Trabajo Fin de Máster

Estudio de asociación de variantes genéticas con la miocardiopatía hipertrófica

Sandra Carballido Regueiro

Máster en Técnicas Estadísticas

Curso 2016-2017

Propuesta de Trabajo Fin de Máster

Título en galego: Estudo de asociación de variantes xenéticas coa miocardiopatía hipertrófica
Título en español: Estudio de asociación de variantes genéticas con la miocardiopatía hipertrófica
English title: Association study of genetics variants in hypertrophic cardiomyopathy
Modalidad: B
Autor/a: Sandra Carballido Regueiro, Universidad de A Coruña
Director/a: Ignacio López De Ullibarri Galparsoro, Universidad de A Coruña
Tutor/a: Lorenzo Monserrat Iglesias, Health in Code S.L.
Breve resumen del trabajo: El objetivo es el descubrimiento de SNPs estadísticamente asociados con la miocardiopatía hipertrófica en un estudio de tipo caso-control. Se emplearán metodologías especialmente adaptadas a la alta dimensionalidad del problema, como la regresión logística penalizada (lasso).
Recomendaciones: Preferiblemente titulado en informática, biología o matemáticas.
Otras observaciones: La empresa desea participar en el proceso de selección.

Agradecimientos

En primer lugar, al equipo de Health in Code, por darme la oportunidad de realizar este proyecto con ellos. En especial, gracias a Lorenzo Monserrat, Pablo Iglesias, Roberto Noya y a todo el departamento de Bioinformática, por la estupenda acogida durante estos meses y por la infinita paciencia conmigo.

A mi director, Ignacio López De Ullibarri Galparsoro, por resolverme todas las dudas que me han ido surgiendo en la elaboración de este trabajo.

A mis padres, a mi hermano y a mi familia en general, por el apoyo incondicional durante todos estos años de estudio.

Por último, a Dani, sin cuyo apoyo y ayuda no habría llegado hasta aquí.

Gracias a todos.

Índice general

Resumen

Resumen en español

La búsqueda de asociaciones entre variantes genéticas y ciertas enfermedades es un tema de gran importancia en la actualidad, sobre todo a medida que se avanza en el conocimiento de técnicas de secuenciación del DNA. Este tipo de técnicas proporcionan una gran cantidad de información que necesita ser analizada con metodologías estadísticas especialmente adaptadas al volumen de datos con el que se trabaja. Con tal propósito, es habitual considerar estudios de tipo caso-control que, mediante el uso de los modelos de regresión apropiados, permitan identificar qué variantes aumentan el riesgo de padecer la enfermedad considerada.

El objetivo de este proyecto es describir cómo funcionan estas técnicas estadísticas y mostrar su aplicación en datos reales en el ámbito de la cardiología. Los datos utilizados provienen de pacientes diagnosticados de diversas cardiopatías congénitas, entre ellas, la miocardiopatía hipertrófica, enfermedad en la que se centran los resultados de este análisis. Esta patología es una de las dolencias cardíacas más comunes y es, en la mayoría de ocasiones, hereditaria, por lo que resulta interesante tratar de identificar la componente genética que desencadena su manifestación.

English abstract

The search for associations between genetic variants and certain diseases is currently an issue of great importance, especially as one advances in the knowledge of DNA sequencing techniques. These types of techniques provide a large amount of information that needs to be analyzed with statistical methodologies especially adapted to the volume of data that we are working with. To this purpose, it is usual to consider case-control studies that, through the use of suitable regression models, allow us to identify which variants increase the risk of suffering the evaluated disease.

The objective of this project is to describe how these statistical techniques work and to show their application in real data in cardiology. The data that we use come from patients diagnosed with different congenital heart diseases, including hypertrophic cardiomyopathy, the illness in which the results of this analysis are focused. This pathology is one of the most common heart diseases and is, in most cases, hereditary, so it is interesting to try to identify the genetic component that triggers its manifestation.

Capítulo 1

Introducción

Un estudio de asociación genética es un análisis cuya finalidad es tratar de identificar asociaciones entre variantes o regiones genéticas y una enfermedad. La metodología utilizada se basa generalmente en un tipo de estudios denominados caso-control, en los que se comparan las frecuencias de aparición de las distintas variantes en dos grupos de pacientes: un grupo control, formado por individuos sanos; y un grupo de casos, formado por individuos que padecen la enfermedad que es objeto de estudio.

Las variantes genéticas más habituales en estos análisis son los polimorfismos de nucleótido simple o SNPs, variantes comunes que se encuentran en al menos el 1 % de la población. Los estudios de asociación genética resultan de gran utilidad cuando se trata de buscar SNPs o genes que puedan estar relacionados con la manifestación de enfermedades complejas, es decir, patologías determinadas por diversas causas (varios genes y factores ambientales). Es el caso de las cardiopatías y, más concretamente, de la miocardiopatía hipertrófica.

La miocardiopatía hipertrófica es una enfermedad cardiovascular hereditaria caracterizada por el engrosamiento de las paredes del corazón. Es una de las cardiopatías más frecuentes y, actualmente, su causa es desconocida. Para tratar de identificar la componente genética asociada a esta dolencia, y así mejorar en el diagnóstico y pronóstico de la misma, es necesario estudiar adecuadamente el DNA de los pacientes afectados. Esa tarea es generalmente llevada a cabo por expertos de diversas instituciones y empresas, como es el caso de Health in Code S.L.

Health in Code S.L. es una spin-off de la Universidad de A Coruña creada en 2006 y especializada en el diagnóstico genético de enfermedades cardiovasculares. Cuenta con un equipo multidisciplinar de más de 60 expertos formado por biólogos, cardiólogos e informáticos. Su base de datos consta de más de 90000 pacientes y 500000 variantes genéticas analizadas. Estos datos se obtienen del análisis de parte del genoma de individuos susceptibles de padecer una enfermedad cardiovascular, bien por la presencia de antecedentes familiares, bien por la manifestación de algún síntoma.

Tras la secuenciación del DNA y su posterior análisis bioinformático, estos datos, utilizados para proporcionar un diagnóstico al paciente, se guardan para otros estudios futuros, como el que se lleva a cabo en este trabajo.

En este proyecto se utilizarán parte de los datos de Health in Code S.L. para realizar un estudio caso-control con la finalidad de determinar asociaciones entre SNPs y la miocardiopatía hipertrófica, a través de la utilización de la metodología estadística apropiada.

En el Capítulo 2 se proporcionan las nociones básicas sobre biología y genética necesarias para comprender el desarrollo posterior del trabajo.

En el Capítulo 3 se describen brevemente los fundamentos de los estudios caso-control. Además, se explican las técnicas empleadas para analizar la base de datos antes de realizar el estudio con el fin de evitar la detección de falsos positivos. En este sentido, existen diversas condiciones que tienen que cumplir los datos y que se deben examinar antes de seguir con el análisis. También, debido a la gran cantidad de tests realizados simultáneamente en algunos modelos empleados, es necesario considerar técnicas de control del error tipo I ante este problema de multiplicidad de contrastes. En este capítulo

se explicará brevemente en qué consisten estos procedimientos.

El Capítulo 4 describe las técnicas estadísticas utilizadas en el estudio caso-control. En primer lugar, se explican los modelos de regresión logística. Un modelo de regresión logística múltiple permite incorporar varias variables explicativas y así analizar el efecto de todas ellas, sin embargo, lo que ocurre a menudo, y sobre todo en estudios genéticos, es que el número de covariables es mayor que el número de individuos, algo que impide la utilización de este tipo de ajustes, excepto si se utilizan métodos de penalización. Los modelos de regresión penalizada descritos en este trabajo son *lasso* y *elastic net*.

Por último, los capítulos 5 y 6 describen la aplicación a los datos de toda la metodología explicada en los capítulos anteriores y los resultados obtenidos en relación a la miocardiopatía hipertrófica.

Capítulo 2

Fundamentos biológicos

2.1. Genética y DNA

El genoma humano está organizado en 23 pares de cromosomas: 22 autosomas y un par de cromosomas sexuales (XX, en mujeres; XY, en hombres). Los miembros de cada par se denominan cromosomas homólogos. En un individuo la mayoría de las células corporales (células diploides) contienen esos 46 cromosomas, excepto las células haploides, como los espermatozoides o los óvulos, que solamente contienen una copia del genoma, es decir, 23 cromosomas.

Los cromosomas se duplican durante la interfase del ciclo celular. Tras esta replicación del DNA, durante la meiosis, se formarán los gametos, en un proceso en el que se llevarán a cabo dos divisiones celulares. Esto da lugar a una reducción del número de cromosomas en cada gameto pasando de ser diploides a haploides. Durante la primera división celular los cromosomas homólogos duplicados se emparejan formando una conexión física en donde tiene lugar la recombinación genética. Así, dos alelos que están en el mismo cromosoma en el padre pueden no estar en el mismo cromosoma en los gametos y, por lo tanto, estos reciben información de ambos homólogos de cada par. Por el contrario, la mitosis (el proceso análogo a la meiosis, pero en células somáticas) da lugar a células genéticamente iguales y, si bien pueden producirse mutaciones debido a errores durante la copia de DNA, estas no se transmitirán a la siguiente generación de individuos, algo que sí ocurre con las que se producen en la meiosis.

Cada cromosoma está formado por dos hebras de DNA enrollado en una doble hélice. Cada hebra es un largo polímero constituido por unidades repetitivas: los nucleótidos. A su vez, cada nucleótido está formado por tres componentes: ácido fosfórico, desoxirribosa y una base nitrogenada. En el DNA existen cuatro bases diferentes: dos purinas, adenina (A) y guanina (G); y dos pirimidinas, citosina (C) y timina (T). Las dos hebras se mantienen ligadas entre sí por la unión de bases opuestas. Esa unión ocurre específicamente entre A y T y entre G y C. La Figura 2.1 ilustra esta disposición de las bases en las hebras de DNA.

Tras la duplicación del DNA tiene lugar un procedimiento denominado transcripción, que consiste en la síntesis de una molécula de RNAm utilizando como molde una hebra de DNA. La información codificada en las bases nitrogenadas del RNA servirá para sintetizar una cadena de aminoácidos que dará lugar a las proteínas en el proceso conocido como traducción. Las bases de los nucleótidos del RNA se encuentran agrupadas funcionalmente en grupos de tres llamados codones, y estos codifican directamente los aminoácidos que forman la proteína objetivo. Una descripción más detallada de este proceso puede verse en Stram et al. (2014).

Cada gen es una región de DNA que influye en una característica particular de un organismo a través de la producción de proteínas. La totalidad del material genético, incluidos todos los genes, es lo que se denomina genoma. Por su parte, el genotipo es la combinación de los dos alelos presentes en los cromosomas homólogos. Menos del 2 % del genoma humano codifica directamente cadenas de aminoácidos. Generalmente, la secuencia codificante para una proteína está constituida en el DNA

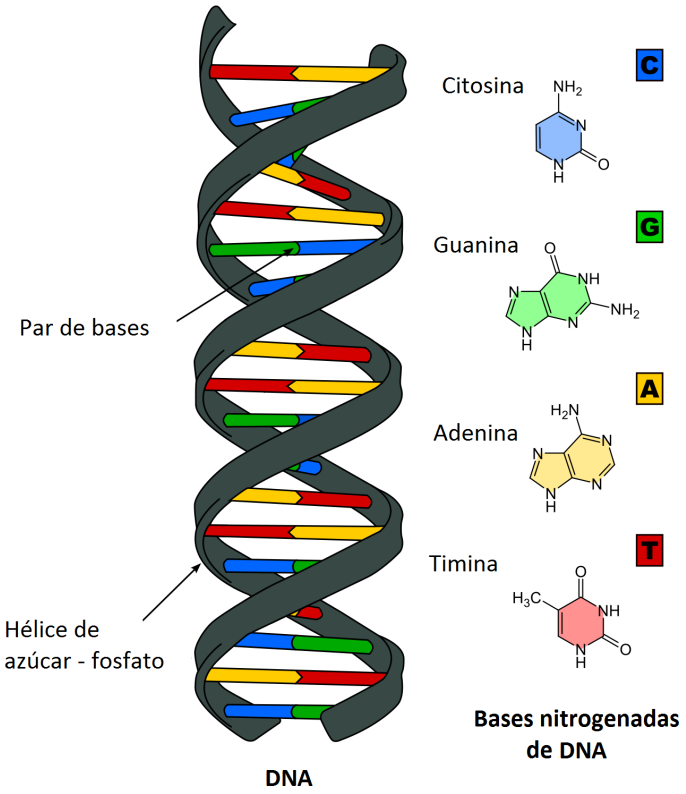

Figura 2.1: Disposición de las bases nitrogenadas en la doble hélice de DNA. Imagen obtenida de http://openbio.cl/biologia-sintetica/el-adn-como-soporte-de-la-informacion-genetica/

por muchos exones intercalados con intrones no codificantes. Los intrones se eliminan en un proceso llamado splicing de RNA.

En el genoma pueden producirse variaciones, es decir, cambios que alteren la secuencia de nucleótidos del DNA. Este tipo de cambios pueden modificar o no la cadena de aminoácidos que compone las proteínas resultantes y, en ocasiones, dar lugar a enfermedades o a cambios en un fenotipo observable.

En concreto, un polimorfismo de nucleótido simple (SNP) es una variación en la secuencia de DNA que afecta a una sola base de una posición específica de una hebra de DNA. Una de estas variaciones debe darse al menos en un 1 % de la población para ser considerada como un SNP. Si no llega al 1 % no se considera SNP, si no mutación puntual. La mayoría de los SNPs en humanos constan de dos alelos A y a, donde A, $a \in \{\text{A,C,G,T}\}$. Las composiciones alélicas posibles serán entonces AA (homocigoto dominante), Aa (heterocigoto) y aa (homocigoto recesivo).

Los SNPs se originan como errores de copia durante la meiosis o bien como resultado de un daño en el DNA, sin embargo, la ocurrencia de una nueva variación en una localización cromosómica dada es un evento raro, es decir, la inmensa mayoría de pares de bases de DNA se copian con completa fidelidad durante cualquier meiosis, y la mayoría del DNA dañado se repara por completo. Debido a la baja frecuencia con que ocurren nuevas mutaciones, la gran mayoría de variantes en el DNA de una persona son heredadas de uno (o ambos) de sus progenitores, que a su vez las heredaron de uno de sus padres, y así sucesivamente. Es mucho más probable que una variante nueva se extinga a que se convierta en frecuente.

2.2. Miocardiopatía hipertrófica

La miocardiopatía hipertrófica (MCH) es una enfermedad del corazón caracterizada por el aumento del grosor de sus paredes (hipertrofia), que no se debe a causas extramusculares.

Esta enfermedad ocurre cuando las células del músculo cardíaco se agrandan y hacen que las paredes de los ventrículos (generalmente el ventrículo izquierdo) se engrosen. Muchas veces el tamaño del ventrículo permanece normal, pero el engrosamiento de las paredes puede bloquear el flujo sanguíneo fuera del ventrículo, lo que se conoce como MCH obstructiva.

A veces, la pared que separa los lados izquierdo y derecho del corazón se hace más densa y se hincha en el ventrículo izquierdo. Esto puede bloquear el flujo sanguíneo fuera de ese ventrículo, lo que hace que necesite un esfuerzo extra para bombear sangre.

La MCH puede también afectar a la válvula mitral del corazón, haciendo que la sangre refluya a través de esta válvula. A veces, el músculo espesado del corazón no bloquea el flujo sanguíneo fuera del ventrículo izquierdo. Esto se conoce como MCH no obstructiva.

En ambos casos, el espesor del músculo hace el interior del ventrículo izquierdo más pequeño, por lo que contendrá menos sangre. Las paredes del ventrículo pueden endurecerse, y como resultado, el ventrículo es menos capaz de relajarse y llenarse de sangre. La Figura 2.2 ilustra la diferencia entre un corazón sano y un corazón con hipertrofia.

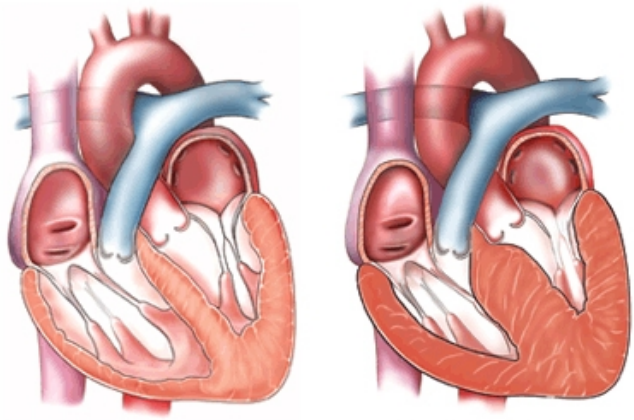

Figura 2.2: A la izquierda, paredes de un corazón no afectado de MCH; a la derecha, paredes del corazón engrosadas debido a la MCH. Imagen obtenida de: http://columbiasurgery.org/conditions-and-treatments/hypertrophic-cardiomyopathy-and-heart-failure

Se trata de una enfermedad cardiovascular muy común que puede afectar a hombres y a mujeres de cualquier edad. Es, de hecho, una de las causas más frecuentes de muerte súbita en jóvenes.

Algunas personas con MCH no muestran síntomas y la enfermedad puede no afectar a su calidad de vida, sin embargo, en muchos casos se producen síntomas y complicaciones severas, como arritmias o incapacidad para hacer ejercicio.

Tal y como describe la American Heart Association, la MCH es generalmente hereditaria. Está causada por un cambio en algunos genes que codifican proteínas del músculo del corazón, aunque también puede desarrollarse con el tiempo debido a alta presión sanguínea o al envejecimiento. Algunas otras enfermedades, como la diabetes, pueden estar asociadas a la MCH, sin embargo, la causa de la enfermedad es desconocida.

Se estima que esta enfermedad se manifiesta en 1 de cada 500 individuos, siendo por tanto una de las cardiopatías más frecuentes.

2.3. Patrones de herencia

A la hora de estudiar una enfermedad hereditaria es importante identificar de qué modo se transmite de generación en generación.

Las enfermedades hereditarias pueden clasificarse inicialmente en dos tipos: mendelianas (o monogénicas) y complejas (o poligénicas). Las enfermedades mendelianas están causadas por la alteración del DNA en un solo gen. Ejemplos de enfermedades mendelianas son el síndrome de Marfan o la fibrosis quística. Por otra parte, las enfermedades complejas se deben a variaciones en diversos genes y a distintos factores ambientales como la edad o el sexo. La MCH es un ejemplo de enfermedad compleja.

En cuanto a la transmisión de la enfermedad, se dice que una enfermedad tiene un patrón de herencia autosómico dominante cuando es suficiente con recibir un alelo anormal de uno de los padres para heredar la enfermedad y además esta herencia es independiente del sexo del descendiente. Se trata de un modelo de herencia común en enfermedades mendelianas.

Por otro lado, cuando se trata de un patrón de herencia aditivo, el riesgo de padecer la enfermedad sería menor si el genotipo es heterocigoto (presencia de la variante en uno de los dos alelos), que si es homocigoto (presencia de la variante en ambos alelos).

Cuando el patrón de herencia de una enfermedad es desconocido, en estudios de asociación como el que se realiza en este trabajo es habitual asumir un modelo de tipo aditivo, por ser el más general. Sin embargo, cuando se trata de la MCH, se presupone habitualmente que su patrón de herencia es de tipo autosómico dominante, aunque se trata de una enfermedad compleja, por lo que asumir un patrón de herencia mendeliano puede no ser del todo adecuado.

Capítulo 3

Conceptos básicos en estudios de asociación genética

Antes de realizar un estudio de asociación genética es necesario evaluar la calidad de los datos que se van a utilizar para así evitar, en la medida de lo posible, la aparición de falsas asociaciones. En este capítulo se describirán los estudios de asociación que se utilizan en genética y las condiciones que deben cumplir los datos para que este tipo de análisis sean válidos.

3.1. Estudios caso-control

Los estudios de asociación genética pretenden buscar regiones genómicas, genes o variantes que puedan contribuir a la presencia de una determinada enfermedad mediante el análisis de la relación estadística existente entre el estatus de la enfermedad y esas variantes genéticas. Las variantes más utilizadas en este tipo de estudios son los polimorfismos de nucleótido simple (SNPs), ya definidos en el Capítulo 2.

Los estudios de asociación son una gran herramienta para identificar una posible relación entre ciertos SNPs y una enfermedad, si bien es cierto que la experiencia en estudios genéticos ha demostrado que en muchas enfermedades complejas (como la diabetes, los trastornos psiquiátricos o las cardiopatías) el riesgo de enfermedad se asocia a muchas variantes diferentes, y cada una de esas variantes tiene por sí sola un efecto muy pequeño.

