Análisis transcriptómicos de la expresión génica

Máster Universitario en Bioinformática

Sesión 4

Dra. Paula Soler Vila
paula.solerv@professor.universidadviu.com

De:
Planeta Formación y Universidades

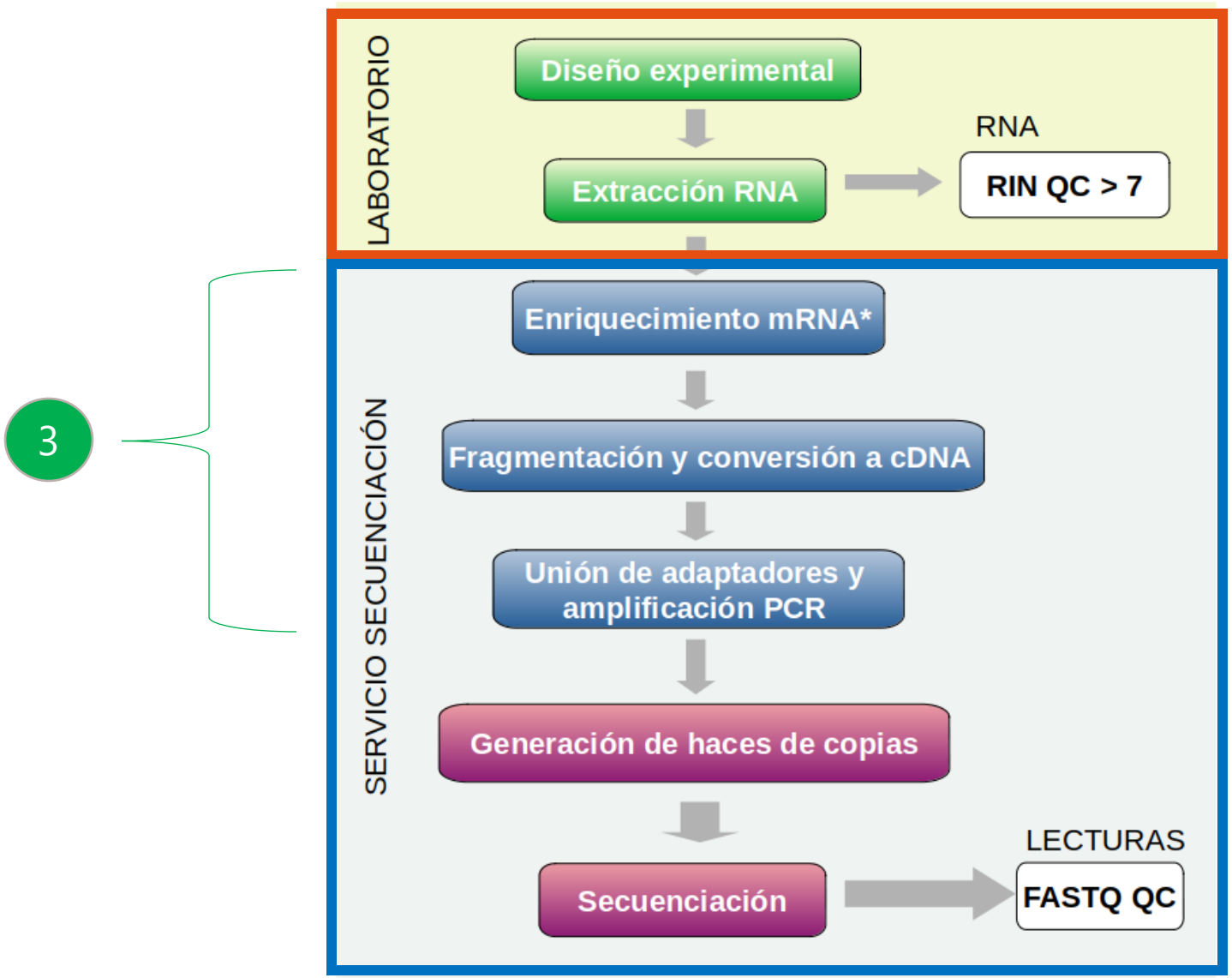
Bloque II: Estudios de expresión génica con datos de NGS

Objetivos de la sesión

1. Conocer cuáles son los puntos claves en el diseño y desarrollo de un experimento de RNA-seq.

- Diseño Experimental
- Proceso de extracción del ARN
- Diseño de librerías
- Secuenciación
  - **Ejercicio práctico:** Tarifas de diseño y secuenciación de una librería de RNA-seq
- Análisis de datos

Flujo de trabajo general de RNA-seq



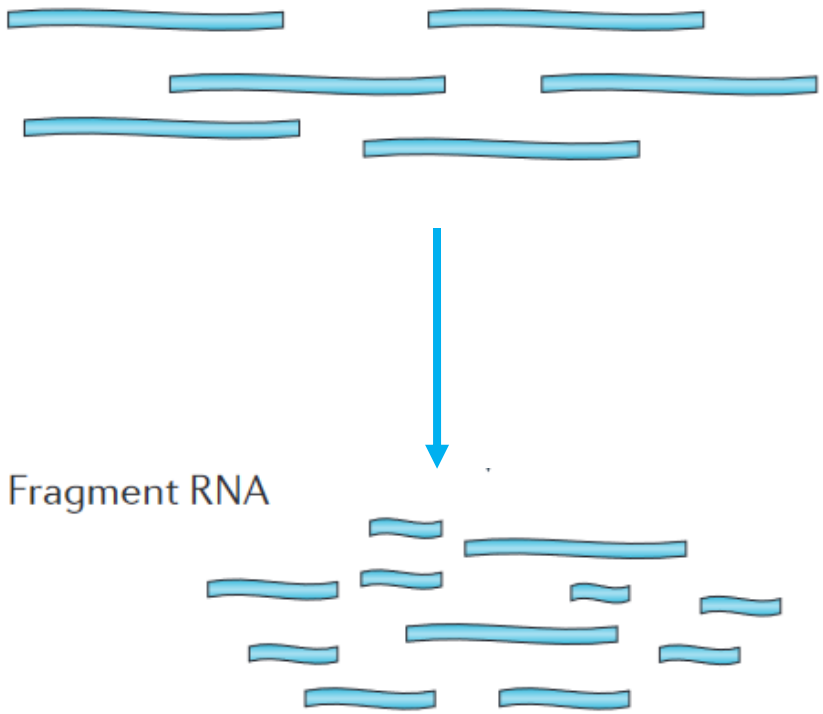
3 Preparación de la librería

La preparación de la librería implica generar una colección de fragmentos de ARN que sean compatibles para secuenciación.

El proceso implica el enriquecimiento del ARN diana , la fragmentación, transcripción inversa (ADNc), adición de adaptadores de secuenciación y amplificación

¿Cómo podemos fragmentar el ARN?

3 Fragmentación del RNA



- **Fragmentación del RNA**
  - Química
  - Enzimática (RNAsas III)
  - Mecánica
    - Ultrasonidos

- **Fragmentación del cDNA**
  - Enzimática (ADNasas)
  - Mecánica

La distribución de tamaños de los fragmentos puede ser amplia, con una cantidad considerable de fragmentos en el rango de **200-500** bases.

3 Transcripción reversa (ADNcomplementario)

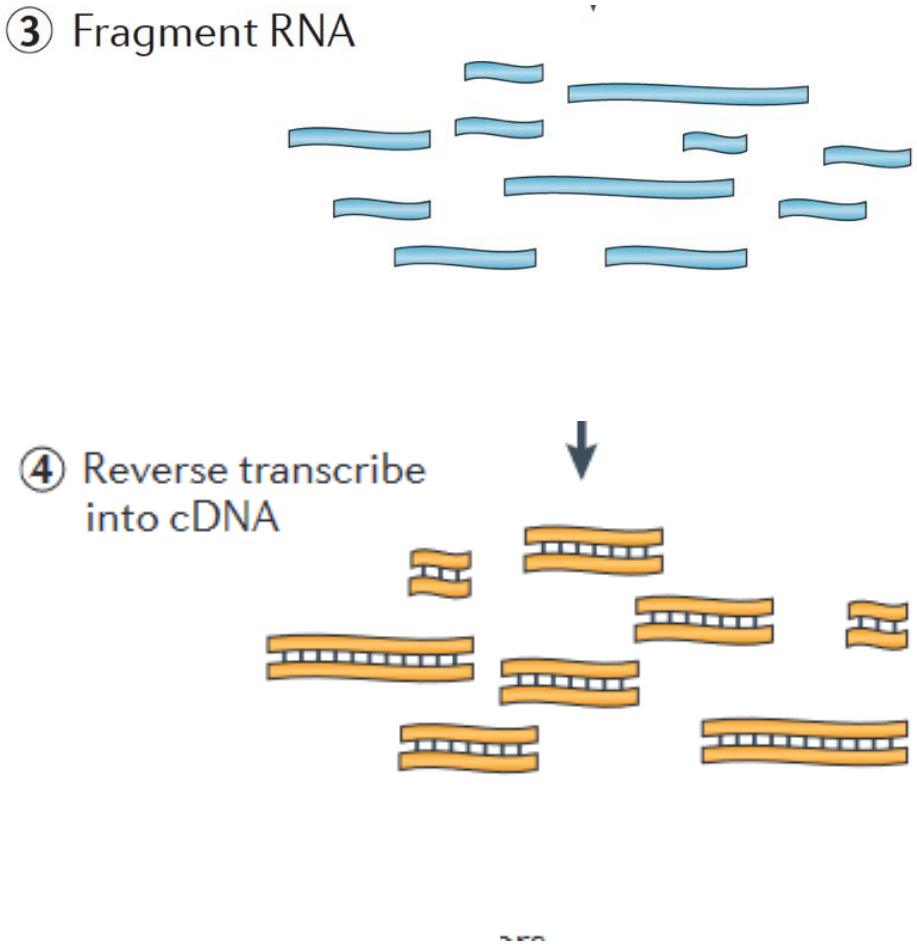
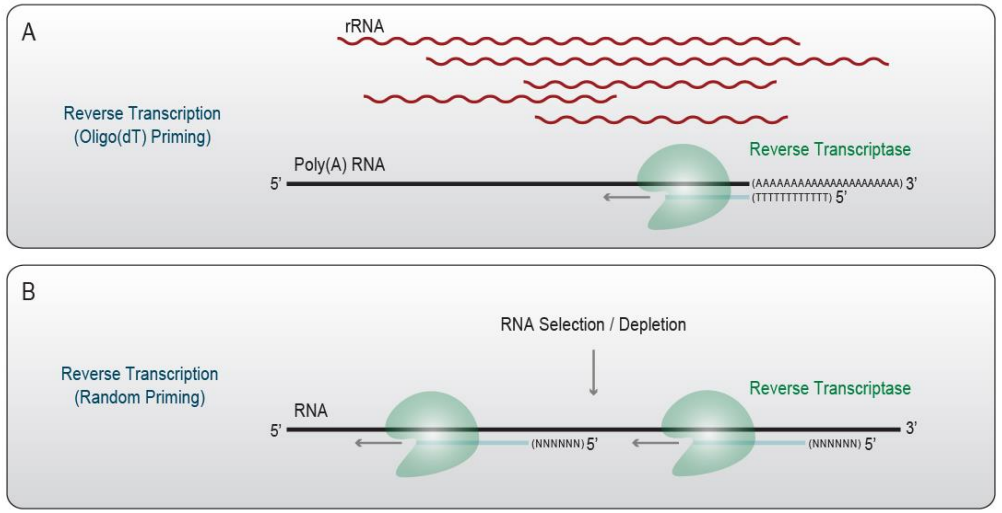


