TRABAJO DE FIN DE MÁSTER

Máster en Estadística e Investigación Operativa

Título: Detección de interacciones genéticas asociadas a enfermedades complejas. Aplicación al cáncer de vejiga.

Autor: Víctor Urrea Gales

Director: María Luz Calle Rosingana

Ponente: Josep Anton Sánchez Espigares

Departamento: Departament d'Estadística i Investigació Operativa

Convocatoria: Febrero 2009

Facultat de Matemàtiques i Estadística

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Detección de interacciones genéticas asociadas a enfermedades complejas. Aplicación al cáncer de vejiga

Master en Estadística e Investigación Operativa
Trabajo Final de Master

Autor: Víctor Urrea Gales
Directora: María Luz Calle Rosingana

Febrero 2009

Mi más sincero agradecimiento a Malu,
por su paciencia y tiempo dedicado

Índice

1. Introducción

Este trabajo resume mi participación en el proyecto "Genetic and Environmental Factors in Bladder Cancer Etiology and Prognosis", financiado por la Fundació La Marató de TV3 y realizado en el Departamento de Biología de Sistemas de la Universidad de Vic bajo la dirección y supervisión de la Dra. M.Luz Calle.

Mi contribución al proyecto ha consistido en la participación en las discusiones metodológicas, la prestación de soporte de carácter técnico y computacional y la aplicación de algunos métodos de minería de datos para la detección de interacciones y patrones genéticos asociados a la susceptibilidad de enfermedades complejas en datos genéticos de elevada dimensión y, en particular, su aplicación al estudio de cáncer de vejiga.

En concreto, el desarrollo de este proyecto ha requerido la elaboración de una serie de rutinas en R para la implementación de las nuevas metodologías propuestas o para la implementación de algunas variaciones en metodologías ya existentes (sección 8). La aplicación de estas técnicas a los datos de nuestro proyecto ha generado una serie de resultados que contribuyen, por un lado, al avance en el conocimiento sobre la componente genética del cáncer de vejiga y, por otro lado, a la mejora de las técnicas estadísticas existentes para el estudio de datos genéticos. Algunos de los resultados obtenidos se han publicado en forma de artículo científico o documento de investigación (Calle et al. 2007 y 2008). Otros resultados están todavía en proceso de redacción.

Este trabajo está organizado en 8 secciones, que comento brevemente. La primera de ellas corresponde a esta misma introducción. En la sección 2 se contextualiza este trabajo, tanto desde el punto de vista de los antecedentes y del origen del estudio, como de su importancia y de los objetivos planteados.

Como paso previo a la exposición del estudio realizado, la sección 3 introduce los conceptos y características importantes propias del tipo de datos de este proyecto. Concretamente, el apartado 3.1 introduce los conceptos biológicos, genéticos y relativos a epidemiología genética mínimos para poder entender el objetivo de este estudio. En el apartado 3.2 se explican los problemas y retos que plantean los estudios con datos genéticos debido a la dimensionalidad de los datos y que plantean la necesidad de técnicas estadísticas distintas a las clásicas. Las técnicas que surgen como respuesta a estos nuevos problemas tienen una serie de particularidades que se esbozan en el apartado 3.3 y que irán apareciendo en las secciones siguientes. En este apartado también se muestra un pequeño resumen de las técnicas más conocidas que permiten abordar nuestro problema desde alguno de sus flancos.

En las secciones de la 4 a la 7 expongo las diferentes aproximaciones al problema en las que he trabajado. Todas ellas tienen un primer apartado en donde hago una

introducción a la metodología utilizada, explico la idea principal en la que se basa el método y comento los aspectos teóricos y prácticos más importantes más importantes.

La sección 4 está dedicada al *Logic Regression*, una metodología que fue desarrollada especificamente para la detección de interacciones con datos genéticos y que ha sido el punto de partida de mi participación en este proyecto, y al *Logic FS*, una extensión del método anterior. Aunque dediqué muchas horas al estudio y aplicación de estas técnicas, tanto en los datos reales como en simulaciones, este esfuerzo no aparece reflejado porque los resultados que obtuvimos no fueron satisfactorios, tal y como se explica en el apartado 4.3. La falta de buenos resultados con estos métodos debidos a la dimensionalidad de los datos motivaron la búsqueda de otras metodologías que permitieran reducir esta dimensión y que desembocaron en el estudio de otras aproximaciones que aparecen en las secciones posteriores.

En la sección 5 se introducen una serie de medidas procedentes de la teoría de la información que son útiles como método de selección de variables. Como estas medidas no estaban implementadas, he elaborado las distintas rutinas y funciones en R necesarias para: (a) el cálculo de las diferentes medidas, cuyas funciones aparecen detalladas y explicadas en la sección 8.1, (b) el proceso de exploración exhaustiva de los datos del estudio a partir de estas medidas, (c) generación de distribuciones de referencia empíricas para la determinación de la significación de los datos, y (d) las simulaciones de datos necesarias para determinar la invariabilidad de esas distribuciones. En el apartado 5.2 comento los resultados obtenidos de la aplicación de dichas medidas a nuestro estudio, y en el 5.3 el estudio de las distribuciones de referencia y la significación de los resultados. Esta metodología nos ha proporcionado un criterio de selección de variables y, por tanto, nos permite reducir la dimensión de los datos como paso previo para el análisis de los datos con otras metodologías para las que esta dimensión suponga un hándicap como, por ejemplo, el método visto en la sección 4.

Los dos primeros apartados de la sección 6 están dedicados a una metodología, el MDR, que ha alcanzado una gran popularidad entre los estudios de la susceptibilidad genética a enfermedades, como el nuestro. En este punto ha sido necesario entender el método y aprender a utilizar los programas e interpretar los resultados. En el apartado 6.2 comento los resultados obtenidos con esta metodología y una serie de limitaciones que hemos observado y que han motivado la propuesta de una nueva metodología bautizada como MB-MDR. Esta alternativa metodológica se detalla en el apartado 6.3. El proceso de elaboración y desarrollo del MB-MDR ha sido largo y ha requerido muchas pruebas y simulaciones que en este trabajo aparecen tan solo resumidas y condensadas. Parte del trabajo realizado en esta metodología ha consistido en su implementación en R, los detalles aparecen en la sección 8.2, y parte ha contribuido a la publicación de dos artículos (Calle et al. 2007 y 2008), uno como documento de investigación y el otro como artículo científico.

La sección 7 presenta otros dos métodos basados en árboles, los CART y el *Random Forest*, que presentan muchas similitudes con los métodos de la sección 4, pero son de construcción completamente distinta. En el apartado 7.1 se describe el funcionamiento de ambos métodos. En el apartado 7.2 se muestran los resultados de su aplicación y, de nuevo, los problemas y limitaciones surgidas. Los puntos 7.3 y 7.4 plantean una propuesta metodológica distinta de exploración de los resultados de *Random Forest* a partir del uso de las curvas ROC y del área bajo la curva (AUC) como medida de evaluación de la capacidad predictiva del modelo y se propone un método de selección de variables a partir de esta aproximación. En este caso también ha sido necesario la programación en R de las funciones necesarias (8.3).

Finalmente, en la sección 8 aparecen detalladas todas las funciones programadas en R que han sido necesarias en el curso de este trabajo.

2. Antecedentes y motivación

2.1. Cáncer de vejiga y estudio EPICURO

El cáncer de vejiga es el quinto cáncer con mayor incidencia en los países industrializados, en España se diagnostican cerca de 8.000 nuevos casos cada año, y es uno de los que presentan una mayor prevalencia. Es una enfermedad crónica con una tasa de supervivencia a los cinco años del 70 %, que requiere de controles médicos estrictos de por vida y que tiene un impacto considerable en la calidad de vida de los pacientes. Es el tumor con un mayor coste sanitario por paciente.

Las principales causas de riesgo para este cáncer son el tabaco y la exposición de determinadas ocupaciones a agentes cancerígenos. Además se conoce que los hombres tienen un mayor riesgo de desarrollar la enfermedad que las mujeres y que este riesgo aumenta con la edad.

Sin embargo varios estudios apuntan también a causas genéticas que pueden tener una cierta influencia en la susceptibilidad al cáncer de vejiga, ya que se ha detectado un aumento del riesgo en los pacientes con antecedentes familiares de cáncer de vejiga.

El estudio de los factores genéticos relacionados con la susceptibilidad a desarrollar este cáncer se ha centrado principalmente en genes que codifican enzimas relacionadas con el metabolismo de xenobióticos (mecanismo de desactivación y eliminación de un tipo de sustancias potencialmente dañinas), pero existen otros tipos de procesos que también son de interés como el transporte xenobiótico, la apoptosis (muerte celular regulada genéticamente), el control del ciclo celular, la angiogénesis (proceso fisiológico de formación de vasos sanguíneos nuevos a partir de vasos preexistentes), la progresión tumoral o el proceso de inflamación.

El proyecto en el que he participado se engloba dentro del Estudio Español de Cáncer de Vejiga/EPICURO, que comenzó en 1997 con el propósito de avanzar en el conocimiento de este cáncer respecto a las causas genéticas y ambientales, prevención, prognosis y tratamiento, y que aglutina esfuerzos de distintos grupos de investigación. El proyecto está coordinado por Núria Malats, jefa del Grupo de Epidemiología Genética y Molecular del CNIO (Centro Nacional de Investigaciones Oncológicas) y es uno de los mayores estudios sobre cáncer de vejiga que se han realizado. Sus objetivos principales son: (a) analizar el riesgo al cáncer de vejiga en relación a factores de susceptibilidad genética, tabaco, ocupación, exposiciones ambientales, dieta, drogas e historial médico; y (b) identificar marcadores moleculares que permitan pronosticar el cáncer de vejiga.

Se trata de un estudio caso-control para el que se seleccionaron individuos de entre 18 hospitales españoles distribuidos en 5 áreas distintas (Asturias, área metropolitana

de Barcelona, Vallès Occidental/Bages, Alicante y Tenerife). Los casos son pacientes entre 21 y 80 años, diagnosticados entre 1998 y 2001, sin antecedentes de tumores en el aparato urinario y descartando los casos que fueran consecuencias secundarias de otras afecciones. Los controles se eligieron en concordancia individual con los casos respecto a la edad, en grupos de 5 años, el sexo, la raza y la región, y fueron reclutados de entre pacientes de los mismos hospitales con diagnósticos no correlacionados con los factores de interés, como el hábito de fumar.

En este trabajo nos hemos centrado en el estudio de la vía de inflamación. La muestra final consta de 1356 casos y 1271 controles, con información del genotipado de 282 SNPs (marcadores genéticos que se definen en el apartado 3.1) en un total de 108 genes involucrados en la vía de inflamación del cáncer, así como de distintas variables ambientales de exposición a factores de riesgo, entre las que destaca el hábito de fumar.

Los participantes se clasificaron en cuatro categorías según su exposición al tabaco; no fumadores si han fumado menos de 100 cigarros en toda su vida, fumadores habituales si han fumado al menos un cigarro al día durante un mínimo de 6 meses, exfumadores si cumplen las especificaciones de fumadores habituales pero hace más de un año que no fuman, y fumadores ocasionales para el resto.

Recientemente, se ha aumentado el número de SNPs genotipados para este estudio a unos 1500 SNPs repartidos en 386 genes y la tecnología permitirá en poco tiempo, genotipar cantidades cada vez mayores de SNPs a un precio asumible, y por tanto se hace necesario desarrollar, adaptar y poner a punto métodos y técnicas estadísticas que permitan manejar y extraer conocimiento de este tipo de datos.

Así pues, este trabajo debe representar un punto de partida para continuar la tarea de investigación de metodologías que permitan la extracción de conocimiento a partir de datos genéticos a gran escala, y debe entenderse, no sólo como la producción de resultados en un estudio concreto, sino como una contribución más, con perspectivas de continuidad, a éste propósito.

2.2. Descripción de los datos

La base de datos sobre la que hemos trabajando ha ido sufriendo pequeños cambios a lo largo de este período, en los que algunas de las variables (SNPs) han sido substituidas por otras o eliminadas del estudio por no estar en equilibrio de Hardy-Weinberg (propiedad sobre la distribución estadística de alelos en un locus). En la versión final de los datos con los que hemos trabajado hay 2314 individuos, 1157 casos y 1157 controles, de los que se tiene información, además de si presentan o no la enfermedad, de la edad, el sexo, la región, la exposición al tabaco y 267 SNPs.

Cada SNP es una variable con tres posibles categorías codificadas como 0, 1 ó 2 que se corresponden con los genotipos homocigoto dominante, heterocigoto i homocigoto recesivo, respectivamente.
La base de datos presenta un 7.8 % de valores faltantes, concentrados prácticamente en un 7.9 % de los individuos, para los cuales no se tiene información de genotipo en más del 85 % de SNPs.

La tabla 1 describe la distribución de individuos respecto a las variables sexo, región, edad y exposición al tabaco. Las tres primeras, son variables para las que se ha realizado el emparejamiento entre casos y controles, no así para el caso del tabaco. Se corrobora que los casos de cáncer de vejiga en hombres son muy superiores que en mujeres.

Tabla 1: Distribución de individuos respecto a las variables sexo, región, edad y exposición al tabaco

	Característica	Casos	Controles	Total
Sexo				
	Hombres	87.4 %	87.2 %	87.3 %
	Mujeres	12.6 %	12.8 %	12.7 %
Edad				
	< 55	14.9 %	16.8 %	15.8 %
	55-64	20.9 %	24.4 %	22.6 %
	65-69	22.0 %	23.0 %	22.5 %
	70-74	22.0 %	19.8 %	20.9 %
	≥ 75	20.2 %	16.1 %	18.2 %
Región				
	Barcelona	18.1 %	20.1 %	19.1 %
	Vallès Occidental/Bages	15.6 %	15.7 %	15.6 %
	Alicante	7.3 %	7.1 %	7.2 %
	Tenerife	18.5 %	16.7 %	17.6 %
	Asturias	40.5 %	40.4 %	40.5 %
Tabaco				
	No fumador	13.8 %	29.4 %	21.6 %
	Ocasional	4.3 %	7.7 %	6.0 %
	Exfumador	38.9 %	37.2 %	38.1 %
	Habitual	43.0 %	25.7 %	34.3 %

La tabla 2 muestra la distribución de casos y controles según la exposición al tabaco, que es el principal factor de riesgo asociado al cáncer de vejiga. Se observa una ratio de aproximadamente el doble de casos que de controles entre los fumadores y que esta ratio se invierte entre los no fumadores y los fumadores ocasionales.

Tabla 2: Distribución de casos y controles según la exposición al tabaco

Categoría	Casos	Controles
No fumador	32.0 %	68.0 %
Ocasional	36.2 %	63.8 %
Exfumador	51.1 %	48.9 %
Habitual	62.6 %	37.4 %

En los resultados expuestos en este trabajo se han agrupado las categorías de no fumadores y fumadores ocasionales debido a la baja proporción de la categoría ocasionales y a las similitudes detectadas al estratificar.

3. Peculiaridades de los datos genéticos

3.1. Conceptos importantes

Los humanos compartimos el 99 % del ADN y, sin embargo, existen millones de diferencias entre el ADN de dos personas. Estas variaciones genéticas pueden ser de diferente tipo y naturaleza, y pueden afectar a la susceptibilidad de una persona a desarrollar algunas enfermedades o afectar a la respuesta a una determinada terapia. Por ello es importante, una vez que se tienen evidencias de una cierta componente genética en la causa o evolución de una enfermedad, conocer qué variaciones genéticas están asociadas con esa enfermedad.

El ADN está formado por una secuencia muy larga de bases, llamadas nucleótidos, que pueden ser de 4 tipos; adenina (A), citosina (C), guanina (G) y timina (T). Hay partes de esta secuencia, denominadas genes, que son las responsables de la codificación de proteínas, de cuya presencia y actividad dependen prácticamente todos los procesos biológicos. Se estima que tenemos entre 30.000 y 40.000 genes.

Frente a las enfermedades Mendelianas, cuya causa principal se debe a mutaciones en un determinado gen, y cuya detección es relativamente sencilla, existe otro tipo de enfermedades comunes causadas tanto por factores genéticos, con varios genes involucrados, como factores ambientales, y que se denominan enfermedades complejas. Algunos ejemplos de enfermedades complejas son la diabetes, la obesidad, enfermedades cardiovasculares o el cáncer. Una de las características de las enfermedades complejas es que el factor genético no viene determinado por un único gen sino por varios y, normalmente, por interacciones complejas de muchos genes.

Existen varios tipos de estudios cuya finalidad es detectar los genes asociados a una determinada enfermedad compleja como, por ejemplo, estudios de asociación, estudios basados en familias, análisis de ligamiento o análisis de haplotipos. Para detectar estos genes se requiere poder detectar su presencia y posición en el genoma, y aquí es donde intervienen los marcadores genéticos, o también denominados polimorfismos, que son posiciones del genoma de las que se conoce su localización y que presentan cierta variabilidad entre los individuos.

El tipo de polimorfismo o marcador genético más común es el que se denomina SNP (Single Nucleotide Polymorphism). Los SNPs, pronunciado "snips", son la forma de variación genética más abundante en el genoma humano y consisten en variaciones en una única base, es decir, una de las bases está substituida por otra. Actualmente, se han identificado más de 10 millones de SNPs. Como una base es substituida por otra, en teoría puede haber 4 variaciones distintas, pero a nivel poblacional se suelen observar solamente dos variantes.

El genoma humano es diploide, es decir, cada célula contiene dos copias de la cadena

de ADN, y esto hace que el genotipo en una posición concreta del genoma conste de dos elementos, correspondientes a cada una de las dos copias de ADN situadas en cromosomas homólogos. Así pues, un SNP viene determinado por dos bases, que se denominan alelos. El alelo más frecuente en la población se denomina mayoritario y lo denotaremos como A, y el menos frecuente se denomina minoritario y lo denotaremos como a. Esto hace que para cada SNP pueda darse tres combinaciones posibles, dando lugar a tres genotipos distintos: que las dos bases sean iguales y correspondan al alelo mayoritario, AA, llamado homocigoto dominante, que sean distintas, Aa, llamado heterocigoto, o que ambas correspondan al alelo minoritario, aa, llamado homocigoto recesivo.

La importancia de los SNPs radica en que son muy frecuentes y están distribuidos por todo el genoma. Existen tres formas en las que un polimorfismo puede estar asociado con una enfermedad:

- *Asociación directa.* El polimorfismo tiene un cierto efecto causal directo con la enfermedad debido a que está situado en una región codificante del genoma y, por tanto, produce cambios en la generación de aminoácidos, que son la base de las proteínas. Aunque este tipo de asociación es la más fácil de analizar, la dificultad radica en la identificación de los polimorfismos candidatos, puesto que no se sabe a priori que SNPs producen este efecto. Además, parece que muchas variaciones genéticas asociadas a enfermedades complejas, son diferencias en zonas no codificantes, por lo que un estudio de asociación directa sólo podría descubrir una parte de las causas genéticas relacionadas con la enfermedad.

- *Asociación indirecta.* El polimorfismo actúa como marcador de una región cercana a él que puede contener un polimorfismo causal. Sin embargo, estos estudios son más difíciles de analizar y tienen menor potencia que los estudios directos.

- *Confusión.* La asociación es debida a una estratificación subyacente de la población o mezcla de poblaciones. En una mixtura de poblaciones con distintas exposiciones ambientales o distinta susceptibilidad genética, zonas con frecuencias alélicas diferentes entre las poblaciones presentarán un cierto grado de asociación con la enfermedad. La presencia de confusión puede causar la detección de falsa asociación o no permitir la detección de una causa real.

