

Transcriptomic analysis

De unos arrays al infinito

Nota del autor:

Yo este trabajo me lo tomé, desde el principio, como lo planteó el mismo Gonzalo Gómez, como si un ‘experimental’ que no tiene ni idea de que contienen sus arrays, decide dármeles para que le saque la máxima información posible. Por ello, apenas he leído el artículo asociado a estos datos, salvo para dos cuestiones:

- Para buscar los datos que aportan y cómo son.
- Para buscar el dataset con el que buscan notch.

Este último, solo lo hice para corroborar que el paper estaba bien, tal y como aconsejó Gonzalo. Por ello, he ido más allá y he intentado recabar la máxima información posible, con las herramientas de las que disponemos.

Todo el código y documentación adicional se encuentra en [Github](#). En este repositorio, se puede encontrar el archivo [main.R](#), en el que se puede observar qué flujo de trabajo se ha seguido.

1. Introducción: los datos

Tenemos dos datasets, uno procedente de ratón cuyo identificador es GSE18351 y otro procedente de humano cuyo identificador es GSE18198. Ambas tienen un caso (SAHM1) y un control (sin SAHM1). Así mismo, este último dataset, se ha probado en dos líneas celulares distintas: KOPT_K1 y HPB_ALL. En este trabajo, sobre todo nos centraremos, en este último dataset, ya que se puede hacer analizar de varias maneras.

2. Primeros pasos

2.1. Analizando SAHM1 vs control

En clase se nos proporcionaron varios archivos con los que trabajar y empezar a “jugar”. Con esto, lo primero que se me ocurrió hacer fue analizar los datos de humano sin diferenciar entre líneas celulares; es decir, casos serán los que tengan el SAHM1 y controles los que no lo tengan, para ver si había algún gen diferencialmente expresado con un *fdr* interesante. Para esta tarea (y haciendo uso del código proporcionado), creamos dos funciones en R ‘normalizeData’ y ‘differentialExpression’.

El primer paso tras leer los datos de los .cel, es normalizar. Este paso se realiza para homogeneizar las muestras existentes en cada array, para que no haya sesgos externos (ej: el

posible fondo y luminosidad de él mismo) y se filtran las que no estén bien. En cualquier caso, si estos datos no se pudieran normalizar habría que hacer algo con ellos, ya que un análisis de estos, podría llevar a resultados incongruentes. Afortunadamente, se normaliza correctamente.

El siguiente paso, teniendo ya los datos normalizados, es buscar genes diferencialmente expresados. Para ello, usaremos modelos lineales con limma para R. Este análisis nos dará que genes han cambiado más de nuestra lista y con su tasa asociada de que sea un falso positivo. Posteriormente, analizaremos si encontramos algún dato significativo con el *fdr* asociado.

Estos dos pasos (normalización y expresión diferencial) son comunes a cada uno de los datasets y sus variaciones utilizados.

2.2. Buscando NOTCH1

Un buen *fdr* como mucho será de 0.05 (o incluso 0.1 he visto), pero estos datos no llegaban ni a eso. El menor de todos ellos era aproximadamente 0.4 (una probabilidad muy alta de que sea falso positivo), por ello, utilizamos GSEA para ver si nuestro dataset tiene algún gene set enriquecido.

El primer paso, como ya hemos mencionado, era comprobar los resultados aportados por el artículo y verificarlos. Por esta razón, lanzamos GSEA con nuestros arrays de expresión. Como base de datos de sets de genes utilizamos “C2 - cgp” (en él se encuentra notch) y “C3 - tft” que está MYCMAX (en dos tandas) y el tipo de permutación por set de genes en vez de fenotipo. En el paper añaden notch1 a C3 - tft. Aquí se hará por separado.

MYCMAX lo encontramos entre los geneset enriquecidos, tal y como menciona el paper.