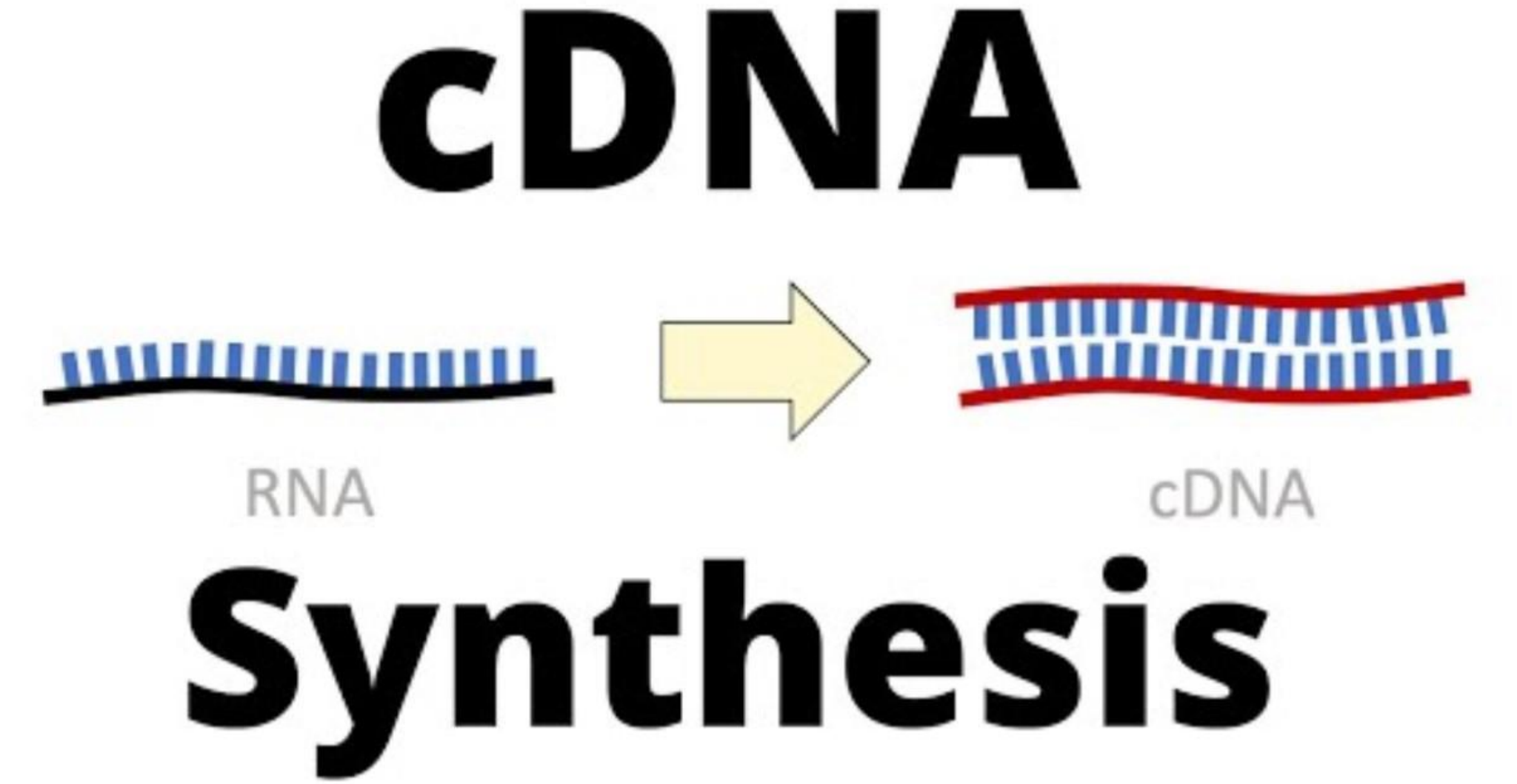
Image credit: Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

- Los fragmentos de ARN son usados como **molde** para la síntesis del ADNc
- Síntesis de la primera hebra mediada por una **retrotranscriptasa.**



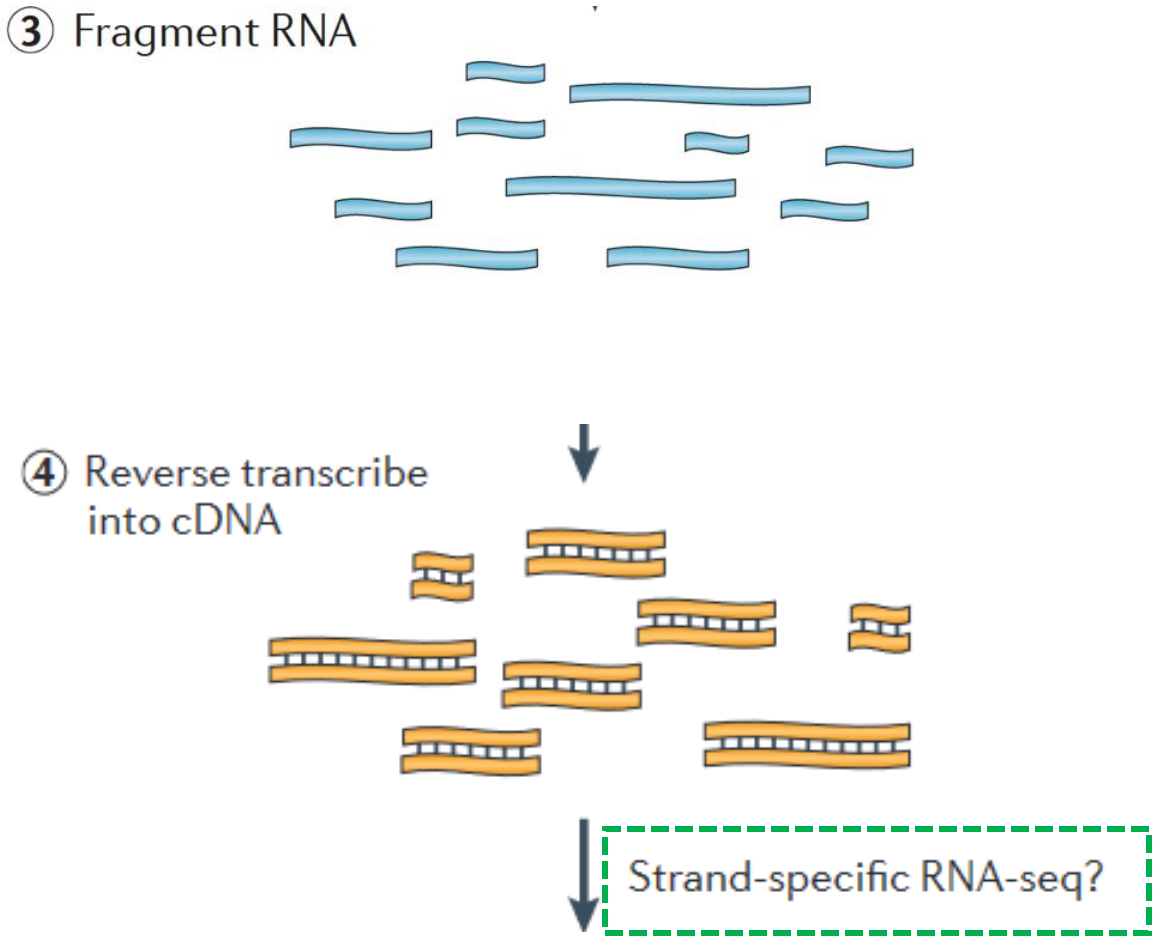
- Síntesis de la segunda hebra mediada por una **ADN polimerasa.**

3 Transcripción reversa (cDNA)



https://www.youtube.com/watch?v=VT_nNVO95eI

3 Transcripción reversa (cDNA)



Stranded

RNA-Seq direccional o específico de cadena, le permite determinar la orientación del transcrito.

- **Introducción de dUTP**

- Identificación de transcriptos *antisentido*
- Anotación y descubrimiento de transcritos nuevos en organismo no modelo.
- Análisis de expresión génica (*overlapping genes*)

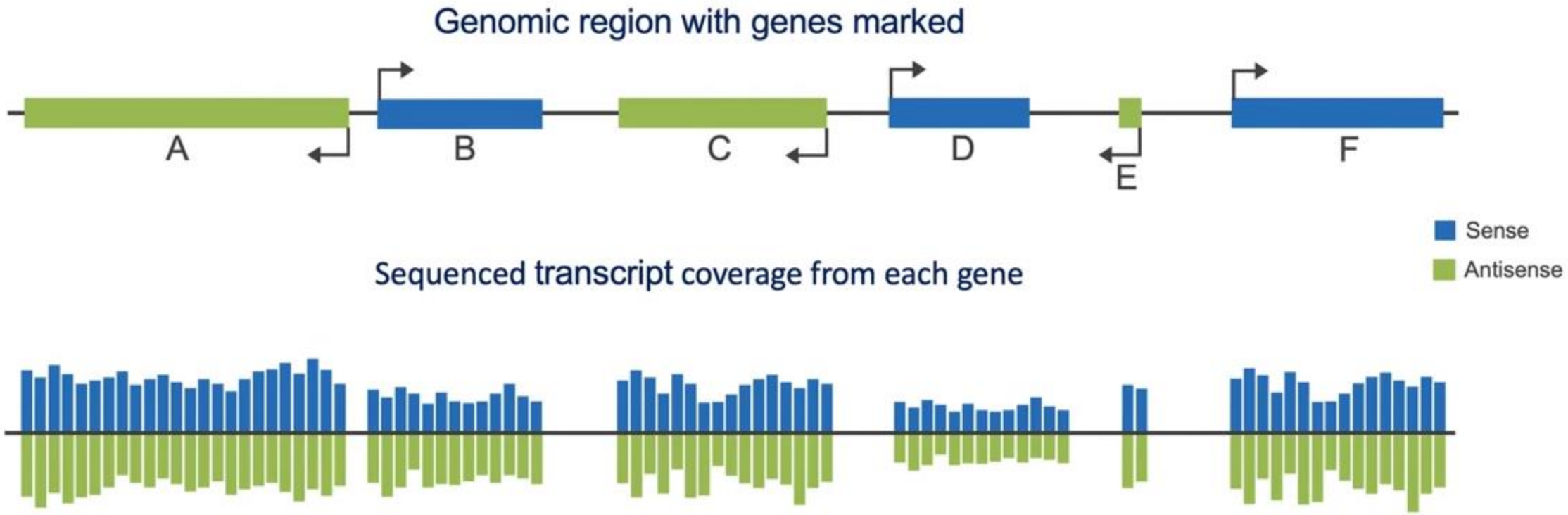
Non-Stranded

RNA-Seq standard o no direccional donde se pierde la información de la orientación del transcrito

