Universidad Europea MADRID

Título del Trabajo:

Análisis mecanístico de Sarcoma de Células Claras del Riñón para el reposicionamiento racional de fármacos mediante Machine Learning

Autora: María del Carmen Prado Zamora

Tutora del trabajo: Dra. María Peña Chilet
Cotutor: Dr. Carlos Loucera

Tutora académica: Dra. María Peña Chilet

Trabajo Fin de Master
Facultad de Ciencias Biomédicas y de la Salud
Titulación: Máster Universitario en Bioinformática
Curso 2021-2022

Índice

1 Resumen 1
2 Introducción 2
2.1 Expresión génica 2
2.2 Análisis de datos de expresión génica mediante análisis mecanístico de rutas de señalización y mapas de enfermedad. 2
2.3 Hipathia 3
2.4 Reposicionamiento de fármacos 5
2.5 Información anotada y curada 6
2.5.1 Gene Ontology (GO) 6
2.5.2 UniProt. 7
2.5.3 Kioto Encyclopedia of Genes and Genomes (KEGG). 7
2.5.4 Drug bank 7
2.6 Repositorios de datos 8
2.6.1 Gene Expression Omnibus (GEO) 8
2.7 Sarcomas 9
2.7.1 Sarcoma de Células Claras del Riñón (SCCR) 9
3 Hipótesis de trabajo 10
4 Objetivos 11
5 Métodos 12
5.1 Datos 12
5.2 Softwares informáticos 12
5.3 Análisis estadístico 12
6 Resultados 15
6.1 Selección del DataSet 15
6.2 Descarga de los datos desde GEO 15
6.3 Importación de los datos a R 16
6.4 Preparación de la matriz de expresión 16
6.5 Preparación de la matriz de diseño experimental 19
6.6 Descarga de rutas señalización humanas 21
6.7 Nivel de activación de las rutas de señalización 21
6.8 Anotación funcional con *GeneOntology* y *UniProt* 25
6.9 Visualización de la activación/inactivación de las funciones anotadas por GO y UniProt de los tres grupos de pacientes. 26
6.10 Comparativa del nivel de activación de las rutas de señalización entre “Clear Cell Sarcoma Kidney” vs “Kidney Control” 27

6.11 Visualización de las rutas 32
6.12 Ruta de señalización PI3K/Akt 33
6.13 Reposicionamiento de fármacos 34
7 Discusión 38
8 Conclusiones 40
9 Bibliografía 41
10 Anexo I:Rutas modificadas significativamente en CCSK 43
11 Anexo II: Circuitos upregulados en SCCR 47
12 Anexo III: Código de R utilizado en el análisis 56
13 Anexo IV: Código para selección KDTs 65

Índice de Figuras

Figura 1. Nivel de expresión de los genes en la matriz de expresión sin normalizar……..18

Figura 2. Nivel de expresión de los genes en la matriz de expresión normalizada……19

Figura 3. Heatmap de la intensidad de activación de los circuitos de señalización calculada mediante el algoritmo hipathia……24

Figura 4. Análisis de Componentes Principales de la actividad de las rutas de señalización en los tres grupos de muestras……25

Figura 5. Heatmap en el que se representan las funciones celulares anotadas por GeneOntology……26

Figura 6. Heatmap en el que se representan las funciones celulares anotadas por UniProt……27

Figura 7. Heatmaps en los que se representan la activación de los circuitos significativos, de los significativos upregulados y de los significativos downregulados……30

Figura 8. VolcanoPlot de la actividad de los circuitos de señalización en SCCR……31

Figura 9. Grafo que representa la ruta de señalización hsa 05214/Glioma. En rojo aparecen los circuitos y los genes activados en CCSK……33

Figura 10. Grafo que representa la ruta de señalización hsa 04961/Reabsorción de calcio. En azul aparecen los circuitos y los genes desregulados en CCSK……33

Figura 11. Grafo que representa la ruta de señalización hsa 04151/PI3K-Akt en las muestras de CCSK……34

Índice de Tablas

Tabla 1. Rutas de señalización con mayor porcentaje de circuitos modificados significativamente en muestras de CCSK. (Listado completo en Anexo I)……32

Abreviaturas

CIE Clasificación Internacional de Enfermedades

FAIR (localizables, accesibles, interoperables y reutilizables)

FK Hígado Fetal (del inglés *Fetal Kidney*)

GEO Gene Expression Omnibus

GO *Gene Ontology*

KDT Diana conocida de fármaco (del inglés *Known Drug-Target*)

KEGG Kioto Encyclopedia of Genes and Genomes

ML Machine Learning

MORF *Random Forest Multi Output*

NCBI Centro Nacional para la Información Biotecnológica

PCA Análisis de Componentes Principales (del inglés *Principal Componet Analysis*)

SCCR Sarcoma de Células claras del Riñón (**CCSK** *Clear Cell Sarcoma Kidney* en inglés)

TW Tumor de Wilms

1 Resumen

Los estudios de expresión génica suelen estar basados en cuantificaciones de la expresión y en anotaciones que tienen en cuenta la distribución de los genes en grupos funcionales. Los análisis mecanísticos añaden información más cercana a la biología de sistemas, con ello se intenta modelizar, no sólo la presencia de los factores de estudio, sino también sus interacciones de un modo dinámico.

El paquete Hipathia, que fue desarrollado y publicado por el Departamento de Genómica Computacional del Centro de Investigación Príncipe Felipe de Valencia, y actualmente sigue en desarrollo continuo por el Área de Bioinformática Clínica del Hospital Virgen del Rocío de Sevilla, analiza la expresión génica con un enfoque mecanístico. Calcula activaciones funcionales mediante un algoritmo iterativo que modeliza la transducción de señal en las rutas de señalización biológica.

Con esta aproximación, hemos re analizado un set de datos de expresión génica de Sarcoma de Células Claras del Riñón depositado en un repositorio público, y hemos podido concluir que, si bién los genes de la ruta de señalización de PI3K/Akt se encuentran sobreexpresados, es posible que dicha sobreexpresión no implique una activación de las funciones celulares controladas por dicha ruta.

El mapa de enfermedad obtenido de la activación diferencial de circuitos de señalización celular sirve como partida de un modelo de Machine Learning de tipo *Random Forest Multi Output* para identificar Genes Diana de Fármacos aprobados por humanos, y para seleccionar posibles fármacos candidatos al reposicionamiento para Sarcoma de Células Claras del Riñón.

Palabras clave: Transcriptómica, Modelos Mecanísticos, Mapa de Enfermedad, *Machine Learning*, Reposicionamiento de Fármacos.

2 Introducción

2.1 Expresión génica

Recordando el longevo dogma de la biología molecular, la expresión génica es la síntesis de RNA mensajero (transcripción) a partir del DNA génico, y es el paso previo a la síntesis proteica (traducción). Por lo tanto, el transcriptoma nos ofrece una visión clave entre genotipo y fenotipo. La transcriptómica es una aproximación fundamental dentro de las técnicas de *high-throughput* para situar dentro de un contexto biológico los especímenes sometidos a este tipo de análisis. Al hacer este tipo de estudios hay que tener presente toda la información que nos aportan otras técnicas: genoma no codificante, codificante no expresado, RNAs no mensajeros, epigenética, proteómica y metabolómica, principalmente. Los nuevos estudios de RNAseq de célula única también aportan una información adicional que queda fuera del alcance de la aproximación más convencional. No obstante, los estudios de expresión génica no dejan de ser una técnica que por sí sola aporta una valiosa información para conocer el contexto funcional, bien sea de distintos tejidos, estatus de enfermedad, de tratamientos, cambios temporales, etc.

La expresión génica comenzó a estudiarse de forma masiva mediante plataformas de microarrays, basadas en hibridación, y posteriormente ha continuado mediante técnicas de RNAseq, basadas en secuenciación. Las plataformas de hibridación se limitan a los genes preestablecidos en la plataforma, mientras que las de secuenciación pueden abarcar secuencias no consideradas de antemano. A su vez, dentro de cada tecnología hay diferentes tipos de plataformas que han ido evolucionando.

2.2 Análisis de datos de expresión génica mediante análisis mecanístico de rutas de señalización y mapas de enfermedad.

Los resultados de una técnica de expresión génica pueden analizarse siguiendo varias aproximaciones (Li, y otros, 2016).

Los análisis de **expresión diferencial** nos ofrecen un listado de genes que se expresan en mayor o menor medida, y con mayor o menor grado de significación,

entre casos y controles, entre grupos discretos de análisis, o con respecto a alguna característica de interés de naturaleza continua.

Un siguiente paso para aportar más información a estos análisis de expresión, es hacer estudios de **enriquecimiento funcional**. En estos estudios introducimos en nuestro análisis información adicional previamente anotada y curada en ontologías o bases de datos. Los análisis de sobre-representación estudian si una lista de genes de nuestro interés (asociada con una determinada funcionalidad) se expresa de forma diferente significativamente entre nuestros grupos de estudio. Los análisis de enriquecimiento de un set de genes estudian si los genes asociados a una determinada función están diferencialmente expresados entre grupos de muestras, sin hacer una preselección previa de los genes y teniendo en cuenta la intensidad de expresión de los genes de un modo ordenado. Por último, los análisis de enriquecimiento de módulos aplican una visión de biología de sistemas, modelizando mediante grafos las redes biológicas, y posteriormente buscando diferencias entre grupos una vez aplicadas las modelizaciones. Este último enfoque es el que utiliza el paquete Hipathia, añadiendo, además del estudio topológico de los grafos, un estudio mecanístico (Rian, y otros, 2021).

2.3 Hipathia

El algoritmo Hipathia (**HI**gh throughput **PATH**way **I**nterpretation and **A**nalysis) (Hidalgo, y otros, 2017) fue desarrollado y publicado por el Departamento de Genómica Computacional del Centro de Investigación Príncipe Felipe de Valencia, y actualmente continúa su desarrollo en el Área de Bioinformática Clínica del Hospital Virgen del Rocío de Sevilla. Hipathia modeliza las redes biológicas de señalización proteica mediante grafos. Los nodos de los grafos representan a proteínas únicas o a conjuntos de proteínas que actúan de modo conjunto, las aristas de los grafos corresponden a interacciones proteicas de transmisión de señal. Esta transmisión tiene sentido (es decir actúa unidireccionalmente desde un nodo hasta el siguiente), y tienen signo (la interacción puede ser positiva; de activación, o negativa; de inhibición). Al comienzo de las rutas se encuentran los nodos receptores que reciben la primera señal y al final los nodos efectores que producen un efecto funcional. Considera que, al comienzo de las rutas, formada por nodos receptores se recibe una señal

a la que se le asigna un valor arbitrario de 1. Dicha señal se va transmitiendo a través del circuito con una intensidad que depende de la estructura de la ruta, del signo de interacción entre nodos (activación-inhibición), y del nivel de expresión del gen representado en cada nodo. La computación de la señal se hace mediante un algoritmo iterativo de propagación recursiva que recorre los circuitos de señalización celular y nos ofrece como resultado un valor de activación final de las funciones que realicen los nodos efectores.

El estudio de las funcionalidades mediante este sistema nos acerca más a la biología real que las aproximaciones que no tienen en cuenta la red de interacciones. Podría ocurrir que, en una ruta determinada, la mayor parte de los genes estén muy expresados, pero haya alguno poco expresado que impida que se transmita la señal correspondiente. En ese caso un análisis de sobrerrepresentación o de enriquecimiento de set de genes nos diría que esa determinada función está activada, mientras que un análisis mecanístico podría detectar que en realidad no lo está.

La herramienta Hipathia puede emplearse como aplicación web de acceso libre (htpp://hipathia.babelomics.org/), en la que se pueden realizar estudios de expresión diferencial, predicción de clases, efecto de perturbación e intérprete de variantes. También puede utilizarse mediante un paquete en R de Bioconductor (https://www.bioconductor.org/pakages/release/bioc/html/hipathia.html), lo que aumenta su personalización, y mediante un pluging de cytoscape Cypathia (https://apps.cytoscape.org/apps/cypathia). Hipahia tabaja con grafos cuya información se encuentra codificada en ficheros .sif y .att que controlan respectivamente la interacción entre nodos y las especificaciones para su visualización.

Hipathia, y la mayoría de métodos de análisis mecanísticos, toman como base la topología de circuitos de KEGG (Enciclopedia de Kioto de Genes y Genomas). Un ejemplo de otros métodos de análisis mecanísticos son TAPPA, DEGraph, SubSPIA o MinePath entre otros. (Amadoz, Hidalgo, C¸ubuk, Carbonell-Caballero, & Dopazo, 2019)

2.4 Reposicionamiento de fármacos

El desarrollo de fármacos conlleva un proceso largo costoso y con muy baja tasa de éxito sobre el total de compuestos probados (Ko, 2020) (Jourdan, Bureau, Rochais, & Dallemagne, 2020). El análisis de la expresión génica con un enfoque mecanístico puede servir como punto de partida para realizar estudios de reposicionamiento de fármacos. El reposicionamiento de fármacos consiste en la búsqueda de nuevas indicaciones para medicamentos ya comercializados (Dudley, Deshpande, & Butte, 2011), (Xue, Li, Xie, & Wang, 2018). Esta estrategia intenta conseguir que la obtención de un nuevo tratamiento para una enfermedad determinada se realice de un modo más rápido, menos costoso y con mayor seguridad (Luo, y otros, 2021), (Orpea & Overington, 2015). La rapidez supone una ventaja para enfermedades emergentes en las que se tiene menos tiempo de reacción. El menor coste hace que este método sea muy adecuado para la búsqueda de tratamiento de enfermedades raras dado que su escasa población diana las hace muy poco atractivas a las grandes inversiones privadas de las empresas farmacéuticas (Cha, y otros, 2018). El tratamiento de diversos tipos de cáncer también se puede ver beneficiado por esta estrategia puesto que, a pesar de la heterogeneidad de este conjunto de enfermedades, hay procesos que pueden ser comunes a varias enfermedades. Las resistencias antimicrobianas son un tema de salud altamente preocupante que también puede ser abordado desde esta perspectiva. Otra ventaja del reposicionamiento de fármacos es que, a partir de fármacos ya comercializados, contamos con mucha evidencia sobre su seguridad. El reposicionamiento de fármacos se puede llevar a cabo mediante métodos experimentales o métodos computacionales (Jarada, Rokne, & Alhajj, 2020), tras los cuales es necesario llevar a cabo una validación de los resultados. (Brown & Patel, 2018)

El paquete Hipathia se ha utilizado en estudios publicados de varias enfermedades, como Covid-19 (Loucera, y otros, 2020) (López-Sánchez, Loucera, Peña-Chilet, & Dopazo, 2022), Anemia de Fanconi. (Esteban-Medina, Peña-Chilet, Loucera, & Dopazo, 2019) En estos estudios con ayuda de Hipathia se modelizaron las enfermedades (Ostaszewski & et al., COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms, 2020), (Ostaszewski & et al., COVID19 Disease Map, a

computational knowledge repository of virus-host interaction mechanisms, 2021) y posteriormente con información de bases de datos de expresión en tejidos y de fármacos aprobados para su uso en humanos y mediante análisis de Machine Learning (*Random Forest multi-output* (MORF)) se seleccionaron fármacos candidatos a ser reposicionados. Posteriormente, de hecho, muchos de estos fármacos están formando parte de ensayos clínicos para dichas enfermedades.

2.5 Información anotada y curada

La anotación y curación de la información es un proceso imprescindible para dotar de significado biológico a la gran cantidad de datos que nos ofrecen las técnicas *high-throughput*, y ponerlos en contexto con todo el conocimiento ya existente. Existen múltiples ontologías y bases de datos que nos ofrecen esa información, cada una de ellas con características diferenciales que las harán más útiles y necesarias en distintas situaciones. Entre ellas están las siguientes:

2.5.1 Gene Ontology (GO)

Gene Ontology (GO) (http://www.geneontology.org), es un consorcio para desarrollar un modelo computacional para sistemas biológicos que forma parte de *Alliance of Genome Resources*. Este consorcio comenzó en 1998, con la puesta en común de las investigaciones del genoma de tres modelos animales: *Drosophila melanogaster*, *Mus musculus* y *Saccharomyces cerevisiae*. Actualmente Incluye información de más de 5000 especies de seres vivos, incluyendo no sólo animales, entre ellos *Homo sapiens* y *Mus musculus*, sino también especies vegetales, hongos, bacterias y virus. Compendia el conocimiento sobre las funciones de los genes de un modo comprensible tanto para máquinas como para humanos. Abarca desde el nivel molecular, pasando por rutas, nivel celular y nivel de organismo. Su ontología incluye más de 40.000 términos válidos, 28.000 procesos biológicos, 11.000 funciones moleculares y 4.000 componentes celulares. A fecha de la presente revisión, incluye 1.503.740 productos génicos anotados de 5.257 especies, 185 de las cuales superan las 1.000 anotaciones. El número total de anotaciones es de 7.694.564. *GOterms* es un vocabulario controlado basado en tres categorías (Proceso biológico, función molecular y componente celular).

2.5.2 UniProt.

UniProt (https://www.uniprot.org) es una base de datos que incluye secuencias proteicas e información funcional. *UniProtKnowledgeBa*se incluye la base de datos *Swiss-prot* con 568.363 resultados manualmente curados y la base de datos *TrEMBL*, con 229.928.140 resultados revisados computacionalmente pendientes de revisión manual. Se pueden hacer búsquedas de proteínas y seleccionar por datos curados o no curados, organismo, longitud de la secuencia y cantidad y calidad de los datos disponibles. Se puede seleccionar una única entrada para hacer una búsqueda directa de similaridad en BLAST, o seleccionar múltiples entradas para alineamientos. Se pueden descargar los datos en varios formatos. *UniProtKB keywords* es un vocabulario controlado desarrollado para las necesidades del contenido de UniProtKB/Swiss-Prot que indexa sus entradas en 10 categorías (Proceso biológico, componente celular, diversidad de secuencia codificante, estadio de desarrollo, enfermedad, dominio, ligando, función molecular, modificación post-transcripcional y términos técnicos). Las keywords son atribuidas a las entradas de UniProtKB/Swiss-Pro de un modo manual. A su vez, también es manual la correlación de las Keywords con otras ontologías como GOterms, EC number, InterPro o HAMAP.

2.5.3 Kioto Encyclopedia of Genes and Genomes (KEGG).

The Kioto Encyclopedia of Genes and Genomes (KEGG) (https://www.genome.jp/kegg) es una base de datos para conocer funcionalidades de alto nivel de los sistemas biológicos como la célula, el organismo y los ecosistemas a partir de información a nivel molecular de secuenciación a gran escala y otras tecnologías experimentales de tipo *high-throughput.* Su información está incluida en 16 bases de datos repartidas entre cuatro categorías (Sistemas, Genómica, Química y Salud). Dentro de la categoría de sistemas se encuentra la base de datos KEGG PATHWAY que contiene los mapas de rutas que utilizan muchos de los modelos mecanísticos anteriormente nombrados.

2.5.4 Drug bank

Drug bank (https://go.drugbank.com) es una base de datos que ofrece información en abierto sobre fármacos, sus mecanismos moleculares, sus

interacciones y sus dianas. Se denominan Dianas de Fármacos Conocidas (KDT, del inglés K*nown Drug Targets* KDT), a las proteínas (y por extensión a sus genes) sobre las que se conoce que actúa un fármaco. La plataforma se creó en 2006 en la universidad de Alberta. En la actualidad incluye un total de 14931 fármacos

2.6 Repositorios de datos

Los resultados de las técnicas *high-throughput,* junto con los metadatos que los acompañan, por lo general se depositan en repositorios en los que poder mantenerlos siguiendo los principios FAIR (localizables, accesibles, interoperables y reutilizables). El depósito de los datos, además de estar cada vez más aceptado por la comunidad científica, en muchos casos, y cada vez más, puede ser obligatorio, sobre todo cuando los datos se han obtenido con fondos público. El depósito, dependiendo del caso, puede ser o no ser de acceso abierto, y tendrá que tener en cuenta los principios éticos y reglamentarios de protección de datos personales. Existen multitud de repositorios, de diversa naturaleza, ajustados en mayor o menor medida a principios FAIR, y que pueden enfocarse preferencialmente a algunos de los aspectos de dichos principios. Por ejemplo, pueden estar más enfocados a proporcionar gran capacidad de almacenamiento, a facilitar la localización y/o la descarga, a permitir el re análisis mediante herramientas incluidas en el propio repositorio, etc.

