

Análisis transcriptómicos de la expresión génica

Máster Universitario en Bioinformática

Sesión 3



**Universidad
Internacional
de Valencia**

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

De:
 Planeta Formación y Universidades



Bloque II: Estudios de expresión génica con datos de NGS



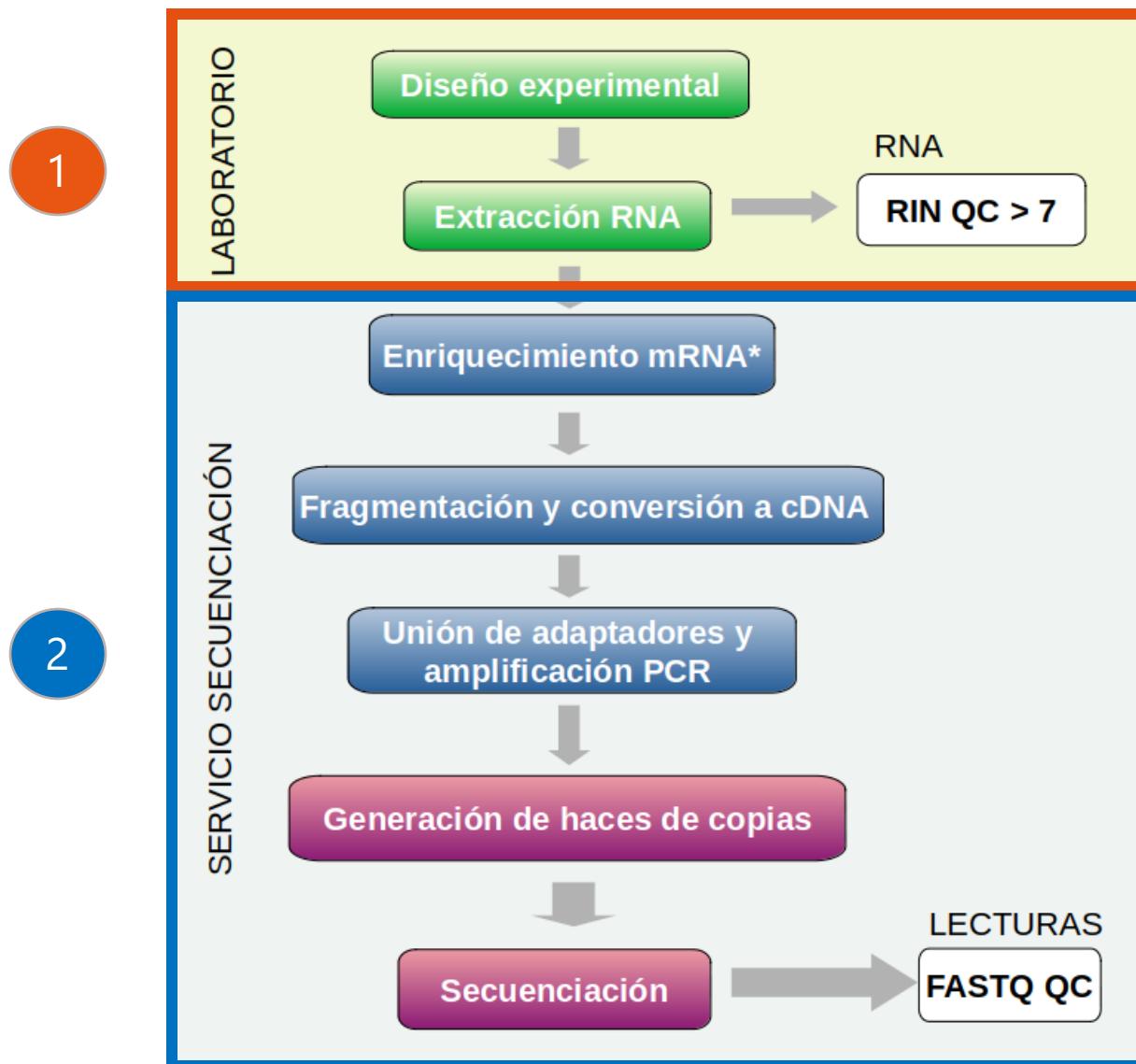
Objetivo de la sesión

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
- Análisis de datos

Flujo de trabajo general de RNA-seq





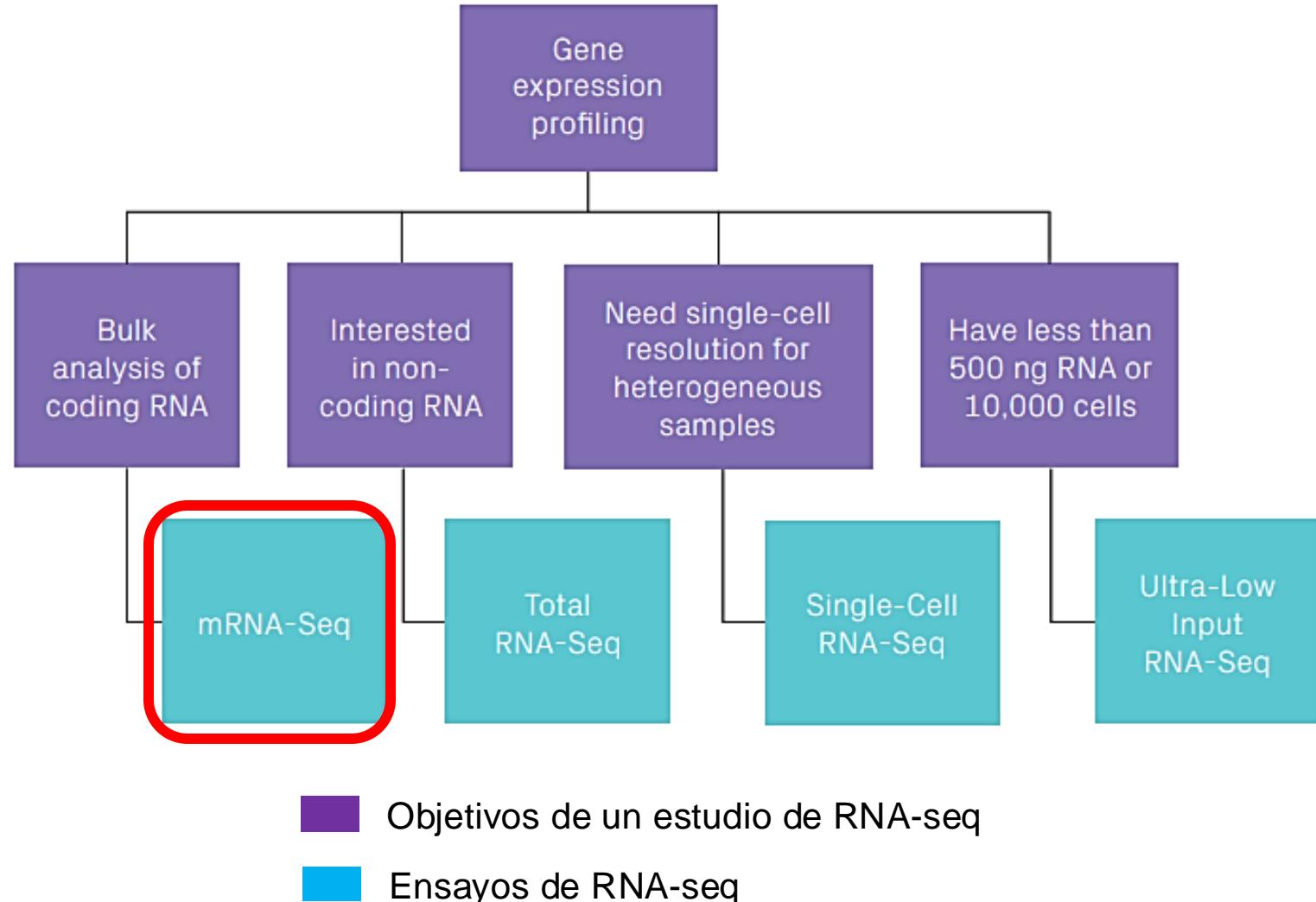


Un requisito crucial para un estudio **exitoso** de RNA-seq es que los datos generados tengan el potencial de **responder** las preguntas biológicas de interés y esto se logra definiendo primero un **buen diseño** experimental

¿Qué tipo de ensayo de RNA-seq vamos a emplear?

¿Cuántas réplicas son necesarias realizar?

¿Qué tipo de ensayo de RNA-seq vamos a emplear?



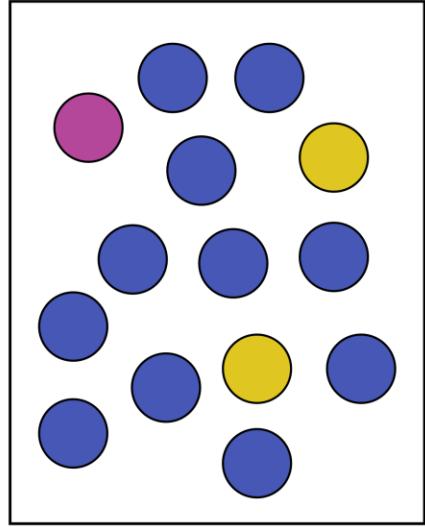


Un requisito crucial para un estudio **exitoso** de RNA-seq es que los datos generados tengan el potencial de **responder** las preguntas biológicas de interés y esto se logra definiendo primero un **buen diseño** experimental

¿Qué tipo de ensayo de RNA-seq vamos a emplear?