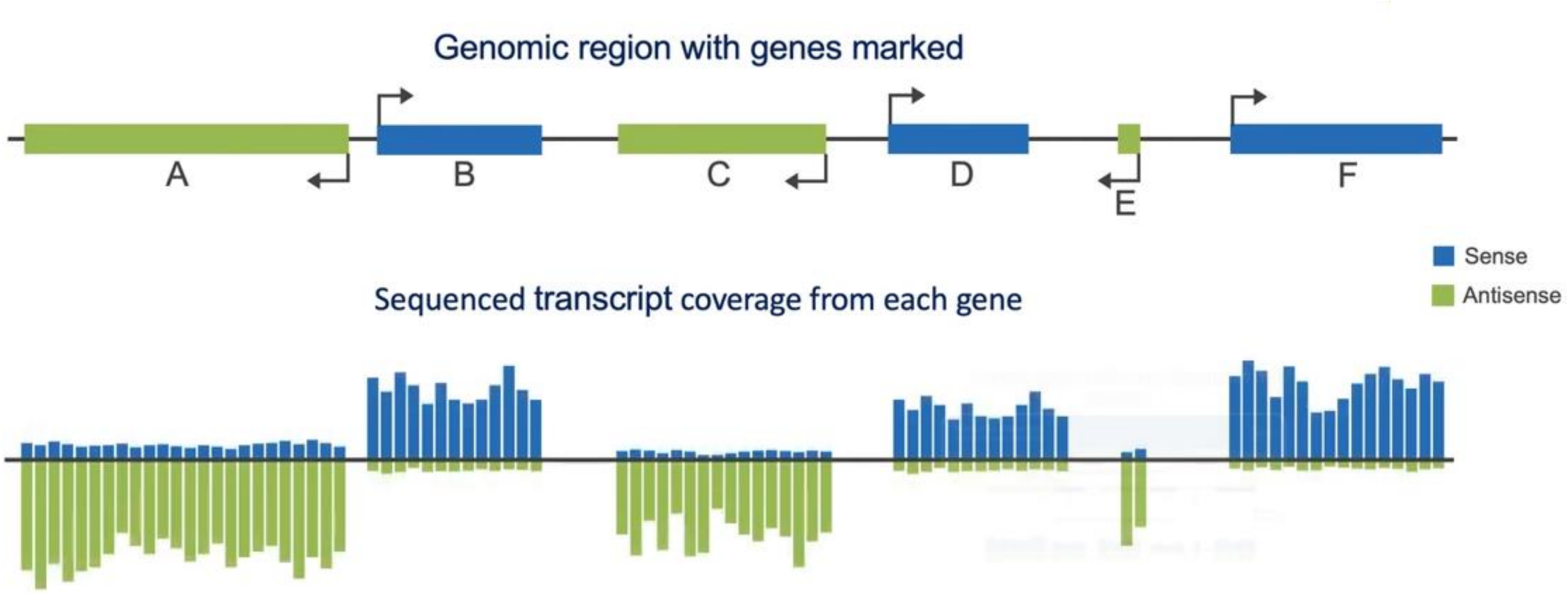
- Más barata
- Análisis de expresión génica (*non-overlapping genes*)

Image credit: Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

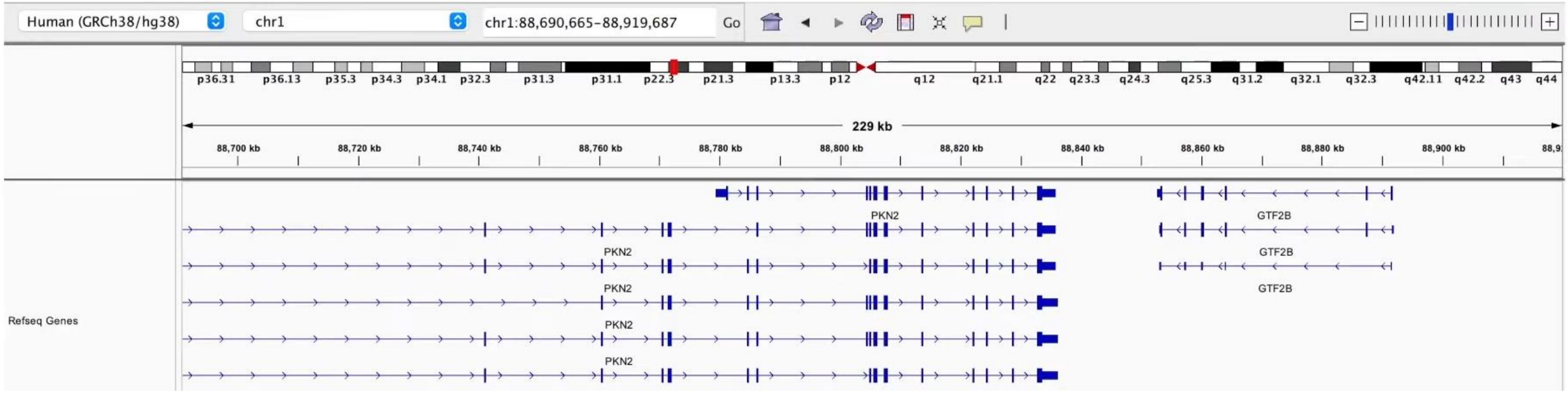
3 Stranded o non-stranded?



3 Stranded o non-stranded?

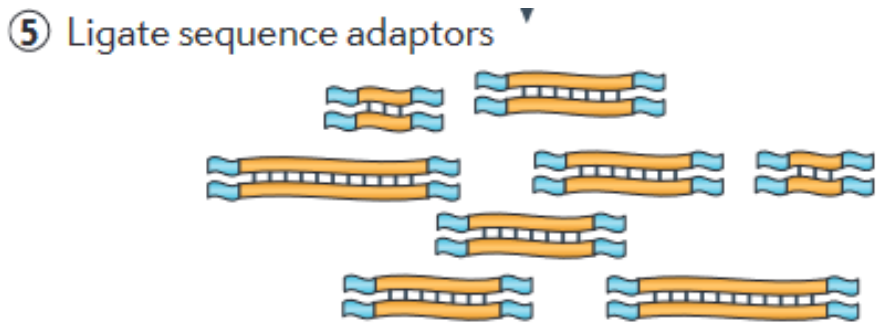


3 Solapamiento genético



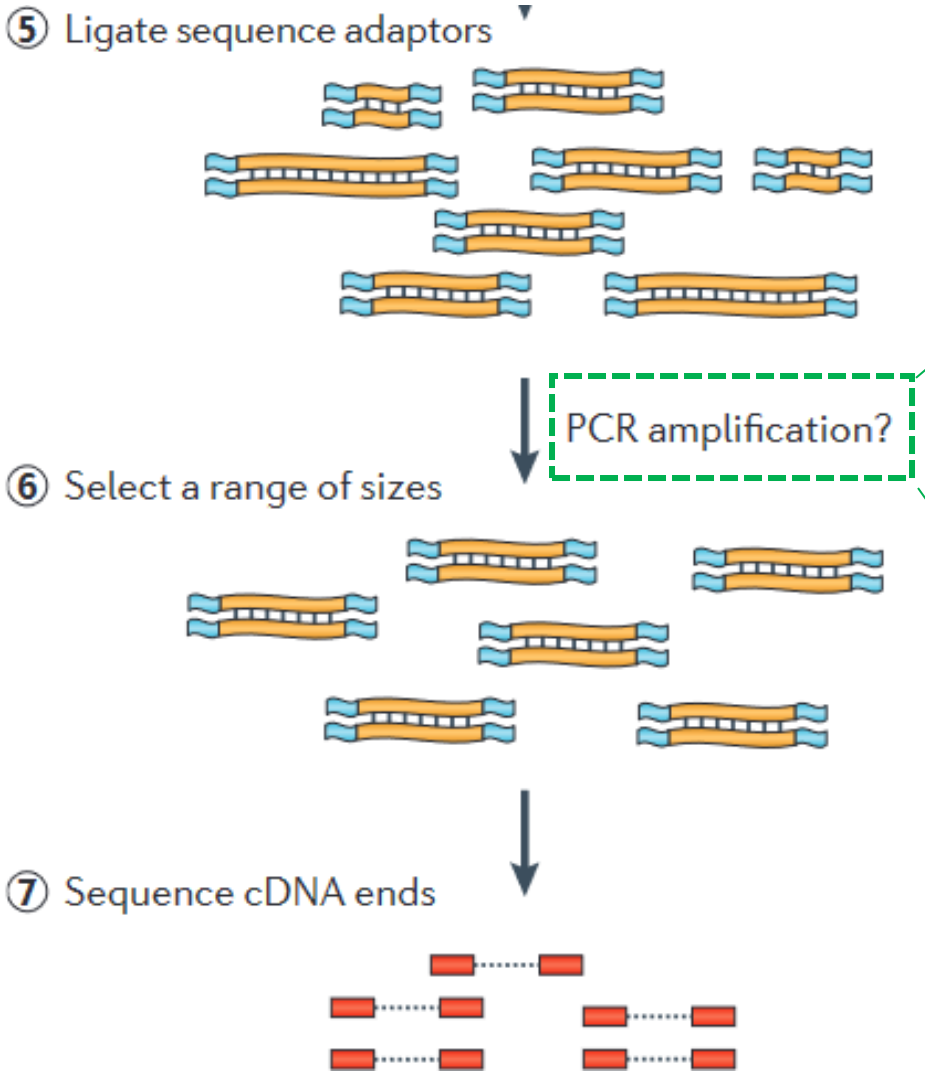
3 Solapamiento genético

3 Ligación de adaptadores



- **P5 y P7**: Secuencias que permiten que la biblioteca se una y genere grupos en la celda de flujo.
- **Secuencias de índice** que son **identificadores de muestra** que permiten la multiplexación/agrupación de varias muestras en un único ciclo de secuenciación.
- **Rd1 SP y Rd2 SP**: Sitios de unión de cebadores para iniciar la secuenciación.

