VNIVERSITAT DE VALÈNCIA

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Unidad de Bioinformática y Bioestadística

Métodos de Metaanálisis de Estudios Ómicos en Biomedicina

Trabajo Final de Máster en Bioestadística

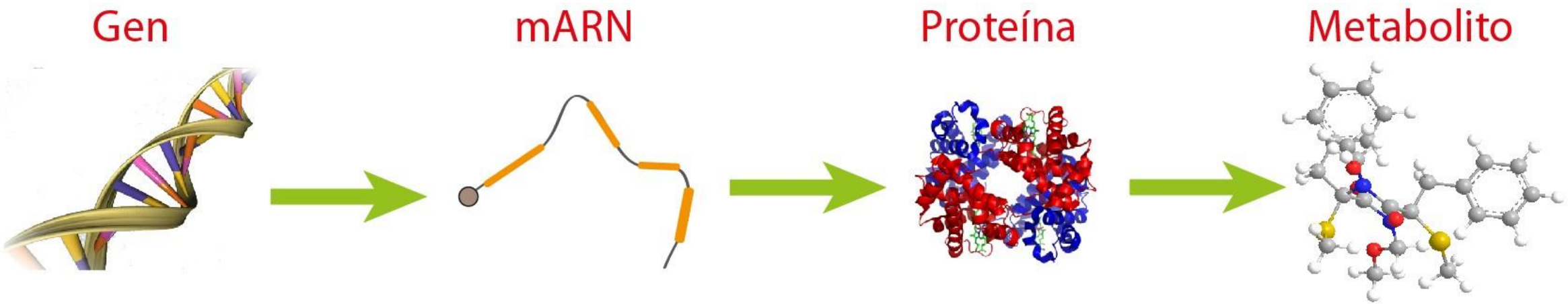
Jesús Gutiérrez Botella

Tutor: Francisco García García
Tutor académico: Antonio López Quílez

23 de septiembre, 2020

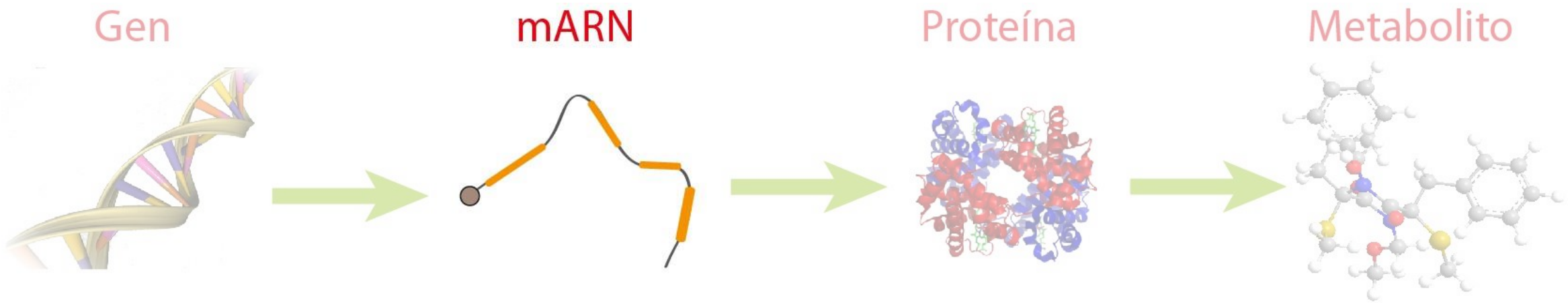
1. Introducción

Ciencias ómicas en estudios Biomédicos



Genómica **Transcriptómica** **Proteómica** **Metabolómica**

Ciencias ómicas en estudios Biomédicos



- **Caracterización funcional** de organismos: expresión de los genes de un individuo en una condición y un tiempo determinados.
- El objetivo suele ser llevar a cabo un **análisis de expresión diferencial**.

Tecnologías de alto rendimiento

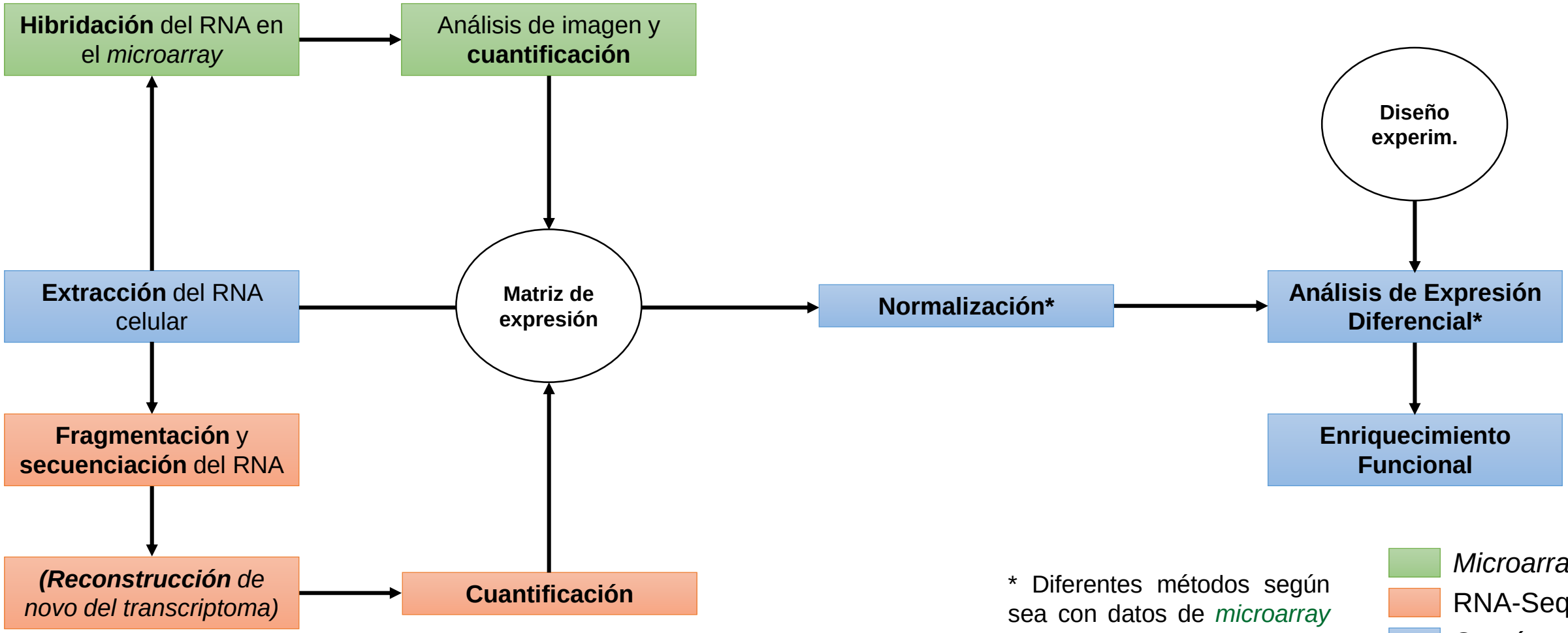
Microarrays

- Basados en **hibridación** del material genético.

