EURL Lm

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

INTRODUCTION TO WGS SEQUENCING PLATFORMS & OUTPUT DATA

Maroua SAYEB
eurl-listeria@anses.fr

Sequencing technologies

EURL Lm

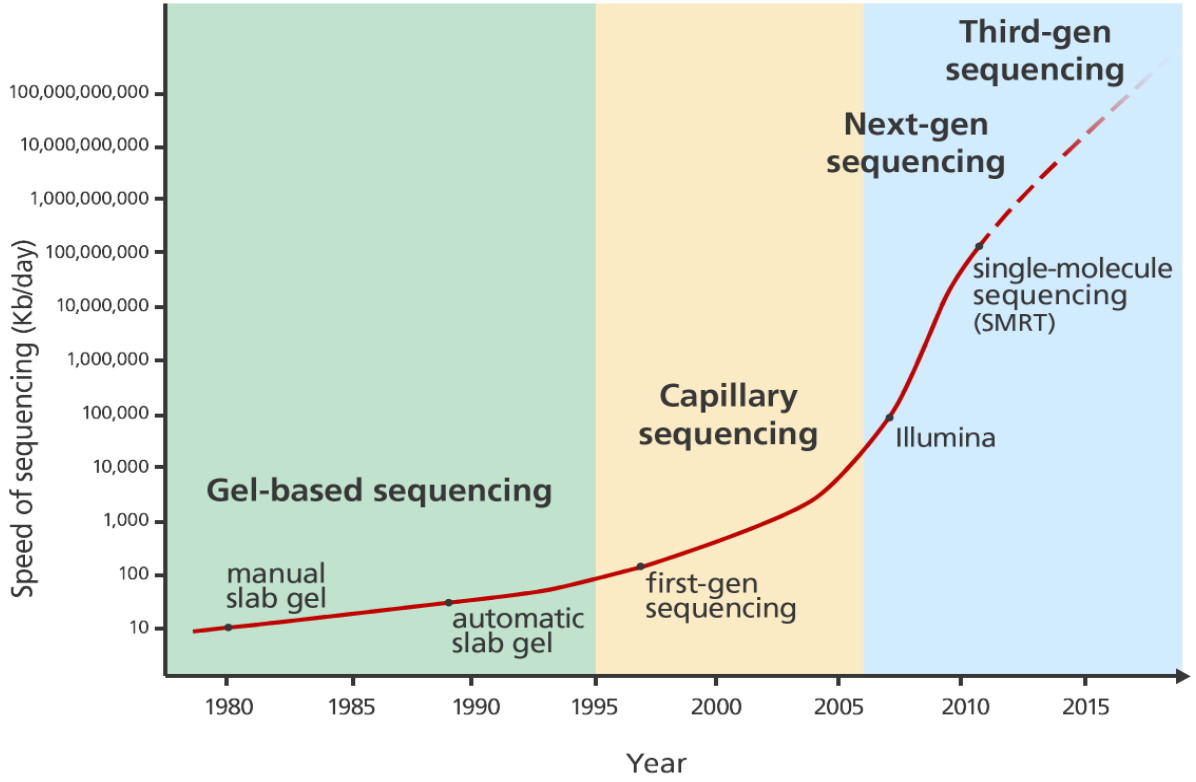
- **First generation: Sanger Sequencing**
  - Most widely used sequencing technology for approximately 25 years

- **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



ABI-Sanger
1996

Illumina-Miseq
2015

Sanger vs NGS

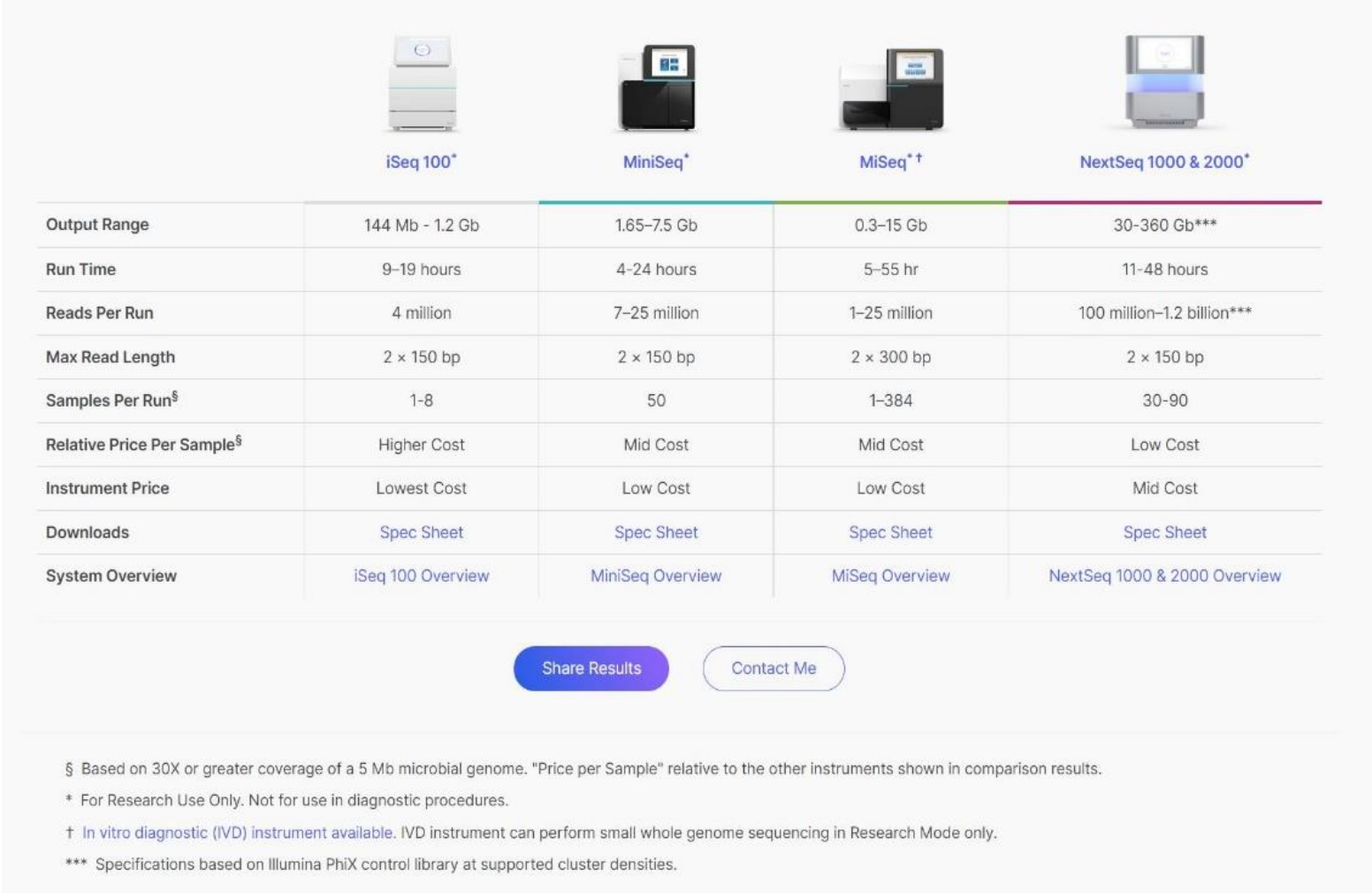
- **Both allows whole genome sequencing**
  - ➢ NGS allows massive parallel sequencing of target genes
- **Both technology standards have their utility in genomics today**
- **Cost:**
- **Sequence read length**
- **Massively Parallel**

