ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

# INTRODUCCIÓN A LA SECUENCIACIÓN DE GENOMA COMPLETO


JOSEFINA CAMPOS
E: jcampos@anlis.gov.ar

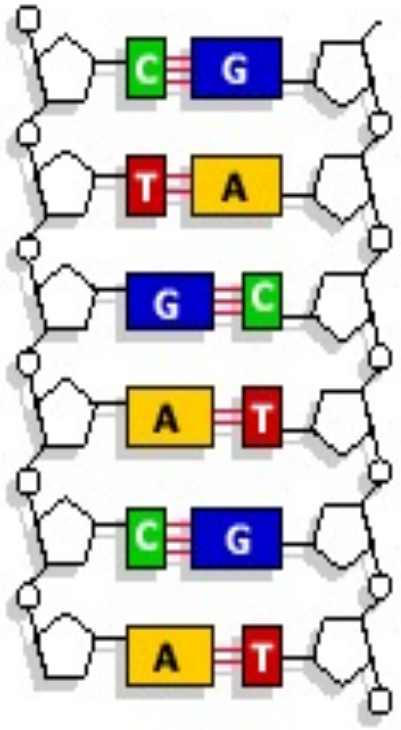
UNIDAD OPERATIVA CENTRO NACIONAL DE GENÓMICA Y BIOINFORMÁTICA-ANLIS

# SECUENCIACIÓN

Conjunto de métodos y técnicas bioquímicas cuya finalidad es la determinación del orden de los nucleótidos (A, C, G y T) en un oligonucléotido de ADN

**Estudio de genes o grupo de genes**

**Estudio contenido genómico**



# PRIMEROS PASOS

ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

Frederick Sanger 1918-2013

- 1975 - Técnica de Sanger
  - 1977 Secuencia de bacteriófago ΦX 174
  - 1995 Secuencias
    - *Haemophilus influenzae*
    - *Micoplasma genitalium*
  - 2001 Primera secuencia borrador de genoma humano

- 2005 - Técnicas de Secuenciación de Nueva Generación
  - ROCHE 454 – Pirosecuenciación
  - Wheeler *et al*. 2007 - 1°secuencia de genoma humano por WGS

James Watson 1928

# CÓMO SE LLAMA

Secuenciación de Nueva Generación

Segunda y Tercera Generación

Secuenciación de genoma completo
- Whole Genome Sequencing - (WGS)

Secuenciación paralela masiva

Secuenciación de alto rendimiento

(“High-throughput”)

# CLASIFICACIÓN

ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

**PRIMERA GENERACIÓN**

Secuanciación de Sanger-Dye terminator

Ej: ABI Applied biosystems

**SEGUNDA GENERACIÓN**

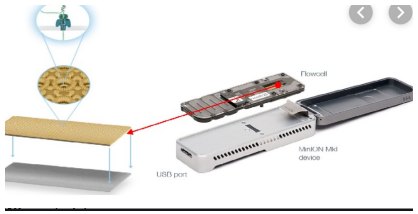
Secuenciación por síntesis paralela masiva

Ej: Roche 454, Illumina, Ion Torrent

**TERCERA GENERACIÓN**

Secuenciación por síntesis en una célula

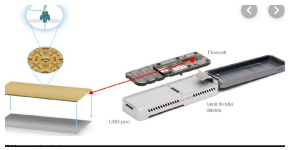
Ej: PacBio, Nanopore



# CLASIFICACIÓN

ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

## Instrumentos "Bench-top"



- 454 GS Junior (ROCHE)
- Ion Proton (LT)
- MiSeq (Illumina)
- MinIon (Nanopre)

## Instrumentos "High-end"

- 454 GS FLX (ROCHE)
- HiSeq 2000/2500 /Novaseq (Illumina)
- 5500xl Solid (Life Tech.)
- Pac Bio RS (Pacific biosc.)
- GridIon (Nanopre)

# PRIMERA GENERACIÓN:SANGER

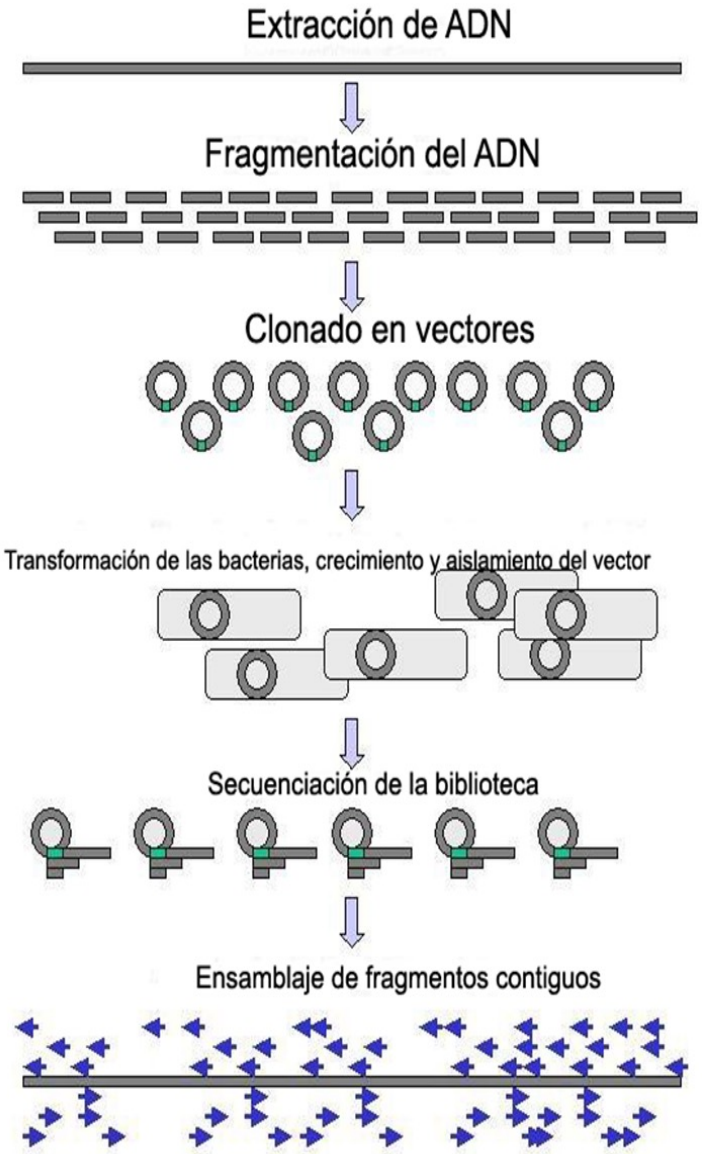
**Fundamentos**

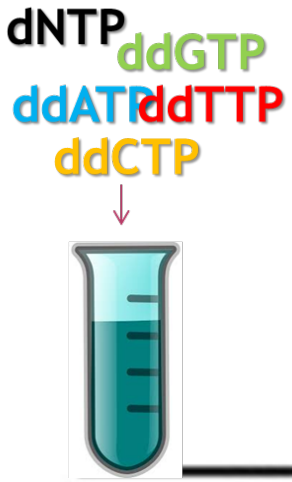
- Se basa en el proceso biológico de replicación del ADN
- Método de terminación de la cadena por utilización de **ddNTP**

  **(dideoxiribonucleótido trifosfato)**

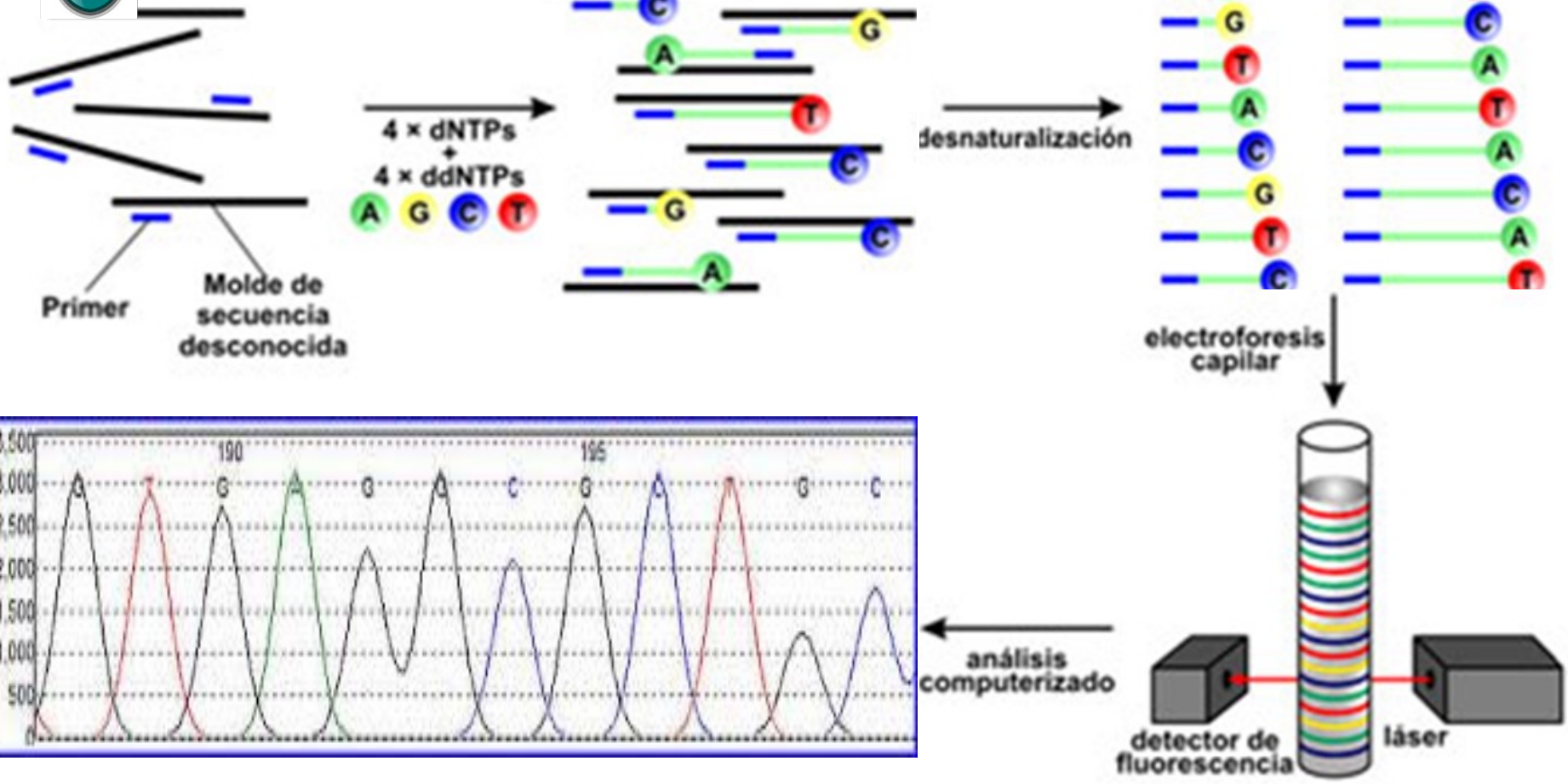
**Procedimiento**

- Preparación de “libraries” por clonado
- Mezcla de reacción / polimerización
  - 1 por cada nucleótido
  - Primers $^{32}P$ / dNTPs $^{35}S$ o $^{32}P$
- Electroforesis en gel de poliacrilamida



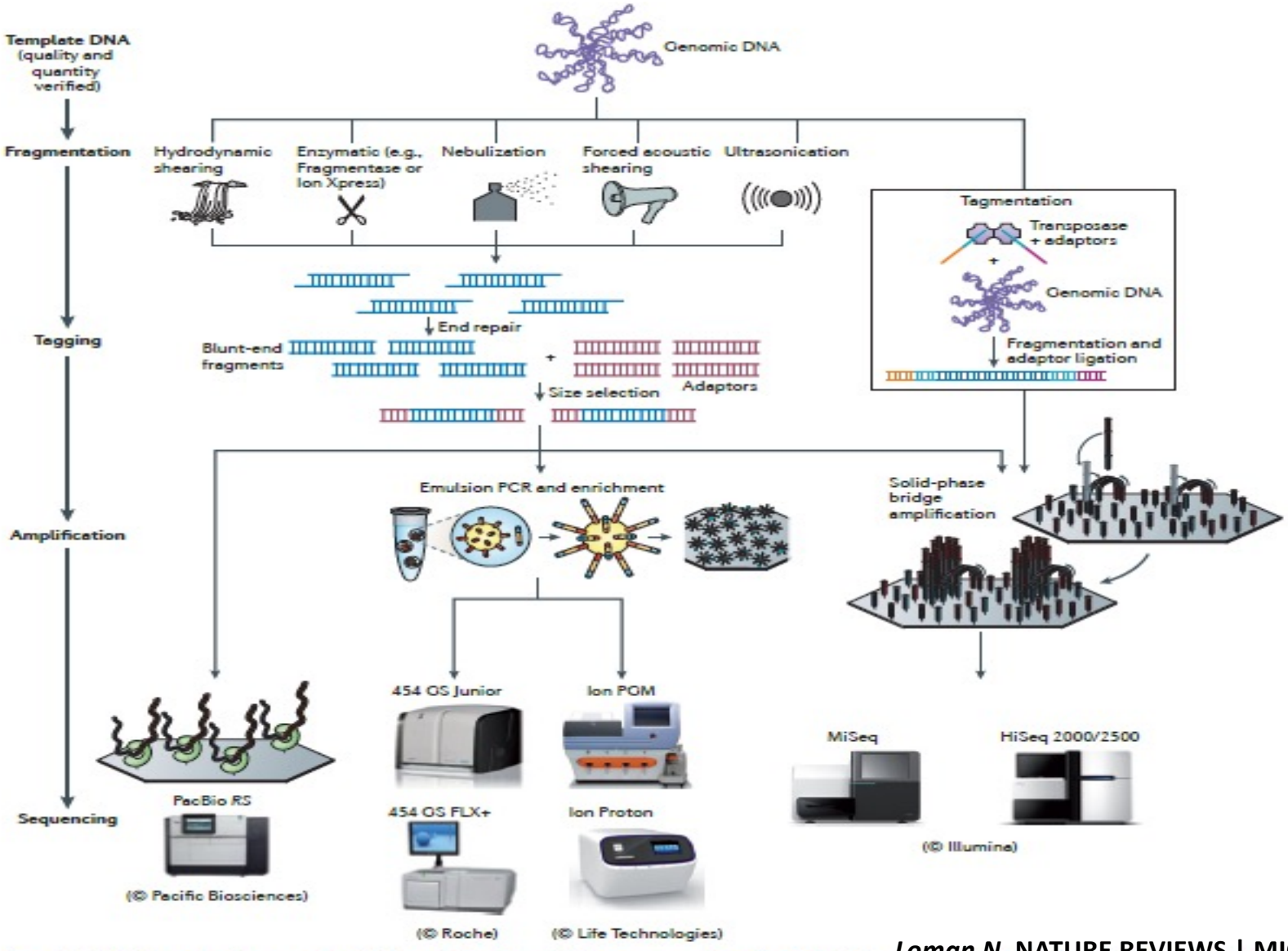



- Marcación terminadores con distinto
  1 mezcla de reacción
- Biopolímeros
- Lector Láser - electroferogramas



