NGS Data Analysis

Clyde Dapat

WHO Collaborating Centre for Reference and Research on Influenza
VIDRL

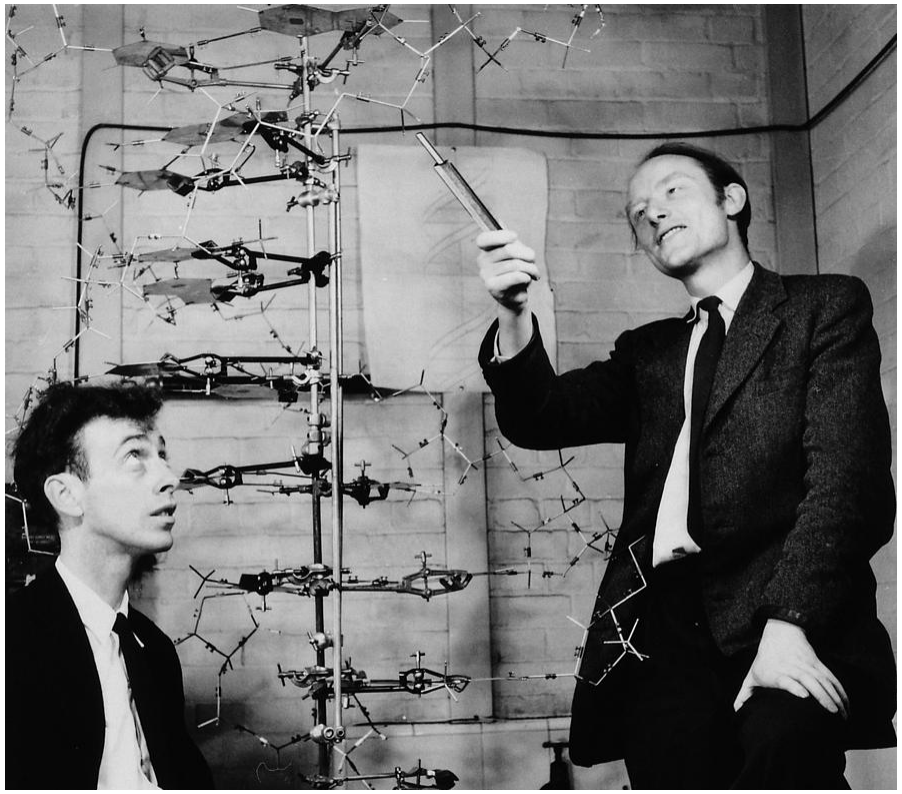
Doherty Institute
THE UNIVERSITY OF MELBOURNE
The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital

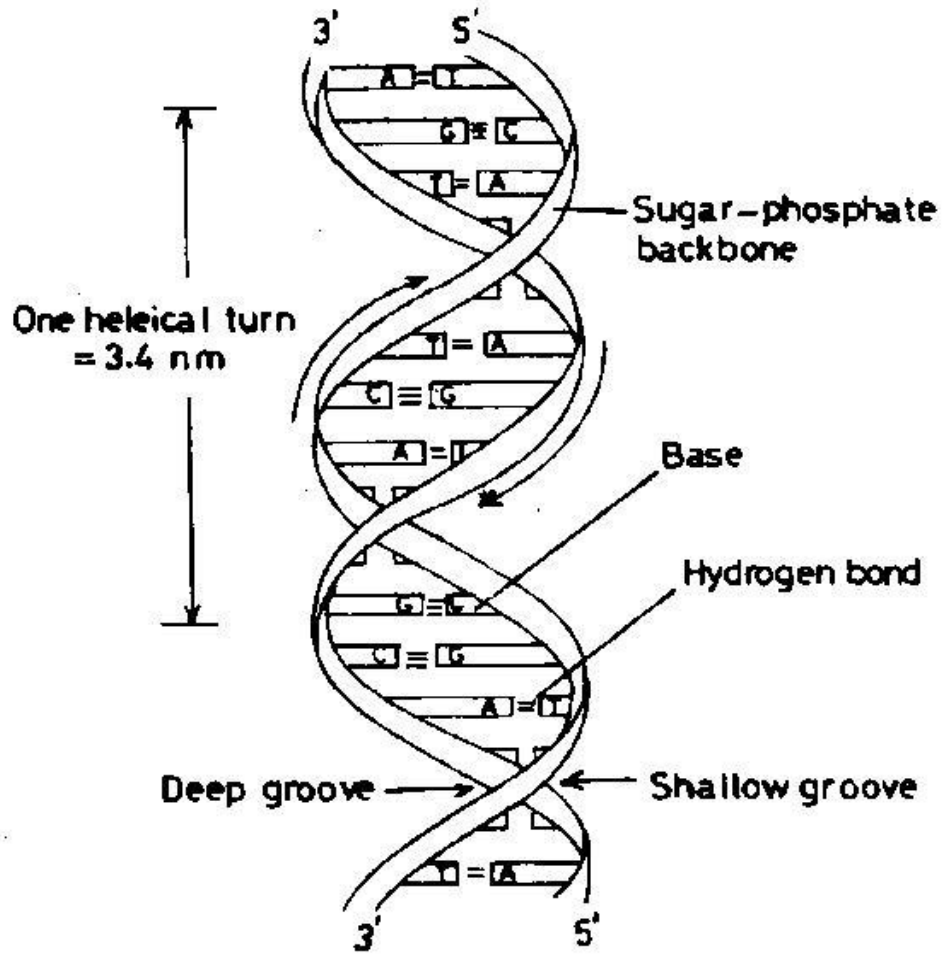
Objectives

- To understand the basics of sequence data analysis
- To provide an overview on the handling and analysis of high throughput NGS data

