

Estudios Transcriptómicos Masivos: Análisis de Microarrays

Fran(cisco J.) Romero-Campero



<https://bit.ly/3lg4pwX>



@greennetworks



@fran_rom_cam

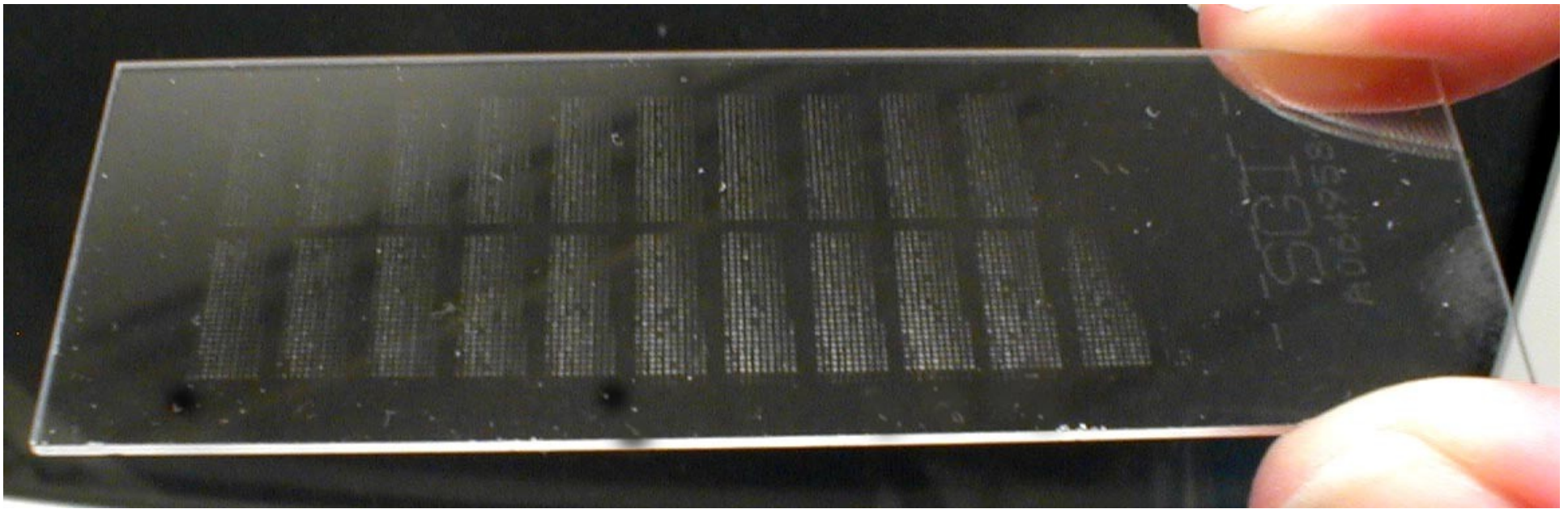


@franromcam

Dpt. de Ciencias de la Computación e
Inteligencia Artificial
Instituto de Bioquímica Vegetal y Fotosíntesis
Universidad de Sevilla

Microarrays o Micromatrices

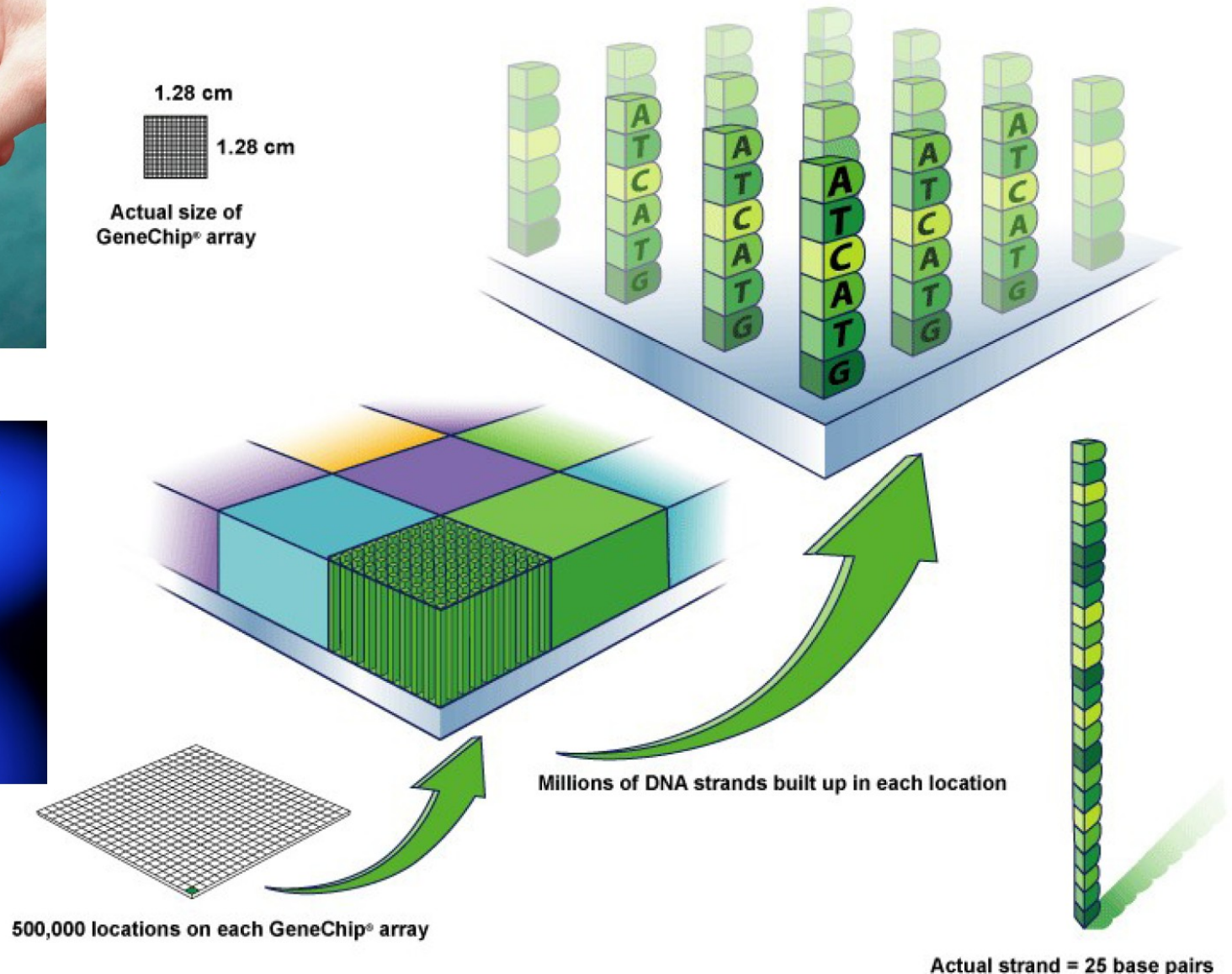
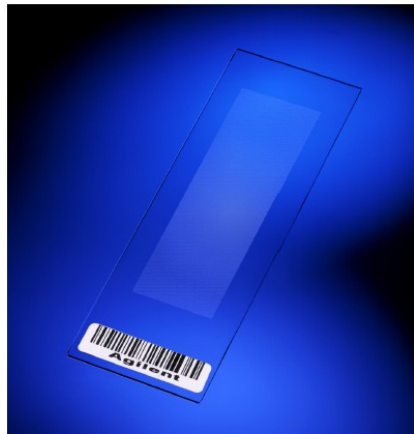
- Los **microarrays o micromatrices** son la tecnología de altas prestaciones clásica que permiten medir los niveles de expresión génicos en una muestra estimando el número de transcritos (mRNA).
- Un **microarray** es un soporte sólido donde se disponen en forma de matriz (en una distribución regular de filas y columnas) secuencias de DNA llamadas sondas (*probes* en inglés) que son complementarias a las secuencias de los transcritos conocidos en una especie en particular.



Microarrays o Micromatrices



1.28 cm
1.28 cm
Actual size of
GeneChip® array



Microarrays o Micromatrices

- El uso más común de los microarrays consiste en **extraer el mRNA de muestras de interés** a comparar. Por ejemplo, un tejido con un tratamiento especial y otro sin el tratamiento o una muestra del tipo silvestre y otra de un mutante.
- Normalmente este mRNA se **convierte a cDNA**, se **etiqueta** con fluorescencia o radioactividad y se coloca sobre el microarray para que **hibride** de forma específica con las correspondientes secuencias de DNA.
- Tras **lavar** el microarray podemos estimar el número de transcritos que han hibridado y así obtener una **medida global del nivel de expresión de los distintos genes a partir de la fluorescencia detectada**.



Microarrays o Micromatrices

Ventajas

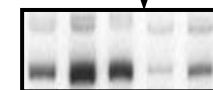
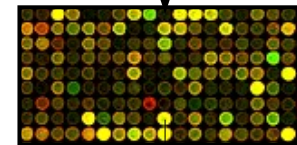
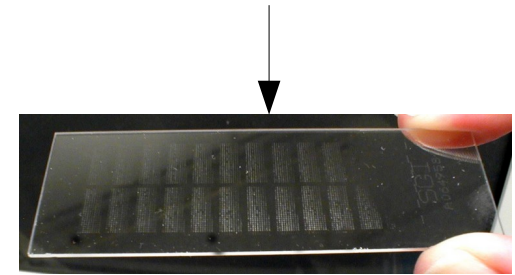
- Relativamente baratos.
- Protocolos de laboratorio bien definidos y depurados.
- Amplia variedad de herramientas de análisis bien establecidas y testeadas.

Desventajas

- Limitaciones a transcritos conocidos.
- Sensibilidad limitada debido a saturación de las sondas.
- Problemas de hibridación.
- Problemas de diseño.

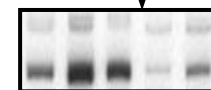
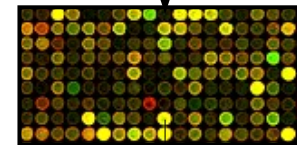
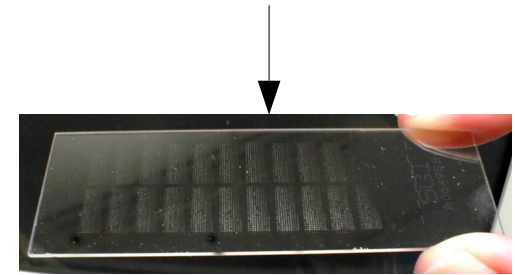
Fases del Estudio Transcriptómico basado en Microarrays

- **Fase 1:** Diseño Experimental.
- **Fase 2:** Extracción del RNA y preparación.
- **Fase 3:** Hibridación de las muestras etiquetadas.
- **Fase 4:** Análisis de imágenes.
- **Fase 5:** Análisis matemático-computacional de los datos.
- **Fase 6:** Confirmación o validación biológica.



Fases del Estudio Transcriptómico basado en Microarrays

- **Fase 1:** Diseño Experimental.
- **Fase 2:** Extracción del RNA y preparación.
- **Fase 3:** Hibridación de las muestras etiquetadas.
- **Fase 4:** Análisis de imágenes.
- **Fase 5:** Análisis matemático-computacional de los datos.
- **Fase 6:** Confirmación o validación biológica.



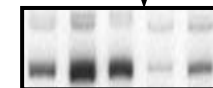
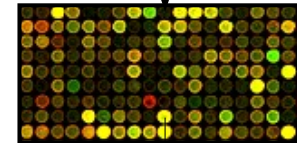
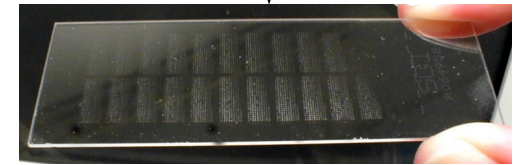
Fase 1: Diseño Experimental

- Esta es la fase más importante del estudio ya que es donde pueden producirse la mayor parte de los **errores irreparables**.
- Aquí se seleccionan las distintas **condiciones a comparar** y se fijan **controles**
 - Muestras de distintos tejidos.
 - Muestras sometidas a distintas condiciones o tratamientos.
 - Muestras provenientes de distintos genotipos.
 - Muestras provenientes de series temporales.
- Decidir y diseñar diferentes **réplicas**:
 - Réplicas técnicas o experimentales
 - Réplicas biológicas



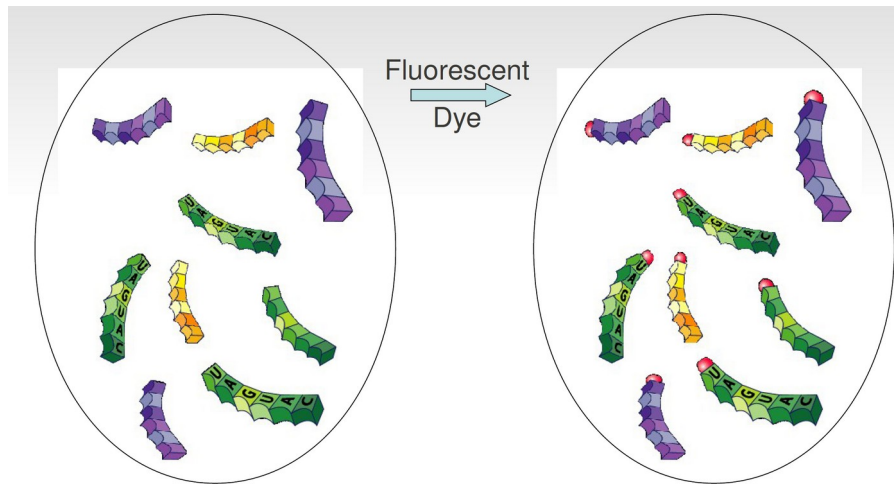
Fases del Estudio Transcriptómico basado en Microarrays

- **Fase 1:** Diseño Experimental.
- **Fase 2:** Extracción del RNA y preparación.
- **Fase 3:** Hibridación de las muestras etiquetadas.
- **Fase 4:** Análisis de imágenes.
- **Fase 5:** Análisis matemático-computacional de los datos.
- **Fase 6:** Confirmación o validación biológica.



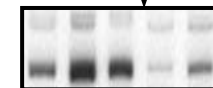
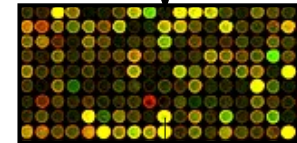
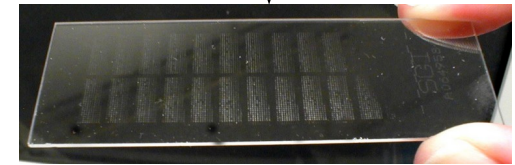
Fase 2: Extracción del RNA y Preparación

- Para la extracción del RNA de una muestra se suelen utilizar reactivos tales como el **TRIzol** (Invitrogen) posiblemente seguidos de pasos extras para la purificación como la **eliminación de DNA genómico o extracción polyA**.
- Es crítico tomar las diferentes muestras y extraer el RNA **bajo las mismas condiciones** que no se analizan en el experimento.
- Es necesario analizar la pureza y calidad del RNA con el espectrofotómetro mediando la **razón a_{260}/a_{280} , a_{260}/a_{230}** , $\sim 1.7 - 2.0$ (absorbancia ácidos nucleicos vs proteína, otros contaminantes), por **electroforesis en gel y RIN (RNA integrity number)**.
- Seguidamente el RNA se convierte en **cDNA** y se etiqueta con un marcador fluorescente.

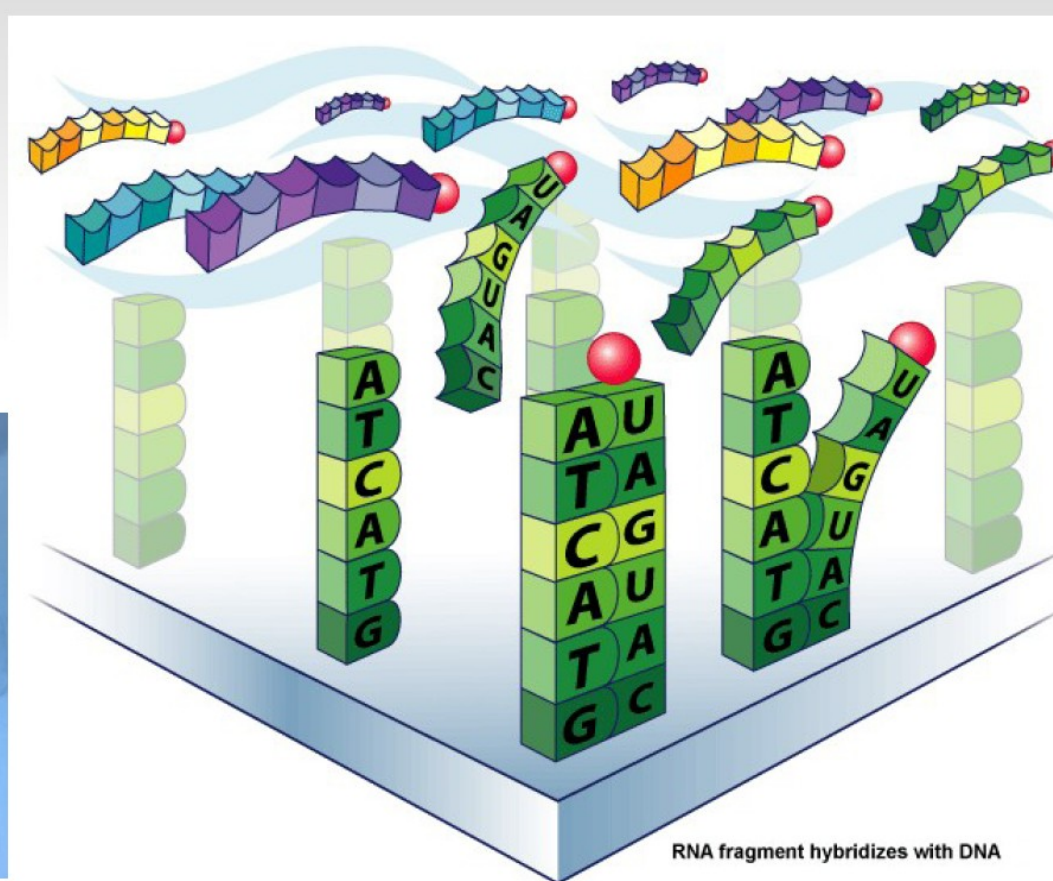


Fases del Estudio Transcriptómico basado en Microarrays

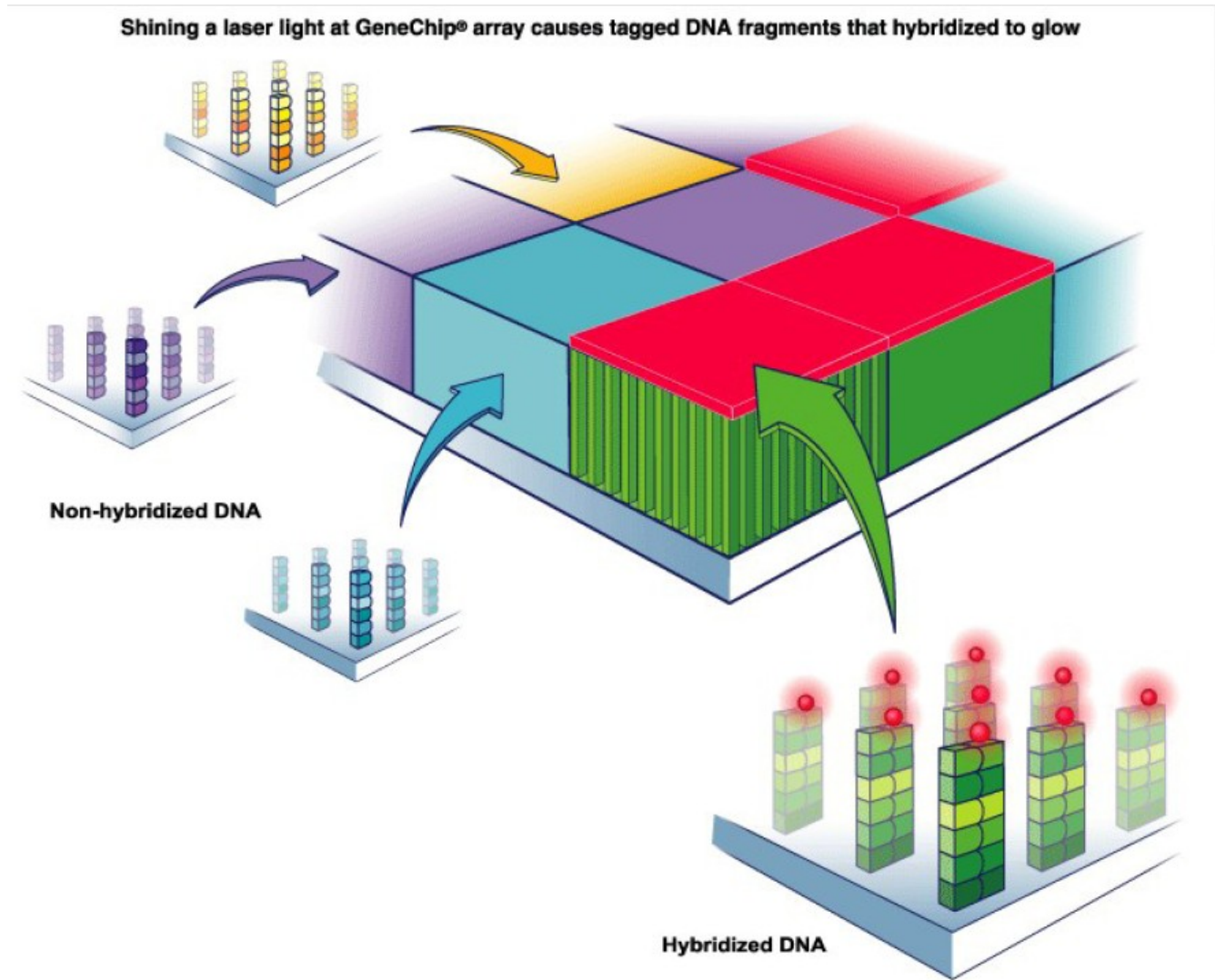
- **Fase 1:** Diseño Experimental.
- **Fase 2:** Extracción del RNA y preparación.
- **Fase 3:** Hibridación de las muestras etiquetadas.
- **Fase 4:** Análisis de imágenes.
- **Fase 5:** Análisis matemático-computacional de los datos.
- **Fase 6:** Confirmación o validación biológica.



Fase 3: Hibridación de las muestras etiquetadas

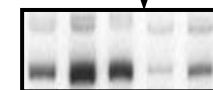
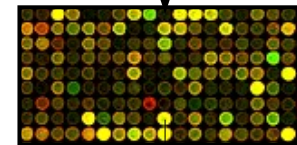
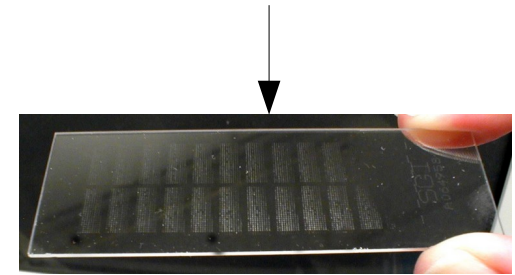


Fase 3: Hibridación de las muestras etiquetadas

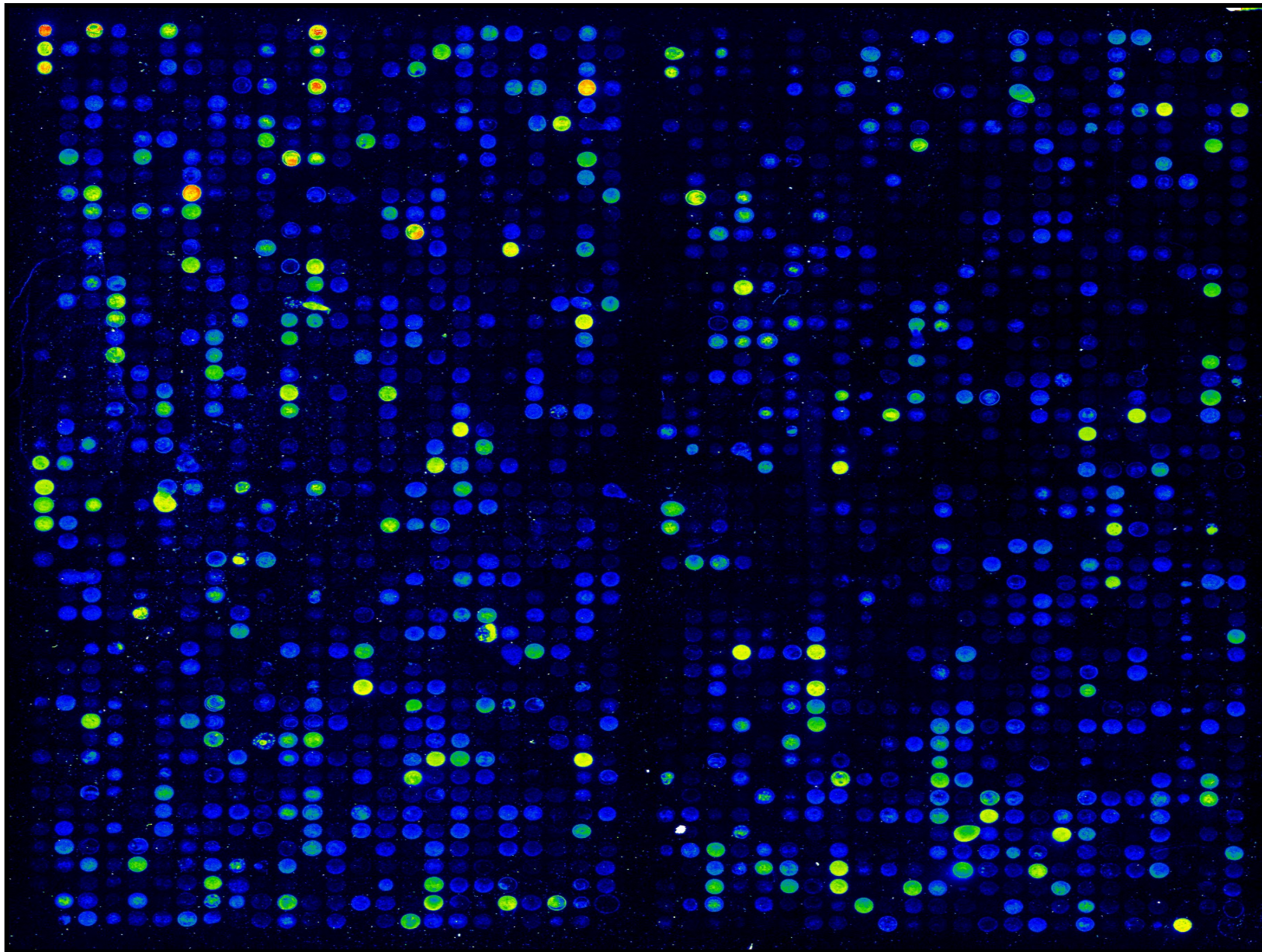


Fases del Estudio Transcriptómico basado en Microarrays

- **Fase 1:** Diseño Experimental.
- **Fase 2:** Extracción del RNA y preparación.
- **Fase 3:** Hibridación de las muestras etiquetadas.
- **Fase 4:** Análisis de imágenes.
- **Fase 5:** Análisis matemático-computacional de los datos.
- **Fase 6:** Confirmación o validación biológica.

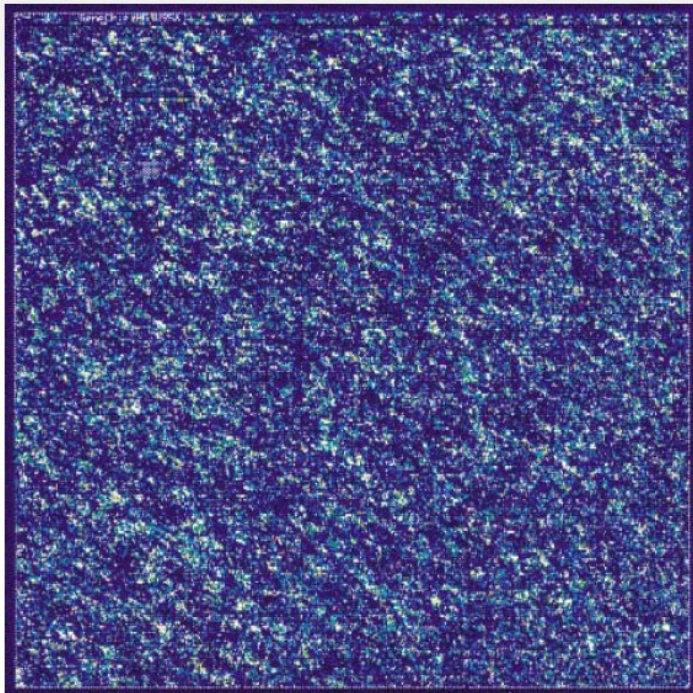


Fase 4: Análisis de Imágenes



Fase 4: Análisis de Imágenes

Para cada muestra
biológica



Obtenemos la
intensidad de
fluorescencia de miles
de transcritos

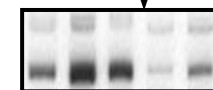
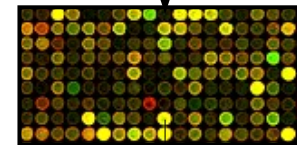
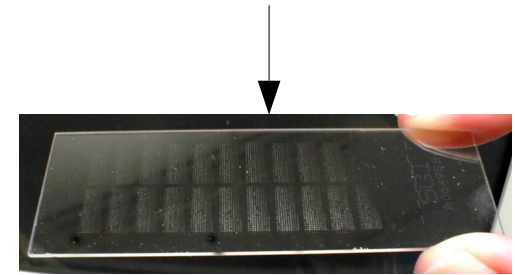
La medida de
intensidad es un
indicador de la
expresión génica

200000_s_at	134.4
200001_at	586.5
200002_at	1868.4
200003_s_at	1232.7
200004_at	1071.6
200005_at	312.8
200006_at	1712.6
200007_at	606.5
200008_s_at	421.9
200009_at	395.6
200010_at	1228.6
200011_s_at	132.5
200012_x_at	2606.3
200013_at	1572.9
200014_s_at	138.7
200015_s_at	124.1
200016_x_at	1058.7
200017_at	889.4
200018_at	3964.2
200019_s_at	1069.9
200020_at	212.1
200021_at	1018.1
200022_at	1254.8
200023_s_at	1202.8
200024_at	2460.6

Trabajaremos con microarrays diseñados por la affymetrix cuyos datos brutos no procesados suelen aparecer en formato CEL

Fases del Estudio Transcriptómico basado en Microarrays

- **Fase 1:** Diseño Experimental.
- **Fase 2:** Extracción del RNA y preparación.
- **Fase 3:** Hibridación de las muestras etiquetadas.
- **Fase 4:** Análisis de imágenes.
- **Fase 5:** Análisis matemático-computacional de los datos.
- **Fase 6:** Confirmación o validación biológica.



Fase 5: Análisis Matemático/Computacional

MATERIALES:

La comunidad científica persigue hacer disponible libremente los datos transcriptómicos en **bases de datos** tales como **GEO**:

<http://www.ncbi.nlm.nih.gov/geo/>

con el doble objetivo de garantizar la reproducibilidad de los resultados científicos y facilitar el desarrollo incremental de la ciencia donde unos resultados se basan en resultados previos.

Series GSE21582		Query DataSets for GSE21582
Status	Public on Dec 01, 2010	
Title	Expression analysis of pye-1 mutants and root pericycle cells to iron sufficient or iron deficient conditions	
Organism	Arabidopsis thaliana	
Experiment type	Expression profiling by array	
Summary	This SuperSeries is composed of the SubSeries listed below.	
Overall design	Refer to individual Series	
Citation(s)	Long TA, Tsukagoshi H, Busch W, Lahner B et al. The bHLH transcription factor POPEYE regulates response to iron deficiency in Arabidopsis roots. <i>Plant Cell</i> 2010 Jul;22(7):2219-36. PMID: 20675571	
Submission date	Apr 28, 2010	
Last update date	Jun 12, 2017	
Contact name	Terri Anita Long	
E-mail(s)	tlong@duke.edu	
Phone	919-613-8202	
Fax	919-613-8177	
Organization name	Duke University	
Department	Biology	
Lab	Philip Benfey	
Street address	Box 90338	
City	Durham	
State/province	NC	
ZIP/Postal code	27707	
Country	USA	
Platforms (1)	GPL198 [ATH1-121501] Affymetrix Arabidopsis ATH1 Genome Array	
Samples (10)	GSM535984 Arabidopsis, WT, +Fe, replicate 1	
More...	GSM535985 Arabidopsis, WT, +Fe, replicate 2	
	GSM535986 Arabidopsis, WT, -Fe, replicate 1	

Fase 5: Análisis Matemático/Computacional

MÉTODOS:

R constituye una **plataforma de software libre** para el *análisis estadístico* y *representación gráfica* de los resultados. Mantenido y desarrollo activamente por una gran comunidad de investigadores. Puede comunicarse de forma sencilla con otros lenguajes de programación como C, Java, FORTRAN, Python, Perl, etc.

Bioconductor es un **proyecto de software libre colaborativo** que consiste de una serie de paquetes de R. Originalmente se desarrolló para el análisis de datos de microarray pero en la actualidad cuenta con paquetes para el **estudio de datos de secuenciación, ensayos celulares, ensayos de altas prestaciones, etc.**

www.bioconductor.org

```
> install.packages("BiocManager") #sólo la primera vez  
> BiocManager::install("nombre_paquete")
```

Fase 5: Análisis Matemático/Computacional

El análisis computacional básico de datos de microarrays se divide en los siguientes pasos:

- **Paso 5.1:** Análisis de la calidad
- **Paso 5.2:** Preprocesamiento de los datos.
- **Paso 5.3:** Estimación de los niveles de expresión.
- **Paso 5.4:** Selección de genes diferencialmente expresados.
- **Paso 5.5:** Anotación funcional.

Fase 5: Análisis Matemático/Computacional

El análisis computacional básico de datos de microarrays se divide en los siguientes pasos:

- **Paso 5.1:** Análisis de la calidad

Quality
control

- **Paso 5.2:** Preprocesamiento de los datos.
- **Paso 5.3:** Estimación de los niveles de expresión.

Data
Processing

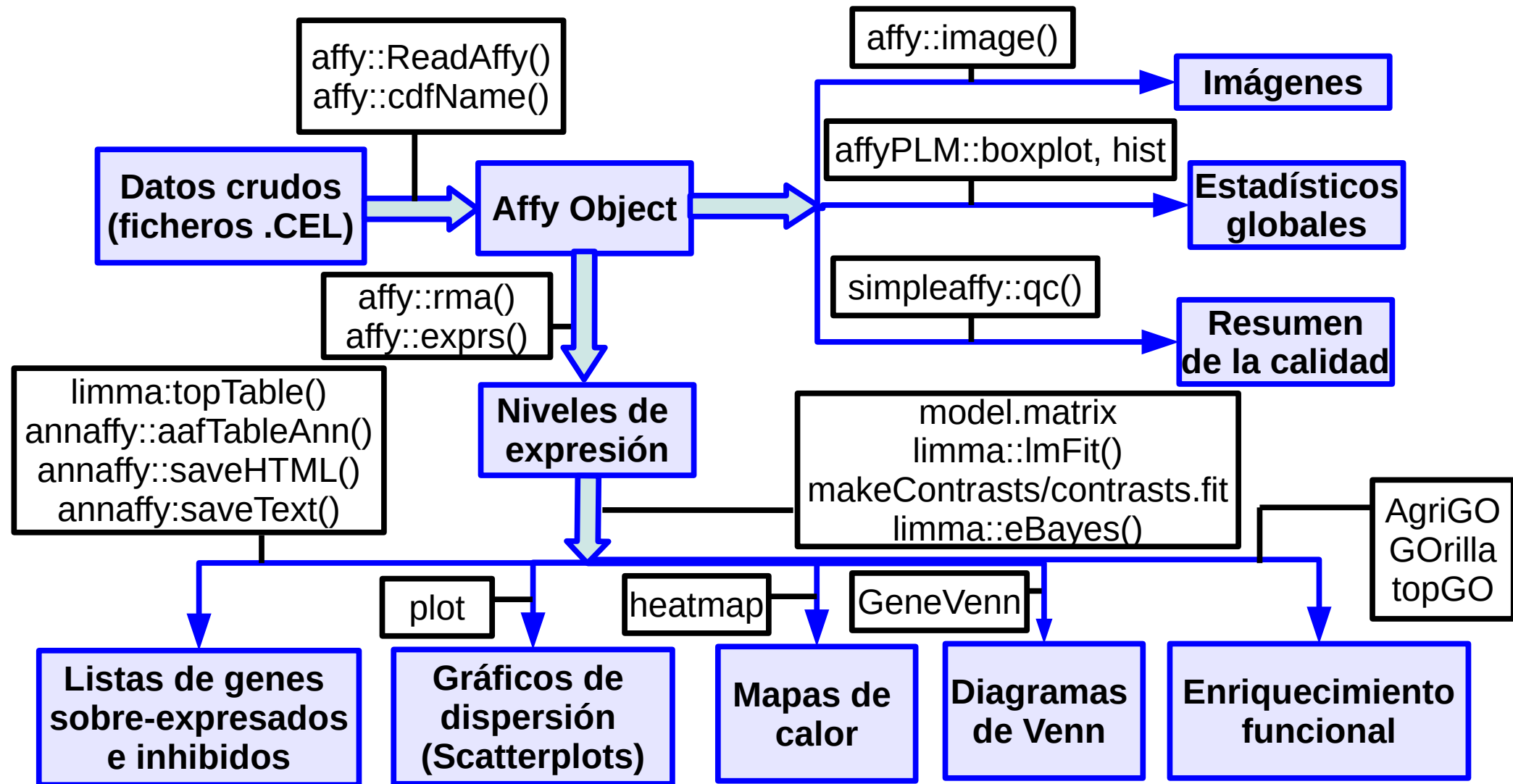
- **Paso 5.4:** Selección de genes diferencialmente expresados.

