

# GUÍA SESIÓN HACKATHON DÍA 1

## Anotación Funcional de la señalización molecular por estrés en *Arabidopsis*

### Motivación



### Resumen

Las poliaminas son compuestos nitrogenados presentes en las plantas que se acumulan en respuestas a estrés. La acumulación de poliaminas específicas en plantas, produce la activación de las defensas, posiblemente a partir de la activación de la ruta metabólica del ácido salicílico.

En un [estudio](#) reciente, investigadores de la Universidad de Barcelona y la Universidad de Düsseldorf han demostrado que existen diferencias en la transcripción desencadenada a partir de la acción de las poliaminas.

El grupo de investigación plantea un experimento de secuenciación a partir de un ensayo de **RNASeq** donde se secuencia tejido de hojas de plantas crecidas en medio de cultivo tratado con distintas poliaminas. En el trabajo se aplica un tratamiento con 5 poliaminas diferentes. Para efectos ilustrativos, en este ejercicio consideraremos sólo los tratamientos con [putrescina](#) y termoespermina. El ensayo se realiza sobre tres repeticiones biológicas, conteniendo cada repetición 3 plantas sobre placas independientes. Como control, se usa un medio tratado con un buffer (MES).

A partir de la extracción de ARN , se preparan las librerías y se secuencian con un equipo HiSeq2000 de Illumina. Las lecturas generadas se mapean contra el genoma de referencia de Arabidopsis y se identifican los genes con expresión diferencial en cada tratamiento.

## Objetivo

Punto de vista biológico : Determinar si existen diferencias cuantitativas / cualitativas en la activación transcripcional de plantas tratadas con putrescina y termoespermina.

Punto de vista computacional : Generar herramientas automáticas de análisis y representación gráfica de la anotación funcional.

## Metodología

La forma directa de estudiar la respuesta molecular de un experimento es comparar el grupo de genes cuya expresión se regula (sobreexpresión / inhibición) por un tratamiento en relación al grupo control. Para caracterizar cada grupo de genes , se realiza la **anotación funcional**, asociando a cada gen, unos identificadores estándar vinculados con su función molecular, proceso biológico y componente celular.

## Datos

La siguiente [Tabla](#) ha sido adaptada a partir de los datos del [artículo](#). La tabla contiene el listado de genes que presentan expresión diferencial en respuesta a la aplicación en el medio de crecimiento de 100  $\mu$ M putrescina (Put) o termoespermina (tSpm). Las diferencias de expresión en cada caso son relativas al tratamiento control. Consideramos expresión diferencial cuando la variación de expresión  $\geq |2|$ -fold y el valor estadístico  $P \leq 0,05$  .

La tabla de datos contiene las siguientes columnas :

**Gene Id.** : identificador del gen ([NCBI](#)).

**Put1-Put3** : variación de la expresión génica en las 3 repeticiones para el trat. Put.

**Put avg Fold-change** : media de la variación de la expresión (Put.)

**Put log Fold-change** : log. media ([base 2](#)).

**Pval** : Pvalue (corrección Bonferroni). Solo se indican aquellos casos  $\leq 0,05$ .

La tabla contiene columnas análogas para el tratamiento con termoespermina.

## Propuesta metodológica

Lo que sigue a continuación es una propuesta para abordar el reto. Puede usarse cualquier otra metodología si permite analizar el resultado del experimento RNASeq y se crea un registro de análisis reproducible y replicable.