DNA



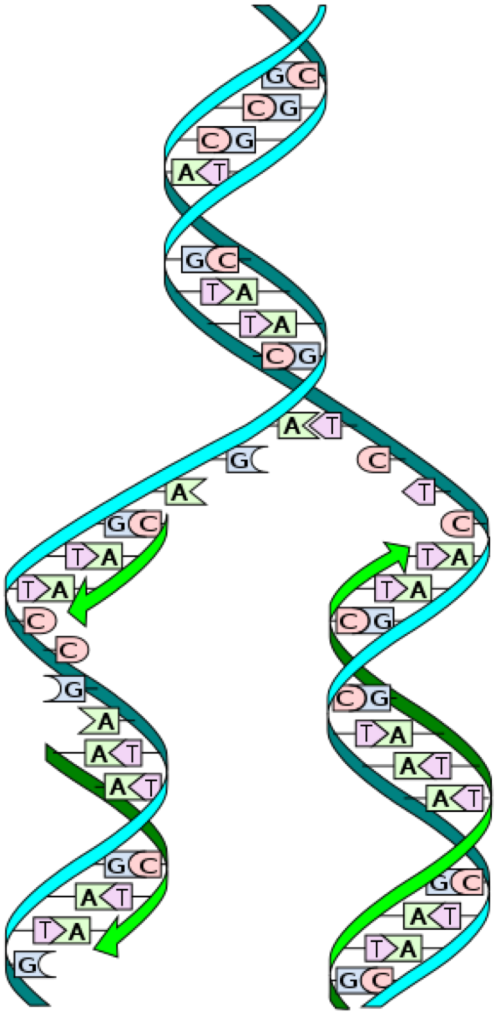
James Watson and Francis Crick



Nature (1953) 171:737-738

DNA sequencing

- Process of determining the order of sequence of nucleotides along the DNA strand

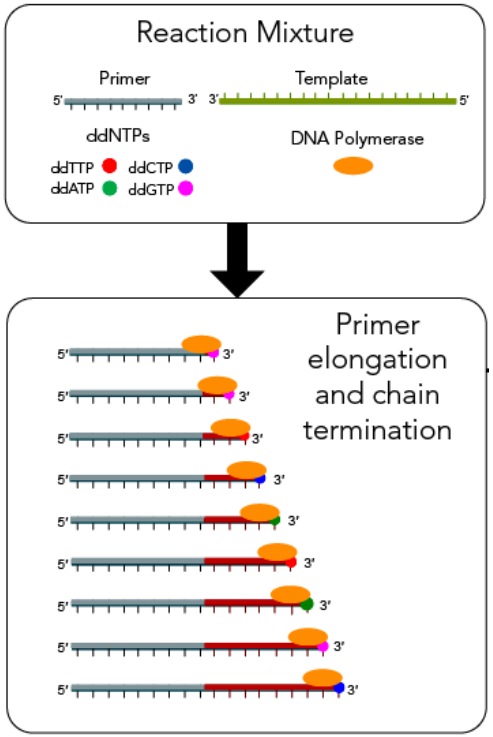


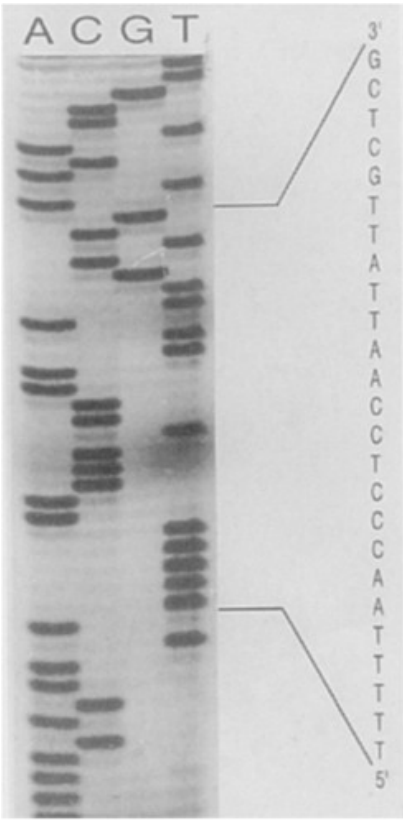
First generation sequencing

- Maxam-Gilbert, 1977
  - Chemical degradation method
  - Chemically breaks down radioactively labeled DNA at specific base positions
- Sanger, 1977
  - Enzyme-based termination of DNA strand elongation
  - Popular method of DNA sequencing

