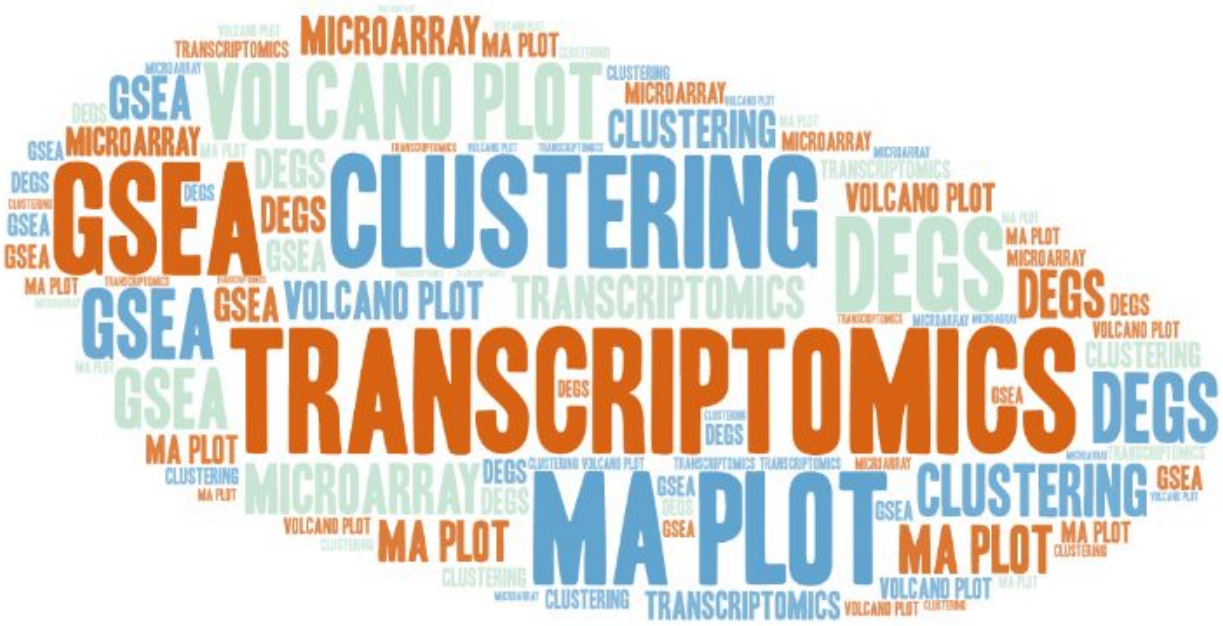
Proyecto Transcriptómica

Transcriptómica, Regulación genómica y Epigenómica - Universidad Autónoma de Madrid- 2019/2020



Álvaro Huertas García

Social media: linkedIn, github

Agradecimientos:

Agradezco a mis compañeros Sara Dorado Alfaro, Alejandro Martín Muñoz y Diego Mañanes Cayero todas las discusiones fructíferas que me han permitido profundizar más en este campo y aprender nuevos conocimientos.

INTRODUCCIÓN

En este proyecto se pretende llevar a cabo el análisis de los datos de microarray pertenecientes a la serie con identificador de la base de datos GEO: GSE18198. Los datos pertenecen al artículo titulado "Direct Inhibition of the NOTCH Transcription Factor Complex" (Moellering et al., 2009) [1].

Para contextualizar el trabajo y los pasos que se realizan, cabe destacar que en este artículo estudian el efecto del péptido sintético SAMH1, obtenidos a partir de la proteína MAML1, sobre las proteínas NOTCH. Las proteínas NOTCH están relacionadas con rutas de diferenciación, proliferación y muerte celular.

Para estudiar el efecto del péptido sintético SAHM1 sobre la actividad transcripcional los autores realizan un análisis del perfil transcripcional y un enriquecimiento de set de genes ("Gene set enrichment analysis", GSEA) en dos líneas celulares humanas de linfocitos T (HPB-ALL y KOPT-K1). Se emplean triplicados de cada línea celular en dos condiciones diferentes, una situación control en la que las muestras son tratadas con el compuesto dimetilsulfóxido (DMSO) y la situación de interés en la que se aplica el péptido sintético SAMH1. De este modo, disponemos de 12 muestras analizadas por la tecnología de microarrays de Affymetrix U133 Plus 2.0.

Es a partir de aquí donde comienza el proyecto de transcriptómica. Se pretende reproducir y analizar los datos procedentes de los microarrays con el objetivo de validar y analizar los resultados del efecto de SAMH1.

MATERIAL

El material de partida son los datos de los 12 microarrays pertenecientes a la serie GSE18198:

- GSM455115 KOPT-K1_DMSO_01
- GSM455116 KOPT-K1_DMSO_02
- GSM455117 KOPT-K1_DMSO_03
- GSM455118 HPB-ALL_DMSO_01
- GSM455119 HPB-ALL_DMSO_02
- GSM455120 HPB-ALL_DMSO_03
- GSM455121 KOPT-K1_SAHM1_01
- GSM455122 KOPT-K1_SAHM1_02

- GSM455123 KOPT-K1_SAHM1_03
- GSM455124 HPB-ALL_SAHM1_01
- GSM455125 HPB-ALL_SAHM1_02
- GSM455126 HPB-ALL_SAHM1_03

Se disponen de dos líneas celulares humanas de linfocitos T (HPB-ALL y KOPT-K1) y dos condiciones (DMSO y SAMH1) con tres réplicas cada una. El tratamiento con DMSO es considerado como control.

Para realizar el análisis de expresión diferencial se emplea R version 3.6.2 y Bioconductor. Para realizar el enriquecimiento de set de genes (GSEA) se emplea la versión de escritorio de GSEA [2]. Para llevar a cabo el análisis de clusterización jerarquizado y no jerarquizado se emplea Morpheus [3]. Para el análisis de predicción de clases mediante métodos supervisado se emplea la herramienta "Expression/Class Prediction" [4] de Babelomics v.5 [5]. El análisis funcional se realiza empleando PANTHER v.14 [6, 7] y las herramientas SNOW [8] y FatiGO [9] de Babelomics v.5.0 [5].

ANÁLISIS DE EXPRESIÓN DIFERENCIAL

El análisis de expresión diferencial consiste en emplear métodos de análisis estadístico para estudiar diferencias cuantitativas en los niveles de expresión entre diferentes condiciones de estudio. En nuestro caso, nos interesa conocer los genes que presentan niveles de expresión significativamente diferentes entre la condición control (DMSO) y la condición experimental (tratada con SAMH1).

La primera cuestión que tenemos que abordar es cómo diseñamos el análisis comparativo. Como se ha indicado anteriormente, disponemos de muestras de ambas condiciones procedentes de dos líneas celulares diferentes (HPB-ALL y KOPT-K1). A pesar de que ambas líneas celulares pertenecen linfocitos T humanos relacionados con leucemia, cabe preguntarse si hemos de emplear ambas líneas celulares conjuntamente o, por el contrario, realizar el estudio de expresión diferencial por separado.

Esta cuestión es clave, dado que si consideramos las muestras de HPB-ALL y KOPT-K1 conjuntamente, todas las diferencias que existan entre estas líneas celulares podrán afectar al análisis de expresión diferencial. Por ejemplo, el artículo de estudio [1] identifica que algunos genes diana de NOTCH1 (*MYC*, *DTX1*, *HES1* y *HES4*) se encuentran sub-regulados cuando son tratados con SAMH1. Para poder asociar este cambio en el nivel de expresión al efecto de SAMH1, es necesario que estos genes no se encuentren diferencialmente expresados en la situación control (DMSO) en ambas líneas celulares.

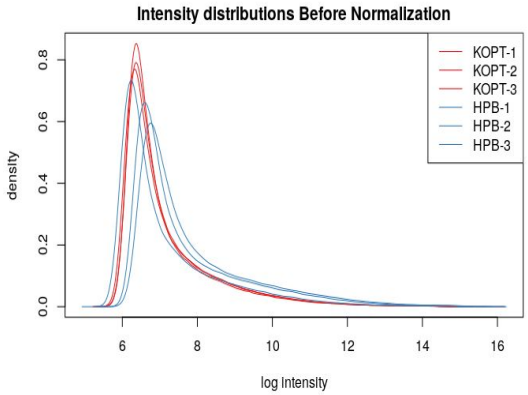
En caso de que las líneas celulares muestren diferencias significativas en los niveles de expresión en la situación control sería necesario estudiar el efecto de SAMH1 en cada línea por separado. En caso contrario enmascararíamos los resultados y atribuiríamos efectos a SAMH1 sobre los niveles de transcripción que realmente no se deben al péptido sintético.

De este modo, el primer análisis de expresión diferencial se realizará entre las muestras control de KOPT-K1 (GSM455115, GSM455116, GSM455117) y HPB-ALL (GSM455118, GSM455119, GSM455120).

Comparación líneas celulares: KOPT-K1 DMSO vs HPB-ALL DMSO

En primer lugar se realiza un análisis exploratorio de los datos control de ambas líneas celulares que se van a comparar. En la Figura 1 se muestran las distribuciones de intensidad y los boxplots de los datos originales de los diferentes microarrays . Podemos observar cómo existen diferencias entre las diferentes muestras, siendo las réplicas de HPB-ALL las que más diferencias muestran entre sí. Estas diferencias señalan la importancia de conocer las fuentes de variación.

A



B

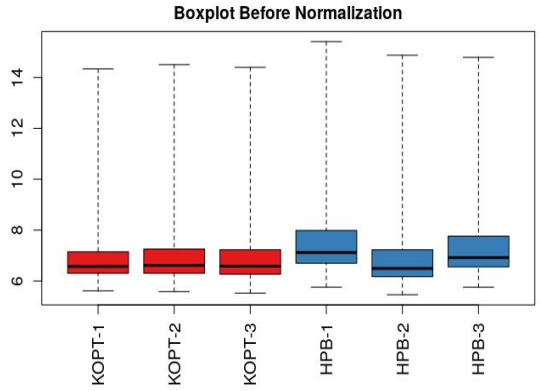


Figura 1 - Distribución de las intensidades (A) y boxplots (B) de las muestras control (DMSO) de ambas líneas celulares antes de normalizar.

Las fuentes de variación en los experimentos de microarrays pueden ser de dos tipos: variación de la técnica y variación biológica. La variación de la técnica procede de la inexactitud de la propia tecnología [10]. Por otro lado, la variación biológica procede de la inexactitud presente en los propios organismos. Las variaciones técnicas, procedente de la limitación experimental y tecnológica, cada vez es menor y en el caso de los

microarrays es conocida.

La variabilidad biológica es más compleja puesto que puede proceder de varias fuentes como la variabilidad entre individuos, la variabilidad en las condiciones experimentales, la variabilidad de factores no controlados y la propia heterogeneidad de la muestra [10]. De este modo, controlar la variabilidad biológica no es tarea fácil y tenemos que ser conscientes de esta limitación cuando obtengamos resultados. No obstante, incluir medidas como protocolos reproducibles y robustos o seleccionar el tejido o células de interés para controlar la heterogeneidad son buenas prácticas que permiten reducir la variabilidad biológica [1].

Como se ha indicado anteriormente, la variabilidad técnica procedente de la tecnología de microarrays es conocida y puede ser controlada a nivel bioinformático. Los equipos empleados para la medición del nivel de hibridación de los microarrays proporcionan una imagen de intensidad de luz. La captación de esta intensidad depende del sistema pero todos comparten una serie de pasos. Primero comprueban la hibridación del cDNA con la sonda del microarray, luego identifican los puntos correspondientes a las diferentes sondas ("spots"), cuantifican la intensidad del spot y la intensidad del fondo del array y finalmente calculan el valor de intensidad media para cada spot [10]. Como consecuencia de este proceso se generan los archivos .CEL, empleados en este trabajo como punto de partida del análisis.

Las variaciones experimentales que afectan a estas lecturas son, por ejemplo, el fondo en las hibridación, la presencia de manchas y las diferentes intensidades entre microarrays. Estas variaciones dificultan la comparación entre microarrays. Para poder solucionar este tipo de problemas y conseguir así que sean comparables los microarrays es necesario eliminar el ruido de fondo y escalar tanto las distintas zonas de un microarrays como los distintos microarrays a la misma escala. Este proceso es el denominado proceso de normalización. En este trabajo se emplea la corrección de fondo RMA junto con la normalización por cuantiles, extrayendo únicamente los perfect match (PM) y empleando "median polish" como el método de cálculo de la expresión génica de cada spot ("summarization").

Para la corrección de fondo es interesante señalar que los microarrays de Affymetrix incluyen sondas diseñadas para que se una perfectamente la secuencia de interés (Perfect Match probe, PM) y sondas diseñadas que contienen una base modificada en el centro de la sonda (Mismatch Probe, MM) [11]. La presencia MM nos permite identificar y cuantificar la cantidad de señal de hibridación no específica que se produce en el microarray. De este modo, la presencia de PM en relación con la cantidad de MM puede

servirnos para diferenciar el ruido de genes que se expresan a bajos niveles [11, 12]. Para esto, se emplea el método de corrección Robust Multi-Array Analysis (RMA) y se seleccionan los PM para calcular el nivel de expresión de los genes. El método de corrección de fondo RMA asume que los PM son combinación del fondo y de la señal real y, por tanto, calcula la señal real como probabilidad posterior dado los PM mediante el teorema de Bayes[11, 12]. Aunque es un método popular, es en algunos casos subestima el ruido de fondo al no considerar la distinta tendencia de los spots a sufrir hibridación no específica [11, 12]. No obstante, se considera un método robusto y ampliamente usado teniendo como ventaja que aplica la corrección de fondo de forma individualizada a cada microarray.

La normalización por cuantiles se emplea para escalar las distribuciones de intensidad de los diferentes microarrays para que así sean comparables. La normalización por cuantiles calcula la distribución empírica procedente de la media de los cuantiles de los diferentes microarrays y la emplea para indexar la distribución empírica original de cada microarray [12]. De este modo, tras la normalización por cuantiles todas las muestras presentarán la misma distribución de intensidades.

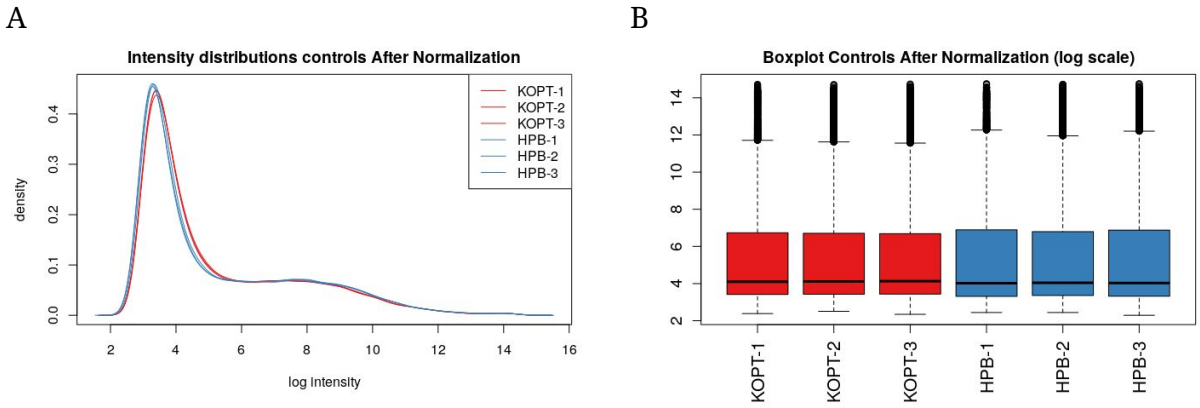


Figura 2 - Distribución de las intensidades (A) y boxplots (B) de las muestras control (DMSO) de ambas líneas celulares después de normalizar y corregir el fondo.