Por otra parte, como un SNP tiene tres posibles genotipos, la relación entre un SNP y un determinado fenotipo puede analizarse bajo diferentes modelos de herencia. Si denotamos por $P(Y|G)$ la distribución de probabilidad de un determinado fenotipo Y ($Y = 1$ para afectados e $Y = 0$ para no afectados) dado el genotipo G, llamada función de penetrancia, los modelos de herencia más habituales son:

- *Modelo Dominante.* Establece dos grupos de comparación, los homocigotos dominantes contra el resto, juntando los heterocigotos con los homocigotos

recesivos. Este modelo supone que basta tener una copia del alelo minoritario para modificar el riesgo, y que tener una segunda copia del alelo minoritario no lo modifica. Es decir, $P(Y = 1|Aa) = P(Y = 1|aa)$.

- *Modelo Recesivo.* Se compara el grupo formado por los homocigotos dominantes y los heterocigotos, contra el grupo de homocigotos recesivos. La suposición en este modelo es que es necesario tener los dos alelos minoritarios para influir en el riesgo, esto es, $P(Y = 1|Aa) = P(Y = 1|AA)$.
- *Modelo Aditivo.* Asume que el riesgo se modifica de forma aditiva por cada copia del alelo minoritario, $P(Y = 1|aa) = 2 \cdot P(Y = 1|Aa)$
- *Modelo Codominante.* Los tres grupos se consideran separadamente, suponiendo que cada genotipo proporciona un riesgo distinto y no aditivo. Este modelo es el más general y $P(Y = 1|Aa) \neq P(Y = 1|Aa) \neq P(Y = 1|aa)$.

A priori no es fácil establecer un criterio para determinar el modelo de herencia más adecuado para un polimorfismo.

3.2. Problema de la dimensionalidad

En los últimos años, gracias al avance de las técnicas para la secuenciación del genoma y los microarrays de expresión génica, se ha producido una gran explosión de la cantidad de información genética disponible, haciendo posibles nuevos tipos de análisis, pero generando también nuevos problemas y retos tanto metodológicos como computacionales.

Uno de los nuevos retos planteados es hacer frente a la alta dimensionalidad de los datos. En muchos estudios con datos genéticos se trabaja con un número muy elevado de marcadores genéticos, produciendo un espacio de datos a examinar con una dimensión tan elevada que se hace computacionalmente imposible examinarlo de forma exhaustiva, requiriendo de algoritmos de búsqueda efectivos. Muchos de estos algoritmos pierden potencia cuando la dimensión del espacio a explorar se hace extremadamente grande y por otro lado, la combinación de exploraciones parciales no proporciona un conocimiento real del conjunto.

Esto hace que se requiera del uso y desarrollo de técnicas de *data mining* para poder abordar la exploración de espacios de dimensión extrema o para poder reducir esta dimensión a valores más tratables.

El gran tamaño de los datos con los que se trabaja en este contexto hace que exista también una gran cantidad de ruido o información irrelevante que dificulta la detección de señales biológicas significativas, que muchas veces se observan mezcladas con factores aleatorios o confusores, incorporando sesgo en las estimaciones.

Otro problema que se plantea es el insuficiente tamaño muestral disponible frente al número de parámetros a estimar. Los métodos estadísticos convencionales requieren de un número de observaciones independientes mucho más grande que el número de parámetros a estimar. En muchos estudios genéticos, el número de factores a considerar simultáneamente se hace tan elevado que se invierte esta relación, existiendo muchos más parámetros a estimar que observaciones disponibles.

El problema de la dimensionalidad se hace especialmente complejo cuando el objetivo es la exploración de interacciones. En muchas enfermedades complejas, las interacciones entre genes y las interacciones genes-ambiente tienen un papel importante, y muchos estudios se centran en detectar qué combinaciones de SNPs aumentan o disminuyen la susceptibilidad o el riesgo a desarrollar estas enfermedades.

3.3. Resumen de las técnicas de *data mining*

Para abordar el problema de la alta dimensionalidad de los datos genéticos existe una amplia variedad de métodos que pueden clasificarse en tres categorías según la estrategia utilizada para afrontar el problema: (a) métodos de reducción de la dimensión, (b) métodos de selección de variables y (c) métodos capaces de analizar un gran número de variables directamente. Esta clasificación no es excluyente, tanto en el sentido que un mismo algoritmo o método puede adaptarse a dos estrategias distintas, como en el sentido que una metodología puede incluir la combinación de más de una de estas estrategias. Son formas complementarias de abordar el problema y en muchos casos es necesario recurrir a más de una estrategia. Por ejemplo, utilizar un método de selección como paso previo a la utilización de otro tipo de método puede hacer aumentar la potencia de este último. A continuación se detallan las características generales de cada una de estas estrategias.

La primera estrategia, ***reducción de la dimensión***, consiste en realizar transformaciones de los datos que permitan reducir la dimensión del espacio de valores a considerar, obteniendo un conjunto de nuevos datos, de tamaño o dimensión menor, que concentre o resuma la información relevante del conjunto de datos original. Puede ser una estrategia muy útil en conjuntos de datos con información redundante.

El objetivo de la segunda estrategia, ***selección de variables***, es seleccionar un subconjunto más pequeño de variables, las más relevantes, y centrar el estudio posterior en ese subconjunto. Para hacer esta selección de variables existen tanto métodos univariantes como multivariantes. Los métodos univariantes analizan cada variable individualmente para establecer un ranking según un criterio de asociación con la variable respuesta y finalmente seleccionar las mejores variables. El problema de los métodos univariantes es que no tienen en cuenta las correlaciones ni las interacciones entre variables, por lo que la selección de las mejores variables individuales no

garantiza obtener el mejor subconjunto. Por contra, los métodos multivariantes se basan en determinar el subgrupo óptimo de variables cuando el espacio de combinaciones posibles es computacionalmente imposible de analizar exhaustivamente. Los métodos multivariantes tratan las posibles correlaciones entre las variables, pero no todos tratan las interacciones. Los principales escollos de los métodos multivariantes son el tiempo computacional, la robustez y la sobre estimación.

Los métodos multivariantes de selección de variables se caracterizan por dos aspectos: el algoritmo de búsqueda en el espacio de posibles subconjuntos y el criterio de evaluación de cada subconjunto considerado. Según estas características hay tres tipos de métodos:

- *Filter.* No incluyen aprendizaje, es decir, no se basan en ningún modelo de clasificación o regresión, sino que utilizan criterios de evaluación independientes como, por ejemplo, la distancia de Mahalanobis.

- *Wrapper.* Incluyen aprendizaje en la selección de variables usando como criterio de evaluación las estimaciones del error de clasificación (o predicción) basadas en un determinado modelo de clasificación (o regresión) pero sin incluir conocimiento de la estructura del modelo en el criterio de evaluación. Un ejemplo de este tipo de modelos es la eliminación recursiva de variables.

- *Embedded.* La parte de aprendizaje y de selección son inseparables. Por ejemplo, en los modelos de clasificación basados en árboles, la selección de las variables forma parte del propio modelo de clasificación.

La tercera estrategia, ***análisis en alta dimensión***, consiste en utilizar métodos que sean capaces de analizar directamente conjuntos de datos en los que el número de variables supera el de observaciones. Hay dos tipos de métodos que utilizan esta estrategia: métodos clásicos basados en penalización y métodos provenientes de la teoría del aprendizaje automático.

La figura 1 presenta un resumen de las características expuestas en esta sección para algunas de las metodologías existentes. Me he centrado exclusivamente en aquellas metodologías que son aplicables a las características específicas de nuestros datos, esto es, que acepten como variables categóricas tanto la respuesta como las variables predictoras. Aún así, no se trata de una lista exhaustiva de todas las técnicas disponibles. También he incluido otro tipo de características que son de interés en nuestro estudio, como la posibilidad de detectar interacciones asociadas con la respuesta, ajustar por variables confusoras y efectos marginales y tratamiento de datos faltantes.

Algunas referencias importantes en este contexto son:

- Hastie T, Tibshirani R, Friedman J *The Elements of Statistical Learning. Data Mining, Inference and Prediction* New York: Springer-Verlag, 2001

- Ripley Brian D *Pattern recognition and neural networks* Cambridge University Press, 1996

Figura 1: Listado de algunas metodologías aplicables a nuestros datos junto con un resumen de las características más importantes.

	Reducir Dimensión	Selección SNPs	Análisis Alta Dimensión	Detección Interacciones	Ajuste Confusion	Ajuste Efectos Marginales	Tratamiento Missings
Regresión Logística				✔	✔	✔	
Regresión Logística Penalizada			✔	✔	✔	✔	
Medidas de Información		✔		✔			
Logic Regression		✔	✔	✔			
Logic Feature Selection		✔	✔	✔			
MDR	✔		✔	✔			
MB-MDR	✔		✔	✔	✔	✔	
CART		✔	✔	✔			✔
Random Forest		✔	✔	✔			✔
Clustering		✔	✔				
Algoritmos genéticos		✔	✔	✔			
Suport Vector Machines			✔	✔			
Neural Network Models			✔				
Bayesian Networks			✔				✔

3.4. Corrección de falsos positivos

Uno de los problemas a que uno se enfrenta en estudios en los que intervienen y hay que analizar muchos factores y sus interacciones, es el estudio de la significación de los resultados obtenidos. Este tipo de estudios conlleva la realización de multitud de pruebas de hipótesis y, como consecuencia, la probabilidad de obtener un falso positivo se incrementa muy por encima del nivel de significación (normalmente 0.05). Esto hace necesario el uso de métodos para controlar la probabilidad de detectar falsos positivos. La tabla 3 introduce la notación que utilizaré en adelante.

Los métodos clásicos, denominados FWE (familywise error-rate), controlan la probabilidad de obtener algún falso positivo, $P(F \geq 1)$. El más conocido es el ajuste de Bonferroni, que garantiza que $P(F \geq 1) \leq \alpha$ tomando niveles de significación de α/m para cada prueba. Pero estos métodos son demasiado conservadores y cuando

Tabla 3: Notación para el resultado de hacer m pruebas de hipótesis.

	Detectadas Significativas	Detectadas No Significativas	Total
Hipótesis nula cierta	F	$m_0 - F$	m_0
Hipótesis alternativa cierta	T	$m_1 - T$	m_1
Total	S	$m - S$	m

el número de pruebas es extremadamente grande, como sucede en los estudios con datos genéticos a gran escala, esto hace que no se detecte ningún resultado significativo, más aún cuando el efecto que se está buscado es moderado o pequeño.

Como alternativa a los métodos FWE existe un punto de vista diferente, denominado FDR (false discovery rate), que consiste en controlar la proporción de falsos positivos entre los positivos detectados, F/S. La idea intuitiva detrás de esta aproximación es que no tiene la misma implicación detectar 4 resultados significativos de los cuales 2 sean falsos positivos (50 % de falsos positivos), que detectar el mismo número de falsos positivos de entre 100 resultados significativos (2 % de falsos positivos). Mientras que, por ejemplo, el método de Bonferroni aplicaría la misma corrección en ambos casos, multiplicando los p-valores obtenidos por el número de pruebas realizadas, estos métodos alternativos son más estrictos al ajustar los p-valores obtenidos en el primer caso que en el ajuste en el segundo caso.

Dada una colección de pruebas de hipótesis, se define el FDR como la proporción esperada de pruebas erróneamente detectadas como significativas de entre todas las detectadas como significativas, $FDR = E[\frac{F}{S}]$.

El procedimiento más popular para controlar el FDR es el método BH (Benjamini et Hochberg 1995). Dada una colección de p-valores ordenados de forma creciente, $p_{(1)} \leq \ldots \leq p_{(m)}$, correspondientes a sendas pruebas con hipótesis nulas $\{H_{(i)}\}_i$, para un nivel q de FDR determinado se compara cada p-valor $p_{(i)}$ con el valor crítico $q \cdot \frac{i}{m}$ y se rechazan $H_{(i)}, ..., H_{(k)}$, donde $k = \max\{i : \ p_{(i)} \leq q \cdot \frac{i}{m}\}$. Este método controla el FDR al nivel q en el sentido que:

$$FDR \leq q \cdot \frac{m_0}{m} \leq q$$

Así pues, el método BH garantiza que la proporción esperada de falsos positivos entre los p-valores aceptados como significativos es menor que el valor q especificado. Este método es válido bajo la suposición de independencia de los múltiples tests o de correlación positiva, y la forma de obtener unos p-valores ajustados por el método BH es a partir de la expresión:

$$q_{BH}\left(p_{(i)}\right) = \min_{k \geq i}\{p_{(k)} \cdot \frac{m}{k}\}$$

Storey y Tibshirani (2003) proponen una alternativa que modifica el método BH ajustando los p-valores como:

$$q\left(p_{(i)}\right) = \min_{k \geq i}\{p_{(k)} \cdot \hat{\pi_0} \cdot \frac{m}{k}\}$$

donde $\hat{\pi_0}$ es una estimación de la proporción de hipótesis nulas ciertas, $\hat{\pi_0} = m_0/m$.

4. *Logic Regression / Logic Feature Selection*

4.1. Descripción del método *Logic Regression*

Logic Regression (Ruczinski et al. 2003) es una metodología para identificar combinaciones lógicas de las variables que mejor predicen la variable respuesta según un determinado modelo de regresión. Está pensada para situaciones en las que las variables predictoras son binarias y se puede aplicar al estudio de interacciones de SNPs, puesto que estos se pueden recodificar fácilmente en dos variables *dummy* binarias (X_{i1}, X_{i2}): siendo $(1,0)$ si el SNP tiene sólo un alelo minoritario, $(0,1)$ si los dos alelos son minoritarios, o $(0,0)$ para el caso en que los dos alelos son mayoritarios.

Más concretamente, este método explora modelos de regresión del tipo:

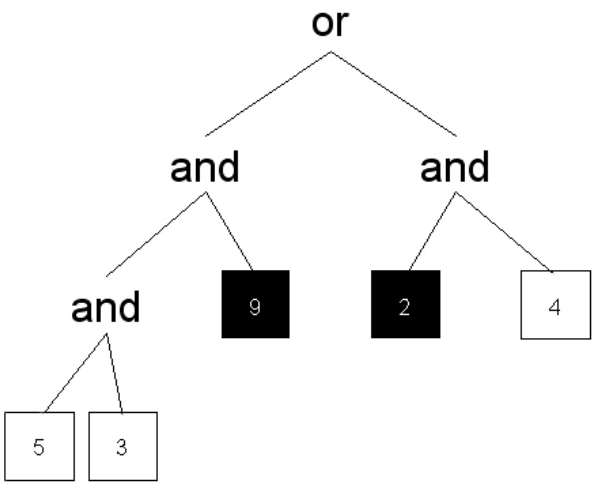
$$g(E(Y)) = \beta_0 + \sum_{j=1}^{t} \beta_j \cdot I_{\{L_j \ es \ cierta\}}$$

donde Y es la variable respuesta, g es una función adecuada y L_j es una expresión lógica de covariables. Un ejemplo de expresión lógica es $L_j = (X_1 \vee X_2) \wedge X_3^c$, donde las variables tienen dos valores posibles, 1 ó 0, indicando si se da o no un determinado factor, en nuestro caso un determinado genotipo, siendo 1 que sí y 0 que no. Estos valores se traducen como verdadero y falso respectivamente, y la evaluación de la expresión L_j tiene dos resultados posibles, que sea cierta, si se cumple que X_3 vale 0 y que X_1 o X_2 valen 1, o que sea falsa, si $X_3 = 1$ o $X_1 = X_2 = 0$.
Estas expresiones lógicas pueden representarse como una estructura en forma de árbol, como muestra la figura 2.

El objetivo del método es encontrar el mejor modelo de regresión, es decir, determinar los L_j y las β_j que permiten una mayor capacidad predictiva. Para ello se utiliza un algoritmo de exploración que permite no tener que comprobar todas las posibles expresiones lógicas, puesto que el espacio de combinaciones lógicas posibles crece de forma exponencial a medida que aumenta el número de variables, y se hace computacionalmente inexplorable en su totalidad.

El *Logic Regression* está implementado como un método iterativo en el que cada estado viene dado por un modelo de regresión, en el que los predictores son expresiones lógicas o también llamados árboles lógicos. La idea básica de este método es realizar, a cada iteración, un cambio en alguno de los árboles lógicos del modelo actual, de forma que la capacidad predictiva del nuevo modelo sea mejor que la del anterior, resiguiendo así (mientras sea posible) un camino desde un estado inicial a un estado final cuyo resultado es un modelo con la mejor capacidad predictiva posible. En el caso de modelos con varios árboles lógicos, a cada iteración se puede realizar un cambio en uno de los árboles, o un cambio en cada uno de ellos.

Figura 2: Representación gráfica con estructura de árbol de la expresión lógica $[(X_5 \wedge X_3) \wedge X_9^c] \vee (X_2^c \wedge X_4)]$ donde los cuadros representan las variables, en negro para indicar el complementario, y las ramas representan las relaciones según el operador indicado.



La estrategia que utiliza el *Logic Regression* se basa en tres puntos: (a) una función para evaluar la capacidad de predicción, (b) la generación de los árboles lógicos a considerar, y (c) el algoritmo de búsqueda utilizado.

(a) Para la función que tiene que evaluar la capacidad predictiva de cada modelo se proponen varias opciones según la naturaleza del modelo:

- En el caso de respuesta dicotómica y un único árbol lógico L, esta expresión lógica clasifica cada individuo a una de las clases si la condición se cumple, y a la otra si no: $Y = I_{\{L \ es \ cierta\}}$. Para este caso, se utiliza la proporción de mal clasificados como medida de evaluación.

- En el caso de modelos de regresión con respuesta continua,

$$Y = \beta_0 + \beta_1 \cdot I_{\{L_1 \ cierta\}} + \ldots + \beta_p \cdot I_{\{L_p \ cierta\}} + \epsilon$$

  con $\epsilon \sim \mathcal{N}(0, \sigma^2)$, el modelo se ajusta por el método de mínimos cuadrados, y la función que se utiliza para evaluarlo es la suma de cuadrados residuales

$$RSS = \sum_{i=1}^{n} (\tilde{Y}_i - Y_i)^2$$

  con $\tilde{Y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 \cdot I_{\{L_1 \ cierta\}} + \ldots + \tilde{\beta}_p \cdot I_{\{L_p \ cierta\}}$

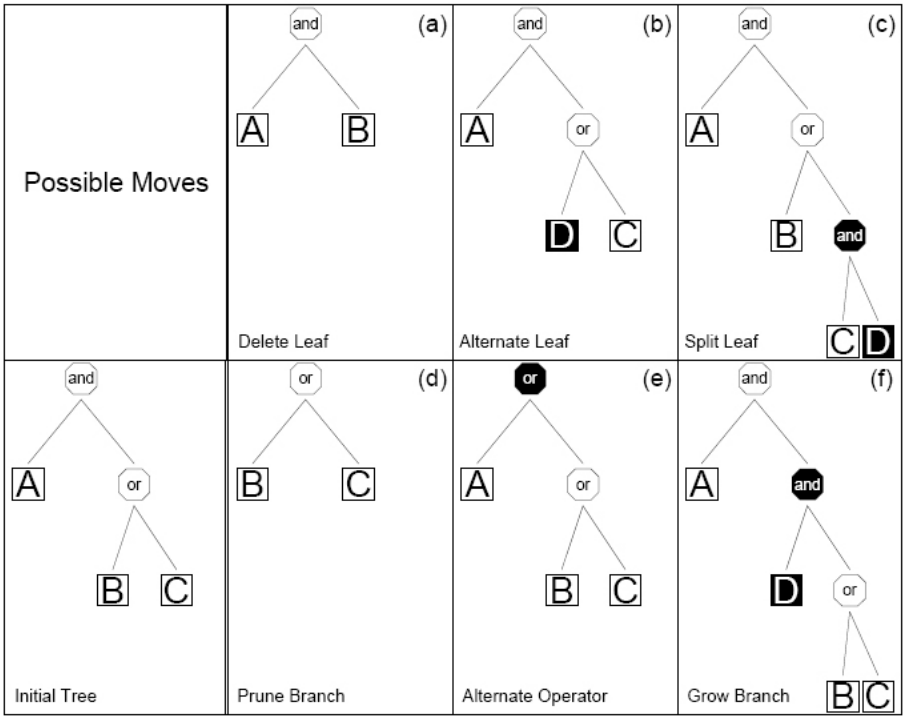
- Para modelos de regresión logística,

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \cdot I_{\{L_1 \ cierta\}} + \ldots + \beta_p \cdot I_{\{L_p \ cierta\}}$$

se utiliza la devianza.

