UNIVERSIDAD DE CANTABRIA

PROGRAMA DE DOCTORADO EN
CIENCIA Y TECNOLOGÍA

TESIS DOCTORAL

Gestión de volúmenes masivos de datos genéticos y análisis de la influencia de su interacción en el desarrollo de cáncer

Realizada por

Camilo Palazuelos Calderón

Dirigida por

Javier Llorca Díaz

Marta Zorrilla Pantaleón

Escuela de Doctorado de la Universidad de Cantabria

Santander, 2019

Agradecimientos

Una tesis doctoral está hecha de tiempo, de ese que uno cree robarles a los que se lo regalan, de ese que uno recobra cuando parece haberlo perdido. Gracias al Gobierno de Cantabria, por la financiación recibida a través del Programa de Personal Investigador en Formación Predoctoral de la Universidad de Cantabria 2014-18, y al estudio MCC-Spain, por la formación impartida y el acceso a los datos de genotipado; a mis directores, Javier Llorca y Marta Zorrilla, por su paciencia, por enseñarme a investigar; a mis mentoras en lo docente, Trinidad Dierssen e Inés Gómez, *impossibles à oublier*, por enseñarme a enseñar; a mis compañeras de despacho, Jéssica Alonso y María Fernández, por lo que hemos aprendido y aprenderemos juntos; y a mi familia y amigos, que no son sino la misma cosa: a Inta y Papa por tanto amor, a mi madre y mis hermanos, también a mi tío y mi primo, por tanto apoyo y tanta comprensión, a Jairo (y los hermanos Fu) por tanta complicidad y tanta ilusión, a Paola, nuestra ahijada, y sus padres, Anita y Jony, por tanta luz a su alrededor, a Sheila,

Lara y Sandra por tantos años (¡veinticinco o veintiséis!) de generosidad, a Brenda y Héctor por tanto cariño y a Patricia por tanta música. A todos, gracias por el tiempo regalado y por el tiempo recobrado.

Resumen

La llegada de los estudios de asociación del genoma completo revolucionó el campo de la genética a partir de la primera década del siglo XXI: de analizar decenas de variantes genéticas, se pasó a analizar millones. Con todo, lo habitual en los resultados arrojados es que, para enfermedades comunes, el efecto que tiene una variante sobre una característica física, un síntoma o una enfermedad sea pequeño y, aun sumando las contribuciones de todas las variantes, no se consigan explicar todos los casos de individuos en que aparece la característica en cuestión.

Los modelos o patrones en las asociaciones entre una variante genética y una enfermedad y entre una interacción de variantes genéticas y una enfermedad, a pesar de la información que proporcionan, se han ignorado en casi todos los estudios de asociación del genoma completo. En esta tesis doctoral, se propuso el estudio de estos modelos o patrones en interacciones de variantes genéticas para explicar parte de esa variación cuyo origen sigue sin desvelarse.

Aunque no todas las variantes genéticas, ni mucho menos todas sus interacciones, presentan un modelo o patrón en su relación con la enfermedad, nuestra hipótesis de partida era que no son tan reducidas en número como parece, por lo que su estudio podía dar lugar a la generación de hipótesis biológicas susceptibles de ser comprobadas experimentalmente. Para demostrar esta hipótesis, (i) se desarrolló un marco de trabajo que permitió evaluar y comparar los niveles de adecuación e incertidumbre de distintos patrones a volúmenes masivos de variables que analizar simultáneamente; (ii) se diseñó e implementó una prueba estadística en el marco de trabajo anterior que permitió decidir qué modelo genético le correspondía —si es que le correspondía alguno— a variantes genéticas e interacciones de estas; y (iii) se confeccionó un protocolo de construcción de redes de interacciones con que se analizaron los datos del estudio MCC-Spain.

El método de análisis propuesto supone una contribución novedosa que permite generar hipótesis biológicas relacionadas con los cinco tipos de cáncer estudiados y su asociación con la variación genética y sus interacciones. En tres de estos cinco tipos de cáncer, se han encontrado asociaciones interacción-enfermedad que han podido refrendarse con descubrimientos científicos de los últimos 5 años, lo que pone de manifiesto tanto la viabilidad del método como su potencial para revelar información oculta en las redes de interacciones de variantes genéticas que conducen a la aparición de enfermedades comunes.

Siglario

ADN Ácido desoxirribonucleico

ARN Ácido ribonucleico

CDF Función de distribución acumulada

EHW Equilibrio de Hardy-Weinberg

GWAS Estudio de asociación del genoma completo

IC Intervalo de confianza

MAF Frecuencia del alelo menor

OR *Odds ratio*

RDC Región de diferencia coherente

REC Región de equivalencia coherente

RR Riesgo relativo

SE Error estándar

SNP Polimorfismo de un solo nucleótido

TD Test de diferencia

TDE Test de diferencia-equivalencia

TE Test de equivalencia

TOST *Two one-sided tests*

Índice general

Índice de figuras

Índice de tablas

Prefacio

Uno de los objetivos fundamentales en el campo de la genética es cuantificar la contribución de la variación en el genoma al desarrollo de rasgos observables, como características físicas, síntomas o enfermedades. El deseo de medir la relación entre ambos fenómenos nace ya en el siglo XIX con los trabajos de Mendel sobre la herencia genética de diferentes variedades de la planta del guisante. Estos trabajos darían lugar a las leyes que llevan su nombre, y que tratan de explicar *completamente* las características de un individuo en función de las de sus progenitores. El énfasis en «completamente» es crucial para entender la evolución de la genética en el último siglo: Mendel estudió características que se transmitían de generación en generación a través de mutaciones en un solo gen que, de encontrarse en un individuo, hacen que muestre la característica en cuestión. No todas las enfermedades, sin embargo, presentan este tipo de comportamiento, por lo que el estudio de aquellas más comunes se ha ido alejando del enfoque para las enfermedades mendelianas.

A raíz de sus observaciones, Mendel distinguió dos tipos de herencia: dominante y recesiva. En la fecundación, los genes del padre y de la madre se entrecruzan, de manera que, para cada gen, todo individuo tiene dos copias, que pueden ser o no iguales. Los patrones de herencia mendeliana se distinguen entre sí por la necesidad de presencia de mutación en ambas copias para el desarrollo de la característica: si el individuo solo la presenta si hay mutación en las dos copias del gen, se trata de una herencia recesiva; si la presenta siempre que haya, al menos, una mutación en una de las copias, dominante.

Con el desarrollo de la genética, los estudios de asociación entre una mutación y una característica o enfermedad pasaron de limitarse de una a varias mutaciones, hasta llegar a los estudios de genes candidatos, en que, bajo sospecha por hipótesis biológica, se obtenía la variación genética de una muestra y se determinaba qué mutaciones afectaban a la aparición de la característica estudiada. Diferentes mutaciones en el gen *CFTR*, por ejemplo, pueden causar fibrosis quística, conclusión a la que se llegó a partir de un análisis de ligamiento: se obtuvo la variación genética —entendida como los genes que se sospechaba que podían estar asociados con la enfermedad— de distintas familias con fibrosis quística y se estudió la distribución de las mutaciones heredadas y la aparición de la enfermedad. Esta técnica ha servido para identificar genes o mutaciones asociados con otras enfermedades relativamente raras, como la enfermedad de Huntington. Sin embargo, al aplicarla en

enfermedades comunes, no se obtienen resultados tan exitosos, debido a que, por tratarse de enfermedades más frecuentes, portar una mutación no implica la aparición de la enfermedad.

La llegada de los GWAS (sigla en inglés de «estudios de asociación del genoma completo») revolucionó el campo de la genética a partir de la primera década del siglo XXI. En estos, se obtiene la variación genética *completa* de una muestra de individuos sin partir de una hipótesis biológica previa, y se buscan mutaciones que estén asociadas con la característica o enfermedad en cuestión. Así, de analizar decenas de mutaciones, se pasó a analizar cientos de miles y aun millones. A pesar de que el genoma humano cuenta con más variación que esos cientos de miles o millones de unidades, la mayoría están relacionadas, es decir, determinadas unas en función de otras, al heredarse en bloque (el entrecruzamiento de los genes del padre y de la madre en la fecundación tiene lugares preferentes). La ventaja más clara que proporciona este hecho es que se necesita estudiar un número reducido de mutaciones —comparado con el de las que realmente existen— para cubrir todo el genoma, aunque las leyes de Mendel, que requieren de la independencia de los genes estudiados, dejan de poder aplicarse.

La ventaja anterior no solo tiene implicaciones computacionales, sino estadísticas: a medida que aumenta el número de hipótesis que someter a prueba, lo hace también la probabilidad de encontrar falsos positivos, es decir, resultados que parecen asociar una mutación con una enfermedad

aun cuando esta asociación no es real. Para controlar este fenómeno, se han establecido niveles de significación estadística estrictos, aunque no tanto —se divide el nivel de significación inicial por un millón, el número de mutaciones que se estima que son independientes entre sí— como si ninguna mutación del genoma estuviera relacionada. A pesar de que las leyes de Mendel no son aplicables para predecir el desarrollo o aparición de una característica común (enfermedades como el infarto o el cáncer), sí pueden relajarse para cuantificar el grado de influencia que una variante genética tiene sobre la enfermedad. De esta forma, los patrones de herencia dominante y recesivo pasarían a referirse a modelos que explican la relación entre los individuos que portan la mutación y aquellos que exhiben la característica o enfermedad.

Los resultados arrojados por los GWAS han sorprendido desde el principio: lo habitual es que, para enfermedades comunes, el efecto que tiene una mutación sobre una característica sea pequeño y, aun sumando las contribuciones de todas las mutaciones, no se consigan explicar todos los casos de individuos en que aparece la característica en cuestión. El estudio de mutaciones raras o de la agregación de los efectos de las mutaciones de un mismo gen ha dado resultados modestos en el avance de este campo de investigación. A finales de los años 2000, se propuso que el estudio de las interacciones de genes o de las mutaciones de estos podría explicar, al menos en parte, toda esa variación cuyo origen sigue sin desvelarse. Con todo, el principal problema al que se enfrenta el estudio

de interacciones es su elevadísimo número en potencia: considerando un millón de variantes para cubrir todo el genoma humano, habría $5 \cdot 10^{11}$ posibles interacciones de primer orden, lo que dificulta el problema desde los puntos de vista computacional y estadístico. Asimismo, el tamaño de las muestras para un estudio de interacciones debería ser mucho más elevado que el de las que se manejan hoy en día.

Así las cosas, se han ido desarrollando métodos que permiten filtrar aquellas interacciones con una probabilidad baja de influir en el desarrollo de la característica o enfermedad bajo estudio. Uno de los filtros que nunca se han considerado ha sido el de comprobar si tanto las mutaciones como sus interacciones siguen un patrón o modelo de herencia genético con respecto a la característica o enfermedad. Aunque no todas las mutaciones, ni mucho menos todas sus interacciones, presentan un modelo o patrón de herencia de los propuestos por Mendel o desde entonces, estos patrones son los que más información dan acerca de su relación con la enfermedad, por lo que se propone el estudio, en esta tesis doctoral, de aquellas mutaciones que, presentando un modelo genético, interactúan con otras a través de un patrón conocido, como el dominante o el recesivo. Nuestra hipótesis es que dichas mutaciones no son tan reducidas en número como parece, por lo que tanto su estudio como la interpretación de los resultados concernientes a su relación con la enfermedad pueden dar lugar a la generación de hipótesis biológicas susceptibles de ser comprobadas experimentalmente.

Con respecto a la gestión de volúmenes masivos de datos, imperan dos enfoques bien diferenciados hoy en día: el relacional y el no relacional. El primero está basado en la lógica de predicados y en la teoría de conjuntos, y es el más utilizado para administrar datos debido a la simplicidad, robustez, flexibilidad y rendimiento que proporciona. La mayoría de los sistemas de gestión de bases de datos comerciales son relacionales, y usan un lenguaje de programación estándar tanto para el diseño como para la administración de bases de datos. Por otra parte, las bases de datos no relacionales se han popularizado en los últimos años debido a la variedad de tipos de datos que admiten y, en particular, a su escalabilidad horizontal. La información almacenada en estas bases de datos no requiere de una estructura fija, como es el caso de las relacionales, y ofrecen un rendimiento superior al de estas al gestionar volúmenes masivos de datos, en detrimento, no obstante, de la integridad de la información. A pesar de que han comenzado a aparecer estudios que defienden el uso de sistemas no relacionales en el ámbito biomédico, tras la práctica totalidad de los paquetes de *software* de epidemiología genética subyace un sistema relacional con índices de mapas de bits, como PLINK, dadas las características de los datos que se manejan.

Capítulo

1 Estudios de asociación del genoma completo

En este capítulo, se pretende familiarizar al lector con el contexto en que se enmarca esta tesis doctoral, para lo cual se introducen tanto los conceptos clave de los estudios de asociación del genoma completo como la problemática metodológica a la que se enfrentan en la actualidad.

1.1 De Mendel a los GWAS

La herencia genética nos ha sido siempre familiar. El parecido entre padres e hijos, tanto en los seres humanos como en otros animales, suponía la constatación de la existencia de mecanismos de transmisión de rasgos biológicos de un ser vivo a sus descendientes, mecanismos que no serían abordados formalmente hasta mediados del siglo XIX por el monje agustino Gregor Mendel.[1] Entre 1857 y 1865, Mendel se dedicó al estudio de los guisantes que plantaba en el huerto de su monasterio, polinizando las plantas y analizando las características resultantes, como el color o la forma de las vainas (véase la FIGURA 1.1).

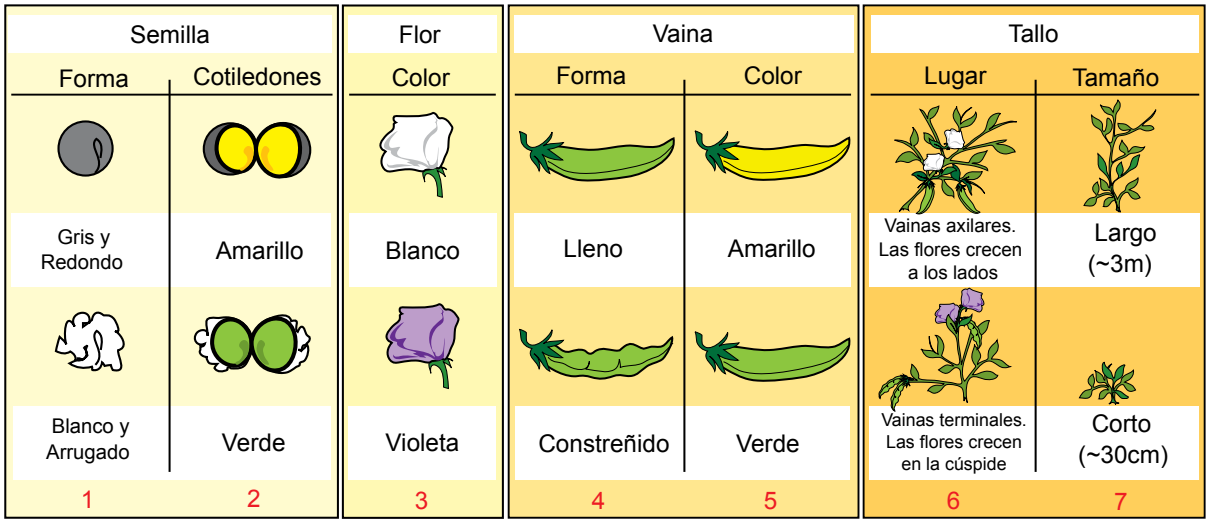

FIGURA 1.1. Características o rasgos biológicos de los guisantes analizados por Mendel en sus experimentos. FUENTE: Wikimedia Commons.

Las leyes de Mendel entonces

Cruzando variedades puras de la planta del guisante, Mendel observó que, en la primera generación, solo surgían descendientes con los rasgos de una de las variedades (por ejemplo, plantas de tallos largo y corto daban lugar a plantas de tallo largo). Sin embargo, cruzando las plantas de la primera generación entre sí, obtuvo las características de ambas variedades con una proporción de 3:1 (por ejemplo, con plantas de flores blancas y violeta, se produjeron 224 y 705, respectivamente).[2]

De estas observaciones, Mendel extrajo las siguientes conclusiones: las características o rasgos biológicos (*fenotipos*) de un organismo vienen dados por «factores» discretos, hoy denominados *genes*; se poseen dos

«versiones» de cada factor, hoy denominadas *alelos*; de cada factor, una de las versiones es *dominante* sobre la otra (*recesiva*), es decir, el organismo expresará su característica asociada si está presente; y las células sexuales contienen, de cada factor, solo una versión elegida al azar.[3]

De manera similar, estudió los patrones de herencia de dos rasgos biológicos al mismo tiempo (por ejemplo, la forma —redonda es la variedad dominante— y el color —en este caso, lo es el amarillo— de las semillas). En la primera generación, todos los guisantes fueron redondos y amarillos, mientras que, en la segunda, se obtuvieron las combinaciones redondo-amarillo (dominante-dominante), redondo-verde (dominante-recesivo), arrugado-amarillo (recesivo-dominante) y arrugado-verde (recesivo-recesivo) con una proporción 9:3:3:1.[4]

Hoy se sabe que cada característica de la planta del guisante estudiada por Mendel se asocia con un gen, es decir, se trataba de características monogénicas: solo variantes de un gen concreto dan lugar a variantes de la característica en cuestión. Asimismo, se sabe que cada uno de esos genes se encuentra en cromosomas distintos, por lo que, *a priori*, son independientes, es decir, no tienen por qué heredarse conjuntamente. Sin embargo, la mayoría de los rasgos biológicos o enfermedades que se investigan hoy en día son poligénicos, cuyos genes implicados a menudo se encuentran en el mismo cromosoma, lo que viola la independencia en la distribución de los alelos y obliga a redefinir la relación de dominancia entre los alelos de un gen.[5]

Las leyes de Mendel hoy

La alcaptonuria (trastorno metabólico congénito que provoca, entre otros síntomas, oscurecimiento de la orina) fue una de las primeras enfermedades a las que se les han atribuido causas de índole genética.[6] Aunque ya en 1908 Garrod había llegado incluso a anticipar su modelo de herencia,[7] no fue hasta los años 90 cuando se halló una mutación con efectos similares a los de la alcaptonuria en ratones,[8] lo que permitió acotar una región del cromosoma 3 humano[9,10] en que se acabaría aislando el gen cuyas variantes se asocian con el trastorno.[11]

Desde entonces, se han descrito miles de asociaciones genotipo-fenotipo, muchas de las cuales se corresponden con la relación entre mutaciones de un mismo gen y una enfermedad concreta; por ejemplo, la fibrosis quística se asocia con múltiples variantes del gen *CFTR*.[12] Para llegar a esta conclusión, se llevó a cabo un análisis de ligamiento por el que se rastreó la heredabilidad de las mutaciones de distintos genes a lo largo de varias familias, y que permitió aislar el gen responsable de la enfermedad. Esto fue posible gracias a que la fibrosis quística, así como la enfermedad de Huntington u otras, cumple las leyes de Mendel.[13] Sin embargo, enfermedades más comunes, como el infarto o el cáncer, no son mendelianas: tienden a ser poligénicas, con mutaciones que, de estar presentes, no garantizan la aparición de la enfermedad, y cuyos efectos se agregan a —e incluso interaccionan con— otros.[14]

Los estudios de genes candidatos, motivados por hipótesis, se postularon como una alternativa a los de ligamiento en un momento en que las tecnologías de genotipado no se habían abaratado tanto como lo harían años más tarde,[15] lo que provocaría una capacidad de análisis del genoma humano sin precedentes. Los estudios de asociación del genoma completo (GWAS, por su sigla en inglés), libres de hipótesis, sustituyeron a los de genes candidatos, explorando conjuntos de cientos de miles o millones de variantes genéticas para identificar aquellas que se asocian con una enfermedad dada,[16,17] lo que involucra el análisis de una muestra de individuos de una población específica.[18,19]

El análisis de GWAS suele realizarse por separado para cada variante comparando la frecuencia con que se encuentra en individuos con y sin la enfermedad[20] y asumiendo que uno de los siguientes patrones (generalmente, el primero) subyace tras la asociación variante-enfermedad: aditivo, dominante y recesivo.[21] En los GWAS, no es que no se asuman las leyes de Mendel —no tiene sentido hacerlo para enfermedades comunes, como ya se ha mencionado—, sino que se aprovecha el hecho de que las variantes genéticas son dependientes entre sí para relajar los niveles de significación estadística necesarios de no ser así. A medida que aumenta el número de hipótesis variante-enfermedad que someter a prueba, lo hace también la probabilidad de encontrar falsos positivos, es decir, resultados que parecen asociar una mutación con una enfermedad aun cuando esta asociación no es real. Se han establecido, por tanto, ni-

veles de significación estadística estrictos, aunque no tanto —se divide el nivel de significación inicial por un millón, el número de mutaciones que se estima que son independientes entre sí— como si ninguna mutación del genoma estuviera relacionada con alguna otra.

1.2 Sobre la voz *asociación*

La cantidad de GWAS que se han llevado a cabo desde el primero[22] es inmensa, lo que ha propiciado la creación de catálogos exhaustivos de asociaciones variante-enfermedad.[23] Los GWAS se apoyan en la hipótesis *variante común-enfermedad común*,[24] que establece que las enfermedades más frecuentes se deben a variantes genéticas también frecuentes en la población,[25] de ahí que el grado de influencia (penetrancia o tamaño del efecto) de cada una de ellas sobre la enfermedad sea necesariamente bajo con respecto al que una menos común tendría sobre una enfermedad más rara.[26] Esto implica que la mayoría de los descubrimientos que provienen de los GWAS supongan un riesgo relativamente sutil para la salud[27] y que, aun combinados, no sean capaces de explicar el componente hereditario de la mayoría de enfermedades.[28] Se sabe que las interacciones de estas variantes con factores ambientales o con otras variantes comprometen las suposiciones de la hipótesis,[29] por lo que el hallazgo de interacciones de variantes genéticas podría tener consecuencias importantes para el avance de los GWAS.

El concepto de interacción presenta al menos dos definiciones que surgen de la inexistencia de una correspondencia precisa entre sus interpretaciones biológica y estadística,[30] tanto es así que el fenómeno que describe la primera —la biológica es la relación física entre las biomoléculas de un individuo que se ven alteradas por la interacción de variantes genéticas—[31] podría darse aun en ausencia del que concierne a la segunda —la interacción estadística es la contribución no aditiva de, al menos, una de las variantes al riesgo de una población de padecer una enfermedad—[32] y viceversa.[33]

En las últimas décadas, el análisis de estas interacciones se ha caracterizado por incorporar el producto de dos variantes genéticas a modelos de regresión.[34,35] Sin embargo, dada la heterogeneidad de los patrones por los que una variante genética puede interactuar con otra,[36-38] el producto de ambas puede que no sea capaz de capturar la complejidad de la arquitectura genética de ciertas enfermedades.[39] Un ejemplo de estos patrones es aquel en que el alelo de riesgo de una de las variantes solo tiene efecto si no está presente el de la otra.[40]

Las interacciones de más de dos variantes genéticas plantean cuestiones fundamentales: por ejemplo, en caso de haber computado todas las interacciones de dos, ¿deberían computarse también las interacciones de 3?, ¿y, en general, de k (tal que $3 < k \leq n$, donde n es el número de variantes totales)?[41] Dado que la complejidad computacional de analizar interacciones de más de dos variantes genéticas crece exponencialmente

con k,[42] ha habido un esfuerzo por adaptar métodos de minería de datos y aprendizaje automático (*machine learning*) para la construcción de modelos que recojan varias.[43] Con todo, la mayor parte de estos métodos no aplican un tratamiento estadístico a las interacciones, y, aun para aquellos que lo hacen, es tan inasequible como para los modelos de regresión convencionales.[44]

Redes de interacciones: SEN y reGAIN

Statistical Epistasis Networks (SEN)[45] y Regression Genetic Association Interaction Networks (reGAIN)[46] son dos metodologías intuitivas que abordan el desafío computacional que supone el estudio de las interacciones analizándolas por medio de grafos. Los grafos se utilizan para modelar sistemas con relaciones binarias (no necesariamente recíprocas) entre los elementos que los componen, donde aquellas se representan mediante aristas, mientras que estos últimos, mediante vértices.[47]

El enfoque de las SEN se basa en la teoría de la información, cuyos conceptos se encuentran entre los más utilizados para la construcción y análisis de redes de interacciones.[48,49] Cada variante genética, X, recibe un peso que cuantifica su capacidad para discriminar los individuos sanos de los enfermos (variable Y) a través de la información mutua,

$$\mathrm{I}(X;Y) = \sum_x \sum_y f_{XY}(x,y) \log \left(\frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} \right) \ ,$$

donde $f_{XY}(\cdot,\cdot)$ es la función de probabilidad conjunta de X e Y, cuyas funciones de probabilidad son, respectivamente, $f_X(\cdot)$ y $f_Y(\cdot)$; cada interacción de dos variantes genéticas, X_1 y X_2, recibe una medida análoga a través de la ganancia de información,

$$\mathrm{GI}(X_1;X_2;Y) = \mathrm{I}(X_1;X_2;Y) - \mathrm{I}(X_1;Y) - \mathrm{I}(X_2;Y) \ ,$$

donde $\mathrm{I}(X_1;X_2;Y)$ es una generalización de la información mutua.[50]

Una vez computados los pesos, se pueden construir tantas redes como se desee: dado un valor umbral, t, el grafo G_t se genera a partir de las variantes que cumplan que $\mathrm{GI}(X_i;X_j;Y) \geq t$, donde $i,j \in \{1,\ldots,n\}$ y $j \neq i$, de modo que, cuanto menor sea t, mayor será el número de aristas de G_t. La cuestión, ahora, es qué valor de t da lugar al grafo que mejor se ajusta a la realidad biológica. Los autores de las SEN, por ejemplo, han apostado por el comportamiento de los componentes conexos del grafo como criterio de calidad, de forma que el valor de t más apropiado es aquel para el que el componente conexo más grande del grafo deja de crecer al mismo ritmo. Esta estrategia tiene la ventaja de ser eficiente y fácil de implementar en un lenguaje de programación, aunque carece del rigor y de la información que proporcionan los conceptos derivados de la significación estadística.

Por otro lado, las reGAIN optan por un modelo de regresión logística múltiple, dadas las ventajas de las que goza, y de las que la teoría de la información carece, como el ajuste por covariables o factores de

confusión, el manejo de *missing data* o la inclusión de efectos aleatorios:[51]

$$\log\left(\frac{\mathrm{E}[Y]}{1-\mathrm{E}[Y]}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 \ . \tag{1.1}$$

La ECUACIÓN 1.1 se aplica a cada par de variantes genéticas, es decir, $n(n-1)/2$ veces. Esto puede ser problemático en el contexto de los GWAS, tanto por el tiempo de cómputo necesario (de orden cuadrático) como por los niveles de significación estadística tan exigentes que deben implantarse para evitar falsos positivos. Por ello, los autores de las reGAIN recomiendan el uso de filtros de variantes genéticas —se selecciona un subconjunto reducido de estas de acuerdo con unos criterios de calidad previamente establecidos— para relajar ambos problemas.

Aunque el enfoque de esta tesis doctoral es similar, sobre todo al de las reGAIN, el método que en esta se propone difiere de lo anterior en (i) explotar la información que proporcionan los patrones o modelos genéticos que subyacen tras las asociaciones variante-enfermedad e interacción-enfermedad para tratar de acercar los fenómenos que describen las interpretaciones biológica y estadística del concepto de interacción, (ii) construir la red de interacciones a partir de un modelo de regresión logística múltiple, en que todas las variantes y sus interacciones se ponderan conjuntamente, para evitar los problemas de que adolecen las reGAIN y (iii) analizar la red de interacciones mediante la teoría de grafos para priorizar variantes genéticas que puedan tener efectos notables.

Estructura de la tesis doctoral

La SECCIÓN 1.2 se dedica al repaso de los conceptos biológicos necesarios para comprender el dominio del problema en que se desenvuelve esta tesis doctoral. Se sobrevuelan el dogma central de la biología molecular, la organización del ADN en las células y la genética de poblaciones. Se recomienda solo para el lector no familiarizado con estos temas; de conocerlos, puede omitir el resto de este CAPÍTULO 1 y comenzar el siguiente.

En el CAPÍTULO 2, se ofrecen los resultados de las primeras investigaciones del estudio MCC-Spain en que participó el autor de esta tesis doctoral. A la presentación de sus conclusiones le siguen la metodología propuesta para demostrar la hipótesis de trabajo que motivaron y los objetivos establecidos.

El CAPÍTULO 3, a través de una serie de resultados técnicos que gira en torno al concepto de coherencia estadística, formaliza el proceso para aplicar un test de diferencia-equivalencia e interpretar sus resultados en presencia de comparaciones múltiples.

En el CAPÍTULO 4, se diseña una prueba estadística en el marco de trabajo de los test de diferencia-equivalencia que permite decidir qué modelo genético le corresponde —si es que le corresponde alguno— a un SNP o una interacción SNP-SNP.

En el CAPÍTULO 5, se confecciona un protocolo de construcción de redes de interacciones SNP-SNP para estudios de casos y controles con que analizar los datos del estudio MCC-Spain. Asimismo, se discuten los resultados obtenidos con referencias biológicas.

En el CAPÍTULO 6, se aporta una perspectiva general tanto de los objetivos alcanzados en esta tesis doctoral como de las contribuciones científicas que se derivan de su consecución. Asimismo, se describen las líneas de investigación futuras.

1.3 Sobre la voz *genoma*

El Diccionario de la Real Academia Española define la voz *genoma* como la «secuencia de nucleótidos que constituye el ADN de un individuo o de una especie». El ADN (sigla de *á*cido *d*esoxirribo*n*ucleico), pues, es una molécula que, formada por otras más pequeñas llamadas nucleótidos, porta el material genético de un organismo, es decir, el conjunto de instrucciones implicado tanto en su desarrollo como en su reproducción.