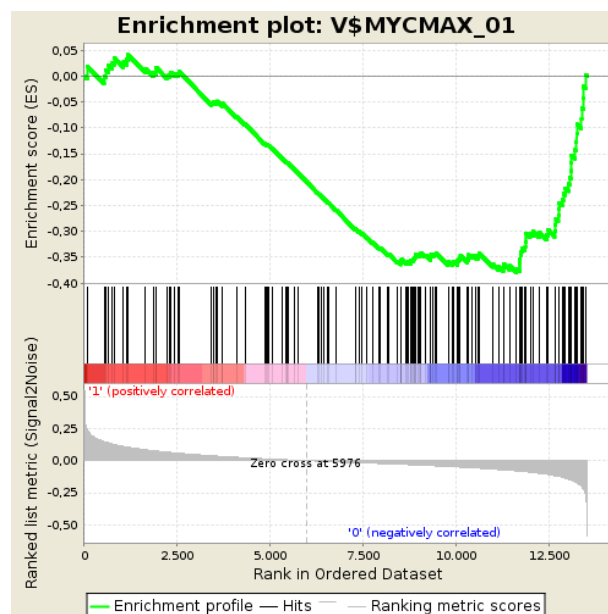


Figura 1. MYCMAX en GSEA.

Y aunque encontramos resultados tan buenos como el de “Leonard_HYPOXIA” con un *fdr* de 0, GSEA no muestra el de “Palomino” utilizado en el paper.

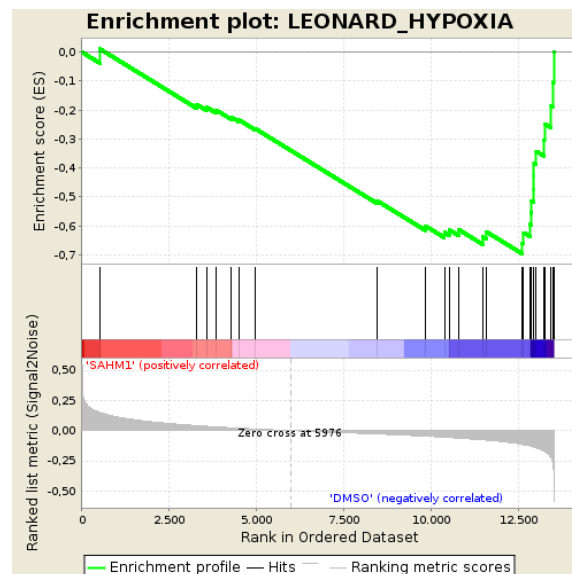


Figura 2. Hypoxia en GSEA.

A pesar de esto, podemos afirmar con rotundidad, que en ambos conjuntos de datos encontramos *gene sets* enriquecidos en nuestros datos. En el dataset de “C2 - gcp” encontramos 428 sets de genes significativamente enriquecidos en DMSO (control) y 19 en SAHM1 (caso), entre ellos, el mencionado en DMSO.

Como la finalidad principal era analizar los datos, nos lanzamos a ello en busca de algo interesante (y sin mucho conocimiento biológico). Lanzamos GSEA contra todos los datasets (C1, C2, ...) y varias variaciones. Pero aunque encontramos en ocasiones muchos *gene sets* enriquecidos, no terminan de ser demasiado claros (Ejemplo: Figura 3).

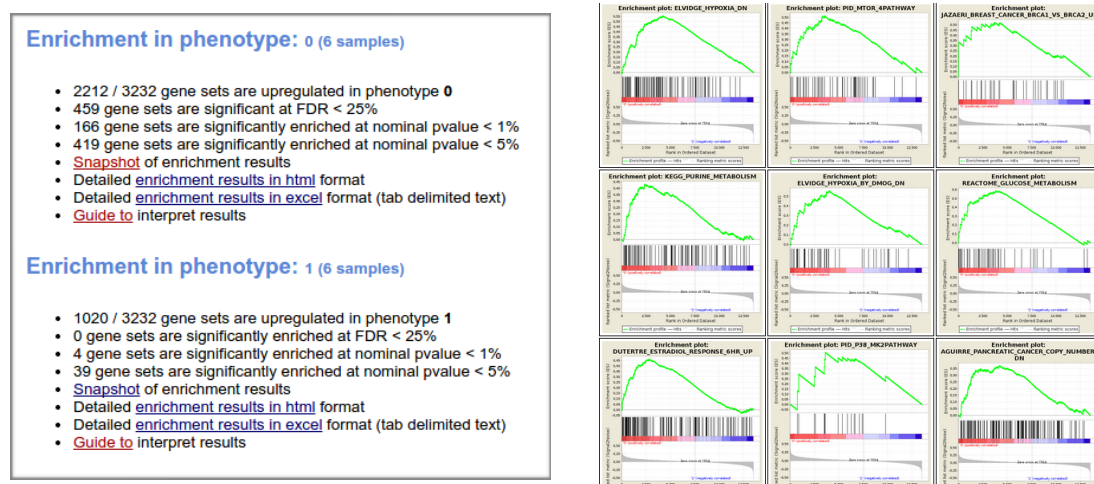


Figura 3. Datos no relevantes de GSEA. 0 es DSM0 y 1 es SAHM1.

