Métodos de selección de variables en estudios de asociación genética. Aplicación a un estudio de genes candidatos en Enfermedad de Parkinson

Máster en Técnicas Estadísticas. Proyecto Fin de Máster

Alumna:

Pilar Cacheiro Martínez

Directoras:

Carmen Cadarso Suárez

M. Luz Calle Rosingana

María Jesús Sobrido Gómez

A Coruña, Julio 2011

Se presenta el trabajo “Métodos de selección de variables en estudios de asociación genética. Aplicación a un estudio de genes candidatos en Enfermedad de Parkinson” como proyecto fin de Máster del Máster en Técnicas Estadísticas e Investigación Operativa de la alumna Pilar Cacheiro Martínez. Este trabajo está dirigido por la profesora Carmen Cadarso Suárez (Departamento de Estadística e Investigación Operativa, Universidade de Santiago de Compostela), la profesora M. Luz Calle Rosingana (Departament de Biologia de Sistemes, Universitat de Vic) y la Dra. María Jesús Sobrido Gómez (Coordinadora del Grupo de Neurogenética de la Fundación Pública Galega de Medicina Xenómica).

En conformidad, autorizan la presentación del proyecto,

A Coruña, 1 de Julio de 2011

Fdo: Carmen Cadarso Suárez

Fdo: M.Luz Calle Rosingana

Fdo: María Jesús Sobrido Gómez

Resumen

En los estudios de asociación genética es habitual trabajar con un elevado número de variables en relación al número de observaciones. Las variables predictoras no son siempre independientes debido a ciertas peculiaridades de los datos genéticos. En este contexto es necesario incorporar técnicas de selección de variables. En este trabajo exploramos dos aproximaciones, en primer lugar una estrategia en dos pasos, que implica selección de variables en un contexto univariante, evaluando la asociación de cada variable predictora con la respuesta de manera independiente, y la posterior incorporación de las variables seleccionadas para ajustar un modelo de regresión logística multivariante. En segundo lugar aplicamos un método de regresión penalizada –LASSO-. Los modelos obtenidos se compararon en base a su capacidad predictiva y su habilidad para identificar variables causales. Para alcanzar este objetivo realizamos un estudio de simulación, donde se contemplaron diferentes escenarios y adicionalmente empleamos datos reales procedentes de un estudio de asociación de genes candidatos en Enfermedad de Parkinson.

Contenidos

1. Introducción

Los estudios de asociación genética tienen como objetivo identificar factores de susceptibilidad a una determinada enfermedad. En muchas ocasiones estos estudios implican el análisis de un elevado número de variables predictoras, que pude llegar a ser muy superior al número de observaciones. Habitualmente este análisis se realiza en un contexto univariante, identificando aquellas variables informativas empleando un test de asociación. Analizarlos en un contexto multivariante permite tener en cuenta el impacto de unos marcadores sobre otros, pero a su vez plantea diferentes retos: el mencionado problema de la dimensionalidad y el hecho de que estas variables puedan estar correlacionadas. Uno de los métodos paramétricos multi-locus empleados es la regresión logística, aunque debido a estas particularidades de los estudios de asociación no siempre puede aplicarse de manera estándar.

Este trabajo aborda el problema de la dimensionalidad, revisando diferentes técnicas de selección de variables en el contexto de la regresión logística. Para ello se ha llevado a cabo en primer lugar un estudio de simulación y posteriormente se han aplicado esas técnicas a datos procedentes de un estudio de asociación en enfermedad de Parkinson.

Así mediante esta aproximación se ha pretendido satisfacer un doble objetivo:

1. Objetivo metodológico: Comparar dos métodos de selección de variables evaluando la capacidad predictiva de los modelos obtenidos y su habilidad para detectar polimorfismos causales.

2. Objetivo aplicado: La identificación de variantes genéticas implicadas en el desarrollo la enfermedad permitirá avanzar en el conocimiento de las bases moleculares de la enfermedad de Parkinson, con posible impacto en el desarrollo de nuevos fármaco y terapias y además la posibilidad de desarrollar modelos predictivos basados en perfiles genéticos.

2. Estudios de asociación genética

2.1 Descripción de los datos genéticos

El ADN se distribuye a lo largo de los cromosomas, que en la especie humana son 22 autosomas o pares de cromosomas homólogos y un par de cromosomas sexuales. Cada uno de los miembros de un par es heredado de uno de los dos progenitores. Un locus – loci en plural- se refiere a una región específica en un cromosoma. Dado que nuestras células son diploides, hay dos posibles secuencias de ADN heredadas independientemente para un locus determinado y un individuo: alelos. Estos dos alelos forman el genotipo de ese individuo para ese locus en particular.

En la mayor parte del genoma, la secuencia de ADN es idéntica para todos los humanos, por lo tanto se trataría de loci monomórficos. Para una proporción de loci se han detectado una serie de mutaciones que se han mantenido a lo largo del tiempo y que consisten en un cambio de base en las secuencias de ADN, lo que implica que existan diferentes alelos para un locus determinado – loci polimórficos.

En los estudios de asociación poblacionales los polimorfismos más habituales objeto de estudio son los SNP – Single Nucleotide Polymorphisms- pronunciado “snip”. Un SNP describe un cambio en una sola base, una de las bases es sustituida por otra. Para que verdaderamente pueda considerarse un polimorfismo, la variación debe aparecer al menos en el 1% de la población.

Estos SNPs, objeto de nuestro estudio son bialélicos, hay 2 bases posibles en el sitio correspondiente dentro de un gen. De modo que un SNP viene determinado por dos bases, que se denominan alelos. Estas bases, también denominadas nucleótidos, pueden ser de 4 tipos: adenina (A), citosina (C), guanina (G) y timina (T). Aunque en teoría puede haber 4 variaciones diferentes debido a la existencia de 4 tipos, en la práctica, en una posición determinada de la secuencia genómica suelen observarse sólo 2 variantes. (Figura 2.1).

Por lo tanto el genotipo de un individuo para un determinado marcador viene dado por la combinación de los dos alelos en ese locus. Si denominamos los dos posibles alelos como “A” y “a”, asumiendo “a” el alelo más frecuente o wild type y “A” el alelo menos frecuente, tenemos tres posibles genotipos: el genotipo “AA” será el homocigoto para el alelo menos frecuente, “Aa” el heterocigoto y “aa” el homocigoto para el alelo más frecuente.

Figura 2.1 Single Nucleotide Polymorphism (SNPs)

Individuo 1	cromosoma paterno	AGCTTGAC T CCATGA	Homocigoto para el alelo más frecuente
	cromosoma materno	AGCTTGAC T CCATGA	
Individuo 2	cromosoma paterno	AGCTTGAC T CCATGA	Heterocigoto
	cromosoma materno	AGCTTGAC G CCATGA	
Individuo 3	cromosoma paterno	AGCTTGAC G CCATGA	Heterocigoto
	cromosoma materno	AGCTTGAC T CCATGA	
Individuo 4	cromosoma paterno	AGCTTGAC G CCATGA	Homocigoto para el alelo menos frecuente
	cromosoma materno	AGCTTGAC G CCATGA	

Como hemos indicado, para cada SNP hay tres posibles genotipos, y podemos evaluar diferentes modelos genéticos bajo los que contemplar esos genotipos. Un modelo genético describe la relación entre el genotipo de un individuo y el fenotipo o rasgo. En el contexto de los estudios de asociación genética, el fenotipo se refiere a si un individuo está o no afectado por una determinada enfermedad, lo denominamos Y, siendo Y=1 afectados (casos), Y=0 no afectados (controles). En el caso de enfermedades mendelianas los modelos genéticos son determinísticos – el genotipo determina exactamente el fenotipo. En el caso de las enfermedades complejas, la relación entre el genotipo y el fenotipo es probabilística, el genotipo influye en las probabilidades de desarrollar la enfermedad.

Dado el genotipo de un individuo (G), el efecto probabilístico del locus en el fenotipo Y viene descrito por P(Y|G), que se denomina función de penetrancia y define la probabilidad de enfermedad condicionada al genotipo [1].

Los cuatro modelos genéticos más habituales son los siguientes:

Modelo codominante. Es el más general. Se consideran los tres genotipos por separado suponiendo que cada uno proporciona un riesgo de enfermedad diferente.

$$P(Y = 1|AA) \neq P(Y = 1|Aa) \neq P(Y = 1|aa)$$

Modelo dominante. Sólo es necesaria una copia del alelo de riesgo para tener un efecto sobre el genotipo. Supone que una única copia del alelo "A" es suficiente para modificar el riesgo y poseer dos copias del alelo lo modifica en igual magnitud.

$$P(Y = 1|AA) = P(Y = 1|Aa)$$

Modelo recesivo. Son necesarias dos copias del alelo de riesgo para modificar el riesgo de enfermedad.

$$P(Y = 1|aa) = P(Y = 1|Aa)$$

Modelo aditivo. Dependiendo de la escala, la probabilidad del genotipo heterocigoto están entre la de los dos homocigotos. El riesgo asociado al heterocigoto es intermedio entre los dos homocigotos.

En escala lineal:

$$P(Y = 1|Aa) = 0.5(P(Y = 1|AA) + P(Y = 1|aa))$$

En escala log (multiplicativa):

$$P(Y = 1|Aa) = \sqrt{P(Y = 1|AA)P(Y = 1|aa)}$$

Estos genotipos, que constituyen variables categóricas son recodificadas como variables numéricas o indicadoras. En la tabla 2.1 se muestra la codificación de los genotipos bajo los diferentes modelos genéticos.

Genotipo	Codominante	Dominante	Recesivo	Aditivo
aa	0 0	0	0	0
Aa	1 0	1	0	1
AA	0 1	1	1	2

Tabla 2.1 Codificación de genotipos

Equilibrio Hardy-Weinberg (HWE)

El equilibrio HW hace referencia a la independencia de alelos en un locus entre los dos cromosomas homólogos. En concreto establece la relación entre las frecuencias alélicas y genotípicas en una población lo suficientemente grande, sin selección, mutación o migración. Si consideramos un locus con dos alelos "A" y "a" con frecuencias $\mathrm{p}A = p$ y $\mathrm{p}a = 1 - p = q,$ si estos dos alelos que porta un individuo se heredan independientemente, el número de copias del alelo A sigue una distribución binomial Bin(2,p). Por lo tanto las probabilidades de los genotipos AA, Aa y aa serán p^2, $2pq$ y q^2 respectivamente. La probabilidad de que un alelo aparezca en un homólogo es independiente de que alelo esté presente en el segundo homólogo.

Dados los genotipos, las frecuencias alélicas observadas y esperadas se pueden comparar empleando diferentes estadísticos (Chi Cuadrado, Test de Fisher,..).

En un estudio de asociación es necesario verificar que los SNPs genotipados se encuentran en equilibrio HW, ya que desviaciones de este equilibrio pueden ser indicativas de errores de genotipado o bien de la existencia de factores que podrían afectar a las frecuencias alélicas (por ejemplo, migración reciente) diferentes del rasgo en estudio. Si no se examina puede dar lugar a asociaciones espurias. La comprobación del HWE se realiza sobre la población control, ya que desviaciones del HWE en casos pueden ser indicativos de una asociación.

Desequilibrio de Ligamiento

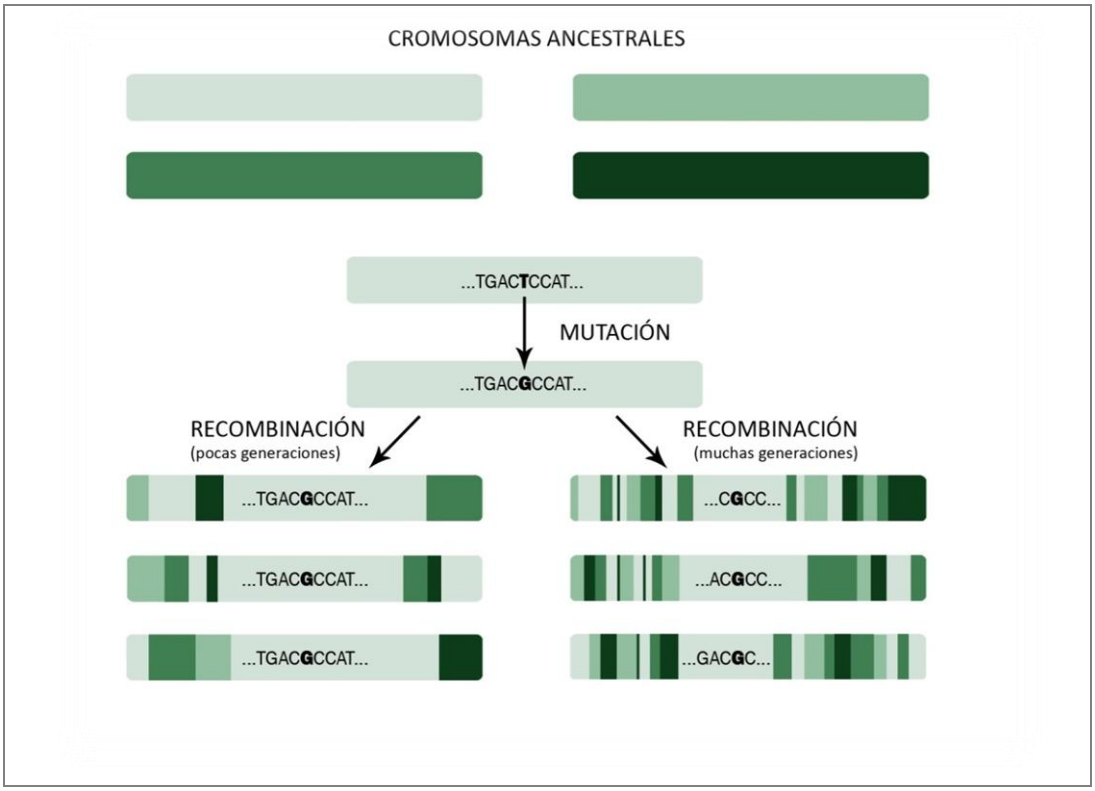
El Equilibrio Hardy-Weinberg hacía referencia a la independencia de alelos en un locus entre cromosomas homólogos, El Desequilibrio de Ligamiento (Linkage disequilibrium

LD) se refiere a la asociación ente alelos de diferentes loci en uno de los cromosomas homólogos.

LD hace por lo tanto referencia a la disposición no casual de alelos en dos loci, de modo que estos alelos serán heredados conjuntamente a lo largo de múltiples generaciones. En la figura 2.2 se ilustra este concepto. Consideramos un locus marcador (**T**). En un determinado momento ocurre una nueva mutación en el genoma. El cromosoma en el que se produce la mutación contiene el background genético de la mutación. Al transmitirse a la descendencia a lo largo de varias generaciones, el fenómeno de recombinación provocará que alelos en loci marcadores que están lejos de la mutación se intercambien. Como resultado el background genético asociado a una mutación será cada vez más pequeño. En general cuanto más cerca está el marcador de la nueva variante, más fuerte será el LD. Cuando el marcador y la nueva variante están muy juntos, la recombinación habrá ocurrido a una tasa tan baja que el alelo marcador y la nueva variante cosegregarán, incluso entre familias [2].

El concepto de LD es fundamental en los estudios de asociación genética, ya que implica que no tenemos que llegar a genotipar la variante causal para detectar asociación, simplemente tenemos que genotipar muy cerca de la verdadera variante.