(b) Dado un árbol concreto, se definen los árboles vecinos como elconjunto de todos los árboles generados a partir de realizar un sólo cambio en el árbol inicial. Estos cambios pueden ser cambiar, eliminar o agregar una variable, expandir o podar una rama, o intercambiar un operador $\wedge$ o $\vee$ por el otro. La figura 3 resume los cambios permitidos para generar el conjunto de árboles vecinos.

Figura 3: Esquema de los cambios utilizados para generar árboles vecinos a un árbol dado. Extraido de : Ruczinski I, Kooperberg C, LeBlanc M (2003) *Logic Regression* Journal of the Computational and Graphical Statistics, 12, 475-511.



(c) También existen varios tipos de implementación del algoritmo de búsqueda. Un primer tipo de búsqueda consiste en construir, a cada iteración, todos los árboles vecinos al actual por medio de los movimientos básicos definidos y seleccionar como nuevo árbol aquel que presente un valor menor para la función de evaluación. Este algoritmo continúa hasta que no se encuentra ningún árbol vecino que tenga un valor de la función de evaluación menor que el árbol considerado, en ese momento se detiene el proceso. Pero este proceso no garantiza encontrar el mejor árbol, y en la práctica no suele dar buenos resultados.

El segundo tipo de búsqueda se basa en el algoritmo conocido como *simulated annealing*, un algoritmo muy popular por sus buenos resultados en una amplia gama de problemas de optimización combinatoria y que consiste en seleccionar uno de los posibles cambios permitidos al azar y evaluar la capacidad predictiva del nuevo árbol. Si esta capacidad predictiva es mejor que la del árbol inicial se acepta el cambio. Si el nuevo árbol tiene una capacidad predictiva menor que la del árbol original, entonces se aceptará el cambio con una cierta probabilidad que dependerá de la disminución de la capacidad de predicción y del número de iteraciones realizadas de forma que, a medida que se van ejecutando iteraciones, esta probabilidad va disminuyendo. Este algoritmo permite evitar escollos como no poder avanzar a partir de un árbol que no mejore su capacidad de predicción al hacer un sólo cambio, pero sí después de dos o más cambios.

En el caso de modelos con varios árboles o expresiones lógicas, independientemente de si a cada iteración se efectúa un cambio en uno sólo de los árboles o en todos, el método busca el mejor modelo ajustando simultáneamente todos los parámetros. Por esta razón se requiere especificar el número de árboles que debe contener el modelo. Para que esto no suponga una limitación, se especifica un número máximo de árboles y el método busca el mejor modelo para cada nivel de complejidad, desde un árbol hasta el valor especificado como máximo, permitiendo después comparar cada modelo mediante el valor obtenido con la función de evaluación. Esta comparación también puede realizarse para otras medidas de complejidad, como el número de variables permitidas en cada árbol o el nivel de profundidad máximo permitido para los árboles.

Finalmente, se puede evaluar la capacidad predictiva del modelo a partir de una validación cruzada, y determinar la significación de los resultados por medio de un test de permutación.

4.2. Descripción del método *Logic Feature Selection*

Existe una ampliación del *Logic Regression*, el *Logic FS* (Logic Feature Selection, H. Schwender et al. 2007), que se basa en construir un número determinando de bootstraps, ajustar un modelo de regresión lógica sobre cada sub-muestra mediante *Logic Regression* y extraer de cada una de las expresiones lógicas las interacciones implicadas.

El algoritmo de *Logic Regression* trabaja con las expresiones lógicas reducidas a las llamadas formas normales disjuntas, que consisten en expresiones donde sólo intervienen operadores $\wedge$ unidas por operadores $\vee$. Por ejemplo,

$$(X_1 \wedge X_3^c \wedge X_6) \vee (X_4^c \wedge X_9) \vee (X_2^c \wedge X_5 \wedge X_6)$$

Todas las expresiones lógicas se pueden reducir a una forma normal disjunta.

Logic FS utiliza las formas normales disjuntas para extraer las interacciones implicadas, que en el ejemplo anterior serían $X_1 \wedge X_3^c \wedge X_6$, $X_4^c \wedge X_9$ y $X_2^c \wedge X_5 \wedge X_6$, para finalmente asignar un valor de importancia a cada una de estas interacciones. Esta medida de importancia se calcula en base a la diferencia de capacidad predictiva de los modelos al quitar o incluir cada una de las interacciones, evaluando esta capacidad predictiva sobre la parte de la muestra que ha quedado excluida en cada boostrap, aproximadamente una tercera parte de la muestra total. Para el caso de modelos con un solo árbol, la medida de importancia para una interacción l se calcula como:

$$VIM_l = \frac{1}{B} \cdot \left(\sum_{b:l \in L_b} (N_b - N_b^-) + \sum_{b:l \notin L_b} (N_b^+ - N_b) \right)$$

donde L_b es el conjunto de todas las interacciones identificadas en la b-ésima iteración, $b = 1, \ldots, B$, N_b es el número de observaciones *out-of-bag* clasificadas correctamente en la b-ésima iteración, y N_b^-/N_b^+ son el número de observaciones *out-of-bag* clasificadas correctamente en la b-ésima iteración después de quitar/agregar la interacción l al modelo.

4.3. Aplicación de *Logic Regression* y *Logic FS* y problemas encontrados

Los modelos basados en árboles son atractivos porque además de proporcionar predicciones sobre nuevas observaciones, permiten explorar la estructura subyacente de los datos. Sin embargo, el resultado de aplicar esta el *Logic Regression* a los datos de nuestro estudio resultó infructuoso debido a la poca robustez y variabilidad de los resultados.

La figura 4 muestra un ejemplo del resultado obtenido al realizar el ajuste de un árbol para los datos de nuestro estudio, repitiendo el proceso 10 veces sobre los mismos datos y con los mismos parámetros. Obtenemos 10 árboles muy distintos en los que la mayoría de las variables sólo aparecen en uno de ellos. Lo mismo ocurre cuando intentamos aplicar un modelo de regresión con varios árboles; distintas ejecuciones del método dan como resultado modelos con árboles muy distintos.

Debido a la poca robustez del método, pensamos que podría ser más apropiada la aproximación por *Logic FS*, ya que, mediante la técnica *bootstrap*, permite extraer combinaciones de SNPs cuya interacción tiene una cierta asociación con la enfermedad a partir de agregar la información extraida de construir muchos árboles distintos, y establece una medida de importancia de estas interacciones.

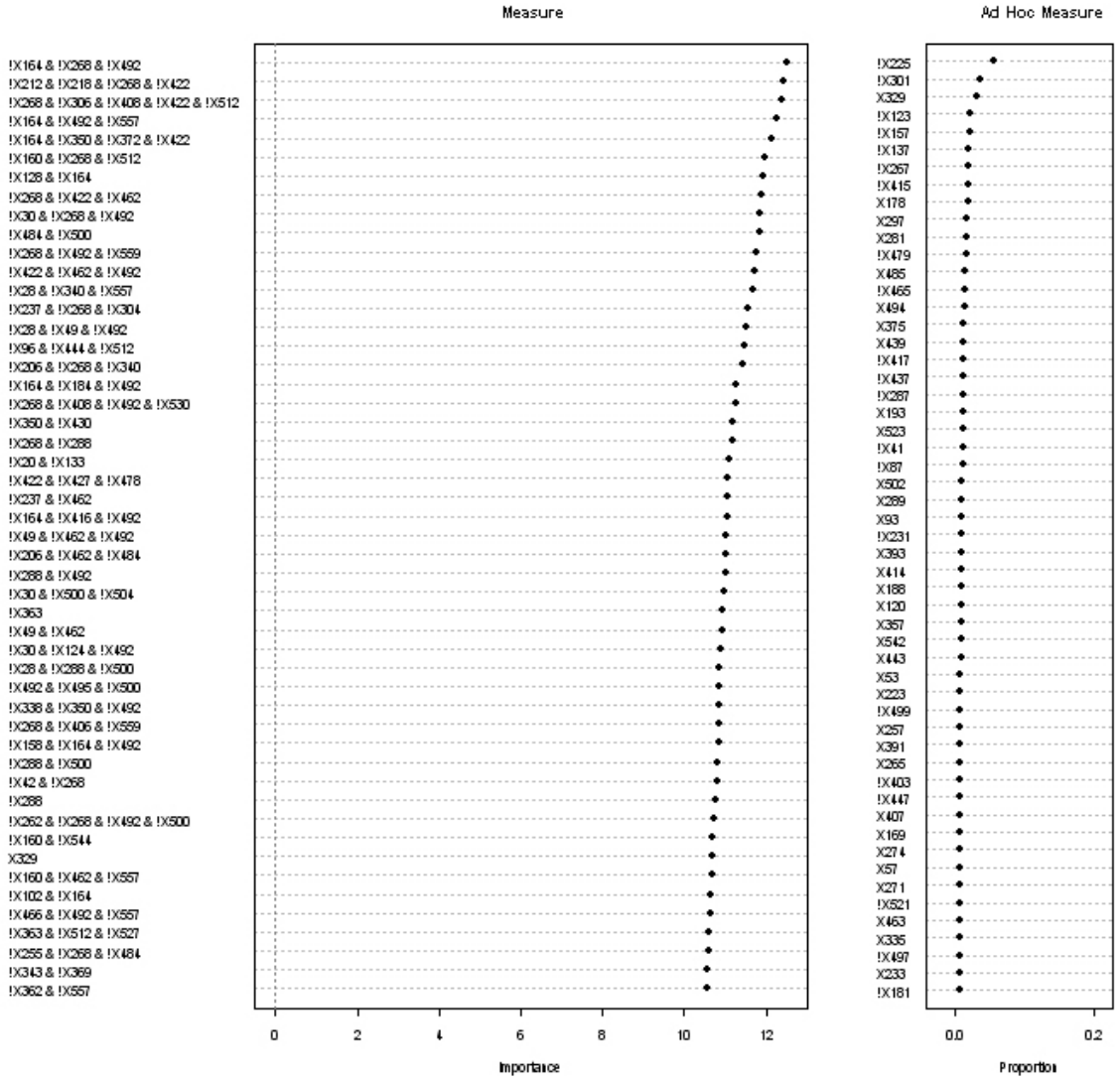
Figura 4: Resultados obtenidos al aplicar *Logic Regression* a los datos de nuestro estudio en 10 ocasiones, todas bajo los mismos parámetros de ejecución y ajustando un modelo con un solo árbol.

```
 [[1]]
 +6.34 * (((X10 and X329) and (X563 and X6)) and ((X325 or (not X724)) or (X470 and (not X983))))

[[2]]
 +6.46 * (((X64 or X930) and (X318 and X32)) and ((X62 and X91) and (X202 and (not X901))))

[[3]]
 -5.47 * (((X589 or (not X8)) or ((not X4) or (not X663))) or (((not X216) or (not X512)) or ((not
X225) and X924)))

[[4]]
 -6.26 * ((((not X19) or (not X176)) or (X293 and (not X888))) or (not X703))

[[5]]
 -6.13 * (((X859 and X49) and (X313 and (not X948))) or (((not X9) or (not X742)) or (not X4)))

[[6]]
 +6.11 * ((X374 and X498) and (((not X106) or (not X728)) and (X113 and X109)))

[[7]]
 -6.24 * (((X530 or X601) and ((not X463) or (not X9))) or (((not X7) or (not X1)) or ((not X52) or
(not X150))))

[[8]]
 -6.29 * ((((not X75) or X615) or ((not X3) or (not X8))) or (((not X41) or (not X184)) or (not
X488)))

[[9]]
 -6.13 * ((((not X24) or X958) or ((not X406) or (not X1))) or ((X85 and (not X720)) or ((not X321)
or (not X37))))

[[10]]
 -6.24 * ((((not X911) and (not X491)) or ((not X10) or (not X49))) or (((not X85) and (not X291)) or
((not X2) or (not X3))))
```

La figura 5 muestra el resultado del *Logic FS* aplicado a nuestros datos a partir de 500 bootstraps y con 500 mil iteraciones del algoritmo *simulated annealing* en la construcción de cada árbol. En la figura se muestra la representación gráfica de las importancia para las 50 principales interacciones detectadas y, por otro lado, la representación del porcentaje de veces, en orden decreciente, que aparece cada variable en alguna de las interacciones para las 50 variables más frecuentes.

De nuevo, los resultados obtenidos son poco robustos por una doble causa. Primero, porque las proporciones en las que aparecen las variables detectadas como más importantes son muy bajas (máximo de 0.056), y segundo porque para cada nueva ejecución del método sobre los mismos datos, tanto las variables como las interacciones detectadas como más importantes son completamente distintas. Además, los resultados de la figura 5 se obtuvieron a partir de modelos con un único árbol y el valor máximo de importancia alcanzado es 12.53, lo que significa que en promedio, para la mejor de las interacciones detectadas, la diferencia de individuos bien clasificados entre considerar o no la interacción es de 12.53. Teniendo en cuenta que este valor se calcula sobre el *out-of-bag*, que es aproximadamente 1/3 de la muestra ($2314/3 = 771.3$), representa un $12.53/771.3 = 1.6\,\%$ de los individuos clasificados.

Figura 5: Resultados de importancia de las 50 mejores interacciones detectadas por *Logic FS* y porcentaje de veces que aparece cada variable en alguna de las interacciones detectadas para las 50 variables más frecuentes.



La implementación existente en R de este método no permite realizar un test de significación de los resultados, ni hemos calculado una distribución nula de la medida de importancia, pero los valores observados y la poca robustez apuntan a una falta de significación de los resultados, que podría ser debida a una insuficiente cantidad de iteraciones del algoritmo de búsqueda, el *simulated annealing*, o un número insuficiente de bootstraps, pero estos parámetros los hemos incrementado hasta los límites de la capacidad computacional de que disponíamos, llegando a efectuar hasta 1 millón de iteraciones del algoritmo de búsqueda, y repitiéndolo sobre 1000 submuestras. El problema probablemente radica en la dimensión del espacio a explorar. Hay que tener en cuenta que para cada SNP se necesitan dos variables, y el número de variables se duplica al considerar las variables complementarias, es decir, si en nuestros datos tenemos 267 SNPs el *Logic FS* trabaja con 1068 variables. Dado este número de variables de partida, el número de modelos posibles se hace prácticamente inexplorable. Esto nos hace pensar que, en la práctica, esta metodología solo será factible para un número reducido de variables. En nuestro caso esto sería posible con la utilización previa de un método de selección o filtrado de variables.

5. Medidas de Sinergia

5.1. Descripción de entropía y sinergia

En esta sección se describe la metodología propuesta por Calle et al.(2008) basada en la utilización de medidas procedentes de la teoría de la información para la selección de un subconjunto de marcadores genéticos.

Una forma de aumentar la capacidad para detectar interacciones entre un gran número de polimorfismos y reducir considerablemente el tiempo de computación necesario sería realizar una selección inicial de polimorfismos susceptibles de formar parte de las interacciones buscadas. Pero esta es una cuestión compleja.

El objetivo es cuantificar la cantidad de información que proporciona conjuntamente un determinado grupo de genes sobre un determinado fenotipo, como es la presencia o no de un tipo de cáncer.

La medida de partida es $I(G_1, G_2, \ldots, G_n; C)$, definida como cantidad de información conjunta y que se interpretaría, en nuestro contexto, como la reducción de la incertidumbre sobre la respuesta C, en nuestro caso indicadora de la enfermedad, al conocer los valores de los genotipos $G_1, \ldots, G_n$.

Esta cantidad se mide a partir de los conceptos de entropía y entropía condicionada.

La entropía de una variable aleatoria C se define como la incertidumbre sobre C, calculada como:

$$H(C) = -\sum_c p(c)\ \log_2 p(c)$$

donde $p(.)$ representa la función de probabilidad para la variable C.

Para nuestro caso, en donde C es una variable dicotómica, esta expresión se convierte en:

$$H(C) = -p\ \log_2 p - (1-p)\ \log_2(1-p)$$

con $p = P(C = 1)$

La entropía condicionada, incertidumbre sobre C conocidos los genotipos $G_1, \ldots, G_n$, se define como:

$$H(C|G_1, \ldots, G_n) = \sum_{(g_1,\ldots,g_n)} p(g_1, \ldots, g_n)\ H(C|g_1, \ldots, g_n)$$

con

$$H(C|g_1, \ldots, g_n) = -\sum_c p(c|g_1, \ldots, g_n)\ \log_2 p(c|g_1, \ldots, g_n)$$

A partir de estas definiciones de entropía y entropía condicionada al conocimiento de ciertos factores, se llega a la formulación de la cantidad de información que cierto conjunto de genotipos aportan sobre el fenotipo como:

$$I(G_1, G_2, \ldots, G_n; C) = H(C) - H(C|G_1, \ldots, G_n)$$

Pero esta medida no discrimina entre las contribuciones marginales y la información atribuible a la propia interacción.
Si detectamos que la cantidad de información proporcionada por un conjunto de genes es mayor que la que cabría esperar por sus contribuciones individuales, esto sería un indicador de que ese aumento de la información se debe a la interacción entre algunos de los genes.

Esto es precisamente lo que se denota por sinergia. Para el caso bivariado se define como:

$$\text{syn}(G_1, G_2; C) = I(G_1, G_2; C) - [I(G_1; C) + I(G_2, C)]$$

y para el caso multivariado como:

$$\text{syn}(G_1, G_2, \ldots, G_n; C) = I(G_1, G_2, \ldots, G_n; C) - \underset{\{S_j\}_j}{\text{máx}} \sum_j I(S_j; C)$$

donde $\{S_j\}_j$ indica el conjunto de todas las posibles particiones del conjunto $S = \{G_1, G_2, \ldots, G_n\}$

Por ejemplo, para el caso $n = 3$ la sinergia sería:

$$\text{syn}(G_1, G_2, G_3; C) = I(G_1, G_2, G_3; C) - \text{máx} \begin{cases} I(G_1; C) + I(G_2; C) + I(G_3; C) \\ I(G_1; C) + I(G_2, G_3; C) \\ I(G_2; C) + I(G_1, G_3; C) \\ I(G_3; C) + I(G_1, G_2; C) \end{cases}$$

5.2. Exploración de la sinergia. Aplicación al cáncer de vejiga

Con el fin de analizar los datos mediante la medida de sinergia, he desarrollado varias funciones en R para el cálculo de la entropía, la entropía condicionada, la información mutua y finalmente la sinergía, para interacciones de orden dos y tres (sección 8.1). De momento nos hemos limitado hasta orden tres porque estamos en una fase de evaluación del potencial de estas medidas y la complejidad de la programación aumenta considerablemente si se desea hacer para cualquier orden de interacción. De todas formas, estas funciones estan programadas con vistas a poder ser utilizadas para una futura ampliación a cualquier orden, manteniendo el máximo de generalización posible y de forma que puedan ser incluidas dentro de una librería

de R.

