Apartado2_GSEA

Marcos Rubio Fernandez

2023-April

Software utilizado

Para el análisis GSEA a realizar se han utilizado la versión de R y los siguientes paquetes/version:

```
## [1] "Versión R: 4.2.2"
```

```
## [1] "Versión tidyverse: 2.0.0"
```

```
## [1] "Versión DESeq2: 1.38.3"
```

```
## [1] "Versión VennDiagram: 1.7.3"
```

Creacion de archivo .rnk

Para este análisis de GSEA vamos a optar por realizar la opción **pre-ranked**, esta opción necesita de:

1. Archivo .gmt que define los gene sets: este archivo nos lo dan y contiene un gen set con los 100 genes más up-regulados en el tratamiento a 48 horas (DPN-Perturbed) y otro gen set con los 100 genes más down-regulated (DPN-UNPERTURBED).
2. Archivo .rnk: archivo de texto que debemos generar y que contiene una lista de genes y sus puntuaciones de clasificación correspondientes. La puntuación de clasificación es una medida que indique la importancia o el grado de expresión del gen en una muestra o condición en particular, en nuestro caso utilizaremos el logFold Change (LFC).

Para crear este archivo cargamos el objeto DESeq2 `dd3` creado en el apartado anterior y realizamos el contraste entre DPN24 horas y Control 24 horas.

```
dds_GSEA <- readRDS("input/dds3.rds")
res <- results(dds_GSEA, alpha = 0.05, contrast = c("group", "DPN24h", "Control24h"))
summary(res)
```

```
##
## out of 24416 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 13, 0.053%
## LFC < 0 (down)     : 74, 0.3%
## outliers [1]       : 0, 0%
## low counts [2]     : 10888, 45%
## (mean count < 28)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Para generación de los archivos .rnk es recomendable “reducir” (“shrunken”) los datos, este concepto se refiere a un método estadístico que busca mejorar la precisión de las estimaciones de los parámetros, reduciendo el error de la varianza. Para ello utilizaremos la función `lfcShrink`, con los siguientes parámetros:

- dds: es el objeto DESeq donde tenemos el modelo y los ajustes realizado
- coef: especifica el índice del coeficiente de la matriz de diseño que corresponde al contraste de interés.
- type: se refiere al método utilizado para realizar el shrunken. En este caso se escoge `apeglm`. El método APEglm (An Adaptive Permutation-based Extreme-value Gene Set Test for RNA-seq Data) utiliza una técnica de ajuste "shrunken" para reducir el ruido y mejorar la precisión de las estimaciones de expresión génica. En particular, APEglm utiliza un enfoque bayesiano empírico para encoger las estimaciones de la expresión génica hacia cero si la evidencia estadística no es lo suficientemente fuerte.
- res: especifica el objeto con los resultados que contiene las estimaciones de los LFC para cada gen.

Si analizamos el resultado de esta función podremos comprobar que el resultado en cuanto a los genes diferencialmente expresados es el mismo.

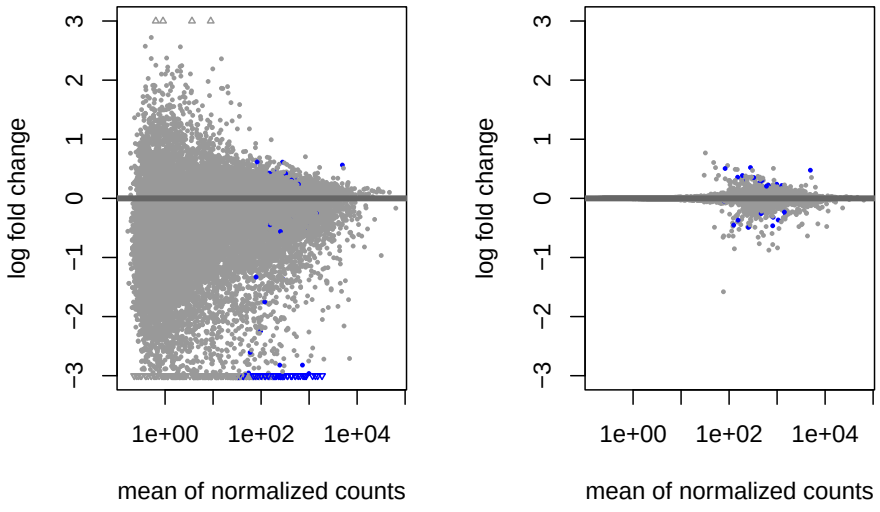
```
res.ape <- lfcShrink(dds = dds_GSEA, coef = "group_DPN24h_vs_Control24h", type = "apeglm",
                     res = res)
```

```
## using 'apeglm' for LFC shrinkage. If used in published research, please cite:
##     Zhu, A., Ibrahim, J.G., Love, M.I. (2018) Heavy-tailed prior distributions for
##     sequence count data: removing the noise and preserving large differences.
##     Bioinformatics. https://doi.org/10.1093/bioinformatics/bty895
```

```
summary(res.ape)
```

```
##
## out of 24416 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 13, 0.053%
## LFC < 0 (down)     : 74, 0.3%
## outliers [1]       : 0, 0%
## low counts [2]     : 10888, 45%
## (mean count < 28)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

De manera gráfica podemos ver la diferencia entre los datos sin "encoger" y los datos "encogidos". La varianza se ha reducido y ahora los datos son más precisos.



Lo último que hacemos es guardar todos estos datos en un archivo con extensión '.rnk'.

```
rnk <- data.frame(Feature = rownames(res.ape), LFC = res.ape$log2FoldChange)
write.table(rnk, file = "./input/DPN-Control_24h.rnk", sep = "\t", quote = FALSE,
            col.names = FALSE, row.names = FALSE)
```

Analisis GSEA

Para realizar el análisis utilizamos la aplicación para escritorio de GSEA, cargando los dos archivos necesarios y procesando la información con la opción pre-ranked. Los parámetros utilizados se pueden ver en el siguiente código:

```
gsea-cli.sh GSEAPreranked \
-gmx /home/vant/Documentos/14_Transcriptomica/transcriptomic-final-exercise/Apartado2/input/DPN_respons
-collapse Remap_Only \
-mode Abs_max_of_probes \
-norm meandiv \
-nperm 1000 \
-rnd_seed 123 \
-rnk /home/vant/Documentos/14_Transcriptomica/transcriptomic-final-exercise/Apartado2/input/DPN-Control_
-scoring_scheme weighted \
-rpt_label DPN_Control \
-create_svgs false \
-include_only_symbols true \
-make_sets true \
-plot_top_x 20 \
-set_max 500 \
-set_min 15 \
-zip_report false \
-out /home/vant/Documentos/14_Transcriptomica/transcriptomic-final-exercise/Apartado2
```

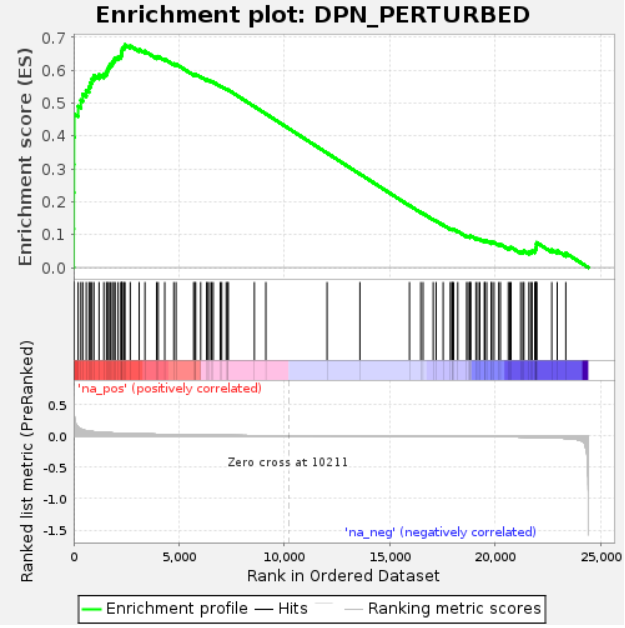
Resultados

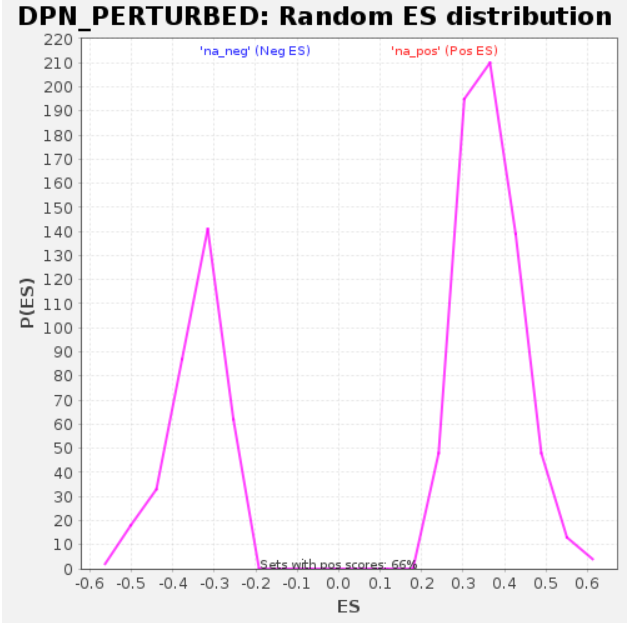
Gene set DPN PERTURBED

Para este gene set, GSEA ha detectado un enriquecimiento. El resumen de resultados mostrado más abajo, nos indica el que valor de `nominal p-value = 0` (que significaria que es 1/1000) y su FDR (valor ajustado a multiple test) es de 0 también. También podemos ver los valores del Enrichment Score (ES) y su valor normalizado (NES).

```
##            NAME        ES      NES NOM.p.val FDR.q.val
## 1 DPN_PERTURBED 0.6771898 1.850796         0         0
```

Si entramos en "Detalles" podemos ver el gráfico característico de este tipo de estudios, el cual nos indica que el gene set está enriquecido en los valores altos de la tabla, es decir, en los que se presupone sobreexpresados. Estos concuerda ya que el gene-set corresponde con los valores de genes sobreexpresados a 48 horas. El otro gráfico muestra la Random ES Distribution, una representación gráfica de la distribución de los Enrichment Scores obtenidos en las permutaciones aleatorias que se utilizan para calcular el ES, su distribución hacia la derecha nos indica que está significativamente enriquecido.





Por último, en esta misma sección de "Details" podemos ver que genes del gene-set forman parte del "core enrichment", es decir conforman el **Leading Edge**. En este gene-set el leading edge lo conforman 29 genes que a continuación podemos ver (filtrados por core.enrichmen == yes), así como su posición en la gene-list y las diferentes métricas del ensayo GSEA.

```
##              SYMBOL RANK.IN.GENE.LIST RANK.METRIC.SCORE RUNNING.ES
## 1  ENSG00000138449                 5         0.50675607  0.1172715
## 2  ENSG00000099194                 9         0.47588217  0.2274680
## 3  ENSG00000171227                15         0.37190548  0.3134781
## 4  ENSG00000226549                18         0.35769469  0.3963173
## 5  ENSG00000197594                33         0.30727932  0.4669756
## 6  ENSG00000128590               214         0.13328229  0.4904708
## 7  ENSG00000157326               343         0.10403246  0.5093238
## 8  ENSG00000151726               433         0.09255262  0.5271193
## 9  ENSG00000219481               598         0.07746407  0.5383326
## 10 ENSG00000166598               730         0.06972019  0.5491079
## 11 ENSG00000211584               789         0.06713055  0.5622850
## 12 ENSG00000198830               869         0.06418245  0.5739150
## 13 ENSG00000099875               963         0.06117212  0.5842714
## 14 ENSG00000173210              1207         0.05408751  0.5868166
## 15 ENSG00000117643              1433         0.04932836  0.5889988
## 16 ENSG00000197976              1559         0.04690869  0.5947326
## 17 ENSG00000168209              1605         0.04618018  0.6035876
## 18 ENSG00000044574              1663         0.04511571  0.6117023
## 19 ENSG00000144712              1740         0.04358725  0.6186812
## 20 ENSG00000121064              1834         0.04204658  0.6246039
## 21 ENSG00000167608              1900         0.04091020  0.6314146
## 22 ENSG00000112294              1970         0.03990458  0.6378278
## 23 ENSG00000176155              2110         0.03804224  0.6409303
## 24 ENSG00000131620              2239         0.03652461  0.6441336
## 25 ENSG00000175356              2275         0.03614455  0.6510733
## 26 ENSG00000066248              2285         0.03596973  0.6590417
## 27 ENSG00000179218              2309         0.03565364  0.6663611
## 28 ENSG00000119242              2393         0.03474127  0.6710015
```

```
## 29 ENSG00000151552              2436        0.03414477  0.6771898
```

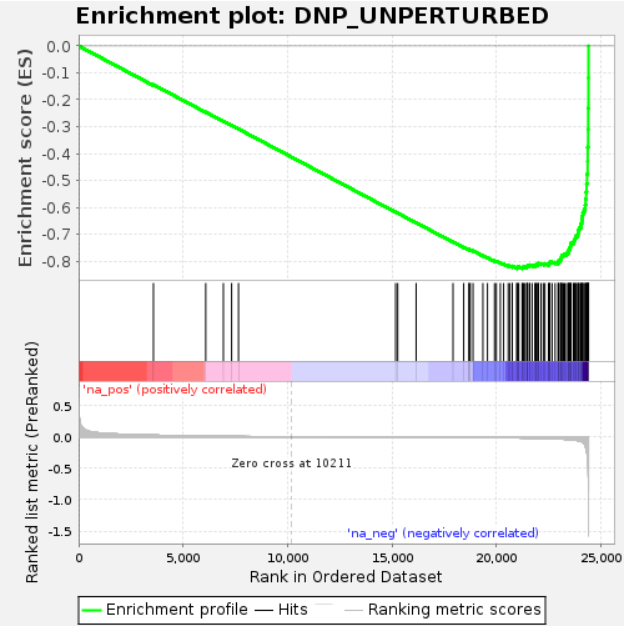
Gene set DPN UNPERTURBED

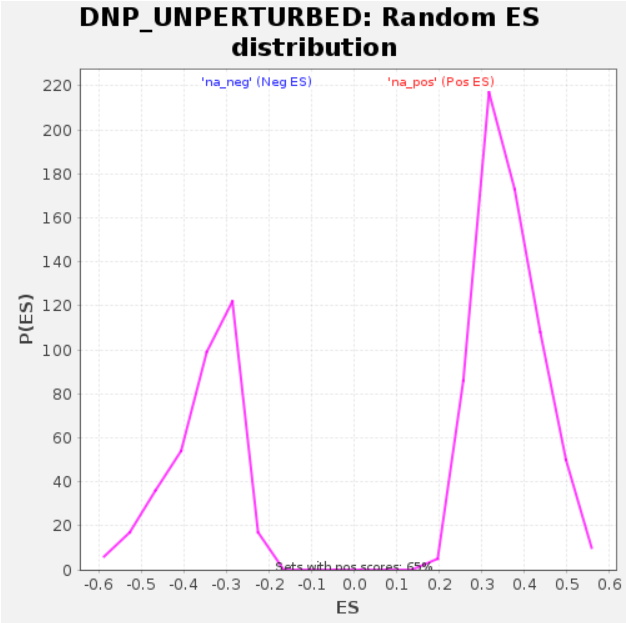
En este segundo Gene Set que contenia los 100 genes más down-regulated del experimento a 48 horas, nuestro análisis GSEA también ha encontrado un enriquecimiento. Como en el gene set anterior, mostramos el resumen general para este enriquecimiento, en el cual comprobamos de nuevo los valores de ES y su normalización (NES). Asimismo, vemo que el valor nominal de p-valor es 0 y su transformación para multiple test (FDR) también, lo cual refleja su significancia.

```
unperturbed <- read.table(file = "./Apartado2_GSEA_DPN.GseaPreranked.1681077764275/gsea_report_for_na_n
unperturbed[c(1,5:8)]
```

```
##               NAME         ES       NES NOM.p.val FDR.q.val
## 1 DNP_UNPERTURBED -0.8272808 -2.334847         0         0
```

En cuanto a los gráficos caracteristicos, en el primero de ellos podemos ver como la curva se desplaza hacia abajo y hacia la derecha, lo cual es un enriquecimento en genes localizados en la parte inferior (zona down-reg) por parte de este gene set; de nuevo, esto concuerda con el gene set aportado. En el segundo gráfico, vemos el desplazamiento a la derecha lo que nos refleja una significancia de ese enriquecimiento.





Al igual que en el caso anterior, mostramos a continuación en formato tabla el grupo de genes pertenecientes al **Leading Edge*, es decir aquellos genes que son el núcleo o los más importantes para el enriquecimiento y que en este caso son los localizados a la derecha del ES. En este grupo encontramos más genes que para el apartado anterior, llegando hasta los 72.

```
DPN48h_unperturbed <- read.table(file = "./Apartado2_GSEA_DPN.GseaPreranked.1681077764275/DNP_UNPERTURBE
DPN48h_unperturbed <- subset(DPN48h_unperturbed, CORE.ENRICHMENT == "Yes")
DPN48h_unperturbed[2:5]
```

```
##              SYMBOL RANK.IN.GENE.LIST RANK.METRIC.SCORE    RUNNING.ES
## 29  ENSG00000155111             21236       -0.01969206 -0.8246086000
## 30  ENSG00000115461             21252       -0.01979029 -0.8225400400
## 31  ENSG00000035928             21307       -0.02012251 -0.8220302000
## 32  ENSG00000149591             21351       -0.02040114 -0.8210302000
## 33  ENSG00000122786             21456       -0.02100726 -0.8224566000
```

```
## 34  ENSG00000168542          21472       -0.02115969 -0.8202022000
## 35  ENSG00000107796          21487       -0.02128690 -0.8178893300
## 36  ENSG00000069869          21500       -0.02138154 -0.8154814000
## 37  ENSG00000148795          21588       -0.02201634 -0.8160717500
## 38  ENSG00000186407          21604       -0.02213128 -0.8136855000
## 39  ENSG00000198959          21715       -0.02282678 -0.8151117600
## 40  ENSG00000107518          21849       -0.02376544 -0.8173564700
## 41  ENSG00000181264          21888       -0.02403468 -0.8156578000
## 42  ENSG00000105655          21896       -0.02409842 -0.8126756000
## 43  ENSG00000011465          21974       -0.02478060 -0.8124795600
## 44  ENSG00000005108          22009       -0.02501485 -0.8104834000
## 45  ENSG00000189060          22015       -0.02507892 -0.8072858500
## 46  ENSG00000153827          22141       -0.02615774 -0.8088770000
## 47  ENSG00000110723          22251       -0.02704364 -0.8096899000
## 48  ENSG00000149968          22328       -0.02762360 -0.8090669500
## 49  ENSG00000108691          22503       -0.02936084 -0.8122385000
## 50  ENSG00000221869          22509       -0.02940636 -0.8084538000
## 51  ENSG00000122641          22556       -0.02980598 -0.8063010000
## 52  ENSG00000113578          22563       -0.02985746 -0.8024961400
## 53  ENSG00000136153          22671       -0.03085777 -0.8027092000
## 54  ENSG00000169213          22816       -0.03255992 -0.8042130000
## 55  ENSG00000152952          22961       -0.03443413 -0.8054623600
## 56  ENSG00000169429          22968       -0.03453509 -0.8010228000
## 57  ENSG00000182836          22985       -0.03479733 -0.7969589000
## 58  ENSG00000020577          23066       -0.03609967 -0.7953503000
## 59  ENSG00000066468          23096       -0.03654617 -0.7915837000
## 60  ENSG00000122862          23108       -0.03671433 -0.7870540600
## 61  ENSG00000196154          23127       -0.03706943 -0.7827641000
## 62  ENSG00000120217          23138       -0.03720191 -0.7781271300
## 63  ENSG00000101974          23215       -0.03827907 -0.7760583000
## 64  ENSG00000137331          23275       -0.03958251 -0.7731134000
## 65  ENSG00000146376          23362       -0.04130083 -0.7710458000
## 66  ENSG00000139372          23440       -0.04290856 -0.7683899000
## 67  ENSG00000173706          23451       -0.04319957 -0.7629390400
## 68  ENSG00000075539          23463       -0.04349310 -0.7574895600
## 69  ENSG00000142192          23510       -0.04479476 -0.7533027500
## 70  ENSG00000205542          23526       -0.04506832 -0.7478040000
## 71  ENSG00000136144          23553       -0.04565543 -0.7426779000
## 72  ENSG00000155330          23556       -0.04574439 -0.7365528300
## 73  ENSG00000100612          23582       -0.04636184 -0.7312897400
## 74  ENSG00000111799          23694       -0.05018931 -0.7290441400
## 75  ENSG00000092969          23722       -0.05145984 -0.7231715300
## 76  ENSG00000075790          23759       -0.05312196 -0.7174435000
## 77  ENSG00000233705          23790       -0.05495274 -0.7112204000
## 78  ENSG00000164597          23827       -0.05636512 -0.7050522600
## 79  ENSG00000072110          23903       -0.06093827 -0.6998675000
## 80  ENSG00000115419          23919       -0.06175620 -0.6921042000
## 81  ENSG00000087053          23925       -0.06192769 -0.6839064400
## 82  ENSG00000069849          23972       -0.06506173 -0.6769695000
## 83  ENSG00000092020          24038       -0.07128559 -0.6699694000
## 84  ENSG00000123200          24050       -0.07299993 -0.6605158400
## 85  ENSG00000241878          24094       -0.07930154 -0.6515232300
## 86  ENSG00000243927          24114       -0.08205108 -0.6411705000
## 87  ENSG00000138772          24125       -0.08383788 -0.6302052000
```

```
## 88  ENSG00000102547            24144       -0.08625926 -0.6192403400
## 89  ENSG00000117632            24202       -0.10442162 -0.6074147000
## 90  ENSG00000100906            24257       -0.14024293 -0.5906049000
## 91  ENSG00000143153            24276       -0.16264535 -0.5692746600
## 92  ENSG00000143799            24302       -0.20010057 -0.5431496500
## 93  ENSG00000111907            24325       -0.23131447 -0.5126657000
## 94  ENSG00000111328            24346       -0.27067155 -0.4767588000
## 95  ENSG00000120129            24372       -0.36637491 -0.4280708700
## 96  ENSG00000101335            24375       -0.38444176 -0.3759854400
## 97  ENSG00000227105            24397       -0.49136135 -0.3101727000
## 98  ENSG00000168014            24401       -0.57628512 -0.2320957800
## 99  ENSG00000091137            24412       -0.83796692 -0.1187972300
## 100 ENSG00000138685            24414       -0.87606597  0.0000413846
```

Conclusion

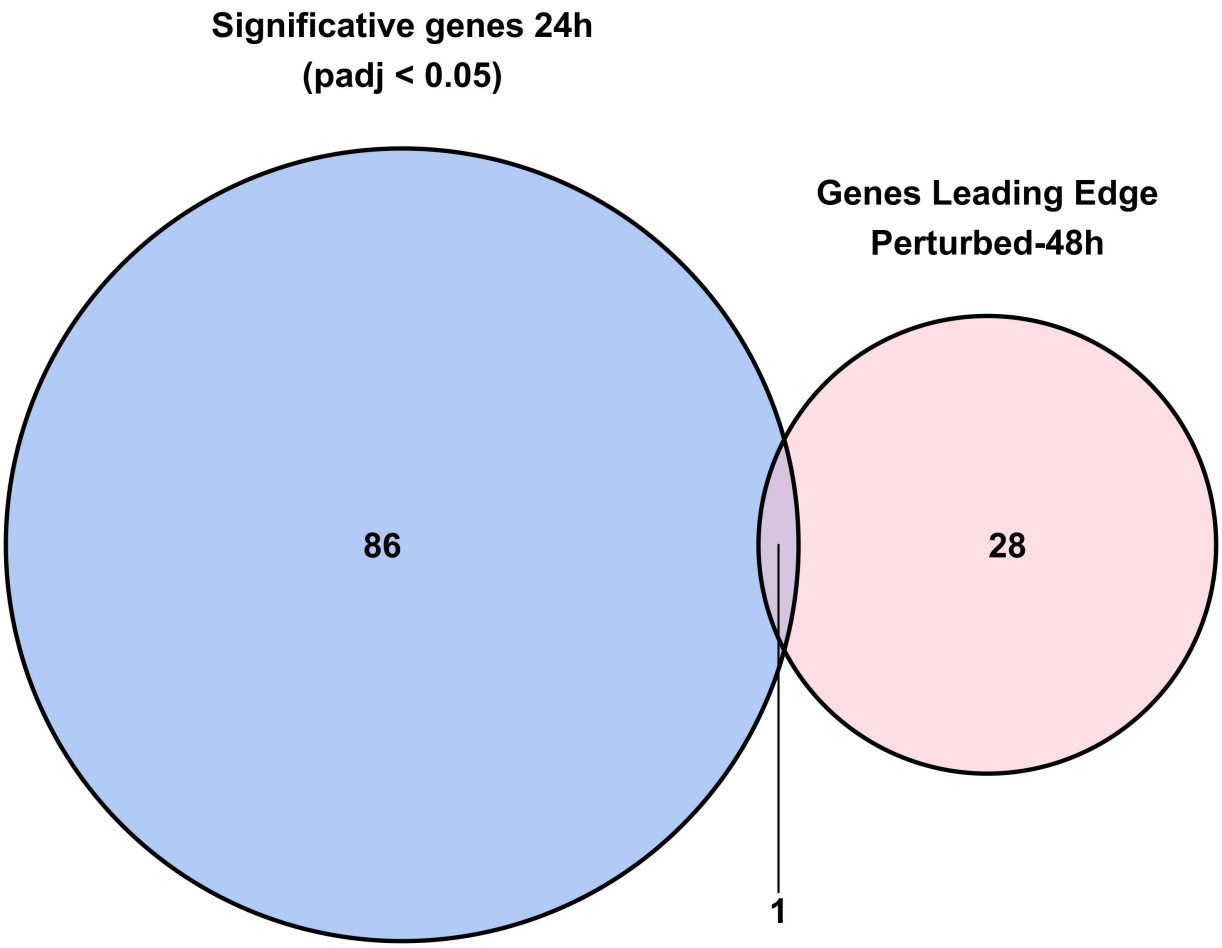
Partiendo de la premisa de que el gen set PERTURBED contenia los 100 up-regulados y el gene set UNPERTURBED los 100 down-regulados, y que el enriquecimiento en estos gene set concuerda con la parte alta y la parte baja de la tabla, podemos interpretar que el tratamiento con DPN tiene un efecto significativo a lo largo del tiempo.

Es decir, los genes que ya se encontraban en la zona de up-regulación (parte alta del .rnk), se encuentran significativamente up-regulados en el tratamiento a 48 horas. Esto mismo, pero sentido contrario, se puede interpretar para la zona baja o down-regulada.

