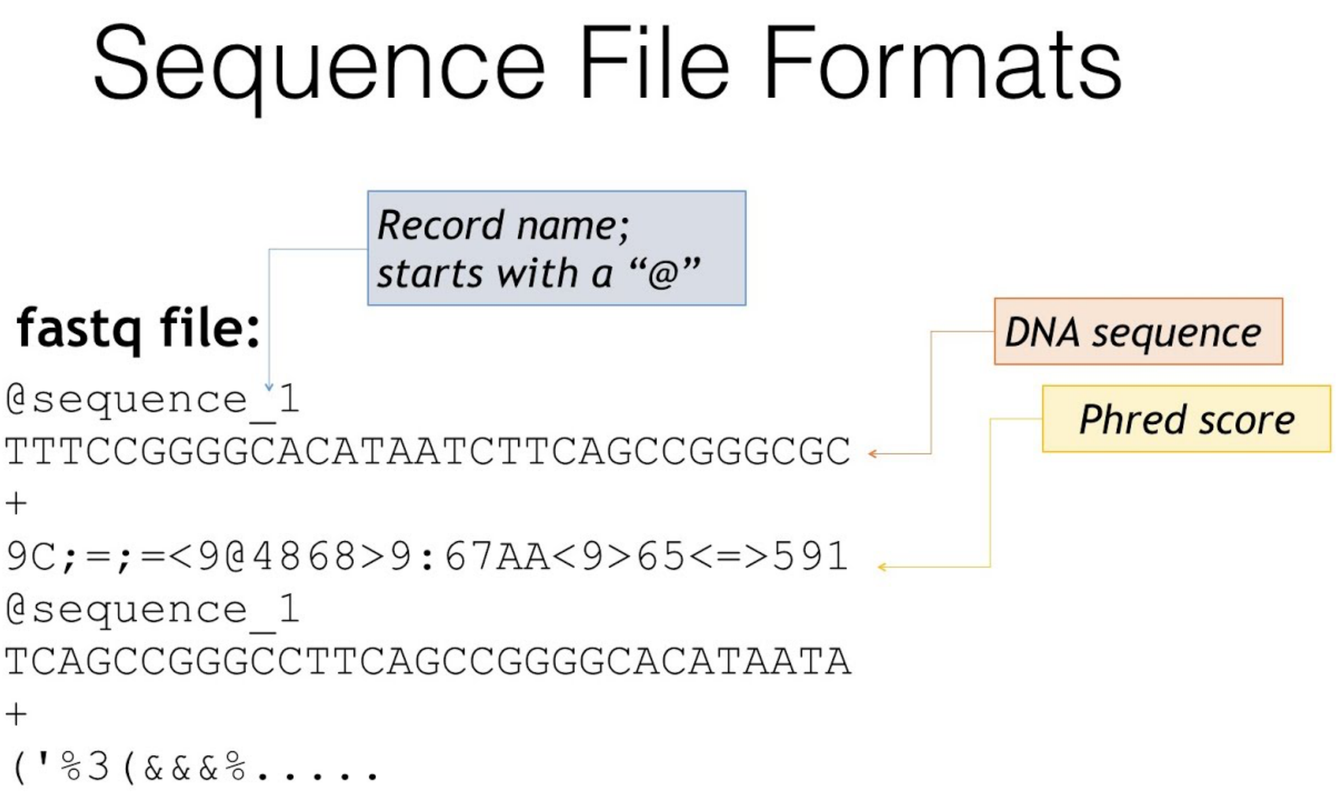
```
>ENSMUSG00000020122|ENSMUST00000138518
> ENSMUSG00000020122|ENSMUST00000125984
>ENSMUSG00000020122|ENSMUST00000125984|epidermal growth factor receptor
>ENSMUSG00000020122|ENSMUST00000125984|Egfr
>ENSMUSG00000020122|ENSMUST00000125984|11|ENSFM00410000138465
```

```
>gene_00284728 length=231;type=dna
GAGAACTGATTCTGTTACCGCAGGGCATTCGGATGTGCTAAGGTAGTAATCCATTATAAGTAACATGCGCGGAATATCCG
GAGGTCATAGTCGTAATGCATAATTATTCCCTCCCTCAGAAGGACTCCCTTGCGAGACGCCAATACCAAAGACTTTCGTA
GCTGGAACGATTGGACGGCCCAACCGGGGGGAGTCGGCTATACGTCTGATTGCTACGCCTGGACTTCTCTT
```