RNA-Seq

- Basado en **secuenciación masiva** del material genético.

Análisis de un estudio de Transcriptómica



Análisis de Expresión Diferencial y Enriquecimiento Funcional

Análisis de Expresión Diferencial

Genes **sobreexpresados**

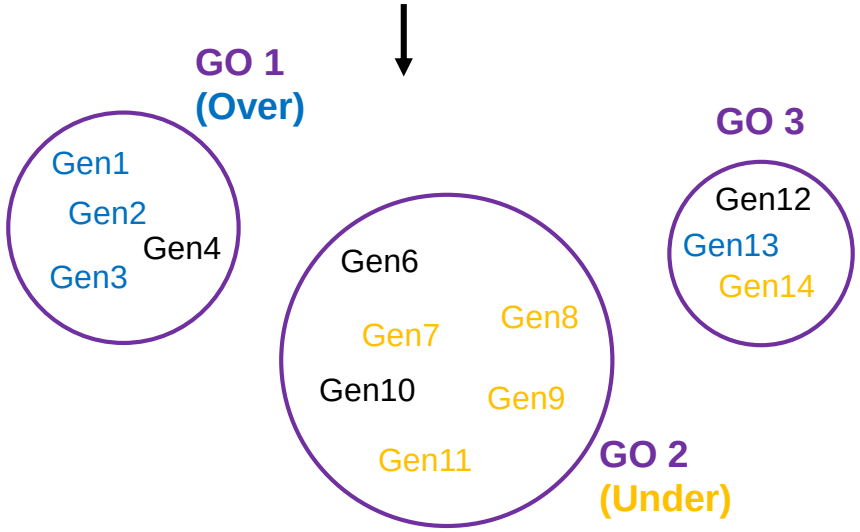
Genes **infraexpresados**

Comparación de los niveles de **expresión génica** entre diferentes condiciones.
Se suele comparar casos vs. controles.

¿**Qué genes** están diferencialmente expresados entre las dos condiciones?

Enriquecimiento funcional

Anotación Funcional (GO)



¿**Qué funciones moleculares** realizan los genes dif. expresados entre condiciones?

- **Tamaño del efecto.** *Odds-ratio, fold-change*.
- **Significación estadística.** *P*-valor.

El Metaanálisis

Los pasos del **metaanálisis**

1. Formulación de objetivos.
2. Recopilación y selección de los estudios.
3. Combinación de los estudios.
   - Combinación de efectos.
   - Combinación de *p*-valores o rangos.
4. Análisis de heterogeneidad y sensibilidad, en caso de combinación de efectos.
5. Cálculo del tamaño del efecto y/o significación estadística combinada.
6. Obtención de conclusiones.

En **transcriptómica…**

- A nivel de gen.
- A nivel de función molecular.

2. Objetivos

Objetivos

Revisar y evaluar algunos métodos de metaanálisis de datos ómicos para conocer las estrategias más adecuadas para integrar la información de estudios biomédicos.

Objetivos específicos:

- Revisar y comparar los métodos de **metaanálisis a nivel de gen**.
- Revisar los métodos de **enriquecimiento funcional**.
- Revisar y comparar los métodos de **metaanálisis a nivel de función**.
- Evaluar **conjuntamente todos los métodos** descritos.
- Aplicar estos métodos a dos **sets de datos** de estudios de transcriptómica: un set de datos de tumores y otro de enfermedades dermatológicas.

3. Metodología

Sets de datos

Disponemos de dos colecciones de estudios:

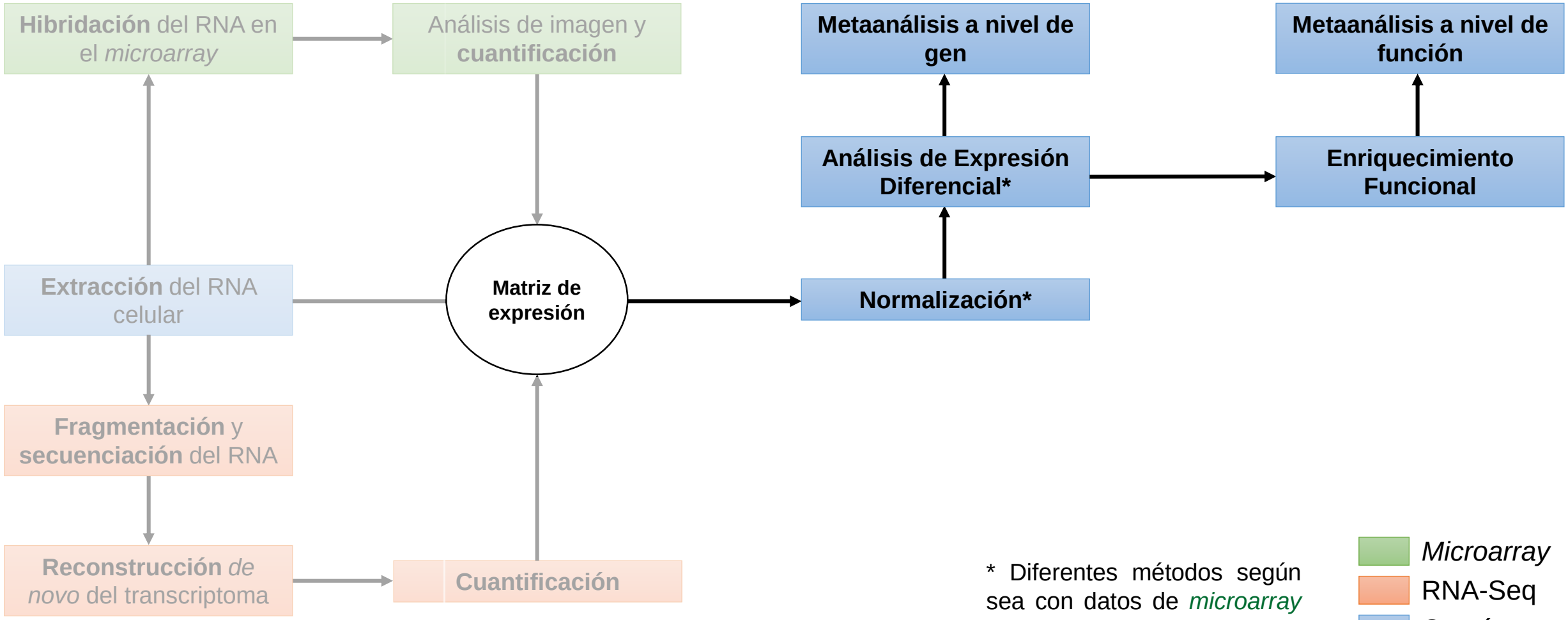
Tumores (TCGA)

- **RNA-Seq.**
- Selección de 17 estudios de la base de datos de The Cancer Genome Atlas (TCGA), relativos a 17 tipos de tumores.
- Punto de partida: matriz de expresión y diseño experimental (se requiere pre-procesado).