- Millions of fragments are sequenced in a single run vs one forward and reverse read

Sequencing platforms

Platform	Read length (bp)	Isolates per run (max)	Run time	Instrument cost	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	$0.93 - $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	$180.00
Sanger	650	96	1 h	$100k	$2800

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

- Read length
- Size of genome being sequenced

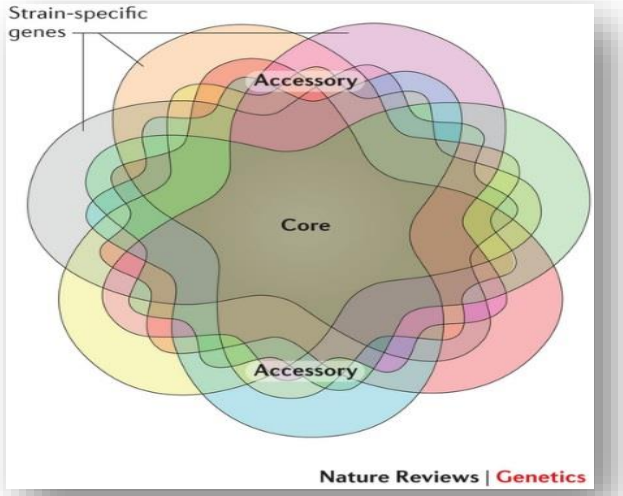
Bacterial Genome

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

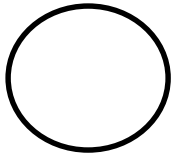
- dsDNA
- Usually single, circular chromosome



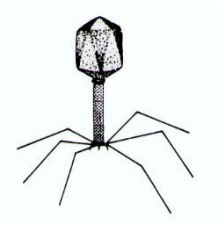
Bacterial Genome Complications

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



Viral Genome

EURL Lm

Compact

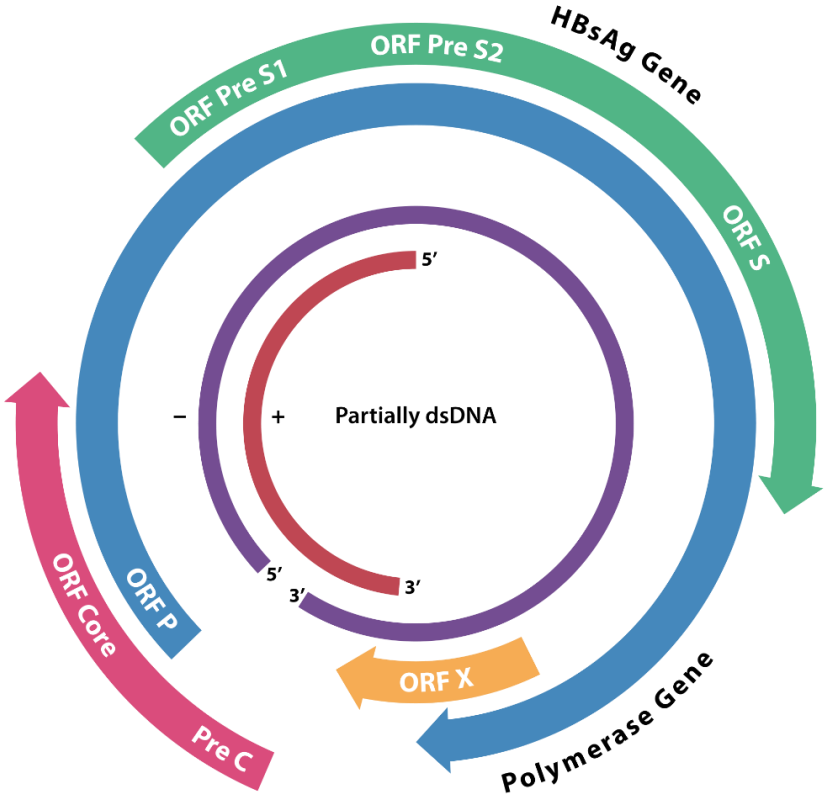
- ~10,000 nucleotides (nt) (typically, ~3,000–200,000)
- Little wasted space

