Análisis y predicción de la tendencia al rechazo de riñón por perfil genético basado en SNP's

Ingeniería Informática
E.T.S. Ingeniería Informática

Alumno: LAURA SEGURA RODA.
Tutor UPV: JUAN MIGUEL GARCÍA GÓMEZ.
Tutor en Empresa: SALVADOR TORTAJADA VELERT.
Empresa: ASOCIACIÓN ITACA.

07/09/2011

Índice

RESUMEN

Justificación del estudio:

El riñón es uno de los órganos encargados de excretar las sustancias tóxicas y seleccionar aquellas útiles, devolviéndolas al organismo. El ser humano dispone de 2 riñones y, a pesar de poder vivir con un único riñón si uno falla, el organismo se resiente. Las sustancias tóxicas son devueltas al organismo y pueden dañarse el resto de órganos. Es por ello que, si se llega a un daño irreversible, es necesario realizar un trasplante. No obstante, el sistema inmunológico del paciente puede reconocer el órgano como un "invasor" y rechazarlo. Dicho rechazo puede producirse de inmediato, a los 3 meses de la operación o pasado un año. En ese caso tanto el injerto como la salud del paciente sufren un gran deterioro y es necesario reemplazar el órgano trasplantado.

Con este trabajo se pretende analizar la relación de los cambios genéticos en el ADN con la presencia del rechazo post-trasplante, así como crear un modelo predictivo que permita predecir si, dado un perfil genético e información clínica del paciente, va a producirse rechazo o no, previniendo dicha problemática y aumentando la calidad de vida de los pacientes que padecen enfermedades renales crónicas con necesidad de trasplante.

Objetivos del estudio:

En primer lugar, se desea realizar una aproximación médica del rechazo de trasplante renal y de cómo se llega a tal situación, viendo la necesidad de mejorar la situación de muchos pacientes afectados por esta enfermedad.

Además, se ve conveniente explicar cómo el perfil genético de los individuos puede influir en la respuesta a un medicamento o ante una enfermedad y, por lo tanto, puede influir en el rechazo post-trasplante.

Para poder crear un modelado predictivo potente con el menor número de variables posible, se deben realizar diversos análisis, tanto estadísticos como genéticos, para poder reducir el conjunto de variables inicial. Una vez obtenido dicho conjunto, se evalúan diversos algoritmos de modelado, con el fin de encontrar el que mejor prediga la aparición o ausencia de rechazo post-trasplante, así como el tipo de rechazo.

Materiales:

La muestra que se ha analizado es de 276 pacientes, con edades comprendidas entre 18 y 87 años. De todos ellos, 118 individuos no han presentado rechazo post-trasplante y sí los 158 pacientes restantes. De entre estos últimos, 58 individuos presentan rechazo crónico, mientras que 100 presentan otros tipos de rechazo.

Todos los pacientes están definidos por 11 variables clínicas y 42 variables genéticas. Respecto a las variables clínicas, éstas son la edad y el sexo del paciente, la edad y el sexo del donante, el tiempo de isquemia durante el proceso de trasplante, la causa de la muerte del donante, la enfermedad primaria del paciente, si éste tuvo uropatía obstructiva previa, así como grado de compatibilidad entre el donante y receptor HLA-A, HLA-B y HLA-DR.

Respecto a las variables genéticas, se trata de un grupo de 42 SNPs que indican los genotipos que presentan para cada paciente.

Métodos:

Para comprobar la existencia de datos anómalos, así como para determinar el número de valores faltantes en alguna variable, se ha utilizado un histograma de distribución. Posteriormente, al observar que era necesario imputar los genotipos faltantes, por un lado, se ha creado un script para realizarlo según el valor más frecuente y, por otro, se ha empleado la herramienta IMPUTE. Tras la imputación, se ha comprobado con los algoritmos K-medias y bietápico si ha habido variación en las proporciones iniciales.

Mediante la herramienta Haploview y la librería haplo.stats de la herramienta R se ha realizado un análisis por haplotipos con la enfermedad. Posteriormente, se ha analizado la asociación de cada SNP por separado y de las interacciones entre pares de SNPs mediante la librería SNPassoc de la herramienta R. Asimismo, se ha realizado un análisis de asociación con la enfermedad de las variables clínicas mediante matrices de contingencia y de correlación, en función de la naturaleza discreta o continua de dichas variables.

Una vez seleccionadas las variables de asociación significativa respecto al rechazo post-trasplante, se ha realizado el modelo predictivo. Para ello se han empleado los algoritmos LDA, SVM y árboles de decisión, de manera que, el que obtenga mejores resultados, será el modelo definitivo. Los modelos se han evaluado mediante validación cruzada para emplear el mayor número de registros como entrenamiento; no obstante, también se ha considerado oportuno realizarlo mediante particiones independientes, con tal de determinar la generalidad de cada modelo.

Resultados:

Mediante los análisis estadísticos de asociación entre las variables, se encontraron numerosas asociaciones entre ellas. Con tal de afinar más dichas asociaciones y poder reducir el conjunto de variables que iban a emplearse en el modelo, se realizaron diversos análisis por haplotipos, encontrando que los haplotipos formados por los SNP's rs1800872, rs1800896 y rs699, del cromosoma 1; rs1143634, rs2234676 y rs419598, del cromosoma 2; rs1801275 y rs243865, del cromosoma 16; rs4586 y rs2107538, del cromosoma 17 y rs1799969 y rs1800471, del cromosoma 19 estaban asociados con la ausencia/presencia del rechazo post-trasplante. Mediante el estudio de asociación univariante no se obtuvieron resultados concluyentes, sin embargo, mediante el análisis de pares de interacciones entre SNP's se encontraron que las interacciones entre los pares rs1143634 - rs1799750, rs1799969 - rs3918226 y rs1800872 - rs2243248 estaban asociadas tanto con la variable dependiente DCTRsi_no como con DCTR_otrDCTR.

Una vez realizados los modelos utilizando los SNP's que formaban interacciones y haplotipos asociados con las variables dependientes, se obtuvo que el mejor modelo era el árbol de decisión formado tanto por dichos SNP's como por las variables clínicas, con un 86% de aciertos en la clasificación mediante validación cruzada.

Discusión:

En primer lugar, se han observado 3 interacciones entre pares de SNP's de diferentes cromosomas asociados con la presencia/ausencia del rechazo post-trasplante, así como 6 haplotipos significativos. Ello indica que ciertas combinaciones del perfil genético de los seres humanos están involucradas en el rechazo del trasplante de riñón, permitiendo, con la ayuda de la información clínica del paciente, poder tener una seguridad del 86% de si se va a producir rechazo o no. No obstante, si se desea detectar si, además de presentar rechazo, el tipo de rechazo que se va a dar, los resultados no son tan satisfactorios, obteniendo un 80% de éxito en la predicción con la ayuda de la información clínica del paciente.

Conclusiones:

Este trabajo ha supuesto un primer avance en la predicción del rechazo post-trasplante mediante el perfil genético del paciente ya que se han encontrado 6 haplotipos significativos y 3 pares de interacciones asociadas a la presencia/ausencia del rechazo. Además, empleando los SNP's que forman dichas interacciones y haplotipos, con ayuda de algunos datos del historial clínico del paciente, se puede realizar una clasificación satisfactoria con tan sólo un 14% de posibilidad de error.

De esta manera, el rechazo post-trasplante, puede ser detectado antes de que aparezca, ganando tiempo para iniciar el correspondiente tratamiento.

1. INTRODUCCIÓN

1.1.Problema médico

La insuficiencia renal crónica produce daños en los riñones y hace que disminuya la habilidad de los mismos para eliminar los productos de desecho y el exceso de líquidos del organismo. Si esta enfermedad se agrava, los productos de desecho que genera el cuerpo humano alcanzan altos niveles y hacen que aparezcan síntomas como altos niveles de presión arterial, anemia, daños del sistema nervioso, huesos débiles, entre otras complicaciones. Además, se incrementa el riesgo de padecer enfermedades del corazón y de los vasos sanguíneos. Estos problemas pueden aparecer lentamente después de un largo periodo de tiempo. Así, la detección temprana y el tratamiento adecuado pueden hacer que la enfermedad no pase a mayores consecuencias; no obstante, si evoluciona, puede conducir a fallo renal y, por consiguiente, a la necesidad de diálisis o de trasplante.

El trasplante renal consiste en una operación en que la persona cuyo riñón ha fallado recibe uno nuevo para reanudar la función renal perdida y continuar con la vida habitual. Sin embargo, pueden ocurrir diversas complicaciones, siendo la más importante el rechazo renal. El sistema inmunológico previene de los ataques de cualquier elemento desconocido, como por ejemplo, bacterias o virus. Dicho sistema puede reconocer el tejido trasplantado como algo desconocido y combatir contra este “invasor”, rechazándolo.

Se tienen diferentes tipos de rechazo, siendo el rechazo crónico (o nefropatía crónica del trasplante, NCT) la causa más frecuente de pérdida tardía del injerto. Actualmente, no se conoce exactamente el origen de la NCT, aceptándose que su desarrollo depende tanto de factores inmunológicos como de otros que no lo son. Además, no se conoce ningún medicamento, tanto para la prevención como para el tratamiento, que tenga una eficacia clínica convincente.

1.2.Mutaciones y SNP’s

EL ADN contiene la información genética de las células de nuestro cuerpo. Dicha información, organizada en unidades llamadas cromosomas, controla el mecanismo durante vida de la célula y hace que se desarrolle con normalidad. El material genético se hereda de padres a hijos, manteniéndose la mayoría de las partes del mismo intactas. A los patrones de estas partes conservadas los llamaremos haplotipos.

Si durante la transmisión de la información genética se producen cambios o bien la cadena de ADN se ve modificada a lo largo de la vida de un ser humano, aparecen las mutaciones. Éstas no siempre causan daños en el individuo ya que dependen del lugar del cromosoma en el que tengan lugar. Una mutación supone que, dada una secuencia de ADN (de una o varias posiciones) predominante en la población, ésta sufre un cambio en un individuo, produciendo una variante distinta.

No obstante, pueden darse diferentes cambios y, por lo tanto, puede haber diversas variantes posibles, dando lugar a individuos con diferentes alteraciones, igualmente válidas, llamadas polimorfismos. La

frontera entre polimorfismo y mutación es que, para el caso de los polimorfismos, la frecuencia de la variante menos común esté presente en más de un 1% de la población; mientras que, para que se considere mutación, la frecuencia debe ser inferior al 1%.

Aquellos polimorfismos que implican únicamente un cambio en una posición de la cadena de ADN se les llama polimorfismo de nucleótido simple, o SNP. Estos cambios son el objetivo principal de los investigadores ya que se está demostrando que dichos cambios, interactuando con los de otros lugares del ADN, podrían explicar muchas enfermedades.

1.3.Bioinformática

Las nuevas tecnologías de secuenciación y microarrays de expresión genética generan grandes cantidades de información, que únicamente se pueden analizar de forma eficaz mediante computadores, siendo la bioinformática quien aporta las herramientas necesarias para el análisis de dicha información.

La bioinformática podría definirse como el conjunto del tratamiento, análisis, predicción y modelado de la información biológica mediante la ayuda de los ordenadores.

Además, es una disciplina que une esfuerzos procedentes de diversas áreas. Así pues, la informática y las telecomunicaciones permiten desarrollar los sistemas, programas e infraestructuras necesarios para solucionar los problemas que la biología, farmacia o medicina puedan proponer. Además, los modelos utilizados emplean la estadística y se optimizan mediante técnicas puramente matemáticas.

Abarca numerosas aplicaciones, desde procesos de detección y modelado de genes para identificar regiones codificantes hasta la creación de herramientas como NCBI o EMBL, que permiten buscar en bases de datos información completa de secuencias de nucleótidos o de proteínas o ayudar a la interpretación de mutaciones en ciertas patologías.

La bioinformática se integra como una componente tecnológica en proyectos dirigidos a la tecnología de los alimentos y la agricultura, así como en proyectos multidisciplinares dirigidos a la Salud, con el fin de apoyar el estudio genético de las enfermedades, su diagnóstico automático, entre otros. Además, la incorporación de factores externos a la experimentación con datos genéticos es crucial para deducir la influencia medioambiental en la expresión y regulación genética.

1.4.Aprendizaje automático

El aprendizaje automático es una técnica que permite elaborar un modelo a partir de la muestra de un problema, de manera que dicho modelo se pueda generalizar a nuevos casos. Incluye diversas aplicaciones como son el análisis, el diagnóstico médico, la clasificación o la detección de fraudes de tarjetas de crédito.

La aplicación que resulta de interés en este trabajo es aquella que predice el diagnóstico o tratamiento adecuado para un paciente a partir de la información que se dispone del mismo, conocida, de manera general, como Sistema de Ayuda a la Decisión (o Computer-based Decision Support System, CDSS).

Para la elaboración de un sistema de ayuda a la decisión, se deben buscar regularidades o patrones en la información proporcionada. Dichos patrones pueden especificarse mediante reglas creadas a partir del conocimiento del dominio del problema. Para ello, se toma un conjunto de datos representativo del mismo, llamado dataset o corpus.

Posteriormente, se emplea un subconjunto de los datos para entrenamiento. En esta etapa, se hace un preprocesado de los datos, incluyendo una serie de pasos como transformación, normalización o cambios de escala para adecuar el conjunto al uso posterior. Además, se realiza una selección y extracción de características, comparando diferentes alternativas, con el fin de obtener una óptima representación de los datos. Una vez realizada la extracción de características, se obtiene el modelo predictivo, que dará lugar a la regla de decisión del sistema.

Con el modelo de predicción creado, se debe comprobar su uso en un entorno médico. Para ello, en esta etapa de validación, se introduce un nuevo corpus de test, realizando el mismo preproceso que el seguido durante el entrenamiento y aplicando el método de selección de características previo.

1.5.Objetivos

Durante la realización del proyecto, se desarrollan los siguientes objetivos:

- Estudiar la insuficiencia renal y el rechazo en los trasplantes de riñón desde un punto de vista médico; haciendo hincapié en la importancia de obtener una solución a esta problemática o, al menos, de encontrar un método que permita la detección del rechazo a tiempo.
- Analizar los cambios genéticos que ocurren en el ADN y cómo éstos influyen en los individuos, originando una tendencia a padecer una enfermedad o produciendo una determinada respuesta ante un medicamento.
- Determinar si, debido a la metodología de extracción de la información genética de los pacientes, existen algunos valores faltantes. En ese caso, es necesario realizar una imputación de los mismos.
- Analizar la información genética mediante métodos que permiten aprovechar al máximo las características intrínsecas de la misma, como son: diferentes formas de herencia, cambios en el ADN durante la misma, entre otros.
- Analizar, mediante métodos estadísticos generales, la información de manera conjunta, con el fin de obtener asociaciones de interés en relación al rechazo.
- Realizar diferentes modelos predictivos con las variables significativas encontradas y evaluar cuál permite predecir el rechazo en los trasplantes de riñón, así como el tipo del mismo, de la manera más eficaz y eficiente.

1.6. Contribuciones

La idea de este proyecto ha surgido de un grupo de investigadores del Instituto Carlos III de Madrid, con quien se ha colaborado para la realización de un proyecto de características similares. En él se trataba de determinar la presencia o ausencia de la enfermedad renal en la población, siendo el grupo de casos enfermos en dicho proyecto el conjunto completo de datos que se ha utilizado para este trabajo. Además, dicha colaboración ha permitido el envío a una revista electrónica de un artículo explicando el proceso llevado a cabo para la realización de dicho proyecto y los resultados obtenidos.

Además, para la realización de este proyecto se ha participado en un programa de prácticas en empresa en el ITACA durante 6 de los meses en los que se ha estado realizando el proyecto.

2. PROBLEMA MÉDICO

En este apartado comenzaremos con una pequeña introducción del funcionamiento del riñón y su estructura para pasar finalmente al concepto de trasplante renal, cómo se llega a tal situación y las posibles complicaciones que se pueden dar.

2.1.El riñón y su estructura

El riñón pertenece al sistema excretor de los seres vertebrados. En concreto, la anatomía y fisiología del riñón humano, que ha evolucionado durante miles de años, le permiten excretar los residuos del organismo a través de la orina, regular los procesos homeostáticos y producir ciertas hormonas.

Los riñones están situados en la parte inferior de la caja torácica, cada uno a un lado de la columna vertebral, estando el izquierdo ligeramente más alto que el derecho. Ambos son idénticos en cuanto a estructura y funcionalidad. Tienen forma de haba y un tamaño alrededor de 10 cm de largo y 6.5 cm de ancho, comprendiendo un córtex exterior y una médula interior (The Human Kidney Structure and Function s.f.).

Cada órgano está formado por aproximadamente un millón de nefronas; su unidad funcional. En ella se pueden distinguir principalmente los siguientes elementos:

- Cápsula de Bowman: constituida por una doble pared de células epiteliales que rodea un conjunto de capilares sanguíneos denominado glomérulo de Malpighi. El conjunto de ambos se denomina corpúsculo de Malpighi.
- Túbulo contorneado proximal: localizado en la zona cortical del riñón, se halla a continuación de la cápsula de Bowman.
- Asa de Henle: es la continuación del túbulo contorneado proximal. Conduce hacia el interior de la zona medular del riñón y en ella se distinguen dos partes: la rama ascendente y la descendente.
- Túbulo contorneado distal: localizado en la zona cortical del riñón, desemboca en el túbulo colector (Concha Gil Soriano, Capítulo 11 2002).

A continuación se muestra la estructura interna, tanto del riñón como de la nefrona, de forma más detallada:

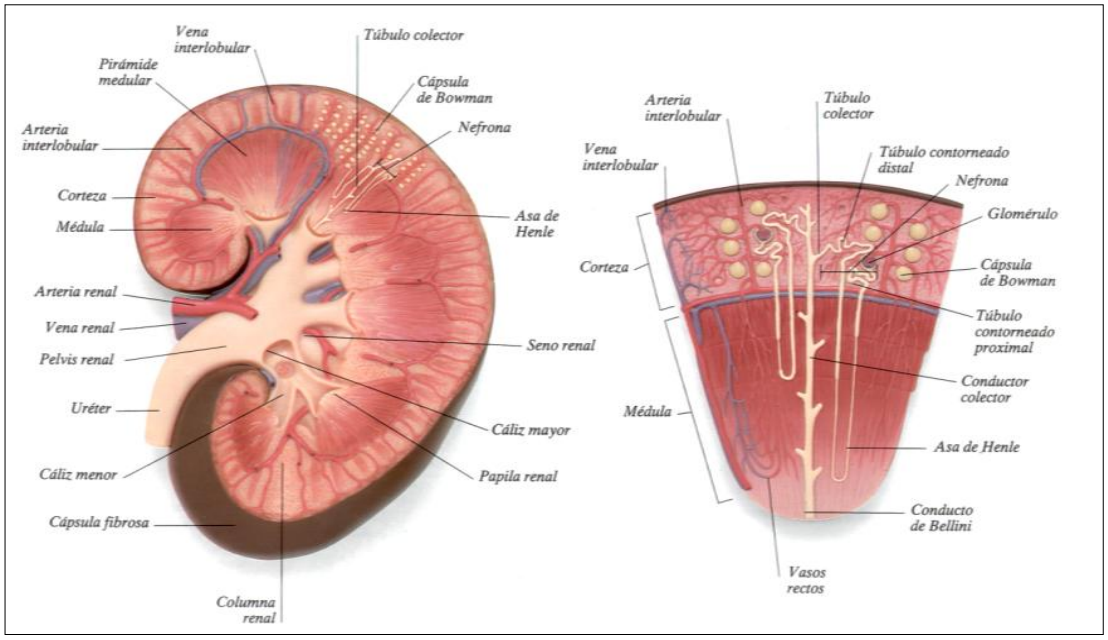


Figura 1 - http://html.rincondelvago.com/enfermedad-poliquistica-del-rinon.html

Una de las principales funciones del riñón es la formación de la orina. Los productos de desecho le llegan a través de la arteria renal, iniciando un proceso que se divide en tres etapas:

- Filtración glomerular: tiene lugar en la cápsula de Bowman, donde, desde los capilares, se filtran hacia el interior de la cápsula productos como agua, sales minerales, glucosa, aminoácidos, vitaminas, entre otros. Por ejemplo, la urea y el ácido úrico son productos de desecho que contienen nitrógeno, fruto de los procesos metabólicos, y deben ser filtrados (National Space Biomedical Research Institute s.f.).

- Reabsorción: se realiza en los túbulos que forman la nefrona y tienen por objetivo recuperar las sustancias aprovechables que se hayan filtrado en la cápsula de Bowman. En cada uno de los tramos se reabsorben diferentes sustancias, que pasan nuevamente a la sangre a través de numerosos capilares que rodean la nefrona. Por ejemplo, el filtrado y reabsorción de glucosa ayudan a mantener correctos niveles de azúcar en sangre (National Space Biomedical Research Institute s.f.).

- Secreción: es un proceso por el que algunas sustancias pasan desde los capilares que rodean la nefrona al interior de esta, especialmente al interior del túbulo distal. Esta secreción tiene importancia en el mantenimiento de la concentración de algunos iones, como hidrógeno (H+) y potasio (K+) (Concha Gil Soriano, Capítulo 11 2002). La secreción de iones de hidrógeno junto con el control los niveles de bicarbonato, mantienen el pH adecuado de la sangre. Cuando la sangre es demasiado ácida, se están secretando demasiados iones de hidrógeno; mientras que si se vuelve alcalina, es porque se ha reducido la secreción de dichos iones (The Human Kidney Structure and Function s.f.) (The Kidney 2011).

Al final se forma la orina, que pasa a los túbulos colectores del riñón. Estos confluyen y terminan en dos uréteres, por donde va la orina hasta la vejiga, en la que se almacena hasta ser eliminada al exterior por la uretra. La sangre filtrada sale por la vena renal y reparte las sustancias aprovechables al resto del cuerpo.

Además de la función excretora y homeostática, los riñones producen la enzima renina, un importante regulador de la presión sanguínea, y liberan dos hormonas a la sangre (The Human Kidney Structure and Function s.f.) (The Kidney 2011). Éstas son:

- Eritropoyetina, que actúa sobre la médula ósea para que aumente la producción de glóbulos rojos.
- Calcitriol, que incrementa la absorción de calcio de los alimentos en el intestino y actúa directamente en los huesos regulando el calcio en el torrente sanguíneo.

Por lo tanto, los riñones son potentes máquinas que realizan las siguientes funciones (National Kidney Federation s.f.):

- Eliminar los productos de desecho del organismo.
- Eliminar medicamentos del organismo.
- Equilibrar los líquidos corporales.
- Liberar hormonas que regulan la presión sanguínea.
- Producir una forma activa de vitamina D que proporciona fuerza y salud a los huesos.
- Controlar la producción de glóbulos rojos.

2.2.Insuficiencia renal

Antes de llegar a la necesidad de un trasplante renal, se pasa por una etapa de insuficiencia renal, en la que los riñones dejan de filtrar la sangre correctamente y no pueden excretar las sustancias tóxicas.

Aunque el cuerpo humano puede sobrevivir únicamente con el funcionamiento de un riñón, el fallo renal se refiere a la pérdida de actividad en ambos riñones.

Los síntomas son muy diferentes de una persona a otra, mientras unos presentan muchos síntomas, otros desconocen que tienen insuficiencia renal. Algunos de estos síntomas son elevados niveles de urea en el torrente sanguíneo, orina más oscura de lo habitual, necesidad de ir al baño por la noche, náuseas, vómitos, diarrea, entre otros.

Principalmente se distinguen 2 tipos de insuficiencia renal, insuficiencia aguda e insuficiencia crónica.

La primera se produce cuando el riñón deja de funcionar repentinamente, haciendo que se acumulen gran cantidad de productos de desecho en la sangre. En función de la afección que causa la insuficiencia y de la gravedad de la misma, se presentan unos problemas u otros (Merck Sharp & Dohme s.f.). Los principales son:

- Disminución del riego sanguíneo a los riñones, debido a una pérdida súbita de gran cantidad de sangre, a lesiones físicas que obstruyan los vasos sanguíneos, a una insuficiencia cardíaca o a una insuficiencia hepatorrenal.

- Obstrucción de la orina excretada. Hay algunas enfermedades, como la hidronefrosis, que producen dicho problema y la orina debe retroceder al interior de los riñones. También puede deberse a algún tumor que esté presionando las vías urinarias, obstruyendo la salida de la vejiga y provocando un aumento de tamaño de la misma.

- Lesión interna de los riñones, debido a un tratamiento prolongado de determinados fármacos; presencia de cristales, proteínas u otro tipo de sedimentos que no deberían depositarse en los riñones.

Además de exámenes físicos o de imagen para comprobar el tamaño de los riñones y vejiga, también se puede determinar la presencia de insuficiencia renal aguda mediante una analítica.

Como se ha comentado anteriormente, los riñones se encargan de regular la excreción de iones de hidrógeno y de potasio; así, si se presentan valores anormales de acidez (debido a iones de hidrógeno) o de potasio, posiblemente se tenga algún problema en la excreción de sustancias.

Otro de los indicios a través de una analítica sería valores fuera de lo regular de urea y creatinina:

- Durante la digestión de las proteínas, algunos aminoácidos se transforman en urea en el hígado y ésta es transportada por la sangre hasta los riñones para ser filtrada y excretada.

  Antes de acabar en la orina, las sustancias pasan por los túbulos para llevar a filtrar los componentes de desecho que transportan y reabsorber las sustancias que puedan ser importantes. Como dichos túbulos son permeables a la urea, si la tasa de filtración glomerular es baja, cuanto más tiempo permanezca el líquido en los túbulos mayor será la cantidad de urea reabsorbida y enviada a la sangre de nuevo (Sterling s.f.).

- Los músculos utilizan para su funcionamiento la creatina y fosfocreatina que, tras degradarse, se convierten en creatinina. Ésta es también una de las sustancias que debe filtrarse por los riñones y ser excretada en la orina (Creatinina s.f.). Si la tasa de filtración glomerular es demasiado baja, aumentarán los niveles de esta sustancia; de la misma manera que si el filtrado es muy elevado, los niveles de creatinina serán demasiado bajos.

Esta insuficiencia puede ser reversible si los riñones no han sufrido demasiado daño, de lo contrario, deriva en fallo renal permanente.

El segundo tipo de insuficiencia se produce cuando hay daño permanente e irreversible en la función de los riñones. Puede ser producida por diversas causas como la complicación de enfermedades renales (glomerulonefritis, enfermedad poliquística renal, obstrucción del tracto urinario); diabetes mellitus o debido a antecedentes familiares (Merck Sharp & Dohme s.f.).

Según los riñones van perdiendo su capacidad para excretar las sustancias de desecho, los niveles de urea y creatinina aumentan considerablemente, derivando en azoemia. Debido a que los riñones no pueden

eliminar el exceso de agua y sal como hacen de costumbre, suele aparecer hipertensión; así como anemia, debido al descenso de producción de los componentes de los glóbulos rojos.

Finalmente, la composición alterada de la sangre llega a producir problemas del sistema nervioso y, conforme la acumulación de sustancias de desecho en la sangre es mayor, también puede haber problemas cutáneos y del aparato digestivo.

Además, puede aparecer un tercer tipo de insuficiencia, aguda sobre crónica, en la que la insuficiencia aguda reside junto con la crónica. Este tipo sólo puede ser detectado si se ha llevado antes un seguimiento médico mediante analíticas y, así, poder realizar una comparación con estados anteriores. Al igual que de forma aislada, la parte aguda puede ser reversible con tratamiento mientras que la crónica, no (Insuficiencia renal s.f.).

Si la función renal se deteriora poco a poco, descubriéndola a tiempo, con una dieta determinada y medicación, se puede retrasar la necesidad de diálisis y trasplante renal. Si, por el contrario, la insuficiencia es (casi) total y permanente, se está en un estado terminal en el que la persona debe someterse obligatoriamente a diálisis o a un trasplante.

Ésta última es la situación de interés.

2.3.Trasplante renal

Los trasplantes de órganos son, a veces, la única alternativa para los pacientes con insuficiencias terminales. Al principio, esta práctica supuso un gran impacto en la sociedad y la opinión pública se dividió entre quienes defendían la nueva práctica y aquellos que la condenaban. El hecho de sustituir un órgano que había dejado de funcionar en una persona con posibilidades de sobrevivir por otro órgano, en buenas condiciones, pero de una persona fallecida, no era aceptado por todos (Daga Ruiz 2008).

Sin embargo, poco a poco, los beneficios que se han ido comprobando a lo largo de los años, han ganado terreno a las opiniones desfavorables y, actualmente, se trata de una práctica común en los países de medianos y altos ingresos.

Además, se han creado leyes para que medien entre el derecho al propio cuerpo y la identidad y la ayuda a la sociedad con la donación de órganos, ya que ésta tiene que ser de forma altruista y voluntaria y supone devolver a la vida a una persona que estaba en peligro o, al menos, mejorar su calidad de vida.

En concreto, los trasplantes renales implican mejorar la situación del paciente.

Cuando una persona presenta insuficiencia renal crónica y la disminución de la función del riñón es muy alta, necesita someterse a diálisis o trasplante.

Esta última opción, a pesar de que la persona tenga que estar siempre sometida a medicación para evitar el rechazo del injerto, garantizaría llevar una vida normal, ya que no tendría que depender de una máquina

para sobrevivir, recuperaría la función renal perdida y no estaría sometido a restricciones severas en su dieta.

En líneas generales, el trasplante renal mejora la calidad de vida, tanto física como psíquica, siendo más destacada en hombres que en mujeres. En concreto, mejoran la actividad física y mental en un 80% y la social en un 60%, aunque siempre existen factores directamente asociados al trasplante renal que, en algunos casos, limitan el aumento de la calidad de vida (Daga Ruiz 2008).

Hoy en día, el trasplante es la solución de elección para la mayoría de los pacientes en estado de insuficiencia renal terminal, principalmente producida por enfermedades primarias como la diabetes, hipertensión o complicaciones de enfermedades renales como glomerulonefritis o genéticas como la poliquistosis renal.

Es el tratamiento más económico en comparación con la diálisis, aunque viene limitado por el número reducido de donantes frente a la gran demanda de enfermos que lo requieren. Esto produce una lista de espera alrededor de 4.500 personas (en España), quienes mientras no pueden ser trasplantadas deben someterse a diálisis durante un periodo de aproximadamente 2 años.

Antes de que se produzca el trasplante se deben evaluar una serie de condiciones, en especial la edad, antecedentes y enfermedades asociadas. Lo mejor sería llevar el seguimiento al paciente y hacerle el estudio pre-trasplante antes de entrar en fase de diálisis, ya que se trata de un tratamiento programable y algunos estudios apuntan a que la supervivencia del injerto y del paciente es mayor si no está en lista de espera con diálisis (P. Marti 2006) (M. Pérez Fontán 2000).

Los trasplantes pueden ser con órganos de donante vivo, que puede estar emparentado con el receptor o no, o de donante fallecido.

Éste último puede ser de dos tipos (Trasplante de riñón s.f.):

- Donante en muerte encefálica: producida principalmente por un accidente cardiovascular agudo (ACVA) o traumatismo craneoencefálico (TCE). Se trata de individuos que han perdido de forma irreversible las funciones cerebrales pero a sus órganos les sigue llegando riego sanguíneo con el bombeo del corazón, mantenido de forma artificial en el hospital.

- Donante en asistolia (“a corazón parado”), individuos que han sufrido paro cardíaco irreversible, haciendo que deje de llegar la sangre a los órganos.

Uno de los motivos por los que se realiza trasplantes de donante vivo es debido a la escasez de donantes cadáver, pero también porque ofrece ciertas ventajas sobre éstos.

Si se trata de donante vivo, debido al escaso tiempo en que el órgano a trasplantar está sin riego sanguíneo (tiempo de isquemia prácticamente inexistente), el riesgo de retraso en la recuperación de la función renal tras el trasplante es menos probable, lo cual implica que la supervivencia del injerto y del paciente sea mayor. De hecho, a los 10 años de haberse realizado el trasplante con un donante vivo, los resultados de supervivencia del injerto son muy buenos y, en general, entre un 15% y 20% mayores que con donante cadáver.

Sin embargo, según fuentes de la ONT (Organización Nacional de Trasplantes 2009), a pesar de que España sea uno de los países líder en trasplantes, solamente se realiza un 10% de trasplantes de donante vivo, una de las cifras más bajas de los países desarrollados, frente a más de un 50% de Estados Unidos o Brasil.

Respecto a los trasplantes con donante cadáver, si el individuo ha sufrido muerte cerebral, sus órganos han pasado menos tiempo sin riego sanguíneo que con muerte cardíaca y la supervivencia del injerto es mayor. No obstante, según la bibliografía encontrada, hay estudios que indican que la muerte encefálica desencadena una serie de mecanismos que incrementan la respuesta inmunológica aguda del receptor tras el trasplante, lo que supondría el rechazo del tejido (Marta Crespo Barrio 2005).

A lo largo de este proyecto se va a tratar el rechazo del órgano trasplantado, que dificulta la supervivencia del tejido y, por tanto, del paciente.

2.4.Rechazo post-trasplante

Estudios realizados durante el año 2000, apuntaban que en algunos países como Chile, de la gente que esperaba un trasplante renal, 30% ya tenía un trasplante previo y el 20% lo esperaba por segunda vez, lo cual implica que el rechazo del injerto provocaba serios problemas, no sólo porque engrosase las listas de espera, sino porque se perdía un riñón y con ello el paciente se debilitaba cada vez más (Mocarquer s.f.).

Tras la realización del trasplante, pueden surgir complicaciones, algunas precoces y otras tardías, siendo la pérdida de la función renal o rechazo del injerto una de las más frecuentes.

El trabajo del sistema inmunológico es luchar contra los invasores del organismo, como por ejemplo, gérmenes u objetos como astillas de madera. El cuerpo reconoce estos cuerpos extraños y los elimina del organismo. La sangre no sólo lleva oxígeno y nutrientes a todas las partes del cuerpo, sino que también lleva defensas allí donde son necesarias. Existen dos tipos: uno son los glóbulos blancos, que eliminan las bacterias, y el otro son los anticuerpos, que también las eliminan y ayudan a los glóbulos blancos a ello.

El sistema inmunológico reconoce qué parte es del organismo y cuál no y, aunque las trasfusiones de sangre se realizan sin rechazo, se trata de una excepción, y órganos tales como los riñones, el hígado, etc. se consideran “invasores”. Aunque estos órganos provengan de la misma especie, cada uno (incluidos aquellos procedentes de gemelos) es diferente, y el organismo puede reconocer estas diferencias. Así, el daño que puede provocar el sistema inmunológico a un riñón trasplantado de una persona a otra, se le llama rechazo (National Kidney Federation 2010).

Incluso cuando dos individuos son compatibles, en términos de grupo sanguíneo y tipo de tejido, es común que se presente algún grado de rechazo. Afortunadamente, existen fármacos, llamados inmunosupresores, que ayudan a prevenir y tratar la evolución del rechazo.

Existen diversos tipos de rechazo, agrupándose principalmente en función de cuándo y cómo se manifiestan de la siguiente forma:

- Rechazo hiperagudo: aparece tan pronto como el órgano es colocado en el cuerpo. Ocurre únicamente si los anticuerpos del receptor reaccionan ante el nuevo órgano debido a

incompatibilidades sanguíneas o de tejido entre el donante y receptor. Casi nunca sucede, ya que los equipos de trasplante comprueban todos los aspectos relacionados con la compatibilidad. No obstante, si se da rechazo hiperagudo, lo más probable es que el receptor muera durante o inmediatamente después de la operación (National Center for Research Resources 2009).

- Rechazo agudo: implica rechazo a corto término y de rápida aparición, necesitando acción inmediata. Puede aparecer durante los primeros meses, en particular, las primeras semanas, después del trasplante. Es muy habitual, cerca del 40%, que las personas experimenten rechazo agudo en los primeros 3 meses después de un trasplante. Si aparece, puede ser un indicador de aparición de rechazo crónico y el paciente es tratado con medicamentos inmunosupresores (National Kidney Federation 2010).

  Por otro lado, si no aparece después de un año de la operación, es poco probable que lo haga posteriormente, siempre y cuando se siga con regularidad el tratamiento.

- Rechazo crónico, nefropatía crónica del injerto (NCI) o glomerulopatía del trasplante: supone rechazo a largo plazo y comienza lentamente. El sistema inmune puede atacar y rechazar el riñón trasplantado, pero de diferente manera que en el caso anterior.

  El efecto se puede expresar como un envejecimiento prematuro del nuevo riñón. Puede aparecer con niveles bajos de rechazo producido por los anticuerpos, por hipertensión, etc. como en cualquier otro rechazo. Si aparece, suele ser a partir de un año después de la operación. Uno de los indicadores es que los niveles de creatinina se incrementan lentamente después de haber estado estables durante un tiempo. La única manera de diagnosticarlo es con una biopsia, al igual que en el tipo anterior, pero no hay ningún tratamiento que garantice el éxito (National Kidney Federation 2010).

  El rechazo crónico varía en función de la gravedad, pudiéndose estabilizar sin causar más consecuencias. Sin embargo, otros tipos más severos conducen al fallo del riñón y, por tanto, el paciente tiene que volver a la diálisis o a necesitar un nuevo trasplante. Puede tardar muchos años en aparecer, pero es la causa más común de fallo del trasplante después del primer año de la operación (Mocarquer s.f.).

Como se ha visto, los trasplantes renales son la mejor solución para la mayoría de personas con insuficiencia renal crónica, pero el tratamiento inmunosupresor, la edad del donante y diversos factores que no están exactamente definidos, producen en gran parte de ocasiones el rechazo del injerto y el regreso a la situación inicial o a una en peores condiciones.

Además, para mejorar la supervivencia del injerto en caso de donante cadáver, hay estudios que apuntan a factores como el sexo femenino de donante y receptor, ausencia de rechazo, edad del receptor (> 14 años) y creatinina del donante por debajo de 2,5 mg/dl en el momento del trasplante (Marta Crespo Barrio 2005).