Existen distintos tipos de estudios genéticos según los datos que se analizan (ver Foulkes, 2009). Los estudios de asociación del genoma completo (GWAS) son los más recientes y consisten en el análisis de las variaciones genéticas a lo largo de todo el genoma. Aunque no todos los SNPs son analizados, sí es necesaria la caracterización de un gran número de ellos distribuidos a través de todo el genoma. Los GWAS están identificando numerosas nuevas asociaciones en genes que a priori no parecían candidatos para tal efecto. Sin embargo, son análisis bastante costosos, por lo que es frecuente seguir realizando estas investigaciones en genes candidatos o bien en determinadas regiones del genoma.

La metodología que siguen estos análisis es bastante simple y se basa en un tipo de estudio denominado estudio caso-control, en el cual se dispone de un grupo de individuos afectados por una determinada enfermedad y un grupo de individuos que no presentan esa misma enfermedad. Para cada uno de esos individuos se estudia su genotipo (o parte de este) y a partir de esos datos se estiman las posibles asociaciones. Además del estudio del genoma de los individuos, puede que alguna otra información sobre los pacientes sea accesible, como por ejemplo, la edad o el sexo. Cuando esto ocurre se pueden obtener inferencias que tengan en cuenta el posible efecto de estas covariables.

Aunque la metodología de los estudios caso-control es, en principio, sencilla, la capacidad de un estudio de asociación genética a la hora de identificar verdaderas asociaciones depende en gran medida de la calidad de los datos (ver Turner et al., 2011).

Incluso las técnicas estadísticas más simples pueden llevar a resultados erróneos si los datos no se eligen correctamente. En este sentido, existen diversas condiciones sobre los datos que se deben comprobar antes de llevar a cabo un estudio caso-control. Estas condiciones son las siguientes: la frecuencia alélica de las variantes y el equilibrio de Hardy-Weinberg. Asimismo, es también importante controlar la estratificación de la población y analizar qué regiones del genoma se encuentran en desequilibrio de ligamiento. Además, al estudiar grandes bases de datos, es necesario tener en cuenta que la realización de múltiples contrastes de forma simultánea puede dar lugar a un incremento de errores de tipo I (o falsos positivos). A continuación, se discuten más en detalle todos estos aspectos.

3.2. Frecuencia del alelo menor

La frecuencia del alelo menor (MAF) se define como la frecuencia con la que aparece el alelo secundario, a, en un determinado locus, dentro de una población. A partir de ese valor se definen los polimorfismos de nucleótido simple (SNPs), considerando únicamente aquellos cambios de base cuya frecuencia es mayor del 1 %. Es importante filtrar variantes y considerar solamente las que se encuentran con una frecuencia mayor del 1 % ya que la potencia estadística es extremadamente baja para variantes raras (ver Turner et al., 2011).

3.3. Estratificación de la población

En estudios de asociación como el realizado en este proyecto se asume que cualquier diferencia genotípica entre casos y controles se debe a la diferencia entre el estatus de la enfermedad y no a la presencia de subpoblaciones. Sin embargo, lo que ocurre generalmente es que los grupos de casos y controles se encuentran estratificados por la estructura genética de la población, lo que derivará en el aumento de falsas asociaciones, como puede inferirse de algunos ejemplos de Tian et al. (2008). Lo que ocurre en esos casos es que la población subyacente es en realidad una mezcla de poblaciones ancestralmente distintas con diferentes valores de prevalencia de enfermedad y de frecuencia alélica. Por ejemplo, considerando el supuesto en que se tienen dos poblaciones y en una de ellas hay una alta prevalencia de cierta enfermedad e, independientemente, una frecuencia más alta de cierto SNP que en la otra población, entonces esto dará lugar a la presencia de más enfermos en la primera población y, consecuentemente, una falsa asociación entre ese SNP y la enfermedad.

Existen diversas técnicas para el estudio y la corrección de la estratificación de la población. Una forma frecuente de evitar el sesgo que proporciona la estratificación es asegurar que las muestras del estudio pertenecen a una sola población. Es habitual disponer de los datos relativos a la etnia de los pacientes, por lo que estos datos podrían utilizarse para identificar los estratos o seleccionar una muestra de individuos homogénea.

Para las situaciones en que esto no es posible, existen otras metodologías estadísticas que permiten detectar y ajustar la presencia de estratificación en la población (Tian et al., 2008).

Una de estas metodologías se denomina “Genomic Control” y consiste en estimar un factor de inflación del test estadístico utilizado en el estudio de asociación para, a continuación, ajustar todos los contrastes según ese factor.

Otra metodología utilizada en este contexto es el software STRUCTURE, que usa los datos genotípicos para inferir la estructura de la población y luego realiza tests de asociación entre cada una de las subpoblaciones detectadas.

Además, cuando el tamaño de la muestra es muy grande y, por tanto, la probabilidad de confusión es alta, se puede utilizar la librería EIGENSTRAT. Este tipo de análisis consiste en calcular componentes principales para detectar los estratos de la población.

3.4. Equilibrio de Hardy-Weinberg

El equilibrio de Hardy-Weinberg (HWE) establece que, en un locus bialélico, las frecuencias alélicas, p (para el alelo dominante: A) y q (para el alelo secundario: a), y las frecuencias genotípicas, $f(AA)$, $f(Aa)$ y $f(aa)$, se mantienen constantes de generación en generación, a menos que algún otro proceso (migraciones, mutaciones, etc.) actúe sobre ellas. Así, el HWE determina que $f(AA) = p^2$, $f(Aa) = 2pq$ y $f(aa) = q^2$.

La fórmula del HWE, $p^2+2pq+q^2 = 1$, se utiliza para calcular las frecuencias genotípicas esperadas y, mediante un test, generalmente basado en una distribución χ^2, estas se comparan con las frecuencias observadas para contrastar si hay diferencias significativas entre ambas.

Cuando un SNP muestra desviación de este equilibrio puede significar que una o varias de las hipótesis del equilibrio no se cumplen, aunque generalmente lo que sugiere son errores de genotipado. Estas desviaciones también pueden ser indicativas de asociaciones a la enfermedad que es objeto de estudio, por este motivo, se debe estudiar el equilibrio de Hardy-Weinberg solamente en la población control.

Aunque para contrastar el equilibrio de Hardy-Weinberg lo más habitual es considerar un estadístico que utilice la distribución χ^2 para comprobar la bondad de ajuste con respecto a las proporciones Hardy-Weinberg, cuando las frecuencias genotípicas son muy bajas puede que este estadístico no sea el más adecuado, pues las suposiciones asintóticas de la distribución χ^2 no se cumplirían. En ese caso, es más razonable utilizar un test exacto, como el test exacto propuesto por Fisher (1922), o su generalización a tablas de contingencia rectangulares, como el test exacto de Levene y Haldane, que se describe en Wigginton et al. (2005).

El p-valor en un test exacto es generalmente calculado como la suma de las probabilidades de todas las posibles tablas de contingencia tanto o más extremas que la actual.

A continuación se describen estos dos tests tal como se aplican al contraste del equilibrio de Hardy-Weinberg.

Test χ^2 de Pearson

El test más usado para estudiar el equilibrio de Hardy-Weinberg es el test χ^2 de Pearson. El procedimiento para la construcción de este test es el que se muestra a continuación.

Sean N_{AA}, N_{aa} y N_{Aa} el número de individuos homocigotos AA, homocigotos aa y heterocigotos observados, respectivamente. Sea $N = N_{AA}+N_{aa}+N_{Aa}$, la suma total de individuos. Con estos valores se calculan las frecuencias alélicas p y q:

$$p = \frac{2N_{AA} + N_{Aa}}{2N}$$
$$q = 1 - p = \frac{2N_{aa} + N_{Aa}}{2N}$$

Los valores esperados para N_{AA}, N_{aa} y N_{Aa} vendrán dados respectivamente por $E_{AA} = Np^2$, $E_{Aa} = 2Npq$ y $E_{aa} = Nq^2$. El estadístico χ^2 se construye del siguiente modo:

$$\sum_{C\in\{AA,\, aa,\, Aa\}} \frac{(N_C - E_C)^2}{E_C} \sim \chi^2_1, \tag{3.1}$$

donde N_C indica los valores observados y E_C los valores esperados para cada composición alélica.

Este estadístico se compara con el correspondiente cuantil de una distribución χ^2 para determinar si la hipótesis nula del HWE se rechaza o no.

Cuando la hipótesis nula es rechazada para algún SNP se concluye que la variante considerada no está en equilibrio de Hardy-Weinberg.

Test exacto

Cuando las frecuencias en una tabla de contingencia son bajas (el valor de alguna celda es menor de 5) es más aconsejable usar un test exacto de Fisher, que no utiliza la aproximación χ^2. El p-valor del test exacto de Fisher se basa en la suma de las probabilidades exactas de ver las frecuencias observadas o un valor más extremo en la dirección de la hipótesis alternativa. El test de Fisher se utiliza en tablas 2×2, pero cuando la tabla de contingencia es rectangular pueden utilizarse otros tests, como por ejemplo un test basado en la aproximación obtenida por Levene y Haldane (ver Weir, 1996).

Este test exacto para el equilibrio de Hardy-Weinberg se basa en la distribución condicional del número de heterocigotos (N_{Aa}) sobre la cantidad de individuos con el alelo secundario ($N_a = 2N_{aa} + N_{Aa}$). Esta distribución viene dada por:

$$P(N_{Aa}|N_a) = \frac{N!N_a!N_A!2^{N_{Aa}}}{\frac{1}{2}(N_a - N_{Aa})!N_{Aa}!\frac{1}{2}(N_A - N_{Aa})!(2N)!}$$

donde $N_A = 2N_{AA} + N_{Aa}$.

El p-valor del test exacto se calcula generalmente como la suma de probabilidades de todas las posibles muestras igual o más extremas que la muestra observada, conocido el recuento de alelos de la muestra observada.

3.5. Desequilibrio de ligamiento

En numerosas ocasiones, en los estudios de asociación, puede ocurrir que un SNP que aparentemente está asociado con la enfermedad objetivo del análisis, en realidad no sea un SNP funcional, es decir, no sea un causante directo de la enfermedad. Cuando esto ocurre probablemente sea debido a que esa variante se encuentra en desequilibrio de ligamiento (LD) con la verdadera variante funcional.

El LD se define como la asociación no aleatoria de alelos en loci adyacentes (ver Stram et al., 2014). Hay ciertas regiones en el genoma donde las recombinaciones son fenómenos raros y otras donde las recombinaciones son comunes. Los alelos entre los que han ocurrido un bajo número de recombinaciones a través de las generaciones se dice que están en desequilibrio de ligamiento.

A continuación se describen las medidas utilizadas para cuantificar el LD entre SNPs.

Medidas de LD

Se consideran dos posiciones en un cromosoma donde A y a son los posibles alelos en uno de los lugares y B y b son los posibles alelos en la otra posición. Una medida utilizada para caracterizar el LD viene dada por $D = p_{AB} - p_A p_B$, donde p_{AB} es la frecuencia de aparición de la combinación AB, y p_A y p_B son las frecuencias de los alelos A y B en sus respectivas posiciones. En presencia de LD, se tiene que el valor de D es distinto de cero.

Para cuantificar con exactitud este desequilibrio, existen dos medidas, funciones de D, que se definen a continuación.

La primera medida considerada es D', y se determina dividiendo D por su máximo valor posible tal que:

$$D' = \frac{|D|}{D_{\text{máx}}},$$

donde $D_{\text{máx}}$ viene dado por

$$D_{\text{máx}} = \begin{cases} \text{mín}(p_A p_b, p_a p_B), & \text{si } D \geq 0, \\ \text{mín}(p_A p_B, p_a p_b), & \text{si } D < 0, \end{cases}$$

con p_a y p_b las frecuencias de los alelos secundarios en cada posición.

Es interesante el hecho de que $D' = 1$ si, y solo si, dos SNPs no fueron separados por recombinación (o mutación recurrente) durante el historial de la muestra. En ese caso, se estaría ante la presencia de LD completo.

La otra medida utilizada es el r^2. Esta medida, basada en el estadístico χ^2 de Pearson, representa la correlación entre alelos de dos lugares diferentes, y se calcula dividiendo D^2 entre el producto de las cuatro frecuencias alélicas en los dos loci, es decir,

$$r^2 = \frac{D^2}{p_A p_B p_a p_b}.$$

Un r^2 cercano a 1 indica que los SNPs están en alto desequilibrio de ligamiento. Para que r^2 sea exactamente 1 se necesitaría, además de la ausencia de recombinación (como en el caso de D'), también que las frecuencias alélicas en las dos posiciones sean idénticas.

Mapas de LD

Cuando se trata de identificar LD entre SNPs, es útil tratar de representar de modo gráfico el valor de las medidas descritas anteriormente, sobre todo para facilitar su rápida interpretación cuando se están utilizando grandes cantidades de datos. En este sentido, es interesante utilizar un mapa de calor. En este tipo de mapas, los valores numéricos altos se representan en celdas con colores más oscuros y los valores más bajos con colores claros. En el caso del LD, los mapas de calor representan el valor de D' o de r^2 entre cada par de SNPs. La Figura 3.1 es una representación del LD mediante un mapa de calor utilizando como medida el r^2. Los SNPs se distribuyen en el mapa según su distancia física en el cromosoma.

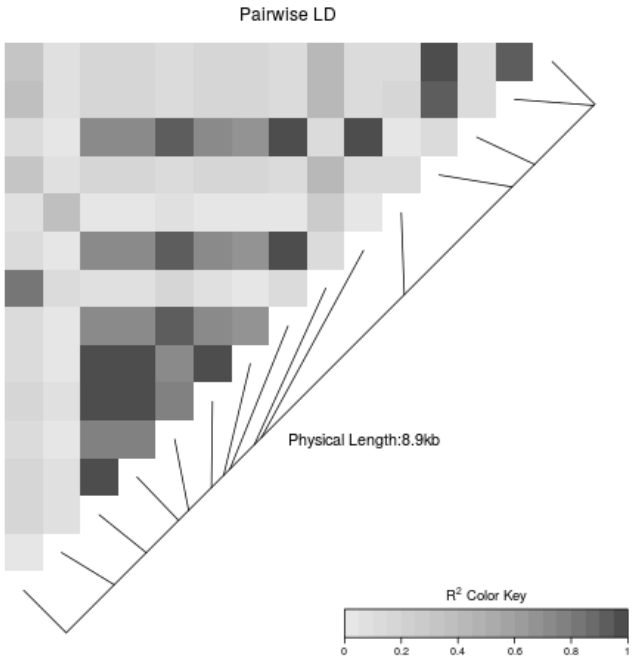

Figura 3.1: Ejemplo de mapa de LD obtenido con R utilizando como medida de desequilibrio el r^2.

3.6. Comparaciones múltiples

En estudios de asociación genética, generalmente se quiere determinar la presencia de asociaciones entre una enfermedad y una gran cantidad de SNPs. Para eso es necesario realizar contrastes de hipótesis asumiendo cierto nivel de error. De entre las hipótesis estudiadas (H_0 o hipótesis nula y H_1 o hipótesis alternativa), el resultado obtenido tras el contraste puede ser o no la hipótesis verdadera. En este sentido, existen los cuatro posibles sucesos de la Tabla 3.1:

		Realidad	
		H_0 es cierta	**H_1 es cierta**
Contraste	**H_0 es cierta**	No hay error	Error tipo II (o falso negativo)
	H_1 es cierta	Error tipo I (o falso positivo)	No hay error

Tabla 3.1: Tipos de errores según los resultados obtenidos tras un constraste de hipótesis.

El p-valor proporciona la probabilidad de alcanzar el resultado obtenido suponiendo que H_0 es cierta. Generalmente se establece un nivel de signficación de $\alpha = 0.05$, y se rechazará H_0 si el p-valor es menor que α; sin embargo, en estudios genéticos, donde se realizan múltiples contrastes, los falsos positivos (o errores tipo I) tienden a acumularse.

Para un buen ajuste de las comparaciones múltiples se utilizan métodos que controlan una de estas dos tasas de error: el *Family-Wise Error Rate* (FWER) o tasa de error a nivel de familia y el *False Discovery Rate* (FDR) o tasa de falsos descubrimientos.

Family-Wise Error Rate

El FWER se define como la probabilidad de cometer por lo menos un error de tipo I, es decir,

$$\text{FWER} = Pr(V \geq 1)$$

donde V es el número de errores de tipo I.

Si los contrastes son independientes cada uno con una tasa de error tipo I de α, entonces el FWER del conjunto de contrastes será $(1 - (1 - \alpha)^M)$, donde M es el número de hipótesis contrastadas. Se tiene, por ejemplo, que para $\alpha = 0.05$ y M relativamente muy grande, $Pr(V \geq 1) \approx 1$. Por eso es importante desarrollar métodos que controlen el FWER a niveles adecuados como, por ejemplo, FWER=0.05.

Un método clásico de control del FWER es la corrección de Bonferroni. Este método calcula el p-valor ajustado a partir de la división de cada p-valor por el número total de comparaciones que se realizan. Así, serán significativos solamente los contrastes en que el p-valor ajustado es menor de 0.05. Este método es bastante conservador y, aunque disminuye el número de falsos positivos, tiende a aumentar los errores de tipo II.

False Discovery Rate

El FDR se define como la proporción esperada de hipótesis nulas que son ciertas sobre las que se han declarado significativas, esto es,

$$\text{FDR} = E\left(\frac{V}{R}\right)$$

donde R es el número de test que son declarados significativos. Si $R = 0$, entonces se define $\dfrac{V}{R} = 0$.

Para controlar el FDR el ajuste más utilizado en genética es el propuesto por Benjamini y Hochberg (1995). Este método consiste en ordenar los p-valores de los contrastes $p_{(1)} \leq \cdots \leq p_{(M)}$ y calcular k tal que

$$k = \text{máx}\left\{ i : p_{(i)} \leq \frac{i}{M}\alpha \right\}, \tag{3.2}$$

donde α es el nivel al que se quiere controlar el FDR. Se rechazarán los contrastes asociados a cada uno de los p-valores $p_{(1)}, \ldots, p_{(k)}$.

El ajuste de Benjamini y Hochberg (B-H) para controlar el FDR asume independencia entre los contrastes. Benjamini y Yekutieli (2001) prueban que este procedimiento permite también controlar el FDR en presencia de diversas estructuras de dependencia entre los estadísticos del contraste.

Este criterio es menos conservador que el FWER, propiedad conveniente en los estudios genómicos, por lo que el FDR será el utilizado en la aplicación a los datos.

Cabe destacar que la presencia de falsos positivos no es solo frecuente en contrastes de asociación, si no que esta problemática se traslada también a cualquier tipo de comparación realizada de forma masiva. Es por tanto necesario realizar una corrección de los p-valores también en otros apartados del estudio, por ejemplo, en el análisis del equilibrio de Hardy-Weinberg, en donde se realizan miles de contrastes para detectar la presencia de desequilibrio.

Capítulo 4

Modelos de regresión en genética

Los modelos lineales generalizados (GLMs) extienden los modelos de regresión ordinarios al caso en que la variable respuesta sigue una distribución no normal. Estos modelos de regresión constan de tres componentes: una componente aleatoria, una componente sistemática y una función link.

La componente aleatoria de un GLM está formada por n variables aleatorias independientes Y_i, con $i = 1, \ldots, n$ y su distribución de probabilidad asociada, la cual pertenece a la familia exponencial.

La componente sistemática se refiere a las variables explicativas que se incluyen en el modelo a través de una combinación lineal, llamada función predictora, y que se puede expresar como un vector $(\eta_1, \ldots, \eta_n)$ tal que

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, \qquad i = 1, \ldots, n$$

donde x_{ij} es el valor del j-ésimo predictor en el i-ésimo individuo y β_j sus coeficientes respectivos. Para simplificar la notación no se ha incluido en los siguientes desarrollos el término constante β_0.

En forma vectorial se tiene

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta},$$

donde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$, $\boldsymbol{X}$ una matriz de covariables de dimensión $n \times p$, con filas $\boldsymbol{x_i} = (x_{i1}, \ldots x_{ip})$, $i = 1, \ldots, n$.