2.6.1 Gene Expression Omnibus (GEO)

Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo) es un repositorio perteneciente al NCBI (Centro Nacional para la Información Biotecnológica) de EEUU. GEO está enfocado al depósito de datos de expresión génica. Sus datos están organizados en varios niveles y con varios métodos de búsqueda. Hay datos de muestras únicas, con identificadores únicos del tipo GSMxxxxx, datos de series que agrupan varias muestras, con id: GSExxxxx y datos agrupados en Datasets (GDSxxxxx). Los datasets incluyen a muestras de series cuya información ha sido curada por GEO. También se pueden hacer búsquedas por plataforma de análisis, por organismo, búsquedas avanzadas de buscadores convencionales y perfiles de genes, transversales a los estudios de la plataforma.

2.7 Sarcomas

Las enfermedades que más se pueden beneficiar de un estudio mecanístico, son aquellas en las que estén desreguladas sus rutas de señalización. El cáncer es un conjunto heterogéneo de enfermedades con una acumulación de mecanismos alterados que se pueden analizar mecanísticamente. A su vez sarcoma es la denominación para los procesos cancerígenos de huesos y tejidos conectivos, que incluye más de 70 variedades. Los sarcomas son tumores raros que corresponden con aproximadamente el 1 % de todos los cánceres en adultos y el 15% en niños. Los sarcomas, son susceptibles de ser estudiados mediante una aproximación mecanística.

2.7.1 Sarcoma de Células Claras del Riñón (SCCR)

El Sarcoma de Células claras del Riñón (SCCR), (en inglés, *Clear Cell Sarcoma Kidney* (CCSK)) suele aparecer en niños de 2-3 años de edad (Zhang, Chu, Ma, Miao, & Diao, 2022) y constituye entre el 3-5% de los tumores renales pediátricos (Ding, Yao, & Chen, 2022). Sus identificaciones por el sistema CIE son: CIE-10: C64, CIE-11: XH0765. Además tiene identificación como enfermedad rara en el sistema de Orphanet: ORPHA:457246. Aparece habitualmente en pacientes menores de 3 años y presenta peor pronóstico que otros tumores renales como el Tumor de Wilms (WT). Su diagnóstico suele ser complicado y el tratamiento no suele evitar las recaídas y metástasis. Las estrategias terapéuticas son la resección quirúrgica y quimioterapia adyuvante, principalmente doxorrubicina, que produce efectos secundarios cardiotóxicos.

3 Hipótesis de trabajo

El estudio de datos transcriptómicos de SCCR depositados en repositorios públicos, nos permite comprender más el mecanismo de la enfermedad e identificar posibles dianas terapeúticas de fármacos ya aprobados, permitiendo su reposicionamiento para el tratamiento de la enfermedad.

4 Objetivos

Para mejorar el conocimiento de los procesos mecanísticos de SCCR, y sentar las bases para realizar un reposicionamiento de fármacos, los objetivos del presente trabajo son:

- Reanalizar datos de RNAseq publicados y depositados en un repositorio público de muestras de SCCR.
- Comprobar si en enfoque mecanístico que realiza Hipathia añade información a los datos ya publicados.
- Obtener una lista de posibles genes diana con potencial terapéutico en SRCC.

5 Métodos

5.1 Datos

Los datos analizados pertenecen al Dataset GDS1282 de Sarcoma de Células Claras del Riñón (SCCR), que se encuentran depositados en GEO en modo abierto. El dataset GDS1282, procede de la serie GSE2712. Incluye 14 muestras de pacientes de CCSK, 15 muestras de WT y 3 muestras de Riñón Fetal que sirven como muestras Control (Cutcliffe, y otros, 2005). El procesamiento original de las muestras publicado es el siguiente: Las muestras de tejido congelado procedían del Banco de Tumores Renales del Grupo de Oncología Infantil del *Children's Memorial Hospital* en Chicago, Illinois, y se utilizaron con autorización del Comité de Revisión correspondiente. Se evaluaron histológicamente y se seleccionaron muestras con celularidad mayor del 80%. Se extrajo el RNA, se sintetizó primero el cDNA y posteriomente el cRNA marcado con biotina. Luego se añadió al array de nucleótidos Affymetrix HG-U113A para su hibridación durante 16 horas. Se lavó y se escaneó con el escáner HP GeneArray con una longitud de onda de excitación de 488 nm. La cantidad de luz emitida a 570 es proporcional a la cantidad de unión a cada posición del microarray. Los datos en crudo se obtienen en archivos tipo cel. Posteriormente los datos se sometieron a procesos de limpieza y normalización.

5.2 Softwares informáticos

Los análisis se realizaron con R versión 4.2.1 "Funny-Looking Kid", utilizando RStudio. Se instalaron los paquetes "BiocManager", "GEOquery", "hipathia", "tidyverse", y "EnhanceVolcano". El paquete de Python "dreml" DREM^{3}L (Drug Repurposing using Mechanistic Models of signal transduction and Machine Learning).

5.3 Análisis estadístico

La activación de las rutas y los circuitos de señalización se calculó mediante "hipathia". Para ello emplea un algoritmo que computa de un modo iterativo la transducción de la señal por rutas de señalización. Este cálculo se realiza

mediante la función “hipathia::results”. Para hacer este cálculo, hipatia tiene en cuenta

Visualización de rutas con la función “hipathia::pathway_comparison_plot(”). El paquete hipathia nos permite visualizar las rutas de señalización como grafos direccionales y con signo: activación (flecha) o de inhibición (T). Sobre los gráfos, el código de colores nos indica si en nuestras muestras las interacciones y/o los genes están upreguladas (rojo) o downreguladas (azul)con respecto al grupo control. Podemos seleccionar la ruta que queramos visualizar, incluyendo los cambios en nuestro grupo de muestras utilizando la función “pathways_summary”.

Visualización mediante heatmaps de la actividad de las rutas en las distintas muestras. Nos muestra garficamente la intensidad de las distintas rutas y nos da una clusterización de las muestras que nos enseña por medio de dendrogramas que nos indican tanto la similitud de las muestras como la de los circuitos. Para ello se utiliza la función “hipathia::heatmap_plot”.

Análisis de Componentes Principales (PCA) mediante la función hipathia::do_pca. El análisis de componentes principales es un método de reducción de la dimensionalidad. Estos métodos nos permiten representar en dos (o tres dimensiones) las características de alta dimensionalidad (como lo es en este caso, la activación de rutas) que presentan cada muestra.

Wilcoxson test con p.valor corregido por el método Benjamin-Hochberg FDR mediante la función hipathia::do_wilcoxon, nos permite hacer una comparativa del cambio de nivel de activación de cada circuito comparando enfermos con controles. Al valor de activación diferencial, le acompaña un p.valor y un p.valor corregido para dar información de la significación de los cambios.

Con los gráficos de tipo Volcano “EnhancedVolcano()”, se representan gráficamente los cambios de nivel entre grupos (bien sean cambios de expresión o de activación), frente a la significación estadística.

Con el paquete “drexml” de Python, (DRExM^{3}L) Drug REpurposing using Mechanistic Models of signal transduction and eXplainable Machine Learning (DOI 10.5281/zenodo.7294338.) (Loucera, Esteban-Medina, & Peña-Chilet, (DRExM3L)

Drug REpurposing using Mechanistic Models of signal transduction and eXplainable Machine Learning (v0.9.5), 2022) utiliza algoritmos de Machine Learning basados en MORF, se hizo una priorización de KDTs para una enfermedad definida por su mapa de señalización.

6 Resultados

6.1 Selección del DataSet

Para seleccionar el DataSet para nuestro estudio hicimos una búsqueda en GEO. Buscamos DataSets de Sarcoma en Humanos ((sarcoma) AND "Homo sapiens"[porgn:__txid9606]). Obtuvimos un total de 40 resultados posibles. Posteriormente hicimos una revisión manual de las características de los estudios para ver cual se adecuaba a nuestros requerimientos. Ordenamos los estudios de mayor a menor número de muestras y seleccionamos los 25 primeros que eran los que tenían 10 o más muestras. A continuación, descartamos los dataset que analizaban líneas celulares y los que analizaban pacientes, pero no incluían muestras de controles. Finalmente, el único DataSet que cumplía con las caracteríaticas requeridas fue el GDS1282 que procede del estudio GSE2712. Este dataset contine un total de 35 muestras, procedentes de pacientes pediátricos: 14 pacientes de CCSK, 15 pacientes de WT (+ 3 duplicados de tres de las muestras) y 3 controles de riñón fetal. Los resultados del estudio de estas muestras se publicaron en el artículo: *Clear Cell Sarcoma of the Kidney: Up-regulation of Neural Markers whit Activation of the Sonic Hedgehog and Akt Pathways.* Colleen Cutcliffe. Clin Cancer Res 2005.

6.2 Descarga de los datos desde GEO

Vamos a analizar el dataset GDS1282. Este dataset procede de información curada por GEO a partir del estudio GSE2712. Están disponibles varios tipos de archivos que contienen información de metadatos y datos con distinto grado de procesamiento.

Los archivos descargables asociados a GDS1282 son los siguientes:

DataSet full SOFT file

DataSet SOFT file

Series family SOFT file

Series family MINiML file

Annotation SOFT file

6.3 Importación de los datos a R

Una vez descargados y desempaquetados los archivos de GEO en nuestro PC, importamos los datos a R utilizando la función GEOquery::getGEO().

```
gds1282 <- getGEO('GDS1282', destdir=".")
```

Obtenemos un objeto de clase “GDS” que tiene información de datos y de metadatos. Contiene información sobre la expresión génica de muestras de SCCR, de TK y controles de riñón fetal. Los datos se han obtenido en la plataforma GPL96, que corresponde con el array de Affymetrix de Genoma Humano U133A.

6.4 Preparación de la matriz de expresión

Importados los datos, creamos la matriz de expresión con la función “GEOquery::Table()”

```
expression_matrix_gds1282_raw<-Table(gds1282)
```

Obtenemos un data.frame de tamaño 22283 X 37 en el que las filas corresponden con los genes analizados y las columnas con las muestras.

```
expression_matrix_gds1282_raw[1:10,1:6]

##       ID_REF IDENTIFIER GSM52495 GSM52496 GSM52497 GSM52463
## 1  1007_s_at    MIR4640  14386.2  10014.2  12416.2  13078.3
## 2    1053_at       RFC2    442.5    721.9   1078.2   1395.8
## 3     117_at      HSPA6    700.3    664.3    412.9   1058.2
## 4     121_at       PAX8  38145.4  26406.2  21787.2  33754.4
## 5  1255_g_at     GUCA1A    507.9    325.2    567.1    781.6
## 6    1294_at    MIR5193   2643.8   2192.4   2147.7   1885.7
## 7    1316_at       THRA   1222.7   1110.0    934.1   1122.9
## 8    1320_at     PTPN21    596.6    555.2    570.0    975.1
## 9  1405_i_at       CCL5     49.6     47.7     42.8    234.3
## 10   1431_at     CYP2E1    740.3    234.1    371.6    607.3
```

El paquete Hipathia trabaja con una matriz de expresión cuyas filas corresponden con los genes y las columnas corresponden con las muestras y con una matriz de diseño experimental de una sola columna con el estatus de enfermedad y la nomenclatura de las muestras como nombre de filas.

Para adecuar esta matriz al formato reconocido por Hipatia, tenemos que hacer algunas modificaciones:

Eliminamos las filas con algún NA. Obtenemos una matriz del mismo tamaño lo que nos indica que no había ninguna fila con NA:

```
expression_matrix_gds1282_na<-expression_matrix_gds1282_raw[complete.cases(expression_matrix_gds1282_raw),]
```

Eliminamos las dos primeras columnas con los nombres de los genes:

```
expression_matrix_gds1282_id<-expression_matrix_gds1282_na[,-c(1,2)]
```

Cambiamos la clase de la matriz de objeto de tipo *data.frame* a un objeto tipo *matrix*:

```
expression_matrix_gds1282_m<-as.matrix(expression_matrix_gds1282_id)
```

Asignamos a los nombres de las filas, los nombres de los genes:

```
row.names(expression_matrix_gds1282_m)<-(expression_matrix_gds1282_na[,1])
```

Cambiamos el nombre de los genes a tipo EntrezID hsa (*Homo sapiesn*), que es el formato que reconoce Hipathia.

```
expression_matrix_gds1282_trans<-translate_data(expression_matrix_gds1282_m,"hsa")
```

Obtenemos una matriz de expresión aún sin normalizar:

```
expression_matrix_gds1282_trans[1:10,1:6]

##     GSM52495  GSM52496 GSM52497 GSM52463  GSM52464   GSM52465
## 2  15299.700 10576.900  11198.4 7036.100 10790.500 31487.4000
## 9   1641.200  1725.900    860.0 1365.200  1166.100  1509.1000
## 10   419.000  1017.400    259.4  512.600   543.400  1039.7000
## 12   299.100   147.500    297.2 1148.900  1070.800  1038.4000
## 13   214.900   146.400    155.6  192.500   121.400   817.3000
## 14  3222.300  2722.500   3306.1 2982.300  2616.400  3228.1000
## 15    44.700    50.800     38.2   91.100    77.500    34.7000
## 16  9717.400  7011.300   8414.1 6616.000  6725.100  9180.1000
## 18  2038.200  1963.000   1581.0 2451.067  2453.667   547.8667
## 19  2320.933  1877.267   1563.6 3077.033  2976.733  2753.6667
```

Visualizamos la matriz de expresión sin normalizar (**Figura 1**). Observamos que el rango de los datos de expresión va desde 0 hasta 400.000 En las

visualizaciones con menor número de muestras comprobamos que los valores mayoritarios están en un orden de magnitud alrededor de 100.

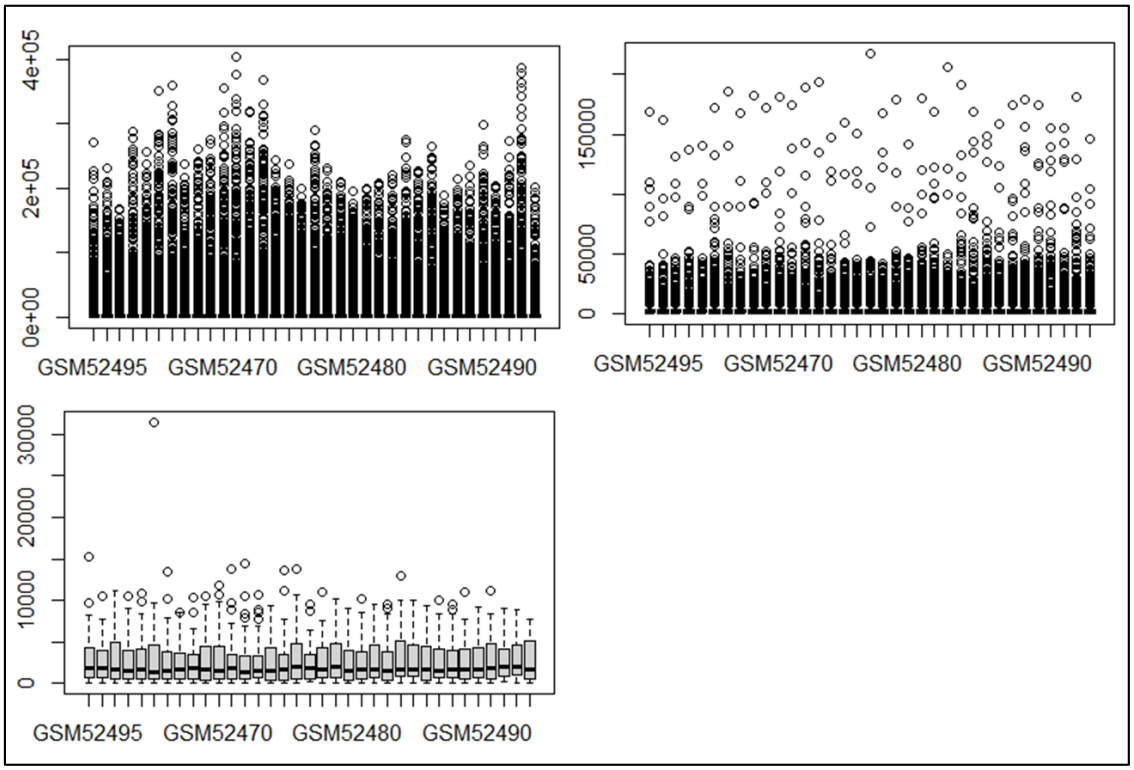

Figura 1. Nivel de expresión de los genes en la matriz de expresión sin normalizar. Representa el nivel de expresión de los genes en cada una de las muestras. En el primer boxplot se muestran todos los genes. En el segundo y tercero se muestran respectivamente los 1000 y los 40 primeros del listado, para ver la distribución en mayor detalle.

Normalizamos la matriz para que los valores de expresión estén entre 0 y 1:

```
expression_matrix_gds1282_nor<-normalize_data(expression_matrix_gds128
2_trans, by_quantiles = TRUE)
```

Comprobamos que efectivamente se ha producido la normalización:

```
expression_matrix_gds1282_nor[1:10,1:5]

##         GSM52495     GSM52496     GSM52497     GSM52463     GSM52464
## 2  0.0721030323 0.0429887947 0.0455442883 0.0273414151 0.0442841661
## 9  0.0053710543 0.0062112461 0.0029748710 0.0049273602 0.0039452191
## 10 0.0011102431 0.0036262037 0.0008621038 0.0017841960 0.0017088451
## 12 0.0007924554 0.0005225469 0.0009993821 0.0041608236 0.0035590429
## 13 0.0005740183 0.0005174576 0.0005287429 0.0006607898 0.0003570605
## 14 0.0118836526 0.0101121623 0.0122842329 0.0112580564 0.0096894856
## 15 0.0001023594 0.0001681871 0.0001346061 0.0002998209 0.0002049062
## 16 0.0423461296 0.0269894497 0.0333666167 0.0256819056 0.0266115255
## 18 0.0069373832 0.0070967633 0.0054961504 0.0092258793 0.0089697875
## 19 0.0080848411 0.0067950630 0.0054333142 0.0116283649 0.0110857614
```

Visualizamos los datos de expresión de la matriz normalizada (**Figura 2**) para comprobar que los valores de expresión se encuentran entre 0 y 1.

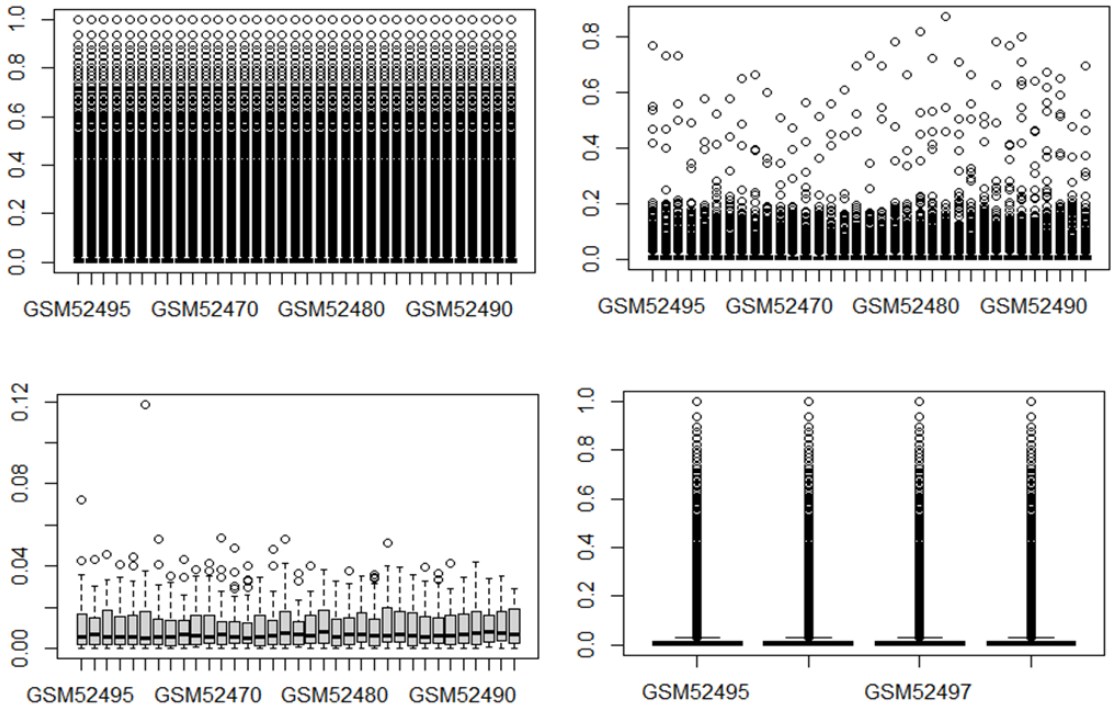

Figura 2. Nivel de expresión de los genes en la matriz de expresión normalizada. Representa el nivel de expresión de los genes en cada una de las muestras. En el primer boxplot se muestran todos los genes. En el segundo y tercero se muestran respectivamente los 1000 y los 40 primeros del listado, para ver la distribución en mayor detalle. En el cuarto boxplot se representan todos los genes de cuatro muestras para apreciar mejor la distribución de valores.

