Análisis de variantes en esclerosis múltiple en el sur de Europa

Adrián Valls Carbó
Máster de bioinformática y bioestadística
Análisis de datos ómicos

Manel Comabella López
Alexandre Sànchez Pla

Fecha de entrega: 24 diciembre 2021

Universitat Oberta de Catalunya

Título del trabajo:	*Análisis de variantes en esclerosis múltiple en el sur de Europa*
Nombre del autor:	*Adrián Valls Carbó*
Nombre del consultor/a:	*Alexandre Sànchez Pla*
Nombre del PRA:	*Manuel Comabella López*
Fecha de entrega (mm/aaaa):	12/2021
Titulación:	*Máster en Bioinformática y bioestadística*
Área del Trabajo Final:	*Análisis de datos ómicos*
Idioma del trabajo:	Castellano
Número de créditos:	15
Palabras clave	*GWAS; Esclerosis múltiple; variantes*

Resumen del Trabajo (máximo 250 palabras):

La esclerosis múltiple (EM) es una enfermedad en cuya etiopatogenia están involucrados factores genéticos y ambientales. Los factores genéticos conocidos hasta el momento solo suponen un 48% de la heredabilidad estimada de la enfermedad. Los estudios genéticos en la población de pacientes del sur de Europa son escasos, por lo que se desconoce si existen marcadores asociados a enfermedad en esta región. Por ello, se realizó un genotipado con microarrays de polimorfismos de un solo nucleótido (SNP) mediante el MSChip de Illumina en 232 pacientes y 232 casos de pacientes afectos de EM identificados en Centre d'Esclerosi Multiple de Catalunya (CEMCAT, España). Se identificaron 11 loci asociados a la enfermedad, todos ellos localizados en la región del complejo mayor de histocompatibilidad, entre los genes HLA-DRA y BTNL2. Una revisión bibliográfica de los loci mostró una asociación con el HLA-DRB1*1501. No se encontraron polimorfismos asociados a enfermedad fuera del complejo mayor de histocompatibilidad.

Abstract (in English, 250 words or less):

Multiple sclerosis (MS) is a disease whose etiopathogenesis involves genetic and environmental factors. Known genetic factors so far explain only 48% of the estimated heritability of the disease. Genetic studies in the population of patients in southern Europe are scarce, so it is associated markers of the disease are unknown in this region. We genotyped with a single nucleotide polymorphisms (SNPs) microarray using Illumina MSChip in 232 patients and 232 cases of MS patients identified at Centre d'Esclerosi Multiple de Catalunya (CEMCAT, Spain). Eleven disease-associated loci were identified, all of them located in the major histocompatibility complex region between the HLA-DRA and BTNL2 genes. A literature review of the loci showed an association with HLA-DRB1*1501. No disease-associated polymorphisms were found outside the major histocompatibility complex.

Tabla de contenido

1

Introducción

A. Contexto y justificación del trabajo

A.1. Descripción general

La esclerosis múltiple (EM) es una enfermedad crónica neurológica caracterizada por la presencia de placas desmielinizantes inflamatorias localizadas en el sistema nervioso central (SNC), que conlleva un proceso neurodegerativo[1]. Se estima que existen 2.8 millones de personas afectadas por la EM en todo el mundo, con una prevalencia que ronda los 35.9 casos por 100.000 habitantes, lo que corresponde con un incremento de un 30% desde 2013[2]. En España se estima que existen aproximadamente 44.000 personas afectas por la enfermedad, con tasas de prevalencia que oscilan entre los 80 y 100 casos por 100.000 habitantes[3].

Aunque la etiología de la enfermedad es por el momento desconocida existen factores ambientales y genéticos que predisponen a la aparición de esta. La heredabilidad de la enfermedad es conocida por diversos estudios, que demuestran concordancias altas en gemelos monocigóticos[4] (si un gemelo padece la enfermedad, el segundo tiene un riesgo del 20-30% de desarrollarla[5]) o la alta frecuencia de casos en determinadas familias (15-20% de los pacientes con EM tiene historia familiar de la enfermedad[6,7]) sin embargo se estima que solo se conoce alrededor del 48% de los genes que explican esta

heredabilidad en poblaciones occidentales[8,9]. Aunque parte de esta heredabilidad puede que no se encuentre a nivel de secuencia genética sino en aspectos regulatorios epigenéticos, aún son necesarios más estudios para determinar todos los genes asociados a la enfermedad.

En la última década, la arquitectura genética la enfermedad sido ampliamente estudiada mediante estudios de asociación genética (GWAS)[9–12], poniendo de manifiesto el papel de diversos elementos del sistema inmune ya conocidos como los linfocitos T y otros más recientes como linfocitos B. Además, recientemente se ha puesto de manifiesto la participación de las células de la microglía, células pertenecientes al sistema inmune innato localizadas en el SNC, con diversos genes expresados en estas células[9]. Sin embargo, y con la excepción de las células microgliales, no se ha conseguido poner de manifiesto la participación de células propias del SNC como astrocitos, oligondendrocitos o neuronas en una enfermedad que afecta primariamente al SNC.

Aunque la mayoría de estudios realizados hasta el momento se han realizado en población estadounidense y del norte de Europa[9–12] y por lo tanto sobre población mayoritariamente descendiente de europeos, el conocimiento de las variantes genéticas asociadas a EM en el sur de Europa es más desconocido. Si bien es esperable que las variantes genéticas de la población americana descendiente de europeos y la población europea actual no difieran en exceso, la propia heterogeneidad genética de la población europea[13] puede dar lugar a la presencia de variantes genéticas minoritarias asociadas a enfermedad en poblaciones concretas. Sin embargo, los estudios genéticos realizados exclusivamente en población europea[14] fallaron en encontrar variantes genéticas asociadas a enfermedad fuera del complejo mayor de histocompatibilidad (MHC por sus siglas en inglés). Aunque recientemente se han publicado estudios con datos de población española, estos estaban más enfocados al estudio de variantes genéticas asociadas a EM en poblaciones hispano americanas y afro descendientes en EEUU, usando los datos de la población española como comparativa con la hispano americana[15].

El objetivo de este trabajo es realizar un estudio de las variantes genéticas asociadas a enfermedad en una cohorte del sur de Europa.

Para ello se realizará el estudio de N casos de pacientes afectados por la enfermedad y N controles.

A.2. Justificación del TFM

Este trabajo se justifica en tres principales pilares. Por un lado, aunque la EM es una enfermedad relativamente frecuente, su etiología sigue siendo desconocida. Se hipotetiza que la interacción de factores ambientales y genéticos puede contribuir a la enfermedad. Se han identificado diversas variantes genéticas asociadas a enfermedad, pero no se conocen todos los factores que determinan la heredabilidad de la enfermedad. El descubrimiento de los factores genéticos asociados a la enfermedad no solo estriba en el mero interés académico, sino en las implicaciones prácticas que puedan producir estos hallazgos.

La existencia de nuevos genes asociados a la enfermedad puede suponer la identificación de dianas moleculares que nos permitan desarrollar nuevos tratamientos para enfermedad. Con ello se podrán desarrollar nuevas terapias para la enfermedad en puntos clave de la patogenia de esta, pudiendo mejorar el control, evitar la progresión de la enfermedad o retrasando el desarrollo de síntomas.

En un segundo término, el descubrimiento de los factores genéticos de la enfermedad puede ayudar a comprender de un modo más certero los diferentes fenotipos de la enfermedad. Aunque la EM es una enfermedad que se caracteriza por los brotes inflamatorios que afectan al SNC, existen diferentes fenotipos de la enfermedad. Clásicamente se consideraban formas recurrentes, en las que los pacientes desarrollaban brotes de la enfermedad con relativa quiescencia de la actividad entre brotes y formas progresivas, en las que el déficit neurológico progresaba sin brotes. En la actualidad se añade una categoría de actividad, para determinar si el paciente presenta o no actividad inflamatoria. Si bien el sustrato clínico-patológico de la EM está bien establecido, es posible que se trate de la expresión común del daño a la vaina de mielina por múltiples mecanismos etiopatogénicos. Por ello, puede tener interés la correlación del genotipo con los diferentes genotipos de la enfermedad.

En último lugar, el descubrimiento de variantes genéticas asociadas a la enfermedad puede tener sentido de cara a la predicción del riesgo de padecer la enfermedad tanto en la población general, familiares de pacientes o pacientes con formas de significado incierto como síndromes radiológicos aislados. El uso de marcadores genéticos en estos subgrupos puede acotar la probabilidad de desarrollar la enfermedad. Sin embargo, el uso de variantes genéticas presentes en otras poblaciones puede que no sea extrapolable a otras poblaciones de pacientes en otras áreas del mundo. Es por ello por lo que determinar las variantes genéticas en nuestra región puede contribuir a mejorar la predicción del riesgo de desarrollar la enfermedad.

B. Objetivos

B.1. Objetivos generales

B.1.1. Describir las variantes genéticas asociadas a esclerosis múltiple en una población del sur de Europa.

B.1.2. Determinar si existe participación de genes asociados a la microglía, oligodendrocitos, astrocitos o neuronas en nuestra muestra.

B.2. Objetivos específicos

B.2.1. Describir y analizar nuevas variantes genéticas asociadas con el control y regulación del sistema inmune asociadas a la EM.

B.2.2. Replicar la presencia de variantes genéticas asociadas a esclerosis múltiple ya descritas en otros estudios previos.

B.2.3. Describir y analizar nuevas variantes genéticas asociadas a células localizadas en el sistema nervioso central.

B.2.4. Describir y analizar los pathways de las nuevas variantes encontradas y relacionarlos con los ya existentes con anterioridad.

B.2.5. Comparar las nuevas variantes genéticas y pathways con los conocidos hasta el momento y determinar si existen dianas moleculares que hasta el momento no hayan sido susceptibles de tratamiento.

B.2.6. Explorar la asociación de determinadas variantes genéticas con los fenotipos clínicos de la enfermedad

C. Enfoque y método a seguir

El trabajo consistirá en el análisis de muestras de pacientes diagnosticados de esclerosis múltiple remitente recurrente en el Centre d'Esclerosi Múltiple de Catalunya (CEMCAT) y controles sanos. A estos sujetos se les realizó un genotipado de polimorfismos de un solo nucleótido usando el MS Chip, un array de genotipado diseñado para detectar variantes de la enfermedad. El array fue diseñado por el consorcio internacional de la genética de la EM (IMSGC) con más de 90.000 marcadores genéticos de interés, además de otras 200 variantes asociadas en el pasado con la enfermedad, a través de la tecnología Illumina Infinum que permite realizar arrays a la carta. El procesamiento de las muestras fue realizado en el CEMCAT Los datos, de acuerdo con la ley orgánica de protección de datos 3/2018 se encuentran totalmente

anonimizados, siéndoles asignado un identificador aleatorio a cada sujeto.

Para realizar el estudio se realizará un análisis de las variantes genómicas detectadas y se compararán con los controles para determinar aquellas asociadas a la enfermedad. Posteriormente se realizará un análisis comparativo con las variantes previamente descritas en la literatura para describir las nuevas variantes encontradas. Se consultarán diferentes bases de datos genómicas para determinar la ontología y función de dichas variantes haciendo un análisis de pathway y línea celular asociada a las nuevas variantes.

El análisis se realizará empleando R y RStudio (software libre) por lo que serán necesarias licencias extras ni prever un presupuesto adicional de gastos.

D. Planificación del trabajo

D.1. Tareas

- *Recopilar información sobre las variantes asociadas a EM hasta el momento.*
  - Duración: 3 días
  - Descripción: se realizará una revisión bibliográfica en diferentes fuentes de bases de datos (Pubmed, OMIM...) para encontrar una base de datos estructurada donde se reporten la lista de genes asociada a EM en el pasado
  - Personas implicadas: AVC[1]
  - Objetivos específicos asociados a la tarea: B.2.1, B.2.2, B.2.3 y B.2.6

- *Obtención de los datos y determinar las infraestructuras necesarias para el análisis y transmisión de la información.*

[1] Adrián Valls Carbó

- Duración: 3 días
- Descripción: se contactará con el centro que proporciona los datos (CEMCAT), el dr. Comabella y el tutor para determinar como se transmitirá la información y como se procesarán los datos para preservar la confidencialidad de los datos.
- Personas implicadas: AVC; MCL[2]; ASP[3]
- Objetivos específicos asociados a la tarea: tarea técnica preliminar (todas).

