Framework para el análisis de datos ómicos

Rubén Sánchez Fernández

Máster universitario en Bioinformática y Bioestadística UOC-UB

Desarrollo de herramientas de soporte a la ómica

Consultor: Antonio Jesús Adsuar Gómez

Profesor responsable: Carles Ventura Royo

4 de junio de 2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Framework para el análisis de datos ómicos*
Nombre del autor:	*Rubén Sánchez Fernández*
Nombre del consultor/a:	*Antonio Jesús Adsuar Gómez*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	06/2019
Titulación:	Máster universitario en Bioinformática y bioestadística
Área del Trabajo Final:	*Desarrollo de herramientas de soporte a la ómica*
Idioma del trabajo:	*Castellano*
Palabras clave	*Ómica, análisis de datos, herramienta web*

Resumen del Trabajo (máximo 250 palabras):

Los recientes avances en las tecnologías de secuenciación masiva han simplificado el proceso de obtención de datos ómicos de forma considerable. La disponibilidad y crecimiento masivo de estos datos requiere de herramientas que simplifiquen el proceso de análisis. Con ese objetivo, una nueva herramienta es presentada en este trabajo para realizar análisis de expresión diferencial a datos procedentes de experimentos de microarray (Affymetrix) y RNA-Seq. La herramienta cubre el procedimiento habitual de análisis de expresión diferencial: exploración visual de los datos crudos, seguido de la identificación de genes diferencialmente expresados, finalizando con un proceso de anotación y análisis de enriquecimiento. Además, la herramienta permite configurar el análisis y exportar los datos. Para asegurar que puede ser utilizada por la mayoría, la herramienta ha sido diseñada con una interfaz de usuario simple e intuitiva mediante el paquete Shiny de R. La aplicación se encuentra públicamente disponible en Github (https://github.com/RubenSanchezF/TFM).

Abstract (in English, 250 words or less):

The recent advances in high-throughput technologies have made the process of obtaining omics data easier than ever. The availability and explosive growth of these data require tools that simplify the process of analysis. Aiming just that, a new tool is presented in this work to perform differential expression analysis to *Affymetrix* microarray and RNA-Seq experiments. The tool covers the usual pipeline of gene expression analysis: a pre-analysis visual exploration of the data, followed by the identification of differential expressed genes, and ending with an annotation and enrichment analysis. In addition, the tool supports analysis configuration and data export. To ensure availability to all audience, the tool has been designed with an easy-to-use user interface as a *Shiny* application. The application is freely available to download in *Github* (https://github.com/RubenSanchezF/TFM).

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Contexto y justificación del trabajo

La aparición de las tecnologías “ómicas” han permitido avances extraordinarios en los campos de la medicina y la biotecnología. Gracias al estudio de la genómica y la proteómica, entre otras, ha sido posible obtener un mejor entendimiento del funcionamiento de ciertas enfermedades.

El futuro de la medicina pasa inevitablemente por el estudio de las tecnologías ómicas. Medicina de precisión [1], mejora en el desarrollo de fármacos [2] mejora en la detección y diagnóstico de enfermedades [3], todo ello son ejemplos de los avances conseguidos gracias al estudio de los datos ómicos.

Debido a las mejoras en las tecnologías de alto rendimiento (*high-throughput*) e iniciativas como *Omics-DI* [4], la riqueza y accesibilidad de los datos ómicos aumenta cada año. De hecho, desde el año 2000 hasta este momento, la cantidad de datos ómicos disponible en la red ha aumentado alrededor de cinco órdenes de magnitud [5]. Por desgracia, la informática no avanza al mismo ritmo. Si ya existen numerosos problemas para lograr almacenar la gran cantidad de datos ómicos que se generan, mucho más para analizarlos [6]. No cabe duda, que el éxito en las ciencias ómicas pasa por la innovación y el desarrollo en infraestructura y software que permita exprimir toda la información contenida.

El análisis y entendimiento de estos datos es necesario para aprovechar la oportunidad presentada. Pero esto no es un proceso sencillo. Esta masiva cantidad de datos requiere metodologías complejas que incluyen minería de datos y técnicas estadísticas, que solo profesionales en estas áreas pueden desarrollar. No solo eso, además la complejidad de estos análisis hace difícil incluso la interpretación de los resultados, dificultando en exceso el trabajo del investigador. Este hecho descubre la necesidad de herramientas efectivas y sencillas de utilizar, que permitan al usuario realizar análisis a un conjunto de datos sin la necesidad de ser un experto bioinformático o estadístico. En la actualidad, existen unas pocas herramientas web que cumplen las

características mencionadas, como *BioStatFlow* (http://biostatflow.org) o MetaboAnalyst 4.0 [7], programas eficientes, pero con algunas limitaciones que no los hacen adaptables a cualquier situación. Por ejemplo, BioStatFlow está limitado en la diversidad de técnicas analíticas que contiene de forma predeterminada. Intenta solucionar esta limitación

permitiendo al usuario integrar nuevas técnicas mediante scripts de R (entre otros), siendo necesario que el usuario tenga conocimientos estadísticos y de programación. Por otro lado, MetaboAnalyst requiere que el usuario realice algo de preprocesado antes de cargar los datos, además al realizar ciertas técnicas de análisis es necesario que expertos comprueben los resultados para asegurar que son correctos. Esto solo son algunos ejemplos de las limitaciones de estas herramientas, herramientas eficientes y útiles en algunos casos, pero no lo suficientemente adaptables para que el investigador pueda utilizarlas en cualquier supuesto.

Entre las diferentes metodologías, el análisis de expresión diferencial (*DEG analysis*, *en inglés Differential Expression Gene* analysis) es una de las más frecuentes. La finalidad del análisis de DEG es encontrar genes cuyo nivel de expresión cambia en diferentes condiciones experimentales. Descubrimiento de genes involucrados en el cáncer de mama [8] o en la obesidad [9] son algunos de los estudios más recientes desarrollados con esta metodología.

El proceso de análisis de DEG no es sencillo. Consta de diferentes pasos que requieren conocimientos y entrenamiento previo, por lo que dificulta la accesibilidad de este tipo de estudio a cualquier investigador. Por ello, este trabajo final de máster pretende crear una aplicación simple e intuitiva que permita al investigador realizar DEG sin necesidad de tener conocimientos técnicos en bioinformática o programación.

1.2. Objetivos del trabajo

1. Diseñar una aplicación que permita realizar un *pipeline* de análisis de expresión diferencial, a conjuntos de datos de microarray (Affymetrix) y RNA-Seq. La aplicación deberá permitir al usuario:

    1.1. Introducir datos de niveles de expresión génica obtenidos con microarrays de Affymetrix o matrices de expresión (counts) obtenidos mediante RNA-Seq.

    1.2. Realizar un análisis gráfico para determinar la calidad de los datos introducidos.

    1.3. Filtrar y normalizar los datos.

1.4. Realizar un análisis estadístico para determinar aquellos genes cuyos niveles de expresión varían a partir de dos o más condiciones experimentales.

1.5. Permitir al usuario realizar un análisis de enriquecimiento a partir de los genes encontrados.

2. Se plantean también como objetivos ciertas características que deberá tener la herramienta:

2.1. **Funcionalidad**: Deberá ser capaz de realizar la función para la que está diseñada.

2.2. **Adaptabilidad**: Uno de los puntos fuertes de la herramienta deberá ser la capacidad de adaptarse a las necesidades del usuario. Para ello se diseñará para dar la máxima libertad posible al usuario. Esto, será, precisamente, lo que la diferencie del resto de soluciones ya disponibles.

2.3. **Escalabilidad**: Del mismo modo, se deberá implementar una solución lo más escalable posible. La herramienta deberá facilitar su ampliación, ya sea para añadir nuevas metodologías de análisis o nuevos tipos de dato.

2.4. **Accesibilidad**: La herramienta deberá permitir que el usuario pueda acceder a ella con facilidad.

1.3. Enfoque y método seguido

Para desarrollar la herramienta, se ha decidido utilizar el lenguaje de programación R [10] por diversos motivos:

- Se trata de un lenguaje *open-source* especialmente potente para realizar análisis estadístico y visualización gráfica de datos, dos áreas especialmente importantes en esta metodología.
- Contiene paquetes específicos para el análisis de datos ómicos, como Bioconductor [11], que facilitan el desarrollo de la herramienta.
- La herramienta Shiny [12] permite desarrollar aplicaciones web interactivas en el propio lenguaje R.

Respecto a la metodología, el *pipeline* de análisis de un estudio de DEG es un proceso bien definido, que queda detallado en los siguientes capítulos de esta memoria. Existen multitudes de técnicas estadísticas que se podrían aplicar en este análisis, e idealmente la herramienta debería implementar la mayoría de ellas, pero debido al tiempo dado para realizar este proyecto no es

un objetivo alcanzable. Por lo que, debido a que han demostrado mejores resultados en algunos estudios comparativos [13], se ha decidido comenzar implementando modelos lineales con moderación empírica de Bayes (paquete *limma* [14]) y la técnica estadística SAM (*Significance Analysis of Microarrays*).

1.4. Planificación del trabajo

La herramienta se ha diseñado en un ordenador con las siguientes características:

- Procesador: Intel Core i5-3570k @ 3.40GHz (4CPUs)
- RAM: 8 GiB DDR4
- Sistema Operativo: Windows 10 Home, 64-bit
- Versión de R: 3.5.3

Para correr la aplicación en local, es necesario que el ordenador tenga R instalado con la versión 3.5.3 y la librería *BiocManager* instalada. Con versiones anteriores, es posible que algunas de las librerías utilizadas no estén disponibles.

1.4.1. Planificación temporal del proyecto

A continuación, se enumeran las tareas realizadas en este proyecto y la duración aproximada de cada una.

Tareas	**Duración**
Análisis de la bibliografía	20h
Desarrollo de la herramienta	124h
Software testing	12h
Redacción del informe de seguimiento (Fase 1)	6h
Desarrollo de la interfaz gráfica	58h
Redacción del informe de seguimiento (Fase 2)	6h
Redacción de la memoria	56h
Elaboración de la presentación	28h
Defensa del trabajo	20h
TOTAL	330h

Tabla 1. División de tareas con su correspondiente duración.

1.4.2. Calendario

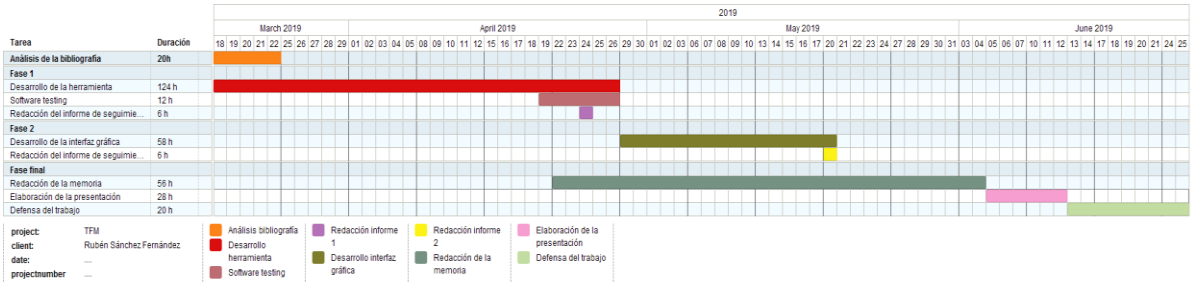


Figura 1. Diagrama de GANTT detallando el calendario del proyecto, con la división de tareas realizadas y la duración de cada una de ellas.

1.4.3. Hitos

A continuación, se detallan los hitos esperados en la elaboración del proyecto, con sus respectivas fechas.

Hitos	**Fecha**
Plan de trabajo (PEC 1)	18/03/2019
Elaboración de la herramienta e informe de seguimiento 1 (PEC 2)	24/04/2019
Cierre de la herramienta, desarrollo interfaz gráfica e informe de seguimiento 2 (PEC 3)	20/05/2019
Memoria del trabajo	04/06/2019
Presentación del trabajo	12/06/2019
Defensa pública	25/06/2019

Tabla 2. Hitos alcanzados durante el desarrollo del proyecto

1.5. Breve sumario de productos obtenidos

Los resultados obtenidos en este proyecto se dividen en:

- **Memoria del proyecto**: La presente memoria detalla los detalles del proyecto, la metodología desarrollada y los resultados obtenidos.
- **Software desarrollado**: El software se presenta en forma de aplicación desarrollada en Shiny, y permite aplicar un *pipeline* de

análisis de expresión diferencial a datos de microarrays de Affymetrix y a datos de RNA-Seq.

La aplicación se encuentra depositada en Github, y está públicamente disponible para descargar mediante el siguiente enlace:

https://github.com/RubenSanchezF/TFM

En el archivo *README.md* dentro del repositorio de Github se detalla el procedimiento para instalar y correr la aplicación. Es importante recordar que la aplicación ha sido desarrollada con la versión de R 3.5.3, por lo que no se asegura el funcionamiento con versiones de R anteriores.

1.6. Breve descripción de los otros capítulos de la memoria

El capítulo 2 se inicia con una introducción a la teoría detrás de los experimentos de microarrays de Affymetrix y a los experimentos de RNA-Seq, y al proceso de análisis de expresión diferencial. A continuación, se presenta la solución propuesta, detallando el diseño de la herramienta y su funcionamiento. Para finalizar el capítulo, se introducen dos ejemplos de uso: un ejemplo de estudio con datos de microarrays, y un segundo ejemplo con datos de RNA-Seq.

En el capítulo 3 se presentan las conclusiones del trabajo, valorando los objetivos alcanzados, y detallando los problemas encontrados durante el desarrollo del proyecto. Además, se introducen los principales aspectos a mejorar de la herramienta, a considerar para futuras líneas de trabajo.

Capítulo 2

Software para el análisis diferencial de datos ómicos

2.1. Introducción

2.1.1. Expresión génica diferencial

Se conoce como expresión génica al proceso por el cuál la información contenida en un gen es utilizada en la síntesis de un producto funcional del mismo [15]. Estos productos suelen ser proteínas, pero también pueden ser ARN funcionales.

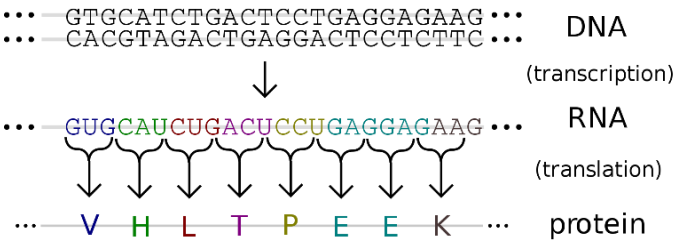


Figura 2. Proceso de síntesis proteica [15]

En los organismos pluricelulares, todas las células contienen la misma información genética, pero la síntesis de proteínas difiere en cada tipo de célula. Esto es debido a que la expresión génica está regulada por características propias de la célula y por causas externas. Respecto al contexto local de la célula, la expresión viene controlada por la presencia de intrones, que son aquellas regiones del gen que no codifican para una proteína, y exones, regiones del gen que si son codificadoras. Por ejemplo, en las células cardiacas, la localización de intrones y exones determina que se expresen

proteínas encargadas de la función cardiaca, y no otro tipo. A este proceso se le llama diferenciación celular, y es el proceso por el cuál una célula madre sufre cambios en su expresión génica hasta convertirse en una determinada célula.

También existen factores externos que pueden influir en la expresión génica, como factores ambientales, endocrinos, mutagénicos, entre otros. La presencia de enfermedades o medicamentos también pueden influir en la expresión génica, esto hace que los estudios de expresión génica diferencial sean especialmente útiles en la investigación de enfermedades.

2.1.2. Tecnologías de cuantificación de expresión génica

En 1995, se publicó el primer estudio realizando análisis de expresión diferencial a partir de datos cuantitativos del nivel de expresión génica [16], obtenidos a partir de una tecnología revolucionaria en la época, los chips (*microarrays*) de ADN. Desde ese momento, esta tecnología se convirtió en la tecnología estándar para cuantificar el nivel de expresión génica entre una o más condiciones. No fue hasta la primera década del siglo 21, que se empezó a explorar el uso de tecnologías de secuenciación de alto rendimiento (*Next Generation Sequencing*) [17] y se empezó a contemplar como reemplazo a los microarrays *de ADN*, ya que permite generar datos más completos y exactos que los microarrays*.*

Con los recientes avances técnicos desarrollados en las tecnologías de secuenciación de alto rendimiento, y la reducción del coste económico que esto ha implicado, la popularidad de esta tecnología ha incrementado exponencialmente. Para estudios de expresión génica, la tecnología estándar de secuenciación de alto rendimiento, *RNA-Seq*, proporciona ventajas técnicas en relación a los microarrays. Microarrays solo devuelven resultados de aquellas regiones para las que el chip ha sido diseñado, mientras que *RNA-Seq* abarca todo el transcriptoma sin necesidad de tener ningún conocimiento previo, por lo que permite el descubrimiento de nuevos tránscritos. Aun así, actualmente la tecnología de microarrays sigue siendo muy popular en estudios de expresión génica, debido a que económicamente sigue siendo más asequible, y a que los investigadores llevan décadas utilizándola, por lo que es una tecnología conocida con la que tienen experiencia.

