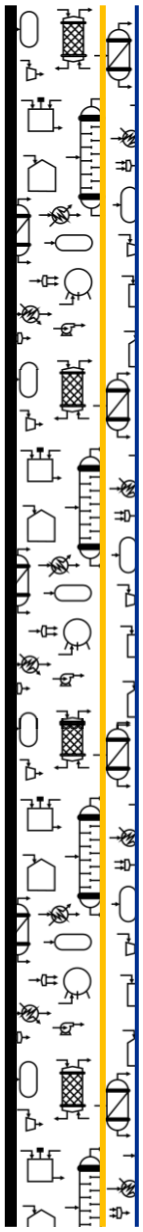
Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Zaragoza

Carrera de Ingeniera Química



Análisis Bioestadístico de las determinantes transcripcionales de una función celular en diversos tejidos humanos

T E S I S

Que para obtener el título de

INGENIERO QUIMICO

P R E S E N T A:

GONZALO RIVERA ORTIZ

Director de Tesis:

Dr. Humberto Gutiérrez González

Asesores:

QFB. Alfonso Macario Luna Vázquez

IQ. José Antonio Zamora Plata

CD. MX. 2023

índice

Agradecimientos

Quiero agradecer especialmente al Dr. Humberto Gutiérrez González, quien despertó en mí un gran interés por las ciencias genómicas. Su guía continua y creciente ha sido invaluable en el desarrollo de mi trabajo.

A la UNAM, quien como dicen, ha sido mi segunda casa, que me ha proporcionado un lugar en donde he conocido gente maravillosa e increíble, cuyos profesores me han preparado de manera excepcional en todo lo que sé que han moldeado mi carácter y mi hambre de saber.

A mis compañeros de instituto, Omar, Selene, Adrián y Fernanda, quiero expresar mi gratitud por su compañía y apoyo a lo largo de este camino. Aunque quizás no estuvieron directamente involucrados en mi trabajo, su presencia y consejos han aligerado la carga y han sido un apoyo invaluable.

También quiero agradecer al INMEGEN por abrirme las puertas de sus laboratorios y brindarme la oportunidad de desarrollarme en una nueva área del conocimiento.

Finalmente, quiero expresar mi agradecimiento a los profesores que forman parte de mi sínodo. Soy consciente de que no es una tarea fácil, mucho menos con las actividades que desempeña cada uno de ellos como docente, pero su dedicación y compromiso solo reafirma la profesionalidad que poseen.

Dedicatorias

Este trabajo y todo el esfuerzo invertido en el están dedicados a todas las personas que han pasado por mi vida. A mis seres queridos, a mis amigos cercanos y a todos aquellos conocidos que han dejado una huella en mi camino. Su presencia ha sido fundamental en este viaje y quiero expresar mi profundo agradecimiento por su apoyo incondicional.

A mis padres y hermanos cuyo amor y aliento han sido una de mis grandes motivaciones, han estado presente en cada paso de este camino. Su fe en mí y el constante apoyo han sido el cimiento sobre el cual he construido mi visión, la determinación y perseverancia que me caracterizan.

A Lisset Monserrath F.V., quien sin mirar atrás me ha apoyado día y noche desde el momento en que nos conocimos, espero algún día poder darte tanto como tu me has dado a mí, gracias por coincidir y decidir acompañarme en mi camino, esto es solo uno de los muchos éxitos que cosecharemos juntos en un futuro, siempre con la voluntad de dios de por medio.

A mis amigos más cercanos y a quienes no puedo olvidar mencionar; Marcos, Karen, Marlene, Nattaly, Jorge G., Eduardo, Valeria M., Antonio, Michelle, Manuel, Ileana, Miguel, Jaqueline Q., Abigail M., Jonathan, Joshua, Javier, Araceli, Rene, Mónica, Abigail J., Juan, Enrique, Miriam, Maricela, Sandra, Elizabeth M., Daniela, Bryan, Anahí, Casandra, Emilio, Brandon, Andrés, Beatriz, Marianna, Itzel, Giselle, Alondra, Diana S., Luis, Axel, Kevin, Nancy, Jorge A., Alexis, Ángel, Evelyn, Paulina, Elizabeth A., Maribel, Ana, Brenda S., Aaron, Jacobo, Brenda E., Sandra L., Sheila, Jaqueline S., Jesús, Johnny, Jordán, Jerry, Valeria S., Alexandra, Fátima, Guadalupe, Fernando, Barbara, Diana I., que sepan que a pesar del tiempo y la distancia, muchas cosas de las que me traen aquí hoy se las debo a ustedes, y siempre sepan que ustedes son los mejores amigos que una persona podía desear.

A mis profesores Prof. Sergio, Profa. Socorro, Profa. Alondra, Prof. Jaime, Profa. Guadalupe, Profa. Isabel, Profa. Josefina, Prof. Ariel, Prof. Alejandro J, Prof. Aldo, Prof. Jorge, Prof. Jose, Prof. Alejandro R., Prof. Crescenciano, Prof. Noé, Prof. Fabricio, Sra. Lorena, Sra. Lucia, Pbro.

Miguel, Pbro. José, que a lo largo de mi historia han cumplido magistralmente en su labor de enseñar y guía.

A todos los conocidos que han cruzado mi camino, quiero expresar mi gratitud por el impacto que han tenido en mi vida. Cada historia, la experiencia y toda la diversidad que representan en mi camino, ha enriquecido mi perspectiva y ha sido una fuente constante de inspiración. Cada conversación, cada interacción, ha sido una oportunidad para aprender y crecer.

A todos ustedes, que han dejado una marca en mi camino, les dedico este pequeño gran éxito, su presencia ha sido una bendición en este mundo tan volátil y lleno de incertidumbre, donde nada es certero y la única constante es el cambio, porque *todo depende de todo*.

Abreviaturas

CC
Ciclo Celular

C
Cerebro

T
Tejidos

L
Líneas Celulares

CB
Cerebelo

CX
Corteza

ST
Estriado

Resumen

Nuestros genes son los encargados de dirigir el desarrollo, mantenimiento y complejidad funcional del organismo. Esto significa que todas las funciones que un sistema vivo puede llevar a cabo no son el resultado de genes aislados actuando de forma independiente, sino que se trata de la colaboración de cientos o miles de ellos, actuando en conjunto y en estrecha coordinación funcional. Sin embargo, se desconoce el mecanismo por el cual esta coordinación funcional tiene lugar al nivel de la expresión de los genes. Específicamente se desconoce si la activación de una determinada función que requiere de la participación de múltiples genes es el resultado del aumento colectivo de la expresión de estos genes o, alternativamente, del nivel de coordinación (correlación) colectiva entre ellos con independencia de su nivel absoluto de expresión. Esto es, ¿Qué tanto una función se encuentra más relacionada con el nivel colectivo de actividad o expresión de los genes asociados en contraposición con qué tan correlacionada colectivamente se encuentra esta actividad o expresión?

Con el objetivo de responder a esta pregunta, se utilizaron datos de expresión génica públicamente disponibles para determinar si un conjunto de genes que se sabe están involucrados en una función bien definida, específicamente los genes involucrados en la función del ciclo celular, muestran un nivel alto de actividad colectiva o, alternativamente, un nivel alto de correlación en su actividad en tejidos con mayor actividad proliferativa. Para esto se compilaron datos de expresión derivados de líneas celulares, tejidos humanos no nervios y tejido nervioso (con niveles alto, medio y bajo actividad proliferativa respectivamente), así como datos de expresión derivados separadamente de tres regiones cerebrales: cuerpo estriado, corteza y cerebelo (alta, media y baja actividad proliferativa, respectivamente); y se midió el nivel colectivo de expresión y coexpresión de estos genes.

Los resultados muestran una más consistente asociación entre la coexpresión y el nivel relativo de actividad proliferativa que la que se observa con el nivel colectivo de expresión, sugiriendo, en general, que el nivel colectivo de coordinación en la expresión de genes asociados a una función celular, es un más robusto indicador de reclutamiento funcional que el nivel colectivo de expresión.

1. Introducción

Está claro que toda función biológica, a nivel celular o fisiológico requiere de la participación de cientos o miles de genes actuando juntos. Por consecuencia, cada gen tiene el potencial de contribuir, en combinación con muchos otros genes, en más de una función y jugar en ellas un papel igualmente crítico[1]. Sin embargo, el modo en que los genes interactúan dinámicamente entre sí para instruir funciones complejas es algo que actualmente se entiende poco.

1.1. Determinantes transcripcionales asociados a una función

En el presente trabajo, entendemos por determinantes transcripcionales a los genes o productos génicos que directamente contribuyen con la integración de alguna función celular o fisiológica. Estos productos génicos, a su vez, pueden constituir por sí mismos, moléculas funcionales (RNAs reguladores, microRNAs etc.), o bien dar lugar, a través de la traducción de proteínas a moléculas tales como factores de transcripción, proteínas reguladoras o estructurales, moléculas de señalización, o modificaciones epigenéticas, que pueden actuar de manera individual o en conjunto para integrar una función celular determinada.

Los factores de transcripción son proteínas que se unen a regiones específicas del ADN, como los promotores y enhancers (“potenciadores”), para iniciar o reprimir la transcripción del gen, razón por la cual reciben su nombre. Estos factores pueden activar o inhibir la transcripción de un gen, y su presencia o ausencia puede determinar la expresión de un gen específico.[2]

Las proteínas reguladoras también pueden afectar la expresión génica al interactuar con los factores de transcripción o directamente con el ADN. Estas proteínas pueden actuar como coactivadores o correpresores de la transcripción, y su presencia o ausencia puede influir en la activación o represión del gen.[3]

Las moléculas de señalización también pueden influir en la expresión génica al activar o reprimir factores de transcripción. Estas moléculas pueden ser producidas por las células mismas o pueden provenir de señales extracelulares, como hormonas o factores de crecimiento.[4]

Por último, en la lista de Determinantes transcripcionales, las modificaciones epigenéticas son cambios en el ADN o en las histonas que pueden afectar la accesibilidad del ADN y, por lo tanto, la expresión génica. Estas modificaciones pueden incluir metilación del ADN, acetilación de histonas y modificaciones de la estructura de la cromatina. [5]

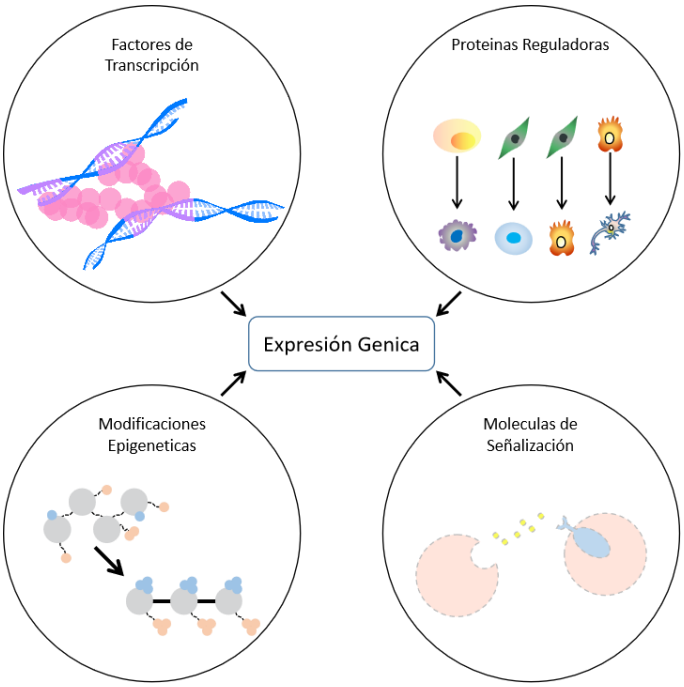


Figura 1. ***Determinantes Transcripcionales*** Reguladores de la expresión génica.

Las determinantes transcripcionales son moléculas que interactúan con los genes para regular o modificar su expresión, y pueden incluir a las moléculas o proteínas anteriormente mencionadas. Reiterando, estas señales pueden actuar de manera individual o en conjunto para controlar la expresión génica y son esenciales para el desarrollo, el crecimiento y la función normal de una célula y de cualquiera de sus funciones.

1.2. Expresión y Coexpresión

La expresión génica se refiere a la cantidad de un gen específico que se transcribe en una célula en particular [6]. Cuando un gen está activo, su ADN se transcribe en ARN y luego se traduce en una proteína que realiza una función específica en la célula. La coexpresión se refiere a la tendencia de dos o más genes a ser transcritos de manera coordinada o sincrónica en la misma célula o tejido. La coexpresión puede sugerir que estos genes comparten una función común o están regulados por los mismos factores de transcripción. [7]

Los factores de transcripción pueden afectar la expresión de un solo gen, de varios o pueden actuar en cascada para afectar la expresión de múltiples genes [8].

La expresión y coexpresión génica se pueden utilizar como medidas de las determinantes transcripcionales. Por ejemplo, si se observa que dos genes están coexpresados en un conjunto de

células, esto puede sugerir que comparten un factor de transcripción o una vía reguladora común. Del mismo modo, si se observa que un gen está sobre expresado en un conjunto de células, esto puede sugerir que la activación de este gen juega un papel determinante en las características fenotípicas de ese conjunto celular [9].

Para analizar la expresión y coexpresión génica, se pueden utilizar diversas técnicas que miden el perfil de expresión de miles de genes, como los microarreglos y la secuenciación masiva (RNA-seq). Estas técnicas permiten analizar la expresión de miles de genes a la vez, y pueden utilizarse para identificar patrones de expresión y coexpresión en diferentes tipos de células y tejidos [10].

1.3. Función Biológica

Las funciones biológicas son los procesos que ocurren en un organismo vivo para mantener su supervivencia, crecimiento y reproducción. Estos procesos incluyen la regulación del ciclo celular, la síntesis y degradación de proteínas, el transporte de nutrientes y la comunicación entre células.

Según Lehninger et al., los procesos biológicos son "acciones coordinadas y reguladas de moléculas, células, tejidos y órganos que permiten el correcto funcionamiento del organismo vivo" [11]. Por su parte, Alberts et al., destacan que los procesos biológicos son "el resultado de la interacción entre moléculas individuales y los sistemas biológicos en los que están incrustados" [6].

1.4. Redes de Coexpresión y las funciones Biológicas

Un aspecto conocido de los perfiles globales de expresión o actividad génica en todos los organismos es que esta se encuentra organizada en subpoblaciones de genes, en donde decenas, cientos o incluso miles de genes muestran un patrón de expresión coordinada [12].

Cuando decimos que los genes tienen una expresión coordinada, nos referimos a que muchos genes diferentes pueden trabajar juntos para llevar a cabo una función específica en el cuerpo. A estas subpoblaciones de genes con expresión altamente correlacionada se les conoce como módulos o redes de coexpresión. Esto significa que cuando un gen en el módulo se activa, es probable que otros genes en el mismo módulo también se activen al mismo tiempo [13]–[15].

Esta organización en módulos de coexpresión es importante porque sugiere que los grupos de genes que trabajan juntos en una función específica también tienen una actividad coordinada y, por lo tanto, es probable que estén regulados de manera similar. Esto es útil para entender cómo se llevan a cabo las funciones en el cuerpo y cómo pueden ser reguladas [16], [17].

