Redes complejas con aplicaciones en biología de sistemas - Trabajo Práctico Final

Análisis de reguladores maestros en cáncer de tiroides

Miranda, Lucas; Salustri, Guido; Schuster, Claudio; Sosa, Ezequiel
12/12/2018

Índice

Abstract **2**

Introducción **3**

Metodología general y resultados **4**

Armado y caracterización de la red 4

Expresión diferencial 6

Análisis de reguladores maestros 7

Discusión **10**

Conclusiones y perspectivas a futuro **12**

Referencias **13**

El presente trabajo se propone identificar los factores de transcripción que se desempeñan como reguladores maestros en la transformación de la célula folicular tiroidea, utilizando una red regulatoria inferida de experimentos de expresión génica masiva (RNAseq).

Abstract

El fenotipo celular es el resultado la interacción entre muchos procesos, entre los cuales tiene especial importancia cómo se encuentra orquestada la regulación de los genes que la célula posee. El presente trabajo se basa en la idea de que, al existir un viraje brusco en el fenotipo como lo es la transformación celular (el proceso mediante el cual una célula se vuelve maligna) son unos pocos factores de transcripción los que desencadenan una respuesta regulatoria que redunda en el cambio y la manutención del fenotipo resultante.

El presente trabajo se propone identificar estos factores de transcripción (de aquí en más **reguladores maestros** o **RM**) en la transformación de las células foliculares tiroideas. Esto se llevará a cabo mediante la inferencia de una red regulatoria a partir de datos de expresión génica masiva y el posterior análisis con software ya desarrollado.

Introducción

La glándula tiroides está localizada debajo del cartílago tiroideo, en la parte delantera del cuello. Esta glándula en forma de mariposa tiene dos lóbulos: el lóbulo derecho y el lóbulo izquierdo, que están unidos por un istmo angosto. El cáncer tiroideo agrupa a un pequeño número de tumores malignos de la glándula tiroides, y es la malignidad más común del sistema endocrino. Por lo general los tumores malignos de la tiroides tienen su origen en el epitelio folicular de la glándula y son clasificados de acuerdo a sus características histológicas. Este trabajo se centra en la descripción del **carcinoma papilar tiroideo** (PTC por su sigla en inglés), un tipo de tumor bien diferenciado que, si bien es el menos agresivo y el de mejor pronóstico, es el más frecuentemente diagnosticado en pacientes.

A modo de resumen, el cáncer se origina cuando un grupo de células proliferan descontroladamente en un determinado tejido. Este comportamiento se origina, entre otras causas, por mutaciones que afectan los mecanismos de reparación de ADN, senescencia y apoptosis, exacerbando los mecanismos de replicación y crecimiento. Dichos cambios redunden en el perfil de expresión génica característico de la célula.

Como el método de detección estándar utilizado en la actualidad se basa meramente en caracteres histológicos, muchas veces el tipo de tumor detectado no se corresponde con el que realmente afecta al paciente, generando falsos positivos (donde pueden realizarse cirugías o tratamientos innecesarios, por ejemplo) o falsos negativos (donde pacientes con tumores graves son tratados en forma más leve). Para entender, diagnosticar correctamente y potencialmente tratar estas enfermedades, es necesario utilizar nuevas fuentes de información que permitan ampliar el análisis. Es aquí donde entran en juego los análisis de expresión génica y de redes regulatorias que desarrollaremos en este informe.

Metodología general y resultados

Armado y caracterización de la red

Para estudiar la problemática descrita se buscó en primer lugar modelar la red de interacción regulatoria de la célula folicular tiroidea humana. Para esto se utilizaron bases de datos públicas con experimentos de expresión génica global (RNASeq) y el programa especializado ARACNe (*Algorithm for the Reconstruction of Accurate Cellular Networks*), desarrollado por el laboratorio de biología de sistemas de la Universidad de Columbia.

Como input se requiere una tabla de datos de expresión en la que las columnas especifican muestras (que deben incluir variabilidad en cuanto a las condiciones en las que fueron obtenidas, no representar un sólo estado del tejido) y las filas especifican genes. Se usaron 560 muestras de RNASeq obtenidas de TCGA (*The Cancer Genome Atlas*). Además, debe proveerse una lista con los factores de transcripción conocidos para la especie en cuestión (obtenidos en este caso de Gene Ontology como todos aquellos genes de *Homo sapiens* etiquetados con *gene regulation* o *transcription factor*).

