Universidad Europea

Máster en Bioinformática

CARACTERIZACIÓN INMUNOGENÉTICA DE LA COVID-19

Autor: David García Valentín-Fernández

Tutor: Roberto Díaz Peña

Curso 2023-24

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ANEXOS

AGRADECIMIENTOS

No quisiera terminar este trabajo sin manifestar mi más profundo agradecimiento a todas aquellas personas que han colaborado ayudado en la realización de este TFM.

En primer lugar, quiero agradecer a mi tutor el Dr. Roberto Díaz Peña, del Grupo de Medicina Xenómica del Instituto de Investigación Sanitaria de Santiago (IDIS) en Santiago de Compostela. Siempre manifestó un gran interés en que me involucrara en su proyecto de investigación y me ha puesto todas las facilidades para que pudiera aprender y entender al máximo en lo que estoy trabajando, pues esta es la mejor manera de poder hacer de tu trabajo tu pasión.

También agradecer a todos los profesores que me han enseñado tanto durante este úlimo año académico, especialmente a María del Rocío González Soltero por su Constancia y predisposición para que el paso de todos los alumnos por el máster sea lo mejor posible, con el objetivo de formar a los que seremos, espero, unos grandes profesionales en nuestros campos de estudio.

Agradezco a todos los pacientes y familiares que han colaborado de forma indirecta en el desarrollo de este trabajo permitiendo que sus datos sean utilizados para potenciar la investigación científica y clínica para mejorar la salud de las personas. Por supuesto, todo ello desde un marco bioético que asegure todos los derechos de los pacientes.

Por último, aunque no menos importante, agradecer el apoyo incondicional de mi familia y amigos, que siempre están presentes en mi desarrollo personal y profesional. Espero poder seguir creciendo en ambos sentidos y poder continuar mi formación durante el programa de doctorado.

ABSTRACT

Background: Clinical manifestations of coronavirus disease 19 (COVID-19) characteristically show great variability with marked differences in severity among infected individuals. Previous studies have associated this variability with genetic factors, including genes encoding human leukocyte antigens (HLA). This study investigates the relationship between HLA allelic variability and COVID-19 severity in Latin American populations.

Methods: We analyzed genetic and clinical data from 4,354 COVID-19 patients from Latin American countries. HLA alleles were imputed using HIBAG software with a Hispanic-specific pre-trained model. Association analyses between HLA alleles and disease severity were performed using logistic regression models adjusted for demographic variables and principal components.

Results: The study identified significant associations between HLA alleles and COVID-19 severity. Four risk alleles were identified: B*50:01 (OR=4.419), DQB1*03:19 (OR=2.945), B*15:03 (OR=2.200), and DRB1*15:01 (OR=2.137); and two protective alleles: DRB1*04:02 (OR=0.419) and C*05:01 (OR=0.599). Demographic analysis revealed significant gender-based differences in disease severity, with males showing higher risk ($p<2.2e-16$), and age-related increased risk ($\beta=0.073554$, $p<2e-16$). The analysis also suggested potential protective effects of African and European ancestry components against severe COVID-19.

Conclusions: This study provides evidence for HLA genetic associations with COVID-19 severity in Latin American populations, identifying both risk and protective alleles characteristic of these populations due to their unique genetic structure. These findings contribute to understanding population-specific genetic factors in COVID-19 outcomes and may help in developing personalized risk assessment strategies.

KEYWORDS: HLA, COVID-19, genetic imputation, HIBAG, severity, Latin American population, risk alleles, protective alleles, genetic ancestry, demographic factors.

RESUMEN

Antecedentes: Característicamente, las manifestaciones clínicas de la enfermedad por coronavirus 19 (COVID-19) muestran una gran variabilidad con marcadas diferencias en su gravedad entre individuos infectados. Esta variabilidad se ha asociado en estudios previos con factores genéticos, entre ellos los genes que codifican los antígenos leucocitarios humanos (HLA). Este estudio investiga la relación entre la variabilidad alélica HLA y la severidad de COVID-19 en poblaciones latinoamericanas.

Métodos: Analizamos datos genéticos y clínicos de 4.354 pacientes con COVID-19 de países latinoamericanos. Los alelos HLA se imputaron utilizando el software HIBAG con un modelo pre-entrenado específico para población hispana. Los análisis de asociación entre alelos HLA y severidad de la enfermedad se realizaron mediante modelos de regresión logística ajustados por variables demográficas y componentes principales.

Resultados: El estudio identificó asociaciones significativas entre alelos HLA y severidad de COVID-19. Se identificaron cuatro alelos de riesgo: B*50:01 (OR=4,419), DQB1*03:19 (OR=2,945), B*15:03 (OR=2,200) y DRB1*15:01 (OR=2,137); y dos alelos protectores: DRB1*04:02 (OR=0,419) y C*05:01 (OR=0,599). El análisis demográfico reveló diferencias significativas en la severidad de la enfermedad según el género, mostrando los hombres mayor riesgo ($p<2,2e-16$), y un riesgo incrementado con la edad ($\beta=0,073554$, $p<2e-16$). El análisis también sugirió potenciales efectos protectores de los componentes de ancestría africana y europea contra COVID-19 severo.

Conclusiones: Este estudio proporciona evidencia de asociaciones genéticas HLA con la severidad de COVID-19 en poblaciones latinoamericanas, identificando alelos tanto de riesgo como protectores característicos de estas poblaciones debido a su estructura genética propia. Estos hallazgos contribuyen a la comprensión de factores genéticos específicos de población en los resultados de COVID-19 y podrían ayudar en el desarrollo de estrategias de evaluación de riesgo personalizadas.

PALABRAS CLAVE: HLA, COVID-19, imputación genética, HIBAG, severidad, población latinoamericana, alelos de riesgo, alelos protectores, ancestría genética, factores demográficos.

1. Introducción

La infección por SARS-CoV-2, causante de la pandemia de enfermedad por coronavirus 19 (COVID-19), ha evidenciado una amplia gama de manifestaciones clínicas, desde infecciones asintomáticas hasta neumonías graves, causando en muchos casos la muerte. La variabilidad en la respuesta clínica a la infección ha sido objeto de numerosos estudios, algunos de los cuales han señalado la influencia de factores genéticos en la severidad de la enfermedad. Entre estos factores, el complejo mayor de histocompatibilidad (MHC) humano, o región HLA (antígeno leucocitario humano) juega un papel crucial debido a su rol en la respuesta inmunitaria mediada por células T (Castro-Santos et al., 2023; Marchal et al., 2024).

El sistema HLA es fundamental para la presentación de antígenos a las células T, una etapa crítica en la respuesta inmunitaria adaptativa. Diversas investigaciones han explorado la asociación entre alelos específicos de HLA y la gravedad de la COVID-19. Sin embargo, los resultados han sido en ocasiones contradictorios. Por ejemplo, mientras algunos estudios han identificado alelos HLA que confieren susceptibilidad o protección contra COVID-19 severa, otros no han encontrado asociaciones significativas (Marchal et al., 2024).

En un estudio reciente, se identificaron los alelos HLA-A*11:01 y HLA-C*04:01 como asociados con formas graves de COVID-19, sugiriendo una posible predisposición genética a desarrollar síntomas más severos en individuos portadores de estos alelos (Castro-Santos et al., 2023). Contrariamente, otro estudio no encontró una asociación significativa entre alelos clásicos de HLA y la infección asintomática por SARS-CoV-2, lo que subraya la complejidad de la relación entre HLA y la respuesta clínica al virus (Marchal et al., 2024).

La presente investigación se centra en imputar los diferentes alelos HLA a partir de datos de genotipado en individuos COVID-19 positivos utilizando el paquete HIBAG de *Bioconductor* en R. El objetivo es investigar si las variaciones en la distribución de los alelos HLA están asociadas con una predisposición a desarrollar síntomas más graves de COVID-19. Este análisis permitirá una comprensión más profunda de los factores genéticos que influyen en la severidad de la COVID-19.

1.1. Antígeno leucocitario humano

El sistema del antígeno leucocitario humano (HLA) es una familia de genes altamente polimórficos que participan en la regulación de la función inmunitaria para distinguir entre lo propio y lo ajeno. Las moléculas de HLA interactúan con los receptores de células T en el timo para modular la respuesta inmunitaria y determinar qué células se reconocen como propias y cuáles no (Madden et al., 2019).

Los genes HLA se encuentran en la región más polimórfica del genoma humano, el complejo mayor de histocompatibilidad, que se encuentra en el cromosoma 6p21.3. Existen casi 40.000 alelos distintos para el MHC que se encuentran estrechamente relacionados entre sí y que codifican para los *loci*, denominados “clásicos”, de clase I (HLA-A, HLA-B y HLA-C) y de clase II (HLA-DR, HLA-DQ y HLA-DP) (Blackwell et al., 2009; Madden et al., 2019).

1.1.1. Clases del HLA. Estructura y función

La región HLA contiene múltiples loci genéticos estructurados en diferentes subregiones, cada una con similitudes estructurales, los cuales están representados en la Figura 1. Estos loci codifican para un heterodímero de la superficie celular en todas las células nucleadas (HLA-I) y

en las células presentadoras de antígenos (Hickey et al., 2016). Su función es unir y presentar péptidos al sistema inmunitario, permitiendo diferrenciar entre los propio y lo ajeno. Hasta la fecha, en todo el mundo se han identificado alrededor de 39.627 alelos en la región de las clases I y II del HLA (HLA Nomenclature Committee, 2024).

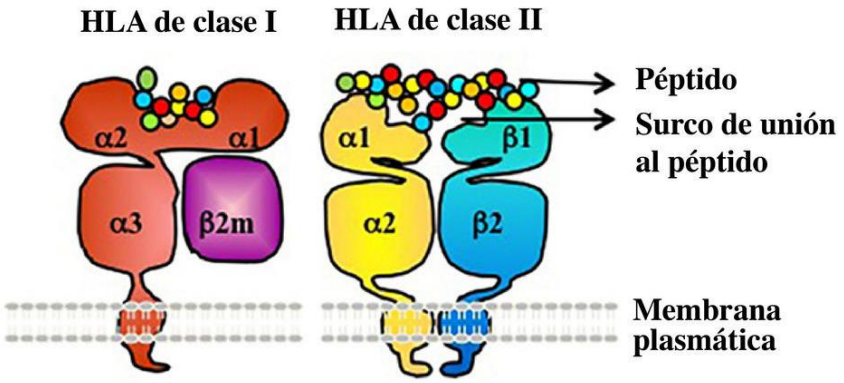


Figura 1: Estructura de las proteínas transmembrana HLA clase I y II (Hickey et al., 2016).

La región de clase I es la parte más telomérica del complejo HLA y contiene tres *loci* "clásicos": HLA-A, B y C; así como cuatro "no clásicos": HLA-E, F, G y H (Yan et al., 2003). Mientras que los genes HLA de clase I "clásicos" son altamente polimórficos y tienen una capacidad de presentación de antígenos distintiva, los "no clásicos" son menos polimórficos y desempeñan diversas funciones (Naito et al., 2022). La clase II incluye los *loci* HLA-DP, HLA-DQ y HLA-DR. La expresión de estas proteínas se limita a células inmunitarias específicas, incluidas las células B, los macrófagos, las células dendríticas y el epitelio tímico (Madden et al., 2019).

Las proteínas de clase I se expresan en la superficie de todas las células nucleadas y están compuestas por una cadena pesada transmembrana con tres dominios extracelulares, denominados α1, α2 y α3; y una cadena ligera de β2-microglobulina que ancla la cadena pesada a la membrana citoplasmática. Estas proteínas presentan los antígenos endógenos a los linfocitos T CD8+ (Madden et al., 2019). Las proteínas de clase II están compuestas por una cadena alfa con dos dominios y una cadena beta con otros dos dominios, ambas ancladas a la membrana plasmática. Estas proteínas se encargan de presentar los antígenos exógenos a los linfocitos T CD4+ (Naito et al., 2022; Madden et al., 2019).

1.2. Proceso de imputación genética

Desde hace tiempo se sabe que es posible inferir o predecir información genética de zonas no analizadas directamente utilizando únicamente un número relativamente modesto de marcadores genéticos por individuo (Li et al., 2009). Para ello se lleva a cabo el proceso de imputación de genotipos, donde se utiliza un panel de referencia de polimorfismos de nucleótido único (SNPs). Los SNPs son variaciones en una sola base del ADN que ocurren en posiciones específicas del genoma; son comunes en la población y pueden influir en características individuales, como la susceptibilidad a enfermedades o la respuesta a ciertos fármacos. A través de la imputación, se busca aumentar la cantidad de SNPs en la muestra de estudio, permitiendo así un análisis más amplio y preciso. La imputación puede generar un aumento de potencia de hasta un 10 % con respecto a la prueba de solo SNP genotipados en un estudio de asociación de todo el genoma (GWA) (Marchini et al., 2010).

El proceso de imputación de regiones genómicas utiliza el denominado "mapeo fino" para aumentar la posibilidad de localizar SNPs causales en relación al análisis que se esté realizando (Marchini et al., 2010; Naito et al., 2022). Esto es de utilidad cuando se quiere realizar un estudio de asociación en una región tan polimórfica y con tantas variaciones estructurales como la región MHC. Estas características hacen que los métodos tradicionales basados en la reacción en cadena de la polimerasa (PCR) y la secuenciación de próxima generación (NGS) sean tan laboriosos, lentos y costosos que no se podrían aplicar al mapeo fino para grandes cohortes de GWAS. De modo que se opta por imputar los genotipos de los alelos HLA utilizando un panel o modelo de referencia de HLA pre-entrenado. La imputación de HLA ha contribuido con éxito al mapeo fino de variantes causales de HLA para delinear la inmunopatología de diversas enfermedades (Naito et al., 2022).

1.2.1. Imputación HLA con asignación de atributos: HIBAG

La imputación de alelos HLA nos permite mapear con precisión la región MHC utilizando un modelo de referencia con genotipos de HLA y SNPs para predecir alelos HLA en base a la información de SNPs proporcionada (Figura 2) (Naito et al., 2022).

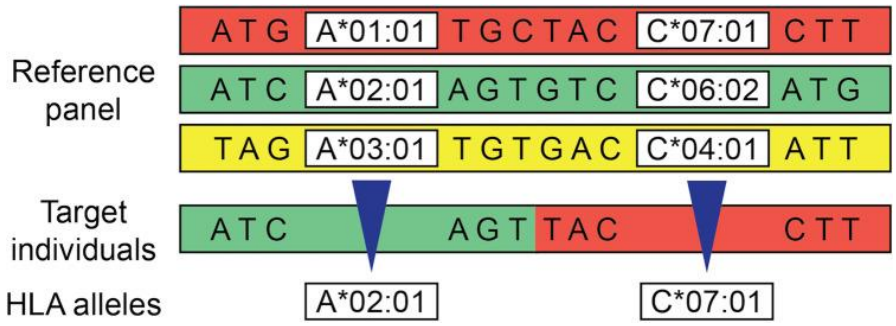


Figura 2: imputación de HLA mediante un panel de referencia que contiene datos individuales de genotipos SNP e información sobre tipificación de HLA. En base a esta información se puede predecir información alélica de individuos sobre los que no se tienen sus genotipos SNP.

Para poder imputar correctamente la complicada estructura molecular de la región MHC existen varios métodos de imputación como HLA*IMP:02 (software independiente), HLA*IMP:03 (aplicación web), SNP2HLA (script de shell), HIBAG (paquete R) o DEEP*HLA (secuencia de comandos Python), entre otros (Naito et al., 2022).

HIBAG es un paquete de software de última generación para imputar tipos de HLA utilizando datos de SNP, y se basa en un conjunto de entrenamiento de genotipos de HLA y SNP. Es un sistema de predicción altamente preciso, computacionalmente manejable y puede utilizarse con estimaciones de parámetros publicadas, eliminando la necesidad de acceder a muestras de entrenamiento grandes. Combina los conceptos de *bagging* de atributos con la deducción de haplotipos a partir de SNPs y tipos de HLA. El *bagging* de atributos es una técnica que mejora la precisión y estabilidad de los modelos, creando varias versiones del mismo al entrenarlos con muestras aleatorias del conjunto de datos original. Luego, los resultados de estos modelos se combinan para obtener una predicción más precisa y confiable (Figura 3) (Zheng et al., 2014).

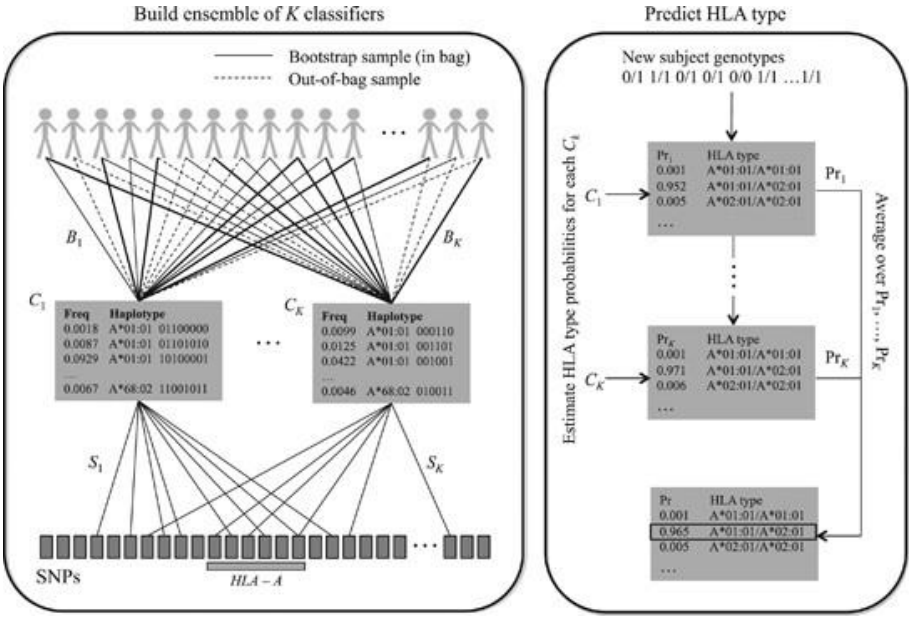


Figura 3: Descripción del funcionamiento del algoritmo HIBAG para la predicción de SNPs.

Para la imputación, se crean varios modelos o clasificadores. Cada modelo trabaja solo con una parte aleatoria de los SNP disponibles para predecir los tipos de HLA y la frecuencia de haplotipos. Al final, se obtienen las predicciones finales al combinar los resultados promedio de todos los modelos (Zheng et al., 2014).