A partir de estas rutinas hemos realizado el cálculo de la sinergia para órdenes de interacción dos y tres, tanto para la muestra completa, como estratificando por exposición de tabaco, es decir, para el grupo de no fumadores (que incluye también los fumadores ocasionales), los exfumadores y los fumadores habituales.

El resultado es una lista exhaustiva de todos los modelos de iteración entre SNPs, de orden 2 y 3, con el valor de sinergia de cada uno y ordenados de mayor a menor. Tal como he indicado anteriormente, la sinergia puede interpretarse como medida del aumento de la información sobre la respuesta atribuido a la interacción, excluyendo la información que aportan los SNPs individualmente.

Tabla 4: Sinergia de dos SNPs en el caso de los no fumadores. Se detallan la entropía, las cantidades de información individual de cada SNP, la información conjunta y la sinergia. $H(C)$ es diferente en cada caso porque no tiene en cuenta las observaciones con misings en los SNPs.

Pos.	G_1	G_1	$H(C)$	$I(G_1;C)$	$I(G_2;C)$	$I(G_1,G_2;C)$	$\mathrm{syn}(\mathrm{G}_1,\mathrm{G}_2;\mathrm{C})$
1	69	37	0.921248	0.001435	0.001313	0.033950	0.033868
2	103	75	0.915917	0.000570	0.001965	0.033280	0.033567
3	**144**	**41**	**0.921801**	**0.000104**	**0.000529**	**0.028852**	**0.030612**
4	174	154	0.918295	0.001660	0.003398	0.032793	0.030202
5	234	152	0.920031	0.000443	0.002796	0.030759	0.029910
6	77	8	0.920031	0.000363	0.000645	0.028505	0.029886
7	217	102	0.907230	0.002033	0.002662	0.031203	0.029218
8	54	28	0.921746	0.005516	0.002157	0.034234	0.028815
9	217	42	0.906897	0.001908	0.000232	0.028143	0.028672
10	175	60	0.921172	0.000698	0.001707	0.028753	0.028602
11	133	86	0.920658	0.002262	0.001286	0.029799	0.028513
12	225	215	0.918295	0.003230	0.000660	0.029892	0.028314
	⋮	⋮					⋮
10101	**261**	**108**	**0.933650**	**0.011800**	**0.023451**	**0.041883**	**0.007102**
	⋮	⋮					⋮

En la tabla 4 se muestra parte del resultado obtenido para la sinergia de dos SNPs en el caso de los no fumadores, ordenado de mayor a menor sinergia. En esta tabla se detallan la entropía (es diferente en cada caso porque está restringida a aquellas observaciones sin ningún dato faltante en los SNPs), las cantidades de información individual de cada SNP, la información conjunta y la sinergia. En la posición 3 tenemos los SNPs 144 y 41, que tienen cantidades de información muy pequeñas,

es decir aportan muy poca información respecto la respuesta y, en cambio, tienen una sinergia muy alta (comparar con la distribución de referencia de la figura 7). Este ejemplo muestra un hecho importante: el estudio de las interacciones génicas no puede restringirse a aquellas combinaciones de SNPs con un efecto marginal significativo. Esta aproximación es muy habitual en estudios clásicos de regresión, pero vemos que puede estar ignorando interacciones importantes. Por contra, mucho más abajo de la tabla, mostramos como ejemplo los SNPs 261 y 108, con cantidades de información mucho mayores pero con poca sinergia. Este ejemplo muestra cómo la cantidad de información conjunta, $I(G_1, G_2; C)$, no es una medida genuina de interacción ya que, como en este caso, puede estar reflejando únicamente efectos marginales.

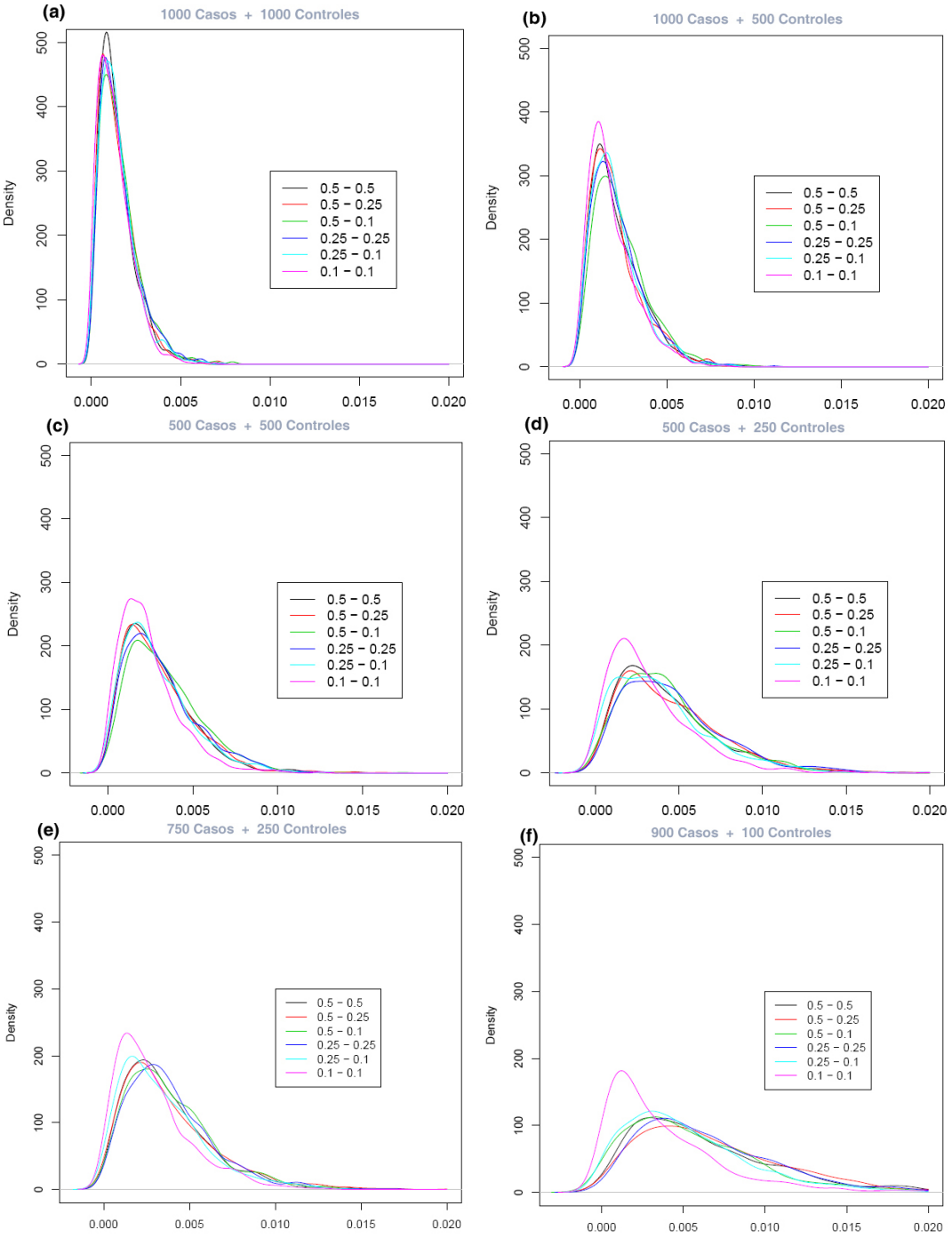
5.3. Estudio de la significación

Para el estudio de la significación, a falta de conocimientos teóricos sobre la función de distribución de esta medida, fué necesario generar una distribución nula, es decir, determinar la función de distribución empírica bajo la hipótesis de no asociación.

Esta distribución la hubiéramos podido generar a partir de un test permutacional, es decir a partir del cálculo de las sinergias en el mismo conjunto de datos de que disponemos en base a multitud de permutaciones aleatorias de la respuesta. Sin embargo, esto significa que para cada nuevo conjunto de datos que quisiéramos analizar deberíamos repetir el proceso de generar la distribución nula, con el coste computacional que ello representa. Así pues, con la intención de tener un conocimiento más general de cuál es el comportamiento de esta medida bajo la hipótesis nula de no asociación y de poder obtener una distribución de referencia más general, hemos calculado empíricamente la distribución nula de la sinergia entre dos SNPs a partir de simulaciones de diferentes escenarios, teniendo en cuenta la proporción de casos y controles, el tamaño muestral y las frecuencias genotípicas de los SNPs. A partir de los resultados en las simulaciones hemos estimado las funciones de densidad de forma no paramétrica y los resultados se muestran en la figura 6.

Como a priori no conocíamos el efecto que pueden tener sobre la sinergia las frecuencias de los genotipos de cada SNP, hemos calculado la distribución de la sinergia para distintas combinaciones de estas frecuencias. Estas frecuencias quedan determinadas a partir de las frecuencias alélicas y de la ley de equilibrio de Hardy-Weinberg, que establece la relación entre las frecuencias genotípicas y las alélicas de un SNP, y las frecuencias alélicas quedan determinadas por la frecuencia de uno de los dos alelos. Así pues, hemos considerado SNPs con tres posibles valores para la frecuencia del alelo minoritario, 0.5, 0.25 y 0.1, y hemos calculado la distribución para las 6 combinaciones resultantes de sinergia entre dos SNPs. Por ejemplo, las densidades en negro, correspondientes a la etiqueta "0.5-0.5", representan la función de densidad cuando ambos SNPs tienen la mismas frecuancias alélicas e iguales a 0.5, las

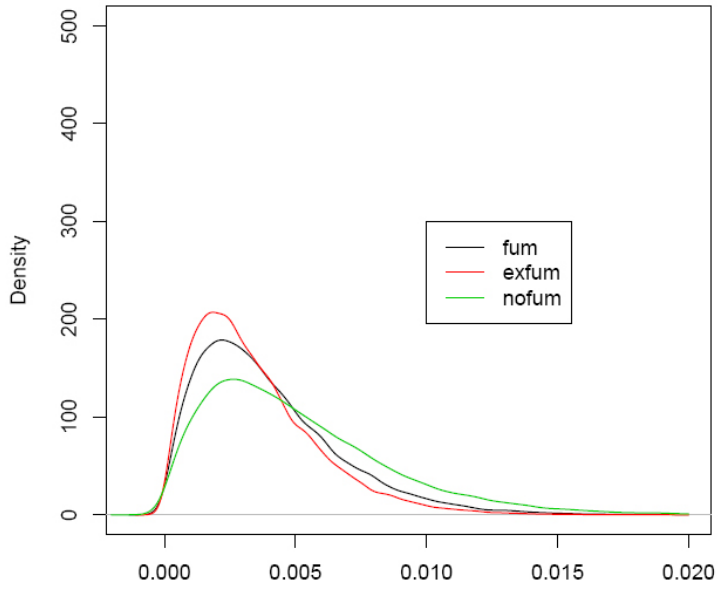
Figura 6: Densidades empíricas de la sinergia de dos SNPs bajo la hipótesis nula de no asociación con la respuesta para distintos escenarios simulados.



densidades en azul claro corresponden al caso en que un SNP tiene una frecuencia de 0.25 para el alelo minoritario y el otro SNP tiene una frecuencia de 0.1 para el alelo minoritario, etc.

Las frecuencias de los alelos parecen no influir en la distribución, sólo cuando los dos SNPs tienen frecuencias muy pequeñas para uno de los alelos la función densidad se diferencia del resto, más cuanto menor es el tamaño muestral. En cambio las proporciones entre casos y controles sí que afectan, como se aprecia en los gráficos (c), (e) y (f), todos generados con 1000 individuos, pero con distinta distribución de casos y controles. Por otra parte, el resultado de intercambiar casos por controles ha der ser necesariamente el mismo, ya que en la práctica se traduce en cambiar la codificación de una variable categórica.

Figura 7: Densidades empíricas de la sinergia de dos SNPs bajo la hipótesis nula de no asociación con la respuesta para cada subgrupo de exposición a tabaco.



Viendo que las distribuciones de referencia no son invariantes, ya que dependen del tamaño muestral y de la proporción de casos y controles, el estudio de la significación en nuestros datos ha requerido la generación de distribuciones específicas a partir de datos simulados pero con el mismo número de casos y controles que en los distintos grupos según la exposición al tabaco; fumadores, exfumadores y no fumadores. Los resultados aparecen en la figura 7. Para determinar la significación de los datos de sinergia obtenidos en los datos reales, calculo la estimación del p-valor para cada valor de sinergia S a partir de los valores de la distribución empírica correspondiente, como la proporción de valores de la distribución empírica que son superiores a

S.

Una vez calculados los p-valores y ajustado mediante FDR para corregir el hecho de estar haciendo múltiples test de hipótesis, el resultado es que no obtenemos valores significativos. Otra vez nos encontramos con el problema de la dimensionalidad: el número de pruebas que estamos realizando es demasiado elevado incluso para el FDR. Así pues, ante la imposibilidad de obtener datos significativos, finalmente optamos por hacer una selección de los 100 pares de SNPs con mayor sinergia para un posterior análisis mediante la metodología que exponemos a continuación (MB-MDR)

6. MDR / MB-MDR

6.1. Descripción del método MDR

MDR (Multifactor Dimensionality Reduction) es una técnica de data mining (Ritchie et al. 2001) para reducir la dimensionalidad en el estudio de las interacciones entre SNPs. Es un método no paramétrico y que no supone ningún modelo de herencia concreto.

Como cada SNP tiene tres genotipos posibles, al intentar analizar las interacciones entre dos SNPs (normalmente bajo el modelo codominante) resulta que hay $3^2 = 9$ posibles combinaciones de genotipos, 27 si analizamos interacciones de tres SNPs, y así sucesivamente. Para poder estudiar estas interacciones con los métodos tradicionales se requieren tamaños muestrales demasiado grandes.

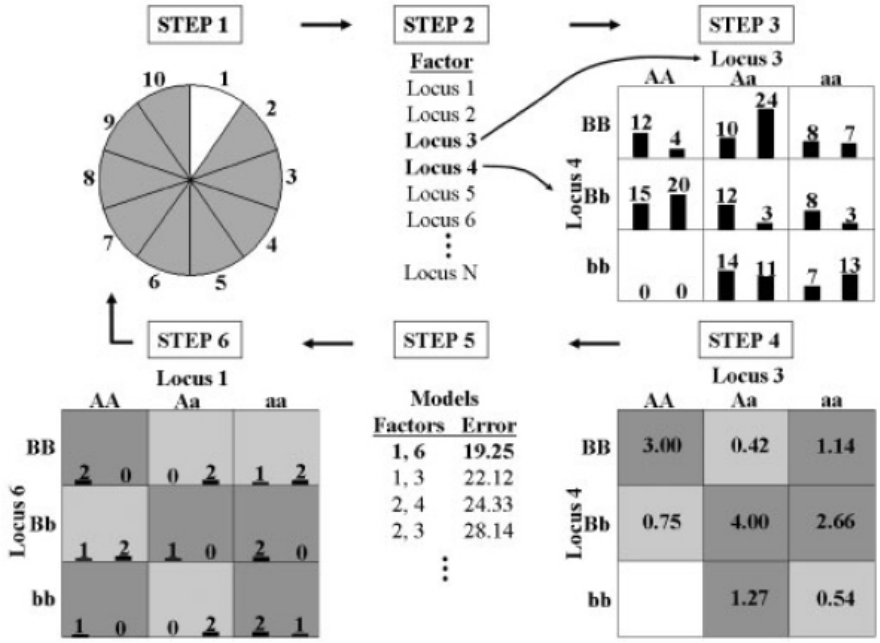
La idea básica del método MDR consiste en examinar la proporción de casos y controles entre los individuos que presentan cada una de las combinaciones de genotipos de la interacción y clasificarlos en dos categorías de riesgo alto o bajo. Esto es, si representamos la interacción en una tabla de contingencia k-dimensional, y examinamos la proporción de casos y controles en cada casilla de la tabla, aquellas casillas que presentan una ratio entre casos y controles igual o mayor que la ratio global de la muestra las catalogamos como casillas de riesgo alto, y aquellas en el que esta ratio es menor que la global, de riesgo bajo. De esta forma obtenemos un modelo de clasificación en el que los individuos en las casillas de riesgo alto se clasifican como casos y los de casillas de riesgo bajo como controles.

Inicialmente se divide el conjunto de datos en n partes iguales (típicamente 10) para hacer una validación cruzada. $n-1$ partes se utilizan para establecer el modelo de clasificación, en base a lo expuesto antes, para cada uno de los modelos de interacción entre SNPs, y la última parte se utiliza para estimar el error de predicción de cada interacción. Esto se repite n veces, dejando cada vez una parte distinta, de las n en que se han divido los datos, para la estimación del error.

La figura 8 resume este proceso de forma esquemática.

Así pues, para cada una de las n partes en que se ha dividido la muestra, y para cada una de las posibles interacciones de un determinado orden l especificado, combinaciones de l SNPs, se construye el modelo de clasificación, utilizando las $n-1$ partes restantes, en base las proporciones de casos y controles en cada casilla y se calculan distintas medidas de precisión. En concreto se calculan la proporción de individuos clasificados correctamente tanto en el conjunto de entrenamiento como en el de test, y la precisión balanceada como el promedio entre sensibilidad y especificidad, también para el conjunto de entrenamiento (ACC) y para el conjunto de test (PRED). La proporción de bien clasificados y la precisión balanceada coinciden

Figura 8: Esquema sobre la metodología MDR extraido de: M.D. Ritchie, L.W. Hahn, J.H. Moore (2003) *Power of Multifactor Dimensionality Reduction for Detecting Gene-Gene Interactions in the Presence of Genotyping Error, Missing Data, Phenocopy, and Genetic Heterogeneity* Genetic Epidemiology, 24, 150-157.



para muestras con igual número de casos y controles. Para muestras no balanceadas es más apropiada la segunda medida.

De estas medidas, se calcula el promedio de entre los n valores obtenidos por la validación cruzada, y se utilizan para determinar cuál es el mejor modelo de interacción del orden especificado. Existen dos implementaciones del MDR, en una se utiliza la media de la precisión balanceada en los conjuntos de entrenamiento para seleccionar el mejor modelo de cada orden, y se utiliza la media de la predicción balanceada (conjuntos de test) para determinar el mejor modelo final, de entre los mejores modelos de cada orden. En la otra implementación, se selecciona el mejor modelo para cada repetición de la validación cruzada en función de la predicción balanceada, y finalmente se escoge el mejor modelo como aquel que aparece seleccionado mas veces entre las n validaciones. Es decir, se calcula otra medida, llamada consistencia de la validación cruzada, que contabiliza el número de veces que una determinada interacción es escogida como mejor modelo entre las n repeticiones de la validación cruzada. Se selecciona el mejor modelo del orden especificado, como aquel que tiene una mayor consistencia.

MDR permite comprobar la significación del resultado mediante un test de permutación, aplicando MDR sobre múltiples permutaciones de la variable respuesta.

6.2. Exploración mediante MDR y problemas detectados.

La principal limitación que se achaca al método MDR es que requiere mucho tiempo de computación, ya que requiere evaluar todas las posibles interacciones del orden especificado, lo que hace que para evaluar interacciones de alto orden se requiera una gran capacidad computacional.

En nuestro caso tenemos 267 SNPs, por lo que la magnitud de modelos a evaluar asciende a algo más de 35.511 para interacciones de orden 2, más de 3 millones para orden 3, unos 207 millones para orden 4 y casi 11 mil millones para orden 5. Además, estos modelos han de ser evaluados para cada repetición de la validación cruzada, lo que multiplica por 10 este número, y en el caso de querer tener una estimación de la significación hay que repetir el proceso un número elevado de veces. Para la ejecución del método disponíamos de un ordenador en dedicación exclusiva, con un procesador dual core a 3 GHz, 2 GB de memoria RAM y con linux como sistema operativo, y el orden máximo de interacción que pudímos evaluar en su totalidad es de 4 SNPs, sin ejecutar el test de significación y con un tiempo de computación ininterrumpida de varias semanas.