Storing sequences

In Fastq format

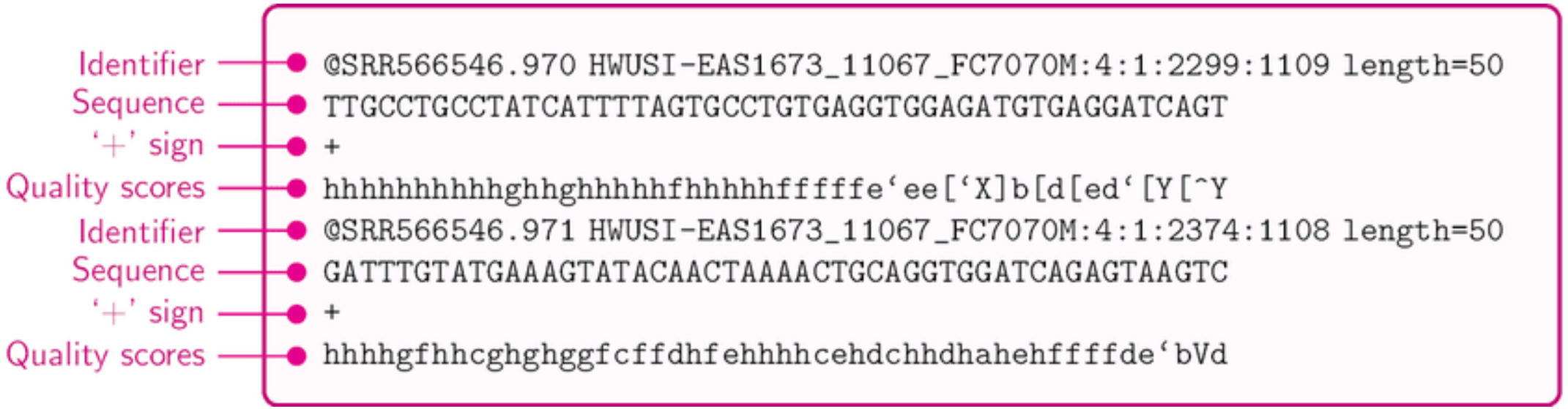
- The FASTQ format extends FASTA by including a numeric quality score to each base in the sequence.
- The FASTQ format is widely used to store high-throughput sequencing data, which is reported with a per-base quality score indicating the confidence of each base call.
- Unfortunately like FASTA, FASTQ has variants and pitfalls that can make the seemingly simple format frustrating to work with.



Storing sequences

In Fastq format

1. The description line, beginning with @. This contains the record identifier and other information.
2. Sequence data, which can be on one or many lines.
3. The line beginning with +, following the sequence line(s) indicates the end of the sequence. In older FASTQ files, it was common to repeat the description line here, but this is redundant and leads to unnecessarily large FASTQ files.
4. The quality data, which can also be on one or many lines, but must be the same length as the sequence. Each numeric base quality is encoded with ASCII characters using a particular scheme.

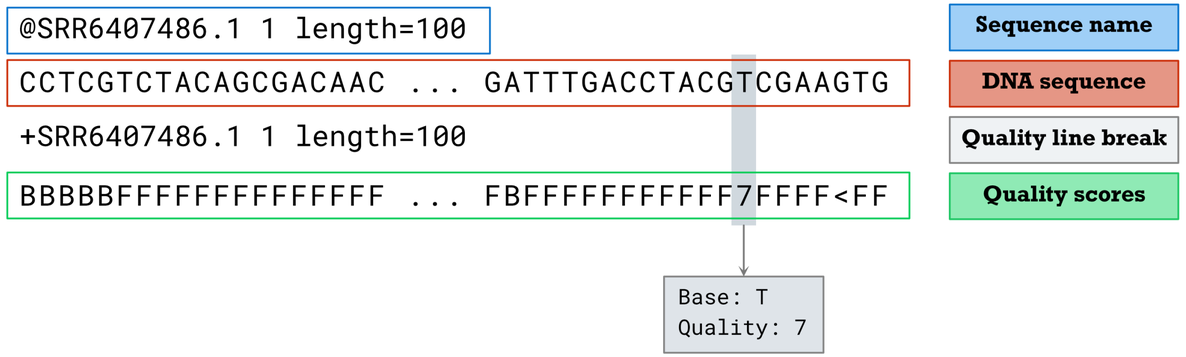


Parsing FASTQ Files Correctly

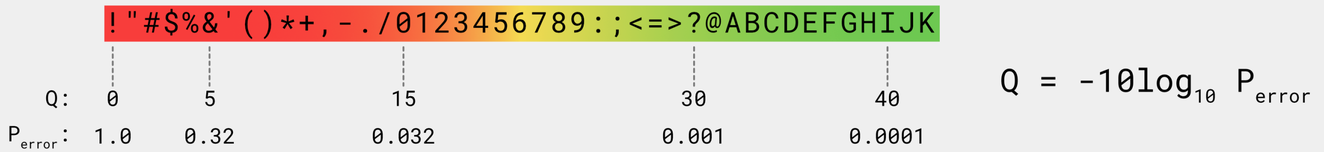
- Common Pitfall: Treating every line starting with @ as a header line is incorrect.
- Quality Scores: @ is a valid quality character, so it may appear in quality lines.
- Line Wrapping: Sequence and quality lines can wrap, so position alone is unreliable.
- Solution: Ensure the number of quality score characters matches sequence characters for accurate parsing.