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



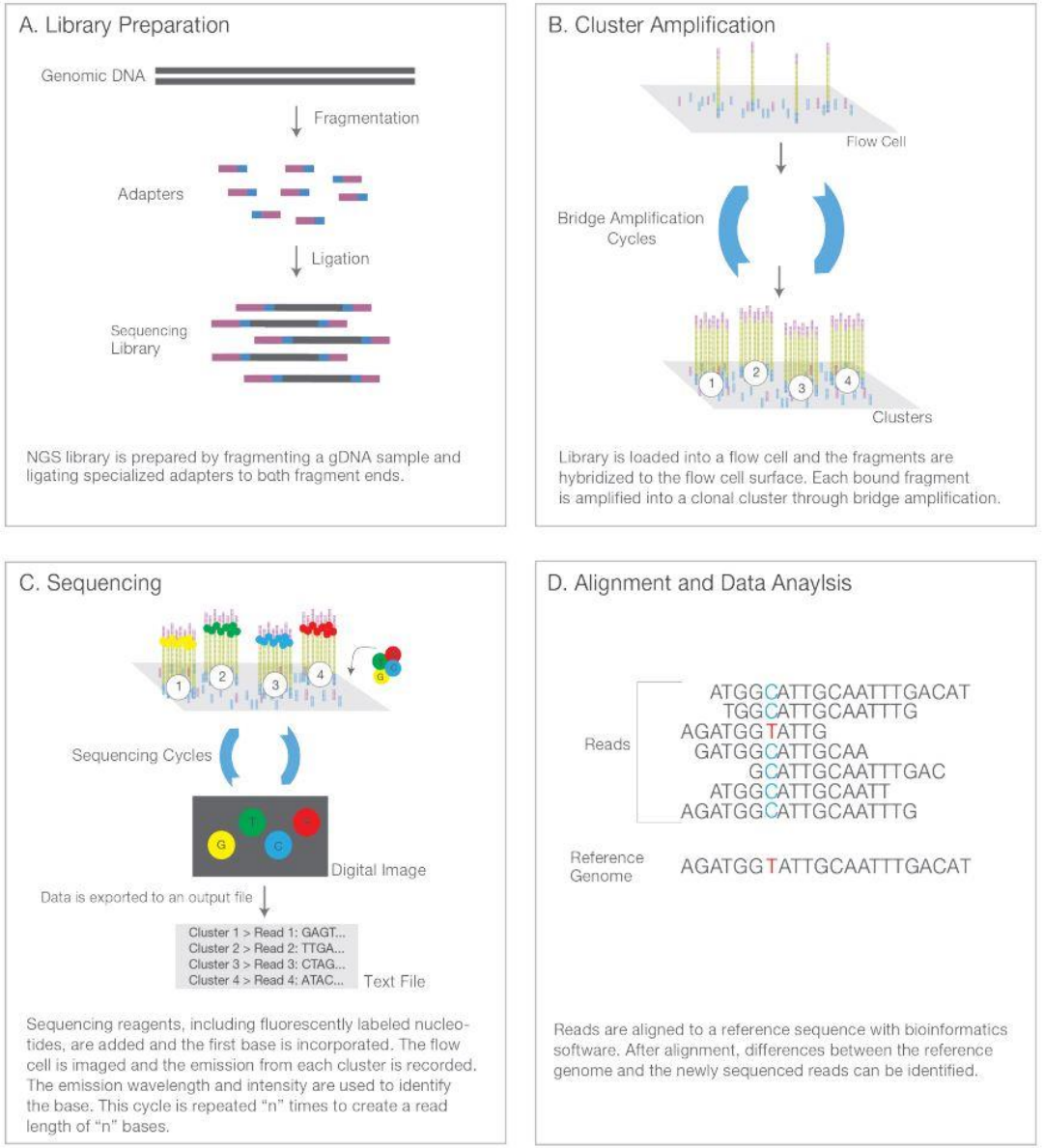
Example: HBV Genome

Sequencing Process Definitions

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

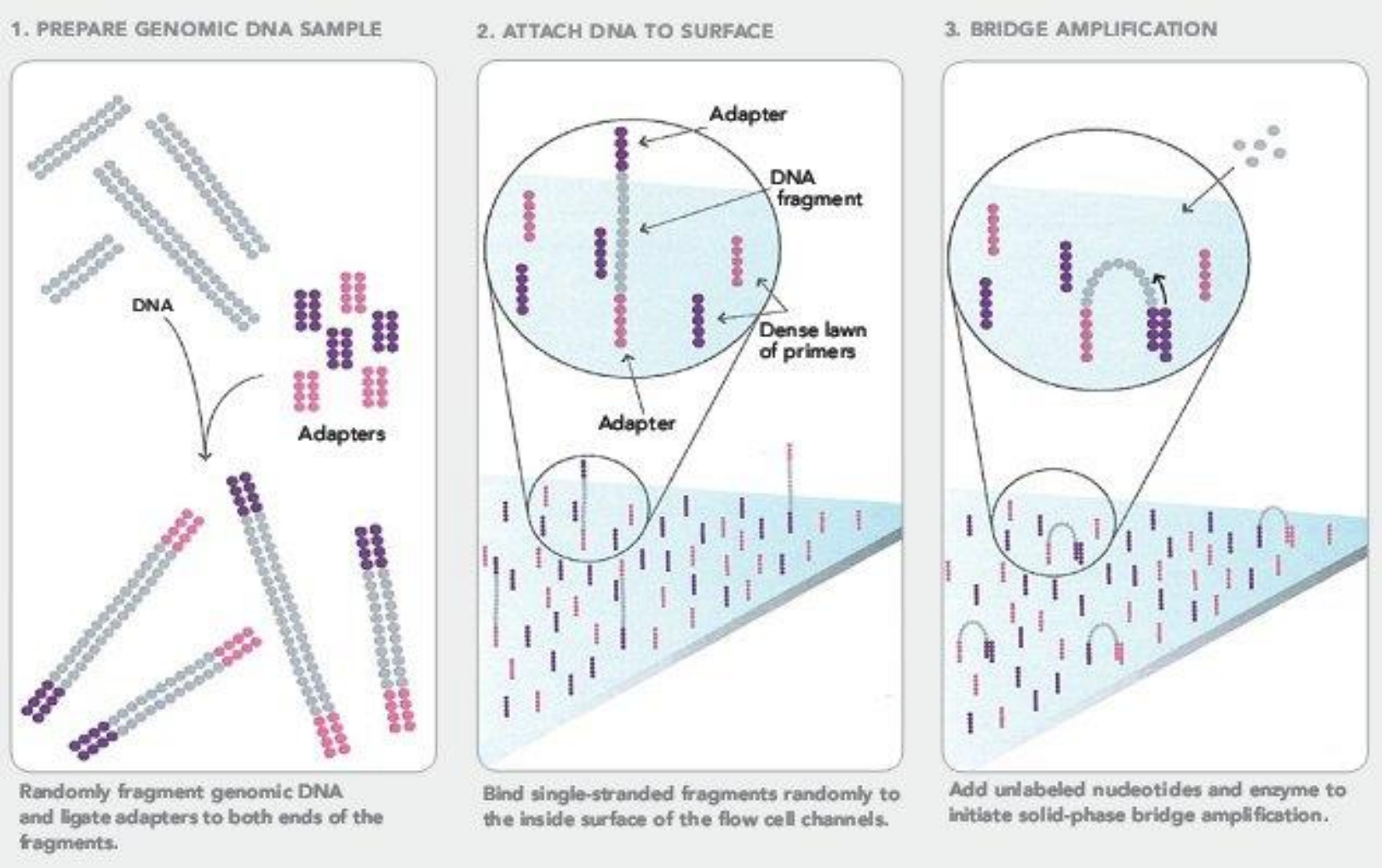
- Reads: sequenced reads of base pairs as you are trying to assemble