La tercera componente del GLM es la función link, que conecta la componente aleatoria con la componente sistemática, y que es una función del valor esperado de Y_i. Sea $\mu_i = E(Y_i)$, $i = 1, \ldots, n$, el modelo relaciona μ_i con η_i a través de $\eta_i = g(\mu_i)$, donde la función g es monótona y diferenciable. Se tiene entonces que

$$g(\mu_i) = \boldsymbol{x_i}\boldsymbol{\beta}, \qquad i = 1, \ldots, n. \tag{4.1}$$

Uno de los GLMs más importantes es el modelo de regresión logística, que se explica a continuación, según el desarrollo de Agresti (2002).

4.1. Regresión logística

Cuando la variable respuesta toma solamente dos posibles valores representando la presencia o ausencia de una cierta característica de interés, como por ejemplo, una enfermedad, los modelos de regresión logística (enmarcados dentro de los GLMs) son los más adecuados.

Sea y_i la respuesta observada para el i-ésimo individuo, $i = 1, \ldots, n$, tal que

$$y_i = \begin{cases} 1 \text{ si el individuo presenta la enfermedad} \\ 0 \text{ en otro caso} \end{cases}$$

Entonces y_i es una observación de la variable aleatoria Y_i que toma los valores uno y cero con probabilidades π_i y $1-\pi_i$, respectivamente, y además, $E(Y_i) = \pi_i$. La variable Y_i sigue una distribución de Bernoulli (dentro de la familia exponencial) con parámetro π_i y se tiene que

$$Pr\{Y_i = y_i\} = \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

para $y_i = 0, 1$.

La probabilidad π_i tiene que estar entre cero y uno, pero el predictor lineal $\boldsymbol{x_i\beta}$ puede tomar cualquier valor real. Para solucionar este problema es necesario transformar la probabilidad. Se considera primero el *odds* tal que

$$odds_i = \frac{\pi_i}{1-\pi_i} \tag{4.2}$$

definido como el ratio entre casos de éxito, $Pr(Y_i = 1)$, y casos de fallo, $Pr(Y_i = 0)$. Se toman ahora logaritmos, calculando así la función *logit*:

$$\eta_i = logit(\pi_i) = \log\frac{\pi_i}{1-\pi_i}.$$

Por tanto, la función link utilizada en este caso es la función *logit* y transforma la Ecuación (4.1) en el modelo de regresión logística:

$$logit(\pi_i) = \boldsymbol{x_i\beta}, \qquad i = 1, \ldots, n. \tag{4.3}$$

Si Y_i es una variable aleatoria con distribución Bernoulli con probabilidad de éxito π_i, el modelo de la Ecuación (4.3) equivale a:

$$\pi_i = Pr(Y_i = 1) = \frac{\exp(\boldsymbol{x_i\beta})}{1+\exp(\boldsymbol{x_i\beta})}. \tag{4.4}$$

Si se considera un modelo con una única variable explicativa, entonces la interpretación del coeficiente β_1 cuando la variable explicativa es continua es la siguiente: por cada unidad que se incrementa x, el *odds* se multiplica por e^{β_1}. Cuando se trata de una variable explicativa cualitativa, β_1 representa la diferencia entre el logaritmo del *odds* entre un grupo y el grupo de referencia.

En el caso en que se tienen varias variables explicativas, el parámetro β_j se refiere al efecto de x_{ij} sobre el logaritmo del $odds_i$ mientras que los restantes predictores permanecen fijos.

4.1.1. Ecuaciones de verosimilitud

La función de masa de probabilidad conjunta es igual al producto de n funciones de probabilidad de Bernoulli y da lugar a la función de verosimilitud:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n}[\pi_i^{y_i}(1-\pi_i)^{1-y_i}].$$

Se calcula la log-verosimilitud aplicando logaritmos a la función anterior y se tiene:

$$L(\boldsymbol{\beta}) = \log l(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i\log\pi_i + (1-y_i)\log(1-\pi_i)]$$

Las ecuaciones de máxima verosimilitud son resultado de igualar a cero la derivada de la log-verosimilitud respecto a $\boldsymbol{\beta}$. Por lo tanto, teniendo en cuenta que

$$\pi_i = \frac{\exp(\boldsymbol{x_i\beta})}{1+\exp(\boldsymbol{x_i\beta})}$$

y

$$1-\pi_i = \frac{1}{1+\exp(\boldsymbol{x_i\beta})}$$

entonces

$$\begin{aligned} L(\boldsymbol{\beta}) &= \sum_{i=1}^{n}[y_i \log \pi_i + (1-y_i)\log(1-\pi_i)] \\ &= \sum_{i=1}^{n}\left[y_i \log\left(\frac{\exp(\boldsymbol{x_i\beta})}{1+\exp(\boldsymbol{x_i\beta})}\right) + (1-y_i)\left(\frac{1}{1+\exp(\boldsymbol{x_i\beta})}\right)\right] \\ &= \sum_{i=1}^{n}\left[y_i \boldsymbol{x_i\beta} - \log(1+\exp(\boldsymbol{x_i\beta}))\right]. \end{aligned}$$

Derivando ahora con respecto a $\beta_j,\ j=1,\ldots,p$, se tiene:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n}\frac{x_{ij}\exp(\boldsymbol{x_i\beta})}{1+\exp(\boldsymbol{x_i\beta})} = \sum_{i=1}^{n} x_{ij}(y_i-\pi_i)$$

y entonces las ecuaciones de verosimilitud serán

$$\sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n}\pi_i x_{ij} = 0, \qquad j=1,\ldots,p,$$

obteniéndose al resolver el sistema de ecuaciones de verosimilitud anterior la estimación de máxima verosimilitud del vector $\boldsymbol{\beta}$.

Estas ecuaciones son no lineales y, por tanto, son necesarios métodos iterativos para su resolución. El método que utilizan muchos programas es el algoritmo de puntuación de Fisher o Fisher scoring.

4.1.2. Algoritmo de Fisher scoring

Sea $\mathcal{I}(\boldsymbol{\beta})$ la matriz de información de Fisher

$$\mathcal{I}(\boldsymbol{\beta}) = E\left[-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right] = \sum_{i=1}^{n}\boldsymbol{x_i x_i'}\pi_i(1-\pi_i) = \boldsymbol{X'WX} \tag{4.5}$$

con $\boldsymbol{W} = diag[\pi_i(1-\pi_i)]$.

El algoritmo de Fisher scoring calcula de modo iterativo

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(t+1)} &= \hat{\boldsymbol{\beta}}^{(t)} + \mathcal{I}^{-1}(\hat{\boldsymbol{\beta}}^{(t)})U(\hat{\boldsymbol{\beta}}^{(t)}) \\ &= \hat{\boldsymbol{\beta}}^{(t)} + \left[\boldsymbol{X'}\widehat{\boldsymbol{W}}^{\boldsymbol{(t)}}\boldsymbol{X}\right]^{-1}\boldsymbol{X'}\left[\boldsymbol{y}-\hat{\boldsymbol{\pi}}^{\boldsymbol{(t)}}\right],\ t=0,1,2,\ldots \end{aligned}$$

donde $U(\hat{\boldsymbol{\beta}})$ es el gradiente con respecto a $\hat{\boldsymbol{\beta}}$, hasta que se cumpla un criterio de parada.

Aquí, $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1,\ldots,\hat{\pi}_n)$ y $\hat{\pi}_i^{(t)}$, la t-ésima aproximación para π_i, se obtiene de $\boldsymbol{\beta}^{(t)}$ a través de

$$\hat{\pi}_i^{(t)} = \frac{\exp(\sum_{j=1}^{p}\hat{\beta}_j^{(t)}x_{ij})}{1+\exp(\sum_{j=1}^{p}\hat{\beta}_j^{(t)}x_{ij})}.$$

Test de razón de verosimilitudes

Un test de razón de verosimilitudes (LRT) es un método de selección entre dos modelos. Un modelo más simple se compara con otro más complejo para ver si este se ajusta a los datos significativamente mejor.

Sea S un subconjunto de $\{1, \ldots, n\}$. Para un modelo de regresión logística, una hipótesis que interesa contrastar es $H_0 : \beta_{H_0} = \{\beta_j\}_{j \in S} = \{0\}$. La hipótesis alternativa sería $H_1 : \beta_{H_1} = \{\beta_j\}_{j \in S} \neq \{0\}$.

Wilks (1938) demostró que, bajo ciertas condiciones de regularidad, $-2 \log \Lambda$, con $\Lambda = L(\hat{\beta}_{H_0})/L(\hat{\beta}_{H_1})$, donde $\hat{\beta}_{H_0}$ y $\hat{\beta}_{H_1}$ son las estimaciones de máxima verosimilitud de β_{H_0} y β_{H_1}, respectivamente, sigue una distribución χ^2 cuando $n \to \infty$. Los grados de libertad, p, son iguales a la diferencia entre la dimensión del vector de parámetros bajo $H_0 \cup H_1$ y bajo H_0. El estadístico del test de razón de verosimilitudes será entonces:

$$\begin{aligned} -2 \log \Lambda &= -2 \log(L(\hat{\beta}_{H_0})/L(\hat{\beta}_{H_1})) \\ &= -2 \sum_{i=1}^{n} \left[y_i \log \left(\frac{\pi(x_i, \hat{\beta}_{H_0})}{\pi(x_i, \hat{\beta}_{H_1})} \right) + (1 - y_i) \log \left(\frac{1 - \pi(x_i, \hat{\beta}_{H_0})}{1 - \pi(x_i, \hat{\beta}_{H_1})} \right) \right] \sim \chi^2_p. \end{aligned}$$

4.1.3. *Odds ratio*

Tal y como se expresa en la Ecuación (4.2), el *odds* se define como el cociente entre la probabilidad de éxito y la probabilidad de fallo, es decir,

$$odds(Y) = \frac{Pr(Y = 1)}{Pr(Y = 0)}.$$

Una medida interesante para ver el incremento de riesgo que producen las variables explicativas en la variable respuesta resulta de calcular su *odds ratio* (OR).

En el caso del modelo, definimos el OR entre dos poblaciones respecto de una variable dicotómica Y como el cociente de odds:

$$OR = \frac{odds(Y|X = 1)}{odds(Y|X = 0)}.$$

Esto hace posible calcular el OR asociado a la j-ésima covariable como $\exp(\beta_j)$.

Los ORs se utilizan para comparar el *odds* relativo de la ocurrencia de la variable respuesta de interés en función de la covariable estudiada. El valor del OR puede ser utilizado para determinar cuándo el valor de una covariable es un factor de riesgo (o un factor protector) para una determinada variable respuesta, y para cuantificar y comparar la magnitud de esos factores en la respuesta. En general, un OR mayor que uno indicaría un aumento de riesgo en la respuesta y un OR menor que uno indicaría que la covariable es un factor protector respecto a la respuesta.

Intervalos de confianza del OR

Los intervalos de confianza se usan para estimar la precisión del OR. Un intervalo de confianza grande indica que el nivel de precisión del OR es bajo, por lo que puede ser necesario un tamaño muestral más grande; mientras que un intervalo pequeño indica una precisión alta. A diferencia del p-valor, los intervalos de confianza no proporcionan una medida de significación estadística.

El intervalo de confianza al 95 % se utiliza frecuentemente como un indicador de la presencia de significación estadística si no contiene al 1. Sin embargo, no se debe interpretar un OR cuyo intervalo contenga al 1 como indicador de evidencia de falta de asociación entre la covariable y la respuesta.

4.1.4. El problema de la separación

Un problema habitual a la hora de utilizar modelos logísticos con respuesta binaria es el fenómeno de la separación. La existencia de los estimadores de máxima verosimilitud descritos en la sección 4.1.2 están condicionados al comportamiento de la variable respuesta. Así, hay veces en que estos estimadores no pueden ser calculados, como ocurre ante la existencia de separación (ver Ensoy, 2015).

El fenómeno de la separación ocurre en modelos de respuesta binaria cuando una variable explicativa predice perfectamente la variable respuesta. En el caso en que la covariable sea binaria, la separación completa se da cuando, disponiendo los datos en una tabla de contingencia 2×2, dos celdas diagonalmente opuestas están vacías.

La separación cuasicompleta se da cuando una variable explicativa binaria predice perfectamenteuno de los grupos de la variable respuesta, pero no ambos. En el caso en que la covariable sea binaria, la separación cuasicompleta se da cuando una celda de un tabla de contingencia 2×2 está vacía.

El solapamiento, el caso ideal, ocurre cuando no existe una variable explicativa que cumpla alguna de las condiciones anteriores. En presencia de solapamiento, la estimación por máxima verosimilitud existe y proporciona estimaciones razonables de los parámetros. Sin embargo, cuando estamos en presencia de algún tipo de separación, no existe una estimación finita de los coeficientes por máxima verosimilitud.

4.2. Regresión logística penalizada

Cuando el número de covariables es superior al número de observaciones en el modelo de regresión, es decir, si el problema cumple que $p > n$, los parámetros no pueden ser estimados de forma única. Particularmente, en el caso de la regresión logística, esto se debe a que la matriz $\boldsymbol{X'WX}$ de la Ecuación (4.5), necesaria para el cálculo de la función de máxima verosimilitud del modelo, es singular, por lo que su inversa no existe.

Es frecuente, sobre todo en datos de tipo genómico, que el número de covariables sea más grande que el número de observaciones.

En ese caso se puede realizar una regresión logística para cada una de las covariables, como se describe en la Sección 4.1, aunque ese tipo de modelos ignora la posibilidad de la multiplicidad de efectos de las covariables.

Para dar solución a estos problemas se pueden utilizar modelos de regresión penalizada.

Existen diversos métodos de regresión penalizada. Al primero de estos métodos se le conoce como regresión *ridge*, fue propuesto por Hoerl y Kennard (1970) y consiste en asignar una restricción del tipo $\sum_{j=1}^{p} \beta_j^2 < c$, para algún $c > 0$ a la hora de estimar los parámetros. Aunque inicialmente este método se pensó para modelos de regresión lineal, puede extenderse al caso de regresión logística (Le Cessie y Van Houwelingen, 1992).

Tibshirani (1995) propone un nuevo método de regresión penalizada para la estimación de los parámetros denominado *lasso*. Asimismo, Zou y Hastie (2005) plantean un nuevo método de penalización llamado *elastic net*. Ambos métodos realizan, además de la estimación de los coeficientes, selección de variables (algo que no proporciona la regresión *ridge*), lo que hace que estas técnicas sean más interesantes desde un punto de vista epidemiológico por su fácil interpretación.

En las siguientes secciones se describirá más detalladamente en qué consisten estos dos tipos de regresión penalizada.

4.2.1. *Lasso*

La regresión *lasso* (*least absolute shrinkage and selection operator*) utiliza, para la estimación de los coeficientes, la minimización de la suma de cuadrados de los residuos sujeto a que la suma de los valores absolutos de los coeficientes sean menores que una constante. Debido a esta penalización, algunos coeficientes valdrán exactamente cero, por lo que la interpretación del modelo es mucho más sencilla. A continuación se puede ver cómo se define este tipo de regresión.

Sea un modelo de regresión lineal

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

donde $\boldsymbol{y}$ es el vector de las n observaciones de la variable respuesta, $\boldsymbol{X} = (x_{i1}, \ldots, x_{ip})$ es una matriz $n \times p$ formada por las variables predictoras, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ son los coeficientes de la regresión y $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ es un vector de los errores asumiendo $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\sigma_\epsilon^2})$. Al igual que en el desarrollo de los modelos de regresión logística clásica, aquí también se ha prescindido del coeficiente constante β_0 en la descripción del modelo.

Si $n > p$, los valores de los parámetros desconocidos, $\boldsymbol{\beta}$, pueden ser estimados minimizando la suma de residuos al cuadrado

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{mín}} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \right\}. \tag{4.6}$$

Generalmente, todos los parámetros estimados en (4.6) serán distintos de cero. Por este motivo, hacer una interpretación final del modelo será una tarea difícil cuando p es grande. De hecho, si $p > n$, como explicamos anteriormente, entonces las estimaciones anteriores no serán únicas: existirán infinitas soluciones que igualan a cero la función objetivo y además estas soluciones seguramente den lugar a un sobreajuste del modelo.

Una solución a este problema es *penalizar* el proceso de estimación. En la regresión *lasso*, los parámetros estimados se calculan resolviendo el problema de minimización siguiente:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{mín}} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \right\} \text{ sujeto a } ||\boldsymbol{\beta}||_1 \leq t \tag{4.7}$$

donde $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|$ es la norma ℓ_1 de $\boldsymbol{\beta}$, y t es un parámetro escogido por el investigador.

Utilizar la norma ℓ_1 permite que, cuando el valor de t es lo suficientemente pequeño, *lasso* proporcione soluciones dispersas (*sparse*), permitiendo que solo un grupo de parámetros sea distinto de cero, algo que no ocurre con ℓ_q, cuando $q > 1$ (como en la regresión *ridge*). Además, $q = 1$ es el valor más pequeño para el cual el problema a resolver es convexo, algo que, computacionalmente, facilita los cálculos.

El valor de t limita la suma de los valores absolutos de las estimaciones de los parámetros y generalmente se calcula por un procedimiento externo como puede ser la validación cruzada (Hastie et al., 2015).

Reescribiendo el problema en forma lagrangiana se tiene:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{mín}} \left[\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]. \tag{4.8}$$

El problema de regresión resultante es no lineal en y y se traduce en un problema de optimización convexa que puede resolverse con algún algoritmo iterativo. El parámetro λ controla la cantidad de contracción de los coeficientes estimados.

En el caso del modelo de regresión logística, se tendría:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{mín}} \left[\sum_{i=1}^{n} (y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)) + \lambda \sum_{j=1}^{p} |\beta_j| \right]. \tag{4.9}$$

En el modelo de regresión logística clásica, la matriz $\boldsymbol{X'WX}$ no era invertible cuando el número de covariables es más grande que el número de observaciones o cuando existe algún tipo de separación. Aplicando una penalización a la diagonal de $\boldsymbol{X'WX}$, su inversión se hace posible pero se traduce en un gran sesgo en el coeficiente estimado del modelo de regresión, mientras que la estimación del coeficiente en la regresión logística ordinaria es insesgada. Por otro lado, la penalización proporciona varianzas más bajas que en el caso del modelo no penalizado.

Estimación del valor de t

Como se ha indicado anteriormente, t es un parámetro necesario en la penalización cuya estimación se realiza por algún procedimiento externo al cálculo de los coeficientes del modelo. Generalmente, este valor se calcula mediante validación cruzada, siguiendo el procedimiento de Hastie et al. (2015) que se describe a continuación.

El valor de t en el criterio *lasso* controla la complejidad del modelo: por un lado, valores grandes de t dejan libres más parámetros y permiten al modelo adaptarse más a los datos; por otro lado, valores pequeños de t restringen más los parámetros, proporcionando estimaciones de los parámetros dispersos y, por tanto, modelos más fácilmente interpretables pero que se ajustan peor a los datos.

Dejando a un lado la interpretabilidad del modelo sería interesante conocer qué valor de t proporciona el modelo que mejor predice al realizar varios tests independientes sobre la misma población. Tal propiedad se conoce como la habilidad de "generalización"del modelo. Un valor de t muy pequeño puede impedir al *lasso* capturar la señal principal de los datos, mientras que un valor grande puede llevar a un sobreajuste del modelo. En ambos casos, el error de predicción estará inflado. Generalmente existe un valor de t que proporciona un buen equilibrio entre estos dos extremos, y al mismo tiempo, da lugar a un modelo con algunos coeficientes iguales a cero, facilitando su interpretación.

Para seleccionar el mejor valor de t se pueden crear muestras de entrenamiento y conjuntos de prueba obtenidos separando aleatoriamente la muestra para, a continuación, realizar la estimación del modelo usando validación cruzada.

El procedimiento es el siguiente: se divide aleatoriamente el conjunto de datos en un número $k > 1$ de grupos, generalmente $k = 5$ o $k = 10$. Se selecciona uno de los grupos como grupo de prueba y los restantes $k - 1$ grupos serán designados como el grupo de entrenamiento. Se aplica después *lasso* al grupo de entrenamiento para un rango de diferentes valores de t y se utiliza cada modelo ajustado para predecir las respuestas en el grupo de prueba, observando los errores medios de predicción al cuadrado para cada valor de t. Se repite ese proceso k veces, con cada uno de los k grupos cambiando el grupo de prueba y usando los restantes $k-1$ grupos como muestra de entrenamiento. De esta forma, se obtienen k estimaciones diferentes del error de predicción para un rango de valores de t. Esas k estimaciones del error de predicción son promediadas para cada valor de t dando lugar a una curva del error por validación cruzada.