¿Cuántas réplicas son necesarias realizar?

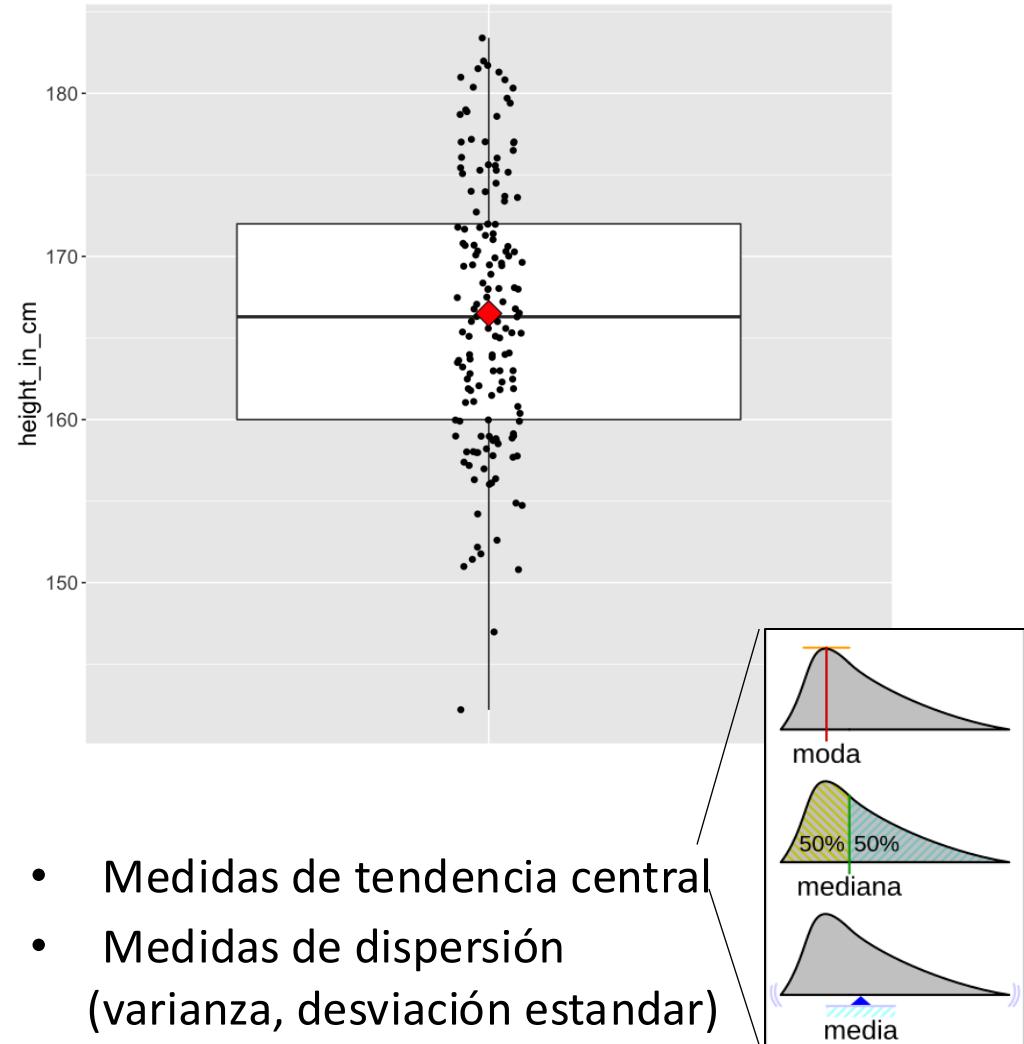
¿Cuál es la población a estudio?



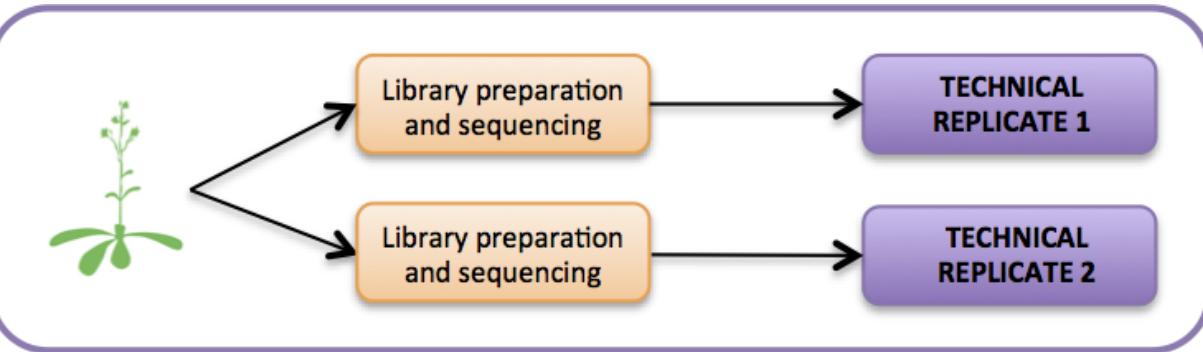
Population

Una muestra es un subconjunto de esa población que debe ser **representativa** de esa población. Por lo tanto, debe seleccionarse **aleatoriamente** entre toda la población.

Altura de la población española ($n = 1000$)



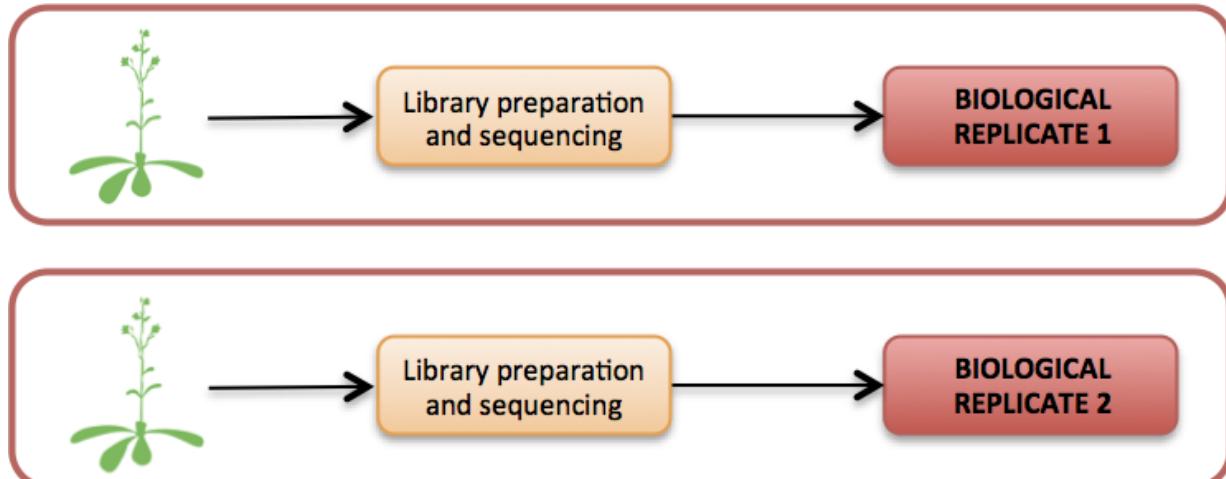
¿Cuántas réplicas son necesarias realizar?



Variabilidad técnica

Las réplicas técnicas son réplicas en las que el material biológico es el mismo, pero se repiten los pasos técnicos utilizados para medir la expresión génica.

Innecesarias las réplicas técnicas (solo controles internos)



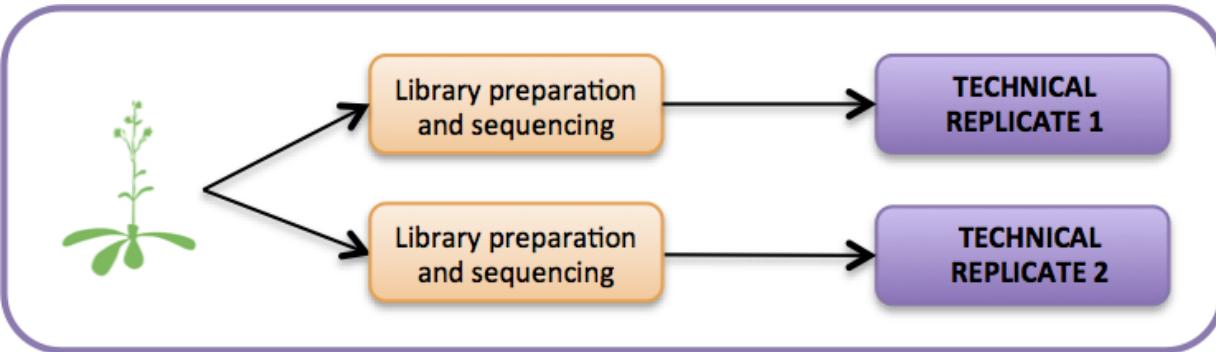
Variabilidad biológica

Las réplicas biológicas consisten en diferentes muestras biológicas que se procesan por separado.