Figura 2.2 Desequilibrio de ligamiento



Por lo tanto es de especial interés determinar si un grupo de alelos están en LD. Mediante la caracterización de regiones con alto LD (LD en una región con múltiples SNPs puede determinarse simplemente como la media de todas las medidas de LD dos a dos), el genoma puede dividirse en bloques de desequilibrio de ligamiento (bloques haplotípicos).

El proyecto Internacional HapMap [3] es una herramienta indispensable en los estudios de asociación. En este proyecto se genotiparon inicialmente 269 individuos de 4 poblaciones con diferente origen étnico. Proporciona un mapa de LD de millones de SNPs a lo largo del genoma y permite, a la hora de diseñar un estudio de asociación, seleccionar para una región del genoma de interés tagSNPs que cubran un bloque haplotípico. El concepto de tag-SNPs se ilustra en la figura 2.3.

Figura 2.3 Tag- SNPs según International HapMap Consortium [3]

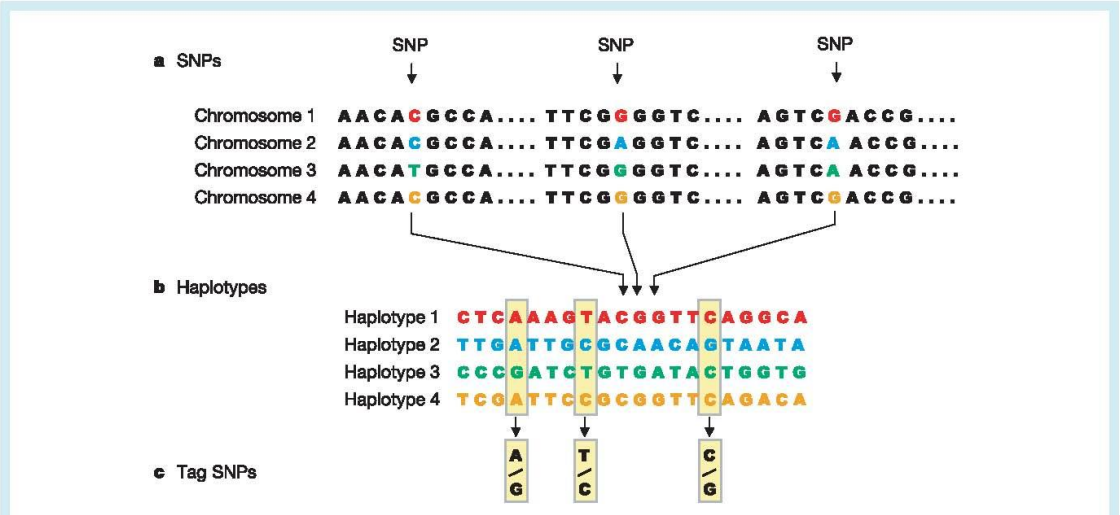
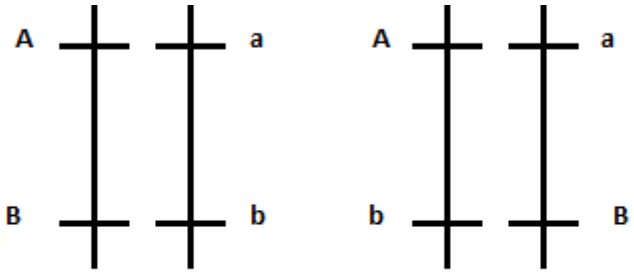


Figure 1 SNPs, haplotypes and tag SNPs. **a**, SNPs. Shown is a short stretch of DNA from four versions of the same chromosome region in different people. Most of the DNA sequence is identical in these chromosomes, but three bases are shown where variation occurs. Each SNP has two possible alleles; the first SNP in panel **a** has the alleles C and T. **b**, Haplotypes. A haplotype is made up of a particular combination of alleles at nearby SNPs. Shown here are the observed genotypes for 20 SNPs that extend across 6,000 bases of DNA. Only the variable bases are shown, including the three SNPs that are shown in panel **a**. For this region, most of the chromosomes in a population survey turn out to have haplotypes 1–4. **c**, Tag SNPs. Genotyping just the three tag SNPs out of the 20 SNPs is sufficient to identify these four haplotypes uniquely. For instance, if a particular chromosome has the pattern A–T–C at these three tag SNPs, this pattern matches the pattern determined for haplotype 1. Note that many chromosomes carry the common haplotypes in the population.

Existen diferentes estadísticos para calcular LD. Si consideremos dos loci bialélicos, de modo que el primer locus tiene los alelos “A” y “a” y el segundo locus los alelos “B” y “b”, existen 4 haplotipos posibles: “AB”, “ab” ,”Ab” y “aB” como se muestra en la figura 2.4.

Figura 2.4 Haplotipos en dos loci bialélicos



Las frecuencia haplotípicas se muestran en la tabla 2.2. Esta tabla reflejaría las frecuencias para cada haplotipo en el caso de que los loci fuesen independientes, donde $p_{AB} = p_A p_B$, $p_{Ab} = p_A p_b$, $p_{aB} = p_a p_B$ y $p_{ab} = p_a p_b$ representan la probabilidad de observar cada haplotipo y p_A, p_a, p_B y p_b representan la freucencia de cada alelo. Desviaciones de esta tabla serían una evidencia de correlación.

	A	a	
B	p_{AB}	p_{aB}	p_B
b	p_{Ab}	p_{ab}	p_b
	p_A	p_a	1

Tabla 2.2. Frecuencias haplotípicas bajo supuesto de loci independientes

Lewontin & Kojima [4] propusieron el coeficiente de desequilibrio de ligamiento *D* como:

$D = p_{AB} - p_A p_B$, o equivalentemente

$D = p_{AB} p_{ab} - p_{AB} p_{aB}$

El desequilibrio de ligamiento es por lo tanto una medida de la asociación entre dos loci que captura cómo de aleatoriamente se distribuyen los alelos. El estadístico D presenta el inconveniente de que su magnitud depende de las frecuencias de los

alelos en los dos loci. Por este motivo Lewontin [5] estandarizó D al máximo valor posible que puede tomar:

$D' = D/Dmax$, donde $Dmax$ es el máximo valor de D para las frecuencias alélicas dadas.

$Dmax = min(p_A p_B, p_a p_b)$ si $D < 0$ y $min(p_A p_b, p_a p_B)$ si $D > 0$

D' toma valores entre 0 y 1 y permite establecer el grado de desequilibrio de ligamiento relativo al máximo valor posible que puede tomar. Normalmente se emplea el valor absoluto de D'

Otra medida de desequilibrio de ligamiento la proporciona r^2, coeficiente de correlación [6]:

$r^2 = D^2/(p_A p_a \mathrm{p_B} p_b)$

r^2 varía entre 0 y 1, $r^2 = 1$ implica LD perfecto , indicando una relación determinista. Por lo tanto el genotipado de dos SNPs en alto grado de desequilibrio de ligamiento es redundante.

2.2 Enfermedades complejas

Las enfermedades mendelianas son aquellas causadas por mutaciones en un único gen y con un modelo de herencia simple. Estas mutaciones generalmente presentan frecuencias bajas en la mayoría de poblaciones y además presentan un espectro alélico complejo, de modo que múltiples mutaciones en un único gen pueden conducir al mismo fenotipo. Las enfermedades complejas no muestran herencia mendeliana atribuible a un único locus. Frente a las enfermedades monogénicas, las enfermedades complejas están causadas por la acción de múltiples loci, cada uno con un pequeño efecto, además estos loci pueden interaccionar entre ellos y a su vez interaccionar con factores ambientales.

El objetivo de los estudios de asociación es identificar esos loci y caracterizar las complejas relaciones que se establecen entre ellos.

Los estudios de asociación trabajan bajo la hipótesis “Enfermedad Común – Variante Común” [7,8] bajo la cual los factores genéticos de susceptibilidad a enfermedades complejas serán alelos comunes en la población. Actualmente, los resultados poco concluyentes de los estudios de asociación realizados hasta la fecha, que en su mayoría tan sólo han permitido identificar variantes con tamaños de efecto muy pequeños, llevan a pensar que es probable que las enfermedades complejas estén causadas por una conjunción de variantes comunes y raras [9,10]. Identificar todas las variantes que contribuyen al riesgo de desarrollar una determinada enfermedad sigue siendo un reto.

2.3 Tipos de Estudios de Asociación

Existen dos aproximaciones a los estudios de asociación genética

- Estudios poblacionales

Estudios caso-control de individuos no relacionados. Se identifica un set de casos diagnosticados según unos criterios especificados (por ejemplo para un estudio en migraña los criterios de la International Headache Society, en enfermedad de Parkinson los criterios del PDBank) y una muestra independiente de individuos no afectados de la misma población. Se comparan las frecuencias genotípicas y/o alélicas en ambos grupos para identificar variantes asociadas con la enfermedad objeto de estudio.

Dos de los principales inconvenientes de este tipo de estudios son 1) la dificultad para encontrar variantes raras y 2) la posibilidad de falsos positivos por problemas de estratificación en la población (población está dividida en diferentes estratos y existen probabilidades distintas de ser seleccionado como caso o control según el estrato).

- Estudios basados en familias.

Emplean información de miembros de una misma familia. Se comparan las frecuencias de alelos transmitidos vs. no transmitidos. Tienen la ventaja de que eliminan el problema de estratificación, como inconveniente tienen menor poder estadístico (TDT, también FBAT que incorpora información de otros miembros de la familia).

Requiere consideraciones estadísticas específicas. Es importante el concepto de fase alélica - alineamiento de nucleótidos en uno de los cromosomas homólogos. Ésta se puede determinar al contener información de miembros de una familia.

En función del número de SNPs analizados y la hipótesis de partida podemos diferenciar entre:

- Estudios de genes candidatos: Se analizan aquellos genes sobre los que existe una hipótesis *a priori* acerca de la funcionalidad, de modo que existe una cierta evidencia experimental de que una determinada ruta biológica puede estar involucrada en el desarrollo de una enfermedad. Normalmente se seleccionan varios SNPs para cada gen. Para la selección de SNPs se emplea el concepto de desequilibrio de ligamiento, que se ha explicado en el apartado anterior, mediante el empleo de tag-SNPs.

- Genome Wide Association Studies (GWAS). Son estudios de asociación de todo el genoma. En este caso no hay hipótesis previa. Se emplean chips de genotipado genéricos que cubren parte o la totalidad del genoma, con mayor o menor cobertura 500k – varios millones de SNPs. El genoma humano comprende aproximadamente $3*10^9$ bases, de modo que se puede capturar gran parte de la variabilidad del genoma mediante el concepto de bloques de desequilibrio de ligamiento.

2.4 Problemática de los estudios de asociación

La premisa de los estudios de asociación es que son variantes comunes (SNPs con frecuencias superiores al 1%) las que modifican el riesgo en la mayoría de enfermedades complejas. A partir de los resultados de GWAS, la mayoría de las variantes que han mostrado asociación significativa con un determinado fenotipo presentan OR entre 1.1 y 1.3.

Estos SNPs constituyen variantes de predisposición (o están en LD con dichas variantes) que aumentan el riesgo de los portadores en un 10-30% respecto al riesgo de los no portadores. Existe la posibilidad de que haya otras variantes no detectadas porque aumentan el riesgo en valores más pequeños, quizás tan bajos como el 1%, La alternativa de que el incremento en el riesgo esté provocado por múltiples variantes

raras con frecuencias inferiores haría que fuesen muy difíciles de identificar con este tipo de diseño.

De la multitud de GWAS realizados hasta la fecha, sólo se han identificado unas pocas variantes comunes implicadas y los SNPs asociados explican tan sólo una pequeña fracción del riesgo genético, incluso si se consideran como un agregado.

Existen por lo tanto, serias dificultades para descifrar el componente genético de las enfermedades complejas. Algunas de las razones de las limitaciones de los GWAS para detectar factores genéticos de susceptibilidad son las siguientes:

- Variantes raras. Que las variantes de susceptibilidad sean variantes que aparecen en menos del 1% de la población, podrían aparecer tan sólo en algunos individuos o familias, por lo tanto estas variantes sólo se observarían en una pequeña fracción de los afectados.
- Heterogeneidad locus. Cualquiera de un número de genes o loci confiere susceptibilidad a la enfermedad de manera independiente.
- Heterogeneidad alélica. Diferentes alelos en un mismo gen están asociados con la enfermedad de manera independiente.
- Epistasis. Ser portador de un determinado genotipo conferirá susceptibilidad hasta un grado dictado por la presencia de otros genotipos.
- Interacciones gen-ambiente. Determinados gen o genes tienen efecto sólo en presencia de determinados factores ambientales.
- Herencia poligénica. Un número de variantes en diferentes loci tienen que presentarse antes de que tenga un efecto sobre el riesgo de enfermedad. Un fenotipo puede venir determinado por múltiples variantes de pequeño efecto a lo largo de una ruta biológica.

2.5 Contexto del trabajo. Antecedentes

Uno de los principales problemas a abordar en los estudios de asociación genética, principalmente GWAS, es el de la dimensionalidad (“curse of dimensionality”) [11], existe un elevado número de variables predictoras en relación al número de observaciones. Estas variables están potencialmente correlacionadas y además

pueden interactuar (en un sentido biológico o estadístico) en su asociación con la variable respuesta.

En los últimos años se han desarrollado nuevas técnicas o adaptado otras existentes para esto tipo de datos, que permiten abordar el problema de la dimensionalidad (CART, Randomforest, LogicRegression, MDR,…). Algunas de ellas permiten reducir la dimensión, otras buscar interacciones, y otras realizan las dos acciones de manera simultánea.

Una de las aproximaciones para abordar este problema es mediante técnicas de selección de variables. Esta selección de variables se puede hacer empleando métodos univariantes o multivariantes.

1) Univariantes. Es la aproximación más habitual en este tipo de estudios. Se establecen unos criterios para medir la asociación de cada una de las variables predictoras con la respuesta, determinar las más significativas en base a un umbral preestablecido. Esta aproximación univariante de selección de variables presenta varios inconvenientes:

- Analizar las distintas variables predictoras en un contexto univariante no permite contemplar la existencia de posibles correlaciones. Además, de este modo no se pueden detectar patrones complejos de manera que no es posible medir el efecto combinado de diferentes loci y se infraestima la contribución genética en presencia de interacciones [12]..

- El problema de las comparaciones múltiples. Al estudiar la asociación de diferentes marcadores de manera independiente la probabilidad de error tipo I se incrementa y hay que realizar ajustes para corregir el nivel de significación (Bonferroni, False Discovery Rate (FDR),..). Esta corrección implica que la probabilidad de detectar un efecto cuando realmente está presente disminuya, de modo que se pueden estar perdiendo asociaciones débiles.

2) Multivariantes. Las variables predictoras son consideradas en un contexto multivariante. Permiten a su vez la posibilidad de incorporar interacciones entre las variables.

En este trabajo compararemos estas dos aproximaciones en el contexto de regresión logística múltiple. En primer lugar emplearemos un método de selección univariante

mediante una estrategia en dos pasos: primero se seleccionan las variables en función de los resultados de un test de asociación individual, a continuación aquellas variables que presentan asociación significativa con la variable respuesta serán incorporaradas a un modelo de regresión logística multivariante. En segundo lugar emplearemos un método particular de regresión penalizada, Least Absolute Shrinkage and Selection Operator (LASSO)[13]. La regresión penalizada ha sido aplicada en diferentes estudios recientes de asociación genética [14,15,16]. Este método supone una alternativa a los modelos de regresión habituales, que bajo los supuestos mencionados (elevado número de predictoras en relación al número de observaciones unido a la posibilidad de correlación entre variables debido al desequilibrio de ligamiento) suelen presentar problemas.