Por todo ello, se ha decidido diseñar esta aplicación para analizar datos obtenidos con las dos tecnologías, datos obtenidos con *RNA-Seq* y datos obtenidos con microarrays de *Affymetrix.*

2.1.3. Microarrays de Affymetrix

Un microarray es un dispositivo, formado por una superficie de cristal, plástico o sílice, en el que se unen múltiples fragmentos que representan genes, proteínas o metabolitos, expuestos a una hibridación con determinadas moléculas diana. Mediante fluorescencia, se cuantifica la cantidad de moléculas diana en cada muestra, y los resultados son visualizados mediante un escáner.

Affymetrix, es una casa comercial que se especializa en la venta de microarrays de un color. Estos microarrays se obtienen sintetizando oligonucleótidos de 25 mer utilizando fotolitografía en una superficie de cuarzo. Este proceso consiste en utilizar ciclos de luz y oscuridad que activan y desactivan unas moléculas que se encuentran en la superficie y que permiten la unión del nucleótido deseado (Figura 3) [18].

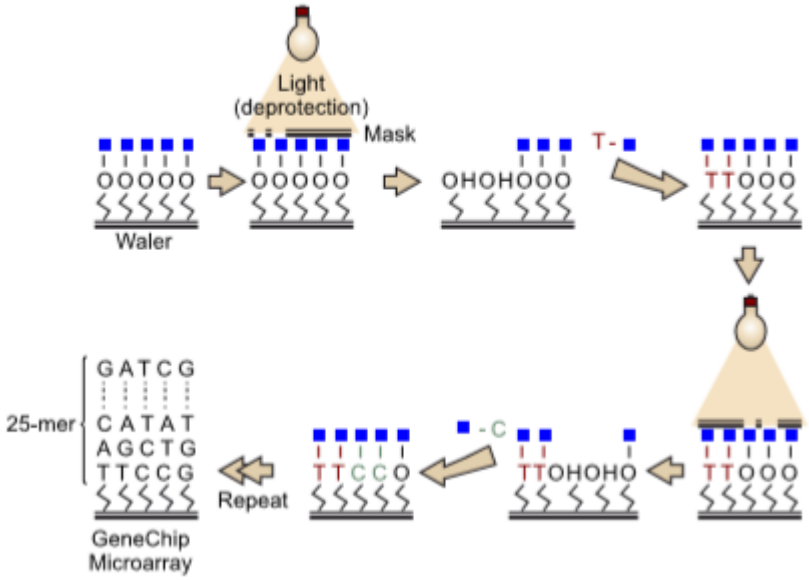


Figura 3. Proceso de fotolitografía utilizado para crear los microarrays de Affymetrix [18]

Para obtener el nivel de expresión génica, el material genético de partida es ARN. Este ARN es sometido a un proceso de transcripción y marcado mediante una molécula fluorescente. El resultado es fragmentado en partes más pequeñas para permitir la hibridación con el chip e iluminado mediante un láser. La intensidad de la fluorescencia es cuantificada mediante un escáner, y determina la cantidad de ARN que se ha unido a cada sonda [18].

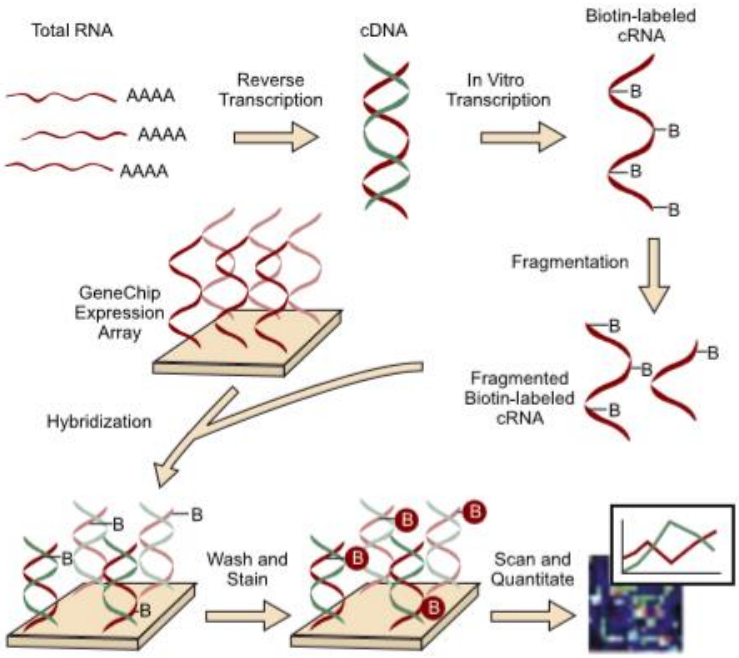


Figura 4. Proceso de cuantificación de los niveles de expresión utilizando un microarray Affymetrix [18].

2.1.4. RNA-Seq

El proceso de secuenciación mediante RNA-Seq se resume en la Figura 4. De misma forma que con los microarrays, la molécula de partida es ARN. El ARN completo, o fragmentado, es sometido a un proceso de transcripción para crear una librería de fragmentos de ADN complementario con adaptadores de secuenciación unidos a uno o ambos extremos de cada fragmento. Estos fragmentos son secuenciados mediante las tecnologías de NGS y se obtienen las secuencias, que son alineadas con un genoma de referencia para determinar si los fragmentos son fragmentos de unión, exones o poly(A). Esta clasificación se utiliza para generar un perfil de expresión para cada gen [19]. El proceso se repite para cada experimento, y al final se obtiene una matriz de conteo, estructurada en una fila por gen y una columna por experimento.

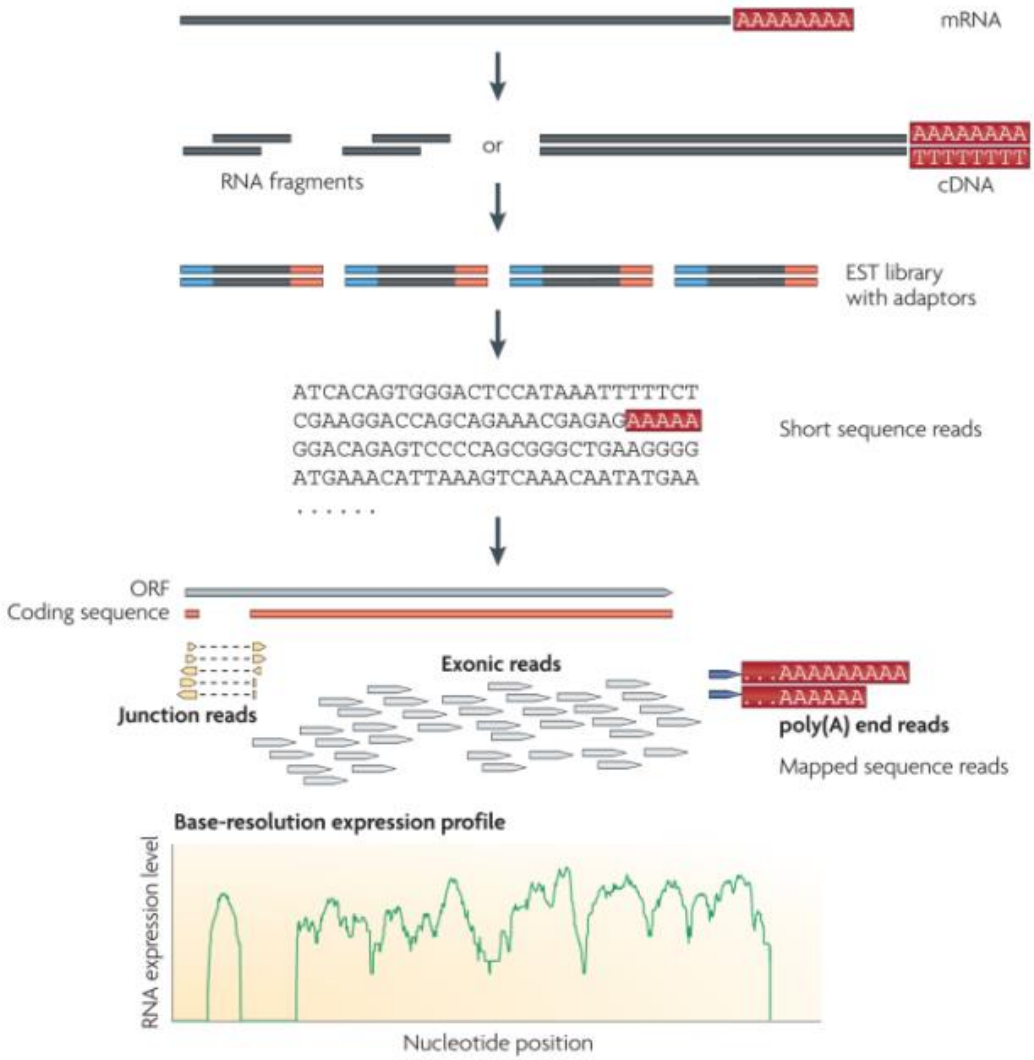


Figura 5. Proceso de cuantificación de niveles de expresión mediante RNA-Seq [19].

La matriz de conteo no solo depende del nivel de expresión génica, también depende de otros factores como la longitud del gen o el tamaño de la librería de fragmentos de ADN complementario. Por ello, es necesario un proceso de normalización para eliminar la dependencia de estos factores, y así obtener una matriz de expresión comparable.

2.1.5. Análisis de expresión diferencial en datos ómicos

Un paso fundamental en el análisis de los niveles de expresión génica es detectar aquellos genes cuya expresión cambia de acuerdo a variaciones fenotípicas o

experimentales. El proceso de expresión génica funciona como un sistema coordinado, por lo que los niveles de expresión normalmente no son independientes. Sin embargo,

debido a que la alta dimensión de los niveles de expresión dificulta una exploración comprensible, y a que todavía no disponemos de un entendimiento completo del funcionamiento de los sistemas biológicos, el proceso comienza con un análisis gen a gen, ignorando las posibles dependencias entre los genes [20].

El proceso de selección de genes diferencialmente expresados (DEG) consiste en realizar una comparación estadística entre las diferentes condiciones experimentales, para encontrar aquellos genes cuyo nivel de expresión difiere de forma significativa. Típicamente, un análisis de DEG suele dividirse en las siguientes etapas [18]:

1. **Control de calidad de los datos**: Estudio de la relación y estructura de los datos mediante resúmenes numéricos y gráficos para determinar si los datos son válidos para el estudio o presentan errores que limitan el resultado. Para este proyecto, se ha decidido basar este control de calidad en un examen visual mediante los siguientes gráficos:

2. **Preprocesado**: Esta etapa suele consistir en la aplicación de una serie de técnicas de normalización y filtrado para transformar los datos con la finalidad de permitir la comparación entre las distintas muestras.

3. **Identificación de genes diferencialmente expresados:** Proceso que conlleva la aplicación de modelos estadísticos para encontrar aquellos genes cuya expresión cambia significativamente entre dos o más condiciones experimentales.

4. **Análisis de los resultados**: Típicamente, el estudio suele finalizar con un proceso de anotación de los resultados, visualización de estos y en algunos casos un análisis básico de enriquecimiento.

En el siguiente apartado, se discutirá con detalle las técnicas escogidas para cada etapa en el diseño de la aplicación

2.1.6. Algoritmos para el análisis de expresión diferencial

En la actualidad, existen multitud de algoritmos y técnicas para encontrar genes diferencialmente expresados. Utilizar una técnica u otra puede ser

determinante en el resultado del análisis. Algunos genes pueden ser clasificados como genes diferencialmente expresados por la mayoría de las técnicas, mientras unos pocos solo lo son por una o pocas técnicas.

Basándonos en la literatura [21][22], utilizar modelos lineales no solo proporciona excelentes resultados, además aporta la flexibilidad de analizar experimentos con más de dos grupos, y con más de una fuente de variabilidad. Además, la técnica de modelos lineales es aplicable tanto a experimentos de microarrays como a experimentos de RNA-Seq. Por todo ello, es interesante incorporar esta metodología en la herramienta diseñada.

La segunda metodología de análisis de expresión diferencial implementada en la herramienta es SAM (*Significance Analysis of Microarrays*). Desde su aparición [23], se ha convertido en una de las metodologías más populares (más de 12500 citaciones del artículo original, en mayo de 2019). SAM asigna un valor de variación a cada gen basado en la desviación estándar, y calcula el *False Discovery Rate* (FDR) mediante permutaciones. El concepto es muy diferente al de modelos lineales, por eso su inclusión en la herramienta es interesante.

Modelos lineales para microarrays

La aplicación de modelos lineales para estudiar los niveles de expresión en datos de microarrays fue propuesto por primera vez por *Smyth* [24][25], ajustando un modelo lineal de los datos de expresión para cada gen asumiendo un valor común de correlación entre replicaciones. En su propuesta, el posible *bias* que se obtiene estimando un valor común de correlación se corrige estimando la varianza entre genes. Además, aplicando una estimación empírica de Bayes permite hacer inferencia en las varianzas incluso con experimentos con pocas replicaciones. Esta mezcla de análisis gen a gen, combinada con estimaciones de parámetros a nivel global, proporciona resultados muy fiables en la mayoría de los experimentos.

A grandes rasgos, la metodología se divide en estos tres pasos [18]:

1. Se propone una solución representando el experimento mediante modelos lineales suponiendo que las variables serán estimadas a partir de la información común de todos los genes.
2. Se ajusta el modelo y se estiman sus parámetros.
3. Se calcula la probabilidad de que un gen esté diferencialmente expresado mediante un *odds-ratio* (O), asociado a un estadístico t-moderado y a su p-valor.

Sea M_{ij} el nivel de expresión del gen *i* y en la réplica *j*:

$$B_{ij} = \log O_{ij} = \log \frac{P[Afectado \mid M_{ij}]}{P[No\ Afectado \mid M_{ij}]}$$

donde B es el estadístico que representa la probabilidad de que un gen esté diferencialmente expresado.

El propio *Smyth* ha desarrollado el paquete *limma* [26] que permite aplicar esta metodología en R.

Modelos lineales para RNA-Seq

El paquete *limma* incorpora varios *pipelines* de análisis para RNA-Seq mediante modelos lineales. Entre ellos, se encuentra la combinación *voom-limma* que ha demostrado proporcionar excelentes resultados [22]. Podemos dividir esta metodología en estos tres pasos:

1. Transformación de los datos *count* normalizados a escala logarítmica.
2. Estimación de la relación media-varianza de forma empírica mediante el método *voom* [27], calculando un "peso" de esta relación para cada observación.
3. Los pesos calculados son incorporados en los modelos lineales, para finalmente realizar el proceso descrito en la metodología para microarrays.

Convenientemente, el basar la metodología de análisis de microarrays y de RNA-Seq en un mismo paquete de R, permite obtener los resultados estadísticos en el mismo formato y las mismas representaciones gráficas.

Ventajas e inconvenientes del modelo lineal

El principal motivo por el que se ha escogido el análisis mediante modelos lineales en este proyecto es la flexibilidad que aporta: esta metodología permite adaptarse a situaciones muy diferentes y complejas sin perder eficacia. Además, mediante la teoría estándar del modelo lineal [28], es posible hacer inferencia de los resultados sobre la totalidad de la población.

Uno de los principales inconvenientes del modelo lineal es la perdida de eficacia en muestras pequeñas. El paquete *limma* hace un buen trabajo en mitigar este problema, utilizando el método paramétrico empírico de Bayes para hacer inferencia [29], de esta forma es capaz de proporcionar buenos resultados incluso con muestras pequeñas. Aun así, se ha decidido implementar SAM, como alternativa al modelo lineal, para proporcionar mayor libertad al usuario a la hora de escoger la metodología de análisis.

Significance analysis of microarrays (SAM)

SAM fue descrito por primera vez por *Tusher, Tibshirani* y *Chu* [23] en 2001, como alternativa a los métodos tradicionales basados en estadísticos *t*. Estos métodos tradicionales no contaban con la posibilidad de detección de falsos positivos, y en experimentos con microarrays, incluso con una probabilidad muy baja, resultados con falsos positivos pueden ser muy significativos. Por ejemplo, una probabilidad de falsos positivos de 0.01, en un conjunto de datos de 10.000 genes, implicaría que el estadístico detectaría 100 genes como significativos sin serlo.

La metodología SAM se basa en el cálculo de una diferencia media del nivel de expresión de los genes en función de la desviación estándar, llamada diferencia relativa *d.* El parámetro *d* está definido por:

$$d(i) = \frac{\widehat{x_I}(i) - \widehat{x_U}(i)}{s(i) + s_0}$$

donde $\widehat{x_I}(i)$ y $\widehat{x_U}(i)$ son los niveles de expresión medios para el gen *i* en las condiciones I y U, s es el valor acumulado de desviación estándar y s_0 es una constante para minimizar el coeficiente de variación.

Para encontrar los genes significativos, por un lado se calculan los valores de *d* y se ordenan los genes en función de la magnitud de ese valor. Se escoge un valor límite (threshold), y aquellos valores por encima de ese límite serán considerados candidatos a genes estadísticamente significativos.

Por otro lado, se realizan permutaciones aleatorias para calcular valores de *d* que se utilizan como “control”. Con los valores de *d* por gen, y los valores de *d* de las permutaciones, se estima el valor de FDR para detectar falsos positivos.