ARACNe implementa un algoritmo que aplica teoría de la información para identificar interacciones transcripcionales entre productos génicos utilizando datos de expresión. Seleccionando las muestras adecuadas, se pueden obtener los pares que componen las interacciones regulatorias para una tejido/fenotipo específico. Esto constituye las bases para posteriores análisis, en nuestro caso la obtención de reguladores maestros.

A diferencia de otras aproximaciones más simples, este algoritmo utiliza un método basado en información mutua, que tiene la capacidad de establecer relaciones de entrada/salida, aun cuando la función entre las mismas es no lineal o irregular.

La información mutua (MI por su sigla en inglés) mide la reducción de la incertidumbre (entropía) de una variable aleatoria X debido al conocimiento del valor de otra variable aleatoria Y. En nuestro caso X e Y son niveles de expresión.

$$I(x_i; y_j) = \log \frac{P(x_i|y_j)}{P(x_i)}$$

El resultado (figura 1) es un grafo dirigido en el cual los nodos representan genes, y las aristas relaciones de regulación (quién regula transcripcionalmente a quién). El acto de regulación transcripcional puede implicar tanto activación como inhibición, y es generalmente debido a la interacción entre la proteína producto del gen del cual parte la arista con el promotor de la secuencia del gen al cual llega.

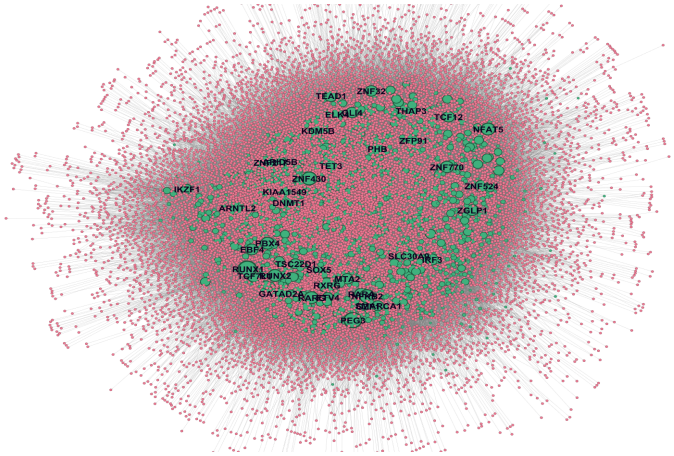

***Figura 1**: Layout de la red regulatoria generada con ARACNe a partir de 560 muestras de RNASeq obtenidas de TCGA. Los nodos verdes corresponden a factores de transcripciòn, y los nodos rosados a genes sin actividad regulatoria conocida. El diámetro de los nodos es función de su grado total (tanto ingresante como egresante).*

Una vez obtenida la red en cuestión se procedió a su caracterización. La tabla 1 muestra algunas de sus características topológicas principales, entre las cuales destacan su muy baja densidad (puesto que sólo hay enlaces que involucran factores de transcripción) y su bajo grado medio (puesto que abundan los genes con grado muy bajo que son regulados por uno o pocos factores de transcripción).

	Aristas	Densidad	Dirigido	Grado_max	Grado_medio	Grado_min	Nodos	Transitividad
ARACNe_net	77181	0.0004	Sí	376	10.853	1	14223	0.0037

***Tabla 1**: características generales de la red inferida.*

Además, y en concordancia con lo encontrado en bibliografía, la distribución de grado obtenida ajusta correctamente a una distribución *power-law* (figura 2). Este hecho, que puede interpretarse como libertad de escala, concuerda con la presencia de pocos nodos de alto grado (factores de transcripción que actúan como *hubs* en la red) y muchos genes de grado escaso, sin actividad regulatoria propia y regulados por uno o pocos factores de transcripción.

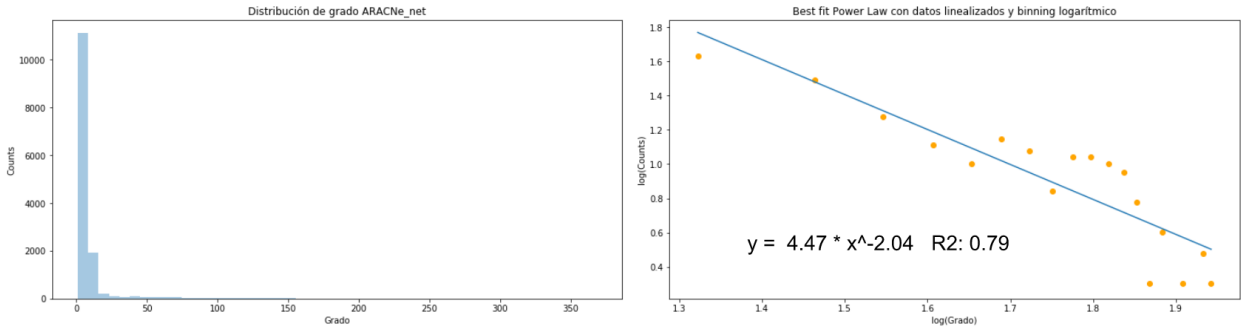

Figura 2: *Izquierda: distribución de grado para la red inferida. Derecha: ajuste de la distribución de grado a una power-law (escala logarítmica en ambos ejes y bineado logarítmico de los datos). Los resultados se condicen con la hipótesis de libertad de escala.*

