WHO Collaborating Centre for Reference and Research on Influenza
VIDRL

World Health Organization
Philippines

KDCA
Korea Disease Control and Prevention Agency

NGS DATA ANALYSIS

FRANCIS A. TABLIZO
Core Facility for Bioinformatics
Philippine Genome Center
University of the Philippines System

April 15, 2024

CENTRE FOR PATHOGEN GENOMICS

Doherty Institute

THE UNIVERSITY OF MELBOURNE

The Royal Melbourne Hospital

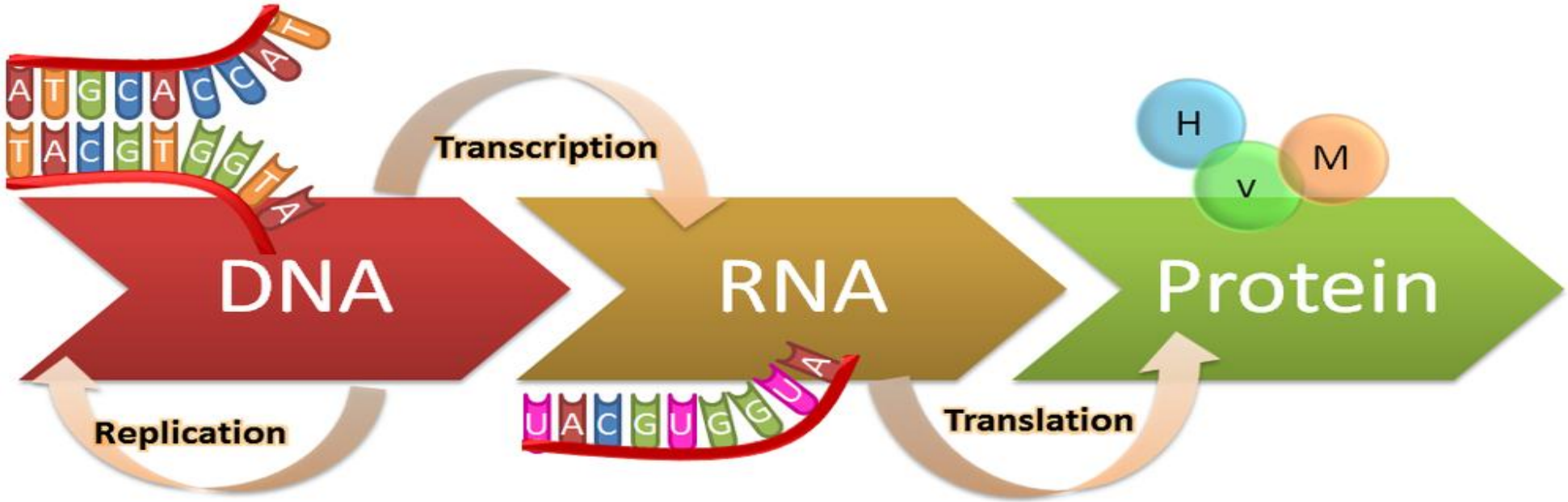
A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Objectives

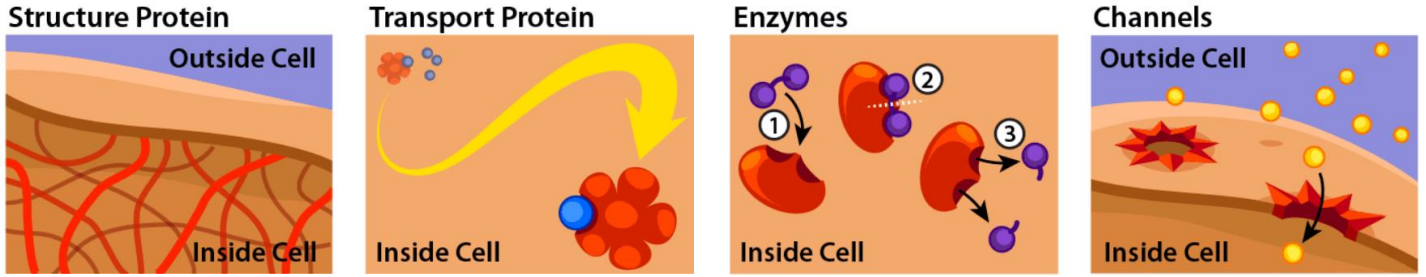
- Go through the currently available sequencing technologies
- Provide an overview on the handling and analysis of high-throughput sequencing data sets, particularly in the field of genomics

Central Dogma of Molecular Biology

CENTRE FOR PATHOGEN GENOMICS
Doherty Institute
THE UNIVERSITY OF MELBOURNE
The Royal Melbourne Hospital
A joint venture between The University of Melbourne and The Royal Melbourne Hospital



https://genius.com/Biology-genius-the-central-dogma-annotated

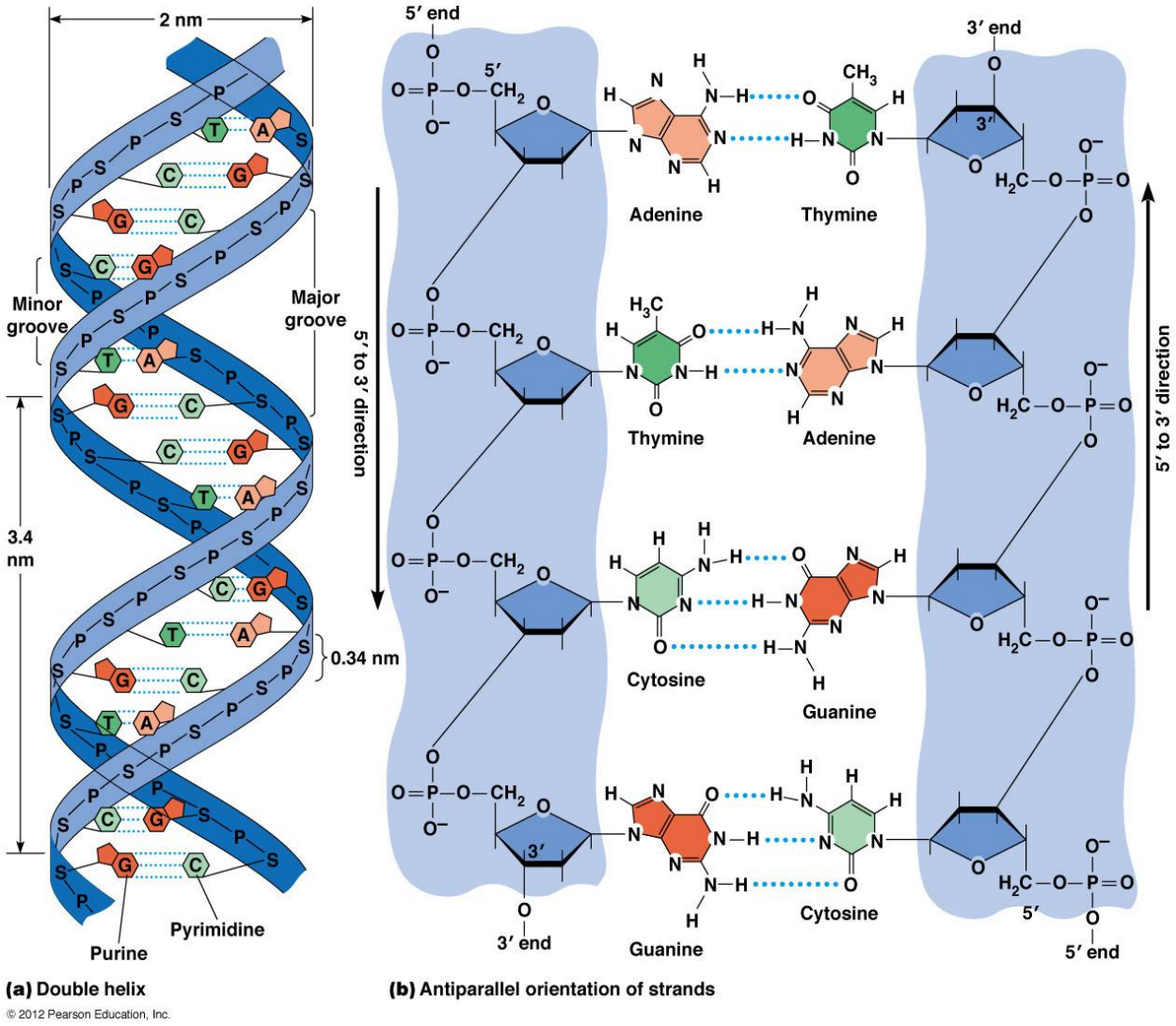


https://askabiologist.asu.edu/venom/what-are-proteins

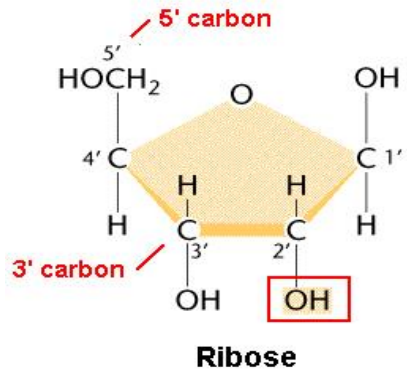
http://www.pixabay.com/

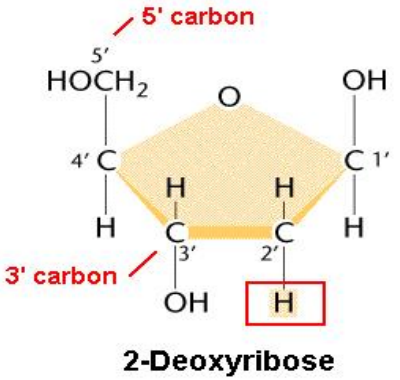
DNA Molecule

Deoxyribonucleic acid



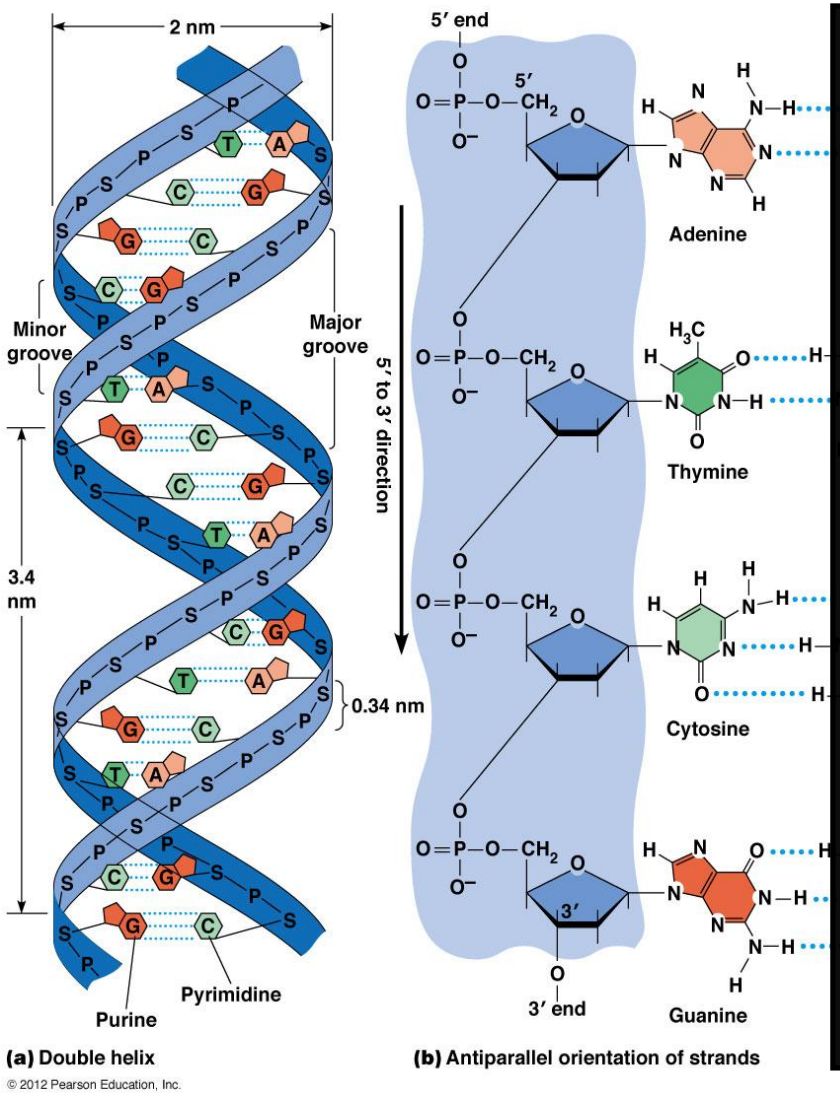
http://www.mun.ca/biology/desmid/brian/BIOL2060/BIOL2060-18/18_04.jpg