Una buena medida de precisión del estimador es el error cuadrático medio, que se define como:

$$ECM(\hat{\boldsymbol{y}}) = E[(\hat{\boldsymbol{y}} - \boldsymbol{y})^2] = Var(\hat{\boldsymbol{y}}) + [Sesgo(\hat{\boldsymbol{y}})]^2.$$

Limitaciones de la regresión *lasso*

A pesar de las ventajas de la metodología *lasso* con respecto a la regresión logística clásica, este tipo de modelos también presenta ciertas limitaciones:

- En el caso de análisis de datos de alta dimensión, es decir, $p > n$, *lasso* selecciona como máximo n variables debido al problema de optimización convexa que tiene que resolver.
- Si existen un grupo de variables con una alta correlación entre ellas, entonces la regresión *lasso* tiende a seleccionar solamente una de las variables del grupo sin prestar atención al resto.

Estes problemas pueden solucionarse con la regresión penalizada *elastic net*, que se describe a continuación.

4.2.2. *Elastic net*

La regresión penalizada *elastic net* (Zou y Hastie, 2005), al igual que el *lasso*, permite realizar contracción de los coeficientes y al mismo tiempo selección de variables pero además soluciona los problemas del *lasso* mencionados en la sección anterior.

La penalización *elastic net* se encuentra a medio camino entre las penalizaciones *ridge* y *lasso* y se define de forma muy parecida a éstas, pero en este caso, se trata de resolver el problema de programación convexa:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmín}} \left[\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda[(1-\alpha)||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1] \right\}, \tag{4.10}$$

donde $\alpha \in [0, 1]$ es un parámetro que debe ser seleccionado.

Cuando $\alpha = 1$, la expresión $(1-\alpha)||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1$ se reduce a la penalización *lasso*, y con $\alpha = 0$ se trataría de la penalización *ridge*.

Por tanto, la penalización *elastic net* tiene un parámetro adicional α que debe ser determinado. Es habitual escoger $\alpha = 0.5$ o bien crear un conjunto de valores de α y escoger uno mediante validación cruzada. El valor de λ, al igual que en *lasso*, se puede obtener también por validación cruzada. La Figura 4.1 muestra una representación de la estimación de los coeficientes de las regresiones *lasso*, *ridge* y *elastic net*.

El problema de minimización (4.10) puede resolverse utilizando diversos algoritmos, uno de los más efectivos es el descenso por coordenadas.

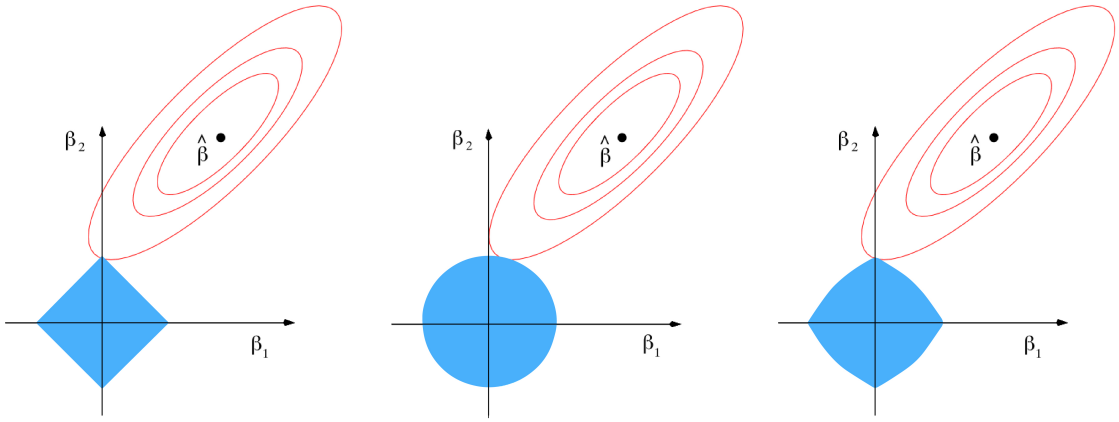

Figura 4.1: Ilustración de la estimación de los coeficientes en regresión *lasso* (izquierda), *ridge* (centro) y *elastic net* (derecha). Las áreas en azul representan las regiones de restricción de los coeficientes y las elipses en rojo representan las curvas de nivel del coeficiente obtenido por mínimos cuadrados.

Capítulo 5

Aplicación a los datos de miocardiopatía hipertrófica

El estudio llevado a cabo en este trabajo pretende detectar la presencia de asociaciones entre SNPs y la MCH. Para ello, en esta parte del trabajo se va a aplicar la metodología explicada en los capítulos anteriores.

En las siguientes secciones se describen los datos que tenemos a nuestra disposición y se realiza el estudio de asociación, llevando a cabo previamente un filtrado inicial de la muestra para cumplir con los requisitos explicados en el Capítulo 3, tales como el equilibrio de Hardy-Weinberg o la estratificación de la población.

5.1. Descripción de los datos

Los datos utilizados en este estudio provienen de pacientes susceptibles de padecer una cardiopatía hereditaria analizados en la empresa Health in Code S.L..

Se ha analizado una librería de genes de una muestra de pacientes mediante secuenciacion NGS (Next-Generation Sequencing). Tras este análisis, y bajo un criterio clínico, se seleccionan los genes candidatos a estar asociados con cardiopatías hereditarias. Un estudio bioinformático después de la secuenciación del DNA del paciente permite detectar la presencia de variantes genéticas mediante la comparación con el genoma de referencia.

Es habitual en este tipo de bases de datos que se disponga de información adicional relevante para su posterior estudio, tales como la edad o el sexo de los pacientes, la frecuencia alélica de las variantes, etc. Sin embargo, en ocasiones, puede que algunos de estos datos no se guarden correctamente o sean desconocidos. Para llevar a cabo este trabajo se han considerado ciertas características y variables adicionales, por lo que se han eliminado del estudio aquellos pacientes y variantes con alguno de estos campos vacío. Existe la posibilidad de que algunos de estos campos aparezcan en la base de datos con categoría de "No disponible". Puesto que no se trata de un campo vacío en sí mismo, no se ha eliminado en este cribado inicial de los datos realizado a través de consultas SQL y, por lo tanto, se verá en las siguientes secciones cómo se han considerado en el estudio.

Se podría haber optado por la posibilidad de imputar los datos faltantes, es decir, estimar los valores ausentes en base a la restante información de la muestra, sin embargo, se ha escogido la opción más simple de descartar los casos con datos incompletos.

En las siguientes secciones se describirán más detalladamente las características de las bases de datos utilizadas, tanto en relación con los pacientes como con las variantes genéticas.

5.1.1. Datos relativos a los estudios de los pacientes

La base de datos de los pacientes analizados en este estudio, tras descartar los individuos con algún campo vacío, consta de 1284 individuos divididos en un grupo de casos y un grupo control. Para este conjunto de pacientes se han analizado variantes que aparecen en 170 genes con posible asociación a cardiopatías familiares.

Las variables disponibles en esta base de datos son las siguientes:

- `STD_RSLT_NUM_ID`: Identificador del resultado del estudio de la variante en el paciente.
- `PAT_NUM_ID`: Identificador del paciente.
- `STD_NUM_ID`: Identificador del estudio de secuenciación en el que se ha evaluado la variante.
- `MUT_NUM_ID`: Identificador de la variante estudiada. Este identificador se basa en un sistema de nomenclatura propio de la empresa Health in Code S.L.
- `ALLELIC`: Variable indicadora de presencia de la variante en los alelos. Puede ser `Heterozygosis`, cuando la variante está presente en uno de los dos alelos; `Homozygosis`, cuando la variante está presente en los dos alelos; o `Hemizygosis`, cuando se trata de los cromosomas sexuales masculinos y por tanto la variante estará presente en una sola copia.
- `SEX`: Variable indicadora del sexo del paciente. Será `Male`, si el paciente es un hombre y `Female`, si la paciente es una mujer.
- `RACE_MOTHER`: Variable que indica la etnia de la progenitora del paciente. Las posibles etnias son las siguientes: `African American`, `Black African`, `Oriental`, `Spanish American`, `Western Caucasian` y `Unavailable`.
- `RACE_FATHER`: Etnia del progenitor del paciente. Las posibles etnias son las mismas que en el caso de `RACE_MOTHER`.
- `AGE`: Variable cuantitativa indicadora de la edad del paciente.

Tal y como ya se ha indicado, estos conjuntos de datos se subdividen en dos grupos: el grupo de casos y el grupo control.

La muestra de casos consta de 973 pacientes afectados de miocardiopatía hipertrófica. Por su parte, la muestra control consta de 311 pacientes también afectados de cardiopatías hereditarias, pero no de MCH. Más concretamente, a estos pacientes se les ha diagnosticado alguna (o varias) de las siguientes enfermedades: miocardiopatía dilatada, miocardiopatía restrictiva, miocardiopatía no compactada, miocardiopatía arritmogénica, enfermedad de Fabry y síndrome de Noonan.

Las Tablas 5.1 y 5.2 muestran un resumen de las características de estas variables.

		Casos	**Controles**
Edad	Mínimo	1	1
	Mediana	59	39
	Media	54.93	39.10
	Máximo	86	90

Tabla 5.1: Resumen de la variable continua *Edad* (`AGE`) relativa a los pacientes de casos y controles.

		Casos	Controles
Composición alélica	Heterocigosis	585	193
	Homocigosis	376	114
	Hemicigosis	12	4
Sexo	Mujer	264	143
	Hombre	689	168
Etnia de la madre	Afroamericana	10	0
	Africana	0	0
	Oriental	0	1
	Hispana	0	3
	Caucásica	648	259
	No disponible	315	48
Etnia del padre	Afroamericana	10	0
	Africana	0	0
	Oriental	0	1
	Hispana	0	3
	Caucásica	648	259
	No disponible	315	48

Tabla 5.2: Resumen de las variables cualitativas *Composición alélica* (`ALLELIC`), *Sexo* (`SEX`), *Etnia de la madre* (`RACE_MOTHER`) y *Etnia del padre* (`RACE_FATHER`) relativas a los pacientes de casos y controles.

5.1.2. Datos relativos a las variantes genéticas

La base de datos de las variantes estudiadas consta de 13407 variantes para las cuales se conocen las siguientes características:

- `MUT_NUM_ID`: Identificador de la variante.
- `GEN_SYMBOL`: Gen en el que se encuentra la mutación.
- `PROTEIN`: Nombre de la proteína codificada.
- `CHROMOSOMIC`: Nombre cromosómico de la variante. En él se proporciona información sobre la variante como lo es el cromosoma en que se encuentra, la posición concreta o el cambio de base. Por ejemplo: `NC_000006.11 : g.44281015T > C` indicaría que la variante se encuentra en el cromosoma 6, en la posición 44281015, y que el cambio se realiza de una base de timina a una de citosina.
- `PATHOGENICITY`: Variable que indica si la variante es patogénica (basado en criterios clínicos). Esta variable toma los valores `Unknown`, `Non pathogenic`, `Unlikely to be pathogenic`, `Likely to be pathogenic`, `Very likely to be pathogenic` o `Pathogenic`.
- `ExAC.Total.Allele.freq.`: Variable numérica indicando la frecuencia del alelo menos común (MAF) en la población. Esta variable no se ha calculado a partir de esta base de datos, si no que ha sido extraída de la base ExAC (Consorcio de Agregación de Exomas), una coalición de investigadores dedicada a reunir los datos de secuenciación de exomas obtenidos en diferentes proyectos internacionales.

5.2. Filtrado previo de los datos

Tal y como se ha descrito en la metodología, el primer paso a la hora de realizar el estudio caso-control es analizar y acotar la base de datos para evitar la detección de falsas asociaciones. Con ese objetivo, en esta sección se aplicarán los métodos descritos en el Capítulo 3.

5.2.1. Frecuencia del alelo menor

Como se ha descrito, la base de datos consta de 13407 variantes diferentes, pero no se sabe si estas se encuentran en la población con una frecuencia alélica de al menos el 1 %, en otro caso no se podrían considerar SNPs, y se deberían eliminar de la muestra.

Se han utilizado las frecuencias disponibles en la base de datos ExAC, que contiene a la gran mayoría de las variantes de este estudio. Cuando una variante no está presente en ExAC es porque en su población no ha sido encontrada, por lo que se asume que su frecuencia es inferior al 1 % y, por tanto, se ha eliminado también de esta base de datos. El número de variantes disponibles clasificadas por su frecuencia puede verse en la Tabla 5.3.

Frecuencia alélica	Menor del 1 %	Mayor o igual al 1 %
Número de variantes	11328 (84.5 %)	2079 (15.5 %)

Tabla 5.3: Distribución de las variantes del estudio según su frecuencia alélica sea mayor o menor que 1 %. El 84.5 % de las variantes tiene una frecuencia baja y, por lo tanto, no serán utilizadas en el estudio.

Como se puede observar, casi un 85 % de las variantes son variantes raras que no se consideran SNPs y que, por tanto, no se utilizarán en este estudio caso-control. El número de SNPs que se considerarán es 2079.

5.2.2. Estratificación de la población

Como se ha explicado en el Capítulo 3, la presencia de subpoblaciones puede dar lugar a asociaciones falsas si las frecuencias alélicas en los grupos de casos y controles difieren.

Para valorar si la población está estratificada, es decir, si existen subpoblaciones, puesto que la etnia de los progenitores de cada paciente es conocida, se han utilizado esos datos para verificar si pudiera haber presencia de distintas etnias.

Como se puede ver en la Tabla 5.2, la etnia del padre y de la madre del paciente coinciden, por lo que es indiferente considerar una u otra. La Tabla 5.4 muestra, de nuevo, la distribución del número de pacientes según la etnia de los progenitores tanto en la población control como en la de casos. La mayoría de individuos, tanto en el grupo de casos como en el de controles, pertenecen a la etnia caucásica. En el grupo de casos existen 315 individuos cuya etnia aparece como “No disponible”, así como 48 en los controles. Aunque este dato no proporciona información sobre la etnia de esos pacientes, se ha considerado que son también caucásicos ya que, según la información proporcionada por la empresa Health in Code S.L., los datos de esos pacientes fueron tomados en población europea no finlandesa. Así, se han eliminado de la base de datos los individuos cuyas etnias son afroamericana, africana, oriental e hispana. En total, el número de pacientes eliminados es de 14, de los cuales 10 estaban en la población de casos y 4 en la de control.

Se podría haber utilizado alguna de las técnicas de control de la estratificación descritas en la Sección 3.3 pero, al disponer de la información étnica de los individuos, se ha optado por esta opción que, además, permite la aplicación de las técnicas de regresión descritas.

Etnia de los progenitores	**Casos**	**Controles**
Afroamericana	10	0
Africana	0	0
Oriental	0	1
Hispana	0	3
Caucásica	648	259
No disponible	315	48
	973	311

Tabla 5.4: Distribución de los pacientes según la etnia de los progenitores.

5.2.3. Equilibrio de Hardy-Weinberg

El hecho de que las frecuencias de las variantes no estén en equilibrio de Hardy-Weinberg puede ser indicativo de errores de genotipado, de estratificación de la población o de asociaciones entre SNP y enfermedad, entre otros. Por este último motivo se realizará el análisis del equilibrio de Hardy-Weinberg solamente en la población control.

Se han realizado test estadísticos para estudiar el equilibrio de Hardy-Weinberg. Se ha realizado un test χ^2 de Pearson sobre las frecuencias alélicas para cada SNP como el descrito en la ecuación (3.1) cuando las frecuencias son mayores que 5 y, análogamente, un test exacto cuando alguna de las frecuencias es más baja. Se considerarán SNPs en desequilibrio de Hardy-Weinberg los que tengan un p-valor menor que 0.05. Así, realizando estos contrastes y corrigiendo los correspondientes p-valores con una corrección del FDR utilizando el ajuste de Benjamini-Hochberg para controlar la presencia de falsos positivos se tiene que, de los 2079 SNPs iniciales, 214 (10.3 %) no están equilibrio de Hardy-Weinberg.

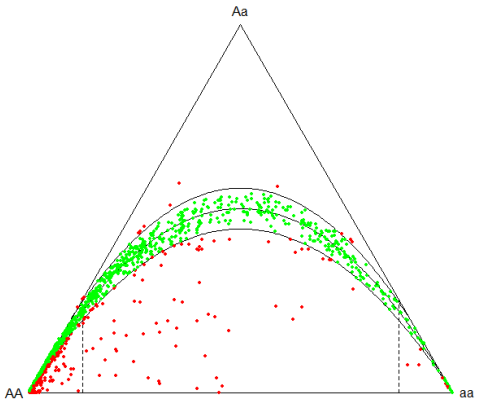

Figura 5.1: Gráfico ternario que representa los p-valores para el contraste del equilibrio de Hardy-Weinberg. En verde los p-valores que se encuentran dentro de la región de confianza del equilibrio de Hardy-Weinberg.

Los SNPs son generalmente bialélicos, es decir, en ellos se observan dos alelos. Esto hace que existan tres composiciones alélicas (AA, Aa y aa) que pueden representarse en un diagrama ternario, al tratarse de datos composiciones (ver Mateu-Figueras et al., 2003). Cuando los SNPs se encuentran en equilibrio

de Hardy-Weinberg, sus frecuencias genotípicas sumarán uno, y entonces estos datos de composición describirán una parábola en tal diagrama. En este sentido, la Figura 5.1 muestra un gráfico ternario en el que se representan en verde los SNPs que se encuentran en equilibrio (dentro de la región de confianza al 95 %), y en rojo los que no, tras aplicar sobre cada uno un contraste de significación.

5.3. Estudio de asociación

En las secciones anteriores se han eliminado de la muestra todos los individuos cuya etnia no era caucásica, utilizando solamente una muestra de pacientes de esa etnia, o bien con etnia "No disponible". Por otro lado, se han descartado también de la muestra aquellos SNPs cuya frecuencia alélica y genotípica no permanece constante a través de los individuos por no encontrarse en equilibrio de Hardy-Weinberg. Además, para poder considerar solamente SNPs en este estudio, se han suprimido las variantes cuya frecuencia en la población es menor del 1 %. Con esto se trata de evitar que estas características genéticas de los individuos lleven a falsas asociaciones entre los SNPs y la MCH. Por tanto, el número de variantes con el que se buscarán las asociaciones es de 1865 y el número de pacientes 1284.

En esta sección se aplicará la metodología explicada en el Capítulo 4 en relación a los estudios tipo caso-control para esos datos. Un primer acercamiento a la búsqueda de asociaciones entre las variantes y la enfermedad es ver si hay algún SNPs que por sí solo esté asociado a la MCH. Como se ha explicado en la primera parte del trabajo, se puede llevar esto a cabo realizando regresiones logísticas sobre cada uno de los SNPs, sin embargo, existen varias limitaciones para esta metodología, por lo que la regresión logística penalizada será más adecuada.

5.3.1. Regresión logística clásica

En este apartado se realizará una regresión logística para cada SNP considerando también otras covariables adicionales y como variable respuesta el hecho de que el individuo esté (o no) afectado de MCH. Se han ajustado estos modelos considerando el patrón genético autosómico dominante y el patrón de herencia aditivo, ambos explicados en el Capítulo 2.

El modelo de regresión logística que se ajustará sobre cada variante consta de las variables explicativas *Composición alélica*, *Sexo* y *Edad* y de la variable respuesta *HCM*, obtenidas a partir de los datos de `ALLELIC`, `SEX`, `AGE` y `HCM`, respectivamente. Se explica a continuación qué representan estas variables:

- *Composición alélica*: Variable indicadora de presencia de la variante en los alelos. La composición alélica puede ser, en general, heterocigosis, homocigosis o hemicigosis, como se ha descrito en la Sección 5.1.1. Sin embargo, cuando se considera patrón de herencia autosómico dominante, heterocigosis y homocigosis serán equivalentes. Según el modelo de herencia, esta variable se codifica del siguiente modo:

  - Considerando patrón de herencia aditivo: Puede tomar los valores 0, 1 y 2. Así, 0 indicaría que esa variante no está presente en este individuo, 1 indica que la variante se encuentra en uno de los alelos y 2 indica que la variante se encuentra en los dos alelos.
  - Considerando patrón de herencia autosómico dominante: Puede tomar los valores 0 ó 1. Aquí, 0 indicaría que esa variante no está presente en este individuo, 1 indica que la variante se encuentra en el individuo (sin especificar si se encuentra en uno o en los dos alelos).

  En los casos en que la variante se encuentre en hemicigosis se ha considerado, por simplicidad, que su comportamiento es equivalente al de una variante en homocigosis.

- *Sexo*: Es la variable indicadora del sexo del paciente. Puede tomar los valores 0 ó 1 según el individuo sea mujer u hombre, respectivamente.