- *Exploración, análisis de calidad y pre-procesado de los datos.*
  - Duración: 4 días
  - Descripción: se realizará un análisis exploratorio de los datos para detectar posibles fallos en la adquisición de estos. Aquellos registros de calidad subóptima serán tratados adecuadamente, realizando un registro continuo de los datos perdidos. Los datos serán pre-procesados para proseguir con el análisis.
  - Personas implicadas: AVC; ASP
  - Objetivos específicos asociados a la tarea: B.2.1, B.2.2 y B.2.3

- *Selección de genes diferencialmente expresados en la muestra.*
  - Duración: 4 días
  - Descripción: mediante los datos pre-procesados de la tarea anterior, se realizará el contraste de hipótesis de expresión diferencial de los diferentes genes contemplados en el array y mediante técnicas de corrección del valor de p se seleccionarán aquellos genes expresados diferencialmente
  - Personas implicadas: AVC; ASP
  - Objetivos específicos asociados a la tarea: B.2.1, B.2.2 y B.2.3

- *Análisis comparativo de lista de la lista de genes (con respecto a genes asociados previamente a la enfermedad).*
  - Duración: 4 días
  - Descripción: Se realizará un análisis comparativo de las listas de genes comparando con aquellos genes descritos

[2] Manel Comabella López
[3] Alexandre Sànchez Pla

como alterados en la enfermedad, para determinar si existen en la muestra genes no descubiertos hasta el momento.
    - Personas implicadas: AVC; ASP
    - Objetivos específicos asociados a la tarea: B.2.1, B.2.2 y B.2.3

- *Análisis ontológico, de significación, función biológica y pathways de la lista de genes.*
    - Duración: 4 días
    - Descripción: mediante la lista de genes se realizará un análisis de la ontología de estos para determinar si existen patrones de funciones biológicas afectadas, rutas celulares o tejidos especialmente afectados en la enfermedad
    - Personas implicadas: AVC; ASP
    - Objetivos específicos asociados a la tarea: B.2.4 y B.2.5

- *Interpretación de los resultados y comparación de los hallazgos con lo publicado en la literatura hasta el momento.*
    - Duración: 5 días
    - Descripción: se revisará la literatura de lo publicado hasta el momento y se compararán los hallazgos con estudios previos para determinar la existencia de nuevas funciones biológicas o tejidos que pudieran encontrarse afectadas en la enfermedad
    - Personas implicadas: AVC; ASP; MCL
    - Objetivos específicos asociados a la tarea: B.2.1 a B.2.6

- *Redacción de los hallazgos.*
    - Duración: 14 días
    - Descripción: se redactarán los resultados encontrados en la memoria y se comenzará la redacción del artículo científico
    - Personas implicadas: AVC
    - Objetivos específicos asociados a la tarea: B.2.1 a B.2.6

- *Realización de gráficas.*
    - Duración: 3 días
    - Descripción: se realizarán gráficas e infografías en diferentes formatos para expresar de forma visual los datos encontrados en el estudio
    - Personas implicadas: AVC

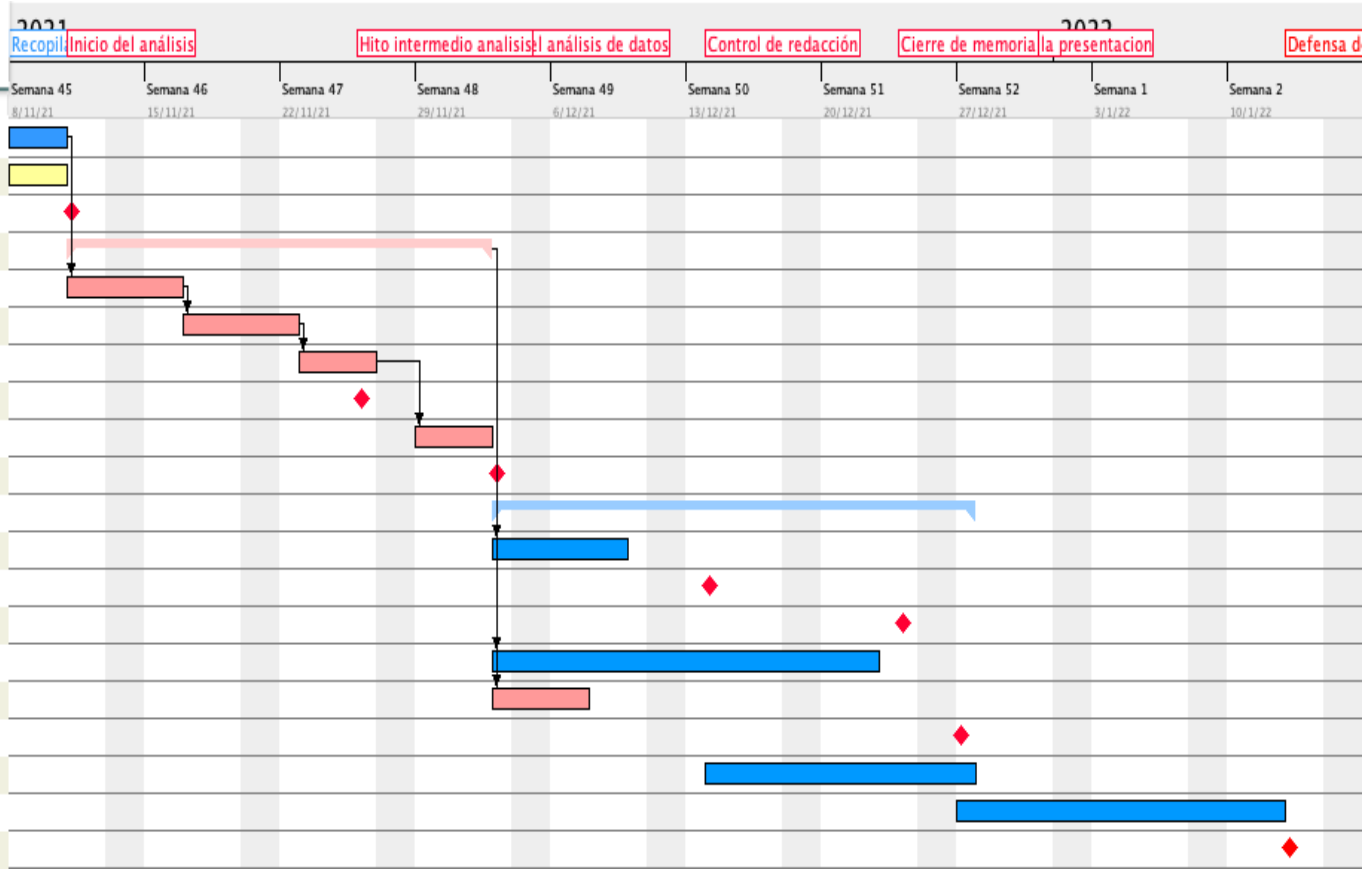
- Objetivos específicos asociados a la tarea: B.2.1 a B.2.6

- *Realización de presentación.*
  - Duración: 13 días
  - Descripción: se confeccionarán las presentaciones para realizar la exposición oral del trabajo.
  - Personas implicadas: AVC
  - Objetivos específicos asociados a la tarea: B.2.1 a B.2.6

- *Preparación para presentación.*
  - Duración: 13 días
  - Descripción: preparación de la presentación pública de los resultados y ultimar los detalles de los datos.
  - Personas implicadas: AVC
  - Objetivos específicos asociados a la tarea: B.2.1 a B.2.6

D.2. Calendario

Nombre	Fecha de inicio	Fecha de fin
Recopilar información	8/11/21	10/11/21
Infraestructuras y recopilar datos	8/11/21	10/11/21
Inicio del análisis	11/11/21	11/11/21
Análisis de los datos	11/11/21	2/12/21
Calidad y preprocesado	11/11/21	16/11/21
Genes expresados diferenci...	17/11/21	22/11/21
Análisis de lista de genes	23/11/21	26/11/21
Hito intermedio analisis	26/11/21	26/11/21
Analisis ontológico, funcion y ...	29/11/21	2/12/21
Fin del análisis de datos	3/12/21	3/12/21
Redacción de resultados	3/12/21	27/12/21
Interpretación de resultados	3/12/21	9/12/21
Control de redacción	14/12/21	14/12/21
Cierre de memoria	24/12/21	24/12/21
Redacción de los hallazgos	3/12/21	22/12/21
Realización de gráficas	3/12/21	7/12/21
Entrega de la presentacion	27/12/21	27/12/21
Realización de presentación	14/12/21	27/12/21
Preparación para exposición	27/12/21	12/1/22
Defensa de presentación	13/1/22	13/1/22



D.2.1. Hitos

- Hito 1: Inicio del análisis.
  - Fecha: 11 de noviembre de 2021.
  - Objetivo: se comenzará el análisis bioinformático de los datos.
  - Elementos necesarios: datos del estudio GWAS
  - Cambios propuestos en caso de desviación: solicitar urgentemente los datos

- Hito 2: Seguimiento intermedio del análisis.
  - Fecha: 26 de noviembre de 2021.
  - Objetivo: se realizará seguimiento del análisis realizado hasta el momento con el tutor para aclarar dudas y problemas en el análisis de los datos.
  - Elementos necesarios: documento de resultados preliminares del análisis de los datos. Lista de genes expresados diferencialmente.
  - Cambios propuestos en el caso de desviación: en caso de no obtener nuevos genes expresados diferencialmente se volverá a realizar el análisis con las indicaciones del tutor.

- Hito 3: Fin del análisis de los datos
  - Fecha: 3 de diciembre de 2021
  - Objetivo: Se expondrán los resultados obtenidos hasta el momento, se comentarán las dudas y se expondrán los errores en el análisis en el caso de existirlos para su posterior corrección
  - Elementos necesarios: documento de resultados finales. Lista de anotaciones de pathways y ontología.
  - Cambios propuestos en caso de desviación: en caso de no obtener nuevas anotaciones se consultará con el tutor y se realizarán las correcciones oporturnas en el análisis.

- Hito 3: Control de redacción.
  - Fecha: 14 de diciembre de 2021
  - Objetivo: Se expondrá el borrador de los resultados y la memoria del proyecto
  - Elementos necesarios: borrador del documento de memoria.

    - Cambios propuestos en caso de desviación: se alargarán 3 días extra para la redacción

- Hito 4: Cierre de memoria
    - Fecha: 24 de diciembre de 2021
    - Objetivo: se entregará la memoria del trabajo realizado hasta el momento
    - Elementos necesarios: documento de memoria de TFM
    - Cambios propuestos: se realizarán los cambios oportunos según el tipo de fallo en el cierre de la memoria.

- Hito 5: Entrega de presentación
    - Fecha: 27 de diciembre de 2021
    - Objetivo: se entregará la presentación para la exposición final del trabajo.
    - Elementos necesarios: documento de presentación del trabajo.
    - Cambios propuestos: se anotarán los cambios de cara a la presentación para la defensa pública del trabajo.

- Hito 6: Defensa de presentación
    - Fecha: 13 de enero de 2022
    - Objetivo: presentación pública de los datos obtenidos
    - Elementos necesarios: documento de presentación del trabajo

D.2.2. Análisis de riesgos

Los riesgos para la puesta en marcha y consecución del proyecto son los siguientes:

- Tiempo: dado el tiempo limitado en la consecución del proyecto y la colaboración de diferentes partes (CEMCAT, estudiante de TFM y tutor) un riesgo importante es el retraso en la entrega de los diferentes materiales del TFM. Aspectos que minimizan el riesgo en este caso es la presencia de unos datos ya obtenidos de la cohorte de pacientes, lo que hace que el retraso en la obtención de los datos sea mínimo. Además, el hecho de que todo el análisis y redacción de la información solo dependa del alumno hace que

la probabilidad de retraso por falta de comunicación entre las partes sea menor.

- Infraestructura: los datos de más de 90.000 sondas procedentes de unos cientos de pacientes hacen que no todos los sistemas de computación puedan procesar esta información adecuadamente en un tiempo razonable. Muchos de los análisis de comparación múltiple requieren capacidad de memoria RAM y velocidad de procesamiento elevada. Aunque el estudiante dispone de un ordenador Mac con CPU 2.3 GHz Intel Core i7 de 4 núcleos y RAM de 16 GB, pueden ser necesarios otros recursos. Sin embargo, existen en la unidad de bioestadística y bioinformática del Hospital Universitario Vall d'Hebron recursos disponibles y suficientes para el análisis de los datos.

- Fallo en la calidad de los datos: aunque los datos han sido obtenidos con gran calidad por técnicos de laboratorio expertos, cabe la posibilidad de que haya habido errores durante el procesamiento de estos que haga que no sean válidos para el análisis. El mayor riesgo que existe es que la cantidad de datos de alta calidad tan bajo que no permita obtener conclusiones sobre el estudio. No obstante, este riesgo es bajo dado que los datos ya han sido examinados previamente en otros estudios con anterioridad, mostrando un porcentaje bajo de pérdidas.

- Tamaño muestral bajo para el descubrimiento de nuevas variantes: los últimos trabajos para el descubrimiento de variantes cuentan con muestras que rondan los 116.000 sujetos en total (47.429 pacientes y 68.374 controles)[9], un tamaño significativamente superior al de nuestro estudio. Sin embargo, se trata de estudios multicéntricos, con los inconvenientes que esto conlleva (heterogeneidad de los criterios de inclusión, del procesamiento de las muestras...) y además el objetivo de estos estudios era algo diferente. Mientras que en los grandes estudios de asociación genética el objetivo era encontrar las variantes asociadas a la enfermedad en la población mundial, el objetivo de nuestro estudio pretende circunscribirse en delimitar las variantes en una población del sur de Europa. Otro posible riesgo de nuestro estudio es el no haber calculado de antemano el tamaño muestral para encontrar las diferencias que pretendemos buscar, por lo que la potencia del análisis puede ser inferior a la necesaria.