Proponemos un análisis en dos partes :

- sección descriptiva
- sección funcional

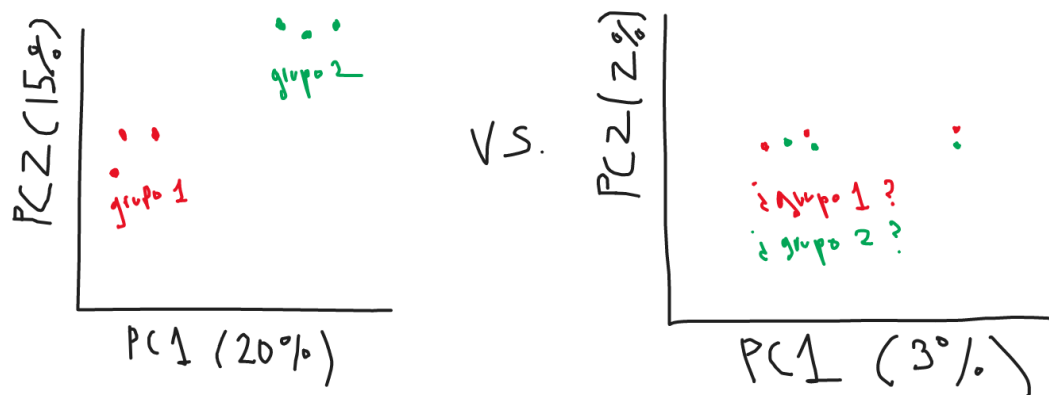
En la parte descriptiva se puede emplear un Análisis de Componentes Principales o un Cluster Analysis. En la sección funcional, se realizará la anotación funcional propiamente dicha.

### Análisis de Componentes Principales

Como primer paso podemos realizar un PCA. El PCA muestra agrupaciones de muestras basándose en su similaridad. Para cada muestra que estamos analizando (Put1, Put2, Put3, tSpm1, tSpm2, tSpm3 ) tenemos información sobre la variación de expresión en 280 genes. Podemos pensar en las 6 muestras como variables de 280 dimensiones. Una forma sencilla de tratar los datos multi-dimensionales es representando las 2 componentes principales que contienen la mayor variación del experimento.

Con esta aproximación podremos ver si :

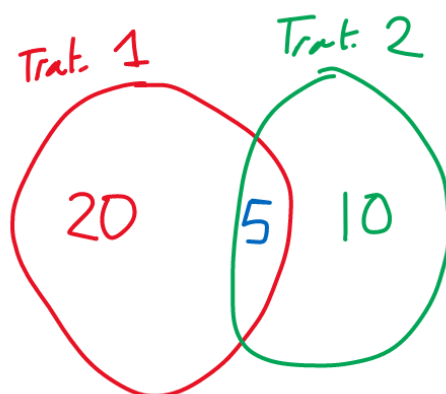
- i) existe coherencia dentro de las repeticiones biológicas ?
- ii) hay diferencias entre tratamientos ? (Put vs tSpem)



**Figura 1.** Análisis de Componentes Principales.

### Análisis cuantitativo de la regulación

A continuación, queremos cuantificar la regulación de cada tratamiento. Podemos crear alguna figura donde aparezca claramente indicado el número de genes regulados de forma específica en cada tratamiento y el número de genes regulados por los dos tratamientos (valor medio  $\geq |2|$ -fold y  $P \leq 0,05$ ). Una forma típica de representación son los diagramas Venn.



**Figura 2.** Genes específicos regulados por los tratamientos 1 y 2 ; genes en común regulados por los dos tratamientos.

### Anotación funcional

El objetivo de la anotación funcional es ordenar los resultados de un experimento de secuenciación de modo que nos permita extraer conclusiones biológicas relevantes.

Genes Ids.



GO terms



GO slim

Para ello trabajaremos con distintas bases de datos y tablas.

### QuickGo

En la sección de búsqueda de [QuickGo](#) podemos introducir el *Gene Id*. La Base de Datos devolverá las anotaciones disponibles para cada gen. En este ejemplo (2 genes) , existen 23 anotaciones para un gen y 10 anotaciones para el otro gen.