https://www.mun.ca/biology/scarr/Fg10_09b_revised.gif

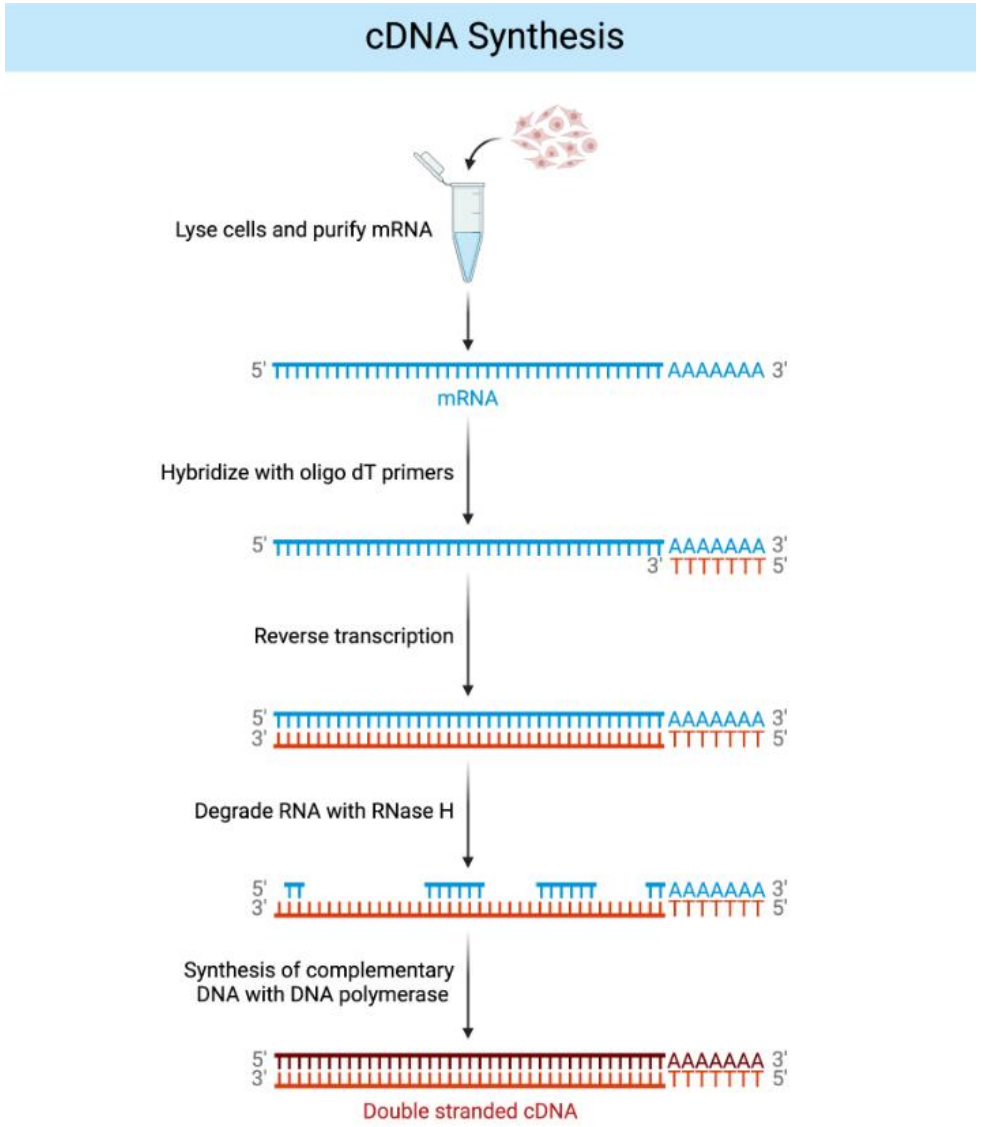
DNA Sequencing



Refers to the process of determining the order or sequence of nucleotides along a DNA strand

http://www.mun.ca/biology/desmid/brian/BIOL2060/BIOL2060-18/18_04.jpg

DNA Sequencing



- Most current sequencing technologies only work with **double-stranded DNA** molecule as starting material.
- Usually, if the starting material is RNA, there is a need to **convert the RNA to its complementary DNA (cDNA) form** prior to sequencing.
- Viruses have **other forms of genetic material** (e.g., ssDNA, dsRNA, etc.). Make sure to take this in consideration before sequencing.

https://www.biorender.com/template/cdna-synthesis

WHO Collaborating Centre for Reference and Research on Influenza
VIDRL

World Health Organization
Philippines

KDCA
Korea Disease Control and Prevention Agency

CURRENT SEQUENCING TECHNOLOGIES

CENTRE FOR PATHOGEN GENOMICS

Doherty Institute

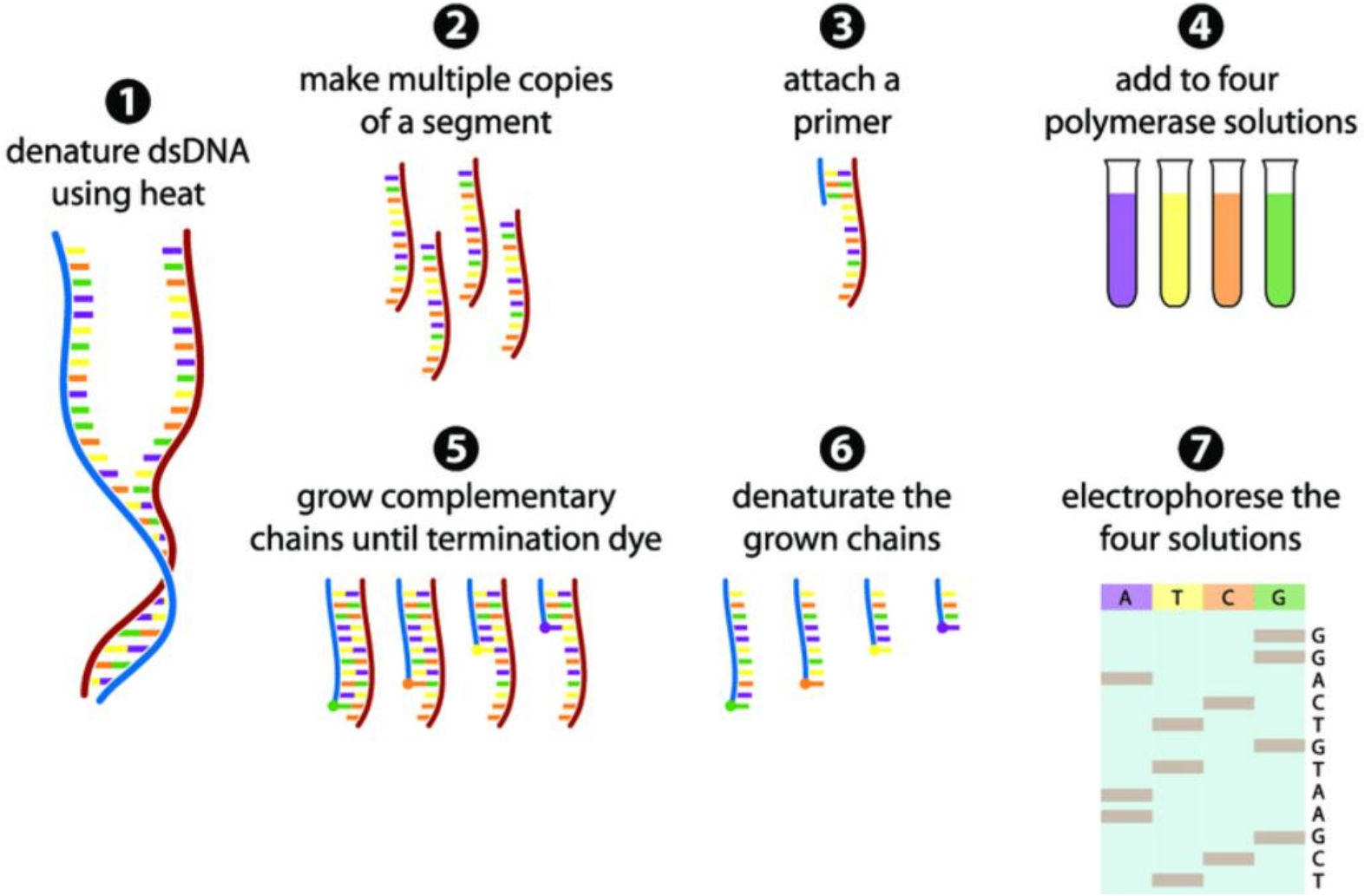
THE UNIVERSITY OF MELBOURNE

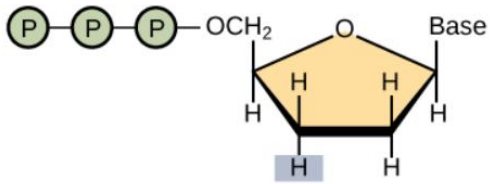
The Royal Melbourne Hospital

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

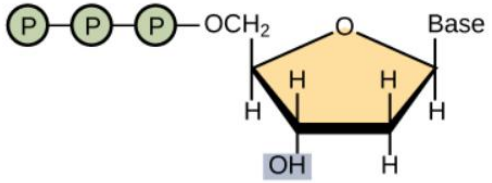
Sanger Sequencing

Dideoxy chain termination method





Dideoxynucleotide (ddNTP)



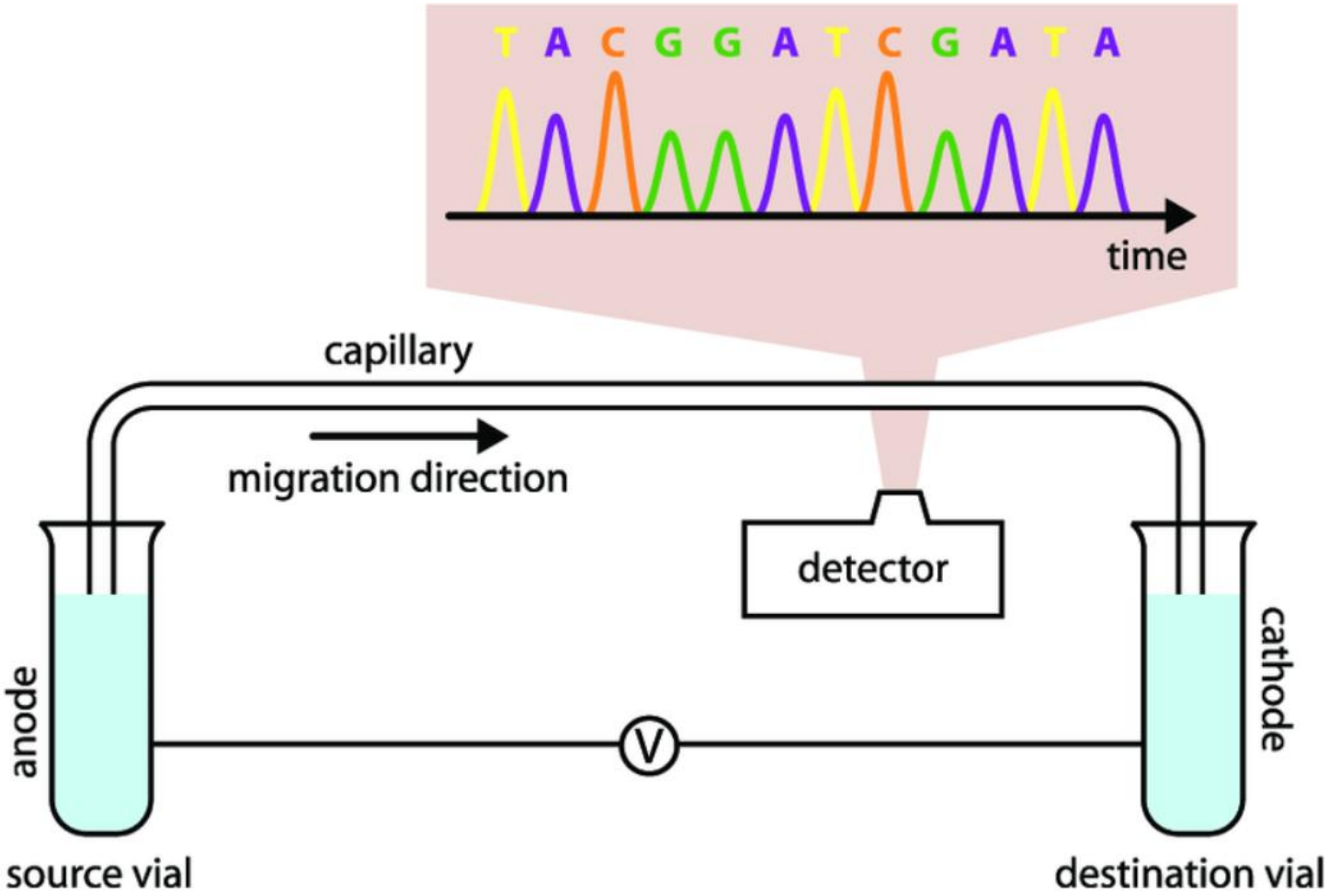
Deoxynucleotide (dNTP)

https://openstax.org/books/biology/pages/17-3-whole-genome-sequencing

https://www.researchgate.net/publication/234248746_Simulation_of_polymer_translocation_through_small_channels_A_molecular_dynamics_study_and_a_new_Monte_Carlo_approach

Capillary Sequencing

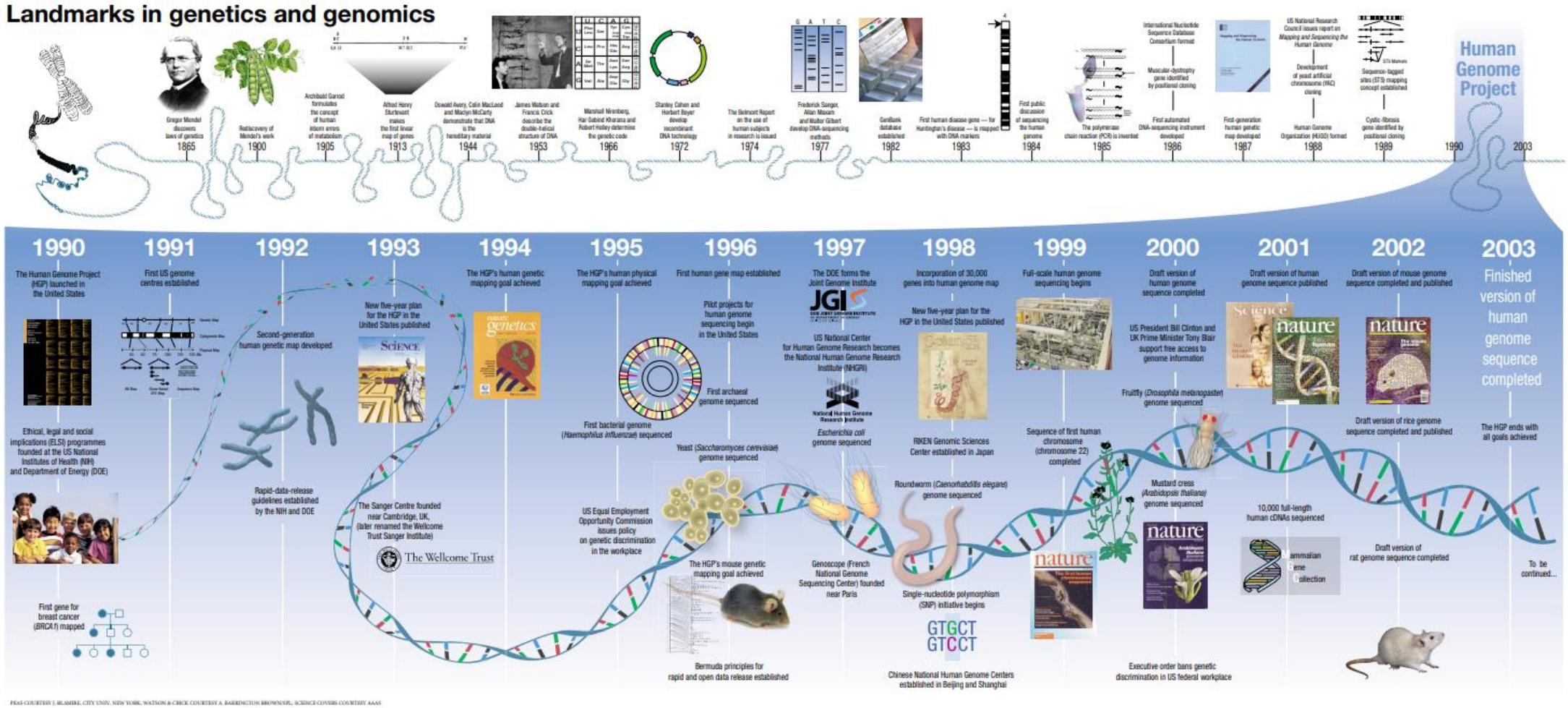
Dideoxy chain termination method



https://www.researchgate.net/publication/234248746_Simulation_of_polymer_translocation_through_small_channels_A_molecular_dynamics_study_and_a_new_Monte_Carlo_approach