6.5 Preparación de la matriz de diseño experimental

Además de la matriz de expresión, necesitamos la matriz de diseño experimental, que nos relacione cada muestra/paciente con el estatus de enfermedad que vayamos a considerar para analizar los datos. Para obtener la matriz de diseño experimental del dataset GDS1282, utilizamos la función Columns() sobre el objeto gds1282.

```
Columns(gds1282)[,1:2]
     sample                    disease.state
1  GSM52495                          control
2  GSM52496                          control
3  GSM52497                          control
4  GSM52463                   Wilms' tumor
5  GSM52464                   Wilms' tumor
6  GSM52465                   Wilms' tumor
7  GSM52466                   Wilms' tumor
8  GSM52467                   Wilms' tumor
9  GSM52468                   Wilms' tumor
```

```
10 GSM52469                     Wilms' tumor
11 GSM52470                     Wilms' tumor
12 GSM52471                     Wilms' tumor
13 GSM52472                     Wilms' tumor
14 GSM52473                     Wilms' tumor
15 GSM52474                     Wilms' tumor
16 GSM52475                     Wilms' tumor
17 GSM52476                     Wilms' tumor
18 GSM52477                     Wilms' tumor
19 GSM52478                     Wilms' tumor
20 GSM52479                     Wilms' tumor
21 GSM52480                     Wilms' tumor
22 GSM52481 clear cell sarcoma of the kidney
23 GSM52482 clear cell sarcoma of the kidney
24 GSM52483 clear cell sarcoma of the kidney
25 GSM52484 clear cell sarcoma of the kidney
26 GSM52485 clear cell sarcoma of the kidney
27 GSM52486 clear cell sarcoma of the kidney
28 GSM52487 clear cell sarcoma of the kidney
29 GSM52488 clear cell sarcoma of the kidney
30 GSM52489 clear cell sarcoma of the kidney
31 GSM52490 clear cell sarcoma of the kidney
32 GSM52491 clear cell sarcoma of the kidney
33 GSM52492 clear cell sarcoma of the kidney
34 GSM52493 clear cell sarcoma of the kidney
35 GSM52494 clear cell sarcoma of the kidney
```

Tal y como hicimos con la matriz de expresión, tenemos que modificar la matriz de diseño experimental para que se adapte al formato que requiere Hipathia. Creamos un data.frame de una sola columna, en el que asignamos a los nombres de las filas, el nombre de las muestras, y creamos la columna *group* en la que especificamos el estatus de enfermedad escrito sin espacios.

```
> exp_design_gds1282
         group
GSM52495     C
GSM52496     C
GSM52497     C
GSM52463    WT
GSM52464    WT
GSM52465    WT
GSM52466    WT
GSM52467    WT
GSM52468    WT
GSM52469    WT
GSM52470    WT
GSM52471    WT
GSM52472    WT
GSM52473    WT
GSM52474    WT
GSM52475    WT
GSM52476    WT
GSM52477    WT
GSM52478    WT
GSM52479    WT
GSM52480    WT
GSM52481  CCSK
GSM52482  CCSK
GSM52483  CCSK
GSM52484  CCSK
GSM52485  CCSK
GSM52486  CCSK
GSM52487  CCSK
GSM52488  CCSK
GSM52489  CCSK
```

```
GSM52490  CCSK
GSM52491  CCSK
GSM52492  CCSK
GSM52493  CCSK
GSM52494  CCSK
```

6.6 Descarga de rutas señalización humanas

Hipathia analiza la expresión de los genes en el contexto de las rutas de señalización de las que forman parte sus proteínas. Para ello, el primer paso es incluir en un objeto la información de las rutas, en nuestro caso humanas ("hsa"):

```
pathways<-load_pathways(species="hsa")
```

Se cargan 146 rutas de señalización, que incluyen rutas de señalización de funciones generales de procesos celulares y rutas de señalización de enfermedades:

```
get_pathways_list(pathways)
```

```
  [1] "hsa03320" "hsa03460" "hsa04010" "hsa04012" "hsa04014" "hsa04015" "hsa04020" "hsa04022"
  [9] "hsa04024" "hsa04062" "hsa04064" "hsa04066" "hsa04068" "hsa04071" "hsa04072" "hsa04110"
 [17] "hsa04114" "hsa04115" "hsa04150" "hsa04151" "hsa04152" "hsa04210" "hsa04211" "hsa04213"
 [25] "hsa04218" "hsa04261" "hsa04270" "hsa04310" "hsa04330" "hsa04340" "hsa04350" "hsa04360"
 [33] "hsa04370" "hsa04380" "hsa04390" "hsa04510" "hsa04520" "hsa04530" "hsa04540" "hsa04550"
 [41] "hsa04610" "hsa04611" "hsa04612" "hsa04620" "hsa04621" "hsa04622" "hsa04623" "hsa04630"
 [49] "hsa04650" "hsa04660" "hsa04662" "hsa04664" "hsa04666" "hsa04668" "hsa04670" "hsa04710"
 [57] "hsa04713" "hsa04720" "hsa04722" "hsa04723" "hsa04724" "hsa04725" "hsa04726" "hsa04727"
 [65] "hsa04728" "hsa04730" "hsa04740" "hsa04742" "hsa04750" "hsa04810" "hsa04910" "hsa04911"
 [73] "hsa04912" "hsa04913" "hsa04914" "hsa04915" "hsa04916" "hsa04917" "hsa04918" "hsa04919"
 [81] "hsa04920" "hsa04921" "hsa04922" "hsa04923" "hsa04924" "hsa04925" "hsa04930" "hsa04931"
 [89] "hsa04932" "hsa04933" "hsa04950" "hsa04960" "hsa04961" "hsa04962" "hsa04970" "hsa04971"
 [97] "hsa04972" "hsa04973" "hsa04976" "hsa05010" "hsa05012" "hsa05014" "hsa05016" "hsa05020"
[105] "hsa05030" "hsa05031" "hsa05032" "hsa05034" "hsa05100" "hsa05110" "hsa05120" "hsa05130"
[113] "hsa05131" "hsa05132" "hsa05133" "hsa05134" "hsa05140" "hsa05142" "hsa05145" "hsa05150"
[121] "hsa05152" "hsa05160" "hsa05161" "hsa05162" "hsa05164" "hsa05166" "hsa05168" "hsa05169"
[129] "hsa05200" "hsa05205" "hsa05210" "hsa05211" "hsa05212" "hsa05213" "hsa05214" "hsa05215"
[137] "hsa05216" "hsa05217" "hsa05218" "hsa05219" "hsa05220" "hsa05221" "hsa05222" "hsa05223"
[145] "hsa05231" "hsa05321"
```

6.7 Nivel de activación de las rutas de señalización

El nivel de activación de las rutas en función de la matriz de expresión, se calcula mediante la función "hipathia::hipathia()". Para ello utiliza un algoritmo iterativo que calcula la transmisión de una señal desde los nodos de input (al comienzo

de cada circuito) a los nodos efectores (final de los circuitos). La transmisión de la señal se considera unidireccional, y en cada caso, de activación o de inhibición. La intensidad y tipo de señal que va a transmitir un *nodo A* a un *nodo B*, va a depender de tres factores: (1) de si la actividad de A sobre B es de activación o de inhibición, (2) del nivel de expresión de A, y (3) de la intensidad de señal que haya recibido A de nodos anteriores. La intensidad de señal del primer nodo de un circuito se establece con un valor de 1.

```
results<-hipathia(expression_matrix_gds1282_nor,pathways,decompose=FAL
SE,verbose = FALSE)
```

Los genes que no se encontraban en la matriz de expresión los añade la función estimando su valor de expresión con la media de toda la matriz de expresión. En este caso se han añadido un total de 741 genes que corresponden con un 5.57% del total de genes. Se obtiene un objeto de clase “MultiArrayExperiment”, que incluye los objetos “paths” y “nodes”:

```
results

## A MultiAssayExperiment object of 2 listed
##  experiments with user-defined names and respective classes.
##  Containing an ExperimentList class object of length 2:
##  [1] paths: SummarizedExperiment with 1876 rows and 35 columns
##  [2] nodes: SummarizedExperiment with 6826 rows and 35 columns
## Functionality:
##  experiments() - obtain the ExperimentList instance
##  colData() - the primary/phenotype DataFrame
##  sampleMap() - the sample coordination DataFrame
##  `$`, `[`, `[[` - extract colData columns, subset, or experiment
##  *Format() - convert into a long or wide DataFrame
##  assays() - convert ExperimentList to a SimpleList of matrices
##  exportClass() - save data to flat files
```

La función “hipathia::get_paths_data” extrae el objeto con los valores de la señal de actividad. Es una matriz en la que las filas son las rutas de señalización y las columnas son las muestras. Los valores numéricos corresponden al nivel de activación de cada sub-ruta en cada muestra.

```
path_vals<-get_paths_data(results,matrix=TRUE)

> path_vals[1:20,1:4]
                    GSM52495     GSM52496     GSM52497     GSM52463
P-hsa03320-37 1.801765e-05 1.070103e-05 2.287399e-05 7.377104e-07
P-hsa03320-61 7.874582e-06 1.514436e-06 6.222047e-06 1.936577e-07
```

```
P-hsa03320-46 7.235299e-07 1.378209e-06 1.428909e-05 3.281982e-07
P-hsa03320-57 1.219821e-06 5.331329e-07 1.577666e-06 6.742614e-07
P-hsa03320-64 1.245240e-05 1.048483e-05 8.607080e-06 3.073348e-06
P-hsa03320-47 4.135780e-05 1.281336e-05 1.962549e-05 2.369771e-06
P-hsa03320-65 3.610837e-06 3.071422e-06 2.713630e-06 9.163865e-07
P-hsa03320-55 4.611191e-05 3.600946e-05 3.253873e-05 6.888647e-06
P-hsa03320-56 1.910072e-05 2.404144e-05 5.125774e-05 1.052410e-05
P-hsa03320-33 1.074678e-06 1.295559e-06 1.343845e-07 3.201500e-07
P-hsa03320-58 1.245240e-05 1.048483e-05 8.607080e-06 3.073348e-06
P-hsa03320-59 2.883607e-05 1.357118e-05 3.301898e-05 3.084138e-06
P-hsa03320-63 1.441453e-04 1.340828e-04 1.215124e-04 1.954832e-05
P-hsa03320-44 2.092040e-06 9.971895e-07 2.665966e-06 1.810478e-07
P-hsa03320-36 2.189782e-05 1.994574e-05 3.148621e-05 4.105617e-06
P-hsa03320-30 1.753622e-05 2.699800e-05 1.528467e-05 3.587557e-06
P-hsa03320-28 3.543440e-05 4.009035e-05 5.663989e-05 7.182162e-06
P-hsa03320-25 5.058408e-06 4.957335e-06 2.517318e-06 7.904712e-07
P-hsa03320-21 3.457592e-05 3.026965e-05 1.653539e-05 5.332535e-06
P-hsa03320-20 3.978328e-05 5.322404e-05 3.438648e-05 8.580298e-06
```

Podemos visualizar la activación de las rutas de señalización en cada muestra mediante un heatmap (**Figura 3**). El dendrograma nos muestra que las muestras de cada uno de los tres grupos se parecen más entre sí que al resto. Además, se observa que las muestras de TW y las muestras Control se parecen más entre sí que a las de SCCR. El mismo resultado se obtiene mediante un análisis de componentes principales (**Figura4**).

```
heatmap_plot(path_vals,group=sample_group, variable_clust = TRUE)
```

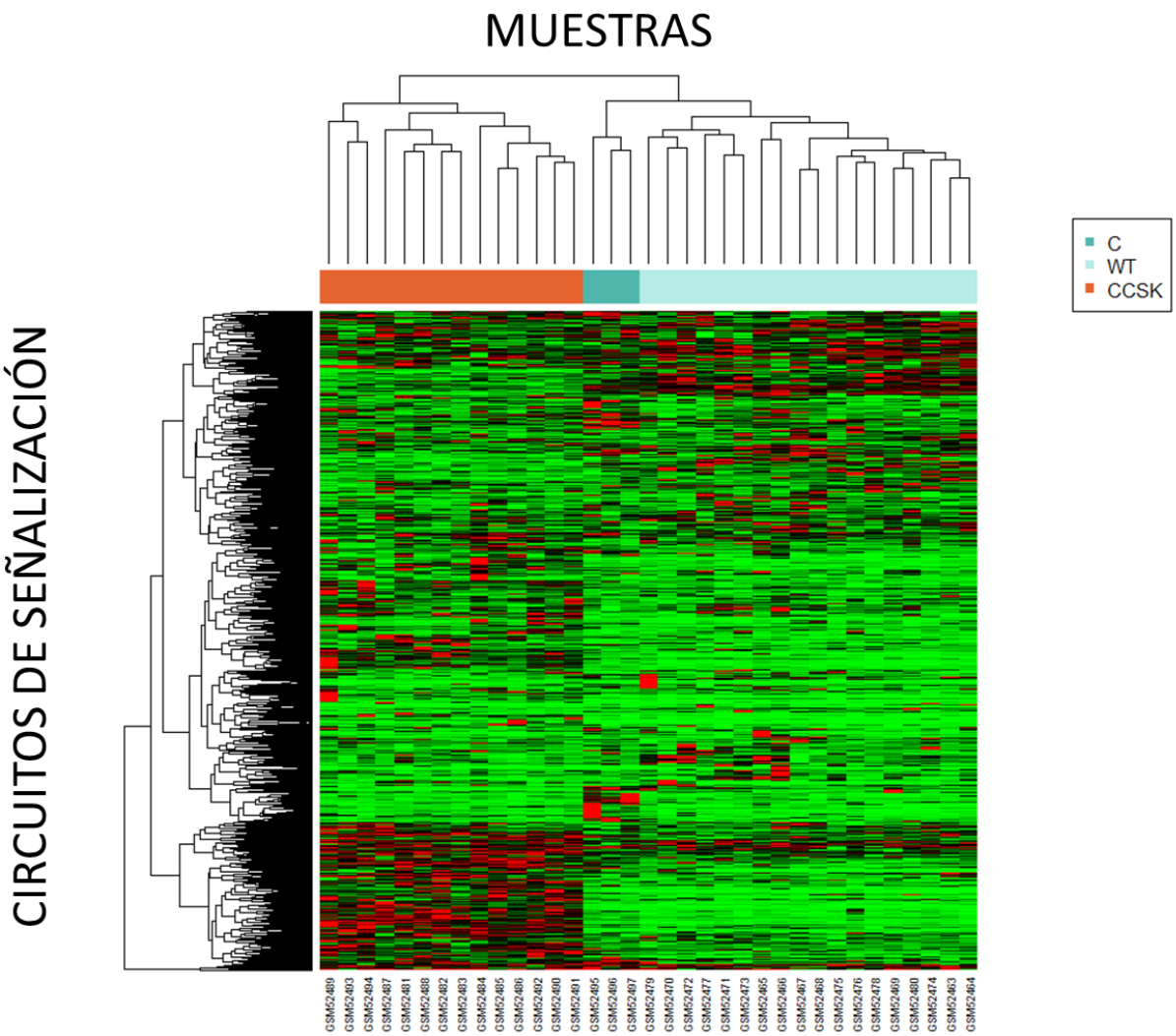

Figura 3. Heatmap de la intensidad de activación de los circuitos de señalización calculada mediante el algoritmo hipathia. Las muestras pertenecen a tres grupos: CCSK (Sarcoma de células Claras del Rinón) WT (Tumor de Wilms) y C (Controles de Riñón Fetal). El color representa la intensidad de activación de cada circuito en cada muestra; en rojo, circuitos muy activados, en verde, circuitos poco activados.

```
pca_plot_multiple<-multiple_pca_plot(pca_model,sample_group,cex=3,plot
_variance = TRUE)
```

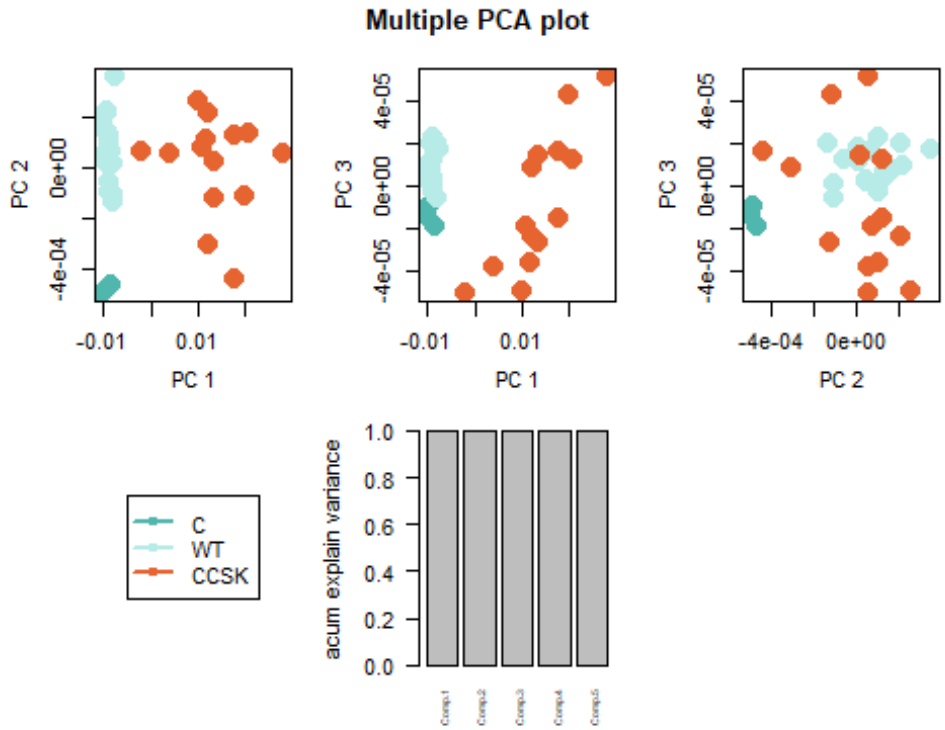

Figura 4. Análisis de Componentes Principales de la actividad de las rutas de señalización en los tres grupos de muestras. Las muestras pertenecen a tres grupos: CCSK (Sarcoma de células Claras del Riñón) WT (Tumor de Wilms) y C (Controles de Riñón Fetal).

6.8 Anotación funcional con *GeneOntology* y *UniProt*

Vamos a realizar una anotación funcional de los circuitos para conocer su funcionalidad utilizando la información de GeneOntology y de UniProt respectivamente.

Distintos circuitos efectores pueden terminar en la misma proteína efectora. A su vez, distintas proteínas efectoras pueden producir una misma función molecular. Para dar más significado biológico a los resultados de activación de las rutas, se hace una anotación funcional utilizando como ontología bien “Gene Ontology” o “UniProt”.

```
go_vals <- quantify_terms(results, pathways, dbannot = "GO")
```

Con ***GeneOntology*** se anotan 1654 funciones.

```
uniprot_vals <- quantify_terms(results, pathways, dbannot = "uniprot")
```

Con ***UniProt*** se anotan 142 funciones.

6.9 Visualización de la activación/inactivación de las funciones anotadas por GO y UniProt de los tres grupos de pacientes.

