

# Métodos de secuenciación masiva

Biocomputación  
Grado en Bioquímica

# Contenido

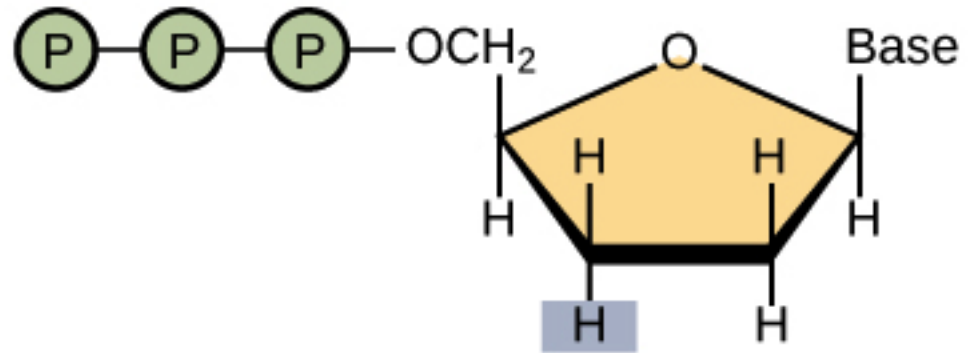
- Principios básicos del método de Sanger
- Secuenciación masiva: *sequencing by synthesis* (Illumina)
- Aplicaciones de la secuenciación masiva
- La salida: el fichero fastq
- Control de calidad:
  - Detección del adaptador
  - Recortar bases con baja calidad

# Método de Sanger

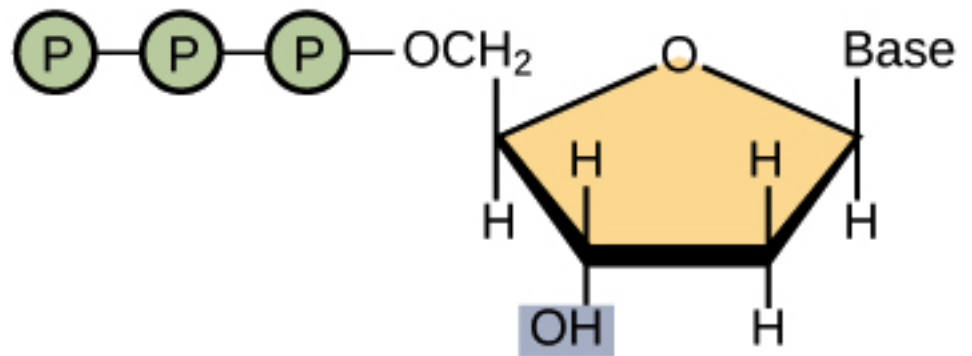
- Este método se basa en que cuando la polimerasa incorpora un dideoxiribonucleótido se para la síntesis del ADN
- No se pueden formar enlaces fosfodiéster por la ausencia del grupo hidroxilo

<https://www.khanacademy.org/science/biology/biotech-dna-technology/dna-sequencing-pcr-electrophoresis/a/dna-sequencing>

Image credit: "Whole-genome sequencing: Figure 1," by OpenStax College, Biology (CC BY 4.0).



**Dideoxynucleotide (ddNTP)**



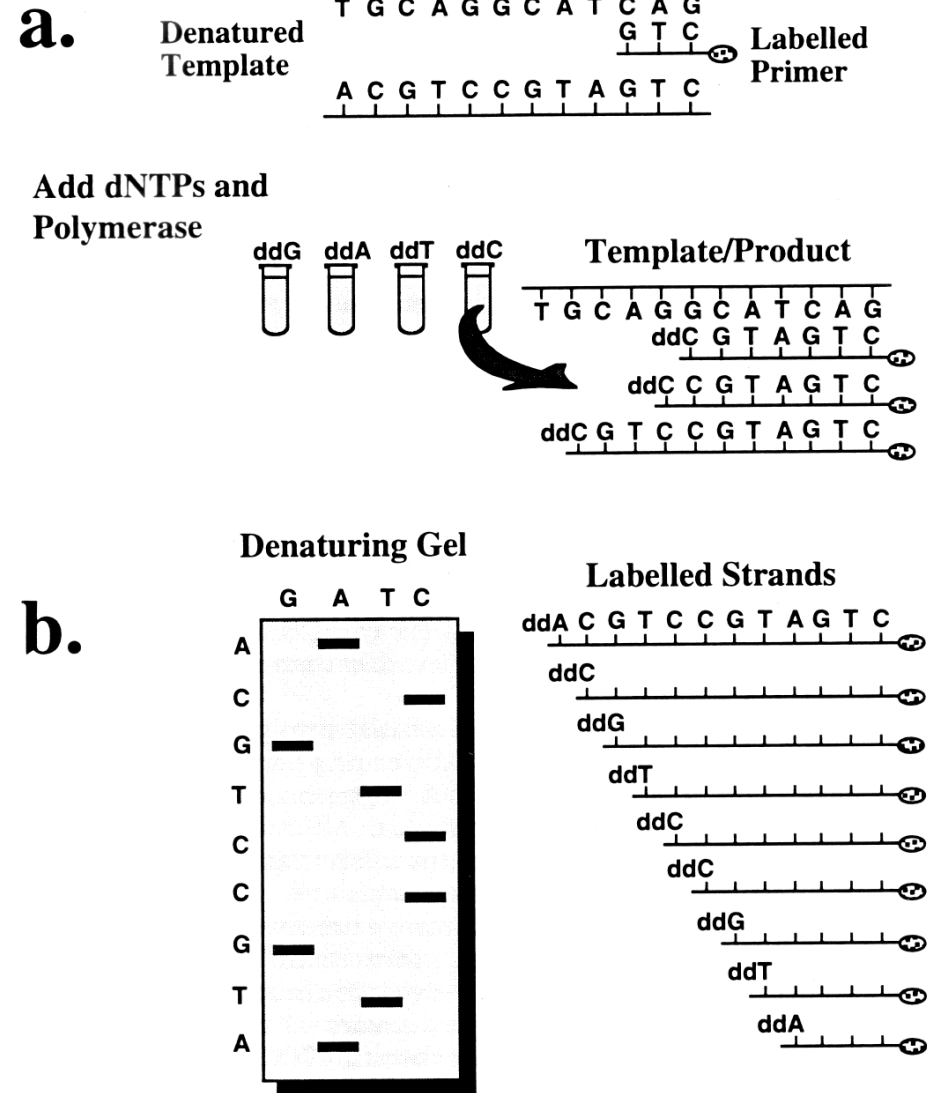
**Deoxynucleotide (dNTP)**

# Método de Sanger

- Molde de DNA desnaturalizado (monocatenario)
- Polimerasa de DNA
- Cebador
- Nucleótidos
- Didesoxinucleótidos
- (didesoxinucleótido trifosfato)

## Definición:

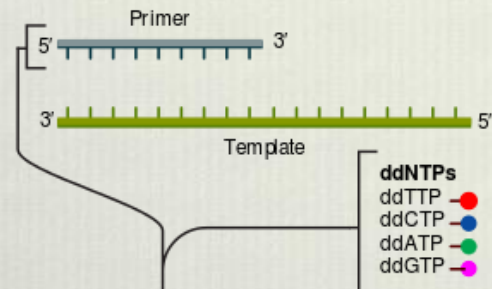
Con secuenciación masiva nos referimos a los métodos que no se basan en el método de Sanger



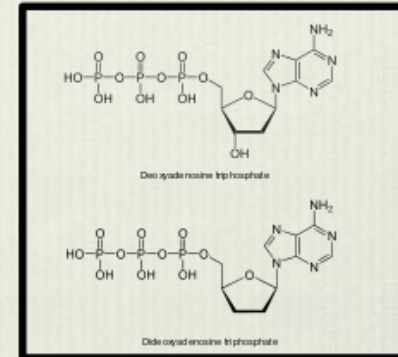
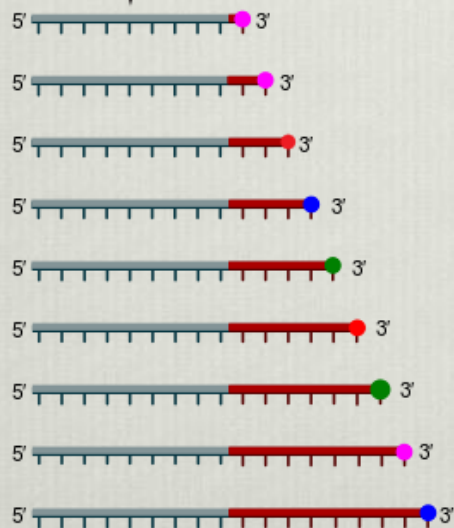
# Método de Sanger

## ① Reaction mixture

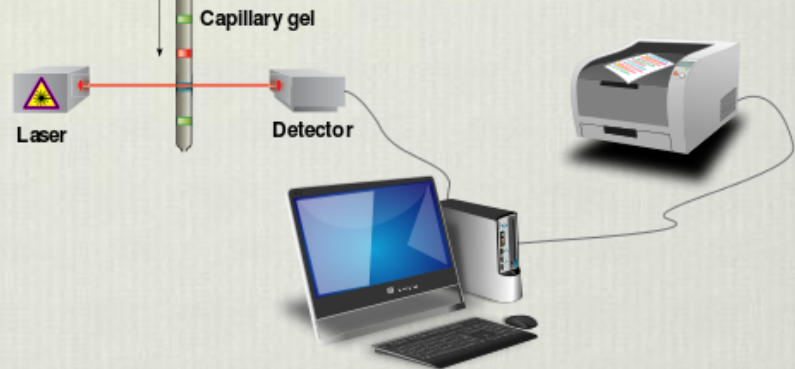
- ▶ Primer and DNA template
- ▶ DNA polymerase
- ▶ ddNTPs with flouorchromes
- ▶ dNTPs (dATP, dCTP, dGTP, and dTTP)