3. Métodos de selección de variables y evaluación de modelos

3.1. Selección de variables mediante test de asociación

En un estudio de asociación genética caso-control la información del genotipo de los individuos se puede resumir en una tabla 2 x k, con k=3 genotipos posibles, como la que se muestra en la tabla 3.1.

Existen diferentes test de asociación para analizar este tipo de tablas, es decir, para evaluar la asociación individual de cada SNP con la respuesta.

	aa	Aa	AA	Total
Casos	r_0	r_1	r_2	r
Controles	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

Tabla 3.1 Distribución de genotipos en un estudio caso-control

3.1.1. Test Chi Cuadrado

Siguiendo la notación de la Tabla 3.1, y siendo $x_i = i$ el número de alelos A que posee un individuo, con $i = 0,1,2,$ el test χ^2 viene determinado por:

$$\chi^2 = \sum_{i=0}^{2} \frac{(r_i - n_i r/n)^2}{n_i r/n} + \sum_{i=0}^{2} \frac{(s_i - n_i s/n)^2}{n_i s/n}$$

Este estadístico sigue una distribución Chi cuadrado con dos grados de libertad (χ_2^2).

En este caso hemos supuesto un modelo codominante, para testar otros modelos genéticos crearíamos tablas de contingencia 2x2 agrupando las columnas. Del mismo modo si quisiéramos emplear las frecuencia alélicas en lugar de genótipicas también obtendríamos una tabla de contingencia 2x2 como se muestra en la tabla 3.2.

	aa	AA	Total
Casos	$2r_0 + r_1$	$r_1 + 2r_2$	2r
Controles	$2s_0 + s_1$	$s_1 + 2s_2$	2s
Total	$2n_0 + n_1$	$n_1 + 2n_2$	2n

Tabla 3.2 Distribución de alelos en un estudio caso-control

3.1.2. Cochran Armitage Trend Test

El test de tendencia de Cochran-Armitage [17] comprueba la tendencia en proporciones binomiales a lo largo de niveles de un factor o covariable bajo un modelo aditivo:

$$P(Y = 1|X) = \beta_0 + \beta_1 X$$

Se emplea en el caso de tablas de contingencia en donde una variable tiene 2 niveles y la otra variable es ordinal. Aplicado a estudios de asociación genética que emplean genotipos, comprueba si la probabilidad de enfermedad aumenta linealmente con el genotipo.

Este test se emplea por lo tanto con datos categóricos para establecer la presencia de asociación entre una variable con 2 categorías y otro variable con k categorías. Modifica el test Chi Cuadrado para incorporar un supuesto orden en los efectos de las k categorías sobre la segunda variable. Se aplica este test a datos de la forma de una tabla de contingencia 2x k (en nuestro caso k=3, número posible de genotipos) como la mostrada en la tabla 3.1. Siguiendo esta notación adoptamos la descripción del estadístico [18]:

Siendo $x_i = i$ el número de alelos A que posee un individuo, para $i = 0,1,2$ el test de Cochran Armitage se puede expresar como $U^2/Var(U)$, donde

$$U = \frac{1}{n}\sum_{i=0}^{2} x_i\,(sr_i - rs_i)$$

Bajo la hipótesis nula de no asociación:

$$\widehat{var(U)} = \frac{rs}{n^3}\left[n\sum_{i=0}^{2} x_i^2 n_i - \left(\sum_{i=0}^{2} x_i^2 n_i\right)^2\right]$$

La distribución asintótica de $Z = U/\sqrt{Var(U)}$ es $N(0,1)$, de modo que el estadístico es Z^2, que sigue una distribución χ_1^2

3.1.3. Likelihood Ratio Test

El Likelihood Ratio Test (LRT) o Test de Razón de Verosimilitud compara las log-verosimilitudes de dos modelos: m_1 (modelo completo) y m_0 (modelo reducido o nulo):

$$LRT = 2(\log Likelihood_{(m_0)} - \log Likelihood_{(m_1)})$$

Este estadístico sigue una distribución χ^2 con grados de libertad igual a la diferencia el número de parámetros entre los dos modelos.

En este caso compara las log-verosimilitudes del modelo nulo y del modelo con cada una de las variables (SNPs).

3.2. Selección de variables en regresión logística

Los modelos de regresión logística son una herramienta habitual en estudios de epidemiología genética para tratar de identificar factores genéticos de susceptibilidad a una determinada enfermedad.

Partiendo de un modelo de regresión lineal

$$E(Y|x) = \beta_0 + \beta_1 x$$

En el caso de datos dicotómicos emplearemos

$$\pi(x) = E(Y|x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

La transformación logit es definida, en términos de $\pi(x)$, como:

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

La estimación de parámetros se realiza mediante máxima verosimilitud. De manera general este método estima los valores de los parámetros que maximizan la probabilidad de obtener los datos observados. Para aplicar este método empleamos la función de verosimilitud, que expresa la probabilidad de los datos observados como función de los parámetros desconocidos. Los estimadores de máxima verosimilitud son aquellos valores que maximizan están función.

A continuación se describe el proceso de máxima verosimilitud para la regresión logística simple, es decir una sola variables predictora x.

Codificando Y como 0 y 1, la expresión de $\pi(x)$ proporciona, para un valor de $\beta = (\beta_0, \beta_1)$, el vector de parámetros, la probabilidad condicionada de que Y sea igual a 1 dado x, es decir $P(Y = 1|x)$. Del mismo modo 1- $\pi(x)$ nos da la probabilidad condicionada de Y igual a 0 dado x, $P(Y = 0|x)$.

Así para aquellos pares (x_i, y_i) donde $y_i = 1$ la contribución a la función de verosimilitud es $\pi(x_i)$, y para los pares donde $y_i = 0$, la contribución es 1- $\pi(x_i)$. Podemos expresar la contribución del par (x_i, y_i) como

$$\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}$$

Dado que se asume que las observaciones son independientes, la función de verosimilitud se obtiene como el producto de los términos dados en la expresión anterior:

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

El principio de máxima verosimilitud establece que la estimación de $\boldsymbol{\beta}$ será el valor que maximice la anterior expresión. Aplicando logaritmo definimos la log-verosimilitud:

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1-\mathrm{y_i})\ln[1-\pi(x_i)]\}$$

Para encontrar el valor de $\boldsymbol{\beta}$ que maximiza $L(\boldsymbol{\beta})$ diferenciamos $L(\boldsymbol{\beta})$ respecto a β_0 y β_1 igualando las expresiones resultantes a 0. Así se obtienen las ecuaciones de verosimilitud:

$$\sum [y_i - \pi(x_i)] = 0$$

$$\sum x_i [y_i - \pi(x_i)] = 0$$

En el caso de regresión logística múltiple, donde tenemos un número p de variables independientes,$x' = (x_1, x_2, \dots x_p)$, la función logit viene dada por la ecuación

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Y el modelo de regresión logística:

$$\pi(x) = \frac{e^{g_x}}{1 + e^{g_x}}$$

Por lo tanto la función de verosimilitud es prácticamente idéntica a la descrita, tan solo cambia la definición de $\pi(x)$. Así, tendremos $p + 1$ ecuaciones de verosimilitud que se obtienen diferenciando la función de log-verosimilitud respecto a los $p + 1$ coeficientes. Las ecuaciones serán:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0$$

$$\sum_{i=1}^{n} x_{ij} [y_i - \pi(x_i)] = 0$$

Para $j = 1,2, \dots, p$

A continuación se describen diferentes métodos de selección de variables en modelos de regresión logística:

3.2.1 Best subset

Encuentra para cada K ∈ {0,1,2,…,p} el subconjunto de tamaño K que proporciona el mejor modelo. Los modelos obtenidos se comparan con el modelo completo con todas las variables empleado LRT, Wald test, Mallow`s Cq, AIC)

Hosmer & Lemeshow [19] propusieron un algoritmo que proporciona un estadístico para cada posible combinación de variables predictoras. De este modo se puede estimar el mejor subconjunto. Este procedimiento permite identificar un grupo de subconjuntos que proporcionan el mejor valor bajo un criterio específico sin que sea necesario ajustar todos los modelos de regresión con todos los subconjuntos.

Para escoger K existen diferentes criterios (el modelo más pequeño que minimiza una estimación del error estándar de predicción esperado). Se puede emplear validación cruzada para escoger K.

Una limitación de esta metodología es que buscar todos los subconjuntos cuando p es mayor que 40 es inviable

3.2.2. Regresión stepwise

Engloba una serie de procedimientos de selección automática de variables significativas basados en la inclusión o exclusión de las mismas en el modelo de una manera secuencial. Se pueden dividir los algoritmos en 2 categorías:

- Selección “forward stepwise”: el algoritmo comienza con el intercept, de modo que todas las predictoras están excluidas del modelo, e incorpora secuencialmente en el modelo aquella variable predictora que mejora el ajuste.

- Eliminación “backward stepwise”: modelo completo que incluye todas las variables predictoras y a cada paso elimina aquellas con menor impacto en el ajuste hasta una determinada regla de parada.

- Selección “stepwise”: este método es una combinación de los dos procedimientos anteriores, comienza como la regresión forward, pero en cada paso se plantea si todas las variables introducidas en el modelo deben de permanecer. El algoritmo termina cuando ninguna variable entra o sale del modelo.

La elección de qué variables introducir o eliminar en el modelo en cada paso viene dado por LRT, score test, Wald Test, AIC,…

3.2.3. Regresión penalizada

Una alternativa a los métodos descritos anteriormente que mantienen un subconjunto de variables predictoras y descartan el resto son los métodos de regresión penalizada.

La idea clave es la penalización, se evita el sobreajuste debido al gran número de variables predictoras imponiendo una penalización sobre fluctuaciones grandes de los parámetros estimados. La elección del parámetro de penalización ($\boldsymbol{\lambda}$) es fundamental, es necesario un procedimiento que estime el valor del parámetro $\boldsymbol{\lambda}$ a partir de los datos.

A continuación se describen tres métodos de regresión logística penalizada:

3.2.3.1 LASSO

En un principio desarrollada parara modelos lineales, Least Absolute Shrinkage and Selection Operator (LASSO) se introdujo para regresión logística [20] y se aplicó en el contexto de estudios de expresión génica [21] y más recientemente en GWAS [14,15]. LASSO combina “shrinkage” y selección de variables, imponiendo una penalización sobre los coeficientes de la regresión, de modo que para valores altos del parámetro de penalización algunos de estos coeficientes se fijan a 0.

LASSO minimiza la log-verosimilitud negativa, sujeta a una penalización en los coeficientes de regresión. Así, siendo β el vector de coeficientes de regresión y L la función log-verosimitud negativa, el estimador lasso se define como:

$$\hat{\beta}_{LASSO} = \underset{\beta}{argmin} \left\{ L(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Donde $\lambda \geq 0$ determina la cantidad de “contracción”. La ventaja es que puede generar modelos dispersos, donde la mayoría de los coeficientes son igual a 0. Se denomina penalización L_1.

3.2.3.2 Ridge

Fue introducida en 1970 por Hoerl & Kennard [22]. Igual que el caso del método LASSO se empleó en primer lugar en el contexto de estudios de expresión génica [23] y posteriormente en GWAS para tratar con situaciones de desequilibrio de ligamiento entre SNPs [24].

En este caso emplea una penalización L_2 , de modo que el estimador de la regresión ridge es:

$$\hat{\beta}_{RR} = \underset{\beta}{argmin} \left\{ L(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

El parámetro de penalización $\lambda \geq 0$ determina la fuerza de la penalización ($\lambda = 0$ no hay contracción y $\lambda \to \infty$ resulta en todos los parámetros a 0). En este caso no puede inferir modelos dispersos.

3.2.3.3. Elastic net

Propuesta por Zou & Hastie [25] se trata de un método de regularización y selección de variables. También ha sido empleado en GWAS y estudios de asociación de genes candidatos [26]. Considerando las penalizaciones L_1: $\sum_{ij}^{p} |\beta_j|$ y L_2: $\sum_{ji}^{p} \beta_j^2$, elastic net es una combinación de las penalizaciones LASSO y ridge:

$$\hat{\beta}_{EN} = \underset{\beta}{argmin} \left\{ L(\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}$$

La penalización puede expresarse de la forma:

$$\lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha) |\beta_j| \right)$$

De modo que supone un compromiso entre la penalización ridge (α=0) y la penalización lasso (α=1).

3.3. Validación cruzada

La validación cruzada es una aproximación que nos permite evaluar y seleccionar modelos. Se emplea un set de entrenamiento (training set) para ajustar un modelo o clasificador y un set de prueba (test set) para evaluar su capacidad predictiva, mediante el error de predicción u otra medida.

La forma en que se aplica la validación cruzada es mediante la división del set de datos disponible de manera aleatoria en k subconjuntos de igual tamaño y mutuamente excluyentes. A continuación se ajusta el modelo k-veces dejando fuera cada una de las veces uno de los subconjuntos, este subconjunto omitido se emplea para evaluar el modelo y computar un criterio de error establecido. (Uno de estos criterios puede ser error de clasificiación, AUC…). La validación cruzada con k=10 es una de las más utilizadas, pero hay que tener en cuenta el número de observaciones del que disponemos.

3.4. Curvas ROC

Si consideramos un clasificador o marcador diagnóstico de modo que, sabiendo el estatus de un individuo,Y=1 enfermo (E) , Y=0, sano $(\bar{E})$, nos dé una predicción de ese estatus $\hat{Y}$, podemos obtener los resultados que se muestran en la tabla 3.3:

		Y		
		1	0	
$\hat{Y}$	1	Verdadero Positivo	Falso Positivo	VP+FP
	0	Falso Negativo	Verdadero Negativo	FN+VP
		VP+FN	FP+VN	

Tabla 3.3. Posibles resultados de un clasificador o marcador diagnóstico

A partir de la tabla anterior se pueden calcular la sensibilidad y especificidad de un modelo predictivo:

Sensibilidad = $\frac{VP}{VP+FN}$ = fracción de verdaderos positivos (FVP)

Especificidad = $\frac{VN}{FP+VN}$ = fracción de verdaderos negativos (FVN)= 1 – fracción de falsos positivos (FFP)

Un clasificador discreto es el que produce una salida que representa sólo la clase de pertenencia. Este clasificador produce un par (FFP,FVP) correspondiente a un único punto en el espacio ROC. Si empleamos un clasificador probabilístico que indica la probabilidad de pertenencia a una de las clases (u otro tipo de marcador continuo), se puede emplear un punto de corte o umbral de modo que si:

$Y \geq c \rightarrow$ resultado positivo
$Y < c \rightarrow$ resultado negativo,

siendo *c* un determinado valor de corte que permite hacer la dicotomización anterior. Por lo tanto, la decisión del test dependerá de los distintos valores de *c*.

Así, cada valor de *c* da lugar a un test binario, para el que se podrá determinar su sensibilidad y especificidad. Cada punto de corte origina un punto diferente en el espacio ROC, que representa un par (FFP; FVP) correspondiente a un nivel de decisión determinado.

Representando el conjunto de pares para todos los posibles valores de *c*

$$\{(\mathrm{FFP(c)}, \mathrm{FVP(c)}), \mathrm{c} \in (-\infty, \infty)\}$$

la curva ROC [27] nos proporciona una representación de la capacidad discriminatoria del test.