Es importante considerar que SAM es especialmente sensible a la variabilidad de los datos. Por lo que procedimientos de filtrado, que normalmente preceden al análisis estadístico, pueden hacer variar considerablemente el resultado de SAM [30].

2.2. Diseño de la herramienta

La aplicación ha sido diseñada siguiendo el *pipeline* discutido en el apartado anterior. Se ha estructurado el código en funciones que desarrollan cada parte de la metodología y se ha diseñado una aplicación con Shiny para llamar a cada función en medida que el usuario interactúa con las distintas ventanas de la aplicación. Este diseño funcional permite ampliar y modificar la herramienta de forma simple.

En la figura 6, se esquematiza el diseño de la herramienta. Cada número simboliza la interacción de una de las funciones creadas para desarrollar la metodología.

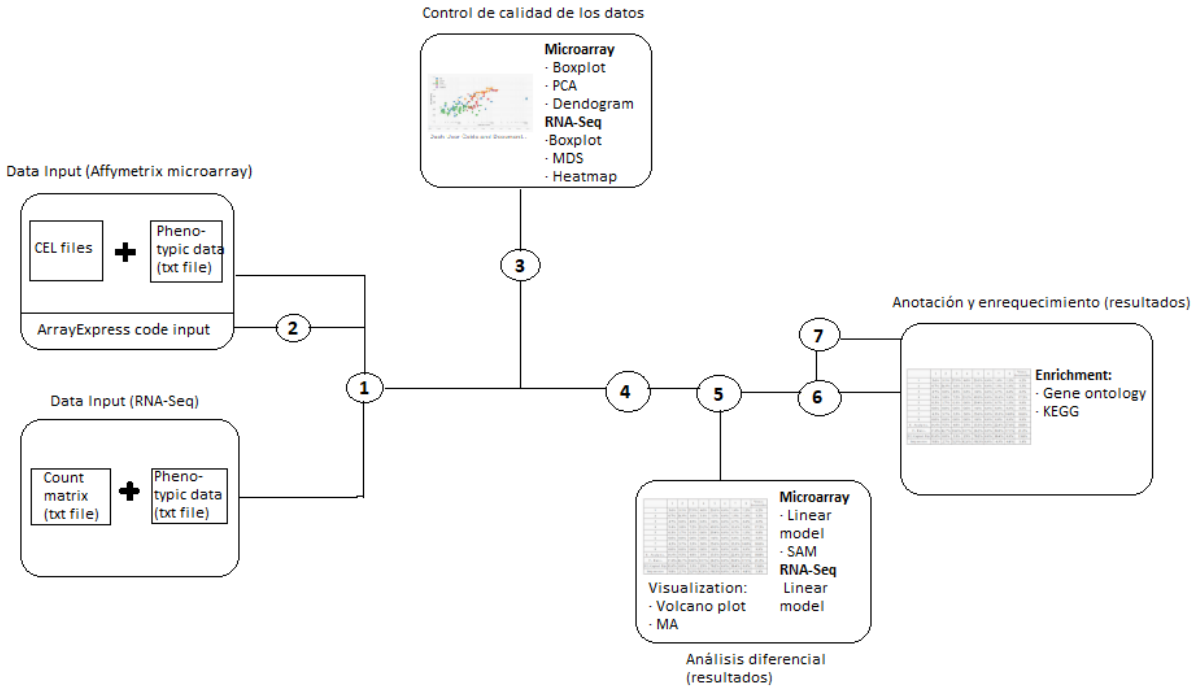


Figura 6. Esquema del diseño de la herramienta

2.2.1. Descripción de las funciones

A continuación, se detallan cada una de estas funciones y el proceso que llevan a cabo.

- Función **reading_files()**: Esta función carga y procesa los datos introducidos por el usuario (Círculo 1, figura 6). Su funcionamiento varía dependiendo de si el usuario quiere realizar análisis a microarrays (Affymetrix) o RNA-Seq. Para analizar datos de microarrays, la función acepta un archivo *ZIP* conteniendo todos los

archivos *CEL*, archivos que devuelve el software de Affymetrix. En el caso de RNA-Seq, la función acepta un archivo de texto conteniendo la matriz de datos de conteo. En ambos casos, el usuario debe introducir un archivo de texto conteniendo los detalles del experimento (nombre de los archivos, condiciones experimentales…) necesario para poder realizar el análisis. Las características requeridas que debe tener este archivo vienen especificadas en la propia aplicación.

- Función **reading_ae()**: Permite acceder a la base de datos de *ArrayExpress* (https://www.ebi.ac.uk/arrayexpress/) mediante el código de acceso y cargar

  y procesar los datos (Círculo 2, figura 6). Esta función es exclusiva para experimentos con datos de Affymetrix.

- Función **exploratory_analysis()**: Esta función procesa la salida de las funciones reading_files() y reading_ae() y crea 5 tipos de gráficos diferentes que permiten comprobar la calidad de los datos (Círculo 3, figura 6). La función crea un diagrama de cajas (*boxplot*), un gráfico de PCA (*Principal Component Analysis*) y un dendograma para los datos de microarray. Para los datos de RNA-Seq, la función devuelve un boxplot, un gráfico MDS (*Multidimensional Scaling*) y un mapa de calor (*heatmap*).

- Función **normalization()**: Función que aplica distintas técnicas de normalización en función de la selección del usuario (Círculo 4, figura 6). La función tiene definido como parámetro la selección del método de normalización. Para los datos de microarray, RMA (*Robust Multiarray Averaging*), GC-RMA (*GeneChip RMA*) o MAS 5.0. (normalización de Affymetrix). Para los datos de RNA-Seq, las técnicas seleccionadas son TMM (*Trimmed Mean of M-values*) o RLE (*Relative Log Expression*), ya que son las más extendidas [31].

- Funciones **analysis()** y **analysis_rnaseq()**: Funciones que procesan la salida de la función normalization() y aplican una serie de técnicas estadísticas para encontrar los genes que cambian en función de distintas condiciones experimentales (Círculo 5, figura 6). La función analysis() (específica microarrays) permite aplicar modelos lineales con moderación empírica de Bayes o la técnica SAM. Además, la función computa automáticamente todos los contrastes posibles introduciendo la columna que contiene las condiciones experimentales, y las condiciones que se quieren contrastar. Esta función devuelve (para cada contraste) las tablas con los genes

seleccionados, y un *volcano plot* para visualizar los resultados. También es posible especificar el valor de FDR (*False Discovery Ratio*) deseado al computar los resultados. En el caso de seleccionar la técnica SAM, la función devuelve la tabla con los resultados y un *Q-Q plot.*
La función analysis_rnaseq() realiza el mismo proceso para experimentos de RNA-Seq. Esta función permite aplicar modelos lineales con moderación empírica de Bayes y devuelve las tablas con los genes seleccionados y un gráfico MA, para cada contraste. También permite especificar el valor de FDR.

- Función **gene_annotation()**: Función específica para datos de Affymetrix. A partir de las tablas de resultados obtenidas con las funciones de análisis, procesa los identificadores de los genes y encuentra los nombres comunes de estos (Círculo 6, figura 6). Como parámetro, la función requiere introducir el nombre de la base de datos del chip utilizado. La función devuelve la misma tabla introducida, con la columna de los nombres de los genes añadida.

- Función **enrichment()**:Función específica para datos de Affymetrix. Permite realizar un simple análisis de enriquecimiento a partir de las tablas con los genes seleccionados. La función devuelve una tabla con los términos de las funciones más frecuentes (*Gene Ontology*) y una tabla con los términos de vías metabólicas más frecuentes (*KEGG*).

El proceso de anotación y enriquecimiento para los datos de RNA-Seq ha sido incluido directamente en la función analysis_rnaseq() debido a incompatibilidades con la librería Shiny.

Además de las funciones principales, también se han implementado funciones para facilitar la metodología desarrollada o la aplicación. La función library_checking() comprueba que todas las librerías necesarias están instaladas en el sistema al iniciar la aplicación. En el caso que no lo estén, las instala. La función get_annotation() es un diccionario que traduce el nombre común de los organismos, al nombre de las bases de datos que contienen su información. Por último, la función html_functions() contiene algunas funciones para mejorar el diseño de la aplicación.

2.2.2 Diseño de la aplicación

La aplicación ha sido diseñada con el objetivo de conseguir una herramienta simple y sencilla de utilizar, a la vez aportando el máximo de configuración posible para adaptarse a las necesidades del usuario. Para poder seguir el proceso de la metodología presentada en los apartados anteriores, la aplicación se estructura en 4 paneles distintos: **Configuration**, **Exploratory analysis**, **Differential analysis**, y **Annotation and enrichment**. La interfaz está configurada de forma dinámica para que cambie en función de la selección del tipo de dato: Affymetrix o RNA-Seq, por lo que cada panel cambia ligeramente en función del tipo de experimento que el usuario quiere analizar.

Panel de configuración (Configuration)

El primer panel es el encargado de configurar la carga de datos. La interfaz para la carga de datos de Affymetrix permite cargar los datos en forma de archivo .zip y un archivo de texto con la información experimental, o introduciendo el código de acceso de la base de datos de ArrayExpress (Figura 7). Al cargar datos propios, una parte crítica es que el archivo con la información experimental cumpla el formato específico: la información debe estar estructurada en forma de tabla, donde cada fila corresponde a un experimento. La primera columna de la tabla debe ser una columna llamada “Name” especificando el nombre del experimento, y además la tabla debe contener una columna

llamada “FileName” especificando el nombre del archivo de cada experimento. Esta información es requerida para crear los gráficos exploratorios que se presentan en el siguiente panel. Como puede verse en la figura 6, esta información viene detallada en la propia aplicación, mostrando además un ejemplo de la tabla requerida.

En el caso de cargar datos mediante la base de datos ArrayExpress, una vez los datos son cargados se muestra en pantalla el archivo con la información experimental, para permitir al usuario comprobar la información contenida.

La interfaz para la carga de datos de RNA-Seq está estructurada de una forma similar. Permite la subida de un archivo de texto con la matriz de datos *count* del experimento y un archivo de texto con la información experimental. De misma forma que con los datos de Affymetrix, la información experimental debe ser estructurada en forma de tabla. Un ejemplo es presentado en esta interfaz para asegurar la correcta introducción de esta información (Figura 8).

Una vez se han cargado los archivos necesarios, el usuario debe clicar en el botón “Load the data” y esperar a que aparezca el mensaje: “Data

loaded!”. Este mensaje indica que los datos han sido cargados correctamente y ya es posible navegar por los paneles siguientes.

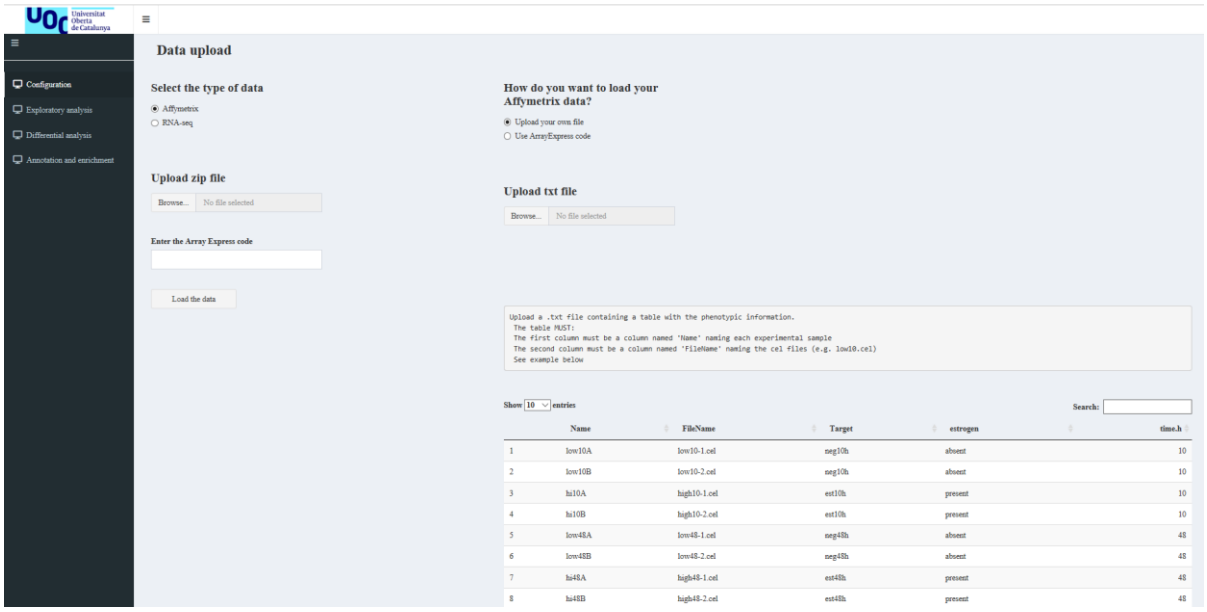


Figura 7. Interfaz del panel "Configuration" para la carga de datos de Affymetrix.

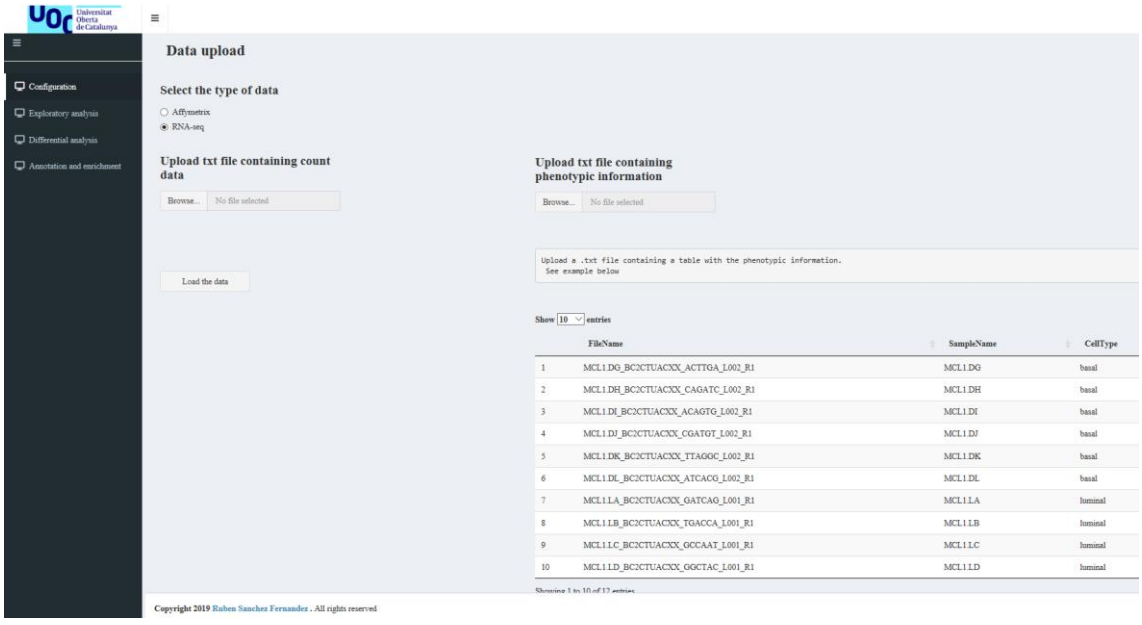


Figura 8. Interfaz del panel "Configuration" para la carga de datos de RNA-Seq.

Panel de análisis exploratorio (Exploratory analysis)

Una vez el usuario ha cargado los datos, accediendo al panel "Exploratory analysis" puede cargar y visualizar una serie de representaciones gráficas para estudiar la calidad de estos. Un panel selector permite al usuario seleccionar el gráfico que se presenta en pantalla. En el caso de la interfaz para estudios con datos de Affymetrix, el usuario simplemente tiene la opción de clicar en el botón "Show" y la aplicación crea automáticamente un gráfico *boxplot*, un gráfico de PCA y un dendograma (Figura 9).

La interfaz para estudios de RNA-Seq mantiene la misma estructura, pero contiene un selector adicional para seleccionar la columna del archivo de información experimental que contiene las condiciones que el usuario quiere contrastar. Los gráficos creados para analizar los datos de RNA-Seq son un *boxplot*, un gráfico MDS (Multidimensional scaling) y un *heatmap* (Figura 10).

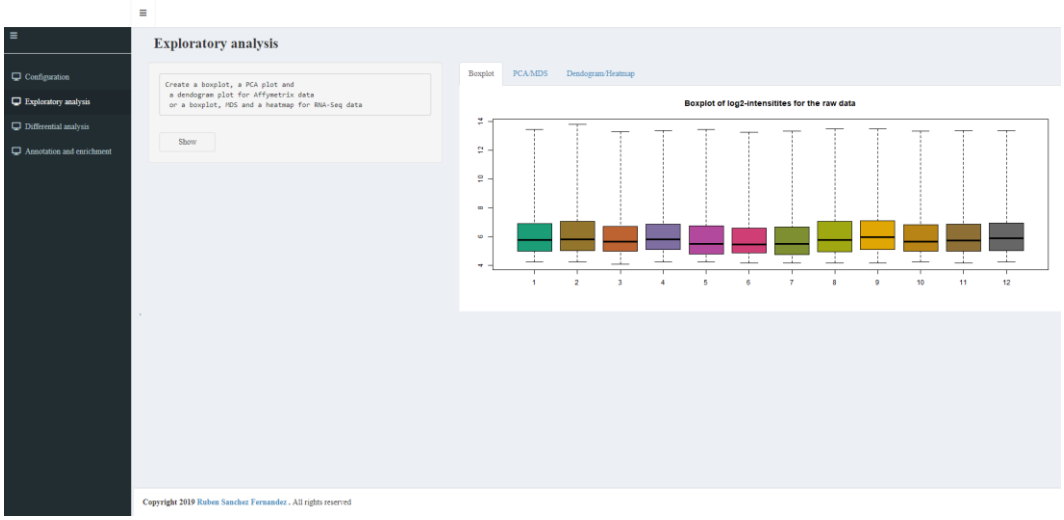


Figura 9. Interfaz del panel "Exploratory analysis" para estudios de Affymetrix.