Sanger sequencing

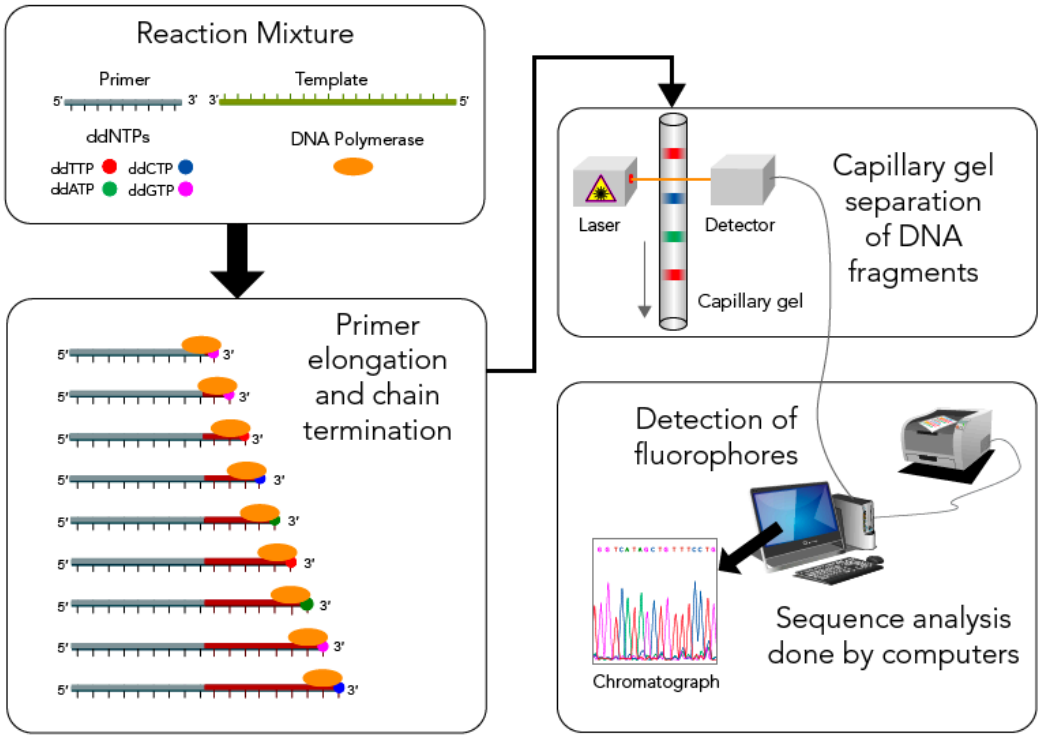
Frederick Sanger



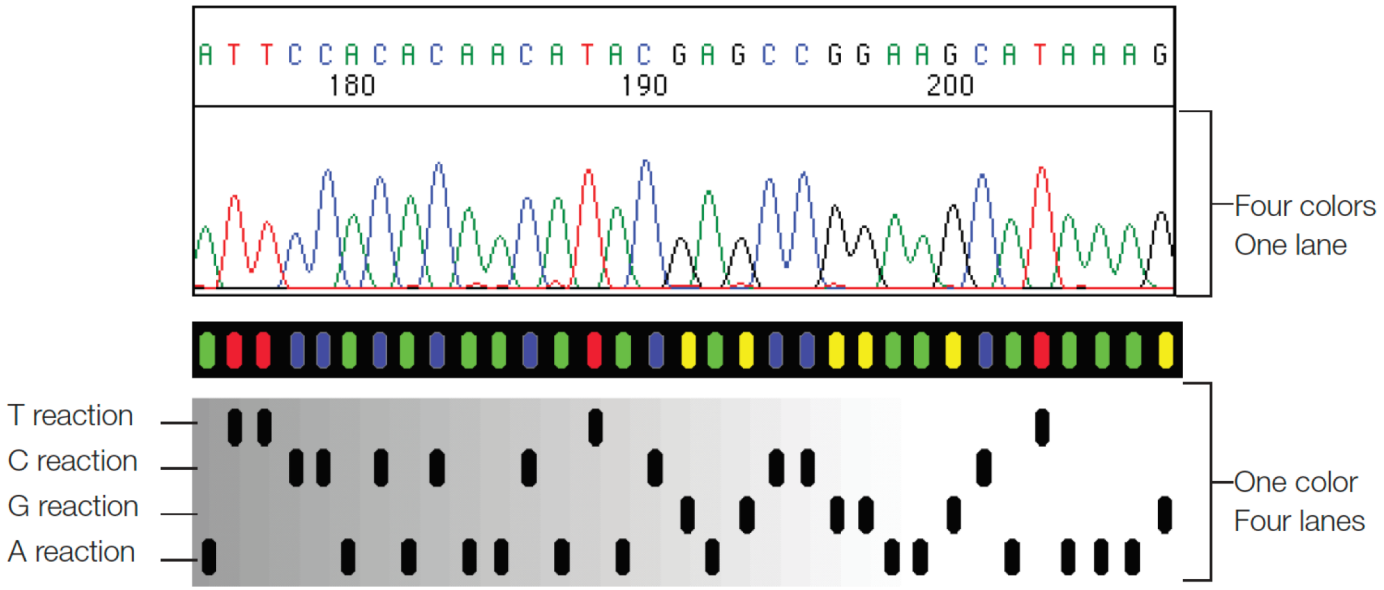


Capillary sequencing

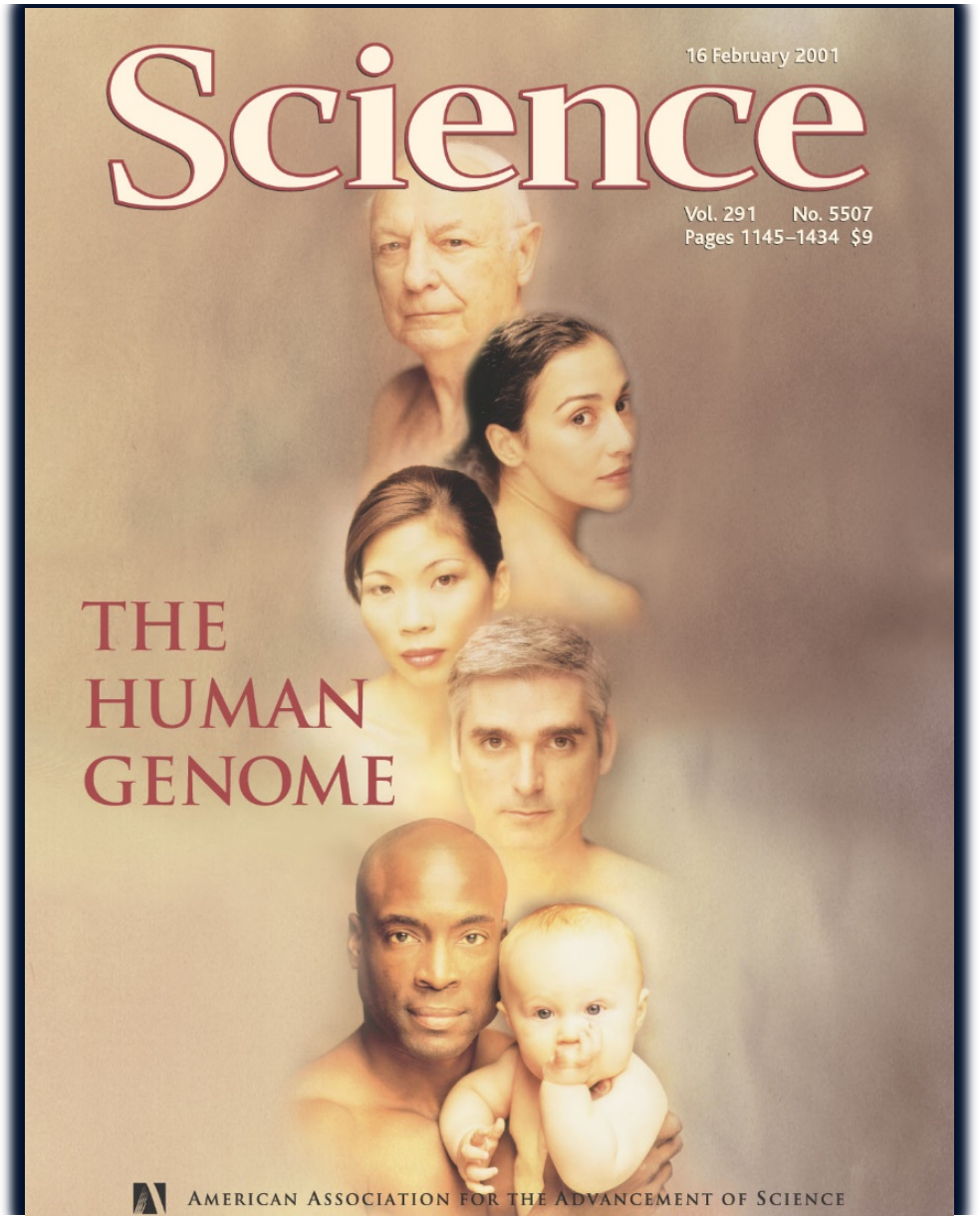
- ABI 310, 1996
- Same underlying concept as Sanger method
- DNA fragments are separated on long thin capillaries
- Small diameter allows for high electric field, faster and efficient separation