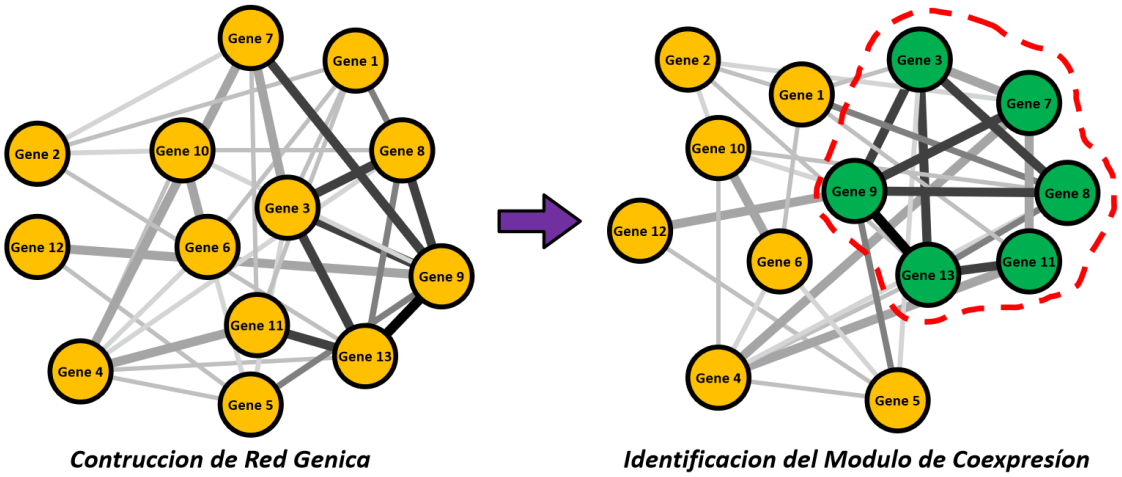


Figura 2. ***Representación esquemática de redes de coexpresión*** en la que cada gen se representa por un círculo y las líneas que los unen representan el nivel de coexpresión (donde el índice de correlación absoluta es proporcional a la intensidad de color de la línea). En esta figura, un módulo de coexpresión se representa como un conjunto de genes altamente correlacionados entres sí (nodos verdes en el esquema de la derecha

Sin embargo, tradicionalmente se ha utilizado el nivel absoluto de actividad o expresión de los genes individuales, asociados a una determinada función, como el indicador maestro de la activación de dicha función [18], [19]. Esto deriva de la suposición de que cuando una actividad génica es requerida, esto debe venir acompañado de un aumento de la expresión del gen o genes asociados. Sin embargo, hasta el momento no hay evidencia empírica consistente que indique que el nivel de activación de alguna función necesariamente vendría acompañado de un aumento en la expresión de los genes asociados, en oposición con un ajuste, en un sentido u otro en sus niveles de expresión. En contraste con el nivel absoluto de expresión, la coexpresión da una medida del grado de coordinación (y potencial corregulación) de un conjunto de genes y, por lo tanto, podría representar una métrica alternativa del nivel de activación de una función, independiente del nivel absoluto de expresión de un conjunto de genes. Sin embargo, hasta el momento, se desconoce si la activación de una función se encuentra más relacionado al nivel de actividad colectiva o

expresión de los genes asociados a dicha función, que al patrón de colectivo de correlación o actividad coordinada del mismo grupo de genes (coexpresión).

1.5. Planteamiento del Problema

Como se ha mencionado previamente, los genes desempeñan un papel crucial en todas las funciones y procesos celulares y fisiológicos [20]. De igual forma ya sabemos que los genes no operan de manera aislada y que la función celular es el resultado de la coordinación entre numerosos genes [21].

El objetivo de este estudio es investigar la relación entre el nivel de actividad o expresión genética (expresión) o, alternativamente, el nivel de coordinación (coexpresión) de un conjunto de genes implicados en una función celular específica y con el nivel de requerimiento de esa función biológica en un tejido particular. En otras palabras, buscamos responder a la pregunta:

¿Cuál de estos dos parámetros, el nivel de expresión o el nivel de coexpresión, de un conjunto de genes implicados en una función, proporciona una indicación más precisa de la activación de esa función?

Para llevar a cabo este estudio, utilizaremos como modelo de función el conjunto de genes conocidos que están implicados en la función de proliferación celular, también conocida como ciclo celular (***CC***). Específicamente, los genes anotados como GO:0007049, según su caracterización y anotación en la base de datos Gene Ontology [22], [23].

La función de ciclo celular (***CC***) es el proceso que experimentan las células para dividirse y crear nuevas células. Este proceso se divide en dos fases principales: la interfase y la fase mitótica. Durante la interfase, la célula crece, se replica el material genético y se prepara para la división celular. En la fase mitótica, la célula se divide en dos células hijas. El ***CC*** está regulado por una serie de proteínas, incluyendo las cinasas dependientes de ciclina (CDK) y las ciclinas. Estas proteínas actúan como interruptores que controlan el avance del ciclo celular a través de las diferentes fases. Las mutaciones en genes que regulan el ciclo celular pueden llevar a una proliferación celular incontrolada y, por tanto, a la formación de tumores cancerosos [24].

Utilizando este modelo y aprovechando una serie de herramientas de genómica computacional [25], generáremos varios algoritmos capaces de evaluar el nivel de expresión y coexpresión de estos genes involucrados en la función de ***CC*** en distintos tejidos con variados niveles de demanda de actividad proliferativa. La coexpresión y expresión colectiva de estos genes se comparará con genes de fondo que no están anotados en esta función, pero que igualmente están presentes y expresándose en los tejidos evaluados.

Nuestro objetivo es generar datos comparables, y para ello mediremos el grado de actividad o expresión colectiva de estos genes, así como la correlación entre ellos.

1.6. Hipótesis

En este trabajo lo que se espera observar es que el requerimiento funcional de un conjunto de genes asociados a cierta función, o, en otras palabras, la demanda de dicha función en los genes involucrados se caracterizará ya sea por un aumento colectivo en la expresión de estos, o por un incremento en el grado de sincronización o coordinación en su expresión.

1.7. Objetivos

1.7.1. Objetivo General

Determinar si la actividad de una función celular, como es la función proliferativa, está más relacionada con el nivel de expresión colectiva de los genes asociados a esta función (***CC***), o si se encuentra más fuertemente asociada al nivel colectivo de expresión coordinada (coexpresión) de estos mismos genes.

1.7.2. Objetivos Particulares

1. Determinar el nivel colectivo de expresión de genes asociados al ciclo celular (n ~ 1700) en diversos tejidos con alta, media y baja actividad proliferativa y contrastar con la expresión media de genes de fondo (no asociados al ciclo celular) en los mismos tejidos

2. Determinar el nivel colectivo de coexpresión (correlación de Spearman media) de genes asociados al ciclo celular en tejidos con alta, media y baja actividad proliferativa, contrastar con la coexpresión media de genes de fondo (no asociados al ciclo celular) en los mismos tejidos

3. Utilizar pruebas pareadas en genes asociados al ciclo celular para cuantificar la significancia estadística en el grado de cambio en expresión de estos genes entre pares de tejidos con alta, media y baja actividad proliferativa.

4. Utilizar pruebas pareadas en genes asociados al ciclo celular para cuantificar la significancia estadística en el grado de cambio en expresión de estos genes entre pares de tejidos con alta, media y baja actividad proliferativa.

5. Utilizar estrategias de análisis de redes para determinar cambios en la red de coexpresión de genes asociados al ciclo celular en tejidos con alta, media y baja actividad proliferativa.

2. Métodos y Herramientas

Como se mencionó anteriormente, para este proyecto se necesita una serie de herramientas computacionales; lenguajes de programación, programas especializados etc., así como algunos métodos estadísticos. A continuación, se describen estas herramientas y métodos a usar en este proyecto:

2.1. Herramientas

2.1.1. R (Lenguaje de Programación Estadístico)

R es un lenguaje de programación estadístico ampliamente utilizado en la bioinformática y la genómica computacional. Es una herramienta con gran poder para el análisis de datos de expresión génica, ya que proporciona una amplia gama de funciones estadísticas y herramientas gráficas para visualizar y analizar datos biológicos.

R permite el procesamiento y análisis de grandes conjuntos de datos de expresión génica de manera eficiente, lo que lo convierte en una herramienta valiosa en la investigación genómica.

2.1.2. Librerías de R

Al ser un lenguaje de programación, permite a los usuarios definir sus propias funciones y por lo tanto generar sus librerías (paquetes de funciones), extendiendo así el potencial de R, ya que de esta forma cuenta con numerosas librerías (paquetes) específicas para el análisis de expresión génica, como DESeq2, edgeR, limma, entre otros. Estas librerías proporcionan métodos y algoritmos avanzados para el análisis de datos de expresión génica, como la normalización, el análisis diferencial de expresión, la identificación de patrones de expresión y la visualización de resultados.

Muchas de estas librerías son ampliamente utilizadas por los investigadores en genómica para realizar análisis estadísticos y obtener información biológica relevante a partir de los datos de expresión génica.

Mencionando algunas de las librerías utilizadas; GGally (Gráficas de Abanico), RCy3 (Conexión con Cytoscape), ggplot2 (Gráficos Avanzados), así como las funciones ya incluidas de forma predeterminada en R.

2.1.3. R Studio

R Studio es un *entorno de desarrollo integrado* (IDE), implementado para el lenguaje de programación R. Este proporciona una interfaz gráfica de usuario y herramientas adicionales que facilitan la escritura, ejecución y visualización de código en R.

R Studio es una herramienta que potencia aún más las capacidades del lenguaje R ya que mejora la productividad y facilita la organización y visualización de los resultados del análisis de expresión génica.

2.1.4. Cytoscape

Es una herramienta de visualización y análisis de redes biológicas. Permite la representación gráfica de redes de interacción molecular, incluyendo redes de expresión génica y redes de interacción proteína-proteína. Cytoscape es utilizado en la investigación genómica para visualizar y analizar las redes biológicas relacionadas con la expresión génica, lo que ayuda a comprender las relaciones funcionales entre genes y proteínas.

2.1.5. GeneOntology (GO)

GO es una base de datos y una ontología que proporciona una anotación funcional estandarizada de genes y proteínas. GO categoriza los genes y proteínas en términos de su función biológica, proceso biológico y localización celular. Los investigadores pueden utilizar GO para interpretar los resultados de los análisis de expresión génica, identificando las funciones biológicas y los procesos celulares asociados a los genes diferencialmente expresados. [22], [23]

2.2. Métodos

2.2.1. Comparación de la Distribución

Este método implica la comparación de las distribuciones de expresión génica entre diferentes grupos o condiciones. Se utiliza para identificar genes cuya expresión difiere significativamente entre dos o más grupos. Esto puede ser útil para identificar genes que están regulados de manera diferencial en diferentes condiciones, como en diferentes tejidos, estados de enfermedad o etapas del desarrollo. Se puede realizar mediante métodos estadísticos como la prueba de Kolmogorov-Smirnov, la prueba de Anderson-Darling o la prueba de Anderson-Darling modificada.

2.2.2. Correlación de Spearman

Este método implica la medición de la correlación entre la expresión de dos genes, basada en la clasificación de sus niveles de expresión en lugar de los valores numéricos exactos. Es una medida no paramétrica de la correlación que puede capturar relaciones no lineales entre genes.

La correlación de Spearman se utiliza para identificar genes que están coexpresados, es decir, que tienen patrones similares de expresión en un conjunto de datos. Esto puede ayudar a identificar genes que pueden estar involucrados en las mismas vías biológicas o procesos celulares.

2.2.3 logFC

Este método implica el cálculo del cambio relativo en la expresión génica entre dos condiciones o grupos. El logFC (logaritmo del cambio relativo en la expresión) se utiliza para identificar genes que muestran cambios significativos en su expresión entre diferentes condiciones experimentales. Se expresa en una escala logarítmica y representa la magnitud del cambio relativo en la expresión de un gen. Se utiliza comúnmente en análisis de expresión génica para identificar genes que están regulados de manera diferencial en diferentes condiciones experimentales.

2.2.4. Test de Wilcoxon

Este método es una prueba estadística no paramétrica que se utiliza para comparar dos grupos o condiciones en términos de sus niveles de expresión génica. Se utiliza cuando los datos no siguen una distribución normal o cuando el tamaño de la muestra es pequeño. La prueba de Wilcoxon se utiliza para identificar genes cuya expresión difiere significativamente entre dos grupos o condiciones experimentales, lo que puede ayudar a identificar genes candidatos relacionados con procesos biológicos específicos o estados de enfermedad.

2.2.5. Método de Aproximación de Redes

Este método implica la construcción y análisis de redes biológicas basadas en la expresión génica. Se utilizan algoritmos y técnicas de análisis de redes para identificar módulos o grupos de genes que están coexpresados y que pueden estar implicados en las mismas vías biológicas o procesos celulares.

La Aproximación por redes permite una visión global y sistémica de la expresión génica, lo que puede ayudar a identificar genes y procesos biológicos clave en un contexto más amplio.

2.3. Bases de Datos.

Para el desarrollo de este trabajo utilizamos datos de expresión de acceso público, derivados de bases de datos generadas por proyectos colaborativos orientados a la compilación de datos de expresión genética en múltiples especies, incluyendo a los humanos. Estas bases de datos o proyectos son una invaluable fuente de información para los investigadores interesados en estudiar la expresión génica en diferentes contextos biológicos y bajo diferentes condiciones experimentales. Proporcionan datos de expresión génica a nivel de transcriptoma, que es la colección de todos los ARN mensajeros (ARNm) presentes en una célula o tejido en un momento dado. Estos datos permiten a los investigadores obtener una visión detallada de qué genes están siendo activados o desactivados en un determinado contexto biológico, lo que a su vez puede ayudar a comprender los procesos biológicos y los mecanismos moleculares involucrados.

Algunos de los principales y más importantes son:

- *Gene Expression Omnibus (GEO)*: Mantenida por el Instituto Nacional de Investigación del Genoma Humano (NHGRI) de Estados Unidos. GEO es una base de datos que contiene una gran cantidad de datos de expresión génica en diversos organismos.

- *ArrayExpress*: Mantenida por el European Bioinformatics Institute (EBI), ArrayExpress es una base de datos que almacena una amplia gama de datos genómicos, incluyendo datos de expresión génica y datos relacionados a diversos organismos.

- *Gene Expression Atlas*: También mantenido por el EBI, Gene Expression Atlas es una base de datos que contiene datos de expresión génica de diferentes tejidos, órganos y condiciones biológicas en diversos organismos.

Para el presente trabajo, utilizamos dos bases de datos reconocidas y con una cuidadosamente curada compilación de valores de expresión en tejidos humanos: FANTOM5, y Brainspan.

2.3.1. FANTOM5

El proyecto FANTOM5 es un consorcio liderado por RIKEN [26], [27], que tiene como objetivo investigar los conjuntos de genes utilizados en prácticamente todos los tipos celulares del cuerpo humano, así como las regiones genómicas que determinan dónde se leen los genes. El proyecto se divide en dos fases:

La Fase 1 utilizó la técnica de Cap Analysis of Gene Expression (CAGE) para mapear los conjuntos de transcripciones, factores de transcripción, promotores y potenciadores activos en la mayoría de los tipos celulares primarios de mamíferos, células cancerosas y tejidos, incluido el ser humano.