Maximizar el número de réplicas biológicas



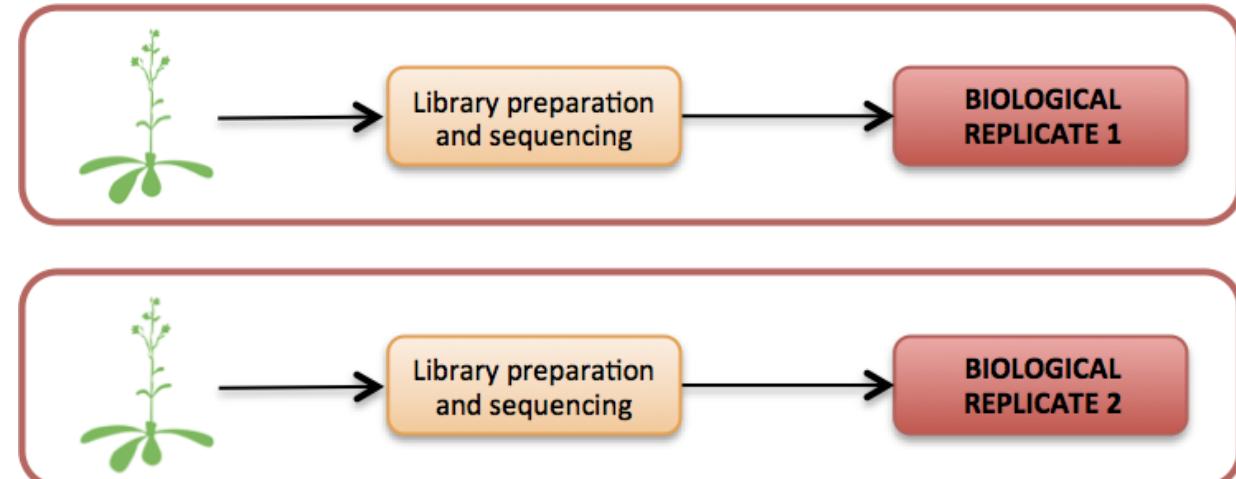
¿Cuántas réplicas son necesarias realizar?



Variabilidad técnica

Las réplicas técnicas son réplicas en las que el material biológico es el mismo, pero se repiten los pasos técnicos utilizados para medir la expresión génica.

Innecesarias las réplicas técnicas (solo controles internos)



Variabilidad biológica

Las réplicas biológicas consisten en diferentes muestras biológicas que se procesan por separado.

Maximizar el número de réplicas biológicas

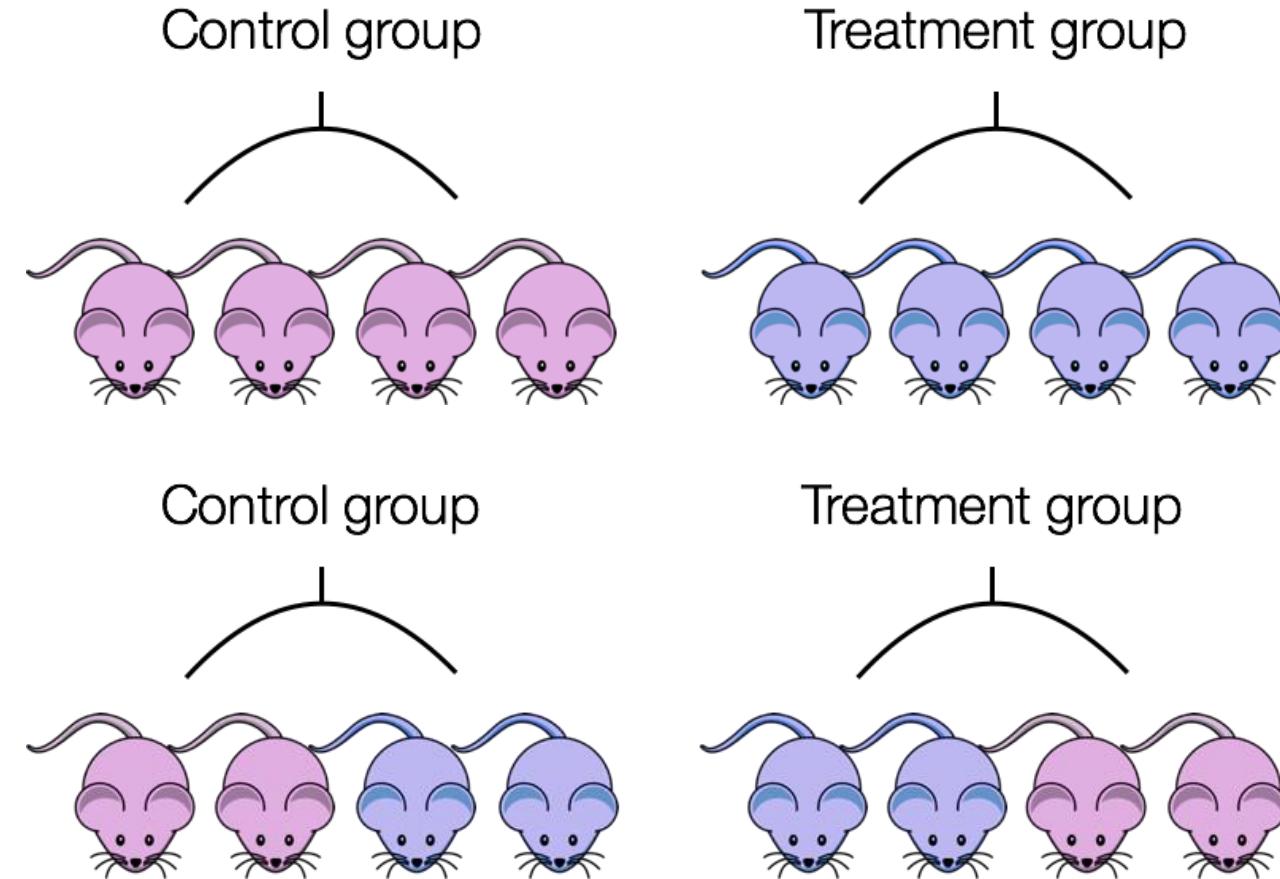
- 1 réplica sola para análisis exploratorios.
- >=2 réplicas para líneas celulares.
- >=3 réplicas para muestras animales.
- 3-6 réplicas para muestras humanas.

Factores de confusión (confounding factors)

- Son variables biológicas externas que pueden afectar la expresión génica y **no** son el foco principal del estudio.
 - **Ejemplos:** edad, sexo, estado de salud, condiciones ambientales, etc.

No podemos diferenciar el efecto del tratamiento del efecto del sexo

Dividir los animales por igual entre condiciones e intentar que comparten la mayor parte de características



Efecto de lote (*Batch effect*)

El efecto de lote representa las **diferencias técnicas sistemáticas** cuando las muestras se procesan y miden en diferentes lotes y que **NO** están relacionadas con ninguna variación biológica.

se refiere a las variaciones sistemáticas en los datos que no están relacionadas con las condiciones biológicas bajo estudio, sino con diferencias técnicas introducidas durante la preparación de las muestras, el procesamiento o la adquisición de los datos.

¿Qué puede causar *Batch effect*?

¿Se realizaron todos los preparativos de la biblioteca el mismo **día**?

¿La misma **persona** realizó el aislamiento/preparación de la biblioteca de ARN para todas las muestras?

¿Utilizó los mismos **reactivos** para todas las muestras?

¿Realizó el aislamiento/preparación de la biblioteca de ARN en el mismo **lugar**?

Efecto de lote (*Batch effect*)

