HKU SPACE and Edinburgh Napier University
Msc Biomedical Science
Applied Medical Microbiology Module

Bioinformatics Tutorial on Medical Microbiology

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

2024-11-30

The Learning Objective

Bioinformatics Tutorial on Medical Microbiology

- Objective: applying bioinformatics to the investigation and diagnosis of infectious diseases.
- Previous knowledge:
  - laboratory techniques (including DNA sequencing) relevant to medical microbiology.
- In these tutorial, we will learn:
  - Reading **Sanger sequencing data** and convert to fasta sequences for mutation analysis.
  - Reading **Next-generation sequencing data** and convert to fasta sequences for mutation analysis.

Rundown

Bioinformatics tutorial

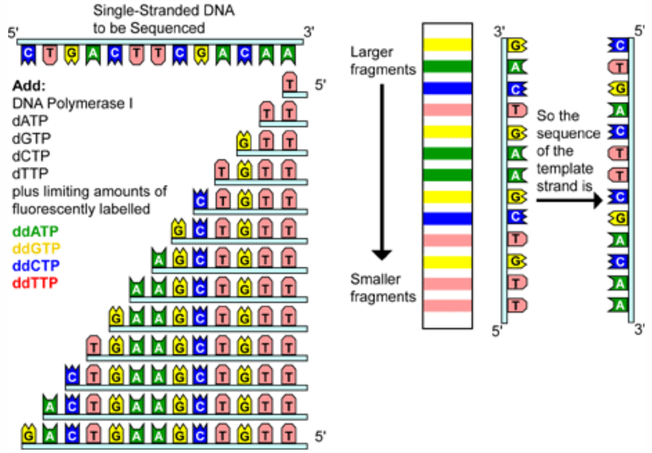
- Recap, samples, data, tools (10 mins)
- Visualising Sanger Sequencing Data (20 mins)
- Generating fasta files for Sequencing Data (30 mins)
- Break (10 mins)
- Checking multiple sequence alignment for Sanger sequencing results (20 mins)
- Introducing NGS data and workflow (30 mins)
- Break (10 mins)
- Working on fast data using galaxy (40 mins)
- Analysing mutations for NGS sequencing results (10 mins)
- Free practice, Q&A (60 mins)