3. Analizando las líneas celulares

La otra variante de análisis que ha resultado ser mucho más interesante ha sido el de dividir el dataset en dos y analizar por separado cada línea celular. Ha sido mucho más interesantes

porque me ha dejado analizar más ampliamente los resultados y, al haber más herramientas para ello, se ha podido estrujar más los datos.

El procedimiento seguido para ambas líneas ha sido el mismo entre ellas:

- A. Normalizar los arrays.
- B. Buscar genes diferencialmente expresados.

Hasta aquí, es el paso común a la variante anterior. El caso es que en ambas, al contrario lo analizado con anterioridad, nos salen genes diferencialmente expresados. Vamos a analizarlos individualmente.

3.1. KOPT_K1

Ya hemos dicho que salen genes diferencialmente expresados, pero en esta línea, y aplicando un threshold en el *fdr* de menores de 0.01, salen 3183 genes. Muchos. Cuando tenemos ya estos genes, debemos anotarlos y ver a que corresponden. Para ello hacemos uso del código ‘*anotateBestGenesHuman*’, el cual nos anotará los valores de nuestros genes (en el caso de que existan) filtrado por *fdr*. Una vez anotados, debemos filtrar los que no hayan sido anotados. Con lo que obtenemos una lista de 1991 nombres de genes (únicos) que se encuentran diferencialmente en nuestro array para esta línea celular.

La anotación está hecha, pero... ¿Y ahora qué? Ahora el siguiente paso, queremos ver qué son nuestros datos, es decir, con qué genes están relacionados y con qué pathways. Por esta razón, nos lanzamos hacia babelomics hacia la sección de enriquecimiento funcional, la cual nos da información bastante variopinta una vez anotados nuestros genes.

Lo primero que hacemos es buscar con qué genes está relacionados nuestros genes encontrados, para ello, utilizamos de Babelomics “Network Enrichment - Snow”.

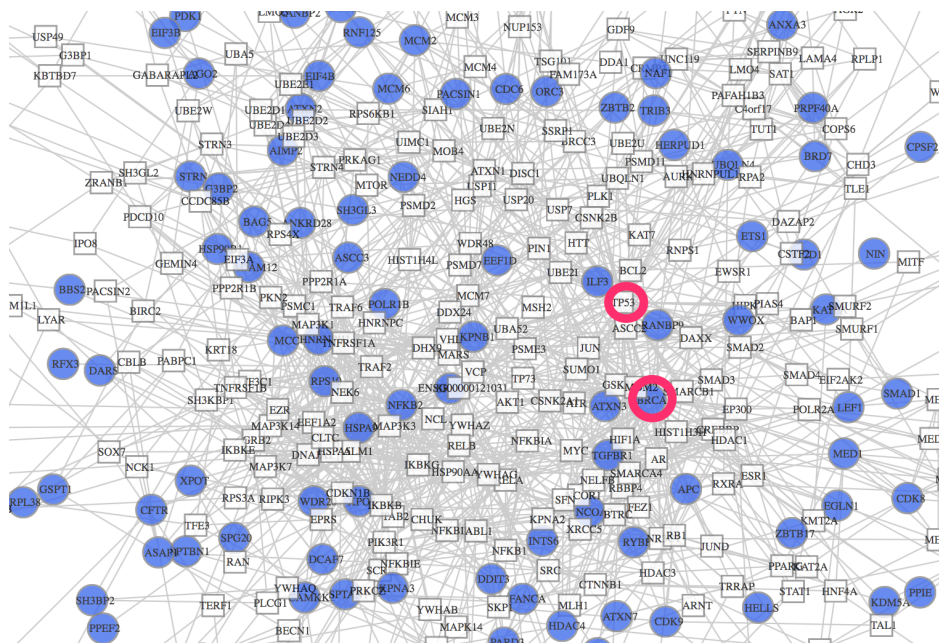


Figura 4. En azul los de nuestra lista y en blanco los puestos por la herramienta. Genes señalados en rojo: TP53 y BRCA1.