Respecto a trasplantes de donante vivo, para un incremento de la supervivencia del injerto se apunta únicamente a la ausencia de rechazo y la edad del receptor (> 14 años) (Marta Crespo Barrio 2005).

También hay otros análisis que asocian a la supervivencia del injerto, además de la ausencia de rechazo, la edad y el sexo del donante, el tiempo de isquemia y la compatibilidad HLA-DR, pero en ningún caso se

obtiene ninguna conclusión completamente definitiva (P. Marti 2006) (Marta Crespo Barrio 2005) (E. Gallego Valcarce 2010).

Así pues, el fin del proyecto en que se engloba este trabajo es determinar diversos factores que permitan predecir dicho rechazo con el menor error posible.

3. EL MODELO BIOLÓGICO

3.1.El ADN

Todos los seres, incluidos los humanos, tienen un genoma que contiene toda la información biológica necesaria para construir y mantener un organismo. Dicha información se codifica gracias al ADN o ácido desoxirribonucleico, las huellas dactilares del ser humano.

Cada célula en el cuerpo de un individuo tiene aproximadamente el mismo ADN. La mayoría, se localiza en el núcleo de las células (llamado ADN nuclear), pero también existe una pequeña cantidad de ADN en las mitocondrias (llamado ADN mitocondrial).

El ADN nuclear se organiza en pequeños cuerpos en forma de palillos, llamados cromosomas. Un cromosoma es un segmento de ADN enroscado, formado por fragmentos involucrados en propósitos estructurales, en la regulación del uso de la información genética y formado por las unidades encargadas de llevar dicha información, conocidas como genes.

Dado que los seres humanos son organismos diploides, se tiene un cromosoma homólogo, formado por dos cromosomas, uno procedente del padre y otro procedente de la madre. Ambos contienen los mismos genes, situados en la misma posición pero con diferente información.

A nivel molecular, la información en el ADN se almacena como un código formado por 4 bases nitrogenadas: adenina (A) y guanina (G), llamadas purinas, y timina (T) y citosina (C), llamadas pirimidinas. El ADN humano consiste en 3 billones de bases, y más del 99% de estas bases son las mismas en toda la población (What is DNA? s.f.). Por lo tanto, un gen es esencialmente una frase formada por las bases A, T, G y C.

Las bases de ADN se emparejan unas con otras, A con T y C con G, para formar unidades llamadas pares de bases. Cada base está también ligada a una molécula de azúcar y a otra de fosfato. A la unión de una base, un grupo azúcar y un grupo fosfato, se le llama nucleótido. Los nucleótidos se disponen en 2 largas hebras que forman una espiral de doble hélice, conocida como la cadena de ADN.

La estructura de doble hélice es algo similar a una escalera, con los pares de base en el interior, formando los peldaños, y las moléculas de azúcar y fosfato en el exterior, formando las varillas verticales de la escalera.

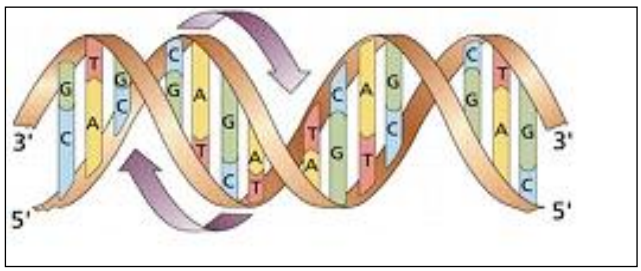


Figura 2: http://www.ehu.es/biomoleculas/an/an41.htm

En concreto, la estructura del ADN es la siguiente (Concha Gil Soriano, Capítulo 1 2002):

- Dos cadenas helicoidales de nucleótidos enrolladas a lo largo de un eje imaginario común.

- Las dos cadenas son antiparalelas ya que van en sentidos opuestos (una cadena se inicia con un extremo 5' libre y acaba en otro libre 3'); la otra cadena se dispondrá en 3'-> 5'.
- Las bases nitrogenadas se dirigen hacia el interior de la doble hélice, mientras que los azúcares y los grupos fosfato forman el esqueleto externo.
- La estructura se mantiene estable gracias a los enlaces de hidrógeno que se forman entre los pares de las bases nitrogenadas complementarias.

El orden, o la secuencia, en que aparezcan las bases nitrogenadas determina la información disponible para construir y mantener un organismo, similar a la manera en que las letras del alfabeto aparecen en un orden determinado para formar palabras y frases. De hecho, si la secuencia de bases cambia, la información del ADN también cambia.

Otra importante propiedad del ADN es su capacidad de replicación, o de hacer copias de sí mismo. Cada hebra de ADN en la doble hélice puede servir como modelo para duplicar las secuencias de bases nitrogenadas. Esto es crucial cuando las células se dividen porque cada nueva célula necesita tener una copia exacta del ADN presente en la célula anterior, lo que hace que la información se herede.

Además, durante el proceso de replicación pueden producirse errores, lo cual justifica la capacidad de mutación del ADN, necesaria para los cambios evolutivos.

3.2.Cambios en el ADN

Todos los individuos tienen cambios en el ADN durante el transcurso de sus vidas. Estos cambios ocurren de muchas maneras. Algunas veces son debidos a errores durante la replicación del ADN; otros, se producen por daños en la estructura debidos a agentes ambientales, como por ejemplo, radiación o humo del tabaco. Las células tienen mecanismos que capturan y reparan la mayoría de los cambios que ocurren durante la replicación del ADN o por daños ambientales. No obstante, con la edad, dichas reparaciones no funcionan de forma tan efectiva y los cambios se van acumulando.

Algunos de los cambios ocurren en las células del cuerpo pero no afectan a las células productoras de gametos y no se transmiten a los hijos, se trata de las mutaciones somáticas. Por otro lado, otros ocurren en el ADN de las células productoras de gametos. Estas mutaciones se llaman germinales y pueden transmitirse de padres a hijos, teniendo éstos dicho error en el ADN de cada célula de su cuerpo. Este tipo de mutación es la responsable de las enfermedades hereditarias (Toland 2001).

Un gen es esencialmente una frase formada por las bases A, T, G y C que describe cómo hacer una proteína. Cualquiera de los cambios en estas instrucciones puede alterar el significado del gen y cambiar la proteína que produce, o cómo o cuándo una célula crea la proteína.

En líneas generales, existen los siguientes tipos de cambios en el ADN (Toland 2001):

- Cambio que ocurre en una única base de la secuencia del gen. Es equivalente a cambiar una letra de una frase.

- Traslocación: en un cambio de este tipo, una o más bases se insertan o eliminan, equivalente a añadir o eliminar letras en una frase. Como las células del ser humano leen el ADN en palabras de 3 letras, llamadas aminoácidos, con una traslocación se cambia una palabra entera. Este tipo de cambio puede hacer que el segmento de ADN carezca de significado y a menudo resulte una proteína reducida.

- Deleción: cambio que resulta de "pérdidas" de alguna parte de ADN. Pueden ser pequeños, tales como cambios de una letra, o más extensos, afectando a un gran número de genes del cromosoma. En ocasiones también causan traslocaciones.

- Inserción: se trata de un cambio que resulta de la adición de un segmento de ADN. Pueden causar traslocaciones y generalmente resultan en una proteína no funcional.

- Inversión: en un cambio por inversión, una sección completa de ADN se invierte. Una pequeña inversión puede implicar a un grupo reducido de bases dentro de un gen, mientras que las grandes implican grandes regiones de un cromosoma conteniendo varios genes.

3.2.1. Los polimorfismos

Las variaciones en la secuencia del ADN se describen en ocasiones como mutaciones y otras veces como polimorfismos. Es necesario especificar la diferencia entre esos términos y cómo se aplican en el genoma humano.

Una mutación se define como cualquier cambio en la secuencia del ADN fuera de lo normal. Ello implica que, dada una población, un individuo pasa a ser una variante anormal si tiene un cambio en un alelo (o más) que era prevalente en la población (Twyman, Mutation or polymorphism? 2003).

Por el contrario, un polimorfismo viene de la combinación de palabras del griego *poli* (muchos) y *morfe* (forma), y se refiere a las múltiples formas de un gen que pueden existir en los individuos. Se trata, por tanto, de un cambio en el ADN que no afecta a la estructura o secuencia, pero sí influye lo suficiente como para producir variaciones entre los individuos (Phillips s.f.).

Existe una frontera arbitraria entre mutación y polimorfismo en la que, para que una modificación del ADN sea considerada como polimorfismo, la variación en la secuencia de nuecleótidos debe estar presente en un 1% o más de la población. Si dicha frecuencia es inferior, la variante se considera mutación, ya que la que ha sufrido el cambio ha sido aquella menos común a todos los individuos (Parma 2009). Además, cualquier variación en la secuencia de ADN que causa directamente una enfermedad en los seres humanos reduce la adaptabilidad de los mismos en la población y pasan a ser una variante inusual.

Sin embargo, no todas las mutaciones provocan enfermedades, ya que cualquier nueva variación en la secuencia, incluso con un efecto neutro o beneficioso, se considerará inicialmente como una mutación inusual.

Las variaciones en la secuencia del ADN consideradas como polimorfismos no causan enfermedades evidentes. Muchas se encuentran fuera de los genes y no tienen ningún efecto. Otras, pueden estar situadas dentro, pero únicamente implican cambios como la estatura o color de ojos en lugar de otras de importancia

médica. No obstante, pueden contribuir a la predisposición de padecer una enfermedad o de tener una respuesta determinada ante un medicamento. Por ejemplo, se dan muchos polimorfismos de CYP 1A1, una de las muchas enzimas[1] del hígado. Aunque las enzimas tienen la misma estructura y secuencia, los polimorfismos para esta enzima pueden influir en cómo cada individuo metaboliza los medicamentos (Phillips s.f.).

Cabe decir que los conceptos mutación y polimorfismo no deben aplicarse de forma estricta ya que una variante causante de una enfermedad en una población puede tratarse de un polimorfismo en otra si ofrece alguna ventaja y tiene una frecuencia de aparición superior. Este es el caso de la alteración en un alelo del gen beta-globina. En la población caucásica causa un grave desorden en la sangre, mientras que en algunos lugares de África es una variante común y confiere resistencia a la malaria (Twyman, Mutation or polymorphism? 2003).

3.2.2. Los SNP's

Un polimorfismo de un único nucleótido (en adelante SNP) es un pequeño cambio genético, o variación, que puede ocurrir dentro de la secuencia de ADN de un individuo. Son los polimorfismos más frecuentes, ya que aparecen una vez cada 300 nucleótidos, lo que quiere decir que hay al menos unos 10 millones de SNP's en todo el genoma humano.

Los SNP's se caracterizan por una variación que ocurre cuando una única base, por ejemplo A, se reemplaza por otra de las 3 posibles, formando un nuevo nucleótido, como se ve en la imagen siguiente:

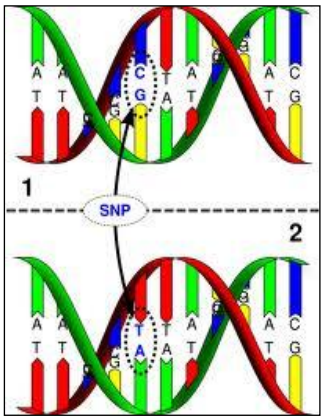


Figura 3: http://es.wikipedia.org/wiki/Polimorfismo_de_nucle%C3%B3tido_simple

De media, los SNP's suceden en la población más de un 1 por ciento de las veces. La mayoría, se encuentran fuera de las secuencias codificantes (en regiones no codificantes del gen o en regiones intergénicas), mientras que de un 3 a 5 por ciento se trata de secuencias de ADN empleadas en la producción de proteínas (SNPs: Variations on a theme 2007).

1 Proteínas que elaboran las células a partir de la información contenida en el ADN de su núcleo. Son las responsables del correcto funcionamiento de la célula.

Los SNP's que se encuentran dentro de una secuencia codificante son de especial interés para los investigadores ya que es más probable que modifiquen la función biológica de una proteína.

Muchas enfermedades comunes en los humanos no son causadas por una variación genética dentro de un único gen, sino que son producidas por complejas interacciones entre múltiples genes, así como por factores ambientales o de forma de vida. Como ambos factores añaden gran incertidumbre en el desarrollo de la enfermedad, resulta difícil medir y evaluar su efecto completo en el proceso de la misma. Sin embargo, la tendencia actual de las investigaciones es la búsqueda de interacciones entre genes relacionadas con las enfermedades (Raquel Iniesta 2005) (Brett A. McKinney 2006).

Los factores genéticos también confieren susceptibilidad o resistencia a una enfermedad y pueden determinar la severidad o progreso de la misma. Además, como también afectan a la respuesta ante medicamentos, los polimorfismos tales como SNP's son útiles para ayudar a los investigadores a determinar y entender por qué los individuos difieren en la capacidad para absorber ciertos medicamentos, así como para determinar por qué un individuo puede experimentar un efecto adverso a un fármaco específico.

Definiendo y entendiendo el rol de los factores genéticos en una enfermedad, también permitirá a los investigadores a evaluar mejor el rol de los factores no-genéticos, tales como comportamiento, dieta, estilo de vida, actividad física, etc.

3.3.Genotipos

Cada posible variante de la secuencia de ADN en una posición del genoma (locus), se le llama alelo. Al tratarse los seres humanos de organismos diploides, éstos poseen dos alelos para una misma posición del cromosoma, uno procedente del padre y otro de la madre (Genotype s.f.). Por lo tanto, podemos referirnos al concepto de genotipo como la combinación de dichos alelos. De hecho, se puede decir que toda la información contenida en los cromosomas se organiza en genotipos, es decir, en pares de alelos.

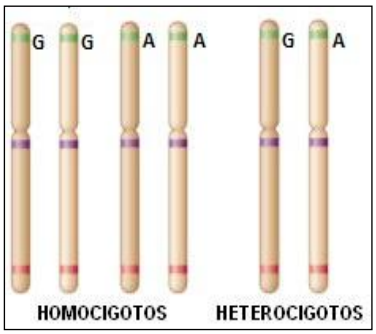


Figura 4: http://www.librosvivos.net/smtc/img/1185_alelos.jpg

Los pares de alelos pueden ser idénticos (homocigotos) o diferentes (heterocigotos), tal y como se muestra en la imagen anterior. Por ejemplo, para los alelos A y G, situados en un determinado locus, los 3 posibles genotipos son: AA, AG/GA, GG. Un individuo con los genotipos AA o GG, es homocigoto y un individuo con el genotipo AG se denomina heterocigoto.

Por otro lado, los organismos cuyos genes difieren en un alelo se dice que tienen genotipos diferentes (Genotype-phenotype distinction s.f.); por tanto, como un SNP consiste en diferentes variantes de alelos en una población para una misma posición del cromosoma, un SNP implica el cambio de un genotipo por otro.

Cabe decir que no se debe confundir genotipo con fenotipo. Este último se refiere a las propiedades observables de un organismo, tales como morfología, desarrollo o comportamiento. Mientras que el genotipo representa la estructura genética, el conjunto de instrucciones para construir y mantener vivo un individuo (Genotype-phenotype distinction s.f.).

La manifestación externa del fenotipo de un organismo es lo que se conoce como expresión génica. Depende del contenido de cada gen o, de manera equivalente, de los alelos que están contenidos en cada gen.

De hecho, algunos alelos reflejan sus características en el fenotipo de manera más dominante que el resto; en otros casos, la expresión génica depende de si el genotipo aparece en un estado homocigoto o heterocigoto. Otros rasgos fenotípicos, en cambio, son una combinación de varios alelos procedentes de diferentes genes.

Anteriormente, determinar la combinación de alelos solía cumplirse únicamente a través de los análisis de genealogías. Sin embargo, este método dejaba muchas preguntas sin responder, en especial aquellas relacionadas con las características resultantes de una interacción entre diferentes genes (What is a genome? s.f.). Actualmente, las técnicas genéticas existentes ayudan a los investigadores a rastrear la herencia de los rasgos en los individuos, identificando la ubicación de genes, variantes alélicas, y determinando aquellas características causadas por múltiples genes. Básicamente se pueden distinguir 4 modelos de herencia (Raquel Iniesta 2005) (What is a genome? s.f.):

- Dominante: Supone que un alelo se expresa siempre, aunque haya una única copia en el genotipo. En este caso, el fenotipo aparece igualmente, tanto si el alelo se encuentra en un genotipo heterocigoto u homocigoto. En algunas ocasiones, se producen efectos que con el alelo restante no ocurrirían, como sucede por ejemplo con la enfermedad de Huntington. Por tanto, en un SNP, si el alelo causante del riesgo de una enfermedad es A, y los genotipos posibles son AA, AT y TT, serán igual de influyentes AA y AT.

- Codominante: Supone que ambos alelos contribuyen de la misma manera en el fenotipo, ninguno es dominante sobre otro, de manera que el genotipo heterocigoto muestra el fenotipo de cada alelo. Esto ocurre en la herencia de los grupos sanguíneos ABO o en la raza de ganado Shortron, en la que si se cruza un toro con pelaje rojo y una vaca con pelaje blanco, la descendencia es ruana (pelo rojo y blanco mezclados, no rosado) (Herencia intermedia y codominancia s.f.).

- Recesivo: Supone que un alelo se expresa sólo si hay dos copias del mismo en el genotipo; de esta manera, el fenotipo de un alelo recesivo aparece cuando ambos alelos son idénticos. Cuando un individuo tiene un alelo dominante y otro recesivo, se expresa el rasgo del alelo dominante. Así, para un SNP, si el alelo recesivo causante de tener tendencia a padecer una enfermedad es A y los genotipos posibles son AA, AT y TT, sólo será influyente el genotipo AA.

- Aditivo: Supone que los alelos contribuyen en el fenotipo en una cantidad ponderada. Es decir, para un SNP, dados los posibles genotipos AA, AT y TT, si el alelo A es causante del riesgo de padecer una

enfermedad, el genotipo TT no tendrá ninguna influencia, AT será un genotipo de riesgo y AA tendrá más riesgo que el anterior.

3.4.Equilibrio de Hardy-Weinberg

Uno de los aspectos que se deben tener en cuenta para realizar una descripción estadística de los polimorfismos es la frecuencia genotípica; es decir, la proporción de individuos que presentan un genotipo u otro.

Así, si en una población de 300 individuos, y con los posibles genotipos AA, TT y AT, aparecen 150 individuos con el genotipo AA, 45 con el genotipo TT y 105 con el genotipo AT, las frecuencias genotípicas son las siguientes:

- 0,50 para el genotipo AA
- 0,15 para el genotipo TT
- 0,35 para el genotipo AT

Respecto a las frecuencias de los alelos, se debe tener en cuenta que cada individuo tiene 2 y, por lo tanto, las proporciones se deben multiplicar por 2 en los casos en que los genotipos sean homocigotos. De esta manera, para la población y genotipos anteriores, las frecuencias alélicas son:

$$frec(A) = \frac{(2 * 0.50) + 0.35}{2} = 0.675$$

$$frec(T) = \frac{(2 * 0.15) + 0.35}{2} = 0.325$$

El equilibrio de Hardy-Weinberg sostiene que las frecuencias genotípicas y alélicas permanecen constantes en una población, generación tras generación, a menos que se introduzcan efectos que alteren dicha población (International HapMap Project s.f.) (Hardy-Weinberg principle s.f.). De hecho, los cambios son los indicadores de la evolución de una especie.

En concreto, el equilibrio de Hardy-Weinberg supone las siguientes características para una población (Hardy-Weinberg principle s.f.) (Kalmes R 2001):

- Panmixia o apareamiento aleatorio. El equilibrio de Hardy-Weinberg afirma que la población tiene las frecuencias genotípicas dadas después de una única generación de apareamiento aleatorio dentro de la población. Cuando ocurren violaciones a esta suposición, la población no está en equilibrio. Tales violaciones son:
    - Endogamia, que causa un incremento de la homocigocidad para todos los genes.
    - Emparejamiento selectivo, que causa un incremento de la homocigocidad sólo en algunos genes. Por ejemplo, una de las teorías que explican el autismo se debe al apareamiento

selectivo entre un hombre y una mujer con los cerebros extremadamente masculinos, carentes de empatía (Assortative mating s.f.) (Autism spectrum disorder s.f.).

- Poblaciones de tamaño pequeño, en las que puede ocurrir deriva génica, es decir, cambios aleatorios en las frecuencias de los alelos que modifican las características de las especies.

Por otro lado, si una población viola uno de los siguientes 3 supuestos, la población puede continuar en cada generación con proporciones de equilibrio de Hardy-Weinberg, pero con variaciones en las frecuencias de los alelos:

- Selección, en general hace que las frecuencias de los alelos cambien, en ocasiones rápidamente. Mientras la selección direccional ocasionalmente conduce a la pérdida de todos los alelos excepto de uno, que es el que se ve favorecido, otras, como la selección balanceada, conduce al equilibrio sin pérdida de alelos.

- Mutación, tiene un efecto muy sutil en las frecuencias alélicas. Los índices de una mutación son del orden de 10^{-4} a 10^{-8}, y el cambio en dichas frecuencias es, como mucho, del mismo orden. La mutación recurrente mantiene los alelos en la población, incluso si hay una fuerte selección "en contra" de ellos.

- Migración, genéticamente junta 2 o más poblaciones. En general, las frecuencias alélicas se convierten más homogéneas entre las poblaciones. Algunos modelos de migración incluyen apareamiento no aleatorio, Para estos modelos, las proporciones de Hardy-Weinberg no son válidas.

Una descripción equivalente para el equilibrio de Hardy-Weinberg es que, para un individuo dado, los alelos de la siguiente generación se eligen de manera aleatoria e independiente unos de otros. En el caso más básico de un único locus con 2 alelos, el alelo dominante se expresa como *A* y el recesivo como *a*, y sus frecuencias se denotan por *p* y *q*, de manera que frec(A) = p y frec(a) = q.

Si la población se encuentra en equilibrio, los genotipos que se forman siguen el siguiente cuadro:

		Mujer	
		A(p)	a(q)
Hombre	A(p)	AA(p^2)	Aa(p ∗ q)
	a(q)	Aa(p ∗ q)	aa(q^2)

De manera que $frec(AA) = p^2$ y $frec(aa) = q^2$ para el caso de genotipos homocigotos y $frec(Aa) = 2 * p * q$ para los heterocigotos. Estas frecuencias son las proporciones del equilibrio de Hardy-Weinberg, que se consiguen en una generación y sólo es necesario que la población sea infinita y panmíctica[2].

En ocasiones, una población se crea uniendo hombres y mujeres con diferentes frecuencias alélicas. En esta situación, el supuesto de única población no se cumple hasta después de la primera generación; por tanto, la primera generación no estará en equilibrio pero sí las sucesivas.

[2] Población en la que se da apareamiento aleatorio.

Cabe decir que, las condiciones de equilibrio de Hardy-Weinberg sólo se pueden dar dentro de un laboratorio, ya cualquiera de las modificaciones señaladas aparecen en la naturaleza. Ello implica que las proporciones ideales sirven como base para estudiar los cambios que se dan en una población y, si dicha población no está en equilibrio, se debe a que durante la transmisión de padres a hijos se ha producido algún tipo de modificación en el ADN que debe estudiarse con detalle.

Siguiendo el ejemplo anterior, dado que tenemos las frecuencias alélicas (frec(A) = 0,675 y frec(T) = 0,325), se pueden calcular las frecuencias genotípicas esperadas para una población ideal y comprobar si realmente está en equilibrio de Hardy-Weinberg o no. Así:

$frec(AA) = 0.675^2 = 0.455$

$frec(AT) = 2 * 0.675 * 0.325 = 0.44$

$frec(TT) = 0.325^2 = 0.105$

Como los resultados son bastante aproximados a las frecuencias genotípicas originales, la población está en equilibrio.

Generalmente, como se comenta más adelante, en los estudios de asociación de polimorfismos con una enfermedad, la comparación de las frecuencias esperadas respecto a las reales se realiza mediante tests estadísticos convencionales, como por ejemplo mediante el test de la χ^2 , y empleando el nivel de significación p-value.

En dichos análisis, se dan dos tipos de poblaciones, una, con muestras de pacientes sanos (control) y otra con pacientes enfermos (caso).

Si, para un polimorfismo, se analiza la existencia de equilibrio en la población de control y éste resulta negativo, puede deberse a sesgos en la interpretación de los resultados durante la extracción de genotipos, ya unos puede que sean más fáciles de detectar que otros. También puede deberse a que dentro de la población se hayan tomado muestras de individuos con relación de parentesco. Por otro lado, si se emplea un nivel de significación estándar del 5%, puede aparecer una falta de ajuste estadístico a pesar de que la población sí esté en equilibrio y deba realizarse un ajuste de dicho valor (International HapMap Project s.f.).

Si la población que se está analizando es la de casos, es necesario revisar también el valor que se considera como nivel de significación. No obstante, si sigue apareciendo desequilibrio en la población para un genotipo dado, posiblemente sea porque está asociado con la enfermedad o factor de estudio.

3.5.Haplotipos

Como se ha comentado anteriormente, a excepción de las células sexuales, los cromosomas en las células humanas se presentan en pares. Una copia se hereda del padre y el otro se hereda de la madre. No obstante, los cromosomas no pasan de generación en generación como copias idénticas, pasan por un proceso conocido como recombinación. Las copias de cada par de cromosomas se unen e intercambian algunos fragmentos. El resultado es un cromosoma híbrido, que contiene partes de ambas copias del par de

cromosomas, y que es transmitido a la siguiente generación (International HapMap Project s.f.) (Twyman, Haplotype mapping 2003).

Durante el transcurso de múltiples generaciones, algunos de los segmentos de los cromosomas de los antepasados permanecen como regiones de ADN compartidas por múltiples individuos. Estos segmentos, que no han sido separados por la recombinación, son los llamados haplotipos.

Un haplotipo (del griego haploûs, "simple"), en genética, es una combinación de alelos de diferentes lugares del cromosoma, que se transmiten juntos. Pueden ser un locus, varios loci, o un cromosoma entero, dependiendo del número de recombinaciones que hayan ocurrido (Haplotype s.f.).

Por otro lado, un haplotipo también es considerado como un conjunto de SNP's estadísticamente asociados pertenecientes a un único cromosoma. Dicha asociación estadística es producida por su transmisión conjunta y se conoce como desequilibrio de enlace. Así, mediante el estudio de dichas asociaciones y la identificación de unos pocos alelos de un bloque haplotípico, se pueden identificar unívocamente el resto de polimorfismos en esa región (Raquel Iniesta 2005) (Haplotype s.f.).

Además, cuando aparece una enfermedad genética, ésta y los SNP's más cercanos tienden a heredarse como un grupo. Así, identificando una zona de desequilibrio de enlace, puede encontrarse la región en la que se ha dado la mutación responsable de una enfermedad.

A continuación se muestra un árbol con 2 cromosomas iniciales (nodo raíz) que se han ido mezclando generación tras generación a través de recombinaciones, cuyo resultado son los nodos hojas.

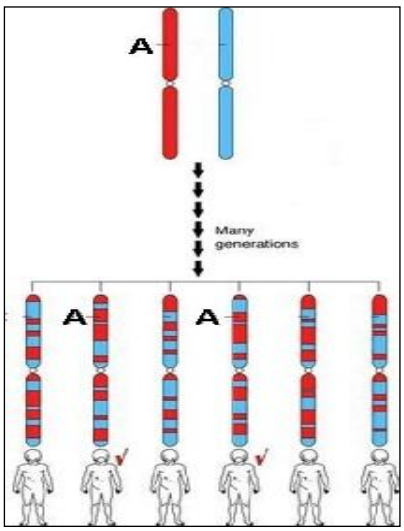


Figura 5: http://hapmap.ncbi.nlm.nih.gov/originhaplotype.html.en

Si la variante genética marcada por A en el cromosoma inicial incrementa el riesgo de padecer una enfermedad, los individuos que hayan heredado dicho fragmento tendrán también ese riesgo. Además, junto con la variante A se encuentran diversos SNP's que pueden utilizarse para identificar la ubicación de dicha variante; es por ello que los SNP's pueden considerarse también como marcadores genéticos.

Obviamente, el análisis de haplotipos permite investigar la genética que hay detrás de las enfermedades comunes, estudiada especialmente por el Proyecto Internacional HapMap, cuyo objetivo es el de desarrollar un mapa haplotípico del genoma humano (HapMap) que describa las variaciones comunes en la secuencia del ADN humano. Dicha cartografía comprende las regiones cromosómicas, incluyendo los SNP que están fuertemente asociados, los haplotipos de dichas regiones y los SNP etiqueta[3]; así como los SNP con asociaciones más débiles (À propos du projet international HapMap s.f.).

El proyecto HapMap es fruto de la colaboración de científicos procedentes de Japón, China, Canadá, EEUU, Nigeria y Reino Unido y se prevé que HapMap se convierta en una fuente clave para los científicos que buscan genes asociados con la salud, las enfermedades y las respuestas a los medicamentos y al entorno (International HapMap Project s.f.) (À propos du projet international HapMap s.f.).

Actualmente, el problema clave es inferir los haplotipos y sus frecuencias correctamente a partir de los datos de los genotipos, ya que, debido a limitaciones tecnológicas, recopilar la información de los haplotipos es generalmente más difícil que recoger la de los genotipos y éstos no siempre pueden determinar de forma unívoca el haplotipo de un individuo.

Por ejemplo, se considera un organismo diploide y 2 loci bialélicos en el mismo cromosoma, como por ejemplo dos SNP's. El primer locus tiene los alelos A y C con 3 posibles genotipos CC, CA y AA, el segundo tiene los alelos G y T, con los posibles genotipos TT, TG y TT. Por tanto, para un individuo, hay 9 posibles combinaciones para los genotipos, como se muestra en el cuadro siguiente. Para el caso de genotipos homocigotos la correspondencia con los haplotipos es unívoca, pero no ocurre lo mismo para el caso de genotipos heterocigotos.

	CC	**CA**	**AA**
TT	CT CT	CT AT	AT AT
TG	CT CG	CT AG / CG AT	AT AG
GG	CG CG	CG AG	AG AG

Hay una gran variedad de métodos que dan resultados bastante aproximados. Unos, se basan en aproximaciones combinatorias, mientras que otros utilizan funciones de probabilidad en combinación con algoritmos de optimización como el de Expectation-Maximization (EM), optimización de cadenas de Markov vía Monte Carlo (MCMC) o modelos ocultos de Markov (HMM) (Haplotype s.f.).

No obstante, la inferencia de haplotipos está en continua mejora ya que no hay un único método que se considere completamente adecuado en todos los entornos.

3 SNP en una región del genoma con alto enlace de desequilibrio. Son útiles en estudios de asociación de SNP ya que, así, es posible identificar una variación genética sin genotipar todos los SNP de una región

3.5.1. Problemas médicos asociados a haplotipos

Actualmente, muchos científicos centran sus estudios de asociación con la enfermedad en el análisis de haplotipos. A continuación se detallan algunos de los estudios recientes:

La enfermedad celíaca:

La enfermedad celíaca se trata de una enfermedad que afecta a la mucosa del intestino, haciendo que los individuos que la padecen presenten intolerancia al gluten. Es considerada como un modelo de enfermedad inmunológica y desde finales de los años 80 se han estado realizando estudios sobre la predisposición genética de los pacientes. Se ha visto que los modelos de las poblaciones sanas son comunes entre sí, pero no ocurre lo mismo con las enfermas.

En concreto, en la población caucásica, el haplotipo HLA-DQ2, codificado con los alelos QA1*0501 y DQB1*0201, está presente en el 90% de los individuos celíacos, mientras que en los individuos sanos se encuentra en un 20%-30% de los casos. Por otro lado, el halotipo HLA-DQ8, codificado por los alelos DQA0103 Y DQB010302, se encuentra en el resto de pacientes que padecen dicha enfermedad (V. Cadahía 2005) (L.N. Karla Melissa Ruiz-Dyck 2010) (F. Fernández-Bañares 2004) (Novo s.f.).

Con esto se puede concluir que, la ausencia del haplotipo HLA-DQ2 y HLA-DQ8 implica, con una elevada probabilidad, que el individuo no va a presentar la enfermedad celíaca.

Por otro lado, los últimos estudios apuntan a que aquellos pacientes de tipo DQ2 negativo presentan el haplotipo HLA-DQ8 (DQA1*0301 y DQB1*0302). No obstante, los análisis de la población que padece la enfermedad celíaca no son del todo concluyentes y los estudios de asociación están en continua evolución.

La obesidad y diabetes mellitus tipo 2:

La obesidad se trata de una enfermedad caracterizada por la acumulación de grasa hasta límites que implican daños severos en la salud, como problemas cardiovasculares, problemas gastrointestinales, determinados tipos de cáncer, entre otros. De hecho, está considerada por la OMS como una enfermedad con características epidémicas a nivel mundial, ya que cada año mueren alrededor de 2,5 millones de personas debido a alguna de las consecuencias que ésta ocasiona (Obesidad s.f.).

Una de estas consecuencias es la diabetes mellitus de tipo 2. Se trata de una enfermedad que altera el metabolismo de manera que los individuos pueden presentar o bien resistencia a la insulina o bien un exceso en la producción de la misma, alteración para la asimilación de grasas o proteínas, entre otras. Además, los individuos que la padecen sufren diversas repercusiones agudas y/o crónicas (Diabetes mellitus tipo 2 s.f.).

Por ello, ambas enfermedades han sido y siguen siendo objeto de estudio. No obstante, científicos franceses du Centre National de la Recherche Scientifique (CNRS) han realizado un avance importante. En concreto, han descubierto un haplotipo del gen ENPP1 que está fuertemente asociado con la obesidad generada durante la infancia y la obesidad mórbida adulta. Adicionalmente, han encontrado que la variación dicho gen produce resistencia a la insulina, lo que asocia ambas enfermedades genéticamente (El gen de los obesos 2005) (Haplotipo del gen ENPP1 (PC-1) asociado con el riesgo de obesidad y diabetes de tipo 2, y sus aplicaciones 2007).

De hecho, la modificación de dicho gen hace que los receptores de insulina no puedan producirla como es debido. De esta manera, el hígado libera glucosa hasta producir un exceso, haciendo que los tejidos absorban la glucosa y se convierta en grasa; implicando, además, el riesgo de padecer diabetes mellitus de tipo 2. A continuación se indica un fragmento de la descripción de dicho artículo:

<< […] *Se ha demostrado que existe una fuerte asociación entre la variante de este gen ENPP1, particularmente un haplotipo de riesgo de tres alelos (K121Q; IVS20 delT-11 y un SNP localizado en la secuencia del dominio 3'UTR (principalmente A>G +1044 TGA, QdeITG), y la obesidad de la niñez y adultos con obesidad mórbida.* […]>>.

<< […] *Además, se encontró que la expresión de una isoforma larga de ARNm de ENPP1, que comprende el SNP A>G +1044 TGA, asociado con la obesidad, era específica de tres tejidos de principal importancia para la homeostasis de la glucosa: las células D pancreáticas de los islotes, los adipositos y el hígado. Estos hallazgos demuestran por primera vez un papel principal para varias variantes de ENPP1 en la mediación de la resistencia a insulina, tanto en el desarrollo de la obesidad como de diabetes de tipo 2 (T2D), proporcionando un mecanismo molecular común de ambas aflicciones extendidas ampliamente.* […] >>.

El estudio realizado incluía 3 generaciones en las que los niños manifestaban obesidad y sus padres y abuelos tenían una tendencia a la gordura y a la diabetes. El riesgo de padecer obesidad y diabetes estaba asociada a al riesgo de tener la variante del gen ENPP1 pero, además, sirvió para demostrar la importancia de los factores ambientales puesto que, mientras que los padres y abuelos presentaban una tendencia a la gordura, los nietos, debido a los hábitos alimenticios, vida sedentaria, etc. presentaban obesidad.

4. MATERIALES Y MÉTODOS

A lo largo de este apartado se describen los métodos empleados para llevar a cabo los análisis de los datos. Para ello, las principales herramientas utilizadas son R y Clemenine 9.0.

R es un entorno de desarrollo especialmente diseñado para manipulación de datos, obtención de estadísticas y gráficos. Además, se pueden añadir librerías para ampliar su funcionalidad y poder trabajar con datos de la rama de la medicina y biotecnología, tales como microarrays o SNP's. Al ser un software de libre distribución, programadores e investigadores también pueden crear paquetes que resultarán útiles para el resto de usuarios.

Así pues, dado que este proyecto trata con variables de tipo genético y es necesaria una buena evaluación de dichos datos, este software resulta de mucha utilidad.

Por otro lado, Clementine es una aplicación integrada de minería de datos. Utiliza técnicas predictivas para obtener patrones que nos puedan ayudar a mejorar procesos actuales y tomar decisiones: herramientas para aplicar reglas de asociación, técnicas de clasificación y agrupación, manipulación de datos, etc. Además, permite integrar datos de diferentes fuentes, generar informes y exportar resultados (SPSS 2000).

Por todo ello, se ha pensado que también es una herramienta adecuada para la elaboración de este proyecto.

4.1.Materiales

Los datos de entrada del proyecto se proporcionaron en una tabla en formato Excel, que permite ser importada tanto a R como a Clementine y éstos, a su vez, permiten crear tablas en dicho formato que tal vez sea de interés para mostrar algunos resultados.

Como se ha dicho anteriormente, a lo largo de este proyecto se va a realizar un análisis del rechazo crónico en los trasplantes de riñón, así como una predicción del mismo.

Esta predicción se enfocará desde dos puntos de vista; por un lado, se hará un modelo para predecir la existencia o no de rechazo y, por otro, un modelo que prediga, además, el tipo de rechazo.

Para la elaboración de ambos modelos se cuenta con muestras de 276 pacientes y, de cada uno, disponemos de 53 variables divididas en clínicas y genéticas. Así, la disposición de los datos es en forma matricial, donde cada paciente corresponde con una fila y cada variable del mismo es una columna. Para visualizar el conjunto de datos inicial, se puede consultar la página 1 del documento Excel “Punto 5.xlsx” de la carpeta “anexos/Resultados” que se adjunta.