1.3. Variabilidad en la región HLA y su relación con el COVID-19

La pandemia de COVID-19 ha tenido un profundo impacto global, con variaciones en la susceptibilidad, gravedad y tasas de mortalidad en diferentes regiones. Se ha visto que la infección por SARS-CoV-2 es heterogénea en cuanto a su presentación clínica y evolución, pudiendo variar desde síntomas muy leves parecidos a los de la gripe hasta el síndrome de dificultad respiratoria aguda asociado con el ingreso en la unidad de cuidados intensivos y una alta mortalidad (Amoroso et al., 2021; Hoseinnezhad et al., 2024; Marchal et al., 2024).

La variabilidad individual en la construcción de una respuesta inmune al virus es una de las hipótesis que pueden justificar esta heterogeneidad fenotípica. Se sabe que el trasfondo genético que controla las respuestas inmunes puede afectar a la capacidad del virus para infectar y del huésped para defenderse. Basándonos en la experiencia con otros patógenos como el VHC (virus de la hepatitis C) o el VIH (virus de la inmunodeficiencia humana), es posible que los polimorfismos en los genes que codifican proteínas de inmunidad innata y adaptativa puedan ser responsables de una respuesta diferente del huésped a la infección y, por lo tanto, de la gravedad de la enfermedad (Amoroso et al., 2021). Si bien muchos factores genéticos pueden contribuir a estas variabilidades poblacionales, las variantes genéticas del HLA han surgido como posibles contribuyentes a los resultados de COVID-19 (Hoseinnezhad et al., 2024). *In silico*, los péptidos del SARS-CoV-2 tienen diferentes afinidades por los alelos HLA, lo que apoya la teoría de que la presencia de diferentes alelos HLA influyen en la evolución de la infección y en sus manifestaciones clínicas (Amoroso et al., 2021).

2. Hipótesis y Objetivos

La hipótesis principal del presente trabajo es que la variabilidad en los alelos HLA juega un papel relevante en la aparición de sintomatología grave en pacientes infectados con SARS-CoV-2.

El objetivo principal del estudio es investigar la relación entre los alelos HLA y la predisposición a desarrollar síntomas graves de COVID-19 en pacientes infectados, utilizando técnicas de imputación y análisis bioinformático y estadístico.

3. Metodología

3.1. Diseño del estudio

Este estudio tiene un diseño observacional y transversal para imputar los alelos HLA en pacientes de COVID-19 utilizando datos de genotipado en formato PLINK y el paquete HIBAG de *Bioconductor* en 'R'. Posteriormente, se analizará la asociación entre los alelos HLA y la gravedad de los síntomas de COVID-19. Todo el proceso y el código está descrito detalladamente en el repositorio público de GitHub: https://github.com/davidgarvalfer/hla-covid-analisis

3.2. Cohorte de pacientes

La población del estudio incluye pacientes diagnosticados con COVID-19 que se dividen fundamentalmente en dos grupos: aquellos que presentaron síntomas leves y aquellos que desarrollaron síntomas graves. Para determinar la gravedad se utilizarán datos de presencia o no de síntomas, hospitaliazción o no y duración, necesidad de ventilación…

En el trabajo se incluyeron un conjunto de datos de 4354 pacientes de Colombia, Brasil y México fundamentalmente, aunque también hay pacientes de República Dominicana, Colombia, Paraguay, Bolivia, Perú, Ecuador, Honduras, Venezuela, Nicaragua, Argentina, Chile, El Salvador y Guatemala. Parte de esta cohorte de Latinoamérica es la utilizada en el estudio llevado a cabo por Almeida et al. (2024). Estos casos clínicos fueron reclutados por 13 centros participantes desde marzo de 2020 hasta julio de 2021.

3.3. Recogida de datos

Las muestras y los datos se recogieron con consentimiento informado tras la aprobación de los Comités de Ética y Científico de los centros participantes y por el Comité de Ética de Galicia Ref 2020/197. El reclutamiento de pacientes del IMSS (en la Ciudad de México), fue aprobado por el Comité Nacional de Investigación Clínica, del Instituto Mexicano del Seguro Social, México (protocolo R-2020-785-082) (Almeida et al., 2024). Además, la cohorte de pacientes específica de Chile cuenta con una Dispensa de Uso de Consentimiento Informado (N°CECDI-02-2) otorgada por el Comité Ético Científico de la Universidad Autónoma de Chile (CEC-UA), siendo el investigador científico responsable el Dr. Roberto Díaz Peña (Anexo I).

Se utilizó la herramienta electrónica de captura de datos *REDCap*, alojada en el Centro de Investigación Biomédica en Red (CIBER) del Instituto de Salud Carlos III (ISCIII), para recopilar y gestionar variables demográficas, epidemiológicas y clínicas. Los sujetos fueron diagnosticados de COVID-19 en función de pruebas de PCR cuantitativa, o según procedimientos clínicos o de laboratorio (pruebas de anticuerpos u otras pruebas microbiológicas) (Almeida et al., 2024).

3.3.1. Datos genéticos

La información genética recopilada de cada uno de los pacientes se divide en tres archivos tipo PLINK (.bed, .bim y .fam).

El archivo BED se encuentra en un formato binario comprimida que almacena la información genotípica, conteniendo la información relacionada con la ausencia o presencia de los alelos para cada marcador genético. A nivel estructural se compone de una matriz donde las filas representan a los individuos y las columnas son los marcadores genéticos.

El archivo BIM contiene texto con información sobre los marcadores genéticos. Se compone de seis columnas que indican el cromosoma, identificador del SNP, distancia genética, posición física y los dos alelos (alelo 1 y alelo 2). Cabe añadir que en este caso es específico para la región HLA ubicada en el cromosoma 6.

Por último, el archivo FAM contiene la información fenotípica de los pacientes, incluyendo el sexo del individuo y los identificadores familiar, individual, paterno y materno. Este archivo es clave para poder vincular los datos genéticos con sus correspondientes variables clínicas.

3.3.2. Datos clínicos

Dentro del *dataframe* de datos clínicos podemos subdividir las variables en diferentes grupos.

Las variables de resultado principal, que son aquellas que aportan descripciones cuantitativas o categóricas sobre aspectos asociados al resultado de la evolución de una enfermedad o al efecto de un tratamiento. En este caso distinguimos:

- Severidad de COVID-19: es una variable dicotómica que categoriza a los pacientes en casos severos (1) y no severos (0). Para la adjudicación de un valor u otro en esta variable se evalúan criterios clínicos como la necesidad de ventilación artificial, necesidad de cuidados intensivos o complicaciones consideradas graves. Esta variable será clave para clasificar a los pacientes en los análisis de riesgo.
- Hospitalización: otra variable dicotómica que indica si el paciente necesito ingresar en el hospital (1) o no necesito ingreso (0).
- Asintomatología: variable dicotómica que difiere entre pacientes con COVID-19 pero que fueron asintomáticos (1), y aquellos que sí presentaron sintomatología (0).
- Tiempo de hospitalización: es una variable numérica que representa el número de días que un paciente estuvo hospitalizado. Está relacionada con la variable binaria de hospitalización, ya que, si no necesitó hospitalización, no hay datos registrados sobre su tiempo en el hospital.

Por otro lado, se incluyeron variables demográficas como la edad, el sexo y el país de origen.

- La edad es una variable continua medida en años, y cabe destacar que el rango de edades para estos datos es de 18 a 78 años.
- El sexo es una variable binaria codificada como “hombre/mujer” que debe ser transformada a valores numéricos binarios para el análisis.
- El país de origen de cada uno de los pacientes es una variable categórica nominal que será importante de cara ver relaciones geográficas respecto de la severidad de la enfermedad.

Por último, es importante tener en cuenta las variables relacionadas con la ancestría.

- Componentes principales (PC1-PC10): diez variables continuas derivadas del análisis genómico de cada paciente. Representan la variación genética poblacional y son importantes de cara al ajuste en modelos estadísticos.
- Proporciones de ancestría: son variables continuas (0-1) que representan la proporción que tiene cada individuo de ancestría europea, africana y nativo-americana.
- La estimación poblacional es una variable categórica que distingue si el paciente se encuentra en una población mezclada (MIXED) o con una clara predominancia nativo-americana (NAM). Se basa en umbrales de ancestría del 80%.

3.4. Software y bibliotecas

El análisis se ejecuta en R, específicamente en la versión 4.3.2. Se utiliza por su alta capacidad en el manejo de conjuntos genéticos y ejecución de análisis estadísticos, además de por implementar una alta variedad de recursos bioinformáticos.

Para la imputación y posterior análisis de asociación entre la variabilidad de la región HLA y la afección por COVID-19 se van a utilizar una serie de bibliotecas especializadas seleccionadas en base a su compatibilidad con los datos genéticos dimensionalmente grandes, capacidad de imputación de HLA utilizando datos PLINK y la capacidad de realizar y visualizar análisis estadísticos de forma eficiente y reproducible.

3.4.1. Paquetes de imputación y análisis genético

El paquete *HIBAG* implementa algoritmos de imputación HLA basados en aprendizaje automático en base a la información de los SNPs, como se explicó anteriormente. Tras la imputación, genera estimaciones de probabilidad posterior para sus predicciones. Para facilitar el procesamiento y manejo de los datos genéticos se utiliza *data.table*.

3.4.2. Paquetes de análisis estadístico

El análisis estadístico se realizó mediante un conjunto de paquetes especializados en estadística en R:

- *pROC* evalúa y compara el rendimiento de los modelos predictivos mediante curvas ROC y el cálculo de áreas bajo la curva (AUC). Es un sistema de puntuación que nos ayuda a saber qué tan bueno es nuestro modelo para hacer predicciones.
- *qvalue* se encarga del control de falsos descubrimientos (FDR) y el ajuste de valores *p* para múltiples comparaciones. Este paquete se suele utilizar cuando se hacen muchas pruebas estadísticas a la vez (en este caso, las pruebas múltiples de alelos HLA), ya que en estos casos aumenta la probabilidad de encontrar resultados significativos por casualidad. De modo que actúa como filtro para distinguir qué resultados son realmente importantes y cuáles podrían ser falsos positivos.
- *glmnet* facilita la implementación de técnicas de regresión regularizada (LASSO, Ridge y Elastic Net) con capacidad de selección automática de variables. Se utiliza cuando existen muchas variables posibles (edad, sexo, alelos HLA…), ayudando a seleccionar cuáles son realmente importantes para la predicción y descartar las que no son relevantes.
- *MASS* proporciona herramientas avanzadas para el análisis estadístico, incluyendo métodos de regresión robusta y análisis de multivariables. Es especialmente útil cuando

los datos no son perfectos o tienen "ruido", ya que incluye métodos que son más resistentes a estos problemas.

- El paquete *caret* se utiliza para el entrenamiento y validación de modelos, implementando validación cruzada y estandarizando las comparaciones entre modelos. Actúa dividiendo los datos en grupos para entrenar y probar el modelo (validación cruzada), y nos ayuda a comparar diferentes modelos de manera justa
- Por último, *brglm2* se emplea para realizar regresiones logísticas con reducción de sesgo, especialmente útil en casos de separación completa, proporcionando estimaciones más robustas y fiables. Es decir, cuando los datos muestran patrones demasiado perfectos, nos da resultados más conservadores, pero más fiables.

3.4.3. Paquetes de visualización

La visualización de los resultados se implementó mediante un conjunto integrado de paquetes de R especializados en representación gráfica: *ggplot2* actúa como el motor principal de la visualización, proporcionando un sistema completo basado en la "gramática de gráficos", permitiendo construir visualizaciones básicas, pero también gráficos complejos con varios tipos de información. *corrplot* se utiliza específicamente para la visualización de matrices de correlación, facilitando la identificación de patrones y asociaciones entre alelos HLA y variables clínicas. *gridExtra* permite combinar múltiples gráficos en presentaciones coherentes y optimizadas, siendo útil para comparar resultados entre diferentes loci HLA. Por último, *ggrepel* optimiza la legibilidad de los gráficos mediante un sistema inteligente de colocación de etiquetas.

3.5. Imputación de alelos HLA mediante HIBAG

3.5.1. Preparación del modelo de imputación y datos genéticos

El modelo pre-entrenado que se utiliza es “Human660W-Hispanic-HLA4-hg19”. Se basa en la plataforma *Human660W* y está optimizado para la región HLA del genoma hg19, incluyendo información clave sobre patrones de desequilibrio de ligamiento que mejoran la precisión de la imputación.

Los datos genéticos en archivo PLINK (BED para los datos binarios de genotipos; BIM para la información de marcadores y FAM para la información de muestras) requieren de una verificación de su calidad, por lo que se hace un control de *missing data* (NA), se verifica la consistencia de los SNPs y se comprueban los formatos de entrada.

3.5.2. Proceso de imputación por *locus*

El procedimiento de imputación se realiza para cada uno de los loci HLA analizados, abarcando tanto los de clase I (A, B, C) como los de clase II (DRB1, DQA1, DQB1, DPA1, DPB1). Sin embargo, hay que tener en cuenta que para DPA1 no hay datos en el modelo, por lo que no se podrá analizar. Cada locus fue procesado individualmente, considerando sus características específicas y su relevancia funcional (Tabla 1).

Tabla 1: Código en R para la preparación del modelo y los datos genéticos; además de para ejecutar la imputación.

```
for(locus in hla_loci) {
    # Obtención del modelo específico
    model <- hlaModelFromObj(model.list[[locus]])

    # Preparación de datos
    sample_ids <- yourgeno$sample.id
    clinical_matched <- datos_clinicos[match(sample_ids, datos_clinicos$id),]

    # Procesamiento de covariables
    covariates <- prepareCovariates(clinical_matched)

    # Imputación
    pred.guess <- predict(model, yourgeno,
                          type = "response+prob",
                          covariates = covariates)
}
```

El control de calidad basado en las probabilidades posteriores tras la imputación se categoriza en tres niveles de confianza:

- Baja confianza (< 0,5): estas predicciones son consideradas poco fiables al estar por debajo del 50% en cuanto a la probabilidad posterior. En muchos casos, estas predicciones podrían ser excluidas del análisis final para evitar introducir ruido en los resultados.
- Media confianza (0,5-0,75): estas predicciones son aceptables, aunque se deben interpretar cuidadosamente.
- Alta confianza (> 0,75): estas son nuestras predicciones de más alta fiabilidad. Representan aquellos casos en que la probabilidad de que el alelo imputado es correcto es del más del 75%.

Finalmente se generan unas métricas de calidad como la tasa de éxito, que calcula qué porcentaje de las predicciones alcanzan cada nivel de confianza. Para ver la distribución de probabilidades se generan histogramas agrupados por frecuencias que muestran como se distribuyen las predicciones según sus probabilidades posteriores para cada uno de los locus.

3.6. Análisis estadístico HLA-COVID

Tras la imputación, se realizan los análisis de asociación entre los datos genéticos de la imputación HLA y los datos clínicos de cada paciente. De modo que se evalúa estadísticamente si existe una relación directa entre la variabilidad alélica de los individuos y la predisposición a padecer COVID-19 de una forma más severa.

3.7. Análisis de asociación sin ajuste

En primer lugar, se hace un análisis de asociación sin ajustar los datos en base a otras variables que podrían ser relevantes; con el objetivo de ver si se aprecian diferencias significativas tras el ajuste en base a dichas variables.

Para este análisis se utiliza el test de Fisher para evaluar la asociación entre cada alelo HLA y la severidad de la enfermedad COVID-19 mediante tablas de contingencia (Tabla 2). Estas tablas de contingencia organizan los datos en filas (presencia o ausencia del alelo) y columnas (severidad o no severidad), y el test proporciona un valor *p* que indica el valor de significancia para las asociaciones.

Tabla 2: Ejecución del test de Fisher en el software R.

```
fisher_test <- function(datos_filtrados, allele_col) {

     freq_table <- table(datos_filtrados[[allele_col]],
                         datos_filtrados$R_SEV4)

     test <- fisher.test(freq_table)

     return(test)

}
```

Además, se aplican varias pruebas de corrección como el método de Bonferroni, por el cual se ajusta el nivel de significancia en base al número total de alelos. También se aplica el FDR para controlar la proporción de falsos positivos. Y, finalmente, se ajusta la significación en base a las estas pruebas realizadas, de modo que: *p-valor < 0.05/n*, donde *n* es igual al número de pruebas realizadas.

3.7.1. Análisis de asociación: *odds ratios*

Para evaluar la asociación entre la variabilidad HLA y la severidad COVID-19 se utiliza el *odds ratio* (OR). Es una medida de asociación estadística que cuantifica la fuerza y relación entre dos variables binarias. Compara dos grupos diferentes evalúa la probabilidad de que un determinado evento ocurra o no ocurra.

Para calcular los valores de OR para cada alelo se sigue la fórmula: $OR = (a \times d) / (b \times c)$, donde a son los casos severos con el alelo; b son los casos no severos con el alelo; c son los casos severos sin el alelo; y d son los casos no severos sin el alelo. Si OR = 1, no existe asociación entre las variables. Si OR > 1 la asociación sería positiva y podríamos estar hablando de un factor de riesgo. Por último, si OR < 1 indica una asociación negativa y podría tratarse de un factor protector.

Es importante destacar que se aplican intervalos de confianza (IC) del 95% a estos valores que proporcionan información sobre la precisión de los OR. Un IC estrecho conlleva una mayor precisión y una estimación confiable, mientras que un IC amplio muestra una menor precisión y confiabilidad. En base a esto, se estima que si el IC incluye el OR = 1 se entiende que la asociación no es estadísticamente significativa, mientras que si no lo incluye sí será una asociación significativa.

Este método permite comparar entre diferentes factores de riesgo y, además, se puede ajustar fácilmente por variables confusoras mediante regresión logística, aspecto a tener en cuenta de cara a hacer un posterior análisis ajustado.

3.8. Análisis de asociación ajustado

Para evaluar más rigurosamente la asociación entre los alelos HLA y la severidad de COVID-19 en los pacientes se hace también un análisis de asociación, pero en este caso de ajusta mediante regresión logística multivariable. Esto permite controlar el efecto de posibles factores de confusión en la asociación de los alelos HLA frente a un fenotipo de severidad de COVID-19. En este trabajo se hace un ajuste en base a las variables de edad, sexo y las diez componentes

principales presentes en los datos clínicos, agrupando así la estructura poblacional en relación a los datos.

3.8.1. Componentes del modelo de regresión logística

En primer lugar, vamos a tener una serie de variables independientes que se van a incluir en el modelo: la edad, el sexo y las componentes principales. Mediante estas variables se realiza un ajuste por factores que podrían afectar a las predicciones de asociación. Además, la variable dependiente en nuestro caso será la severidad de COVID-19, codificada de forma binaria para casos severos (1) y no severos (0).

3.8.2. Modelo de regresión logística

En la fórmula del modelo de regresión logística se especifica la relación entre la variable dependiente de interés (severidad de COVID-19) y las variables independientes. De este modo, se analiza cada alelo HLA por separado, ajustando de forma sistemática las variables de edad, sexo y las componentes principales (Tabla 3).