FASTQ file sample:

```
@SRR6407486.1 1 length=100
CCTCGTCTACAGCGACAACGTCCAGACCCGCGAACGGGTGATGCGGGCCCTGGGCAAACGGTTGCACCCGGATCTGCCCGATTTGACCTACGTCGAAGTG
+SRR6407486.1 1 length=100
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF<FFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFBFFFFFFFFFF7FFFF<FF
```



Quality scores as ASCII characters:



Nucleotide codes

IUPAC scheme

- A, T, C, G represent the nucleotides adenine, thymine, cytosine, and guanine.
- Degenerate (or ambiguous) nucleotide codes are used to represent two or more bases.
- For example, N is used to represent any base. The International Union of Pure and Applied Chemistry (**IUPAC**) has a standardized set of nucleotides, both unambiguous and ambiguous.

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

Codon table

coding for amino acids

- **Codon**: A sequence of three nucleotides coding for a specific amino acid or stop signal.
- **Codon Table**: Maps 64 codons to amino acids or stop signals, essential for interpreting the genetic code.
- **Redundancy**: Multiple codons can code for the same amino acid (e.g., leucine has six codons).
- **Start Codon**: AUG, signaling the beginning of translation (codes for methionine).
- **Stop Codons**: UAA, UAG, UGA, marking the end of protein synthesis.
- **Universal Code**: Nearly universal across all organisms, with few exceptions in mitochondria and some microorganisms.

Standard DNA codon table [edit]

Amino-acid biochemical properties	Nonpolar (np)	Polar (p)	Basic (b)	Acidic (a)	Termination: stop codon *	Initiation: possible start codon ⇒

Standard genetic code[17][note 3]

1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	(Phe/F) Phenylalanine (np)	TCT	(Ser/S) Serine (p)	TAT	(Tyr/Y) Tyrosine (p)	TGT	(Cys/C) Cysteine (p)	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine (np)	TCA		TAA	Stop (*Ochre*) *[note 2]	TGA	Stop (*Opal*) *[note 2]	A
	TTG ⇒		TCG		TAG	Stop (*Amber*) * [note 2]	TGG	(Trp/W) Tryptophan (np)	G
C	CTT		CCT	(Pro/P) Proline (np)	CAT	(His/H) Histidine (b)	CGT	(Arg/R) Arginine (b)	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine (p)	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine (np)	ACT	(Thr/T) Threonine (p)	AAT	(Asn/N) Asparagine (p)	AGT	(Ser/S) Serine (p)	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine (b)	AGA	(Arg/R) Arginine (b)	A
	ATG ⇒	(Met/M) Methionine (np)	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine (np)	GCT	(Ala/A) Alanine (np)	GAT	(Asp/D) Aspartic acid (a)	GGT	(Gly/G) Glycine (np)	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid (a)	GGA		A
	GTG ⇒		GCG		GAG		GGG		G

Base qualities

Schemes

- Phred+33 (Sanger): Commonly used in Illumina and modern NGS data. Scores range from 0 to 93, with ASCII characters ! to ~.
- Phred+64 (Illumina 1.3–1.7): Older Illumina format, now mostly obsolete. Scores range from 0 to 62, with ASCII characters @ to h.
- Solexa/Illumina: Used in early Solexa sequencing. Negative scores are possible, less common today.
- Ensure compatibility with software by knowing the scheme used.
- Quality Interpretation: Higher scores = better quality (probability of correct base call).