Capillary Sequencing

Main technology available during the Human Genome Project



https://www.genome.gov/sites/default/files/media/files/2020-09/HGP_Timeline.pdf

High-Throughput Sequencing

First NGS Platform on the Market: 454 Life Sciences, Pyrosequencing, 2004

Proof of Concept: “Project Jim”, Published in 2008

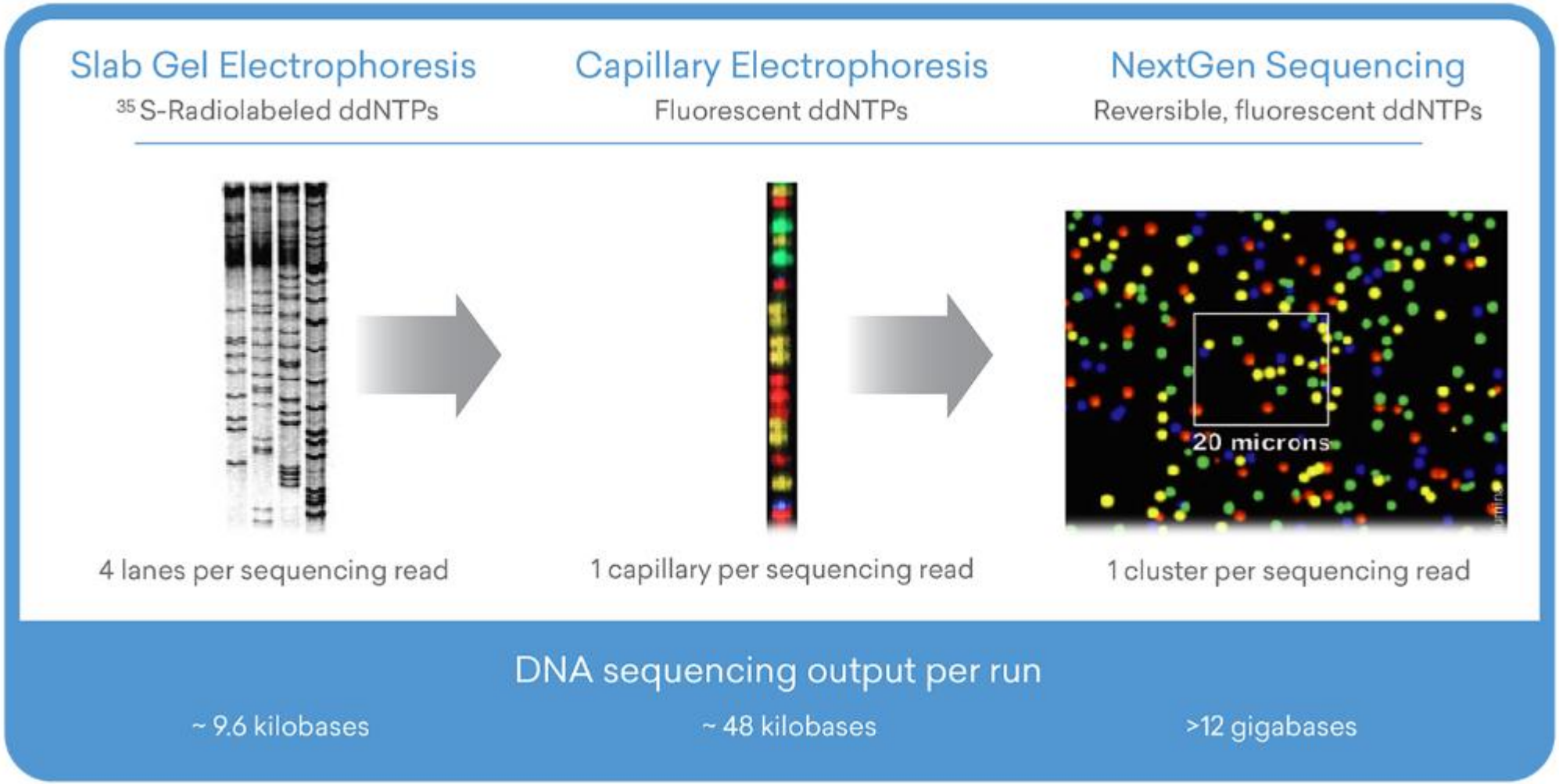
Jim Watson	Human Genome Project
454 Life Sciences, Pyrosequencing Technology, analyzed by BCM-HGSC	Sanger Capillary Sequencing
2 months, 3 instruments	10-13 Years
$1-2 million $ 250,000 with GS Titanium FLX	$ 100 million - $ 2.7 billion
8x coverage	7.5x coverage
250 bp read length 400 bp with GS Titanium FLX	500-800 bp

Source: History of DNA Sequencing & Current Applications. Roche Applied Sciences.

https://www.yumpu.com/en/document/read/25815413/history-of-dna-sequencing-amp-roche

High-Throughput Sequencing

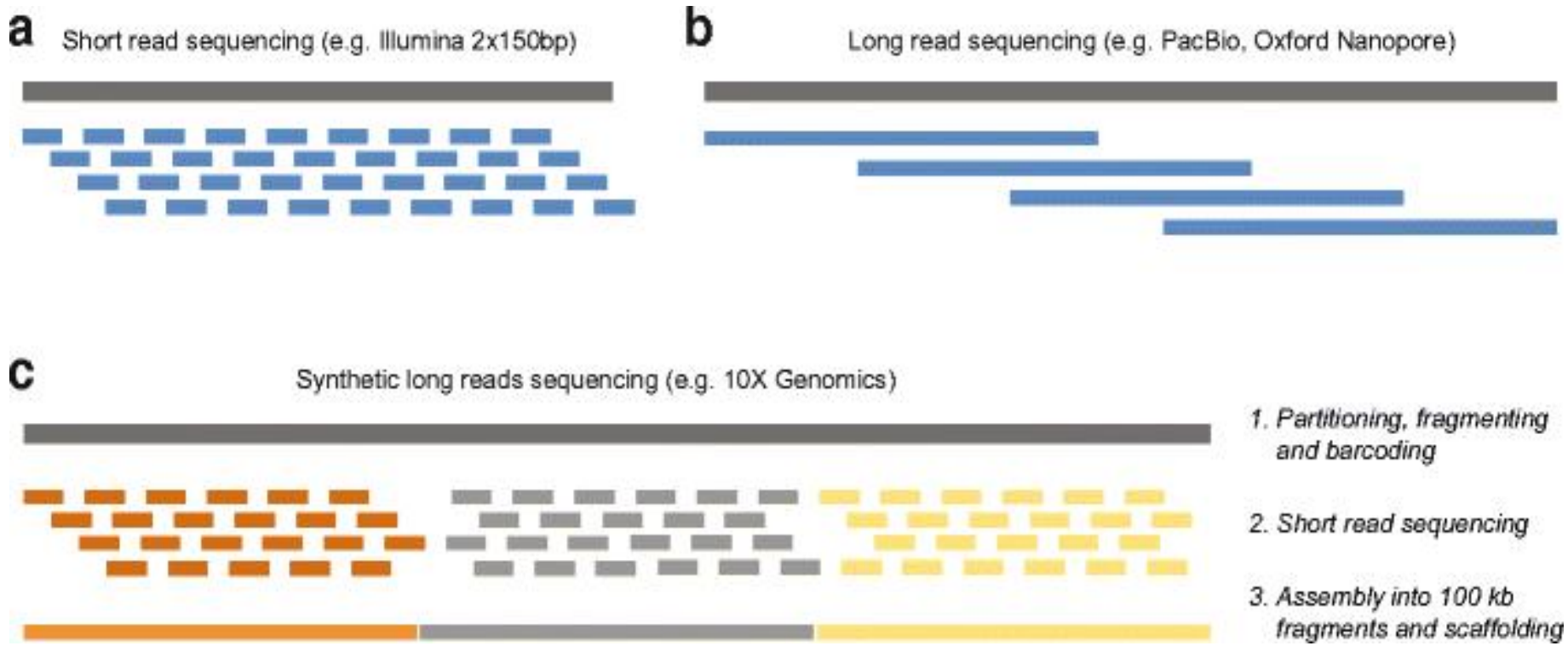
Simultaneous sequencing of thousands to even billions of DNA fragments in a single run



https://seqwell.com/sanger-method-the-mp3-player-of-sequencing/

Important Concepts

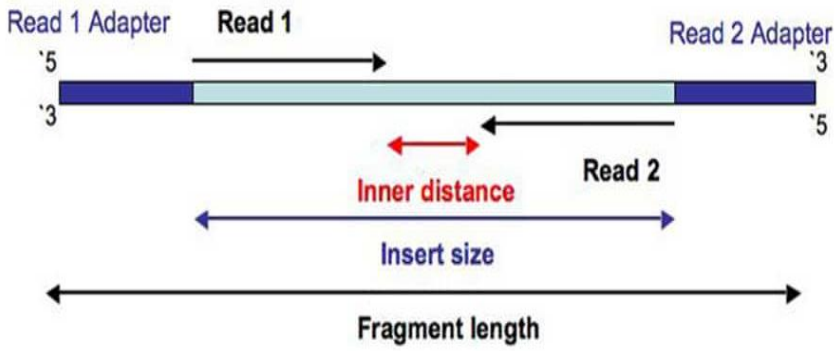
Short vs. Long Read Sequencing



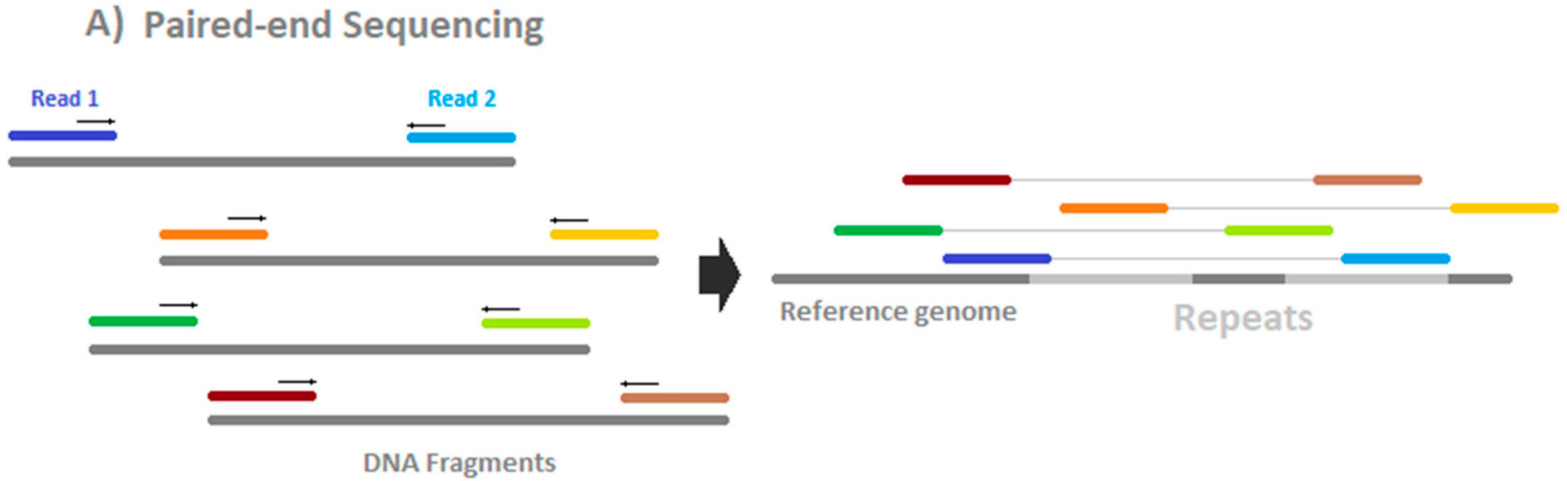
https://media.springernature.com/full/springer-static/image/art%3A10.1208%2Fs12248-017-0161-x/MediaObjects/12248_2017_161_Fig5_HTML.gif?as=webp

Important Concepts

Single-end vs. Paired-end Data (Short Reads)