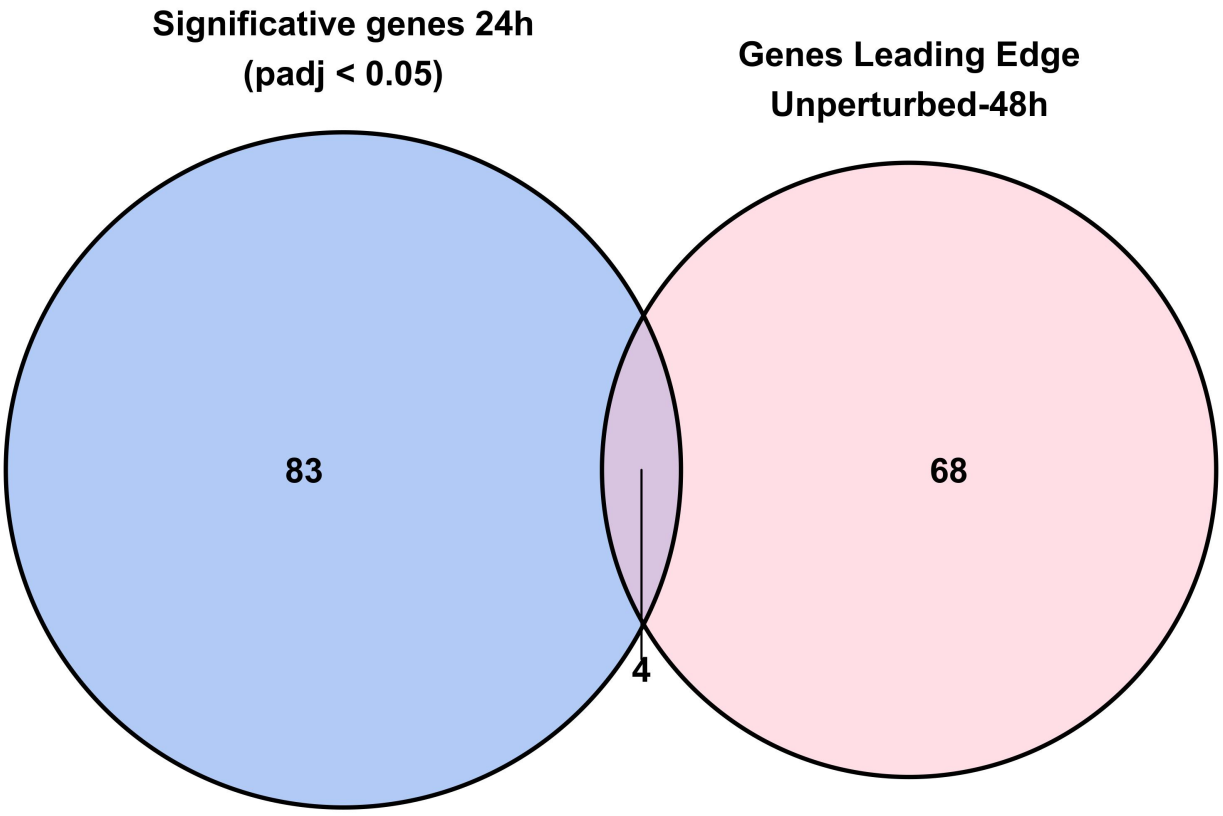
Un análisis complementario que podemos realizar consiste en comprobar si los genes del Leading Edge se superponen con los genes significativamente sobreexpresados o infraexpresados que obtuvimos del experimento a 24 horas.

A continuación mostramos los gráficos de Venn que muestran la intersección entre los genes que tenian un p-valor ajustado significativo en el experimento a 24 horas y los leading edge de cada gene set. Como vemos, aunque son muy pocos (1 y 4 genes), existen algunas concordancias. Estos genes que interseccionan podrían ser interesantes ya que supondrían genes que podrían estar actuando desde el comienzo del tratamiento y estar regulando otros, que podrían ser aquellos que no interseccionan. Una posible aproximación sería estudiar estos genes y las redes de interacción de los dos conjuntos para ver las concordancias entre ellos y si existe alguna “linea temporal” o procesos biológicos comunes.

Intersection DNP_24h vs PERTURBED



Intersection DNP_24h vs UNPERTURBED



“‘