3 Amplificación por PCR



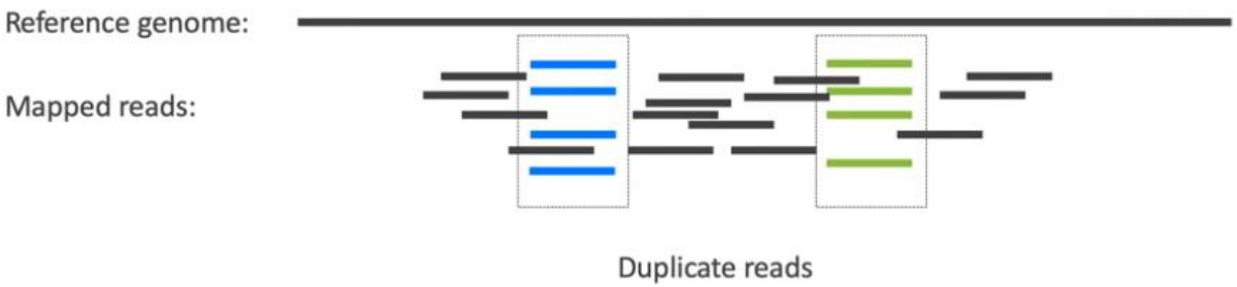
Para alcanzar una mínima cantidad de material de partida, toda la biblioteca de cDNA es amplificada por PCR hasta cierto nivel.

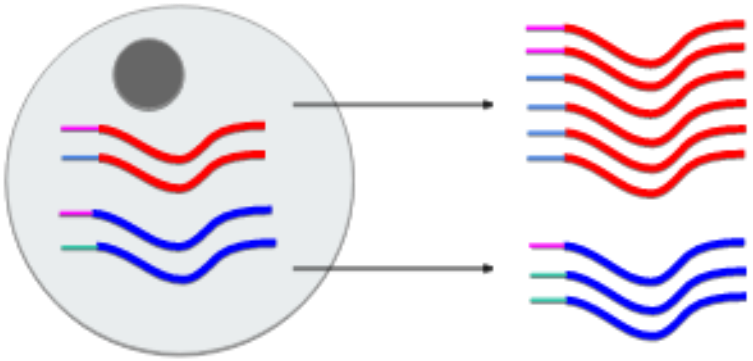
Esta amplificación puede evitarse eligiendo una modalidad de **"PCR-free"** al servicio de secuenciación, lo cual evita algunos artefactos como los **duplicados por PCR** en casos específicos.

3 Amplificación por PCR : Duplicados PCR

Duplicados técnicos -> artefactos de PCR / Secuenciación

Duplicados biológicos -> sobreexpresión de genes/secuencias repetitivas.



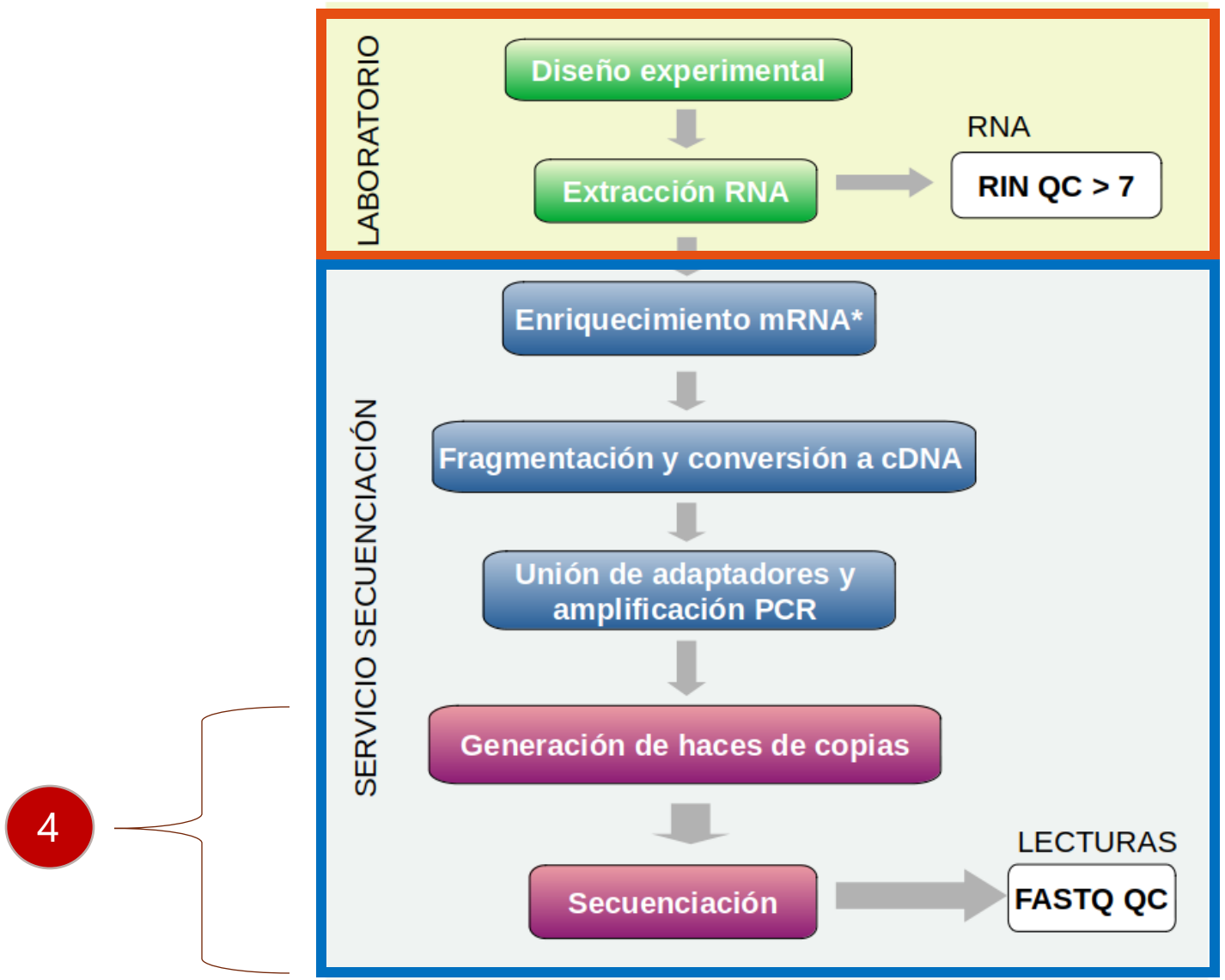


Duplicados por PCR

- Ignorar para datos de RNA-Seq.
- Use kits de preparación de bibliotecas sin PCR.
- Usar *Unique molecular identifiers* (UMI) .

Muchos experimentos de RNA-seq se analizan **sin eliminación** de los duplicados

Flujo de trabajo general de RNA-seq



4 Secuenciación

Los parámetros para la secuenciación, como la longitud de lectura, la configuración y la cantidad de lectura, dependen del objetivo de su proyecto e influirán en su elección del instrumento y la química de secuenciación

- Tipo de plataforma de secuenciación
- Paired-end o single-end sequencing
- Longitud de las lecturas
- Número total de lecturas

4 Secuenciación: Tipo de plataforma

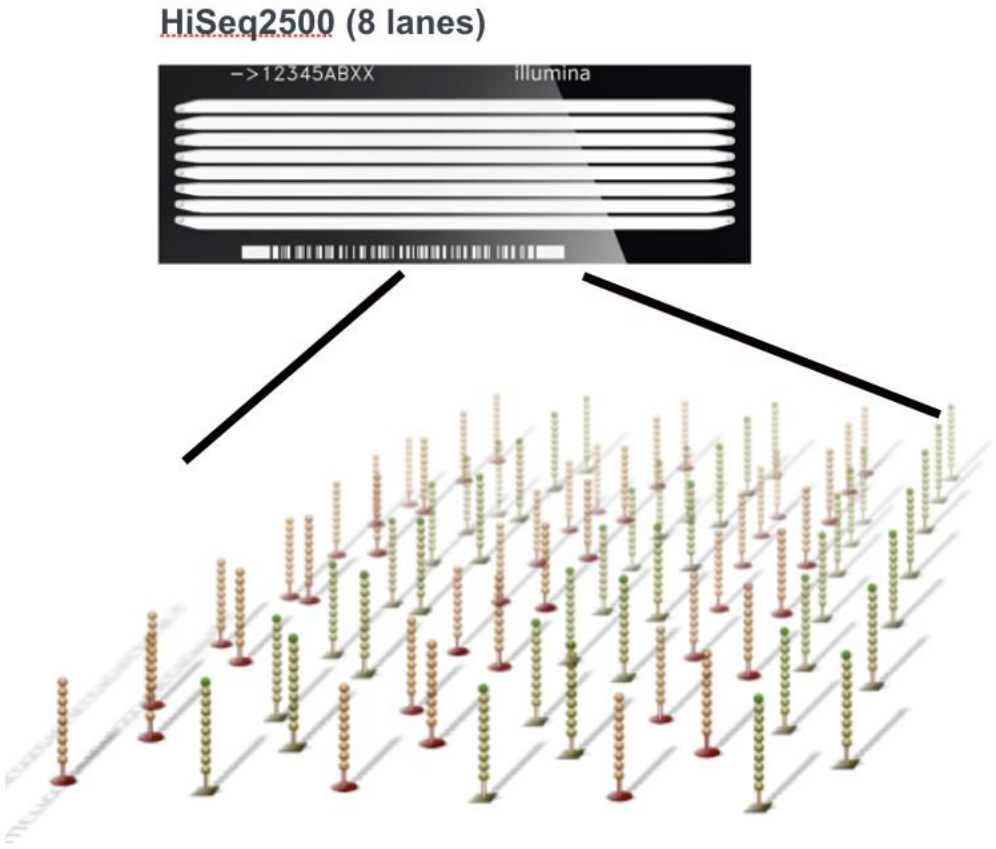
Type	Platform Provider	Advantages	Disadvantages
Short-read sequencing	Illumina	• High throughput • High accuracy • Low cost per base • Widely available	• Maximum single-end read length of 300 bp (600 bp paired-end)
	BGI/MGI DNBSEQ™	• Low cost per base	• Not available in US or Europe • Maximum single-end read length of 150 bp (300 bp paired-end)
Long-read sequencing	PacBio	• Contiguous, full-length transcript sequencing up to 10 kb • Lower error rate compared to other long-read sequencing technologies	• Higher cost • Low-throughput
	Oxford Nanopore	• Real-time, high-throughput transcript sequencing up to >20 kb • Cost-effective and portable	• Higher error rate compared to other long-read sequencing technologies

4 Secuenciación: Tipo de Secuenciador

Longitud de las lecturas - **calidad** de las lecturas - **número total de lecturas** obtenidas por ejecución - la cantidad de **tiempo** necesario para secuenciar las bibliotecas.