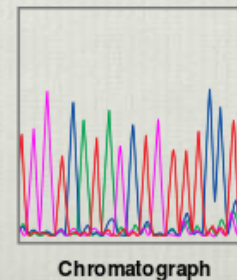
## ② Primer elongation and chain termination



## ③ Capillary gel electrophoresis separation of DNA fragments



## ④ Laser detection of flouorchromes and computational sequence analysis



# Comparación Sanger vs. Illumina

Podemos distinguir 3 pasos

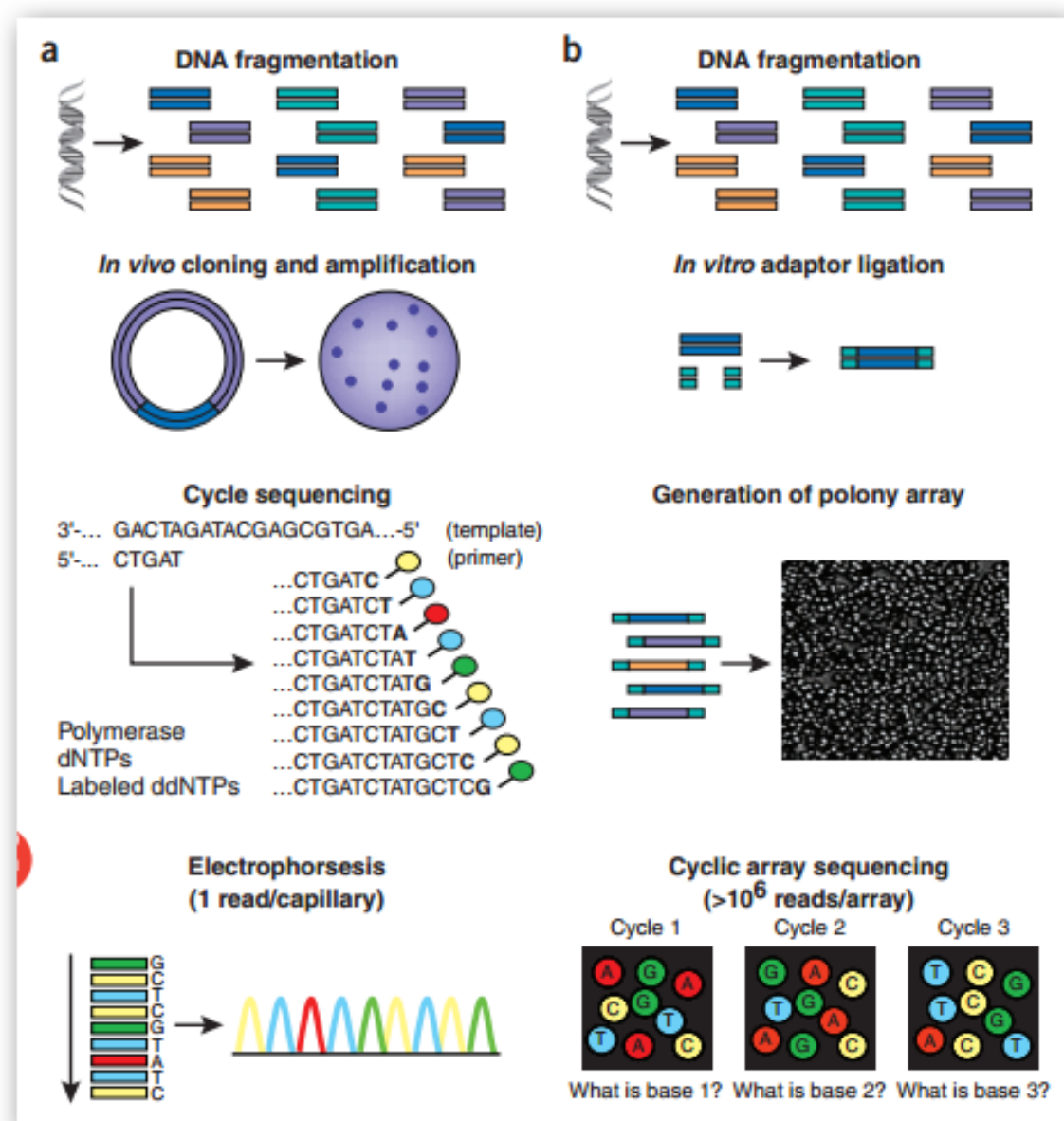
## *Preparación de la muestra:*

- Fragmentar el ADN
- Ligar los adaptadores
- Seleccionar la longitud

*¡Este paso es distinto para los diferentes protocolos!*

## *Generación de clusters y amplificación*

## *Secuenciación (sequencing by synthesis)*



Review

## Next-generation DNA sequencing

Jay Shendure & Hanlee Ji

Nature Biotechnology 26, 1135–1145 (2008)

doi:10.1038/nbt1486

[Download Citation](#)

Published online: 09 October 2008

# Clusters

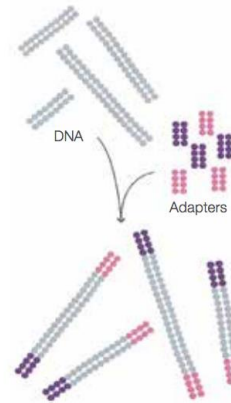
- El material de entrada es ADN de doble cadena ligado a unos adaptadores
- Los adaptadores son complementarios a oligos fijados en la superficie del flowcell
- Se desnaturaliza el ADN y este se une al *flowcell* formando puentes mediante la hibridación con los oligos

## Figuras de Illumina:

[https://www.illumina.com/documents/products/techspotlights/techspotlight\\_sequencing.pdf](https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf)

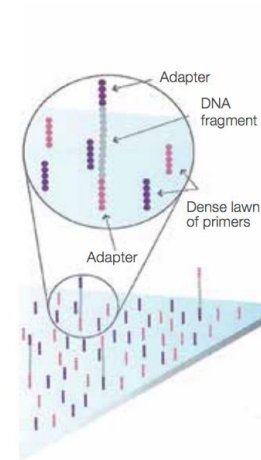
<https://www.youtube.com/watch?v=womKfikW1xM>

Figure 2: Prepare Genomic DNA Sample



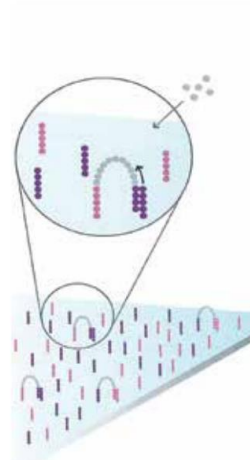
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Figure 3: Attach DNA to Surface



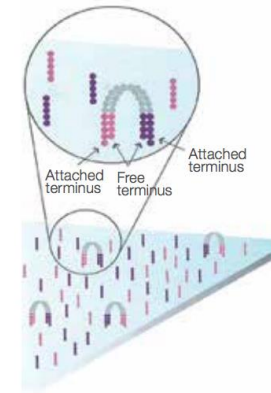
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

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.

# Método de Illumina

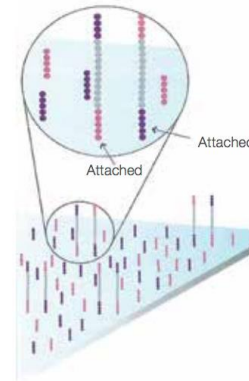
- Desnaturalización para generar un molde de cadena sencilla
- Este proceso se repite hasta obtener un cluster con la secuencia original amplificado
- Añadir los nucleótidos marcados con fluorocromo y con un terminador en 3', polimerasa y cebadores
- Tomar la imagen de todo el flowcell
- Eliminar el terminador y el fluorocromo y iniciar el próximo ciclo

## Figuras de Illumina:

[https://www.illumina.com/documents/products/techspotlights/techspotlight\\_sequencing.pdf](https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf)

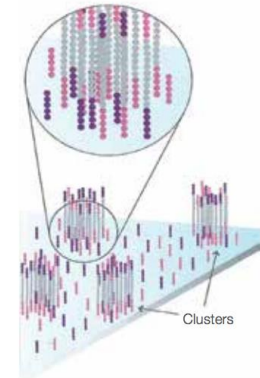
<https://www.youtube.com/watch?v=womKfikW1xM>

Figure 6: Denature the Double-Standed Molecules



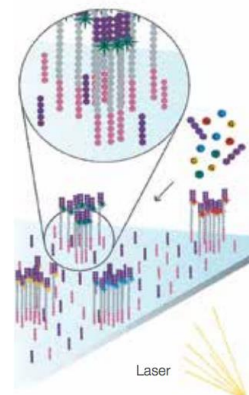
Denaturation leaves single-stranded templates anchored to the substrate.

