

# Bioinformática y análisis de datos ómicos

## TEMA 6: TECNOLOGÍAS DE ANÁLISIS MOLECULARES MASIVOS



**Ignacio Varela Egocheaga**

DEPARTAMENTO DE BIOLOGÍA MOLECULAR

Este material se publica bajo la siguiente licencia:

[Creative Commons BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)

# **Tema 6: Tecnologías de análisis moleculares masivos**

6.1 Chips genéticos

6.2 Principios de la secuenciación

6.3 Generación de librerías genómicas

6.4 Plataformas de secuenciación

6.5 Técnicas de secuenciación de célula única

6.6 Control de calidad de la secuenciación

# **Tema 6: Tecnologías de análisis moleculares masivos**

## **6.1 Chips genéticos**

## 6.2 Principios de la secuenciación

## 6.3 Generación de librerías genómicas

## 6.4 Plataformas de secuenciación

## 6.5 Técnicas de secuenciación de célula única

## 6.6 Control de calidad de la secuenciación

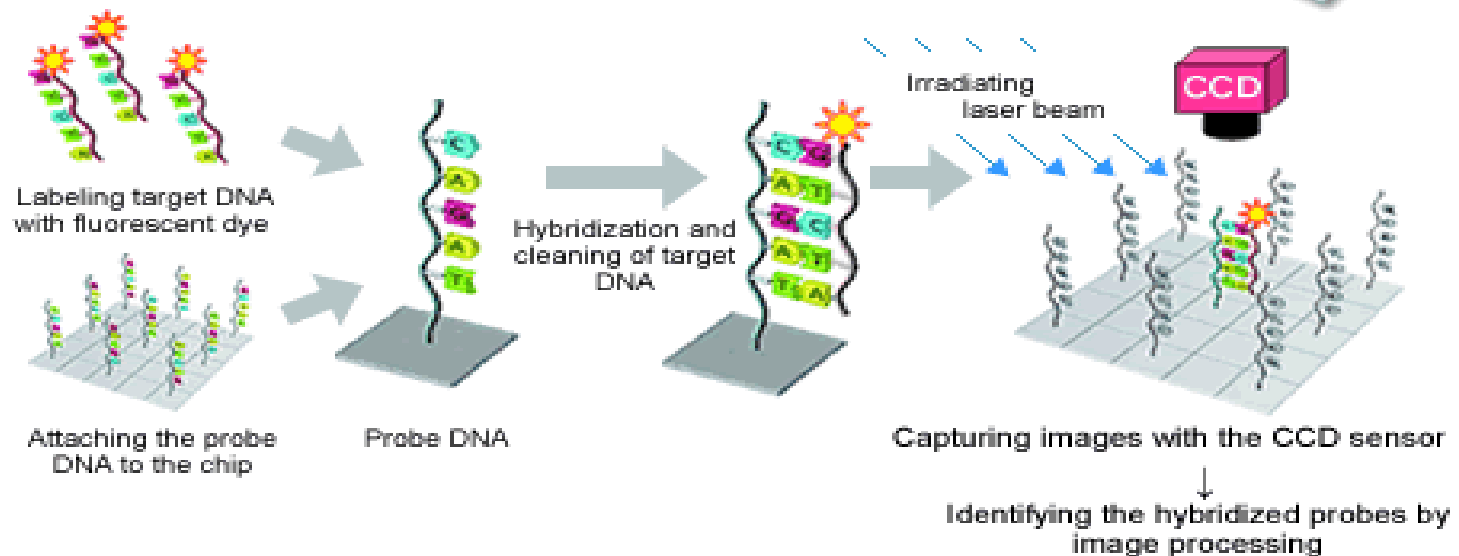
# Fundamentos de un chip genético

Un Chip de DNA es un soporte sólido que contiene, en distintas localizaciones espaciales, oligonucleótidos (sondas) con una secuencia conocida complementaria a una región de interés en una muestra biológica.

Una muestra de DNA/RNA se marca típicamente con un componente fluorescente y se hibrida con el chip. Aquellas regiones con secuencia complementaria a alguna de las sondas del chip se quedarán fuertemente unidas a estas sondas resistiendo los lavados y provocando una señal estable en esta sonda directamente proporcional al número de moléculas unidas.

## Fluorescence-Detection DNA Chip

The four bases A, T, G, and C bind A to T or G to C. A target DNA sequence is analyzed by checking which bases the target DNA bases bind.

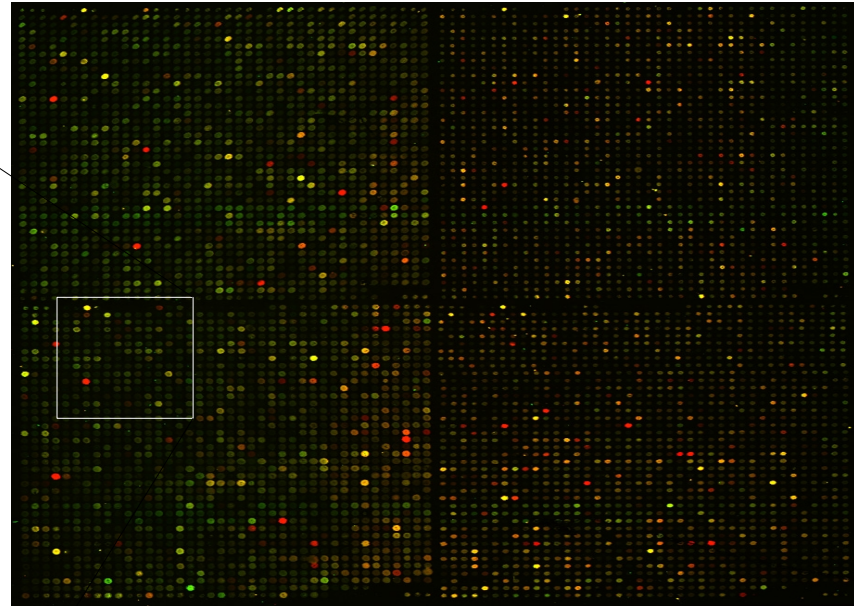
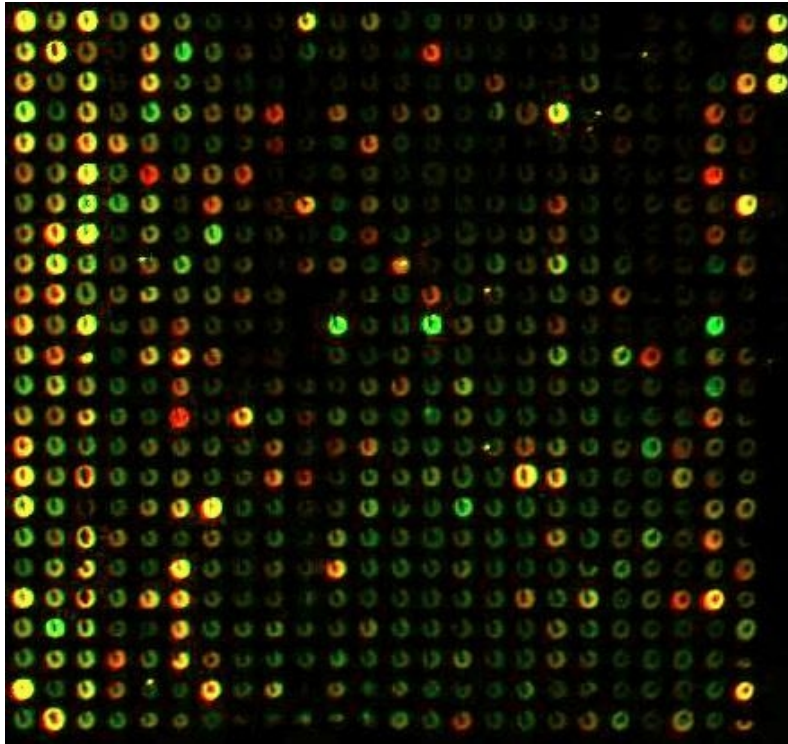


<http://science.nayland.school.nz/graemeb/yr13%20work/biotech/biotechnology.htm>



# Fundamentos de un chip genético

<http://www.actionbioscience.org/esp/educacion/guilfoile.html>

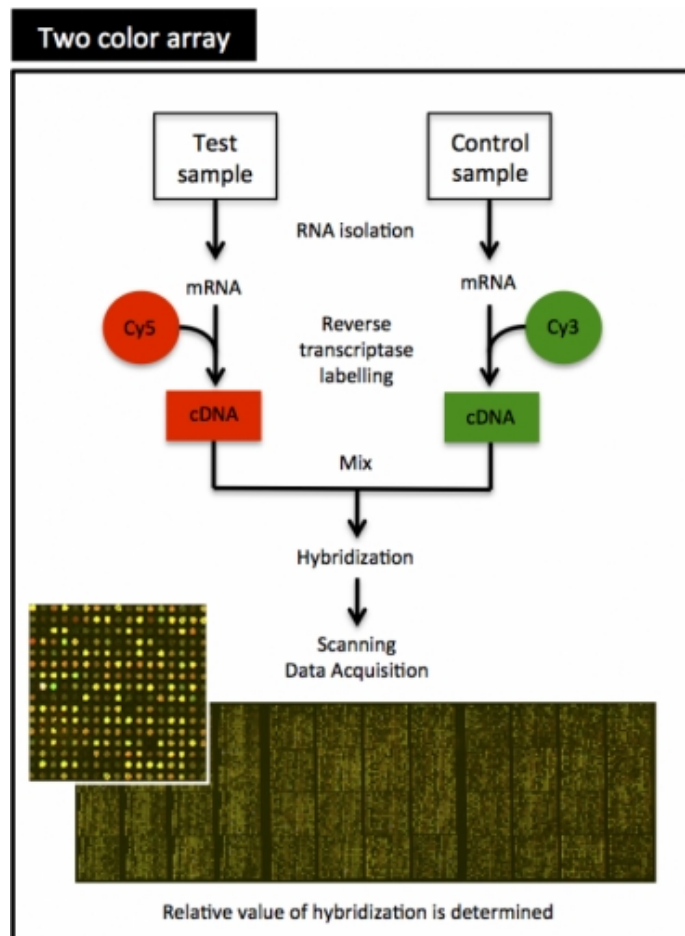


<http://www.seedgenenetwork.net/annotate>

# Fundamentos de un chip genético

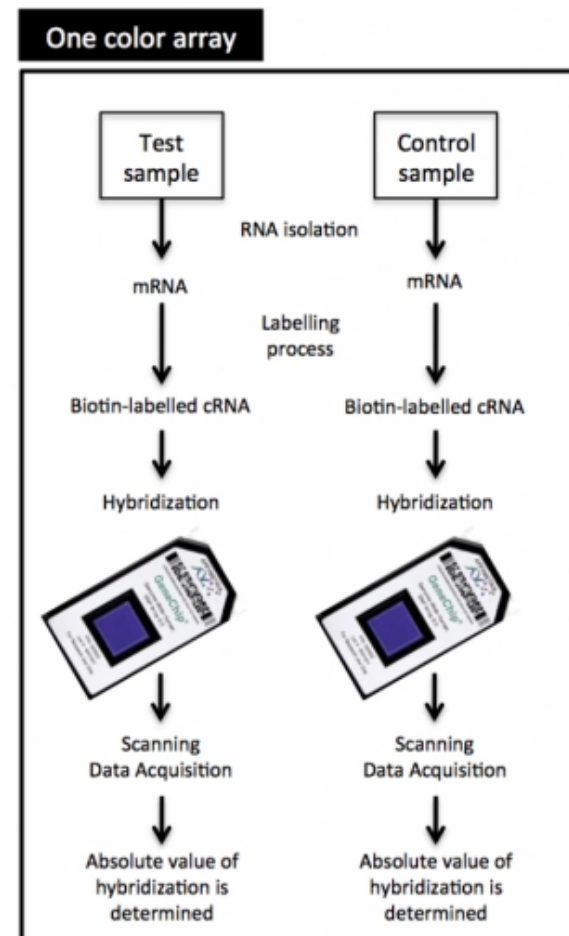
## Hibridación competitiva

Se basa en la hibridación competitiva de dos muestras marcadas con marcadores fluorescentes distintos sobre el mismo chip. Se mide la señal relativa de cada marcador.