Expresión diferencial

Como fue mencionado anteriormente, la red generada posee información sobre una gran variedad de estados que pueden afectar a la célula folicular tiroidea humana. Para estudiar el comportamiento de la red en la transformación al fenotipo canceroso, se buscó reducir el problema a las muestras relacionadas con los fenotipos de interés.

Para esto es necesario mencionar que muchos de los tejidos reportados en TCGA se encuentran a*pareados*. Es decir, que para un mismo paciente con la patología se extraen datos de expresión tanto de las células malignas como de sus vecinas no afectadas por el tumor, a fin de disminuir la variabilidad entre individuos a la hora de hacer el análisis. Teniendo esto en cuenta se filtraron las muestras disponibles, reteniendo sólo aquellas que correspondiesen a tumores PTC y a sus correspondientes tejidos normales apareados. Además, como distintos perfiles mutacionales pueden originar perfiles de expresión diferentes, se decidió retener sólo aquellas muestras provenientes de pacientes con mutaciones en el gen BRAF. Este gen, uno de los más comúnmente mutados en este tipo de cáncer, corresponde a una kinasa que forma parte de una vía muy conocida de proliferación, cuya sobreactivación se cree capaz de generar los desbalances observados a nivel regulatorio. Teniendo todo esto en cuenta, el análisis de expresión diferencial se llevó a cabo con 33 muestras tumorales (BRAF-like PTC) y las 33 muestras apareadas correspondientes de tejido normal. Un simple análisis de reducción dimensional (figura 3) muestra las diferencias globales entre los tejidos.

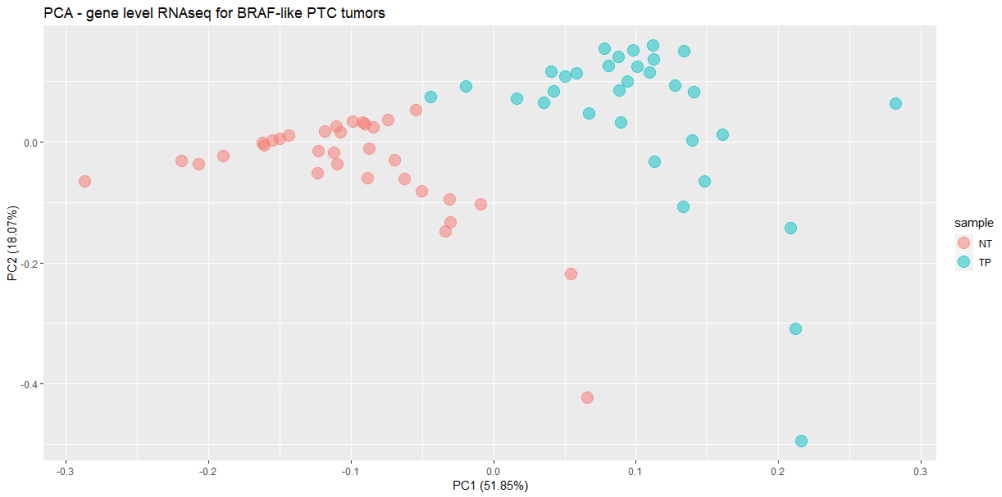

***Figura 3**. PCA que analiza las muestras para ambos tipos de cáncer de tiroides de acuerdo a las dos componentes principales. Los puntos rojos corresponden a muestras de tejido normal, y los puntos celestes a tejido tumoral.*

Para analizar qué genes se encuentran diferencialmente expresados entre ambos grupos de muestras, se realizó una comparación exacta de Fischer para cada uno de los genes. Los valores p obtenidos fueron corregidos para comparaciones múltiples por el método de Bonferroni.

La figura 4 consiste de un *heatmap* que representa los resultados obtenidos, en el cual cada columna representa una muestra y cada fila un gen.