La expresión de las curvas ROC empleando las funciones de supervivencia de sanos y enfermos es:

$$ROC(t) = S_E\left(S_{\bar{E}}^{-1}(t)\right), t \in (0,1)$$

donde S_E y $S_{\bar{E}}$ son:

$$S_E(y) = P[Y \geq c|E] = FVP(c)$$

$$S_{\bar{E}}(y) = P[Y \geq c \,|\, \bar{E}] = FFP(c)$$

El área bajo la curva ROC o AUC (Area Under the Curve) se define como:

$$AUC = \int_0^1 ROC(t)dt$$

El AUC es una medida de la capacidad diagnóstica del test o clasificador, los valores oscilan entre 0.5 y 1, de manera que un valor de AUC=0.5 indica clasificación al azar y un valor de AUC=1 clasificador perfecto.

4. Estudio de simulación

4.1. Descripción del estudio de simulación

Se realizó un estudio de simulación con el objetivo de comparar los dos métodos de selección de variables propuestos: selección basada en análisis de asociación univariante vs. selección basada en regresión penalizada.

Para este estudio se han utilizado los datos simulados por Calle et al. [28].Los detalles de la simulación se encuentran en la anterior referencia. A continuación se describen las características principales de estos datos y de los distintos escenarios considerados:

Se generó un set de k SNPs causales y 1000-k SNPs no causales, siguiendo una estrategia similar a Janseens et al.[29]. Esta estrategia asume independencia entre SNPs por lo que no se dan situaciones de Desequilibrio de Ligamiento. Además asume un modelo genético multiplicativo.

Otras asunciones de la simulación son las siguientes: todos los SNPs causales están en equilibrio HW, tienen el mismo tamaño de efecto en la respuesta y las mismas frecuencias genotípicas. En este caso, para cada SNP causal se fijó el RR del heterocigoto (RR1), siendo el riesgo relativo del homocigoto para el alelo más frecuente $RR2=(RR1)^2$

En este estudio se han contemplado nueve escenarios diferentes, considerando distintas prevalencias de la enfermedad (p) y diferentes frecuencias del alelo menor (MAF). Asimismo se han mantenido constantes el Riesgo Relativo (RR) y el tamaño muestral, tratándose en todos los casos de un estudio balanceado, donde el número de casos (N1) y controles (N0) es igual a 2000. Para todos los posibles escenarios el número de SNPs causales (k) se fijó en 10. En la tabla 4.1 se detallan los 12 escenarios.

Para cada uno de estos escenarios se generaron dos sets de datos: un learning set y un test set (de igual tamaño). El learning set se empleó para construir los modelos y el test set para valorar la capacidad predictiva de los modelos obtenidos.

Escenario	P	MAF	RR	K	N0	N1
I	0.01	0.1	1.3	10	2000	2000
II	0.01	0.2	1.3	10	2000	2000
III	0.01	0.3	1.3	10	2000	2000
IV	0.1	0.1	1.3	10	2000	2000
V	0.1	0.2	1.3	10	2000	2000
VI	0.1	0.3	1.3	10	2000	2000
VII	0.2	0.1	1.3	10	2000	2000
VIII	0.2	0.2	1.3	10	2000	2000
IX	0.2	0.3	1.3	10	2000	2000
X	0.3	0.1	1.3	10	2000	2000
XI	0.3	0.2	1.3	10	2000	2000
XII	0.3	0.3	1.3	10	2000	2000

Tabla 4.1 Escenarios contemplados en el estudio de simulación

Así, para cada escenario se generaron 100 learning sets con los correspondientes 100 test sets. Cada uno de estos sets contiene genotipos de 1000 SNPS, codificados como 0,1 y 2 en función del número de alelos de riesgo (considerando como alelo de riesgo el alelo menos frecuente) para un total de 4000 observaciones, 2000 casos y 2000 controles

Para el método de selección de variables empleando test de asociación (univariante) se aplicó un procedimiento en dos etapas:

1) Empleamos el Cochran Armitage Trend Test descrito en el apartado 3.1.2. e implementado en el paquete *Rassoc* [30,31], especialmente diseñado para estudios de asociación genética. En este caso hemos considerado un modelo genético aditivo. Para identificar las variables que presentan asociación con la variable respuesta se seleccionaron dos niveles de significación:

$\alpha=0.05$, que se correspondería al nivel de significación establecido para evaluar la asociación de un SNP en particular y

$\alpha=0.05/1000=0.00005$, teniendo en cuenta una corrección para comparaciones múltiples. En este estudio estamos empleando la misma muestra para evaluar la asociación en 1000 SNPs, dado que estos SNPs son independientes podemos aplicar

corrección para comparaciones múltiples empleando el método de Bonferroni, corrigiendo así el nivel de significación en función del número de test realizados.

2) Las variables seleccionadas para cada uno de los niveles de significación se incluyeron en un modelo de regresión logística multivariante.

Para evaluar la capacidad predictiva de los modelos de regresión ajustados empleamos las curvas ROC. El modelo de regresión constituye un clasificador probabilístico, ya que nos proporciona las probabilidades de pertenencia a la clase I (Y=1, casos). Empleamos como indicador el AUC con los correspondientes IC al 95% [32].

Para el método de selección de variables mediante regresión penalizada empleamos el paquete *glmnet* [33]. Seleccionamos el valor del parámetro de penalización mediante validación cruzada, mediante la función *cv.glmnet*. En concreto, hemos realizado una validación cruzada (5 fold) de modo que se selecciona el mayor valor de lambda tal que el error está dentro de 1 error estándar del mínimo

Para cada uno de los escenarios y de los métodos considerados se han calculado:1) los AUC promedio de las 100 simulaciones, 2) la media y desviación estándar del número de SNPs seleccionados en cada simulación y 3) el número de veces que son seleccionados cada uno de los 10 SNPs causales a lo largo de todas las simulaciones.

4.2 Resultados y discusión

Los resultados para los distintos escenarios se muestran en las tablas 4.2 – 4.13. Contienen información del AUC promedio de las 100 simulaciones, el promedio de SNPs seleccionados, la frecuencia (absoluta) con la que son seleccionados cada uno de los 10 SNPs causales, así como la frecuencia promedio para los 10 SNPs causales.

ESCENARIO I **p=0.01 MAF=0.1**		Selección pvalue		Selección regresión LASSO
		α=0.05	α=0.00005	λ=1.se
AUC		0.559	**0.557**	**0.575**
Promedio SNPs seleccionados en cada simulación (sd)		59.29 (6.63)	**3.78 (1.45)**	**23.77 (64.61)**
Frecuencia con la que son seleccionados los SNPs causales	SNP1	98	36	79
	SNP2	97	45	80
	SNP3	96	40	80
	SNP4	98	32	74
	SNP5	95	41	78
	SNP6	97	39	77
	SNP7	96	34	73
	SNP8	98	48	80
	SNP9	94	30	73
	SNP10	95	37	71
Frecuencia promedio para los 10 SNPs causales (sd)		96.40 (1.43)	**37.20 (5.14)**	**76.30 (3.27)**

Tabla 4.2. Resultados para el escenario I

ESCENARIO II **p=0.01 MAF=0.2**		Selección pvalue		Selección regresión LASSO
		α=0.05	α=0.00005	λ=1.se
AUC		0.594	**0.620**	**0.617**
Promedio SNPs seleccionados en cada simulación (sd)		58.07 (6.78)	**7.90 (1.30)**	**13.46 (15.00)**
Frecuencia con la que son seleccionados los SNPs causales	SNP1	100	77	96
	SNP2	100	75	91
	SNP3	100	83	88
	SNP4	100	85	96
	SNP5	100	78	95
	SNP6	100	78	90
	SNP7	100	80	94
	SNP8	100	78	94
	SNP9	100	76	87
	SNP10	100	76	93
Frecuencia promedio para los 10 SNPs causales (sd)		100 (0.00)	**78.60 (3.20)**	**92.40 (3.24)**

Tabla 4.3. Resultados para el escenario II

ESCENARIO III **p=0.01 MAF=0.3**		Selección pvalue		Selección regresión LASSO
		α=0.05	α=0.00005	λ=1.se
AUC		0.611	**0.647**	**0.641**
Promedio SNPs seleccionados en cada simulación (sd)		58.94 (6.46)	**9.47 (0.82)**	**11.53 (4.40)**
Frecuencia con la que son seleccionados los SNPs causales	SNP1	100	92	97
	SNP2	100	89	93
	SNP3	100	97	98
	SNP4	100	86	94
	SNP5	100	93	96
	SNP6	100	98	98
	SNP7	100	98	100
	SNP8	100	95	97
	SNP9	100	97	99
	SNP10	100	93	97
Frecuencia promedio para los 10 SNPs causales (sd)		100 (0.00)	**93.80 (4.02)**	**96.90 (2.13)**

Tabla 4.4. Resultados para el escenario III

ESCENARIO IV **p=0.1 MAF=0.1**		Selección pvalue		Selección regresión LASSO
		α=0.05	α=0.00005	λ=1.se
AUC		0.570	**0.573**	**0.587**
Promedio SNPs seleccionados en cada simulación (sd)		59.58 (7.26)	**4.90 (1.70)**	**14.05 (17.92)**
Frecuencia con la que son seleccionados los SNPs causales	SNP1	99	90	85
	SNP2	99	53	86
	SNP3	97	41	77
	SNP4	100	58	88
	SNP5	96	46	83
	SNP6	100	46	79
	SNP7	98	45	77
	SNP8	97	45	73
	SNP9	98	50	80
	SNP10	98	40	79
Frecuencia promedio para los 10 SNPs causales (sd)		98.20 (1.32)	**48.40 (6.75)**	**80.70 (4.69)**

Tabla 4.5. Resultados para el escenario IV

ESCENARIO V **p=0.1 MAF=0.2**		Selección pvalue		Selección regresión LASSO
		α=0.05	α=0.00005	λ=1.se
AUC		0.606	**0.637**	**0.632**
Promedio SNPs seleccionados en cada simulación (sd)		58.47 (6.12)	**8.84 (0.91)**	**11.98 (4.99)**
Frecuencia con la que son seleccionados los SNPs causales	SNP1	100	87	94
	SNP2	100	85	93
	SNP3	100	92	98
	SNP4	100	88	98
	SNP5	100	90	95
	SNP6	100	93	97
	SNP7	100	85	93
	SNP8	100	84	96
	SNP9	100	90	93
	SNP10	100	87	96
Frecuencia promedio para los 10 SNPs causales (sd)		100 (0.00)	**88.10 (3.07)**	**95.30 (2.00)**

Tabla 4.6. Resultados para el escenario V

ESCENARIO VI **p=0.1 MAF=0.3**		Selección pvalue		Selección regresión LASSO
		α=0.05	α=0.00005	λ=1.se
AUC		0.623	**0.656**	**0.652**
Promedio SNPs seleccionados en cada simulación (sd)		59.94 (6.59)	**9.73 (0.69)**	**11.86 (4.55)**
Frecuencia con la que son seleccionados los SNPs causales	SNP1	100	92	99
	SNP2	100	95	95
	SNP3	100	96	100
	SNP4	100	98	99
	SNP5	100	97	100
	SNP6	100	98	98
	SNP7	100	100	99
	SNP8	100	94	97
	SNP9	100	96	98
	SNP10	100	96	100
Frecuencia promedio para los 10 SNPs causales (sd)		100 (0.00)	**96.20 (2.25)**	**98.50 (1.58)**

Tabla 4.7. Resultados para el escenario VI

ESCENARIO VII p=0.2 MAF=0.1		Selección pvalue		Selección regresión LASSO
		α=0.05	α=0.00005	λ=1.se
AUC		0.585	**0.601**	**0.608**
Promedio SNPs seleccionados en cada simulación (sd)		59.61 (7.17)	**6.81 (1.55)**	**13.72 (12.04)**
Frecuencia con la que son seleccionados los SNPs causales	SNP1	100	59	87
	SNP2	99	69	91
	SNP3	100	66	90
	SNP4	100	59	85
	SNP5	100	76	96
	SNP6	100	71	95
	SNP7	100	61	87
	SNP8	99	72	90
	SNP9	99	67	88
	SNP10	100	76	95
Frecuencia promedio para los 10 SNPs causales (sd)		99.70 (0.48)	**67.60 (6.40)**	**90.40 (3.84)**

Tabla 4.8. Resultados para el escenario VII

ESCENARIO VIII p=0.2 MAF=0.2		Selección pvalue		Selección regresión LASSO
		α=0.05	α=0.00005	λ=1.se
AUC		0.622	**0.658**	**0.650**
Promedio SNPs seleccionados en cada simulación (sd)		60.51 (7.09)	**9.63 (0.68)**	**11.12 (3.24)**
Frecuencia con la que son seleccionados los SNPs causales	SNP1	100	94	97
	SNP2	100	98	97
	SNP3	100	95	99
	SNP4	100	95	98
	SNP5	100	96	100
	SNP6	100	96	97
	SNP7	100	96	97
	SNP8	100	98	99
	SNP9	100	98	99
	SNP10	100	93	95
Frecuencia promedio para los 10 SNPs causales (sd)		100 (0.00)	**95.90 (1.73)**	**97.80 (1.48)**

Tabla 4.9. Resultados para el escenario VIII

ESCENARIO IX **p=0.2 MAF=0.3**		Selección pvalue		Selección regresión LASSO
		α=0.05	α=0.00005	λ=1.se
AUC		0.644	**0.680**	**0.671**
Promedio SNPs seleccionados en cada simulación (sd)		59.49 (6.63)	**9.97 (0.39)**	**10.60 (3.01)**
Frecuencia con la que son seleccionados los SNPs causales	SNP1	100	97	98
	SNP2	100	99	100
	SNP3	100	99	100
	SNP4	100	99	100
	SNP5	100	98	98
	SNP6	100	100	98
	SNP7	100	100	100
	SNP8	100	99	99
	SNP9	100	100	100
	SNP10	100	100	98
Frecuencia promedio para los 10 SNPs causales (sd)		100 (0.00)	**99.10 (0.99)**	**99.10 (0.99)**

Tabla 4.10. Resultados para el escenario IX

ESCENARIO X **p=0.3 MAF=0.1**		Selección pvalue		Selección regresión LASSO
		α=0.05	α=0.00005	λ=1.se
AUC		0.604	**0.631**	**0.630**
Promedio SNPs seleccionados en cada simulación (sd)		59.56 (6.97)	**8.53 (1.10)**	**12.20 (6.31)**
Frecuencia con la que son seleccionados los SNPs causales	SNP1	100	83	95
	SNP2	99	83	95
	SNP3	100	87	97
	SNP4	100	90	98
	SNP5	100	83	96
	SNP6	100	88	97
	SNP7	100	84	92
	SNP8	99	83	97
	SNP9	100	87	96
	SNP10	100	82	97
Frecuencia promedio para los 10 SNPs causales (sd)		99.80 (0.42)	**85.00 (2.75)**	**96.00 (1.70)**

Tabla 4.11. Resultados para el escenario X

ESCENARIO XI **p=0.3 MAF=0.2**		Selección pvalue		Selección regresión LASSO
		α=0.05	α=0.00005	λ=1.se
AUC		0.646	**0.681**	**0.673**
Promedio SNPs seleccionados en cada simulación (sd)		59.63 (6.43)	**9.96 (0.40)**	**12.20 (6.31)**
Frecuencia con la que son seleccionados los SNPs causales	SNP1	100	99	100
	SNP2	100	97	99
	SNP3	100	99	100
	SNP4	100	99	100
	SNP5	100	100	100
	SNP6	100	98	99
	SNP7	100	100	100
	SNP8	100	99	99
	SNP9	100	99	100
	SNP10	100	100	100
Frecuencia promedio para los 10 SNPs causales (sd)		100 (0.00)	**99.00 (0.94)**	**99.70 (0.48)**

Tabla 4.12. Resultados para el escenario XI

ESCENARIO XII **p=0.3 MAF=0.3**		Selección pvalue		Selección regresión LASSO
		α=0.05	α=0.00005	λ=1.se
AUC		0.667	**0.700**	**0.693**
Promedio SNPs seleccionados en cada simulación (sd)		60.66 (6.82)	**10.00 (0.28)**	**10.29 (1.12)**
Frecuencia con la que son seleccionados los SNPs causales	SNP1	100	98	98
	SNP2	100	99	100
	SNP3	100	100	100
	SNP4	100	100	99
	SNP5	100	99	100
	SNP6	100	100	100
	SNP7	100	100	99
	SNP8	100	100	100
	SNP9	100	100	100
	SNP10	100	100	100
Frecuencia promedio para los 10 SNPs causales (sd)		100 (0.00)	**99.60 (0.70)**	**99.60 (0.70)**

Tabla 4.13. Resultados para el escenario XII

Número de SNPs seleccionados

En primer lugar analizamos el número de SNPs seleccionados por cada modelo (Tabla 4.14). Los dos supuestos de mayor interés son aquellos que se corresponden con la selección por pvalue teniendo en cuenta un nivel de significación corregido en función del número de test, y regresión LASSO con elección del parámetro de penalización mediante validación cruzada.