Capillary sequencing



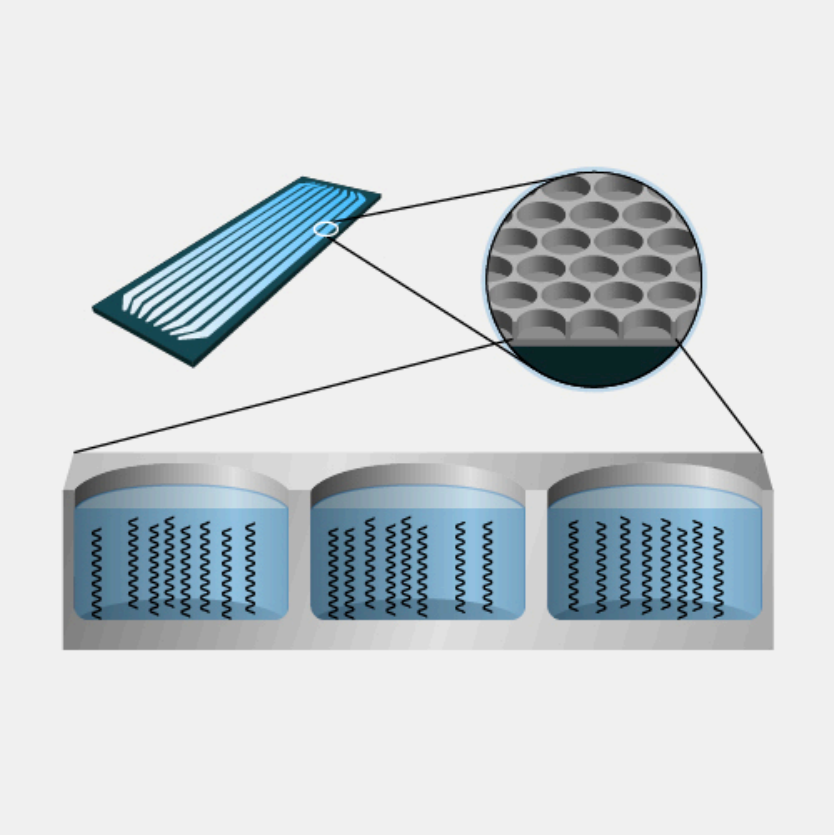

15 February 2001
nature
$10.00
www.nature.com
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special


16 February 2001
Science
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
American Association for the Advancement of Science