Enfermedades dermatológicas

- ***Microarray*.**
- Selección de 41 estudios de psoriasis y dermatitis, descargados de la base de datos Gene Expression Omnibus (GEO).
- Punto de partida: resultados del análisis de expresión diferencial (no requiere pre-procesado).

Flujo de trabajo



Modelo para datos de conteo de RNA-Seq

Matriz de expresión

	M_1	M_2	…	M_m
Gen 1	5	1	…	7
Gen 2	9	8	…	4
Gen 3	0	17	…	3
…	…	…	…	…
Gen *i*	1	4	…	16

Para **datos de conteo** de RNA-Seq el modelo es:

$$K_{ij} \sim NB\left(\mu_{ij}, \alpha_i\right) \qquad \mu_{ij} = s_j q_{ij}$$

Es necesario **normalizar** la matriz de expresión debido a los sesgos que pueden contener los datos de RNA-Seq:

- Longitud de los genes.
- Genes con nivel de expresión extremo.
- …

Se introducen **factores de normalización** s_j calculados con el método de la mediana de las ratios:

$$s_j = \underset{i:\, K_i^R \neq 0}{\text{mediana}} \frac{K_{ij}}{K_i^R} \qquad K_i^R = \left(\prod_{j=1}^{m} K_{ij}\right)^{\frac{1}{m}}$$

Análisis de Expresión Diferencial de datos de RNA-Seq

Condición 1 (referencia)

Condición 2 (contraste)

$$\log FC = \log_2 \frac{E_{contraste}}{E_{referencia}}$$

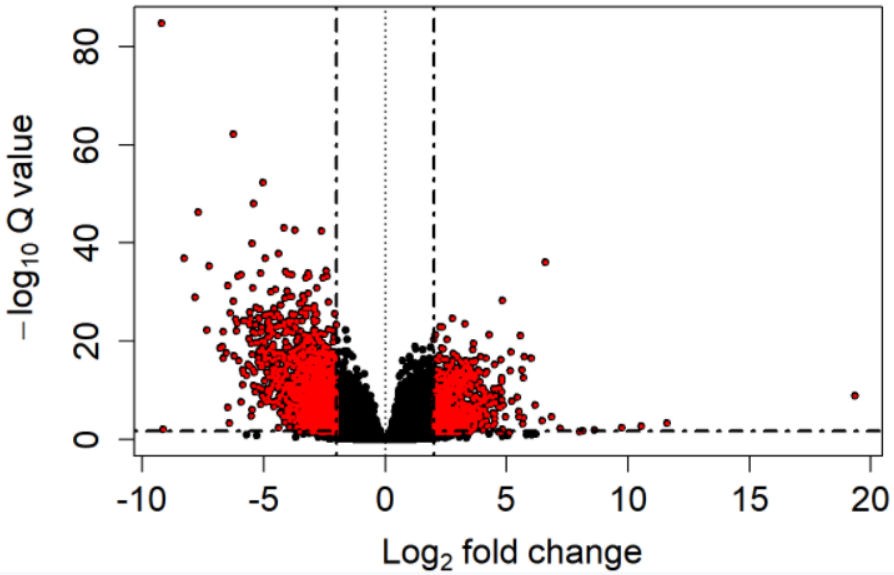
p-valor y *p*-valor corregido (FDR, Benjamini & Hochberg)

Recordemos el modelo anterior para el gen *i* en la muestra *j*:

$$K_{ij} \sim NB\left(\mu_{ij}, \alpha_i\right) \qquad \mu_{ij} = s_j q_{ij}$$

$$\log_2 q_{ij} = \sum_r x_{jr} \beta_{ir} \longrightarrow \text{LogFC}$$

Matriz de diseño



Gen sobreexpresado

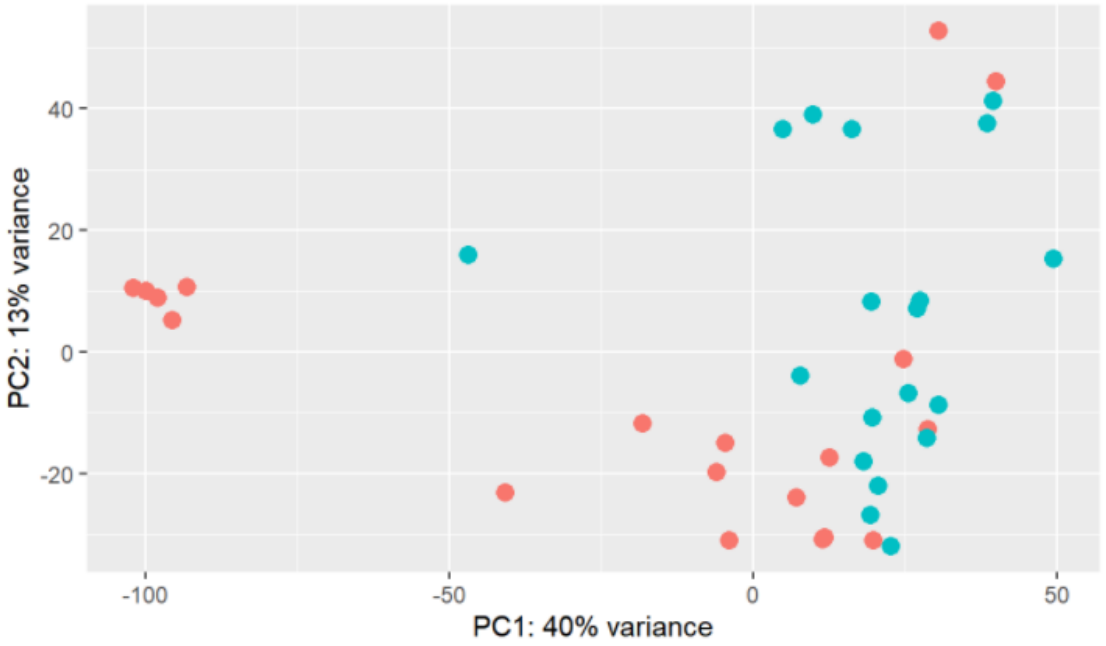
- logFC > 1
- *p*-valor corregido < 0.05