- Problemas en el diseño del array: el array fue diseñado de forma específica para detectar variantes genéticas que posiblemente pudieran asociarse a EM. Para ello se determinaron una serie de loci que previsiblemente estarían alterados en la enfermedad. Sin embargo, estos loci fueron seleccionados de antemano, por lo que un riesgo posible es que aquellos loci de interés en nuestro estudio no hayan sido seleccionados por el array y por lo tanto no podamos encontrarlos diferencialmente expresados. Este riesgo es inherente a los estudios realizados mediante microarrays y es relativamente menor, dado que la selección de sondas fue realizada por un consorcio de expertos.

D.2.3. Breve sumario de productos obtenidos

Se espera como resultado del trabajo obtener:

- Memoria: donde se recopilarán el método, resultados y conclusión del trabajo realizado.
- Presentación virtual: el documento recogerá la introducción al tema, métodos, resultados y discusión de un modo gráfico para la audiencia.

2

Metodología

A. Muestras del estudio

Para el estudio se disponen de 464 muestras del Centre d'Esclerosi Multiple de Catalunya (CEMCAT), de los cuales 232 pacientes que cumplían criterios de McDonald para el diagnóstico de esclerosis múltiple y 232 muestras procedían de controles sanos.

Las muestras de ADN fueron obtenidas a partir de extracción de sangre completa. Las muestras de ADN fueron genotipadas mediante el MS chip, un array personalizado de Illumina. El MS Chip fue diseñado por el International Multiple Sclerosis Genetics Consortium (IMSGC) conteniendo >90.000 SNPs, incluyendo 200 variantes asociadas a la enfermedad conocidas hasta el momento y otros marcadores genéticos asociados a la enfermedad.

El genotipado de las muestras fue realizado en el Center for Genome Technology del Instituto John P. Haussman para la genómica humana de la universidad de Miami mediante el software GenomeStudio v2.0.

B. Control de calidad

B.1. Control de calidad de las muestras

De acuerdo con artículos anteriores, para obtener muestras de calidad adecuada se emplearon los siguientes parámetros:

- **Ratio de genotipado correcto (genotype call rate):** aquellas muestras que presentaban menos del 98% de los polimorfismos de un solo nucleótido (SNP) correctamente genotipados fueron excluidos.

- **Discrepancia entre el sexo reportado y el genotipado:** se estimó el género genotipado mediante la heterocigosidad del cromosoma X. Aquellos sujetos que presentaban una heterocigosidad para el cromosoma X 3 veces superior o inferior a la desviación estándar a la media de su sexo reportado fueron excluidos del análisis.

- **Outliers de heterocigosidad autosómica:** las muestras que presentaban una heterocigosidad que se alejaba más de 3 desviaciones típicas de la media de heterocigosidad de la muestra, fueron eliminadas del análisis.

- **Muestras genéticamente relacionadas:** se estimó de cada muestra la identidad por descendencia mediante el método del momento descrito en el software PLINK[16]. Aquellas muestras que presentaban un parentesco superior a 0.1 fueron eliminadas del análisis. Se estimó el índice de endogamia, eliminando aquellas muestras que presentaban un índice superior a 0.1.

- **Muestras extremas:** se estimaron las 10 primeras componentes principales de cada una de las muestras. Aquellas muestras que se alejaban en alguna de las 10 primeras componentes principales más de 6 desviaciones estándar de la media de la componente principal fueron eliminadas del análisis.

B.2. Control de calidad del genotipado

Se realizó un control de calidad del genotipado de las muestras eliminando de los análisis aquellos genotipados que no cumplían los siguientes criterios de calidad:

- **Bajo ratio de genotipado:** se eliminaron aquellos SNPs que tenían:
    - *Ratio de genotipado <99.5% y cuya frecuencia del alelo minoritario (MAF) era menor o igual al 5%*
    - *Ratio de genotipado <99% y MAF entre el 5 y 10%*
    - *Ratio de genotipado <98% y MAF superior al 10%*

- **Valores extremos en el equilibrio de Hardy Weinberg:** se estimó el equilibrio de Hardy Weinberg para cada uno de los SNPs en los controles. Aquellos SNPs que presentaban un valor inferior a $p<0.000001$ fueron eliminados del análisis.

C. Análisis estadístico

Para el análisis de los datos se ajustó una regresión logística con cada uno de los SNPs seleccionados, ajustando como covariables por el sexo y las 10 primeras componentes principales como proxy de la variabilidad genómica, intentando predecir el estatus del paciente.

Se tomó como punto de corte para considerar estadísticamente significativo un SNP el ajuste de Bonferroni. Dado que se realizaron N contrastes, el punto de corte del valor de p escogido para la muestra fue de $5 \cdot 10^{-7}$. Todo el control de calidad, análisis de los datos y representación gráfica fue realizado en lenguaje R[17] y la plataforma RStudio[18].

D. Análisis de listas de SNPs y de genes

De aquellos SNPs que superaron el punto de corte del análisis se buscó el gen al que pertenecían en bases de datos accesibles a través de la web

(National Center for Biotechnology Information [NCBI] y Ensembl). Solo de aquellos SNPs que se encontraban dentro de un gen conocido hasta el momento se extrajo el gen en cuestión. De los SNPs se obtuvieron también las enfermedades y las publicaciones en las que se encontraron asociaciones con dichos SNPs.

Además, se buscaron en la lista de SNPs proporcionada en el último artículo de revisión disponible sobre los genes de la enfermedad.

E. Análisis de ontologías y pathways

De aquellos genes obtenidos en el paso anterior se realizó un estudio de la ontología y los pathways en los que estos genes se encuentran involucrados. Para el análisis de ontologías se realizó una búsqueda manual de los procesos biológicos asociados a dichos genes en la base de datos proporcionada por GeneOntology[4].

Para el análisis de enriquecimiento de las ontologías y pathways se empleó el paquete XGR[19] disponible en R y en web[5]. Este software permite la realización de análisis de enriquecimiento y de similaridad de los SNPs de forma directa, sin ser necesario buscar de forma manual cada uno de los genes asociados a los SNPs. Para calcular las distancias entre los diferentes SNPs se empleó el cálculo por pares de similaridad semántica de acuerdo con la "experimental factor ontology" (EFO[20]) descritos en el articulo de Fang[19].

[4] http://geneontology.org

[5] http://galahad.well.ox.ac.uk:3030

3

Resultados

A. Análisis del control de calidad

A.1. Análisis de la calidad de las muestras

De un total de 464 muestras, se eliminaron 47 de ellas por presentar criterios de calidad bajos de acuerdo con lo especificado en apartados anteriores. Se eliminaron del análisis 1 muestra por bajo ratio de genotipado, 20 muestras por presentar discrepancia entre el sexo reportado y el sexo genómico (figura 1), 7 muestras por ser outliers de heterocigosidad, 9 muestras por encontrarse genéticamente relacionadas y 10 muestras por presentar valores extremos en alguno de los componentes principales. La figura 2 muestra los componentes principales tras la eliminación de las muestras que no cumplían los criterios de calidad. Se aprecia que existe un cluster homogéneo, si bien algunas de las muestras presentan cierto grado de divergencia. En total para el análisis se emplearon 417 muestras (210 pacientes, 207 controles).

A.2. Análisis de la calidad de los SNPs

Se genotiparon inicialmente 319.950 SNPs. Se eliminaron 5.511 SNPs por presentar un bajo ratio de genotipado y 1.512 SNPs por presentar valores extremos en el equilibrio de Hardy Weinberg (figura 3). De forma global fueron analizados en el estudio 312.927 SNPs y se eliminaron 7.023 del análisis.

F. SNPs y genes asociados a enfermedad

Del estudio se encontraron 11 SNPs que se asociaron de forma significativa a esclerosis múltiple, localizados todos ellos en el cromosoma 6 en la región comprendida entre 32.399.240 y 32.445.768 pb, abarcando 46.528 pares de bases (Figuras 5 y 6).

De estos SNPs, solo dos de ellos han sido descritos con anterioridad en otros estudios asociándolos con la enfermedad en cuestión, aunque ninguno fue detectado en el mayor estudio de asociación genética de la enfermedad hasta el momento excepto la asociación con el HLA DRB1*1501.

No se encontraron asociaciones significativamente estadísticas con el umbral de p seleccionado en otros loci fuera de la región del complejo mayor de histocompatibilidad.

La tabla 1 muestra los SNPs que mostraron una significación estadística con la enfermedad.

Cuando estudiamos el desequilibrio de ligamiento de los polimorfismos significativos, se aprecia que debido a su proximidad genética estos se segregan de forma conjunta (figura 7), presentando valores altos de desequilibrio de ligamiento (mínimo de 0.6, máximo 1, mediana 0.95). Esto se aprecia en la figura 8.

En el estudio de la localización de los SNPs significativos, encontramos que estos SNPs se encotraban en 2 genes diferentes: BTNL2 y HLA-DRA. Además, el polimorfismo rs3117116 se encuentra sobre en gen antisentido de TSBP1 (figura 9).

SNP	Posición	-log(p)	OR	Alelos	Gen	Asociación conocida EM	Asociación con otras enfermedades
rs3117116	Chr6:32.399.240	6.86	2.86 [3.23-6.96]	A|G	BTNL2	Sí	Sarcoidosis, CU, AR, miositis, DM 1, LES, SCA y cáncer de próstata
rs3135352	Chr6:32.425.129	6.83	2.93 [3.26-7.13]	A|C	-	No	Déficit de IgA
rs3135350	Chr6:32.425.204	6.85	2.99 [3.28-7.29]	A|G	-	No	IgG antigliadina
rs3129971	Chr6:32.425.458	6.83	2.93 [3.26-7.13]	C|G	-	No	IgG antigliadina
rs3129860	Chr6:32.433.302	6.86	2.86 [3.23-6.96]	G|A	-	No	Cáncer de pulmón en asiáticos
rs3129865	Chr6:32.436.271	6.86	2.86 [3.23-6.96]	G|C	-	No	-
rs3129868	Chr6:32.436.600	6.86	2.86 [3.23-6.96]	C|A	-	No	Anticoagulante lúpico en SAF
rs9268635	Chr6:32.438.802	6.83	2.93 [3.25-7.13]	G|A	HLA-DRA	No	-
rs7197	Chr6:32.444.803	7.17	2.28 [3.01-5.58]	G|A	HLA-DRA	No	Uveítis anterior
rs3135388	Chr6:32.445.274	6.99	2.99 [3.29-7.23]	G|A	HLA-DRA (DRB1*1501)	Sí	-
rs3129889	Chr6:32.445.768	6.99	2.99 [3.29-7.23]	A|G	HLA-DRA (DRB1*1501)	Sí	-

Tabla 1: Polimorfismos genéticos asociados a esclerosis múltiple. SNP=polimorfismo de un solo nucleótido, -log(p)= -logaritmo en base 10 del valor de p, EM= esclerosis múltiple, OR=odds ratio, CU=colitis ulcerosa, AR=artritis reumatoide, DM1=diabetes mellitus 1, SAF=síndrome antifosfolípido, LES=lupus, SCA=síndrome coronario agudo, HLA=antígeno leucocitario de histocompatibilidad. En la columna de alelos, la primera letra representa el alelo salvaje y la segunda representa el alelo asociado a enfermedad.

G. Análisis de ontología y pathways

Las ontologías asociadas a procesos biológicos de los genes significativos se muestran en la tabla 2. La figura 11 muestra las relaciones de acuerdo a las distancias de similaridad semántica de los términos EFO de cada polimorfismo significativo. Se aprecia como el SNP rs3135350 se encuentra fuera del cluster del resto de polimorfismos. La figura 12 muestra las relaciones de los SNPs significativos con en resto de los SNPs genotipados en el array.

En el estudio de los genes asociados a los SNPs de acuerdo con la distancia y el valor de p de dichos polifmorfismos se encontraron 7 genes asociados: HLA-DRA (p=0.42), BTNL2 (p=0.83), HCG23 (p=0.89), C6orf10 (p=0.97), HLA-DRB6 (p=0.99), HLA-DRB5 (p=0.99), HLA-DQA1 (p=0.99), si bien ninguno de ellos alcanzó la significación estadística.

No se encontró ninguna ontología ni vía enriquecida en el análisis de enriquecimiento realizado usando los polimorfismos o los genes encontrados.

Gen	Términos GO	Descripción
BTNL2	GO:0001817	Regulación en la producción de citosinas
	GO:0050776	Regulación de la respuesta inmune
	GO:0050852	Vía de señalización del receptor T
HLA-DRA	GO:0002250	Respuesta inmune adaptativa
	GO:0002503	Ensamblado de péptido con HLA de clase II
	GO:0002504	Procesamiento y presentación de antígenos en HLA de clase II
	GO:0006955	Respuesta inmune
	GO:0016032	Proceso viral
	GO:0019886	Procesamiento y presentación de antígenos exógenos en HLA de clase II
	GO:0050890	Cognición
	GO:0060333	Vía de señalización mediada por interferón gamma

Tabla 2: Ontología de los procesos biológicos asociados a los genes encontrados.