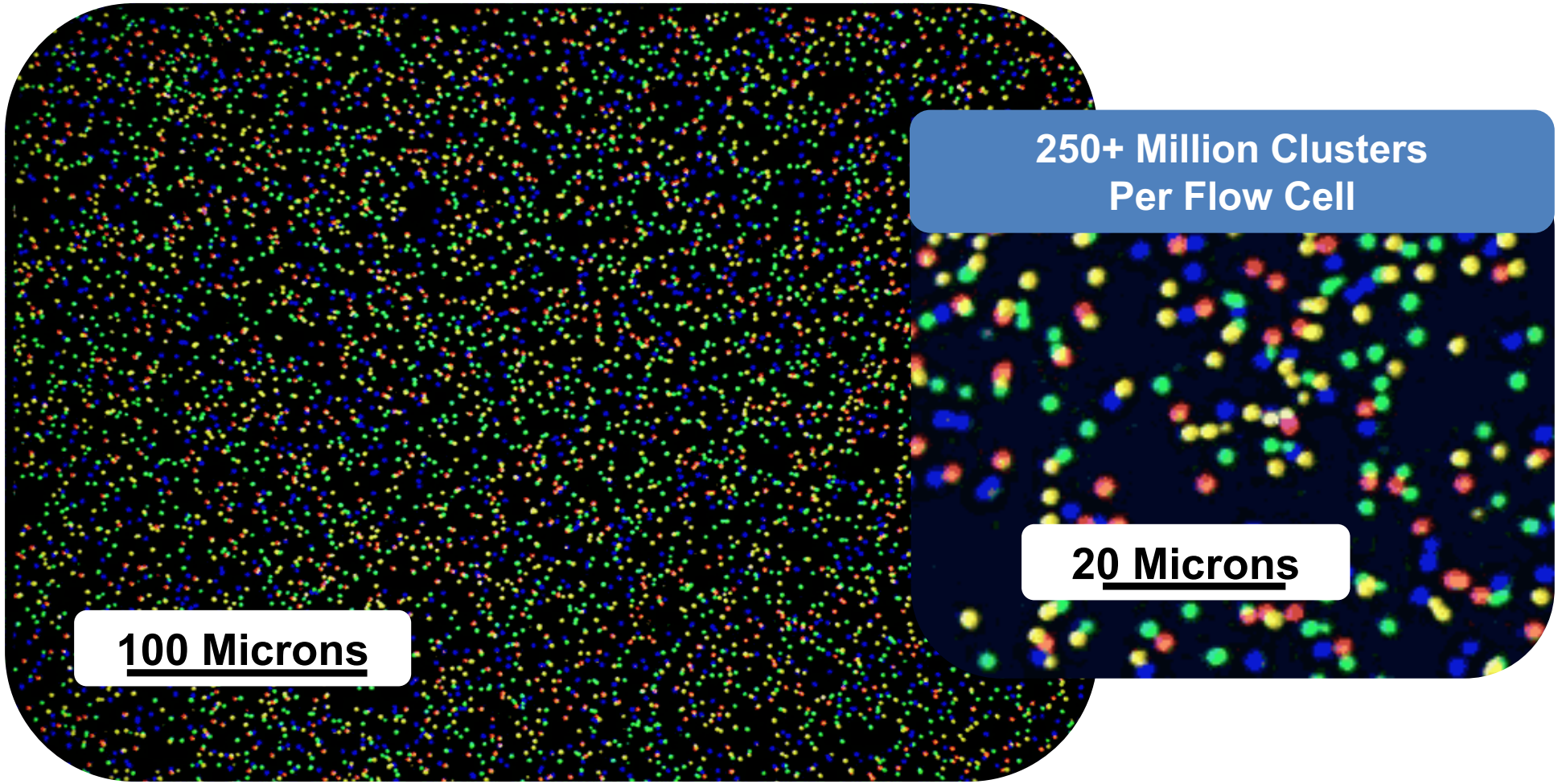
Next generation sequencing

Massively parallel

- Solexa (Illumina), 2006
- Uses micro- and nanotechnology for simultaneous sequencing of millions of short DNA fragments at a time