Teniendo en cuenta que estos 267 SNPs representan sólo una pequeña muestra de la cantidad de datos que se espera poder disponer en poco tiempo para este tipo de estudios, actualmente ya se dispone de unos 1500 SNPs para este estudio, y que en el caso de enfermedades complejas como la que estamos estudiando, probablemente el factor genético venga determinado por las interacciones de varios genes, se hace evidente que es necesaria una selección de variables previa al análisis con MDR.

Pero durante el análisis de los resultados de MDR, detectamos otras limitaciones que presenta este método (Calle et al. 2007), algunas más generales y otras más específicas del tipo de estudio concreto. La más destacada de estas limitaciones es que interacciones importantes pueden pasar desapercibidas.

Por ejemplo, al examinar la interacción entre los SNPs 40 y 252 obtenemos la tabla 5. Vemos que el genotipo 11, correspondiente a los individuos heterocigotos para ambos SNPs, presenta una ratio entre casos y controles de 2.59, muy por encima del 1.12 de la muestra, y con un número considerable de individuos con este genotipo, que indica una cierta asociación con la enfermedad. Pero al reducir estos datos a dos categorías, tal como hace MDR, los individuos con genotipo 11 se agrupan con los de genotipo 00 y 20. El genotipo 20 no es influyente por tener muy pocos individuos, pero el genotipo 00 tiene muchos individuos y no presenta asociación con la enfermedad. El resultado es que al agrupar los tres genotipos la asociación del genotipo 11

queda diluida y el MDR no detecta la interacción entre estos SNPs como importante.

Tabla 5: Exploración de la interacción SNP40 x SNP252 con MDR

Genotipos	Casos	Controles	Ratio	Categoría MDR
00	88	77	1,14	H
01	102	114	0,89	L
02	38	34	1,11	L
10	50	59	0,84	L
11	96	37	2,59	H
12	18	28	0,64	L
20	12	6	2,00	H
21	14	18	0,77	L
22	6	6	1,00	L
TOTAL	424	379	1,12	

Otra limitación importante, tal como muestran los resultados de un estudio de simulación realizado por Ritchie et al.(2003), es la baja potencia de MDR en datos con presencia de heterogeneidad genética, que se refiere al hecho que un determinado fenotipo puede estar causado por distintas causas genéticas, un fenómeno característico de las enfermedades complejas.

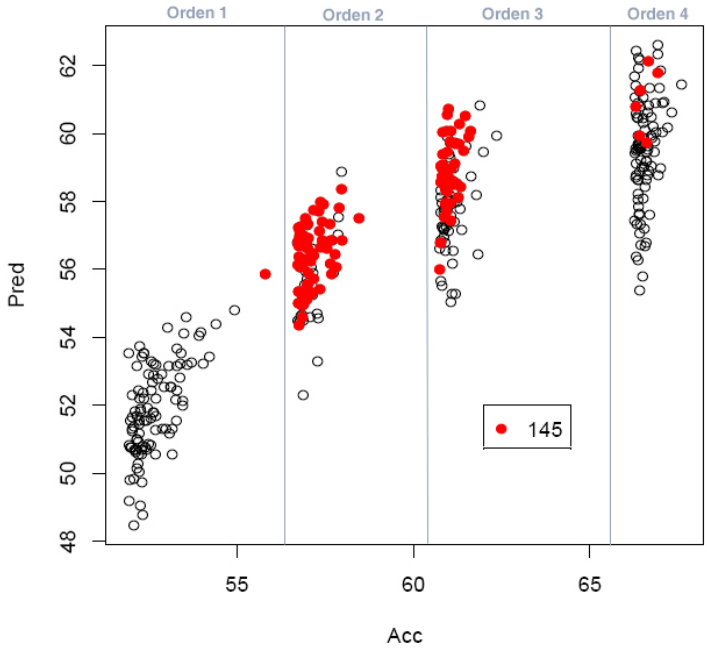
Otras carencias, que son comunes a los métodos no paramétricos, pero que en nuestro caso son importantes, son la imposibilidad de ajustar por efectos principales, es decir por el efecto observado que se puede atribuir a un SNP individual, y la imposibilidad de ajustar por variables confusoras. La segunda obliga a estratificar la muestra si se desea eliminar esa confusión, multiplicando los esfuerzos de análisis y computación y disminuyendo la potencia del método.

El no poder ajustar por efectos principales hace que sea difícil discriminar si una determinada interacción detectada por MDR, lo es simplemente por la influencia de uno de los SNPs o por la propia interacción. En la tabla 6 aparecen las interacciones significativas a un nivel de 0.05 detectadas por MDR después de realizar un test permutacional. En este caso la significación se ha calculado de forma diferente a como lo hace el MDR; se ha calculado la distribución empírica de la medida de predicción balanceada media (sobre los conjuntos de test) bajo la hipótesis nula de no asociación, calculada a partir de permutaciones aleatorias de la variables respuesta, y se ha evaluado la significación de los resultados a partir de esta distribución. En la tabla se aprecia como todas las interacciones de orden 2 y dos de las de orden 3 detectadas, contienen uno de los SNPs detectados individualmente (145 ó 151). Probablemente, estos resultados son fruto del efecto marginal y son consecuencia de

Tabla 6: Interacciones significativas identificadas por MDR hasta orden 3

Orden de interacción	SNP1	SNP2	SNP3
1	145		
	27		
	151		
	230		
	46		
2	151	21	
	169	145	
	179	145	
	151	72	
	145	129	
	209	145	
3	230	64	17
	239	179	145
	263	88	81

Figura 9: PRED vs ACC para los 100 mejores modelos detectados por MDR, para interacciones de orden 1 a 4



no poder separar este efecto.

Otro ejemplo de la importancia de ajustar por efectos marginales se muestra en la figura 9 donde he representado las medidas ACC y PRED, que son las medidas que utiliza MDR para seleccionar los mejores modelos, para los 100 modelos con mayor ACC y para interacciones de orden 1 a 4. He representado en rojo los modelos en los que aparece el SNP 145, que tiene un efecto marginal importante. Por ejemplo, para los modelos de orden 2, entre los 100 mejores modelos detectados por MDR, en 64 aparece el SNP 145.

6.3. MB-MDR como alternativa a MDR

Para intentar solventar algunos de los inconvenientes del MDR planteados, trabajamos en una aproximación distinta, pero basada en la idea principal del MDR, llamada MB-MDR, Model Based MDR (Calle et al. 2007). Esta nueva aproximación tiene dos diferencias principales respecto al MDR; solo agrupa genotipos con una cierta evidencia de asociación y utiliza un modelo paramétrico para estudiar esta asociación.

En la primera fase del MDR se clasifican las distintas casillas de la tabla de contingencia de la interacción en alto riesgo (H) y bajo riesgo (L), en función de la ratio entre casos y controles, para después agrupar todos los genotipos de la misma categoría. En este punto, el MB-MDR, introduce una nueva categoría de no evidencia (0). Aquellas casillas que no muestran evidencia de asociación o tienen un tamaño insuficiente de muestra no se agrupan con el resto sino que se agrupan en esta nueva categoría.

La forma de clasificar ahora los genotipos en tres categorías se basa en dos pasos. Sea j el número de genotipos resultantes de considerar interacciones de un determinado orden l, $j = 1, \ldots, 3^l$ y denotemos por c_j la casilla j-ésima de la correspondiente tabla l-dimensional.

Primero se realiza un test de asociación para cada casilla de la tabla, siendo la hipotesis nula que $OR_j = 1$, donde OR_j se refiere al *odds ratio* de comparar el grupo de individuos en la casilla c_j, individuos que presentan el genotipo determinado por esa casilla, con el resto de individuos. En principio este test lo realizamos de forma paramétrica, a partir de una regresión logística, que permite ajustar por efectos principales y variables confusoras, pero que también podría realizarse de forma no paramétrica, como por ejemplo a partir de una prueba ji-cuadrado.

Para este primer paso se escoge un nivel de significación conservador, de 0.1, ya que la potencia para detectar asociación con una sola casilla es baja. Así pues las casillas con $OR > 1$ y un p-valor inferior a 0.1 se agrupan en la categoría H, las que tienen

un $OR < 1$ y p-valor inferior a 0.1 se agrupan en la categoría L, y el resto, las que no presentan un *odds ratio* significativamente distinto de 1, se agrupan en la categoría 0.

Una vez hecha la agrupación, se realiza un nuevo test de asociación para calcular los *odds ratio* para las categorías H y L, es decir, se calcula el *odds ratio* del grupo clasificado como H contra el resto, OR_H, y el *odds ratio* del grupo L contra el resto, OR_L. De nuevo aquí se puede optar por un modelo paramétrico o no paramétrico. En nuestro caso, optamos por sendas regresiones logísticas.

La significación para estos nuevos *odds ratio* se establece a partir del estadístico de Wald del que en este caso desconocemos la distribución teórica ya que la manipulación de combinar casillas convenientemente no nos permite considerar que su función de distribución sea una ji-cuadrado como en el caso de la regresión estándar. Así pues, para poder establecer la significación de los resultados calculamos experimentalmente las distribuciones de referencia bajo la hipótesis nula de no asociación, a partir de repetir multitud de veces este proceso pero partiendo de permutaciones aleatorias de la variable respuesta. Concretamente calculamos sendas distribuciones nulas en función del orden de la interacción y del número de casillas agrupadas.

La tabla 7 muestra los resultados para el mismo ejemplo considerado en la sección anterior, la interacción entre los SNPs 40 y 252, donde ahora sí se detecta esta interacción como significativa.

Tabla 7: Exploración de la interacción SNP40 x SNP252 con MB-MDR

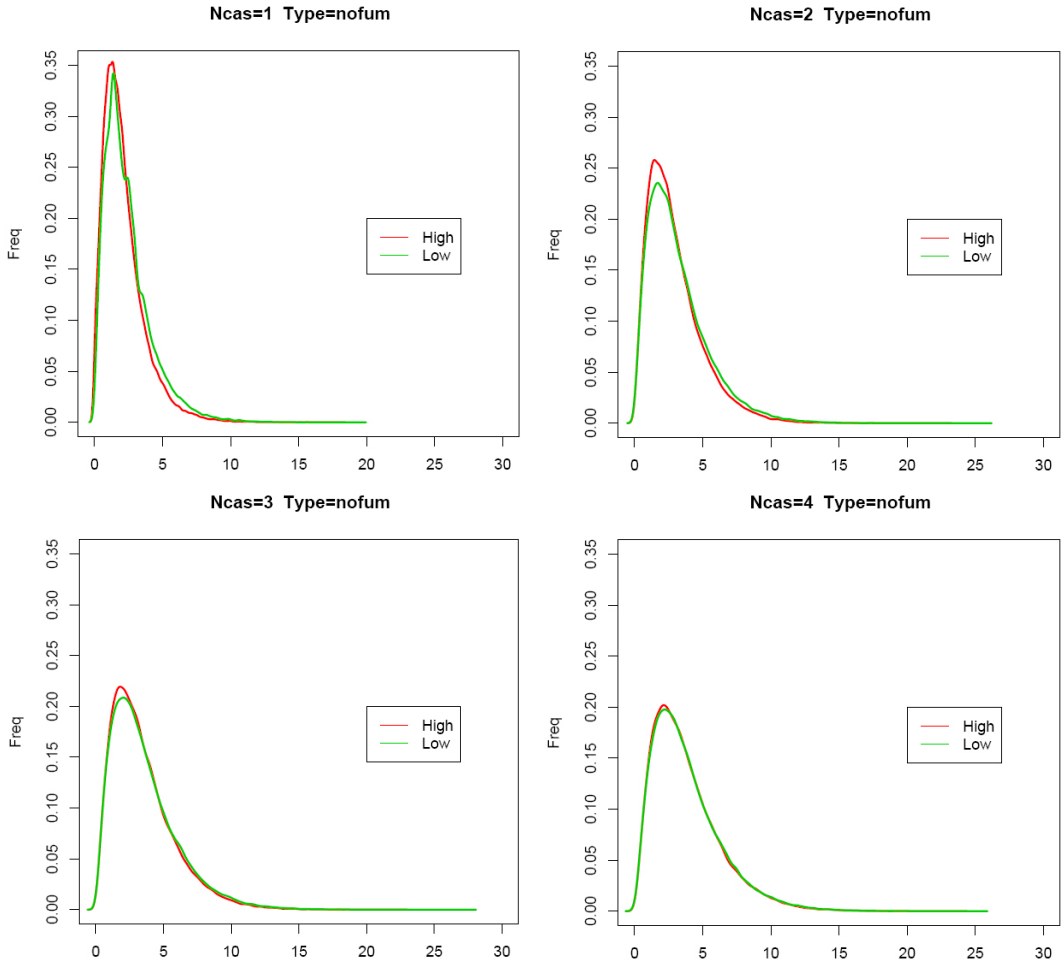
		Casos	Controles	OR	p-valor	Categoría
genotipos (Paso 1)	00	88	77	1,01	0,9303	0
	01	102	114	0,73	0,0562	L
	02	38	34	0,98	1,0000	0
	10	50	59	0,76	0,1229	0
	11	96	37	2,68	0,0000	H
	12	18	28	0,55	0,0675	L
	20	12	6	1,99	0,3399	0
	21	14	18	0,67	0,3668	0
	22	6	6	0,84	1,0000	0
categorías (Paso 2)	H	96	37	2,68	3.34 e-05	
	L	120	142	0,65	7.45 e-02	
	0	208	200			

A la hora de establecer las distribuciones de referencia, comprobamos que éstas solo

dependían del orden de interacción considerado y del número de casillas que se agrupan, y que son invariantes respecto a la categoría de riesgo (alta o baja) al tamaño muestral y a la proporción de casos y controles, tal como se muestra a continuación.

Por ejemplo, la figura 10 representa las estimaciones, de forma no paramétrica, de las distribuciones empíricas para el grupo de no fumadores, según la categoría de riesgo y el número de genotipos agrupados, y muestran que son invariantes respecto la categoría de riesgo. Sólo se muestran hasta 4 a modo de ejemplo, pero los hemos calculado para valores superiores con idéntico resultado.

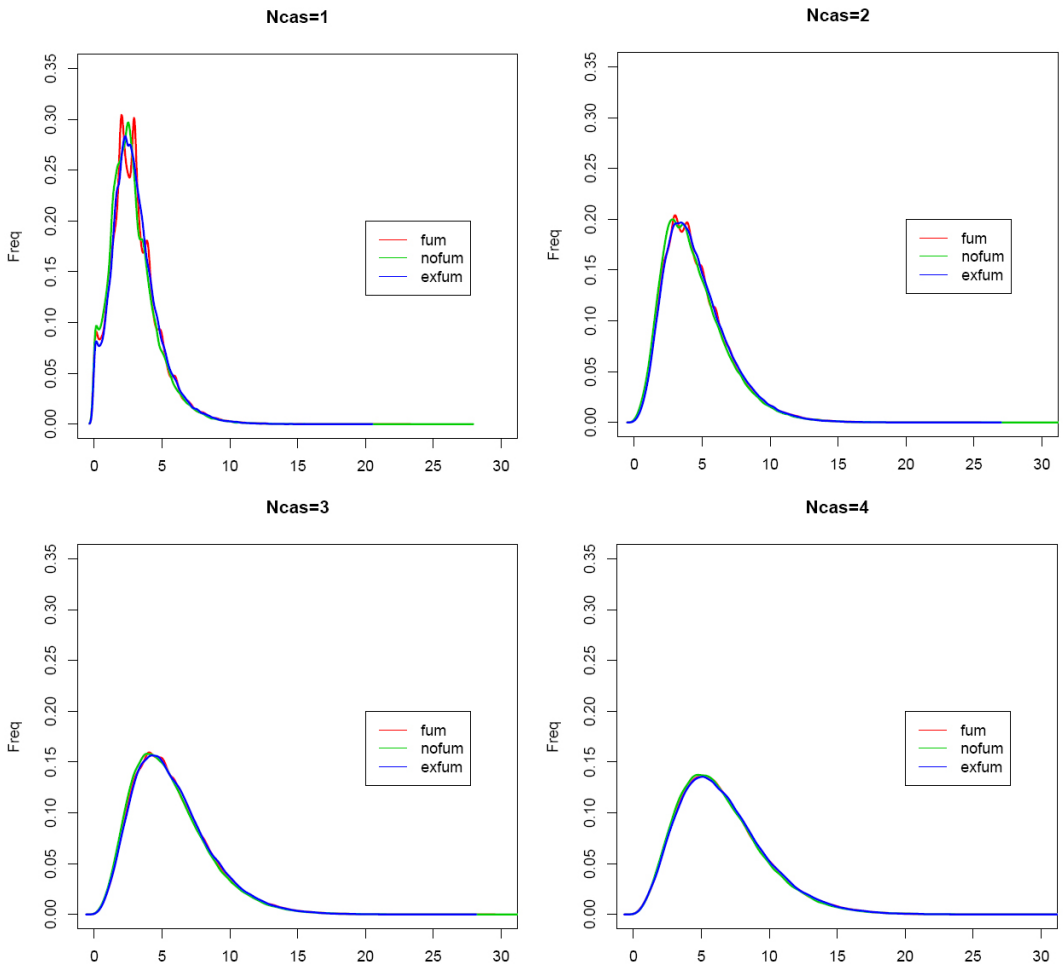
Figura 10: Distribuciones empíricas bajo la hipótesis nula de no asociación para el grupo de fumadores, según la categoría de riesgo.



La figura 12 muestra el resultado de las distribuciones para interacciones de orden 3 y distintos valores de genotipos agrupados, separando entre los tres grupos de exposición a tabaco. Se observa como tampoco hay diferencias entre los tres grupos. En este caso, los tres grupos son tres conjuntos de datos distintos, independientes,

con distinto tamaño muestral y con muy distinta proporción entre casos y controles. El hecho que la distribución se comporte de forma tan similar en estos tres grupos nos induce a pensar que la distribución del estadístico también es invariante respecto a la configuración de la muestra, cosa que implicaría no tener que realizar un test permutacional para establecer la significación de los resultados en cada nuevo conjunto de datos en los que se aplicara el método MB-MDR.

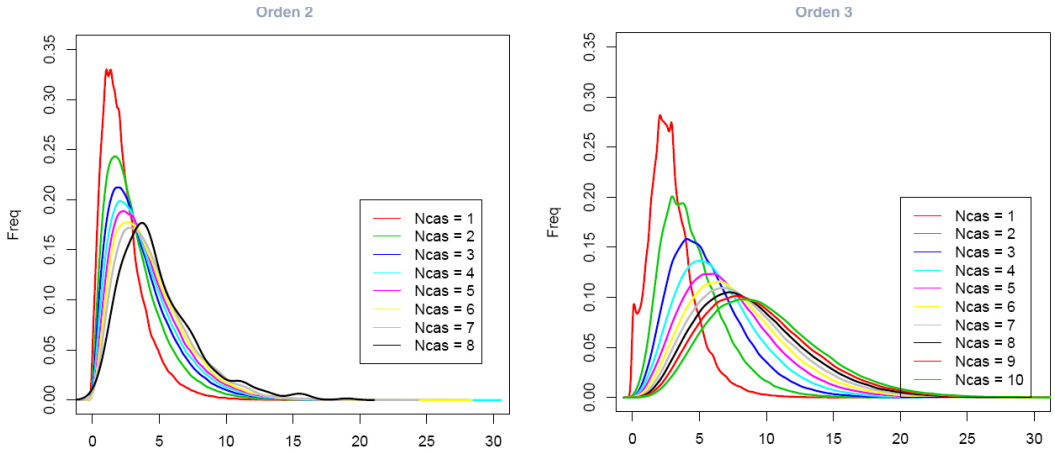
Figura 11: Distribuciones empíricas bajo la hipótesis nula de no asociación para interacciones de orden 3, según el número de genotipos agrupados.