Comparamos los resultados para el método de selección (1) pvalue<0.05, (2) pvalue<0.00005 y (3) regresión LASSO (λ=1.se).

Escenario	**p() MAF()**	**α=0.05**	**α=0.00005**	**λ=1.se**
I	p(0.01) MAF (0.1)	59.29 (6.63)	3.78 (1.45)	23.77 (64.61)
II	p(0.01) MAF (0.2)	58.07 (6.78)	7.90 (1.30)	13.46 (15.00)
III	p(0.01) MAF (0.3)	58.94 (6.46)	9.47 (0.82)	11.53 (4.40)
IV	p(0.1) MAF (0.1)	59.58 (7.26)	4.90 (1.70)	14.05 (17.92)
V	p(0.1) MAF (0.2)	58.47 (6.12)	8.84 (0.91)	11.98 (4.99)
VI	p(0.1) MAF (0.3)	59.91 (6.59)	9.73 (0.69)	11.86 (4.55)
VII	p(0.2) MAF (0.1)	59.61 (7.17)	6.81 (1.55)	13.72 (12.04)
VIII	p(0.2) MAF (0.2)	60.51 (7.09)	9.63 (0.68)	11.12 (3.24)
IX	p(0.2) MAF (0.3)	59.49 (6.63)	9.97 (0.39)	10.60 (3.01)
X	p(0.3) MAF (0.1)	59.56 (6.97)	8.53 (1.10)	12.20 (6.31)
XI	p(0.3) MAF (0.2)	59.63 (6.43)	9.96 (0.4)	12.20 (6.31)
XII	p(0.3) MAF (0.3)	60.66 (6.82)	10.00 (0.28)	10.29 (1.12)

Tabla 4.14. Media y desviación estándar del número de SNPs seleccionados en cada escenario

(1) α=0.05

- Se selecciona un número elevado de variables, muy superior al número de SNPs causales, por lo que se obtuvienen muchos falsos positivos. Podemos observar que el número de SNPs seleccionados está dentro del rango (58 - 61). Esto evidencia que el test empleado Cochran-Amitage-Trend Test aproxima bien el nivel de significación bajo la hipótesis nula: al seleccionar para todos los escenarios en torno a 60 SNPs, de los cuales 10 son causales, los 50 restantes constituyen el porcentaje de falsos positivos esperados para el nivel de significación establecido.

(2) α=0.00005

- Ajustando por “multiple testing” con el método de Bonferroni la selección es muy conservadora. El número de SNPs seleccionados está en el rango 3-10, es decir, es igual o inferior al número de SNPs causales. Esta estrategia asegura no hay falsos positivos a costa de perder algunos verdaderos positivos.

(3) λ=1.se

- En el caso de la regresión LASSO se selecciona un número de variables muy próximo al número de variables causales, en este caso 10. Al seleccionar el parámetro de penalización mediante validación cruzada, penaliza a 0 los coeficientes de la mayoría de variables no informativas, en este caso el número de variables con coeficientes distintos de 0 es en todos los escenarios >=10, aunque presenta mayor varianza.

Efecto MAF

- Empleando el método (2), al aumentar la MAF (para una misma prevalencia) aumenta el número de SNPs seleccionados, de modo que se aproxima más al número real de SNPs causales. En el caso del método (3) al aumentar MAF disminuye el número de SNPs seleccionados, pero del mismo modo se acerca más al número de SNPs causales.

  Los dos métodos de selección de variables funcionan mejor para MAF elevadas. Este hecho parece ser independiente de la prevalencia.

Prevalencia

- No parece ejercer un gran efecto. Aunque en condiciones de baja prevalencia y y bajas frecuencias del alelo menor, los métodos (2) y (3) funcionan peor.

Frecuencia con la que son seleccionados los SNPs causales

A continuación analizamos el número de SNPs causales seleccionados, tomando la frecuencia promedio para los 10 SNPs (Tabla 4.15)

Escenario	p( ) MAF ( )	α=0.05	α=0.00005	λ=1.se
I	p(0.01) MAF (0.1)	96.14 (1.43)	37.20 (5.14)	76.30 (3.27)
II	p(0.01) MAF (0.2)	100 (0.00)	78.60 (3.20)	92.40 (3.24)
III	p(0.01) MAF (0.3)	100 (0.00)	93.80 (4.02)	96.90 (2.13)
IV	p(0.1) MAF (0.1)	98.20 (1.32)	48.40 (6.75)	80.70 (4.69)
V	p(0.1) MAF (0.2)	100 (0.00)	88.10 (3.07)	95.30 (2.00)
VI	p(0.1) MAF (0.3)	100 (0.00)	96.20 (2.25)	98.50 (1.58)
VII	p(0.2) MAF (0.1)	99.70 (0.48)	67.60 (6.40)	90.40 (3.84)
VIII	p(0.2) MAF (0.2)	100 (0.00)	95.90 (1.73)	97.80 (1.48)
IX	p(0.2) MAF (0.3)	100 (0.00)	99.10 (0.99)	99.10 (0.99)
X	p(0.3) MAF (0.1)	99.80 (0.42)	85.00 (2.75)	96.00 (1.70)
XI	p(0.3) MAF (0.2)	100 (0.00)	99.00 (0.94)	99.70 (0.48)
XII	p(0.3) MAF (0.3)	100 (0.00)	99.60 (0.70)	99.60 (0.70)

Tabla 4.15. Frecuencia promedio de selección para los 10 SNPs causales

(1) α=0.05

- Estableciendo este umbral de significación se seleccionaron siempre todas las variables causales (a excepción de aquellos escenarios con MAF y prevalencia bajas). Esto es a costa de tener entre las variables seleccionadas un número muy elevado de falsos positivos. Veremos en la siguiente tabla que este hecho reduce considerablemente la capacidad predictiva de los modelos obtenidos.

(2) α=0.00005

- Corregir el nivel de significación en función del número de test realizados implica perder potencia en la selección de variables causales, de modo que ser tan estrictos para evitar falsos positivos implica no poder identificar todas las variables que muestran asociación con la respuesta.

(3) λ=1.se

- Este método proporciona un equilibrio entre una alta potencia de selección de variables causales (para la mayoría de los escenarios con una frecuencia >= 90%) y un bajo número de falsos positivos.

Efecto MAF

- Para valores altos de MAF (0.2, 0.3) no existen apenas diferencias entre los métodos (2) y (3) en cuanto a su capacidad para identificar variables causales. Para valores de MAF=0.1 la regresión LASSO tiene mayor habilidad para seleccionar SNPs asociados con la respuesta.

Prevalencia

- Igual que en el caso anterior, no se observa un gran efecto de la prevalencia, aunque sí parece apreciarse una cierta tendencia: a medida que aumenta la prevalencia es más fácil identificar SNPs causales. Asimismo p(0.01) y MAF(0.1) representan el peor escenario posible para cualquiera de los métodos empleados.

Capacidad predictiva. AUC

Por último evaluamos la capacidad predictiva de los 3 métodos (Tabla 4.16)

(1) α=0.05

- Claramente este método es el menos efectivo a nivel de capacidad predictiva. El ruido introducido por el gran número de falsos positivos reduce considerablemente la capacidad predictiva del modelo de regresión construido con las variables seleccionadas.

(2) α=0.00005 y (3) λ=1.se

- Ambos métodos presentaron unos valores de AUC similares. En este caso el AUC no nos permite discriminar entre los dos métodos de selección de variables, los podemos considerar prácticamente equivalentes en cuanto a su capacidad predictiva.

Escenario	p () MAF ()	α=0.05	α=0.00005	λ=1.se
I	p(0.01) MAF (0.1)	0.559 [0.536, 0.582]	0.557 [0.527, 0.585]	0.575 [0.536, 0.605]
II	p(0.01) MAF (0.2)	0.593 [0.573, 0.615]	0.620 [0.587, 0.647]	0.617 [0.587,0.646]
III	p(0.01) MAF (0.3)	0.610 [0.589, 0.633]	0.647 [0.626, 0.668]	0.641 [0.611, 0.663]
IV	p(0.1) MAF (0.1)	0.570 [0.550, 0.587]	0.573 [0.529, 0.606]	0.587 [0.548, 0.614]
V	p(0.1) MAF (0.2)	0.606 [0.590, 0.627]	0.636 [0.615, 0.658]	0.641 [0.611, 0.663]
VI	p(0.1) MAF (0.3)	0.623 [0.604, 0.645]	0.659 [0.636, 0.679]	0.652 [0.630, 0.673]
VII	p(0.2) MAF (0.1)	0.585 [0.563, 0.606]	0.601 [0.569, 0.628]	0.608 [0.572, 0.632]
VIII	p(0.2) MAF (0.2)	0.622 [0.600, 0.644]	0.658 [0.636, 0.677]	0.650 [0.623, 0.673]
IX	p(0.2) MAF (0.3)	0.644 [0.624, 0.665]	0.680 [0.658, 0.698]	0.671 [0.645, 0.690]
X	p(0.3) MAF (0.1)	0.604 [0.583, 0.623]	0.631 [0.599, 0.654]	0.630 [0.602, 0.651]
XI	p(0.3) MAF (0.2)	0.646 [0.626, 0.665]	0.681 [0.662, 0.700]	0.673 [0.652, 0.695]
XII	p(0.3) MAF (0.3)	0.667 [0.648, 0.686]	0.700 [0.684, 0.717]	0.693 [0.669, 0.713]

Tabla 4.16. AUCs (IC 95%) para los diferentes escenarios

Efecto MAF

- A medida que aumenta MAF (independientemente de la prevalencia) aumenta la capacidad predictiva de los modelos generados, esto es así para cualquiera de los métodos de selección de variables empleados.

Efecto prevalencia

- Para un valor dado de MAF y manteniendo el resto de parámetros constantes, el valor de AUC se incrementa al aumentar la prevalencia.

5. Aplicación a un estudio en enfermedad de Parkinson

5.1. Descripción de los datos

Los datos proceden de un estudio de asociación de genes candidatos en enfermedad de Parkinson, llevado a cabo por el grupo de Neurogenética de la Fundación Pública Galega de Medicina Xenómica (FPGMX) en colaboración con neurólogos del Complejo Hospitalario Universitario de Santiago de Compostela (CHUS). Se trata de un estudio poblacional caso-control.

Los criterios clínicos se han establecido en base a la *UK PD Society Brain Bank* mediante examen de neurólogos especializados en trastornos del movimiento. En los controles se descartó la presencia de antecedentes de parkinsonismo. Todos los participantes pertenecían a la población gallega, que desde el punto de vista genético no presenta estratificación signifcativa y por lo tanto sin problemas de subestructura poblacional que puedan dar lugar a falsos positivos por esa causa.

Partimos del estudio de 587 individuos y 115 SNPs. Los datos sobre los que finalmente trabajamos (525 individuos, 71 SNPs) son aquellos que han pasado diferentes filtros de control de calidad referentes a:

1. Tasa de genotipado: Todos los SNPs están genotipados en al menos el 95% de las muestras y para todas las muestras se han genotipado al menos el 90% de los SNPs. Las muestras y SNPs con tasas de genotipado inferiores a las mencionadas fueron eliminadas del estudio.

2. HWE: Se estudió el equilibro Hardy-Weinberg en la población control. Se estableció un nivel de significación, $\alpha=0.05$. Aquellos SNPs que no se encontraban en HWE en controles fueron asimismo eliminados del estudio.

3. MAF: Para todos los SNPs la frecuencia del alelo menor es superior al 10% en la población control.

Los análisis relativos al control de calidad no se muestran en el presente trabajo.

Dado que estos resultados están pendientes de publicación se han recodificado los nombres de los genes y SNPs analizados.

Estos 71 SNPs objeto del estudio pertenecen a 15 genes, tal y como se muestra en la tabla 5.1

gen A	snp27	snp42							
gen B	snp20	snp70							
gen C	snp02	snp03	snp04	snp05	snp11	snp22	snp30	snp32	snp51
	snp58								
gen D	snp50								
gen E	snp64								
gen F	snp12	snp17							
gen G	snp35								
gen H	snp06								
gen I	snp01	snp19	snp28	snp69					
gen J	snp45	snp53							
gen K	snp10	snp14	snp16	snp18	snp21	snp23	snp24	snp26	snp33
	snp52	snp59	snp60	snp68					
gen L	snp08	snp15	snp29	snp31	snp36	snp40	snp41	snp43	snp46
	snp47	snp48	snp49	snp54	snp56	snp62	snp63	snp65	snp66
	snp67	snp71							
gen M	snp09	snp25	snp57						
gen N	snp13	snp34	snp37	snp38	snp39	snp44	snp55	snp61	
gen O	snp07								

Tabla 5.1. Genes y SNPs objeto de estudio

5.2 Análisis exploratorio

5.2.1 Análisis de asociación univariante

En primer lugar empleamos un test de asociación individual, en este caso Likelihood Ratio Test, implementado en el paquete *SNPassoc* [34] mediante la función *WGassociation.* En la tabla 5.2 se muestran los pvalues obtenidos al aplicar este test para aquellos SNPs que mostraron asociación significativa para α=0.1 considerando un modelo log-aditivo. En este análisis inicial no consideramos ningún tipo de corrección para comparaciones múltiples.