Flow cells and imaging

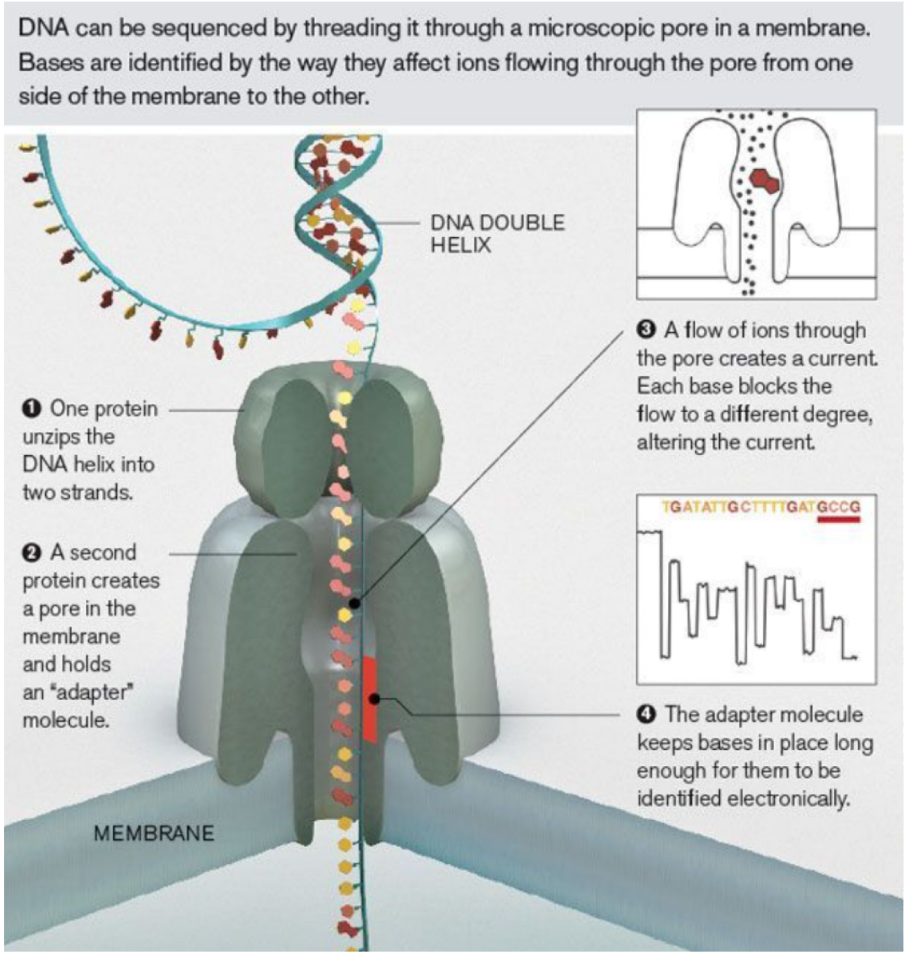


Third generation sequencing

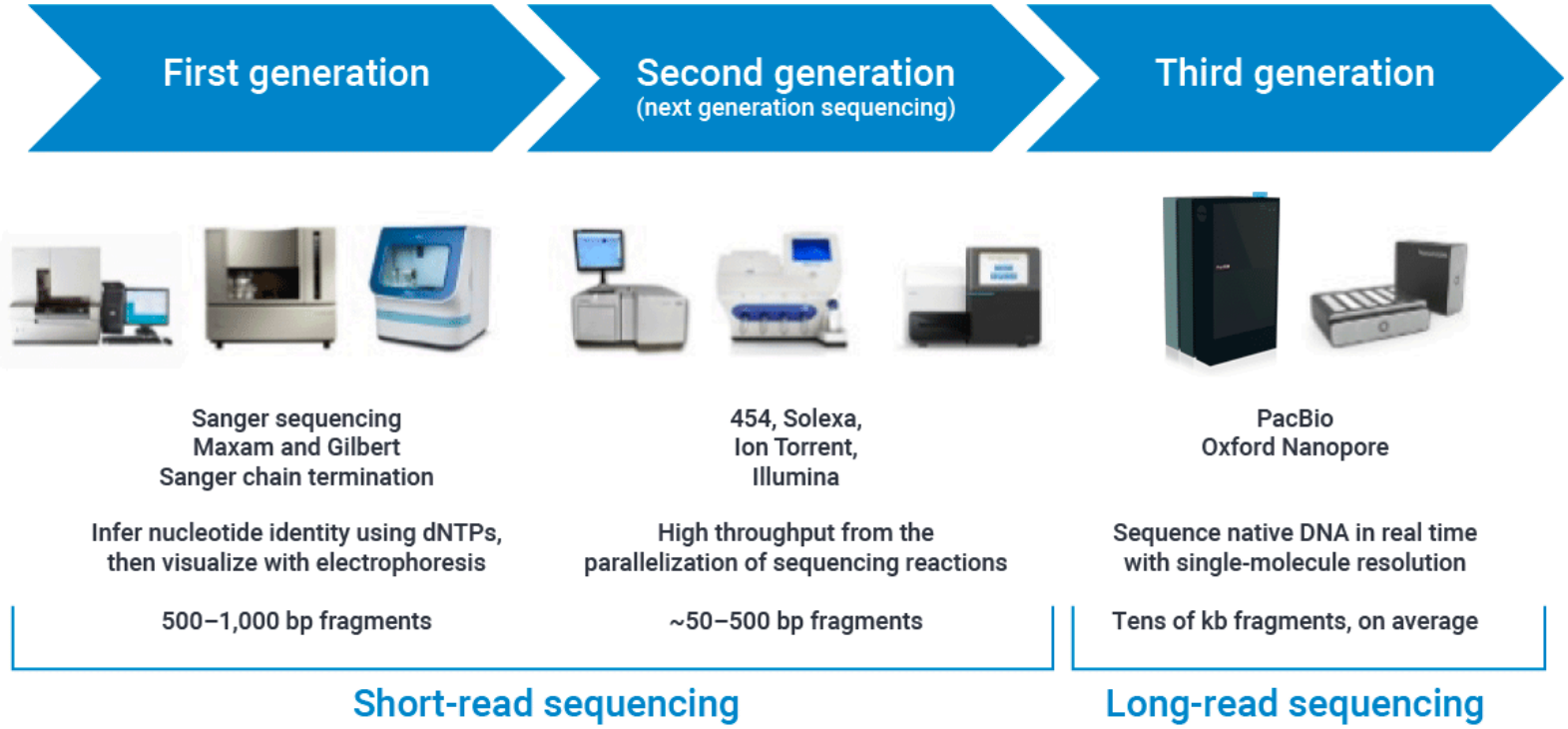
- Single molecule sequencing
- Obviates the need for DNA amplification
- Longer read lengths
- Minimal sample preparation
- Higher error rates in base calling

Nanopore sequencing

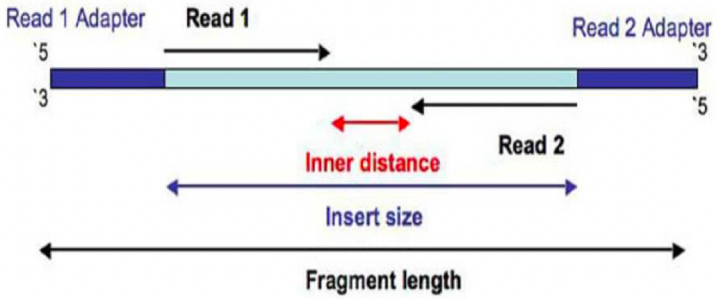
- Nanopore sequencing is based on threading a DNA strand through a pore (e.g. bacterial alpha-hemolysin) and
- The base sequence is inferred based on the shape and size of the nucleotides passing through the center of the pore
- Scalable and portable
- Real-time long read sequencing



