EURL Lm
European Union Reference Laboratory for
Listeria monocytogenes
http://eurl-listeria.anses.fr

INTRODUCTION TO WGS SEQUENCING PLATFORMS & OUTPUT DATA

Joint Training Course of the inter EURLs Working Group on NGS:
Introduction to Bioinformatics for genomic data mining
20-21 June 2023, Bilthoven

Maroua SAYEB & Adrien ASSERE
eurl-listeria@anses.fr

Sequencing technologies

EURL Lm

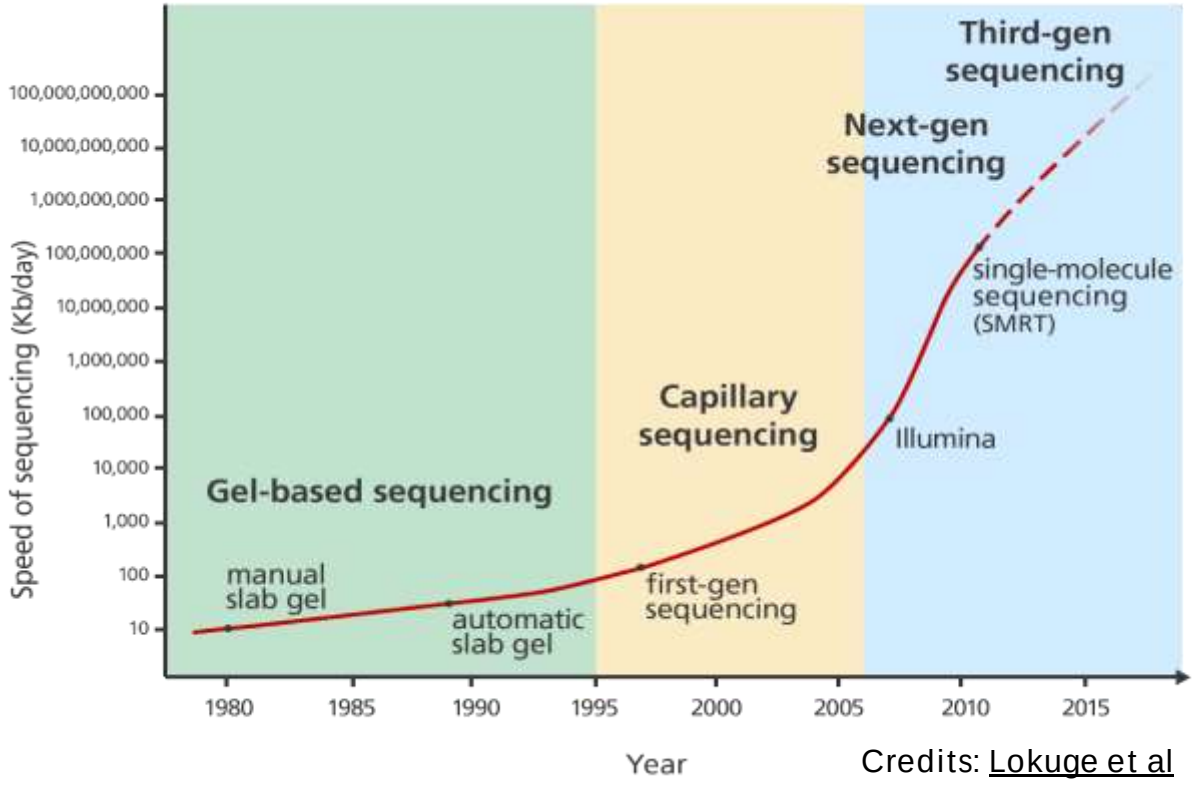
- First generation: Sanger Sequencing
  - Most widely used sequencing technology for approximately 25 years

- Next Generation Sequencing

"AKA" high-throughput sequencing Includes most sequencing technologies that came after Sanger sequencing

  - Second Generation: Short-read
    - Illumina
    - Ion Torrent
    - 454 pyrosequencing (*Legacy Technology*)

  - Third Generation: Long-read
    - Pac-Bio
    - Oxford Nanopore



Credits: Lokuge et al

ABI-Sanger
1996

Illumina-Miseq
2015

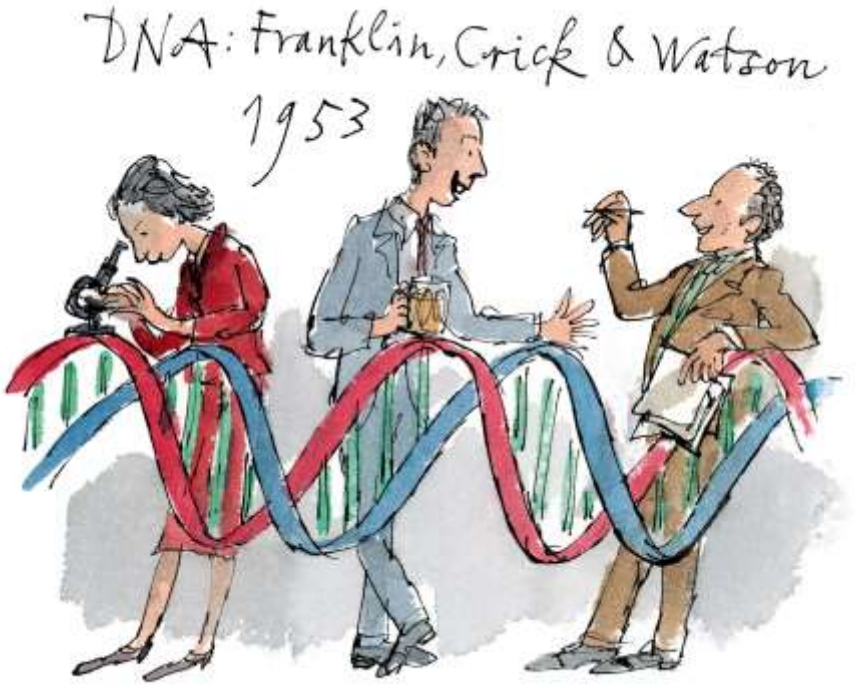
NGS vs 1st gen (Sanger)

Sanger	NGS
• targeted sequencing, • highly accurate sequence data, • confirmation of variants in pathogenic bacteria.	• massive parallel sequencing of target genes, • comprehensive genomic analysis, • outbreak investigations, • antimicrobial resistance profiling, • transcriptomics, • metagenomics studies.

✓ The choice between Sanger sequencing and NGS depends on the specific objectives, the scale of the study, the desired resolution, and the depth of genomic information required.