Table 10-2. FASTQ quality schemes (adapted from Cock et al., 2010 with permission)

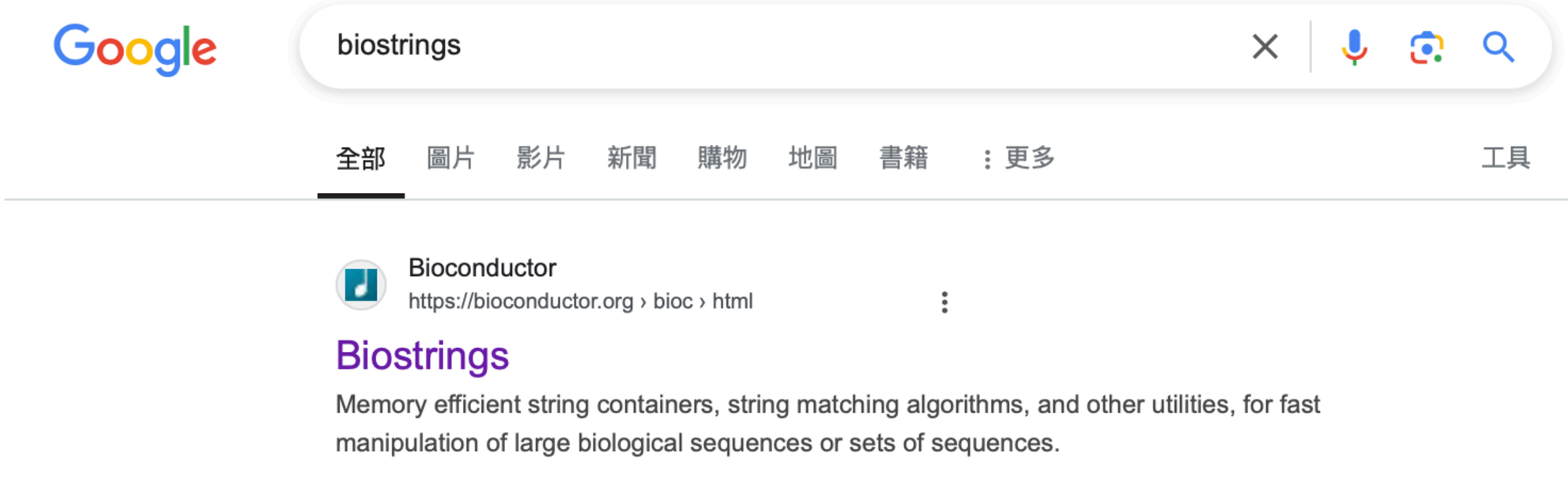
Name	ASCII character range	Offset	Quality score type	Quality score range
Sanger, Illumina (versions 1.8 onward)	33–126	33	PHRED	0–93
Solexa, early Illumina (before 1.3)	59–126	64	Solexa	5–62
Illumina (versions 1.3–1.7)	64–126	64	PHRED	0–62

Working with sequencing data in R

Google colab

New notebook

- Create a new notebook with proper name.
- Change the runtime to R.
- Install the “Biostrings” library.