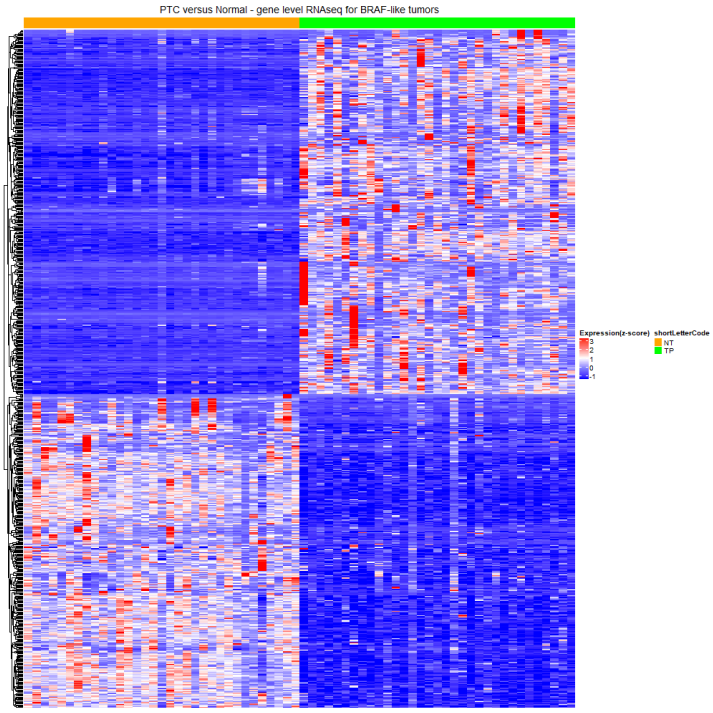

Figura 4. *Heatmap de expresión diferencial de tejido normal (franja naranja) y tejido canceroso BRAF-like (franja verde). Cada fila es un gen, cada columna una muestra. Color rojo representa sobreexpresión y color azul subexpresión, respecto de la media de la fila.*

Con estos resultados se confirmó que las muestras bajo estudio poseen efectivamente diferencias en su expresión génica. Es el momento entonces de volcar esta *signature* sobre la red, a fin de elucidar los mencionados reguladores maestros que operan en el sistema.

Análisis de reguladores maestros

Como fue mencionado, la actividad regulatoria en el contexto de un fenotipo celular específico puede ser investigada a través de redes de interacción o interactoma. En nuestro caso, el mismo es la red de expresión el construida por ARACNe.

El análisis de reguladores maestros (MRA por su sigla en inglés) es utilizado para identificar a los factores de transcripción que están enriquecidos en un determinado perfil (lista de genes diferencialmente expresados). Cruzando los datos del contexto regulatorio con los genes diferencialmente expresados, es posible identificar a los reguladores maestros responsables por coordinar dicha actividad la transformación y la manutención del fenotipo resultante.

En resumen, dados una red de interacciones *I*, un potencial regulador maestro *A* y un conjunto de genes diferencialmente expresados *S*, MRA computa el enriquecimiento de genes diferencialmente expresados de el regulón de A, donde él mismo se define como sus vecinos en la red de interacción *I*.

Para llevar a cabo esta tarea se utilizó el paquete de R *Viper*, que implementa dos algoritmos: MARINa (*MAster Regulator INference algorithm*) y VIPER (*Virtual Inference of Protein-activity by Enriched Regulon analysis*).

El primero es de nuestro particular interés, mientras que del segundo solo diremos que permite inferir la actividad de una proteína, de manera individual, a partir de datos de perfiles de expresión.

MARINa (rebautizado en versiones recientes como msVIPER) utiliza GSEA (*Gene Set Enrichment Analysis*) para calcular si el regulón de un factor de transcripción está enriquecido en genes diferencialmente expresados entre los dos perfiles de interés. El *cut-off* para determinarlo está dado por el p-value del método. Los genes con mayor puntaje son posteriormente reprocesados teniendo en cuenta dos aspectos:

- *Shadow effect* (o efecto sombra): chequea si existen escenarios en los cuales el regulón de algún factor de transcripción es representativo de la intersección de otros dos de alto puntaje. Este tipo de escenarios sirve para descartar casos en los cuales el regulón de un factor de transcripción se encuentra diferencialmente expresado en forma significativa, pero no por causa de una activación diferencial del mismo.
- Sinergia: si el set de genes corregulado por dos factores de transcripción se encuentra más significativamente enriquecido en genes diferencialmente expresados que la suma de los regulones individuales, el par es reportado como regulador maestro sinérgico.

Los autores demostraron en varios trabajos que analizando las redes de factores de transcripción inferidas por ARACNe con MARINa se identifican de manera eficiente los genes que dirigen fenotipos celulares específicos.