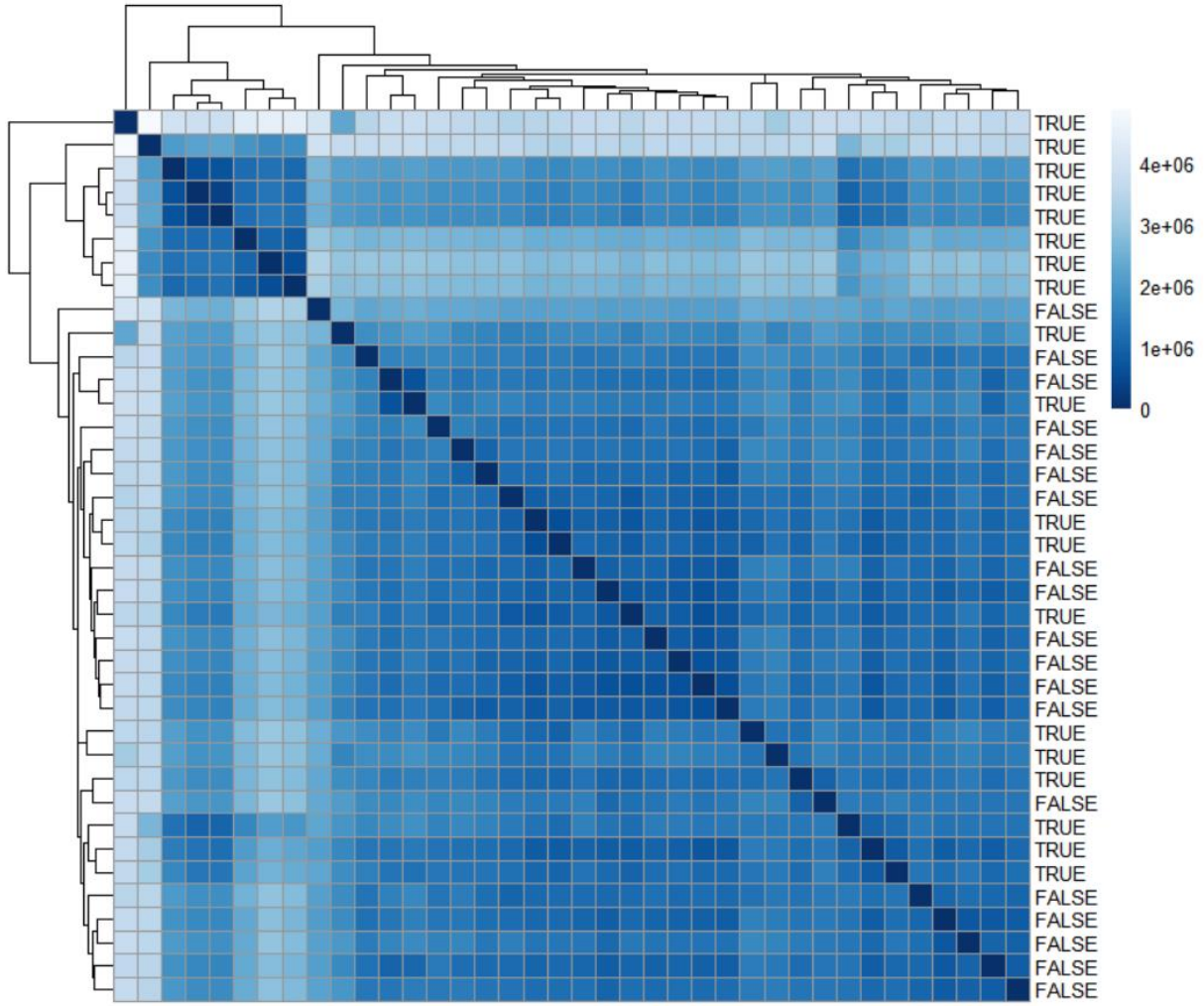
Gen infraexpresado

- logFC < -1
- *p*-valor corregido < 0.05

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with **DESeq2**. *Genome biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Análisis de Expresión Diferencial de datos de RNA-Seq





Análisis de Enriquecimiento Funcional. Método de Grupos de Genes

Hay que anotar los genes con **términos GO** (BioMart).

1. Se obtiene una **lista ordenada** *x* de todos los genes según su estadístico de contraste de la expresión diferencial.

P	**Gen**	**E.Contraste**
1	Gen1	EC_1
2	Gen2	EC_2
3	Gen3	EC_3
…		
k	Genk	EC_k

$$EC_1 > EC_2 > \ldots > EC_k$$

2. Se definen los **sets de genes** *c*. Cada set representa todos los genes anotados con el mismo término GO (con la misma función).

3. Para cada *c* se aplica un modelo logístico:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$$

LOR

Test de Wald

H_0. $\beta = 0$ $\quad$ H_1. $\beta \neq 0$

Bajo H_0, el estadístico de contraste *W* sigue una distribución χ^2_1.

$$W = \left(\frac{\hat{\beta}}{s_{\hat{\beta}}}\right)^2$$

GO enriquecido

- LOR > 0
- *p*-valor corregido < 0.05

GO infrarrepresentado

- LOR < 0
- *p*-valor corregido < 0.05

Montaner D, Dopazo J (2010) Multidimensional Gene Set Analysis (**MDGSA**) of Genomic Data. PLoS ONE 5(4): e10348. https://doi.org/10.1371/journal.pone.0010348

Métodos de Metaanálisis

Aplicamos tres métodos de metaanálisis tanto a los genes diferencialmente expresados como a las funciones moleculares enriquecidas:

Combinación de *p*-valores

- Combinación simple de *p*-valores sin tener en cuenta el tamaño del efecto.
- La combinación puede ser sin ponderar o ponderada.

metaRNAseq

Rau, A., Marot, G. & Jaffrézic, F. Differential meta-analysis of RNA-seq data from multiple studies. *BMC Bioinformatics* **15,** 91 (2014). https://doi.org/10.1186/1471-2105-15-91

Combinación de rangos

- Test no paramétrico.
- A partir de una lista ordenada de genes/funciones y se estima un *p*-valor combinado experimental.

RankProd

Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*. 2006;22(22):2825-2827. doi:10.1093/bioinformatics/btl476

Combinación del tamaño del efecto

- Modelos estadísticos para combinar los tamaños del efecto de cada estudio en una métrica combinada.
- Permite modelizar heterogeneidad y estimar parámetros de un modelo.

metafor

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. Journal of Statistical Software, 36(3), 1 - 48. doi: http://dx.doi.org/10.18637/jss.v036.i03

Metaanálisis (I). Combinación de *p*-valores

Para cada gen o término GO *g*, se definen los estadísticos…

Método de la normal inversa (ponderado)

$$N_g = \sum_{s=1}^{S} w_s \phi^{-1}\left(1 - p_{gs}\right)$$

El vector de pesos w_s da más importancia a aquellos estudios con más muestras (Marot & Meyer, 2009):

$$w_s = \sqrt{\frac{n(s)}{\sum_{i=1}^{S} n(i)}}$$

Bajo H_0, N_g sigue una distribución normal estándar.

Método de Fisher (no ponderado)

$$F_g = -2\sum_{s=1}^{S} \ln\left(p_{gs}\right)$$

Bajo H_0, F_g sigue una distribución χ^2 con $2S$ grados de libertad.

Consideraciones

- *p*-valores uniformemente distribuidos de todos los genes y GOs: filtrado.
- Revisión de patrones conflictivos de tamaño del efecto en genes y términos GO.
- Los *p*-valores finales deben ser corregidos por comparaciones múltiples.

Metaanálisis (II). Combinación de rangos

Para cada gen o término GO *g*…

1. Se obtienen *s* **listas ordenadas** (rangos), de 1 a *g* según su LogFC en el caso de genes o según su LOR en el caso de términos GO.

