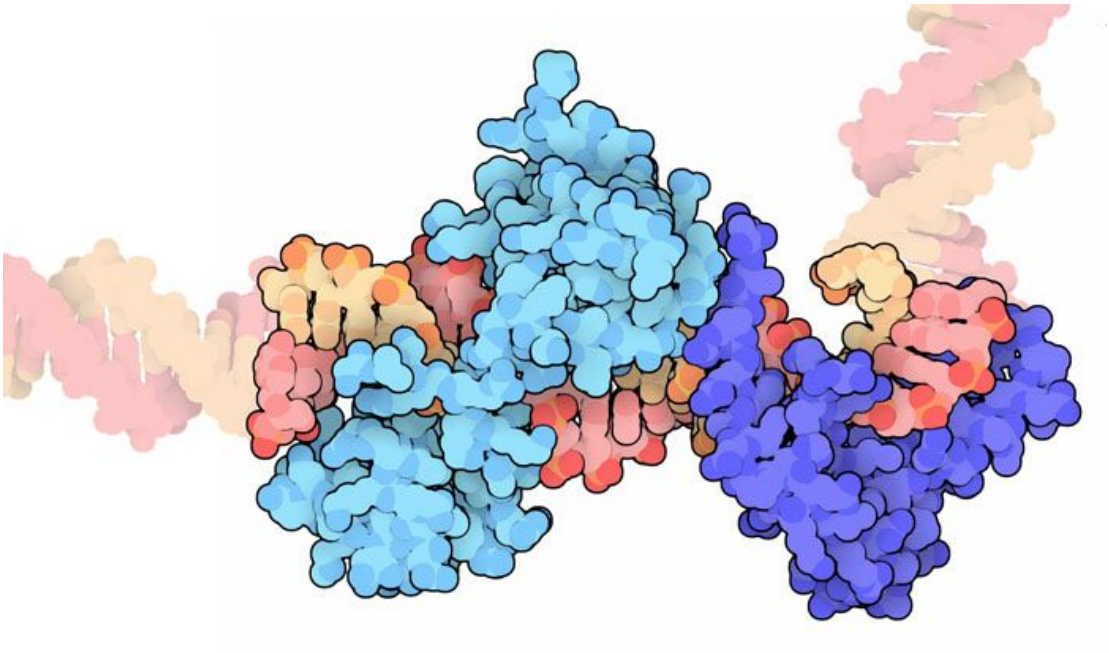
Análisis de datos de arrays

Direct inhibition of the NOTCH transcription factor complex.



Sara Dorado Alfaro

INTRODUCCIÓN

Los factores de transcripción tienen un papel muy importante en la regulación del estado celular. Por ejemplo, son los responsable de mantener la especificación de tejido en cáncer. Son elementos muy interesantes para el descubrimiento de ligandos, pero químicamente difíciles de tratar. Por ello, hay muy pocos inhibidores específicos para factores de transcripción humanos. En este trabajo se demuestra la inhibición directa del complejo del factor de transcripción NOTCH por medio del péptido SAHM1 (stapled α-helical MAML1) [1].

Las proteínas NOTCH participan directamente en las rutas de señalización de la diferenciación, la proliferación y la muerte celular. Su unión resulta en el complejo ICN-CSL-MAML, que recluta la maquinaria de transcripción de los genes dependientes de NOTCH (genes diana de NOTCH). Normalmente la duración y la fuerza de la señalización NOTCH está fuertemente controlada. Se ha demostrado que mutaciones de pérdida de función que afectan a las rutas de NOTCH producen diversas enfermedades, mientras que mutaciones de ganancia de función se relacionan causalmente con el cáncer. En particular, más del 50% de pacientes con leucemia linfoblástica aguda en células T (T-ALL) presentan mutaciones activadoras de NOTCH. También se han encontrado este tipo de mutaciones en otros tipos de cáncer. Es por ello que la inhibición de NOTCH vía SAHM1 podría ser un tratamiento para pacientes con estas patologías.

Péptidos dirigidos al complejo NOTCH

El diseño del péptido SAHM1 es posible gracias al descubrimiento de un fragmento negativo de MAML1, denominado dnMAML1. Si bien es sabido que MAML1 activa NOTCH, el fragmento negativo presenta una función antagónica en la señalización de NOTCH y la proliferación celular en líneas T-ALL. El polipéptido dnMAML1 tiene la forma de una α-hélice casi continua, que se engancha en una ranura alargada formada por el complejo ICN-CSL.

El objetivo es evitar la unión del complejo de activación transcripcional a pesar de la presencia de una señal NOTCH aguas arriba. Gracias a la sabida estructura de dnMAML1, se crean distintos péptido sintéticos de carácter más helicoidal que MAML1, estableciendo un escenario competitivo para la unión de NOTCH, pero sin afectar a otras moléculas. En total, se diseñan 6 péptidos capaces de cubrir la superficie de contacto de MAML1 e ICN-CSL. De entre los péptidos testados, SAHM1 es el mejor en la evaluación

funcional, inhibiendo directamente el reclutamiento del complejo ICN-CSL.