Al analizar los datos de una red de las primeras cosas que se debe hacer de manera muy básica, es ver que nodos son más importantes. En este caso, eso nos lo da tanto la centralidad (betweenness), como el grado del nodo.





| Input id | id | Type | Betweenness ^ | Clustering | grado nodo |
|----------|-----------------|----------|---|------------|--|
| TP53 | ENSG00000141510 | external | 0.13  | 0.041 | 52  |
| BRCA1 | ENSG0000012048 | list | 0.084  | 0.059 | 46  |
| NFKB2 | ENSG00000077150 | list | 0.047 | 0.11 | 41 |
| MAP3K3 | ENSG00000198909 | external | 0.032 | 0.074 | 30 |

Figura 5. Top genes por centralidad.

Tal y como hemos reflejado en la red anterior, tenemos expresados en nuestro array conectado a muchos otros el gen BRCA1 y el programa ha añadido TP53, también muy conectado con otros. Ambos genes son oncogenes, con lo que la probabilidad de que ocurra algo con el cáncer en nuestro array es media-alta.

Seguimos haciendo tests y ahora buscamos genes relacionados a nivel gene set, con la herramienta “Gene set network enrichment - Network miner”. Sobre todo, buscamos intentar replicar los resultados anteriores, en busca de oncogenes o similares.







| Input id | id | Type | Betweenness ^ | Clustering | Connections |
|----------|-----------------|----------|---|------------|-------------|
| RELA | ENSG00000173039 | external | 0.12  | 0.11 | 10 |
| TARS | ENSG00000113407 | list | 0.11  | 0.071 | 8 |
| LARS | ENSG00000133706 | list | 0.11  | 0.048 | 7 |
| HSP90AA1 | ENSG00000080824 | external | 0.11  | 0.2 | 11 |
| JUN | ENSG00000177606 | list | 0.1  | 0.13 | 11 |
| TP53 | ENSG00000141510 | external | 0.094  | 0.071 | 8 |

Figura 6. Resultados GSNE. Verde: Nuevos oncogenes. Rojo: Replicado en anterior.

Comprobamos los resultados de la centralidad con el cytoscape que los confirma mayormente. Por tanto, en verde encontramos nuevos genes relacionados con nuestra lista que son oncogenes, con una centralidad alta, pero con un grado del nodo, no tan alto; lo cual, quiero decir que los resultados son consistentes. Por otro lado, volvemos a encontrar en el top 6, de nuevo, al gen TP53. A partir de estos resultados nos deja lanzar, tanto FatiGo, como el Network Enrichment, y lanzamos ambos. Con el segundo, no obtenemos nada nuevo. Y con el primero lo analizaremos después de la siguiente explicación.

Por otro lado, con nuestra lista de genes anotados, también podemos lanzar FatiGo para ver con qué pathways están relacionados estos. Lo lanzamos contra el resto del genoma. Ordenamos todos por menor *fdr* y observamos que los primeros resultados con un 2% de genes (que serían alrededor de 20 de los casi 2000 genes que hay en la lista) corresponden con procesos metabólicos de tRNA y helicasas, que, al parecer, ayudan en ese proceso.

| Term | Term size | Term size(in genome) | Annotations lists | Annotated ids list | Odds ratio (log e) | Pvalue | Adj. Pvalue v |
|--|-----------|----------------------|--------------------------------|--|--------------------|---------|---------------|
| Helicase_C(IPR001650) | 294 | 335 | List 1: 1.36% List 2: 0.14% | List 1: ZNRANB3 DDX17 DDX18 DDX21 DDX31 List 2: ENST00000061: ENST00000037: ENST00000042: ENST00000026: | 2.28 | 1.7e-17 | 6.9e-14 |
| Helicase_ATP-bd(IPR014001) | 375 | 421 | List 1: 1.36% List 2: 0.18% | List 1: ZNRANB3 DDX17 DDX18 DDX21 DDX31 List 2: ENST00000061: ENST00000059: ENST00000037: ENST00000037: ENST00000026: | 2.02 | 7.1e-15 | 1.5e-11 |

Figura 7. Interpro most significant terms.

| | | | | | | | |
|--|-----|-----|--------------------------------|---|------|---------|---------|
| helicase activity(GO:0004386) | 480 | 549 | List 1: 2.06% List 2: 0.23% | List 1: XRCC5 MCM2 MCM6 DDX17 DDX18 List 2: ENST000000581 ENST000000615 ENST000000315 ENST000000434 ENST000000595 | 2.21 | 1.4e-24 | 2.6e-21 |
| transcription corepressor activity(GO:0003714) | 397 | 444 | List 1: 1.61% List 2: 0.19% | List 1: TBL1X DDIT3 RLIM SCAI NPAT List 2: ENST000000212 ENST000000263 ENST000000366 ENST000000615 ENST000000394 | 2.14 | 8.8e-19 | 8.6e-16 |
| protein complex binding(GO:0032403) | 304 | 315 | List 1: 1.41% List 2: 0.14% | List 1: SPTBN1 MED1 HADHB MYSM1 IQGAP1 List 2: ENST000000537 ENST000000263 ENST000000263 ENST000000252 ENST000000270 | 2.29 | 4e-18 | 2.6e-15 |

Figura 8. GOSlim GOA most significant terms.

| | | | | | | | |
|---|-----|-----|--------------------------------|--|------|---------|---------|
| tRNA metabolic process(GO:0006399) | 453 | 581 | List 1: 2.06% List 2: 0.22% | List 1: DARS EARS2 IARS MTO1 WDR4 List 2: ENST000000566 ENST000000434 ENST000000377 ENST000000428 ENST000000366 | 2.27 | 1.5e-25 | 8.3e-22 |
| response to drug(GO:0042493) | 400 | 407 | List 1: 1.86% List 2: 0.19% | List 1: XRCC5 USP47 P2RX7 KCNJ11 EGR1 List 2: ENST000000252 ENST000000434 ENST000000219 ENST000000343 ENST000000327 | 2.3 | 1.5e-23 | 3.8e-20 |
| regulation of gene expression(GO:0010468) | 235 | 253 | List 1: 1.51% List 2: 0.11% | List 1: IRX5 KDM6A SETD2 NPAT DICER1 List 2: ENST000000249 ENST000000366 ENST000000263 ENST000000249 ENST000000249 | 2.66 | 2e-23 | 3.8e-20 |

Figura 9. GO molecular function most significant terms.

| | | | | | | | | |
|--|-----|-----|--------------------------------|--|--|------|---------|-------|
| tRNA metabolic process(GO:0006399) | 453 | 581 | List 1: 2.06% List 2: 0.22% | List 1: DARS EARS2 IARS MTO1 WDR4 | List 2: ENST000005667 ENST000004342 ENST000003777 ENST000004280 ENST000003665 | 2.27 | 1.5e-25 | 7e-24 |
| helicase activity(GO:0004386) | 479 | 548 | List 1: 2.06% List 2: 0.23% | List 1: XRCC5 MCM2 MCM6 DDX17 DDX18 | List 2: ENST000005818 ENST000006157 ENST000003152 ENST000004342 ENST000005950 | 2.21 | 1.3e-24 | 3e-23 |

Figura 10. GO biological process most significant terms.

Volviendo a los datos procesados de la red, lanzamos FatiGo con eso datos que, mayoritariamente nos daban genes oncogénicos. En contraposición a estos, los datos vuelven a ser similares a los anteriores mostrados con FatiGo también, pero solo con la lista de genes, en vez de con la red. Pero, a los términos encontrados anteriormente, se añaden bastante y con mucha significancia en un 6.4% de los genes: la actividad de ligasa. Dando, así, más fuerza al metabolismo de tRNA.