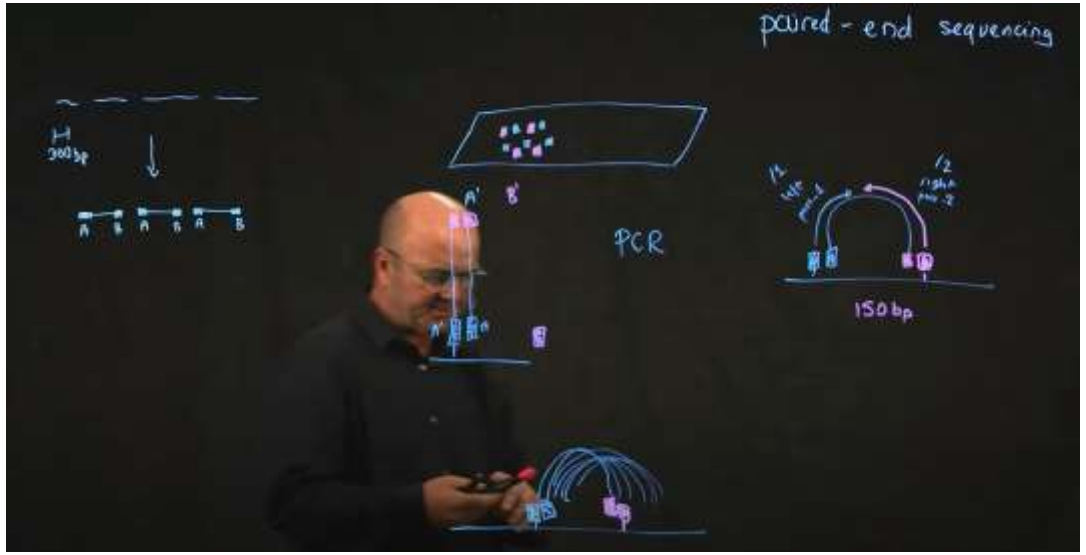
Different kinds of reading the sequences

- Illumina
- Ion Torrent
- Pacific Biosciences (PacBio)
- Oxford Nanopore



Credits: DNA double helix breakthrough (cam.ac.uk)

Illumina Sequencing:



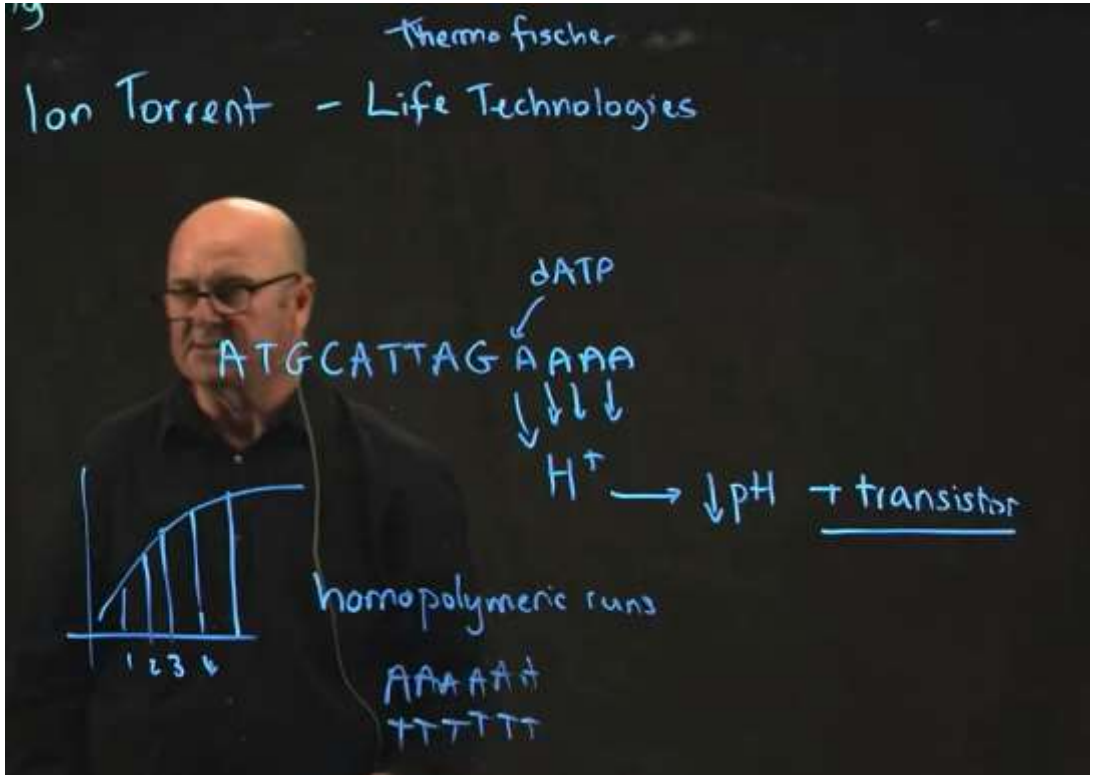
RobEdwards

illumina paired end sequencing - YouTube

- widely used NGS platform known for its high throughput and accuracy.
- based on sequencing-by-synthesis technology and reversible dye-terminators that enable the identification of single bases as they are introduced into DNA strands.
- Output data from Illumina sequencing typically consists of short reads in the form of FASTQ files. The read lengths can range from a few dozen to a few hundred bases, depending on the specific sequencing chemistry and instrument used.
- Illumina platforms can generate millions to billions of reads per sequencing run, resulting in high coverage and enabling various genomic applications.

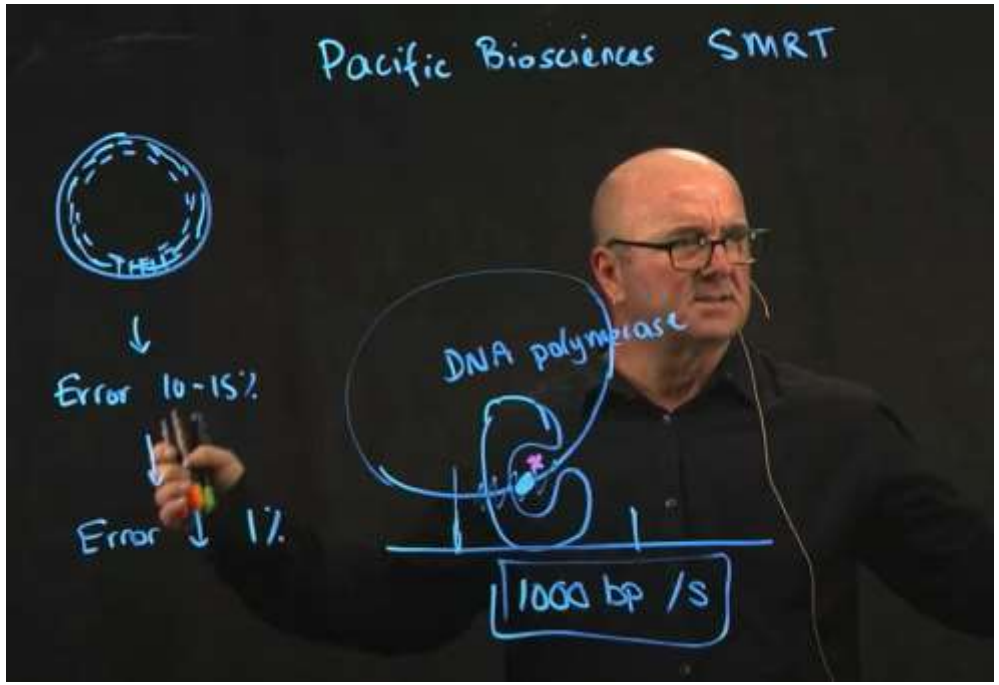
Ion Torrent Sequencing:

- developed by Thermo Fisher Scientific,
- is based on semiconductor technology.
- It detects the release of hydrogen ions during nucleotide incorporation.
- Output data from Ion Torrent sequencing includes short reads in FASTQ format, similar to Illumina sequencing. Read lengths typically range from around 100 to 400 bases.
- Ion Torrent platforms offer relatively fast turnaround times and are suited for applications such as targeted sequencing and small-scale projects.