Objetivos

Se comprueba la inhibición génica de las dianas de NOTCH en células T-ALL humanas mediante el uso del péptido SAHM1. Para ello, se realizarán varios análisis de enriquecimiento, comparando los niveles de expresión génica tras aplicar el tratamiento con SAHM1 y los controles, que únicamente han recibido el tratamiento con DMSO (dimethylsulphoxide). Se buscarán, por tanto, cambios globales en la expresión génica entre los casos (tratados con SAHM1) y los controles.

Además, se pretende comparar el rendimiento del tratamiento con SAHM1 contra el del tratamiento tradicional con GSI, que presenta varios inconvenientes:

- Debe establecerse una dosis límite por riesgo de intoxicación gastrointestinal.
- Algunas mutaciones activadoras de NOTCH1 son resistentes al GSI.

ANÁLISIS REALIZADOS

Se miden cambios globales en la expresión génica y cambios específicos de expresión en los genes diana de NOTCH tras el tratamiento con SAHM1 en dos líneas celulares de T-ALL: HPB-ALL y KOPT-K1. Los datos disponibles son un conjunto de 12 ficheros .CEL de cultivos triplicados de células KOPT-K1 y HBP-ALL que fueron tratadas únicamente con DMSO, que se utiliza como grupo control, o con DMSO y SAHM1. Pueden descargarse de la página Gene Expression Omnibus [2] asociados a la serie GSE18198.

El ARN extraído de las células T-ALL se analiza con arrays de expresión de Affymetrix U133 Plus 2.0. Recordamos que en cada fichero .CEL están los datos de intensidad de arrays relativos a cada experimento.

Además, en el material suplementario se proporciona un conjunto de transcritos subregulados por GSI en células T-ALL. Nos referiremos a este conjunto de genes como GSI-NOTCH, ya que son los genes inhibidos por el tratamiento tradicional con GSI en pacientes de NOTCH.

METODOLOGÍA Y ANÁLISIS BIOINFORMÁTICO

En esta sección se especifican los métodos bioinformáticos utilizados para el análisis de los datos, así como la extracción de algunas conclusiones. Se explica el procedimiento de carga de datos, normalización y preprocesamiento, análisis de expresión diferencial y análisis de expresión diferencial de conjuntos de genes.

Carga de datos

Al tratarse de un array de Affymetrix, utilizamos el paquete affy [4]. Primero debemos leer el fichero targets.txt, que contiene la información relativa a cada fichero .CEL. Este fichero se crea a partir de la información de la serie de GEO (ver Tabla 1) y se lee con la función readTargets() del paquete limma.

Después se cargan los ficheros .CEL por medio de la función `ReadAffy(file)`. Se analizarán por separado los ficheros correspondientes a cada línea celular.

Fichero	Línea Celular	Class
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

Tabla 1: Fichero targets.txt

Preprocesamiento

Debemos efectuar la corrección de fondo y la normalización de la intensidad de los ficheros de arrays con intensidades crudas, obteniendo niveles de expresión comparables para cada muestra. En la Figura 1 se muestran los histogramas que evidencian las diferencias en las intensidades entre las distintas muestras.

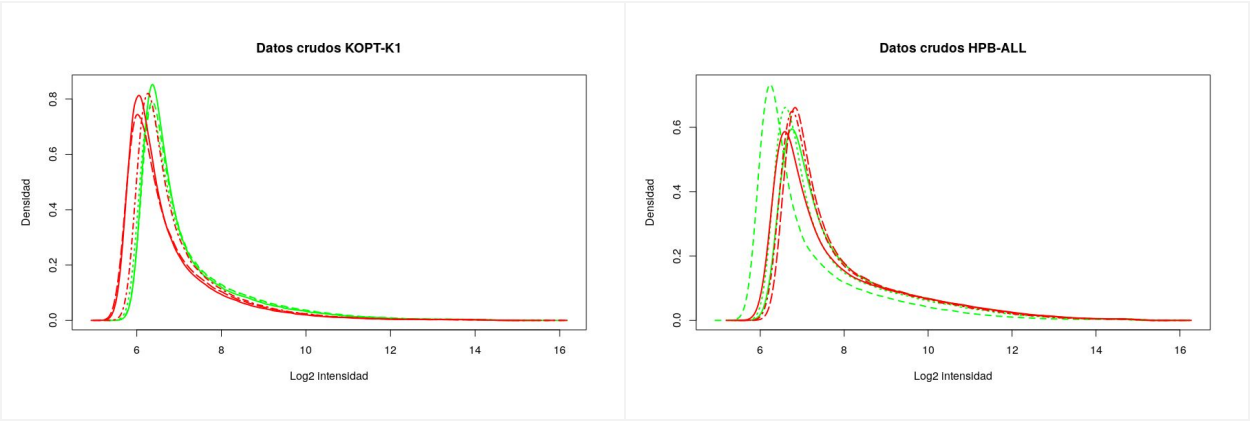


Figura 1: histogramas de los datos de microarrays para cada una de las líneas celulares. Rojo son muestras tratadas con DMSO y en verde aparecen las tratadas con DMSO + SAMH1.