Podemos observar que todos los SNPs que muestran asociación individual a un nivel de significación α=0.05 pertenecen a tres genes. Por este motivo nos centraremos en describir los patrones de desequilibrio de ligamiento de los genes K, M y N.

snp	gen	pvalue
snp55	gen N	0.00060788
snp38	gen N	0.00870665
snp37	gen N	0.00876914
snp39	gen N	0.01203832
snp10	gen K	0.01728450
snp57	gen M	0.02782538
snp25	gen M	0.03813298
snp13	gen N	0.04236906
snp09	gen M	0.05680599
snp35	gen G	0.06375716
snp58	gen C	0.06601989
snp18	gen K	0.08701018
snp50	gen D	0.09800406

Tabla 5.2 Resultados de LRT para α=0.1

5.2.2 Estudio del Desequilibro de Ligamiento

En el capítulo 2 describíamos el desequilibrio de ligamiento y diferentes medidas para evaluarlo. Existen diferentes aplicaciones gratuitas que permiten calcular estas medidas y a su vez proporcionan representaciones gráficas de las mismas en la forma de "heatmaps"" teniendo en cuanta las posiciones relativas entre SNPs. Uno de los más extendidos es el desarrollado por el Broad Institue of MIT and Harvard, Haploview

4.2 [35]. En este caso se ha empleado el software R, que dispone del paquete *LDheatmap* [36], con una función con el mismo nombre disponible que permite calcular los estadísticos D' y r^2 así como mapas de desequilibrio de ligamiento para una serie de SNPs en una determinada región genómica, dadas sus distancias físicas o posiciones genéticas y los genotipos correspondientes.

De este modo obtenemos las medidas de desequilibrio de ligamiento por parejas correspondientes a D' y r^2 y los correspondientes gráficos con códigos de colores según estas medidas. En la figura 5.1 se muestran los mapas de desequilibrio de ligamiento tomando los valores de D' y r^2 para todos los SNPs analizados pertenecientes al gen N. Asimismo en las tablas 5.3 y 5.4 se muestran los valores de los estadísticos D' y r^2 respectivamente para cada par de SNPs.

	snp39	snp34	snp13	snp37	snp44	snp55	snp61
snp38	1.000	0.819	0.999	1.000	0.998	0.999	0.987
snp39		0.821	0.999	1.000	0.998	0.999	0.974
snp34			0.400	0.819	0.011	0.283	0.386
snp13				0.999	0.418	0.582	0.029
snp37					0.998	0.999	0.988
snp44						0.910	0.287
snp55							0.807

Tabla 5.3 Medidas de D' para todos los SNPs analizados del gen N

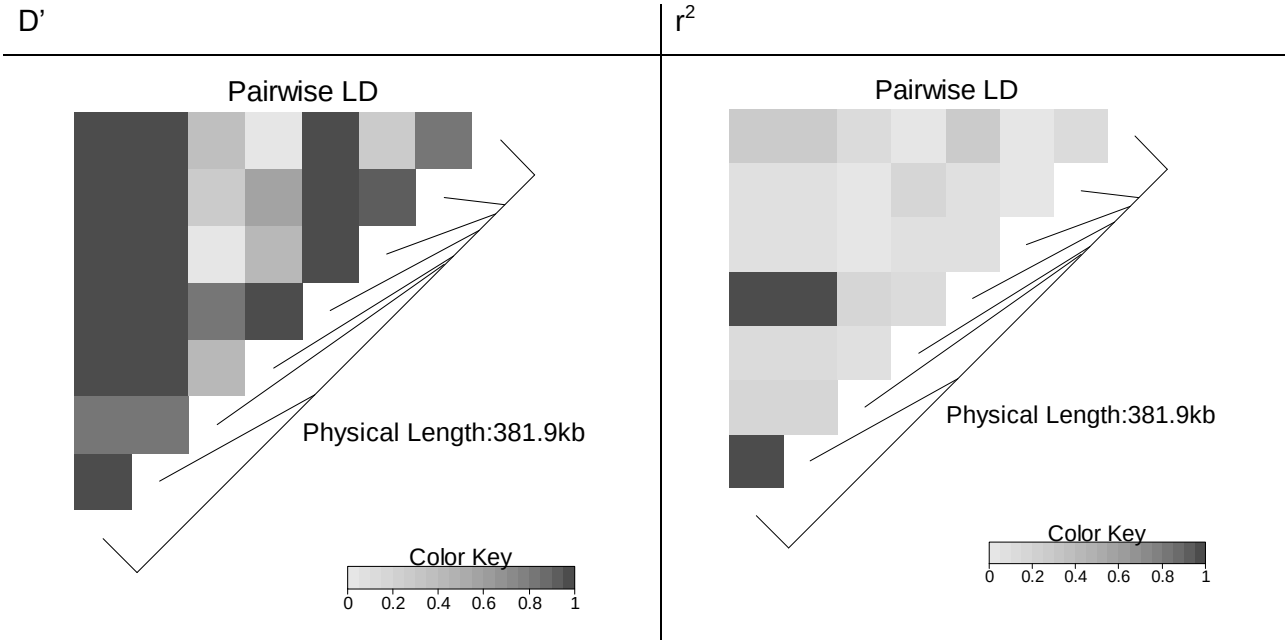
	snp39	snp34	snp13	snp37	snp44	snp55	snp61
snp38	0.999	0.152	0.134	0.994	0.067	0.069	0.257
snp39		0.154	0.135	0.997	0.067	0.069	0.252
snp34			0.050	0.153	3.3e-05	0.013	0.128
snp13				0.135	0.088	0.174	2.9e-04
snp37					0.067	0.069	0.259
snp44						0.039	0.021
snp55							0.122

Tabla 5.4 Medidas de r^2 para todos los SNPs analizados del gen N

D' presenta una gran dependencia del tamaño muestral, y puede estar sesgado si el tamaño muestral es pequeño. Dado que r^2 es una medida más restrictiva, será la que empleemos como referencia.

En base a los valores de r^2 podemos observar que existe un fuerte LD entre los siguientes pares de snps: snp37:snp38, snp37:snp39 y snp38:snp39. Destacamos aquellos valores de $r^2 \geq 0.9$. Dado que estos tres SNPs están entre los que muestran asociación significativa según se muestra en la tabla 5.2, serán objeto de especial atención en los análisis posteriores.

Figura 5.1. Mapas de desequilibrio de ligamiento para el gen N empleando los valores de *D'* y r^2

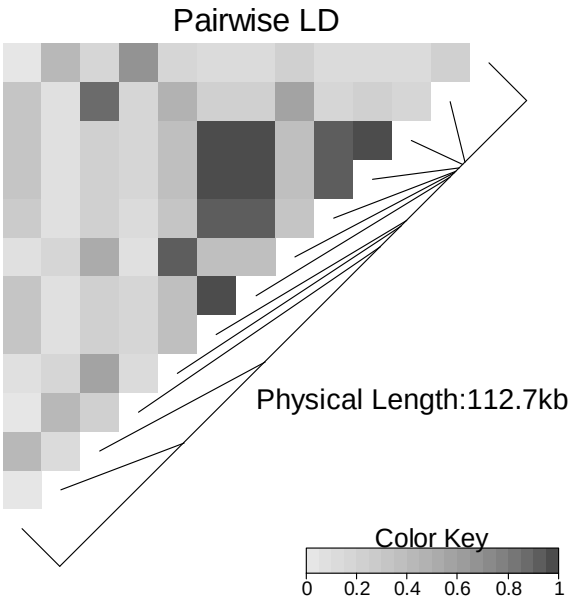


Realizamos el mismo análisis para el gen K, en este caso mostramos los valores correspondientes a r^2 en la tabla 5.5 y el mapa de desequilibrio de ligamiento en la figura 5.2

	snp18	snp14	snp26	snp24	snp60	snp59	snp16	snp14	snp68	snp52	snp21	snp23
snp10	0,04	0,41	0,00	0,09	0,31	0,31	0,07	0,29	0,30	0,30	0,33	0,03
snp18		0,10	0,43	0,18	0,07	0,07	0,15	0,06	0,07	0,07	0,07	0,45
snp14			0,23	0,57	0,22	0,22	0,51	0,20	0,21	0,21	0,88	0,17
snp26				0,10	0,15	0,15	0,07	0,14	0,15	0,15	0,18	0,69
snp24					0,37	0,37	0,91	0,34	0,36	0,36	0,50	0,17
snp60						1,00	0,37	0,93	0,98	0,97	0,21	0,12
snp59							0,37	0,93	0,98	0,97	0,21	0,12
snp16								0,34	0,36	0,36	0,56	0,21
snp14									0,95	0,93	0,19	0,13
snp68										0,99	0,20	0,13
snp52											0,20	0,13
snp21												0,20

Tabla 5.5. Valores de r^2 para los SNPs del gen K

Figura 5.2 . Mapa de desequilibrio de ligamiento para el gen K empleando los valores de r^2



En este caso hay varios SNPs con valores de $r^2 \geq 0.9$: snp24:snp16, snp60:14, snp60:snp68, snp60:52, snp59:snp14, snp59:snp52, snp14:snp52, snp68:snp52.

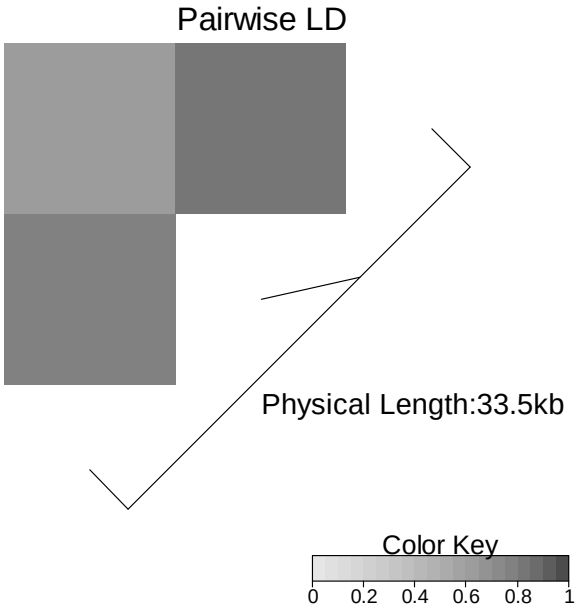
Aunque ninguno de ellos está entre los SNPs que presentan asociación significativa, evaluaremos su comportamiento en los posteriores modelos.

Por último realizamos los mismos cálculos para el gen M, para el que tan sólo tenemos genotipos de tres SNPs. Los valores de r^2 se muestran en la tabla 5.6 y el mapa correspondiente en la figura 5.3.

	snp25	snp09
snp57	0.771	0.633
snp25		0.810

Tabla 5.6. Valores de r^2 para los SNPs del gen M

Figura 5.3. Mapa de desequilibrio de ligamiento para el gen K empleando los valores de y r^2



En este caso no hay SNPs que estén en alto desequilibrio de ligamiento. Todos los valores de r^2 son inferiores a 0.9

5.3. Selección de variables y evaluación de la capacidad predictiva

Selección de variables mediante test de asociación univariante

Se seleccionan la variables predictoras que presentaron una asociación significativa con la variable respuesta aplicando un test de asociación. Se empleó de nuevo el Likelihood Ratio Test (LRT) incorporado en la función *WGassociation*. Esta función permite evaluar la asociación bajo diferentes modelos genéticos, en este caso escogemos el modelo log-aditivo. Tan sólo consideramos los SNPs como variables predictoras, descartando otras variables como sexo, edad, etc.

De este modo se obtiene el pvalue correspondiente al LRT. Es necesario establecer un umbral de significación para decidir qué variables se incorporarán al modelo de regresión multivariante. Mickey & Greenland [37] recomiendan incluir aquellas variables cuyo test univariante arroje un p-valor <0.25. Establecen esta aproximación conservadora de modo que se puedan identificar todas aquellas variables importantes.

En este estudio se consideraron tres umbrales de significación: α=0.05, α=0.1 y α=0.25 que constituyen los tres métodos que denominamos a.1) a.2) y a.3) respectivamente. Si quisiéramos ser más exigentes y establecer un nivel de significación teniendo en cuanta el número de test realizados, no sería apropiado emplear el método de Bonferroni, ya que hemos verificado que muchos SNPs no son independientes ya que está en fuerte desequilibrio de ligamiento. Si observamos los resultados obtenidos en el análisis preliminar tan sólo hay tres SNPs con pvalues <0.01, los tres pertenecen al mismo gen y están el alto LD. Este hecho, unido a que el número de observaciones del que disponemos es relativamente pequeño y además vamos a emplear validación cruzada, y por lo tanto se verá todavía más reducido, hace que no se hayan considerado umbrales más restrictivos.

Una vez seleccionados los SNPs con pvalues inferiores a los distintos niveles de significación establecidos, se construyeron los correspondientes modelos de regresión mediante la función *glm* del paquete *stats* [38], empleando estos subconjuntos de SNPs como variables independientes.

Como indicábamos con anterioridad, hemos empleado la técnica de validación cruzada para evaluar los modelos obtenidos.

En este proceso que estamos describiendo y que consta de dos etapas: 1) selección de variables y 2) ajuste de un modelo de regresión logística múltiple con las variables

seleccionadas en el paso anterior, la validación cruzada tiene que ser aplicada a la secuencia entera de pasos [39] :

1) Se divide la muestra en k grupos (en nuestro estudio k=5). Para cada k, seleccionamos todas las muestras excepto aquellas en k, y buscamos aquellas variables predictoras cuya asociación individual con la variable respuesta sea más fuerte.

2) Empleando ese subconjunto de predictoras se construye el clasificador multivariante, utilizando las mismas muestras que en el caso anterior (todas excepto aquellas en k).

3) Usamos el clasificador para predecir la etiqueta de clase de las muestras en el grupo k (set de validación/test).

En este estudio hemos incorporado la validación cruzada con k=5. De este modo obtenemos 5 clasificadores, que en nuestro caso son modelos de regresión logística múltiple. Calculamos las predicciones para cada uno de los cinco modelos y para evaluar el rendimiento de los modelos, en lugar de emplear el error de predicción hemos construido las curvas ROC correspondientes y calculado el AUC con los correspondientes IC al 95%. Como medida global, representamos la curva ROC y estimamos el AUC agregando las predicciones de los 5 modelos obtenidos mediante validación cruzada.

Selección mediante regresión logística penalizada

Como indicábamos en el capítulo 3 la idea fundamental de este método es la penalización, de modo que se evita el sobreajuste imponiendo una penalización sobre fluctuaciones grandes de los parámetros estimados. La regresión LASSO "encoge" algunos coeficientes a 0 (en este caso SNPs no informativos), de este modo actúa también como método de selección de variables. La elección de la constante de penalización (λ) es fundamental. Es necesario un procedimiento que estime el valor de este parámetro a partir de los datos.

Existen diferentes criterios para seleccionar la constante de penalización, empleando validación cruzada lo más habitual es la regla 1 error estándar. Otra posibilidad es fijar el número máximo de variables a incluir en el modelo.

La elección del parámetro de penalización controla el número de variables seleccionadas. Cuanto mayor sea el valor del parámetro, menor el número de predictoras incluidas en el modelo. De nuevo empleamos el paquete *glmnet* y las funciones *glmnet* y *cv.glmnet*, que ajustan un modelo de regresión logística vía máxima verosimilitud penalizada.

En este estudio hemos empleado diferentes métodos para seleccionar el valor del parámetro de penalización:

(b.1) Mediante validación cruzada (incorporada en la función *cv.glmnet*). Selecciona el valor del parámetro λ= 1 s.e (mayor valor de λ tal que el error está dentro de un error estándar del mínimo). Separamos un set de validación y calculamos el AUC para las predicciones sobre este set de validación.

(b.2) Del mismo modo que en el caso anterior pero con λ= min (valor de λ que produce el mínimo de los errores medios de validación cruzada).

(b.3) Incorporación de validación cruzada manualmente. Se fija un número máximo de variables. Con cada training set construimos un modelo. Análogamente al caso anterior agregamos las predicciones de los 5 modelos y representamos una única curva ROC. Fijamos a 10 el número máximo de variables a incluir en el modelo, de modo que para cada uno de los grupos seleccionamos el valor de λ correspondiente.