Figure 7: Complete Amplification



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

Figure 8: Determine First Base



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

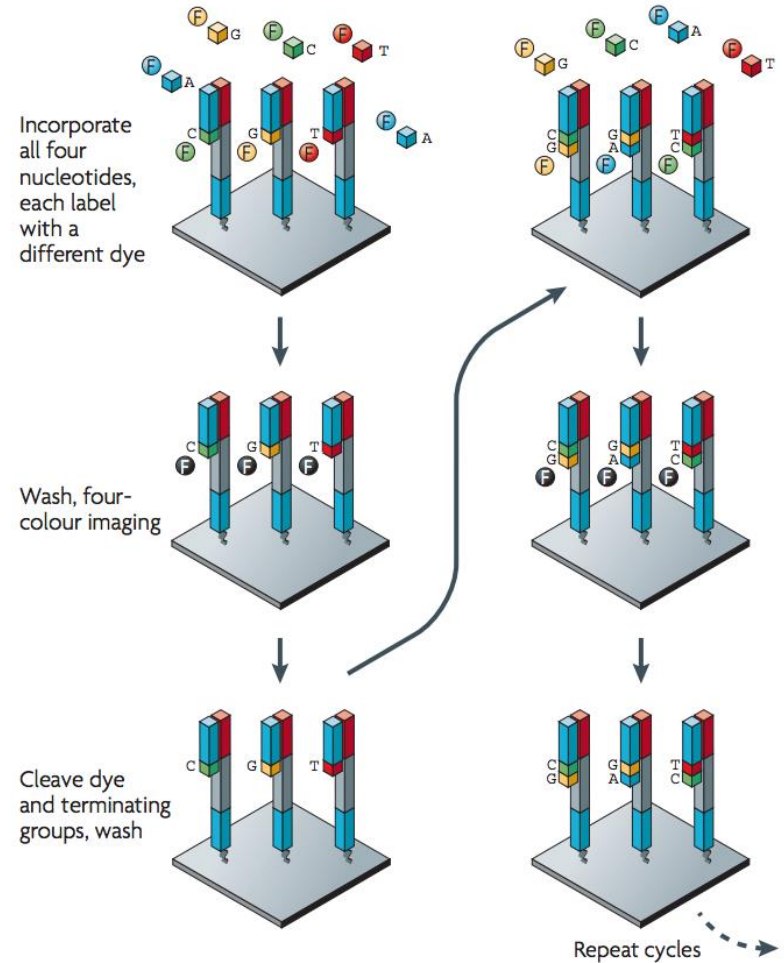
Figure 9: Image First Base



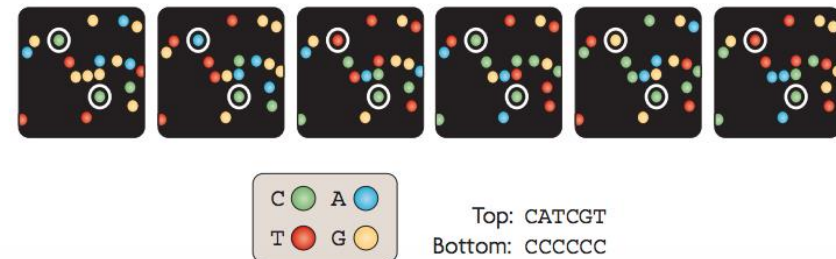
After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

# Método de Illumina

## a Illumina/Solexa — Reversible terminators



## b



## Sequencing technologies — the next generation

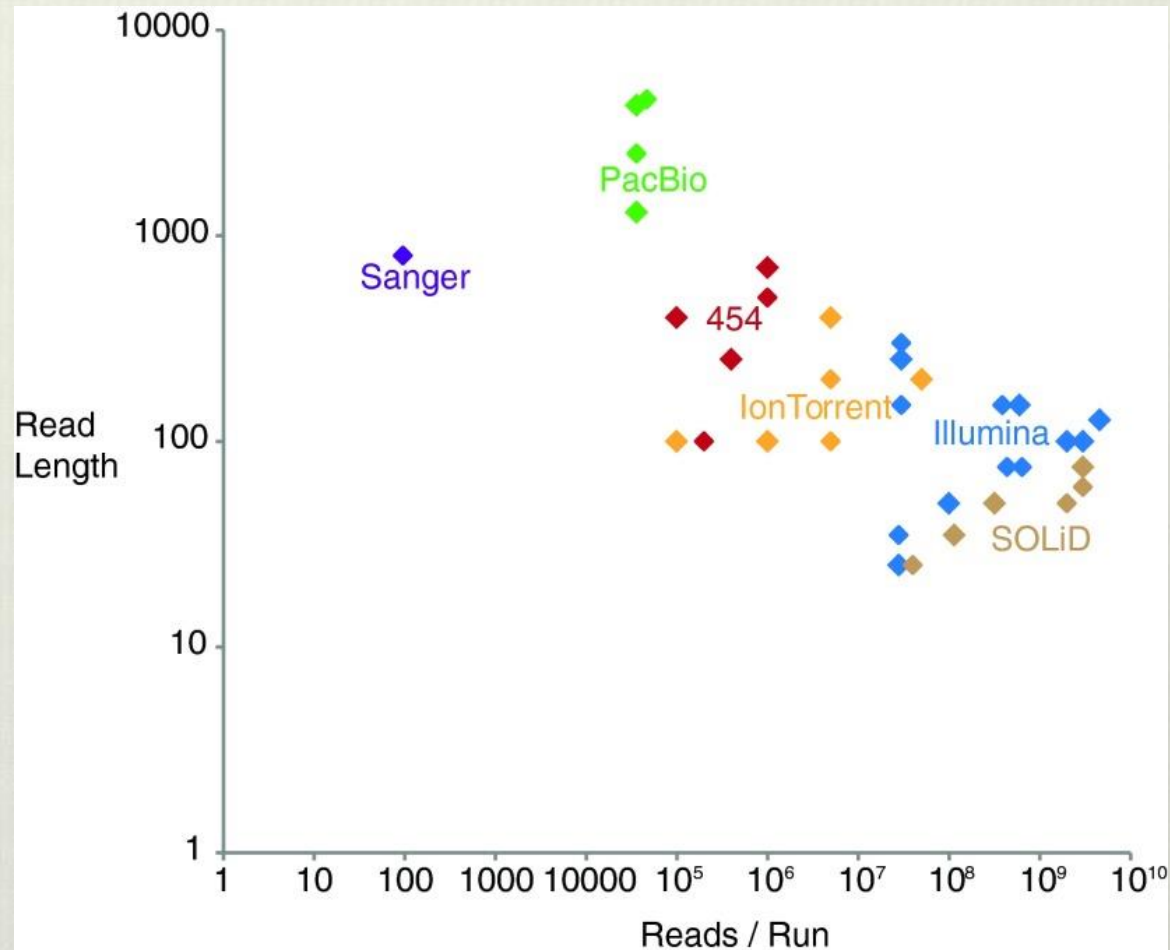
Michael L. Metzker\*\*

# Comparación de métodos

**Cada método en el mercado tiene sus fuertes y debilidades**

## Principales diferencias:

- La longitud de la lectura
- El número de lecturas
- El tiempo que lleva un proceso
- Los tipos de errores que podemos esperar
  - 454 indels
  - Illumina cambios de una base



ADVANCES IN WOUND CARE

Many Annals, Inc. publishes

Journals

Search

Alerts

Adv Wound Care (New Rochelle). 2015 Jan 1; 4(1): 50-58.  
doi: [10.1089/wound.2014.0542](https://doi.org/10.1089/wound.2014.0542)

PMCID: PMC4281878

**Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research**

Brendan P. Hodkinson and Elizabeth A. Grice\*

# APLICACIONES

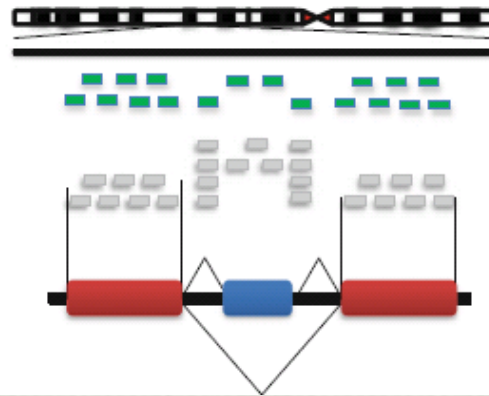
## Secuenciación del ADN genómico



ACCCGTTACGTAAACGTTT C AGATGACGATGACCAAGGTTGACGA  
ACCCGTTACGTAAACGTTT  
ACCCGTTACGTAAACGTTT G AGA  
ACCCGTTACGTAAACGTTT G AGATGAC  
ACCCGTTACGTAAACGTTT G AGATGACGATA  
ACCCGTTACGTAAACGTTT G AGATGACGATGACCA  
CGTTACGTAAACGTTT G AGATGACGATGACCAAGG  
ACGTAAACGTTT G AGATGACGATGACCAAGGTTGA  
TAAACGTTT G AGATGACGATGACCAAGGTTGACGA  
CGTTT G AGATGACGATGACCAAGGTTGACGA  
T G AGATGACGATGACCAAGGTTGACGA