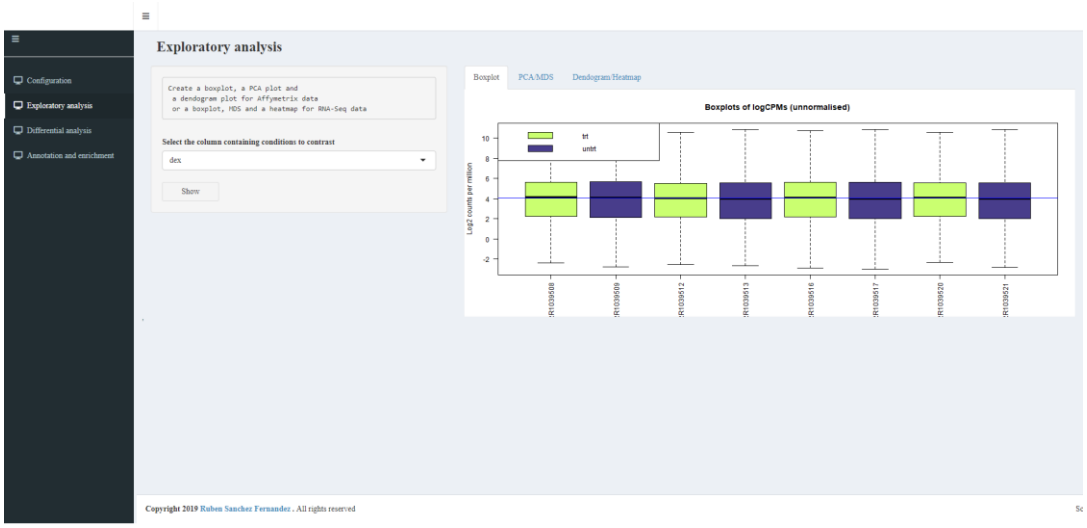


Figura 10. Interfaz del panel "Exploratory analysis" para estudios de RNA-Seq.

Panel de análisis diferencial (Differential analysis)

Una vez el usuario ha determinado si los datos son de suficiente calidad, accediendo al panel “Differential analysis” puede configurar y lanzar el análisis estadístico para encontrar los genes diferencialmente expresados. Mediante esta interfaz, el usuario puede seleccionar el tipo de normalización a aplicar, la técnica estadística, seleccionar la columna conteniendo las condiciones a contrastar, seleccionar las condiciones a contrastar, y seleccionar el valor de FDR (Figura 11). Una vez se han seleccionado los detalles del análisis, clicar en el botón “Calculate” inicia el proceso de análisis. Como se ha detallado en las secciones previas, los resultados son presentados en forma de tabla, y en forma gráfica. El usuario puede descargar la tabla en formato *.csv* clicando en el botón “Download results”. Además, debido a que el software computa automáticamente los resultados para todos los contrastes posibles a partir de las condiciones que selecciona el usuario, en la interfaz podemos encontrar un selector para seleccionar los resultados de cada contraste (Figura 12).

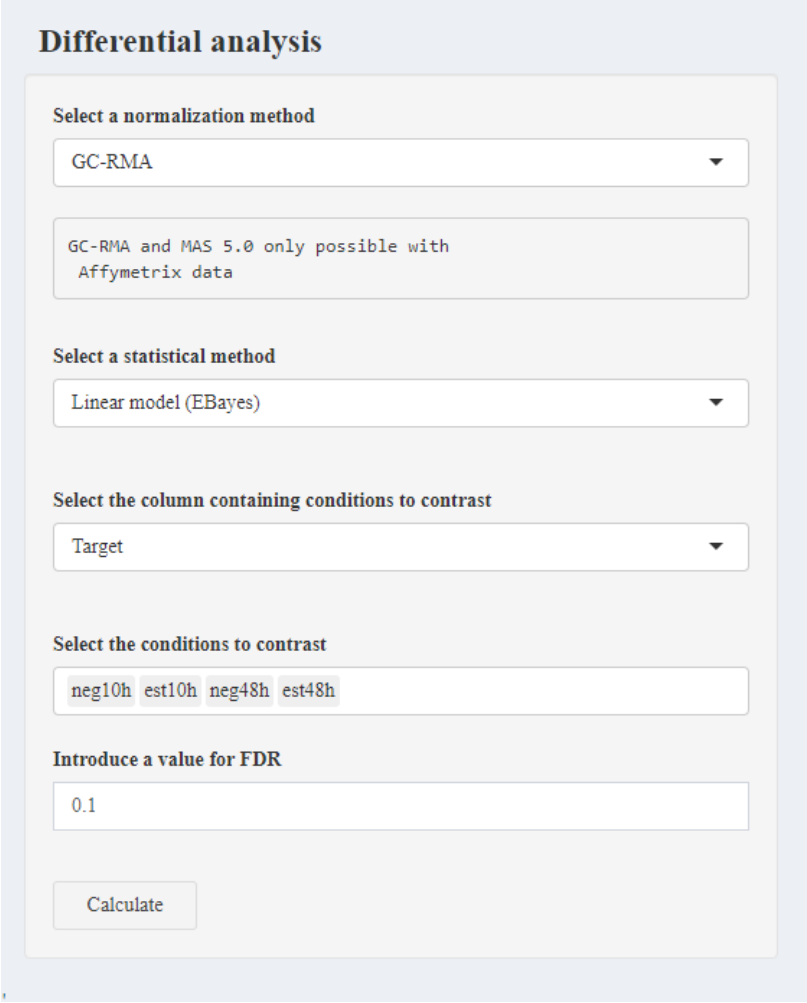
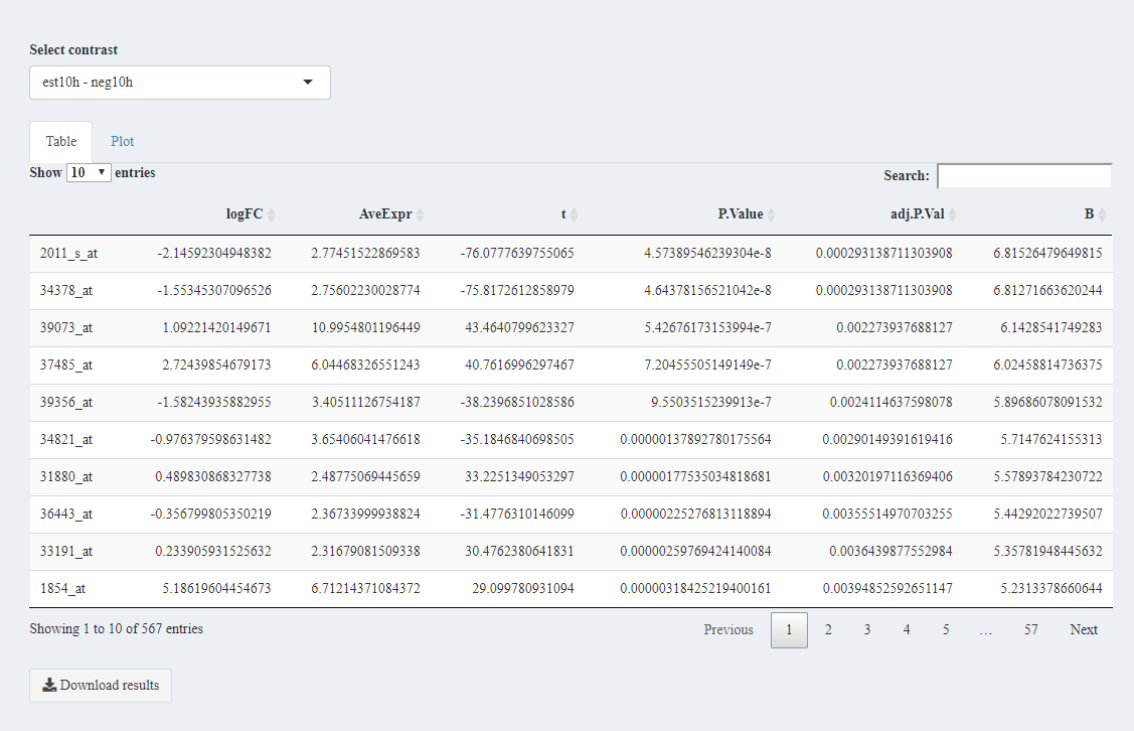


Figura 11. Parámetros para configurar el análisis en el panel “Differential analysis”


Select contrast

est10h - neg10h

Table Plot

Show 10 entries Search:

	logFC	AveExpr	t	P.Value	adj.P.Val	B
2011_s_at	-2.14592304948382	2.77451522869583	-76.0777639755065	4.57389546239304e-8	0.000293138711303908	6.81526479649815
34378_at	-1.55345307096526	2.75602230028774	-75.8172612858979	4.64378156521042e-8	0.000293138711303908	6.81271663620244
39073_at	1.09221420149671	10.9954801196449	43.4640799623327	5.42676173153994e-7	0.002273937688127	6.1428541749283
37485_at	2.72439854679173	6.04468326551243	40.7616996297467	7.20455505149149e-7	0.002273937688127	6.02458814736375
39356_at	-1.58243935882955	3.40511126754187	-38.2396851028586	9.5503515239913e-7	0.0024114637598078	5.89686078091532
34821_at	-0.976379598631482	3.65406041476618	-35.1846840698505	0.000001378927801755 64	0.00290149391619416	5.7147624155313
31880_at	0.489830868327738	2.48775069445659	33.2251349053297	0.000001775350348186 81	0.00320197116369406	5.57893784230722
36443_at	-0.356799805350219	2.36733999938824	-31.4776310146099	0.000002252768131188 94	0.00355514970703255	5.44292022739507
33191_at	0.233905931525632	2.31679081509338	30.4762380641831	0.000002597694241400 84	0.0036439877552984	5.35781948445632
1854_at	5.18619604454673	6.71214371084372	29.099780931094	0.000003184252194001 61	0.00394852592651147	5.2313378660644

Showing 1 to 10 of 567 entries Previous 1 2 3 4 5 … 57 Next

Download results

Figura 12. Presentación de los resultados en la interfaz "Differential analysis"

Panel de anotación y análisis de enriquecimiento (Annotation and enrichment analysis)

Después de obtener los genes estadísticamente significativos, el usuario puede acceder al panel “Annotation and enrichment analysis” para añadir a la tabla de resultados el nombre común de los genes y para realizar el análisis de enriquecimiento.

La interfaz para experimentos de microarray (Figura 13) permite al usuario introducir el nombre del paquete de anotación del chip de microarray utilizado en el experimento (se puede acceder al enlace de la página de Bioconductor para consultar el nombre desde la propia aplicación), permite seleccionar el contraste para generar los resultados de anotación y enriquecimiento, y en el caso de querer realizar enriquecimiento permite introducir la especie y el p-valor.

En el caso de la interfaz para experimentos de RNA-Seq, el usuario puede seleccionar el contraste y en el caso de querer realizar el análisis de enriquecimiento dispone del campo para introducir el nombre de la especie y el p-valor deseado (Figura 14).

Para exportar los resultados, en la interfaz se disponen de los botones respectivos para descargar la tabla con los resultados de anotación, la tabla con los resultados para GO y la tabla con los resultados para KEGG, en formato *.csv*.

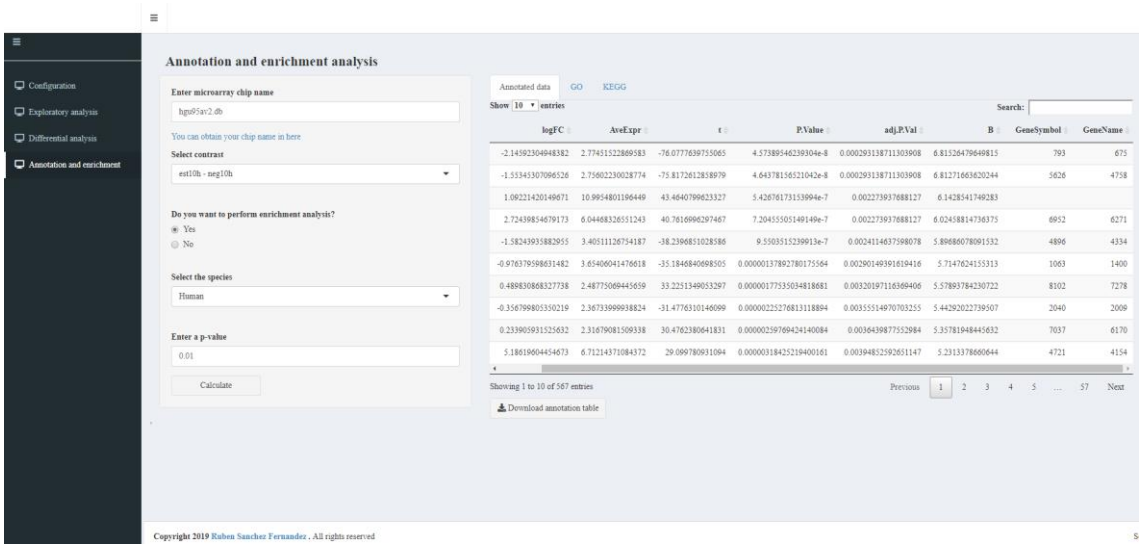


Figura 13. Interfaz del panel "Annotation and enrichment" para experimentos de Affymetrix.

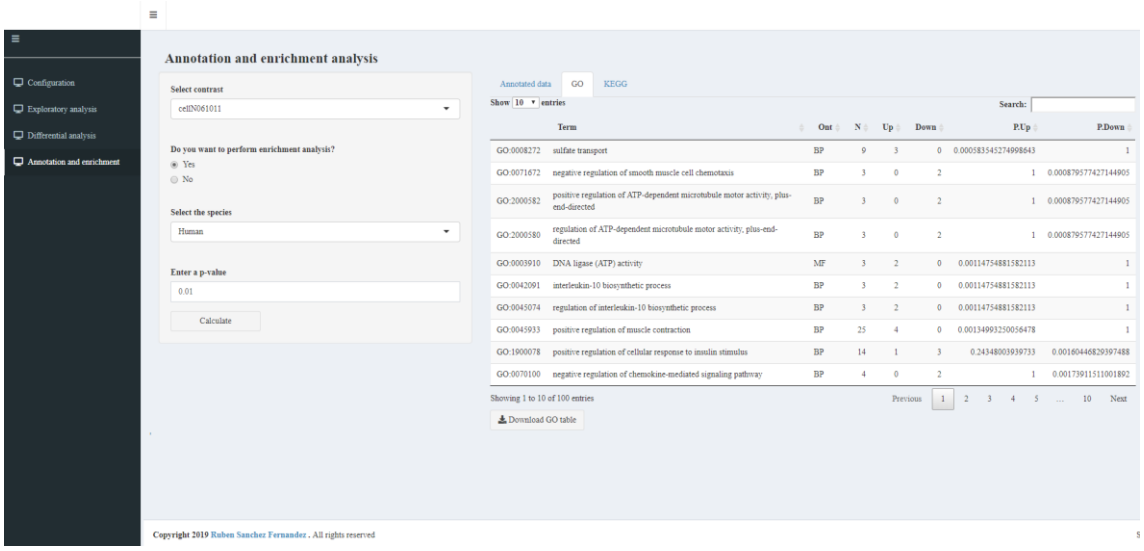


Figura 14. Interfaz del panel "Annotation and enrichment" para experimentos de RNA-Seq.

2.3. Ejemplo de aplicación de la herramienta

A continuación, se presentarán dos casos ejemplo de estudios de análisis diferencial mediante la aplicación desarrollada. Se demostrará su funcionamiento mediante un estudio con datos de microarray de Affymetrix y un estudio con datos de RNA-Seq. Los datos utilizados en los análisis se pueden encontrar en la carpeta *datasets* en el repositorio de Github.

2.3.1. Ejemplo microarrays: Estudio del efecto de la Camptotecina en los niveles de expresión del genoma humano

Los datos utilizados en esta demostración fueron publicados en 2004 por *Carson et al* [32]. Estos datos contienen los perfiles de expresión génica de cultivos de células HeLa, un subconjunto de células tratadas con camptotecina y otro subconjunto sin tratamiento. Los datos fueron obtenidos extrayendo el RNA de las células y realizando una hibridación con chips de Affymetrix Human Genome U133A conteniendo sondas para 22,283 tránscritos. La intensidad de señal fue cuantificada utilizando el software MAS 5.0 (Affymetrix).

La camptotecina es un fármaco citotóxico que actúa inhibiendo la enzima nuclear topoisomerasa I, enzima clave en el proceso de compactación y descompactación del ADN. Se ha demostrado que la camptotecina es un fármaco muy potente en el tratamiento de quimioterapia, con especial efectividad contra cáncer de ovario, cáncer de cuello uterino, entre otros [33].