(b.4) De manera análoga al método (b.3), en este caso se establece el número máximo de variables en 20. De igual modo, para obtener una medida global para cada uno de los métodos, agregamos las predicciones de los 5 modelos para obtener una única curva ROC y el correspondiente AUC.

Estos valores (10 y 20) fueron escogidos porque es el rango en el que se mueve el número de variables seleccionadas mediante test de asociación individual al variar el nivel de significación de 0.05 a 0.25.

Dado que la función *glmnet* no permite la existencia de “missing values” realizamos una imputación de los datos faltantes mediante la función *rfImpute* del paquete *randomForest* [40]. Esta función imputa los datos faltantes empleando la matriz de proximidad de randomForest. La recodificación de los SNPs para considerlos bajo un modelo aditivo la realizamos con la función *recodeSNPs* del paquete *scrime* [41].

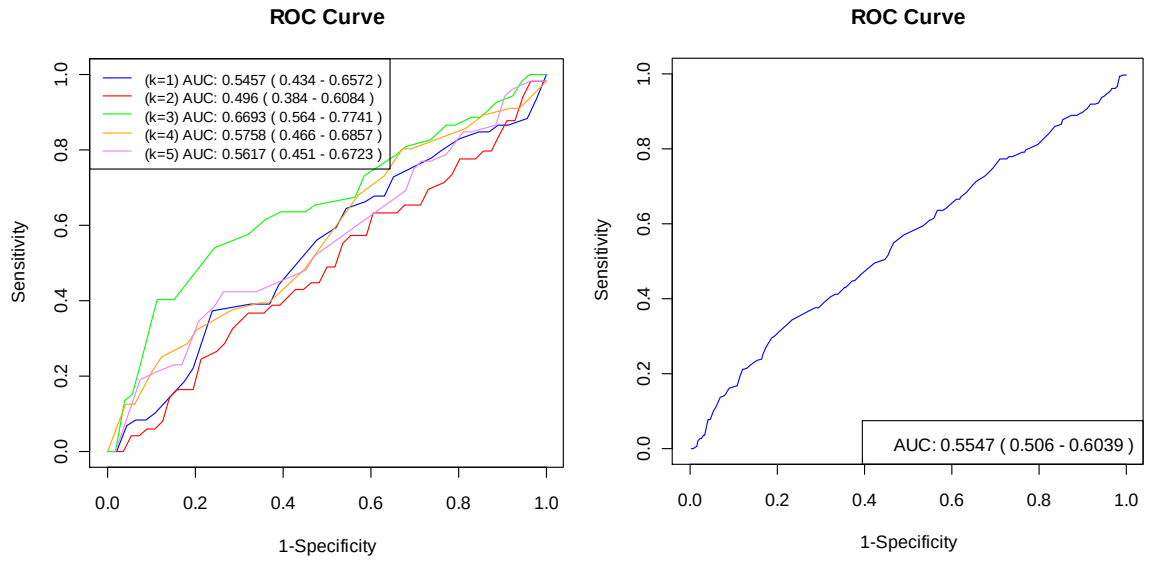
5.4. Resultados y discusión

Selección mediante test de asociación

Una vez ajustados los modelos de regresión con las variables seleccionadas para los tres niveles de significación establecidos, evaluamos su capacidad predictiva. Las curvas ROC, junto con los índices AUC y los correspondientes I.C. 95% obtenidos para cada uno de los subconjuntos, tras aplicar validación cruzada con k=5 (a.1.1), y para la agregación de predicciones de los 5 subconjuntos (a.1.2), se muestran en la figuras 5.4, 5.5 y 5.6 para α=0.05, α=0.1 y α=0.25 respectivamente.

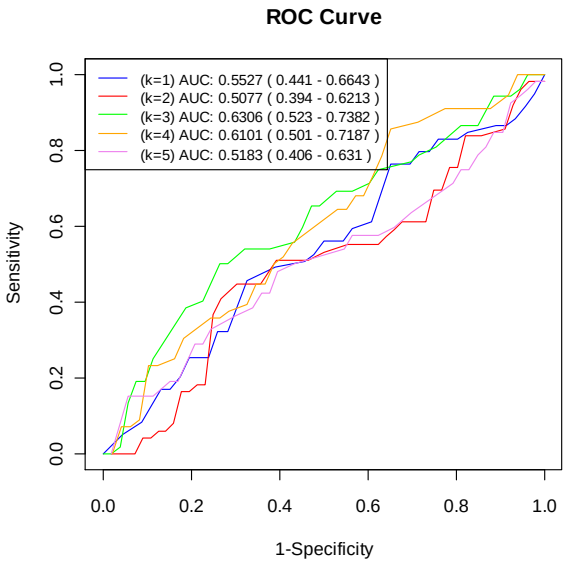
Los SNPs seleccionados para cada uno de los 5 sets de entrenamiento para los distintos niveles de significación se muestran en la tabla 5.6.

Figura 5.4. ROC y AUC para α=0.05

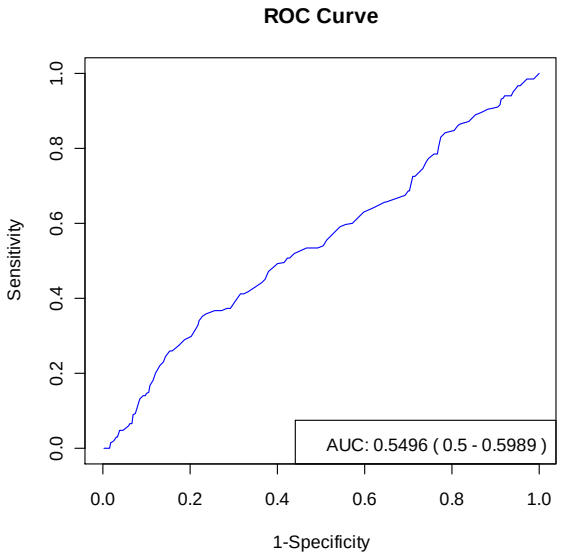


(a.1.1) (a.1.2)

Figura 5.5. ROC y AUC para α=0.1

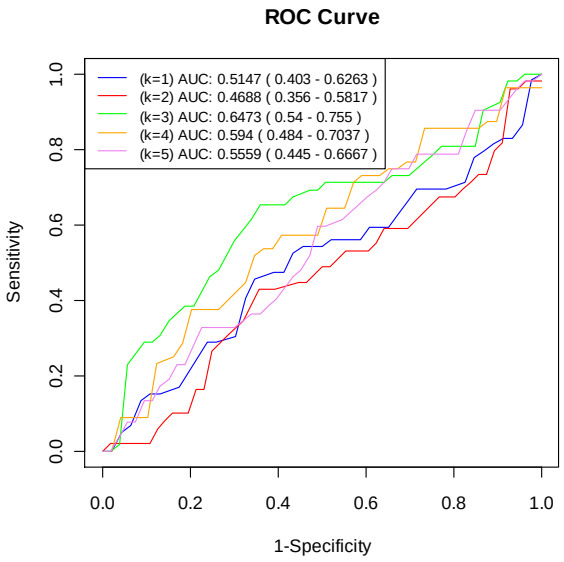


(a.2.1)

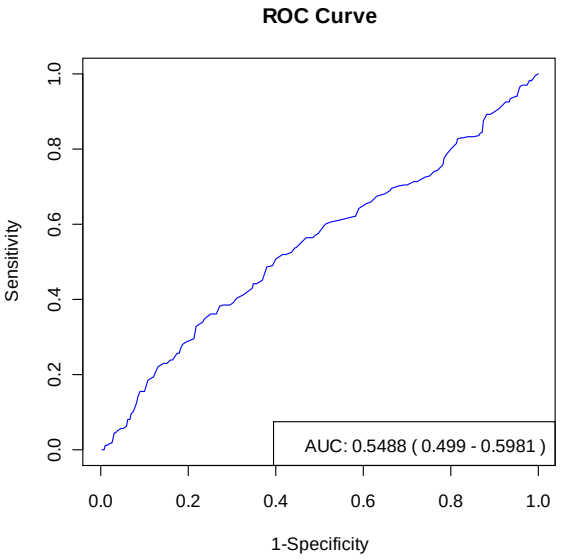


(a.2.2)

Figura 5.6. ROC y AUC para α=0.25



(a.3.1)



(a.3.2)

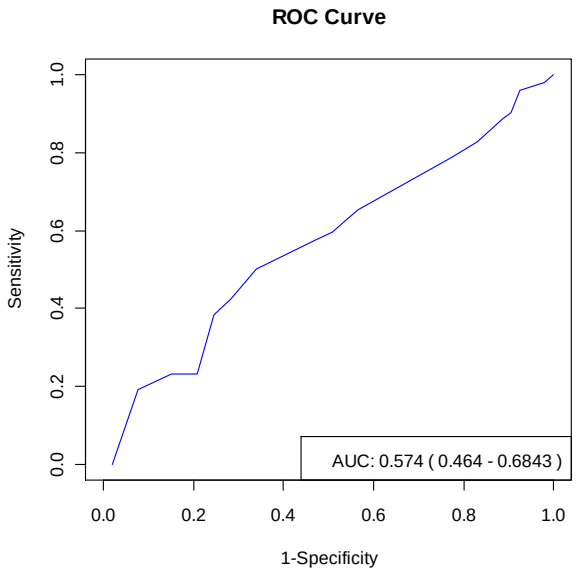
	k=1			k= 2			k= 3			k=.4			k= 5			Total		
	α 0.05	**α 0.1**	**α 0.25**	**α 0.05**	**α 0.1**	**α 0.25**	**α 0.05**	**α 0.1**	**α 0.25**	**α 0.05**	**α 0.1**	**α 0.25**	**α 0.05**	**α 0.1**	**α 0.25**	**α 0.05**	**α 0.1**	**α 0.25**
snp04																	1	2
snp05																	1	2
snp09																2	3	4
snp10																4	5	5
snp13																2	3	5
snp15																		1
snp16																		1
snp18																2	4	4
snp19																		1
snp21																1	1	1
snp23																		2
snp24																		2
snp25																2	3	4
snp26																		3
snp27																		1
snp29																		1
snp33																	1	1
snp35																1	2	5
snp36																		1
snp37																5	5	5
snp38																5	5	5
snp39																5	5	5
snp41																	1	1
snp45																		2
snp50																2	3	4
snp51																		3
snp54																		1
snp55																5	5	5
snp57																3	4	5
snp58																1	2	5
snp62																	1	1
snp63																1	1	3
snp64																		1
Total	8	11	19	12	15	19	8	10	18	6	11	18	7	9	18			

Tabla 5.6. SNPs seleccionados mediante validación cruzada para los niveles de significación α=0.05, α=0.1 y α=0.25

Selección mediante regresión penalizada LASSO

(b.1) Separamos inicialmente 4/5 de la muestra que constituirán el set de entrenamiento. Mediante validación cruzada con k=4, estimamos el valor del parámetro de penalización que produce 1 s.e. Sobre el set de validación, 1/5 de la muestra inicial, calculamos las predicciones para evaluar el modelo ajustado. Las curvas ROC y AUCs correspondientes para ese valor de λ se muestran en la figura 5.7.

Figura 5.7. ROC y AUC para λ=1.se



(b.2) De modo análogo λ=min. Figura 5.8

(b.3) Validación cruzada incorporada manualmente con k=5. Para cada subconjunto se estima el valor de λ para incluir un número máximo de 10 variables en el modelo. Calculamos las predicciones para cada uno de los sets de validación empleando el parámetro de penalización correspondiente. Las curvas ROC y AUCs con I.C 95% para cada uno de los sets de entrenamiento se muestran en la figura 5.9 (b.3.1), así como las obtenidas para la agregación de predicciones de los 5 sets. (b.3.2)

Figura 5.8. ROC y AUC para λ=1.se

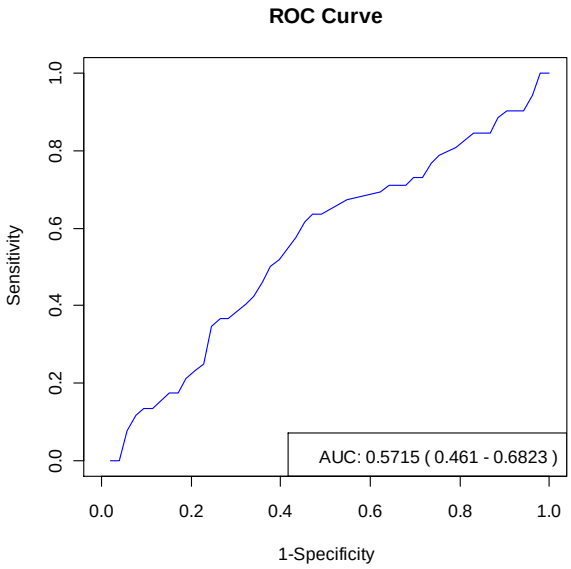
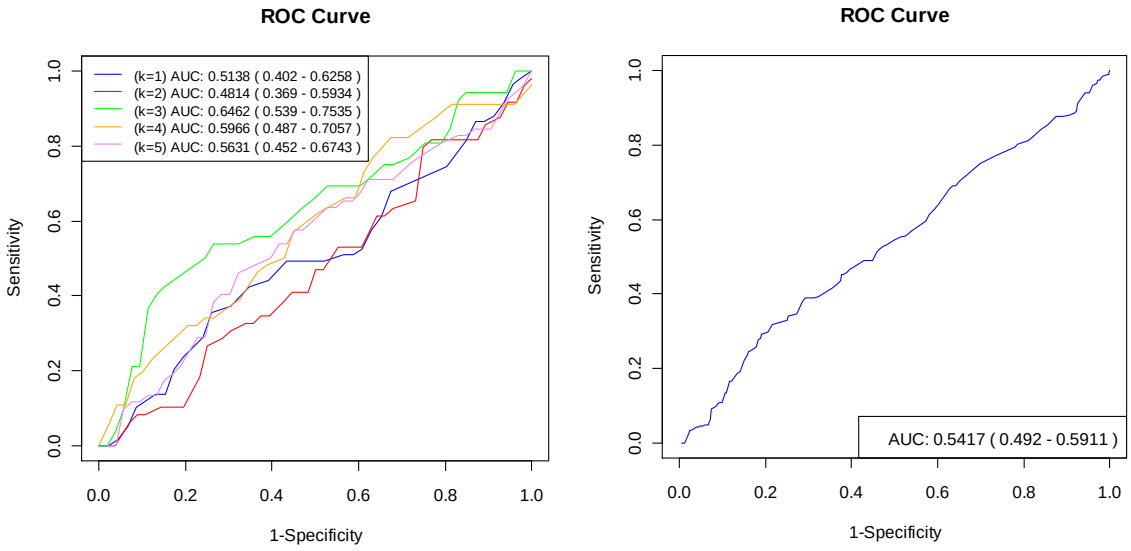


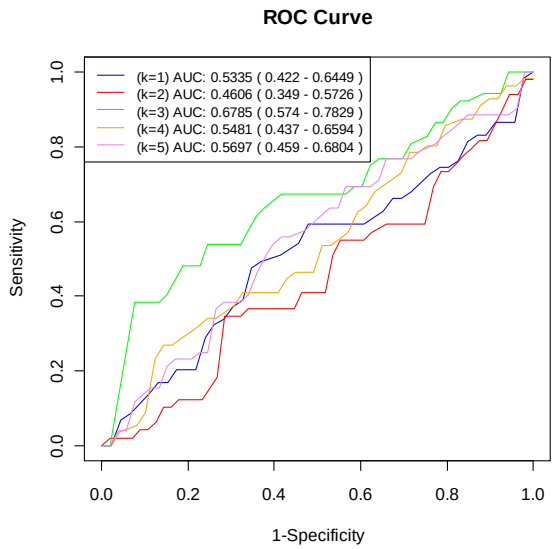
Figura 5.9. ROC y AUC para un máximo de 10 variables



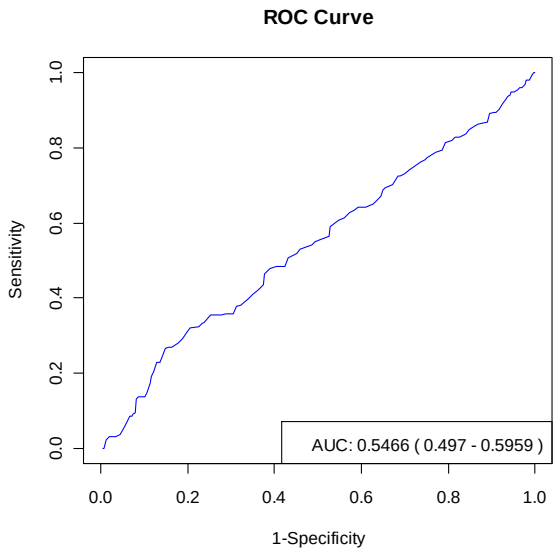
(b.3.1) (b.3.2)

(b.4) De manera análoga al método anterior estimamos el valor de λ para incluir un número máximo de 20 variables en el modelo. Resultados en figura 5.10.