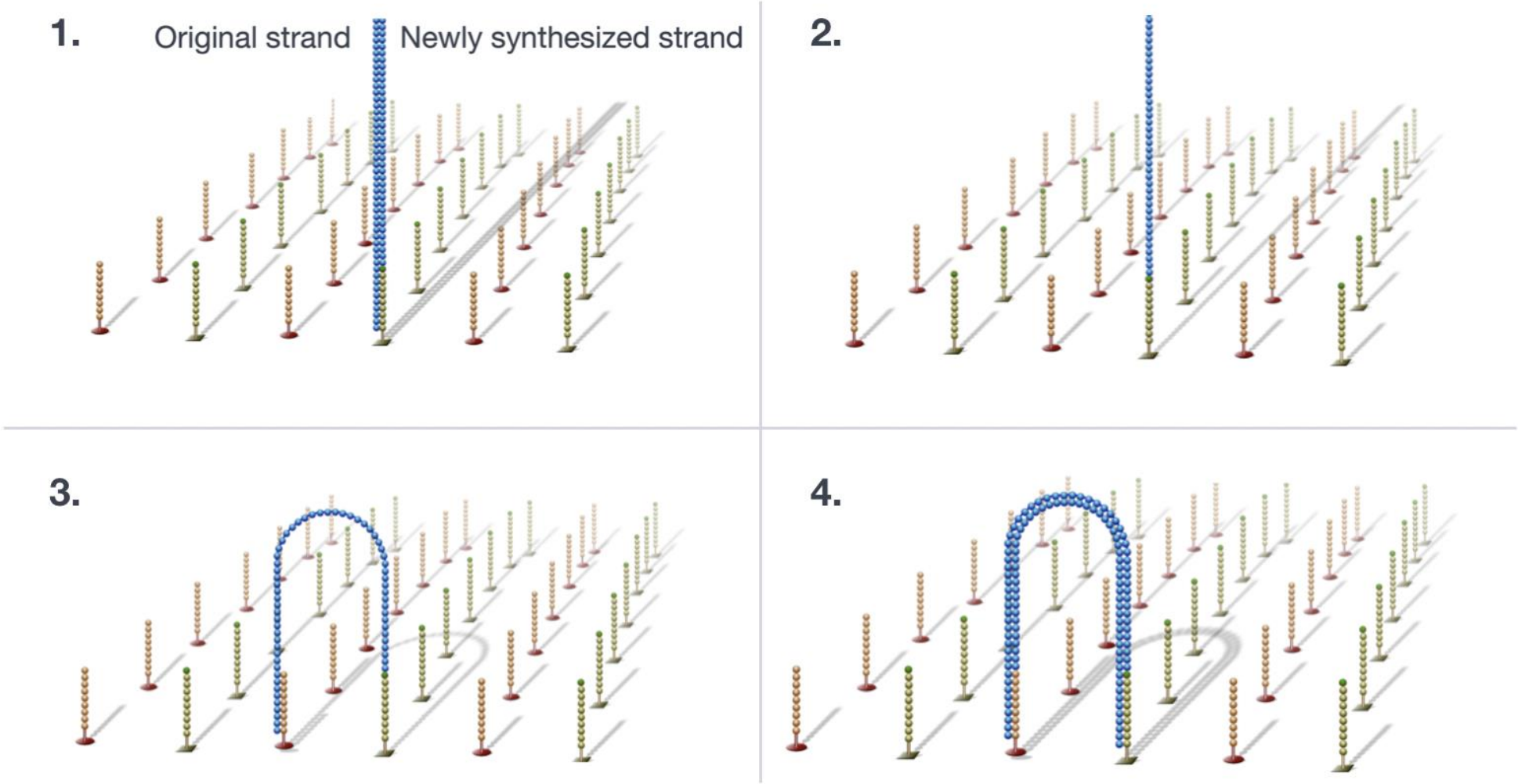
4 Secuenciación: Celda de flujo (*Flow cell*)

Celda de flujo (*Flow cell*): es una superficie de vidrio recubierta oligonucleótidos complementarios a los adaptadores agregados a sus moléculas de plantilla. La celda de flujo es donde tienen lugar las reacciones de secuenciación.

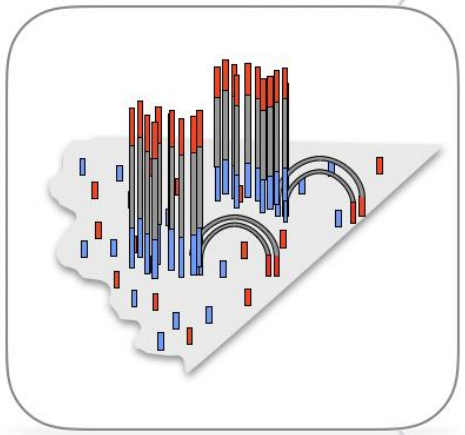


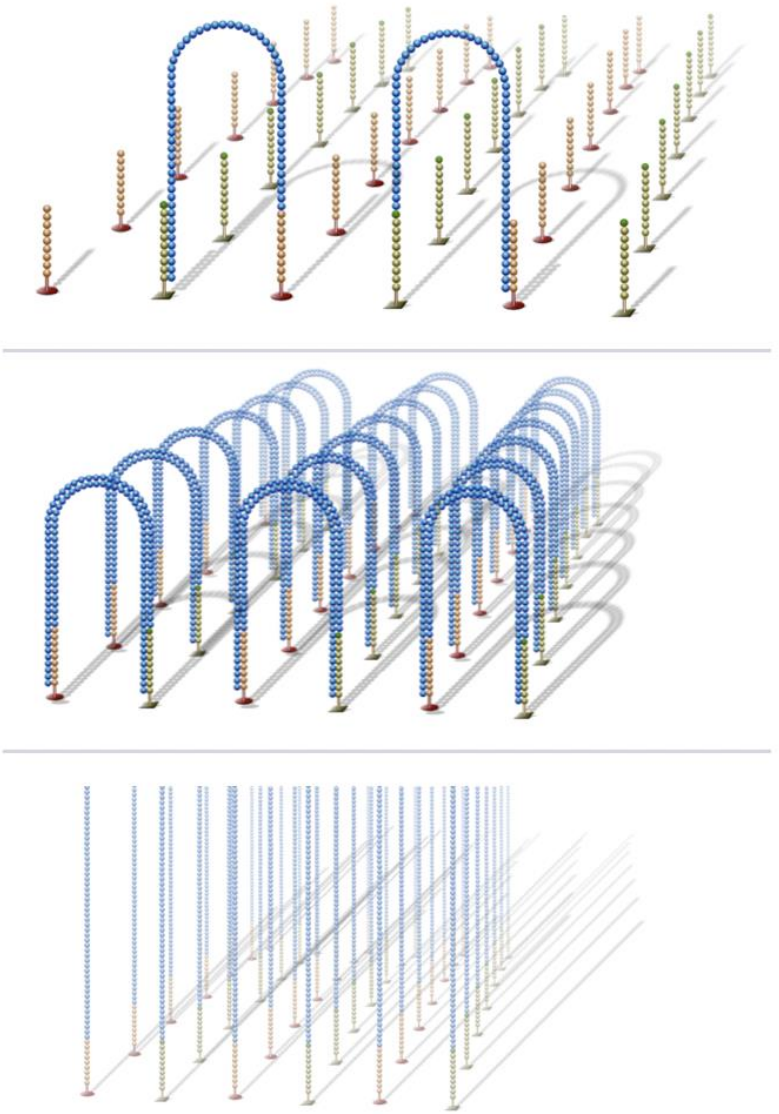
4 Secuenciación por síntesis: Ilumina

Generación de clústeres o de haces de copias



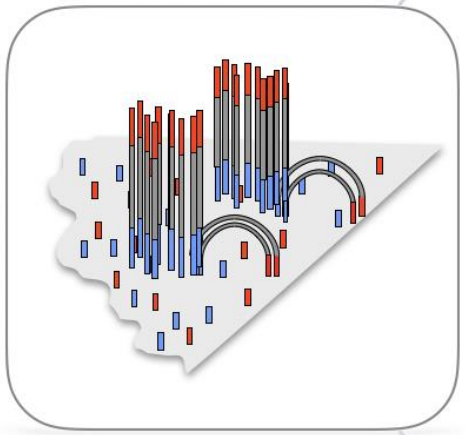
4 Secuenciación por síntesis: Ilumina

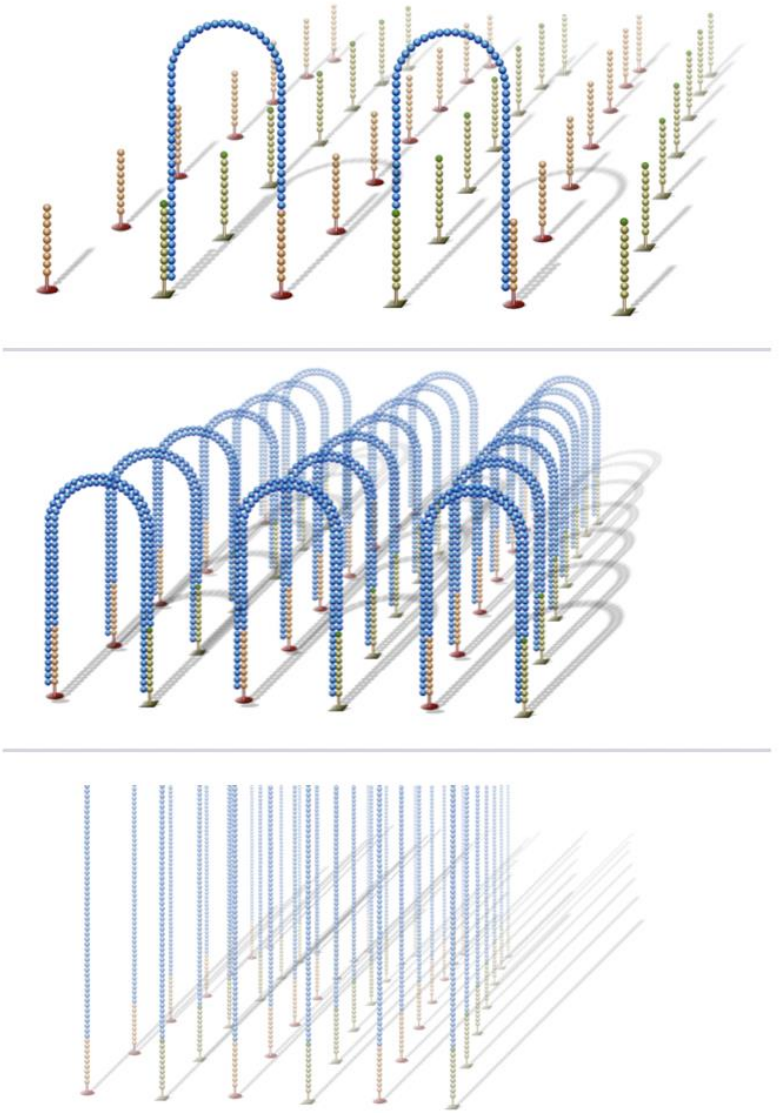




Tras la amplificación de la biblioteca, cada transcrito original está representado por cerca de mil copias de su cDNA.