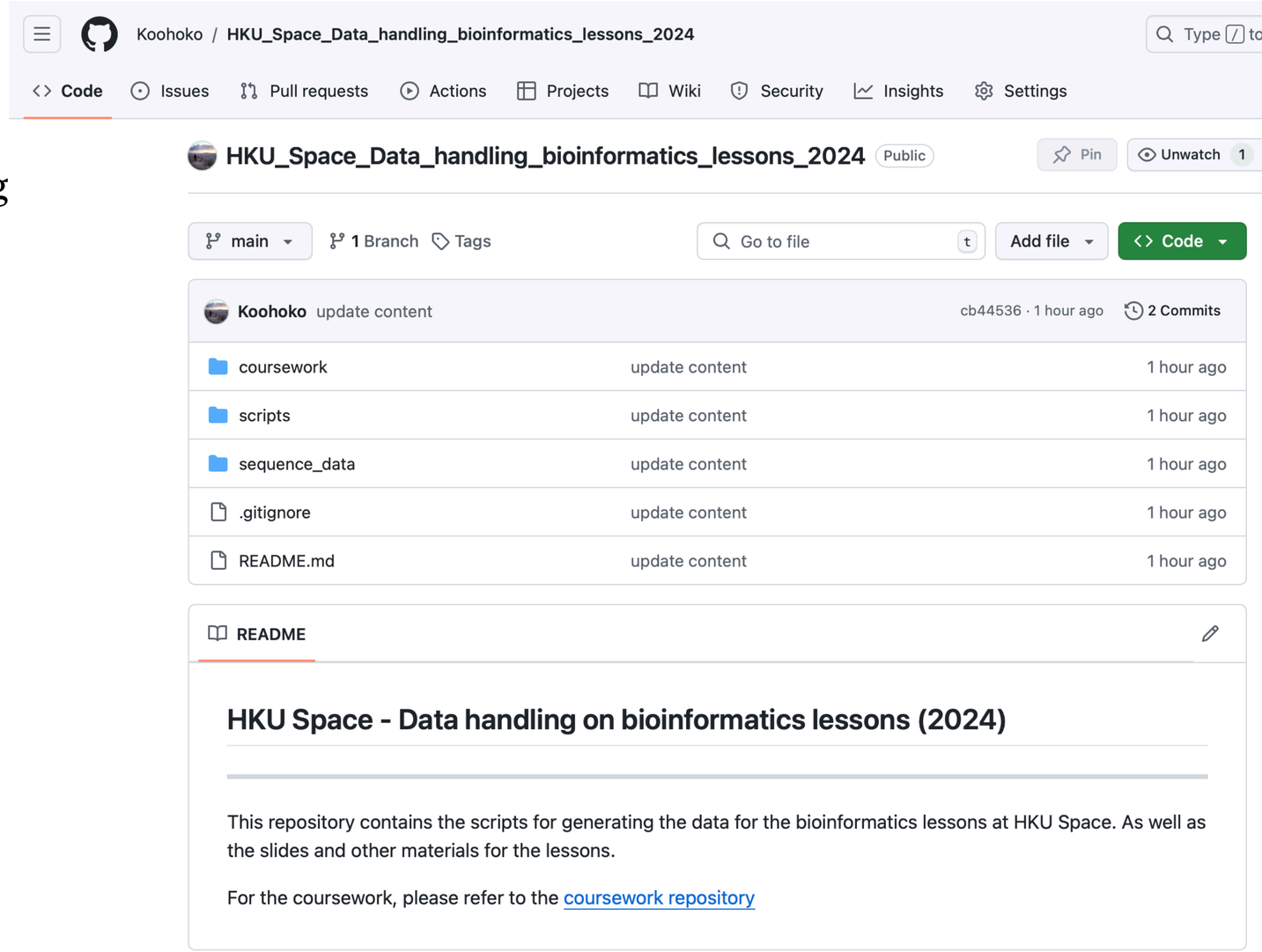


https://bit.ly/40DmtJm

Github repository

for this lesson

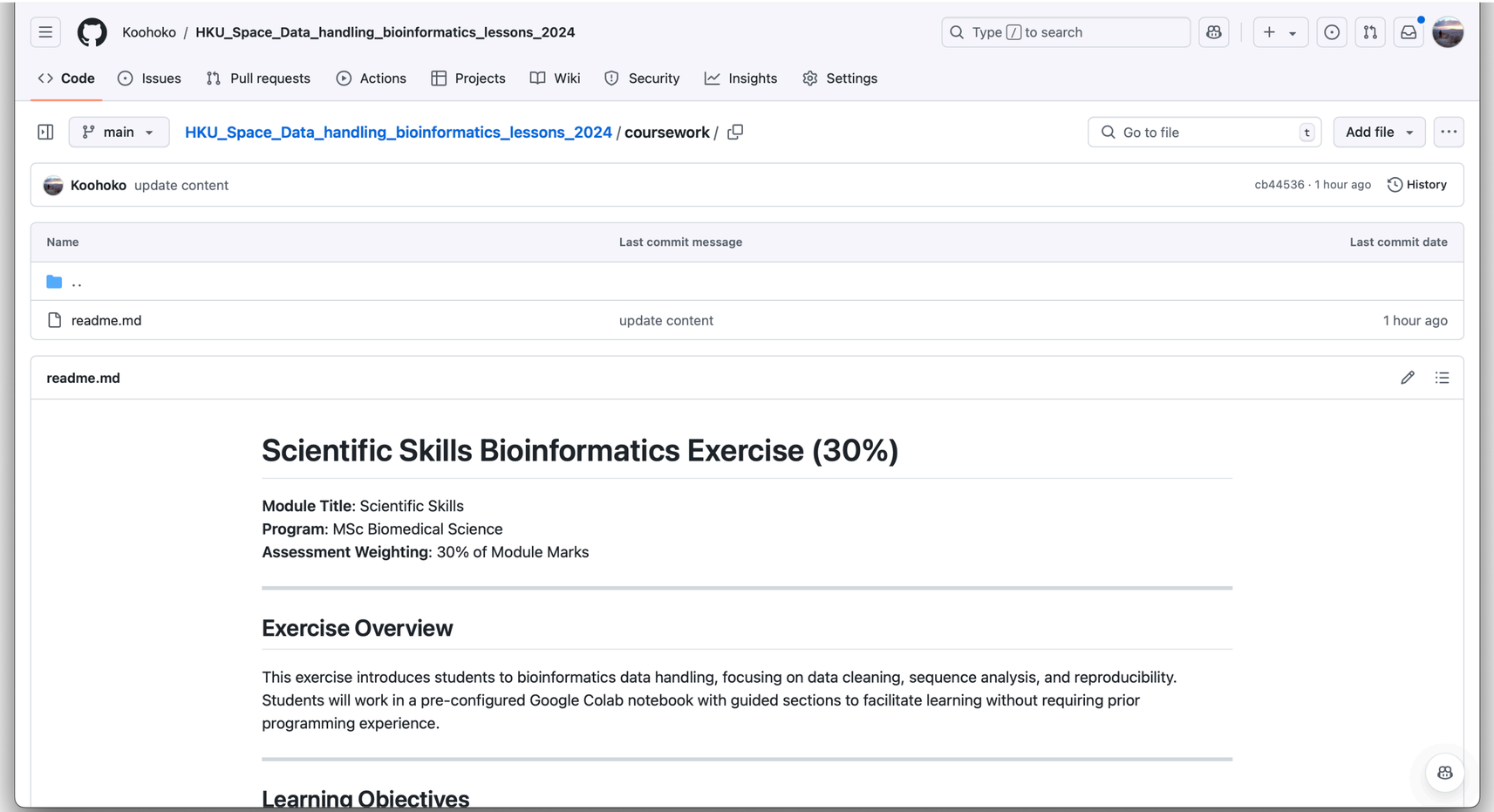
- **Access Course Resources**: All materials, including scripts and datasets, are hosted on GitHub.
- **Repository Structure**: Organized for easy navigation, with separate folders for code, data, and documentation.
- **Hands-On Practice**: Follow examples and exercises directly from the provided resources.
- **Version Control**: Using GitHub ensures that you always have access to the latest updates and materials.
- **Contribute and Collaborate**: Students are encouraged to explore and even contribute if they have ideas or improvements.