Inferential
Statistics

- **Paso 5.5:** Anotación funcional.

Exploratory
Analysis

Fase 5: Análisis Matemático/Computacional



Ejemplo: Estudio del Efecto Global en el Transcriptoma de *Arabidopsis thaliana* del gen **POPEYE** y el hambre de hierro

POPEYE codifica un factor de transcripción del tipo bHLH (basic helix loop helix) que en *Arabidopsis thaliana* regula la respuesta a deficiencia de hierro.



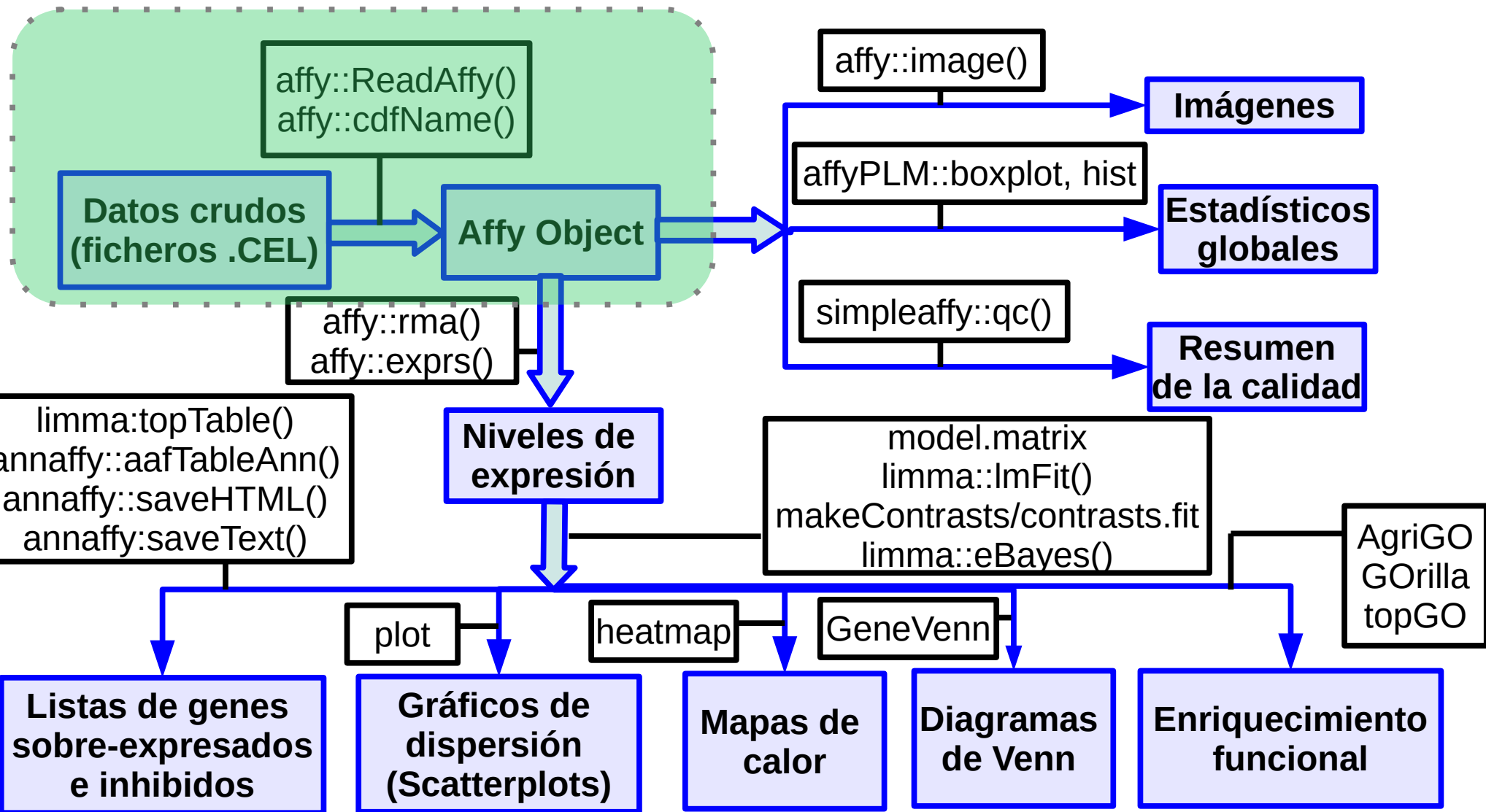
Estudiamos el efecto del gen POPEYE y la deficiencia de hierro. Tomamos las siguientes muestras con dos réplicas biológicas:

- WT con hierro
- WT sin hierro
- POPEYE con hierro
- POPEYE sin hierro

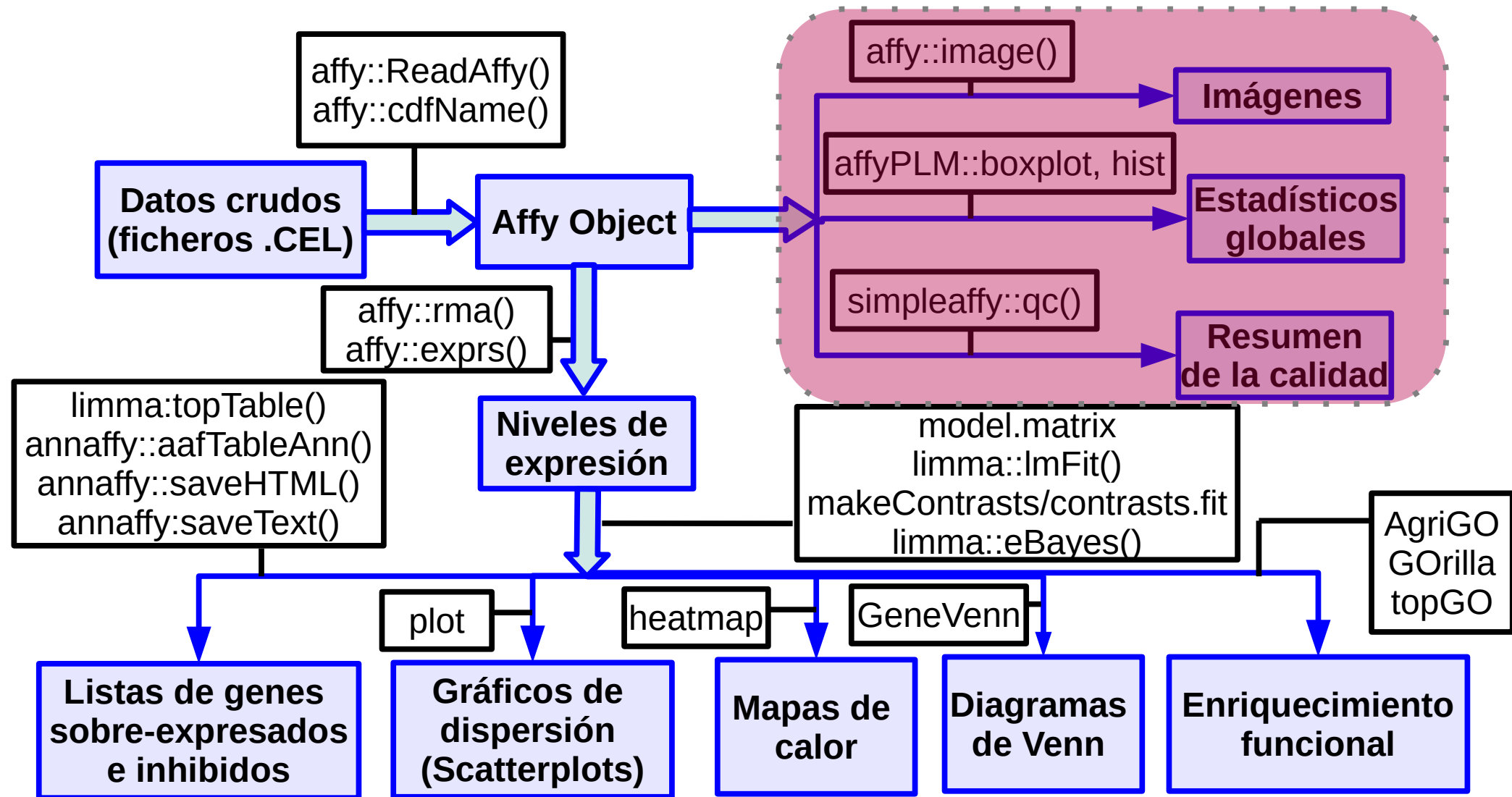


Long TA et al. The bHLH transcription factor POPEYE regulates response to iron deficiency in *Arabidopsis* roots. *Plant Cell* 2010 Jul;22(7):2219-36.
Accession Number: GSE21582

Paso 5.0: Lectura de Datos

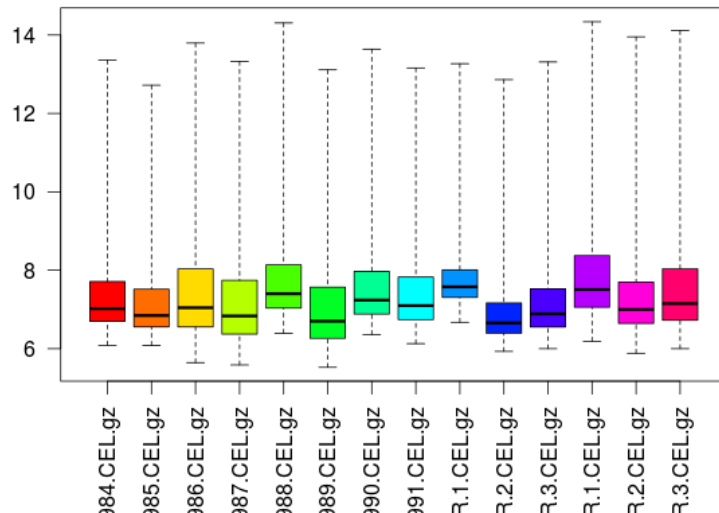
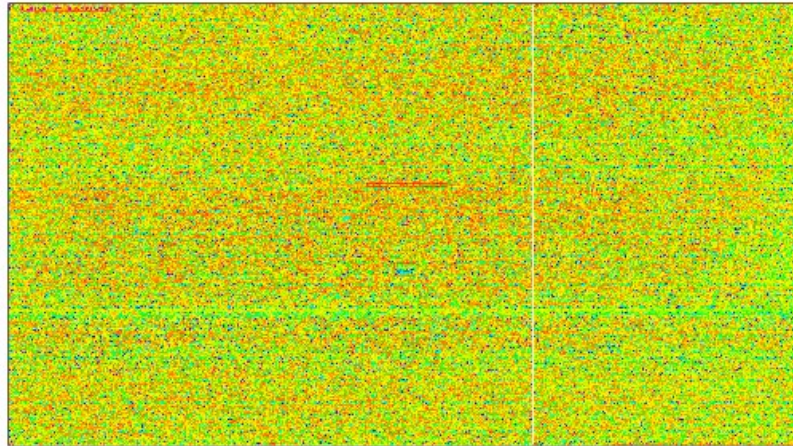


Paso 5.1: Análisis de la Calidad

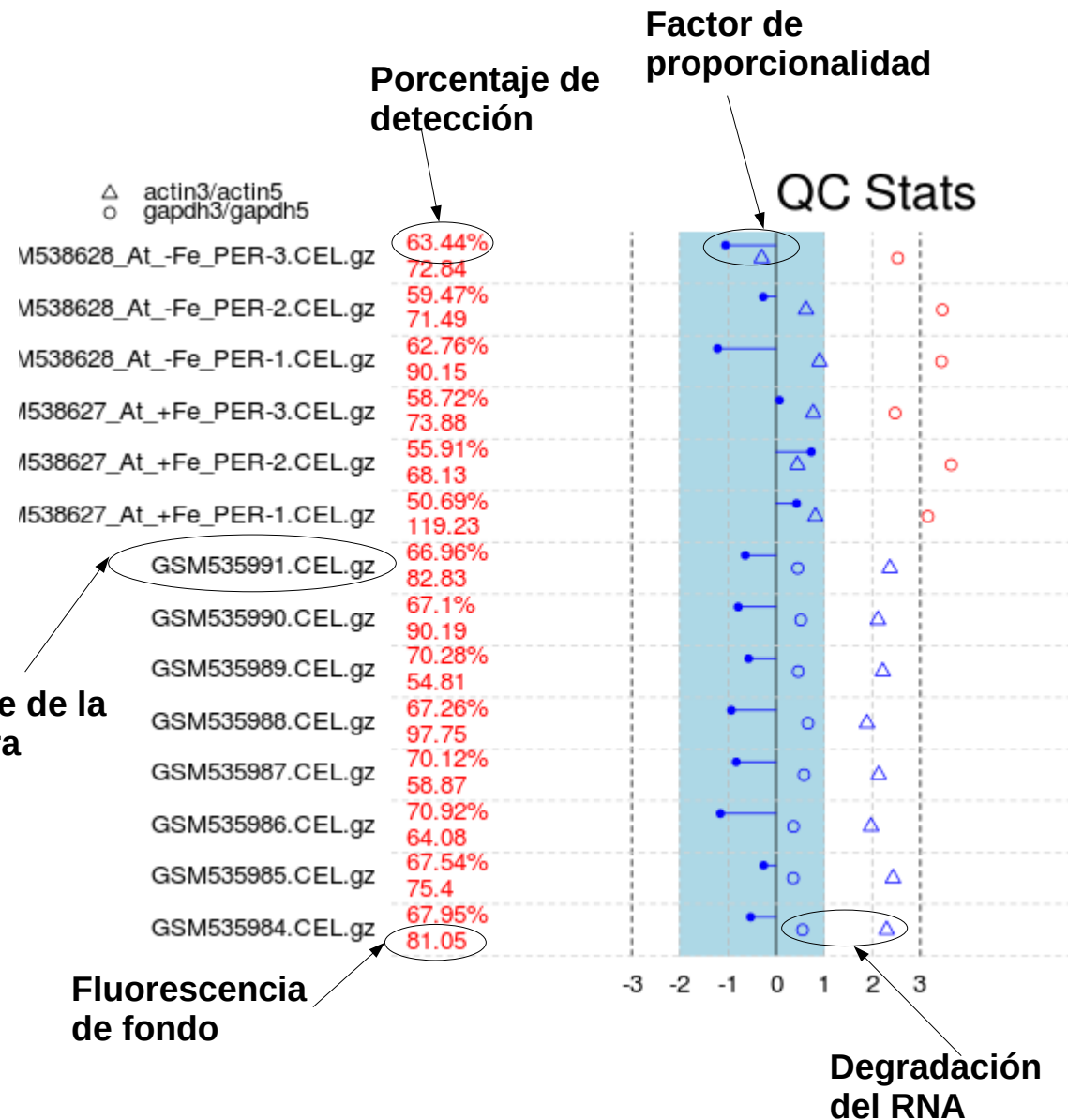


Paso 5.1: Análisis de la Calidad

GSM538628_At_-Fe_PER-3.CEL.gz

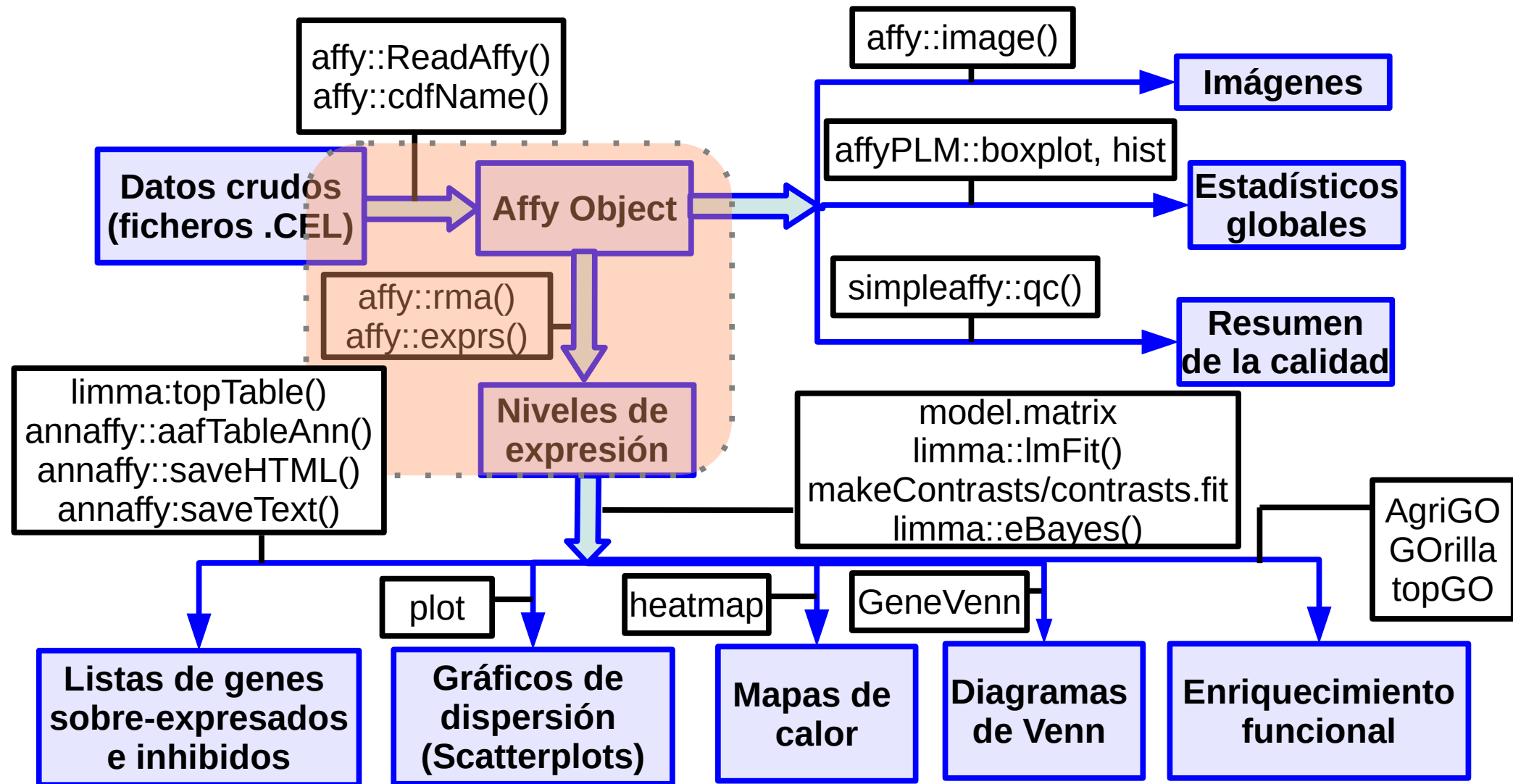


Nombre de la muestra



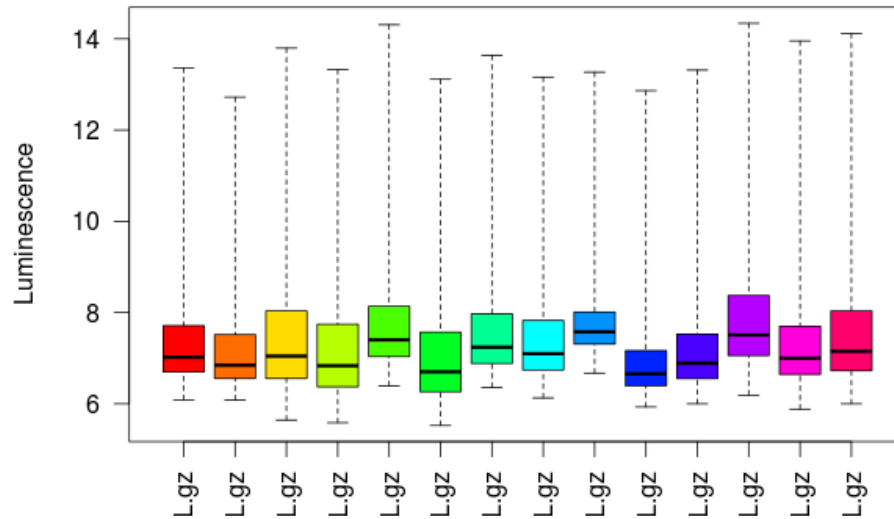
Paso 5.2: Preprocesamiento

Paso 5.3: Estimación Niveles de Expresión



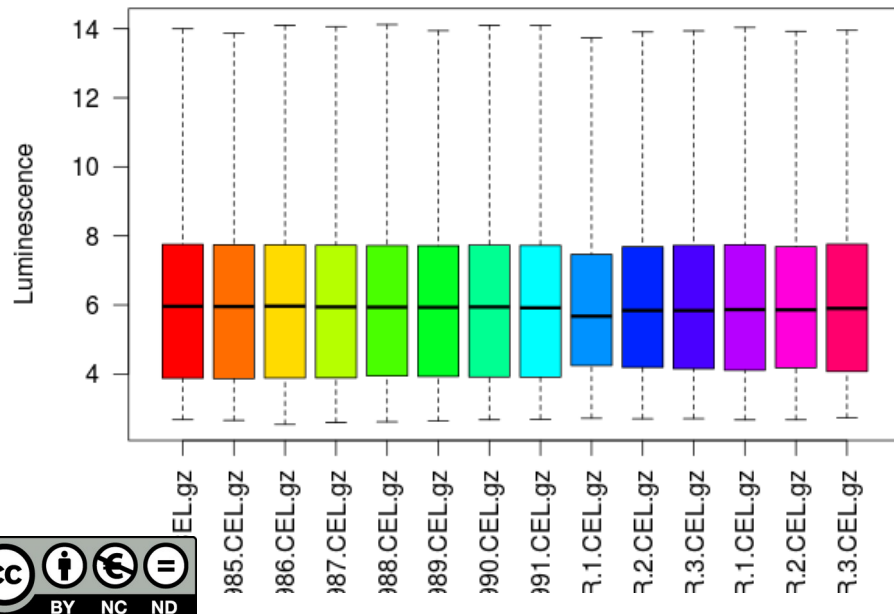
Paso 5.2: Preprocesamiento

Paso 5.3: Estimación Niveles de Expresión



El principal objetivo del preprocesamiento de los datos consiste en eliminar el ruido producido por el sesgo de la técnica experimental utilizada (variabilidad experimental) a la vez que se conserva la variabilidad generada por la condición o fenotipo estudiado (variabilidad biológica).

Robust Multiarray Average (RMA) es un algoritmo que realiza una corrección o sustracción de la fluorescencia de fondo, normalización de la fluorescencia de cada sonda y una estimación de los niveles de expresión de los genes representados por las distintas sondas del microarray utilizado.



Paso 5.2: Preprocesamiento

Paso 5.3: Estimación Niveles de Expresión

Las muestras se organizan por columnas

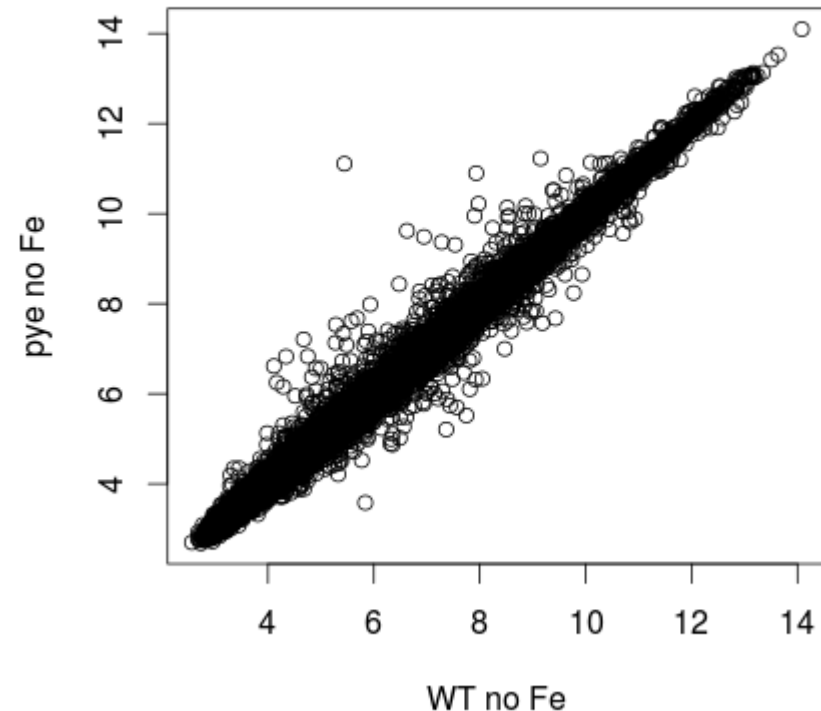
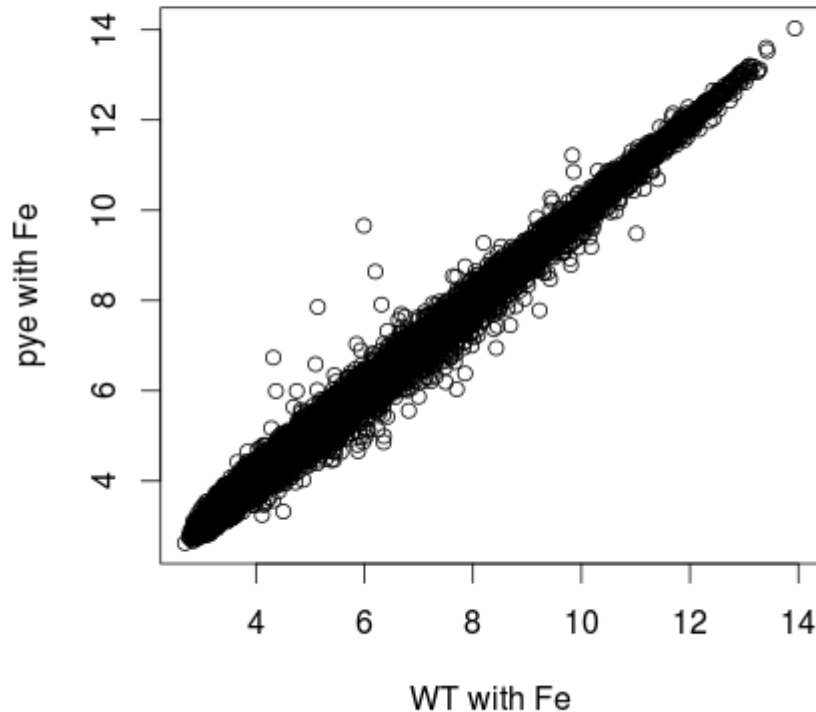
```
> head(expression.level)
```

	WT_with_Fe_1	WT_with_Fe_2	WT_no_Fe_1	WT_no_Fe_2	pye_with_Fe_1	pye_with_Fe_2
▶ 244901_at	4.191814	4.481378	3.839588	3.993979	3.822802	4.330534
▶ 244902_at	4.727383	4.933480	4.537309	4.762214	4.601498	4.939710
▶ 244903_at	5.569595	6.274849	5.160593	6.061207	5.390620	6.119615
▶ 244904_at	5.146491	5.113851	5.117658	5.138793	4.966478	5.067616
▶ 244905_at	3.869215	3.793270	3.876314	3.776267	3.998623	4.182826
▶ 244906_at	5.746207	5.815738	5.826470	5.709536	5.698041	6.150307

Los identificadores de las sondas que representan genes se organizan por filas

Paso 5.2: Preprocesamiento

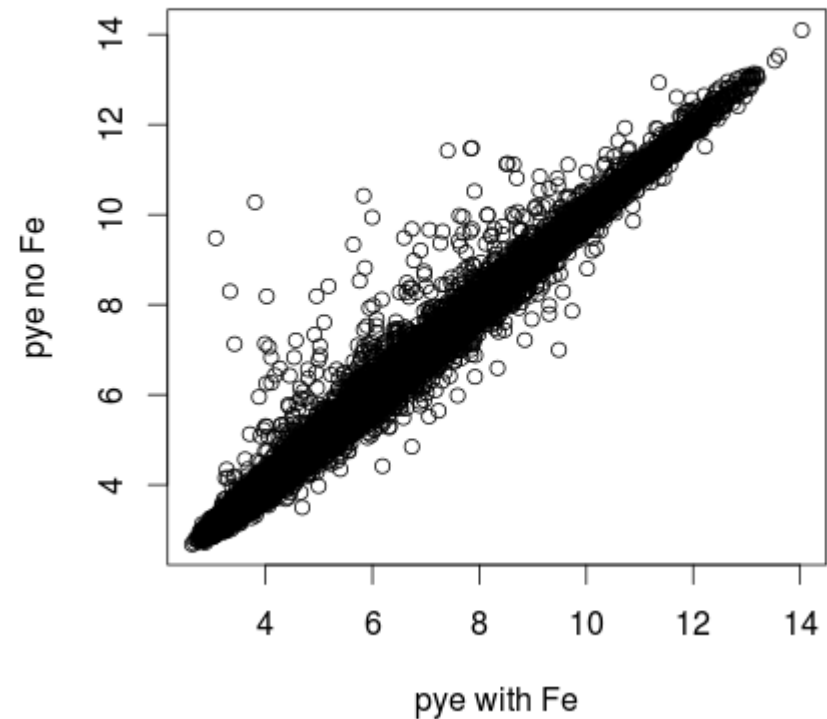
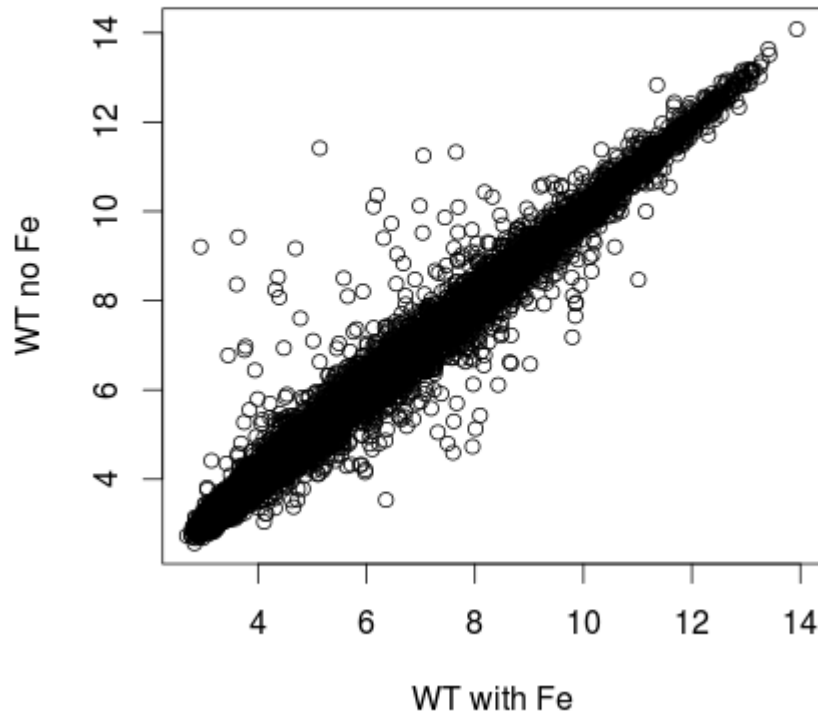
Paso 5.3: Estimación Niveles de Expresión



Comparación de dos genotipos en la misma condición aporta información sobre el efecto del cambio genotípico en respuesta a la condición estudiada.

Paso 5.2: Preprocesamiento

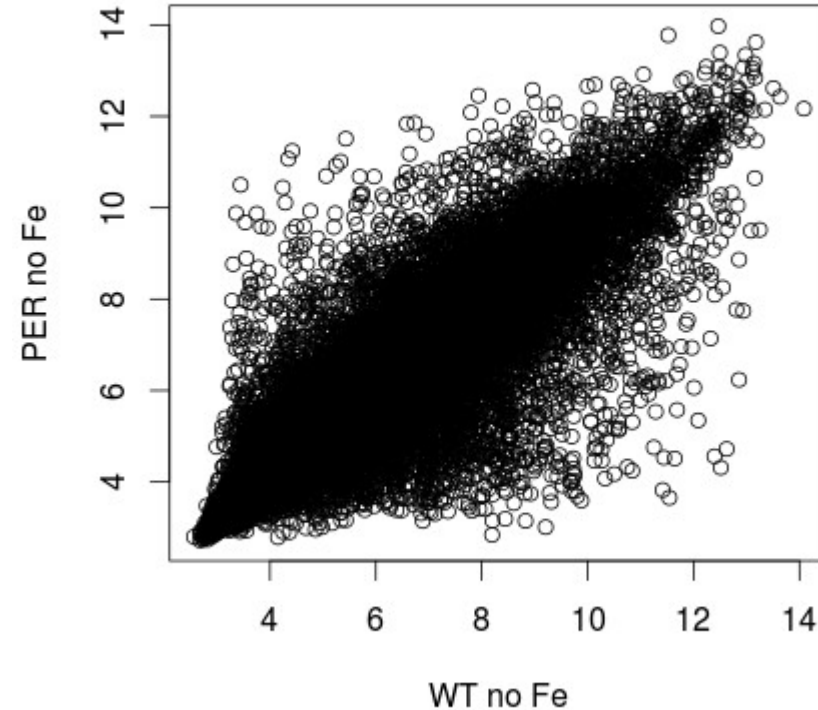
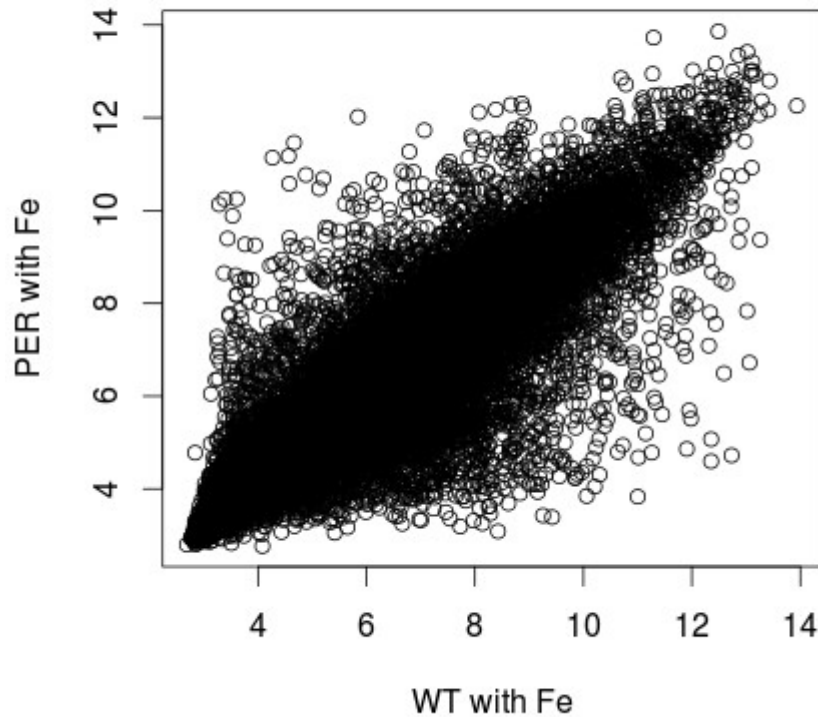
Paso 5.3: Estimación Niveles de Expresión



Comparación de un mismo genotipo en dos condiciones distintas aporta información sobre la respuesta del genotipo estudiado al cambio en las condiciones analizadas.