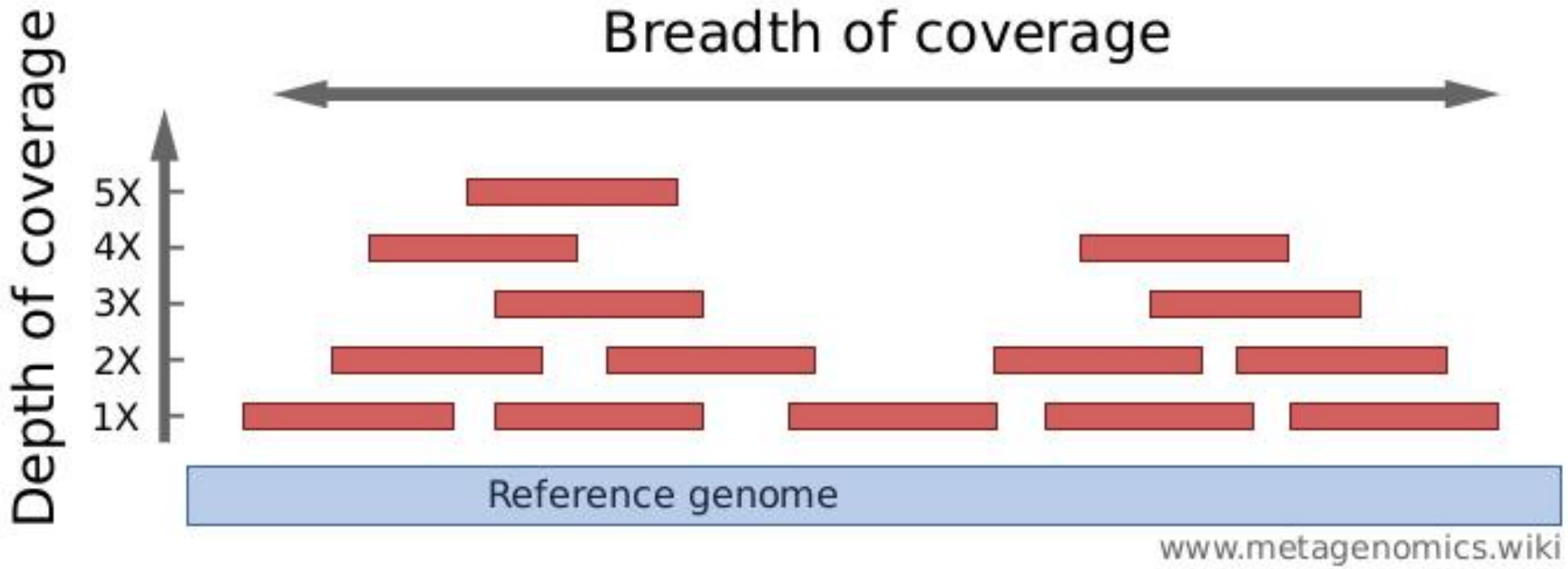
https://thesequencingcenter.com/knowledge-base/what-are-paired-end-reads/



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

Important Concepts

Breadth vs. Depth of Coverage



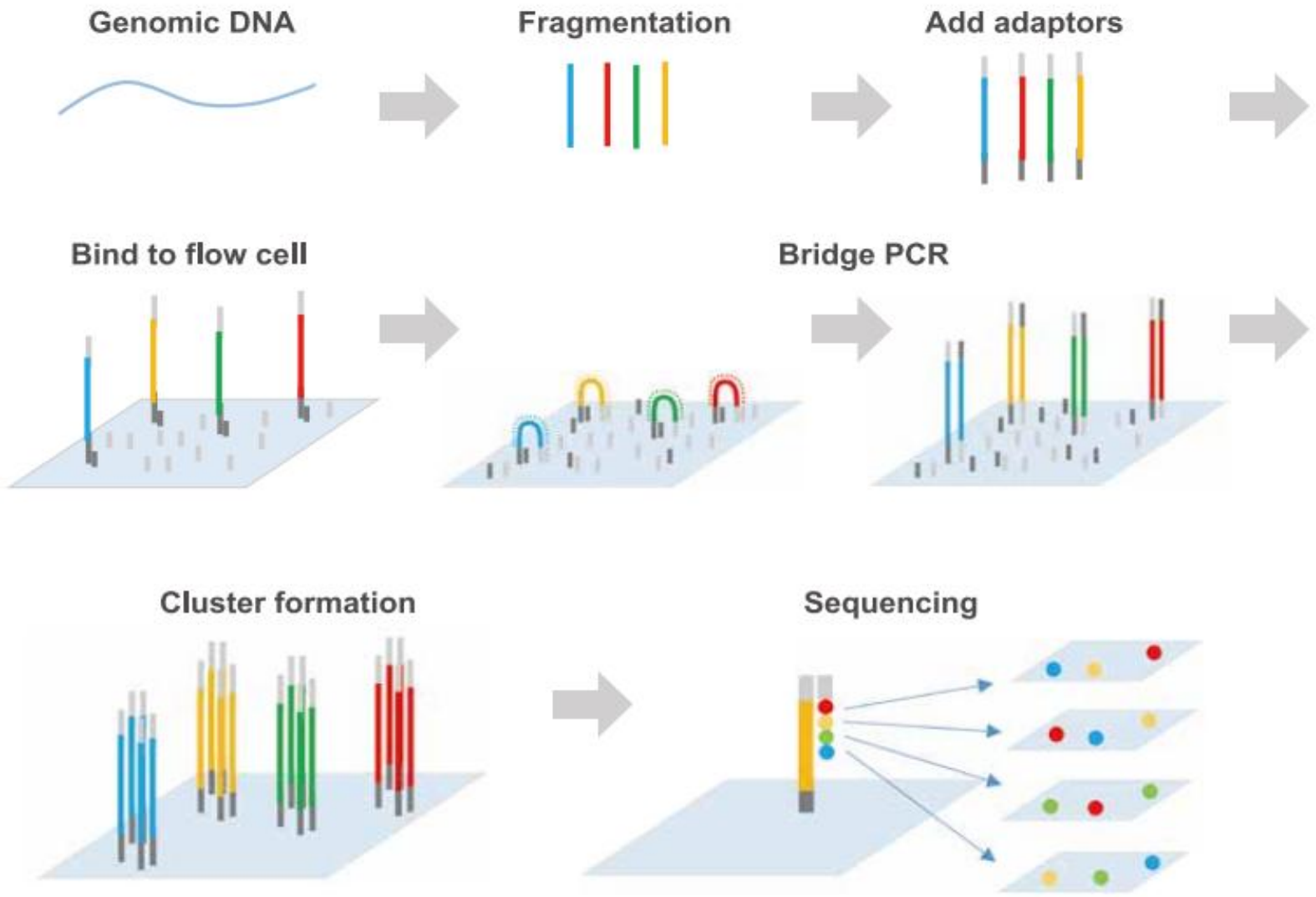
https://i.stack.imgur.com/hZqvF.png

Next (2nd) Generation Sequencing

Platform: Illumina

Sequencing by synthesis

- Read Length: 50-300 bp
- Read Fragments: Paired-end, Single-end
- Throughput: 1.2GB to 8TB



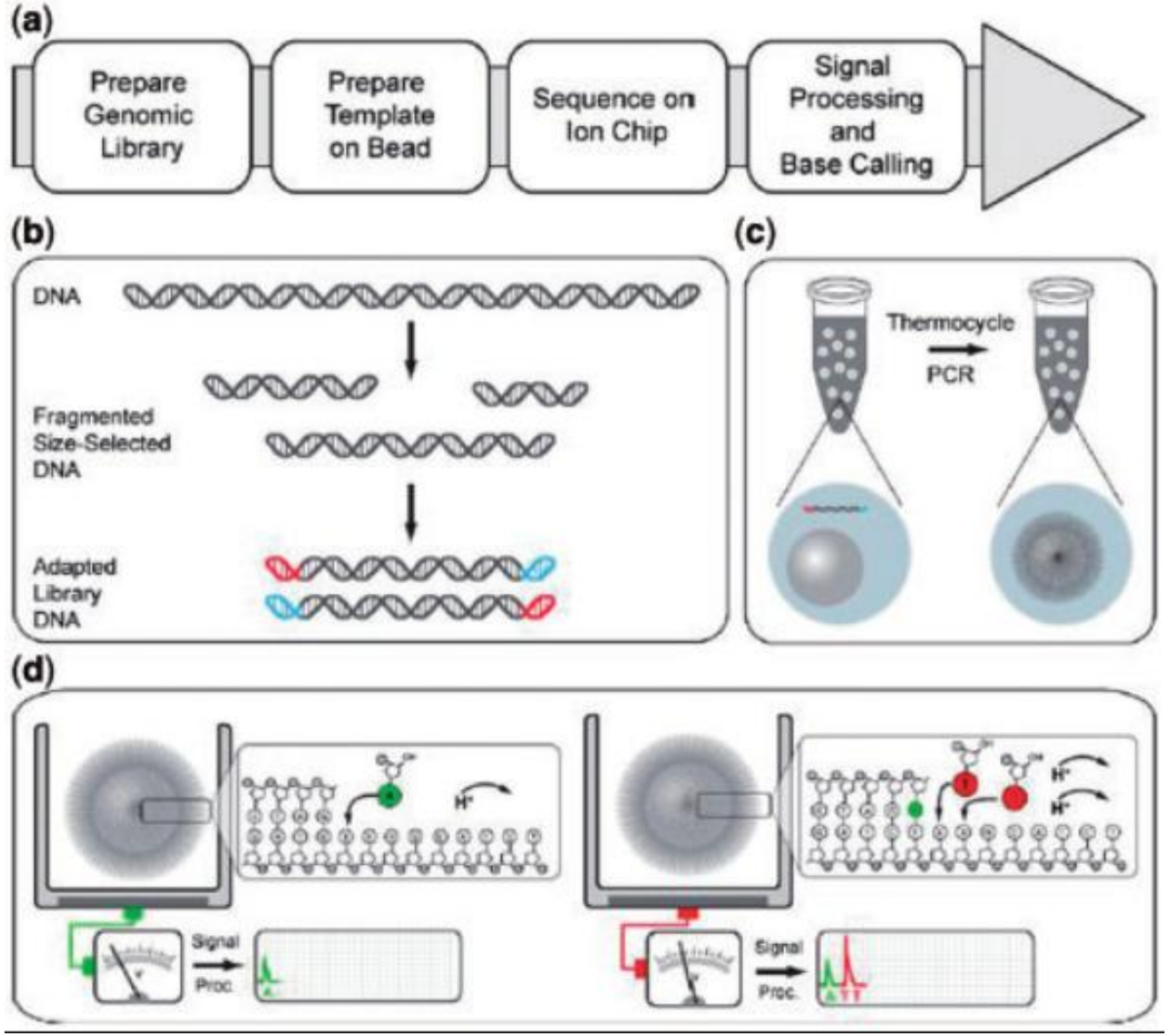
https://www.genscript.com/advancing-genomics-medicine-and-health-together-by-semiconductor-dna-synthesis-technology-summary.html

Next (2nd) Generation Sequencing

Platform: Ion

Semiconductor sequencing

- Read Length: 200-400 bp
- Read Fragments: Single-end, Paired-end (requires additional preparation steps)
- Throughput: 400MB to 26GB



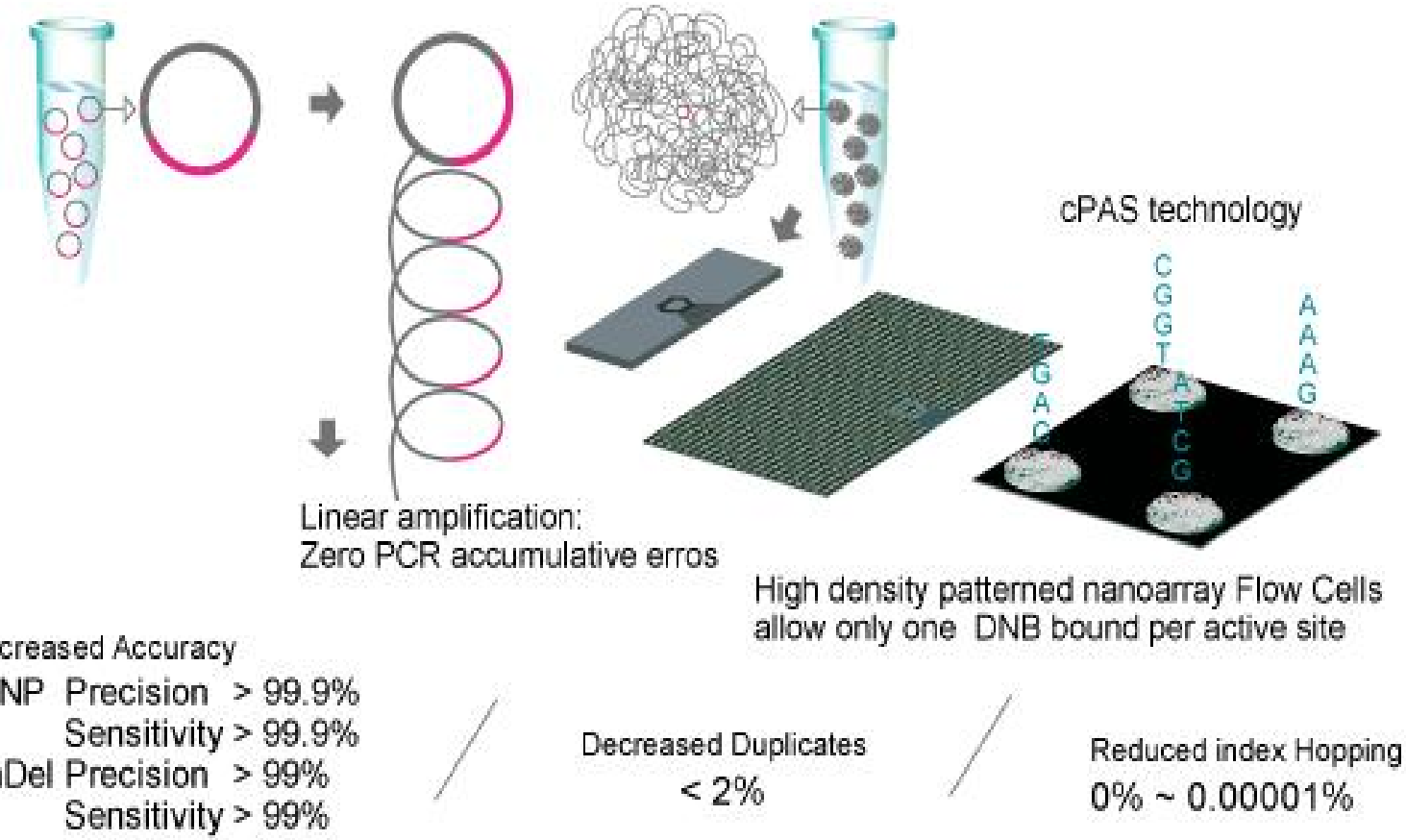
https://www.researchgate.net/publication/243969229/figure/fig1/AS:340876061626368@1458282594524/Ion-sequencing-work-flow-The-overall-workflow-is-shown-in-a-A-genome-library-is.png