4

Discusión

En el presente estudio se han encontrado 11 polimorfismos diferentes asociados a la EM, localizados todos ellos en el cromosoma 6, en una región adyacente al complejo mayor de histocompatibilidad. En una búsqueda de los genes asociados a estos loci se han encontrado 2 genes (BTNL2 y HLA-DRA), los cuales todos ellos han sido relacionados con la enfermedad.

- **Gen HLA-DRA:** se trata del gen que codifica la cadena alfa del complejo mayor de histocompatibilidad en humanos (HLA) de tipo II encargado de la presentación de antígenos exógenos por parte de células presentadoras de antígenos a linfocitos. El HLA de clase II tiene diferentes moléculas, HLA-DR, DP y DQ. En el caso del HLA-DR, la molécula se trata de un heterodímero compuesto por dos cadenas, una alfa y otra beta. La cadena alfa solo está codificada por un gen (HLA-DRA), mientras que la cadena beta está codificada por diferentes genes (HLA-DRB1, HLA-DRB3, HLA-DRB4, HLA-DRB5). Según la combinación de cadenas beta que exprese el individuo tendremos los diferentes haplotipos de HLA-DR disponible en humanos. De forma general, casi todos los haplotipos presentan un gen funcional HLA-DRB1 y otro más entre el resto de cadenas beta. Las cadenas más polimórficas del HLA suelen ser las cadenas beta, y es en ellas donde encontramos la mayoría de mutaciones que se asocian a distintos rasgos fenotípicos.

  Los polimorfismos en los genes del HLA se han asociado con múltiples enfermedades autoinmunes en el pasado[21]. Dado su papel en la interacción y reconocimiento por parte de las células del sistema inmune adaptativo de antígenos exógenos, existe una plausibilidad biológica en la que alteraciones en el reconocimiento de estos antígenos podría justificar la activación de las células del

sistema inmunitario para que atacasen células del propio individuo, desencadenando enfermedades autoinmunes.

En el caso de la EM, la asociación de la enfermedad con polimorfismos en la región del HLA es de sobra conocida[9]. De hecho, se considera que la región del HLA alberga el principal locus de susceptibilidad para la EM[22], recogiendo el 10.5% del riesgo genético de la enfermedad[10]. Concretamente la asociación de la enfermedad con el haplotipo HLA-DRB1*1501 se sabe que comporta una OR alrededor de 3.08.

Nuestros datos se encuentran en línea con la literatura previa por diversos motivos. Por un lado, tal como podemos ver en la figura 6, el grueso de polimorfismos que presentan valores bajos de p se encuentran alrededor de la región del HLA en el cromosoma 6. De estos, podemos apreciar como solo unos pocos alcanzan la significación estadística. Por otro lado, aunque los polimorfismos encontrados relacionados con el HLA se encuentran localizados en el gen de la cadena alfa del HLA-DR, una región mucho más conservada, esto no va en contra de la literatura. El polimorfismo *rs3135388* ya se ha asociado en el pasado con la EM[23–25], confiriendo el alelo A una OR entre 1.99 a 2.75 para el desarrollo de la enfermedad. En diversos estudios se ha demostrado que existen altos niveles de desequilibrio de ligamiento entre SNPs fuera de genes del HLA con alelos de la cadena beta del HLA, por lo que estos polimorfismos pueden suponer marcadores subrogados de determinados HLA[26]. En este caso, el polimorfismo *rs3135388* y *rs3129889* a pesar de encontrarse en el gen del HLA-DRA presenta una alta correlación con el haplotipo HLADRB1*1501, un conocido marcador asociado a la enfermedad en estudio[27–30].

Respecto a los otros polimorfismos localizados en el gen HLA-DRA existen datos que relacionan el polimorfismo *rs7197* con la uveítis anterior[31], la cual constituye una forma de presentación de múltiples enfermedades autoinmunes

No existen publicaciones al respecto de *rs9268635*.

- **Gen Butyrophilin-Like Molecule (BTNL2): el** gen BTNL2 es un gen ligado al complejo mayor de histocompatibilidad de tipo II que codifica para una proteína transmembrana, compuesta por dos cadenas con homología con las inmunoglobulinas. Esta proteína se encarga de inhibir la proliferación de células T y la activación del receptor de la las células T por el factor nuclear, por lo que se ha implicado en la regulación de linfocitos intraepiteliales intestinales[32,33]. Mutaciones en este gen se han asociado con enfermedades como la sarcoidosis[34], artritis reumatoide[35,36], colitis ulcerosa[37,38], miositis por cuerpos de inclusión[39], diabetes mellitus tipo 1[35], lupus eritematoso sistémico[35], síndrome coronario agudo[40] y cáncer de próstata[41]. Aunque este gen fue relacionado inicialmente con la EM, en un estudio[42] se demostró que dicha asociación era mediada por el desequilibrio de ligamiento de este gen con el HLA-DRB1*1501[27]. Parte de la secuencia del gen es compartida con el ARN antisentido del gen TSBP-1

- **Región intrónica 6: 32.407.128- 32.439.887:** Se trata de la región que se encuentra entre el gen BTNL2 y el HLA-DRA. En esta región han sido 6 los polimorfismos asociados a EM, los cuales detallaremos a continuación.

    - *rs3135352:* en un estudio se le ha asociado como factor protector del déficit de IgA, aunque en este mismo estudio se le empleaba como marcador subrogado del haplotipo DRB1*1501[43]. No se han encontrado otras asociaciones con la EM.

    - *rs3135350:* se encontró una asociación estadísticamente significativa en pacientes con aumento de la IgG antigliadina tras el consumo de trigo[44]. No se han encontrado asociaciones con EM.

    - *rs3135391:* ha sido empleado en algunos estudios como marcador de HLA-DRB1*1501 por su gran desequilibrio de ligamiento con este haplotipo. Por ejemplo en este estudio[45] sobre variables de respuesta al copaxone en pacientes con EM, se empleó este SNP como marcador de la presencia de HLA-DRB1*1501. En otro estudio se asoció este polimorfismo con la respuesta de IgG a la gliadina[44].

- *rs3129860:* se ha relacionado con la susceptibilidad a padecer cáncer de pulmón en población asiática por su asociación con el HLA DQB1*0401[46]. Además se ha asociado este polimorfismo con la respuesta frente al virus de Epstein Barr, en concreto con la presencia de anticuerpos frente al antígeno nuclear del virus (EBNA-1)[47] .

- *rs3129865*: no se encontraron publicaciones asociadas

- *rs3129868:* se le ha asociado con la presencia de anticoagulante lúpico en pacientes con síndrome antifosfolípido[48] así como con el lupus eritematoso sistémico[49].

En definitiva, los hallazgos del estudio parecen indicar que, en la población estudiada, la asociación de la EM con la región del complejo mayor de histocompatibilidad, especialmente con el alelo HLA-DR1*1501, es consistente.

Los datos del presente estudio deben ser tomados con cautela dadas las limitaciones del diseño experimental y del análisis. En primer lugar, hemos de tener en cuenta el bajo tamaño muestral para un estudio de estas características. Para el control de la tasa de falsos negativos en un estudio GWAS en el que se realizan miles de tests, es preciso tomar límites estrictos del valor de p. En nuestro caso, dado el bajo tamaño muestral, el emplear la corrección de Bonferroni ha restado potencia al estudio a costa de reducir el porcentaje de falsos positivos. De cara a realizar el estudio no se calculó de antemano el tamaño muestral para alcanzar el poder estadístico deseado, lo cual es otra de las limitaciones del estudio.

Dada la baja potencia del estudio hemos sido incapaces de encontrar polimorfismos asociados a la enfermedad fuera de la región adyacente al complejo mayor de histocompatibilidad. Aunque ya ha sido demostrado en anteriores ocasiones la implicación de la esta región en la etiopatogenia de la EM, en este caso es posible que muchos de los loci encontrados como asociados a la enfermedad puedan atribuirse al desequilibrio de ligamiento, especialmente con HLA-DR1*1501. Hemos podido comprobar que muchos de estos loci significativos presentan una correlación elevada, por lo que conforman un haplotipo. Para poder estudiar si estos loci de forma independiente se asocian a la enfermedad son necesarios estudios de

segregación que permitan estudiar estos genes de forma aislada como causa de la enfermedad.

Las estrategias que se pueden tomar en el futuro para aumentar la potencia del estudio son varias. De una parte, es posible realizar una imputación de las variantes genómicas no tipadas en el estudio. La imputación genómica es un paso frecuente en los diversos estudios GWAS y puede en muchos casos aumentar la potencia estadística[50], aumentando el número de loci que se incluyen en el análisis al universo de loci conocidos en la actualidad en plataformas como 1000 Genomes[51]. Sin embargo, este paso tiene un coste computacional elevado y la realización e interpretación de los datos obtenidos en la imputación genómica requieren conocimiento específico que queda fuera del ámbito de este trabajo.

Otra solución es aumentar el número de muestras para el estudio, especialmente con muestras étnicamente similares. Si dispusiéramos de más muestras, en regiones adyacentes (otras regiones de España, Italia, Portugal o el sur de Francia) se podría aumentar la potencia estadística de los tests realizados.

Por otro lado, de cara a descubrir variantes genómicas asociadas a EM fuera de la región del complejo mayor de histocompatibilidad se pueden seguir algunas estrategias que mejoren el análisis. Una de ellas es el método empleado en el último GWAS realizado por el consorcio internacional de EM[9]. En este estudio se retiró una zona de 12Mbps alrededor de la región del complejo mayor de histocompatibilidad, y posteriormente se intentó realizar el análisis en el genoma restante. Tras esto, se identificó el polimorfismo más significativo del análisis, retirando 1Mbps alrededor de él, volviendo a repetir el análisis y retirando aquellas regiones con el polimorfismo más significativo de forma iterativa. Posteriormente en cada una de estas regiones identificadas como significativas, se realizó una regresión logística, añadiendo el SNP identificado en primer lugar. Al SNP más significativo de cada región se le denomina SNP de efecto. De forma iterativa vamos realizando el análisis en cada región, añadiendo el SNP de efecto significativo anterior hasta que el análisis no añada nuevos SNP significativos. Sin embargo, por cuestiones de limitación temporal, este análisis iterativo no pudo llevarse a cabo.

5

Conclusiones

A. Conclusiones

Planteadas las hipótesis iniciales, nuestro estudio no ha conseguido los objetivos que se planteaba.

Sin embargo, a pesar de las limitaciones que presenta el estudio, podemos llegar a unas conclusiones

- La región del complejo mayor de histocompatibilidad presenta una asociación significativa con la esclerosis múltiple

- La asociación de esta región en la población estudiada no difiere de la encontrada en otras cohortes de pacientes de países occidentales.

- Los genes asociados en nuestra cohorte con la enfermedad son el HLA-DRA y el BTNL-2 si bien no se puede descartar que esta asociación se deba al ligamiento con HLA-DRB1*1501, un conocido haplotipo asociado a la enfermedad.

Las causas de la no consecución de los objetivos son, como ya dijimos en la discusión, de causa metodológica y influyendo algunos de los riesgos que se mencionaron en la planificación del trabajo. Probablemente mejorando algunos aspectos del análisis nuevos locus podrían ser detectados con el mismo set de datos.

B. Líneas de futuro

De cara a mejorar el análisis en el futuro se plantean las siguientes propuestas:

- Aumentar el tamaño muestral recopilando datos en el centro y solicitando datos a otras regiones.

- Realizar imputación genómica de los datos para aumentar la potencia de los tests.

- Realizar el análisis empleando una estrategia estratificada por SNP más significativo, tal como detallamos anteriormente

C. Seguimiento de planificación

La planificación del estudio, aunque supuso una forma de organizar el tiempo disponible para la realización del análisis no fue posible de ser llevada a cabo en todos los hitos planteados. Las causas de esto son varias. Por un lado, dado el tiempo ajustado el planning en muchos casos fue inexacto. Aunque gran parte del tiempo planificado inicialmente estaba dirigido a la realización del análisis y la redacción de los hallazgos, finalmente gran parte del tiempo fue empleado en averiguar sobre los diferentes formatos de los datos y como aprender a leer las estructuras de los datos proporcionadas. La bibliografía disponible sobre los formatos y cómo realizar el análisis no es fácilmente accesible y requirió mucho tiempo de investigación.

Por otro lado, aunque los tutores siempre estuvieron disponibles cuando se requirió su ayuda, debido a lo ajustado de los tiempos, la comunicación entre el alumno y los tutores no fue siempre posible. Una mejor comunicación entre ambos podría haber acelerado el análisis y mejorado algunos aspectos.