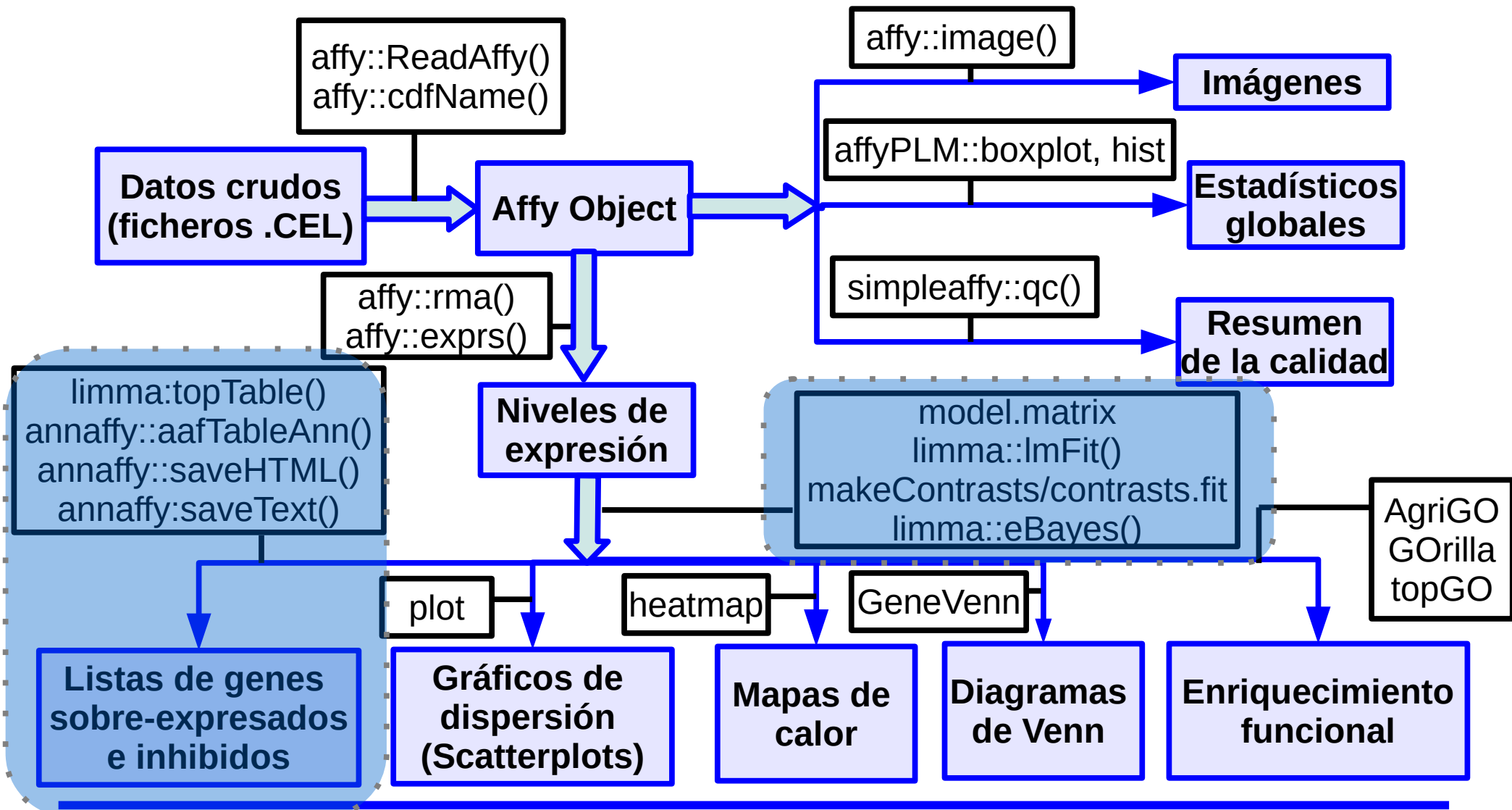
Paso 5.2: Preprocesamiento

Paso 5.3: Estimación Niveles de Expresión



La comparación de dos tejidos diferentes en un organismos multicelular normalmente muestra una masiva diferencia en expresión génica.

Paso 5.4: Selección de Genes Diferencialmente Expresados



Paso 5.4: Selección de Genes Diferencialmente Expresados

Cuando se comparan los transcriptomas de dos genotipos diferentes o de un mismo genotipo bajo distintas condiciones existen diversos métodos para determinar genes expresados de forma diferencial o *differentially expressed genes* (DEGs) en inglés:

- **Método basado en el *fold-change*** (factor de proporcionalidad): Se fija un umbral para el fold-change típicamente 2, 4 u 8 que en log2 corresponde a 1, 2 ó 3. Los DEGs son aquellos que incrementan (o decrementan) su expresión por encima de dicho umbral (por debajo de menos dicho umbral). Este método es biológicamente interpretable de forma directa y no requiere un alto número de réplicas biológicas. Se aplica especialmente a estudios con organismo modelos donde no son necesarias muchas réplicas.
- **Método basado en inferencia estadística:** Para aplicar este método es necesario tener un alto número de réplicas biológicas. Para cada gen y para cada pareja de genotipos/condiciones a comparar se formula un contraste de hipótesis sobre igualdad de medias. Normalmente este contraste de hipótesis utiliza un estadístico similar a la t-student. Se fija un nivel de significancia y se calcula el correspondiente p-valor (y p-valor corregido para el testeo múltiple o q-valor). Si dicho p-valor (o q-valor) es menor que el nivel de significancia se asume que el correspondiente gen se expresa de forma diferencial en los genotipos/condiciones estudiadas.
- **Combinación de los dos anteriores métodos**

Paso 5.4: Selección de Genes Diferencialmente Expresados

Cuando se comparan los transcriptomas de dos genotipos diferentes o de un mismo genotipo bajo distintas condiciones existen diversos métodos para determinar genes expresados de forma diferencial o *differentially expressed genes* (DEGs) en inglés:

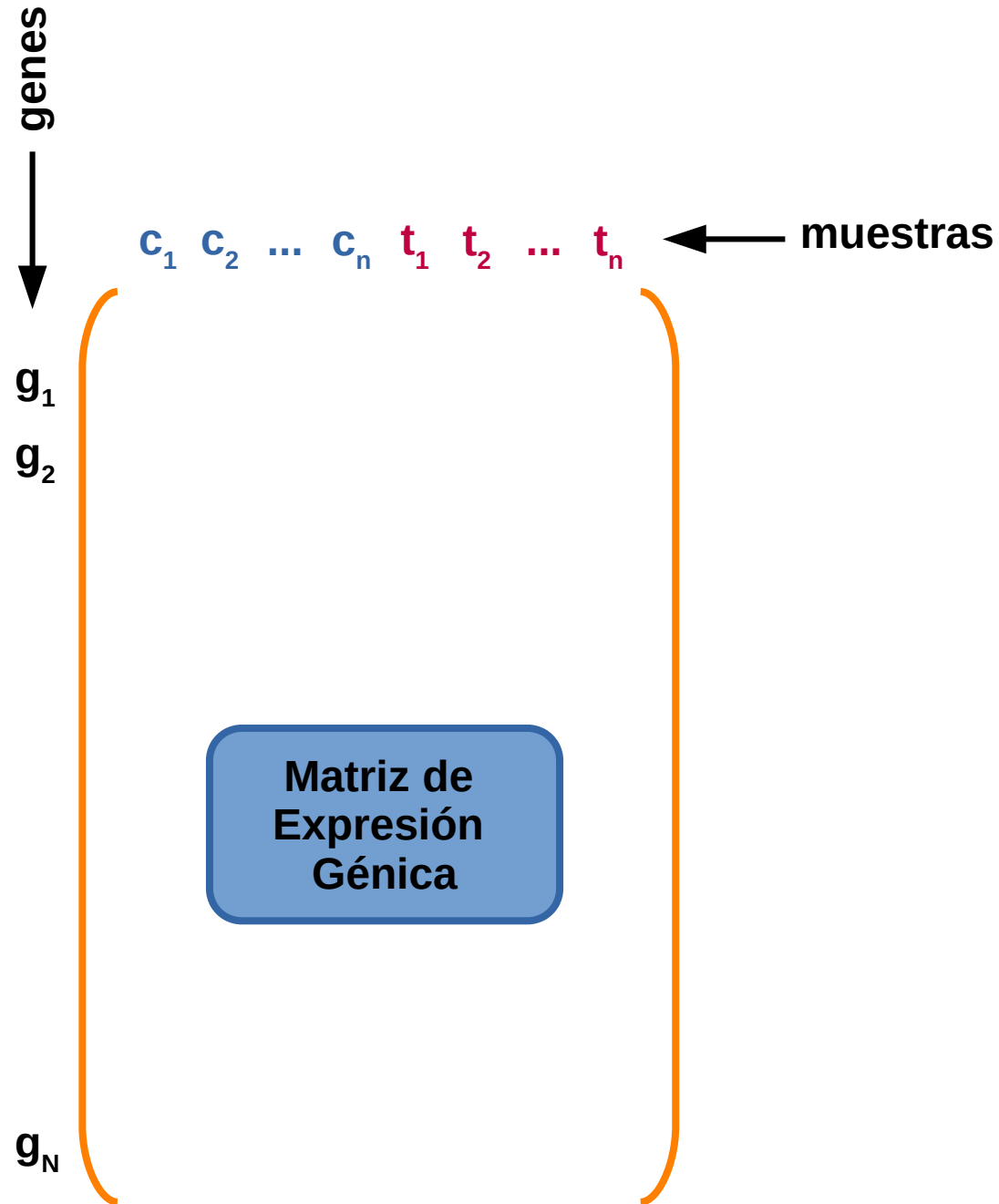
- **Método basado en el *fold-change*** (factor de proporcionalidad): Se fija un umbral para el fold-change típicamente 2, 4 u 8 que en log2 corresponde a 1, 2 ó 3. Los DEGs son aquellos que incrementan (o decrementan) su expresión por encima de dicho umbral (por debajo de menos dicho umbral). Este método es biológicamente interpretable de forma directa y no requiere un alto número de réplicas biológicas. Se aplica especialmente a estudios con organismo modelos donde no son necesarias muchas réplicas.
- **Método basado en inferencia estadística:** Para aplicar este método es necesario tener un alto número de réplicas biológicas. Para cada gen y para cada pareja de genotipos/condiciones a comparar se formula un contraste de hipótesis sobre igualdad de medias. Normalmente este contraste de hipótesis utiliza un estadístico similar a la t-student. Se fija un nivel de significancia y se calcula el correspondiente p-valor (y p-valor corregido para el testeo múltiple o q-valor). Si dicho p-valor (o q-valor) es menor que el nivel de significancia se asume que el correspondiente gen se expresa de forma diferencial en los genotipos/condiciones estudiadas.
- **Combinación de los dos anteriores métodos**

Paso 5.4: Selección de Genes Diferencialmente Expresados

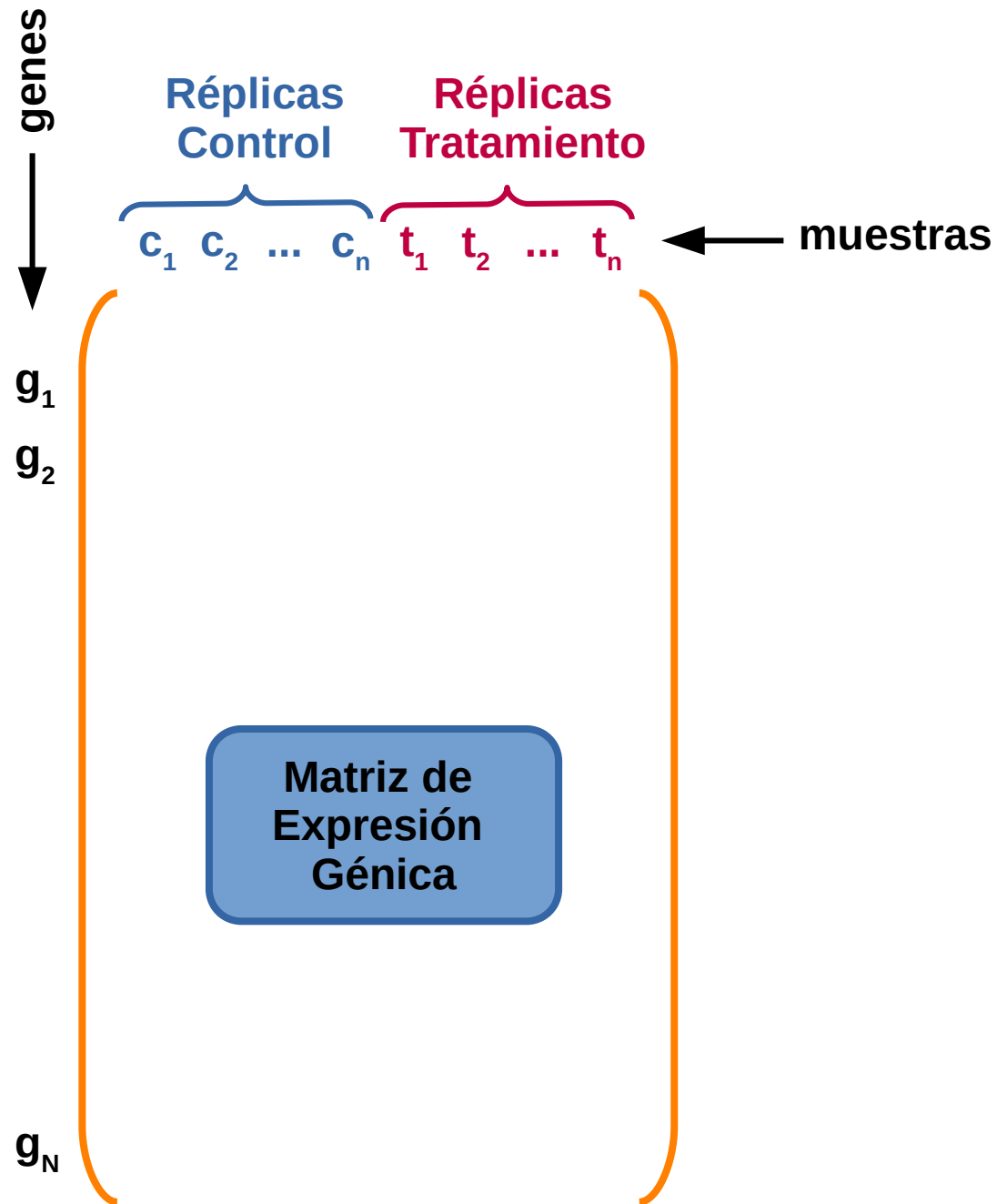


Matriz de
Expresión
Génica

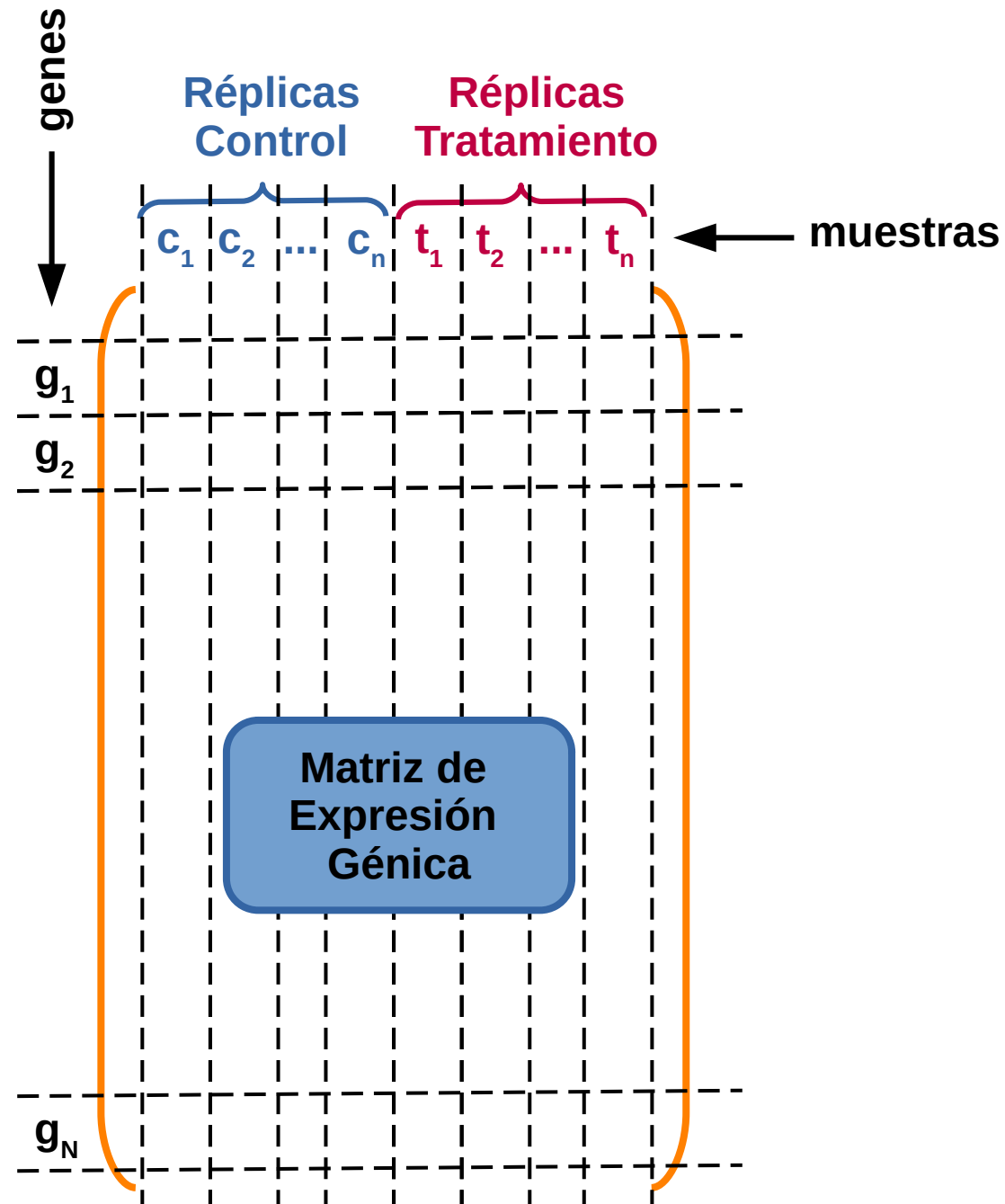
Paso 5.4: Selección de Genes Diferencialmente Expresados



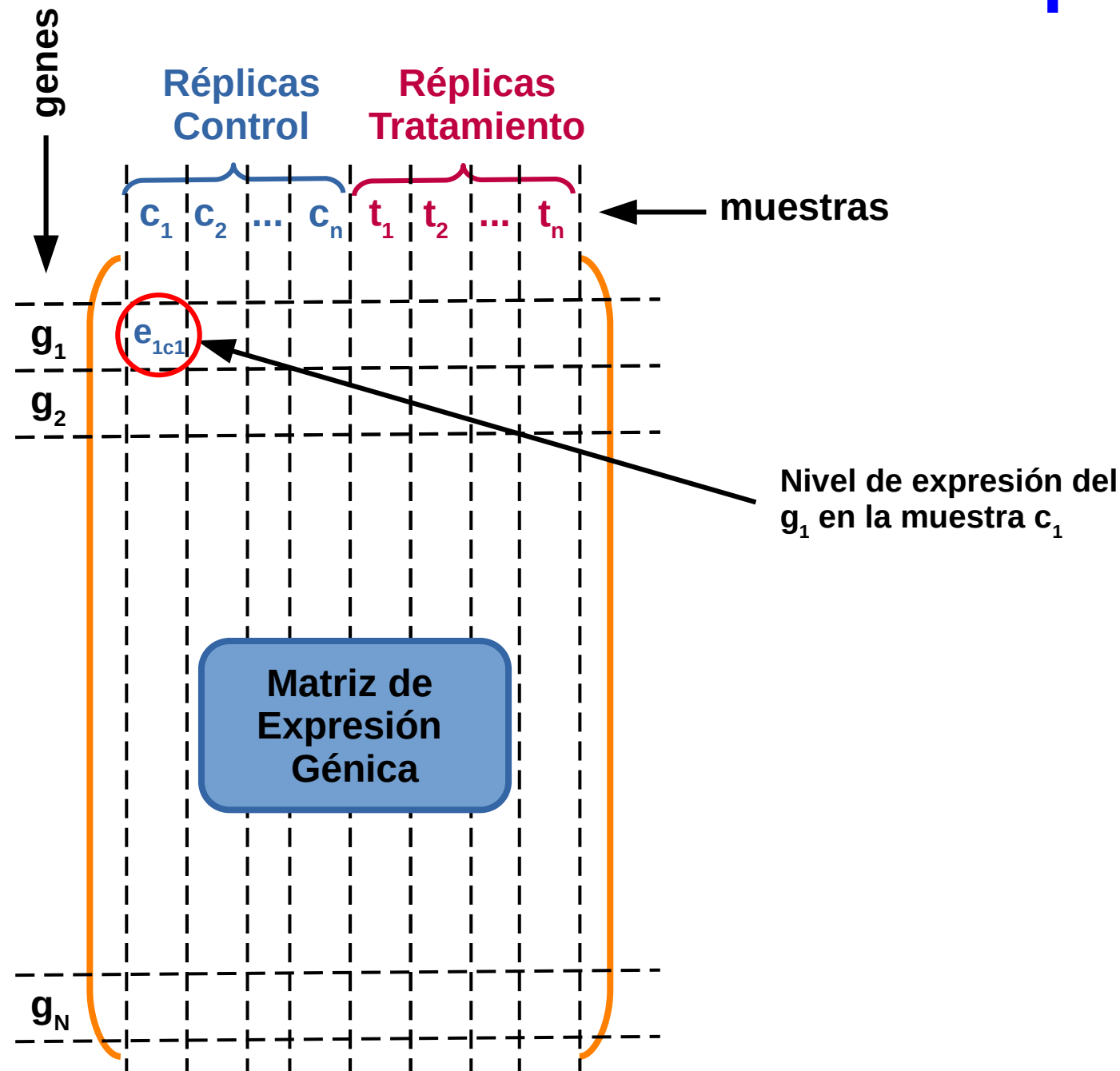
Paso 5.4: Selección de Genes Diferencialmente Expresados



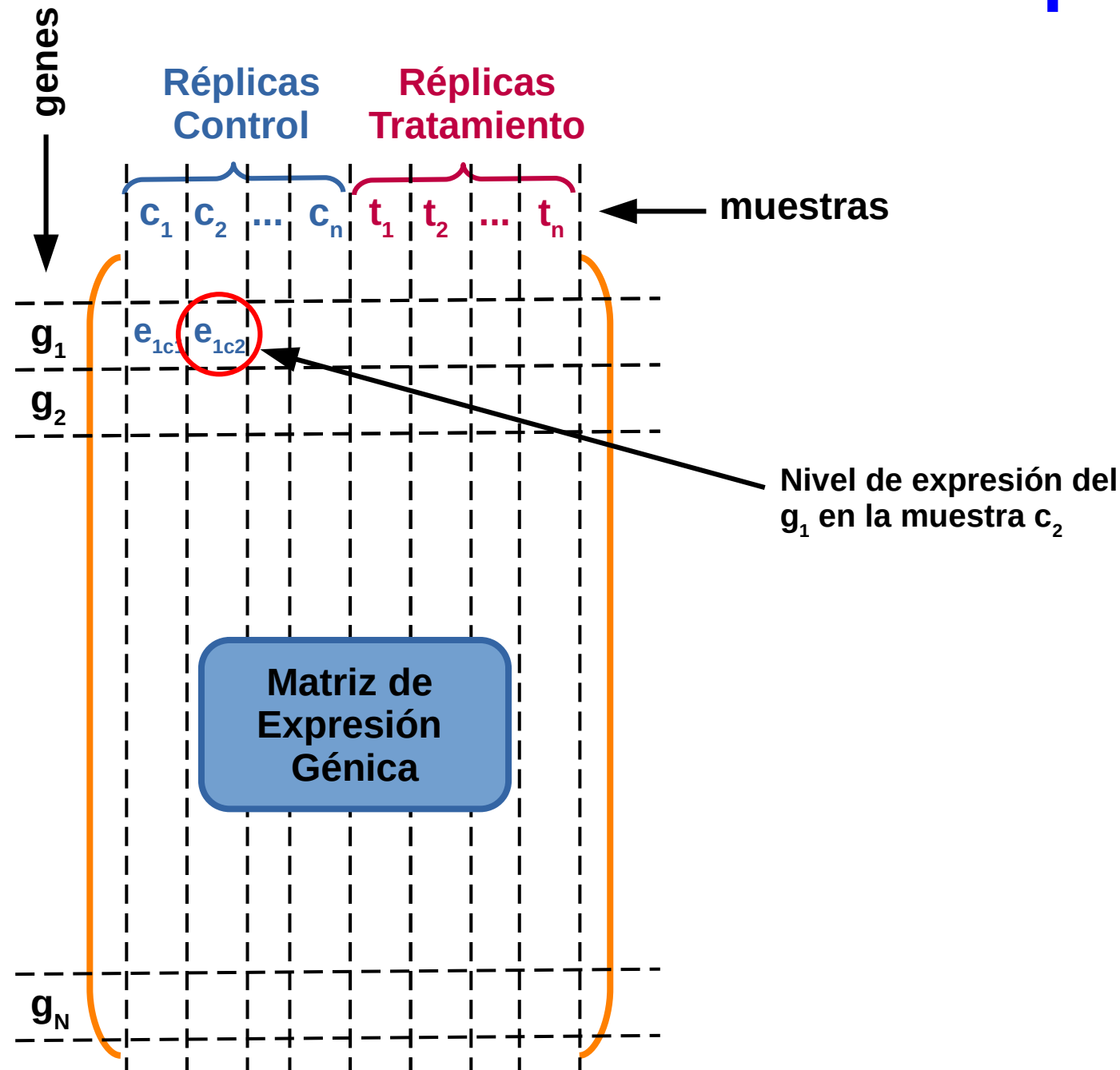
Paso 5.4: Selección de Genes Diferencialmente Expresados



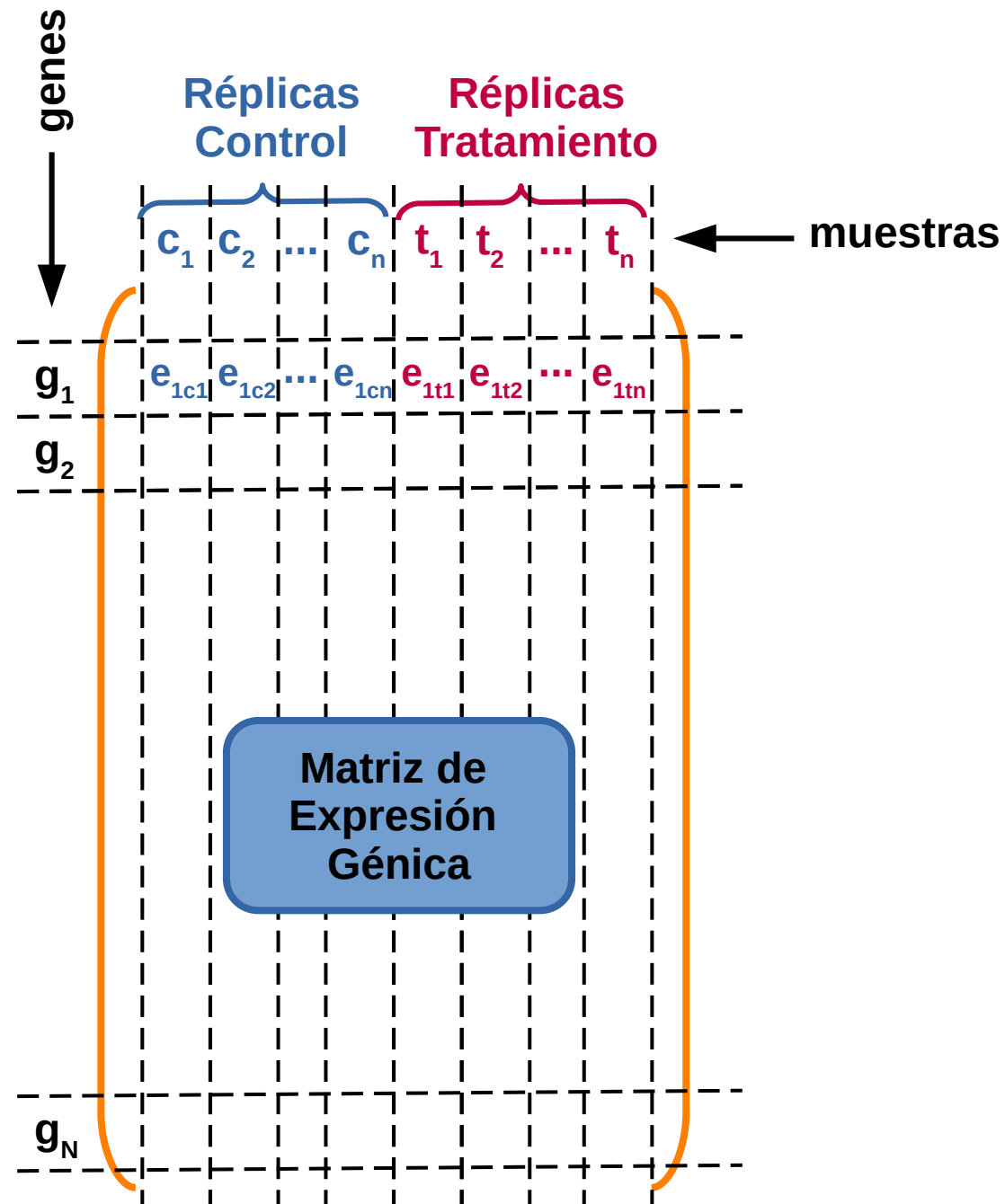
Paso 5.4: Selección de Genes Diferencialmente Expresados



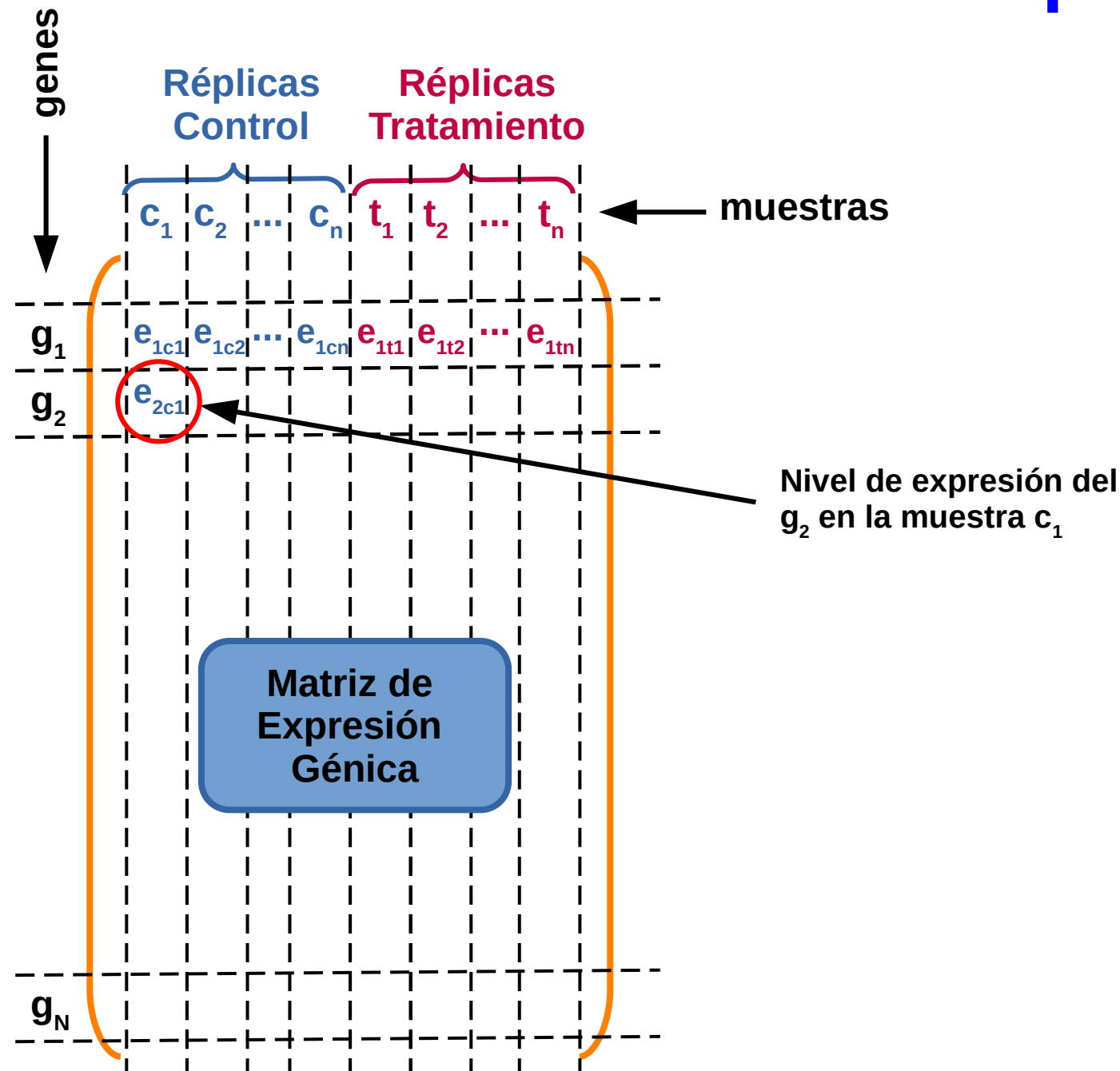
Paso 5.4: Selección de Genes Diferencialmente Expresados



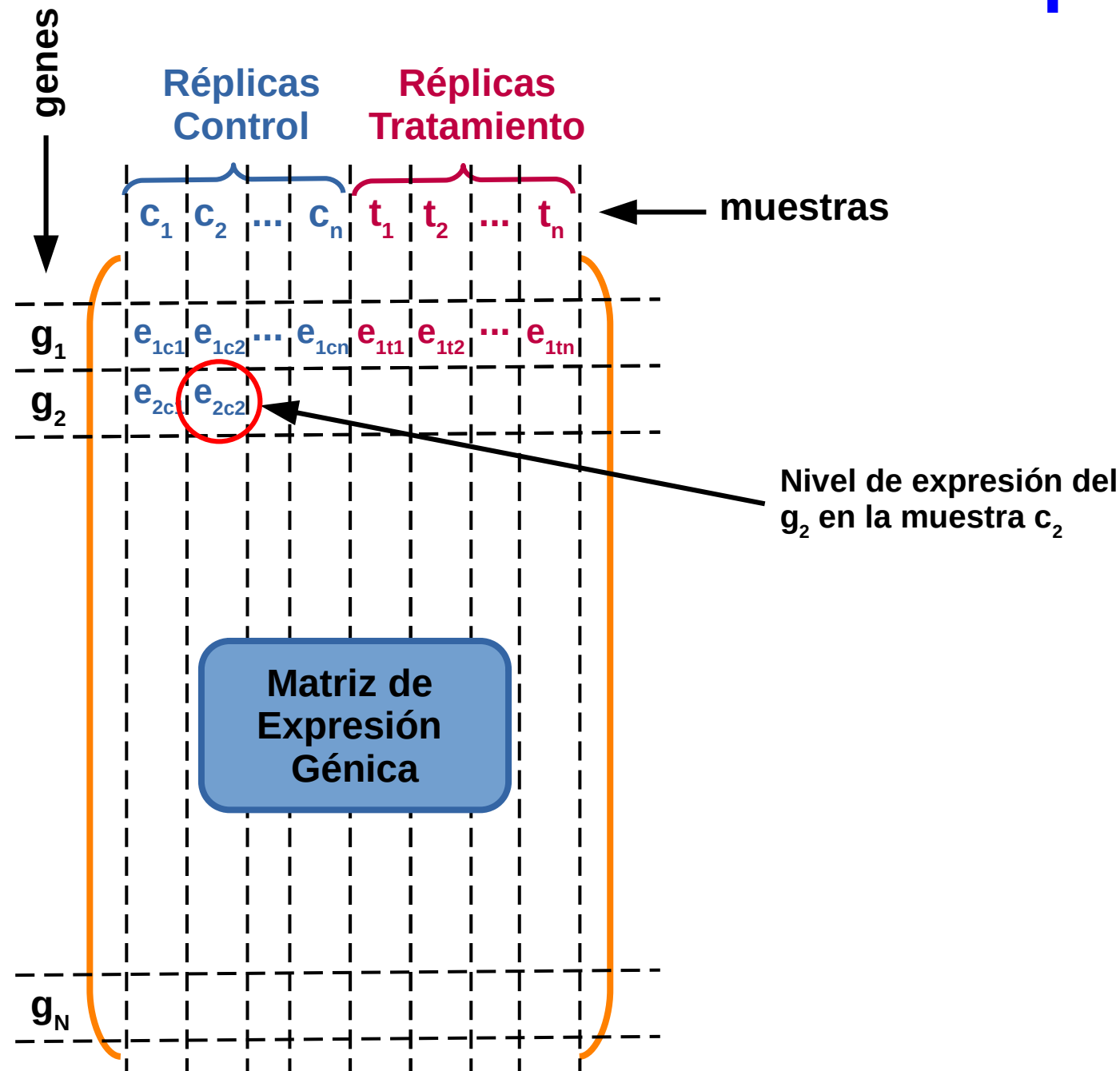
Paso 5.4: Selección de Genes Diferencialmente Expresados