La figura 4 muestra los doce reguladores maestros más significativos obtenidos para la transformación de la célula tiroidea al fenotipo PTC, en orden de activación.

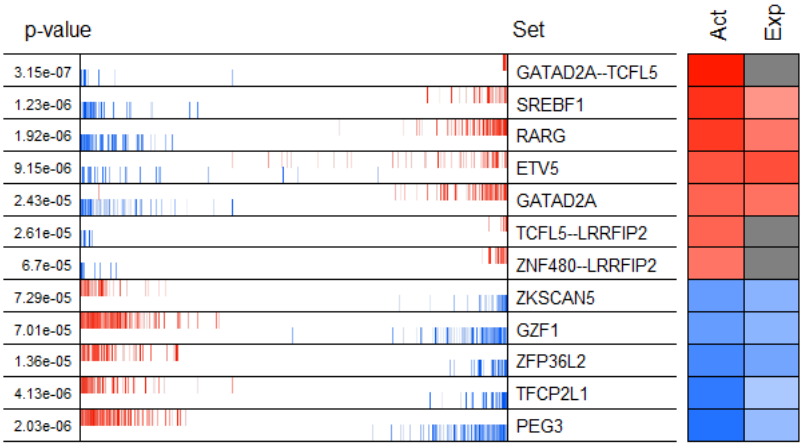

Figura 5: *Output de MARINa/msVIPER. Muestra un gen enriquecido en vecinos diferencialmente expresados, ordenados por actividad decreciente. La primera columna muestra la significancia estadística, la segunda el perfil de genes, donde rojo implica regulado positivamente y azul negativamente mientras que la tercera es el nombre del gen y la cuarta/quinta el nivel de actividad/expresión con el mismo código de colores que el perfil, respectivamente.*

Una vez obtenidos los reguladores maestros, algo interesante a corroborar es que tanto correlaciona su significancia con las métricas de grado. ¿Un regulador maestro lo es sólo porque tiene mucho alcance en la red (regula a muchos genes)? ¿O hay realmente vías específicas que están siendo activadas independientemente del grado?

Si bien esto no fue testeado estadísticamente, se procedió a desarmar la red teniendo en cuenta el grado decreciente, *betweenness centrality* decreciente (una medida que indica cuántos de los caminos más cortos entre nodos de la red pasan por un nodo en particular), y la significancia como RM de los factores de transcripción (también en forma decreciente).

Como es de esperar, la componente gigante se desarma más lentamente por remoción de nodos por azar que por cualquiera de los otros criterios (figura 6, izquierda). Por otro lado, si bien este gráfico sugiere que la componente gigante se desarma al mismo ritmo tanto por grado, como centralidad o por reguladores maestros en cáncer BRAF-like y RAS-like (otro perfil mutacional importante en cáncer de tiroides PTC, también analizado pero removido del informe por cuestiones de espacio), una observación detallada vuelve evidente que el desarmado es sutilmente más lento (y cualitativamente distinto) por criterio de reguladores maestros en ambos tipos de cáncer que por centralidad o grado del nodo. Esto es fácilmente interpretable: si bien los reguladores maestros son factores de transcripción, y por el modo en el que está armada la red estos tienden a tener de por sí un grado mucho más alto que los nodos tomados al azar, el enriquecimiento en genes diferencialmente expresados no correlaciona en forma lineal con el grado. Hay más factores a tener en cuenta que el mero alcance de los reguladores en la red (por ejemplo, especificidad en las vías que cada uno regula).

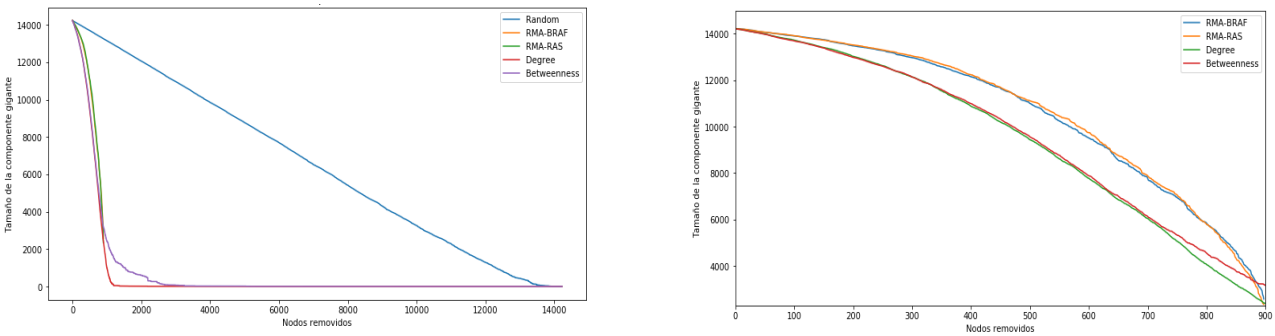

Figura 6. *Izquierda: Tamaño de la componente gigante de la red en función de la cantidad de nodos removidos de esta según distintos criterios. Derecha: Los primeros 900 nodos removidos de la red. La curva correspondiente a remoción de nodos por azar no fue incluída.*