## Hibridación no competitiva

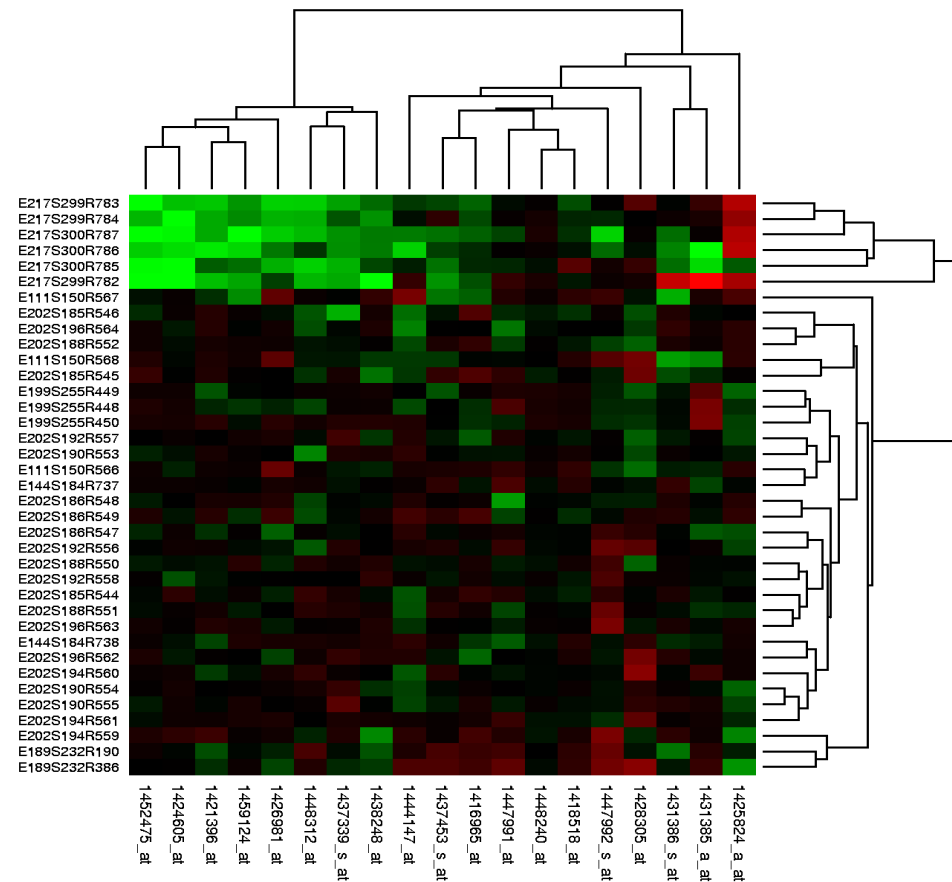
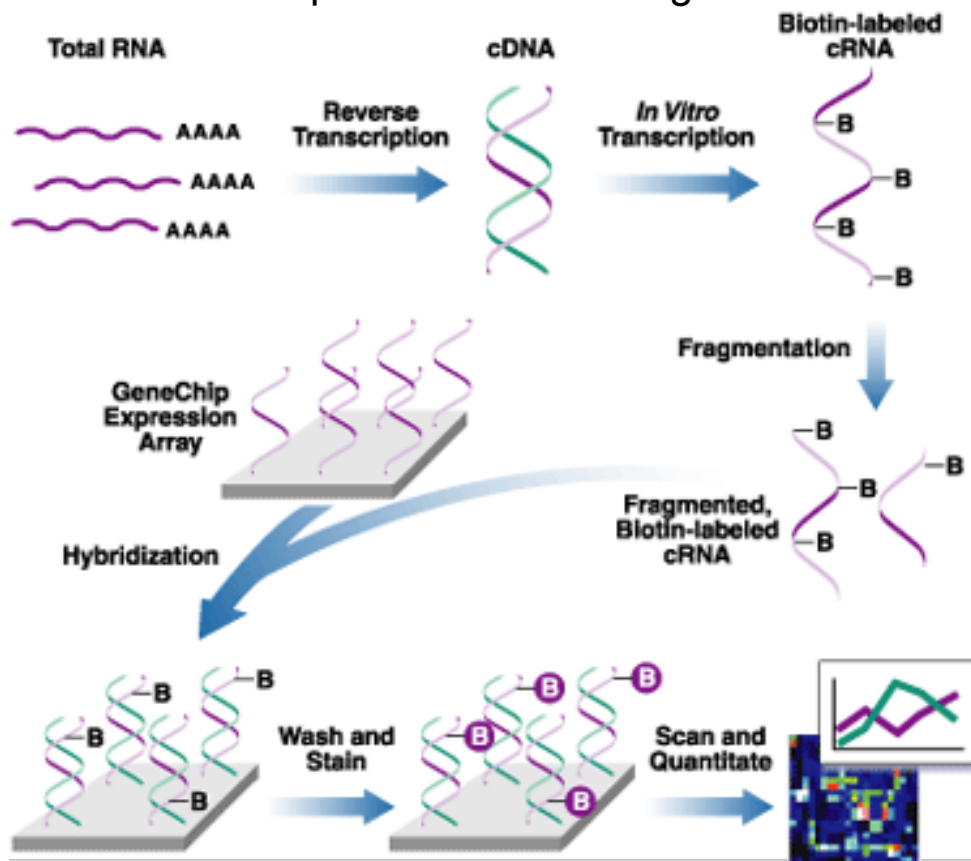
Se hibrida una sola muestra en cada chip. Permite la comparación de varios grupos de muestras y la agrupación de las mismas ("clustering"). Hay que realizar múltiples correcciones para garantizar la comparabilidad del experimento.



# Aplicaciones de los chips genéticos

## Chips de expresión genética

Sondas complementarias a las regiones codificantes de los genes de interés. Típicamente varias sondas correspondientes a cada gen.



<http://dept.stat.lsa.umich.edu/~kshedden/Courses/Stat545/Notes/microarray.html>

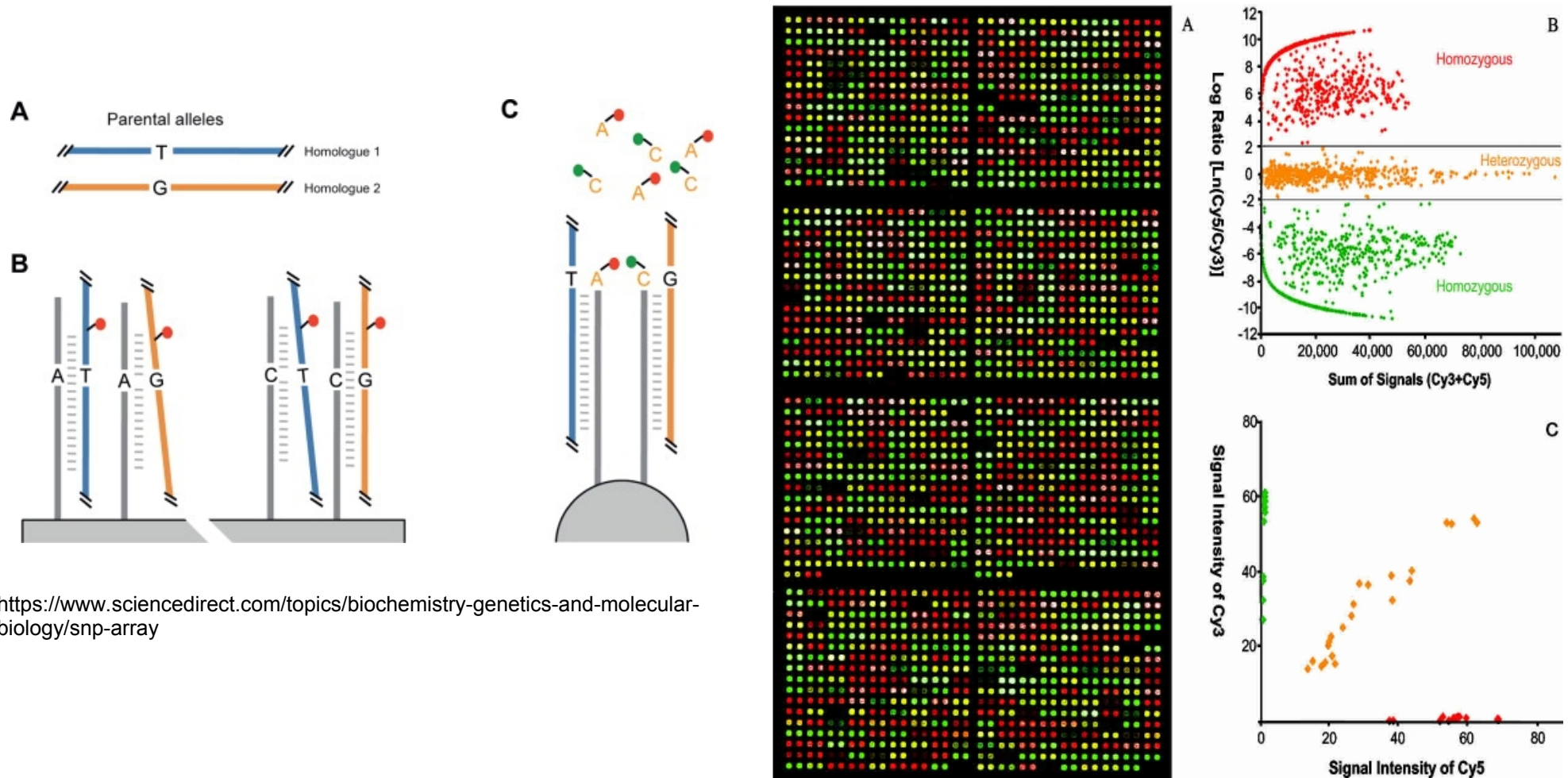
[http://en.wikipedia.org/wiki/DNA\\_microarray](http://en.wikipedia.org/wiki/DNA_microarray)



# Aplicaciones de los chips genéticos

## Chips de genotipado

Sondas que permiten distinguir variables alélicas en posiciones concretas de interés.

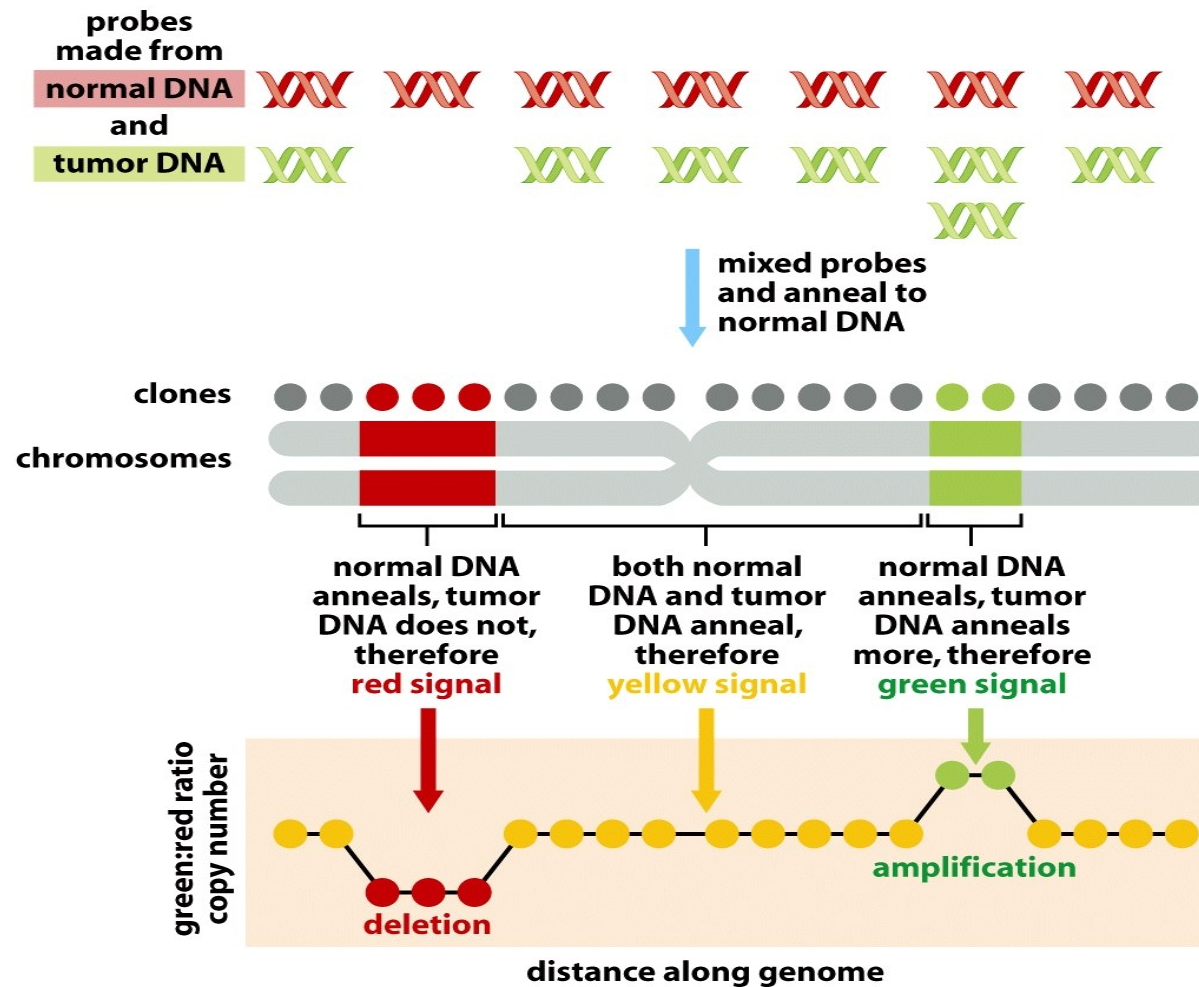


<https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/snp-array>

# Aplicaciones de los chips genéticos

## Hibridación comparativa de genoma (CGH)

Sondas complementarias a regiones distribuidas uniformemente a lo largo del genoma.



Robert Weinberg. *The Biology of Cancer*. Garland Science(2007 )