RobEdwards
Ion Torrent Sequencing - YouTube

Pacific Biosciences (PacBio) Sequencing:



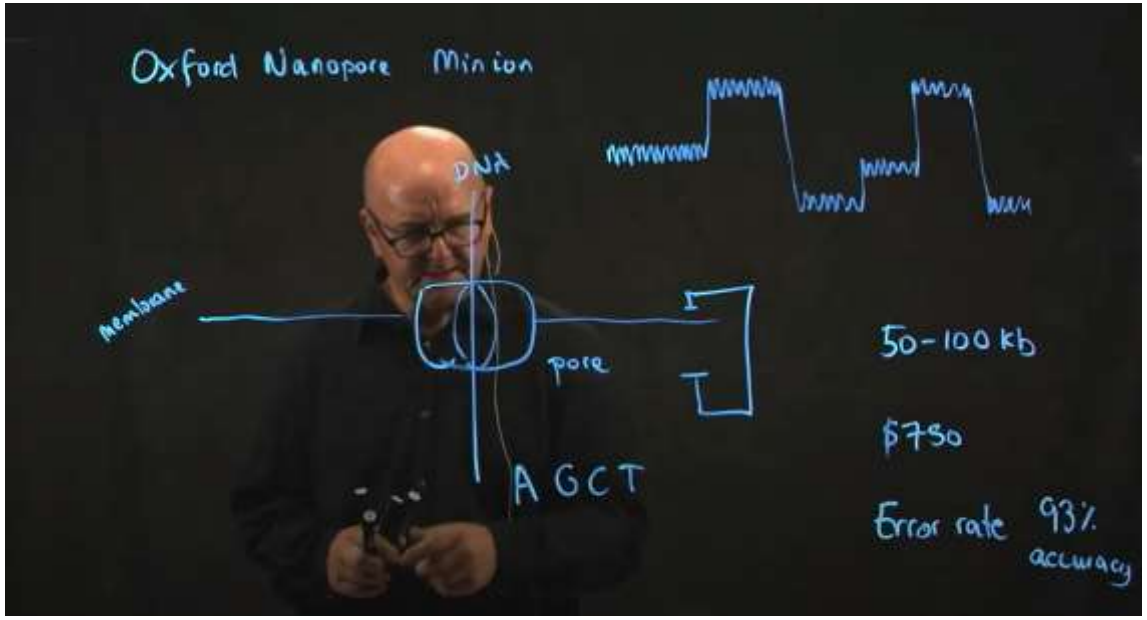
RobEdwards

Pacific Biosciences Sequencing - YouTube

- known as Single-Molecule Real-Time (SMRT) sequencing,
- utilizes circular consensus sequencing. It involves the real-time monitoring of DNA polymerase activity. PacBio platforms generate long reads, spanning thousands to tens of thousands of bases. The output data from PacBio sequencing consists of long reads in the form of FASTQ or BAM files.
- PacBio sequencing is advantageous for applications requiring accurate characterization of complex genomic regions, such as de novo genome assembly and structural variant detection.
- Cost 0,5 to 1M€
- size

Oxford Nanopore Sequencing

- Oxford Nanopore Technologies (ONT) offers nanopore-based sequencing technology.
- It involves passing DNA strands through nanopores and detecting changes in electrical current as the DNA bases pass through the nanopore.
- The output data from ONT sequencing includes long reads in FASTQ or FAST5 formats. Read lengths can vary from a few thousand bases to tens of thousands of bases, with potential for even longer reads.
- ONT platforms provide real-time sequencing and are suitable for various applications, including rapid pathogen detection, metagenomics, and mobile genetic element analysis.



RobEdwards

Oxford Nanopore ONT Sequencing - YouTube

Sequencing platforms

Platform	Read length (bp)	Isolates per run (max)	Run time	Instrument cost (k€)	Cost (€)/ Mb
Illumina HiSeq 2500	150	600-1000	5-11 d	740k€	0,05
Illumina MiSeq	150, 250, 300	12-16	26 h, 36 h 65 h	99k€	1,37
Illumina NextSeq	75, 150	96	29 h	250k€	0.03-0.07
IonTorrent PGM (314, 316, 318)	200, 400	1-10	2-8h	75k€	0,9 -7,5
Ion Proton	100-200	96	2-4 h	245k€	0,02
PacBio RSII	10 000-40 000	8 /smrt cell	0,5-2 h	750k€	180
Sanger	650	96	1 h	100k€	2800
Oxford Nanopore	>20k		1-sev d	1k€	<750

Example: Sequencing Platform Comparison Tool for Illumina

	iSeq 100*	MiniSeq*	MiSeq*†	NextSeq 1000 & 2000*
Output Range	144 Mb - 1.2 Gb	1.65–7.5 Gb	0.3–15 Gb	30-360 Gb***
Run Time	9–19 hours	4-24 hours	5–55 hr	11-48 hours
Reads Per Run	4 million	7–25 million	1–25 million	100 million–1.2 billion***
Max Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp
Samples Per Run§	1-8	50	1–384	30-90
Relative Price Per Sample§	Higher Cost	Mid Cost	Mid Cost	Low Cost
Instrument Price	Lowest Cost	Low Cost	Low Cost	Mid Cost
Downloads	Spec Sheet	Spec Sheet	Spec Sheet	Spec Sheet
System Overview	iSeq 100 Overview	MiniSeq Overview	MiSeq Overview	NextSeq 1000 & 2000 Overview

Share Results | Contact Me

§ Based on 30X or greater coverage of a 5 Mb microbial genome. "Price per Sample" relative to the other instruments shown in comparison results.

* For Research Use Only. Not for use in diagnostic procedures.

† In vitro diagnostic (IVD) instrument available. IVD instrument can perform small whole genome sequencing in Research Mode only.