# TECNOLOGÍAS NGS

ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"




*Loman N.* NATURE REVIEWS | MICROBIOLOGY
VOLUME 10 | SEPTEMBER 2012 | 599

# LECTURAS CORTAS

ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

Short-read Sequencing Platforms and various characteristics.

| Company | Illumina | | | | | | | | ThermoFisher | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| System Platform | iSeq | Miniseq | MiSeq | NextSeq550 | NextSeq 1000&2000 | NovaSeq 6000 | MiSeqDx | NextSeq550 Dx | GeneStudio S5 | Genexus | Ion PGM-Dx |
| Sequencing Principle | Sequence by Synthesis | | | | | | | | | | |
| Detection | Fluorescent | | | | | | | | Ion | | |
| Applications | Small WGS, TS, Small RNA sequencing | | Small WGS, TS, ChIP-Seq, Small RNA sequencing | TS, small WGS, exome and transcriptome sequencing | | TS, WGS, WES, transcriptome and epigenome sequencing | TS, Small WGS | TS, exome and transcriptome sequencing | TS, epigenetic, exome, and transciptome sequencing | TS | TS |
| Maximum Read length (bases) | 2 × 150 bp | | 2 × 300 bp | 2 × 150 bp | 2 × 150 bp | 2 × 250 bp | 2× 300 bp | 2× 150 bp | 600 bp | 400 bp | 200 bp |
| Flow cells/device | 1 | | | | | 2 | 1 | | | | |
| Output (per flow cell) | 1.2 Gb | 7.5 Gb | 15 Gb | 120 Gb | 330 Gb | 3000 Gb | ≥5 Gb | ≥90 Gb | 15 Gb | 24 Gb | 1 Gb |
| Sequencing Run time | 9.5-19 hr | 5-24 hr | 4-56 hr | 11-29 hr | 11-48 hr | 13-44 hr | 24 hr | ≤35 hr | 4.5-21.5 hr | 14-31 hr | 4.4 hr |
| Accuracy/Quality Score | Q30≥ 80% (2 × 150 bp) | | Q30≥ 70% (2 × 300 bp) | Q30≥ 75% (2 × 150 bp) | | Q30≥ 75% (2 × 250 bp) | Accuracy >99.66%, Q30> 80% | Accuracy ≥99.98%, Q30≥75% | Accuracy ≥99% | Accuracy ≥99% | Accuracy ≥99% |
| Equipment Cost (USD) | $19,900 | $49,500 | $99,000 | $275,000 | $335,000 | on request | | | | | |

Hu, 2021

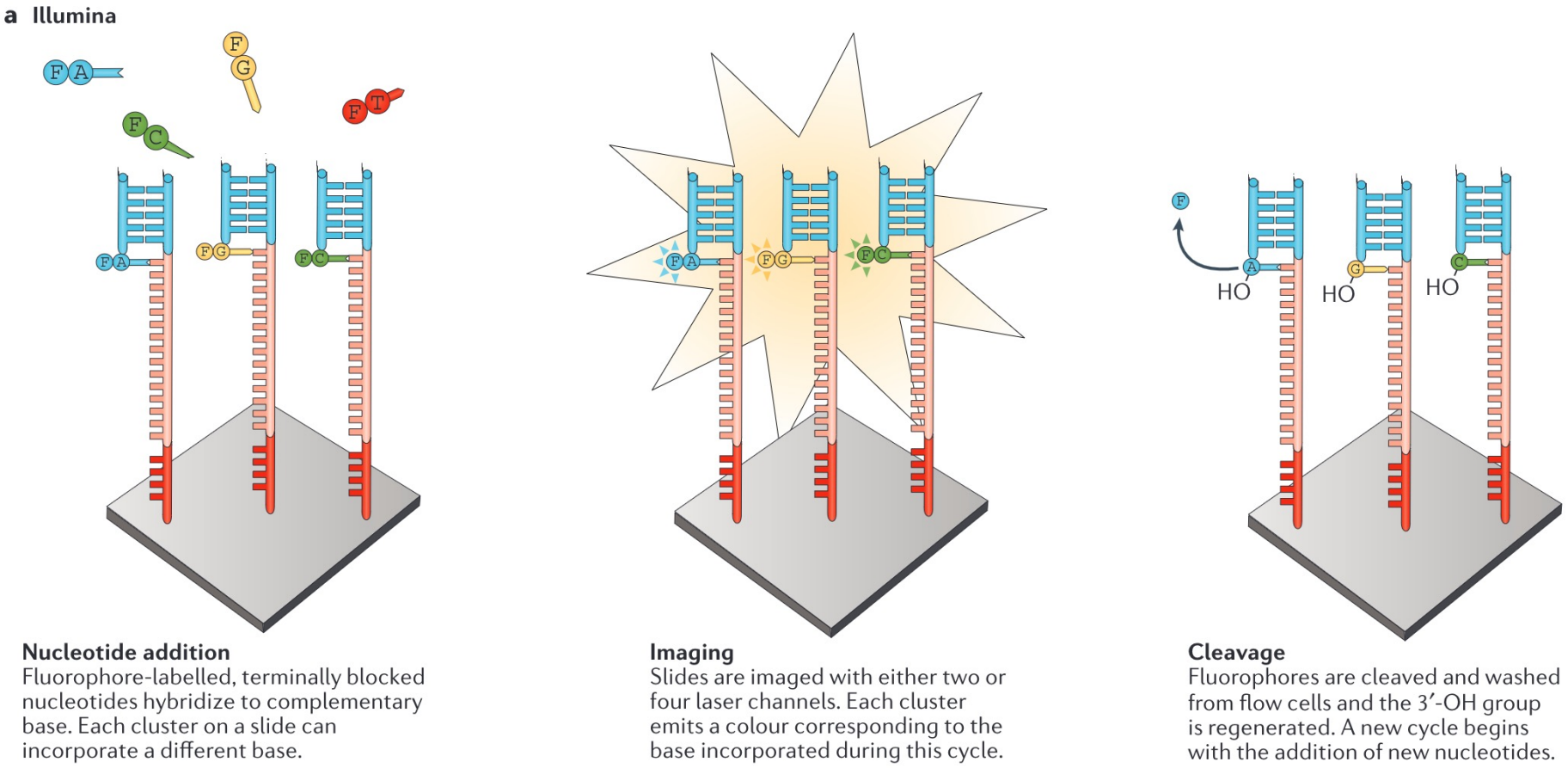
# LECTURAS LARGAS

ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

| System Platform | Sequel | Sequel II | Sequel IIe | Flongle | MinION | GridION | PromethION |
|---|---|---|---|---|---|---|---|
| Sequencing Principle | PacBio Single Molecule Sequencing | | | Nanopore Single molecule Sequencing | | | |
| Detection | Fluorescent | | | Electrical Conductivity | | | |
| Applications | Whole genome *de novo* assembly, variant detection, structural variation detection, full length transcript sequencing, targeted/amplicon sequencing, metagenomics sequencing | | | DNA, amplicons, cDNA, Direct RNA sequencing | | | |
| Maximum Read length (bases) | 300 kb | | | Longest read so far: > 4 Mb | | | |
| Flow cells/device | 12 SMRT Cells 1M can be used at a time, and 8 SMRT Cell 8M can be used serially | | | 1 (126 channels per flow cell) | 1 (512 channels per flow cell) | 5 (512 channels per flow cell) | 24 or 48 (3000 channels per flow cell) |
| Output (per flow cell) | 75 Gb | 600 Gb | | 1 - 2 Gb[a] | 10 - 30 - 50 Gb[a] | | 100 - 200 - 300 Gb[a] |
| Sequencing Run time | Up to 20 hr | Up to 30 hr | | 1 min - 16 hr | 1 min - 72 hr | | |
| Accuracy/Quality Score | Number of HiFi Reads >99% Accuracy: Up to 5,000,000 reads | Number of HiFi Reads >99% Accuracy: Up to 4,000,000 reads | | **Single Molecule:** R9 modal Accuracy >98.3%, R10 modal Accuracy >97.5%. New chemistry Accuracy >99% (coming soon) **Consensus:** R9.4.1: Current best Q45 (>99.99%) R10: Current Best Q50 (99.999%) | | | |
| Equipment Cost (USD) | approximately $525,000 | | | $1,460 (12 flow cells included) | $9,300 | $69,955 | 24 flow cells: $335,455<br>48 flow cells: $530,000 |