En la Fase 2, los científicos del consorcio utilizaron un análisis exhaustivo de la expresión de ARN en diferentes tipos celulares para investigar cómo se producen los cambios fenotípicos en las células, y descubrieron que la activación inicial de los genes que determinan la diferenciación celular ocurre en regiones de ADN llamadas "potenciadores", que son interruptores regulatorios que se encuentran típicamente lejos de los genes que activan.

El objetivo final del proyecto es construir modelos regulatorios de transcripción para cada tipo celular primario del cuerpo humano, utilizando esta información.

El proyecto ha generado numerosas bases de datos para diversas especies, Humanos, Ratones, Ratas, Perros, Gallinas y Macacos. Estas bases de datos incluyen los datos de expresión génica de varios tejidos, así como los análisis CAGE y otros tipos de datos genómicos. Además, en algunos casos, los datos brutos (sin procesar) también están disponibles para su análisis posterior. [26], [27]

2.3.2. Brainspan

El proyecto Brainspan es un esfuerzo colaborativo para crear un mapa detallado de la expresión génica y la regulación del desarrollo cerebral humano desde la concepción hasta la edad adulta [28]. Se basa en la premisa de que la comprensión de los mecanismos moleculares subyacentes al desarrollo del cerebro humano es esencial para comprender las causas de los trastornos neurológicos y psiquiátricos.

El proyecto Brainspan inició en 2011 y es una colaboración entre múltiples instituciones y laboratorios de investigación. Los datos se generan a partir de muestras de tejido cerebral humano post-mortem y se analizan utilizando una variedad de técnicas de secuenciación y análisis bioinformático avanzado. El proyecto también incluye datos de otras especies animales, como el ratón y el chimpancé, para permitir la comparación con modelos animales.

Los datos del proyecto Brainspan están disponibles en una base de datos pública y accesible llamada Brainspan Atlas of the Developing Human Brain, que contiene información detallada sobre la expresión génica y la regulación en diferentes regiones del cerebro humano durante el desarrollo. [28]

3. Caso de Estudio

Con base en lo expuesto anteriormente, para este proyecto se requiere trabajar con tejidos que presenten diferentes niveles de actividad proliferativa. La función modelo definida será el Ciclo Celular, y los genes asociados a esta función serán identificados utilizando la anotación ontológica derivada del proyecto GeneOntology (GO [22], [23]).

Para la obtención de datos de expresión, se utilizarán las bases de datos de los proyectos FANTOM5 [26], [27] y Brainspan [28], que contienen datos de expresión para el ser humano en diferentes "condiciones" de actividad proliferativa. De la base de expresión Fantom5, se extrajeron datos de expresión para líneas celulares (alta actividad proliferativa), tejidos no nerviosos (actividad proliferativa media), y tejido nervioso [29] (actividad proliferativa baja). De la base de expresión de Brainspan, se extrajeron datos procedentes regiones cerebrales con distinto contenido relativo de células gliales, ya que estas son las únicas células con actividad proliferativa en el sistema nervioso. De este modo utilizamos el contenido relativo de glía como un indicador de actividad proliferativa general en el tejido de interés. De acuerdo con esto, extrajimos datos de expresión del cuerpo estriado (alto contenido relativo de Glía), corteza (contenido de Glía medio) y cerebelo (bajo contenido de Glía) [30], [31].

3.1. Procedimientos numéricos y gráficos

3.1.1. Obtención de Datos de Trabajo

Se realizó la obtención de datos de expresión génica, a partir de las dos bases de datos arriba mencionados. Específicamente se utilizaron los datos de expresión, de alrededor de 20 mil genes humanos, derivados de la base de datos FANTOM5, normalizados en tags por millón (TPM) y se trabajó únicamente con las muestras de genes disponibles de tejidos (nervioso y no nervioso) y Líneas Celulares. Por otra parte, en dichas bases se identificaron los genes pertenecientes a la función de Ciclo Celular (GO:0007049, Gene ontology project [22], [23]).

3.1.2. Análisis preliminar de datos

Como análisis preliminar del perfil general de expresión de las muestras de tejidos a analizar se llevó a cabo un examen de componentes principales para determinar si existe una diferencia global entre las muestras, al nivel de los genes del ciclo celular.

3.1.3. Pretratamiento de los Datos

Intentando evitar artefactos asociados al número de muestras de genes disponibles por grupo (tejido), presentes en las bases de datos, para cada grupo se utilizó un número de muestras igual al menor número de muestras presentes entre los tejidos analizados. En aquellos grupos de tejidos en los que existe un mayor número de muestras, estas se seleccionaron aleatoriamente.

3.1.4. Medición colectiva de los Datos de Expresión

Para determinar si los tejidos con mayor actividad proliferativa muestran un mayor nivel colectivo de expresión de los genes asociados al ciclo celular, a cada uno de estos genes se les determinó la mediana, a lo largo de las muestras contenidas en cada grupo (mediana de cada fila), y se calculó a continuación la mediana de estas medianas, obteniendo finalmente una mediana estos genes por cada grupo (Mediana de Expresión). Para determinar si la Mediana de Expresión de los genes asociados al ***CC*** es igual, mayor o menor de lo esperado al azar entre el resto de los genes, se tomaron 5000 muestras aleatorias de genes del mismo tamaño que el número de genes del ***CC*** presentes en cada grupo, y se les calculó la Mediana de Expresión de la misma forma, para así obtener la distribución de medianas esperadas al azar (fondo).

3.1.5. Eliminación de Genes con Varianza 0

En la determinación de los datos de coexpresión se usó el método “Spearman”, el cual es independiente de si la relación entre dos variables es o no lineal. Para ello, inicialmente se removieron del análisis aquellos genes con varianza igual a cero.

3.1.6. Determinación de los niveles de Coexpresión

Tomando los datos obtenidos previamente, se extrajeron los datos de expresión correspondientes a los genes del ***CC*** y se obtuvo la matriz de coexpresión entre todos los pares posibles de éstos, para después, a estas matrices calcularles las medianas a lo largo de los genes (mediana de las filas) y finalmente la mediana de estas medianas para obtener así la Mediana de Coexpresión de los genes del ***CC***. Con objeto de determinar en qué medida los valores de coexpresión de los genes del Ciclo Celular difieren de lo esperado al azar, se tomaron 5000 muestras de genes aleatorios en cada grupo de tejidos, siendo el tamaño de muestra el mismo que el número de genes del ***CC*** presentes en cada grupo (tejidos). A estas muestras aleatorias se les calculó la Mediana de Coexpresión de la misma forma, obteniendo la matriz de coexpresión de

estas muestras, a partir de ellas se obtuvieron las medianas de las medianas de coexpresión a lo largo de los genes. Obteniendo la correspondiente distribución de estos 5000 valores esperados. Estas distribuciones de valores medios de coexpresión esperados se utilizaron para determinar la significancia estadística de la coexpresión colectiva de los genes del ciclo celular en cada uno de los tejidos analizados.

3.1.7. Reportar los datos gráficamente

Con los datos obtenidos en ambos sets se realizó 1 gráfica por cada uno de los análisis anteriormente descritos, cada una de ellas mostrando la distribución de medianas (fondo), así como la mediana de los genes del ciclo celular en cada grupo de tejidos.

3.1.8. Análisis por método Wilcoxon

De manera complementaria, se llevaron a cabo comparaciones pareadas de Wilcoxon en los niveles de expresión y coexpresión de los genes del ciclo celular, comparando todos los pares posibles de grupos de tejidos con alta, media y baja actividad proliferativa.

3.1.9. Determinación del logFC

En conjunción con las pruebas de Wilcoxon, se calculó el Log Fold Change (logFC), de los genes identificados como relacionados a la función del ciclo celular, comparando pares de tejidos con menor o mayor actividad proliferativa, para determinar la tendencia colectiva de los genes del ciclo celular hacia un aumento o disminución en el nivel de expresión o coexpresión. Estos datos se plasmaron en gráficas de abanico en las que se midió el incremento o decremento en expresión o coexpresión de cada uno de los genes individuales involucrados en la función del ciclo celular.

3.1.10. Generación de la Red Génica

Con objeto profundizar en la caracterización de los cambios en la dinámica de coexpresión de los genes asociados al ciclo celular, al comparar tejidos de mayor y menor actividad proliferativa, se llevó a cabo un análisis de estos genes utilizando una aproximación de teoría de redes [7]. Se tomó la matriz de coexpresión de los genes asociados al ciclo celular en cada tejido examinado y se utilizó como matrices de adyacencia para la generación de redes de interacción entre ellos. Con las redes obtenidas se determinaron cambios en algunos de los parámetros, propiedades asociadas a la red de coexpresión, tales como nivel de agrupamiento (coeficiente de

clustering), número de interacciones de coexpresión (grado medio de la red de coexpresión), distancias topológicas, etc.

Como resultado de este proyecto se quiere determinar si la función de un determinado conjunto de genes se encuentra preferencialmente asociada, ya sea a la coordinación en su expresión o bien a su nivel colectivo de expresión.

3.2. Resultados

De acuerdo con los procedimientos enumerados en las secciones anteriores, se realizó la obtención de datos de expresión, públicamente disponibles de alrededor de 20 mil genes humanos derivados de la Base de Datos "FANTOM5", normalizados en tags por millón (TPM) y paralelamente con los datos obtenidos de la Base de datos de "Brainspan", normalizados en lecturas por kilobase por millón de cuentas (RPKM). En donde estos datos se encuentran organizados en tablas con los genes en las filas y el número de muestras en las columnas.

Para la base de datos de FANTOM5, únicamente usamos las muestras disponibles de tejidos nervioso y no nervioso, así como de líneas celulares. Particularmente, dentro del tejido nervioso, comparamos entre las muestras disponibles en la base de datos de Brainspan, de cerebelo, corteza y cuerpo estriado. Se agruparon estos datos de expresión de acuerdo con el tejido que pertenecen, formando así 6 grandes grupos de muestras, en la base de datos de FANTOM5; cerebro (tejido nervioso); tejidos (tejido no nervioso) y líneas celulares, y en la base de datos de Brainspan; cerebelo, corteza y cuerpo estriado.

3.2.1. Análisis de Componentes Principales (*PCA*)

Como se mencionó en los procedimientos, previo a entrar en los análisis de expresión y coexpresión, surgió la duda de si los datos de expresión de los genes del Ciclo Celular se diferenciaban de forma pronunciada o no, al comparar entre los tres grupos de tejidos con alta, media y baja actividad proliferativa, por lo que para determinar esto se practicó un análisis del perfil general de expresión de los genes asociados al ciclo celular.

En cada uno de los tres grupos de tejidos, se identificó a los genes asociados a la función del ciclo celular (***CC***, GO:0007049), y se extrajeron únicamente los datos de expresión de éstos, para llevar a cabo el correspondiente análisis de componentes principales. Este análisis es un método

de reducción de dimensiones que nos permite representar en sólo dos dimensiones el modo en el que se agrupan las muestras individuales como función del perfil de expresión de cada una de ellas (en este caso sólo tomando en cuenta la expresión de los genes asociados al ciclo celular. Para este análisis se utilizaron los valores de expresión directamente obtenidos de la base sin ningún preprocesamiento o renormalización.

En la Figura 4, se puede observar una muy ligera diferencia entre el nivel de expresión de los genes identificados como asociados al Ciclo Celular entre los diferentes grupos, sin embargo, éstos se sobreponen unos a otros, mostrando una muy pobre separación entre los tejidos, o las muestras de alta, media y baja actividad proliferativa. Este resultado sugiere que el nivel colectivo de expresión de genes asociados al ciclo celular no necesariamente discrimina entre poblaciones celulares con diferentes niveles de actividad proliferativa.

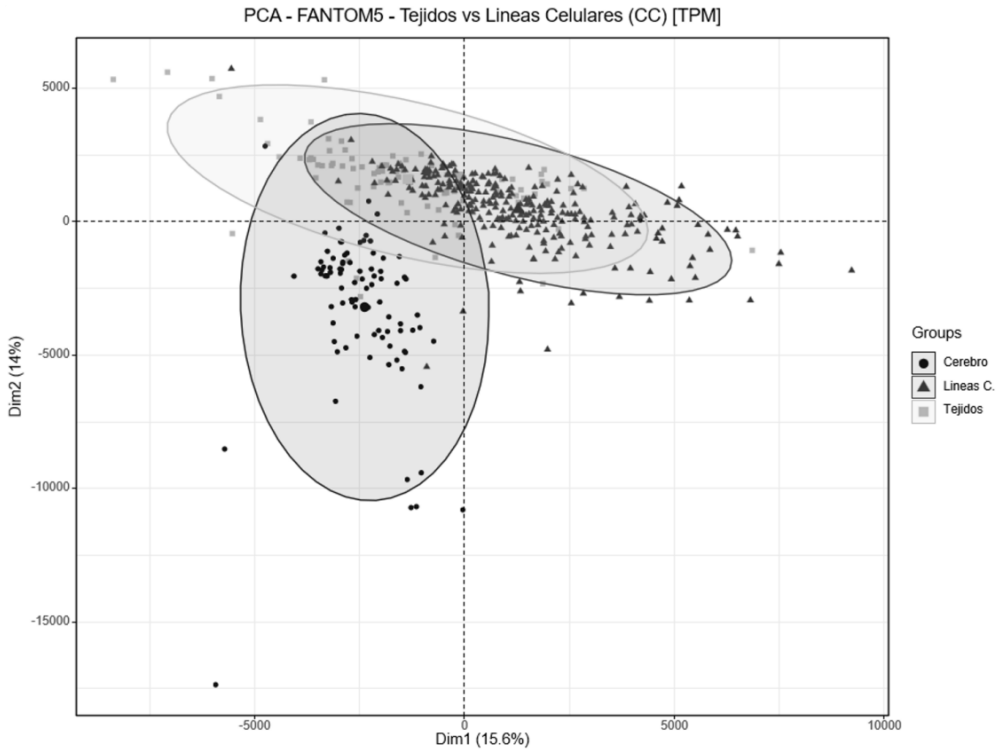


FIGURA 4. Análisis de Componentes Principales (PCA) Aplicado a los genes del *CC* identificados en los grupos de tejidos de alta, media y baja actividad proliferativa. Utilizando la base de datos de FANTOM5,se identificaron los genes relacionados a la función del ***CC*** (GO:0007049) en los tres tipos de tejidos indicados en la gráfica (C, T, L), y se aplicó el PCA. En la figura se observa una pobre separación debida a los niveles de expresión de estos genes entre los distintos grupos de tejidos)

3.2.2. Pretratamiento de Datos

Dado que en ambas bases de datos consultadas (FANTOM5 y Brainspan), existe una diferencia en cuanto al número de muestras de genes disponibles entre grupos de tejidos y las mediciones de correlación, que proyectamos hacer, son sensibles al tamaño de muestra, se optó por utilizar el mismo número de muestras por grupo. Tomando como punto de referencia el grupo de datos con el menor número de muestras, en su base de datos correspondiente, y a partir de éste, seleccionar de manera aleatoria, muestras de genes en los grupos restantes para igualar su tamaño.