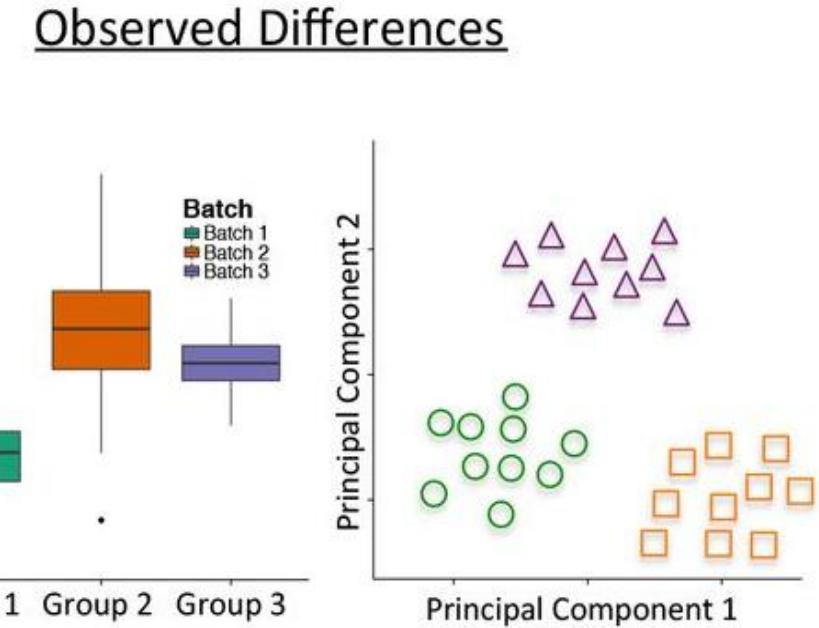
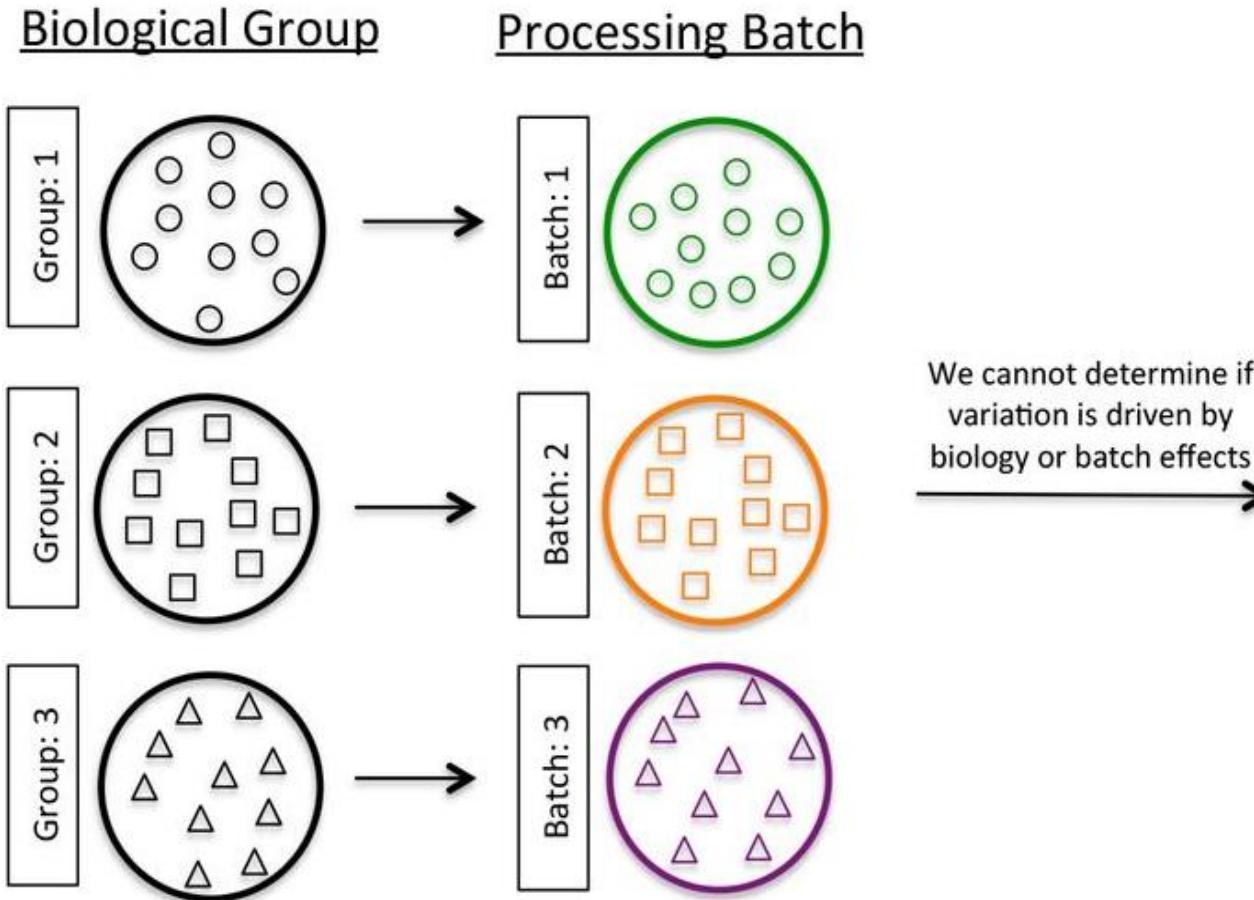
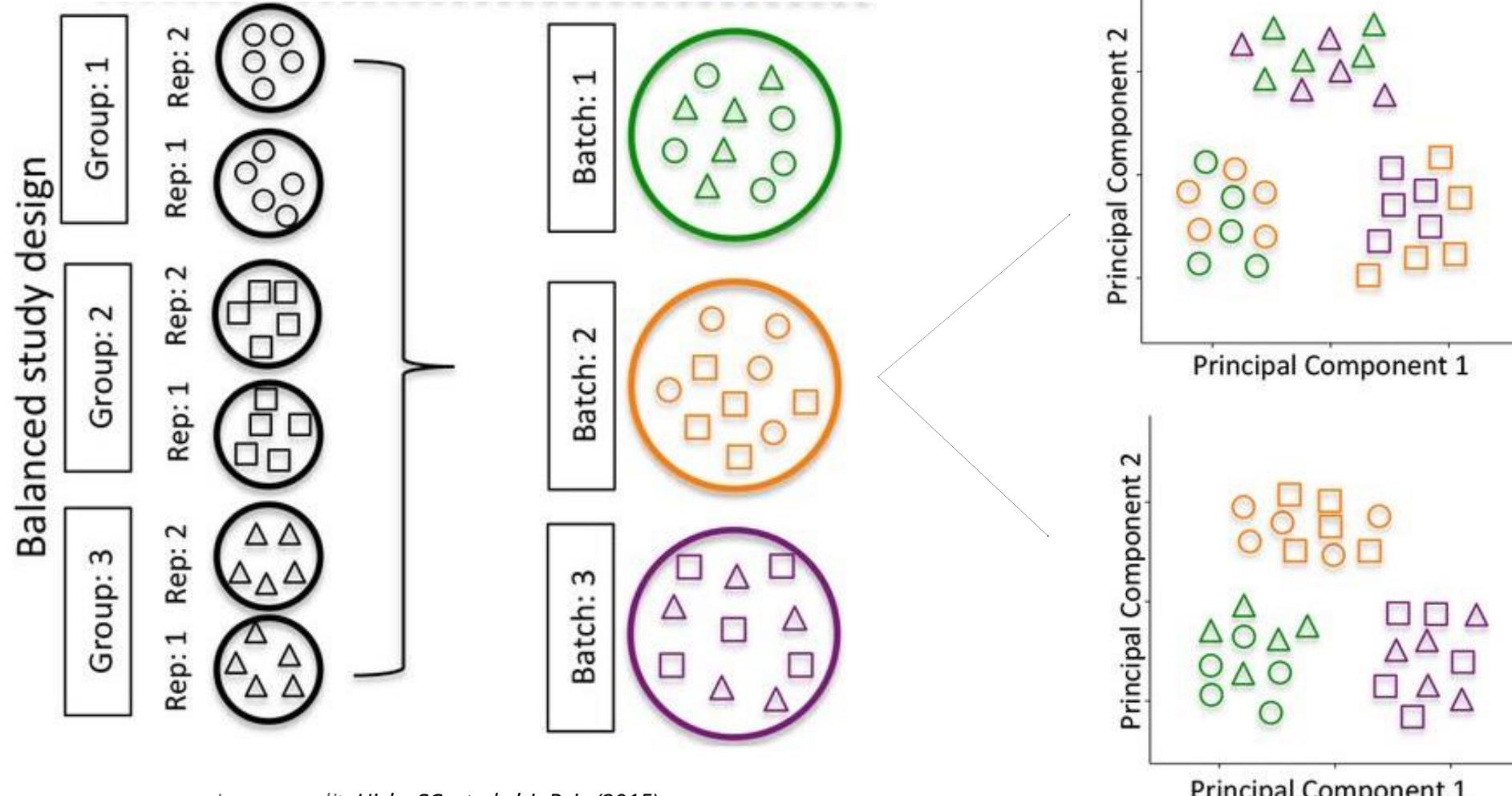
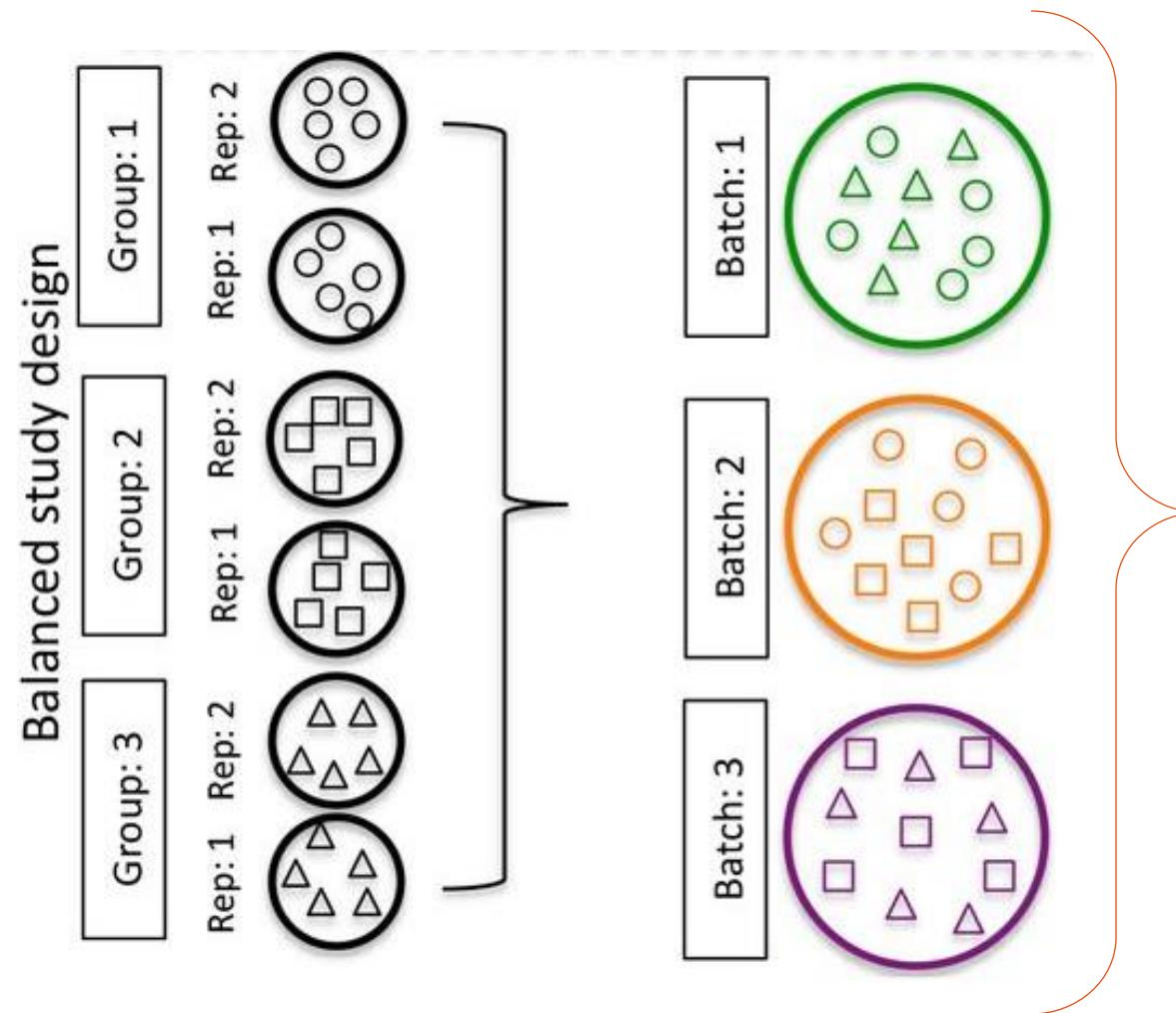