ADN, ARN y proteínas

Los nucleótidos del ADN están compuestos por un azúcar (la desoxirribosa), un grupo fosfato y una base nitrogenada, que puede ser adenina (A), citosina (C), guanina (G) o timina (T).[52] Estos se disponen en cadenas

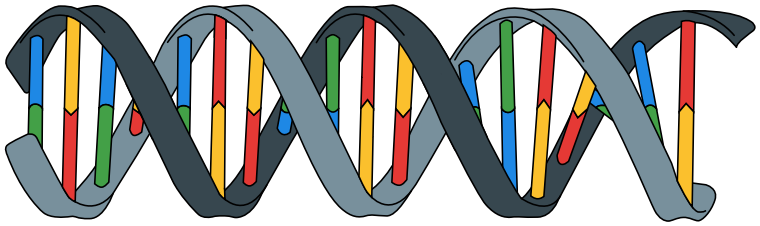

FIGURA 1.2. Segmento de una molécula de ADN. En *negro* y en *gris*, las uniones azúcar-fosfato de las cadenas de nucleótidos en los sentidos $5' \rightarrow 3'$ y $3' \rightarrow 5'$, respectivamente. En el resto de colores, los pares de bases nitrogenadas **A**-**T** y **C**-**G**. FUENTE: Wikimedia Commons. Material adaptado.

con direccionalidad —por consenso, los extremos inicial y final se corresponden con los que se identifican como $5'$ y $3'$, respectivamente— en que el grupo fosfato de cada nucleótido forma un enlace con su azúcar, y este, a su vez, tanto con su base nitrogenada como con el grupo fosfato del siguiente.[53] La molécula de ADN consta de dos cadenas de nucleótidos unidas entre sí a través de sus bases nitrogenadas (véase la FIGURA 1.2): A-T y C-G son los únicos pares de bases que pueden presentarse, lo que implica que ambas cadenas son complementarias, es decir, cualquiera de ellas puede reconstruirse a partir de la otra.[54] Existe una correspondencia unívoca entre los tipos de nucleótidos y sus bases nitrogenadas, por lo que basta enumerar las bases nitrogenadas de una cadena de nucleótidos para describirla. Por ejemplo, la cadena en sentido $5' \rightarrow 3'$ de la FIGURA 1.2 es **CAC AGA TCG TAG CAT CCT**.

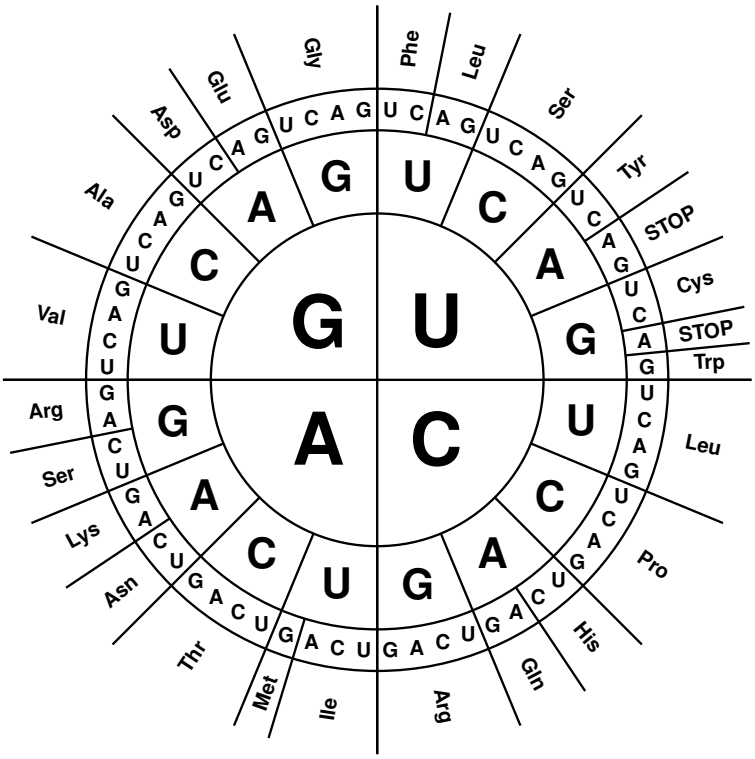

FIGURA 1.3. El código genético. De mayor a menor tamaño de las letras A, C, G y U (*de dentro afuera*), el sentido 5′→3′ de la cadena de nucleótidos del ARN. Cada combinación de tres (codón o triplete) en dicho sentido codifica un aminoácido, que puede resultar, a su vez, de otras combinaciones. Por ejemplo, mientras que solo el codón UGG da lugar al triptófano (*Trp*), la cisteína (*Cys*) viene dada tanto por el triplete UGC como por el UGU.

El orden en que se disponen los nucleótidos en una molécula de ADN determina qué proteínas puede producir, es decir, qué instrucciones puede llevar a cabo. La expresión génica o síntesis de proteínas consta de dos etapas: la transcripción del ADN y la traducción del ARN.[55] En la primera, una enzima crea una copia en ARN de la cadena en sentido 5′→3′ de la molécula de ADN; la secuencia de bases de esta copia es idéntica a la original, salvo por las T, que se sustituyen por uracilos (U).[56] Por ejemplo, la molécula de ARN que resulta de la transcripción del ADN de la FIGURA 1.2 es **CAC AGA UCG UAG CAU CCU**. Por otra parte, la

segunda etapa de la síntesis de proteínas interpreta la molécula de ARN de acuerdo con el código genético (véase la FIGURA 1.3), que asigna un aminoácido —las proteínas son básicamente sucesiones de aminoácidos unidos entre sí— a cada posible combinación de tres bases (codón o triplete).[57] Así, los 6 codones de la molécula de ARN que resulta de la transcripción del ADN de la FIGURA 1.2 codifican sendos aminoácidos: histidina, arginina, serina, STOP, histidina y prolina. Sin embargo, los dos últimos no forman parte de la proteína que surge de la expresión génica de este ejemplo, ya que la traducción finaliza con la lectura de cualquiera de los codones de terminación **UAA**, **UAG** y **UGA**.[58,59] Análogamente, **AUG** constituye el codón de inicio de la traducción, lo que hace que las proteínas dispongan de una metionina en su extremo 5′.

Genes, cromosomas y SNP

La expresión génica solo ocurre en posiciones concretas del ADN que, *nomen est omen*, se denominan genes.[60] Los genes cuentan con dos tipos de regiones que se alternan entre sí, los intrones y los exones, cuya diferencia radica en que, antes de la traducción, los intrones se eliminan de la molécula de ARN transcrita en la primera etapa de la expresión génica.[61] Cada exón contiene, por tanto, un fragmento de la cadena de nucleótidos que da lugar a la proteína codificada en el gen.[62] Dicha proteína, con todo, no es la única que el gen es capaz de producir: a

través de un *splicing* alternativo, la eliminación de dos intrones puede llevar consigo la del exón entre ellos, lo que provoca que tanto la proteína resultante como su función sean distintas a las originales.[63]

Los genes se organizan en cromosomas localizados en el núcleo de las células.[64] En el ser humano, las células no sexuales, diploides, presentan 23 pares de cromosomas homólogos, el último de los cuales hace referencia a los sexuales X e Y: los pares XX y XY se hallan en mujeres y varones, respectivamente.[65] Por otra parte, los gametos (óvulos y espermatozoides) son haploides, ya que solo disponen de una copia del conjunto de cromosomas, mientras que los glóbulos rojos son las únicas células humanas que carecen de ADN.[66] Cada cromosoma, excepto los sexuales en varones, posee los mismos genes en las mismas posiciones (*loci*, y *locus* en singular) que su homólogo; esto implica que ambos compartan las funciones que llevan a cabo, pero no que sean idénticos: debido a la reproducción sexual, uno proviene del padre y el otro de la madre, lo que introduce, inevitablemente, variaciones entre los cromosomas homólogos.[67]

El SNP (sigla en inglés de *p*olimorfismo de un *s*olo *n*ucleótido), que se erige como el tipo de variante genética más común en el ser humano, es un cambio en un par de bases del genoma que puede observarse en una parte significativa de la población.[68] El ser humano, en todos los *loci* de sus cromosomas homólogos, tiene dos bases que, o bien son iguales, o bien son diferentes. La FIGURA 1.4 muestra un *locus* en que

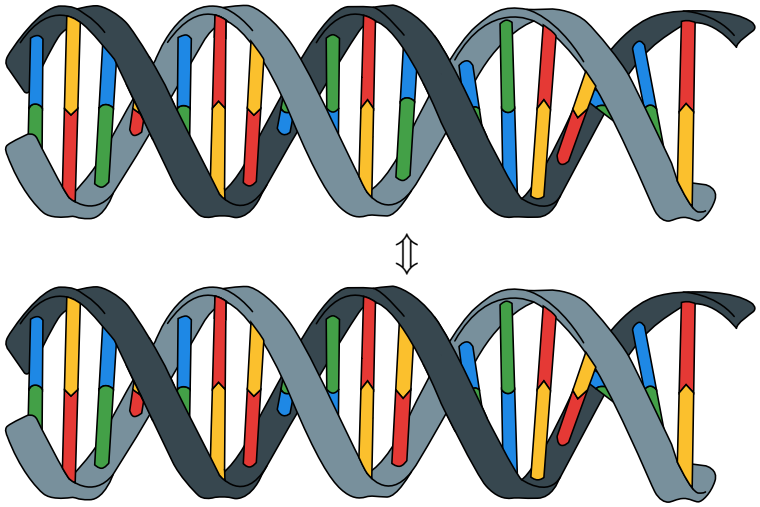

FIGURA 1.4. *Locus* en que un organismo diploide es heterocigoto. En el cromosoma *superior*, la 11.ª base (alelo) de la cadena en sentido $5' \to 3'$ (⇳) es **C**, mientras que en su cromosoma homólogo (el *inferior*) es **A**. En este *locus*, por tanto, el organismo es heterocigoto (con genotipo **CA** o **AC**). FUENTE: Wikimedia Commons. Material adaptado.

una de las bases es distinta en un cromosoma (alelo C) en lo que se refiere a su homólogo (alelo A), por lo que el individuo es heterocigoto —con bases iguales, el individuo sería homocigoto— en dicho *locus*, con genotipo CA o AC.[69] Los términos comparativos alelo mayor y alelo menor hacen referencia a la relación que se establece entre la frecuencia en la población de un alelo con respecto a la del otro, y determinan la tipología de la variante genética: si la MAF (sigla en inglés de *f*recuencia del *a*lelo *m*enor) es, al menos, del 1 %, se trata de un SNP; si no, de una variante rara o mutación.[70]

Los SNP, y en general todas las variantes genéticas, se localizan a lo largo del genoma.[71] En los exones de los genes, los SNP son capaces de alterar la proteína codificada en estos —en la FIGURA 1.4, el cuarto codón

es UCG, serina, en un cromosoma y UAG, STOP, en su homólogo, que produce una proteína truncada en comparación con la otra—, mientras que, en los intrones, son susceptibles de afectar al *splicing* alternativo. Asimismo, pueden encontrarse SNP en regiones intergénicas relacionados con enfermedades o caracteres biológicos.[72]

Ligamiento y haplotipos

En el proceso de formación de los óvulos y espermatozoides, se produce un entrecruzamiento cromosómico (o recombinación genética) por el cual los pares homólogos intercambian segmentos de ADN.[73] De esta manera, cada gameto del individuo porta un genoma que, al estar formado por genes tanto del padre como de la madre, no solo es distinto al del resto de gametos, sino también al del propio organismo.[74] La FIGURA 1.5 ofrece un ejemplo de dicho proceso en tres generaciones de una familia para un par de cromosomas homólogos dado. El entrecruzamiento puede darse en cualquier *locus* del cromosoma (véanse las marcas ❶ y ❷ en la figura), por lo que el ligamiento entre dos *loci* determina su frecuencia de recombinación: cuanto más cercanos, es decir, a mayor ligamiento, menor probabilidad de que un entrecruzamiento cromosómico los separe y no se hereden juntos en forma de haplotipo.[75]

El desequilibrio de ligamiento describe en qué medida se da este fenómeno en una población y un momento concretos,[76] de ahí que puede

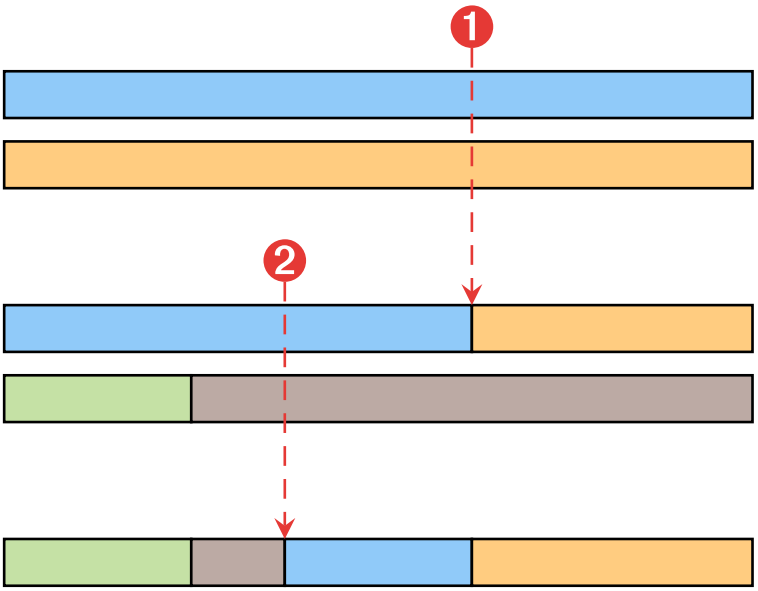

FIGURA 1.5. Ligamiento y haplotipos en tres generaciones de una familia. La pareja de *rectángulos bajo la marca* ❶ representa dos cromosomas homólogos de un individuo que, una vez recombinados (*rectángulo azul y naranja*) y en ausencia de mutaciones, conforman uno de los cromosomas del mismo par en el hijo; el homólogo (*rectángulo verde y marrón*) proviene del otro progenitor que no se muestra. La *marca* ❷ indica el *locus* del entrecruzamiento en el hijo que constituye uno de los cromosomas del mismo par en el nieto (*rectángulo inferior*).

que dos *loci* (por ejemplo, dos SNP) hayan tendido a heredarse juntos en la mayor parte de los individuos de una población, es decir, estén en un desequilibrio de ligamiento elevado, y no en los de otra:[77] las subpoblaciones europeas, por ejemplo, poseen regiones de alto desequilibrio de ligamiento de mayor tamaño que las de las subpoblaciones africanas, ya que las primeras, al ser más recientes, han sufrido menos

recombinaciones genéticas que las segundas.[78] La existencia de estas regiones de alto desequilibrio de ligamiento, en que el genotipo de un *locus* es capaz de determinar en gran medida los genotipos de otros *loci*, ofrece la posibilidad de estudiar el genoma completo a través de un conjunto reducido de SNP representativos (o *tag* SNP) que, al igual que dichas regiones, depende de cada población.[79] El proyecto HapMap proporcionó indicios de que, al menos, el 80 % de la variabilidad de los SNP en subpoblaciones europeas puede recrearse a partir de un conjunto de hasta un millón de SNP repartidos por todo el genoma.[80]

Capítulo

2 Hacia el análisis de redes de interacciones SNP-SNP

En este capítulo, se ofrecen los resultados de las primeras investigaciones del estudio MCC-Spain en que participó el autor de esta tesis doctoral. A la presentación de sus conclusiones le siguen la metodología propuesta para demostrar la hipótesis de trabajo que motivaron y los objetivos establecidos.

En 1598, diez años después de la publicación de su relato sobre una epidemia de peste en Cerdeña,[81] Quinto Tiberio Angelerio introducía la voz *epidemiología* no solo en el nuevo título de la segunda edición de su obra,[82] sino también en el léxico médico.[83] Con los siglos, la epidemiología ha madurado hasta convertirse en una disciplina que integra el formalismo de la estadística y el empirismo de la medicina basada en pruebas para identificar factores relacionados con la salud y cuantificar su asociación. La epidemiología genética se encarga del estudio de las asociaciones genotipo-fenotipo,[84] para lo cual existen dos diseños de estudios fundamentales: los de cohortes y los de casos y controles.[85]

En los estudios de cohortes, se realiza el seguimiento de al menos un grupo de personas expuestas y otro de no expuestas a un factor para comparar después la frecuencia de aparición de un efecto en uno y otro grupo, es decir, «de la exposición al efecto».[86] En cambio, en los de casos y controles, los grupos se definen de acuerdo con la presencia o ausencia del efecto en las personas de que están formados, y más tarde se investiga si estas han estado expuestas al factor, es decir, «del efecto a la exposición».[87] Cada enfoque presenta ventajas —los estudios de cohortes permiten calcular el riesgo relativo (cociente entre las probabilidades de aparición del efecto en los grupos expuesto y no expuesto) y los de casos y controles, estudiar enfermedades raras con una inversión de tiempo y dinero reducida— e inconvenientes —los estudios de casos y controles son más propensos a sesgos, como el de memoria,[88] y los de cohortes, proclives a la obtención tardía y costosa de resultados—[89] que obedecen a cómo se construyen los grupos, sobre todo el enfoque de casos y controles, en que el grupo de controles deben constituirlo personas que, en ausencia del efecto, sean representativas de la población en riesgo de convertirse en casos.[90]

2.1 Interacciones gen-ambiente (MCC-Spain)

El estudio MCC-Spain[91] —MCC es la sigla de *m*ulti*c*aso-*c*ontrol— es un estudio de casos y controles de base poblacional que, entre 2008 y

2013, reclutó 10 106 individuos,[92] de los que casi el 60 % (6008) estaba compuesto por casos de algunos de los cánceres más comunes en España: colorrectal, de mama, de próstata, gástrico y leucemia linfática crónica; los casos y los controles se emparejaron por edad, sexo y provincia de residencia. El reclutamiento de los casos —se trataba de casos incidentes: pasaban a formar parte del estudio en el momento del diagnóstico— se llevó a cabo en 23 hospitales de 12 provincias españolas (Asturias, Barcelona, Cantabria, Gerona, Granada, Guipúzcoa, Huelva, León, Madrid, Murcia, Navarra y Valencia), mientras que los controles se seleccionaron al azar a partir de los registros de centros de atención primaria de estas provincias. El objetivo principal del estudio era —y sigue siendo— investigar la influencia de factores ambientales y de su interacción con factores genéticos en el desarrollo de cáncer.

Tanto a los casos como a los controles se les realizó una entrevista de una hora y media de duración que recogía, entre otras, preguntas sobre factores sociodemográficos, exposiciones ambientales y antecedentes familiares, al final de la cual, además de entregarles un cuestionario validado de frecuencia alimentaria de 140 elementos, se les tomaron medidas antropométricas, así como muestras de sangre, orina, pelo y uña. De los casos, se obtuvo información sobre los síntomas (tipo y fecha de aparición) procedente de las historias clínicas, los métodos de diagnóstico, el estadio tumoral, etc. En 2011, se genotiparon los casos y controles de que entonces se disponía (aproximadamente, el 60 % del

total de individuos) mediante un *array* de exoma de Illumina® con más de 200 000 variantes genéticas en que se incluyeron unas 5000 adicionales de genes implicados en rutas biológicas asociadas con los cánceres colorrectal, de mama, de próstata, gástrico y leucemia linfática crónica.

El estudio MCC-Spain ha permitido identificar variantes genéticas asociadas con los cánceres colorrectal, de mama y de próstata.[93-95] Sin embargo, este estudio, así como los GWAS en general, debe lidiar con el problema del error tipo I —se comete al establecer una asociación que, en realidad, no existe—[96] inflado por comparaciones múltiples.[97] Para abordarlo, se han implantado niveles de significación estadística muy exigentes, por lo que no se ha podido prestar atención a los centenares de miles de variantes genéticas que, o bien no alcanzan esos niveles de significación tan estrictos, o bien se erigen, por su frecuencia en la población, en mutaciones o variantes raras.[98,99]

A continuación, se presentan las contribuciones del autor de este documento al estudio MCC-Spain antes de y al inicio de sus estudios de doctorado, lo que coincidió con el arranque del análisis de los datos de genotipado del estudio. Así, el autor fue una de las dos personas que llevaron a cabo el control de calidad de los datos de genotipado de manera independiente, y participó en las siguientes publicaciones contribuyendo al planteamiento de alternativas al enfoque convencional de analizar SNP independientemente —esto motivó la hipótesis de esta tesis doctoral— y a la redacción de los artículos correspondientes.[100-102]

Control de calidad

El control de calidad del genotipado de los individuos del MCC-Spain se realizó, de manera independiente y con el protocolo que se detalla a continuación, en los nodos de Barcelona y Cantabria. En este último, fue el autor de esta tesis doctoral el que lo llevó a cabo utilizando la aplicación por línea de comandos PLINK.[103] Una vez realizado, se compararon los individuos y los SNP que se habían clasificado como incluidos y excluidos en ambos nodos para comprobar que coincidían.

PASO 1: Se excluyeron 12 885 SNP y 84 individuos (véase la TABLA 2.1). El grupo de SNP incluía 6678 exclusiones previstas. Los individuos, por otra parte, se descartaron por estar etiquetados como «no elegibles» (29), no tener datos epidemiológicos asociados (17), estar duplicados (14) y no ser casos confirmados de los cánceres de mama o gástrico (24).

PASO 2: Se llevó a cabo el control de calidad de los individuos. Para ello, se comprobó que, de cada uno, se desconociera más del 5 % de sus SNP (217), que el sexo especificado no se correspondiera con el de sus cromosomas sexuales (64), que el número de SNP con genotipo heterocigoto fuera excesivo (35 fuera de la media $\pm$ 4 desviaciones estándar) y que tuviera parentesco con, al menos, otro individuo de la muestra (11). Asimismo, para evitar la estratificación de la

población, es decir, que los individuos tengan diferente origen étnico, lo que suele ser contraproducente en el análisis posterior, se computaron los dos primeros componentes principales teniendo en cuenta 11 731 SNP considerados de interés para ello previamente. La TABLA 2.2 recoge el número individuos excluidos por origen por esta razón (53). En total, sin contar las exclusiones del mismo individuo en categorías diferentes, se descartaron 328 SNP.

PASO 3: Se llevó a cabo el control de calidad de los SNP. Para ello, se comprobó que, de cada uno, el número de individuos con genotipo desconocido fuera mayor del 5 % (2619) y que no estuvieran en equilibrio de Hardy-Weinberg (770), por lo que, sin contar las exclusiones del mismo SNP por ambas razones, se descartó un total de 2921 SNP.

PASO 4: Se llevó a cabo un control de calidad adicional de los SNP, consistente en comprobar si alguno no se encontraba en bases de datos externas, como dbSNP (10), o su MAF difería mucho de la presente en el proyecto 1000 Genomes para la población europea (52).

PASO 5: Se filtraron las variantes raras, para lo cual se estableció que los SNP finales habían de tener una MAF superior al 5 %.

	Tamaño	
Paso	SNP	Individuos
1	235 379	7185
2	235 379	6857
3	232 458	6857
4	232 396	6857
5	29 473	6857

TABLA 2.1. SNP e individuos después de cada paso del control de calidad. El control de calidad se inició con 248 264 SNP y 7269 individuos.

País	**Individuos**
Argentina	6
Brasil	1
Chile	5
China	1
Colombia	18
Cuba	1
Ecuador	5
Guatemala	1
Honduras	1
México	1
Marruecos	2
Nicaragua	1
Paraguay	5
Perú	5

TABLA 2.2. Exclusión de individuos por país. Se computaron sus dos primeros componentes principales y se descartaron aquellos individuos fuera de la media $\pm$ 4 desviaciones estándar para cualquiera de los componentes principales.

Cáncer colorrectal

La detección sistemática de cáncer colorrectal por la prueba de sangre oculta en heces ha demostrado ser eficaz para reducir tanto la mortalidad como la incidencia de esta enfermedad.[104] Sin embargo, la eficiencia de esta estrategia depende de las características —la edad (mayor de 50 años) suele ser la única—[105] que definen la población de riesgo, cuya correcta identificación evita realizar pruebas innecesarias para los individuos sanos o inadecuadas para los enfermos. Considerar otras características, como la historia familiar y factores de riesgo ambientales y genéticos, podría mejorar, por tanto, la eficiencia de los programas de detección sistemática de cáncer colorrectal. Los modelos predictivos para cáncer colorrectal que se han propuesto, incluso aquellos que han incluido factores de riesgo genéticos,[106] presentan una capacidad de discriminación limitada.[107] Esto se debe a que cada factor por sí mismo se asocia con un pequeño aumento en el riesgo de cáncer colorrectal,[108] pero ¿y si se combinaran los factores de riesgo ambientales y genéticos?

En el estudio publicado en *Scientific Reports*, **7** (2017), con 1336 casos de cáncer colorrectal y 2744 controles,[100] se desarrollaron modelos predictivos para esta enfermedad en que se consideraron factores con amplios indicios de asociación con ella: el índice de masa corporal, el uso de ácido acetilsalicílico y antiinflamatorios no esteroideos, el consumo de alcohol, la actividad física, la dieta (ingesta de carne roja y de verduras) y

la historia familiar, además de 21 SNP recogidos en la literatura científica. Para mitigar posibles sesgos, como el de selección, estos modelos predictivos se ajustaron por una puntuación de propensión,[109] obtenida de la predicción individual de un modelo de regresión logística múltiple con la edad, el sexo, la provincia de residencia y el nivel educativo como términos principales, así como edad-sexo y provincia-sexo como términos de interacción. Asimismo, se definió una puntuación de riesgo genética[110] como la suma de los alelos de riesgo —cada variante contribuye con 0, 1 o 2 copias del alelo de riesgo— de los 21 SNP. La capacidad de los factores anteriores para el diagnóstico de cáncer colorrectal se evaluó con el área bajo la curva ROC[111] ajustada por la puntuación de propensión.

La FIGURA 2.1 muestra las contribuciones tanto individuales como acumuladas (de cada factor junto con aquellos de contribuciones individuales menores) a la predicción de cáncer colorrectal. La capacidad de los 7 factores ambientales combinados para el diagnóstico de cáncer colorrectal, que dio como resultado un área bajo la curva ROC de 0,6, es mayor que la de la puntuación de riesgo genética por sí sola (0,56). Añadir la historia familiar incrementó la capacidad de diagnóstico hasta 0,61, y con la subsiguiente incorporación de la puntuación de riesgo genética solo se obtuvo una mejora de 0,02. Estos resultados son coherentes con los disponibles en la literatura científica, que recoge áreas bajo la curva ROC de entre 0,56 y 0,74 con puntuaciones de riesgo genéticas de hasta 27 SNP.[112-115] No obstante, se han descrito más de 60 asociaciones entre

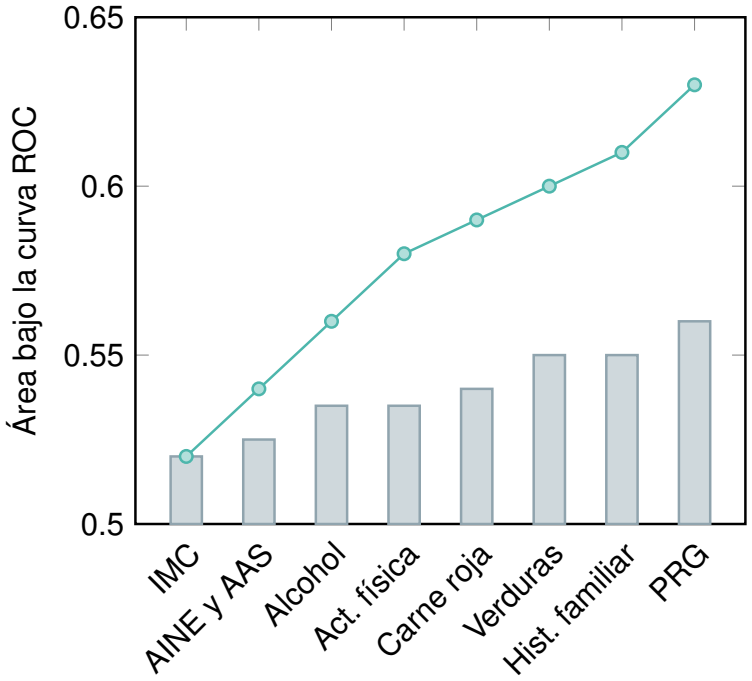

FIGURA 2.1. Capacidad de distintos factores para el diagnóstico de cáncer colorrectal. El diagrama de *barras* y el de *líneas* representan, respectivamente, las contribuciones individuales y acumuladas a la predicción de cáncer.

IMC es la sigla de índice de masa coporal; *AINE*, de antiinflamatorios no esteroideos; *AAS*, de ácido acetilsalicílico; y *PRG*, de puntuación de riesgo genética.

SNP y cáncer colorrectal,[116] por lo que cabe preguntarse si la capacidad de las puntuaciones de riesgo genéticas para el diagnóstico de cáncer colorrectal satura antes de incluir todas estas asociaciones y, sea como fuere, qué papel desempeñan las interacciones SNP-SNP.

Cáncer de próstata

El de próstata es, tras el de pulmón, el cáncer más común en varones de todo el mundo, con una incidencia que aumenta cada año, especialmente en los países occidentales, tanto por el envejecimiento de la población como por los avances en la detección sistemática y el diagnóstico de la

enfermedad.[117] El cáncer de próstata adolece —sobre todo, si se compara con otros cánceres— de una etiología aún por descubrir: solo la edad, la raza y la historia familiar se han establecido como factores de riesgo,[118-120] mientras que el índice de masa corporal, el peso, la dieta, el consumo de alcohol o la diabetes mellitus tipo 2 se siguen discutiendo.[121,122] En el ámbito genético, desde 2006, se han descrito más de 70 asociaciones entre SNP y cáncer de próstata que, en conjunto, explican el 30 % de los casos en la población.[123] La identificación de los alelos de riesgo correspondientes y su combinación con potenciales factores de riesgo podría mejorar la sensibilidad y la especificidad de las pruebas diagnósticas para cáncer de próstata actuales.

En el estudio publicado en *Scientific Reports*, **7** (2017), con 818 casos de cáncer de próstata y 1006 controles,[101] se desarrollaron modelos predictivos para esta enfermedad en que se consideraron factores con algún indicio de asociación con ella, además de la historia familiar. Para mitigar posibles sesgos, como el de selección, estos modelos predictivos se ajustaron por una puntuación de propensión, obtenida de la misma manera que en el estudio de cáncer colorrectal. Asimismo, se definió una puntuación de riesgo ambiental como la suma de las estimaciones de los coeficientes de 5 modelos de regresión logística, con el índice de masa corporal, el peso, la dieta (ingesta de carne roja), el consumo de alcohol y la diabetes mellitus tipo 2 como términos principales. Análogamente, la puntuación de riesgo genética se construyó con 56 SNP recogidos

en la literatura científica. Se encontraron asociaciones estadísticamente significativas entre el cáncer de próstata y los tres grupos de factores: las puntuaciones de riesgo genética (*odds ratio* (OR) $=$ 2,05; intervalo de confianza (IC) al 95 %: de 1,79 a 2,36) y ambiental (OR $=$ 2,47; IC al 95 %: de 1,62 a 3,76) y la historia familiar (OR $=$ 3,32; IC al 95 %: de 2,34 a 4,71). A partir de las fuerzas de asociación entre cáncer de próstata y los tres grupos de factores, se puede obtener una estimación del riesgo relativo (RR), en comparación con la población, de un individuo con puntuaciones de riesgo genética y ambiental G y A, así como ausencia o presencia ($F = 0$ y $F = 1$, respectivamente) de antecedentes de cáncer de próstata en la historia familiar:

$$\text{RR} = 2{,}05^{G-6{,}98}\, 2{,}47^{A-0{,}94}\, 3{,}32^{F} \ .$$

Así, el RR de un individuo con la exposición promedio en la población a estos factores ($G = 6{,}98$, $A = 0{,}94$ y sin antecedentes de la enfermedad en la historia familiar) es igual a 1. Calculado de esta manera, el RR puede utilizarse junto con las tasas de incidencia en una población para la estimación del riesgo absoluto.

