UNIVERSIDAD DE ZARAGOZA

FACULTAD DE CIENCIAS

GRADO EN FÍSICA

MÉTODOS COMPUTACIONALES PARA LA CARACTERIZACIÓN DE RELACIONES CAUSALES ENTRE GENOTIPO Y FENOTIPO: HEREDABILIDAD DE LA EXPRESIÓN Y COEXPRESIÓN GENÉTICA

Autor: Regina Santesteban Azanza
Tutores: Dr. Joaquín Sanz Remón y Dr. Pierpaolo Bruscolini

Zaragoza
2020

Índice

1. Resumen

En este trabajo, se propone una metodología para inferir redes causales de regulación genética a partir de datos transcriptómicos de RNA-seq, combinados con datos genéticos recolectados en los mismos individuos.

Para la implementación de dicha metodología, se estudian e integran tres tipos de análisis distintos que se aplican normalmente de manera independiente: 1. la caracterización de redes de coexpresión, 2. el mapado de EQTLs y 3. la randomización mendeliana. Partiendo de relaciones simétricas de coexpresión entre genes (1), se pretende usar los efectos genéticos independientes que ambos genes sufren, asociados a la presencia de variantes genéticas en sus vecindades (cis-eQTLs, 2) para discernir la direccionalidad de la eventual relación causal existente entre los genes coexpresados. Este último paso se lleva a cabo aplicando randomización mendeliana (3). De este modo, se pretende caracterizar qué fracción de los niveles de correlación encontrados entre diversos pares de genes es explicado por efectos genéticos sobre cada uno de los genes que participan en la interacción, o dicho de otro modo, cuál es la componente de dichas correlaciones que puede rastrearse a un efecto genético (i.e. es heredable). Para ello se analizarán, a modo de ejemplo, datos transcriptómicos correspondientes a macrófagos humanos, extraídos de un panel de 90 donantes voluntarios, cuyos genotipos fueron caracterizados independientemente en el contexto de un proyecto de investigación en curso en el que participan los tutores de este TFG.

A nivel formativo, este TFG tiene como principal objetivo la exposición de las principales metodologías experimentales en genómica contemporánea, especialmente transcriptómica y caracterización genotípica de variantes genéticas, así como la exploración de herramientas computacionales de amplia implantación en el campo, combinadas alrededor del pipeline analítico descrito.

2. Palabras clave

RNA-seq, Single Nucleotide Polymorfism (SNP), Matrix-eQTL, Mendelian Randomization.

3. Objetivos

La literatura de inferencia de redes de regulación es vasta [10] [11], pero, a menudo es complicado inferir relaciones causales (dirigidas), a partir de patrones de coexpresión (no dirigidas) [12], ya que a veces las interacciones no están claramente mediadas por sólo un mecanismo molecular, entre otros problemas. Muchos de estos métodos se basan en la integración de experimentos de tipo knock-out (experimentos donde se bloquea la actividad de un gen, para estudiar su funcionalidad, observando las consecuencias experimentadas por la muestra en consecuencia de esa desactivación) los cuales son costosos, difícilmente permiten una caracterización sistémica y generan, a veces, condiciones fisiológicas artificiales, cuya representatividad del rol del factor regulador alterado resulta discutible [13]. Un recurso es el uso de series temporales [14], pero esto a veces no es posible, o presenta, asimismo, limitaciones.

Aquí proponemos analizar al mismo tiempo datos de expresión y genotípicos para discernir las relaciones de direccionalidad causal a partir de patrones de coexpresión. Para ello adaptaremos la randomización mendeliana, una técnica estadísticas que se utiliza en epidemiología generalmente, para el estudio del efecto causal, a partir de datos experimentales, de un factor de riesgo de una enfermedad. La idea ha sido ya propuesta en la literatura [15], pero su aplicación ha sido solo marginalmente explorada en [16] [17].

En este TFG proponemos su implementación y aplicación para el análisis de redes de regulación en macrófagos humanos.

4. Introducción: Ingredientes

4.1. Redes de coexpresión genética

Las Redes de coexpresión genética, se ocupan de describir los patrones de correlación de la expresión de los genes a través de conjuntos de muestras en las cuales se ha medido simultáneamente la expresión de un número grande de ellos. En este TFG analizaremos datos de RNA-seq, un tipo de datos cuyas bases experimentales y principales características técnicas describimos en el documento anexo a esta memoria. En esta técnica, el ARN de cada muestra se trocea en pequeños fragmentos, que después se secuencian y mapean con respecto a un genoma de referencia, obteniéndose, como principal medida de expresión, el número de fragmentos correspondientes a cada gen en cada muestra. Esta medida de expresión cruda tiene que normalizarse, y transformarse tal como se explica en el anexo, antes de poder ser analizada, para encontrar pares de genes coexpresados, tal como describimos a continuación

4.1.1. Eliminación de factores técnicos y Estabilización de la varianza

Antes de comenzar con el análisis de las correlaciones entre los diferentes genes, debemos tener en cuenta que existe una relación sistemática entre el número promedio de lecturas que podemos detectar de un gen determinado y la dispersión técnica (ruido) sujeta a las mediciones de cada individuo para cada gen. Este hecho representa un problema técnico importante, pues puede introducir sesgos e incrementar el ruido en los datos, impidiendo caracterizar relaciones de coexpresión reales, e induciendo espuriamente otras. Para solucionar este problema, procederemos a realizar la llamada estabilización de la varianza [5].

La relación a la que hacemos referencia en el párrafo anterior viene descrita por:

$$var(\log_2 r) = \frac{1}{E(r)} + \phi$$

Siendo var la varianza, r los reads o lecturas, $E(r)$ la expresión de los genes (número esperado de reads) y ϕ una constante dependiente del estudio.

Para tratar este problema debemos ahondar en el análisis de modelos lineales. Para describir estos modelos se construyen cuatro matrices diferentes. La primera, la llamamos matriz de diseño D donde las filas corresponderán con los individuos de nuestro estudio y las columnas a las variables del modelo, que son aquellas características de cada muestra acerca de las cuales existe una expectativa en virtud de la cual esperamos que exista una correlación significativa entre ellas y la expresión de una fracción de los genes a analizar. En nuestro experimento, estamos analizando un conjunto de muestras esencialmente homogéneo, la única variable de nuestro diseño es técnica, y se corresponde a los lotes experimentales en lo que las muestras fueron procesadas. En Biología, y especialmente en genómica, tener en cuenta la variación en los resultados experimentales que se verifica entre lotes es un tema absolutamente central [18]. La segunda corresponde con nuestros datos de expresión de genes llamada matriz de modelización Y. Teniendo los datos de expresión de cada gen(fila) para cada individuo(columna). La cuarta va ligada a la estadística y corresponde a los residuos de nuestros datos, se pueden sacar gracias a funciones como $lmFit$ encargada de hacer ajustes por mínimos cuadrados y por $eBayes$ encargada de hacer tests estadísticos. Con estas tres y gracias al modelo:

$$Y^T = D \cdot B + E$$

Podemos calcular la matriz B que representa la variación de expresión de cada gen en relación con las variables del modelo y E corresponde con la matriz de residuales.

Pasando a nuestro estudio en concreto, tenemos una matriz de expresión de los genes descrita por:

$$y_{ij} = \log_2(\frac{r_{ij} + 0,5}{R_j + 1} \cdot 10^6)$$

Formada gracias a la herramienta de *voom* encargada de realizar la TMM normalization, descrita anteriormente. El término R_j es igual a $N_j * TMM_j$ donde el factor TMM lo hemos utilizado para normalizar las variaciones relativas en el output total entre muestras. Para visualizar la relación, representamos dos magnitudes en una gráfica. En el eje de las X representamos a raíz de lo anterior $x = \langle \log_2(r_{ij} + 0,5) \rangle$ y para el eje de las Y ayudándonos de la matriz de los residuos, calculamos la desviación típica como:

$$\sigma_i = \sqrt{\frac{1}{GL}} \sqrt{\sum_j (\epsilon_{ij} - \langle \epsilon_i \rangle)^2}$$

Donde GL son los grados de libertad. Una vez calculada, representamos la raíz de la desviación típica frente al eje X ya mencionado, pudiendo así visualizar la relación descrita anteriormente(*Figura 1*).

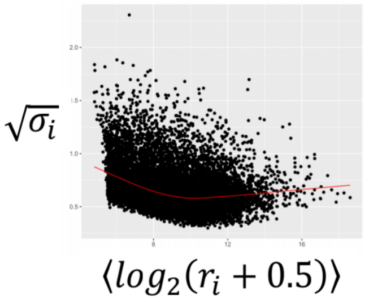

Figura 1: Ejemplo: Visualización relación varianza-expresión.

Para representar la línea en rojo que podemos diferenciar en la figura y estudiar la tendencia de nuestra nube de puntos, debemos hablar de la función *lowess*. Es una función que nos permite realizar una regresión local, combinando la regresión lineal por mínimos cuadrados con la regresión no lineal, al utilizar la primera en subconjuntos de puntos, para crear una función correspondiente a la parte sistemática de la variación de los datos.

Sabemos que la raíz de la varianza puede entonces interpretarse como la suma de dos componentes: una varianza técnica, asociada a la relación dicha anteriormente, mas una posible desviación por motivos biológicos $\sqrt{\sigma_i} = \sqrt{\sigma_{tech}} + \delta_{Bio}$. La línea roja representa la $\sqrt{\sigma_{tech}}$ que expresa la predicción de escala del tamaño de los residuos, en función del valor medio de la expresión de los genes. Es decir, haciendo $f(\log_2(r_i + 0,5))$ siendo $f()$ la función *lowess* calculamos cual debe ser el tamaño de los residuos. Por lo que δ_{Bio} será positiva si el gen es más variable entre muestras de lo que se espera de él en función de la media de su expresión o negativa si es menos variable.

Para acabar con esa relación sistémica, debemos convertir la curva de la varianza en una recta. A lo que llamamos estabilizar la varianza. Para ello, redefinimos nuestra matriz y_{ij} como $y_{ij}^{Model} = y_{ij} - \epsilon_{ij}$ y definimos los llamados residuos escalados $\widetilde{\epsilon_{ij}} = \epsilon_{ij}/\sigma_{tech}$, esto hará que la varianza asociada a los residuos escalados se estabilice, habiéndonos desecho de la dependencia con la media de expresión. En lo sucesivo, nuestra variable dependiente será esta matriz de residuos escalados, que puede interpretarse como los niveles de expresión transformados tras 1) eliminar los efectos de la variación técnica entre lotes. Y 2) corregir las asociaciones sistemáticas entre expresión media y varianza que son características de los datos de RNAseq. Ambas son modificaciones que, se sabe, contribuyen a la obtención de resultados más robustos en análisis de coexpresión [24].