The screenshot shows the QuickGo search interface. At the top is a 'Search' bar with the input 'AT3G63150 AT4G01630'. Below it is a 'Terms' section. The 'Gene products' section displays two results: 'F4J0W4 Mitochondrial Rho GTPase 2' with '23 annotations' and 'Q9ZS11 Putative expansin-A17' with '10 annotations'. A link 'Show all 2 results' is present. Below this is a table with the following data:

Database	ID	Name	Type	Taxon	Annotations
UniProtKB	F4J0W4	Mitochondrial Rho GTPase 2	PROTEIN	Arabidopsis thaliana	23 annotations
UniProtKB	Q9ZS11	Putative expansin-A17	PROTEIN	Arabidopsis thaliana	10 annotations


**Figura 3.** Esquema de trabajo en Anotación Funcional.

Debemos hacer click en las anotaciones de cada gen y descargarnos el fichero en formato TSV. Solo necesitaremos la columna "Symbol".

The screenshot shows the QuickGo export interface. At the top are 'Customise' and 'Export' buttons, with '23 annotations' displayed. The 'Choose' section has four checkboxes: 'Symbol' (checked), 'Qualifier', 'Evidence', and 'Reference'. Below these is a 'Reset' button. The 'Export' section has a 'Format' dropdown menu set to 'Tab-delimited (TSV)', a 'Limit (maximum 2,000,000):' input field with '1000', and a 'Go' button. On the right side, there is a vertical list of column headers: 'ate', 'ame', 'ynonym', and 'ype'.

**A efectos de simplificar el reto , hemos descargado la anotación de los genes expresados diferencialmente en cada experimento. Puedes acceder a los ficheros [aquí](#).**

Tras descargar las anotaciones disponibles para cada gen construiremos un fichero único con 2 columnas : GENE PRODUCT ID (o SYMBOL) y el GO term. Este archivo contendrá todas las anotaciones posibles en las 3 ontologías (BP, MF, CC) para cada gen que se encuentre anotado en la base de datos QuickGo. El fichero único de anotaciones GO lo crearemos para cada tratamiento. **Pista : es importante comprobar que cada fila de este fichero es única y que no estamos repitiendo la misma anotación varias veces.** Si no comprobamos esto, corremos el riesgo de sobreestimar la anotación real de nuestro experimento.



F4J0W4	GO:0000166
F4J0W4	GO:0003924
F4J0W4	GO:0005509
F4J0W4	GO:0005525
F4J0W4	GO:0005739
F4J0W4	GO:0005741
F4J0W4	GO:0007005
F4J0W4	GO:0009737
F4J0W4	GO:0010821
F4J0W4	GO:0016020
F4J0W4	GO:0016021
F4J0W4	GO:0016787
F4J0W4	GO:0031307
F4J0W4	GO:0046872
Q9ZSI1	GO:0005576
Q9ZSI1	GO:0005618
Q9ZSI1	GO:0009664
Q9ZSI1	GO:0009828
Q9ZSI1	GO:0010311
Q9ZSI1	GO:0016020
Q9ZSI1	GO:0071555

### *GOSlim*

Una vez tenemos el archivo de anotación para cada tratamiento, el siguiente paso consiste en **simplificar las anotaciones**. El vocabulario GO está jerarquizado de modo que podemos reducir los términos ascendiendo en la escala. Por ejemplo, el ‘mantenimiento de la identidad del meristemo’ (GO:0010074) forma parte del nivel superior ‘mantenimiento de la población de células madre’ (GO:0019827), que a su vez forma parte del nivel superior ‘mantenimiento del número celular’ (GO:0098727), que a su vez está en el primer nivel de la jerarquía ‘Proceso Biológico’ (GO:0008150). La base de datos [AgBase](#) contiene distintas herramientas para el análisis funcional de productos de plantas y animales. La herramienta GOSlim Viewer permite obtener los niveles superiores en jerarquía de la anotación GO.

Para cada tratamiento, cargaremos nuestro archivo de términos GO y seleccionaremos el set GOSlim de plantas.