Como variables clínicas, se tienen de tipo cuantitativo:

- T_isqm, el tiempo que tarda un órgano en tener el aporte de riego sanguíneo óptimo.
- Edad, edad del receptor.

- Edad_donante, edad del donante.

También se tienen variables clínicas de tipo cualitativo que, para poder trabajar con ellas, cuentan con la siguiente codificación:

- Enf_primaria, se trata de la enfermedad que padeció el receptor antes de llegar a la necesidad de un trasplante renal. Dependiendo de dicha enfermedad, la variable admitirá los valores:
    - 1 – glomerulonefritis, es un tipo de enfermedad renal en la que los glomérulos se inflaman, produciendo pérdida de sangre y proteínas en la orina. Las causas son muy diversas, desde diabetes hasta exposición a hidrocarburos (Glomerulonefritis 2009).
    - 2 – hipertensión arterial, cuando el corazón late bombea sangre a las arterias, produciendo presión sobre ellas. Esta presión es la que consigue que circule sangre por todo el cuerpo. Si sobrepasa los límites normales, se produce hipertensión arterial (Hipertension s.f.).
    - 3 – nefropatía diabética, es una enfermedad renal que se produce por una complicación de la diabetes, ya que la acumulación de demasiado azúcar puede dañar los glomérulos y, por tanto, producir filtraciones en la orina, como la albúmina (proteína) (Nefropatía diabética s.f.).
    - 4 – nefritis túbulo intersticial, enfermedad que consiste en la inflamación de los túbulos e intersticios, sin afectar a los glomérulos. Puede producirse por muchas causas, entre ellas, por efectos secundarios de ciertos medicamentos (Nefritis intersticial s.f.).
    - 5 – uropatía obstructiva, afección en la que la orina no drena por el uréter, haciendo que ésta se acumule en el riñón y produzca la lesión de uno o ambos riñones. Sus consecuencias dependen de la duración y gravedad y de si la obstrucción es uni o bilateral. Si se diagnostica y corrige rápidamente el daño renal es mínimo o reversible, independientemente de que afecte a uno o dos riñones; si no, puede derivar en fallo renal permanente (Uropatía obstructiva s.f.).
    - 7 – causas vasculares, todas aquellas enfermedades relativas a los vasos sanguíneos, como la arteriosclerosis o trombosis. La insuficiencia renal crónica está ligada a patologías cardiovasculares ya que éstas pueden ocasionarla, y viceversa. Además, los pacientes que presentan problemas vasculares tienen una mortalidad de un 10% a un 30% superior que el resto de pacientes. (P. Sierra s.f.).
    - 8 – poliquistosis renal, enfermedad hereditaria en la cual se forman racimos de quistes en los riñones. La acción que desencadena la formación de dichos quistes se desconoce, pero puede ser de forma autosómica dominante o recesiva. En forma dominante, si uno de los padres presenta esta afección, los hijos tienen un 50% de padecerla; se presenta tanto en niños como en adultos, siendo ésta última más probable ya que sus síntomas no aparecen hasta una mediana edad. Por otro lado, en forma recesiva, aparece durante la lactancia o niñez, siendo más grave que la anterior y de rápido progreso ya que suele producir daño renal permanente y la muerte a temprana edad (MedlinePlus s.f.).
    - 6 – desconocida/otras.

- Sexo, indica el sexo del receptor y podrá tener los valores:
    - 1 – masculino.
    - 2 – femenino.
- Sexo_donante, se trata del sexo del donante. Los valores podrán ser:
    - 0 – masculino.
    - 1 – femenino.
- Causa_muerte_donante, como indica el nombre, se refiere a la causa de la muerte del donante. Se codificará como:
    - 5 – traumatismo cráneo encefálico (TCE), lesión cerebral ocasionada por un golpe fuerte en la cabeza, bien por un accidente de tráfico, una caída, un accidente laboral, etc. Produce una alteración de la conciencia o una pérdida de la misma y puede provocar pérdida de las habilidades y funciones físicas. Hay muchos niveles, desde muy leve hasta el coma y muerte (Traumatismo craneal 2011). En este caso, los órganos dañados son los propios del accidente, de las maniobras para salvar la vida a la persona o debido a la demora en el diagnóstico de su muerte.
    - 2 – accidente cardiovascular (AVC), lesión producida por la falta de oxígeno en alguna de las zonas del sistema nervioso central, debido a que un vaso sanguíneo en dicha zona se ha bloqueado o roto. Los signos y síntomas de esta lesión dependen de la causa que los haya provocado, de la cantidad de tejido afectada y, en especial, de la región en la que se haya producido. Puede darse en el hemisferio derecho, en el izquierdo, en el cerebelo o en el tallo cerebral, siendo esta última altamente mortal ya que controla funciones como la respiración o el latido del corazón (Qué es una embolia o ictus s.f.).
    - 0 – otros.
- UO, indica si el receptor presentó uropatía obstructiva (UO = 1) o no (UO = 0).
- Matches-A, Matches-B, Matches-DR, indican el nivel de compatibilidad donante-receptor. Los HLA (Antígenos Leucocitarios Humanos), son moléculas que se encuentran en los glóbulos blancos de la sangre y en la superficie de las células de casi todos los tejidos. Se encargan del reconocimiento inmunológico y garantizan la respuesta inmune. Este conjunto de moléculas y la forma en la que se transmiten de padres a hijos constituyen el complejo de histocompatibilidad: el sistema HLA.

  En dicho sistema hay unas zonas especiales (A, B, C, DR, DQ, etc.) y, dependiendo del tipo de molécula (antígeno) que haya en dichas zonas, habrá un nivel de compatibilidad u otro. Así, debe presentarse una combinación molecular idéntica o con determinadas coincidencias para ser compatible y, por tanto, que los tejidos se "acepten". Además, dependiendo del tipo de trasplante se requerirán diferentes grados de coincidencia (Banco Nacional de Órganos y Tejidos 2004) (National Center for Research Resources 2009).

Por tanto, Matches-A se refiere a la compatibilidad en el HLA-A, Matches-B a la compatibilidad en el HLA-B y Matches-DR a la compatibilidad en el HLA-DR.

Dependiendo del grado de compatibilidad, las variables podrán tener los valores siguientes:

- 0 – compatibilidad alta.
- 1 – compatibilidad media.
- 2 – compatibilidad nula.

Respecto a las variables genéticas, se dispone de 42 nucleótidos de polimorfismo simple para cada paciente en los que aparecen los 2 alelos del genotipo sin separación.

Finalmente, como variables respuesta se tiene:

- DCTRsi_no, para indicar si hay rechazo crónico (DCTRsi_no = 1) o no (DCTRsi_no = 0).
- DCTR_otrDCTR, para determinar el tipo de rechazo crónico:
  - Si no hay rechazo crónico, DCTR_otrDCTR = 0.
  - Si el rechazo crónico es de tipo nefropatía crónica del injerto (NCI), DCTR_otrDCTR = 1.
  - Si se trata de otro tipo, DCTR_otrDCTR = 2.

4.2.Métodos de análisis de calidad

Para empezar a conocer los datos con los que se va a trabajar, se ha considerado oportuno realizar, en primer lugar, un estudio de la calidad de los mismos. Dicho estudio se ha realizado empleando aplicaciones de la herramienta Clementine.

Por un lado, es importante conocer si la prevalencia de los datos es equilibrada o, de lo contrario, hay muchos casos de una clase y muy pocos de otra. Esto puede provocar que el modelo predictivo tome como ruido la clase minoritaria y tenga en cuenta sólo la mayoritaria. De esta manera, el modelo tiene un buen porcentaje de aciertos, pero simplemente debido a la proporción de datos, no porque realmente se estén evaluando todas las clases. Ante esta situación, se debe modificar la muestra con la que se van a entrenar los datos, dejando una proporción similar de casos y de control (Orallo, Tema 2: El proceso KDD. Técnicas de minería de datos. 2010). Para ello se puede optar por realizar uno de los siguientes cambios:

- Submuestreo, reduciendo la clase que contiene más valores.
- Sobremuestreo, aumentando, mediante repetición, la clase que contiene menos valores.

También es posible realizar la modificación en la etapa de test. Esto es, en lugar de emplear la proporción de aciertos/errores como técnica de evaluación, utilizar la Macromedia:

$$macromedia(h) = \frac{\frac{aciertos_{clase1}}{total_{clase1}} + \frac{aciertos_{clase2}}{total_{clase2}} + \ldots + \frac{aciertos_{clase\,m}}{total_{clase\,m}}}{m}$$

Así, para determinar si va a ser necesario realizar algunos de los cambios anteriormente citados, se crea una tabla para cada variable dependiente en la que se indica la proporción de cada clase. Para ello, se ha utilizado la herramienta Clementine.

Por otro lado, debido a fallos en la transcripción o en la realización de las extracciones de SNP's, pueden encontrarse datos faltantes, en especial en algunas casillas correspondientes a las variables genéticas.

Como para la correcta realización de los análisis posteriores no se pueden manejar datos con valores faltantes, se ha considerado oportuno obtener, para cada variable, el número de valores perdidos.

Finalmente, se realiza una tabla resumen para cada variable en la que aparecen los siguientes aspectos:

- Histograma de distribución, lo que permite revisar la forma de la distribución y si hay algún dato anómalo que no siga la distribución del resto.
- Rango de los datos que contiene la variable, cuando el tipo de la misma lo permita.
- Tipo de la variable, ya que en función de éste se pueden realizar unos análisis u otros. Posteriormente se describe con más detalle este punto.

Además, para algunas de las variables, se muestran elementos como la media o la desviación típica (SPSS 2000).

Todo lo comentado en este apartado es de utilidad para poder realizar un tipo de análisis u otro, o para saber si es necesario realizar alguna modificación inicial de la muestra.

4.3.Métodos de visualización

El objetivo en este apartado es poder visualizar algún tipo de asociación entre las variables dependientes y la variable que se desea predecir (DCTRsi_no, DCTR_otrDCTR, según el caso). Para ello se emplea la aplicación malla de la herramienta Clementine (SPSS 2000).

Con ella se pueden visualizar las asociaciones fuertes y poder empezar a tener una idea de la relación que hay entre los datos. En un principio, aparecen todas las variables relacionadas con la dependiente mediante líneas, creando una malla, pero es necesario fijar un límite para obtener aquellas que son más importantes.

Para cada variable, la medida de la asociación se realiza en función de las veces que, para un mismo valor, la variable dependiente coincide. Por ejemplo, dada una muestra de 150 casos, si se tiene un SNP con un genotipo "GG" que aparece en 130 de los casos, se considera que tiene una asociación fuerte con la variable dependiente y así queda reflejado en la posible malla creada.

Además, los resultados también se muestran en una tabla en la que se indica el valor de la variable independiente para el que se considera que la asociación es relevante, así como el de la variable que se desea predecir.

4.4.Imputación

Ante la existencia de valores faltantes, dado que la mayoría de las herramientas estadísticas necesitan trabajar con los datos completos, se debe aplicar alguna metodología que permita completar los valores perdidos y, así, poder utilizar la información en su totalidad.

El conjunto de esos métodos forman la imputación de datos.

Una alternativa sería eliminar todas las muestras que presentan valores nulos o las variables que no tengan toda la información completa. Esta práctica es la más fácil pero, dado que se dispone de un número relativamente pequeño de pacientes y la gran mayoría de las variables presentan datos faltantes, la probabilidad de perder información útil sería bastante alta.

Otra alternativa consiste en estimar los valores faltantes. Así, a través de la información que se tiene se "predecirán" dichos valores, completando todos los datos.

La imputación puede ser simple, utilizando técnicas como la media, mediana o moda sobre los valores conocidos (Ramírez s.f.). Así, por ejemplo, para una variable dada, los datos faltantes se estimarían realizando la media de los valores conocidos y el resultado obtenido sería el valor que se imputaría en todos los vacíos de dicha variable.

También se puede utilizar imputación múltiple, con métodos más complejos como el algoritmo de Monte Carlo.

En cualquier caso, no hay ninguna técnica cien por cien adecuada, ya que se tratan de datos aproximados, no reales, y cuando hay un gran número de espacios vacíos pueden modificar la información original y los resultados asociados.

4.4.1. Imputación mediante el valor más frecuente

Este método es una manera de estimar los valores faltantes de las variables. Para cada una, se obtiene el valor más común y éste es el elegido para sustituir a los espacios en blanco.

Esta metodología se ha elegido por su simplicidad pero, además, porque al ser la mayoría de las variables SNP's y, por tanto, categóricas, no se puede estimar la media o realizar cálculos numéricos convencionales.

No obstante, se sabe que mediante este tipo de imputación se pueden producir sesgos en la información, obteniendo resultados con posibles desviaciones en exceso o en defecto hacia uno u otro valor, en especial si hay un gran número de espacios en blanco.

Por ello, al finalizar el análisis, se realizan comparativas entre los resultados obtenidos con los datos originales y con los obtenidos mediante imputación.

La herramienta utilizada en este apartado es R. Tras realizar las conversiones de tipo oportunas, para cada variable, se diseña el algoritmo qua lleva a cabo la imputación.

4.4.2. Imputación mediante el modelo de la cadena de Markov Monte Carlo

Este segundo tipo de imputación consiste en aplicar el método de la cadena de Markov Monte Carlo mediante la herramienta IMPUTE para rellenar los valores faltantes. Para poder alcanzar una mayor comprensión de la metodología utilizada, en primer lugar, se realiza una pequeña explicación de dicho método y, posteriormente, su aplicación a la herramienta utilizada.

4.4.2.1. El modelo matemático

En algunos procesos, en cada instante de tiempo t se registra un valor diferente para una misma variable. Por ejemplo, si cada día se registra un número diferente de ventas de un producto, la variable "ventas de un producto" varía en cada instante t = día.

Si, dada una población, sobre cada uno de los individuos hay una variable asociada que toma diferentes valores en cada instante de tiempo, se tendrá un proceso estocástico (Rafael Romero Villafranca 2005).

Un proceso estocástico será markoviano si, para conocer el valor en el estado futuro sólo es necesario el valor en el presente, sin importar el valor que haya tomado en instantes anteriores. Es decir, la distribución de probabilidad condicional del estado futuro del proceso, dados los estados presente y pasado, depende únicamente del presente. Por ejemplo, dada la variable X(t) y la distribución condicional $X(t+u)\big/X(t)=a,\ X(t-1)=b\ \ldots$, dicha distribución sólo dependerá del valor más reciente, es decir,

de X(t) = a.

Un modelo oculto de Markov es un modelo de Markov en el que aparecen estados ocultos, desconocidos.

A continuación se muestra un dibujo de un modelo de Markov típico:

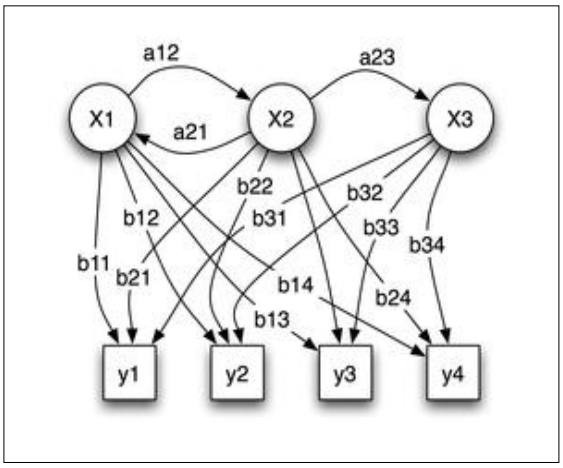


Figura 6: http://en.wikipedia.org/wiki/Hidden_Markov_model

- En un modelo regular de Markov los estados X_i son visibles y los únicos parámetros desconocidos son las probabilidades entre estados a_i.
- En un modelo oculto, el estado X_i no es directamente visible, pero la salida, dependiente del estado, sí lo es. Cada estado tiene una distribución de probabilidad b_i sobre los posibles tokens de salida. Por lo tanto, la secuencia de tokens generada por un HMM da algo de información sobre la secuencia de estados (Hidden Markov Model s.f.).

Cada nodo del diagrama representa el valor que puede tomar una variable aleatoria X(t). Ésta es el estado oculto en el instante t; en el dibujo $X(t) \in (x_1, x_2, x_3)$. La variable aleatoria Y(t) es la observación en el instante t, en el dibujo $Y(t) \in (y_1, y_2, y_3)$. Los arcos en el dibujo serían las dependencias condicionales.

En el diagrama también se aprecia que la distribución de probabilidad condicional de la variable oculta X(t) en el instante t, dados los valores de la variable en todos los instantes, depende sólo del valor de la variable oculta X(t-1). De la misma manera, el valor de la variable observada Y(t) depende únicamente de X(t), ambas en el mismo instante de tiempo t.

Por otro lado, una cadena de Markov se refiere a la secuencia de valores de una variable aleatoria X_i $(X_0, X_1, \ldots, X_n)$ generada por un proceso de Markov, manteniendo las mismas propiedades que un proceso de Markov general.

Normalmente no es difícil construir una cadena de Markov con las propiedades deseadas (un proceso de Markov que se modela como una cadena y tiene un espacio de estados finito). El problema más difícil es determinar cuántos pasos son necesarios para que el modelo converja a una distribución con un error aceptable.

4.4.2.2. Aplicación a IMPUTE

Los métodos de imputación trabajan combinando un panel de individuos de referencia (caracterizados por un conjunto de SNP's genotipados) y una muestra de estudio procedente de una población similar, de la que

se va a evaluar un subconjunto de los SNP's del panel de referencia. Los métodos de imputación predicen los genotipos no observados en la muestra de estudio utilizando un modelo que extrapola las correlaciones alélicas medidas en el panel de referencia.

IMPUTE es una herramienta que estima los genotipos no observados en los estudios de caso-control genómicos (Background on IMPUTE s.f.). En el momento de realizar este trabajo, existían 2 versiones de IMPUTE:

- IMPUTE v1, diseñada para usarse con un panel de referencia de haplotipos conocidos, tales como los proporcionados por el proyecto Internacional HapMap, y una muestra de estudio genotipada como un subconjunto de SNPs del panel de referencia.

  Un posible escenario es el siguiente (Bryan N. Howie 2009):

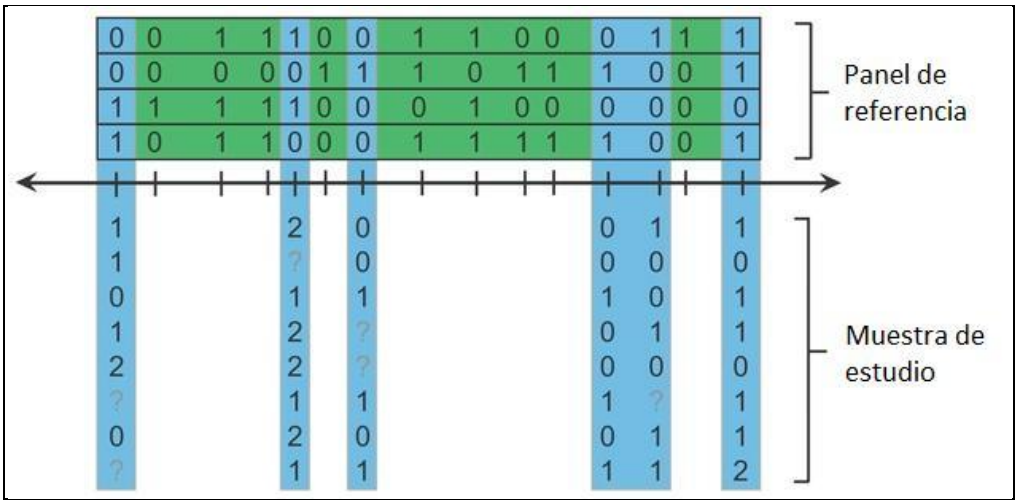


Figura 7:
http://www.plosgenetics.org/article/slideshow.action?uri=info:doi/10.1371/journal.pgen.1000529&imageURI=info:doi/10.1371/journal.pgen.1000529.g001#

En el esquema se ve el panel de referencia y el que se va a inferir. En el panel de referencia, los haplotipos se representan mediante filas de 0's y 1's, que indican si aparece el alelo que se codifica como 0 o el que se codifica como 1 (explicado con más detalle posteriormente). Las columnas son los SNP's, que se dividen en dos conjuntos disjuntos:

  - Un conjunto T, azul: SNP's que están genotipados tanto en el conjunto de referencia como en la muestra de estudio.
  - Un conjunto U, verde: SNP's que únicamente están genotipados en el panel de referencia.

  La muestra de estudio contiene los genotipos codificados con 0's y 2's, representando un estado homocigoto; 1's, representando un estado heterocigoto y ¿?, representando un genotipo perdido.

  El objetivo de la imputación en esta situación es estimar los genotipos de los SNP's del conjunto U en la muestra de estudio.

- IMPUTE v2, que es la versión utilizada para este trabajo, se basa en el mismo modelo genético que IMPUTE v1, pero incluyendo un contexto estadístico que permite incrementar la precisión (usando más la información intrínseca de los datos) y manejando una variedad más amplia de conjuntos para la imputación.

  La principal diferencia respecto a la versión anterior es que se puede trabajar con 2 paneles de referencia: un conjunto de haplotipos conocidos y un conjunto de genotipos sin habérseles determinado los haplotipos previamente, siendo un subconjunto de los SNP's del panel superior. También puede ser aplicado en otras situaciones que, con IMPUTE v1, dan problemas, como puede ser trabajar con paneles de referencia que contengan numerosos cromosomas.

  Además de poder aplicar IMPUTE v2 en el escenario anterior, con IMPUTE v2 se puede tener el siguiente esquema:

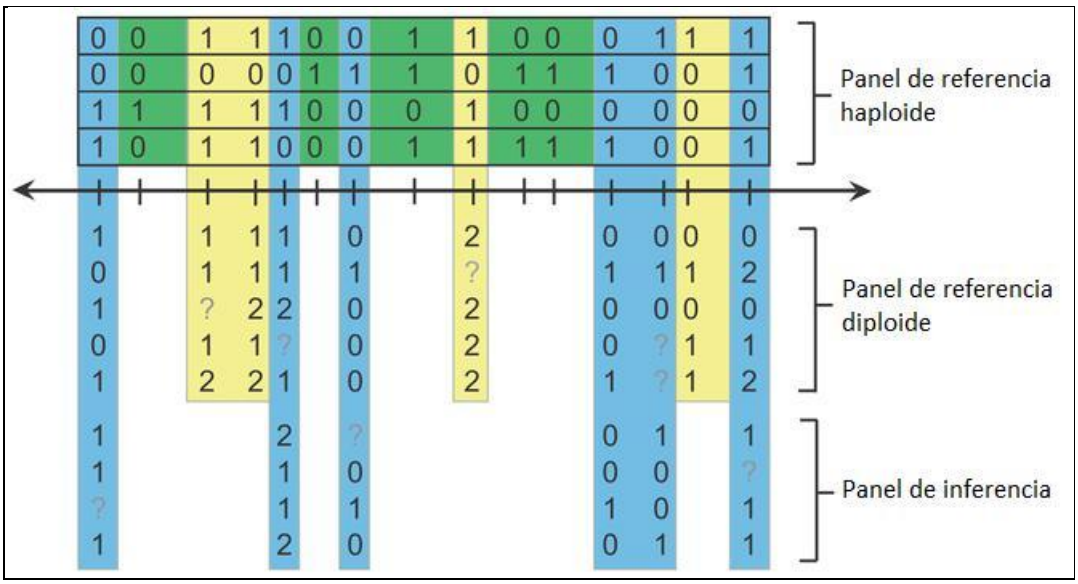


Figura 8:
http://www.plosgenetics.org/article/slideshow.action?uri=info:doi/10.1371/journal.pgen.1000529&imageURI=info:doi/10.1371/journal.pgen.1000529.g001#

Los haplotipos están representados mediante filas que contienen 0's y 1's que, al igual que en el escenario anterior, indican los alelos que corresponden a cada SNP. Los genotipos se representan mediante columnas de 0's y 2's para representar el estado homocigoto; 1's, que señalan el estado heterocigoto y ¿?, que corresponde a un genotipo faltante. Los SNP's son las columnas divididas en 3 conjuntos disjuntos:

- Un conjunto T, azul: en el que los SNP's están genotipados en todos los paneles.
- Un conjunto U_2, amarillo: SNP's que están genotipados en los paneles de referencia, pero no en la muestra de estudio.
- Un conjunto U_1, verde: conjunto de SNP's que sólo aparecen genotipados en el panel de referencia haploide.

El objetivo de la imputación es estimar los genotipos de los SNP's del conjunto U_2 y U_1 en la muestra de estudio y, si se desea, en el panel de referencia diploide.

Para aplicar el anterior escenario, se puede emplear también una población de referencia y la muestra de estudio. En este caso, los individuos de la muestra que tienen genotipados la mayoría (o todos) los SNP's se utilizan como un panel de referencia diploide, obteniendo así el segundo panel de referencia necesario para la imputación.

Para la realización de este trabajo se ha aplicado IMPUTE v2 en un escenario con un único panel de referencia, siguiendo el primer esquema.

Para explicar el proceso de imputación en dicho escenario es necesario definir:

- $H_R^{T,U}$, el conjunto de haplotipos de referencia conocidos de los SNP's en T y U (es decir, el panel de referencia completo).
- H_R^T, el conjunto de haplotipos de referencia de los SNP's en T.
- H_S^T, el conjunto de haplotipos de estudio no observados de los SNP's en T. Si hay $\boldsymbol{N_S}$ individuos en la muestra de estudio, sus haplotipos en los SNP's de T pueden representarse como $H_S^T = \{H_{S,1}^T, \dots, H_{S,N_S}^T\}$, donde $H_{S,i}^T$ es el haplotipo para el individuo de estudio i.

El método empieza estimando inicialmente los haplotipos en H_S^T, creando haplotipos consistentes con los genotipos observados. Entonces se realizan diversas iteraciones del modelo cadena de Markov Monte Carlo. Cada iteración actualiza cada muestra de estudio i (en orden arbitrario) siguiendo los 2 pasos siguientes:

1. Se crea un nuevo haplotipo $H_{S,i}^T$ para el genotipo observado $G_{S,i}^T$ del individuo i en los SNP's de T. Ello se realiza con la distribución condicional $Prob(H_{S,i}^T / G_{S,i}^T, H_{S,(-i)}^T, H_R^T, \rho)$, donde:

   - $G_{S,i}^T$ es el genotipo del individuo i para el SNP en T.
   - $H_{S,(-i)}^T$ contiene las estimaciones de los haplotipos actuales para los SNP's en T para todos los individuos excepto el i.
   - H_R^T contiene los haplotipos del panel de referencia para los SNP's en T.
   - ρ es el valor del índice de recombinación para la región de interés, que equivale a la función de probabilidad de transición entre pares de SNP's.

   El espacio de estados del modelo incluye todos los haplotipos conocidos en H_R^T y las actuales estimaciones de los haplotipos $H_{S,(-i)}^T$.

2. Se imputan nuevos genotipos para los SNP's en U, en función de $H_{S,i}^T$, $H_R^{T,U}$ y ρ. El espacio de estados para el HMM incluye los haplotipos del panel de referencia $H_R^{T,U}$. La imputación se realiza mediante

un algoritmo de inferencia 'forward-backward' para Modelos Ocultos de Markov independientemente del haplotipo que haya en $H_{S,i}^{T}$.

El método se aplica un relativo pequeño número de iteraciones que afectan solo al paso de creación de haplotipos (en nuestro caso 10 iteraciones), seguidas de un número de iteraciones principales que incluyen ambos pasos y que contribuyen a las probabilidades de imputación finales (en nuestro trabajo 30 iteraciones del MCMC).

Informalmente, se crean los haplotipos a partir de los genotipos de los SNP's de la muestra de estudio (conjunto T) y se buscan perfectas o similares coincidencias entre los haplotipos resultantes y los correspondientes haplotipos parciales en el panel de referencia, ya que los haplotipos que coinciden con los SNP's en T se asume que también coinciden con los SNP's en U. Una vez creados los haplotipos, se estiman los genotipos en U.

Los argumentos de entrada para IMPUTE son (IMPUTE2 2010):

- Fichero haps: Fichero con los haplotipos conocidos para la población de referencia, con una fila por SNP y una columna por haplotipo. Los alelos deben codificarse como 0 ó 1.

  Cada archivo de haplotipos necesita un archivo correspondiente de leyenda. Un ejemplo de fichero haps es la siguiente captura de pantalla, correspondiente a una parte de los haplotipos de un SNP:

```
0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0
0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1
0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0
0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 1 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
```

- Fichero legend: Contiene la información de los SNP's que aparecen en el archivo "haps". Cada fila tiene 4 columnas: identificador del SNP, su posición en el cromosoma (en pares de base), el alelo que se codificará como 0 y el alelo codificado como 1 en el correspondiente archivo "haps". Estos alelos pueden ser {A, T, C, G}. A continuación se muestra un ejemplo:

```
ID pos allele0 allele1
7-16692 16692 G C
7-16719 16719 G A
rs62052724 19701 T C
7-23912 23912 T A
7-26356 26356 A G
rs55911008 30599 T C
rs55742255 30658 A C
7-31273 31273 G C
```

- Fichero maps : Mapa de recombinación de la región que se va a analizar. Es un archivo formado por tres columnas. La primera es la posición que ocupa el SNP dentro del cromosoma (en pares de base),

la segunda es el índice de recombinación entre la posición actual y la siguiente posición en el mapa genético (en cM[4] /Mb[5]) y la última se refiere a la posición en el mapa genético (en cM). A continuación se muestra un ejemplo del mapa de recombinación para el cromosoma 7:

```
35411 0 0
40483 0.8765228552 0.0044457239215744
40852 0.8788667762 0.0047700257619922
41421 0.8613687356 0.0052601445725486
41892 0.860446802 0.0056654150162906
42920 0.8565783444 0.0065459775543338
43259 0.8544385007 0.0068356322060711
44167 0.8789872976 0.0076337526722919
44408 0.9350248461 0.007859093660202
44846 1.3095690725 0.008432684913957
```

- Fichero gens: Se trata del archivo que contiene los genotipos para el estudio que queremos imputar. Un posible ejemplo es el siguiente:

```
7 rs1800795 22766645 C G 0 1 0 1 0 0 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 1
0 1 0 0 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 0
1 0 0 1 0 0 1 1 0 0 0 1 0 0 1 0 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 0 0 1

7 rs1800796 22766246 C G 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 1 0 0

7 rs1800797 22766221 A G 0 1 0 1 0 0 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 0
1 1 0 0 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 0
1 0 0 1 0 0 1 1 0 0 0 1 0 0 1 0 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 0 0 1

7 rs3918226 150690176 C T 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
0 1 0 0 1 0 0 1 0 0 0 1 0 1 0 0 0 0 1 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0
0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0

7 rs7830 150709571 A C 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 0 1 0 0 1 0 1 0
0 1 0 1 0 0 0 0 1 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1
0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1
```

Las columnas que contiene son:

1. Identificador de la fila, generalmente es el número del cromosoma.
2. Identificador del SNP.
3. Posición del SNP en pares de base.
4. Los dos alelos que contiene el SNP.

4 Centimorgan, Ud. de medida de la frecuencia de recombinación. Equivale al 1% de probabilidad de que un marcador situado en un locus genético sea separado de un marcador en un segundo locus debido al entrecruzamiento en una generación individual. En los seres humanos, 1cM equivale, de media, a 1 millón de pares de bases.

5 Megabase, Ud. de medida de longitud comúnmente utilizada para describir la longitud de una molécula de ADN/ARN (aprox. un millón de pares de bases).

5. Los siguientes valores son grupos de 3 números que se refieren a las probabilidades de aparición de los 3 genotipos; es decir, si un alelo es A y el otro a, aparecerían las probabilidades de aparición de AA, Aa y aa para cada individuo.

Así, para la primera línea del ejemplo, se trata del SNP rs1800795, perteneciente al cromosoma 7 y que ocupa la posición 2276645, con genotipos formados por los alelos C y G. El primer individuo del fichero, al tener las probabilidades "0 1 0", presentará el genotipo CG; el siguiente individuo, con las probabilidades "1 0 0", tendrá el genotipo CC y, así, de esta manera, con toda la población.

- Int: Intervalo genómico que se usa para la inferencia. Se especifican los límites inferior y superior en pares de base. De esta manera se restringe la región que se quiere analizar o se crean regiones más manejables.
- Ne: Se utiliza para indicar el número de individuos en la población de referencia. Para la población caucásica, es 15.000.
- Fichero de salida: Archivo en el que se almacenan los resultados de la imputación. El formato es el mismo que el de ".gens" pero con las probabilidades imputadas. Además, también genera un archivo proporcionando una estimación de la calidad de la imputación para cada SNP.

Un ejemplo de código que se le pasaría a la herramienta IMPUTE es el siguiente:

```
./impute2 -m genetic_map_chr7_combined_b37.txt -h EUR.chr7.impute.hap -l
EUR.chr7.impute.legend -g 7/snps7.gen -int 22766210 22766655 -Ne 15000 -o
7/chr7.impute1000G -fix_strand_g -pgs_miss
```

Como se puede ver:

El fichero maps es: genetic_map_chr7_combined_b37.txt.

El fichero haps es: EUR.chr7.impute.hap.

El fichero legend es: EUR.chr7.impute.legend.

El fichero con los SNP's que se desean imputar es: snps7.gen, restringiendo el conjunto a aquellos situados en el intervalo 22766210-22766655.

El fichero donde se almacenan los resultados es: chr7.impute1000G.

4.4.2.3. Herramientas adicionales

Para poder convertir nuestra base de datos inicial a un formato que pueda aceptar IMPUTE, ha sido necesario utilizar la herramienta GTOOL. Ésta se encarga de transformar un conjunto de datos genotípicos en un formato adecuado para usar los programas SNPTEST e IMPUTE (GTOOL s.f.).

Para ello, se utilizan 2 archivos, uno .ped y otro .map.

El primero consiste en un fichero con tantas filas como individuos, con las siguientes columnas:

1. Identificador del individuo.

2. Identificador del padre.

3. Identificador de la madre.

4. Identificador del sexo del individuo.

5. Indicador de si es un individuo enfermo (1) o sano (0).

Como de los puntos 2 y 3 no se dispone de la información necesaria y el sexo del individuo no es relevante, se introducen 3 columnas con 0's a continuación del identificador del usuario. Estos datos se añaden a los genotipos que presenta cada individuo, añadiendo tantas columnas como SNP's haya. Finalmente, los espacios en blanco deben rellenarse con '00'.

Es decir, dados los siguientes SNP's:

```
SNP 1: A/T
SNP 2: C/A
SNP 3: C/T
SNP 4: G/A
SNP 5: -/G
SNP 6: A/G
```

El fichero .ped debe quedar como sigue, incluyendo al principio una columna con el número de fila del fichero:

```
1     TR10  0     0     0     1     AA    00    CC    GG    -G    GG
2     TR100 0     0     0     1     AA    CC    CT    GG    GG    GG
3     TR102 0     0     0     1     AA    CC    CC    AA    --    00
4     TR104 0     0     0     1     AA    CC    CT    AG    GG    GG
5     TR105 0     0     0     1     AA    CC    CC    GG    GG    GG
6     TR106 0     0     0     1     AA    CC    CC    GG    -G    GG
7     TR109 0     0     0     1     AA    CC    CC    GG    --    GG
8     TR110 0     0     0     1     AA    CC    CT    AG    GG    GG
9     TR112 0     0     0     1     AA    AC    00    AA    --    AG
```

Respecto al fichero .map, está formado por tantas filas como SNP's, dispuestos en el mismo orden que en el fichero .ped. Las columnas que contiene son las siguientes:

1. Número del cromosoma al que pertenece el SNP.

2. Identificador del SNP.

3. Distancia genética. En este caso, es un dato desconocido para nosotros y no influye en los resultados.

4. Posición del SNP en pares de base, coincidiendo con la versión que aparece en los paneles de referencia de IMPUTE. Este punto es muy importante, ya que los datos genómicos se actualizan constantemente y hay muchas versiones diferentes, incluyendo diferentes versiones para las situaciones de los SNP's. Así, este valor tiene que ser de la misma versión que la de los ficheros de referencia, si no, no se reconocerán los SNP's al pasarlos a IMPUTE.

Un ejemplo de fichero .map es la imagen siguiente:

```
7       rs1800795       0       22766645
7       rs1800796       0       22766246
7       rs1800797       0       22766221
7       rs3918226       0       150690176
7       rs7830          0       150709571
```

Con los dos argumentos de entrada comentados, GTOOL crea un fichero de salida .gen con el formato utilizado por IMPUTE. Una posible línea de código es la siguiente:

```
./gtool -P --ped definitivo/snps.ped --map definitivo/snps.map --og
definitivo/snps.gen --os definitivo/snps.sample
```

Una vez se haya realizado la imputación, todos los ficheros .gen se vuelven a convertir mediante GTOOL, creando un fichero .map y otro .ped, que se pasa directamente a una nueva tabla Excel y con la que se puede seguir trabajando.

4.5.Agrupamiento

El siguiente paso es ver cómo se agrupa cada muestra únicamente por la información que las variables le pueden aportar. Así, puede determinarse algún tipo de relación entre ellas o algún patrón de agrupamiento que sirva de ayuda para la predicción.

Con las técnicas de agrupamiento, no se tiene en cuenta la variable dependiente de salida, ya que se trata de medir la capacidad del modelo para crear grupos sin conocerlos de antemano. A todo el conjunto de métodos que no comparan los resultados obtenidos con las agrupaciones reales, en caso de conocerlas, se les denomina métodos de aprendizaje no supervisado.

El método escogido es el K-medias, que divide aleatoriamente los datos en k grupos, calculando el centro (o punto medio) de cada conjunto. De forma iterativa asigna elementos a cada conglomerado y recalcula el centro de los conjuntos hasta que no se puede mejorar el modelo.

Para aplicar el método K-medias se ha utilizado la herramienta Clementine, que permite mostrar las variables del modelo según la importancia que han tenido en el momento de realizar la agrupación.

Una de las desventajas de este tipo de métodos es que, no utiliza factores o covariables para realizar el agrupamiento, por lo que los grupos encontrados no tienen por qué corresponder a la clasificación que vamos buscando, sino a otros factores ocultos.