Discusión

La metodología empleada permitió obtener en forma satisfactoria postulantes a reguladores maestros entre los dos fenotipos de interés células normales y tejidos afectados por carcinoma papilar tiroideo). De la lista mostrada anteriormente, se ejemplifican los resultados con los factores de transcripción PEG3 y RARG (figuras 7 y 8). Estos fueron seleccionados por su comportamiento antagónico (uno es inhibido y el otro activado) y por los resultados obtenidos al analizar el set de genes que regulan.

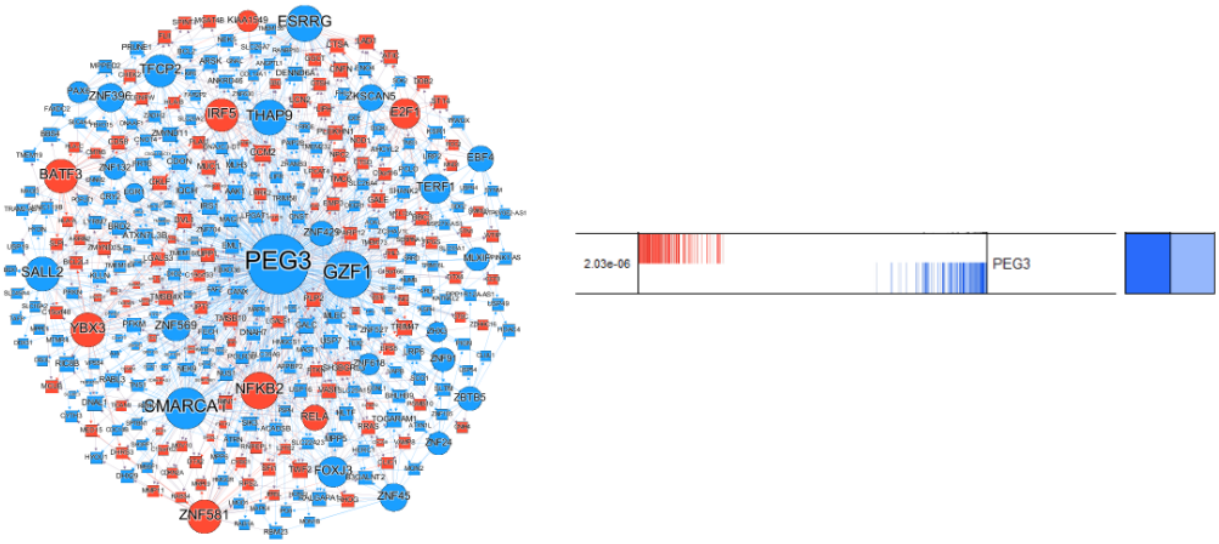

***Figura 7**: Red ego del factor de transcripción PEG3 en fenotipo canceroso. Los nodos con contorno circular corresponden a factores de transcripción, mientras que aquellos que no lo son se muestran con un contorno rectangular. Azul y rojo implican subexpresión y sobreexpresión en cáncer con respecto al tejido normal, respectivamente. El tamaño de los nodos es función de su conectividad total.*

Como puede observarse, la mayoría de los vecinos de ambos ejemplos está expresado diferencialmente. En el caso de PEG3, se ve que está subexpresado en cáncer con respecto al tejido normal. Como todos los genes normalmente regulados positivamente por él pierden su condición de activación, se infiere que los genes que este factor de transcripción activa son los que figuran en azul en el grafo (figura 7, nodos azules a la izquierda y líneas verticales azules en la segunda columna de la derecha). Los genes normalmente reprimidos, por el contrario, pierden la restricción a su expresión y pasan a estar sobreexpresados en el tejido canceroso. Vemos como en este caso, la afección de un sólo factor de transcripción repercute **directamente** en la expresión de otros 350 genes. El alcance neto es aún mayor si tenemos en cuenta que varios de los genes afectados son factores de transcripción también, por lo que el efecto se propaga aún más a través de la red (queda pendiente cuantificar cuánto).