## GOSlimViewer

GOSlimViewer is used to provide a high level summary of functions for a dataset. The output can be charted in Excel to obtain publication quality figures. Note that records without annotation are not analyzed by GOSlimViewer.

An example of an input file is given [here](#).

For help using the tool click [here](#).

An online tutorial for this tool is available [here](#).

More information about the tool can be found [here](#).

A standalone version of the tool can be found [here](#).

The screenshot shows the GOSlimViewer web interface. At the top, there is a section for uploading a GO Summary File, with a 'Choose file' button and the text 'No file chosen'. Below this is a 'Select GOSlim Set:' dropdown menu, which is currently open, showing a list of options: 'Generic' (checked), 'GOA Whole Proteome', 'Metagenomics', 'PANTHER', 'PIR', 'Plant' (highlighted in blue), 'Yeast', and 'TIGR Prokaryote (BP only)'. To the left of the dropdown is a 'Search' button. Below the dropdown, there is a note: '\*\*Note: The GO Summary i... The input file has to contain...'. Below the note, there are two example input lines: 'A0JNB7 GO:0006826' and 'A0JNB7 GO:0008021'. To the right of the dropdown, there is a label 'records are listed below:'.

**Figura 4.** GOSlim Viewer (AgBase database).

El resultado de GOSlim presenta varios formatos para el análisis. Alguno de estos formatos pueden abrirse en Excel, lo que se desaconseja encarecidamente porque produciría la pérdida instantánea de 500 puntos en el Biohackathon.

En el siguiente resultado GOSlim, se analizan algunas entradas de la ontología 'Proceso Biológico'. El ejemplo tiene 3 posibles entradas :

- desarrollo del organismo multicelular, con 1 gen (que tiene 1 anotación GO).
- proceso biológico, con 2 genes (que tienen 3 y 4 anotaciones GO, respectivamente)
- estructura anatómica de la morfogénesis, con 1 gen (1 anotación GO)



## Biological Process Accessions [Back to top](#)

<b>Slim id: GO:0007275</b>	<b>multicellular organism development</b>	<b>1</b>
Q9ZSI1	GO:0010311	lateral root formation
<b>Slim id: GO:0008150</b>	<b>biological_process</b>	<b>2</b>
F4J0W4	GO:0010821	regulation of mitochondrion organization
F4J0W4	GO:0007005	mitochondrion organization
F4J0W4	GO:0009737	response to abscisic acid
Q9ZSI1	GO:0009664	plant-type cell wall organization
Q9ZSI1	GO:0009828	plant-type cell wall loosening
Q9ZSI1	GO:0010311	lateral root formation
Q9ZSI1	GO:0071555	cell wall organization
<b>Slim id: GO:0009653</b>	<b>anatomical structure morphogenesis</b>	<b>1</b>
Q9ZSI1	GO:0010311	lateral root formation

**Figura 5.** Resultado del GOSlim Viewer (AgBase database).

El siguiente paso se trata de desarrollar una función que nos permita calcular de forma automática el número de anotaciones GO de cada entrada. Siguiendo el ejemplo, la anotación 'desarrollo del organismo multicelular' tiene 1 GO, mientras que la anotación 'proceso biológico' tiene 7 entradas. Las diferencias cuantitativas/cualitativas de las entradas al comparar varios experimentos, nos sugieren procesos biológicos, funciones moleculares y componentes celulares diferentes implicados en cada caso.

Para completar este paso, es recomendable trabajar con el formato de texto plano que nos ofrece AgBase.