Rank	***G***	**Efecto**
1	g_1	E_1
2	g_2	E_2
3	g_3	E_3
…		
k	g_k	E_k

El procedimiento se repite **dos veces**:

· $E_1 > E_2 > \ldots > E_k$ Para detectar genes sobreexpresados o GOs enriquecidos.

· $E_1 < E_2 < \ldots < E_k$ Para detectar genes infraexpresados o GOs infrarrepresentados.

2. Para cada *g* se calcula el producto RP_g:

$$RP_g = \prod_{s=1}^{S} \left(r_{sg}\right)^{\frac{1}{S}}$$

3. Análisis de significación estadística: cálculo de *p*-valores utilizando una estrategia basada en permutaciones: *¿cómo de probable es observar este RP o uno mayor?*

   a) Generar *p* permutaciones de cada lista.
   b) Calcular RP de los *g* genes o términos GO en cada permutación.
   c) Contar cuántas veces los RP en cada una de las *p* permutaciones son iguales o más pequeños que el RP real y estimar un *p*-valor.
   d) Corrección de *p*-valores por FDR.

Metaanálisis (III). Combinación del tamaño del efecto

Para cada gen o término GO se ajusta el siguiente modelo:

$$y_i = \theta_i + \varepsilon_i,\ \varepsilon_i \sim N\left(0, s_i^2\right)$$

$$\theta_i = \mu + \delta_i,\ \delta_i \sim N\left(0, \tau^2\right)$$

- **μ** Tamaño medio del efecto de la expresión diferencial / enriquecimiento funcional.
- **s_i** Error de muestreo en el estudio *i.*
- **τ^2** Variación entre estudios.

Modelo de Efectos Fijos (FEM) y de Efectos Aleatorios (REM)

En un **REM** se modeliza la variabilidad entre estudios. Se usa cuando los estudios son muy heterogéneos.

Un **FEM** considera que $\tau^2 = 0$ y se utiliza cuando el conjunto de estudios es homogéneo.

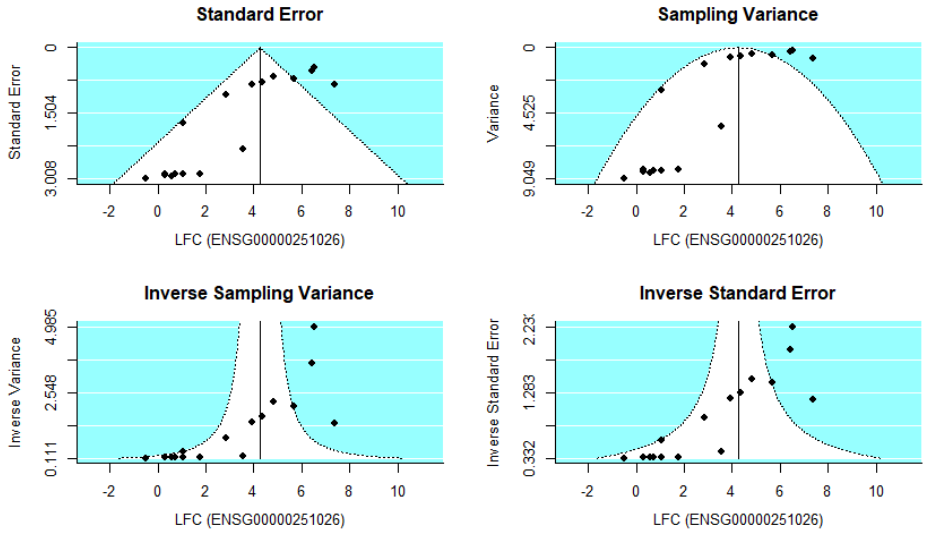
Para comprobar si los estudios son o no heterogéneos:

1. **Test de Cochran**

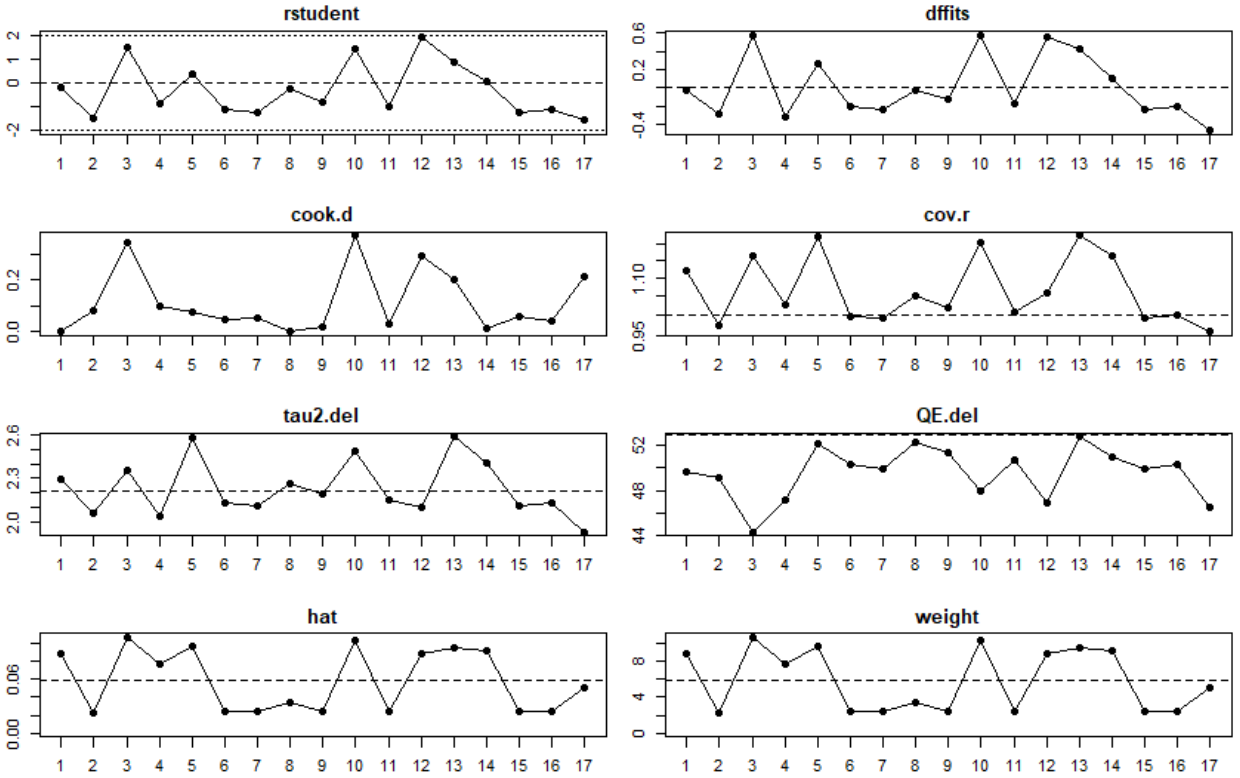
$$Q = \sum_{i=1}^{k} w_i \left(y_i - \hat{\mu}\right)$$