- *Edad*: Es la edad del paciente. Es una variable cuantitativa que toma valores enteros comprendidos entre 1 y 90.
- *HCM*: Es una variable respuesta de tipo binario que toma el valor 0 si el paciente procede de la población control, es decir, si no está afectado de MCH; y 1 si el paciente procede de la población de casos y es por tanto portador de la enfermedad.

Regresión logística en R

Se ha utilizado R para llevar a cabo estas regresiones. La función `glm` permite ajustar esos modelos y proporciona, entre otros datos, el valor de los coeficientes estimados, $\hat{\boldsymbol{\beta}}$, de la regresión en cada caso. En este estudio se ha considerado un modelo por cada SNP del siguiente modo:

$$logit(\pi_i) = \beta_0 + \beta_1 I(Composición\,alélica_i) + \beta_2 I(Sexo_i) + \beta_3 Edad_i. \tag{5.1}$$

Si bien cuando se trata de realizar un test de razón de verosimilitudes para elegir un modelo el ajuste del coeficiente constante no es necesario, tanto en este ejemplo como en el resto de los modelos ajustados, se ha incluido.

Se debe indicar en la función `glm` la función link que se usará en el modelo que, en este caso, por tratarse de una regresión logística, será `family="binomial"`. A continuación se muestra, a modo de ejemplo, el resultado obtenido en R tras ajustar un modelo de regresión logística con uno de los SNPs estudiados.

```
> fit<-glm(dat$HCM~dat$ALLELIC+dat$SEX+dat$AGE,family="binomial",data=dat)
> fit

Call:  glm(formula = dat$HCM ~ dat$ALLELIC + dat$SEX + dat$AGE,
family = "binomial", data = dat)

Coefficients:
(Intercept)  dat$ALLELIC1  dat$ALLELIC2   dat$SEX1      dat$AGE
-1.15154      -1.20166      -1.25870       0.77049       0.03922

Degrees of Freedom: 1269 Total (i.e. Null);  1264 Residual
Null Deviance:     1759
Residual Deviance: 1258  AIC: 1268
```

Este ajuste proporciona, entre otros datos, las estimaciones de los coeficientes de la regresión, β_0, β_1, β_2 y β_3, para la constante, la composición alélica de la variante, el sexo y la edad, respectivamente.

En este ejemplo, el OR que compara la presencia de la variante en heterocigosis con la ausencia de la variante, ajustando también el efecto de la edad y el sexo, sería $\exp(-1.20166) = 0.30069$. Por tanto, ese SNP en heterocigosis sería un factor protector para la enfermedad.

Análogamente, puede obtenerse un ajuste para un modelo más simple, en el que sólo se consideran las variables *Sexo* y *Edad*:

```
> fit1<-glm(dat$HCM~dat$SEX+dat$AGE,family="binomial",data=dat)
> fit1

Call:  glm(formula = dat$HCM ~ dat$SEX + dat$AGE, family = "binomial",
data = dat)

Coefficients:
(Intercept)  dat$SEX1    dat$AGE
```

```
-1.19723     0.76929     0.03924

Degrees of Freedom: 1268 Total (i.e. Null);  1266 Residual
Null Deviance:     1407
Residual Deviance: 1259  AIC: 1265
```

La función `anova` permite realizar un test de razón de verosimilitudes como el descrito en la Sección 4.1 para contrastar si el primer modelo ajustado es significativamente distinto del modelo básico (que no incluye la covariable *Composición alélica*) o no.

A continuación se muestra el resultado obtenido en R a través de la función `anova`:

```
> anova(fit, fit1, test="LRT")
Analysis of Deviance Table

Model 1: dat$HCM ~ dat$ALLELIC + dat$SEX + dat$AGE
Model 2: dat$HCM ~ dat$SEX + dat$AGE
   Resid. Df   Resid. Dev    Df     Deviance    Pr(>Chi)
1   1264        1257.7
2   1266        1258.5       -2     -0.82428      0.6622
```

Así, se obtiene un p-valor para el contraste entre ambos modelos. En el caso considerado en este ejemplo se tiene un p-valor de 0.6622 que, para un nivel de significación $\alpha = 0.05$, indicaría que los modelos ajustados no son significativamente diferentes.

Se ha trabajado con los dos patrones de herencia descritos anteriormente: patrón de herencia aditivo y patrón de herencia autosómico dominante.

Patrón de herencia aditivo

Se ha restringido previamente el conjunto de datos a aquellas variantes en que no ocurra fenómeno de separación completa o cuasicompleta. Para ello, se han considerado tablas de contingencia en función del número de casos y controles y de la composición alélica de la variante. En el caso en que algún factor de la tabla sea 0, se considera que existe algún tipo de separación. En este caso, tras eliminar esas variantes, la base de datos consta de 999 SNPs.

Se consideran, para cada variante, dos modelos: un modelo que incorpora las covariables *Composición alélica*, *Sexo* y *Edad* y un modelo más simple integrado solamente por las covariables *Sexo* y *Edad*. Puesto que se considera un patrón de herencia aditivo, aquí la variable *Composición alélica* puede tomar tres valores distintos. En el análisis se utiliza, por cada variante, un test de razón de verosimilitudes para contrastar si los dos modelos considerados son significativamente distintos. Se rechazará la hipótesis de que los modelos son iguales para un nivel de significación $\alpha = 0.05$ si el p-valor del contraste es menor que α.

La Figura 5.2 representa el -logaritmo de los p-valores esperados en una distribución χ^2 frente al -logaritmo de los p-valores observados en los contrastes mediante un Q-Q plot. Puntos alejados de la recta estarían indicando diferencias significativas en los modelos considerados en ese contraste. Se puede observar que solamente un punto se aleja considerablemente de la recta, mientras que los restantes están sobre ella o muy cerca.

Los resultados de los tests considerados muestran que, de 999 contrastes, 62 tendrían, en principio, una asociación significativa a un nivel $\alpha = 0.05$.

La Tabla 5.5 muestra los resultados de las 30 primeras regresiones ordenados de menor a mayor según el p-valor obtenido en el contraste, para las regresiones en que este valor resultó menor que 0.05 (los restantes resultados pueden verse en el Apéndice A). No se incluyen las estimaciones de los coeficientes de la regresión pero sí los *odds ratio* y sus intervalos de confianza respectivos. OR_1 es el odds ratio de la presencia de la variante en heterocigosis con respecto a la no presencia del SNP, y

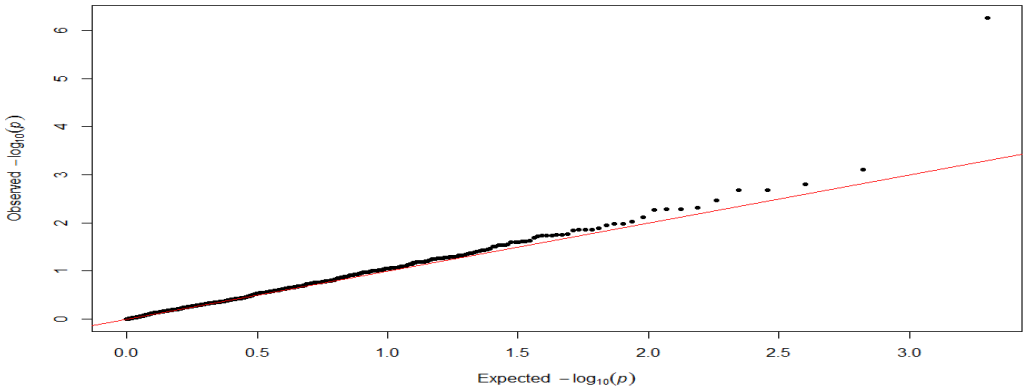

Figura 5.2: Q-Q plot del -logaritmo de los p-valores esperados frente al -logaritmo de los p-valores observados en los contrastes con patrón de herencia aditivo.

OR_2 es el *odds ratio* entre la presencia de la variante en homocigosis con respecto a la no presencia del SNP.

La multiplicidad de contrastes implica que muchos de esos resultados no serán realmente estadísticamente significativos. Para ajustar los p-valores teniendo en cuenta la multiplicidad de contrastes se ha utilizado la corrección basada en FDR empleando el procedimiento de Benjamini-Hochberg. En la última columna se muestra el p-valor ajustado.

Tabla 5.5: SNPs que muestran asociación en regresión logística clásica considerando patrón de herencia aditivo. Por columnas, de izquierda a derecha: identificador de la variante, gen en que se encuentra la variante, OR de la presencia de la variante en heterocigosis con respecto a la no presencia de la variante, intervalo de confianza al 95 % para el OR anterior, OR de la presencia de la variante en homocigosis con respecto a la no presencia de la variante, intervalo de confianza al 95 % para el OR anterior, p-valor del test y p-valor ajustado.

SNP ID	GEN	OR_1	IC para OR_1	OR_2	IC para OR_2	p-val.	p-val. ajust.
19318	BAG3	1.55	(1.15, 2.08)	7.33	(2.57, 20.92)	1.455e-06	0.0015
419280	SCN10A	1.43	(0.98, 2.08)	2.32	(1.54, 3.49)	0.0001	0.0626
340420	SCN10A	1.52	(1.04, 2.21)	2.29	(1.53, 3.44)	0.0002	0.0753
23551	RBM20	0.48	(0.32, 0.72)	0.63	(0.41, 0.97)	0.0008	0.2081
1541	SCN5A	1.43	(1.07, 1.92)	2.50	(1.24, 5.04)	0.0034	0.4424
44321	TTN	2.22	(1.28, 3.84)	0.14	(0.01, 2.72)	0.0035	0.4424
24404	TBX20	1.47	(1.10, 1.97)	1.92	(1.21, 3.05)	0.0039	0.4424
24134	TBX20	1.47	(1.10, 1.97)	1.92	(1.21, 3.05)	0.0039	0.4424
429390	GATA6	1.85	(1.01, 3.39)	8.99	(1.10, 73.10)	0.0040	0.4424
160857	FHOD3	7.33	(1.62, 33.25)	0.85	(0.07, 10.07)	0.0054	0.4909
6107	LDLR	1.10	(0.81, 1.49)	1.90	(1.25, 2.90)	0.0058	0.4909

(Continúa en la página siguiente)

SNP ID	GEN	OR_1	IC para OR_1	OR_2	IC para OR_2	p-val.	p-val. ajust.
6106	LDLR	1.02	(0.75, 1.39)	1.85	(1.21, 2.82)	0.0059	0.4909
11373	SCN1B	0.65	(0.46, 0.91)	2.04	(0.83, 5.03)	0.0073	0.5203
18966	GATA4	1.04	(0.72, 1.50)	13.24	(1.59, 110.55)	0.0075	0.5203
22541	RBM20	0.96	(0.72, 1.29)	3.11	(1.36, 7.11)	0.0078	0.5203
23688	PSEN2	0.48	(0.23, 0.99)	0.71	(0.35, 1.46)	0.0097	0.5869
13146	AGL	0.84	(0.61, 1.16)	1.45	(0.98, 2.16)	0.0100	0.5869
310891	SGCA	1.91	(1.17, 3.12)	0.27	(0.04 , 1.98)	0.0107	0.5869
9821	LDLR	0.87	(0.64, 1.20)	1.55	(1.02, 2.35)	0.0125	0.5869
28515	OBSL1	1.59	(1.17, 2.17)	1.40	(0.96, 2.06)	0.0126	0.5869
268014	CTNNA3	1.43	(0.94, 2.19)	0.18	(0.04, 0.79)	0.0142	0.5869
5877	SCN5A	1.48	(1.10, 2.00)	1.92	(0.84, 4.36)	0.0148	0.5869
23800	AGL	0.86	(0.61, 1.23)	1.40	(0.94, 2.09)	0.0151	0.5869
2686	LMNA	0.96	(0.68, 1.37)	5.98	(1.33, 26.94)	0.0151	0.5869
5698	RYR2	0.74	(0.50, 1.07)	1.15	(0.76, 1.73)	0.0152	0.5869
21826	PSEN2	0.48	(0.23, 1.01)	0.70	(0.34, 1.44)	0.0153	0.5869
5875	SCN5A	1.46	(1.08, 1.99)	2.07	(0.82, 5.23)	0.0179	0.6635
159757	FHOD3	1.47	(1.09, 1.97)	1.54	(1.00, 2.39)	0.0197	0.7022
23087	PSEN2	0.50	(0.26, 0.96)	0.70	(0.37, 1.32)	0.0216	0.7170
5892	LDLR	0.93	(0.68, 1.27)	1.57	(1.04, 2.38)	0.0217	0.7170

Exceptuando el SNP con ID 19318, los p-valores son, aunque menores a 0.05, bastante altos. En efecto, si se observan los p-valores corregidos se tiene que solamente ese SNP tiene un p-valor ajustado menor que 0.05 y, por tanto, solamente ese tendría una asociación significativa con la enfermedad.

Además, solamente con el dato relativo al p-valor del contraste LRT se desconoce si esa asociación implica que la presencia de la variante aumenta el riesgo de padecer la enfermedad o bien lo disminuye. Los coeficientes de la regresión proporcionarían esa información pero, en estudios epidemiológicos, es habitual utilizar el OR para ver de qué tipo de asociación se trata. ORs superiores a 1 indicarían riesgo de enfermedad, mientras que los que se encuentran entre 0 y 1 indicarían que se trata de un factor protector. En el caso del SNP con ID 19318 se tiene que, en cualquiera de las dos composiciones alélicas posibles, actúa como factor de riesgo. Sin embargo, también se observa que el intervalo de confianza relativo al OR_2 es bastante grande, indicando así que las estimaciones no son del todo precisas.

Patrón de herencia autosómico dominante

En este caso también se llevaron a cabo regresiones logísticas pero considerando un patrón de herencia autosómico dominante, por lo que la covariable *Composición alélica* puede tomar solamente dos valores. Ahora el número de regresiones a realizar es de 1723, tras eliminar los SNPs que están afectados de algún tipo de separación.

La Figura 5.3 representa el -logaritmo de los p-valores esperados frente al -logaritmo de los p-valores observados mediante un Q-Q plot. En este caso, parece que hay menos puntos sobre la recta, y los que

no están sobre la recta se alejan moderadamente de ésta.

La Tabla 5.6 muestra los 30 resultados con p-valores más bajos obtenidos en este caso. El resto de resultados pueden verse en el Apéndice A. Como en el modelo anterior, también se incluye el OR junto con el intervalo de confianza al 95 %, el p-valor y el p-valor ajustado.

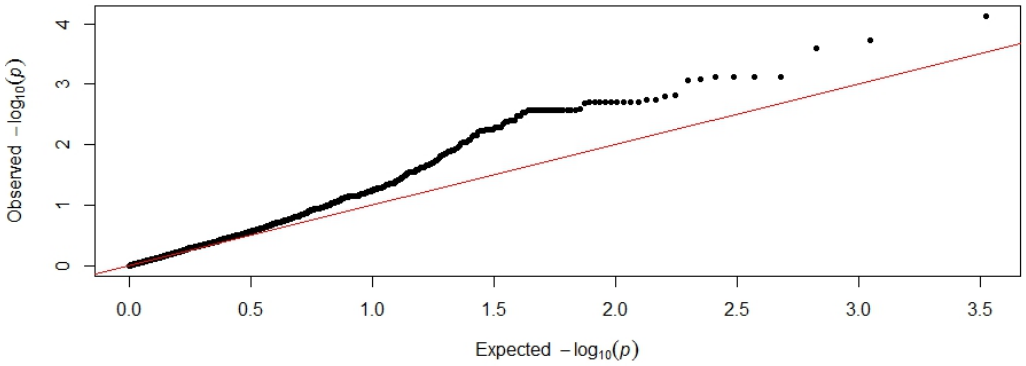

Figura 5.3: Q-Q plot de los p-valores esperados frente a los p-valores observados en los contrastes con patrón de herencia autosómico dominante.

Tabla 5.6: SNPs con asociación significativa en regresión logística clásica considerando patrón de herencia autosómico dominante. Por columnas, de izquierda a derecha: identificador de la variante, gen en que se encuentra la variante, OR de la presencia de la variante con respecto a la no presencia de la variante, intervalo de confianza al 95 % para el OR, p-valor del test y p-valor ajustado.

SNP ID	GEN	OR	IC para OR	p-val.	p-val. ajust.
19318	BAG3	1.78	(1.33, 2.37)	0.0001	0.1204
51429	TTN	6.03	(1.82, 19.97)	0.0002	0.1204
51984	TTN	2.67	(1.50, 4.72)	0.0002	0.1204
51676	TTN	4.49	(1.57, 12.86)	0.0008	0.1204
51697	TTN	4.49	(1.57, 12.86)	0.0008	0.1204
51853	TTN	4.49	(1.57, 12.86)	0.0008	0.1204
51964	TTN	4.49	(1.57, 12.86)	0.0008	0.1204
23551	RBM20	0.53	(0.36, 0.78)	0.0008	0.1204
269056	TTN	4.44	(1.55, 12.75)	0.0008	0.1204
446918	TXNRD2	5.66	(1.59, 20.11)	0.0015	0.1204
340420	SCN10A	1.80	(1.26, 2.57)	0.0016	0.1204
24134	TBX20	1.56	(1.18, 2.06)	0.0018	0.1204
24404	TBX20	1.56	(1.18, 2.06)	0.0018	0.1204

(Continúa en la página siguiente)

SNP ID	GEN	OR	IC para OR	p-val.	p-val. ajust.
51164	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
51473	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
51475	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
51720	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
51876	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
51938	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
144713	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
144756	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
51303	TTN	3.67	(1.41, 9.54)	0.0020	0.1204
3831	SCN5A	4.47	(1.49, 13.42)	0.0020	0.1204
51664	TTN	3.55	(1.37, 9.22)	0.0025	0.1204
51148	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51213	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51521	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51626	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51642	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51661	TTN	3.53	(1.36, 9.16)	0.0027	0.1204

En este caso se han encontrado 152 SNPs cuyo p-valor del contraste LRT es menor de 0.05. Tras realizar un ajuste de los p-valores para corregir la multiplicidad de contrastes se obtiene que, para un $\alpha = 0.05$, ningún SNP tiene asociación significativa con la variable respuesta.

Como se puede observar en los resultados de la Tabla 5.6 (y, en menor medida, en la Tabla 5.5), hay varias variantes (del mismo gen) en los que los ORs coinciden y, en esos casos, también los correspondientes p-valores son iguales. Por ejemplo, en la Tabla 5.5, esto ocurre con los SNPs con ID 24404 y 24134; y en la Tabla 5.6, con los SNPs con ID 51676, 51697, 51853 y 51964, entre otros. Esto parece ser indicativo de que esos SNPs están altamente correlacionados o se encuentran en una región de alto desequilibrio de ligamiento, por lo que se heredan juntos. El hecho de heredarse conjuntamente lleva a que no se pueda identificar cuál es el verdadero SNP causal y se seleccionen más variantes que las que realmente podrían estar asociadas. En este sentido, como se ha explicado en la metodología, existen técnicas de regresión múltiple que tendrían en cuenta la posible correlación entre las variantes, como se verá a continuación.

5.3.2. Regresión logística penalizada: *lasso*

Los modelos de regresión logística clásica resultan útiles para realizar una primera aproximación a las posibles asociaciones, sin embargo, cuando se trata de muchos ajustes simultáneos, este tipo de análisis lleva a la aparición de numerosas falsas asociaciones, si no se corrige el p-valor del contraste. Además, podría ser interesante combinar la información genética entre loci y, en ese aspecto, este tipo de modelos ignoran por completo la posibilidad de múltiples pequeños efectos que contribuyan a la respuesta, algo que sí se consigue con los modelos de regresión multivariante en la que se incluyan todos los SNPs, como ya se ha explicado. Además, añadir una penalización en estos modelos permite

la estimación de coeficientes cuando el número de covariables es mayor que el número de observaciones, algo habitual en genómica.

Se ha realizado una regresión logística penalizada considerando un modelo de tipo *lasso*, como el descrito en la Sección 4.2.1 incluyendo como variables explicativas todas las variantes. En este tipo de modelos se hace selección de variables además de la estimación de los coeficientes de regresión, por lo que su interpretación es mucho más sencilla que en el caso de los modelos logísticos ordinarios.

Al igual que en regresión logística clásica puede obtenerse el OR relativo al efecto de cada variante incluida en el modelo a través de $\exp(\beta_j)$, donde β_j es, como en los casos anteriores, el coeficiente de la regresión correspondiente a la j-ésima variante. Sin embargo, en este tipo de modelos no es posible obtener intervalos de confianza para estos ORs ni p-valores que aporten información sobre la significación estadística de las variables incluidas.