El análisis de los datos tras la normalización puede observarse en la Figura 2. En esta figura observamos el efecto de la normalización, siendo ahora la distribución mucho más homogénea entre las muestras.

Uno de los gráficos más simples, pero muy útil, para comparar los niveles de expresión de genes entre dos condiciones son los denominados “scatter-plot”. En este tipo de gráfico cada eje corresponde a un microarray y los puntos a los distintos genes distribuidos en el espacio en función de su nivel de expresión. Este tipo de gráfico es muy

útil dado que los genes con niveles de expresión similar se sitúan en la diagonal del gráfico, existiendo más diferencias entre los microarrays de estudio cuantos más genes se separen de la diagonal.

En la Figura 3A se muestra el scatter-plot correspondiente a la comparación de los microarrays de HPB-ALL en DMSO y KOPT-K1 en DMSO. Se emplea la mediana de las 3 réplicas por cada línea celular para generar el microarray consenso de cada línea celular. La mediana es preferida a la media por su robustez [14]. En el scatter-plot podemos comprobar cómo algunos genes se encuentran separados de la diagonal.

En la Figura 3B se muestra el MA-plot. El MA-plot es otro tipo de gráfico con la misma utilidad que el scatter-plot, pero en este caso, se representa el ratio de la intensidad (M) frente a la intensidad media (A). De este modo, los genes que no muestran una expresión diferencial entre ambas líneas celulares se situarán en y = 0, mientras que los genes que se separan por encima o por debajo de este valor se encontrarán diferencialmente expresados. De nuevo, en la Figura 3B observamos que una parte de los genes se encuentran diferencialmente expresados.

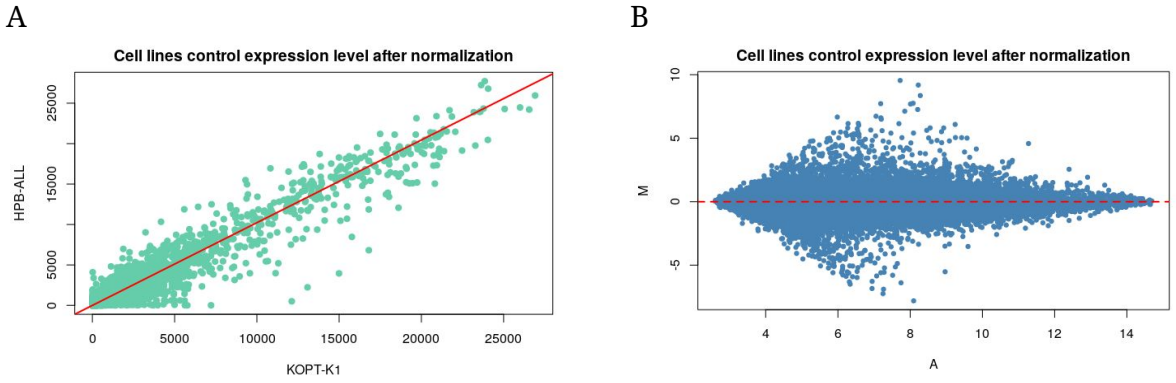


Figura 3 - Scatter-plot (A) y MA-plot (B) de las muestras control (DMSO) de ambas líneas celulares después de normalizar y corregir el fondo.

Otra herramienta muy útil para analizar las diferencias entre los diferentes microarrays es el uso del gráfico de componentes principales (Principal component analysis, PCA). Este tipo de visualización es muy útil sobre todo para visualizar datos que contienen una gran cantidad de dimensiones y un número reducido de muestras [13]. En nuestro caso las dimensiones serían los diferentes genes analizados en el microarray (54675 genes), por lo que este tipo de visualización es necesaria.

En la Figura 4 podemos ver como los dos componentes principales retienen el 92% de la varianza de los datos, de modo que la disposición de los datos en el espacio es fiable.

Podemos ver el primer componente de la PCA (87% varianza) separa los microarrays en función del tipo celular. Esto indica que efectivamente hay diferencias entre ambos tipos celulares. Igualmente, el segundo componente de la PCA (4%) muestra como los microarrays de las líneas celulares HPB-ALL presentan más diferencias entre sí que los microarrays pertenecientes a la línea KOPT-K1, dado que estos se mantienen agrupados. A pesar de que las réplicas de HPB-ALL presenten heterogeneidad entre sí, no consideramos que sea necesario eliminar ninguno de los microarrays, puesto que la fiabilidad estadística mejora con la cantidad de información que dispongamos y despreciar una muestra supone perder información. Además, en el principal componente que retiene el 87% de la varianza estas muestras no muestran diferencias.

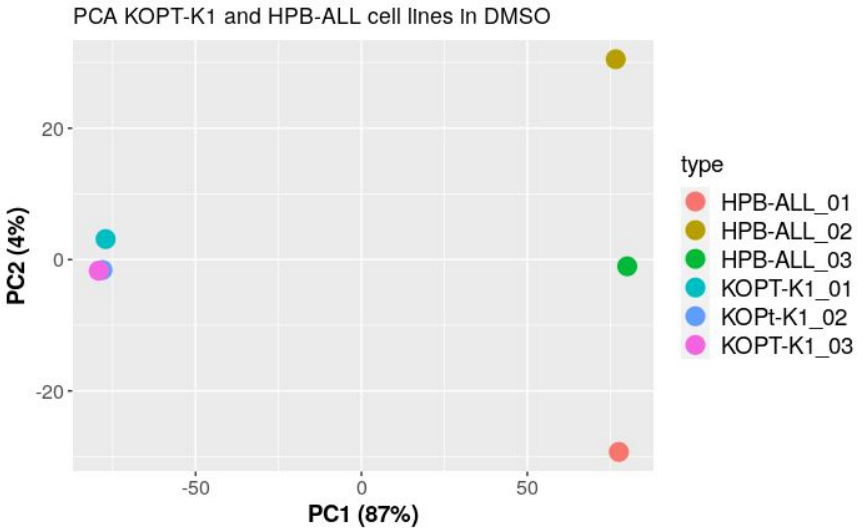


Figura 4 - Gráfico de los 2 componentes principales (91.6% de la varianza) de los microarrays en condición control de las dos líneas celulares.

Con el análisis llevado a cabo hasta ahora ya podemos sospechar que existen diferencias entre ambas líneas celulares. No obstante, estos gráficos no están apoyados por ningún test estadístico que nos indique si esas diferencias son o no significativas. En consecuencia, se lleva a cabo una comparación estadística con “empirical Bayes” del paquete limma de Bioconductor [15]. Para facilitar los cálculos, el análisis estadístico sólo se lleva a cabo sobre los genes se encuentran por encima del rango intercuartílico 0.5, consiguiendo así emplear sólo genes que muestran variabilidad y que son susceptibles de presentar diferentes niveles de expresión en ambas líneas celulares. Con este filtrado pasamos de tener que comparar 54675 genes a comparar 27335 genes.

Es importante señalar que al comparar un gran número de genes entre dos líneas celulares es posible que por azar algún p-valor sea significativo sin serlo realmente. Para

evitar esta situación se emplea la corrección FDR de Benjamini-Hochberg (BH). Los genes que presenten un p-valor ajustado menor o igual al 0.05 son considerados como genes diferencialmente expresados (DEGs). Para poder conocer qué genes forman parte de los DEGs se emplea la notación asociada a la tecnología Affymetrix del artículo de referencia [1]: `hgu133plus2.db`. Se emplea la función `mapId` del paquete AnnotationDb dado que se comprobó que incorpora mayor información sobre los spots del microarray.

Se comprobó que los genes diana de NOTCH1 (*MYC*, *DTX1*, *HES1* y *HES4*) identificados por los autores del artículo de referencia [1] como sub-regulados cuando son tratados con SAMH1, ya se encuentran entre los genes diferencialmente expresados con un p-valor ajustado menor o igual que 0.05. El número total de genes diferencialmente expresados entre ambas líneas celulares es de 21828. De estos 1865 sobre-expresados considerando un log Fold Change superior a 1 y 3016 sub-expresados considerando un log Fold Change inferior a -1. En la Figura 5 se muestra un volcano-plot donde se puede observar que estos genes identificados por los autores ya se encuentran diferencialmente expresados y de forma significativa entre ambas líneas celulares. De este modo, hemos comprobado que el efecto de SAMH1 debe ser estudiado por separado en cada línea celular.

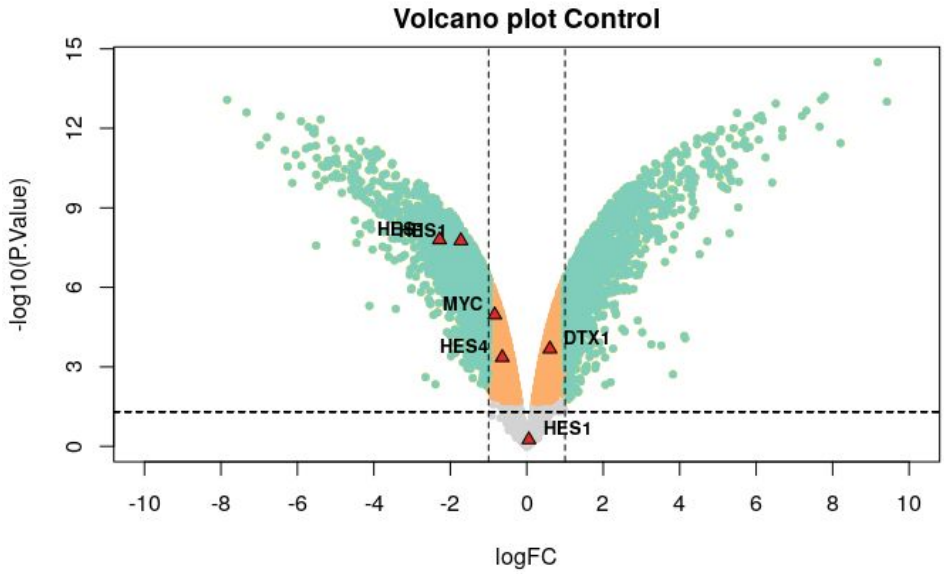


Figura 5 - Volcano plot que muestra en azul los genes con un log Fold Change superior/inferior a ±1 y un p-valor ajustado inferior a 0.05; en naranja los genes con un Fold Change menor a 1 y con un p-valor ajustado inferior a 0.05; y en gris los genes con un Fold Change menor que 1 y un p-valor ajustado mayor a 0.05. Con triángulos se señalan la posición en el volcano plot de los genes identificados en el paper como sub-regulados al aplicar SAMH1.

Estudio del efecto de SAMH1 en la línea celular KOPT-K1 mediante expresión diferencial

El proceso realizado para comparar la situación control tratada con DMSO entre las dos líneas celulares se repite para comparar el tratamiento con DMSO y el tratamiento con SAMH1 en la línea celular KOPT-K1. En la Figura 6A 6B se muestran los datos de intensidad sin normalizar mientras que en la Figura 6C 6D se muestran los datos normalizados.

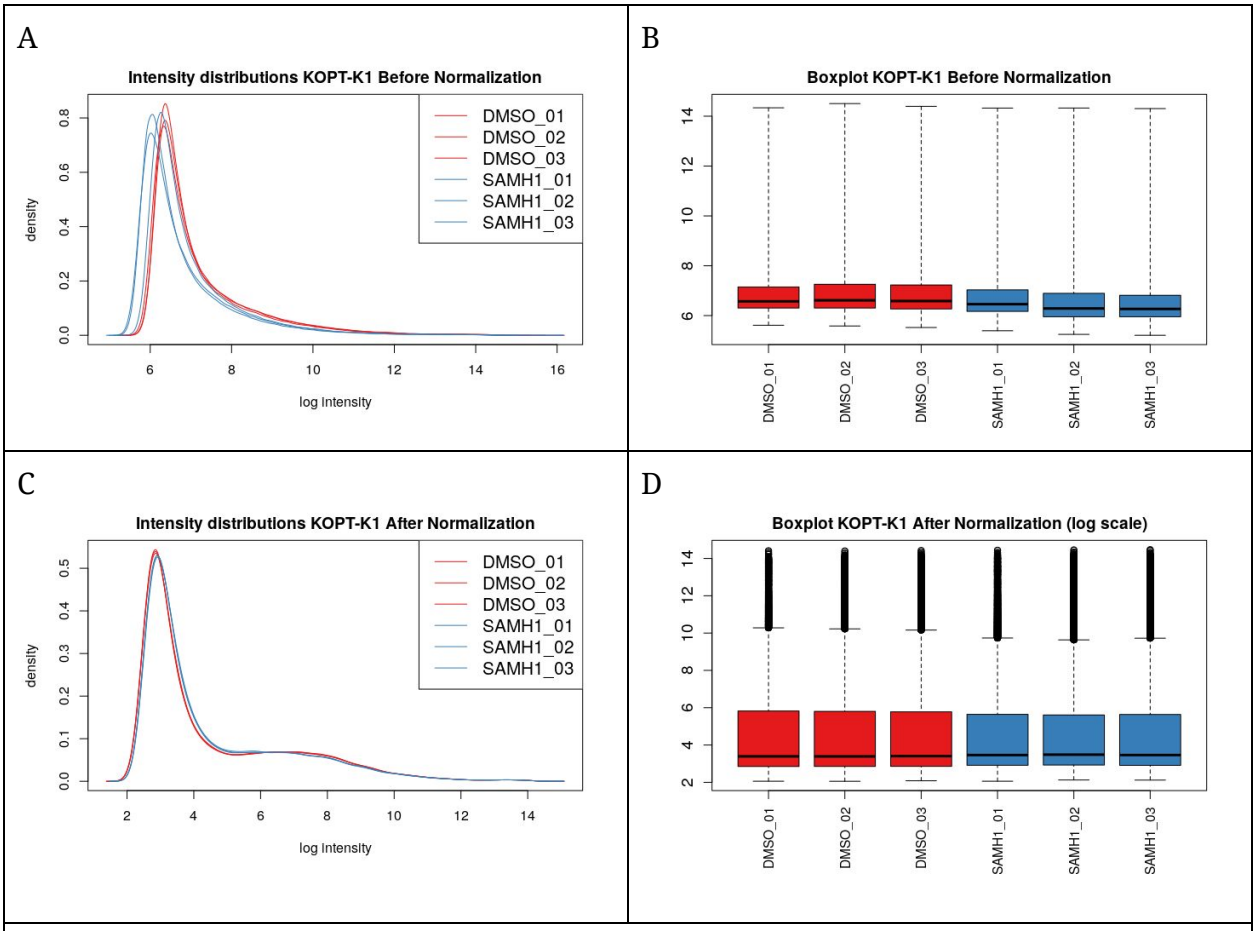


Figura 6 - Distribución de las intensidades (A) y boxplots (B) de la línea celular KOPT-K1 antes de normalizar y corregir el fondo. Distribución de las intensidades (C) y boxplots (D) de la línea celular KOPT-K1 después de normalizar y corregir el fondo.

