

Genomeeting 2016

Análisis de datos RNA-Seq



**Introducción a las tecnologías NGS
(Next Generation Sequencing)**

M. en C. Israel Aguilar Ordóñez (iaquilar@wintergenomics.com)



Sesión 1

- Secuenciación de ácidos nucleicos
 - Secuenciación Masiva
- Construcción de bibliotecas RNA-Seq



CONTENIDO

- Fundamentos de Secuenciación
- Tecnologías NGS
- RNA-Seq

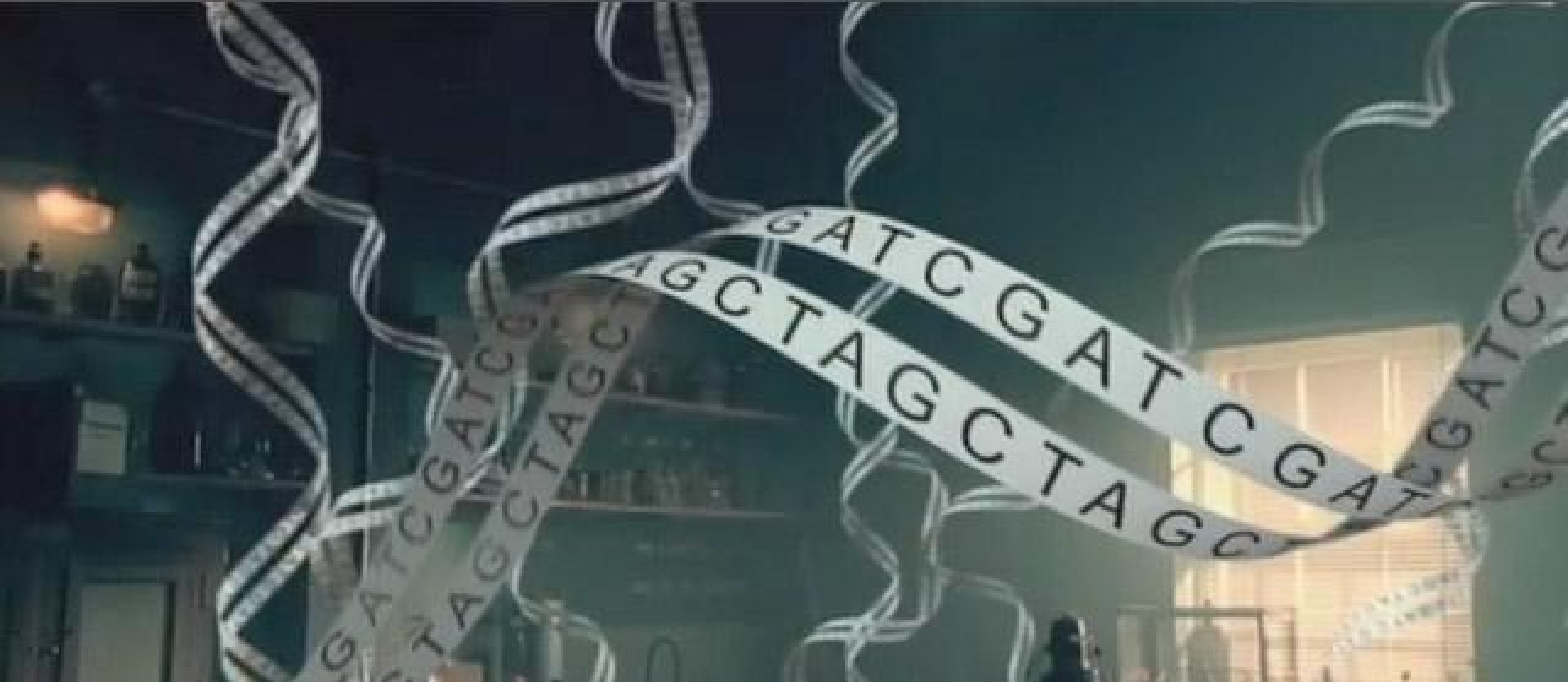


CONTENIDO

- Al finalizar la sesión entenderemos:

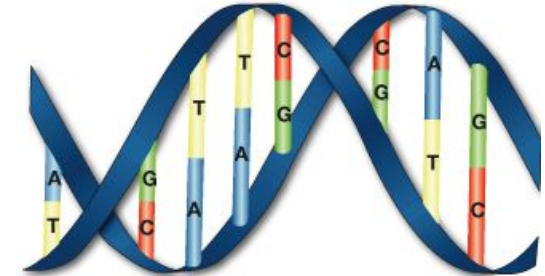
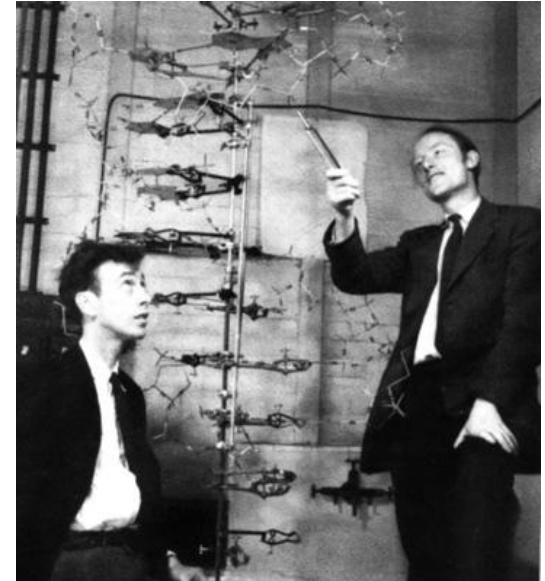
¿Cómo funciona un secuenciador NGS?

¿Cómo se construye una biblioteca para RNA-Seq?



FUNDAMENTOS DE SECUENCIACIÓN DE ÁCIDOS NUCLEICOS

El trabajo de Watson y Crick sentó las bases conceptuales para la secuenciación de DNA, al establecer el concepto de **complementariedad de bases**.



Propuestas de Watson & Crick, 1953:

- «...si se conociera el orden [de los nucleótidos] de una de las cadenas, uno podría anotar el orden exacto de las bases en la otra, debido al apareamiento específico...»
- «...en una molécula [de DNA] extensa muchas permutaciones son posibles, por tanto parece probable que la secuencia precisa de las bases es el código que transmite la información genética...»

De lo anterior se puede inferir que:

- Si hay un código en el DNA, **debe haber una forma de leerlo.**
- Gracias a la complementariedad de bases la información contenida en la molécula de DNA trae un respaldo integrado, puesto que **una de las cadenas es suficiente para revelar la secuencia de la otra.**

Pregunta

Si definimos secuenciar como la capacidad de revelar el orden de los nucleótidos en una molécula de ácido nucléico...

- ¿Cuándo surgió la primer maquinaria de secuenciación?

Pregunta

Si definimos secuenciar como la capacidad de revelar el orden de los nucleótidos en una molécula de ácido nucléico...

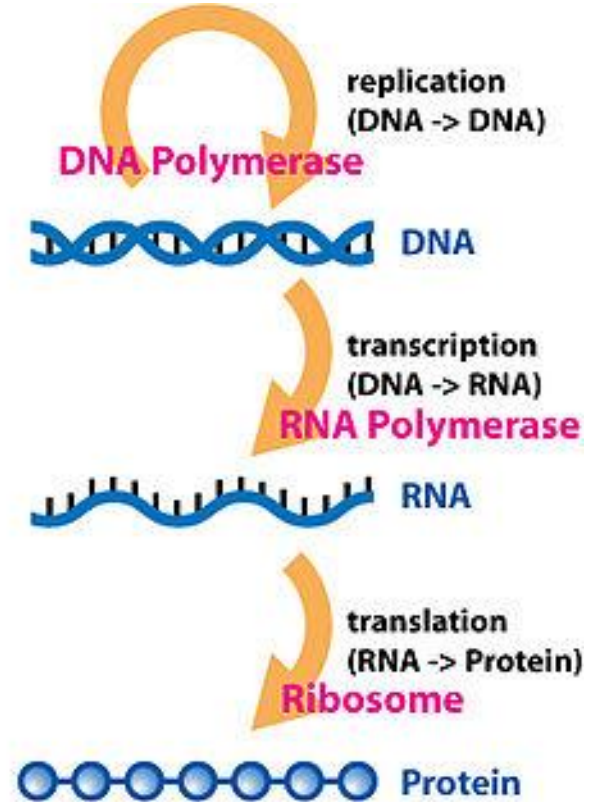
- ¿Cuándo surgió la primer maquinaria de secuenciación?

Respuesta:

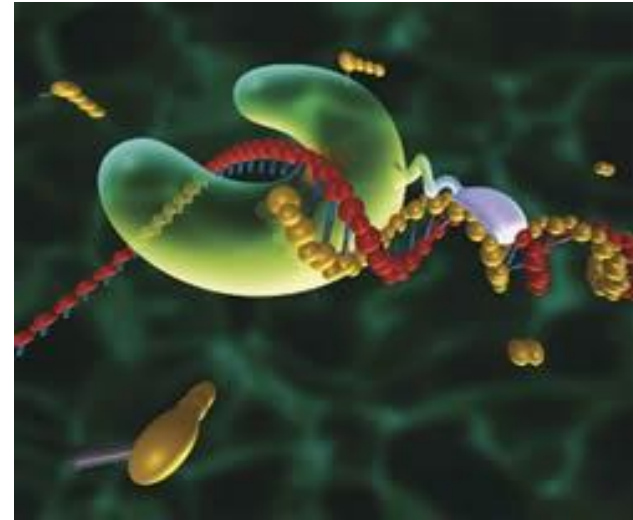
Hace aprox. 3,500 millones de años, con las primeras células capaces de transmitir su información genética.