El análisis visual de la actividad de las funciones anotadas nos da un resultado similar al que nos daba el análisis de la activación de las rutas. Visualizando los niveles de actividad de las funciones con ambas anotaciones (UniProt y GO), vemos que en los dendogramas las muestras de cada grupo están próximas entre sí, y que las de TW se parecen más a los controles que a las de SCCR. Vemos que con ambas ontologías, GO (**Figura 5**) y UniProt (**Figura 6**), la distribución de la funcionalidad entre los grupos de muestras, mantiene una concordancia entre grupos similar a la obtenida con las rutas de señalización.

```
heatmap_plot(go_vals,group=sample_group,colors="classic",variable_clust = TRUE)
```

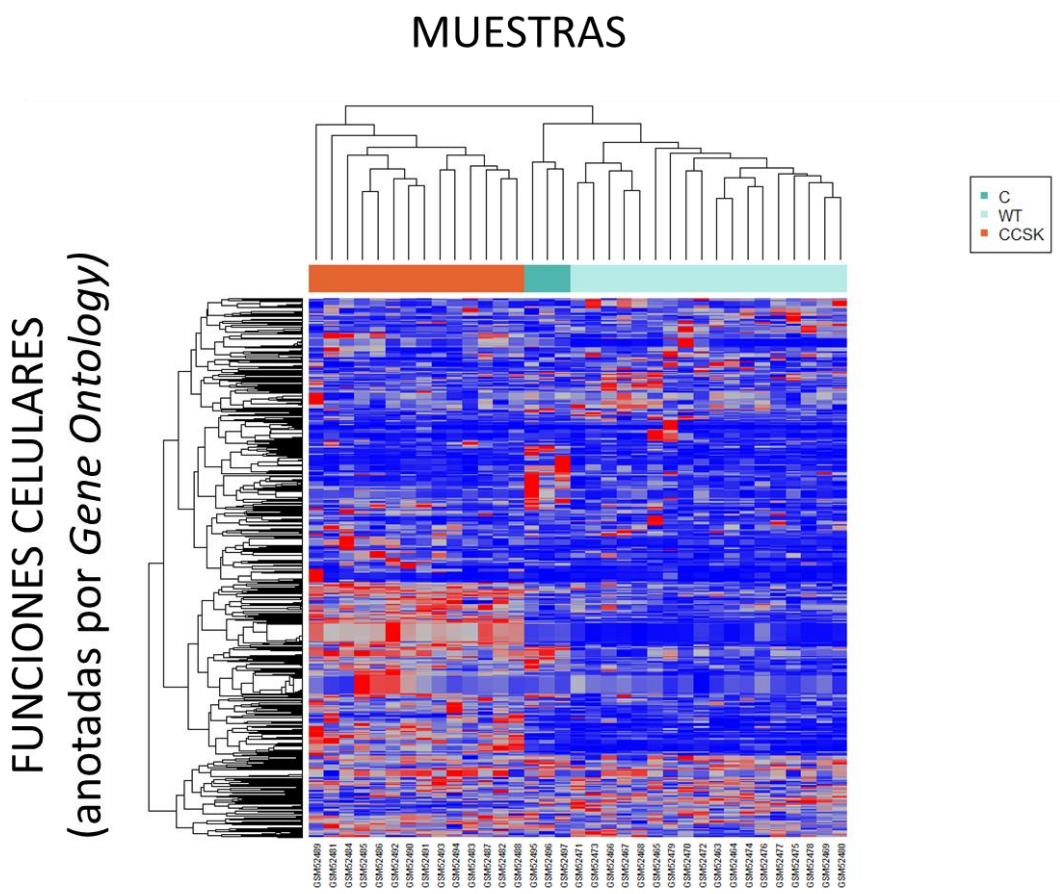

Figura 5. Heatmap en el que se representan las funciones celulares anotadas por GeneOntology. Las muestras pertenecen a tres grupos: CCSK (Sarcoma de células Claras del Rinón) WT (Tumor de Wilms) y C (Controles de Riñón Fetal). El color representa la intensidad de activación de cada función en cada muestra; en rojo, funcioness muy activados, en azul funciones poco activadas.

```
heatmap_plot(uniprot_vals,group=sample_group,colors="classic",variable
_clust = TRUE)
```

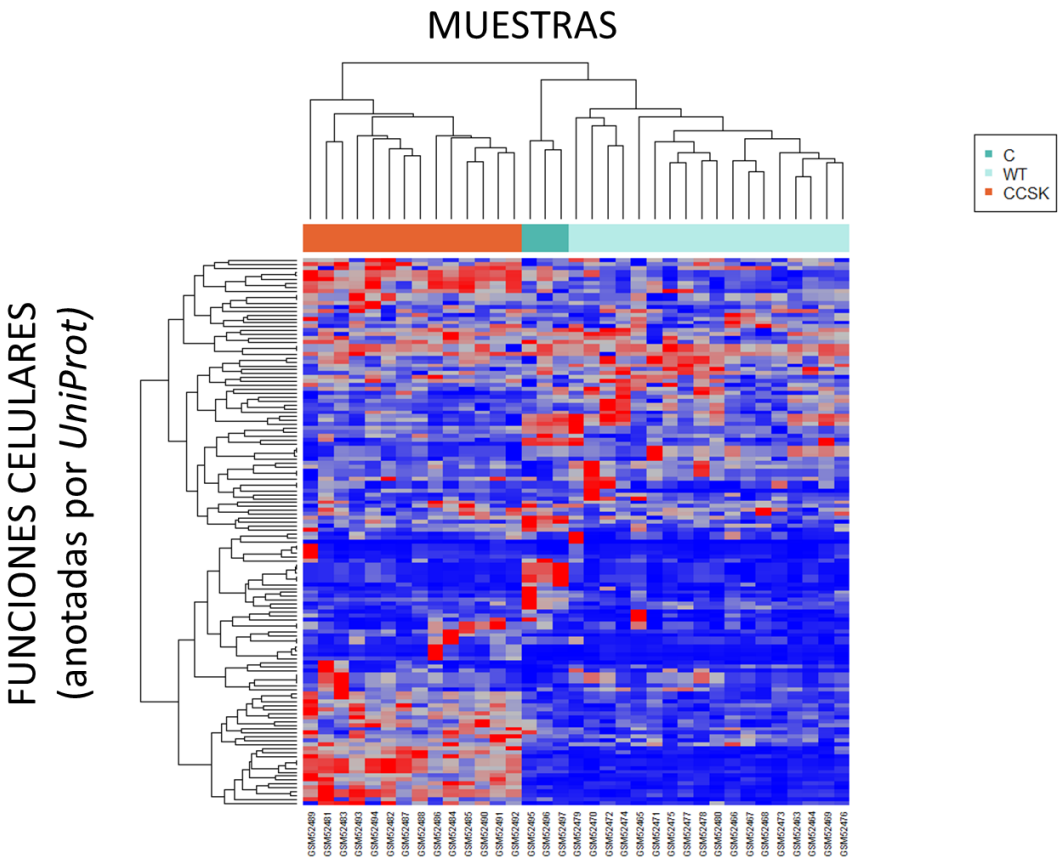

Figura 6. Heatmap en el que se representan las funciones celulares anotadas por UniProt. Las muestras pertenecen a tres grupos: CCSK (Sarcoma de células Claras del Rinón) WT (Tumor de Wilms) y C (Controles de Riñón Fetal). El color representa la intensidad de activación de cada función en cada muestra; en rojo, funciones muy activadas, en azul funciones poco activadas.

6.10 Comparativa del nivel de activación de las rutas de señalización entre "Clear Cell Sarcoma Kidney" vs "Kidney Control"

Una vez que tenemos valores con la activación de cada circuito en cada muestra, y valores de la activación de cada función en cada muestra, podemos hacer comparaciones, agrupando las muestras por características comunes. En nuestro caso vamos a agrupar las muestras por estatus de enfermedad (columna "group", para separar en CCSK, WT y C), analizado la significación estadística de las diferencias mediante un test de Wilcoxon.

Creamos el objeto “sample_group” de clase “factor”, para introducir la información de los grupos que queremos analizar:

```
sample_group<-exp_design_gds1282[colnames(path_vals),"group"]
class(sample_group)
```

Comparamos por Wilconson, corregido por el método Benjamin-Hochberg FDR, la diferencia de activación de los circuitos entre C y CCSK:

```
comp_paths<-do_wilcoxon(path_vals,sample_group,g1="CCSK",g2="C")
```

Obtenemos una matriz con las diferencias de activación de 1876 circuitos del grupo de pacientes de CCSK con respecto al grupo Control, y la significación estadística de dicha modificación, sin corregir y corregida:

```
hhead(comp_paths,20)

##                UP/DOWN  statistic     p.value FDRp.value
## P-hsa03320-37    DOWN -2.6457513 0.002941176 0.02006417
## P-hsa03320-61    DOWN -2.5197632 0.005882353 0.03245675
## P-hsa03320-46    DOWN -1.3858697 0.197058824 0.38873013
## P-hsa03320-57    DOWN -1.8898224 0.067647059 0.18773060
## P-hsa03320-64    DOWN -2.5197632 0.005882353 0.03245675
## P-hsa03320-47    DOWN -2.5197632 0.005882353 0.03245675
## P-hsa03320-65    DOWN -2.6457513 0.002941176 0.02006417
## P-hsa03320-55    DOWN -2.1417987 0.032352941 0.10935877
## P-hsa03320-56    DOWN -1.3858697 0.197058824 0.38873013
## P-hsa03320-33    DOWN -0.5039526 0.676470588 0.88129085
## P-hsa03320-58    DOWN -2.5197632 0.005882353 0.03245675
## P-hsa03320-59    DOWN -2.0158105 0.047058824 0.14520124
## P-hsa03320-63    DOWN -2.3937750 0.011764706 0.05383070
## P-hsa03320-44    DOWN -2.6457513 0.002941176 0.02006417
## P-hsa03320-36    DOWN -2.6457513 0.002941176 0.02006417
## P-hsa03320-30    DOWN -0.6299408 0.591176471 0.80893294
## P-hsa03320-28    DOWN -1.7638342 0.091176471 0.23145745
## P-hsa03320-25    DOWN -2.5197632 0.005882353 0.03245675
## P-hsa03320-21    DOWN -2.3937750 0.011764706 0.05383070
## P-hsa03320-20    DOWN -1.3858697 0.197058824 0.38873013
```

Para conocer cuantas rutas tienen cambios significativos hacemos una selección de la matriz “comp_paths” de las rutas con cambios (tanto up como down) significativos (FDRp.value<0.05),

```
 comp_paths_sig<-subset(comp_paths, FDRp.value < 0.05)
```

Obtenemos que, del total de 1876 circuitos analizados, 340 circuitos están significativamente alterados en CCSK con respecto a los controles. 212 circuitos están upregulados y 128 downregulados:

```
> head(comp_paths_sig,20)
              UP/DOWN statistic     p.value FDRp.value
P-hsa03320-37    DOWN -2.645751 0.002941176 0.02006417
P-hsa03320-61    DOWN -2.519763 0.005882353 0.03245675
P-hsa03320-64    DOWN -2.519763 0.005882353 0.03245675
P-hsa03320-47    DOWN -2.519763 0.005882353 0.03245675
P-hsa03320-65    DOWN -2.645751 0.002941176 0.02006417
P-hsa03320-58    DOWN -2.519763 0.005882353 0.03245675
P-hsa03320-44    DOWN -2.645751 0.002941176 0.02006417
P-hsa03320-36    DOWN -2.645751 0.002941176 0.02006417
P-hsa03320-25    DOWN -2.519763 0.005882353 0.03245675
P-hsa03320-27      UP  2.645751 0.002941176 0.02006417
P-hsa03320-26    DOWN -2.645751 0.002941176 0.02006417
P-hsa03320-22    DOWN -2.645751 0.002941176 0.02006417
P-hsa03320-60    DOWN -2.645751 0.002941176 0.02006417
P-hsa03320-31    DOWN -2.645751 0.002941176 0.02006417
P-hsa03320-10    DOWN -2.645751 0.002941176 0.02006417
P-hsa03320-40    DOWN -2.645751 0.002941176 0.02006417
P-hsa03320-53    DOWN -2.645751 0.002941176 0.02006417
P-hsa03320-38    DOWN -2.645751 0.002941176 0.02006417
P-hsa03320-9     DOWN -2.645751 0.002941176 0.02006417
P-hsa03320-7     DOWN -2.645751 0.002941176 0.02006417
```

Ordenamos las filas por orden decreciente de significación (FDFp.value creciente), y vemos los 20 primeras rutas más significativas:

```
comp_paths_ordered<-comp_paths[order(comp_paths$FDRp.value,decreasing=
FALSE),]
```

Vemos que los valores de cambio y de significación son iguales en estos 20 primeros circuitos:

```
head(comp_paths_ordered,20)

##                  UP/DOWN statistic     p.value FDRp.value
## P-hsa03320-37       DOWN -2.645751 0.002941176 0.02006417
## P-hsa03320-65       DOWN -2.645751 0.002941176 0.02006417
## P-hsa03320-44       DOWN -2.645751 0.002941176 0.02006417
## P-hsa03320-36       DOWN -2.645751 0.002941176 0.02006417
```

```
## P-hsa03320-27          UP  2.645751 0.002941176 0.02006417
## P-hsa03320-26        DOWN -2.645751 0.002941176 0.02006417
## P-hsa03320-22        DOWN -2.645751 0.002941176 0.02006417
## P-hsa03320-60        DOWN -2.645751 0.002941176 0.02006417
## P-hsa03320-31        DOWN -2.645751 0.002941176 0.02006417
## P-hsa03320-10        DOWN -2.645751 0.002941176 0.02006417
## P-hsa03320-40        DOWN -2.645751 0.002941176 0.02006417
## P-hsa03320-53        DOWN -2.645751 0.002941176 0.02006417
## P-hsa03320-38        DOWN -2.645751 0.002941176 0.02006417
## P-hsa03320-9         DOWN -2.645751 0.002941176 0.02006417
## P-hsa03320-7         DOWN -2.645751 0.002941176 0.02006417
## P-hsa03320-8         DOWN -2.645751 0.002941176 0.02006417
## P-hsa03460-16 49       UP  2.645751 0.002941176 0.02006417
## P-hsa03460-16 34       UP  2.645751 0.002941176 0.02006417
## P-hsa04012-24          UP  2.645751 0.002941176 0.02006417
## P-hsa04012-57          UP  2.645751 0.002941176 0.02006417
```

La representación en Heatmap de la selección de circuitos significativos, y de los significativos upregulados y downregulados, nos permite ver la distribución de la actividad de los circuitos con más detalle, y los cambios en la distribución de las muestras en el dendograma de grupos.

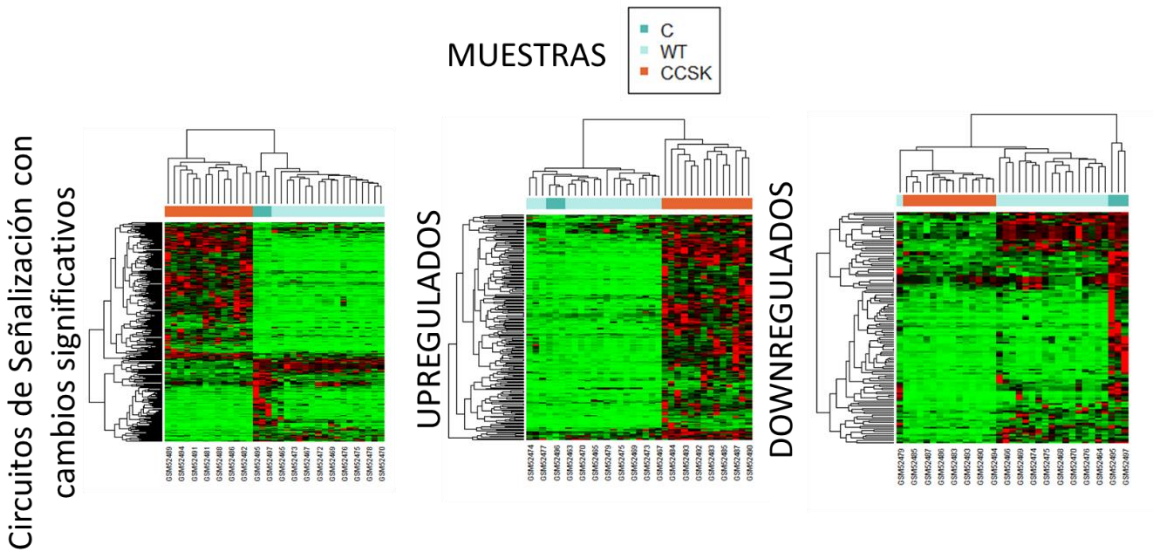

Figura 7. Heatmaps en los que se representan la activación de los circuitos significativos, de los significativos upregulados y de los significativos downregulados. Las muestras pertenecen a tres grupos: CCSK (Sarcoma de células Claras del Rinón) WT (Tumor de Wilms) y C (Controles de Riñón Fetal). El color representa la intensidad de activación de cada circuito en cada muestra; en rojo, circuitos muy activados, en verde, circuitos poco activados.

Representando estos resultados con un vulcano plot (**Figura 7**), vemos que la imagen es diferente a la que se suele obtener cuando representamos genes individuales. Los valores de activación que se obtienen aparecen agrupados en

grupos de valores discretos. Nónese que cada ponto que se ve en la gráfica, es en realidad un grupo de puntos superpuestos.

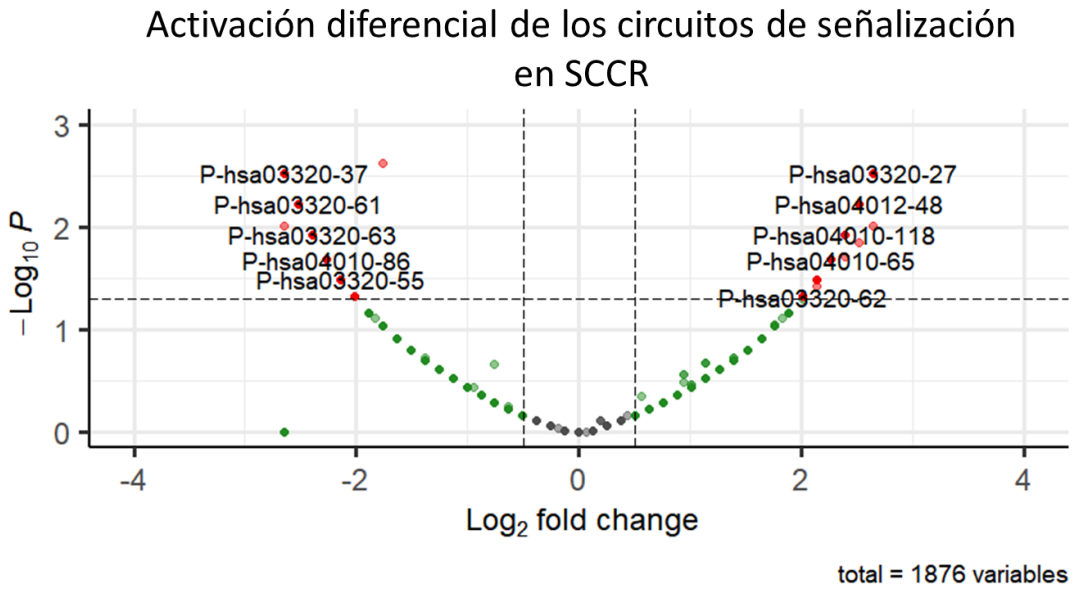

Figura 8. VolcanoPlot de la actividad de los circuitos de señalización en SCCR. Gris: circuitos sin cambios (con Fold Change<0.5) . Verde: Circuitos con cambios no significativos (p.valor>0.05). Rojo: Circuitos con cambios significativos. Izquierda: Reprimidos. Derecha: Activados. Los circuitos aparecen sobreposicionados, las etiquetas muestran una pequeña parte del total.

La función “get_pathways_summary” nos da un resumen de los resultados de la función “hipathia”. Nos resume el porcentaje de circuitos activados en cada ruta de señalización. (**Tabla 1**). La ruta de señalización con mayor porcentaje de circuitos activados es la ruta de *Glioma* (hsa05214). Esta ruta está compuesta por seis circuitos, y en las muestras de SCCR, todos ellos están upregulados. La ruta de señalización con mayor porcentaje de circuitos reprimidos es la *ruta de reabsorción de calcio regulada de modo endocrino y por otros factores* (hsa04961). Esta ruta también está formada por seis circuitos, de los que cuatro están reprimidos, uno está activado y otro no tiene cambios significativos.