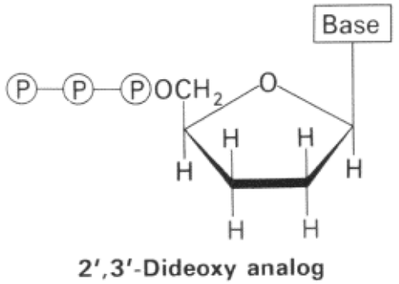
Recap on Sanger sequencing samples

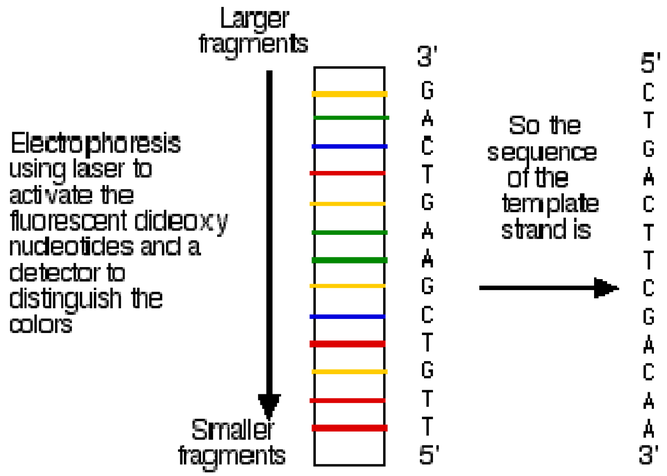
Sanger Method

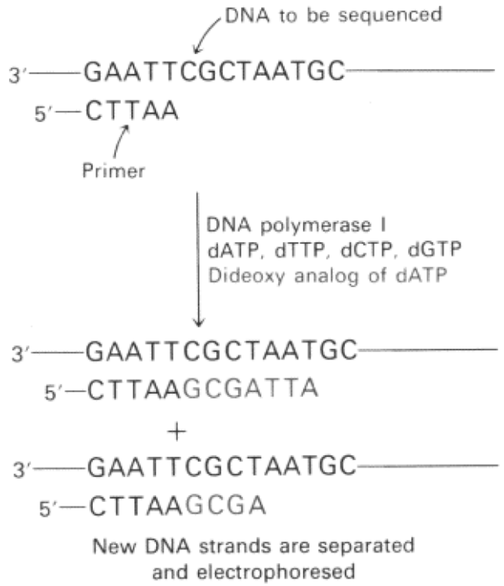
- Uses dideoxy-nucleotides (ddNTP) that inhibits DNA elongation after integration
- Requires primer, DNA polymerase, template, mixture of deoxyribonucleic acids (dNTP) and ddNTP
- Incorporation of ddNTP into growing strand to terminate DNA synthesis during **cycle sequencing**
- The synthesized strand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

Sanger sequencing





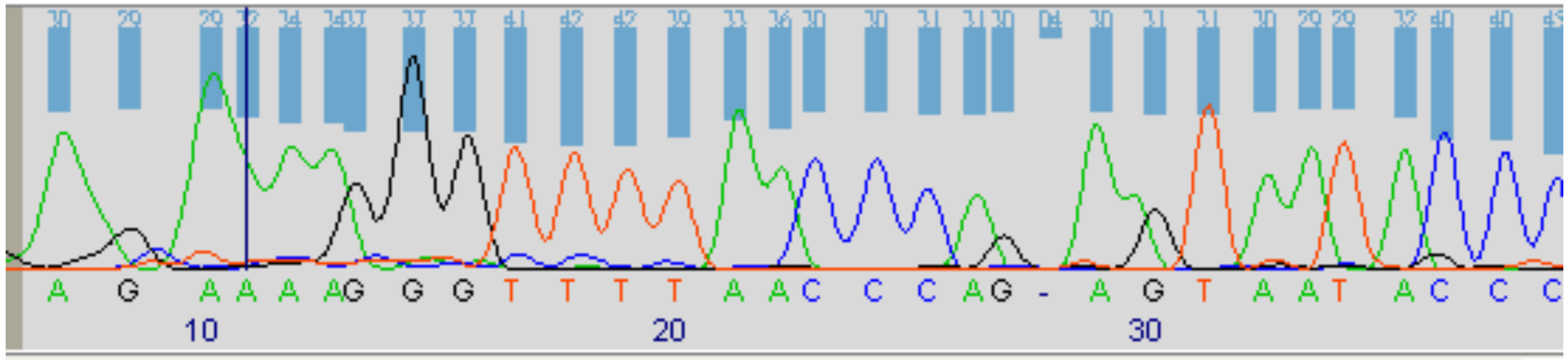




Previous note by Kingsley Tam

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

Samples

from previous the lesson

- We sent the samples of two group (Group B2, D2) for the reactions.
- Sanger sequencing of the HA and NA gene of one H1N1 and one H3N2 sample.
- HA (3 sequencing reactions): HA-universal Forward primer (HA-1F), H1 / H3 internal forward primer (336F), Universal reverse primer (NS-890R)
- NA (2 sequencing reactions): NA-universal forward and reverser primers (NA-1F, NA-1413R)