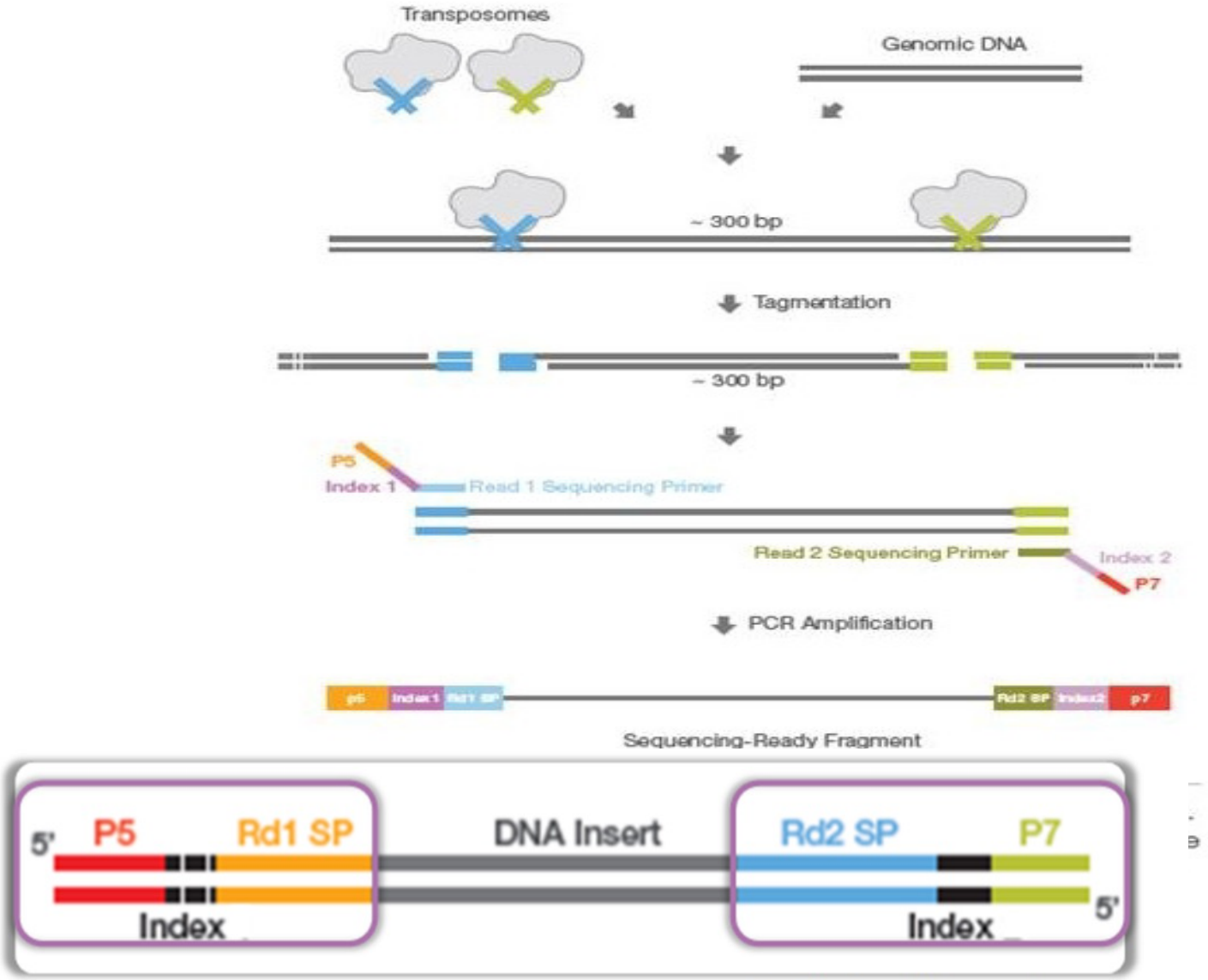
Hu, 2021

# SEGUNDA GENERACIÓN ILLUMINA



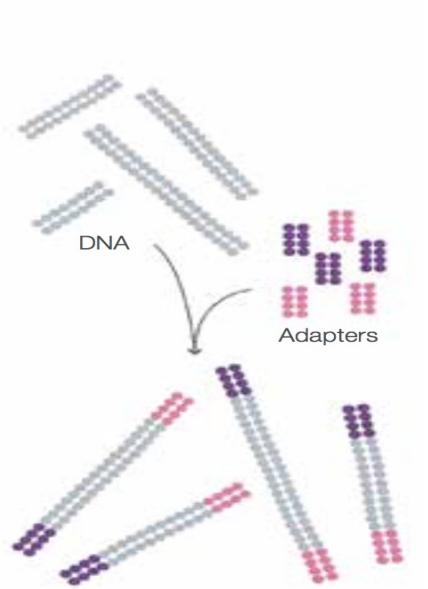

Goodwin, 2016. Nature Reviews

# SEGUNDA GENERACIÓN ILLUMINA



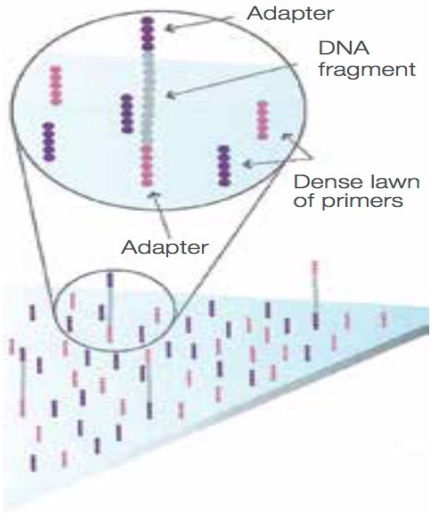
# SEGUNDA GENERACIÓN ILLUMINA

**Figure 2: Prepare Genomic DNA Sample**




Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

**Figure 3: Attach DNA to Surface**

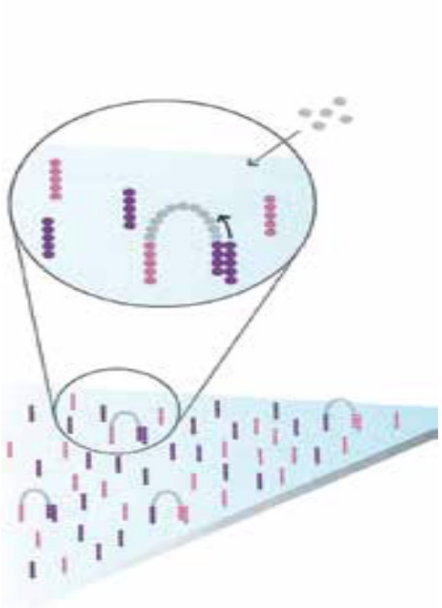



Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

# SEGUNDA GENERACIÓN ILLUMINA

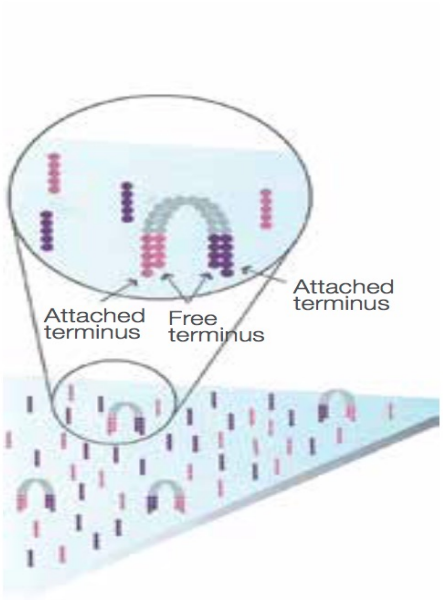
ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

**Figure 4: Bridge Amplification**



Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

**Figure 5: Fragments Become Double Stranded**

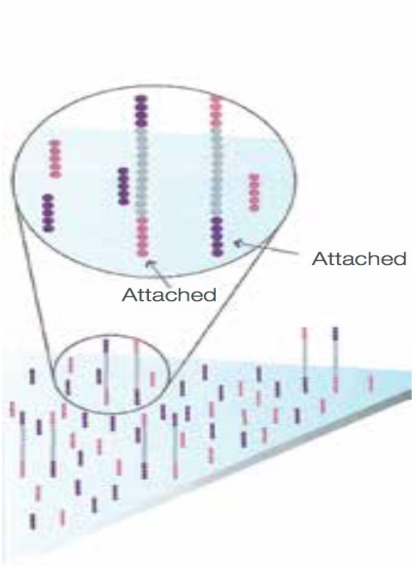



The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

# SEGUNDA GENERACIÓN ILLUMINA

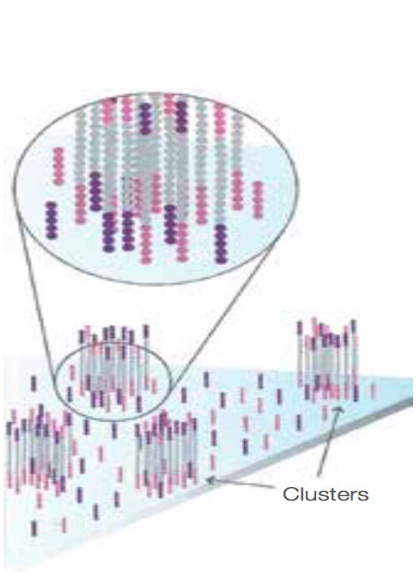
ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

**Figure 6: Denature the Double-Standed Molecules**




Denaturation leaves single-stranded templates anchored to the substrate.

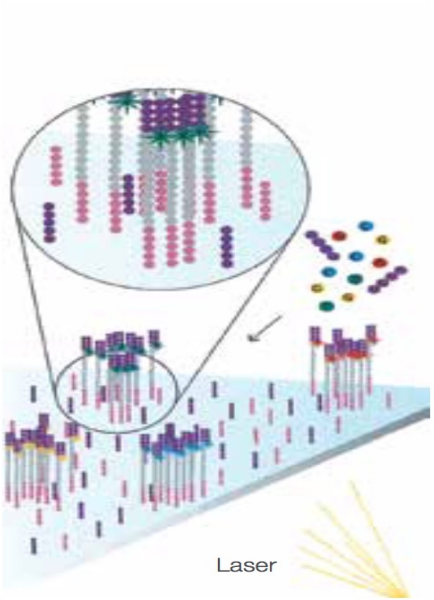
**Figure 7: Complete Amplification**




Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

# SEGUNDA GENERACIÓN ILLUMINA

**Figure 8: Determine First Base**




The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

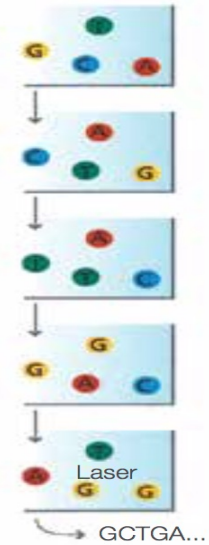
**Figure 9: Image First Base**


After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

# SEGUNDA GENERACIÓN ILLUMINA

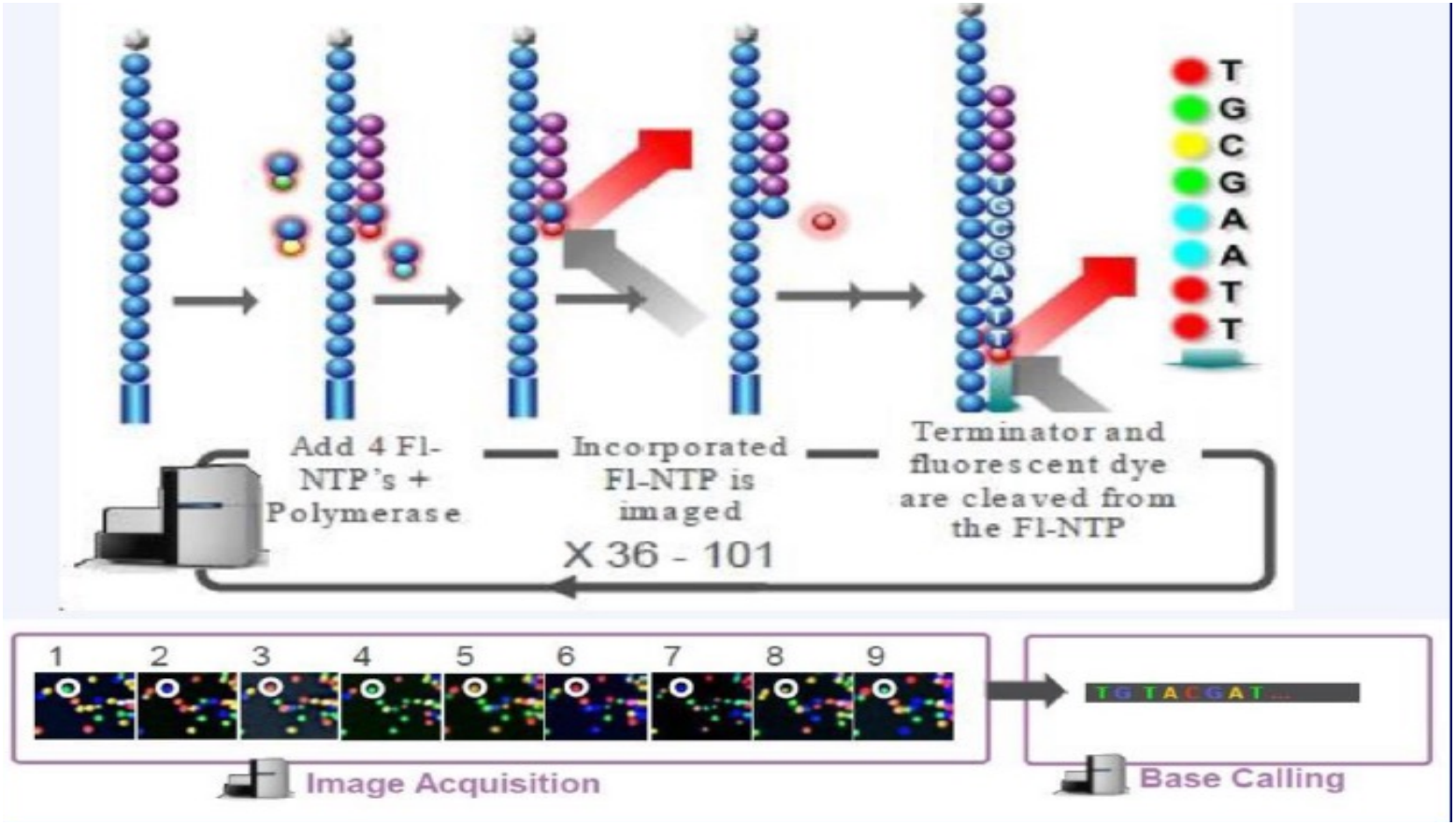
ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

**Figure 12: Sequencing Over Multiple Chemistry Cycles**




The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

# SEGUNDA GENERACIÓN ILLUMINA



- https://www.youtube.com/watch?annotation_id=annotation_228575861&feature=iv&src_vid=womKfikWlxM&v=fCd6B5HRaZ8

