Whole Genome Sequencing and Bioinformatics SeqAfrica Training

Marco van Zwetselaar

Niamh Lacy-Roberts

Day 1

The Fleming Fund
Regional Grants

Introductions and Goals

Days 1-2

- Understanding sequencing technologies
- Using bioinformatics online tools
- Understanding data quality and performing quality control (QC) and assembly

Days 3-4

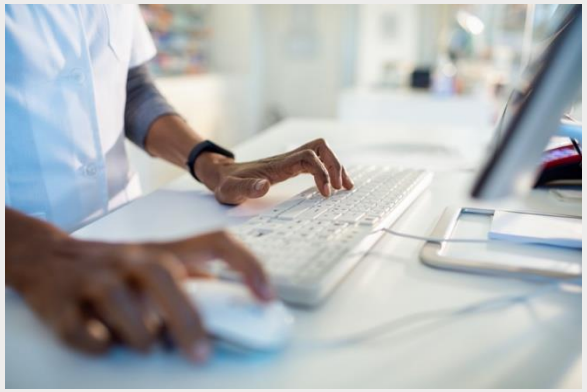
- Understanding antimicrobial resistance (AMR)
- Performing bacterial typing and detecting AMR genes
- Understanding phylogenetics and visualising phylogenetic relationships

https://github.com/seqafrica/zambia-course

Agenda

- Day 1: Introduction to WGS and Online bioinformatics Basics
- Day 2: Quality Control and Assembly
- Day 3: Typing and AMR analysis using Online Platforms
- Day 4: Typing and Phylogenetic Analysis via Online Tools







Introduction to Whole Genome Sequencing (WGS) technologies

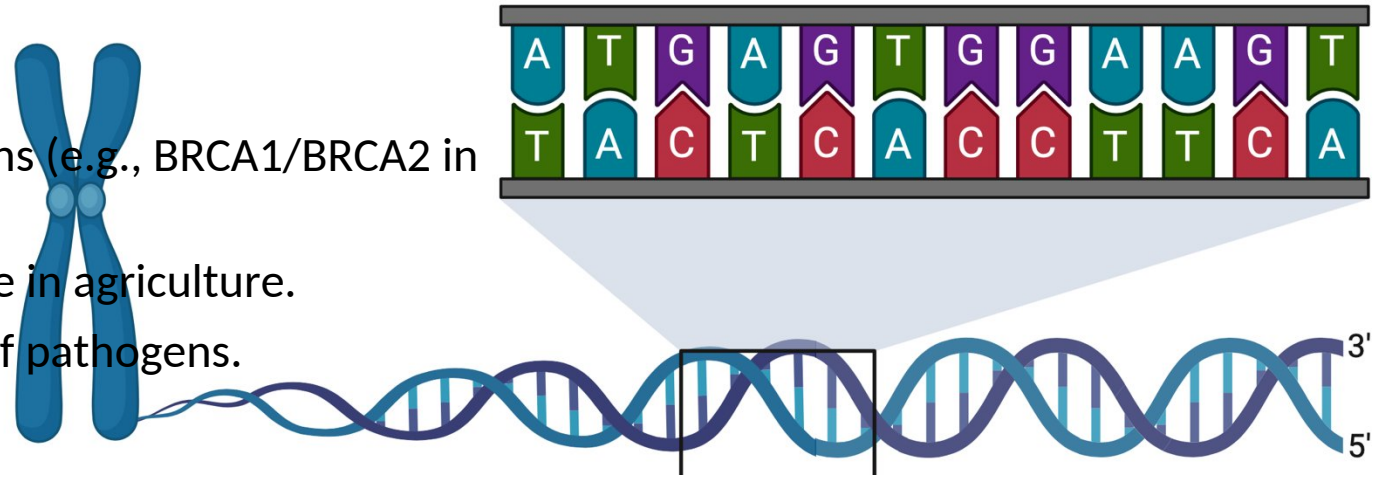
Genomics

- Genomics is the branch of biology that focuses on the structure, function, evolution, and mapping of entire genomes (the complete set of DNA in an organism, including all its genes).

- Key Aspects:
  - Sequencing genomes to decode the DNA blueprint of life.
  - Understanding gene interactions and regulatory mechanisms.
  - Studying genetic variations and their effects on organisms.

- Applications:
  - Identifying disease-causing mutations (e.g., BRCA1/BRCA2 in cancer).
  - Enhancing crop yields and resistance in agriculture.
  - Tracking the spread and evolution of pathogens.

Figure created with BioRender.com



Bioinformatics

- Bioinformatics combines biology, computer science, and statistics to manage, analyse, and interpret biological data, particularly large datasets like DNA sequences.

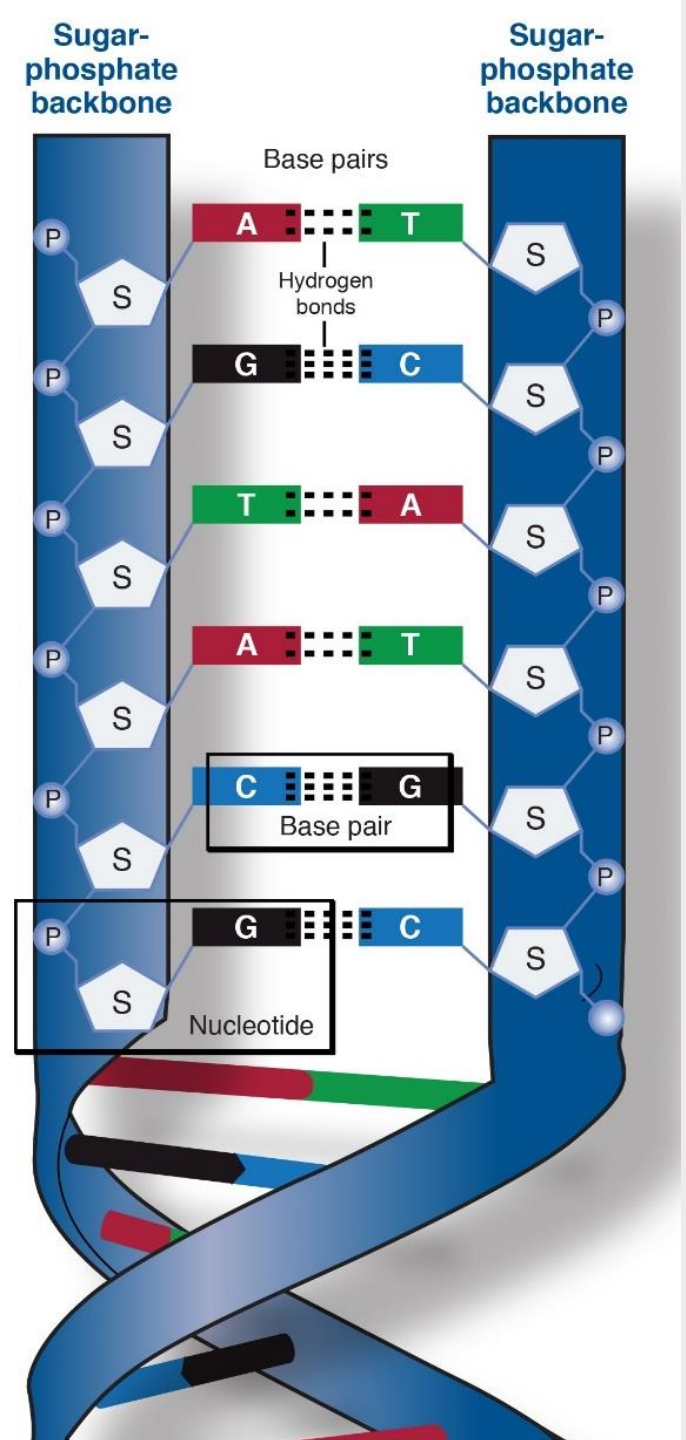
- **Core Tasks**:
  - Storing and organizing genomic data.
  - Developing algorithms for data analysis (e.g., BLAST for sequence alignment).
  - Creating software tools for genomics workflows.

- **Why It's Needed**:
  - Modern sequencing generates terabytes of data that cannot be processed manually.

Figure created with BioRender.com

DNA

- blueprints for all cell components are stored as a long polymer of DeoxyriboNucleic Acid (DNA).
- Each unit or building block in DNA is referred to as a nucleotide.
- Each nucleotide consists of two parts a sugar-phosphate and one of four different bases, Adenine (A), Thymine (T), Guanine (G) or Cytosine (C).
- The (covalently) bound sugar-phosphates forms the backbone of the DNA molecule.
- DNA in cells are stored as a double stranded helix, each nucleotide facing a complementary base on the opposite strand.
- Because bases on each strand are complementary, stating the base on one strand, informs us of it the identity of its partner. The base pair is essentially the fundamental unit of DNA



Courtesy: National Human Genome Research Institute https://www.genome.gov/genetics-glossary/Phosphate-Backbone, *This image is a work of the National Institutes of Health and is in the public domain*

Bacterial genome

- In bacteria, DNA is stored in circular chromosomes.
- Most bacteria have 1 chromosome containing essential genes for survival, with notable exceptions e.g. *Vibrio cholerae*.
- Bacteria can also contain additional smaller circular DNA molecules, referred to as plasmids.
- Plasmids usually contain gene non-essential genes, which can confer advantages to the cell in specific cases, such as antimicrobial resistance genes.
- The genome refers to all the DNA in the cell
- Genome = Chromosome(s) + Plasmids