Finalmente, la figura **??** muestra las distribuciones para orden de interacción de 2 y 3, según el número de genotipos agrupados, sin diferenciar ni por categoría de riesgo, ni por estrato de tabaco.

El MB-MDR constituye una aproximación similar al MDR, en el sentido de reducir la dimensión del problema, pero con la ventaja de poder ajustar los resultados por

Figura 12: Distribuciones empíricas bajo la hipótesis nula de no asociación para interacciones de orden 2 y 3, según el número de genotipos agrupados.



efectos marginales y covariables y ganando en potencia, tal y como se muestra en Calle et al.(2007) al comparar los resultados obtenidos con los mismos datos de simulación utilizados por Ritchie et al.(2003), con los resultados del MDR.

Además, aunque el coste computacional de evaluar cada interacción es mayor, los resultados obtenidos apuntan a que no es necesario hacer un test permutacional cada vez, puesto que se puede estimar una distribución de referencia genérica.

No obstante, para conjuntos de datos con un número muy elevado de variables, o para poder explorar interacciones de órdenes superiores, el coste computacional que requiere este método hace necesario un filtro previo de variables para reducir el número de interacciones a explorar. Esta selección la hemos realizado a partir de la medida de sinergia comentada en la sección 5. Para ello hemos seleccionado las 100 interacciones con mayor sinergia para cada tipo de exposición a tabaco, para orden de interacción de 2 y 3 SNPs.

Así pues, hemos analizado con el método MB-MDR el conjunto de datos de cáncer de vejiga, estratificando por exposición a tabaco, para detectar interacciones de orden 2 y 3 asociadas con la enfermedad, aplicándolo a las 100 interacciones de mayor sinergia para cada caso.

La significación la hemos evaluado a partir de las distribuciones nulas comentadas antes, y finalmente hemos ajustado los p-valores mediante FDR para controlar el efecto de hacer múltiples test de hipótesis.

A continuación se muestran a modo de ejemplo los resultados obtenidos para el estrato de los fumadores. Se han detectado 8 interacciones de orden 2 significativas

y 14 interacciones de orden 3, que se muestran en la siguiente tabla:

Tabla 8: Interacciones significativas

	SNPs			W	p-valor ajustado
Orden 2	224	151		19.01	5.38 e-05
	165	80		16.71	7.72 e-05
	282	157		18.12	1.20 e-04
	220	112		14.66	1.60 e-04
	259	112		14.56	1.90 e-04
	239	192		17.23	2.00 e-04
	173	99		16.69	2.50 e-04
	201	6		16.10	2.80 e-04
Orden 3	219	165	80	32,73	0
	218	130	82	33,32	1,68 e-06
	256	173	103	31,31	3,37 e-06
	241	224	151	31,82	6,85 e-06
	173	104	59	27,23	2,69 e-05
	235	155	90	24,07	4,18 e-05
	238	116	87	26,29	4,88 e-05
	208	170	50	28,53	5,31 e-05
	153	103	35	25,86	6,57 e-05
	235	218	192	29,26	7,23 e-05
	190	156	101	22,92	9,02 e-05
	188	153	112	27,21	9,59 e-05
	150	131	33	26,80	0,00012
	250	237	190	26,55	0,00013

Para el estudio realizado en esta sección ha sido necesario implementar el método MB-MDR en R. Concretamente he desarrollado dos funciones de R, una con la propia implementación del método, y otra que permite generar las distribuciones de referencia bajo la hipótesis nula mediante permutaciones de la variable respuesta para cada orden de interacción y en función del número de casillas agrupadas. Los detalles de estas funciones aparecen en la sección 8.2.

7. CART / RandomForest

7.1. Descripción del método CART

Los árboles de clasificación y regresión (CART), es una técnica no paramétrica basada en la generación de un modelo con estructura de árbol que permita explicar o predecir una determinada variable respuesta (Breiman et al. 1984), que puede ser tanto categórica (árboles de clasificación) como continua (árboles de regresión).

Este modelo en forma de árbol se construye a partir de la división sucesiva de la muestra en subgrupos sobre el espacio de variables. Se utiliza un método recursivo en que cada grupo de datos es dividido en dos subgrupos en base a una regla binaria (con dos posibles valores). En cada división sólo interviene una variable, que se escoge cada vez mediante la exploración exhaustiva de todas las posibilidades de forma que los subgrupos resultantes sean lo más homogéneos posible. Para variables continuas y variables ordinales, las reglas de división son del tipo $x_j < t$ y $x_j >= t$ para un determinado valor t, obteniendo un subgrupo de observaciones que presentan un valor menor a t para la variable x_j, y otro subgrupo con el resto. Para variables categóricas la división se realiza decidiendo, para cada categoría de la variable, a cuál de los dos subgrupos se asignan las observaciones que presentan ese valor, quedando definida la regla de división mediante una partición de las categorías en dos conjuntos disjuntos. El proceso se selección de cada división requiere determinar conjuntamente la mejor variable y la regla de división (el valor t si es continua o la partición si es categórica), analizando todas las posibles reglas en función de cómo son los dos subgrupos resultantes. Para cada variable categórica hay que analizar $2^{(K-1)} - 1$ posibles divisiones. El resto de variables puede tratarse como numéricas discretas, incluyendo las continuas puesto que sólo se observa un número finito de valores en la muestra, teniendo que comprobar $n - 1$ posibles valores, siendo n el número de observaciones, correspondientes a los puntos medios de cada dos valores consecutivos observados.

En el caso de árboles de clasificación, que es el que nos interesa en nuestro estudio caso-control, el criterio que se utiliza para seleccionar la mejor división de cada grupo se basa en el índice de Gini, que es una medida de la impureza de cada nodo (subgrupo). El índice de Gini para el nodo m se define como:

$$I_m = \sum_{k=1}^{K} p_{m,k}(1 - p_{m,k}) = 1 - \sum_{k=1}^{K} (p_{m,k})^2$$

donde $k = 1, \ldots, K$ son las categorías de la variable respuesta y $p_{m,k}$ es la proporción de elementos de la categoría k en el nodo m.

Esta medida alcanza su mínimo en 0, cuando todos los elementos de un nodo pertenecen a una misma clase. Para cada nodo se escoge la variable y regla de división que minimiza la suma ponderada de los índices de Gini de los subnodos

generados.

La determinación del árbol final se consigue en dos pasos. Primero generando un árbol con el máximo número de subdivisiones, que permite explicar de forma muy precisa la muestra pero que peca de un acusado sobre ajuste y, por tanto, muy poca capacidad de generalización. En segundo lugar, a partir de este árbol se utilizan técnicas de poda para ir reduciendo el tamaño y complejidad del árbol con el fin de reducir el sobre ajuste y ganar en poder de generalización, pero manteniendo un buen poder predictivo.

La figura 13 contiene un ejemplo de CART, generado a partir de datos de nuestro estudio, en donde el primer nodo ha sido dividido por la variable *cigcat* que contiene información sobre el habito de fumar en cuatro categorías; 0 para no fumadores, 1 para fumadores ocasionales, 2 para exfumadores y 3 para fumadores habituales. Esta primera división deja en el subnodo de la izquierda a los individuos con categorías 0 ó 1, y al de la derecha el resto. El subnodo de la derecha es dividido según el sexo, dejando los hombres a la izquierda (gender=1) y las mujeres a la derecha. El subnodo de la izquierda es dividido de nuevo por la variable *cigcat*, enviando a la izquierda los exfumadores (cigcat=2) y a la derecha los fumadores. A su vez, estos dos últimos nodos son divididos por la variable edad, con valores óptimos para la división de 66,5 y 55,5. Por ejemplo, el nodo con los exfumadores se divide entre los menores de 66,5 años (a la izquierda) y el resto (a la derecha).

Sin embargo, el resultado de un único árbol tiene algunos inconvenientes:

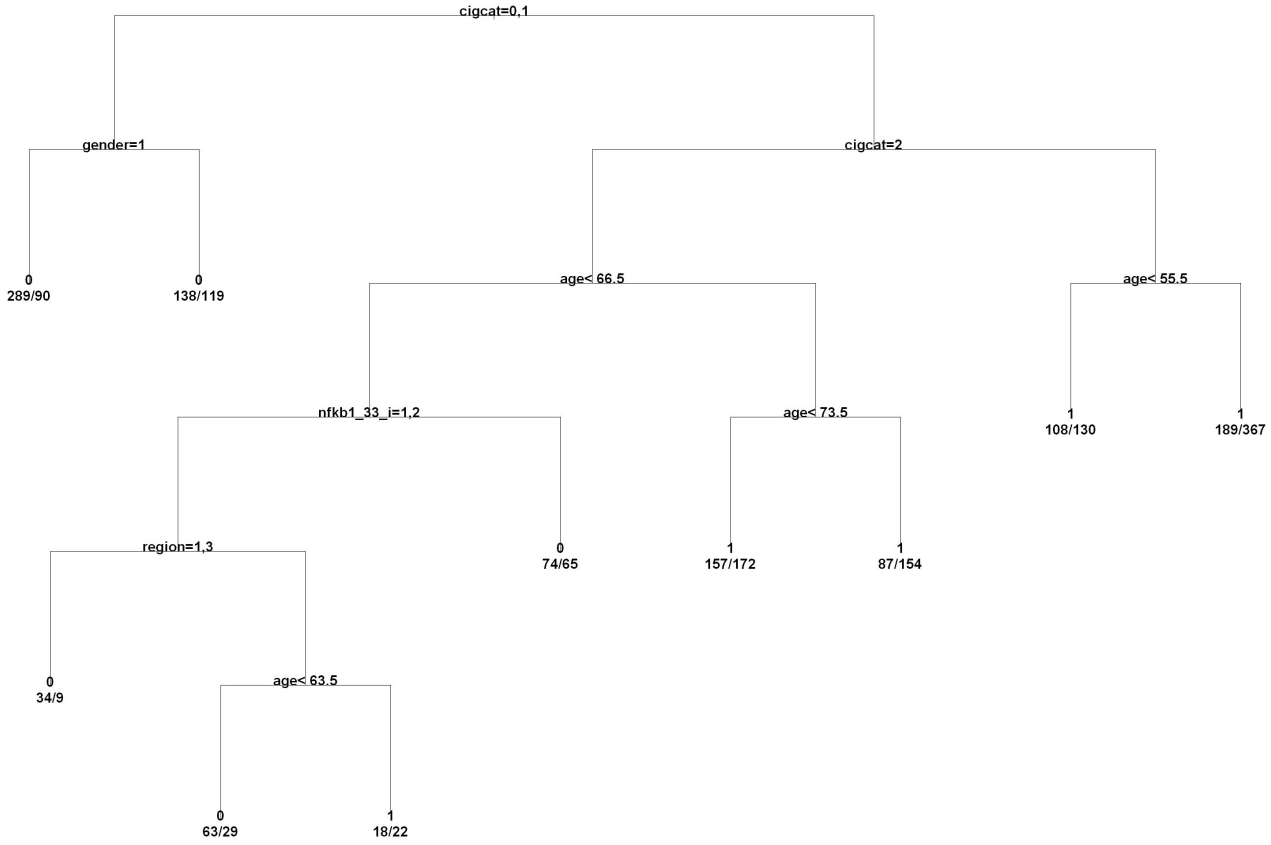
- Cada nodo del árbol es fruto de una serie de divisiones y por tanto, las divisiones posteriores están afectadas por las divisiones precedentes.
- Es poco robusto. Pequeños cambios en los datos pueden originar árboles muy distintos.
- Para cada división puede existir un conjunto de variables con un rendimiento muy similar, información que se pierde al escoger sólo una de ellas.

Existe una metodología alternativa basada en los CART que soluciona estos problemas; el *random forest* (Breiman et al. 2001). La idea principal de *random forest* consiste en generar multitud de árboles distintos a partir de los cuales se establece la clasificación de los datos por votación, es decir, cada caso se clasifica según la categoría mayoritaria a partir de la clasificación de cada árbol.

Las principales ventajas de esta metodología son:

- Proporciona una buena capacidad predictiva incluso cuando hay más variables que observaciones y cuando la mayoría de las variables son ruido.
- No sobre ajusta los datos

Figura 13: Ejemplo de CART a partir de nuestros datos. En cada rama aparece la condición que cumplen las observaciones enviadas al nodo de la izquierda; para los de la derecha es el complementario. Para los nodos finales aparece la categoria predecida para los individuos de ese nodo (0 ó 1) y el número de controles/casos en el nodo.



- Proporciona un ranking de importancia de las variables.

La idea básica del *random forest* se centra en cuatro puntos:

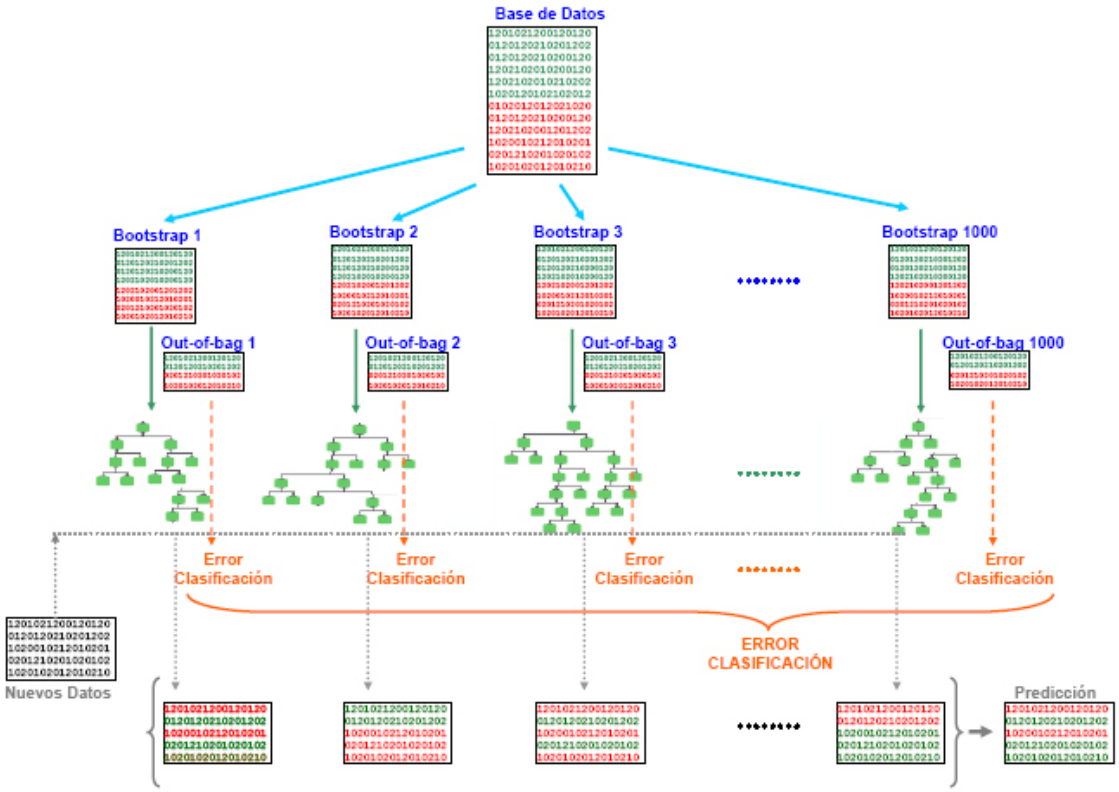
- No se genera un único árbol sino un gran número de ellos. Además se generan árboles maximales, sin poda.
- Los árboles se construyen a partir de muchos conjuntos de datos similares generados mediante bootstrap de la muestra original, es decir, haciendo remuestreo con reposición. De esta forma se consiguen dos cosas, primero corregir el error de predicción debido a la selección específica del conjunto de datos y, segundo, disponer para cada árbol de una muestra independiente (emphout-of-bag) para la estimación del error de clasificación, puesto que aproximadamente un tercio de la muestra original queda excluida de cada muestra generada por bootstrap.
- Para cada división de un nodo, no se selecciona la mejor variable de entre todas, sino que se selecciona al azar un subconjunto de variables del tamaño especificado y se restringe la selección de la variable a este subconjunto. De esta forma se incluye una mayor variabilidad de árboles y se reduce la dependencia del resultado con las divisiones precedentes.
- El *Random Forest*, a diferencia del CART, no proporciona una representación gráfica de las interrelaciones entre las variables, sino que establece un ranking de la importancia de las variables en la predicción de la variable respuesta.

La figura 14 muestra un esquema en el que se resumen las ideas principales de *random forest*.

La cuestión de la medida de importancia de las variables es un punto crucial y delicado porque la importancia de una variable está condicionada a su interacción, posiblemente compleja, con otras variables. El RandomForest calcula dos medidas de importancia distintas.

La primera, denominada MDA (Mean Decrease Accuracy), se basa en la contribución de la variable al error de predicción, es decir, al porcentaje de mal clasificados. El error de clasificación de cada árbol se calcula a partir de la parte de la muestra que ha quedado excluida de la submuestra utilizada en la construcción del árbol, generada por remuestreo. Para calcular la importancia de cada una de las variables que aparecen en un árbol se permutan aleatoriamente los valores de esa variable, dejando intactos el resto de variables, y se vuelven a clasificar los mismos individuos según el mismo árbol pero ahora con la variable permutada. La importancia en ese árbol se calcula como el aumento en el error de predicción resultante. Finalmente se calcula la medida MDA, como la media de estos incrementos en todos los árboles en donde interviene la variable.

Figura 14: Esquema de la metodología *random forest*.



La segunda medida de importancia, denominada MDG (Mean Decrease Gini), se calcula a partir del índice de Gini. Éste es el criterio que se utiliza para seleccionar la variable en cada partición en la construcción de los árboles y que comporta una disminución de esta medida. La importancia de una variable en un árbol se mide como la suma de los decrementos atribuidos a esa variable y la importancia final, como la media en todos los árboles.

7.2. Aplicación de *Random Forest* y limitaciones

La implementación escogida para trabajar con este método es una librería de R llamada *randomForest*. Se trata de una interfície para R, de los programas en Fortran originales de Breiman (Breiman et al. 2001). Inicialmente se requiere ajustar una serie de parámetros a valores óptimos para el conjunto de datos a tratar, como son el número de árboles a generar o el tamaño de variables a considerar para cada división. Estos ajustes se realizan experimentalmente a partir de ciertas indicaciones propuestas por Breiman hasta obtener los mejores resultados posibles de forma estable. A modo de ejemplo, los valores determinados como óptimos en nuestro estudio para los dos parámetros más importantes fueron 1000 para el número de árboles a generar y 128 para el número de variables a considerar en cada división.

Una segunda cuestión importante que se presenta de entrada es el tratamiento de los datos faltantes. Aunque en la metodología de los árboles de regresión y clasificación es posible gestionar los datos faltantes sin tener que recurrir a su imputación, en la generación del *random forest* no es posible. La forma en que hemos imputado estos datos es a partir de una función, *rfImpute*, que utiliza el propio método del *random forest* para hacer las imputaciones. A partir del resultado de un *random forest* se calcula una medida de similitud o de proximidad entre dos individuos computando el número de árboles del *random forest* que clasifican en el mismo nodo final a ambos. La función comienza haciendo una primera imputación en base a la moda para variables categóricas o la media para las continuas. A continuación se genera un *random forest* y se calcula la matriz de distancias entre los individuos. En base a esta matriz se actualizan las imputaciones hechas, de forma que para las variables continuas se imputa la media ponderada de los valores conocidos, estableciendo los pesos como las distancias, y para los datos categóricos, se calcula la media de las distancias para cada categoría y se asigna la categoría con menor distancia media. A partir de aquí se vuelve a generar un nuevo *random forest* y se itera el proceso.