Image credit: Hicks SC, et al., *bioRxiv* (2015)

Efecto de lote (*Batch effect*)



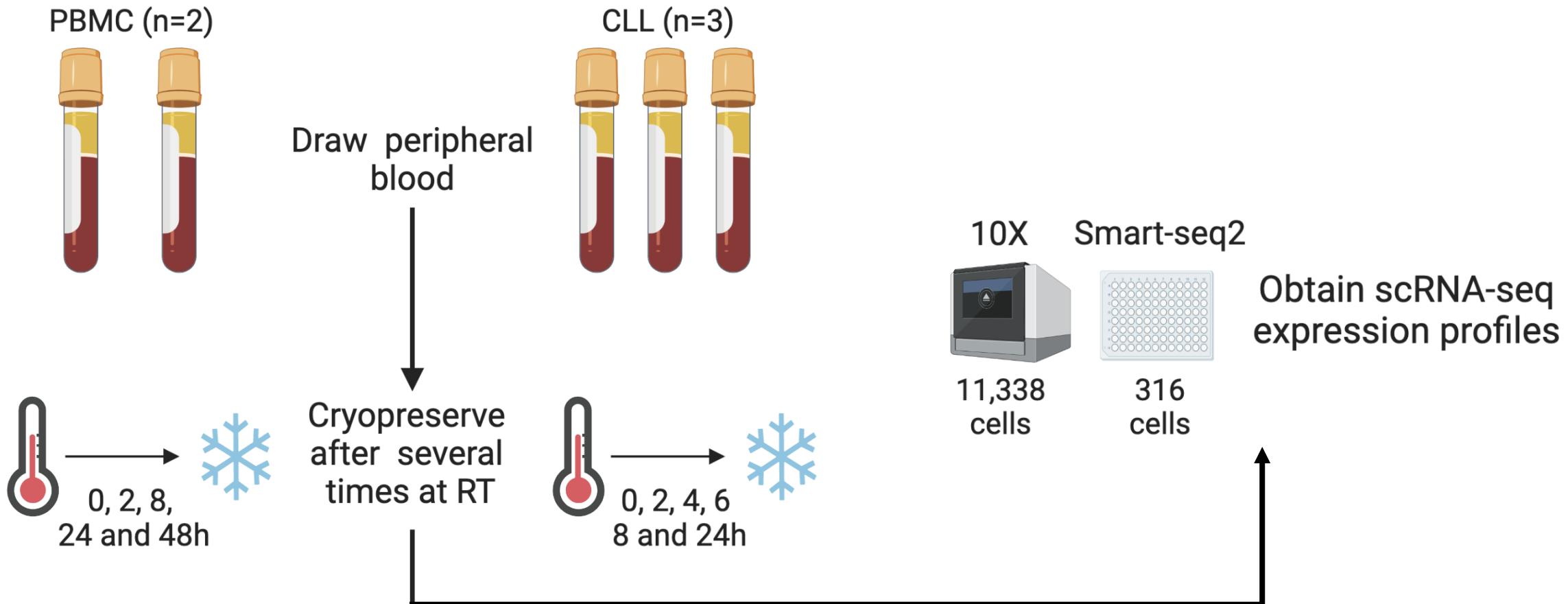
Efecto de lote (Batch effect)



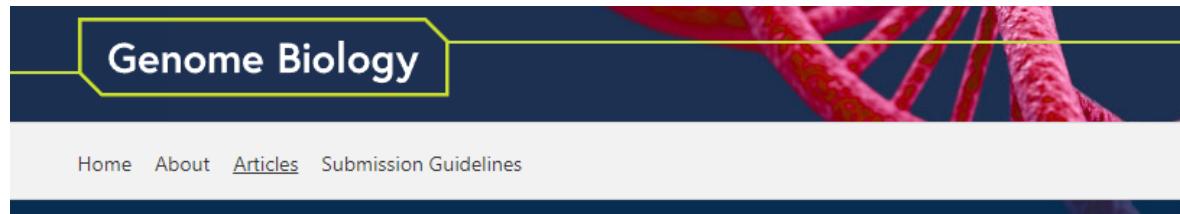
Metadata				
sample	replicate	condition	batch	
sample1	1	control	1	
sample2	2	control	1	
sample3	3	control	2	
sample4	4	control	2	
sample5	1	treatment1	1	
sample6	2	treatment1	1	
sample7	3	treatment1	2	
sample8	4	treatment1	2	
sample9	1	treatment2	1	
sample10	2	treatment2	1	
sample11	3	treatment2	2	
sample12	4	treatment2	2	

Image credit: Hicks SC, et al., bioRxiv (2015)

Batch effect : Sampling time



Sampling time-dependent artifacts in single-cell genomics studies



Abstract

Robust protocols and automation now enable large-scale single-cell RNA and ATAC sequencing experiments and their application on biobank and clinical cohorts. However, technical biases introduced during sample acquisition can hinder solid, reproducible results, and a systematic benchmarking is required before entering large-scale data production. Here, we report the existence and extent of gene expression and chromatin accessibility artifacts introduced during sampling and identify experimental and computational solutions for their prevention.



Results and Discussion

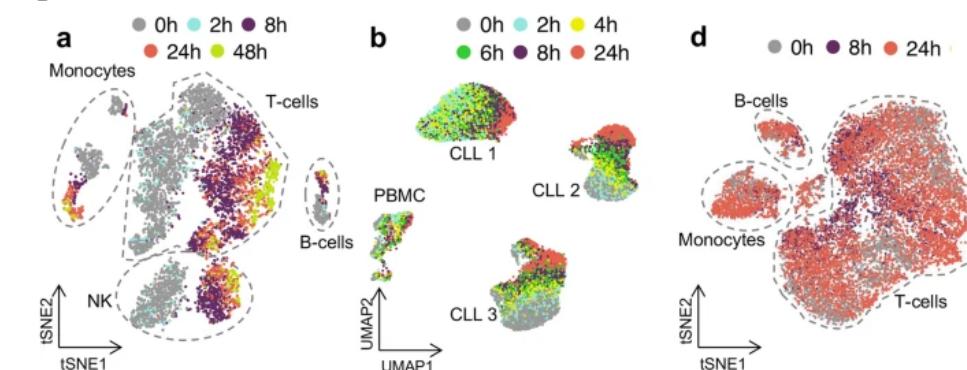
<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02032-0>

Sampling time-dependent artifacts in single-cell genomics studies

Results and discussion

We generated transcriptome and epigenome profiles for 71,064 and 76,146 high-quality cells, respectively. To evaluate the effect of sampling time on single-cell gene expression profiles, we initially obtained fresh PBMC from 2 healthy donors and 3 CLL patients. To simulate local processing, we stored cells prior to cryopreservation at room temperature (RT) for various time intervals up to 8 h. Additionally, we stored cells for 24 h and 48 h, common sampling times for central sample processing. Following scRNA-seq, we detected a marked effect of the sampling time on single-cell transcriptome profiles, initiating after 2 h and increasing in a time-dependent manner (Fig. 1a). This effect was reproducible across all blood cell subtypes from healthy donors and neoplastic cells from CLL patient samples (Fig. 1a,b, Additional file 1: Fig. S1a) and across scRNA-seq technologies (Additional file 1: Fig. S1b). Sampling time correlated with the first principal component (PC1) for all cell subtypes (Fig. 1c), explaining between 15.3% (T cells) and 8.4% (B cells) of the variance contained in the first 50 PC (Additional file 1: Fig. S2a). Moreover, sampling time followed cell type and patient variability as the greatest driver of variance in the PBMC and CLL datasets, respectively (Additional file 1: Fig. S2b-d), and surpassing batch and donor for different cell types (Additional file 1: Fig. S2d-f). Although gene expression profiles varied notably, viable cells did not show signs of reduced RNA integrity across the time points (Additional file 1: Fig. S3).