Next (2nd) Generation Sequencing

Platform: MGI / BGI

DNA nanoball sequencing (DNBSeq)

- Read Length: 50-150 bp
- Read Fragments: Paired-end, Single-end
- Throughput: 7.5GB to 76.8TB



https://genebay.co.jp/wp-content/uploads/2020/04/5e7e14571f8684deb577245d105792f2.png

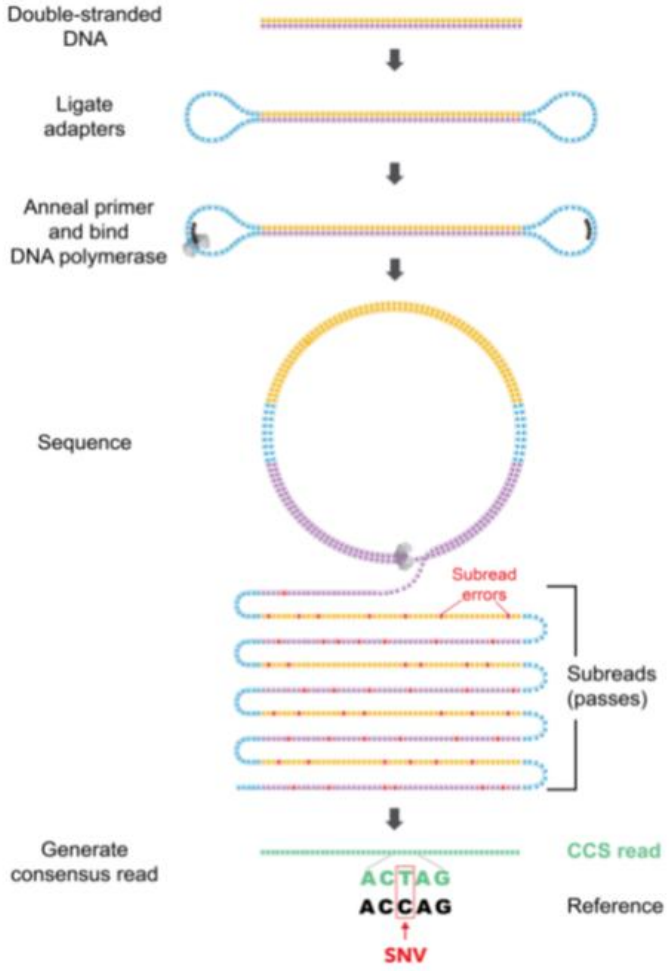
Third Generation Sequencing

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

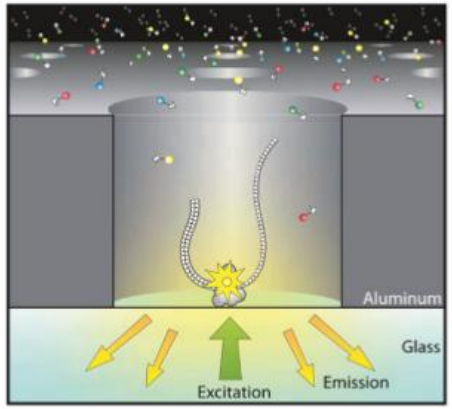
Platform: Pacific Biosciences (Pacbio)

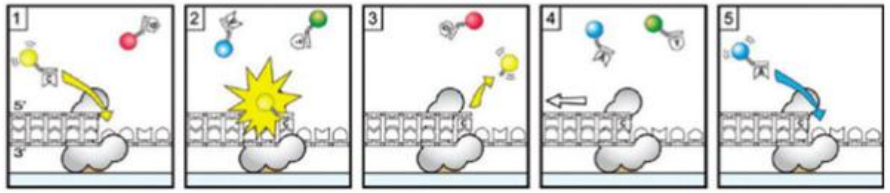
Single Molecule, Real Time (SMRT) Sequencing

- Read Length: 15-25 kbp
- Throughput: 24Gb to 360Gb



Single Molecule, Real-Time (SMRT) Sequencing technology delivers the highest consensus accuracy with unprecedented read lengths





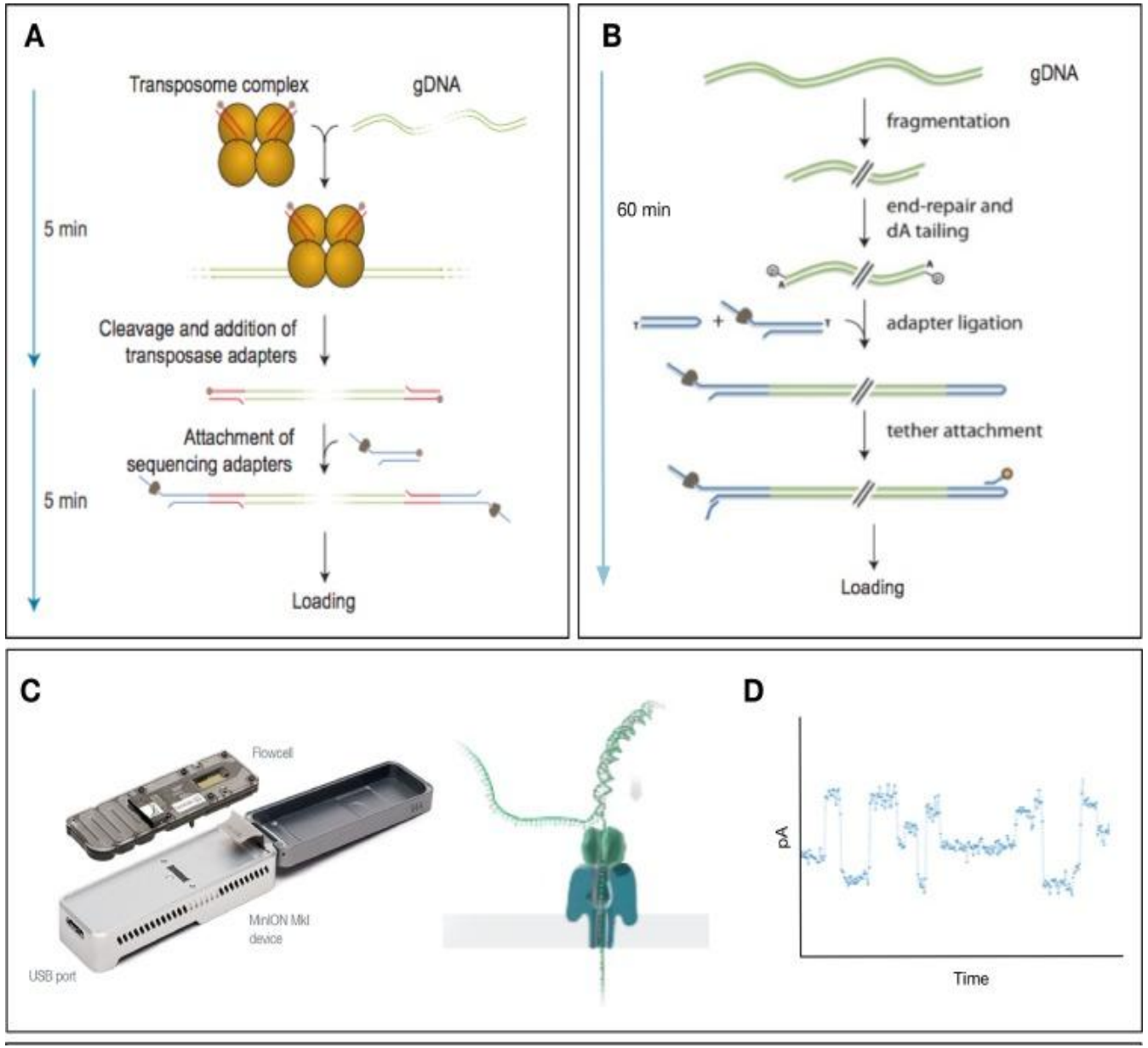
https://www.bmkgene.com/uploads/7-1.png

Third Generation Sequencing

Platform: Oxford Nanopore

Nanopore Sequencing

- Read Length:
  - Short Fragment: >20 bp
  - Standard: 5-30kbp
  - Ultra-long: >50kbp
- Throughput: 35Gb to 240Gb



https://ars.els-cdn.com/content/image/1-s2.0-S2214552418300245-gr1.jpg

Comparison of Technologies

	First generation	Second generation[a]	Third generation[a]
Fundamental technology	Size-separation of specifically end-labeled DNA fragments, produced by SBS or degradation	Wash-and-scan SBS	SBS, by degradation, or direct physical inspection of the DNA molecule
Resolution	Averaged across many copies of the DNA molecule being sequenced	Averaged across many copies of the DNA molecule being sequenced	Single-molecule resolution
Current raw read accuracy	High	High	Moderate
Current read length	Moderate (800–1000 bp)	Short, generally much shorter than Sanger sequencing	Long, 1000 bp and longer in commercial systems
Current throughput	Low	High	Moderate
Current cost	High cost per base Low cost per run	Low cost per base High cost per run	Low-to-moderate cost per base Low cost per run
RNA-sequencing method	cDNA sequencing	cDNA sequencing	Direct RNA sequencing and cDNA sequencing
Time from start of sequencing reaction to result	Hours	Days	Hours
Sample preparation	Moderately complex, PCR amplification not required	Complex, PCR amplification required	Ranges from complex to very simple depending on technology
Data analysis	Routine	Complex because of large data volumes and because short reads complicate assembly and alignment algorithms	Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges
Primary results	Base calls with quality values	Base calls with quality values	Base calls with quality values, potentially other base information such as kinetics

[a]There are many TGS technologies in development but few have been reduced to practice. While there is significant potential of TGS to radically improve current throughput and read-length characteristics (among others), the ultimate practical limits of these technologies remain to be explored. Furthermore, there is active development of SGS technologies that will also improve read-length and throughput characteristics.

WHO Collaborating Centre for Reference and Research on Influenza
VIDRL

World Health Organization
Philippines

KDCA
Korea Disease Control and Prevention Agency

NGS DATA ANALYSIS AND CONSIDERATIONS

CENTRE FOR PATHOGEN GENOMICS

Doherty Institute

THE UNIVERSITY OF MELBOURNE

The Royal Melbourne Hospital

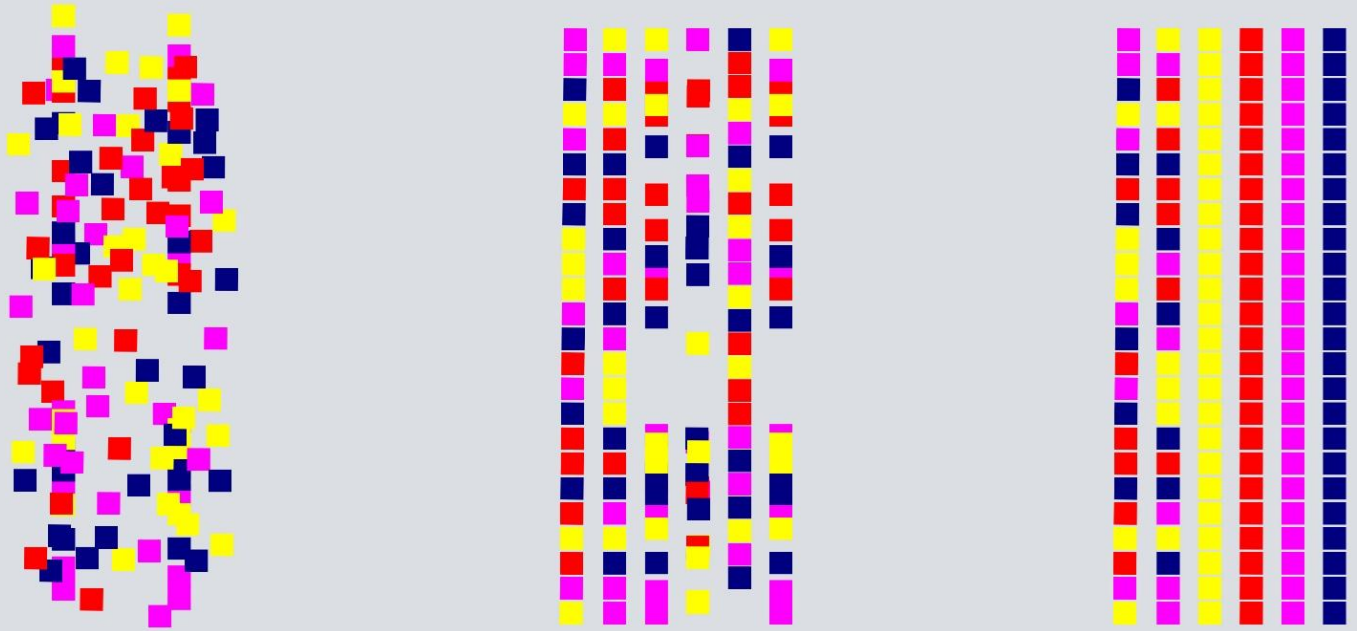
A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Big Data

Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

https://www.gartner.com/en/information-technology/glossary/big-data

BIG DATA ANALYTICS **DECISIONS**



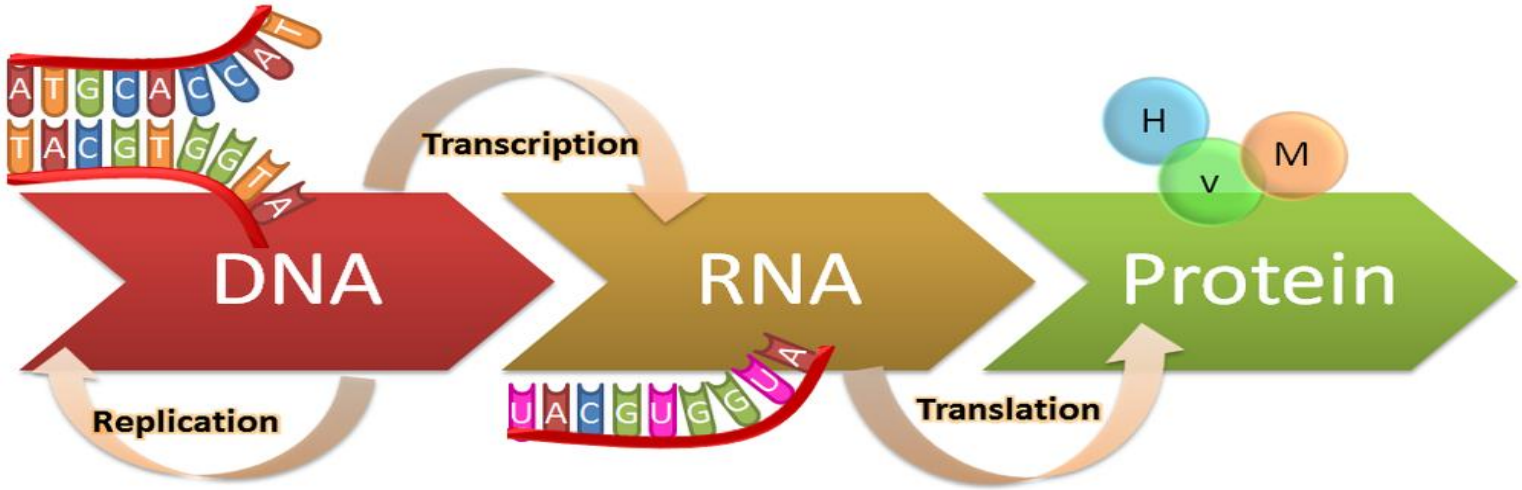
https://www.informationbuilders.com/big-data-analytics

Omics - The Big Data Biology

CENTRE FOR PATHOGEN GENOMICS
Doherty Institute
THE UNIVERSITY OF MELBOURNE
The Royal Melbourne Hospital
A joint venture between The University of Melbourne and The Royal Melbourne Hospital

-ome, -omics

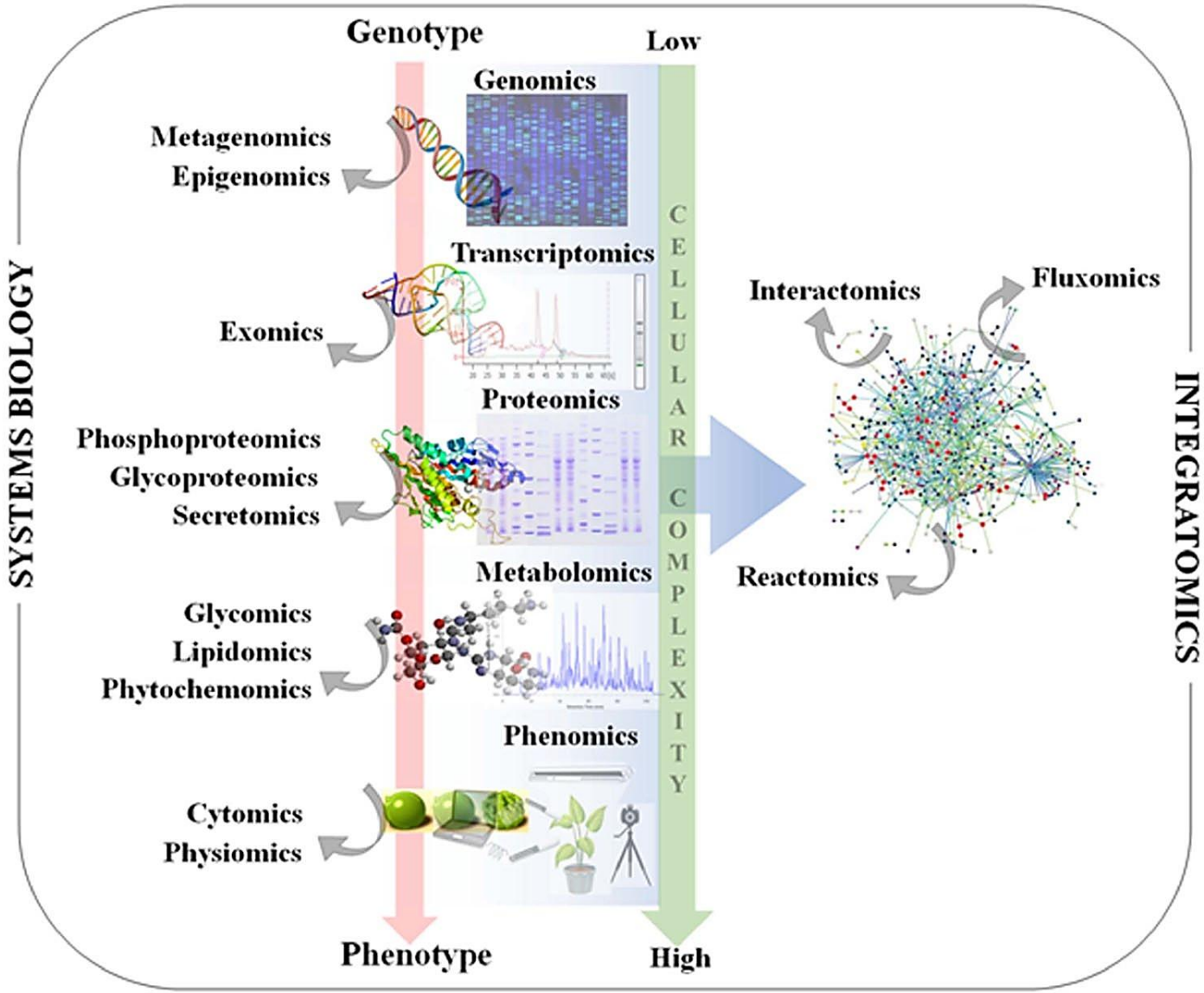
In molecular biology, suffixes used to refer to a ***totality*** of some sort



GENOME **TRANSCRIPTOME** **PROTEOME**

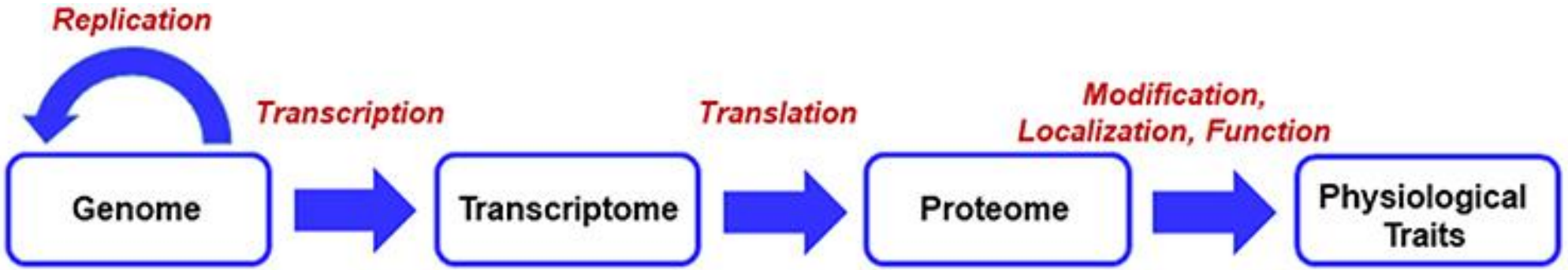
https://genius.com/Biology-genius-the-central-dogma-annotated

Omics - The Big Data Biology



https://bioone.org/ContentImages/Journals/wees/66/6/wsc.2018.33/graphic/f01_681.jpg

Omics - The Big Data Biology



SNPs, CNVs, SVs

Technologies:
PCR-RFLP, TaqMan assay
SNP/CNV arrays
NGS

mRNA expression

Technologies:
Northern blotting
qRT-PCR
SAGE
CAGE
cDNA/oligo gene arrays
Exon arrays
Tiling arrays
RNA-seq

Protein expression

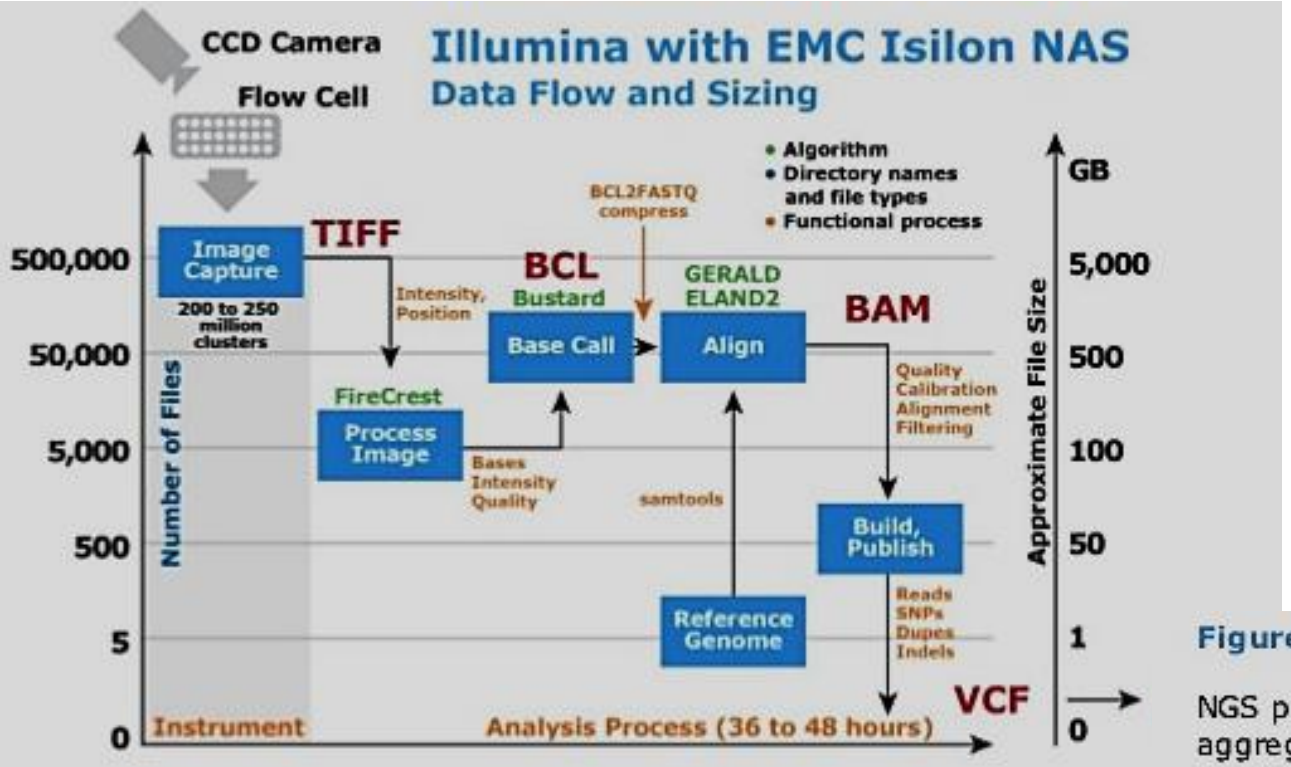
Technologies:
Western blotting
2D-PAGE
2D-DIGE
ICAT, iTRAQ, SILAC
LC-MS/MS
Antibody arrays
Aptamer arrays

Continuous phenotypes

Examples:
Height
BMI
BMD
Fasting glucose
HDL-cholesterol
LDL-cholesterol
Triglycerides
SBP
DBP

https://www.frontiersin.org/files/Articles/386438/fgene-10-00110-HTML/image_m/fgene-10-00110-g001.jpg

How huge are NGS data sets?



"… the process generates over **500,000 [files]** having aggregate size of greater than **5 Tb** over the course of the **48-hour run**."

Figure 4. Data flow using Illumina NGS process

NGS production processes generate potentially millions of files with terabytes of aggregate storage impacting the capacity and manageability limits of existing file server structures.

Figure 4 shows the data flow including a file number and capacity summary of an actual NGS process using Illumina sequencer and Isilon scale-out NAS storage. As can be seen, the process generates over 500,000 having aggregate size of greater than 5 TB over the course of the 48-hour run.

https://image.slidesharecdn.com/h10961-wp-ngs-sizing-perf-120912232830-phpapp01/95/white-paper-nextgeneration-genome-sequencing-using-emc-isilon-scaleout-nas-sizing-and-performance-guidelines-10-728.jpg?cb=1347492622

How huge are NGS data sets?

IEEE Spectr. Author manuscript; available in PMC 2014 Jun 9.
Published in final edited form as:
IEEE Spectr. 2013 Jul; 50(7): 26–33.
doi: 10.1109/MSPEC.2013.6545119

PMCID: PMC4048922
NIHMSID: NIHMS563699
PMID: 24920863

The DNA Data Deluge

Fast, efficient genome sequencing machines are spewing out more data than geneticists can analyze

Michael C. Schatz and Ben Langmead

The roughly 2000 sequencing instruments in labs and hospitals around the world can collectively sequence 15 quadrillion nucleotides per year, which equals about 15 petabytes of compressed genetic data. A petabyte is 2^{50} bytes, or in round numbers, 1000 terabytes. To put this into perspective, if you were to write this data onto standard DVDs, the resulting stack would be more than 2 miles tall. And with sequencing capacity increasing at a rate of around three- to fivefold per year, next year the stack would be around 6 to 10 miles tall. At this rate, within the next five years the stack of DVDs could reach higher than the orbit of the International Space Station.

Handling NGS data sets

Bioinformatics (Applied)

"Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data."