Tabla 3: Función para la fórmula y ajuste del modelo de regresión logística.

```
analyze_allele_adjusted <- function(datos, alelo) {
    # Fórmula del modelo
    formula <- as.formula(paste("R_SEV4 ~", alelo,
                                "+ edad_scaled + sexo_num + ",
                                paste(paste0("PC", 1:10, "_scaled"),
                                      collapse = " + ")))

    # Ajuste del modelo
    modelo <- glm(formula,
                  family = binomial(link = "logit"),
                  data = datos)

    return(modelo)
}
```

Este modelo sigue los principios de la regresión logística binaria, puesto que se quiere predecir un resultado para dos posibles valores, en este caso asociados a la severidad (severo/no severo). Para ello se utiliza la función de enlace “logit” como transformador matemático que mantiene los valores de las predicciones en base a las tres variables independientes entre 0 y 1.

3.8.3. Cálculo de OR ajustados

De un modo similar a como se procedió para el análisis sin ajuste mediante variables independientes, se calculan los OR ajustados con sus correspondientes intervalos de confianza (Tabla 4). La interpretación de valores significativos se lleva a cabo con los mismos criterios estadísticos.

Tabla 4: Código en R para el cálculo de los OR ajustados con sus respectivos IC 95%.

```
calculate_adjusted_OR <- function(modelo) {
    # Extraer coeficientes
    coef <- summary(modelo)$coefficients

    # Calcular OR e IC95%
    OR <- exp(coef[2,1])  # Coeficiente del alelo
    IC_lower <- exp(coef[2,1] - 1.96*coef[2,2])
    IC_upper <- exp(coef[2,1] + 1.96*coef[2,2])
```

```
    p_valor <- coef[2,4]

    return(list(or = or,
                IC_lower = IC_lower,
                IC_upper = IC_upper,
                p_valor = p_valor))
}
```

3.8.4. Control de calidad del modelo

En primer lugar, se hace un análisis de residuos para comprobar las predicciones del modelo de asociación HLA-COVID. De modo que se buscan valores que se alejen del patrón esperado de residuos, detectando anomalías que indiquen una relación incorrecta entre los alelos HLA y la severidad de COVID-19. También se hace un análisis de Leverage para identificar combinaciones inusuales de alelos o características de cada paciente; y un análisis de valores influyentes mediante la distancia de Cook, con el objetivo de identificar aquellos casos que afectan a las conclusiones del estudio utilizando las características del individuo (como Leverage) y el resultado (severidad COVID). Por último, se evalúa globalmente el modelo mediante el test de bondad de ajuste Hosmer-Lemeshow. La ejecución de estos controles post-análisis se aprecia en la Tabla 5.

Tabla 5: Código en R para ejecutar el control de calidad del modelo tras la realización del análisis de asociación.

```
check_model_diagnostics <- function(modelo) {
    # Residuos
    residuos <- residuals(modelo, type = "deviance")

    # Valores influyentes
    cook_dist <- cooks.distance(modelo)

    # Leverage
    leverage <- hatvalues(modelo)

    # Test de bondad de ajuste
    hoslem_test <- hoslem.test(modelo$y, fitted(modelo))

    return(list(residuos = residuos,
                cook = cook_dist,
                leverage = leverage,
                hoslem = hoslem_test))
}
```

3.9. Análisis Multivariable de Variables Clínicas

Se realiza un análisis multivariable para examinar las correlaciones entre las tres principales variables clínicas (severidad, hospitalización y estado asintomático), ajustando por los mismos factores de confusión (edad, sexo y componentes principales).

Para ello se implementan tres modelos de regresión logística que se ejecutan de forma paralela (Tabla 6). Por lo que se calculan, para cada alelo, las probabilidades ajustadas para severidad, hospitalización y asintomatología.

Tabla 6: Modelos de regresión logística para análisis multivariable

```
# Modelo para severidad
modelo_sev <- glm(R_SEV4 ~ sexo_num + edad_scaled + PC1_scaled + ... +
PC10_scaled,
                  family = binomial)

# Modelo para hospitalización
```

```
modelo_hosp <- glm(R_HOSP ~ sexo_num + edad_scaled + PC1_scaled + ... +
PC10_scaled,
                   family = binomial)

# Modelo para estado asintomático
modelo_asint <- glm(ASINT ~ sexo_num + edad_scaled + PC1_scaled + ... +
PC10_scaled,
                    family = binomial)
```

Para el análisis estadístico se calculan las matrices de correlación para cada binomio de variables: severidad-hospitalización, severidad-asintomatología y hospitalización-asintomatología. Al igual que en análisis anteriores, se verifica la normalidad de las probabilidades y se detectan valores atípicos.

Este análisis multivariable permite una evaluación más precisa de las relaciones existentes entre las diferentes variables clínicas de COVID-19, debidamente ajustadas por factores demográficos y ancestrales potencialmente confusores.

4. Resultados

4.1. Imputación de HLA mediante HIBAG

La imputación de alelos HLA representa un desafío en genética de poblaciones, sobre todo si lo aplicamos en poblaciones con ancestría mixta como las latinoamericanas. Los resultados obtenidos en este trabajo muestran una clara variabilidad en los porcentajes de confianza para los loci analizado (Tabla 6), reflejando tanto la complejidad del sistema HLA como las particularidades de la estructura genética de la población objeto de estudio.

Tabla 7: Porcentaje de confianza en las predicciones para cada uno de los *loci* HLA.

	Porcentaje de confianza en las predicciones (%)		
	< 50%	**50-75%**	**> 75%**
HLA-A	18,93	24,89	56,19
HLA-B	69,07	16,61	14,33
HLA-C	15,09	22,19	62,72
HLA-DPB1	28,59	30,11	41,31
HLA-DQA1	33,03	38,43	28,54
HLA-DQB1	26,47	35,30	38,22
HLA-DRB1	69,64	18,42	11,94

HLA-C muestra el mejor rendimiento, con un 62,72% de las predicciones por encima del 75% de confianza; seguido de HLA-A, con un 56,19%. Por otro lado, HLA-B y HLA-DRB1 presentan los niveles más bajos de predicciones de alta confianza (14,33% y 11,94% respectivamente), además de unos valores significativamente altos en el intervalo de bajas confianza (69,07% y 69,64%, respectivamente). Los loci restantes (DQA1, DQB1 y DPB1) muestran un rendimiento intermedio en sus predicciones, teniendo un reparto muy similar entre los tres niveles de confianza. Podemos destacar DPB1, que tiene el 41,31% de sus predicciones de alta confianza.

4.2. Análisis de severidad: variables demográficas

El análisis de los factores demográficos en relación con la severidad de la enfermedad reveló asociaciones significativas tanto para la edad como para el sexo, así como una relación importante entre ambas variables.

El sexo parece tener una fuerte influencia en la severidad de la enfermedad, con diferencias significativas entre hombres y mujeres ($p < 2{,}2e\text{-}16$). Los hombres presentaron una proporción considerablemente mayor de casos severos, en torno al 30%, mientras que en las mujeres se redujo al 10%. El OR de 0,3564 (IC 95%: 0,3008 - 0,4216) indica un efecto protector en mujeres, presentando, aproximadamente, una tercera parte del riesgo del fenotipo severo en comparación con los hombres.

La edad muestra una relación positiva y significativa con la severidad de la enfermedad, que se refleja en el modelo de regresión logística ($\beta = 0{,}073554$ y $p < 2e\text{-}16$). Por cada año de edad que se incrementa, se observa un aumento en la expresión logarítmica del OR para la severidad, lo que parece indicar un aumento exponencial del riesgo a medida que pasan los años.

El análisis combinado de edad y sexo muestra patrones más complejos que los efectos individuales. Ambas variables no solo tienen efectos independientes, sino que también interactúan de manera significativa ($\beta = 0{,}025394$ y $p = 0{,}000175$). El efecto protector del sexo femenino ($\beta = -2{,}405688$ y $p = 3{,}91e\text{-}10$) se mantiene significativo incluso después de ajustar por edad, mientras que el efecto de la edad continúa siendo positivo y significativo ($\beta = 0{,}062665$ y $p < 2e\text{-}16$).

Al graficar los resultados por grupos de edad se pueden observar patrones de interacción entre ambas variables (Figura 4). En el grupo de 0-20 años, las diferencias entre sexos son mínimas. Esta diferencia empieza a aumentar en el rango de 21-40 años, llegando a su máxima expresión en el grupo de 61-80 años. A partir de los 80 años se ve una diferencia menor, primando la edad como factor determinante por encima edl sexo.

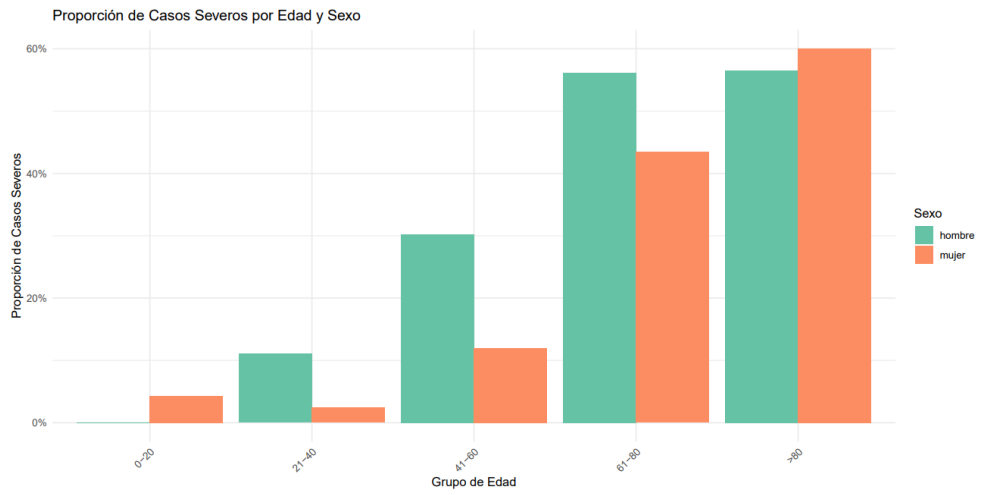


Figura 4: Gráfico de proporción de casos severos respecto del sexo y la edad de los individuos.

Un factor determinante que afecta a la precisión de la imputación en poblaciones latinoamericanas es su compleja estructura poblacional. Como muestran Arrieta-Bolaños et al. (2012), las poblaciones latinoamericanas presentan tres patrones principales de mezcla en relación a la ancestría: un componente amerindio o nativoamericano, un componente caucásico

o europeo y otro componente africano. Esta heterogeneidad genética impacta directamente en la precisión de la imputación y puede ser crucial de cara a análisis de asociación como el que ha realizado el presente trabajo.

Los datos utilizados en este estudio describen porcentajes de ancestría para las tres componentes ya mencionadas asociadas a cada paciente. Con el objetivo de comprobar si sería importante tener en cuenta estas variables para el estudio de asociación por COVID, se ha hecho un breve análisis de relación entre las componentes de ancestría y la severidad a COVID-19.

En la Figura 5 se puede ver un claro efecto protector del porcentaje de ancestría africana y europea, siendo más claro en el primer caso. Parece apreciarse que cuanto más porcentaje de ancestría europea o africana presenta un individuo, menor es la probabilidad de que desarrolle un cuadro grave de COVID-19.

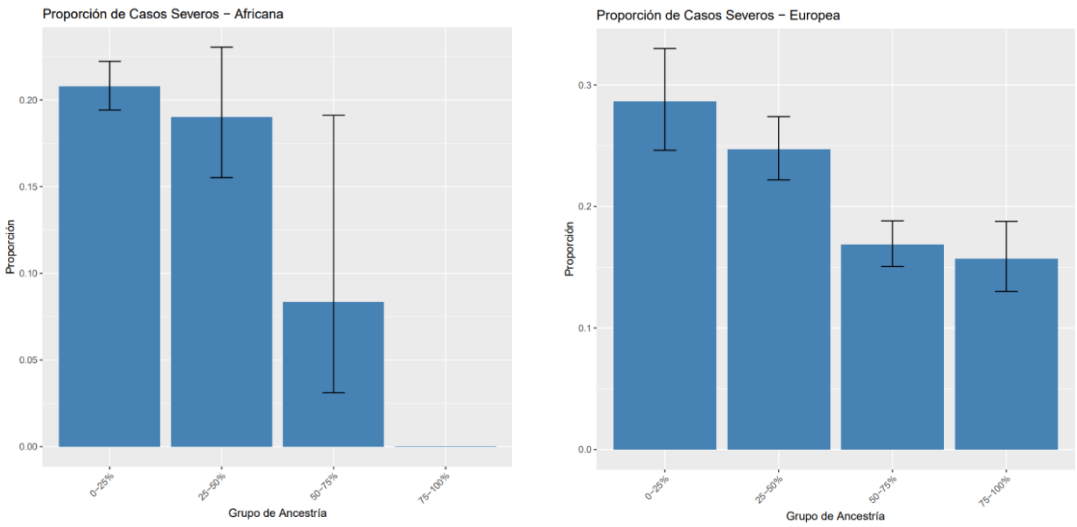


Figura 5: Proporción de casos severos de COVID-19 en relación al porcentaje de ancestría africana y europea.

Por otro lado, la Figura 6 muestra un claro efecto de riesgo para aquellos individuos que tienen un mayor porcentaje de ancestría nativoamericana.

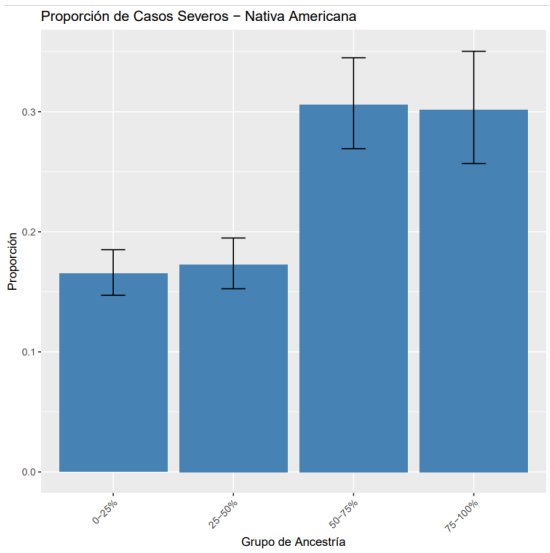


Figura 6: Proporción de casos severos de COVID-19 en relación al porcentaje de ancestría nativa americana.

4.3. Análisis de asociación HLA-COVID

4.3.1. Estudio de los alelos HLA con la severidad de COVID-19

En el análisis de asociación entre la variabilidad alélica y la severidad de COVID-19 se han identificado una serie de alelos para cada locus HLA que parecen representar un factor de riesgo o protección para los individuos. Para este estudio se ha tomado como límite para la significación $p < 0,5$; y un $OR > 1$ para llegar a considerar como riesgo u $OR < 1$ para considerarlo como alelo protector.

Para HLA-A se han analizado 18 alelos distintos, siendo los más frecuentes: A*02:01, A*24:02 y A*01:01. Al comparar los resultados antes y después del ajuste por variables independientes (sexo, edad y PCs), dos alelos han pasado de ser factores significativos de riesgo a no significativos. Es el caso del alelo A*02:06, que ha cambiado de riesgo (OR = 3,430 y p = 0,0004) a no significativo (OR = 1,408 y p = 0,3957); y del alelo A*31:01, que ha cambiado también de riesgo (OR = 1,777 y p = 0,0167) a no significativo (OR = 1,650 y p = 0,0605) (Anexo II).

Para la región HLA-B se han analizado 42 alelos diferentes, con una mayor frecuencia para los alelos B*07:02, B*44:02 y B*08:01. En este caso se han encontrado dos asociaciones significativas relacionadas con el riesgo a desarrollar sintomatología severa de COVID-19. Los alelos asociados a este riesgo son: B*15:03 (OR = 2,200 y p = 0,0103) y B*50:01 (OR = 4,419 y p = 0,0097). Antes del ajuste estos alelos no se habían identificado como factor de riesgo (OR = 1,583 y p = 0,0800; OR = 1,831 y p = 0,2474); y, además, se habían identificado otros alelos como significativos protectores o de riesgo: B*15:15 (de OR = 3,541 y p = 0,0114 a OR = 0,790 y p = 0,6686), B*15:16 (de OR = 2,653 y p = 0,0443 a OR = 1,557 y p = 0,4271), B*40:02 (de OR = 1,475 y p = 0,0382 a OR = 1,476 y p = 0,0673) y B*48:01 (de OR = 2,141 y p = 0,0213 a OR = 1,345 y p = 0,4334) (Anexo III).

En la región HLA-C se han analizado 20 alelos en total, siendo los más frecuentes: C*07:01, C*07:02 y C*04:01. Se ha encontrado que el alelo C*05:01 (OR = 0,599 y p = 0,0317) se asocia de manera significativa como factor protector para la severidad de COVID-19. Es además destacable que este alelo, antes del ajuste, no era significativo (OR = 0,718 y p = 0,1118); al contrario de como ocurre con los alelos no significativos que sí lo eran previo al ajuste: C*01:02 ha perdido significación de riesgo (de OR = 1,333 y p = 0,0149 a OR = 1,207 y p = 0,1741) y C*08:02 ha perdido efecto protector (de OR = 0,503 y p = 0,0422 a OR = 0,593 y p = 0,1723) (Anexo IV).

Pasando a los loci de clase II, para HLA-DPB1 se han analizado 13 alelos, con mayor frecuencia para los alelos DPB1*04:01, DPB1*04:02 y DPB1*02:01. El alelo DPB1*04:02 había mostrado un efecto de riesgo significativo antes del ajuste (OR = 1,251 y p = 0,0277), pero perdió al hacer el análisis ajustado (OR = 0,987 y p = 0,9126) (Anexo V).

Para la región HLA-DQA1 se analizaron también 13 alelos distintos, siendo los más frecuentes: DQA1*01:02, DQA1*05:01 y DQA1*01:01. En cuanto a la asociación con la severidad de COVID-19, el alelo DQA1*05:03 pasó a ser no un factor de riesgo significativo (OR = 1,601 y p = 0,0333) a ser no significativo (OR = 1,068 y p = 0,7895) (Anexo VI).

Se analizaron 13 alelos para HLA-DQB1, con más frecuencia para los alelos DQB1*03:01, DQB1*02:01 y DQB1*05:01. En este caso se ha encontrado un alelo con una asociación significativa de riesgo: DQB1*03:19 (OR = 2,945 y p = 0,0348) (Anexo VII).

Por último, se analizaron 28 alelos de la región HLA-DRB1, con mayores frecuencias en los alelos DRB1*07:01, DRB1*13:01 y DRB1*15:01. Tras el ajuste se perdió el efecto de riesgo significativo para los alelos DRB1*08:02 (de OR = 1,390 y p = 0,0222 a OR = 1,070 y p = 0,6929), DRB1*14:02 (de OR = 1,763 y p = 0,0149 a OR = 1,440 y p = 0,1674) y DRB1*14:06 (de OR = 3,947 y p = 0,0346 a OR = 1,506 y p = 0,5679); y el factor protector del alelo DRB1*11:01 (de OR = 0.537 y p = 0,0197 a OR = 0,566 y p = 0,0615). Sin embargo, dos alelos que no tenían asociaciones significativas tras el ajuste, pasaron a presentar un efecto protector y de riesgo, respectivamente: DRB1*04:02 (OR = 0,419 y p = 0,0202) y DRB1*15:01 (OR = 2,137 y p = 0,0360) (Anexo VIII).

En definitiva, el análisis ajustado por variables demográficas y componentes principales mostró seis asociaciones significativas entre alelos HLA y la severidad de COVID-19, con un predominio de alelos de riesgo sobre protectores. Entre los alelos de riesgo, B*50:01 obtuvo un valor de asociación más fuerte (OR = 4,419), seguido por DQB1*03:19 (OR = 2,945), B*15:03 (OR = 2,200) y DRB1*15:01 (OR = 2,137). Por otro lado, se identificaron dos alelos protectores: DRB1*04:02 (OR = 0,419) y C*05:01 (OR = 0,599), que se asociaron con una reducción significativa en el riesgo de severidad.

4.3.2. Análisis de correlación multivariable

Tras analizar la relación entre la severidad de COVID-19 y la variabilidad alélica HLA, también se puede visualizar la asociación con las variables clínicas de hospitalización y asintomatología.

En el caso de los genes HLA de clase I, el locus HLA-A muestra una correlación positiva entre la proporción de casos severos y hospitalizados (Figura 7). El alelo A*02:01, siendo el más frecuente en la población estudiada, presenta una proporción media de casos severos y hospitalizados. Destacamos el alelo A*02:06, que, aunque perdió su significación estadística tras el ajuste (pasando de OR = 3,430, p=0.0004 a OR = 1,408 y p = 0,3957), muestra una alta proporción de casos severos y hospitalizados en el análisis de correlación. Del mismo modo, A*31:01 parece mantener una correlación positiva entre severidad y hospitalización.

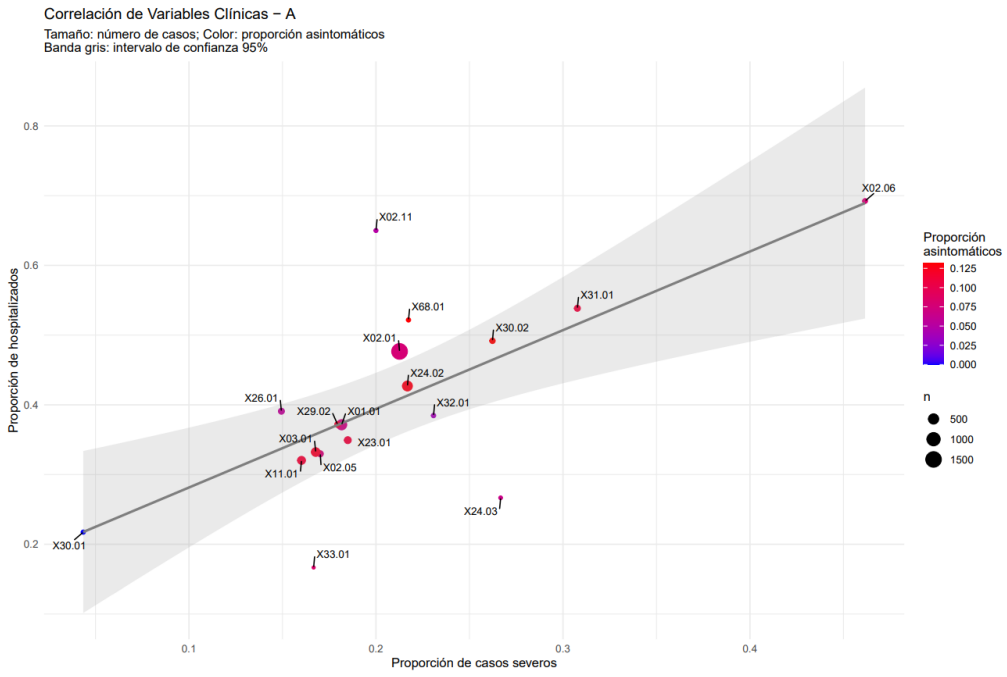


Figura 7: Gráfico de correlación de las variables clínicas severidad, hospitalización y asintomatología, en relación a los alelos HLA-A.

El locus HLA-B presenta la mayor variabilidad en las correlaciones observadas (Figura 8). Los alelos B*15:03 y B*50:01, clasificados estadísticamente como alelos de riesgo (OR = 2,200ç y p = 0,0103 y OR = 4,419 y p = 0,0097; respectivamente), muestran una elevada proporción de casos severos y de hospitalizados. Además, estos alelos presentan una menor proporción de casos asintomáticos.

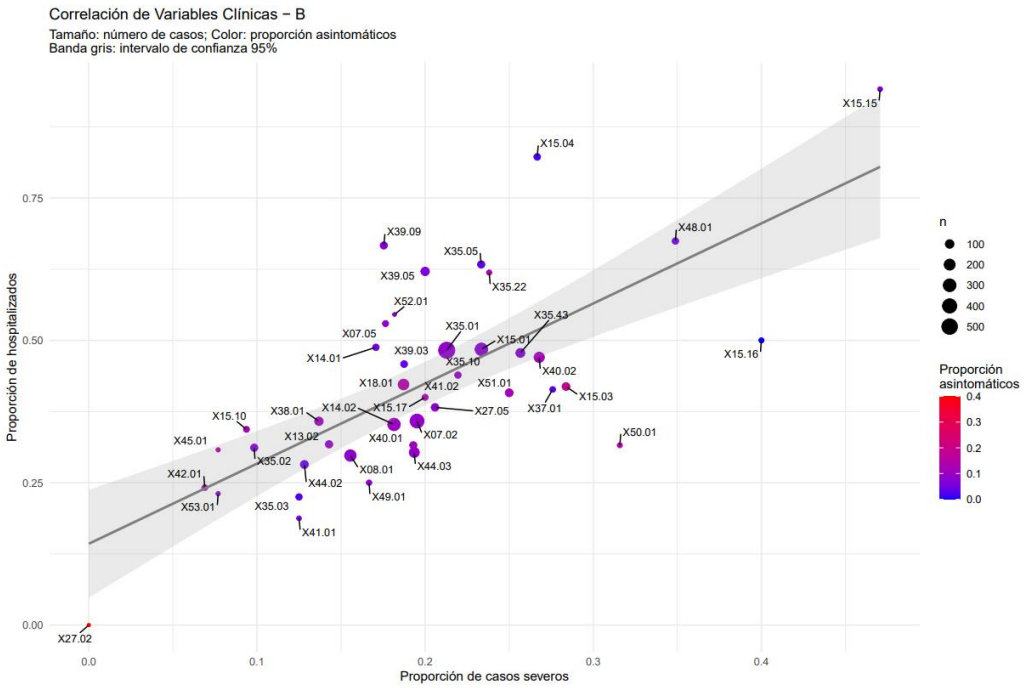


Figura 8: Gráfico de correlación de las variables clínicas severidad, hospitalización y asintomatología, en relación a los alelos HLA-B.

En cuanto a HLA-C, se observa una tendencia en la correlación positiva entre severidad y hospitalización (Figura 9). El alelo C*05:01, identificado como protector (OR = 0,599 y p = 0,0317), muestra una baja proporción de casos severos y hospitalizados.

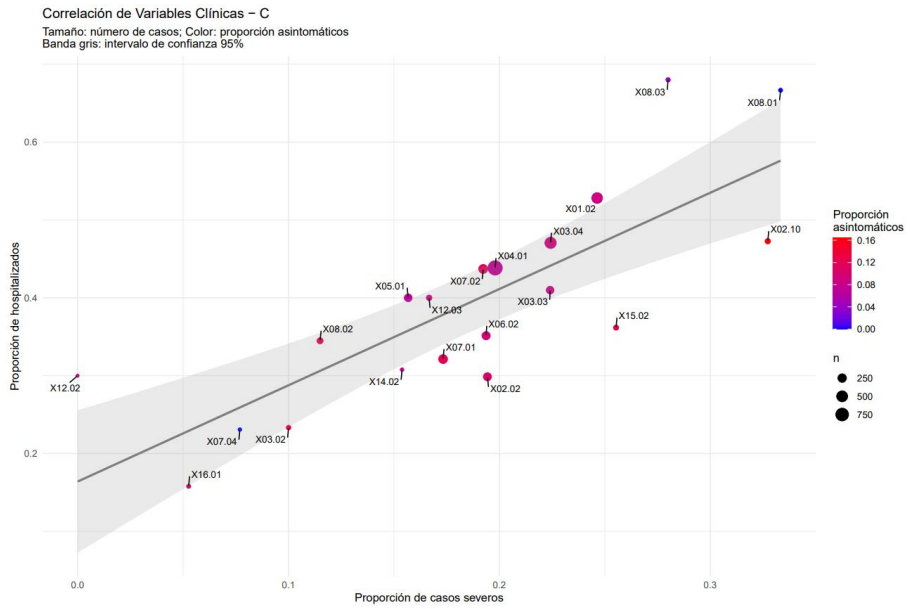


Figura 9: Gráfico de correlación de las variables clínicas severidad, hospitalización y asintomatología, en relación a los alelos HLA-C.

Los genes HLA de clase II muestran correlaciones más fuertes y consistentes que los de clase I. HLA-DPB1 muestra un patrón de correlación donde los alelos más frecuentes presentan proporciones moderadas de severidad y hospitalización, con una correlación positiva más fuerte que la observada en los genes de clase I (Figura 10).

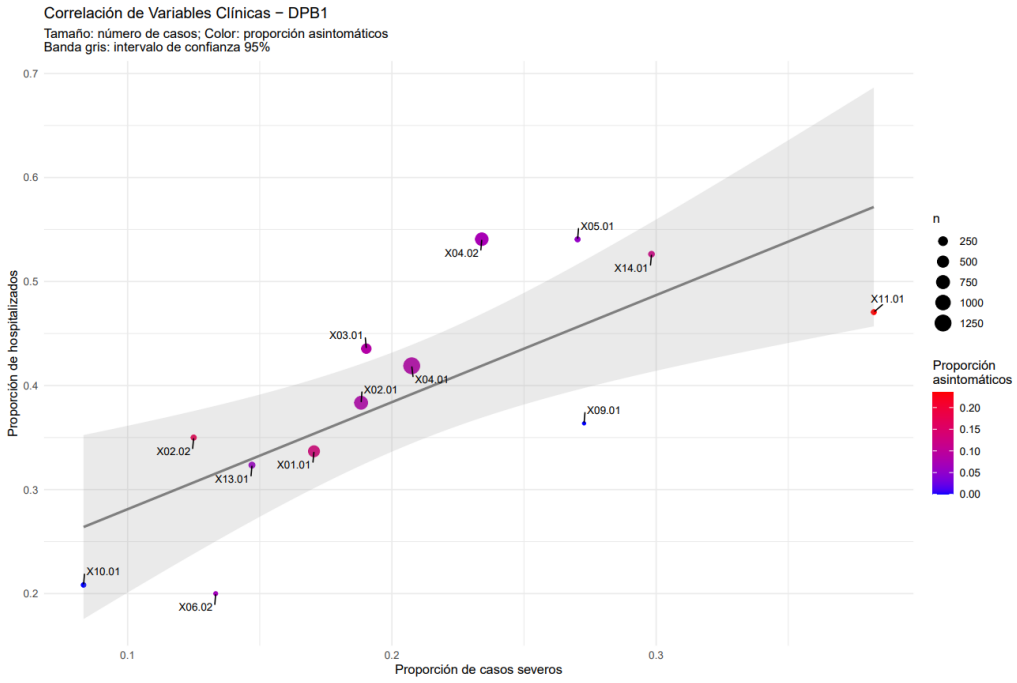


Figura 10: Gráfico de correlación de las variables clínicas severidad, hospitalización y asintomatología, en relación a los alelos HLA-DPB1.

El locus HLA-DQA1 presenta una correlación más clara entre severidad y hospitalización (Figura 11). Destacamos el alelo DQA1*05:03, que, aunque perdió su significación estadística tras el ajuste, mantiene una proporción moderada-alta de casos severos. Se observa además una relación inversa entre la proporción de asintomáticos y la severidad de la enfermedad.

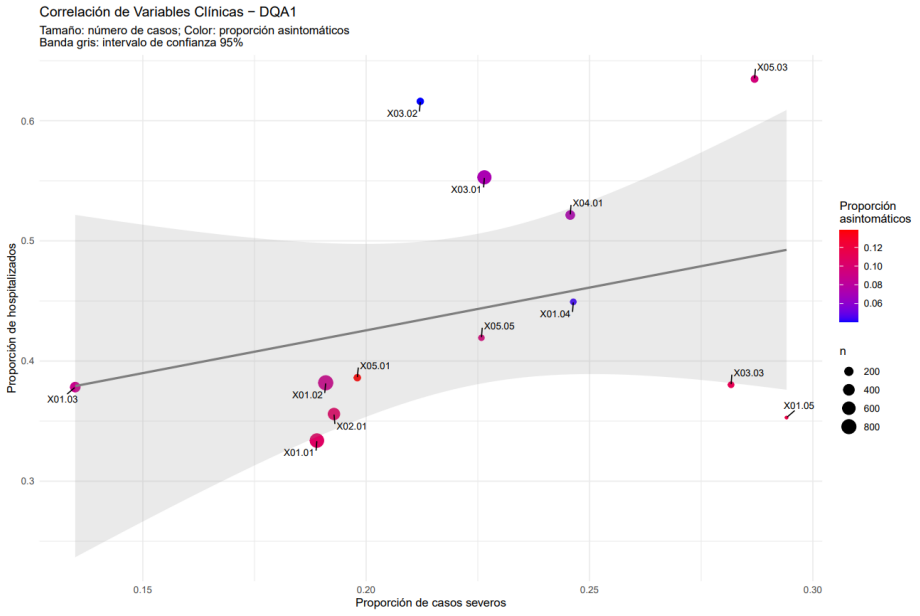


Figura 11: Gráfico de correlación de las variables clínicas severidad, hospitalización y asintomatología, en relación a los alelos HLA-DQA1.

En HLA-DQB1 destaca el alelo DQB1*03:19, identificado como de riesgo tras el ajuste (OR = 2.945 y p = 0.0348), mostrando una alta proporción de casos severos y hospitalizados (Figura 12). Sin embargo, la correlación general en este locus es más débil que en otros, lo que es concuerda con su test de independencia no significativo (p = 0.3281).

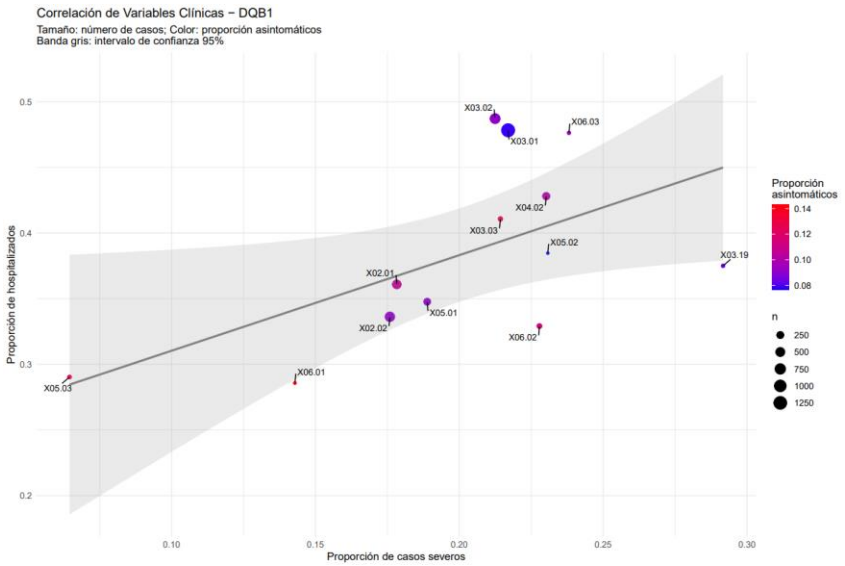


Figura 12: Gráfico de correlación de las variables clínicas severidad, hospitalización y asintomatología, en relación a los alelos HLA-DQB1.

Por último, el locus HLA-DRB1 muestra los patrones más claros de correlación (Figura 13). El alelo DRB1*04:02, identificado como protector (OR = 0,419 y p = 0,0202), presenta una baja proporción de casos severos y hospitalizados. Mientras que, en contraposición, el alelo DRB1*15:01, identificado como de riesgo (OR = 2,137 y p = 0,0360), muestra una fuerte correlación positiva entre severidad y hospitalización. Además, en general este locus presenta la correlación más clara entre la proporción de casos asintomáticos y los niveles de severidad y hospitalización.

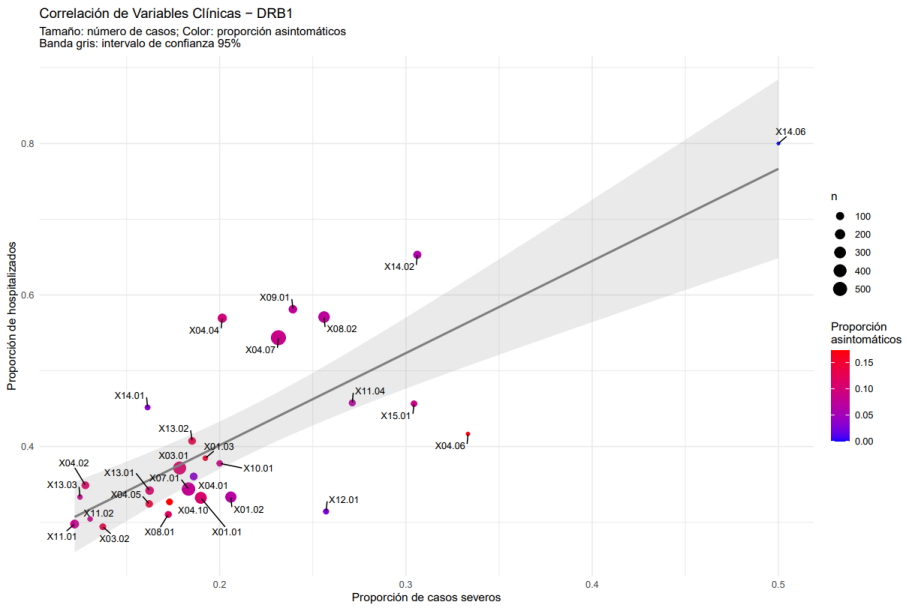


Figura 13: Gráfico de correlación de las variables clínicas severidad, hospitalización y asintomatología, en relación a los alelos HLA-DRB1.

5. Discusión

5.1. Imputación de HLA mediante HIBAG

5.1.1. Genes HLA de clase I

HLA-A muestra una confianza considerablemente alta (56,19%) que se puede apreciar en la Figura 14, lo cual concuerda con estudios previos que describen una menor diversidad alélica en este locus comparado con HLA-B y HLA-C (Robinson et al., 2020).

Este mejor rendimiento en las predicciones también se debe a regiones flanqueantes mejor conservadas y patrones de desequilibrio de ligamiento más estables (Zheng et al., 2014). Otro aspecto importante descrito por Degenhardt et al. (2019) es que HLA-A un mayor número de SNPs marcadores para las predicciones, proporcionando una base firme de cara a la imputación. Por otro lado, en poblaciones con componente asiático, alelos específicos como A*02:01 y A*02:03, pueden presentar problemas para generar buenos resultados de imputación (Dilthey et al., 2013).

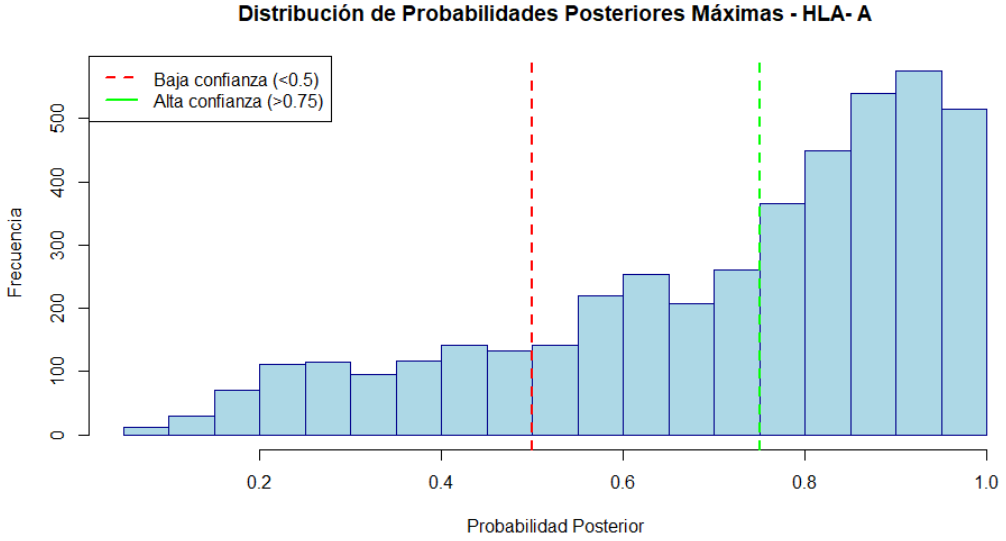


Figura 14: Distribución de probabilidades posteriores de imputación de la región HLA-A.

HLA-B presenta el menor valor de alta confianza de imputación (14,33%) entre los genes de clase I (Figura 15). Esto era un hecho esperable de acuerdo con el trabajo de Robinson et al. (2020), que describe a HLA-B como el locus más polimórfico del sistema HLA, con más de 7,000 alelos descritos. Este rendimiento tan bajo en las predicciones es debido a una alta tasa de recombinación, posiblemente ocasionada por la variabilidad en la ancestría en los diferentes países, resultando todo ello en patrones de desequilibrio de ligamiento más complejo y que dependerá de cada país. En definitiva, los resultados de este trabajo concuerdan con estudios similares que demuestran, cada vez más, una menor precisión de imputación para HLA-B en diversas poblaciones y que se necesitan paneles de referencia adecuados (Zhou et al., 2016).

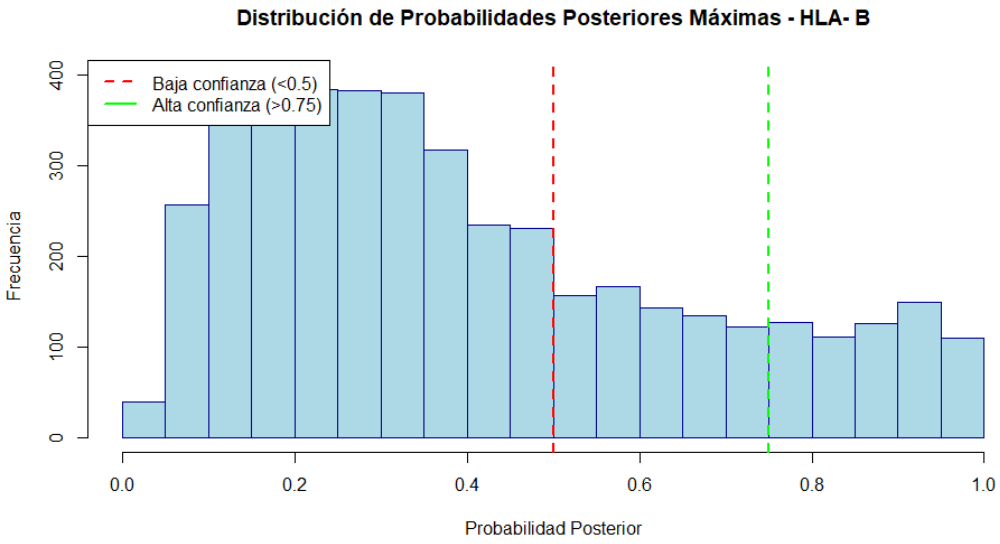


Figura 15: Distribución de probabilidades posteriores de imputación de la región HLA-B.

HLA-C muestra el valor más alto de confianza de imputación (62,72%) entre los genes de clase I (Figura 16). Esto se debe a su menor diversidad alélica en comparación con HLA-B, además de un fuerte desequilibrio de ligamiento. Cabe añadir que esta región genómica es más estable evolutivamente, lo que facilita la identificación de SNPs informativos y facilita su imputación incluso en poblaciones mezcladas (Dilthey et al., 2013; Karnes et al.,2017).

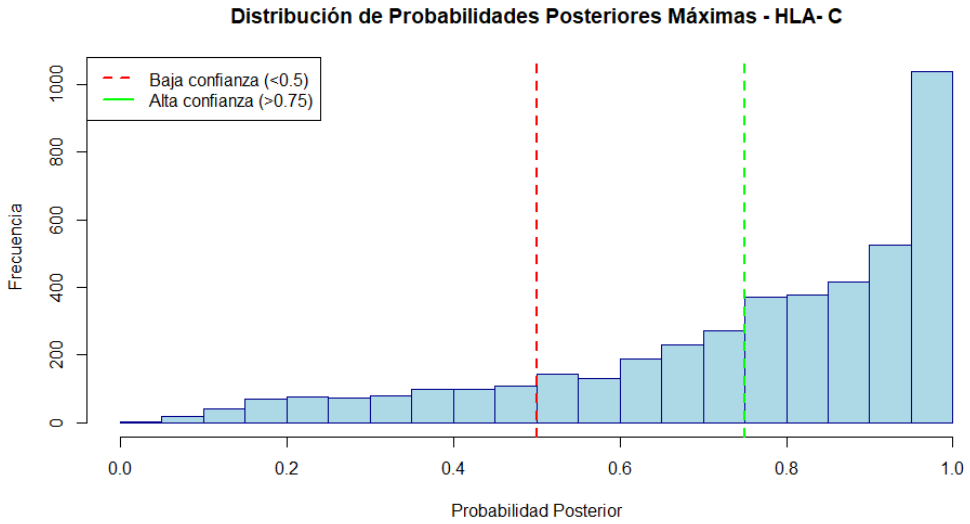


Figura 16: Distribución de probabilidades posteriores de imputación de la región HLA-A.

5.1.2. Genes HLA de clase II

HLA-DRB1 presenta un porcentaje bajo de predicciones de alta confianza (11,94%), como se aprecia en la Figura 17. Este hecho se explica por su alto polimorfismo, como ocurre con HLA-B, y la compleja estructura de la región DR, que incluye múltiples pseudogenes. Específicamente está descrito en estudios como el de Degenhardt et al. (2019) que la imputación de DRB1 se dificulta con los alelos DRB1*04:03/04:04 y DRB1*11:01/11:04. Este efecto se ve incrementado en poblaciones latinoamericanas debido a patrones de desequilibrio de ligamiento complejos

que dependen del país de origen, lo que conllevará un porcentaje de ancestría diferente (Okada et al., 2015).

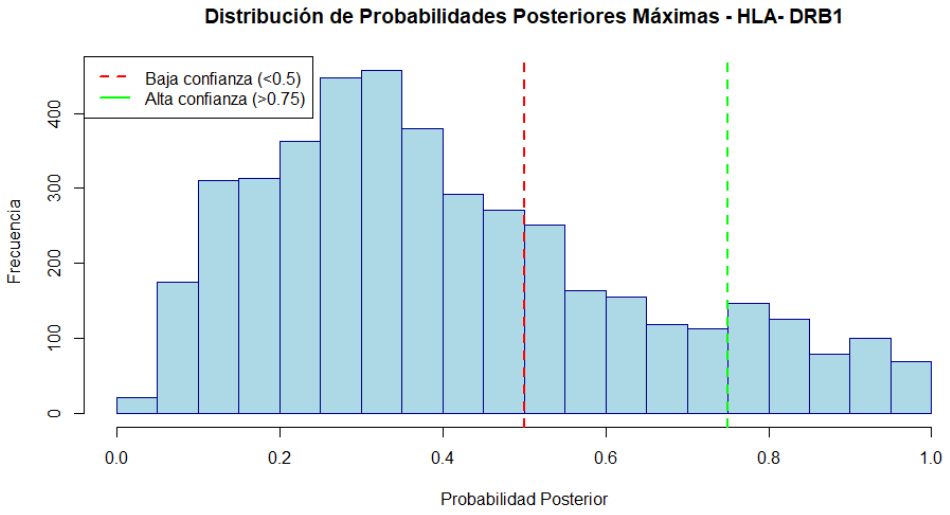


Figura 17: Distribución de probabilidades posteriores de imputación de la región HLA-DRB1.

Los genes HLA-DQA1 y DQB1 muestran unos valores intermedios para las predicciones de alta confianza: 28,54% y 38,22%, respectivamente (Figura 18 y 19). Ambos loci presentan un fuerte desequilibrio de ligamiento con DRB1, lo que podría facilitar la imputación, pero también puede arrastrar los errores de imputación ocasionados en DRB1 (Dilthey et al., 2013). De acuerdo a estudios anteriores, estos loci presentan un mejor rendimiento en la imputación en poblaciones mezcladas, debido a una menor diversidad alélica respecto a DRB1 (Degenhardt et al., 2019; Zhou et al., 2016).

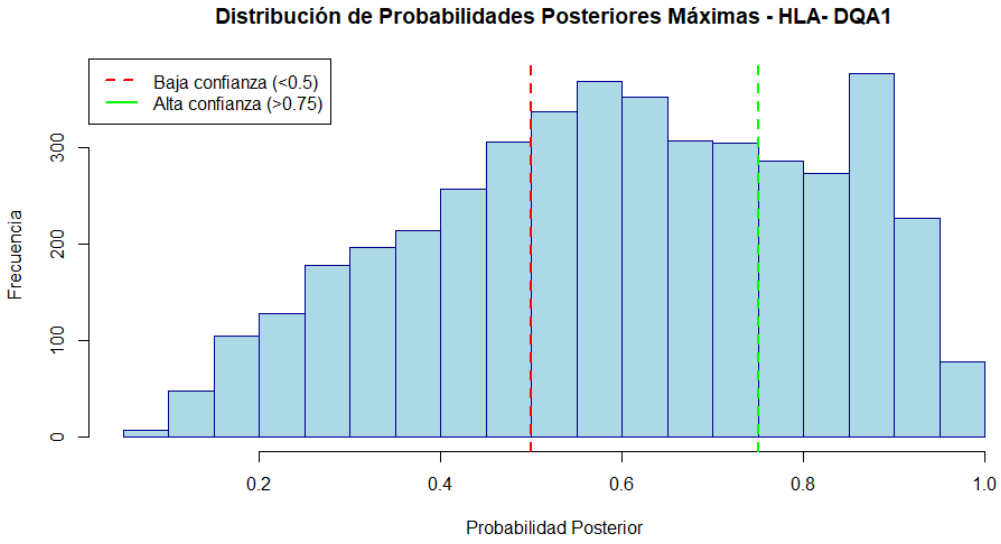


Figura 18: Distribución de probabilidades posteriores de imputación de la región HLA-DQA1.

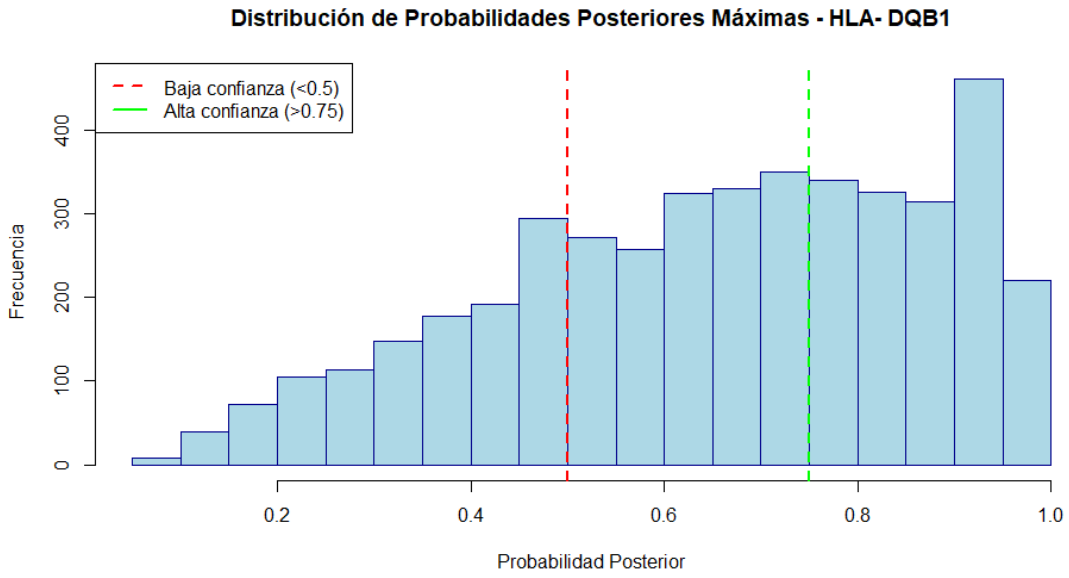


Figura 19: Distribución de probabilidades posteriores de imputación de la región HLA-DQB1.

HLA-DPB1, con un porcentaje de confianza alta del 41,31%, presenta una menor diversidad alélica en comparación con otros loci de clase II. Según los datos de la base IPD-IMGT/HLA (Robinson et al., 2020), mientras que HLA-DRB1 tiene más de 2,800 alelos descritos, HLA-DPB1 tiene aproximadamente 1,000 alelos, lo que simplifica el proceso de imputación e incrementa su rendimiento final (Figura 20).

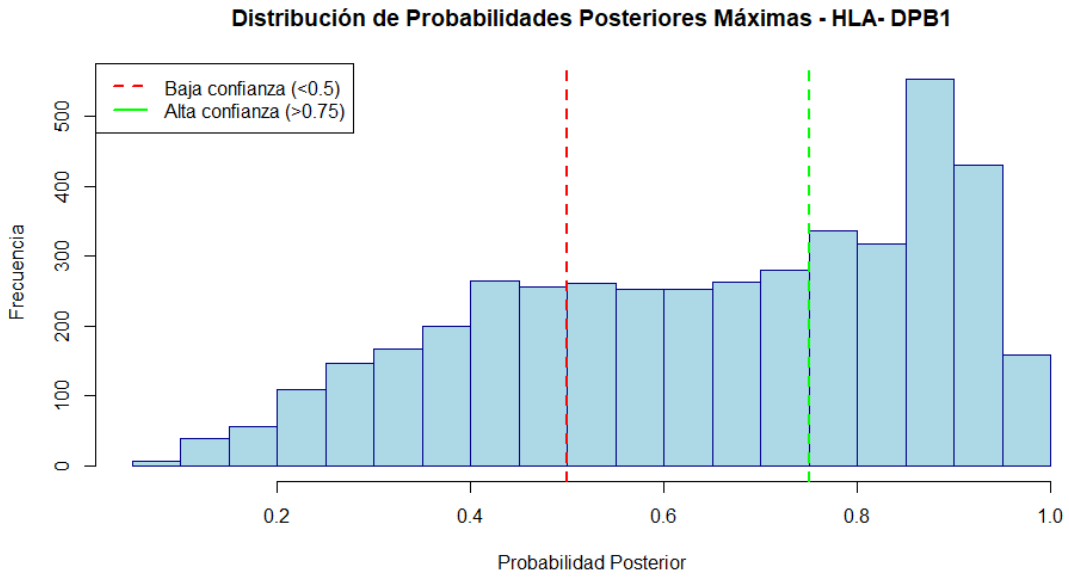


Figura 20: Distribución de probabilidades posteriores de imputación de la región HLA-DPB1.

Un factor técnico importante que contribuye a la mejor imputación de DPB1 es la presencia de un mayor número de SNPs informativos. Como describieron Okada et al. (2015), HLA-DPB1 tiene 51 SNPs *on target*, uno de los que más tiene respecto a otros loci HLA, lo que mejora

significativamente la capacidad de capturar la variación alélica mediante imputación. Además, Karnes et al. (2017) demuestran que la precisión de la imputación de DPB1 mantiene su rendimiento incluso en poblaciones mezcladas; aunque si se utilizan paneles de referencia específicos de población, como en este trabajo, la precisión de la imputación aumentar significativamente (Zhou et al., 2016).

5.1.3. Limitaciones técnicas de la imputación

La calidad de la imputación también se ve afectada por características genómicas específicas, como las tasas variables de recombinación entre loci, la selección de SNPs informativos, y la presencia de pseudogenes. Además, se ha demostrado que la selección cuidadosa de SNPs marcadores puede mejorar significativamente la precisión de la imputación (Okada et al., 2015). Para optimizar el rendimiento de la imputación en estas poblaciones se podría utilizar umbrales de probabilidad posterior específicos para cada loci y población (Zheng et al., 2014).

Las limitaciones actuales incluyen el tamaño relativamente pequeño de los paneles de referencia para ciertas poblaciones y la dificultad para capturar toda la diversidad HLA. Como se indica en el trabajo de Degenhardt et al. (2019), se necesitan paneles de referencia más grandes y diversos para mejorar la precisión de la imputación, especialmente en poblaciones no europeas, como ocurre en este estudio, que se ve limitado por los paneles de referencia existentes. Idealmente se deberían usar paneles específicos de población, donde se tenga en cuenta la variabilidad propia de la región analizada.

5.2. Análisis de severidad: variables demográficas

Los datos mostrados anteriormente muestran una fuerte asociación entre el sexo y la severidad ($p < 2,2e-16$), mostrando un claro efecto protector en mujeres (OR = 0,3564, IC 95%: 0,3008-0,4216). Este dimorfismo sexual puede darse por varios motivos, como las diferencias en la respuesta inmune innata y adaptativa, los efectos de las hormonas sexuales en la expresión de ACE2 o variaciones en la respuesta inflamatoria; y se ha observado para distintas poblaciones (Klein y Morgan, 2020; Takahashi et al., 2020).

La relación positiva y significativa entre edad y severidad ($\beta = 0,073554$, $p < 2e-16$) puede deberse a una gran diversidad de factores poblacionales y biológicos (Mueller et al., 2020). Además, el incremento logarítmico que se ha observado en este estudio concuerda con lo descrito por Zhang et al. (2021), donde se demostró que la relación edad-severidad sigue una tendencia exponencial.

La interacción significativa entre edad y sexo ($\beta = 0.025394$, $p = 0.000175$) puede deberse a varios factores que se pueden agrupar en variaciones hormonales dependientes de la edad y diferencias en comorbilidades (Scully et al., 2020). Además, la evolución descrita en los resultados de este estudio de la brecha entre los sexos según se incrementa la edad coincide con lo descrito en estudios como el de Li et al. (2020); observándose diferencias mínimas en edades tempranas, que se incrementan en edades intermedias y persisten en edades avanzadas.

Tener estas variables en consideración de cara a un análisis de correlación entre la variabilidad HLA y la severidad COVID-19 es clave, hecho por el cual se hizo un ajuste en base a estas variables junto con los componentes principales.

5.3. Análisis de asociación HLA-COVID

5.3.1. Estudio de los alelos HLA con la severidad de COVID-19

Los datos extraídos estudio parecen indicar ciertas asociaciones significativas entre varios alelos HLA y la severidad de COVID-19 en población latinoamericana.

Entre los alelos de riesgo identificados, B*50:01 muestra la asociación más fuerte a nivel estadístico, destacando que adquiere significación como factor de riesgo tras el ajuste por variables de confusión, igual que ocurre con B*15:03. Estos resultados concuerdan con análisis *in silico* previos que han demostrado que B*15:03 tiene una capacidad excepcional para unir péptidos virales con alta afinidad, siendo uno de los alelos HLA con mayor capacidad de presentación de péptidos del SARS-CoV-2 (Nguyen et al., 2020; Barquera et al., 2020).

En los genes HLA de clase II, DQB1*03:19 se clasifica estadísticamente como un factor de riesgo significativo tras el ajuste. Esta asociación no ha sido descrita en estudios similares como el de Migliorini et al. (2021). A diferencia del alelo DRB1*15:01, también asociado a severidad de COVID-19, pero que este sí es respaldado por estudios previos que asociaron este alelo con mayor susceptibilidad a la COVID-19, aunque en este caso se utilizó una cohorte italiana (Novelli et al., 2020).

Por otro lado, se han identificado dos alelos protectores significativos: DRB1*04:02 y C*05:01. La asociación protectora de DRB1*04:02 también se describe en el trabajo de Langton et al. (2021), quienes establecieron que este alelo se asociaba con síntomas menos severos en población británica. Por otro lado, el establecimiento del alelo C*05:01 como protector, se opone a estudios previos realizados con población italiana, donde se asoció con mayor riesgo (Correale et al., 2020).

DRB1*15:01 también se identifica como alelo de riesgo, del mismo modo que se detectó estudios en población italiana (Novelli et al., 2020). Sin embargo, hay diferencias en cuanto a la magnitud de la asociación, siendo más fuerte en nuestra cohorte latinoamericana.

Se identificaron dos alelos protectores significativos: DRB1*04:02 y C*05:01. La asociación protectora de DRB1*04:02 concuerda con estudios realizados con una cohorte europea británica (Langton et al., 2021), sugiriendo un efecto que prevalece en distintas poblaciones. Sin embargo, este efecto protector es más notable en el presente estudio, como ocurría con con el alelo DRB1*15:01.

Es importante destacar el caso del alelo C*05:01, puesto que para nuestra cohorte de pacientes latinoamericanos se clasifica estadísticamente como protector, pero en estudios con una cohorte europea italiana se define como un alelo asociado a riesgo de severidad de COVID-19 (Correale et al., 2020). Esta aparente discrepancia podría explicarse por las diferencias en el contexto genético entre poblaciones europeas y latinoamericanas, como han sugerido estudios recientes sobre la arquitectura genética en poblaciones mixtas (Shi et al., 2021). Las diferencias en los efectos causales de las variantes genéticas entre poblaciones se pueden atribuir a múltiples factores, como la estructura poblacional específica y las ambientales propias de cada población.

En definitiva, estos resultados muestran seis asociaciones significativas tras el ajuste por variables demográficas y componentes principales. Los alelos de riesgo B*50:01, B*15:03,

DQB1*03:19 y DRB1*15:01, junto con los alelos protectores DRB1*04:02 y C*05:01, configuran un perfil genético que difiere en ciertos aspectos de los estudiados en otras poblaciones. Estas diferencias, junto con los cambios observados tras el ajuste por variables de confusión, evidencian la importancia de considerar el contexto genético poblacional en estudios de asociación HLA.

6. Perspectivas de futuro

Este trabajo ha permitido establecer una base inicial en el análisis de la relación entre la variabilidad de alelos HLA y la gravedad de COVID-19 en una muestra representativa de pacientes. Sin embargo, existen múltiples direcciones en las que se puede ampliar y profundizar este análisis para obtener una comprensión más detallada y aplicada de los factores inmunogenéticos asociados a la severidad de la enfermedad.

- **Análisis específico de la hospitalización y asintomatología**: este estudio se centra en la severidad global de los casos, pero un análisis más detallado que separe los casos severos en función de la hospitalización y el estado asintomático podría mostrar cómo la variabilidad de HLA impacta en las diferentes manifestaciones clínicas de COVID-19. Este tipo de análisis permitiría una caracterización más precisa de los alelos de riesgo y protección, además de la relación existente entre todas estas variables clínicas.
- **Análisis según el país de procedencia y el porcentaje de ancestría**: sería interesante tener en consideración el país de procedencia de cada individuo y su porcentaje de ancestría (europea, africana y nativoamericana), ya que se ha visto que la variabilidad poblacional es un aspecto clave en la imputación de alelos HLA y su posterior análisis de asociación con variables clínicas.

7. Conclusiones

A continuación, se detallan las principales conclusiones derivadas de este trabajo:

- Para los alelos con más variabilidad, HLA-B y HLA-DRB1, el éxito en las predicciones tras la imputación es muy bajo.
- Los factores demográficos de edad y sexo demuestran ser determinantes críticos en la severidad de COVID-19.
- Se describe una asociación genética entre los alelos B*50:01, DQB1*03:19, B*15:03 y DRB1*15:01; y la severidad de la enfermedad en el grupo estudiado. Así como una asociación entre los alelos DRB1*04:02 y C*05:01, con una severidad menor.
- Las diferencias observadas en las asociaciones alélicas entre poblaciones latinoamericanas y europeas enfatizan la importancia del contexto genético poblacional en la respuesta a COVID-19 y la necesidad de estudios específicos por población.

8. Declaración de uso responsable de herramientas IA

Se declara que para la realización del presente trabajo se han usado diferentes herramientas que utilizan la inteligencia artificial para diversos fines. Como ya se ha descrito anteriormente, se ha utilizado el paquete HIBAG de *R*, que utiliza una forma de inteligencia artificial basada en modelos de *bagging*, específicamente Attribute Bagging. También se ha utilizado el motor de búsqueda bibliográfica basado en inteligencia artificial *Consensus* para optimizar la recopilación de literatura científica en relación al estudio. Se ha utilizado el asistente de conversación *Claude* para corrección de código en *R*, comprobación de coherencia en los resultados estadísticos y explicación de conceptos de estadística asociados a los resultados. Por último, se ha utilizado el asistente de conversación *ChatGPT* para sintetizar textos, acalaración de conceptos y búsqueda de palabras clave en texto.

A continuación, se citan las herramientas en formato APA, el uso de estas herramientas se realizó en distintas fechas desde el comienzo del estudio hasta el día 6 de noviembre de 2024.

Zhang, H., & Stram, D. O. (2016). *HIBAG: HLA Genotype Imputation with Attribute Bagging*. R package version 1.18.0. https://www.bioconductor.org/packages/HIBAG/

Consensus. (s.f.). *Consensus: An AI-powered research tool*. https://consensus.app/

Anthropic. (2023). *Claude: A conversational AI assistant*. https://www.anthropic.com/

OpenAI. (2023). *ChatGPT: An AI language model*. https://chat.openai.com/

9. Bibliografía

Almeida Silvia Diz-de, Cruz Raquel, Luchessi Andre D., Lorenzo-Salazar José M., de Heredia Miguel López, Quintela Inés, González-Montelongo Rafaela, Silbiger Vivian N., Porras Marta Sevilla, Castaño Jair Antonio Tenorio, Nevado Julian, Aguado Jose María, Aguilar Carlos, Aguilera-Albesa Sergio, Almadana Virginia, Almoguera Berta, Alvarez Nuria, Andreu-Bernabeu Álvaro, Arana-Arri Eunate, Arango Celso, Arranz María J., Artiga Maria-Jesus, Baptista-Rosas Raúl C., Sánchez María Barreda-, Belhassen-Garcia Moncef, Bezerra Joao F., Bezerra Marcos A.C., Boix-Palop Lucía, Brion María, Brugada Ramón, Bustos Matilde, Calderón Enrique J., Carbonell Cristina, Castano Luis, Castelao Jose E., Vicente Rosa Conde-, Cordero-Lorenzana M. Lourdes, Cortes-Sanchez Jose L., Corton Marta, Darnaude M. Teresa, Martino-Rodríguez Alba De, del Campo-Pérez Victor, de Bustamante Aranzazu Diaz, Domínguez-Garrido Elena, Eirós Rocío, Fariñas María Carmen, Fernandez-Nestosa María J., Fernández-Robelo Uxía, Fernández-Rodríguez Amanda, Fernández-Villa Tania, Gago-Domínguez Manuela, Gil-Fournier Belén, Arrue Javier Gómez-, Álvarez Beatriz González, de Quirós Fernan Gonzalez Bernaldo, González-Neira Anna, González-Peñas Javier, Gutiérrez-Bautista Juan F., Herrero María José, Herrero-Gonzalez Antonio, Jimenez-Sousa María A., Lattig María Claudia, Borja Anabel Liger, Lopez-Rodriguez Rosario, Mancebo Esther, López Caridad Martín-, Martín Vicente, Martinez-Nieto Oscar, Martinez-Lopez Iciar, Martinez-Resendez Michel F., Martinez-Perez Ángel, Mazzeu Juliana F., Macías Eleuterio Merayo, Minguez Pablo, Cuerda Victor Moreno, Oliveira Silviene F., Ortega-Paino Eva, Parellada Mara, Paz-Artal Estela, Santos Ney P.C., Matute Patricia Pérez-, Perez Patricia, Pérez-Tomás M. Elena, Perucho Teresa, Abuin Mel·lina Pinsach-, Pita Guillermo, Pompa-Mera Ericka N., Porras-Hurtado Gloria L., Pujol Aurora, León Soraya Ramiro, Resino Salvador, Fernandes Marianne R., Rodríguez-Ruiz Emilio, Rodriguez-Artalejo Fernando, Rodriguez-Garcia José A., Ruiz-Cabello Francisco, Ruiz-Hornillos Javier, Ryan Pablo, Soria José Manuel, Souto Juan Carlos, Tamayo Eduardo, Tamayo-Velasco Alvaro, Taracido-Fernandez Juan Carlos, Teper Alejandro, Torres-Tobar Lilian, Urioste Miguel, Valencia-Ramos Juan, Yáñez Zuleima, Zarate Ruth, de Rojas Itziar, Ruiz Agustín, Sánchez Pascual, Real Luis Miguel, SCOURGE Cohort Group , Guillen-Navarro Encarna, Ayuso Carmen, Parra Esteban, Riancho José A., Rojas-Martinez Augusto, Flores Carlos, Lapunzina Pablo, Carracedo Ángel (2024) Novel risk loci for COVID-19 hospitalization among admixed American populations eLife 13:RP93666 https://doi.org/10.7554/eLife.93666.1

Amoroso, A., Magistroni, P., Vespasiano, F., Bella, A., Bellino, S., Puoti, F., Alizzi, S., Vaisitti, T., Boros, S., Grossi, P. A., Trapani, S., Lombardini, L., Pezzotti, P., Deaglio, S., Brusaferro, S., Cardillo, M., & Italian Network of Regional Transplant Coordinating Centers (2021). HLA and AB0 Polymorphisms May Influence SARS-CoV-2 Infection and COVID-19 Severity. *Transplantation*, *105*(1), 193–200. https://doi.org/10.1097/TP.0000000000003507

Arrieta-Bolaños, E., Madrigal, J. A., & Shaw, B. E. (2012). Human leukocyte antigen profiles of latin american populations: differential admixture and its potential impact on hematopoietic stem cell transplantation. *Bone marrow research*, *2012*, 136087. https://doi.org/10.1155/2012/136087

Barquera, R., Collen, E., Di, D., Buhler, S., Teixeira, J., Llamas, B., Nunes, J. M., & Sanchez-Mazas, A. (2020). Binding affinities of 438 HLA proteins to complete proteomes of seven

pandemic viruses and distributions of strongest and weakest HLA peptide binders in populations worldwide. *HLA*, *96*(3), 277–298. https://doi.org/10.1111/tan.13956

Blackwell, J. M., Jamieson, S. E., & Burgner, D. (2009). HLA and infectious diseases. *Clinical microbiology reviews*, *22*(2), 370–385. https://doi.org/10.1128/CMR.00048-08

Castro-Santos, P., Rojas-Martinez, A., Riancho, J. A., Lapunzina, P., Flores, C., Carracedo, Á., Díaz-Peña, R., & Scourge Cohort Group (2023). HLA-A*11:01 and HLA-C*04:01 are associated with severe COVID-19. *HLA*, *102*(6), 731–739. https://doi.org/10.1111/tan.15160

Correale, P., Mutti, L., Pentimalli, F., Baglio, G., Saladino, R. E., Sileri, P., & Giordano, A. (2020). HLA-B*44 and C*01 Prevalence Correlates with Covid19 Spreading across Italy. *International journal of molecular sciences*, *21*(15), 5205. https://doi.org/10.3390/ijms21155205

Degenhardt, F., Wendorff, M., Wittig, M., Ellinghaus, E., Datta, L. W., Schembri, J., Ng, S. C., Rosati, E., Hübenthal, M., Ellinghaus, D., Jung, E. S., Lieb, W., Abedian, S., Malekzadeh, R., Cheon, J. H., Ellul, P., Sood, A., Midha, V., Thelma, B. K., Wong, S. H., … Franke, A. (2019). Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles. *Human molecular genetics*, *28*(12), 2078–2092. https://doi.org/10.1093/hmg/ddy443

Dilthey, A., Leslie, S., Moutsianas, L., Shen, J., Cox, C., Nelson, M. R., & McVean, G. (2013). Multi-population classical HLA type imputation. *PLoS computational biology*, *9*(2), e1002877. https://doi.org/10.1371/journal.pcbi.1002877

Hickey, M. J., Valenzuela, N. M., & Reed, E. F. (2016). Alloantibody Generation and Effector Function Following Sensitization to Human Leukocyte Antigen. Frontiers in immunology, 7, 30. https://doi.org/10.3389/fimmu.2016.00030

HLA Nomenclature Committee. (2024, octubre 2). *HLA allele frequency statistics*. HLA Allele Nomenclature. https://hla.alleles.org/nomenclature/stats.html

Hoseinnezhad, T., Soltani, N., Ziarati, S., Behboudi, E., & Mousavi, M. J. (2024). The role of HLA genetic variants in COVID-19 susceptibility, severity, and mortality: A global review. Journal of clinical laboratory analysis, 38(1-2), e25005. https://doi.org/10.1002/jcla.25005

Karnes, J. H., Shaffer, C. M., Bastarache, L., Gaudieri, S., Glazer, A. M., Steiner, H. E., Mosley, J. D., Mallal, S., Denny, J. C., Phillips, E. J., & Roden, D. M. (2017). Comparison of HLA allelic imputation programs. *PloS one*, *12*(2), e0172444. https://doi.org/10.1371/journal.pone.0172444

Klein, S. L., & Morgan, R. (2020). The impact of sex and gender on immunotherapy outcomes. *Biology of sex differences*, *11*(1), 24. https://doi.org/10.1186/s13293-020-00301-y

Langton, D. J., Bourke, S. C., Lie, B. A., Reiff, G., Natu, S., Darlay, R., Burn, J., & Echevarria, C. (2021). The influence of HLA genotype on the severity of COVID-19 infection. *HLA*, *98*(1), 14–22. https://doi.org/10.1111/tan.14284

Li, X., Xu, S., Yu, M., Wang, K., Tao, Y., Zhou, Y., Shi, J., Zhou, M., Wu, B., Yang, Z., Zhang, C., Yue, J., Zhang, Z., Renz, H., Liu, X., Xie, J., Xie, M., & Zhao, J. (2020). Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *The Journal of allergy and clinical immunology*, *146*(1), 110–118. https://doi.org/10.1016/j.jaci.2020.04.006

Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype imputation. *Annual review of genomics and human genetics*, *10*, 387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242

Madden, K., & Chabot-Richards, D. (2019). HLA testing in the molecular diagnostic laboratory. *Virchows Archiv : an international journal of pathology*, *474*(2), 139–147. https://doi.org/10.1007/s00428-018-2501-3

Marchal, A., Cirulli, E. T., Neveux, I., Bellos, E., Thwaites, R. S., Schiabor Barrett, K. M., Zhang, Y., Nemes-Bokun, I., Kalinova, M., Catchpole, A., Tangye, S. G., Spaan, A. N., Lack, J. B., Ghosn, J., Burdet, C., Gorochov, G., Tubach, F., Hausfater, P., COVID Human Genetic Effort, COVIDeF Study Group, … Bolze, A. (2024). Lack of association between classical HLA genes and asymptomatic SARS-CoV-2 infection. *HGG advances*, *5*(3), 100300. https://doi.org/10.1016/j.xhgg.2024.100300