*** Specifications based on Illumina PhiX control library at supported cluster densities.

Source: Illumina

Platforms and Pathogens

- Read length
- Size of genome being sequenced

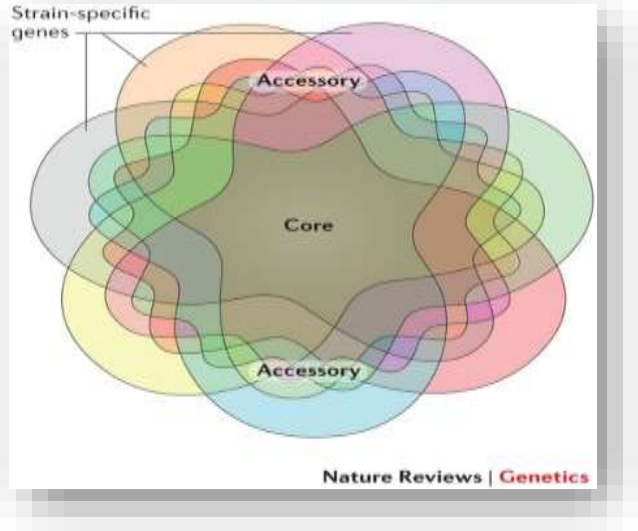
Bacterial Genome

Larger

- ~5MB (2–10MB)
- Pan-genome
  - •Core genome: ~3,000–5,000 genes, present in most strains of a given specie
  - •Accessory genome: up to thousands of genes, not always present

Structure

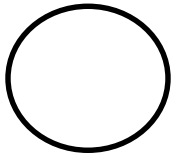
- dsDNA
- Usually single, circular chromosome



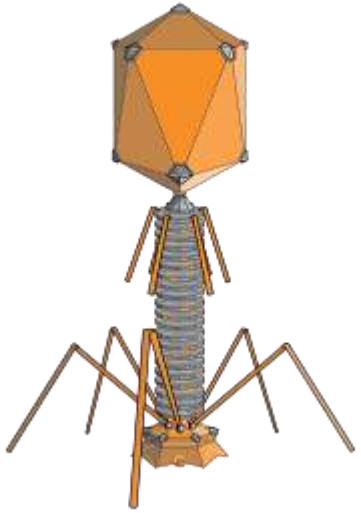
Bacterial Genome Complications

Complications: more complex, repetitive elements. Size: 1,000,000MB

- Plasmids
  - •Circular, dsDNA structures that replicate independently from chromosome
  - •Often carry resistance or virulence genes
  - •Can be passed from one bacterium to another



- Phages
  - •Viruses that infect bacteria
  - •Genome can integrate into chromosome



Credits: Wikipedia

Viral Genome

Compact

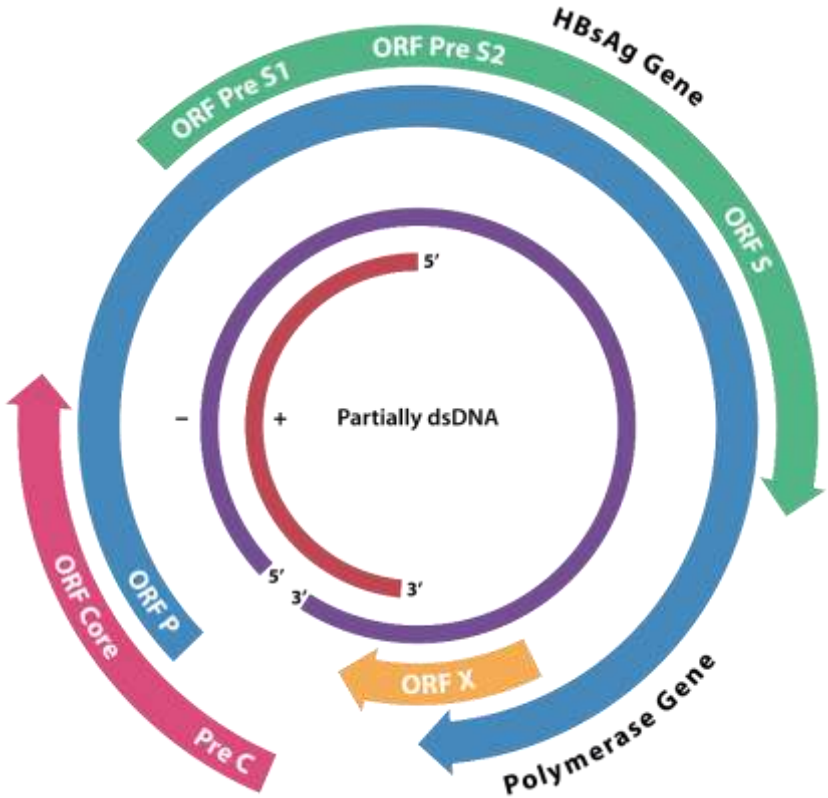
- ~10,000 nucleotides (nt) (typically, ~3 000–200 000)
- Little wasted space

Variable composition

- DNA; RNA
- Single-stranded; double stranded
- Linear; circular
- Single; segmented