Generación Zero de Secuenciadores

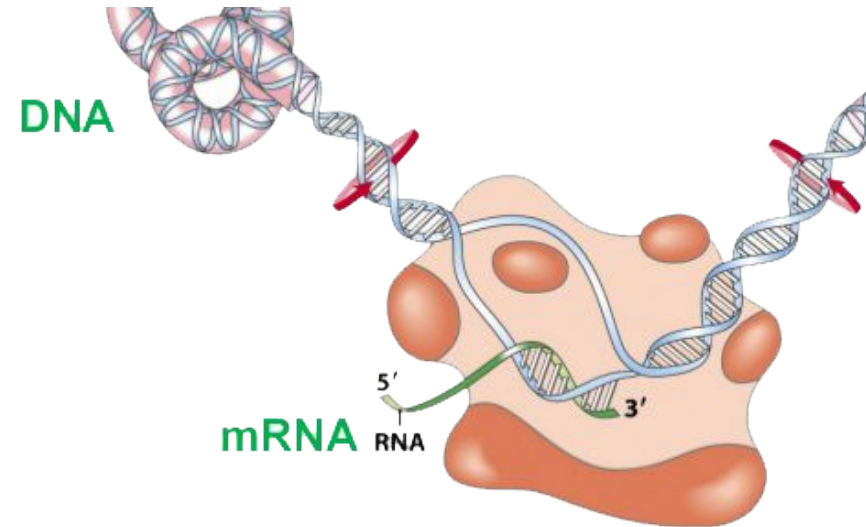
- Los primeros secuenciadores de ácidos nucleicos (DNA o RNA) fueron diseñados por la evolución.
- Éstos son las **proteínas** que median los procesos fundamentales del dogma de la biología molecular.



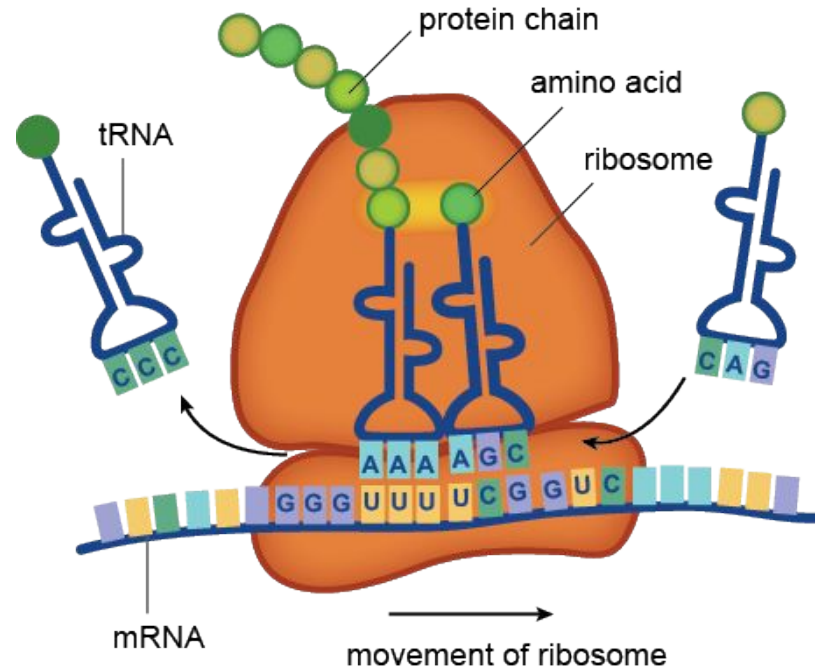
- La DNA polimerasa **lee y copia** la información contenida en el genoma, un nucleótido a la vez (**mismo formato, pero secuencia complementaria**).
- Incluso es capaz de corregir errores *in situ* a través del **proofreading**.



La RNA polimerasa **lee** el DNA y **convierte** la **información en otro formato**: RNA (muy similar al DNA).



El ribosoma **lee** el mRNA y lo **traduce** a un formato muy **distinto**: polipéptidos.



Maquinaria	Formato de entrada (Input)	Formato de salida (Output)
DNA pol	DNA	DNA
RNA pol	DNA	RNA
Ribosoma	mRNA	Polipéptidos

Pregunta

- ¿Qué otras máquinas biológicas son capaces de “leer” ácidos nucleicos?

Pregunta

- ¿Qué otras máquinas biológicas son capaces de “leer” ácidos nucleicos?

Respuesta:

Retrotranscriptasas, ligasas, enzimas de restricción, ribozimas, etc.

El concepto es el mismo: existen **mecanismos de reconocimiento molecular** que permiten identificar la composición del DNA y RNA (por tanto, su secuencia).

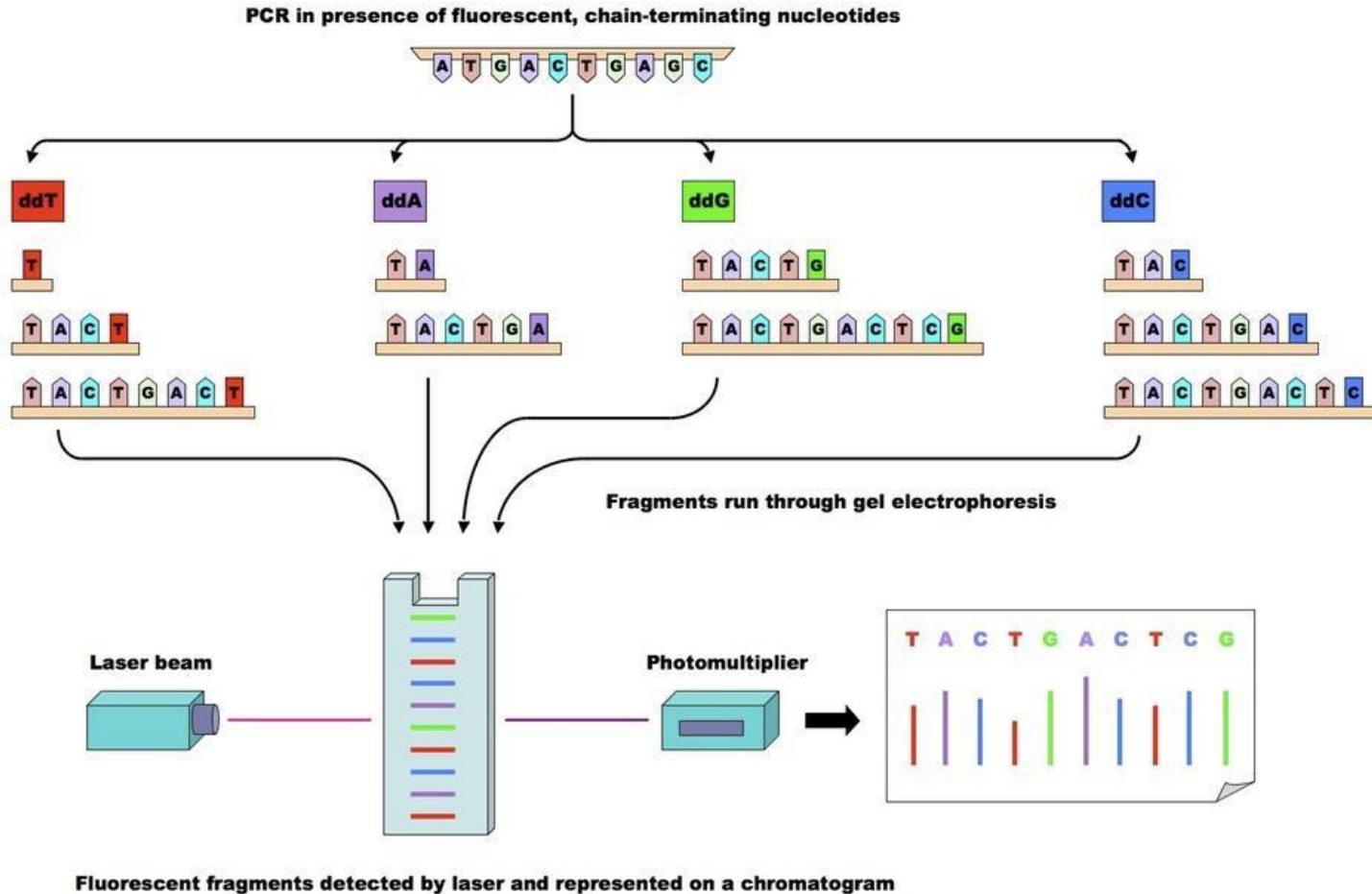
A partir de este idea común se desarrollaron las principales técnicas de secuenciación diseñadas por el hombre.

Secuenciadores de Primera Generación

En la década de los 70's se desarrollaron dos técnicas pioneras de secuenciación de DNA.

- **Maxam & Gilbert** – Secuenciación por degradación.
- **Sanger** – Secuenciación por síntesis parcial.