Github repository

Important resources

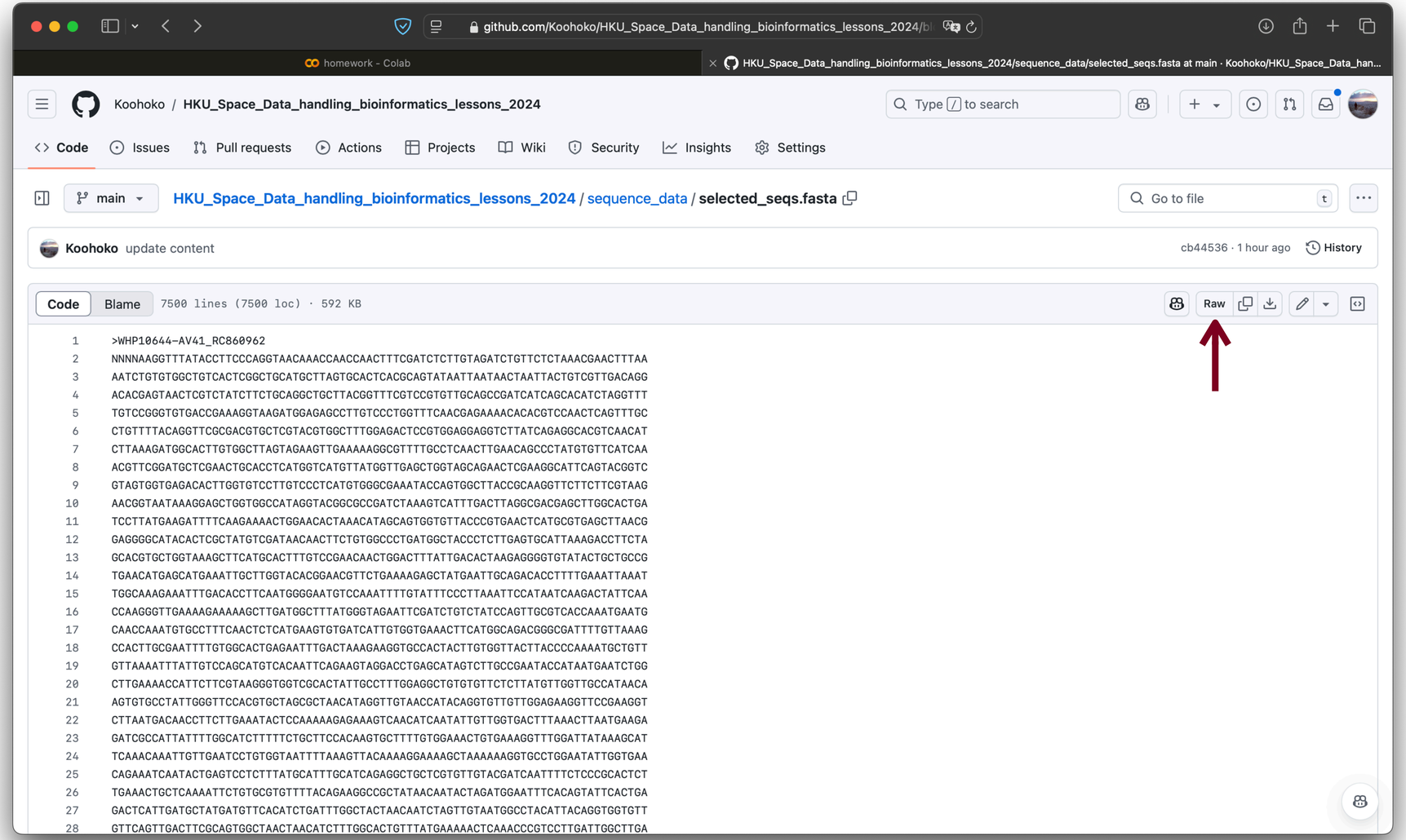
- Coursework requirements.
- Input data.
- Reference scripts.