Often highly variable

- Particularly true of ssRNA viruses
- Quasispecies. Example: hepatitis C virus



Example: HBV Genome

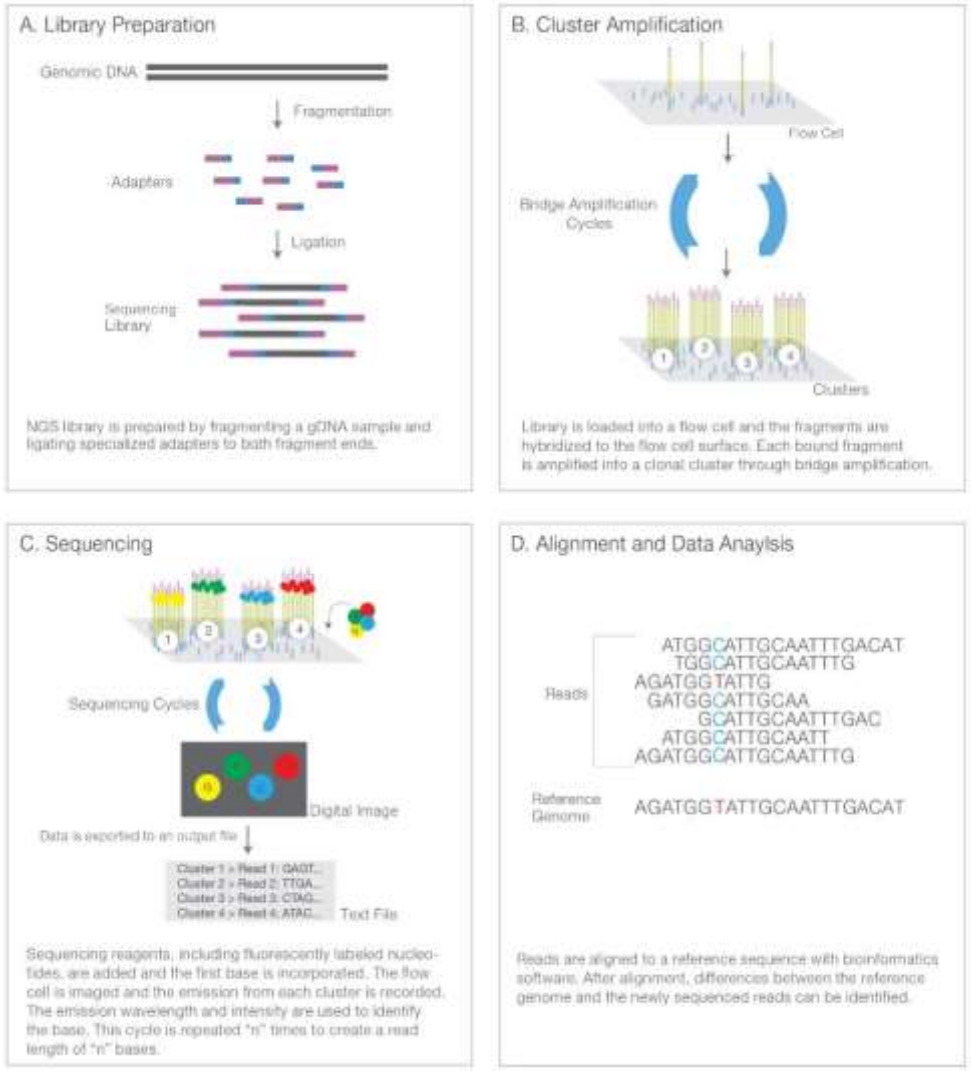
Sequencing Process Definitions

EURL Lm

- Sequence : generic name describing order of biological letters (DNA/RNA).
- Both reads and contigs are DNA/RNA sequences:

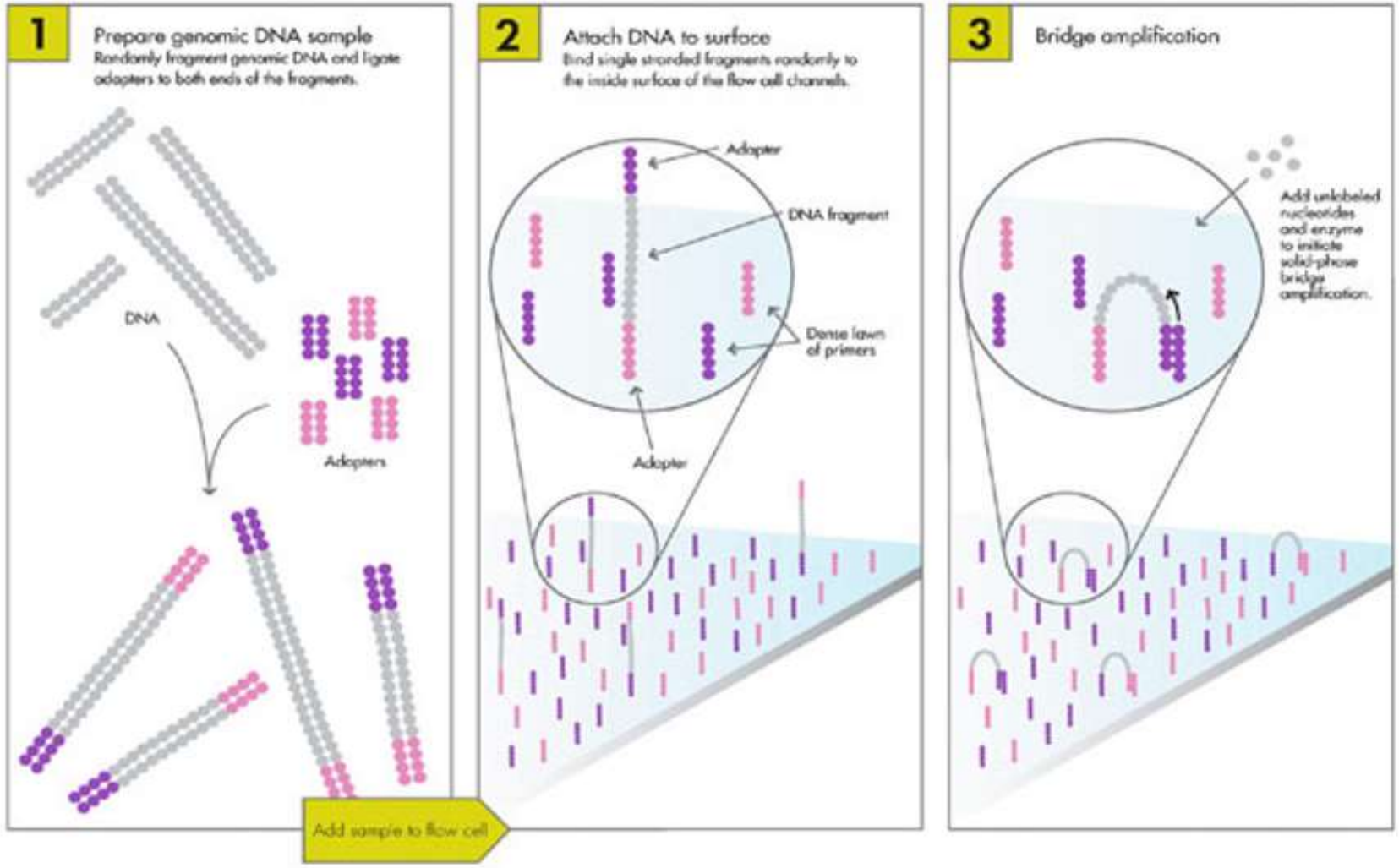
- Reads: sequenced reads of base pairs as you are trying to assemble