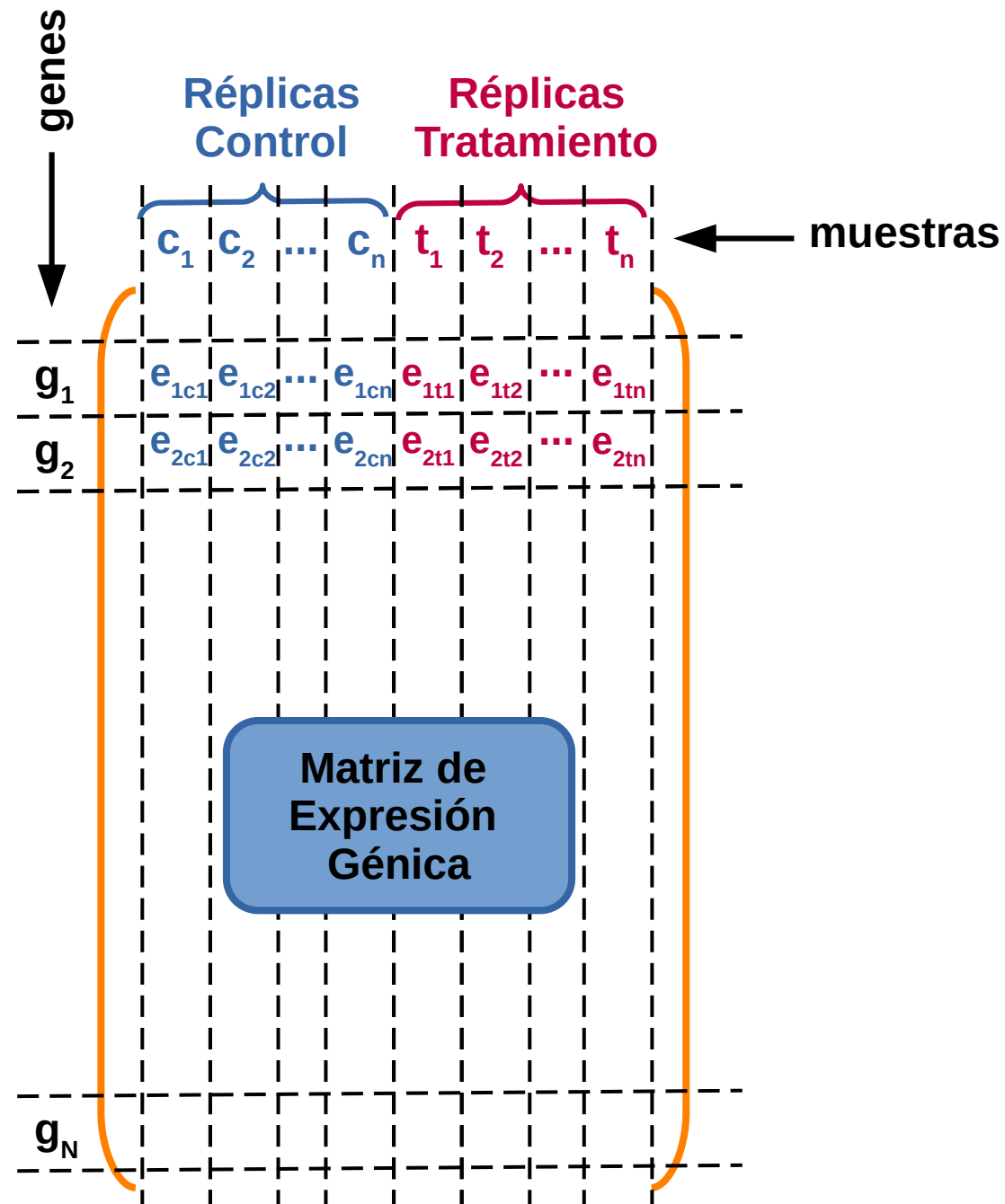
Paso 5.4: Selección de Genes Diferencialmente Expresados



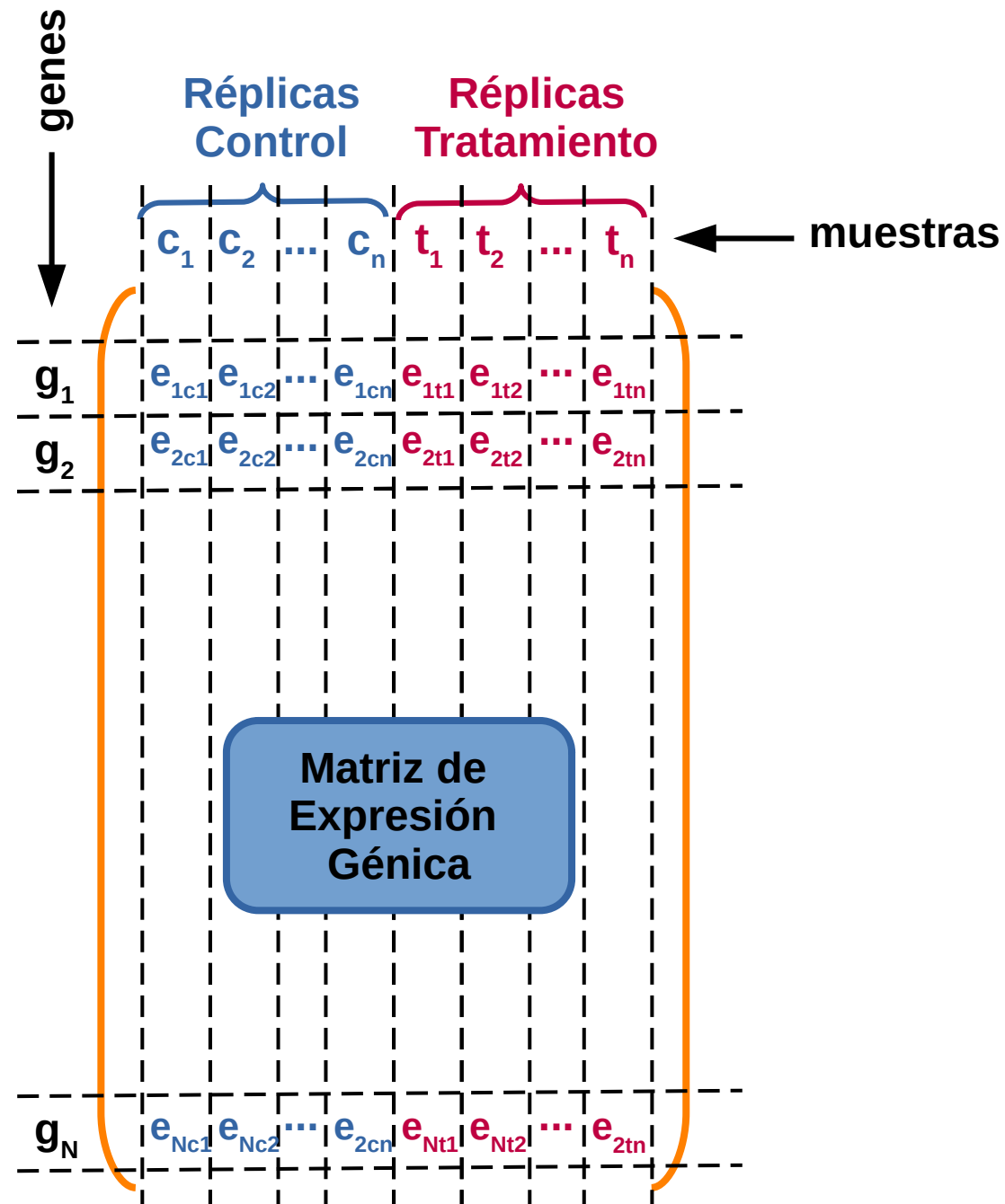
Paso 5.4: Selección de Genes Diferencialmente Expresados



Paso 5.4: Selección de Genes Diferencialmente Expresados

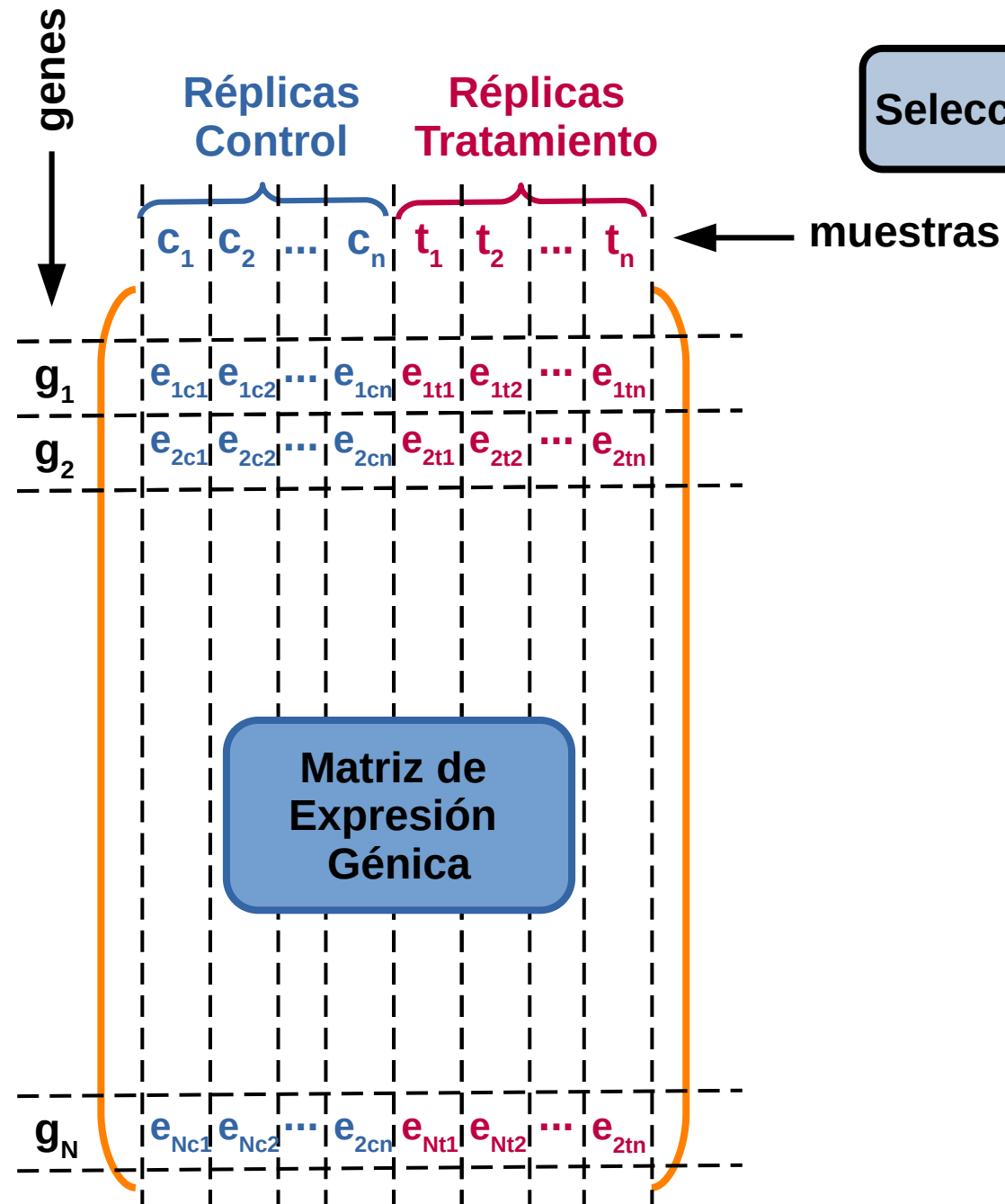


Paso 5.4: Selección de Genes Diferencialmente Expresados



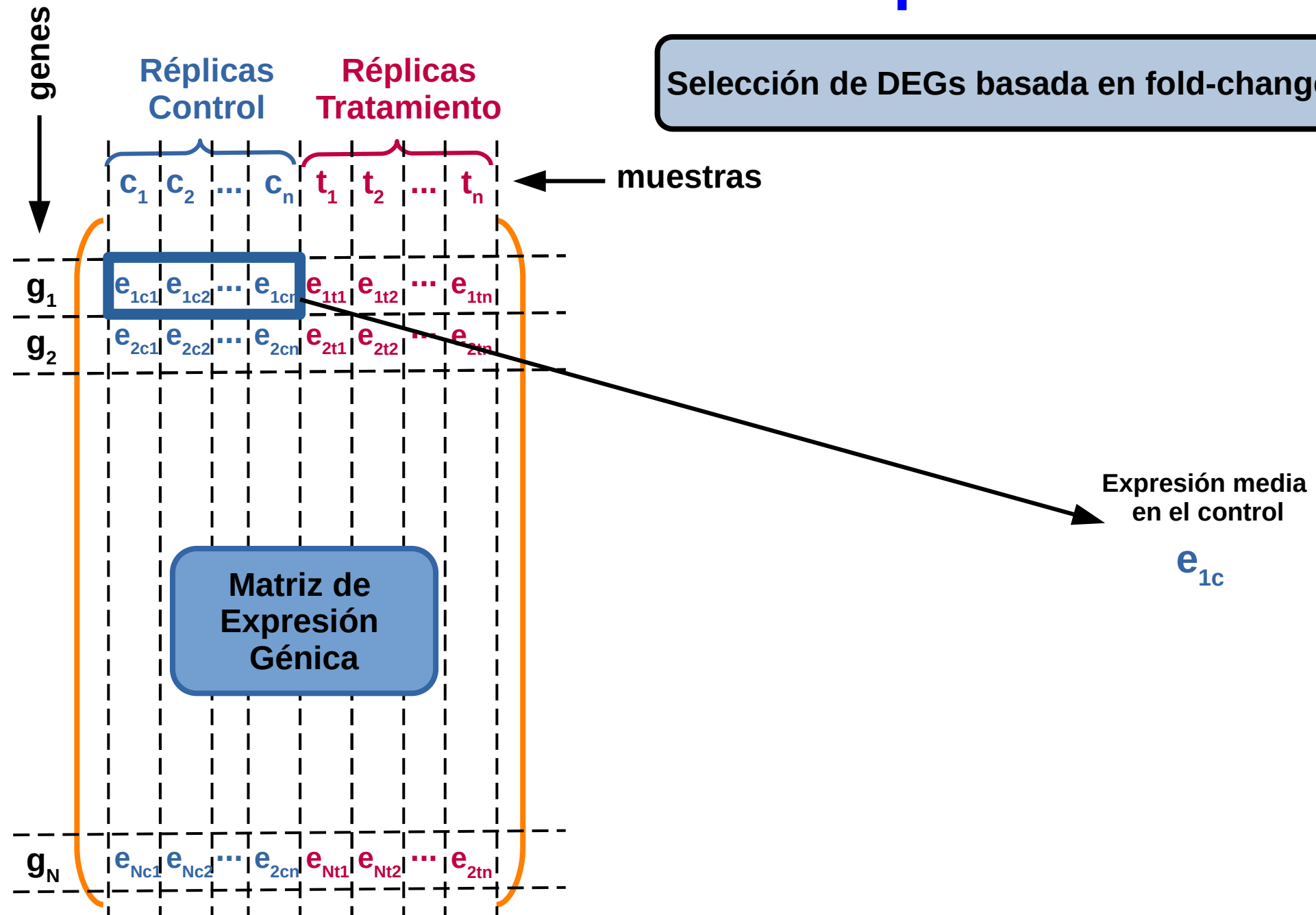
Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en fold-change(FC)



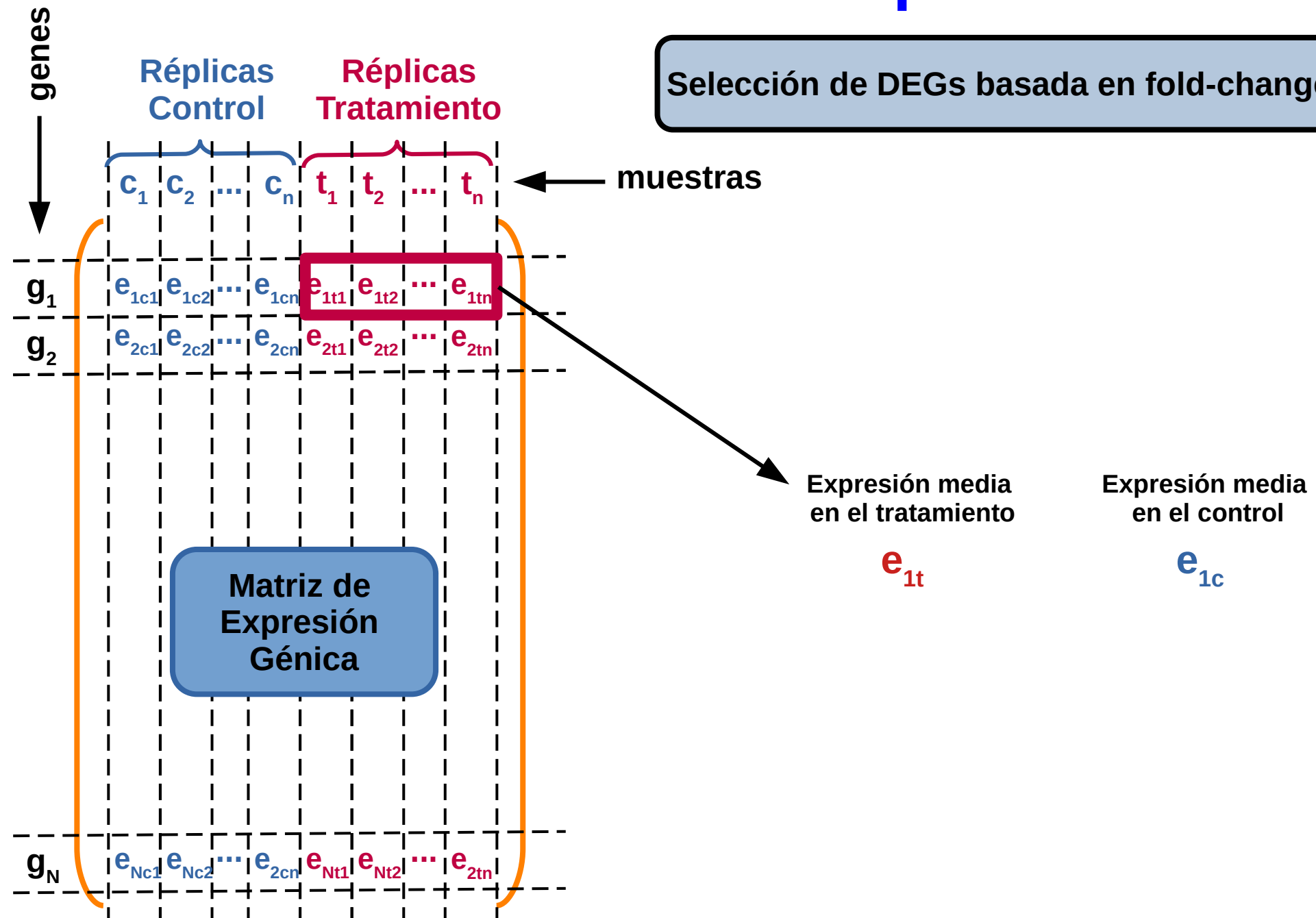
Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en fold-change(FC)



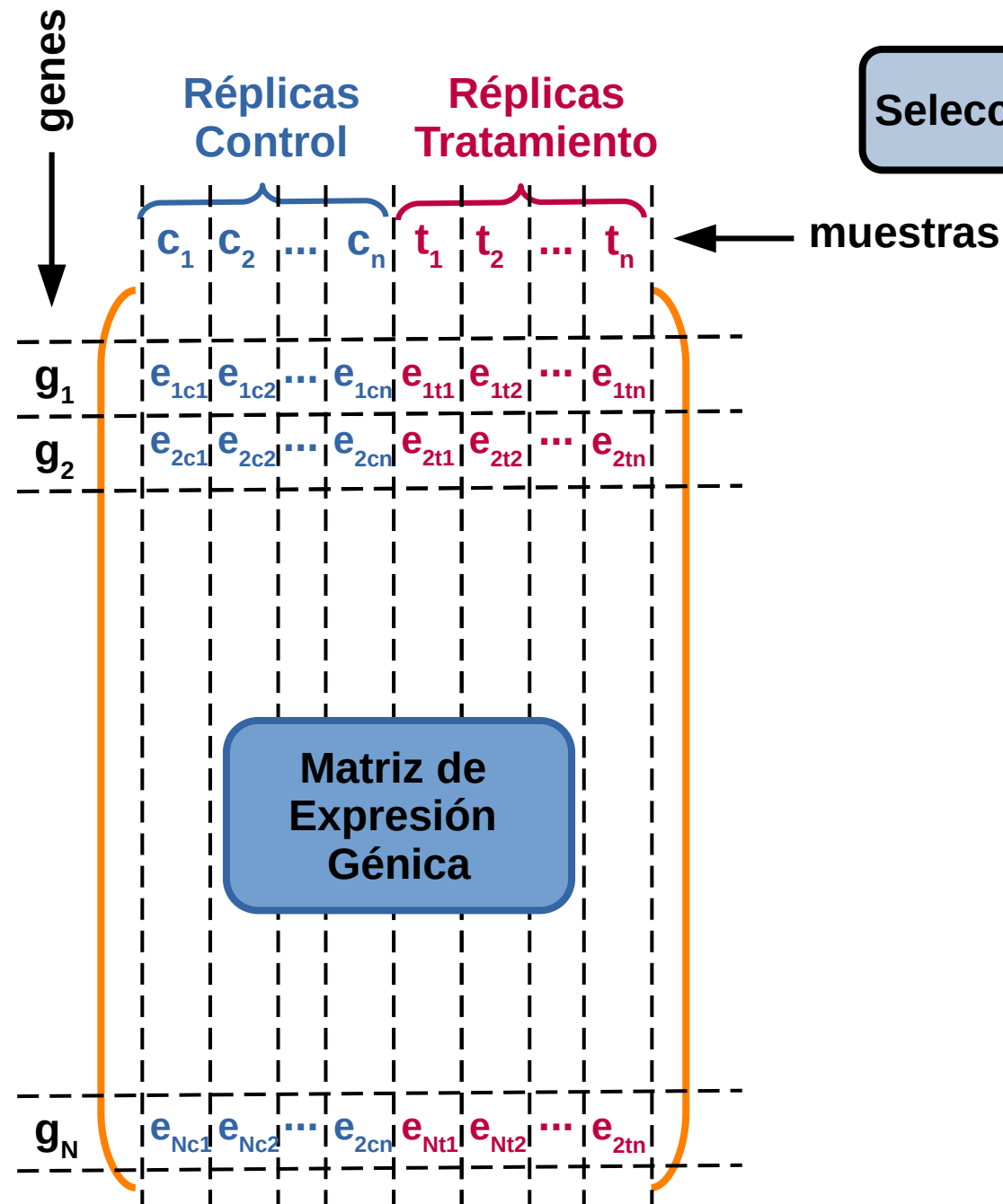
Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en fold-change(FC)



Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en fold-change(FC)



Expresión media
en el tratamiento

e_{1t}

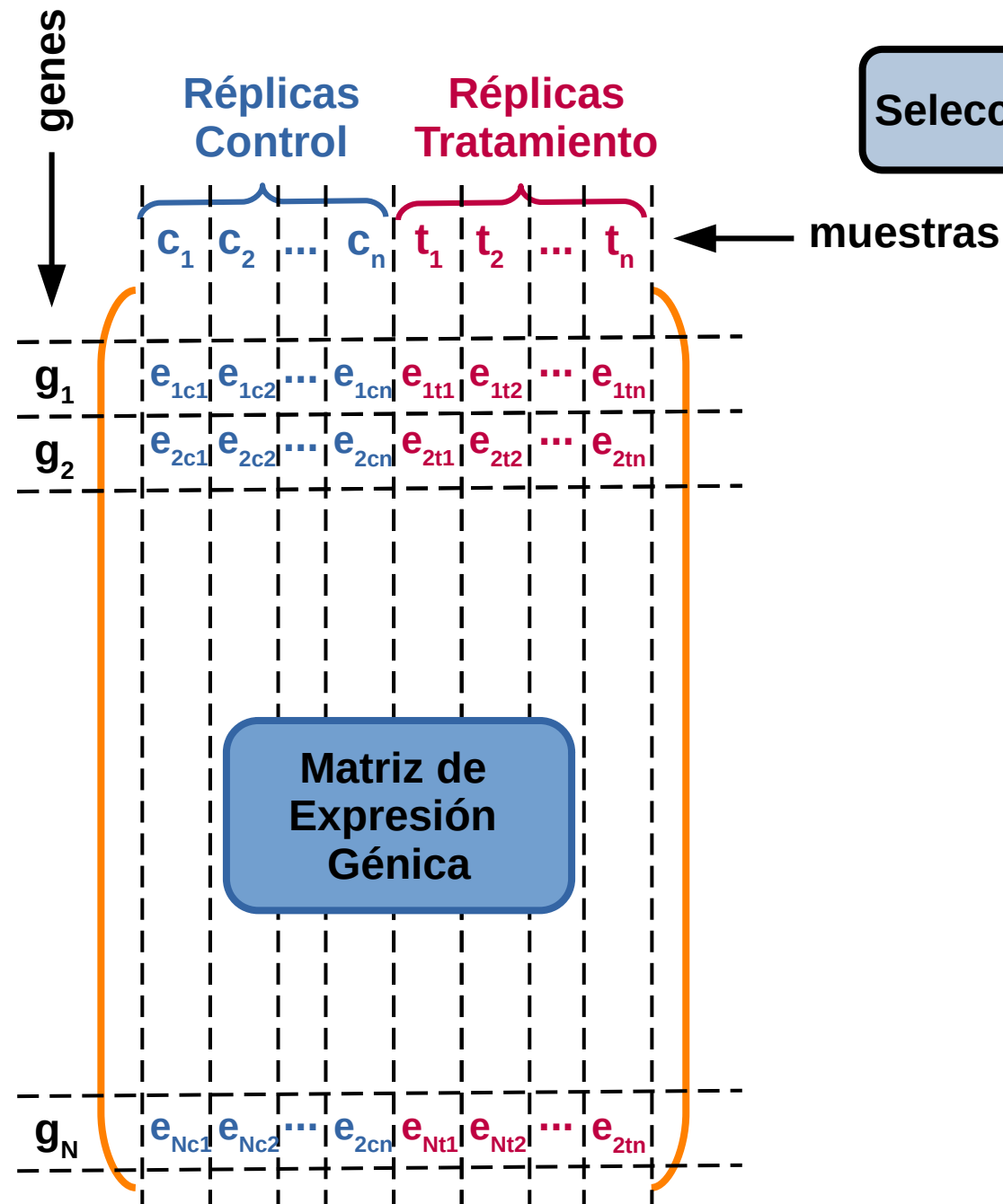
Expresión media
en el control

e_{1c}

$$fc_1 = \frac{e_{1t}}{e_{1c}}$$

Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en fold-change(FC)



Expresión media
en el tratamiento

e_{1t}

Expresión media
en el control

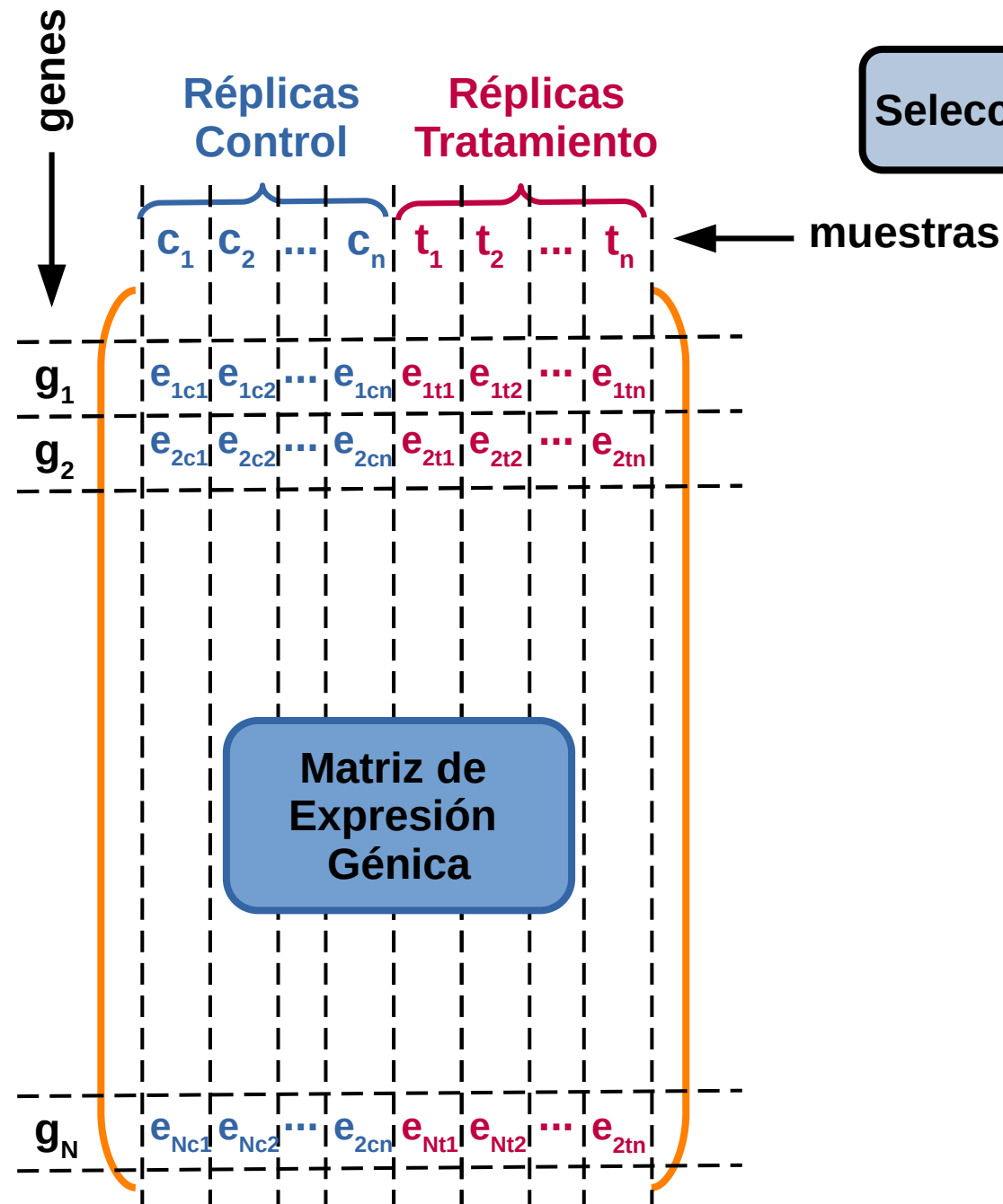
e_{1c}

Datos transformados
por \log_2

$$\log_2\left(\frac{a}{b}\right) = \log_2(a) - \log_2(b)$$

Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en fold-change(FC)



Expresión media
en el tratamiento

e_{1t}

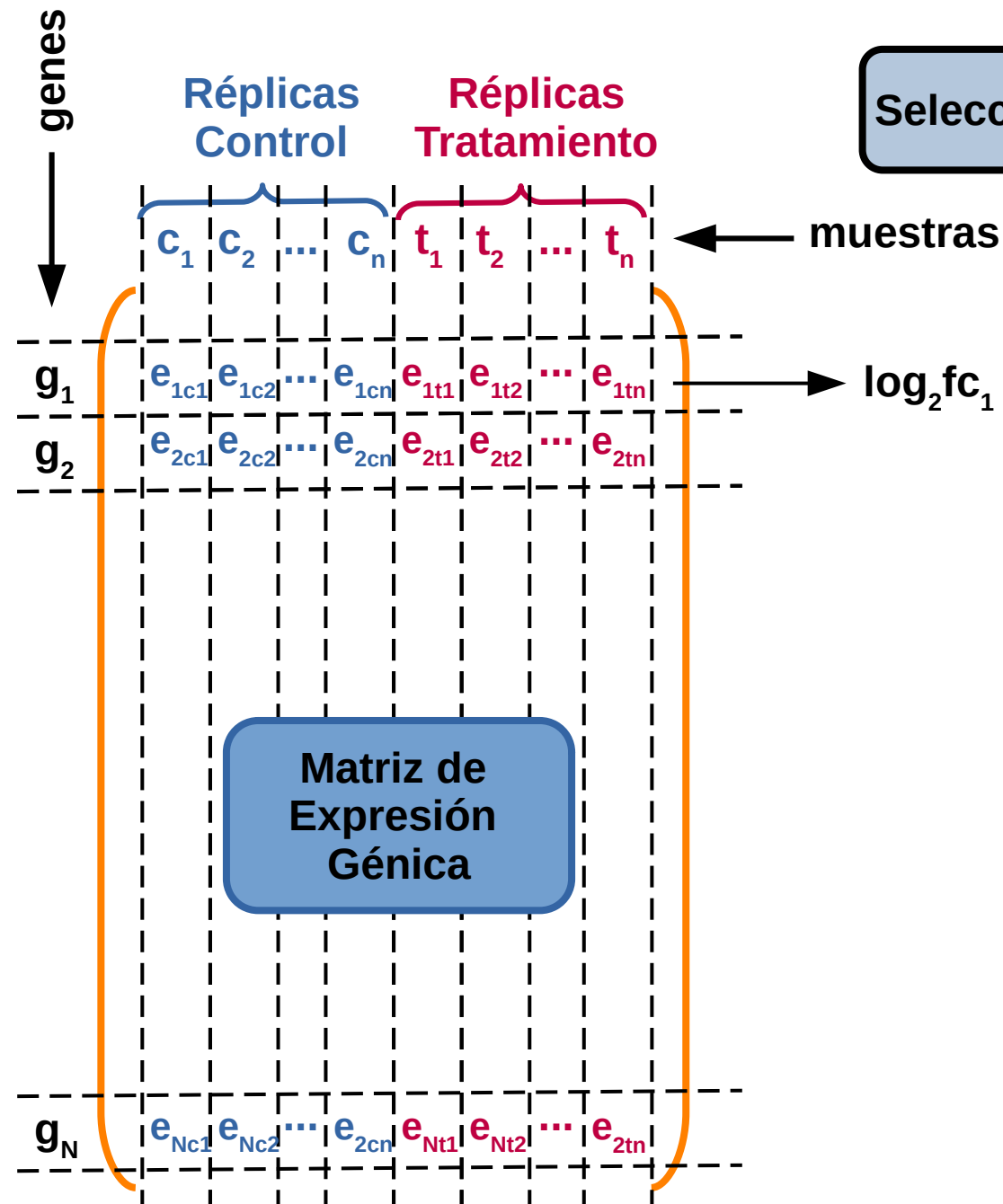
Expresión media
en el control

e_{1c}

$$\log_2 fc_1 = e_{1t} - e_{1c}$$

Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en fold-change(FC)