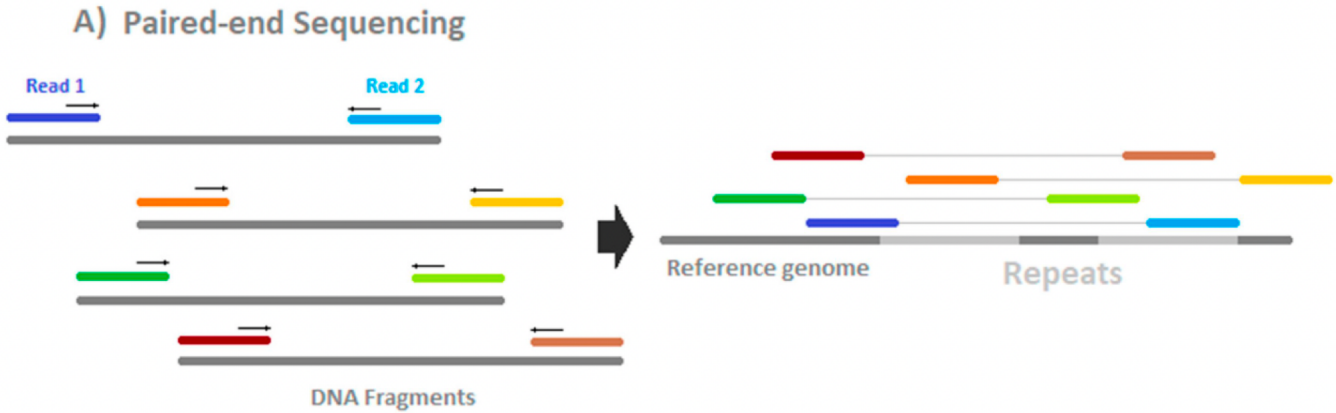
NGS technologies

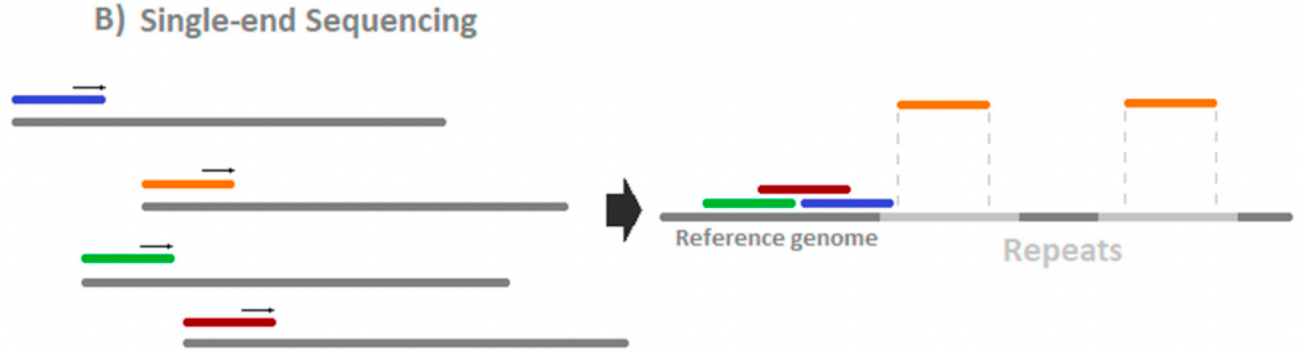


Single-end vs paired-end reads



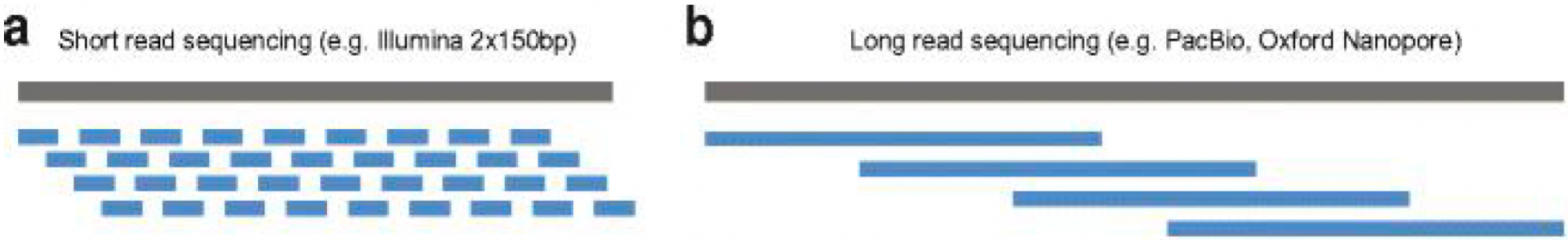
https://thesequencingcenter.com/knowledge-base/what-are-paired-end-reads/





https://www.mdpi.com/epigenomes/epigenomes-02-00021/article_deploy/html/images/epigenomes-02-00021-g004.png

Short vs long reads



Sequence data analysis

- Problem

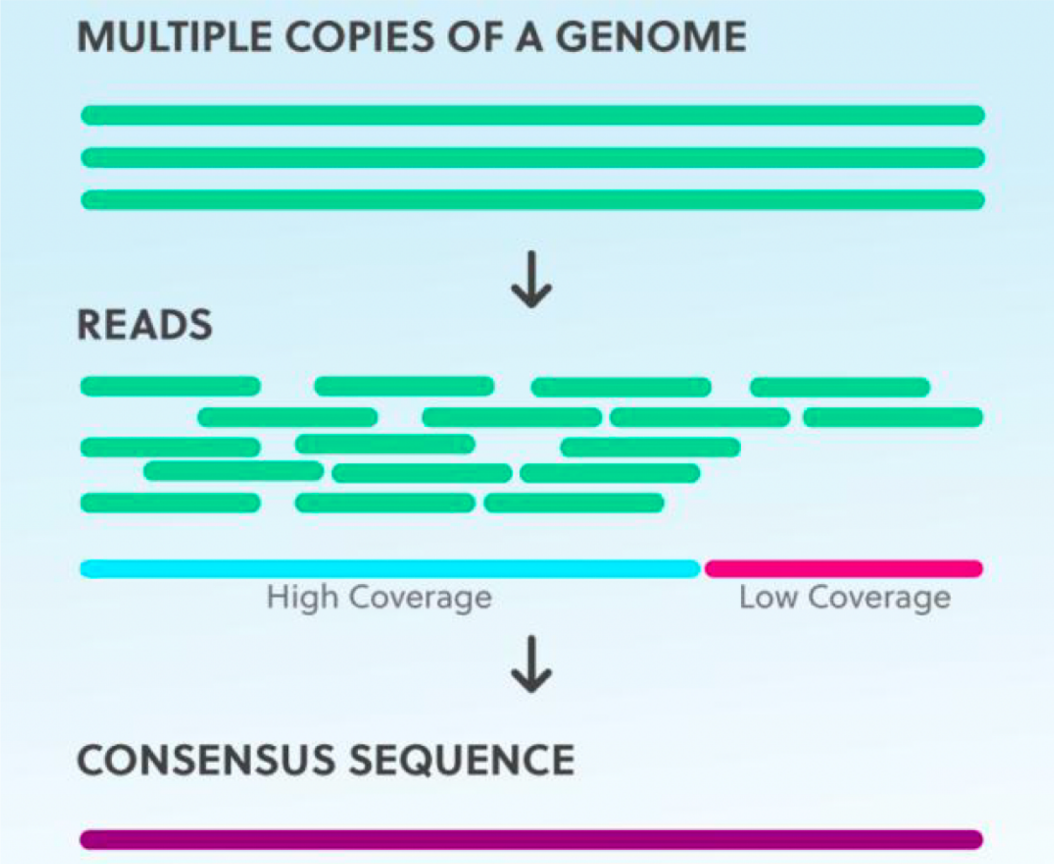
copies of target genome

sequence reads

reconstruct genomic sequence

- Assembly – process of putting individual short reads in the right order to create a consensus sequence