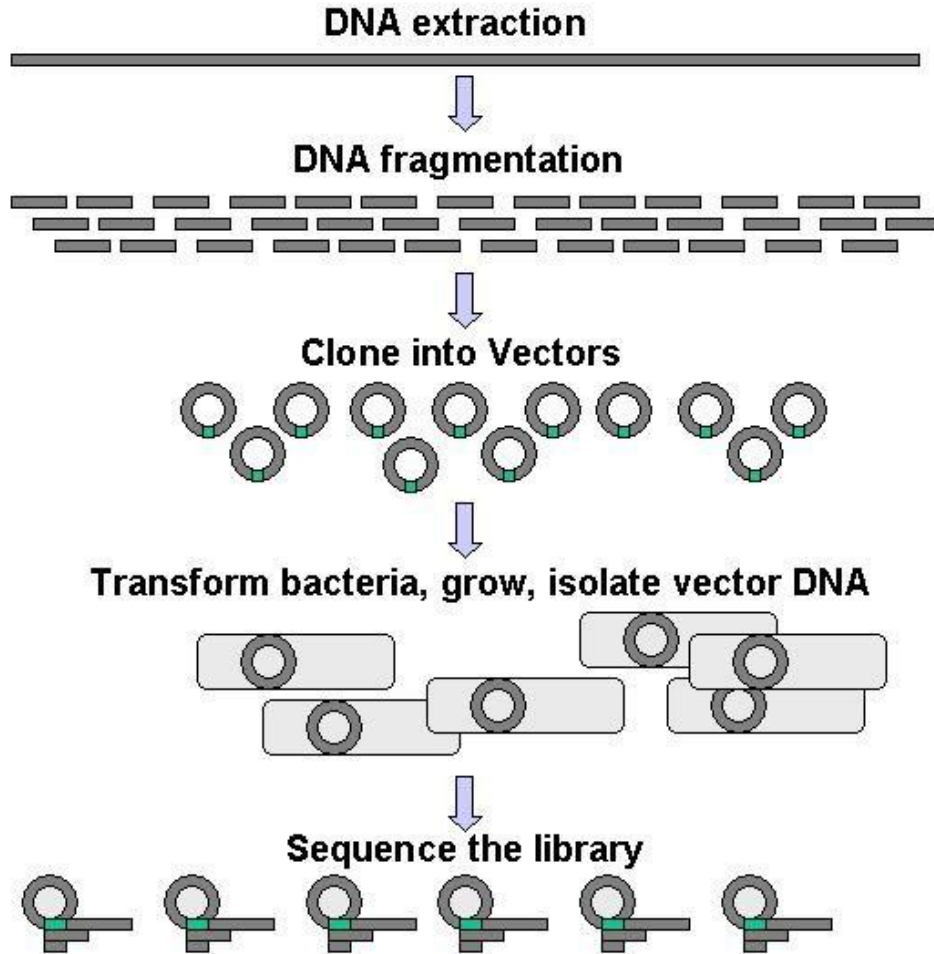
Eventualmente Sanger se convirtió en el estándar, **favorecido por la automatización del proceso.**



Durante décadas el método Sanger fue rey, pero el **Proyecto del Genoma Humano** evidenció su limitación en cuanto a rendimiento.

Con suerte, un secuenciador Sanger automatizado puede producir **384 secuencias con lecturas de hasta 1,000 pares de bases (1 kb)**.

Se requerirían casi 3 millones de reacciones Sanger para secuenciar el genoma humano completo (**3 Gb aprox.**)



Uno de las principales limitantes de los secuenciadores Sanger es la necesidad de **clonar los fragmentos y almacenarlos en vectores microbianos**.

Aunque las **lecturas son largas**, en proyectos grandes no se compensa el rendimiento del instrumento (costo de bases secuenciadas por corrida del aparato).

Estos fueron los principales problemas que pretendieron resolver las tecnologías **NGS (Next Generation Sequencing)**.

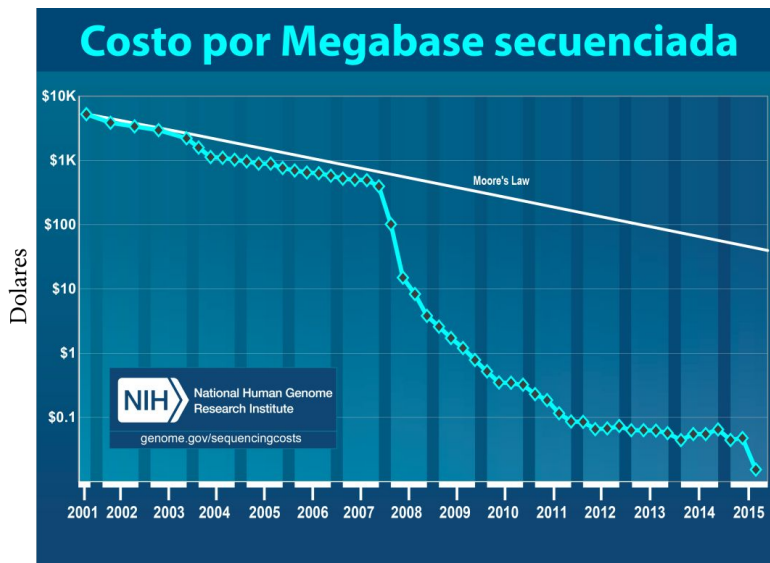
- **Costos**
- **Practicidad**
- **Velocidad**



Tecnologías NGS

(o secuenciadores de Segunda Generación)

Tecnologías NGS



Las tecnologías NGS **aceleraron el ritmo** de los estudios y aplicaciones genómicas.

La reducción de costos permitió la **democratización** de esta poderosa herramienta.

Tecnologías de Secuenciación

Primera Generación

Sanger

Segunda Generación

454

SOLiD

Ion Torrent

Illumina

Complete Genomics

Tercera Generación

Pacific Biosciences

Nanopore

Tecnologías NGS

Actualmente el NIH reporta el costo de secuenciar un genoma humano en \$1,300 aprox.

¡Y existen estímulos y propuestas tecnológicas para alcanzar el precio de **\$100!**

(Estos costos no toman en cuenta el precio del procesamiento bioinformático).

Tecnologías NGS

Muchos avances tecnológicos dieron paso a las tecnologías NGS:

- Microfluidos
- Robótica
- Óptica
- Herramientas de biología molecular

**¡Un secuenciador masivo es una maravilla tecnológica,
digna de un Nobel!**

Tecnologías NGS

Las principales características de los secuenciadores NGS son:

- **Miniaturización** - reduce costos al utilizar menos reactivos por secuencia.
- **Paralelización** - permite digitalizar millones de secuencias al mismo tiempo.

Tecnologías NGS

Unas cuantas empresas **iniciaron la carrera tecnológica NGS**, ofreciendo plataformas diferentes, químicas diferentes, archivos de salida diferentes, etc...

The logo for Illumina, featuring the word "illumina" in a lowercase, sans-serif font, with a cluster of yellow dots above the letters "i" and "n".The Roche logo, consisting of the word "Roche" in a blue, sans-serif font, enclosed within a blue hexagonal border.The Life Technologies logo, featuring the word "life" in a white, cursive script font, with the word "technologies" in a smaller, white, sans-serif font below it, all set against a dark blue rectangular background.The Pacific Biosciences logo, featuring a red circle with the letters "pb" in white, followed by the words "PACIFIC BIOSCIENCES" in a white, sans-serif font, all on a dark brown rectangular background.The Helicos Biosciences logo, featuring a stylized orange and red "H" icon, followed by the word "Helicos" in a large, black, sans-serif font, and "BioSciences Corporation" in a smaller, black, sans-serif font below it.The Ion Torrent logo, featuring the words "ion torrent" in a bold, black, sans-serif font, with a row of colorful geometric shapes (a purple teardrop, a blue star, an orange triangle, a white circle, a blue X, a green square, a red plus, and a black wavy line) below it.The Oxford Nanopore Technologies logo, featuring a blue circular icon with a stylized "N" inside, followed by the word "Oxford" in a small, blue, sans-serif font, and the words "NANOPORE Technologies" in a large, blue, sans-serif font.

Tecnologías NGS

Algunas perecieron, otras se fusionaron, y muchas otras continúan apareciendo.



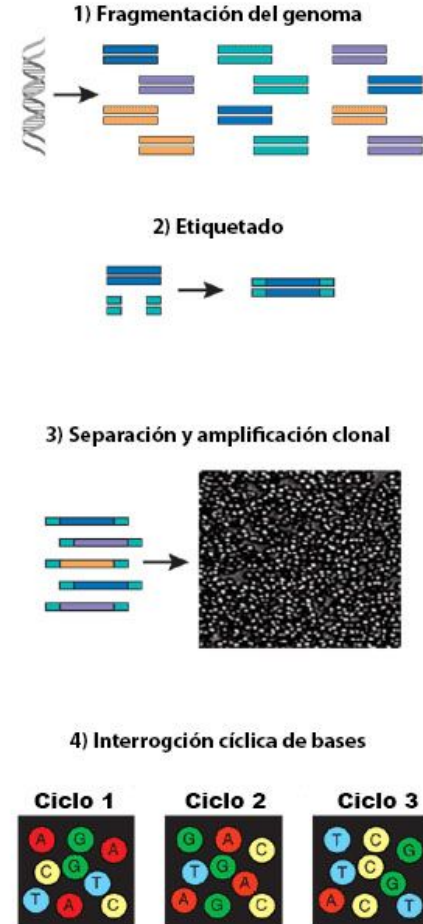
Tecnologías NGS

Haremos énfasis en la tecnología **illumina**, por ser la plataforma de mayor uso actualmente.

The logo for Illumina, featuring a stylized lowercase 'i' in orange and the word 'llumina' in a grey sans-serif font, followed by a registered trademark symbol (®).

Proceso general NGS

1. Fragmentación de la muestra (genoma o amplicones)
2. Etiquetado y separación de fragmentos.
3. Amplificación clonal.
4. Interrogación cíclica de bases.