En todo este estudio, los métodos descritos anteriormente (exceptuando el análisis de componentes principales), y los resultados que se muestran a continuación, se trabajó con un total de 89 muestras y un máximo de 23,274 genes en cada uno de los grupos de datos de líneas celulares, tejido nervioso (cerebro), y tejidos no nerviosos (FANTOM5), así como 23 Muestras y un máximo de 22,327 genes para cada uno de los grupos de datos de cuerpo estriado, corteza y cerebelo (Brainspan), correspondiendo a tejidos de alta, media y baja actividad proliferativa respectivamente. Así mismo, dentro del desarrollo del trabajo se fueron realizando diferentes agrupaciones para trabajar con estos datos, en la figura 5 se muestra cómo se organizaron algunos de estos grupos.

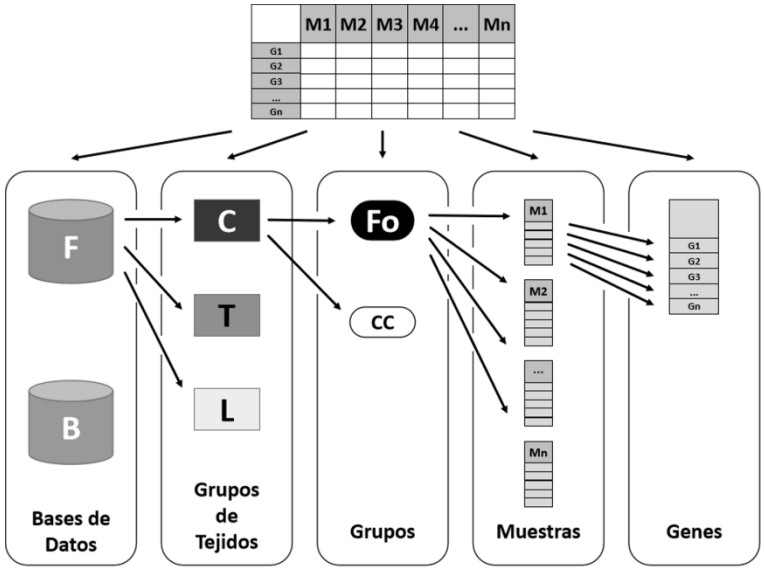


Figura 5. Organización de los datos de expresión durante el estudio. Partiendo de dos bases de datos independientes (FANTOM5 y Brainspan, F y B respectivamente), organizados como se indica en la matriz superior (los genes, Gi, se muestran en las filas y las muestras, Mi, en las columnas), se identificaron las muestras correspondientes a tejidos de alta, media y baja actividad proliferativa (Cerebro, Tejido no nervioso, o líneas celulares en el ejemplo de la figura, C, T y L respectivamente). Dentro de esas muestras se extrajeron los datos para los genes asociados a ciclo celular (***CC***) o subgrupos de genes del mismo tamaño, extraídos del fondo (Fo). De este modo cualquier grupo de tejidos analizados (alta, media o baja actividad proliferativa) consiste en un grupo de muestras, cada una de las cuales está integrada por una lista de genes

3.2.3. Análisis de Expresión

Teniendo los datos tratados (contando con el mismo número de muestras por grupo) y clasificados de acuerdo a su requerimiento de la función proliferativa (en grupos de tejidos con alta, media o baja actividad proliferativa), queríamos conocer de qué forma se diferencian los genes asociados a esta función entre los diferentes grupos de tejidos e Con este objetivo, se identificaron los genes asociados a la función del ***CC***, seguido de ello y para determinar el comportamiento de sus valores colectivos de expresión, de acuerdo al procedimiento descrito en la sección anterior (ver sección **3.1.4**), se les determinó la mediana de expresión, e identificamos como CC_C a la mediana expresión de los genes del Ciclo Celular en Cerebro, CC_T y CC_L a las medianas de expresión de estos mismos genes para tejido no nervioso y líneas celulares respectivamente en la Base de datos de FANTOM5. Similarmente, para el análisis de regiones cerebrales, se identificó como CC_{CB}, CC_{CX}, y CC_{ST} a las medianas de expresión de los genes del CC en los grupos de Tejidos de Cerebelo, Corteza y Estriado respectivamente.

Así mismo y con objeto de determinar en qué medida o con qué intensidad estos valores colectivos de expresión son mayores o menores a la expresión de cualquier grupo arbitrario del mismo número de genes en el tejido (fondo), así como su significancia estadística, se tomaron 5000 muestras aleatorias de genes de manera que el tamaño de cada muestra aleatoria fuese igual al del número de genes asociados a ***CC*** encontrados en cada grupo de tejidos. A cada uno de estos muestreos se le calculó la mediana de expresión de la misma forma que a los genes del ***CC***, obteniendo así la distribución de 5000 medianas de expresión colectiva esperada al azar (fondo) en cada uno de los grupos de tejidos.

Sin embargo, se observó que las distribuciones de los valores colectivos esperados, tanto en los genes del ***CC***, así como los valores del Fondo, se encuentran desplazadas unas respecto a la otras al comparar los resultados entre los grupos de tejidos o entre distintas bases de datos. Por esta razón se optó por corregir los valores de expresión en ambas bases de datos para poder transformarlos en valores comparables. Esto se llevó a cabo dividiendo las medianas de expresiones calculadas en cada grupo, tanto las medianas aleatorias (Fondo) como las medianas de los genes del ***CC***, por la mediana global de expresión del tejido respectivo.

Estas medianas globales de expresión se calcularon de la misma forma que las medianas de expresión del Fondo y las de los genes del ***CC***, salvo que en lugar de trabajar con muestras dentro de los genes disponibles o solo con cierto número de estos, se trabajó con los datos de expresión de todos los genes disponibles en cada grupo. Así, en cada grupo de tejidos, se tomó la mediana de expresión de cada gen a lo largo de todas las muestras disponibles (mediana de cada fila), seguido de esto se calculó la mediana de estas medianas. Haciendo este procedimiento se obtuvieron 6 medianas globales de expresión, una por cada grupo estudiado.

Para realizar la corrección del desplazamiento de las medianas de los valores colectivos de expresión del Fondo, así como las de los genes del ***CC***, en cada tejido, se dividieron las medianas de ambos resultados entre sus Medianas Globales Respectivas. Esta transformación permite obtener distribuciones comparables de medianas entre los grupos con diferentes niveles de actividad proliferativa en su respectiva base de datos.

Como se observa en la Figura 6, en donde se graficaron los valores corregidos de medianas de expresión colectiva, tanto los valores de Fondo así como los de los genes del ***CC***, en la base de FANTOM5 (que contiene a los grupos; Cerebro, Tejidos, Líneas Celulares) las medianas de expresión de los genes del Ciclo Celular se encuentran muy por encima de la distribución de valores esperados (Fondo), en donde también podemos apreciar que los valores de estas medianas están ordenados de acuerdo al nivel creciente de actividad proliferativa, es decir Cerebro (CC_C), < Tejidos (CC_T) < Líneas Celulares (CC_L), donde el valor de la mediana de este último, se encuentra incluso aún más alejado que el de los dos primeros. Sin embargo este ordenamiento no ocurre al interior de las distintas regiones cerebrales, es decir con los valores de expresión obtenidos de la base de Brainspan (Agrupados en; Cerebelo, Corteza, Estriado) en donde de inicio se puede destacar que, a pesar de que las medianas de expresión de los genes del ***CC*** si se encuentran por encima de la distribución esperada (Fondo), el nivel colectivo de expresión por grupo no se ordena de acuerdo al nivel creciente de actividad proliferativa, dado por el contenido de Glía en la región cerebral, en donde: Cerebelo < Corteza < Estriado, en contraste con esto, el nivel colectivo de expresión de genes del ***CC*** en el cerebro muestra un ordenamiento distinto: Estriado(CC_{ST}) < Cerebelo (CC_{CB}) < Corteza (CC_{CX})

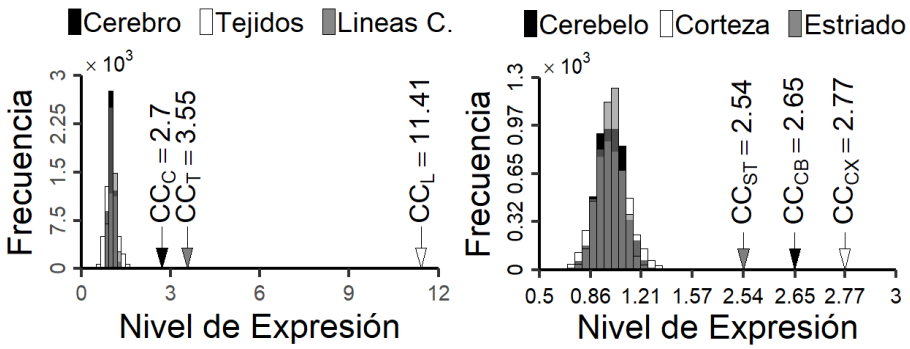


FIGURA 6. Nivel colectivo de Expresión de genes asociados a CC en tejidos de alta, media y baja actividad proliferativa. Se identificaron los genes relacionados a la función del CC (GO:0007049) en en grupos de tejidos con diferencias en niveles de actividad proliferativa (cerebro, C, tejidos no nerviosos, T, líneas celulares, L,, por un lado y núcleo estriado, ST, cerebelo ,CB y corteza, CX), y se determinó, para cada uno, la mediana, de expresión (Flechas). Se tomaron 5000 muestras de genes aleatorios (Con una n del mismo tamaño que los genes del CC presentes en su respectivo grupo), para obtener la distribución esperada al azar de medianas (Histogramas). Se puede observar que en todos los casos la mediana de expresión colectiva de los genes asociados al ciclo celulares es significativamente mayor a la esperada en genes de fondo. Sin embargo, la expresión colectiva no necesariamente se alinea con el nivel de demanda de actividad proliferativa.

Este resultado demuestra que *el nivel colectivo de expresión* de los genes asociados al ***CC*** *no necesariamente se corresponde con el nivel de demanda de actividad proliferativa* en poblaciones de células.

3.2.4. Análisis de Coexpresión

A diferencia del análisis realizado para nivel de colectivo de expresión de un cierto grupo de genes, lo que se intentó cuantificar en este análisis fue el nivel de coordinación en la expresión de dicho grupo de genes. Para este análisis y con ese objetivo en mente, se obtuvo la correlación entre todos los pares posibles, de los genes asociados al ciclo celular. Específicamente utilizamos correlaciones de "Spearman", teniendo esto en cuenta y dado que las correlaciones no pueden definirse para series de datos con varianza cero, se realizó un tratamiento a los datos de cada grupo en sus respectivas bases de datos, en donde se removieron los genes con varianza 0.

A continuación, se identificaron nuevamente los genes que están relacionados a la función del ***CC*** y que permanecieron en las muestras de cada grupo de tejidos después de haber realizado la remoción de los genes con varianza 0. Se obtuvieron entonces, correlacionando estos genes con

el método ya mencionado, las 6 matrices de correlación, que contienen las correlaciones de todos los pares posibles de los genes asociados al ***CC*** en su respectivo grupo.

De las matrices de correlación, se obtuvo para cada una de ellas, una mediana global la cual se determinó para cada tejido, calculando la mediana de las filas de la matriz de correlación (es decir, la mediana de cada gen) y por último la mediana de estas medianas, obteniendo así un solo valor para cada grupo (ver procedimiento en sección **3.1.6**).

Como en el análisis anterior, se calculó qué tanto estos valores globales de coexpresión por tejido, difieren de lo esperado al azar en grupos aleatorios de genes del fondo. Para ello, se tomaron 5000 muestras de genes elegidos de forma aleatoria, haciendo coincidir el tamaño del muestreo con el número de genes asociados a ***CC***, encontrados en su respectivo grupo, se obtuvo la matriz de coexpresión para cada una de estas muestras por el método de Spearman , obteniendo las matrices de correlación a las cuales se les calculó la mediana de coexpresión de la misma forma, calculando la mediana de las filas de la matriz y sucesivamente la mediana de estas medianas, obteniendo así la distribución de los valores de coexpresión esperada al azar (fondo).

En este caso también fue necesario aplicar una corrección similar a la efectuada con los resultados del análisis de expresión de la sección anterior, ya que las distribuciones de genes de fondo se encuentran desplazadas entre los diferentes grupos de tejidos, en sus respectivas Bases de Datos. Por lo tanto dividimos estos valores, la distribución mediana aleatoria así como las Medianas de Coexpresión de los genes del ***CC*** entre las medianas globales de coexpresión correspondientes.

Estas Medianas Globales de Coexpresión se calcularon con el mismo método que las medianas globales de expresión, esto es, se obtuvieron las matrices de coexpresión por el método de Spearman para absolutamente todos los genes presentes, y de éstas se obtuvo la mediana de cada una de sus filas (mediana de cada gen) y seguido de esto, se calculó la mediana de estas medianas. Realizando el procedimiento a cada grupo se obtuvieron 6 valores de medianas globales de coexpresión, una por cada uno de los tejidos.

Teniendo los valores de las medianas globales de coexpresión, para cada grupo de tejidos, se realizó la corrección dividiendo tanto las medianas de genes aleatorios como las de los genes

asociados al ciclo celular entre sus Medianas Globales Respectivas. Esta corrección hace que las distribuciones del fondo se vuelvan comparables entre los distintos grupos de tejidos con alta, media y baja actividad proliferativa.

Como se puede apreciar en la Figura 7, tanto en la base de FANTOM5 (Grupos de tejidos; Cerebro, Tejidos, Líneas Celulares) como en la de Brainspan (Grupos de tejidos; Cerebelo, Corteza, Estriado), los genes asociados al ciclo celular, muestran que su coexpresión colectiva es mucho mayor a la distribución esperada al azar. Dicho de otro modo, en ambas gráficas podemos identificar claramente los genes del ***CC*** en los diferentes tejidos; cerebro (CC_C), tejidos (CC_T) y líneas celulares (CC_L), para la base de datos de FANTOM5 y cerebelo (CC_{CB}), corteza (CC_{CX}) y cuerpo estriado (CC_{ST}), así como también las distribuciones aleatorias (Fondo). En contraste con el análisis de expresión, en este análisis podemos notar que los valores de coexpresión para los genes del ***CC*** se encuentran ordenados de acuerdo al nivel creciente de actividad proliferativa de cada grupo de tejidos, es decir la coexpresión colectiva observada fue; Cerebro (CC_C) < Tejidos (CC_T) < Líneas Celulares (CC_L), lo mismo ocurre para las distintas subregiones cerebrales: Cerebelo (CC_{CB}) < Corteza (CC_{CX}) < Estriado (CC_{ST}). Por lo que este resultado muestra que, en los grupos de tejidos examinados, el nivel de coexpresión de los genes asociados al ***CC*** se corresponde con el nivel de demanda de actividad proliferativa en poblaciones de células.