- Contigs: reads that have been assembled together; final product



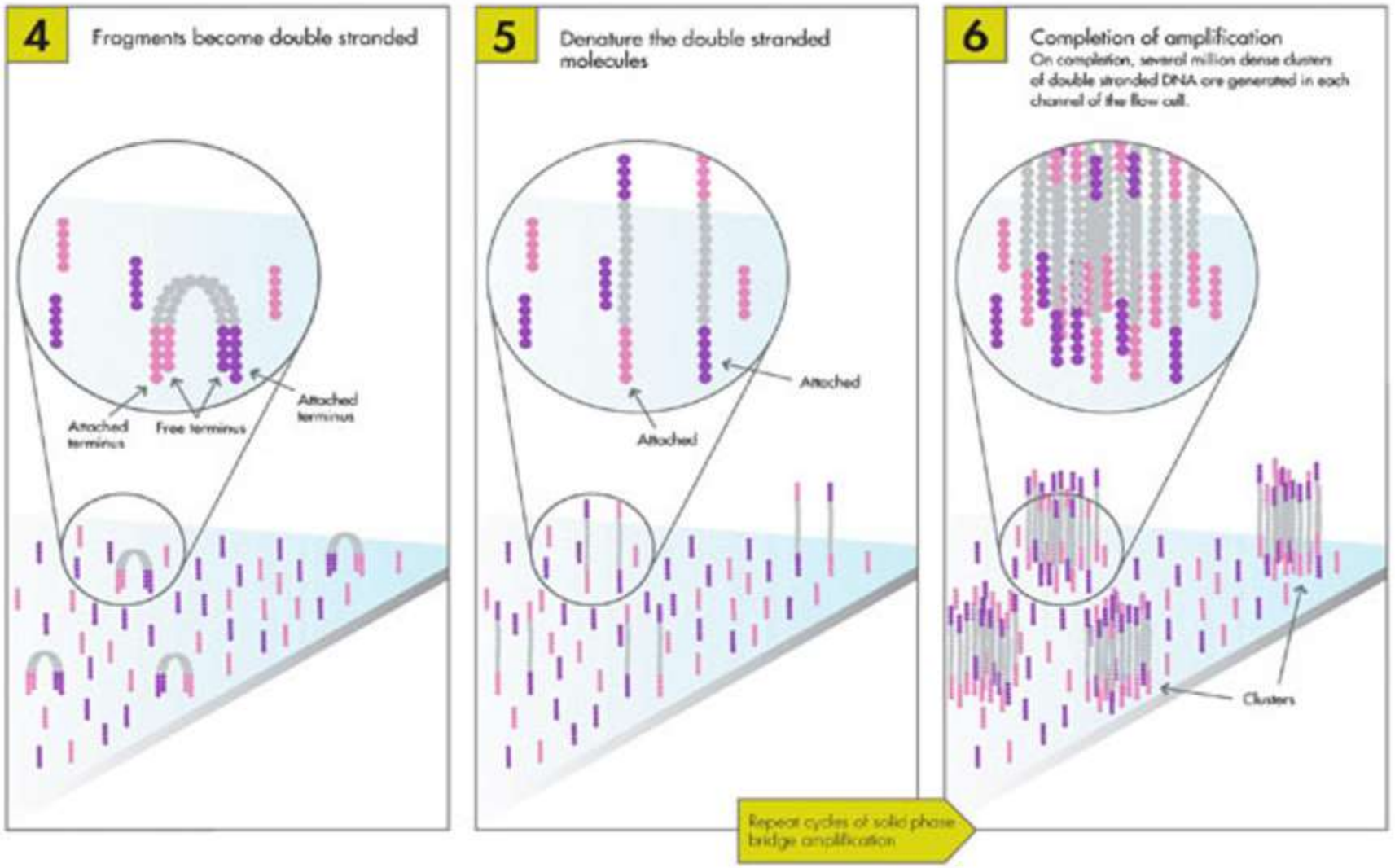
Credits: Intro to NGS (illumina.com)

Example: Illumina Sequencing in details



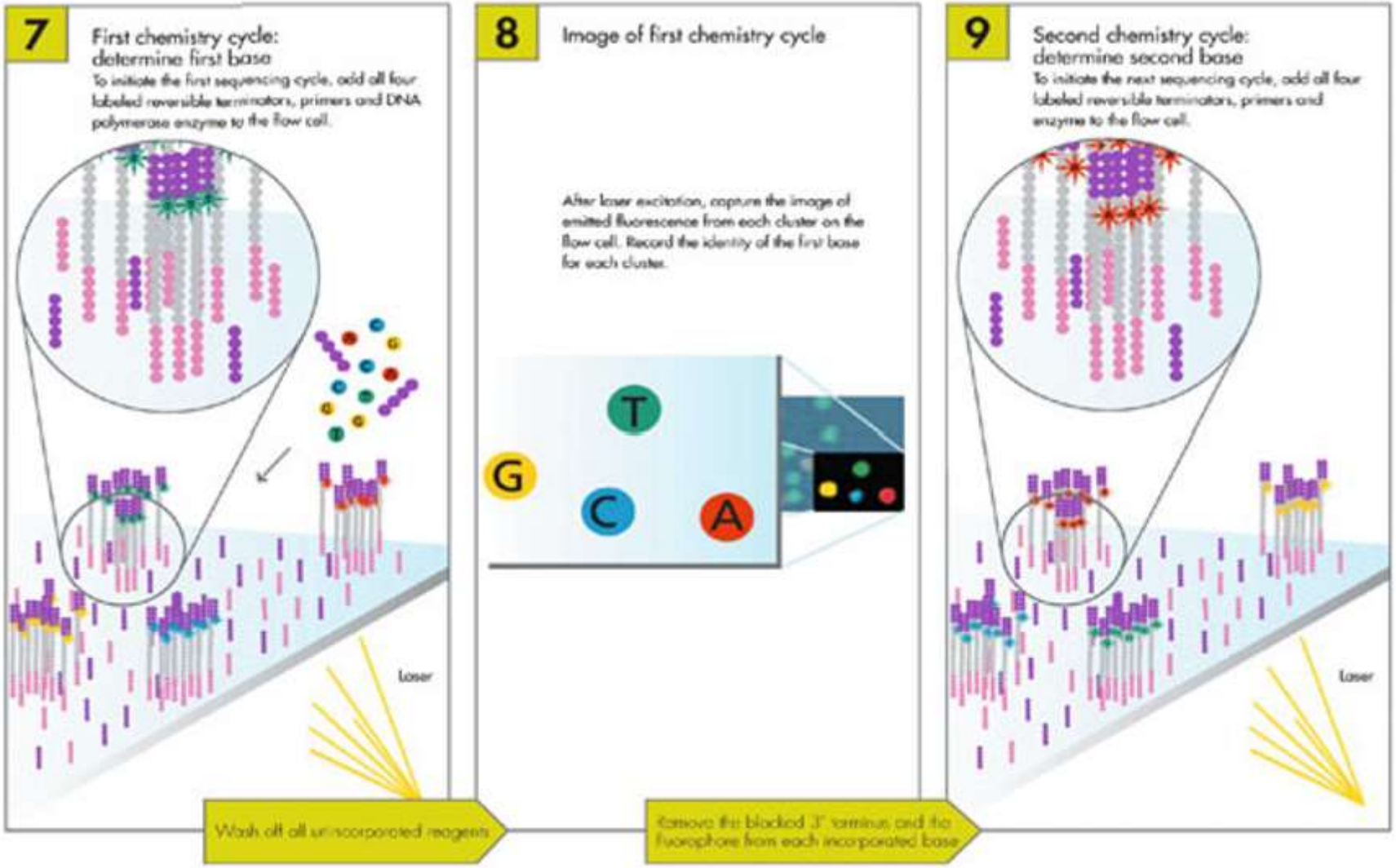
Credits: The next-generation sequencing technology and application (Zhou et al, 2010)