Algo remarcable que se encontró indagando el set de genes directamente regulados por este factor de transcripción es que se visualiza la activación de la vía NF-kB, directamente implicada en proliferación celular y ampliamente estudiada en cáncer en general. La figura 9 muestra cómo ciertos genes clave en esta vía son alcanzados directamente por PEG3, dando sustento a su relación con cáncer.

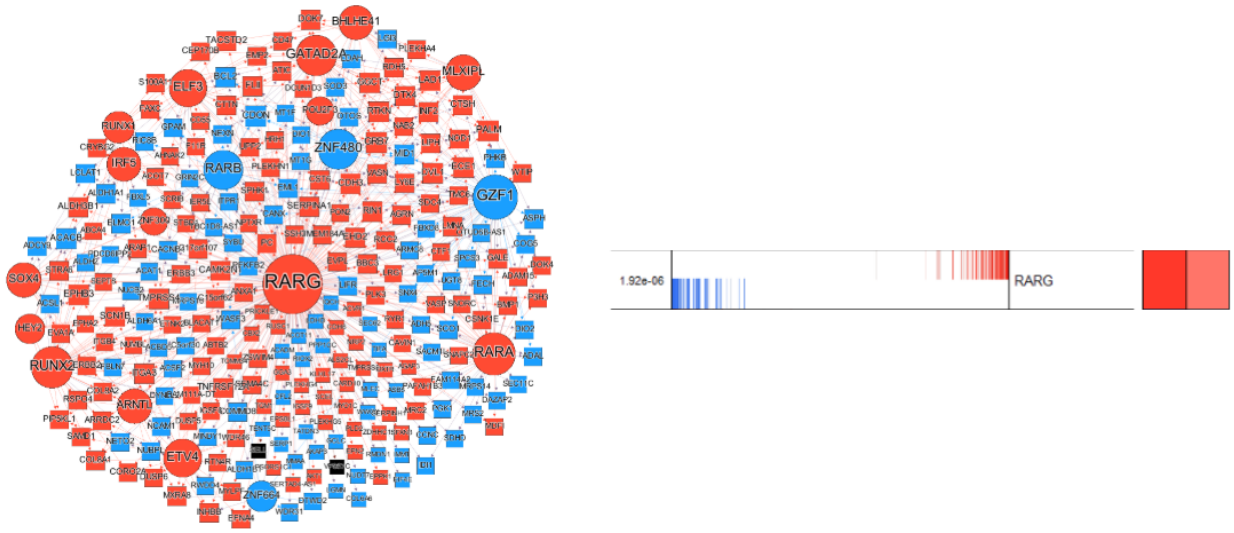

***Figura 8**: Red ego del factor de transcripción RARG en fenotipo canceroso. Los nodos con contorno circular corresponden a factores de transcripción, mientras que aquellos que no lo son se muestran con un contorno rectangular. Azul y rojo implican subexpresión y sobreexpresión en cáncer con respecto al tejido normal, respectivamente. El tamaño de los nodos es función de su conectividad total.*

En el caso de RARG, por el contrario, se observa una sobreexpresión en los tejidos tumorales con respecto al control. En el fenotipo canceroso bajo estudio, RARG está activando a sus vecinos sobreexpresados e inhibiendo a sus vecinos subexpresados. En cambio en el fenotipo normal, los vecinos subexpresados por RARG pasan a estar activados (debido a la inhibición de RARG) y viceversa con los activados en fenotipo canceroso. El alcance directo en este caso es de 268 genes, de nuevo propagables aún más a través de los factores de transcripción regulados por este gen.

Algo interesante para comentar respecto de este subgrafo, es que un análisis de ontología génica (realizado con la plataforma en línea DAVID), reveló un enriquecimiento significativo en genes relacionados con carcinoma tiroideo (lo que es de esperar, y no hace más que validar la expresión diferencial calculada entre nuestras muestras).