4 Secuenciación por síntesis: Ilumina





La [concentración] de nuestra librería

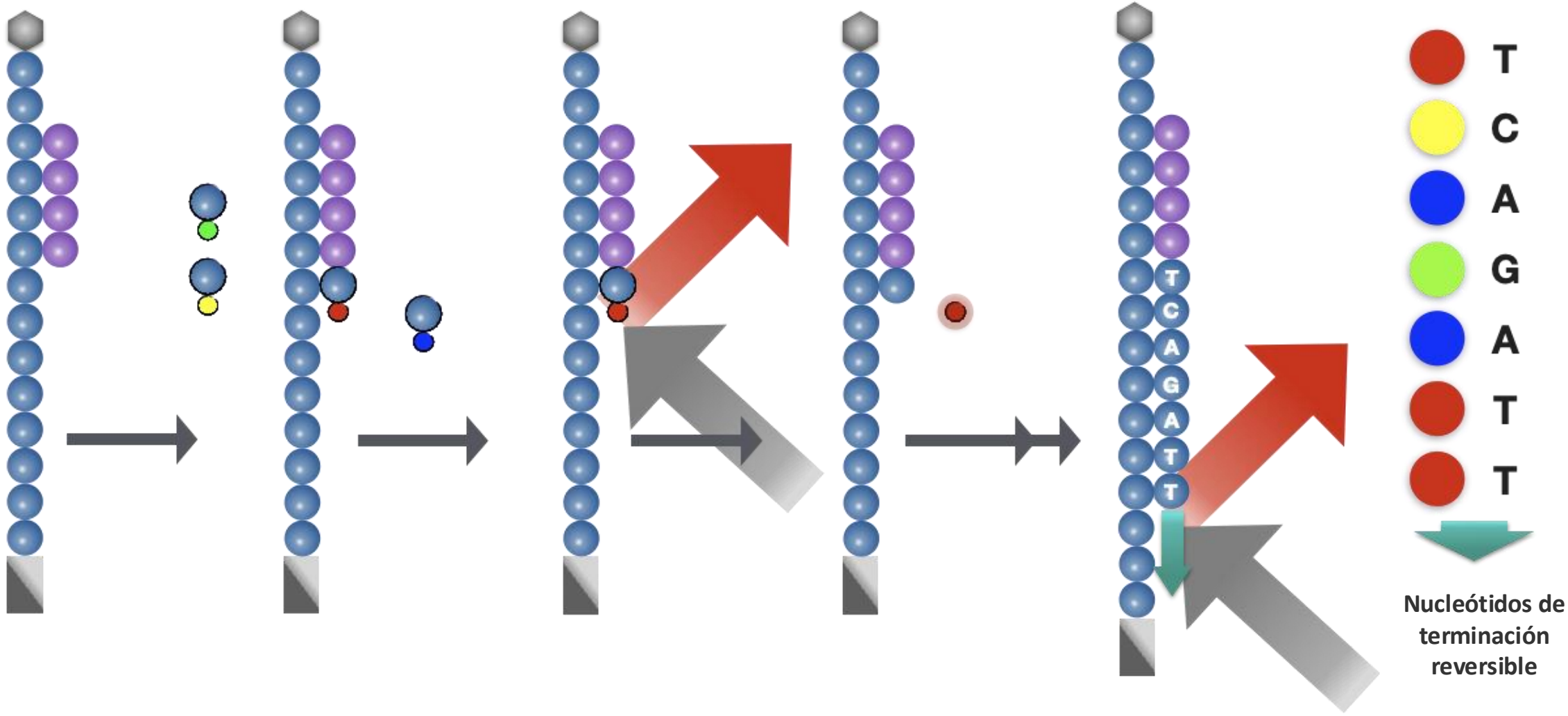
Overclustering

La densidad de los clusters es muy alta y puede hacer la secuenciación falle.

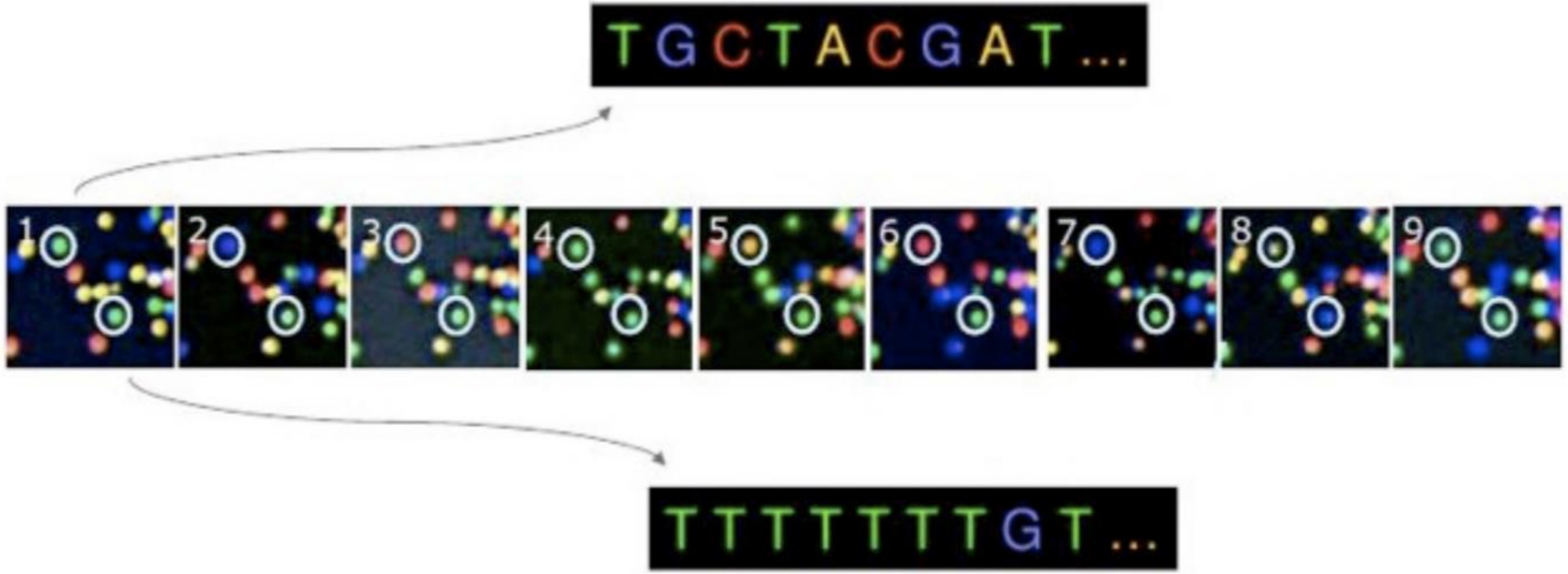
Underclustering

La densidad de los clusters es muy baja y comporta a un bajo número de lecturas finales.

4 Secuenciación por síntesis: Ilumina



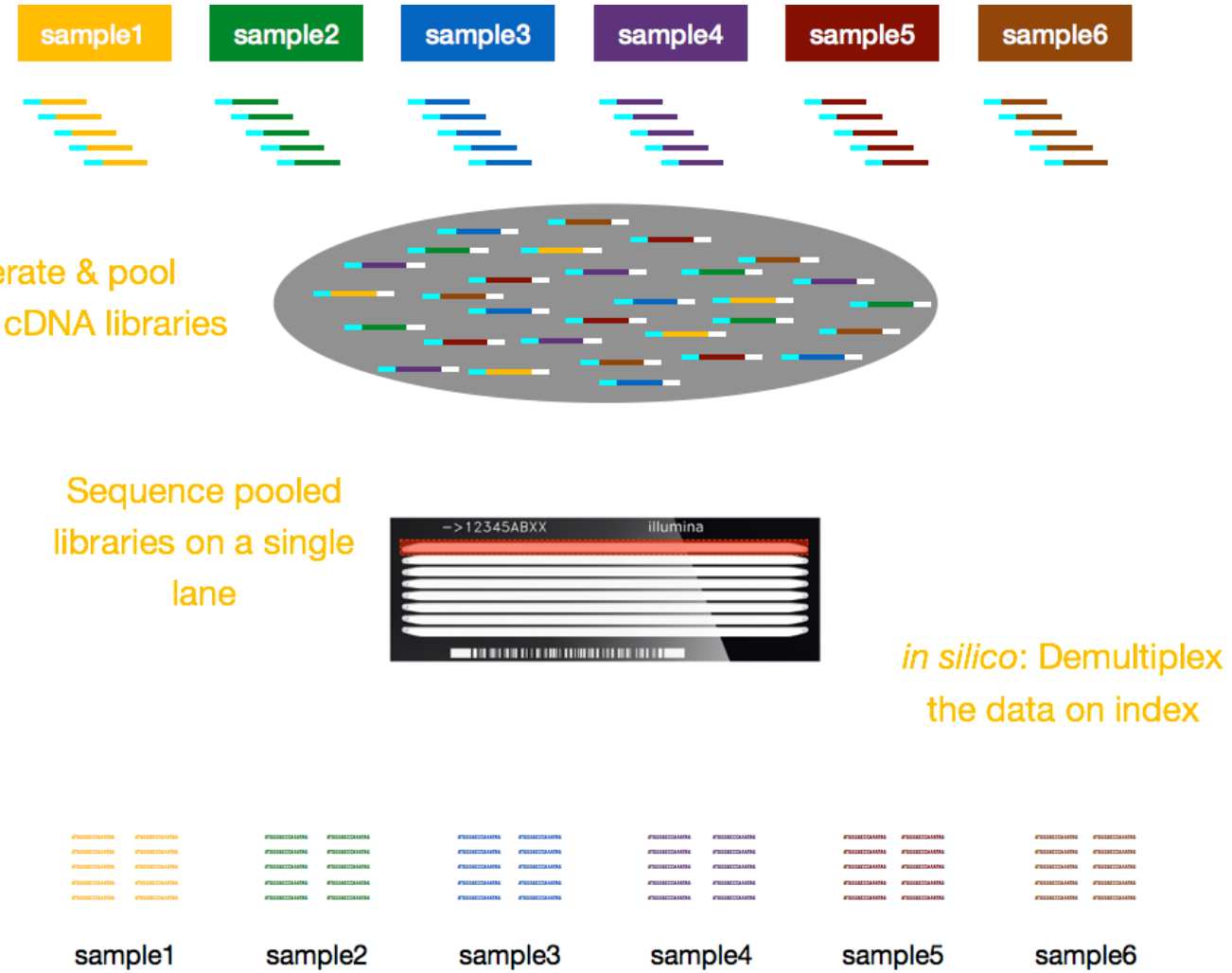
4 Secuenciación por síntesis: Ilumina



Número de clusters -> Número de lecturas

Número de ciclos -> La longitud de las secuencias

4 Secuenciación por síntesis: Ilumina Multiplexing



4 Secuenciación

Los parámetros para la secuenciación, como la longitud de lectura, la configuración y la cantidad de lectura, dependen del objetivo de su proyecto e influirán en su elección del instrumento y la química de secuenciación