Se aplica la normalización estándar para arrays de Affymetrix utilizando la función `expresso()` de la librería `affy`. Siguiendo el consenso, aplicamos el método de corrección de fondo entre arrays RMA (Robust Multi-array Average) y normalización por cuantiles. En la Figura 2 se muestran los boxplot de los resultados de la normalización en cada una de las líneas celulares. Puede observarse que la normalización ha igualado las medias y medianas entre ficheros.

Además de la normalización, es importante eliminar los genes que tienen un comportamiento plano en el experimento. Para ello filtramos los datos usando el rango intercuartílico con la función `varFilter`. Esto sirve para eliminar genes que tienen poca varianza entre experimentos o que, en general, dan poca señal, facilitando el posterior análisis de expresión diferencial. Estableciendo un `cutoff` de 0.5 reducimos el tamaño del conjunto de datos de 54675 observaciones a 27337. La reducción es casi del 50%.

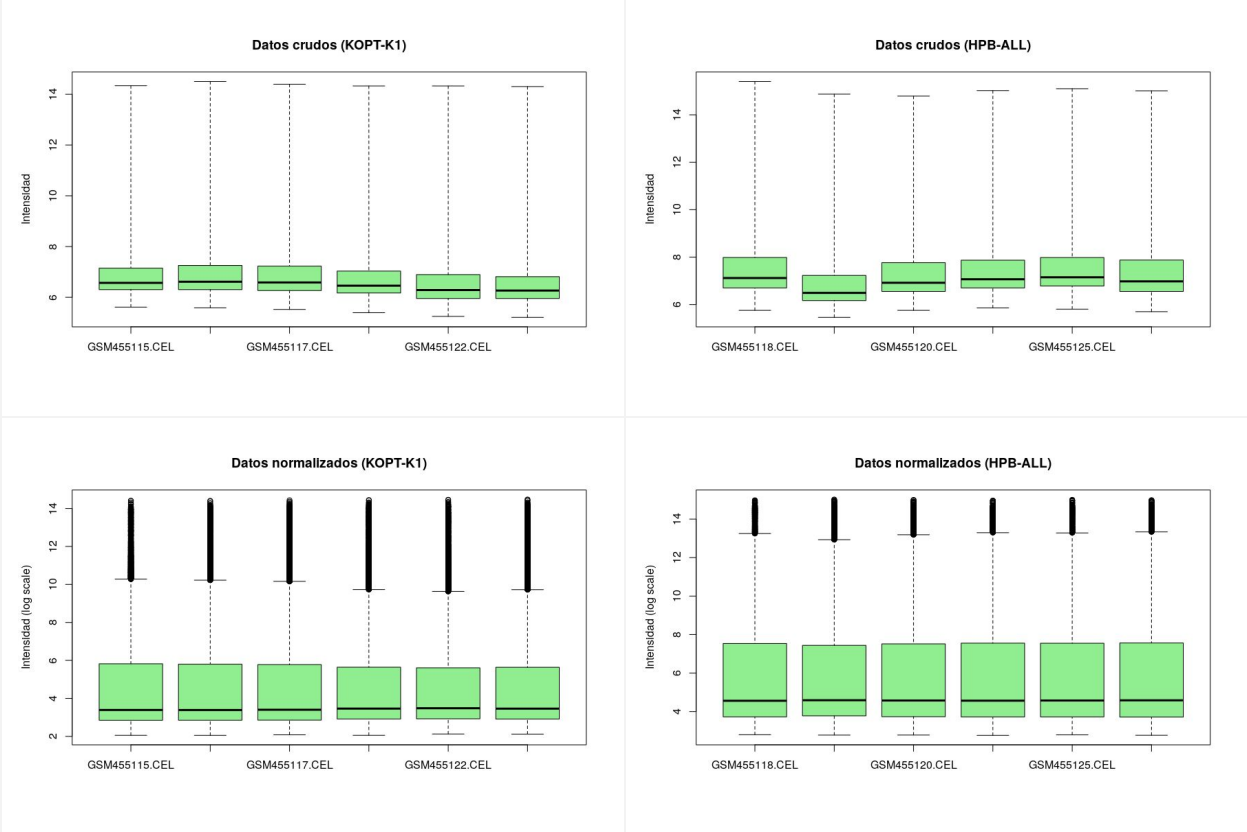


Figura 2: datos crudos (arriba) y normalizados (abajo) para la línea celular KOPT-K1 (izquierda) y HPB-ALL (derecha).