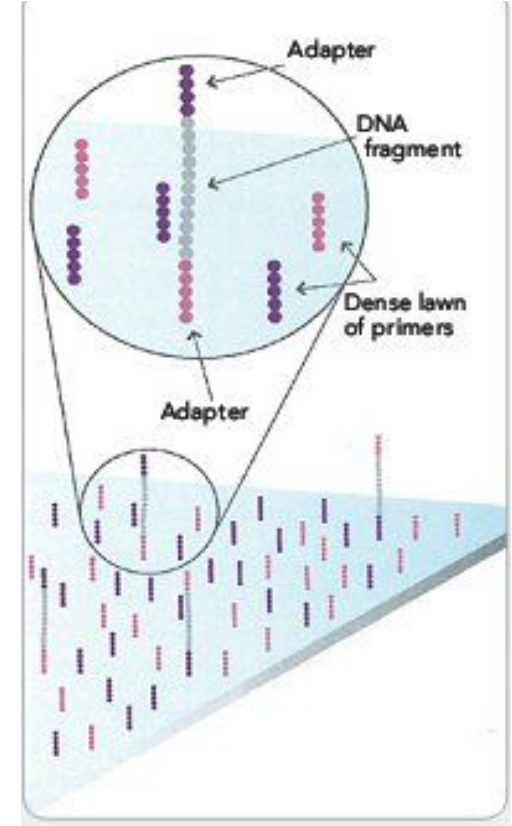
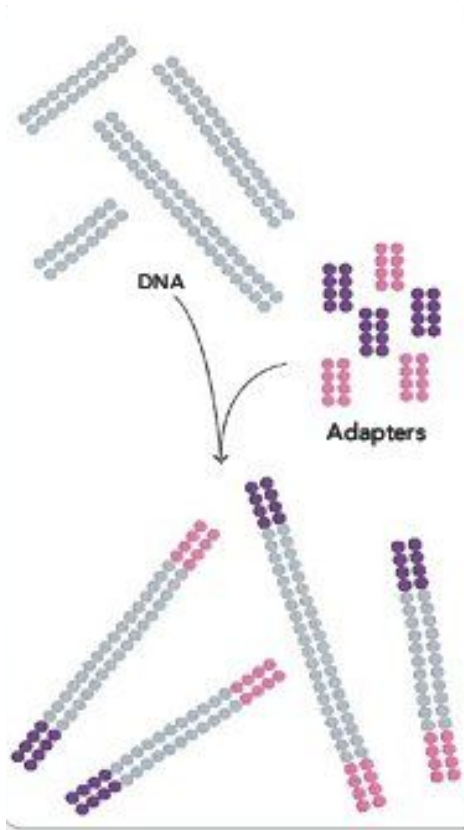
Fragmentación

La longitud de los fragmentos DNA secuenciados es **importante**. Mientras más grandes las piezas, más fácil es armar el rompecabezas.

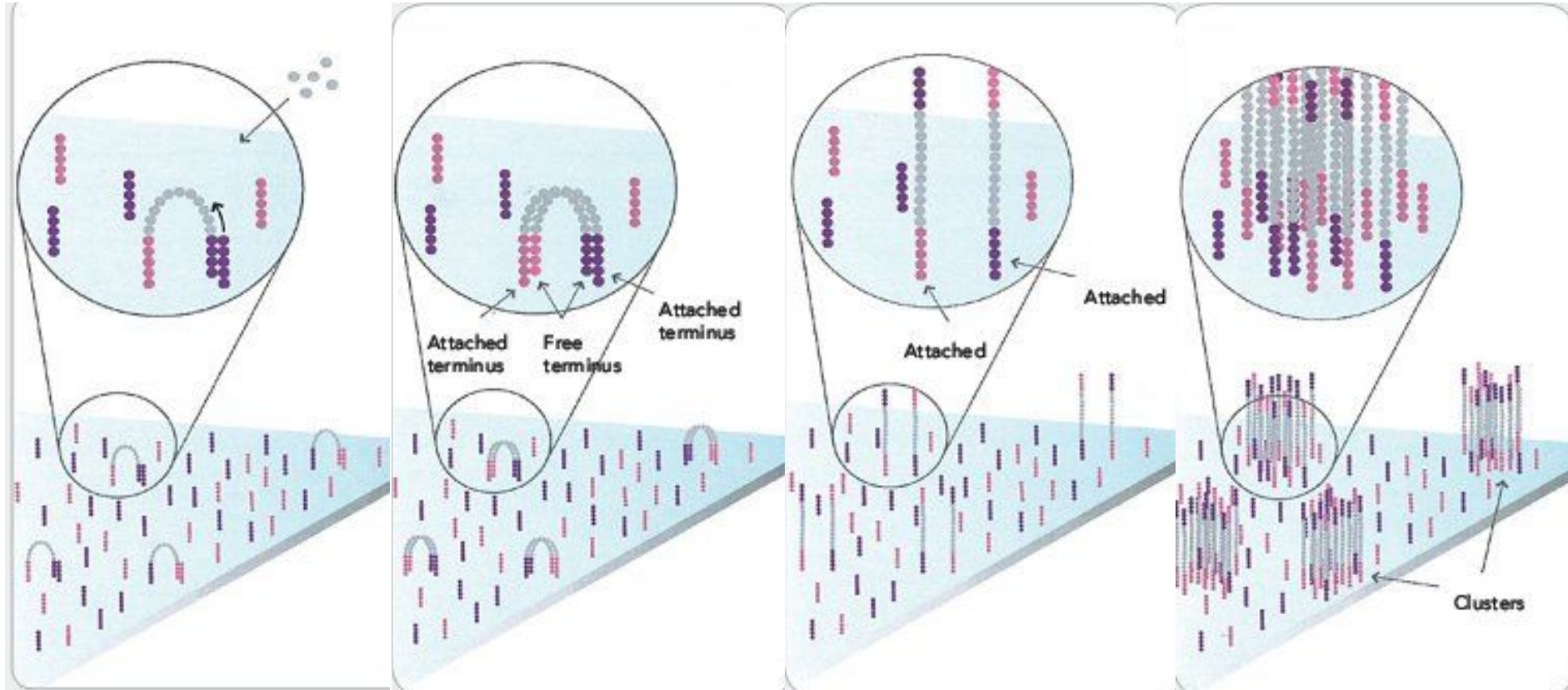
Los métodos de fragmentación pueden ser:

- **Enzimáticos** (endonucleasas, o transposasas)
- **Mecánicos** (presión hidráulica, o estrés generado por ondas acústicas)

Etiquetado y separación



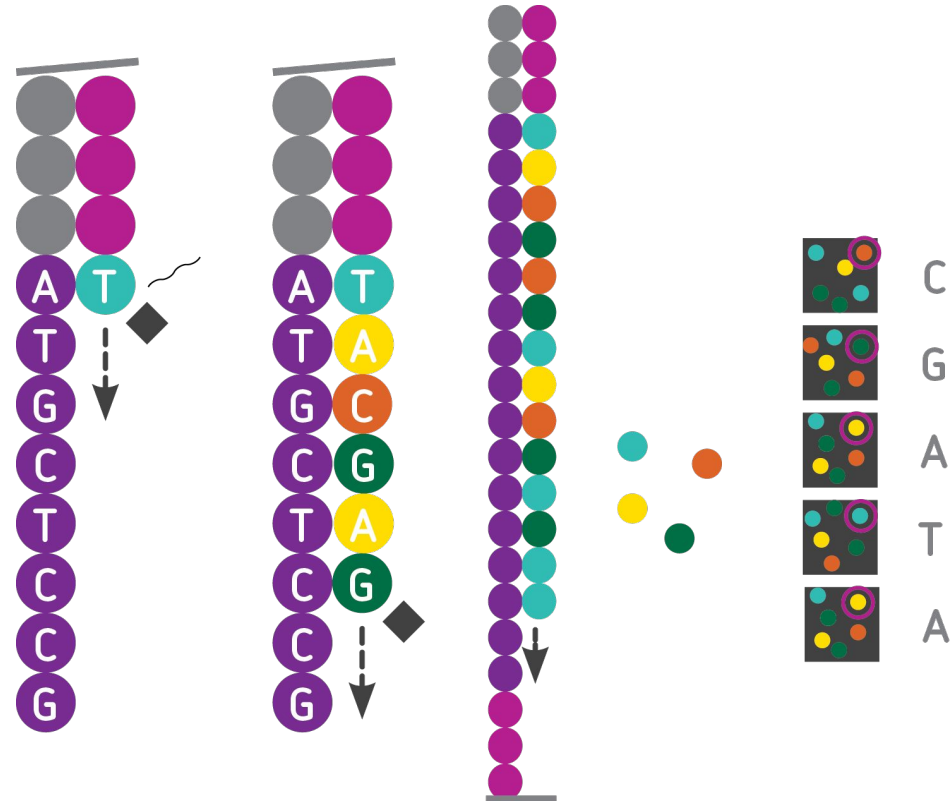
Amplificación clonal



Interrogación de bases

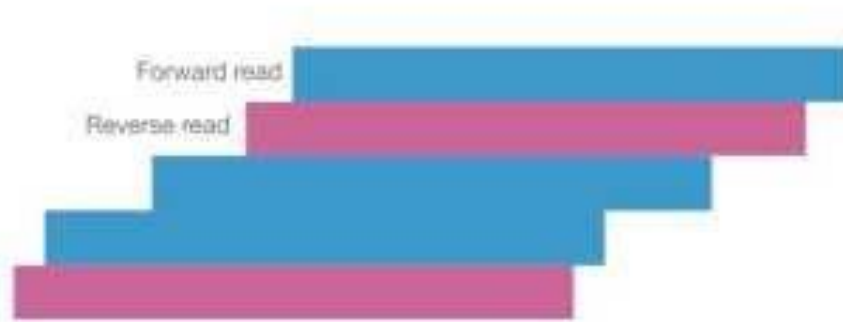
En cada ciclo de secuenciación:

1. Se agregan una solución de DNA polimerasa + los 4 nucleótidos (nt) bloqueados, acoplados a un fluoróforo.
2. Solo un nt se incorpora.
3. Se lavan los no incorporados.
4. Se excitan los fluoróforos, se toma una foto para digitalizar el nt incorporado.
5. Se desbloquea el nt, y se elimina su fluoróforo.