Una forma de averiguarlo es obteniendo la distancia entre los centros de cada conglomerado y compararla con la distancia de cada elemento al centro del conglomerado que se le ha asignado. Si esta distancia es menor, la asignación a dicho conglomerado es correcta, de lo contrario, no.

Por otro lado, otra forma es comparar 2 métodos diferentes de aprendizaje no supervisado. En función del número de coincidencias en la agrupación, ésta es correcta o no (Orallo, Práctica 3 de Minería de Datos. Validación con el Clementine 2010).

Como método utilizado para la comparación de los resultados obtenidos mediante K-medias, se utiliza el algoritmo bietápico. Se denomina así porque realiza las agrupaciones en 2 grandes fases. En la primera, crea subgrupos con los datos iniciales y, posteriormente, emplea agrupamiento jerárquico para unir de manera progresiva los grupos en otros más grandes.

Para ello, también se ha empleado la herramienta Clementine, donde aparecen las variables en función de su importancia para realizar los grupos.

Una vez obtenidos los resultados mediante ambos métodos, se agrupan en una tabla de contingencia para obtener las coincidencias en cada conglomerado.

Dado que se quiere conocer con seguridad si dichas coincidencias son relevantes, se aplica el método de la χ^2 mediante la herramienta R (Chi-square with R 2008).

El test de la χ^2 mide la diferencia entre los valores observados y los esperados, siendo éstos últimos correspondientes al caso en que las variables sean independientes (Asociación de variables cualitativas: test de Chi-cuadrado s.f.):

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Así, cuanto mayor sea el valor del test, mayor es la diferencia entre los valores observados y los correspondientes a si las variables fueran independientes; y, por lo tanto, mayor es la posibilidad de que las variables estén asociadas.

Se tendrán en cuenta dos hipótesis para realizar el test de la χ^2:

- Hipótesis nula: las coincidencias son aleatorias, no hay relación entre ambos métodos.
- Hipótesis alternativa: los métodos están relacionados, las coincidencias no se deben al azar.

Para rechazar la hipótesis nula se tiene en cuenta el p-valor. Rechazar dicha hipótesis supone admitir que la diferencia entre los valores observados y los esperados es elevada y la posibilidad de obtener un resultado mayor es muy baja.

Generalmente, el p-valor estándar para rechazar la hipótesis nula es 0.05, pero en ocasiones los valores no se ajustan exactamente a una distribución χ^2 y se puede introducir cierto grado de error. Es por ello que debe utilizarse una corrección (en este caso, corrección de Yates), y así reducir la posibilidad de que se produzca sobreajuste. Dicha corrección debe aplicarse en muestras pequeñas, cuando la frecuencia esperada es menor de un determinado valor, entre otras. Dado que puede tender a la sobre corrección y dejar fuera resultados interesantes, se utiliza sólo para el caso de tablas de dimensiones 2*2 (Yates' correction for continuity s.f.).

Por lo tanto, a través del test de la χ^2 se evalúan las matrices de contingencia donde se asocian los resultados obtenidos mediante K-medias y bietápico, ajustando los valores con la corrección de Yates. Así, se determina si realmente los resultados obtenidos con los métodos de aprendizaje no supervisado dan conclusiones similares.

4.6.Métodos de estudio de dependencias

Como se ha dicho en el apartado **4.1**, se tienen 3 variables cuantitativas de tipo continuo (t_isqm, Edad, Edad_donante) puesto que contienen valores dentro de un intervalo real. También se tienen variables cualitativas, cuyos valores son de tipo nominal pero aparecen codificadas mediante números. La principal diferencia y, por tanto, que se tiene que tener en cuenta para realizar éste y posteriores estudios, es el significado de cada tipo. Así, Edad expresa con los valores 34 y 45 que la persona con 45 años tiene 11 más que la de 34 años, Causa_muerte_donante, con los valores 2 y 5 no está indicando una relación de superioridad de 3 veces más, sino un tipo diferente de causa de la muerte.

Por este motivo, en primer lugar se analizan las variables cuantitativas y las cualitativas mediante técnicas diferentes y, posteriormente, se transforman las de tipo cuantitativo para poder aplicar una técnica común y observar las relaciones entre todas (Síntesis de las principales técnicas estadísticas aplicadas en la investigación sanitaria 1998).

El método elegido para evaluar la relación entre las variables cuantitativas es construir una matriz de correlación, en la que se mide cuán intensa es su asociación mediante los coeficientes de correlación. Dichos coeficientes, toman un valor dentro de [-1, +1] de forma que, cuanto más cercanos estén los valores a los extremos del intervalo, mayor será la asociación.

Mediante la herramienta R se ha obtenido la matriz de correlación para cada par de variables.

Respecto al análisis de dependencias de las variables cualitativas, el método empleado es la elaboración de una matriz de contingencia (Manual de Estadística. Capítulo III: DISTRIBUCIONES BIDIMENSIONALES s.f.). En ella, aparecen el número de elementos que se ha encontrado para cada categoría. Cada celda es el punto de cruce de las filas y de las columnas y, por tanto, el cruce entre las diferentes categorías que refleja el número de elementos coincidentes para cada categoría de las variables.

Para evaluar si las coincidencias entre las variables suponen una dependencia, se emplea el test χ^2, explicado anteriormente, utilizando el p-valor para aceptar una de las siguientes hipótesis:

- Hipótesis nula: las variables a estudiar son independientes, no existe ninguna relación entre ellas.
- Hipótesis alternativa: las variables a estudiar están relacionadas entre sí.

Además, se aplica la corrección de Yates en los casos de matrices de 2*2 para evitar sobre ajustes.

Finalmente, se evalúan las dependencias de las variables cualitativas con las cuantitativas. Para ello, se transforman las variables cuantitativas en cualitativas, agrupando sus valores en intervalos. Esta técnica es muy sencilla pero puede presentarse el inconveniente de que las asociaciones que aparezcan no sean del todo correctas, bien porque dependan de los intervalos realizados o porque se pierda información.

El método que se aplica posteriormente es el mismo que para el análisis de dependencias entre variables cualitativas, es decir, una vez realizada la segmentación de las variables numéricas, se realiza el test de χ^2 con la corrección de Yates para las matrices de 2*2.

4.7.Método de análisis de haplotipos

Como se ha descrito anteriormente, cuando los cromosomas se transmiten de padres a hijos, una parte de ellos es una copia idéntica, a excepción de las zonas que son el resultado de recombinaciones. La frecuencia de aparición de este fenómeno es de ¼ respecto al total de manera que la probabilidad de que en dos posiciones consecutivas de un cromosoma haya recombinación es pequeña y, por lo tanto, es posible que se observe correlación entre SNP's, conocida como desequilibrio de enlace. Ello también produce que, si un polimorfismo resulta modificado por una mutación, es posible que los que están situados a su alrededor también hayan sufrido variación y aparezcan asociados a la enfermedad, cuando realmente no es así.

Mediante el análisis de haplotipos se pretende, por un lado, determinar el conjunto de alelos que se transmite a la vez en una región del cromosoma, observando si presentan desequilibrio de enlace y, por otro, determinar los alelos que se tienen en cuenta para la creación de los haplotipos, reduciendo el número de SNP's que se van a utilizar en el análisis de asociación posterior.

El primer paso para la realización del análisis es determinar la posición de cada SNP en el cromosoma, ya que aquellos que estén situados en el mismo gen pueden sugerir que no han sufrido recombinaciones.

Dada la importancia del conocimiento del genoma humano, existen numerosas páginas en las que los investigadores aportan sus descubrimientos y permiten explorar en profundidad cada gen y cada SNP mediante aplicaciones de acceso libre, como son NCBI o HapMap Project (International HapMap Project s.f.) (NCBI. National Center for Biotechnology Information s.f.).

Es por ello que para obtener la posición de cada SNP y cómo está genotipado se ha recurrido a dichas aplicaciones.

Por otro lado, dado que HapMap se trata de un proyecto que está en desarrollo no se han encontrado todos las secuencias genotipadas, lo que ha dificultado algunos análisis posteriores.

Para la estimación de los haplotipos, como se ha comentado anteriormente, no hay ningún método 100% fiable debido a las posibles ambigüedades que se pueden presentar en la estimación a partir de los genotipos. Para este proyecto se ha utilizado la herramienta Haploview, cuya metodología de estimación es mediante el algoritmo de Esperanza Maximización, EM (Glaucoma-Associated CYP1B1 Mutations Share Similar Haplotype Backgrounds in POAG and PACG Phenotypes 2007).

Éste, se encarga de encontrar una estimación de la máxima verosimilitud para los parámetros de modelos estadísticos que dependen de variables inferidas[6]. En este caso, dichas variables serían el porcentaje de sujetos con cada haplotipo que presenta ambigüedad.

Es un método iterativo que alterna dos pasos (Expectation-maximization algorithm s.f.) (Raquel Iniesta 2005):

- Esperanza (E), en el que se calcula la esperanza de la verosimilitud incluyendo variables inferidas como si fueran observadas. En este caso se trataría de incluir en el modelo unas frecuencias para cada haplotipo como si fuesen las correctas. Con esto, se calcula la frecuencia de cada combinación de genotipos.

- Maximización (M), en el que se calculan los parámetros maximizando la verosimilitud estimada que se ha encontrado en el paso E. Estos parámetros estimados se utilizan para determinar la distribución de las variables inferidas en el siguiente paso E.

  Es decir, con las frecuencias esperadas para los genotipos que se han calculado en el paso anterior, se obtienen las nuevas frecuencias de cada haplotipo, maximizando la función de verosimilitud. Se comparan las frecuencias obtenidas y se repiten los pasos hasta que se obtienen unos valores estables.

De esta manera, una vez obtenidas las posiciones de los SNP's y sus genotipos, se introduce dicha información en la herramienta Haploview, agrupando los SNP's por genes. Así, se obtiene si presentan desequilibrio de enlace, las frecuencias de los haplotipos formados y si hay algún SNP que no se emplea para la creación del haplotipo.

Como una segunda etapa del análisis de haplotipos, se realiza un análisis de asociación de los mismos con la enfermedad mediante el paquete haplo.stats de la herramienta R (Sinnwell JP 2009).

El análisis que se realiza es una asociación lineal con la variable dependiente mediante una función *logit-link*. Ésta se trata de la transformación matemática de un modelo de regresión logística estándar de la siguiente forma (Link Functions and the Generalized Linear Model 2010):

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1 X$$

6 Variables no observadas en el modelo que se infieren mediante métodos matemáticos a través de otras que sí lo están.

Donde X son las variables haplotipo y Bi los coeficientes de regresión para cada haplotipo.

Como se trata de un estudio caso-control y se dispone de 2 posibles valores (0 o 1), se le debe indicar a la aplicación mediante el argumento “family=binomial”.

También se incluye la frecuencia mínima de aparición que deben tener los haplotipos en nuestra población. Es decir, todos aquellos haplotipos que estén por debajo de dicha frecuencia se incluirán en un mismo grupo llamado “rare”.

Al igual que en la herramienta Haploview, para determinar los haplotipos resultantes de la combinación de los posibles alelos de los SNP's, se emplea el algoritmo de Esperanza-Maximización, EM.

Si no convergiesen las frecuencias de los haplotipos con las de los genotipos, se buscaría el SNP causante de dicho error y no se tendría en cuenta para el análisis.

Los resultados obtenidos quedan de la siguiente forma:

- Tabla de Coeficientes: para cada haplotipo se indica el coeficiente de regresión estimado (coef), el error estándar (se), el t-estadístico correspondiente (t-stat) y el p-valor (pval).
- Tabla de Haplotipos: para cada haplotipo aparece su frecuencia de aparición y los alelos que lo forman, así como el nombre de los SNP's que aportan dichos alelos.

4.8.Modelos de asociación

El objetivo de este trabajo es poder determinar la presencia/ausencia de rechazo y cuantificar la relación entre dicho suceso con las variables genéticas dadas.

Las técnicas tipo con las que se evalúa la probabilidad de que aparezca un suceso y cuantificar cómo influye en él la presencia de diversos factores, son las técnicas de regresión.

Con la regresión lineal, que tal vez sea la idea inicial, se evalúa el suceso obteniendo cualquier tipo de valor, en lugar de restringirse a los valores de las categorías. Como necesitamos evaluar un suceso con varias posibles categorías se empleará regresión logística.

4.8.1. Modelo de asociación univariante

Los factores que van a influir en el suceso son SNP’s, variables categóricas con unas propiedades diferentes a las estadísticas tradicionales.

Como se ha comentado anteriormente, un SNP está formado por 2 alelos y puede tomar 3 genotipos diferentes, 2 homocigotos y uno heterocigoto. Uno de los alelos puede ser el causante de que el SNP se asocie con la enfermedad, dependiendo de la presencia o nº de copias de dicho alelo en el genotipo que codifique el SNP.

Por tanto, para aplicar la regresión logística, se van que tener en cuenta las diferentes formas de herencia posibles, ya que determinan cómo tiene que presentarse el alelo causante del riesgo de la enfermedad para que haga que el SNP sea un factor de riesgo o no (Raquel Iniesta 2005).

Así, se realizan los estudios de asociación mediante regresión logística para cada modelo genético.

Se tienen 2 hipótesis, la hipótesis nula (los SNP's no están asociados) y la alternativa (los SNP's sí están asociados), una de las cuales se debe rechazar mediante el p-value.

Para poder emplear regresión logística teniendo en cuenta diferentes modelos de herencia, se emplea la herramienta R; en concreto, el paquete SNPassoc.

Además, proporciona el p-value corregido por el método de Bonferroni (Juan R. González 2009).

La corrección de Bonferroni se aplica en problemas de múltiples comparaciones.

El problema que se está estudiando es si los pacientes presentan rechazo o no utilizando un nº de variables. Éste se podría considerar como un problema de múltiples comparaciones, ya que se quiere determinar las diferencias entre ambos grupos a través de las variables proporcionadas. Si, por ejemplo, se estuviese comparando una nueva forma de enseñar a escribir a los niños con la forma estándar mediante atributos como la gramática, parte oral, organización, tipo de contenido, etc. cuantos más atributos hubiera, más probable sería encontrar diferencias entre los grupos caso-control y encontrar alguno asociado a dichas diferencias. Ello sucede porque se considera un conjunto de propiedades simultáneamente; pudiendo llegar a conclusiones erróneas y rechazar la hipótesis nula cuando no se debía.

La corrección de Bonferroni prueba cada posible hipótesis de manera individual; es decir, si hay **n** características, para mantener un buen nivel de significancia, se divide éste por **n**. Así, la probabilidad de obtener resultados correctos no depende del número de pruebas a realizar (Bonferroni correction s.f.).

En este análisis, para un nivel de significación de 0.05, dado que se tienen 40 SNP válidos (2 de ellos no se consideran por ser monomórficos), el p-value corregido sería (0.05/40) 0.00125.

Sin embargo, la corrección de Bonferroni asume que las variables a analizar son independientes y, si en los análisis previos se demuestra que existe dependencia entre las variables genéticas, la corrección no se va a aplicar en los resultados.

Dado que se va a tener un modelo de regresión logística para cada modelo de herencia, hay que decantarse por uno de los cuatro. Para ello, se realiza una comparativa entre:

- Los p-valor que se obtendrían para un análisis de regresión logística realizado suponiendo que los SNP's siguen el modelo de herencia ideal.

- Los p-valor obtenidos con el análisis de regresión logística ajustando los SNP's al modelo de herencia, es decir, los resultados obtenidos.

El modelo de regresión logística que más se ajuste al modelo de herencia ideal, es el modelo que se elige como resultado (Steen s.f.). Si, para dicho modelo, hay SNP's que resultan significativos, implica que están asociados con la enfermedad y se incluirán en los modelos predictivos posteriores.

4.8.2. Modelo de asociación de pares de interacciones

Como se ha comentado anteriormente, las complejas interacciones entre genes y los factores ambientales juegan un gran papel en las enfermedades humanas. Cada vez hay más estudios que demuestran que, en lugar de las pequeñas modificaciones clásicas relativas a las leyes de Mendel, las interacciones tienen un efecto predominante.

Utilizando el paquete SNPassoc de la herramienta R, se pueden estudiar las interacciones entre pares de SNP's en función de cada modelo de herencia (Juan R. González 2009).

El método utilizado es un test de la razón de verosimilitud (LR) que permite comparar dos modelos de regresión logística, uno completo, con un número de variables, y otro reducido, con un subconjunto de las variables presentes en el primero (Pedro Larrañaga s.f.).

El objetivo del presente análisis es evaluar la importancia respecto a la enfermedad de pares de interacciones entre SNP's. Así, la comparación realizada en el test es entre modelos formados por combinaciones de pares de SNP's, obteniendo un p-value para cada par analizado.

El p-value permite aceptar o rechazar una de las siguientes hipótesis:

- Hipótesis nula: los parámetros que corresponden a las variables que forman parte del modelo completo pero no del reducido, valen 0.

- Hipótesis alternativa: las variables del modelo completo diferentes a las del reducido tienen parámetros diferentes de 0.

Al obtener un p-value para cada par de interacciones entre SNP's analizado, el resultado se muestra en forma de matriz, con los SNP's como filas y columnas y los p-value correspondientes a cada par analizado como los valores de las celdas de la matriz (Juan R. González 2009):

- En la diagonal principal aparecen los p-value que indican el nivel de asociación con la enfermedad de cada SNP de forma aislada. Hacen referencia al siguiente cociente de verosimilitudes (o resta, tomando logaritmos):

  $\mathrm{LRT}_{ii} = -2(\ln L(i) - \ln L0)$, donde:

  - $L(i)$ es el modelo completo que incluye el SNP i.

- $L0$ es un modelo nulo.

- En el triángulo superior están los p-value que indican el nivel de asociación de la interacción entre cada par de SNP's (lo que deseamos obtener). Hace referencia al cociente de verosimilitudes siguiente:

  $LRT_{ij} = -2(\ln L_f(i,j) - \ln L_a(i,j))$, donde:

  - $L_f(i,j)$ es el modelo completo correspondiente a la interacción entre el par de SNP's i y j, junto con el ajuste de ambos por separado.

  - $L_a(i,j)$ es el modelo reducido correspondiente a la suma de los SNP i, j.

- En el triángulo inferior se sitúan los p-value que comparan el modelo aditivo con el mejor de los modelos de cada SNP de manera aislada. Se refieren a:

  $LRT_{ji} = -2(\ln L_a(i,j) - \ln \max(L(i) - L(j)))$, donde:

  - $L_a(i,j)$ es el modelo aditivo con los SNP i, j.

  - $L(i)$ y $L(j)$ son los modelos de los SNP's i y j de manera aislada, respectivamente.

Como en el apartado anterior, el análisis de asociación de cada par de interacciones se realiza para cada modelo de herencia y, después, se elige cuál se ajusta más al modelo de herencia ideal. Los pares de SNP's cuya interacción resulte significativa para el modelo elegido, se seleccionan para incluirlos como variables en los modelos predictivos posteriores.

4.9.Modelos predictivos

Con la información recopilada en los pasos previos, se construyen los modelos predictivos. Se han elegido los algoritmos LDA, SVM y árboles de decisión por su eficiencia y la sencillez en la interpretación de los resultados que obtienen. Los dos primeros se realizan mediante la herramienta R, mientras que para los árboles de decisión se emplea la herramienta Clementine. Dichos métodos se aplican tanto para predecir la variable DCTRsi_no, indicadora de la presencia o ausencia de rechazo, como para predecir la variable DCTR_otrDCTR, indicadora de la ausencia y del tipo de rechazo.

Los tres algoritmos se van a aplicar en 3 ocasiones: primero, para crear un modelo únicamente con las variables clínicas significativas; otro únicamente con los SNP's significativos y, finalmente, un tercero con los SNP's y las variables clínicas como ajuste.

Las variables que se incluyen son todas las que se han encontrado como significativas en los pasos anteriores; esto es, SNP's que han aparecido como significativos de forma aislada, SNP's pertenecientes a pares de interacciones significativas, SNP's que dan lugar a haplotipos significativos y variables clínicas que han aparecido asociadas a la enfermedad.

4.9.1. LDA

El análisis discriminante lineal, LDA, es un algoritmo de aprendizaje supervisado que trata de explicar y predecir la pertenencia de un individuo a una clase con la ayuda de variables predictivas. Mediante este algoritmo se encuentra una combinación lineal de dichas variables, para obtener mejores resultados en la predicción que con las variables por separado (Linear Discriminant Analysis s.f.) (Mauricio Delbracio 2006). Además, la transformación lineal maximiza la distancia entre clases, tal y como se muestra en la imagen siguiente:

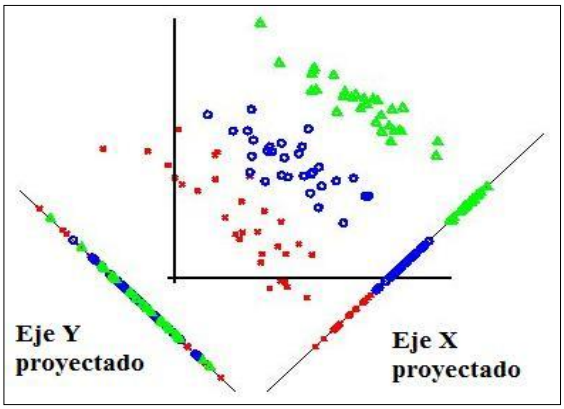


Figura 9: http://www.dtreg.com/lda.htm

En la imagen, la proyección utilizando LDA (eje X) da una mejor separación entre las clases que una proyección que no utiliza la etiqueta de clase (eje Y).

La combinación resultante puede ser utilizada como un clasificador lineal o, más comúnmente, para una reducción de la dimensionalidad previa a la clasificación, de manera que haya una mejor separación entre clases.

Para este proyecto se utiliza como método de predicción puesto que ya se habrá hecho una reducción de variables con los análisis realizados en los pasos previos.

4.9.2. SVM

Support Vector Machine, SVM, es un algoritmo de aprendizaje supervisado que toma un conjunto de datos de entrada y, para cada uno de ellos, predice a cuál de las posibles clases pertenece.

Para ello, de forma intuitiva, se puede considerar que representa las muestras como puntos en el espacio, separándolas, en función de la categoría a la que pertenecen, por una frontera tan amplia como se pueda. Hay una cantidad infinita de posibles líneas, la cuestión es determinar cuál es la mejor y cómo definirla (Xie 2007) (SVM - Support Vector Machines s.f.).

Tal y como aparece en la imagen siguiente, las líneas discontinuas marcan la distancia entre la frontera y los puntos más cercanos a dicha línea. La distancia entre las líneas discontinuas se conoce como margen, mientras que los vectores (puntos) son los que restringen la anchura del mismo:

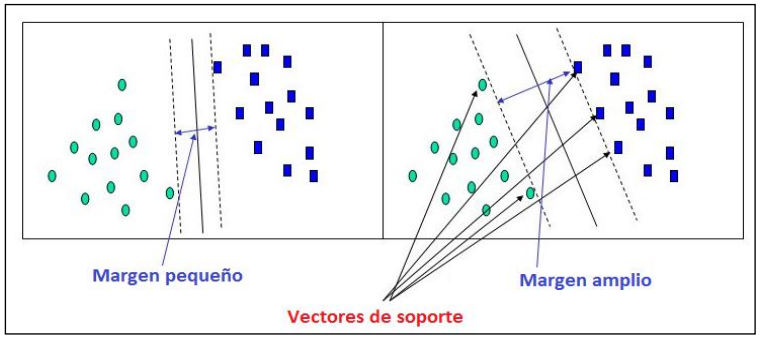


Figura 10: http://www.dtreg.com/svm.htm

Un análisis SVM encuentra la línea (o, en general, hiperplano) que se orienta de tal manera que el margen entre los vectores de soporte sea máximo (SVM - Support Vector Machines s.f.). En la figura anterior, el panel de la derecha es el que tiene una separación máxima entre clases.

4.9.3. Árboles de decisión

Se ha elegido el método de los árboles de decisión porque proporcionan sistemas de clasificación, de manera que se puede predecir y tomar decisiones a través de las reglas que se crean con los datos iniciales.

Así, los árboles de decisión se pueden ver como una colección de reglas con las que, en función del valor que tomen los atributos, clasifican los registros en uno u otro grupo, tal y como se puede ver en la imagen siguiente:

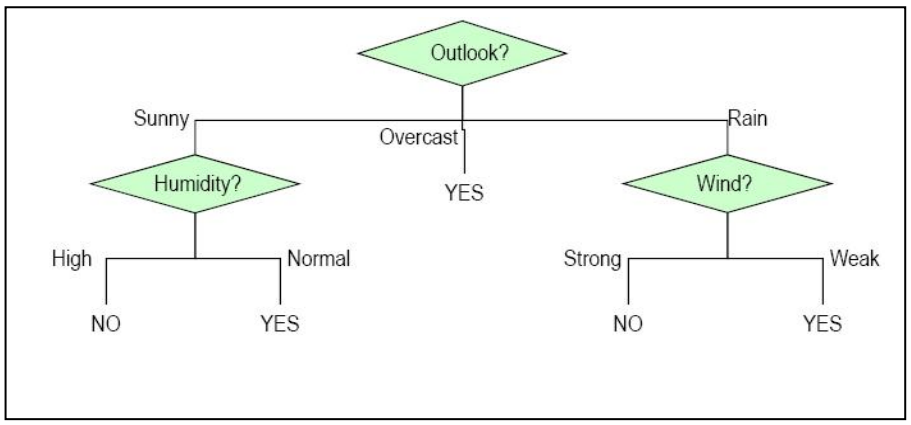


Figura 11: (Orallo, Tema 2: El proceso KDD. Técnicas de minería de datos. 2010)

El algoritmo utilizado es el llamado C.5 (SPSS 2000). Mediante este método, los datos se separan cada vez en función de un campo para obtener la máxima ganancia de información. Las ramas del árbol se dividen de forma que haya el máximo número de valores pertenecientes a la misma categoría. El proceso continúa

hasta que no se pueden realizar más subdivisiones. Una vez realizado esto, se evalúan de nuevo las divisiones del nivel inferior y, si hay alguna que no es relevante, se elimina.

Este método puede generar dos tipos de modelos: conjunto de reglas y árbol de decisión. En este caso, se crea el segundo tipo, que consiste en una descripción de las particiones encontradas mediante el algoritmo utilizado y permite realizar un pronóstico para cada registro.

Para aumentar la precisión del modelo se puede aplicar una opción de aumento, basada en el algoritmo Boosting propuesto por Yoav Freund y Robert Schapire (SPSS 2000) (AdaBoost s.f.). Con ella se genera un primer modelo y, en función de los errores de éste, se genera otro y así sucesivamente. Para clasificar cada registro se tienen en cuenta todos los modelos generados de forma ponderada, resultando un modelo global.

4.10. Estrategias de evaluación

Todos los resultados, independientemente del algoritmo utilizado, se muestran en función del número de aciertos y fallos en la predicción. Éstos se pueden expresar mediante una matriz de confusión, en la que las filas corresponden a los casos/controles reales y las columnas a aquellos obtenidos en el test. A continuación se muestra un ejemplo de los posibles resultados en una predicción:

	Test control (0)	**Test caso (1)**
Real control (0)	115 (49,78%) c	116 (50,22%) d
Real caso (1)	24 (9,2%) a	237 (90,80%) b

Mediante la matriz de confusión anterior se pueden extraer las conclusiones siguientes (Understanding and using sensitivity, specificity and predictive values 2008) (Valoración de pruebas diagnósticas s.f.):

- Sensibilidad: se trata de la bondad del modelo para predecir los casos, aquellos verdaderos positivos, como por ejemplo, pacientes enfermos. En la tabla anterior se trata de un 90,8%.
- Especificidad: se refiere a la bondad del modelo para predecir los controles, aquellos verdaderos negativos, como puede ser, pacientes sanos. En la matriz anterior, es un 49,78%.
- Balanced accuracy: es el valor medio entre la sensibilidad y especificidad. Con la matriz anterior, es 70,29%.
- Valor predictivo positivo (PPV): es la probabilidad, cuando la evaluación es positiva, de que corresponda a la presencia verdadera de la enfermedad. Por tanto, dada la matriz anterior, el cálculo es el siguiente: $\frac{b}{(b+d)}$, obteniendo un 67,13%.

- Valor predictivo negativo (NPV): es la probabilidad, cuando la evaluación resulta negativa, de que corresponda a la ausencia real de la enfermedad. Así, dada la tabla anterior, el cálculo es el siguiente: $\frac{c}{(c+a)}$, obteniendo un 82,73%.

Además, mediante el porcentaje de aciertos obtenidos en el test, se tiene la precisión del modelo (accuracy).

Por otro lado, las anteriores medidas, pueden representarse mediante un intervalo de confianza que indique, con cierta probabilidad de éxito o nivel de confianza (generalmente del 95%), que el valor de la medida correspondiente se encontrará dentro de dicho intervalo.

A continuación se describen dos formas de realizar la evaluación y obtener los resultados del modelo.

4.10.1. Entrenamiento-test

Si se tiene un gran número de datos, se puede dividir el conjunto en dos grupos de manera aleatoria, una para entrenar los datos y, así, generar un modelo, y otra para evaluar si dicho modelo se puede aplicar de manera general y con unos resultados óptimos. La partición implica una mayor proporción de datos para entrenamiento, por ejemplo un 85% respecto al total, y el resto para test.

El algoritmo de aprendizaje se aplica al conjunto de datos de entrenamiento, de manera que se obtienen unos coeficientes o reglas para las variables empleadas (conjunto conocido como modelo) que permiten predecir el valor de la variable dependiente. Para verificar que el modelo sea adecuado con cualquier dato y permita predecir obteniendo unos resultados óptimos, se utiliza con nuevos datos, es decir, con el conjunto de test.

4.10.2. Validación cruzada

Si se dispone de un número relativamente pequeño de registros para hacer una partición independiente, una para entrenamiento y otra test, tal vez el entrenamiento se realice con muy pocos datos y los resultados no sean fiables. Es por ello que, ante dichas situaciones, se emplea validación cruzada.

Mediante esta metodología, los datos se dividen en subconjuntos, por ejemplo 10, uno de ellos es para la evaluación y el resto para el entrenamiento (en este caso 9) y se van intercambiando los roles, de manera que todos se usen una única vez como test. Así, si se realizan 10 subconjuntos, son necesarias 10 iteraciones.

El error estimado para el modelo es la media de los errores obtenidos en la evaluación de cada iteración, lo cual implica unos resultados más generales y, por tanto, más fiables.

5. PREPARACIÓN DE LOS DATOS

Tal y como se ha descrito en los apartados 4.2 y 4.3, es necesario hacer una primera revisión de los datos con los que se va a trabajar para empezar a conocer la información que se dispone. Para ello, la herramienta empleada es Clementine. En primer lugar, se ven las características de las variables y, posteriormente, se realiza una primera aproximación de las relaciones entre las mismas de manera visual.

5.1.Calidad de los datos

De acuerdo a lo comentado en el apartado 4.2, es necesario realizar una serie de pasos para empezar a conocer los datos con los que se va a trabajar.

En primer lugar, se ha visto recomendable determinar la proporción de casos y control que se tienen para cada variable dependiente, ya que si dichas proporciones estuviesen desequilibradas se deberían realizar los cambios comentados en el apartado 4.2.

Para la variable dependiente DCTRsi_no, indicadora de la ausencia (0) y presencia (1) del rechazo del injerto, se han obtenido los siguientes porcentajes:

Valor	Proporción	%	Recuento
0.0		42,75	118
1.0		57,24	158

Para la variable dependiente DCTR_otrDCTR, indicadora de la ausencia de rechazo (0), rechazo crónico (1) y otros tipos de rechazo (2) se han obtenido las proporciones siguientes:

Valor	Proporción	%	Recuento
0.0		42,75	118
1.0		21,01	58
2.0		36,23	100

En ambos casos las proporciones de las clases se puede considerar que son adecuadas y no va a ser necesario realizar modificaciones para equilibrarlas o emplear la macromedia como medida de evaluación.

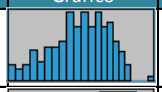
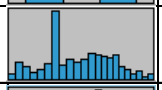
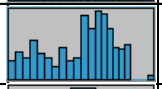
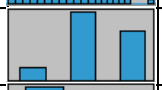
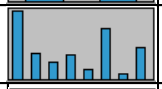
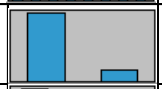
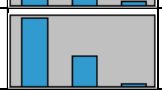
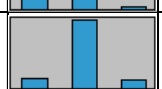
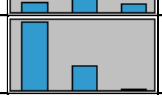
En segundo lugar, se determina la existencia de valores faltantes en las variables, ya que puede ser necesaria la imputación de dichos valores para el correcto análisis de asociación con la enfermedad. Los resultados se muestran en la tabla siguiente:

Campo	% completado	Registros válidos
Causa_muerte_donante	100.0	276
DCTR_otrDCTR	100.0	276
DCTRsi_no	100.0	276
Edad	100.0	276
Edad_donante	100.0	276
Enf_primaria	99.28	274
Matches_A	99.64	275
Matches_B	99.64	275
Matches_DR	99.64	275
Sexo	100.0	276
Sexo_donante	100.0	276
UO	100.0	276
rs1036199	99.64	275
rs10515746	98.55	272
rs1143634	94.93	262
rs12449782	99.64	275
rs175176	95.29	263
rs1799750	100.0	276
rs1799969	89.86	248
rs1800471	94.2	260
rs1800629	99.28	274
rs1800764	98.91	273
rs1800795	100.0	276
rs1800796	100.0	276
rs1800797	100.0	276
rs1800825	98.55	272
t_isqm	99.64	275

Campo	% completado	Registros válidos
rs1800871	91.3	252
rs1800872	99.28	274
rs1800896	99.64	275
rs1801275	97.1	268
rs2070874	99.64	275
rs2071231	99.28	274
rs2107538	98.19	271
rs2234676	99.28	274
rs2243248	100.0	276
rs2430561	99.64	275
rs243865	100.0	276
rs301640	98.91	273
rs3918226	99.64	275
rs41297579	98.91	273
rs419598	99.28	274
rs4311	85.87	237
rs4586	97.83	270
rs4696480	99.28	274
rs470206	100.0	276
rs4986790	93.84	259
rs4986791	100.0	276
rs5186	100.0	276
rs5743708	99.64	275
rs5749511	100.0	276
rs699	89.49	247
rs699947	93.48	258
rs7830	100.0	276
rs833061	98.91	273

Se han marcado en color verde las variables que no tienen ningún dato faltante. Así, como se puede ver, la mayoría, en especial las variables genéticas, tienen valores faltantes. De hecho, algunas sobrepasan el 10%, lo cual indica que va a ser necesario imputar.

En tercer lugar, se evalúan las características de las variables. A continuación, aparece la tabla que se ha obtenido para algunas de ellas:

Campo	Gráfico	Tipo	Mín	Máx	Media	Desv. típica	Asimetría	Únicos	Válidos
Edad		range	18.000	87.000	50.014	13.067	-0.283	--	276
Sexo		set	1.000	2.000	--	--	--	2	276
t_isqm		range	245.000	1.440.000	784.978	267.787	0.125	--	275
Edad_donante		range	14.000	82.000	45.877	15.078	-0.452	--	276
Causa_muerte_donante		set	0.000	5.000	--	--	--	3	276
Sexo_donante		set	0.000	1.000	--	--	--	2	276
Enf_primaria		set	1.000	8.000	--	--	--	8	274
UO		set	0.000	1.000	--	--	--	2	276
Matches_A		orderedSet	0.000	2.000	--	--	--	3	275
Matches_B		orderedSet	0.000	2.000	--	--	--	3	275
Matches_DR		orderedSet	0.000	2.000	--	--	--	3	275
rs1036199		set	--	--	--	--	--	3	275

Gracias a esta tabla, por ejemplo, se puede ver si el tipo de las variables que ha cargado a la herramienta Clementine es el adecuado o es necesario modificarlo para que los análisis posteriores se hagan de manera correcta. Los valores que aparecen son:

- Gráfico: Se trata de un histograma con la distribución de los datos. Así, para una variable dada, se ve la proporción de pacientes que tienen un valor dado.

  Por ejemplo, para el caso de la variable t_isqm, se observa que hay un valor predominante sobre el resto.

  En el caso de las variables Edad y Edad_donante, los datos siguen una distribución bastante uniforme, viendo que es más frecuente el trasplante en pacientes alrededor de 50 años. No obstante, hay un valor correspondiente a 87 y 82 años, respectivamente, que "rompe" la uniformidad de la distribución. No obstante, debido a la posibilidad de tratarse de un error en la extracción de los datos y tener que prescindir de dicho paciente, se consultó al ISC III y se confirmó que los datos eran correctos.

- Tipo: Según las variables se debe especificar un tipo u otro:

- Las variables discretas, se deben especificar como ‘flag’, si tienen dos posibles valores, o ‘set’, si pueden tomar más de 2 valores. Este es el caso de todos los SNP’s, y las variables siguientes: Sexo, Causa_muerte_donante, Sexo_donante, Enf_primaria y UO.

- Las variables Matches_A, Matches_B y Matches_DR son categóricas pero sus valores tienen una relación de orden descendente; por lo tanto deben especificarse como conjunto ordenado o ‘orderedSet’.

- Las variables numéricas Edad, t_isqm y Edad_donante se especifican como ‘range’.

- Rango de valores: Para el caso de variables categóricas cuyas clases son números y para las numéricas se indica el mínimo y máximo valor que toman. Éste es un buen indicador de la existencia de un dato anómalo, como pudiera ser una edad negativa.

- Media: Campo que indica la media de los valores de una variable.

- Desviación típica: Con la desviación típica se determina la variación de los datos respecto a la media; de manera que, cuanta más diferencia haya entre los valores de la variable, mayor será su desviación típica. Por lo tanto, la variable t_isqm presenta mayor diferencia entre sus valores que la Edad o Edad_donante, estando los valores de éstas últimas bastante próximos a la media.