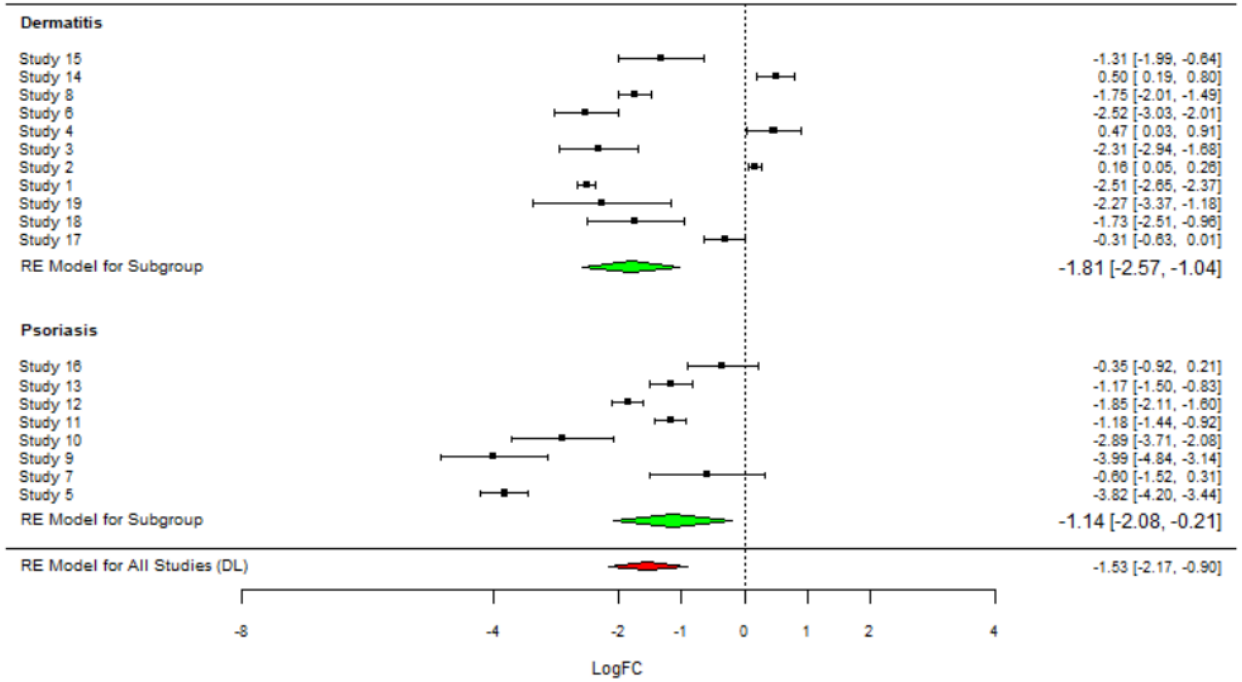
Bajo la **H_0** de no heterogeneidad ($\tau^2 = 0$), el estadístico *Q* sigue una distribución χ^2 con *k* – 1 g.l.

2. **Gráficos de embudo**



Metaanálisis (III). Combinación del tamaño del efecto

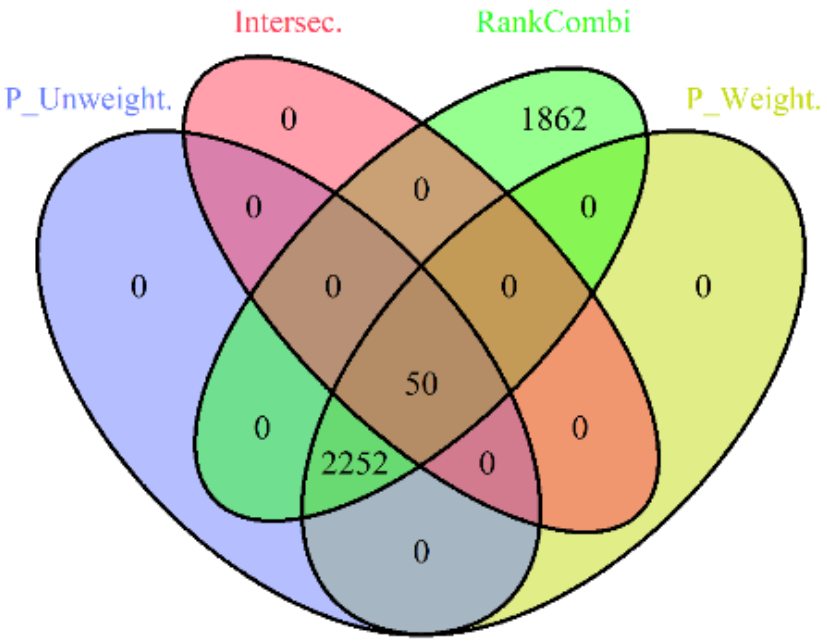


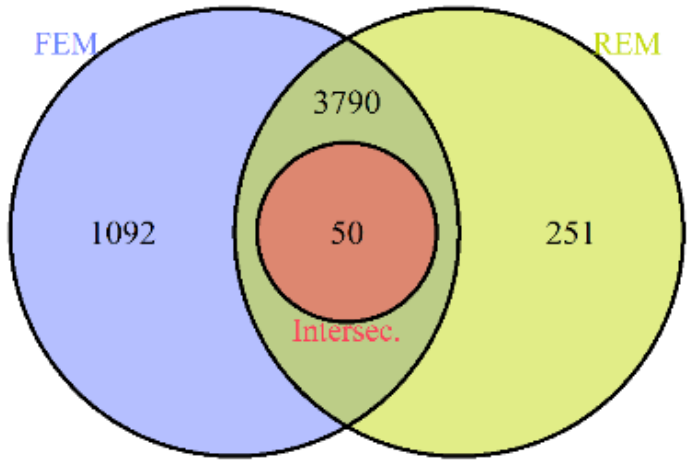


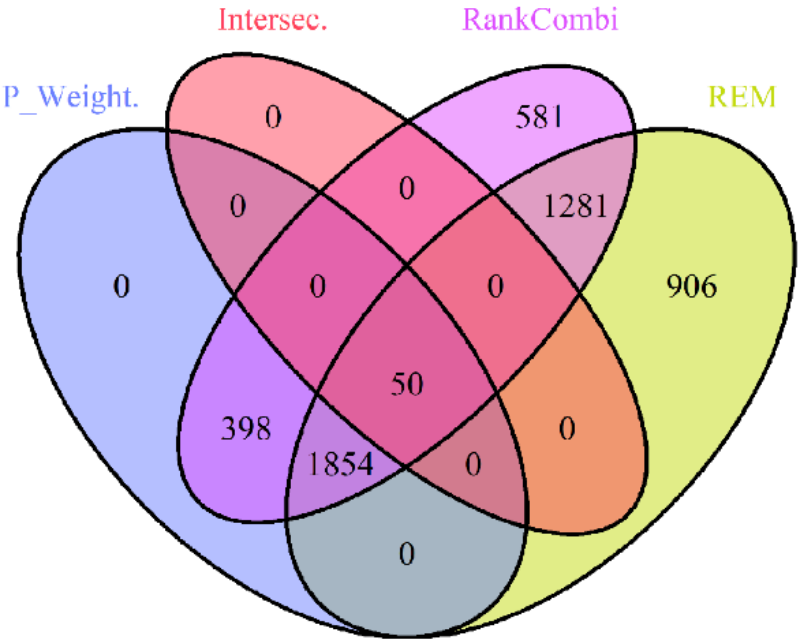
4.