6

Figuras

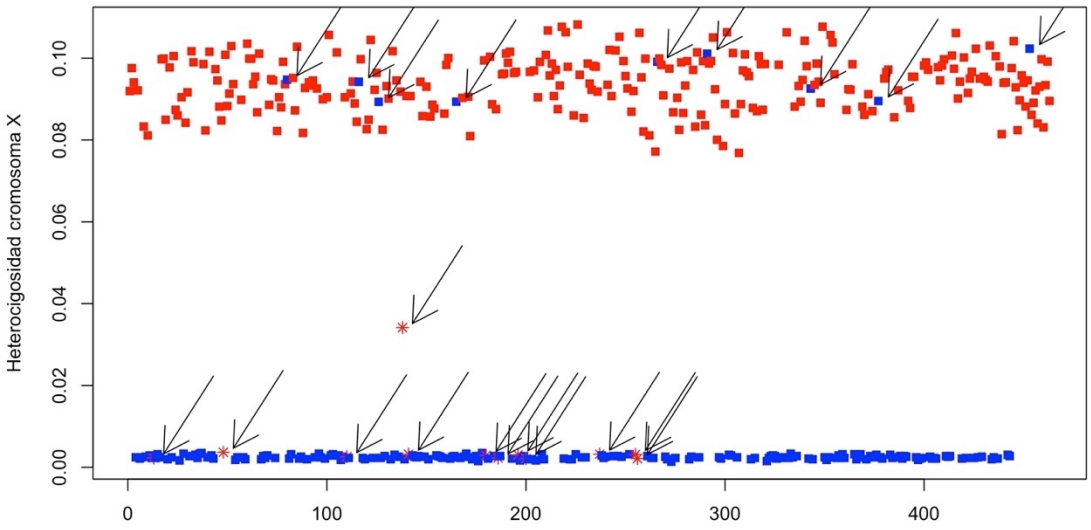


Figura 1: Outliers de heterocigosidad para el cromosoma X. Los cuadrados rojos representan muestras etiquetadas como procedentes de mujeres, mientras que los cuadrados azules representan a las muestras masculinas. Se aprecia cómo existen muestras tanto masculinas como femeninas que no cumplen los estándares de heterocigosidad para el cromosoma X.

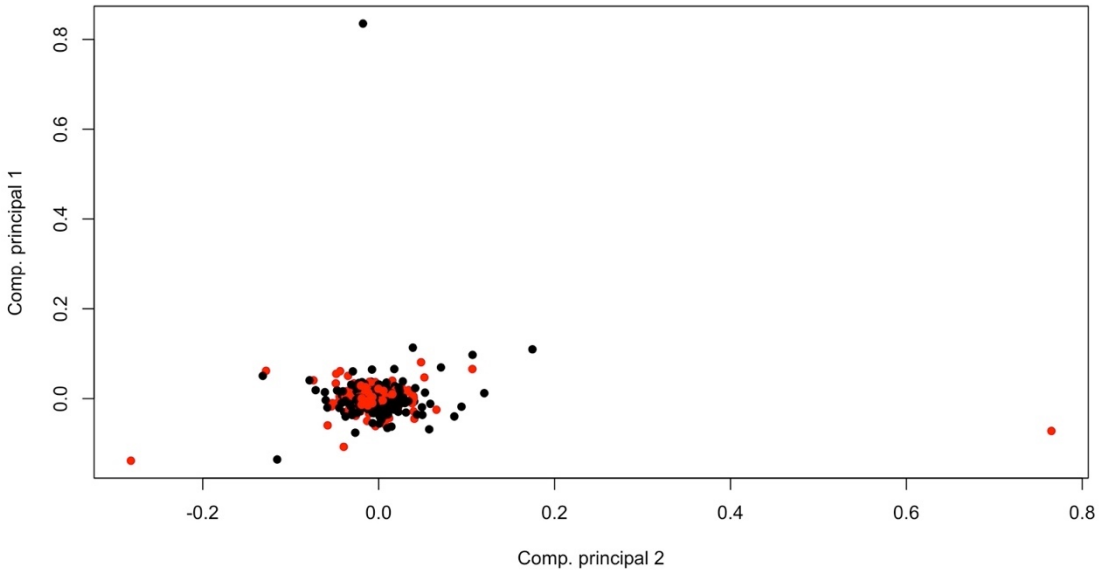


Figura 2: Gráfico de componentes principales tras el control de calidad. Los puntos rojos representan controles y los puntos negros representan casos de la enfermedad. Se aprecia como existe una relativa homogeneidad de los puntos, aunque existen todavía algunos outliers.

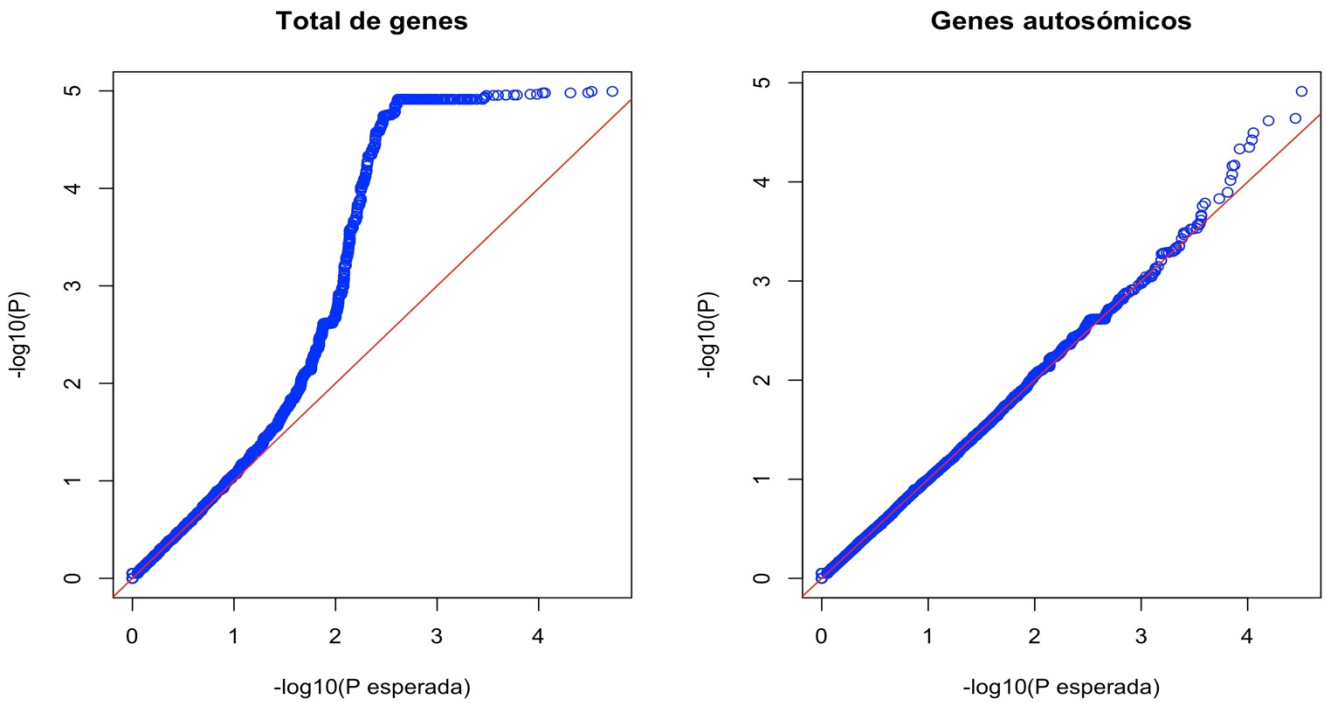


Figura 3: Equilibrio de Hardy Weinberg para todos los genes y para los genes autosómicos.

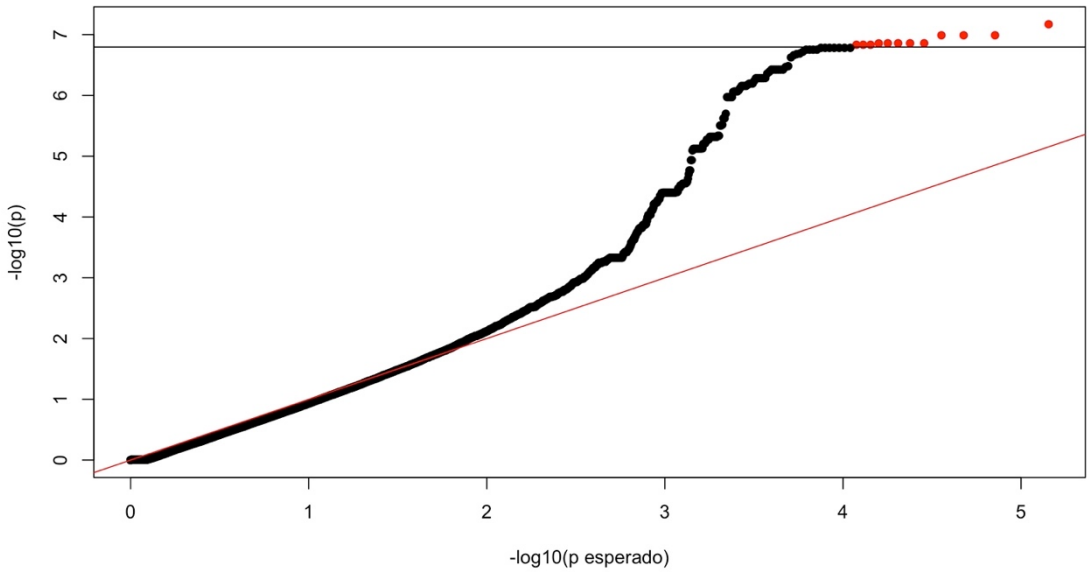


Figura 4: Valores de p teóricos frente a los observados en el análisis. Los valores mostrados en rojo representan aquellos polimorfismos que superan el umbral de significación definido anteriormente. Se aprecia como solo unos pocos valores alcanzan la significación estadística, si bien son muchos los valores que se alejan del valor teórico.

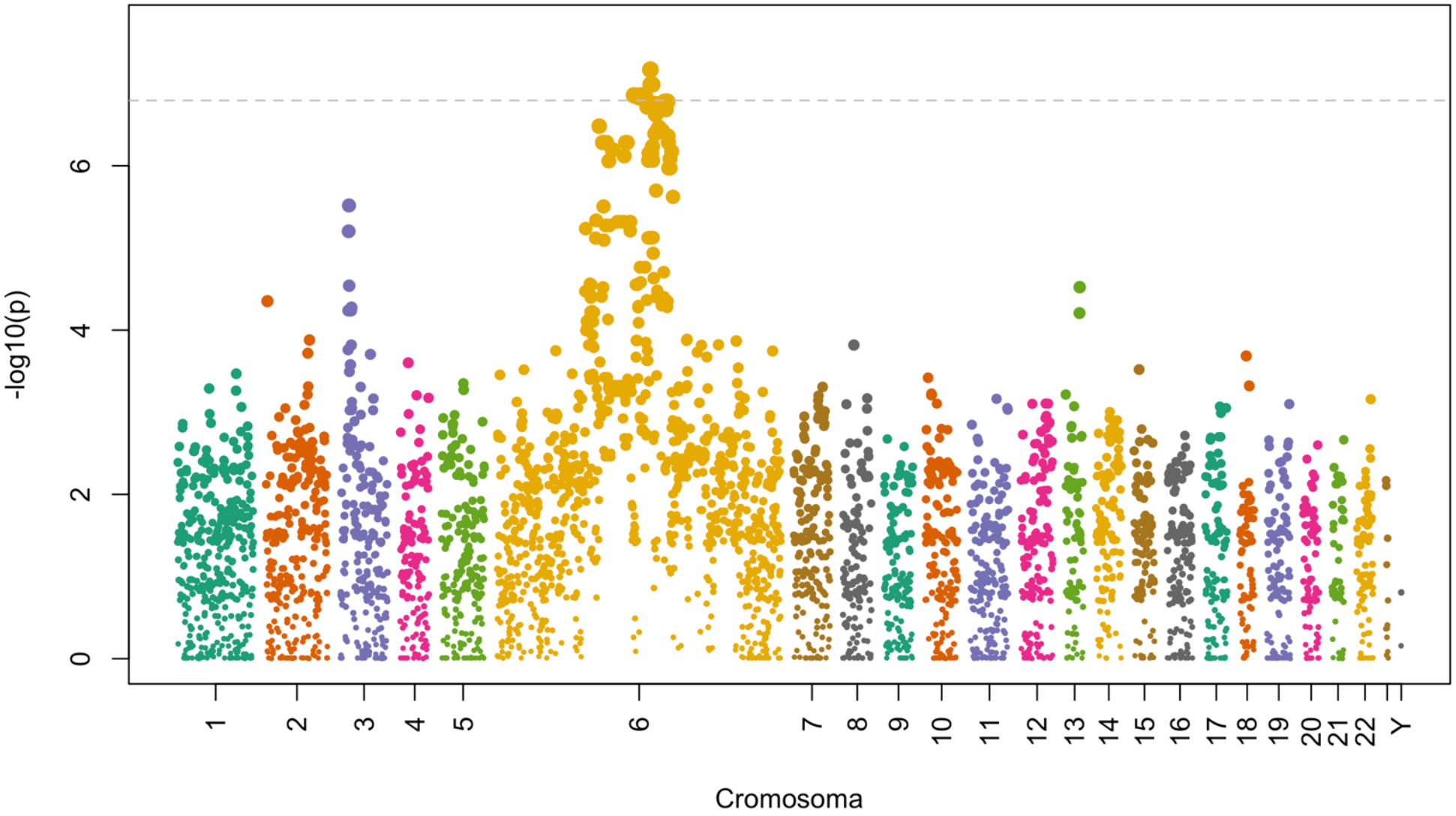


Figura 5: Manhattan plot de todos los loci estudiados. Se aprecia como solo alcanzan la significación estadística los loci localizados en el cromosoma 6.