Expresión media
en el tratamiento

e_{1t}

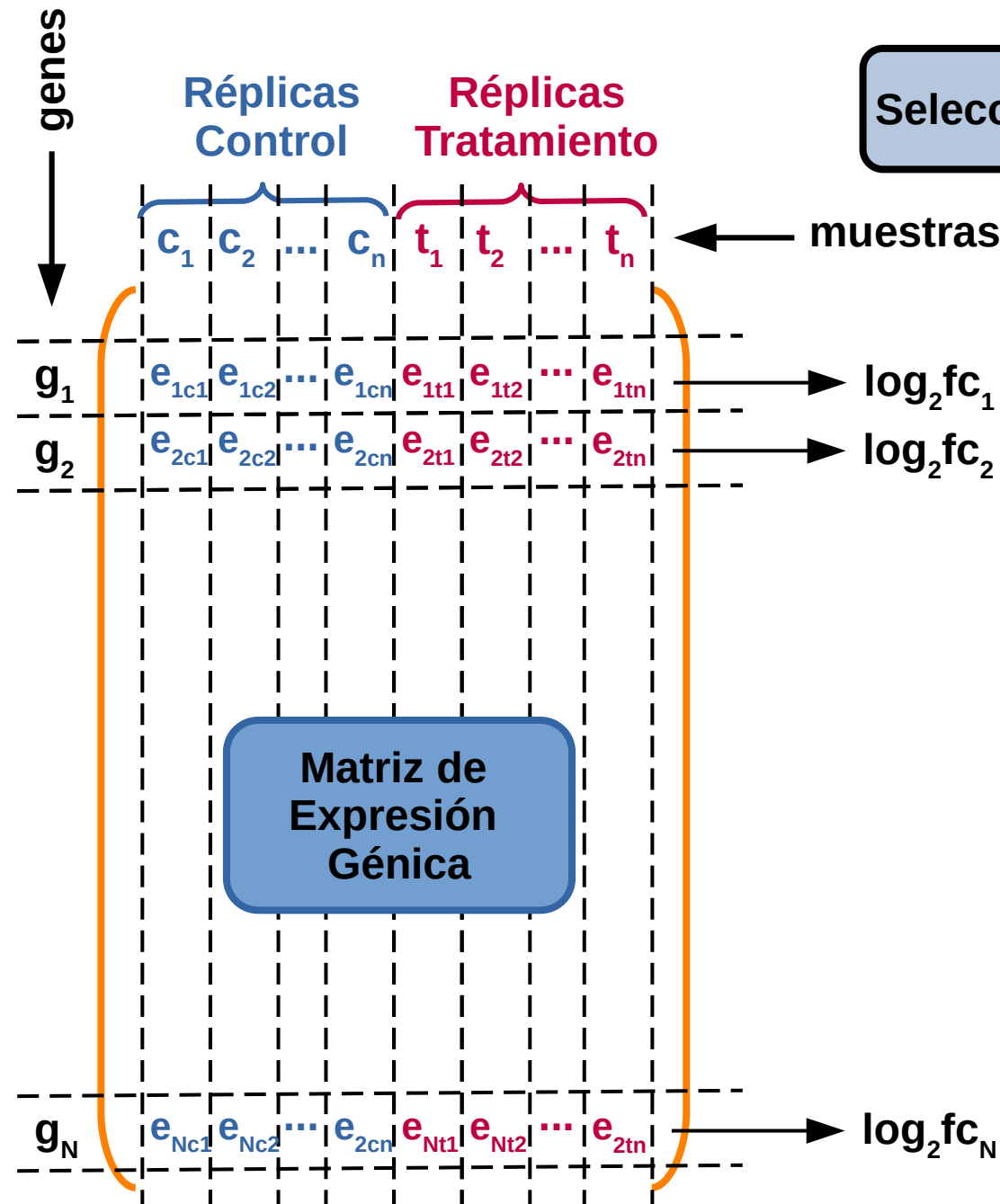
Expresión media
en el control

e_{1c}

$$\log_2 fc_1 = e_{1t} - e_{1c}$$

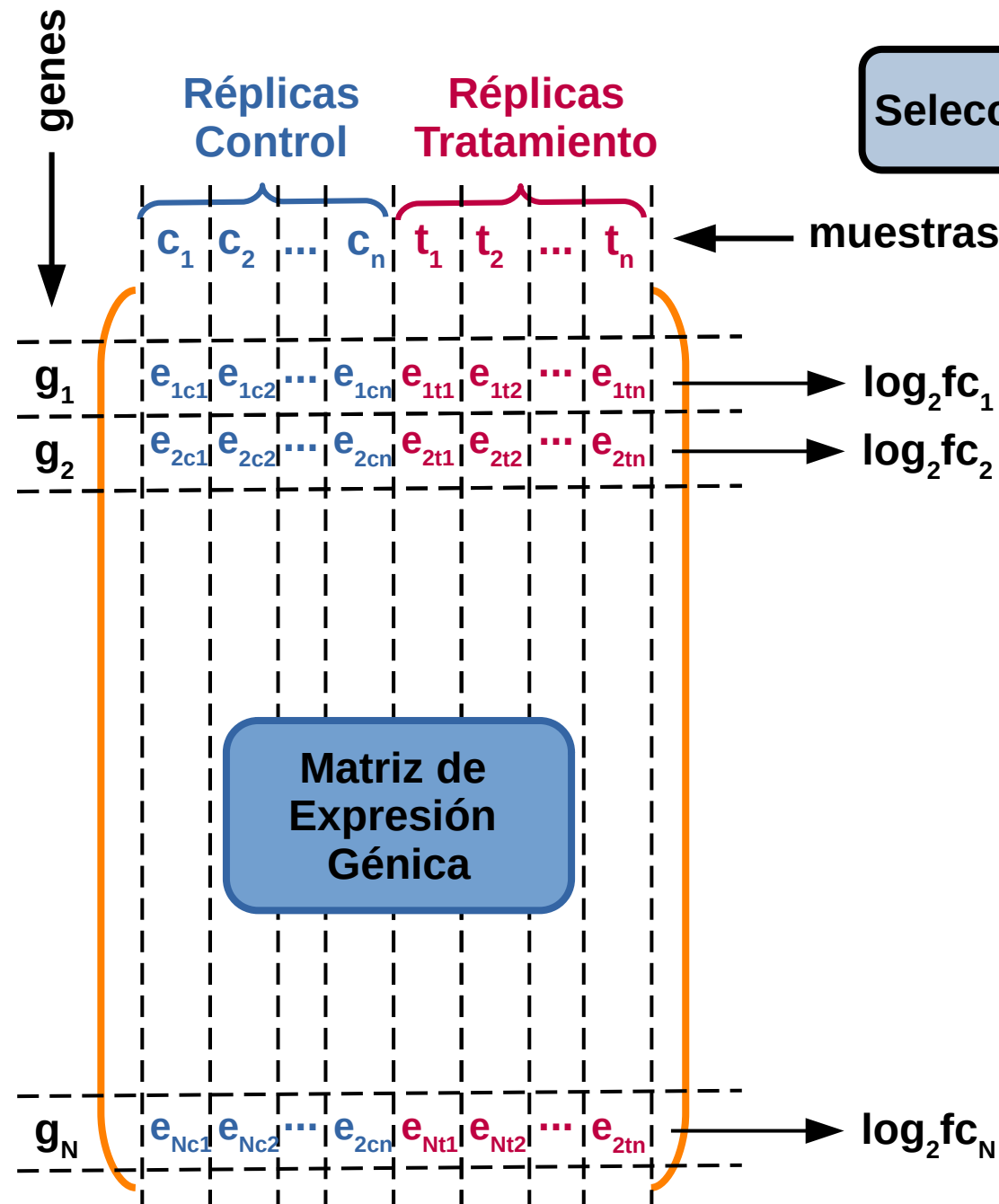
Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en fold-change(FC)



Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en fold-change(FC)



Se fija un umbral para el fold-change, típicamente 2, 4 u 8, y se determinan:

Genes activados

$$fc_k > 2, 4, 8 \rightarrow \log_2 fc_k > 1, 2, 3$$

Genes reprimidos

$$fc_k < \frac{1}{2}, \frac{1}{4}, \frac{1}{8} \rightarrow \log_2 fc_k < -1, -2, -3$$

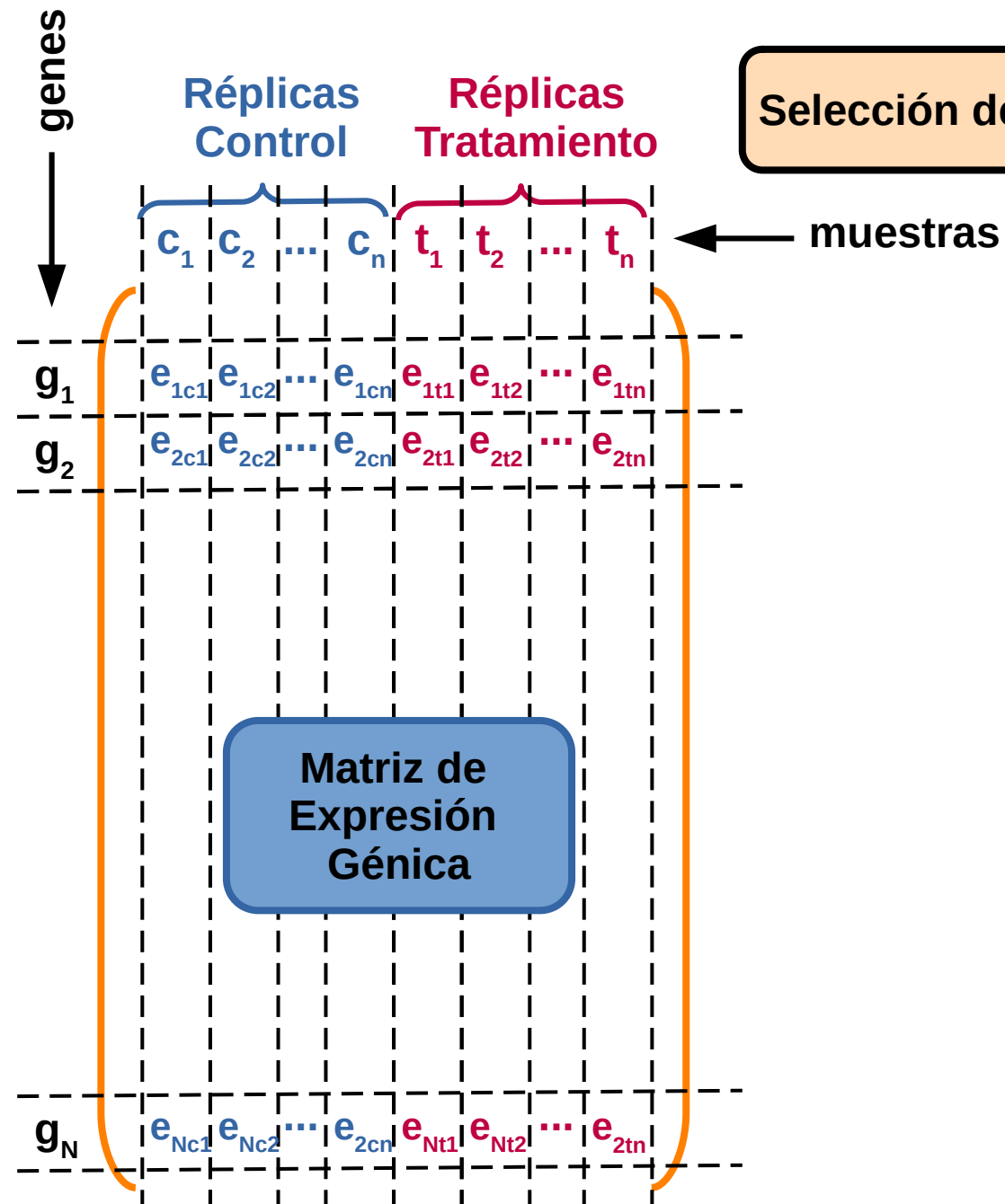
Paso 5.4: Selección de Genes Diferencialmente Expresados

Cuando se comparan los transcriptomas de dos genotipos diferentes o de un mismo genotipo bajo distintas condiciones existen diversos métodos para determinar genes expresados de forma diferencial o *differentially expressed genes* (DEGs) en inglés:

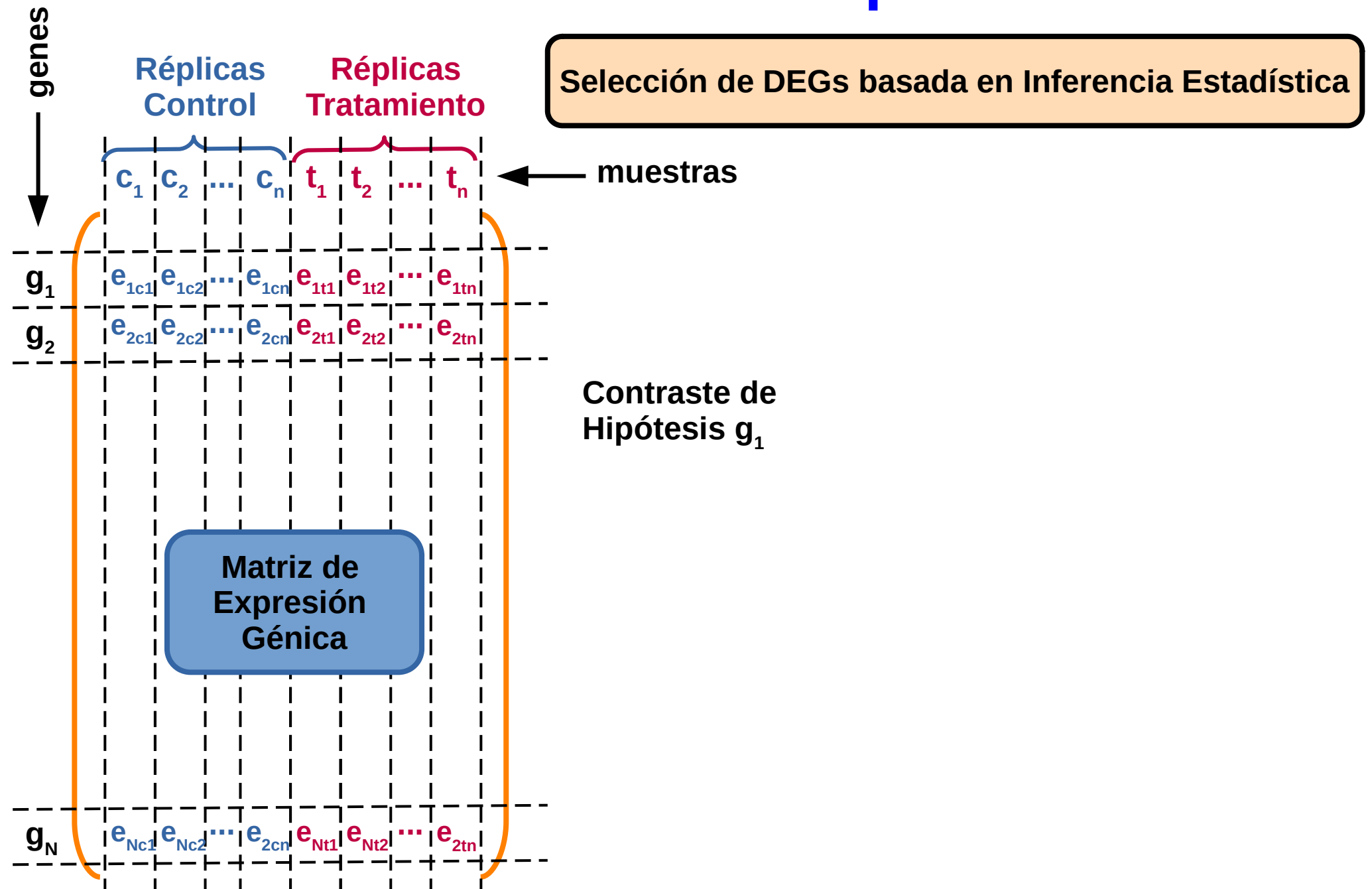
- **Método basado en el *fold-change*** (factor de proporcionalidad): Se fija un umbral para el fold-change típicamente 2, 4 u 8 que en log2 corresponde a 1, 2 ó 3. Los DEGs son aquellos que incrementan (o decrementan) su expresión por encima de dicho umbral (por debajo de menos dicho umbral). Este método es biológicamente interpretable de forma directa y no requiere un alto número de réplicas biológicas. Se aplica especialmente a estudios con organismo modelos donde no son necesarias muchas réplicas.
- **Método basado en inferencia estadística:** Para aplicar este método es necesario tener un alto número de réplicas biológicas. Para cada gen y para cada pareja de genotipos/condiciones a comparar se formula un contraste de hipótesis sobre igualdad de medias. Normalmente este contraste de hipótesis utiliza un estadístico similar a la t-student. Se fija un nivel de significancia y se calcula el correspondiente p-valor (y p-valor corregido para el testeo múltiple o q-valor). Si dicho p-valor (o q-valor) es menor que el nivel de significancia se asume que el correspondiente gen se expresa de forma diferencial en los genotipos/condiciones estudiadas.
- **Combinación de los dos anteriores métodos**

Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en Inferencia Estadística

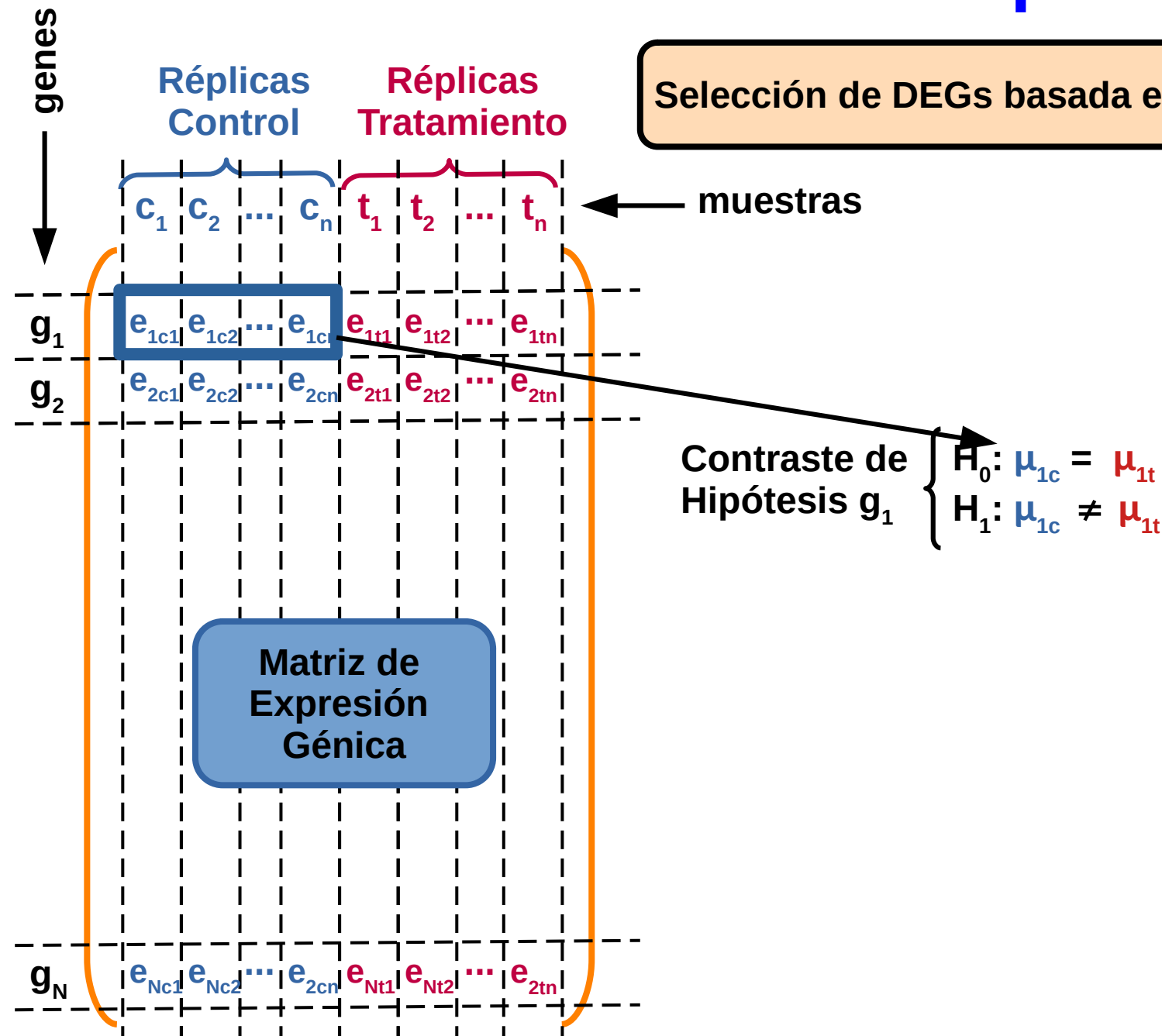


Paso 5.4: Selección de Genes Diferencialmente Expresados



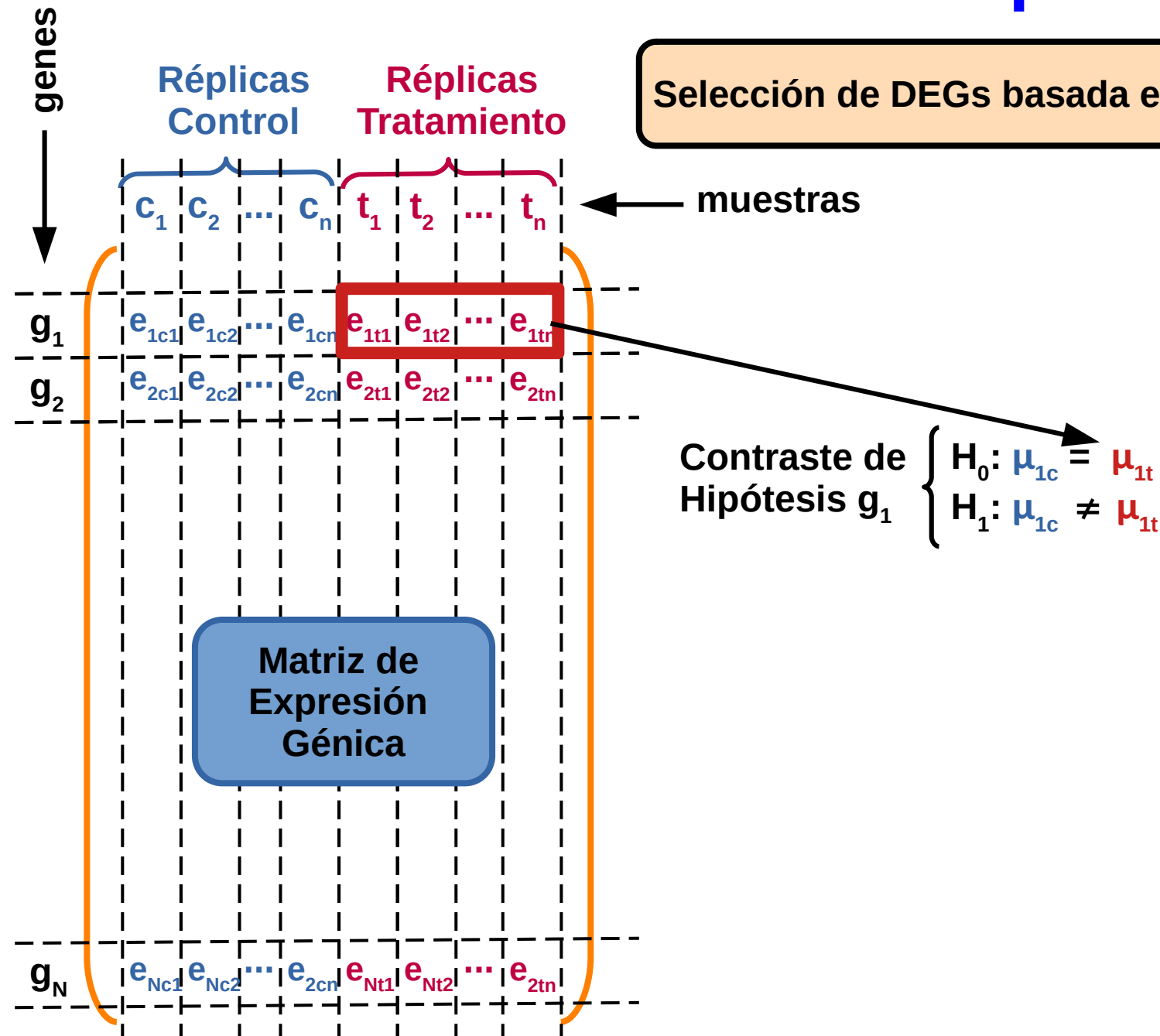
Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en Inferencia Estadística



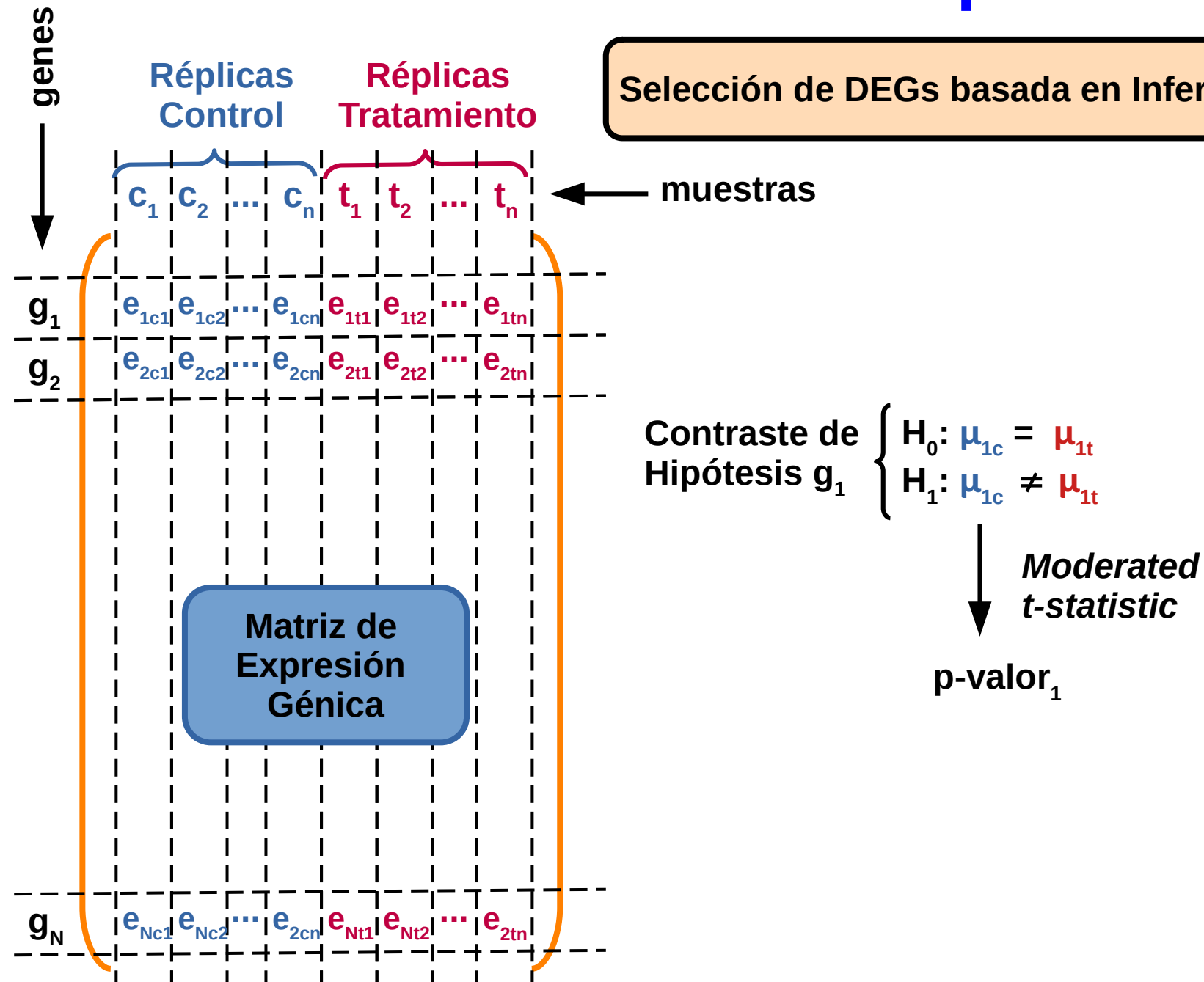
Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en Inferencia Estadística



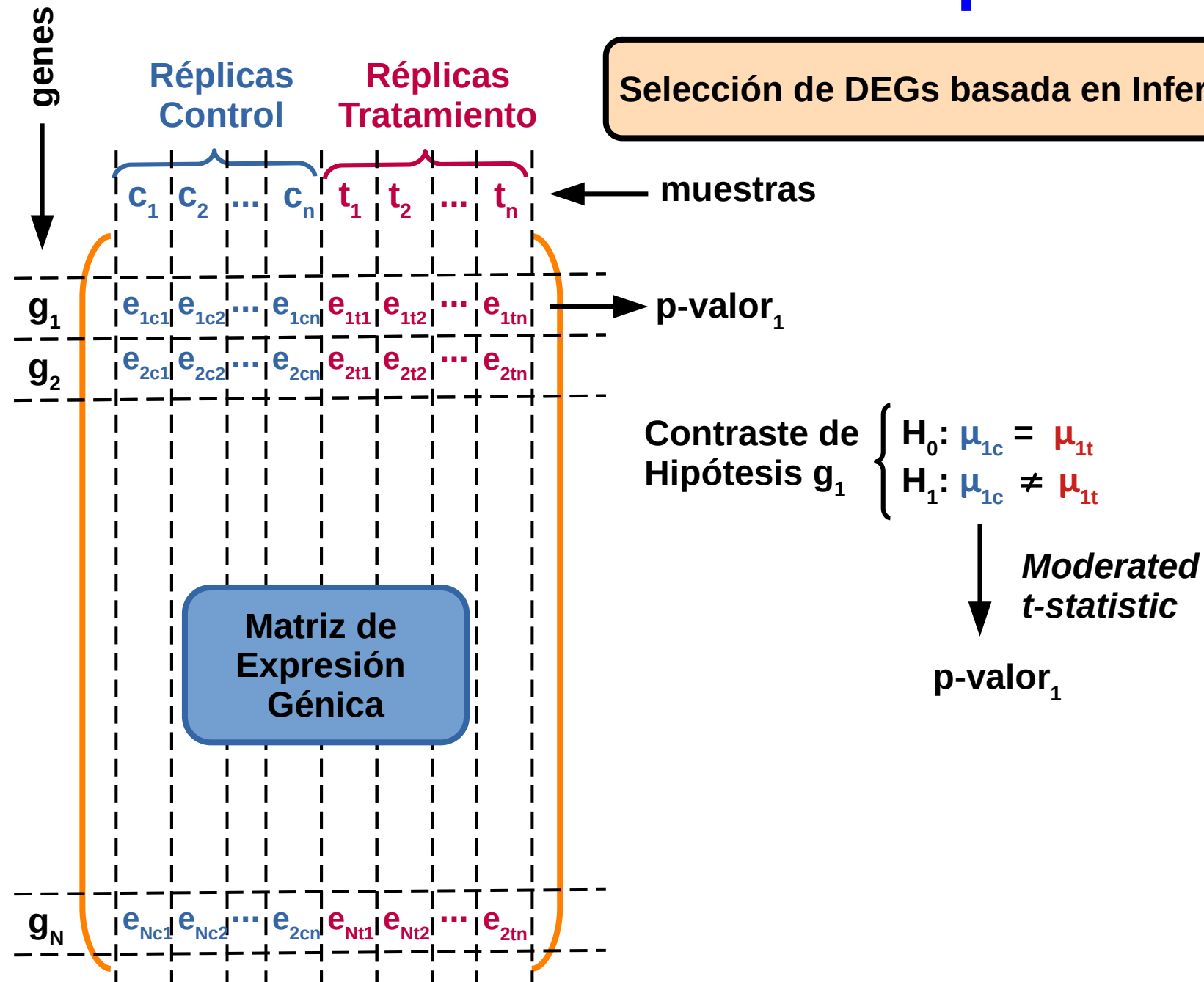
Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en Inferencia Estadística



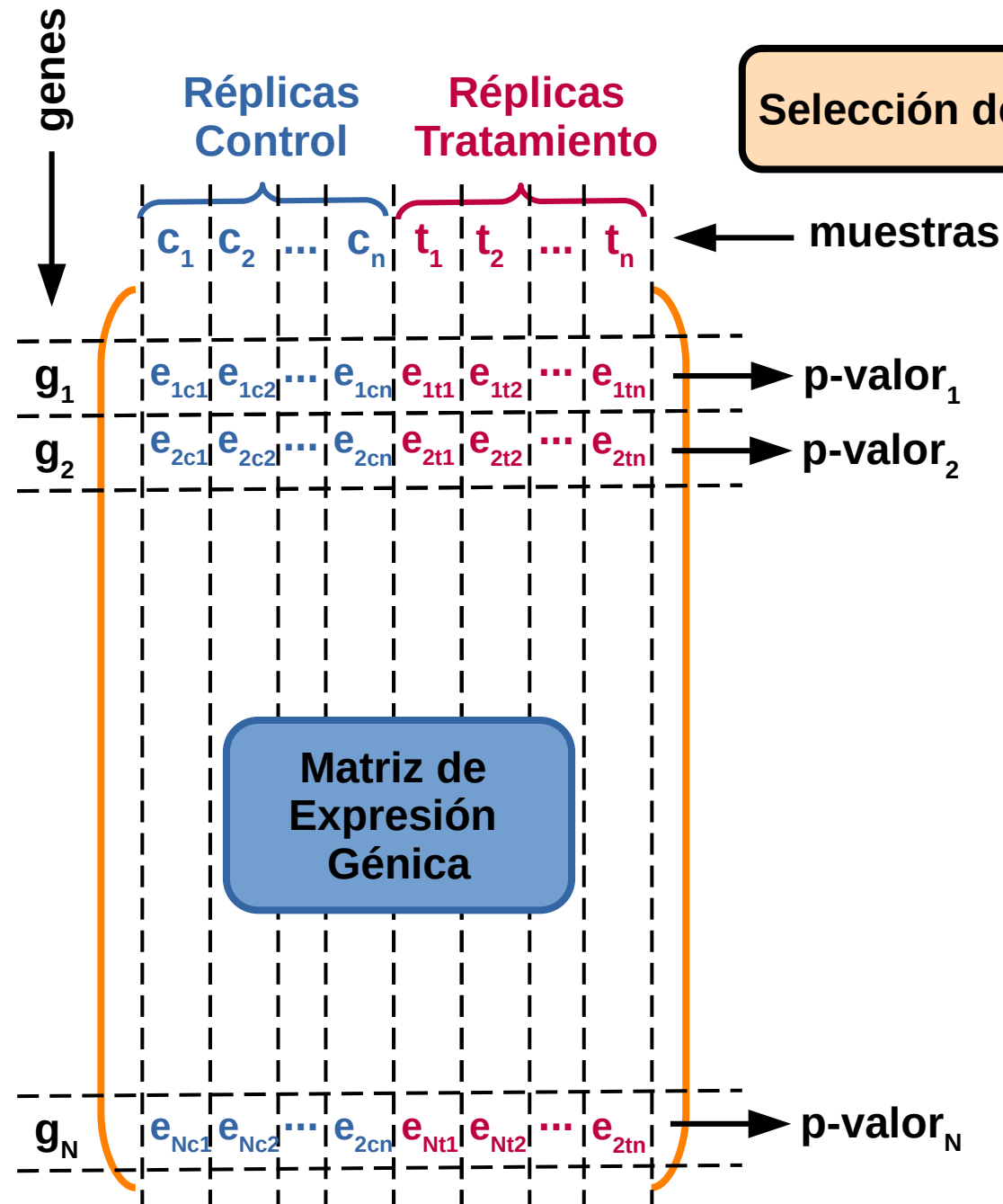
Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en Inferencia Estadística

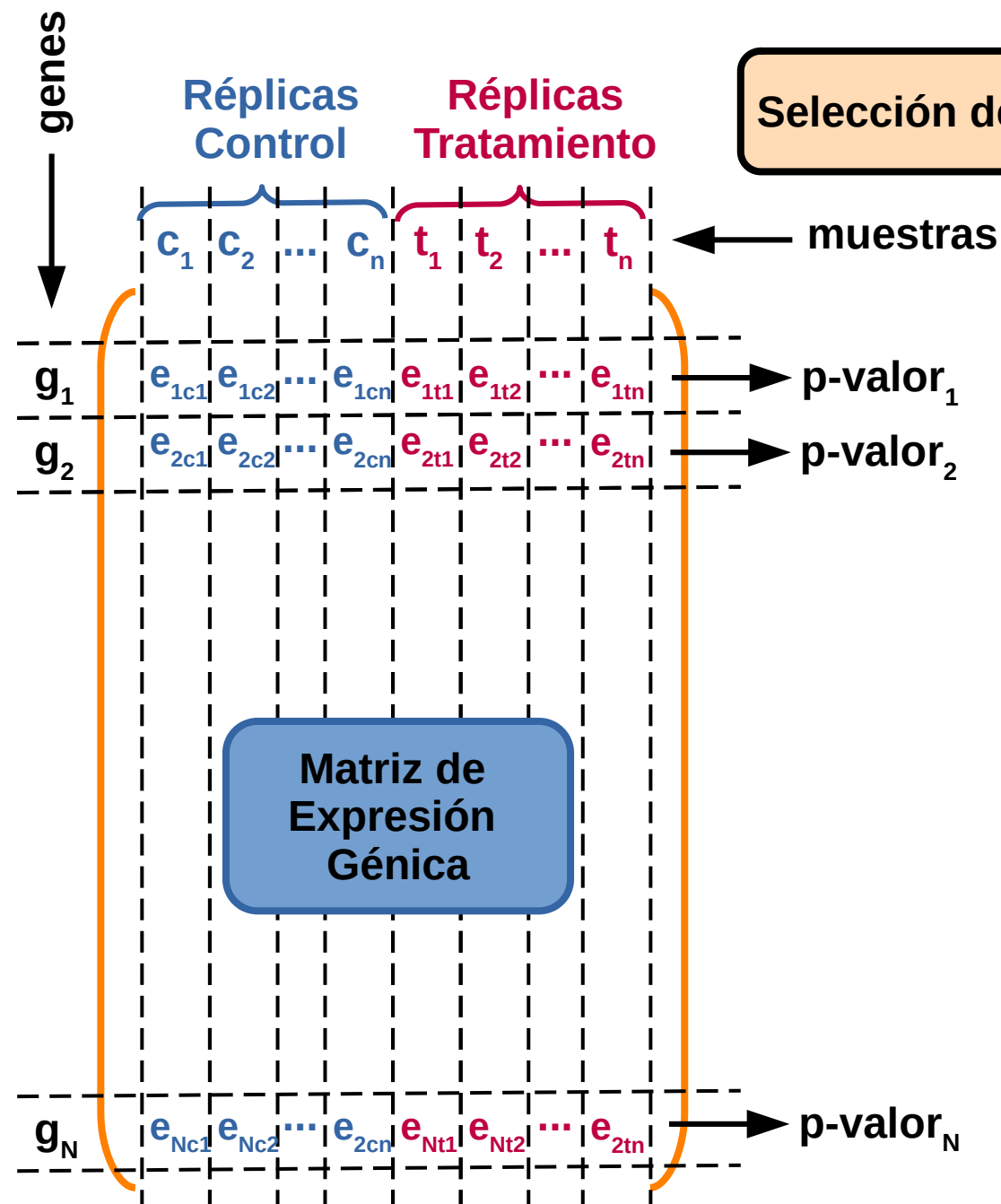


Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en Inferencia Estadística



Paso 5.4: Selección de Genes Diferencialmente Expresados



Selección de DEGs basada en Inferencia Estadística

Problemas con el testeo múltiple:

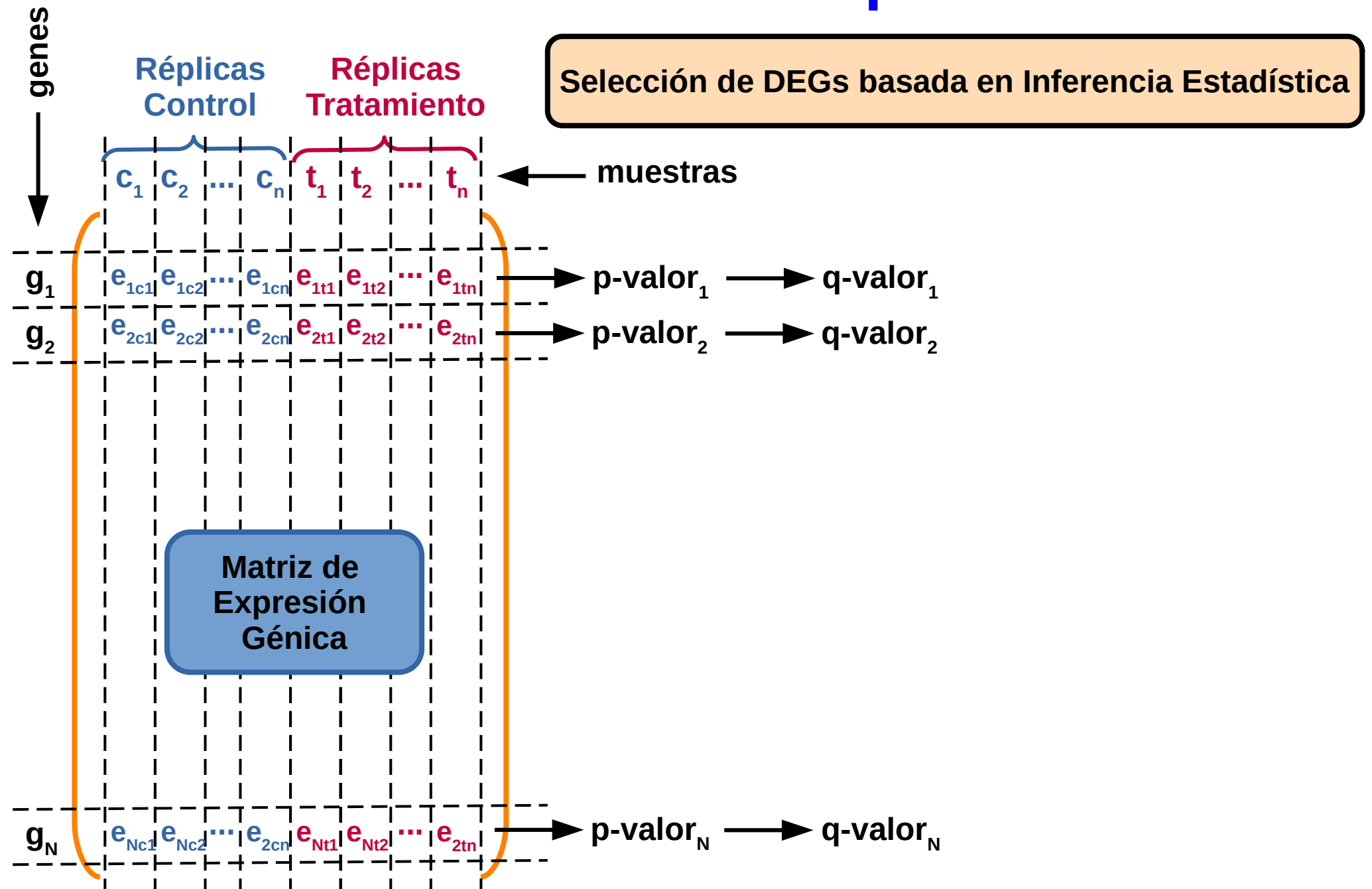
Si fijamos como nivel de significancia 0.01.