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature reviews. Genetics*, *11*(7), 499–511. https://doi.org/10.1038/nrg2796

Migliorini, F., Torsiello, E., Spiezia, F., Oliva, F., Tingart, M., & Maffulli, N. (2021). Association between HLA genotypes and COVID-19 susceptibility, severity and progression: a comprehensive review of the literature. *European journal of medical research*, *26*(1), 84. https://doi.org/10.1186/s40001-021-00563-1

Mueller, A. L., McNamara, M. S., & Sinclair, D. A. (2020). Why does COVID-19 disproportionately affect older people?. *Aging*, *12*(10), 9959–9981. https://doi.org/10.18632/aging.103344

Naito, T., & Okada, Y. (2022). HLA imputation and its application to genetic and molecular fine-mapping of the MHC region in autoimmune diseases. Seminars in immunopathology, 44(1), 15–28. https://doi.org/10.1007/s00281-021-00901-9

Nguyen, A., David, J. K., Maden, S. K., Wood, M. A., Weeder, B. R., Nellore, A., & Thompson, R. F. (2020). Human Leukocyte Antigen Susceptibility Map for Severe Acute Respiratory Syndrome Coronavirus 2. *Journal of virology*, *94*(13), e00510-20. https://doi.org/10.1128/JVI.00510-20

Novelli, A., Andreani, M., Biancolella, M., Liberatoscioli, L., Passarelli, C., Colona, V. L., Rogliani, P., Leonardis, F., Campana, A., Carsetti, R., Andreoni, M., Bernardini, S., Novelli, G., & Locatelli, F. (2020). HLA allele frequencies and susceptibility to COVID-19 in a group of 99 Italian patients. *HLA*, *96*(5), 610–614. https://doi.org/10.1111/tan.14047

Okada, Y., Momozawa, Y., Ashikawa, K., Kanai, M., Matsuda, K., Kamatani, Y., Takahashi, A., & Kubo, M. (2015). Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. *Nature genetics*, *47*(7), 798–802. https://doi.org/10.1038/ng.3310

Robinson, J., Barker, D. J., Georgiou, X., Cooper, M. A., Flicek, P., & Marsh, S. G. E. (2020). IPD-IMGT/HLA Database. *Nucleic acids research*, *48*(D1), D948–D955. https://doi.org/10.1093/nar/gkz950

Scully, E. P., Haverfield, J., Ursin, R. L., Tannenbaum, C., & Klein, S. L. (2020). Considering how biological sex impacts immune responses and COVID-19 outcomes. *Nature reviews. Immunology*, *20*(7), 442–447. https://doi.org/10.1038/s41577-020-0348-8

Shi, H., Gazal, S., Kanai, M., Koch, E. M., Schoech, A. P., Siewert, K. M., Kim, S. S., Luo, Y., Amariuta, T., Huang, H., Okada, Y., Raychaudhuri, S., Sunyaev, S. R., & Price, A. L. (2021). Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nature communications*, *12*(1), 1098. https://doi.org/10.1038/s41467-021-21286-1

Takahashi, T., Ellingson, M. K., Wong, P., Israelow, B., Lucas, C., Klein, J., Silva, J., Mao, T., Oh, J. E., Tokuyama, M., Lu, P., Venkataraman, A., Park, A., Liu, F., Meir, A., Sun, J., Wang, E. Y., Casanovas-Massana, A., Wyllie, A. L., Vogels, C. B. F., … Iwasaki, A. (2020). Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature*, *588*(7837), 315–320. https://doi.org/10.1038/s41586-020-2700-3

Yan, C., Wang, R., Li, J., Deng, Y., Wu, D., Zhang, H., Zhang, H., Wang, L., Zhang, C., Sun, H., Zhang, X., Wang, J., Yang, H., & Li, S. (2003). HLA-A gene polymorphism defined by high-resolution sequence-based typing in 161 Northern Chinese Han people. *Genomics, proteomics & bioinformatics*, *1*(4), 304–309. https://doi.org/10.1016/s1672-0229(03)01036-2

Zhang, J. J., Dong, X., Cao, Y. Y., Yuan, Y. D., Yang, Y. B., Yan, Y. Q., Akdis, C. A., & Gao, Y. D. (2020). Clinical characteristics of 140 patients infected with SARS-CoV-2 in Wuhan, China. *Allergy*, *75*(7), 1730–1741. https://doi.org/10.1111/all.14238

Zheng, X., Shen, J., Cox, C., Wakefield, J. C., Ehm, M. G., Nelson, M. R., & Weir, B. S. (2014). HIBAG--HLA genotype imputation with attribute bagging. *The pharmacogenomics journal*, *14*(2), 192–200. https://doi.org/10.1038/tpj.2013.18

Zhou, F., Cao, H., Zuo, X., Zhang, T., Zhang, X., Liu, X., Xu, R., Chen, G., Zhang, Y., Zheng, X., Jin, X., Gao, J., Mei, J., Sheng, Y., Li, Q., Liang, B., Shen, J., Shen, C., Jiang, H., Zhu, C., … Zhang, X. (2016). Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. *Nature genetics*, *48*(7), 740–746. https://doi.org/10.1038/ng.3576

10. Anexos

Anexo I

UNIVERSIDAD AUTONOMA DE CHILE

Comité de Ética

N°CEC DI-02-24

ACTA DE DISPENSA
PROYECTO DE INVESTIGACIÓN

En Temuco, a 06 de mayo de 2024, el Comité Ético Científico de la Universidad Autónoma de Chile (CEC-UA) en sesión expedita, presidida por la Dra. Marlen Barreto, informan han sesionado la petición de "Dispensa de uso de consentimiento informado" para el Proyecto de Investigación titulado: "Caracterización inmunogenética de la COVID-19" cuyo investigador responsable es el Dr. Roberto Diaz Peña. El objetivo de este estudio es analizar la presencia e influencia de los alelos HLA clase I y II en pacientes con la COVID-19, mediante su imputación a partir de polimorfismos de nucleótido único (SNP, single nucleotide polymorphims), de datos de genotipado, y su relación con el desarrollo de enfermedad grave y/o enfermedad asintomática (datos fenotípicos). Estos datos se requerirán durante un período de dos años (2024-2026).
El CEC-UA ha evaluado y sancionado la pertinencia de dispensar el uso del consentimiento informado para el uso de datos con fines de investigación almacenados en la Red de Biobancos y Biorepositorios de C19-GenoNet (RedBB C19-GenoNet).

Esta sanción se rige bajo las siguientes consideraciones éticas:

1. El Protocolo de Investigación se enmarca en los principios de respeto a los Derechos Humanos no transgrediendo la confiabilidad de los pacientes que autorizaron el almacenamiento de sus datos y muestras.
2. El uso de datos obtenidos a partir de Red de Biobancos y Biorepositorios de C19-GenoNet (RedBB C19-GenoNet) no vulnera la dignidad de los sujetos que pudiesen verse involucrados, asegura el derecho a la privacidad y anonimato, garantiza la protección de la confidencialidad de los datos y define la custodia de la información y el uso que se le dará.
3. No habrá un nuevo acceso a los pacientes para solicitar el consentimiento informado para el uso de la información.
4. No se observan conflictos de interés.

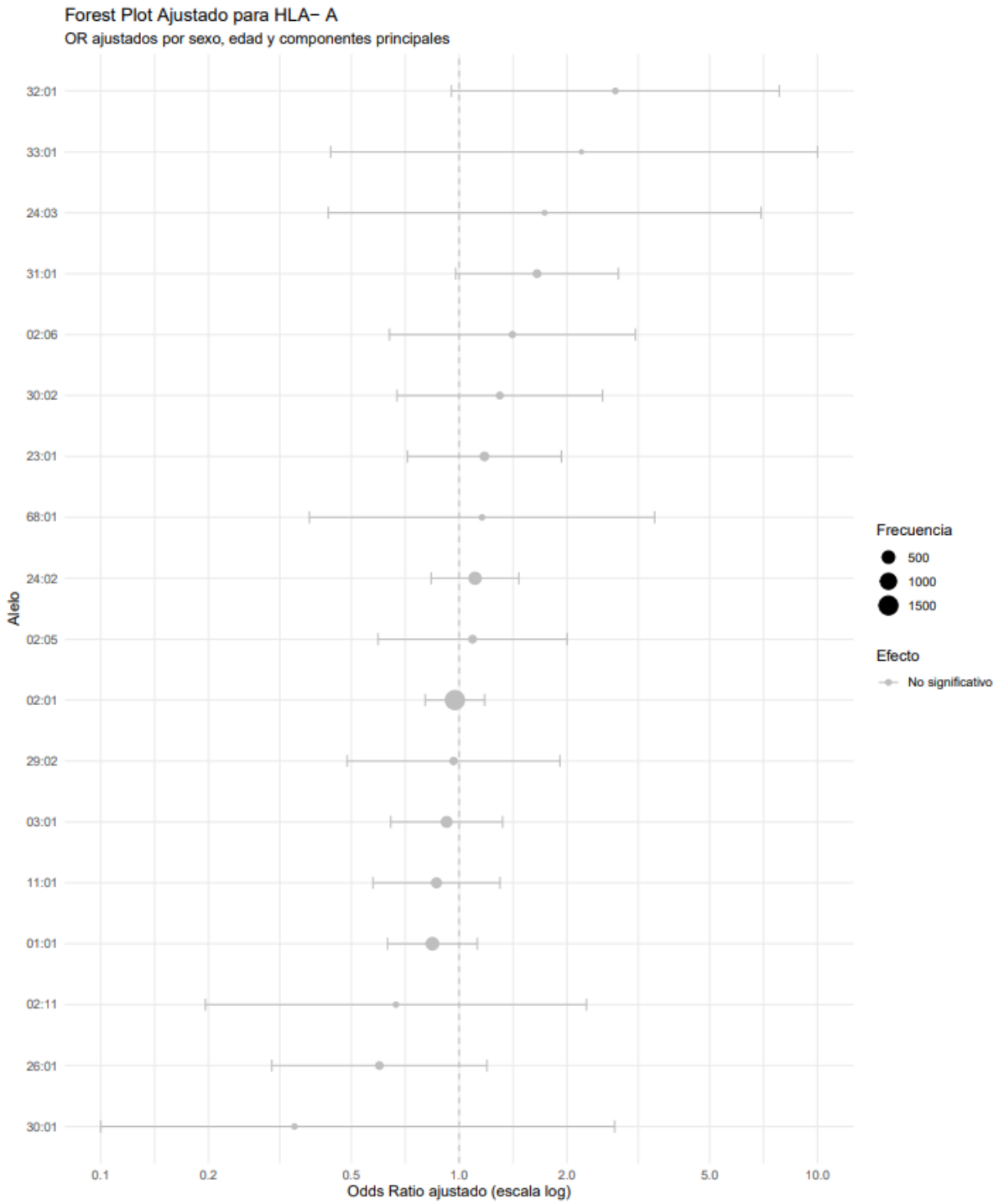
Para constancia firma:

UNIVERSIDAD AUTONOMA DE CHILE
Comité de Etica

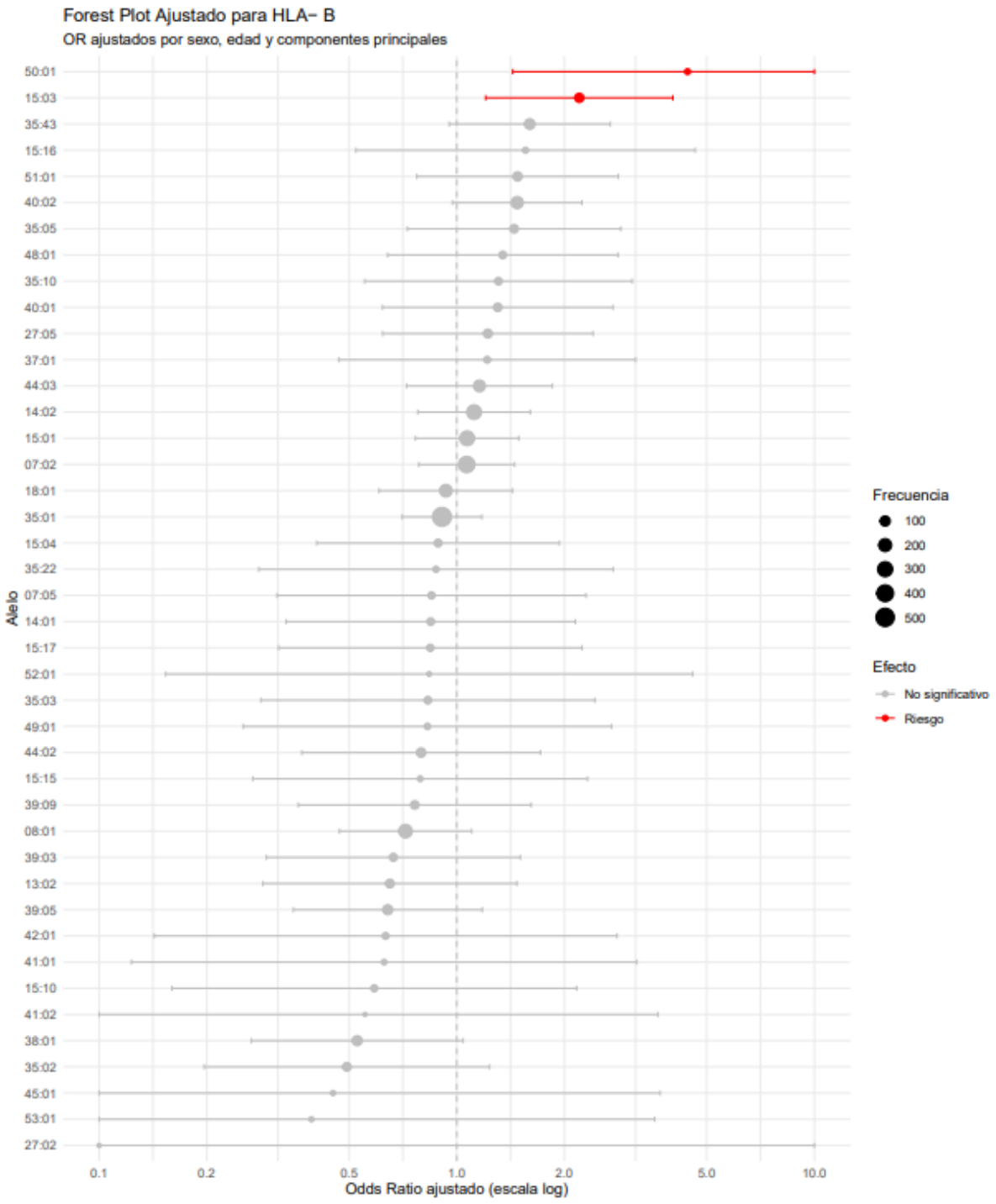
Dra. Marlén Barreto
Presidenta Comité Ético Científico Institucional

El Comité Ético Científico de la Universidad Autónoma de Chile está conformado por 11 miembros, representados en sus tres sedes

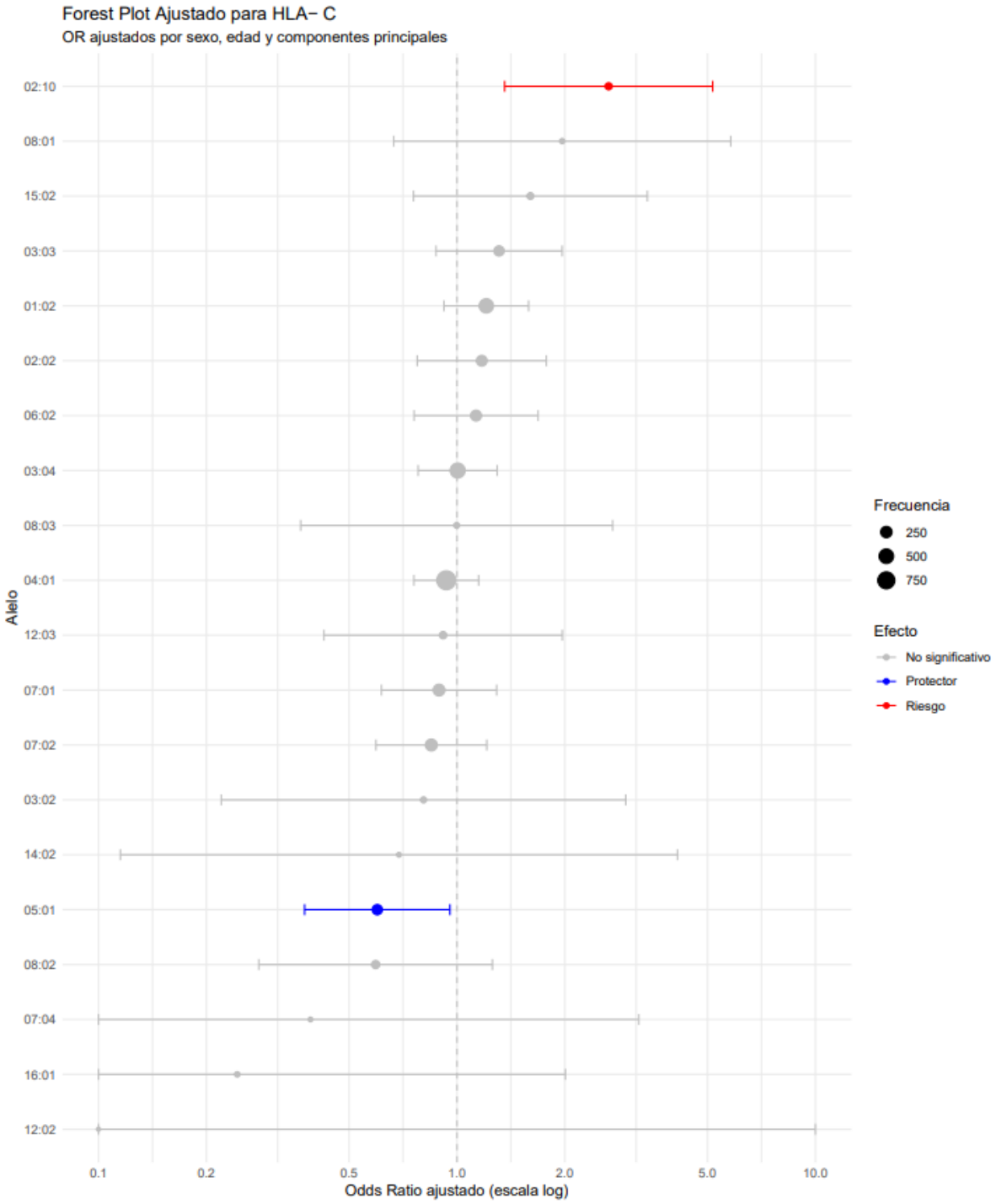
Anexo II


Forest Plot Ajustado para HLA− A
OR ajustados por sexo, edad y componentes principales
Alelo
32:01
33:01
24:03
31:01
02:06
30:02
23:01
68:01
24:02
02:05
02:01
29:02
03:01
11:01
01:01
02:11
26:01
30:01
0.1
0.2
0.5
1.0
2.0
5.0
10.0
Odds Ratio ajustado (escala log)
Frecuencia
500
1000
1500
Efecto
No significativo