Example: Illumina Sequencing



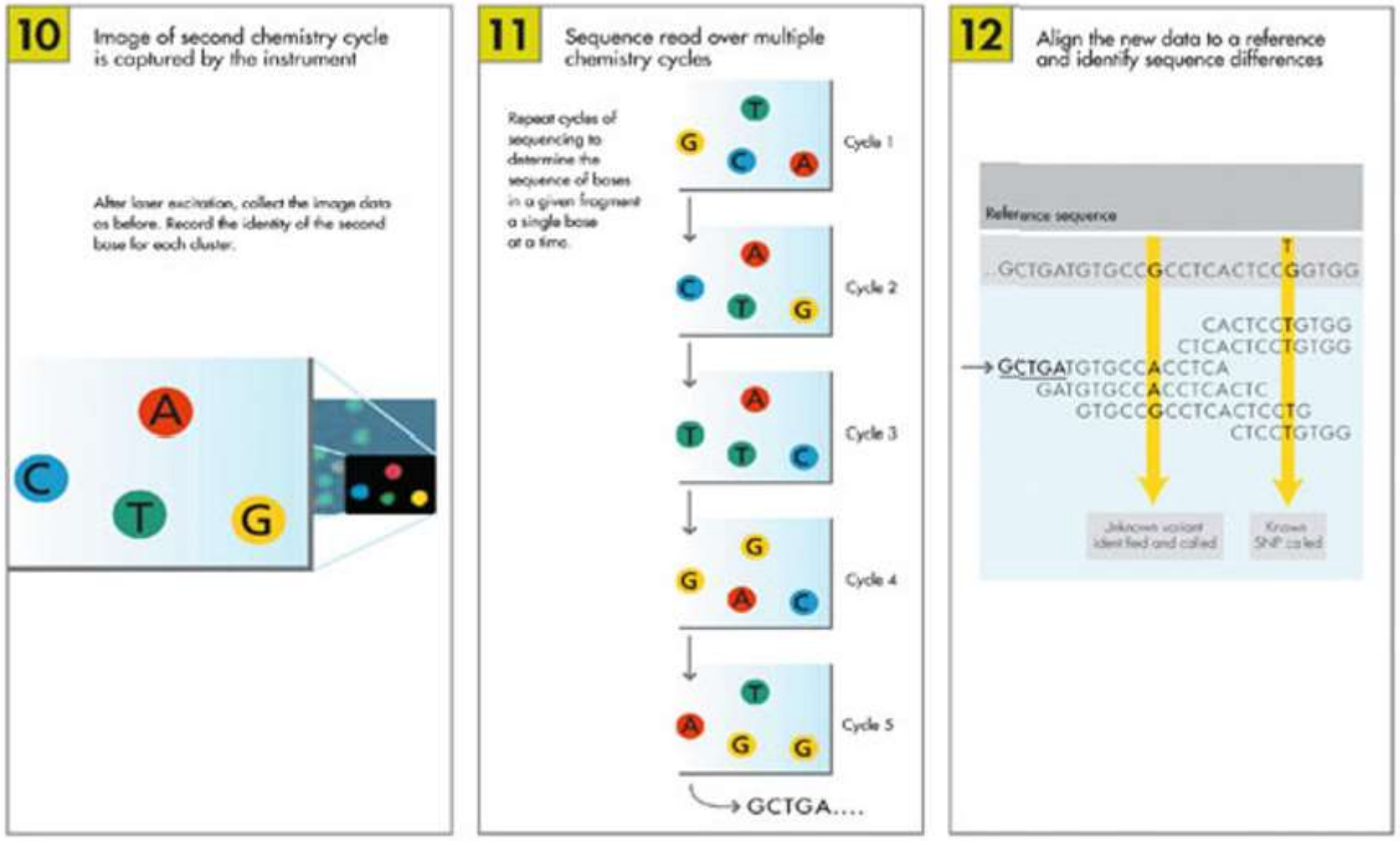
Credits: The next-generation sequencing technology and application (Zhou et al, 2010)

Example: Illumina Sequencing



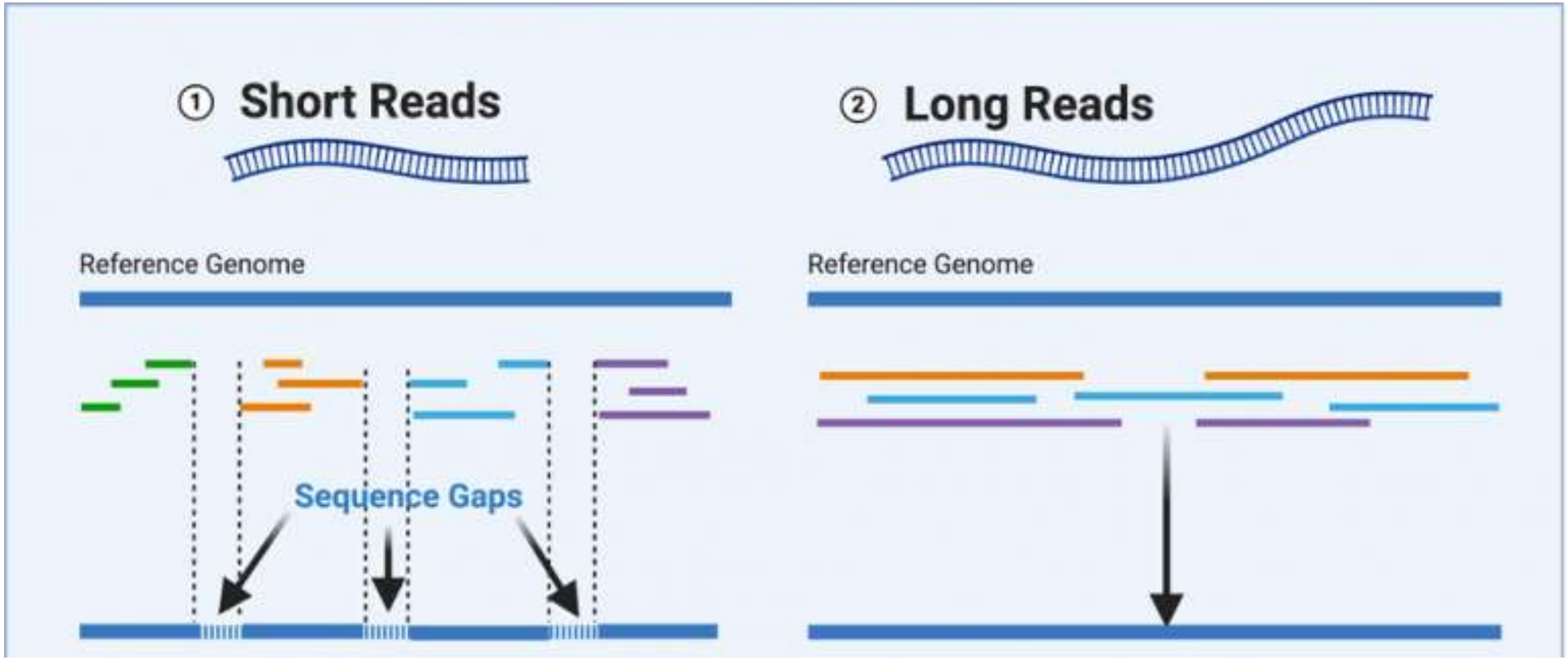
Credits: The next-generation sequencing technology and application (Zhou et al, 2010)