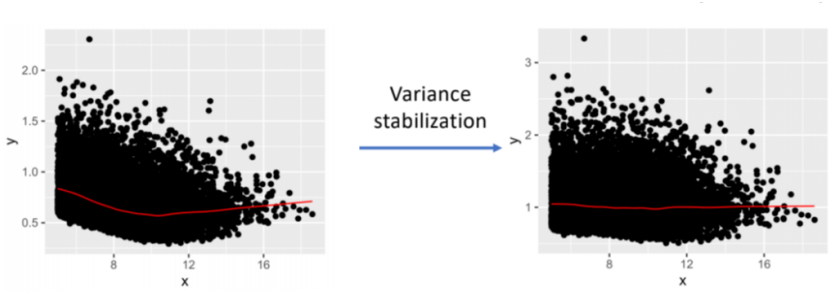

Figura 2: Ejemplo: Estabilización de la varianza.

4.1.2. Correlación Parcial

Una vez que los datos de expresión han sido pre-procesados, el primer paso del proceso será obtener una matriz de correlaciones entre pares de genes. Sin embargo, los tests de correlación habituales no pueden usarse directamente, puesto que las variables, tras haber eliminado los efectos de lote, no son totalmente independientes.

De este modo, en lugar de una correlación estándar entre los niveles de expresión independientes de un par de genes, tendremos que correlacionar los residuos de cada una de estas variables con respecto al factor "lote", lo que se conoce como un test de correlación parcial.

Para ello, lo primero que haremos es calcular la correlación de Pearson r para cada par de vectores de residuos, pertenecientes a cada par de genes. Más adelante hacemos el test estadístico calculando:

$$t_{n-k-n} \sim r\sqrt{\frac{n-k-2}{1-r^2}}$$

Siendo n los grados de libertad y k el número de covariables en nuestro modelo. Después, se calculan los pvalue correspondientes a ese modelo.

Y por último, se aplica una corrección para las comparaciones múltiples. En nuestro caso utilizaremos la corrección de Benjamini-Hochberg [7]. Al hacer múltiples medidas crece la probabilidad de encontrarnos con un evento raro y por consecuencia, crece también la probabilidad de rechazar la hipótesis nula, a esto se le denomina error de Tipo I. Teniendo más probabilidades de llegar a un resultado significativo por puro azar. Esta corrección se aplica calculando el llamado False Discovery Rate (FDR). Procedimiento que comienza obteniendo la distribución de p-values del conjunto de tests analizados, $f(p)$, y se compara con la que tendríamos si la hipótesis nula fuese cierta, que sería uniforme, $f_0(p)$ (*Figura 3*).
Para cada p, se calcula el número de tests que aceptamos bajo ese umbral de significancia (área bajo la curva $f(p)$ a la izquierda de p), y lo comparamos con el número de tests que obtendríamos si aceptásemos como significativos los tests con p-value menor que ese en un ensayo aleatorio, que serían falsos positivos (área a la izquierda de p bajo $f_0(p)$. El ratio entre la segunda área dividida entre la primera es el FDR, sobre el cual impondremos en nuestro caso un umbral del 5 %. Esto significa que esperamos que el 5 % de nuestros enlaces (y no más), sean falsos positivos.

Una vez aplicadas las correcciones y tests estadísticos anteriores, podemos comenzar con la construcción y el análisis de las redes de co-expresión.

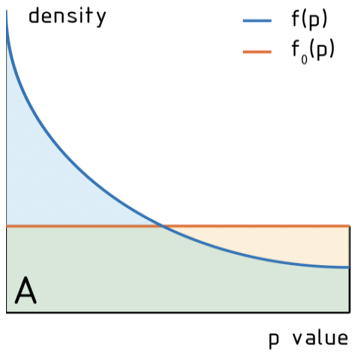

Figura 3: Método Benjamini-Hochberg.

4.2. Mapado relaciones genotipo-expresión: Matrices eQTL (Expression Quantitative Trait Loci)

Las eQTL (Expression Quantitative Trait Loci) [1] [25] son loci genómicos que describen la variación de nivel de expresión del ARN mensajero (ARNm), que viene a ser la cantidad de transcripciones que sufre o la presencia de una proteína. El análisis de las eQTL, enlaza la variación de niveles de expresión al genotipo, esencialmente se busca una regresión linear para describir esa asociación. Para este análisis computacional utilizaremos un software de amplia implantación, llamado Matrix-EQTL, que utiliza una implementación ultra rápida que permite la realización de cientos de miles de estos tests de regresión en unos pocos minutos en una máquina promedio.

El principal objetivo de este análisis es identificar los SNPs que están significativamente relacionados con expresiones de genes conocidos. Estas asociaciones nos ayudan a descubrir los factores genéticos que desencadenan ciertas enfermedades y determinar las vías que se ven afectadas por ellas. Además, nos ayudan a determinar los puntos críticos, que son regiones del ADN que afectan a los niveles de expresión de múltiples genes, a construir redes de causalidad y descubrir subclases de fenotipos. En el documento anexo a esta memoria, adjunto una descripción detallada de los métodos experimentales que se utilizan para obtener datos genotípicos como los que hemos utilizado en este trabajo.

4.2.1. Conceptos clave

Presentamos algunos de los ingredientes principales para la realización de este análisis:

Alelo menor y mayor

Para este tipo de análisis, el verdadero interés reside en la frecuencia de los alelos. Cuando observamos los alelos de una población, llamaremos alelo mayor al que sea más frecuente en una posición en concreto para distintos individuos, mientras al de menor frecuencia lo llamaremos alelo menor. Siendo el alelo mayor aquel alelo más probable.

Trans-eQTL y Cis-eQTL

Existen dos modos fundamentales en organismos haploides a través de los cuales una variante genética puede regular la actividad transcripcional de un gen. Estas se denominan regulación -cis y regulación -trans.

En un evento -cis, la variante sólo afecta la expresión del gen localmente, es decir, en el mismo cromosoma donde se encuentra la variante. Un ejemplo es un SNP que aparece cerca del promotor de un gen modificando el binding de un factor de transcripción en el mismo cromosoma donde la variante yace.

En trans, sin embargo, la presencia de la variante en una hebra contribuye a variar el nivel de expresión de un gen en ambos cromosomas. Un ejemplo es un SNP que cambia el aminoácido de un factor de transcripción, el cual, a su vez, modifica la actividad del mismo sobre un segundo gen diana. Si encontramos el SNP en uno de los dos cromosomas, al menos una parte del pool de factor de transcripción tendrá la variante, lo cual afectará a la transcripción del gen diana desde los dos cromosomas.

Aunque a veces es complicado distinguir variantes que actuan en -cis de las que lo hacen en -trans, una diferencia esencial es que las primeras suelen localizarse cerca del origen del gen que regulan, mientras que, para las variantes -trans, no existe correlación entre la posición de la variante y su target.

4.2.2. Métodos

El análisis eQTL [2] utiliza, para identificar asociaciones entre SNPs y los valores de expresión de sus genes diana, la modelización del impacto del genotipo como un aditivo lineal (modelo de mínimos cuadrados) o categórico (modelo ANOVA). Los modelos pueden contener covariables para tener en cuenta factores como pueden ser el sexo, estructura poblacional o variables clínicas. El análisis admite errores heterocedásticos (donde la varianza de éstos no permanece constante para todas las observaciones) y correlacionados para tener en cuenta la relación de las muestras. Al ser un análisis de datos que puede llegar a involucrar la realización de $10^9 - 10^{10}$ tests estadísticos se utilizan varias optimizaciones.

A continuación,el algoritmo para el modelo de regresión lineal simple, que no incluye covariables y supone errores homoscedasticos, y no correlacionados, que es el que utilizaremos en este TFG.

REGRESIÓN LINEAL SIMPLE

Es uno de los métodos más comunes. Consiste en que para cada par gen-SNP con el SNP codificado por 0,1 y 2 según la frecuencia del alelo menor, 0 correspondería a un genotipo homocigoto del alelo más común en la población, 1 al heterocigoto y 2 al homocigoto del alelo menos común. Se supone que la asociación entre la expresión génica g y el genotipo s es lineal:

$$g = \alpha + \beta s + \epsilon$$

Este análisis implica el cálculo de los valores de las medias de las muestras como $\bar{g}$, $\bar{s}$, la pendiente $\hat{\beta}$, la ordenada en el origen $\hat{\alpha}$ y de los residuos ϵ_i que son variables aleatorias e independientes sujetas a $N(0, \sigma^2)$, además de la suma total de los cuadrados SST y la suma de los cuadrados de los residuos SSE.Todo esto acompañado de un test estadístico a nuestra elección, que calcula el $p - value$. Para una computación más rápida y unos outputs más ligeros, donde puede elegirse que el programa no devuelva el $p - value$ de todas las parejas SNP-gen. Se compara el resultado del test estadístico elegido para cada pareja mencionada con un umbral calculado y se devuelve por output el $p - value$ solo de aquellas parejas que lo excedan. Obteniendo sólo el $p-value$ para aquellas parejas significativas ahorrándonos memoria.

La elección del test estadístico es importante. Para la regresión lineal simple tenemos los siguientes: t, F, R^2 y LR, y pueden escribirse en función de la correlación de las muestras $r = cor(g, s)$:

$$t = \sqrt{n-2}\frac{r}{1-r^2} \quad ; \quad F = t^2 = (n-2)\frac{r^2}{1-r^2} \quad ; \quad R^2 = r^2 \quad ; \quad LR = -n\log(1-r^2)$$

Quedándonos el cálculo de nuestra correlación como:

$$r_{gs} = cor(s, g) = \frac{\sum(s_i - \bar{s})(g_i - \bar{g})}{\sqrt{\sum(s_i - \bar{s})^2(g_i - \bar{g})^2}} = \sum s_i g_i = \langle s, g \rangle$$

Donde $\langle s, g \rangle$ representa el producto interno de los dos vectores. Si tenemos una matriz llamada S que contiene las medidas de SNP de la muestra y G la matriz de la expresión de los genes donde cada valor contiene la expresión de un único gen. Podemos construir la matriz de correlación como la multiplicación de las dos matrices ilustrada en la *Figura 4*. A partir de dicha matriz, se ejecutan los tests estadísticos descritos antes: nosotros usaremos el primero, el t-test.