- Asimetría: Con un valor positivo indica que las frecuencias más altas en la distribución de una variable numérica se encuentran a la izquierda de la media, mientras que a la derecha están las frecuencias más bajas. Un ejemplo de asimetría positiva es la variable t_isqm, en la que la mayoría de los valores se encuentran a la izquierda de la media, es decir, la mayoría de los trasplantes han presentado tiempos de isquemia por debajo de 784 minutos.

  Un valor negativo apunta a que las frecuencias más bajas están a la izquierda de la media, mientras que las más altas se encuentran a la derecha. Un ejemplo de este tipo son las variables Edad y Edad_donante, en las que, la mayoría de los donantes –receptor tienen edades por encima de los 50 y 45 años, respectivamente.

- Únicos: Este campo señala el número de posibles valores que toma la variable, en el caso de que sea discreta.

- Válidos: Indica el número de valores válidos que tiene la variable. Si no se contabiliza un valor es posible que se trate de un dato faltante o porque tenga un tipo erróneo.

Con la elaboración de esta tabla no se ha visto ninguna anomalía en las variables y no es necesario realizar ninguna modificación. Además, el número de valores válidos de cada variable coincide con el número de valores completos; por lo tanto, no se considera erróneo ningún otro dato.

Para comprobarlo y ver la tabla obtenida para todas las variables, se tiene la página 4 del documento “Punto 5.xlsx” de “anexos/Resultados”; además, en las variables de tipo continuo se incluyen enlaces a histogramas más visuales.

5.2.Visualización de la información

Una vez realizados los pasos anteriores, se utiliza la aplicación malla de la herramienta Clementine para visualizar las relaciones entre las variables. Los resultados que se muestran son las asociaciones más fuertes que la herramienta ha encontrado en función de, dadas 2 variables, el número de veces que para un valor de la primera, se da siempre el mismo valor para la segunda.

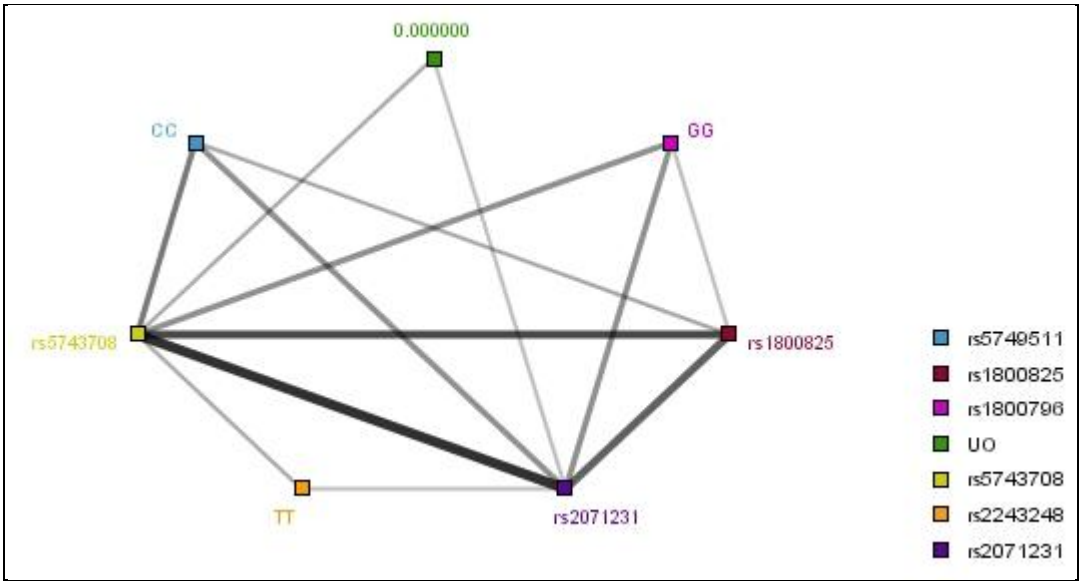
Respecto a la relación entre las variables, se ha obtenido la siguiente tabla:

Enlaces	Campo 1	Campo 2
263	rs2071231 = "TT"	rs5743708 = "GG"
256	rs1800825 = "TT"	rs5743708 = "GG"
251	rs1800825 = "TT"	rs2071231 = "TT"
246	rs5743708 = "GG"	rs5749511 = "CC"
244	rs1800796 = "GG"	rs5743708 = "GG"
241	rs1800796 = "GG"	rs2071231 = "TT"
241	rs2071231 = "TT"	rs5749511 = "CC"
235	rs1800825 = "TT"	rs5749511 = "CC"
235	UO = "0"	rs5743708 = "GG"
235	rs2243248 = "TT"	rs5743708 = "GG"
234	rs1800796 = "GG"	rs1800825 = "TT"
232	UO = "0"	rs2071231 = "TT"
230	rs2071231 = "TT"	rs2243248 = "TT"

El campo 'Enlace' indica el número de veces que el par de variables presentan los mismos valores. Los campos 'Campo 1' y 'Campo 2' corresponden a las 2 variables asociadas.

Dado que la muestra de la que se dispone es de 276 pacientes, éste valor será el máximo número de enlaces. Así, por ejemplo, en 263 individuos se da la situación en la que siempre que la variable rs2071231 sea igual a TT, la variable rs5743708 será GG.

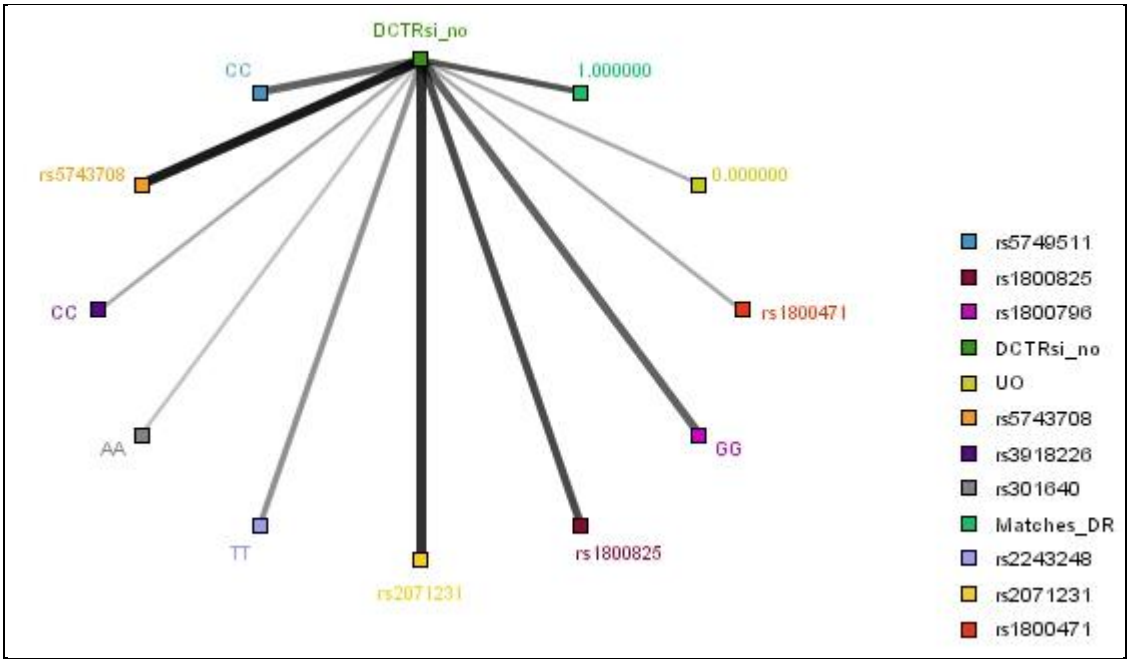
Dichas relaciones quedan reflejadas en el gráfico siguiente:



Para la variable dependiente indicadora de la presencia/ausencia de rechazo DCTRsi_no se han obtenido las siguientes asociaciones:

Enlaces	Campo 1	Campo 2
156	rs5743708 = "GG"	DCTRsi_no = "1"
153	rs2071231 = "TT"	DCTRsi_no = "1"
145	rs1800825 = "TT"	DCTRsi_no = "1"
142	rs5749511 = "CC"	DCTRsi_no = "1"
140	rs1800796 = "GG"	DCTRsi_no = "1"
134	rs2243248 = "TT"	DCTRsi_no = "1"
130	Matches_DR = "1"	DCTRsi_no = "1"
128	rs1800471 = "GG"	DCTRsi_no = "1"
126	rs3918226 = "CC"	DCTRsi_no = "1"
125	UO = "0"	DCTRsi_no = "1"
123	rs301640 = "AA"	DCTRsi_no = "1"

Representadas mediante el siguiente gráfico:

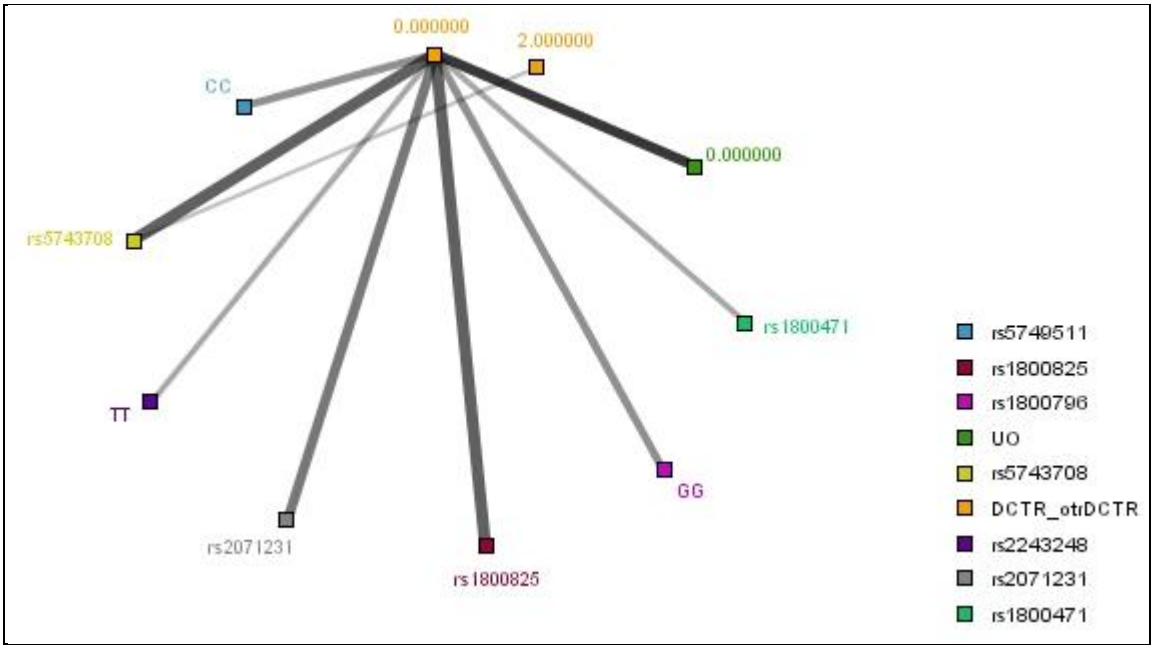


Así, por ejemplo, de las 158 ocasiones en que la variable DCTRsi_no = 1, en 142 el SNP rs5749511 toma el valor CC.

Para la variable DCTR_otrDCTR, indicadora del tipo de rechazo, se tienen las relaciones principales siguientes:

Enlaces	Campo 1	Campo 2
116	rs5743708 = "GG"	DCTR_otrDCTR = "0"
115	rs1800825 = "TT"	DCTR_otrDCTR = "0"
114	rs2071231 = "TT"	DCTR_otrDCTR = "0"
113	UO = "0"	DCTR_otrDCTR = "0"
108	rs1800796 = "GG"	DCTR_otrDCTR = "0"
108	rs5749511 = "CC"	DCTR_otrDCTR = "0"
104	rs2243248 = "TT"	DCTR_otrDCTR = "0"
100	rs1800471 = "GG"	DCTR_otrDCTR = "0"
99	rs5743708 = "GG"	DCTR_otrDCTR = "2"

Representadas en la malla siguiente:



En este caso, los SNP's asociados son los mismos que para la variable anterior, empleando los mismos valores. Además, el SNP rs5743708 aparece asociado con la variable DCTR_otrDCTR tanto para la ausencia (0) como para determinar otros tipo de rechazo (2). Ello es posible a que tal vez los SNP's tengan el mismo valor para la mayoría de los pacientes y el número de enlaces no sea debido a una asociación con la enfermedad, sino a que la variable tiene un valor predominante.

Por lo tanto, esto nos ha permitido tener una primera idea de las relaciones entre las variables, pero sin obtener nada concluyente.

6. IMPUTACIÓN DE LOS DATOS

Debido a la gran cantidad de datos faltantes que se ha visto en el apartado 5.1, es necesario realizar la imputación de dichos valores. Los métodos que se emplean son los descritos en el punto 4.4, imputación por el valor más frecuente y mediante la herramienta IMPUTE.

6.1. Imputación mediante el valor más frecuente

La imputación por el valor más frecuente se realiza mediante la herramienta R. En primer lugar es necesario leer los datos de la tabla Excel en la que están y cargarlos adecuadamente. Para ello, se han tenido que eliminar 2 columnas espaciadoras y la primera columna en la que se indicaba el nombre de cada fila.

Una vez cargados los datos, se especifica el tipo de cada variable. La herramienta reconoce de forma automática si se trata de variables categóricas, cuyos posibles valores sean de tipo carácter, o variables de tipo continuo. No obstante, es necesario especificar aquellas que sean categóricas pero sus posibles niveles sean numéricos y, si es el caso, especificar la relación de orden entre los niveles.

Para la imputación por el valor más frecuente se ha creado un script llamado 'rellena' que revisa, para cada variable, todos los valores, creando una lista con el número de veces que aparece cada uno. Una vez revisados todos los datos de la variable, se rellenan los espacios en blanco mediante el valor que se haya contabilizado más veces. A continuación se muestra un fragmento del código:

```
for(i in 1:ncol(snp)){
      cont[i] <- 0; ## se inicializa a 0 el nº de "" que tiene cada columna
      for(j in 1:nrow(snp)){
            if(snp[j,i]=="" | snp[j, i]=="NaN")
                  cont[i] <- cont[i] +1;
            }
      if(cont[i] > 0){
            frec <- table(snp[,i])  ## se obtienen las frecuencias de cada
nivel del factor
            ind <- which.max(frec)  ## se obtiene el indice de la frecuencia
maxima
            snp[,i][which(snp[,i]=="" | snp[,i]=="NaN")]<- levels(snp[,
i])[ind]  ## para toda la columna, se sustituyen los espacios y NaN por el
indice de maxima frecuencia de aparición
      ## se recalculan los niveles de los factores
            snp[, i] <- factor(snp[, i])
            }
```

Para la correcta ejecución de dicho script, se han pasado a tipo lista las variables numéricas, ya que de esta manera se facilita la contabilización de valores diferentes para la variable.

Una vez se ejecuta 'rellena', se valida su correcta ejecución mediante el script 'prueba', con el que se contabilizan el número de espacios en blanco que se puedan haber quedado pendientes de imputar. Un fragmento del código es el siguiente:

```
for(i in 1:ncol(snp)){
      cont[i] <- 0; ## se inicializa a 0 el nº de "" que tiene cada columna
      for(j in 1:nrow(snp)){
            if(snp[j,i]=="" | snp[j,i]=="NaN")
                  cont[i] <- cont[i] +1;
            }
      }
```

Finalmente, se devuelven a tipo numérico las variables que se habían modificado para la ejecución del script y las que eran de tipo discreto con relación de orden se establecen como tal (punto 1 y 2 de Scripts).

La nueva tabla con los datos imputados se convierte a Excel para que pueda almacenarse de forma más visual y pueda ser evaluada por la herramienta Clementine.

A continuación se incluye una muestra de los datos que se han imputado mediante la herramienta R. La tabla original es la siguiente:

Edad	Sexo	t_isqm	Matches_A	Matches_B	Matches_DR	rs1143634	rs12449782	rs1799750	rs1799969	rs1800471
47	2	600	0	0	1	CC	AA	GG		GG
68	2	792	1	0	0	CT	AG	GG	GG	GG
48	2	856				CT	GG	GG	GG	
59	1	600	1	1	1	CC	AG	GG	GG	CG
59	1	625	0	0	1	CC	AG	--	AG	GG
44	1	1200	1	0	1	CC	AA	--	AG	GG
50	2	510	1	1	1	CC	GG	GG	GG	GG
43	1	735	2	0	1	TT	GG	-G	GG	GG
45	1	840	0	1	1	CT	GG	-G	GG	GG
35	2	1140	0	0	1	CC	GG	--	GG	GG
59	2	2,6965E+308	0	1	1	CC	AG	--	GG	GG
70	2	960	1	1	1		GG	GG		GG

Tras la imputación queda:

Edad	Sexo	t_isqm	Matches_A	Matches_B	Matches_DR	rs1143634	rs12449782	rs1799750	rs1799969	rs1800471
47	2	600	0	0	1	CC	AA	GG	GG	GG
68	2	792	1	0	0	CT	AG	GG	GG	GG
48	2	856	0	0	1	CT	GG	GG	GG	GG
59	1	600	1	1	1	CC	AG	GG	GG	CG
59	1	625	0	0	1	CC	AG	--	AG	GG
44	1	1200	1	0	1	CC	AA	--	AG	GG
50	2	510	1	1	1	CC	GG	GG	GG	GG
43	1	735	2	0	1	TT	GG	-G	GG	GG
45	1	840	0	1	1	CT	GG	-G	GG	GG
35	2	1140	0	0	1	CC	GG	--	GG	GG
59	2	600	0	1	1	CC	AG	--	GG	GG
70	2	960	1	1	1	CC	GG	GG	GG	GG

6.2.Imputación mediante la herramienta IMPUTE

Para la imputación mediante la herramienta IMPUTE es necesario preparar los datos en el formato adecuado. Con este tipo de imputación sólo se rellenan los datos faltantes de las variables genéticas, haciendo que permanezcan los espacios en blanco del resto de variables.

En el punto 5.1 se ha visto que falta 1 valor para Matches_A, Matches_B, Matches_DR y t_isqm y 2 valores para Enf_primaria. Aun siendo un número relativamente reducido de datos faltantes, se ha visto a qué pacientes corresponden dichos valores viendo que el paciente que no consta de Matches_A, tampoco tiene Maches_B ni Matches_DR.

Los pacientes para los que no aparece registrado el tiempo de isquemia (t_isqm) ni la enfermedad primaria (Enf_primaria) son diferentes, siendo necesario, por tanto, eliminar en total 4 pacientes, número que no hará variar los resultados para la predicción en caso de eliminarse. No obstante, para la imputación de las variables genéticas sí que se mantienen, ya que para la mayoría de valores de variables genéticas conservan los datos y pueden ayudar en la imputación.

Como se ha comentado anteriormente, el primer paso consiste en emplear la aplicación GTOOL.

Una muestra de la tabla inicial de las variables genéticas es:

rs1036199	rs10515746	rs1143634	rs12449782	rs175176	rs1799750	rs1799969	rs1800471
AA		CC	GG	TT	-G	GG	CG
AA	CC	CT	GG	TT	GG	GG	GG
AC	AC	CT	AG	TT	-G	GG	GG
AA	CC	CC	AA	TT	--		
AA	CC	CC	AG	TT	-G	GG	GG
AA	CC	CT	AG	TT	GG	GG	GG
AA	CC	CC	GG	TT	GG	GG	GG
AA	CC	CC	GG	TT	-G	GG	GG
AA	CC		GG	TT	-G	GG	
AC	AC	CC	AG	TT	GG	GG	GG

Tras la primera modificación necesaria para pasar los datos a GTOOL, rellenando las casillas vacías con '00' e incluyendo número e identificador del individuo, así como el identificador de su padre y de su madre, identificador del sexo del individuo y de si se trata de un paciente enfermo o sano, la tabla queda:

1	TR10	0	0	0	0	AA	00	CC	GG	TT	-G	GG	CG
2	TR100	0	0	0	0	AA	CC	CT	GG	TT	GG	GG	GG
3	TR102	0	0	0	1	AC	AC	CT	AG	TT	-G	GG	GG
4	TR104	0	0	0	0	AA	CC	CC	AA	TT	--	00	00
5	TR105	0	0	0	1	AA	CC	CC	AG	TT	-G	GG	GG
6	TR106	0	0	0	0	AA	CC	CT	AG	TT	GG	GG	GG
7	TR109	0	0	0	0	AA	CC	CC	GG	TT	GG	GG	GG
8	TR110	0	0	0	0	AA	CC	CC	GG	TT	-G	GG	GG
9	TR112	0	0	0	1	AA	CC	00	GG	TT	-G	GG	00
10	TR113	0	0	0	1	AC	AC	CC	AG	TT	GG	GG	GG

Una vez realizadas las modificaciones, se pasa a formato de texto plano con la extensión '.ped'. Este es el fichero '.ped' que necesita como argumento la herramienta GTOOL.

Las cabeceras con los nombres de los SNP's se han eliminado ya que, tanto el número del cromosoma al que pertenece el SNP, como su identificador, distancia genética y posición en el cromosoma son los datos que deben indicarse en un fichero adicional. Así, se crea una tabla como la siguiente:

5	rs1036199	0	156531736
5	rs10515746	0	156536568
2	rs1143634	0	113590390
17	rs12449782	0	61576249
22	rs175176	0	20128685
11	rs1799750	0	102670496
19	rs1799969	0	10394792
19	rs1800471	0	41858876

Esta tabla se debe pasar a formato de texto plano con extensión '.map', que se corresponde con el fichero '.map' asociado al '.ped' creado en el paso previo.

Con el comando indicado en el apartado 4.4 GTOOL crea el fichero '.gen' cuyo formato es el que se ha explicado para los ficheros '.gen' de IMPUTE. A continuación se muestra un ejemplo:

```
5 rs1036199 156531736 A C 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0
1 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0
0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 0 1 1 0
0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 1 0 0 0
1 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0 1 0 0 1
0 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1
0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1
0 0 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0
0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1
1 0 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 1 0 1
0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0
1 0 0 1 0 1 0 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 1
0 0 1 0 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0
1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0
0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1
1 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0
1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 1 0 0 1
0 0 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1
0 1 0 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0
0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 1 0 0 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1
0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0
1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0
5 rs10515746 156536568 A C 0 0 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1
0 1 0 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1
0 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 0 1 1 0 0 0
0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1
0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 1 0 1 0 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0
0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1
0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0
1 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0
1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1
0 0 1 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0
0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1
1 0 0 1 0 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0
1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 1 0 0 1
0 0 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1
0 1 0 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0
0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 1 0 0 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1
0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0
1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0
```

Como se puede ver, el primer triplete de ceros para el segundo SNP, rs10515746, corresponde con la primera casilla rellenada con '00' del fichero '.ped' ya que se trata de un dato faltante, e igualmente sucede con el resto de tripletes subrayados en verde.

Para la imputación, se deben crear tantos ficheros '.gen' como cromosomas haya, puesto que IMPUTE trabaja cromosoma a cromosoma. Además, dentro del fichero '.gen' es necesario ordenar cada SNP en función de la posición que ocupa en el cromosoma.

Dado que no todos los SNP's tienen datos faltantes, únicamente se imputan aquellos que lo necesitan mediante la especificación del intervalo que ocupan dentro del cromosoma.

Tras la imputación, se genera un fichero '.gen' pero, donde anteriormente había un triplete de 0's, ahora hay un triplete de 0's y 1's que indica qué genotipo imputado le corresponde al paciente. Así, para la anterior muestra, ahora se encuentra como sigue:

```
5 rs1036199 156531736 C A 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0
0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 1 0 0 0 0
1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0
1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1
0 0 1 0 1 0 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0
0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1
0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 1
0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
0 1 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1
0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0 1
0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 1 0 0 0 1
0 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
0 0 1 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1
0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1
0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0 1
0 0 1 0 0 1 0 0 1 0 0 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1
0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 1 0 0 1 0
0 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
1 0 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1
0 0 1 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0
1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 1 0 0 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1
0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1
0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
5 rs10515746 156536568 A C 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
1 0 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0
0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 1 0 0 0 0
1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0
0 0 1 0 1 0 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0
1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0
0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1
0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 1
0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
0 1 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1
0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0 0 1 1
0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1
0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 1 0 0 1 0
0 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
1 0 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1
0 0 1 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 0
1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 1 0 0 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0 1 0 0 0 1 0
0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 0 0 1
0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
```

Tras la imputación de los SNP's, se agruparon tanto aquellos imputados como los que no para formar el fichero '.ped' mediante GTOOL y poder convertirse a una tabla Excel similar a la original.

No obstante, durante la imputación aparecieron errores para el SNP rs175176 y no se pudo imputar. En estos casos, IMPUTE no modifica el triplete de 0's y, al convertirse a un fichero '.ped', en la casilla correspondiente aparece 'N N'.

Dado que se trata de un SNP monomórfico y todos aquellos que tengan esta característica se eliminan del análisis, no supone ningún inconveniente y la imputación se ha realizado correctamente.

Una parte del fichero '.ped' tras la imputación, convertido a una tabla Excel, es:

0	TR10	0	0	-9	-9	A A	C C	G G	G G	T T	G G	G G	C G
1	TR100	0	0	-9	-9	A A	C C	G A	G G	T T	G G	G G	C C
2	TR102	0	0	-9	-9	C A	A C	G A	G A	T T	G G	G G	C C
3	TR104	0	0	-9	-9	A A	C C	G G	A A	T T	G G	G G	C C
4	TR105	0	0	-9	-9	A A	C C	G G	G A	T T	G G	G G	C C
5	TR106	0	0	-9	-9	A A	C C	G A	G A	T T	G G	G G	C C
6	TR109	0	0	-9	-9	A A	C C	G G	G G	T T	G G	G G	C C
7	TR110	0	0	-9	-9	A A	C C	G G	G G	T T	G G	G G	C C
8	TR112	0	0	-9	-9	A A	C C	G G	G G	T T	G G	G G	C C
9	TR113	0	0	-9	-9	C A	A C	G G	G A	T T	G G	G G	C C

Las columnas correspondientes al sexo del individuo y al indicador de enfermo o sano tienen el valor '-9' que, junto con el '0' referente al identificador de la madre y del padre del individuo, indican que se trata de campos desconocidos.

Adicionalmente, si se compara con la tabla del fichero '.ped' inicial, se observa que los genotipos de algunos SNP's han cambiado. En la muestra, es el caso de la 3ª, 6ª y 8ª columnas, correspondientes a los SNP's rs1143634, rs1799750 y rs1800471, respectivamente. Esto ocurre porque durante la imputación, en algunos casos, la herramienta IMPUTE emplea las bases complementarias de los alelos que se le pasan en el fichero de entrada. En dichos casos, es necesario realizar la conversión a los alelos originales una vez se encuentren en la tabla Excel.

Finalmente, se añaden las columnas de las variables clínicas a las de los SNP's, eliminando los 4 pacientes que carecían de valor para el tiempo de isquemia (t_isqm), enfermedad primaria (Enf_primaria) y compatibilidad A, B y DR (Matches_A, Matches_B y Matches_DR).

7. ANÁLISIS DE LOS DATOS

Para el análisis de los datos se eliminan, tanto de los datos imputados como de los originales, los SNP's rs175176 y rs470206 por ser monomórficos.

Además, durante la realización del trabajo, se informó que las variables clínicas Enf_primaria, Casua_muerte_donante y UO no estaban tomadas correctamente y podrían producir conclusiones erróneas en la predicción, por lo tanto, se eliminan también de los grupos de datos que se van a analizar.

7.1.Agrupamiento

Como se ha comentado en el apartado 4.5, se va a analizar si los datos se agrupan arbitrariamente o si siguen algún patrón antes de realizar cualquier modificación de los mismos. Además, también se puede determinar si, tras la imputación, hay diferencias significativas en la forma en que cada algoritmo agrupa los datos.

Para ello se emplean los algoritmos Kmedias y biétapico de Clementine, analizando si ambos métodos coinciden en la agrupación. Además, para evaluar la aleatoriedad de dichas coincidencias, se presentan en una matriz de contingencia a la que se realiza el test de la χ^2 mediante la herramienta R.

- Con los datos iniciales, pero eliminando los registros con algún dato faltante, se ha obtenido:
    - Para la realización de 2 grupos:

$KM-K-Medias	conglomerado-1	conglomerado-2
conglomerado-1	44	34
conglomerado-2	29	45

Pearson's Chi-squared test with Yates' continuity correction

data: .Table

X-squared = 3.8482, df = 1, p-value = 0.0498

    - Para la realización de 3 grupos:

$KM-K-Medias	conglomerado-1	conglomerado-2	conglomerado-3
conglomerado-1	27	18	31
conglomerado-2	10	19	12
conglomerado-3	2	33	0

Pearson's Chi-squared test

data: .Table3

X-squared = 48.3876, df = 4, p-value = 7.835e-10

En ambos casos se ha obtenido un p-valor por debajo del estándar 0.05 y, por tanto, las coincidencias no son arbitrarias. Para el primer caso, teniendo en cuenta las celdas que maximizan las coincidencias, (34+29)63 registros de 152 tendrían una asignación diferente. No obstante, para el caso de 3 conglomerados, el p-valor también es inferior a 0.05 pero no hay coincidencias definidas para todos los grupos.

Algunas de las variables que han resultado significativas con el método Kmedias y bietápico para la agrupación en 2 clusters son: rs1800797, rs10515746 y rs4986790, siendo también significativas las 2 primeras para la agrupación en 3 clusters mediante Kmedias.

No obstante, todos los detalles de este apartado se encuentran en la página 1 de "Punto 7.xlsx" que se adjunta en "anexos/Resultados".

- Con los datos imputados por el valor más frecuente se ha obtenido:
  - Para la agrupación en 2 conglomerados:

$KM-K-Medias	conglomerado-1	conglomerado-2
conglomerado-1	46	110
conglomerado-2	95	25

Pearson's Chi-squared test with Yates' continuity correction

data: .Table21

X-squared = 65.0176, df = 1, p-value = 7.423e-16

  - Para la agrupación en 3 conglomerados:

$KM-K-Medias	conglomerado-1	conglomerado-2	conglomerado-3
conglomerado-1	25	55	36
conglomerado-2	43	9	20
conglomerado-3	52	34	2

Pearson's Chi-squared test

data: .Table31

X-squared = 59.5563, df = 4, p-value = 3.596e-12

En este caso, las coincidencias tampoco se consideran aleatorias; además, para el agrupamiento en 2 clusters, tan sólo (46+25) 71 registros de 276 se agruparían de forma errónea. No obstante, para la situación de 3 conglomerados las coincidencias siguen sin estar claramente definidas. Tal vez el aumento de las coincidencias se deba a los efectos de la imputación, por lo que no es fácil determinar la buena o mala calidad de la agrupación.

- Utilizando los datos imputados mediante IMPUTE los resultados han sido:
  - Para la realización de 2 conglomerados:

$KM-K-Medias	conglomerado-1	conglomerado-2
conglomerado-1	56	96
conglomerado-2	99	21

Pearson's Chi-squared test with Yates' continuity correction

data: .Table22
X-squared = 55.1831, df = 1, p-value = 1.098e-13

  - Para la agrupación en 3 clusters:

$KM-K-Medias	conglomerado-1	conglomerado-2	conglomerado-3
conglomerado-1	62	57	1
conglomerado-2	21	14	44
conglomerado-3	29	29	15

Pearson's Chi-squared test

data: .Table32
X-squared = 84.1647, df = 4, p-value < 2.2e-16

Con los datos imputados mediante IMPUTE la situación es similar. Para la agrupación en 2 conglomerados, además de no ser aleatorias las coincidencias, considerando aquellas celdas que las maximizan, tan sólo 77 de 272 registros no estarían correctamente agrupados. Respecto a la agrupación en 3 clusters, siguen sin haber coincidencias tan definidas como para la separación en 2 grupos.

Como conclusión a este apartado, se puede decir que ambas imputaciones mantienen la tendencia que presentan los datos originales: para una agrupación en 2 conglomerados tienden a agrupar los registros con un éxito de más del 50% sin haberlos tratado previamente; mientras que para una agrupación en 3 conglomerados no se garantiza un agrupamiento fiable, a pesar de tratarse de coincidencias no aleatorias.

7.2.Análisis de dependencias

A lo largo de este apartado se van a determinar si existen dependencias entre las variables, evaluándolas en función de su naturaleza discreta o continua. Para visualizar mejor los resultados, se pueden consultar las páginas 2, 3 y 4 de “Punto7.xlsx” que se incluye en “anexos/Resultados”.

7.2.1. Variables cuantitativas

En este apartado se estudian las asociaciones entre las variables de naturaleza continua: Edad, Edad_donante y t_isqm. Para ello, con la herramienta R, se crea una matriz de correlación; así, mediante los coeficientes de dicha matriz, se determina la intensidad de la asociación entre cada par de variables.

Las variables Edad y Edad_donante no tenían valores faltantes, pero sí que los tenía t_isqm; es por ello que la matriz de correlación se ha obtenido empleando tanto con los datos imputados mediante el valor más frecuente como mediante IMPUTE. Los resultados obtenidos son los siguientes:

- Con la imputación por el valor más frecuente:

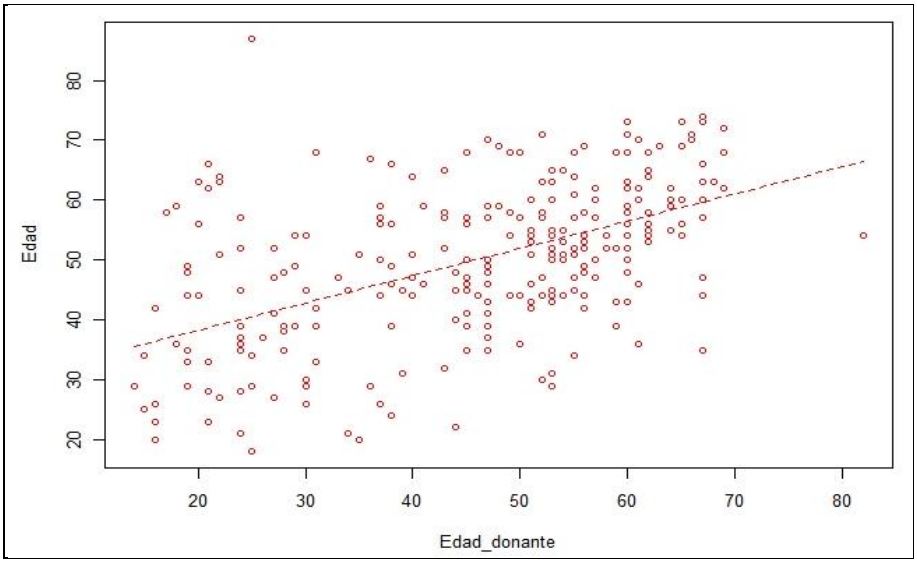
	Edad	Edad_donante	t_isqm
Edad	1.00	0.52	0.11
Edad_donante	0.52	1.00	0.03
t_isqm	0.11	0.03	1.00

- Con la imputación mediante IMPUTE:

	Edad	Edad_donante	t_isqm
Edad	1.00	0.53	0.11
Edad_donante	0.53	1.00	0.03
t_isqm	0.11	0.03	1.00

Como se puede apreciar, en ambos casos se obtiene un coeficiente de correlación por encima de 0.5 para el par de variables Edad-Edad_donante, que indica cierto grado de asociación positiva entre ambas.

Además, se ha realizado una gráfica que muestra la ligera asociación entre dichas variables, representando mediante la línea de puntos la función que mejor se ajusta a los datos según el mínimo error cuadrático:



Dado el carácter lineal de la asociación se puede intentar mejorar la precisión del modelo empleando como variable el cociente Edad/Edad_donante, en lugar de las 2 variables por separado.

7.2.2. Variables cualitativas

Para determinar las asociaciones entre las variables discretas se ha empleado, para cada par de variables, una matriz de contingencia, utilizando el test de la χ^2 para determinar si las coincidencias son significativas o no.

Con la herramienta R se ha creado el script "contingenciaInt". Este script realiza una tabla de contingencia para cada par de variables, aplicando la corrección de Yates cuando las dimensiones de las tablas son de 2*2. Una vez creada la tabla, guarda aquella que tenga un p-valor inferior al valor que se le pasa previamente como argumento al script, obteniendo únicamente las asociaciones significativas.

A continuación se muestra un ejemplo de código:

```
.Table <- xtabs(~datos[, i] + datos[,j], data=datos)
      if((length(levels(datos[, i])) > 2) | (length(levels(datos[, j])) > 2) )
            .Test <- chisq.test(.Table, correct=FALSE)
      else
            .Test <- chisq.test(.Table, correct=TRUE)
      if(.Test$p.value < sig){
            dep<- dep +1;
            mensaje <- paste("datos[, i] = ", names(datos[i]), "; datos[, j] =
", names(datos[j]))
```

El script "contingenciaInt" se ha ejecutado asumiendo una significación de 0.05. Además, se ha aplicado tanto para los datos imputados por el valor más frecuente como para la imputación mediante IMPUTE, obteniendo 73 y 66 asociaciones significativas, respectivamente.

Utilizando los datos imputados mediante ambos métodos, se han dado asociaciones entre el mismo par de variables. Algunos ejemplos son los siguientes:

- Mediante los datos imputados por el valor más frecuente:

		rs10515746		
		AA	AC	CC
rs1036199	AA	0	7	196
	AC	5	65	1
	CC	2	0	0

Pearson's Chi-squared test

data: .Table
X-squared = 310.9258, df = 4, p-value < 2.2e-16

		rs1800764		
		CC	CT	TT
rs12449782	AA	2	7	34
	AG	13	92	38
	GG	48	35	7

Pearson's Chi-squared test

data: .Table
X-squared = 125.1343, df = 4, p-value < 2.2e-16

- Mediante los datos imputados con IMPUTE:

		rs10515746		
		AA	AC	CC
rs1036199	AA	0	7	193
	AC	5	65	0
	CC	2	0	0

Pearson's Chi-squared test

data: .Table
X-squared = 311.0133, df = 4, p-value < 2.2e-16

		rs1800764		
		CC	CT	TT
rs12449782	AA	2	6	35
	AG	12	90	38
	GG	47	34	8

Pearson's Chi-squared test

data: .Table
X-squared = 126.3732, df = 4, p-value < 2.2e-16

De 47 variables utilizadas, más de 60 asociaciones entre ellas es un número bastante elevado, tal vez producido por falsas asociaciones producidas por las imputaciones. No obstante, anteriormente se ha explicado cómo los SNP's pueden estar asociados entre sí debido al proceso de transmisión de la información hereditaria, lo que lleva a concluir que va a ser necesario emplear métodos de análisis genéticos con el fin de obtener resultados más precisos.