- Tipo de plataforma de secuenciación
- Paired-end o single-end sequencing
- Longitud de las lecturas
- Número total de lecturas

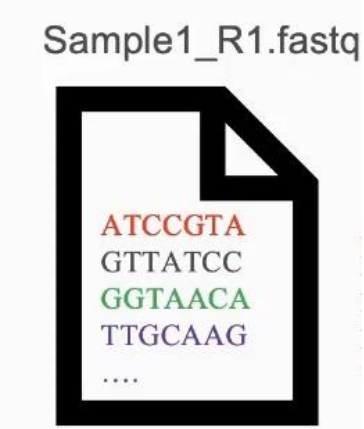
4 *Paired-end* o *single-end* sequencing

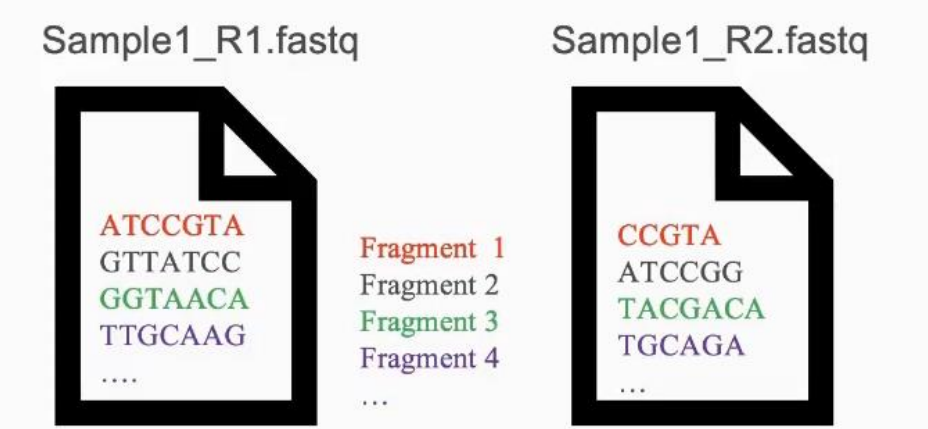
Single-end (SE)

- Conjunto de datos de extremo único
- Solo lectura 1 (1X)

Paired-end (PE)

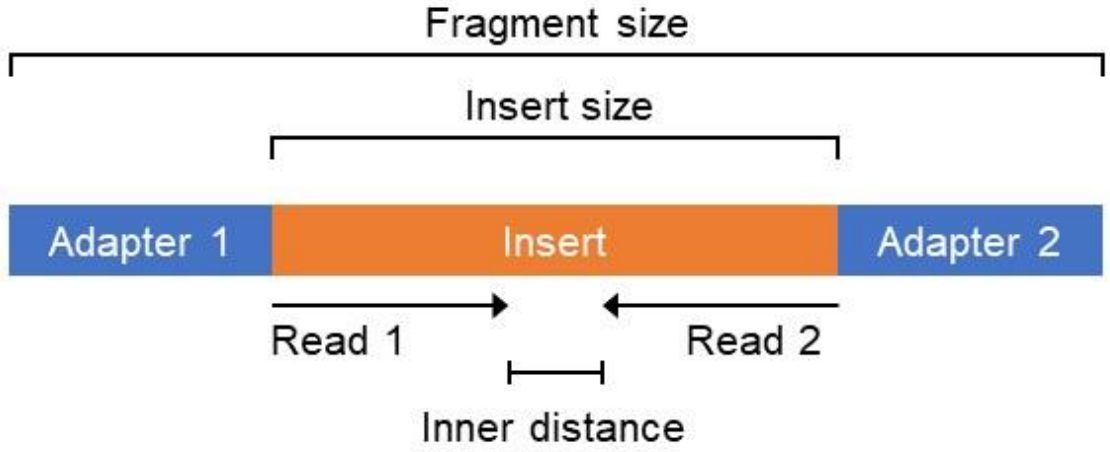
- Conjunto de datos de extremo emparejado
- Lectura 1 + Lectura 2 (2X)

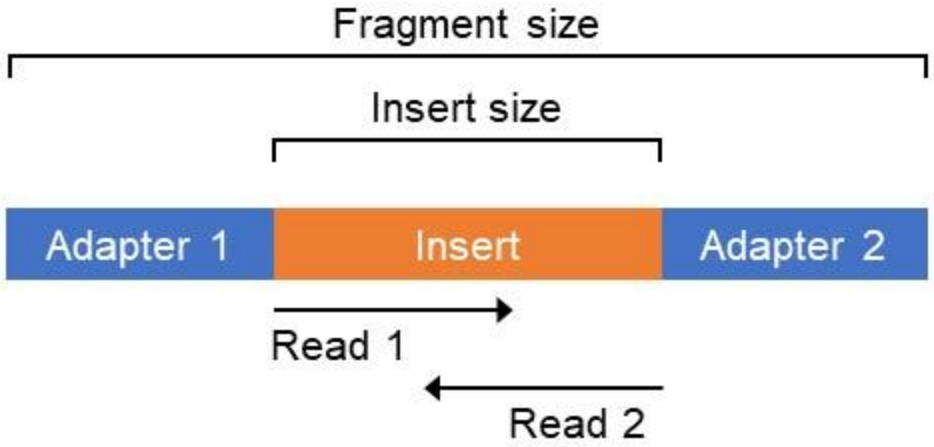


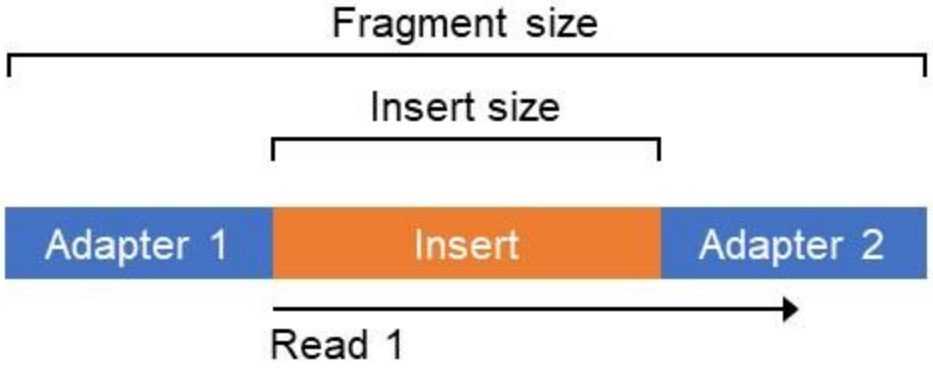


Las lecturas cortas *single-end* (SE) son suficientes para los estudios de los niveles de expresión génica en organismos bien anotados, mientras que las lecturas de *paired-end* (PE) son preferibles para descubrimiento de transcripciones de novo.

4 *Paired-end* o *single-end* sequencing







4 Longitud de las lecturas

La longitud de lectura se refiere al número de pares de bases (pb) secuenciados de un fragmento de ADN

50bp

Análisis de ARN pequeños

75/100nt

Análisis de expresión génica

150nt

Análisis de nuevos transcriptomas

4 Tamaño de la biblioteca (*library size*)

El **tamaño de la biblioteca (*library size*)** se refiere al número de lecturas secuenciadas para una muestra determinada

- Los experimentos de perfiles de expresión génica que buscan una **instantánea rápida** de genes altamente expresados pueden necesitar solo de 5 a 25 millones de lecturas por muestra.

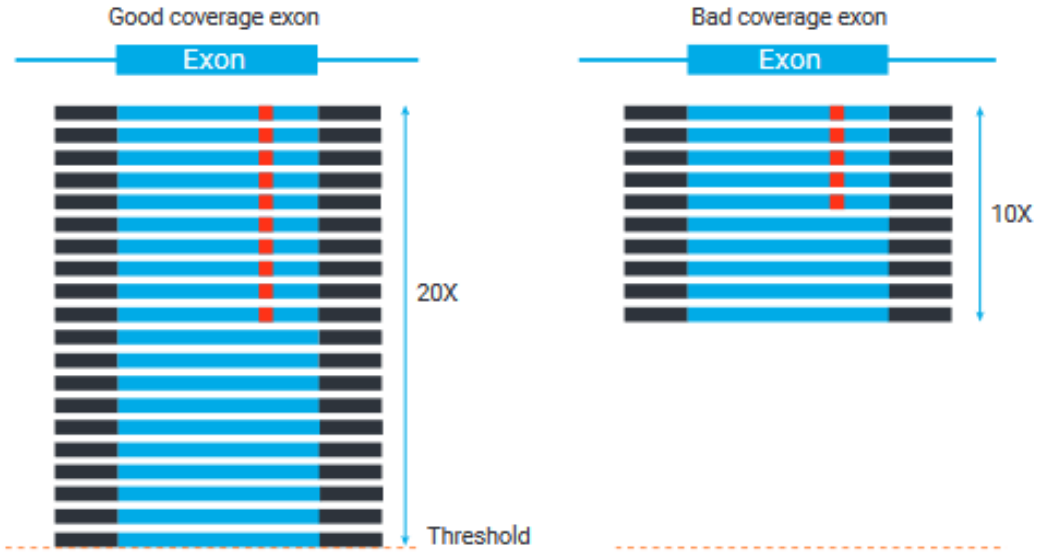
- Los experimentos que buscan una visión más global de la expresión génica y alguna información sobre el empalme alternativo normalmente requieren de 30 a 50 millones de lecturas por muestra. Este rango abarca la **mayoría de los experimentos de RNA-Seq** publicados para la secuenciación de ARNm/transcriptoma completo.