La FIGURA 2.2 muestra cómo la edad y el RR moldean las curvas de incidencia estimada de cáncer de próstata en España. Se observa cómo, entre los 50 y los 65 años, la incidencia crece exponencialmente en aquellos individuos con RR > 1, mientras que también crece, aunque con un comportamiento más lineal, para RR ≤ 1. Así, el grupo con RR $= 2$

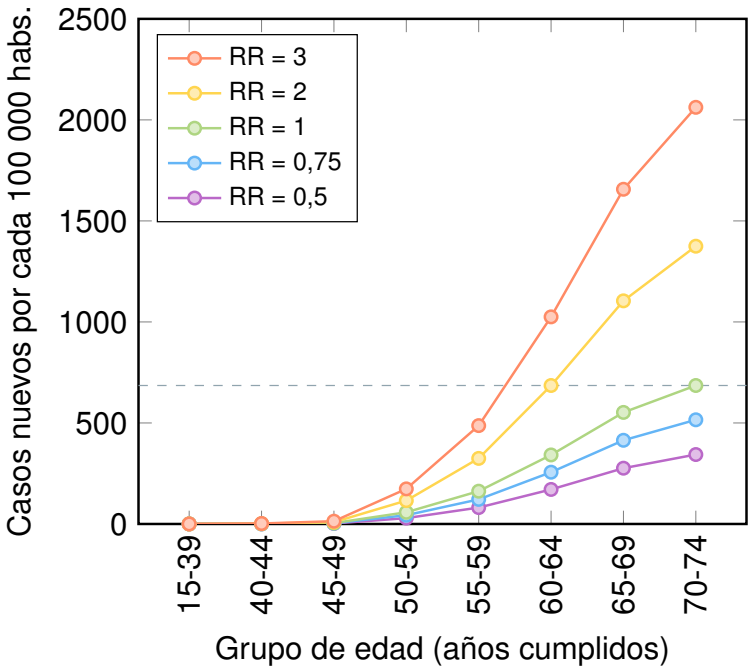

FIGURA 2.2. Incidencia estimada de cáncer de próstata en España por edad y riesgo relativo. La *línea discontinua* se incluye como apoyo al texto principal, en que, dada la misma incidencia estimada, se compara la edad de los individuos con riesgos relativos (*RR*) iguales a 1 y 2.

alcanza, 10 años antes, el nivel de incidencia más alto en los individuos con la exposición promedio en la población, es decir, aquellos con RR = 1 (véase la línea discontinua en la FIGURA 2.2). Esto indica que los modelos predictivos desarrollados podrían ser útiles para estimaciones de RR altas, pero quizá no tanto para aquellas más moderadas, es decir, entre 1 y 2. Téngase en cuenta que el de próstata es el cáncer con un componente hereditario más marcado —en concreto, el 42 % de los casos se atribuyen a causas de índole genética—,[124] que comprende los efectos tanto individuales como combinados de SNP y sus interacciones.[125] La caracterización, por tanto, de los patrones que rigen las interacciones SNP-SNP podría ser fundamental para explicar parte del 58 % restante de los casos de cáncer de próstata.

2.2 Hipótesis y objetivos

La hipótesis de esta tesis doctoral está basada en la enunciada por Moore, que establece que las interacciones SNP-SNP son «un componente ubicuo de la arquitectura genética de las enfermedades humanas comunes y [...] son más importantes que los efectos principales independientes de cualquier gen de susceptibilidad».[126] Ante un escenario así, en que debe ponderarse la asociación entre cada par de SNP y la enfermedad, el uso de la teoría de grafos es óptimo como herramienta de cálculo y visualización eficientes en redes.

Hipótesis: definición y metodología

Codifíquese un SNP de un individuo, X, de acuerdo con cuántos alelos menores presenta la suma de sus dos alelos, por lo que el conjunto de valores que puede tomar es $R_X = \{0, 1, 2\}$. Cada uno de los patrones clásicos que rigen las asociaciones SNP-enfermedad, recogidos en la TABLA 2.3, recodifica el SNP siguiendo la función genotípica correspondiente. Así, el modelo aditivo indica que el incremento del riesgo de enfermedad por el genotipo BB es el doble que por el genotipo AB; el dominante, que el incremento del riesgo de enfermedad por los genotipos AB y BB es el mismo; el sobredominante, que se requiere una copia de cada alelo para el incremento del riesgo de enfermedad; y el recesivo, dos copias del alelo B para el incremento del riesgo de enfermedad. El

	Genotipo		
Modelo genético	AA	AB	BB
Aditivo	0	1	2
Dominante	0	1	1
Sobredominante	0	1	0
Recesivo	0	0	1

TABLA 2.3. Funciones genotípicas de los modelos genéticos. *A* y *B* denotan, respectivamente, el alelo mayor y el menor con respecto a la población.

concepto de patrón o modelo genético es clave para la hipótesis, que conjetura que un porcentaje significativo de los SNP y las interacciones SNP-SNP sigue uno de ellos en su asociación con la enfermedad.

Sean $\{X_1, \ldots, X_n\}$ un conjunto de SNP y $\mathcal{F} = \{f_\mathrm{A}, f_\mathrm{D}, f_\mathrm{S}, f_\mathrm{R}\}$, el de las funciones genotípicas clásicas. Así, el siguiente modelo de regresión logística múltiple serviría para ponderar las interacciones:[b]

$$\log\left(\frac{\mathrm{E}[Y]}{1-\mathrm{E}[Y]}\right) = \beta_0 + \sum_i \beta_i f_i(X_i) + \sum_j \beta_{ij} f_{ij}(f_i(X_i), f_j(X_j)) \ , \tag{2.1}$$

donde $f_i(\cdot), f_j(\cdot) \in \mathcal{F}$ y $f_{ij}(\cdot,\cdot) \in \mathcal{F} \setminus \{f_\mathrm{A}\}$ —como modelo genético de una interacción, el aditivo daría lugar a colinealidad— tal que $i, j \in \{1, \ldots, n\}$ y $j > i$. La estimación de los efectos principales de los SNP, además de que ajusta los efectos de interacción, se justifica por el hecho de que uno no puede asumir, en general, la existencia de los segundos en ausencia de los primeros, es decir, que $\beta_k = 0, \forall k \in \{1, \ldots, n\}$.[127]

[b] Nótese que, en las expresiones de tipo $f(\cdot)$ de la ECUACIÓN 2.1, se comete un abuso de notación que permite simplificar $f(\mathrm{alelo}_1(\cdot), \mathrm{alelo}_2(\cdot))$.

Compárese el modelo de regresión logística múltiple de reGAIN (véase la ECUACIÓN 1.1) con el de la ECUACIÓN 2.1: el primero incorpora los SNP X_1 y X_2 y su producto,X_1X_2, para representar los términos principales y de interacción, respectivamente, mientras que, en el segundo, cada término queda matizado por los modelos genéticos correspondientes mediante $f(\cdot)$ y $f(\cdot,\cdot)$. Decidir, pues, qué modelo le corresponde —si es que le corresponde alguno— a cada SNP e interacción es básico para el método que se propone en esta tesis doctoral. Existen varios para la selección de modelos genéticos, entre los que destacan el MAX y el MERT.[128] Ambos se basan en la prueba de tendencia de Cochran-Armitage,[129,130] la cual permite asignar, a través de los «pesos» de la TABLA 2.3, un modelo genético a la asociación SNP-enfermedad bajo estudio. Los estadísticos de contraste de los métodos MAX y MERT,

$$t_{\text{MAX}} = \text{máx}\left\{|a|,|d|,|r|\right\} \quad \text{y} \quad t_{\text{MERT}} = \frac{d+r}{\sqrt{2\left(1+\widehat{\text{Corr}}(D,R)\right)}} \ ,$$

donde $\widehat{\text{Corr}}(D,R)$ es el estimador de la correlación entre los modelos dominante y recesivo, contienen los de Cochran-Armitage correspondientes a los modelos aditivo, dominante y recesivo (a, d y r, respectivamente). Así, ninguno de los dos métodos —en realidad, podría incluirse cualquiera de los recogidos en revisiones recientes—[131] considera el modelo sobredominante ni la posibilidad de que la asociación SNP-enfermedad no esté regida por un modelo genético. Esto último es crucial, ya que la

hipótesis de esta tesis doctoral conjetura que

> *un porcentaje significativo de las asociaciones SNP-enfermedad e interacción-enfermedad están regidas por uno de los siguientes modelos: dominante, sobredominante, recesivo y, solo en el caso SNP-enfermedad, aditivo,*

no que «todas» las asociaciones lo estén. Por ello, el método que se propone en esta tesis doctoral no solo debe decidir qué modelo le corresponde a cada asociación, sino que, previamente, ha de ser capaz de filtrar aquellas asociaciones sin modelo genético.

La segunda parte de la hipótesis dice que,

> *sin embargo, las redes en que, de forma estadísticamente significativa, todas las asociaciones SNP-enfermedad e interacción-enfermedad siguen uno de los modelos anteriores están formadas por componentes conexos pequeños.*

De ser así, esto dificultaría la tarea de priorizar SNP con posibles efectos notables, ya que las métricas derivadas de la teoría de grafos se benefician de lo conexo de la red.

Hipótesis: simulación

En el 30.º IEEE International Symposium on Computer-Based Medical Systems, celebrado en Salónica a finales de junio de 2017, el autor de esta tesis doctoral defendió una simulación en que se investigaba la eficacia de la metodología esbozada en la sección anterior para el estudio de la

hipótesis presentada.[132] En este análisis, se habían generado al azar 50 SNP con frecuencia del alelo menor entre 0,05 y 0,5, cada uno de ellos con modelo dominante o recesivo en su asociación con la enfermedad, para 10 000 casos y controles; las asociaciones entre las interacciones SNP-SNP y la enfermedad estaban regidas también por uno de ambos modelos. Ajustando el modelo lineal de la ECUACIÓN 2.1, se habían estimado los efectos de las interacciones y calculado sus valores de p, para después descartar aquellas con $p \geq 0{,}05$ y construir la red con las restantes.

La FIGURA 2.3 muestra la red de interacciones de los 50 SNP simulados, donde los vértices se corresponden con los SNP y las aristas y su grosor, con las interacciones y su grado ponderado, respectivamente. El grado ponderado del SNP 44 es mayor que el del resto, por lo que el efecto conjunto (la suma) de sus interacciones con otros SNP es el más destacado. Si el grado ponderado de un SNP es menor que 0, se dice que el efecto conjunto de sus interacciones con otros SNP es de protección; si es mayor, de riesgo.

En esta simulación, los SNP 20, 26, 36 y 44 constituyeron los factores de riesgo más relevantes, mientras que los SNP 3, 9, 13 y 42 son aquellos que presentaron un efecto de protección más acusado. Nótese que esta medida, el grado ponderado, se basa en la acumulación de los efectos de interacción con otros SNP, por lo que un SNP cuyo efecto principal es de protección podría ser un factor de riesgo, y viceversa, si se considera el grado ponderado como criterio.

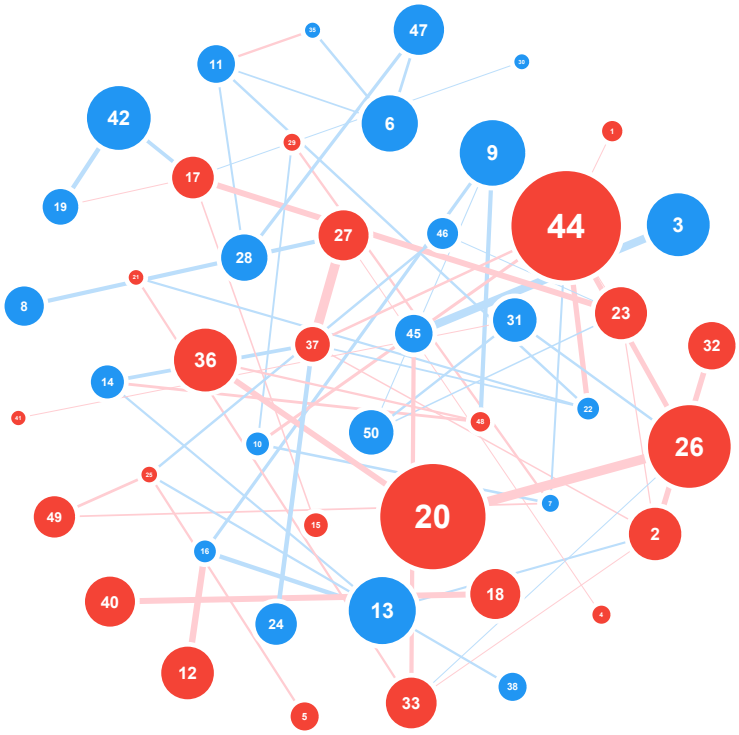

FIGURA 2.3. Red de interacciones de 50 SNP simulados. Los *vértices* y las *aristas azules* representan factores e interacciones, respectivamente, de protección; en *rojo*, de riesgo. El *grosor de las aristas* captura la importancia de los efectos de interacción. El *tamaño de los vértices* se establece en función del valor absoluto de su grado ponderado.

Debe tenerse en cuenta que, en la simulación, subyacía un modelo genético tras «todas» —en lugar de tras «un porcentaje significativo», como dice la primera parte de la hipótesis— las asociaciones SNP-enfermedad e interacción-enfermedad, lo cual pudo influir en el hecho de que la red exhibiera un componente conexo y no varios pequeños, como conjetura la segunda parte de la hipótesis.

Objetivos

Los objetivos que se desprenden tanto de la hipótesis de esta tesis doctoral como de la metodología que adoptar para demostrarla están en consonancia con la estructura en capítulos del resto del documento:

OBJETIVO 1: Desarrollar un marco de trabajo o modelo estadístico que permita evaluar y comparar el nivel de adecuación —debe contemplarse desde la no adecuación hasta la sobreadecuación, así como la incertidumbre— de distintos patrones a volúmenes masivos de datos («problema de las comparaciones múltiples») potencialmente redundantes. Se corresponde con el CAPÍTULO 3.

OBJETIVO 2: Diseñar e implementar una prueba estadística en el marco de trabajo del OBJETIVO 1 que permita decidir qué modelo genético le corresponde —si es que le corresponde alguno— a un SNP o una interacción (véanse las funciones $f(\cdot)$ y $f(\cdot,\cdot)$ en la ECUACIÓN 2.1). Se corresponde con el CAPÍTULO 4.

OBJETIVO 3: Construir y analizar redes de interacciones SNP-SNP a partir de los datos del estudio MCC-Spain con la metodología esbozada en este capítulo, basada en la prueba estadística resultante del OBJETIVO 2, discutiendo los resultados con referencias biológicas. Se corresponde con el CAPÍTULO 5.

Capítulo

3 Contraste de múltiples hipótesis de diferencia-equivalencia

En este capítulo, a través de una serie de resultados técnicos que gira en torno al concepto de coherencia estadística, se formaliza el proceso para aplicar un test de diferencia-equivalencia e interpretar sus resultados en presencia de comparaciones múltiples.

El contraste de hipótesis de equivalencia, que ha regido el proceso de aprobación de medicamentos en Estados Unidos durante las últimas décadas,[133] está ganando adeptos en ámbitos cada vez más alejados de su reino,[134] ya que viene a llenar el vacío metodológico que se ha prolongado por la incapacidad del contraste de hipótesis de diferencia para demostrar semejanza o similitud.[135] Esta incapacidad radica en su falta de equilibrio:[136] mientras rechazar la hipótesis nula (H_0) lleva a la conclusión de que la hipótesis alternativa (H_1) es verdadera, no ser capaz de rechazar H_0 no implica que ella misma sea verdadera.[137]

Así, el contraste de hipótesis de diferencia no podría demostrar semejanza o similitud —es H_0 la hipótesis equipada con la noción de

equivalencia ($H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$, donde θ y θ_0 son, respectivamente, el parámetro poblacional bajo estudio y un valor de referencia con que compararlo)— aun si H_0 y H_1 intercambiaran los signos de igualdad y desigualdad, para lo cual no se conoce un test razonable.[138] El de equivalencia supera este obstáculo añadiendo a su formulación los extremos de un intervalo que depende del problema, $\theta_0 - \delta_1$ y $\theta_0 + \delta_2$ ($\delta_1, \delta_2 \in \mathbb{R}$ tal que $\delta_1, \delta_2 \geq 0$), para determinar si las estimaciones del parámetro poblacional, $\widehat{\theta}$, se encuentran entre ellos (equivalencia), es decir, $H_0 : \theta \notin (\theta_0 - \delta_1, \theta_0 + \delta_2)$, $H_1 : \theta \in (\theta_0 - \delta_1, \theta_0 + \delta_2)$.[139]

La comparación de dos medias es un caso paradigmático del contraste de hipótesis de equivalencia ($\theta = \mu_A - \mu_B$, casi siempre con $\theta_0 = 0$),[140] lo que justifica algunos de los nombres que, además de «margen de equivalencia», han recibido δ_1 y δ_2 en la literatura científica, como «diferencia irrelevante» y «distancia significativa mínima».[141] En cuanto al intervalo $(\theta_0 - \delta_1, \theta_0 + \delta_2)$, cuya definición tiene un gran impacto en la potencia estadística y el tamaño muestral necesarios,[142] siempre que $\delta_1 = \delta_2 = \delta$, es conveniente reformular el problema:

$$\begin{aligned} H_0 &: \theta \notin (\theta_0 - \delta, \theta_0 + \delta) \ , \\ H_1 &: \theta \in (\theta_0 - \delta, \theta_0 + \delta) \ . \end{aligned} \tag{3.1}$$

El procedimiento TOST (del inglés t*wo* o*ne-sided t*ests) es el soberano del contraste de hipótesis de equivalencia,[143] lo cual se debe, en parte, al apoyo de la Food and Drug Administration de Estados Unidos.[144]

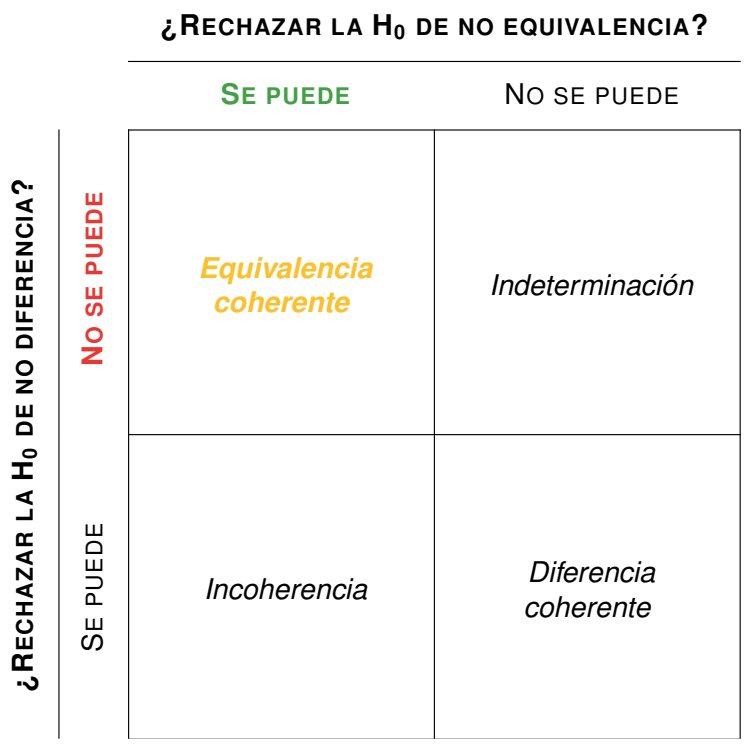

FIGURA 3.1. Los cuatro resultados de un TDE.

Este método descompone el problema en dos test de diferencia (TD) unilaterales o de una cola que arrojan sendos valores de p ($\overrightarrow{p_e}$ y $\overleftarrow{p_e}$ para los test de las colas derecha e izquierda, respectivamente):

$$\begin{aligned} \overrightarrow{H_0} &: \theta \leq \theta_0 - \delta \ , \\ \overrightarrow{H_1} &: \theta > \theta_0 - \delta \ ; \end{aligned} \quad \text{y} \quad \begin{aligned} \overleftarrow{H_0} &: \theta \geq \theta_0 + \delta \ , \\ \overleftarrow{H_1} &: \theta < \theta_0 + \delta \ . \end{aligned} \tag{3.2}$$

La combinación de estos valores de p permite a TOST aceptar la hipótesis de equivalencia (H_1 en la ECUACIÓN 3.1) a nivel α si por separado se rechazan tanto $\overrightarrow{H_0}$ como $\overleftarrow{H_0}$ a nivel α (máx $\{\overrightarrow{p_e}, \overleftarrow{p_e}\} < \alpha$). Dado que los dos valores de p son dependientes entre sí, cada hipótesis individual no necesita comprobarse a nivel $\alpha/2$.[145]

Lo que subyace tras el procedimiento TOST es el enfoque de Ander-

son y Hauck,[146,147] así como otros equivalentes a este.[148-150] La principal discrepancia entre ellos y TOST reside en la regla de decisión tomada para demostrar semejanza o similitud: los primeros aceptan H_1 a nivel α si $|\overrightarrow{p_e} - \overleftarrow{p_e}| < \alpha$, lo que conlleva una percepción más liberal del problema.[151] Sin embargo, estos métodos apenas se usan en la práctica, ya que adolecen de regiones de rechazo no acotadas para las que es posible aceptar la hipótesis de equivalencia incluso si $\widehat{\theta} \notin (\theta_0 - \delta, \theta_0 + \delta)$, siempre y cuando la variabilidad sea suficientemente grande.[152] En este capítulo, TOST es el test de equivalencia (TE) considerado, por lo que ambos términos se tratan como sinónimos.

Aplicar conjuntamente un TD y un TE de hipotésis nulas de no diferencia y no equivalencia, respectivamente, es decir, un test de diferencia-equivalencia (TDE), puede complementar el alcance de ambos test combinando sus conclusiones, de lo que derivan cuatro posibles resultados (véase la FIGURA 3.1) a condición de que dichas hipótesis nulas compartan θ y θ_0 en sus formulaciones y sean contrastadas en la misma muestra al mismo nivel de significación.[153] Uno de los resultados que puede parecer paradójico es el de haber demostrado diferencia además de equivalencia (incoherencia), cuyas causas incluyen grandes tamaños muestrales y el problema de las comparaciones múltiples.[154]

El problema de las comparaciones múltiples es inherente a cualquier contraste de más de una hipótesis, ya sean TD o TE: cuando cada una de varias hipótesis nulas se contrasta a nivel α, la probabilidad de que

algunas se rechacen incorrectamente es mayor que α, lo que puede provocar un abundante número de falsos positivos o errores tipo I.[155] A pesar de que la literatura científica está repleta de procedimientos que impiden que esto ocurra tanto en TD[156] como en TE,[157] no se ha prestado atención al problema de las comparaciones múltiples en TDE.[158]

3.1 Equivalencia coherente

A partir del fructífero trabajo de Rogers, Howard y Vessey,[159] ha habido un flujo lento de publicaciones sobre el uso de TDE en áreas tan variadas como la epidemiología,[160] la sociología[161] y la ingeniería de software.[162] A pesar de esta variedad, la estrategia de interpretación de los TDE rara vez ha sido distinta de la que representa la FIGURA 3.1; esta hace posible una lectura conjunta de los dos test que constituyen un TDE mediante la intersección de sus regiones de rechazo y no rechazo.

El concepto de coherencia es esencial para la búsqueda de un tratamiento no sesgado de la equivalencia. Como ya se ha mencionado, en el contraste de hipótesis de diferencia, la hipótesis equipada con la noción de semejanza o similitud es $H_0 : \theta = \theta_0$, mientras que en el de equivalencia es $H_1 : \theta \in (\theta_0 - \delta, \theta_0 + \delta)$. De esta manera, no poder rechazar la primera y aceptar la segunda es lo que se denomina equivalencia coherente (véase la FIGURA 3.2); análogamente, la diferencia coherente implica rechazar la primera y no aceptar la segunda.

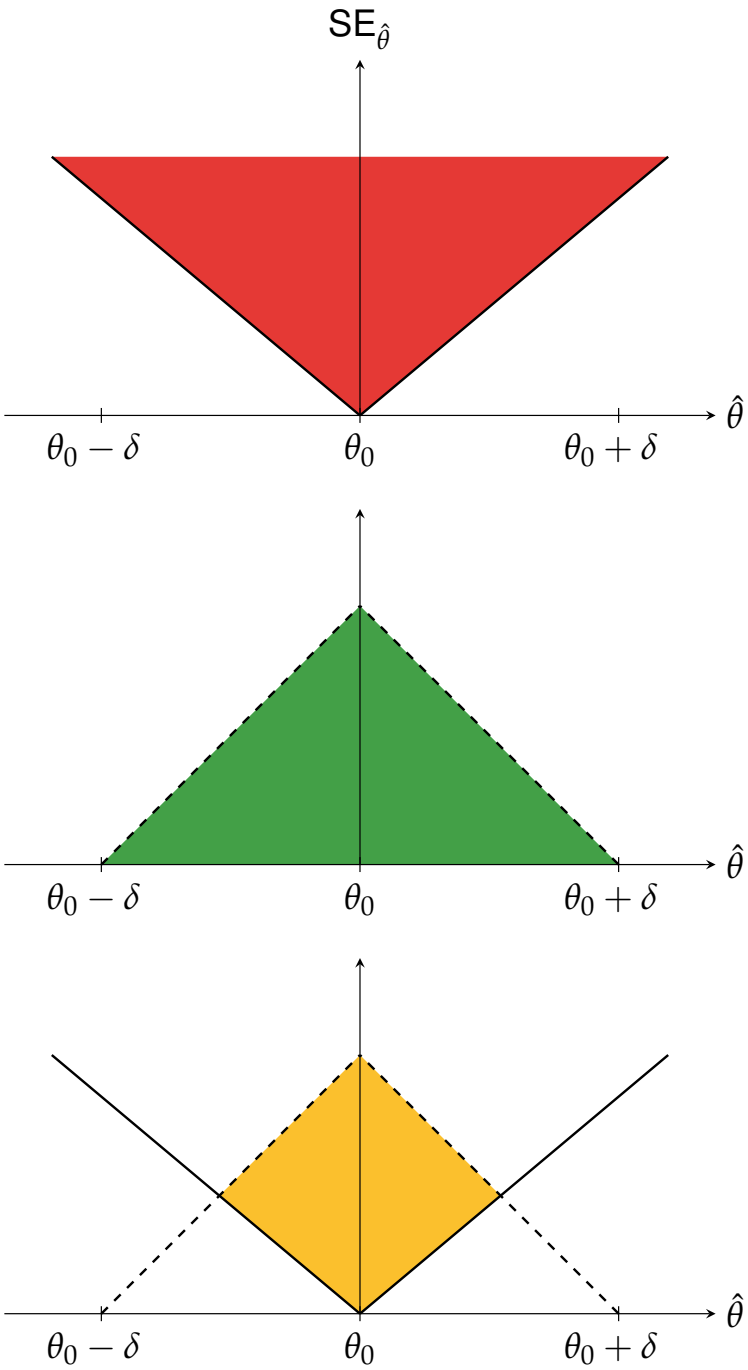

FIGURA 3.2. Elementos de la equivalencia coherente. (*Parte superior*) Conjunto de valores de $\widehat{\theta}$ y $SE_{\widehat{\theta}}$ que no llevan a rechazar la hipótesis nula de no diferencia usando un TD a nivel α. (*Parte central*) Conjunto de valores de $\widehat{\theta}$ y $SE_{\widehat{\theta}}$ que llevan a rechazar la hipótesis nula de no equivalencia usando TOST a nivel α. (*Parte inferior*) Intersección de los conjuntos anteriores, que se corresponde con no demostrar diferencia usando un TD además de demostrar equivalencia usando TOST. Nótese que los conjuntos incluyen y excluyen los valores representados por las líneas continuas y discontinuas, respectivamente.

Estadísticos de contraste y valores de *p*

Por una parte, no se puede rechazar la hipótesis nula de no diferencia $\mathrm{H}_0 : \theta = \theta_0$ si $|t_\mathrm{d}| \leq w_{\alpha/2}$ (o $p_\mathrm{d} \geq \alpha$), donde

$$t_\mathrm{d} = \frac{\widehat{\theta} - \theta_0}{\mathrm{SE}_{\widehat{\theta}}} \tag{3.3}$$

y $w_{\alpha/2}$ es el $(\alpha/2)$-ésimo cuantil superior de la distribución de T_d, la variable aleatoria de la que t_d es una realización, es decir, $w_{\alpha/2} = Q_{T_\mathrm{d}}(1 - \alpha/2)$. El valor de p asociado con la ECUACIÓN 3.3 para un test bilateral o de dos colas es

$$p_\mathrm{d} = 2\Pr(T_\mathrm{d} \geq |t_\mathrm{d}| : \mathrm{H}_0 \text{ es verdadera}) = 2\,F_{T_\mathrm{d}}(-|t_\mathrm{d}|) \ , \tag{3.4}$$

donde $F_{T_\mathrm{d}}(\cdot)$ y $Q_{T_\mathrm{d}}(\cdot)$ son la CDF (sigla en inglés de *f*unción de *d*istribución a*c*umulada) y la CDF inversa de T_d, respectivamente.

Por otra parte, se puede rechazar la hipótesis nula de no equivalencia $\mathrm{H}_0 : \theta \notin (\theta_0 - \delta, \theta_0 + \delta)$ si se rechazan $\overrightarrow{\mathrm{H}_0}$ y $\overleftarrow{\mathrm{H}_0}$ en la ECUACIÓN 3.2, es decir, si $\overrightarrow{t_\mathrm{e}}, -\overleftarrow{t_\mathrm{e}} > w_\alpha$ (o $\overrightarrow{p_\mathrm{e}}, \overleftarrow{p_\mathrm{e}} < \alpha$), donde

$$\overrightarrow{t_\mathrm{e}} = \frac{\widehat{\theta} - (\theta_0 - \delta)}{\mathrm{SE}_{\widehat{\theta}}} \quad \text{y} \quad \overleftarrow{t_\mathrm{e}} = \frac{\widehat{\theta} - (\theta_0 + \delta)}{\mathrm{SE}_{\widehat{\theta}}} \ . \tag{3.5}$$

Asimismo, w_α es el α-ésimo cuantil superior de la distribución de $\overrightarrow{T_\mathrm{e}}$ y $\overleftarrow{T_\mathrm{e}}$, las variables aleatorias de las que $\overrightarrow{t_\mathrm{e}}$ y $\overleftarrow{t_\mathrm{e}}$ son realizaciones, es decir, $w_\alpha = Q_{T_\mathrm{e}^{\rightarrow}}(1 - \alpha) = Q_{T_\mathrm{e}^{\leftarrow}}(1 - \alpha)$. Los valores de p asociados con la

ECUACIÓN 3.5 para test unilaterales son

$$\overrightarrow{p_{\mathrm{e}}} = \Pr(\overrightarrow{T_{\mathrm{e}}} \geq \overrightarrow{t_{\mathrm{e}}} : \overrightarrow{\mathrm{H}_0} \text{ es verdadera}) = F_{T_{\mathrm{e}}^{\rightarrow}}(-\overrightarrow{t_{\mathrm{e}}}) \text{ y}$$
$$\overleftarrow{p_{\mathrm{e}}} = \Pr(\overleftarrow{T_{\mathrm{e}}} \leq \overleftarrow{t_{\mathrm{e}}} : \overleftarrow{\mathrm{H}_0} \text{ es verdadera}) = F_{T_{\mathrm{e}}^{\leftarrow}}(\overleftarrow{t_{\mathrm{e}}}) \ .$$

Los paralelismos entre $\overrightarrow{t_{\mathrm{e}}}$ y $\overleftarrow{t_{\mathrm{e}}}$ y entre $\overrightarrow{p_{\mathrm{e}}}$ y $\overleftarrow{p_{\mathrm{e}}}$ motivan los siguientes resultados.

Proposición 1. *El estadístico de contraste para el procedimiento TOST es*

$$t_{\mathrm{e}} = \frac{\delta - |\widehat{\theta} - \theta_0|}{\mathrm{SE}_{\widehat{\theta}}} \ .$$

Demostración. El procedimiento TOST rechaza H_0 en la ECUACIÓN 3.1 a nivel α si máx $\{\overrightarrow{p_{\mathrm{e}}}, \overleftarrow{p_{\mathrm{e}}}\} < \alpha$ o, lo que es lo mismo, mín $\{\overrightarrow{t_{\mathrm{e}}}, -\overleftarrow{t_{\mathrm{e}}}\} > w_\alpha$. Reorganizando la ECUACIÓN 3.5,

$$\overrightarrow{t_{\mathrm{e}}} = \frac{\delta + (\widehat{\theta} - \theta_0)}{\mathrm{SE}_{\widehat{\theta}}} \quad \text{y} \quad -\overleftarrow{t_{\mathrm{e}}} = \frac{\delta - (\widehat{\theta} - \theta_0)}{\mathrm{SE}_{\widehat{\theta}}} \ ,$$

donde $\widehat{\theta}, \theta_0, \delta, \mathrm{SE}_{\widehat{\theta}} \in \mathbb{R}$ tal que $\delta \geq 0$ y $\mathrm{SE}_{\widehat{\theta}} > 0$. Por tanto, el signo de $\widehat{\theta} - \theta_0$ determina el valor de mín $\{\overrightarrow{t_{\mathrm{e}}}, -\overleftarrow{t_{\mathrm{e}}}\}$:

$$\text{mín}\,\{\overrightarrow{t_{\mathrm{e}}}, -\overleftarrow{t_{\mathrm{e}}}\} = \begin{cases} \overrightarrow{t_{\mathrm{e}}} & \text{si } \widehat{\theta} \leq \theta_0 \ , \\ -\overleftarrow{t_{\mathrm{e}}} & \text{si } \widehat{\theta} \geq \theta_0 \ , \end{cases}$$

que es igual a t_{e}. □

Proposición 2. *El valor de p para el procedimiento TOST es*

$$p_{\mathrm{e}} = \Pr(T_{\mathrm{e}} \geq t_{\mathrm{e}} : \overrightarrow{\mathrm{H}_0} \text{ y } \overleftarrow{\mathrm{H}_0} \text{ son verdaderas}) = F_{T_{\mathrm{e}}}(-t_{\mathrm{e}}) \ .$$

Demostración. TOST rechaza H_0 en la ECUACIÓN 3.1 a nivel α si $t_e > w_\alpha$, que es la regla de decisión para los test de la cola derecha. Así, el valor de p para este test es $\Pr(T_e \geq t_e : H_0 \text{ es verdadera}) = F_{T_e}(-t_e)$. □

Calcular el valor de p para el resultado de equivalencia coherente puede requerir de suposiciones estadísticas arriesgadas, como la de la independencia de $\widehat{\theta}$ y $SE_{\widehat{\theta}}$, que lleva a error a menos que $\theta \sim N(\mu, \sigma^2)$.[163]

Regiones de no rechazo

Dada una estimación del parámetro poblacional, $\widehat{\theta}$, las líneas continuas y discontinuas de la FIGURA 3.2 representan los valores más extremos de $SE_{\widehat{\theta}}$ tales que $|t_d| = w_{\alpha/2}$ (o $p_d = \alpha$) y $t_e = w_\alpha$ (o $p_e = \alpha$), respectivamente. Por tanto, la ecuación de las líneas continuas es

$$SE^*_{\widehat{\theta}} = \frac{|\widehat{\theta} - \theta_0|}{w_{\alpha/2}} \tag{3.6}$$

—no se puede rechazar la hipótesis nula de no diferencia si $SE_{\widehat{\theta}} \geq SE^*_{\widehat{\theta}}$ (en rojo en la FIGURA 3.2)—, mientras que para las líneas discontinuas es

$$SE^*_{\widehat{\theta}} = \frac{\delta - |\widehat{\theta} - \theta_0|}{w_\alpha} \tag{3.7}$$

—se rechaza la hipótesis nula de no equivalencia si $SE_{\widehat{\theta}} < SE^*_{\widehat{\theta}}$ (en verde en la FIGURA 3.2)—, para todo $\widehat{\theta} \in \mathbb{R}$. De esta manera, si

$$\frac{|\widehat{\theta} - \theta_0|}{w_{\alpha/2}} \leq SE_{\widehat{\theta}} < \frac{\delta - |\widehat{\theta} - \theta_0|}{w_\alpha} \ , \tag{3.8}$$

la tupla $(\widehat{\theta}, \mathrm{SE}_{\widehat{\theta}})$ define un punto en $\{(x, y) \in \mathbb{R}^2 : y \geq 0\}$ dentro del deltoide amarillo en la FIGURA 3.2, que se corresponde con el resultado de equivalencia coherente, o no demostrar diferencia usando un TD además de demostrar equivalencia usando el procedimiento TOST. Los siguientes resultados se derivan de esto.

Teorema 3. *Si la* ECUACIÓN *3.8 se cumple,* $\widehat{\theta} \in (\theta_0 - \delta_\alpha, \theta_0 + \delta_\alpha)$*, para todo* $\alpha \in (0, 2/3)$*, donde*

$$\delta_\alpha = \delta \left(1 + \frac{w_\alpha}{w_{\alpha/2}}\right)^{-1} > \frac{\delta}{2}$$

y w_α *y* $w_{\alpha/2}$ *son el* α*-ésimo y el* $(\alpha/2)$*-ésimo cuantiles superiores de la distribución normal estándar.*

Demostración. Comparando las partes izquierda y derecha de la ECUACIÓN 3.8,

$$\frac{|\widehat{\theta} - \theta_0|}{w_{\alpha/2}} < \frac{\delta - |\widehat{\theta} - \theta_0|}{w_\alpha} \iff |\widehat{\theta} - \theta_0| < \delta \left(1 + \frac{w_\alpha}{w_{\alpha/2}}\right)^{-1},$$

se tiene que $\widehat{\theta} \in (\theta_0 - \delta_\alpha, \theta_0 + \delta_\alpha)$ si la ecuación se cumple.

Para la distribución normal estándar, $w_{1/2} = 0$ y $w_\alpha = -w_{1-\alpha}$, y $w_\alpha \to \infty$ y $w_\alpha \to -\infty$ a medida que $\alpha \to 0^+$ y $\alpha \to 1^-$, respectivamente. Sea $(\widehat{\theta}^*, \mathrm{SE}^*_{\widehat{\theta}})$ el punto de intersección de las líneas representadas por las partes izquierda y derecha de la ECUACIÓN 3.8. Estas líneas son paralelas entre sí si $w_{\alpha/2} = -w_\alpha \iff \alpha/2 = 1 - \alpha \iff \alpha = 2/3$. Por tanto,

$$\lim_{\alpha \to 2/3^-} \delta_\alpha = \infty \quad \text{y} \quad \lim_{\alpha \to 2/3^+} \delta_\alpha = -\infty \ ,$$

lo que asegura que para todo $\alpha \in (0, 2/3)$, $\mathrm{SE}^*_{\widehat{\theta}} > 0$, y que para todo $\alpha \in (2/3, 1)$, $\mathrm{SE}^*_{\widehat{\theta}} < 0$, respectivamente. El hecho de que el error estándar de un parámetro es siempre positivo restringe el dominio de α a $(0, 2/3)$. En cuanto al dominio de δ_α, su ínfimo viene dado por

$$\lim_{\alpha \to 0^+} \delta_\alpha = \frac{\delta}{1 + \underbrace{\lim_{\alpha \to 0^+} \frac{w_\alpha}{w_{\alpha/2}}}_{\text{LEMA 5}}} = \frac{\delta}{2} \ .$$

Así, de acuerdo con el LEMA 5, $\delta_\alpha \in (\delta/2, \infty)$. □

Corolario 4. *Para todo* $\alpha \in (0, 2/3)$, $(\theta_0 - \delta_\alpha, \delta/(w_\alpha + w_{\alpha/2}))$ *y* $(\theta_0 + \delta_\alpha, \delta/(w_\alpha + w_{\alpha/2}))$ *son los puntos de intersección de las líneas representadas por las* ECUACIONES *3.6 y 3.7.*

Demostración. Las ECUACIONES 3.6 y 3.7 son las partes izquierda y derecha de la ECUACIÓN 3.8, respectivamente. Sustituyendo $\widehat{\theta}$ por $\theta_0 - \delta_\alpha$ o $\theta_0 + \delta_\alpha$ en cualquiera de ellas, por ejemplo en la ECUACIÓN 3.6,

$$\mathrm{SE}^*_{\widehat{\theta}} = \frac{|\pm \delta_\alpha|}{w_{\alpha/2}} = \frac{\delta}{w_{\alpha/2}} \left(1 + \frac{w_\alpha}{w_{\alpha/2}}\right)^{-1} = \frac{\delta}{w_\alpha + w_{\alpha/2}} \ ,$$

se tiene que $(\theta_0 - \delta_\alpha, \mathrm{SE}^*_{\widehat{\theta}})$ y $(\theta_0 + \delta_\alpha, \mathrm{SE}^*_{\widehat{\theta}})$ son los puntos de intersección de las líneas representadas por las ECUACIONES 3.6 y 3.7. □

La relevancia del TEOREMA 3 y su corolario se sustenta en el hecho de que prescriben los límites de la equivalencia coherente (véanse las curvas de puntos en la FIGURA 3.3). El siguiente lema es primordial para la demostración de este teorema.

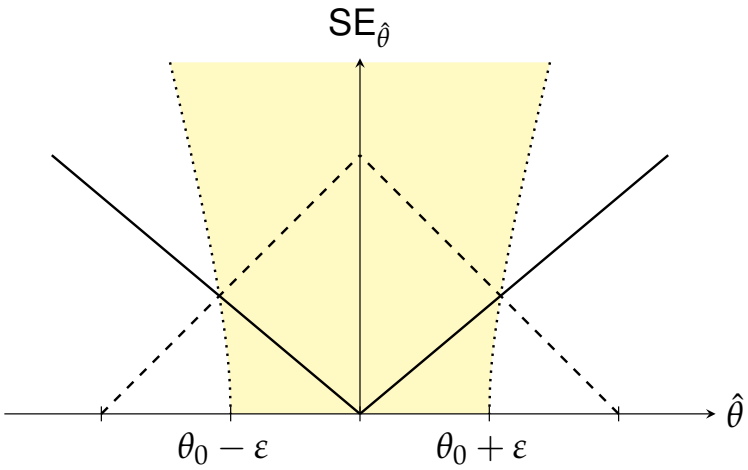

FIGURA 3.3. Límites de la equivalencia coherente. (*Línea continua*) Gráfica de la ECUACIÓN 3.6 a nivel α, para $\alpha \in \mathbb{R}$ tal que $0 < \alpha < 1/2$. (*Línea discontinua*) Gráfica de la ECUACIÓN 3.7 a nivel α. (*Línea de puntos*) Curvas definidas por los puntos de intersección de las líneas representadas por las ECUACIONES 3.6 y 3.7 a todos los niveles $\alpha \in (0, 2/3)$, para los cuales $p_{\mathrm{d}} = p_{\mathrm{e}}$. (*Sombreado*) Conjunto de valores de $\widehat{\theta}$ y $\mathrm{SE}_{\widehat{\theta}}$ que llevan a equivalencia coherente a algún nivel α, para los cuales $p_{\mathrm{d}} > p_{\mathrm{e}}$. Nótese que $\varepsilon = \lim_{\alpha \to 0^+} \delta_\alpha = \delta/2$ (véanse el TEOREMA 3 y el LEMA 5).

Lema 5. *Sea $\Phi^{-1}(\cdot)$ la CDF inversa de la distribución normal estándar. Entonces, para todo $a, b \in \mathbb{R}$ tales que $a, b > 0$,*

$$\lim_{p \to 0^+} \frac{\Phi^{-1}(1-ap)}{\Phi^{-1}(1-bp)} = 1 \ .$$

Demostración. Sea w_p el p-ésimo cuantil superior de la distribución de una variable aleatoria continua X, es decir, $w_p = Q_X(1-p)$, tal que la segunda derivada de $Q_X(\cdot)$ existe,

$$\lim_{p \to 0^+} w_p = \infty \ , \quad \lim_{p \to 0^+} f_X(w_p) = 0 \quad \text{y} \quad \frac{\mathrm{d}w_p}{\mathrm{d}p} = \frac{1}{f_X(w_p)} \ ,$$

donde $f_X(\cdot)$ es la función de densidad de probabilidad de la distribución de X. Así, aplicando la regla de L'Hôpital dos veces,

$$\lim_{p \to 0^+} \frac{w_{ap}}{w_{bp}} = \lim_{p \to 0^+} \frac{a f_X(w_{bp})}{b f_X(w_{ap})} = \lim_{p \to 0^+} \frac{g_X(w_{bp})}{g_X(w_{ap})} \ ,$$

donde

$$g_X(x) = -\frac{\mathrm{d}}{\mathrm{d}x} \log f_X(x) = -\frac{1}{f_X(x)} \frac{\mathrm{d}}{\mathrm{d}x} f_X(x) \ .$$

Supóngase que X tiene una distribución χ^2 con 1 grado de libertad. Por tanto,

$$g_X(x) = \frac{x+1}{2x} \Longrightarrow \lim_{p \to 0^+} g_X(w_p) = \frac{1}{2} \ ,$$

de ahí que

$$\lim_{p \to 0^+} \frac{w_{ap}}{w_{bp}} = \cdots = \lim_{p \to 0^+} \underbrace{\frac{w_{bp}+1}{2w_{bp}}}_{g_X(w_{bp})} \underbrace{\frac{2w_{ap}}{w_{ap}+1}}_{1/g_X(w_{ap})} = \frac{1}{2} \times 2 = 1 \ .$$

Así,

$$\lim_{p\to 0^+} \frac{\Phi^{-1}(1-ap)}{\Phi^{-1}(1-bp)} = \lim_{p\to 0^+} \sqrt{\frac{w_{2ap}}{w_{2bp}}} = \sqrt{\lim_{p\to 0^+} \frac{w_{2ap}}{w_{2bp}}} = \sqrt{1} = 1 \ .$$

□

3.2 Coeficiente de coherencia

Waldhoer y Heinzl demostraron que aplicar conjuntamente un TD y un TE (ambos a nivel α) para llevar a cabo una comparación mantiene la probabilidad de error tipo I a α.[164] Sin embargo, es inevitable que aplicar un TDE para llevar a cabo comparaciones múltiples plantee cierto riesgo de cometer falsos positivos. Aunque no se ha derivado un valor de p para la equivalencia coherente, el TEOREMA 5 y su corolario caracterizan el comportamiento de la región de equivalencia coherente (REC, el deltoide amarillo en la FIGURA 3.2) como función de α, lo que basta para proporcionar una corrección por comparaciones múltiples.

Considérese el problema de aplicar m TDE simultáneos que compartan θ_0 y δ en sus formulaciones, para $m \in \mathbb{R}$ tal que $m > 1$, lo que puede tener consecuencias para la probabilidad de error tipo I si otro problema, el de las comparaciones múltiples, se pasa por alto. La corrección de Bonferroni controla la probabilidad de obtener al menos un falso positivo a α aplicando cada uno de los m TDE a nivel α/m,[165] lo que lo hace adecuado para experimentos que carecen de validación independiente.[166]

«Where ligh[t]ning leaps from the *numbulous*»

Las REC para los m TDE simultáneos se ajustan a una de las REC de la FIGURA 3.4 si estos se aplican, o bien no ajustando por comparaciones múltiples, o bien usando la corrección de Bonferroni. La superposición de ambas regiones revela dos conjuntos disjuntos de valores de $\widehat{\theta}$ y $\mathrm{SE}_{\widehat{\theta}}$ que llevan a equivalencia coherente a nivel α, cuya separación radica en si sus elementos también llevan a equivalencia coherente a nivel α/m o no (en amarillo más oscuro y más claro, respectivamente, en la parte inferior de la FIGURA 3.4).

Para los elementos del conjunto amarillo oscuro, incluidos en los dos REC, la parte derecha de la ECUACIÓN 3.8 se cumple al más estricto nivel α/m. La severidad en la parte izquierda de la ecuación es, sin embargo, lo que causa que ciertos valores de $\widehat{\theta}$ y $\mathrm{SE}_{\widehat{\theta}}$ salten del resultado de incoherencia al de equivalencia coherente a medida que el nivel al que se aplican los m TDE, $\widetilde{\alpha}$, desciende de α a α/m (azul oscuro si $\widetilde{\alpha} = \alpha/m$ y azul claro si $\alpha/m < \widetilde{\alpha} < \alpha$ en la parte inferior de la FIGURA 3.4), de ahí la cita, arriba, de *Finnegans Wake*, de James Joyce.

¿Cuántos TDE simultáneos se tienen que aplicar para que un resultado incoherente a nivel $\widetilde{\alpha} = \alpha$ salte de la incoherencia? Recuérdese que las líneas continuas a lo largo de este capítulo, que gobiernan los límites entre equivalencia coherente e incoherencia, representan los valores de $\widehat{\theta}$ y $\mathrm{SE}_{\widehat{\theta}}$ para los que $p_{\mathrm{d}} = \widetilde{\alpha}$. Así, comparar p_{d} con α responde a la pregunta:

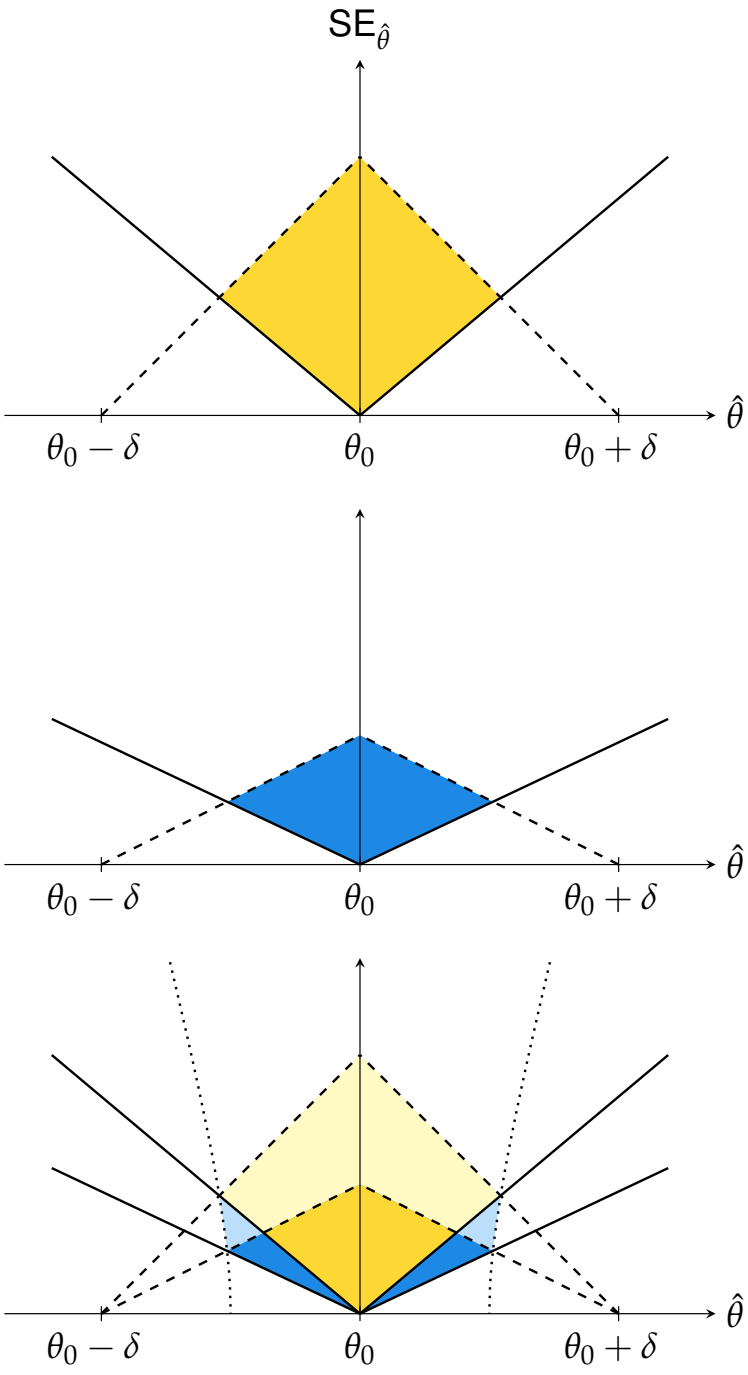

FIGURA 3.4. La región de equivalencia coherente como función de α. (*Parte superior*) Región de equivalencia coherente para un TDE aplicado a nivel α, para $\alpha \in \mathbb{R}$ tal que $0 < \alpha < 1/2$. (*Parte central*) Región de equivalencia coherente para el mismo TDE aplicado a nivel α/m, para $m \in \mathbb{R}$ tal que $m > 1$. (*Parte inferior*) Representación de la estela que la región de equivalencia coherente deja tras de sí a medida que el nivel al que se aplica el TDE desciende desde α hasta α/m.

un resultado incoherente a nivel α permanece fuera de las primeras $c-1$ de las m REC, cada una a nivel α/k, para todo $k \in \mathbb{N}$ tal que $1 \leq k \leq m$, donde $c = \text{inc}(p_\text{d})$ (véase la ECUACIÓN 3.9a).

¿Y cuántos TDE simultáneos se tienen que aplicar para que un resultado de equivalencia coherente (o de incoherencia) a nivel $\widetilde{\alpha} = \alpha$ salte a la indeterminación? En este caso, las líneas discontinuas a lo largo de este capítulo, que gobiernan los límites entre equivalencia coherente e indeterminación, representan los valores de $\widehat{\theta}$ y $\text{SE}_{\widehat{\theta}}$ para los que $p_\text{e} = \widetilde{\alpha}$. Así, comparar p_e con α responde a la pregunta: un resultado de equivalencia coherente (o de incoherencia) a nivel α permanece fuera de las últimas $m-d$ de las m REC, cada una a nivel α/k, donde $d = \text{ind}(p_\text{e})$ (véase la ECUACIÓN 3.9b).

$$\text{inc}(p) = \frac{\alpha}{\text{mín}\,\{p, \alpha\}} = \begin{cases} \alpha/p & \text{si } p < \alpha \ , \\ 1 & \text{si } p \geq \alpha \ ; \end{cases} \tag{3.9a}$$

$$\text{ind}(p) = \frac{\alpha}{\text{máx}\,\{p, \alpha/m\}} = \begin{cases} \alpha/p & \text{si } p > \alpha/m \ , \\ m & \text{si } p \leq \alpha/m \ . \end{cases} \tag{3.9b}$$

De esta manera, ξ, el número de REC, cada una a nivel α/k, que contienen cualquier elemento de los cuatro conjuntos coloreados en la parte inferior de la FIGURA 3.4, es igual a $m - ((c-1) + (m-d)) = d - c + 1$. La combinación de los casos contemplados en las ECUACIONES 3.9a y 3.9b determina las condiciones bajo las que $1 \leq \xi \leq m$: $p_\text{d} \geq p_\text{e}$,

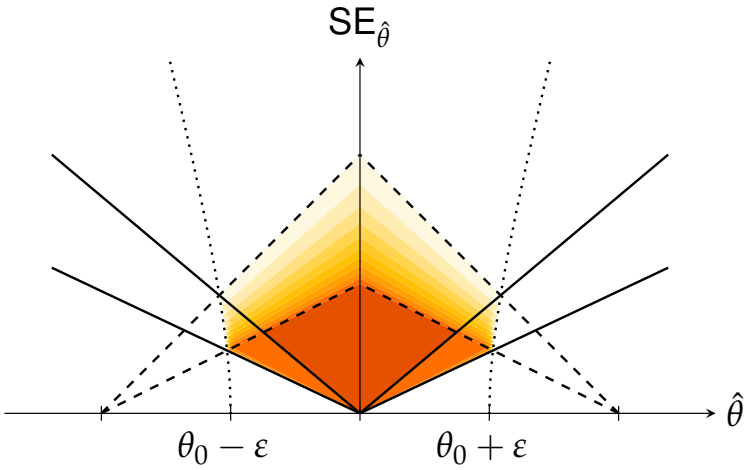

FIGURA 3.5. Coeficiente de coherencia como función de ξ. Para los valores de $\widehat{\theta}$ y $\mathrm{SE}_{\widehat{\theta}}$ que llevan a equivalencia coherente a algún nivel entre α/m y α, sombreados en tonos de naranja, $\log_m \xi$ va desde 0 (*más claro*) hasta 1 (*más oscuro*), que se corresponden con equivalencia coherente débil y fuerte, respectivamente.

$p_{\mathrm{d}} \geq \alpha/m$, $p_{\mathrm{e}} \leq \alpha$ y $m \geq 1$. Estas delimitan los valores de $\widehat{\theta}$ y $\mathrm{SE}_{\widehat{\theta}}$ que llevan a equivalencia coherente a algún nivel entre α/m y α, por lo que el coeficiente de coherencia, $\log_m \xi$, sirve como medida de hasta qué punto la equivalencia coherente prevalece sobre la incoherencia y la indeterminación para cualquier elemento de los cuatro conjuntos coloreados mencionados anteriormente (véase la FIGURA 3.5).

Diferencia coherente

Por claridad lectora, la discusión sobre coherencia se ha centrado en la faceta del concepto que tiene que ver con la equivalencia, en detrimento de la que se refiere a la diferencia. No obstante, de haber optado por la segunda como enfoque de la explicación, apenas ningún concepto

se habría visto alterado. La región de diferencia coherente (RDC), por ejemplo, se define como el conjunto de valores de $\widehat{\theta}$ y $\mathrm{SE}_{\widehat{\theta}}$ tal que

$$\frac{\delta - |\widehat{\theta} - \theta_0|}{w_\alpha} \leq \mathrm{SE}_{\widehat{\theta}} < \frac{|\widehat{\theta} - \theta_0|}{w_{\alpha/2}} \tag{3.10}$$

(compárese con la ECUACIÓN 3.8), de lo cual surge el siguiente resultado, cuyo corolario coincide con el del TEOREMA 3.

Teorema 6. *Si la* ECUACIÓN *3.10 se cumple,* $\widehat{\theta} \notin [\theta_0 - \delta_\alpha, \theta_0 + \delta_\alpha]$*, para todo* $\alpha \in (0, 2/3)$*, donde*

$$\delta_\alpha = \delta \left(1 + \frac{w_\alpha}{w_{\alpha/2}}\right)^{-1} > \frac{\delta}{2}$$

y w_α *y* $w_{\alpha/2}$ *son el* α*-ésimo y el* $(\alpha/2)$*-ésimo cuantiles superiores de la distribución normal estándar.*

Demostración. Comparando las partes izquierda y derecha de la ECUACIÓN 3.10,

$$\frac{|\widehat{\theta} - \theta_0|}{w_{\alpha/2}} > \frac{\delta - |\widehat{\theta} - \theta_0|}{w_\alpha} \iff |\widehat{\theta} - \theta_0| > \delta \left(1 + \frac{w_\alpha}{w_{\alpha/2}}\right)^{-1},$$

se tiene que $\widehat{\theta} \notin [\theta_0 - \delta_\alpha, \theta_0 + \delta_\alpha]$ si la ecuación se cumple. El resto de la demostración es idéntica a la del TEOREMA 3. □

Asimismo, ζ, el número de RDC, cada una a nivel α/k, que contienen cualquier elemento de los cuatro conjuntos análogos a los coloreados en la parte inferior de la FIGURA 3.4, es igual a $\mathrm{ind}(p_\mathrm{d}) - \mathrm{inc}(p_\mathrm{e}) + 1$ (compárese con $\xi = \mathrm{ind}(p_\mathrm{e}) - \mathrm{inc}(p_\mathrm{d}) + 1$). Así, $\log_m \zeta$ puede incorporarse al

coeficiente de coherencia, $C(p_{\mathrm{d}}, p_{\mathrm{e}})$, como medida de hasta qué punto la diferencia coherente prevalece sobre la incoherencia y la indeterminación para cualquier elemento de los conjuntos mencionados anteriormente (véase la ECUACIÓN 3.11, y nótese que el signo menos permite distinguir entre $\log_m \xi$ y $\log_m \zeta$).

$$C(p_{\mathrm{d}}, p_{\mathrm{e}}) = \begin{cases} \log_m \xi & \text{si } \alpha/m \leq p_{\mathrm{d}} \geq p_{\mathrm{e}} \leq \alpha \ , \\ -\log_m \zeta & \text{si } \alpha/m \leq p_{\mathrm{e}} \geq p_{\mathrm{d}} \leq \alpha \ , \\ 0 & \text{en otro caso} \ . \end{cases} \tag{3.11}$$

3.3 Resultados

Wellek, Goddard y Ziegler[167] ilustraron su test de equilibrio de Hardy-Weinberg analizando los controles de 39 estudios de asociación con SNP recogidos en una revisión bibliográfica.[168] En esta sección, se amplía el alcance de su análisis teniendo en cuenta tanto el problema de las comparaciones múltiples como el concepto de coherencia.

El equilibrio de Hardy-Weinberg

El equilibrio de Hardy-Weinberg (EHW) es un principio que constituye uno de los criterios de calidad fundamentales de los SNP de un GWAS. Dado un SNP con dos alelos, A y B, con frecuencias alélicas π_{A} y π_{B}, las

frecuencias genotípicas, π_{AA}, π_{AB} y π_{BB}, están en EHW si, bajo ciertas suposiciones, $\pi_{AA} = \pi_A^2$, $\pi_{AB} = 2\pi_A\pi_B$ y $\pi_{BB} = \pi_B^2$.[169]

El enfoque convencional con que se evalúa el EHW en un SNP concreto se centra en demostrar el *desequilibrio*,[170] para lo que se necesita un TD de hipótesis nula de no diferencia. Mientras rechazar esta hipótesis nula lleva a la conclusión de que las frecuencias genotípicas *no* están en EHW, no ser capaz de rechazarla no implica que las frecuencias genotípicas realmente estén en EHW.

Wellek, Goddard y Ziegler observaron que $\omega_{AB} = 2\sqrt{\pi_{AA}\pi_{BB}}$ es la frecuencia genotípica esperada de AB si π_{AA}, π_{AB} y π_{BB} están en EHW, de donde $\theta = \log(\pi_{AB}/\omega_{AB})$ surge como medida del exceso de frecuencia genotípica de AB. Así, puede aplicarse un TE para demostrar *equilibrio*, con $\theta_0 = 0$, $\delta = \log(1 + 2/5)$ y

$$\sigma_{\widehat{\theta}} = \sqrt{\frac{1}{n}\left(\frac{1}{4\pi_{AA}} + \frac{1}{\pi_{AB}} + \frac{1}{4\pi_{BB}}\right)} \ . \tag{3.12}$$

«Los treinta y nueve TDE»

La TABLA 3.1 muestra, en orden ascendente de $\widehat{\theta}$, el número de genotipos AA, AB y BB, la estimación y el error estándar del exceso de frecuencia genotípica de AB, $\widehat{\theta}$ y $SE_{\widehat{\theta}}$, y el valor del coeficiente de coherencia, C, de cada una de las 39 muestras, resaltadas en la tabla y presentadas en la FIGURA 3.6 si $SE_{\widehat{\theta}} < 0{,}3$.

TABLA 3.1. Valores del coeficiente de coherencia (C) para los 39 TDE.

PMID	AA	AB	BB	$\widehat{\theta}$	$SE_{\widehat{\theta}}$	C
11156391	79	10	3	–1.12	0.43	–0.46
11124296	95	38	33	–1.08	0.19	–1.00
12631667	5	22	122	–0.81	0.31	–0.45
10792336	231	16	1	–0.64	0.56	0.00
11468325	180	40	8	–0.64	0.24	–0.51
10869806	34	22	12	–0.61	0.27	–0.19
10680782	221	64	15	–0.59	0.18	–0.99
11781417	42	32	16	–0.48	0.23	–0.09
10189842	21	47	62	–0.43	0.19	–0.18
11027931	148	33	4	–0.39	0.31	0.00
10843185	3	23	75	–0.27	0.36	0.00
12753258	75	35	6	–0.19	0.27	0.00
11045785	61	212	237	–0.13	0.10	0.29
9844142	66	80	31	–0.12	0.16	0.00
10794488	297	258	71	–0.12	0.09	0.50
11097227	57	192	197	–0.10	0.10	0.41
10712418	197	349	186	–0.09	0.07	1.00
12105308	58	97	48	–0.08	0.14	0.08
10231446	77	41	6	–0.05	0.26	0.00
11122322	193	393	215	–0.04	0.07	1.00
11402126	98	233	144	–0.02	0.09	1.00
12787424	23	83	67	0.06	0.16	0.04
10964048	62	65	15	0.06	0.19	0.00
11786085	16	76	79	0.07	0.18	0.00
11142420	112	132	33	0.08	0.13	0.17
12675860	27	26	5	0.11	0.31	0.00
11889073	114	560	533	0.13	0.07	1.00
11914402	174	45	2	0.19	0.39	0.00
12221172	98	77	9	0.26	0.21	0.00
10430441	154	203	36	0.31	0.12	–0.51
11231353	21	60	23	0.31	0.20	0.00
10504487	29	123	68	0.33	0.14	–0.22
12034804	15	107	95	0.35	0.17	–0.07
10781645	43	150	62	0.37	0.13	–0.71
15338456	63	83	12	0.41	0.19	–0.12
11074789	260	124	6	0.45	0.23	–0.03
10815136	3	24	18	0.49	0.37	0.00
10698474	44	127	29	0.58	0.15	–1.00
9607207	8	48	19	0.67	0.26	–0.46

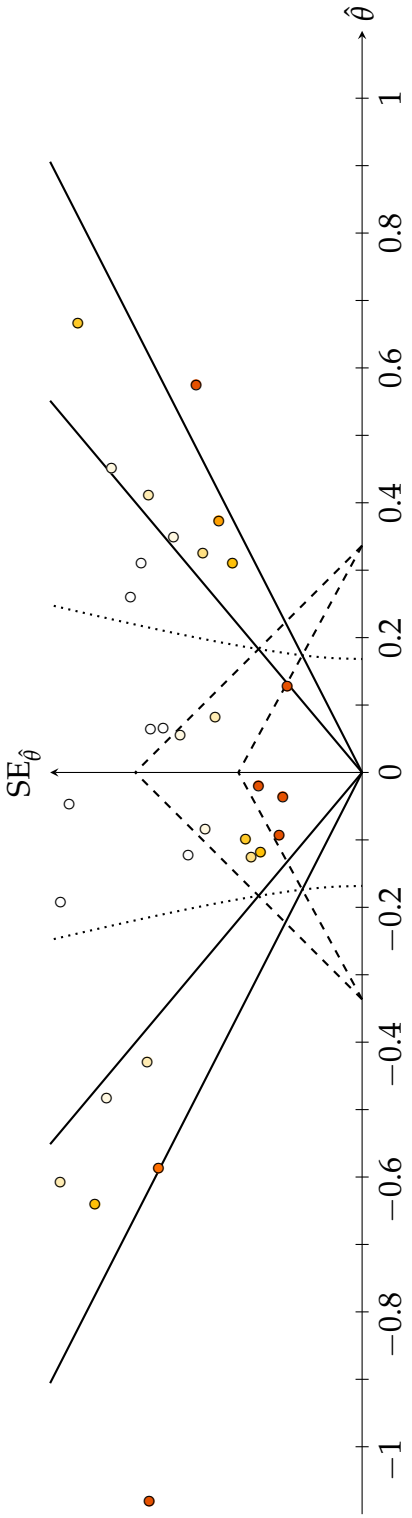

FIGURA 3.6. Diagrama de dispersión de los valores resaltados en la TABLA 3.1. (*Línea continua*) Gráficas de la ECUACIÓN 3.6 a niveles α y α/m, para $\alpha = 0{,}05$, $m = 39$ y $\theta_0 = 0$. (*Línea discontinua*) Gráficas de la ECUACIÓN 3.7 a niveles α y α/m, para $\alpha = 0{,}05$, $m = 39$, $\theta_0 = 0$ y $\delta = \log(1 + 2/5)$. (*Línea de puntos)* Curvas que delimitan el signo de $p_\mathrm{d} - p_\mathrm{e}$, para todo $p_\mathrm{d}, p_\mathrm{e} \in [0, 1]$: entre ellas, $p_\mathrm{d} - p_\mathrm{e} > 0$; más allá de ellas, $p_\mathrm{d} - p_\mathrm{e} < 0$. (*Sombreado*) Colores de los valores de C en la TABLA 3.1 (véase la FIGURA 3.5).

Para cada fila de la tabla, p_{d} y p_{e} vienen dados por la ECUACIÓN 3.4 y la PROPOSICIÓN 2, respectivamente, con $\widehat{\theta} = \log\left(n_{\mathrm{AB}} / \left(2\sqrt{n_{\mathrm{AA}} n_{\mathrm{BB}}}\right)\right)$, $\theta_0 = 0$, $\delta = \log(1 + 2/5)$, $\alpha = 0{,}05$, $m = 39$ y

$$\mathrm{SE}_{\widehat{\theta}} = \sqrt{\frac{1}{4n_{\mathrm{AA}}} + \frac{1}{n_{\mathrm{AB}}} + \frac{1}{4n_{\mathrm{BB}}}} \ ,$$

donde $n_{\mathrm{AA}} + n_{\mathrm{AB}} + n_{\mathrm{BB}} = n$ y $\mathrm{SE}_{\widehat{\theta}}$ es un estimador consistente de $\sigma_{\widehat{\theta}}$ (compárese con la ECUACIÓN 3.12), lo que implica que tanto T_{d} como T_{e} convergen en distribución a una variable aleatoria normal estándar.

Dos pares de valores de $\widehat{\theta}$ y $\mathrm{SE}_{\widehat{\theta}}$ en la TABLA 3.1 se corresponden con desequilibrio a nivel α/m (véanse las filas en que $C = -1$). ¿Puede probarse la falacia que asume que el resto de muestras están en equilibrio? El signo de $p_{\mathrm{d}} - p_{\mathrm{e}}$, del que depende el de C, permite distinguir los pares de valores de $\widehat{\theta}$ y $\mathrm{SE}_{\widehat{\theta}}$ que no pueden dar equivalencia coherente como resultado a ningún nivel, para los que $p_{\mathrm{d}} - p_{\mathrm{e}} < 0$ (véanse las primeras 10 y las últimas 11 filas de la tabla), de los que pueden a algún nivel, para los que $p_{\mathrm{d}} - p_{\mathrm{e}} > 0$. Aquellos dejan la falacia en entredicho —14 de los 21 pares dan como resultado diferencia coherente a algún nivel entre α/m y α— y estos, la que asume que no ser capaz de demostrar equilibrio significa demostrar desequilibrio.

Capítulo

4 Test de diferencia-equivalencia para la selección de modelos genéticos

En este capítulo, se diseña una prueba estadística en el marco de trabajo de los test de diferencia-equivalencia que permite decidir qué modelo genético le corresponde —si es que le corresponde alguno— a un SNP o una interacción SNP-SNP.

Para el contraste de hipótesis de diferencia-equivalencia, al contrario que para el de diferencia y al igual que para el de equivalencia, no solo debe indicarse la forma de construir el estadístico de contraste, sino también definirse el margen de equivalencia, δ.[141] Esta definición tiene un gran impacto en la potencia estadística y el tamaño muestral necesarios: si el valor de δ es muy elevado, cualquier declaración de equivalencia coherente carecerá de credibilidad; si, por el contrario, es demasiado bajo, apenas se producirán dichos resultados de equivalencia coherente.[142] Para expresar el valor de δ, pueden emplearse las unidades del parámetro poblacional o de magnitudes derivadas de este, aunque se desaconsejan aquellas en que intervienen estimaciones de la varianza.[133,144]

4.1 Parámetros poblacionales y valor de δ

Sean, en el marco de un estudio de casos y controles, X un SNP y $R_x = \Pr(\text{efecto} \mid X = x)$, donde $x \in \{0, 1, 2\}$, es decir, el riesgo de ser caso en función del genotipo.[♯] La relación entre los riesgos relativos (RR) $\text{RR}_1 = R_1/R_0$ y $\text{RR}_2 = R_2/R_0$, siempre que $R_0 \neq 0$, permite abordar el problema de comprobar si un modelo genético dado rige la asociación SNP-enfermedad.[171] Así, el modelo aditivo, que indica que los riesgos relativos entre AB y AA ($\text{RR}_1 = R_1/R_0$) y entre BB y AB (no $\text{RR}_2 = R_2/R_0$, sino R_2/R_1) son iguales, se da si $\text{RR}_1 \neq 1$, $\text{RR}_2 \neq 1$ y

$$\frac{R_1}{R_0} = \frac{R_2}{R_1} \iff \frac{R_1^2}{R_0} = R_2 \iff \left(\frac{R_1}{R_0}\right)^2 = \frac{R_2}{R_0} \iff \text{RR}_1^2 = \text{RR}_2 \ ;$$

el dominante, si $\text{RR}_1 \neq 1$, $\text{RR}_2 \neq 1$ y $\text{RR}_1 = \text{RR}_2$; el sobredominante, si $\text{RR}_2 = 1$ y $\text{RR}_1 \neq \text{RR}_2$; y el recesivo, si $\text{RR}_1 = 1$ y $\text{RR}_1 \neq \text{RR}_2$.

Parámetros poblacionales lineales

La FIGURA 4.1 muestra la relación entre los RR de los genotipos AB y BB con respecto al homocigoto más frecuente, AA, por modelo genético. De esta manera, las curvas representan aquellos pares de valores que evidencian un patrón clásico en la asociación SNP-enfermedad. Sería deseable que la intersección de las cuatro curvas dividiera el plano en

[♯] Recuérdese que un SNP se codifica de acuerdo con cuántos alelos menores presenta la suma de sus dos alelos, A (el mayor) y B (el menor), por lo que 0, 1 y 2 representan los genotipos AA, AB y BB, respectivamente.

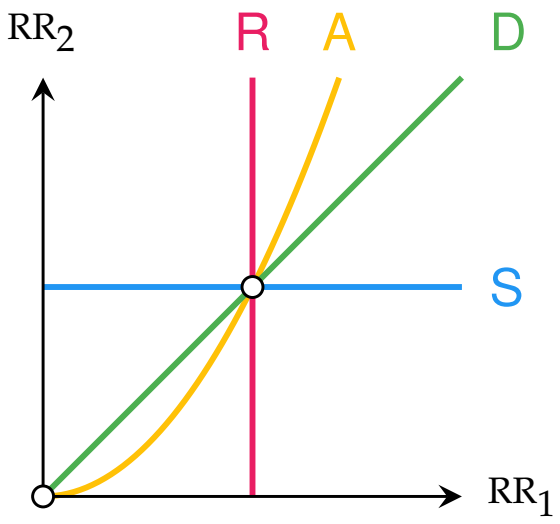

FIGURA 4.1. Relación entre los riesgos relativos por modelo genético. Las curvas A ($y = x^2$), D ($y = x$), S ($y = 1$) y R ($x = 1$), donde $x = \mathrm{RR}_1$ e $y = \mathrm{RR}_2$, representan la relación entre los riesgos relativos en los modelos aditivo, dominante, sobredominante y recesivo, respectivamente.

regiones abiertas, pero el carácter no lineal del modelo aditivo —la relación entre RR_1 y RR_2 forma una parábola— lo dificulta. Esto hace imposible establecer un valor de δ adecuado (y, por tanto, de seleccionar un modelo genético de entre dos) en regiones como la que dejan entre sí los modelos aditivo y dominante para RR_1 y RR_2 menores que 1.

La transformación logarítmica de los RR no solo permite obtener relaciones lineales entre ellos, sino que es conveniente, además, en el marco de un estudio de casos y controles. Nótese que, a diferencia de en los estudios de cohortes, en los de casos y controles no pueden calcularse tasas de incidencia ni, por tanto, riesgos o RR;[86] en cambio, es posible derivar, a partir de los individuos expuestos en los grupos de casos y de controles, *odds ratios* (OR), que mejor estiman el RR cuanto menor es la tasa de incidencia del efecto,[87] y cuyo logaritmo se corresponde con el

coeficiente del término principal del modelo de regresión logística. Con la estimación de β_1 y β_2 en

$$\log\left(\frac{\mathrm{E}[Y]}{1-\mathrm{E}[Y]}\right) = \beta_0 + \beta_1 \mathbb{1}_{\{X=1\}} + \beta_2 \mathbb{1}_{\{X=2\}} \tag{4.1}$$

se obtiene, pues, la de las transformaciones logarítmicas de RR_1 y RR_2, respectivamente, siempre que la tasa de incidencia del efecto sea baja.[172] Así, el modelo aditivo se da si $\beta_1 \neq 0$, $\beta_2 \neq 0$ y

$$\begin{aligned} \log\left(\mathrm{OR}_1^2\right) = \log\left(\mathrm{OR}_2\right) &\iff \log\left(\exp\left\{\beta_1\right\}^2\right) = \log\left(\exp\left\{\beta_2\right\}\right) \\ &\iff 2\beta_1 - \beta_2 = 0 \ ; \end{aligned} \tag{4.2}$$

el dominante, si $\beta_1 \neq 0$, $\beta_2 \neq 0$ y

$$\log\left(\mathrm{OR}_1\right) = \log\left(\mathrm{OR}_2\right) \iff \beta_1 - \beta_2 = 0 \ ; \tag{4.3}$$

el sobredominante, si $\beta_1 \neq \beta_2$ y

$$\log\left(\mathrm{OR}_2\right) = \log 1 \iff \beta_2 = 0 \ ; \tag{4.4}$$

y el recesivo, si $\beta_1 \neq \beta_2$ y

$$\log\left(\mathrm{OR}_1\right) = \log 1 \iff \beta_1 = 0 \ . \tag{4.5}$$

La FIGURA 4.2 muestra la relación entre los coeficientes β_1 y β_2 de la ECUACIÓN 4.1 por modelo genético (véanse las ECUACIONES 4.2-4.5), que pueden representarse, incluida la del modelo aditivo, por rectas. Estas cuatro rectas se cruzan en el mismo punto, por lo que subdividen el

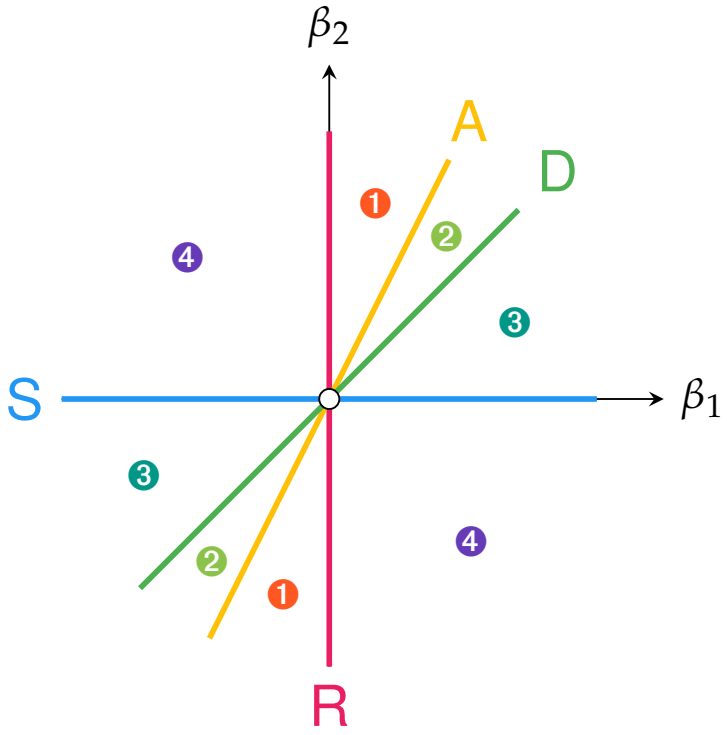

FIGURA 4.2. Relación entre los logaritmos de los riesgos relativos por modelo genético. Las rectas A ($y = 2x$), D ($y = x$), S ($y = 0$) y R ($x = 0$), donde $x = \beta_1$ e $y = \beta_2$, representan la relación entre las transformaciones logarítmicas de RR_1 y RR_2 en los modelos aditivo, dominante, sobredominante y recesivo.

plano en 8 regiones abiertas, en cada una de las cuales dos modelos genéticos compiten por regir la asociación SNP-enfermedad. Por ejemplo, la marca ❶ indica las regiones en que los puntos (β_1, β_2) pueden evidenciar el modelo aditivo o el recesivo. Otra forma de interpretar la figura es la siguiente: cuanto más cercanos sean los puntos de las regiones ❶ y ❷ a la recta A, mayor probabilidad de que se trate de un modelo aditivo, mientras que, según se alejen hacia uno u otro lado, la probabilidad irá a favor del modelo recesivo o del dominante. Esto significa que cualquier punto de las regiones ❸ y ❹ nunca podrá representar el modelo aditivo, lo cual debe controlarse a través del valor de δ.

Parámetros poblacionales no lineales con δ común

En principio, dado que ningún par de regiones adyacentes a un modelo genético es igual a otro en tamaño, debería definirse un valor de δ para cada modelo genético. Sin embargo, si se aplica una transformación no lineal sencilla, como el cociente, a las ECUACIONES 4.2-4.5, puede obtenerse un valor de δ único. Para ello, selecciónese, como parámetro poblacional de un modelo genético, el cociente de la parte izquierda de su ecuación entre la parte izquierda de la del modelo con que no comparte regiones adyacentes. De esta manera, el parámetro poblacional del modelo aditivo es

$$\theta_{\mathrm{A}} = f_{\mathrm{A}}(\beta_1, \beta_2) = \frac{2\beta_1 - \beta_2}{\beta_2} \ , \tag{4.6}$$

donde el numerador es la parte izquierda de la ECUACIÓN 4.2 y el denominador, la del modelo sobredominante, con el que no comparte regiones adyacentes. Obsérvese que $\theta_{\mathrm{A}} = 0$ si el modelo es aditivo. Sustituyendo β_1 y β_2 por cualquier punto del modelo dominante ($\beta_1 = \beta_2$) y del recesivo ($\beta_1 = 0$), cuyas rectas delimitan las regiones adyacentes al modelo aditivo, se tiene, respectivamente, que

$$\theta_{\mathrm{A}} = \frac{2\beta_2 - \beta_2}{\beta_2} = 1 \quad \text{y} \quad \theta_{\mathrm{A}} = \frac{2 \cdot 0 - \beta_2}{\beta_2} = -1 \ ,$$

por lo que $\delta = 1$ es la opción adecuada para un TDE con el parámetro poblacional de la ECUACIÓN 4.6.

Lo interesante de este método de construcción de los parámetros poblacionales de los modelos genéticos es que, como puede comprobarse a continuación, siempre da lugar a un valor de δ igual a 1.

El parámetro poblacional del modelo dominante es

$$\theta_{\mathrm{D}} = f_{\mathrm{D}}(\beta_1, \beta_2) = \frac{\beta_1 - \beta_2}{\beta_1} \ , \tag{4.7}$$

donde el numerador es la parte izquierda de la ECUACIÓN 4.3 y el denominador, la del modelo recesivo, con el que no comparte regiones adyacentes. Obsérvese que $\theta_{\mathrm{D}} = 0$ si el modelo es dominante. Sustituyendo β_1 y β_2 por cualquier punto del modelo sobredominante ($\beta_2 = 0$) y del aditivo ($2\beta_1 = \beta_2$), cuyas rectas delimitan las regiones adyacentes al modelo dominante, se tiene, respectivamente, que

$$\theta_{\mathrm{D}} = \frac{\beta_1 - 0}{\beta_1} = 1 \quad \text{y} \quad \theta_{\mathrm{D}} = \frac{\beta_1 - 2\beta_1}{\beta_1} = -1 \ .$$

El parámetro poblacional del modelo sobredominante es

$$\theta_{\mathrm{S}} = f_{\mathrm{S}}(\beta_1, \beta_2) = \frac{\beta_2}{2\beta_1 - \beta_2} \ , \tag{4.8}$$

donde el numerador es la parte izquierda de la ECUACIÓN 4.4 y el denominador, la del modelo aditivo, con el que no comparte regiones adyacentes. Obsérvese que $\theta_{\mathrm{S}} = 0$ si el modelo es sobredominante. Sustituyendo β_1 y β_2 por cualquier punto del modelo dominante ($\beta_1 = \beta_2$) y del recesivo ($\beta_1 = 0$), cuyas rectas delimitan las regiones adyacentes al modelo sobredominante, se tiene, respectivamente, que

$$\theta_S = \frac{\beta_2}{2\beta_2 - \beta_2} = 1 \quad \text{y} \quad \theta_S = \frac{\beta_2}{2 \cdot 0 - \beta_2} = -1 \ .$$

El parámetro poblacional del modelo recesivo es

$$\theta_R = f_R(\beta_1, \beta_2) = \frac{\beta_1}{\beta_1 - \beta_2} \ , \tag{4.9}$$

donde el numerador es la parte izquierda de la ECUACIÓN 4.5 y el denominador, la del modelo dominante, con el que no comparte regiones adyacentes. Obsérvese que $\theta_R = 0$ si el modelo es recesivo. Sustituyendo β_1 y β_2 por cualquier punto del modelo sobredominante ($\beta_2 = 0$) y del aditivo ($2\beta_1 = \beta_2$), cuyas rectas delimitan las regiones adyacentes al modelo recesivo, se tiene, respectivamente, que

$$\theta_R = \frac{\beta_1}{\beta_1 - 0} = 1 \quad \text{y} \quad \theta_R = \frac{\beta_1}{\beta_1 - 2\beta_1} = -1 \ .$$

4.2 Contraste de hipótesis

Dados un vector de estimaciones de 2 parámetros, $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2)$, y una transformación que se le aplica, $f : \mathbb{R}^2 \to \mathbb{R}$, la varianza estimada de $f(\widehat{\boldsymbol{\beta}})$ viene definida por[173]

$$\widehat{\text{Var}}(f(\widehat{\boldsymbol{\beta}})) = \mathbf{J}\,\boldsymbol{\Sigma}\,\mathbf{J}^\top \ ,$$

donde $\mathbf{J} \in \mathbb{R}^{1\times 2}$ es la matriz jacobiana, es decir,

$$\mathbf{J} = \left(\frac{\partial f(\boldsymbol{\beta})}{\partial \beta_1} \quad \frac{\partial f(\boldsymbol{\beta})}{\partial \beta_2} \right) \Bigg|_{\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}} \ ,$$

y $\boldsymbol{\Sigma} \in \mathbb{R}^{2\times 2}$ es la matriz de covarianza, o sea,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \widehat{\mathrm{Var}}(\beta_1) & \widehat{\mathrm{Cov}}(\beta_1, \beta_2) \\ \widehat{\mathrm{Cov}}(\beta_1, \beta_2) & \widehat{\mathrm{Var}}(\beta_2) \end{pmatrix} \Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} ,$$

por lo que

$$\widehat{\mathrm{Var}}(f(\widehat{\boldsymbol{\beta}})) = J_{11}\left(J_{11}\,\widehat{\mathrm{Var}}(\widehat{\beta}_1) + J_{12}\,\widehat{\mathrm{Cov}}(\widehat{\beta}_1, \widehat{\beta}_2)\right) + J_{12}\left(J_{11}\,\widehat{\mathrm{Cov}}(\widehat{\beta}_1, \widehat{\beta}_2) + J_{12}\,\widehat{\mathrm{Var}}(\widehat{\beta}_2)\right) \quad . \quad (4.10)$$

Las cuatro ECUACIONES 4.6-4.9 constituyen sendas transformaciones que pueden aplicarse a los parámetros β_1 y β_2 una vez estimados, tal y como ocurre con la función f al principio de esta sección. A continuación, se estima la varianza de cada uno de los parámetros poblacionales de los modelos genéticos.

Las matrices jacobianas de los modelos aditivo y sobredominante son, respectivamente,

$$\mathbf{J}_{\mathrm{A}} = \begin{pmatrix} \dfrac{2}{\widehat{\beta}_2} & -\dfrac{2\widehat{\beta}_1}{\widehat{\beta}_2^2} \end{pmatrix} \quad \text{y} \quad \mathbf{J}_{\mathrm{S}} = \begin{pmatrix} -\dfrac{2\widehat{\beta}_2}{(2\widehat{\beta}_1 - \widehat{\beta}_2)^2} & \dfrac{2\widehat{\beta}_1}{(2\widehat{\beta}_1 - \widehat{\beta}_2)^2} \end{pmatrix} ,$$

por lo que sus varianzas estimadas (véase la ECUACIÓN 4.10) son

$$\widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{A}}) = \frac{4\widehat{\gamma}^2}{\widehat{\beta}_2^4} \quad \text{y} \quad \widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{S}}) = \frac{4\widehat{\gamma}^2}{(2\widehat{\beta}_1 - \widehat{\beta}_2)^4} \quad , \quad (4.11)$$

donde $\widehat{\gamma}^2 = \widehat{\beta}_1^2\,\widehat{\mathrm{Var}}(\widehat{\beta}_2) + \widehat{\beta}_2^2\,\widehat{\mathrm{Var}}(\widehat{\beta}_1) - 2\widehat{\beta}_1\widehat{\beta}_2\,\widehat{\mathrm{Cov}}(\widehat{\beta}_1, \widehat{\beta}_2)$, expresión que se repite en las varianzas estimadas de cada uno de los parámetros poblacionales de los modelos genéticos.

Las matrices jacobianas del resto de modelos son

$$\mathbf{J}_{\mathrm{D}} = \left(\frac{\widehat{\beta}_2}{\widehat{\beta}_1^2} \quad -\frac{1}{\widehat{\beta}_1} \right) \text{ y } \mathbf{J}_{\mathrm{R}} = \left(-\frac{\widehat{\beta}_2}{(\widehat{\beta}_1 - \widehat{\beta}_2)^2} \quad \frac{\widehat{\beta}_1}{(\widehat{\beta}_1 - \widehat{\beta}_2)^2} \right) ,$$

por lo que sus varianzas estimadas son

$$\widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{D}}) = \frac{\widehat{\gamma}^2}{\widehat{\beta}_1^4} \text{ y } \widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{R}}) = \frac{\widehat{\gamma}^2}{(\widehat{\beta}_1 - \widehat{\beta}_2)^4} \ . \tag{4.12}$$

Estadísticos de contraste

La variable aleatoria de la que el estadístico

$$t_{\mathrm{d}} = \frac{f(\widehat{\boldsymbol{\beta}}) - \theta_0}{\sqrt{\widehat{\mathrm{Var}}(f(\widehat{\boldsymbol{\beta}}))}} \tag{4.13}$$

es una realización sigue una distribución normal estándar.[174] Aplicar, pues, un test de hipótesis nula de no diferencia $\mathrm{H}_0 : \theta = \theta_0, \mathrm{H}_1 : \theta \neq \theta_0$, donde $\theta = f(\boldsymbol{\beta})$, es sencillo, de cuyo caso especial $\theta_0 = 0$ se deriva el siguiente resultado.

Lema 7. *Si $\theta_0 = 0$, los valores del estadístico t_{d} son iguales para las parejas de modelos genéticos aditivo-sobredominante y dominante-recesivo.*

Demostración. De acuerdo con la ECUACIÓN 4.13, por una parte, para el modelo aditivo,

$$t_{\mathrm{d}} = \frac{\widehat{\theta}_{\mathrm{A}} - 0}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{A}})}} = \frac{\dfrac{2\widehat{\beta}_1 - \widehat{\beta}_2}{\widehat{\beta}_2}}{\dfrac{2\widehat{\gamma}}{\widehat{\beta}_2^2}} = \frac{\widehat{\beta}_2(2\widehat{\beta}_1 - \widehat{\beta}_2)}{2\widehat{\gamma}}$$

y, para el sobredominante,

$$t_{\mathrm{d}} = \frac{\widehat{\theta}_{\mathrm{S}} - 0}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{S}})}} = \frac{\dfrac{\widehat{\beta}_2}{2\widehat{\beta}_1 - \widehat{\beta}_2}}{\dfrac{2\widehat{\gamma}}{(2\widehat{\beta}_1 - \widehat{\beta}_2)^2}} = \frac{\widehat{\beta}_2(2\widehat{\beta}_1 - \widehat{\beta}_2)}{2\widehat{\gamma}} \ .$$

Por otra parte, para el modelo dominante,

$$t_{\mathrm{d}} = \frac{\widehat{\theta}_{\mathrm{D}} - 0}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{D}})}} = \frac{\dfrac{\widehat{\beta}_1 - \widehat{\beta}_2}{\widehat{\beta}_1}}{\dfrac{\widehat{\gamma}}{\widehat{\beta}_1^2}} = \frac{\widehat{\beta}_1(\widehat{\beta}_1 - \widehat{\beta}_2)}{\widehat{\gamma}}$$

y, para el recesivo,

$$t_{\mathrm{d}} = \frac{\widehat{\theta}_{\mathrm{R}} - 0}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_{\mathrm{R}})}} = \frac{\dfrac{\widehat{\beta}_1}{\widehat{\beta}_1 - \widehat{\beta}_2}}{\dfrac{\widehat{\gamma}}{(\widehat{\beta}_1 - \widehat{\beta}_2)^2}} = \frac{\widehat{\beta}_1(\widehat{\beta}_1 - \widehat{\beta}_2)}{\widehat{\gamma}} \ .$$

□

Análogamente, las variables aleatorias de las que los estadísticos

$$\overrightarrow{t_{\mathrm{e}}} = \frac{f(\widehat{\boldsymbol{\beta}}) - (\theta_0 - \delta)}{\sqrt{\widehat{\mathrm{Var}}(f(\widehat{\boldsymbol{\beta}}))}} \quad \text{y} \quad \overleftarrow{t_{\mathrm{e}}} = \frac{f(\widehat{\boldsymbol{\beta}}) - (\theta_0 + \delta)}{\sqrt{\widehat{\mathrm{Var}}(f(\widehat{\boldsymbol{\beta}}))}} \ . \tag{4.14}$$

son realizaciones siguen una distribución normal estándar.[174] Aplicar, pues, un test de hipótesis nula de no superioridad $\overrightarrow{\mathrm{H}}_0 : \theta \leq \theta_0 - \delta, \overrightarrow{\mathrm{H}}_1 : \theta > \theta_0 - \delta$ o de no inferioridad $\overleftarrow{\mathrm{H}}_0 : \theta \geq \theta_0 + \delta, \overleftarrow{\mathrm{H}}_1 : \theta < \theta_0 + \delta$, donde $\theta = f(\boldsymbol{\beta})$, es sencillo, de cuyo caso especial $\theta_0 = 0$ y $\delta = 1$ se derivan los siguientes resultados.

Lema 8. *Si θ_0 y $\delta = 1$, los valores de los estadísticos $\overrightarrow{t_e}$ y $\overleftarrow{t_e}$ son iguales para las parejas de modelos genéticos dominante-sobredominante y aditivo-recesivo, respectivamente.*

Demostración. De acuerdo con la ECUACIÓN 4.14, por una parte, para el modelo dominante,

$$\overrightarrow{t_e} = \frac{\widehat{\theta}_D - (0-1)}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_D)}} = \frac{\dfrac{\widehat{\beta}_1 - \widehat{\beta}_2}{\widehat{\beta}_1} + 1}{\dfrac{\widehat{\gamma}}{\widehat{\beta}_1^2}} = \frac{\widehat{\beta}_1(2\widehat{\beta}_1 - \widehat{\beta}_2)}{\widehat{\gamma}}$$

y, para el sobredominante,

$$\overrightarrow{t_e} = \frac{\widehat{\theta}_S - (0-1)}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_S)}} = \frac{\dfrac{\widehat{\beta}_2}{2\widehat{\beta}_1 - \widehat{\beta}_2} + 1}{\dfrac{2\widehat{\gamma}}{(2\widehat{\beta}_1 - \widehat{\beta}_2)^2}} = \frac{\widehat{\beta}_1(2\widehat{\beta}_1 - \widehat{\beta}_2)}{\widehat{\gamma}} \ .$$

Por otra parte, para el modelo aditivo,

$$\overleftarrow{t_e} = \frac{\widehat{\theta}_A - (0+1)}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_A)}} = \frac{\dfrac{2\widehat{\beta}_1 - \widehat{\beta}_2}{\widehat{\beta}_2} - 1}{\dfrac{2\widehat{\gamma}}{\widehat{\beta}_2^2}} = \frac{\widehat{\beta}_2(\widehat{\beta}_1 - \widehat{\beta}_2)}{\widehat{\gamma}}$$

y, para el recesivo,

$$\overleftarrow{t_e} = \frac{\widehat{\theta}_R - (0+1)}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_R)}} = \frac{\dfrac{\widehat{\beta}_1}{\widehat{\beta}_1 - \widehat{\beta}_2} - 1}{\dfrac{\widehat{\gamma}}{(\widehat{\beta}_1 - \widehat{\beta}_2)^2}} = \frac{\widehat{\beta}_2(\widehat{\beta}_1 - \widehat{\beta}_2)}{\widehat{\gamma}} \ .$$

□

Lema 9. *Si $\theta_0 = 0$ y $\delta = 1$, los valores de los estadísticos $\overrightarrow{t_e}$ y $\overleftarrow{t_e}$ son opuestos, es decir, $\overrightarrow{t_e} = -\overleftarrow{t_e}$, para las parejas de modelos genéticos aditivo-dominante y sobredominante-recesivo.*

Demostración. De acuerdo con la ECUACIÓN 4.14, por una parte, para el modelo aditivo,

$$\overrightarrow{t_e} = \frac{\widehat{\theta}_A - (0-1)}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_A)}} = \frac{\dfrac{2\widehat{\beta}_1 - \widehat{\beta}_2}{\widehat{\beta}_2} + 1}{\dfrac{2\widehat{\gamma}}{\widehat{\beta}_2^2}} = \frac{\widehat{\beta}_1\widehat{\beta}_2}{\widehat{\gamma}}$$

y, para el dominante,

$$\overleftarrow{t_e} = \frac{\widehat{\theta}_D - (0+1)}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_D)}} = \frac{\dfrac{\widehat{\beta}_1 - \widehat{\beta}_2}{\widehat{\beta}_1} - 1}{\dfrac{\widehat{\gamma}}{\widehat{\beta}_1^2}} = -\frac{\widehat{\beta}_1\widehat{\beta}_2}{\widehat{\gamma}} \ .$$

Por otra parte, para el modelo sobredominante,

$$\overleftarrow{t_e} = \frac{\widehat{\theta}_S - (0+1)}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_S)}} = \frac{\dfrac{\widehat{\beta}_2}{2\widehat{\beta}_1 - \widehat{\beta}_2} - 1}{\dfrac{2\widehat{\gamma}}{(2\widehat{\beta}_1 - \widehat{\beta}_2)^2}} = -\frac{(\widehat{\beta}_1 - \widehat{\beta}_2)(2\widehat{\beta}_1 - \widehat{\beta}_2)}{\widehat{\gamma}}$$

y, para el recesivo,

$$\overrightarrow{t_e} = \frac{\widehat{\theta}_R - (0-1)}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}_R)}} = \frac{\dfrac{\widehat{\beta}_1}{\widehat{\beta}_1 - \widehat{\beta}_2} + 1}{\dfrac{\widehat{\gamma}}{(\widehat{\beta}_1 - \widehat{\beta}_2)^2}} = \frac{(\widehat{\beta}_1 - \widehat{\beta}_2)(2\widehat{\beta}_1 - \widehat{\beta}_2)}{\widehat{\gamma}} \ .$$

□

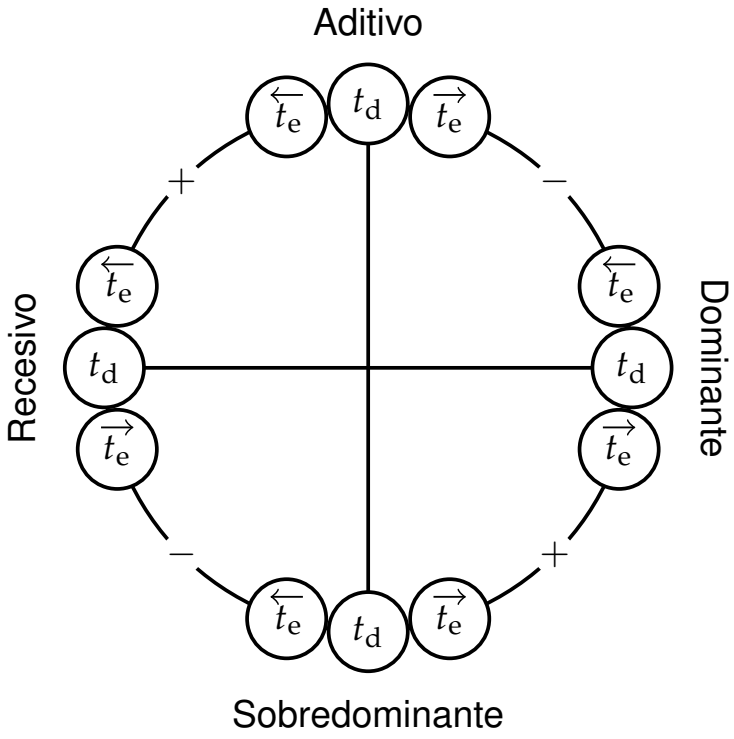

FIGURA 4.3. Relaciones entre los estadísticos de contraste. Las *líneas rectas* y las *curvas con un* $+$ representan relaciones de igualdad, mientras que aquellas *con un* $-$, de simetría con respecto a la suma.

La FIGURA 4.3 recoge gráficamente las relaciones a que hacen referencia los LEMAS 7-9. Nótese que los estadísticos de contraste de cualquier combinación de tres modelos genéticos proporcionan la información necesaria para reconstruir los del modelo genético restante. Por ello, la corrección de Bonferroni, que controla la probabilidad de obtener al menos un falso positivo a α aplicando cada uno de los 4 TDE a nivel $\widetilde{\alpha} = \alpha/4$, sería demasiado conservadora en este caso. Para decidir, pues, qué modelo genético (aditivo, dominante, recesivo o sobredominante) le corresponde —si es que le corresponde alguno— a un SNP o a una interacción SNP-SNP aplicando los 4 TDE simultáneamente, $\widetilde{\alpha} \geq \alpha/3$. La siguiente subsección trata de dilucidar cuán cerca está $\widetilde{\alpha}$ de α.

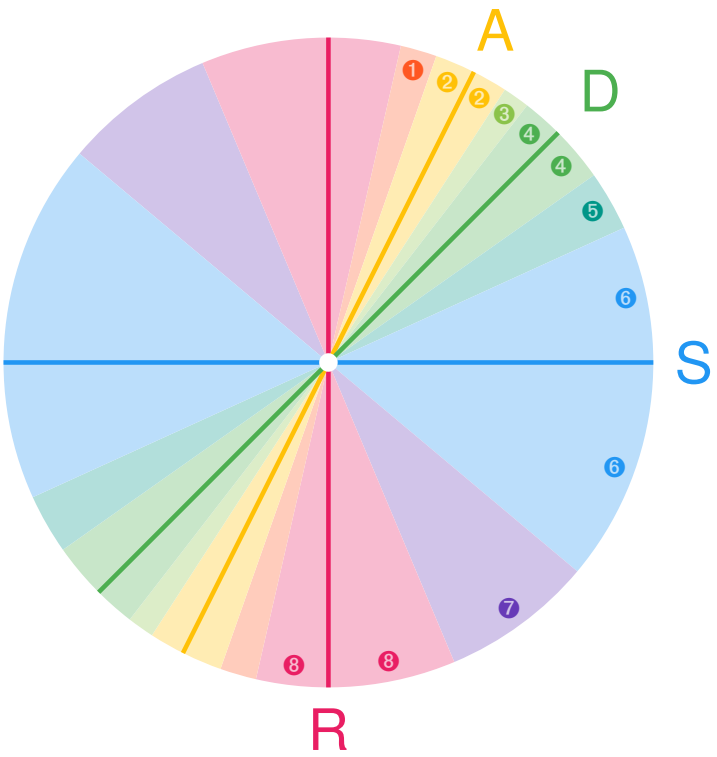

FIGURA 4.4. Límites de la equivalencia coherente por modelo genético. Las regiones con *marcas pares* están formadas por valores susceptibles de ser coherentemente equivalentes al modelo genético de la recta (A, D, S o R) que contienen. Las regiones con *marcas impares*, por valores susceptibles de serlo a los modelos entre cuyas rectas se encuentran.

Comparaciones múltiples

Compárense las FIGURAS 4.2 y 4.4. En la primera, se observan la relación entre β_1 y β_2 por modelo genético y las 8 regiones abiertas en que estas relaciones subdividen el plano; en la segunda, los dominios de cada uno de los modelos genéticos en dichas regiones de acuerdo con el TDE (a nivel $\alpha = 0{,}05$) que puede aplicarse a partir de los resultados de la sección anterior: las marcas pares designan valores susceptibles de ser coherentemente equivalentes al modelo genético de la recta (A, D, S o R) que contienen; las marcas impares, valores susceptibles de serlo a los modelos entre cuyas rectas se encuentran.

Los límites de la equivalencia coherente por modelo genético se han obtenido igualando el parámetro poblacional correspondiente (θ_A, θ_D, θ_S o θ_R) a $\pm\delta_\alpha$ (véase la SECCIÓN 3.1): para el modelo aditivo ($\theta_A = (2x-y)/y$, donde $x = \beta_1$ e $y = \beta_2$), se tienen las rectas

$$\frac{2x-y}{y} = \pm\delta_\alpha \Longleftrightarrow y = \frac{2x}{1\pm\delta_\alpha} \ ; \tag{4.15}$$

para el dominante,

$$\frac{x-y}{x} = \pm\delta_\alpha \Longleftrightarrow y = (1\mp\delta_\alpha)\,x \ ; \tag{4.16}$$

para el sobredominante,

$$\frac{y}{2x-y} = \pm\delta_\alpha \Longleftrightarrow y = \frac{2x}{1\pm\delta_\alpha^{-1}} \ ; \tag{4.17}$$

y, para el recesivo,

$$\frac{x}{x-y} = \pm\delta_\alpha \Longleftrightarrow y = (1\mp\delta_\alpha^{-1})\,x \ . \tag{4.18}$$

Recuérdese que, a medida que $\alpha \to 0^+$, $\delta_\alpha \to \delta/2$, valor para el que las ECUACIONES 4.15-4.18 dan las regiones con marcas impares de tamaño menor, y que el ángulo menor (en radianes) entre dos rectas con pendientes m_1 y m_2 viene dado por

$$\arctan\left|\frac{m_1-m_2}{1+m_1m_2}\right| \ .$$

Así, la región con marca ❶ está delimitada por las rectas de las ECUACIONES 4.15 y 4.18 para $\delta_\alpha = -1/2$, por lo que su tamaño es

$$\arctan\left|\frac{4-3}{1+4\cdot 3}\right| = \arctan\left(\frac{1}{13}\right) \ ;$$

la región con marca ❸, por las de las ECUACIONES 4.15 y 4.16 para $\delta_\alpha = 1/2$ y $\delta_\alpha = -1/2$, respectivamente, por lo que su tamaño es

$$\arctan\left|\frac{\frac{4}{3}-\frac{3}{2}}{1+\frac{4}{3}\cdot\frac{3}{2}}\right| = \arctan\left(\frac{1}{18}\right) \ ;$$

la región con marca ❺, por las de las ECUACIONES 4.16 y 4.17 para $\delta_\alpha = 1/2$, por lo que su tamaño es

$$\arctan\left|\frac{\frac{1}{2}-\frac{2}{3}}{1+\frac{1}{2}\cdot\frac{2}{3}}\right| = \arctan\left(\frac{1}{8}\right) \ ;$$

y la región con marca ❼, por las de las ECUACIONES 4.17 y 4.18 para $\delta_\alpha = -1/2$ y $\delta_\alpha = 1/2$, respectivamente, por lo que su tamaño es

$$\arctan\left|\frac{-2-(-1)}{1+(-2)\cdot(-1)}\right| = \arctan\left(\frac{1}{3}\right) \ .$$

Sumando los tamaños de cada región y dividiendo esta suma por los π radianes de la semicircunferencia, se halla la proporción mínima del plano que ocupan los valores de β_1 y β_2 para los cuales es necesario aplicar dos TDE:

$$\frac{1}{\pi}\left[\arctan\left(\frac{1}{13}\right)+\arctan\left(\frac{1}{18}\right)+\arctan\left(\frac{1}{8}\right)\right.$$
$$\left.+\arctan\left(\frac{1}{3}\right)\right] = \frac{1}{\pi}\arctan\left(\frac{126}{193}\right) \approx 18{,}41\,\% \ .$$

Asumiendo, pues, que los valores de β_1 y β_2 se distribuyen al azar y que m, el número de SNP o interacciones SNP-SNP a los que se quiere aplicar uno o dos TDE para la selección de modelos genéticos, es

elevado, el nivel de significación estadística de cada TDE ha de ser $\widetilde{\alpha} \leq \alpha/(1{,}1841m)$ para mantener la probabilidad de error tipo I a α.

Capítulo

5 Redes de interacciones en el estudio MCC-Spain

En este capítulo, se confecciona un protocolo de construcción de redes de interacciones SNP-SNP para estudios de casos y controles con que analizar los datos del estudio MCC-Spain. Asimismo, se discuten los resultados obtenidos con referencias biológicas.

Como ya se anticipó en la SECCIÓN 2.1, el estudio MCC-Spain genotipó, aproximadamente, el 60 % de los 10 106 individuos mediante un *array* de exoma de Illumina® con más de 200 000 variantes genéticas en que se incluyeron unas 5000 adicionales de genes implicados en rutas biológicas asociadas con cáncer. Para la construcción de las redes de interacciones SNP-SNP, las variantes genéticas se filtraron de acuerdo con su frecuencia del alelo menor (MAF, por su sigla en inglés), de manera que se descartaron aquellas con una MAF $< 0{,}1$, lo que supone el análisis de 23 806 SNP y sus interacciones. A continuación, se detalla el método de generación de redes de interacciones SNP-SNP en el estudio MCC-Spain y se discuten las posibles implicaciones de los resultados obtenidos.

5.1 Protocolo de construcción

Se parte, pues, de 23 806 SNP comunes a los cánceres colorrectal, de mama, de próstata, gástrico y leucemia linfática crónica, con abreviaturas `COL`, `MAM`, `PRO`, `GAS` y `LLC`, respectivamente, cuyo número irá reduciéndose a lo largo de este capítulo a medida que se les apliquen los filtros que se exponen a continuación.

Filtrado de SNP

Para cada SNP y tipo de cáncer, se computan sus coeficientes de coherencia ($m = 10^6$, aunque bastaría con $m = \lceil 1{,}1841 \cdot 23806 \rceil = 28189$ según la SECCIÓN 4.2) con el equilibrio de Hardy-Weinberg en controles, C_{HW}, y con los modelos genéticos dominante, sobredominante y recesivo. El coeficiente de coherencia del modelo genético que subyace tras la asociación SNP-cáncer, C_{MG}, es el máximo de los tres anteriores. Para cada tipo de cáncer, se descartan aquellos SNP (véase la segunda columna de la TABLA 5.1) para los que

$$\text{mín}\left\{\text{máx}\left\{0, C_{\mathrm{HW}}\right\}, C_{\mathrm{MG}}\right\} \leq 0 \ . \tag{5.1}$$

A los SNP que pasan el filtro se les aplica la transformación que se corresponde con su modelo genético. De esta manera, todos los SNP resultantes se convierten en variables binarias. Para controlar la potencial redundancia entre SNP, acrecentada ahora por la reducción del rango de

	Filtro	
Efecto	ECUACIÓN 5.1	ECUACIÓN 5.2
COL	5831	4631
MAM	5097	4161
PRO	4971	4023
GAS	5906	4706
LLC	5276	4264

TABLA 5.1. SNP no descartados por efecto y filtro. La aplicación de los filtros es secuencial, es decir, la ECUACIÓN 5.2 se comprueba sobre los SNP no descartados por la ECUACIÓN 5.1.

las variables, se utiliza un coeficiente de similitud emparentado con el índice de Jaccard. Así, para cada par de SNP, se comprueba si

$$\frac{n_{01} + n_{10}}{n} < 0{,}01 \ , \tag{5.2}$$

donde n es el número total de individuos y $n_{01} + n_{10}$, el de aquellos en que los genotipos de los dos SNP son diferentes.

El problema de encontrar el mínimo número de SNP tal que, al ser eliminados, ningún par de SNP entre los restantes cumpla la ECUACIÓN 5.2 es NP-completo.[175] Sin embargo, existe un algoritmo aproximado sencillo que elimina, a lo sumo, el doble de SNP de lo necesario en tiempo cuadrático con respecto al número de estos.[176] Para ello, asigna un peso igual a 1 a cada SNP. Entonces, para cada par de SNP X_1 y X_2, los pesos correspondientes se actualizan a partir del peso mínimo entre X_1 y X_2: $w_1 = w_1 - \text{mín}\{w_1, w_2\}$ y $w_2 = w_2 - \text{mín}\{w_1, w_2\}$, donde w_1 y w_2 son los pesos de X_1 y X_2, respectivamente. Finalmente, aquellos SNP con peso igual a 0 se eliminan (véase la tercera columna de la TABLA 5.1).

TABLA 5.2. Interacciones no descartadas por efecto y filtro. La aplicación de los filtros es secuencial, es decir, $C_{MG} \leq 0$ se comprueba sobre las interacciones no descartadas por la ECUACIÓN 5.3.

	Filtro	
Efecto	ECUACIÓN 5.3	$C_{MG} \leq 0$
COL	9 477 266	3 755 487
MAM	7 692 028	2 812 402
PRO	7 070 041	2 493 346
GAS	7 876 634	2 739 893
LLC	7 211 881	2 542 417

Filtrado de interacciones

Para cada par de SNP, se comprueba si

$$\bigvee_{a,b} (n_{ab} = 0) \ , \tag{5.3}$$

donde $(a, b) \in \{0, 1\}^2$, es decir, si hay alguna combinación de genotipos (00, 01, 10 o 11) para la que no haya individuos que la presenten. De ser así, se descarta el análisis de la interacción potencial entre el par de SNP (véase la segunda columna de la TABLA 5.2).

Para cada par de SNP, se suman en una variable ternaria de interacción y se computan, por tipo de cáncer, su coeficiente de coherencia ($m = 5 \cdot 10^{11}$) con los modelos genéticos dominante, sobredominante y recesivo. El coeficiente de coherencia del modelo genético que subyace tras la asociación interacción-cáncer, C_{MG}, es el máximo de los tres anteriores. Para cada tipo de cáncer, se descartan aquellas interacciones (véase la tercera columna de la TABLA 5.2) para las que $C_{MG} \leq 0$.

Con cada interacción, se ajusta el siguiente modelo de regresión logística, que no incluye los efectos principales de los SNP involucrados en ellas:

$$\log\left(\frac{\mathrm{E}\left[Y\right]}{1-\mathrm{E}\left[Y\right]}\right) = \beta_0 + \sum_{i,j} \beta_{ij} f_{ij}(f_i(X_i), f_j(X_j)) \ ,$$

tal que $i, j \in \{1, \ldots, n\}$ y $j > i$, con $\tilde{\alpha} = 3{,}5 \cdot 10^{-9}$, que se obtiene de dividir $\alpha = 0{,}05$ entre el total de interacciones SNP-SNP que analizar (sumando la tercera columna de la TABLA 5.2, se obtienen 14 343 545 interacciones SNP-SNP).

Con las interacciones cuyo valor de p es menor que $3{,}5 \cdot 10^{-9}$, que son 4, 0, 2, 20 y 3 para los cánceres colorrectal, de mama, de próstata, gástrico y leucemia linfática crónica, respectivamente, se ajusta el siguiente modelo lineal, que incluye tanto los efectos principales de cada SNP como los de interacción con otros SNP:

$$\log\left(\frac{\mathrm{E}\left[Y\right]}{1-\mathrm{E}\left[Y\right]}\right) = \beta_0 + \sum_i \beta_i f_i(X_i) + \sum_j \beta_{ij} f_{ij}(f_i(X_i), f_j(X_j))$$

(véase la SECCIÓN 2.2), con $\tilde{\alpha} = 0{,}0125$, que se obtiene de dividir $\alpha = 0{,}05$ entre 4, que es el número de modelos que van a ajustarse (el correspondiente al cáncer de mama se omite por no haber interacciones que cumplieran el criterio anterior).

5.2 Resultados

Los resultados que se presentan a continuación, obtenidos a través del método propuesto en esta tesis doctoral, suponen un primer paso en la generación de hipótesis biológicas relacionadas con los tipos de cáncer estudiados y su asociación con la variación genética. Nótese que la mayoría de los trabajos que se citan para justificar los resultados obtenidos se corresponden con descubrimientos científicos de los últimos 5 años, lo que podría indicar el potencial del método.

Cáncer colorrectal

Las 4 interacciones incluidas en el modelo de cáncer colorrectal fueron estadísticamente significativas, 3 de las cuales —todas son de riesgo— formaban parte del componente conexo más grande: el SNP rs61744949 (gen *VN1R1*) interactúa con rs2158041 (gen *AHR*), rs1350058 (gen *NKAIN3*) y rs1536475 (gen *RXRA*).

El gen *VN1R1* codifica un receptor de sustancias olfativas[177] y transduce su señal a la vía de señalización del adenosín monofosfato cíclico (cAMP).[178] Gran parte del funcionamiento de esta vía depende de la fosforilación de varios genes, como el *CREB* o el *NR4A1*, por lo que el funcionamiento de la ATPasa sodio/potasio (ATP) es crucial.[179] La proteína transmembrana codificada por el gen *NKAIN3* interactúa con subunidades β1 de la ATPasa sodio/potasio,[180] por lo que mutaciones

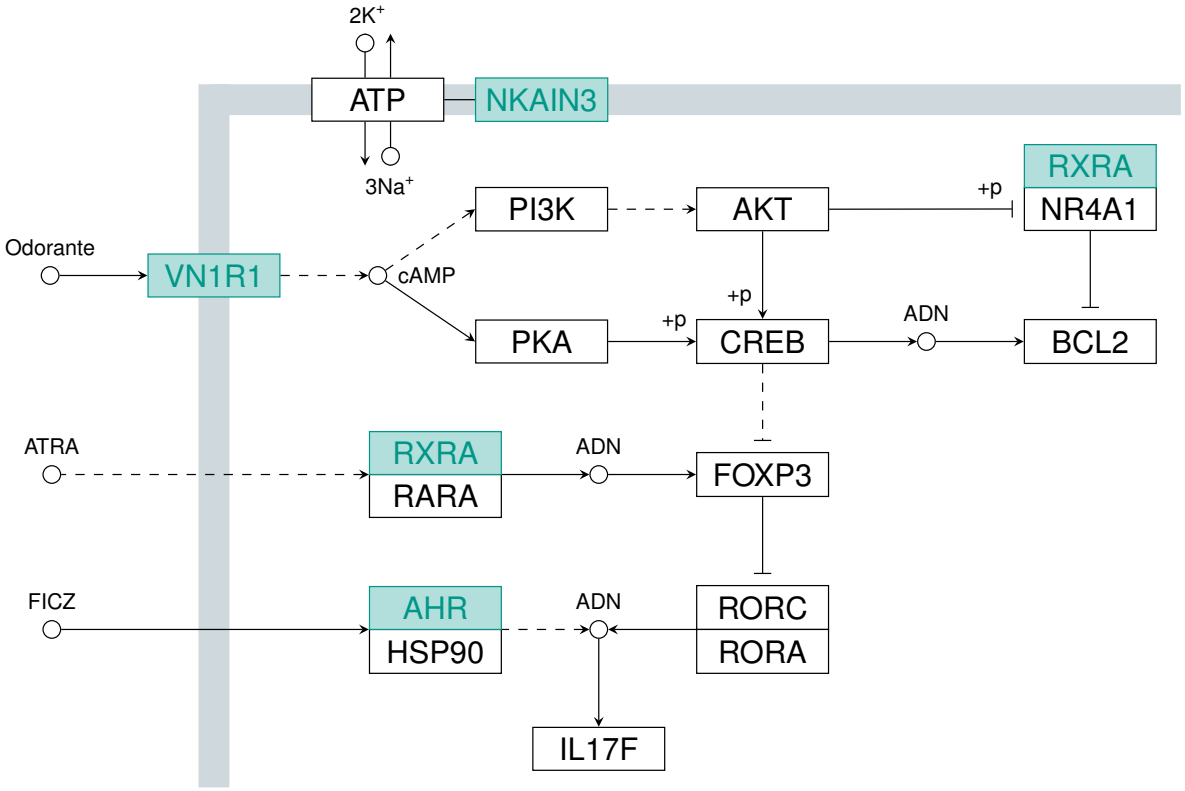

FIGURA 5.1. Relación entre los genes involucrados en las interacciones del cáncer colorrectal. Los rectángulos representan proteínas y los círculos, moléculas; las líneas continuas, una relación directa, y las discontinuas, indirecta. Dos rectángulos juntos simbolizan un complejo proteico, mientras que si los une una línea continua sin flecha, se trata de una interacción. La banda gris representa la membrana celular.

en el primero podrían afectar a la liberación de fósforos libres en el interior de la célula y, por tanto, a la fosforilación de los genes antes mencionados.[181]

BCL2 es una proteína antiapoptótica, es decir, que no favorece la muerte celular, por lo que sintetizada en exceso podría contribuir al desa-

rrollo de cáncer. La proteína producida por *RXRA* forma un complejo proteico con NR4A1 para inhibir la producción de BCL2. Mutaciones en la primera podrían no impedir la creación de BCL2 en demasía. Por otra parte, RXRA, formando otro complejo proteico con RARA, favorece la expresión de *FOXP3* que, a su vez, inhibe la formación de interleucina 17F, que se trata de una citoquina proinflamatoria. Mutaciones en *RXRA* podrían favorecer la formación de IL17F y, de esta manera, una respuesta inflamatoria relacionada con el cáncer colorrectal.

La proteína del gen *AHR* está relacionada con el cáncer colorrectal de la misma manera, a través de la IL17F. El complejo proteico que forma con HSP90 favorece la producción de IL17F, por lo que mutaciones en *AHR* podrían determinar la unión del complejo proteico.

Cáncer gástrico

De las 20 interacciones incluidas en el modelo de cáncer gástrico, 8 fueron estadísticamente significativas, con 3 componentes conexos de 2 interacciones cada uno. Se analiza a continuación aquel cuya suma de los tamaños del efecto (de riesgo) es mayor. El SNP rs1864193 (gen *INSR*) interactúa con rs1998598 (gen *DENND1B*) y rs6696888 (gen *ASH1L*).

El gen *INSR* codifica un receptor de la insulina que, mutado, glicosila el gen de la e-caderina, el *CDH1*.[182] Este gen está relacionado con cáncer gástrico y con la supresión de tumores, por lo que una glicosilación en él

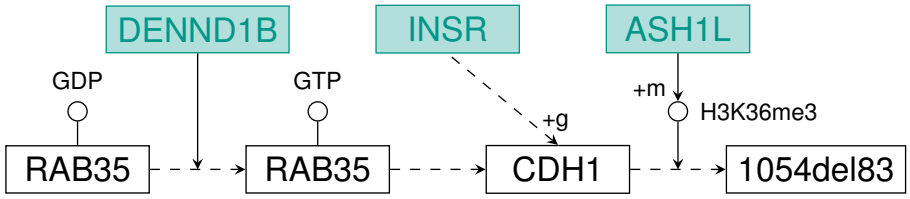

FIGURA 5.2. Relación entre los genes involucrados en las interacciones del cáncer gástrico. *+g* y *+m* representan glicosilación y metilación, respectivamente. 1054del83 es una forma alternativa del gen *CDH1* en cuyo exón 8 hay una deleción.

puede motivar la invasión de células tumorales. La expresión del *CDH1* se controla a través de la activación (paso de GDP a GTP) del RAB35.[183] En esta activación está involucrado, entre otros genes, el *DENND1B*.[184] Mutaciones en este gen podrían afectar a la cantidad de e-caderina que poder glicosilar por parte del *INSR*.

Por otro lado, *ASH1L*, a través del proceso de metilación, regula el *splicing* alternativo del gen *CDH1* acortándolo en el exón 8, lo que quizá determine la funcionalidad o incluso la producción de e-caderina en el organismo.[185]

Leucemia linfática crónica

Las 3 interacciones incluidas en el modelo de leucemia linfática crónica fueron estadísticamente significativas, 2 de las cuales —de protección— formaban parte del componente conexo más grande: el SNP rs9301584 (gen *LINC00353*) interactúa con rs4355801 (gen *OPG*) y rs11559146 (gen *ZDHHC4*).

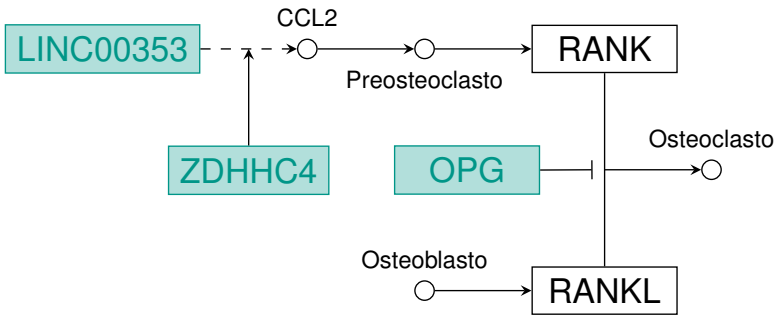

FIGURA 5.3. Relación entre los genes involucrados en las interacciones de la leucemia linfática crónica.

El gen *LINC00353* se ha visto asociado con los niveles de CCL2, la proteína quimioatrayente de monocitos 1, cuya función está relacionada con el reclutamiento de monocitos en el sistema inmune. En esta asociación, se ha visto involucrado, asimismo, el gen *ZDHHC4*, que se expresa en una palmitoiltransferasa que, en líneas celulares derivadas de monocitos de pacientes con leucemia, ayuda al reclutamiento de monocitos por parte de CCL2.[186]

Entre los monocitos reclutados por CCL2 están los preosteoclastos, responsables de activar la expresión de *RANK*. Este es un gen que, en conjunción con *RANKL* (activado por osteoblastos), se encarga de la formación de osteoclastos, que, en exceso, se han visto asociados con un mayor riesgo de padecer leucemia linfática crónica.[187] El gen *OPG*, precisamente, es el que regula la interacción entre *RANK* y *RANKL* en la formación de osteoclastos. La activación de *OPG* inhibe parcialmente la formación de osteoclastos, lo que proporciona protección frente a

leucemia linfática crónica.[188] Mutaciones en *OPG* podrían llevar a una sobreexpresión del gen y, por tanto, a una sobreinhibición de la formación de osteoclastos, protegiendo así ante la leucemia linfática crónica.

Capítulo

6 Conclusiones y líneas de investigación futuras

En este capítulo, se aporta una perspectiva general tanto de los objetivos alcanzados en esta tesis doctoral como de las contribuciones científicas que se derivan de su consecución. Asimismo, se describen las líneas de investigación futuras.

6.1 Conclusiones

La hipótesis de trabajo de esta tesis doctoral conjeturaba, por un lado, que un porcentaje significativo de las asociaciones SNP-enfermedad e interacción-enfermedad estarían regidas por uno de los siguientes modelos: aditivo, dominante, sobredominante y recesivo. Con respecto a las primeras, en torno al 25 % de los SNP analizados en relación con los cánceres colorrectal, de mama, de próstata, gástrico y leucemia linfática crónica presentó un patrón compatible con alguno de los modelos (véase la TABLA 5.1); en cuanto a las segundas, este porcentaje se situó entre el 35 y el 40 % de las interacciones analizadas (véase la TABLA 5.2). Ambos porcentajes confirman esta parte de la hipótesis.

Por otro lado, la misma hipótesis aventuraba que las redes en que todas las asociaciones SNP-enfermedad e interacción-enfermedad siguieran uno de los modelos anteriores estarían formadas por componentes conexos relativamente pequeños. Las redes de interacciones SNP-SNP generadas en el estudio MCC-Spain se dispusieron en subgrafos estrella: las cuatro interacciones estadísticamente significativas del modelo de cáncer colorrectal se agruparon en el subgrafo de 3 aristas discutido en la SECCIÓN 5.2 y en otro de 1; las ocho del modelo de cáncer gástrico, en el de 2 aristas discutido, en dos isomorfos a este y en dos más de 1 arista cada uno; y las tres del modelo de leucemia linfática crónica, en el de 2 aristas también discutido y en otro de 1. Estas redes de interacciones SNP-SNP confirman esta parte de la hipótesis.

El primero de los objetivos que se definieron para verificar la hipótesis anterior era el de desarrollar un marco de trabajo que permitiera evaluar y comparar los niveles de adecuación e incertidumbre de distintos patrones a volúmenes masivos de datos potencialmente redundantes. El marco de trabajo idóneo para ello es el de los test de diferencia-equivalencia, por lo que, para alcanzar este objetivo, se formalizó el proceso para aplicar un test de diferencia-equivalencia e interpretar sus resultados en presencia de comparaciones múltiples. A falta de procedimientos fiables para el cálculo del valor de p para los resultados de diferencia y equivalencia coherentes al corregir por comparaciones múltiples, se desarrolló, en esta tesis doctoral, el coeficiente de coherencia, medida de

hasta qué punto estos resultados prevalecen sobre los de incoherencia e indeterminación. Los valores del coeficiente de coherencia obtenidos en la SECCIÓN 3.3 han dejado en entredicho las falacias que asumen que no ser capaz de demostrar diferencia significa demostrar equivalencia y que no ser capaz de demostrar equivalencia significa demostrar diferencia, falacias en que el investigador puede incurrir de no aplicar test de diferencia-equivalencia.

El segundo de los objetivos que se definieron para verificar la hipótesis de trabajo de esta tesis doctoral era el de diseñar e implementar una prueba estadística que permitiera decidir qué modelo genético le corresponde a un SNP o una interacción. La prueba estadística fruto de la consecución de este objetivo permite decidir, mediante la aplicación de uno o dos test de diferencia-equivalencia (con una proporción 4:1), qué patrón subyace —si es que subyace alguno— tras asociaciones SNP-enfermedad o interacción-enfermedad en volúmenes masivos de datos. El estudio de las relaciones entre los riesgos relativos de los modelos genéticos ha permitido construir estadísticos de contraste de hipótesis nulas de no equivalencia con margen de equivalencia común. Asumiendo que estos riesgos relativos se distribuyen al azar y que el número de SNP o interacciones SNP-SNP que analizar es elevado, las probabilidades de tener que aplicar uno o dos test de diferencia-equivalencia para la selección de modelos genéticos son del 81,59 % y del 18,41 %, respectivamente (véase la SECCIÓN 4.2).

El último de los objetivos que se definieron para verificar la hipótesis de trabajo de esta tesis doctoral era el de construir y analizar redes de interacciones SNP-SNP a partir de los datos del estudio MCC-Spain. Para alcanzarlo, se confeccionó un protocolo de construcción de redes de interacciones SNP-SNP para estudios de casos y controles, que permite generar hipótesis biológicas relacionadas con la contribución de la variación en el genoma al desarrollo de enfermedades comunes. El método de análisis propuesto supone un primer paso en la generación de hipótesis biológicas relacionadas con los tipos de cáncer estudiados y su asociación con la variación genética y sus interacciones. Nótese que la mayoría de los trabajos que se han citado para justificar los resultados obtenidos se corresponden con descubrimientos científicos de los últimos 5 años, lo que podría indicar el potencial del método.

6.2 Líneas de investigación futuras

TEST DE DIFERENCIA-EQUIVALENCIA: El contraste de hipótesis de diferencia-equivalencia representa, para la inferencia estadística, la unificación de dos enfoques complementarios. En esta tesis doctoral, se ha formalizado el proceso para aplicar un test de diferencia-equivalencia a través de una serie de resultados, entre los que el LEMA 5 destaca por su utilidad en una gran variedad de áreas. Aunque este lema hace referencia a una distribución de probabili-

dad específica, su demostración sugiere que se cumple para otras unidimensionales. El coeficiente de coherencia ha permitido, en esta tesis doctoral, describir el efecto de corregir por comparaciones múltiples en test de diferencia-equivalencia. Derivar la distribución muestral del coeficiente de coherencia está entre nuestros trabajos futuros, así como generalizar su definición para el contraste de hipótesis nulas de no equivalencia con márgenes de equivalencia asimétricos, es decir, $\delta_1 \neq \delta_2$.

SELECCIÓN DE MODELOS Y REDES DE INTERACCIONES: Tanto la posibilidad de que las asociaciones SNP-enfermedad o interacción-enfermedad estén regidas por un modelo que no sea aditivo, dominante, sobredominante o recesivo como la de que no estén regidas por ningún modelo suelen ignorarse en la literatura científica. La prueba estadística diseñada en esta tesis doctoral pondera estas alternativas, por lo que compararlas con la estrategia de asignar el «modelo más probable» está entre nuestros trabajos futuros. En cuanto a las redes de interacciones SNP-SNP, la demostración o refutación y la reproducibilidad de las hipótesis biológicas generadas suponen objetivos que cumplir a medio o largo plazo.

Bibliografía

[1] G. MENDEL. Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, **4** (1866), 3-47.

[2] K. PANOUTSOPOULOU y E. WHEELER. Key concepts in genetic epidemiology. *Genetic Epidemiology*. 2018, 7-24.

[3] T. H. N. ELLIS, J. M. I. HOFER, G. M. TIMMERMAN-VAUGHAN, C. J. COYNE y R. P. HELLENS. Mendel, 150 years on. *Trends in Plant Science*, **16** (2011), 590-596.

[4] M. K. BHATTACHARYYA, A. M. SMITH, T. H. N. ELLIS, C. HEDLEY y C. MARTIN. The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. *Cell*, **60** (1990), 115-122.

[5] S. E. ANTONARAKIS, A. CHAKRAVARTI, J. C. COHEN y J. HARDY. Mendelian disorders and multifactorial traits: The big divide or one for all? *Nature Reviews Genetics*, **11** (2010), 380-384.

[6] A. E. GARROD. The incidence of alkaptonuria: A study in chemical individuality. *The Lancet*, **160** (1902), 1616-1620.

[7] A. E. GARROD. The Croonian lectures on inborn errors of metabolism, II: Alkaptonuria. *The Lancet*, **172** (1908), 73-79.

[8] X. MONTAGUTELLI, A. LALOUETTE, M. COUDÉ, P. KAMOUN, M. FOREST y col. *aku*, a mutation of the mouse homologous to human alkaptonuria, maps to chromosome 16. *Genomics*, **19** (1994), 9-11.

[9] M. R. POLLAK, Y. H. W. CHOU, J. J. CERDA, B. STEINMANN, B. N. LA DU y col. Homozygosity mapping of the gene for alkaptonuria to chromosome 3q2. *Nature Genetics*, **5** (1993), 201-204.

[10] S. JANOCHA, W. WOLZ, S. SRSEN, K. SRSNOVA, X. MONTAGUTELLI y col. The human gene for alkaptonuria (*AKU*) maps to chromosome 3q. *Genomics*, **19** (1994), 5-8.

[11] J. M. FERNÁNDEZ-CAÑÓN, B. GRANADINO, D. BELTRÁN-VALERO DE BERNABÉ, M. RENEDO, E. FERNÁNDEZ-RUIZ y col. The molecular basis of alkaptonuria. *Nature Genetics*, **14** (1996), 19-24.

[12] B. KEREM, J. M. ROMMENS, J. A. BUCHANAN, D. MARKIEWICZ, T. K. COX y col. Identification of the cystic fibrosis gene: Genetic analysis. *Science*, **245** (1989), 1073-1080.

[13] M. E. MACDONALD, A. NOVELLETTO, C. LIN, D. TAGLE, G. BARNES y col. The Huntington's disease candidate region exhibits many different haplotypes. *Nature Genetics*, **1** (1992), 99-103.

[14] H. K. TABOR, N. J. RISCH y R. M. MYERS. Candidate-gene approaches for studying complex genetic traits: Practical considerations. *Nature Reviews Genetics*, **3** (2002), 391-397.

[15] I. R. GIZER, C. FICKS e I. D. WALDMAN. Candidate gene studies of ADHD: A meta-analytic review. *Human Genetics*, **126** (2009), 51-90.

[16] J. N. HIRSCHHORN y M. J. DALY. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, **6** (2005), 95-108.

[17] W. Y. S. WANG, B. J. BARRATT, D. G. CLAYTON y J. A. TODD. Genome-wide association studies: Theoretical and practical concerns. *Nature Reviews Genetics*, **6** (2005), 109-118.

[18] J. N. HIRSCHHORN. Genome-wide association studies: Illuminating biologic pathways. *The New England Journal of Medicine*, **360** (2009), 1699-1701.

[19] J. HARDY y A. SINGLETON. Genome-wide association studies and human disease. *The New England Journal of Medicine*, **360** (2009), 1759-1768.

[20] C. A. ANDERSON, F. H. PETTERSSON, G. M. CLARKE, L. R. CARDON, A. P. MORRIS y col. Data quality control in genetic case–control association studies. *Nature Protocols*, **5** (2010), 1564-1573.

[21] G. M. CLARKE, C. A. ANDERSON, F. H. PETTERSSON, L. R. CARDON, A. P. MORRIS y col. Basic statistical analysis in genetic case–control studies. *Nature Protocols*, **6** (2011), 121-133.

[22] K. OZAKI, Y. OHNISHI, A. IIDA, A. SEKINE, R. YAMADA y col. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, **32** (2002), 650-654.

[23] A. D. JOHNSON y C. J. O'DONNELL. An open access database of genome-wide association results. *BMC Medical Genetics*, **10** (2009), 6-22.

[24] D. E. REICH y E. S. LANDER. On the allelic spectrum of human disease. *Trends in Genetics*, **17** (2001), 502-510.

[25] J. H. MOORE y M. D. RITCHIE. The challenges of whole-genome approaches to common diseases. *Journal of the American Medical Association*, **291** (2004), 1642-1643.

[26] A. G. CLARK, E. BOERWINKLE, J. HIXSON y C. F. SING. Determinants of the success of whole-genome association testing. *Genome Research*, **15** (2005), 1463-1467.

[27] L. A. HINDORFF, P. SETHUPATHY, H. A. JUNKINS, E. M. RAMOS, J. P. MEHTA y col. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, **106** (2009), 9362-9367.

[28] B. MCKINNEY y N. PAJEWSKI. Six degrees of epistasis: Statistical network models for GWAS. *Frontiers in Genetics*, **2** (2012), 109-114.

[29] P. C. PHILLIPS. Epistasis: The essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, **9** (2008), 855-867.

[30] H. J. CORDELL. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, **11** (2002), 2463-2468.

[31] W. BATESON. *Mendel's Principles of Heredity*. Cambridge University Press, Cambridge, 1909.

[32] R. A. FISHER. The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, **52** (1919), 399-433.

[33] J. H. MOORE. A global view of epistasis. *Nature Genetics*, **37** (2005), 13-14.

[34] Z. B. ZENG, T. WANG y W. ZOU. Modeling quantitative trait loci and interpretation of models. *Genetics*, **169** (2005), 1711-1725.

[35] H. J. CORDELL. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, **10** (2009), 392-404.

[36] R. J. NEUMAN, J. P. RICE y A. CHAKRAVARTI. Two-locus models of disease. *Genetic Epidemiology*, **9** (1992), 347-365.

[37] W. LI y J. REICH. A complete enumeration and classification of two-locus disease models. *Human Heredity*, **50** (2000), 334-349.

[38] I. B. HALLGRÍMSDÓTTIR y D. S. YUSTER. A complete classification of epistatic two-locus models. *BMC Genetics*, **9** (2008), 17-31.

[39] W. H. WEI, G. HEMANI y C. S. HALEY. Detecting epistasis in human complex traits. *Nature Reviews Genetics*, **15** (2014), 722-733.

[40] C. CASTILLEJO-LÓPEZ, A. M. DELGADO-VEGA, J. WOJCIK, S. V. KOZYREV, E. THAVATHIRU y col. Genetic and physical interaction of the B-cell systemic lupus erythematosus-associated genes *BANK1* and *BLK*. *Annals of the Rheumatic Diseases*, **71** (2011), 136-142.

[41] J. MARCHINI, P. DONNELLY y L. R. CARDON. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, **37** (2005), 413-417.

[42] J. H. MOORE, F. W. ASSELBERGS y S. M. WILLIAMS. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, **26** (2010), 445-455.

[43] D. F. SCHWARZ, I. R. KÖNIG y A. ZIEGLER. On safari to Random Jungle: A fast implementation of Random Forests for high-dimensional data. *Bioinformatics*, **26** (2010), 1752-1758.

[44] X. PANG, Z. WANG, J. S. YAP, J. WANG, J. ZHU y col. A statistical procedure to map high-order epistasis for complex traits. *Briefings in Bioinformatics*, **14** (2012), 302-314.

[45] T. HU, N. A. SINNOTT-ARMSTRONG, J. W. KIRALIS, A. S. ANDREW, M. R. KARAGAS y col. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics*, **12** (2011), 364-378.

[46] N. A. DAVIS, C. A. LAREAU, B. C. WHITE, A. PANDEY, G. WILEY y col. Encore: Genetic Association Interaction Network centrality pipeline and application to SLE exome data. *Genetic Epidemiology*, **37** (2013), 614-621.

[47] S. FORTUNATO. Community detection in graphs. *Physics Reports*, **486** (2010), 75-174.

[48] J. H. MOORE, J. C. GILBERT, C. T. TSAI, F. T. CHIANG, T. HOLDEN y col. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology*, **241** (2006), 252-261.

[49] J. KNIGHTS, J. YANG, P. CHANDA, A. ZHANG y M. RAMANATHAN. SYMPHONY, an information-theoretic method for gene–gene and gene–environment interaction analysis of disease syndromes. *Heredity*, **110** (2013), 548-559.

[50] W. MCGILL. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, **4** (1954), 93-111.

[51] O. COMBARROS, M. CORTINA-BORJA, A. D. SMITH y D. J. LEHMANN. Epistasis in sporadic Alzheimer's disease. *Neurobiology of Aging*, **30** (2009), 1333-1349.

[52] P. A. LEVENE y W. A. JACOBS. Über inosinsäure. *Berichte der deutschen chemischen Gesellschaft*, **42** (1909), 1198-1203.

[53] P. A. LEVENE. The structure of yeast nucleic acid, IV: Ammonia hydrolysis. *Journal of Biological Chemistry*, **40** (1919), 415-424.

[54] J. D. WATSON y F. H. C. CRICK. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, **171** (1953), 737-738.

[55] F. H. C. CRICK. Central dogma of molecular biology. *Nature*, **227** (1970), 561-563.

[56] M. NIRENBERG, P. LEDER, M. BERNFIELD, R. BRIMACOMBE, J. TRUPIN y col. RNA codewords and protein synthesis, VII: On the general nature of the RNA code. *Proceedings of the National Academy of Sciences of the United States of America*, **53** (1965), 1161-1168.

[57] F. H. C. CRICK. The origin of the genetic code. *Journal of Molecular Biology*, **38** (1968), 367-379.

[58] S. BRENNER, A. O. W. STRETTON y S. KAPLAN. Genetic code: The 'nonsense' triplets for chain termination and their suppression. *Nature*, **206** (1965), 994-998.

[59] S. BRENNER, L. BARNETT, E. R. KATZ y F. H. C. CRICK. UGA: A third nonsense triplet in the genetic code. *Nature*, **213** (1967), 449-450.

[60] W. M. JOU, G. HAEGEMAN, M. YSEBAERT y W. FIERS. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, **237** (1972), 82-88.

[61] L. T. CHOW, R. E. GELINAS, T. R. BROKER y R. J. ROBERTS. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, **12** (1977), 1-8.

[62] A. J. BERK y P. A. SHARP. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell*, **12** (1977), 721-732.

[63] S. M. BERGET, C. MOORE y P. A. SHARP. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, **74** (1977), 3171-3175.

[64] W. S. SUTTON. The chromosomes in heredity. *Biological Bulletin*, **4** (1903), 231-250.

[65] J. H. TJIO y A. LEVAN. The chromosome number of man. *Hereditas*, **42** (1956), 1-6.

[66] E. SKUTELSKY y D. DANON. Comparative study of nuclear expulsion from the late erythroblast and cytokinesis. *Experimental Cell Research*, **60** (1970), 427-436.

[67] T. H. MORGAN. *The Mechanism of Mendelian Heredity*. Henry Holt & Co., NY, 1915.

[68] THE INTERNATIONAL HAPMAP 3 CONSORTIUM. Integrating common and rare genetic variation in diverse human populations. *Nature*, **467** (2010), 52-58.

[69] W. BATESON y E. R. SAUNDERS. The facts of heredity in the light of Mendel's discovery. *Reports to the Evolution Committee of the Royal Society*, **1** (1902), 125-160.

[70] THE 1000 GENOMES PROJECT CONSORTIUM. A map of human genome variation from population-scale sequencing. *Nature*, **467** (2010), 1061-1073.

[71] N. RISCH y K. MERIKANGAS. The future of genetic studies of complex human diseases. *Science*, **273** (1996), 1516-1517.

[72] D. G. WANG, J. B. FAN, C. J. SIAO, A. BERNO, P. YOUNG y col. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280** (1998), 1077-1082.

[73] F. A. JANSSENS. Spermatogénèse dans les batraciens, V: La théorie de la chiasmatypie. Nouvelle interprétation des cinèses de maturation. *La Cellule*, **25** (1909), 389-411.

[74] E. B. WILSON y T. H. MORGAN. Chiasmatype and crossing over. *The American Naturalist*, **54** (1920), 193-219.

[75] H. B. CREIGHTON y B. MCCLINTOCK. A correlation of cytological and genetical crossing-over in Zea mays. *Proceedings of the National Academy of Sciences of the United States of America*, **17** (1931), 492-497.

[76] B. DEVLIN y N. RISCH. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, **29** (1995), 311-322.

[77] S. W. GUO. Linkage disequilibrium measures for fine-scale mapping: A comparison. *Human Heredity*, **47** (1997), 301-314.

[78] M. LI, C. LI y W. GUAN. Evaluation of coverage variation of SNP chips for genome-wide association studies. *European Journal of Human Genetics*, **16** (2008), 635-643.

[79] C. M. LEWIS. Genetic association studies: Design, analysis and interpretation. *Briefings in Bioinformatics*, **3** (2002), 146-153.

[80] THE INTERNATIONAL HAPMAP CONSORTIUM. A haplotype map of the human genome. *Nature*, **437** (2005), 1299-1320.

[81] Q. T. ANGELERIO. *Ectypa pestilentis status Algheriae Sardiniae*. F. Guarnerio, Cagliari, 1588.

[82] Q. T. ANGELERIO. *Epidemiologia sive tractatus de peste*. Ex Typographia Regia, Madrid, 1598.

[83] C. BUCK, A. LLOPIS, E. NÁJERA y M. TERRIS. *The Challenge of Epidemiology*. Pan American Health Organization, DC, 1988.

[84] N. E. MORTON. Genetic epidemiology. *Annals of Human Genetics*, **61** (1997), 1-13.

[85] H. J. CORDELL y D. G. CLAYTON. Genetic association studies. *The Lancet*, **366** (2005), 1121-1131.

[86] D. A. GRIMES y K. F. SCHULZ. Cohort studies: Marching towards outcomes. *The Lancet*, **359** (2002), 341-345.

[87] K. F. SCHULZ y D. A. GRIMES. Case–control studies: Research in reverse. *The Lancet*, **359** (2002), 431-434.

[88] M. DELGADO-RODRÍGUEZ y J. LLORCA. Bias. *Journal of Epidemiology & Community Health*, **58** (2004), 635-641.

[89] D. A. GRIMES y K. F. SCHULZ. Bias and causal associations in observational research. *The Lancet*, **359** (2002), 248-252.

[90] D. A. GRIMES y K. F. SCHULZ. An overview of clinical research: The lay of the land. *The Lancet*, **359** (2002), 57-61.

[91] G. CASTAÑO-VINYALS, N. ARAGONÉS, B. PÉREZ-GÓMEZ, V. MARTÍN, J. LLORCA y col. Population-based multicase–control study in common tumors in Spain (MCC-Spain): Rationale and study design. *Gaceta Sanitaria*, **29** (2015), 308-315.

[92] G. CASTAÑO-VINYALS, N. ARAGONÉS, B. PÉREZ-GÓMEZ, V. MARTÍN, J. LLORCA y col. Corrigendum to: Population-based multicase–control study in common tumors in Spain (MCC-Spain): Rationale and study design. *Gaceta Sanitaria*, **32** (2018), 501.

[93] J. ALONSO-MOLERO, C. GONZÁLEZ-DONQUILES, C. PALAZUELOS, T. FERNÁNDEZ-VILLA, E. RAMOS y col. The rs4939827 polymorphism in the *SMAD7* gene and its association with Mediterranean diet in colorectal carcinogenesis. *BMC Medical Genetics*, **18** (2017), 122.

[94] T. DIERSSEN-SOTOS, C. PALAZUELOS-CALDERÓN, J. J. JIMÉNEZ-MOLEÓN, N. ARAGONÉS, J. M. ALTZIBAR y col. Reproductive risk factors in breast cancer and genetic hormonal pathways: A gene–environment interaction in the MCC-Spain project. *BMC Cancer*, **18** (2018), 280.

[95] I. GÓMEZ-ACEBO, T. DIERSSEN-SOTOS, C. PALAZUELOS, P. FERNÁNDEZ-NAVARRO, G. CASTAÑO-VINYALS y col. Pigmentation phototype and prostate and breast cancer in a select Spanish population: A Mendelian randomization analysis in the MCC-Spain study. *PLOS ONE*, **13** (2018), 1-15.

[96] E. J. C. G. VAN DEN OORD. Controlling false discoveries in genetic studies. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, **147** (2008), 637-644.

[97] W. S. BUSH y J. H. MOORE. Genome-wide association studies. *PLOS Computational Biology*, **8** (2012), e1002822.

[98] I. MENASHE, D. MAEDER, M. GARCIA-CLOSAS, J. D. FIGUEROA, S. BHATTACHARJEE y col. Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. *Cancer Research*, **70** (2010), 4453-4459.

[99] J. P. A. IOANNIDIS, P. CASTALDI y E. EVANGELOU. A compendium of genome-wide associations for cancer: Critical synopsis and reappraisal. *Journal of the National Cancer Institute*, **102** (2010), 846-858.

[100] G. IBÁÑEZ-SANZ, A. DÍEZ-VILLANUEVA, M. H. ALONSO, F. RODRÍGUEZ-MORANTA, B. PÉREZ-GÓMEZ y col. Risk model for colorectal cancer in Spanish population using environmental and genetic factors: Results from the MCC-Spain study. *Scientific Reports*, **7** (2017), 43263.

[101] I. GÓMEZ-ACEBO, T. DIERSSEN-SOTOS, P. FERNÁNDEZ-NAVARRO, C. PALAZUELOS, V. MORENO y col. Risk model for prostate cancer using environmental and genetic factors in the Spanish multi-case–control (MCC) study. *Scientific Reports*, **7** (2017), 8994.

[102] T. DIERSSEN-SOTOS, I. GÓMEZ-ACEBO, C. PALAZUELOS, P. FERNÁNDEZ-NAVARRO, J. M. ALTZIBAR y col. Validating a breast cancer score in Spanish women. The MCC-Spain study. *Scientific Reports*, **8** (2018), 3036.

[103] S. PURCELL, B. NEALE, K. TODD-BROWN, L. THOMAS, M. A. R. FERREIRA y col. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, **81** (2007), 559-575.

[104] P. HEWITSON, P. GLASZIOU, E. WATSON, B. TOWLER y L. IRWIG. Cochrane systematic review of colorectal cancer screening using the fecal occult blood test (hemoccult): An update. *The American Journal of Gastroenterology*, **103** (2008), 1541-1549.

[105] D. A. LIEBERMAN, J. L. WILLIAMS, J. L. HOLUB, C. D. MORRIS, J. R. LOGAN y col. Race, ethnicity, and sex affect risk for polyps >9 mm in average-risk individuals. *Gastroenterology*, **147** (2014), 351-358.

[106] U. PETERS, S. BIEN y N. ZUBAIR. Genetic architecture of colorectal cancer. *Gut*, **64** (2015), 1623-1636.

[107] J. A. USHER-SMITH, F. M. WALTER, J. D. EMERY, A. K. WIN y S. J. GRIFFIN. Risk prediction models for colorectal cancer: A systematic review. *Cancer Prevention Research*, **9** (2016), 13-26.

[108] B. MÜLLER, A. WILCKE, A. L. BOULESTEIX, J. BRAUER, E. PASSARGE y col. Improved prediction of complex diseases by common genetic markers: State of the art and further perspectives. *Human Genetics*, **135** (2016), 259-272.

[109] R. MÅNSSON, M. M. JOFFE, W. SUN y S. HENNESSY. On the estimation and use of propensity scores in case–control and case–cohort studies. *American Journal of Epidemiology*, **166** (2007), 332-339.

[110] N. CHATTERJEE, J. SHI y M. GARCÍA-CLOSAS. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, **17** (2016), 392-406.

[111] T. FAWCETT. An introduction to ROC analysis. *Pattern Recognition Letters*, **27** (2006), 861-874.

[112] M. G. DUNLOP, A. TENESA, S. M. FARRINGTON, S. BALLEREAU, D. H. BREWSTER y col. Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42 103 individuals. *Gut*, **62** (2013), 871-881.

[113] J. M. YARNALL, D. J. M. CROUCH y C. M. LEWIS. Incorporating non-genetic risk factors and behavioural modifications into risk prediction models for colorectal cancer. *Cancer Epidemiology*, **37** (2013), 324-329.

[114] K. J. JUNG, D. WON, C. JEON, S. KIM, T. I. KIM y col. A colorectal cancer prediction model using traditional and genetic risk scores in Koreans. *BMC Genetics*, **16** (2015), 49.

[115] L. HSU, J. JEON, H. BRENNER, S. B. GRUBER, R. E. SCHOEN y col. A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology*, **148** (2015), 1330-1339.

[116] J. MACARTHUR, E. BOWLER, M. CEREZO, L. GIL, P. HALL y col. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, **45** (2016), D896-D901.

[117] J. FERLAY, I. SOERJOMATARAM, R. DIKSHIT, S. ESER, C. MATHERS y col. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, **136** (2015), E359-E386.

[118] K. J. PIENTA y P. S. ESPER. Risk factors for prostate cancer. *Annals of Internal Medicine*, **118** (1993), 793-803.

[119] E. D. CRAWFORD. Epidemiology of prostate cancer. *Urology*, **62** (2003), 3-12.

[120] D. G. BOSTWICK, H. B. BURKE, D. DJAKIEW, S. EULING, S. HO y col. Human prostate cancer risk factors. *Cancer*, **101** (2004), 2371-2490.

[121] L. N. KOLONEL. Fat, meat, and prostate cancer. *Epidemiologic Reviews*, **23** (2001), 72-81.

[122] L. N. KOLONEL, D. ALTSHULER y B. E. HENDERSON. The multiethnic cohort study: Exploring genes, lifestyle and cancer risk. *Nature Reviews Cancer*, **4** (2004), 519-527.

[123] R. EELES, C. GOH, E. CASTRO, E. BANCROFT, M. GUY y col. The genetic epidemiology of prostate cancer and its clinical implications. *Nature Reviews Urology*, **11** (2014), 18-31.

[124] P. LICHTENSTEIN, N. V. HOLM, P. K. VERKASALO, A. ILIADOU, J. KAPRIO y col. Environmental and heritable factors in the causation of cancer: Analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England Journal of Medicine*, **343** (2000), 78-85.

[125] D. J. SCHAID. The complex genetic epidemiology of prostate cancer. *Human Molecular Genetics*, **13** (2004), R103-R121.

[126] J. H. MOORE. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*, **56** (2003), 73-82.

[127] R. CULVERHOUSE, B. K. SUAREZ, J. LIN y T. REICH. A perspective on epistasis: Limits of models displaying no main effect. *The American Journal of Human Genetics*, **70** (2002), 461-471.

[128] B. FREIDLIN, G. ZHENG, Z. LI y J. L. GASTWIRTH. Trend tests for case–control studies of genetic markers: Power, sample size and robustness. *Human Heredity*, **53** (2002), 146-152.

[129] W. G. COCHRAN. Some methods for strengthening the common chi-squared tests. *Biometrics*, **10** (1954), 417-451.

[130] P. ARMITAGE. Tests for linear trends in proportions and frequencies. *Biometrics*, **11** (1955), 375-386.

[131] P. G. BAGOS. Genetic model selection in genome-wide association studies: Robust methods and the use of meta-analysis. *Statistical Applications in Genetics and Molecular Biology*, **12** (2013), 285-308.

[132] C. PALAZUELOS, M. ZORRILLA y J. LLORCA. Toward a network-based approach to modeling epistatic interactions in genome-wide association studies. *30th IEEE International Symposium on Computer-Based Medical Systems*, (2017), 225-230.

[133] M. MEYNERS. Equivalence tests: A review. *Food Quality and Preference*, **26** (2012), 231-245.

[134] I. D. BROSS. Why proof of safety is much more difficult than proof of hazard. *Biometrics*, **41** (1985), 785-793.

[135] S. P. MILLARD. Proof of safety vs. proof of hazard. *Biometrics*, **43** (1987), 719-725.

[136] D. G. ALTMAN y J. M. BLAND. Absence of evidence is not evidence of absence. *The British Medical Journal*, **311** (1995), 485.

[137] P. ALDERSON. Absence of evidence is not evidence of absence. *The British Medical Journal*, **328** (2004), 476-477.

[138] W. J. WESTLAKE. Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences*, **61** (1972), 1340-1341.

[139] W. J. WESTLAKE. Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, **32** (1976), 741-744.

[140] R. A. CRIBBIE, J. A. GRUMAN y C. A. ARPIN-CRIBBIE. Recommendations for applying tests of equivalence. *Journal of Clinical Psychology*, **60** (2004), 1-10.

[141] E. WALKER y A. S. NOWACKI. Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, **26** (2011), 192-196.

[142] J. OCAÑA, M. P. SÁNCHEZ, A. SÁNCHEZ y J. L. CARRASCO. On equivalence and bioequivalence testing. *Statistics and Operations Research Transactions*, **32** (2008), 151-176.

[143] D. J. SCHUIRMANN. On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics*, **37** (1981), 617.

[144] D. J. SCHUIRMANN. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, **15** (1987), 657-680.

[145] C. LAUZON y B. CAFFO. Easy multiplicity control in equivalence testing using two one-sided tests. *The American Statistician*, **63** (2009), 147-154.

[146] S. ANDERSON y W. W. HAUCK. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics: Theory and Methods*, **12** (1983), 2663-2692.

[147] W. W. HAUCK y S. ANDERSON. A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics*, **12** (1984), 83-91.

[148] H. I. PATEL y G. D. GUPTA. A problem of equivalence in clinical trials. *Biometrical Journal*, **26** (1984), 471-474.

[149] D. M. ROCKE. On testing for bioequivalence. *Biometrics*, **40** (1984), 225-230.

[150] A. MARTÍN ANDRÉS. On testing for bioequivalence. *Biometrical Journal*, **32** (1990), 125-126.

[151] R. L. BERGER y J. C. HSU. Bioequivalence trials, intersection–union tests and equivalence confidence sets. *Statistical Science*, **11** (1996), 283-319.

[152] M. D. PERLMAN y L. WU. The emperor's new tests. *Statistical Science*, **14** (1999), 355-369.

[153] C. J. MECKLIN. A comparison of equivalence testing in combination with hypothesis testing and effect sizes. *Journal of Modern Applied Statistical Methods*, **2** (2003), 329-340.

[154] W. W. TRYON y C. LEWIS. An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*, **13** (2008), 272-277.

[155] T. HOTHORN, F. BRETZ y P. WESTFALL. Simultaneous inference in general parametric models. *Biometrical Journal*, **50** (2008), 346-363.

[156] J. RÖHMEL. On familywise type I error control for multiplicity in equivalence trials with three or more treatments. *Biometrical Journal*, **53** (2011), 914-926.

[157] S. Y. HUA, S. XU y R. B. D'AGOSTINO. Multiplicity adjustments in testing for bioequivalence. *Statistics in Medicine*, **34** (2015), 215-231.

[158] H. CAMPBELL y P. GUSTAFSON. Conditional equivalence testing: An alternative remedy for publication bias. *PLOS ONE*, **13** (2018), e0195145.

[159] J. L. ROGERS, K. I. HOWARD y J. T. VESSEY. Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, **113** (1993), 553-565.

[160] L. E. BARKER, E. T. LUMAN, M. M. MCCAULEY y S. Y. CHU. Assessing equivalence: An alternative to the use of difference tests for measuring disparities in vaccination coverage. *American Journal of Epidemiology*, **156** (2002), 1056-1061.

[161] A. DINNO. Comment on "The effect of same-sex marriage laws on different-sex marriage: Evidence from the Netherlands". *Demography*, **51** (2014), 2343-2347.

[162] J. J. DOLADO, M. C. OTERO y M. HARMAN. Equivalence hypothesis testing in experimental software engineering. *Software Quality Journal*, **22** (2014), 215-238.

[163] E. LUKACS. A characterization of the normal distribution. *The Annals of Mathematical Statistics*, **13** (1942), 91-93.

[164] T. WALDHOER y H. HEINZL. Combining difference and equivalence test results in spatial maps. *International Journal of Health Geographics*, **10** (2011), 3.

[165] J. P. SHAFFER. Multiple hypothesis testing. *Annual Review of Psychology*, **46** (1995), 561-584.

[166] J. J. GOEMAN y A. SOLARI. Multiple hypothesis testing in genomics. *Statistics in Medicine*, **33** (2014), 1946-1978.

[167] S. WELLEK, K. A. B. GODDARD y A. ZIEGLER. A confidence-limit-based approach to the assessment of Hardy–Weinberg equilibrium. *Biometrical Journal*, **52** (2010), 253-270.

[168] J. K. WITTKE-THOMPSON, A. PLUZHNIKOV y N. J. COX. Rational inferences about departures from Hardy–Weinberg equilibrium. *The American Journal of Human Genetics*, **76** (2005), 967-986.

[169] S. WELLEK. Tests for establishing compatibility of an observed genotype distribution with Hardy–Weinberg equilibrium in the case of a biallelic locus. *Biometrics*, **60** (2004), 694-703.

[170] J. E. WIGGINTON, D. J. CUTLER y G. R. ABECASIS. A note on exact tests of Hardy–Weinberg equilibrium. *The American Journal of Human Genetics*, **76** (2005), 887-893.

[171] N. L. DIMOU, K. D. TSIRIGOS, A. ELOFSSON y P. G. BAGOS. GWAR: Robust analysis and meta-analysis of genome-wide association studies. *Bioinformatics*, **33** (2017), 1521-1527.

[172] D. J. BALDING. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **7** (2006), 781-791.

[173] P. C. B. PHILLIPS y J. Y. PARK. On the formulation of Wald tests of nonlinear restrictions. *Econometrica*, **56** (1988), 1065-1083.

[174] G. W. OEHLERT. A note on the delta method. *The American Statistician*, **46** (1992), 27-29.

[175] R. M. KARP. Reducibility among combinatorial problems. *Complexity of Computer Computations*. 1972, 85-103.

[176] R. BAR-YEHUDA y S. EVEN. A linear-time approximation algorithm for the weighted vertex cover problem. *Journal of Algorithms*, **2** (1981), 198-203.

[177] D. KRAUTWURST. Human olfactory receptor families and their odorants. *Chemistry & Biodiversity*, **5** (2008), 842-852.

[178] B. PAVAN y A. DALPIAZ. Odorants could elicit repair processes in melanized neuronal and skin cells. *Neural Regeneration Research*, **12** (2017), 1401-1404.

[179] X. WANG, L. NI, D. CHANG, H. LU, Y. JIANG y col. Cyclic AMP-responsive element-binding protein (CREB) is critical in autoimmunity by promoting Th17 but inhibiting Treg cell differentiation. *EBioMedicine*, **25** (2017), 165-174.

[180] S. GOROKHOVA, S. BIBERT, K. GEERING y N. HEINTZ. A novel family of transmembrane proteins interacting with β subunits of the Na,K-ATPase. *Human Molecular Genetics*, **16** (2007), 2394-2410.

[181] P. ROMANIA, A. CASTELLANO, C. SURACE, A. CITTI, M. A. DE IORIS y col. High-resolution array CGH profiling identifies Na/K transporting ATPase interacting 2 (NKAIN2) as a predisposing candidate gene in neuroblastoma. *PLOS ONE*, **8** (2013), e78481.

[182] I. DONNER, T. KIVILUOTO, A. RISTIMÄKI, L. A. AALTONEN y P. VAHTERISTO. Exome sequencing reveals three novel candidate predisposition genes for diffuse gastric cancer. *Familial Cancer*, **14** (2015), 241-246.

[183] S. CHARRASSE, F. COMUNALE, S. DE ROSSI, A. ECHARD y C. GAUTHIER-ROUVIÈRE. Rab35 regulates cadherin-mediated adherens junction formation and myoblast fusion. *Molecular Biology of the Cell*, **24** (2013), 234-245.

[184] P. D. ALLAIRE, M. S. SADR, M. CHAINEAU, E. S. SADR, S. KONEFAL y col. Interplay between Rab35 and Arf6 controls cargo recycling to coordinate cell adhesion and migration. *Journal of Cell Science*, **126** (2013), 722-731.

[185] X.-W. LI, B.-Y. SHI, Q.-L. YANG, J. WU, H.-M. WU y col. Epigenetic regulation of CDH1 exon 8 alternative splicing in gastric cancer. *BMC Cancer*, **15** (2015), 954.

[186] N. TAMEHIRO, Z. MUJAWAR, S. ZHOU, D. Z. ZHUANG, T. HORNEMANN y col. Cell polarity factor Par3 binds SPTLC1 and modulates monocyte serine palmitoyltransferase activity and chemotaxis. *Journal of Biological Chemistry*, **284** (2009), 24881-24890.

[187] B. J. SCHMIEDEL, C. A. SCHEIBLE, T. NUEBLING, H.-G. KOPP, S. WIRTHS y col. RANKL expression, function, and therapeutic targeting in multiple myeloma and chronic lymphocytic leukemia. *Cancer Research*, **73** (2013), 683-694.

[188] S. D'ORONZO, J. BROWN y R. COLEMAN. The role of biomarkers in the management of bone-homing malignancies. *Journal of Bone Oncology*, **9** (2017), 1-9.

Código fuente

Este apéndice contiene el código fuente con que se implementó, en el paquete de *software* estadístico Stata® 14, el protocolo de construcción de redes de interacciones SNP-SNP para estudios de casos y controles confeccionado en la SECCIÓN 5.1.

script_1.do

Calcula las frecuencias genotípicas absolutas por efecto (caso/control) de un conjunto de SNP y las guarda en el fichero `‘mcc’.st1`, donde `‘mcc’` es la ruta de la base de datos de Stata® que contiene el genotipado de los pacientes del estudio MCC-Spain.

```
use "‘mcc’"

tempname out
file open ‘out’ using "‘mcc’.st1", write

ds Identifier - Phenotype, not
```

```
foreach SNP of varlist `r(varlist)' {
  tabulate `SNP' Phenotype, matcell(A)
  file write `out' "`SNP'" _n

  forvalues col = 1 / 2 {
    forvalues row = 1 / 3 {
      file write `out' (A[`row',`col']) _n
    }
  }
}
```

script_2_6.do

Según se ejecute después de `script_1.do` o `script_5.do` (`stn' = 1` y 5, respectivamente), selecciona los SNP de `mcc'.st1` o las interacciones SNP-SNP de `mcc'-*.st5` que siguen un modelo genético en su asociación con la enfermedad. Guarda los primeros en `mcc'.st2`, solo si están en equilibrio de Hardy-Weinberg, y las interacciones en `mcc'-*.st6`.

```
local alpha = 0.05
local   m   = 1E06

if `stn' != 1 {
  local m   = 5E11
}
```

```
tempname in
file open ‘in’  using "‘mcc’.st‘stn’", read

tempname out
file open ‘out’ using "‘mcc’.st‘=‘stn’+1’", write

input _Pheno _SNP_1 _SNP_2
     0      0      0
     0      1      0
     0      0      1
     1      0      0
     1      1      0
     1      0      1
     0      0      0
     0      1      0
     1      0      0
     1      1      0
end

generate _Count = .

file read ‘in’ line
while !r(eof) {
  local SNP = "‘line’"

  forvalues i = 1 / 6 {
    file read ‘in’ line
```

```
  replace _Count = `line' in `i'
}

logit _Pheno _SNP_1 _SNP_2 [fw = _Count] if _n <= 6, iterate(20)

if e(converged) {
  local    X    (_b[_SNP_1])
  local    Y    (_b[_SNP_2])
  matrix  Cov = e(V)
  local  Var_X  ( Cov[1,1] )
  local  Var_Y  ( Cov[2,2] )
  local Cov_X_Y ( Cov[1,2] )
  local  Z  = sqrt(`Y'^2 * `Var_X' - 2 * `X' * `Y' * `Cov_X_Y' + ///
                  `X'^2 * `Var_Y')

  local A_D =                                  `Y' * `X' /      `Z'
  local A_O =  (2 * `X' - `Y') *               `Y'       / (2 * `Z')
  local A_R = -                  (`X' - `Y') * `Y'       /      `Z'
  local D_O =  (2 * `X' - `Y') *                     `X' /      `Z'
  local D_R =                    (`X' - `Y') *       `X' /      `Z'
  local O_R =  (2 * `X' - `Y') * (`X' - `Y')             /      `Z'

  local TdA = `A_O'
  local TdD = `D_R'
  local TdO = `A_O'
  local TdR = `D_R'
  local TeA = min(`A_D', `A_R')
```

```
local TeD = min(‘D_O’, ‘A_D’)
local TeO = min(‘D_O’, ‘O_R’)
local TeR = min(‘O_R’, ‘A_R’)

local CCM = 0
local Mod = ""
foreach model in "A" "D" "O" "R" {
  local Pd = 2 * normal(-abs(‘Td‘model’’))
  local Pe =     normal(-    ‘Te‘model’’ )
  local CC = 0

  if !missing(‘Pd’) & !missing(‘Pe’) {
    if ‘alpha’ / ‘m’ <= ‘Pd’ & ‘Pd’ >= ‘Pe’ & ‘Pe’ <= ‘alpha’ {
      local CC =  log(‘alpha’ / max(‘Pe’, ‘alpha’ / ‘m’) - ///
                      ‘alpha’ / min(‘Pd’, ‘alpha’) + 1) / log(‘m’)
    }
    else if ‘alpha’ / ‘m’ <= ‘Pe’ & ‘Pe’ >= ‘Pd’ & ‘Pd’ <= ‘alpha’ {
      local CC = -log(‘alpha’ / max(‘Pd’, ‘alpha’ / ‘m’) - ///
                      ‘alpha’ / min(‘Pe’, ‘alpha’) + 1) / log(‘m’)
    }
  }

  if ‘CC’ > ‘CCM’ {
    local CCM = ‘CC’
    local Mod = "‘model’"
  }
}
```

```
if ‘CCM’ > 0 & "‘Mod’" != "A" {
  local a = _Count[1]
  local b = _Count[2]
  local c = _Count[3]

  if ‘stn’ == 1 {
    local delta = log(1 + 2/5)
    local   X   = log(‘b’ / (2 * sqrt(‘a’ * ‘c’)))
    local   Y   = sqrt(1 / (4 * ‘a’) + 1 / ‘b’ + 1 / (4 * ‘c’))
    local  Td_  = ‘X’ / ‘Y’
    local  Te_  = (‘delta’ - abs(‘X’)) / ‘Y’
    local  Pd   = 2 * normal(-abs(‘Td_’))
    local  Pe   =     normal(-    ‘Te_’ )
    local  HWE  = 0

    if !missing(‘Pd’) & !missing(‘Pe’) {
      if ‘alpha’ / ‘m’ <= ‘Pd’ & ‘Pd’ >= ‘Pe’ & ‘Pe’ <= ‘alpha’ {
        local HWE =  log(‘alpha’ / max(‘Pe’, ‘alpha’ / ‘m’) - ///
                         ‘alpha’ / min(‘Pd’, ‘alpha’) + 1) / log(‘m’)
      }
      else if ‘alpha’ / ‘m’ <= ‘Pe’ & ‘Pe’ >= ‘Pd’ & ‘Pd’ <= ‘alpha’ {
        local HWE = -log(‘alpha’ / max(‘Pd’, ‘alpha’ / ‘m’) - ///
                         ‘alpha’ / min(‘Pe’, ‘alpha’) + 1) / log(‘m’)
      }
    }

    local scr = min(max(0, ‘HWE’), ‘CCM’)
```

```
  if ‘scr’ > 0 {
    file write ‘out’ "‘SNP’" _n "‘Mod’" _n (‘scr’) _n
  }
}
else {
  local d = _Count[4]
  local e = _Count[5]
  local f = _Count[6]

  if "‘Mod’" == "D" {
    replace _Count =    ‘a’    in 7
    replace _Count = ‘b’ + ‘c’ in 8
    replace _Count =    ‘d’    in 9
    replace _Count = ‘e’ + ‘f’ in 10
  }
  else if "‘Mod’" == "O" {
    replace _Count = ‘a’ + ‘c’ in 7
    replace _Count =    ‘b’    in 8
    replace _Count = ‘d’ + ‘f’ in 9
    replace _Count =    ‘e’    in 10
  }
  else {
    replace _Count = ‘a’ + ‘b’ in 7
    replace _Count =    ‘c’    in 8
    replace _Count = ‘d’ + ‘e’ in 9
    replace _Count =    ‘f’    in 10
  }
```

```
        logit _Pheno _SNP_1 [fw = _Count] if _n > 6, iterate(20)

        if e(converged) {
          local p = 2 * normal(-abs(_b[_SNP_1] / _se[_SNP_1]))
          file write ‘out’ "‘SNP’" _n "‘Mod’" _n (‘p’) _n
        }
      }
    }
  }

  file read ‘in’ line
}
```

script_3.do

Ejecutado después de `script_2_6.do`, guarda los SNP de ‘`mcc`’`.st2` en ‘`mcc`’`.st3` ordenados de acuerdo con una puntuación de calidad basada en el mínimo de los coeficientes de coherencia con el modelo genético seleccionado y el equilibrio de Hardy-Weinberg.

```
tempname in
file open ‘in’ using "‘mcc’.st2", read

local obs = 0
file read ‘in’ line
while !r(eof) {
  local ++obs
```

```
  file read ‘in’ line
}

file close ‘in’
local obs = ‘obs’ / 3
set   obs   ‘obs’

generate  SNP  = ""
generate  Mod  = ""
generate Score = .

local n = 1
file open ‘in’ using "‘mcc’.st2", read
file read ‘in’ line
while !r(eof) {
  if missing(SNP[‘n’]) {
    replace SNP = "‘line’" in ‘n’
  }
  else {
    if missing(Mod[‘n’]) {
      replace  Mod  = "‘line’" in ‘n’
    }
    else {
      replace Score = -‘line’  in ‘n’
      local ++n
    }
  }
```

```
  file read ‘in’ line
}

sort Score

tempname out
file open ‘out’ using "‘mcc’.st3", write
forvalues i = 1 / ‘obs’ {
  file write ‘out’ (SNP[‘i’]) _n (Mod[‘i’]) _n
}
```

script_4.do

Ejecutado después de `script_3.do`, recodifica los SNP de `‘mcc’.st3` siguiendo las funciones genotípicas correspondientes e identifica aquellos potencialmente redundantes. Calcula las frecuencias genotípicas absolutas por efecto (caso/control) de todas las interacciones SNP-SNP posibles y las guarda en `‘mcc’.st4`.

```
use "‘mcc’"

tempname in
file open ‘in’ using "‘mcc’.st3", read

local i = 0
file read ‘in’ SNP
```

```
while !r(eof) {
  file read ‘in’ Mod
  char ‘SNP’[SNP] ‘SNP’
  char ‘SNP’[Mod] ‘Mod’
  char ‘SNP’[Wei]  1

  local ++i
  char ‘SNP’[Num] ‘i’
  rename ‘SNP’ _V_‘i’

  if "‘Mod’" == "R" {
    replace _V_‘i’ = _V_‘i’ == 2 if !missing(_V_‘i’)
  }
  else if "‘Mod’" == "O" {
    replace _V_‘i’ = _V_‘i’ == 1 if !missing(_V_‘i’)
  }
  else {
    replace _V_‘i’ = _V_‘i’ != 0 if !missing(_V_‘i’)
  }

  file read ‘in’ SNP
}

order _V_*, after(Phenotype) sequential
keep Identifier - _V_‘i’

tempname out
```

```
file open ‘out’ using "‘mcc’.st4", write

generate _Inter = .

local p = ‘i’
forvalues i = ‘=‘p’-1’ (-1) 1 {
  forvalues j = ‘p’ (-1) ‘=‘i’+1’ {
    replace  _Inter = _V_‘i’ + _V_‘j’
    tabulate _Inter Phenotype, matcell(C)

    file write ‘out’ "_W_‘i’_‘j’" _n

    forvalues col = 1 / 2 {
      forvalues row = 1 / 3 {
        local c = C[‘row’,‘col’]

        if missing(‘c’) {
          local c = 0
        }

        file write ‘out’ (‘c’) _n
      }
    }

    tabulate _V_‘i’ _V_‘j’, matcell(A)
    if (A[1,2] + A[2,1]) / (A[1,1] + A[1,2] + A[2,1] + A[2,2]) < 0.01 {
      local eps = min(‘_V_‘i’[Wei]’, ‘_V_‘j’[Wei]’)
```

```
      char _V_‘i’[Wei] ‘=‘_V_‘i’[Wei]’-‘eps’’
      char _V_‘j’[Wei] ‘=‘_V_‘j’[Wei]’-‘eps’’
    }
  }
}

drop _Inter
save "‘mcc’", replace
```

script_5.do

Ejecutado después de `script_4.do`, elimina las interacciones con SNP potencialmente redundantes y divide las restantes en ficheros `‘mcc’-*.st5` de un máximo de un millón de interacciones SNP-SNP.

```
local n = 5E05
set obs ‘n’

generate SNP_1 = .
generate SNP_2 = .

forvalues col = 0 / 1 {
  forvalues row = 0 / 2 {
    generate C‘col’‘row’ = .
  }
}
```

```
tempname in
file open ‘in’ using "‘mcc’.st4", read

local i = 1
file read ‘in’ inter
while !r(eof) {
  local ids = substr("‘inter’", 4, .)
  local pos = strpos("‘ids’", "_")

  if ‘i’ > ‘n’ {
    local n = ‘n’ + 5E05
    set obs ‘n’
  }

  replace SNP_1 = real(substr("‘ids’", 1, ‘pos’ - 1)) in ‘i’
  replace SNP_2 = real(substr("‘ids’", ‘pos’ + 1, .)) in ‘i’

  forvalues col = 0 / 1 {
    forvalues row = 0 / 2 {
      file read ‘in’ count
      replace C‘col’‘row’ = ‘count’ in ‘i’
    }
  }

  local ++i
  file read ‘in’ inter
}
```

```
drop if missing(SNP_1) & missing(SNP_2)

compress
save "‘mcc’-aux"

use "‘mcc’"
foreach SNP of varlist _V_* {
  if !‘‘SNP’[Wei]’ {
    use "‘mcc’-aux"

    local id = real(substr("‘SNP’", 4, .))
    drop if SNP_1 == ‘id’ | SNP_2 == ‘id’

    save "‘mcc’-aux", replace
    use  "‘mcc’"
  }
}

use "‘mcc’-aux"
keep if C00 & C01 & C02 & C10 & C11 & C12

local out = 0
file open f‘out’ using "‘mcc’-‘out’.st5", write

local n = 1
forvalues k = 1 / ‘=_N’ {
  if ‘k’ > ‘n’ * 1E06 {
```

```
    file close f‘out’

    local ++out
    local ++n

    file open f‘out’ using "‘mcc’-‘out’.st5", write
  }

  local i = SNP_1[‘k’]
  local j = SNP_2[‘k’]

  file write f‘out’ "_W_‘i’_‘j’" _n

  forvalues col = 0 / 1 {
    forvalues row = 0 / 2 {
      file write f‘out’ (C‘col’‘row’[‘k’]) _n
    }
  }
}

file close f‘out’
```

script_7.do

Ejecutado después de `script_2_6.do`, guarda las interacciones SNP-SNP de ‘`mcc`’-*.`st6` con valor de $p < 3{,}5 \cdot 10^{-9}$ en ‘`mcc`’.`st7`.

```
local n = 5E05
set obs ‘n’

generate Int = ""
generate Mod = ""
generate  P  = .

local files : dir "‘dir’" files "‘mcc’*.st6"
local   i   = 1
local   k   = 1
foreach file in ‘files’ {
  local in = "in‘k’"
  tempname   ‘in’
  file open ‘‘in’’ using "‘dir’\\‘file’", read

  file read ‘‘in’’ int
  while !r(eof) {
    file read ‘‘in’’ mod
    file read ‘‘in’’  p

    if ‘i’ > ‘n’ {
      local n = ‘n’ + 5E05
      set obs ‘n’
    }

    replace Int = "‘int’" in ‘i’
    replace Mod = "‘mod’" in ‘i’
```

```
    replace  P  =   ‘p’   in ‘i’

    local ++i
    file read ‘‘in’’ int
  }

  local ++k
}

drop if missing(Int)

compress
save "‘dir’\\‘mcc’-aux2"

tempname out
file open ‘out’ using "‘dir’\\‘mcc’.st7", write

keep if P < 3.5E-9
forvalues i = 1 / ‘=_N’ {
  local int = Int[‘i’]
  local ids = substr("‘int’", 4, .)
  local pos = strpos("‘ids’", "_")
  local s_1 = substr("‘ids’", 1, ‘pos’ - 1)
  local s_2 = substr("‘ids’", ‘pos’ + 1, .)

  file write ‘out’ "‘s_1’" _n "‘s_2’" _n (Mod[‘i’]) _n
}
```

script_8.do

Ejecutado después de `script_7.do`, construye una red de interacciones SNP-SNP a partir de un modelo de regresión logística múltiple, en que las interacciones de `‘mcc’.st7` y sus SNP se ponderan conjuntamente, y la guarda en `‘mcc’.st8`.

```
use "‘mcc’"

local k = 0
foreach SNP of varlist _V_* {
  char ‘SNP’[Reg] 0
  local ++k
}

tempname in
file open ‘in’ using "‘mcc’.st7", read
file read ‘in’ i
while !r(eof) {
  file read ‘in’  j
  file read ‘in’ Mod

  char _V_‘i’[Reg] 1
  char _V_‘j’[Reg] 1

  if "‘Mod’" == "D" {
```

```
    generate _W_‘i’_‘j’ = _V_‘i’ + _V_‘j’ -     _V_‘i’ * _V_‘j’
  }
  else if "‘Mod’" == "0" {
    generate _W_‘i’_‘j’ = _V_‘i’ + _V_‘j’ - 2 * _V_‘i’ * _V_‘j’
  }
  else {
    generate _W_‘i’_‘j’ =                      _V_‘i’ * _V_‘j’
  }

  char  _W_‘i’_‘j’[Mod]   "‘Mod’"
  char  _W_‘i’_‘j’[Source]   ‘i’
  char  _W_‘i’_‘j’[Target]   ‘j’
  order _W_‘i’_‘j’, after(_V_‘k’)

  file read ‘in’ i
}

local nu ""
foreach SNP of varlist _V_* {
  if ‘‘SNP’[Reg]’ {
    local nu "‘nu’‘SNP’ "
  }
}

foreach Int of varlist _W_* {
  local nu "‘nu’‘Int’ "
}
```

```
tempname out
file open  ‘out’ using "‘mcc’.st8", write
file write ‘out’ "Source Target Type Weight Risk" _n

logit Phenotype ‘nu’, level(98.75)
foreach Int of varlist _W_* {
  if normal(-abs(_b[‘Int’] / _se[‘Int’])) < 0.0125 / 2 {
    local source = "‘_V_‘‘Int’[Source]’[SNP]’"
    local target = "‘_V_‘‘Int’[Target]’[SNP]’"
    local weight = abs(_b[‘Int’])
    local  risk  = 0 < _b[‘Int’]

    file write ‘out’ "‘source’ ‘target’ Undirected ‘weight’ ‘risk’" _n
  }
}
```