No se realiza un único testeo donde la probabilidad de FP sería 0.01 sino N testeos conduciendo a una tasa de FP $0.01 \cdot N$

Por ejemplo para $N=22810$ tendríamos el número no asumible de 230 FP.

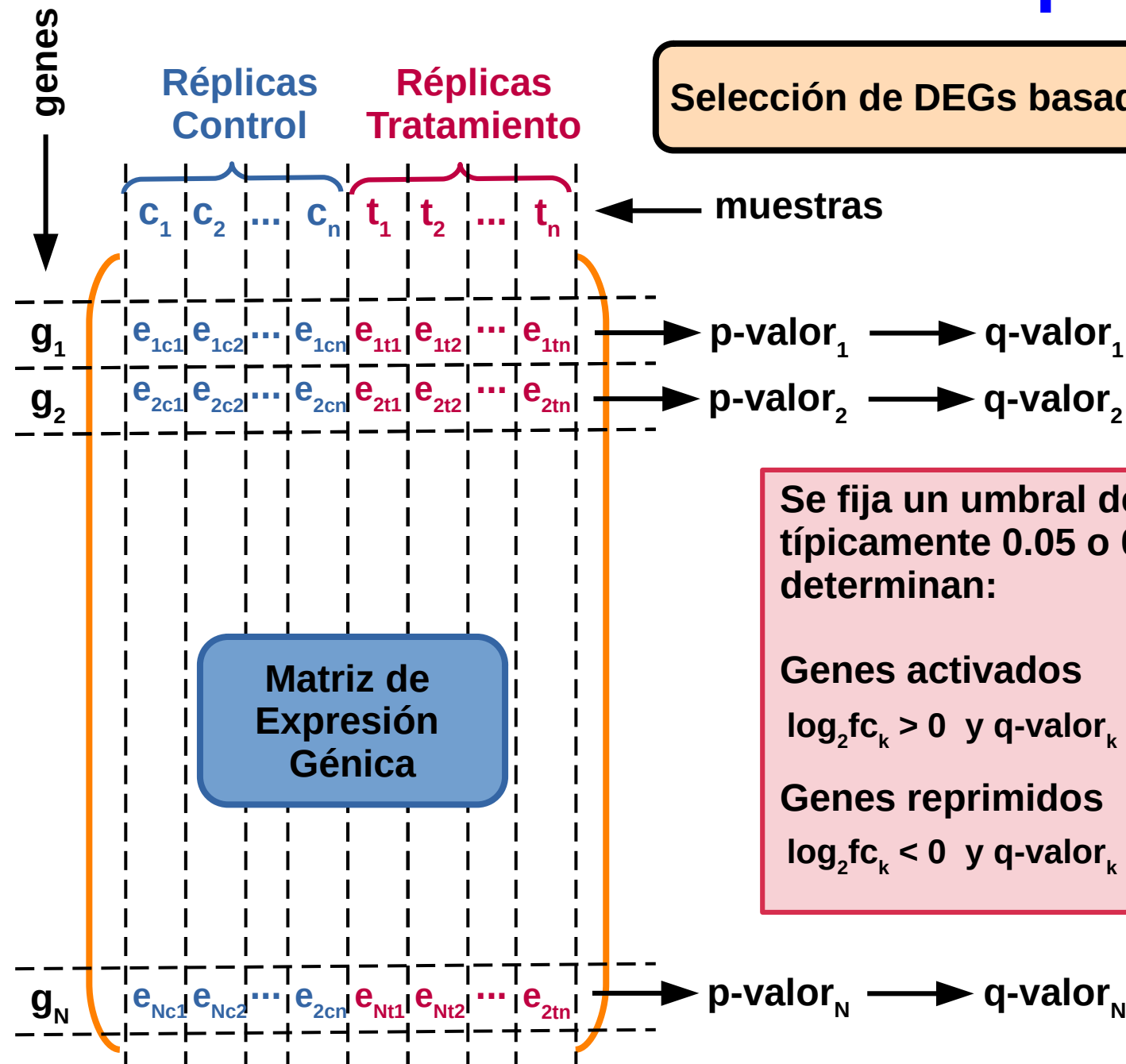
Es necesario corregir el p-valor por el testeo múltiple con técnicas como Benjamini-Hochberg para generar un q-valor o FDR (False Discory Rate).

Paso 5.4: Selección de Genes Diferencialmente Expresados



Paso 5.4: Selección de Genes Diferencialmente Expresados

Selección de DEGs basada en Inferencia Estadística



Se fija un umbral de significancia típicamente 0.05 o 0.01 y se determinan:

Genes activados

$$\log_2 fc_k > 0 \text{ y } q\text{-valor}_k < 0.05, 0.01$$

Genes reprimidos

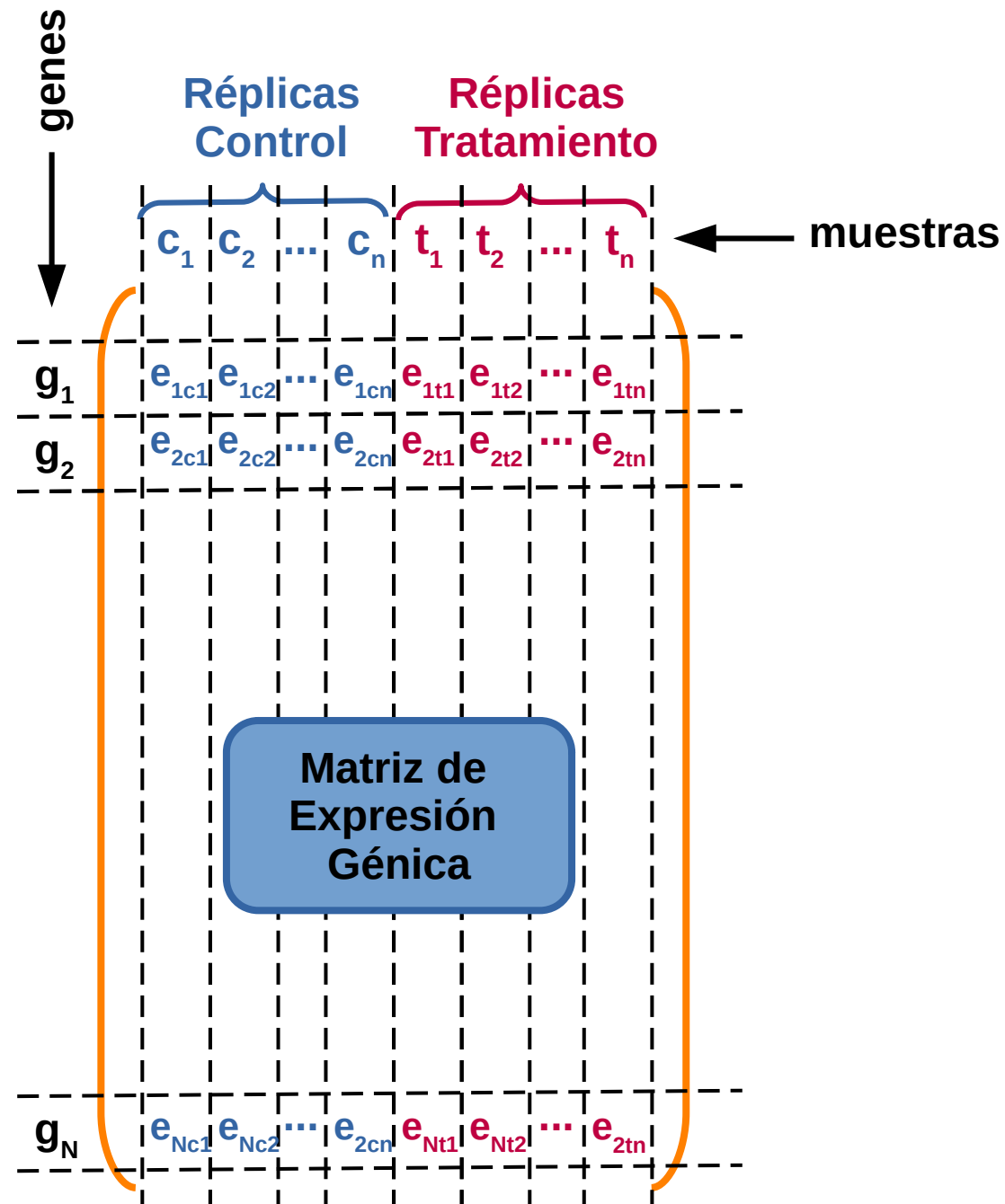
$$\log_2 fc_k < 0 \text{ y } q\text{-valor}_k < 0.05, 0.01$$

Paso 5.4: Selección de Genes Diferencialmente Expresados

Cuando se comparan los transcriptomas de dos genotipos diferentes o de un mismo genotipo bajo distintas condiciones existen diversos métodos para determinar genes expresados de forma diferencial o *differentially expressed genes* (DEGs) en inglés:

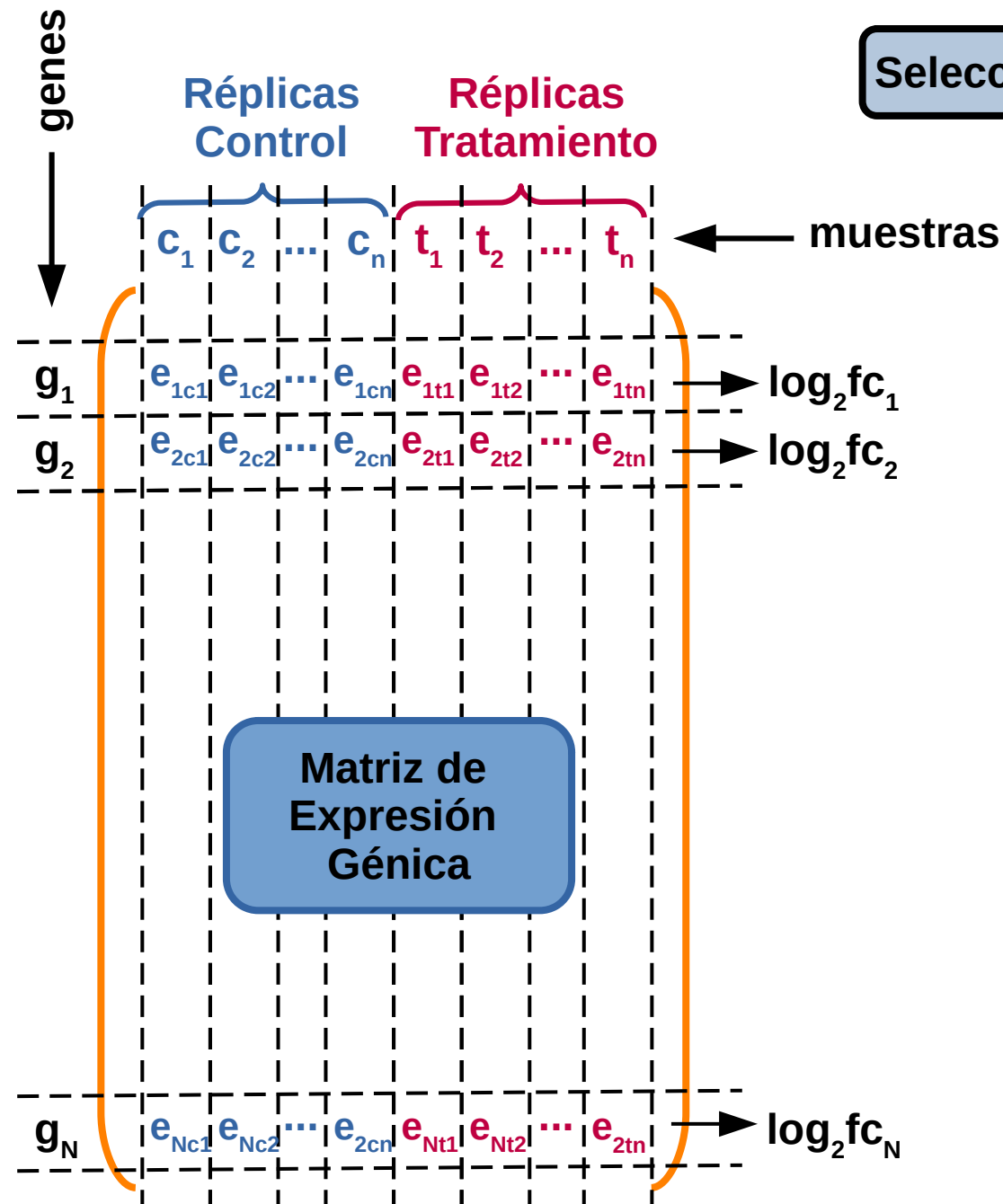
- **Método basado en el *fold-change*** (factor de proporcionalidad): Se fija un umbral para el fold-change típicamente 2, 4 u 8 que en log2 corresponde a 1, 2 ó 3. Los DEGs son aquellos que incrementan (o decrementan) su expresión por encima de dicho umbral (por debajo de menos dicho umbral). Este método es biológicamente interpretable de forma directa y no requiere un alto número de réplicas biológicas. Se aplica especialmente a estudios con organismo modelos donde no son necesarias muchas réplicas.
 - **Método basado en inferencia estadística:** Para aplicar este método es necesario tener un alto número de réplicas biológicas. Para cada gen y para cada pareja de genotipos/condiciones a comparar se formula un contraste de hipótesis sobre igualdad de medias. Normalmente este contraste de hipótesis utiliza un estadístico similar a la t-student. Se fija un nivel de significancia y se calcula el correspondiente p-valor (y p-valor corregido para el testeo múltiple o q-valor). Si dicho p-valor (o q-valor) es menor que el nivel de significancia se asume que el correspondiente gen se expresa de forma diferencial en los genotipos/condiciones estudiadas.
- **Combinación de los dos anteriores métodos**

Paso 5.4: Selección de Genes Diferencialmente Expresados

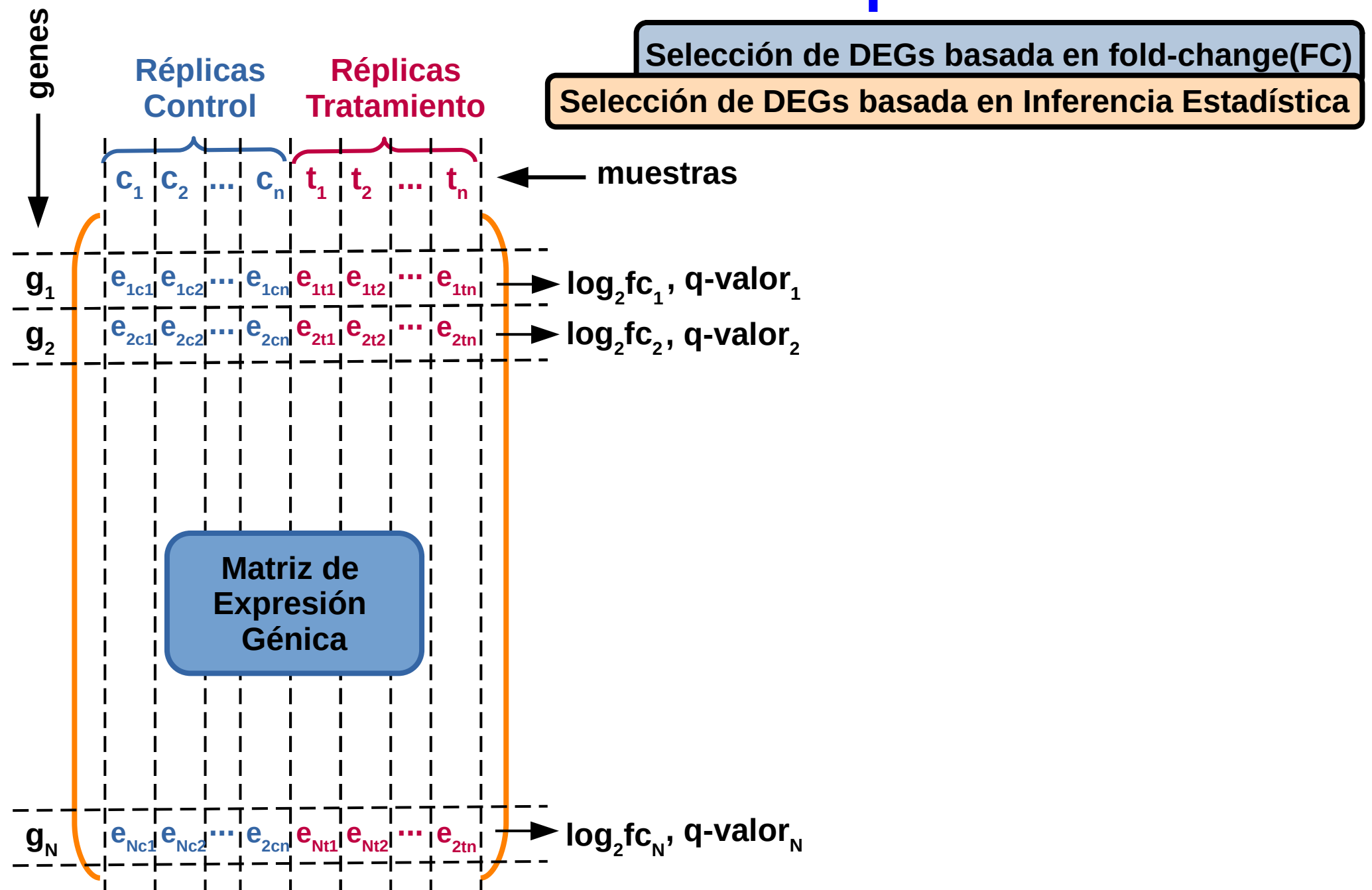


Paso 5.4: Selección de Genes Diferencialmente Expresados

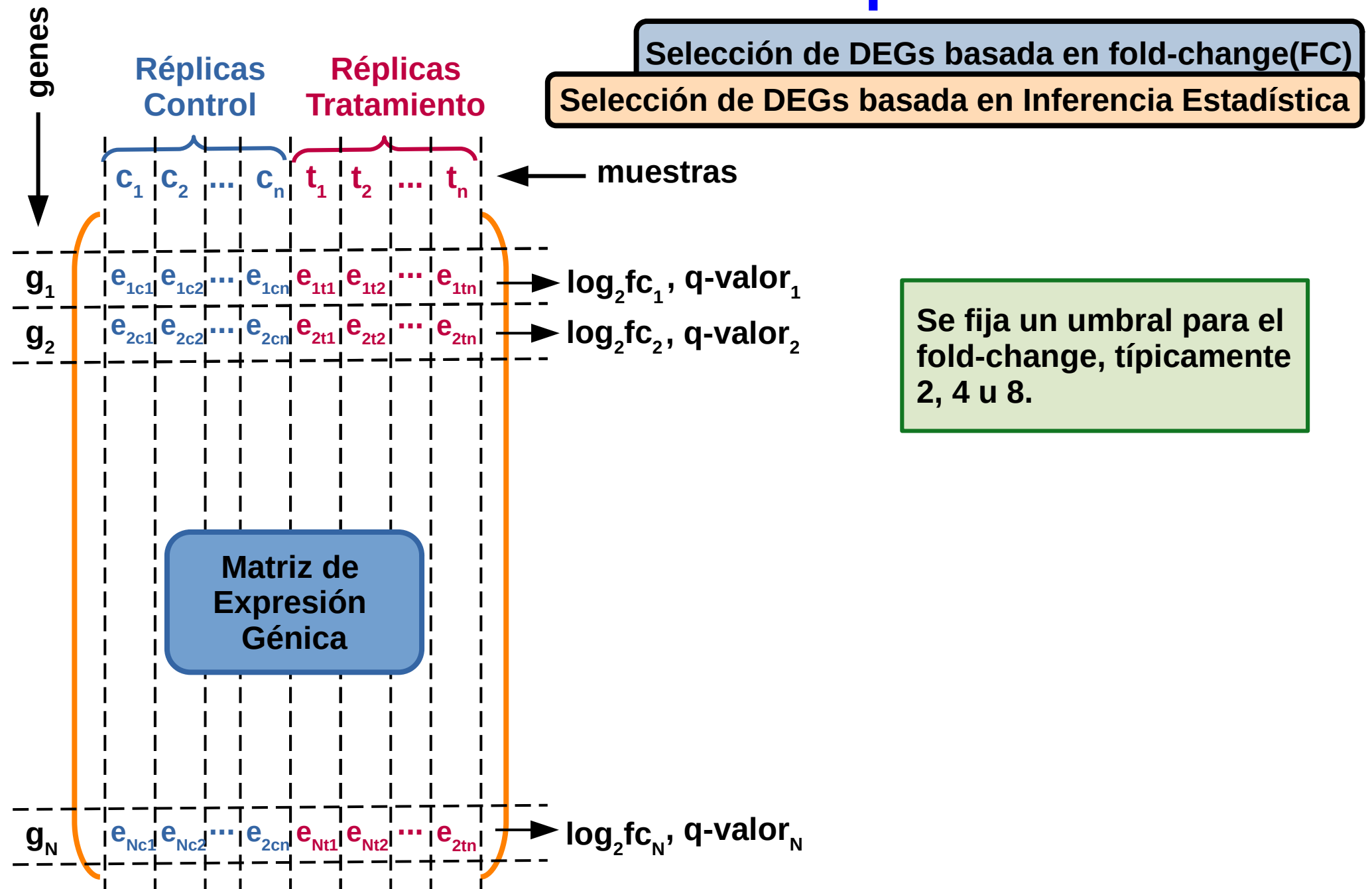
Selección de DEGs basada en fold-change(FC)



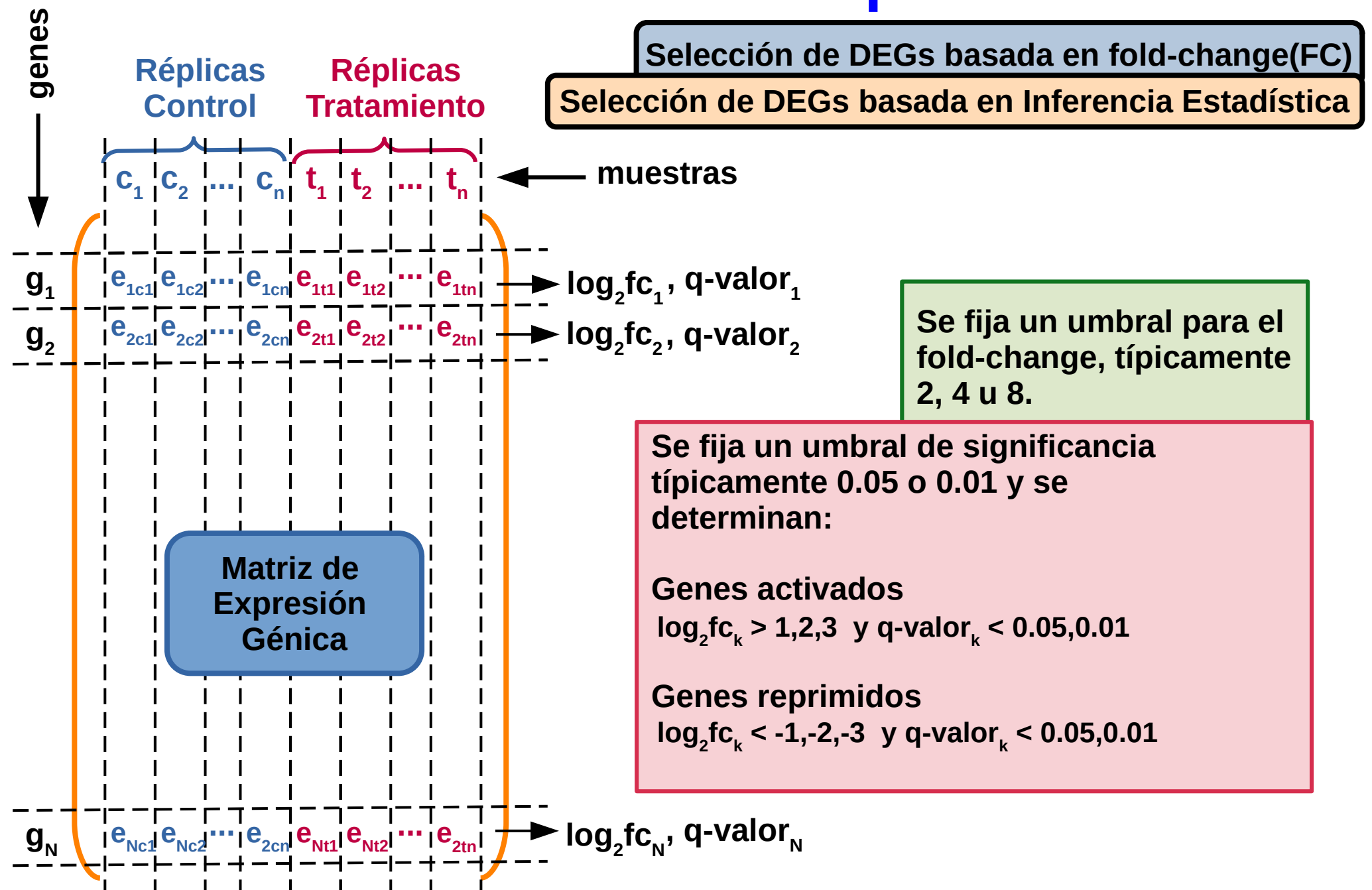
Paso 5.4: Selección de Genes Diferencialmente Expresados



Paso 5.4: Selección de Genes Diferencialmente Expresados



Paso 5.4: Selección de Genes Diferencialmente Expresados

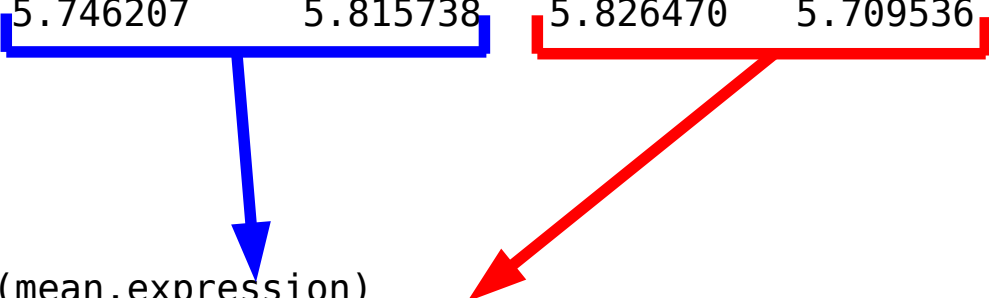


Paso 5.4: Selección de Genes Diferencialmente Expresados

Método basado en fold-change

```
> head(expression.level)
```

	WT_with_Fe_1	WT_with_Fe_2	WT_no_Fe_1	WT_no_Fe_2	pye_with_Fe_1	pye_with_Fe_2
244901_at	4.191814	4.481378	3.839588	3.993979	3.822802	4.330534
244902_at	4.727383	4.933480	4.537309	4.762214	4.601498	4.939710
244903_at	5.569595	6.274849	5.160593	6.061207	5.390620	6.119615
244904_at	5.146491	5.113851	5.117658	5.138793	4.966478	5.067616
244905_at	3.869215	3.793270	3.876314	3.776267	3.998623	4.182826
244906_at	5.746207	5.815738	5.826470	5.709536	5.698041	6.150307



```
> head(mean.expression)
```

	WT_with_Fe	WT_no_Fe	pye_with_Fe	pye_no_Fe	per_with_Fe	per_no_Fe
244901_at	4.336596	3.916784	4.076668	4.229890	5.618332	6.330688
244902_at	4.830432	4.649761	4.770604	4.804530	7.792671	7.630835
244903_at	5.922222	5.610900	5.755118	5.577170	8.063286	8.201758
244904_at	5.130171	5.128225	5.017047	5.257470	6.714227	6.563183
244905_at	3.831243	3.826290	4.090725	3.913093	4.740712	4.666626
244906_at	5.780972	5.768003	5.924174	5.851086	8.216157	8.541953

Paso 5.4: Selección de Genes Diferencialmente Expresados

La primera columna contiene los identificadores de las sondas que representan genes

> head(WT.With.no.Fe)

	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
9650	254550_at	6.277207	6.783913	9.83369	3.645075e-07	4.157208e-04	6.467981
8602	253502_at	6.265717	4.816715	33.08769	2.400984e-13	5.476646e-09	13.629654
12235	257135_at	5.794304	5.653286	24.50304	9.001222e-12	1.026589e-07	12.580207
10624	255524_at	4.754619	5.830145	20.50022	7.605795e-11	5.782940e-07	11.731873
17473	262373_at	4.478431	6.170963	18.56509	2.470630e-10	1.408877e-06	11.181815
6083	250983_at	4.200265	7.327805	16.54205	9.640295e-10	3.299739e-06	10.471151

Paso 5.4: Selección de Genes Diferencialmente Expresados

La segunda columna contiene para cada sonda/gen el fold-change en **log2** entre las condiciones/genotipos estudiados.

```
> head(WT.with.no.Fe)
```


	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
9650	254550_at	6.277207	6.783913	9.83369	3.645075e-07	4.157208e-04	6.467981
8602	253502_at	6.265717	4.816715	33.08769	2.400984e-13	5.476646e-09	13.629654
12235	257135_at	5.794304	5.653286	24.50304	9.001222e-12	1.026589e-07	12.580207
10624	255524_at	4.754619	5.830145	20.50022	7.605795e-11	5.782940e-07	11.731873
17473	262373_at	4.478431	6.170963	18.56509	2.470630e-10	1.408877e-06	11.181815
6083	250983_at	4.200265	7.327805	16.54205	9.640295e-10	3.299739e-06	10.471151

Paso 5.4: Selección de Genes Diferencialmente Expresados

Las columnas 5 y 6 contienen los p-valores y p-valores corregidos (según el FDR, false discovery rate) para los contrastes de hipótesis realizados entre las replicas de las distintas condiciones/genotipos

```
> head(WT.with.no.Fe)
```

	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
9650	254550_at	6.277207	6.783913	9.83369	3.645075e-07	4.157208e-04	6.467981
8602	253502_at	6.265717	4.816715	33.08769	2.400984e-13	5.476646e-09	13.629654
12235	257135_at	5.794304	5.653286	24.50304	9.001222e-12	1.026589e-07	12.580207
10624	255524_at	4.754619	5.830145	20.50022	7.605795e-11	5.782940e-07	11.731873
17473	262373_at	4.478431	6.170963	18.56509	2.470630e-10	1.408877e-06	11.181815
6083	250983_at	4.200265	7.327805	16.54205	9.640295e-10	3.299739e-06	10.471151



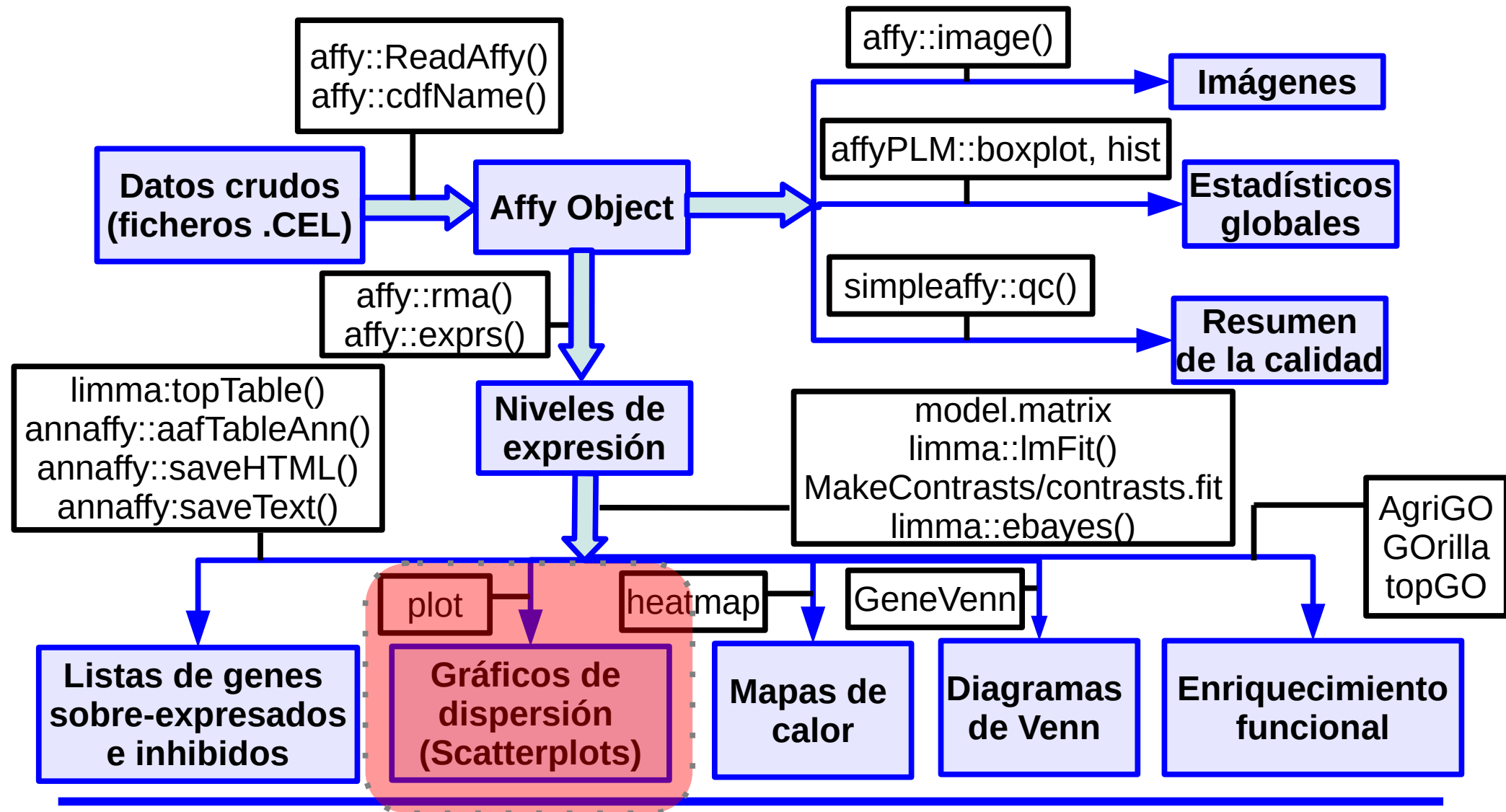
Paso 5.4: Selección de Genes Diferencialmente Expresados

Criterio más restrictivo a aplicar cuando hay mucho ruido experimental o se trabaja con especies no modelos. No es necesario con organismos modelos.