Github repository

Important resources

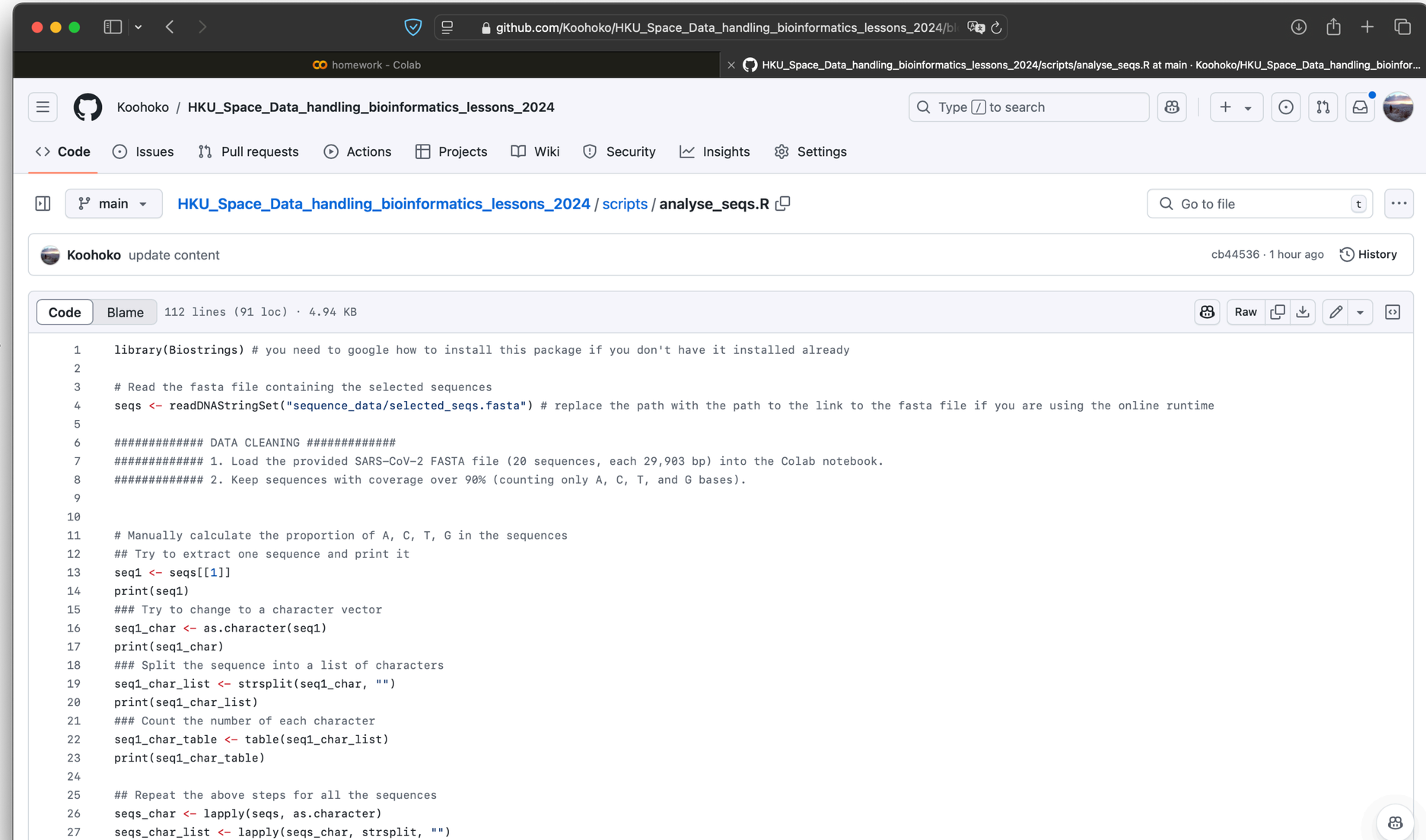
- Coursework requirements.
- Input data.
- Reference scripts.