Fig. 1



Sections

Figures

References

[Abstract](#)

[Background](#)

[Results and discussion](#)

[Conclusions](#)

[Methods](#)

[Availability of data and materials](#)

[References](#)

[Funding](#)

[Author information](#)

[Ethics declarations](#)

[Additional information](#)

[Supplementary information](#)

[Rights and permissions](#)

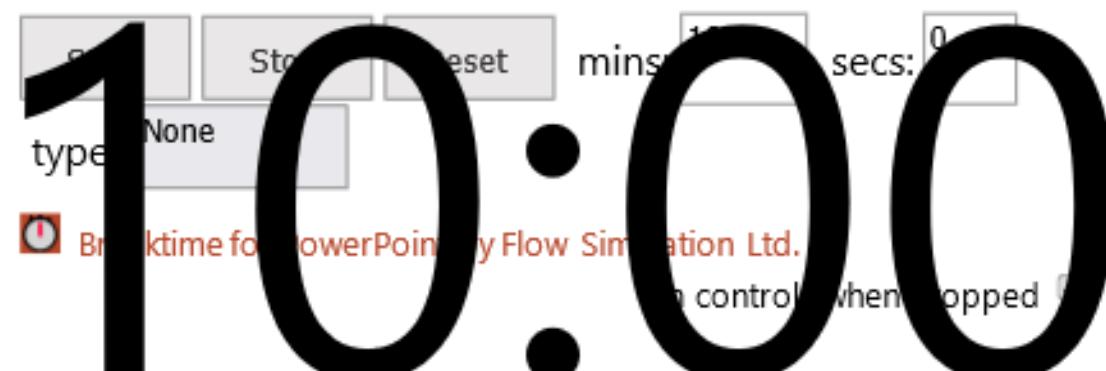
[About this article](#)



Results and Discussion

Sampling time-dependent artifacts in single-cell genomics studies

1. ¿Cuántas células de alta calidad se generaron en total para los perfiles transcriptómicos?
2. ¿Qué se detectó en los perfiles de expresión a nivel en relación con el tiempo de muestreo?
3. ¿Qué efecto se observó en los perfiles de cromatina?
4. ¿De forma general, la expresión de los genes aumenta o disminuye con el tiempo?
5. ¿Qué genes se encontraron entre los más sobreexpresados en ambos conjuntos de datos?
6. ¿Qué alternativas experimentales se propusieron para reducir los efectos del muestreo?



Sampling time-dependent artifacts in single-cell genomics studies

The screenshot shows the 'Genome Biology' journal website. At the top, there's a navigation bar with 'Genome Biology' in a yellow box, followed by 'Home', 'About', 'Articles' (underlined), and 'Submission Guidelines'. Below the navigation is a header with 'Short Report | Open Access | Published: 11 May 2020'. The main title of the article is 'Sampling time-dependent artifacts in single-cell genomics studies'. Below the title, the authors are listed: Ramon Massoni-Badosa, Giovanni Iacono, Catia Moutinho, Marta Kulis, Núria Palau, Domenica Marchese, Javier Rodríguez-Ubreva, Esteban Ballestar, Gustavo Rodriguez-Esteban, Sara Marsal, Marta Aymerich, Dolors Colomer, Elias Campo, Antonio Julià, José Ignacio Martín-Subero & Holger Heyn. There are links for 'Genome Biology' volume 21, article 112 (2020), 'Cite this article', '9930' accesses, '24' citations, '69' Altmetric, and 'Metrics'.

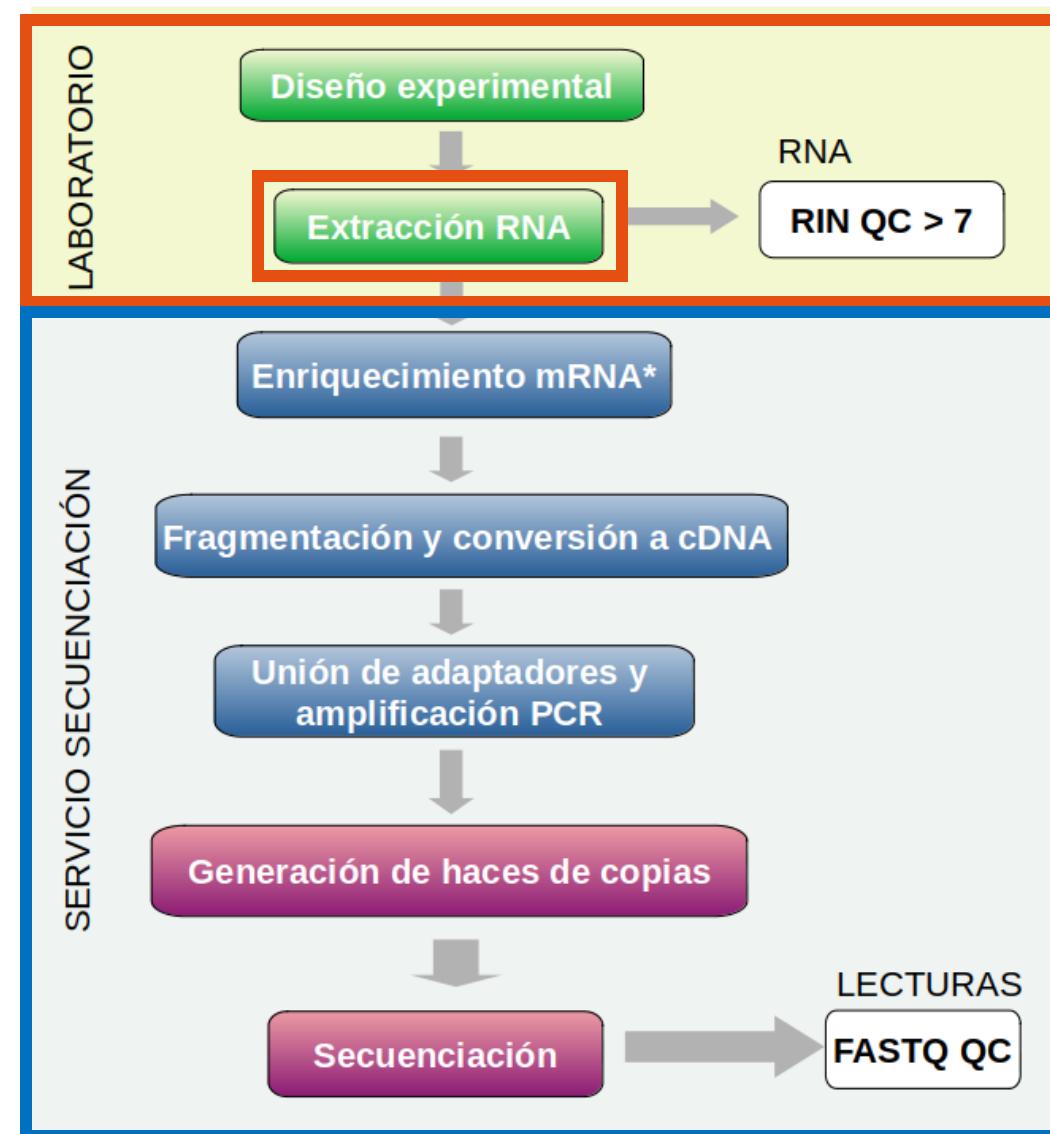
Abstract

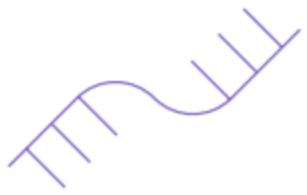
Robust protocols and automation now enable large-scale single-cell RNA and ATAC sequencing experiments and their application on biobank and clinical cohorts. However, technical biases introduced during sample acquisition can hinder solid, reproducible results, and a systematic benchmarking is required before entering large-scale data production. Here, we report the existence and extent of gene expression and chromatin accessibility artifacts introduced during sampling and identify experimental and computational solutions for their prevention.

Conclusiones

1. El tiempo de almacenamiento es un artefacto que puede aparecer en los datos de scRNA-seq caracterizados por una respuesta de estrés de ciertos genes claves.
2. El almacenamiento a 4°C minimiza los artefactos de conservación antes de la criopreservación.
3. No importa cuán impresionante sea una tecnología, en ciencia siempre necesitamos volver a lo básico.

Flujo de trabajo general de RNA-seq





El primer paso para caracterizar el transcriptoma consiste en **aislar** y **purificar** el ARN celular. La calidad y cantidad de material tiene un impacto significativo en la calidad de los datos.

¿Cómo vamos a aislar el RNA?

¿Cómo vamos a medir la cantidad de ARN y la calidad del mismo?

El aislamiento del ARN **intacto** es esencial para las técnicas de biología molecular empleadas para el análisis de la expresión génica

Molécula especialmente lábil

- Rápidamente degradada por las **RNAasas** (ribonucleasas)
 - Células/Tejidos
 - Fuentes externas a la muestra -> laboratorio
- **Condiciones ambientales**
 - Altas temperaturas
 - pH extremos
 - Presencia de otras sustancias químicas reactivas



Extracción del RNA



- Mantener las muestras en las condiciones adecuadas de temperatura y pH -> muestras de ARN a **4°C** durante el procesamiento de las mismas.
- Trabajar siempre en un **entorno libre de ARNasas** -> empleo de guantes, inhibidores de ARNasas,etc.

- **Criogenia:** Congelación utilizando nitrógeno líquido y posterior almacenamiento a -80°C.
- **Reactivos químicos:** Reactivos estabilizadores disponibles comercialmente (p. ej., RNAlater®, RNAProtect® y tubos de estabilización de ARN) inhiben las ARNasas y previenen el daño de la hidrólisis y oxidación del ARN.
- **Liofilización:** Ciertas estrategias de estabilización de ARN deshidratan el ARN en presencia de una matriz de estabilización. Esto evita la desnaturación por calor y la digestión por parte de las RNasas, además de permitir que las muestras se almacenen y envíen a temperatura ambiente.