Análisis de expresión diferencial

En esta sección se muestran los resultados del análisis de expresión diferencial de las dos líneas celulares, KOPT-K1 y HPB-ALL. Así, se obtiene una lista con los genes diferencialmente expresados (GDE) en las muestras sometidas tratamiento con SDMSO+SAHM1 y aquellas tratadas únicamente con DMSO. Para ello, se ajusta un modelo lineal con el paquete `limma` [3] siguiendo los pasos:

1. Creación de las matrices de diseño. Se generan a mano, e indica a qué muestra pertenece cada fichero. En nuestro caso distingue las muestras control DMSO y los casos DMSO+SAHM1.
2. Matrices de contraste. Prepara los contrastes estadísticos, asignando qué muestras pertenecen a cada contraste. En nuestro caso, queremos SAHM1vsDMSO=SAHM1-DMSO.
3. Obtención de los genes diferencialmente expresados. Se ajusta un modelo lineal

(`lmFit`) y se calculan los estadísticos de la expresión diferencial (p-valor, estadístidos t y F y foldChange) con Bayes empírico (`eBayes`).

Siguiendo el tratamiento del trabajo original, se filtra el conjunto de genes obtenidos con el p-valor ajustado. El valor de corte elegido es $\rho < 0.001$, que es el valor de filtrado del artículo original. Esto nos permite centrarnos en genes significativos en el análisis. La cuantificación de la expresión diferencial puede verse en la columna logFC.

Finalmente, la anotación se hace con las librerías `annotate` y `hgu133plus2.db` (específico para el genoma humano) de R y el método `getSymbol`.

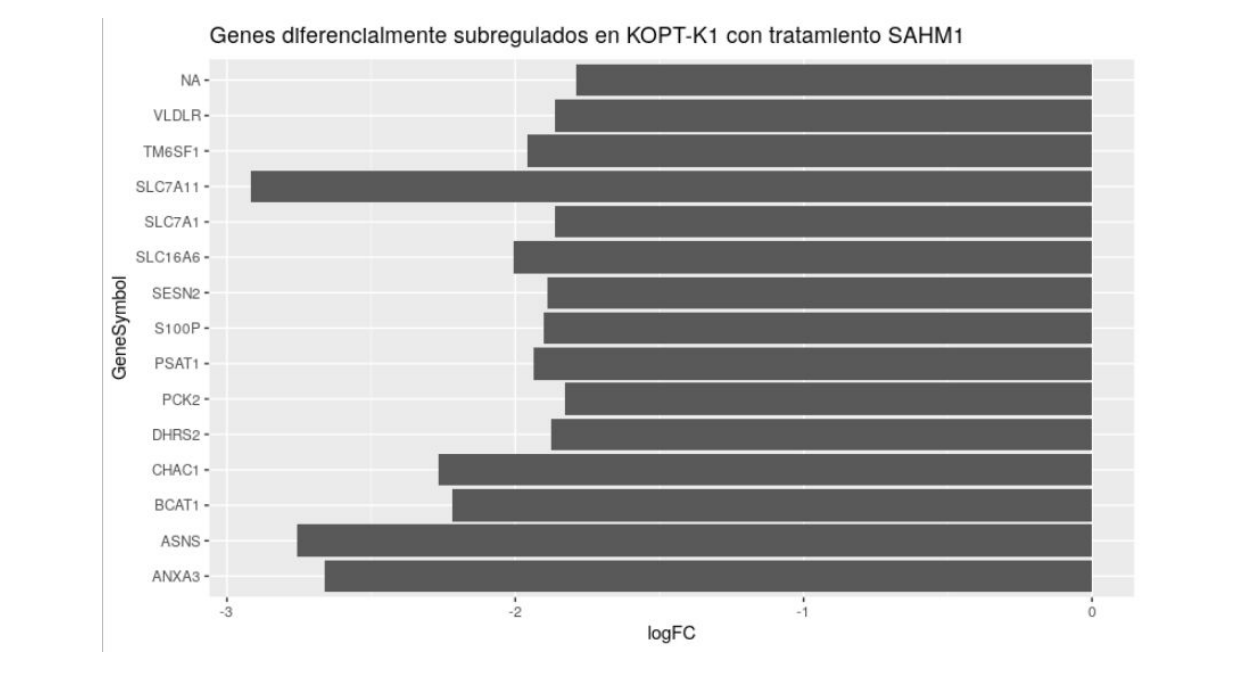


Figura 3: Genes subregulados por el tratamiento con SAHM1 en KOPT-K1.

En las Figuras 3 y 4 se muestran los 15 genes diferencialmente más subregulados para ambas líneas celulares. Es decir, aquellos con menor valor de logFC. Aunque hay algunos conjuntos de genes sobrerregulados en cada una de las líneas celulares, estos deciden no mostrarse, ya que no hay ningún gen comúnmente sobrerregulado en la intersección de las expresiones diferenciales de KOPT-K1 y HPB-ALL. Por lo tanto, no se considera que el tratamiento con SAHM1 produzca efectos de sobrerregulación.