Tabla 1. Rutas de señalización con mayor porcentaje de circuitos modificados significativamente en muestras de CCSK. (Listado completo en Anexo I)

		CIRCUITOS						
		Nº Absolutos				Porcentajes		
RUTAS DE SEÑALIZACIÓN	ID Rutas	Total	SIG	UP	DOWN	SIG	UP	DOWN
Glioma	hsa05214	6	6	6	0	100	100	0
FoxO signaling pathway	hsa04068	30	25	25	0	83,33	83,33	0
Endocrine and other factor-regulated calcium reabsorption	hsa04961	6	5	1	4	83,33	16,67	66,67
Basal cell carcinoma	hsa05217	6	5	5	0	83,33	83,33	0
Bacterial invasion of epithelial cells	hsa05100	4	3	0	3	75	0	75
PPAR signaling pathway	hsa03320	42	22	1	21	52,38	2,38	50
Calcium signaling pathway	hsa04020	10	5	2	3	50	20	30
VEGF signaling pathway	hsa04370	10	5	0	5	50	0	50
Vasopressin-regulated water reabsorption	hsa04962	2	1	1	0	50	50	0
Carbohydrate digestion and absorption	hsa04973	2	1	1	0	50	50	0
Endometrial cancer	hsa05213	6	3	2	1	50	33,33	16,67
Melanogenesis	hsa04916	7	3	1	2	42,86	14,29	28,57
Neurotrophin signaling pathway	hsa04722	12	5	4	1	41,67	33,33	8,33
Adrenergic signaling in cardiomyocytes	hsa04261	17	7	7	0	41,18	41,18	0
Colorectal cancer	hsa05210	10	4	4	0	40	40	0
Thyroid cancer	hsa05216	5	2	1	1	40	20	20
ErbB signaling pathway	hsa04012	18	7	6	1	38,89	33,33	5,56
Acute myeloid leukemia	hsa05221	13	5	5	0	38,46	38,46	0
Proteoglycans in cancer	hsa05205	47	18	9	9	38,3	19,15	19,15
Rap1 signaling pathway	hsa04015	14	5	4	1	35,71	28,57	7,14

6.11 Visualización de las rutas

Como ya habíamos visto anteriormente, los circuitos de la ruta hsa05214/Glioma están sobreexpresados y aparecen en el grafo en color rojo (**Figura 8**). Por su parte los circuitos de hsa04961/reabsorción de calcio están reprimidos y aparecen en azul (**Figura 9**).

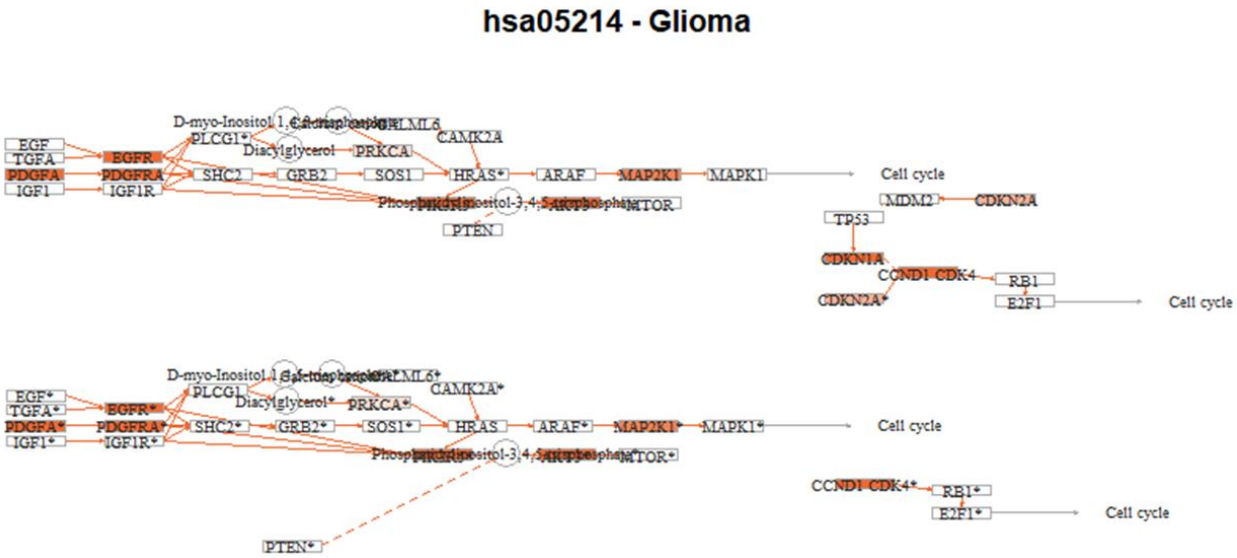

Figura 9. Grafo que representa la ruta de señalización hsa 05214/Glioma. En rojo aparecen los circuitos y los genes activados en CCSK.

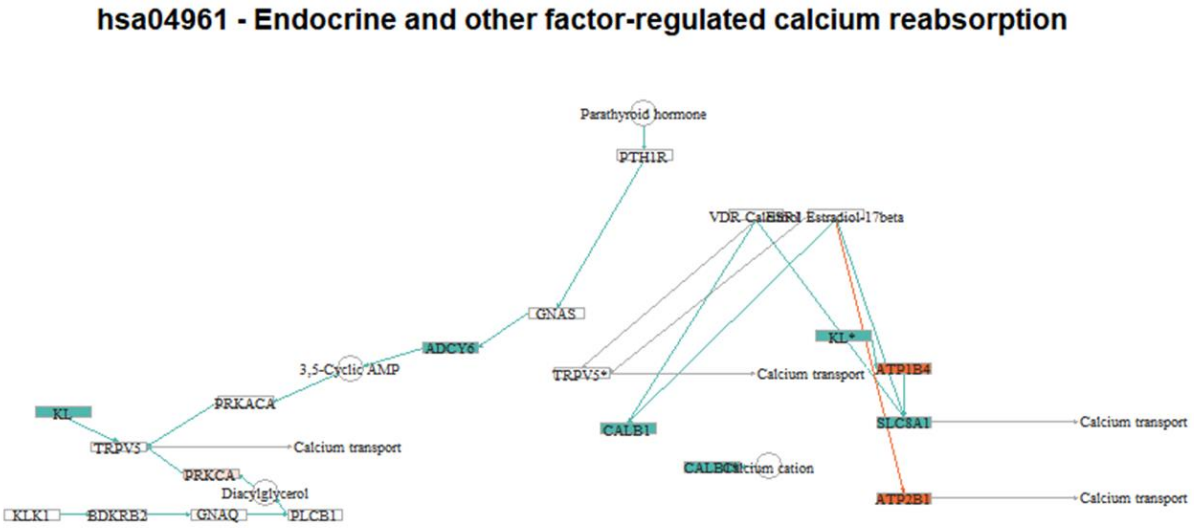

Figura 10. Grafo que representa la ruta de señalización hsa 04961/Reabsorción de calcio. En azul aparecen los circuitos y los genes desregulados en CCSK.

6.12 Ruta de señalización PI3K/Akt

Vamos a representar la ruta PI3K/Akt, que es una de las rutas con más genes sobreexpresados, para comparar este hecho con el resultado del análisis mecanístico. En la representación gráfica de esta ruta (**Figura 10**) comprobamos que, aparecen muchos de los nodos en rojo (sobreexpresados). En cambio, cuando analizamos dicha expresión desde un punto de vista mecanístico con Hipatia, comprobamos que esa expresión no se traduce en un aumento de actividad de la ruta. Analizando los resultados de la ruta (**Ver anexo I**),

comprobamos que la ruta del PI3K-Atk está formada por 28 circuitos de entre los cuales hay uno upregulado, dos downregulados y el resto no tiene cambios significativos

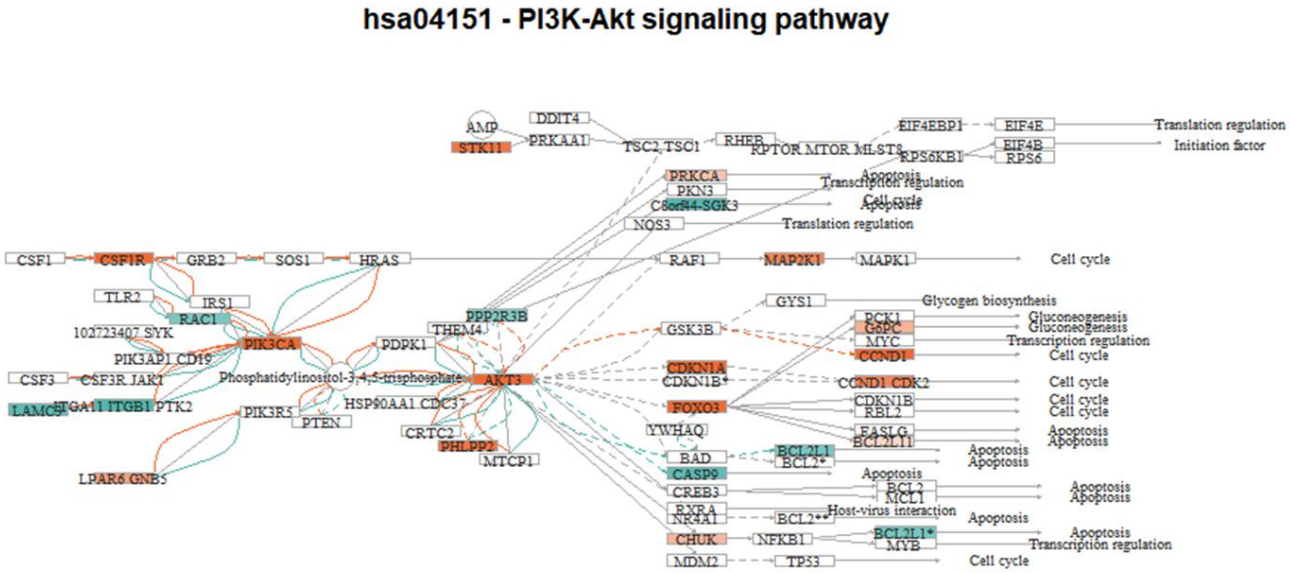

Figura 11. Grafo que representa la ruta de señalización hsa 04151/PI3K-Akt en las muestras de CCSK.

6.13 Reposicionamiento de fármacos

A partir del listado de circuitos de señalización significativos para SCCR, el grupo del Área de Bioinformática Clínica del Hospital Virgen del Rocío de Sevilla (en concreto el Dr Carlos Loucera y Marina Esteban-Medina) realizaró una modelización de Machine Learning con el paquete drexml de Python. (script del anexo IV). Este modelo infiere posibles efectos regulatorios de dianas génicas de fármacos sobre la actividad del mapa de señalización asociado a la enfermedad. Obtenemos el resultado en tres archivos: (1) Circuito X Gen: con puntuaciones de relevancia, (2) Circuito X Gen: con un operados booleano, que nos indica si el gen ha sido elegido o no para ese circuito, y (3) Circuito x medidas: medidas de calidad que nos indican cómo de estable y preciso es el modelo para cada circuito.

La matriz que relaciona los KDTs significativos para esta enfermedad. y sus scores de relevancia sobre cada circuito. "shap_relevant_stable_matrix", tiene

una dimensión de 186 circuitos por 116 KDTs (Dianas Génicas de Fármacos, del inglés *Known Drug Targets*):

```
> shap_relevant_stable_matrix[1:10,1:6]

# A tibble: 20 × 6

   circuit_name                          ACP3 ACTN1    ADH1B    ADRB2    AGTR1

   <chr>                                <dbl> <dbl>    <dbl>    <dbl>    <dbl>

 1 PPAR signaling pathway: HMGCS2   0.000559      0 0        0        0

 2 PPAR signaling pathway: APOA1   -0.000280      0 0        0        0

 3 PPAR signaling pathway: APOA5   -0.0000775     0 0        0        0

 4 PPAR signaling pathway: PLTP     0             0 0        0        0

 5 PPAR signaling pathway: ME1      0             0 0        0        0

 6 PPAR signaling pathway: CYP8B1   0             0 0.000162 0        0

 7 PPAR signaling pathway: FABP1    0             0 0        0        0

 8 PPAR signaling pathway: SLC27A4  0             0 0        0        0

 9 PPAR signaling pathway: EHHADH   0.000127      0 0        0        0

10 PPAR signaling pathway: CPT1C    0             0 0        0        0
```

Podemos obtener de esta matriz el listado de KDTs para SCCR:

```
names(shap_relevant_stable_matrix)

  [1] "circuit_name" "ACP3"         "ACTN1"        "ADH1B"        "ADRB2"        "AGTR1"

  [7] "SLC25A4"      "ANXA1"        "APOC3"        "AR"           "ARL2"         "AVPR1A"

 [13] "AXL"          "C1R"          "C1S"          "CA4"          "CA12"         "CACNA1B"

 [19] "CACNA1C"      "CACNA1D"      "CACNA1S"      "CACNA2D1"     "CACNB1"       "CACNG1"

 [25] "CAT"          "CFTR"         "CHRNB2"       "CKM"          "CKMT2"        "COL3A1"

 [31] "CSF3R"        "S1PR1"        "EDNRA"        "EGFR"         "EPHA2"        "ERBB2"

 [37] "PTK2B"        "FCER1G"       "FCGR2A"       "FCGR3B"       "FGFR1"        "FGFR3"

 [43] "GABRA1"       "GABRB3"       "GABRG2"       "GHR"          "GRIN1"        "HBB"

 [49] "IL1R1"        "IL2RG"        "ITGAX"        "ITGB1"        "ITGB7"        "KCNA7"

 [55] "KCNC1"        "KCNH2"        "KCNJ8"        "KCNJ11"       "KCNQ1"        "LPL"

 [61] "TACSTD2"      "MAP1A"        "MAP2"         "MAP4"         "MAPT"         "MB"

 [67] "MAP3K1"       "NFKB1"        "NPR2"         "DDR2"         "PDE3A"        "PDGFRA"

 [73] "PIK3CD"       "PPARG"        "PTGER3"       "PTH1R"        "RARG"         "RYR1"

 [79] "SCN1B"        "SCN2A"        "SCN4A"        "SCN5A"        "SCN8A"        "SCNN1A"

 [85] "SCNN1B"       "SCNN1G"       "SELP"         "SLC6A1"       "SLC8A1"       "SNAP25"

 [91] "SRD5A1"       "ELOVL4"       "SYK"          "TFPI"         "TGFBR2"       "THRB"

 [97] "TNNC1"        "VWF"          "XDH"          "AOC3"         "SLC22A6"      "SV2A"
```

```
[103] "ABCC9"        "RAMP2"        "RAMP1"        "RAMP3"        "TUBB3"        "HRH3"
[109] "NTSR2"        "CALY"         "ANO1"         "SLC12A5"      "NOD2"         "DGAT2"
[115] "VKORC1L1"     "CKMT1A"
```

Dependiendo de si un KDT actúa sobre un circuito activado o reprimido, y de si su efecto sobre el circuito es positivo o negativo, el fármaco requerido para actuar sobre ese KDT podrá ser de activación o de inhibición.

Separamos los circuitos activados y reprimidos en dos submatrices procedentes de la anterior:

```
shap_relevant_stable_matrix_up<-
shap_relevant_stable_matrix|>filter(circuit_name%in%path_names_up)
```

Seleccionando los circuitos upregulados, obtenemos una matriz de 96 circuitos y 116 KDTs:

```
> dim(shap_relevant_stable_matrix_up)

[1]  96 116

> shap_relevant_stable_matrix_up[1:10,1:6]

# A tibble: 10 × 6

   circuit_name                          ACP3 ACTN1 ADH1B ADRB2 AGTR1

   <chr>                                <dbl> <dbl> <dbl> <dbl> <dbl>

 1 PPAR signaling pathway: CPT1C   0              0     0     0     0

 2 ErbB signaling pathway: CDKN1A  0              0     0     0     0

 3 ErbB signaling pathway: ABL1    0              0     0     0     0

 4 ErbB signaling pathway: CAMK2A  0              0     0     0     0

 5 ErbB signaling pathway: PRKCA   0              0     0     0     0

 6 ErbB signaling pathway: STAT5A* 0              0     0     0     0

 7 ErbB signaling pathway: ELK1*   0.00000658     0     0     0     0

 8 Ras signaling pathway: NFKB1    0              0     0     0     0

 9 Rap1 signaling pathway: ITGA2B  0              0     0     0     0

10 Rap1 signaling pathway: MAPK14  0              0     0     0     0
```

Obtenemos la submatriz de los circuitos downregulados:

```
shap_relevant_stable_matrix_down<-
shap_relevant_stable_matrix|>filter(circuit_name%in%path_names_down)
```

Sus dimensiones son de 86 circuitos por 116 KDTs:

```
> dim(shap_relevant_stable_matrix_down)

[1]  84 116

> shap_relevant_stable_matrix_down[1:10,1:6]

# A tibble: 10 × 6

   circuit_name                          ACP3 ACTN1    ADH1B ADRB2 AGTR1

   <chr>                                <dbl> <dbl>    <dbl> <dbl> <dbl>

 1 PPAR signaling pathway: HMGCS2   0.000559      0 0            0     0

 2 PPAR signaling pathway: APOA1   -0.000280      0 0            0     0

 3 PPAR signaling pathway: APOA5   -0.0000775     0 0            0     0

 4 PPAR signaling pathway: PLTP     0             0 0            0     0

 5 PPAR signaling pathway: ME1      0             0 0            0     0

 6 PPAR signaling pathway: CYP8B1   0             0 0.000162     0     0

 7 PPAR signaling pathway: FABP1    0             0 0            0     0

 8 PPAR signaling pathway: SLC27A4  0             0 0            0     0

 9 PPAR signaling pathway: EHHADH   0.000127      0 0            0     0

10 PPAR signaling pathway: CPT2     0             0 0            0     0
```

Dentro de cada matriz, tenemos el efecto de los KDTs sobre los circuitos.

En la matriz de circuitos activados, los KDTs con valor positivo requieren de fármacos inhibidores para contrarrestar el efecto del circuito. Y al contrario los KDTs con valor negativo sobre el circuito requieren de fármacos activadores para producir un descenso de actividad.

En la matriz de circuitos reprimidos, los KDTs con valor positivo requieren de activadores y los de valor negativo requieren de inhibidores, para producir en ambos casos un aumento de actividad del circuito reprimido.

7 Discusión

Mediante el reanálisis de un dataset público de Sarcoma de Células Claras de Riñón (CCSK) hemos obtenido resultados equivalentes a los ya publicados en los que vemos que los patrones de expresión génica de muestras de Tumor de Wilms y muestras controles de Riñón Fetal se parecen más entre sí que a los de las muestras de CCSK.

El perfil de expresión génico de las muestras puede reflejar el origen, desarrollo y patogeneidad del tumor. Esta diferencia de las muestras de CCSK con los otros dos grupos podría ser un reflejo del peor pronóstico que sufren estos pacientes pediátricos comparado con el de otros tumores renales.

Cuando representamos en Heatmap sólo los circuitos significativos, y los significativos upregulados y downregulados por separado, vemos que en los circuitos downregulados hay menos diferencia de expresión entregrupos, y los grupos no están tan definido. El perfil de los circuitos significativos downregulados es más similar entre SCCR y WT que con las muestras control. Esto podría ser indicativo de que esa downregulación puede ser derivada de la patogeneidad, pero no sería específica del mal pronóstico del SCCR. Por otro lado, no hay que olvidar que las muestras control son de origen fetal. Es posible que estos circuitos, en realidad, no estén downregulados en las muestras de pacientes, si no que sean circuitos upregulados en las muestras control. Podría tratarse de circuitos activados transitoriamente en el periodo fetal, específicos del desarrollo embrionario.

Por lo tanto, una limitación de este estudio es el tipo de muestra control que se ha utilizado. Es evidente la dificultad de obtener muestras control de individuos sanos de la misma edad, pero el que las muestras control sean fetales, añade variables no controlables que dificultan la interpretación de los resultados.

En el estudio que publicó este dataset originalmente se determinó que las muestras de CCSK tenían diferencialmente expresados genes de 4 categorías: (a) marcadores neurales, (b) miembros de la ruta Sonic hedgehog (c) miembros de la ruta de proliferación celular fosfoinositol-3-kinasa/Akt, y (d) dianas terapeúticas conocidas (Cutcliffe, y otros, 2005).

El artículo original en el que se publicó el estudio del dataset GDS1282 nombra la vía de señalización Akt como una de las más upreguladas.

El análisis que obtenemos con Hipatia coincide con los resultados publicados en cuanto a la expresión alta de genes en la ruta, encambio la funcionaliadd no correlaciona.

Mediante nuestro re análisis hemos podido detectar una ruta que, si bien sus genes se encuentran sobre expresados con ambos métodos, con el análisis mecaníastico, además hemos podido comprobar que su funcionalidad podría no ser tan alta como se consideraba.

Podemos comprobar como en este caso el enfoque mecanístico de hipatia añande información adicional que no podemos encontrar fuera de dicho enfoque.