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

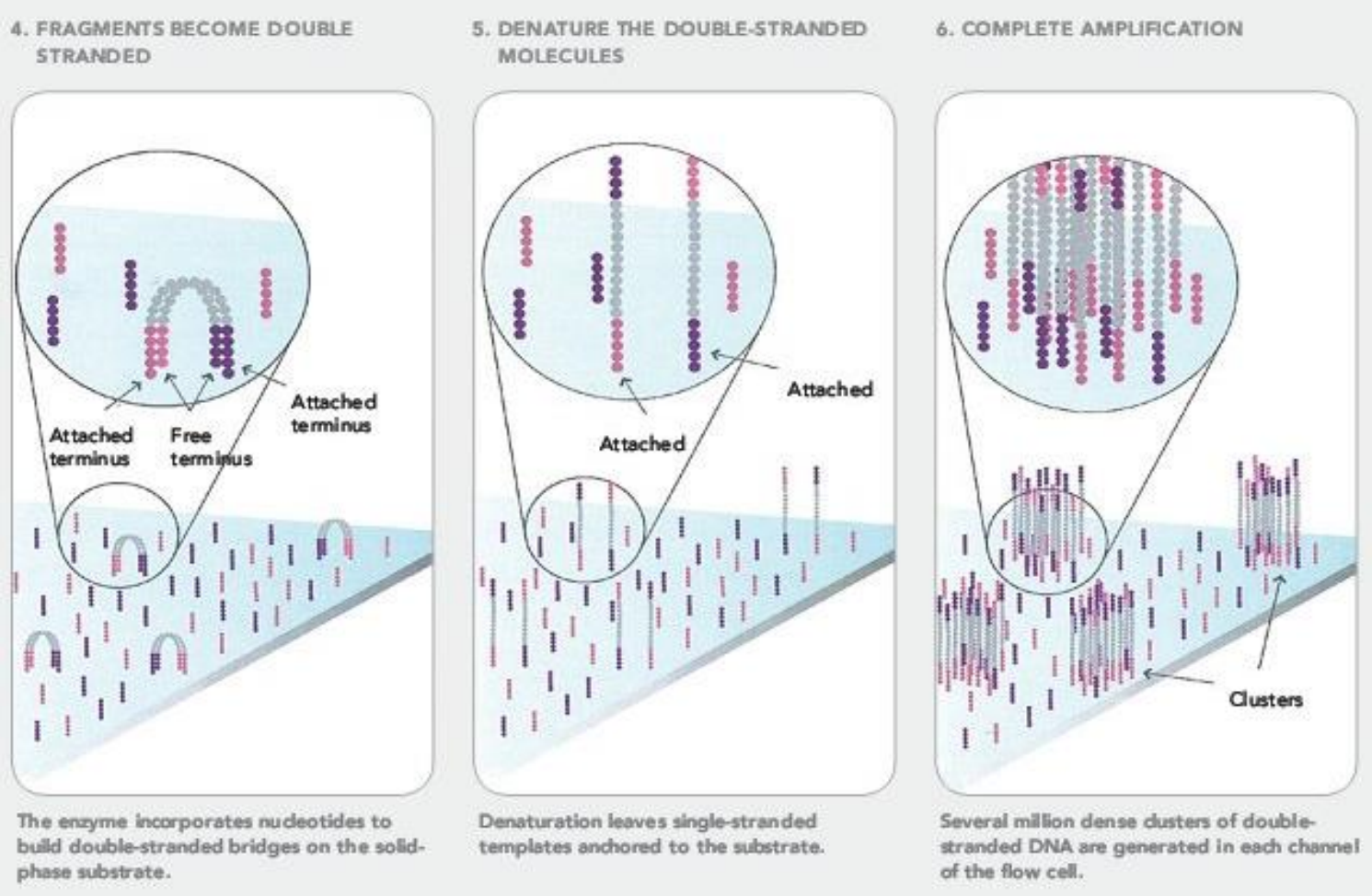


Next-Generation Sequencing Chemistry

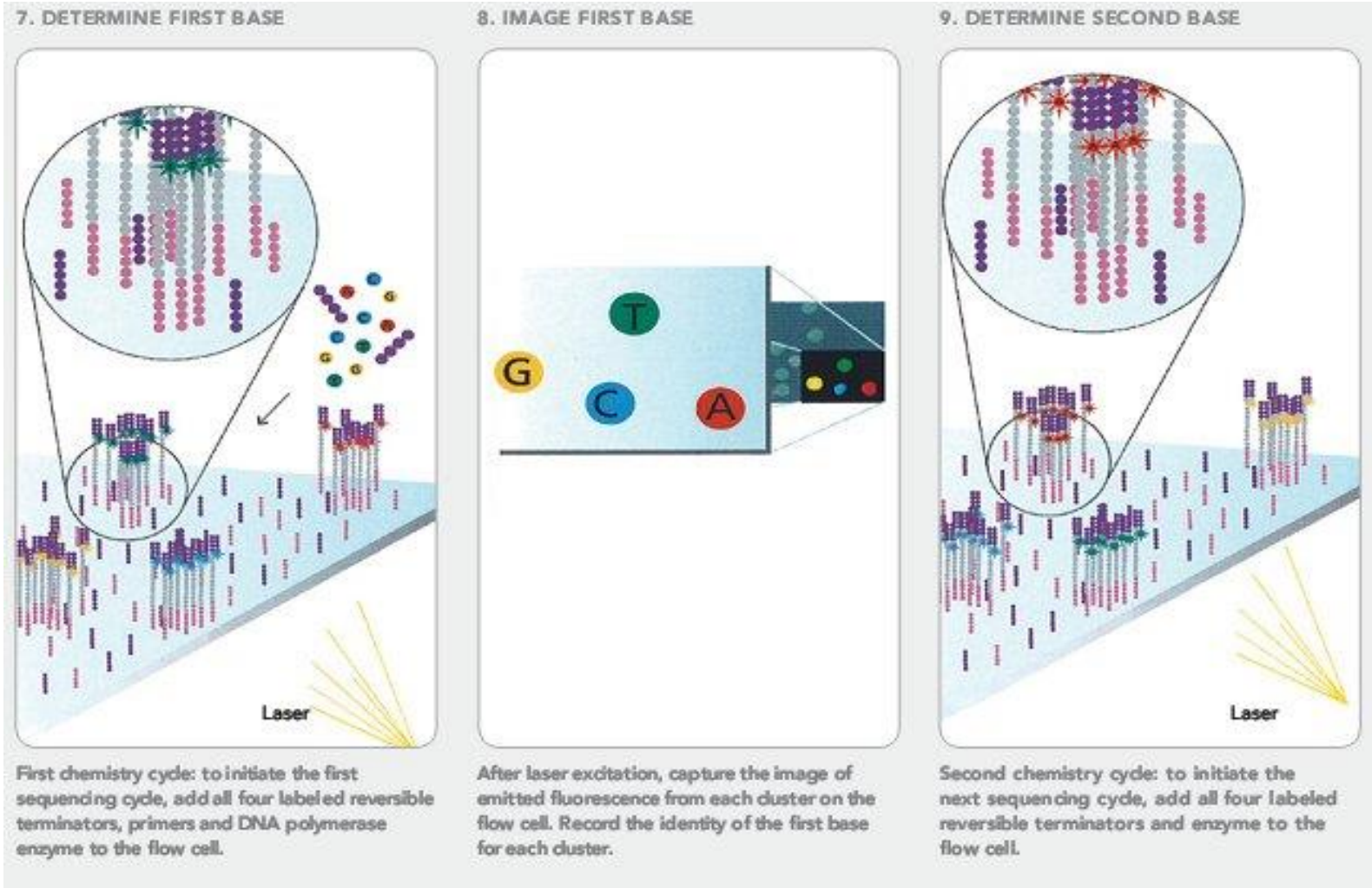
Example: Illumina Sequencing



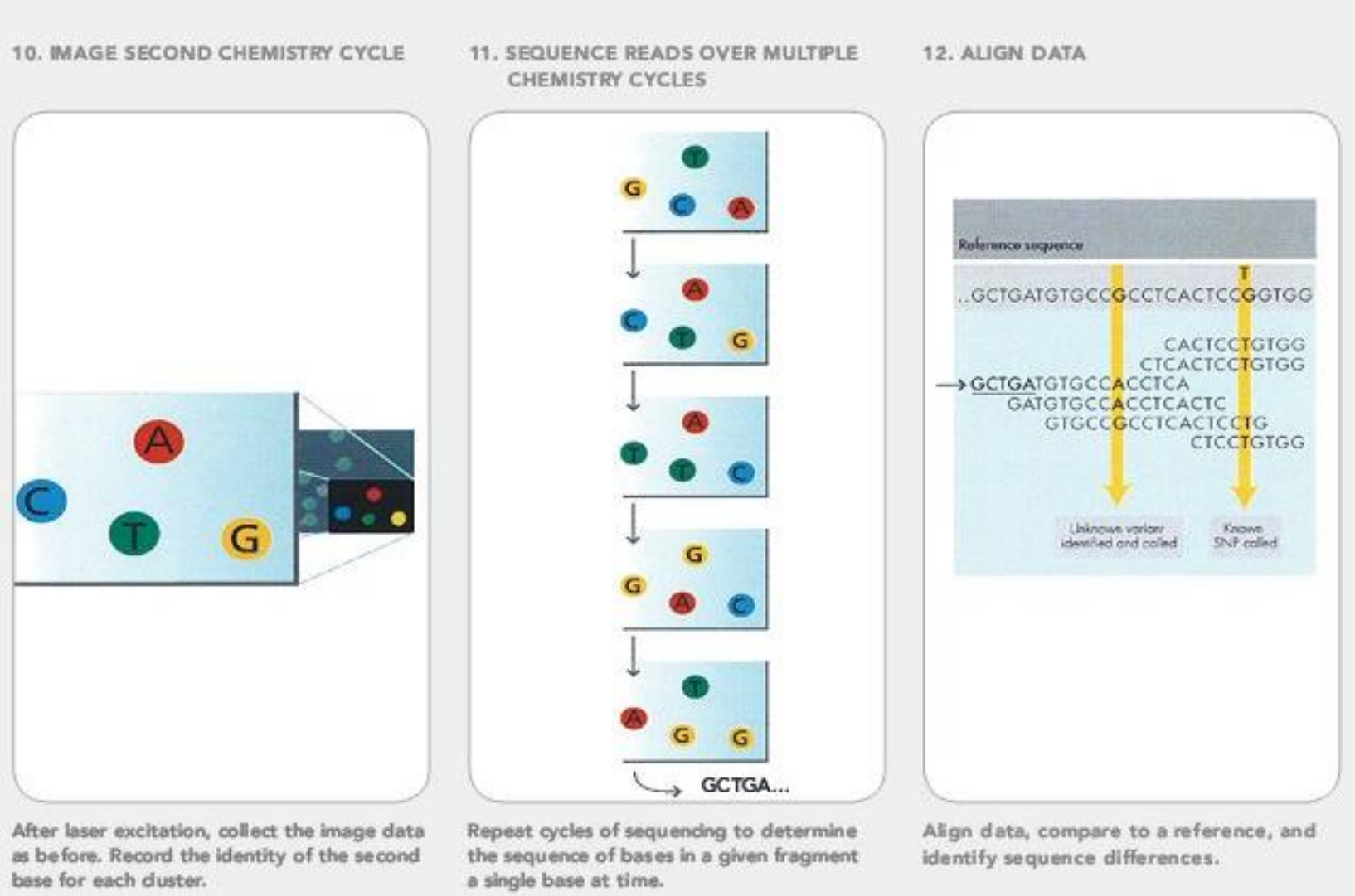
Example: Illumina Sequencing



Example: Illumina Sequencing



Example: Illumina Sequencing



Outputs of sequencing: raw data = Fastq

EURL Lm

- FASTQ format stores sequences and Phred qualities in a single file. It is concise and compact.

- FASTQ is first widely used in the Sanger Institute and therefore we usually take the Sanger specification and the standard FASTQ format, or simply FASTQ format.