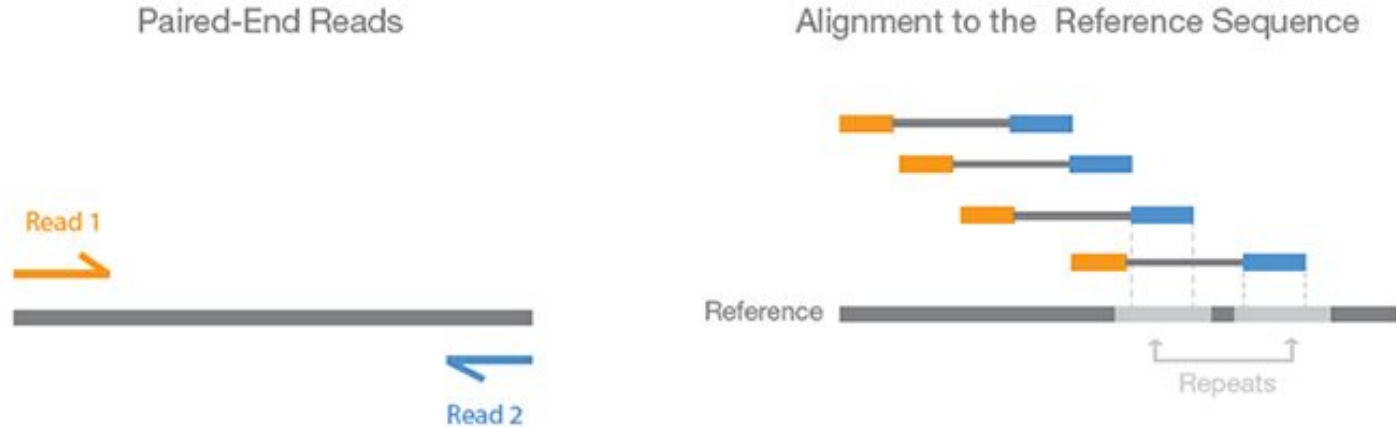


Create contiguous sequences



Paired-end

Figure 4. Paired-End Sequencing and Alignment



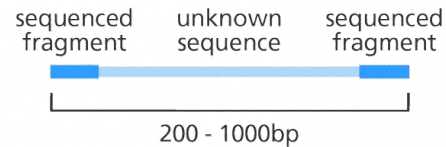
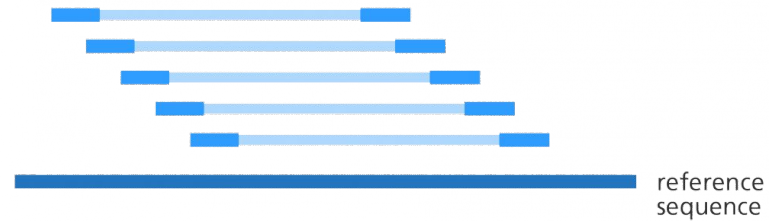
Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Tecnologías NGS

Single-end reads



Paired-end reads

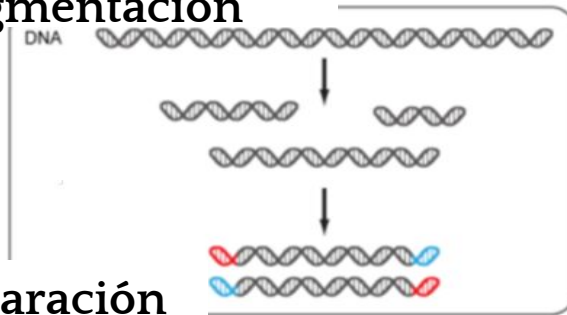


Otras tecnologías NGS actuales

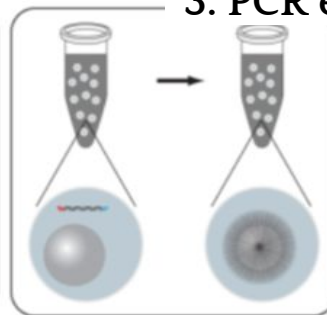
Ion Torrent

En Ion Torrent lo que se detecta no es fluorescencia, sino la liberación del protón cuando un nucleótido es integrado a la cadena naciente DNA.

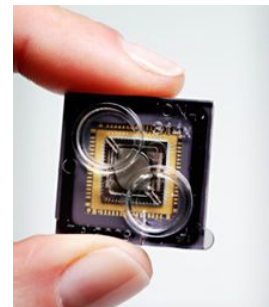
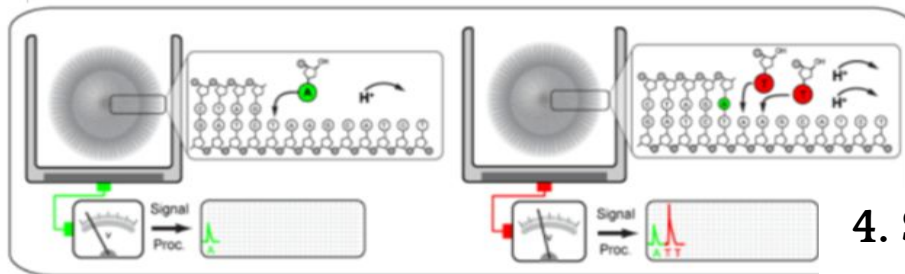
1. Fragmentación



3. PCR en emulsión



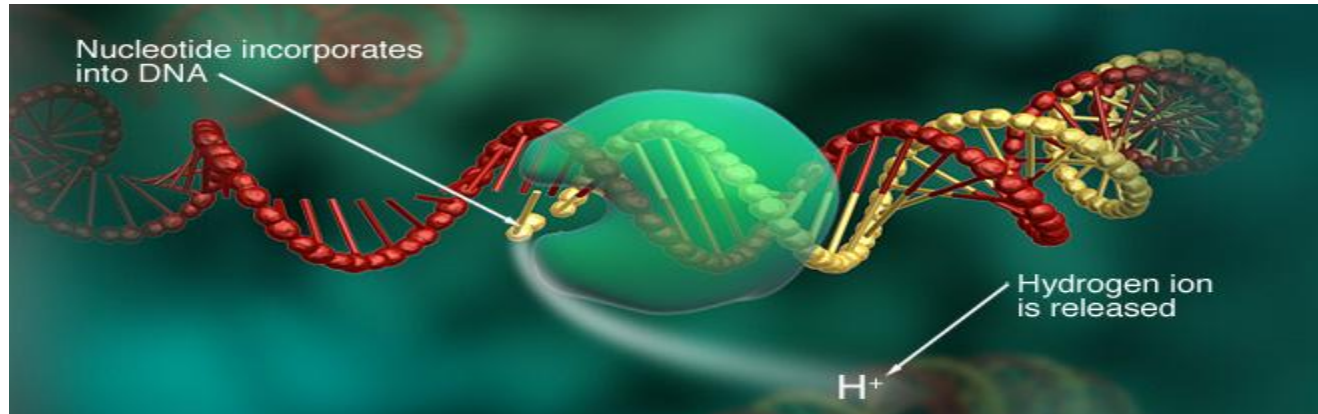
2. Etiquetado y separación



4. Secuenciación por semiconductores

Ion Torrent

- Al no utilizar nucleótidos secuenciados ni sistemas ópticos de detección, su funcionamiento es más barato que la competencia.
- Por otro lado, el hecho de no utilizar dNTP modificados hace al sistema proclive a errores en trectos homopoliméricos.

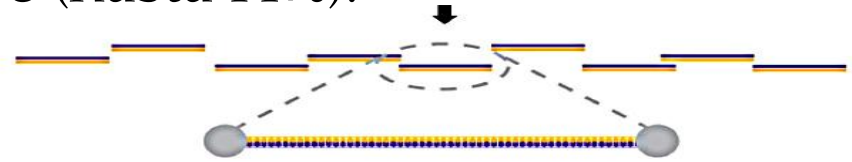


Tercera generación de Secuenciadores (NGS)

PacBio

Los adaptadores generan templados circulares, con la finalidad de que una misma molécula sea leída varias veces para amortiguar las tasas de error base (hasta 14%).