Debido a su efectividad, y a su extendido uso, es necesario un análisis completo del transcriptoma después del tratamiento con camptotecina, y así comprobar el efecto que tiene esta medicina en el nivel de expresión génica.

Carga de datos

Comenzamos el análisis accediendo al panel “Configuration” y cargando el archivo *.zip* conteniendo los archivos *.CEL* y cargando el archivo de texto con la información experimental. Una vez hemos cargado los archivos, debemos clicar el botón “Load the data” y esperar a que aparezca el mensaje “Data loaded!”, indicando que el proceso ha sido exitoso (Figura 15).

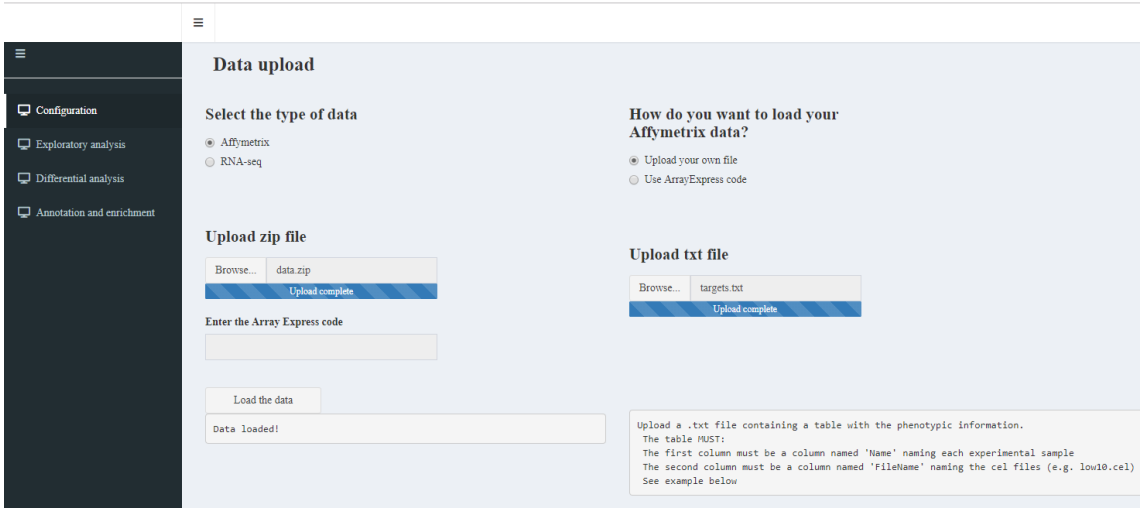


Figura 15. Cargamos el archivo .zip y el archivo .txt, clicamos en el botón "Load the data" y esperamos al mensaje "Data loaded!" indicando que el proceso ha sido exitoso.

Análisis exploratorio

El siguiente paso es comprobar que los datos no presentan inconsistencias que dificulten el proceso de análisis. Para ello accedemos al panel “Exploratory analysis” y visualizamos los gráficos exploratorios.

Con el gráfico *boxplot* estudiamos la distribución de los datos y comprobamos si existen diferencias importantes entre los distintos arrays (Figura 16).

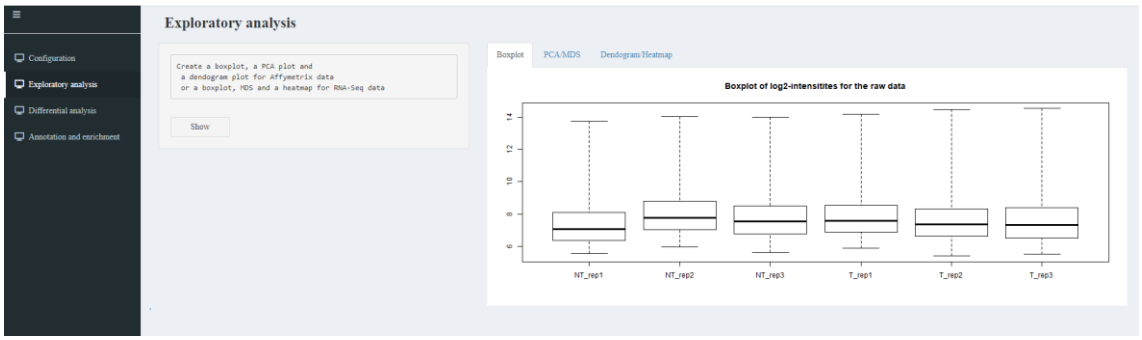


Figura 16. Boxplot para comprobar la distribución de los datos.

Con el gráfico de PCA podemos comprobar si las muestras se agrupan en función de su grupo o no siguen una distribución clara (Figura 17). Este gráfico suele ser muy indicativo de problemas técnicos, como por ejemplo el efecto *batch*.

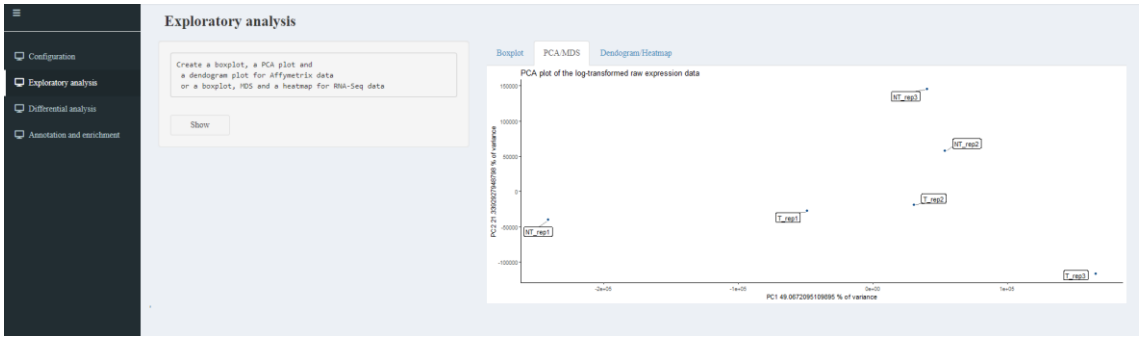


Figura 17. Gráfico de componentes principales para detectar posibles fuentes de variabilidad debido a problemas técnicos.

Por último, visualizamos como se agrupan los datos en base a un clúster jerárquico (Figura 18). De misma forma que en el gráfico de componentes principales, si las muestras se agrupan en base a una condición es indicativo de buena calidad de los datos.

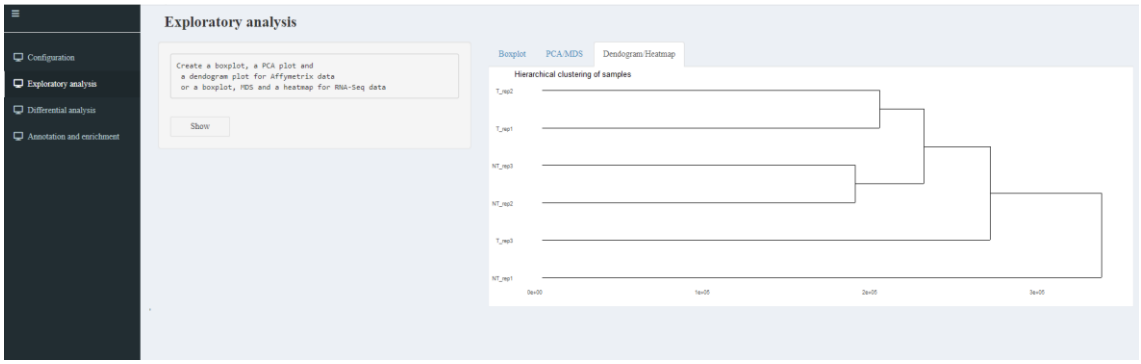


Figura 18. Dendograma para estudiar la agrupación de los datos en base a un clúster jerárquico.

En este ejemplo, podemos ver como una de las muestras (“NT_rep1”) se diferencia de forma considerable del resto de muestras. Esto podría ser indicativo de algún problema técnico en ese array que determina esta variabilidad.

Al final, es siempre decisión del investigador si en base a estos gráficos, desea continuar con el análisis, o eliminar primero esta muestra de los datos. En este caso, para continuar con la demostración, mantendremos ese array y continuaremos el análisis.

Análisis diferencial

Accedemos al panel “Differential analysis” y seleccionamos las opciones de análisis que nos interesen. En este caso, realizaremos un análisis mediante modelos lineales normalizando con GeneChip RMA. Por lo tanto, seleccionamos “GC-RMA” como método de normalización y seleccionamos “Linear model (EBayes)” como técnica estadística. En el selector siguiente seleccionamos el nombre de la columna en el archivo de texto conteniendo las condiciones experimentales, en este caso “Target”, y como condiciones seleccionamos “Treated” y “Untreated”. Por último, el valor de FDR lo mantendremos en 0.1. Una vez hemos configurado el análisis, clicamos en el botón “Calculate” (Figura 19).

Al clicar este botón, aparece el mensaje “Calculating…” indicando que se están calculando los resultados.

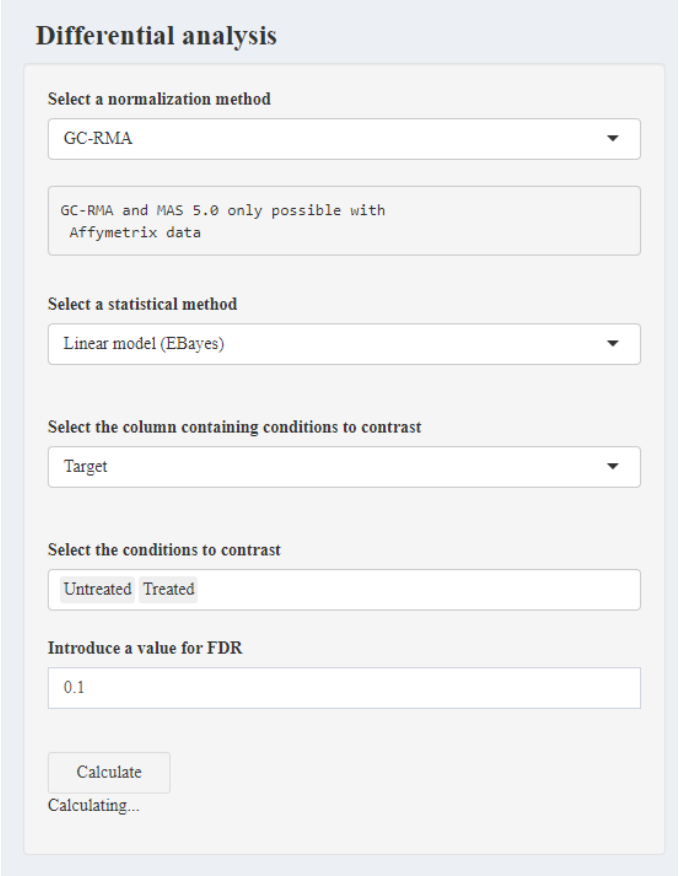


Figura 19. Configuración de análisis escogida en el ejemplo 1,

Una vez el proceso ha finalizado, se presenta una tabla con el resultado del análisis estadístico para cada gen (Figura 20) y un gráfico *volcano* remarcando los 10 genes más significativos (Figura 21).

Select contrast

Treated - Untreated

Table | Plot

Show 10 entries | Search:

	logFC	AveExpr	t	P.Value	adj.P.Val	B
36711_at	6.62006643616333	5.68064570220007	92.5693933231114	2.65028119719727e-8	0.000295281079585734	10.2773590076056
209627_s_at	-1.12929917647223	2.7748012256519	-78.31736426409	5.46323320940229e-8	0.000405790752017037	9.706502815097
205207_at	2.21941755488266	3.73191933028938	45.7427185191415	5.58881448165734e-7	0.00301617250024004	7.47746405880953
38037_at	3.49901229491592	5.57574779460836	41.939722035969	8.13191556244247e-7	0.00301617250024004	7.07557001935978
203704_s_at	-2.87475517346297	4.65957253177468	-40.5418027919656	9.41430574503109e-7	0.00301617250024004	6.91630575967529
214855_s_at	-2.15963824567424	5.12243200203555	-40.4814997035835	9.47502917097352e-7	0.00301617250024004	6.9092847218066
219832_s_at	2.81808862731065	6.7711707465303	37.2486176754948	0.0000013573009171624	0.00378059204214123	6.5132288589758
220742_s_at	-1.68423599821727	7.08584517398116	-33.9550782163455	0.00000202404224712383	0.0050113037102956	6.06543239007262
203094_at	2.87439111401632	7.70938259710934	32.5197708766691	0.00000243865410333822	0.00536051577284091	5.85422401689705
219228_at	0.897106029114045	6.62800443385283	31.9097670273677	0.00000264621790159539	0.00536051577284091	5.76120507617444

Showing 1 to 10 of 2,030 entries | Previous 1 2 3 4 5 … 203 Next

Figura 20. Tabla conteniendo los resultados del análisis estadístico para cada gen. Clicando en la columna "adj.P.Val" podemos ordenar estos genes de más a menos significativo.

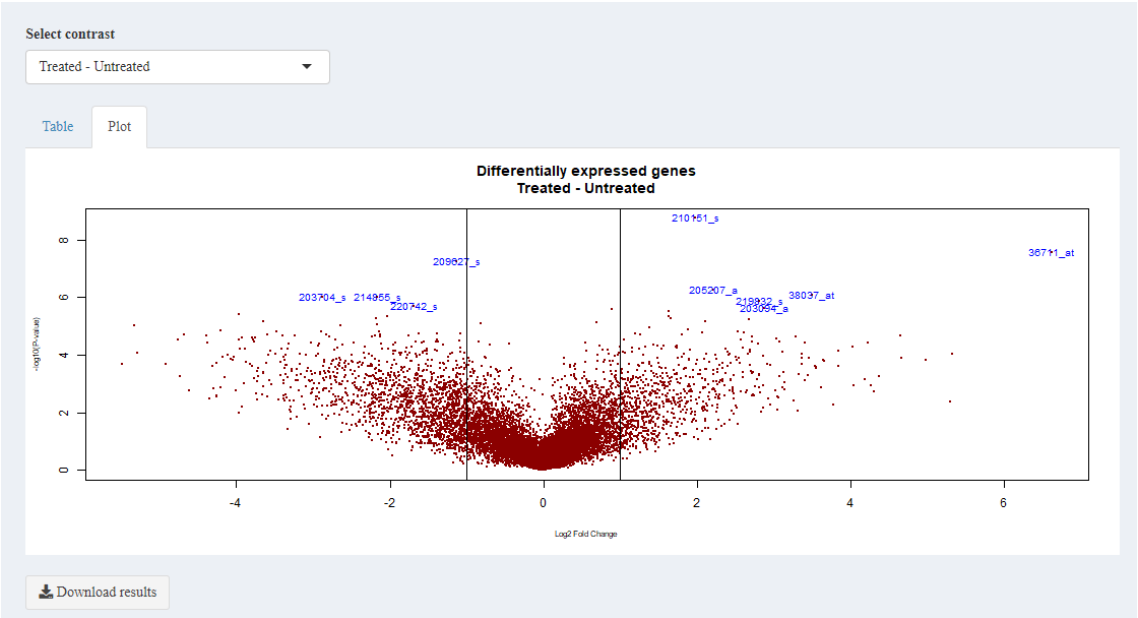


Figura 21. Gráfico volcano con los resultados del análisis. Los genes en azul representan los 10 genes estadísticamente más significativos.

Dependiendo de las necesidades del investigador, éste puede descargar directamente la tabla en formato *.csv* clicando en el botón “Download results”, o puede acceder al panel “Annotation and enrichment” para añadir el nombre común de los genes.

Anotación de los resultados y análisis de enriquecimiento

En el panel de anotación y enriquecimiento, introducimos el nombre de la base de datos del chip, en este caso *hgu133a.db*. El siguiente panel nos permite seleccionar el contraste, en este caso solo hemos realizado un contraste, “Treated – Untreated”. A continuación, seleccionamos “Yes” en “Do you want to perform enrichment analysis?” para configurar el análisis de enriquecimiento. Seleccionamos “Human” como especie y 0.05 como p-valor. Clicamos en “Calculate” y esperamos que se computen los resultados (Figura 22).