# **Tema 6: Tecnologías de análisis moleculares masivos**

6.1 Chips genéticos

**6.2 Principios de la secuenciación**

6.3 Generación de librerías genómicas

6.4 Plataformas de secuenciación

6.5 Técnicas de secuenciación de célula única

6.6 Control de calidad de la secuenciación

# Secuenciación Sanger

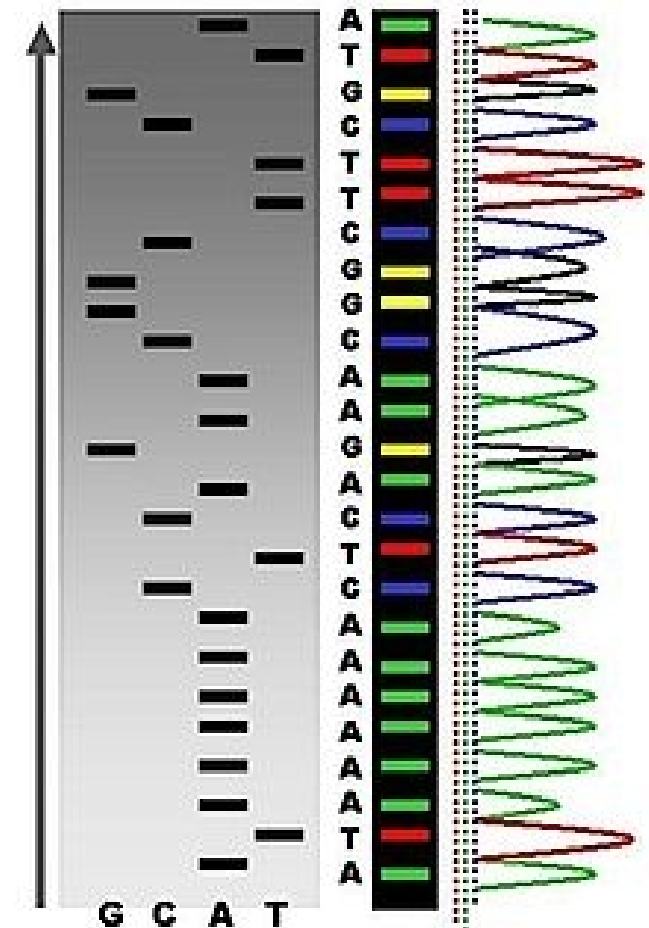
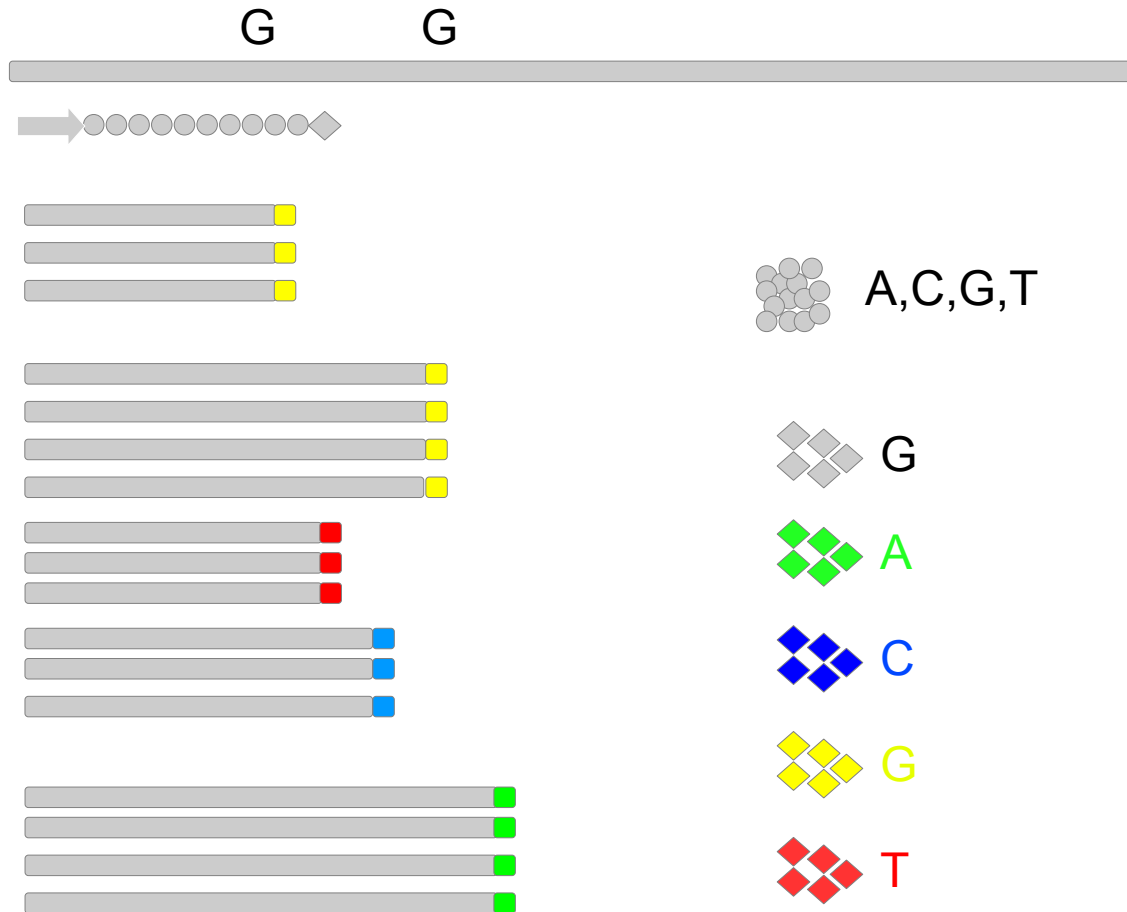
*Proc. Natl. Acad. Sci. USA*  
Vol. 74, No. 12, pp. 5463–5467, December 1977  
Biochemistry

## **DNA sequencing with chain-terminating inhibitors**

(DNA polymerase/nucleotide sequences/bacteriophage  $\phi$ X174)

F. SANGER, S. NICKLEN, AND A. R. COULSON

Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England





# Secuenciación de nueva generación

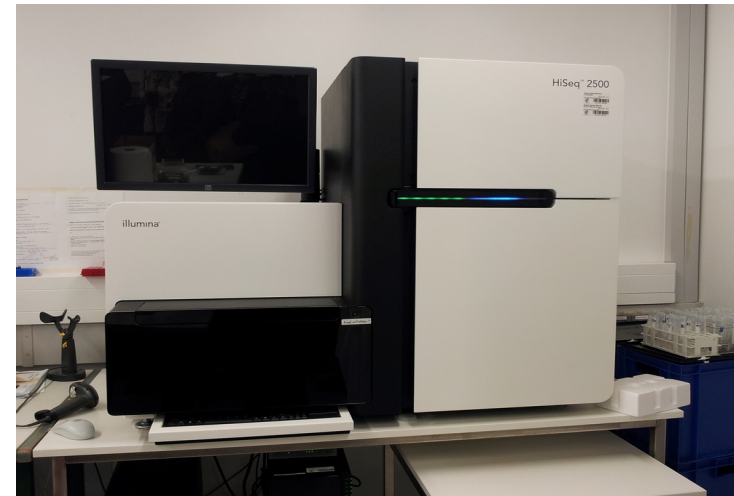
(1990-2003)



<http://www.blog.gurukpo.com/wp-content/uploads/2012/08/metal-gear-solid-human-genome-project-big.jpg>

- Costó ~2.700 millones de dólares
- Se tardaron 13-15 años
- Se llevó a cabo gracias a la colaboración de cientos de investigadores de al menos 6 países distintos.

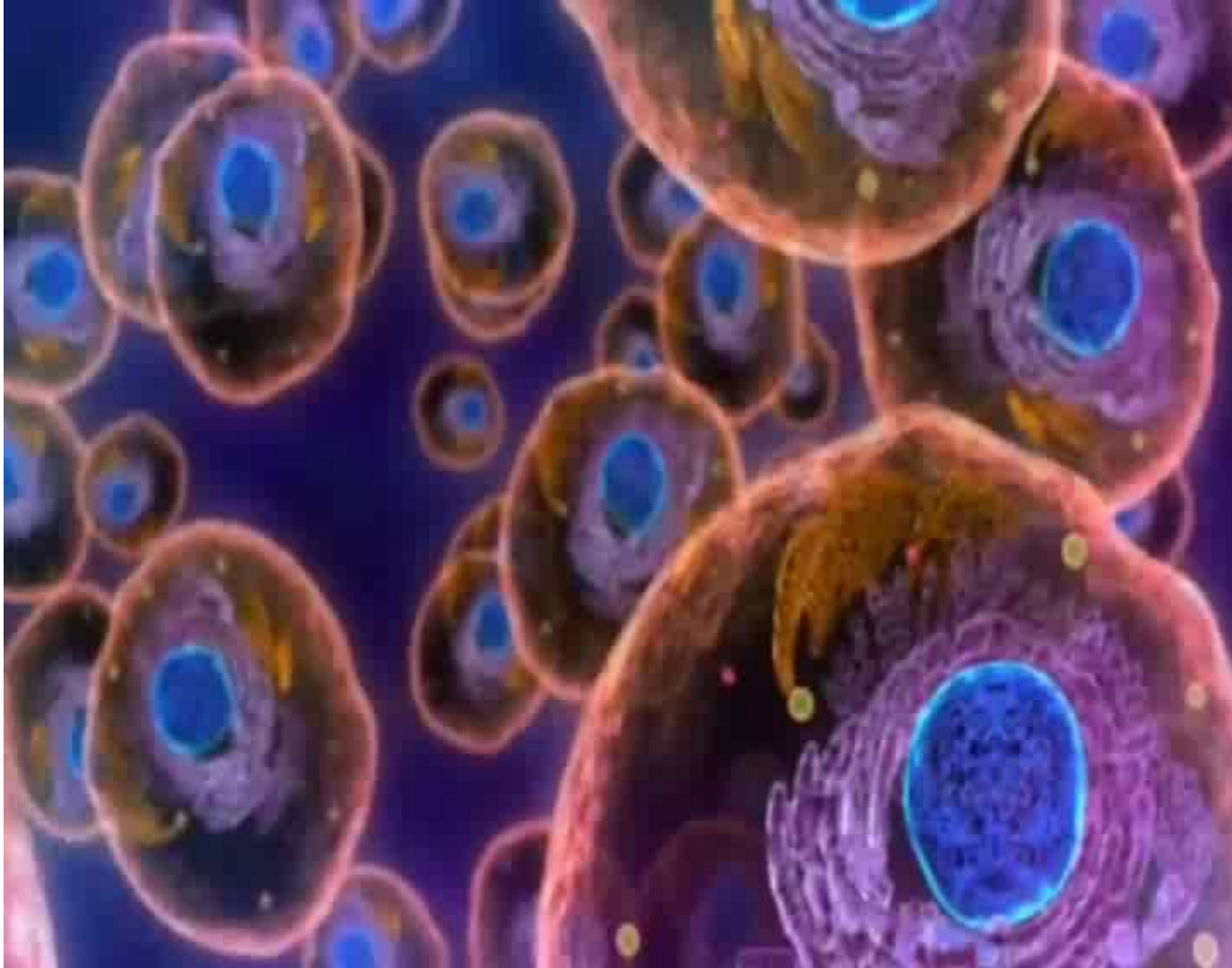
(2010...)



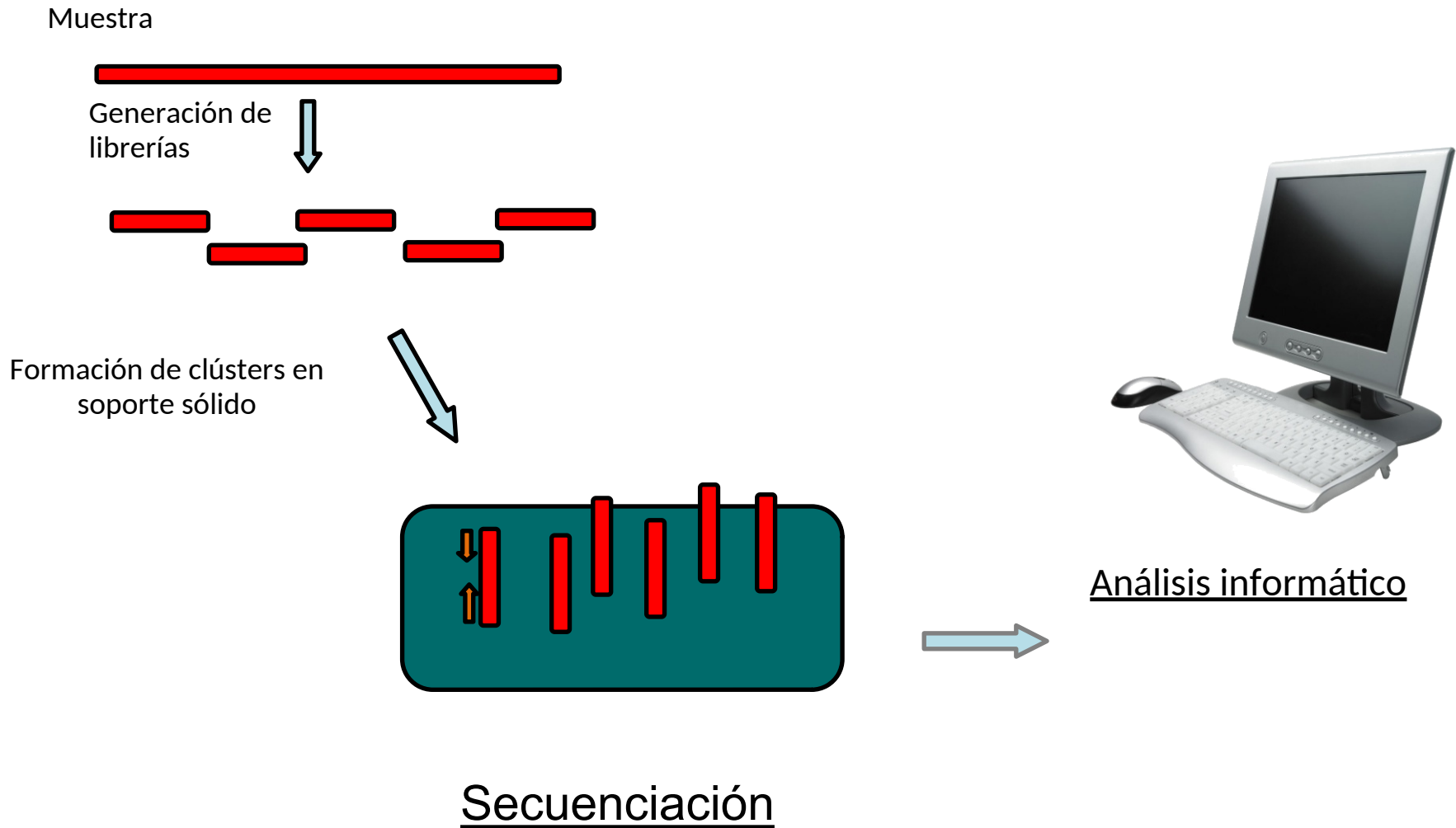
[http://en.wikipedia.org/wiki/DNA\\_sequencing](http://en.wikipedia.org/wiki/DNA_sequencing)