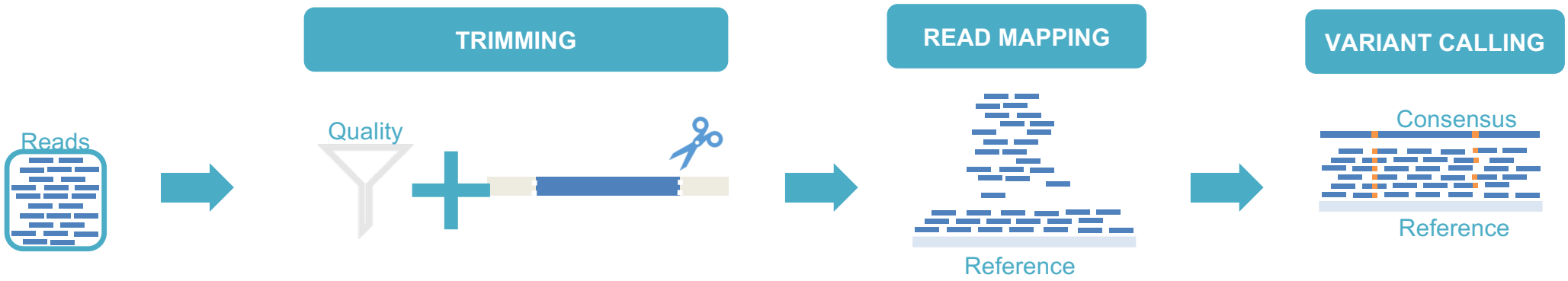
Sequence data analysis



Sequence Reads

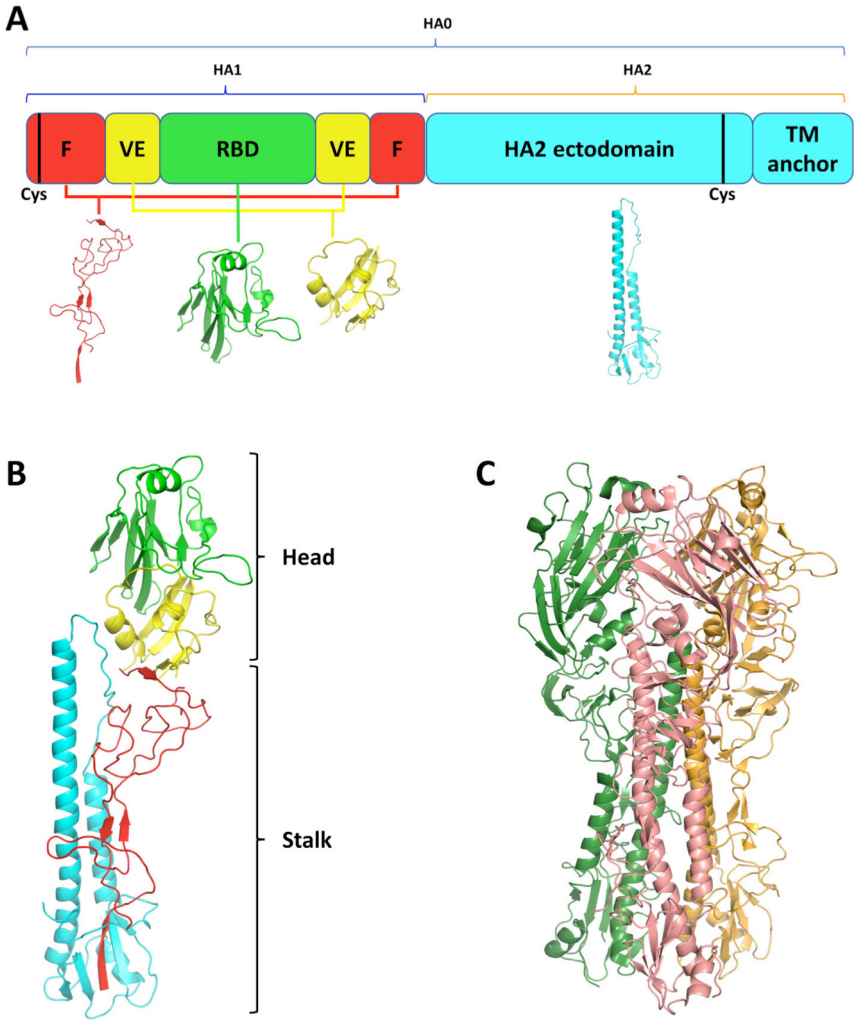
- Fragments of the original genome or any template sequence (FASTQ format)
- One of the initial goals of bioinformatics is to reconstruct the template sequence based on the read fragment information (sequence assembly)

Sequence data analysis



Annotation

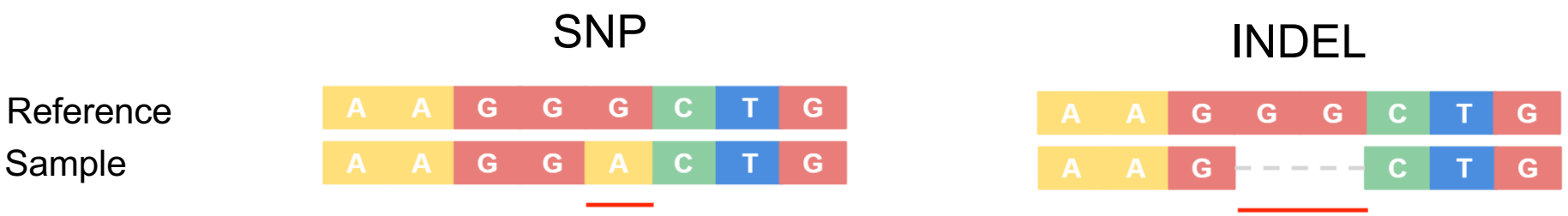
- Genome annotation is the process of identifying structure and function of the components of a genome



Sci Rep (2023) 13:6940

Variant calling

- Variant calling is the process of identifying single nucleotide polymorphisms (SNPs), insertions and deletions (indels) from NGS data



Thank you !

clyde.dapat@influenzacentre.org

WHO Collaborating Centre for Reference and Research on Influenza
VIDRL

Doherty Institute
THE UNIVERSITY OF MELBOURNE
The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital