PEC1

Vargas Parra Esteban

25/4/2020

Resumen

De referencia para el estudio, se tomaron unas matrices de diferentes líneas celulares. Esto con el objetivo de comparar aquellas líneas que eran sensibles a prexasertib con aquellas que eran resistentes, pues esta es una molécula que interfire en el ciclo celular, permitiendo que células cancerígenas continuen con su respectiva división. Debido a esto, se hizo un análisis donde se pudo obtener los resultados de aquellos genes que se expresaron diferencialmente; para esto se implementó el software R con la ayuda del paquete Bioconductor[1].

Palabras claves

Líneas celulares, cancerígenas, genes, R, Bioconductor.

Introducción

El proyecto Pan-Cáncer surgió como la idea de reunir el conjunto de datos de The Cancer Genome Atlas (TCGA), los cuales fuesen coherentes y consistentes en todas las clases de tumores, así como lo fueran en todas las plataformas para poder ser interpretados. Como objetivo fundamental de este proyecto, se desea determinar la alteraciones presentes en diferentes líneas tumorales para poder diseñar terapias efectivas en algún tipo de cáncer y así poder aplicarlas en otros perfiles tumorales semejantes Weinstein et al. (2013).

Se sabe que una de las razones por las que se presentan tumores, es debido al estrés de replicación (ER), el cual se define como el desacoplamiento del desarrollo impulsado por helicasa y el avance de las ADN polimerasas en la bifurcación de replicación del ADN. Como resultado, aquellos cánceres que se encuentran bajo ER, impulsan a la proliferación celular continua e impulsan la inestabilidad genómica Zhao, Watkins, and Piwnica-Worms (2002). La activación de la quinasa de punto de control 1(CHK1) en respuesta al daño excesivo por ER, da como resultado la fosforilación mediada por CHK1 de la fosfatasa CDK2 CDC25A, dirigiéndose a la destrucción proteolítica. En ausencia de CDC25A, la fosforilación inhibitoria de CDK2 en el residuo de tirosina 15 se mantiene deteniendo así la fase S para permitir la

[1] https://github.com/EstebanVargasParra/PEC1.git

reparación del daño del ADN (DDR) y la resolución de conflictos de replicación del ADN Smith et al. (2010).

Prexasertib es un inhibidor de CHK1, evitando así la reparación del ADN dañado. Esto puede conllevar a la acumulación de ADN dañado, promoviendo la indestabilidad genómica. Además, prexasertib potencia la citotoxicidad de los genes que dañan el ADN revirtiendo la resistencia de las células tumorales a los agentes quimioterapéuticos Lowery et al. (2017). Por otra parte, en estudios clínicos se ha visto la alta relación entre la expresión de la ciclina E con prexasertib, en una población de pacientes con cáncer de ovario seroso de alto grado altamente tratado, lo curioso, es que la alta expresión de éstos se ha relacionado con el estrés de replicación mejorado Jones et al. (2013).

Materiales

Para este trabajo se tomó los reultados del estudio "A pan-cancer transcriptome analysis to identify the molecular mechanism of prexasertib resistance [microarray]". Los datos se encuentran disponibles con la entrada de serie GSE143007.

Software

El análisis de este trabajo se realizó con la versión de R 3.6.3 (2020-02-09). R es un leguaje para computación, estadísticas y gráficos, el cual proporciona una variedad de técnicas estadísticas y gráficas. R-Studio es una interfaz que facilita el uso, debido que posee un espacio más cómodo y gráfico.

Datos

Los datos fueron tomados de un estudio publicado por Blosser et al. (2020), se encuentran en la base de datos Gene Expression Omnibus (GEO), un repositorio de datos genómicos funcionales públicos el cual guarda y comparte libremente la expresión génica de alto rendimiento y otros conjuntos de datos genómicos funcionales. Los datos escogidos se encuentran identificados con el código de acceso **GSE143007**.

Como fundamento del estudio, se observó la respuesta prexasertib en una variedad de tumores que eran de interés clínico. Por lo tanto se buscó identificar marcadores de sensibilidad prexasertib y definir los mecanismos moleculares de resistencia intrínseca y adquirida utilizando modelos preclínicos que representan múltiples tipos de tumores. Para ello, se generaron líneas celulares resistentes a prexasertib de diferentes tipos de cáncer utilizando un protocolo de escalado de concentración de fármacos a largo plazo. Con al menos 3 réplicas biológicas en líneas resistentes y sensibles. El experimento tomó en cuenta dos factores, siendo el primero la línea celular del cual se tomó, para este estudio tená tres niveles lo cuales fueron RH41, NHI-H520 y SJCRH30; el otro factor es la respuesta, siendo de dos niveles, las líneas resistentes a prexasertib (PR) y las líneas sensibles a prexasertib (PS). teniendo como

resultado un diseño factorial 3x2 (CellLine y Response). Se usaron microarrays tipo Clariom S Human de Affymetrix para identificar genes expresados diferencialmente.

- Línea celular
- RH41
- NHIH520
- SJCRH30
- Respuesta
- Prexasertib Resistant (PR)
- Prexasertib Sensitive (PS)

Por otra parte, el manual limma nos indica que para parametrizar los datos, es más sencillo en un sólo factor aunque sea menos intuitivo, ya que nos deja realizar de una forma más simple las cuestiones en las que más interés hay sobre la investigación. Por lo cual quedaremos con modelo de un factor y seis niveles

- RH41.PR
- RH41.PS
- NHIH520.PR
- NHIH520.PS
- SJCRH30.PR
- SJCRH30.PS

Métodos

Para el siguiente estudio se siguieron los pasos del Pipelin Statistical Analysis of Microarray data Based on Gonzalo, Ricardo and Sanchez-Pla, Alex (2019) March 29, 2020.

Preparación de datos:

Como se dijo anteiormente, los datos fueron tomados de la base de datos Gene Expression Omnibus. Estos fueron descargados como archivos .CEL en cada muestra del experimento. Además, se realizó una archivo .csv el cual se denominó targets.

Tanto los archivos .CEL como el archivo targets.csv, se guardaron en una carpeta denominada “data” en el repositorio creado. Así como se creó una carpeta nombrada “results” donde llegarán todos los resultados de nuestro trabajo.

Preparación de datos para su análisis

El archivo targets, se encarga de relacionar cada nombre de los archivos .CEL, con su categoría en el experimento, consta de cinco columnas con el siguiente orden:

- FileName: indica el nombre del archivo .CEL.
- Grupo: es la categoría en el experimento para cada muestra.
- CellLines: indica a la línea celular a la que pertenece cada muestra.
- Response: nos señala si la muestra responde con sensibilidad o resistencia al prexasertib.
- ShotName: es una etiqueta corta para cada muestra.

Content of the targets file used for the current analysis

FileName	Group	CellLine	Response	ShortName
GSM4248426	RH41.PS	RH41	PS	RH41.PS.1
GSM4248427	RH41.PS	RH41	PS	RH41.PS.2
GSM4248428	RH41.PS	RH41	PS	RH41.PS.3
GSM4248429	RH41.PR	RH41	PR	RH41.PR.1
GSM4248430	RH41.PR	RH41	PR	RH41.PR.2
GSM4248431	RH41.PR	RH41	PR	RH41.PR.3
GSM4248432	RH41.PR	RH41	PR	RH41.PR.4
GSM4248433	RH41.PR	RH41	PR	RH41.PR.5
GSM4248434	NCIH520.PS	NCIH520	PS	NCIH520.PS.1
GSM4248435	NCIH520.PS	NCIH520	PS	NCIH520.PS.2
GSM4248436	NCIH520.PS	NCIH520	PS	NCIH520.PS.3
GSM4248437	NCIH520.PR	NCIH520	PR	NCIH520.PR.1
GSM4248438	NCIH520.PR	NCIH520	PR	NCIH520.PR.2
GSM4248439	NCIH520.PR	NCIH520	PR	NCIH520.PR.3
GSM4248440	SJCRH30.PS	SJCRH30	PS	SJCRH30.PS.1
GSM4248441	SJCRH30.PS	SJCRH30	PS	SJCRH30.PS.2
GSM4248442	SJCRH30.PS	SJCRH30	PS	SJCRH30.PS.3
GSM4248443	SJCRH30.PR	SJCRH30	PR	SJCRH30.PR.1
GSM4248444	SJCRH30.PR	SJCRH30	PR	SJCRH30.PR.2
GSM4248445	SJCRH30.PR	SJCRH30	PR	SJCRH30.PR.3

Paquetes de instalación en R

Lo primero que se debe realizar es la instalación de los paquetes necesarios para que se pueda llevar a cabo el análisis de los resultados. Los paquete más comunes serán los CRAN y los de Bioconductor.

Para esto, se realizar primero una instalación de “BiocManager”, como indica el siguiente código:

Y enseguida, se hará la instalación de los paquete CRAN y Bioconductor para empezar con nuestro estudio.

Leer los archivos .CEl

Como primer paso leeremos nuestros archivos .CEL, para poder asociarlos con nuestro archivo *targets* y almacenarlo en una sola variable, con el objetivo de combinar las fuentes de información en una sola estructura apropiada.

```
ExpressionFeatureSet (storageMode: lockedEnvironment)
assayData: 1 features, 20 samples
  element names: exprs
protocolData
  rowNames: RH41.PS.1 RH41.PS.2 ... SJCRH30.PR.3 (20 total)
  varLabels: exprs dates
  varMetadata: labelDescription channel
phenoData
  rowNames: RH41.PS.1 RH41.PS.2 ... SJCRH30.PR.3 (20 total)
  varLabels: Group CellLine Response ShortName
  varMetadata: labelDescription channel
featureData: none
experimentData: use 'experimentData(object)'
Annotation: pd.clariom.s.human
```

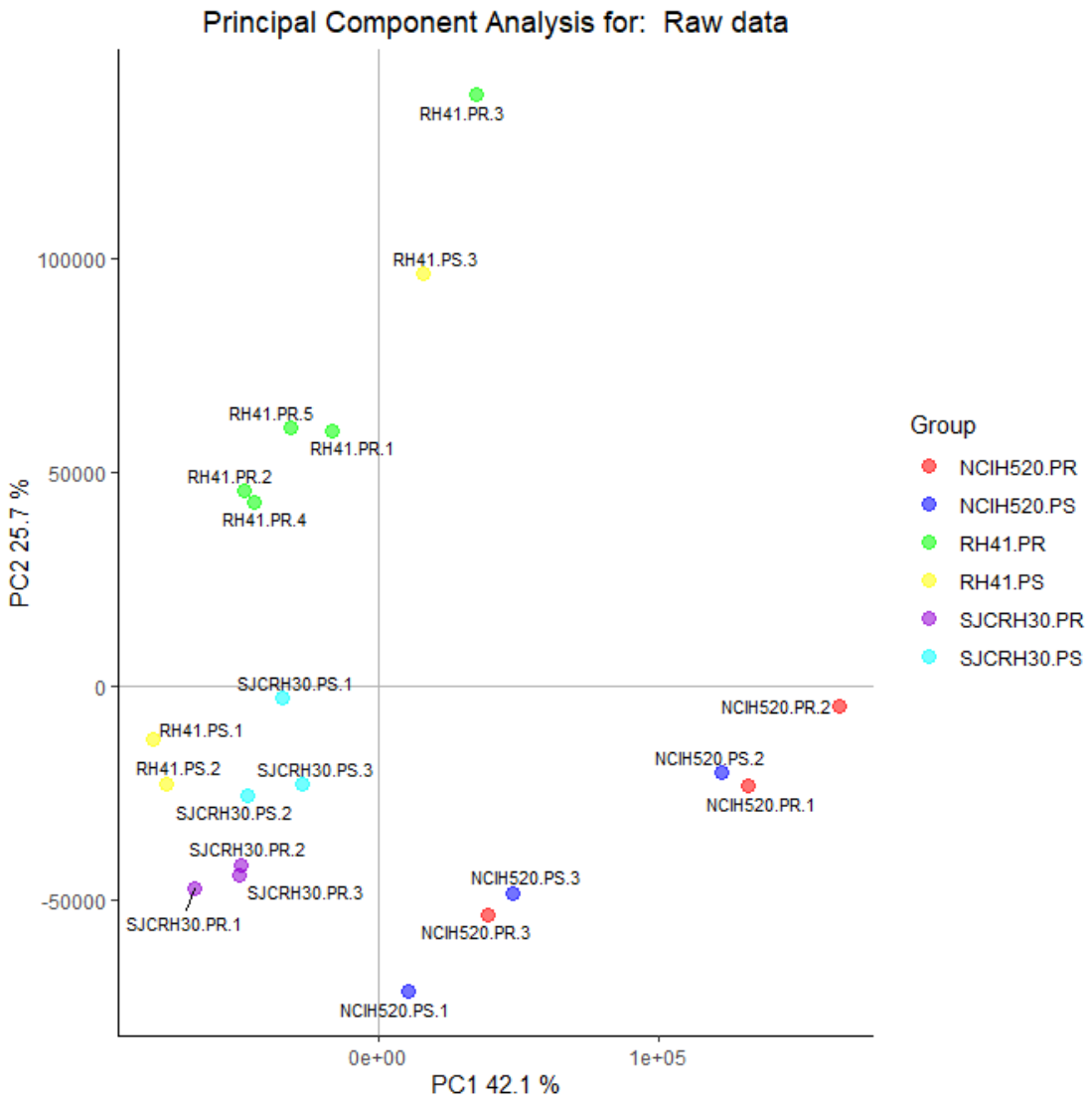
Control de calidad de los datos

En el momento que los datos están cargados, es importante rectificar la calidad para proceder a la normalización. Para ello, revisamos los resultados de nuestra matriz y concluimos cuáles resultados del experimento se encuentran por encima de un umbral definido.

	array	sampleNames	*1	*2	*3	Group	CellLine	Response	ShortName
☐	1	RH41.PS.1				RH41.PS	RH41	PS	RH41.PS.1
☐	2	RH41.PS.2				RH41.PS	RH41	PS	RH41.PS.2
☐	3	RH41.PS.3				RH41.PS	RH41	PS	RH41.PS.3
☐	4	RH41.PR.1				RH41.PR	RH41	PR	RH41.PR.1
☐	5	RH41.PR.2				RH41.PR	RH41	PR	RH41.PR.2
☑	6	RH41.PR.3	x		x	RH41.PR	RH41	PR	RH41.PR.3
☐	7	RH41.PR.4				RH41.PR	RH41	PR	RH41.PR.4
☑	8	RH41.PR.5			x	RH41.PR	RH41	PR	RH41.PR.5
☐	9	NCIH520.PS.1				NCIH520.PS	NCIH520	PS	NCIH520.PS.1
☑	10	NCIH520.PS.2			x	NCIH520.PS	NCIH520	PS	NCIH520.PS.2
☐	11	NCIH520.PS.3				NCIH520.PS	NCIH520	PS	NCIH520.PS.3
☑	12	NCIH520.PR.1			x	NCIH520.PR	NCIH520	PR	NCIH520.PR.1
☑	13	NCIH520.PR.2	x	x	x	NCIH520.PR	NCIH520	PR	NCIH520.PR.2
☐	14	NCIH520.PR.3				NCIH520.PR	NCIH520	PR	NCIH520.PR.3
☐	15	SJCRH30.PS.1				SJCRH30.PS	SJCRH30	PS	SJCRH30.PS.1
☐	16	SJCRH30.PS.2				SJCRH30.PS	SJCRH30	PS	SJCRH30.PS.2
☐	17	SJCRH30.PS.3				SJCRH30.PS	SJCRH30	PS	SJCRH30.PS.3
☐	18	SJCRH30.PR.1				SJCRH30.PR	SJCRH30	PR	SJCRH30.PR.1
☐	19	SJCRH30.PR.2				SJCRH30.PR	SJCRH30	PR	SJCRH30.PR.2
☐	20	SJCRH30.PR.3				SJCRH30.PR	SJCRH30	PR	SJCRH30.PR.3

Aspect of the summary table, in the index.html file, produced by the arrayQualityMetrics package on the raw data.

Lo datos sin procesar indicaron que hay 5 grupos que tienen valores atípicos, principalmente *NCIH520.PR.2*, que presentó valores atípicos en los tres grupos de análisis. Por otra parte, *RH41.PR.3* presentó diferencia en los grupos de análisis de las *ditancias entre los arrays* y el *MAPlots*. Por último *RH41.PR.5*, *NCIH520.PS.2* y *NCIH520.PR.1*, obtuvieron valores atípicos únicamente en el *MAPlots*



Visualización de los dos primeros componentes principales para datos sin procesar

El análisis se componentes principales, representa el 42.1% de variablidad de nuestras muestras, esta variabilidad se destaca principalmente por la línea ceular de donde se obtuvo los resultados del experimento.

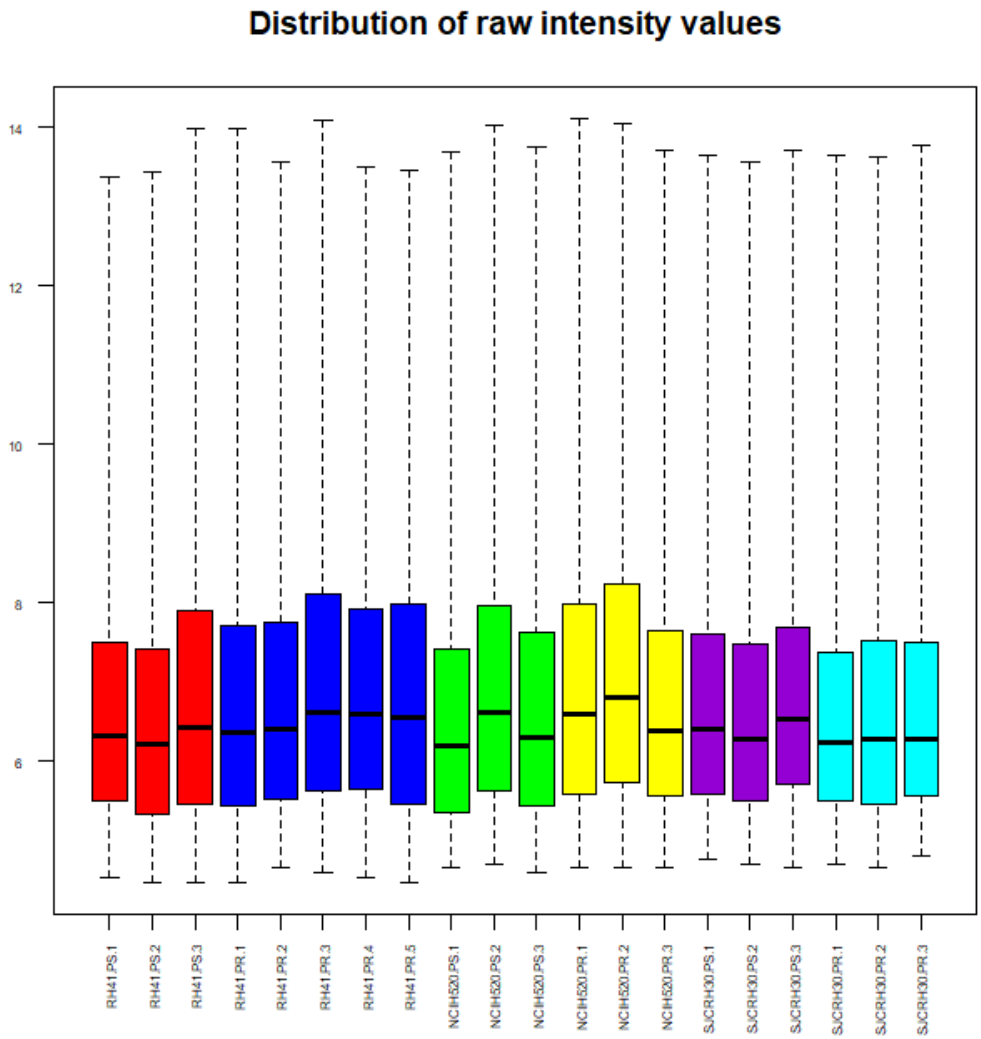
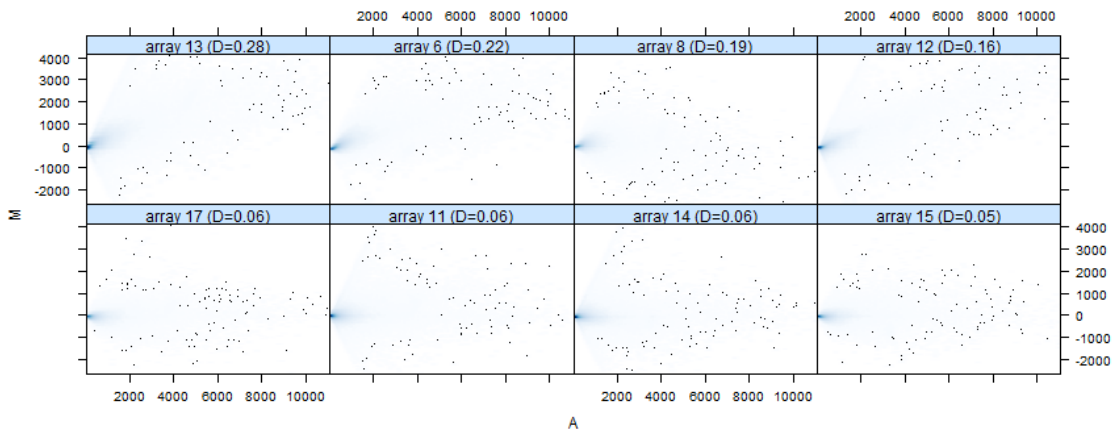


Diagrama de caja (Raw Data)

El boxplot representa las distribuciones de intensidad de señal de cada matriz, en donde cada caja representa un grupo. Se espera que las distribuciones tengan rangos y anchos similares, algo que no está ocurriendo en este instante.



El *MAPlots*, define la distirbución de la masa, si la tendencia muestra un rango infeior *A*, señala que las matrices poseen diferentes intensidades de fondo como sucede en nuestra respectiva gráfica.

Normalización de datos

Como nuestras matrices poseen cierta diferencias entre ellas, para poder realizar una análisis de expresión diferencial se buscará reducir o eliminar toda la variabilidad presente. Así, procedmeos a la normalización, donde se corregirán todos lo posibles errores sistemáticos y la variabilidad de las muestras, logrando comparaciones bajo las mismas condiciones.

```
Background correcting
Normalizing
Calculating Expression
```

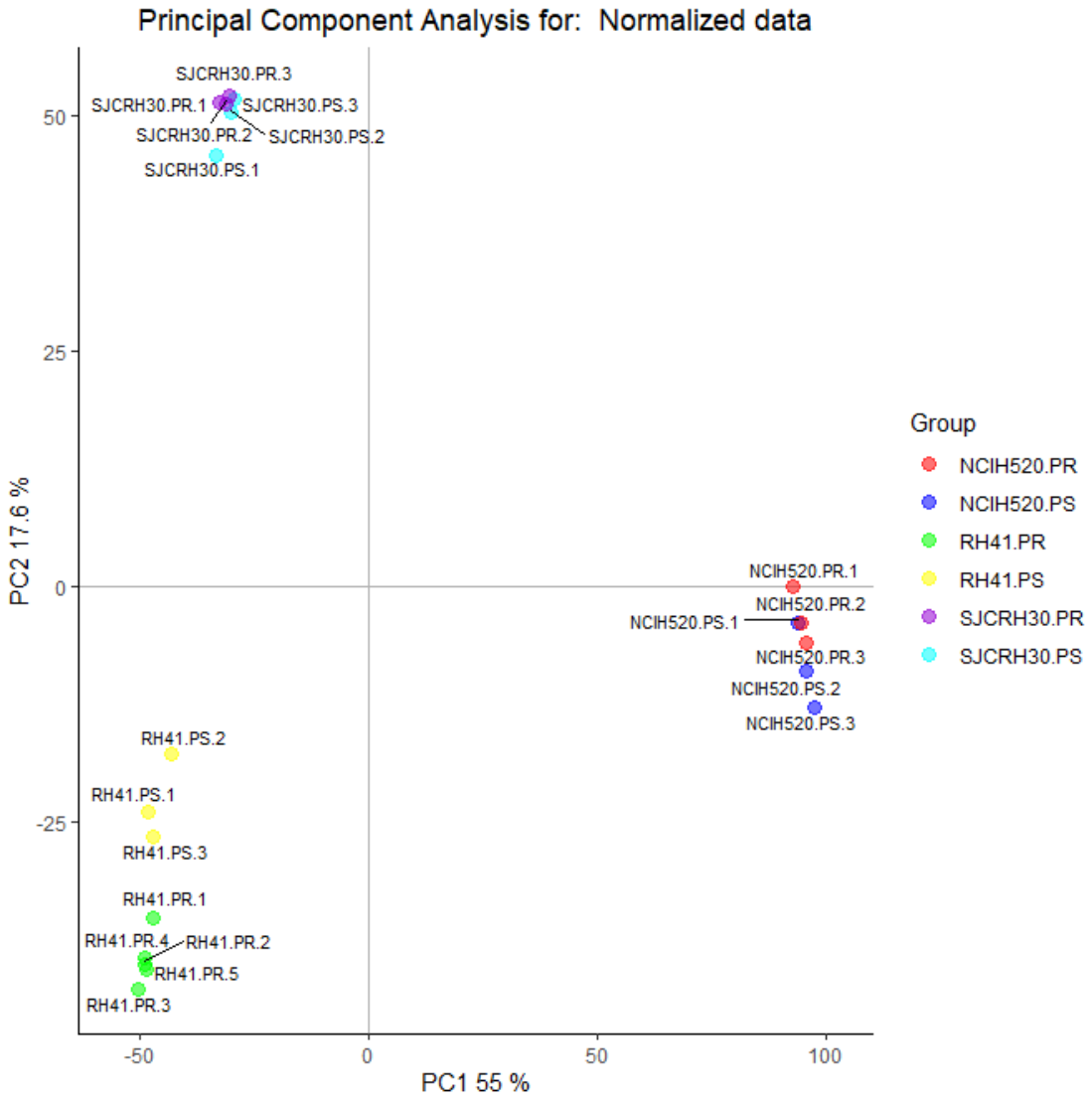
Control de calidad de los datos normalizados

Después de normalizar los datos, se realiza de nuevo el control de calidad para observar los datos como se hizo anteriormente.

	array	sampleNames	*1	*2	*3	Group	CellLine	Response	ShortName
☐	1	RH41.PS.1				RH41.PS	RH41	PS	RH41.PS.1
☐	2	RH41.PS.2				RH41.PS	RH41	PS	RH41.PS.2
☐	3	RH41.PS.3				RH41.PS	RH41	PS	RH41.PS.3
☐	4	RH41.PR.1				RH41.PR	RH41	PR	RH41.PR.1
☐	5	RH41.PR.2				RH41.PR	RH41	PR	RH41.PR.2
☐	6	RH41.PR.3				RH41.PR	RH41	PR	RH41.PR.3
☐	7	RH41.PR.4				RH41.PR	RH41	PR	RH41.PR.4
☐	8	RH41.PR.5				RH41.PR	RH41	PR	RH41.PR.5
☐	9	NCIH520.PS.1				NCIH520.PS	NCIH520	PS	NCIH520.PS.1
☐	10	NCIH520.PS.2				NCIH520.PS	NCIH520	PS	NCIH520.PS.2
☐	11	NCIH520.PS.3				NCIH520.PS	NCIH520	PS	NCIH520.PS.3
☐	12	NCIH520.PR.1				NCIH520.PR	NCIH520	PR	NCIH520.PR.1
☐	13	NCIH520.PR.2				NCIH520.PR	NCIH520	PR	NCIH520.PR.2
☐	14	NCIH520.PR.3				NCIH520.PR	NCIH520	PR	NCIH520.PR.3
☐	15	SJCRH30.PS.1				SJCRH30.PS	SJCRH30	PS	SJCRH30.PS.1
☐	16	SJCRH30.PS.2				SJCRH30.PS	SJCRH30	PS	SJCRH30.PS.2
☐	17	SJCRH30.PS.3				SJCRH30.PS	SJCRH30	PS	SJCRH30.PS.3
☐	18	SJCRH30.PR.1				SJCRH30.PR	SJCRH30	PR	SJCRH30.PR.1
☐	19	SJCRH30.PR.2				SJCRH30.PR	SJCRH30	PR	SJCRH30.PR.2
☐	20	SJCRH30.PR.3				SJCRH30.PR	SJCRH30	PR	SJCRH30.PR.3

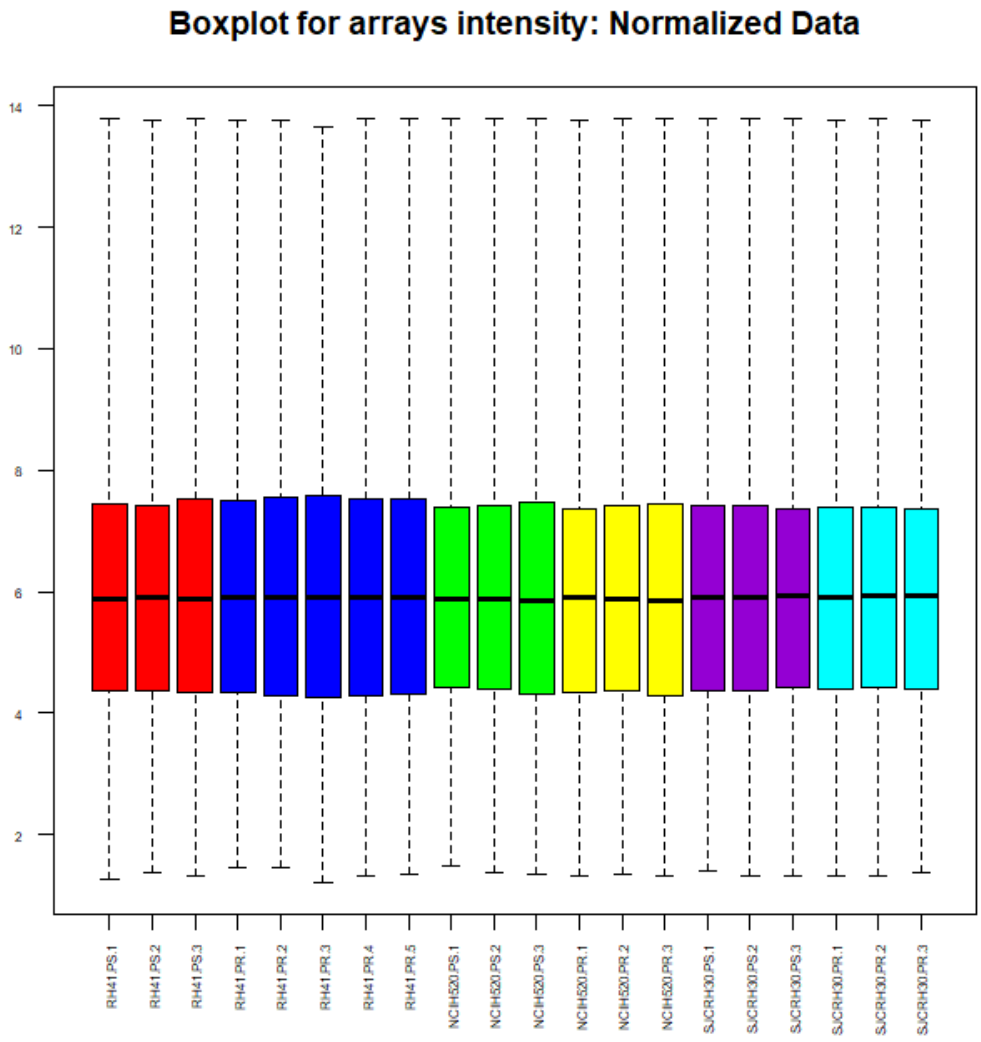
Aspecto de la tabla de resumen, en el archivo index.html, producido por el paquete arrayQualityMetrics en datos normalizados

A diferencia de los datos anteriores, ya no se resalta una variablidad en alguna de las tres categorías. Así, las matrices están más acordes para entrar en el análisis de expresión diferencial.



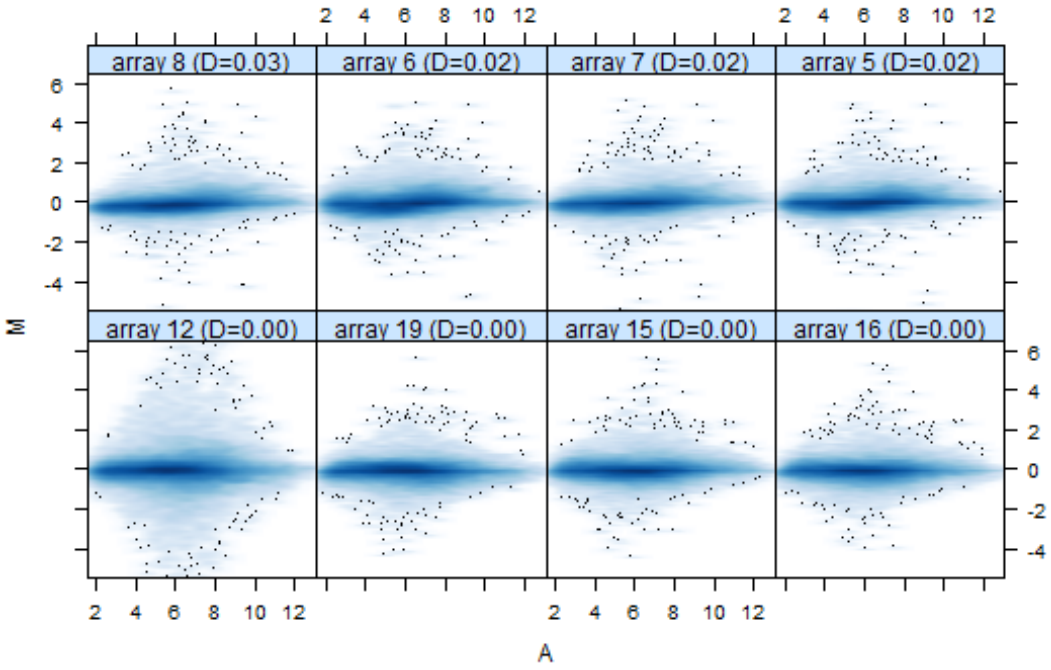
Visualización de los dos primeros componentes principales para datos normalizados

El primer componente representa el 55% de la variabilidad total, mientras que el segundo representa un 17.6%. Si lo comparamos con los datos sin procesar, el primer componene aumentó, mientras que el segundo ha disminuido, conglomerando tres grupos en el nivel de línea celular.



Distribución de intensidades para datos normalizados

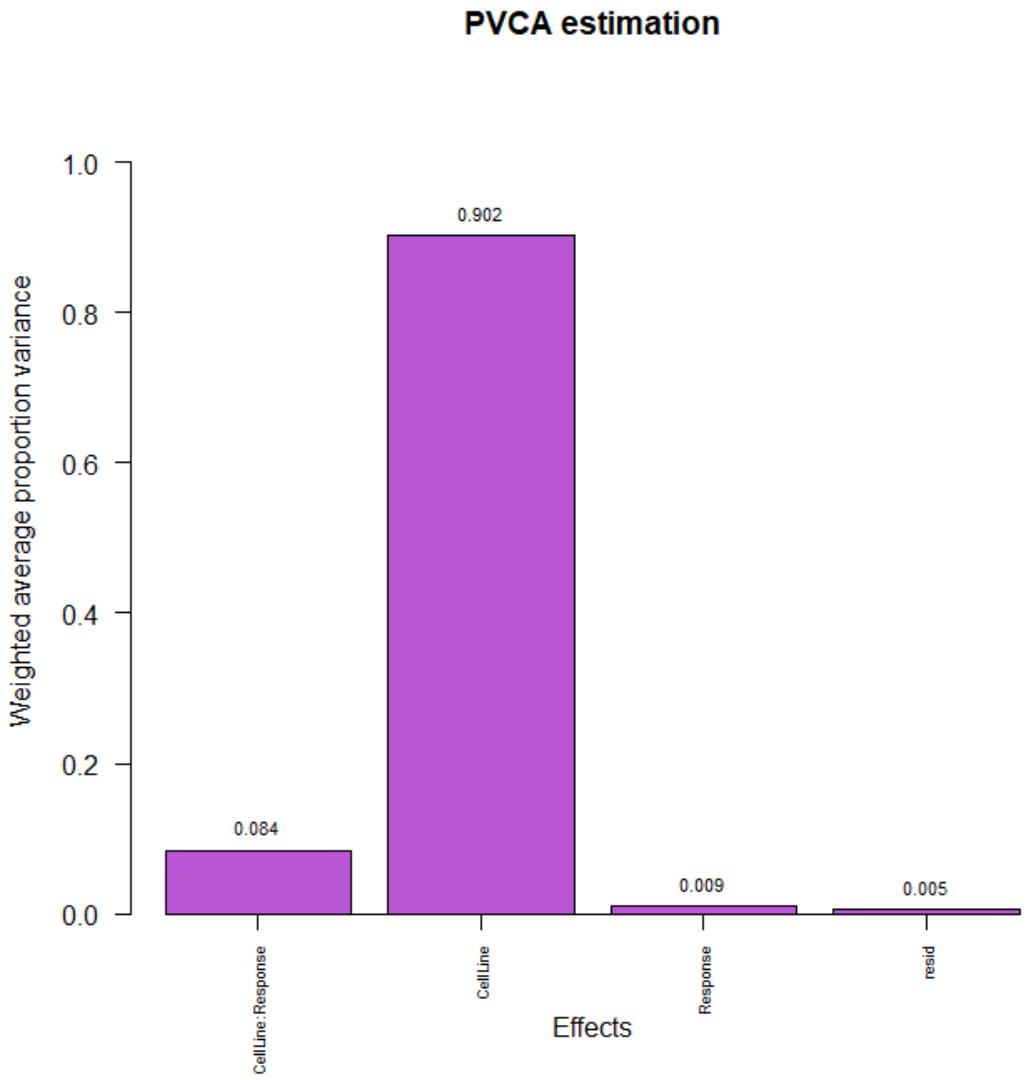
Los datos normalizados, ya muestran unas cajas con la misma distribución y el mismo ancho.



Lo mismo ocurre con el *MAPlots*, teniendo una distribución de las masas en *M=0* y un rango igual a *A*, indicando que las matrices poseen la misma intensidad de fondo.

Detección de lotes.

Los efectos por lotes se presentan en los datos de microarrays. El enfoque de PVCA, se usa como herramienta de detección para encontrar qué fuentes de variabilidad ya sea biológica, técnica u otra, son más prominentes en un conjunto de datos de microarrays dado Bushel and Li (2013).

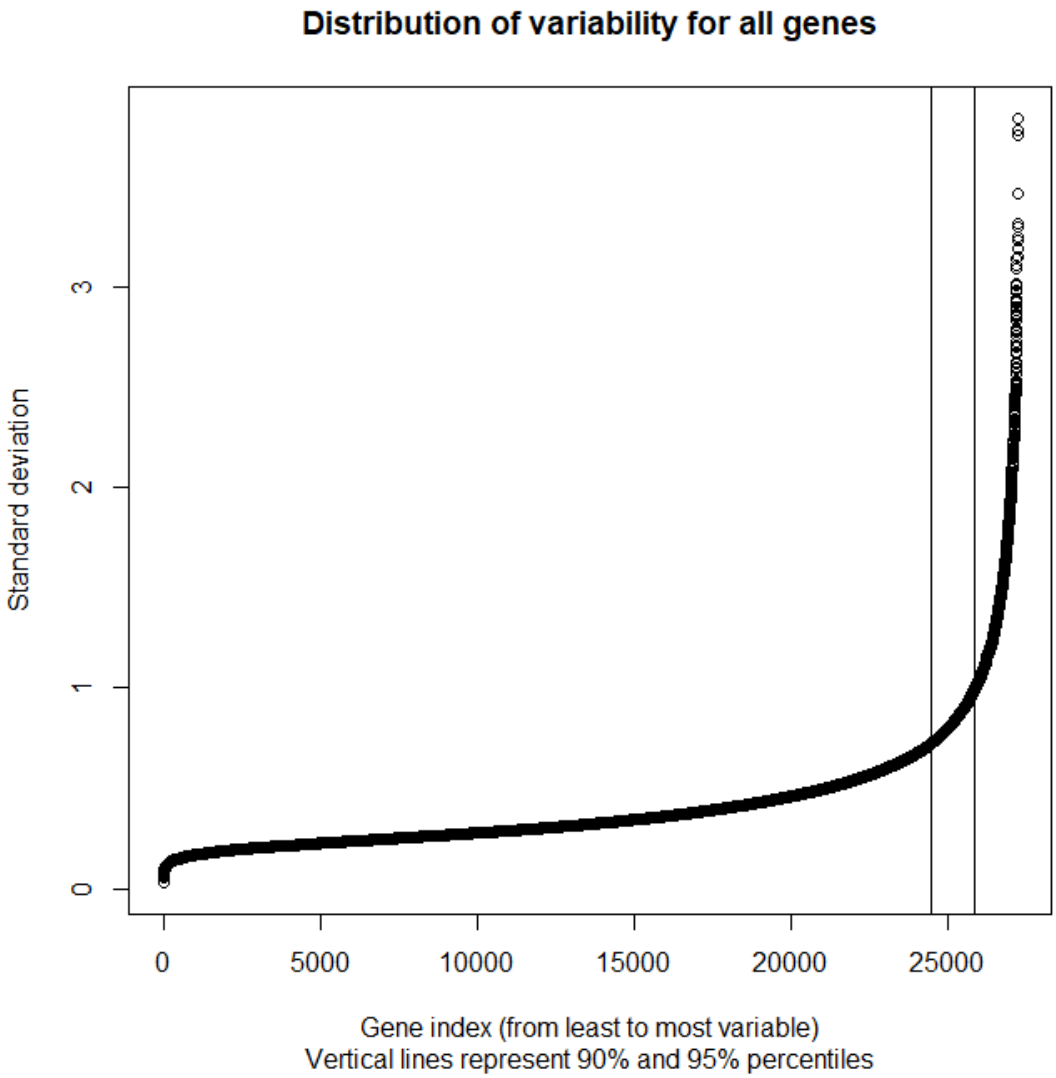


Importancia relativa de los diferentes factores -genotipo, que afectan la expresión génica

La figura indica un diagrama, donde cada barra es una fuente de variación incluida en el análisis. La principal fuente de variación es *CellLine*, lo cual también se observó en el PCA de datos sin procesar y normalizados. Es importante resaltar, que no es un factor lote, sino un factor experimental.

Detectar genes más variables.

Cuando un gen se expresa de manera diferencial, su varianza es mayor que la de aquellos genes que no poseen una expresión diferencial.



Los valores de las desviaciones estándar abarcan todas las muestras para todos los genes ordenados de menor a mayor

Los genes de mayor variabilidad son aquellos que están con una desviación estándar superior al 90-95% de todas las desviaciones estándar.

Filtrar genes con menos variables

El filtrado de genes, es una práctica común no solo porque puede aumentar nuestra confianza en los genes descubiertos expresados diferencialmente, sino también porque puede aumentar el número total de estos en un experimento. Así como como también permite identificar aquellos genes que su variabilidad fue aleatoria y no se esperaba que se expresaran diferencialmente Bourgon, Gentleman, and Huber (2010).

```
$numDupsRemoved
[1] 86
```

```
$numLowVar
[1] 13890

$numRemoved.ENTREZID
[1] 8583
```

Después de filtrar, quedan 4630 genes. Tener en cuenta que los genes almacenados queda en a variable *eset_filtered*

Guardar datos normalizados y filtrados

Los datos filtrados normalizados son el punto de partida para los análisis. Guardarlos en este punto es clave, por si se quiere revisar algunos valores de expresión génica específicos.

Matriz de diseño

La matriz de diseño se puede tomar de una variable de factor introducida en el archivo *targets*, este fue el objetivo de su creación. Para el presente estudio, la variable Group es una combinación de las condiciones "RH41, NCIH520, SJCRH30" y "PS, PR", que se representan simultáneamente como un factor de seis niveles. Como resultado, se tiene una matriz de 20x6.

Definiendo comparaciones con la Matriz de Contrastes

Con la matriz de contraste, se hizo una comparación entre grupos. El número de columnas es igual a la cantidad de comparaciones y el número de filas es igual a la cantidad de grupos. El "1" y "-1" están en las filas de los grupos a comparar y el "0" el resto. En este estudio se quiere comparar la expresión diferencial génica entre líneas celulares por separado (RH41, NCIH520, SJCRH30) para la resistencia a prexasertib o la sensibilidad a éste.

```
            Contrasts
Levels       NCIH520.PRvsNCIH520.PS RH41.PRvsRH41.PS
SJCRH30.PRvsSJCRH30.PS INT
  NCIH520.PR                      1               0
0   1
  NCIH520.PS                     -1               0
0  -1
  RH41.PR                         0               1
0  -1
  RH41.PS                         0              -1
0   1
  SJCRH30.PR                      0               0
1  -1
  SJCRH30.PS                      0               0
-1   1
```

Estimación del modelo y selección de genes

Ya definida la matriz de diseño y los contrastes, se estima el modelo y los contrastes para realizar las pruebas de significación, que permitirán decidir si cada gen en comparación puede considerarse expresados diferencialmente. El análisis contempla loas valores-p ajustados, para ordenar los genes del que más se exprese al de menor expresión diferencial.

Los falsos positivos se controlan, ajustando el valor-p para tener control sobre la tasa de falsos positivos utilizados.

```
[1] "MArrayLM"
attr(,"package")
[1] "limma"
```

Obtención de listas de genes expresados diferencialmente

La lista de genes diferencialmente expresados se obtiene ordenada, desde el valor-p más pequeño al más grande y que se puede considerar como más o menos expresado diferencialmente. En cada gen se obtiene los siguientes resultados:

- `logFC`: diferencia media entre grupos.
- `AveExpr`: expresión promedio de todos los genes en la comparación.
- `t`: estadística t moderada.
- `P.Value`: prueba p - valor.
- `adj.P.Val`: valor p ajustado
- `B`: estadística B: probabilidad del registro posterior del gen del ser vs no ser diferencialmente expresado.

Para la comparación 1 (NCIH520.PRvsNCIH520.PS): Genes que cambian su expresión en la línea celular NCIH520 entre los resistentes y sensibles a prexasertib.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
TC1000006891.hg.1	-3.368062	9.185505	-29.20949	0	0	30.64423
TC0500013261.hg.1	-2.823415	7.719235	-28.14823	0	0	29.97636
TC0100017947.hg.1	-3.667875	5.722903	-27.60971	0	0	29.62613
TC1000007199.hg.1	-3.488094	8.245939	-27.16069	0	0	29.32802

TC0800011064.hg.1	4.304902	5.561647	24.24205	0	0	27.24336
TC1600011574.hg.1	2.273131	7.669633	23.95481	0	0	27.02325

Para la comparación 2 (RH41.PRvsRH41.PS): Genes que cambian su expresión en la línea celular RH41 entre los resistentes y sensibles a prexasertib.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
TC0X00006631.hg.1	4.401999	10.64726	36.71538	0	0	34.90320
TC1300009980.hg.1	2.938994	7.497045	27.68956	0	0	29.78471
TC0200015242.hg.1	2.829245	9.844361	27.44198	0	0	29.61804
TC0300011172.hg.1	3.347105	7.147172	27.31740	0	0	29.53354
TC0200015607.hg.1	3.057394	6.325555	26.22498	0	0	28.77374
TC0100013445.hg.1	4.194641	8.674719	26.15551	0	0	28.72425

Para la comparación 3 (SJCRH30.PRvsSJCRH30.PS): Genes que cambian su expresión en la línea celular SJCRH30 entre los resistentes y sensibles a prexasertib.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
TC1600009217.hg.1	-4.987123	5.159140	-27.44537	0	0	28.96612
TC0100017844.hg.1	5.465643	6.092721	25.53992	0	0	27.75290
TC0X00010207.hg.1	-3.349400	4.828032	-18.73943	0	0	22.32946
TC0100013534.hg.1	3.877785	6.480764	17.71833	0	0	21.32699
TC0500010615.hg.1	-3.100156	4.960280	-16.61152	0	0	20.17015
TC1700008123.hg.1	-3.21816	7.599251	-16.2811	0	0	19.80968

5 8

Para la comparación 4 (INT): Genes que se comportan de manera diferente entre la comparación 1, 2 y 3.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
TC0500008736.hg.1	-5.875602	5.075763	-26.22580	0	0	28.25364
TC0500013261.hg.1	-3.363305	7.719235	-20.03842	0	0	23.54676
TC0100017947.hg.1	-4.376930	5.722903	-19.68964	0	0	23.23326
TC1800007963.hg.1	-5.032156	6.007954	-19.60975	0	0	23.16060
TC0X00006631.hg.1	-4.111411	10.64726 1	-18.32968	0	0	21.95089
TC1600011398.hg.1	-3.067991	8.364270	-18.23695	0	0	21.85979

La primera columna de cada topTab contiene la identificación del fabricante (Affrymetix) para cada conjunto de sondas. El paso siguiente es saber qué gen corresponde a cada ID de Affymetrix, o sea se lleva a cabo la anotación.

Anotación de genes

En este punto, lo que se busca es información asociada a los identificadores que aparecen en cada topTab, generalmente correspondientes a sondas o transcripciones que dependen del tipo de matriz, con nombres más familiares como el Símbolo del gen, el Identificador del gen Entrez o la descripción del gen.

La anotación permite que las tablas sean más comprensibles. El siguiente resultado es sólo una muestra de cómo se observan las anotaciones para la comparación NCIH520.PRvsNCIH520.PS (Se tendrán en cuenta sólo las primeras cuatro columnas).

Annotations added to results “topTable” for the comparison “NCIH520.PRvsNCIH520.PS”
PROBEID
SYMBOL
ENTREZID

GENENAME
TC0100006494.hg.1
B3GALT6
126792
beta-1,3-galactosyltransferase 6
TC0100006550.hg.1
PRKCZ
5590
protein kinase C zeta
TC0100006565.hg.1
SKI
6497
SKI proto-oncogene
TC0100006698.hg.1
PHF13
148479
PHD finger protein 13
TC0100006761.hg.1
CA6
765
carbonic anhydrase 6

	PROBEID	SYMBOL	ENTREZID	GENENAME
1	TC0100006494.hg.1	B3GALT6	126792	beta-1,3-galactosyltransferase 6
2	TC0100006550.hg.1	PRKCZ	5590	protein kinase C zeta
3	TC0100006565.hg.1	SKI	6497	SKI proto-oncogene
4	TC0100006698.hg.1	PHF13	148479	PHD finger protein 13
5	TC0100006761.hg.1	CA6	765	carbonic anhydrase 6

Visualizando la expresión diferencial

El gráfico de volcán es una plot que ordena los genes en dos ejes, el eje x representado por *Fold Change* que es la dimensión biológica y el *P-value* que es la dimensión estadística. Por ende el eje horizontal nos indica el impacto biológico del cambio y el eje vertical la fiabilidad del cambio.

Cada diagrama de volcán muestra la confrontación entre las matrices, basados en el diseño experimental. Los nombres de los 10 genes con mayor epresión diferencial en cada comparación, se encuentran en la gráfica.

Comparaciones múltiples

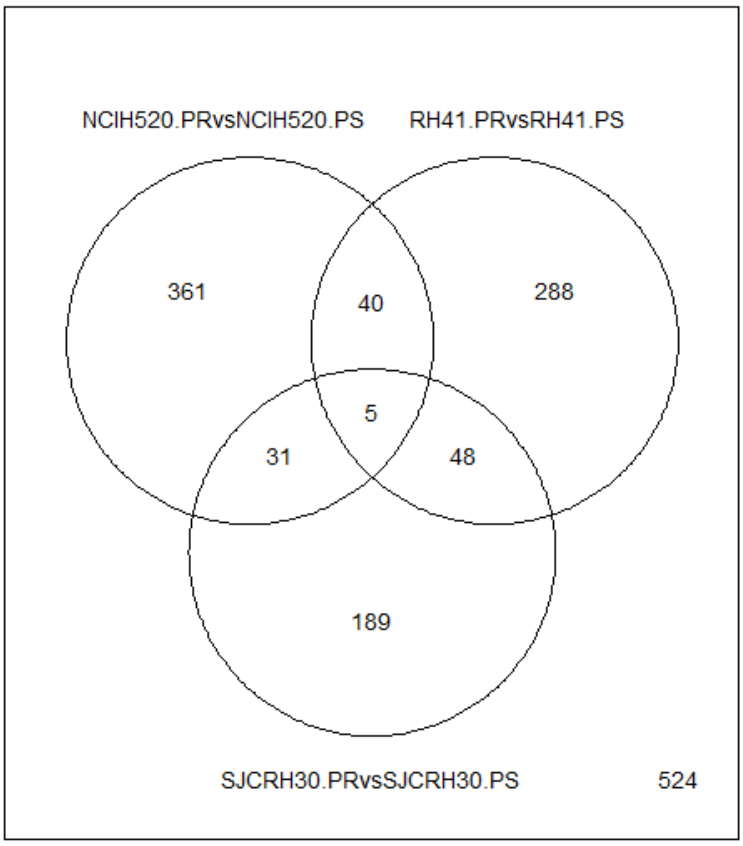
Cuando uno selecciona genes en varias comparaciones, generalmente es interesante saber qué genes se han seleccionado en cada comparación. A veces, los genes biológicamente relevantes serán aquellos que se seleccionan en uno de ellos pero no

en otros. En otras ocasiones, su interés radicará en los genes que se seleccionan en todas las comparaciones Gonzalo and Sanchez-Pla (2020).

Este objeto tiene tantas columnas como comparaciones y tantas filas como genes. Por cada gen y comparación, un "+1" denota una regulación significativamente alta (valores de la prueba t > 0, FDR < punto de corte seleccionado), un "-1" significativamente baja (valores de la prueba t < 0, FDR < corte seleccionado) y un "0" un diferencia no significativa (FDR > corte seleccionado).

```
       NCIH520.PRvsNCIH520.PS RH41.PRvsRH41.PS SJCRH30.PRvsSJCRH30.PS
INT
Down                      187              85                    153
692
NotSig                   4193            4249                   4357
3431
Up                        250             296                    120
507
```

En la fila *DOWN* nos apareceran todos los genes downregulated, mientras que la fila *UP* nos indica todos los genes upregulated.

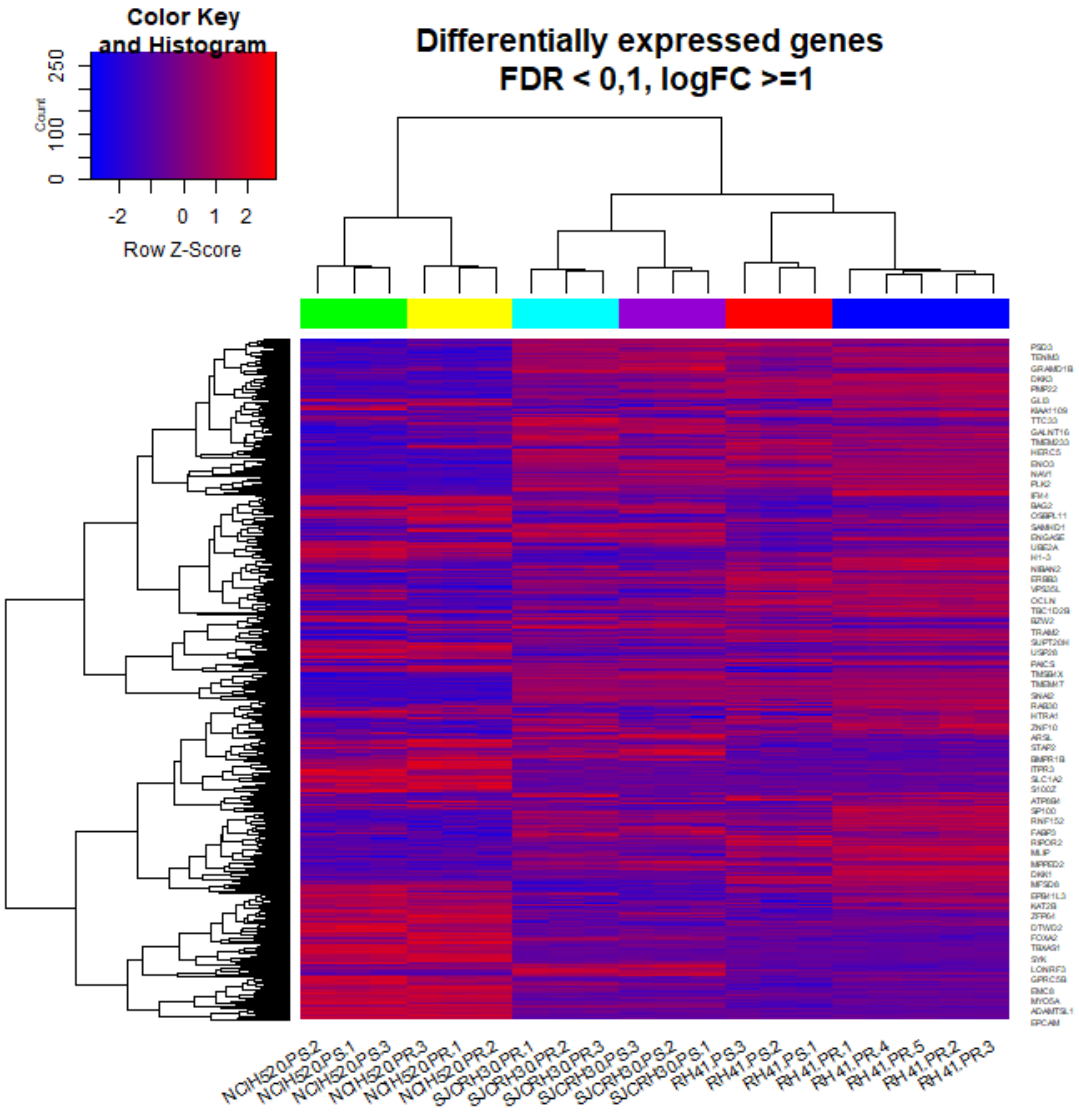


Venn diagram showing the genes in common between the three comparisons performed

Con el diagrama de Venn, se puede observar el número de genes seleccionados en cada comparación, y cuándtos hay en común entre las comparaciones dos a dos y o con las tres.

Heatmaps

Aquellos genes que se seleccionaron como diferencialmente expresados, se logran visualizar en un mapa de calor. Estos mapas no muestran un orden específico, pero generalmente se prefiere trazarlos haciendo un agrupamiento jerárquico en genes (filas) o columnas (muestras) para encontrar grupos de genes con patrones comunes de variación que eventualmente puede asociarse a los diferentes grupos que se comparan Gonzalo and Sanchez-Pla (2020).



Mapa de calor para expresión de datos que agrupan genes (filas) y muestras (columnas) por su similitud

Mapa de calor producido para todos los genes seleccionados con los mismos criterios descritos anteriormente (FDR <0.1 y logFC> 1) donde los genes y las muestras se ven obligados a agruparse por fila y columna de forma similar.

Significado biológico de los resultados

El análisis de conjuntos de genes es una herramienta valiosa para resumir datos de expresión génica de alta dimensión en términos de conjuntos biológicamente relevantes. Esta es un área activa de investigación y se han desarrollado numerosos métodos de análisis de conjuntos de genes. A pesar de esta popularidad, los estudios comparativos sistemáticos han sido de alcance limitado Alex (n.d.).

Con la lista de genes diferencialmente expresados entre dos condiciones, se busca establecer si las funciones, procesos biológicos o vías moleculares que los caracterizan aparecen en esta lista con más frecuencia que entre el resto de los genes analizado Gonzalo and Sanchez-Pla (2020), determinando diferencias concordantes estadísticamente significativas entre dos estados biológicos (por ejemplo, fenotipos).

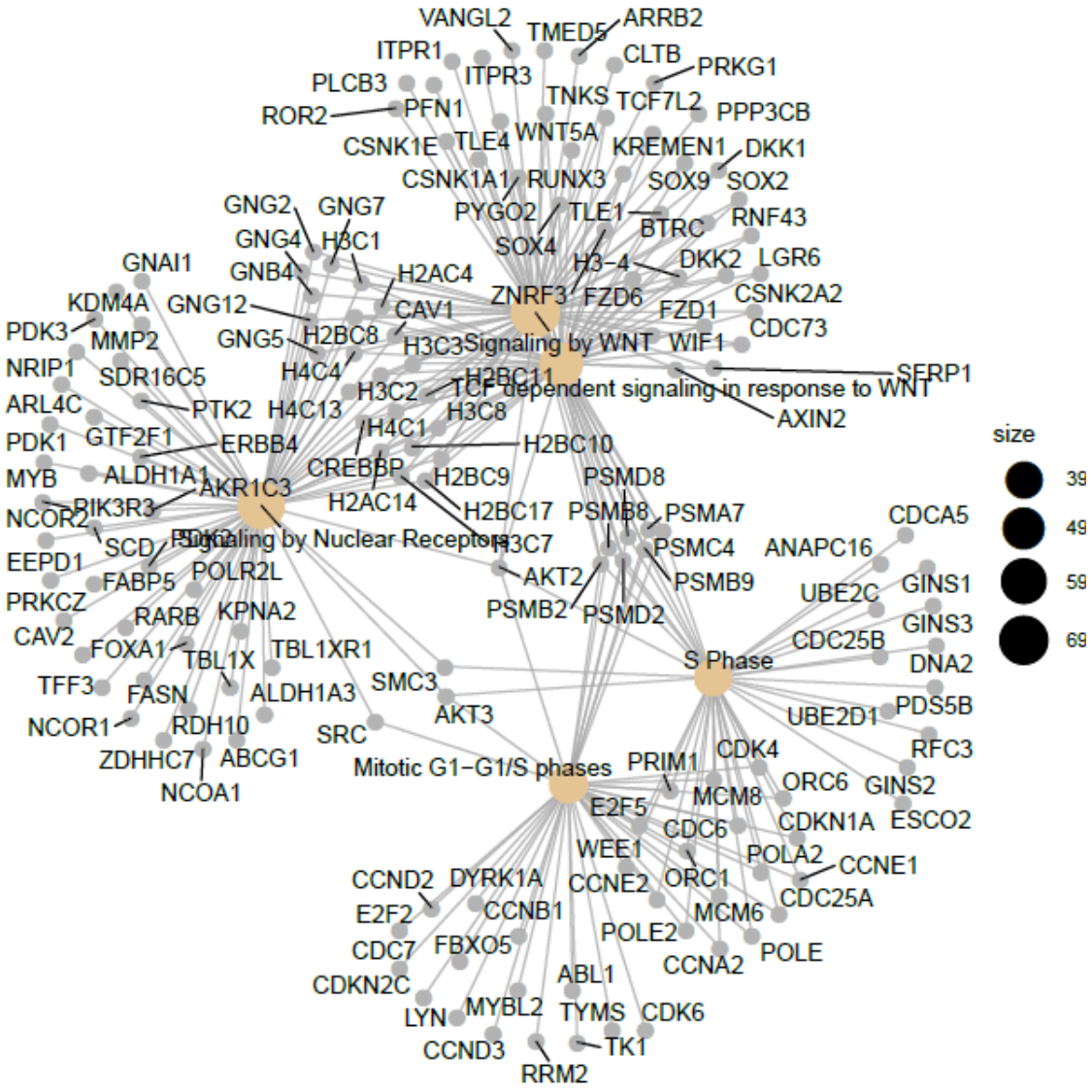
```
NCIH520.PRvsNCIH520.PS       RH41.PRvsRH41.PS SJCRH30.PRvsSJCRH30.PS
                  2359                   2568                    2237
                   INT
                  2565
```

El análisis también requiere tener los identificadores de Entrez para todos los genes analizados. Es una discusión abierta si lo que se debe usar es “todos los genes analizados”, es decir, los genes que se han retenido en el análisis y son parte de la “topTab”, o todos los genes disponibles. En este caso, se utiliza la segunda opción y se define el universo como todos los genes que tienen al menos una anotación en la ontología genética Gonzalo and Sanchez-Pla (2020).

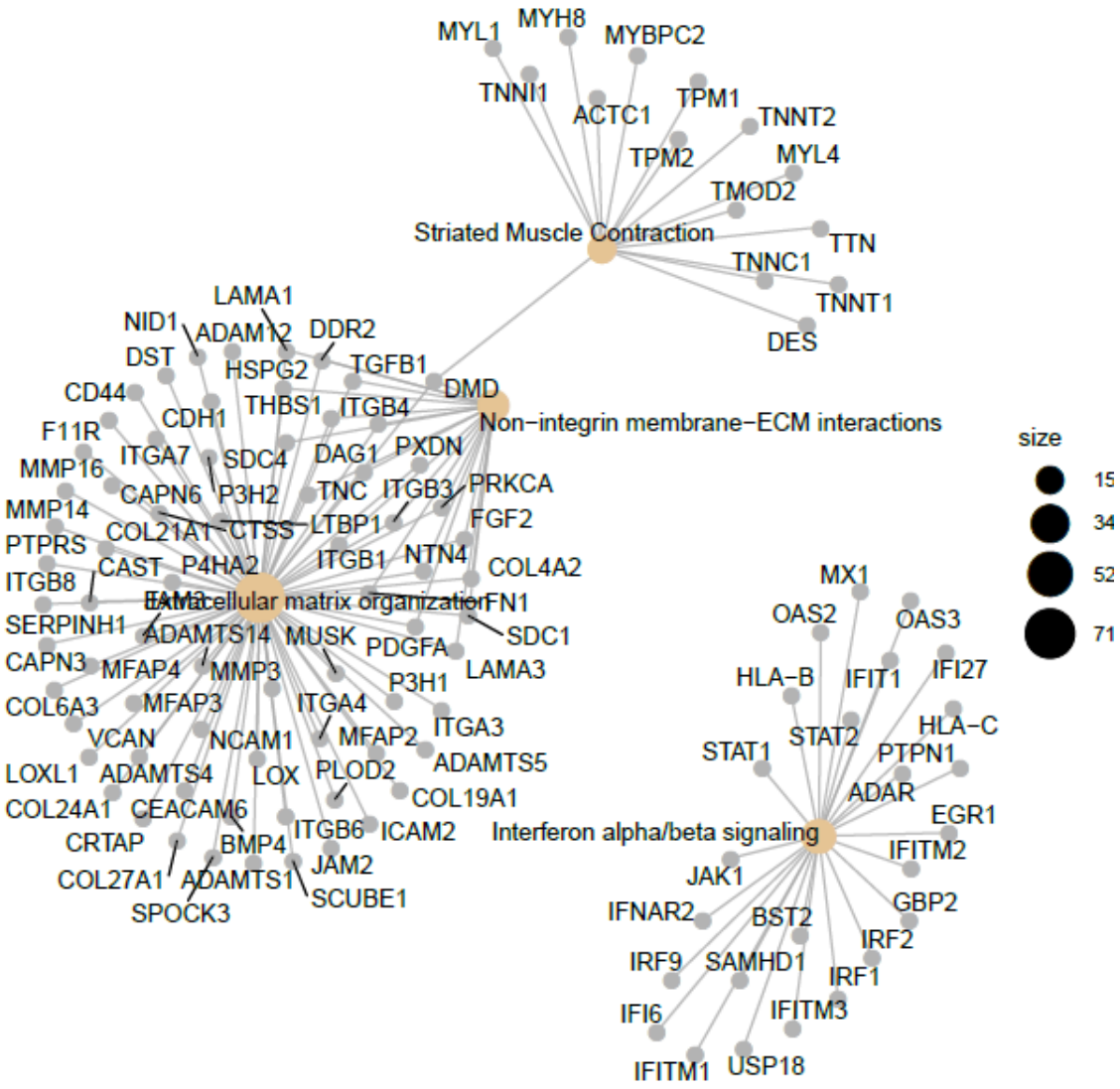
```
###################################
Comparison:  NCIH520.PRvsNCIH520.PS
                         ID                                  Description
R-HSA-453279   R-HSA-453279                      Mitotic G1-G1/S phases
R-HSA-195721   R-HSA-195721                            Signaling by WNT
R-HSA-201681   R-HSA-201681 TCF dependent signaling in response to WNT
R-HSA-9006931 R-HSA-9006931             Signaling by Nuclear Receptors
R-HSA-69242     R-HSA-69242                                     S Phase
R-HSA-3000171 R-HSA-3000171     Non-integrin membrane-ECM interactions
              GeneRatio   BgRatio       pvalue    p.adjust      qvalue
R-HSA-453279    42/1429 149/10616 1.478428e-06 0.001964831 0.001739876
R-HSA-195721    69/1429 329/10616 8.836369e-05 0.036902939 0.032677903
R-HSA-201681    52/1429 231/10616 1.012164e-04 0.036902939 0.032677903
R-HSA-9006931   63/1429 298/10616 1.393924e-04 0.036902939 0.032677903
R-HSA-69242     39/1429 161/10616 1.465070e-04 0.036902939 0.032677903
R-HSA-3000171   19/1429  59/10616 1.666047e-04 0.036902939 0.032677903

geneID
R-HSA-453279
CCNE2/RRM2/AKT3/TK1/CDC25A/CDC7/CDC6/MYBL2/CDKN1A/TYMS/FBXO5/WEE1/LYN/MCM
6/POLE2/ORC6/SRC/POLA2/ORC1/E2F5/CCNB1/AKT2/CCNA2/CCNE1/PSMB9/DYRK1A/CCND
3/PRIM1/E2F2/CDKN2C/POLE/PSMD8/CCND2/CDK4/ABL1/PSMB2/PSMB8/PSMD2/PSMC4/PS
MA7/MCM8/CDK6
R-HSA-195721
DKK2/ITPR1/WIF1/H2BC8/H4C13/WNT5A/LGR6/H3C1/GNG4/GNG12/TLE1/H2AC4/H4C1/DK
K1/GNB4/H3C2/FZD6/H2BC11/SOX9/AKT2/ITPR3/H2BC9/TLE4/CSNK2A2/H2BC17/PSMB9/
PFN1/GNG2/KREMEN1/CSNK1E/H3C8/SFRP1/PSMD8/GNG7/TCF7L2/ZNRF3/H3C7/RUNX3/H3
C3/PSMB2/H2AC14/GNG5/FZD1/PSMB8/CREBBP/PLCB3/H4C4/H2BC10/PSMD2/BTRC/RNF43
/PSMC4/CLTB/TNKS/PYGO2/PSMA7/PRKG1/ROR2/ARRB2/VANGL2/SOX2/CSNK1A1/AXIN2/T
MED5/CAV1/H3-4/SOX4/CDC73/PPP3CB
R-HSA-201681
DKK2/WIF1/H2BC8/H4C13/WNT5A/LGR6/H3C1/TLE1/H2AC4/H4C1/DKK1/H3C2/FZD6/H2BC
```

```
11/SOX9/AKT2/H2BC9/TLE4/CSNK2A2/H2BC17/PSMB9/KREMEN1/CSNK1E/H3C8/SFRP1/PS
MD8/TCF7L2/ZNRF3/H3C7/RUNX3/H3C3/PSMB2/H2AC14/FZD1/PSMB8/CREBBP/H4C4/H2BC
10/PSMD2/BTRC/RNF43/PSMC4/TNKS/PYGO2/PSMA7/SOX2/CSNK1A1/AXIN2/CAV1/H3-
4/SOX4/CDC73
R-HSA-9006931
MMP2/H2BC8/AKT3/KPNA2/H4C13/PDK3/KDM4A/H3C1/GNG4/TFF3/ZDHHC7/PDK1/ERBB4/G
NG12/AKR1C3/FOXA1/SDR16C5/SRC/H2AC4/H4C1/SCD/RDH10/ALDH1A3/GNB4/H3C2/H2BC
11/FASN/AKT2/H2BC9/H2BC17/TBL1X/PDK2/NCOR1/GNAI1/GNG2/SMC3/ALDH1A1/MYB/H3
C8/RARB/NCOR2/POLR2L/NRIP1/GNG7/CAV2/GTF2F1/H3C7/FABP5/H3C3/H2AC14/NCOA1/
GNG5/CREBBP/H4C4/EEPD1/H2BC10/ARL4C/PRKCZ/TBL1XR1/ABCG1/PIK3R3/CAV1/PTK2
R-HSA-69242
GINS2/CCNE2/AKT3/CDC25A/GINS1/CDC6/CDKN1A/RFC3/CDCA5/WEE1/UBE2C/MCM6/ESCO
2/POLE2/ORC6/POLA2/ORC1/CDC25B/E2F5/GINS3/AKT2/CCNA2/CCNE1/PSMB9/DNA2/PRI
M1/SMC3/POLE/PSMD8/UBE2D1/ANAPC16/PDS5B/CDK4/PSMB2/PSMB8/PSMD2/PSMC4/PSMA
7/MCM8
R-HSA-3000171
FN1/SDC4/THBS1/COL5A1/NTN4/PDGFA/DMD/ITGB1/TGFB1/DAG1/LAMB1/COL4A1/SDC1/T
NC/LAMA2/ITGAV/SDC3/DDR2/COL5A2
             Count
R-HSA-453279     42
R-HSA-195721     69
R-HSA-201681     52
R-HSA-9006931    63
R-HSA-69242      39
R-HSA-3000171    19

###################################
Comparison:  RH41.PRvsRH41.PS
                        ID                             Description
GeneRatio
R-HSA-909733   R-HSA-909733        Interferon alpha/beta signaling
25/1542
R-HSA-1474244 R-HSA-1474244      Extracellular matrix organization
71/1542
R-HSA-390522   R-HSA-390522            Striated Muscle Contraction
15/1542
R-HSA-3000171 R-HSA-3000171 Non-integrin membrane-ECM interactions
20/1542
                BgRatio       pvalue   p.adjust      qvalue
R-HSA-909733   69/10616 6.017386e-06 0.00805728 0.007645248
R-HSA-1474244 301/10616 1.516641e-05 0.01015391 0.009634659
R-HSA-390522   36/10616 6.879463e-05 0.03070534 0.029135130
R-HSA-3000171  59/10616 1.474382e-04 0.04935493 0.046831022

geneID
R-HSA-909733
STAT1/IFI6/IFIT1/IFITM1/IRF9/SAMHD1/USP18/IFITM3/JAK1/IRF1/IRF2/BST2/IFIT
M2/ADAR/OAS3/GBP2/OAS2/IFI27/IFNAR2/EGR1/STAT2/HLA-C/PTPN1/HLA-B/MX1
R-HSA-1474244
ITGB8/SDC1/ITGA3/NID1/CTSS/MMP16/COL19A1/MUSK/CAST/THBS1/LAMA1/ADAMTS4/MM
```