# SEGUNDA GENERACIÓN ILLUMINA

ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

Extracción ADN

Preparación de la Librería

Amplificación
Secuenciación

# Preparación de la librería

- 1 día antes preparar el cultivo e incubar ON ~30 minutes

B. Purificación del ADN

- 1 h & 1 h tiempo de corrida Qiacube ~ 2hrs

***---- Good stopping point -----***

Cuantificación

~30 min

D. Dilución

~1.5 h

E. Tagmentation

~30 min

F. PCR Amplificación

- 1 h & 40 min PCR Total ~ 1 h 40 min

G. PCR Cleanup

~45 min

***---- Good stopping point -----***


Normalización de la librería

~2 hours

I. Pooling

# Mi Seq V2 Reagent Kits
# Box 1 / box 2

**Temperatura de conservación**

**BOX 2**

**4°C**

**BOX 1**

**-20°C**


- Librería preparada
- Último paso de pooling

**PAL**: 5 ul de cada biblioteca en un tubo

**DAL**: 24 ul + 576 ul de HT1

-Mezclar con pipeta 3-5 veces

-Vortexear a max vel

-Incubar 96°C 2 min

-Mezclar por inversión

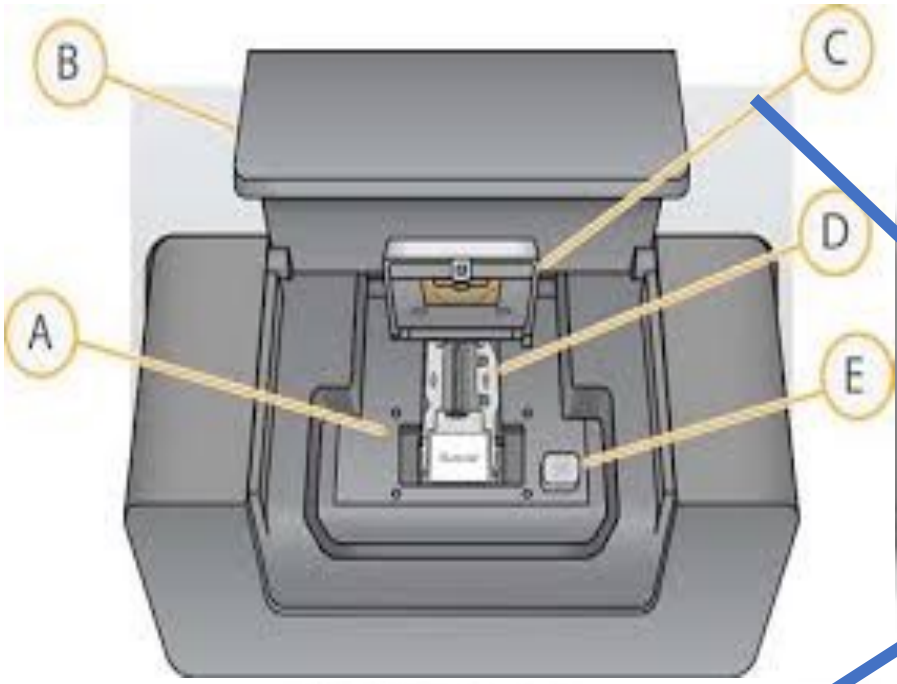
-Colocar en baño de hielo/agua (3/1) 5 min

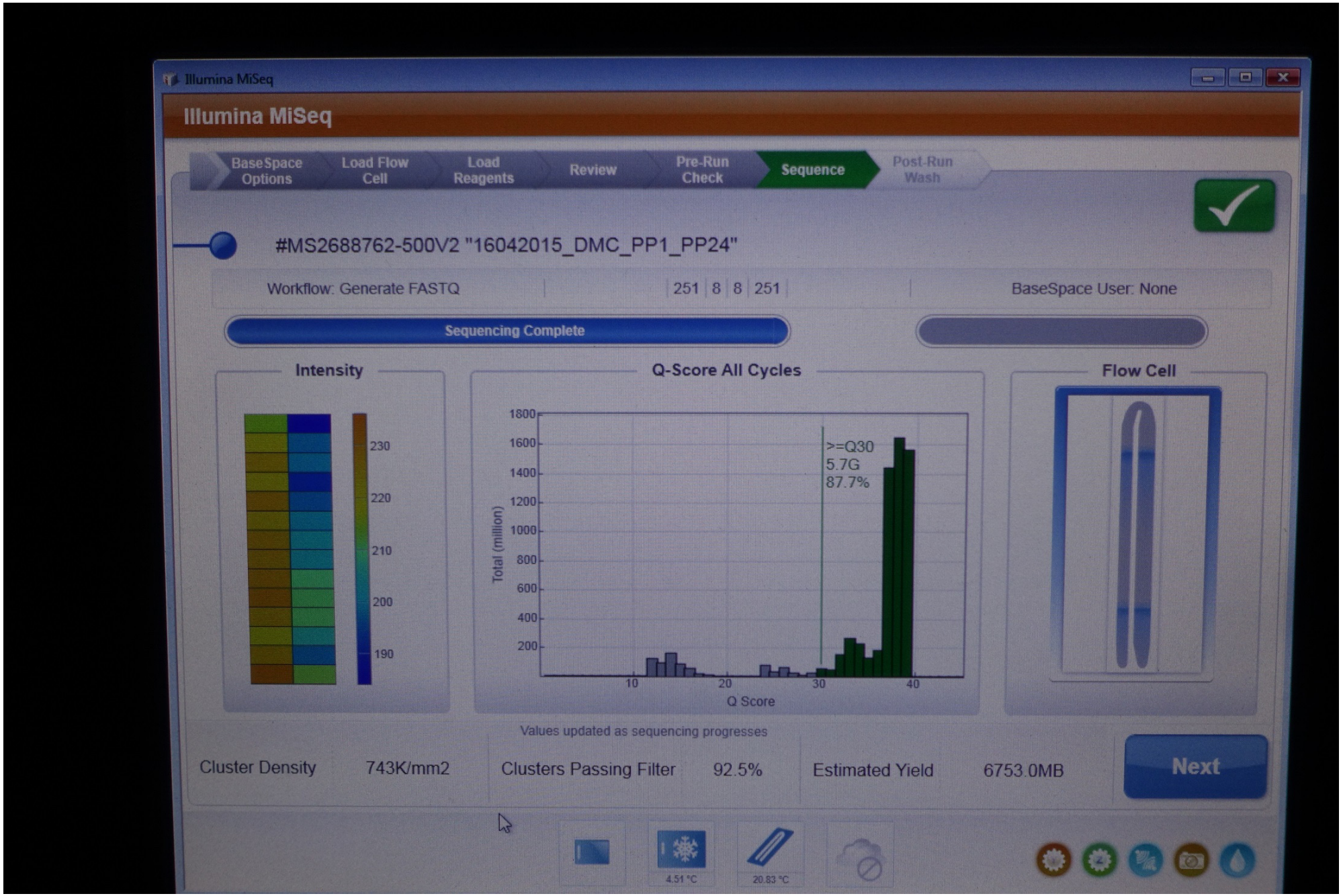
-Cargar 600 ul en el cartucho

# Cargar la muestra en el cartucho

- Antes de cargar el cartucho asegurarse que no hayan burbujas en los compartimentos
- Cargar el cartucho en la celda señalada

# MiSeq Illumina





Illumina MiSeq
Illumina MiSeq
BaseSpace Options
Load Flow Cell
Load Reagents
Review
Pre-Run Check
Sequence
Post-Run Wash
#MS2688762-500V2 "16042015_DMC_PP1_PP24"
Workflow: Generate FASTQ
251 8 8 251
BaseSpace User: None
Sequencing Complete
Intensity
230
220
210
200
190
Q-Score All Cycles
1800
1600
1400
1200
1000
800
600
400
200
Total (million)
>=Q30
5.7G
87.7%
10
20
30
40
Q Score
Flow Cell
Values updated as sequencing progresses
Cluster Density 743K/mm2
Clusters Passing Filter 92.5%
Estimated Yield 6753.0MB
Next
4.51 °C
20.83 °C

# TERCERA GENERACIÓN

ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

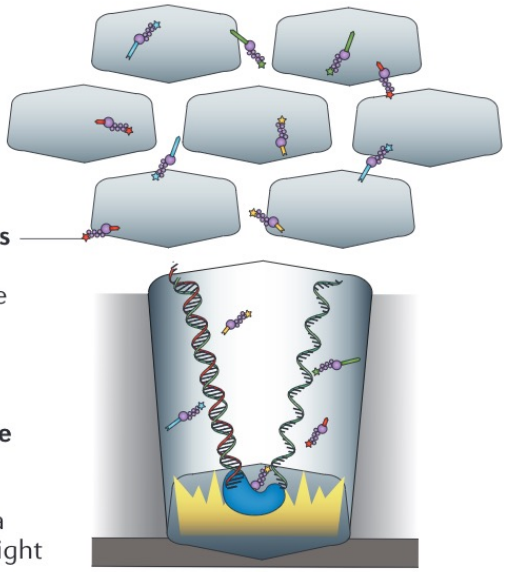
**A** Real-time long-read sequencing

**Aa** Pacific Biosciences

**SMRTbell template**
Two hairpin adapters allow continuous circular sequencing

**ZMW wells**
Sites where sequencing takes place

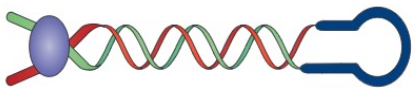
**Labelled nucleotides**
All four dNTPs are labelled and available for incorporation



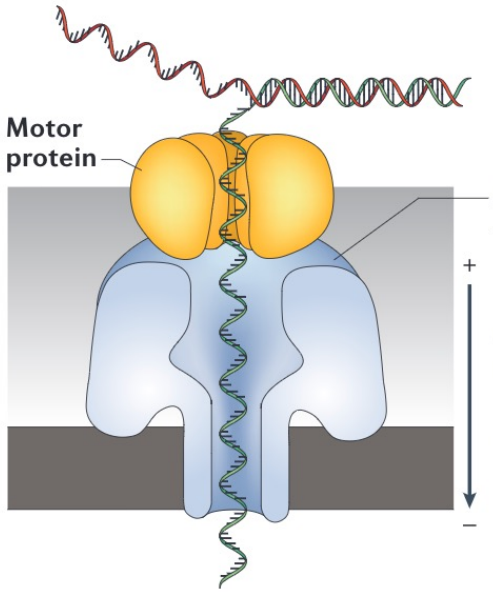
**Modified polymerase**
As a nucleotide is incorporated by the polymerase, a camera records the emitted light

**PacBio output**
A camera records the changing colours from all ZMWs; each colour change corresponds to one base

**Ab** Oxford Nanopore Technologies

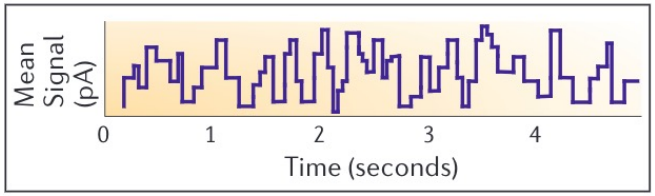


**Leader–Hairpin template**
The leader sequence interacts with the pore and a motor protein to direct DNA, a hairpin allows for bidirectional sequencing




**Alpha-hemolysin**
A large biological pore capable of sensing DNA

**Current**
Passes through the pore and is modulated as DNA passes through




**ONT output (squiggles)**
Each current shift as DNA translocates through the pore corresponds to a particular k-mer

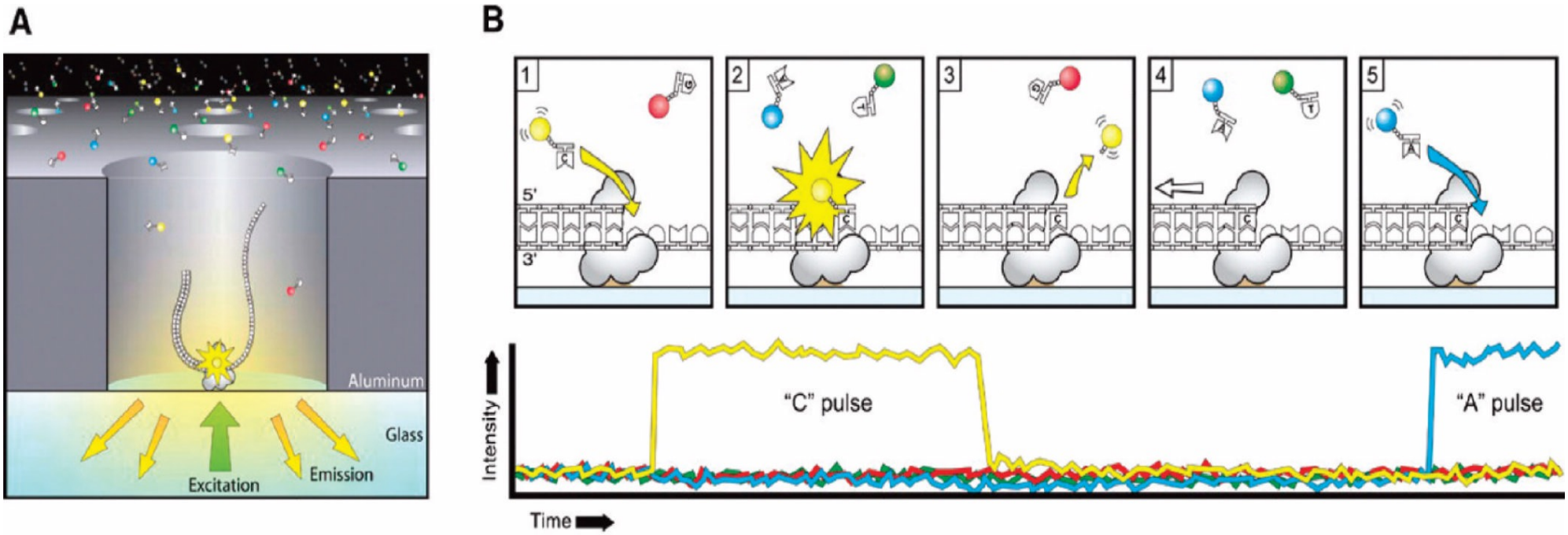
Goodwin, 2016. Nature Reviews

# TERCERA GENERACIÓN

ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

PacBio
SMRT Single-Molecule in Real-Time

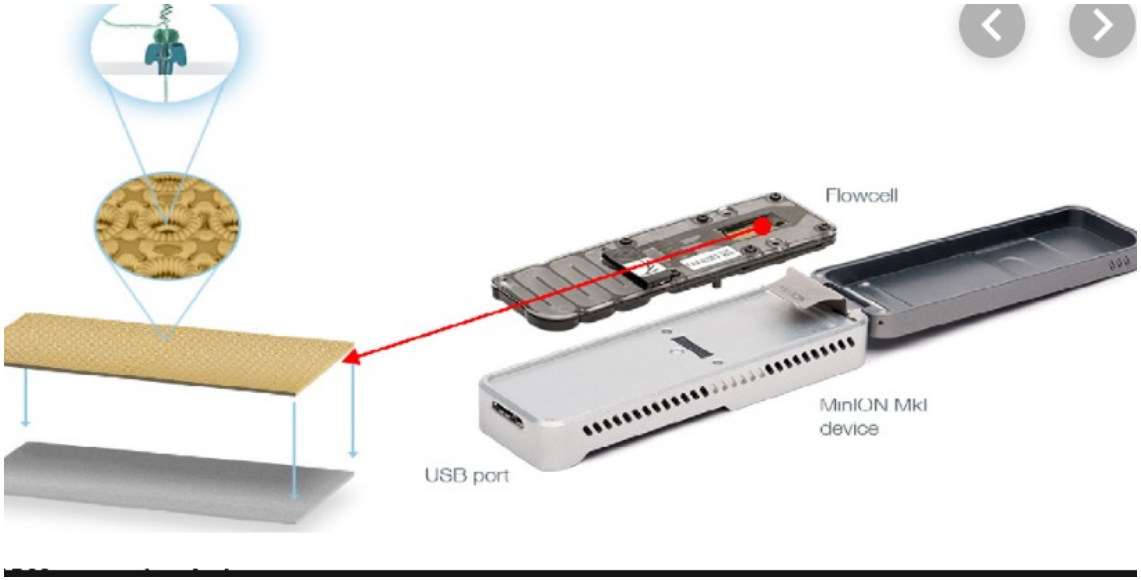
Secuenciación de molécula única en tiempo real