Comprobamos que la normalización y corrección de fondos son útiles para escalar los microarrays de la línea celular KOPT-K1 en la condición control DMSO y con el tratamiento SAMH1.

La presencia de genes diferencialmente expresados entre ambas condiciones en la línea celular KOPT-K1 puede observarse en el scatter-plot y MA-plot (Figura 7). En ambos gráficos podemos observar que la mayor parte de los genes no se encuentran

diferencialmente expresados. No obstante, existen algunos genes que sí muestran diferencias en los niveles de expresión entre ambas condiciones, predominando los genes sub-regulados como se puede ver en el MA-plot.

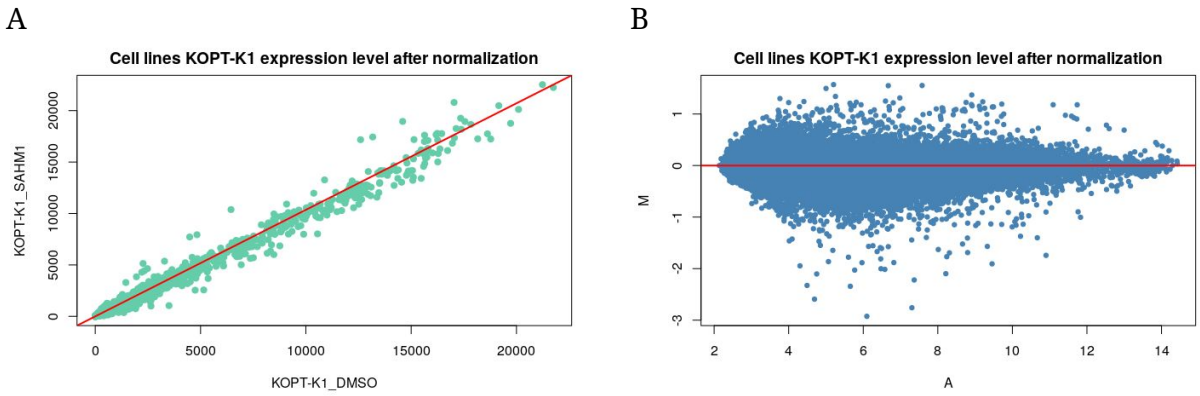


Figura 7 - Scatter-plot (A) y MA-plot (B) de las muestras tratadas con DMSO y con SAMH1 de la línea celular KOPT-K1.

La presencia de diferencias en la línea celular KOPT-K1 cuando es tratada con SAMH1 con respecto a la situación control también se observa en la PCA de la Figura 8. Podemos ver cómo el primer principal componente (55.78%) separa las muestras tratadas con DMSO de las tratadas con SAMH1. También se observa que las muestras tratadas con SAMH1 presentan mayor heterogeneidad, siendo separadas a lo largo del segundo principal componente. No obstante, no se desprecia ningún microarray dado que este segundo componente tan solo retiene el 13.43% de la varianza.

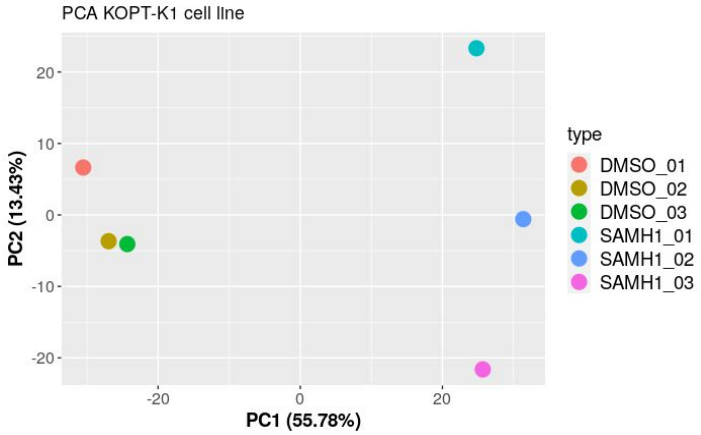


Figura 8 - Gráfico de los 2 componentes principales (69% de la varianza) de los microarrays tratados con DMSO y SAMH1 en la línea celular KOPT-K1.

De nuevo, la mejor forma de asegurar la presencia de genes diferencialmente expresados es mediante el apoyo de un test estadístico. Adicionalmente, se aplica el filtro del rango intercuartílico que reduce el número de genes a comparar de 54675 a 27337. Ajustando el p-valor mediante FDR de BH y seleccionando aquellos genes con un valor menor o igual a 0.05, reducimos finalmente el número de genes diferencialmente expresados a 7765 genes.

Para validar los resultados encontrados por los autores del paper en la Figura 9 se muestra el volcano-plot correspondiente a la expresión diferencial de genes entre DMSO y SAMH1 en la línea celular KOPT-K1. En este gráfico podemos comprobar cómo los genes diana de NOTCH1 (*MYC*, *DTX1*, *HES1* y *HES4*) identificados por los autores del artículo de referencia como sub-regulados cuando son tratados con SAMH1 efectivamente se encuentran sub-expresados al presentar un log FC menor que -0.5.

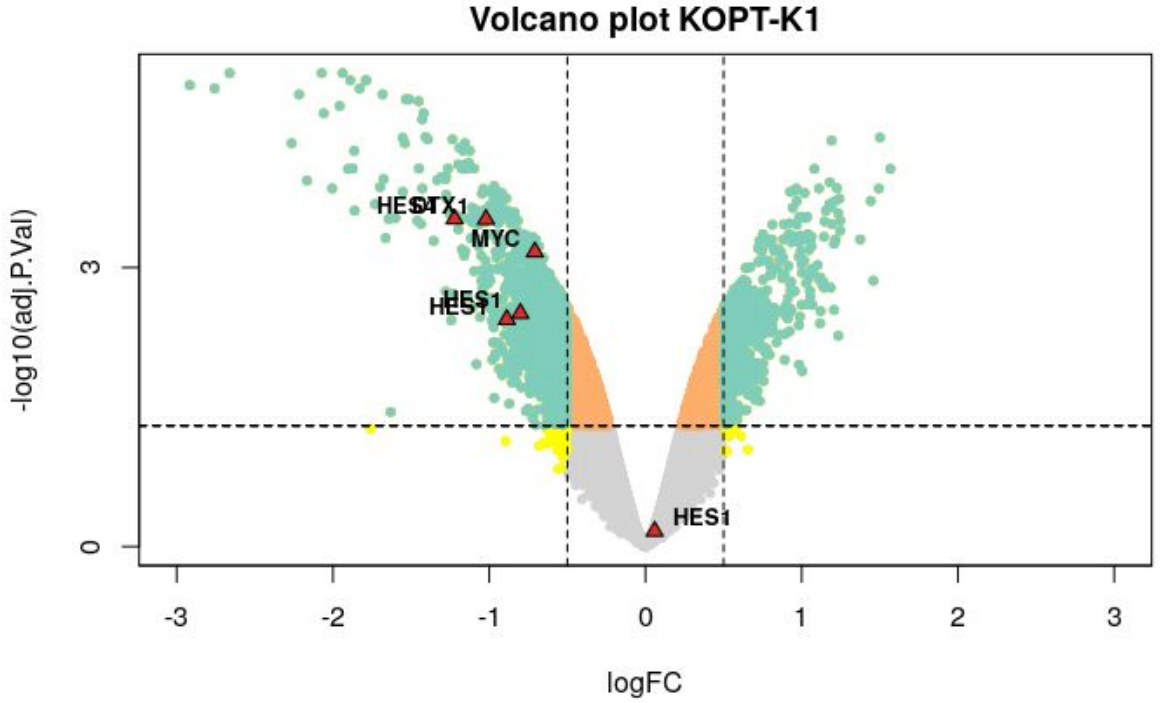


Figura 9 - Volcano plot que muestra en azul los genes con un log Fold Change superior/inferior a ±0.5 y un p-valor ajustado inferior a 0.05; en naranja los genes con un Fold Change entre -0.5 y 0.5 y con un p-valor ajustado inferior a 0.05; y en gris los genes con un Fold Change entre -0.5 y 0.5 y un p-valor ajustado mayor a 0.05. Con triángulos se señalan la posición en el volcano plot de los genes identificados en el paper como sub-regulados al aplicar SAMH1.

Entre el tratamiento con DMSO y SAMH1 en la línea celular KOPT-K1, considerando un log FC de ±0.5 existen 576 sobre-expresados cuando se aplica SAMH1 y 1579 genes sub-expresados, entre ellos los identificados por los autores. Por lo tanto, hemos corroborado los resultados obtenidos por el artículo de referencia para la línea KOPT-K1.

Estudio del efecto de SAMH1 en la línea celular HPB-ALL mediante expresión diferencial

De forma análoga al análisis de la línea celular KOPT-K1 se lleva a cabo el análisis de expresión diferencial de la línea celular HPB-ALL.

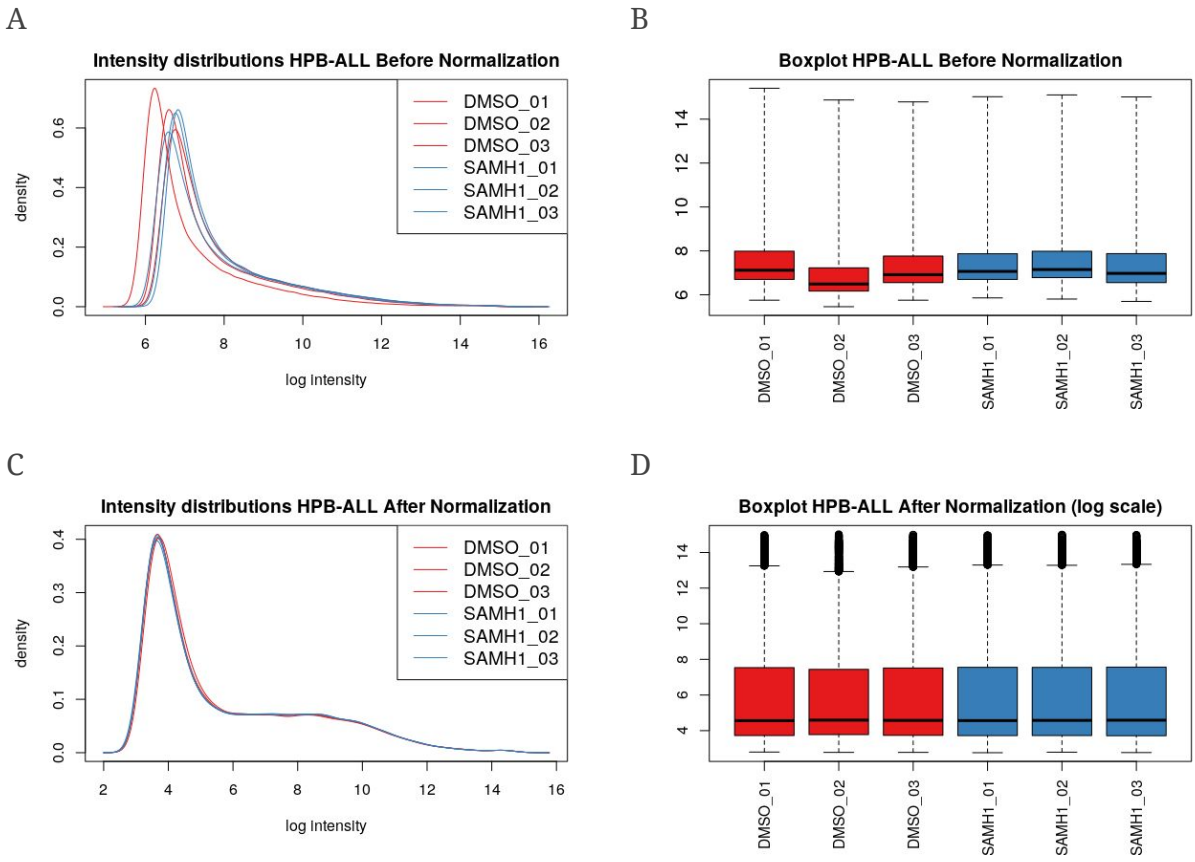


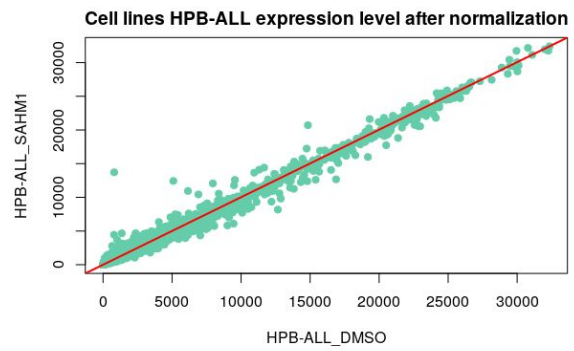
Figura 10 - Distribución de las intensidades (A) y boxplots (B) de la línea celular HPB-ALL antes de normalizar y corregir el fondo. Distribución de las intensidades (C) y boxplots (D) de la línea celular HPB-ALL después de normalizar y corregir el fondo.

Comprobamos que la normalización y corrección de fondos son útiles para escalar los microarrays de la línea celular HPB-ALL en la condición control DMSO y con el tratamiento SAMH1.

La presencia de genes diferencialmente expresados entre ambas condiciones en la línea celular HPB-ALL puede observarse en el scatter-plot y MA-plot (Figura 11). En ambos

gráficos podemos observar que la mayor parte de los genes no se encuentran diferencialmente expresados. No obstante, existen algunos genes que sí muestran diferencias en los niveles de expresión entre ambas condiciones, predominando los genes sobre-regulados como se puede ver en el MA-plot.

A



B

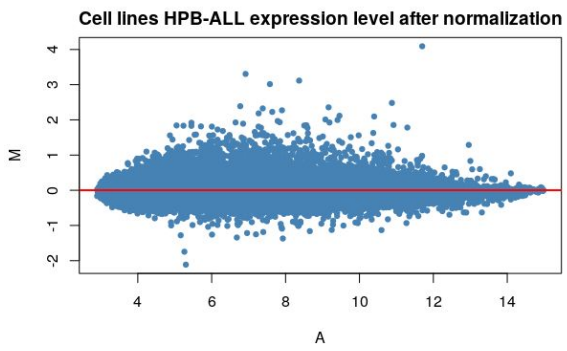


Figura 11 - Scatter-plot (A) y MA-plot (B) de las muestras tratadas con DMSO y con SAMH1 de la línea celular KOPT-K1.