- Coste ~2.000 €.
- Un secuenciador necesita < 1 día

# Secuenciación de nueva generación



# Secuenciación de nueva generación



# **Tema 6: Tecnologías de análisis moleculares masivos**

6.1 Chips genéticos

6.2 Principios de la secuenciación

**6.3 Generación de librerías genómicas**

6.4 Plataformas de secuenciación

6.5 Técnicas de secuenciación de célula única

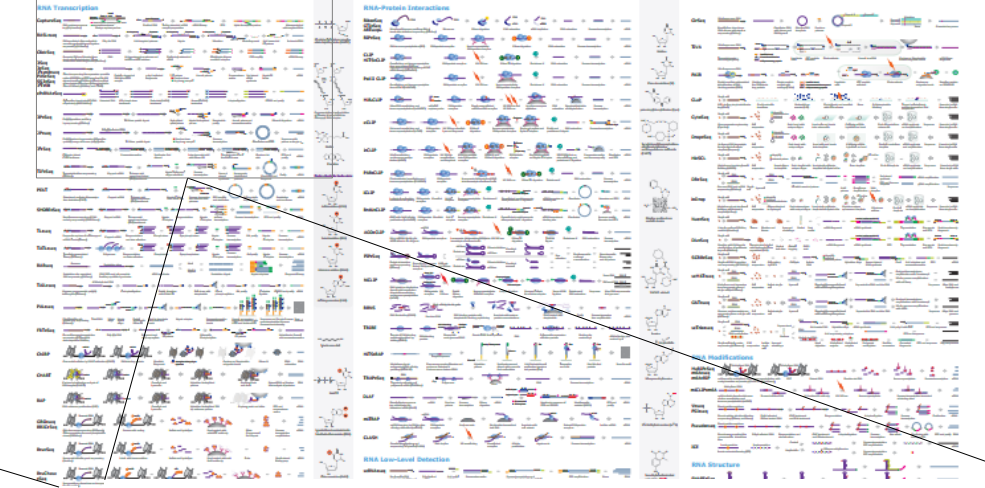
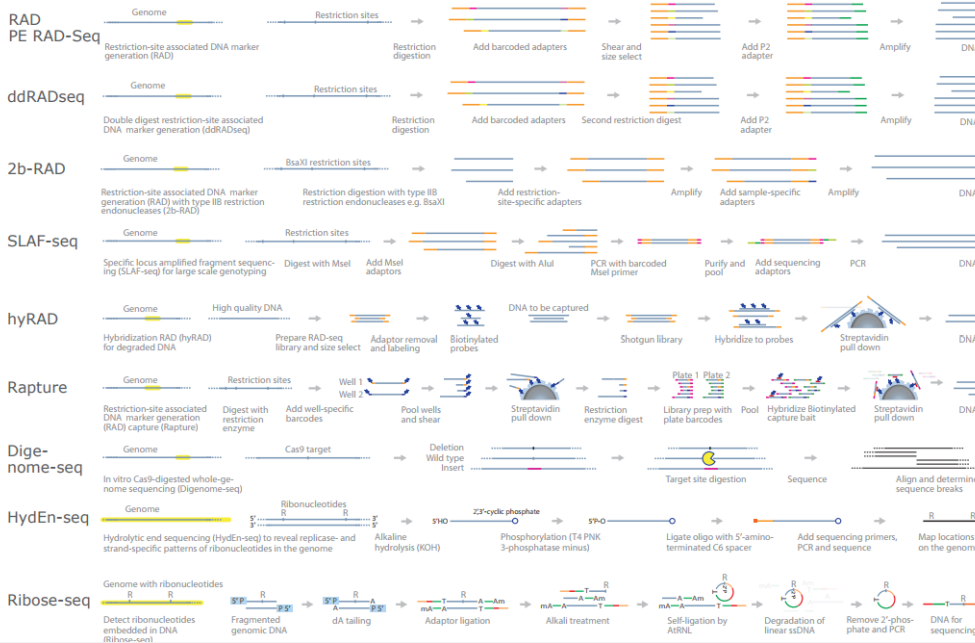
6.6 Control de calidad de la secuenciación

## Aplicaciones posibles

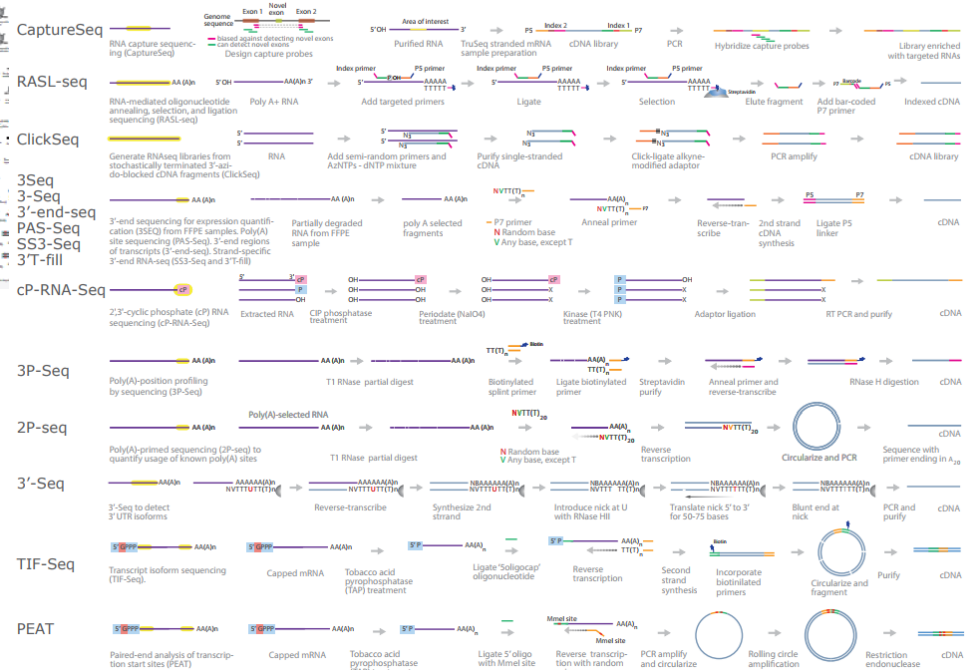
## ¡Más de 200 aplicaciones y creciendo!