# TERCERA GENERACIÓN

ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

https://nanoporetech.com/how-it-works

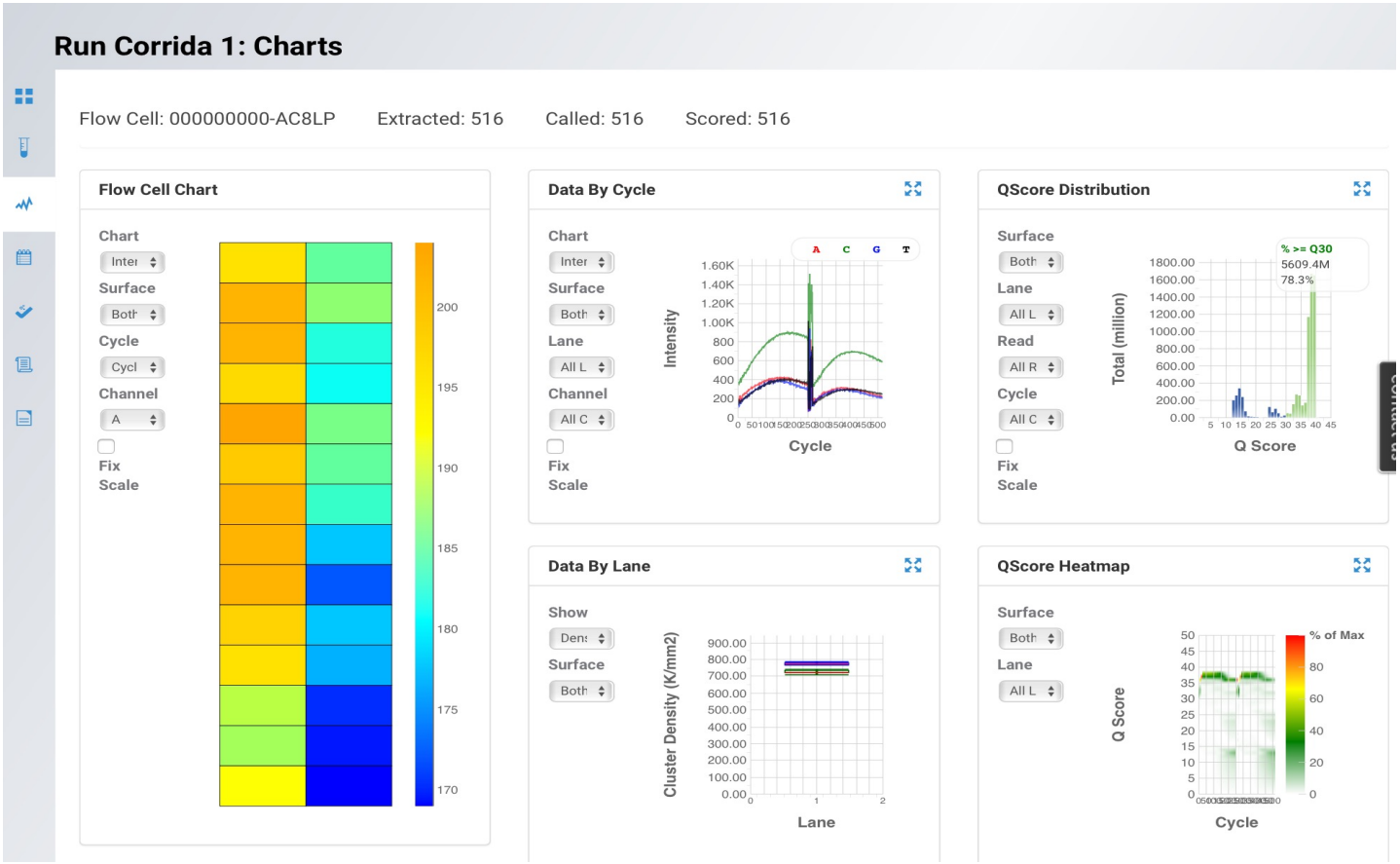


# Conclusiones

- Selección de la Plataforma – Kit cerrados. Depende del tamaño del genoma y de la demanda de muestras.
- Tener en cuenta el objetivo
- Cobertura-tamaño del genoma – MULTIPLEX
- Cantidad de genomas a procesar simultaneamente
- Capacitación y estandarización en análisis e interpretación de los resultados
- Secuenciación Genómica será una metodología ampliamente utilizada en los laboratorios de salud desplazando a los métodos convencionales

# CÓMO OBTENEMOS LOS RESULTADOS: CALIDAD

# CALIDAD DE LA CORRIDA



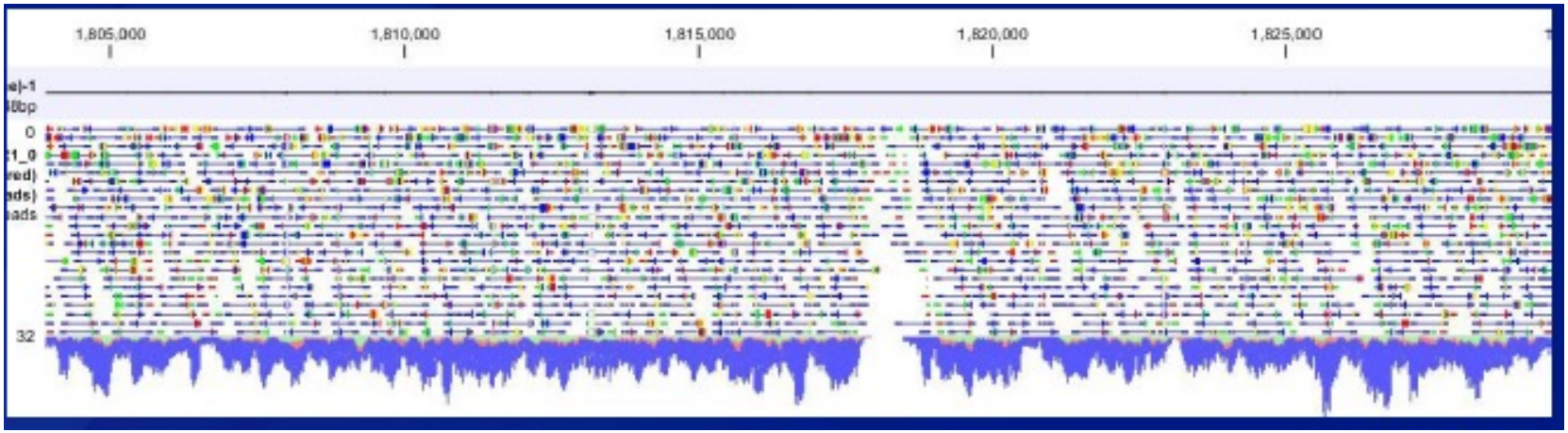
# READS O LECTURAS

ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

- Dato crudo (Raw read)

  Secuencia que sale del equipo, el largo va a depender de la química de la secuenciación

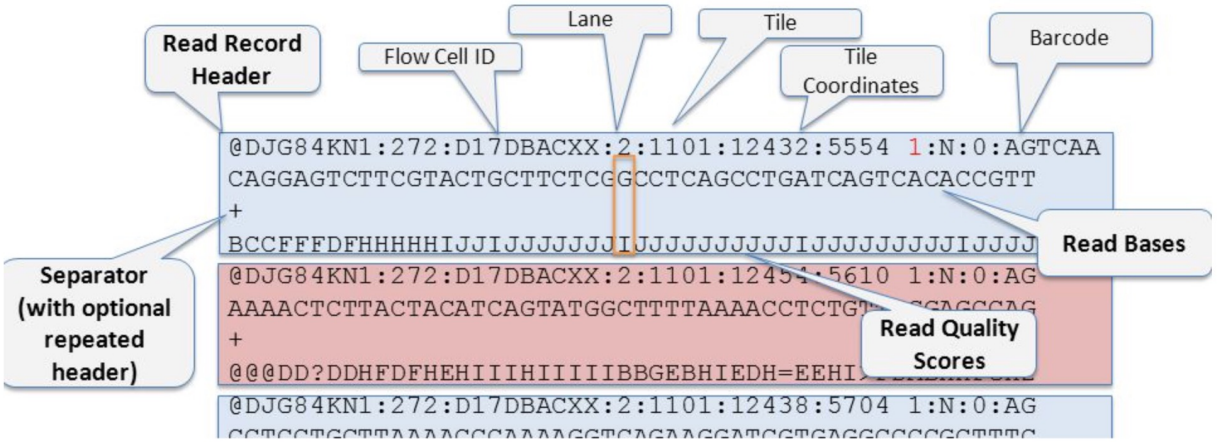
  De 100000 a millones de raw reads se generan por aislamiento



# ARCHIVO: FASTQ

ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

## FASTQ Format (Illumina Example)



# ASCII

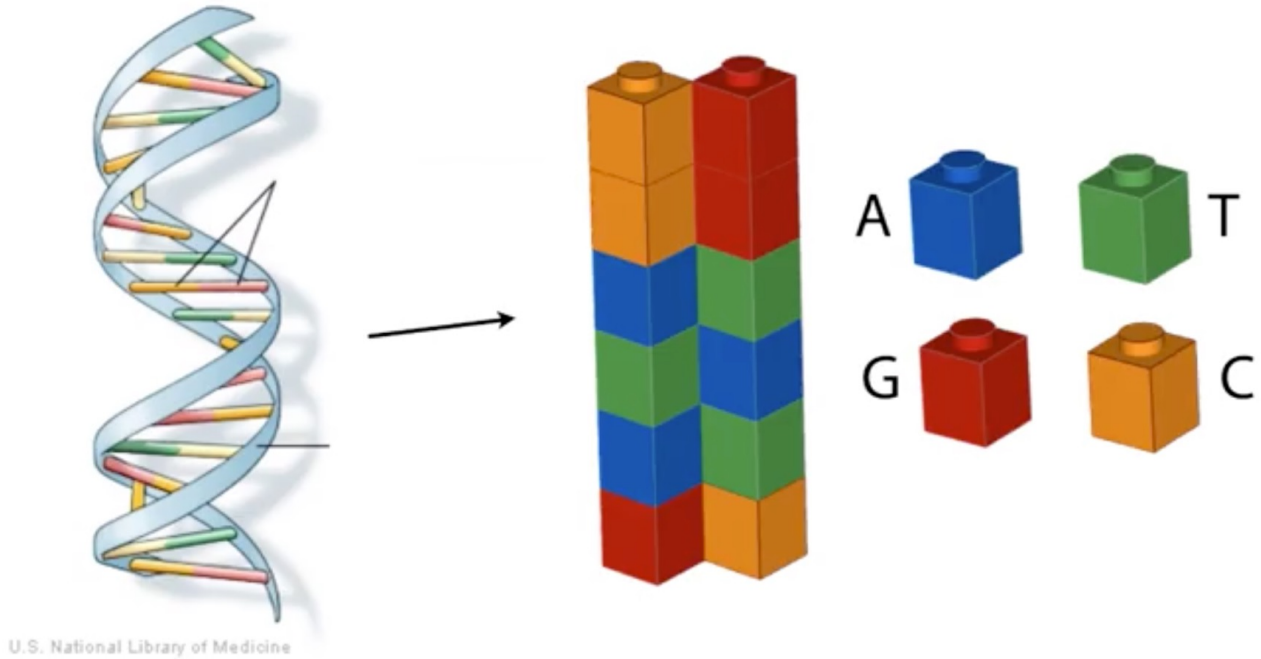
ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