Para eliminar el efecto de una determinada imputación de datos faltantes y asegurar la robustez de los resultados hemos repetido el proceso de imputación 100 veces, por lo que contamos con cien conjuntos iniciales de datos. Para los resultados que se detallan en esta sección, se ha repetido el cálculo sobre los 100 conjuntos de datos y se ha promediado el resultado. Para pasos intermedios en los que computacionalmente no era factible realizar 100 repeticiones se ha efectuado el promedio de entre 20 de

los conjuntos.

El resultado del *random forest* es, por un lado, un método de predicción para nuevos casos a partir del conjunto de árboles generado, y para el que se da una estimación del error de predicción y, por otro lado, un ranking de importancia de las variables, que proporciona información sobre qué variables juegan un papel determinante en la predicción de la variable respuesta.

La siguiente tabla es un resumen de los resultados preliminares (no satisfactorios) obtenidos con RandomForest, aplicado tanto a la muestra total como estratificando por la categoría de exposición al tabaco.

Tabla 9: Resultados preliminares obtenidos con RandomForest

	Proporción mal clasificados		
Categoría	Casos	Controles	Error Predicción
Todos	0,2368	0,4978	36,73 %
No fumadores	0,8612	0,0631	32,50 %
Exfumadores	0,1639	0,6939	41,43 %
Fumadores	0,0262	0,8746	34,26 %

Observamos que en todos los casos se obtiene un error de predicción aproximadamente entre 30 % y 40 %, sin embargo hay diferencias importantes entre los diferentes estratos. En el caso de los no fumadores, el error de predicción es del 32,5 %, pero si observamos las proporciones de mal clasificados separando entre casos y controles, resulta que el método está clasificando erróneamente un 86 % de los casos, frente a un escaso 6 % de los controles, es decir, prácticamente está clasificando todos los individuos como controles. Algo similar, pero a la inversa, ocurre con el grupo de fumadores, para el que prácticamente se están clasificando todos los individuos como casos.

Es evidente pues, que estos resultados no son útiles y que el error de predicción no representa en este caso una buena medida. Este problema se debe a que en las muestras de estos subgrupos no hay un equilibrio entre el número de casos y de controles. Entre los no fumadores hay 32 % de casos frente a un 68 % de controles, proporciones que se invierten en el grupo de fumadores, con un 62,6 % de casos y un 37,4 % de controles.

La propia implementación de *random forest* que hemos escogido incluye dos opciones para evitar esta situación.

La primera opción consiste en especificar una distribución de probabilidades a priori para las categorías de la variable respuesta mediante el parámetro *classwt*. Esta dis-

tribución a priori se utiliza en el cálculo de la reducción del índice de Gini, afectando en la selección de la mejor variable en cada división, y por tanto, tiene un rol decisivo en la construcción de los árboles.

La segunda opción consiste en corregir el sesgo producido por tener una muestra no equilibrada, en la fase final de clasificación de un individuo. Esta opción no tiene ningún efecto en la construcción de los árboles, sino solamente en la clasificación final. Para decidir en qué categoría se clasifica un individuo, el *random forest* construye una tabla de votos, es decir una tabla con la clasificación sugerida por cada uno de los árboles para los que ese individuo quedaba fuera del remuestreo. Por defecto se clasifica el individuo en la categoría con más votos. Esta opción por defecto se puede modificar mediante el parámetro *cutoff*, mediante el que se especifica una probabilidad a cada categoría, de forma que un individuo se clasifica en aquella categoría con una mayor ratio entre la proporción de votos obtenidos y el valor del *cutoff* asignado a la categoría. En nuestro caso hay dos posibles categorías de la variable respuesta, 0 para los controles y 1 para los casos, por lo tanto, el *cutoff* tendrá dos componentes que denotaré como $cutoff = (1-p,\ p)$ con p entre 0 y 1. Entonces, un individuo se clasifica como $\hat{Y} = 1$ si $\frac{\#(\text{votos}=0)}{1-p} < \frac{\#(\text{votos}=1)}{p}$.

Una primera aproximación para esta corrección es especificar uno de estos parámetros, *cutoff* o *classwt*, como las proporciones observadas de casos y controles en cada grupo. Pero a priori, desconocemos la influencia real de estos parámetros y la robustez de los resultados bajo ligeras modificaciones de estos parámetros. En distintas pruebas que realizamos pudimos comprobar que la especificación de estos parámetros como las proporciones observadas de casos y controles no siempre daba buenos resultados. Estas dificultades nos han llevado a la propuesta de una nueva estrategia metodológica de exploración de *random forest* que permite el tratamiento de datos no equilibrados y la selección del parámetro *cutoff* óptimo. El material que se describe a continuación, en los apartados 7.3 y 7.4, es un material inédito que se someterá a revisión próximamente.

7.3. Exploración de *Random Forest* mediante AUC

La metodología propuesta consiste en la utilización de las curvas ROC y del área bajo la curva (AUC) para evaluar la capacidad predictiva del *random forest* y la determinación del *cutoff* óptimo para conjuntos de datos desequilibrados.

Tal y como se ha visto en la tabla 9, la medida de predicción global no nos proporciona un buen método de evaluación de *random forest* y, por tanto, es necesario evaluar la capacidad predictiva en casos y controles por separado, es decir, en base a la sensibilidad (proporción de enfermos bien clasificados) y a la especificidad (proporción de sanos bien clasificados).

Las curvas ROC son representaciones gráficas de la sensibilidad versus 1-especificidad, o lo que es lo mismo, de la proporción de verdaderos positivos (VP) versus la proporción de falsos positivos (FP). Una curva ROC tiene dos utilidades muy importantes: (a) permite determinar el punto de equilibrio óptimo entre sensibilidad y especificidad, (b) da una medida de la capacidad predictiva mediante el cálculo del área bajo la curva. El área bajo la curva ROC permite evaluar la capacidad del método para clasificar correctamente, y puede interpretarse como la probabilidad de que ante un par de individuos, uno sano y otro enfermo, el método los clasifique correctamente a los dos. Un valor de 1 para el área significa que el método es perfecto; un valor de 0.5 indica que el método es inútil y valores intermedios miden la capacidad del método para discriminar entre casos y controles, que será mayor cuanto más se aproxime a 1.

A partir de un *random forest* podemos construir la curva ROC variando el valor del *cutoff* $= (1-p,\ p)$ con p entre 0 y 1. Si hacemos tender p a 0, la razón entre el número de votos para clasificar un individuo como caso y p (valor de *cutoff* para los casos) tenderá a infinito para todos los individuos, y por tanto para valores de p próximos a 0, todos los individuos se clasificaran como casos. Análogamente, si hacemos tender p a 1, entonces $1-p$ tenderá a 0 y por tanto, para valores de p próximos a 1 todos los individuos se clasificarán como controles. Es decir, para $p=0$ tendremos $VP=1$ y $FP=1$, para $p=1$ tendremos $VP=0$ y $FP=0$, y al variar la p entre 0 y 1 iremos obteniendo la curva ROC del *random forest* generado.

En la figura 15 se muestran las curvas ROC generadas considerando distintos valores para el *cutoff*, tanto para toda la muestra como para los distintos subgrupos de exposición al tabaco. La tabla 10 muestra las correspondientes áreas bajo la curva. Se observa cómo, después de estratificar por tabaco, el área de la curva ROC para los fumadores es prácticamente de 0.5, lo que indica que la contribución a la predicción del resto de variables ambientales y de los factores genéticos es prácticamente nula, y los posibles efectos que pudieran tener quedan enmascarados por el efecto del tabaco, que es el principal factor de riesgo. Es decir, el riesgo asociado a la exposición al tabaco es tan elevado que hace despreciable el incremento del riesgo que pueda aportar la susceptibilidad genética. En el otro extremo tenemos el grupo de no fumadores, con AUC = 0.67, que, aunque tampoco es un valor muy elevado, muestra cierta evidencia de la componente genética. Estos resultados sugieren que la búsqueda de factores genéticos tiene más posibilidades de éxito si se lleva a cabo entre individuos que no han estado sometidos a factores de riesgo ambientales importantes, en nuestro caso el grupo de no fumadores.

Por otro lado, observando otra vez la figura 15 se ve claramente cómo pequeñas diferencias en los valores del *cutoff* pueden producir resultados muy distintos. Por ejemplo, observando el gráfico para toda la muestra, una p de 0.54 para el *cutoff* produce un error de predicción cerca del 30 % para los controles y superior al 50 % en los casos. Estos errores prácticamente se invierten al considerar una p de 0.5, y si consideramos una $p=0.58$, los errores para casos y controles son del orden del

80 % y 10 % respectivamente.

A partir de las curvas ROC podemos obtener una estimación del *cutoff* óptimo en cada caso y los valores del área bajo la curva, que se muestran en la tabla 10. El valor *cutoff* considerado como óptimo puede depender de las consecuencias reales que tenga hacer una predicción incorrecta de un enfermo como sano y viceversa. En nuestro caso hemos considerado ambas situaciones igual de importantes y hemos determinado el valor óptimo para el *cutoff* como aquel que equilibra ambas situaciones, determinando el punto de corte entre la curva ROC y la recta $y = 1 - x$

Tabla 10: Valores del área bajo la curva (AUC) para las curvas ROC correspondientes a la figura 15.

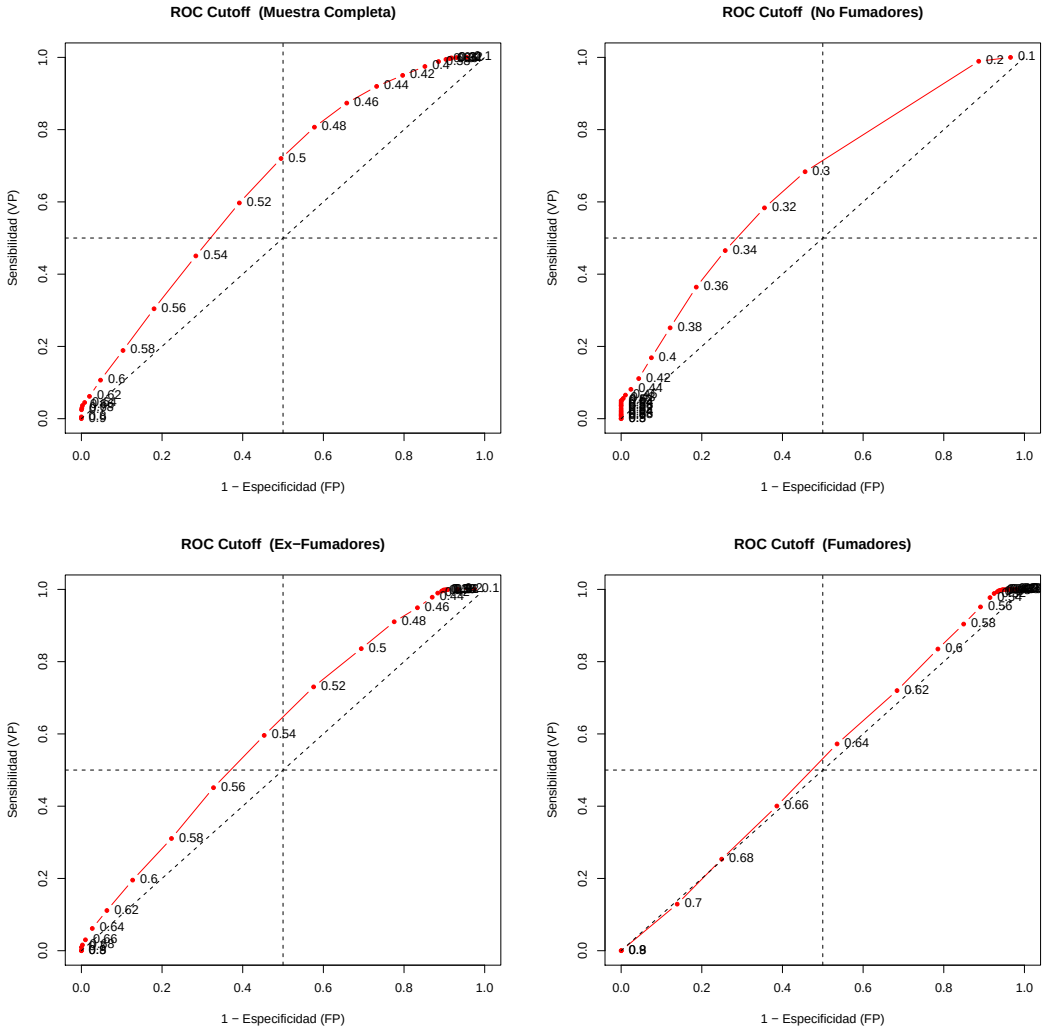
	AUC
Muestra Completa	0.6526
No fumadores	0.6670
Exfumadores	0.6063
Fumadores	0.5199

7.4. *Random Forest* como método de selección de variables

A parte del papel predictivo del resultado del *random forest*, éste también nos proporciona una forma de determinar la importancia de las variables con el fin de poder determinar qué variables juegan un papel en la predicción, y lo que nos proponemos es utilizar los resultados del *random forest* como método de selección de las variables más relevantes para un estudio posterior más detallado. En concreto, estamos interesados en identificar los marcadores genéticos para el grupo de no fumadores, para los que los resultados del *random forest* muestran cierta susceptibilidad genética.

Esta aproximación de selección mediante *random forest* ya ha sido utilizada para la selección de genes a partir de datos de microarrays (Díaz-Uriarte et al. 2006). La estrategia propuesta por Díaz-Uriarte se basa en un proceso iterativo de eliminación hacia atrás. Consiste en ajustar un *random forest*, y a partir de una medida del error de predicción y un ranking de importancia de las variables obtenidos, eliminar las variables con menor importancia y repetir el proceso. Finalmente se escoge aquel grupo de variables que hayan generado el *random forest* con menor error de predicción. Para el proceso de eliminación de variables utilizaron la medida MDA (Mean Decrease Accuracy) y eliminaron un 20 % de las variables a cada iteración. Para evitar un sobre ajuste de los datos, la medida de importancia la calculan al principio y utilizan ese ranking durante todo el proceso; no la recalculan a cada iteración.

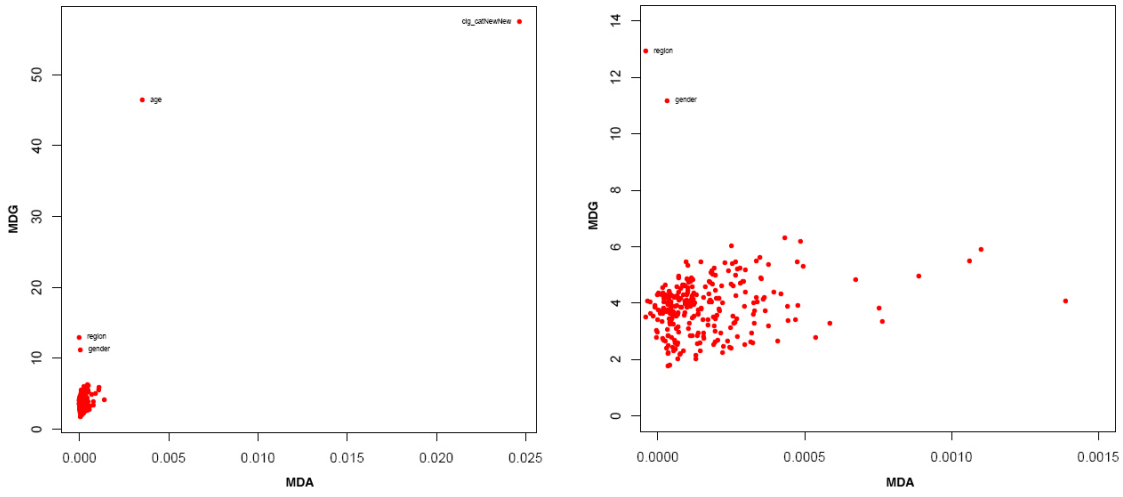
Figura 15: Curvas ROC generadas a partir de distintos valores del parámetro *cutoff* para los distintos estratos según la exposición al tabaco. Estas curvas corresponden a sendos *random forest* generados incluyendo todos los SNPs y las variables edad, sexo, región y exposición al tabaco (en el caso de la muestra competa).



La metodología que proponemos se basa en el misma idea principal de selección de variables por eliminación utilizando los resultados de *random forest*, pero superando tres limitaciones que, a nuestro entender, presenta la metodología antes citada: (a) utiliza sólo la medida de MDA para la eliminación de variables, (b) utiliza la medida del error de predicción global para seleccionar el subgrupo óptimo, (c) para datos no balanceados requiere de la especificación de parámetros de corrección, *cutoff*, a priori.

(a) Determinar la medida de importancia más adecuada no es una cuestión trivial. Hay dos medidas de importancia principales que se pueden calcular mediante *random forest*, el MDA y el MDG, y ambas miden aspectos distintos del buen ajuste del modelo. La primera mide el peso que tiene cada variable a la hora de hacer una predicción y la segunda mide la contribución de cada variable en la construcción del modelo. El problema de la selección de la medida de importancia no sería tal, o sería menor, si ambas medidas presentaran una alta correlación, es decir, si el ranking de variables generado por una medida fuera similar al generado por la otra, pero hemos comprobado que esto no sucede en nuestro caso. En la figura 16 aparece representado un diagrama de dispersión de las variables de nuestros datos según los valores de importancia de ambas medidas. Se aprecia como la segunda y tercera variables con más importancia según el índice de Gini quedan relegadas a las últimas posiciones según la disminución del error de predicción y variables de la parte alta del ranking según MDA pasan a la cola del ranking según el índice de Gini. Cabe destacar también cómo la exposición al tabaco es el factor más importante detectado por cualquiera de las variables de forma indiscutible, seguido de la edad.

Figura 16: Diagrama de dispersión de las variables según los valores de importancia MDA y MDG calculadas mediante *random forest* utilizando toda la muestra. El gráfico de la derecha es una ampliación de la parte más cercana al origen.



(b) La medida global de predicción no siempre es un buen criterio de selección del mejor modelo, sobretodo con datos no equilibrados, tal y como muestra la tabla 9.

(c) Como hemos visto en la sección anterior, al trabajar con datos desequilibrados es necesario utilizar un parámetro de corrección de las predicciones como el *cutoff*, por lo que se requiere establecer un valor a priori de este parámetro, cuyo valor óptimo puede ser diferente para cada subconjunto de variables.

Para solventar estas limitaciones, proponemos utilizar conjuntamente ambas medidas de importancia en el proceso de eliminación de variables y utilizar, en lugar del error de predicción global, el área bajo la curva (AUC) como medida de determinación del subconjunto óptimo, en base a las curvas ROC generadas a partir del parámetro *cutoff*. La metodología propuesta es la siguiente:

- Ajustamos un *random forest* con todas las variables y extraemos el ranking de importancia de las variables según cada una de las dos medidas, MDA y MDG. En este punto es importante notar que el ranking de importancia de cada medida no está influido por el parámetro especificado como *cutoff*, puesto que este sólo se utiliza como corrección en la predicción por votos mediante todos los árboles, mientras que las medidas de importancia se calculan como la media de la importancia en cada árbol individualmente. Esto no se cumpliría si en lugar del *cutoff* utilizáramos el parámetro *classwt*.