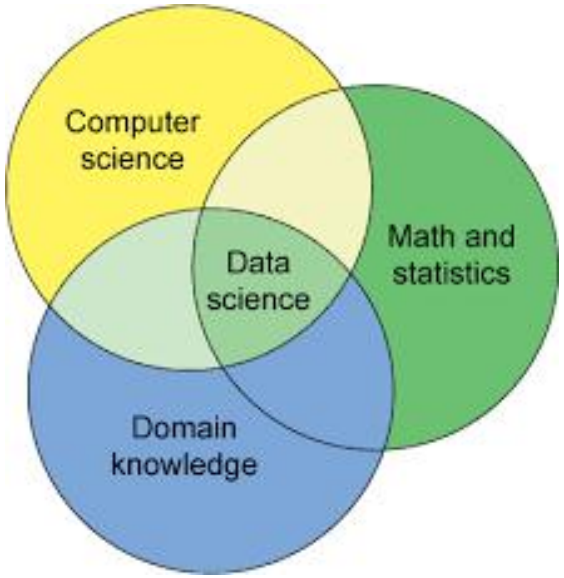
Computational Biology (Theoretical)

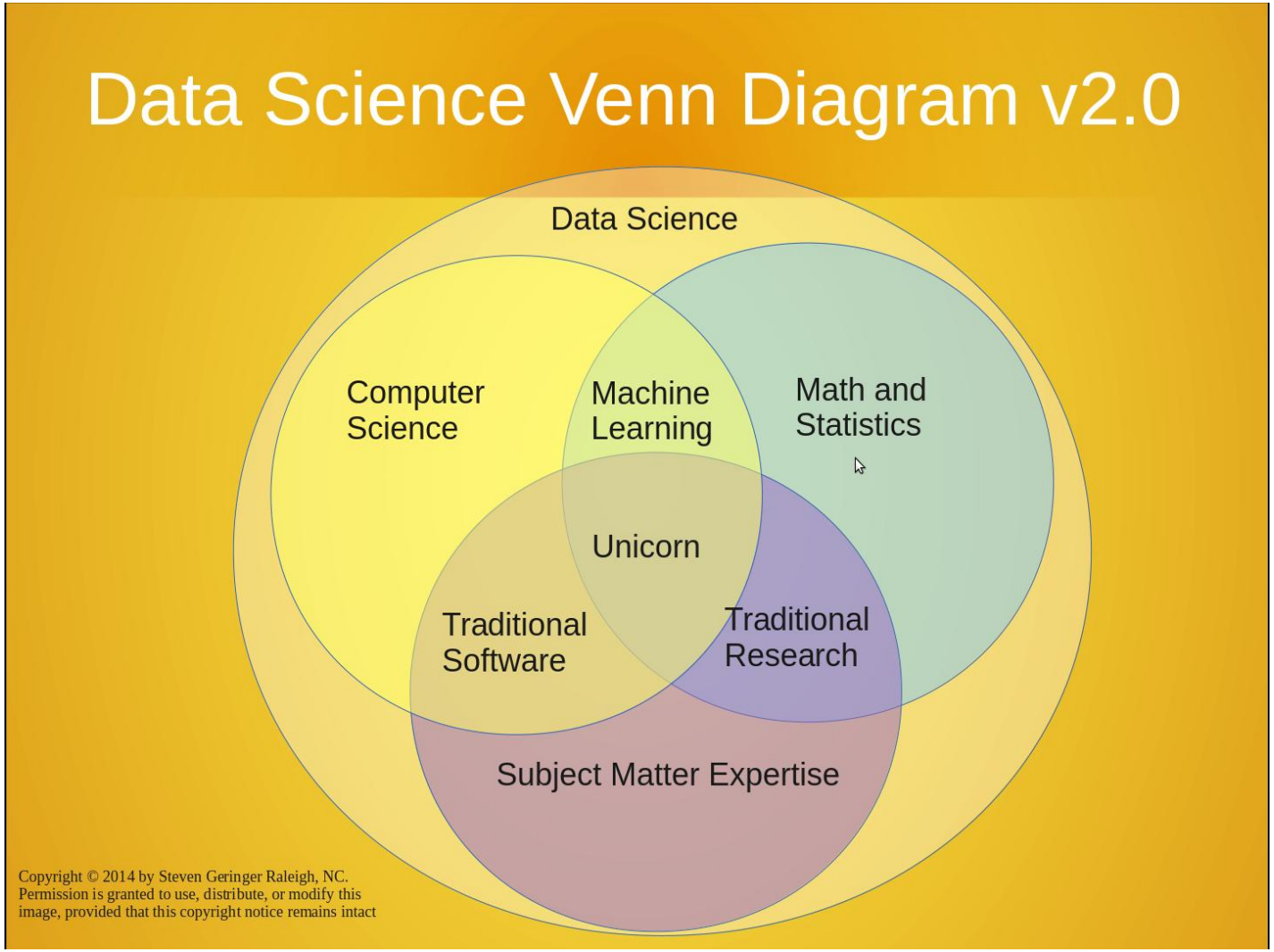
"The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems."

NIH Working Definition of Bioinformatics and Computational Biology (Huerta et al. 2000)

Handling NGS data sets

Data Science





https://www.stat.wisc.edu/~wahba/binyu.on.datascience.pdf
http://www.anlytcs.com/2014/01/data-science-venn-diagram-v20.html

NGS Analogy

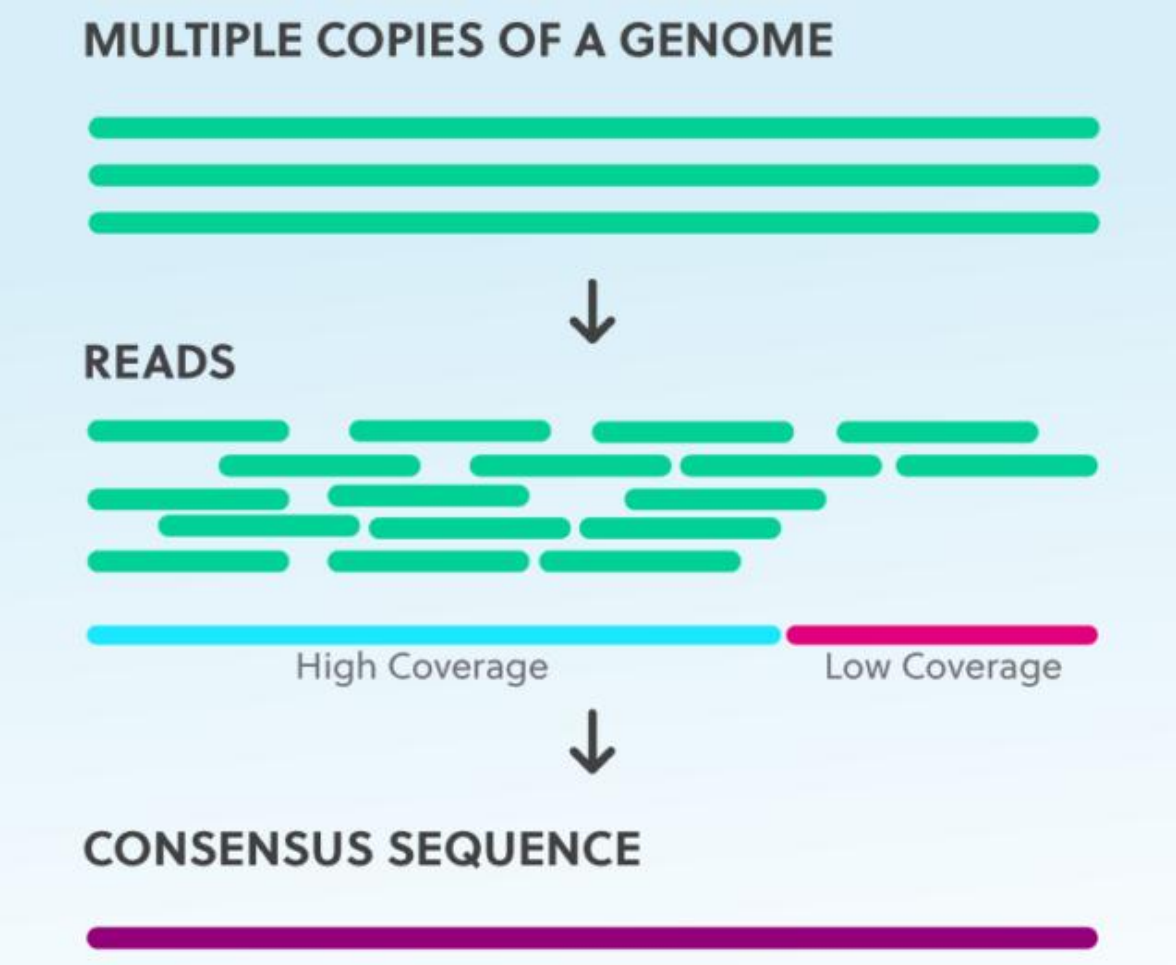
Multiple copies of template genome

Sequence reads / fragments

Reconstructed sequence with readable features

Adapted from: https://www.slideshare.net/AustralianBioinformatics/de-novo-assembly-of-ngs-data-torsten-seemann-imb-4-jul-2011

NGS Analysis



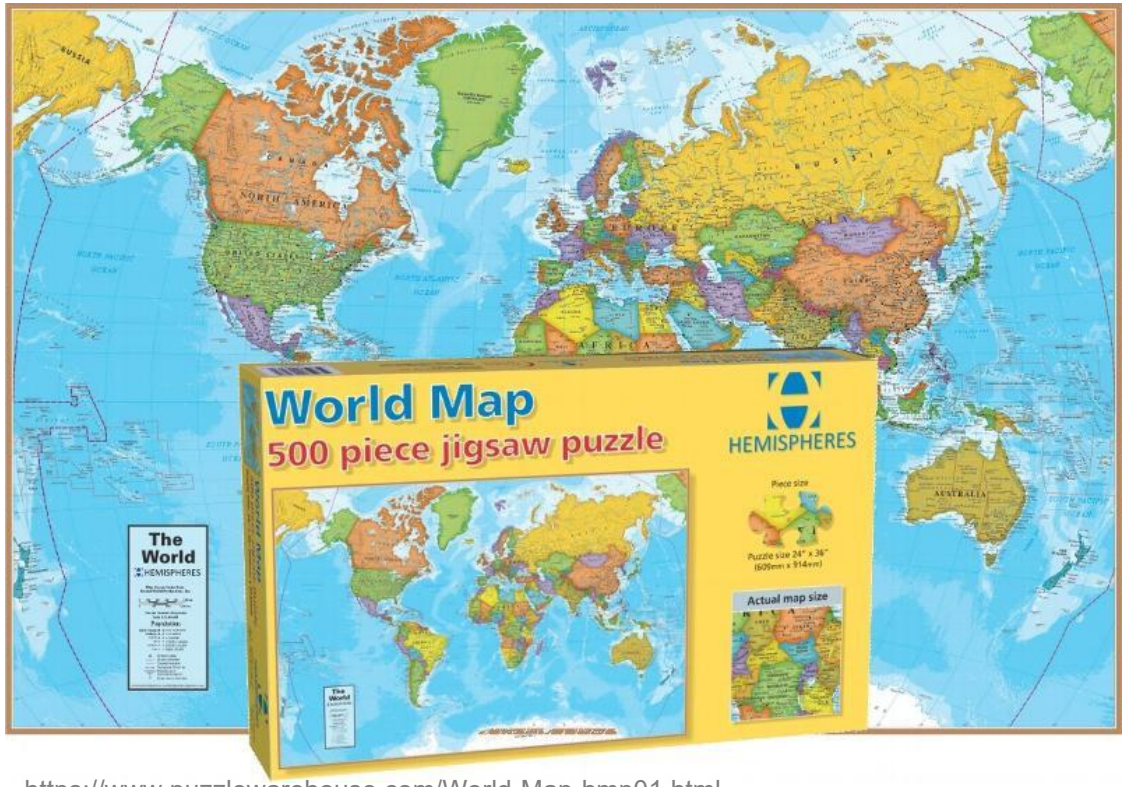
Sequence Reads

- 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)

https://irepertoire.com/wp-content/uploads/2020/05/NGS-Considerations-Sequencing-Workflow-1024x768.png

NGS Analysis

https://www.123rf.com/photo_104112068_stock-illustration-missing-jigsaw-puzzle-pieces-in-unfinished-work-concept-white-pattern-texture-background-3d-illustra.html