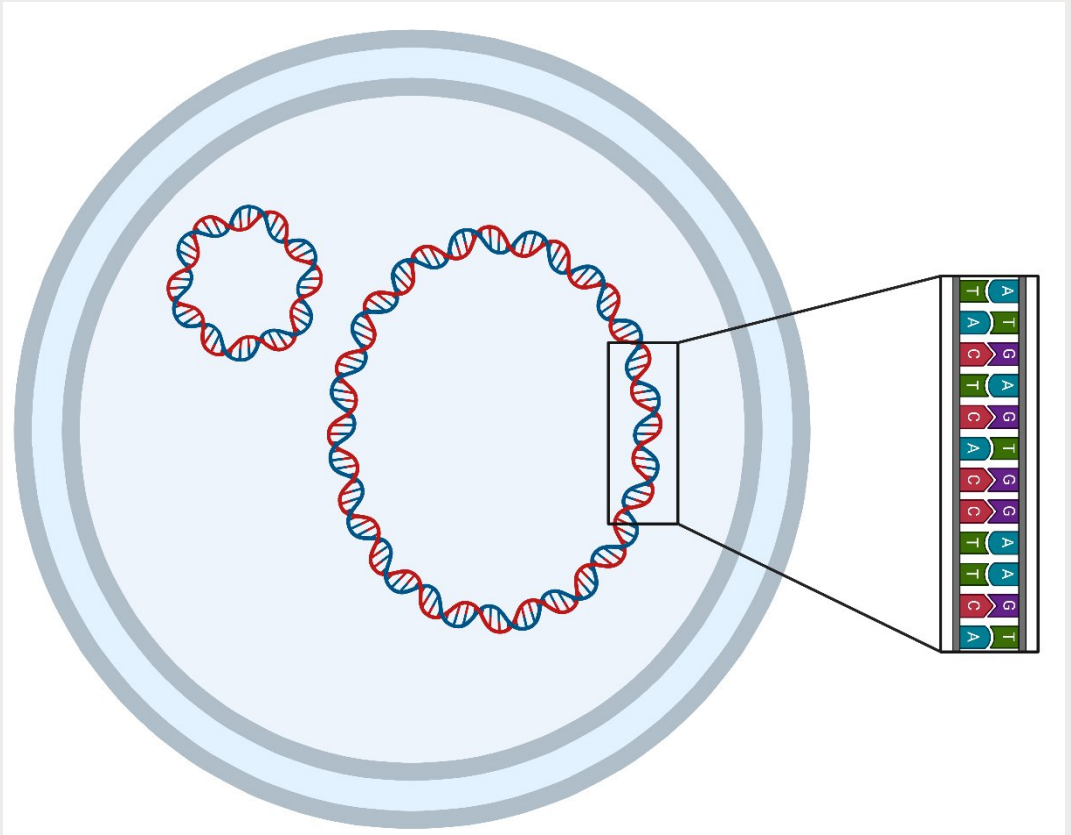


Figure created with BioRender.com

Central Dogma of Molecular Biology

- **Concept:** DNA → RNA → Protein.
- Genomics focuses on the first step: decoding DNA sequences.
- Bioinformatics connects these steps by analysing how genetic data translates into cellular function.
- **Examples**:
- DNA mutation → Faulty protein → Disease.
- DNA mutation → Protein conveying resistance → AMR.

Figure created with BioRender.com

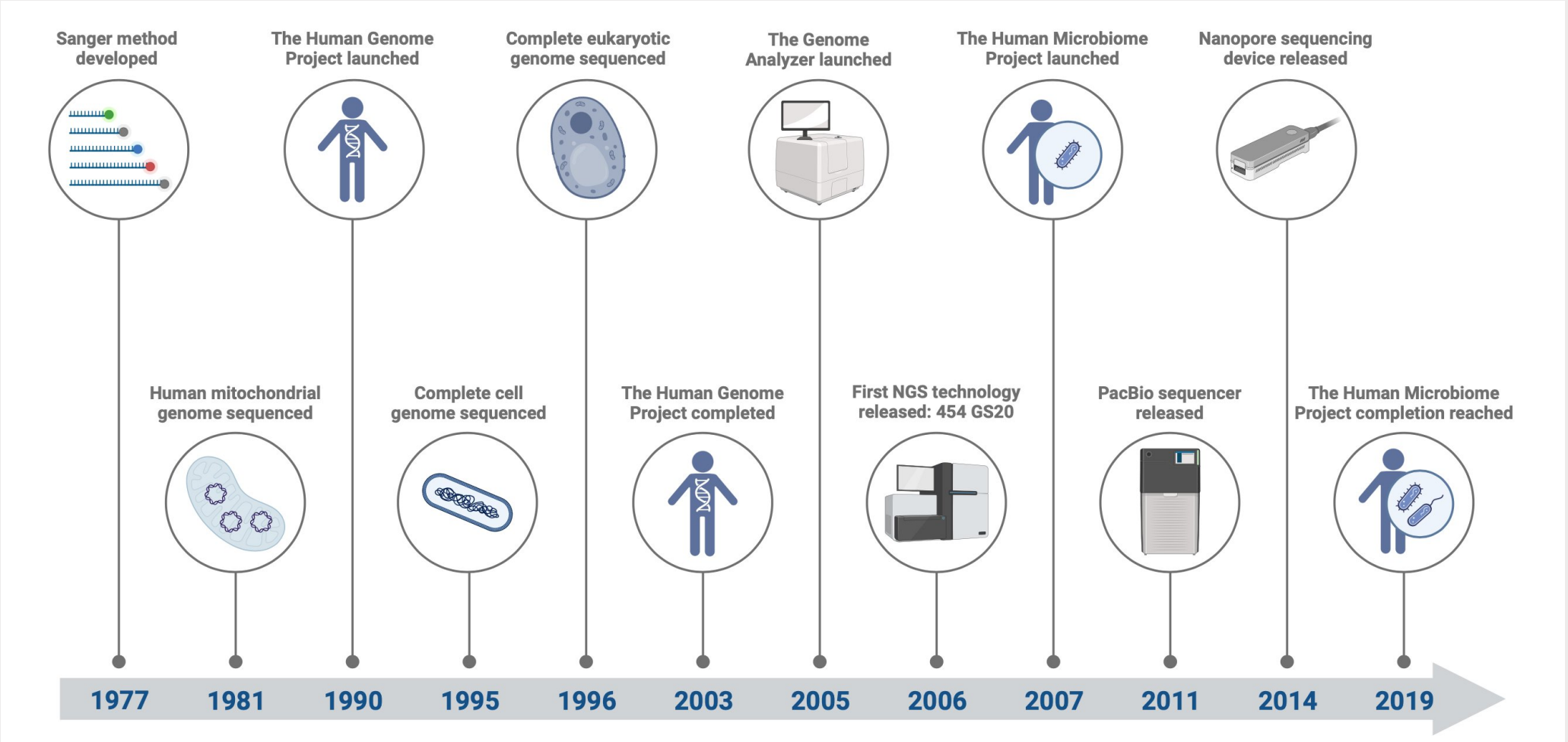
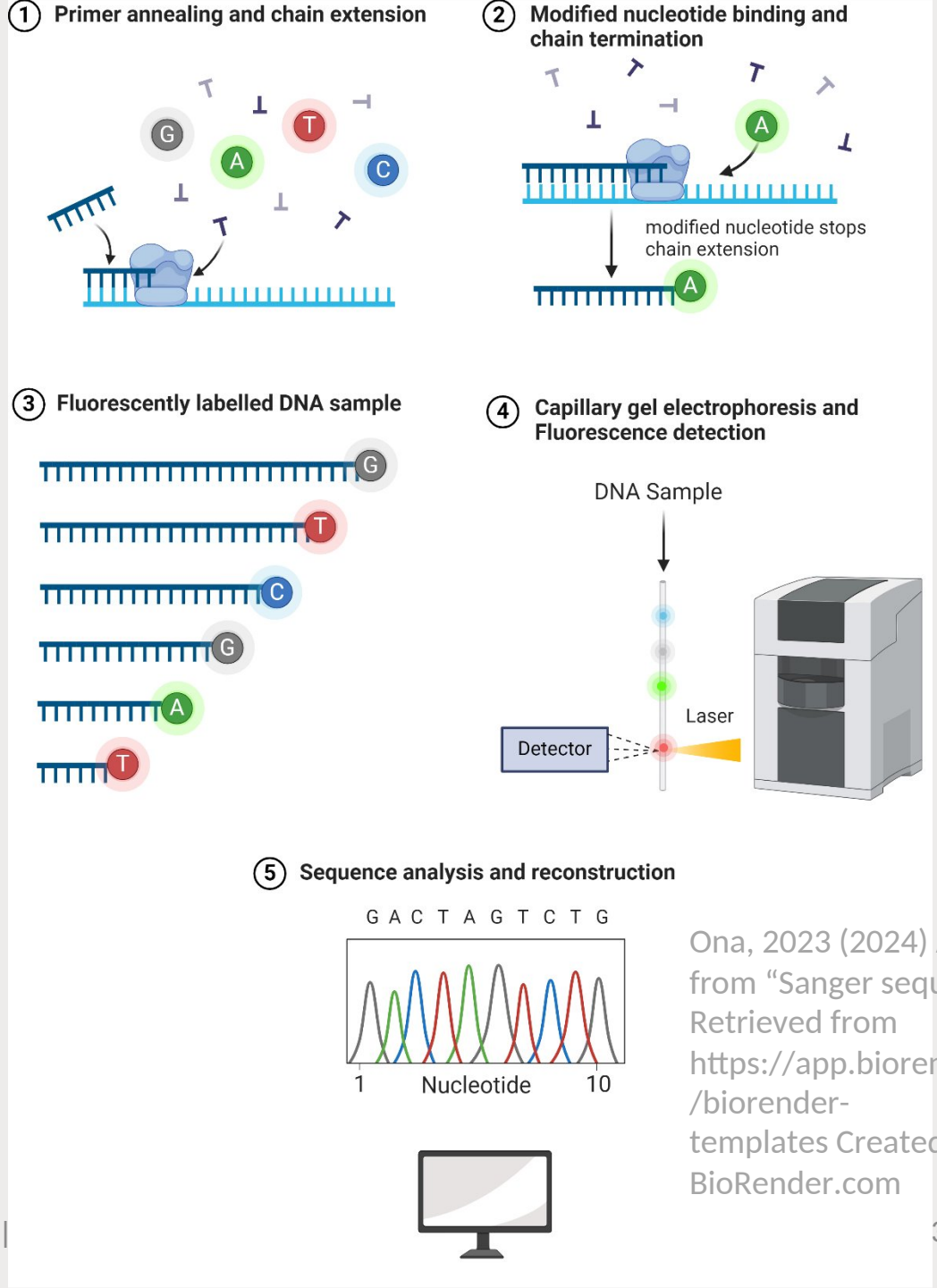


Figure created with BioRender.com

The Fleming Fund
Regional Grants

DNA Sequencing

- Sequencing refers to the process of determining the sequence of the individual basepairs in a stretch of DNA.
- Modern sequencing started with the Sanger-method first described in 1977.
- Sanger sequencing utilized the enzymes from the natural DNA replication machinery of cells combined with modified nucleotides to prematurely stop synthesis.
- Modified nucleotides contain a fluorescent dye which can be exited to elicit light. Different colors attached to each of the four bases (A, T, G and C) allows distinction between bases based on the color of the light.
- This method thus utilizes "sequencing-by-synthesis"



Ona, 2023 (2024) Adapted from "Sanger sequencing". Retrieved from https://app.biorender.com/biorender-templates Created in BioRender.com

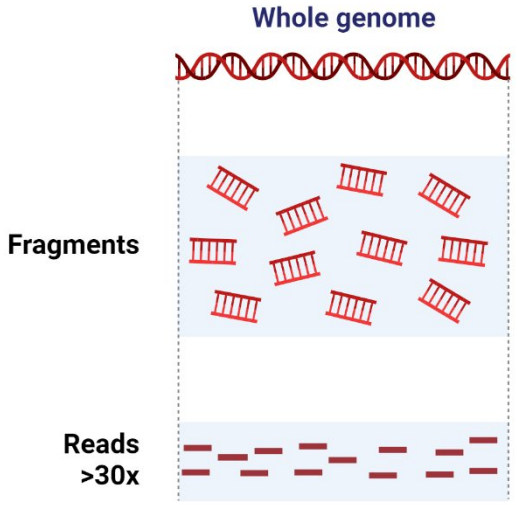
The Fleming Fund
Regional Grants

Sequencing approaches

- Host of different approaches developed for sequencing at different levels and specificities.