| | | | | | | | | |
|--|-----|------|--------------------------------|---|---|------|---------|---------|
| ligase activity(GO:0016874) | 299 | 1388 | List 1: 6.4% List 2: 0.65% | MDM2 RLIM MARS IARS BTRC | ENSG00000170 ENSG00000170 ENSG00000128 ENSG00000157 ENSG00000145 | 2.34 | 2.5e-14 | 1.6e-11 |
| aminoacyl-tRNA ligase activity(GO:0004812) | 30 | 174 | List 1: 2.44% List 2: 0.05% | List 1: MARS IARS TARS SARS EPRS | List 2: ENSG00000145 ENSG00000113 ENSG00000115 MARS IARS | 3.88 | 5.3e-11 | 1.8e-8 |
| nucleotide binding(GO:000166) | 231 | 1748 | List 1: 3.35% List 2: 0.52% | List 1: TUT1 MARS IARS TARS SARS | List 2: ENSG00000135 ENSG00000095 ENSG00000177 ENSG00000145 ENSG00000188 | 1.9 | 1.9e-6 | 3.4e-4 |
| transcription factor binding(GO:0008134) | 333 | 1174 | List 1: 3.96% List 2: 0.75% | List 1: ATF2 FOS CEBPB ESR1 TAF9 | List 2: ENSG00000165 ENSG00000155 ENSG00000170 ENSG00000174 ATF2 | 1.7 | 2.1e-6 | 3.4e-4 |

Figura 11. Go molecular function most significant terms.

| | | | | | | | | |
|--|-----|-----|--------------------------------|---|---|------|---------|---------|
| TRIF-dependent toll-like receptor signaling pathway(GO:0035666) | 81 | 225 | List 1: 3.66% List 2: 0.16% | List 1: ATF2 FOS NFKB2 UBE2D3 UBE2D2 | List 2: ENSG0000016 ENSG0000012 ATF2 ENSG0000005 FOS | 3.15 | 1.4e-12 | 8.6e-10 |
| toll-like receptor signaling pathway(GO:0002224) | 100 | 305 | List 1: 3.96% List 2: 0.2% | List 1: ATF2 FOS NFKB2 UBE2D3 UBE2D2 | List 2: ENSG0000016 ENSG0000012 ATF2 ENSG0000016 ENSG0000005 | 3.01 | 9.2e-13 | 8.6e-10 |
| MyD88-independent toll-like receptor signaling pathway(GO:0002756) | 81 | 230 | List 1: 3.66% List 2: 0.16% | List 1: ATF2 FOS NFKB2 UBE2D3 UBE2D2 | List 2: ENSG0000016 ENSG0000012 ATF2 ENSG0000005 FOS | 3.15 | 1.4e-12 | 8.6e-10 |
| toll-like receptor 4 signaling pathway(GO:0034142) | 94 | 278 | List 1: 3.96% List 2: 0.19% | List 1: ATF2 FOS NFKB2 UBE2D3 UBE2D2 | List 2: ENSG0000016 ENSG0000012 ATF2 ENSG0000005 FOS | 3.08 | 4e-13 | 8.6e-10 |
| toll-like receptor 3 signaling pathway(GO:0034138) | 82 | 231 | List 1: 3.66% List 2: 0.16% | List 1: ATF2 FOS NFKB2 UBE2D3 UBE2D2 | List 2: ENSG0000016 ENSG0000012 ATF2 ENSG0000016 ENSG0000005 | 3.14 | 1.6e-12 | 8.6e-10 |

Figura 12. Go biological process most significant terms.

| | | | | | | | | |
|---|-----|------|--------------------------------|--|---|------|---------|---------|
| ligase activity(GO:0016874) | 299 | 1389 | List 1: 6.4% List 2: 0.65% | MARS IARS BTRC TARS ASS1 SARS | ENSG00000012 ENSG00000015 ENSG00000014 ENSG00000012 ENSG00000001 ENSG00000018 | 2.34 | 2.5e-14 | 2.7e-12 |
| cellular amino acid metabolic process(GO:0006520) | 153 | 1203 | List 1: 3.96% List 2: 0.33% | List 1: MARS IARS TARS SHMT2 ASS1 TARS SARS EPRS | List 2: ENSG00000014 ENSG00000016 ENSG00000014 ENSG00000011 ENSG00000012 ENSG00000011 ENSG00000011 MARS | 2.53 | 2.1e-10 | 1.2e-8 |
| tRNA metabolic process(GO:0006399) | 66 | 581 | List 1: 2.44% List 2: 0.14% | YARS LARS CARS | ENSG00000018 IARS ENSG00000013 ENSG00000017 | 2.91 | 4.1e-8 | 1.5e-6 |