Con los datos disponibles se podría repetir para el Tumor de Wilms todo el análisis realizado para el Sarcoma de Células Claras del Riñón. En su comparación con las muestras de riñón fetal, previsiblemente encontraríamos menos circuitos y menos funciones diferencialmente activados en WT, de los que aparecen en SCCR, ya que en los dendrogamas realizados, TW y C aparecen más cercanos entre sí que con SCCR.

Estos hallazgos podrían ser base para reorientar la búsqueda de dianas terapeúticas y de posibles fármacos para el tratamiento de CCSK.

Los circuitos de señalización celular con cambios significativos de actividad en las muestras de CSSK, nos ofrece un mapa de la enfermedad con enfoque mecanístico. El modelado de dicho mapa mediante ML nos permite obtener un listado de KDTs que pueden ser partida para seleccionar fármacos comercializados para uso humano, que hayan sido aprobados para otras indicaciones y que podrían ser candidatos a ser incluidos en estudios más específicos que valorasen su utilidad real en esta enfermedad.

El reposicionamiento de fármacos supone una ventaja para estudiar esta enfermedad, ya que al tratarse de una enfermedad rara, la investigación por parte de la industria farmaceútica es más escasa que para otras enfermedades más prevalentes.

8 Conclusiones

- Podemos reutilizar un dataset de muestras de pacientes de SCCR depositado en un repositorio público, para realizar un reanálisis de la expresión génica con enfoque mecanístico, que nos determine la actividad de los circuitos delulares de señalización en dicha enfermedad.
- En pacientes de SCCR, la ruta de señalización celular PI3K/Atk tiene una alta proporción de genes sobreexpresados, pero en cambio un análisis mecanístico nos demuestra que la mayor parte de sus circuitos no se encuentran activados.
- El mapa de la enfermedad de SCCR obtenido con el listado de circuitos activados, resultado del análisis mecanístico, se puede utilizar como punto de partida para un modelado de ML (MORF) que nos selecciona los KDTs relevantes en la enfermedad para encontrar los fármacos ya comercializados para uso humano, candidatos a ser reposicionados para el tratamiento de SCCR.

9 Bibliografía

Amadoz, A., Hidalgo, M., C¸ubuk, C., Carbonell-Caballero, J., & Dopazo, J. (2019). A comparison of mechanistic signaling pathway. *Briefings in Bioinformatics*, 1655-1668.

Brown, A., & Patel, C. (2018). A review of validation strategies for computational. *Briefings in Bioinformatics*, 174-177.

Cha, Y., Erez, T., Reynolds, I., Kumar, D., Ross, J., Koytiger, G., . . . Laifenfeld, D. (2018). Drug repurposing from the perspective of pharmaceutical companies. *British Journal of Pharmacology*, 168-180.

Cutcliffe, C., Kersey, D., Huang , C.-C., Zeng, Y., Walterhouse, D., & J Perlman, E. (2005). Clear cell sarcoma of the kidney: up-regulation of neural markers with activation of the sonic hedgehog and Akt pathways. *Clinical Cancer Research*, 7986-7994.

Ding, J., Yao, H., & Chen, Q. (2022). A Nomogram-Based Risk Classification System Predicting the Overall Survival of Childhood with Clear Cell Sarcoma of the Kidney Based on the SEER Database. *Evid Based Complement Alternat Med*.

Dudley, J., Deshpande, T., & Butte, A. (2011). Exploiting drug^disease relationships for computational drug repositioning. *BRIEFINGS IN BIOINFORMATICS*, 303-311.

Esteban-Medina, M., Peña-Chilet, M., Loucera, C., & Dopazo, J. (2019). Exploring the druggable space around the Fanconi anemia pathwway using machine learning and mechanistic models. *BMC Bioinformatics*.

Hidalgo, M., Cubuk, C., Amadoz, A., Salavert, F., Carbonell-Caballero, J., & Dopazo, J. (2017). High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes. *Oncotarget*, 5160-5178.

Jarada, T., Rokne, J., & Alhajj, R. (2020). A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *Journal of Cheminformatics*.

Jourdan, J.-P., Bureau, R., Rochais, C., & Dallemagne, P. (2020). Drug repositioning: a brief overview. *Journal of Pharmacy and Pharmacology*, 1145-1151.

Ko, Y. (2020). Computational Drug Repositioning: Current Progress. *Applied Sciences*.

Li, J., Zheng, S., Chen, B., Butte, A., Swamidass, S., & Lu, Z. (2016). A survey of current trends in computational. *Briefings in Bioinformatics*.

López-Sánchez, M., Loucera, C., Peña-Chilet, M., & Dopazo, J. (2022). Discovering potencial interactions between rare diseases and COVID-19 by combining mechanistic models of viral infection with statistical modeling. *Human Molecular Genetics*.

Loucera, C., Esteban-Medina, M., & Peña-Chilet, M. (2022). (DRExM3L) Drug REpurposing using Mechanistic Models of signal transduction and eXplainable Machine Learning (v0.9.5). *Zenodo*.

Loucera, C., Esteban-Medina, M., Rian, K., Falco, M., Dopazo, J., & Peña-Chilet, M. (2020). Drug repurposing for COVID-19 using machine learning and mechanistic models of signal transduction cicuits related to SARS-Cov-2 infection. *Nature*.

Luo, H., Li, M., Yang, M., Wu, F.-X., Li, Y., & Wang, J. (2021). Biomedical data and computational models for drug repositioning: a comprehensive review. *Briefings in Bioinformatics*, 1604–1619.

Orpea, T. I., & Overington, J. P. (2015). Computational and Practical Aspects of Drug Repositioning. *ASSAY and Drug Development Technologies*, 299-306.

Ostaszewski, M., & et al. (2020). COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Nature-ScientificData*.

Ostaszewski, M., & et al. (2021). COVID19 Disease Map, a computational knowledge repository of virus-host interaction mechanisms. *Molecular Systems Biology*.

Rian, K., Esteban-Medina, M., Hidalgo, M., Cubuk, C., Falco, M., Loucera, C., . . . Dopazo, J. (2021). Mechanistic modeling of the SARS-CoV-2 disease map. *BioData Minig*.

Xue, H., Li, J., Xie, H., & Wang, Y. (2018). Review of Drug Repositioning Approaches and Resources. *International Journal of Biological Sciences*, 1232-1244.

Zhang, Y., Chu, Q., Ma, Y., Miao, C., & Diao, J.-J. (2022). Overall survival nomogram and relapse-related factors of clear cell sarcoma of the kidney: A study based on published patients. *Frontiers in Pediatrics*.

10 Anexo I:Rutas modificadas significativamente en CCSK

Rutas de señalización con mayor porcentaje de circuitos modificados significativamente en muestras de CCSK.

		CIRCUITOS						
		Nº Absolutos				Porcentajes		
RUTAS DE SEÑALIZACIÓN	ID Rutas	Total	SIG	UP	DOWN	SIG	UP	DOWN
Glioma	hsa05214	6	6	6	0	100	100	0
FoxO signaling pathway	hsa04068	30	25	25	0	83,33	83,33	0
Endocrine and other factor-regulated calcium reabsorption	hsa04961	6	5	1	4	83,33	16,67	66,67
Basal cell carcinoma	hsa05217	6	5	5	0	83,33	83,33	0
Bacterial invasion of epithelial cells	hsa05100	4	3	0	3	75	0	75
PPAR signaling pathway	hsa03320	42	22	1	21	52,38	2,38	50
Calcium signaling pathway	hsa04020	10	5	2	3	50	20	30
VEGF signaling pathway	hsa04370	10	5	0	5	50	0	50
Vasopressin-regulated water reabsorption	hsa04962	2	1	1	0	50	50	0
Carbohydrate digestion and absorption	hsa04973	2	1	1	0	50	50	0
Endometrial cancer	hsa05213	6	3	2	1	50	33,33	16,67
Melanogenesis	hsa04916	7	3	1	2	42,86	14,29	28,57
Neurotrophin signaling pathway	hsa04722	12	5	4	1	41,67	33,33	8,33
Adrenergic signaling in cardiomyocytes	hsa04261	17	7	7	0	41,18	41,18	0
Colorectal cancer	hsa05210	10	4	4	0	40	40	0
Thyroid cancer	hsa05216	5	2	1	1	40	20	20
ErbB signaling pathway	hsa04012	18	7	6	1	38,89	33,33	5,56
Acute myeloid leukemia	hsa05221	13	5	5	0	38,46	38,46	0
Proteoglycans in cancer	hsa05205	47	18	9	9	38,3	19,15	19,15
Rap1 signaling pathway	hsa04015	14	5	4	1	35,71	28,57	7,14
Maturity onset diabetes of the young	hsa04950	14	5	4	1	35,71	28,57	7,14
Oocyte meiosis	hsa04114	3	1	0	1	33,33	0	33,33
Wnt signaling pathway	hsa04310	12	4	4	0	33,33	33,33	0
Parkinson's disease	hsa05012	3	1	1	0	33,33	33,33	0
Morphine addiction	hsa05032	12	4	4	0	33,33	33,33	0
Pancreatic cancer	hsa05212	12	4	2	2	33,33	16,67	16,67
Choline metabolism in cancer	hsa05231	10	3	0	3	30	0	30
Fanconi anemia pathway	hsa03460	7	2	2	0	28,57	28,57	0
Signaling pathways regulating pluripotency of stem cells	hsa04550	14	4	3	1	28,57	21,43	7,14
GABAergic synapse	hsa04727	7	2	2	0	28,57	28,57	0
Prostate cancer	hsa05215	14	4	3	1	28,57	21,43	7,14
AMPK signaling pathway	hsa04152	29	8	6	2	27,59	20,69	6,9
Serotonergic synapse	hsa04726	11	3	0	3	27,27	0	27,27
Glucagon signaling pathway	hsa04922	11	3	1	2	27,27	9,09	18,18
Chagas disease (American trypanosomiasis)	hsa05142	11	3	3	0	27,27	27,27	0
Non-small cell lung cancer	hsa05223	11	3	2	1	27,27	18,18	9,09
Hedgehog signaling pathway	hsa04340	15	4	3	1	26,67	20	6,67
Phospholipase D signaling pathway	hsa04072	4	1	1	0	25	25	0
Focal adhesion	hsa04510	12	3	1	2	25	8,33	16,67
Adherens junction	hsa04520	16	4	1	3	25	6,25	18,75

Rutas de señalización con mayor porcentaje de circuitos modificados significativamente en muestras de CCSK.

		CIRCUITOS						
		Nº Absolutos				Porcentajes		
RUTAS DE SEÑALIZACIÓN	ID Rutas	Total	SIG	UP	DOWN	SIG	UP	DOWN
Circadian entrainment	hsa04713	4	1	0	1	25	0	25
Long-term depression	hsa04730	4	1	1	0	25	25	0
Renin secretion	hsa04924	8	2	1	1	25	12,5	12,5
Pathogenic Escherichia coli infection	hsa05130	4	1	0	1	25	0	25
Melanoma	hsa05218	4	1	1	0	25	25	0
Pathw ays in cancer	hsa05200	58	14	11	3	24,14	18,97	5,17
Longevity regulating pathw ay - mammal	hsa04211	13	3	2	1	23,08	15,38	7,69
cGMP-PKG signaling pathw ay	hsa04022	22	5	3	2	22,73	13,64	9,09
Thyroid hormone signaling pathw ay	hsa04919	31	7	6	1	22,58	19,35	3,23
Small cell lung cancer	hsa05222	9	2	2	0	22,22	22,22	0
Leukocyte transendothelial migration	hsa04670	14	3	0	3	21,43	0	21,43
Hepatitis B	hsa05161	28	6	4	2	21,43	14,29	7,14
cAMP signaling pathw ay	hsa04024	35	7	5	2	20	14,29	5,71
Platelet activation	hsa04611	5	1	0	1	20	0	20
Adipocytokine signaling pathw ay	hsa04920	15	3	1	2	20	6,67	13,33
Aldosterone-regulated sodium reabsorption	hsa04960	5	1	0	1	20	0	20
Vibrio cholerae infection	hsa05110	10	2	1	1	20	10	10
Bladder cancer	hsa05219	5	1	0	1	20	0	20
Epstein-Barr virus infection	hsa05169	16	3	3	0	18,75	18,75	0
Sphingolipid signaling pathw ay	hsa04071	11	2	2	0	18,18	18,18	0
Natural killer cell mediated cytotoxicity	hsa04650	11	2	2	0	18,18	18,18	0
Taste transduction	hsa04742	17	3	2	1	17,65	11,76	5,88
Fc epsilon RI signaling pathw ay	hsa04664	6	1	0	1	16,67	0	16,67
Non-alcoholic fatty liver disease (NAFLD)	hsa04932	12	2	1	1	16,67	8,33	8,33
Gastric acid secretion	hsa04971	6	1	0	1	16,67	0	16,67
Axon guidance	hsa04360	31	5	3	2	16,13	9,68	6,45
Tight junction	hsa04530	25	4	0	4	16	0	16
HTLV-I infection	hsa05166	19	3	2	1	15,79	10,53	5,26
Chemokine signaling pathw ay	hsa04062	13	2	0	2	15,38	0	15,38
Cellular senescence	hsa04218	7	1	1	0	14,29	14,29	0
Vascular smooth muscle contraction	hsa04270	7	1	0	1	14,29	0	14,29
Glutamatergic synapse	hsa04724	14	2	0	2	14,29	0	14,29
Regulation of lipolysis in adipocytes	hsa04923	7	1	1	0	14,29	14,29	0
Huntington's disease	hsa05016	7	1	1	0	14,29	14,29	0
Epithelial cell signaling in Helicobacter pylori infection	hsa05120	7	1	1	0	14,29	14,29	0
AGE-RAGE signaling pathw ay in diabetic complications	hsa04933	24	3	2	1	12,5	8,33	4,17
Alzheimer's disease	hsa05010	8	1	1	0	12,5	12,5	0
Cocaine addiction	hsa05030	8	1	1	0	12,5	12,5	0
Measles	hsa05162	16	2	1	1	12,5	6,25	6,25
Renal cell carcinoma	hsa05211	17	2	1	1	11,76	5,88	5,88

Rutas de señalización con mayor porcentaje de circuitos modificados significativamente en muestras de CCSK.

		CIRCUITOS						
		Nº Absolutos				Porcentajes		
RUTAS DE SEÑALIZACIÓN	ID Rutas	Total	SIG	UP	DOWN	SIG	UP	DOWN
B cell receptor signaling pathway	hsa04662	9	1	0	1	11,11	0	11,11
Cholinergic synapse	hsa04725	9	1	1	0	11,11	11,11	0
PI3K-Akt signaling pathway	hsa04151	28	3	1	2	10,71	3,57	7,14
Longevity regulating pathway - multiple species	hsa04213	10	1	1	0	10	10	0
T cell receptor signaling pathway	hsa04660	10	1	1	0	10	10	0
Jak-STAT signaling pathway	hsa04630	11	1	1	0	9,09	9,09	0
Retrograde endocannabinoid signaling	hsa04723	11	1	1	0	9,09	9,09	0
Oxytocin signaling pathway	hsa04921	22	2	2	0	9,09	9,09	0
Toxoplasmosis	hsa05145	11	1	1	0	9,09	9,09	0
Chronic myeloid leukemia	hsa05220	11	1	0	1	9,09	0	9,09
Fc gamma R-mediated phagocytosis	hsa04666	12	1	0	1	8,33	0	8,33
Regulation of actin cytoskeleton	hsa04810	12	1	0	1	8,33	0	8,33
Legionellosis	hsa05134	12	1	1	0	8,33	8,33	0
Tuberculosis	hsa05152	24	2	2	0	8,33	8,33	0
NOD-like receptor signaling pathway	hsa04621	13	1	0	1	7,69	0	7,69
Inflammatory mediator regulation of TRP channels	hsa04750	14	1	1	0	7,14	7,14	0
Amphetamine addiction	hsa05031	14	1	1	0	7,14	7,14	0
Inflammatory bowel disease (IBD)	hsa05321	14	1	1	0	7,14	7,14	0
Insulin signaling pathway	hsa04910	15	1	1	0	6,67	6,67	0
Ras signaling pathway	hsa04014	32	2	1	1	6,25	3,12	3,12
Prolactin signaling pathway	hsa04917	17	1	1	0	5,88	5,88	0
Insulin resistance	hsa04931	18	1	1	0	5,56	5,56	0
Influenza A	hsa05164	18	1	1	0	5,56	5,56	0
Hippo signaling pathway	hsa04390	23	1	0	1	4,35	0	4,35
NF-kappa B signaling pathway	hsa04064	31	1	0	1	3,23	0	3,23
Apoptosis	hsa04210	33	1	1	0	3,03	3,03	0
p53 signaling pathway	hsa04115	35	1	0	1	2,86	0	2,86
MAPK signaling pathway	hsa04010	28	0	0	0	0	0	0
HIF-1 signaling pathway	hsa04066	31	0	0	0	0	0	0
Cell cycle	hsa04110	7	0	0	0	0	0	0
mTOR signaling pathway	hsa04150	3	0	0	0	0	0	0
Notch signaling pathway	hsa04330	3	0	0	0	0	0	0
TGF-beta signaling pathway	hsa04350	6	0	0	0	0	0	0
Osteoclast differentiation	hsa04380	9	0	0	0	0	0	0
Gap junction	hsa04540	5	0	0	0	0	0	0
Complement and coagulation cascades	hsa04610	10	0	0	0	0	0	0
Antigen processing and presentation	hsa04612	3	0	0	0	0	0	0
Toll-like receptor signaling pathway	hsa04620	18	0	0	0	0	0	0
RIG-I-like receptor signaling pathway	hsa04622	7	0	0	0	0	0	0
Cytosolic DNA-sensing pathway	hsa04623	4	0	0	0	0	0	0

Rutas de señalización con mayor porcentaje de circuitos modificados significativamente en muestras de CCSK.