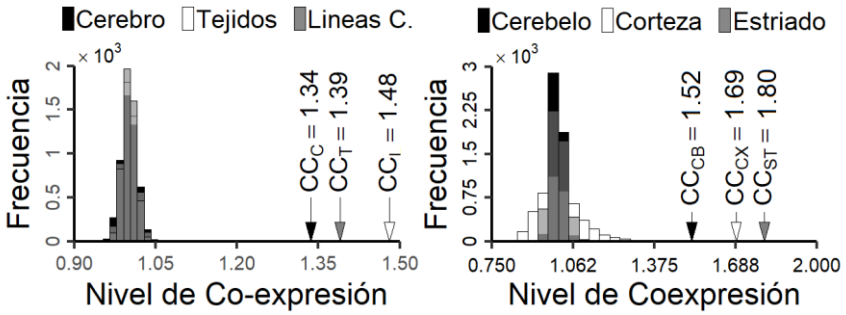


FIGURA 7. Nivel Coexpresión de genes asociados a *CC* en tejidos de alta, media y baja actividad proliferativa. Se identificaron los genes relacionados a la función del CC (GO:0007049) en grupos de tejidos con diferencias en niveles de actividad proliferativa (cerebro, C, tejidos no nerviosos, T, líneas celulares, L,, por un lado y núcleo estriado, ST, cerebelo ,CB y corteza, CX), correlacionándolos y determinando, para cada uno, la mediana, de coexpresión (normalizada respecto a la coexpresión del transcriptoma completo, Flechas). Además, se tomaron 5000 muestras de genes aleatorios (Con una n del mismo tamaño que los genes del CC presentes en su respectivo grupo), para obtener la distribución esperada al azar de medianas (Histogramas). Podemos ver que en los 6 grupos la mediana de coexpresión de los genes asociados al ciclco celular es maor de lo que se espera al azar en los genes de fondo, además correspondiendo con el nivel de demanda de actividad proliferativa.

3.2.5. Análisis pareado de los genes del *CC* en los diferentes tejidos

Dado que el análisis anterior es un análisis general del comportamiento colectivo de los genes del ***CC*** , aún quedaba la incógnita de cómo es que estos genes, los identificados como asociados a la función del Ciclo Celular, se comportan de forma individual en los diferentes tejidos con niveles alto, medio y bajo de actividad proliferativa. Por esta razón se practicó un análisis a nivel individual, es decir un análisis del cambio de cada gen individual entre tejidos con mayor y menor actividad proliferativa.

Para llevar a cabo este análisis se calculó, para los genes en cuestión, el logaritmo de la relación de cambio (log fold change o logFC por sus siglas en inglés), que es el logaritmo base 2, de la relación de expresión mediana de expresión cada gen individual en dos grupos de tejidos (o o mediana de coexpresión de cada gen individual en dos grupos de tejidos). De este modo se comparó la expresión mediana de cada uno de los genes del ***CC***, entre pares de condición de actividad proliferativa (Tejidos), baja vs. alta, media vs. alta y baja vs. media, comparando estas condiciones en cada base de datos por separado.

Tratándose de un análisis pareado, se requiere un valor un valor de expresión o coexpresión por gene (por tejido o subregión analizada). Y para ello se tomó, para el caso del análisis pareado de expresión, la mediana de expresión de cada gen, mientras que para la coexpresión se tomó la mediana de cada fila de la matriz correspondiente de coexpresión (mediana de coexpresión por gen, ver secciones **3.1.4** y **3.1.6**). Para evitar divisiones por cero (resultado de la existencia de genes con una mediana de expresión igual a cero), se realizó una corrección de los datos consistente en sumar a los todos los valores de expresión/coexpresión en cada muestra, el inverso del valor máximo de la relación de cambio (fold change).

Las comparaciones de los valores de expresión y coexpresión para cada uno de los genes asociados al ciclo celular se representan en forma de abanico en las figuras 8 y 9 respectivamente, en donde cada línea representa el incremento o decremento de estos valores, (medidos en LogFC) de cada gen, en un tejido de alta actividad proliferativa en relación a su expresión, o coexpresión, en un tejido de baja actividad proliferativa.

3.2.5.1. Análisis pareado para los valores de expresión

En la figura 8 se muestran los cambios de expresión entre los diferentes pares de tejidos de la Base de datos de FANTOM5, Cerebro vs. Líneas, Tejidos vs. Líneas y Tejidos vs. Cerebro, así

como para la Base de Brainspan, Cerebelo vs. Estriado, Corteza vs. Estriado y Corteza vs. Cerebelo, en estas figuras el tejido de baja actividad proliferativa se representa en el vértice de las gráficas de abanico (siempre de lado izquierdo), de modo que el incremento o decremento en la expresión de cada gen (líneas individuales) aparece en referencia a dicho tejido. Como se puede observar, la mayoría de los genes asociados al ciclo celular, incrementan su expresión en el tejido de mayor actividad proliferativa. La única excepción se observa en la comparación entre corteza y estriado (panel derecho) en el que hay un mayor número de genes que muestran un decremento en la expresión en la condición de mayor actividad proliferativa.

Con el fin de determinar la significancia estadística en estos sesgos de incremento o decremento de expresión para todos los genes del ciclo celular, llevamos a cabo una prueba pareada de Wilcoxon La probabilidad asociada se indica en la parte superior de cada comparación.

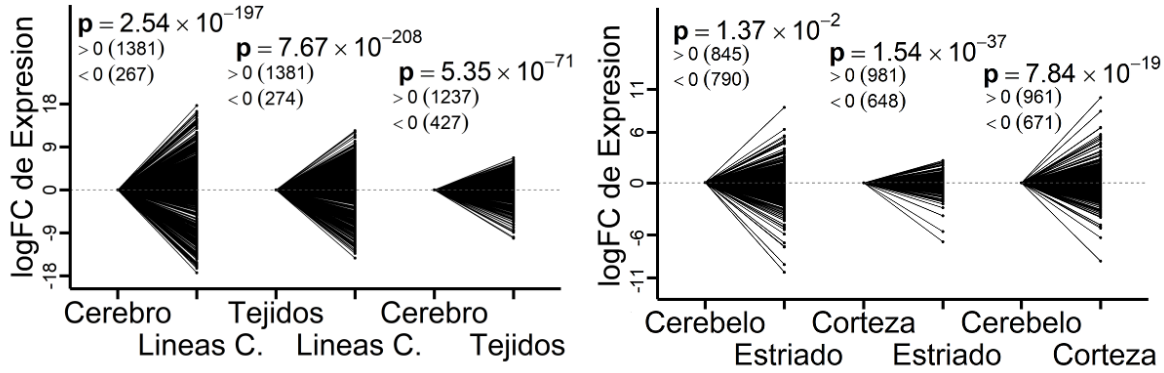


FIGURA 8. Cambios en el Nivel de Expresión individual de genes asociados al CC en los diferentes Tejidos a Analizar. Se realizó la prueba de Wilcoxon para comprobar que hay una diferencia significativa entre los genes del ***CC***, entre cada par de tejidos comparados y en su respectiva Base de datos. El valor de probabilidad se muestra en la parte superior. Se calculó el logaritmo de la proporción de cambio del tejido con mayor actividad proliferativa, (de lado derecho del abanico) comparado con el de menor actividad proliferativa (de lado izquierdo del abanico, ver sección de resultados). Cada línea representa el logaritmo del cambio en proporción de expresión de cada uno de los genes del ***CC***.

3.2.5.2. Análisis pareado para los valores de coexpresión

Para el análisis pareado de los valores de coexpresión se usó el mismo método que para los valores de expresión, como ya se dijo los valores de coexpresión de cada gen se obtuvieron resumiendo las matrices de coexpresión de los tejidos de acuerdo con lo descrito en la sección **3.1.6**. En la figura 9 se encuentran graficadas las comparaciones entre las medianas de coexpresión

de los genes asociados al ciclo celular en los tejidos con diferentes niveles de actividad proliferativa. En primer término, se compararon , Cerebro vs. Líneas, Tejidos vs. Líneas y Tejidos vs. Cerebro (panel izquierdo, figura 9) En segundo término se compararon las subregiones Cerebelo vs. Estriado, Corteza vs. Estriado y Corteza vs. Cerebelo. Como se observa en la figura, en todos los casos la mayoría de los genes, asociados al CC incrementan su nivel de coexpresión en el tejido con mayor actividad proliferativa y este sesgo es en todos los casos estadísticamente significativo de acuerdo con la prueba de Wilcoxon.

Tomados en conjunto, estos resultados demuestran que el nivel de coexpresión y no así el nivel colectivo de expresión, de los genes asociados al ciclo celular, se asocia robustamente a las variaciones en actividad proliferativa en distintas poblaciones celulares o tejidos.

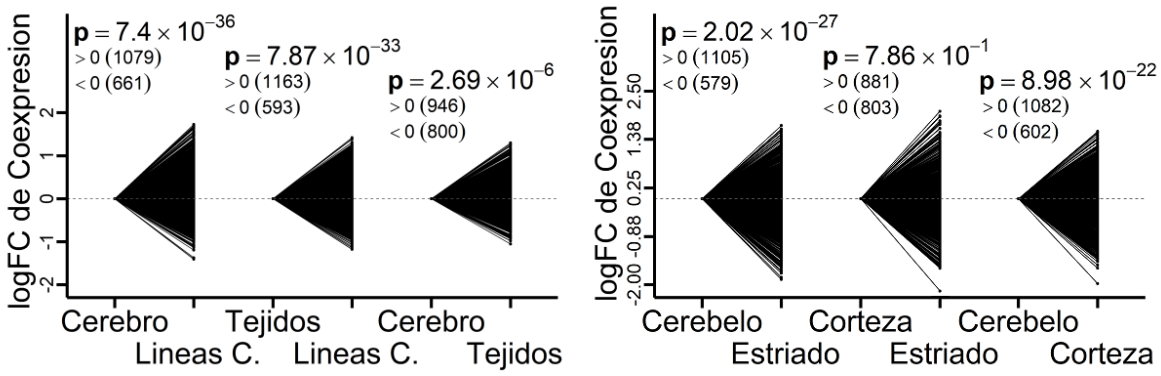


FIGURA 9. Cambios en el Nivel de Coexpresión individual en genes asociados al CC en los diferentes Tejidos a Analizar. Se realizó la prueba de Wilcoxon para comprobar que hay una diferencia significativa entre los genes del ***CC***, entre cada par de tejidos comparados y en su respectiva Base. El resultado del p-value se muestra en la parte superior. Se calculó el logaritmo de la proporción de cambio del tejido con mayor actividad proliferativa, (de lado derecho del abanico) comparado con el de mayor actividad (de lado izquierdo del abanico). Cada línea representa el cambio en proporción de coexpresión de cada uno de los genes del ***CC***.

3.2.6 Análisis de Redes de Coexpresión.

Habiendo establecido que el nivel colectivo de coexpresión entre los genes asociados al ciclo celular refleja de modo más robusto el nivel de actividad de esta función, decidimos a continuación utilizar una aproximación de análisis de redes para caracterizar mejor los cambios en las dinámicas de coexpresión de los genes asociados al ciclo celular en los distintos tejidos estudiados. La cual

consistió en generar las redes, diagramas en donde mediante nodos y aristas, se representan los diferentes niveles de interacción (aristas) entre los genes analizados (nodos), de modo que es más práctico visualizar las interacciones deseadas.

En la aproximación de redes, se representa a los elementos que forman parte de ellas (los genes en este caso) como nodos y se representa a las interacciones (relaciones de coexpresión) como aristas. Esta representación nos permite examinar y cuantificar las características topológicas de esta red de interacciones, así como los cambios en estas estructuras al comparar condiciones distintas.

3.3.6.1 Determinación del Umbral

Dado que las redes de coexpresión se generan a partir de las matrices de correlación, y que estas se componen de alrededor de 1,444,150 (el número total de pares posibles de genes asociados al CC), se necesita seleccionar un umbral en el valor numérico de las correlaciones a partir del cual se consideran los pares de genes que formarán parte de la construcción de la red.

En el presente estudio, se eligió como umbral la mediana de coexpresión del tejido con mayor actividad proliferativa. Específicamente tomamos el valor de coexpresión correspondiente a la mediana de líneas celulares e identificamos todos los pares de genes cuyas coexpresiones fueran superiores a este umbral tanto en la matriz de líneas celulares como de tejidos no nerviosos y cerebro, resultando en tres listas de pares de genes. Con respecto a las subregiones cerebrales, tomamos la mediana de coexpresión del estriado e identificamos todos los pares de genes cuyas coexpresiones fueran superiores a este umbral tanto en la matriz del estriado como en la de corteza y cerebelo, resultando también en tres listas de pares de genes.

3.3.6.2 Generación de Redes Principales

Las listas resultantes de pares de genes descritas en la sección precedente superaban las 500,000 interacciones, por lo que decidimos tomar una muestra aleatoria del 1 % de las interacciones derivadas del tejido con actividad proliferativa más alta (líneas celulares en un caso o estriado en el otro) e identificamos cuáles de estas interacciones se encontraban presentes en los otros dos tejidos (tejidos no nerviosos y cerebro en un caso, o corteza y cerebelo en el otro). Las listas de pares así obtenidas se importaron al programa de Cytoscape, mediante la librería de R; RCy3, con la cual, el programa Cytoscape, género 1 red por lista de pares, para cada tejido de alta,

media y baja actividad proliferativa respectivamente. En las figuras 10 y 11 se muestra el comportamiento de cien interacciones aleatorias derivadas de la red de líneas celulares y estriado respectivamente, y el comportamiento de estas mismas interacciones en los tejidos de media y baja actividad proliferativa respectivos. Como puede verse, existe un aumento gradual en el número de interacciones desde el tejido de más baja actividad proliferativa hacia el de más alta, indicando que las interacciones de coexpresión entre genes asociados al ciclo celular aumentan conforme aumenta la actividad proliferativa de una población celular.

3.2.7 Análisis de Red

Cómo ya se mencionó, para los resultados mostrados en la figura 10 y 11, se tomó el 1% de todos los pares de genes disponibles en la matriz de correlación, este procedimiento se repitió 50 veces (cincuenta muestras aleatorias independientes) y se observó un resultado idéntico y consistente. Con objeto de llevar a cabo un análisis global más riguroso, obtuvimos cinco métricas que describen distintos aspectos de las redes globales de coexpresión aquí examinadas (1% aleatorio de las redes globales de coexpresión originales). Estas métricas son

- ***Numero de Nodos***: El cual indica el número de genes que están interactuando en la red, siendo el máximo el número de genes identificados como relacionados al **CC** (~1700)
- ***Interacciones***: Este parámetro indica el número total de interacciones presentes en la red, es decir, cuántas

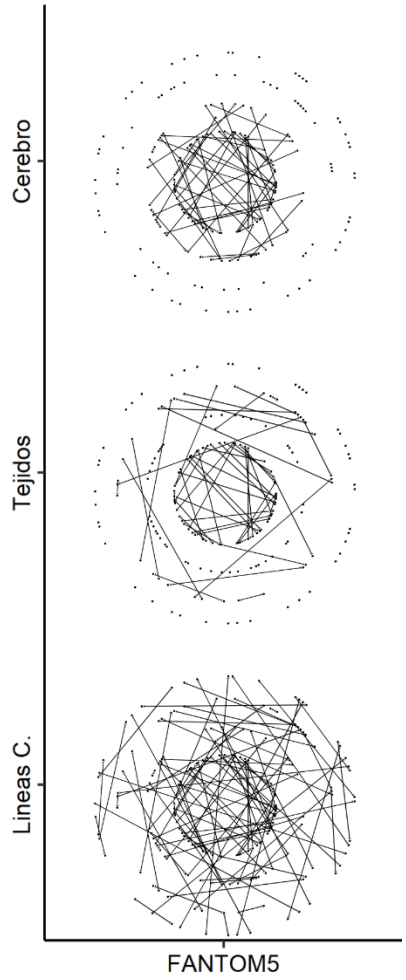


FIGURA 10. Muestra representativa de una red de cien interacciones en la Red de Líneas Celulares, identificando las interacciones correspondientes en Cerebro y Tejidos no nerviosos. Zoom hecho con 100 interacciones aleatorias de la Red generada cortando las Lista de pares, de la matriz de coexpresión, según la mediana de expresión del tejido de mayor actividad (Líneas Celulares), y seleccionando el 1% de estas interacciones al azar, las cuales se intersectarón en los otros dos tejidos

interacciones entre los diversos genes existen en el tejido representado, nuevamente siendo el máximo en el 1% de las interacciones totales.