GO_Type	Slim_GO_ID	Slim_GO_Name	Input_Accession	Input_GOID	Input_GO_Name
C	GO:0005575	cellular_component	F4J0W4	GO:0016021	integral component of membrane
C	GO:0005575	cellular_component	F4J0W4	GO:0005741	mitochondrial outer membrane
C	GO:0005575	cellular_component	F4J0W4	GO:0031307	integral component of mitochondrial outer membrane
C	GO:0005576	extracellular region	Q9ZSI1	GO:0005576	extracellular region
C	GO:0005618	cell wall	Q9ZSI1	GO:0005618	cell wall
C	GO:0005622	intracellular	F4J0W4	GO:0005741	mitochondrial outer membrane
C	GO:0005622	intracellular	F4J0W4	GO:0031307	integral component of mitochondrial outer membrane
C	GO:0005623	cell F4J0W4	GO:0005741	GO:0005741	mitochondrial outer membrane
C	GO:0005623	cell F4J0W4	GO:0031307	GO:0031307	integral component of mitochondrial outer membrane
C	GO:0005737	cytoplasm	F4J0W4	GO:0005741	mitochondrial outer membrane
C	GO:0005737	cytoplasm	F4J0W4	GO:0031307	integral component of mitochondrial outer membrane
C	GO:0005739	mitochondrion	F4J0W4	GO:0005741	mitochondrial outer membrane
C	GO:0005739	mitochondrion	F4J0W4	GO:0031307	integral component of mitochondrial outer membrane
C	GO:0005739	mitochondrion	F4J0W4	GO:0005739	mitochondrion
C	GO:0016020	membrane	F4J0W4	GO:0016021	integral component of membrane
C	GO:0016020	membrane	F4J0W4	GO:0005741	mitochondrial outer membrane
C	GO:0016020	membrane	F4J0W4	GO:0031307	integral component of mitochondrial outer membrane
C	GO:0016020	membrane	F4J0W4	GO:0016020	membrane

**Figura 6.** Resultados GOSlim obtenidos con AgBase.

Sobre ese archivo plano, podemos también desarrollar algunas funciones que nos digan rápidamente :



- nº genes anotados en la ontología BP ?
- nº genes anotados en la ontología MF ?
- nº genes anotados en la ontología CC ?
- nº genes sin anotación ? (eso suele ser típico de secuencias desconocidas)
- nº genes anotados con 2 ontologías ?
- nº genes anotados con 3 ontologías ?

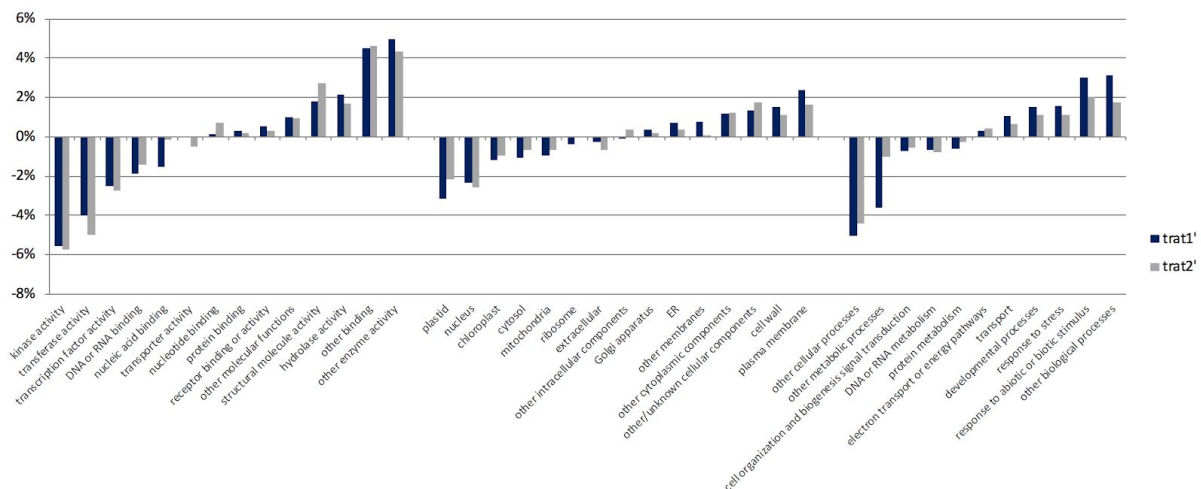
### *Etiquetas*

El último paso de la anotación consiste en **asignar** a cada anotación GOSlim una **etiqueta** que usaremos para crear figuras informativas. Las etiquetas asociadas a cada GOSlim puedes encontrarlas en este [archivo](#).