Example

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

FASTQ Format Specification

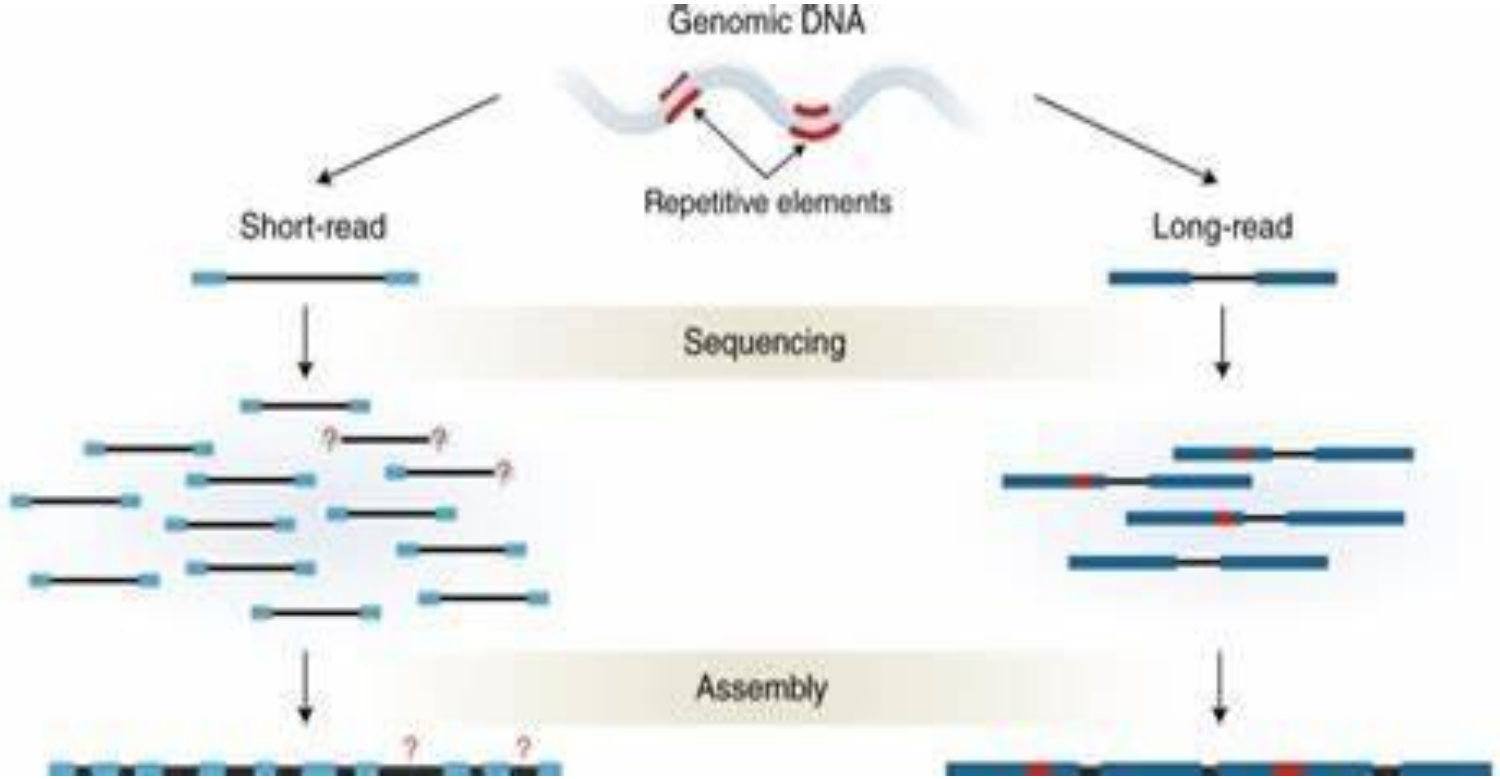
Notations

- <fastq>, <blocks> and so on represents non-terminal symbols.
- Characters in red are regex-like operators.
- '\n' stands for the Return key.

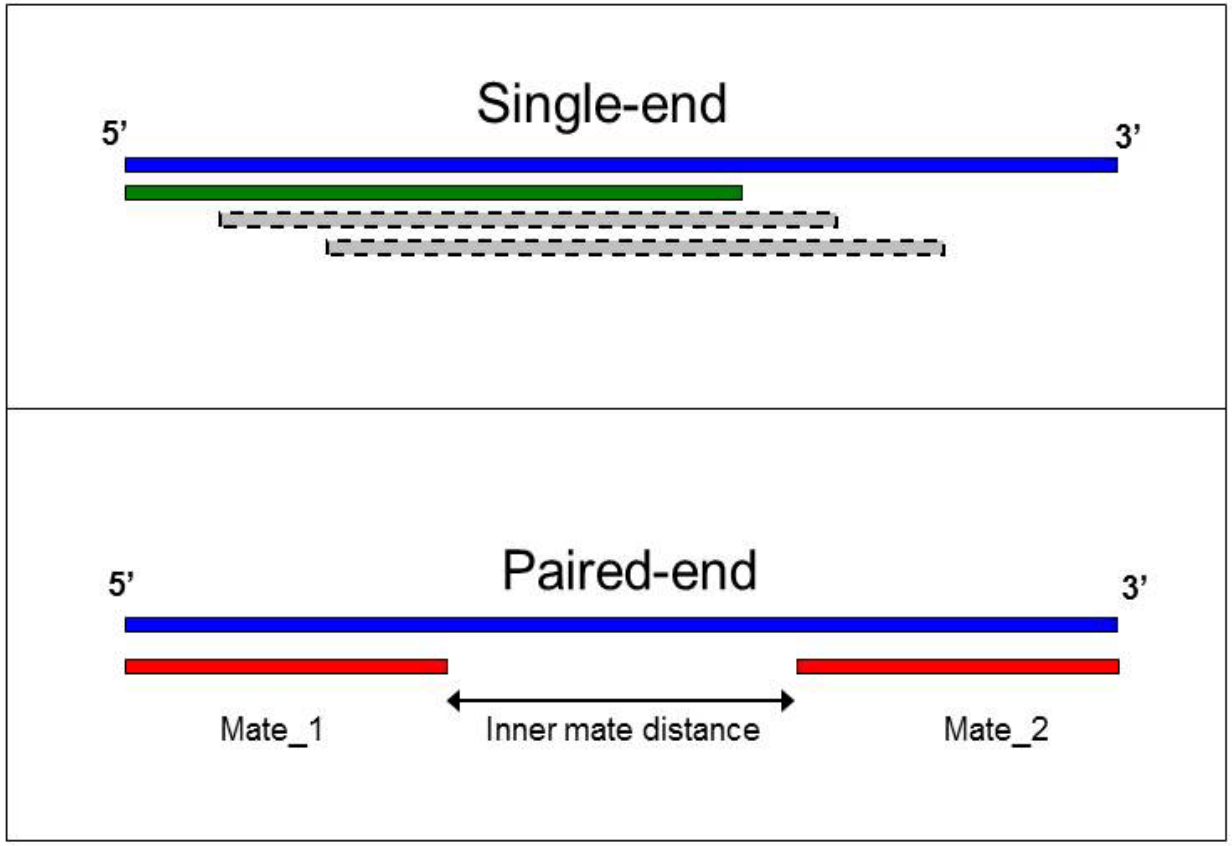
Syntax

```
<fastq>     := <block>+
<block>     := @<seqname>\n<seq>\n+[<seqname>]\n<qual>\n
<seqname>   := [A-Za-z0-9_.:-]+
<seq>       := [A-Za-z\n\.~]+