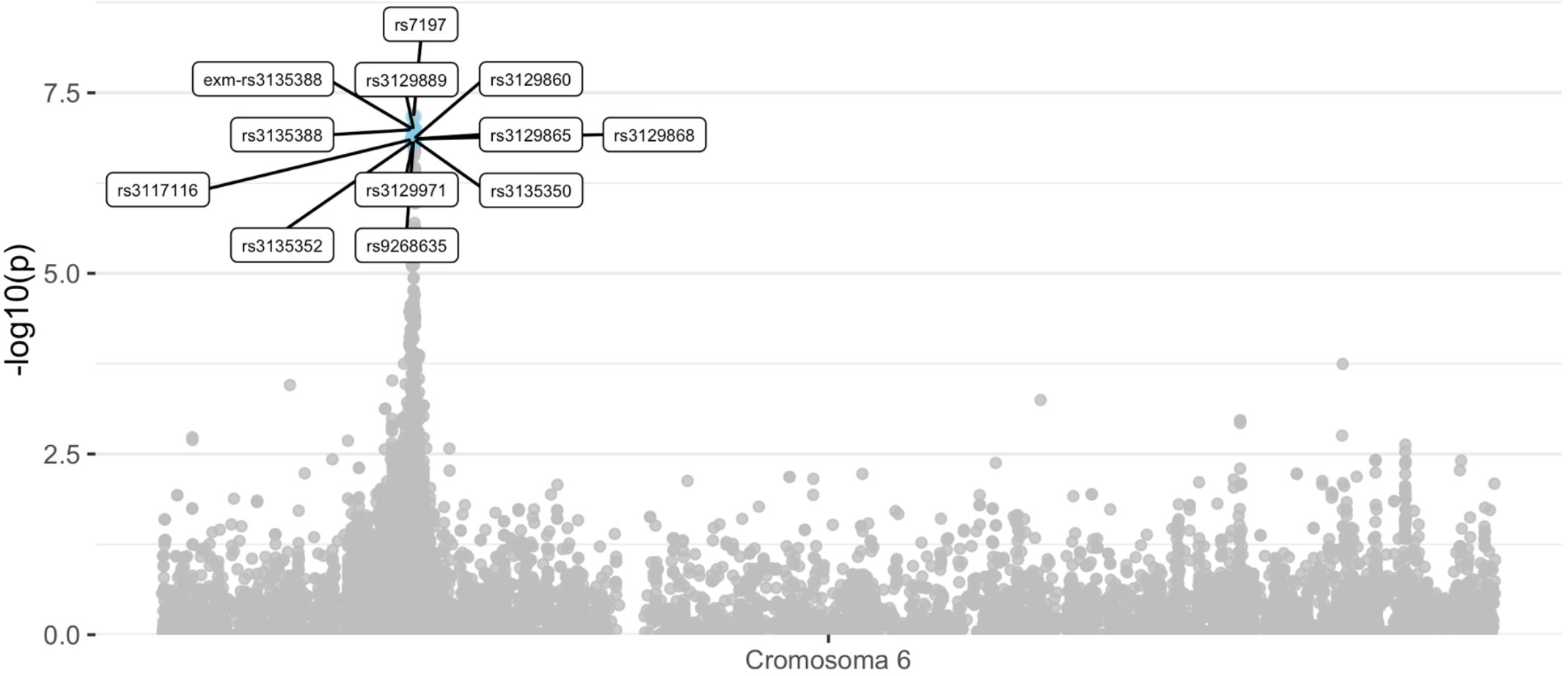


Figura 6: Manhattan plot del cromosoma 6. Se aprecia como casi todos los loci significativos se encuentran localizados en unas pocas pares de bases. Nota: el polimorfismo exm-rs3135338 es equivalente al rs3135338.

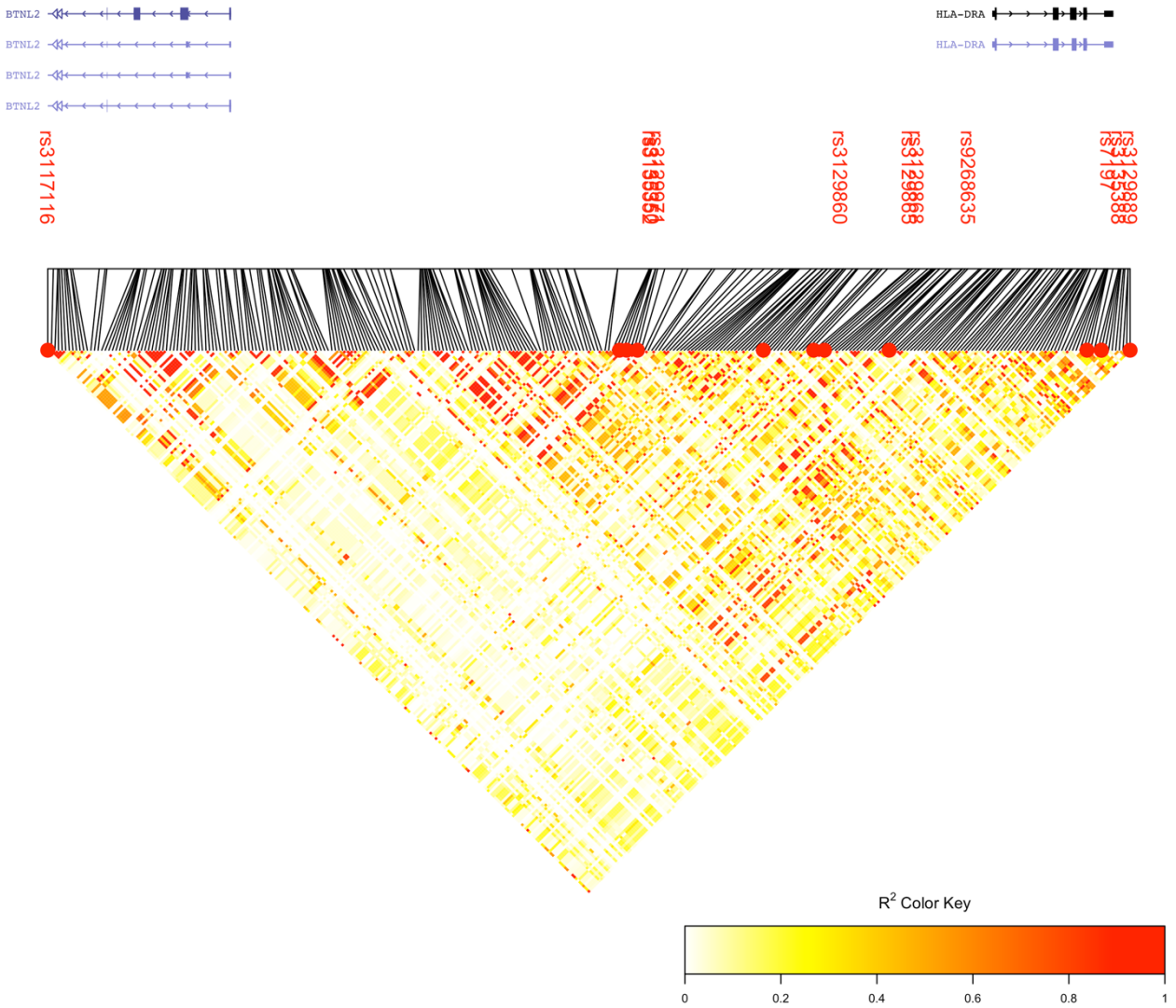


Figura 7: Desequilibrio de ligamiento de la región Chr6:32.399.240-32.445.768 junto con la representación de los genes en dichas localizaciones. Se aprecia como existe un desequilibrio de ligamiento, especialmente en la región situada cerca del HLA-DRA.

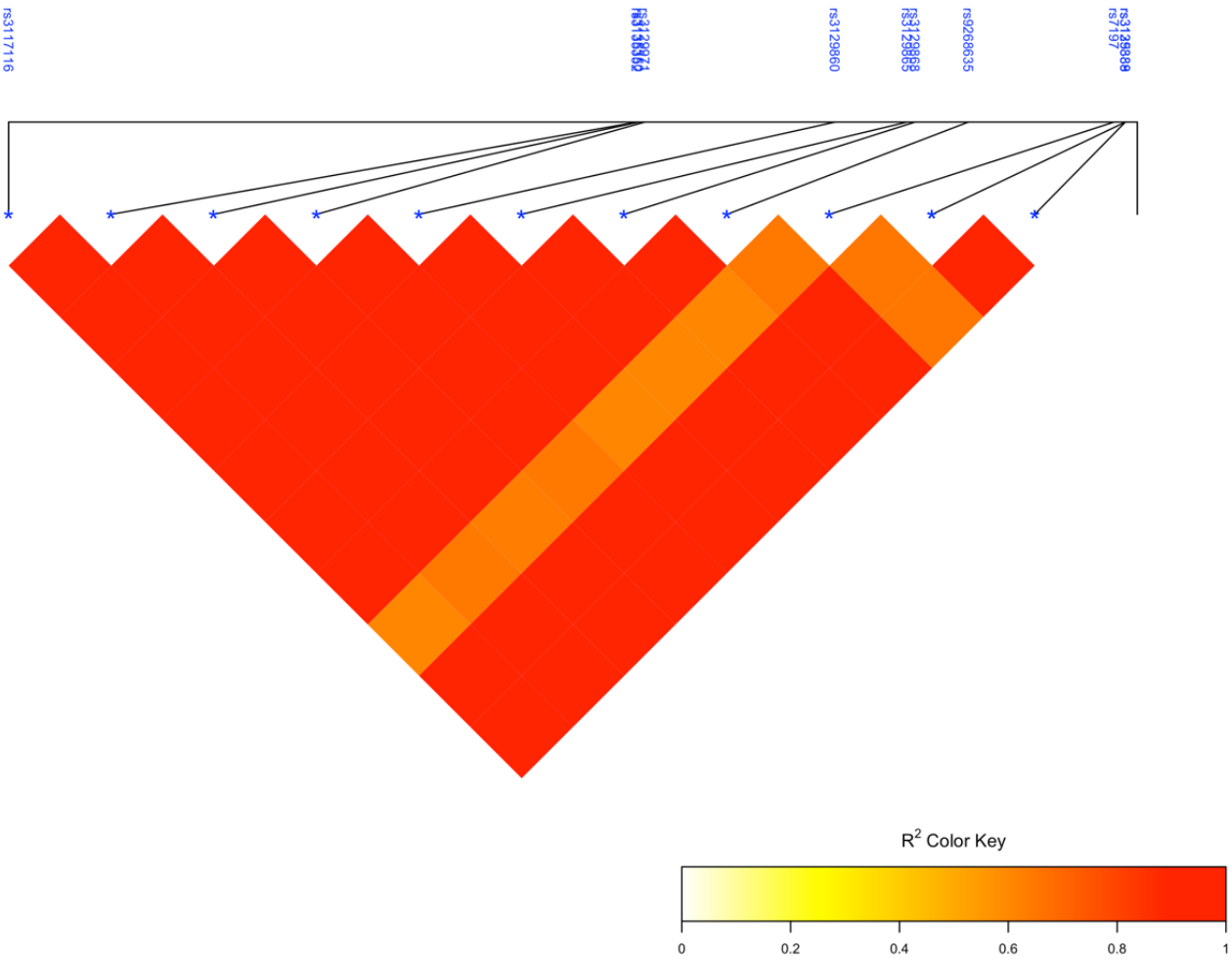


Figura 8: Desequilibrio de ligamiento de los polimorfismos seleccionados. Se aprecia como los loci en cuestión tienen una segregación común.