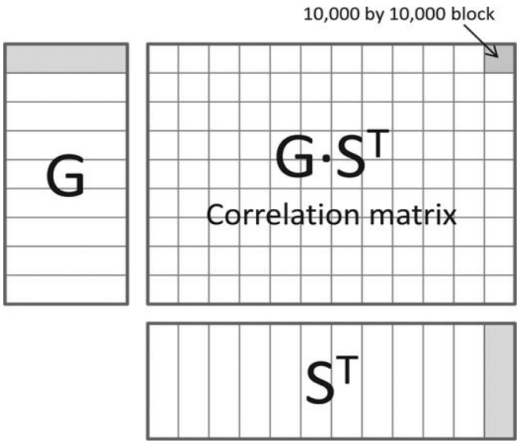

Figura 4: Matriz de Correlación

Para facilitar el análisis de estas macromatrices, cuyo cálculo constituye sin duda el paso más costoso computacionalmente del problema, se agrupan en bloques de 10.000 variables tal como aparece en la figura y se lleva a cabo en parejas de bloques. En resumen, el algoritmo de $Matrix - eQTL$ para la regresión lineal separa el input de las matrices en bloques de 10.000 variables y para cada par de bloques, calcula su respectivo bloque de correlación a través de la multiplicación de matrices, halla las correlaciones que exceden del umbral calculado y las somete a un test estadístico que calcula su $p - value$. El algoritmo permite la identificación de tests -cis, a partir de la definición de un umbral de distancia, permitiendo sólo el cálculo de los tests para los que el genotipo esté dentro de esa distancia con respecto a la posición del gen diana (mapado de cis-EQTL), lo cual reduce inmensamente la cantidad de tests a hacer, y por tanto la memoria necesaria y el tiempo de computación. En nuestro trabajo, nos hemos limitado al mapado de cis-EQTLs, definiendo para ello una distancia máxima entre el gen diana y el SNP a testar de 100 kilo-bases. Solo los SNPs dentro de estas ventanas, definidas para cada gen, son testados.

4.3. Randomización mendeliana para dar el salto de correlación a causalidad en epidemiología

La Randomización mendeliada [3] es una técnica epidemiológica de aproximación para el estudio del efecto causal de los factores de riesgo que existen en una enfermedad a partir de datos observacionales. Utiliza la variación cuantificada de genes de funciones conocidas, para estimar la presencia o ausencia de

un el efecto causal de una variable modificable (como puede ser el entorno) sobre un fenotipo (enfermedad).

4.3.1. Ingredientes

La aproximación se construyó a partir de la idea principal de que el genotipo se asigna aleatoriamente a causa de la meiosis, pudiendo considerarlo como una variable instrumental (G). Es decir, se trata a las variantes genéticas como instrumentos. Para poder considerarlos como tal, esas variables deben cumplir una serie de características:

- Debe ser independiente a los factores de confusión U.
- Debe ir asociada a un factor de riesgo X.
- Debe ser independiente del resultado Y (enfermedad), relacionado con el factor de riesgo X y los factores de confusión U, es decir, no debe estar directamente relacionado con el resultado.

Para ilustrarlo mejor veremos el DAG (Direct Acyclic Graph) asociado a este método computativo representado en le *Figura 5*. Donde podemos identificar SNPs como variables instrumentales.

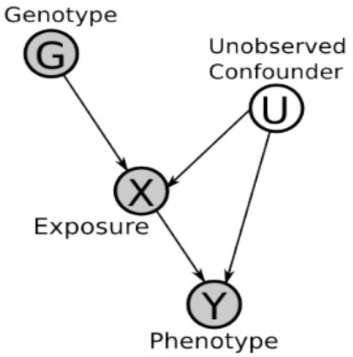

Figura 5: DAG Randomización Mendeliana.

4.3.2. Motivación y problemas

Para entender mejor el funcionamiento de esta aproximación, pondremos el ejemplo que llevó al desarrollo de este método. Un científico llamado Martijn Katan en 1986 hizo un estudio sobre si el bajo nivel de colesterol (factor de riesgo, o exposición, X) aumentaba la probabilidad de padecer un cáncer (variable resultado, o enfermedad, Y) [19]. Supuso que tanto los niveles de colesterol como la probabilidad de tener cáncer podrían estar relacionados con la dieta (factor de confusión U). También existía la posibilidad de que el cáncer provocara niveles de colesterol bajos, a esto se le denomina causalidad inversa.

En el curso de su estudio, Katan observó que los pacientes que padecían abetalipoproteinemia (incapacidad de absorber colesterol), no presentaban una predisposición a padecer cáncer. Esto condujo a Katan a la idea de encontrar un gran grupo de personas que tuviesen una predisposición genética a tener un nivel bajo de colesterol, para comparar su riesgo de padecer cancer con aquél asociado a una población sin esa característica genética. Nótese, que no quiere decir que las variables genéticas sean las causas directas de la enfermedad. Pero se utiliza esa asociación para posibilitar la inferencia de los efectos ambientales de riesgo (en este contexto, los niveles de colesterol) sobre la probabilidad de padecer cáncer.

El gen que codifica la Apolipoproteína E se denomina *ApoE* mientras que la variación de éste asociada al bajo nivel de colesterol es *ApoE*2. Los pacientes que nacen con este alelo, tienen un nivel bajo de colesterol en el suero sanguíneo desde su nacimiento. Sabiendo que la meiosis es un proceso aleatorio, se asume que los individuos que contienen el gen *ApoE* o la variante *ApoE*2 presentan ninguna característica significativa que sea sistemáticamente distinta entre ellos, lo cual asemeja el estudio a un ensayo clínico aleatorizado. Si el bajo nivel de colesterol es un factor de riesgo para el cáncer, los pacientes del grupo *ApoE*2 deben tener más riesgo de padecer cancer.. Katan hizo esta sugerencia, pero más adelante algunas conclusiones fueron refutadas por ensayos aleatorios controlados. Para entenderlo bien, debemos indagar en las suposiciones biológicas relacionadas con esta conclusión y los problemas que plantean.

Estratificación poblacional

Hace referencia a la diferencia sistemática en la frecuencia de alelos entre subpoblaciones debido a la ascendencia. Por ejemplo, la separación física conduce a la ausencia de aleatoreidad de apareamiento, que conlleva a una diferente deriva genética entre distintas subpoblaciones, haciendo que el genotipo no sea una variable completamente aleatoria entre ellas. En otras palabra: en el ejemplo de Katan, los sujetos con el alelo APOE2 vs los sujetos que no lo presentan, tienen distinta probabilidad de presentar alelos diferentes en otros muchos alelos simultáneamente, cuya herencia no es del todo independiente. En la medida de la importancia cuantitativa de este problema (que dependerá mucho de la muestra y de las variables instrumentales usadas), este puede ser un problema que imposibilite llegar a ninguna conclusión válida; al menos, si la estratificación poblacional no se contempla de manera adecuada en los modelos estadísticos.

Pleiotropía

Representa el hecho de que un gen puede afectar en varios fenotipos diferentes, hecho que la randomización mendeliana no tiene en cuenta. Hacemos la suposición que el genotipo observado solo influye en el fenotipo por la vía que estamos estudiando. Volviendo al ejemplo, asumimos que el gen *ApoE*2 solo afecta a los niveles de colesterol y no puede influir en el desarrollo de un cáncer por otras vías. Esta suposición debe estar acompañada por un conocimiento previo. Teniendo múltiples variables instrumentales (múltiples SNP) podemos atenuar este problema, ya que al ser todos consistentes, es poco probable que todos tengan otras vías que causen el mismo cambio.

Podemos concluir que el verdadero DAG no es exactamente el mostrado en el apartado anterior siendo este una aproximación, el verdadero grafo, que ya no sería un DAG, sería, teniendo presentes todos los factores biológicos anteriormente mencionados el relativo a la (*Figura 6*).

En cambio, los problemas anteriormente mencionados no presentan un impacto significativo en este estudio. La razón por la cual es razonable asumir que no existen relaciones de pleiotropía entre las variables instrumentales identificadas en el mapeo de EQTLs, es que, típicamente, un SNP es EQTL de un solo gen target; y, de serlo en dos, estos habrían de aparecer cerca en el mismo cromosoma. Específicamente, al requerir que las parejas coexpresadas tengan que estar más alejadas de 200 KB entre sí(requisito implementado en nuestro código), estamos imposibilitando que un cis-EQTL de uno de los miembros de una pareja lo sea a la vez del otro miembro de la misma. Podríamos tener, eso sí, casos en los que una variable instrumental es trans-EQTL de otro gen, no relacionado con el par, el cual, a su vez, estaría correlacionado con el segundo gen del par. Sin embargo, los EQTL en trans son mucho menos numerosos, y difíciles de caracterizar con significancia estadística, por lo que su mapado cae más allá del objetivo de este TFG.

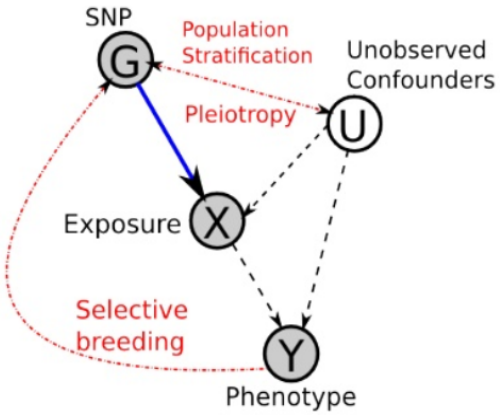

Figura 6: Grafo causal completo.

En cuanto a la estratificación poblacional, es un problema que no se presenta en la cohorte analizada, tal como se pudo comprobar en experimentos de control de calidad que se realizaron previamente al comienzo de este TFG.

4.3.3. Método cuantitativo

La idea fundamental detrás de nuestro TFG, como se expuso en la sección de motivación, consiste en interpretar los pares de vectores de expresión correlacionados que conectan genes en una red de coexpresión no-dirigida como pares de variables exposición/enfermedad en un test de randomización mendeliana. En este contexto, la caracterización genotípica de los individuos, nos ofrece las variables instrumentales asociadas a cada par de manera masiva, a través del mapado de EQTLs. [4] [8] [9].

El DAG de nuestra randomización Mendeliana estaría representado por la *Figura 7*. Queriendo testar si la coexpresión de los genes A y B es consecuencia de una influencia causal del gen A sobre el gen B. Donde nuestras variables instrumentales corresponden con las cis-EQTLs.

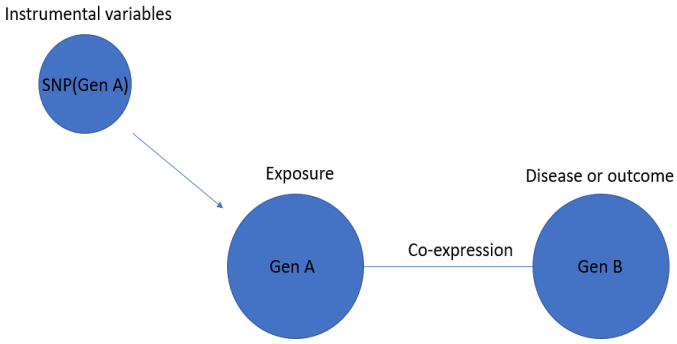

Figura 7: DAG de nuestro estudio.