La presencia de diferencias en la línea celular HPB-ALL cuando es tratada con SAMH1 con respecto a la situación control también se observa en la PCA de la Figura 12. En este caso son las muestras control las que mayor heterogeneidad muestran. Podemos ver cómo el primer principal componente (52%) separa las muestras tratadas con DMSO de las tratadas con SAMH1.

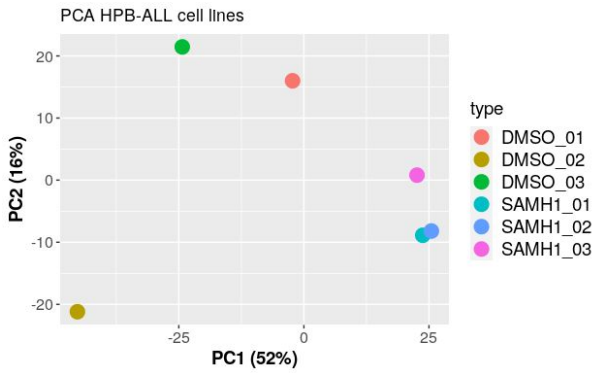


Figura 12 - Gráfico de los 2 componentes principales (69% de la varianza) de los microarrays tratados con DMSO y SAMH1 en la línea celular KOPT-K1.

Finalmente, la expresión diferencial de genes en la línea celular HPB-ALL se muestra con el volcano-plot (Figura 13). En esta representación volvemos a comprobar que los genes identificados por los autores se vuelven a encontrar sub-expresados en HPB-ALL cuando es tratada con SAMH1. De este modo podemos concluir que el efecto de SAMH1 es independiente de la línea celular.

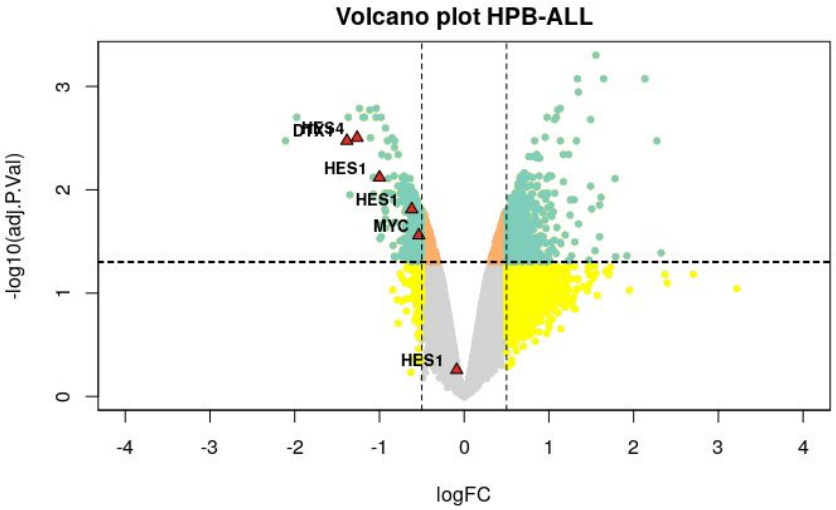


Figura 13 - Volcano plot que muestra en azul los genes con un log Fold Change superior/inferior a ±0.5 y un p-valor ajustado inferior a 0.05; en naranja los genes con un Fold Change entre -0.5 y 0.5 y con un p-valor ajustado inferior a 0.05; y en gris los genes con un Fold Change entre -0.5 y 0.5 y un p-valor ajustado mayor a 0.05. Con triángulos se señalan la posición en el volcano plot de los genes identificados en el paper como sub-regulados al aplicar SAMH1.

CLUSTERIZACIÓN

Otra estrategia de análisis de los datos procedentes de microarrays es el análisis de clasificación no supervisada mediante clustering. El término "clasificación no supervisada" hace referencia al hecho de que se emplean algoritmos que clasifican las muestras o genes según su nivel de similitud sin conocer previamente el grupo al que pertenecen [10].

El clustering permite encontrar relaciones desconocidas tanto entre las muestras como entre genes. Dentro del clústering encontramos el clustering jerárquico y el clustering no jerárquico. El primero de ellos permite analizar la relación de dependencia entre los grupos sin necesidad de definir previamente el número de clústeres que se han de

encontrar, un ejemplo sería el algoritmo UPGMA. Por el contrario, el segundo de ellos no indica ninguna relación entre grupos y necesita tener definido previamente el número de clústeres en los que clasificar las muestras, por ejemplo K-means [10].

Con la herramienta Morpheus del Broad Institute (https://software.broadinstitute.org/morpheus/) se pueden realizar tanto clusterizaciones jerárquicas como no jerárquicas.

Ambos métodos de clústeres dependen de la medida de la distancia entre grupos para poder inferir la similitud entre los mismos. Existen dos grandes grupos de distancias, las distancias absolutas (euclídeas) y distancias basadas en tendencias (correlacionadas). Según el criterio de clasificación un tipo de distancia es más adecuada que la otra [10].

Las distancias absolutas emplean la distancia geométrica entre dos puntos. Este tipo de distancias es sensible a la multi-dimensionalidad, dado que a medida que aumenta el número de dimensiones más vacío se encuentra el espacio (maldición de la dimensionalidad). En nuestro caso, las dimensiones serían los valores de expresión para cada uno de los genes analizados en el microarray. No obstante, este tipo de distancias nos permite analizar las muestras en función de similitudes globales.

Por su parte, las distancias basadas en tendencias miden la correlación que existen entre dos dimensiones de la muestra y, por lo tanto, permiten agrupar los genes según su patrón de expresión. En otras palabras, este tipo de distancias analizan las tendencias en la expresión de genes en distintas muestras para llevar a cabo la clasificación.

Para analizar los microarrays se emplearán tanto métodos de clusterización jerárquicos como métodos de clusterización no jerárquicos (K-means). Los métodos jerárquicos se realizan en primer lugar para conocer el número de clústeres presentes en los datos y poder así dirigir la clasificación en los métodos no jerárquicos. Igualmente, se emplea la distancia euclídea (absolutas) para verificar si los triplicados de las diferentes condiciones se agrupan conjuntamente, y la distancia basada en el coeficiente de Pearson (correlación) para detectar genes con el mismo patrón de expresión entre diferentes condiciones.

Comparación líneas celulares: KOPT-K1 DMSO vs HPB-ALL DMSO

Para apoyar el hecho de que las líneas celulares KOPT-K1 y HPB-ALL muestran diferencias significativas en la propia situación control (DMSO), se realiza una clusterización jerárquica y no jerárquica de los 35 genes significativamente más diferencialmente expresados según el p-valor ajustado por BH mediante la distancia

euclídea (Figura 14). En este caso se emplea una distancia absoluta porque nos interesa conocer si las similitudes globales de los distintos microarrays permite agrupar o no conjuntamente las dos líneas celulares. En esta Figura podemos ver como claramente ambas líneas celulares se agrupan en conjuntos diferentes.

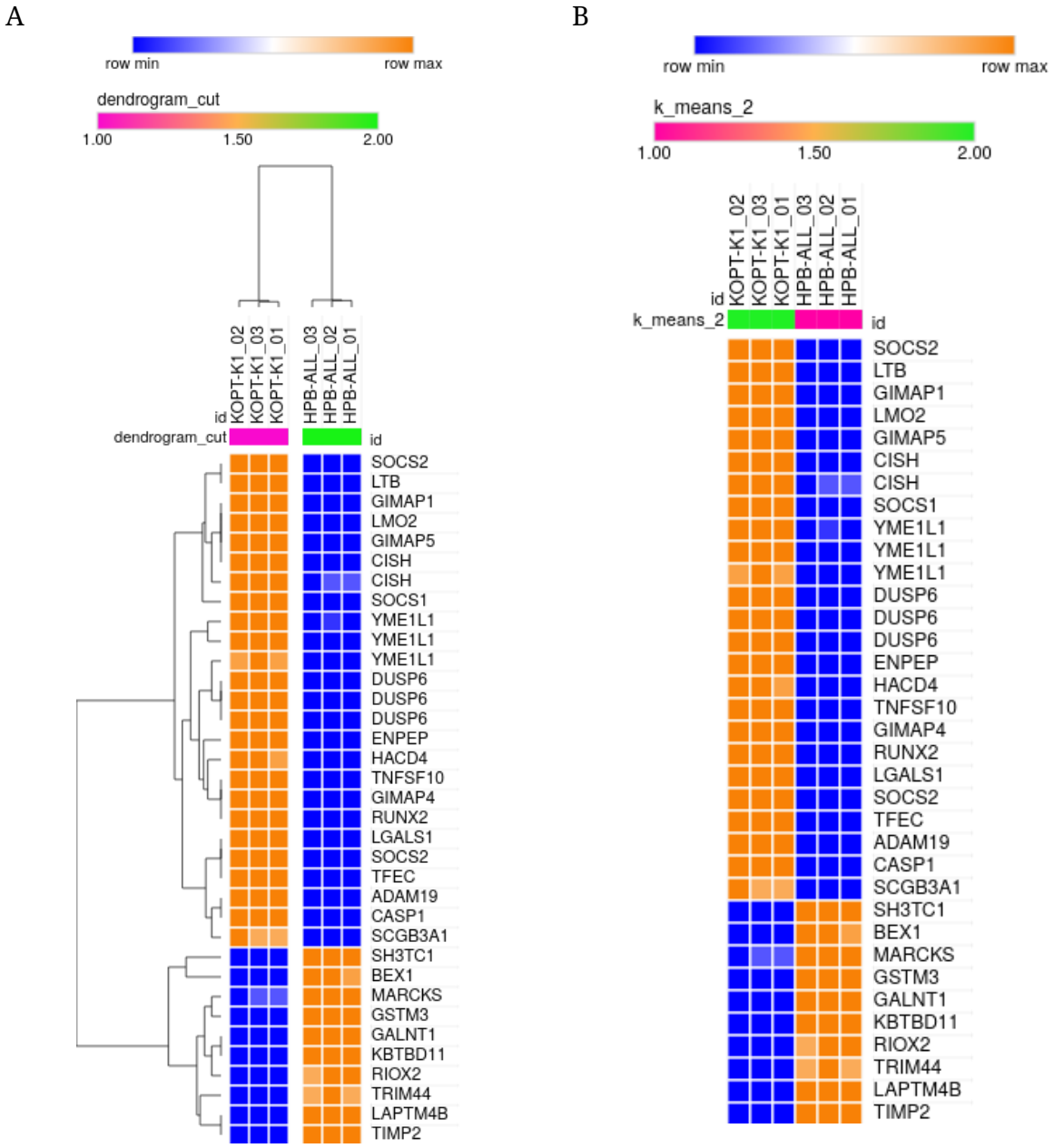


Figura 14 - Clusterización jerárquica (A) y no jerárquica (B) realizada con Morpheus [3] con distancia euclídea de las muestras control (DMSO) de dos líneas celulares en función de los 35 top genes con un p-valor ajustado menor a 0.05. Estos genes presentan un p-valor ajustados menor a 0.05. La clusterización no jerárquica se realiza mediante K-means con K = 2.

Igualmente, seleccionando los genes *MYC*, *HES1*, *HES4* y *DTX1* señalados en el artículo de referencia como genes diana de NOTCH-1 sub-regulados en presencia de SAMH1 (Figura 15) podemos ver cómo las muestras procedentes de una misma línea celular se agrupan entre sí, pero se diferencian de la otra línea celular.

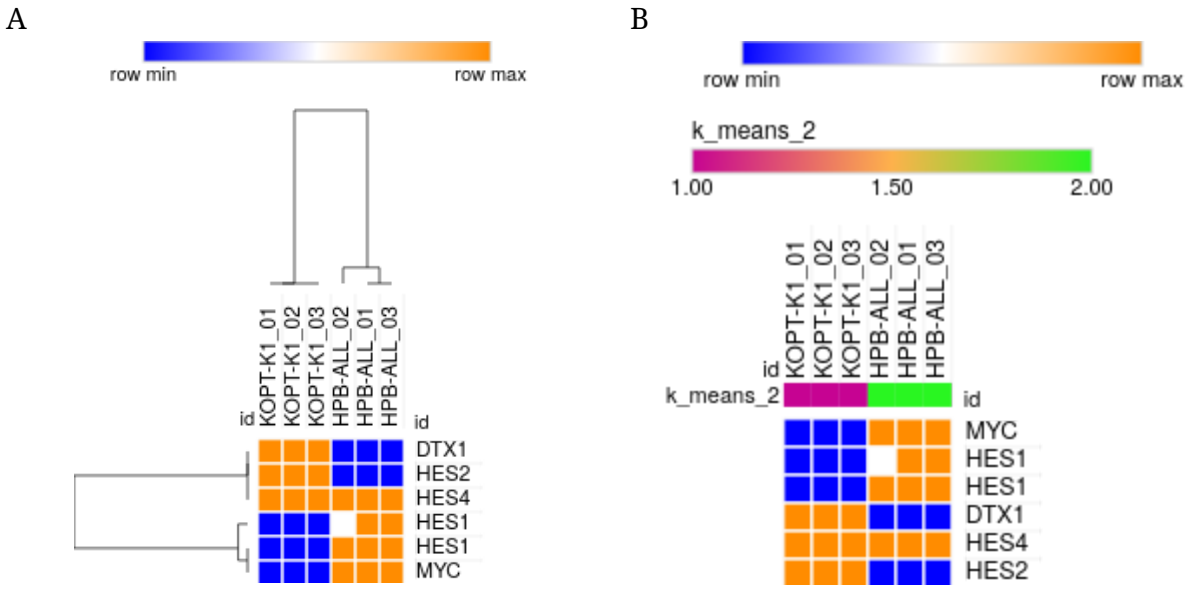


Figura 15 - Clusterización jerárquica (A) y no jerárquica (B) realizada con Morpheus [3] de las muestras control (DMSO) de dos líneas celulares en función de los genes diana de NOTCH-1 señalados en el artículo de referencia (*MYC*, *HES1*, *HES4* y *DTX1*). Estos genes presentan un p-valor ajustados menor a 0.05. La clusterización no jerárquica se realiza mediante K-means con K = 2.

Estudio del efecto de SAMH1 en la línea celular KOPT-K1

De forma análoga se analiza la clusterización jerárquica y no jerárquica sobre la línea celular KOPT-K1 en la tratadas con DMSO y tratadas con SAMH1.