Figura 13. GOSlim GOA most significant terms

3.2. ¹HPB_ALL

Siguiendo los mismos (o casi los mismos) pasos que en la línea celular anterior, obtenemos 287 genes diferencialmente expresados y de esos, después de anotar, tenemos 247 anotaciones de genes únicas. Ahora procederemos a buscar información funcional a estos datos y, si fuera posible, buscar concordancia con la línea anterior.

Al igual que en la línea celular anterior, con la lista de identificadores de las anotaciones lanzamos “Network enrichment - Snow”.





| Input id | id | Type | Betweenness ^ | Clustering | Connections |
|----------|-----------------|----------|---|------------|-------------|
| TP53 | ENSG00000141510 | external | 0.13  | 0.073 | 36 |
| JUN | ENSG00000177606 | list | 0.082  | 0.062 | 26 |
| MDM2 | ENSG00000135679 | list | 0.068  | 0.09 | 25 |
| HSP90AA1 | ENSG00000080824 | external | 0.057  | 0.18 | 21 |

Figura 14. Genes con más centralidad en la red. En rosa: ya salieron antes.

Siguimos en la misma línea de la información dada por los datos y vemos que el nodo con más centralidad es TP53, el oncogen por excelencia.

Al hacer el “Gene Set Network Enrichment” no aporta demasiada información, pero es relevante, para confirmar que así salen.

¹ T cell tumor producing iL-2 and iL-6. Tissue: Lymphocyte. Fenotipo: supresión.

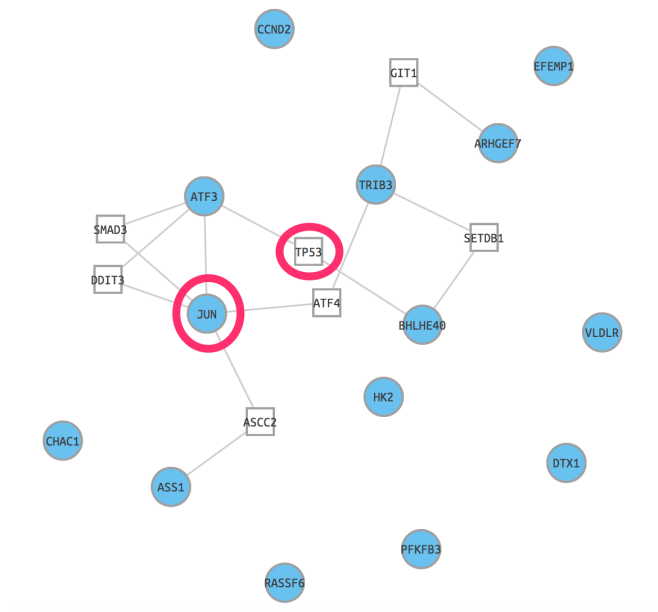


Figura 15. Red sacada del GSN. Remarcados: JUN y TP53.

Con mayor centralidad aparece JUN y ATF3, oncogenes también. Haciendo NE-Snow, obtenemos algún encogen más. Pero haciendo FatiGo después de GSNE, ampliamos más información. Aunque, al ser menos genes, ampliamos menos. Tanto en Gen Ontology molecular función y GO biological function sale “regulación de la transcripción por la RNA polimerasa II”.