Matemáticamente podemos describir la variante genética j (SNP) con el factor de riesgo (expresión del gen A) representado por $\hat{\beta}_{Xj}$ con un error estándar $se(\hat{\beta}_{Xj})$ y el outcome (expresión del gen B) representado por $\hat{\beta_{Yj}}$ con un error estándar de $se(\hat{\beta}_{Yj})$. El efecto causal del factor de riesgo en el resultado $\hat{\theta}_j$ y su error estándar pueden estimarse como:

$$\hat{\theta}_j = \frac{\hat{\beta}_{Yj}}{\hat{\beta}_{Xj}} \quad ; \quad se(\hat{\theta}_j) = \frac{se(\hat{\beta}_{Yj})}{\hat{\beta}_{Xj}}$$

Siendo esta estimación de causalidad válida para una variable instrumental, cuando poseemos varias variables instrumentales por relación de causalidad, siendo este nuestro caso al tener varias cis-EQTL por gen, podemos obtener una variable más precisa de estimación causal, utilizando toda la información de nuestras variables instrumentales.

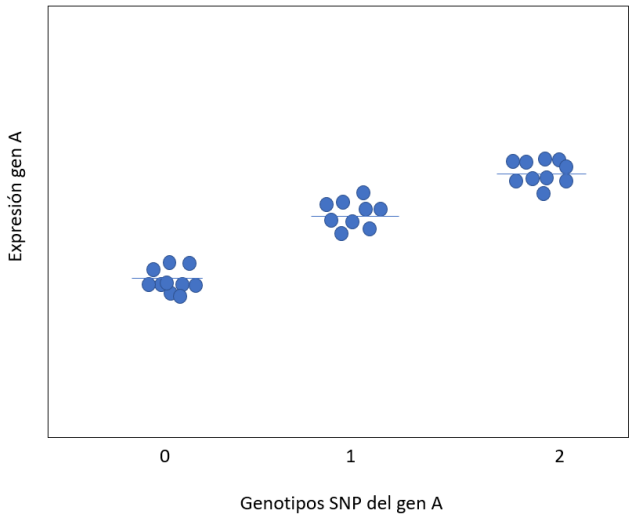

Figura 8: Gráfico expresión vs genotipo.

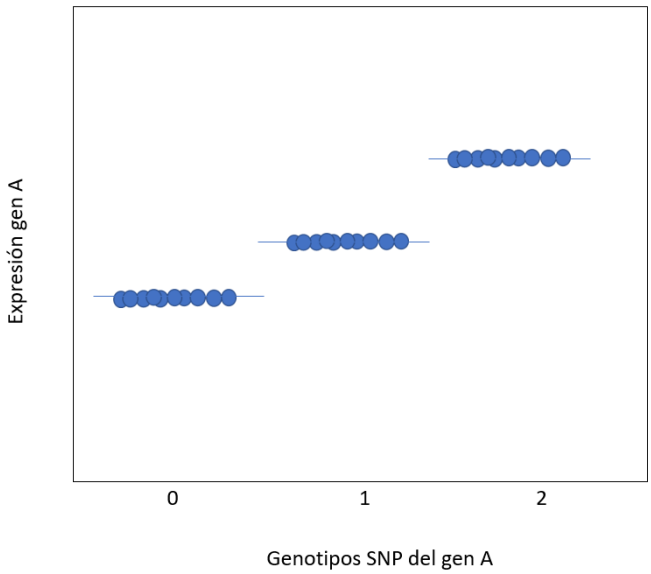

Figura 9: Gráfico expresión ajustada vs genotipo.

El estimador causal, construido como la ratio de los coeficientes de la regresión de la variable instrumental sobre los dos genes, coincide con el efecto que puede estimarse a través de una regresión de dos pasos.

Según este procedimiento, un primer paso consiste en una regresión de la variable instrumental sobre el gen A (*Figura 8*); cuyos residuales son sustraídos de la expresión del gen A para obtener la componente de su varianza A_t que puede serle atribuida a la variable instrumental (*figura 9*).

El segundo paso del procedimiento consiste en ajustar un modelo donde la expresión del gen B (variable dependiente) es ajustada en función de A_t (variable independiente). La pendiente obtenida a través de este procedimiento coincide con el estimador θ calculado más arriba.

Por último, realizaremos el mismo procedimiento en sentido contrario, tomando las cis-EQTLs del gen B como variables instrumentales y estudiando la causalidad de la expresión del gen B sobre el gen A.

Cabe destacar que debemos realizar un proceso de integración a causa de la gran cantidad de variables instrumentales que poseemos en este estudio, correspondientes a las diversos cis-EQTLs que puede tener cada gen coexpresado incluido en el estudio. Cada SNP generará un estimador del efecto causal distinto, para llegar a un estimador único el método más apropiado, siempre que el número de EQTLs a integrar sea al menos tres, es el de los efectos aleatorios [8]. Donde cada uno de los estimadores asociados a cada SNP es una observación que viene de una distribución normal centrada en el valor "real" del estimador del efecto A→ B, donde los valores con más varianza se pesarán menos que los valores cercanos. Acción

implementada en el paquete *MendelianRandomization*. Corresponde a la realización de una regresión pesada que sigue el siguiente modelo:

$$\hat{\beta}_{Yj} = \theta\hat{\beta}_{Xj} + \epsilon_j \; ; \; \epsilon_j \sim N(0, \phi^2 se(\hat{\beta}_{Yj})^2)$$

Donde hemos añadido un término adicional ϕ que puede ser estimado, siendo éste el error estándar residual en el modelo de regresión, que representa la sobredispersión de la varianza de la estimación de causalidad. Pudiendo sacar de este modelo el *pvalue* de la estimación final.

El resultado final de este estudio, consistirá en dos estimadores de causalidad de A→ B y viceversa, con una probabilidad asociada a cada uno p1 y p2. El análisis computacional de esas dos probabilidades nos debería proporcionar la información necesaria para asegurar la dirección de la causalidad y la fiabilidad de la medida.

5. Desarrollo

Como ya hemos adelantado, en este TFG pretendemos caracterizar una red de coexpresión genética analizando datos de RNA-seq, para después utilizar la randomización mendeliana para inferir direccionalidad causal en tantos pares de genes coexpresados como sea posible. Para ello, nos serviremos de datos genotípicos obtenidos para los mismos individuos tal como se detalla en el anexo. Las muestras de RNA-seq proceden de macrófagos humanos, extraídos de 90 donantes voluntarios en el curso de un experimento en el que participan los tutores de este TFG. En el anexo a esta memoria se describe en detalle las características técnicas del dataset utilizado como ejemplo en este TFG. Hemos dividido nuestro código en 4 partes diferenciadas.

5.1. Filtrado de genes

En este apartado, se procede a ordenar y filtrar los datos, de manera que podamos contar con los más relevantes.

En primer lugar, nos deshacemos de aquellos genes que no codifican proteínas y aquellos autosómicos (genes que no se localizan en los cromosomas sexuales), información presente en la base de datos utilizada.

Utilizando la mediana de expresión de todos los individuos de cada gen. Seleccionamos aquellos que estén más expresados calculando para cada combinación gen-muestra, $log2((Reads + 0{,}5))$ y quedándonos con aquellos genes cuya mediana a través de los 90 individuos sea mayor que 6. La escala usada para la implementación de este criterio es la misma que *voom* utiliza como variable independiente para la estimación de la relación de media y varianza. Al final, obtenemos dos tablas una llamada "genes" con los datos de la posición y nombre de los genes seleccionados. Y otra llamada "reads" con la expresión de cada gen seleccionado registrada en cada individuo.

5.2. Pre-tratamiento de los datos de expresión

En esta parte del código, construiremos las matrices de coexpresión y analizaremos las relaciones entre pares de genes.

Estabilización de la varianza

Para esta parte, a la que hacemos referencia en el apartado 4.3.1 comenzamos normalizando las muestras usando TMM, y transformando los niveles de expresión a escala $log_2(CPM)$ a través de *voom*. La razón

del uso de estos métodos, y la descripción de la transformación de los datos que estos introducen están descritas en detalle en el anexo. Una vez hemos hecho esto, definimos el diseño experimental asociado al modelo lineal que usaremos para estimar los efectos de lote (matriz D en pag. 21), y resolvemos el modelo asociado usando $lmFit$, estimando las diferencias de expresión entre lotes (matriz B), y los residuales, que representan la parte de la expresión no afectada por los lotes, y que es aquella que nos interesa más a nosotros. Más adelante, realizamos el plot de la varianza, calculando correctamente las magnitudes de los dos ejes $x = \langle \log_2(r_{ij} + 0,5) \rangle$ y $y = \sqrt{\sigma_i}$, donde σ_i corresponde con la desviación típica de los residuales de cada gen. Una vez comprobado, que existe una relación sistémica al visualizar la curva, pasamos a escalar los residuos. Una vez escalados, volvemos a realizar la representación para visualizar la correcta estabilización de la varianza.

Correlación Parcial

En esta parte se aplica la corrección mencionada en el apartado 4.3.2. Se comienza calculando la matriz de correlaciones con el método de Pearson. se hace el test estadístico descrito en el apartado mencionado y se saca la matriz de los pvalues. Esta matriz que obtenemos es una matriz simétrica, por lo que utilizamos una función llamada *uperTriangle* que nos permite, convertir el triángulo superior de una matriz en un vector de manera ultra eficiente computacionalmente. Ese sería nuestro vector de pvalues. Después, aplicamos la corrección de Benjamini-Hochberg a este vector.

Una vez filtrados los datos gracias a los procesos anteriores. Recopilamos los pares de genes involucrados en las interacciones, los tests estadístidos de r y t junto con sus *pvalue* ordenados de menor a mayor. Todo ello, lo guardaremos en un *output* llamado *network*

5.3. Filtrado de SNPs

En esta parte realizaremos un filtrado de genotipos al igual que hicimos anteriormente con los genes. Lo haremos en en función del valor del Minor allele frequency (MAF) siendo esta la frecuencia del alelo menos común en la población, en una determinada localización del cromosoma. Seleccionando aquellas SNPs con $MAF < 0{,}05$.

Además, eliminaremos aquellos SNPs que estén en perfecto Linkage Desequilibrium (LD) con algún otro. Se dice que dos SNPs están en Linkage Desequilibrium cuando la frecuencia de asociación de sus diferentes alelos es mayor o menor de lo que se esperaría si los loci fueran independientes y se asociaran aleatoriamente. Es decir el valor de su correlación toma el valor 1.