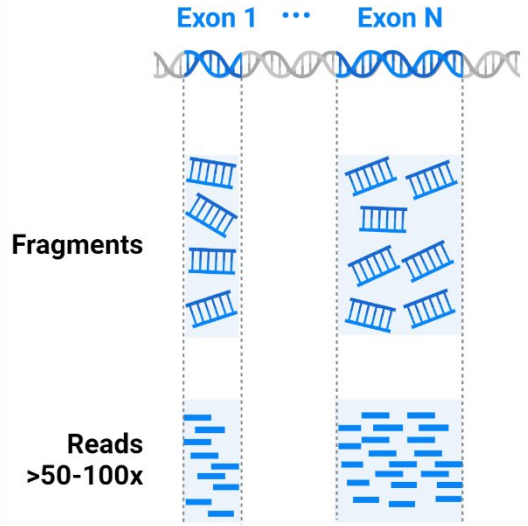
- For AMR surveillance and pathogens, the most relevant methods are WGS and targeted gene panels,
- single isolate and metagenomic sequencing.

Whole Genome Sequencing



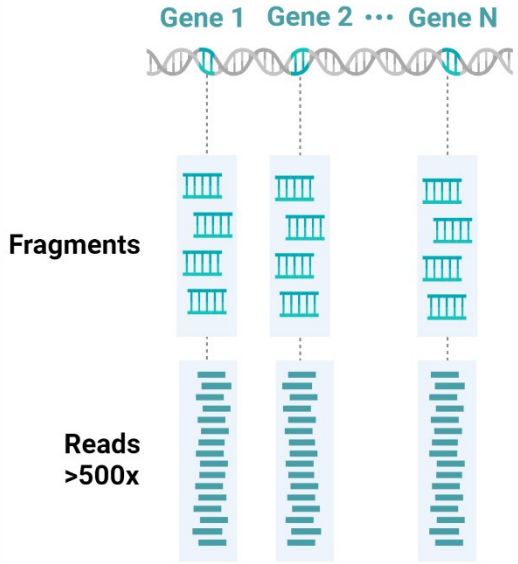
Coverage: All genes and non-coding DNA
Time: Longest turnaround time
Cost: Most expensive

Exome Sequencing



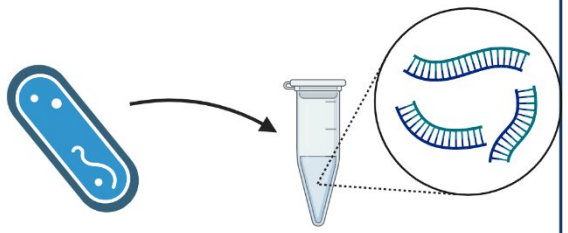
Coverage: Entire exome (20-25k genes)
Time: Long turnaround time
Cost: Cost-effective

Targeted Gene Panel



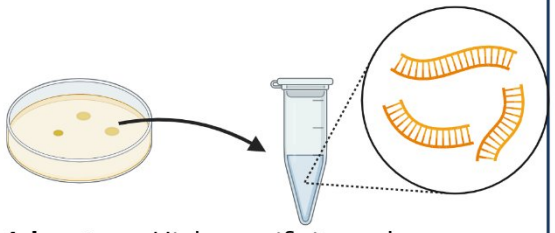
Coverage: 10-500 genes
Time: Relatively Rapid turnaround time
Cost: Most cost-effective

Single Cell



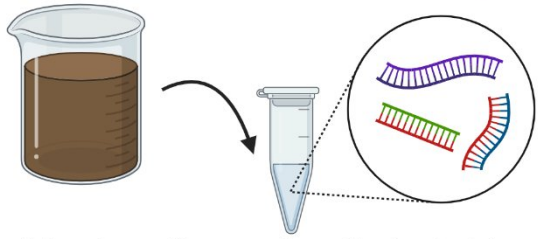
Advantage: Very high specificity
Disadvantage: Low representativeness and technically difficult

Single Isolate



Advantage: High specificity and easy analysis.
Disadvantage: strain must be culturable and pure

Metagenomic



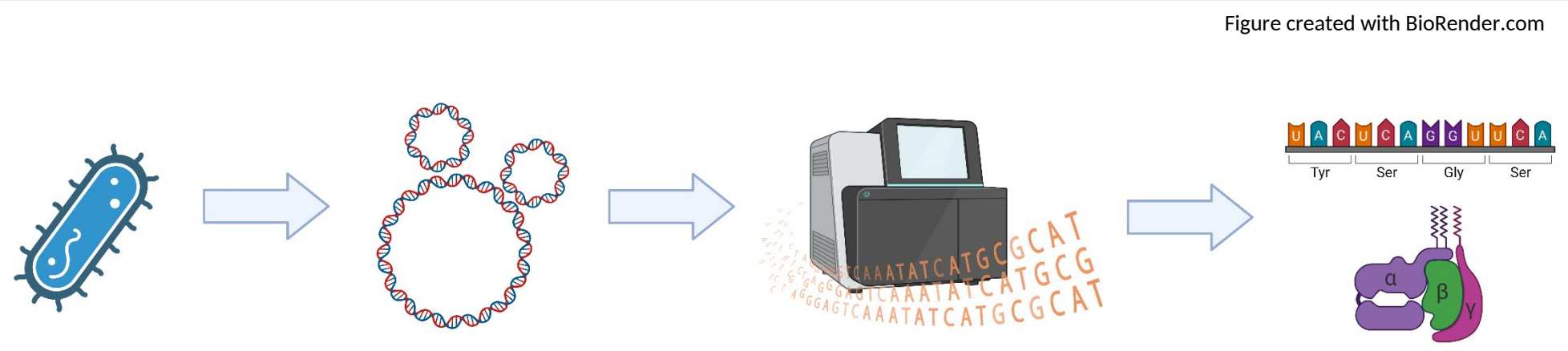
Advantage: Represents entire bacterial population, even unculturable spp.
Disadvantage: Very difficult to analyse and ensure proper coverage

Figure created with BioRender.com

Whole Genome Sequencing (WGS)

- In WGS we aim to sequence the entire genetic content of a bacteria (or human, virus etc…)

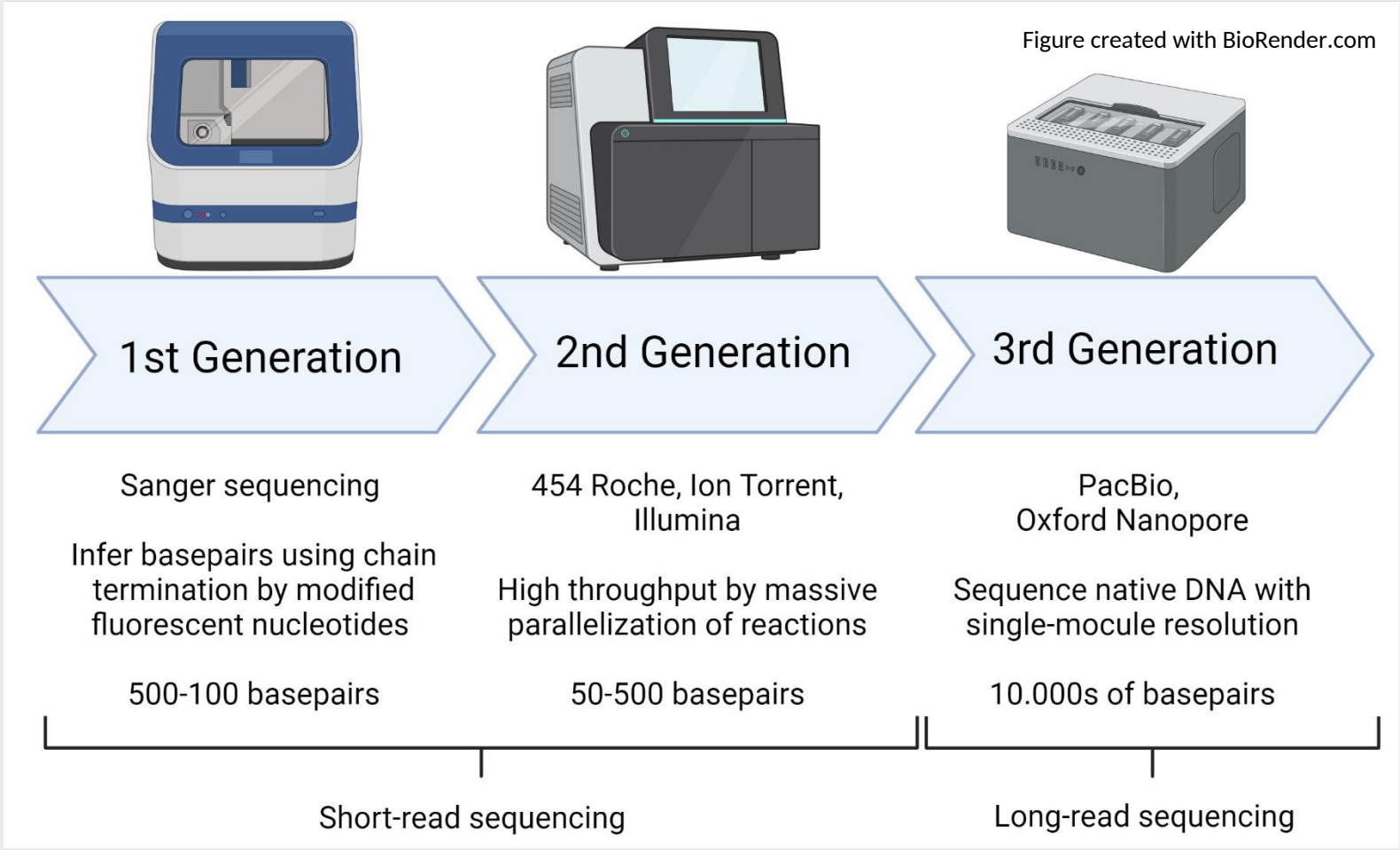
- Advantages of WGS:
- The entire genetic repertoire of the cell is available for analysis.
- Non-coding changes are captured.
- Large and small variations which might be missed by targeted sequencing are captured.
- Phylogenetic comparison at high depth (down to the single basepair)
- Rapid re-analysis for epidemiological investigations or when new knowledge becomes available