## DNA Rearrangements and Markers

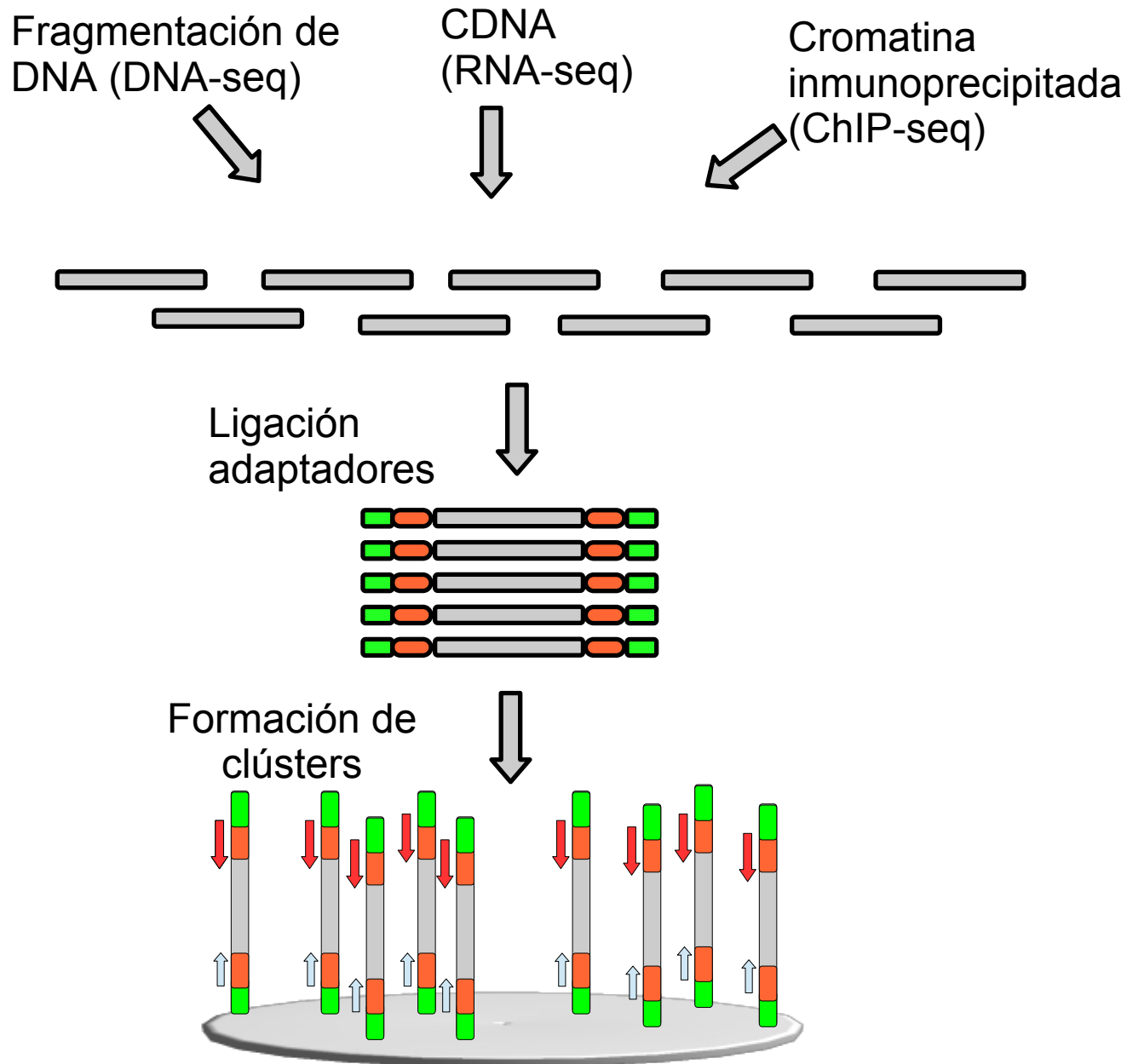


## RNA Transcription

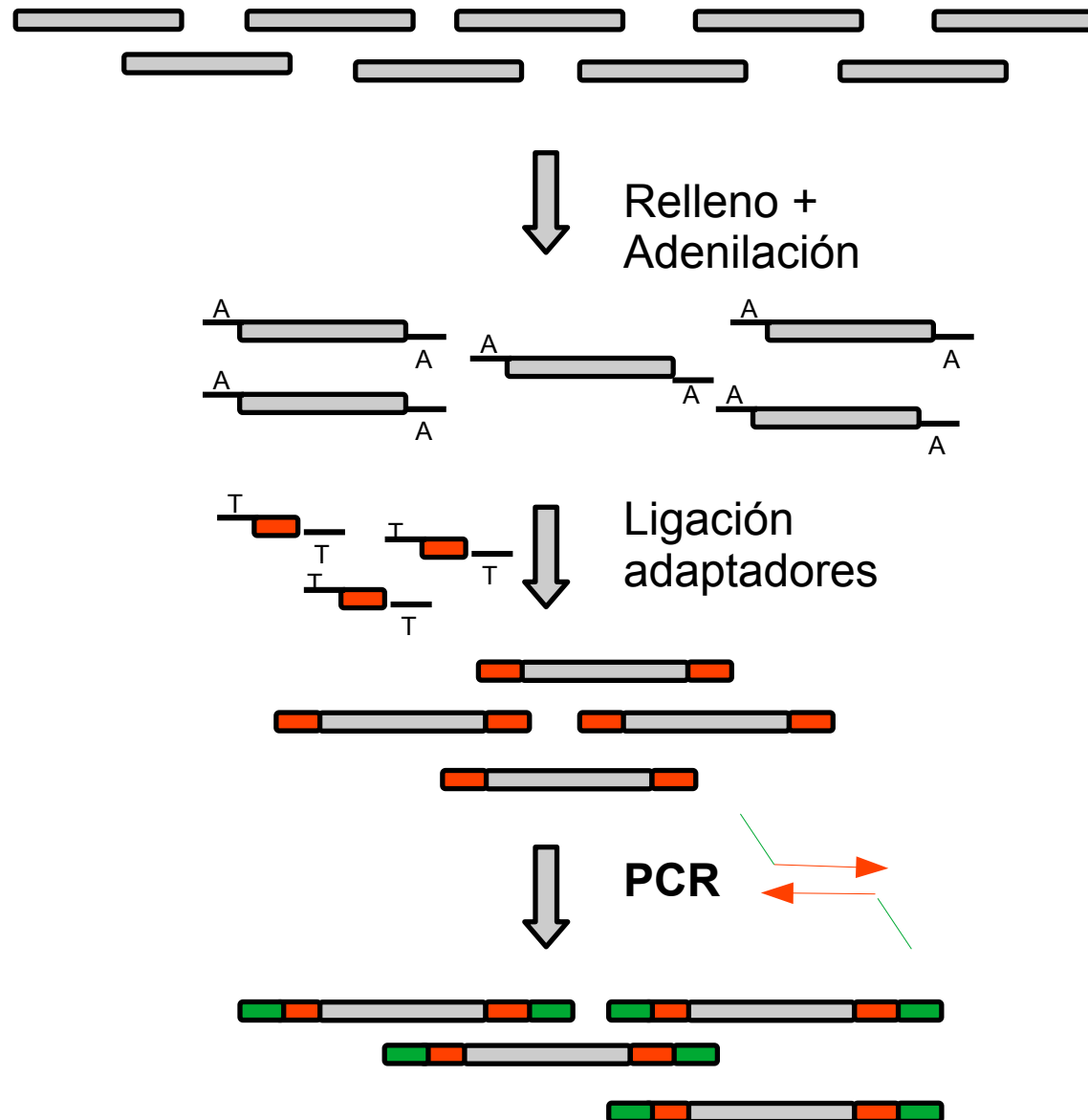




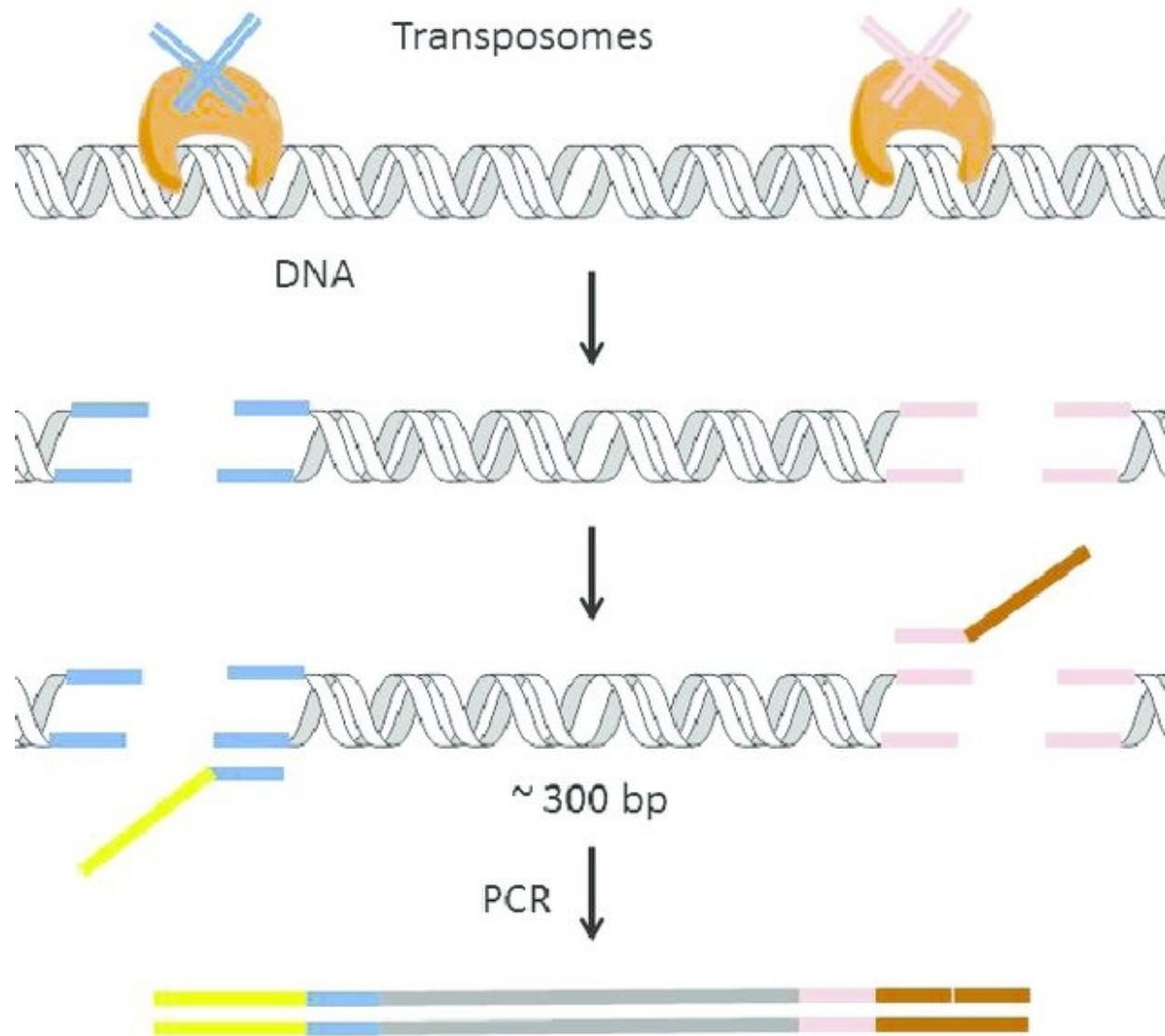
# Preparación de librerías



# Protocolo tradicional de generación de librerías

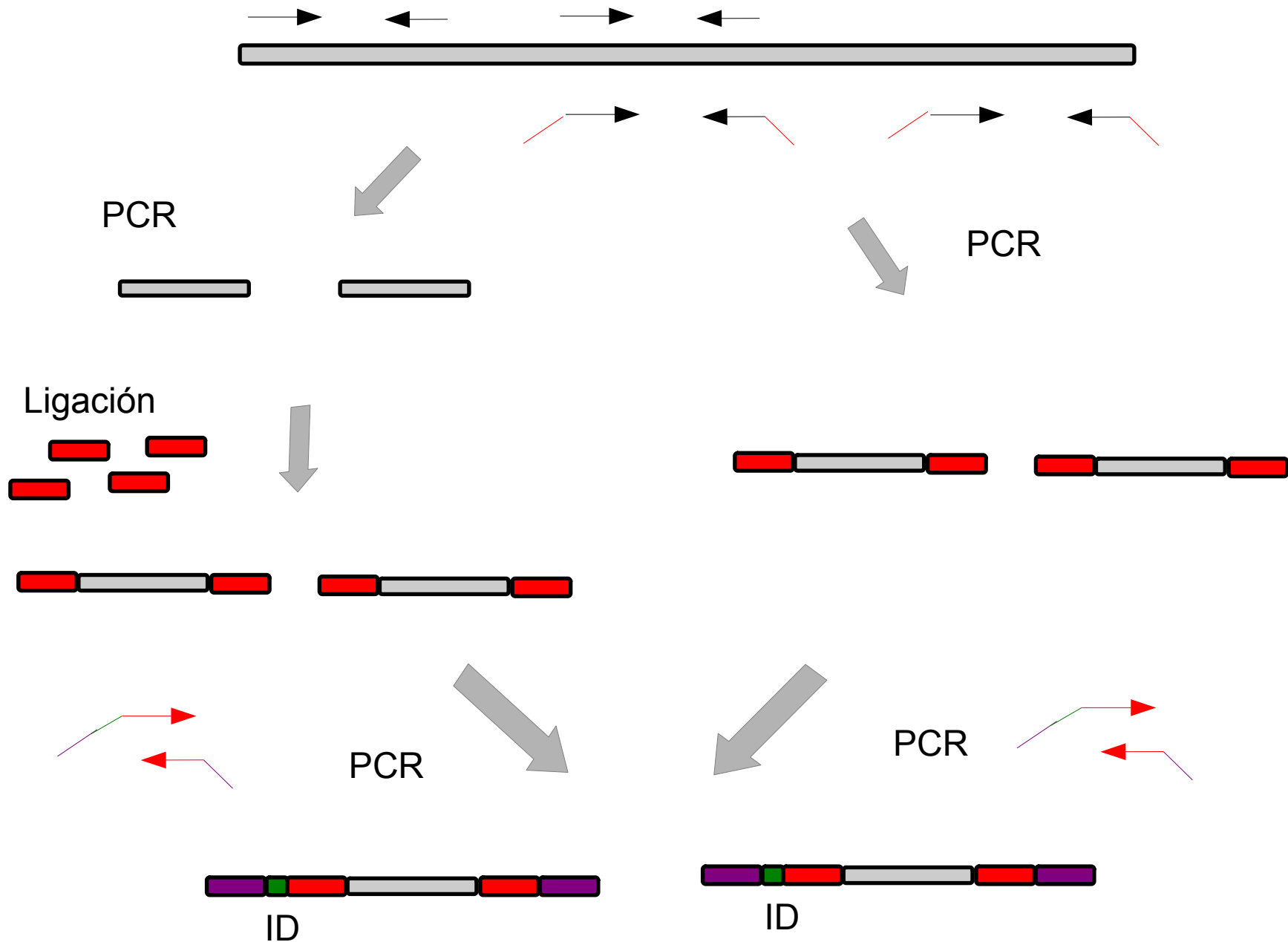


# Variante mediante el uso de una trasposasa



Head SR *et al.* Biotechniques 56:61

# Variante utilizando PCR



# **Tema 6: Tecnologías de análisis moleculares masivos**

6.1 Chips genéticos

6.2 Principios de la secuenciación

6.3 Generación de librerías genómicas

**6.4 Plataformas de secuenciación**

6.5 Técnicas de secuenciación de célula única

6.6 Control de calidad de la secuenciación



# Plataformas de secuenciación

illumina®

**High-Seq**



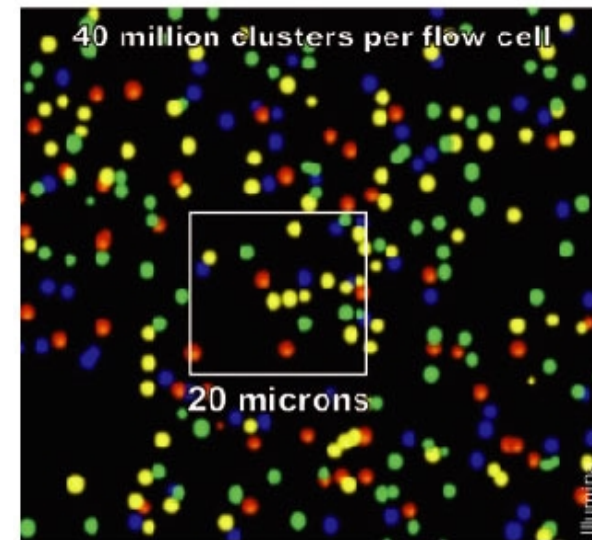
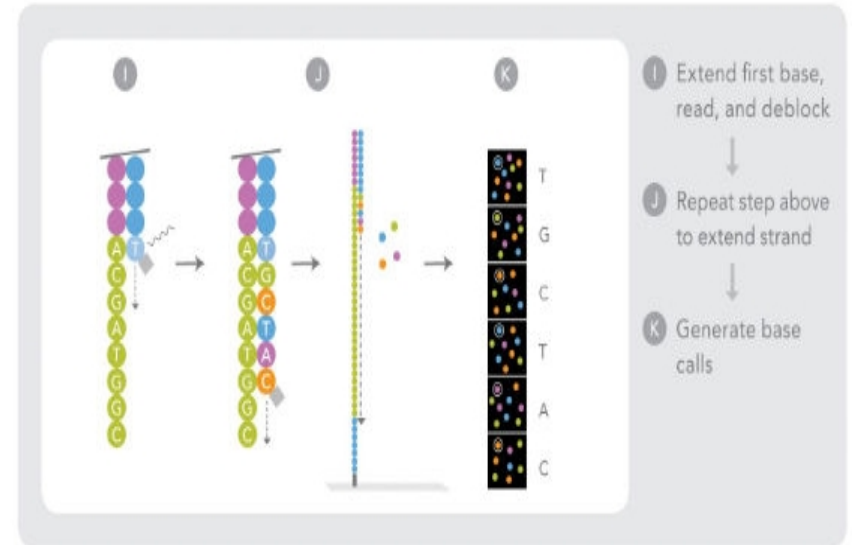
**Mi-Seq**



**Next-Seq**

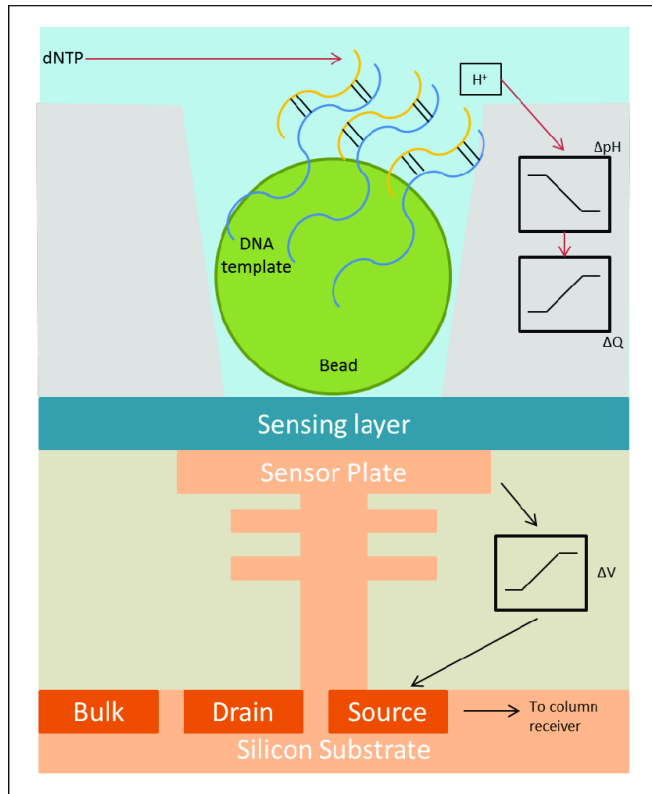


**Nova-Seq**



En cada ciclo se añade los 4 dNTPs y se valora el color de la fluorescencia

# Plataformas de secuenciación



*life* technologies™

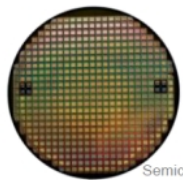
**Ion-torrent PGM**



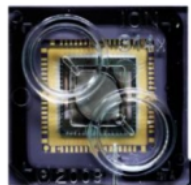
**Ion-proton**



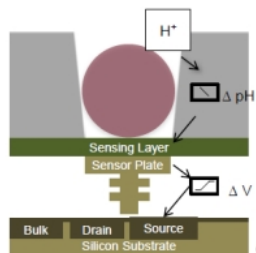
En cada ciclo se añade solo uno de los 4 dNTPs y se valora la incorporación o no en cada pocillo



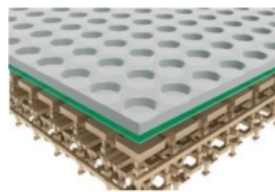
Wafer  
Semiconductor Manufacturing



Chip  
Semiconductor Packaging



Single Sensor  
Chemical to Digital Sequence



Millions of Sensors  
Semiconductor Design

# Plataformas de secuenciación

illumina®

High-Seq



Mi-Seq



Next-Seq



- Tamaño de secuencias constante.
- Mayor calidad de secuencia.
- Secuencias de menor tamaño.
- ¿Más caro?.
- Pocos kits de aplicación rápida.

life  
technologies™

Ion-torrent PGM



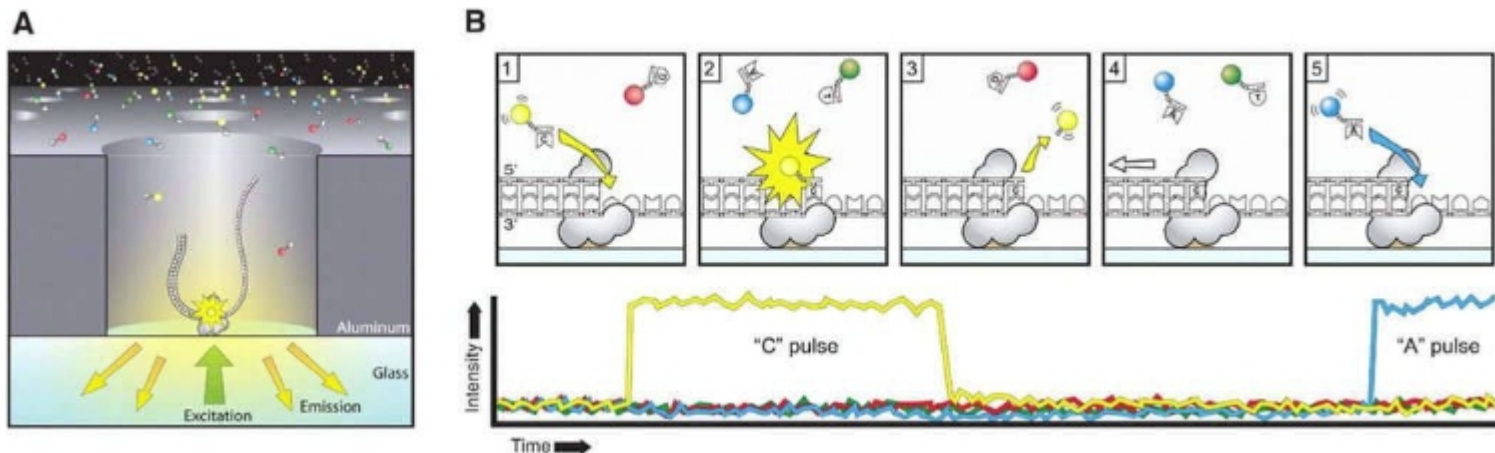
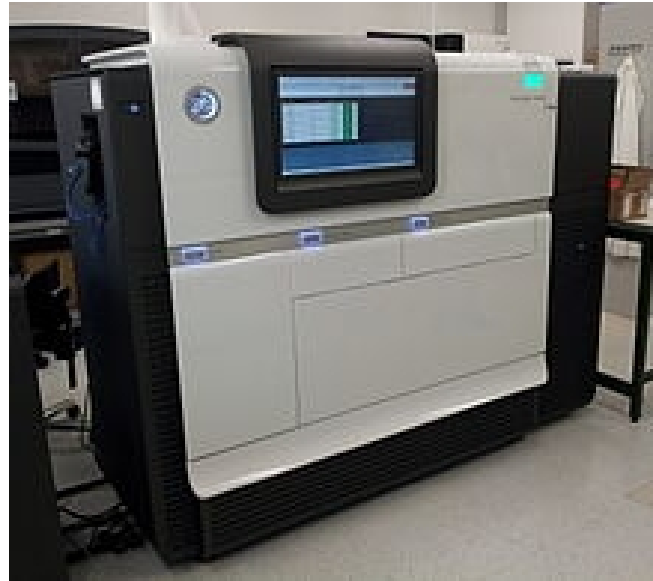
Ion-proton



- Tamaño de secuencias variable.
- Mayor longitud de secuencias.
- Problemas con repeticiones de una misma base.
- Gran cantidad de kits clínicos de aplicación rápida.