| Dec | Hex | Oct | Chr | Dec | Hex | Oct | HTML | Chr | Dec | Hex | Oct | HTML | Chr | Dec | Hex | Oct | HTML | Chr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 000 | NULL | 32 | 20 | 040 | &#032; | Space | 64 | 40 | 100 | &#064; | @ | 96 | 60 | 140 | &#096; | ` |
| 1 | 1 | 001 | Start of Header | 33 | 21 | 041 | &#033; | ! | 65 | 41 | 101 | &#065; | A | 97 | 61 | 141 | &#097; | a |
| 2 | 2 | 002 | Start of Text | 34 | 22 | 042 | &#034; | " | 66 | 42 | 102 | &#066; | B | 98 | 62 | 142 | &#098; | b |
| 3 | 3 | 003 | End of Text | 35 | 23 | 043 | &#035; | # | 67 | 43 | 103 | &#067; | C | 99 | 63 | 143 | &#099; | c |
| 4 | 4 | 004 | End of Transmission | 36 | 24 | 044 | &#036; | $ | 68 | 44 | 104 | &#068; | D | 100 | 64 | 144 | &#100; | d |
| 5 | 5 | 005 | Enquiry | 37 | 25 | 045 | &#037; | % | 69 | 45 | 105 | &#069; | E | 101 | 65 | 145 | &#101; | e |
| 6 | 6 | 006 | Acknowledgment | 38 | 26 | 046 | &#038; | & | 70 | 46 | 106 | &#070; | F | 102 | 66 | 146 | &#102; | f |
| 7 | 7 | 007 | Bell | 39 | 27 | 047 | &#039; | ' | 71 | 47 | 107 | &#071; | G | 103 | 67 | 147 | &#103; | g |
| 8 | 8 | 010 | Backspace | 40 | 28 | 050 | &#040; | ( | 72 | 48 | 110 | &#072; | H | 104 | 68 | 150 | &#104; | h |
| 9 | 9 | 011 | Horizontal Tab | 41 | 29 | 051 | &#041; | ) | 73 | 49 | 111 | &#073; | I | 105 | 69 | 151 | &#105; | i |
| 10 | A | 012 | Line feed | 42 | 2A | 052 | &#042; | * | 74 | 4A | 112 | &#074; | J | 106 | 6A | 152 | &#106; | j |
| 11 | B | 013 | Vertical Tab | 43 | 2B | 053 | &#043; | + | 75 | 4B | 113 | &#075; | K | 107 | 6B | 153 | &#107; | k |
| 12 | C | 014 | Form feed | 44 | 2C | 054 | &#044; | , | 76 | 4C | 114 | &#076; | L | 108 | 6C | 154 | &#108; | l |
| 13 | D | 015 | Carriage return | 45 | 2D | 055 | &#045; | - | 77 | 4D | 115 | &#077; | M | 109 | 6D | 155 | &#109; | m |
| 14 | E | 016 | Shift Out | 46 | 2E | 056 | &#046; | . | 78 | 4E | 116 | &#078; | N | 110 | 6E | 156 | &#110; | n |
| 15 | F | 017 | Shift In | 47 | 2F | 057 | &#047; | / | 79 | 4F | 117 | &#079; | O | 111 | 6F | 157 | &#111; | o |
| 16 | 10 | 020 | Data Link Escape | 48 | 30 | 060 | &#048; | 0 | 80 | 50 | 120 | &#080; | P | 112 | 70 | 160 | &#112; | p |
| 17 | 11 | 021 | Device Control 1 | 49 | 31 | 061 | &#049; | 1 | 81 | 51 | 121 | &#081; | Q | 113 | 71 | 161 | &#113; | q |
| 18 | 12 | 022 | Device Control 2 | 50 | 32 | 062 | &#050; | 2 | 82 | 52 | 122 | &#082; | R | 114 | 72 | 162 | &#114; | r |
| 19 | 13 | 023 | Device Control 3 | 51 | 33 | 063 | &#051; | 3 | 83 | 53 | 123 | &#083; | S | 115 | 73 | 163 | &#115; | s |
| 20 | 14 | 024 | Device Control 4 | 52 | 34 | 064 | &#052; | 4 | 84 | 54 | 124 | &#084; | T | 116 | 74 | 164 | &#116; | t |
| 21 | 15 | 025 | Negative Ack. | 53 | 35 | 065 | &#053; | 5 | 85 | 55 | 125 | &#085; | U | 117 | 75 | 165 | &#117; | u |
| 22 | 16 | 026 | Synchronous idle | 54 | 36 | 066 | &#054; | 6 | 86 | 56 | 126 | &#086; | V | 118 | 76 | 166 | &#118; | v |
| 23 | 17 | 027 | End of Trans. Block | 55 | 37 | 067 | &#055; | 7 | 87 | 57 | 127 | &#087; | W | 119 | 77 | 167 | &#119; | w |
| 24 | 18 | 030 | Cancel | 56 | 38 | 070 | &#056; | 8 | 88 | 58 | 130 | &#088; | X | 120 | 78 | 170 | &#120; | x |
| 25 | 19 | 031 | End of Medium | 57 | 39 | 071 | &#057; | 9 | 89 | 59 | 131 | &#089; | Y | 121 | 79 | 171 | &#121; | y |
| 26 | 1A | 032 | Substitute | 58 | 3A | 072 | &#058; | : | 90 | 5A | 132 | &#090; | Z | 122 | 7A | 172 | &#122; | z |
| 27 | 1B | 033 | Escape | 59 | 3B | 073 | &#059; | ; | 91 | 5B | 133 | &#091; | [ | 123 | 7B | 173 | &#123; | { |
| 28 | 1C | 034 | File Separator | 60 | 3C | 074 | &#060; | < | 92 | 5C | 134 | &#092; | \ | 124 | 7C | 174 | &#124; | \| |
| 29 | 1D | 035 | Group Separator | 61 | 3D | 075 | &#061; | = | 93 | 5D | 135 | &#093; | ] | 125 | 7D | 175 | &#125; | } |
| 30 | 1E | 036 | Record Separator | 62 | 3E | 076 | &#062; | > | 94 | 5E | 136 | &#094; | ^ | 126 | 7E | 176 | &#126; | ~ |
| 31 | 1F | 037 | Unit Separator | 63 | 3F | 077 | &#063; | ? | 95 | 5F | 137 | &#095; | _ | 127 | 7F | 177 | &#127; | Del |

asciicharstable.com

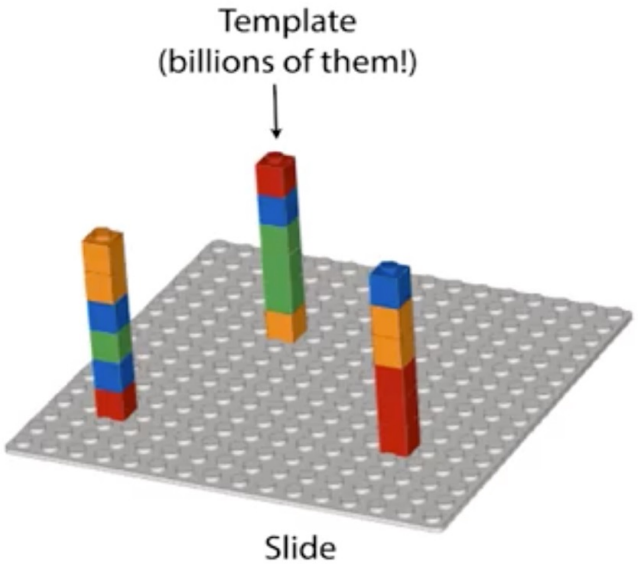
# A CONSTRUIR…

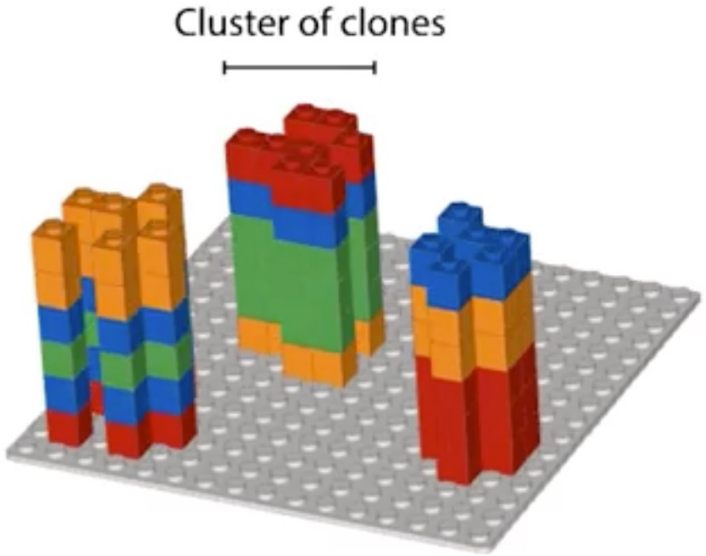
ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

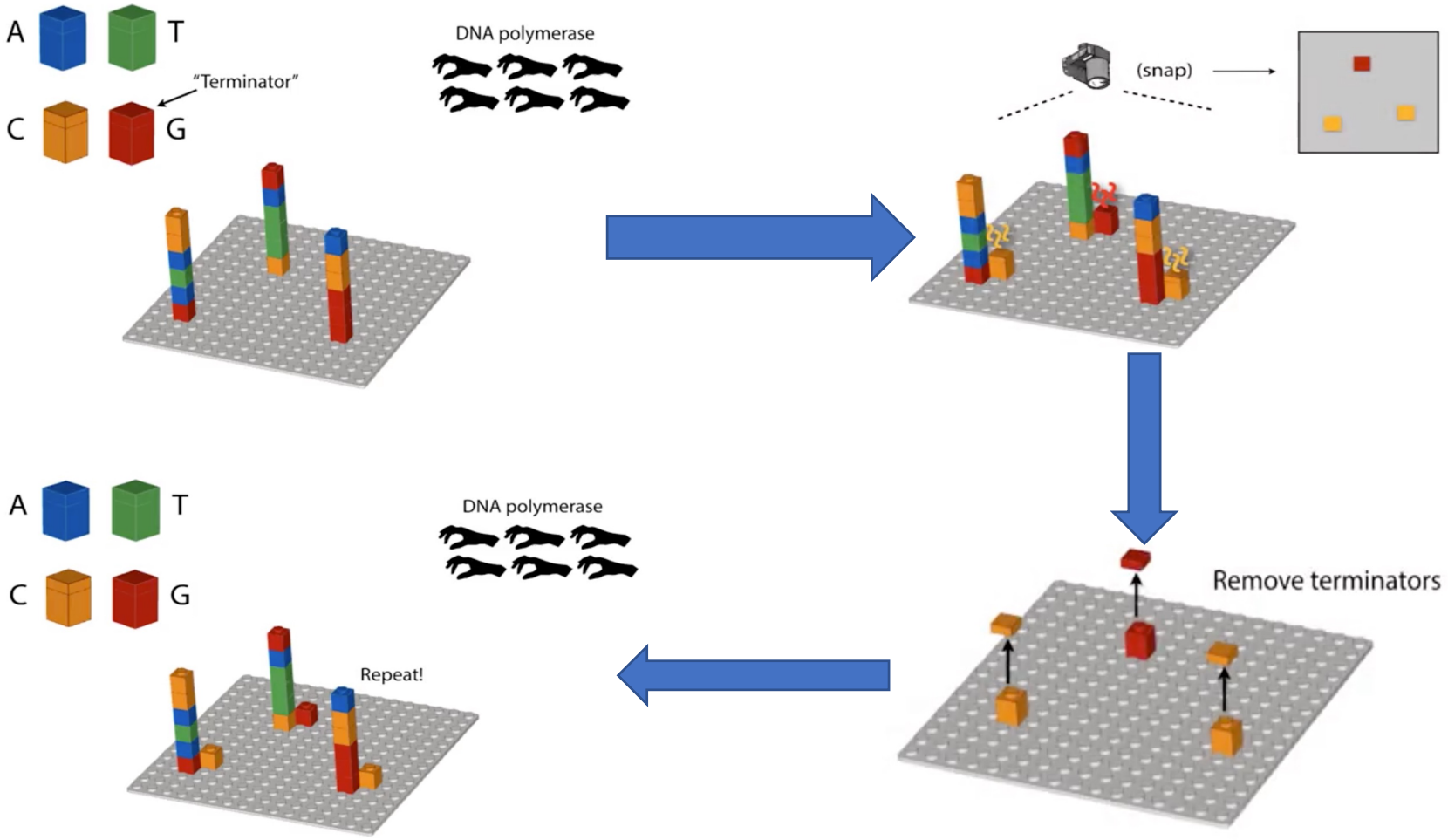


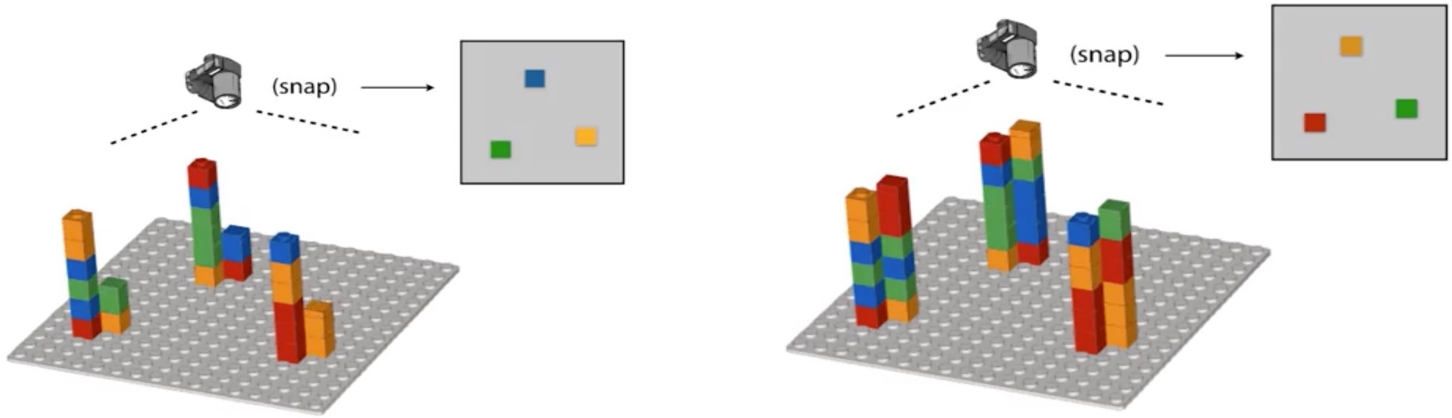

Double stranded DNA (double helix)