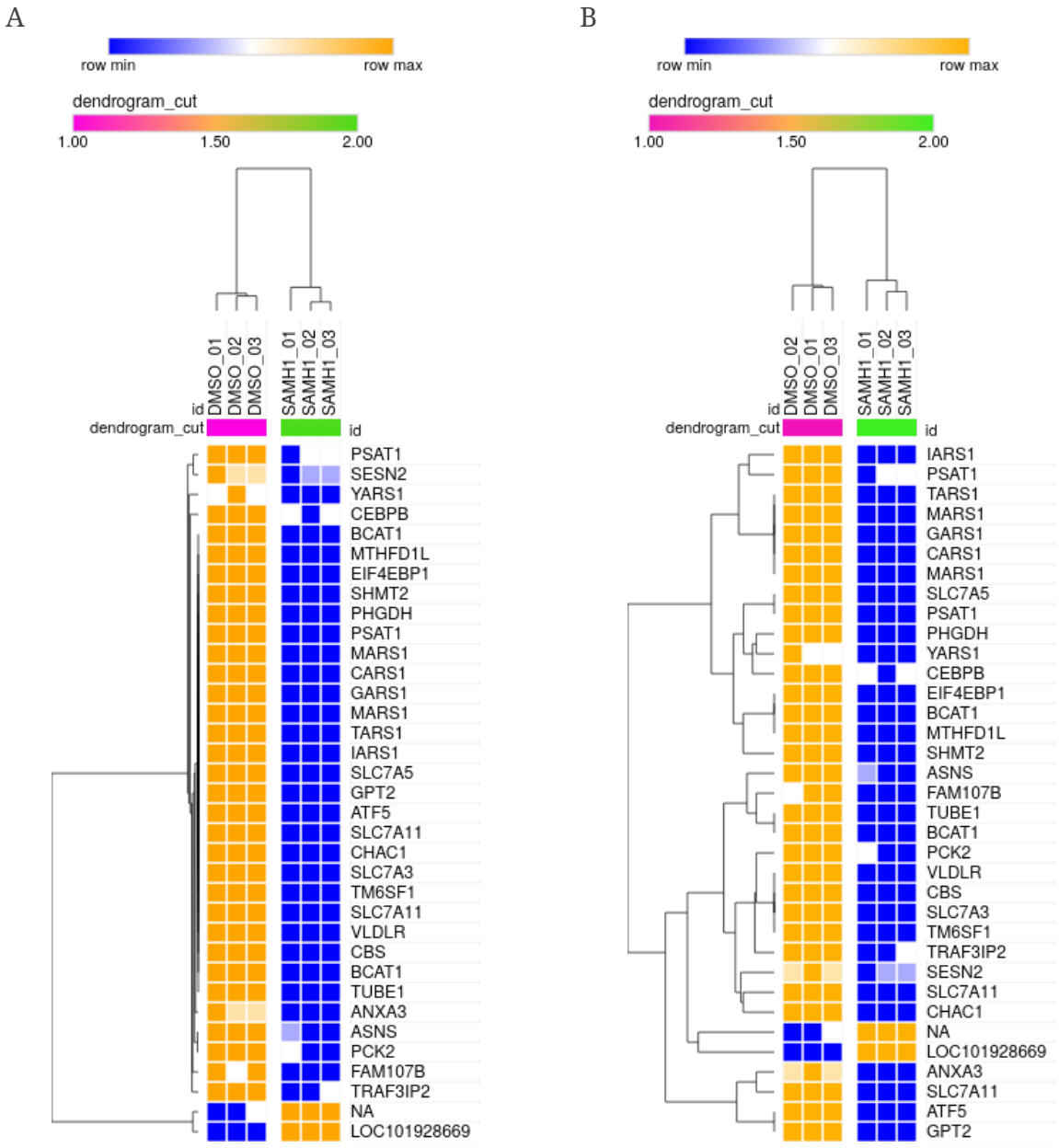


Figura 16 - Clusterización jerárquica con coeficiente de Pearson (A) y distancia euclídea (B) realizada con Morpheus [3] del top 35 genes de la línea KOPT-K1 tratadas con DMSO y SAMH1. Estos genes presentan un p-valor ajustados menor a 0.05.

En la Figura 16 podemos comparar el efecto de la distancia empleada en la clusterización jerarquizada. La Figura 16 A muestra la clusterización basada en el coeficiente de Pearson mientras que la Figura 16 B muestra la clusterización basada en la distancia euclídea. A la hora de clasificar las muestras ambas agrupan las réplicas de cada tratamiento de forma conjunta. No obstante, el efecto de la distancia se observa en mayor medida a la hora de clasificar los genes. Podemos comprobar como la distancia euclídea encuentra muchas sub-agrupaciones a menor nivel mientras que el coeficiente de Pearson encuentra dos grandes grupos en función del patrón de expresión. Esto nos indica que la mayoría de los top 35 genes diferencialmente expresados presentan la misma tendencia a sufrir sub-expresión en presencia de SAMH1.

Igualmente, de entre los genes diferencialmente expresados podemos clusterizar las muestras en función de los genes identificados por los autores (Figura 17). Podemos ver cómo estos genes se encuentran sub-regulados en presencia de SAMH1. No obstante, tanto la clusterización jerárquica como la no jerárquica agrupa las muestras tratadas con DMSO con dos de las tres muestras tratadas con SAMH1, lo que corrobora la heterogeneidad que observábamos en estas muestras en la Figura 8.

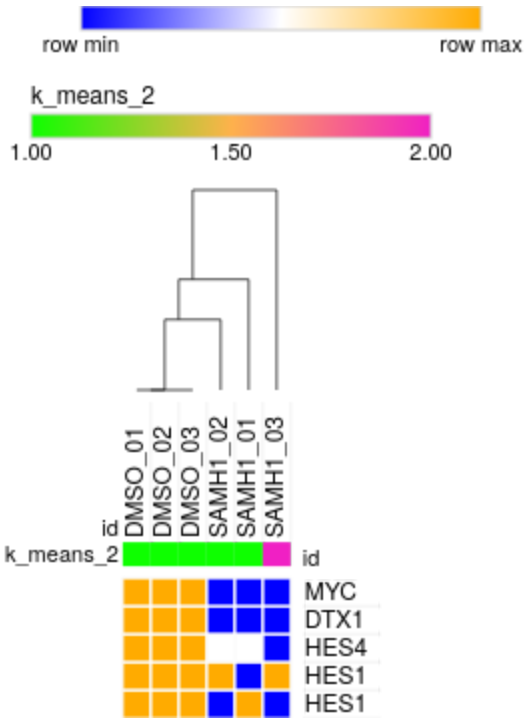


Figura 17 - Clusterización jerárquica con coeficiente de Pearson y no jerárquica realizada con Morpheus [3] de los genes diana de NOTCH-1 señalados en el artículo de referencia (*MYC*, *HES1*, *HES4* y *DTX1*) de la línea KOPT-K1 tratadas con DMSO y SAMH1. Estos genes presentan un p-valor ajustados menor a 0.05.

Estudio del efecto de SAMH1 en la línea celular HPB-ALL

El mismo análisis se lleva a cabo para la línea celular HPB-ALL. De nuevo el análisis del top 35 genes diferencialmente expresados permite agrupar las réplicas conjuntamente y separar los tratamientos entre sí, tanto en clusterización jerárquica (Figura 18A) como no jerárquica (Figura 18B). Cabe destacar que entre estos top 35 genes se encuentra DTX1 y HES4.

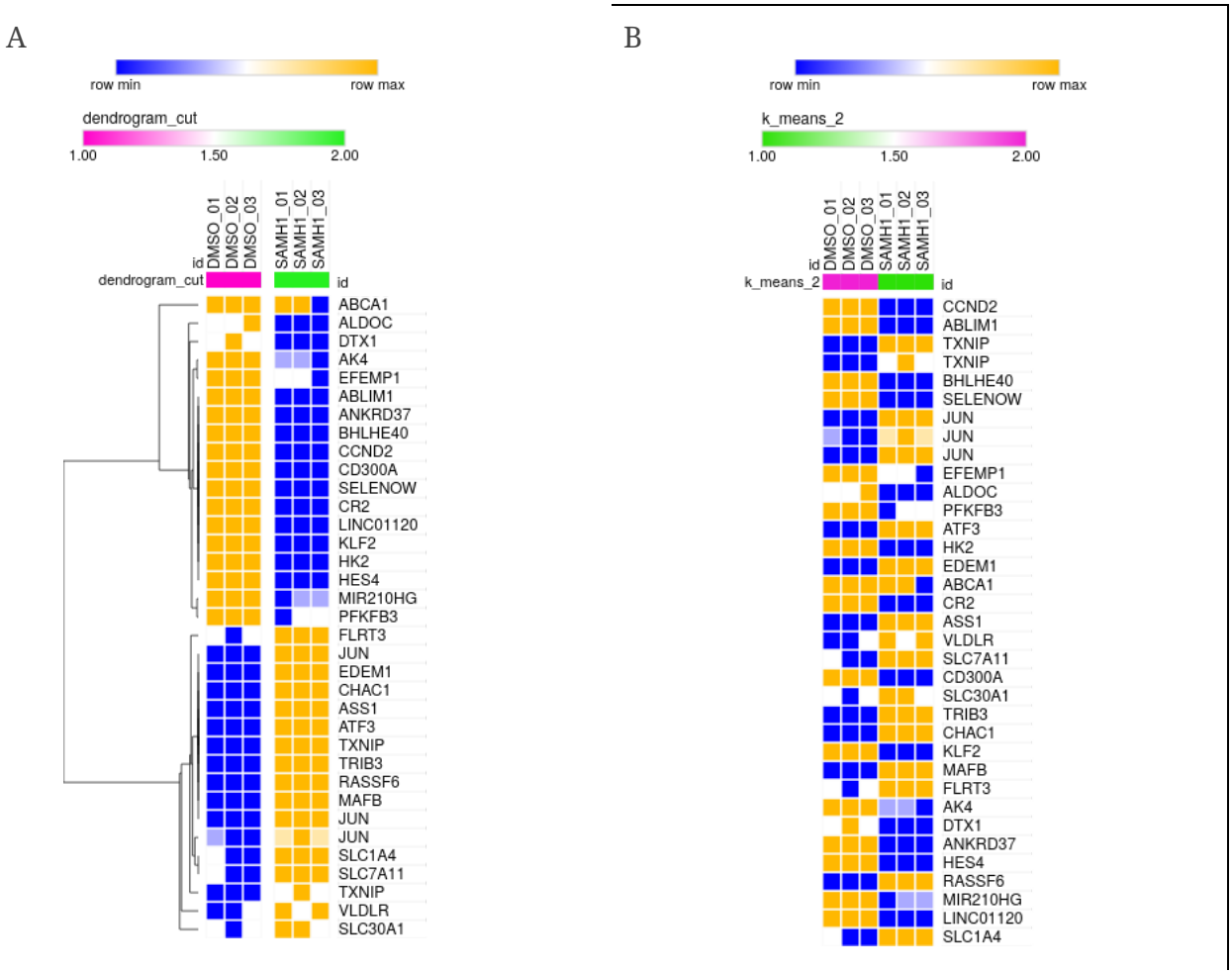


Figura 18 - Clusterización jerárquica con coeficiente de Pearson (A) y no jerárquica (B) realizada con Morpheus [3] del top 35 genes de la línea HPB-ALL tratadas con DMSO y SAMH1. Estos genes presentan un p-valor ajustados menor a 0.05. La clusterización no jerárquica se realiza mediante K-means con K = 2.

En el caso de la línea HPB-ALL los genes diana de NOTCH1 sub-regulados en presencia de SAMH1 identificados por los autores permite separar correctamente las muestras (Figura 19).

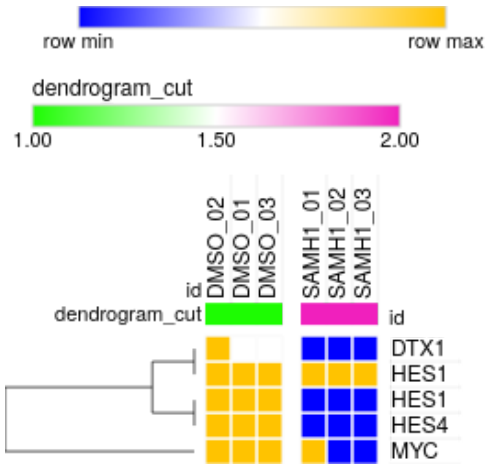


Figura 19 - Clusterización jerárquica con coeficiente de Pearson y no jerárquica realizada con Morpheus [3] de los genes diana de NOTCH-1 señalados en el artículo de referencia (*MYC*, *HES1*, *HES4* y *DTX1*) de la línea HPB-ALL tratadas con DMSO y SAMH1. Estos genes presentan un p-valor ajustados menor a 0.05.

PREDICCIÓN DE CLASE

También se ha probado a generar un clasificador mediante aprendizaje supervisado. La utilidad de los métodos de aprendizaje supervisado es que, una vez entrenados y comprobado que el error de generalización es el deseado, se pueden emplear para clasificar otras condiciones similares [16]. Tiene gran importancia en farmacología, ya que por ejemplo, podemos emplear el efecto del péptido SAMH1 en los niveles de expresión de diferentes genes como referencia para detectar otros compuestos con el mismo efecto. Cuando nos referimos al mismo efecto queremos decir que el clasificador entrenado con los datos procedentes de SAMH1 será capaz de clasificar los nuevos datos procedentes del empleo de otro compuesto.

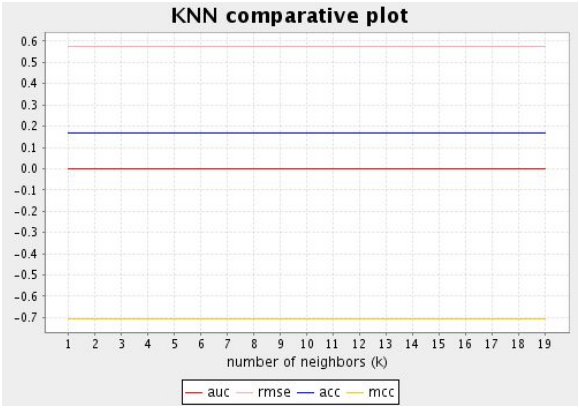
Adicionalmente, durante el proceso de entrenamiento supervisado del clasificador se puede llevar a cabo la selección de genes más importantes para realizar correctamente la clasificación [16]. De este modo, no sólo conseguimos un clasificador robusto capaz de clasificar correctamente las muestras en función de su nivel de expresión sino que además facilita el análisis de los datos al identificar los genes relevantes para el estudio.

Para realizar este tipo de análisis supervisado se empleado la herramienta “Expression/Class Prediction” [4] de Babelomics v.5 [5]. Se emplearon los algoritmos Support Vector Machines (SVM), K-Nearest Neighbor (KNN) y Random Forest. Para la

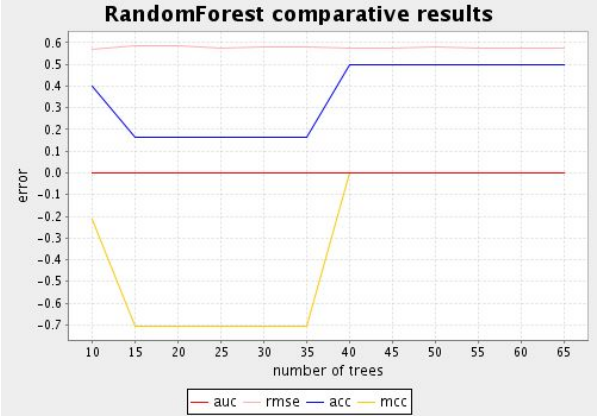
estimación del error de generalización se emplea la validación cruzada con 10 pliegues y 5 repeticiones. Finalmente para la selección de genes se emplea la selección basada en la correlación (CFS).