		CIRCUITOS						
		Nº Absolutos				Porcentajes		
RUTAS DE SEÑALIZACIÓN	ID Rutas	Total	SIG	UP	DOWN	SIG	UP	DOWN
TNF signaling pathway	hsa04668	14	0	0	0	0	0	0
Circadian rhythm	hsa04710	4	0	0	0	0	0	0
Long-term potentiation	hsa04720	2	0	0	0	0	0	0
Dopaminergic synapse	hsa04728	12	0	0	0	0	0	0
Olfactory transduction	hsa04740	8	0	0	0	0	0	0
Insulin secretion	hsa04911	8	0	0	0	0	0	0
GnRH signaling pathway	hsa04912	9	0	0	0	0	0	0
Ovarian steroidogenesis	hsa04913	12	0	0	0	0	0	0
Progesterone-mediated oocyte maturation	hsa04914	8	0	0	0	0	0	0
Estrogen signaling pathway	hsa04915	9	0	0	0	0	0	0
Thyroid hormone synthesis	hsa04918	2	0	0	0	0	0	0
Aldosterone synthesis and secretion	hsa04925	5	0	0	0	0	0	0
Type II diabetes mellitus	hsa04930	8	0	0	0	0	0	0
Salivary secretion	hsa04970	8	0	0	0	0	0	0
Pancreatic secretion	hsa04972	6	0	0	0	0	0	0
Bile secretion	hsa04976	8	0	0	0	0	0	0
Amyotrophic lateral sclerosis (ALS)	hsa05014	10	0	0	0	0	0	0
Prion diseases	hsa05020	10	0	0	0	0	0	0
Alcoholism	hsa05034	6	0	0	0	0	0	0
Shigellosis	hsa05131	2	0	0	0	0	0	0
Salmonella infection	hsa05132	10	0	0	0	0	0	0
Pertussis	hsa05133	9	0	0	0	0	0	0
Leishmaniasis	hsa05140	8	0	0	0	0	0	0
Staphylococcus aureus infection	hsa05150	8	0	0	0	0	0	0
Hepatitis C	hsa05160	16	0	0	0	0	0	0
Herpes simplex infection	hsa05168	13	0	0	0	0	0	0

11 Anexo II: Circuitos upregulados en SCCR

```
> path_names_up
 [1] "PPAR signaling pathway: CPT1C"
 [2] "Fanconi anemia pathway: FANCM C19orf40"
 [3] "Fanconi anemia pathway: FANCM STRA13"
 [4] "ErbB signaling pathway: CDKN1A"
 [5] "ErbB signaling pathway: ABL1"
 [6] "ErbB signaling pathway: CAMK2A"
 [7] "ErbB signaling pathway: PRKCA"
 [8] "ErbB signaling pathway: STAT5A*"
 [9] "ErbB signaling pathway: ELK1*"
[10] "Ras signaling pathway: NFKB1"
[11] "Rap1 signaling pathway: ITGA2B"
[12] "Rap1 signaling pathway: MAPK14"
[13] "Rap1 signaling pathway: AKT3"
[14] "Rap1 signaling pathway: PRKCI PARD6A PARD3"
[15] "Calcium signaling pathway: Sphingosine 1-phosphate"
[16] "Calcium signaling pathway: ADCY1"
[17] "cGMP-PKG signaling pathway: MYH7"
[18] "cGMP-PKG signaling pathway: NPPB"
[19] "cGMP-PKG signaling pathway: TRPC6*"
[20] "cAMP signaling pathway: PTCH1"
[21] "cAMP signaling pathway: F2R"
[22] "cAMP signaling pathway: AMH"
[23] "cAMP signaling pathway: RYR2"
[24] "cAMP signaling pathway: HCN4"
[25] "FoxO signaling pathway: CCNB3"
[26] "FoxO signaling pathway: CCND1"
[27] "FoxO signaling pathway: CCNG2"
[28] "FoxO signaling pathway: CDKN2D"
[29] "FoxO signaling pathway: CDKN1A"
[30] "FoxO signaling pathway: CDKN1B"
[31] "FoxO signaling pathway: RBL2"
[32] "FoxO signaling pathway: PLK4"
[33] "FoxO signaling pathway: GADD45G"
[34] "FoxO signaling pathway: FASLG"
```

```
[35] "FoxO signaling pathway: BCL2L11"
[36] "FoxO signaling pathway: TNFSF10"
[37] "FoxO signaling pathway: BCL6"
[38] "FoxO signaling pathway: ATG12"
[39] "FoxO signaling pathway: GABARAP"
[40] "FoxO signaling pathway: SOD2"
[41] "FoxO signaling pathway: GADD45G*"
[42] "FoxO signaling pathway: ATM"
[43] "FoxO signaling pathway: PCK1"
[44] "FoxO signaling pathway: G6PC"
[45] "FoxO signaling pathway: BCL6*"
[46] "FoxO signaling pathway: IL7R"
[47] "FoxO signaling pathway: S1PR1"
[48] "FoxO signaling pathway: RAG1"
[49] "FoxO signaling pathway: FBXO32"
[50] "Sphingolipid signaling pathway: BAX"
[51] "Sphingolipid signaling pathway: SMPD1"
[52] "Phospholipase D signaling pathway: Sphingosine 1-phosphate"
[53] "PI3K-Akt signaling pathway: CCND1"
[54] "AMPK signaling pathway: FOXO1"
[55] "AMPK signaling pathway: CPT1C"
[56] "AMPK signaling pathway: G6PC"
[57] "AMPK signaling pathway: FASN"
[58] "AMPK signaling pathway: SCD"
[59] "AMPK signaling pathway: CCND1"
[60] "Apoptosis: HRK"
[61] "Longevity regulating pathway - mammal: SOD2"
[62] "Longevity regulating pathway - mammal: ATG5"
[63] "Longevity regulating pathway - multiple species: CRYAB"
[64] "Cellular senescence: NFATC1"
[65] "Adrenergic signaling in cardiomyocytes: SCN1B"
[66] "Adrenergic signaling in cardiomyocytes: RYR2"
[67] "Adrenergic signaling in cardiomyocytes: ATP1B4"
[68] "Adrenergic signaling in cardiomyocytes: SLC9A1"
[69] "Adrenergic signaling in cardiomyocytes: KCNE1"
[70] "Adrenergic signaling in cardiomyocytes: ATP2B1"
[71] "Adrenergic signaling in cardiomyocytes: AKT3"
[72] "Wnt signaling pathway: NFATC1"
[73] "Wnt signaling pathway: CAMK2A"
[74] "Wnt signaling pathway: PRKCA"
```

```
 [75] "Wnt signaling pathway: CCND1"
 [76] "Hedgehog signaling pathway: PTCH1**"
 [77] "Hedgehog signaling pathway: PTCH1***"
 [78] "Hedgehog signaling pathway: CCND1*"
 [79] "Axon guidance: RASA1"
 [80] "Axon guidance: FYN"
 [81] "Axon guidance: PLXNC1"
 [82] "Focal adhesion: BIRC2"
 [83] "Adherens junction: CTNND1**"
 [84] "Signaling pathways regulating pluripotency of stem cells: SMAD2"
 [85] "Signaling pathways regulating pluripotency of stem cells: AKT3*"
 [86] "Signaling pathways regulating pluripotency of stem cells: SMAD2 SMAD4"
 [87] "Jak-STAT signaling pathway: CDKN1A"
 [88] "Natural killer cell mediated cytotoxicity: TNF"
 [89] "Natural killer cell mediated cytotoxicity: CSF2"
 [90] "T cell receptor signaling pathway: NFATC1"
 [91] "Neurotrophin signaling pathway: BAX"
 [92] "Neurotrophin signaling pathway: JUN"
 [93] "Neurotrophin signaling pathway: NGFRAP1 YWHAE"
 [94] "Neurotrophin signaling pathway: NFKB1*"
 [95] "Retrograde endocannabinoid signaling: ADCY1"
 [96] "Cholinergic synapse: BCL2"
 [97] "GABAergic synapse: ADCY1"
 [98] "GABAergic synapse: GABRA1 GPHN"
 [99] "Long-term depression: GNA13*"
[100] "Taste transduction: GRM1"
[101] "Taste transduction: PKD1L3 PKD2L1"
[102] "Inflammatory mediator regulation of TRP channels: ASIC2"
[103] "Insulin signaling pathway: FASN"
[104] "Melanogenesis: TYRP1"
[105] "Prolactin signaling pathway: CCND1**"
[106] "Thyroid hormone signaling pathway: TBC1D4"
[107] "Thyroid hormone signaling pathway: ATP1B4"
[108] "Thyroid hormone signaling pathway: CCND1"
[109] "Thyroid hormone signaling pathway: ATP2A2"
[110] "Thyroid hormone signaling pathway: WNT4"
[111] "Thyroid hormone signaling pathway: BMP4"
[112] "Adipocytokine signaling pathway: N-Acylsphingosine"
[113] "Oxytocin signaling pathway: CCND1"
[114] "Oxytocin signaling pathway: CDKN1A"
```

```
[115] "Glucagon signaling pathway: CPT1C*"
[116] "Regulation of lipolysis in adipocytes: PDE3B"
[117] "Renin secretion: PPP3CA"
[118] "Insulin resistance: D-Glucose*"
[119] "Non-alcoholic fatty liver disease (NAFLD): CEBPA"
[120] "AGE-RAGE signaling pathway in diabetic complications: CCND1"
[121] "AGE-RAGE signaling pathway in diabetic complications: NFATC1"
[122] "Maturity onset diabetes of the young: PAX6*"
[123] "Maturity onset diabetes of the young: INS"
[124] "Maturity onset diabetes of the young: IAPP"
[125] "Maturity onset diabetes of the young: NR5A2"
[126] "Endocrine and other factor-regulated calcium reabsorption: ATP2B1"
[127] "Vasopressin-regulated water reabsorption: ARHGDIA"
[128] "Carbohydrate digestion and absorption: SLC2A5"
[129] "Alzheimer's disease: MAPT"
[130] "Parkinson's disease: PRKACA*"
[131] "Huntington's disease: CASP3"
[132] "Cocaine addiction: SLC6A3"
[133] "Amphetamine addiction: PPP3CA"
[134] "Morphine addiction: GABRR3"
[135] "Morphine addiction: GABRR3*"
[136] "Morphine addiction: GABRR3**"
[137] "Morphine addiction: 3,5-Cyclic AMP*"
[138] "Vibrio cholerae infection: ADCY3"
[139] "Epithelial cell signaling in Helicobacter pylori infection: EGFR"
[140] "Legionellosis: HSPA1A"
[141] "Chagas disease (American trypanosomiasis): ADCY1"
[142] "Chagas disease (American trypanosomiasis): AKT3"
[143] "Chagas disease (American trypanosomiasis): NFKB1*"
[144] "Toxoplasmosis: STAT3"
[145] "Tuberculosis: PPP3CA"
[146] "Tuberculosis: CASP3"
[147] "Hepatitis B: NFKBIA"
[148] "Hepatitis B: MMP9*"
[149] "Hepatitis B: BCL2"
[150] "Hepatitis B: CDKN1A"
[151] "Measles: IFNAR1 GNB2L1"
[152] "Influenza A: IL33"
[153] "HTLV-I infection: NFKB1"
[154] "HTLV-I infection: NFATC1"
```

```
[155] "Epstein-Barr virus infection: CDKN1A"
[156] "Epstein-Barr virus infection: NFKB1"
[157] "Epstein-Barr virus infection: STAT3"
[158] "Pathways in cancer: BCL2"
[159] "Pathways in cancer: BIRC8"
[160] "Pathways in cancer: CCND1**"
[161] "Pathways in cancer: CDKN1A*"
[162] "Pathways in cancer: CCND1***"
[163] "Pathways in cancer: CCND1"
[164] "Pathways in cancer: KLK3"
[165] "Pathways in cancer: CSF3R*"
[166] "Pathways in cancer: IL6"
[167] "Pathways in cancer: E2F1*"
[168] "Pathways in cancer: MYC**"
[169] "Proteoglycans in cancer: MAPK1******"
[170] "Proteoglycans in cancer: AKT3"
[171] "Proteoglycans in cancer: MAPK1*"
[172] "Proteoglycans in cancer: RAC1*"
[173] "Proteoglycans in cancer: CCND1"
[174] "Proteoglycans in cancer: CDKN1A*"
[175] "Proteoglycans in cancer: CASP3"
[176] "Proteoglycans in cancer: VEGFA*"
[177] "Proteoglycans in cancer: AKT3*"
[178] "Colorectal cancer: CCND1"
[179] "Colorectal cancer: BAD"
[180] "Colorectal cancer: CASP9"
[181] "Colorectal cancer: CCND1*"
[182] "Renal cell carcinoma: AKT3"
[183] "Pancreatic cancer: NFKB1"
[184] "Pancreatic cancer: E2F1"
[185] "Endometrial cancer: ELK1"
[186] "Endometrial cancer: CCND1"
[187] "Glioma: MAPK1"
[188] "Glioma: MTOR"
[189] "Glioma: MAPK1*"
[190] "Glioma: MTOR*"
[191] "Glioma: E2F1"
[192] "Glioma: E2F1*"
[193] "Prostate cancer: CCND1"
[194] "Prostate cancer: CDKN1A"
```

```
[195] "Prostate cancer: KLK3"
[196] "Thyroid cancer: CCND1"
[197] "Basal cell carcinoma: PTCH1*"
[198] "Basal cell carcinoma: BMP2"
[199] "Basal cell carcinoma: HHIP"
[200] "Basal cell carcinoma: GLI1"
[201] "Basal cell carcinoma: PTCH1"
[202] "Melanoma: CCND1"
[203] "Acute myeloid leukemia: PIM1"
[204] "Acute myeloid leukemia: PIM2"
[205] "Acute myeloid leukemia: NFKB1"
[206] "Acute myeloid leukemia: RPS6KB1"
[207] "Acute myeloid leukemia: MAPK1"
[208] "Small cell lung cancer: CCND1"
[209] "Small cell lung cancer: BIRC8"
[210] "Non-small cell lung cancer: CCND1*"
[211] "Non-small cell lung cancer: CCND1"
[212] "Inflammatory bowel disease (IBD): FOXP3*"
```

Anexo III: Circuitos downregulados en SCCR

```
> path_names_down
 [1] "PPAR signaling pathway: HMGCS2"
 [2] "PPAR signaling pathway: APOA1"
 [3] "PPAR signaling pathway: APOA5"
 [4] "PPAR signaling pathway: PLTP"
 [5] "PPAR signaling pathway: ME1"
 [6] "PPAR signaling pathway: CYP8B1"
 [7] "PPAR signaling pathway: FABP1"
 [8] "PPAR signaling pathway: SLC27A4"
 [9] "PPAR signaling pathway: EHHADH"
[10] "PPAR signaling pathway: CPT2"
[11] "PPAR signaling pathway: ACADM"
[12] "PPAR signaling pathway: CYP27A1"
[13] "PPAR signaling pathway: CD36"
[14] "PPAR signaling pathway: PLIN1"
[15] "PPAR signaling pathway: FABP4"
[16] "PPAR signaling pathway: ADIPOQ"
[17] "PPAR signaling pathway: SORBS1"
[18] "PPAR signaling pathway: PCK1"
[19] "PPAR signaling pathway: AQP7"
[20] "PPAR signaling pathway: GK"
[21] "PPAR signaling pathway: ACADL"
[22] "ErbB signaling pathway: ERBB3 ERBB3"
[23] "Ras signaling pathway: RAC1**"
[24] "Rap1 signaling pathway: CDH1"
[25] "Calcium signaling pathway: Sodium cation*"
[26] "Calcium signaling pathway: Sodium cation"
[27] "Calcium signaling pathway: MYLK4"
[28] "cGMP-PKG signaling pathway: MYL9"
[29] "cGMP-PKG signaling pathway: CNGB1"
[30] "cAMP signaling pathway: MYL9"
[31] "cAMP signaling pathway: ATP1B4 FXYD1"
[32] "Chemokine signaling pathway: PARD3 PRKCZ TIAM1"
[33] "Chemokine signaling pathway: MAPK1"
[34] "NF-kappa B signaling pathway: BCL2L1"
[35] "Oocyte meiosis: REC8"
[36] "p53 signaling pathway: CDK2 CCNE1"
[37] "PI3K-Akt signaling pathway: BCL2L1"
[38] "PI3K-Akt signaling pathway: CASP9"
[39] "AMPK signaling pathway: PPARG"
[40] "AMPK signaling pathway: FBP1"
```

```
[41] "Longevity regulating pathway - mammal: PPARG"
[42] "Vascular smooth muscle contraction: ACTA2"
[43] "Hedgehog signaling pathway: SMO"
[44] "Axon guidance: CDC42*"
[45] "Axon guidance: ABLIM3"
[46] "VEGF signaling pathway: NOS3"
[47] "VEGF signaling pathway: RAC1"
[48] "VEGF signaling pathway: SHC2"
[49] "VEGF signaling pathway: PTK2"
[50] "VEGF signaling pathway: PXN"
[51] "Hippo signaling pathway: ID1"
[52] "Focal adhesion: FLNA ITGB1 ITGA11"
[53] "Focal adhesion: ACTN4 ITGB1 ITGA11"
[54] "Adherens junction: CDH1*"
[55] "Adherens junction: SNAI2"
[56] "Adherens junction: CDH1 CTNNB1"
[57] "Tight junction: ACTB MYL12B"
[58] "Tight junction: EPB41 TJP2 ACTB"
[59] "Tight junction: IGSF5 IGSF5"
[60] "Tight junction: CLDN24 TJP1 TJP2"
[61] "Signaling pathways regulating pluripotency of stem cells: MAPK1"
[62] "Platelet activation: MYL12B"
[63] "NOD-like receptor signaling pathway: NLRC4 PYCARD"
[64] "B cell receptor signaling pathway: VAV3"
[65] "Fc epsilon RI signaling pathway: MAPK14"
[66] "Fc gamma R-mediated phagocytosis: ARPC5"
[67] "Leukocyte transendothelial migration: CDH5"
[68] "Leukocyte transendothelial migration: RAC1*"
[69] "Leukocyte transendothelial migration: MYL12B*"
[70] "Circadian entrainment: PER1"
[71] "Neurotrophin signaling pathway: RHOA"
[72] "Glutamatergic synapse: TRPC1"
[73] "Glutamatergic synapse: CACNA1A"
[74] "Serotonergic synapse: PRKACA"
[75] "Serotonergic synapse: PLA2G4B"
[76] "Serotonergic synapse: TRPC1"
[77] "Taste transduction: PRKACA"
[78] "Regulation of actin cytoskeleton: MYL12B MYH9 ACTB"
[79] "Melanogenesis: CAMK2A"
[80] "Melanogenesis: PRKACA"
[81] "Thyroid hormone signaling pathway: CASP9"
[82] "Adipocytokine signaling pathway: Fatty acid"
[83] "Adipocytokine signaling pathway: Diacylglycerol"
[84] "Glucagon signaling pathway: PFKL"
[85] "Glucagon signaling pathway: FBP1"
```

```
 [86] "Renin secretion: REN"
 [87] "Non-alcoholic fatty liver disease (NAFLD): PPARA RXRA"
 [88] "AGE-RAGE signaling pathway in diabetic complications: F3"
 [89] "Maturity onset diabetes of the young: HNF1B"
 [90] "Aldosterone-regulated sodium reabsorption: KCNJ1"
 [91] "Endocrine and other factor-regulated calcium reabsorption: TRPV5"
 [92] "Endocrine and other factor-regulated calcium reabsorption: CALB1*"
 [93] "Endocrine and other factor-regulated calcium reabsorption: SLC8A1"
 [94] "Endocrine and other factor-regulated calcium reabsorption: CALB1"
 [95] "Gastric acid secretion: MYLK4"
 [96] "Bacterial invasion of epithelial cells: ACTB"
 [97] "Bacterial invasion of epithelial cells: CTNNA1"
 [98] "Bacterial invasion of epithelial cells: DNM1"
 [99] "Vibrio cholerae infection: Potassium cation"
[100] "Pathogenic Escherichia coli infection: CTNNB1"
[101] "Hepatitis B: CDK2"
[102] "Hepatitis B: CDK2*"
[103] "Measles: IL12A"
[104] "HTLV-I infection: JUN"
[105] "Pathways in cancer: CASP9"
[106] "Pathways in cancer: FIGF"
[107] "Pathways in cancer: CTBP1 HDAC1"
[108] "Proteoglycans in cancer: HSPB2"
[109] "Proteoglycans in cancer: KDR"
[110] "Proteoglycans in cancer: ITGA2*"
[111] "Proteoglycans in cancer: MAPK1*****"
[112] "Proteoglycans in cancer: CTNNB1**"
[113] "Proteoglycans in cancer: PTK2"
[114] "Proteoglycans in cancer: PRKCA"
[115] "Proteoglycans in cancer: NUDT16L1"
[116] "Proteoglycans in cancer: PXN"
[117] "Renal cell carcinoma: TGFA"
[118] "Pancreatic cancer: RAC1*"
[119] "Pancreatic cancer: CASP9"
[120] "Endometrial cancer: CASP9"
[121] "Prostate cancer: CASP9"
[122] "Thyroid cancer: CDH1"
[123] "Bladder cancer: ERBB2"
[124] "Chronic myeloid leukemia: HDAC1 CTBP1"
[125] "Non-small cell lung cancer: CASP9"
[126] "Choline metabolism in cancer: CHKA"
[127] "Choline metabolism in cancer: PCYT1A"
[128] "Choline metabolism in cancer: WASF2"
```