Double stranded DNA (lego version)

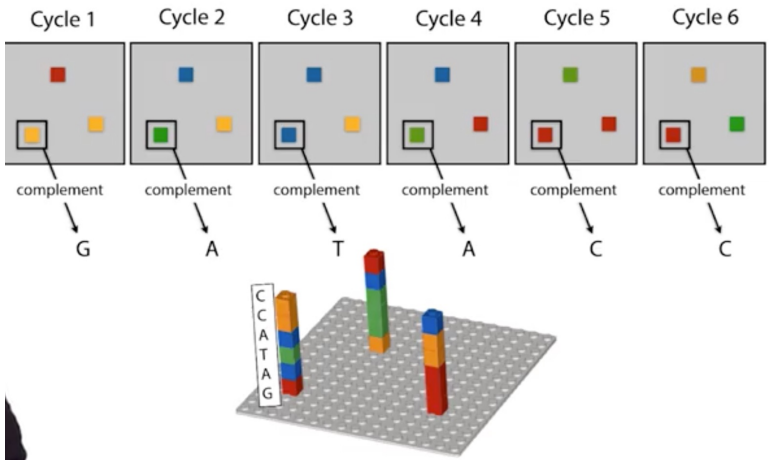

Template
(billions of them!)
Slide


Cluster of clones

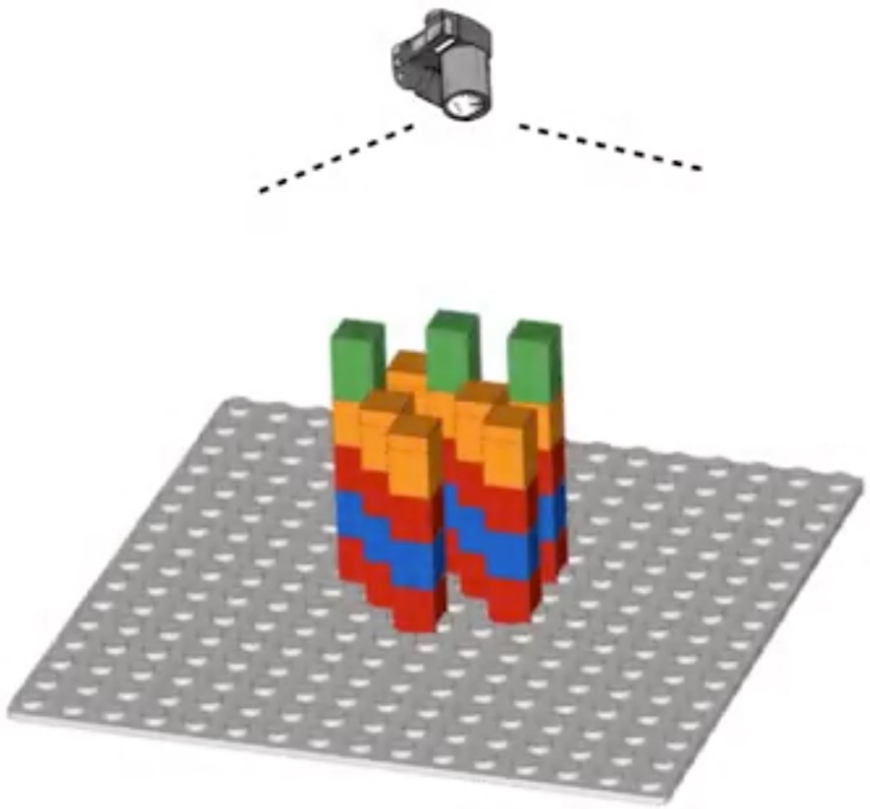

A
T
C
G
"Terminator"
DNA polymerase
(snap)
Remove terminators
DNA polymerase
Repeat!




## Sequencing by synthesis



ANLIS
MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS
E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"



Call: orange (C)

Estimate $p$, probability incorrect:

non-orange light / total light

$p$ = 3 green / 9 total = 1/3

$Q = -10 \log_{10} 1/3 = 4.77$

ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

$$q = -10 \times log_{10}(p)$$

**Donde:**

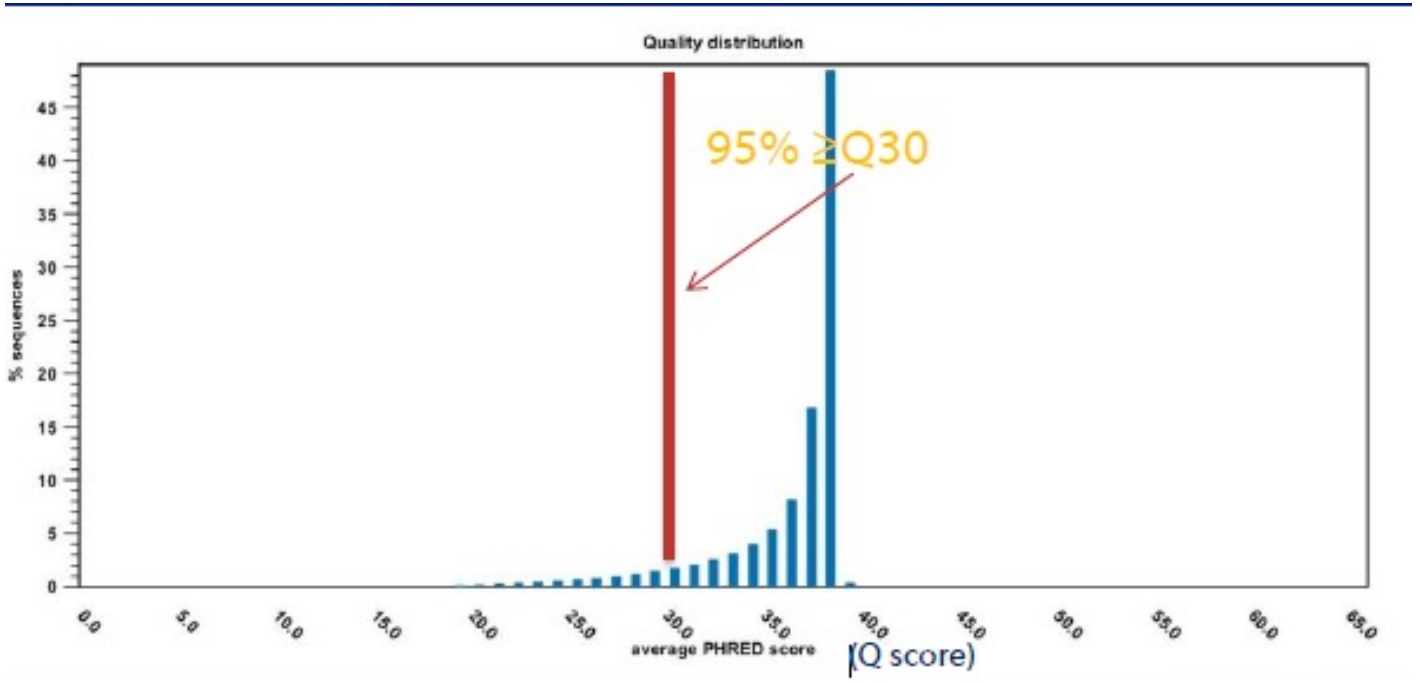
- **q** = quality value
- **p** = estimated probability error for a base call

**Ejemplos:**

- **q** = 20 significa **p** = $10^{-2}$ (1 error cada 100 bases)
- **q** = 30 significa **p** = $10^{-3}$ (1 error cada 1000 bases)
- **q** = 40 significa **p** = $10^{-4}$ (1 error cada 10000 bases)

# CALIDAD

- Q30: Posisbilidad de equivocarse en el Base calling 1:1000



ANLIS MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"

# FASTQC

FastQC es una herramienta para evaluar la calidad (Babraham Bioinformatics) para secuenciación de nueva generación. Da de manera gráfica el set de análisis para un rápido control de calidad de las secuencias crudas (R1 or R2) .

Las principals funciones:

- Importar los datos de BAM, SAM o FastQ de R1 o R2. Las métricas de R1 y R2 se ven al mismo tiempo
- Una rápida visualización que permite identificar problemas
- Provee gráficos y tablas que resumen la información rápidamente

# CALIDAD FASTQC

- FASTQC

**FastQC Report** — Wed 25 Mar 2015 — good_sequence_short.txt

**Summary**

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

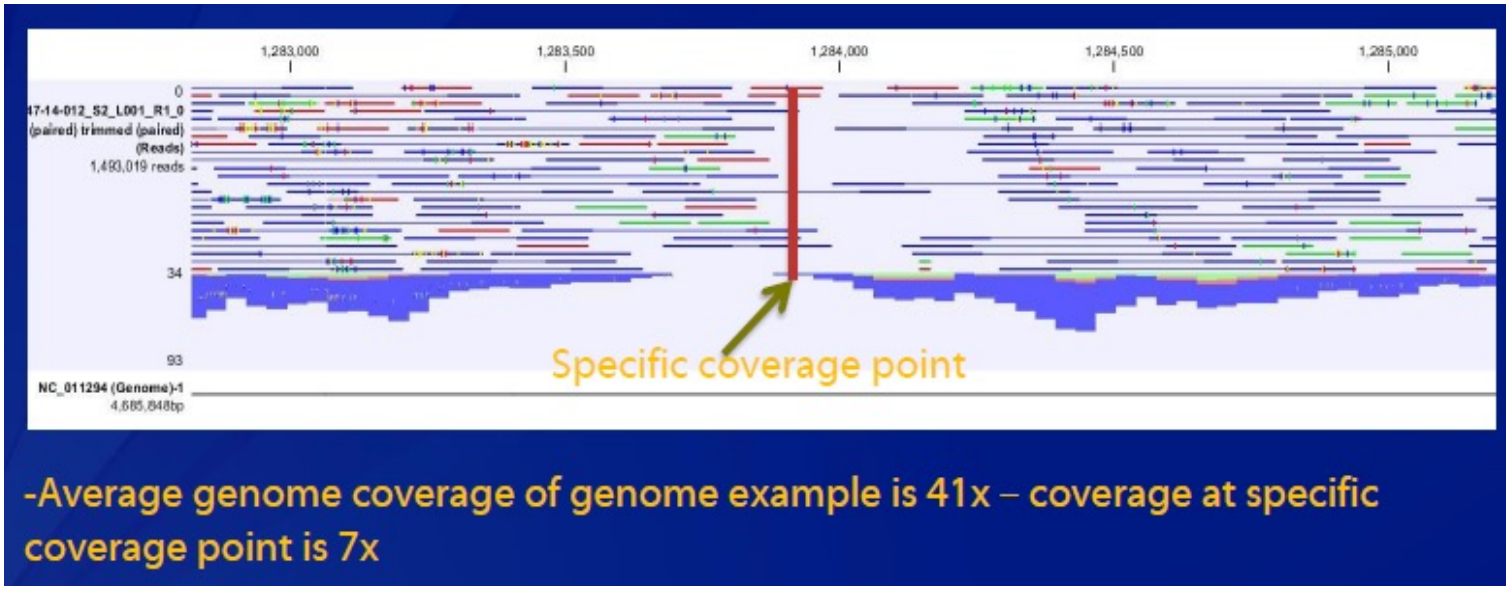
**Basic Statistics**

| Measure | Value |
|---|---|
| Filename | good_sequence_short.txt |
| File type | Conventional base calls |
| Encoding | Illumina 1.5 |
| Total Sequences | 250000 |
| Sequences flagged as poor quality | 0 |
| Sequence length | 40 |
| %GC | 45 |