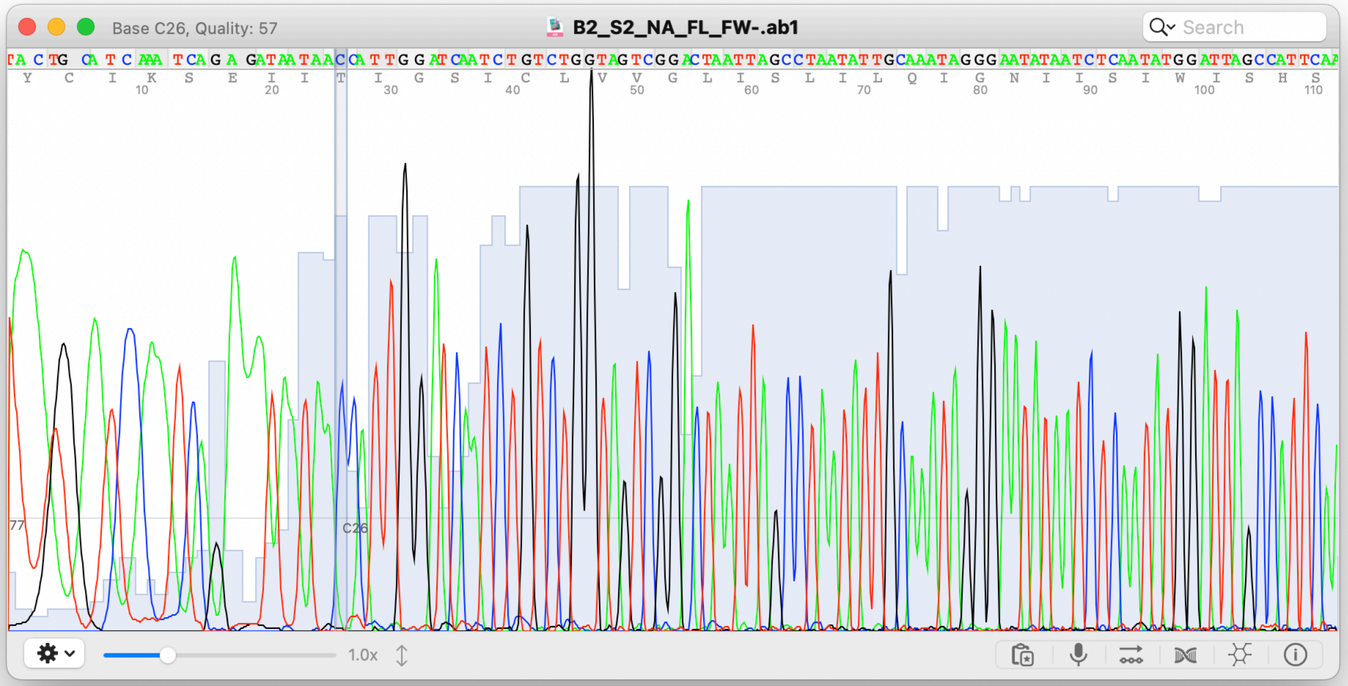
Name	Date Modified	Size	Kind
B2_S2_H1_HA_Int_FW-.ab1	19 Nov 2024 at 23:34	276 KB	ABI
B2_S2_HA_FL_FW-.ab1	19 Nov 2024 at 23:34	273 KB	ABI
B2_S2_HA_FL_RV-.ab1	19 Nov 2024 at 23:34	272 KB	ABI
B2_S2_NA_FL_FW-.ab1	19 Nov 2024 at 23:34	270 KB	ABI
B2_S2_NA_FL_RV-.ab1	19 Nov 2024 at 23:34	271 KB	ABI
B2_S3_H3_HA_Int_FW-.ab1	19 Nov 2024 at 23:34	274 KB	ABI
B2_S3_HA_FL_FW-.ab1	19 Nov 2024 at 23:34	249 KB	ABI
B2_S3_HA_FL_RV-.ab1	19 Nov 2024 at 23:34	267 KB	ABI
B2_S3_NA_FL_FW-.ab1	19 Nov 2024 at 23:34	273 KB	ABI
B2_S3_NA_FL_RV-.ab1	19 Nov 2024 at 23:34	274 KB	ABI
D2_S2_H1_HA_Int_FW-.ab1	19 Nov 2024 at 23:34	271 KB	ABI
D2_S2_HA_FL_FW-.ab1	19 Nov 2024 at 23:34	276 KB	ABI
D2_S2_HA_FL_RV-.ab1	19 Nov 2024 at 23:34	272 KB	ABI
D2_S2_NA_FL_FW-.ab1	19 Nov 2024 at 23:34	270 KB	ABI
D2_S2_NA_FL_RV-.ab1	19 Nov 2024 at 23:34	272 KB	ABI
D2_S3_H3_HA_Int_FW-.ab1	19 Nov 2024 at 23:34	277 KB	ABI
D2_S3_HA_FL_FW-.ab1	19 Nov 2024 at 23:34	271 KB	ABI
D2_S3_HA_FL_RV-.ab1	19 Nov 2024 at 23:34	273 KB	ABI
D2_S3_NA_FL_FW-.ab1	19 Nov 2024 at 23:34	277 KB	ABI
D2_S3_NA_FL_RV-.ab1	19 Nov 2024 at 23:34	275 KB	ABI