Github repository

Important resources

- Coursework requirements.
- Input data.
- Reference scripts.

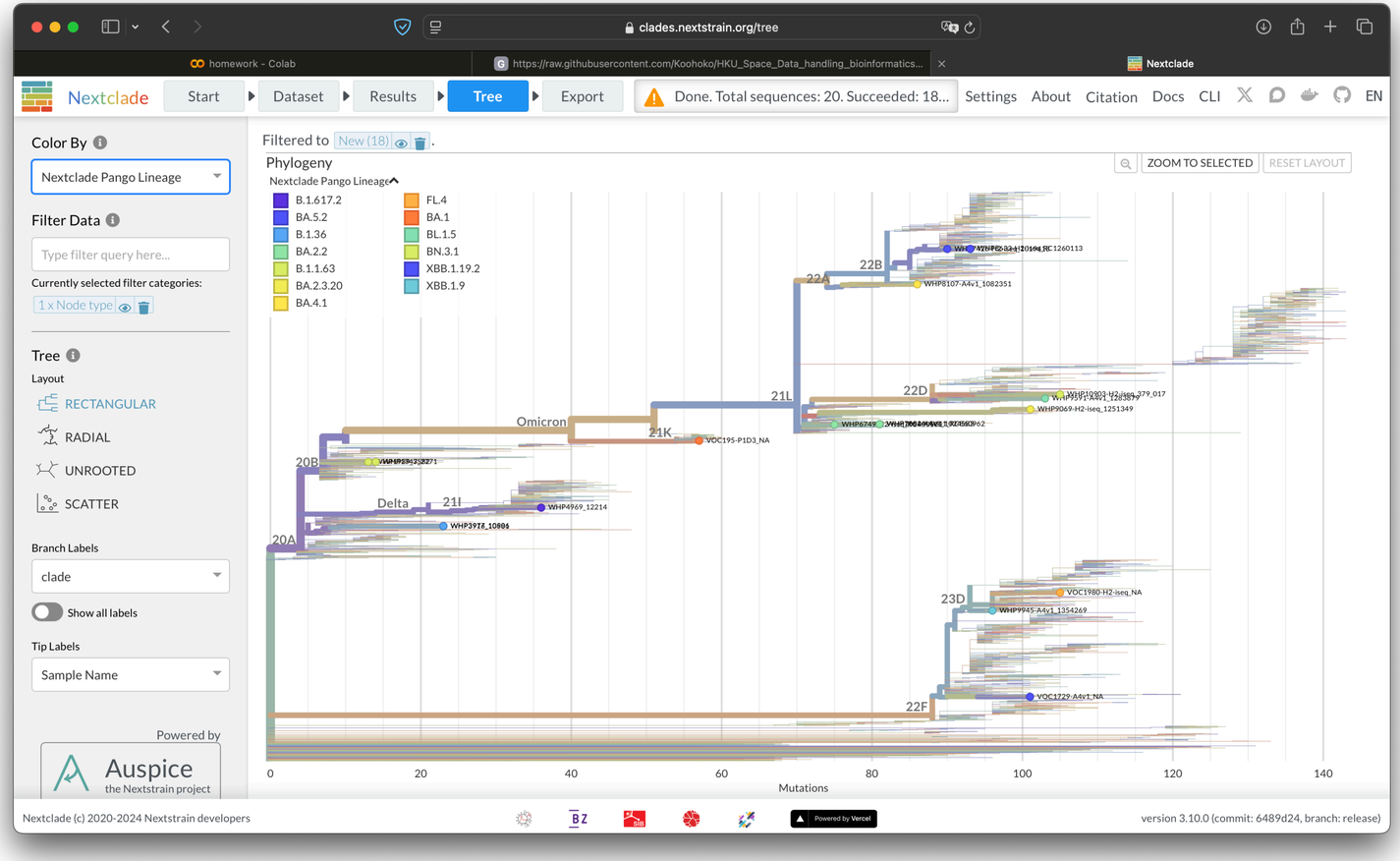


Checking the sequences using nextclade

Nextclade

online tool

- **Purpose**: Nextclade is a tool for the analysis of SARS-CoV-2 genetic sequences, focusing on the assignment of variants and detection of mutations.
- **Variant Classification**: It classifies viral genomes into clades and assigns them to specific SARS-CoV-2 lineages based on mutation patterns.
- **Mutation Detection**: Nextclade identifies both synonymous and non-synonymous mutations, helping to track the evolution of the virus.
- **Alignment-Based**: The tool compares input sequences to a reference genome, providing insights into the sequence variation.



End of lesson II