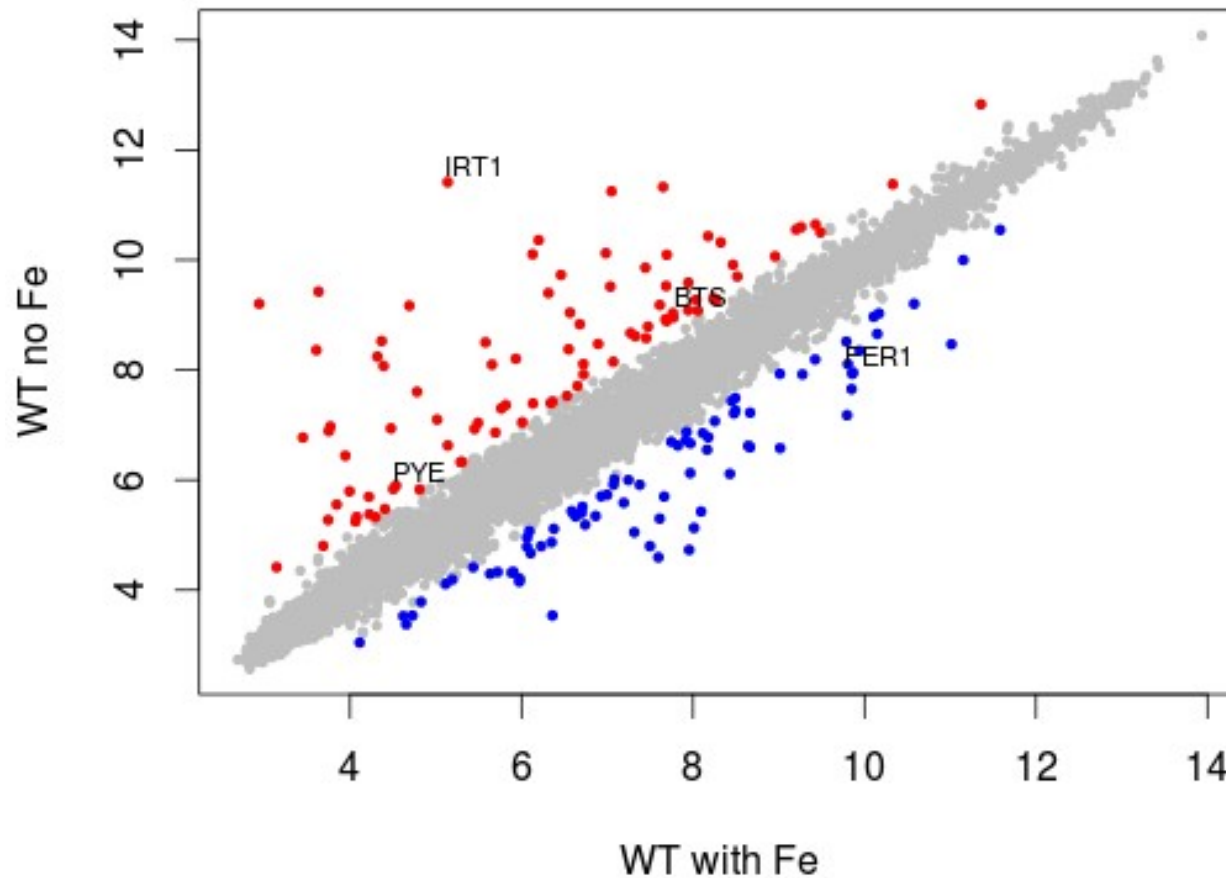


Comparación	Criterio del Fold-Change		Inferencia Estadística		Criterio conjunto	
	Genes Activados	Genes Reprimidos	Genes Activados	Genes Reprimidos	Genes Activados	Genes Reprimidos
WT -Fe/+Fe	90	80	35	7	35	7
PYE -Fe/+Fe	144	29	49	4	49	4
WT/PYE +Fe	22	12	0	1	0	1
WT/PYE -Fe	43	72	4	17	4	17

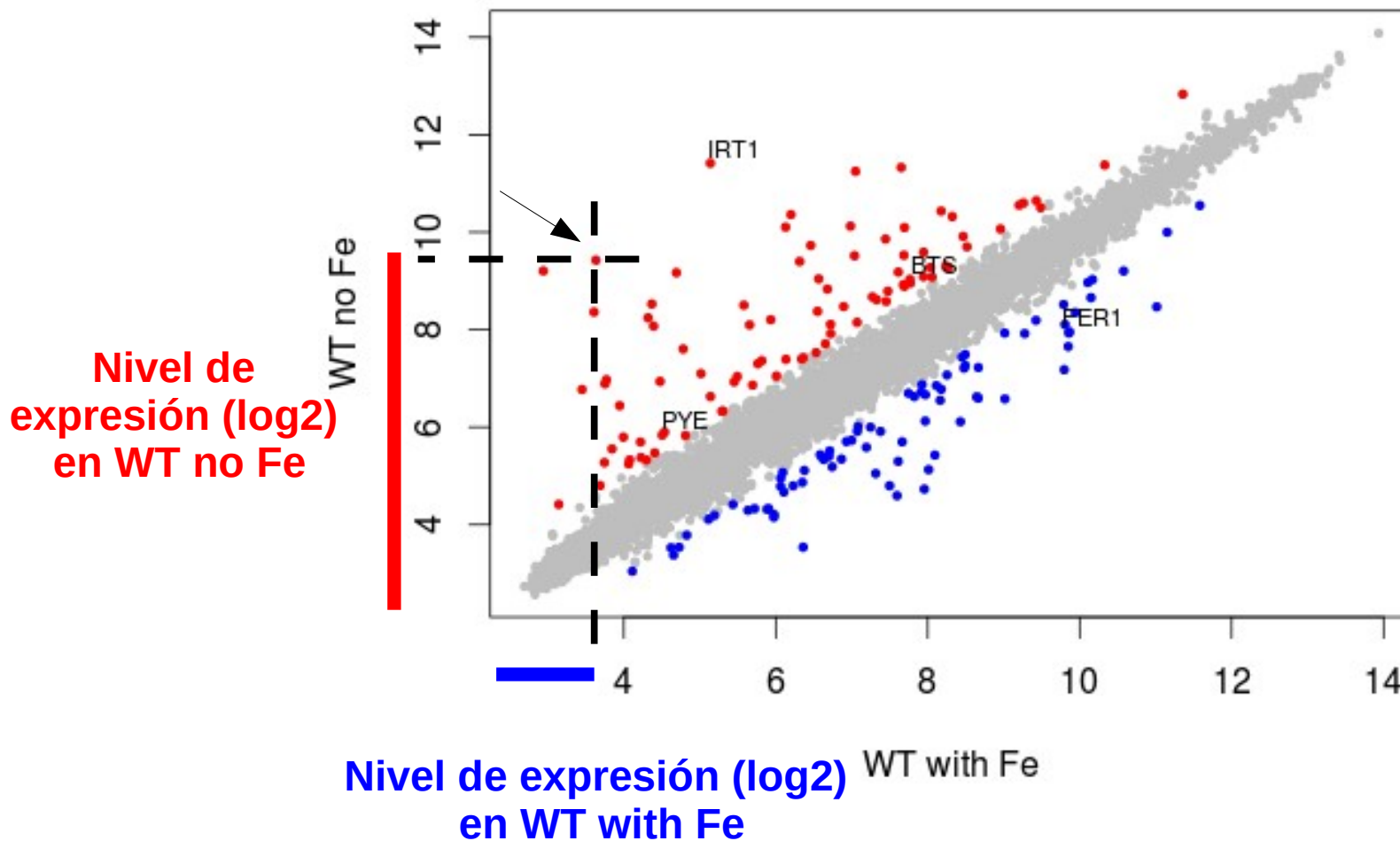
Análisis Explorativo: Gráficos de dispersión y de volcán



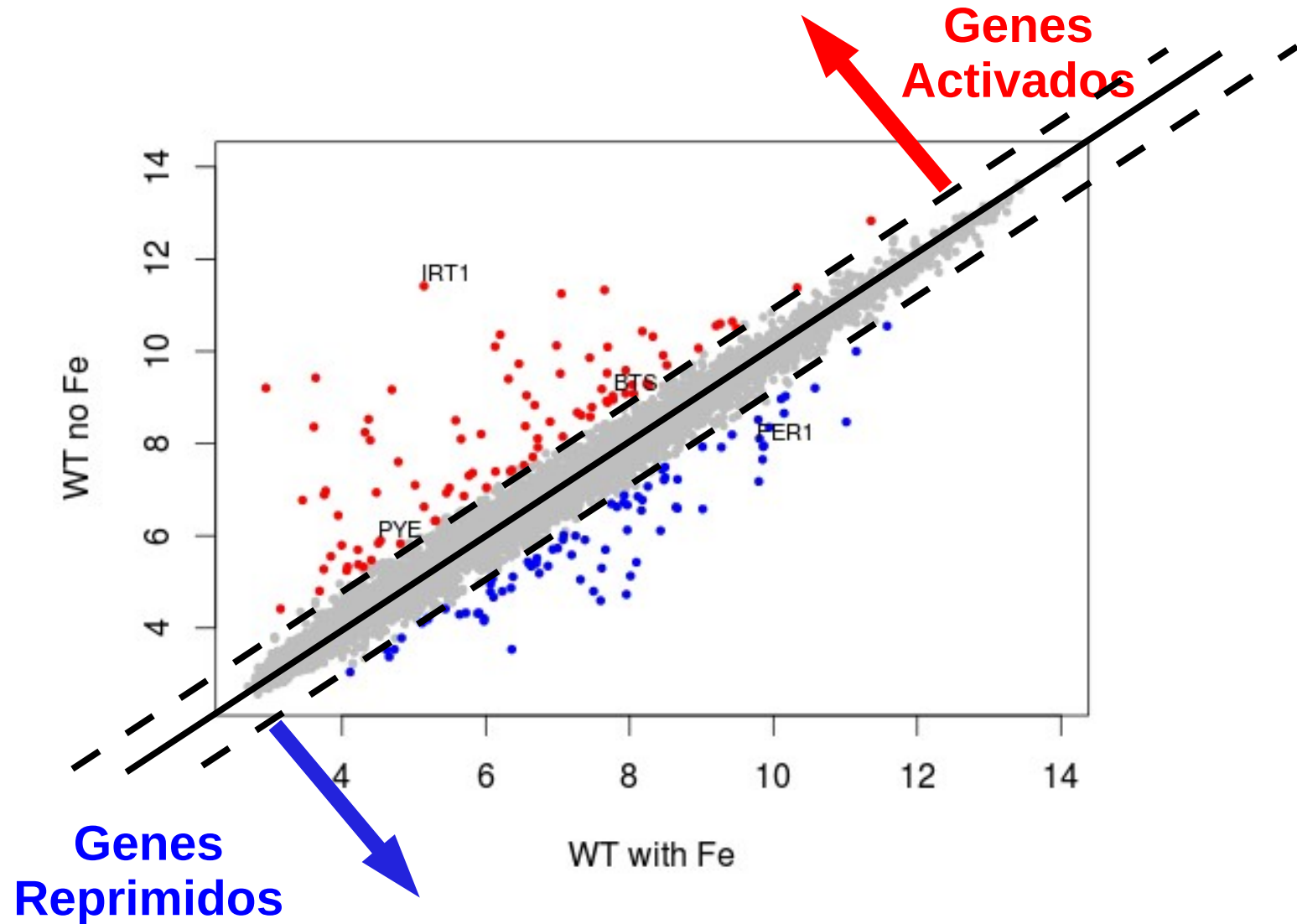
Análisis Explorativo: Gráficos de dispersión y de volcán



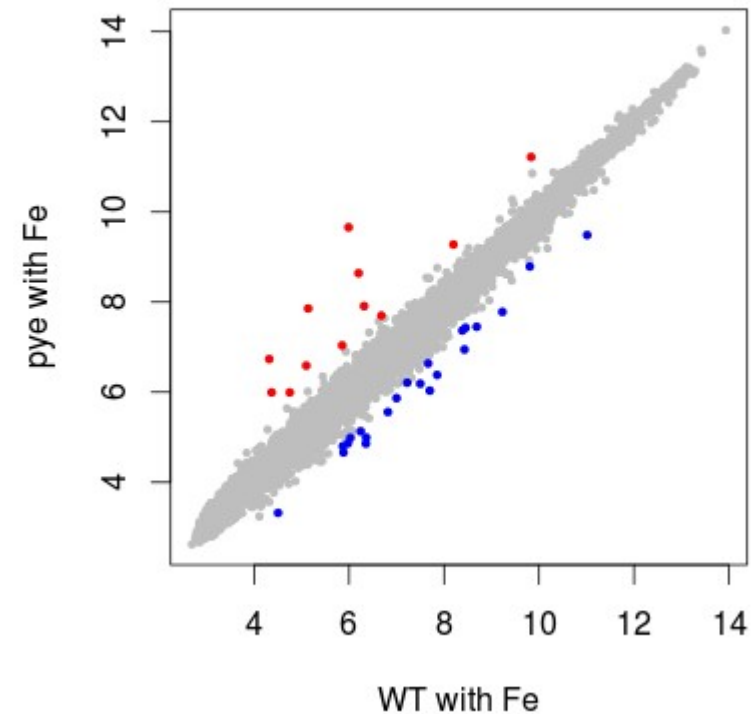
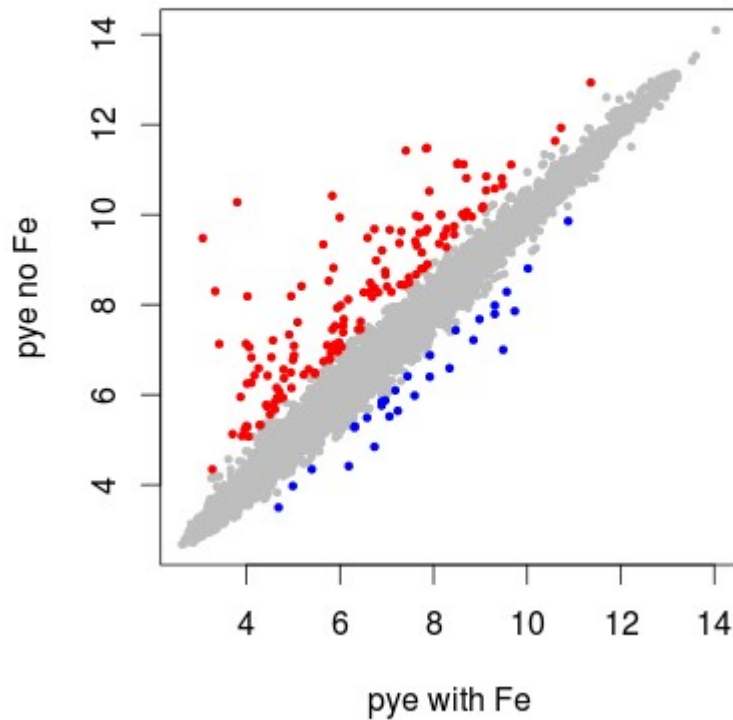
Análisis Explorativo: Gráficos de dispersión y de volcán



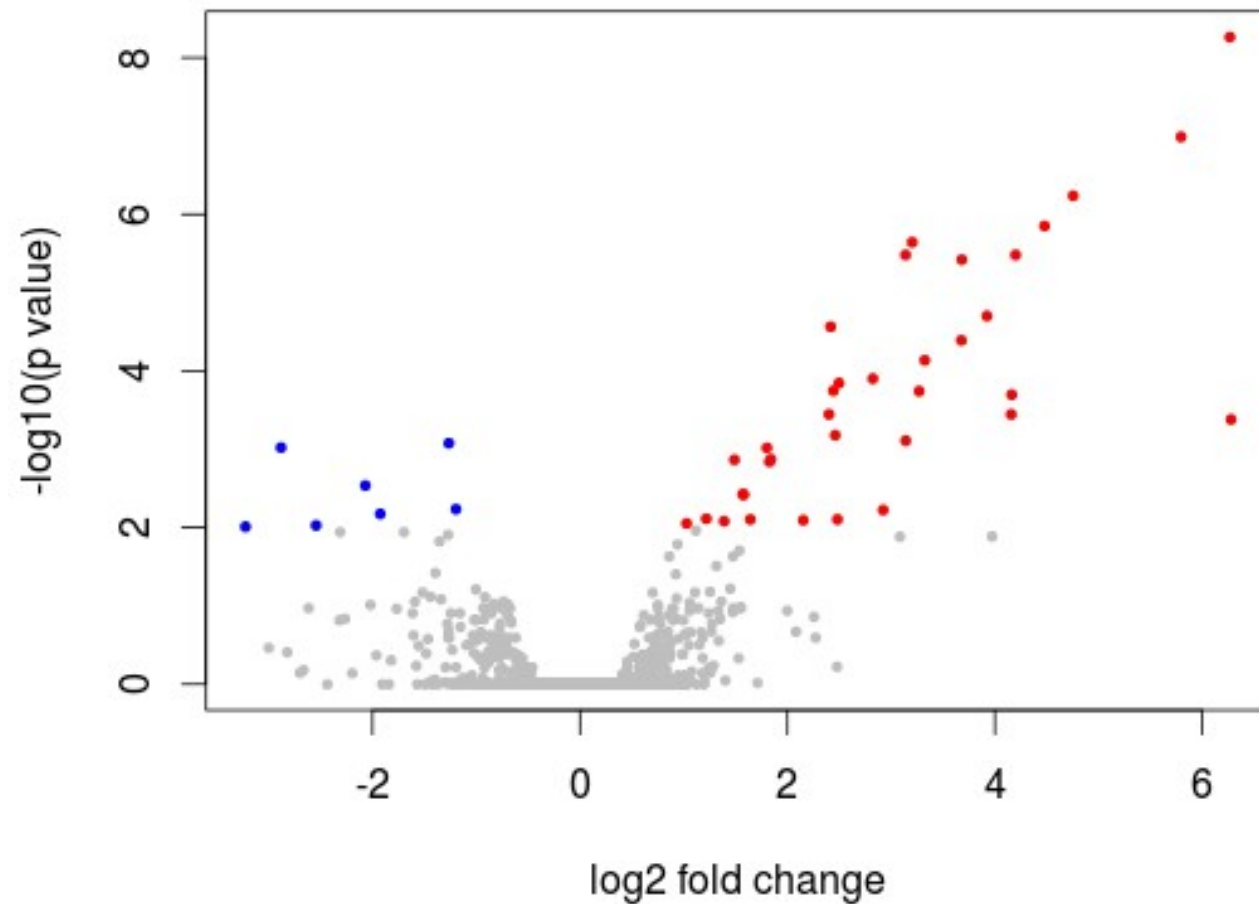
Análisis Explorativo: Gráficos de dispersión y de volcán



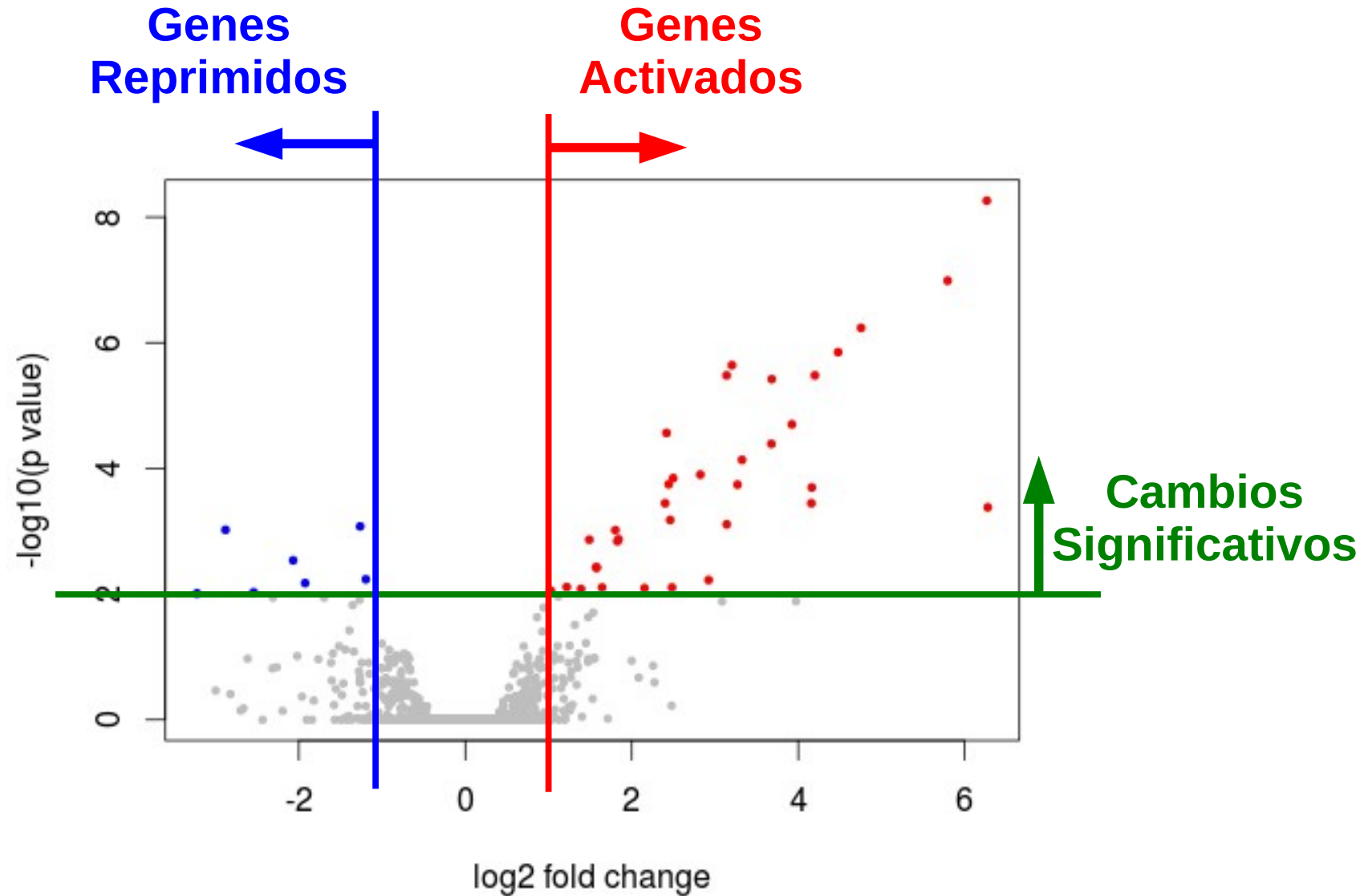
Análisis Explorativo: Gráficos de dispersión y de volcán



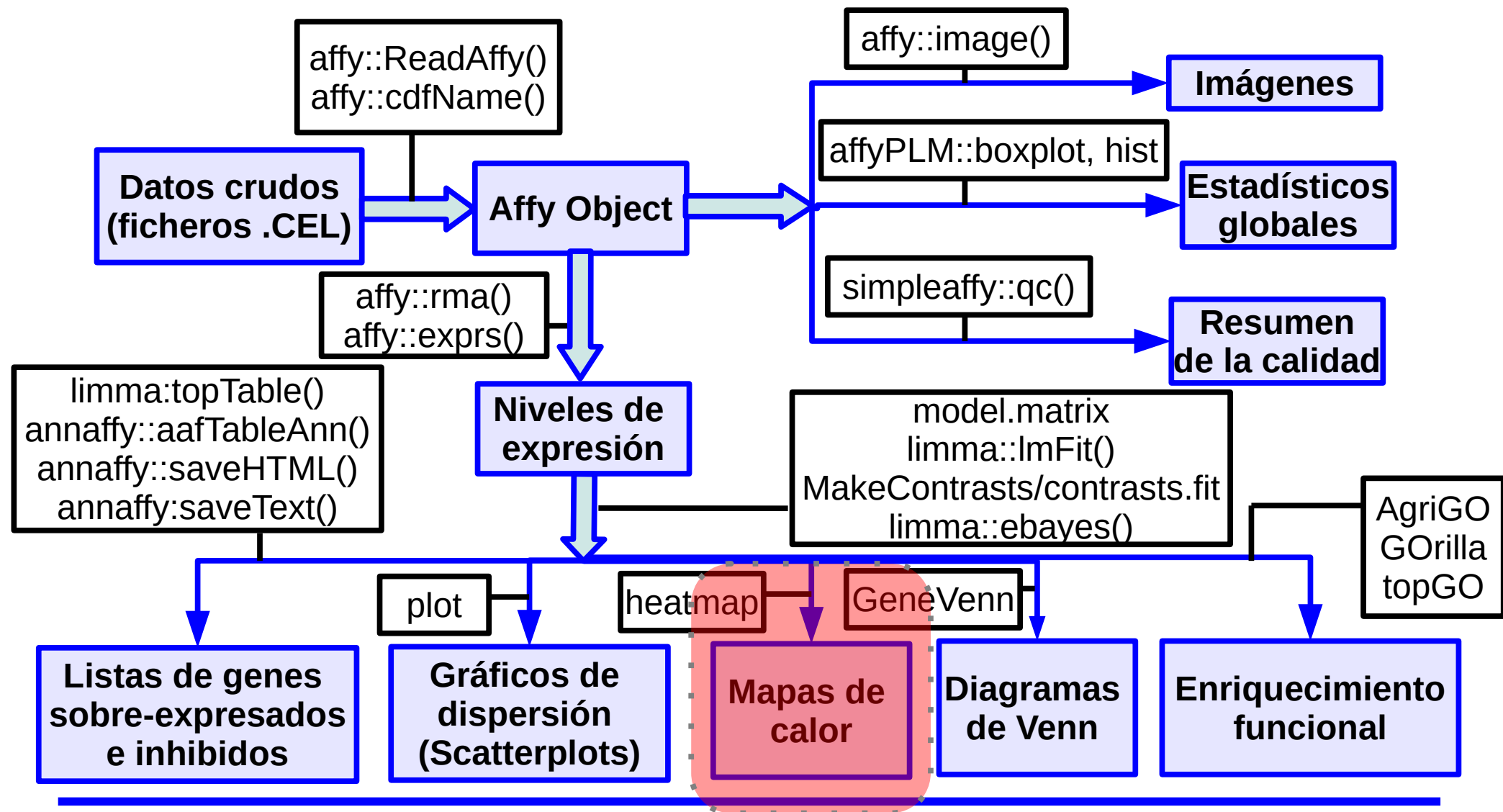
Análisis Explorativo: Gráficos de dispersión y de volcán



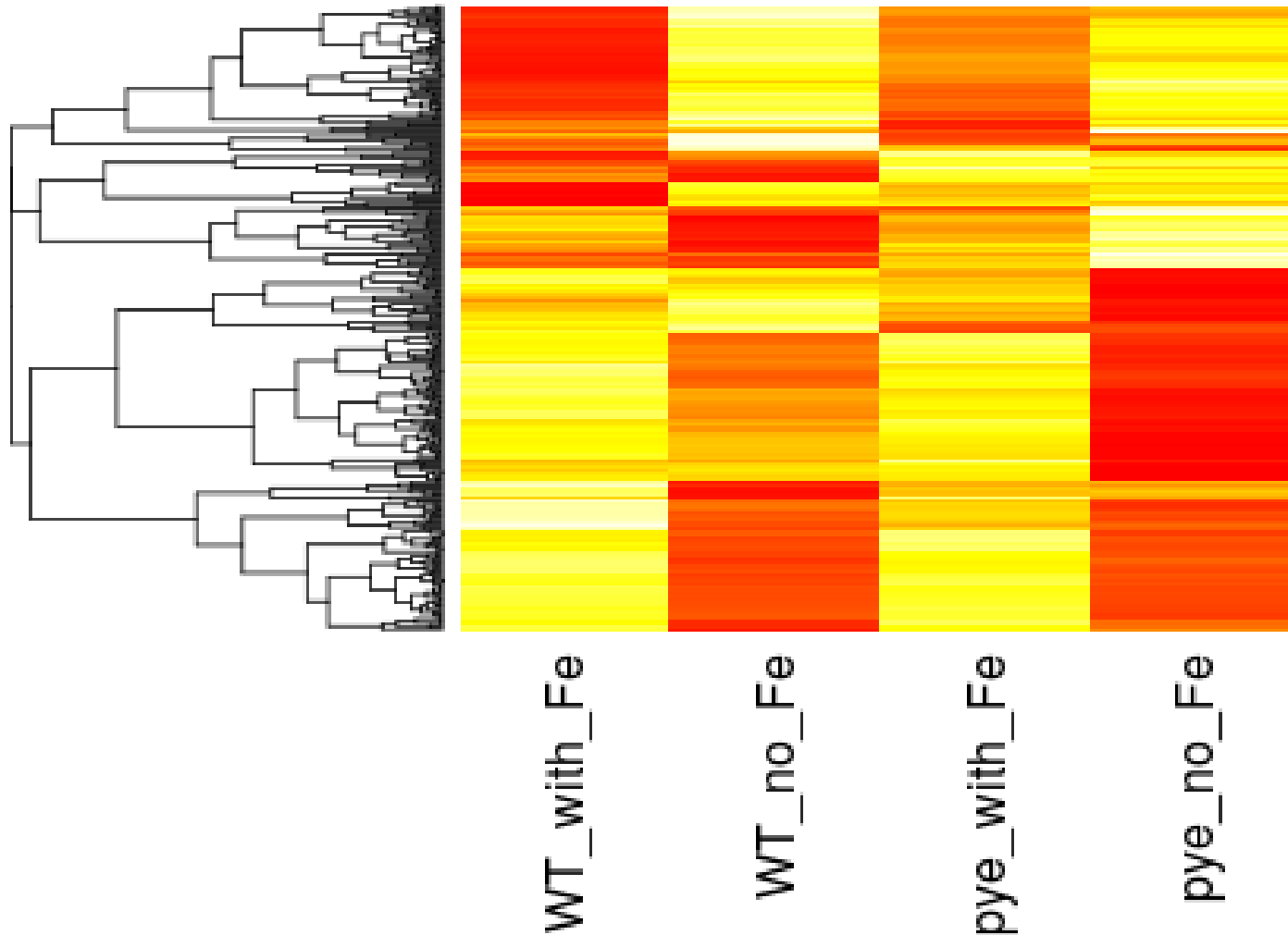
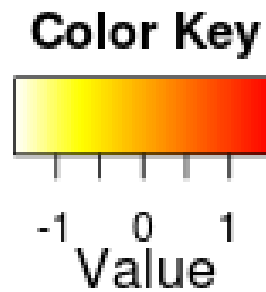
Análisis Explorativo: Gráficos de dispersión y de volcán