5.4. Identificación de EQTLs

Para el análisis de estas matrices, necesitamos un formato muy específico de los datos por lo que reordenamos los datos necesarios, incluyendo la selección de los genes que participan en las correlaciones filtradas en el script anterior.

Se necesita que todos los datos que vamos a utilizar para las Matrices EQTL estén almacenados en archivos de texto. Por ello, guardaremos los datos de expresión, genes que intervienen, SNPs y genotipos en un archivo de texto en nuestra máquina.

Más adelante, ejecutamos la función de $Matrix\ eQTL$ para construir la matriz. Aunque antes se debe precisar características muy específicas de cómo se va a llevar a cabo esa construcción. Especificando cómo es el formato de nuestros datos guardados y cómo queremos que los lea (en nuestro caso de 2.000 en 2.000). Además, establecemos que se haga el análisis para las parejas de gen-SNP de regulación -cis,

obteniendo cis-EQTLs que nos servirán para la variable instrumental en la Randomización Mendeliana. Una vez identificadas dichas variables instrumentales, es importante considerar que necesitaremos estimar también los efectos de los mismos SNPs sobre aquellos genes que aparecen coexpresados con sus dianas. Es decir, si, para un gen A que aparece significativamente coexpresado con otros dos en el dataset: (gen B y gen C), encontramos dos cis-EQTLs, (SNP_{A1} y SNP_{A2}), necesitaremos, para tener todos los ingredientes necesarios para la randomización Mendeliana, saber cuál es el efecto que esos dos SNP_{A1} y SNP_{A2} ejercen sobre los niveles de expresión de los genes B y C.

Teniendo en cuenta lo anterior, es necesario una segunda ejecución de matrixEQTL, focalizada en estimar dichos efectos conjugados. Sobre dicha ejecución, dos consideraciones son esenciales: primera: los tests a ejecutar son tests en trans, no en cis, pues los genes B y C pueden incluso estar en cromosomas diferentes al gen A y sus SNPs. Por esto, es muy importante filtrar solo los SNPs y los genes estrictamente necesarios (pues estimar todos los posibles efectos en trans es muy costoso en memoria y tiempo de computación). La segunda consideración es que, aunque los resultados de dichos tests conjugados no sean significativos, son igualmente necesarios, por lo que no se puede poner un umbral de pvalor para el resgistro de resultados aquí.

5.5. Realización de tests de coexpresión

Antes de utilizar la función *Matrix EQTL* debemos filtrar nuestra matriz de coexpresión, para sólo llevar a cabo los tests que vayamos después a poder utilizar para implementar la randomización mendeliana. Con ese fin, filtramos las parejas de genes con un False Discovery Rate menor del 5 % ejecutando $FDR < 0{,}05$.

5.6. Randomización Mendeliana

Antes de correr la función de la Randomización Mendeliana, en primer lugar, sometemos a las parejas de genes a un test de vecindad y eliminaremos aquellos genes que no estén suficientemente alejados. Esta es una de las partes más importantes de nuestro código, ya que de esta manera nos aseguramos que las SNPs que afectan al gen A no afectan al gen B, corrigiendo el error de pleiotropía. Posteriormente, de entre esos pares significativos, seleccionaremos aquellos para los cuales contemos con al menos un cis-EQTL para cada gen del par. Una vez filtradas las parejas de genes, debemos seleccionar las cis-EQTLs más relevantes, al igual que en el apartado de la correlación, las sometemos al test de Benjamini-Hochberg, quedándonos con aquellos que son estadísticamente significativos: $FDR < 0{,}05$. Una vez seleccionados los datos de interés, debemos ordenarlos de tal forma que seamos capaces de seleccionar como variables instrumentales las cis-EQTL y el trans-EQTL como el outcome, con sus correspondientes estadísticos. Una vez organizada toda la información, podemos correr la función *mr_input* de la cual obtendremos los p-values correspondientes con las relaciones de causalidad. El esquema, pues, es el siguiente:

1. Seleccionamos cada par de genes significativamente expresados, uno por uno. Llamemos a estos dos genes gen A y gen B.

2. Para cada par, empezamos interrogando la presencia de un efecto causal de A hacia B. Para ello, los cis-EQTL del gen A son las variables instrumentales. Para realizar el test de la randomización en esta dirección, también necesitamos conocer los efectos de esos mismos SNPs sobre el gen B.

3. Con esos ingredientes, ejecutamos la randomización mendeliana en dirección A→B; y repetimos, en sentido contrario, los pasos 2 y 3 para la dirección B→A.

4. Repetimos para todos los pares coexpresados. Finalmente, tomamos los vectores de p-values de la randomización en ambos sentidos, y aplicamos Benjamini-Hochberg.

6. Resultados

6.1. Filtrado de datos

En este diagrama hemos querido resumir el proceso de filtrado de datos, donde seleccionamos aquellos necesarios para nuestro estudio. Podemos obtener la información del proceso seguido y de los datos finales directamente y de forma visual.

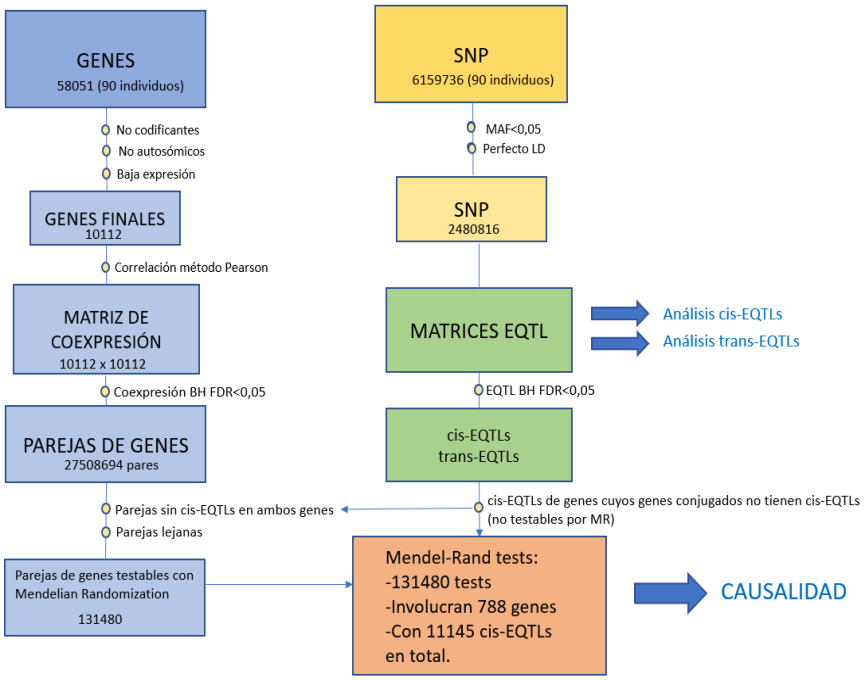

Figura 10: Diagrama de filtrado de datos.

6.2. Coexpresión de parejas de genes

El primer test a realizar fue el de coexpresión. En este test, de un total de 10112*10111/2 test, obtuvimos correlaciones significativas en 27508694, más de un 50 % d del total de pares. Para desarrollar una intuición visual de la potencia de estas asociaciones, es interesante visualizar los niveles de coexpresión de distintos pares en función de su nivel de significancia estadística (FDR)

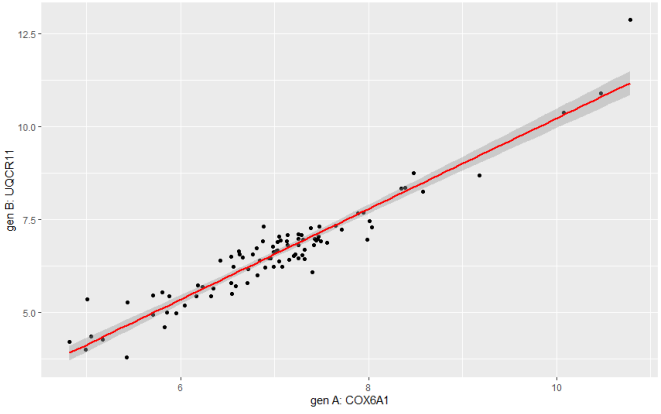

Figura 11: Coexpresión FDR orden de 10^{-38}

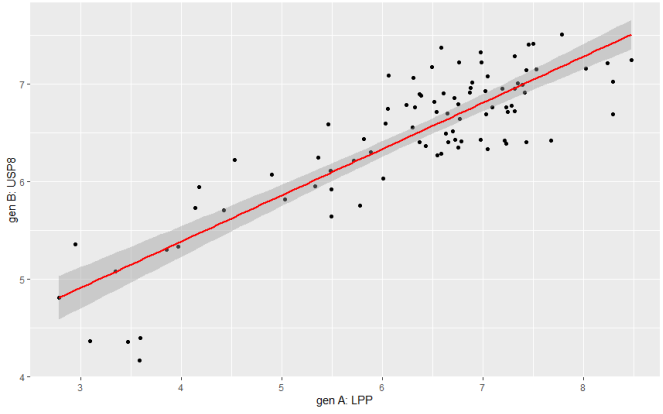

Figura 12: Coexpresión FDR orden de 10^{-20}

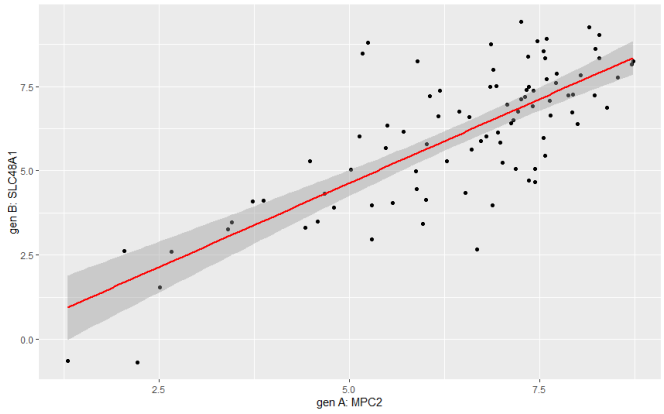

Figura 13: Coexpresión FDR orden de 10^{-15}

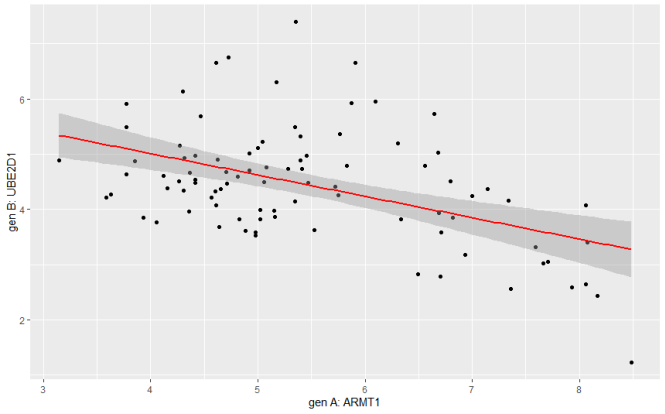

Figura 14: Coexpresión FDR orden de 10^{-5}

6.3. Correlaciones

Para visualizar las correlaciones y sus cambios representamos $0{,}5xln((1+r)/(1-r))$ siendo r el coeficiente de correlación de Pearson vs $-log10(FDR)$. Esto nos proporciona un plot del tipo de los *Volcano plot* siendo estos muy comunes en experimentos ómicos como la genómica, donde se representa el tamaño del estadístico bajo estudio frente a su significancia estadística. En estos plots a menudo se tiene una lista de miles de puntos de datos replicados entre dos condiciones y se desea identificar rápidamente los cambios más significativos, lo que permite una identificación visual rápida de aquellos puntos de datos (genes, en nuestro caso) que muestran cambios de gran magnitud (de correlación positiva o negativa).