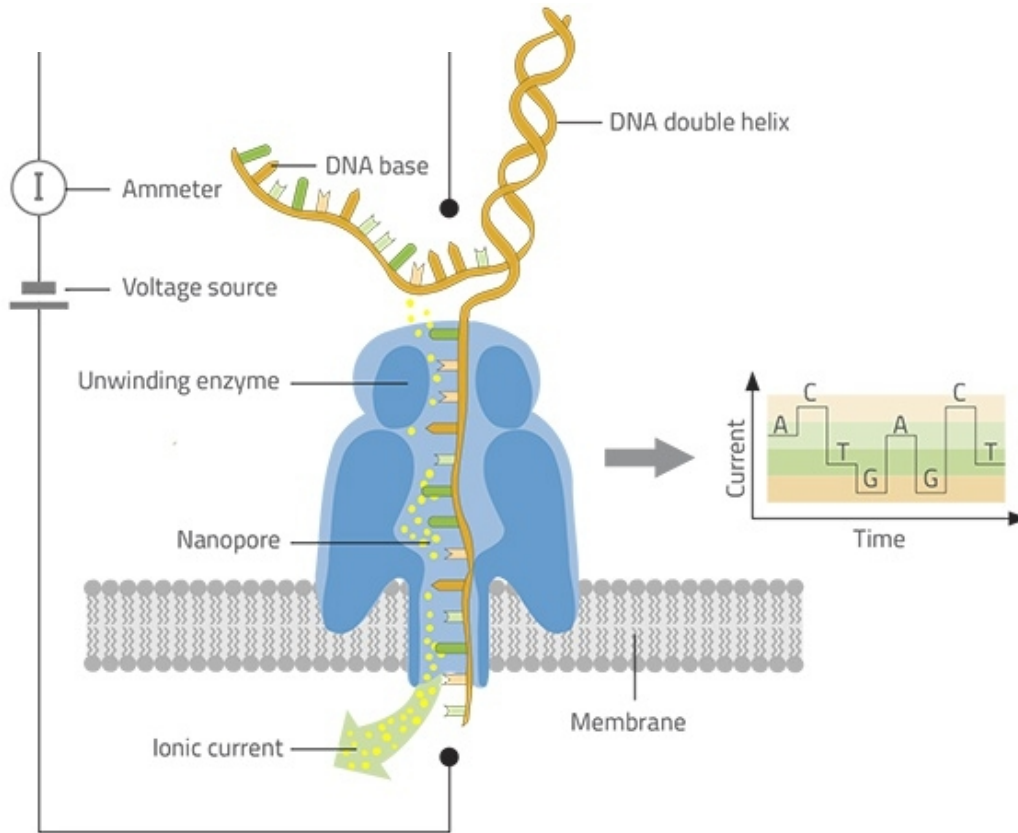
# Plataformas de secuenciación

# PacBio



- Secuencias de gran longitud.
- Baja calidad de secuencia que se compensa con la Lectura múltiple y la generación de consensos.

# Plataformas de secuenciación



<https://www.quora.com/How-does-a-nanopore-DNA-sequencer-like-Oxford-Nanopore-work>

- Secuencias de gran longitud.
- Baja calidad de secuencia.
- Gran portabilidad y no necesita gran infraestructura

# **Tema 6: Tecnologías de análisis moleculares masivos**

6.1 Chips genéticos

6.2 Principios de la secuenciación

6.3 Generación de librerías genómicas

6.4 Plataformas de secuenciación

**6.5 Técnicas de secuenciación de célula única**

6.6 Control de calidad de la secuenciación



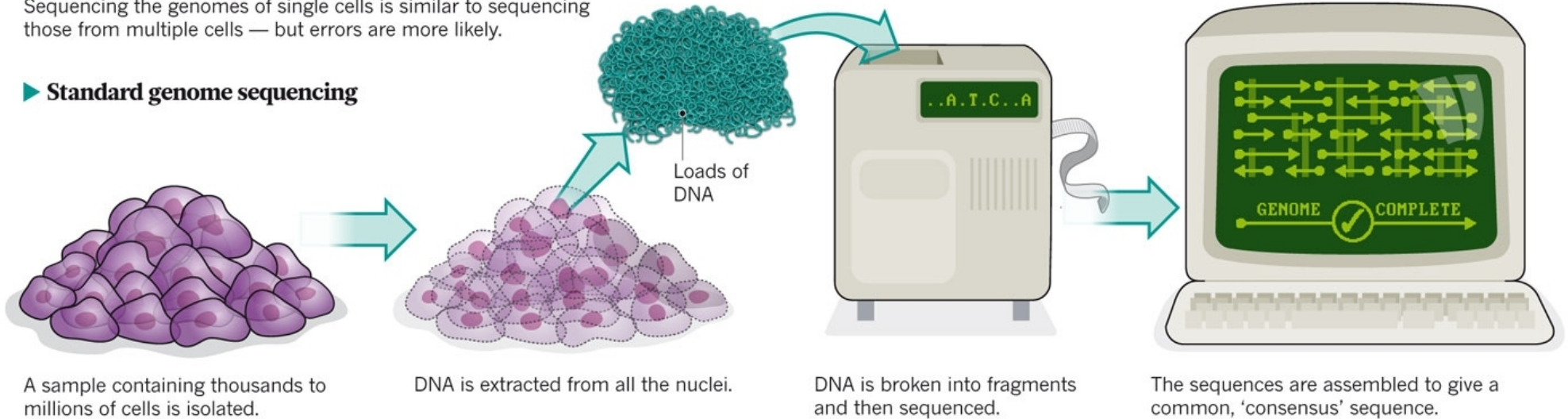
# El problema de las muestras heterogéneas

En las técnicas de secuenciación tradicionales, solo podemos capturar las alteraciones presentes en un porcentaje importante de la población de células

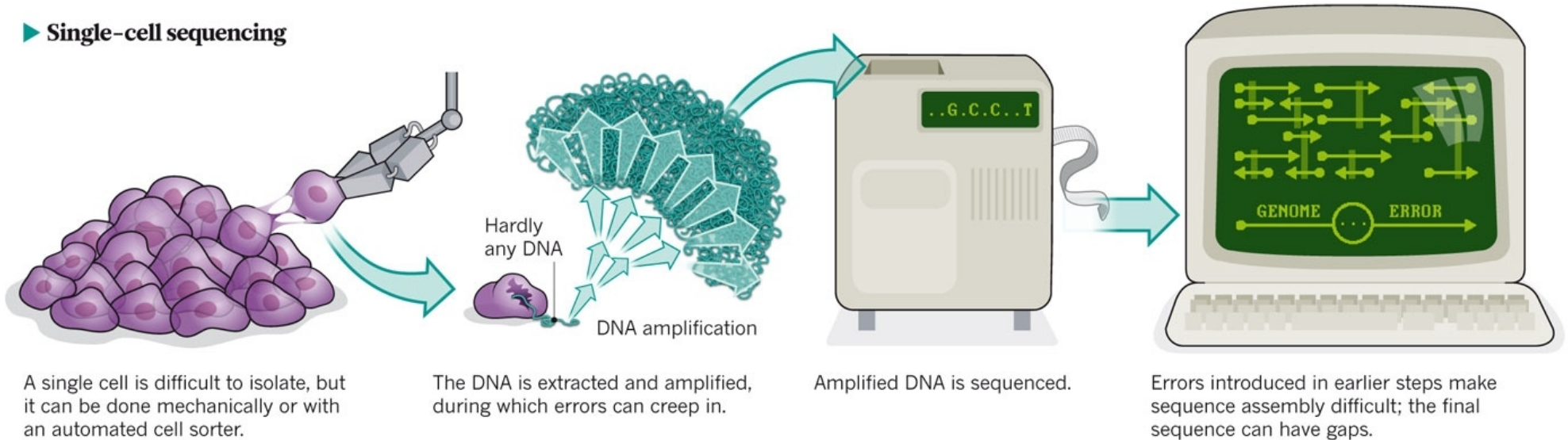
## ONE GENOME FROM MANY

Sequencing the genomes of single cells is similar to sequencing those from multiple cells — but errors are more likely.

### ► Standard genome sequencing

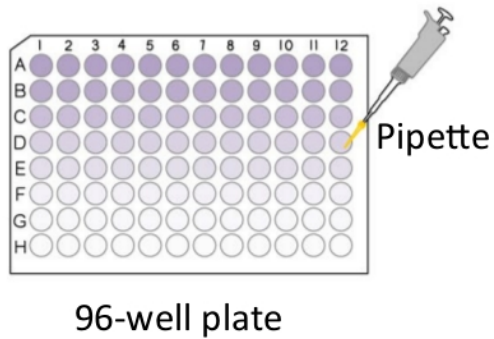


### ► Single-cell sequencing

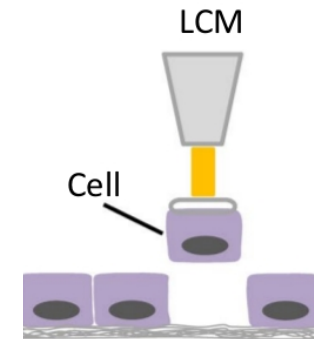
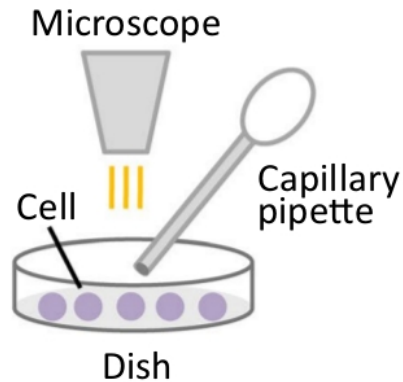


# Aislamiento de células únicas

## Métodos manuales



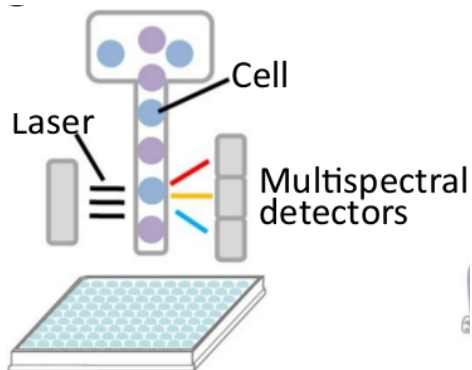
### *Dilución seriada*



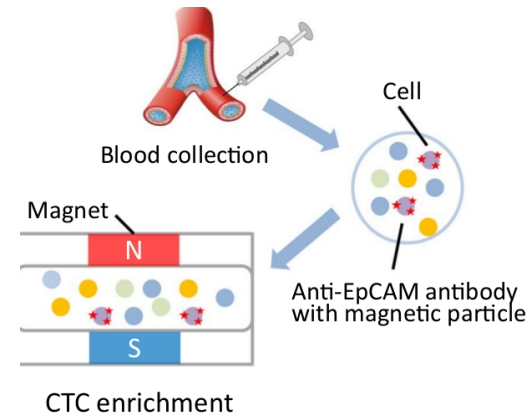
### *Selección manual con microdissección*