Resultados

TCGA. Comparativa del metaanálisis a nivel de gen







P_Unweight. Combinación de p-valores sin ponderar (normal inversa)

P_Weight. Combinación de p-valores ponderado (Fisher)

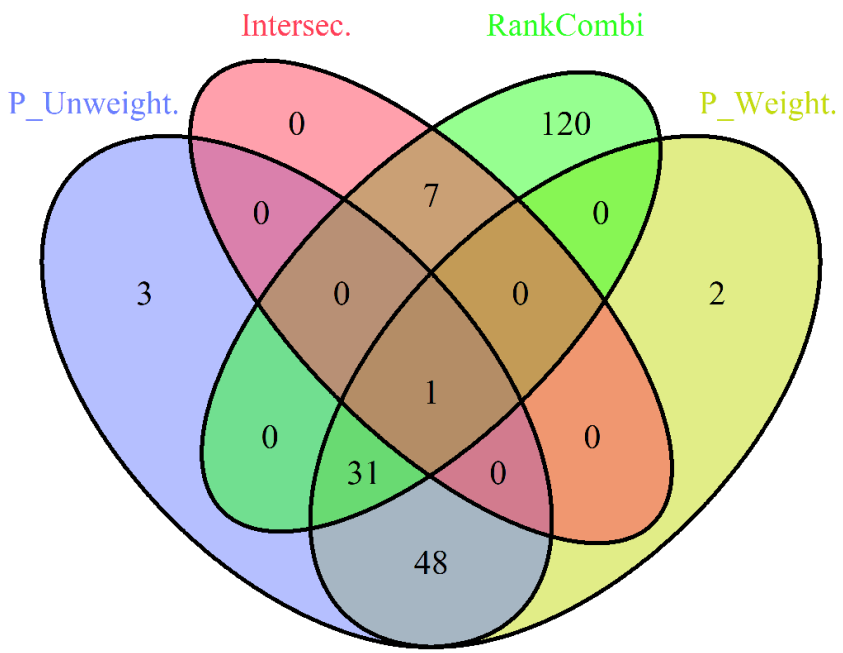
FEM. Combinación de tamaño del efecto: modelo de efectos fijos

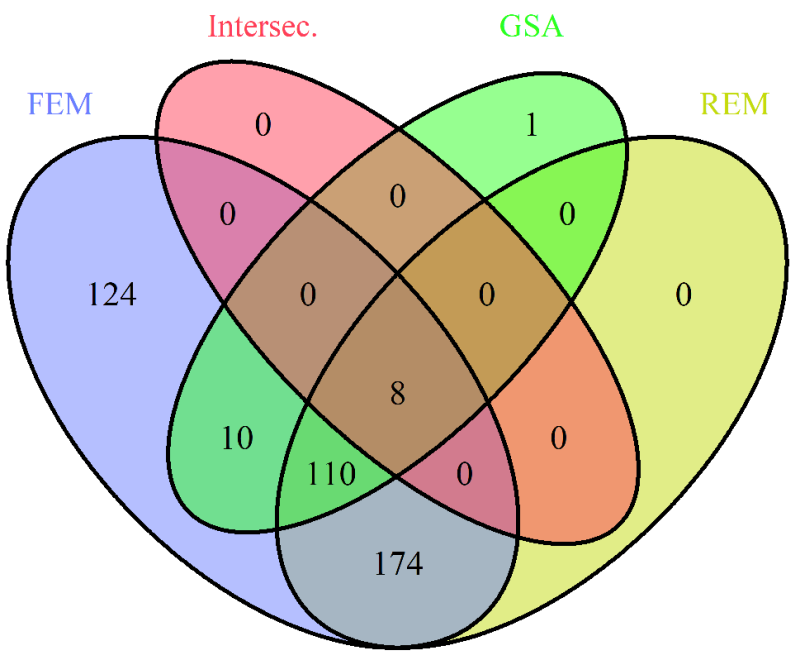
REM. Combinación de tamaño del efecto: modelo de efectos aleatorios

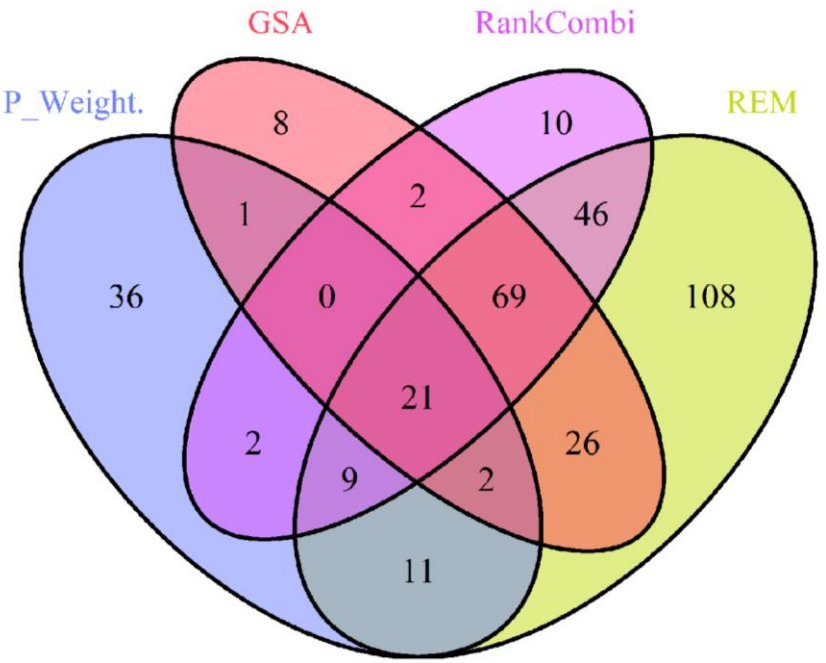
RankCombi. Combinación de rangos

Intersec. Intersección de genes diferencialmente expresados en todos los estudios

TCGA. Comparativa del metaanálisis a nivel de función







P_Unweight. Combinación de p-valores sin ponderar (normal inversa)

P_Weight. Combinación de p-valores ponderado (Fisher)