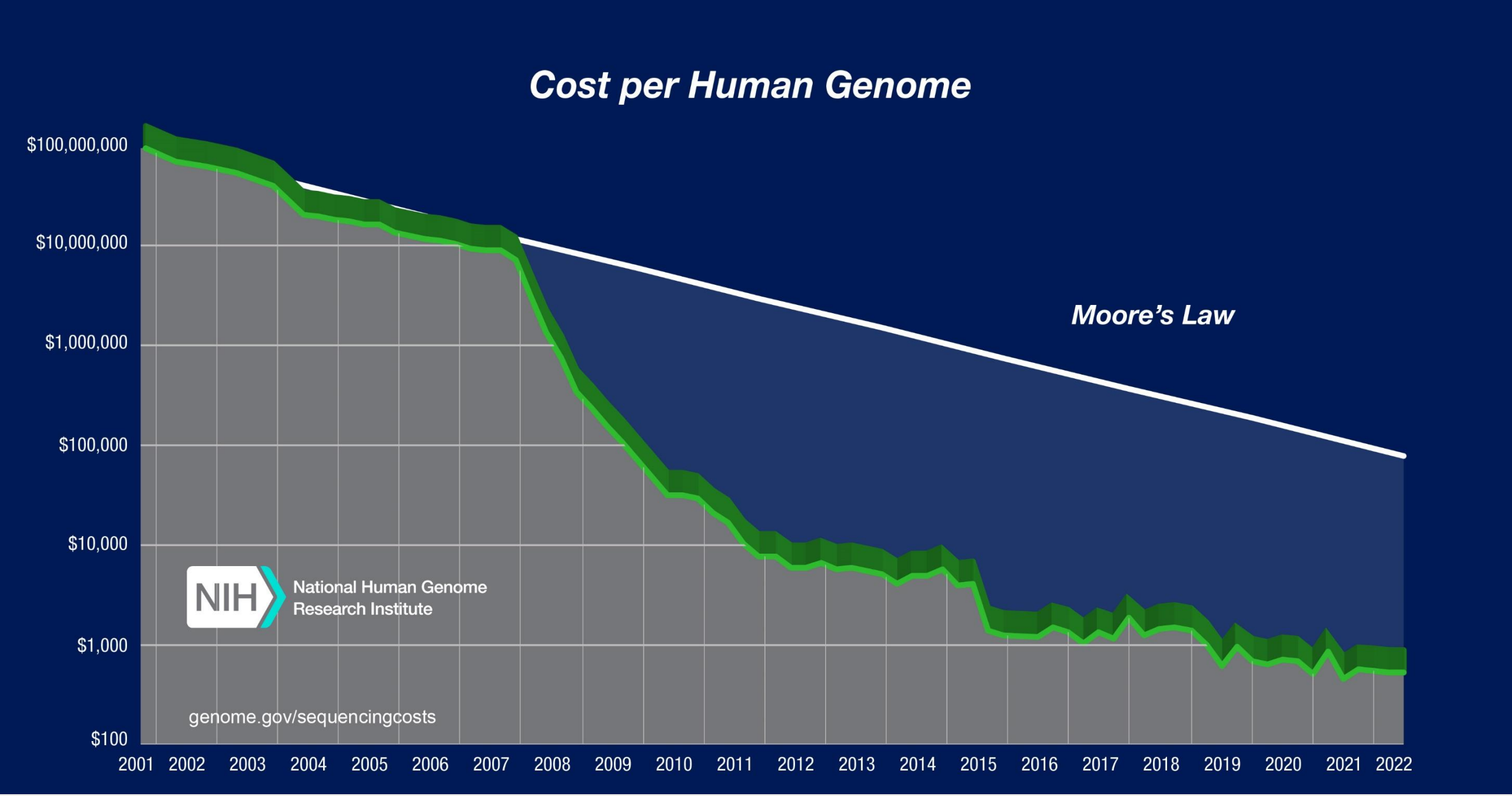


The Fleming Fund
Regional Grants

Sequencing platforms

- There exists a variety of sequencing platforms divided into generations based on their methodology.
- 1st generations were based on the original Sanger sequencing method. Not suitable for WGS.
- 2nd generation improved sequencing by running millions of reactions in parallel, increasing the throughput of data
- 3rd generation utilized new methods to sequence ultra long segments of DNA



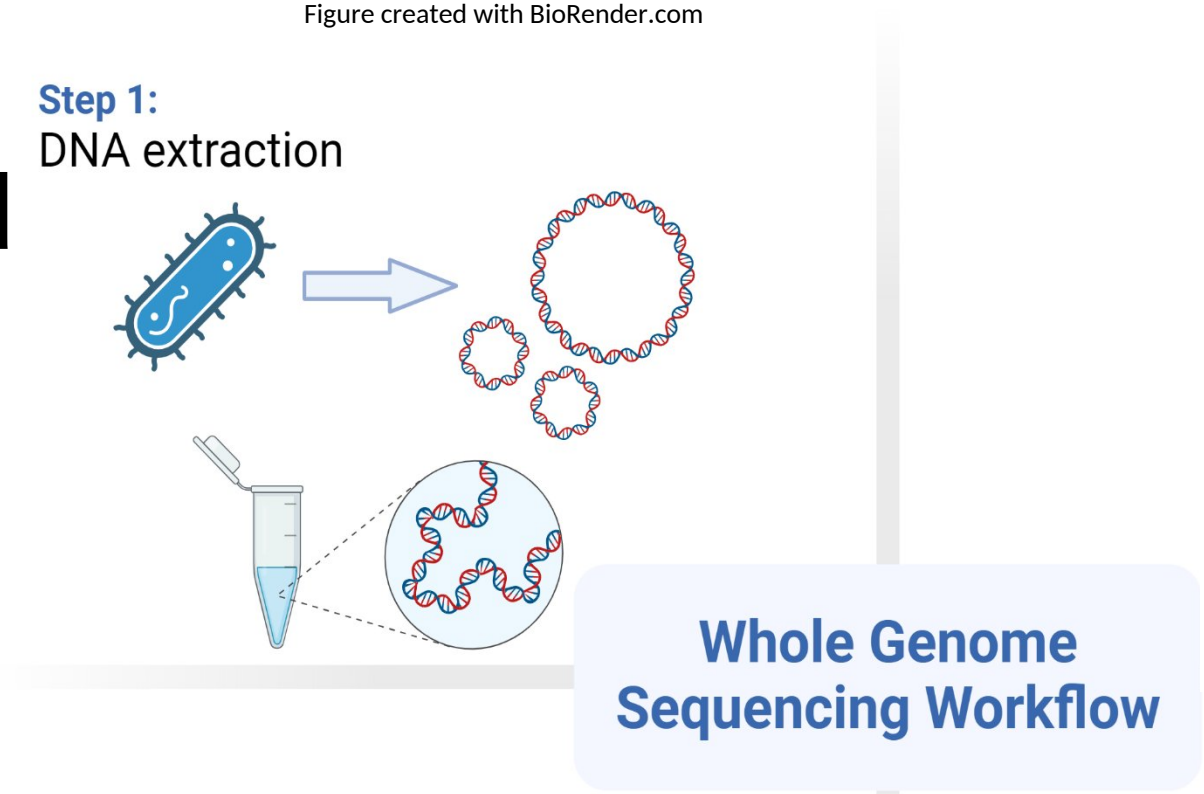


Courtesy: National Human Genome Research Institute, Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: **www.genome.gov/sequencingcostsdata**. Accessed [25th Sep. 2024]

Figure created with BioRender.com

WGS generalized

1) Cells from a pure culture broken open. DNA is extracted, cleaned for proteins and cell debris.

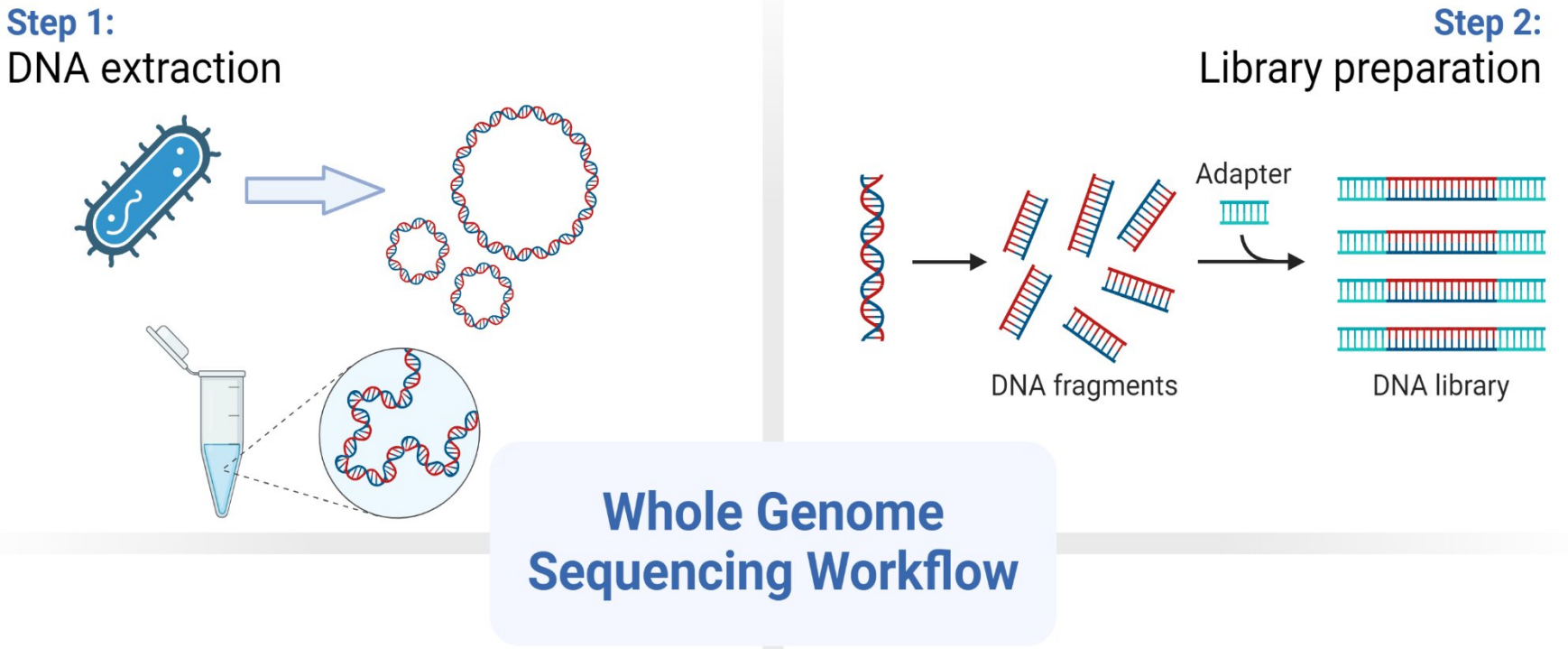


Ona, 2023 (2024) Adapted from “Next Generation Sequencing Workflow”. Retrieved from https://app.biorender.com/biorender-templates Created in BioRender.com