7.2.3. Variables cuantitativas – cualitativas

En este apartado se va a utilizar la misma metodología que en el anterior, pero evaluando las asociaciones entre diferentes tipos de variables. Para poder llevar a cabo el análisis se ha segmentado cada una de las variables continuas en 5 intervalos uniformes, de manera que cada intervalo se pueda tratar como una variable discreta. Esto se ha realizado empleando los datos imputados por ambos métodos y aplicando posteriormente el script "contingenciaInt" con un nivel de significancia de 0.05.

De la misma manera que en el análisis de asociación de variables cualitativas, se ha obtenido un número muy elevado de asociaciones, en concreto, 88 con los datos imputados por el valor más frecuente y 82 con los datos imputados mediante IMPUTE.

Un ejemplo de asociaciones coincidentes mediante ambos tipos de datos es:

- Empleando los datos imputados por el valor más frecuente:

		DCTRsi_no	
		0	1
Edad_donante	[14, 27]	27	23
	(27, 39]	20	17
	(39, 49]	27	24
	(49, 58]	26	49
	(58, 82]	18	45

Pearson's Chi-squared test

data: .Table
X-squared = 13.8585, df = 4, p-value = 0.007761

		DCTR_otrDCTR		
		0	1	2
Edad_donante	[14, 27]	27	9	14
	(27, 39]	20	8	9
	(39, 49]	27	10	14
	(49, 58]	26	15	34
	(58, 82]	18	16	29

Pearson's Chi-squared test

data: .Table
X-squared = 15.6697, df = 8, p-value = 0.04736

- Empleando los datos imputados mediante IMPUTE:

		DCTRsi_no	
		0	1
Edad_donante	[14, 27]	30	23
	(27, 39]	19	18
	(39, 49]	29	20
	(49, 58]	22	48
	(58, 82]	18	45

Pearson's Chi-squared test

data: .Table
X-squared = 19.4085, df = 4, p-value = 0.0006532

		DCTR_otrDCTR		
		0	1	2
Edad_donante	[14, 27]	30	8	15
	(27, 39]	19	8	10
	(39, 49]	29	9	11
	(49, 58]	22	15	33
	(58, 82]	18	16	29

Pearson's Chi-squared test

data: .Table
X-squared = 21.0016, df = 8, p-value = 0.007143

7.3.Análisis genético

Durante los análisis anteriores se ha visto que las conclusiones obtenidas mediante los datos imputados por el valor más frecuente y mediante IMPUTE eran muy similares. Es por ello que para los posteriores análisis sólo se va a emplear uno de los dos ya que se estima que los resultados también serán semejantes.

Dado que IMPUTE es la herramienta empleada por los grupos de investigación para imputar, se ha optado por el grupo de datos imputados mediante dicho método.

El análisis genético se ha realizado mediante la herramienta R, cargando, en primer lugar, las variables genéticas para que se les pueda aplicar las técnicas de análisis adecuadas (punto 3.3 de Scripts).

7.3.1. Equilibrio de Hardy-Weinberg

Como se ha comentado en el apartado 3.4, un indicador para saber si ha ocurrido una mutación es determinar si la población se encuentra en equilibrio de Hardy-Weinberg. Por un lado, para un SNP, si la población correspondiente a los controles está en equilibrio indica que no hay ninguna relación de consanguinidad entre los individuos y las muestras están tomadas correctamente. Por otro lado, si la población de casos está en equilibrio, indica que el SNP no está asociado con la enfermedad.

Para saber si la población de estudio, dados los SNP's que se dispone, cumple el equilibrio de Hardy-Weinberg, se emplea el paquete SNPassoc de la herramienta R.

La población se divide en casos y control para poder realizar los análisis correctamente. Inicialmente, para cada SNP, la herramienta obtiene la frecuencia de los genotipos y la de los alelos que lo forman. Posteriormente, calcula la frecuencia genotípica esperada si se diera el equilibrio de Hardy-Weinberg siguiendo las fórmulas: $frec(AA) = p^2$, $frec(aa) = q^2$ para el caso de genotipos homocigotos y $frec(Aa) = 2 * p * q$ para los heterocigotos. Empleando el test de la χ^2 se comparan los valores obtenidos con los esperados, dando lugar a una tabla en la que cada SNP tiene el p-valor obtenido en el test.

En primer lugar se ha analizado la población de control, obteniendo que el SNP rs1800796 no está en equilibrio de Hardy-Weinberg (p-valor = 0.025). Como se trata de la muestra de control, puede deberse a posibles sesgos en la genotipificación, produciendo falsos resultados en los estudios de asociación. Es por ello que dicho SNP se elimina del conjunto de variables genéticas.

A continuación se muestra la tabla con los valores obtenidos:

SNP	HWE (p value)	flag
rs1036199	-	
rs10515746	-	
rs1143634	1.00	
rs12449782	0.3375	
rs1799750	10.000	
rs1799969	-	
rs1800471	-	
rs1800629	0.7171	
rs1800764	0.7099	
rs1800795	0.5262	
rs1800796	0.0250	<-
rs1800797	0.3970	
rs1800825	-	
rs1800871	0.1158	
rs1800872	0.1158	
rs1800896	0.0879	
rs1801275	0.4174	
rs2070874	0.6629	
rs2071231	-	
rs2107538	1.00	

SNP	HWE (p value)	flag
rs2234676	0.4050	
rs2243248	0.3777	
rs2430561	0.7110	
rs243865	0.8119	
rs301640	1.00	
rs3918226	0.0754	
rs41297579	0.5201	
rs419598	0.5342	
rs4311	1.00	
rs4586	1.00	
rs4696480	0.5802	
rs4986790	-	
rs4986791	-	
rs5186	0.8330	
rs5743708	-	
rs5749511	-	
rs699	0.4493	
rs699947	0.2650	
rs7830	1.00	
rs833061	0.3525	

Posteriormente, se ha analizado la población de casos, en la que el SNP rs1801275 no cumple el equilibrio de Hardy-Weinberg (p-valor = 0.0288). Dado que se trata de la muestra de casos, este resultado se tendrá en cuenta para mantener dicho SNP en el conjunto de variables del modelo predictivo. La tabla con los resultados para todos los SNP's es la siguiente:

SNP	HWE (p-value)	flag
rs1036199	0.7397	
rs10515746	0.0809	
rs1143634	1.00	
rs12449782	0.8695	
rs1799750	0.6286	
rs1799969	-	
rs1800471	-	
rs1800629	0.0887	
rs1800764	0.6311	
rs1800795	0.4870	
rs1800796	0.4463	
rs1800797	0.3865	
rs1800825	-	
rs1800871	1.00	
rs1800872	1.00	
rs1800896	0.4946	
rs1801275	0.0288	<-
rs2070874	0.1786	
rs2071231	-	
rs2107538	0.6056	

SNP	HWE (p-value)	flag
rs2234676	0.8422	
rs2243248	0.2859	
rs2430561	1.00	
rs243865	0.0703	
rs301640	1.00	
rs3918226	1.00	
rs41297579	1.00	
rs419598	0.8493	
rs4311	0.4166	
rs4586	0.8540	
rs4696480	0.1392	
rs4986790	-	
rs4986791	-	
rs5186	0.2777	
rs5743708	-	
rs5749511	0.3366	
rs699	0.2232	
rs699947	0.4194	
rs7830	0.5864	
rs833061	0.4194	

7.3.2. Análisis de haplotipos

Dado el elevado número de asociaciones entre SNP's obtenido durante el análisis de las variables cualitativas, se puede pensar que es debido a la existencia de desequilibrio de enlace entre ellos, anteriormente comentado. Así, como se ha dicho en el apartado 3.5, mediante el análisis de haplotipos se van a determinar los alelos que se trasmiten conjuntamente en una región del cromosoma y aquellos utilizados para la creación de los haplotipos.

Las zonas de los cromosomas que reúnen información similar son aquellas que se encuentran próximas y no se han visto afectadas por recombinaciones, por lo tanto, el primer paso es situar a cada SNP en su cromosoma y gen correspondiente ya que aquellos SNP's situados en el mismo gen y que no hayan sufrido entrecruzamientos sugieren que codifican la misma información.

En la tabla siguiente se indica el gen al que pertenece cada SNP, obtenido mediante la aplicación NCBI:

SNP	CROM.	GEN
rs1800872	1	IL10
rs1800871	1	IL10
rs1800896	1	IL10
rs699	1	AGT
rs1143634	2	IL-1b
rs2234676	2	IL1RN
rs419598	2	IL1RN
rs5186	3	AGTR1
rs4696480	4	TL-R2
rs5743708	4	TL-R2
rs2243248	5	IL4
rs2070874	5	IL4
rs41297579	5	HAVCR1
rs1036199	5	HAVCR2
rs10515746	5	HAVCR2
rs1800629	6	TNF
rs699947	6	VEGFA
rs833061	6	VEGFA
rs3918226	7	NOS 3
rs7830	7	NOS 3

SNP	CROM.	GEN
rs1800795	7	IL 6
rs1800797	7	IL 6
rs4986790	9	TLR 4
rs4986791	9	TLR 4
rs2071231	11	MMP 1
rs1799750	11	MMP 1
rs2430561	12	IFNG
rs301640	13	MMP3
rs1801275	16	IL4R
rs243865	16	MMP2
rs4586	17	CCL2
rs1800825	17	CCL5
rs2107538	17	CCL5
rs1800764	17	ACE
rs4311	17	ACE
rs12449782	17	ACE
rs1799969	19	ICAM1
rs1800471	19	TGFB1
rs5749511	22	TIMP-3

7.3.2.1 Análisis mediante Haploview

Para la estimación de los haplotipos se emplea la herramienta Haploview (Broad Institute 2010). Para ello, se necesita, además de conocer el gen al que pertenece cada SNP, cómo están genotipados. Mediante la página web del proyecto HapMap se obtiene dicha información, agrupando, en un mismo fichero, los datos correspondientes a los SNP's que pertenecen a un mismo cromosoma. No obstante, no se ha podido obtener el genotipado para todos los SNP's ya que el proyecto HapMap está en desarrollo y durante el periodo de elaboración de este trabajo no disponía de la información completa.

Una vez creados los ficheros para cada cromosoma, se pasan a Haploview, que agrupa de forma gráfica los SNP's en genes, permitiendo ver si presentan desequilibrio de enlace, las frecuencias de los haplotipos y si hay algún SNP que no se emplea en la creación del haplotipo.

Los valores que muestra son los siguientes:

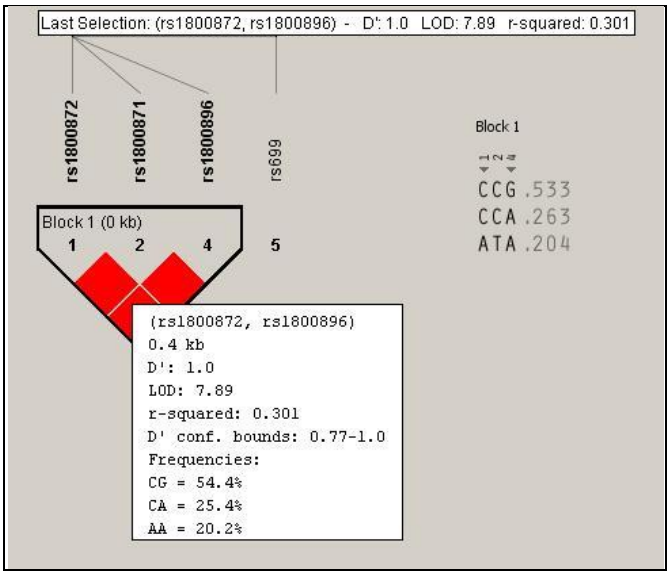
- D': representa el grado de desequilibrio de enlace (LD) entre un par de SNP's; cuanto más cercano esté a 1, mayor será el LD entre ellos.
- LOD: indica el grado de fiabilidad del valor de D'; así, si $LOD \geq 2$el valor de D' será seguro.
- r^2: es 1 si los SNP pertenecen a la misma rama y no se han visto modificados por la recombinación. Se trata del coeficiente de correlación.

Además, al realizar el análisis, los parámetros señalados aparecen mediante los siguientes colores:

	D' < 1	D' = 1
LOD < 2	Blanco	Azul
LOD ≥ 2	Rosa	Rojo

A continuación se muestran los resultados obtenidos:

- Para el cromosoma 1:



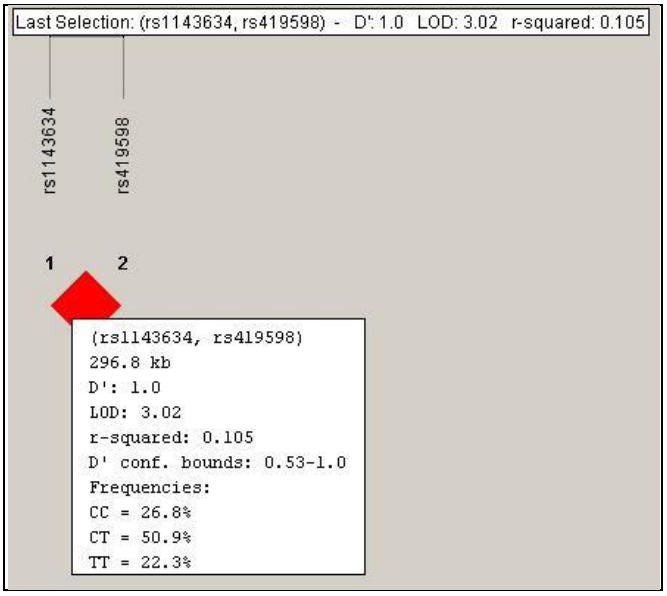
Se han dividido los SNP's en dos bloques, ya que los 3 primeros pertenecen a un mismo gen, mientras que rs699 no. Tal y como queda reflejado en la imagen, los 3 primeros SNP's presentan desequilibrio de enlace, obteniendo 3 haplotipos:

  - CCG, con una frecuencia de aparición de 53,3%.
  - CCA, con una frecuencia de 26,3%.
  - ATA, con frecuencia de aparición de 20,4%.

Además, como SNP's significativos para la creación de los haplotipos han resultado el primero (rs1800872) y el cuarto[7] (rs1800896), presentando, además, un coeficiente de correlación de 1 el par rs1800872-rs1800871. Por ello, rs1800871 no se tendrá en cuenta para análisis posteriores.

- Para el cromosoma 2:

[7] En nuestra muestra el SNP rs1800896 ocupa la tercera posición en el gen, pero en la realidad entre rs1800871 y rs1800896, se encuentra el SNP rs1800894, que no entra en nuestro estudio, pero que Haploview contabiliza igualmente.

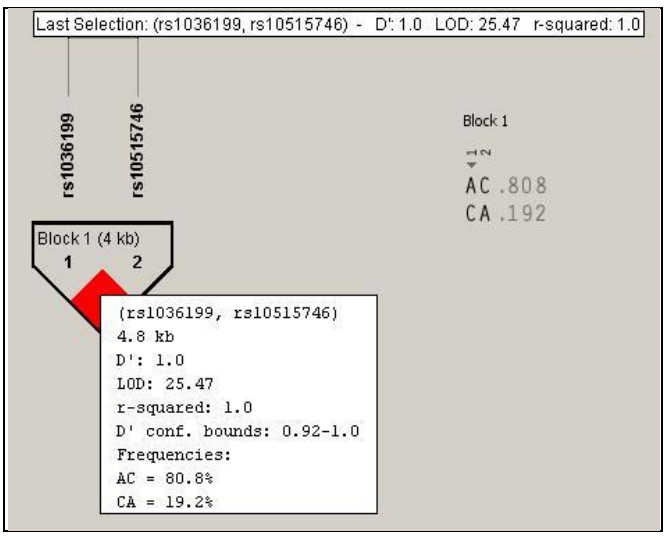


Se tienen 3 SNP's repartidos en 2 genes. En este caso, el genotipado para el SNP rs2234676 no se encontraba disponible en la página web del proyecto HapMap y sólo se ha realizado el análisis a los 2 SNP's restantes. Se ha encontrado que presentan desequilibrio de enlace a pesar de pertenecer a genes diferentes y ambos son necesarios para la creación de los haplotipos correspondientes.

- Para el cromosoma 3: Dado que sólo hay un SNP perteneciente a dicho cromosoma (rs5186), este cromosoma queda excluido del análisis de haplotipos.

- Para el cromosoma 4: En la muestra empleada para este trabajo hay 2 SNP's pertenecientes a dicho cromosoma, pero ninguno de los 2 tenían disponible el genotipado en la página web del proyecto HapMap.

- Para el cromosoma 5: A excepción del SNP rs41297579, el resto sí se han podido analizar mediante Haploview. El primer par de SNP's, rs2243248 y rs2070874, pertenecientes al mismo gen, no presentan desequilibrio de enlace; al contrario que los 2 SNP's siguientes:

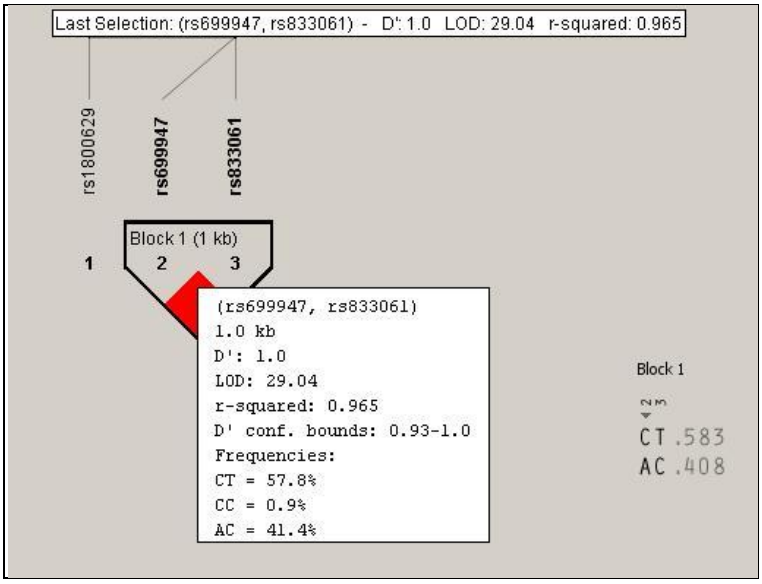


Como se puede ver, el par de SNP's presenta desequilibrio de enlace, obteniendo los haplotipos:

- AC, con una frecuencia de aparición de 80,8%.
- CA, con una frecuencia de 19,2%.

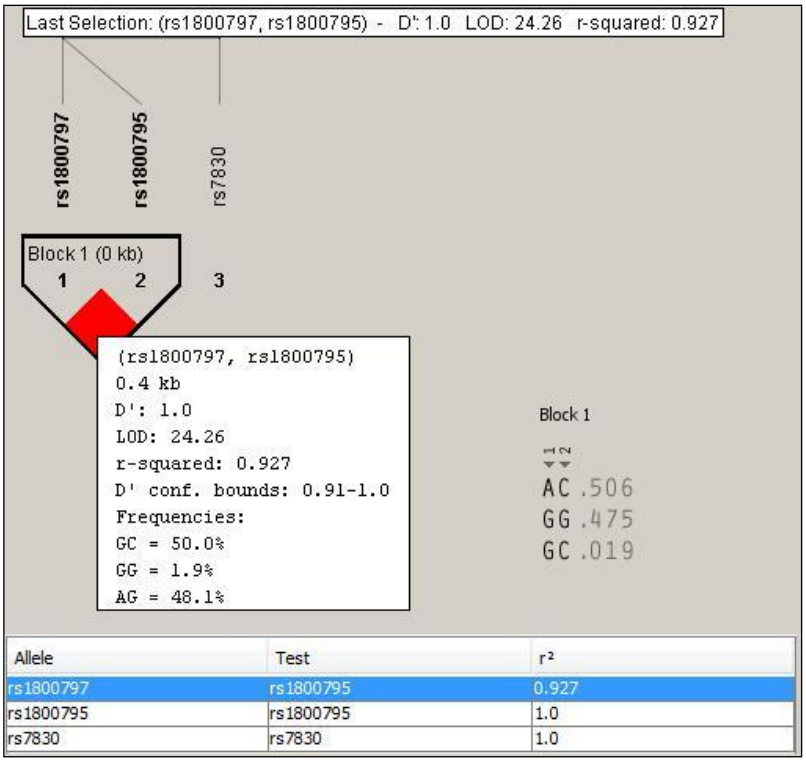
Para la formación de dichos haplotipos sólo es significativo el SNP rs1036199, por lo tanto, se prescindirá de rs10515746 para posteriores análisis.

- Para el cromosoma 6:



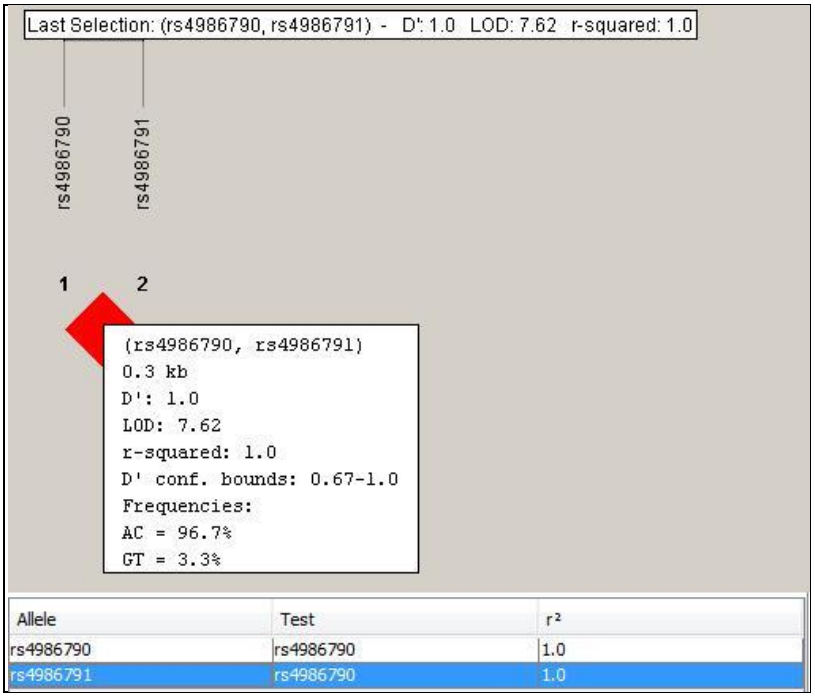
Los 3 SNP's de la muestra de estudio se han podido genotipar. Haploview los ha dividido en 2 bloques, ya que el primero pertenece a un gen diferente del de los 2 restantes. Éstos, presentan desequilibrio de enlace, formando los correspondientes haplotipos en los que el SNP rs833061 no resulta relevante.

- Para el cromosoma 7: Se han analizado todos los SNP's menos rs391826 que no estaba genotipado.



Como se puede ver, se han separado los SNP's en 2 bloques, en función del gen al que pertenecen. Los 2 primeros pertenecen a un mismo gen, formando los haplotipos correspondientes. Por otro lado, el coeficiente de correlación sugiere que no se han visto afectados por la recombinación (r^2=0.927), siendo rs1800795 el *representante*. Así pues, se prescindirá de rs1800797.

- Para el cromosoma 9:



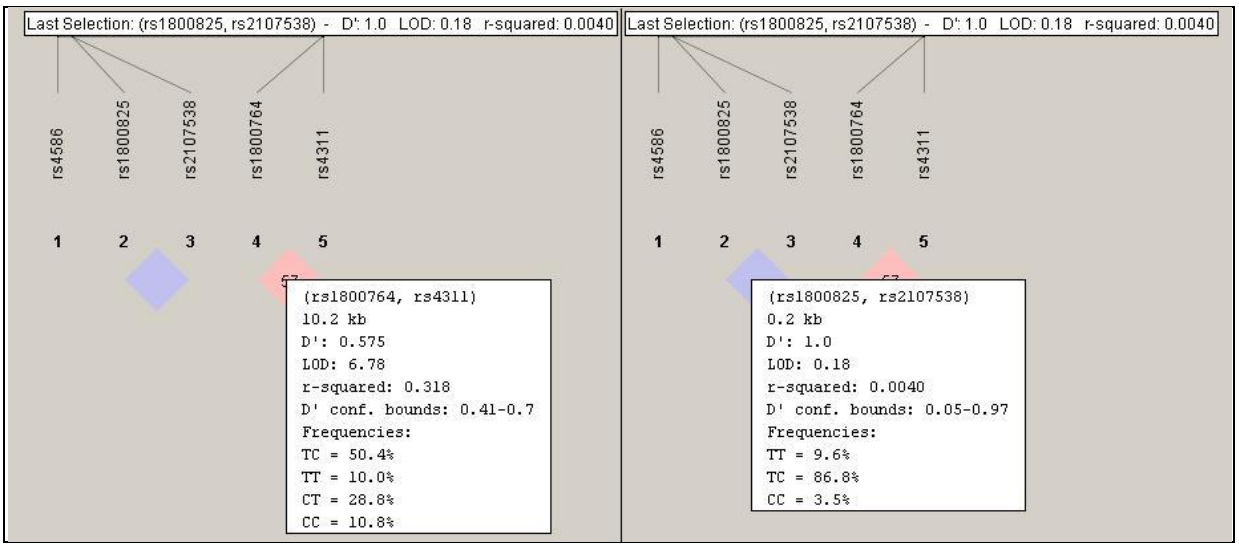
Tras analizar los 2 SNP's pertenecientes al cromosoma 9, se ha obtenido que ambos presentan desequilibrio de enlace, además de estar fuertemente correlacionados. Al ser rs4986790 el *representante* de dicho par, rs4986791 no se tendrá en cuenta para posteriores análisis.

- Para el cromosoma 11: De los 2 SNP's pertenecientes a dicho cromosoma, únicamente uno estaba genotipado, por lo tanto, no se ha podido realizar el análisis mediante Haploview.

- Para el cromosoma 12 y 13: en la muestra de estudio, ambos cromosomas contienen 1 SNP respectivamente y, por lo tanto, se excluyen del análisis de haplotipos.

- Para el cromosoma 16: Tras analizar los 2 SNP's pertenecientes al cromosoma 16, situados en diferentes genes, se ha obtenido que no hay ningún tipo de asociación entre ellos.

- Para el cromosoma 17: Excepto para el SNP rs12449782, se ha podido realizar el análisis mediante Haploview como se muestra a continuación:



Haploview ha agrupado los SNP's en función del gen al que pertenecen, obteniendo que ningún SNP presenta desequilibrio de enlace y ninguno de ellos presenta una correlación significativa.

- Para el cromosoma 19: uno de los 2 SNP's de los que se dispone no presentaba genotipado en la página web del proyecto HapMap y no se ha realizado el análisis mediante Haploview.

- Para el cromosoma 22: en nuestra muestra únicamente pertenece a dicho cromosoma el SNP rs5749511 y, por lo tanto, se excluye del análisis de haplotipos.

Así, durante la realización del análisis mediante Haploview, se han encontrado SNP's que pertenecían a la misma región del gen y, al no sufrir modificaciones durante la transmisión de padres a hijos, la información era similar a la de sus 'vecinos', pudiendo prescindir de ellos. Estos SNP's eliminados son:

- rs1800871, rs10515746, rs833061, rs1800797, rs4986791.

7.3.2.2. Análisis mediante R

Como una segunda etapa del análisis de haplotipos, se realiza un análisis de asociación de los mismos con la enfermedad mediante el paquete haplo.stats de la herramienta R. Este estudio se realiza a través de una función *logit-link*, es decir, a través de un modelo de regresión logística tal y como se ha explicado en el punto 4.7.

La asociación se va a estudiar para la variable dependiente DCTRsi_no indicadora de la ausencia/presencia de rechazo ya que el método empleado únicamente permite estudios de caso-control o variables dependientes de tipo cuantitativo con una distribución normal, con lo que queda excluida la variable DCTR_otrDCTR, de tipo discreto con 3 posibles categorías.

Mediante la herramienta R se indica que se trata de un estudio caso-control, así como la frecuencia mínima de aparición en la muestra para considerar minoritarios los haplotipos formados (en el análisis, dicha frecuencia se ha establecido en 5%).

A continuación se indican los resultados obtenidos:

- Para el cromosoma 1: Como se ha visto mediante el análisis del Haploview, los SNP's rs1800872 y rs1800896 tienen desequilibrio de enlace. No obstante, el grado de correlación no es muy elevado (r^2 = 0,301). Para poder estudiar dichos SNP's con más detalle, se ha realizado un análisis de asociación con la enfermedad (variable dependiente DCTRsi_no) con los haplotipos de los SNP's que se tienen para el cromosoma 1.

  En los resultados que se muestran a continuación, ha aparecido significativo el haplotipo 6 (C G C), presentando una asociación inversa con la enfermedad:

```
Coefficients:
                coef    se  t.stat   pval
(Intercept)  0.58555 0.338  1.7314 0.0845
geno.1      -0.11687 0.334 -0.3502 0.7265
geno.2      -0.20288 0.288 -0.7034 0.4824
geno.4       0.00695 0.299  0.0233 0.9815
geno.5      -0.06958 0.318 -0.2187 0.8270
geno.6      -0.72581 0.351 -2.0657 0.0398

Haplotypes:
           rs1800872 rs1800896 rs699 hap.freq
geno.1             A         A    C    0.108
geno.2             A         A    T    0.172
geno.4             C         A    C    0.144
geno.5             C         A    T    0.180
geno.6             C         G    C    0.136
haplo.base         C         G    T    0.261
```

- Para el cromosoma 2: Para los SNP's rs1143634 y rs419598 se obtuvo mediante Haploview que presentaban desequilibrio de enlace pero con una correlación débil (r^2 = 0.105). Tras el estudio realizado mediante R se tiene, como se puede ver en la imagen siguiente, que el haplotipo de referencia[8] (C G T) tiene significancia positiva con la enfermedad:

```
Coefficients:
              coef    se t.stat   pval
(Intercept)  0.489 0.216  2.262 0.0245
geno2.1     -0.280 0.217 -1.292 0.1975
geno2.8     -0.277 0.284 -0.976 0.3300
geno2.rare   0.143 0.470  0.304 0.7611

Haplotypes:
           rs1143634 rs2234676 rs419598 hap.freq
geno2.1            C         A        C   0.2647
geno2.8            T         G        T   0.1586
geno2.rare         *         *        *   0.0553
haplo.base         C         G        T   0.5215
```

- Para el cromosoma 4: Como se puede ver en la siguiente imagen, el haplotipo de referencia (A G) tiene significancia positiva con la enfermedad. Por otro lado, el SNP rs5743708 presenta el mismo alelo (G) tanto para el haplotipo significativo como para el haplotipo 4, que no lo es. Por lo tanto, se elimina para posteriores análisis.

```
Coefficients:
              coef    se t.stat  pval
(Intercept)  0.455 0.200  2.270 0.024
geno4.4     -0.194 0.171 -1.137 0.257
geno4.rare  -1.085 1.236 -0.878 0.381

Haplotypes:
           rs4696480 rs5743708 hap.freq
geno4.4            T         G  0.45221
geno4.rare         *         *  0.00551
haplo.base         A         G  0.54228
```

- Para el cromosoma 5: Mediante Haploview, los SNP's rs2243248 y rs2070874 no presentaban desequilibrio de enlace y rs41297579 no se pudo analizar. Con R se han analizado los haplotipos de los SNP's del cromosoma 5 y su asociación con la variable dependiente DCTRsi_no:

[8] Haplotipo de referencia es aquel con mayor frecuencia de aparición. En la tabla de coeficientes se le llama "Intercept", mientras que en la tabla de haplotipos se le llama "haplo.base".

```
Coefficients:
               coef    se  t.stat  pval
(Intercept)  0.1383 0.206  0.6717 0.502
geno5.3      0.1010 0.371  0.2724 0.786
geno5.6     -0.0295 0.296 -0.0996 0.921
geno5.9      0.1225 0.369  0.3318 0.740
geno5.11     0.4427 0.417  1.0623 0.289
geno5.rare   0.2331 0.430  0.5424 0.588

Haplotypes:
           rs2243248 rs2070874 rs41297579 rs1036199 hap.freq
geno5.3            G         C          G         A   0.0613
geno5.6            T         C          A         A   0.1306
geno5.9            T         C          G         C   0.0955
geno5.11           T         T          G         A   0.0764
geno5.rare         *         *          *         *   0.0758
haplo.base         T         C          G         A   0.5604
```

En la imagen anterior se ha visto no hay ningún haplotipo significativo.

- Para el cromosoma 6: Aunque mediante Haploview se estudiaron los haplotipos de los SNP's del cromosoma 6, también se ha analizado su asociación con la variable DCTRsi_no mediante R. Tal y como se muestra a continuación, no hay ningún haplotipo asociado con la variable dependiente:

```
Coefficients:
              coef    se t.stat  pval
(Intercept) 0.0705 0.229  0.308 0.758
geno6.1     0.1169 0.421  0.278 0.781
geno6.2     0.0195 0.375  0.052 0.959
geno6.3     0.2253 0.199  1.134 0.258

Haplotypes:
           rs1800629 rs699947 hap.freq
geno6.1            A        A   0.0636
geno6.2            A        C   0.0816
geno6.3            G        A   0.3978
haplo.base         G        C   0.4570
```

- Para el cromosoma 7: A continuación queda reflejado que no hay ningún haplotipo del cromosoma 7 asociado con la variable dependiente DCTRsi_no:

```
Coefficients:
              coef    se t.stat  pval
(Intercept)  0.402 0.277  1.453 0.147
geno7.1      0.576 0.400  1.438 0.152
geno7.2     -0.281 0.280 -1.002 0.317
geno7.3     -0.340 0.278 -1.221 0.223
geno7.8     -0.407 0.423 -0.961 0.337
geno7.rare   0.929 0.641  1.450 0.148

Haplotypes:
           rs3918226 rs7830 rs1800795 hap.freq
geno7.1            C      A         C   0.1036
geno7.2            C      A         G   0.2195
geno7.3            C      C         C   0.2026
geno7.8            T      C         G   0.0632
geno7.rare         *      *         *   0.0416
haplo.base         C      C         G   0.3694
```

- Para el cromosoma 11: Para este cromosoma, los haplotipos formados tampoco presentan ninguna asociación significativa con la enfermedad:

```
Coefficients:
              coef    se t.stat  pval
(Intercept) 0.216 0.214 1.0074 0.315
geno11.4    0.051 0.178 0.2874 0.774
geno11.rare 0.045 0.778 0.0579 0.954

Haplotypes:
            rs2071231 rs1799750 hap.freq
geno11.4            T         G   0.4866
geno11.rare         *         *   0.0129
haplo.base          T         -   0.5005
```

No obstante, el SNP rs2071231 presenta el mismo alelo (T) en aproximadamente el 98% de las muestras, por lo que se supone que no aporta información y se descarta del estudio.

- Para el cromosoma 16: Mediante Haploview se obtuvo que los 2 SNP's pertenecientes a dicho cromosoma no presentaban desequilibrio de enlace. Tras realizar el análisis de asociación mediante R se ha obtenido que uno de los haplotipos formados por los SNP's rs1801275 y rs248565 (en concreto, el haplotipo de referencia A C) está asociado con la variable dependiente DCTRsi_no. A continuación se muestran los resultados:

```
Coefficients:
               coef    se  t.stat   pval
(Intercept)  0.5034 0.197  2.5499 0.0113
geno16.2    -0.2936 0.247 -1.1887 0.2356
geno16.3    -0.3630 0.262 -1.3878 0.1663
geno16.rare  0.0556 0.631  0.0881 0.9299

Haplotypes:
            rs1801275 rs243865 hap.freq
geno16.2            A        T   0.2077
geno16.3            G        C   0.1599
geno16.rare         *        *   0.0349
haplo.base          A        C   0.5974
```

- Para el cromosoma 17: En la muestra de estudio se tienen 6 SNP's pertenecientes al cromosoma 17. En un primer lugar se analizaron los haplotipos de los 6 SNP's pero se obtuvo que el conjunto de los haplotipos menos frecuentes (agrupados como $geno_i$.rare) aparecían en alrededor del 30% de las muestras, mientras que 5 haplotipos suponían el 70% restante. Es por ello que se ha decidido separar los SNP's en 2 grupos y analizarlos por separado.

    - Para el primer grupo se tienen los siguientes resultados:

```
Coefficients:
               coef    se t.stat  pval
(Intercept)   0.442 0.220  2.004 0.046
geno171.2    -0.320 0.228 -1.401 0.162
geno171.3     0.185 0.488  0.379 0.705
geno171.7    -0.194 0.309 -0.628 0.530
geno171.rare  0.779 0.686  1.136 0.257

Haplotypes:
            rs4586 rs1800825 rs2107538 hap.freq
geno171.2        C         T         C   0.2721
geno171.3        C         T         T   0.0608
geno171.7        T         T         T   0.1341
geno171.rare     *         *         *   0.0221
haplo.base       T         T         C   0.5110
```

    El haplotipo de referencia (T T C) presenta asociación significativa con la variable dependiente DCTRsi_no. Por otro lado, el SNP rs1800825 presenta el mismo alelo (T) tanto para los haplotipos que no están asociados con la variable dependiente como para el que sí que lo está. Por lo tanto, rs1800825 se elimina del estudio.