- ***Vecinos promedio***: El parámetro nos indica la media de interacciones directas que un gen establece con otros. Este parámetro nos puede indicar que tan densa (cuantas interacciones están presentes respecto al número de genes) es una red.

- ***Distancia topológica media***: La distancia topológica entre dos nodos en una red, es el número mínimo de interacciones que separa a estos nodos. La distancia media, es el valor medio de estas distancias para todos los pares posibles de nodos en la red. cortas" de la red, las cuales se dividen entre el número total de estas, obteniendo así la distancia topológica media, la cual indica de forma inversa el tamaño de la red ya que a más interacciones y genes la distancia entre estos se vuelve más pequeña respecto al número de interacciones por lo que el valor promedio va en decremento, con valores más grandes en tejidos con baja actividad proliferativa, y los valores más pequeños en tejidos de actividad alta.

- ***Coeficiente de clustering***: Indica el promedio de los coeficientes de agrupamiento (clustering) de los nodos de la red. El coeficiente de clustering para un nodo se define como la proporción existente realmente, del total de interacciones posibles entre los vecinos inmediatos de un nodo. Este es un número que varía entre 0 y 1 y representa el grado de cohesión o agrupamiento interno entre los vecinos de un determinado nodo en una red. Un aumento en este valor representaría un aumento en el grado de coordinación interna entre los nodos de la red

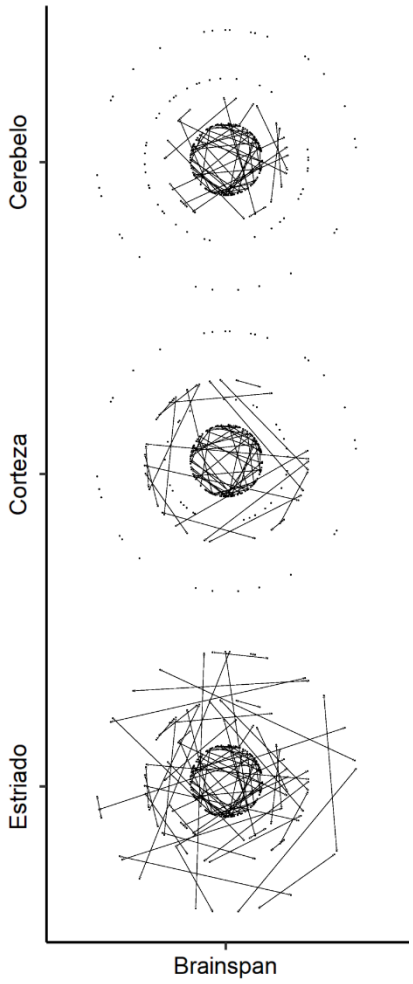


FIGURA 11. Muestra representativa de una red de cien interacciones en la Red de núcleo estriado identificadas en la red de corteza y cerebelo. Zoom hecho con 100 interacciones aleatorias de la Red generada cortando las Lista de pares, de la matriz de coexpresión, según la mediana de expresión del tejido de mayor actividad (Cuerpo Estriado), y seleccionando el 1% de estas interacciones al azar, las cuales se intersectarón en las otras dos subregiones

- ***Densidad***: Es una medida del número total de interacciones en la red respecto al máximo posible de estas y varia también entre 0 y 1. Cuando mas alto es este valor, más profusa son las interacciones entre los nodos de una red.

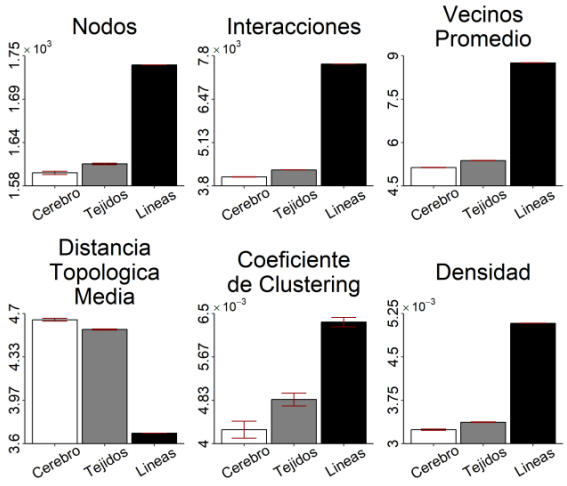


FIGURA 12. Análisis de Redes en los tejidos de la Base de Datos de FANTOM5 (cerebro, tejidos no nerviosos y líneas celulares). Generado con los datos de alrededor de 50 muestras independientes del 1% de la red global de coexpresión (ver resultados) . A cada una de las iteraciones del proceso de creación de redes se les calculó los parámetros mostrados (y error estándar basado en las cincuenta muestras independientes).

Como se aprecia en sus respectivas gráficas los valores máximos se encuentran en los tejidos de mayor actividad proliferativa, Líneas y Estriado, disminuyendo con respecto a esta en los tejidos de actividad media, Tejidos y Corteza, y siendo mínimos en Cerebro y Cerebelo, los de actividad proliferativa baja. El único parámetro que se encuentra invertido es la distancia topológica media, dado que este indica el promedio del número mínimo de relevos entre 2 nodos de una red, y al ser redes con menos interacciones presentes (las de tejidos de actividad baja) al dividir entre el número total de relevos, el valor del parámetro es más grande. Al ser datos promedio, obtenidos de analizar 50 redes del mismo tipo, también se graficó el error estándar, el cual en su mayoría resulta ser muy pequeño.

Estos resultados demuestran que el aumento observado en la coexpresión global de los genes asociados al ciclo celular, en tejidos de mayor actividad proliferativa, no es sólo el resultado de un aumento en el valor de las correlaciones entre todos los pares de genes posibles sino además un aumento en la densidad y cohesión topológica de las interacciones existentes entre pares de genes.

3.3. Discusión

Queda claro que existe una diferencia notable al trabajar con los genes del ***CC*** tanto en el nivel de expresión como en el análisis de coexpresión. Hablando de las gráficas del análisis de

expresión (Figura 6), donde se muestra esto, la mediana de la expresión de los genes del ***CC*** (Flechas) son en promedio superiores a la expresión media de muestras de genes aleatorias del mismo tamaño (Histogramas), es decir que la actividad de estos genes ya sea absoluta o coordinada, indica la activación funcional en el tejido analizado.

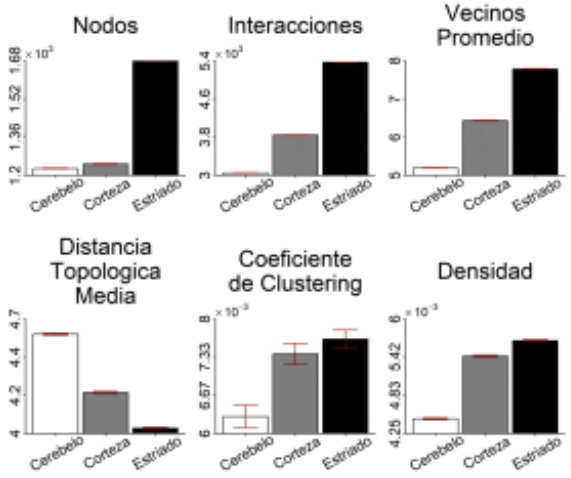


FIGURA 13. Análisis de Redes en los tejidos de la Base de Datos Brainspan. Generado con los datos de alrededor de 50 muestras independientes de 1% de las interacciones globales. . A cada una de las iteraciones del proceso de creación de redes se les calculó los parámetros mostrados, las barras de error se basan las cincuenta muestras independientes

Sin embargo se destaca en el mismo análisis (medianas de expresión) el hecho de que, en los tejidos analizados para la base de datos de Brainspan, el nivel de actividad colectiva de estos genes no es consistente con el nivel de actividad proliferativa del tejido en el que estan siendo medidos. , ya que el orden colectivo de expresión no corresponde con el órden de actividad proliferativa: Cerebelo (CB) < Corteza (CX) < Estriado (ST), y sin embargo la mediana de la expresión para los genes el tejido estriado (ST) resulta ser mucho menor que en Cerebelo (CB) y Corteza (CX). Ya desde este resultado podemos decir que el análisis de las medianas de expresión nos dice menos sobre la actividad proliferativa que el análisis de las medianas de coexpresión (Figura 7) en donde en ambos gráficos las flechas (Medianas de coexpresión del ***CC***) se encuentran dispuestas en el orden correspondiente a la actividad proliferativa de los tejidos.

En el análisis individual ($logFC$) podemos decir que se corrobora esta afirmación, ya que como se plantea al final de dicho análisis, a pesar de ser un análisis representativo debido a los diversos resultado según las muestras tomadas para el análisis, al tratarse de los mismos genes empleados, los del ***CC***, este análisis demuestra que la tendencia prevalece en el promedio de los casos, en donde para los valores de mediana de coexpresión de los genes del ***CC*** aumentan siempre en el tejido de mayor actividad proliferativa reforzando este enfoque respecto a la métrica de coexpresión.

Finalmente, la redes de coexpresión, nos muestran una vez más la concordancia de esta métrica con la naturaleza del requerimiento proliferativo de los tejidos analizados, tanto en la base de datos de FANTOM5 como en la base de datos de Brainspan en donde al realizar el corte de las interacciones a la misma altura (Mediana más grande, que coincidió con el tejido más proliferativo) y generar las redes a partir del 1% aleatorio, se puede notar que en base a las 50 iteraciones de esta selección, se mantiene la tendencia en donde hay siempre más interacciones presentes en el tejido más proliferativo, y menos del tejido que casi no prolifera.

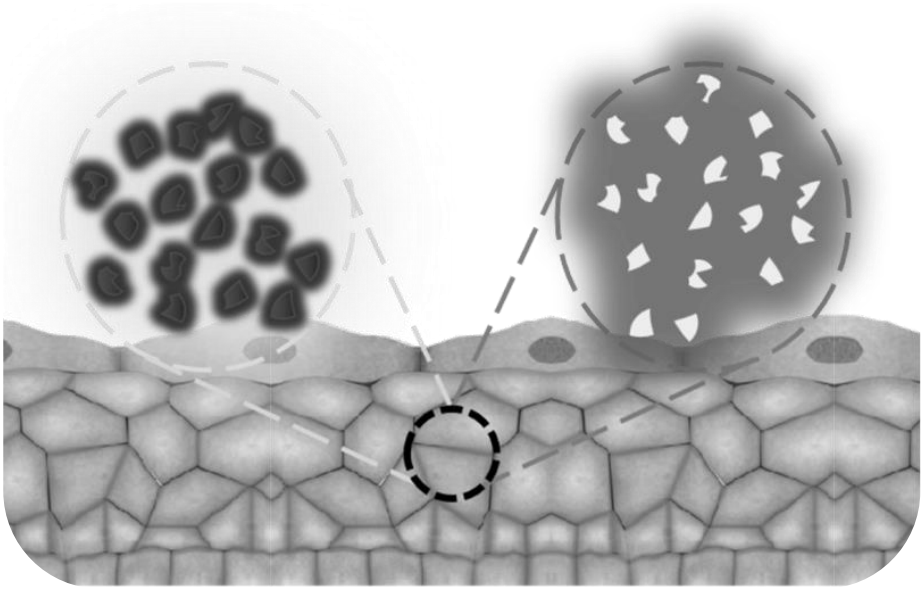


Figura 14. *Expresión vs. Coexpresión: dos perspectivas para indicar una actividad funcional*

4. Conclusiones

Este proyecto se enfocó en analizar mediante la bioestadística y genómica computacional, la relación entre dos métricas independientes de actividad colectiva de un ensamble de genes y el reclutamiento de la función con la cual se encuentran asociados estos genes. En el presente estudio, la función estudiada fue el ciclo celular (o la función celular proliferativa), utilizamos el análisis de medidas colectivas de expresión, así como las de coexpresión, con el objetivo de determinar cuál de estos es el indicador más fuertemente asociado al requerimiento de una función celular en específico. En este caso el ciclo celular.

La función de proliferación es fácilmente distinguible en los tejidos seleccionados: en promedio y mayormente los tejidos no nerviosos tienen una mayor actividad proliferativa que los tejidos nerviosos, y las líneas celulares tienen una actividad proliferativa aún mayor que los dos anteriores. En la comparación entre regiones dentro del cerebro, aquí, en las muestras postnatales utilizadas en este estudio, la proliferación no ocurre en neuronas, y sólo se da en células gliales, de modo que la actividad proliferativa corresponde con el contenido de células gliales en las distintas regiones, en donde el contenido de éstas es menor en cerebelo, mayor en corteza y mucho mayor en el núcleo estriado.

Después de haber realizado tales pruebas en los tejidos de interés, y comparado los resultados arrojados por cada una de ellas, hemos observado que el indicador que tiene una mayor consistencia, así como un mayor peso en la identificación del requerimiento funcional de la actividad proliferativa, resulta ser la coexpresión ya que en los resultados arrojados por las 3 pruebas se alinean de una manera más consistente con el requerimiento de esta función. Lo cual se haya en contraste con la métrica que tradicionalmente se usa, la expresión.

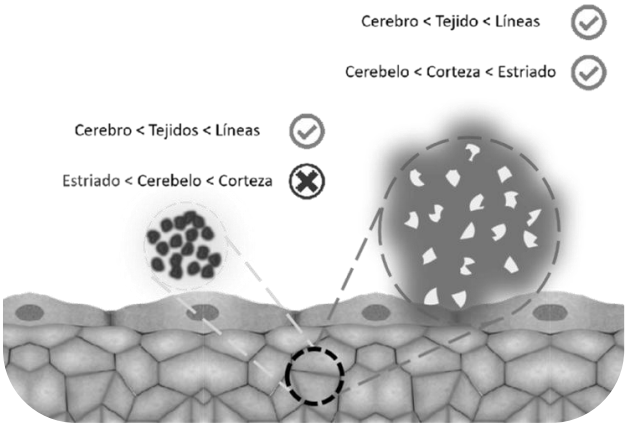


Figura 15. *Coexpresión, el indicador más robusto de la activación funcional, de acuerdo con su consistencia*

4.1. Principales Aportaciones

En el presente trabajo hemos demostrado que es la coexpresión colectiva y no el nivel de expresión, el indicador más robusto de reclutamiento funcional de genes asociados al ciclo celular y se sugiere que esta conclusión es válida para cualquier función celular. Este hallazgo se haya en contraste con el uso convencional del nivel de expresión colectiva de un ensamble de genes como indicador de reclutamiento funcional. Nuestro estudio por lo tanto propone el uso del nivel de coexpresión como un indicador más robusto de reclutamiento funcional.

Así también en el desarrollo del trabajo se han propuesto diversas técnicas o métodos tanto para el análisis de la expresión como para el análisis de la coexpresión, que pueden ser usadas de forma complementaria en estudios similares.