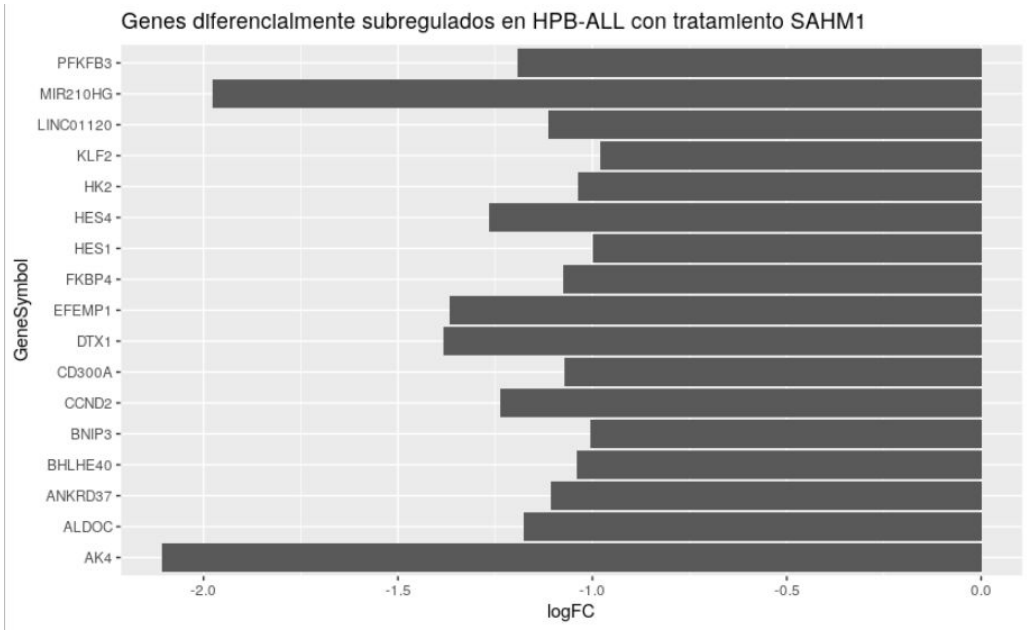


Figura 4: Genes subregulados por el tratamiento con SAHM1 en HPB-ALL.

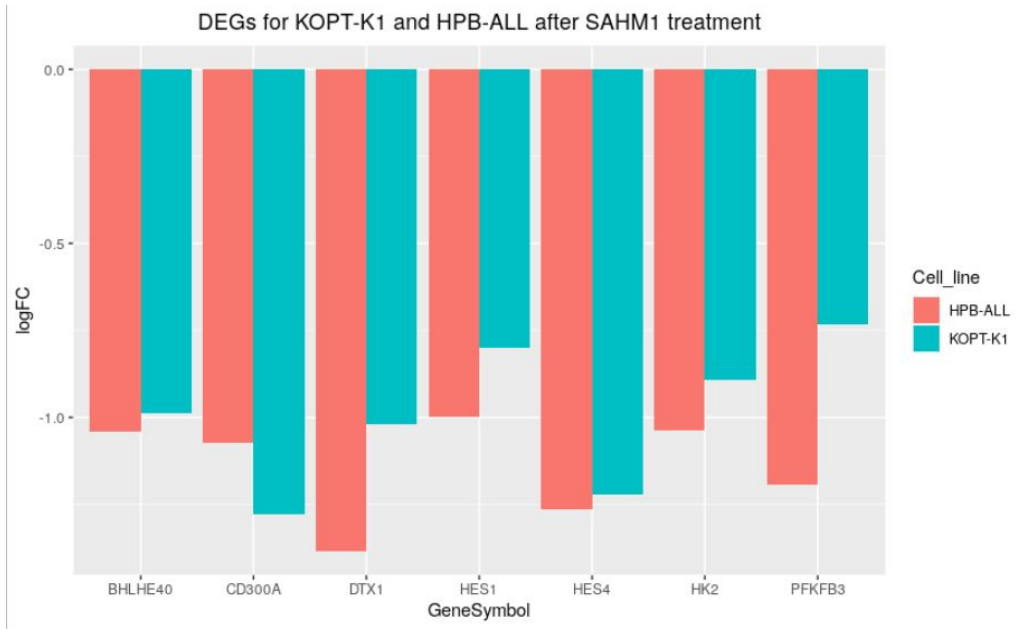


Figura 5: Intersección de genes subregulados en ambas líneas celulares tras el tratamiento con SAHM1.

Además, en la Figura 5 se muestra la figura con la intersección de los genes

diferencialmente expresados en ambas líneas celulares. Cabe destacar que todos estos genes se encuentran subregulados, ya que todos presentan un valor de logFC negativo. Estos son, probablemente, los genes destino del tratamiento con SAHM1. En la siguiente sección compararemos estos genes con los que tradicionalmente se regulan con el tratamiento tradicional con GSI. Cabe destacar que no hay intersección en genes sobrerregulados entre ambas líneas celulares.

Comparación con el conjunto de genes GSI-NOTCH

Para demostrar la validez de SAHM1 como tratamiento, en esta sección compararemos los resultados obtenidos en el análisis de expresión diferencial con el conjunto de genes diana del tratamiento GSI-NOTCH. En el material suplementario del artículo los autores proporcionan la lista de genes GSI-NOTCH subregulados por el tratamiento tradicional con GSI, que es el material que usamos para realizar esta comparación.