Data and tools

Sanger sequencing

AB1 files

- AB1 files are binary files produced by Applied Biosystems' Sanger sequencing machines.
- They store raw data from sequencing runs, including chromatograms and metadata.
- Contents of an AB1 File:
  - Chromatogram Data: Electropherogram peaks for A, T, C, G.
  - Base Calls: The called DNA sequence.
  - Quality Scores: PHRED scores for each base.
  - Metadata: Information about the sequencing run (e.g., instrument settings, sample info).



Sanger sequencing

AB1 files

- Reading AB1 files:
  - GUI-Based tools:
    - Tools with graphical interfaces are ideal for beginners or non-programmers.
  - Command line tools:
    - Ideal for advanced users and automation. Requires coding knowledge.

Sanger Data Analysis

	Free/Paid	Trace File Viewer	Multiple Sequence Alignment & Contig Assembly	Windows	Mac	Linux	In-browser
4Peaks	Free	✓	-	-	✓	-	-
Sequence Scanner	Free	✓	-	✓	-	-	-
Chromas Lite	Free	✓	-	✓	-	-	-
SnapGene Viewer	Free	✓	-	✓	✓	✓	-
GeneStudio Pro	Free	✓	✓	✓	-	-	-
UGENE	Free	✓	✓	✓	✓	✓	-
Chromaseq	Free	✓	✓	✓	✓	✓	-
BioLign (Note: See bottom half of page for description and download link)	Free	-	✓	✓	-	-	-
CAP3 (Note: Only for contig assembly of sequences in FASTA format)	Free	-	✓	-	-	-	✓
Geneious	Paid	✓	✓	✓	✓	✓	-
SeqMan Pro	Paid	✓	✓	✓	✓	-	-
Sequencher	Paid	✓	✓	✓	✓	-	-
Vector NTI Express	Paid	✓	✓	✓	✓	-	-
Chromas Pro	Paid	✓	✓	✓	-	-	-
CodonCode Aligner	Paid	✓	✓	✓	✓	-	-
DNA Baser	Paid	✓	✓	✓	-	-	-