La herramienta emplea 5 métricas para valorar los distintos clasificadores: área bajo la curva ROC (auc), root-mean-square error (RMSE), Matthews correlation coefficient (MCC) y la precisión (accuracy). Los resultados de las métricas para los distintos clasificadores se muestra en la Figura 20.

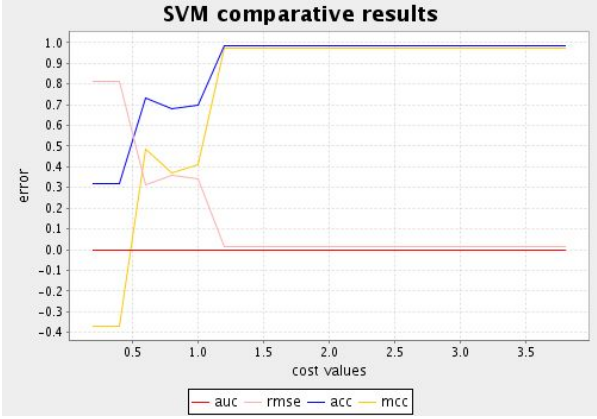
A



B



C



D

Classifier	Parameters	Accuracy	MCC	RMSE	AUC	Selected genes
SVM	cost=0.4, features=1	0.32	-0.37	0.81	0	PXK
SVM	cost=0.6, features=1	0.73	0.48	0.31	0	PXK
SVM	cost=0.8, features=1	0.68	0.37	0.36	0	PXK
SVM	cost=1, features=1	0.7	0.41	0.34	0	PXK
SVM	cost=1.2, features=1	0.98	0.97	0.014	0	PXK
SVM	cost=1.4, features=1	0.98	0.97	0.014	0	PXK
SVM	cost=1.6, features=1	0.98	0.97	0.014	0	PXK
SVM	cost=1.8, features=1	0.98	0.97	0.014	0	PXK
SVM	cost=2, features=1	0.98	0.97	0.014	0	PXK
SVM	cost=2.2, features=1	0.98	0.97	0.014	0	PXK

Figura 20 - Métricas de los clasificadores generados a partir del entrenamiento supervisado con los datos procedentes de la línea celular KOPT-K1.

Se emplean los datos de nivel de expresión procedentes de KOPT-K1 dado que mostraban menos heterogeneidad que los datos procedentes de la línea celular HPB-ALL. Se contempló el empleo de los datos de ambas líneas celulares de forma conjunta, pero la

presencia de variación en los datos procedentes de HPB-ALL incorporaría ruido al entrenamiento y el clasificador tendría riesgo de sobre-ajustar.

En la Figura 20 podemos comprobar cómo el algoritmo SVM con el hiper parámetro de regularización C >= 1.2 muestra una precisión del 98% en la clasificación de las muestras. Este algoritmo clasifica mejor que KNN (17% precisión) y que Random Forest (50% precisión). Por lo tanto el clasificador que emplearemos para clasificar futuras muestras será SVM. Además, para alcanzar esta precisión en la clasificación se a empleado el gen PXK ("PX domain containing serine/threonine kinase like") cuyo mal funcionamiento está relacionado con el Lupus Eritematoso (GeneCards). De este modo, facilitamos el análisis del efecto de compuestos similares a SAMH1 mediante el estudio del nivel de expresión en PXK.

ANÁLISIS FUNCIONAL

Finalmente, para contextualizar el papel biológico de los genes diferencialmente expresados identificados en línea celular KOPT-K1 y HPB-ALL al aplicar el tratamiento con SAMH1, se lleva a cabo un análisis funcional de los mismos. Se emplea el paquete AnnotationDb para anotar los genes diferencialmente expresados y, posteriormente, se emplea la herramienta "Gene List Analysis" [6] de PANTHER v.14 [7] para incorporar anotación y representación gráfica de los términos GO y Pathways presentes en los genes diferencialmente expresados. Es importante señalar que este análisis funcional no es el mismo que lleva a cabo GSEA. En este tipo de análisis empleamos la notación de forma individualizada, mientras que en GSEA se emplean set completos de genes para estudiar el enriquecimiento.

Análisis Funcional de los genes diferencialmente expresados en la línea celular KOPT-K1

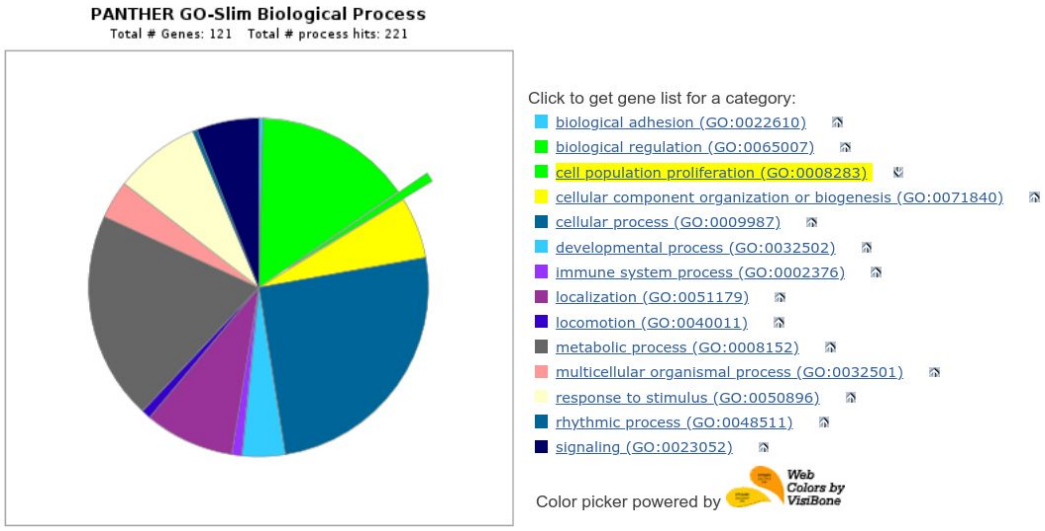
El análisis funcional se lleva a cabo sobre los 576 genes sobre-expresados y 1579 genes sub-expresados que presentan un logFC superior a 0.5 en valor absoluto y un p-valor ajustado inferior a 0.05. De este modo, se analizan los genes que superan un umbral de expresión diferencial determinado y que además son significativos.

Para facilitar la interpretación de los resultados, de entre los genes sub-expresados se emplearon los top 170 entre los que se encontraban *MYC, HES4* y *DTX1*. Lo mismo se realiza sobre los genes sobreexpresados. Los resultados del análisis funcional con PANTHER v.15 se muestran a continuación.

Podemos observar en la Figura 21 que uno de los términos GO presenten entre los 170 genes sub-expresados con p-valor ajustado más significativo es el relacionado con la proliferación celular. También hay otros términos relevantes como la regulación biológica o procesos rítmicos y la señalización. Todos estas funciones se encuentran significativamente sub-reguladas en presencia de SAMH1 en la línea KOPT-K1.

En cuanto a las rutas biológicas alteradas por el tratamiento con SAMH1 la ruta de más abundante es la correspondiente a la biosíntesis de la serina-glicina (P02776) seguida de la ruta de la señalización mediada por el receptor de acetilcolina (P00044). Es interesante señalar que entre las rutas a las que pertenecen los genes sub-rexpresados encontramos la ruta de señalización de NOTCH (Figura 21B). Esto corrobora los resultados encontrados por los autores y los expuestos en el presente trabajo.

A



B

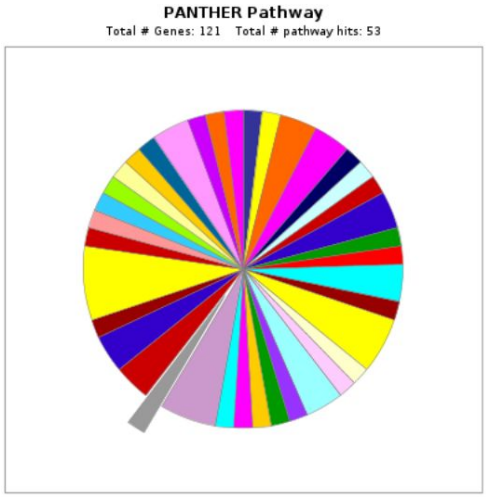


Figura 21 - Visualización gráfica de los términos GO relacionados con los procesos biológicos (A) y rutas biológicas presentes entre los top 170 genes sub-expresados en la línea celular KOPT-K1. En (A) se señala el término GO “cell population proliferation process”. En (B) se destaca la ruta NOTCH signaling pathway (P00045).

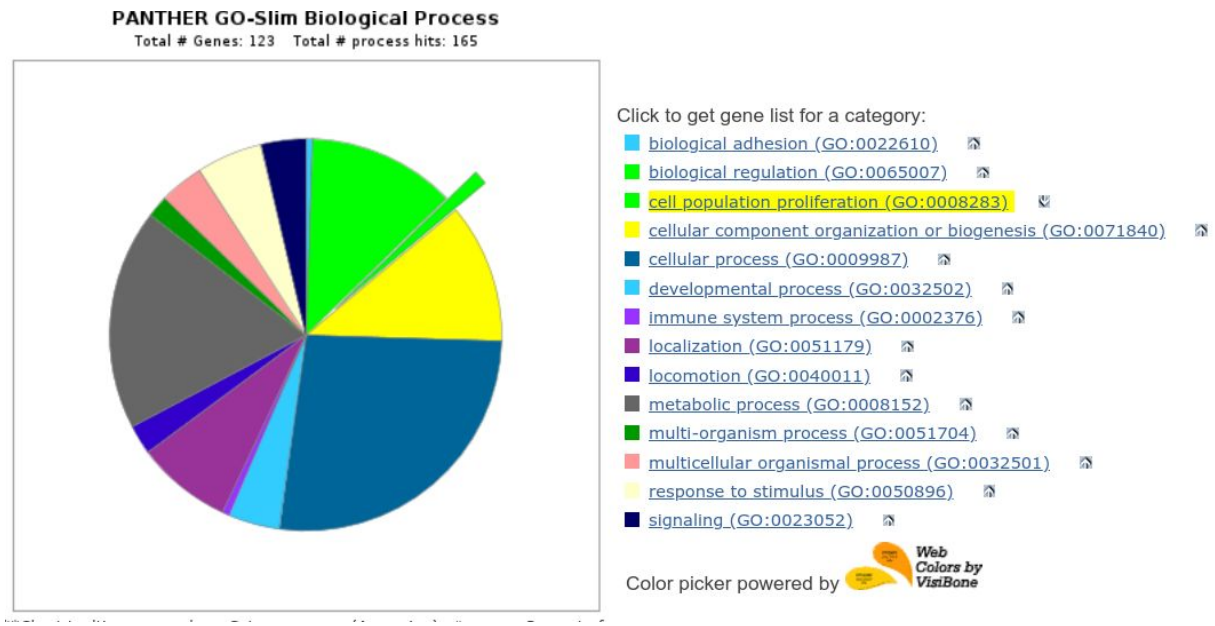


Figura 22 - Visualización gráfica de los términos GO relacionados con los procesos biológicos (A) presentes entre los top 170 genes sub-expresados en la línea celular KOPT-K1. Se señala el término GO "cell population proliferation process".

En el caso de los genes sobreexpresados observamos que los términos GO de los procesos biológicos son muy similares a los presentes en los genes sub-expresados, encontrando de nuevo el término de la proliferación celular.

En cuanto a las rutas biológicas las rutas biológicas con mayor número de genes sobreexpresados en presencia de SAMH1 corresponde a la Toll receptor signaling pathway (P00054), Huntington disease (P00029) y Apoptosis signaling pathway (P00006). Es interesante encontrar la ruta de muerte celular programada (apóptosis) entre estas rutas, puesto que el efecto de SAMH1 podría estar mediado a través de esta vía. También es interesante señalar que entre las rutas a las que pertenecen los genes sobreexpresados no se encuentra la ruta de señalización de NOTCH.

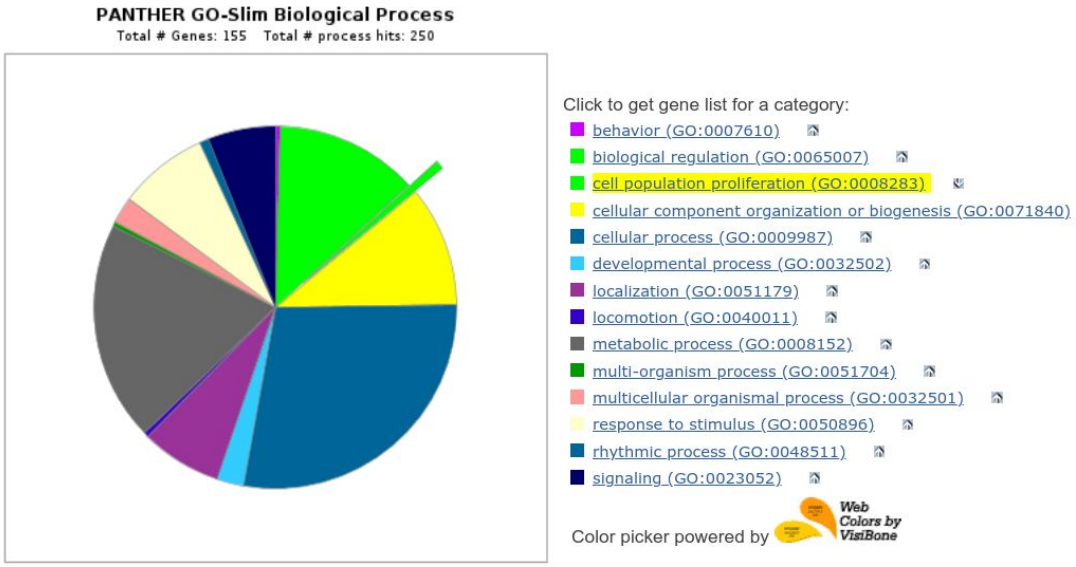
Análisis Funcional de los genes diferencialmente expresados en la línea celular HPB-ALL

En el caso de la línea celular HPB-ALL se disponen de 495 genes sobreexpresados con un logFC superior a 0.5 y un p-valor ajustado menor a 0.05; y de 297 genes sub-expresados con un logFC inferior a -0.5 y un p-valor ajustado menor a 0.05.

En el caso de la anotación GO relacionada con los procesos biológicos del top 170 genes sub-expresados observamos resultados muy similares a la línea KOPT-K1, encontrándose de nuevo el término relacionado con el control de la proliferación celular (Figura 23A). Igualmente hay que destacar la presencia de la ruta de señalización de NOTCH entre las

rutas biológicas asociadas a los genes sub-regulados (Figura 23B).