- En los genes sobrerregulados por el tratamiento con SAHM1 no hay ningún gen del conjunto GSI-NOTCH. Esto tiene sentido, ya que comparamos precisamente con genes conocidos que son inhibidos por el tratamiento con GSI.
- Entre los genes subregulados, encontramos 12 genes que intersecan con el conjunto GSI-NOTCH. Se muestran en la Figura 6.

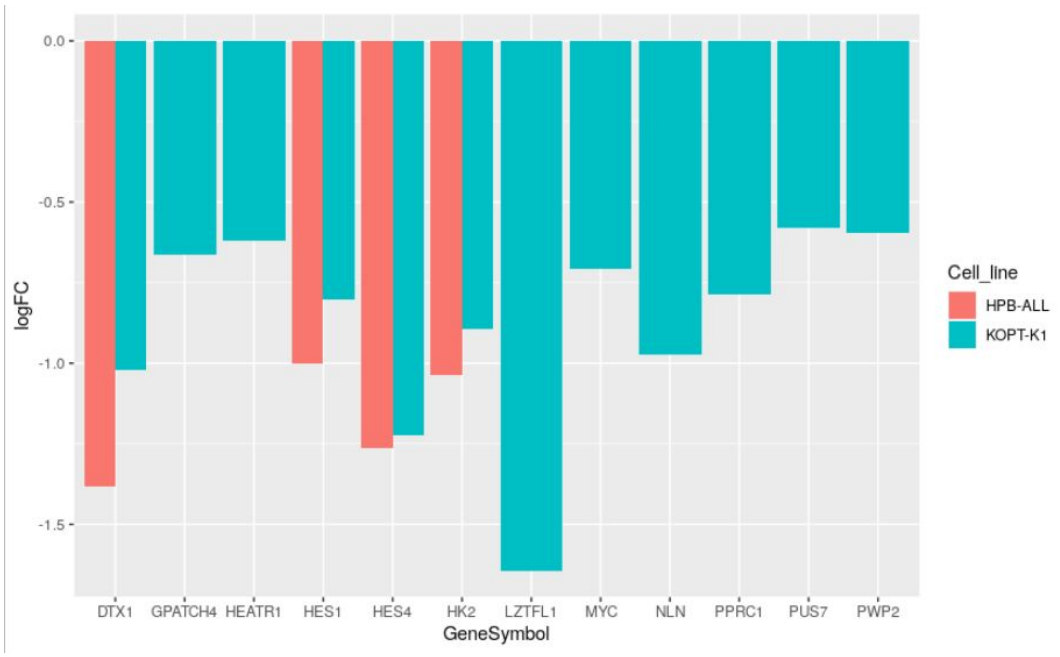


Figura 6: Intersección de los genes diferencialmente expresados en las líneas celulares KOPT-K1 (azul) y HPB-ALL (rosa) y el conjunto de genes GSI. Filtrado para $\rho < 0.001$.

Comparando con los genes comunes de la Figura 5, encontramos los genes PFKB3, CD300A y BHLHE40, que son subregulados en ambas líneas celulares para el tratamiento con SAHM1, pero no para el tratamiento con GSI. Esto debería ser revisado y en caso de confirmarse, buscar posibles efectos adversos. También puede ser la explicación de por qué en algunas ocasiones, como dicen los autores del paper, el tratamiento con GSI no tiene efectos, pero sí el tratamiento con SAHM1.

De los 12 genes mostrados en la Figura 6, sólo 4 son significativos para las líneas celulares HPB-ALL y KOPT-K1. Estos son los genes DTX1, HES1, HES4 y HK2. Este resultado coincide con el obtenido por los autores del artículo. Ellos también reportan el gen MYC, que puede encontrarse en la Figura 6, pero sólo diferencialmente subregulado en la línea celular KOPT-K1. Estas diferencias pueden deberse a los diferentes procesos de normalización y filtrado de los datos.

Tomados en conjunto, estos datos establecen que SAHM1 ejerce un efecto antagonista específico sobre la expresión génica dirigida por NOTCH.

Análisis de enriquecimiento de conjuntos de genes (GSEA)

En esta sección se realiza el estudio de enriquecimiento en conjuntos de genes con la herramienta `GSEA`. Los genes no actúan como elementos aislados, por lo que el objetivo es testear conjuntos de genes relacionados.