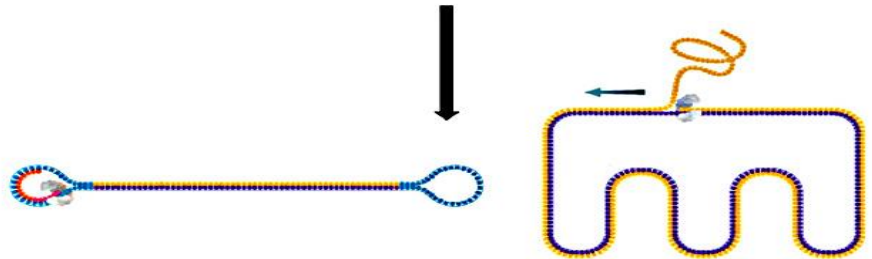
1. Fragmentación



2. Etiquetado y separación

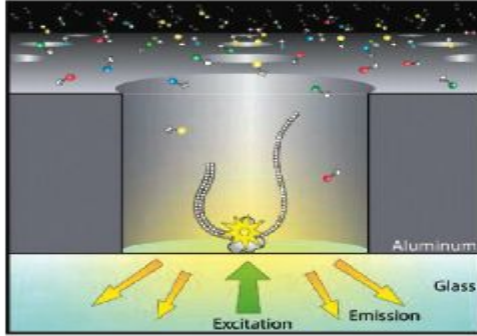


3. Secuenciación por síntesis en tiempo real

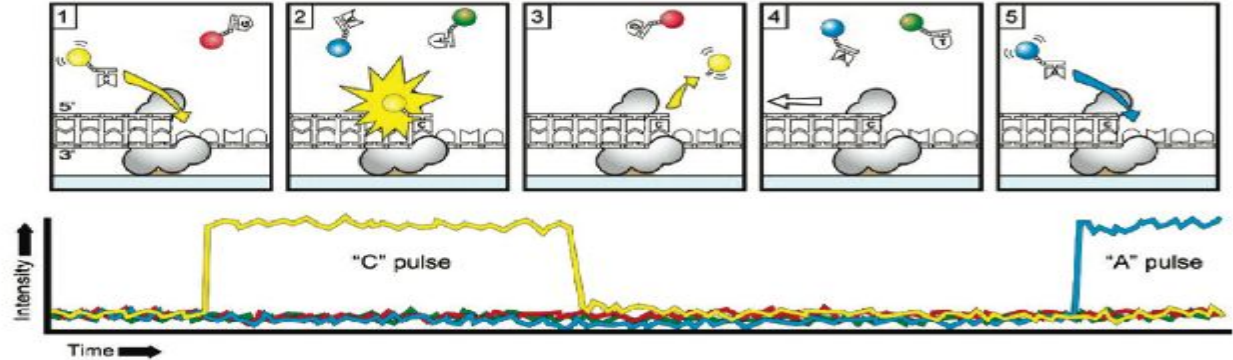


PacBio

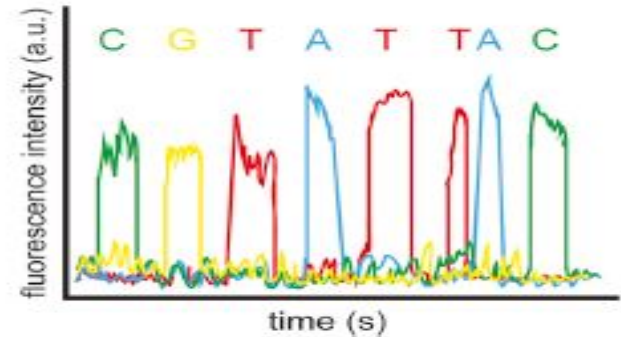
A



B



La incorporación de bases es monitoreada en **video en tiempo real**.



Nanopore

La plataforma insignia es el **MinION**.

Recientemente liberado al mercado, hacen falta más reseñas por parte de la comunidad.

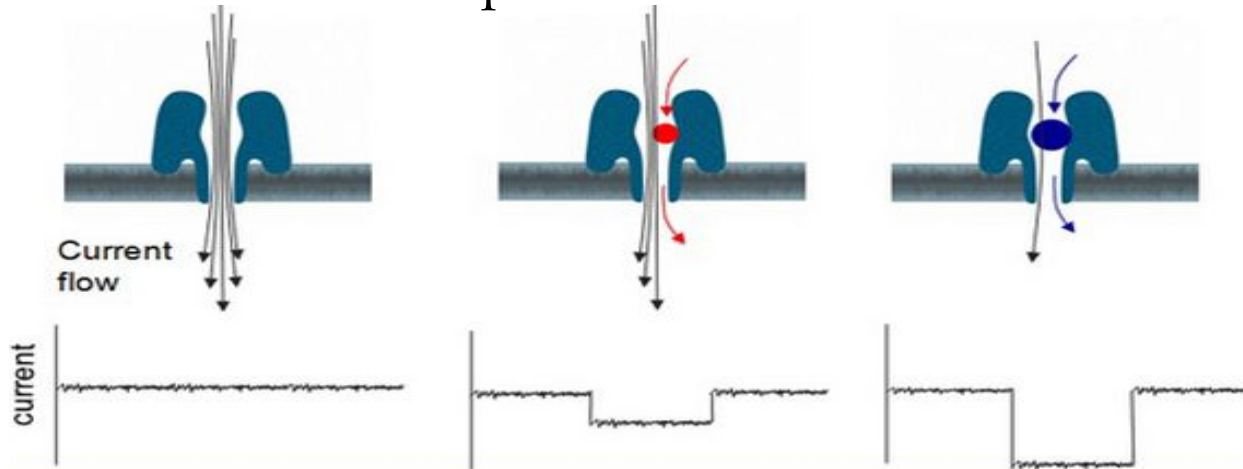
Trae el concepto **Lab-on-chip** a NGS.



Nanopore

Monitoreo en tiempo real.

Sensores detectan cambios en la corriente eléctrica provocados por la obstrucción que provoca el fragmento DNA al transitar por el canal interno del nanoporo.

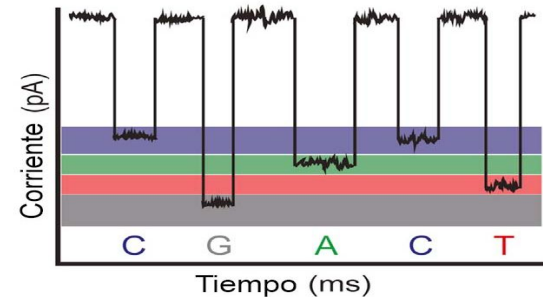
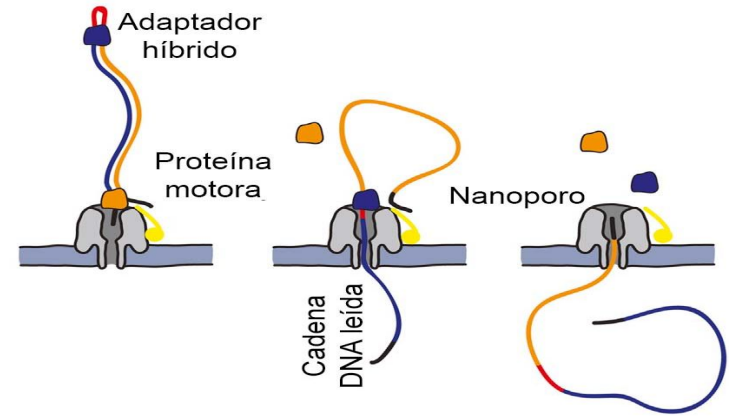


Cada nucleótido produce una señal diferente.

Nanopore

Los **adaptadores** son híbridos DNA-proteína que **guían** el **fragmento** hacia el **poro**.

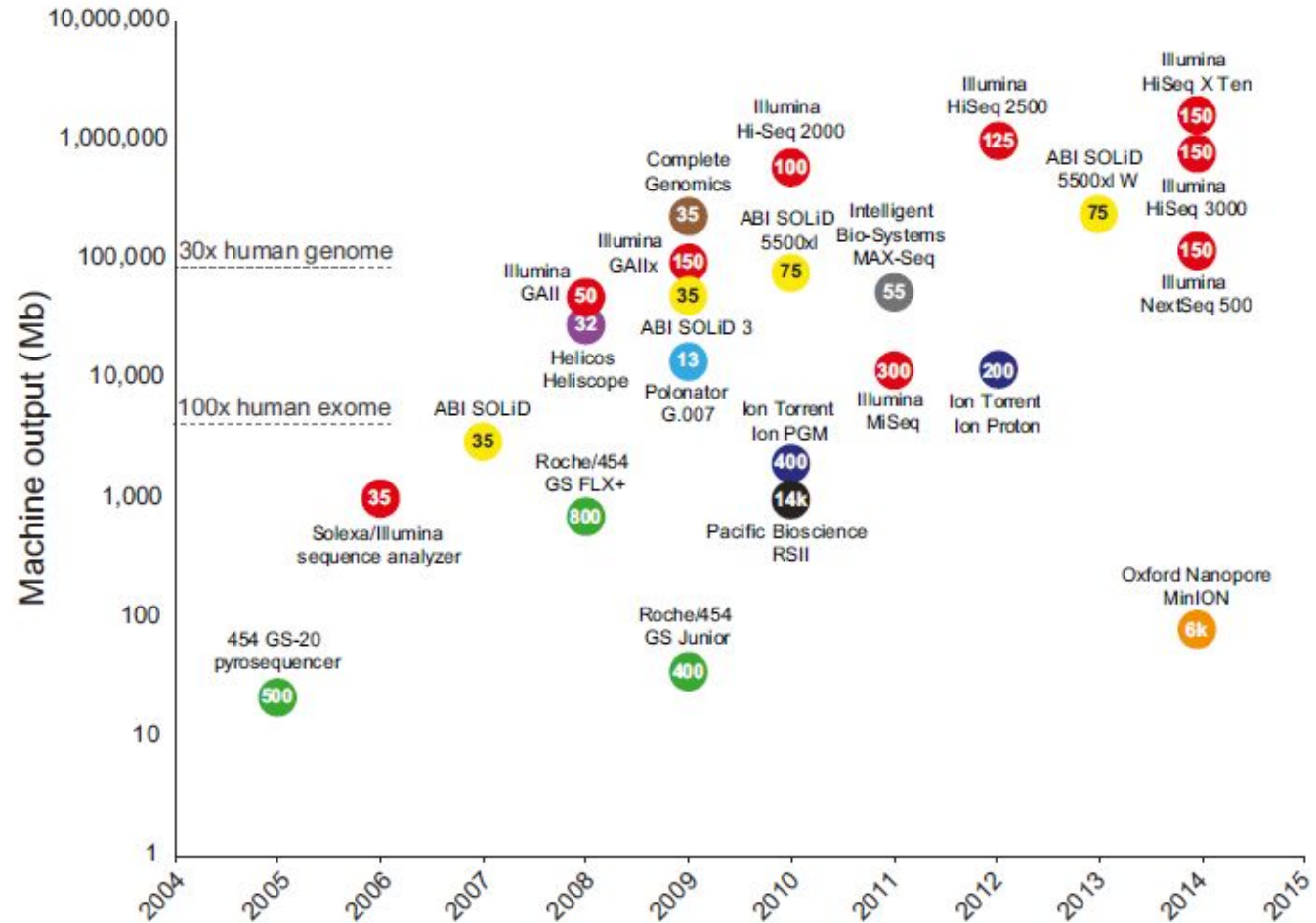
Los adaptadores son en forma de **ovillo**, lo cual conecta ambas cadenas del fragmento original. De esta manera, **la secuenciación es bidireccional**.