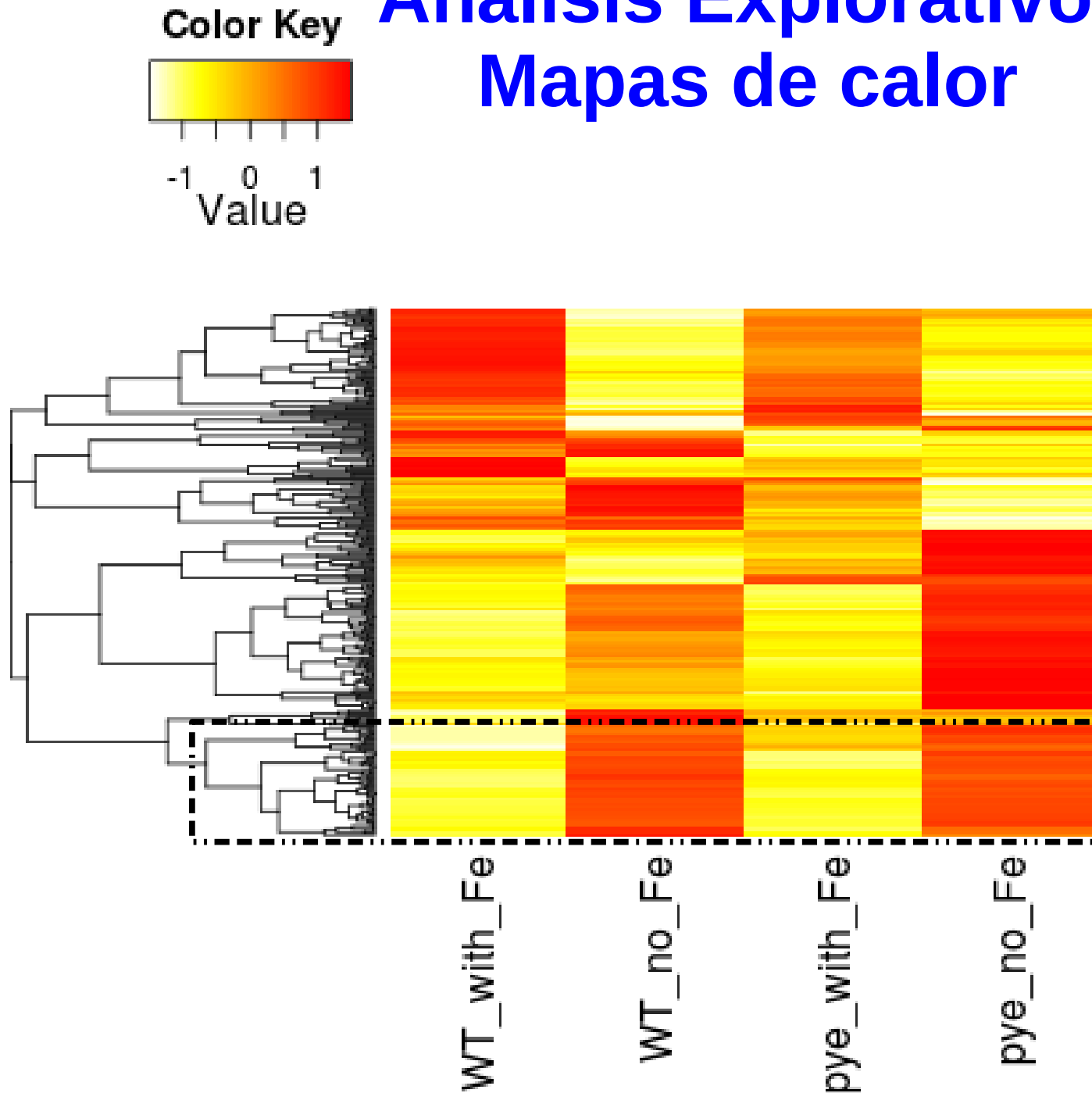
Análisis Explorativo: Mapas de calor



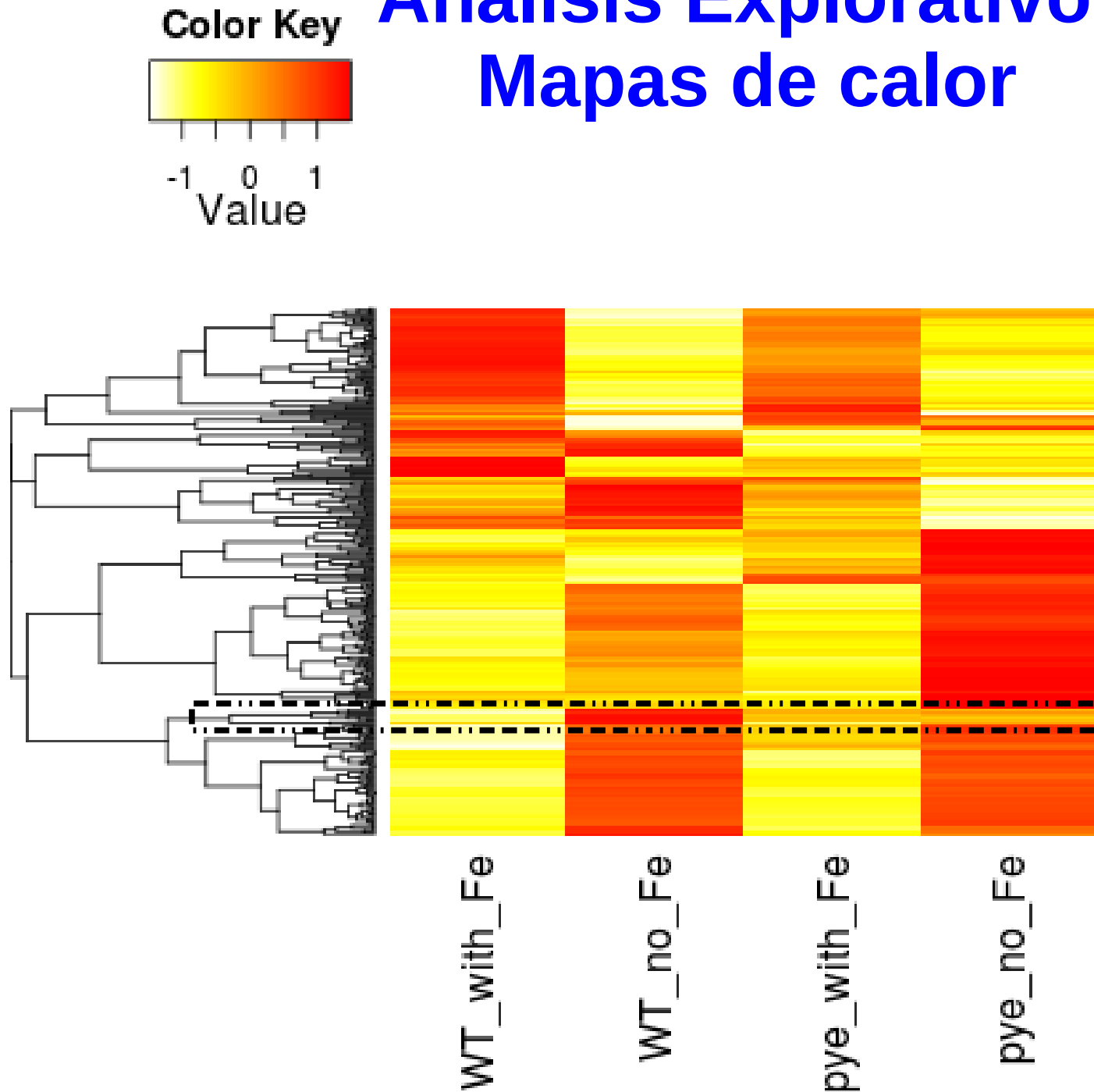
Análisis Explorativo: Mapas de calor



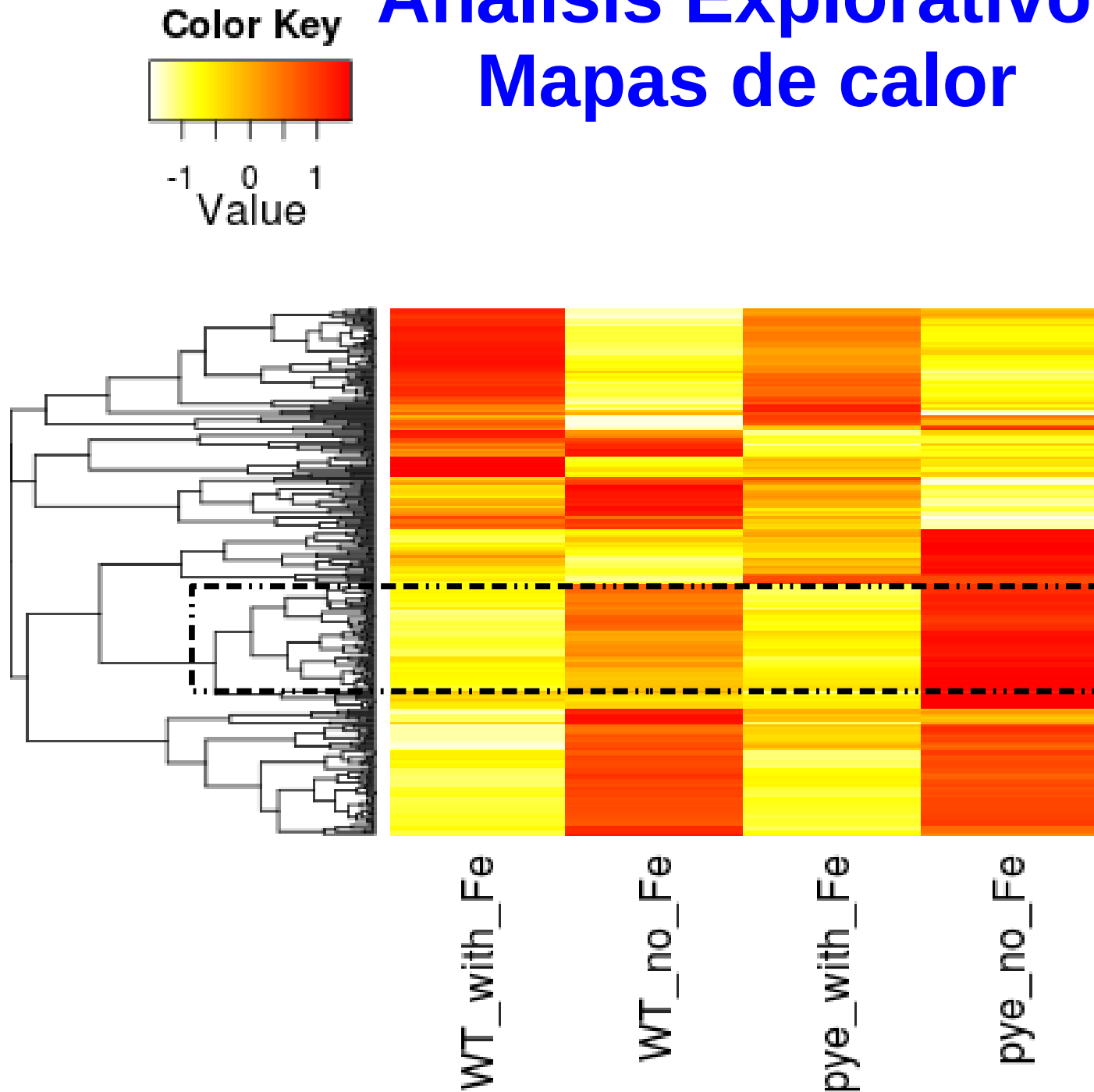
Análisis Explorativo: Mapas de calor



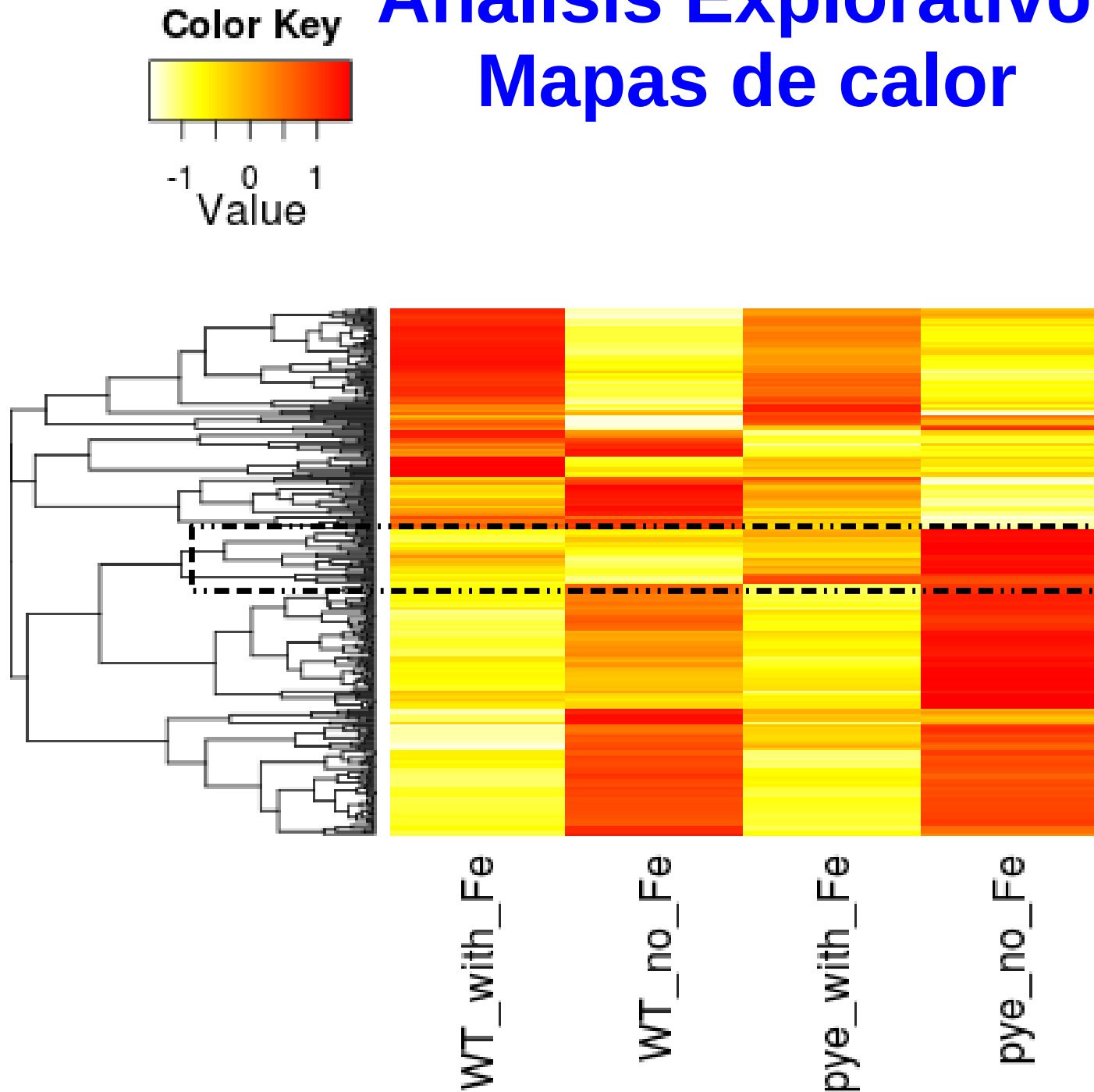
Análisis Explorativo: Mapas de calor



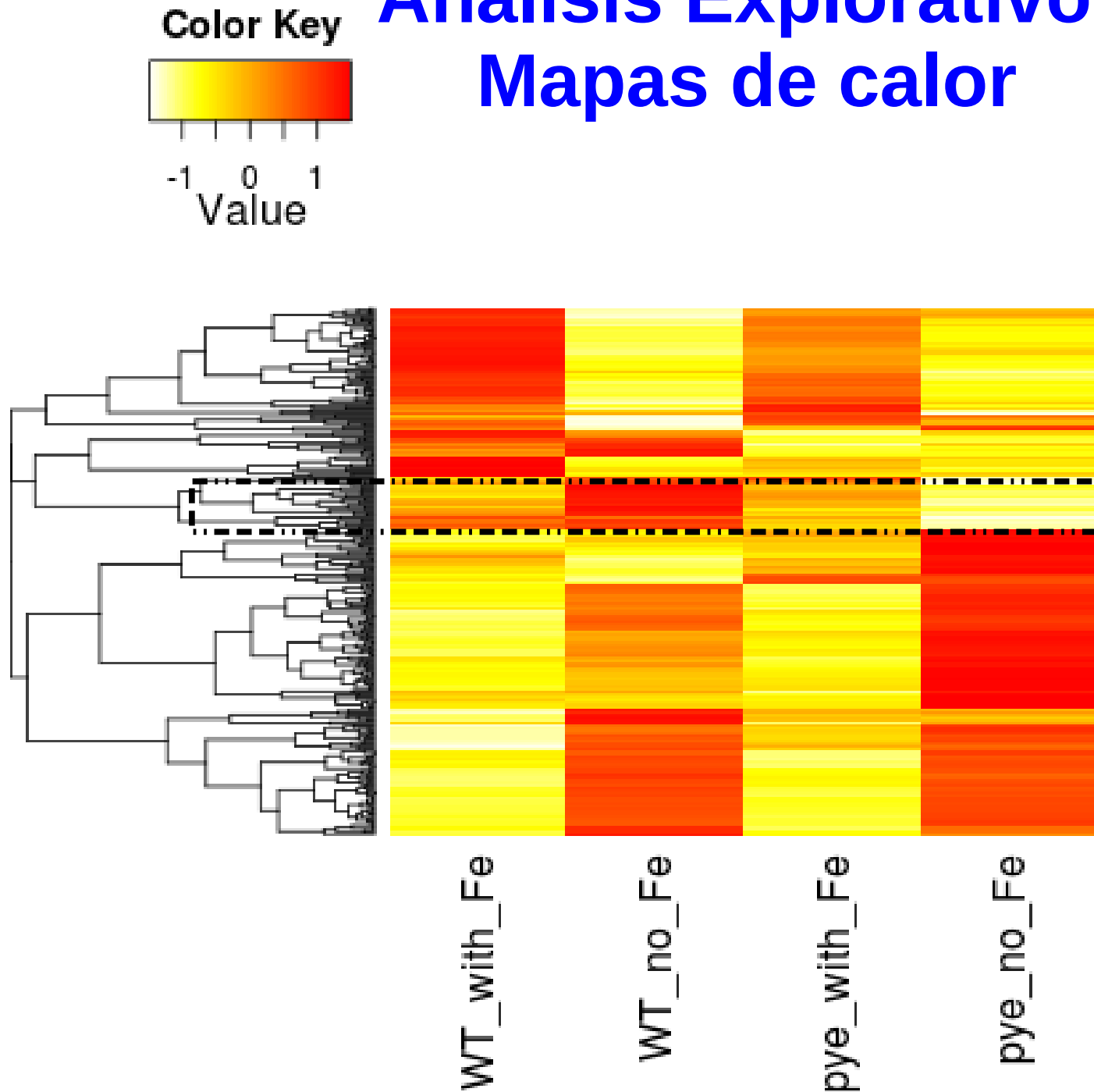
Análisis Explorativo: Mapas de calor



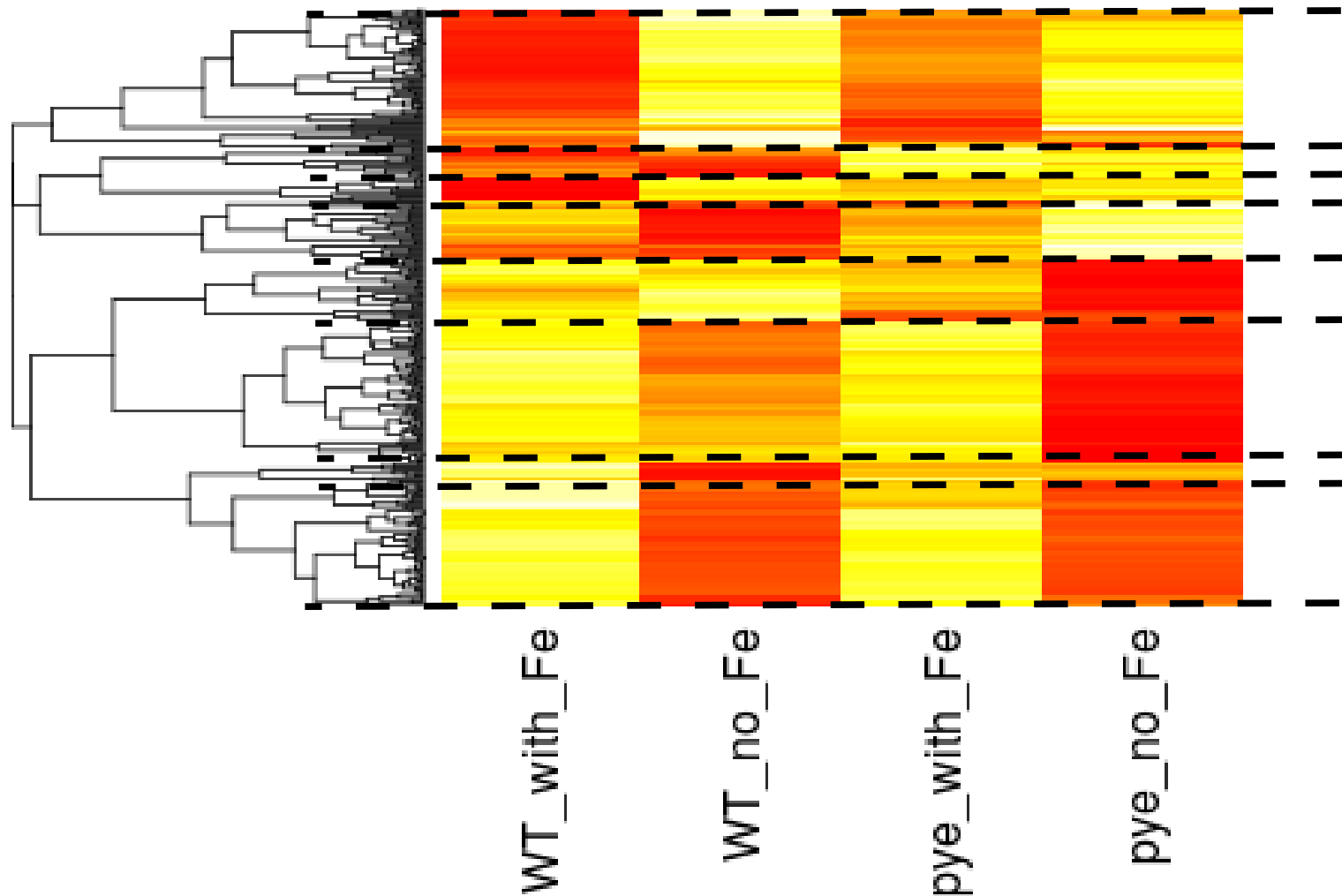
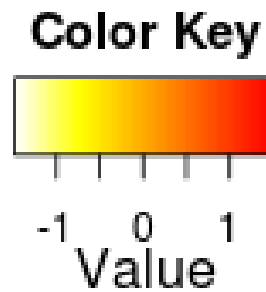
Análisis Explorativo: Mapas de calor



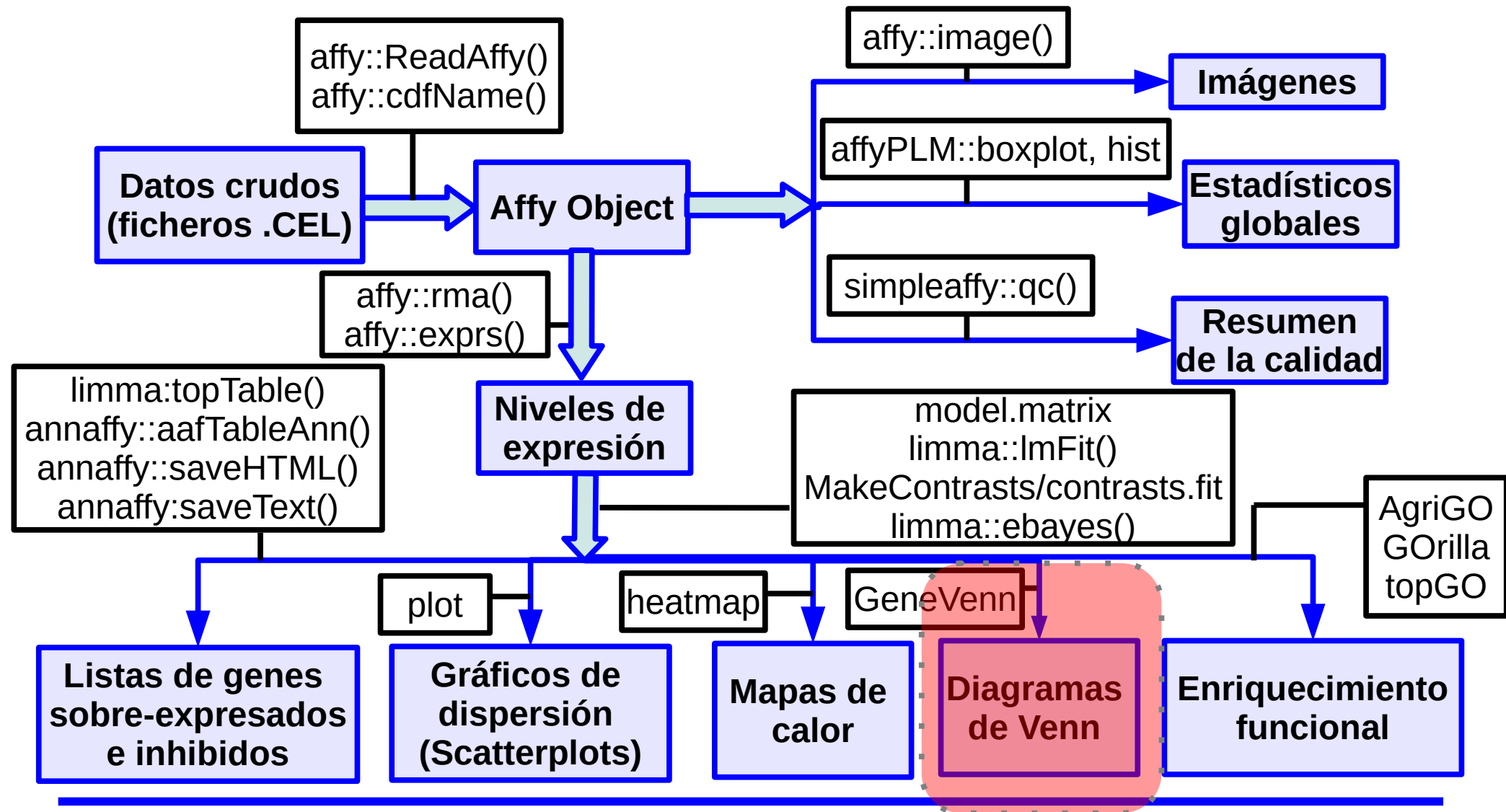
Análisis Explorativo: Mapas de calor



Análisis Explorativo: Mapas de calor



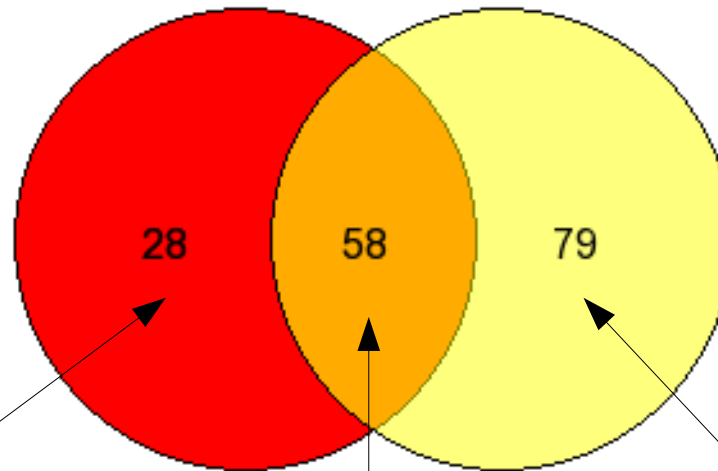
Análisis Explorativo: Diagramas de Venn



Análisis Explorativo: Diagramas de Venn

Genes Activados

WT Fe/-Fe pye Fe/-Fe



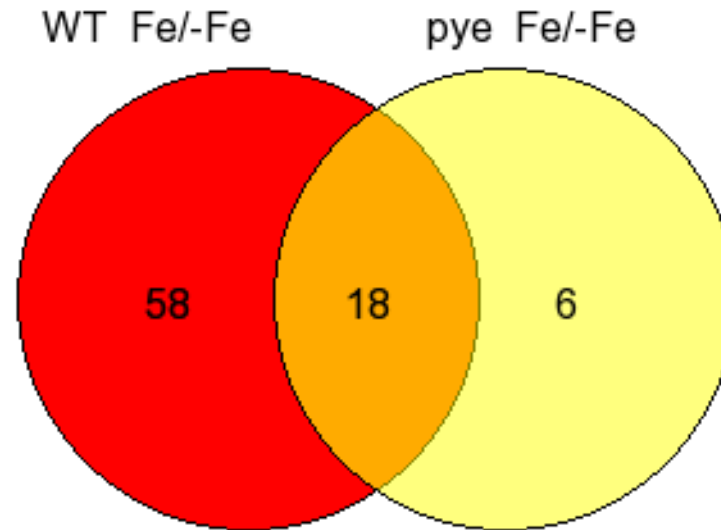
Genes activados por
falta de hierro específicamente
por la presencia de PYE

Genes activados por
falta de hierro
independientes de PYE

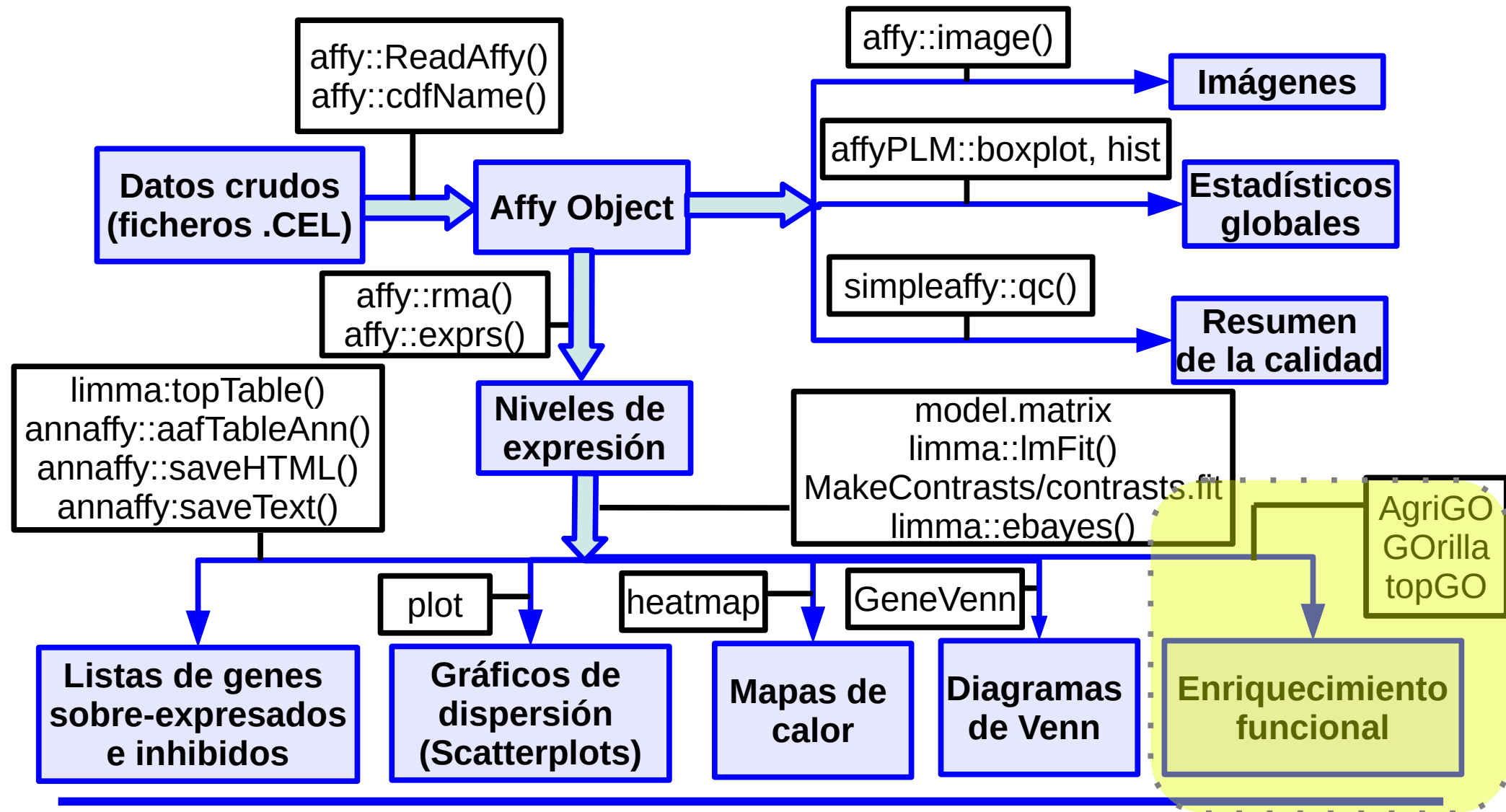
Genes activados por
falta de hierro específicamente
por la mutación de PYE

Análisis Explorativo: Diagramas de Venn

Genes
Reprimidos



Análisis Explorativo: Diagramas de Venn



Análisis Explorativo: Enriquecimiento de términos de Ontologías de Genes

Las ontologías de genes se desarrollaron para posibilitar la anotación (incorporación de información) a genes de forma sistemática e inequívoca.

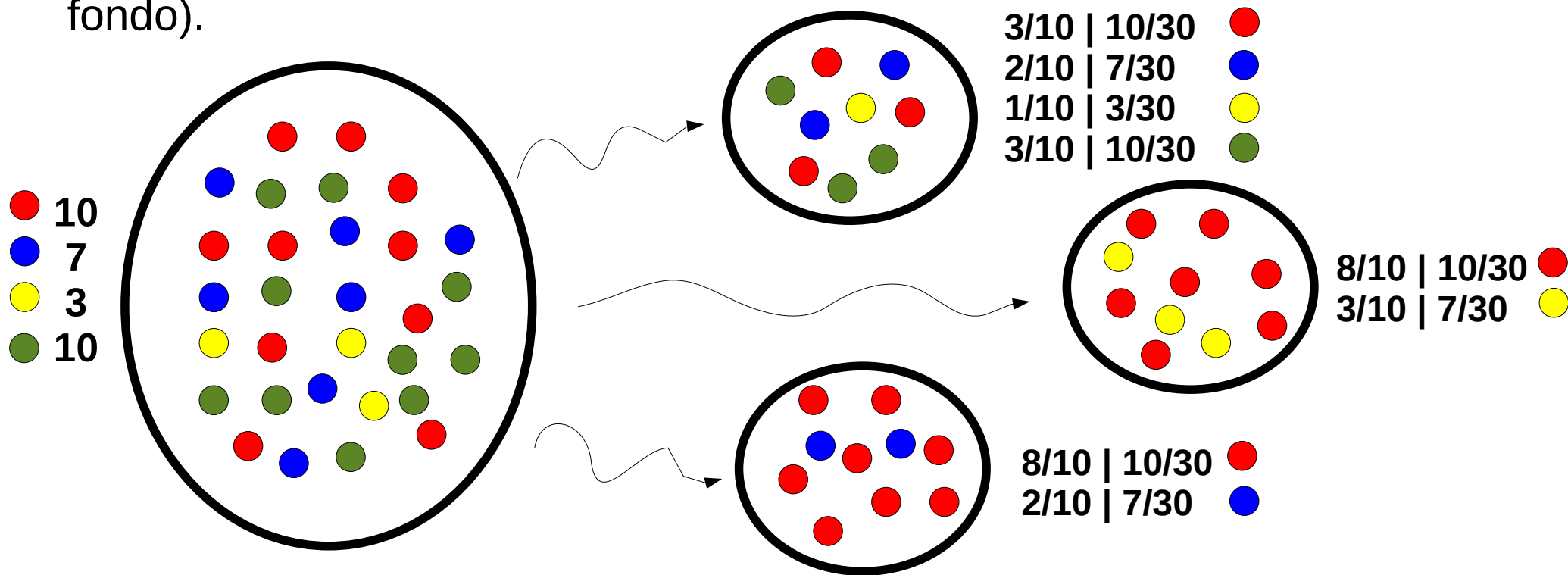
Las **ontologías de genes** consisten en un vocabulario estructurado y controlado de términos que describen productos génicos según:

- **Procesos biológicos:** Una serie de eventos moleculares con principio y fin.
- **Componentes celulares:** Localización en estructuras celulares o macromoleculares.
- **Funciones moleculares:** Actividades moleculares tales como actividades enzimáticas.

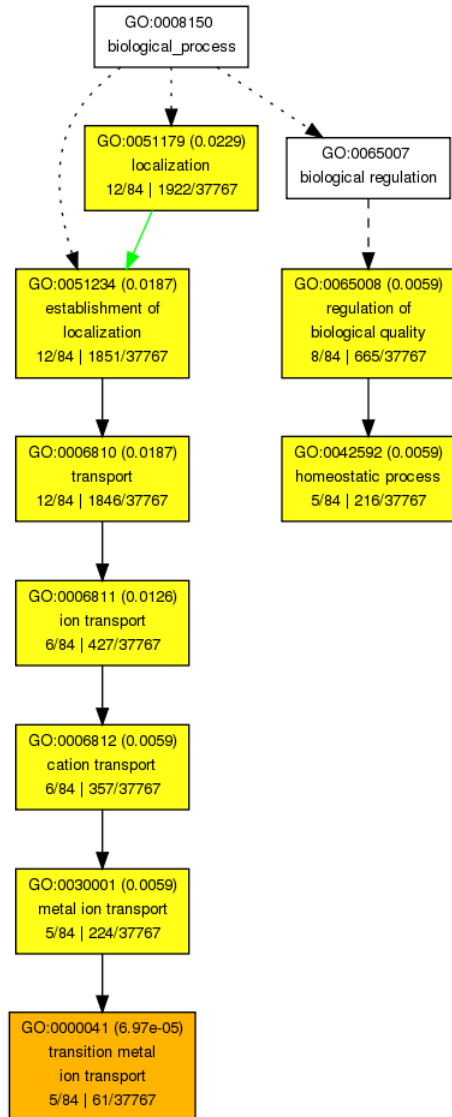
<http://www.geneontology.org/>

Análisis Explorativo: Enriquecimiento de términos de Ontologías de Genes

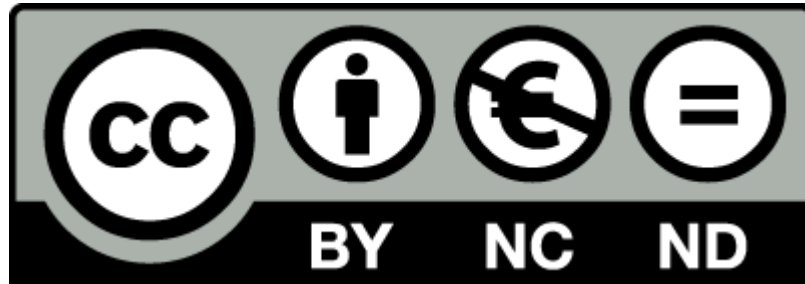
Gene Set Enrichment Analysis (GSEA) es un método matemático/computacional que determina si en un conjunto de genes dados hay términos de GO que aparecen con significancia estadística con respecto a un conjunto de genes que representan el universo total (o fondo).



Análisis Explorativo: Enriquecimiento de términos de Ontologías de Genes



GO term	Description	Representative genes	p-value
GO:0000041	transition metal ion transport	IREG2 IRT1 COPT2	3.8e-07
GO:0030001	metal ion transport	IREG2 IRT1 COPT2	0.00016
GO:0006811	ion transport	IREG2 IRT1 COPT2	0.00041
GO:0006812	cation transport	IREG2 IRT1 COPT2	0.00016
GO:0042592	Homeostatic process	GH3.3 ZIF1	0.00014



This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/>.