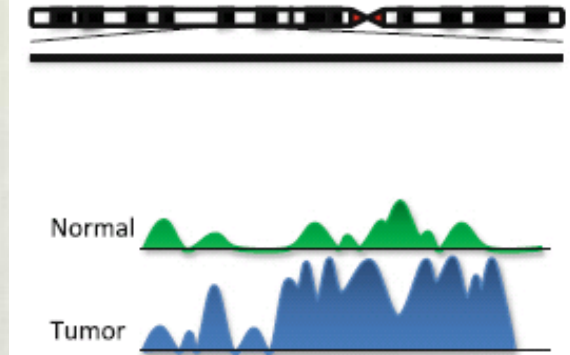
- Ensamblar un genoma
- SNVs and CNVs
- Inserciones y deleciones

## Expresión génica y su regulación



- Expresión génica
- ARNs pequeños
- TFBSs

## Epigenómica

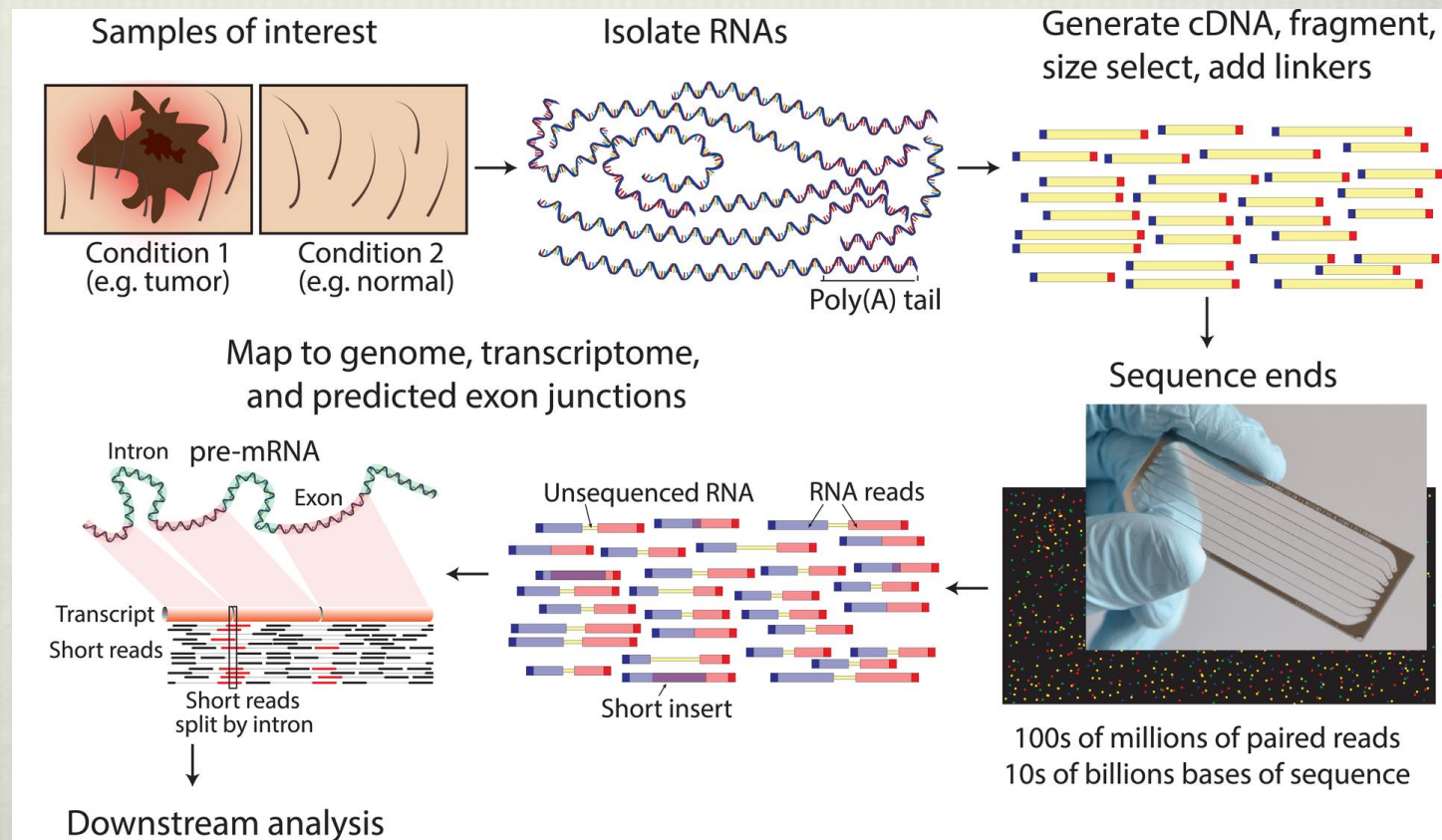


- Metilación del ADN
- Modificaciones químicas en histonas
- Cromatina

# RNA-seq

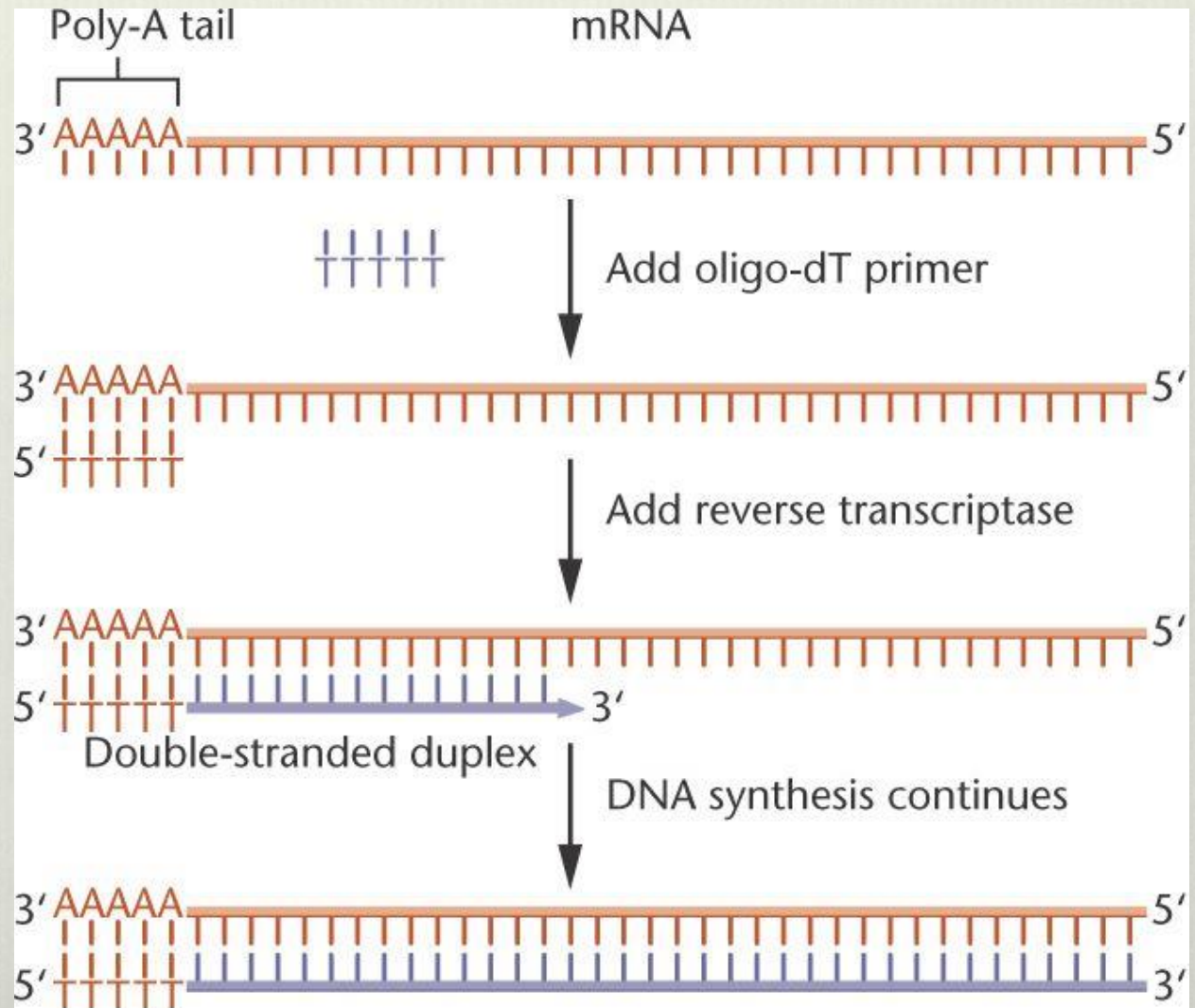
## Permite:

- Detectar transcritos raros sin conocer el gen previamente
- Detectar splicing alternativo
- Detectar variación de secuencia