Evolución de las tecnologías NGS

Mejoras NGS en la última década

* El número dentro del círculo señala la longitud máxima de lecturas.



Instrumentos NGS en desarrollo. (El futuro es ahora)



Q-POC

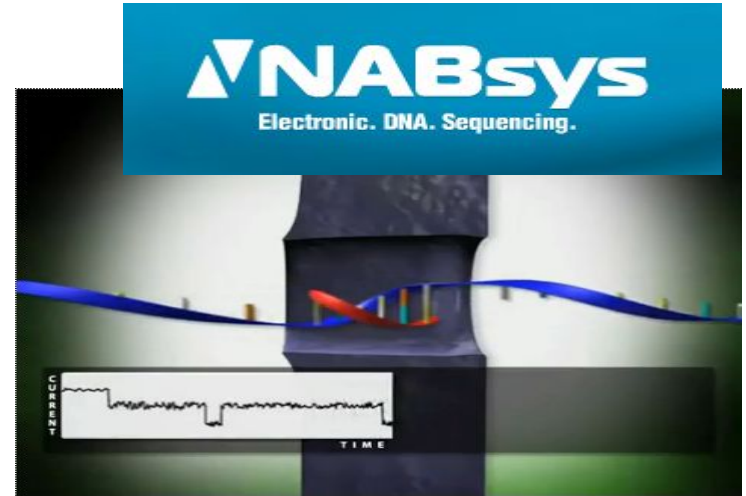
Lector de secuencias cortas (50 pb)

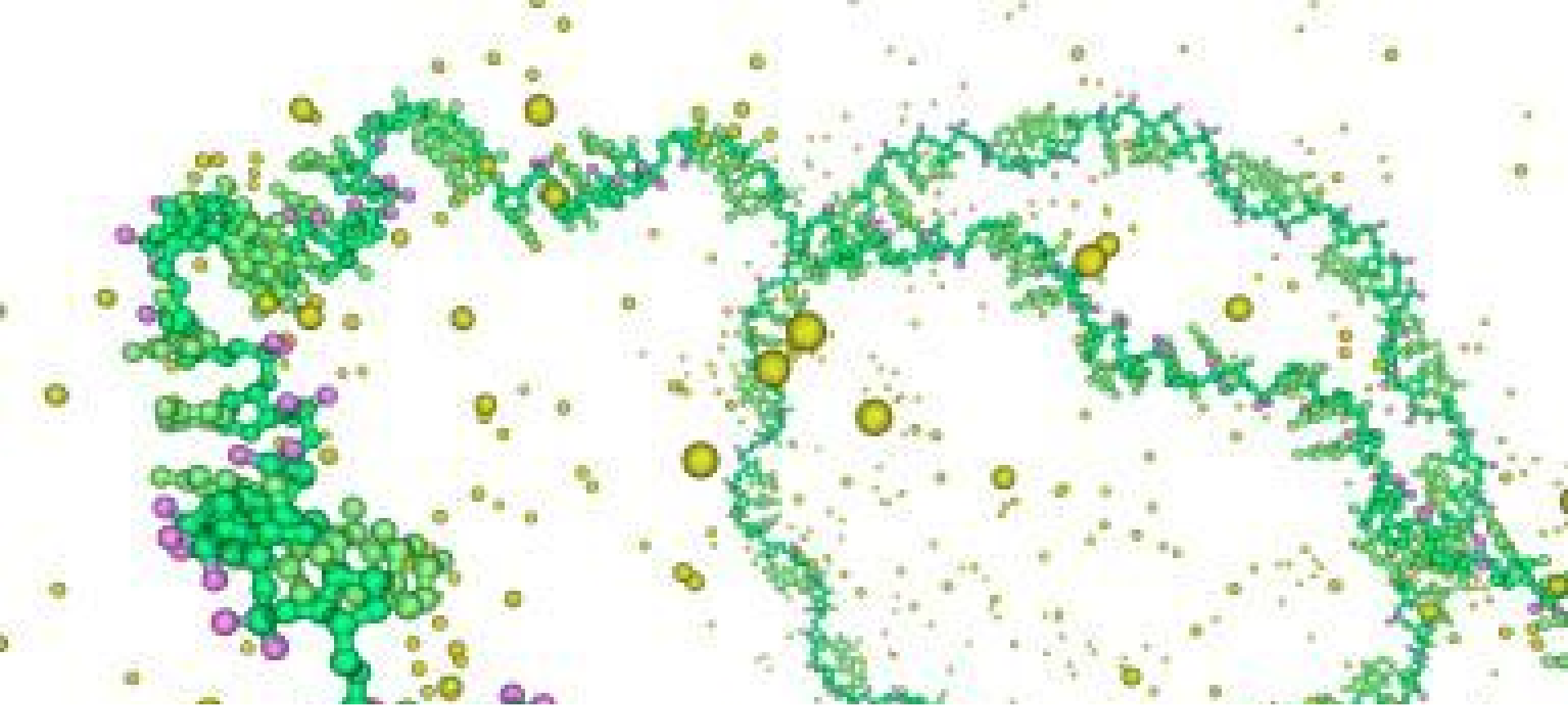
Detecta la adición de nucleótidos por medio de **nanocables**.

NABsys

¡Lector de secuencias de **hasta 100 kb**!

Basado en nanoporos e hibridación de sondas DNA.





RNA-Seq

Para secuenciar RNA por NGS, **es necesario convertirlo previamente en cDNA**. Después de eso la secuenciación se lleva a cabo de manera normal.

Ninguna de las plataformas comercialmente establecidas acepta **RNA directamente como muestra inicial**.

¿Por qué?

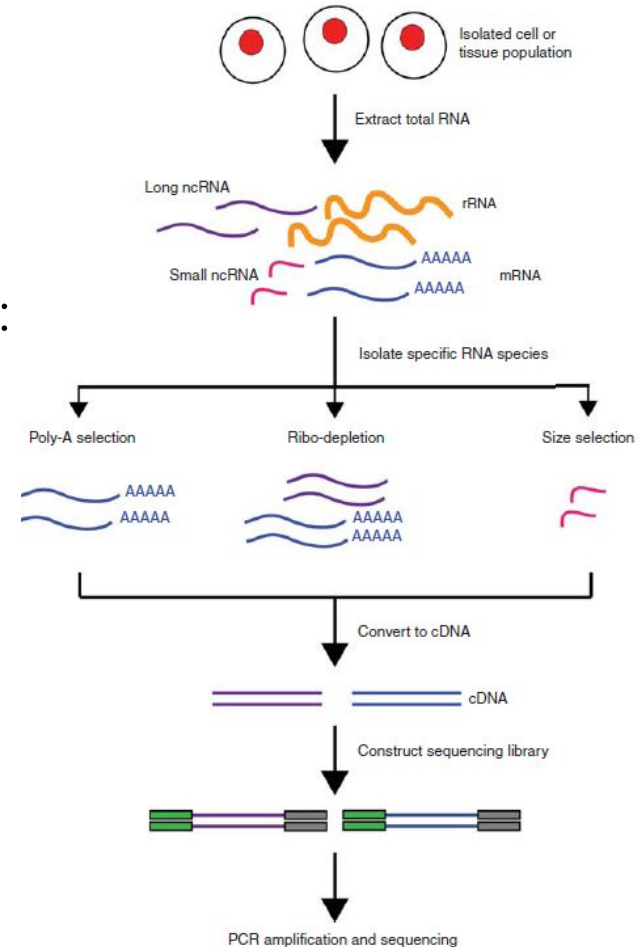
RNA-Seq

Construcción de bibliotecas RNA-Seq

Paso 1. Extraer RNA

Paso 2. Seleccionar la subpoblación deseada:

- RNA total
- mRNA
- lncRNA
- small RNA
- miRNA
- piRNA
- siRNA
- y la lista sigue...