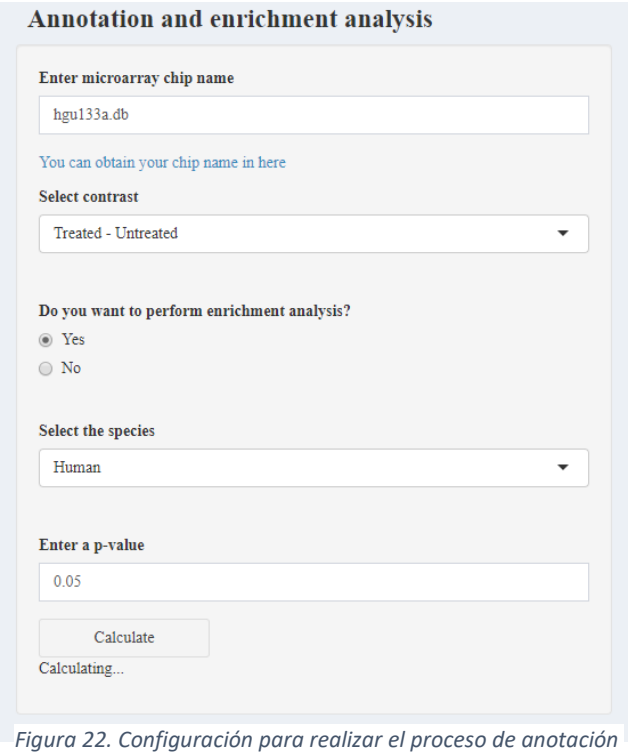
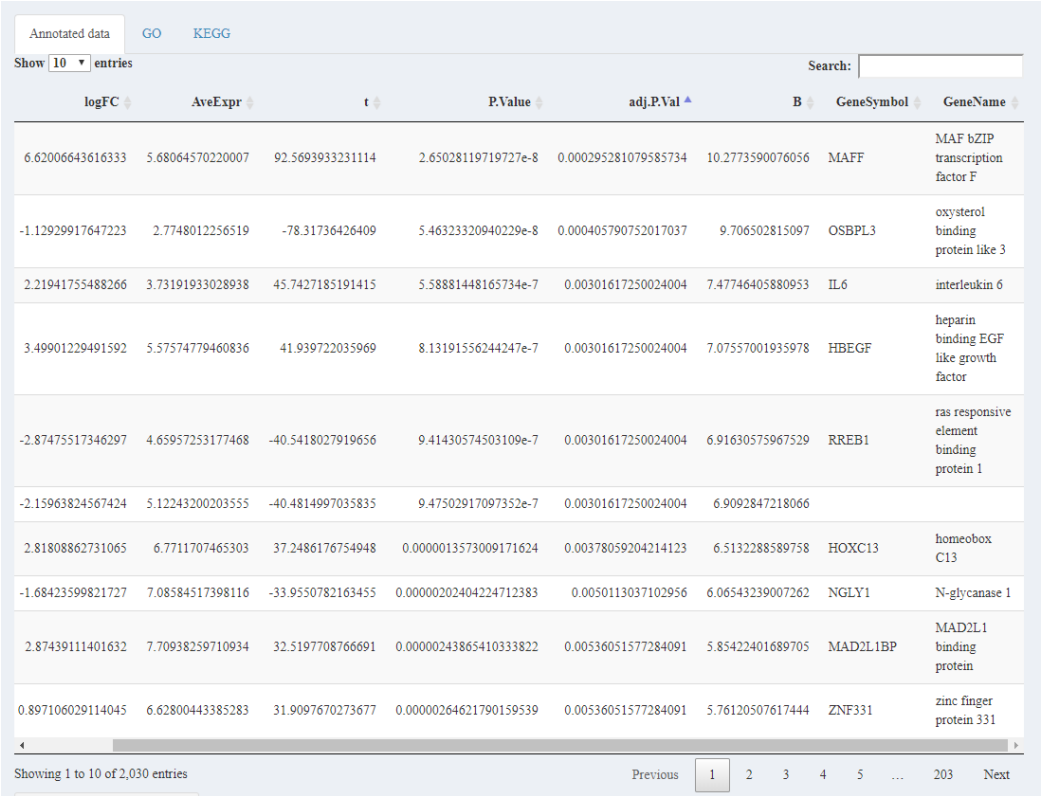


Figura 22. Configuración para realizar el proceso de anotación y análisis de enriquecimiento en el ejemplo 1.

La primera tabla que se presenta en los resultados es la misma tabla obtenida en el panel anterior, pero con dos columnas nuevas, una columna con el símbolo común del gen y otra columna con el nombre común (Figura 23).



Annotated data | GO | KEGG

Show 10 entries — Search:

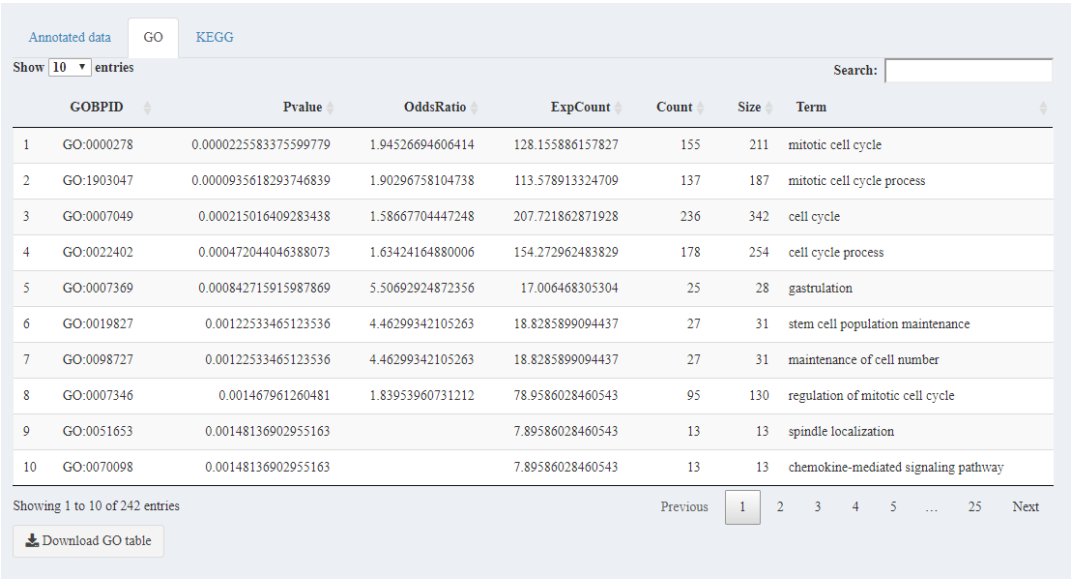
logFC	AveExpr	t	P.Value	adj.P.Val	B	GeneSymbol	GeneName
6.62006643616333	5.68064570220007	92.5693933231114	2.65028119719727e-8	0.000295281079585734	10.2773590076056	MAFF	MAF bZIP transcription factor F
-1.12929917647223	2.7748012256519	-78.31736426409	5.46323320940229e-8	0.000405790752017037	9.706502815097	OSBPL3	oxysterol binding protein like 3
2.21941755488266	3.73191933028938	45.7427185191415	5.58881448165734e-7	0.00301617250024004	7.47746405880953	IL6	interleukin 6
3.49901229491592	5.57574779460836	41.939722035969	8.13191556244247e-7	0.00301617250024004	7.07557001935978	HBEGF	heparin binding EGF like growth factor
-2.87475517346297	4.65957253177468	-40.5418027919656	9.41430574503109e-7	0.00301617250024004	6.91630575967529	RREB1	ras responsive element binding protein 1
-2.15963824567424	5.12243200203555	-40.4814997035835	9.47502917097352e-7	0.00301617250024004	6.9092847218066		
2.81808862731065	6.7711707465303	37.2486176754948	0.0000013573009171624	0.00378059204214123	6.5132288589758	HOXC13	homeobox C13
-1.68423599821727	7.08584517398116	-33.9550782163455	0.0000020240224712383	0.0050113037102956	6.06543239007262	NGLY1	N-glycanase 1
2.87439111401632	7.70938259710934	32.5197708766691	0.0000024386541033822	0.00536051577284091	5.85422401689705	MAD2L1BP	MAD2L1 binding protein
0.897106029114045	6.62800443385283	31.9097670273677	0.0000026462179015953 9	0.00536051577284091	5.76120507617444	ZNF331	zinc finger protein 331

Showing 1 to 10 of 2,030 entries — Previous 1 2 3 4 5 … 203 Next

Figura 23. Tabla con los resultados estadísticos para cada gen, con el símbolo y nombre común añadidos, en el ejemplo 1.

Como hemos configurado la opción de análisis de enriquecimiento, también obtenemos una tabla con los resultados para procesos biológicos (Figura 24) y otra tabla con los resultados para vías metabólicas (Figura 25).

Las tres tablas obtenidas están disponibles para descargar clicando en los respectivos botones debajo de éstas.



Annotated data | GO | KEGG

Show 10 entries — Search:

	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
1	GO:0000278	0.0000225583375599779	1.94526694606414	128.155886157827	155	211	mitotic cell cycle
2	GO:1903047	0.0000935618293746839	1.90296758104738	113.578913324709	137	187	mitotic cell cycle process
3	GO:0007049	0.000215016409283438	1.58667704447248	207.721862871928	236	342	cell cycle
4	GO:0022402	0.000472044046388073	1.63424164880006	154.272962483829	178	254	cell cycle process
5	GO:0007369	0.000842715915987869	5.50692924872356	17.006468305304	25	28	gastrulation
6	GO:0019827	0.00122533465123536	4.46299342105263	18.8285899094437	27	31	stem cell population maintenance
7	GO:0098727	0.00122533465123536	4.46299342105263	18.8285899094437	27	31	maintenance of cell number
8	GO:0007346	0.001467961260481	1.83953960731212	78.9586028460543	95	130	regulation of mitotic cell cycle
9	GO:0051653	0.00148136902955163		7.89586028460543	13	13	spindle localization
10	GO:0070098	0.00148136902955163		7.89586028460543	13	13	chemokine-mediated signaling pathway

Showing 1 to 10 of 242 entries — Previous 1 2 3 4 5 … 25 Next

Download GO table

Figura 24. Resultados del análisis de enriquecimiento para funciones biológicas.

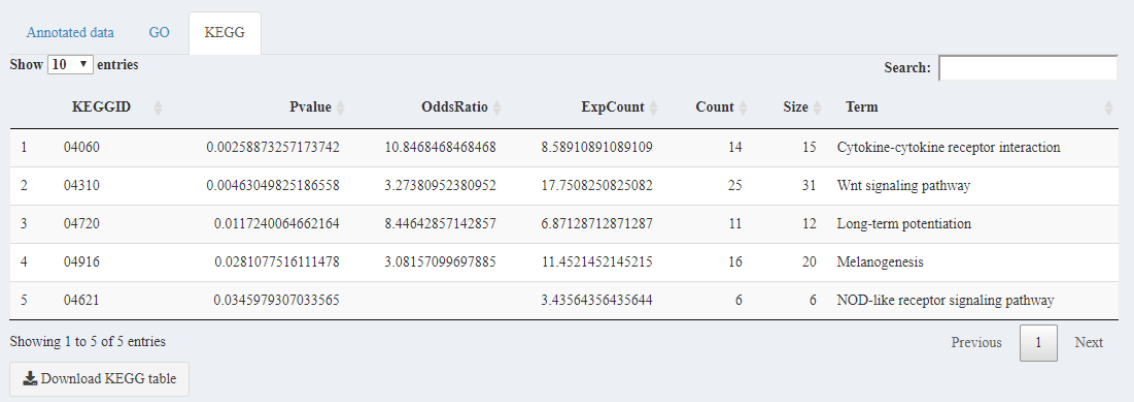


Figura 25. Resultados del análisis de enriquecimiento para vías metabólicas.

2.3.2. Ejemplo RNA-Seq: Estudio de variabilidad en los niveles de expresión génica a partir de células basales y luminales en hembras de ratón vírgenes, lactantes y embarazadas.

Los datos utilizados para este ejemplo fueron publicados en 2015, en un artículo de la revista *Nature Cell Biology*, por *Fu et al.* [34]. En este artículo se estudian los perfiles de expresión de células basales y de células luminales, extraídas de las glándulas mamarias de hembras de ratón, en condiciones de embarazo, lactancia o vírgenes. Por lo tanto, disponemos de 6 grupos diferentes, combinando el tipo de célula con el estado del ratón. Cada grupo experimental es replicado 1 vez, obteniendo 12 muestras diferentes representando esos 6 grupos.

Carga de datos

Comenzamos el análisis accediendo al panel “Configuration” y cargando el archivo de texto con los datos *count* y el archivo de texto con la información experimental. Pulsamos el botón “Load” y esperamos a que aparezca el mensaje “Data loaded!”. Una vez hemos cargado los datos, aparece la tabla con la información experimental en la interfaz (Figura 26).

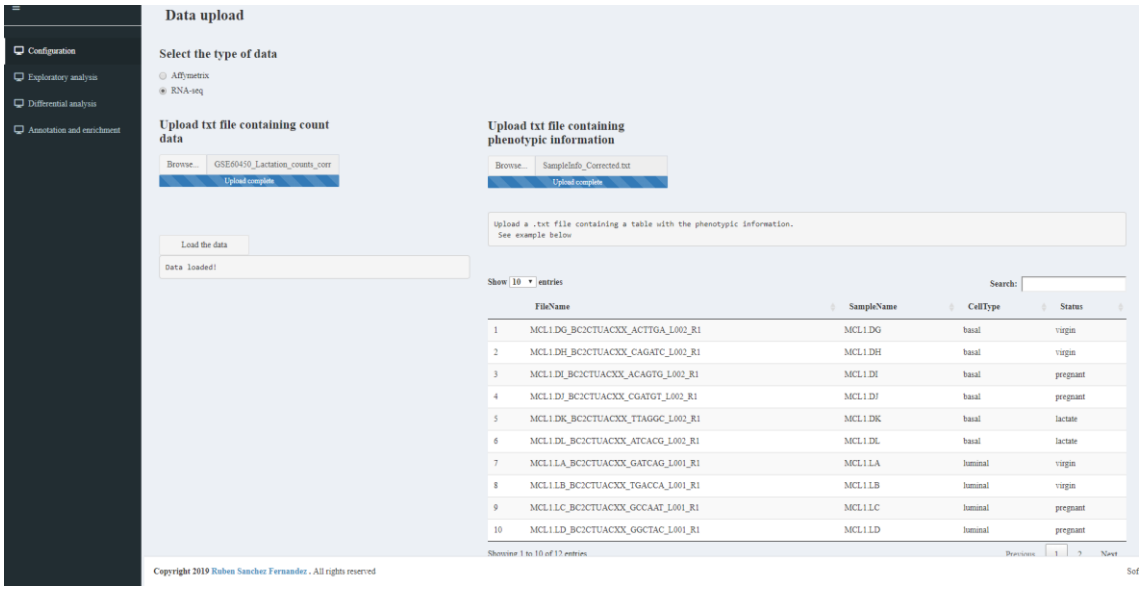


Figura 26. Proceso de carga de datos para el ejemplo 2.

Análisis exploratorio

Una vez hemos cargado los datos, podemos acceder al panel "Exploratory analysis" para visualizar los gráficos de control de calidad. Para experimentos de RNA-Seq, podemos seleccionar la columna con las condiciones que queremos comparar. En

este caso, realizaremos el análisis exploratorio seleccionando la columna con el estado del ratón (*Status*).

En figura 27, figura 28 y figura 29 se presentan, respectivamente, el *boxplot*, el gráfico MDS y el *heatmap* obtenidos para este ejemplo.

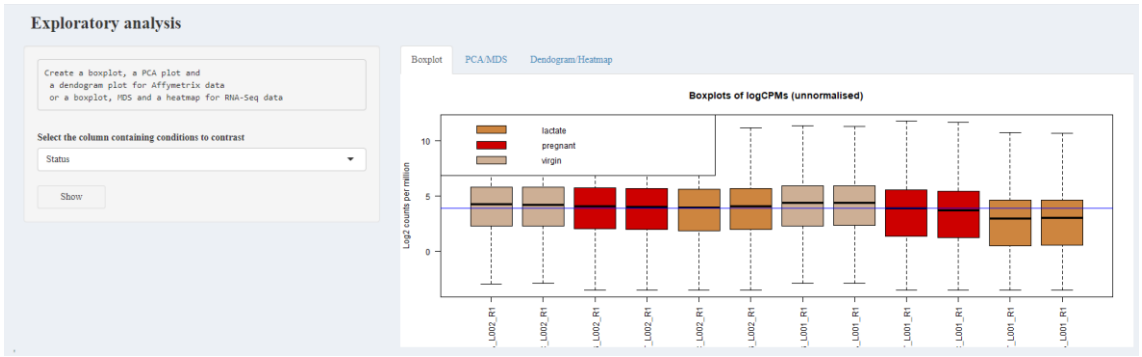


Figura 27. Boxplot obtenido en el panel "Exploratory analysis" para el ejemplo 2.

Figura 28. Gráfico MDS obtenido en el panel "Exploratory analysis" para el ejemplo 2.

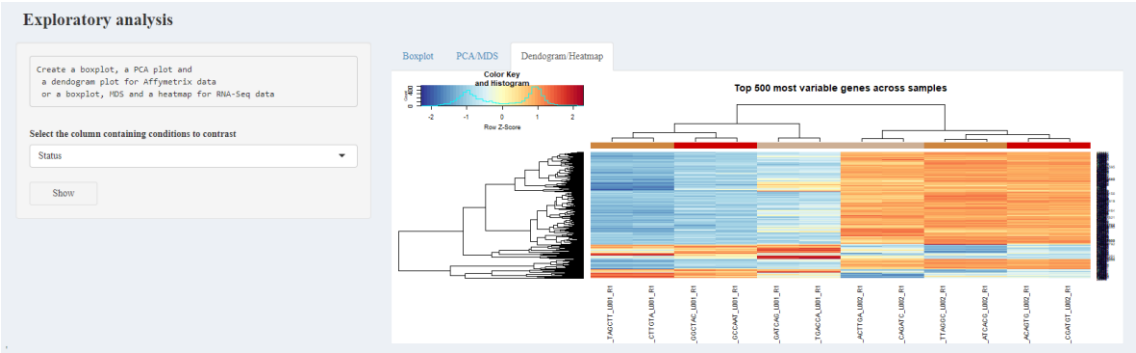


Figura 29. Heatmap obtenido en el panel "Exploratory analysis" para el ejemplo 2.