EXTRACCIÓN TOTAL DE ARN

EXTRACCIÓN ORGÁNICA

Utilización de solventes orgánicos

Versátil

Escalable

Económico

PURIFICACIÓN CON COLUMNAS

Evita el uso de solventes orgánicos

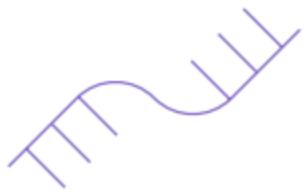
Rápido

Mayor coste económico (Kits comerciales)

Métodos Magnéticos, Métodos basados en membranas ...



Extracción del RNA



El primer paso para caracterizar el transcriptoma consiste en **aislar** y **purificar** el ARN celular. La calidad y cantidad de material tiene un impacto significativo en la calidad de los datos.

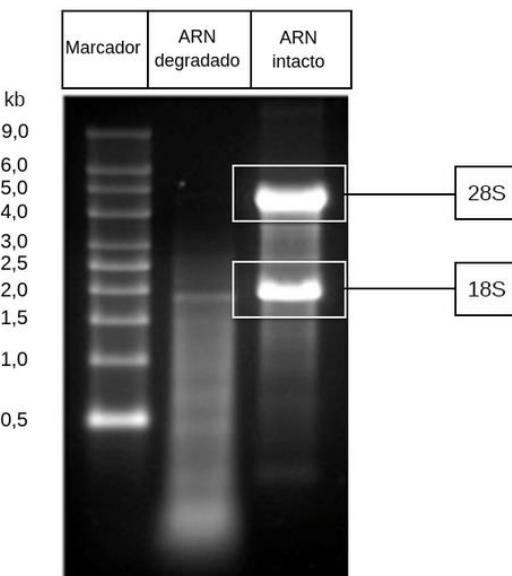
¿Cómo vamos a aislar el RNA?

¿Cómo vamos a medir la cantidad de ARN y la calidad del mismo?



Integridad

El fragmento del ARN se debe mantener **intacto**, de forma que no se ha degradado por la acción de las RNAsas o el calor.



RNA Integrity Number (RIN)

- Análisis de las proporciones de las bandas ribosómicas 28S a 18S.
- Escala del 1-10 -> **RIN > 7**
- **No** sirve en todas las especies

Pureza

Determina que no haya ningún tipo de **contaminante**, como pueden ser proteínas o algún reactivo resultado de la extracción del ARN.

Se determina midiendo relaciones de absorbancia usando un **espectrofotómetro**:

- 260/280nm (ARN/Proteínas) ~ 2.0
- 260/230nm (ARN/fenol) (~2.0 - 2.2).



NanoDrop™

Cantidad de ARN



Métodos basados en fluorescencia

- Qubit®
- RiboGreen™ / PicoGreen™

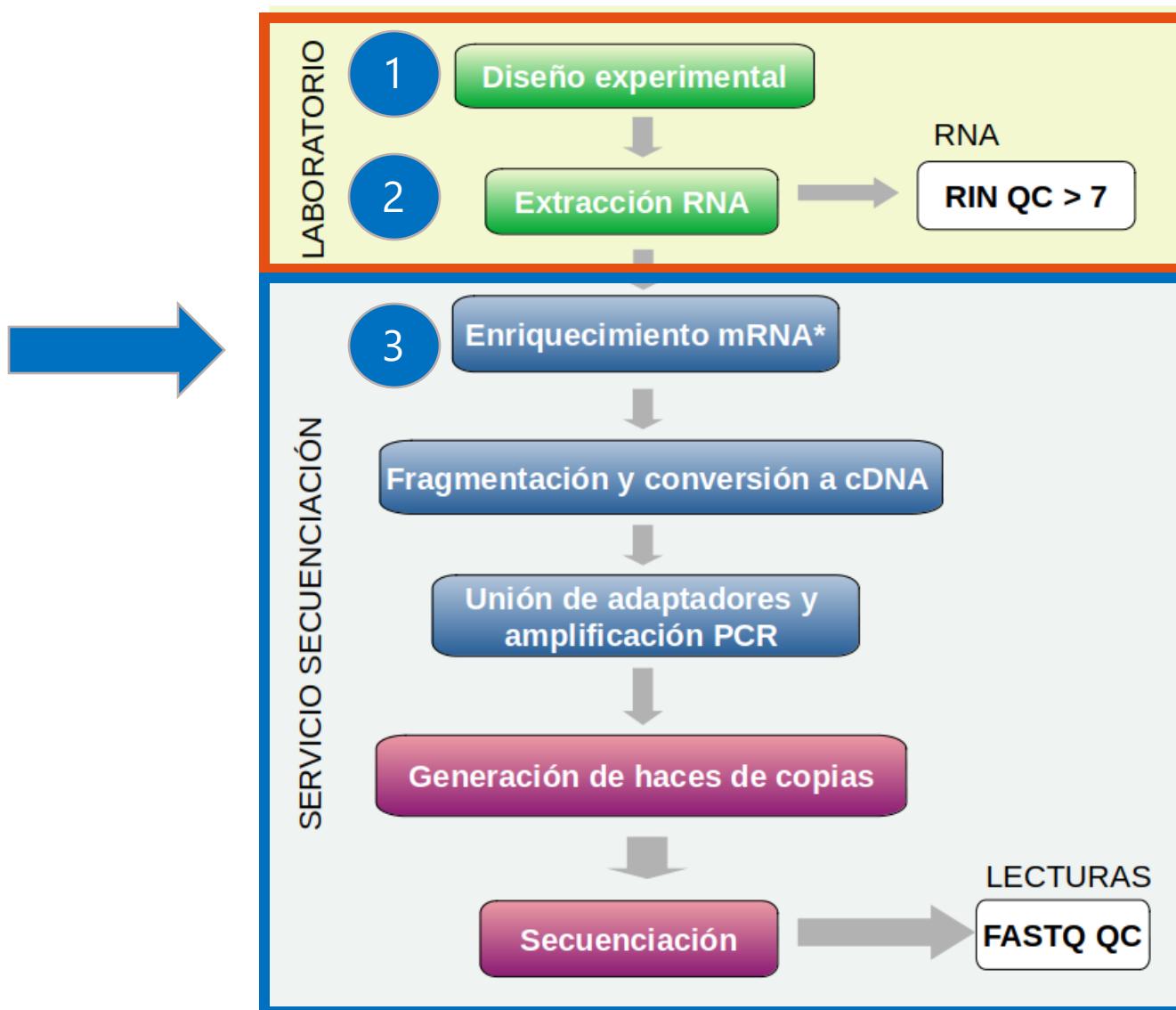
Métodos basados en espectroscopia

- NanoDrop™

Necesitaremos alcanzar cantidades de $\sim > 500$ ng de RNA total purificados por microlitros muestra con una integridad mínima de 7 RIN.

Idealmente, debe mantenerse el mismo rango en todas las muestras para que las bibliotecas finales sean más homogéneas.

Flujo de trabajo general de RNA-seq



Preparación de la librería

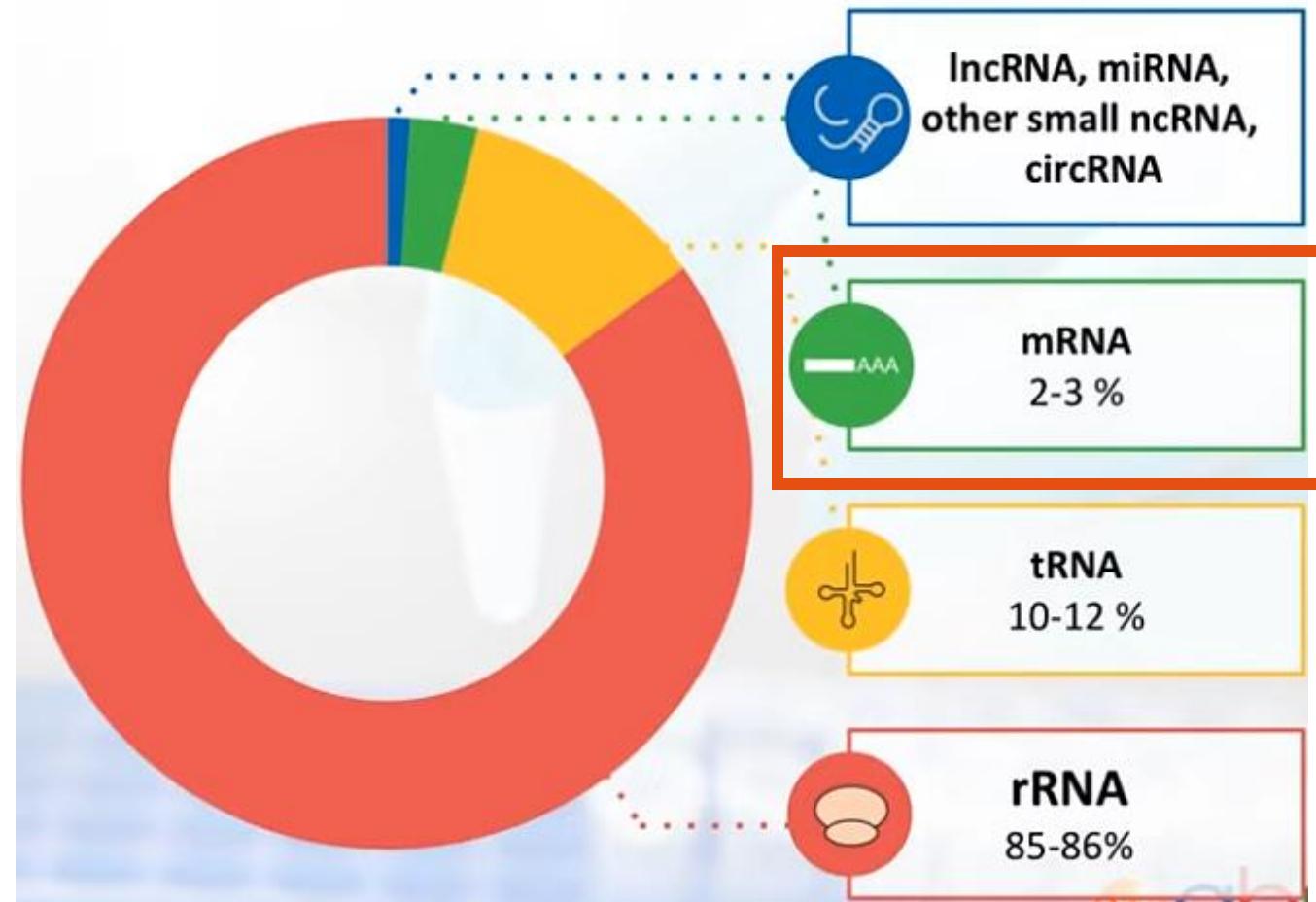


La preparación de la librería implica generar una colección de fragmentos de ARN que sean compatibles para secuenciación.