# RNA-seq

## Transcripción inversa usando la poli-A

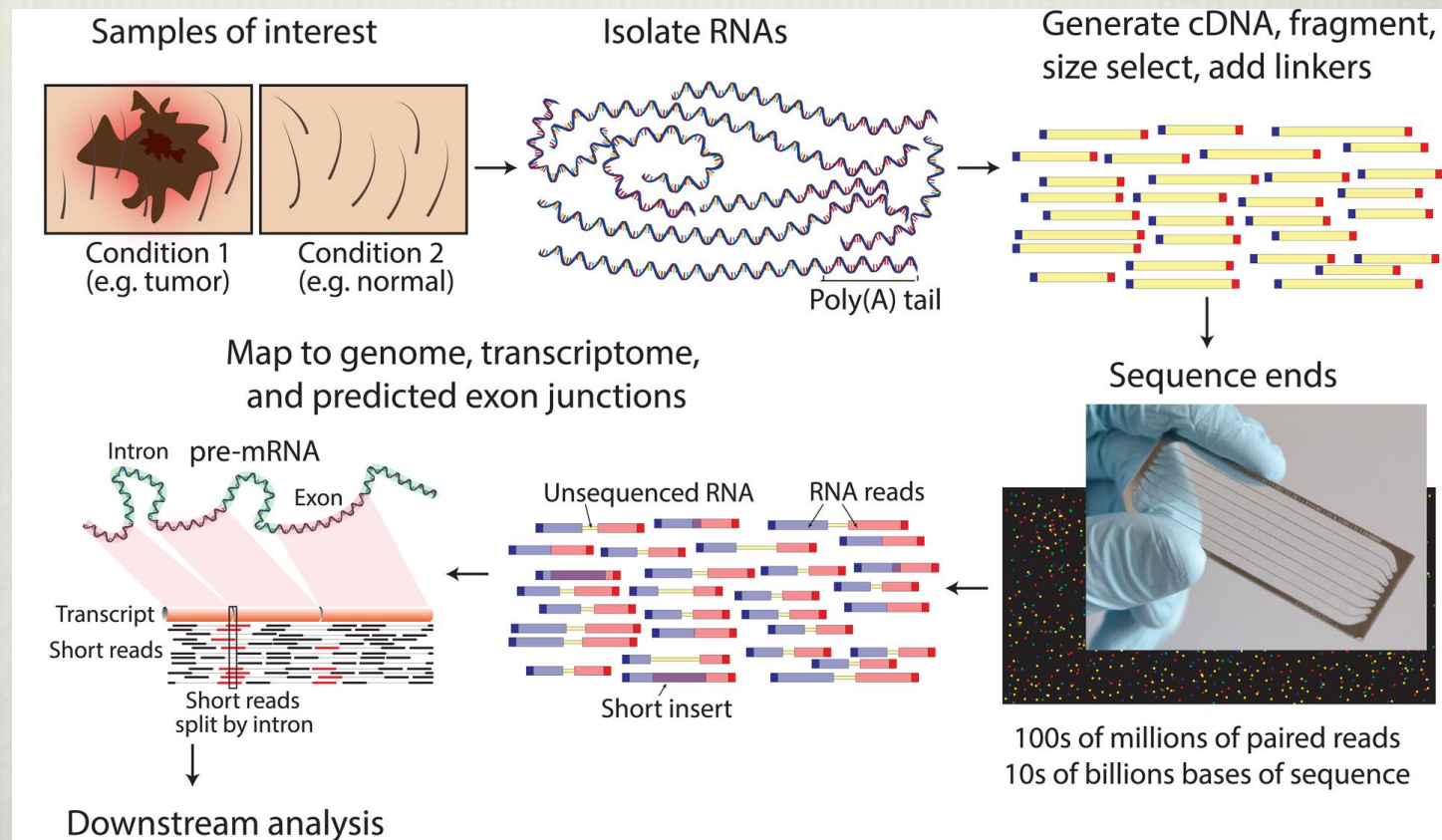


<http://bio3400.nicerweb.com/Locked/media/ch19/cDNA.html>

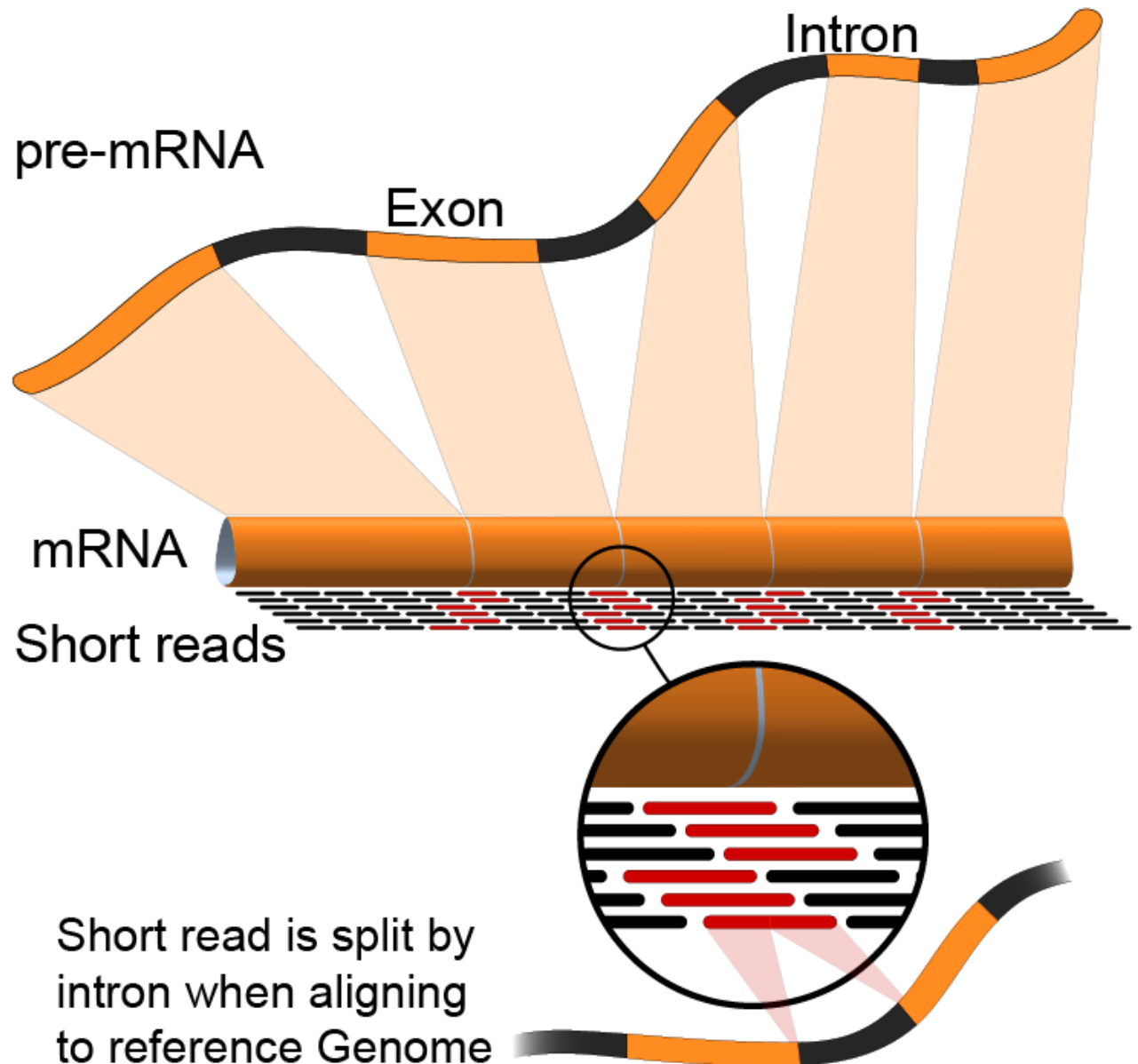
# RNA-seq

## Permite:

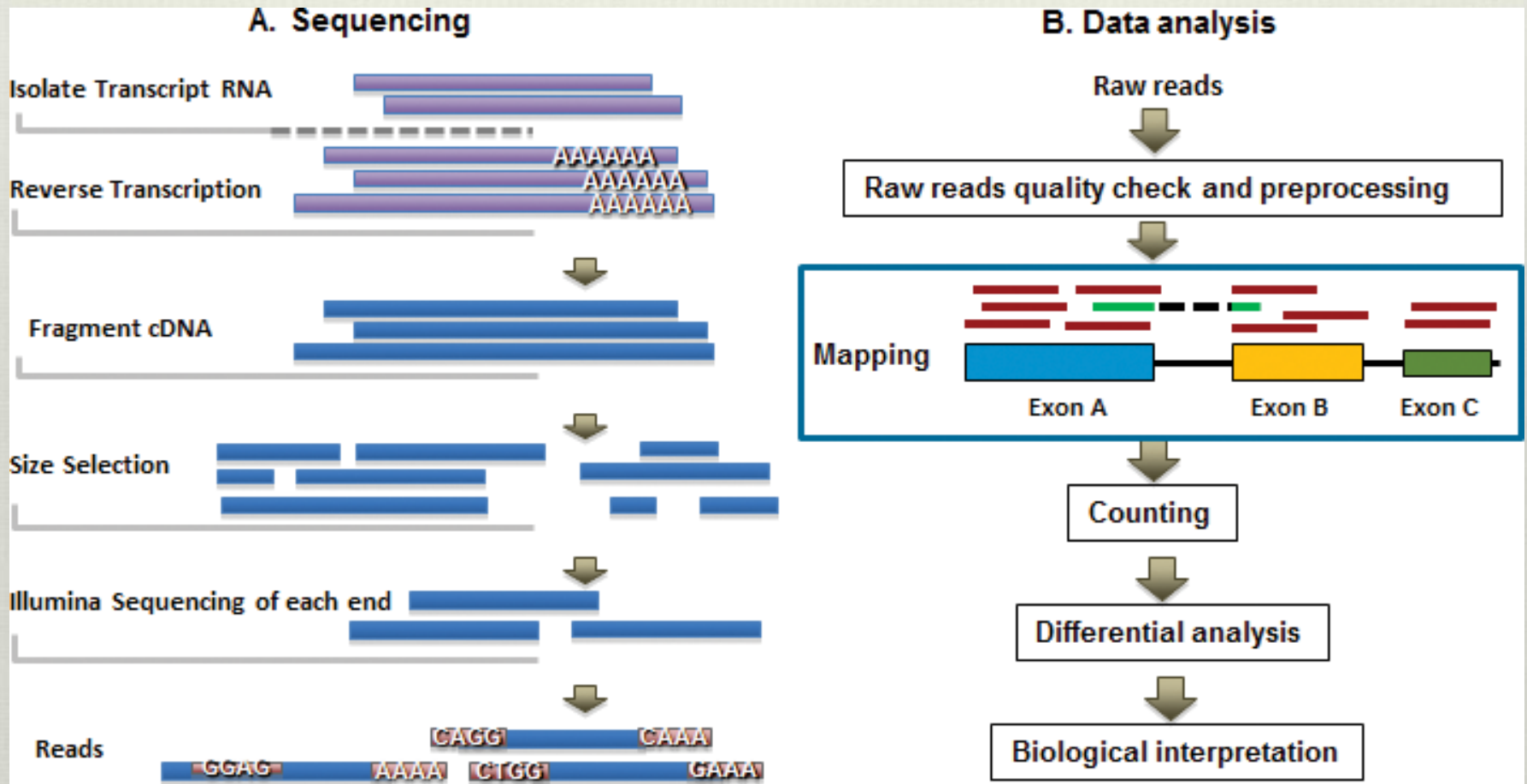
- Detectar transcritos raros sin conocer el gen previamente
- Detectar splicing alternativo
- Detectar variación de secuencia