Example: Illumina Sequencing



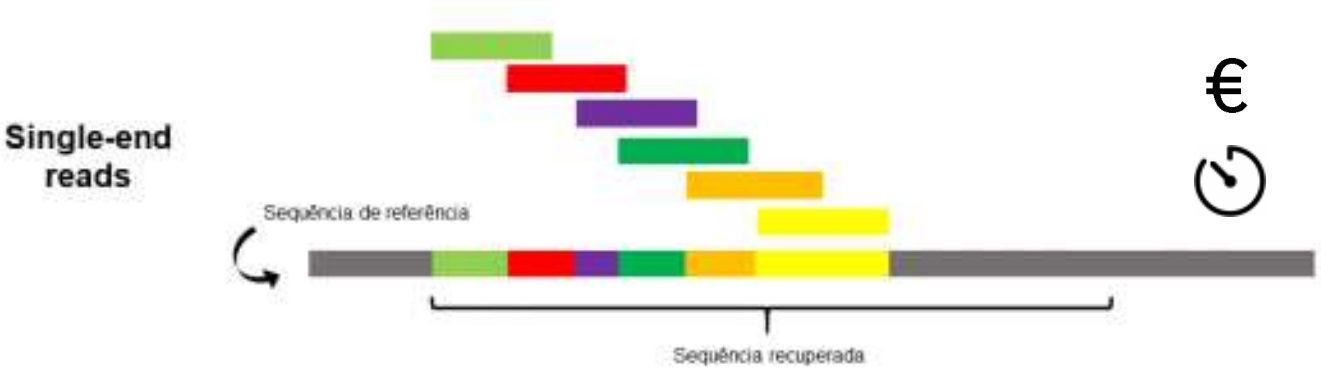
Credits: The next-generation sequencing technology and application (Zhou et al, 2010)

Reads: short vs long



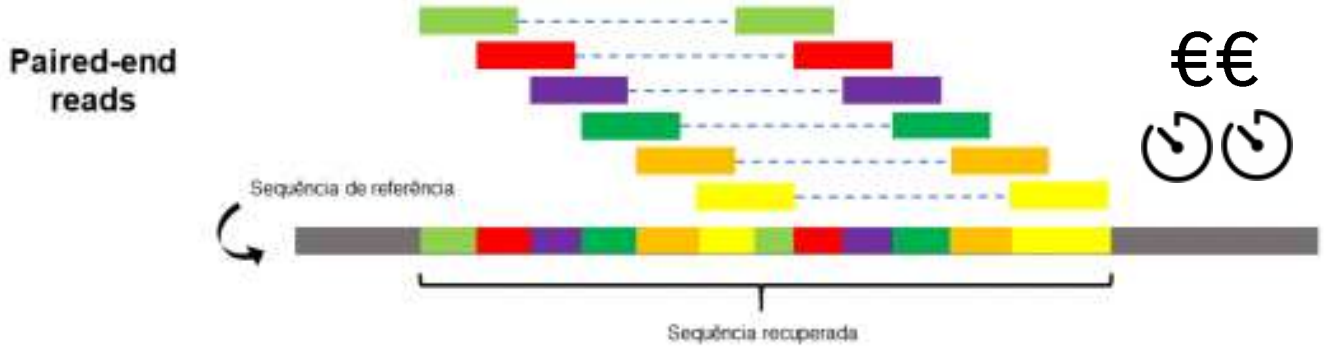
Credits: HudsonAlpha

Reads: single-end or paired-end



-> reads fragment from only one end to the other

Single-end accuracy may not be sufficient



1. starts at one direction at until specified read length
2. then starts the opposite end until specified read length

Paired-end improves accuracy for :

1. identifying relative positions of various reads -> more effective in resolving structural rearrangements:
   - gene insertions
   - deletions
   - or inversions
2. assembly of repetitive regions

Credits: Igor Paim

Outputs of sequencing: raw data = Fastq

FASTQ format stores:

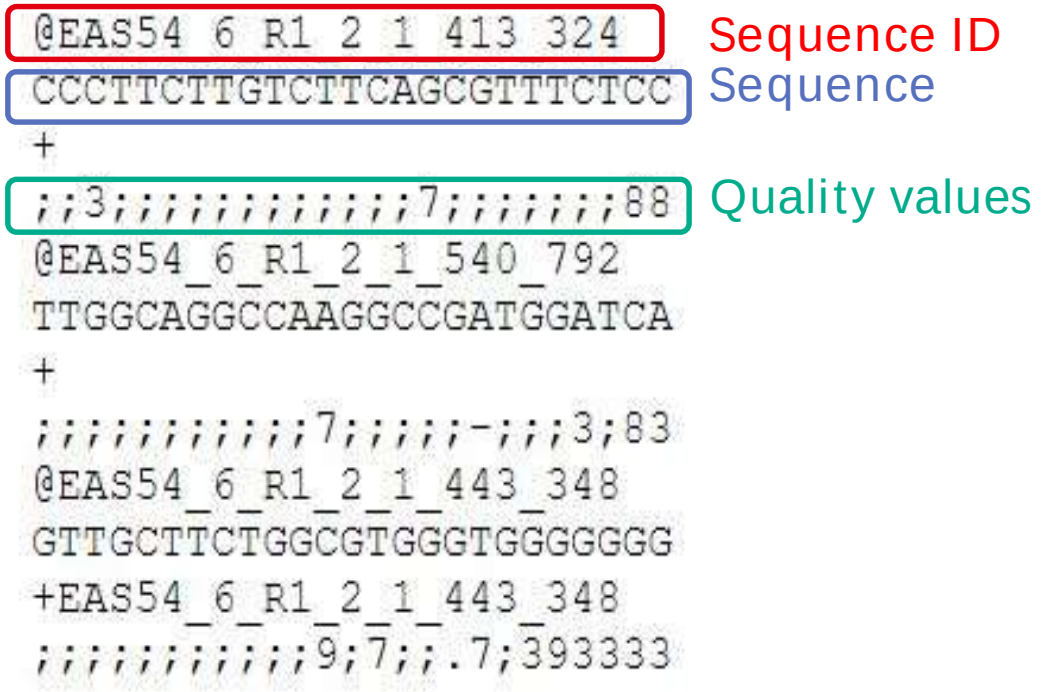
- sequences and
- Phred qualities

in a single file. => It is concise and com

Originally developed at the

Wellcome Trust Sanger Institute

Example



Choice of platform

The choice of sequencing platform depends on factors such as desired:

- read length,
- throughput,
- accuracy,
- project scale,
- and budget.

Different platforms offer distinct advantages and are suitable for specific applications within the field of genomics and molecular biology.