Wang J *et al.* Clin Trans Med 2017

## Métodos automáticos



### *FACS*

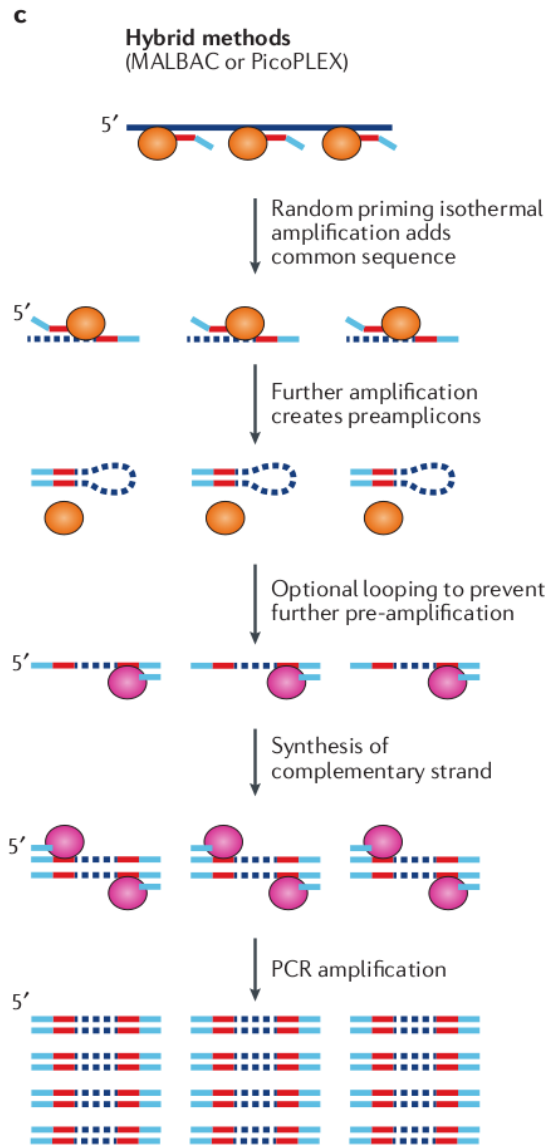


### *Aislamiento magnético*

# Amplificación de ácidos nucleicos

**UMI:** Unique molecular identifier

## Amplificación de DNA

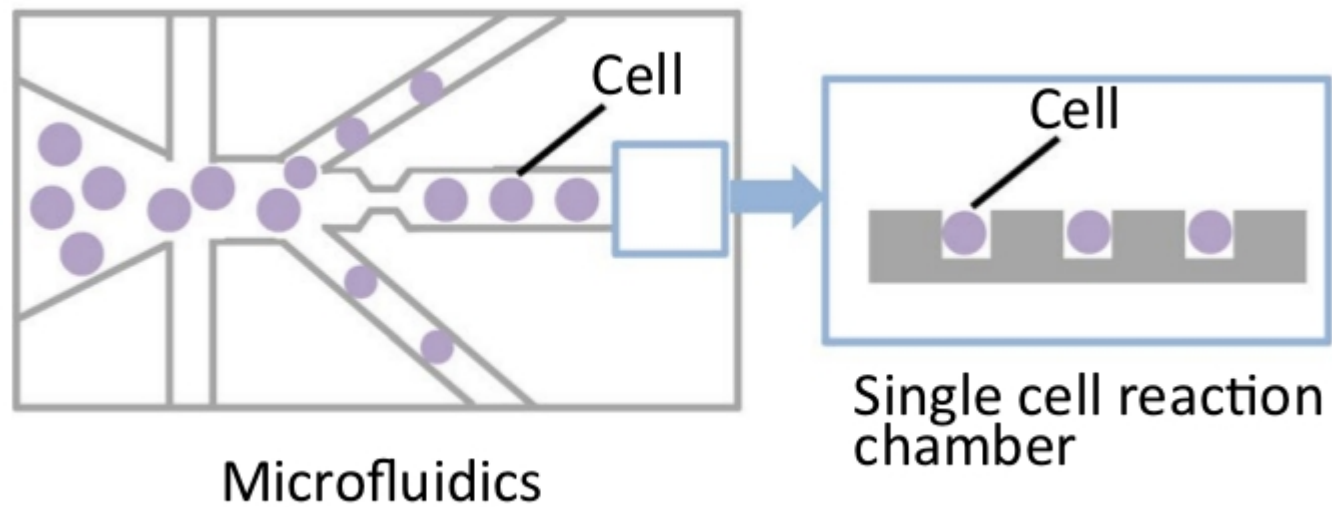
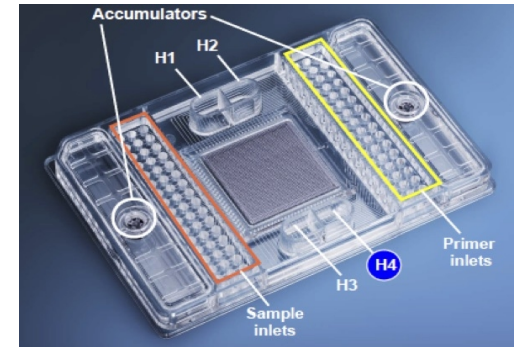
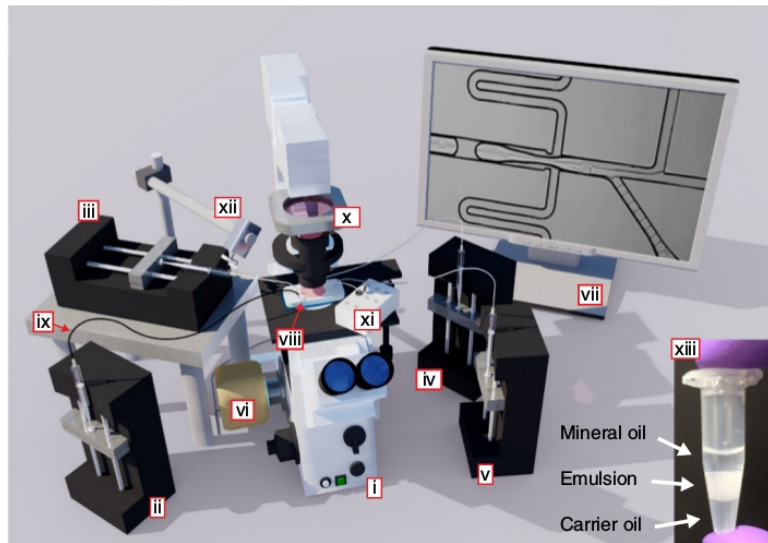


## Amplificación de RNA

|                                       | A       | B        | C          |
|---------------------------------------|---------|----------|------------|
|                                       | CEL-seq | STRT-seq | SMART-seq2 |
| polyA+ mRNA capture                   |         |          |            |
| Reverse transcription                 |         |          |            |
| cDNA amplification                    |         |          |            |
| Fragmentation and library preparation |         |          |            |
| Sequencing                            |         |          |            |
| Coverage                              |         |          |            |
| UMI possible                          | ✓       | ✓        | ✗          |
| Strand information                    | ✓       | ✓        | ✗          |

Gawad C et al. Nat Review 2016

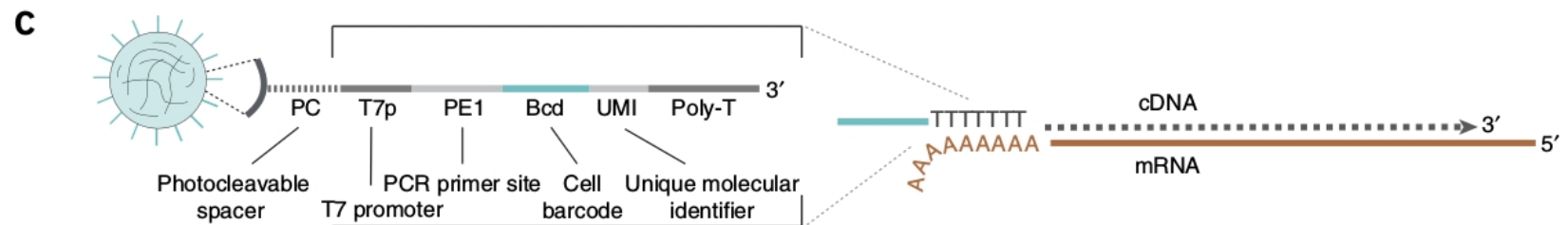
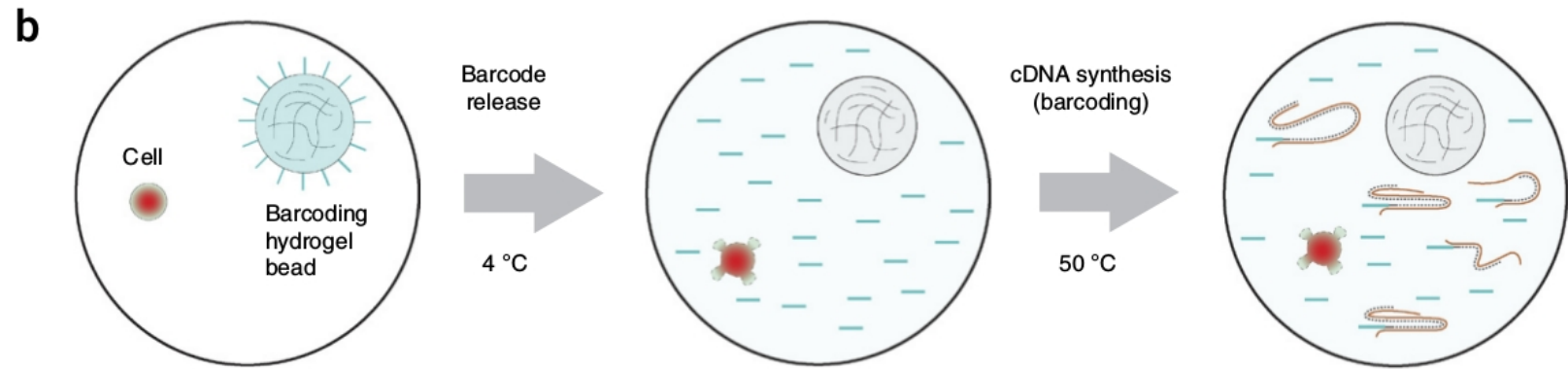
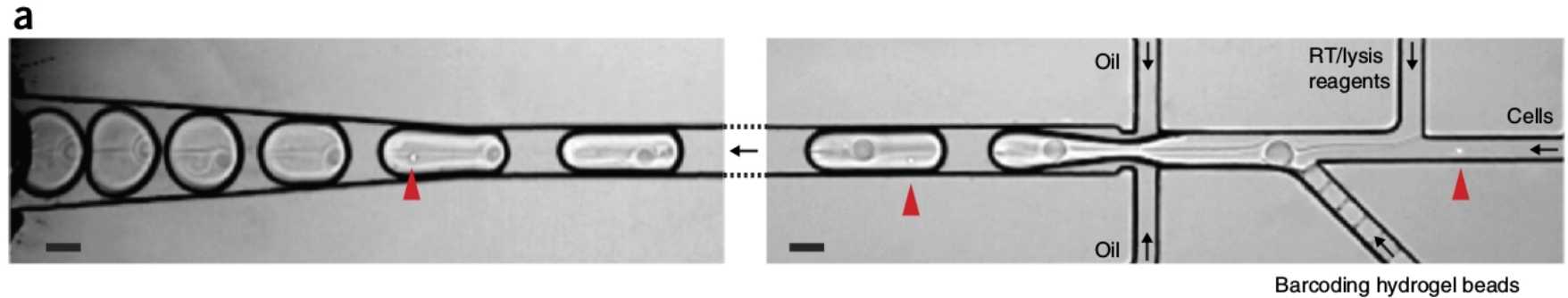
# Microfluídica



Wang J *et al.* Clin Trans Med 2017



# Secuenciación de célula única basada en la generación de droplets



# Plataformas comerciales

**10x**  
GENOMICS®

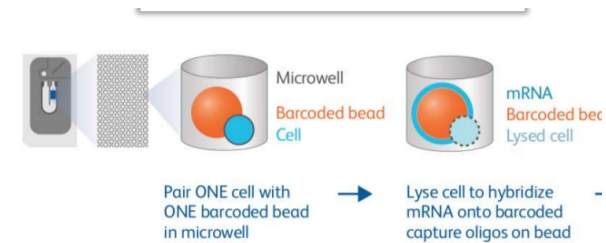
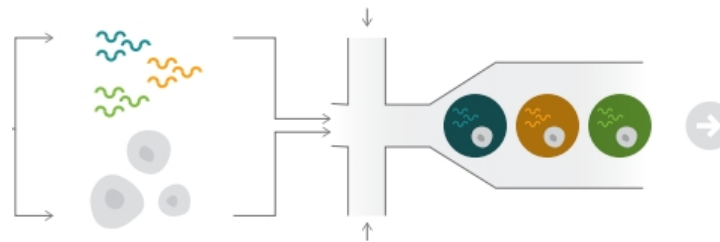
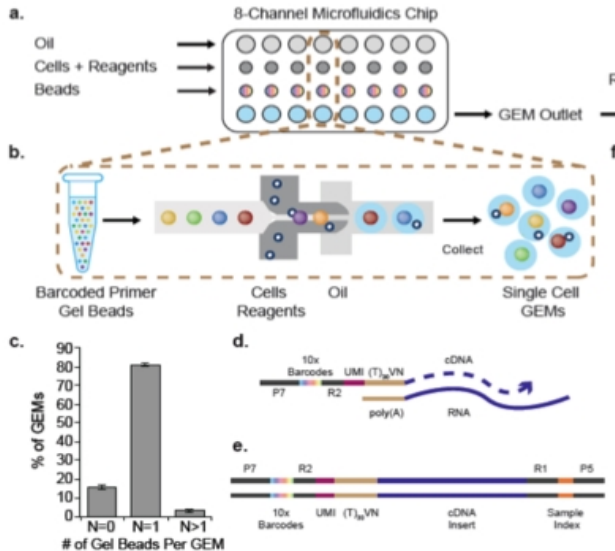


**BIO-RAD**

**ddSEQ Single-Cell Isolator**



**BD Rhapsody™**





# **Tema 6: Tecnologías de análisis moleculares masivos**

6.1 Chips genéticos

6.2 Principios de la secuenciación

6.3 Generación de librerías genómicas

6.4 Plataformas de secuenciación

6.5 Técnicas de secuenciación de célula única

**6.6 Control de calidad de la secuenciación**

# Archivos con la información de secuenciación

## Formato de secuencia: Fastq

```
@M03913:60:000000000-D46KB:1:1101:19075:1885 1:N:0:TTAGTCAT
CATACGAGATCGGAAGAGCGGTTCAGCAGGAATTCCGAGACCGATCTCTTAGCCATTCTCTTATTCCCTCTTCT
+
11>1>1111>D11?AEF0B00AAA12110000B11A0////////A/B2DF111A1B1@2B@2@1B222110>B1@
M03913:60:000000000-D46KB:1:1101:13680:1886 1:N:0:TTAGGCAT
TGGATGTATTTGGAAGAGAAAGATCGGAAGAGCGGTTCAGCAGGAATGCCGTGACCGATCTCTTAGTCATTCTCT
+
>111>BFFFFFFFGBA1111B1BE0000000A00BAA12110//0B11A/////1///AB@@F112B2B2@@@2
M03913:60:000000000-D46KB:1:1101:14309:1891 1:N:0:TTAGGCAT
TCTCCCTTCTCAGGATTCTACAGGTATCAAGTCGTATTTGATGGTGAAACCTGTCTCTTGTATATTCTCTACAC
+
>1>>>CFFFFFFDB11BF13F3B11B113A331D10A00BD222211111A0B0DABABD111D2DD2B222110
M03913:60:000000000-D46KB:1:1101:16455:1892 1:N:0:TTAGGCAT
ATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCTTAGGCATATCTCGTATTCGGTCTTCTGCTTTTCA
+
1>>>1A1111CA?EGFGBAA0000B11B000/AA//AAADFG1111B1BBDA0B/B12@0BBB?@E11BB1122
M03913:60:000000000-D46KB:1:1101:12318:1895 1:N:0:TTAGGCAT
TTGTGTTTAATTAGGGGTTGTGAGATCCGAAGATCGTTTCATCATTAATTCCGTTACCTTTCTCTTATTCATTTC
+
>>1>1BCFD3DFD1D1EEAAE0313DB10000011000BB22A22221A2A1011//111B1A2D212222B2A2
M03913:60:000000000-D46KB:1:1101:15198:1899 1:N:0:TTAGGCAT
TTTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGATACCGTTCTCTTAGTCATATCTCGTTTGCCGTCTTCTTCT
+
1>>111BD111111AFCCCEEC1B110000B1110//A1//A0A2BF112A222A2D20A/B/11/B>0?2B12@
M03913:60:000000000-D46KB:1:1101:15952:1910 1:N:0:TTAGGCAT
TTGTGTTTAATTAGGGGTTGTGAGATCGGAAGAGCGGTTCGCAGGAATTCCGAGTCCGATCTCTTAGTCTTTTC
```