Figura 5.10. ROC y AUC para un máximo de 20 variables



(b.4.1)



(b.4.2)

Los resultados en relación al número de variables seleccionadas en función del método empleado se muestran en la tabla 5.7.

			k= 1		k=2		k=3		k= 4		k= 5		Tot	Tot
	1se	min	10	20	10	20	10	20	10	10	10	20	10	20
snp01														1
snp04													1	2
snp05													1	2
snp07														2
snp09													1	2
snp10													5	5
snp13													1	2
snp14														3
snp16													1	
snp18													4	4
snp19														1
snp24														3
snp25													1	1
snp26														1
snp27														1
snp29														2
snp30														1
snp35													2	5
snp36													1	4
snp37													4	5
snp38													4	5
snp39														1
snp41													1	2
snp42														1
snp44														0
snp45														2
snp49														4
snp50													4	4
snp51														2
snp52														1
snp54														2
snp55													5	5
snp56														
snp57													2	4
snp58													4	5
snp62														
snp63													1	4
snp64														2
snp66														1
snp69														2
snp71														1
Total snps	5	12	9	20	7	20	8	20	10	17	9	20		

Tabla 5.7 SNPs seleccionados mediante validación cruzada para los diferentes métodos empleando penalización LASSO.

Capacidad predictiva

Con el método de selección por pvalue no se aprecian diferencias en cuanto a la capacidad predictiva estableciendo diferentes umbrales de significación. Aumentar el número de SNPs no afecta a la capacidad predictiva. En este caso no podemos saber qué asociaciones son reales y cuales son falsos positivos. Puede que pvalues entre 0.1 y 0.25 correspondan a asociaciones débiles.

La máxima capacidad predictiva se obtiene al emplear regresión penalizada LASSO con selección del parámetro de penalización mediante validación cruzada. Los polimorfismos incluidos en este modelo pertenecen a tres genes K, M y N que previamente han sido identificados como factores de susceptibilidad en enfermedad de Parkinson y otras enfermedades relacionadas, por lo que existen indicios sólidos de que se trata de asociaciones reales y no falsos positivos. Es posible que la incorporación de un número mayor de variables disminuya la capacidad predictiva de los modelos por incorporar variables no causales.

Esta puede ser la razón de que no se aprecien apenas diferencias entre los AUCs obtenidos al emplear regresión LASSO y pasar de permitir como máximo 10 variables en el modelo a 20 (0.5417 frente a 0.5466). Del mismo modo al ser menos restrictivos con el nivel de significación en el caso de selección por pvalue tampoco se aprecian muchas diferencias (0.5547 para α=0.05, 0.5496 para α=0.1 y 0.5488 para α=0.25). Aunque sí parece haber una tendencia a disminuir el AUC al permitir la entrada de variables más débilmente asociadas con la respuesta. Si aplicásemos corrección para comparaciones múltiples considerando False Discovery Rate (paquete *qvalue*) [42], tan solo un SNP sería significativo: snp55 (qvalue=0.0431) y sería el único que pasaría a formar parte del modelo.

La capacidad predictiva del mejor de los modelos es muy baja. Aunque un marcador presente una asociación significativa esto no implica que sea un buen clasificador, para que así fuera debería presentar Odds Ratio de magnitudes muy superiores a las encontradas en los estudios de asociación [43].

SNPs seleccionados

En el caso de selección de variables mediante pvalue, si tenemos en cuenta los SNPs seleccionados en los 5 subconjuntos, observamos que son 4 para α=0.05 (snp37, snp38, snp39 y snp55), 5 para α=0.1 (snp10, snp37, snp38, snp39 y snp55) y 9 par α=0.25 (snp10, snp13, snp35, snp37, snp38, snp39, snp55, snp57 y snp58). En las dos primeras situaciones todos los SNPs pertenecen a los genes mencionados, por lo tanto existen evidencias de que sean variantes causales (o que estén en LD con las verdaderas variantes causales).

En el caso de la regresión LASSO, para estimación de λ por validación cruzada, tenemos un modelo con 5 SNPs (snp18, snp37, snp38, snp55 y snp57). Igualmente estos SNPs pertenecen a los tres genes K, M y N.

Si analizamos los dos métodos teniendo en cuenta las situaciones de desequilibrio de ligamiento entre las variables observamos lo siguiente: al emplear la técnica de selección por pvalue, en todos los casos se han seleccionados los snps 37, 38 y 39. Al ajustar el modelo de regresión logística múltiple nos da el aviso "prediction from a rank-deficient fit may be misleading" de modo que no existe solución única y algunas de las estimaciones de los parámetros son NAs. Eliminamos manualmente la variables correspondientes del modelo, que en la totalidad de los casos se corresponde con el snp38. En la regresión LASSO podemos observar que en todos los casos excepto uno, se elimina el snp39. El coeficiente de correlación entre el snp38 y snp39 es prácticamente 1 (0.999), esto es indicativo de la presencia de colinealidad exacta entre ambos SNPs. El método LASSO "encoge" el coeficiente de una de las variables a 0. En cuanto a la relación entre el snp37 y snp38, que también presenta fuerte colinealidad, las estimaciones correspondientes de los coeficientes de regresión aplicando LASSO, son -4.367078e-02 y -3.868657e-23 respectivamente, por lo tanto la regresión LASSO "encoge" el coeficiente del segundo SNP prácticamente a 0.

Tanto al ajustar un modelo de regresión logística múltiple como al aplicar penalización LASSO, se consiguen identificar situaciones de colinealidad exacta – desequilibrio de ligamiento completo-. En el primer caso se eliminan las variables manualmente del modelo y en el segundo caso lo hace automáticamente. En cuanto a otras variables que presentan fuerte colinealidad, la regresión LASSO "contrae" los coeficientes prácticamente a 0. Existen varias referencias en cuanto a la capacidad de la regresión penalizada LASSO para detectar situaciones de desequilibrio de ligamiento [25,44]. En

ellas destacan que en presencia de variables altamente correlacionadas, LASSO tiene a seleccionar tan solo una variable del grupo sin importar cuál es la seleccionada. Hay evidencias que en situaciones donde n>p, y en presencia de fuertes correlaciones entre las variables predictoras, la capacidad predictiva de la regresión ridge es superior al LASSO. Regresión penalizada mediante Elastic net podría ser una alternativa, ya que combina selección de variables al igual que la regresión LASSO, y contrae conjuntamente los coeficientes de variables correlacionadas al igual que la regresión ridge. Este método ha sido aplicado en GWAS [26] y podría ser una propuesta para continuar investigando métodos de selección de variables.

6. Conclusiones

Respecto al objetivo metodológico

1. Ambos métodos (LASSO y selección mediante test univariante) detectaron con eficiencia similar la existencia de asociación de los genes K, M y N a la enfermedad de Parkinson.

2. Mientras que la capacidad predictiva de los modelos obtenidos con los dos métodos es muy similar en el estudio de simulación, en la aplicación a datos reales, el mayor valor de AUC se obtiene al emplear regresión LASSO con estimación del parámetro de penalización mediante validación cruzada.

3. La capacidad para identificar SNPs causales bajo diferentes escenarios en el estudio de simulación empleando regresión LASSO fue superior que empleando el método de selección por pvalue.

4. La capacidad predictiva de los perfiles genéticos en enfermedades complejas es escasa. Aunque un marcador presente una fuerte asociación esto no implica que sea un buen clasificador, para lo cual debería presentar Odds Ratio de magnitudes muy superiores a las encontradas habitualmente en los estudios de asociación genética.

Respecto al objetivo aplicado

1. Los genes K, M y N se asocian con la susceptibilidad a desarrollar enfermedad de Parkinson. Aunque sólo uno de los polimorfismos estudiados supera corrección FDR para comparaciones múltiples en nuestra investigación, la existencia de evidencias previas sugestivas de asociación con marcadores en los mismos genes hace probable que se trate de asociaciones verdaderas.

2. Son necesarios estudios adicionales para elucidar si las variantes genéticas asociadas identificadas tienen un papel causal en la patología o bien se trata de polimorfismos en desequilibrio de ligamiento con las verdaderas variantes causales.

3. Los modelos predictivos basados exclusivamente en los marcadores polimórficos estudiados en el presente trabajo no son de utilidad en la práctica clínica rutinaria para el diagnóstico de la enfermedad de Parkinson debido a su escaso valor predictivo. Son

necesarias investigaciones adicionales para identificar nuevos factores de susceptibilidad así como patrones genéticos más complejos que permitan desarrollar modelos predictivos basados en perfiles genéticos para esta enfermedad.

Referencias

[1] Laird, N.M. and Lange,C. (2011).*The Fundamentals of Modern Statistical Genetics*. 1st EditionXIV, 226 p.

[2] Neale, B.M. (2008). *Statistical Genetics: Gene Mapping Through Linkage and Association*. Taylor & Francis Group, 574 p.

[3] International HapMap Consortium. (2003).The International HapMap Project. *Nature,* 426:789-96.

[4] Lewontin, R.C. and Kojima, K.(1960). The evolutionary dynamics of complex polymorphisms. *Evolution*, 14:458-472.

[5] Lewontin, R.C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models". *Genetics,* 49: 49-67.

[6] Hill, W.G. and Robertson, A. (1968). Linkage disequilibrium in finite populations.*Theor Appl Genet*, 38: 226-231.

[7] Lander, E.S. (1996). The new genomics: global views of biology. *Science,* 274: 536–539.

[8] Chakravarti, A. (1999) Population genetics –making sense out of sequence. *Nat Genet*, 22: 56–60

[9] Pritchard, J.K.(2001). Are Rare Variants Responsible for Susceptibility to Complex Diseases? *Am J Hum Genet,* 69:124–137.

[10] Cirulli, E.T. and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet,* 11(6):415-425.

[11] Bellman, R. (1961) *Adaptive Control Processes.* Princeton NJ: Princeton University Press.

[12] Heidema, A.G., Boer, J.M., Nagelkerke, N,, Mariman, E.C., van der A, D.L. and Feskens, E.J. (2006) .The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases. *BMC Genet.*, 7:23.

[13] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J R Statis Soc*, 58:267-288.

[14] Ayers, K.L. and Cordell, H.J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet Epidemiol.* 34(8):879-891

[15] Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E. and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics,*25(6):714-721.

[16] Park, M.Y., Hastie,T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30-50.

[17] Armitage, P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics,* 11:375–386.

[18] Slager, S.L. and Schaid, D.J. (2001). Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Hum Hered,*52(3):149-53.

[19] Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*, 1st edition. New.York: John Wiley & Sons, Inc.

[20] Lokhorst, J. (1999).The lasso and generalised linear models. *Technical Report*. University of Adelaide, Adelaide.

[21] Shevade, S. and Keerthi, S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19: 2246–2253.

[22] Hoerl, A.E. and Keenard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12: 55-67.

[23] Hastie, T. and Tibshirani, R. (2004). Efficient quadratic regularization for expression arrays. *Biostatistic,.* 5(3):329-340.

[24] Malo, N., Libiger, O. and Schork ,N.J. (2008). Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am J Hum Genet,* 82(2):375-385.

[25] Zou, H. and and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J R Statist Soc* B, 67(2):301–320.

[26] Cho, S., Kim, H., Oh. S., Kim, K. and Park,T. (2009).Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC Proc,*3 Suppl 7:S25.

[27] Swets, J.A. and Pickett, R.M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection theory*. New York: Academic

[28] Calle, M.L., Urrea V., Boulesleix A.L. and Malats, N. (2011). AUC-RF: A new strategy for genomic profiling with Random Forest (*submitted manuscript*).

[29] Janssens, A.C., Aulchenko, Y.S., Elefante,S., Borsboom, G.J., Steyerberg, E.W. and van Duijn, C.M. (2006). Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med,*8(7):395-400.

[30] Zang, Y., Fung, W.K. and Zheng, G.(2010). Simple algorithms to calculate the asymptotic null distributions of robust tests in case-control genetic association studies in R. *Journal of Statistical Software* 33(8).

[31] Zheng, G., Freidlin, B., Li, Z. and Gastwirth, J.L. (2003). Choice of scores in trend tests for case control studies of candidate-gene associations. *Biometrical Journal,* 45,:335-348.

[32] DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988). Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, *Biometrics*, 44: 837-845.

[33] Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, 33(1):1-22.

[34] González, J.R., Armengol, L., Solé, X., Guinó, E., Mercader, J.M., Estivill, X. and Moreno, V. (2007). SNPassoc: an R package to perform whole genome association studies. *Bioinformatics,* 23(5):644-645.

[35] Barrett, J.C., Fry,B., Maller, J. and Daly, M.J. (2005).Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21:263-265.

[36] Shin, J-H., Blay, S., McNeney, B. and Graham, J. (2006). LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms. *Journal of Statistical Software,* 16 Code Snippet 3.

[37] Mickey, R.M. and Greenland, S. (1989). The impact of confounder selection criteria on effect estimation. *American Journal of Epidemiology*, 129:125-137.

[38] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

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

[40] Breiman,L. (2003). Manual for Setting Up, Using, and Understanding Random Forest V4.0. http://oz.berkeley.edu/users/breiman/Using_random_forests_v4.0.pdf

[41] Schwender, H. and with a contribution of Fritsch, A. (2011). scrime: Analysis of High-Dimensional Categorical Data such as SNP Data. R package version 1.2.6. http://CRAN.R-project.org/package=scrime

[42] Storey JD. (2002) A direct approach to false discovery rates. *J R Statist Soc B*, 64: 479-498.

[43] Pepe, M.S., Janes, H., Longton, G., Leisenring,W. and Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*, 159:882–890.

[44] Park, M.Y. and Hastie*, T. (2007). L*1-regularization path algorithm for generalized linear models. *J R Statist Soc B* 69(4)*:*659–677