A



B

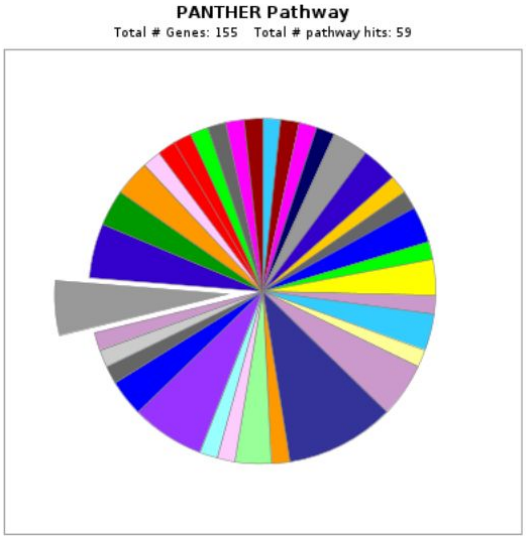


Figura 23 - Visualización gráfica de los términos GO relacionados con los procesos biológicos (A) presentes entre los top 170 genes sobreexpresados en la línea celular HPB-ALL. En (A) se señala el término GO “cell population proliferation process”. En (B) se destaca la ruta NOTCH signaling pathway (P00045).

En el caso de los genes sobreexpresados, los términos GO relacionados con los procesos biológicos volvemos a observar la similitud con lo términos de los genes sub-expresados (Figura 24). Entre las rutas biológicas la más abundantes son la ruta de liberación de gonadotropina (P0664), respuesta a estrés oxidativo (P00046) y la ruta de señalización de

p53 (P00059). También es interesante señalar que la ruta de señalización de NOTCH se encuentra entre las rutas asociadas a los genes sobreexpresados, en concreto, por la sobreexpresión del gen *NCOR2*. La función de *NCOR2* es regular negativamente la unión a DNA mediante la condensación de cromatina y así controlar la proliferación de los linfocitos. De este modo, un posible mecanismo de acción de SAMH1 podría explicarse mediante la sobreexpresión de *NCOR2*. El efecto producido por SAMH1 a la hora de reducir los niveles de expresión de genes diana de NOTCH1 podría explicarse por la sobreexpresión de *NCOR2*, el cual provocaría la condensación de la cromatina y la no expresión de estos genes diana. Esto es tan sólo una hipótesis, dado que es muy poco probable que el efecto de SAMH1 se deba sólo a un gen.

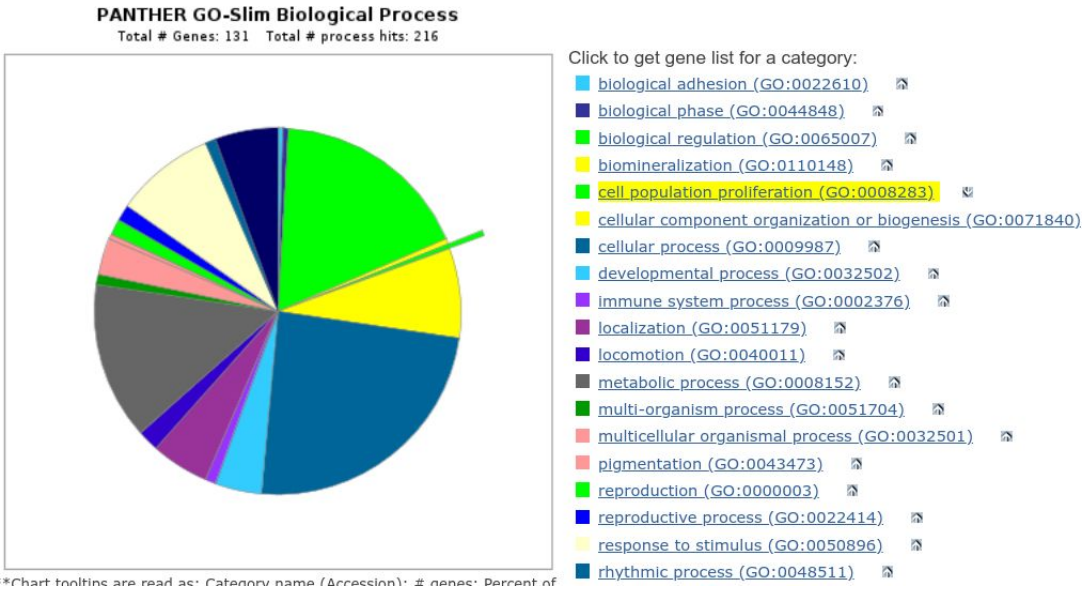


Figura 24 - Visualización gráfica de los términos GO relacionados con los procesos biológicos (A) presentes entre los top 170 genes sobreexpresados en la línea celular HPB-ALL. Se señala el término GO “cell population proliferation process”.

También se han empleado otras herramientas de análisis funcional, como SNOW [8] y FatiGO [9] cuyos resultados se recogen en el Apéndice A.

GENE SET ANALYSIS (GSEA)

El análisis de set de genes (“Gene Set Analysis”, GSEA) es otra aproximación útil para el análisis de datos de expresión génica. Esta aproximación es especialmente potente en los casos donde el número de genes analizados es elevado (alta dimensionalidad), ya que permite resumir el gran número de genes en sets de genes biológicamente relevantes.

La potencia de este tipo de análisis reside precisamente en que los genes no son analizados de forma individual sino que el análisis se lleva a cabo teniendo en cuenta la participación conjunta de los mismos. Esto facilita la interpretación de los resultados y la captación de interacciones entre genes que no se detectan mediante un análisis de expresión diferencial.

En el artículo de referencia se emplea la herramienta GSEA del Broad Institute junto a Universidad de San Diego [2]. El funcionamiento de GSEA se basa en el empleo de una lista de genes (L) que es ordenada mediante alguna métrica (por ejemplo el log FC). Sobre la lista de genes de interés (L) se calcula si un set de genes curado (S) se encuentra significativamente asociado a los genes de la lista (L) rankeados en la parte superior o inferior [16]. Algo importante a tener en cuenta es que GSEA ordena la lista de genes en función de la capacidad que tiene cada uno de los genes en separar las dos condiciones experimentales independientemente de la métrica empleada. Para conocer si un set de genes se encuentra asociado a genes en la parte inferior o superior de la lista calcula un score de enriquecimiento normalizado por FDR (Normalized Enrichment Score, NES). Por ejemplo, en caso de que NES sea positivo indica que el set de genes curado se asocia a genes ordenados en la parte superior de la lista. Normalmente, en la parte superior de la lista se sitúan los genes sobreexpresados y en la parte inferior los genes sub-expresados.

Para reproducir los datos emplearemos como set de genes curados los correspondientes a la colección C3 TFT compuesta por 1137 set relacionados con las dianas de los factores de transcripción procedente de la base de datos MSigDB. Además, los autores incluyeron el set de genes denominado GSI-NOTCH asociado al efecto del fármaco GSI en la regulación de NOTCH. Todos estos set de genes se recopilaron en un mismo archivo *.gmt.

Igualmente, para generar el archivo *.gct se emplearon los valores de intensidades de los diferentes microarrays normalizadas por RMA y la media de los controles DMSO. En este caso el análisis GSEA se lleva a cabo de forma conjunta en ambas líneas celulares, dado que el análisis GSEA no es un análisis comparativo individual y lo importante es que los valores que se emplean para ordenar los genes sean homogéneos, es decir, que la variabilidad en todos ellos sea equivalente de modo que la ordenación no se encuentre sesgada. Además, la normalización por la media de los controles de DMSO permite controlar la variabilidad inter-array. Para el procedimiento de análisis también es interesante señalar que el archivo *.chip es generado en R a partir de los spots del objeto AffyBatch() y la notación extraída mediante AnnotationDbi.

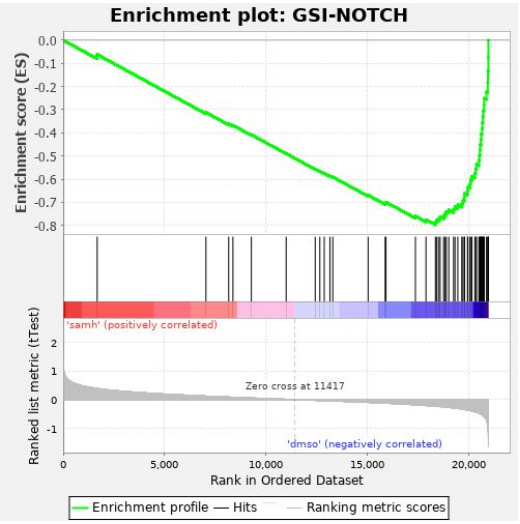
Para estudiar el enriquecimiento de las muestras tratadas con SAMH1 en los sets de

genes de interés se emplean los siguientes parámetros (reproduciendo los pasos del artículo referencia): collapse = True, phenotype = SAMH1 vs DMSO, permutations = 1000, gene set size: 15 < n < 500. El tipo de permutación empleado es “gene_set” dado que disponemos de menos de 7 muestras por condición.

Se emplea la métrica t-test para generar la lista de genes ordenados y el estadístico de enriquecimiento “weighted”. El estadístico de enriquecimiento es clave para determinar si un set de genes se encuentra enriquecido en tu muestra, puesto que establece la importancia que se da a la posición de los genes pertenecientes al gene set en la lista ordenada. El método “weighted” da más importancia a los genes del set de genes localizados en la parte superior o inferior de la lista ordenada. De este modo, se evita el problema de dar la misma importancia a genes localizados en el centro de la lista. Igualmente, el valor empleado para ordenar los genes es el real y no el valor absoluto.

Los resultados obtenidos se muestran en la Figura 25. En esta Figura reproducimos los resultados obtenidos por los autores. Encontramos que el set de genes GSI-Notch efectivamente es el más enriquecido con un p-valor ajustado = 0.0 y un valor de enriquecimiento normalizado (NES) = -3.27. Este valor corrobora que el efecto de SAMH1 se produce sub-expresando genes diana de Notch. Igualmente, encontramos en tercera posición el set de genes MYC / MAX con un p-valor ajustado = 4e-4 y NES = -2.217.

A



B

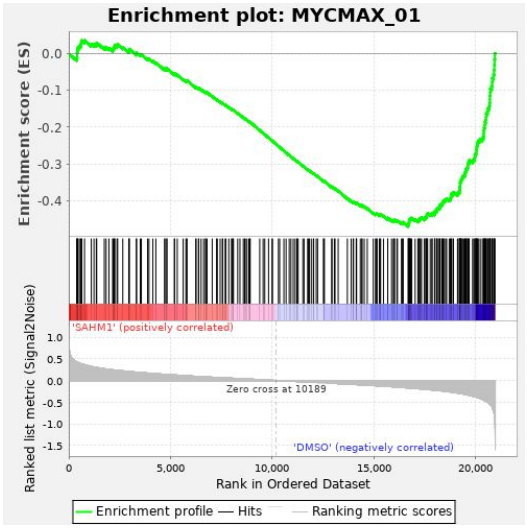


Figura 25 - Gráficos de enriquecimiento del set de genes perteneciente a GSI-Notch (A) y MYC / MAX (B), primer y tercer gene set más enriquecidos en el análisis de GSEA de las líneas celulares KOPT-K1 y HPB-ALL.

Reproduciendo los resultados del artículo de referencia podemos decir que los genes que

son sub-expresados en presencia del fármaco GSI, también aparecen sub-expresados en presencia del péptido sintético SAMH1 obtenido por los autores.

Adicionalmente, se llevó a cabo el mismo análisis pero añadiendo los siguientes gene sets:

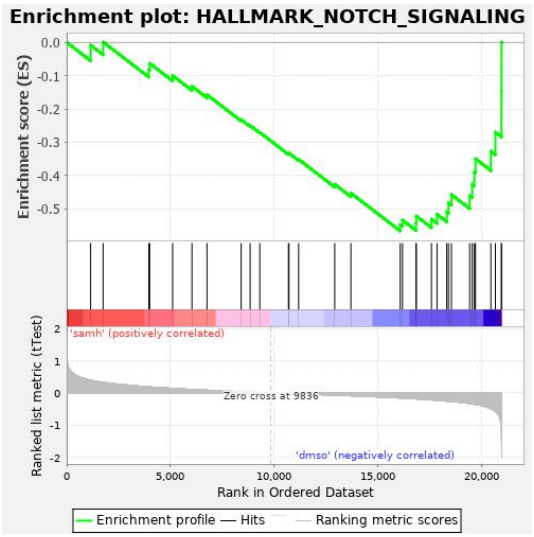
- GO_NOTCH_SIGNALING_PATHWAY perteneciente a la colección C5
- GO_NEGATIVE_REGULATION_OF_NOTCH_SIGNALING_PATHWAY perteneciente a la colección C5
- HALLMARK_NOTCH_SIGNALING
- REACTOME_SIGNALING_BY_NOTCH1 perteneciente a la colección C2
- REACTOME_SIGNALING_BY_NOTCH2 perteneciente a la colección C2
- REACTOME_SIGNALING_BY_NOTCH3 perteneciente a la colección C2
- REACTOME_SIGNALING_BY_NOTCH4 perteneciente a la colección C2
- REACTOME_SIGNALING_BY_NOTCH perteneciente a la colección C2

De nuevo es el set de genes GSI-Notch el que aparece en primera posición, seguido en tercer lugar de MYC / MAX. De los set de genes incorporados nuevamente sólo “HALLMARK NOTCH SIGNALING” se encuentra enriquecido en la muestra con un p-valor corregido inferior al 0.05 (FDR q-value = 0.041, NES = -1.94) (Figura 26 A).

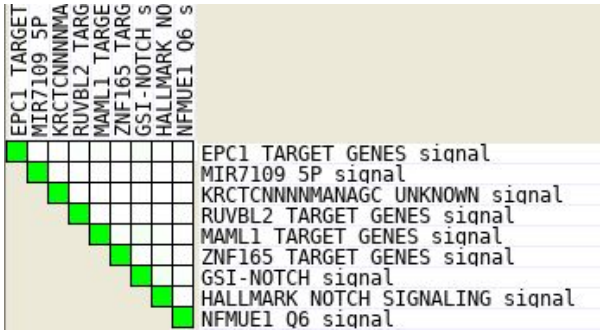
Para analizar los resultados se emplea la herramienta “Leading edge analysis” que nos permite conocer el solapamiento de genes entre set de genes seleccionados. En nuestro caso estudiamos el solapamiento entre los set de genes con un valor de FDR q-value inferior a 0.05. En la Figura 26 B podemos observar como los principales set de genes no se encuentran solapados en gran medida. En la Figura 26 C podemos ver que precisamente son los genes HES1, HES4 y DTX1 los que se encuentran compartidos por más set de genes.