A partir de los gráficos observamos como las muestras se agrupan dos a dos (replicaciones), pero no parece que se agrupen en función de la condición seleccionada.

Análisis diferencial

Obtenemos la selección de genes estadísticamente significativos accediendo al panel “Differential analysis” y configurando el análisis. Seleccionamos el método de normalización deseado, en este caso seleccionamos “TMM”, “VOOM + EBayes “como técnica estadística, seleccionamos las columnas a contrastar, en este caso seleccionamos “CellType” y “Status”, y por último escogemos el valor de FDR deseado, en este caso 0.1 (Figura 30).

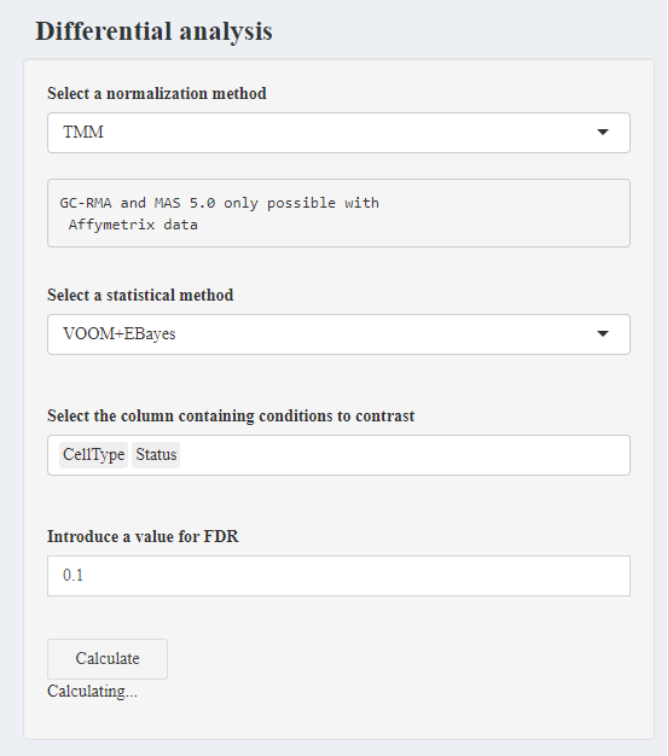
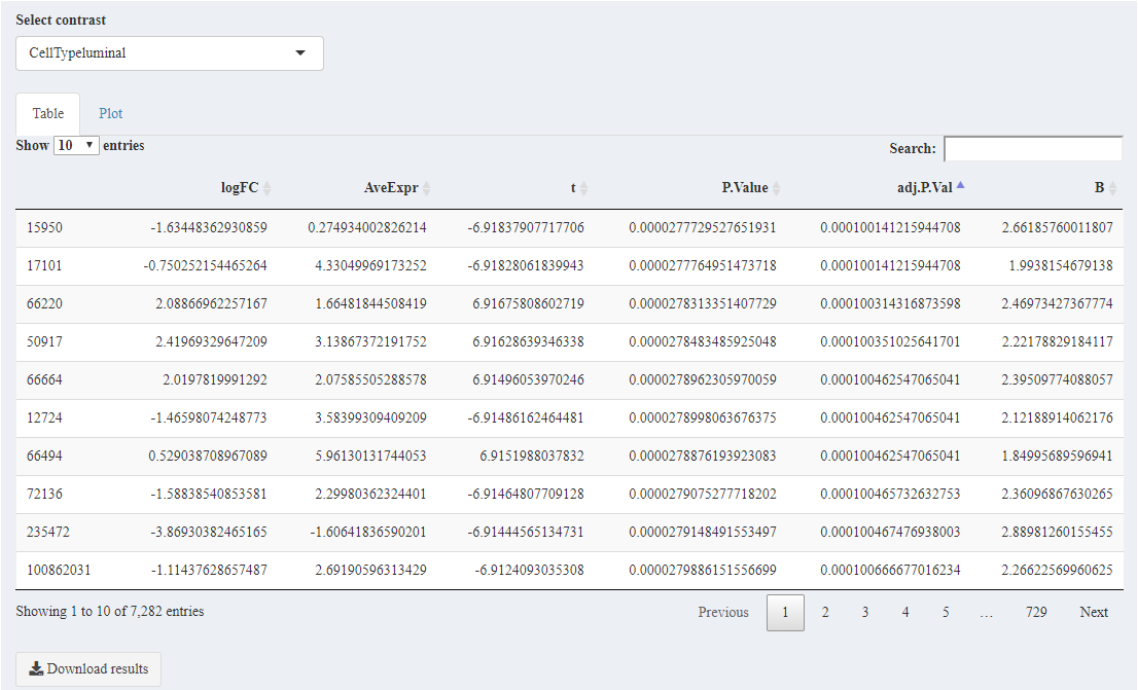


Figura 30. Configuración escogida en el panel "Differential analysis" para el ejemplo 2..

Los resultados se presentan en forma de tabla conteniendo los resultados del test estadístico para cada gen, y el gráfico MD destacando los genes estadísticamente significativos. Podemos seleccionar para que contraste queremos visualizar los resultados, mediante el selector del contraste. Además, se pueden descargar las tablas en formato .csv pulsando el botón correspondiente.

En la figura 30 se muestra la tabla obtenida para el contraste “CellType”, es decir, con el tipo de célula (luminal/basal) como contraste. En la figura 31, se muestra el gráfico MD para este resultado.



	logFC	AveExpr	t	P.Value	adj.P.Val	B
15950	-1.63448362930859	0.274934002826214	-6.91837907717706	0.0000277729527651931	0.000100141215944708	2.66185760011807
17101	-0.750252154465264	4.33049969173252	-6.91828061839943	0.0000277764951473718	0.000100141215944708	1.9938154679138
66220	2.08866962257167	1.66481844508419	6.91675808602719	0.0000278313351407729	0.000100314316873598	2.46973427367774
50917	2.41969329647209	3.13867372191752	6.91628639346338	0.0000278483485925048	0.000100351025641701	2.22178829184117
66664	2.0197819991292	2.07585505288578	6.91496053970246	0.0000278962305970059	0.000100462547065041	2.39509774088057
12724	-1.46598074248773	3.58399309409209	-6.91486162464481	0.0000278998063676375	0.000100462547065041	2.12188914062176
66494	0.529038708967089	5.96130131744053	6.9151988037832	0.0000278876193923083	0.000100462547065041	1.84995689596941
72136	-1.58838540853581	2.29980362324401	-6.91464807709128	0.0000279075277718202	0.000100465732632753	2.36096867630265
235472	-3.86930382465165	-1.60641836590201	-6.91444565134731	0.0000279148491553497	0.000100467476938003	2.88981260155455
100862031	-1.11437628657487	2.69190596313429	-6.9124093035308	0.0000279886151556699	0.000100666677016234	2.26622569960625

Figura 31. Tabla conteniendo los resultados del test estadístico para cada gen, con el tipo de célula como condición.

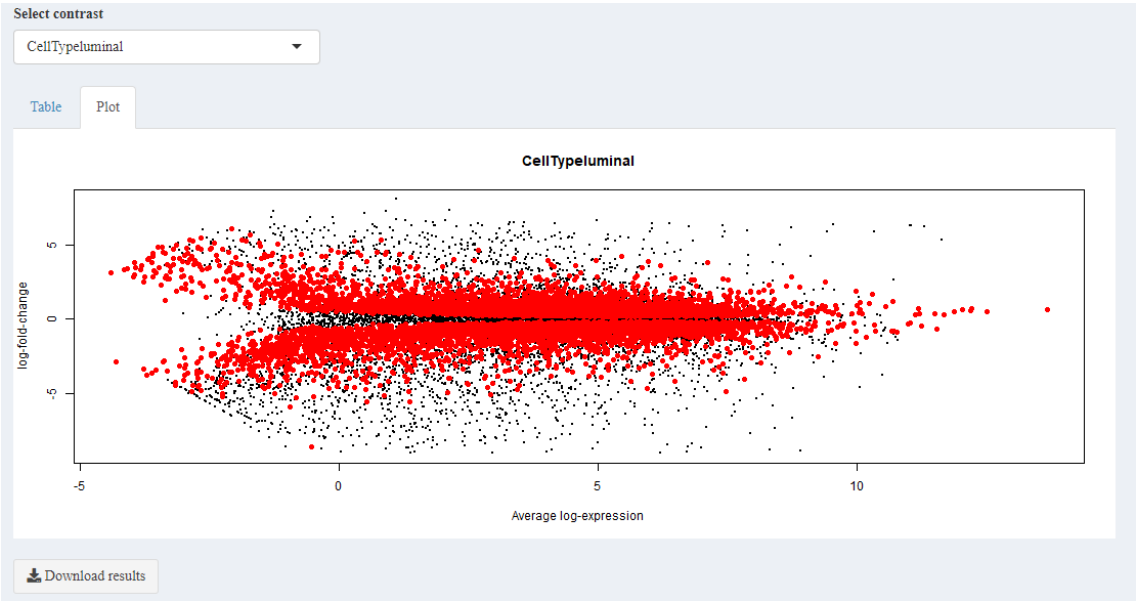


Figura 32. Gráfico MD mostrando los genes estadísticamente significativos para la condición Tipo de célula.

Anotación de los resultados y análisis de enriquecimiento

Finalizamos el análisis accediendo al panel “Annotation and enrichment analysis” para computar los nombres comunes de los genes estadísticamente significativos y para realizar el análisis de enriquecimiento.

Primero, seleccionamos el contraste que nos interesa, en este caso seleccionaremos el contraste “CellTypeluminal” (lumina-basal), marcamos “Yes” en el selector “Do you want to perform enrichment analysis?”, seleccionamos la especie “Mouse” y un p-valor de 0,01 (Figura 33). Clicamos en “Calculate” y esperamos a que se generen los resultados.

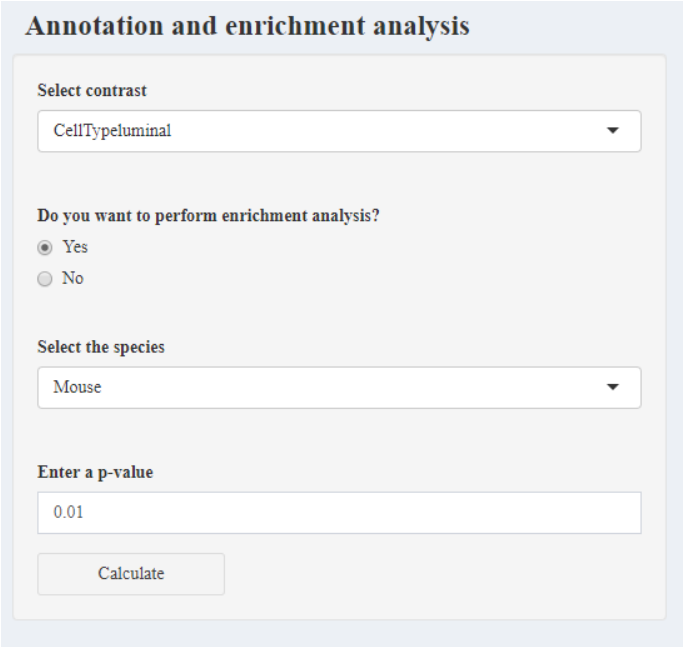
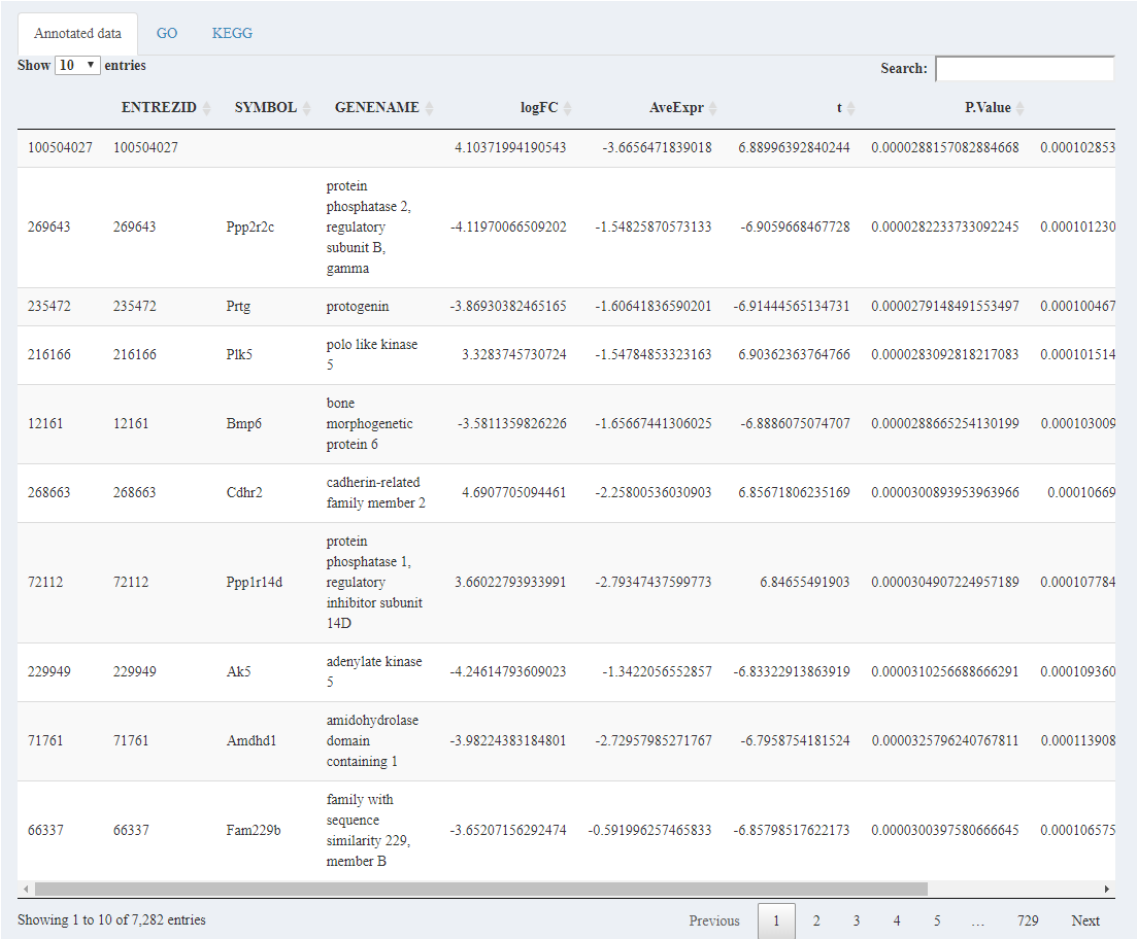


Figura 33. Configuración del panel "Annotation and enrichment analysis" para el ejemplo 2..

Como en el ejemplo anterior, los resultados se presentan en forma de tablas. La primera tabla que obtenemos (Figura 34), contiene los genes estadísticamente significativos con sus respectivos símbolos y nombres comunes. Seleccionando la pestaña “GO”, accedemos a la tabla con las funciones biológicas más representadas (Figura 35) y accediendo a la pestaña “KEGG” accedemos a la tabla con las vías metabólicas con mayor representación (Figura 36). Para finalizar, el usuario puede descargar todos los resultados en formato *.csv* clicando en los botones que se encuentran debajo de cada tabla.