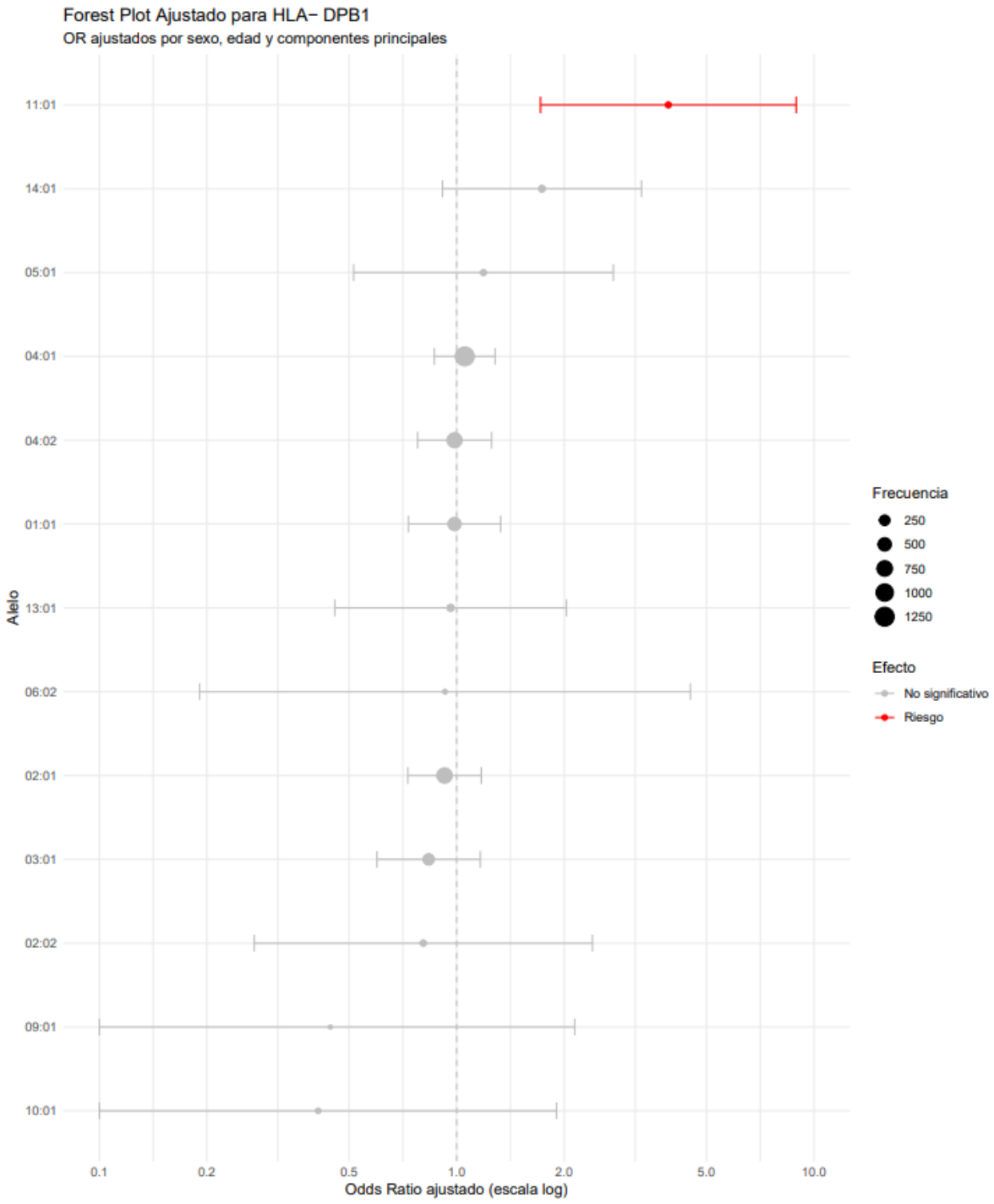
Anexo III


Forest Plot Ajustado para HLA− B
OR ajustados por sexo, edad y componentes principales
Alelo
50:01
15:03
35:43
15:16
51:01
40:02
35:05
48:01
35:10
40:01
27:05
37:01
44:03
14:02
15:01
07:02
18:01
35:01
15:04
35:22
07:05
14:01
15:17
52:01
35:03
49:01
44:02
15:15
39:09
08:01
39:03
13:02
39:05
42:01
41:01
15:10
41:02
38:01
35:02
45:01
53:01
27:02
0.1
0.2
0.5
1.0
2.0
5.0
10.0
Odds Ratio ajustado (escala log)
Frecuencia
100
200
300
400
500
Efecto
No significativo
Riesgo

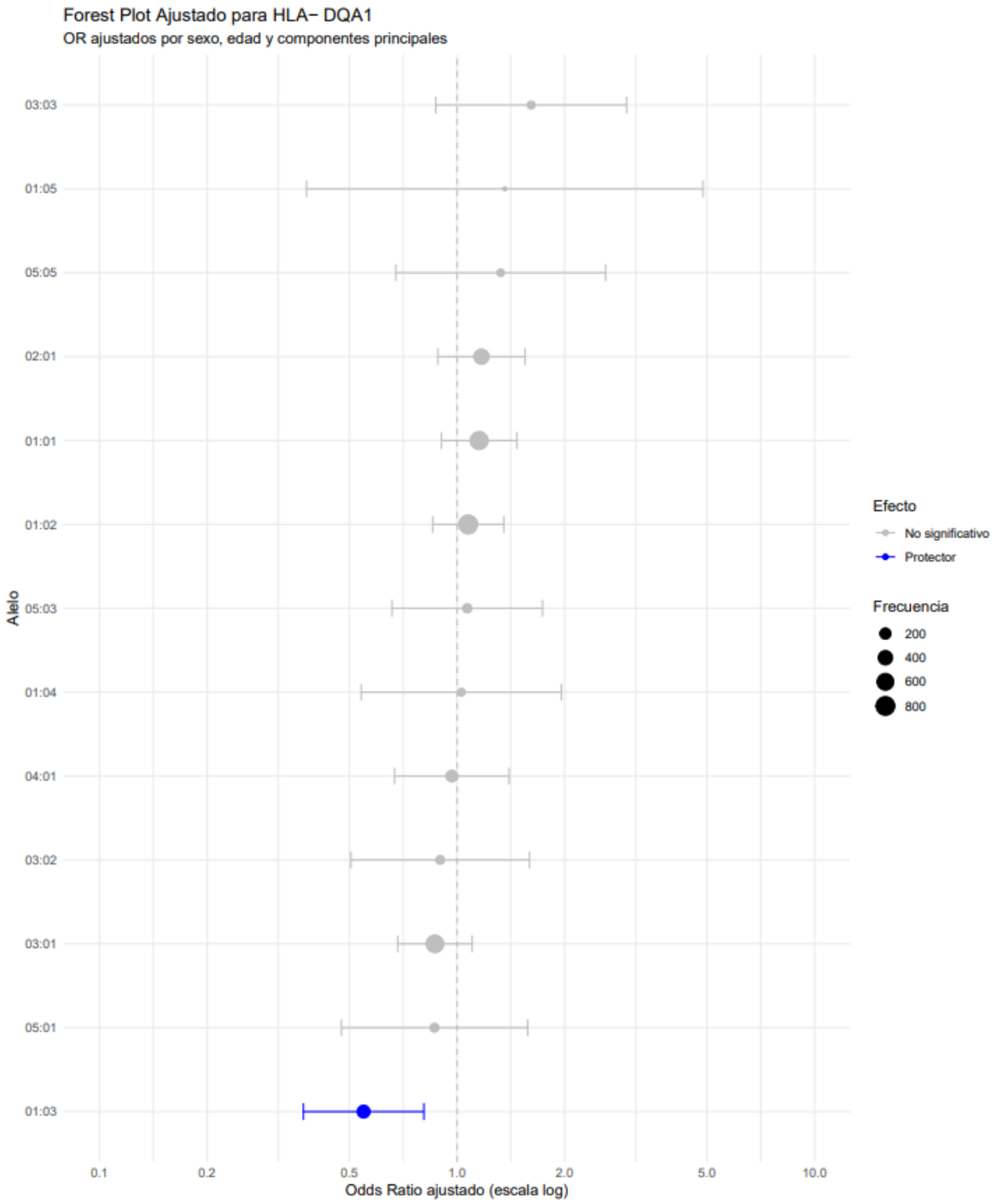
Anexo IV


Forest Plot Ajustado para HLA− C
OR ajustados por sexo, edad y componentes principales
Alelo
02:10
08:01
15:02
03:03
01:02
02:02
06:02
03:04
08:03
04:01
12:03
07:01
07:02
03:02
14:02
05:01
08:02
07:04
16:01
12:02
0.1
0.2
0.5
1.0
2.0
5.0
10.0
Odds Ratio ajustado (escala log)
Frecuencia
250
500
750
Efecto
No significativo
Protector
Riesgo

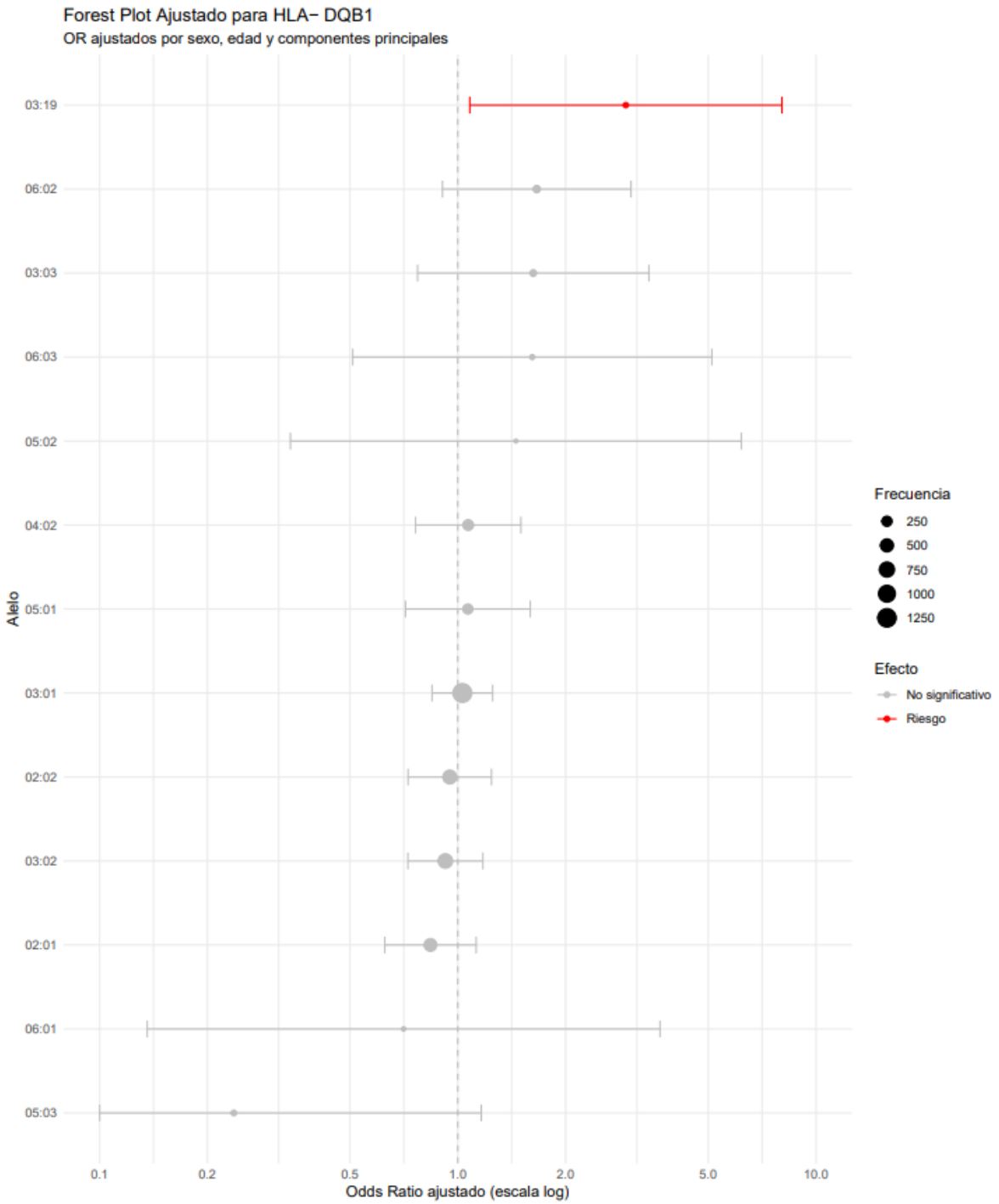
Anexo V


Forest Plot Ajustado para HLA− DPB1
OR ajustados por sexo, edad y componentes principales
11:01
14:01
05:01
04:01
04:02
01:01
13:01
06:02
02:01
03:01
02:02
09:01
10:01
Alelo
0.1
0.2
0.5
1.0
2.0
5.0
10.0
Odds Ratio ajustado (escala log)
Frecuencia
250
500
750
1000
1250
Efecto
No significativo
Riesgo

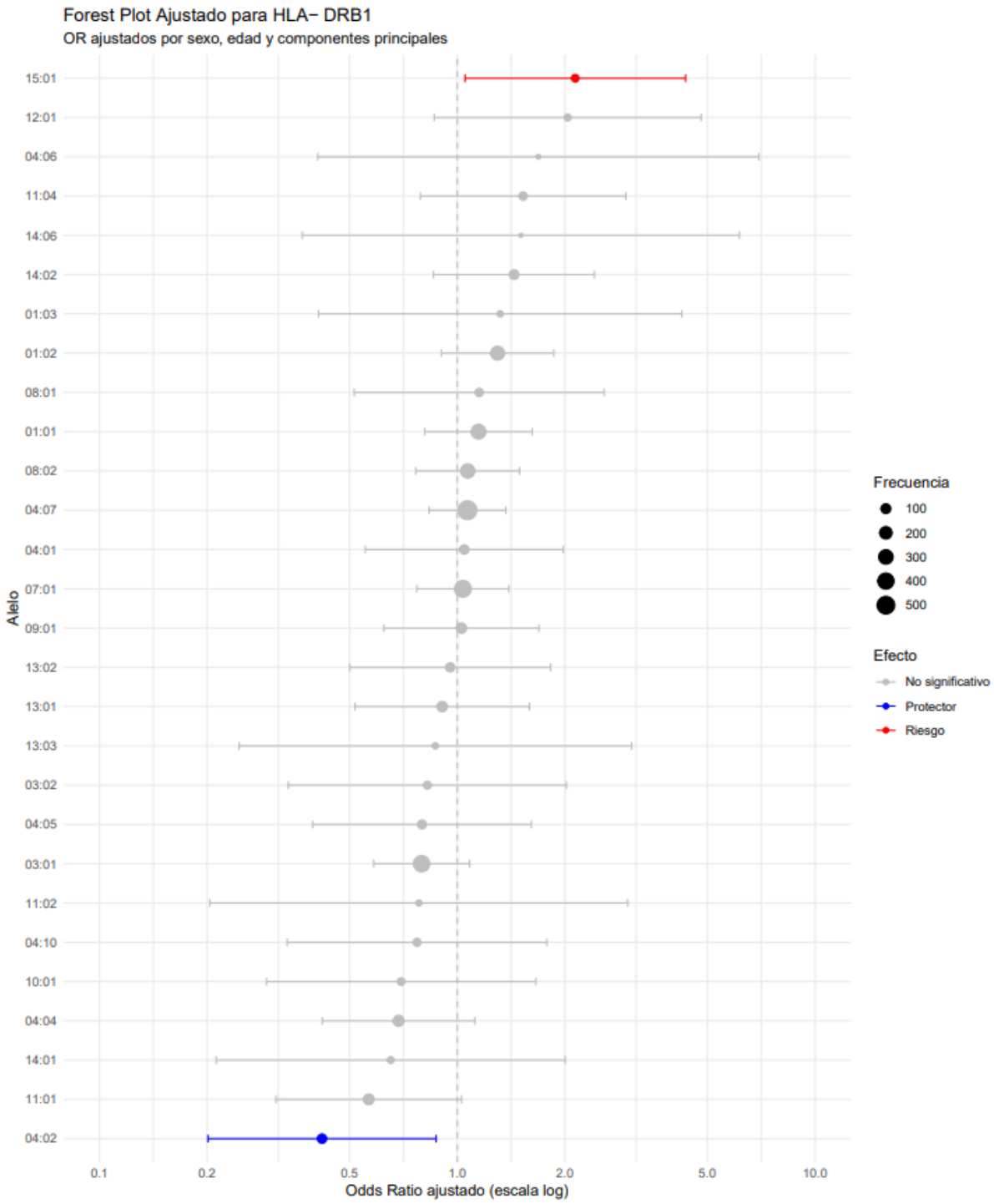
Anexo VI


Forest Plot Ajustado para HLA− DQA1
OR ajustados por sexo, edad y componentes principales
Alelo
03:03
01:05
05:05
02:01
01:01
01:02
05:03
01:04
04:01
03:02
03:01
05:01
01:03
0.1
0.2
0.5
1.0
2.0
5.0
10.0
Odds Ratio ajustado (escala log)
Efecto
No significativo
Protector
Frecuencia
200
400
600
800

Anexo VII


Forest Plot Ajustado para HLA− DQB1
OR ajustados por sexo, edad y componentes principales
Alelo
03:19
06:02
03:03
06:03
05:02
04:02
05:01
03:01
02:02
03:02
02:01
06:01
05:03
0.1
0.2
0.5
1.0
2.0
5.0
10.0
Odds Ratio ajustado (escala log)
Frecuencia
250
500
750
1000
1250
Efecto
No significativo
Riesgo

Anexo VIII


Forest Plot Ajustado para HLA− DRB1
OR ajustados por sexo, edad y componentes principales
15:01
12:01
04:06
11:04
14:06
14:02
01:03
01:02
08:01
01:01
08:02
04:07
04:01
07:01
09:01
13:02
13:01
13:03
03:02
04:05
03:01
11:02
04:10
10:01
04:04
14:01
11:01
04:02
Alelo
0.1
0.2
0.5
1.0
2.0
5.0
10.0
Odds Ratio ajustado (escala log)
Frecuencia
100
200
300
400
500
Efecto
No significativo
Protector
Riesgo