Para escoger el valor del parámetro de penalización λ se han considerado técnicas de validación cruzada considerando el λ que minimiza el error cuadrático medio.

Se ha utilizado el paquete `glmnet` de R para llevar a cabo la validación cruzada y la regresión *lasso*.

El lasso en R

El paquete `glmnet` es uno de los más eficientes y utilizados en R para llevar a cabo modelos de regresión logística penalizada, tanto para *lasso* como para *elastic net*.

Dentro de esta librería, la función `cv.glmnet` lleva a cabo la validación cruzada necesaria para buscar el valor óptimo de λ según diversos criterios, como por ejemplo el ECM. Esta función utiliza, entre otros, los siguientes argumentos: `x`, la matriz de covariables en formato `model.matrix`, es decir, como una matriz de modelo que transforma los factores en variables dummy; `y`, el vector de la variable respuesta; `family`, el tipo de regresión, que en este caso será `family=binomial`, por tratarse de una regresión logística; el criterio utilizado en la validación cruzada, que en este caso será el ECM, `type.measure=mse`; y el parámetro `alpha`, que puede tomar distintos valores entre 0 y 1. Será `alpha=1` cuando se ajusta un modelo *lasso*.

A continuación se muestra, a modo de ejemplo, una salida de R, una vez transformada la base de datos hasta conseguir el formato requerido por esta función:

```
> lasso<-cv.glmnet(x, y, family="binomial", alpha=1, type.measure="mse")
> lasso$glmnet.fit

Call:  glmnet(x = x, y = y, family = "binomial", alpha = 1)

      Df       %Dev    Lambda
[1,]   0 -1.883e-14 0.0453700
[2,]   1  9.079e-04 0.0433100
[3,]   3  2.849e-03 0.0413400
[4,]   5  6.050e-03 0.0394600
[5,]   6  9.301e-03 0.0376700
[6,]   7  1.299e-02 0.0359600
[7,]  10  1.735e-02 0.0343200
[8,]  11  2.236e-02 0.0327600
[9,]  16  2.799e-02 0.0312700
[10,]  25  3.552e-02 0.0298500
[11,]  31  4.519e-02 0.0284900
[12,]  49  5.642e-02 0.0272000
[13,]  63  7.079e-02 0.0259600
[14,]  71  8.594e-02 0.0247800
[15,]  87  1.018e-01 0.0236600
[16,]  98  1.190e-01 0.0225800
```

```
[17,] 113  1.371e-01 0.0215600
[18,] 128  1.560e-01 0.0205800
[19,] 145  1.752e-01 0.0196400
[20,] 161  1.950e-01 0.0187500
```

En la columna de la izquierda (`Df`) se muestran los grados de libertad del modelo, que son los coeficientes distintos de cero. La columna de la derecha proporciona el valor de λ (`Lambda`) en cada caso. Por defecto, la función muestra 100 valores de λ, aunque en este ejemplo sólo se han incluido los 20 primeros.

Aunque existen otras funciones dentro de esta librería para obtener los coeficientes de la regresión según el valor de λ seleccionado, también es posible extraerlos a partir de esta función. Así, siguiendo con el ejemplo anterior, si se considera el λ que minimiza el error cuadrático medio en la validación cruzada, la estimación de los coeficientes de regresión según la posición que ocupan en el vector de coeficientes es la que se muestra a continuación. Aunque la extracción de todos los coeficientes es sencilla, aquí, al tratarse de un vector bastante grande, se incluyen solamente los coeficientes distintos de cero.

```
> pos_estimadores<-which(as.vector(coef(lasso, s="lambda.min"))[-1] !=0)
 [1]   21   53  288  340  762  967  1092 2112 2228 2829
> estimadores<-(coef(lasso, s="lambda.min")[-1])[pos_estimadores]
> estimadores
 [1] -0.11160006 -0.16202586 -0.76977808  0.02286335  0.31836121  0.04102083
 [7] -0.03887722  0.26498669 -0.23943519  0.06339404
```

Se han estudiado los mismos patrones de herencia que en el apartado anterior: aditivo y autosómico dominante. A continuación de muestran los resultados obtenidos en ambos casos.

Patrón de herencia aditivo

Considerando un patrón de herencia aditivo y, puesto que este tipo de regresión realiza selección de variables, puede ocurrir (y de hecho sucede en este caso) que se seleccione alguna variante en heterocigosis y que no se escoja en homocigosis, aunque biológicamente es algo que ocurre en muy raras ocasiones. Generalmente, cuando una variante en heterocigosis es un factor de riesgo para determinada enfermedad, también lo será en homocigosis.

La Figura 5.4 muestra la trayectoria de la estimación de los coeficientes según el valor del logaritmo de λ. Cada curva se corresponde con una variante. En el eje superior se indica el número de variantes cuyo coeficiente es distinto de cero según $\log(\lambda)$ (eje inferior). En este caso se tiene que el valor que minimiza el error cuadrático medio se corresponde con un λ escogido por validación cruzada de 0.0328 ($\log(\lambda) = -3.4185$), para el cual se seleccionan 10 variantes.

La Tabla 5.7 muestra los SNPs seleccionados, el gen en que se encuentra, la estimación del coeficiente de regresión para esa covariable, el OR asociado, y el modo en que está presente (en heterocigosis, denotado por 1; o en homocigosis, denotado por 2).

Los coeficientes negativos indicarían que la variante actúa como un factor protector para la enfermedad, mientras que los coeficientes positivos indicarían que estos actúan como factores de riesgo, algo de lo que también proporciona información el OR.

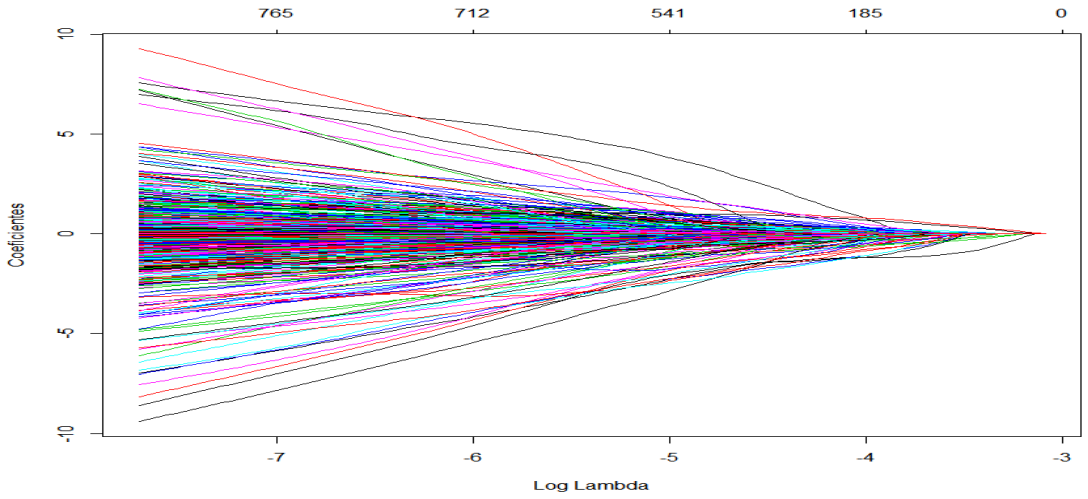

Figura 5.4: Trayectoria de las estimaciones de los coeficientes de la regresión *lasso* según el valor del logaritmo de λ considerando patrón de herencia aditivo.

SNP ID	GEN	Coeficiente	*OR*	Comp. alélica
1067	PKP2	−0.1116	0.8944	1
2052	KCNH2	−0.1620	0.8504	2
4475	CASQ2	−0.7698	0.4631	1
6106	LDLR	0.0229	1.0231	2
19318	BAG3	0.3184	1.3749	2
22541	RBM20	0.0410	1.0419	2
23688	PSEN2	−0.0389	0.9619	1
51429	TTN	0.2650	1.3034	1
53838	NOTCH1	−0.2394	0.7871	1
419280	SCN10A	0.0634	1.0654	2

Tabla 5.7: SNPs seleccionados mediante *lasso* considerando patrón de herencia aditivo. Por columnas, de izquierda a derecha: identificador de la variante, gen en que se encuentra, coeficiente estimado en la regresión para esa variante, OR de la presencia de la variante con respecto a la no presencia de la variante y composición alélica para la cual se ha seleccionado.

Patrón de herencia autosómico dominante

Se ha considerado en este caso que la herencia de la enfermedad sigue un modelo autosómico dominante.

La Figura 5.5 muestra la trayectoria de la estimación de los coeficientes según el valor del logaritmo de λ. Aquí, el valor de λ escogido por validación cruzada es de 0.0289 ($\log(\lambda) = -3.5449$), para el cual se seleccionan 25 variantes.

En la Tabla 5.8 pueden verse los SNPs seleccionados, el gen en que se encuentran, la estimación del coeficiente de regresión para esa covariable y su OR asociado.

SNP ID	GEN	Coeficiente	OR
976	TMPO	−0.5075	0.6020
1067	PKP2	−0.2891	0.7489
3350	DSP	0.0326	1.0332
4475	CASQ2	−0.2232	0.7999
12970	KCNA5	−0.0079	0.9921
14779	CACNA1C	−0.1963	0.8217
19318	BAG3	0.2467	1.2799
23444	MAP2K1	0.0690	1.0715
23551	RBM20	−0.2045	0.8150
25000	NPPA	0.0474	1.0486
39387	TTN	−0.0212	0.9790
41936	TTN	0.0061	1.0062
45675	EYA4	0.0134	1.0135
51030	CACNA1D	−0.0026	0.9974
51429	TTN	0.4043	1.4982
51837	TRDN	−0.0336	0.9669
52965	KLF10	0.0073	1.0074
53335	TTN	−0.0646	0.9374
53838	NOTCH1	−0.3380	0.7132
69629	KLF10	0.0201	1.0203
79142	ANK2	0.0591	1.0609
158975	HFE	−0.0118	0.9882
159757	FHOD3	0.0923	1.0967
340420	SCN10A	0.0195	1.0197
423960	SCN10A	0.0046	1.0046

Tabla 5.8: SNPs seleccionados mediante *lasso* considerando patrón de herencia autosómico dominante. Por columnas, de izquierda a derecha: identificador de la variante, gen en que se encuentra, coeficiente estimado en la regresión para esa variante y OR de la presencia de la variante con respecto a la no presencia de la variante.

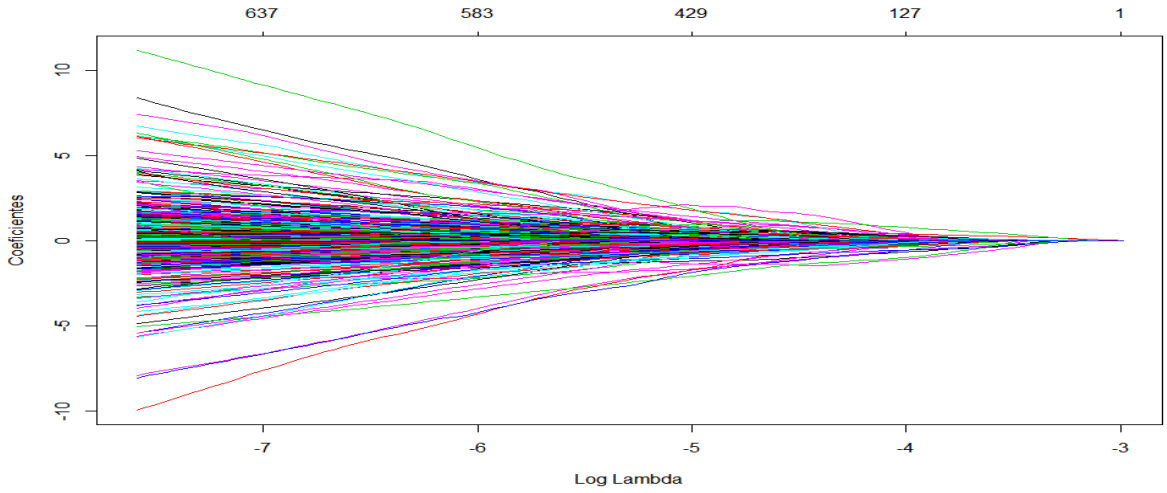

Figura 5.5: Trayectoria de las estimaciones de los coeficientes de la regresión *lasso* según el valor del logaritmo de λ considerando un patrón de herencia autosómico dominante.

5.3.3. Regresión logística penalizada: *elastic net*

En esta sección se ha realizado una regresión logística penalizada considerando una modelo de tipo *elastic net*, como el descrito en la Sección 4.2.2. Al igual que en la regresión *lasso*, aquí también se hace selección de variables además de la estimación de los coeficientes de regresión. Si bien este modelo es muy similar al *lasso*, una de las ventajas con respecto a él es que, en el caso de covariables altamente correlacionadas, se seleccionan todas ellas, mientras que *lasso* selecciona una de ellas e ignora el resto.

El parámetro de penalización λ se ha seleccionando como en los casos anteriores considerando el λ que minimiza el error cuadrático medio y el modelo ha sido ajustado también mediante la librería de **R** `glmnet` y la función `cv.glmnet`, pero considerando en este caso `alpha=0.5`.

Se muestran a continuación los resultados obtenidos en ambos casos considerando los mismos patrones de herencia que se han descrito en las secciones anteriores.

Patrón de herencia aditivo

En la Figura 5.6 se puede ver la trayectoria de la estimación de los coeficientes según el valor del logaritmo de λ. En este caso se tiene que el valor que minimiza el error cuadrático mínimo se corresponde con un λ de 0.0686 ($\log(\lambda) = -2.6788$), para el cual se seleccionan 10 variantes.

La Tabla 5.9 muestra los SNPs seleccionados, el gen en que se encuentra, la estimación del coeficiente de regresión para esa covariable, el OR y la composición alélica en que está presente (en heterocigosis, denotado por 1; o en homocigosis, denotado por 2).

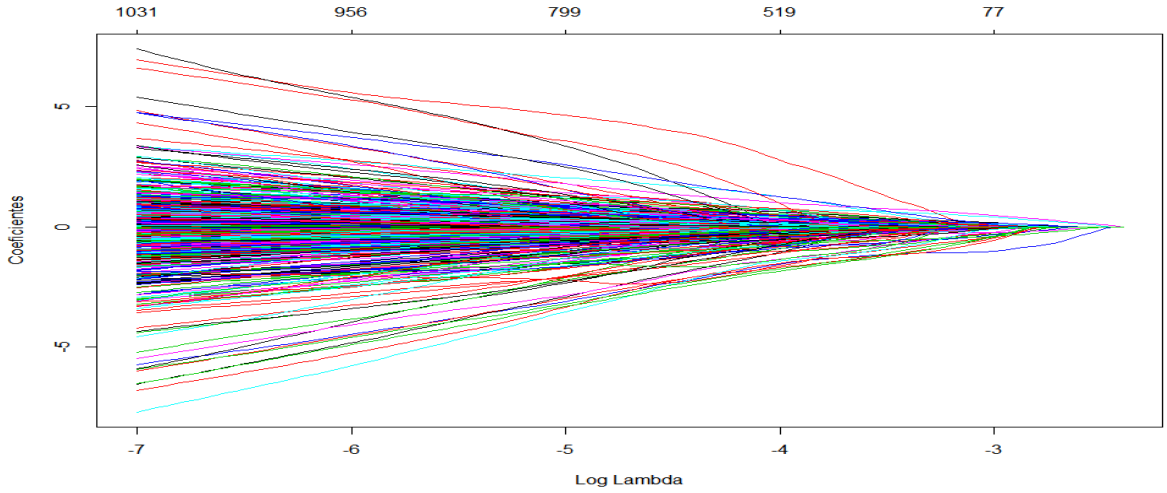

Figura 5.6: Trayectoria de las estimaciones de los coeficientes de la regresión *elastic net* según el valor del logaritmo de λ considerando un patrón de herencia aditivo.

SNP ID	GEN	Coeficiente	OR	Comp. alélica
1067	PKP2	-0.0423	0.9586	1
2052	KCNH2	-0.1184	0.8884	2
4475	CASQ2	-0.5789	0.5605	1
6106	LDLR	0.0018	1.0018	2
19318	BAG3	0.2274	1.2553	2
22541	RBM20	0.0116	1.0116	2
23688	PSEN2	-0.0205	0.9797	1
51429	TTN	0.1857	1.2040	1
53838	NOTCH1	-0.1670	0.8462	1
419280	SCN10A	0.0364	1.0371	2

Tabla 5.9: SNPs seleccionados mediante *elastic net* considerando patrón de herencia aditivo. Por columnas, de izquierda a derecha: identificador de la variante, gen en que se encuentra, coeficiente estimado en la regresión para esa variante, OR de la presencia de la variante con respecto a la no presencia de la variante y composición alélica para la cual se ha seleccionado.

Patrón de herencia autosómico dominante

Se ha considerado en este caso que la herencia de la enfermedad sigue un patrón autosómico dominante.

En la Figura 5.7 se puede ver la trayectoria de la estimación de los coeficientes según el valor del logaritmo de λ. El valor de λ es 0.0551 ($\log(\lambda) = -2.8983$), para el cual se seleccionan 30 variantes.

La Tabla 5.10 muestra los SNPs seleccionados, el gen en que se encuentran, la estimación del coeficiente de regresión para esa covariable y su correspondiente OR.

SNP ID	GEN	Coeficiente	OR
976	TMPO	−0.5383	0.5837
1067	PKP2	−0.2873	0.7503
3350	DSP	0.0581	1.0599
4475	CASQ2	−0.2746	0.7599
12970	KCNA5	−0.0333	0.9672
14779	CACNA1C	−0.2495	0.7792
19318	BAG3	0.2260	1.2535
23444	MAP2K1	0.0730	1.0757
23551	RBM20	−0.1911	0.8260
25000	NPPA	0.0580	1.0597
39387	TTN	−0.0371	0.9636
41936	TTN	0.0401	1.0409
45675	EYA4	0.0396	1.0404
45961	MYLK2	0.0171	1.0172
51030	CACNA1D	−0.0196	0.9806
51429	TTN	0.3440	1.4106
51837	TRDN	−0.0514	0.9499
51984	TTN	0.0360	1.0367
52965	KLF10	0.0288	1.0292
53335	TTN	−0.0788	0.9242
53838	NOTCH1	−0.3240	0.7233
69629	KLF10	0.0296	1.0301
74261	TTN	−0.0264	0.9739
79142	ANK2	0.0781	1.0812
158975	HFE	−0.0213	0.9789
159757	FHOD3	0.0931	1.0975
310891	SGCA	0.0027	1.0027
340420	SCN10A	0.0306	1.0311
423956	SCN10A	0.0091	1.0091
423960	SCN10A	0.0166	1.0168

Tabla 5.10: SNPs seleccionados mediante *elastic net* considerando patrón de herencia autosómico dominante. Por columnas, de izquierda a derecha: identificador de la variante, gen en que se encuentra, coeficiente estimado en la regresión para esa variante y OR de la presencia de la variante con respecto a la no presencia de la variante.

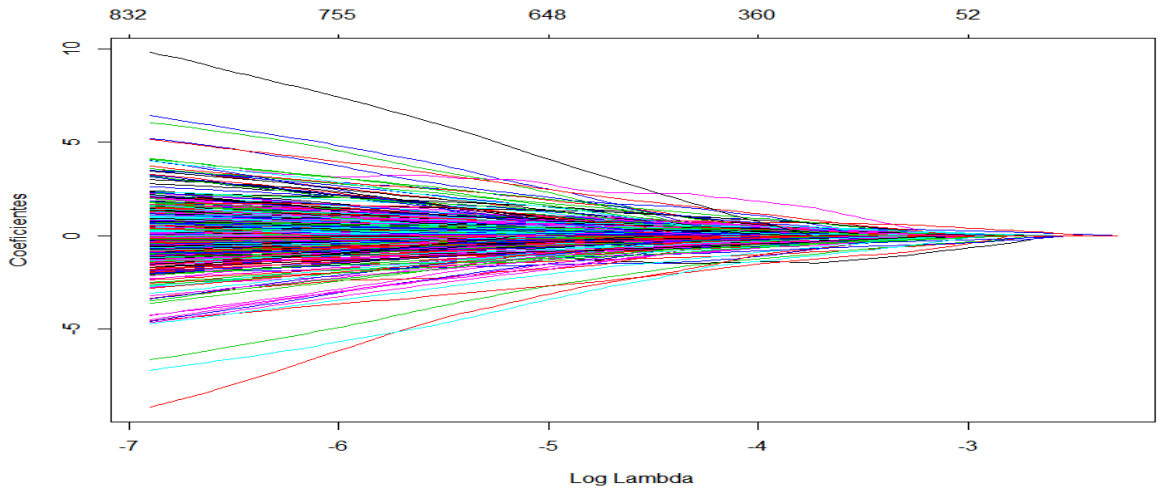

Figura 5.7: Trayectoria de las estimaciones de los coeficientes de la regresión *elastic net* según el valor del logaritmo de λ considerando patrón de herencia autosómico dominante.