    - Para el segundo grupo se obtienen los siguientes resultados:

```
Coefficients:
                coef    se t.stat  pval
(Intercept)   0.3455 0.257  1.344 0.180
geno172.5    -0.0528 0.203 -0.259 0.795
geno172.8    -0.2247 0.284 -0.792 0.429
geno172.rare  0.0696 0.284  0.245 0.807

Haplotypes:
             rs1800764 rs4311 rs12449782 hap.freq
geno172.5            T      C          A    0.361
geno172.8            T      T          G    0.131
geno172.rare         *      *          *    0.134
haplo.base           C      T          G    0.373
```

En este caso no se han obtenido haplotipos asociados con la enfermedad.

- Para el cromosoma 19: Como se puede apreciar a continuación, el haplotipo de referencia G G tiene asociación positiva con la variable dependiente DCTRsi_no:

```
Coefficients:
              coef    se t.stat   pval
(Intercept)  0.283 0.142   2.00 0.0468
geno19.2    -0.620 0.349  -1.78 0.0762
geno19.3     0.699 0.416   1.68 0.0944

Haplotypes:
           rs1799969 rs1800471 hap.freq
geno19.2           A         G   0.0735
geno19.3           G         C   0.0588
haplo.base         G         G   0.8676
```

Durante el análisis mediante R, se han encontrado SNP's que presentan el mismo alelo tanto para haplotipos asociados con la variable dependiente DCTRsi_no, como para haplotipos que no presentan asociación. Además, se han encontrado SNP's que en más del 90% de las muestras contribuyen en los haplotipos con el mismo alelo. Los SNP's con dichas características y que no se van a tener en cuenta en análisis posteriores son:

- rs5743708, rs2071231, rs1800825

Por otro lado, los SNP's que no participan en los haplotipos significativos se mantienen en el conjunto de variables para ver su asociación con la enfermedad mediante interacciones con otros SNP's aunque no pertenezcan al mismo cromosoma.

7.3.3. Análisis de asociación univariante

Una vez realizada una reducción de SNP's mediante el análisis de haplotipos, se realiza un análisis de asociación univariante.

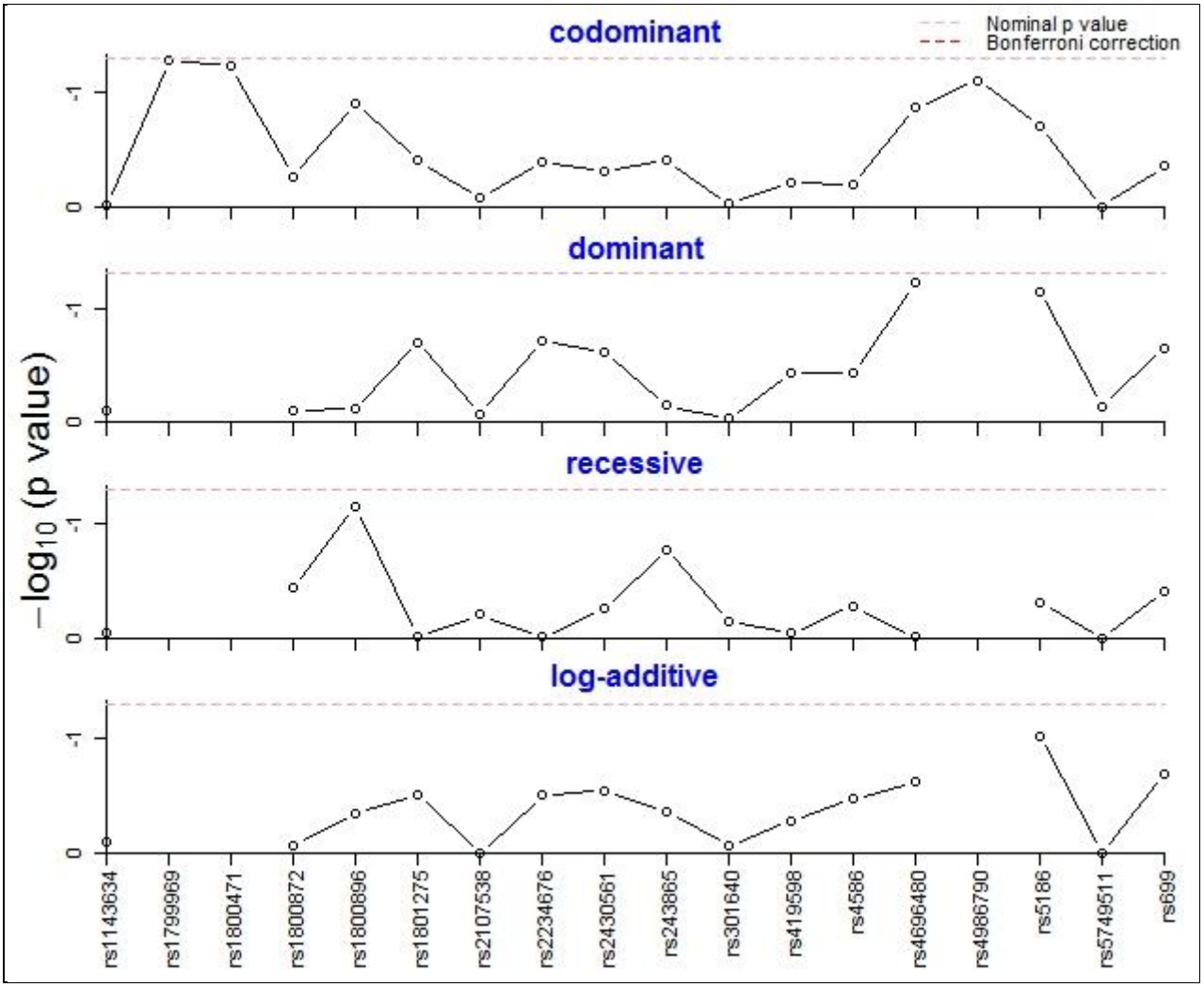
Este análisis se realiza para todos los SNP's que forman haplotipos asociados con la variable dependiente. Es necesario ver si dicha asociación significativa es producida por un único SNP o por la combinación de todos los SNP's implicados en el haplotipo ya que, cuando aparece un cambio genético en una zona del cromosoma, ésta y las regiones cercanas se heredan como un grupo.

Anteriormente se han tratado los SNP como variables estadísticas, junto con las variables clínicas, con unas categorías posibles, construyendo una tabla de contingencia y evaluando el nivel de asociación mediante un test de χ^2.

A continuación se estudian las asociaciones de los SNP's con las variables dependientes DCTRsi_no y DCTR_otrDCTR mediante regresión logística, teniendo en cuenta, además, las posibles formas de herencia de los alelos que forman los SNP's. Así, se estudia que tal vez un alelo y la forma en que se haya heredado sean los causantes de que el SNP esté asociado con la variable dependiente.

Los SNP's que se van a incluir en este estudio son aquellos que participan en haplotipos asociados con la variable DCTRsi_no, así como aquellos que no se han podido evaluar en el punto 7.3.2 al ser los únicos pertenecientes a un cromosoma.

Para cada SNP, el resultado obtenido es el p-valor que indica si se acepta la hipótesis nula (el SNP no está asociado con la enfermedad) o se rechaza (el SNP sí está asociado). A continuación, se muestran dichos valores para cada modelo de herencia en escala ln.



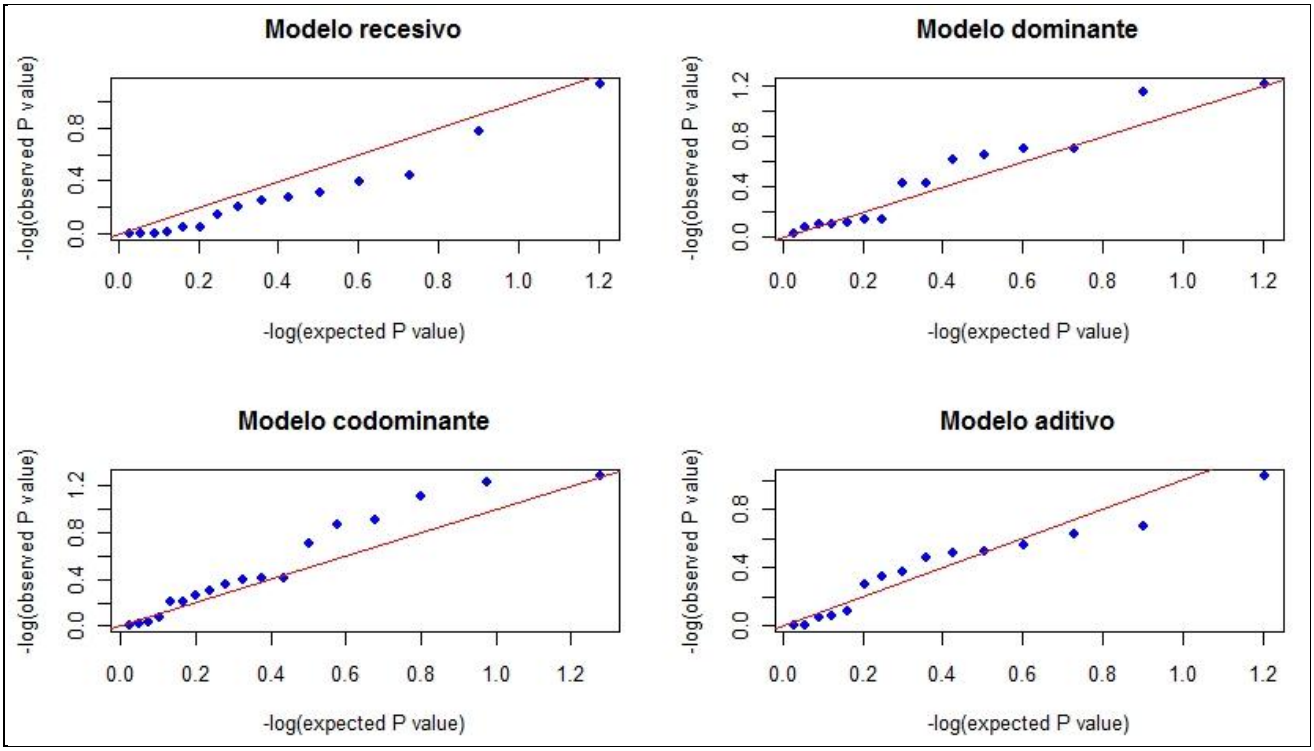
Dado que tenemos un modelo de regresión logística para cada modelo de herencia, hay que decantarse por uno de los cuatro.

Para ello, se realiza una comparativa para cada modelo de herencia entre:

- Los p-valor que se obtendrían para un análisis de regresión logística realizado con SNP's que siguen el modelo de herencia ideal.
- Los p-valor obtenidos con el análisis de regresión logística ajustando los SNP's al modelo de herencia, es decir, los resultados obtenidos.

Como se puede ver a continuación, el modelo de herencia que mejor se ajusta al ideal, es el de herencia aditiva. Para dicho modelo no se ha obtenido ningún SNP asociado de forma significativa con la enfermedad, lo cual indica que es necesaria la combinación de éstos para la obtención de los haplotipos asociados con la enfermedad.

A continuación se muestran las gráficas con las comparativas mencionadas:



7.3.4. Análisis por pares de interacciones

El objetivo de este apartado es determinar si existen interacciones entre SNP's, incluso entre aquellos pertenecientes a diferentes cromosomas, que estén asociadas de forma significativa con la variable dependiente DCTRsi_no o con la variable dependiente DCTR_otrDCTR. Así, se incluyen todos los SNP's que no se hayan eliminado del análisis de haplotipos, aunque no formen haplotipos asociados con la variable dependiente.

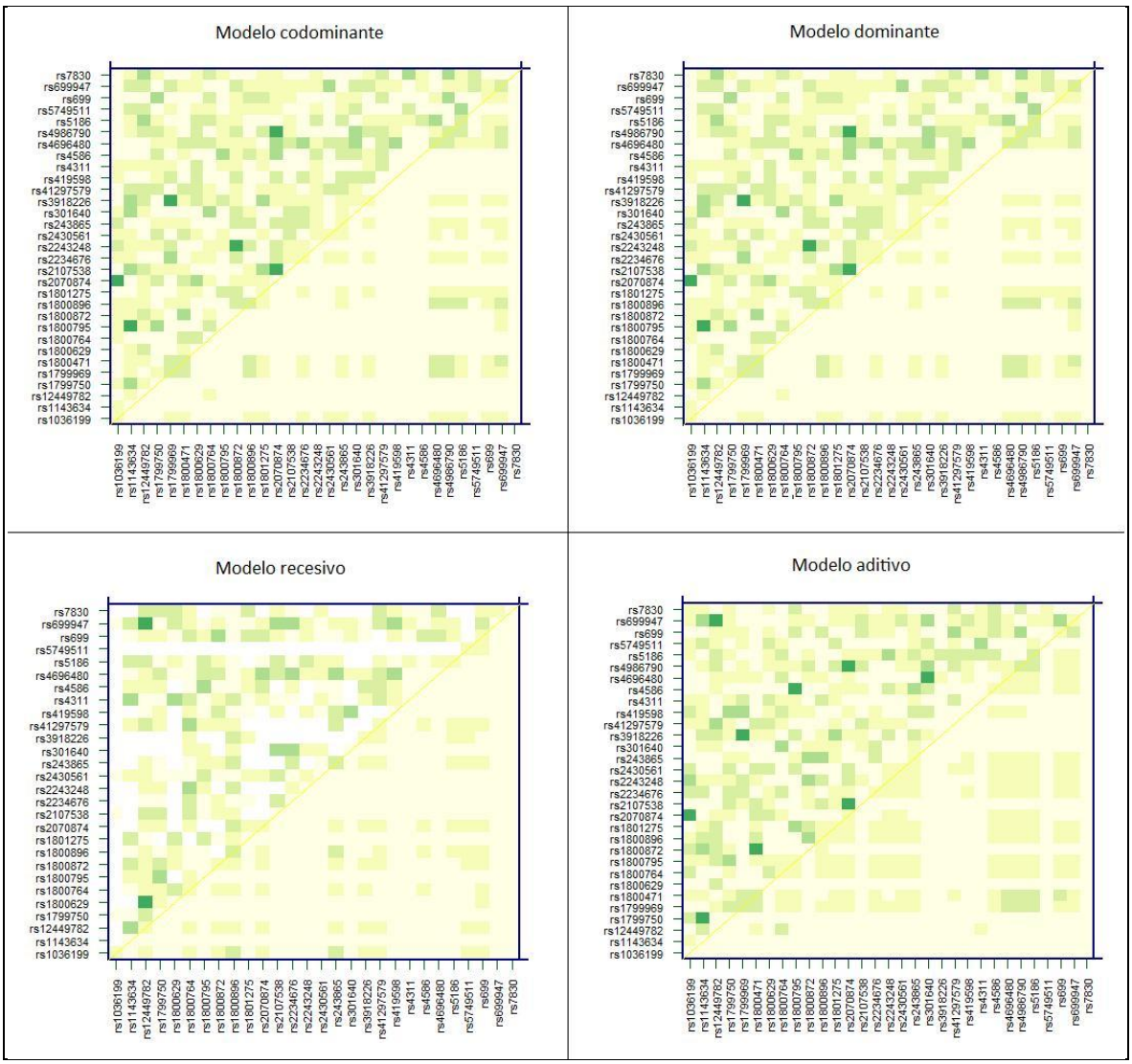
Para ello, se utiliza el paquete SNPassoc de la herramienta R tal y como se ha explicado en el apartado 4.8.2. Los resultados obtenidos se muestran en forma de matriz, donde las celdas corresponden a los p-valor obtenidos para cada par de SNP's evaluados. Si la celda está situada en la diagonal de dicha matriz, corresponde a la asociación del SNP con la variable dependiente de forma aislada; si la celda corresponde al triángulo superior, se trata de la asociación de la interacción del par de SNP's; y si la celda está situada en el triángulo inferior, se trata de la asociación con la variable dependiente del SNP de la fila i ajustado por el SNP de la columna j.

Para ver de una manera más visual dichos resultados, se muestran en una matriz en la que los p-valor se representan mediante colores como sigue:

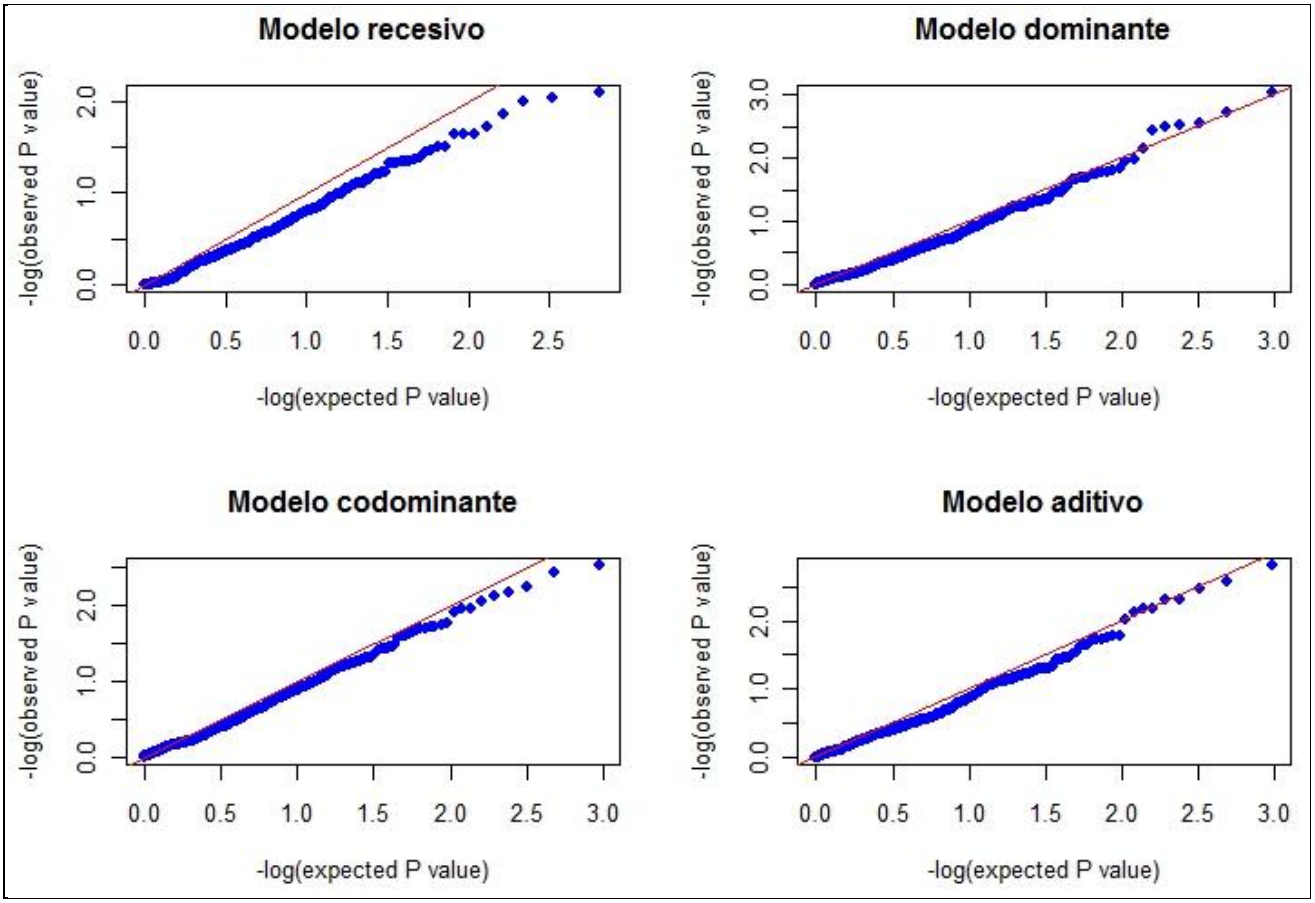
- Rosado: p-valor entre 1 y 0.3.
- Amarillo: p-valor entre 0.3 y 0.1.
- Verde claro: p-valor entre 0.1 y 0.05.

- Verde: p-valor entre 0.05 y 0.025, la interacción del par de SNP's está asociada con la enfermedad.
- Verde oscuro/muy oscuro: p-valor por debajo de 0.025, la interacción del par de SNP's está fuertemente asociada con la variable dependiente.

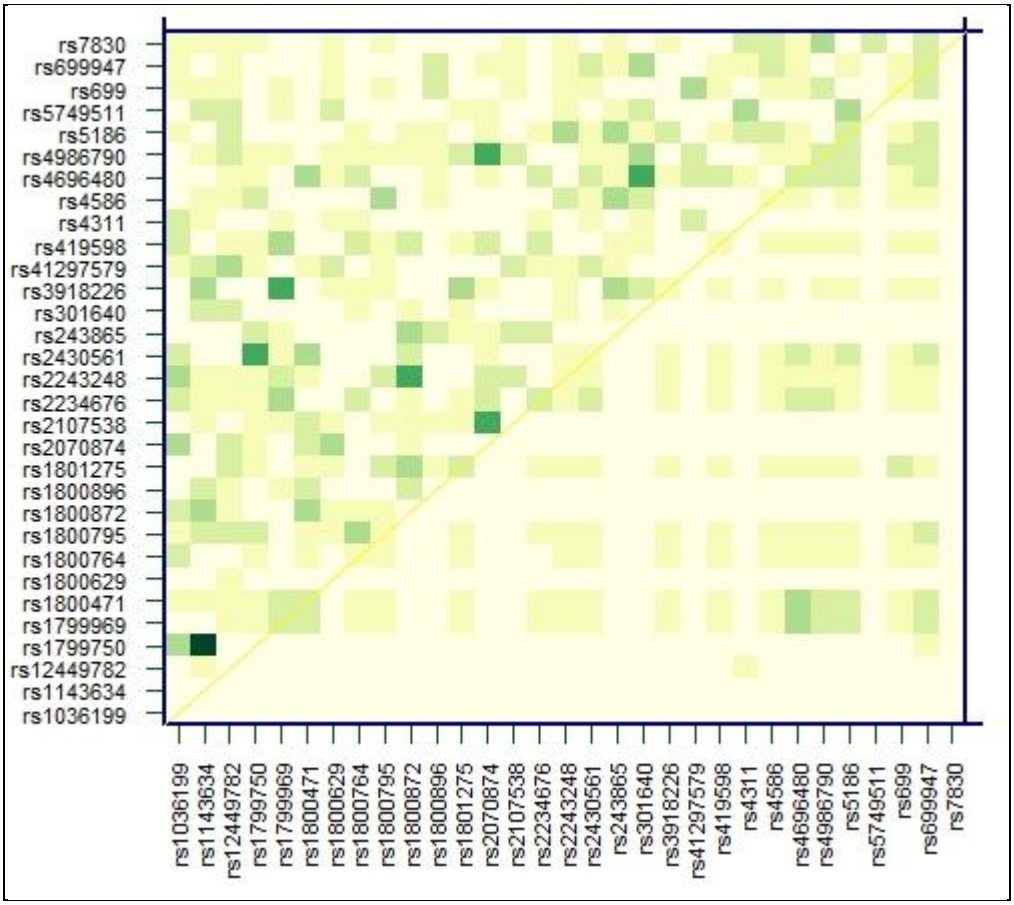
En primer lugar, se han analizado las interacciones para la variable dependiente DCTRsi_no para cada modelo de herencia obteniendo las siguientes matrices:



Al igual que para el análisis de asociación univariante, se debe elegir el modelo de herencia que mejor se ajusta al ideal correspondiente. Para ello, para cada modelo de herencia, se ha realizado la comparativa entre los p-valor obtenidos y los esperados suponiendo un modelo de herencia ideal. Las gráficas resultantes son las siguientes:



Tras las comparativas realizadas, se puede decir que los p-valor obtenidos mediante el modelo dominante son los que mejor se ajustan al modelo ideal correspondiente. A continuación se muestra con más detalle la matriz:

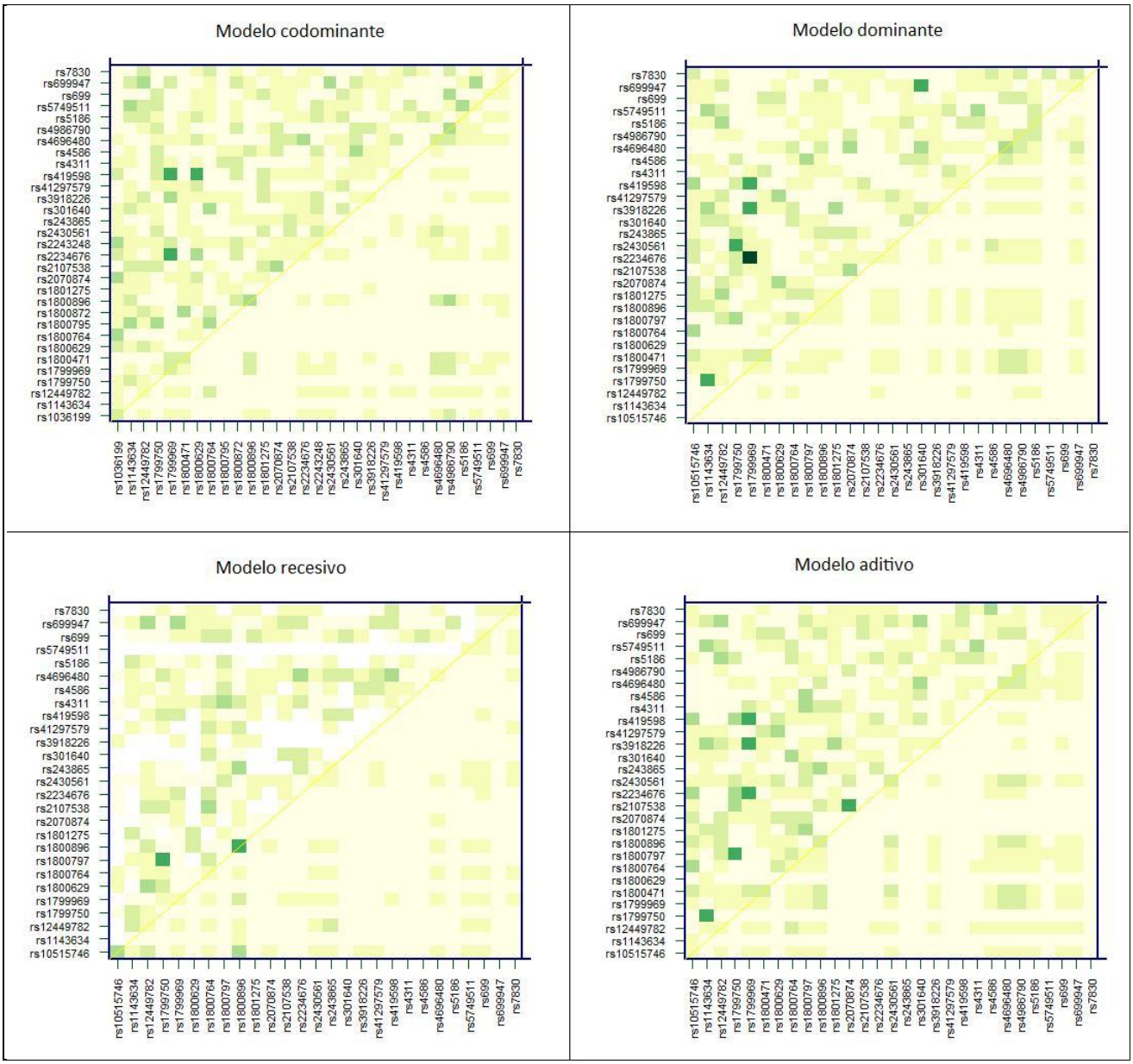


A simple vista, se distinguen 7 celdas en el triángulo superior de la matriz de color verde oscuro y, por tanto, 7 interacciones entre pares de SNP's asociadas de forma significativa con la variable dependiente DCTRsi_no. Para ver con más exactitud qué pares de interacciones han resultado significativas, se ha creado el script 'int'. Dada la matriz con los p-valor obtenidos, el nivel de significancia por debajo del cual se van a considerar significativos los p-valor y los nombres de los SNP's empleados, se obtienen los pares de interacciones significativos y el p-valor asociado. Un fragmento del mismo se indica a continuación:

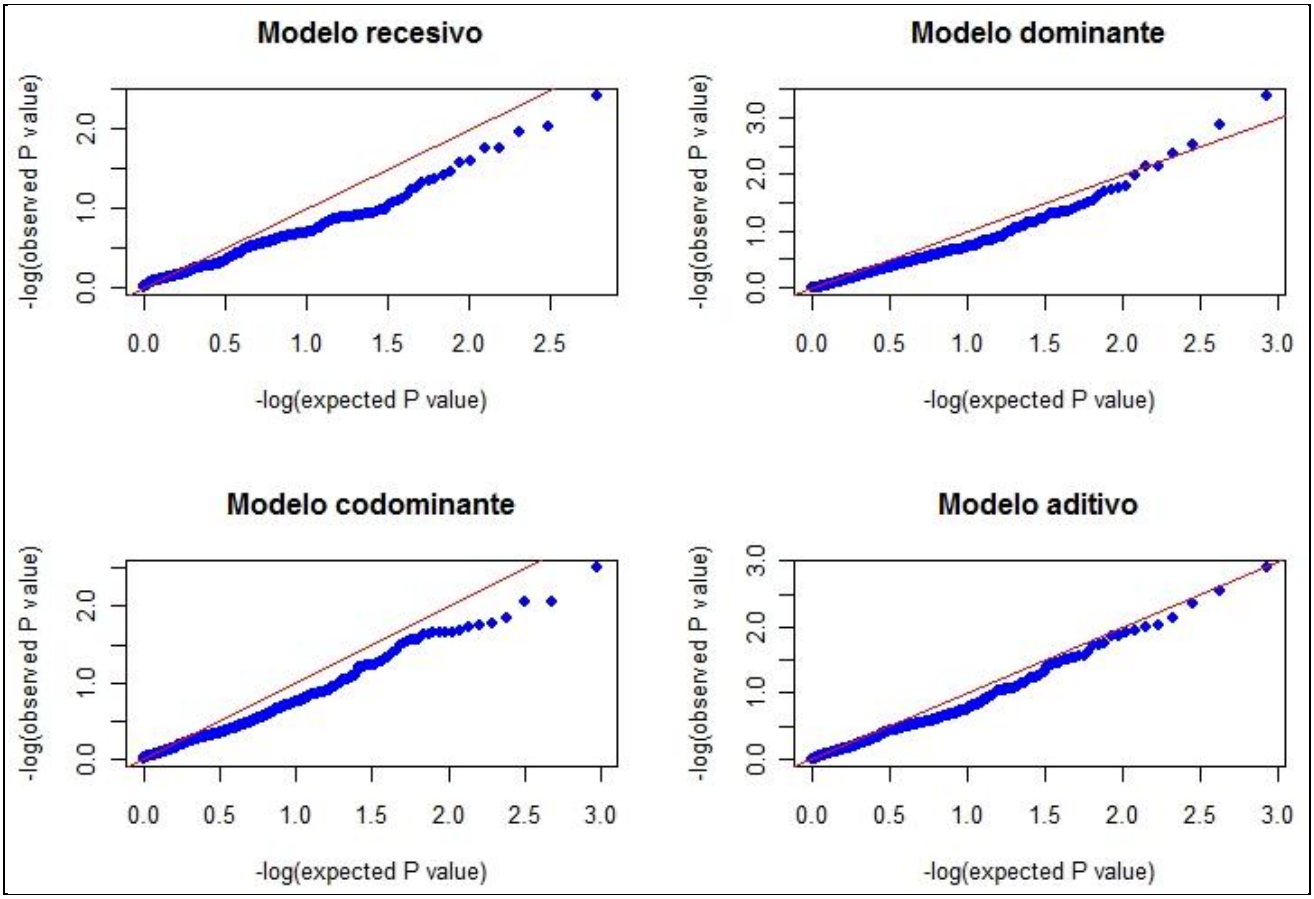
```
for (i in 1: nrow(inter)){
      for(j in 1:ncol(inter)){
            if((!is.na(inter[i,j])) & (inter[i,j] <= sig)){
            cont <- cont+1
            mensaje[cont] <- paste(labels[i], labels[j])
            print(mensaje[cont])
            print(inter[i,j])
            }
          }
```

De esta manera, se han encontrado 36 pares de interacciones entre SNP's asociados con la variable dependiente con un p-valor inferior a 0.05, 21 pares con un p-valor inferior a 0.025 y 7 pares de interacciones con un p-valor inferior a 0.01. Todas ellas se reflejan con más detalle en la página 10 del documento "Punto 7.xlsx" de "anexos/Resultados".

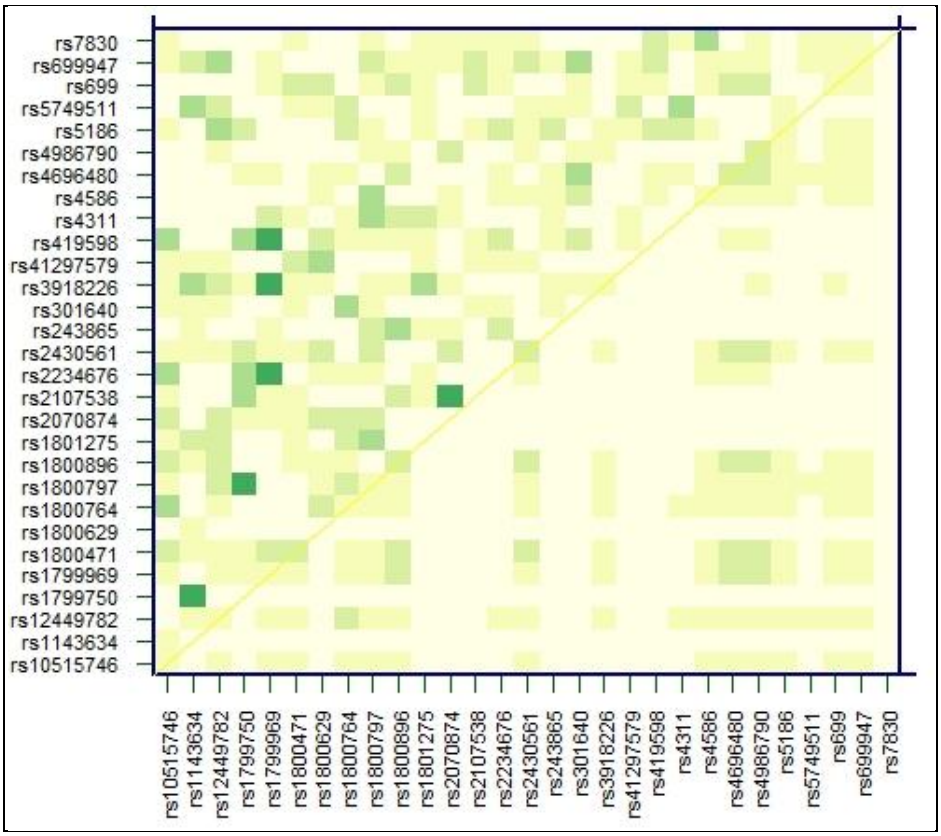
Para el análisis de asociación con la variable dependiente DCTR_otrDCTR, se procede de la misma manera que para la variable DCTRsi_no. Las matrices para cada modelo de herencia obtenidas son las siguientes:



Para determinar el modelo de herencia más adecuado, se han realizado gráficas comparativas similares a las de la variable DCTRsi_no. Éstas son las siguientes:



Como se puede ver, los p-valor que mejor se ajustan al modelo ideal son los correspondientes al modelo de herencia aditivo. A continuación, se muestra la matriz de dicho modelo con más detalle:



Así, se aprecia 6 celdas de color verde oscuro correspondientes a pares de interacciones asociados con la variable DCTR_otrDCTR. Para obtener exactamente todas las interacciones significativas se emplea el script creado 'int', obteniendo 27 pares de interacciones con p-valor inferior a 0.05, 14 pares con p-valor inferior a 0.025 y 6 pares de SNP's con un p-valor inferior a 0.01.

El objetivo de este apartado era obtener SNP's asociados con la variable dependiente al interactuar con otros SNP's con el fin de poder reducir al máximo el grupo de variables genéticas que se van a usar para el modelado predictivo a la vez que maximizar el éxito en la predicción. Por ello, se ha visto conveniente la implementación de un script ('intCom') que, dadas las 2 variables con los resultados anteriores, obtiene los pares de SNP's que, al interactuar, presentan asociación significativa con ambas variables dependientes. Un fragmento de dicho script se muestra a continuación:

```
for(i in 1:length(int2)){
      if(!is.na(match(int2[i], int1))){
            coinc[j] <- int2[i]
            j <- j+1;
            }
      }
```

De esta manera se han obtenido las siguientes coincidencias:

p-valor	Pares de interacciones
0.05	rs1143634 - rs1799750, rs1799969 - rs3918226, rs1800872 - rs2243248
0.025	rs1143634 - rs1799750, rs1799969 - rs3918226
0.01	rs1143634 - rs1799750, rs1799969 - rs3918226

7.4.Discusión

La muestra de estudio presenta cierta tendencia a agrupar los registros en 2 grupos con una probabilidad de éxito de 60% a 70%, en función de si se emplea el conjunto de datos inicial eliminando los registros con datos faltantes o si se emplea el conjunto de datos imputado. Por otro lado, si se trata de realizar una agrupación en 3 clusters, la probabilidad de éxito es muy baja.

Para poder obtener un modelo que mejore la probabilidad de éxito en la clasificación de los registros, se ha intentado reducir el conjunto de variables, manteniendo aquellas que contribuyen a obtener un valor para la variable dependiente y eliminando aquellas que puedan interferir en el poder predictivo de otras o que no aporten información relevante al modelo y únicamente aumenten la complejidad del mismo.

Es por ello que se ha hecho un primer análisis de asociación entre todas las variables. Para las variables clínicas de tipo continuo se ha encontrado una asociación lineal entre Edad y Edad_donante, de manera que para la realización del modelo predictivo se va a emplear la variable derivada Edad/Edad_donante con el fin de reducir el número de variables a la vez que mantienen o mejoran los resultados.

El resto de variables clínicas, han presentado asociaciones entre sí y con algunas de las variables genéticas, por lo que se van a mantener para el modelado predictivo.