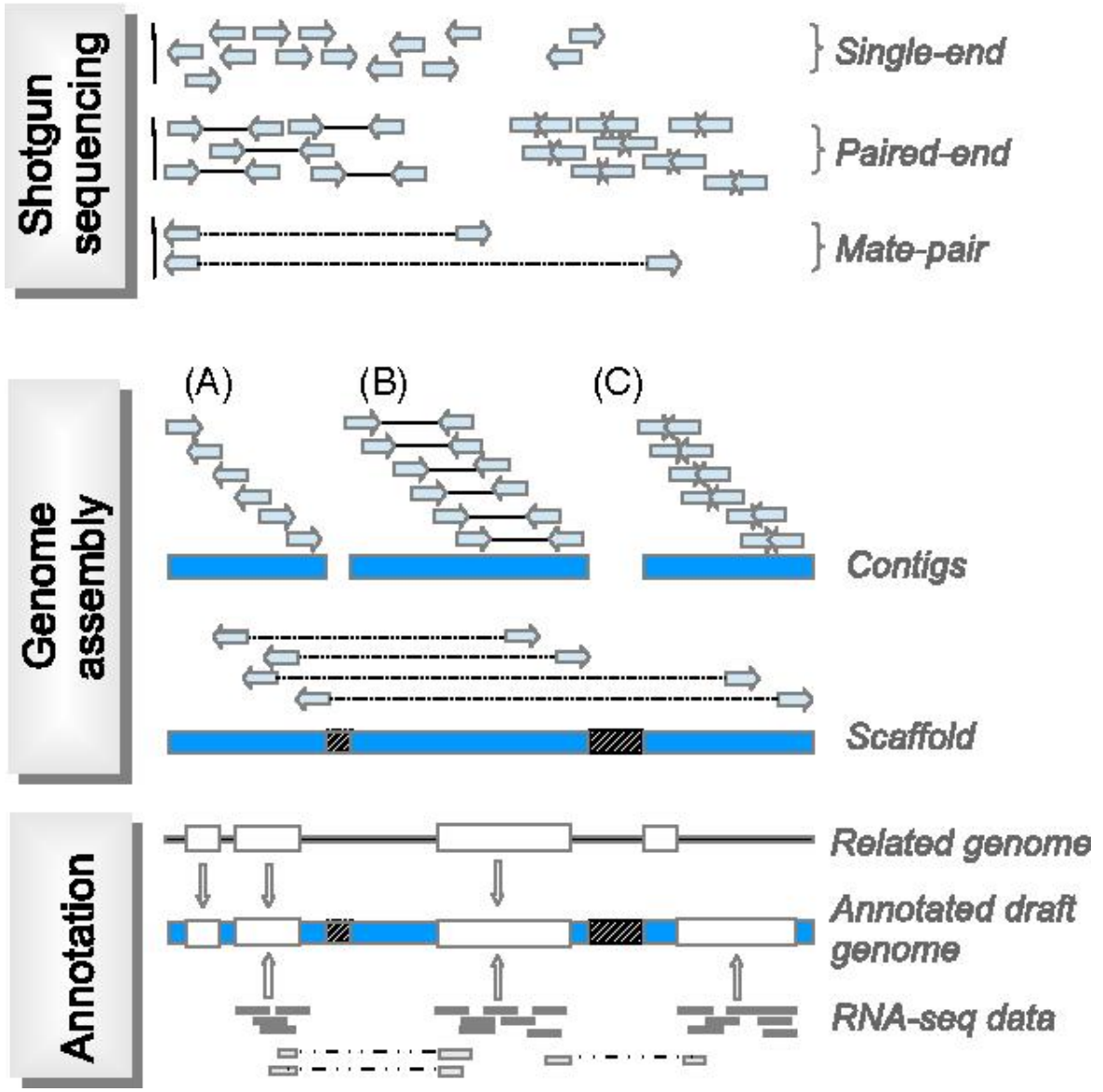
https://www.puzzlewarehouse.com/World-Map-hmp01.html

De novo

Reference-guided

NGS Analysis

***De novo* Workflow**



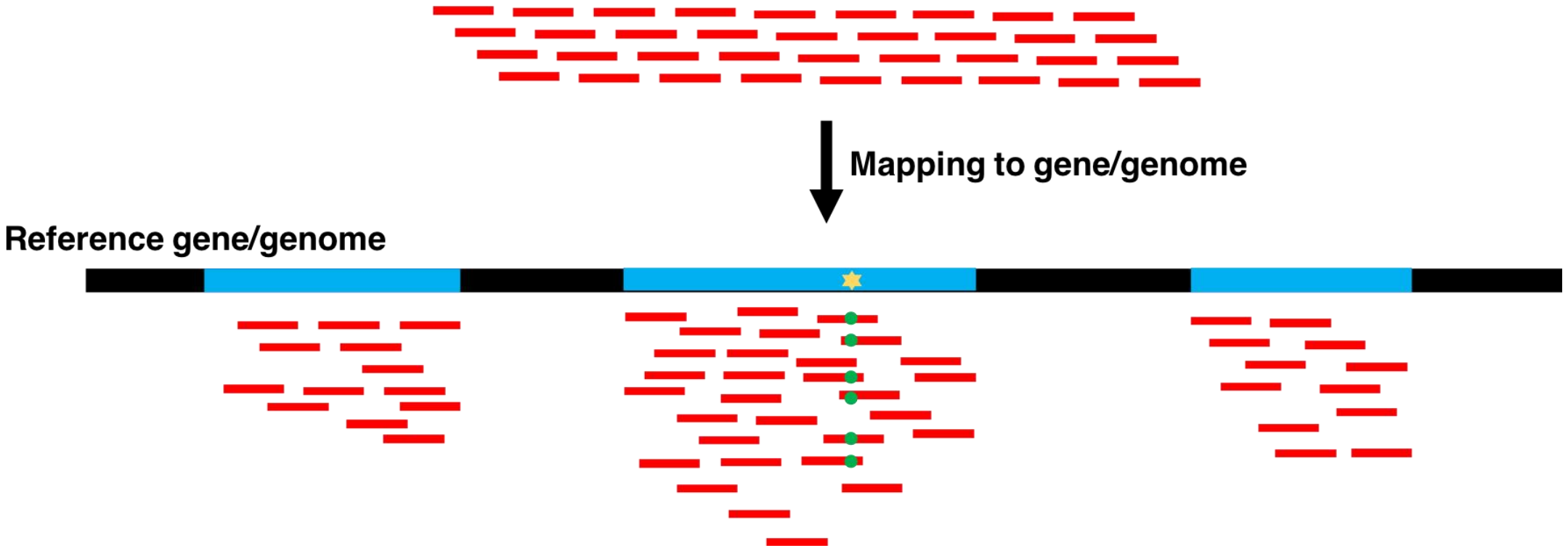
Output: FASTA file

Output: GFF, GenBank, EMBL

https://ai2-s2-public.s3.amazonaws.com/figures/2017-08-08/f0fe28758c4e9e519e66f048df30ec211499b5a8/5-Figure2-1.png

NGS Analysis

Reference-Guided Workflow



Output: SAM, BAM

https://galaxyproject.github.io/training-material/topics/proteomics/images/variant_calling.png

NGS Analysis

Note on file format standardization

- In bioinformatics, many of the processes involve input and output files with standard formats
- File format standardization enables process automation

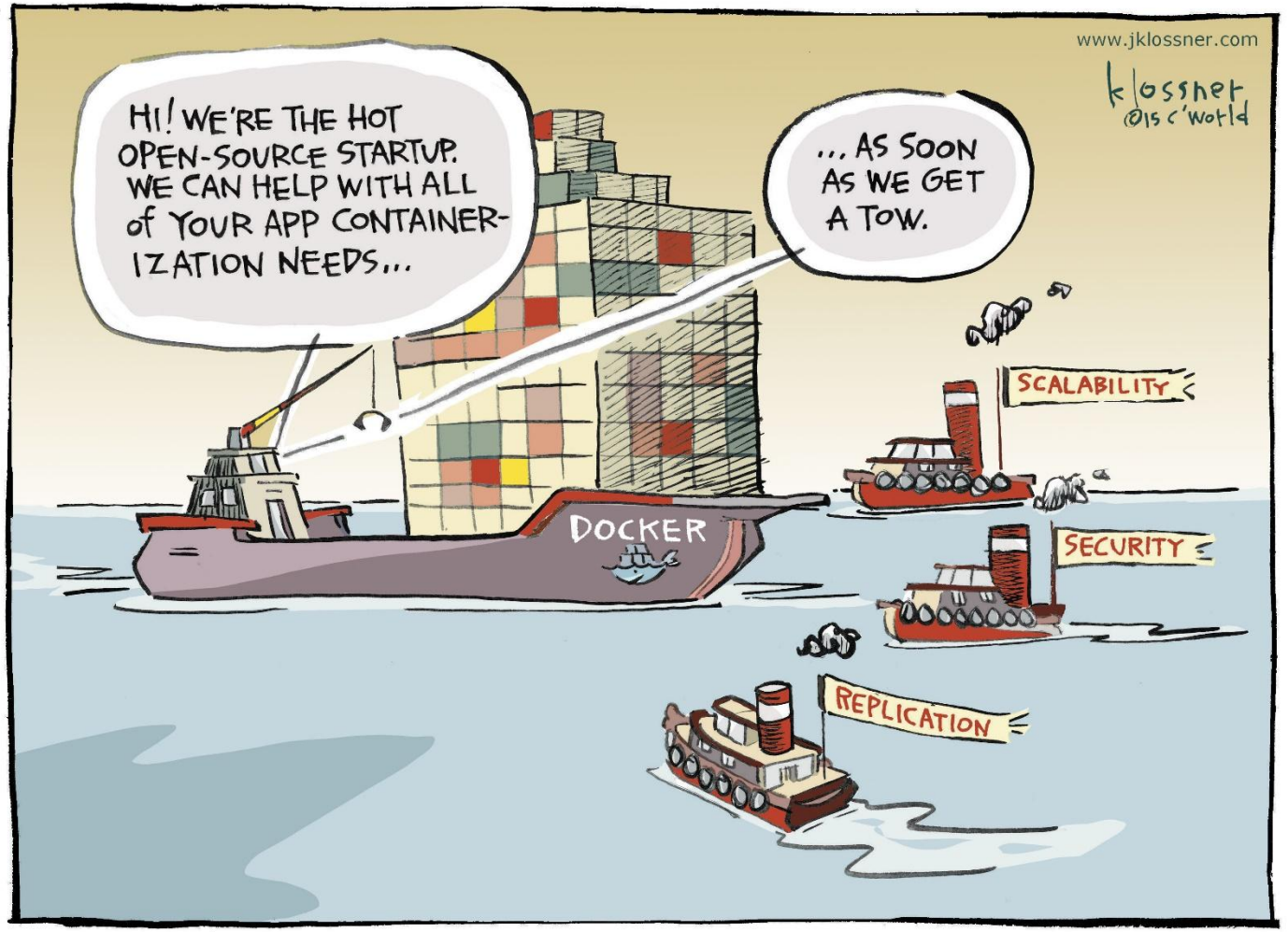
Process	Input	Output
Quality Control	FASTQ	FASTQ
Sequence Assembly	FASTQ	FASTA
Feature Annotation	FASTA	GFF, GenBank, EMBL
Read Mapping	FASTQ	SAM / BAM
Variant Calling	SAM / BAM	VCF
Alignment	Multi-FASTA	Multi-FASTA / Phylip
Phylogenetics	Multi-FASTA / Phylip	Newick / Nexus

Considerations in NGS Data Analysis

CENTRE FOR PATHOGEN GENOMICS
Doherty Institute
THE UNIVERSITY OF MELBOURNE
The Royal Melbourne Hospital
A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Infrastructure

- Computing power
- Storage
- Network
- Security

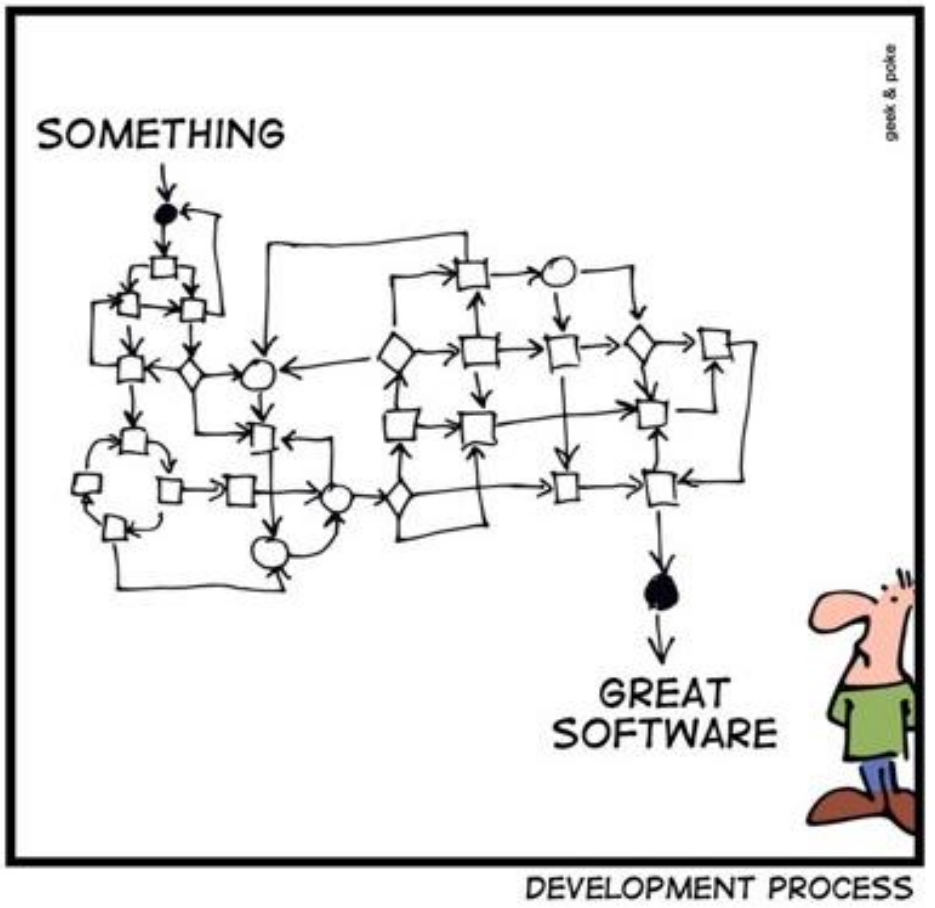


https://images.techhive.com/images/article/2015/12/cw_2015_02_klossner_docker_containerization_onetimeuse_rsz-100633562-orig.jpg

Considerations in NGS Data Analysis

Algorithms / Implementation

- Statistical / Mathematical models (accuracy)
- Databases and Data Management
- Speed
- Usage of resources



http://www.datamation.com/news/tech-comics-software-development-2.html

Considerations in NGS Data Analysis

CENTRE FOR PATHOGEN GENOMICS
Doherty Institute
THE UNIVERSITY OF MELBOURNE
The Royal Melbourne Hospital
A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Skillset / Learning Curve / Manpower

- Biologists trying to understand computational concepts
- Computational scientists trying to understand biology

Biologists be like …

Computational scientists be like …

Computational scientists: https://www.slideshare.net/iliasfl/data-science-at-atlassian-the-transition-towards-a-datadriven-organisation
Biologists: http://www.inf.fu-berlin.de/w/SE/ThesisFOSSErrors

World Health Organization
Philippines

CENTRE FOR PATHOGEN GENOMICS

WHO Collaborating Centre for Reference and Research on Influenza
VIDRL

QUESTIONS?

fatablizo@up.edu.ph
bioinformatics@pgc.up.edu.ph

KDCA
Korea Disease Control and Prevention Agency

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

THE UNIVERSITY OF MELBOURNE

The Royal Melbourne Hospital