- Los experimentos que buscan obtener una vista detallada del transcriptoma o ensamblar nuevas transcripciones pueden requerir entre 100 y 200 millones de lecturas.

4 Otros parámetros a considerar

Profundidad de cobertura *(Depth of coverage)*

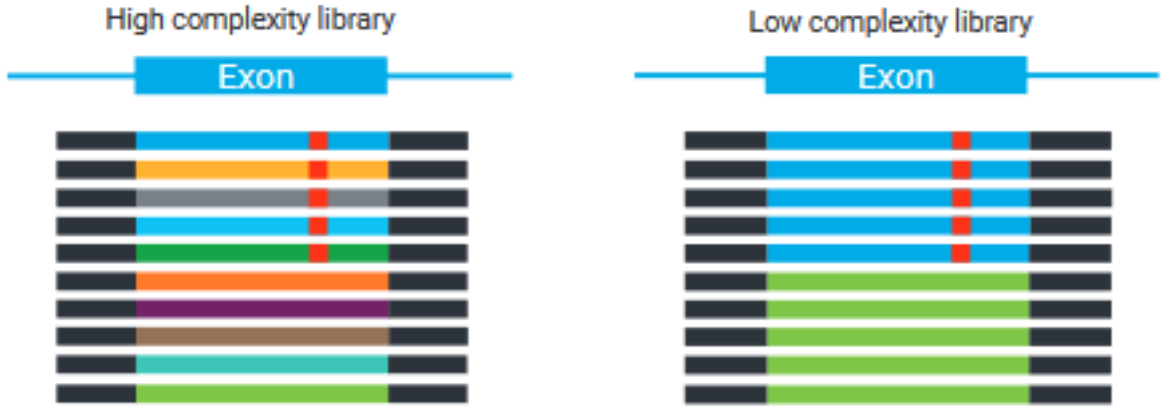
La profundidad de cobertura se define como el número promedio de veces que se lee una base. durante la secuenciación. Un mayor número de lecturas aumentará la confianza en datos de secuenciación



Complejidad de la librería *(library complexity)*

La diversidad de fragmentos de ADN únicos en nuestra librería

- Mayor representatividad muestral
- Mayor especificidad y sensibilidad.



4 Tarifas de preparación + secuenciación

Universidad Internacional de Valencia

1 Tarifas

Recursos y materiales/03. Materiales del profesor

cnag

Genomics Services

2021 Price list

Rev.1

Valid from: 01/07/2021

VAT not included

Item (CRG Code)	Item (CNAG Code)	Description	Ut of measure	Fees		
				Public Institutions subsidized by the Generalitat de Catalunya and/or Spanish Government	Other Public Institutions	Private Organizations
Quality control						
SERGIC100	SERCNAG127	Bioanalyzer run (DNA high sensitivity)	sample	6,09 €	6,83 €	7,39 €
SERGIC101	SERCNAG128	Bioanalyzer run (DNA standard sensitivity)	sample	4,19 €	4,70 €	5,11 €
SERGIC102	SERCNAG129	Bioanalyzer run (RNA)	sample	4,65 €	5,22 €	5,67 €
SERGIC266	SERCNAG208	DNA Concentration Normalization (96-well plate)	plate	61,15 €	68,49 €	73,44 €
SERGIC268	SERCNAG210	DNA quantification (96-well plate)	plate	130,20 €	145,84 €	156,35 €
SERGIC106	SERCNAG46	Library Quality Control	sample	25,15 €	28,18 €	30,28 €
SERGIC206	SERCNAG151	Library Quantification (qPCR)	sample	18,35 €	20,55 €	22,02 €
SERGIC104	SERCNAG44	Sample Quality Control (RNA)	sample	9,45 €	10,60 €	11,45 €
SERGIC103	SERCNAG43	Sample Quality Control (DNA)	sample	7,81 €	8,76 €	9,48 €
SERGIC105	SERCNAG45	Sample quality Control (low integrity samples)	sample	31,82 €	35,64 €	38,24 €
SERGIC196	SERCNAG131	DNA Purification Beads	mL	171,51 €	192,52 €	210,13 €
SERGIC295	SERCNAG238	Genomic DNA extraction	sample	71,80 €	80,42 €	86,19 €
SERGIC299	SERCNAG242	Sample Quality Control HMW gDNA	sample	46,88 €	52,56 €	56,82 €
SERGIC320	SERCNAG256	Sample Quality Control (DNA, 96 well-plate)	set of samples	324,05 €	363,46 €	394,16 €
SERGIC321	SERCNAG257	Sample Quality Control (Low integrity samples, 96 well-plate)	set of samples	1.250,02 €	1.400,27 €	1.502,51 €
SERGIC322	SERCNAG258	Sample Quality Control (RNA, 96 well-plate)	set of samples	310,14 €	347,52 €	373,81 €
SERGIC355	SERCNAG294	Nuclei extraction	sample	64,62 €	72,40 €	77,77 €

2

4 Tarifas de preparación + secuenciación

Número de muestras

- **15 muestras**

Protocolo para la preparación de la librería

- En primer lugar, se ha llevado a cabo un **enriquecimiento del mRNA** mediante las colas poli(A) y posteriormente la transcripción a **cDNA específica de cadena** o *stranded* mediante el kitNEBNext Ultra II Directional RNA de Illumina.
- La calidad de la librería se ha analizado en un Bioanalizador usando un kit de DNA de alta sensibilidad y la cantidad de la librería empleando qPCR.

Protocolo de secuenciación

- Las muestras agrupadas se secuenciaron empleando HiSeq v4 de Illumina (1 lane promociona unos 350 M de reads) en el modo de 1 X 100 pb single-end. Se obtuvieron una media de 22 millones de lecturas por muestra.

Institución pública subvencionada por el gobierno de España *(primera columna)*

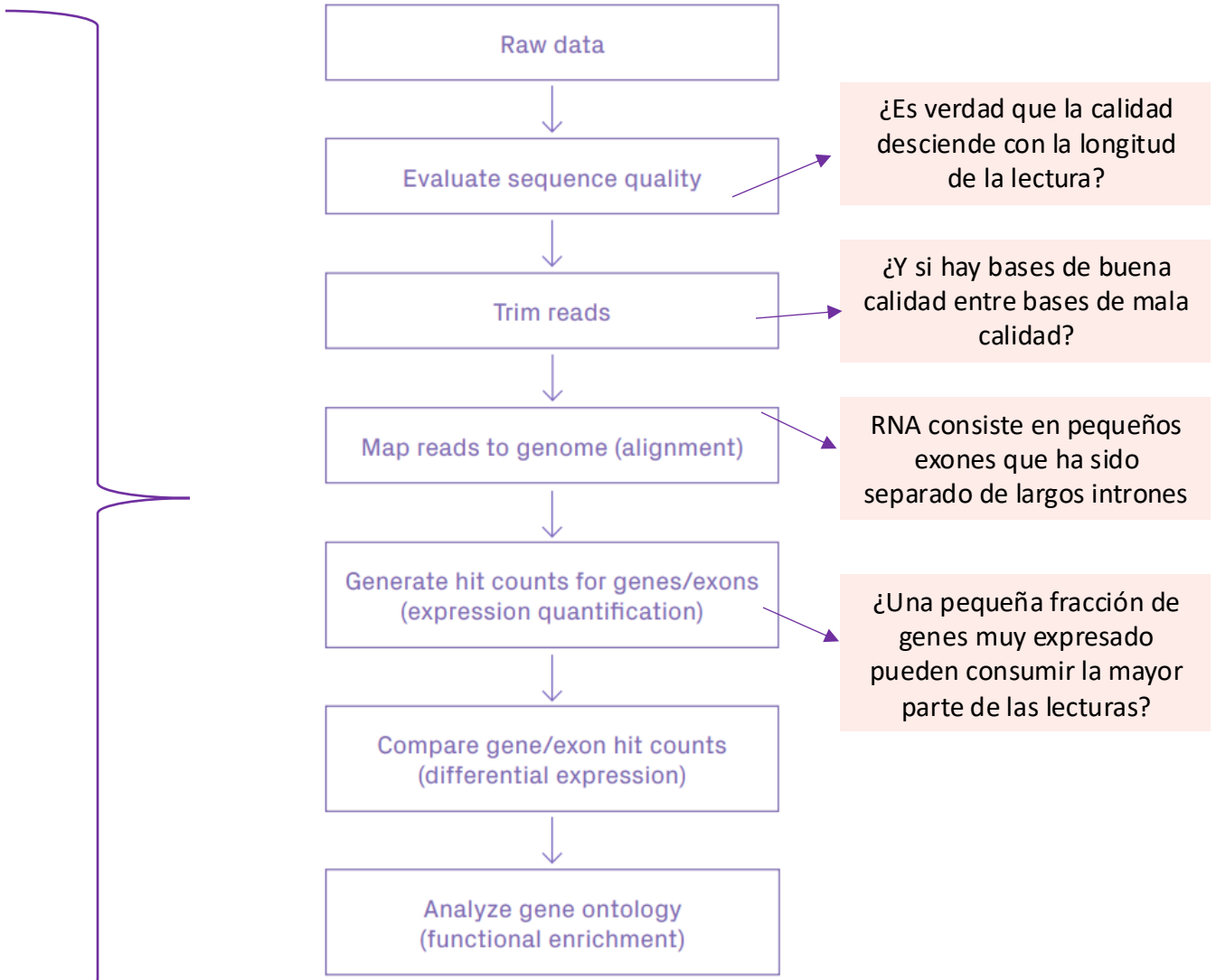
4 Tarifas de preparación + secuenciación

Concepto	Precio	Total
Preparación de las librerías de mRNA (stranded)	85,04€/muestra *15	1275,6
Análisis de la calidad (Bioanalyzer alta sensibilidad)	5,74 €/muestra *15	86,1
Cuantificación de la librería (qPCR)	18,52 €/muestra *15	277,8
Sequencing lane HiSeq v4 (1* 100)	963,31 € * 1 lane	963,31
	Total neto	2602,81
	IVA (21%)	
	TOTAL	**3149,4**

5 Análisis de los datos

La evaluación de la calidad de sus datos y extracción información biológicamente relevante es el paso más gratificante en un experimento de RNA-Seq.

Es importante discutir su proyecto con un bioinformático experimentado para encontrar y determinar el mejor flujo de trabajo computacional.



Universidad
Internacional
de Valencia

universidadviu.com

De:
Planeta Formación y Universidades