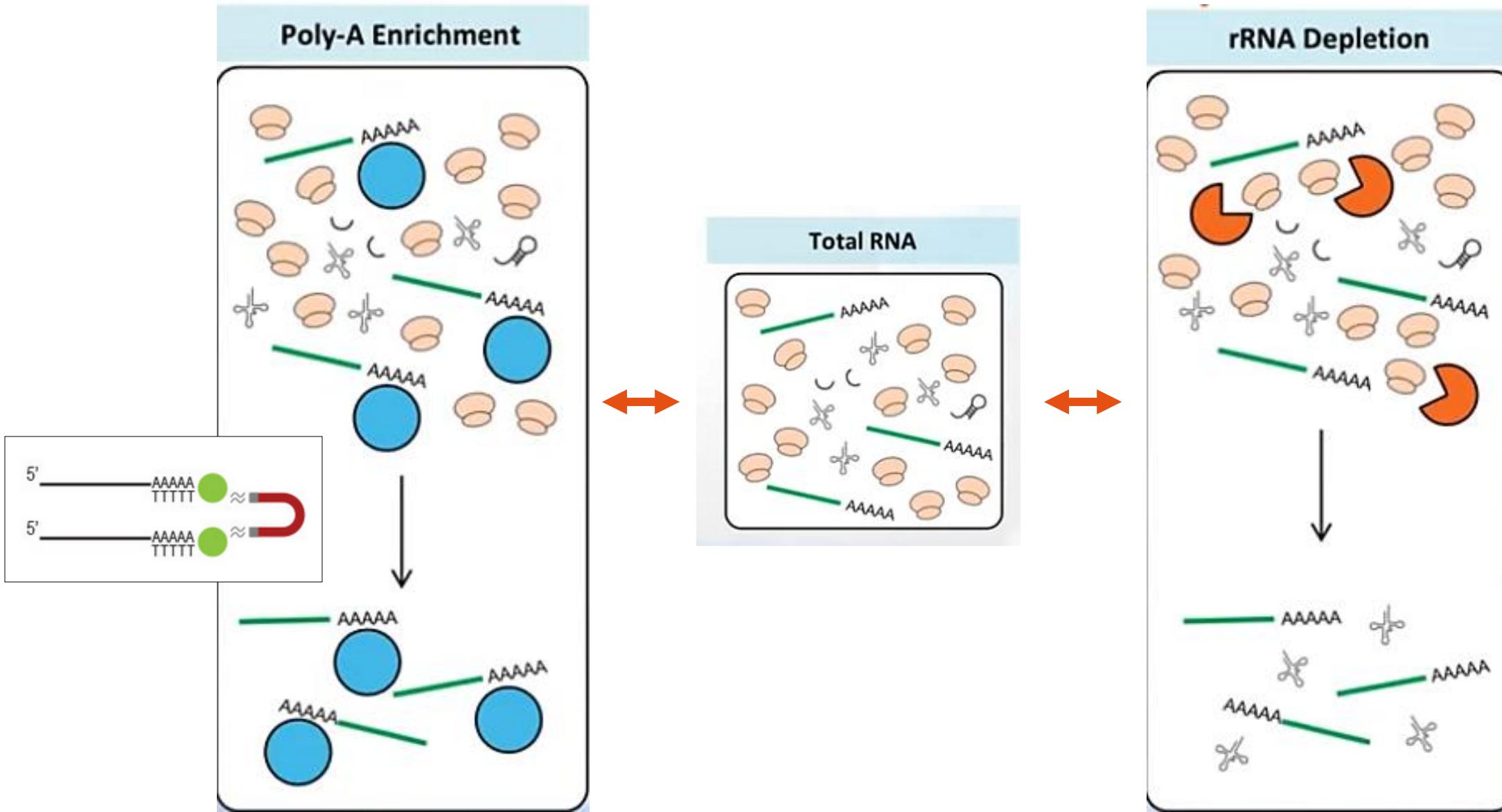
El proceso implica el enriquecimiento del ARN diana , la fragmentación, transcripción inversa (ADNc), adición de adaptadores de secuenciación y amplificación

Enriquecimiento del ARN diana
Enriquecimiento Poly(A) o eliminación del ARNr

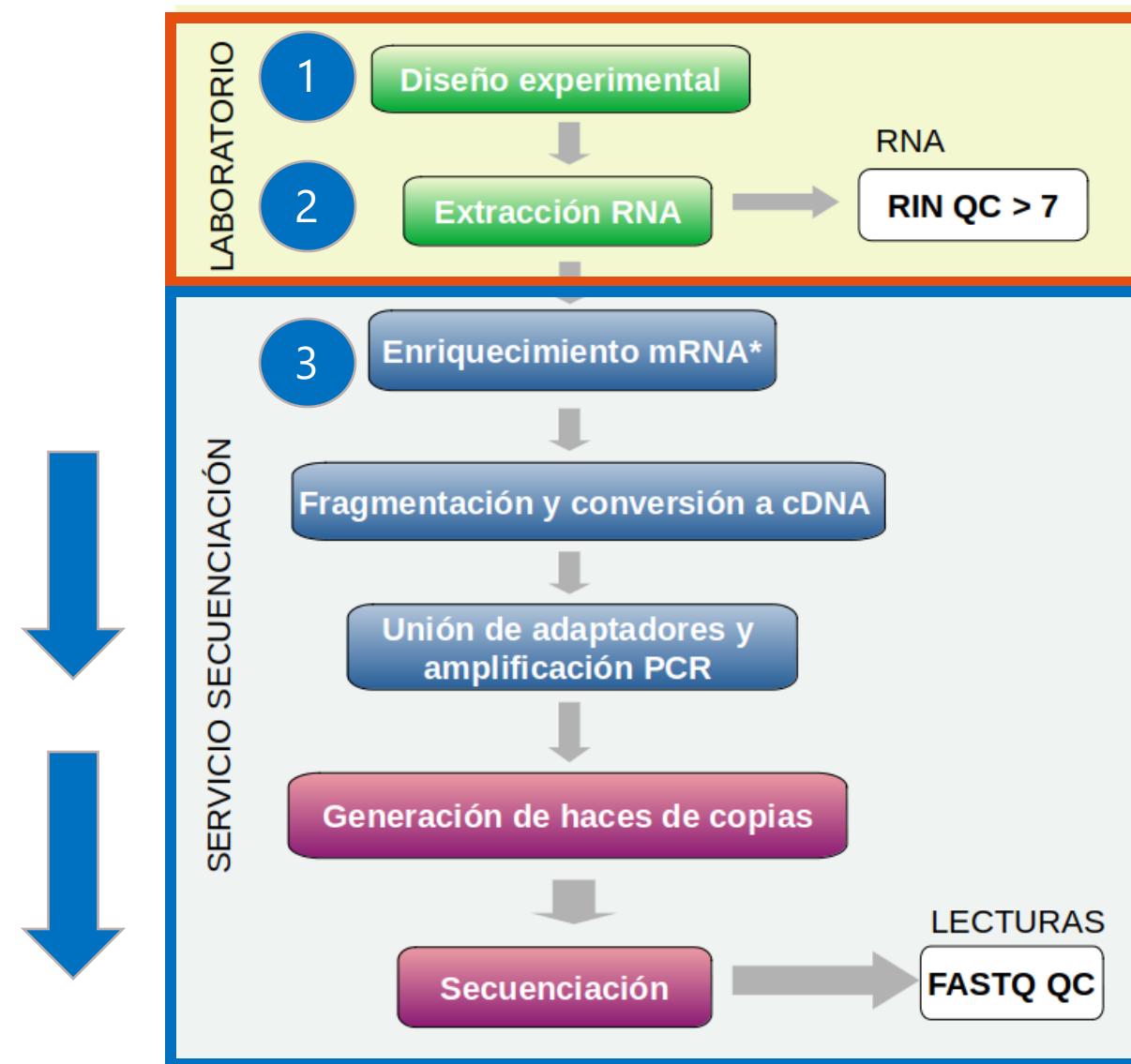
La mayor parte de ARN en la célula **NO** es ARNm



Tipo de librería



Flujo de trabajo general de RNA-seq





viu

Universidad
Internacional
de Valencia

universidadviu.com

De:
 Planeta Formación y Universidades