<qual>      := [!-~\n]+
```

Reads: short vs long

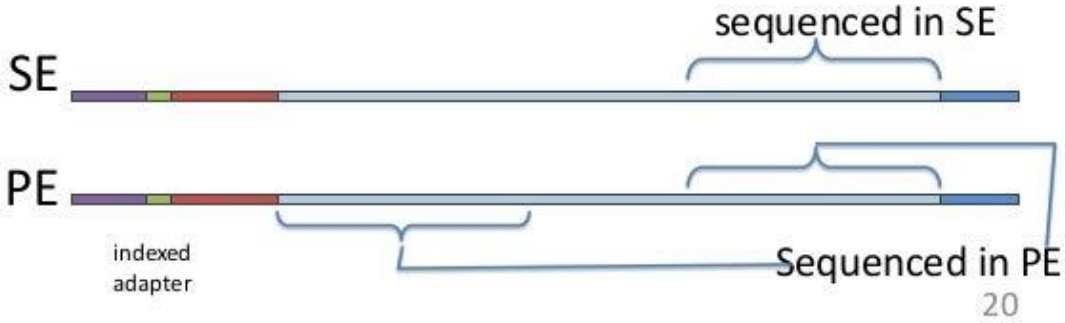


Reads: single end or paired-end



Review: Paired End vs Single End

- Single End – **cheaper**
- Paired End – more expensive
  - two reads per fragment
  - counting fragments, not reads
  - call normalized counts *FPKM* not *RPKM*



20

Outputs of sequencing

Single-end vs Paired-end Sequencing

Once you have your nucleic acids ready to go you can then choose whether you want single-end or paired-end data.