Figure created with BioRender.com

WGS generalized

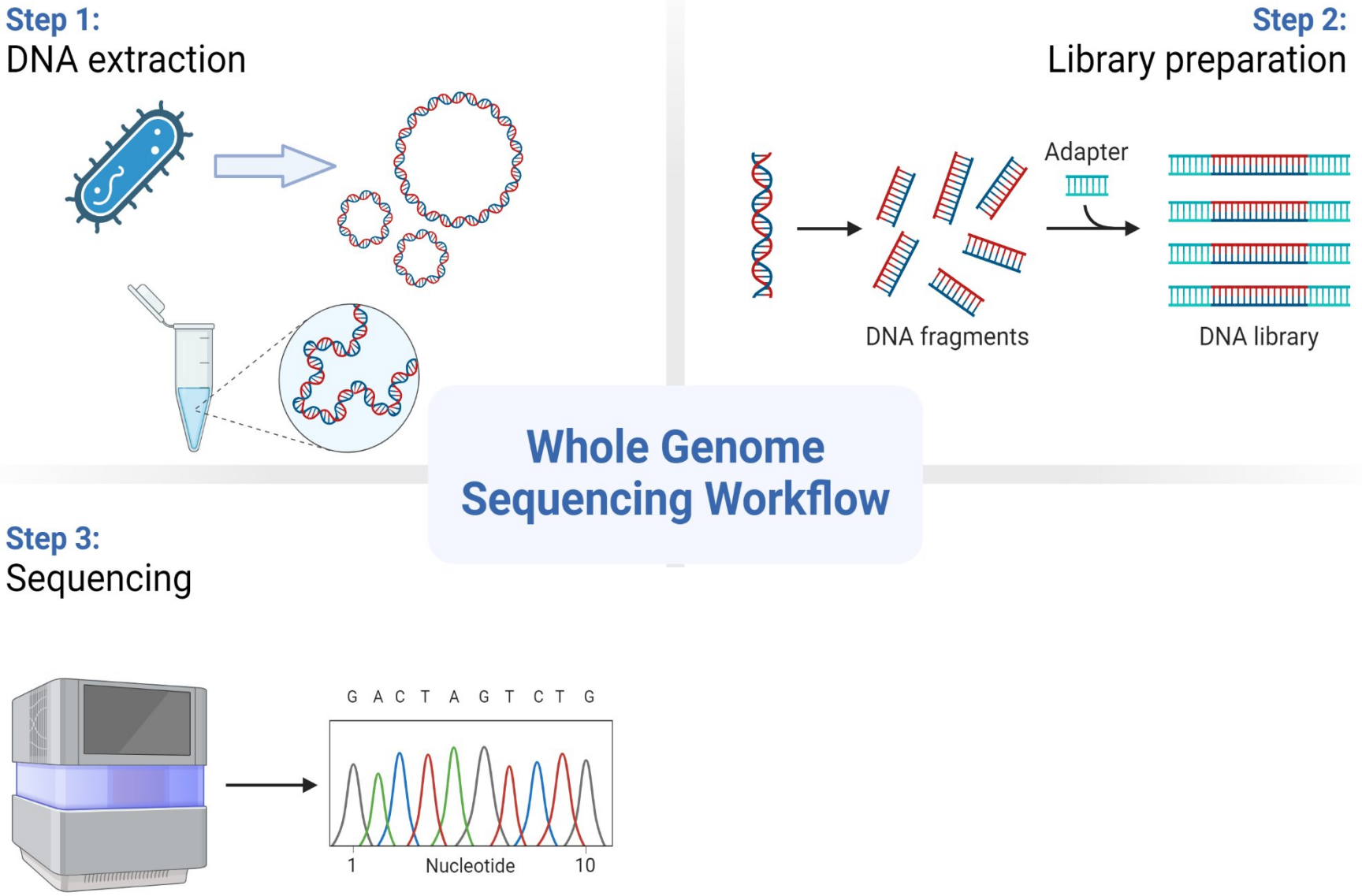
1) Cells from a pure culture broken open. DNA is extracted, cleaned for proteins and cell debris.
2) DNA is fragmented to smaller pieces and adapters are attached.



Ona, 2023 (2024) Adapted from “Next Generation Sequencing Workflow”. Retrieved from https://app.biorender.com/biorender-templates Created in BioRender.com

WGS generalized

1) Cells from a pure culture broken open. DNA is extracted, cleaned for proteins and cell debris.
2) DNA is fragmented to smaller pieces and adapters are attached.
3) DNA library is loaded to sequencing platform and the sequence of nucleotides in each fragment determined.

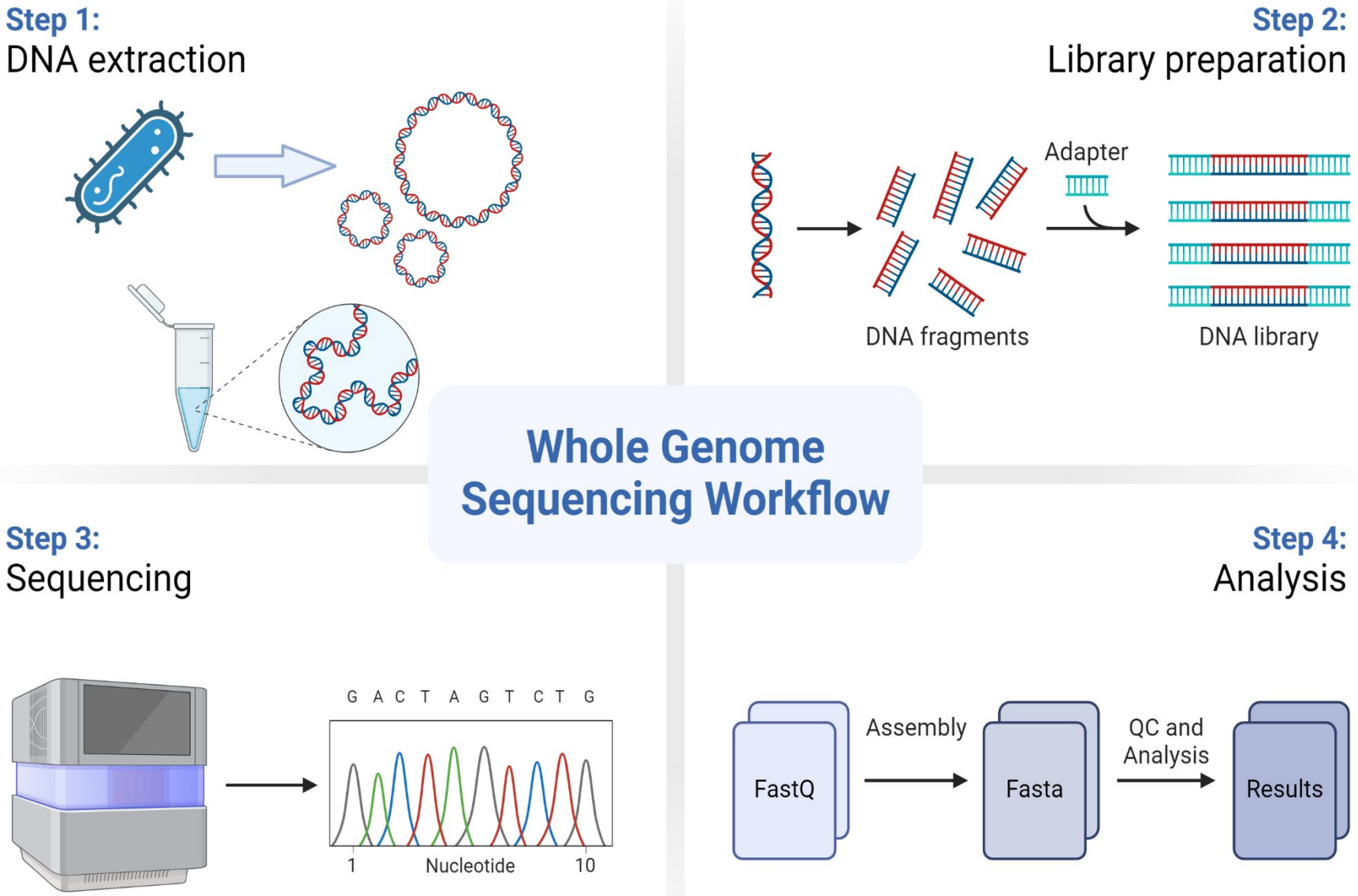


Ona, 2023 (2024) Adapted from "Next Generation Sequencing Workflow". Retrieved from https://app.biorender.com/biorender-templates Created in BioRender.com

Figure created with BioRender.com

WGS generalized

1) Cells from a pure culture broken open. DNA is extracted, cleaned for proteins and cell debris.
2) DNA is fragmented to smaller pieces and adapters are attached.
3) DNA library is loaded to sequencing platform and the sequence of nucleotides in each fragment determined.
4) The machine outputs each fragment as a “read”. Post-processing and quality control (QC) are conducted before analysis.



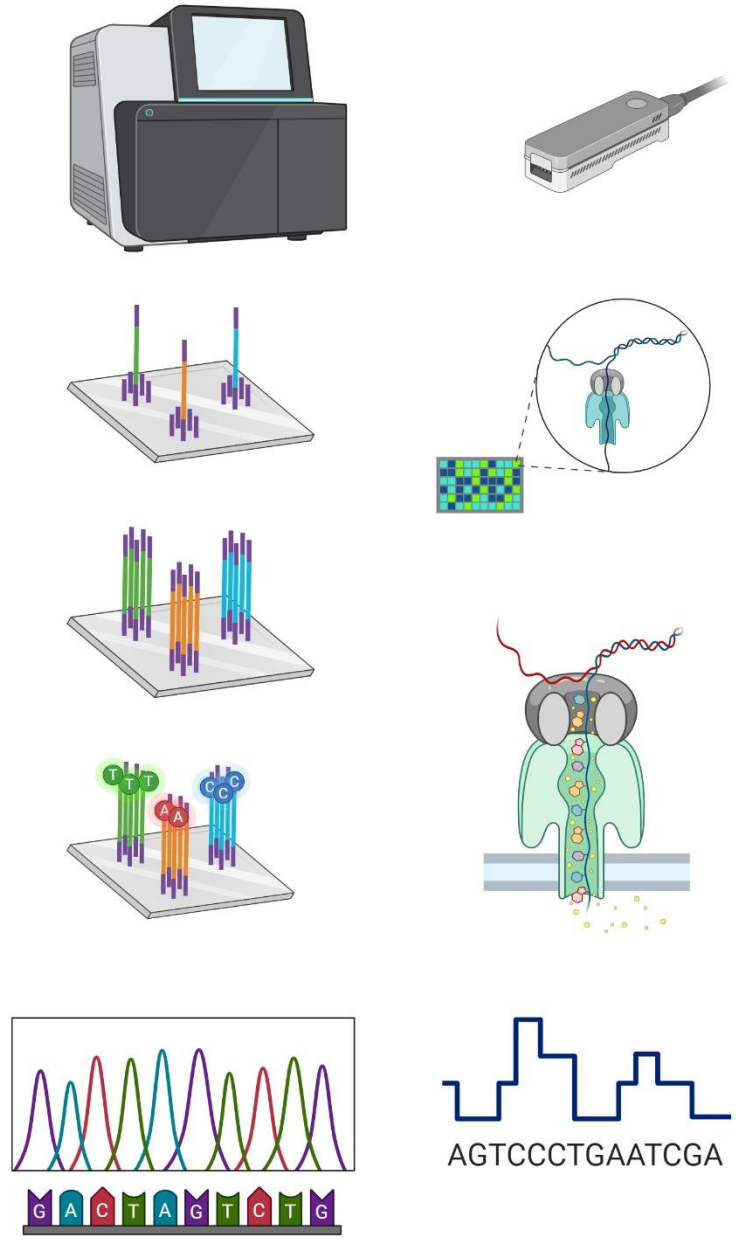
Ona, 2023 (2024) Adapted from “Next Generation Sequencing Workflow”. Retrieved from https://app.biorender.com/biorender-templates Created in BioRender.com

Illumina vs. Nanopore WGS Sequencing

- In Illumina sequencing:
  - Each fragment inputted into the machine is sequenced-by-synthesis.
  - The machine records every added nucleotide to the growing DNA polymer using modified fluorescent nucleotides.
  - Each nucleotide has a distinct color, meaning bases can be called by the color of light emitted.