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	GSI-NOTCH	Details ...	56	-0.83	-3.15	0.000	0.000	0.000	2558	tags=86%, list=12%, signal=97%
2	EPC1_TARGET_GENES	Details ...	354	-0.49	-2.40	0.000	0.000	0.000	3875	tags=40%, list=18%, signal=48%
3	MYCMAX_01	Details ...	255	-0.47	-2.22	0.000	0.000	0.001	4285	tags=44%, list=20%, signal=54%
4	RUVBL2_TARGET_GENES	Details ...	35	-0.62	-2.14	0.000	0.003	0.010	1936	tags=37%, list=9%, signal=41%
5	HIF1_Q3	Details ...	231	-0.45	-2.09	0.000	0.004	0.015	3925	tags=34%, list=19%, signal=42%
6	NMYC_01	Details ...	272	-0.44	-2.08	0.000	0.004	0.020	4256	tags=37%, list=20%, signal=46%
7	ZNF165_TARGET_GENES	Details ...	75	-0.51	-2.02	0.000	0.009	0.051	3499	tags=33%, list=17%, signal=40%
8	MIR7109_5P	Details ...	75	-0.49	-1.92	0.000	0.028	0.165	5265	tags=51%, list=25%, signal=67%
9	CSHL1_TARGET_GENES	Details ...	210	-0.41	-1.89	0.000	0.038	0.243	4481	tags=34%, list=21%, signal=43%
10	MIR6813_5P	Details ...	40	-0.55	-1.89	0.000	0.035	0.245	3885	tags=40%, list=19%, signal=49%
11	KTGGYRSGAA_UNKNOWN	Details ...	74	-0.48	-1.89	0.000	0.033	0.253	4088	tags=42%, list=20%, signal=52%
12	ZNF239_TARGET_GENES	Details ...	37	-0.54	-1.88	0.000	0.034	0.283	1761	tags=22%, list=8%, signal=24%
13	MIR6887_5P	Details ...	133	-0.43	-1.87	0.000	0.033	0.297	5611	tags=49%, list=27%, signal=66%
14	MIR6085	Details ...	39	-0.56	-1.86	0.002	0.035	0.331	3885	tags=41%, list=19%, signal=50%
15	NCOA2_TARGET_GENES	Details ...	431	-0.38	-1.86	0.000	0.035	0.350	3968	tags=33%, list=19%, signal=40%
16	FOXR2_TARGET_GENES	Details ...	24	-0.59	-1.85	0.007	0.036	0.375	5103	tags=63%, list=24%, signal=83%
17	ZNF704_TARGET_GENES	Details ...	68	-0.47	-1.83	0.000	0.045	0.453	3950	tags=37%, list=19%, signal=45%
18	HIF1_Q5	Details ...	246	-0.39	-1.83	0.000	0.043	0.460	3893	tags=29%, list=19%, signal=35%
19	MAX_01	Details ...	263	-0.39	-1.83	0.000	0.042	0.467	5289	tags=39%, list=25%, signal=51%
20	MAML1_TARGET_GENES	Details ...	253	-0.39	-1.81	0.000	0.047	0.515	5383	tags=39%, list=26%, signal=51%

Figura 7: Conjuntos de genes enriquecidos en el fenotipo DMSO (6 muestras).

Para poder utilizar `GSEA`, primero debemos convertir los ficheros en formato `.CEL` de `Affymetrix` en el fichero de expresión estándar .gcd. Para ello, el manual de `GSEA` recomienda utilizar el módulo `ExpressionFileCreator` de `GenePattern`, que es lo que

hacen los autores del artículo. En esta sección se analizan y normalizan los 12 ficheros `.CEL` de manera simultánea, tal y como se indica en el trabajo original. La normalización, al igual que en las secciones anteriores, se hace con el método RMA y normalizando por cuantiles. Para simular el experimento del trabajo original, se ejecuta el análisis de enriquecimiento funcional en todos los conjuntos de datos de genes diana de factores de transcripción s (C3 TFT) con el conjunto GSI-NOTCH. Establecemos el parámetro `permutation_type` a `gene_set`, ya que no disponemos de más de 7 muestras de cada fenotipo para el análisis.

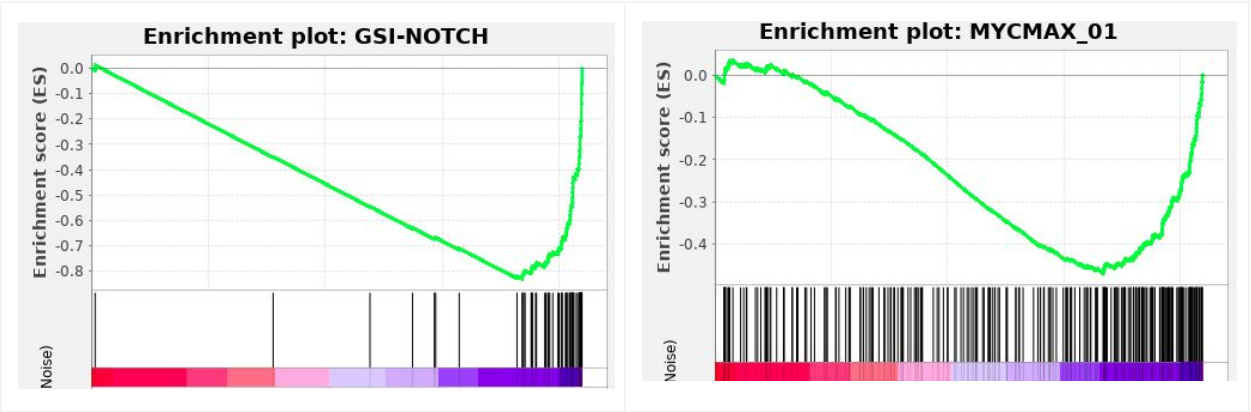


Figura 8: Comparación cuantitativa de los genes subregulados por GSI (GSI-NOTCH y MYC/MAX) con SAHM1 por GSEA.