Analysing Sanger sequencing data via colab

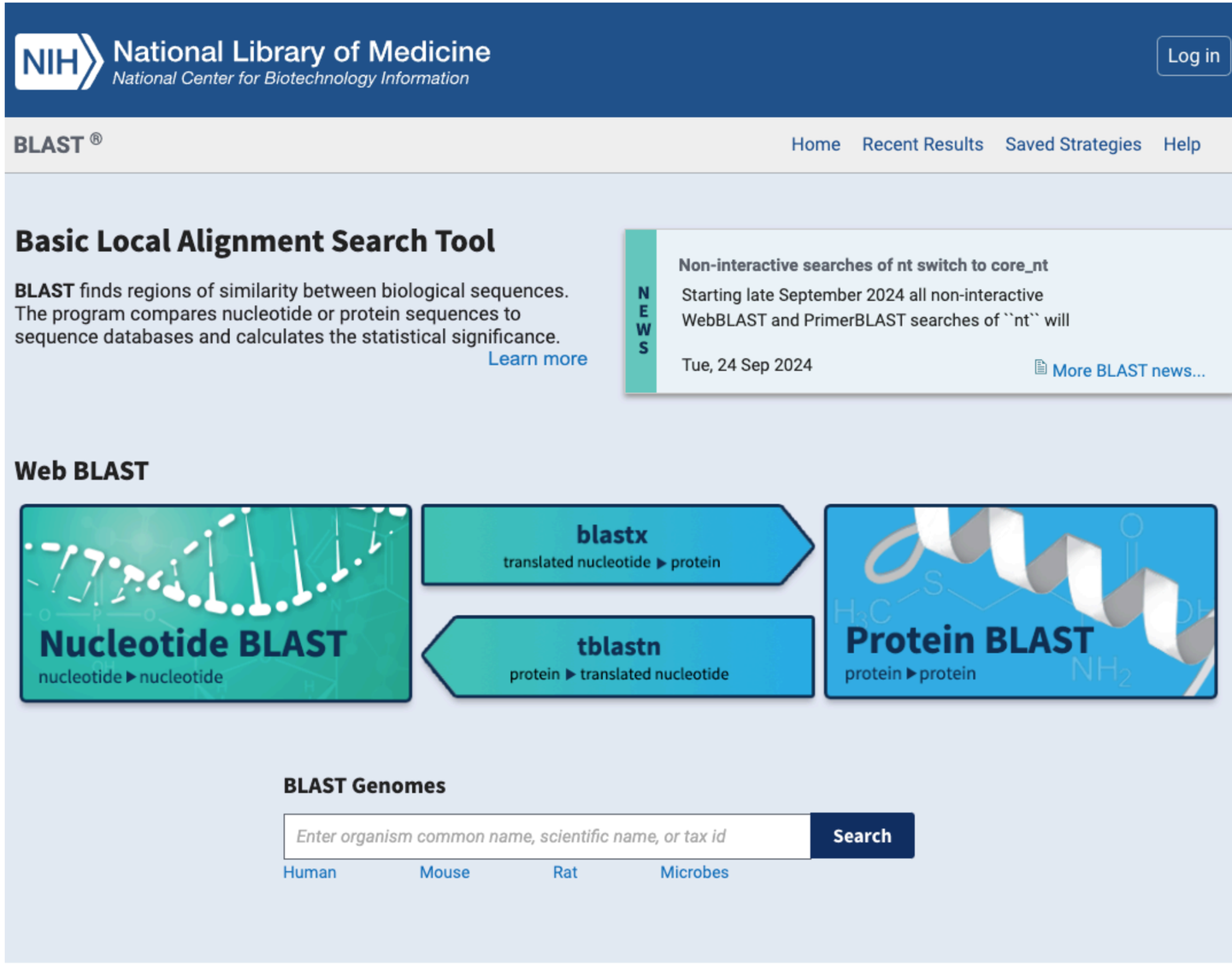
https://bit.ly/AMMBIO

BLAST the fasta

BLAST

trimmed sequences

- **BLAST** (Basic Local Alignment Search Tool)
- **Purpose**: Perform local alignments of your contigs against a reference database to find close relatives.
- **Limitations**: BLAST provides local alignments but does not directly generate a consensus genome.