- Base calling in Illumina machines are determined by the intensity and color off light emitted.

Figure created with BioRender.com



- In Nanopore sequencing
  - In the flow cell, a membrane separate one side of the flow cell from the other, each side connected by protein pores.
  - Osmotic pressure forces DNA molecules to enter pores.
  - Adapters bind to the pore and one string of DNA moves through the pore.
  - A weak electric current runs through the membrane and the bases of the nucleotides disrupts the current.

- Bases produce a characteristic disruption pattern in the electric flow, which is interpreted for base calling.

Illumina vs. Nanopore WGS Sequencing

Illumina platforms are the most widely used platform for genomic surveillance.

Nanopore produces much longer reads, simplifying post-processing and analysis.

Illumina machines are more expensive, but sequencing is comparably cheaper than on Nanopore platforms.

Nanopore platforms are historically less accurate at correctly calling basepairs, but have improved considerably the last few

Metric	Illumina Nextseq	Illumina Novaseq	Nanopore MinION	Nanopore PromethION
Read size	Short (50-300 bp)		Long (1000-10.000s bp)	
Machine price $	~250.000	~1.000.000	1.000-2.000	~450.000
Reagent price $ (per Gbp)	Medium-high	Low	High	Medium-low
Accuracy %	99.9		95-98.9 (depending on chemistry)	

Prices may not be accurate

The Fleming Fund
Regional Grants

Let’s take a break ☺️

Image generated with https://chatgpt.com

Basic Bioinformatics: file formats and sequence data types

Basecalling, what is it?

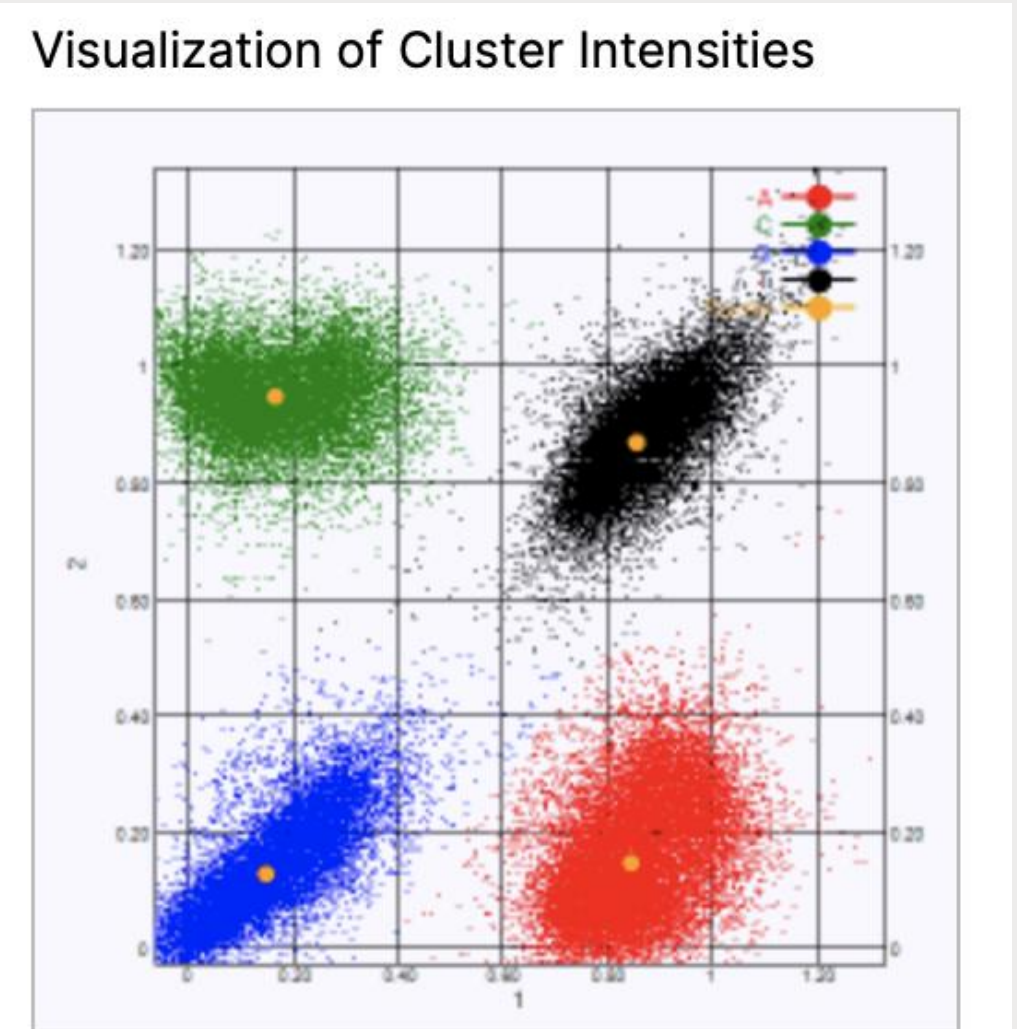
- Base calling is the process of assigning nucleobases to chromatogram peaks, light intensity signals, or electrical current changes resulting from nucleotides passing through a nanopore. From: Wikipedia

- Base calling is the process of algorithmically deciding the incorporated nucleotide from the signal intensities that are detected during sequencing process. From: Encyclopedia of Bioinformatics and Computational Biology, 2019.

The Fleming Fund
Regional Grants

Illumina Basecalling

- iSeq 100 Sequencing System
- Base calling determines a base (A, C, G, or T) for every cluster of a given tile at a specific cycle.
- The iSeq 100 uses one-dye sequencing, which requires one dye and two images to encode data for the four bases.
- Intensities extracted from one image and compared to a second image result in four distinct populations, each corresponding to a nucleotide.
- Base calling determines which population each cluster belongs to.



Base Calls in One-Dye Sequencing

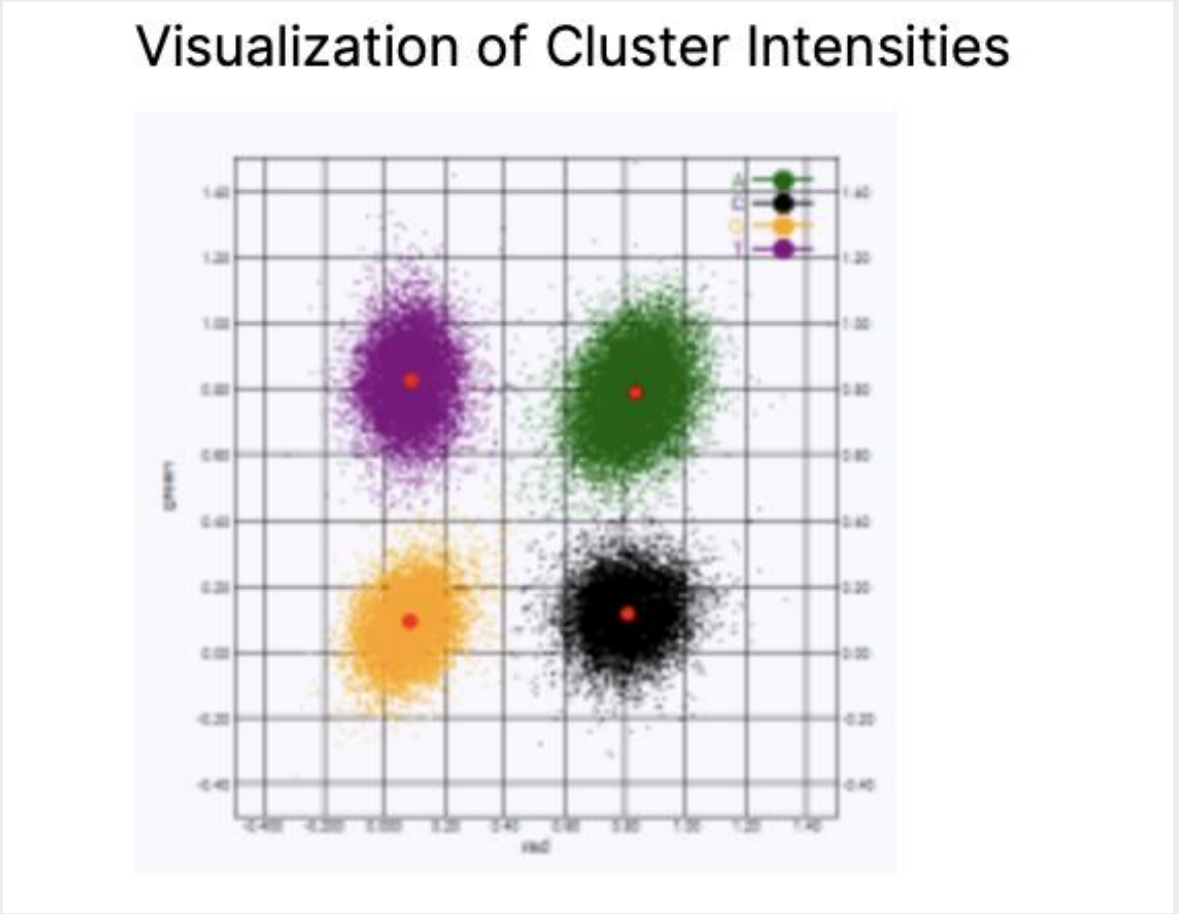
Base	Dye in First Image	Dye in Second Image	Conclusion From Compared Images
T	On	On	Clusters that show intensity in both images are T bases.
A	On	Off	Clusters that show intensity in the first image only are A bases.
C	Off	On	Clusters that show intensity in the second image only are C bases.
G	Off	Off	Clusters that show intensity in neither image are G bases.

Images accessed from https://support-docs.illumina.com/IN/iSeq100/Content/IN/iSeq/BaseCalling_fISQ.htm.

The Fleming Fund
Regional Grants

Illumina Basecalling

- NextSeq 500 and 550 Sequencing Systems
- Base calling determines a base for every cluster of a given tile at a specific cycle.
- The NextSeq 550 uses two-channel sequencing, which requires only two images to encode the data for four DNA bases, one from the red channel and one from the green channel.
- Intensities extracted from an image compared to another image result in four distinct populations, each corresponding to a nucleotide.
- The base calling process determines to which population each cluster belongs.