Finalmente, en este trabajo también se desarrolló una metodología completa para la construcción de las redes génicas, basadas en muestras representativas de redes globales de coexpresión y una estrategia para su comparación entre distintas condiciones y/o tejidos.

4.2. Ingeniera Química y la Incursión en ámbitos Biológicos

Como estudiante de la carrera de ingeniería química, he de decir que fue algo difícil incursionar en este tema, ya que nuestra preparación no está centrada en procesos biológicos, sin embargo si llegamos a tocar temas relacionado a la biología, como son los procesos biológicos “simples”, entre comillas, que se llevan a cabo en los biorreactores, al trabajar con bacterias, materias primas de origen vegetal, animal, alimenticio, etc. o en diversos proyectos que requieren del componente biológico, sin embargo nunca hemos estado limitados en ningún sentido, ya que también gracias a nuestra formación, siempre profundizamos en la metodología y en cada uno de los aspectos generales que rodean a los procesos, yendo siempre de lo general a lo particular, es así como nos es posible abrirnos caminos y diversificarse en las diferentes áreas de todas las ciencias.

Este es el caso del presente proyecto, ya que debido justamente a la formación como IQ, es que he podido desempeñarme de manera fluida en esta área, la bioinformática, donde los temas son tan diversos e importantes, como lo es el análisis de las determinantes transcripcionales y su utilidad en el entendimiento de los procesos biológicos que nos componen a los seres humanos. En particular en este estudio hemos desarrollado una metodología que nos permite entender cómo los ensambles de genes contribuyen con una función biológica determinada, herramienta de potencial aplicación para el entendimiento y manipulación de funciones biológicas de interés.

Un IQ debe estar preparado en diversos aspectos para poder adentrarse en la bioinformática, tanto en el entendimiento básico de la bioquímica, como en el uso de diferentes aproximaciones y herramientas computacionales, así como ser capaz de realizar tareas complejas y, dado que siempre existen múltiples acercamientos o metodologías, encontrar siempre la forma adecuada y más conveniente de llegar al resultado deseado.

4.3. Trabajo Futuro

Como se ha comentado anteriormente, en este trabajo usamos una batería de aproximaciones genómico-computacionales para examinar la relación entre la activación relativa de una función celular y dos métricas de actividad colectiva de ensambles asociados de genes. Sin embargo, este estudio ofrece oportunidades de ampliación y generalización. En particular se proponen las siguientes líneas de investigación futura:

- Examinar la relación entre función, expresión y coexpresión utilizando una función Celular Diferente, o varias al mismo tiempo
- Examen detallado de la equivalencia entre distintas métricas de coexpresión (correlación de Spearman, correlación de Pearson, información mutua, etc).
- Exploración de cambios en coexpresión para la identificación de genes involucrados en una función cuya base génica se desconozca.
- Realizar las redes génicas completas sin sesgar ninguna interacción para analizar su comportamiento

4.4. Difusión

Como punto final, cabe mencionar que este trabajo también forma parte de un manuscrito actualmente en preparación para su envío a una revista académica especializada y eventual publicación, bajo la dirección del Dr. Humberto Gutiérrez del Instituto Nacional de Medicina Genómica, en coautoría con el Q.F.B. Omar Franco Rodríguez.

5. Referencias

[1] N. Monroy Jaramillo y M. E. Alonso Vilatela, “La influencia de los genes del envejecimiento”, *Ciencia - Academia Mexicana de Ciencias*, México, pp. 26–31, marzo de 2011.

[2] S. A. Lambert *et al.*, “The Human Transcription Factors.”, *Cell*, vol. 172, núm. 4, pp. 650–665, feb. 2018, doi: 10.1016/j.cell.2018.01.029.

[3] T. I. Lee y R. A. Young, “Transcriptional regulation and its misregulation in disease.”, *Cell*, vol. 152, núm. 6, pp. 1237–51, mar. 2013, doi: 10.1016/j.cell.2013.02.014.

[4] M. E. Levine *et al.*, “Low protein intake is associated with a major reduction in IGF-1, cancer, and overall mortality in the 65 and younger but not older population.”, *Cell Metab*, vol. 19, núm. 3, pp. 407–17, mar. 2014, doi: 10.1016/j.cmet.2014.02.006.

[5] G. E. Zentner y S. Henikoff, “Regulation of nucleosome dynamics by histone modifications.”, *Nat Struct Mol Biol*, vol. 20, núm. 3, pp. 259–66, mar. 2013, doi: 10.1038/nsmb.2470.

[6] B. Alberts *et al.*, *Biología molecular de la célula*, 6a. ed. Barcelona: Omega, 2013.

[7] H. Yu y M. Gerstein, “Genomic analysis of the hierarchical structure of regulatory networks.”, *Proc Natl Acad Sci U S A*, vol. 103, núm. 40, pp. 14724–31, oct. 2006, doi: 10.1073/pnas.0508637103.

[8] D. S. Latchman, “Transcription factors: An overview”, *Int J Biochem Cell Biol*, vol. 29, núm. 12, pp. 1305–1312, dic. 1997, doi: 10.1016/S1357-2725(97)00085-X.

[9] P. Carmona-Saez, M. Chagoyen, F. Tirado, J. M. Carazo, y A. Pascual-Montano, “GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists”, *Genome Biol*, vol. 8, núm. 1, p. R3, ene. 2007, doi: 10.1186/gb-2007-8-1-r3.

[10] Z. Wang, M. Gerstein, y M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics”, *Nat Rev Genet*, vol. 10, núm. 1, pp. 57–63, ene. 2009, doi: 10.1038/nrg2484.

[11] A. L. Lehninger, D. L. Nelson, y M. M. Cox, *Lehninger principios de bioquímica*, 5a. ed. Barcelona: Omega, 2009.

[12] W. Yin, L. Mendoza, J. Monzon-Sandoval, A. O. Urrutia, y H. Gutierrez, “Emergence of co-expression in gene regulatory networks.”, *PLoS One*, vol. 16, núm. 4, p. e0247671, 2021, doi: 10.1371/journal.pone.0247671.

[13] D. J. Allocco, I. S. Kohane, y A. J. Butte, “Quantifying the relationship between co-expression, co-regulation and gene function.”, *BMC Bioinformatics*, vol. 5, p. 18, feb. 2004, doi: 10.1186/1471-2105-5-18.

[14] A. Marco, C. Konikoff, T. L. Karr, y S. Kumar, “Relationship between gene co-expression and sharing of transcription factor binding sites in Drosophila melanogaster.”, *Bioinformatics*, vol. 25, núm. 19, pp. 2473–7, oct. 2009, doi: 10.1093/bioinformatics/btp462.

[15] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, y J. P. de Magalhães, “Gene co-expression analysis for functional classification and gene–disease predictions”, *Brief Bioinform*, p. bbw139, ene. 2017, doi: 10.1093/bib/bbw139.

[16] B. Zhang y S. Horvath, “A general framework for weighted gene co-expression network analysis.”, *Stat Appl Genet Mol Biol*, vol. 4, p. Article17, 2005, doi: 10.2202/1544-6115.1128.

[17] H. J. Kang *et al.*, “Spatio-temporal transcriptome of the human brain.”, *Nature*, vol. 478, núm. 7370, pp. 483–9, oct. 2011, doi: 10.1038/nature10523.

[18] N. H. Tolia y L. Joshua-Tor, “Strategies for protein coexpression in Escherichia coli.”, *Nat Methods*, vol. 3, núm. 1, pp. 55–64, ene. 2006, doi: 10.1038/nmeth0106-55.

[19] A. Peixoto, M. Monteiro, B. Rocha, y H. Veiga-Fernandes, “Quantification of Multiple Gene Expression in Individual Cells”, *Genome Res*, vol. 14, núm. 10a, pp. 1938–1947, oct. 2004, doi: 10.1101/gr.2890204.

[20] A. Castillo-Morales, J. Monzón-Sandoval, A. O. Urrutia, y H. Gutiérrez, “Postmitotic cell longevity-associated genes: a transcriptional signature of postmitotic maintenance in neural tissues.”, *Neurobiol Aging*, vol. 74, pp. 147–160, feb. 2019, doi: 10.1016/j.neurobiolaging.2018.10.015.

[21] I. Seim, S. Ma, y V. N. Gladyshev, “Gene expression signatures of human cell and tissue longevity.”, *NPJ Aging Mech Dis*, vol. 2, p. 16014, 2016, doi: 10.1038/npjamd.2016.14.

[22] M. Ashburner *et al.*, “Gene Ontology: tool for the unification of biology”, *Nat Genet*, vol. 25, núm. 1, pp. 25–29, may 2000, doi: 10.1038/75556.

[23] Gene Ontology Consortium, “The Gene Ontology resource: enriching a GOld mine.”, *Nucleic Acids Res*, vol. 49, núm. D1, pp. D325–D334, ene. 2021, doi: 10.1093/nar/gkaa1113.

[24] M. Malumbres y M. Barbacid, “Cell cycle, CDKs and cancer: a changing paradigm”, *Nat Rev Cancer*, vol. 9, núm. 3, pp. 153–166, mar. 2009, doi: 10.1038/nrc2602.

[25] Dr. Agustino Martínez Antonio, “Genómica computacional”, *Revista Hypatia No.12* , Morelos, abril de 2004. Consultado: el 27 de marzo de 2022. [En línea]. Disponible en: https://revistahypatia.org/genomica-computacional.html

[26] M. Lizio *et al.*, “Gateways to the FANTOM5 promoter level mammalian expression atlas”, *Genome Biol*, vol. 16, núm. 1, p. 22, dic. 2015, doi: 10.1186/s13059-014-0560-6.

[27] I. Abugessaisa *et al.*, “FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs”, *Nucleic Acids Res*, vol. 49, núm. D1, pp. D892–D898, ene. 2021, doi: 10.1093/nar/gkaa1054.

[28] S. Ding *et al.*, “Cellular resolution anatomical and molecular atlases for prenatal human brains”, *Journal of Comparative Neurology*, vol. 530, núm. 1, pp. 6–503, ene. 2022, doi: 10.1002/cne.25243.

[29] M. C. Oldham *et al.*, “Functional organization of the transcriptome in human brain.”, *Nat Neurosci*, vol. 11, núm. 11, pp. 1271–82, nov. 2008, doi: 10.1038/nn.2207.

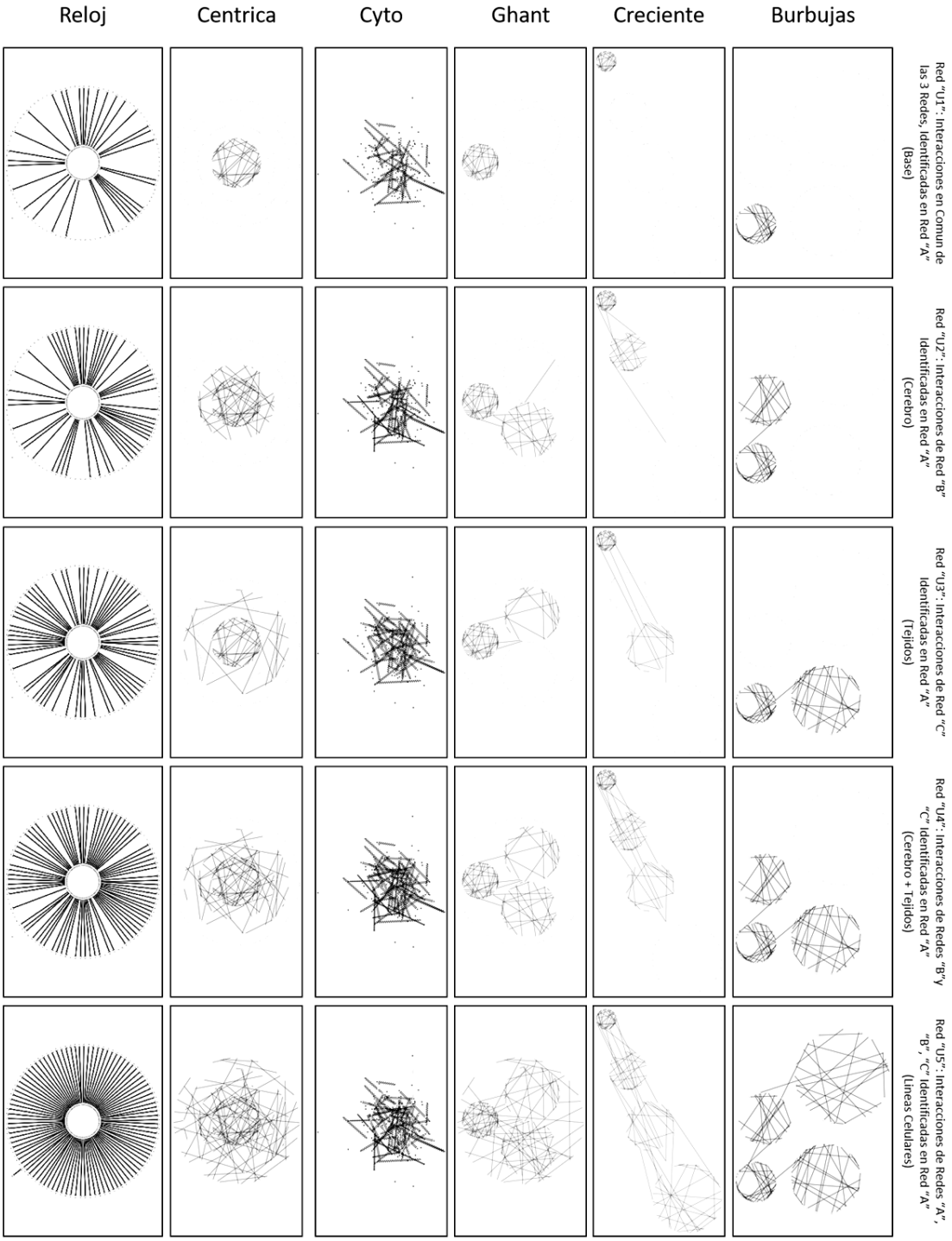
[30] F. A. C. Azevedo *et al.*, “Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain”, *J Comp Neurol*, vol. 513, núm. 5, pp. 532–541, abr. 2009, doi: 10.1002/cne.21974.

[31] S. Herculano-Houzel, “The glia/neuron ratio: How it varies uniformly across brain structures and species and what that means for brain physiology and evolution”, *Glia*, vol. 62, núm. 9, pp. 1377–1391, sep. 2014, doi: 10.1002/glia.22683.