5.3.4. Desequilibrio de ligamiento

El desequilibrio de ligamiento es otro de los factores que puede llevar a identificar erróneamente algún SNP como causante de la enfermedad del estudio ya que es un indicador de que existe una región en que ciertas variantes están correlacionadas. Ante la presencia de un grupo de SNPs correlacionados, la regresión *lasso* selecciona solamente una de las variantes de ese grupo sin tener el cuenta el resto. Por otro lado, la regresión *elastic net*, ante la misma situación, seleccionaría todas las variantes del grupo. Aunque, en teoría, la regresión *elastic net* ya tiene en cuenta la correlación de las variables, es interesante utilizar el LD, ya que de este modo de realiza un análisis más pormenorizado de la correlación existente.

Para observar la presencia de regiones con SNPs en LD es interesante observar mapas de desequilibrio de ligamiento. En R pueden obtenerse estos mapas con la librería `LDheatmap` si se conoce, además del cambio de nucleótido, la posición y el cromosoma en que se encuentra la variante.

En la muestra de SNPs utilizados en este estudio, tras analizar por separado cada cromosoma, se han observado regiones con presencia de alto desequilibrio de ligamiento donde se encuentra alguna de las variantes seleccionadas, más concretamente, esto se ha observado en una región del cromosoma 9. En la Figura 5.8, correspondiente a ese cromosoma, se puede observar una región con alto LD. La Figura 5.9 muestra el mapa correspondiente a las variantes de esa región concreta. En esa región se encuentra el SNP con ID 53838 del gen NOTCH1 seleccionado tanto en *lasso* como en *elastic net*. Esto parece indicar que quizás ese SNP no sea realmente el SNP causante de la enfermedad, si no que el causante podría ser alguno de los que se encuentran en la misma región de alto LD que esa. En la Tabla 5.11 se pueden ver los valores de la medida r^2 entre todos los pares de SNPs de esa región. Se considera que existe LD cuando los valores de r^2 son mayores de 0.2, sin embargo, en este caso, se puede ver que los valores son muy superiores, indicando así que el LD es muy alto.

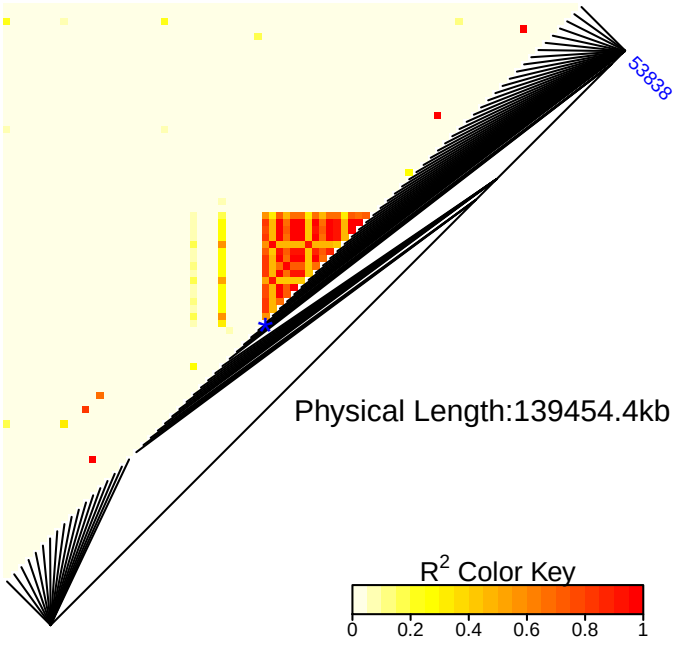

Figura 5.8: Mapa de desequilibrio de ligamiento para el cromosoma 9.

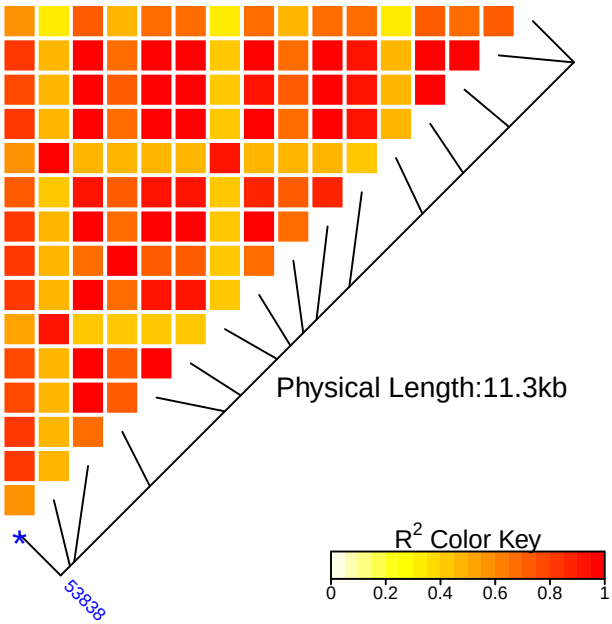

Figura 5.9: Mapa de desequilibrio de ligamiento para la región estudiada en el cromosoma 9, que incluye el SNP con ID 53838 seleccionada en las regresiones *lasso* y *elastic net*.

SNP ID	53838	149031	23267	24307	102396	30038	23681	23964	47901	23926	23785	316693	20693	50184	66390	25408
53838		0.56	0.81	0.83	0.79	0.79	0.53	0.81	0.84	0.81	0.75	0.56	0.81	0.79	0.81	0.57
149031			0.46	0.46	0.45	0.45	0.94	0.46	0.46	0.46	0.42	1.00	0.46	0.45	0.46	0.33
23267				0.67	0.95	0.95	0.44	0.96	0.68	0.97	0.92	0.46	0.99	0.95	0.99	0.73
24307					0.71	0.71	0.44	0.67	0.98	0.67	0.71	0.46	0.67	0.71	0.67	0.48
102396						0.99	0.43	0.95	0.72	0.95	0.93	0.45	0.95	0.99	0.95	0.69
30038							0.43	0.95	0.72	0.95	0.93	0.45	0.95	0.99	0.95	0.69
23681								0.44	0.44	0.44	0.40	0.94	0.44	0.43	0.44	0.31
23964									0.68	0.96	0.89	0.46	0.96	0.95	0.96	0.69
47901										0.68	0.72	0.46	0.68	0.72	0.68	0.48
23926											0.90	0.46	0.97	0.95	0.97	0.70
23785												0.42	0.92	0.93	0.92	0.67
316693													0.46	0.45	0.46	0.33
20693														0.95	0.99	0.73
50184															0.95	0.69
66390																0.73
25408																

Tabla 5.11: Medidas de r^2 entre cada par de SNPs para las variantes de la región en LD analizada en el cromosoma 9.

5.4. Resultados

Aunque se podría haber llevado a cabo un análisis más exhaustivo de los resultados, puesto que la principal finalidad de este estudio es identificar las variantes que pueden estar asociadas con la MCH, se ha dejado a un lado la interpretación de los coeficientes del modelo en cuanto a la cuantificación del efecto que pudiese tener cada variante. Se ha hecho hincapié en la selección de las variantes y si estas protegen de la enfermedad o, en caso contrario, incrementan el riesgo de padecerla.

Así, considerando un patrón de herencia aditivo, en el caso de la regresión logística clásica por variante, se ha encontrado, inicialmente, asociación con la enfermedad (bien como factor de riesgo, bien como factor protector) en 62 SNPs. Sin embargo, tras el ajuste de los p-valores para detectar falsos positivos debido a la multiplicidad de contrastes, se tiene que solamente una variante estaría asociada con la MCH. En el caso de la regresión *lasso*, se han seleccionado 10 variantes, las mismas que se han seleccionado utilizando la regresión *elastic net*. Entre esas variantes seleccionadas se encuentra el SNP con asociación significativa en regresión logística clásica. Además, los ORs para cada variante son muy similares en ambos modelos. Es destacable el hecho de que los ORs obtenidos en los modelos penalizados son siempre cercanos a 1, indicando así que el efecto sobre la respuesta es pequeño.

En el caso del patrón de herencia dominante, considerando una regresión logística por cada variante, se obtiene, en principio, que 152 variantes podrían estar asociadas con la MCH. Tras la corrección de los p-valores, ninguno de los contrastes proporcionaría un p-valor ajustado menor que 0.05, por lo que ninguna de las variantes tendría asociación con la MCH. Cuando se utilizan modelos de regresión penalizada sí se seleccionan variantes. En concreto, en el caso de la regresión *lasso*, se han seleccionado 25 variantes y, en el caso de *elastic net*, 30. Todas las variantes seleccionadas en el *lasso* se han seleccionado también en *elastic net*. Además, como cabría esperar, los ORs son muy similares en los dos casos. Aunque *elastic net* selecciona cinco SNPs más, se puede observar que los coeficientes estimados en esos casos son, en valor absoluto, los más pequeños. En cualquier caso, como en el patrón de herencia aditivo, aquí también se tiene que los ORs son cercanos a 1, por lo que, de nuevo, el efecto sobre la respuesta es bastante reducido.

Teóricamente, parece que la regresión *elastic net* es la mejor opción en este tipo de estudios pues tiene en cuenta la posibilidad de que existan múltiples variantes con un pequeño efecto en la variable respuesta y además presta atención a la correlación entre las variables, algo en lo que *lasso* fallaría. Sin embargo, a la vista de los resultados de este estudio se puede observar que ambos modelos se comportan de forma muy semejante y que, de hecho, *elastic net* no logra identificar correctamente los grupos de variantes correlacionadas, como se ha comprobado con la cuantificación del LD en una región del cromosoma 9.

En definitiva, aunque las técnicas de regresión logística penalizada resultan muy útiles para realizar estudios caso-control, es importante destacar el hecho de que, en este caso, los resultados obtenidos no son demasiado concluyentes. También es fundamental tener en cuenta que, cuando se habla de asociaciones de variantes genéticas con un fenotipo, el efecto asignable a cada variante suele ser muy pequeño, como demuestran los numerosos estudios publicados recientemente, algo que explica también el hecho de que en la regresión logística SNP a SNP se obtengan, en general, p-valores tan altos. Quizás considerar bases de datos más grandes pueda proporcionar resultados más determinantes.

Capítulo 6

Conclusiones

A lo largo de este trabajo se han descrito las principales técnicas existentes para la búsqueda de asociaciones entre variantes genéticas y la MCH en un estudio de tipo caso-control. Además, se ha realizado una aplicación de estas técnicas considerando dos patrones de herencia: aditivo y autosómico dominante. Una de las dificultades encontradas a la hora de realizar la aplicación a los datos ha sido transformar la base de datos en cada caso para adaptarla a los distintos paquetes desarrollados en R, ya que la implementación de algunas de las técnicas usadas es bastante compleja, por lo que resulta más eficiente utilizar los recursos disponibles. Asimismo, en algunos casos, por ejemplo, a la hora de generar matrices y gráficos de LD entre regiones de muchos SNPs, las exigencias computacionales hacen que el tiempo de ejecución del proceso se alargue.

En primer lugar, se ha analizado la base de datos disponible para controlar que esta verifique ciertas condiciones. En un acercamiento inicial a la búsqueda de asociaciones se ha empleado un modelo de regresión logística para cada SNP ajustando el efecto en la respuesta por las covariables sexo y edad. Utilizando esta metodología y, tras considerar técnicas de control del FDR para detectar la presencia de falsos positivos, solamente se ha seleccionado un SNP en el caso de patrón de herencia aditivo, y ninguno en el caso del modelo autosómico dominante.

Un problema de considerar un ajuste por cada SNP es que no se contempla la posibilidad de que existan múltiples pequeños efectos que contribuyan a la respuesta, algo que sí tienen en cuenta los modelos de regresión multivariante. No obstante, cuando el número de observaciones es mayor que el número de individuos, como en este estudio, es necesario usar modelos penalizados. Se han considerado dos modelos de regresión logística penalizada, *lasso* y *elastic net*, también para ambos patrones de herencia. Como este tipo de modelos realizan selección de variables, su interpretación es mucho más sencilla. En ambos modelos se han obtenidos asociaciones entre numerosas variantes de varios genes diferentes.

Las variantes analizadas en este estudio se encuentran, debido a la naturaleza de la base de datos, en genes susceptibles de estar asociados con la MCH y con otras cardiopatías. La mayor parte de los genes que actualmente se asocian con la MCH son genes que codifican proteínas del sarcómero cardíaco, por ejemplo, los genes MYH7 (cadena pesada de beta miosina) y MYBPC3 (proteína C de unión a miosina). Sin embargo, debido a la funcionalidad de las proteínas codificadas, estos genes también se pueden asociar a otros fenotipos con características parecidas, como ocurre en el caso de la Miocardiopatía Dilatada (MCD) (ver Monserrat y Dumont, 2007). Por ende, la elección del grupo control en este estudio condiciona bastante los resultados obtenidos. De hecho, se puede observar que no se ha seleccionado ningún SNP de esos dos genes. Por tanto, aunque resulta interesante identificar variantes que permitan esclarecer si se está ante la presencia de MCH o de MCD (u otra cardiopatía hereditaria), es una tarea compleja. Quizás considerar una muestra de individuos más grande pueda facilitar esa labor.

Aunque los resultados obtenidos no sean demasiado significativos, estos pueden ser útiles para focalizar los esfuerzos médicos en el estudio de las variantes que han mostrado cierta asociación,

aunque el efecto, a primera vista, sea muy débil.

Otra limitación importante de este tipo de metodologías es que no tienen en cuenta las variantes poco frecuentes, sin embargo, puede ser que estas estén implicadas en la manifestación de la enfermedad.

A pesar de las limitaciones de estos estudios, este tipo de técnicas resultan de gran utilidad en el campo de la genética y han sido capaces de identificar numerosos SNPs, mutaciones y regiones genéticas asociadas a diversas enfermedades. Además, el abaratamiento en los costes de la secuenciación del DNA proporciona hoy en día una gran cantidad de información que necesita ser analizada con metodologías estadísticas y computacionales adecuadas. En consecuencia, Health in Code S.L. trabaja actualmente en la implementación de algoritmos y técnicas de base estadística que faciliten la búsqueda de asociaciones. Debido a la complejidad de las técnicas empleadas en este proyecto no resulta fácil su desarrollo y automatización en lenguajes de programación no estadísticos. Sin embargo, para agilizar la aplicación de estos métodos y ayudar a seleccionar las bases de datos adecuadas, durante la estancia en la empresa, se ha trabajado, en colaboración con el departamento de Bioinformática, en la implementación de algún proceso de este tipo. Así, se han implementado tests estadísticos que permitan detectar diferencias significativas en las frecuencias de las variantes según el fenotipo observado en los pacientes. Con tal fin, se han programado e incorporado en el software propio de la empresa tests χ^2 de Pearson y tests exactos que detectan de forma automática diferencias entre las frecuencias de aparición de cada variante. En el Apéndice B se puede ver una imagen de la interfaz gráfica que muestra el resultado de esta técnica para una variante genética. Es necesario observar estos resultados con cautela pues, tal y como se ha explicado anteriormente, la realización de cualquier test de forma masiva lleva a la acumulación de falsos positivos.

Bibliografía

[1] Agresti A (2002) *Categorical Data Analysis*. Wiley, New York.

[2] American Heart Association. Hypertrophic Cardiomyopathy. http://www.heart.org/HEARTORG/Conditions/More/Cardiomyopathy/Hypertrophic-Cardiomyopathy_UCM_444317_Article.jsp#.WR7jXOuLTIV Accedido 15 de abril de 2017.

[3] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B 57:289-300.

[4] Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 29:1165-1188.

[5] Ensoy C, Rakhmawati TW, Faes C, Aerts M (2015) Separation issues and possible solutions: Part I - Systematic literature review on logistic models. EFSA Journal.

[6] Fisher RA (1922). On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85:87-94.

[7] Foulkes AS (2009) *Applied Statistical Genetics with R*. Springer, Nueva York.

[8] Hastie T, Tibshirani R, Wainwright M (2015) *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, Florida.

[9] Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55-67.

[10] Le Cessie S, Van Houwelingen JC (1992) Ridge estimators in logistic regression. Applied Statistics 41:191-201.

[11] Mateu-Figueras G, Martín-Fernández JA, Pawlowsky-Glahn V, Barceló-Vidal C (2003) El problema del análisis estadístico de datos composicionales. 27 Congreso Nacional de Estadística e Investigación Operativa, Lleida.

[12] Monserrat L, Dumont CA (Octubre, 2007) Utilidad de la genética en la miocardiopatía hipertrófica. Documento presentado en el 5º Congreso Virtual de Cardiología.

[13] Tian C, Gregersen PK, Seldin, MF (2008) Accounting for ancestry: population substructure and genome-wide association studies. Human Molecular Genetics 17:143-150.

[14] Turner S, Armstrong LL, Bradford Y et al. (2011) Quality control procedures for genome wide association studies. Current protocols in human genetics 1. Unit 1.

[15] Stram DO (2014) *Design, Analysis and Interpretation of Genome-Wide Association Scans*. Springer, Londres.

[16] Weir, BS (1996) *Genetic Data Analysis II*. Sinauer Associates, Massachusetts.

[17] Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy-Weinberg equilibrium. American Journal of Human Genetics 76:887-893.

[18] Wilks, SS (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics 9:60-62.

[19] Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 67(2):301-320.

Apéndice A

Tablas

Regresión logística clásica. Patrón de herencia aditivo.

Tabla A.1: SNPs que muestran asociación en regresión logística clásica considerando patrón de herencia aditivo. Por columnas, de izquierda a derecha: identificador de la variante, gen en que se encuentra la variante, OR de la presencia de la variante en heterocigosis con respecto a la no presencia de la variante, intervalo de confianza al 95 % para el OR anterior, OR de la presencia de la variante en homocigosis con respecto a la no presencia de la variante, intervalo de confianza al 95 % para el OR anterior, p-valor del test y p-valor ajustado.