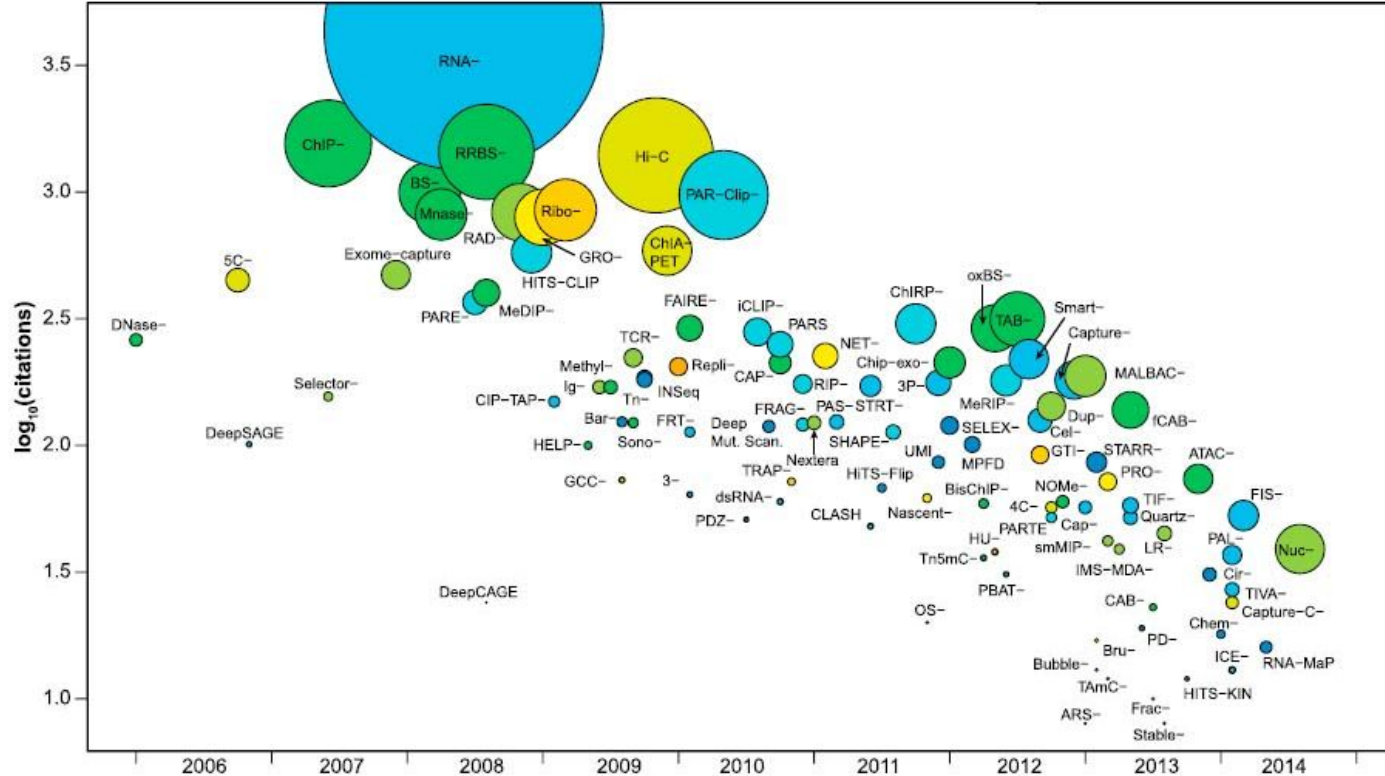


Cerca del 90 % del RNA total extraído de células es rRNA.

Esto hace necesario el uso de estrategias para **enriquecer** las subpoblaciones de interés (**mRNA**, etc).

Como regla tácita, si existe una **proteína que interactúa** con tu subpoblación RNA de interés, es posible enriquecer la muestra.

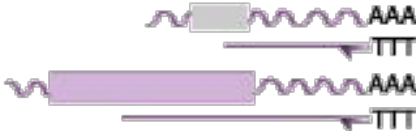
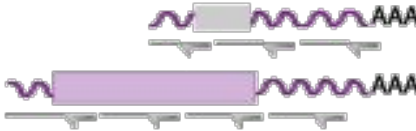
Alcance de las publicaciones de nuevos métodos de enriquecimiento NGS



RNA-Seq

Paso 3. Síntesis de cDNA por transcripción reversa.

La forma más sencilla de generar cDNA es a través de oligos **poli-T**, y oligos **Random**.

	Oligo dT	Random (pdN ₆)
		
Advantage	Specifically primes poly A ⁺ mRNA in total RNA preps	Generates a large representative cDNA library of sequences
Disadvantage	Library will be under-represented for gene coding sequences	cDNA products can be small, poly A ⁺ templates are primed

RNA-Seq

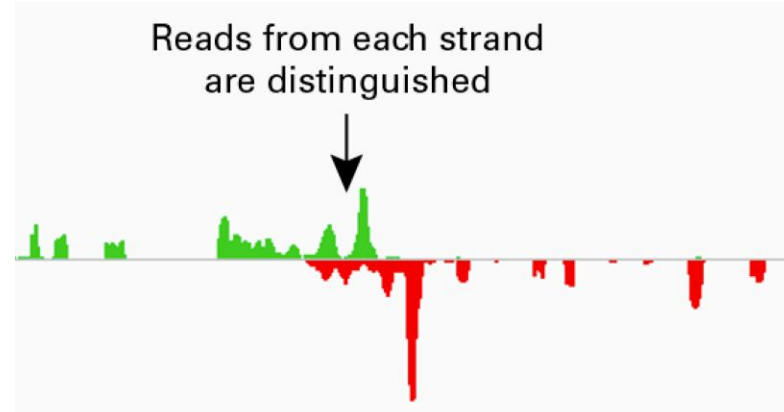
Para experimentos de expresión diferencial, es importante que la biblioteca RNA-Seq se haya hecho con un protocolo direccional.

Esto significa que debe ser posible reconocer cuál era el sentido original del transcrito secuenciado (5' - 3', o 3' - 5').

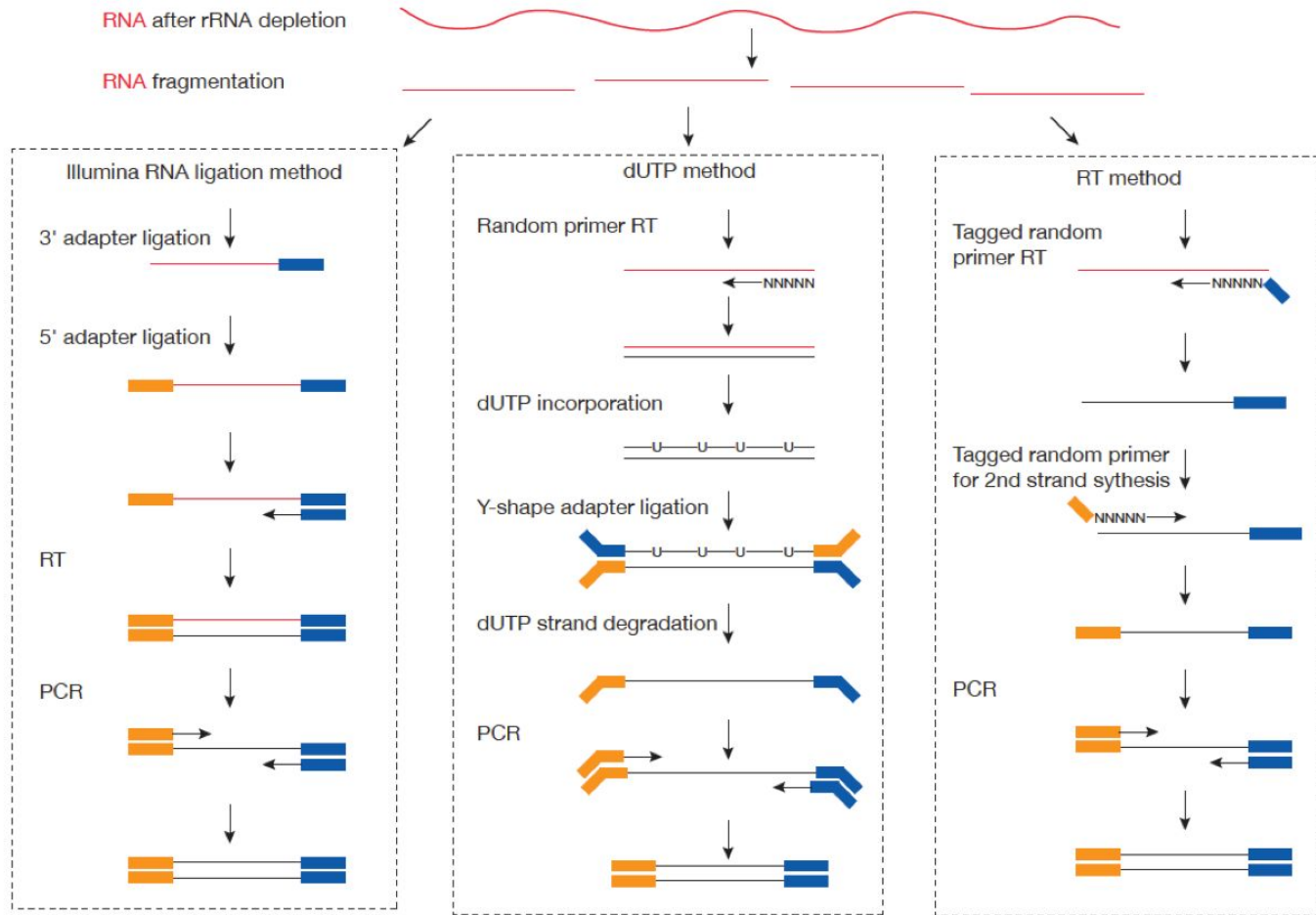
Sentido

ó

Antisentido



Métodos para dar direccionalidad a la biblioteca



Proceso general RNA-Seq

1. Enriquecimiento de muestra y síntesis de cDNA.
2. Etiquetado y separación de fragmentos.
3. Amplificación clonal.
4. Interrogación cíclica de bases.

Referencias

Watson, J. (1993). Genetical implications of the structure of deoxyribonucleic acid. *JAMA*, 269(15),p.1967.

Biotech.missouri.edu, (2015). *MU DNA Core Facility*. [online] Available at: <http://biotech.missouri.edu/dnacore/sangersequencing.html> [Accessed 12 Nov. 2015].

Genome.gov, (2015). *Genome Technology Program*. [online] Available at: <http://www.genome.gov/10000368> [Accessed 12 Nov. 2015].

Bennett, G. and Moran, N. (2013). Small, Smaller, Smallest: The Origins and Evolution of Ancient Dual Symbioses in a Phloem-Feeding Insect. *Genome Biology and Evolution*, 5(9), pp.1675-1688.

van Dijk, E., Auger, H., Jaszczyszyn, Y. and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9), pp.418-426.

Reuter, J., Spacek, D. and Snyder, M. (2015). High-Throughput Sequencing Technologies. *Molecular Cell*, 58(4), pp.586-597.

> **Turning data into forefront knowledge**



CONTACTO

Manizales No.906, Colonia Lindavista.
México, Distrito Federal.

Tel/fax: (52) (55) 5119-0240
(52) (55) 5119-5624

E-mail: contacto@wintergenomics.com

www.wintergenomics.com