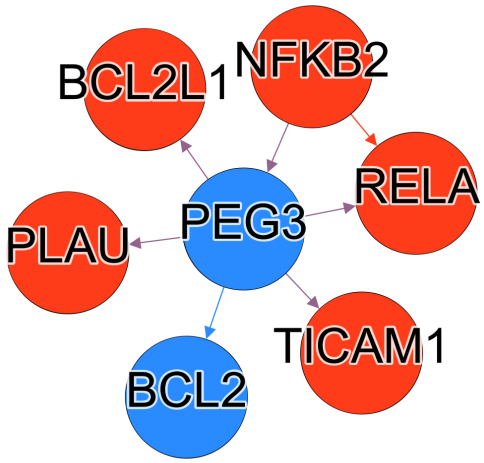

Figura 9: *Genes relacionados directamente con la vía de proliferación activada por el complejo regulatorio NF-kB, conocido mediador de proliferación celular, directamente regulados transcripcionalmente por el gen PEG3, identificado como regulador maestro en el análisis presentado. Se observan activadas NFKB2 y RELA, subunidades del complejo mencionado y sobreexpresados TICAM2 y PLAU, genes modulados también por NF-kB y relacionados con proliferación celular. BCL y BCL2L1 son genes pro y antiapoptóticos respectivamente. Su modulación en esta condición redunda en una reducción de la apoptosis celular, fenotipo esperable en un entorno canceroso.*

Conclusiones y perspectivas a futuro

Como primera conclusión se puede decir que fue exitosa la obtención de la red de interacción génica a partir de los datos de RNAseq. Así mismo fue posible también establecer un listado de reguladores maestros coherente con la biología previamente conocida de las células cancerosas en cuestión.

Por otro lado, el desarmado de la red dio a entender que los reguladores maestros no tienen una correlación estrictamente directa con su grado o centralidad en la red en cuestión.

Si bien estos análisis son prometedores, no debe dejar de tenerse en cuenta que las redes están inferidas bioinformáticamente, y que, si bien estos métodos están ampliamente validados en la bibliografía citada, la validación experimental de las redes no es un detalle menor.

Para terminar, cabe mencionar que a partir de este tipo de estudios pueden indagarse funcionalmente los reguladores maestros y su relación con el fenotipo canceroso (de hecho, este informe contiene datos que van a ser usados para esto en serio). En esta línea, la idea principal es revertir la expresión de los RM seleccionados en líneas celulares representativas del fenotipo tumoral. Esto puede llevarse a cabo por expresión del gen exógeno cuando el target está subexpresado, o por el silenciamiento del mensajero por RNA de interferencia cuando el regulador maestro se expresa en exceso. La reversión del patrón de expresión asociado debería, hipotéticamente, verse reflejada en variables típicas del fenotipo maligno tales como proliferación y migración.

Referencias

1. Pacifico F, Paolillo M, Chiappetta G, Crescenzi E, Arena S, Scaloni A, Monaco M, Vascotto C, Tell G, Formisano S, Leonardi A. RbAp48 is a target of nuclear factor-kappaB activity in thyroid cancer. J Clin Endocrinol Metab 2007; 92:1458-1466
2. Aubry, Soline, William Shin, John F. Crary, Roger Lefort, Yasir H. Qureshi, Celine Lefebvre, Andrea Califano, and Michael L. Shelanski. 2015. “Assembly and Interrogation of Alzheimer’s Disease Genetic Networks Reveal Novel Regulators of Progression.” *PloS One* 10 (3): e0120352.
3. Lefebvre, Celine, Presha Rajbhandari, Mariano J. Alvarez, Pradeep Bandaru, Wei Keat Lim, Mai Sato, Kai Wang, et al. 2010. “A Human B-Cell Interactome Identifies MYB and FOXM1 as Master Regulators of Proliferation in Germinal Centers.” *Molecular Systems Biology* 6 (June): 377.
4. Chen, James C., Mariano J. Alvarez, Flaminia Talos, Harshil Dhruv, Gabrielle E. Rieckhof, Archana Iyer, Kristin L. Diefes, et al. 2016. “Identification of Causal Genetic Drivers of Human Disease through Systems-Level Analysis of Regulatory Networks.” *Cell* 166 (4): 1055.
5. Lachmann, Alexander, Federico M. Giorgi, Gonzalo Lopez, and Andrea Califano. 2016. “ARACNe-AP: Gene Network Reverse Engineering through Adaptive Partitioning Inference of Mutual Information.” *Bioinformatics* 32 (14): 2233–35.
6. Califano, Andrea. 2014. “Predicting Protein Networks in Cancer.” *Nature Genetics* 46 (12): 1252–53.
7. Nazar M, Nicola JP, Velez ML, Pellizas CG, Masini-Repiso AM. Thyroid peroxidase gene expression is induced by lipopolysaccharide involving nuclear factor (NF)-kappaB p65 subunit. Endocrinology 2012; 153:6114-6125
8. Nicola JP, Nazar M, Mascanfroni ID, Pellizas CG, Masini-Repiso AM. NF-kappaB p65 subunit mediates lipopolysaccharide-induced Na(+)/I(-) symporter gene expression. Mol Endocrinol 2010; 24:1846-1862