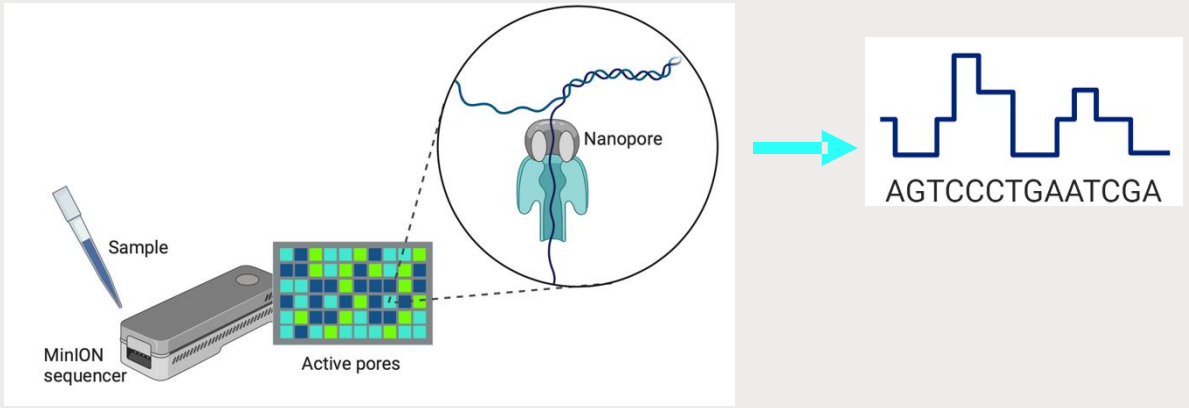


Base Calls in Two-channel Sequencing

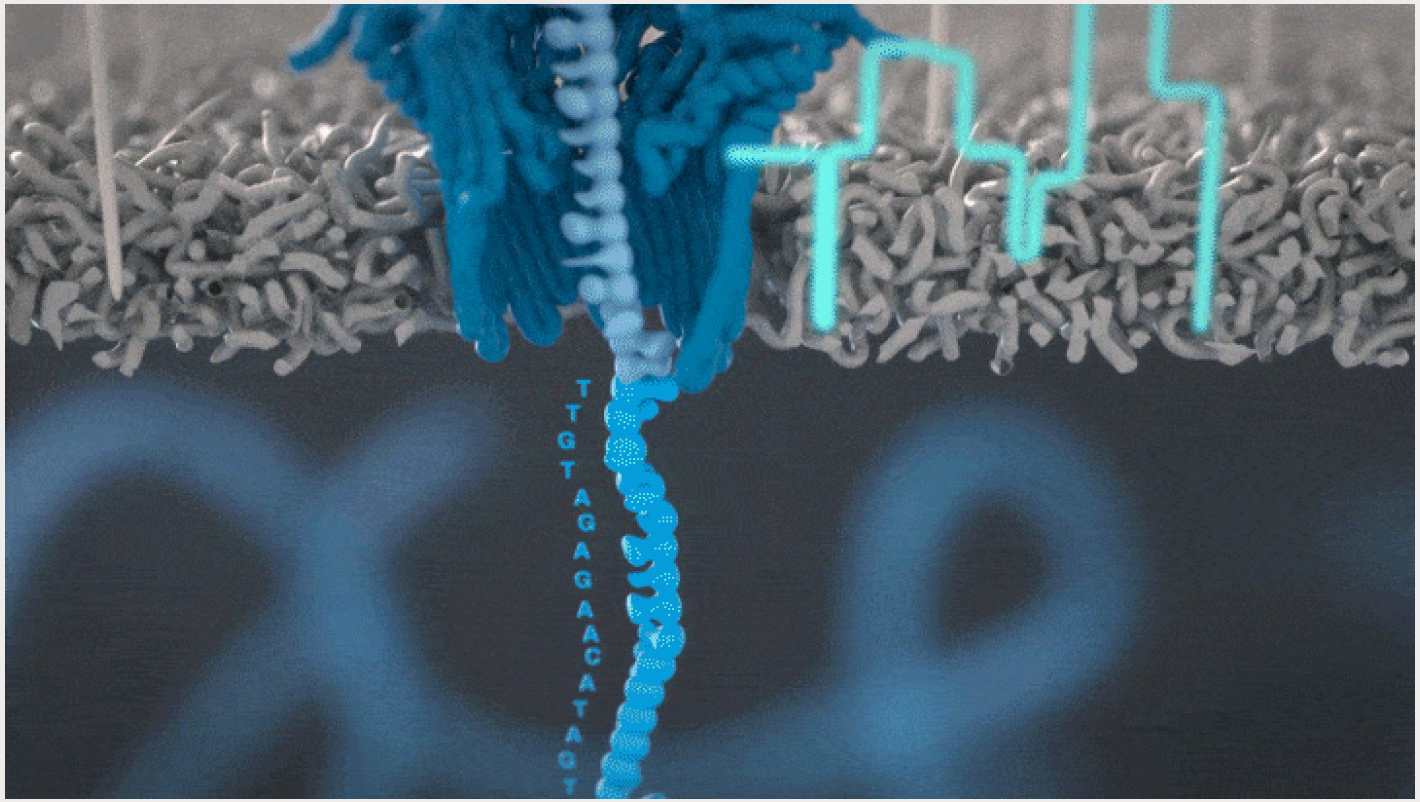
Base	Red Channel	Green Channel	Result
A	1 (on)	1 (on)	Clusters that show intensity in both the red and green channels.
C	1 (on)	0 (off)	Clusters that show intensity in the red channel only.
G	0 (off)	0 (off)	Clusters that show no intensity at a known cluster location.
T	0 (off)	1 (on)	Clusters that show intensity in the green channel only.

Images accessed from https://support-docs.illumina.com/IN/iSeq100/Content/IN/iSeq/BaseCalling_fISQ.htm.

Figure created with BioRender.com



ONT basecalling



- When sequencing DNA or RNA through nanopores, the characteristic electrical signals are recorded by MinKNOW™, the software that controls Oxford Nanopore Technologies sequencing devices.
- This entire characteristic electrical signal is known as a ‘squiggle’. MinKNOW processes the squiggle into reads in real time — each read corresponding to a single strand of sequenced DNA or RNA.

GIF accessed from https://nanoporetech.com/platform/technology/basecalling

ONT Basecalling

- The structure of the nanopore determines the information contained within a squiggle: the raw signal that reflects the molecules that have passed through the nanopore before basecalling.
- Different nanopores contain different ‘readers’.
- The previous R9 nanopore had a single reader in the middle of the barrel
- The new R10 nanopore has two readers spaced along its length, meaning more bases within a DNA or RNA strand can contribute to the squiggle at any one time.
- This leads to improvements in capturing signals around homopolymer regions, where multiples of the same nucleotide appear one after the other on a DNA or RNA strand.