- ➔ Nombre único (@)
- ➔ Secuencia
- ➔ Línea libre (comentarios, +)
- ➔ Calidad de secuencia

Seguridad de la  
secuencia en esa  
posición. Escala Phred  
logarítmica codificada  
con ASCII - 32 / ASCII  
- 64

# Escala Phred

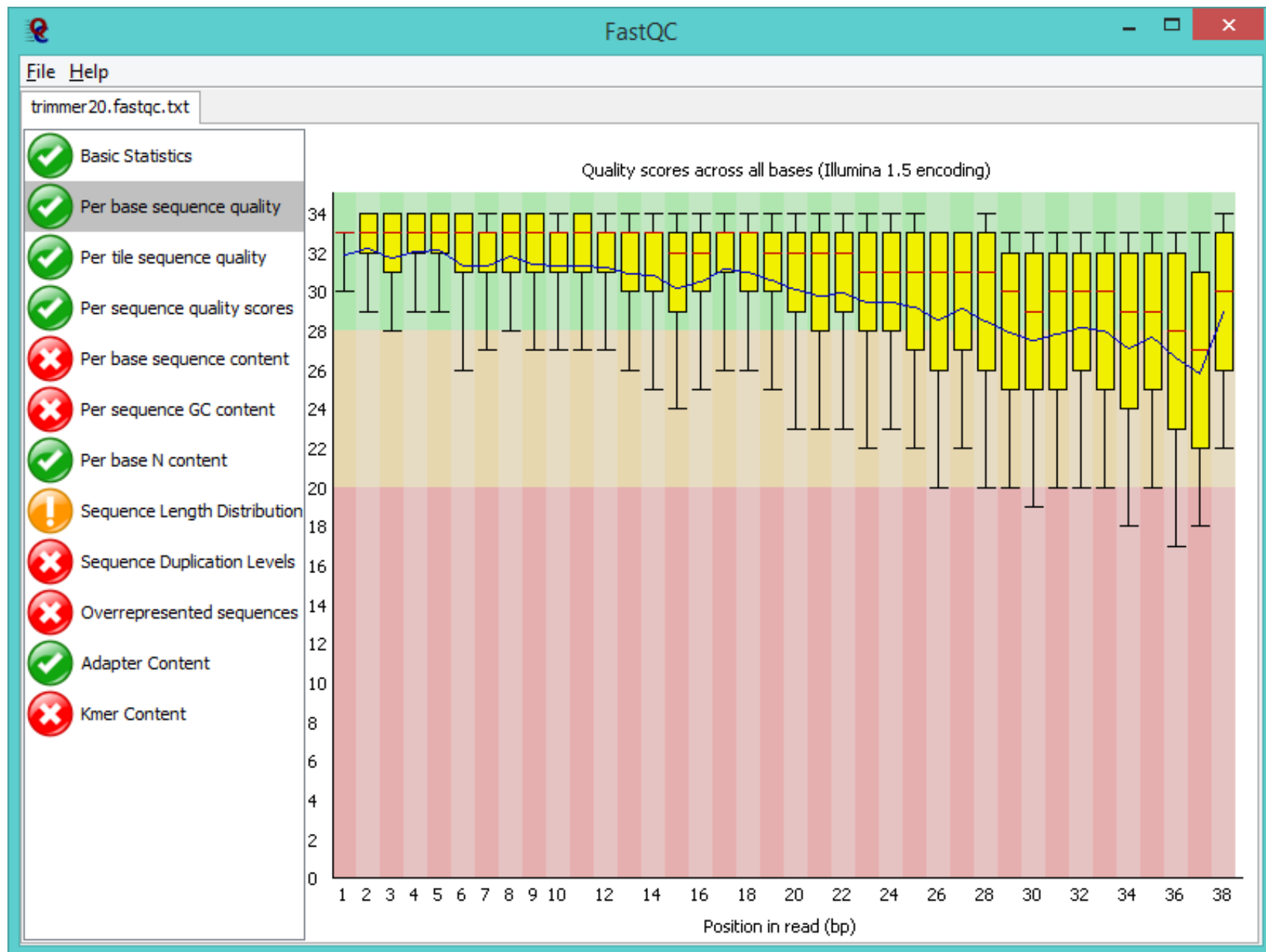
10 → 90 % seguridad.  
 20 → 99 % seguridad.  
 30 → 99.9 % seguridad

....

| Dec | Hx | Oct | Char                               | Dec | Hx | Oct | Html  | Chr          | Dec | Hx | Oct | Html  | Chr      | Dec | Hx | Oct | Html   | Chr        |
|-----|----|-----|------------------------------------|-----|----|-----|-------|--------------|-----|----|-----|-------|----------|-----|----|-----|--------|------------|
| 0   | 0  | 000 | <b>NUL</b> (null)                  | 32  | 20 | 040 | &#32; | <b>Space</b> | 64  | 40 | 100 | &#64; | <b>@</b> | 96  | 60 | 140 | &#96;  | <b>`</b>   |
| 1   | 1  | 001 | <b>SOH</b> (start of heading)      | 33  | 21 | 041 | &#33; | <b>!</b>     | 65  | 41 | 101 | &#65; | <b>A</b> | 97  | 61 | 141 | &#97;  | <b>a</b>   |
| 2   | 2  | 002 | <b>STX</b> (start of text)         | 34  | 22 | 042 | &#34; | <b>"</b>     | 66  | 42 | 102 | &#66; | <b>B</b> | 98  | 62 | 142 | &#98;  | <b>b</b>   |
| 3   | 3  | 003 | <b>ETX</b> (end of text)           | 35  | 23 | 043 | &#35; | <b>#</b>     | 67  | 43 | 103 | &#67; | <b>C</b> | 99  | 63 | 143 | &#99;  | <b>c</b>   |
| 4   | 4  | 004 | <b>EOT</b> (end of transmission)   | 36  | 24 | 044 | &#36; | <b>\$</b>    | 68  | 44 | 104 | &#68; | <b>D</b> | 100 | 64 | 144 | &#100; | <b>d</b>   |
| 5   | 5  | 005 | <b>ENQ</b> (enquiry)               | 37  | 25 | 045 | &#37; | <b>%</b>     | 69  | 45 | 105 | &#69; | <b>E</b> | 101 | 65 | 145 | &#101; | <b>e</b>   |
| 6   | 6  | 006 | <b>ACK</b> (acknowledge)           | 38  | 26 | 046 | &#38; | <b>&amp;</b> | 70  | 46 | 106 | &#70; | <b>F</b> | 102 | 66 | 146 | &#102; | <b>f</b>   |
| 7   | 7  | 007 | <b>BEL</b> (bell)                  | 39  | 27 | 047 | &#39; | <b>'</b>     | 71  | 47 | 107 | &#71; | <b>G</b> | 103 | 67 | 147 | &#103; | <b>g</b>   |
| 8   | 8  | 010 | <b>BS</b> (backspace)              | 40  | 28 | 050 | &#40; | <b>(</b>     | 72  | 48 | 110 | &#72; | <b>H</b> | 104 | 68 | 150 | &#104; | <b>h</b>   |
| 9   | 9  | 011 | <b>TAB</b> (horizontal tab)        | 41  | 29 | 051 | &#41; | <b>)</b>     | 73  | 49 | 111 | &#73; | <b>I</b> | 105 | 69 | 151 | &#105; | <b>i</b>   |
| 10  | A  | 012 | <b>LF</b> (NL line feed, new line) | 42  | 2A | 052 | &#42; | <b>*</b>     | 74  | 4A | 112 | &#74; | <b>J</b> | 106 | 6A | 152 | &#106; | <b>j</b>   |
| 11  | B  | 013 | <b>VT</b> (vertical tab)           | 43  | 2B | 053 | &#43; | <b>+</b>     | 75  | 4B | 113 | &#75; | <b>K</b> | 107 | 6B | 153 | &#107; | <b>k</b>   |
| 12  | C  | 014 | <b>FF</b> (NP form feed, new page) | 44  | 2C | 054 | &#44; | <b>,</b>     | 76  | 4C | 114 | &#76; | <b>L</b> | 108 | 6C | 154 | &#108; | <b>l</b>   |
| 13  | D  | 015 | <b>CR</b> (carriage return)        | 45  | 2D | 055 | &#45; | <b>-</b>     | 77  | 4D | 115 | &#77; | <b>M</b> | 109 | 6D | 155 | &#109; | <b>m</b>   |
| 14  | E  | 016 | <b>SO</b> (shift out)              | 46  | 2E | 056 | &#46; | <b>.</b>     | 78  | 4E | 116 | &#78; | <b>N</b> | 110 | 6E | 156 | &#110; | <b>n</b>   |
| 15  | F  | 017 | <b>SI</b> (shift in)               | 47  | 2F | 057 | &#47; | <b>/</b>     | 79  | 4F | 117 | &#79; | <b>O</b> | 111 | 6F | 157 | &#111; | <b>o</b>   |
| 16  | 10 | 020 | <b>DLE</b> (data link escape)      | 48  | 30 | 060 | &#48; | <b>0</b>     | 80  | 50 | 120 | &#80; | <b>P</b> | 112 | 70 | 160 | &#112; | <b>p</b>   |
| 17  | 11 | 021 | <b>DC1</b> (device control 1)      | 49  | 31 | 061 | &#49; | <b>1</b>     | 81  | 51 | 121 | &#81; | <b>Q</b> | 113 | 71 | 161 | &#113; | <b>q</b>   |
| 18  | 12 | 022 | <b>DC2</b> (device control 2)      | 50  | 32 | 062 | &#50; | <b>2</b>     | 82  | 52 | 122 | &#82; | <b>R</b> | 114 | 72 | 162 | &#114; | <b>r</b>   |
| 19  | 13 | 023 | <b>DC3</b> (device control 3)      | 51  | 33 | 063 | &#51; | <b>3</b>     | 83  | 53 | 123 | &#83; | <b>S</b> | 115 | 73 | 163 | &#115; | <b>s</b>   |
| 20  | 14 | 024 | <b>DC4</b> (device control 4)      | 52  | 34 | 064 | &#52; | <b>4</b>     | 84  | 54 | 124 | &#84; | <b>T</b> | 116 | 74 | 164 | &#116; | <b>t</b>   |
| 21  | 15 | 025 | <b>NAK</b> (negative acknowledge)  | 53  | 35 | 065 | &#53; | <b>5</b>     | 85  | 55 | 125 | &#85; | <b>U</b> | 117 | 75 | 165 | &#117; | <b>u</b>   |
| 22  | 16 | 026 | <b>SYN</b> (synchronous idle)      | 54  | 36 | 066 | &#54; | <b>6</b>     | 86  | 56 | 126 | &#86; | <b>V</b> | 118 | 76 | 166 | &#118; | <b>v</b>   |
| 23  | 17 | 027 | <b>ETB</b> (end of trans. block)   | 55  | 37 | 067 | &#55; | <b>7</b>     | 87  | 57 | 127 | &#87; | <b>W</b> | 119 | 77 | 167 | &#119; | <b>w</b>   |
| 24  | 18 | 030 | <b>CAN</b> (cancel)                | 56  | 38 | 070 | &#56; | <b>8</b>     | 88  | 58 | 130 | &#88; | <b>X</b> | 120 | 78 | 170 | &#120; | <b>x</b>   |
| 25  | 19 | 031 | <b>EM</b> (end of medium)          | 57  | 39 | 071 | &#57; | <b>9</b>     | 89  | 59 | 131 | &#89; | <b>Y</b> | 121 | 79 | 171 | &#121; | <b>y</b>   |
| 26  | 1A | 032 | <b>SUB</b> (substitute)            | 58  | 3A | 072 | &#58; | <b>:</b>     | 90  | 5A | 132 | &#90; | <b>Z</b> | 122 | 7A | 172 | &#122; | <b>z</b>   |
| 27  | 1B | 033 | <b>ESC</b> (escape)                | 59  | 3B | 073 | &#59; | <b>;</b>     | 91  | 5B | 133 | &#91; | <b>[</b> | 123 | 7B | 173 | &#123; | <b>{</b>   |
| 28  | 1C | 034 | <b>FS</b> (file separator)         | 60  | 3C | 074 | &#60; | <b>&lt;</b>  | 92  | 5C | 134 | &#92; | <b>\</b> | 124 | 7C | 174 | &#124; | <b> </b>   |
| 29  | 1D | 035 | <b>GS</b> (group separator)        | 61  | 3D | 075 | &#61; | <b>=</b>     | 93  | 5D | 135 | &#93; | <b>]</b> | 125 | 7D | 175 | &#125; | <b>}</b>   |
| 30  | 1E | 036 | <b>RS</b> (record separator)       | 62  | 3E | 076 | &#62; | <b>&gt;</b>  | 94  | 5E | 136 | &#94; | <b>^</b> | 126 | 7E | 176 | &#126; | <b>~</b>   |
| 31  | 1F | 037 | <b>US</b> (unit separator)         | 63  | 3F | 077 | &#63; | <b>?</b>     | 95  | 5F | 137 | &#95; | <b>_</b> | 127 | 7F | 177 | &#127; | <b>DEL</b> |

Source: [www.asciitable.com](http://www.asciitable.com)

# FASTQC



# Bioinformática y análisis de datos ómicos