Para visualizar este solapamiento se emplea la herramienta “Enrichment Map Visualization” que permite representar en un grafo los diferentes set de genes (nodos) conectados entre sí (aristas) en función del grado de solapamiento (overlapping). En la Figura 27 podemos comprobar cómo empleando un filtro de FDR q-value = 0.05 y un solapamiento = 0.01 observamos cómo el set GSI-NOTCH se encuentra conectado con el set de genes “MAML1-TARGET-GENES”. Este resultado es lógico dado que SAMH1, el péptido sintético generado por los autores, se obtiene a partir del molde de un fragmento de la proteína MAML1 con capacidad de antagonizar la señalización mediada por NOTCH.

A



B



C

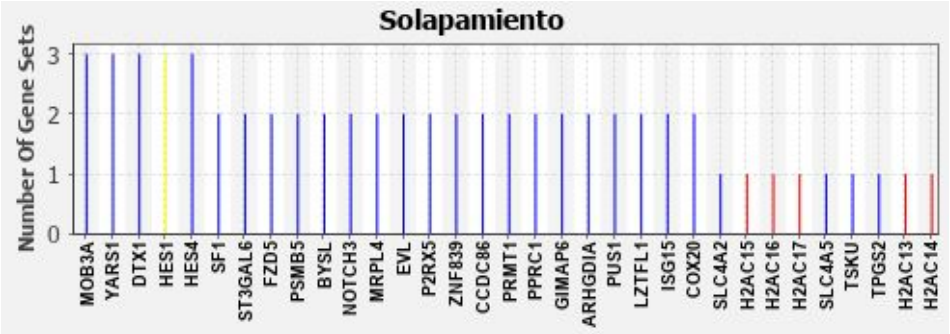


Figura 26 - Gráficos de enriquecimiento del set de genes perteneciente a HALLMARK NOTCH SIGNALING (A), matriz de solapamiento entre los gene sets con un p-valor ajustado menor a 0.05 (B) y genes con mayor solapamiento entre los gene sets con un p-valor ajustado menor a 0.05 (C) en el análisis de GSEA de las líneas celulares KOPT-K1 y HPB-ALL ampliado.

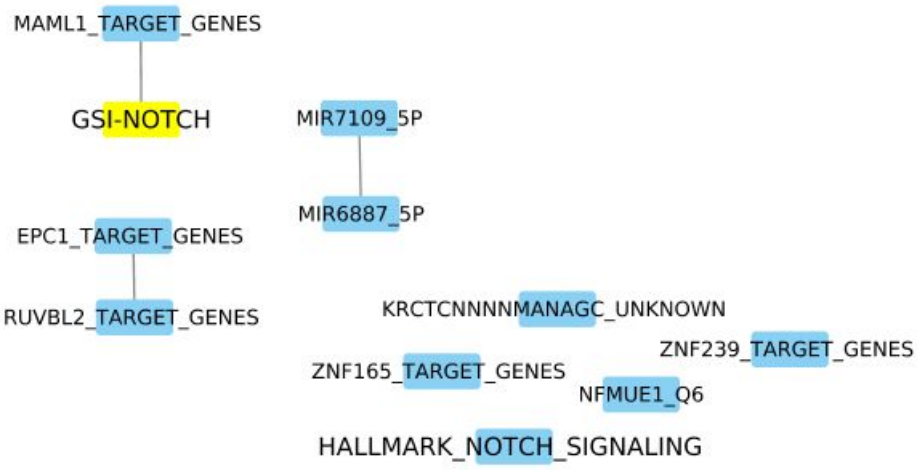


Figura 27 - Visualización del grado de enriquecimiento procedente de los gene sets con un p-valor ajustado FDR menor a 0.05 del análisis de GSEA en las líneas celulares KOPT-K1 y HPB-ALL tratadas con SAMH1. Se emplea un umbral de solapamiento = 0.01. En amarillo se señala el set de genes GSI-Notch.

CONCLUSIONES

1. Hemos comprobado que el efecto de SAMH1 debe ser estudiado por separado en cada línea celular.

2. Hemos validado el efecto de SAMH1 a la hora de provocar la sub-expresión de los genes identificados por los autores del artículo tanto en la línea celular KOPT-K1 como en HPB-ALL. De este modo, el efecto de SAMH1 es independiente de la línea celular.

3. El análisis de clusterización ha permitido validar el hecho de que entre las líneas celulares KOPT-K1 y HPB-ALL existen suficientes diferencias como para ser clasificadas en grupos diferentes, mediante el empleo tanto de distancias absolutas (euclídea) como de tendencias (coeficiente de Pearson).

4. Se ha generado un clasificador supervisado basado en el algoritmo de SVM capaz de clasificar muestras en función del tratamiento DMSO o SAMH1 con un 98% de precisión empleando el gen PXK.

5. El análisis funcional muestra que la ruta de señalización de NOTCH1 se encuentra entre las rutas asociadas a los genes sub-regulados en presencia de SAMH1 tanto en la línea celular KOPT-K1 como en la línea celular HPB-ALL.

6. Hemos reproducido los resultados en GSEA del artículo de referencia encontrando el set de genes GSI-Notch como el principal gene set enriquecido y el set de genes MYC / MAX en tercera posición. Ambos sub-expresados en la condición SAMH1 cuando son comparados con respecto a la situación control.

7. La visualización del grafo de los set de genes enriquecidos en GSEA ha permitido demostrar que GSI-Notch presenta genes solapados con el set de genes “MAML1-TARGET-GENES”, lo que corrobora el efecto de SAMH1 a la hora de antagonizar la señalización mediada por NOTCH.

REFERENCIAS

1. Moellering, R., Cornejo, M., Davis, T., Bianco, C., Aster, J., & Blacklow, S. et al. (2009). Direct inhibition of the NOTCH transcription factor complex. Nature, 462(7270), 182-188. doi: 10.1038/nature08543

2. Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., & Gillette, M. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings Of The National Academy Of Sciences*, *102*(43), 15545-15550. doi: 10.1073/pnas.0506580102

3. Morpheus. (2020). Retrieved 3 June 2020, from https://software.broadinstitute.org/morpheus/

4. Medina I, Montaner D, Tárraga J, Dopazo J. (2007) Prophet, a web-based tool for class prediction using microarray data. [Bioinformatics] (http://www.ncbi.nlm.nih.gov/pubmed/17138587) 23(3):390-1

5. Alonso R, Salavert F, Garcia-Garcia F, Carbonell-Caballero J, Bleda M, Garcia-Alonso L, Sanchis-Juan A, Perez-Gil D, Marin-Garcia P, Sanchez R, Cubuk C, Hidalgo MR, Amadoz A, Hernansaiz-Ballesteros RD, Alemán A, Tarraga J, Montaner D, Medina I, Dopazo J. (2015) Babelomics 5.0: functional interpretation for new generations of genomic data. [Nucleic Acids Res.] (http://www.ncbi.nlm.nih.gov/pubmed/25897133) (2015 Apr 20. pii: gkv384.)

6. Mi, H., Muruganujan, A., Huang, X., Ebert, D., Mills, C., Guo, X., & Thomas, P. (2019). Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nature Protocols*, *14*(3), 703-721. doi: 10.1038/s41596-019-0128-8

7. Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. (2018). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, *47*(D1), D419-D426. doi: 10.1093/nar/gky1038

8. Minguez P, Götz S, Montaner D, Al-Shahrour F, Dopazo J.(2009) SNOW, a web-based tool for the statistical analysis of protein-protein interaction networks. [Nucleic Acids Res.] (http://www.ncbi.nlm.nih.gov/pubmed/19454602) 37(Web Server issue):W109-14

9. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. [Bioinformatics] (http://www.ncbi.nlm.nih.gov/pubmed/14990455) 20(4):578-80.

10. Cañizares, J. (2010). VI BioMur módulo 3 — BioMur modulo3 v1.1 documentation. Retrieved 3 June 2020, from http://personales.upv.es/jcanizar/modulo_3/index.html

11. Binder, H., & Preibisch, S. (2005). Specific and Nonspecific Hybridization of Oligonucleotide Probes on Microarrays. *Biophysical Journal*, *89*(1), 337-352. doi: 10.1529/biophysj.104.055343

12. Deshmukh, S., & Purohit, S. (2011). *Microarray data*. Oxford: Alpha Science.

13. Amat Rodrigo, J. (2017). RPubs - Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. Retrieved 3 June 2020, from https://rpubs.com/Joaquin_AR/287787

14. Rousseeuw, P. and Verboven, S. (2002). Robust estimation in very small samples. Computational Statistics & Data Analysis, 40(4), pp.741-758.

15. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic Acids Research*, **43**(7), e47. doi: 10.1093/nar/gkv007.

16. Ringnér, M., Peterson, C., & Khan, J. (2002). Analyzing array data using supervised methods. *Pharmacogenomics*, *3*(3), 403-415. doi: 10.1517/14622416.3.3.403

17. GSEA User Guide. (2020). Retrieved 3 June 2020, from https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html

APÉNDICE A

En este Apéndice se incorporan figuras adicionales correspondientes a la exploración de nuevas herramientas de análisis vistas durante el curso.

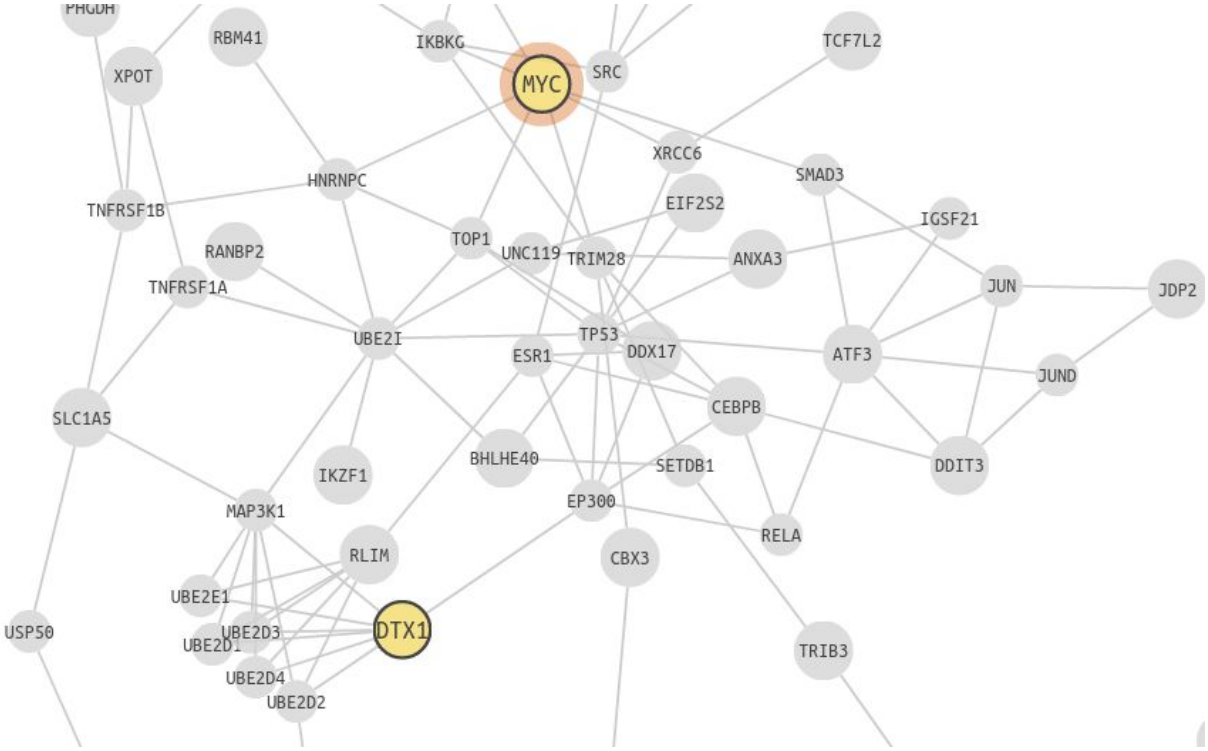


Figura A 1 - Visualización de los genes *MYC*, *DTX1* en el grafo de interacción correspondiente al top 170 genes sub-expresados en la línea celular KOPT-K1. Grafo creado con la herramienta SNOW [8] de Babelomics 5.0.

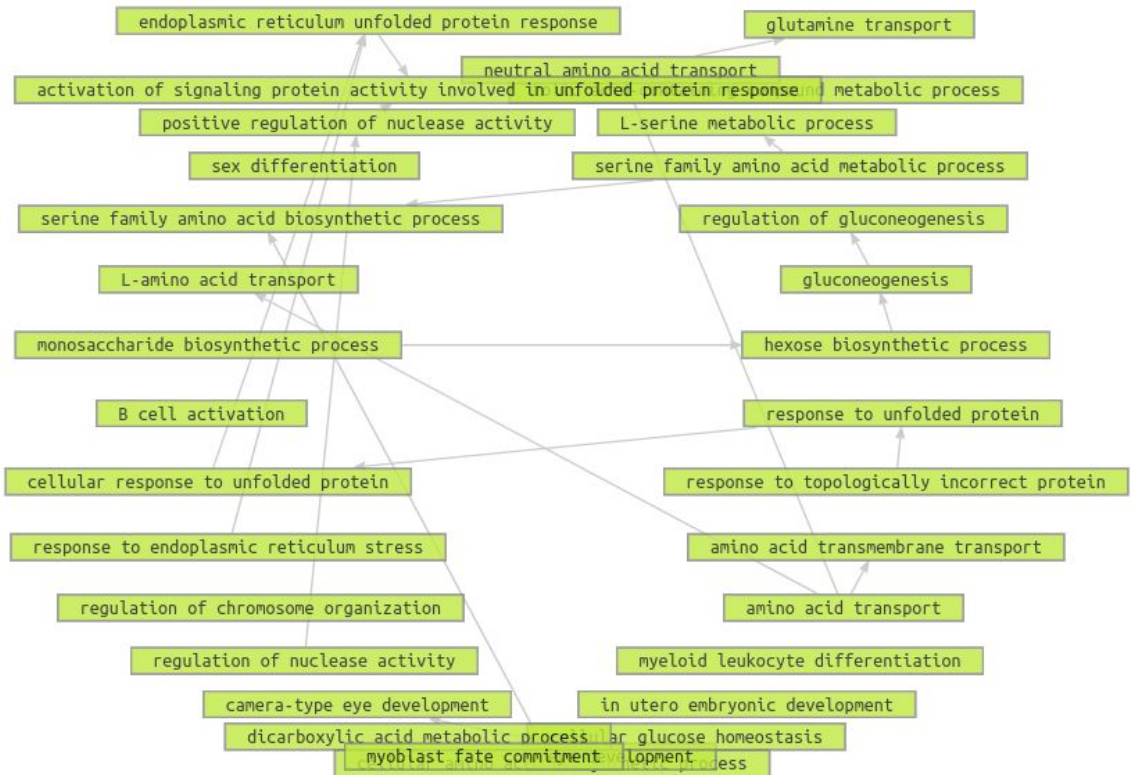


Figura A 2 - Visualización de los términos GO de función molecular enriquecidos en los top 170 genes sub-expresados en la línea celular KOPT-K1 con respecto al resto del genoma. Grafo creado con la herramienta FatiGO [9] de Babelomics 5.0.