# Detectar fronteras entre exones y intrones



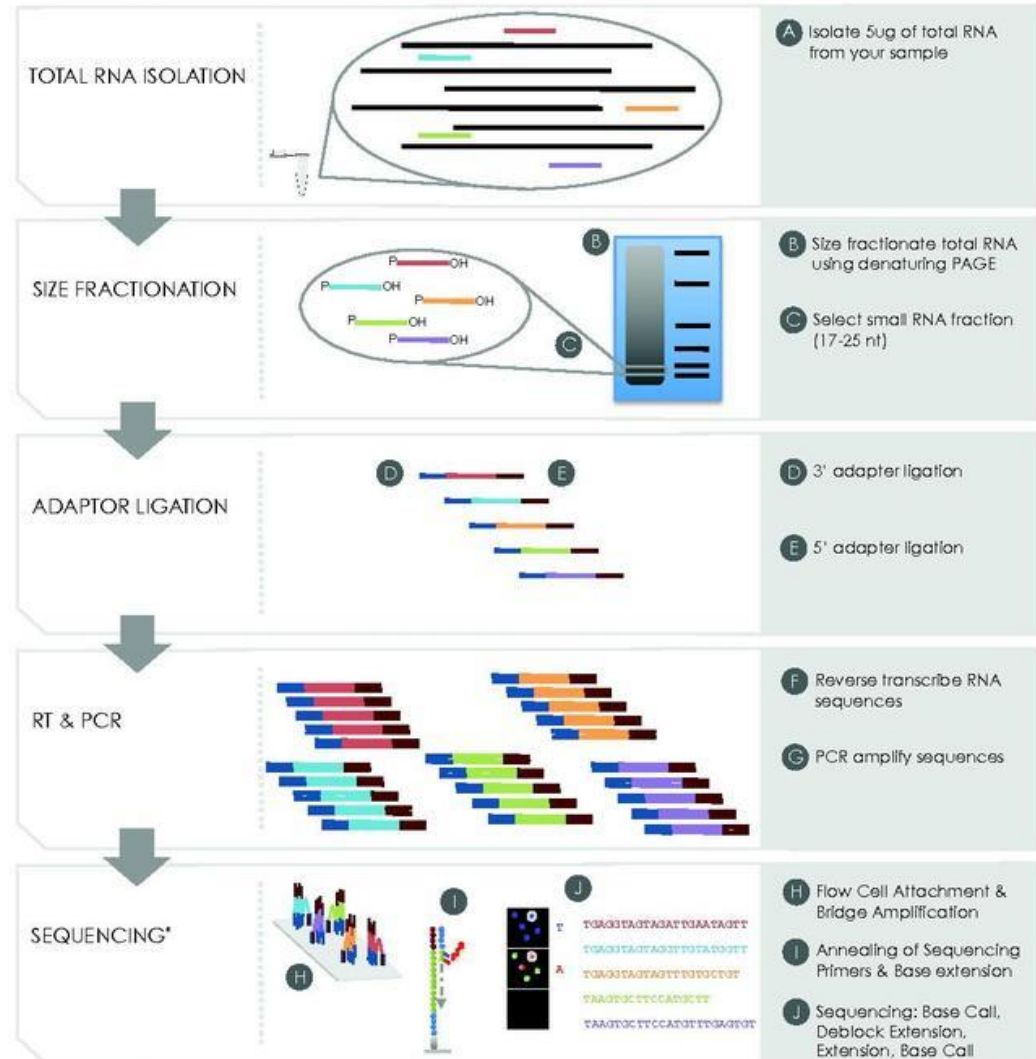
# RNA-seq



# miRNA-seq

- Extracción del RNA total
- Ligar adaptadores
- Seleccionar por longitud
- Generar una librería de cDNA
- Secuenciar
- Análisis bioinformático

## MIRNA-SEQ LIBRARY PREPARATION



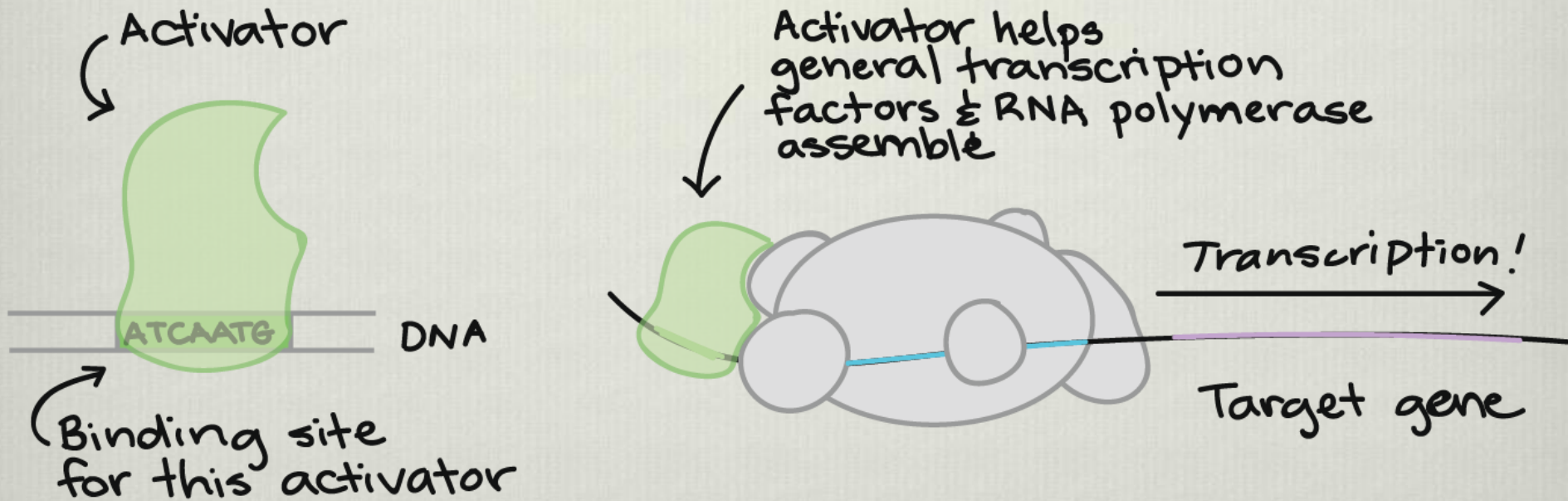
\*Illumina sequencing method depicted however other sequencing platforms can also be used.