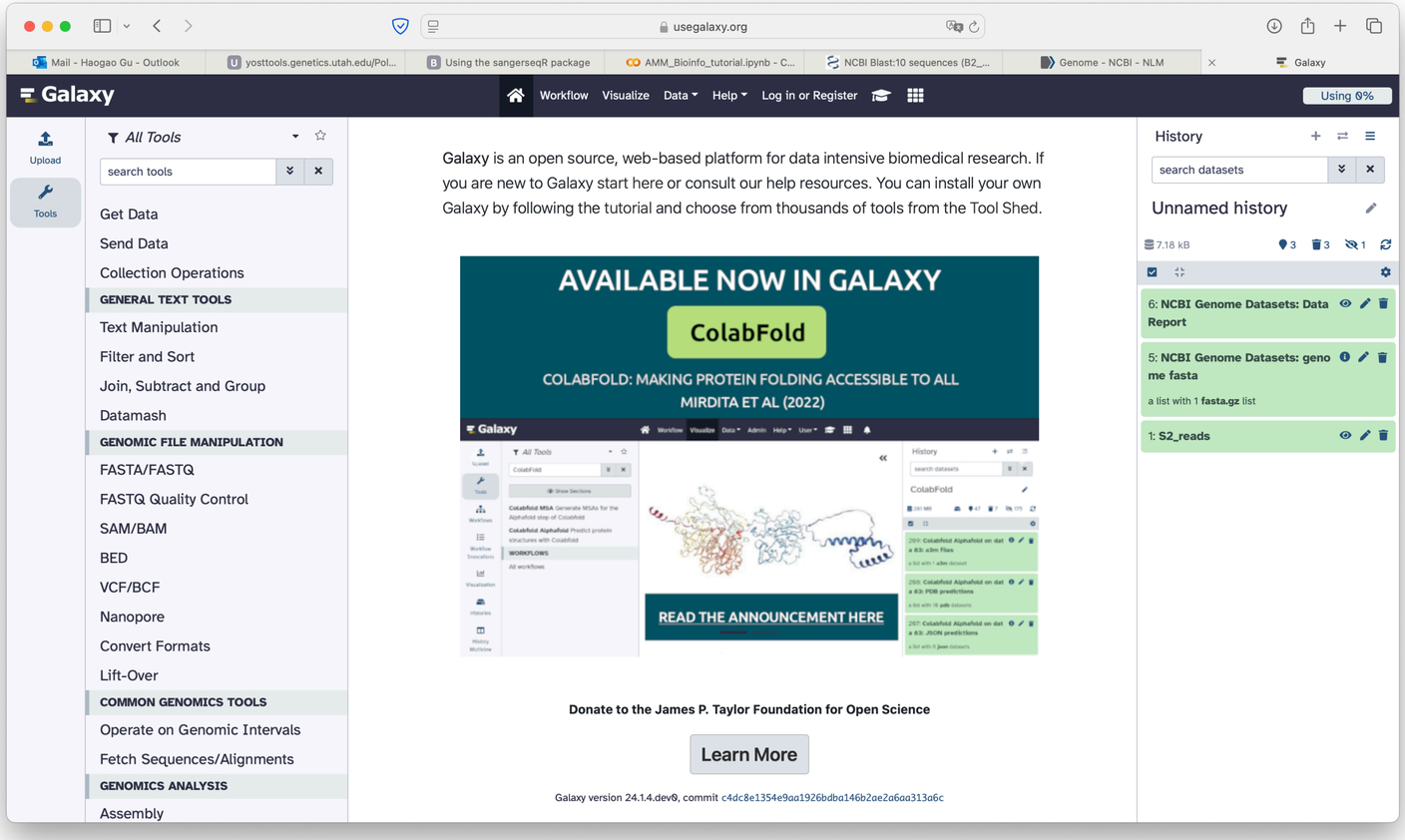


Galaxy

GALAXY

trimmed sequences

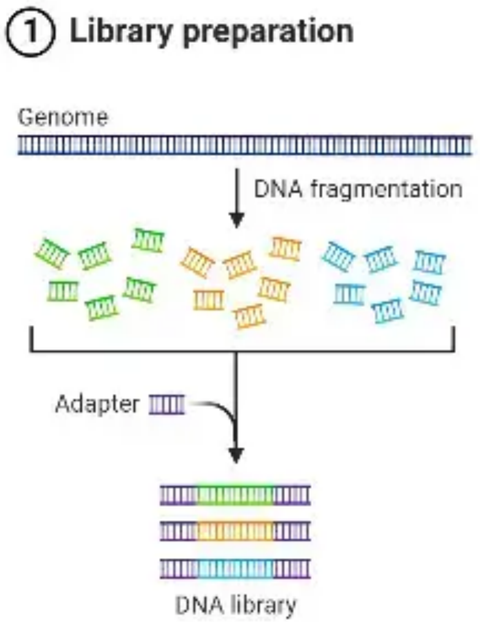
- A online platform for sequence analysis, including mapping, alignment, and downstream.
- Import your own Sanger data.
- Download reference genome data according to blast results.