- Considerando el espacio de coordenadas bidimensional generado por las dos medidas de importancia, construimos un nuevo ranking de las variables de la siguiente forma: (1) reescalamos las medidas dividendo por su desviación estándar, (2) calculamos el valor mínimo observado para cada medida y situamos el origen de coordenadas en el punto con esas coordenadas y (3) establecemos el ranking en función de la distancia euclidea de cada punto al origen de coorndenadas.

- Mediante distintos valores de *cutoff*, generamos la curva ROC asociada al *random forest* y calculamos el AUC.

- Eliminamos el 20 % de las variables con menor valor de importancia según el ranking generado en el segundo punto, construimos un nuevo *random forest* con las variables no eliminadas, calculamos la curva ROC y su AUC, y reiteramos el proceso hasta agotar todas las variables.

- Seleccionamos el subconjunto de variables que optimice el AUC.

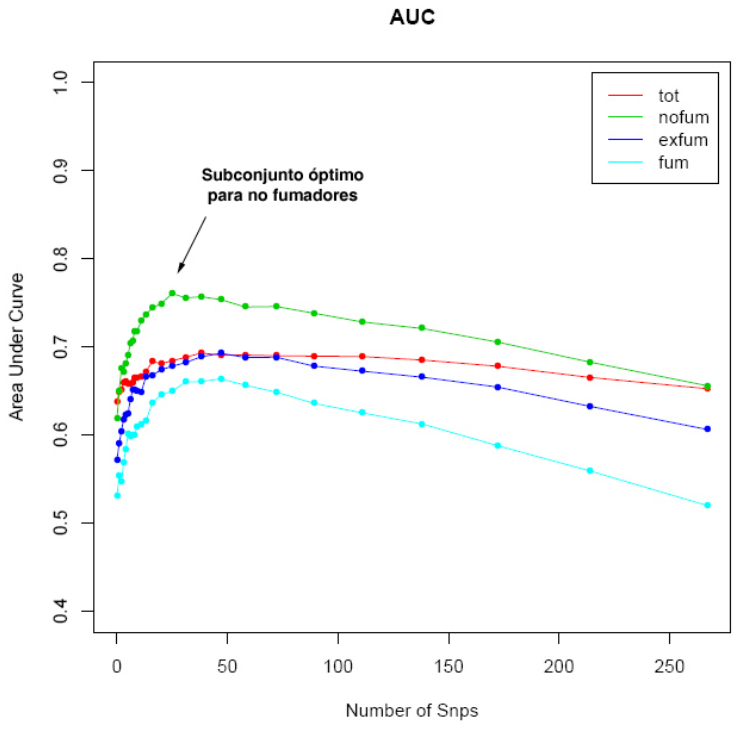
- Finalmente, calculamos el *cutoff* óptimo para el grupo de variables seleccionado.

En la aplicación de este proceso a nuestros datos hemos tenido en cuenta las siguientes consideraciones:

- El objetivo del análisis es seleccionar los SNPs candidatos a estar asociados con la enfermedad. El resto de variables incluidas (edad, sexo, región, exposición al tabaco) son factores de riesgo conocidos y que ya han sido analizados en estudios previos, por lo que lo que realmente nos interesa es estudiar si podemos extraer un mayor conocimiento incorporando el factor genético. Por este motivo, todas las covariables no genéticas se incluyen en todos los *random forest* generados, y únicamente vamos excluyendo SNPs, en el proceso descrito antes.
- El ranking que generamos para decidir el orden en que se van eliminando las variables, lo construimos una sola vez, al principio, y lo utilizamos durante todo el proceso con el fin de no sobre ajustar los datos (Díaz-Uriarte et al. 2006)
- A partir de que sólo quedan 10 SNPs, se van eliminando uno a uno.

En el gráfico de la figura 17 se representa el área bajo la curva en función del número de SNPs seleccionados. Hay que tener en cuenta que, aunque se han representado los cuatro estratos para tabaco en un mismo gráfico, el conjunto de SNPs en cada uno de los subgrupos seleccionados es distinto para cada estrato.

Figura 17: Gráfico del valor del AUC en función del número de SNPs considerados siguiendo la metodología propuesta en esta sección.



A partir de las áreas AUC determinamos el subconjunto óptimo de SNPs como aquel que alcanza el área máxima. Para el grupo de no fumadores, en el que estamos interesados, el óptimo se alcanza con 25 SNPs. A partir de una nueva curva ROC calculada para este conjunto de SNPs se puede determinar el *cutoff* óptimo y calcular los errores de clasificación, que dan un porcentage del 30,74 % (para casos, para controles y para el total, puesto que consideramos óptimo, a falta de otro criterio, el que equilibra los porcentajes de casos y controles). Sin embargo, esta medida seguramente está sobre ajustada y probablemente también lo esté el valor del nuevo AUC (aproximadamente de 0.77). Por tanto, a falta de un método que permita reajustar estos valores, no podemos asumirlos como ciertos y solo podemos considerar este método para la selección de SNPs. Así pues, aunque no podamos asegurar que el poder predictivo del conjunto de los 25 SNPs seleccionados sea superior al poder predictivo al considerarlos todos (AUC = 0.67), sí que hemos conseguido un método de reduccion de la dimensión, reduciendo los 267 SNPs iniciales a solo 25.

Para el estudio realizado en esta sección ha sido necesario implementar rutinas para el cálculo de las curvas ROC, el cálculo del área bajo la curva (AUC) y para el procedimiento de eliminación recursiva. Los detalles de estas funciones aparecen en la sección .

8. Rutinas en R

Para el desarrollo del trabajo realizado con las distintas metodologías expuestas en este documento ha sido preciso desarrollar diferentes utilidades bajo el lenguaje de programación R. En esta sección se detallan cuáles han sido estas rutinas, así como su descripción y forma de uso, en un formato similar al de la ayuda de las librerías de R.

8.1. Rutinas de exploración de la sinergia

Para la exploración de los datos de nuestro estudio mediante la medida de sinergia (sección 5), ha sido necesario programar las funciones para el cálculo de las distintas medidas que intervienen. Estas funciones, que se detallan más abajo, han sido desarrolladas de forma genérica, sin restricciones específicas de nuestros datos, para poder ser incluídas en una librería de R.

- ***CondEntropy***

  Descripción: Función interna para el cálculo de la entropía de una variable categórica C condicionada al conocimiento de k factores $G_1, \ldots, G_k$.

  Uso: CondEntropy(datos)

  Argumentos:

  datos : *data frame* con $k+1$ columnas formado por las variables categoricas $G_1, \ldots, G_k$ y C y en donde la última columna debe contener la variable respuesta C.

  Detalles: El resultado es un valor numérico con el valor de $H(G_1, \ldots, G_k; C)$; entropía de C condicionada a $G_1, \ldots, G_k$.

- ***Synergy2***

  Descripción: Función interna para el cálculo de la sinergia bidimensional de dos factores G_1 y G_2 respecto a otra variable categórica C.

  Uso: Synergy2(datos)

  Argumentos:

  ***datos*:** *data frame* con 3 columnas formado por las variables categoricas G_1, G_2 y C, y en donde la última columna debe contener la variable respuesta C.

  Detalles: El resultado es un valor numérico con el valor de syn($G_1, G_2; C$); sinergia de G_1 y G_2 respecto C.

- ***Synergy3***

  Descripción: Función interna para el cálculo de la sinergia tridimensional de tres factores G_1, G_2 y G_3 respecto a otra variable categórica C.

  Uso: Synergy3(datos)

  Argumentos:

  ***datos*:** *data frame* con 4 columnas formado por las variables categoricas G_1, G_2, G_3 y C, y en donde la última columna debe contener la variable respuesta C.

  Detalles: El resultado es un valor numérico con el valor de $\mathrm{syn}(\mathrm{G}_1, \mathrm{G}_2, \mathrm{G}_3; \mathrm{C})$; sinergia de G_1 y G_2 respecto C.

- ***Synergy***

  Descripción: Cálculo de la sinergia a partir de un conjunto de factores de todos los modelos de orden especificado respecto una cierta variable C también categórica. Hace uso de las funciones anteriores.

  Uso: Synergy(ordre,datos,Ycol=1,keep=100,batch=900)

  Argumentos:

  ***ordre*:** Orden de interacción de sinergia, puede ser 1 ó 2.

  ***datos*:** *data frame* con los factores y la variable respuesta, todos categóricos.

  ***Ycol*:** Especificación de la columna que contiene la variable respuesta C. Por defecto 1.

  ***keep*:** Número de modelos con mayor sinergia a devolver. Por defecto devuelve los 100 mejores modelos.

  ***batch*:** Número de modelos a guardar en memoria durante el proceso de calculo. Simplemente tiene valor computacional, cuanto mayor sea menos tiempo de cómputo será necesario pero hará falta más memoria.

  Detalles: Calcula la sinergia de todas las combinaciones posibles del orden especificado respecto la variable C y retorna un *data frame* con los k mejores modelos y su sinergia, siendo k el valor especificado por el argumento "keep" (por defecto 100). Para el cálculo de la entropía, no tiene en cuenta las observaciones con missings para alguno de los factores analizados.

8.2. Implementación del MB-MDR

La parte de programación que ha comportado más tiempo y trabajo es, sin duda, la implementación del método MB-MDR. Por una parte representa el trabajo de implementar y depurar el código, y por otra parte, representa un reto de programación porque, aunque su implementación es sencilla, requiere de mucha manipulación de datos, para las que es necesario reducir al máximo el tiempo de computación puesto que se deben iterar millones de veces.

- ***MBMDR***

  Descripción: Implementación del método MB-MDR propuesto en Calle et al. 2007 para la detección de interacciones en datos genéticos de gran dimensión.

  Uso: MBMDR(y, data, covar, pval=0.1, order, first.model=NULL, output="mbmdr.out", adjust=0, list.models=NULL, correction=T)

  Argumentos:

  ***y*:** Vector con la variable respuesta. Debe ser dicotómica.

  ***data*:** Matriz o *data frame* con el conjunto de SNPs codificados como 0, 1 y 2, para los tres genotipos. Los missings pueden estar codificados como NA o -1.

  ***covar*:** Matriz o *data frame* con el conjunto de variables de ajuste.

  ***pval*:** p-valor utilizado en la primera fase del MB-MDR. Por defecto es 0.1

  ***order*:** Orden de la interacción, es decir, número de SNPs intervinientes en los modelos a analizar.

  ***first.model*:** Modelo en donde iniciar la exploración.
  Por defecto se analizan todos los modelos de interacción posibles en un orden concreto. Si la matriz de datos contiene m SNPs y el orden de interacción es 3, se comienza por el modelo $[m, m-1, m-2]$, donde cada índice indica la columna correspondiente, y se hacen disminuir el tercer índice hasta llegar a 1, luego se disminuye el segundo en una unidad, se reinicializa el tercer índice a uno menos que el segundo y se repite la operación, etc. Es decir, se prueban todos de forma que un índice siempre es superior a los posteriores, y siempre se hace disminuir el índice de más a la derecha posible.
  Este argumento es útil en caso de error o stop del procedimiento para reiniciar el proceso a partir del primer modelo no analizado.

  ***output*:** Nombre para el fichero de salida

  ***adjust*:** Indica el nivel de ajuste deseado en las regresiones. 0 para no ajustar, 1 para ajustar por covariables, 2 para ajustar por efectos marginales y 3 para ajustar tanto por covariables como por efectos marginales. Para 1 y 3 es necesario especificar el parámetro *covar*.

***list.models*:** Argumento para especificar el/los modelos a analizar. Por defecto se analizan todos los modelos posibles, pero si se especifica este parámetro sólo se analizaran los modelos determinados. Este parámetro puede ser un vector de dimensión igual a *order* para indicar un único modelo, una matriz con los modelos como vectores fila de dimensión *order* o un string para indicar el nombre del fichero (formato texto separado por espacios) que contiene la especificación de los modelos.

***correction*:** Valor lógico (TRUE o FALSE) para indicar si se quiere aplicar la regresión corregida para las situaciones en que la respuesta sólo presenta una categoría.

Detalles: El análisis mediante MB-MDR consta de dos fases:

En la primera fase se analiza cada combinación de genotipos de la interacción de k SNPs, representados por cada una de las casillas de la tabla de contingencia k-dimensional, mediante una regresión logística, ajustando o no, de la respuesta en función de la exposición, entendiendo como exposición el presentar o no la combinación de genotipos a analizar. Si el factor de exposición resulta significativo se marca la casilla como significativa. En esta primera fase se fija un nivel de significación conservador, por defecto 0.1. Todas las casillas se clasifican en tres categorías según si es significativa o no y según la ratio entre casos y controles en cada casilla: (a) High risk, si la casilla es significativa y la ratio casos/controls es superior a la ratio para toda la muestra; (b) Low risk, si la casilla es significativa la ratio casos/controls es inferior a la ratio para toda la muestra; y (c) Null, si la casilla es no significativa.

En la segunda fase se agrupan todas las casillas clasificadas en la misma categoría, reduciendo la dimensión multifactorial de la interacción a un factor con tres categorías; High, Low y Null. Para las categorías High y Low, si es que no estan vacías, se realizan sendas regresiones logísticas de la respuesta respecto la exposició, donde ahora la exposición es pertenecer a una de las casillas clasificadas como High o Low respectivamente. El resultado de estas regresiones forma parte del resultado retornado.

Si para alguna de las regresiones la respuesta sólo tiene un factor, es decir, todos son casos o todos son controles, se utilitza la función *logistf* que realiza una regresión corregida.

El resultado se va guardando en un fichero con el nombre especificado para el argumento *output*, por defecto es "mbmdr.out". Al final este fichero contiene un registro por cada regresión efectuada en la segunda fase, de la que se registra la siguiente información: los SNPs de la interacción, la

categoría de riesgo analizada (H o L), el número de casillas que agrupa dicha categoría y los valores de la regresión correspondientes al coeficiente del factor de exposición, su error estándar, el correspondiente estadístico de Wald y su p-valor.

- ***MBMDR_NullDist***

  Descripción: Rutina para analizar la significación de los resultados del MB-MDR.

  Uso: MBMDR_NullDist(y, data, covar, pval=0.1, order, first.model=NULL, output="mbmdr.out", adjust=0, list.models=NULL, correction=T, B=500)

  Argumentos:

  ***B*:** Número de permutaciones a realizar.
  El resto de parámetros son los mismos que para MBMDR

  Detalles: A partir de permutaciones aleatorias de la respuesta, genera una distribución nula bajo la hipótesis de no asociación en función del número de casillas agrupadas. Esta distribución permite calcular los p-valores ajustados para los resultados del MB-MDR.

8.3. Rutinas de exploración de *random forest*

Para la exploración mediante *random forest* también ha sido necesario construir una serie de funciones que actualmente están implementadas de forma personalizada a las características de nuestros datos, por lo que todavía no están en formato de librería de R pero lo estarán próximamente. Estas funciones són las siguientes:

- ***Curvas ROC***

  Descripción: Rutina para generar la curva ROC para *random forest* a partir de distintos valores del parámetro *cutoff* y hacer su representación gráfica.

- ***AUC***

  Descripción: Rutina para calcular el área bajo la curva (AUC) a partir del resultado de una curva ROC.

- ***Selección de SNPs***

  Descripción: Función que implementa el método de eliminación recursiva de SNPs, en base a los resultados de medidas de importancia de *random forest*, y de selección del subgrupo óptimo en base al área AUC bajo las curvas ROC.

9. Bibliografía

Anastassiou D (2007) *Computational analysis of the synergy among multiple interacting genes.* Mol Syst Biol, 3, 83.

Benjamini Y, Hochberg Y (1995) *Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing Journal Of The Royal Statistical Society Series B-Methodological* Journal Of The Royal Statistical Society Series B-Methodological, 57, 289-300.

Boulesteix AL, Strobl C, Augustin T, Daumer M (2008) *Evaluating Microarray-based Classifiers: An Overview* Cancer Informatics, 6, 77-97.

Breiman L (2001) *Random forests* Machine Learning, 45, 5-32.

Breiman L, Friedman J, Olshen R, Stone C (1984) *Classification and regression tress* Chapman & Hall, New York.

Calle ML, Sánchez JA (2007) *Árboles de clasificación y regresión en la investigación biomédica* Medicina Clínica, 129, 702-6.

Calle ML, Urrea V, Malats N, Van Steen K (2007) *MB-MDR: Model-Based Multifactor Dimensionality Reduction for detecting interactions in high-dimensional genomic data* Technical Report n.24. Department of Systems Biology. Universitat de Vic. http://www.uvic.cat/recerca/ca/documents experimentals.html

Calle ML, Urrea V, Vellalta G, Malats N, Steen KV (2008) *Improving strategies for detecting genetic patterns of disease susceptibility in association studies* Statistics in Medicine

Díaz-Uriarte R, Alvarez de Andrés S (2006) *Gene selection and classification of microarray data using random forest* BMC Bioinformatics, 7, 3.

Hahn LW, Ritchie MD, Moore JH (2003) *Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions* Bioinformatics, 19, 376-382

Hastie T, Tibshirani R, Friedman J (2001) *The Elements of Statistical Learning. Data Mining, Inference and Prediction* New York: Springer-Verlag.

Iniesta R, Guinó E, Moreno V (2005) *Statistical analysis of genetic polymorphisms in epidemiological studies* Gac Sanit, 19, 333-41

Murta-Nascimento C, Silverman DT, Kogevinas M et al. (2007) *Risk of Bladder Cancer Associated with Family History of Cancer: Do Low-Penetrance Polymorphisms Account for the Increase in Risk?* Cancer Epidemiol Biomarkers Prev, 16, 1595-1600.

Park MY, Hastie T (2008) *Penalized logistic regression for detecting gene interactions.* Biostatistics, 9, 30-50.

Pepe MS, Cai T, Longton G (2006) *Combining Predictors for Classification Using the Area under the Receiver Operating Characteristic Curve* Biometrics, 62, 221-229.

Puente D, Malats N, Cecchini L, Tardón A, García-Closas R et al. (2003) *Gender-Related Differences in Clinical and Pathological Characteristics and Therapy of Bladder Cancer* European Urology , 43, 53-62.

Ripley Brian D (1996) *Pattern recognition and neural networks* Cambridge University Press.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) *Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer*, American Journal of Human Genetics, 69, 138-147.

Ritchie MD, Hahn LW, Moore JH (2003) *Power of Multifactor Dimensionality Reduction for Detecting Gene-Gene Interactions in the Presence of Genotyping Error, Missing Data, Phenocopy, and Genetic Heterogeneity* Genetic Epidemiology, 24, 150-157.

Ritchie W, Granjeaud S, Puthier D, Gautheret D (2008) *Entropy measures quantify global splicing disorders in cancer* PLoS Comput Biol, 4.

Ruczinski I, Kooperberg C, LeBlanc ML (2003) *Logic Regression* Journal of the Computational and Graphical Statistics, 12, 475-511.

Ruczinski I, Kooperberg C, LeBlanc ML (2004) *Exploring interactions in highdimensional genomic data: an overview of LogicRegression* Journal of Multivariate Analysis, 90, 178-195.

Storey JD, Tibshirani R (2003) *Statistical significance for genomewide studies.* Proc Natl Acad Sci USA, 100, 9440–9445.

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) *Bias in random forest variable importance measures: illustrations, sources and a solution* BMC Bioinformatics

Tarca AD, Carey VJ, Chen X, Romero R, Draghici S (2007) *Machine learning and its applications to biology* PLoS Comput Biol, 3, 116.

Ziegler A, König IR (2006) *A Statistical Approach to Genetic Epidemiology* Wiley