Source:

[http://en.wikipedia.org/wiki/MicroRNA\\_Sequencing](http://en.wikipedia.org/wiki/MicroRNA_Sequencing)

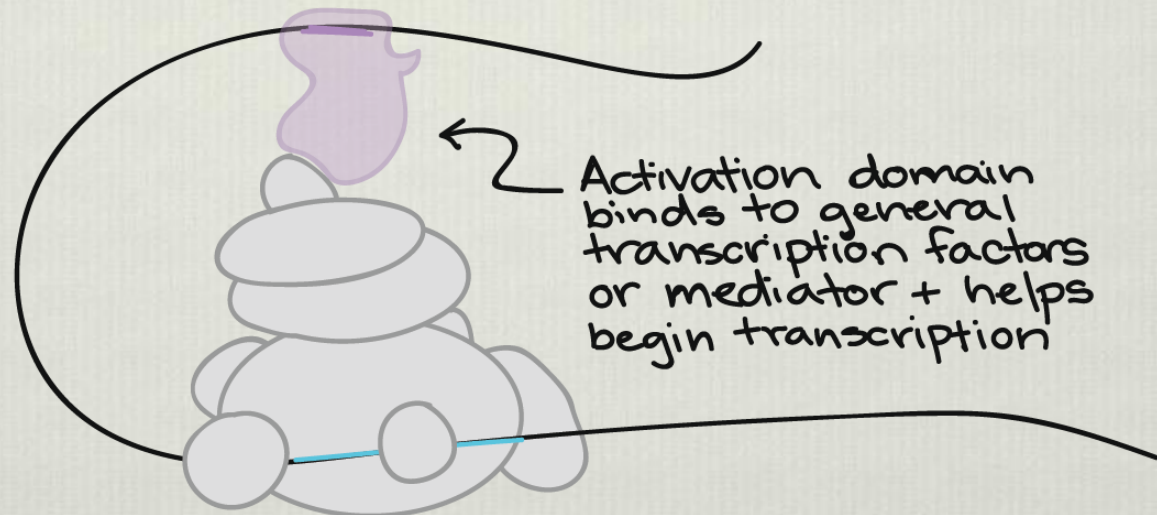
# CHIP-seq

- Un factor de transcripción (TF) es una proteína que se une a una secuencia de DNA específica
- Controlan la expresión génica
- Pueden tanto promover (activadores - *activator*) o bloquear (represores- *repressor*) la unión de la polimerasa de RNA
- Estas proteínas tienen dominios específicos que les permite unirse al DNA
- Los TF se unen o al promotor o a un *enhancer* (“potenciador”)



# CHIP-seq

- Un factor de transcripción (TF) es una proteína que se une a una secuencia de DNA específica
- Controlan la expresión génica
- Pueden tanto promover (activadores - *activator*) o bloquear (represores- *repressor*) la unión de la polimerasa de RNA
- Estas proteínas tienen dominios específicos que les permite unirse al DNA
- Los TF se unen o al promotor o a un *enhancer* (“potenciador”)

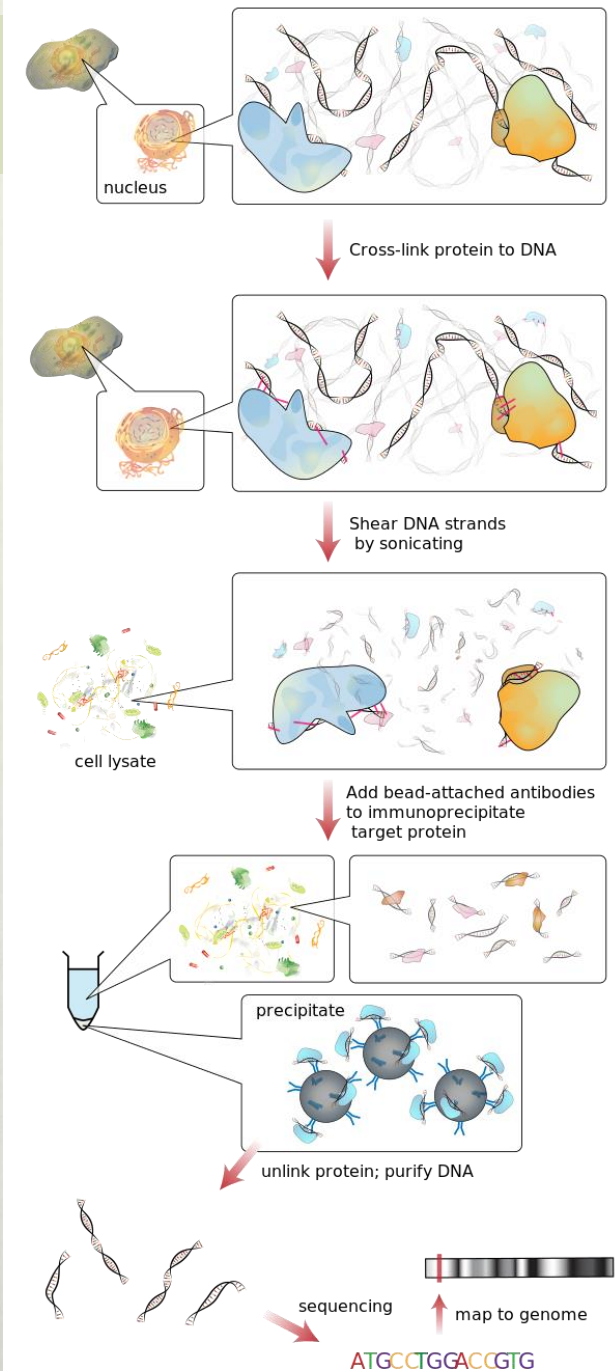


# CHIP-seq

## Detección experimental de los sitios de unión a factores de transcripción

1. Enlaces covalentes entre ADN y proteínas mediante formaldehído (reversible)
2. Inmunoprecipitación mediante un anticuerpo específico al factor de transcripción
3. Extraer /enriquecer las secuencias de ADN a las que se había unido una proteína
4. Secuenciar el ADN
5. Localizar el ADN en el genoma

<https://en.wikipedia.org/wiki/ChIP-sequencing>



# La salida: el formato fastq

## La salida del secuenciador: Los reads (lecturas) en formato fastq

```
@SRR037876.8543926 GSM522374_1:1:148:931:861
TAGTTCTACAGTCCGACGATCTCGTATGCCGTCTTC
+
BB@+?0:4@B@-@/A<3A7@-=@<1=@87=?<==9#
@SRR037876.8543927 GSM522374_1:1:148:931:517
AGCTACATTGTCTGCTGGGTTTCTCGTATGCCGTCT
+
BBC@3<1=872661.@C;@A93+?:;.2.?386<;>
@SRR037876.8543928 GSM522374_1:1:148:931:648
TGAGGTAGTAGGTTGTGTGGTTAATCGTATGCCGCT
+
5@059)@6?' :9>0<@@)@=BA8)99@3258?####
@SRR037876.8543929 GSM522374_1:1:148:931:770
GCTACATTGTCTGCTGGGTTTCTCGTATGCCGTCTT
+
```

Identificador

Secuencia de la lectura ('read')

Calidad de la lectura'

### La calidad

Los caracteres se pueden convertir en un score (Q, Phred score) de calidad

# La calidad

Línea de calidad:

BB@+?0:4@B@-@/A<3A7

ASCII codificación

B = 66

@ = 64

+ = 43

...

Codificación depende del fabricante  
(por ejemplo codificación Sanger para la  
línea de calidad de de arriba)

$Q(B) = 66 - 33 = 33$  (primer base)

$Q(@) = 64 - 33 = 31$  (tercer base)

$Q(+)= 43 - 33 = 10$  (cuarta base)

...

| Dec | Hx | Oct | Html  | Chr | Dec | Hx | Oct | Html | Chr | Dec | Hx | Oct | Html  | Chr |
|-----|----|-----|-------|-----|-----|----|-----|------|-----|-----|----|-----|-------|-----|
| 32  | 20 | 040 | Space |     | 64  | 40 | 100 | #64; | 0   | 96  | 60 | 140 | #96;  | `   |
| 33  | 21 | 041 | !     |     | 65  | 41 | 101 | #65; | A   | 97  | 61 | 141 | #97;  | a   |
| 34  | 22 | 042 | "     |     | 66  | 42 | 102 | #66; | B   | 98  | 62 | 142 | #98;  | b   |
| 35  | 23 | 043 | #     |     | 67  | 43 | 103 | #67; | C   | 99  | 63 | 143 | #99;  | c   |
| 36  | 24 | 044 | \$    |     | 68  | 44 | 104 | #68; | D   | 100 | 64 | 144 | #100; | d   |
| 37  | 25 | 045 | %     |     | 69  | 45 | 105 | #69; | E   | 101 | 65 | 145 | #101; | e   |
| 38  | 26 | 046 | &     |     | 70  | 46 | 106 | #70; | F   | 102 | 66 | 146 | #102; | f   |
| 39  | 27 | 047 | '     |     | 71  | 47 | 107 | #71; | G   | 103 | 67 | 147 | #103; | g   |
| 40  | 28 | 050 | (     |     | 72  | 48 | 110 | #72; | H   | 104 | 68 | 150 | #104; | h   |
| 41  | 29 | 051 | )     |     | 73  | 49 | 111 | #73; | I   | 105 | 69 | 151 | #105; | i   |
| 42  | 2A | 052 | *     |     | 74  | 4A | 112 | #74; | J   | 106 | 6A | 152 | #106; | j   |
| 43  | 2B | 053 | +     |     | 75  | 4B | 113 | #75; | K   | 107 | 6B | 153 | #107; | k   |
| 44  | 2C | 054 | ,     |     | 76  | 4C | 114 | #76; | L   | 108 | 6C | 154 | #108; | l   |
| 45  | 2D | 055 | -     |     | 77  | 4D | 115 | #77; | M   | 109 | 6D | 155 | #109; | m   |
| 46  | 2E | 056 | .     |     | 78  | 4E | 116 | #78; | N   | 110 | 6E | 156 | #110; | n   |
| 47  | 2F | 057 | /     |     | 79  | 4F | 117 | #79; | O   | 111 | 6F | 157 | #111; | o   |
| 48  | 30 | 060 | 0     |     | 80  | 50 | 120 | #80; | P   | 112 | 70 | 160 | #112; | p   |
| 49  | 31 | 061 | 1     |     | 81  | 51 | 121 | #81; | Q   | 113 | 71 | 161 | #113; | q   |
| 50  | 32 | 062 | 2     |     | 82  | 52 | 122 | #82; | R   | 114 | 72 | 162 | #114; | r   |
| 51  | 33 | 063 | 3     |     | 83  | 53 | 123 | #83; | S   | 115 | 73 | 163 | #115; | s   |
| 52  | 34 | 064 | 4     |     | 84  | 54 | 124 | #84; | T   | 116 | 74 | 164 | #116; | t   |
| 53  | 35 | 065 | 5     |     | 85  | 55 | 125 | #85; | U   | 117 | 75 | 165 | #117; | u   |
| 54  | 36 | 066 | 6     |     | 86  | 56 | 126 | #86; | V   | 118 | 76 | 166 | #118; | v   |
| 55  | 37 | 067 | 7     |     | 87  | 57 | 127 | #87; | W   | 119 | 77 | 167 | #119; | w   |
| 56  | 38 | 070 | 8     |     | 88  | 58 | 130 | #88; | X   | 120 | 78 | 170 | #120; | x   |
| 57  | 39 | 071 | 9     |     | 89  | 59 | 131 | #89; | Y   | 121 | 79 | 171 | #121; | y   |
| 58  | 3A | 072 | :     |     | 90  | 5A | 132 | #90; | Z   | 122 | 7A | 172 | #122; | z   |
| 59  | 3B | 073 | ;     |     | 91  | 5B | 133 | #91; | [   | 123 | 7B | 173 | #123; | (   |
| 60  | 3C | 074 | <     |     | 92  | 5C | 134 | #92; | \   | 124 | 7C | 174 | #124; | !   |
| 61  | 3D | 075 | =     |     | 93  | 5D | 135 | #93; | ]   | 125 | 7D | 175 | #125; | )   |
| 62  | 3E | 076 | >     |     | 94  | 5E | 136 | #94; | ^   | 126 | 7E | 176 | #126; | ~   |