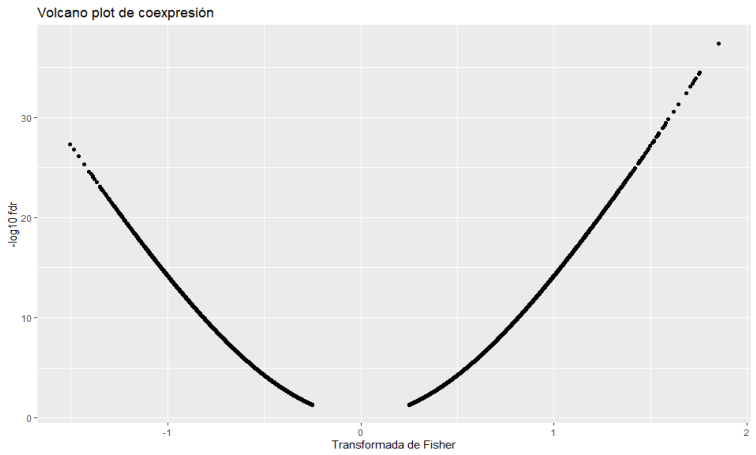

Figura 15: Volcano plot de cambio de correlación.

6.4. Identificación de EQTLs

El siguiente test a visualizar es el mapado de EQTLs. En nuestros análisis encontramos 14409 cis-EQTLs significativos, a un FDR del 5 %. En las *Figuras 16-17* hemos representado dos ejemplos, caracterizados por una elevada MAF que permite una estimación robusta de los efectos. Para ello, realizamos los llamados *Violin plots* donde representaremos el nivel de expresión del gen frente al genotipo de su SNP asociada.

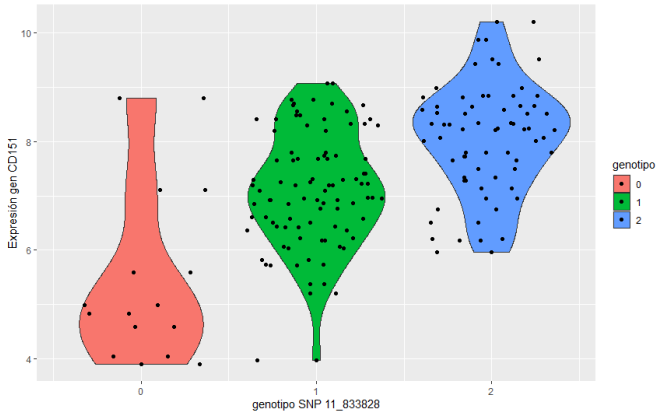

Figura 16: Efecto EQTL positivo

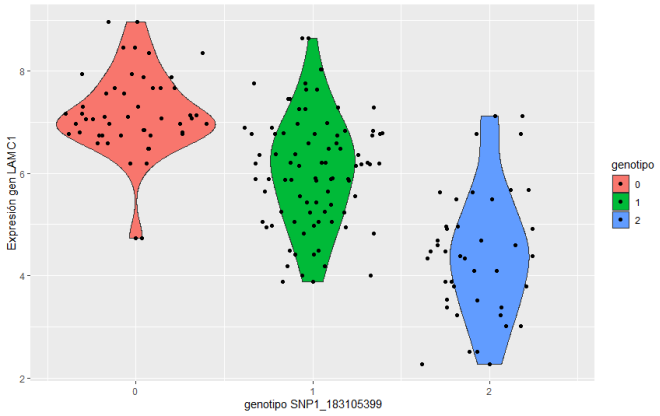

Figura 17: Efecto EQTL negativo

Hemos seleccionado dos pares de SNP-gen con efectos de EQTL contrarios. Para la *Figura 16* seleccionamos un efecto positivo, con un FDR de $1,2 \cdot 10^{-5}$ donde podemos observar en la figura que para el genotipo tipo 2 (homocigoto del alelo menor) la expresión del gen toma valores elevados y para el genotipo 0 (homocigoto del alelo mayor), además de ser menos frecuente, el valor de expresión del gen es reducido. En cambio en la *Figura 17*, donde hemos escogido una pareja gen-SNP con un FDR asociado de $2,9 \cdot 10^{-9}$ ocurre el efecto contrario.

Una vez hemos identificado los EQTLs, en nuestro dataset, filtramos los pares de genes coexpresados en los cuales ambos genes involucrados presentan al menos un cis-EQTL; pues estos serán los únicos para los cuales podremos utilizar la randomización mendeliana. Ello, combinado con el último requerimento, que excluye parejas de genes que se encuentran demasiado cercanas, establece el número total de parejas a testear, 131480.

6.5. Randomización Mendeliana

Después de la implementación de la función de Randomización Mendeliana para el estudio de causalidad, analizaremos los datos obtenidos y desarrollaremos las conclusiones derivadas de estos.

En primer lugar, tras ejecutar la randomización mendeliana para cada test, obtenemos un p valor y un false discovery rate asociado a la plausibilidad estadística de la existencia de una relación causal de un gen al otro del par y/o viceversa. Por tanto, tras uno de estos tests bi-direccionales, podemos clasificar los pares de acuerdo a cuál es el modelo causal más plausible:

- Relación causal unidireccional ($FDR < 0{,}05$ para una dirección, y $> 0{,}05$ para la otra, cuelsquiera estas sean).
- Relación causal mutua (o feedback): $FDR < 0{,}05$ en ambas direcciones.
- Relación no causal (confounding) $FDR > 0{,}05$.

Representamos el número de casos obtenidos para cada opción de direccionalidad (*Figura 18*). Como se puede apreciar en la figura, obtenemos un mayor porcentaje para los casos en el que no tenemos causalidad, representados por la sección azul. Seguido de la sección verde, donde se ven representadas las parejas

de genes en las que hemos podido identificar una dirección de causalidad única, o bien del gen A→gen B, o bien del gen A←gen B. Podemos observar también, un porcentaje más pequeño de una causalidad bidireccional, representada en tono salmón. Cabe destacar que obtenemos también un porcentaje de un 0,2 % casi imperceptible en la grádica donde obtenemos error en el cálculo del pvalue correspondiente. Hemos identificado a que se debe a un incorrecto filtrado de SNPs. Siendo este error no significativo, podría ser una mejora a tomar en cuenta en vista de estudios posteriores sobre esta materia. El error se debe al filtrado escueto de SNPs en relación con el Perfect Linkage Desequilibrium, ya que retiramos aquellas SNPs con una correlación perfecta de 1, no obstante, deberíamos haber elimano aquellas con una correlación perfecta de -1. Esto hace que en 268 de los más de 131000 pares, la presencia de pares de SNPs perfectamente decorrelacionados nos haya impedido estimar la causalidad en estos casos. En a figura 25 vemos que su presencia es del todo marginal en los datos, mientras que en la 26, los hemos sustraído del dataset para renormalizar la frecuencia de enlaces de cada tipo que hemos podido detectar.

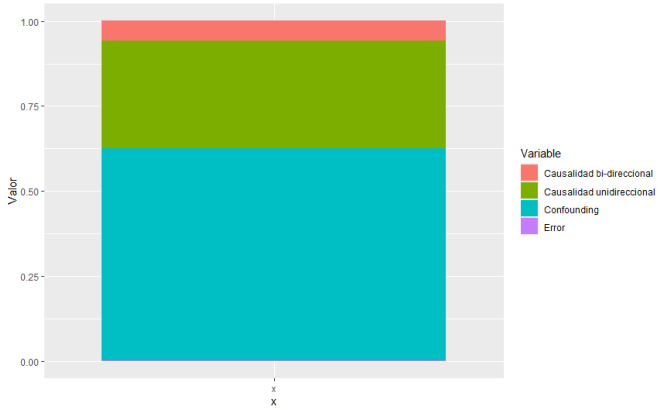

Figura 18: Fracción de enlaces testados.

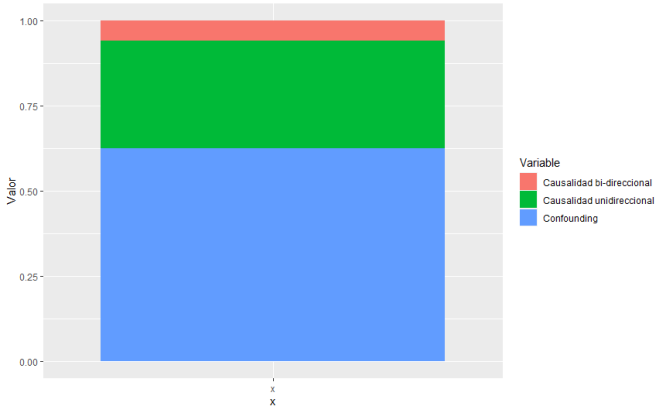

Figura 19: Fracción de enlaces testados.

6.5.1. Análisis de datos

La intencionalidad de esta parte del estudio, consiste en entender el resultado de nuestros datos y reflexionar sobre las conclusiones que podemos derivar de ellos. Con ese fin, representamos $-\log_{10}(FDR_{fw})$ donde FDR_{fw} corresponde con el FDR de la direccionalidad gen A→gen B a la que hemos denominado $forward$ frente a $-\log_{10}(FDR_{Bw})$, correspondiente a la direccionalidad inversa ($backward$). Obteniendo la gráfica correspondiente a la (*Figura 20*).