12 Anexo III: Código de R utilizado en el análisis

```
#########################################################################
#################Reanálisis de GDS1282 con Hipathia######################
#########################################################################

##  Instalación de paquetes.
#"GEOquery" importa información desde archivos de Gene Expression Omnibus (GEO).
#"Hipathia" analisis funcional mecanístico de expresión génica.
#"tydiverse" para manejo de "data.frames".
#"EnhancedVolcano" para volcanoplots
#"openxlsx" para trabajar con excel

if(!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")
#BiocManager::version()
#BiocManager::install("GEOquery",force=TRUE)
#BiocManager::install("hipathia",force=TRUE)
#install.packages("tidyverse")
#tinytex::install_tinytex()
#BiocManager::install('EnhancedVolcano')
#install.packages("openxlsx")

## Apertura de paquetes

library(GEOquery)
library(tidyverse)
library(Biobase)
library(hipathia)
library(tinytex)
library(EnhancedVolcano)
library(openxlsx)
library(readr)

## Importación de datos de GDS1282 (Estudio GSE2712)

gds1282 <- getGEO('GDS1282', destdir=".")
```

```
class(gds1282)

Meta(gds1282)$channel_count

Meta(gds1282)$description

Meta(gds1282)$feature_count

Meta(gds1282)$platform

Meta(gds1282)$sample_count

Meta(gds1282)$sample_organism

Meta(gds1282)$sample_type

Meta(gds1282)$title

Meta(gds1282)$type


## Matriz de expresión génica. Extracción y modificación


#Extraemos la matriz de Expresión génica


expression_matrix_gds1282_raw<-Table(gds1282)


class(expression_matrix_gds1282_raw)

dim(expression_matrix_gds1282_raw)

names(expression_matrix_gds1282_raw)

colnames(expression_matrix_gds1282_raw)

head(expression_matrix_gds1282_raw)

tail(expression_matrix_gds1282_raw)

expression_matrix_gds1282_raw[1:10,1:6]


#Eliminamos filas con algún NA


expression_matrix_gds1282_na<-
expression_matrix_gds1282_raw[complete.cases(expression_matrix_gds1282_raw),]


class(expression_matrix_gds1282_na)

dim(expression_matrix_gds1282_na)

head(expression_matrix_gds1282_na)

tail(expression_matrix_gds1282_na)

expression_matrix_gds1282_na[1:10,1:6]


#Eliminamos las columnas con los nombres de los genes


expression_matrix_gds1282_id<-expression_matrix_gds1282_na[,-c(1,2)]


class(expression_matrix_gds1282_id)
```

```
dim(expression_matrix_gds1282_id)
colnames(expression_matrix_gds1282_id)
row.names(expression_matrix_gds1282_id)
head(expression_matrix_gds1282_id)
tail(expression_matrix_gds1282_id)
expression_matrix_gds1282_id[1:10,1:6]

#Transformamos de "data.frame" a "matrix"

expression_matrix_gds1282_m<-as.matrix(expression_matrix_gds1282_id)

class(expression_matrix_gds1282_m)
dim(expression_matrix_gds1282_m)
colnames(expression_matrix_gds1282_m)
expression_matrix_gds1282_m[1:10,1:6]

#Asignamos a los nombres de las filas,los nombres de los genes

row.names(expression_matrix_gds1282_m)<-(expression_matrix_gds1282_na[,1])

class(expression_matrix_gds1282_m)
dim(expression_matrix_gds1282_m)
expression_matrix_gds1282_m[1:10,1:6]

#Transformamos en la matriz de expresión los IDs de los genes a Entrez IDs

expression_matrix_gds1282_trans<-translate_data(expression_matrix_gds1282_m,"hsa")

class(expression_matrix_gds1282_trans)
dim(expression_matrix_gds1282_trans)
expression_matrix_gds1282_trans[1:10,1:6]

#Visualizamos la matriz de expresión sin normalizar.

boxplot(expression_matrix_gds1282_trans)
boxplot(expression_matrix_gds1282_trans[1:1000,])
boxplot(expression_matrix_gds1282_trans[1:40,])

#Normalizamos la matriz de expresión entre 0 y 1.
```

```
expression_matrix_gds1282_nor<-normalize_data(expression_matrix_gds1282_trans, by_quantiles =
TRUE)


class(expression_matrix_gds1282_nor)

dim(expression_matrix_gds1282_nor)

expression_matrix_gds1282_nor[1:10,1:5]


#Visualizamos la matriz de expresión normalizada.


boxplot(expression_matrix_gds1282_nor)

boxplot(expression_matrix_gds1282_nor[1:1000,])

boxplot(expression_matrix_gds1282_nor[1:40,])

boxplot(expression_matrix_gds1282_nor[,1:4])


## Matriz de diseño experimental


#Importamos la matriz de diseño experimental


Columns(gds1282)[,1:2]

exp_design_gds1282<-data.frame(Columns(gds1282)[,2])

exp_design_gds1282


#creamos una columna de estatus de enfermedad sin espacios


exp_design_gds1282$group<-c("C","C","C",
                            "WT","WT","WT","WT","WT","WT","WT","WT","WT","WT",
                            "WT","WT","WT","WT","WT","WT","WT","WT",

"CCSK","CCSK","CCSK","CCSK","CCSK","CCSK","CCSK","CCSK","CCSK","CCSK",
                            "CCSK","CCSK","CCSK","CCSK")


exp_design_gds1282<-data.frame(exp_design_gds1282[,2])

colnames(exp_design_gds1282)<-c("group")


#Asiganamos a los nombres de las filas, los nombres de las muestras


row.names(exp_design_gds1282)<-(Columns(gds1282)[,1])

exp_design_gds1282


class(exp_design_gds1282)

dim(exp_design_gds1282)
```

```
colnames(exp_design_gds1282)

row.names(exp_design_gds1282)

exp_design_gds1282


## creación de SummarizedExperiment: objeto que encapsula las dos matrices


gds1282_sum_exp<-SummarizedExperiment(assays =
SimpleList(raw=expression_matrix_gds1282_nor),colData=exp_design_gds1282)


class(gds1282_sum_exp)


## Descarga de 146 rutas de señalización humanas ("hsa")


pathways<-load_pathways(species="hsa")

get_pathways_list(pathways)


## cálculo del nivel de activación de los circuitos de las muestras

#Se obtiene un objeto "MultiArrayExperiment", que incluye los objetos "paths" y "nodes"

#Se estiman un total de 741 genes 5.57% de la muestra


results<-hipathia(expression_matrix_gds1282_nor,pathways,decompose=FALSE,verbose = FALSE)

results


## Matriz de nivel de actividad.

#Las filas son los circuitos

#Las columnas son las muestras


path_vals<-get_paths_data(results,matrix=TRUE)

path_vals[1:10,1:4]


## Visualización del nivel de activación de los circuitos en los tres tipos de muestra: c ccsk
wt


#Creación de los tres grupos de analisis que queremos visualizar según el estatus de enfermedad


sample_group<-exp_design_gds1282[colnames(path_vals),"group"]

class(sample_group)

sample_group


#Heatmap
```

```
heatmap_plot(path_vals,group=sample_group,variable_clust = TRUE,colors = "redgreen")


## Anotación funcional con Gene Ontology y ___UniProt___


#Con GeneOntology se anotan 1654 funciones:


go_vals <- quantify_terms(results, pathways, dbannot = "GO")


#Con Uniprot se anotan 142 funciones:


uniprot_vals <- quantify_terms(results, pathways, dbannot = "uniprot")


# Visualización de la activación de las funciones anotadas por go y uniprot de los tres grupos


heatmap_plot(uniprot_vals,group=sample_group,colors="classic",variable_clust = TRUE)


heatmap_plot(go_vals,group=sample_group,colors="classic",variable_clust = TRUE)


## Comparación de activación de circuitos en "Clear Cell Sarcoma Kidney" vs "Kidney Control"


#Comparativa entre grupos


comp_paths<-do_wilcoxon(path_vals,sample_group,g1="CCSK",g2="C")

dim(comp_paths)

hhead(comp_paths,20)


#Seleccionamos los circuitos con cambios significativos (FDRp.value<0.05) 340/1876, y obtenemos
un listado con su nombre.


comp_paths_sig<-subset(comp_paths, FDRp.value < 0.05)

dim(comp_paths_sig)

head(comp_paths_sig,20)

path_names_sig<-get_path_names(pathways,rownames(comp_paths_sig))


#Ordenamos los circuitos de mayor a menor significación (FDRp.value creciente)


comp_paths_ordered<-comp_paths_sig[order(comp_paths_sig$FDRp.value,decreasing=FALSE),]

head(comp_paths_ordered,20)


#Calculamos cuantos circuitos están upregulados y downregulados.(212 up, 128 down), y sacamos un
listado con los nombres
```

```
comp_paths_sig_up<-subset(comp_paths_sig,statistic>0)
comp_paths_sig_down<-subset(comp_paths_sig,statistic<0)

dim(comp_paths_sig_up)
dim(comp_paths_sig_down)

path_names_up<-get_path_names(pathways,rownames(comp_paths_sig_up))
path_names_down<-get_path_names(pathways,rownames(comp_paths_sig_down))

#Hacemos vectores con con los nombres de los circuitos sig, sig_up y sig_down

hsa_ids_circuitos_sig<-row.names(comp_paths_sig)
hsa_ids_circuitos_sig_up<-row.names(comp_paths_sig_up)
hsa_ids_circuitos_sig_down<-row.names(comp_paths_sig_down)

#Seleccionamos de la matriz de valores de actividad los circuitos sig, sig_up y sig_down

path_vals_sig<-path_vals[hsa_ids_circuitos_sig,]
path_vals_sig_up<-path_vals[hsa_ids_circuitos_sig_up,]
path_vals_sig_down<-path_vals[hsa_ids_circuitos_sig_down,]

#Hacemos heatmap de las rutas sig sig_up y sig_down

heatmap_plot(path_vals_sig,group=sample_group,variable_clust = TRUE,colors = "redgreen")

heatmap_plot(path_vals_sig_up,group=sample_group,variable_clust = TRUE,colors = "redgreen")

heatmap_plot(path_vals_sig_down,group=sample_group,variable_clust = TRUE,colors = "redgreen")

#Exportamos la matrices "comp_paths" y "comp_paths_sig", como un archivo de texto con formato
tsv:

write.table(comp_paths,file = "C:/Users/comp_paths.txt", sep ="\t" ,row.names = TRUE, col.names
= TRUE)
write.table(comp_paths_sig,file = "C:/Users/comp_paths_sig.txt", sep ="\t" ,row.names = TRUE,
col.names = TRUE)

#Resumen de los resultados de la comparativa de actividad de los circutitos entre C y CCSK

pathways_summary<-get_pathways_summary(comp_paths,pathways)
head(pathways_summary,20)
```

```
#Exportamos la matriz resunen de resultados


write.xlsx(pathways_summary, file = "My_openxlsx_File.xlsx")


##PCA de los circuitos de activación en los tres grupos


ranked_path_vals<-path_vals[order(comp_paths$p.value,decreasing=FALSE),]
pca_model<-do_pca(ranked_path_vals[1:ncol(ranked_path_vals),])


#PCA de dos dimensiones


pca_plot(pca_model,sample_group,legend=TRUE)
pca_plot(pca_model,sample_group,legend=FALSE)


#Vamos a verlo con un PCA con tres dimensiones:


pca_plot_multiple<-multiple_pca_plot(pca_model,sample_group,cex=3,plot_variance = TRUE)


## volcano plot


EnhancedVolcano(comp_paths,
                lab = rownames(comp_paths),
                x = 'statistic',
                y = 'p.value',
                xlim = c(-4, 4),
                ylim = c(0, 3),
                pCutoff = 0.05,
                FCcutoff = 0.5,)


## Visualización de las rutas


#hsa05214/Glioma


colors_de_hipathia <- node_color_per_de(results, pathways, sample_group,
"CCSK", "C", colors = "hipathia")
pathway_comparison_plot(comp_paths, metaginfo = pathways, pathway = "hsa05214",
node_colors = colors_de_hipathia, colors = "hipathia")


#hsa04961/Reabsorción de calcio
colors_de_hipathia <- node_color_per_de(results, pathways, sample_group,
```

```
"CCSK", "C", colors = "hipathia")

pathway_comparison_plot(comp_paths, metaginfo = pathways, pathway = "hsa04961",

node_colors = colors_de_hipathia, colors = "hipathia")


#hsa04151/PI3K-Atk


colors_de_hipathia <- node_color_per_de(results, pathways, sample_group,

"CCSK", "C", colors = "hipathia")

pathway_comparison_plot(comp_paths, metaginfo = pathways, pathway = "hsa04151",

node_colors = colors_de_hipathia, colors = "hipathia")


###REPOSICIONEMIENTO DE FARMACOS


##Importamos la matriz "shap_relevant_stable_matrix", Matriz 186 X 116 (Circuitos X Genes),
procedente de la modelización por machine learning


shap_relevant_stable_matrix<-read_tsv('C:/Users/mcprado/Desktop/Mª CARMEN/MCPZ/mpz/MASTER
BIOINFORMATICA/TFM/MODELO REPOSICIONEMIENTO/prado-zamora/shap_relevant_stable_matrix.tsv')


dim(shap_relevant_stable_matrix)

shap_relevant_stable_matrix[1:10,1:6]


#Seleccionamos la submatrices con los circuitos upregulados y downregulados:

#shap_relevant_stable_matrix_up Tiene 96circuitos X 116 KDR

#shap_relevant_stable_matrix_down Tiene 84circuitos X 116 KDR


shap_relevant_stable_matrix_up<-
shap_relevant_stable_matrix|>filter(circuit_name%in%path_names_up)

shap_relevant_stable_matrix_down<-
shap_relevant_stable_matrix|>filter(circuit_name%in%path_names_down)


dim(shap_relevant_stable_matrix_up)

dim(shap_relevant_stable_matrix_down)


shap_relevant_stable_matrix_up[1:10,1:6]

shap_relevant_stable_matrix_down[1:10,1:6]


> sessionInfo ()

R version 4.2.1 (2022-06-23 ucrt)

Platform: x86_64-w64-mingw32/x64 (64-bit)

Running under: Windows 10 x64 (build 19043)
```

13 Anexo IV: Código para selección KDTs

```
############################################################

######    SHAP MODELS  WITH SIGN INTERPRETATION   PRADO-ZAMORA ########

############################################################


# CODIGO REALIZADO POR CARLOS LOUCERA Y MARINA ESTEBAN-MEDINA







if (!require(pacman, quietly = TRUE)){

  install.packages(pacman)

}else{

  library(pacman)

}


pacman::p_load("here", "rentrez","reutils","hipathia", "biomaRt", "utils", "stringr",
"AnnotationDbi", "org.Hs.eg.db",

               "dplyr","tidyr", "openxlsx", "data.table", "scales", "BiocManager", "NMF")

#

# BiocManager::install("clusterProfiler")

# BiocManager::install("org.Hs.eg.db")


pathways <- load_pathways("hsa")




stability <- fread(file = file.path(here("stability_results_symbol.tsv"))) %>% as.data.frame(.)

rownames(stability) <-stability$circuit_name

stability <- stability[-1,-1]


### Load shap relevances and filter those that overcome the threshold #


shap <- fread(file = file.path(here("shap_summary_symbol.tsv")), header = T) %>% as.data.frame()

rownames(shap)<- shap$circuit_name

shap <- shap[ ,-1]
```

```
shap_stable <- shap[rownames(shap) %in% rownames(stability)[stability$stability >= 0.4],] ### No
funciona porque los circuitos en stability son códigos



threshold <- fread(file = file.path(here("shap_selection_symbol.tsv")), header = T) %>%
as.data.frame()

rownames(threshold) <- threshold$circuit_name

threshold <- threshold[ ,-1]

threshold_stable <-  threshold[rownames(threshold) %in% rownames(stability)[stability$stability
>= 0.4],]




## Subset  only the shap values which are relevant for at least 1 circuit in threshold



shap[which(threshold == 0, arr.ind = T)] <- 0

shap_relevant <- shap[,(apply(threshold, 2, function(y) any(y == 1)))]

dim(shap_relevant)



shap_stable[which(threshold_stable == 0, arr.ind = T)] <- 0

shap_relevant_stable <- shap_stable[,(apply(threshold_stable, 2, function(y) any(y == 1)))]







targets_shap_rel <- data.frame( Gene_symbol = colnames(shap_relevant),

                                Gene_name = mapIds(org.Hs.eg.db, keys, keys =
colnames(shap_relevant), column = "GENENAME", keytype = "SYMBOL"),

                                Entrez = mapIds(org.Hs.eg.db, keys, keys =
colnames(shap_relevant), column = "ENTREZID", keytype = "SYMBOL"),

                                stringsAsFactors = F)



targets_shap_rel_stable <- data.frame( Gene_symbol = colnames(shap_relevant_stable),

                                       Gene_name = mapIds(org.Hs.eg.db, keys, keys =
colnames(shap_relevant_stable), column = "GENENAME", keytype = "SYMBOL"),

                                       Entrez = mapIds(org.Hs.eg.db, keys, keys =
colnames(shap_relevant_stable), column = "ENTREZID", keytype = "SYMBOL"),

                                       stringsAsFactors = F)




write.xlsx(targets_shap_rel, file = here("relevant_targets.xlsx"))

write.xlsx(targets_shap_rel_stable, file = here("relevant_targets_above04_stable.xlsx"))
```

```
####### PLOT HEAT MAP AS RELEVANCES #######


## First we have to rescale by circuit (rows) each SHAP matrix to make the circuits comparable


mat <- as.data.frame(t(apply(shap_relevant, 1, function(x) x/max(abs(x)))))


mat_stable <- as.data.frame(t(apply(shap_relevant_stable, 1, function(x) x/max(abs(x)))))







#########################################################

##### USE DRUGBANK DATA FOR DRUGS RELEVANT  FUN ANNOT ######

#########################################################


## Load Drugbank database


data_folder2 <- "~/INFO_PROYECTO/drugbank"

fname2 <- "drugbank_drug-bindings_v5.1.8.tsv"

fpath2 <-file.path(data_folder2,fname2)


drugbank_alltar <- read.delim(file = fpath2, sep = "\t" )




drugbank_effects_tar <- drugbank_alltar[which(drugbank_alltar$entrez_id %in%
targets_shap_rel_stable$Entrez), c(2,8,10,13) ]

# drugbank_effects_tar$actions[is.na(drugbank_effects_tar$actions)] <- "other"

drugbank_effects_tar$actions[which(drugbank_effects_tar$entrez_id %in% "3688")] <- "other" ##
Sustituimos el único gen que solo tiene como anotación NA

length(unique(drugbank_effects_tar$entrez_id))


library(tidyverse)


drug_bygenes <- data.frame(aggregate(cbind(as.character(drugbank_effects_tar$actions)) ~
drugbank_effects_tar$entrez_id, data = drugbank_effects_tar , FUN = paste))

colnames(drug_bygenes) <- c("gene", "drug_action")






## Aquí filtramos la acción que más aparece para cada target PARA EL PLOT
```

```
drug_bygenes$drug_action <- sapply(drug_bygenes$drug_action, function(x)
names(which.max(table(x))))



# targets_shap_rel_stable$Entrez[which(!targets_shap_rel_stable$Entrez %in% drug_bygenes$gene )]
## Sale el "3688"



## DATA FRAME CON TARGETS ENTREZ ID Y EFECTO MÄS FRECUENTE

drug_ef <- data.frame( entrez = targets_shap_rel_stable$Entrez[match(colnames(mat),
targets_shap_rel_stable$Gene_symbol)],

                       symbol = colnames(mat))

drug_ef$Drug_effect = drug_bygenes$drug_action[match(drug_ef$entrez, drug_bygenes$gene)]



drug_ef <- drug_ef[which(!is.na(drug_ef$Drug_effect)),]



any(is.na(drug_ef))



table(drug_ef$Drug_effect)



# Simplify the drug action vector

simplfied_drugact <- function(annotations){

  Drug_effect = ifelse(annotations$Drug_effect ==
"activator|agonist,inducer|inducer|activator,modulator|substrate|stimulator|", "Activator",

                       ifelse(annotations$Drug_effect == "ligand", "Ligand",

                              ifelse(annotations$Drug_effect == "antibody","Antibody",

                                     ifelse(annotations$Drug_effect == "agonist" , "Activator",

                                            ifelse(annotations$Drug_effect == "potentiator" ,
"Activator",

                                                   ifelse(annotations$Drug_effect == "ligand",
"Activator",

                                                          ifelse(annotations$Drug_effect ==
"binder", "Binder",

                                                                 ifelse(annotations$Drug_effect
== "other|neutralizer|incorporation into and destabilization", "NA","Inhibitor"))))))))

  return(Drug_effect)

}




drug_ef$Drug_effect <- simplfied_drugact(drug_ef)

dim(drug_ef)

table(drug_ef$Drug_effect)



# FUNCION DE HEATMAP CON ANOTACIONES
```

```
annot_testAheatmap <- function(matrix, annotations, title, anot_colors) {

  aheatmap(matrix, color = "-RdYlBu", border_color = "white", annCol=
annotations["Drug_effect"], Rowv = F, Colv = F,

           cexCol = 4, cexRow = 10, main =  title,

           annColors = anot_colors )

}




## DO IT FOR ALL THE TARGETS ###



## Selecionamos los colores que van a anototar los efectos de las drogas

annot.color.col <- list('Drug_effect'=c('green','lightblue', "purple", "black", "grey"))



png(here("heatmap_PradoZamora_KDTrelstable_drugEffect.png"), height = 8000, width = 8000, res =
500)

annot_testAheatmap(matrix = mat_stable, annotations = drug_ef[drug_ef$symbol %in%
colnames(mat),], anot_colors = annot.color.col,

                   title = "RELEVANCE SCORE OF RELEVANT KDTs OVER STABLE CIRCUITS \n Score of
how relevant is each KDT gene for the activity of each stable circuit with sign \n Drug effect
annotation according the most frequent effect")



dev.off()
```