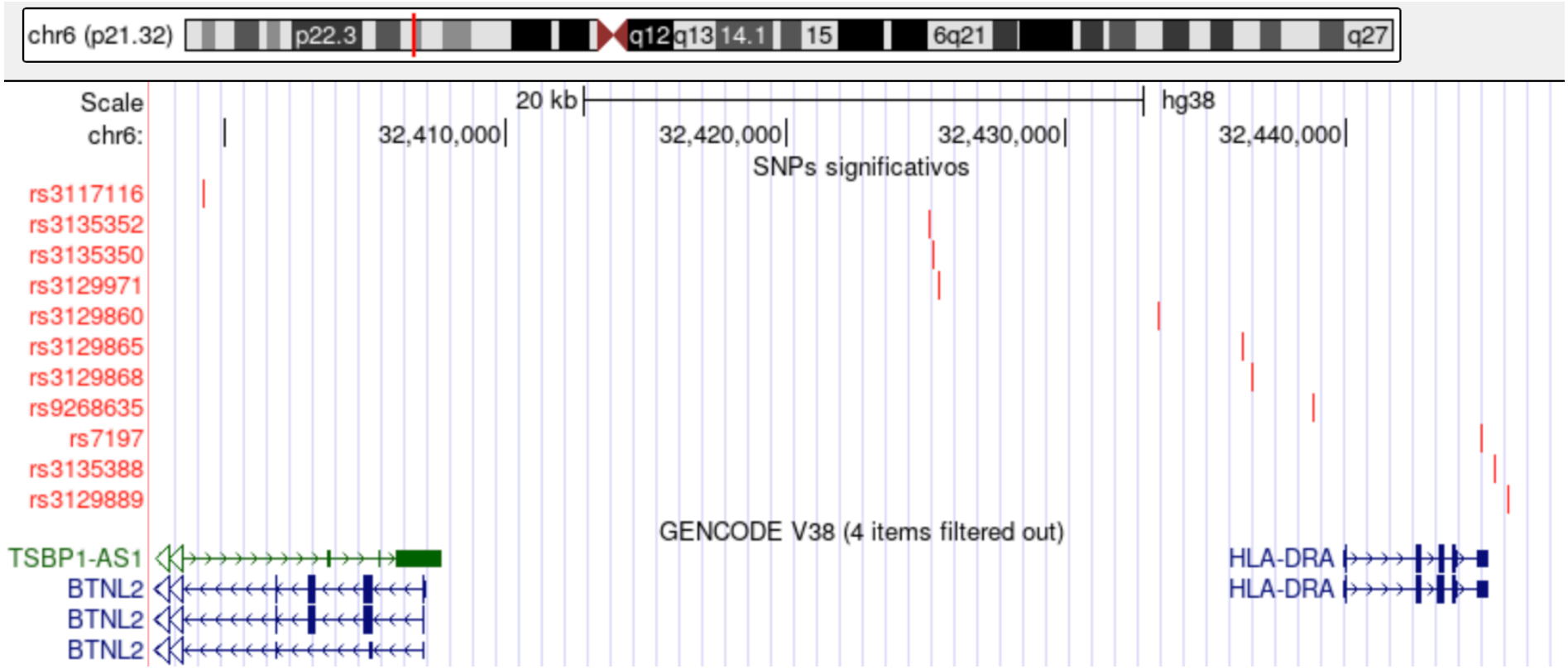


Figura 9: Distancias y localización de los genes BTNL2, HLA-DRA y HLA-DRB1 y de los SNPs significativos. Se muestra la región chr6:32.397.451-32.448.051 de la versión hg38 del genoma humano. Imagen obtenida a partir del UCSC Genome Browser[6].

[6] http://genome.ucsc.edu

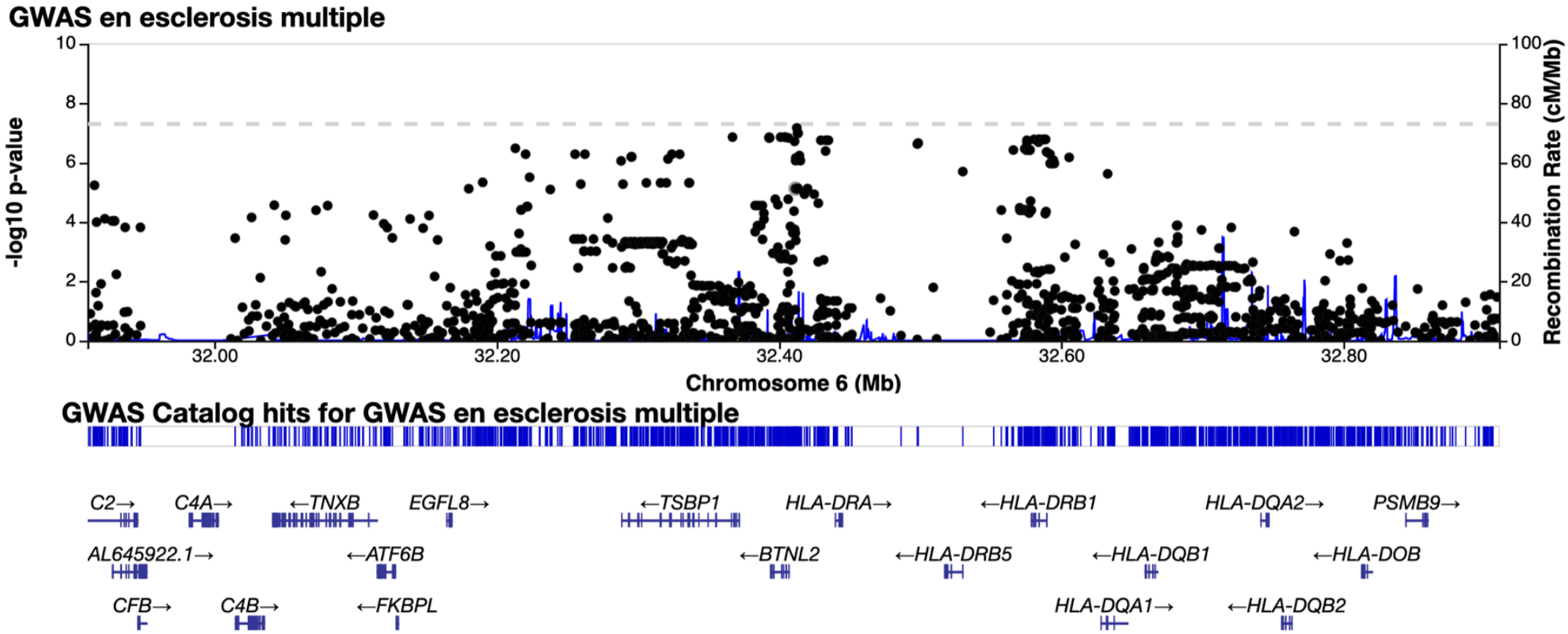


Figura 10: Región del complejo mayor de histocompatibilidad con los con los valores de p obtenidos en el estudio para cada uno de los SNPs de la región. Se aprecia como se acumulan los valores con un menor valor de p en torno a esta región. Imagen obtenida con el software web LocusZoom[52].

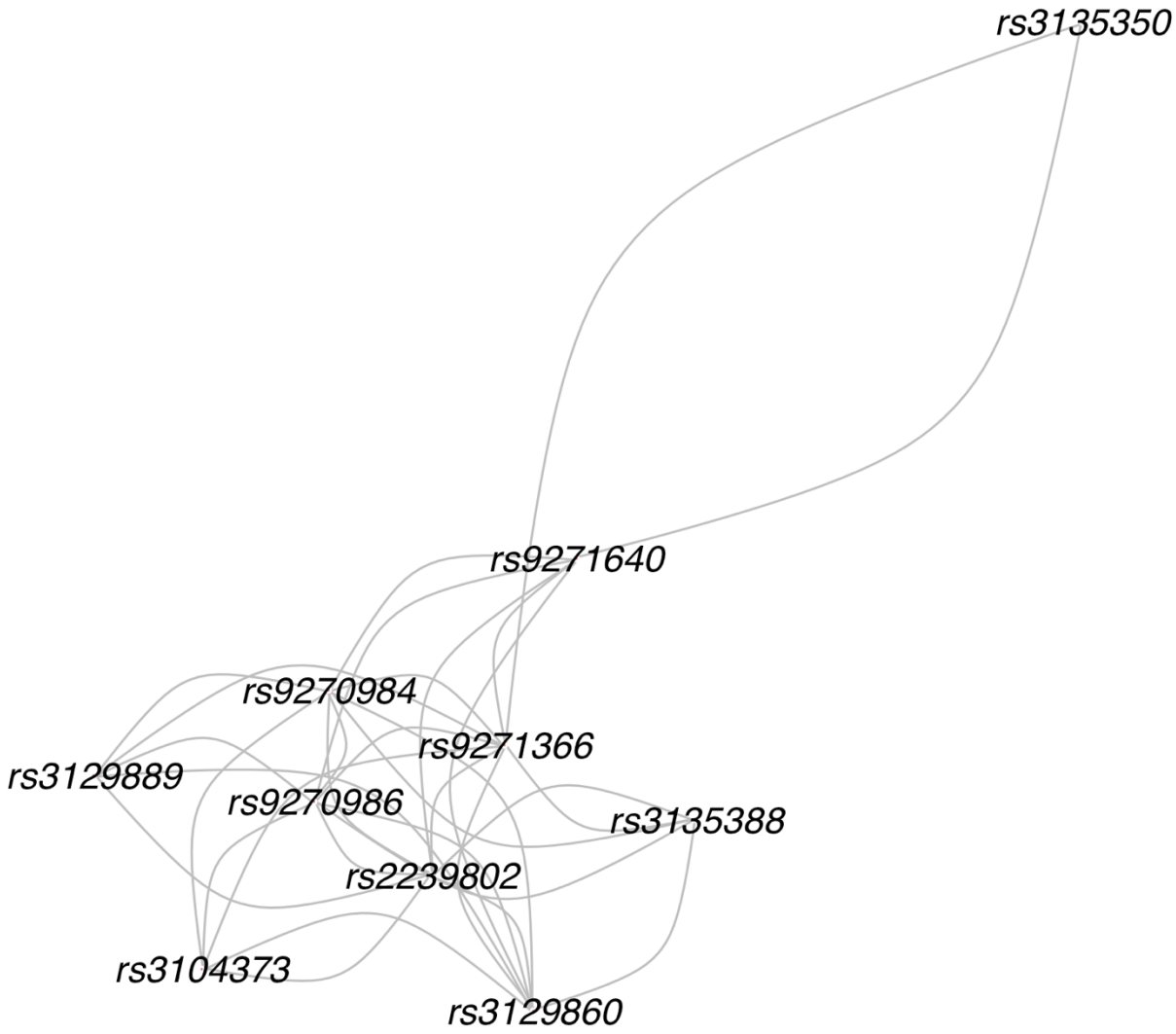


Figura 11: Gráfico de distancias entre los diferentes polimorfismos significativos de acuerdo con los términos EFO. Se aprecia como el SNP rs3135350 se encuentra más alejado del resto.

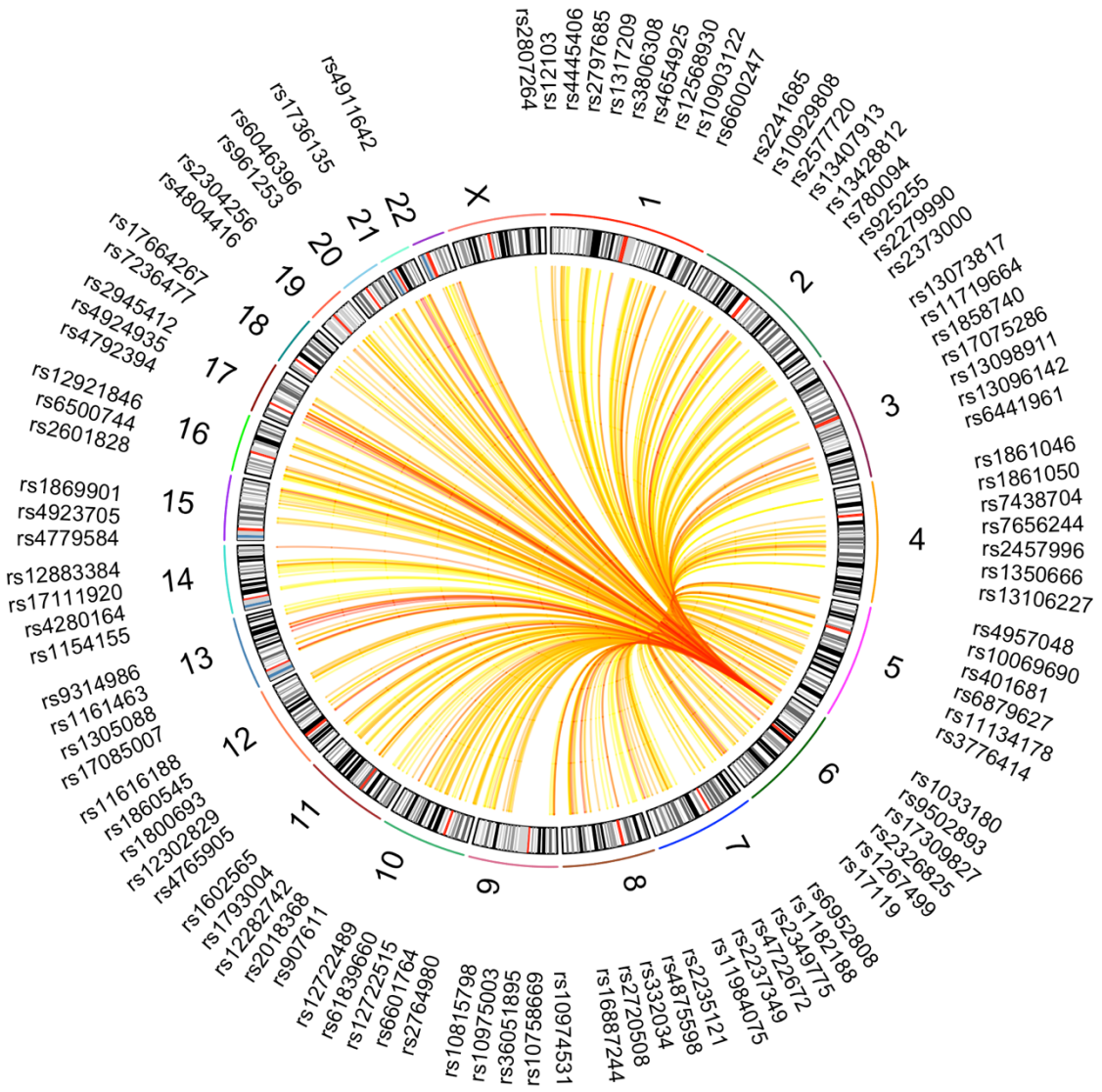


Figura 12: Gráfico de las distancias de los polimorfismos significativos respecto al resto de polimorfismos tipados en el estudio. Se aprecia la conexión con el cromosoma 17 y el 13.