El conjunto GSI-NOTCH aparece como el valor estadístico más atípico en los perfiles SAHM1 (Figura 7) con un FDR<0.25. También, al igual que en el trabajo original, podemos ver que activadores transcripcionales como MYC/MAX, que han sido previamente identificados como objetivos aguasabajo de NOTCH. Las snapshot de los conjunto GSI-NOTCH y MYC/MAX se encuentran en la Figura 8. En el heatmap de la Figura 9 se muestran genes más significativamente subregulados por SAHM1. Es interesante encontrar genes de la sección anterior como HES4 y DTX1.

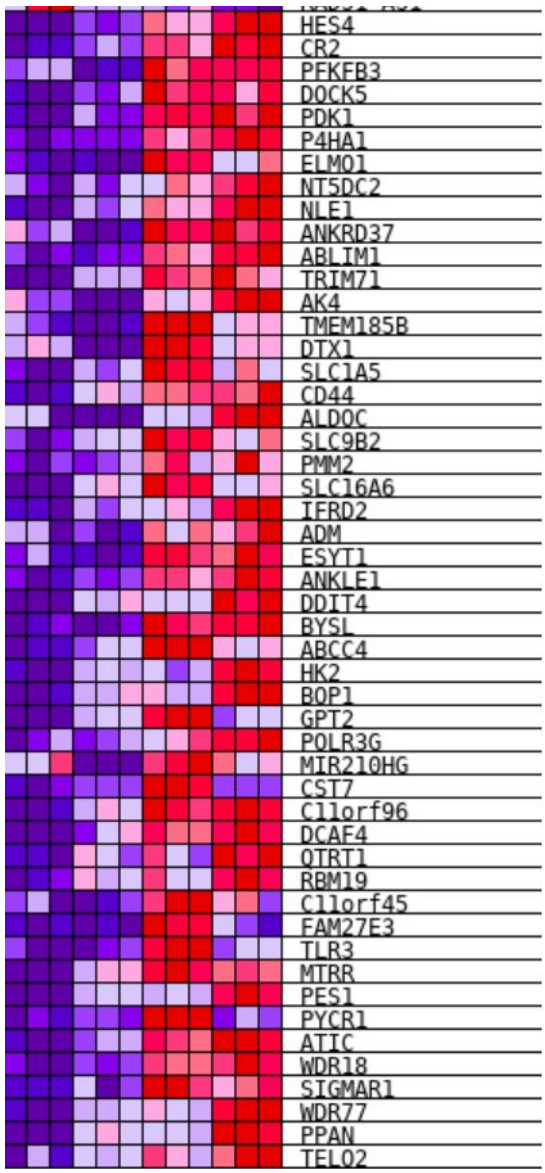


Figura 9: heatmap con el top 50 de los genes más significativamente subregulados por el tratamiento con SAHM1.

CONCLUSIONES

Descubrir tratamientos que actúen sobre factores de transcripción es complicado por la alta complejidad química de estas moléculas. En este trabajo se intenta evaluar si el tratamiento con SAHM1 podría dirigirse al factor de transcripción NOTCH. Para ello, se lleva a cabo un análisis de expresión diferencial en muestras tratadas con DMSO o DMSO+SAHM1 en dos líneas celulares de T-ALL: HPB-ALL y KOPT-K1.

En el análisis de expresión diferencial de las dos líneas celulares se observa que parte de los genes comúnmente subregulados (Figura 6) se encuentran en el conjunto de genes de GSI-NOTCH, indicando que SAHM1 podría ser un buen tratamiento para la inhibición de los genes activados por NOTCH. No obstante, también se han encontrado diferencialmente regulados algunos genes que no se encuentran en el conjunto de genes inhibidos por el tratamiento tradicional con GSI, lo que podría causar efecto en otras rutas celulares, por lo que es imposible descartar de manera concluyente cualquier actividad fuera del objetivo del tratamiento con SAHM1.

Del análisis de enriquecimiento de conjuntos de genes (GSEA) podemos concluir que SAHM1 ejerce un efecto antagonista específico sobre la expresión génica dirigida por NOTCH (Figura 7), ya que se proporciona una sorprendente correlación entre los efectos de expresión de SAHM1 y GSI. Por lo tanto, se puede concluir que el principal objetivo de SAHM1 es la ruta de señalización de NOTCH.

Esto nos permite concluir que SAHM1 es un posible tratamiento dirigido para cánceres controlados por NOTCH, como la leucemia linfoblástica aguda en células T.

BIBLIOGRAFÍA

1. Moellering, R. E., Cornejo, M., Davis, T. N., Del Bianco, C., Aster, J. C., Blacklow, S. C., ... & Bradner, J. E. (2009). Direct inhibition of the NOTCH transcription factor complex. Nature, 462(7270), 182-188.
2. Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 30(1), 207-210.
3. 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.
4. Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics, 20(3), 307-315.

MATERIAL

El código desarrollado durante este trabajo puede encontrarse en el repositorio:

`https://github.com/SaraLite/TRRPP_GSEA.git`