S - Sanger Phred+33, raw reads typically (0, 40)  
X - Solexa Solexa+64, raw reads typically (-5, 40)  
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)  
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)  
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)  
(Note: See discussion above).  
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

# La calidad

## Phred Score y probabilidad de un error de secuenciación

$$Q = -10 \log_{10} P$$

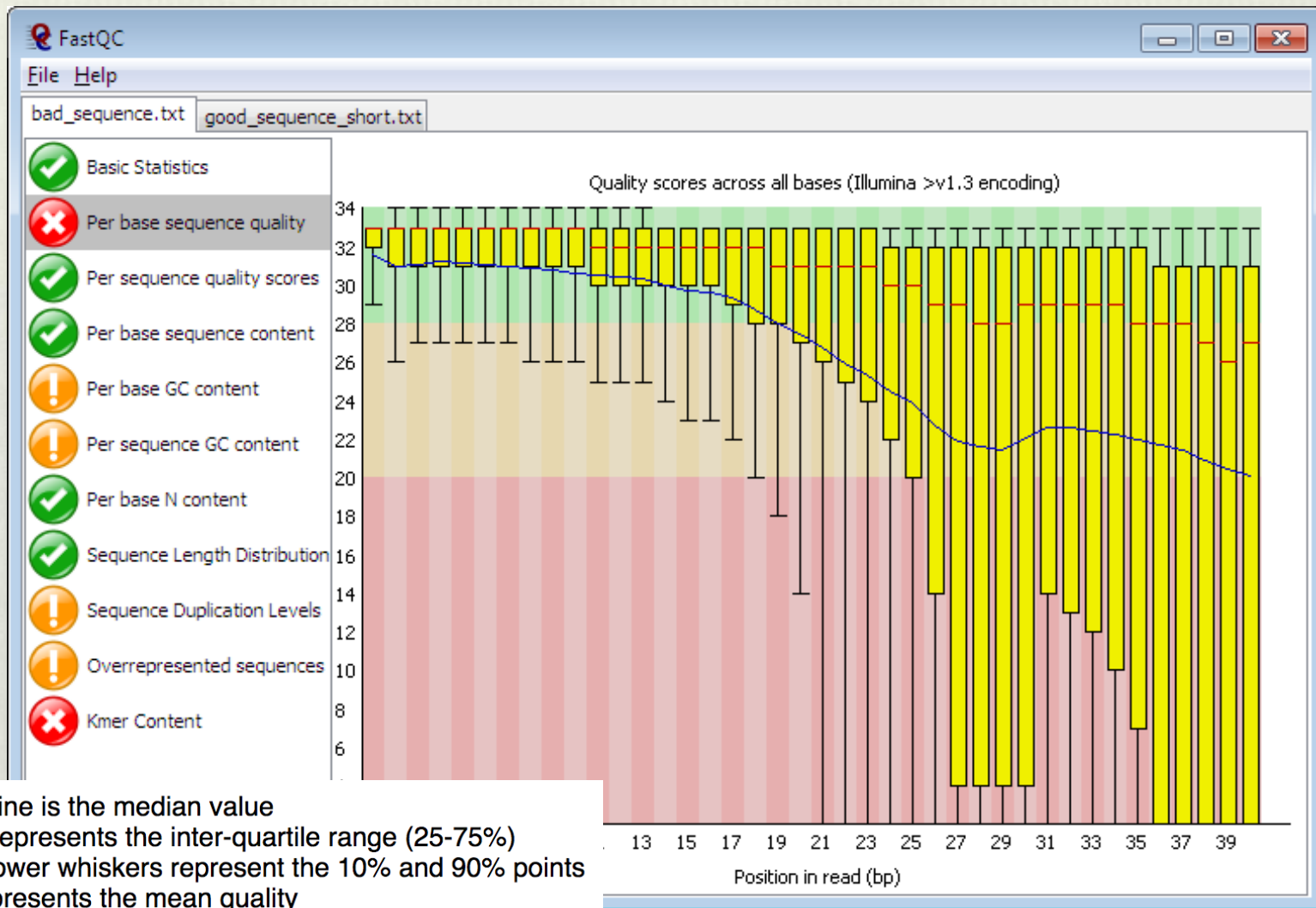
or

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

### Phred quality scores are logarithmically linked to error probabilities

| Phred Quality Score | Probability of incorrect base call | Base call accuracy |
|---------------------|------------------------------------|--------------------|
| 10                  | 1 in 10                            | 90 %               |
| 20                  | 1 in 100                           | 99 %               |
| 30                  | 1 in 1000                          | 99.9 %             |
| 40                  | 1 in 10000                         | 99.99 %            |
| 50                  | 1 in 100000                        | 99.999 %           |

# Control de calidad: FastQC



- The central red line is the median value
- The yellow box represents the inter-quartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% points
- The blue line represents the mean quality

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality.html>

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

# Adaptadores y lecturas con baja calidad

**Si la molécula de RNA/DNA o el fragmento de DNA es mas corto que el número de ciclos**

→ Se secuencia parte del 3' adaptador

→ Hay que encontrar esta secuencia y eliminarla de la lectura

**La calidad de la secuenciación va bajando hacia el extremo 3'**

→ Cortar la parte con menor calidad en 3'

Trimmomatic:

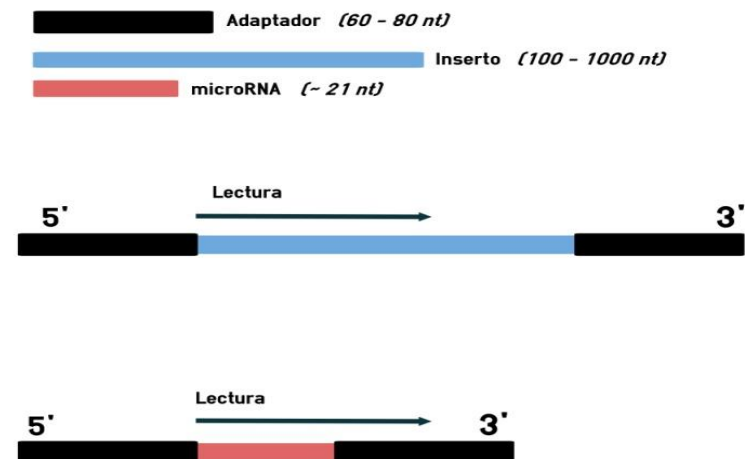
<http://www.usadellab.org/cms/?page=trimmomatic>

Cutadapt:

<https://cutadapt.readthedocs.io/en/stable/>

Trim\_galore:

[https://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)



# Adaptadores y lecturas con baja calidad

La idea básica es eliminar las bases del extremo 3' que están por debajo de cierto umbral de calidad → ¿usar la información válida del extremo 5'!

Para eso, cutadapt usa la sumación parcial desde el extremo 3':

Suponemos un umbral de calidad de  $Q=10$  y la siguiente serie de valores  $Q$

|     |     |     |     |    |    |     |    |    |   |
|-----|-----|-----|-----|----|----|-----|----|----|---|
| A   | C   | T   | G   | A  | C  | T   | G  | A  | C |
| 42, | 40, | 26, | 27, | 8, | 7, | 11, | 4, | 2, | 3 |

Se resta primero el umbral de los valores  $Q$ :

32, 30, 16, 17, -2, -3, 1, -6, -8, -7

Se calcula la suma parcial (empezando en el extremo 3')

(70), (38), 8, -8, -25, -23, -20, -21, -15, -7

- El proceso se para en un valor  $> 0$
- La lectura se corta en la posición que presenta el mínimo en la suma parcial
- La lectura final se quedaría en las primeras 4 nucleótidos: ACTGACTGAC → ACTG