7

Glosario

AR: Artritis reumatoide

BTNL2: Butyrophilin-Like Molecule 2

CEMCAT: Centre d'Esclerosi Múltiple de Catalunya

Chr: Cromosoma

CU: Colitis ulcerosa

DM1: Diabetes mellitus tipo 1

EBNA-1: Epstein Barr Nuclear Antigen 1

EFO: Experimental factor ontology

EM: Esclerosis múltiple

GWAS: Genome Wide Association Study

HLA: Human Leukocyte Antigen

IgA: Inmunoglobulina A

IgG: Inmunoglobulina G

IMSGC International Multiple Sclerosis Genetics Consortium

LES: Lupus eritematoso sistémico

MHC: Major histocompatibility complex

NCBI: National Center for Biotechnological Information

OR: Odds ratio

pb: pares de bases

SAF: Síndrome antifosfolípido

SCA: Síndrome coronario agudo

SNC: Sistema nervioso central

SNP: Single nucleotid polimorphism

TSBP-1: Testis expressed basic protein

UCSC: University of California, Santa Cruz

9

Bibliografía

1. Iwanowski P, Losy J. Immunological differences between classical phenothypes of multiple sclerosis. *J Neurol Sci*. 2015;349(1-2):10-14. doi:10.1016/j.jns.2014.12.035
2. Walton C, King R, Rechtman L, et al. Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition. *Mult Scler*. 2020;26(14):1816-1821. doi:10.1177/1352458520970841
3. Perez-Carmona N, Fernandez-Jover E, Sempere AP. [Epidemiology of multiple sclerosis in Spain]. *Rev Neurol*. 2019;69(1):32-38. doi:10.33588/rn.6901.2018477
4. Fagnani C, Neale MC, Nisticò L, et al. Twin studies in multiple sclerosis: A meta-estimation of heritability and environmentality. *Mult Scler*. 2015;21(11):1404-1413. doi:10.1177/1352458514564492
5. Kuusisto H, Kaprio J, Kinnunen E, Luukkaala T, Koskenvuo M, Elovaara I. Concordance and heritability of multiple sclerosis in Finland: study on a nationwide series of twins. *Eur J Neurol*. 2008;15(10):1106-1110. doi:10.1111/j.1468-1331.2008.02262.x
6. Alonso A, Hernán MA, Ascherio A. Allergy, family history of autoimmune diseases, and the risk of multiple sclerosis. *Acta Neurol Scand*. 2008;117(1):15-20. doi:10.1111/j.1600-0404.2007.00898.x
7. Esposito F, Guaschino C, Sorosina M, et al. Impact of MS genetic loci on familial aggregation, clinical phenotype, and disease prediction. *Neurol Neuroimmunol neuroinflammation*. 2015;2(4):e129. doi:10.1212/NXI.0000000000000129
8. Gusev A, Bhatia G, Zaitlen N, et al. Quantifying missing heritability at known GWAS loci. *PLoS Genet*. 2013;9(12):e1003993. doi:10.1371/journal.pgen.1003993
9. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science*. 2019;365(6460). doi:10.1126/science.aav7188

10. Sawcer S, Hellenthal G, Pirinen M, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*. 2011;476(7359):214-219. doi:10.1038/nature10251
11. Patsopoulos NA, Esposito F, Reischl J, et al. Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Ann Neurol*. 2011;70(6):897-912. doi:10.1002/ana.22609
12. Beecham AH, Patsopoulos NA, Xifara DK, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet*. 2013;45(11):1353-1360. doi:10.1038/ng.2770
13. Nelis M, Esko T, Mägi R, et al. Genetic structure of Europeans: a view from the North-East. *PLoS One*. 2009;4(5):e5472. doi:10.1371/journal.pone.0005472
14. Barcellos LF, Thomson G. Genetic analysis of multiple sclerosis in Europeans. *J Neuroimmunol*. 2003;143(1-2):1-6. doi:10.1016/j.jneuroim.2003.08.004
15. Beecham AH, Amezcua L, Chinea A, et al. The genetic diversity of multiple sclerosis risk among Hispanic and African American populations living in the United States. *Mult Scler*. 2020;26(11):1329-1339. doi:10.1177/1352458519863764
16. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW DM& SP. PLINK: a toolset for whole-genome association and population-based linkage analysis. Published online 2007.
17. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2020. doi:https://www.R-project.org/
18. Team Rs. RStudio: Integrated Development Environment for R. RStudio,. Published online 2020. http://www.rstudio.com/
19. Fang H, Knezevic B, Burnham KL, Knight JC. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. *Genome Med*. 2016;8(1):129. doi:10.1186/s13073-016-0384-y
20. Malone J, Holloway E, Adamusiak T, et al. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*. 2010;26(8):1112-1118. doi:10.1093/bioinformatics/btq099
21. Gough SCL, Simmonds MJ. The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. *Curr Genomics*. 2007;8(7):453-465. doi:10.2174/138920207783591690
22. O'Gorman C, Lin R, Stankovich J, Broadley SA. Modelling genetic susceptibility to multiple sclerosis with family data. *Neuroepidemiology*. 2013;40(1):1-12. doi:10.1159/000341902

23. De Jager PL, Jia X, Wang J, et al. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat Genet*. 2009;41(7):776-782. doi:10.1038/ng.401
24. Risk Alleles for Multiple Sclerosis Identified by a Genomewide Study. *N Engl J Med*. 2007;357(9):851-862. doi:10.1056/NEJMoa073493
25. Goris A, Pauwels I, Gustavsen MW, et al. Genetic variants are major determinants of CSF antibody levels in multiple sclerosis. *Brain*. 2015;138(Pt 3):632-643. doi:10.1093/brain/awu405
26. de Bakker PIW, McVean G, Sabeti PC, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet*. 2006;38(10):1166-1172. doi:10.1038/ng1885
27. Alcina A, Abad-Grau MDM, Fedetz M, et al. Multiple sclerosis risk variant HLA-DRB1*1501 associates with high expression of DRB1 gene in different human populations. *PLoS One*. 2012;7(1):e29819. doi:10.1371/journal.pone.0029819
28. Restrepo NA, Butkiewicz M, McGrath JA, Crawford DC. Shared Genetic Etiology of Autoimmune Diseases in Patients from a Biorepository Linked to De-identified Electronic Health Records. *Front Genet*. 2016;7:185. doi:10.3389/fgene.2016.00185
29. Zhang L, Pan Q, Wang Y, Wu X, Shi X. Bayesian Network Construction and Genotype-Phenotype Inference Using GWAS Statistics. *IEEE/ACM Trans Comput Biol Bioinforma*. 2019;16(2):475-489. doi:10.1109/TCBB.2017.2779498
30. Hadjigeorgiou GM, Kountra P-M, Koutsis G, et al. Replication study of GWAS risk loci in Greek multiple sclerosis patients. *Neurol Sci Off J Ital Neurol Soc Ital Soc Clin Neurophysiol*. 2019;40(2):253-260. doi:10.1007/s10072-018-3617-6
31. Márquez A, Cordero-Coma M, Martín-Villa JM, et al. New insights into the genetic component of non-infectious uveitis through an Immunochip strategy. *J Med Genet*. 2017;54(1):38-46. doi:10.1136/jmedgenet-2016-104144
32. Nguyen T, Liu XK, Zhang Y, Dong C. BTNL2, a butyrophilin-like molecule that functions to inhibit T cell activation. *J Immunol*. 2006;176(12):7354-7360. doi:10.4049/jimmunol.176.12.7354
33. Panea C, Zhang R, VanValkenburgh J, et al. Butyrophilin-like 2 regulates site-specific adaptations of intestinal γδ intraepithelial lymphocytes. *Commun Biol*. 2021;4(1):913. doi:10.1038/s42003-021-02438-x

34. Lin Y, Wei J, Fan L, Cheng D. BTNL2 gene polymorphism and sarcoidosis susceptibility: a meta-analysis. *PLoS One*. 2015;10(4):e0122639. doi:10.1371/journal.pone.0122639
35. Orozco G, Eerligh P, Sánchez E, et al. Analysis of a functional BTNL2 polymorphism in type 1 diabetes, rheumatoid arthritis, and systemic lupus erythematosus. *Hum Immunol*. 2005;66(12):1235-1241. doi:10.1016/j.humimm.2006.02.003
36. Mitsunaga S, Hosomichi K, Okudaira Y, et al. Exome sequencing identifies novel rheumatoid arthritis-susceptible variants in the BTNL2. *J Hum Genet*. 2013;58(4):210-215. doi:10.1038/jhg.2013.2
37. Pathan S, Gowdy RE, Cooney R, et al. Confirmation of the novel association at the BTNL2 locus with ulcerative colitis. *Tissue Antigens*. 2009;74(4):322-329. doi:10.1111/j.1399-0039.2009.01314.x
38. Mochida A, Kinouchi Y, Negoro K, et al. Butyrophilin-like 2 gene is associated with ulcerative colitis in the Japanese under strong linkage disequilibrium with HLA-DRB1*1502. *Tissue Antigens*. 2007;70(2):128-135. doi:10.1111/j.1399-0039.2007.00866.x
39. Price P, Santoso L, Mastaglia F, et al. Two major histocompatibility complex haplotypes influence susceptibility to sporadic inclusion body myositis: critical evaluation of an association with HLA-DR3. *Tissue Antigens*. 2004;64(5):575-580. doi:10.1111/j.1399-0039.2004.00310.x
40. Sinisalo J, Vlachopoulou E, Marchesani M, et al. Novel 6p21.3 Risk Haplotype Predisposes to Acute Coronary Syndrome. *Circ Cardiovasc Genet*. 2016;9(1):55-63. doi:10.1161/CIRCGENETICS.115.001226
41. Fitzgerald LM, Kumar A, Boyle EA, et al. Germline missense variants in the BTNL2 gene are associated with prostate cancer susceptibility. *Cancer Epidemiol biomarkers Prev a Publ Am Assoc Cancer Res cosponsored by Am Soc Prev Oncol*. 2013;22(9):1520-1528. doi:10.1158/1055-9965.EPI-13-0345
42. Traherne JA, Barcellos LF, Sawcer SJ, et al. Association of the truncating splice site mutation in BTNL2 with multiple sclerosis is secondary to HLA-DRB1*15. *Hum Mol Genet*. 2006;15(1):155-161. doi:10.1093/hmg/ddi436
43. Rioux JD, Goyette P, Vyse TJ, et al. Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proc Natl Acad Sci U S A*. 2009;106(44):18680-18685. doi:10.1073/pnas.0909307106
44. Rubicz R, Yolken R, Alaedini A, et al. Genome-wide genetic and transcriptomic investigation of variation in antibody response to

dietary antigens. *Genet Epidemiol*. 2014;38(5):439-446. doi:10.1002/gepi.21817
45. Ross CJ, Towfic F, Shankar J, et al. A pharmacogenetic signature of high response to Copaxone in late-phase clinical-trial cohorts of multiple sclerosis. *Genome Med*. 2017;9(1):50. doi:10.1186/s13073-017-0436-y
46. Ferreiro-Iglesias A, Lesseur C, McKay J, et al. Fine mapping of MHC region in lung cancer highlights independent susceptibility loci by ethnicity. *Nat Commun*. 2018;9(1):3927. doi:10.1038/s41467-018-05890-2
47. Rubicz R, Yolken R, Drigalenko E, et al. A genome-wide integrative genomic study localizes genetic factors influencing antibodies against Epstein-Barr virus nuclear antigen 1 (EBNA-1). *PLoS Genet*. 2013;9(1):e1003147. doi:10.1371/journal.pgen.1003147
48. Kamboh MI, Wang X, Kao AH, et al. Genome-wide association study of antiphospholipid antibodies. *Autoimmune Dis*. 2013;2013:761046. doi:10.1155/2013/761046
49. Fernando MMA, Freudenberg J, Lee A, et al. Transancestral mapping of the MHC region in systemic lupus erythematosus identifies new independent and interacting loci at MSH5, HLA-DPB1 and HLA-G. *Ann Rheum Dis*. 2012;71(5):777-784. doi:10.1136/annrheumdis-2011-200808
50. Shi S, Yuan N, Yang M, et al. Comprehensive Assessment of Genotype Imputation Performance. *Hum Hered*. 2018;83(3):107-116. doi:10.1159/000489758
51. Siva N. 1000 Genomes project. *Nat Biotechnol*. 2008;26(3):256. doi:10.1038/nbt0308-256b
52. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*. 2010;26(18):2336-2337. doi:10.1093/bioinformatics/btq419