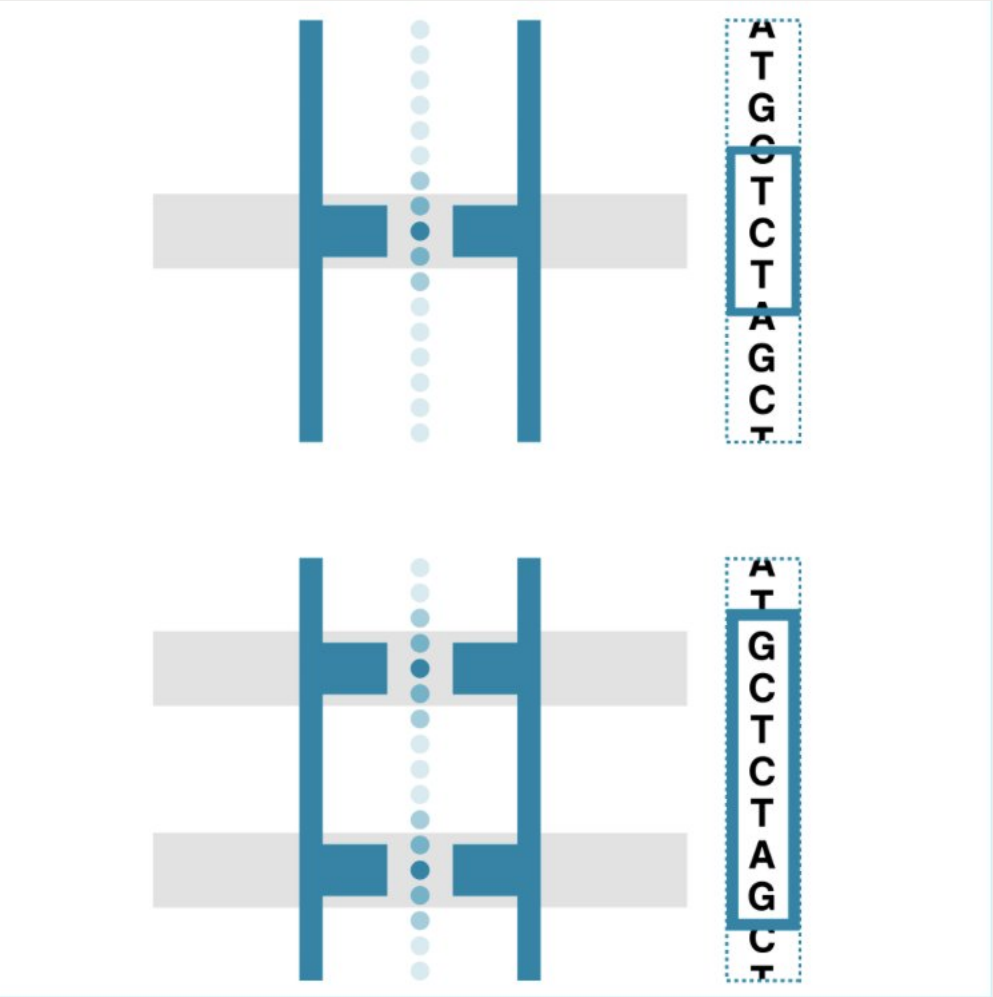


Image accessed from https://nanoporetech.com/platform/technology/basecalling

ONT Basecalling

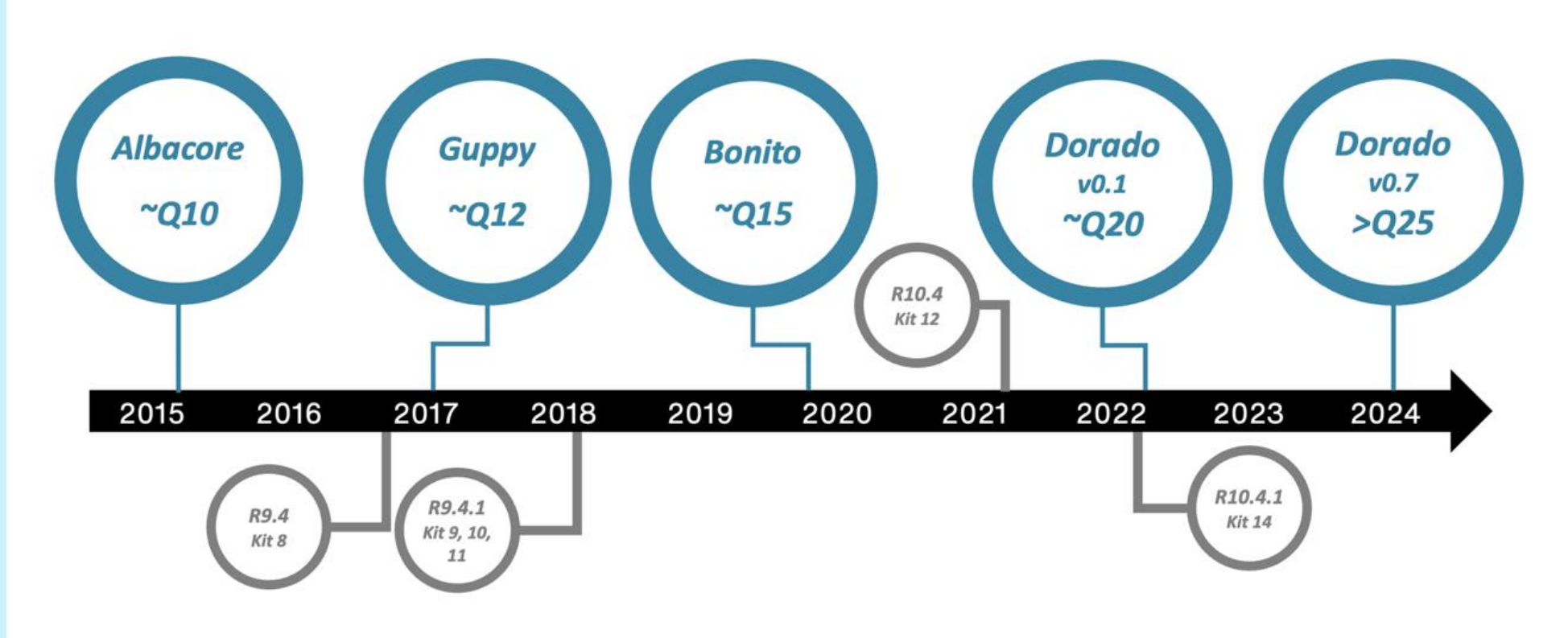


Image accessed from https://nanoporetech.com/platform/technology/basecalling

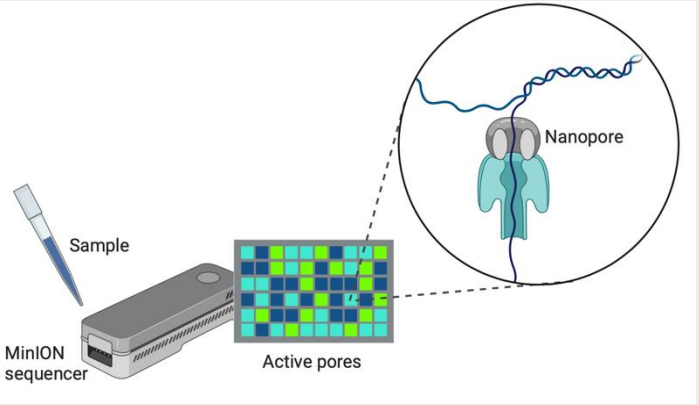
The Fleming Fund
Regional Grants

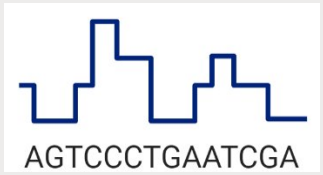
Data processing

Base calling

Fastq file containing millions of reads

```
@SRR1770413.1 1/1
CACCCGGCATCAGGTGCGGTACTTTTGCGCCTCCCAGCCGGACCGGCCCTGCGGCGTAATA
CCAGCCTCACATCCCTCGCTGCCTGCGTATCCAGCTCACTCTCCCTGGTTGCCGCCTACAT
GCTCCCTCCCGCTGTTCCACCCCTTTGCACCCCCCCTCTGCCCCTCCTGCTCGCCAGCCCC
+
CCCCCECFCEFC@8F8C77B7BFEFD,C+@@@BCB##########################
##############################################################
##############################################################
```





Base calling

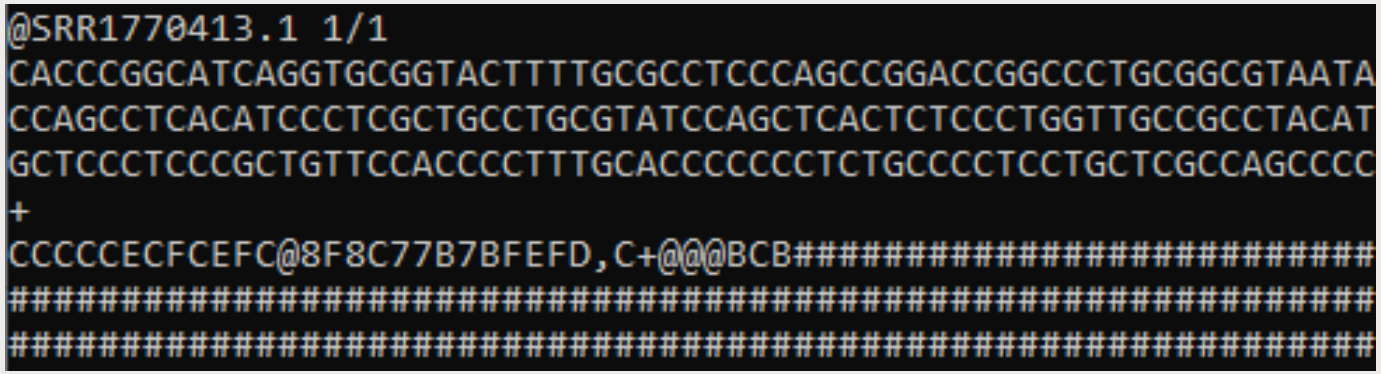
Fastq files containing 4000 reads (default)

```
@SRR1770413.1 1/1
CACCCGGCATCAGGTGCGGTACTTTTGCGCCTCCCAGCCGGACCGGCCCTGCGGCGTAATA
CCAGCCTCACATCCCTCGCTGCCTGCGTATCCAGCTCACTCTCCCTGGTTGCCGCCTACAT
GCTCCCTCCCGCTGTTCCACCCCTTTGCACCCCCCCTCTGCCCCTCCTGCTCGCCAGCCCC
+
CCCCCECFCEFC@8F8C77B7BFEFD,C+@@@BCB##########################
##############################################################
##############################################################
```

Figure created with BioRender.com

What is fastq?

- Fastq is the file format DNA reads are stored in after the sequencing machine does base-calling.
- It has a particular format:
  - ➢ Header
    - Contains info on the run, depends on machine
    - Unique ID
  - ➢ Called bases
    - Sequence
  - ➢ Spacer line
    - Not used or additional info
  - ➢ Base quality scores
    - Phred-score giving the probability that the base call is incorrect.



Phred scores

- The Phred quality score given as one of the 127 standard ASCII characters.
- Traditionally the scale is off-set, with different sequencing machines using different scales.
- Most modern machines use the sanger scale.
- The base quality score is important in correctly calling Single Nucleotide Polymorphisms (SNP), used in phylogeny and outbreak detection

Table 1 ASCII Characters Encoding Q-scores 0–40

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	/	47	14	=	61	28
"	34	1	0	48	15	>	62	29
#	35	2	1	49	16	?	63	30
$	36	3	2	50	17	@	64	31
%	37	4	3	51	18	A	65	32
&	38	5	4	52	19	B	66	33
'	39	6	5	53	20	C	67	34
(	40	7	6	54	21	D	68	35
)	41	8	7	55	22	E	69	36
*	42	9	8	56	23	F	70	37
+	43	10	9	57	24	G	71	38
,	44	11	:	58	25	H	72	39
-	45	12	;	59	26	I	73	40
.	46	13	<	60	27			

Courtesy: Omion, Bioinformatics for Beginners – File Formats Part 2. – Short reads. Available at: https://www.omixon.com/bioinformatics-for-beginners-file-formats-part-2-short-reads/ . Accessed [30th Sep. 2024]

The probability of error

- The Phred quality score is a logarithmic score based on the probability that the base call (nucleotide) is incorrect

- Q10 = 1/10 risk of incorrect base
- Q20 = 1/100 risk of incorrect base
- Q30 = 1/1000 risk of incorrect base

- This means that in a sequence of 100 bp at Q20, there will most likely be at least 1 error.

$$Q = -10 \bullet \log_{10}(P)$$

or in terms of probability

$$P = 10^{-\frac{Q}{10}}$$

Where

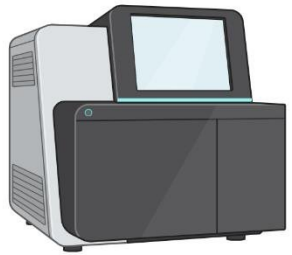
P = probability of incorrect base call

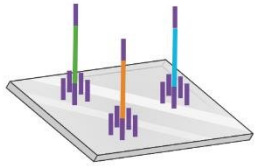
Q = Phred quality score

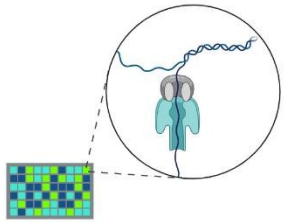
Phred quality score	Probability of incorrect base call	Probability of being correct
10	0.1	90%
20	0.01	99%
30	0.001	99.9%

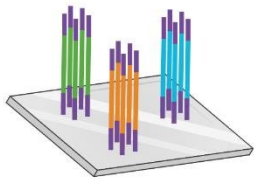
Why do errors occur?

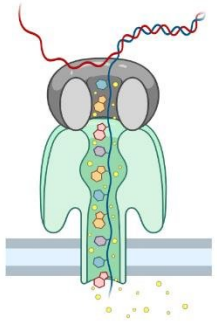
- In Illumina sequencing:
  - As multiple rounds of sequencing are conducted, the probability of erroneous base calls increases
  - Every time a new base is called an error may occur, meaning the signal for the correct base gets weaker
  - Degradation of enzymes used in the reaction may introduce more errors
  - Stretches of repetitive nucleotide calls are difficult to call.

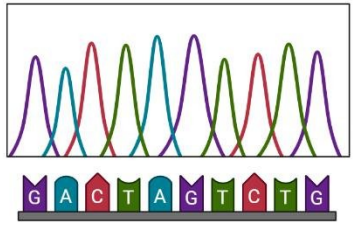












- In nanopore sequencing:
  - The pattern obtained from the nanopore needs to be interpreted.
  - Interpretation is based on machine learning models.
  - Signal varies depending on neighboring nucleotides in the polymer and condition of pore.
  - DNA string may slip in the pore.
  - Stretches of repetitive nucleotide calls are difficult to call.
  - Newer chemistry improves significantly on accuracy.

Created in BioRender.com

The Fleming Fund
Regional Grants

Let’s take a break ☺️

The Fleming Fund
Regional Grants

Thank you

This programme is being funded by the UK Department of Health and Social Care.
The views expressed do not necessarily reflect the UK Government's official policies.