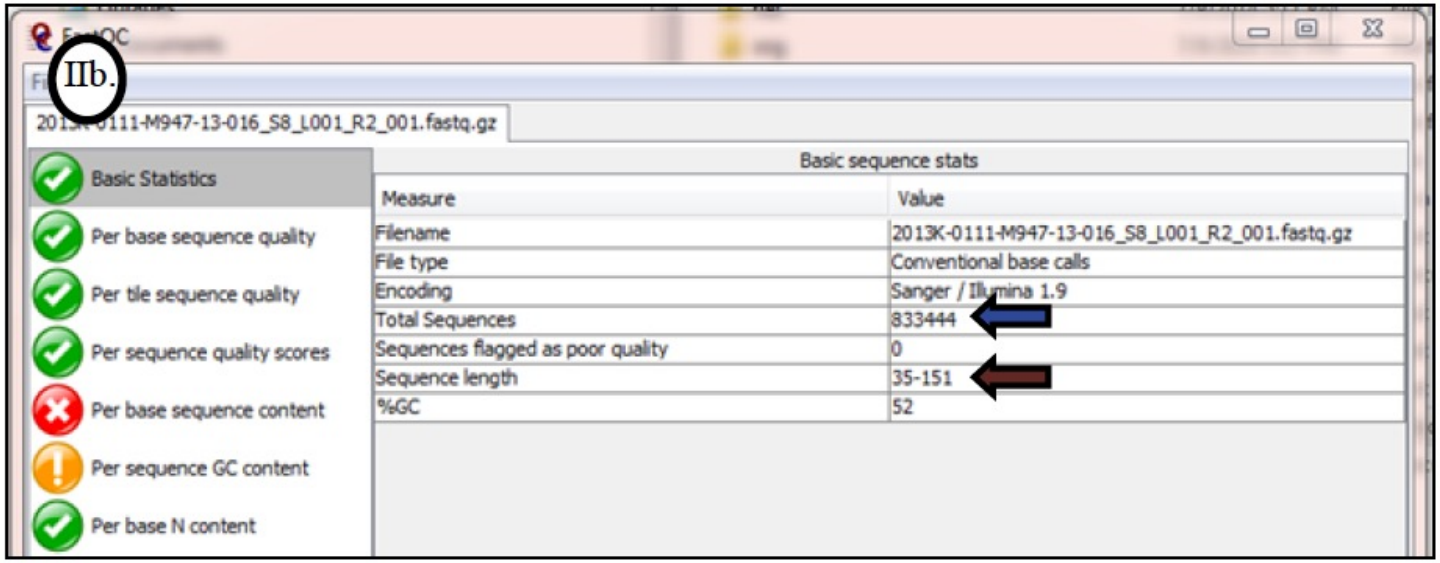
**Per base sequence quality**

# CALIDAD: COBERTURA

- Promedio: divide el total de # de bases por el tamaño del genoma (ej: 156.000.000 número total de bases del secuenciador/ 3.000.000 tamaño del genoma= 52x)
- Específica para una lectura



# CALIDAD: COBERTURA




$$\frac{\text{Max Read length} * \text{Total Sequences} * \text{Sequencing Chemistry (1 for single end, 2 for paired end)}}{\text{Genome size (bp)}} = \text{Coverage}$$

**Example:**

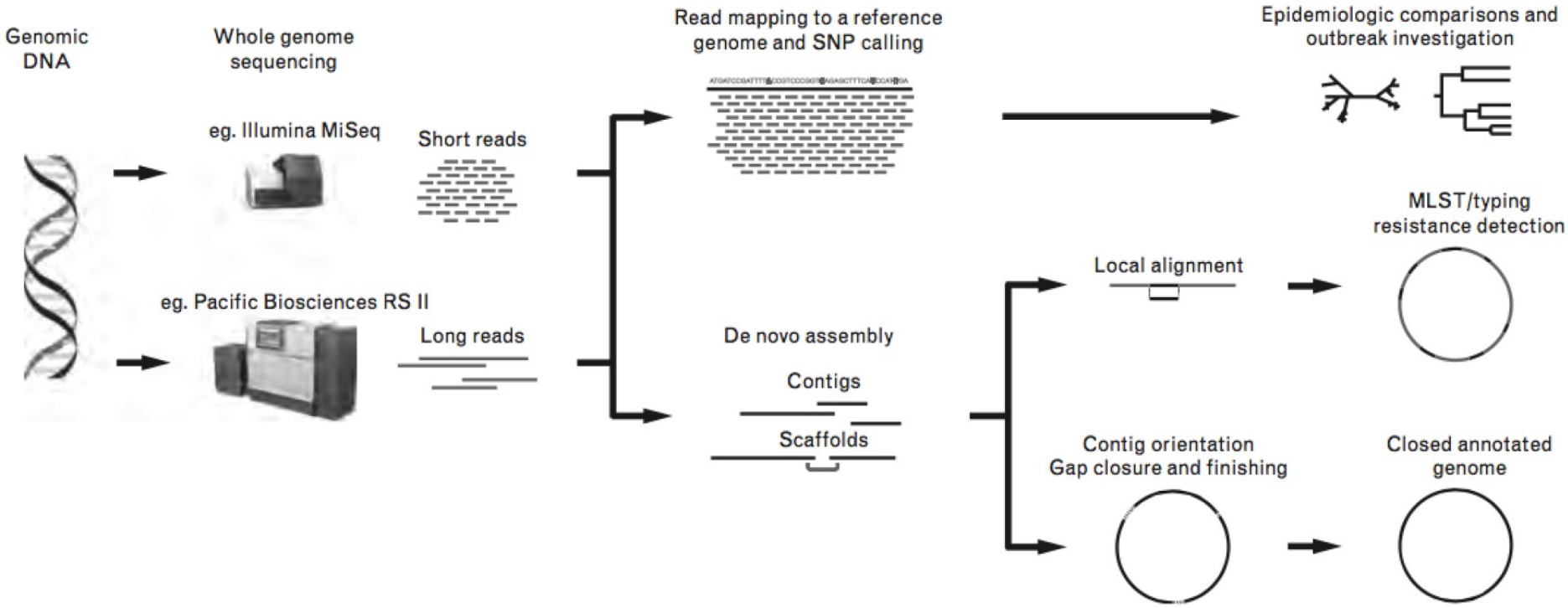
$$\frac{150 * 833{,}444 * 2}{5{,}000{,}000 \text{ (\# bp of Salmonella)}} = 50.01\text{x}$$

# CALIDAD COBERTURA

| | Organism | | | |
|---|---|---|---|---|
| | *Listeria* | *E. coli & Shigella* | *Salmonella* | *Campylobacter* |
| **Genome Size (bp)** | 3,000,000 | 5,000,000 | 5,000,000 | 1,600,000 |
| **Target Coverage** | ≥20X | ≥40X | ≥30X | ≥20X |

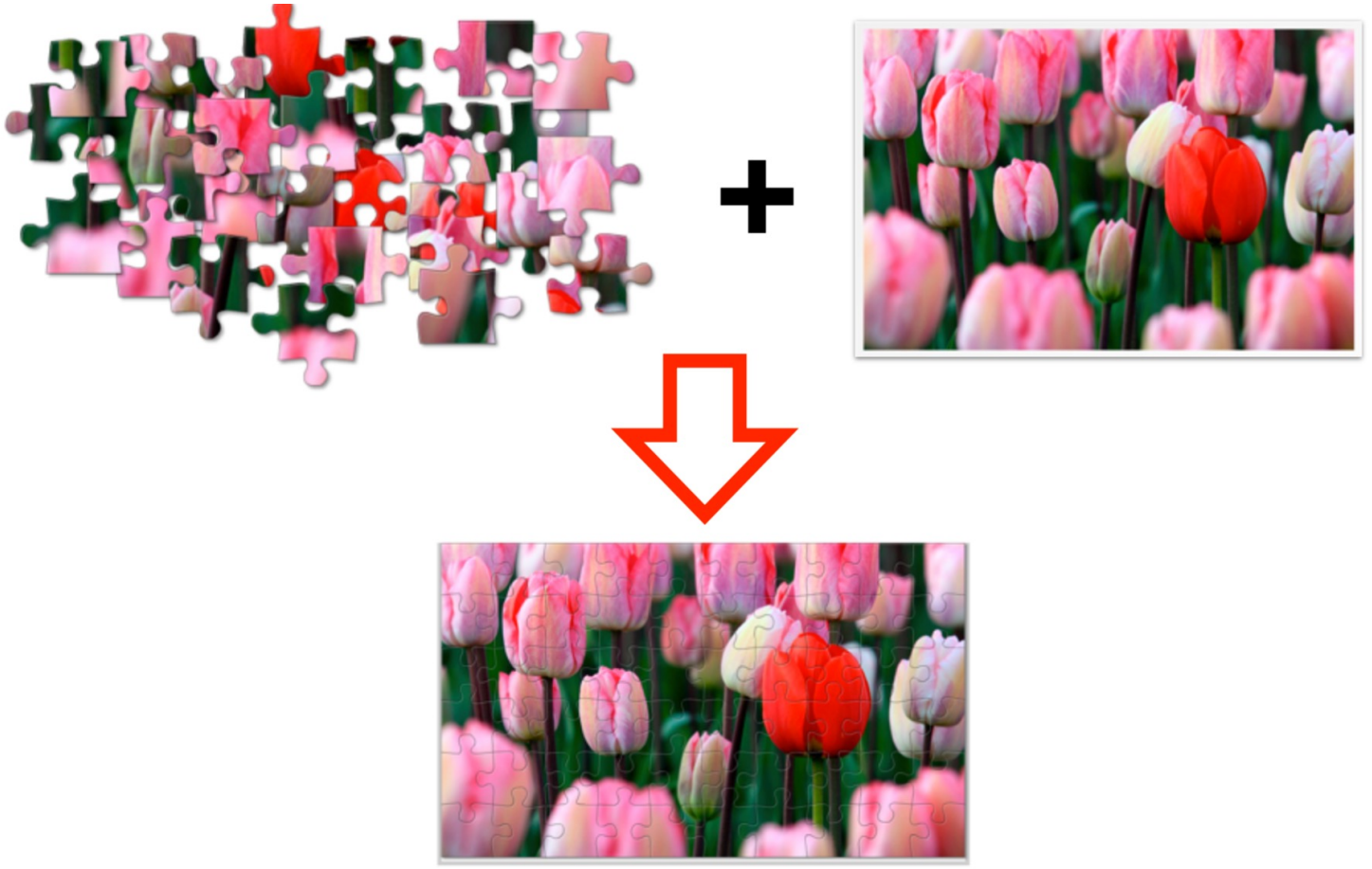
# CÓMO OBTENEMOS LOS RESULTADOS: ESTRATEGIAS DE ANÁLISIS

# ANÁLISIS

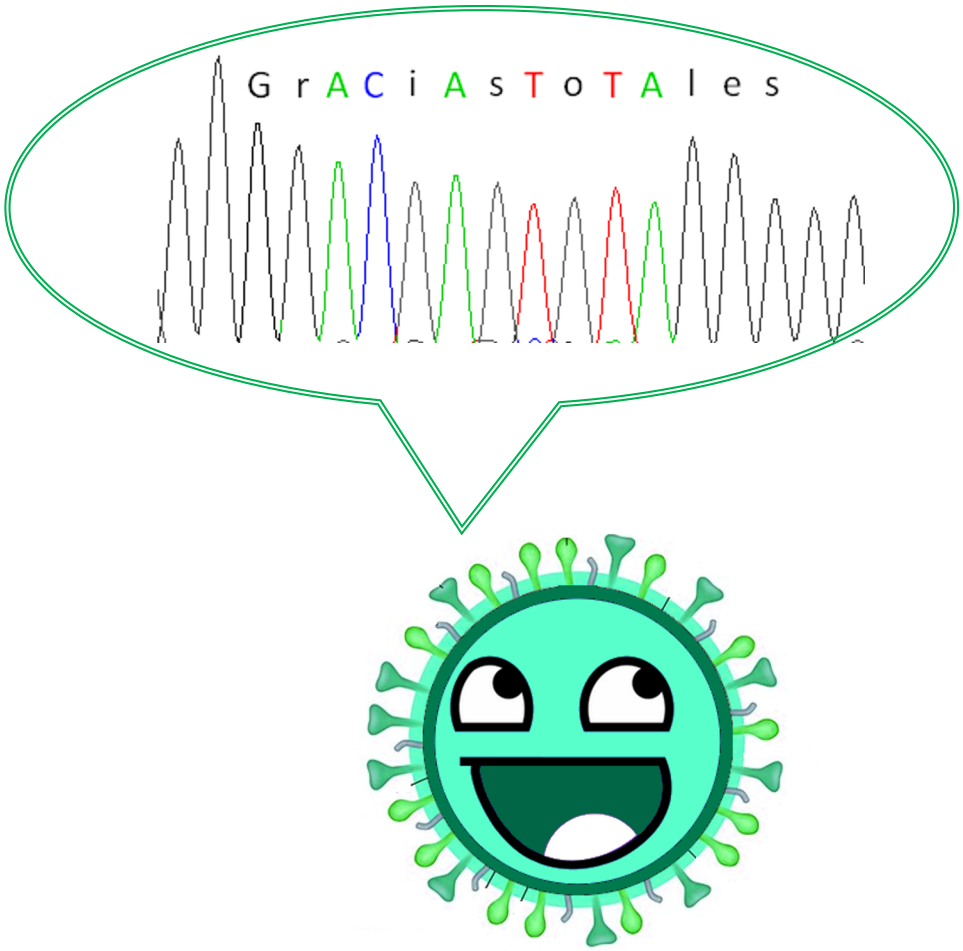



Kwong *et al.*, 2015 Pathology

ANLIS
MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS
E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"



ANLIS
MALBRÁN
ADMINISTRACIÓN NACIONAL DE LABORATORIOS
E INSTITUTOS DE SALUD "DR. CARLOS G. MALBRÁN"


GrACiAsToTAles