Respecto a las variables clínicas, tras obtener un elevado número de asociaciones entre sí, se ha realizado un estudio atendiendo a las características genéticas de las mismas. Así, en primer lugar, se ha visto si la población de control estaba en equilibrio de Hardy-Weinberg para todos los SNP's, encontrando que no era así para rs1800796, por lo que se ha eliminado del conjunto de variables. Respecto a la población de casos se ha encontrado que rs1801275 tampoco estaba en equilibrio, lo que se ha tenido en consideración para los siguientes análisis.

Posteriormente, tras un análisis de haplotipos mediante Haploview, se ha determinado que algunos SNP's se heredaban junto con otros sin sufrir recombinaciones y, por lo tanto, al presentar desequilibrio de enlace, la información que aportaban era similar y se podía prescindir de ellos. Éste ha sido el caso de los SNP's: rs1800871, rs10515746, rs833061, rs1800797, rs4986791.

Debido a las limitaciones para la obtención del genotipado de algunos SNP's a través del proyecto HapMap, se ha visto conveniente realizar un análisis de asociación con la variable dependiente DCTRsi_no mediante haplotipos. Algunos han presentado asociación significativa con la variable dependiente, por lo que los SNP's que los forman se van a incluir en el modelo predictivo. Éstos son:

- rs1800872, rs1800896, rs699, del cromosoma 1
- rs1143634, rs2234676, rs419598, del cromosoma 2
- rs4696480, del cromosoma 4
- rs1801275, rs243865, del cromosoma 16
- rs4586, rs2107538, del cromosoma 17
- rs1799969, rs1800471, del cromosoma 19

Durante dicho análisis también se han encontrado SNP's que aportaban el mismo alelo tanto para un haplotipo que presentaba asociación con la enfermedad como para otros que no, con lo que podía distorsionar el poder predictivo del resto de variables genéticas. Éste ha sido el caso de rs5743708 y rs1800825. También se ha encontrado que el SNP rs2071231 presentaba el mismo valor en más de un 90% de las ocasiones y contribuía a aumentar la complejidad del modelo sin aportaban información de interés.

Con los SNP's que participaban en haplotipos asociados con la variable DCTRsi_no, se ha realizado un análisis de asociación univariante ya que dicha asociación podía deberse a que era uno de los SNP's el que estaba asociado. No obstante, no se obtuvo ningún SNP asociado de manera aislada con la variable dependiente.

Finalmente, dado que las interacciones entre SNP's han demostrado que pueden ser relevantes para la predicción de enfermedades, se ha realizado un análisis de pares de interacciones con las variables dependientes DCTRsi_no y DCTR_otrDCTR. En ambos casos se han obtenido interacciones asociadas con la enfermedad, algunas comunes y otras particulares para cada variable dependiente. Con el fin de obtener un único modelo predictivo, se van a emplear los SNP's que participan en los pares de interacciones comunes con asociación significativa; estos son:

- rs1143634 - rs1799750, del cromosoma 2 y 11, respectivamente.
- rs1799969 - rs3918226, del cromosoma 19 y 7, respectivamente.
- rs1800872 - rs2243248, del cromosoma 1 y 5, respectivamente.

Por lo tanto, el grupo de variables que se van a utilizar para crear el modelo predictivo son las variables genéticas cuyas interacciones con otras hayan resultado significativas, las variables clínicas iniciales con la variable derivada Edad/Edad_donante y las variables genéticas cuyos haplotipos están asociados con la variable dependiente DCTRsi_no.

8. MODELADO PREDICTIVO

En este apartado se van a emplear 3 métodos de modelado para predecir las variables dependientes DCTRsi_no y DCTR_otrDCTR. Para cada uno, además, se van a crear 3 modelos:

- Un modelo utilizando únicamente las variables clínicas.
- Un segundo modelo usando las variables genéticas seleccionadas previamente.
- Un tercer modelo con las variables clínicas y genéticas.

De esta manera se puede comparar qué grupo de variables aporta información más valiosa en la predicción. A continuación se muestran las variables que se van a utilizar:

variables clínicas	variables genéticas	
Edad/Edad_donante	rs1800872	rs3918226
Sexo	rs1800896	rs1799750
Sexo_donante	rs699	rs1801275
t_isqm	rs1143634	rs243865
Matches A	rs2234676	rs4586
Matches B	rs419598	rs2107538
Matches DR	rs4696480	rs1799969
	rs2243248	rs1800471

Todos los resultados que se muestran a continuación aparecen de forma más detallada en el documento Excel “Punto 8.xlsx” de la carpeta “anexos/Resultados” que se adjunta.

8.1.LDA

Mediante el análisis discriminante lineal, LDA, se pretende encontrar una combinación lineal de variables tal y como se describió en el apartado 4.9.1. La herramienta empleada para realizar este modelo ha sido R, con la librería MASS. En primer lugar, se ha probado este método para la variable dependiente DCTRsi_no, evaluando el modelo mediante validación cruzada:

SNP's + variables clínicas:

	0	1
0	48	70
1	56	98

Correctos	146	53.67%
Erróneos	126	46.33%
Total	272	

variables clínicas:

	0	1
0	38	80
1	36	118

Correctos	156	57.36%
Erróneos	116	42.64%
Total	272	

SNP's :

	0	1
0	41	77
1	53	101

Correctos	142	52.20%
Erróneos	130	47.80%
Total	272	

Como se puede ver, el porcentaje de aciertos es superior a 50% utilizando tanto las variables al completo, como por separado. No obstante, no supera los aciertos obtenidos mediante aprendizaje no supervisado. Para comprobar la generalidad del modelo, a pesar del relativamente reducido número de muestras, también se ha evaluado realizando una partición para entrenamiento y otra para test; en concreto, el grupo de entrenamiento corresponde al 85% de las muestras, mientras que el de test es el 15% restante. Los resultados obtenidos han sido similares, con un número bastante elevado de falsos positivos mediante ambas evaluaciones:

SNP's + variables clínicas:

	0	**1**
0	9	13
1	6	13

Correctos	22	53.66%
Erróneos	19	46.34%
Total	41	

variables clínicas:

	0	**1**
0	8	14
1	5	14

Correctos	22	53.66%
Erróneos	19	46.34%
Total	41	

SNP's:

	0	**1**
0	3	19
1	4	15

Correctos	18	43.90%
Erróneos	23	56.10%
Total	41	

Respecto al modelo predictivo para la variable DCTR_otrDCTR, también se ha evaluado, en primer lugar, mediante validación cruzada para los 3 grupos de variables, obteniendo los siguientes resultados:

SNP's + variables clínicas:

	0	**1**	**2**
0	58	18	42
1	27	8	21
2	44	15	39

Correctos	105	38.60%
Erróneos	167	61.40%
Total	272	

variables clínicas:

	0	**1**	**2**
0	81	3	34
1	30	0	26
2	46	1	51

Correctos	132	48.53%
Erróneos	140	51.47%
Total	272	

SNP's:

	0	**1**	**2**
0	59	15	44
1	27	5	24
2	54	13	31

Correctos	95	34.93%
Erróneos	177	65.07%
Total	272	

En este caso, tampoco mejoran los resultados obtenidos mediante aprendizaje no supervisado ya que el porcentaje de aciertos en la predicción no alcanza el 50%.

Al igual que para la variable DCTRsi_no, se ha realizado la evaluación utilizando dos grupos independientes para entrenamiento y test, con la misma proporción de muestras que en el caso anterior. Como se puede comprobar, el porcentaje de aciertos es algo inferior al obtenido mediante validación cruzada, llegando únicamente al 20% de aciertos para el modelo formado por las variables genéticas:

SNP's + variables clínicas:

	0	**1**	**2**
0	10	6	6
1	4	0	0
2	5	5	5

Correctos	15	36.58%
Erróneos	26	63.42%
Total	41	

variables clínicas:

	0	**1**	**2**
0	13	0	9
1	4	0	0
2	9	1	5

Correctos	18	43.90%
Erróneos	23	56.10%
Total	41	

SNP's:

	0	**1**	**2**
0	6	8	8
1	3	0	1
2	6	6	3

Correctos	9	21.95%
Erróneos	32	78.05%
Total	41	

Además, mediante ambas evaluaciones, el número de registros predichos para DCTR_otrDCTR = 1, es nulo o muy inferior en comparación con el resto de valores de la variable dependiente.

Así, dados los resultados anteriores, se puede concluir que el método de discriminante lineal no ha encontrado una combinación lineal de las variables que permita clasificar con cierta fiabilidad los registros para la variable dependiente DCTRsi_no ni para la variable dependiente DCTR_otrDCTR, siendo de inferior calidad la clasificación obtenida para esta última variable.

8.2. SVM

Mediante Support Vector Machine, SVM, se han creado los correspondientes modelos de clasificación para las variables dependientes DCTRsi_no y DCTR_otrDCTR. Al igual que mediante LDA, se ha realizado la evaluación en primer lugar por validación cruzada y posteriormente mediante dos particiones independientes de entrenamiento y test, con una proporción de muestras de 85% y 15%, respectivamente.

Para la evaluación mediante validación cruzada, se han dividido las muestras en 10 subconjuntos que se irán alternando entre entrenamiento y test durante las 10 iteraciones correspondientes. Para ello, se ha empleado la librería e1071 de la herramienta R, que aporta el porcentaje de aciertos tanto global como el obtenido en cada iteración pero, por el contrario, no permite obtener la matriz de contingencia resultante.

A continuación se indican los porcentajes obtenidos:

SNP's + variables clínicas: aciertos: 0.5404412

variables clínicas: aciertos: 0.5294118

SNP's: aciertos: 0.5551471

En los 3 modelos se supera el 50% de los aciertos, siendo mejor el modelo formado únicamente por variables genéticas.

Para la evaluación mediante las particiones independientes de entrenamiento y test se han obtenido las matrices siguientes:

SNP's + variables clínicas:

	0	**1**
0	8	14
1	6	13

Correctos	21	51.22%
Erróneos	20	48.78%
Total	41	

variables clínicas:

	0	**1**
0	7	15
1	5	14

Correctos	21	51.22%
Erróneos	20	48.78%
Total	41	

SNP's:

	0	**1**
0	4	18
1	4	15

Correctos	19	46.34%
Erróneos	22	53.66%
Total	41	

El porcentaje de aciertos es bastante similar al obtenido mediante validación cruzada y, al igual que mediante el método LDA, el porcentaje de falsos positivos es bastante elevado: 60% para el modelo formado por ambos grupos de variables, y 80% para el modelo formado únicamente por SNP's.

Con el modelo que clasifica los registros para la variable dependiente DCTR_otrDCTR, evaluado mediante validación cruzada, se han obtenido porcentajes de aciertos que no alcanzan el 50% de éxito en la clasificación, tal y como se muestra a continuación:

SNP's + variables clínicas: aciertos: 0.4191176

variables clínicas: aciertos: 0.4705882

SNP's: aciertos: 0.3125

Posteriormente se ha realizado la evaluación mediante las particiones independientes de entrenamiento y test, obteniendo porcentajes de acierto algo inferiores en comparación a validación cruzada:

SNP's + variables clínicas:

	0	1	2
0	8	3	11
1	4	0	0
2	6	3	6

Correctos	14	34.15%
Erróneos	27	65.85%
Total	41	

variables clínicas:

	0	1	2
0	13	0	9
1	4	0	0
2	11	0	4

Correctos	17	41.46%
Erróneos	24	58.54%
Total	41	

SNP's:

	0	1	2
0	7	7	8
1	3	0	1
2	6	6	3

Correctos	10	24.40%
Erróneos	31	75.60%
Total	41	

Además, se ha dado la misma tendencia que con LDA, en la que el porcentaje de aciertos en la predicción para DCTR_otrDCTR = 1 es nulo.

Ante estos resultados, SVM no ha podido separar los registros mediante una frontera lineal de manera clara y el clasificador no alcanza los resultados deseados para las variables dependientes DCTRsi_no ni DCTR_otrDCTR.

8.3.Árboles de decisión

Para la realización del modelo mediante árboles de decisión se ha utilizado la herramienta Clementine, creando un modelo para cada variable dependiente y empleando los mismos grupos de variables que en los dos métodos anteriores.

Mediante esta herramienta se puede ver si durante la realización del modelo ha habido alguna variable que no se ha utilizado, en cuyo caso se elimina y se prueba el nuevo conjunto de variables.

Los resultados obtenidos en la realización de un árbol de decisión para la variable DCTRsi_no han sido los siguientes:

SNP's:

	0	1
0	73	45
1	15	139

Correctos	212	77.94%
Erróneos	60	22.06%
Total	272	

variables clínicas:

	0	1
0	93	25
1	37	117

Correctos	210	77.21%
Erróneos	62	22.79%
Total	272	

SNP's + variables clínicas:

	0	1
0	101	17
1	19	135

Correctos	236	86.76%
Erróneos	36	13.24%
Total	272	

Mediante este método los resultados son notablemente mejores, superando en los 3 casos el 75% de aciertos en la clasificación de los registros. El mejor modelo es el formado por las variables genéticas y clínicas, con un 86% de aciertos; además, dicho modelo presenta elevados porcentajes de sensibilidad y

especificidad (87,66% y 85,59%, respectivamente), por lo que predice de forma equilibrada tanto los pacientes sanos como los enfermos.

A pesar de estos resultados, se vio que para el primer modelo, algunas de las variables genéticas del conjunto inicial no aportaban información para el árbol de decisión, por lo que se eliminaron. Dichas variables eran: rs1801275, rs2107538 y rs419598. Los resultados obtenidos mediante los 13 SNP's restantes han sido los siguientes:

	0	**1**
0	73	45
1	17	137

Correctos	210	77.21%
Erróneos	62	22.79%
Total	272	

Respecto al modelo formado por SNP's y variables clínicas, también se observó ciertas variables que no se empleaban para la elaboración del modelo y se descartaron: Matches_DR, rs419598 y rs699, por lo que el modelo queda con 14 variables genéticas y 6 variables clínicas. A continuación se ve que el porcentaje de aciertos se mantiene, así como la precisión para detectar pacientes sanos o enfermos:

	0	**1**
0	103	15
1	19	135

Correctos	238	87.5%
Erróneos	34	12.5%
Total	272	

Al igual que en el caso de LDA y SVM, se ha evaluado el modelo mediante particiones independientes, dedicando un 85% de las muestras a entrenamiento y el 15% restante para test. En este caso, el conjunto de variables utilizado ha sido el reducido en el paso anterior, obteniendo los siguientes resultados:

SNP's:

	0	**1**
0	7	9
1	1	23

Correctos	30	75%
Erróneos	10	25%
Total		

variables clínicas:

	0	**1**
0	9	7
1	4	19

Correctos	28	71.79%
Erróneos	11	28.21%
Total	39	

SNP's + variables clínicas:

	0	**1**
0	13	4
1	6	17

Correctos	30	75%
Erróneos	10	25%
Total	40	

El porcentaje de aciertos en los 3 casos es inferior al obtenido mediante validación cruzada pero tanto en el formado únicamente por SNP's como en el que tiene en cuenta, además, las variables clínicas, mantienen porcentajes de 75% de aciertos y un total de 13 y 14 variables genéticas, respectivamente en lugar de las 42 empleadas durante el aprendizaje no supervisado. También cabe destacar que, pese a que los porcentajes sean iguales, el modelo ajustado por variables clínicas presenta mejores resultados, debido a una predicción similar tanto para casos como para controles, a diferencia del primero, con un 56,25% de falsos positivos.

Para la variable dependiente DCTR_otrDCTR, se ha empleado la misma metodología, obteniendo mediante validación cruzada los siguientes resultados:

SNP's:

	0	1	2
0	99	5	14
1	16	32	8
2	25	8	65

Correctos	196	72.06%
Erróneos	76	27.94%
Total	272	

variables clínicas:

	0	1	2
0	97	6	15
1	12	28	16
2	27	3	68

Correctos	193	70.96%
Erróneos	79	29.04%
Total	272	

SNP's + variables clínicas:

	0	1	2
0	105	2	11
1	16	37	3
2	14	4	80

Correctos	222	81.62%
Erróneos	50	18.38%
Total	272	

El porcentaje de aciertos es inferior al obtenido para la variable dependiente DCTRsi_no, no obstante, supera en los 3 modelos el 70% de éxito en la clasificación. Los mejores resultados se han obtenido para el modelo formado por las variables clínicas y genéticas, con un 81.62% de aciertos, seguido del modelo formado únicamente por variables genéticas con 72% de aciertos.

Tras determinar que en los modelos que incluían variables genéticas había algunas que no se empleaban en el árbol de decisión, para el modelo formado sólo por SNP's se han eliminado las variables rs243865 y rs419598; y para el ajustado por variables clínicas se han descartado los SNP's rs1800471, rs2243248 y rs419598. Los resultados han sido:

SNP's:
tras eliminar variables no utilizadas:

	0	1	2
0	99	5	14
1	16	32	8
2	25	8	65

Correctos	196	72.06%
Erróneos	76	27.94%
Total	272	

SNP's + variables clínicas:
tras eliminar variables no utilizadas:

	0	1	2
0	104	3	11
1	13	40	3
2	13	4	81

Correctos	225	82.72%
Erróneos	47	17.28%
Total	272	

Así, con 14 variables genéticas se puede obtener un 72% de aciertos en la clasificación y con 13 variables genéticas y 7 clínicas, un 82% de aciertos.

Para la evaluación mediante particiones independientes de entrenamiento y test, empleando el conjunto de variables ya reducido, se han obtenido las siguientes matrices de confusión:

SNP's:

	0	1	2
0	15	0	2
1	1	4	3
2	4	1	9

Correctos	28	71.79%
Erróneos	11	28.21%
Total	39	

variables clínicas:

	0	1	2
0	12	0	2
1	7	2	0
2	3	0	16

Correctos	30	71.43%
Erróneos	12	28.57%
Total	42	

SNP's + variables clínicas:

	0	1	2
0	16	0	0
1	6	7	0
2	7	0	12

Correctos	35	72.92%
Erróneos	13	27.08%
Total	48	

En todos los modelos se supera el 70% de los aciertos, obteniendo una muy buena clasificación para registros con DCTR_otrDCTR = 0.

8.4.Discusión

Durante la etapa de modelado se han probado 3 técnicas diferentes: LDA, SVM y árboles de decisión, empleando en todas ellas 3 grupos diferentes de variables para obtener cuál es el que ofrece mejores resultados. Además, dado que se disponía de un reducido número de variables, se evaluaron los modelos mediante validación cruzada, ya que así se realiza el entrenamiento con suficientes registros. No obstante, para probar cuán general es el modelo, se ha optado por realizar también la evaluación mediante particiones independientes.

Para el modelo realizado con LDA, los aciertos alcanzaban el 50% para una evaluación con validación cruzada y apenas el 40% para la evaluación por particiones independientes. Dichos resultados se han obtenido empleando 7 variables clínicas y 16 variables genéticas, por lo que el modelo emplea más de la mitad de las variables explicativas iniciales.

Para el modelo realizado con SVM, el porcentaje de aciertos era similar al obtenido mediante LDA, empleando el mismo número de variables, tanto para la evaluación por validación cruzada, como para particiones independientes.

Respecto a los modelos obtenidos mediante árboles de decisión los resultados han sido satisfactorios. En la clasificación de registros para la variable DCTRsi_no se ha conseguido un modelo formado por 14 variables genéticas y 6 variables clínicas con el 86% de aciertos mediante validación cruzada, mientras que mediante particiones independientes se ha alcanzado el 75% de los aciertos. En cualquier caso, con un reducido número de variables se alcanzan predicciones fiables. Las variables que permiten dichos resultados son:

- Variables clínicas: Edad/Edad_donante, Sexo, Sexo_donante, t_isqm, Matches_A y Matches_B.
- Variables genéticas: rs1800872, rs1800896, rs1143634, rs2234676, rs4696480, rs2243248, rs3918226, rs1799750, rs1801275, rs243865, rs4586, rs2107538, rs1799969, rs1800471.

Para la clasificación de los registros en función de la variable DCTR_otrDCTR, los mejores resultados se han obtenido mediante el modelo formado por variables genéticas ajustado por las variables clínicas, con resultados de 80% de aciertos mediante validación cruzada y 70% mediante particiones independientes. Los resultados no son tan notables como para la variable DCTRsi_no pero, dada la clasificación inicial obtenida con el algoritmo K-medias y posteriormente mediante LDA y SVM, los presentes resultados pueden considerarse satisfactorios. El conjunto de variables utilizado ha sido:

- Variables clínicas: Edad/Edad_donante, Sexo, Sexo_donante, t_isqm, Matches_A, Matches_B y Matches_DR.
- Variables genéticas: rs1800872, rs1800896, rs699, rs1143634, rs2234676, rs4696480, rs3918226, rs1799750, rs1801275, rs243865, rs4586, rs2107538 y rs1799969.

Para una estimación final del error de generalización deberíamos disponer de un conjunto independiente de casos, que en el momento de presentar el proyecto no había sido adquirido, pero que se ha presentado como interesante por los resultados obtenidos.

9. CONCLUSIONES

Al igual que en otras enfermedades, se ha demostrado que el perfil genético está asociado directamente con el rechazo post-trasplante. En concreto, se ha visto que 6 haplotipos están asociados con la aparición del rechazo y 6 SNP's, al interactuar, presentan asociación, además, con el tipo de rechazo que se va a dar.

Por otro lado, si se toman los SNP's que dan lugar a los haplotipos significativos, así como aquellos cuyas interacciones son significativas, se puede construir un árbol de decisión para determinar la ausencia y aparición de rechazo con un 87% de aciertos. Dado que a la fecha de realización del proyecto no se disponía de más muestras, se ha comprobado su efectividad mediante particiones independientes de entrenamiento y test, dando lugar a un 75% de aciertos. Ello implica que el rechazo post-trasplante puede detectarse antes de que aparezca, permitiendo iniciar las medidas oportunas para evitarlo.

Además, si se considera el modelo predictivo creado únicamente con los SNP's seleccionados, se ha obtenido un 77% de aciertos mediante validación cruzada y un 75% de aciertos mediante particiones independientes, lo cual implica que, sin tener datos del historial clínico del paciente obtenidos durante la operación, como es el caso del tiempo de isquemia, se puede predecir la aparición de rechazo post-trasplante únicamente mediante el perfil genético. Así, antes de realizarse el trasplante puede conocerse si el paciente va a presentar rechazo.

Es obvio que la investigación sobre el rechazo del injerto tras un trasplante renal está aún en sus inicios ya que dicha práctica empezó a realizarse a partir de la segunda mitad del siglo XX. No obstante, el hecho de que el perfil genético pueda determinar con un 75% de aciertos si va a aparecer rechazo post-trasplante supone un paso significativo y puede aportar algo de luz para futuras investigaciones.

Bibliografía

MedlinePlus. http://www.nlm.nih.gov/medlineplus/spanish/ency/article/000502.htm (último acceso: Mayo de 2011).

AdaBoost. http://en.wikipedia.org/wiki/AdaBoost (último acceso: Marzo de 2011).

Creatinina. http://html.rincondelvago.com/creatinina.html (último acceso: Mayo de 2011).

El gen de los obesos. 24 de Julio de 2005. http://www.infobae.com/notas/198298-El-gen-de-los-obesos.html (último acceso: Junio de 2011).

À propos du projet international HapMap. http://hapmap.ncbi.nlm.nih.gov/abouthapmap.html.fr (último acceso: Mayo de 2011).

Asociación de variables cualitativas: test de Chi-cuadrado. http://www.fisterra.com/mbe/investiga/chi/chi.asp (último acceso: Marzo de 2011).

Assortative mating. http://en.wikipedia.org/wiki/Assortative_mating (último acceso: Junio de 2011).

Autism spectrum disorder. http://en.wikipedia.org/wiki/Autism_spectrum_disorder (último acceso: Junio de 2011).

Background on IMPUTE. https://mathgen.stats.ox.ac.uk/impute/impute_background.html (último acceso: Julio de 2011).

Banco Nacional de Órganos y Tejidos. «Compatibilidad HLA.» Uruguay, 2004.

Bonferroni correction. http://en.wikipedia.org/wiki/Bonferroni_correction (último acceso: Mayo de 2011).

Brett A. McKinney, David M. Reif, Marylyn D. Ritchie, Jason H. Moore. *Machine Learning for Detecting Gene-Gene Interactions.* Biomedical Genomics and Proteomics, 2006.

Broad Institute. 2010. http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/tutorial (último acceso: Marzo de 2011).

Bryan Howie, Jonathan Marchini. «Instructions for IMPUTE version 2.» 2009.

Bryan N. Howie, Peter Donnelly, Jonathan Marchini. «A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies.» *PLoS Genetics.* 19 de Junio de 2009. http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000529 (último acceso: Julio de 2011).

Chi-square with R. 2008. http://courses.statistics.com/software/R/Rchisq.htm (último acceso: Marzo de 2011).

«Capítulo 1.» En *Proyecto Ecosfera. Biología y geología 1*, de José M. Gómez de Salazar García Galiano, Emilio Pedrinaci Rodríguez Concha Gil Soriano. Madrid: SM, 2002.

«Capítulo 11.» En *Proyecto Ecosfera. Biología y geología 1*, de José M. Gómez de Salazar García Galiano, Emilio Pedrinaci Rodríguez Concha Gil Soriano. Madrid: SM, 2002.

Daga Ruiz, D., Fernández Aguirre, C., Segura González, F., Carballo Ruiz, M. «Indicaciones y resultados a largo plazo de los trasplantes de órganos sólidos. Calidad de vida en pacientes trasplantados.» Málaga, 2008.

Diabetes mellitus tipo 2. http://es.wikipedia.org/wiki/Diabetes_mellitus_tipo_2 (último acceso: Junio de 2011).

E. Gallego Valcarce, A. Ortega Cerrato, F. Llamas Fuentes, J. Masiá Mondéjar, G. Martínez Fernández, E. López Rubio, A. López Montes, J. Pérez Martínez, M. Martínez Villaescusa, C. Gómez Roldán. «Nefrologia.» *El tiempo de isquemia fría corto optimiza los resultados de los trasplantes renales efectuados con donantes con criterios expandidos.* 22 de Febrero de 2010. http://www.revistanefrologia.com/modules.php?name=articulos&idarticulo=416&idlangart=ES (último acceso: Junio de 2011).

Expectation-maximization algorithm. http://en.wikipedia.org/wiki/Expectation-maximization_algorithm (último acceso: Marzo de 2011).

F. Fernández-Bañares, M. Esteve-Comas, M. Rosinach. *Cribado de la enfermedad celíaca en grupos de riesgo.* Barcelona: Procesos en Gastroenterología, 2004.

Genotype. http://en.wikipedia.org/wiki/Genotype (último acceso: Junio de 2011).

Genotype-phenotype distinction. http://en.wikipedia.org/wiki/Genotype-phenotype_distinction (último acceso: Junio de 2011).

«Glaucoma-Associated CYP1B1 Mutations Share Similar Haplotype Backgrounds in POAG and PACG Phenotypes.» *IOVS.* 2007. http://www.iovs.org/content/48/12/5439.full (último acceso: Abril de 2011).

«Glomerulonefritis.» *MedlinePlus.* 8 de Diciembre de 2009. http://www.nlm.nih.gov/medlineplus/spanish/ency/article/000484.htm (último acceso: Marzo de 2011).

GTOOL. http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html (último acceso: Mayo de 2011).

«Haplotipo del gen ENPP1 (PC-1) asociado con el riesgo de obesidad y diabetes de tipo 2, y sus aplicaciones.» *Patentados.com.* 2007. http://patentados.com/patente/haplotipo-gen-enpp1-pc-1-asociado-riesgo/ (último acceso: Junio de 2011).

Haplotype. http://en.wikipedia.org/wiki/Haplotype (último acceso: Mayo de 2011).

Hardy-Weinberg principle. http://en.wikipedia.org/wiki/Hardy%E2%80%93Weinberg_principle (último acceso: Mayo de 2011).

Herencia intermedia y codominancia. http://webcache.googleusercontent.com/search?q=cache:nxNV7u9jsVEJ:www.iesbanaderos.org/html/departamentos/bio-geo/Apuntes/Bio/T14_Genet/3%2520herencia%2520Intermedia.htm+modelo+codominante+herencia&cd=1&hl=es&ct=clnk&gl=es&source=www.google.es (último acceso: Julio de 2011).

Hidden Markov Model. http://en.wikipedia.org/wiki/Hidden_Markov_model (último acceso: Mayo de 2011).

«Hipertension.» *GeoSalud.* http://www.geosalud.com/hipertension/que%20es%20hipert.htm (último acceso: Marzo de 2011).

IMPUTE2. 8 de Diciembre de 2010. https://mathgen.stats.ox.ac.uk/impute/impute_v2.html (último acceso: Mayo de 2011).

Insuficiencia renal. http://es.wikipedia.org/wiki/Insuficiencia_renal (último acceso: Mayo de 2011).

International HapMap Project. http://hapmap.ncbi.nlm.nih.gov/ (último acceso: Junio de 2011).

International HapMap Project. *The Origins of Haplotypes.* http://hapmap.ncbi.nlm.nih.gov/originhaplotype.html.en (último acceso: Junio de 2011).

Juan R. González, Lluís Armengol, Elisabet Guinó, Xavier Solé, Víctor Moreno. «Package 'SNPassoc'.» *SNPs based whole genome association studies.* 2009.

Kalmes R, Huret JL. *Modelo de Hardy-Weinberg.* Febrero de 2001. http://atlasgeneticsoncology.org/Educ/HardySp.html (último acceso: Mayo de 2011).

L.N. Karla Melissa Ruiz-Dyck, Dr. Norberto Sotelo-Cruz N, Dra. Ana María Calderón de la Barca. «Tipificación de Haplotipos que predisponen a enfermedad celiaca en sangre de cordón umbilical de niños sonorenses.» Sonora, 2010.

«Linear Discriminant Analysis.» *DTREG.* http://www.dtreg.com/lda.htm (último acceso: Junio de 2011).

«Link Functions and the Generalized Linear Model.» 2010. http://www.upa.pdx.edu/IOA/newsom/da2/ho_link.pdf (último acceso: Mayo de 2011).

M. Pérez Fontán, A. Rodríguez-Carmona, F. Valdés. *Diálisis peritoneal antes del trasplante renal, ¿procedimiento de elección o factor de riesgo?* A Coruña: Servicio Nefrología. Hospital Juan Canalejo, 2000.

Manual de Estadística. Capítulo III: DISTRIBUCIONES BIDIMENSIONALES . http://www.eumed.net/cursecon/libreria/drm/1f.htm (último acceso: Marzo de 2011).

Markov chain Monte Carlo. http://en.wikipedia.org/wiki/Markov_chain_Monte_Carlo (último acceso: Mayo de 2011).

Marta Crespo Barrio, Nuria Esforzado Armengol, Maria José Ricart Brulles, Federico Oppenheimer Salinas. *Resultados a largo plazo del trasplante renal de donante vivo: supervivencia de injerto y receptor.* Barcelona: Servicio de Nefrología. Unidad Trasplante Renal. Hospital Clínico de Barcelona., 2005.

Mauricio Delbracio, Matías Mateu. «Trabajo Final de Reconocimiento de Patrones:Identificación utilizando PCA, ICA y LDA.» 2006.

Merck Sharp & Dohme. «Trastornos del riñón y de las vias urinarias.» *Merck Sharp & Dohme.* http://www.msd.es/content/index.html (último acceso: Junio de 2011).

Mocarquer, Alfredo. «Daño Crónico del Injerto en Trasplante Renal.» *Medwave.* http://www.mednet.cl/link.cgi/Medwave/Cursos/trasplantes/III/3431 (último acceso: Mayo de 2011).

National Center for Research Resources. «Kidney Trasplant Rejection.» Rochester, 2009.

National Kidney Federation. *Kidney disease.* http://www.kidney.org/kidneydisease/ (último acceso: Junio de 2011).

—. «What is trasplant rejection?» *National Kidney Federation.* Marzo de 2010. http://www.kidney.org.uk/Medical-Info/transplant/txrej.html (último acceso: Mayo de 2011).

National Space Biomedical Research Institute. *Human Physiology in Space.* http://www.nsbri.org/HumanPhysSpace/focus4/ep-kidney.html (último acceso: Junio de 2011).

NCBI. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/ (último acceso: Junio de 2011).

«Nefritis intersticial.» *MedlinePlus.* http://www.nlm.nih.gov/medlineplus/spanish/ency/article/000464.htm (último acceso: Marzo de 2011).

«Nefropatía diabética.» *MedlinePlus.* http://www.nlm.nih.gov/medlineplus/spanish/ency/article/000494.htm (último acceso: Marzo de 2011).

non linear dynamics. http://www.nonlinear.com/support/progenesis/samespots/faq/pq-values.aspx (último acceso: Febrero de 2011).

Novo, Mª Dolores García. «Estudio genético de la enfermedad celíaca.»

Obesidad. http://es.wikipedia.org/wiki/Obesidad (último acceso: Junio de 2011).

Orallo, José Hernández. «Práctica 3 de Minería de Datos. Validación con el Clementine.» Valencia, 2010.

—. «Tema 2: El proceso KDD. Técnicas de minería de datos.» Valencia, 2010.

Organización Nacional de Trasplantes. *Memoria actividad Trasplante Renal 2009.* Organización Nacional de Trasplantes, 2009.

P. Marti, P.Errasti. *Trasplante renal - Kidney transplant.* Pamplona: Departamento de Nefrología. Clínica Universitaria, 2006.

P. Sierra, C. Monsalve, O. Comps, E. Andrés. «Valoración preoperatoria del paciente con Enfermedad renal crónica.»

Parma, Diana L. «Genética del asma.» 2009.

Pedro Larrañaga, Iñaki Inza, Abdelmalik Moujahid. «Tema 7: Regresión logística.» País Vasco: Departamento de Ciencias de la Computación e Inteligencia artificial.

Phillips, Theresa. *Polymorphism.* http://biotech.about.com/od/glossary/g/polymorphism.htm (último acceso: Mayo de 2011).

Qué es una embolia o ictus. http://www.trasplante.es/doc.php?op=leer&id=91 (último acceso: Marzo de 2011).

Rafael Romero Villafranca, Luisa Zúnica Ramajo. «Procesos estocásticos y teoría de colas.» Valencia: UPV, 2005.

Ramírez, Guillermo. «Imputación de datos.» Venezuela.

Raquel Iniesta, Elisabet Guinó, Víctor Moreno. *Análisis estadístico de polimorfismos genéticos en estudios epidemiológicos.* Barcelona: La Gaceta Sanitaria, 2005.

Sinnwell JP, Schaid DJ. «Package 'haplo.stats'.» *Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous.* 2009.

Síntesis de las principales técnicas estadísticas aplicadas en la investigación sanitaria. 10 de Junio de 1998. http://www.investigalia.com/tecnica_estad.html (último acceso: Marzo de 2011).

SNPs: Variations on a theme. 20 de Septiembre de 2007. http://www.ncbi.nlm.nih.gov/About/primer/snps.html (último acceso: Mayo de 2011).

SPSS. *Referencia de nodos Clementine 9.0.* SPSS. Chicago, 2000.

Steen, Kristel Van. «Bioinformatics. Chapter 6: Population-based genetic association studies.»

Sterling, Edmundo. «Determinación de urea en sangre.» *Monografias.com.* http://www.monografias.com/trabajos17/urea/urea.shtml (último acceso: Mayo de 2011).

«SVM - Support Vector Machines.» *DTREG.* http://www.dtreg.com/svm.htm (último acceso: Junio de 2011).

«The Human Kidney Structure and Function.» *General Medicine suite101.* http://www.suite101.com/content/the-human-kidney-structure-and-function-a75153 (último acceso: Junio de 2011).

The Kidney. 20 de Abril de 2011. http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/K/Kidney.html (último acceso: Junio de 2011).

Toland, Amanda Ewart. *DNA Mutations.* 3 de Junio de 2001. http://www.geneticheath.com/g101_changes_in_dna.shtml (último acceso: Mayo de 2011).

Trasplante de riñón. http://es.wikipedia.org/wiki/Trasplante_de_ri%C3%B1%C3%B3n (último acceso: Mayo de 2011).

Traumatismo craneal. 12 de Enero de 2011. http://www.nlm.nih.gov/medlineplus/spanish/ency/article/000028.htm (último acceso: Marzo de 2011).

Twyman, Richard. *Haplotype mapping.* 20 de Marzo de 2003. http://genome.wellcome.ac.uk/doc_WTD020781.html (último acceso: Junio de 2011).

—. *Mutation or polymorphism?* 20 de Marzo de 2003. http://genome.wellcome.ac.uk/doc_WTD020780.html (último acceso: Mayo de 2011).

«Understanding and using sensitivity, specificity and predictive values.» *Indian Journal of Ophthalmology.* 2008. http://www.ijo.in/article.asp?issn=0301-4738;year=2008;volume=56;issue=1;spage=45;epage=50;aulast=Parikh (último acceso: Junio de 2011).

«Uropatía obstructiva.» *MedlinePlus.* http://www.nlm.nih.gov/medlineplus/spanish/ency/article/000507.htm (último acceso: Marzo de 2011).

V. Cadahía, L. Rodrigo, D. Fuentes, S. Riestra, R. de Francisco y M. Fernández. «Enfermedad celíaca (EC), colitis ulcerosa (UC) y colitis esclerosante primaria (CEP) asociadas al mismo paciente: estudio familiar.» *Scielo.* Diciembre de 2005. http://scielo.isciii.es/scielo.php?pid=S1130-01082005001200007&script=sci_arttext&tlng=es (último acceso: Junio de 2011).

«Valoración de pruebas diagnósticas.» *Asociación de la Sociedad Española de Hipertensión.* http://www.seh-lelha.org/pdiagnos.htm (último acceso: Junio de 2011).

What is a genome? http://www.ncbi.nlm.nih.gov/About/primer/genetics_genome.html (último acceso: Junio de 2011).

What is DNA? http://ghr.nlm.nih.gov/handbook/basics/dna (último acceso: Mayo de 2011).

Xie, Yihui. «An Introduction to Support Vector Machine and Implementation in R.» Beijing, 2007.

Yates' correction for continuity. http://en.wikipedia.org/wiki/Yates'_correction_for_continuity (último acceso: Marzo de 2011).