Podemos observar en la figura que obtenemos valores de FDR para una direccionalidad extremadamente pequeños (representados en la gráfica por valores extremadamente grandes en escala logarítmica) y valores elevados de FDR de su direccionalidad contraria conjuntamente para un número muy elevado de casos. La interpretación más evidente es que el método tiende a asignar direccionalidad a los enlaces con gran significancia en uno, u en otro sentido, pero no en ambos a la vez. Si bien resulta tentador interpretar este resultado en clave biológica, tenemos primero que fijarnos en las características de estos casos, donde la FDR es marcadamente menor en una dirección que en la otra.

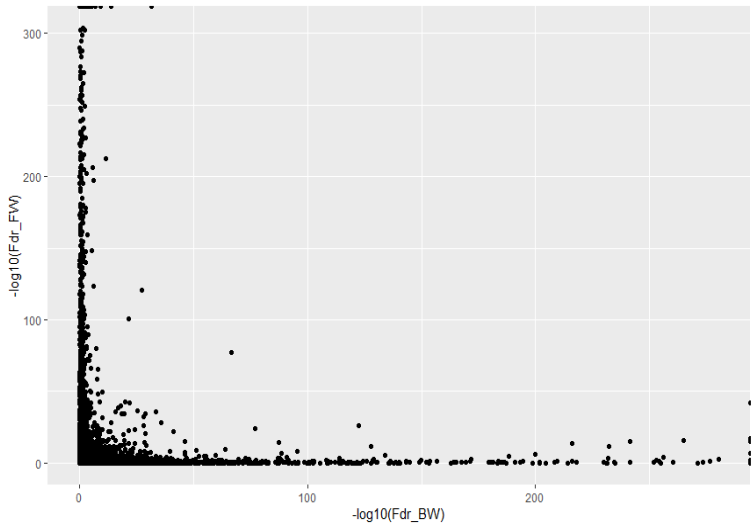

Figura 20: $-\log_{10}(FDR_{fw})$ vs $-\log_{10}(FDR_{Bw})$

6.5.2. Volcano Plot

Recurrimos al Volcano Plot, donde estudiaremos la naturaleza de nuestros datos y podremos obtener la dispersión de nuestras direccionalidades, en este caso la de la dirección *forward*. Representamos el anterior utilizado $-Log_{10}(FDR_{fw})$ frente al estadístico θ, donde hemos coloreado el plot, en función de las SNPs contribuyentes en la obtención de la causalidad (*Figura 21*).

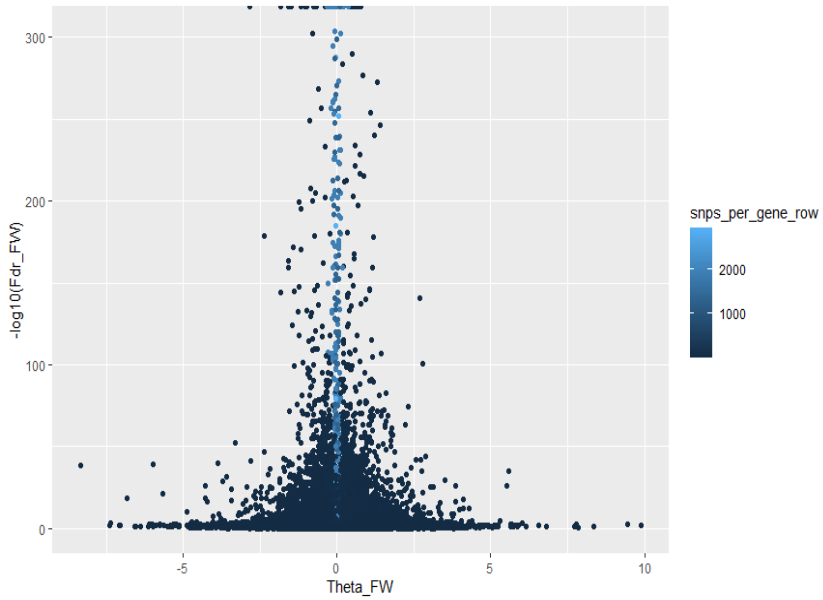

Figura 21: Volcano plot para la visualización

Podemos apreciar una clara relación, entre el número de SNPs involucrados en la primer gen de la pareja, es decir, el número de cis-EQTL que utilizamos como variables instrumentales para cada par de genes, el valor del estadístico θ y su FDR. Específicamente, vemos que, cuando el número de SNPs disponibles para estimar el efecto causal es elevado, 1)el estimador causal tiende a mantenerse pequeño en valor absoluto, y 2) su FDR puede llegar a dispararse a valores ultra-significativos, siendo prácticamente todos los casos asociados a una significancia extrema, asociados a casos donde tenemos muchos SNPs disponibles. Esto nos lleva a deducir que no podemos basarnos en el criterio que hemos seleccionado para evaluar la

direccionalidad, ya que el valor FDR se ve afectado por el número de variables instrumentales empleadas. Esto conlleva a una alteración en el resultado, obteniendo un FDR mucho menor para genes con mayor número de cis-EQTLs, no por un hecho biológico de direccionalidad, sino por el hecho de poseer esa cantidad de información adicional, que en cambio, su pareja de gen al poseer una menor cantidad de cis-EQTL, no tiene.

6.5.3. Densidad de cis-EQTL en dirección *forward* vs *backward*

En vista de lo anterior, procedemos a comprobar nuestra hipótesis, representando la distribución de variables instrumentales empleadas para nuestra randomización mendeliana para los casos donde hemos obtenido valores extremos de la FDR, concretamente representaremos aquellos resultados menores de 10^{-50} para la causalidad en las dos direcciones *forward* y *backward* (*Figuras 22-23*) y (*Figuras 24-25*) respectivamente.

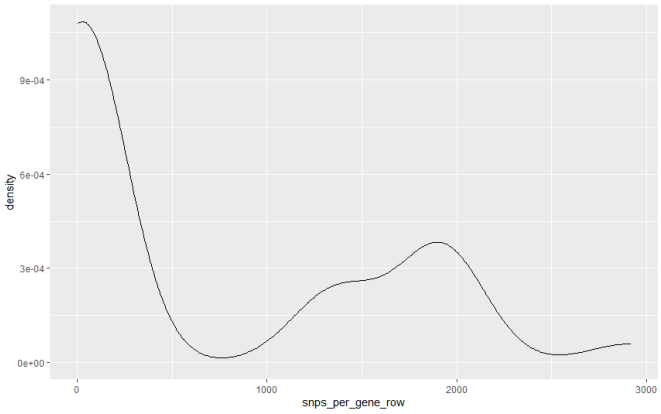

Figura 22: Distribuición de cis-EQTL del gen A *foward*

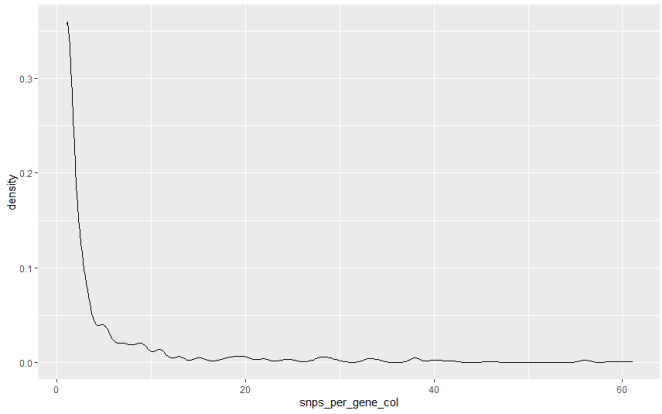

Figura 23: Distribuición de cis-EQTL del gen B *foward*

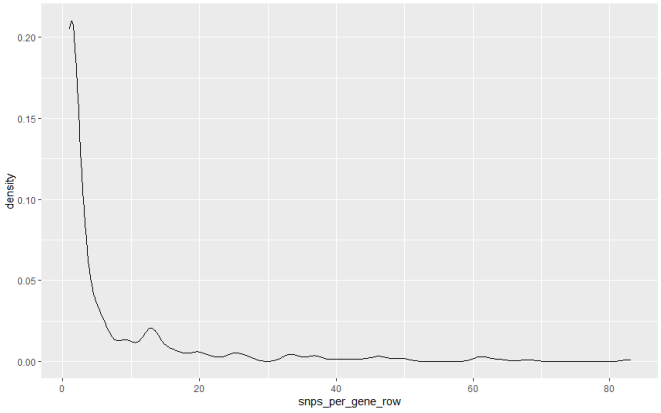

Figura 24: Densidad de cis-EQTL del gen A *backward*

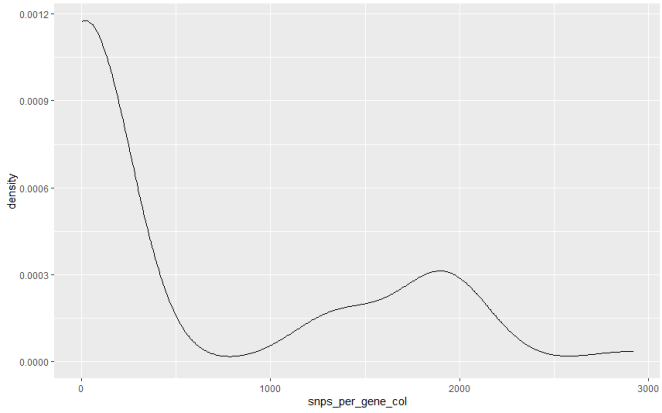

Figura 25: Densidad de trans-EQTL del gen B *backward*

Como se puede observar en nuestras figuras, podemos asegurar que existe una relación entre el número de variables elegidas como variables instrumentales en nuestra Randomización Mendeliana y el nivel de significancia estadística con el que podemos determinar la presencia de un flujo de causalidad. En las dos direccionalidades estudiadas, donde hemos identificado una direccionalidad *forward* con un $FDR < 10^{-50}$ obtenemos un mayor número de pares de genes con un número elevado de cis-EQTL en *forward* en comparación con el número de cis-EQTLs con el que contamos para completar el test en la dirección opuesta. Mientras que en los pares de genes dónde se ha identificado una direccionalidad *backward*, obtenemos el

caso simétricamente contrario, obtenemos un $FDR < 10^{-50}$ donde utilizamos un número también elevado de cis-EQTL del segundo gen del par mucho mayor que del primero.