Siguiendo el ejemplo de la Figura 5 :

- desarrollo del organismo multicelular (1 anotación GO) pertenece a la etiqueta 'procesos del desarrollo'.
- proceso biológico (7 anotaciones GO) pertenece a la etiqueta 'otros procesos biológicos'
- estructura anatómica de la morfogénesis (1 anotación GO) pertenece a la etiqueta 'procesos del desarrollo'.

Para cada etiqueta, estimamos el % de anotaciones GOSlim que contiene y lo representamos mediante una figura. La representación puede hacerse de forma individual para cada tratamiento o combinando varios tratamientos. También puede representarse en figuras independientes las etiquetas de genes que se sobreexpresan por un lado, y las de los genes que se inhiben por otro, o puede combinarse todo en una misma figura como en el ejemplo siguiente. Otros ejemplos de visualizaciones típicas en anotación funcional puedes encontrarlos [aquí](#).



**Figura.** Anotación funcional en 2 tratamientos para las ontologías BP, CC y MF. En eje X se representan las etiquetas de cada ontología. En eje Y se representa el % de anotaciones GOSlim asociadas a cada etiqueta.

Es importante **pensar bien el diseño de este último archivo** porque además de permitirnos realizar una visualización general de los genes que se expresan de forma diferencial en nuestro experimento, nos va a permitir, en un momento dado (mediante el diseño de algunas funciones), localizar información relevante para seguir con nuestros estudios. Por ej., si estamos trabajando con defensas moleculares, puede interesarnos saber específicamente los genes que se han anotado con la etiqueta '*response to abiotic or biotic stimulus*'. Una vez localizados estos genes, podríamos plantear futuros experimentos para inhibir su expresión, aumentarla, o colocar ese gen detrás de un regulador que haga que se exprese en un tejido diferente, etc, ...

Otras preguntas de interés (p. ej.) :

- búsqueda por etiqueta :
  - qué genes están anotados con la etiqueta '*response to stress*'
  - qué genes están anotados con la etiqueta '*transcription factor activity*'
- búsqueda por GOSlim :
  - qué genes están anotados como '*photosynthesis*' ?
  - qué genes se asocian con '*cell wall*' ?

## ¿Qué hay que presentar (jueves 4 marzo) ?

Además de las preguntas que se han señalado en esta guía como sugerencias para analizar el experimento RNASeq, de cara a la presentación de los resultados (2-3 diapositivas) y a la definición o creación de funciones para analizarlos, podríamos seguir este esquema general :

- En general, el tratamiento de poliaminas , ¿produce alguna respuesta en la planta ?
- Si produce respuesta, se consigue mayoritariamente mediante la activación o mediante la inhibición de genes ?
- La respuesta de la planta a las poliaminas analizadas (putrescina y termoespermina), ¿es similar / diferente ?
- ¿Qué tratamiento produce una mayor regulación de genes ?

Para desarrollar este reto, proponemos que trabajéis de forma colaborativa con las herramientas de Google :

- Presentación : Google Slides
- Flujo de trabajo : combinando descripción de cada paso con código : [Google Colab](#)

### Valoración

Cuando tengáis estos 2 archivos listos, compartirlos con nosotros. Valoraremos la claridad a la hora de documentar cada paso del análisis (Google Colab) y la interpretación biológica que hagáis en vuestra presentación (Google Slides) a partir de las figuras realizadas.

**¡ Os deseamos mucha suerte !**  
**El equipo organizador.**