6. Anexos

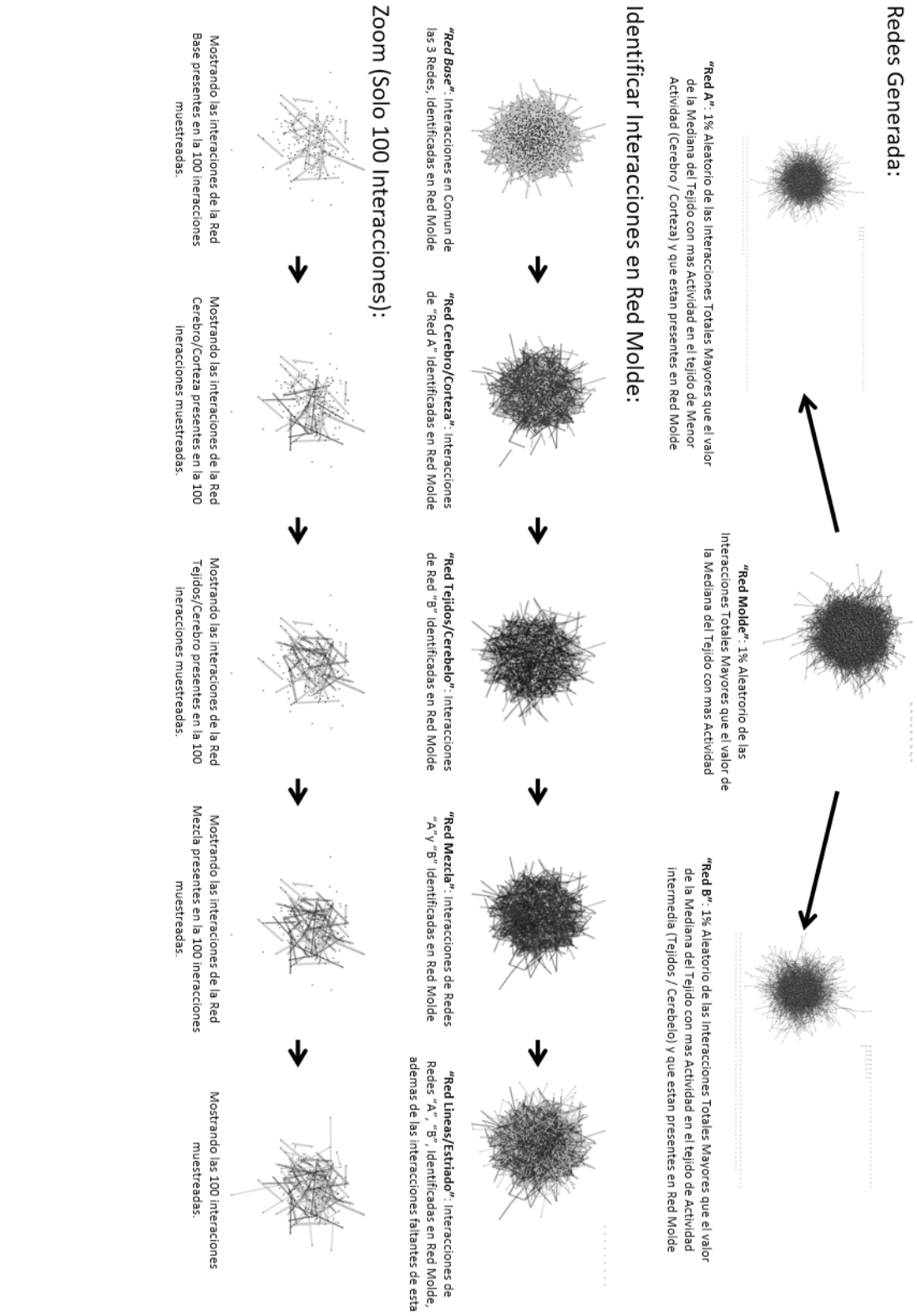
Anexo A: Diferentes Visualizaciones de la Red Génica

En este anexo, se muestran diferentes arreglos surgidos de trabajar y modificar las opciones de visualización de Cytoscape, de forma que se pueda apreciar como la red se completa con las interacciones



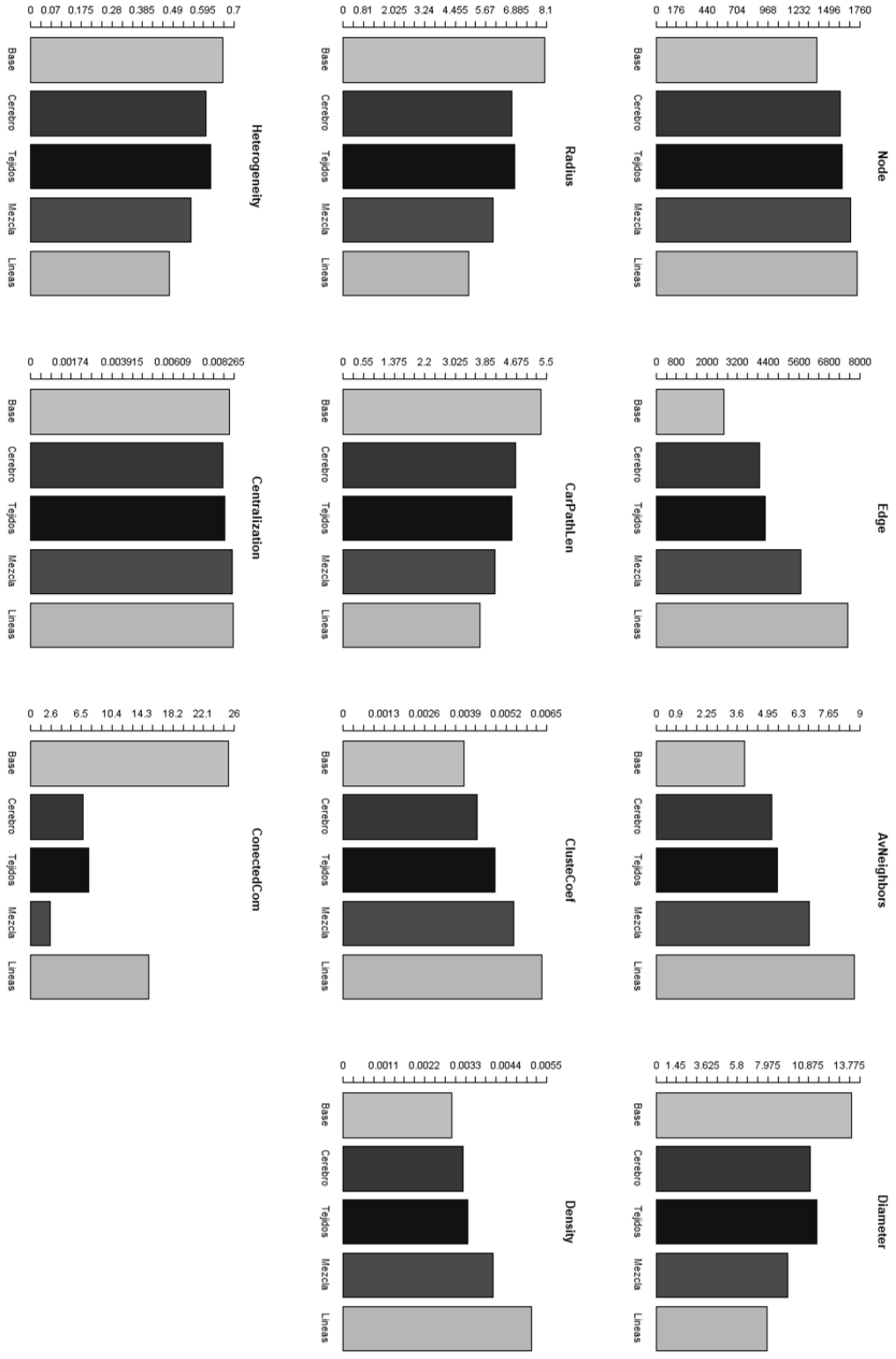
Anexo B: Gráfico de la Identificación de Redes

Se muestra gráficamente como se fueron "completando" las redes a medida que se identificaban las interacciones correspondientes

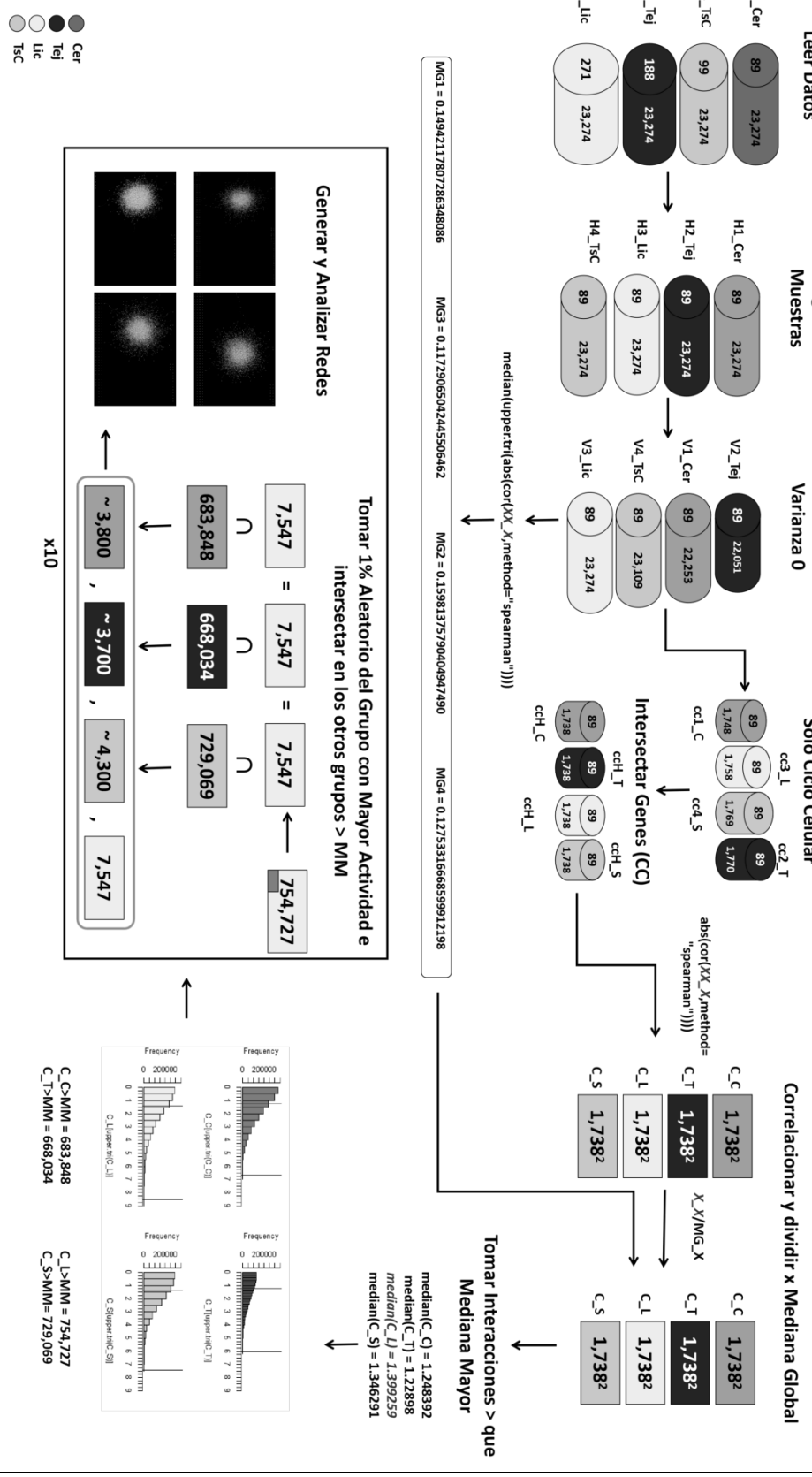


Anexo C: Datos del Análisis de Red Completo

Se muestran en las gráficas las medias de todos los parámetros arrojados por el programa Cytoscape, además de que se muestran las gráficas de los parámetros de las subredes “base” y “mezcla”.

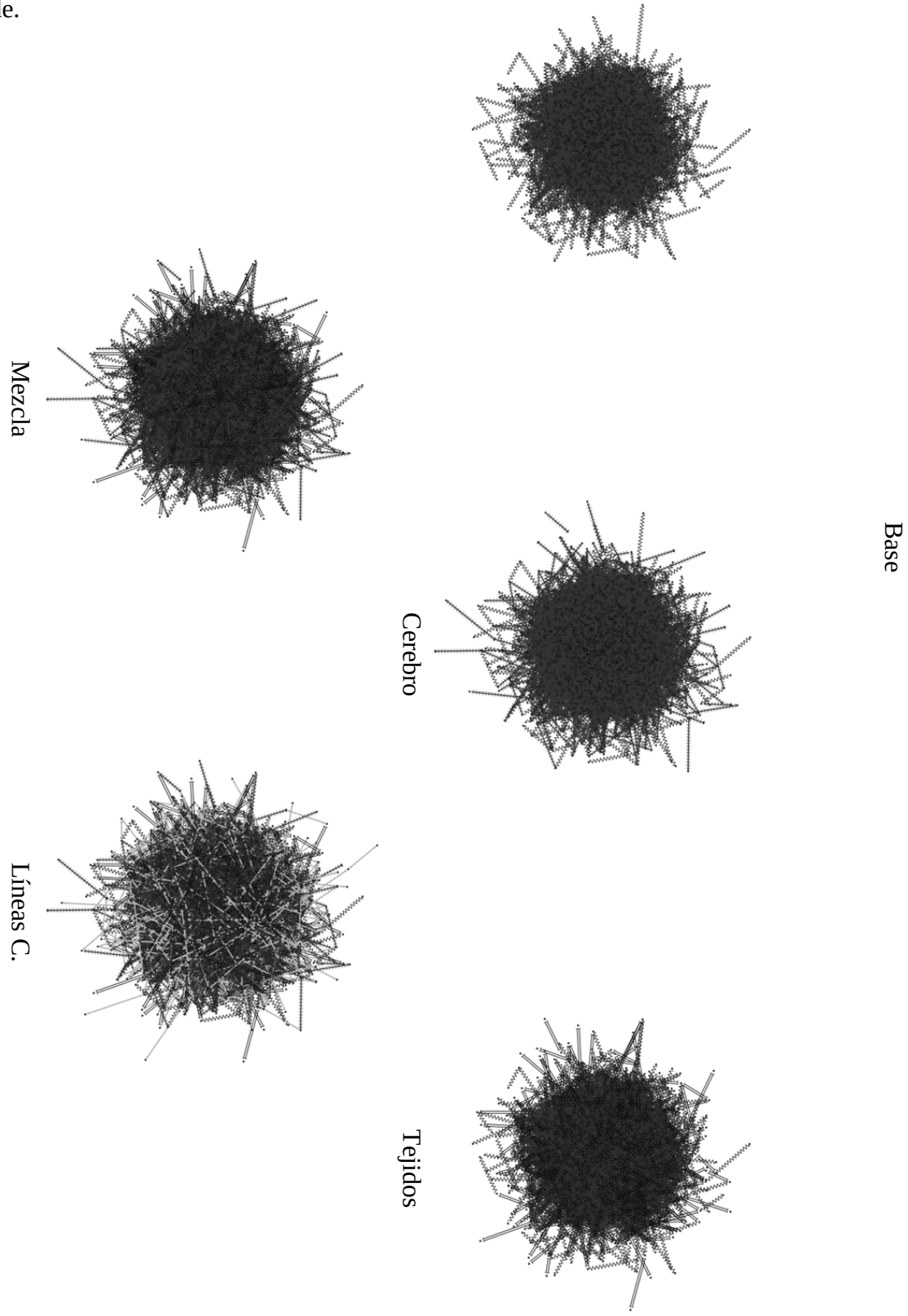


Anexo D: Diagrama de flujo del proceso de análisis de datos



Anexo E: Algunas de las redes Completas

Se muestran uno de los grupos de subredes generadas, con las interacciones identificadas en la red Molde.



Anexo F: Anexo F: Zoom de Algunas de las Redes Completas

Se muestra el zoom realizado a uno de los grupos de subredes generadas, con las interacciones identificadas en la red Molde.