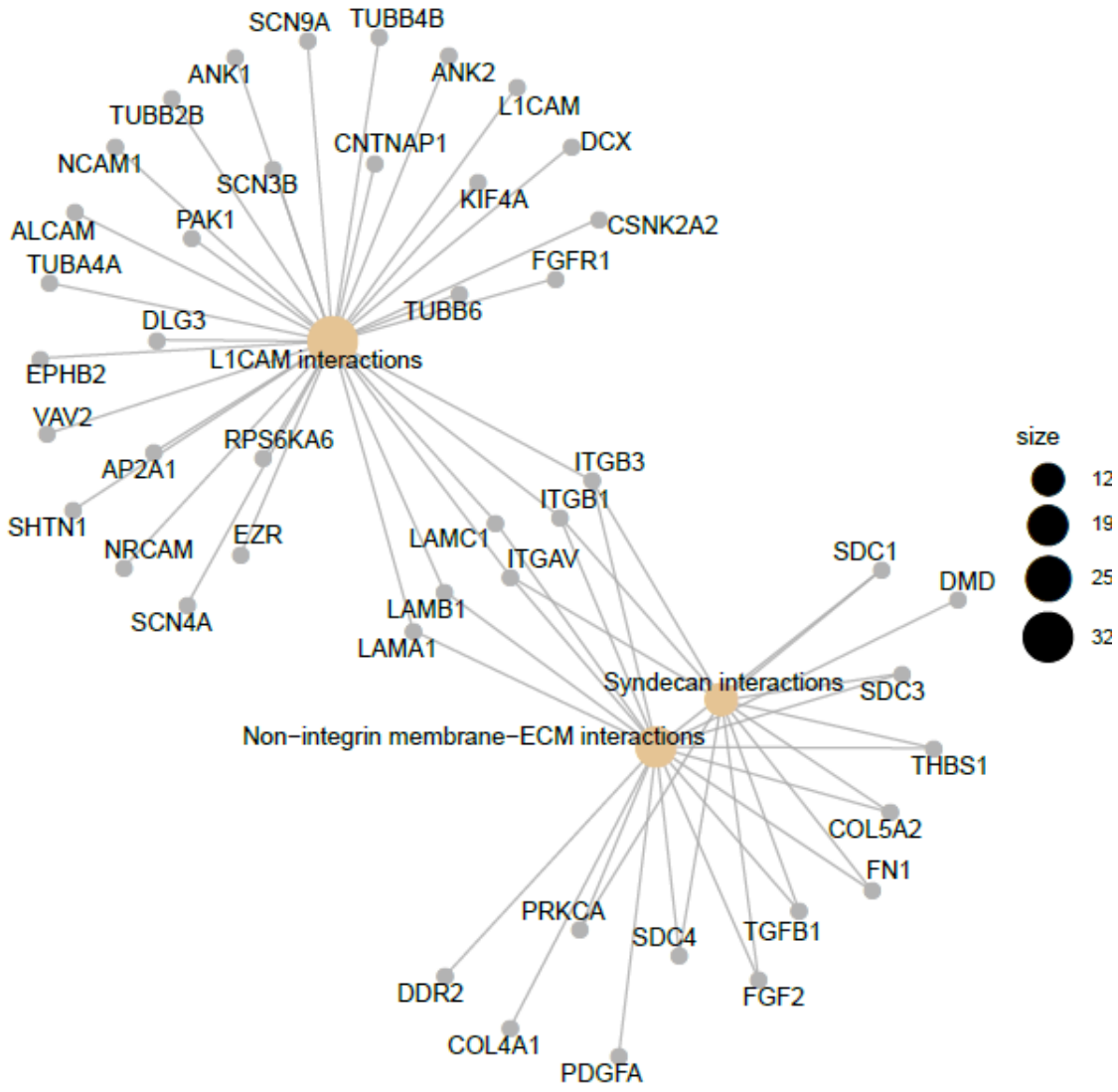
```
P3/ICAM2/JAM2/MFAP3/TNC/COL21A1/P3H1/CRTAP/DST/JAM3/DMD/ITGB6/SPOCK3/PRKC
A/LOX/P4HA2/PXDN/LAMA3/TGFB1/NCAM1/COL6A3/DAG1/ADAM12/FGF2/ADAMTS1/BMP4/M
MP14/DDR2/ADAMTS14/HSPG2/PTPRS/NTN4/LOXL1/LTBP1/CAPN6/ITGA7/PLOD2/FN1/SDC
4/ITGB3/CAPN3/ITGA4/ADAMTS5/CD44/COL4A2/ITGB4/PDGFA/COL24A1/CEACAM6/ITGB1
/VCAN/SCUBE1/MFAP2/P3H2/F11R/CDH1/COL27A1/SERPINH1/MFAP4
R-HSA-390522
TNNC1/MYL1/TNNI1/TPM1/MYH8/TTN/TNNT2/MYL4/ACTC1/DES/DMD/MYBPC2/TNNT1/TPM2
/TMOD2
R-HSA-3000171
SDC1/THBS1/LAMA1/TNC/DMD/PRKCA/LAMA3/TGFB1/DAG1/FGF2/DDR2/HSPG2/NTN4/FN1/
SDC4/ITGB3/COL4A2/ITGB4/PDGFA/ITGB1
              Count
R-HSA-909733     25
R-HSA-1474244    71
R-HSA-390522     15
R-HSA-3000171    20

###################################
Comparison:  SJCRH30.PRvsSJCRH30.PS
                         ID                            Description
GeneRatio
R-HSA-373760   R-HSA-373760                      L1CAM interactions
32/1355
R-HSA-3000170 R-HSA-3000170                   Syndecan interactions
12/1355
R-HSA-3000171 R-HSA-3000171 Non-integrin membrane-ECM interactions
19/1355
                BgRatio       pvalue   p.adjust     qvalue
R-HSA-373760  119/10616 2.442901e-05 0.03244923 0.03146249
R-HSA-3000170  27/10616 4.875917e-05 0.03244923 0.03146249
R-HSA-3000171  59/10616 8.111225e-05 0.03598680 0.03489250

geneID
R-HSA-373760
DCX/ITGB3/LAMC1/ANK1/TUBB2B/SCN9A/PAK1/DLG3/NRCAM/CSNK2A2/EPHB2/L1CAM/KIF
4A/CNTNAP1/FGFR1/TUBB4B/ITGAV/TUBA4A/LAMB1/SCN3B/ANK2/RPS6KA6/TUBB6/SCN4A
/VAV2/NCAM1/ALCAM/SHTN1/LAMA1/EZR/ITGB1/AP2A1
R-HSA-3000170
ITGB3/FN1/THBS1/SDC4/ITGAV/SDC1/SDC3/TGFB1/PRKCA/COL5A2/FGF2/ITGB1
R-HSA-3000171
ITGB3/FN1/LAMC1/THBS1/SDC4/ITGAV/LAMB1/DMD/SDC1/SDC3/DDR2/TGFB1/PRKCA/COL
5A2/COL4A1/FGF2/LAMA1/PDGFA/ITGB1
              Count
R-HSA-373760     32
R-HSA-3000170    12
R-HSA-3000171    19
```



Red obtenida del análisis de enriquecimiento de Reactome en la lista obtenida de la comparación entre NCIH520.PRvsNCIH520.PS. Se encontraron cinco vías enriquecidas: *Signaling by WNT*, *Signaling by Nuclear Receptors*, *TCF dependent signaling in response to WNT*, *Mitotic G1–G1/S phases* y *S Phase*.



Red obtenida del análisis de enriquecimiento de Reactome en la lista obtenida de la comparación entre RH41.PRvsRH41.PS Se encontraron cuatro vías enriquecidas: *Extracellular matrix organization*, *Interferon alpha/beta signaling*, *Non−integrin membrane−ECM interactions* y *Striated Muscle Contraction*.



Red obtenida del análisis de enriquecimiento de Reactome en la lista obtenida de la comparación entre RH41.PRvsRH41.PS Se encontraron tres vías enriquecidas: *L1CAM interactions*, *Non−integrin membrane−ECM interactions* y *Syndecan interactions*.

Conclusiones

- Aunque todas las matrices poseen sus réplicas, es importante resaltar que la matriz RH41.PR posee dos más que el resto, o sea cinco. Quizás hubiese sido más simple haber tomado sólo tres, si al momento de observar el control de calidad, estos no poseían errores.
- Es importante realizar el control de datos, para determinar la calidad de los resultados y lograr hacer las comparaciones bajo las mismas condiciones.
- La agrupación de los resultados se obtiene a partir de la línea celular de donde se tomó.
- La matriz de contraste permite determinar cuales grupos serán comparados entre sí.

- Al momento de tener la lista de genes expresados diferencialmente, es importante realizar la anotación, que permite poner nombres más familiares a cada gen.
- La lista de genes diferencialmente expresados se obtiene desde el p-valor más pequeño hasta el más grande.
- Con el mapa de calor podemos ver aquellos genes que se seleccionaron como diferencialmente expresados.
- La comparación entre NCIH520.PRvsNCIH520.PS es la que más vías enriquecidas evidenció.

References

Alex, Sánchez. n.d. "An Introduction to Pathway Enrichment Analysis."

Blosser, Wayne D., Jack A. Dempsey, Ann M. McNulty, Xi Rao, Philip J. Ebert, Caitlin D. Lowery, Philip W. Iversen, et al. 2020. "A pan-cancer transcriptome analysis identifies replication fork and innate immunity genes as modifiers of response to the CHK1 inhibitor prexasertib." *Oncotarget* 11 (3): 216–36. https://doi.org/10.18632/oncotarget.27400.

Bourgon, Richard, Robert Gentleman, and Wolfgang Huber. 2010. "Independent filtering increases detection power for high-throughput experiments." *Proceedings of the National Academy of Sciences of the United States of America* 107 (21): 9546–51. https://doi.org/10.1073/pnas.0914005107.

Bushel, Pierre R, and Jianying Li. 2013. "Estimating batch effect in Microarray data with Principal Variance Component Analysis (PVCA) method," 1–5.

Gonzalo, Ricardo, and Alex Sanchez-Pla. 2020. "Statistical Analysis of Microarray data." https://github.com/ASPteaching/Omics{_}Data{_}Analysis-Case{_}Study{_}1-Microarrays.

Jones, R. M., O. Mortusewicz, I. Afzal, M. Lorvellec, P. García, T. Helleday, and E. Petermann. 2013. "Increased replication initiation and conflicts with transcription underlie Cyclin E-induced replication stress." *Oncogene* 32 (32): 3744–53. https://doi.org/10.1038/onc.2012.387.

Lowery, Caitlin D., Alle B. VanWye, Michele Dowless, Wayne Blosser, Beverly L. Falcon, Julie Stewart, Jennifer Stephens, Richard P. Beckmann, Aimee Bence Lin, and Louis F. Stancato. 2017. "The checkpoint kinase 1 inhibitor prexasertib induces regression of preclinical models of human neuroblastoma." *Clinical Cancer Research* 23 (15): 4354–63. https://doi.org/10.1158/1078-0432.CCR-16-2876.

Smith, Joanne, Lye Mun Tho, Naihan Xu, and David A. Gillespie. 2010. *The ATM-Chk2 and ATR-Chk1 pathways in DNA damage signaling and cancer*. 1st ed. Vol. 108. C. Elsevier Inc. https://doi.org/10.1016/B978-0-12-380888-2.00003-0.

Weinstein, John N., Eric A. Collisson, Gordon B. Mills, Kenna R.Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Chris Sander, et al. 2013. “The cancer genome atlas pan-cancer analysis project.” *Nature Genetics* 45 (10): 1113–20. https://doi.org/10.1038/ng.2764.

Zhao, Hui, Janis L. Watkins, and Helen Piwnica-Worms. 2002. “Disruption of the checkpoint kinase 1 / cell division cycle 25A pathway abrogates ionizing radiation-induced S and G2 checkpoints.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (23): 14795–14800. https://doi.org/10.1073/pnas.182557299.