Single-Read Sequencing

Single-read sequencing involves sequencing DNA from only one end, and is the simplest way to utilize Illumina sequencing. By leveraging proprietary reversible terminator chemistry and a novel polymerase, this solution delivers large volumes of high-quality data, rapidly and economically.

Single-Read Sequencing Highlights

- **Simple Library Preparation:** Follows standard molecular biology methods; compatible with robotics
- **Low Input DNA Requirements:** As little as 100 ng genomic DNA or cDNA
- **Economical:** 1/100th the cost of traditional Sanger sequencing
- **Simplified Data Analysis:** High-quality sequence assemblies with short-insert libraries

Paired-End Sequencing Highlights

- **Simple Paired-End Libraries:** Simple workflow allows generation of unique ranges of insert sizes
- **Efficient Sample Use:** Requires the same amount of DNA as single-read gDNA or cDNA sequencing
- **Broad Range of Applications:** Does not require methylation of DNA or restriction digestion; can be used for bisulfite sequencing
- **Simplified Data Analysis:** Higher quality sequence assemblies with short-insert libraries. A simple modification to the standard single-read library preparation process facilitates reading both the forward and reverse template strands of each cluster during one paired-end read. Both reads contain long-range positional information, allowing for highly precise alignment of reads.

Example: performance analysis of sequencing platforms Illumina

Platform	Read length (bp)	Output (Gb)	Coverage	Run time (h)	No. of reads	Cost per Gb ($)	Consumables cost ($)	Instrument cost ($)	Error rate	Dimensions (width × depth × ht) (cm)	Source(s) (reference[s])
Illumina MiniSeq Mid Output	2 × 150c	2.1–2.4c	8.6	17c	14 million–16 millionc	2,584–2,953d	6,201c	55,411c	0.1% in >80% of base callsc	45.6 × 48 × 51.8c	Illumina
Illumina MiniSeq High Output	1 × 75c	1.7–1.9c	6.8	7c	22 million–25 millionc	3,264–3,648d	6,201c	55,411c	0.1% in >85% of base callsc	45.6 × 48 × 51.8c	Illumina
	2 × 75c	3.3–3.8c	13.6	13c	44 million–50 millionc	1,632–1,879d	6,201c	55,411c	0.1% in >85% of base callsc	45.6 × 48 × 51.8c	Illumina
	2 × 150c	6.6–7.5c	26.9	24c	44 million–50 millionc	827–940d	6,201c	55,411c	0.1% in >80% of base callsc	45.6 × 48 × 51.8c	Illumina
Illumina MiSeq Reagent kit v2	1 × 36c	0.54–0.61c	2.2	4c	12 million–15 millionc	7,946–8,976d	4,847c	108,244c	0.1% in >90% of base callsc	68.6 × 56.5 × 52.3c	Illumina
	2 × 25c	0.75–0.85c	3.1	5.5c	24 million–30 millionc	5,702–6,463d	4,847c	108,244c	0.1% in >90% of base callsc	68.6 × 56.5 × 52.3c	Illumina
	2 × 150c	4.5–5.1c	18.3	24c	24 million–30 millionc	950–1,077d	4,847c	108,244c	0.1% in >80% of base callsc	68.6 × 56.5 × 52.3c	Illumina
	2 × 250c	7.5–8.5c	30.5	39c	24 million–30 millionc	570–646d	4,847c	108,244c	0.1% in >75% of base callsc	68.6 × 56.5 × 52.3c	Illumina
Illumina MiSeq Reagent kit v3	2 × 75c	3.3–3.8c	13.6	21c	44 million–50 millionc	1,362–1,568d	5,174c	108,244c	0.1% in >85% of base callsc	68.6 × 56.5 × 52.3c	Illumina
	2 × 300c	13.2–15c	53.8	56c	44 million–50 millionc	345–392d	5,174c	108,244c	0.1% in >70% of base callsc	68.6 × 56.5 × 52.3c	Illumina
Ilumina NextSeq 500 Mid Output	2 × 75c	16.3–20c	71.8	15c	<260 millionc	318–391d	6,369c	266,835c	0.1% in >75% of base callsc	53.3 × 63.5 × 58.4c	Illumina
	2 × 150c	32.5–39c	140	26c	<260 millionc	163–196d	6,369c	266,835c	0.1% in >80% of base callsc	53.3 × 63.5 × 58.4c	Illumina
Illumina NextSeq 500 High Output	1 × 75c	25–30c	107.7	11c	<400 millionc	312–374d	9,347c	266,835c	0.1% in >80% of base callsc	53.3 × 63.5 × 58.4c	Illumina
	2 × 75c	50–60c	215.3	18c	<800 millionc	156–187d	9,347c	266,835c	0.1% in >80% of base callsc	53.3 × 63.5 × 58.4c	Illumina
	2 × 150c	100–120c	430.6	29c	<800 millionc	78–93d	9,347c	266,835c	0.1% in >75% of base callsc	53.3 × 63.5 × 58.4c	Illumina

Conclusion

Choice of sequencing platform depending on the needs

Listeria

Short paired end reads – typing by cgMLST

E coli

Campylobacter

?

Thank you for your attention !