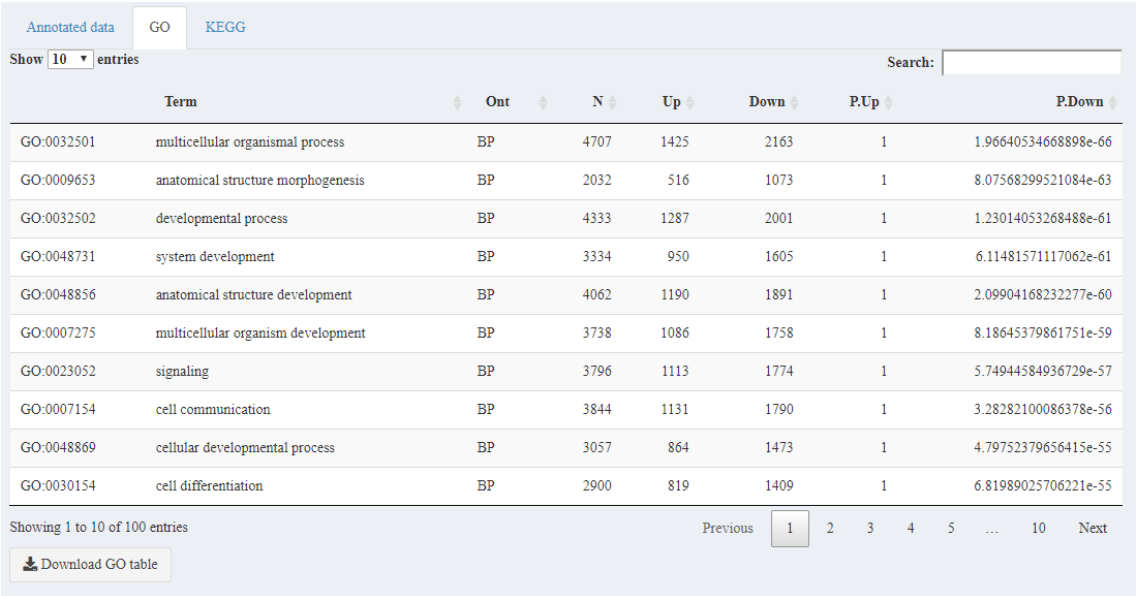
Annotated data | GO | KEGG

Show 10 entries | Search:

	ENTREZID	SYMBOL	GENENAME	logFC	AveExpr	t	P.Value	
100504027	100504027			4.10371994190543	-3.6656471839018	6.88996392840244	0.0000288157082884668	0.000102853
269643	269643	Ppp2r2c	protein phosphatase 2, regulatory subunit B, gamma	-4.11970066509202	-1.54825870573133	-6.9059668467728	0.0000282233733092245	0.000101230
235472	235472	Prtg	protogenin	-3.86930382465165	-1.60641836590201	-6.91444565134731	0.0000279148491553497	0.000100467
216166	216166	Plk5	polo like kinase 5	3.3283745730724	-1.54784853323163	6.90362363764766	0.0000283092818217083	0.000101514
12161	12161	Bmp6	bone morphogenetic protein 6	-3.5811359826226	-1.65667441306025	-6.8886075074707	0.0000288665254130199	0.000103009
268663	268663	Cdhr2	cadherin-related family member 2	4.6907705094461	-2.25800536030903	6.85671806235169	0.0000300893953963966	0.00010669
72112	72112	Ppp1r14d	protein phosphatase 1, regulatory inhibitor subunit 14D	3.66022793933991	-2.79347437599773	6.84655491903	0.0000304907224957189	0.000107784
229949	229949	Ak5	adenylate kinase 5	-4.24614793609023	-1.3422056552857	-6.83322913863919	0.0000310256688666291	0.000109360
71761	71761	Amdhd1	amidohydrolase domain containing 1	-3.98224383184801	-2.72957985271767	-6.7958754181524	0.0000325796240767811	0.000113908
66337	66337	Fam229b	family with sequence similarity 229, member B	-3.65207156292474	-0.591996257465833	-6.85798517622173	0.0000300397580666645	0.000106575

Showing 1 to 10 of 7,282 entries | Previous 1 2 3 4 5 … 729 Next

Figura 34. Resultados del proceso de anotación para el contraste "Tipo de célula"



Annotated data | GO | KEGG

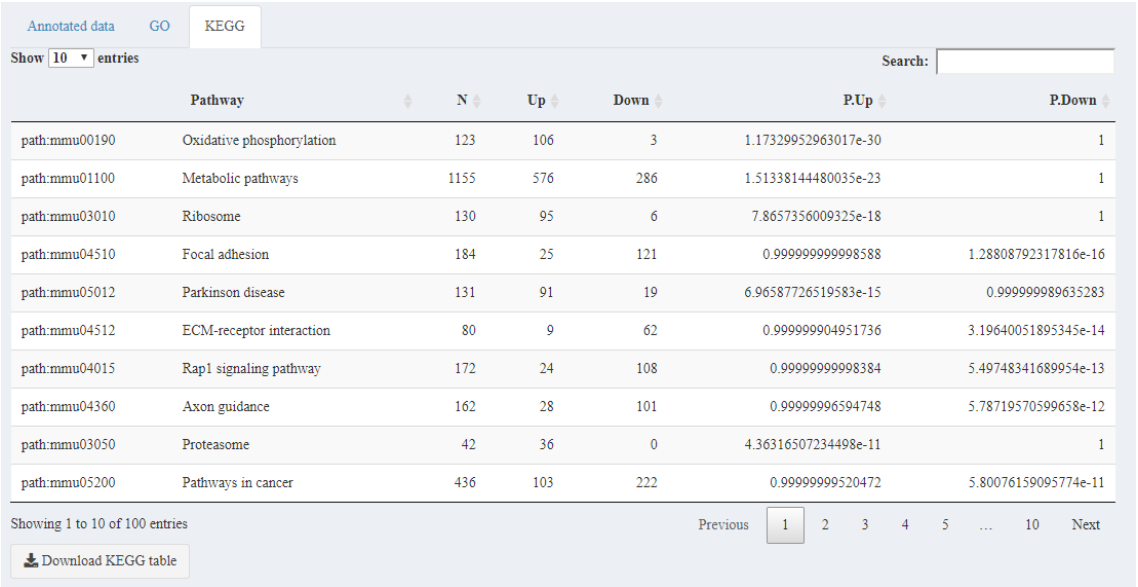
Show 10 entries | Search:

	Term	Ont	N	Up	Down	P.Up	P.Down
GO:0032501	multicellular organismal process	BP	4707	1425	2163	1	1.96640534668898e-66
GO:0009653	anatomical structure morphogenesis	BP	2032	516	1073	1	8.07568299521084e-63
GO:0032502	developmental process	BP	4333	1287	2001	1	1.23014053268488e-61
GO:0048731	system development	BP	3334	950	1605	1	6.11481571117062e-61
GO:0048856	anatomical structure development	BP	4062	1190	1891	1	2.09904168232277e-60
GO:0007275	multicellular organism development	BP	3738	1086	1758	1	8.18645379861751e-59
GO:0023052	signaling	BP	3796	1113	1774	1	5.74944584936729e-57
GO:0007154	cell communication	BP	3844	1131	1790	1	3.28282100086378e-56
GO:0048869	cellular developmental process	BP	3057	864	1473	1	4.79752379656415e-55
GO:0030154	cell differentiation	BP	2900	819	1409	1	6.81989025706221e-55

Showing 1 to 10 of 100 entries | Previous 1 2 3 4 5 … 10 Next

Download GO table

Figura 35. Resultados del proceso de análisis de enriquecimiento para funciones biológicas.



Annotated data | GO | KEGG

Show 10 entries — Search:

	Pathway	N	Up	Down	P.Up	P.Down
path:mmu00190	Oxidative phosphorylation	123	106	3	1.17329952963017e-30	1
path:mmu01100	Metabolic pathways	1155	576	286	1.51338144480035e-23	1
path:mmu03010	Ribosome	130	95	6	7.8657356009325e-18	1
path:mmu04510	Focal adhesion	184	25	121	0.999999999998588	1.28808792317816e-16
path:mmu05012	Parkinson disease	131	91	19	6.96587726519583e-15	0.999999989635283
path:mmu04512	ECM-receptor interaction	80	9	62	0.999999904951736	3.19640051895345e-14
path:mmu04015	Rap1 signaling pathway	172	24	108	0.99999999998384	5.49748341689954e-13
path:mmu04360	Axon guidance	162	28	101	0.99999996594748	5.78719570599658e-12
path:mmu03050	Proteasome	42	36	0	4.36316507234498e-11	1
path:mmu05200	Pathways in cancer	436	103	222	0.99999999520472	5.80076159095774e-11

Showing 1 to 10 of 100 entries — Previous 1 2 3 4 5 … 10 Next

Download KEGG table

Figura 36. Resultados del proceso de análisis de enriquecimiento para vías metabólicas.

Capítulo 3

Conclusiones

Mediante el proyecto presentado, se ha explorado una metodología de diseño y aplicación de una herramienta para realizar análisis diferencial a datos ómicos. Debido a la naturaleza del propio análisis, plantearse un proyecto de estas características en un espacio relativamente corto de tiempo ha supuesto un reto considerable. Aun así, se ha conseguido un producto funcional, con una interfaz simple e intuitiva, que cumple el objetivo más importante del proyecto: simplificar el proceso de análisis y abrirlo a cualquier investigador sin requerirle conocimientos técnicos en programación.

Como hemos visto, existen una serie de pasos establecidos que los analistas deben seguir para implementar este tipo de análisis. Aunque sigamos este *guión*, el criterio del analista puede hacer variar el proceso de análisis notablemente. Por ejemplo, en este caso, en el análisis de microarrays, se ha decidido no implementar un proceso de filtraje para evitar posibles sesgos en los resultados, pero si se hubiera implementado seguramente procesos posteriores como el análisis estadístico se hubieran visto condicionados. También encontramos variabilidad de método en el propio proceso de análisis estadístico e incluso en el proceso de normalización. Todo esto debe ser considerado a la hora de comparar resultados entre diferentes estudios, y sobre todo se debe tener en cuenta que el análisis de expresión diferencial es un proceso todavía muy abierto, en el que continuamente aparecen nuevos estudios proponiendo nuevas herramientas y métodos.

En el caso de disponer de un tiempo ilimitado para realizar el proyecto, toda esta variabilidad no supondría problema alguno. De hecho, permitiría ampliar la herramienta y proporcionar mayor libertad al investigador para configurar el análisis que desee. Como este no era el caso, se han tenido que tomar decisiones a la hora de escoger métodos de normalización, y especialmente a la hora de escoger métodos de análisis estadístico. Escoger el análisis mediante modelos lineales como piedra angular de la herramienta ha permitido abarcar un gran número de posibilidades permitiendo al usuario configurar el diseño experimental y los contrastes, además de ampliar la herramienta a varios tipos de experimentos (microarray y RNA-Seq).

Respecto a la aplicación presentada, algunos detalles a mejorar merecen ser comentados. Lo primero, debido al gran número de procesos que la aplicación puede realizar, numerosos paquetes de R son utilizados. Algunos de estos paquetes han sido desarrollados con versiones de R relativamente nuevas, lo que dificulta que la herramienta pueda ser utilizada en sistemas operativos sin acceso a las últimas versiones. Un ejemplo es Linux. En la realización del proyecto, se intentó instalar la aplicación en una máquina virtual Ubuntu de *Amazon Web Services* (AWS) para permitir el acceso remoto a la aplicación vía web. El proceso no tuvo éxito, en parte, debido a que la última versión de R en Ubuntu es anterior a la versión de R requerida por algunos paquetes. Además, al instalar múltiples paquetes, la aplicación requiere un mínimo de memoria que sobrepasa la memoria proporcionada por los servidores gratuitos de AWS. Por lo tanto, este requerimiento de memoria es otro de los problemas actuales de la aplicación. Una posible solución a estos problemas sería disminuir el uso de paquetes, desarrollando funciones propias para sustituir a estos. Debido a la falta de tiempo, esta solución no ha sido viable en este momento, y ha sido necesario aprovechar los distintos paquetes que proporciona R y Bioconductor para el análisis de datos ómicos.

Por lo tanto, como mejoras futuras, se propone la sustitución de algunos paquetes, principalmente los que limitan el uso de la herramienta, con funciones propias. Respecto a la interfaz gráfica, aun siendo funcional, tiene espacio para mejora en la parte visual y de diseño. Una vez se han solucionado estos inconvenientes, se puede trabajar en ampliar la herramienta, añadiendo nuevas metodologías de análisis, y ampliando también a otros tipos de datos.

Si nos centramos en la planificación del proyecto, algunos errores se han cometido que han dificultado el éxito de este. Quizás, uno de los errores más importantes ha sido no haber tenido un planteamiento concreto de lo que iba a ser el proyecto hasta unas semanas después del inicio del mismo. Debido a esto, las primeras semanas se centraron en pensar que se iba a hacer, en lugar de plantear y desarrollar la solución. Aun así, la planificación planteada en el inicio del proyecto se ha podido seguir sin problemas, incluso permitiendo incorporar la metodología de análisis para datos de RNA-Seq, no prevista en el diseño inicial. La buena planificación del calendario, en parte configurado a partir de las entregas de evaluación continua, ha permitido dedicarle el tiempo necesario a cada fase, y ha permitido finalizar el proyecto a tiempo, incluso después de un inicio algo tardío.

Capítulo 4

Glosario

ADN: ácido desoxirribonucleico

ARN: ácido ribonucleico

Counts: datos de conteo de RNA-Seq

DEG: Differential expressed gene/s (gen o genes diferencialmente expresados)

FDR: False Discovery Rate (término que hace referencia a la proporción de errores de tipo I)

GC-RMA: GeneChip-Robust Multiarray Averaging (técnica de normalización de microarrays)

GO: Gene Ontology (ontología génica)

KEGG: Kyoto Encyclopedia of Genes and Genomes (base de datos utilizada para localizar vías metabólicas)

MAS 5.0: MicroArray Suite 5.0 (técnica de normalización de microarrays)

Microarrays: chips para la secuenciación de ADN o ARN

RLE: Relative Log Expression (técnica de normalización de RNA-Seq)

RMA: Robust Multiarray Averaging (técnica de normalización de microarrays)

RNA-Seq: RNA-Sequencing (secuenciación de ARN)

SAM: Significance analysis of microarrays (análisis de significación de microarrays)

TMM: Trimmed Mean of M-values (técnica de normalización de RNA-Seq)

Capítulo 5

Bibliografía

[1] Tebani, A., Afonso, C., Marret, S., & Bekri, S. (2016). Omics-based strategies in precision medicine: toward a paradigm shift in inborn errors of metabolism investigations. *International journal of molecular sciences*, *17*(9), 1555.

[2] Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., ... & Djoumbou, Y. (2010). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research*, *39*(suppl_1), D1035-D1041.

[3] Abu-Asab, M., Chaouchi, M., & Amri, H. (2008). Evolutionary medicine: a meaningful connection between omics, disease, and treatment. *PROTEOMICS–Clinical Applications*, *2*(2), 122-134.

[4] Perez-Riverol, Y., Bai, M., da Veiga Leprevost, F., Squizzato, S., Park, Y. M., Haug, K., ... & del-Toro, N. (2017). Discovering and linking public omics data sets using the Omics Discovery Index. *Nature biotechnology*, *35*(5), 406.

[5] Gross, M. (2011). Riding the wave of biological data.

[6] Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., & Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis. *Nature reviews genetics*, *11*(9), 647.

[7] Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., ... & Xia, J. (2018). MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic acids research*, *46*(W1), W486-W494.

[8] Yu, S., Jiang, X., Li, J., Li, C., Guo, M., Ye, F., ... & Guo, B. (2019). Comprehensive analysis of the GATA transcription factor gene family in breast carcinoma using gene microarrays, online databases and integrated bioinformatics. *Scientific reports*, *9*(1), 4467.

[9] de Luis, D. A., Almansa, R., Aller, R., Izaola, O., & Romero, E. (2018). Gene expression analysis identify a metabolic and cell function alterations as a hallmark of obesity without metabolic syndrome in peripheral blood, a pilot study. *Clinical Nutrition*, *37*(4), 1348-1353.

[10] Team, R. C. (2013). R: A language and environment for statistical computing.

[11] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., ... & Hornik, K. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, *5*(10), R80.

[12] Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2017). shiny: Web Application Framework for R. R package version 1.0. 0. *R Found. Stat. Comput., Vienna. https://CRAN. R-project. org/package= shiny (accessed 19 May 2019)*.

[13] Assefa, A. T., De Paepe, K., Everaert, C., Mestdagh, P., Thas, O., & Vandesompele, J. (2018). Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing data. *Genome biology*, *19*(1), 96.

[14] Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor* (pp. 397-420). Springer, New York, NY.

[15] Wikipedia. Gene Expression, 2019.

[16] Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*(5235), 467-470.

[17] Hall, N. (2007). Advanced sequencing technologies and their wider impact in microbiology. *Journal of experimental biology*, *210*(9), 1518-1525.

[18] Ruíz de Villa, M.Carmen, Sánchez-Pla, Alex. Apuntes académicos asignatura *Análisis de datos ómicos (2019).* Universitat Oberta de Catalunya .

[19] Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57.

[20] Scholtens D., Heydebreck AV. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* -Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S, eds. (2005) New York: Springer. 229-248.

[21] Palejev, D. (2017). Comparison of RNA-seq differential expression methods. *Cybernetics and Information Technologies*, *17*(5), 60-67.

[22] Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PloS one*, *12*(12), e0190152.

[23] Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, *98*(9), 5116-5121.

[24] Smyth, G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*3 Article 3.

[25] Smyth, G. K., Michaud, J., & Scott, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, *21*(9), 2067-2075.

[26] Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor* (pp. 397-420). Springer, New York, NY.

[27] Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, *15*(2), R29.

[28] J. J. Faraway (2004). Linear Models with R (1.a ed.). Chapman and Hall/CRC.

[29] Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American statistical Association*, *78*(381), 47-55.

[30] Larsson, O., Wahlestedt, C., & Timmons, J. A. (2005). Considerations when using the significance analysis of microarrays (SAM) algorithm. *BMC bioinformatics*, *6*(1), 129.

[31] Maza, E. (2016). In papyro comparison of TMM (edgeR), RLE (DESeq2), and MRN normalization methods for a simple two-conditions-without-replicates RNA-seq experimental design. *Frontiers in genetics*, *7*, 164.

[32] Carson, J. P., Zhang, N., Frampton, G. M., Gerry, N. P., Lenburg, M. E., & Christman, M. F. (2004). Pharmacogenomic identification of targets for adjuvant therapy with the topoisomerase poison camptothecin. *Cancer research*, *64*(6), 2096-2104.

[33] Dancey, J., & Eisenhauer, E. A. (1996). Current perspectives on camptothecins in cancer treatment.

[34] Fu, N. Y., Rios, A. C., Pal, B., Soetanto, R., Lun, A. T., Liu, K., ... & Strasser, A. (2015). EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival. *Nature cell biology*, *17*(4), 365.