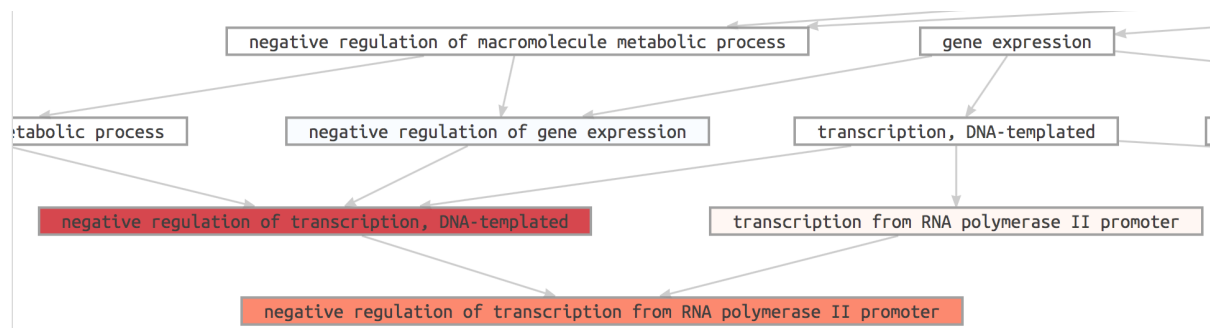


Figura 16. Red de temas para GO biological process, en rojo los genes más importantes.

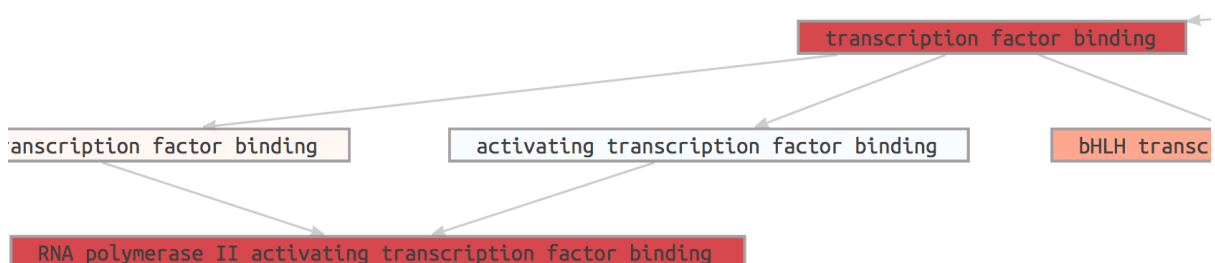


Figura 17. Red de temas para GO molecular function, en rojo los genes más importantes.

Algo que había llamado la atención anteriormente, en los resultados de la anterior línea celular, aunque no se había mostrado, porque parecía carecer de relevancia. Pero es algo que comparten ambas líneas, junto al tema de oncogenes. Por otro lado, también comparten el

proceso glicolítico, que estaría relacionado con el ATP (al igual que el metabolismo) anteriormente relacionado.

3.3. Relación entre ambas líneas

Otra opción que tiene FatiGo, es que puedes lanzar una lista de IDs contra otra. He de decir que no se encuentran muchos resultados con FDRs significativos, pero uno que si aparece, y que anteriormente lo ha hecho también, es la glicólisis y el proceso metabólico de la glucosa.

3.4. Y con todos estos datos, ¿qué hacemos?

Pues hacemos una búsqueda en algún buscador (tipo ecosia o Google para los más tradicionales), para los términos: metabolismo, toll-like receptores y cancer. En los cuales encontramos cosas como:

- ▶ [\[1\]](#) Toll-like receptors: linking inflammation to metabolism.
- ▶ [\[2\]](#) Toll-like receptors and cancer.

Y para los términos de la segunda línea: “negative regulation transcription rna polymerase II cancer”.

- ▶ [\[3\]](#) Bacterial control of host gene expression through RNA polymerase II
- ▶ [\[4\]](#) RNA polymerase III transcription in cancer: the BRF2 connection

Uniendo todo esto, quizás, un poco con pinzas (o no).

4. Conclusión

Una muy mala conclusión hubiera sido lo que hacen muchos en multitud de papers, arrojando y sentenciando que los resultados aquí mostrados o las posibles conjeturas sacadas de los mismos son válidas o no, no seré tan *naïve*.

Lo que está claro de todo esto, es que cuando haces un análisis de datos sin buscar cosas concretas, las cuales tengas alguna evidencia de ellas y quieras demostrarla, es muy complicado que encuentres algo sólido.

Por esta razón, yo he intentado algo, que, quizás es lo que más me gusta, sacar información de los datos, he intentado estrujarlos y ver si se puede sacar algo en claro, o medianamente claro. Lo haya conseguido con sentido o no, es otro cosa.

Los datos te inundan de información, esto es algo bueno si sabes nadar. Pero, cuidado amigo, no tragues muchas agua, que al final te puedes ahogar.