- Alignment with MAFFT-add.
- Visualise the multiple sequence alignment.

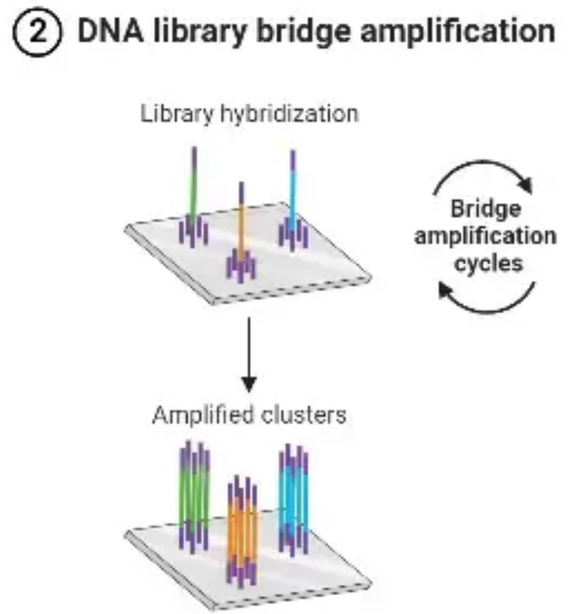


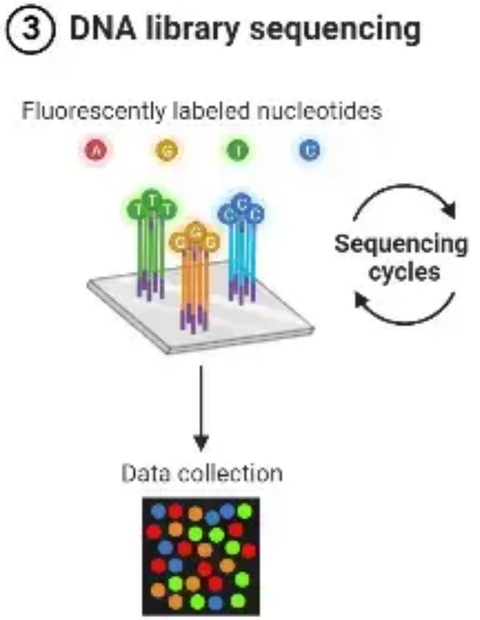
NGS

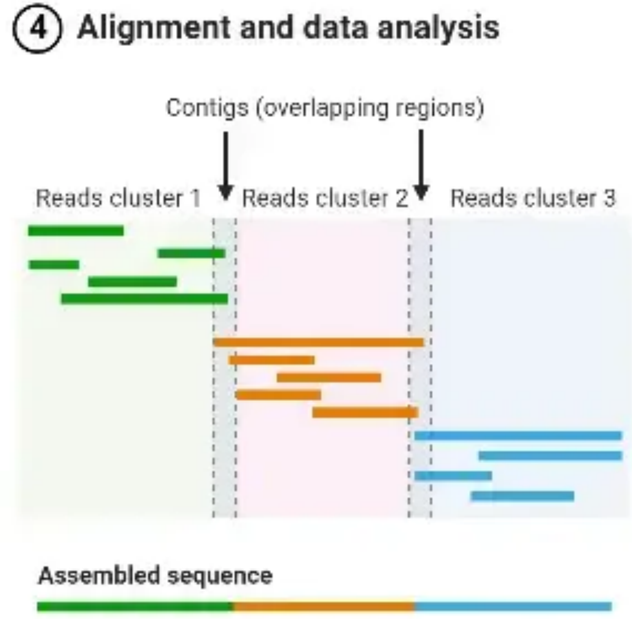
High-throughput sequencing

- 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.