SNP ID	GEN	OR_1	IC para OR_1	OR_2	IC para OR_2	p-val.	p-val. ajust.
19318	BAG3	1.55	(1.15, 2.08)	7.33	(2.57, 20.92)	1.455e-06	0.0015
419280	SCN10A	1.43	(0.98, 2.08)	2.32	(1.54, 3.49)	0.0001	0.0626
340420	SCN10A	1.52	(1.04, 2.21)	2.29	(1.53, 3.44)	0.0002	0.0753
23551	RBM20	0.48	(0.32, 0.72)	0.63	(0.41, 0.97)	0.0008	0.2081
1541	SCN5A	1.43	(1.07, 1.92)	2.50	(1.24, 5.04)	0.0034	0.4424
44321	TTN	2.22	(1.28, 3.84)	0.14	(0.01, 2.72)	0.0035	0.4424
24404	TBX20	1.47	(1.10, 1.97)	1.92	(1.21, 3.05)	0.0039	0.4424
24134	TBX20	1.47	(1.10, 1.97)	1.92	(1.21, 3.05)	0.0039	0.4424
429390	GATA6	1.85	(1.01, 3.39)	8.99	(1.10, 73.10)	0.0040	0.4424
160857	FHOD3	7.33	(1.62, 33.25)	0.85	(0.07, 10.07)	0.0054	0.4909
6107	LDLR	1.10	(0.81, 1.49)	1.90	(1.25, 2.90)	0.0058	0.4909
6106	LDLR	1.02	(0.75, 1.39)	1.85	(1.21, 2.82)	0.0059	0.4909
11373	SCN1B	0.65	(0.46, 0.91)	2.04	(0.83, 5.03)	0.0073	0.5203
18966	GATA4	1.04	(0.72, 1.50)	13.24	(1.59, 110.55)	0.0075	0.5203
22541	RBM20	0.96	(0.72, 1.29)	3.11	(1.36, 7.11)	0.0078	0.5203
23688	PSEN2	0.48	(0.23, 0.99)	0.71	(0.35, 1.46)	0.0097	0.5869
13146	AGL	0.84	(0.61, 1.16)	1.45	(0.98, 2.16)	0.0100	0.5869

(Continúa en la página siguiente)

SNP ID	GEN	OR_1	IC para OR_1	OR_2	IC para OR_2	p-val.	p-val. ajust.
310891	SGCA	1.91	(1.17, 3.12)	0.27	(0.04 , 1.98)	0.0107	0.5869
9821	LDLR	0.87	(0.64, 1.20)	1.55	(1.02, 2.35)	0.0125	0.5869
28515	OBSL1	1.59	(1.17, 2.17)	1.40	(0.96, 2.06)	0.0126	0.5869
268014	CTNNA3	1.43	(0.94, 2.19)	0.18	(0.04, 0.79)	0.0142	0.5869
5877	SCN5A	1.48	(1.10, 2.00)	1.92	(0.84, 4.36)	0.0148	0.5869
23800	AGL	0.86	(0.61, 1.23)	1.40	(0.94, 2.09)	0.0151	0.5869
2686	LMNA	0.96	(0.68, 1.37)	5.98	(1.33, 26.94)	0.0151	0.5869
5698	RYR2	0.74	(0.50, 1.07)	1.15	(0.76, 1.73)	0.0152	0.5869
21826	PSEN2	0.48	(0.23, 1.01)	0.70	(0.34, 1.44)	0.0153	0.5869
5875	SCN5A	1.46	(1.08, 1.99)	2.07	(0.82, 5.23)	0.0179	0.6635
159757	FHOD3	1.47	(1.09, 1.97)	1.54	(1.00, 2.39)	0.0197	0.7022
23087	PSEN2	0.50	(0.26, 0.96)	0.70	(0.37, 1.32)	0.0216	0.7170
5892	LDLR	0.93	(0.68, 1.27)	1.57	(1.04, 2.38)	0.0217	0.7170
426974	TXNRD2	1.04	(0.76, 1.42)	8.22	(1.08, 62.83)	0.0223	0.7170
2052	KCNH2	1.29	(0.95, 1.73)	0.58	(0.33, 1.03)	0.0236	0.7218
21871	CACNA1D	0.30	(0.08, 1.06)	0.25	(0.07, 0.86)	0.0238	0.7218
423960	SCN10A	2.26	(1.17, 4.37)	2.34	(0.24, 23.15)	0.0263	0.7435
3216	DSP	1.17	(0.86, 1.58	0.68)	(0.45, 1.01)	0.0277	0.7435
26676	TRIM63	0.73	(0.54, 0.98)	0.53	(0.28, 0.98)	0.0278	0.7435
28514	JPH2	1.36	(1.01, 1.82)	0.77	(0.48, 1.24)	0.0296	0.7435
423956	SCN10A	2.19	(1.14, 4.24)	2.34	(0.24, 23.07)	0.0328	0.7435
2287	DSC2	4.17	(1.19, 14.61)	0.59	(0.03, 10.25)	0.0329	0.7435
32880	FLNC	1.04	(0.77, 1.42)	2.85	(1.20, 6.72)	0.0331	0.7435
21969	AGL	0.84	(0.61, 1.16)	1.35	(0.91, 2.02)	0.0331	0.7435
14735	CACNA1C	1.02	(0.76, 1.36)	0.42	(0.22, 0.80)	0.0333	0.7435
8413	KCNH2	1.47	(1.09, 1.98)	1.12	(0.72, 1.74)	0.0348	0.7435
9833	LDLR	1.17	(0.85, 1.61)	1.69	(1.12, 2.53)	0.0351	0.7435
23290	TTN	1.55	(1.05, 2.28)	0.46	(0.13, 1.63)	0.0353	0.7435
23787	CACNA1D	1.11	(0.69, 1.79)	0.76	(0.48, 1.21)	0.0353	0.7435
22498	CACNA1C	1.50	(0.99, 2.27)	1.05	(0.70, 1.58)	0.0355	0.7435
4070	MYL3	1.00	(0.73, 1.36)	2.63	(1.17, 5.92)	0.0366	0.7435
24828	CACNA1D	0.95	(0.68, 1.32)	1.44	(0.98, 2.13)	0.0390	0.7435
27206	LAMA2	1.42	(1.05, 1.90)	0.89	(0.54, 1.47)	0.0395	0.7435

(Continúa en la página siguiente)

SNP ID	GEN	OR_1	IC para OR_1	OR_2	IC para OR_2	p-val.	p-val. ajust.
23465	TRDN	4.63	(1.20, 17.88)	3.08	(0.85, 11.23)	0.0399	0.7435
2611	PRKAG2	0.66	(0.46, 0.96)	0.34	(0.09, 1.33)	0.0400	0.7435
158975	HFE	0.69	(0.51, 0.94)	0.68	(0.44, 1.04)	0.0401	0.7435
24640	PSEN2	0.53	(0.25, 1.09)	0.73	(0.35, 1.49)	0.0403	0.7435
23444	MAP2K1	1.48	(1.00, 2.17)	2.04	(0.83, 5.00)	0.0410	0.7435
39387	TTN	0.54	(0.33, 0.87)	0.39	(0.02, 7.53)	0.0423	0.7435
27822	NEBL	1.01	(0.76, 1.35)	2.25	(1.14, 4.46)	0.0427	0.7435
159008	ALMS1	0.61	(0.39, 0.95)	0.57	(0.36, 0.91)	0.0432	0.7435
24084	TBX5	1.33	(0.83, 2.15)	1.72	(1.07, 2.78)	0.0460	0.7730
973	HFE	2.05	(1.10, 3.82)	2.11	(0.19, 23.81)	0.0479	0.7730
2047	KCNH2	1.28	(0.95, 1.72)	0.62	(0.34, 1.14)	0.0488	0.7730
4069	MYL3	1.02	(0.74, 1.40)	2.56	(1.13, 5.78)	0.0492	0.7730

Regresión logística clásica. Patrón de herencia autosómico dominante.

Tabla A.2: SNPs con asociación significativa en regresión logística clásica considerando patrón de herencia autosómico dominante. Por columnas, de izquierda a derecha: identificador de la variante, gen en que se encuentra la variante, OR de la presencia de la variante con respecto a la no presencia de la variante, intervalo de confianza al 95 % para el OR, p-valor del test y p-valor ajustado.

SNP ID	GEN	OR	IC para OR	p-val.	p-val. ajust.
19318	BAG3	1.78	(1.33, 2.37)	0.0001	0.1204
51429	TTN	6.03	(1.82, 19.97)	0.0002	0.1204
51984	TTN	2.67	(1.50, 4.72)	0.0002	0.1204
51676	TTN	4.49	(1.57, 12.86)	0.0008	0.1204
51697	TTN	4.49	(1.57, 12.86)	0.0008	0.1204
51853	TTN	4.49	(1.57, 12.86)	0.0008	0.1204
51964	TTN	4.49	(1.57, 12.86)	0.0008	0.1204
23551	RBM20	0.53	(0.36, 0.78)	0.0008	0.1204
269056	TTN	4.44	(1.55, 12.75)	0.0008	0.1204
446918	TXNRD2	5.66	(1.59, 20.11)	0.0015	0.1204
340420	SCN10A	1.80	(1.26, 2.57)	0.0016	0.1204
24134	TBX20	1.56	(1.18, 2.06)	0.0018	0.1204
24404	TBX20	1.56	(1.18, 2.06)	0.0018	0.1204
51164	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
51473	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
51475	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
51720	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
51876	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
51938	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
144713	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
144756	TTN	3.65	(1.41, 9.46)	0.0020	0.1204
51303	TTN	3.67	(1.41, 9.54)	0.0020	0.1204
3831	SCN5A	4.47	(1.49, 13.42)	0.0020	0.1204
51664	TTN	3.55	(1.37, 9.22)	0.0025	0.1204
51148	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51213	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51521	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51626	TTN	3.53	(1.36, 9.16)	0.0027	0.1204

(Continúa en la página siguiente)

SNP ID	GEN	OR	IC para OR	p-val.	p-val. ajust.
51642	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51661	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51695	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51754	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51824	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51969	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51982	TTN	3.53	(1.36, 9.16)	0.0027	0.1204
51559	TTN	3.04	(1.34, 6.91)	0.0027	0.1204
51602	TTN	3.04	(1.34, 6.91)	0.0027	0.1204
51712	TTN	3.04	(1.34, 6.91)	0.0027	0.1204
51892	TTN	3.04	(1.34, 6.91)	0.0027	0.1204
419280	SCN10A	1.74	(1.22, 2.48)	0.0029	0.1233
1541	SCN5A	1.53	(1.15, 2.03)	0.0029	0.1233
74261	TTN	0.39	(0.21, 0.71)	0.0033	0.1344
160857	FHOD3	4.98	(1.41, 17.64)	0.0034	0.1344
28515	OBSL1	1.53	(1.15, 2.04)	0.0039	0.1482
86864	VCL	4.18	(1.37, 12.70)	0.0040	0.1482
87312	VCL	4.18	(1.37, 12.70)	0.0040	0.1482
429390	GATA6	2.23	(1.25, 3.98)	0.0042	0.1517
51576	TTN	3.36	(1.29, 8.75)	0.0042	0.1517
5877	SCN5A	1.52	(1.13, 2.03)	0.0045	0.1597
12853	TTN	2.46	(1.24, 4.86)	0.0051	0.1679
51493	TTN	2.46	(1.24, 4.86)	0.0051	0.1679
51847	TTN	2.46	(1.24, 4.86)	0.0051	0.1679
159757	FHOD3	1.48	(1.13, 1.96)	0.0052	0.1679
79101	ANK3	5.86	(1.28, 26.90)	0.0055	0.1679
12854	TTN	2.44	(1.23, 4.83)	0.0056	0.1679
51329	TTN	2.44	(1.23, 4.83)	0.0056	0.1679
51593	TTN	2.44	(1.23, 4.83)	0.0056	0.1679
3350	DSP	2.49	(1.23, 5.04)	0.0057	0.1679
44321	TTN	2.02	(1.19, 3.42)	0.0058	0.1679
51293	TTN	2.85	(1.24, 6.54)	0.0059	0.1679
51709	TTN	2.99	(1.24, 7.22)	0.0059	0.1679

(Continúa en la página siguiente)

SNP ID	GEN	OR	IC para OR	p-val.	p-val. ajust.
5875	SCN5A	1.51	(1.12, 2.03)	0.0062	0.1725
108077	DES	8.48	(1.07, 66.95)	0.0069	0.1873
423960	SCN10A	2.26	(1.20, 4.28)	0.0070	0.1873
445181	DLD	9.13	(1.11, 75.08)	0.0071	0.1873
161120	FHOD3	5.44	(1.21, 24.44)	0.0072	0.1873
4597	LMNA	8.91	(1.09, 73.06)	0.0083	0.2131
508389	BSCL2	8.32	(1.05, 65.92)	0.0084	0.2131
41936	TTN	3.05	(1.18, 7.87)	0.0089	0.2154
299279	TTN	8.30	(1.04, 66.15)	0.0090	0.2154
423956	SCN10A	2.20	(1.17, 4.16)	0.0090	0.2154
53838	NOTCH1	0.38	(0.18, 0.77)	0.0090	0.2154
445180	DLD	5.62	(1.20, 26.38)	0.0094	0.2217
90506	PSEN2	5.38	(1.17, 24.83)	0.0104	0.2425
47132	MYOT	2.60	(1.17, 5.79)	0.0109	0.2503
158975	HFE	0.69	(0.52, 0.92)	0.0113	0.2566
69473	TTN	3.00	(1.14, 7.86)	0.0117	0.2621
8698	MYH7	7.46	(0.95, 58.78)	0.0122	0.2666
39387	TTN	0.54	(0.33, 0.86)	0.0122	0.2666
446630	GLB1	0.42	(0.22, 0.81)	0.0125	0.2682
26676	TRIM63	0.70	(0.53, 0.93)	0.0128	0.2732
159008	ALMS1	0.59	(0.39, 0.91)	0.0131	0.2761
973	HFE	2.05	(1.12, 3.76)	0.0137	0.2819
21871	CACNA1D	0.26	(0.07, 0.90)	0.0138	0.2819
73847	EYA4	0.31	(0.13, 0.76)	0.0139	0.2819
23444	MAP2K1	1.55	(1.08, 2.22)	0.0149	0.2943
2045	KCNE2	1.95	(1.11, 3.41)	0.0150	0.2943
310891	SGCA	1.76	(1.09, 2.83)	0.0150	0.2943
440075	TXNRD2	3.51	(1.13, 10.92)	0.0156	0.3021
1067	PKP2	0.36	(0.16, 0.82)	0.0167	0.3191
3181	DSP	0.56	(0.35, 0.90)	0.0178	0.3340
2611	PRKAG2	0.64	(0.44, 0.92)	0.0180	0.3340
99300	TTN	2.98	(1.10, 8.12)	0.0180	0.3340
45675	EYA4	3.05	(1.05, 8.80)	0.0192	0.3526

(Continúa en la página siguiente)

SNP ID	GEN	OR	IC para OR	p-val.	p-val. ajust.
2287	DSC2	3.25	(1.07, 9.85)	0.0199	0.3614
25000	NPPA	1.63	(1.06, 2.48)	0.0202	0.3620
9247	MYH7	6.90	(0.85, 55.79)	0.0210	0.3709
28053	ANK2	1.47	(1.05, 2.04)	0.0211	0.3709
8413	KCNH2	1.38	(1.05, 1.82)	0.0213	0.3709
96260	TBX5	7.72	(0.89, 66.66)	0.0219	0.3758
120084	KCND3	7.57	(0.89, 64.77)	0.0220	0.3758
25397	MYH11	1.39	(1.05, 1.84)	0.0227	0.3841
99149	TTN	2.88	(1.05, 7.90)	0.0234	0.3895
14779	CACNA1C	0.13	(0.02, 0.80)	0.0238	0.3895
69629	KLF10	2.29	(1.05, 5.01)	0.0241	0.3895
419555	FOXD4	0.71	(0.52, 0.95)	0.0241	0.3895
79142	ANK2	2.12	(1.05, 4.30)	0.0242	0.3895
33313	PSEN1	0.67	(0.47, 0.96)	0.0245	0.3915
158748	ALMS1	0.62	(0.40, 0.95)	0.0253	0.3989
26334	ANK3	0.73	(0.55, 0.96)	0.0255	0.3989
24675	CRYAB	0.73	(0.55, 0.96)	0.0263	0.4077
51030	CACNA1D	0.64	(0.44, 0.95)	0.0275	0.4077
102396	NOTCH1	3.06	(1.01, 9.28)	0.0276	0.4077
35581	MYH11	1.81	(1.05, 3.12)	0.0279	0.4077
35885	MYH11	1.81	(1.05, 3.12)	0.0279	0.4077
93141	MAP2K2	6.00	(0.77, 46.46)	0.0279	0.4077
93275	MAP2K2	6.00	(0.77, 46.46)	0.0279	0.4077
158985	ALMS1	0.64	(0.42, 0.96)	0.0281	0.4077
629320	SURF1	5.98	(0.77, 46.46)	0.0282	0.4077
4514	RYR2	0.74	(0.56, 0.97)	0.0292	0.4162
52965	KLF10	1.90	(1.03, 3.49)	0.0292	0.4162
24421	ANK3	0.74	(0.56, 0.97)	0.0295	0.4163
28020	TRIM63	0.73	(0.56, 0.97)	0.0318	0.4420
3313	KCNQ1	4.34	(0.91, 20.61)	0.0318	0.4420
102460	FHL1	6.76	(0.78, 58.23)	0.0332	0.4575
87509	KCNJ8	6.72	(0.78, 57.78)	0.0336	0.4593
13436	KCNQ1	4.27	(0.90, 20.34)	0.0350	0.4729

(Continúa en la página siguiente)

SNP ID	GEN	OR	IC para OR	p-val.	p-val. ajust.
1906	PKP2	2.48	(0.99, 6.22)	0.0351	0.4729
53335	TTN	0.53	(0.29, 0.95)	0.0373	0.4949
99278	TTN	2.04	(1.00, 4.14)	0.0373	0.4949
72674	TTN	2.08	(1.00, 4.33)	0.0384	0.4972
98779	TTN	2.08	(1.00, 4.33)	0.0384	0.4972
99310	TTN	2.08	(1.00, 4.33)	0.0384	0.4972
17121	MAP2K2	1.37	(1.02, 1.83)	0.0398	0.5112
419826	BSCL2	0.48	(0.23, 1.02)	0.0403	0.5112
268996	CTNNA3	5.13	(0.68, 38.43)	0.0403	0.5112
39334	RYR2	1.69	(1.00, 2.88)	0.0431	0.5286
34248	OBSL1	0.48	(0.24, 0.95)	0.0433	0.5286
159275	HFE	0.70	(0.49, 1.00)	0.0433	0.5286
45961	MYLK2	1.50	(1.00, 2.26)	0.0437	0.5286
120256	SCN5A	4.02	(0.85, 19.03)	0.0443	0.5286
16944	NEXN	0.75	(0.56, 0.99)	0.0445	0.5286
31191	ANK2	0.72	(0.53, 0.99)	0.0445	0.5286
2781	DSG2	1.43	(1.00, 2.05)	0.0447	0.5286
159324	ALMS1	0.66	(0.44, 1.00)	0.0448	0.5286
24252	OBSL1	0.65	(0.43, 1.00)	0.0448	0.5286
20800	MYPN	1.65	(0.99, 2.75)	0.0462	0.5387
51837	TRDN	0.59	(0.36, 0.98)	0.0463	0.5387
25320	TTN	1.65	(0.99, 2.76)	0.0469	0.5420
438124	GUSB	3.12	(0.89, 11.02)	0.0475	0.5459
2111	MYH7	2.65	(0.90, 7.83)	0.0489	0.5583
15292	NPPA	1.37	(1.00, 1.88)	0.0498	0.5644

Apéndice B

Ilustración de la interfaz gráfica

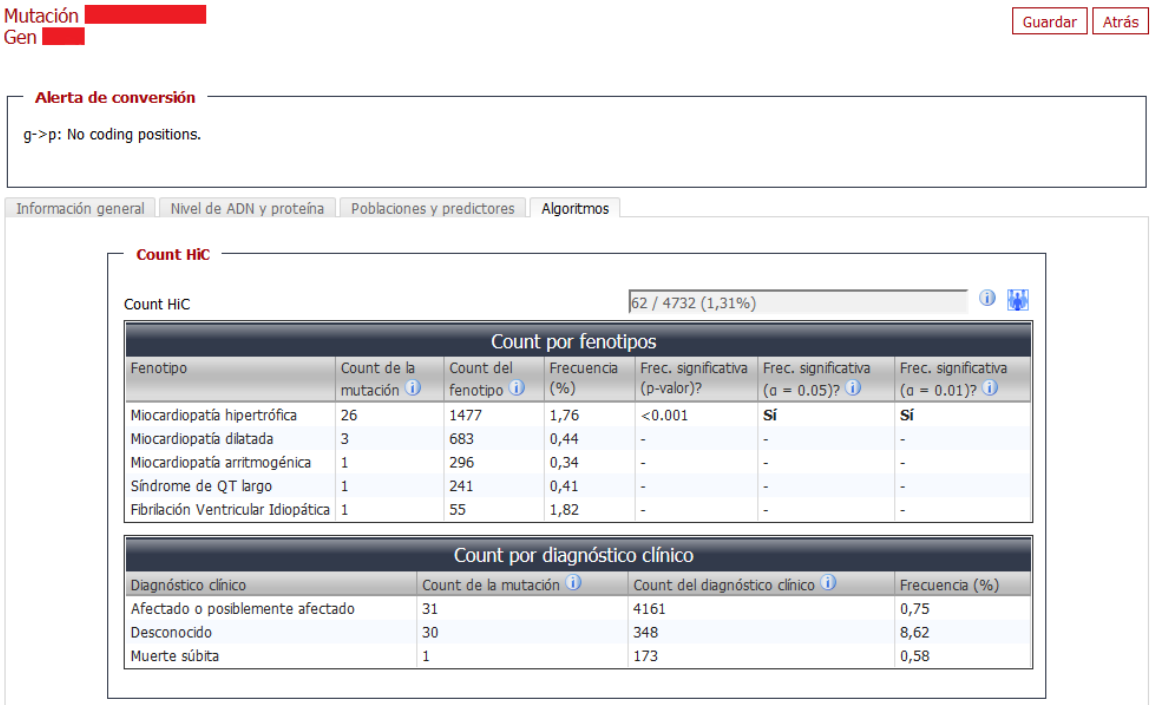

Figura B.1: Parte de la interfaz gráfica que muestra el resultado de la realización automática de un contraste de frecuencias para una variante genética. Por motivos de confidencialidad, se ha ocultado el nombre de la variante y el gen en que se encuentra.