6.5.4. Corrección de Volcano Plot

Una vez identificado el problema de nuestro método utilizado, si el problema es la presencia de pares de coexpresión con un número de cis-EQTLs disponibles para cada uno de los genes extremadamente diferente, podemos preguntarnos cómo cambiarían nuestros resultados, si descartásemos esos casos. Esto podría conseguirse, al menos provisionalmente, imponiendo un rango estrecho en el número de SNPs por gen, para considerar un par de genes coexpresados. Así, si restringimos el análisis a pares de genes en los que ambos genes cuentes con entre 1 y 5 cis-EQTLs, nos aseguramos de contar con una cantidad de información más homogénea de información para las dos direccionalidades. Obteniendo dos Volcano Plot (*Figuras 26-27*) cada uno para una direccionalidad, donde hemos representado al igual que en el primer Volcano Plot donde detectamos el problema $-Log_{10}(FDR)$ frente al estadístico θ.

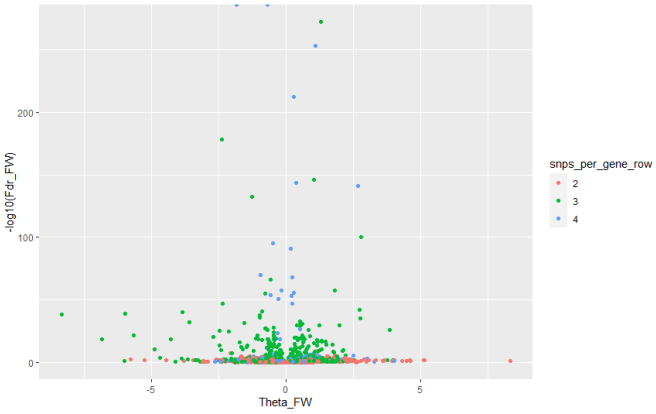

Figura 26: Volcano Plot *foward*

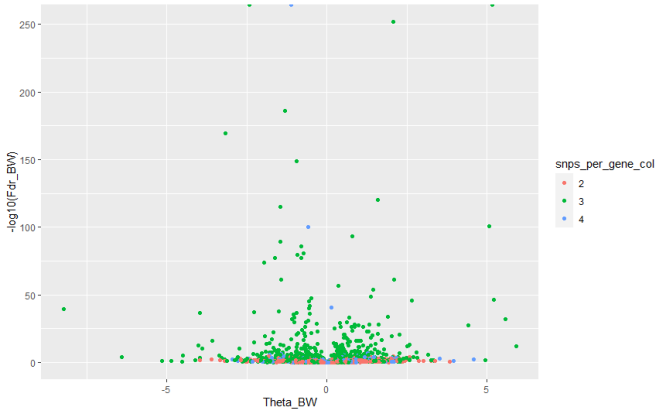

Figura 27: Volcano Plot *backward*

Es notoria la diferencia con respecto a la primera representación realizada. Estas gráficas se ajustan de una manera más reconocible a la forma habitual de este tipo de representaciones. Podemos ver una franja diferenciada en torno a los menores valores del estadístico θ, en valor absoluto, al igual que la desaparición de valores desorbitados de FDR.

7. Conclusiones y trabajo futuro

7.1. Conclusiones

Uno de los objetivos principales del estudio de la evolución biológica consiste en analizar los cambios genómicos, tanto en contexto de su caracterización, como en su relación con la expresión génica para el desarrollo de distintas funciones biológicas. En este estudio, hemos descrito y aplicado una serie de metodologías para la caracterización de estas relaciones. En concreto, utilizamos la técnica estadística llamada Randomización Mendeliana para su análisis, utilizando para su implementación las matrices EQTL como variables instrumentales y la causalidad de correlación de expresión de pares de genes como motivo de estudio.

Como resultado, hemos comprobado que nuestro método es factible, y que, con menos de cien individuos, permite la estimación de relaciones causales significativas en miles de pares de coexpresión. No obstante, se requiere de más trabajo, puesto que discernir la direccionalidad de un enlace requiere comparar modelos, y hemos visto que la cantidad de SNPs disponibles para los modelos $forward$ y $backward$, si es muy

distinta, puede sesgar la significancia de la estimación de la direccionalidad de los enlaces de manera sustancial.
En resumen, la metodología propuesta es prometedora, y para refinarla hasta convertirla en una herramienta efectiva para la inferencia de redes de regulación causal se requiere más investigación.

7.2. Trabajo Futuro

Para el correcto discernimiento de direccionalidad, proponemos una alternativa al desarrollo del método empleado en este trabajo. Donde en los genes influenciados por un elevado número de SNPs, de los que obtendríamos un elevado número de cis-EQTLs, se aplicase una función capaz de elegir aleatoriamente un número de cis-EQTLS similar en ambas direcciones. Esas cis-EQTLs aleatorias corresponderían con las variables instrumentales de nuestra Randomización Mendeliana, de la cual, obtendríamos un valor estadístico de su causalidad libre del problema experimentado en este trabajo. Esta metodología debería repetirse varias veces, para elecciones de cis-EQTLs diferentes y debería ir acompañada de un último estudio estadístico para la interpretación de las diferentes causalidades obtenidas para cada par de genes. Otra alternativa consistiría en la selección de los N SNPs asociados más significativamente con cada gen en el dataset, sin importar su nivel de significancia estadística. En esta segunda opción, debería tenerse en cuenta que ello puede llevar a testar SNPs cuya asociación en cis con sus respectivas dianas sea muy distinta en intensidad, y ello pueda, quizás, llevar a problemas similares a los aquí detectados.

Otro aspecto que proponemos a desarrollar en propósito de un modelo más completo, es aquel en referencia a la temporalidad de las muestras. Integrando la información que se obtendría a partir del método mencionado anteriormente, junto con datos sobre series temporales de las muestras, nos abriría la posibilidad de elaborar una red de causalidad de manera mas completa y dinámica. Estas redes se pueden utilizar para identificar las redes causales a través de las cuales un tratamiento externo (i.e. un proceso de infección, o un gradiente externo que atraviese a las muestra [22] [23]) afecta a una fracción del genoma. En resumen, el estudio y caracterización de estas redes, nos ayuda a identificar genes que pueden causar, al menos de forma parcial, distintas enfermedades, un propósito a largo plazo, sería tener la posibilidad de prevenirlas.

Referencias

[1] Getting started with Matrix eQTL. (2020). Consultado 1 Junio 2020, de http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/runit.html

[2] Andrey A. Shabalin, Matrix eQTL: ultra fast eQTL analysis via large matrix operations, *Bioinformatics*, Volumen 28, Tema 10, 15 Mayo 2012, Páginas 1353–1358, https://doi.org/10.1093/bioinformatics/bts163

[3] Mendelian Randomisation. (2020). Consultado 1 Junio 2020, from https://es.slideshare.net/jamesmcm03/mendelian-randomisation

[4] Yavorska, O., y Burgess, S. (2017). MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *International Journal Of Epidemiology, 46*(6), 1734-1739. doi: 10.1093/ije/dyx034

[5] Durbin, B., Hardin, J., Hawkins, D., y Rocke, D. (2002). A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics, 18*(Suppl 1), S105-S110. doi: 10.1093/bioinformatics/18.suppl_1.s105

[6] 6.3 - Testing for Partial Correlation — STAT 505. (2020). Retrieved 13 Junio 2020, from https://online.stat.psu.edu/stat505/lesson/6/6.3

[7] Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1), 289-300

[8] Burgess, S., Foley, C., Allara, E., Staley, J., y Howson, J. (2020). A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nature Communications, 11*(1). doi: 10.1038/s41467-019-14156-4

[9] Burgess, S., Small, D., y Thompson, S. (2015). A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods In Medical Research, 26*(5), 2333-2355. doi: 10.1177/0962280215597579

[10] Banf, M., y Rhee, S. Y.(2017). Computational inference of gene regulatory networks:approaches, limitations and opportunities. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 1860*(1), 41-52.

[11] Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E., y **Guthke, R.** (2009). Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems, 96*(1),86-103

[12] Altay, G.,y Emmert-Streib, F.(2010). Inferring the conservative causal core of gene regulatory networks.*BMC systems biology,4*(1),132.

[13] Chen, X., Li, M., Zheng, R., Zhao, S., Wu, F. X., Li, Y., y Wang, J.(2019). A novel method of gene regulatory network structure inference from gene knock-out expression data. *Tsinghua Science and Technology,24*(4), 446-455

[14] Tung, T. Q., Ryu, T., Lee, K. H., y Lee, D.(2007, Junio). Inferring gene regulatory networks from microarray time series data using transfer entropy. In Twentieth IEEE International. *Symposium on Computer-Based Medical Systems(CBMS'07)*(pp. 383-388). IEEE

[15] Chen, L. S.(2012). Using eQTLs to reconstruct gene regulatory networks. In Quantitative Trait Loci (QTL) (pp. 175-189). *Humana Press*

[16] Peñagaricano, F., Valente, B. D., Steibel, J. P., Bates, R. O., Ernst, C. W., Khatib, H., y Rosa, G. J. (2015). Exploring causal networks underlying fat deposition and muscularity in pigs through the integration of phenotypic, genotypic and transcriptomic data.*BMC systems biology,9*(1),58

[17] Zhong, W., Dong, L., Poston, T. B., Darville, T., Spracklen, C. N., Wu, D., y Zheng, X.(2020). Inferring regulatory networks from mixed observational data using directed acyclic graphs.*Frontiers in Genetics,11*(8)

[18] Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., y Irizarry, R. A.(2010). Tackling the widespread and critical impact of batch effects in high-throughput data.*Nature Reviews Genetics, 11*(10), 733-739

[19] Katan MB. Apolipoprotein E isoforms, serum cholesterol, and cancer. Lancet 1986;i:507–08.

[20] Acerca Del Proyecto Internacional Hapmap. (2020). Consultado el 10 Julio 2020, de https://www.genome.gov/acerca-del-proyecto-internacional-hapmap

[21] Peripheral blood mononuclear cell. (2020). Consultado el 4 Julio 2020, de https://en.wikipedia.org/wiki/Peripheral_blood_mononuclear_cell

[22] Nédélec, Y., Sanz, J., Baharian, G., Szpiech, Z. A., Pacis, A., Dumaine, A., y Sabourin, A. P.(2016). Genetic ancestry and natural selection drive population differences in immune responses to pathogens.*Cell,167*(3), 657-669

[23] Snyder-Mackler, N., Sanz, J., Kohn, J. N., Brinkworth, J. F., Morrow, S., Shaver, A. O., y Barreiro, L. B. (2016). Social status alters immune regulation and response to infection in macaques. *Science,354*(6315), 1041-1045

[24] WGCNA: R package for performing Weighted Gene Co-expression Network Analysis. (2020). Consultado 1 Junio 2020, de https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/

[25] Dündar, F., Skrabanek, L., y Zumbo,P.(2015-2019). Introduction to differential gene expression analysis using RNA-seq. *Applied Bioinformatics Core—Weill Cornell Medical College*,1-97.