Analysing NGS sequencing data via galaxy

Storing sequences

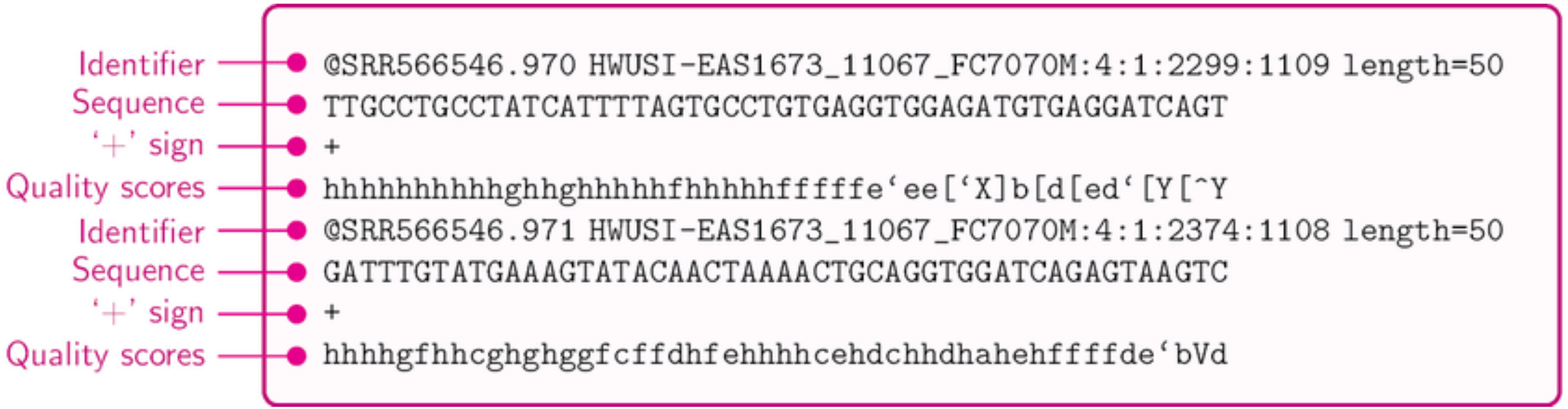
In Fasta format

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

Storing sequences

In 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.



Avian influenza viral strain analysis from gene segment sequencing data

Authors: Wolfgang Maier

Overview

Questions:

- With reassortment of gene segments being a common event in avian influenza virus (AIV) evolution, does it make sense to use a reference-based mapping approach for constructing consensus genome sequences for AIV samples?
- Is it possible to reuse existing tools and workflows developed for the analysis of sequencing data from other viruses?
- How can we obtain meaningful phylogenetic insight from AIV consensus sequences?

Objectives:

- Determine how reassortment impacts reference-based mapping approaches
- Use a collection of per-segment reference sequences to construct a hybrid reference genome that is sufficiently close to a sequenced sample to be useful as a reference for mapping
- Construct a sample consensus genome from mapped reads
- Generate per-segment phylogenetic trees of AIV consensus sequences

Requirements:

- Introduction to Galaxy Analyses
- Slides: Quality Control
- Hands-on: Quality Control
- Slides: Mapping
- Hands-on: Mapping

Time estimation: 4 hours

Level: Intermediate

Supporting Materials:

Input Histories | FAQs | Available on these Galaxies

https://bit.ly/3CMhEDw

End of tutorial