FEM. Combinación de tamaño del efecto: modelo de efectos fijos

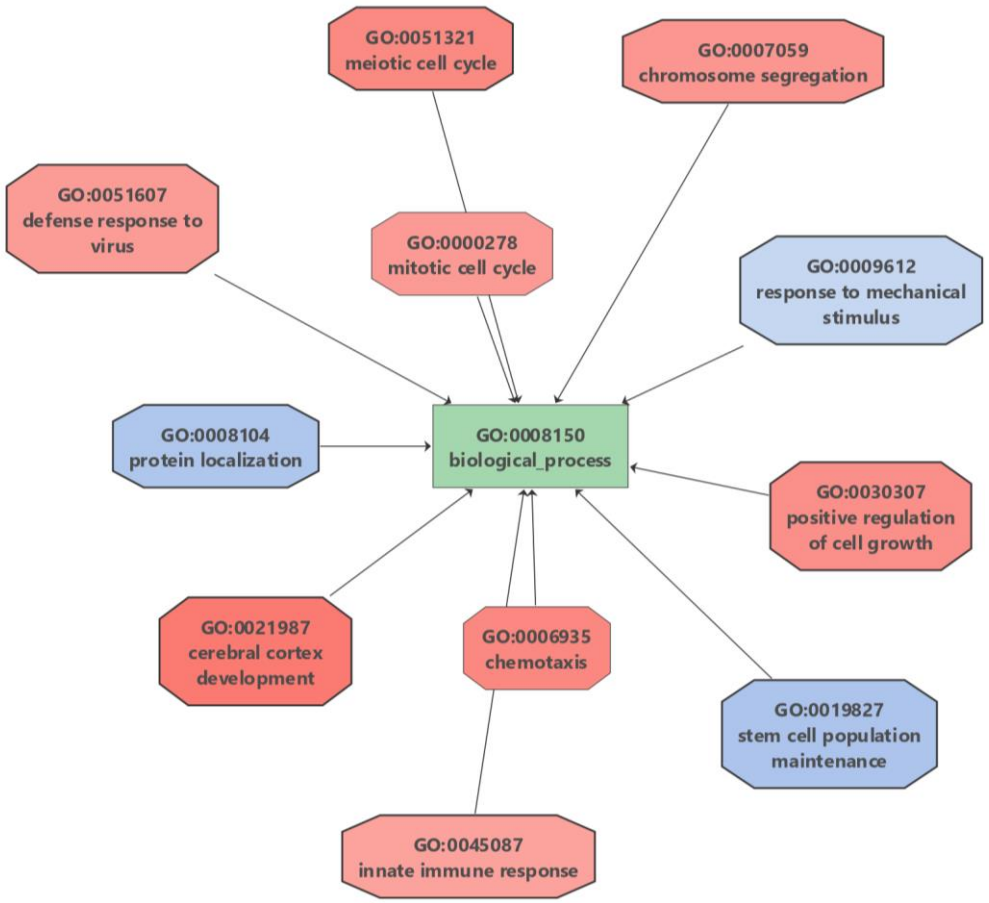
REM. Combinación de tamaño del efecto: modelo de efectos aleatorios

Intersec. Intersección de genes diferencialmente expresados en todos los estudios

RankCombi. Combinación de rangos

GSA. Metaanálisis a nivel de gen + Enriquecimiento


GO:0033209
tumor necrosis factor-mediated signaling pathway
GO:0031146
SCF-dependent proteasomal ubiquitin-dependent protein catabolic process
GO:0010972
negative regulation of G2/M transition of mitotic cell cycle
GO:0043161
proteasome-mediated ubiquitin-dependent protein catabolic process
GO:0038061
NIK/NF-kappaB signaling
GO:0030177
positive regulation of Wnt signaling pathway
GO:0000278
mitotic cell cycle
GO:0042178
xenobiotic catabolic process
GO:0006521
regulation of cellular amino acid metabolic process
GO:0001889
liver development
GO:0008150
biological_process
GO:0000082
G1/S transition of mitotic cell cycle
GO:0010719
negative regulation of epithelial to mesenchymal transition
GO:0006281
DNA repair
GO:1901687
glutathione derivative biosynthetic process
GO:0006974
cellular response to DNA damage stimulus
GO:0007059
chromosome segregation
GO:0001824
blastocyst development
GO:0070527
platelet aggregation
GO:0090090
negative regulation of canonical Wnt signaling pathway


GO:0051321
meiotic cell cycle
GO:0007059
chromosome segregation
GO:0051607
defense response to virus
GO:0000278
mitotic cell cycle
GO:0009612
response to mechanical stimulus
GO:0008104
protein localization
GO:0008150
biological_process
GO:0030307
positive regulation of cell growth
GO:0021987
cerebral cortex development
GO:0006935
chemotaxis
GO:0019827
stem cell population maintenance
GO:0045087
innate immune response

5.

Discusión y valoración de los métodos descritos

Valoración de los métodos de Metaanálisis empleados

Combinación de *p*-valores	Combinación de rangos	Combinación tamaño del efecto
Sencillo. Sólo se requiere el *p*-valor de la expresión diferencial / enriq. funcional. **Tamaño del efecto**. No presta atención al tamaño del efecto, por lo que hay que atender a comportamientos conflictivos. **Parámetros**. No constituye un modelo, no hay parámetros estimados y no resulta explicativo.	**Tamaño del efecto**. La ordenación en rankings obedece al tamaño del efecto. **Parámetros**. Al basarse en permutaciones, no se construye un modelo estadístico y el *p*-valor por gen o función solo se calcula de manera experimental.	**Tamaño del efecto**. Permite obtener una estimación global del tamaño del efecto. **Parámetros**. Construye un modelo donde se pueden estimar parámetros útiles para comprobar la heterogeneidad entre estudios, variabilidad intra-estudios, etc. **Validación**. Los parámetros permiten validar el modelo comprobando estudios influyentes y analizando sensibilidad y heterogeneidad.

6.

Conclusiones

Conclusiones

1. Los metaanálisis han demostrado ser una herramienta muy prometedora para combinar la información de estudios con datos similares.

2. En el campo de la expresión génica, el metaanálisis a nivel de función es de mayor utilidad para obtener una panorámica de las reacciones bioquímicas que ocurren en cada condición experimental y, salvo casos puntuales, ofrece más información que el metaanálisis a nivel de gen.

3. Todos los métodos empleados dieron buenos resultados, pero el modelo de efectos aleatorios parecen ser los más adecuados: permiten ajustar un modelo y estimar todos los parámetros, lo que los hace más interpretables. Además, modelizan la variabilidad intraestudio y la heterogeneidad entre estudios, muy común en datos ómicos que provienen de diferentes laboratorios.

4. Aunque cada vez el metaanálisis empieza a utilizarse más, el paso crítico continúa siendo la selección de estudios. Se hace cada vez más necesario emplear principios comunes para almacenar resultados de experimentos en bases de datos biológicas.

VNIVERSITAT DE VALÈNCIA

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Unidad de Bioinformática y Bioestadística

Métodos de Metaanálisis de Estudios Ómicos en Biomedicina

Trabajo Final de Máster en Bioestadística

Jesús Gutiérrez Botella

Tutor: Francisco García García
Tutor académico: Antonio López Quílez

23 de septiembre, 2020