ELSEVIER

*Rev. Esp. Antrop. Fís.* (2023)
Vol. 48: 40-50

# Sistema de predicción de riesgo a COVID-19 grave basado en Polygenic Risk Score

Rafael Farias, Neskuts Izagirre y Santos Alonso

Departamento de Genética, Antropología Física y Fisiología Animal. Facultad de Ciencia y Tecnología, UPV/EHU. Barrio Sarriena s/n 48940 Leioa Bizkaia

*Corresponding Author:* rfarias001@ikasle.ehu.eus

*Palabras claves:*

Polygenic Risk Score (PRS)
COVID-19 grave
PLINK
PRSice-2
Poblaciones europeas
Componente genético neandertal

Recibido: 24-01-2024
Aceptado: 20-02-2024

RESUMEN

El método de Puntuación de Riesgo Poligénico (PRS) evalúa simultáneamente los efectos de múltiples SNPs asociados a una enfermedad de base genética compleja para obtener una puntuación individual que predice el riesgo de padecerla. Así, el PRS permite estratificar los individuos según su riesgo genético a padecer una enfermedad, lo cual es de utilidad clínica. En los últimos años, la pandemia de COVID-19 ha provocado millones de muertes globalmente. Las manifestaciones clínicas de la enfermedad son muy variables y se ha demostrado la relevancia de ciertos factores ambientales en la gravedad de la enfermedad. Sin embargo, los factores de riesgo genéticos también pueden ser importantes, incluido el componente genético neandertal global de cada individuo. En este contexto, el objetivo de este trabajo es utilizar el método PRS para estimar el riesgo poligénico de desarrollar COVID-19 grave en diferentes poblaciones. Para ello, se obtuvieron summary statistics de SNPs significativamente asociados a COVID-19 grave y después, se adquirieron datos genotípicos de 5 poblaciones europeas del Proyecto 1000 Genomas (P1000G) y de la población española. Además, para esta última se obtuvieron datos fenotípicos y covariantes del consorcio SCOURGE. Mediante los programas PLINK y PRSice-2, se calcularon PRS para las distintas poblaciones. Los resultados revelaron que la población finlandesa del P1000G mostraban un riesgo promedio mayor de COVID-19 grave y que el sexo no era una variable relevante en el riesgo genético. Por otro lado, los PRS calculados se correlacionaron débilmente con el componente genético neandertal global.

*Keywords:*

Polygenic Risk Score (PRS)
Severe COVID-19
PLINK
PRSice-2
European populations
Neanderthal genetic component

ABSTRACT

The Polygenic Risk Score (PRS) method evaluates the effects of multiple SNPs associated with a complex genetically based disease to obtain an individual score that predicts the risk of having the disease. Thus, PRS allows stratification of individuals according to their genetic risk for a disease, which is clinically useful. In the last three years, the COVID-19 pandemic has caused millions of deaths globally. The clinical manifestations of the disease are highly variable and the relevance of certain environmental factors in the severity of the disease has been demonstrated. However, genetic risk factors may also be important, including the overall Neanderthal genetic component of each individual. In this context, the aim of this work is to use the PRS method to estimate the polygenic risk of developing severe COVID-19 in different populations. For this purpose, summary statistics of SNPs significantly associated with severe COVID-19 were obtained and then, genotypic data were acquired from 5 European populations of the 1000 Genomes Project and from the Spanish population. In addition, for the latter, phenotypic and covariate data were obtained from the SCOURGE consortium. Using PLINK and PRSice-2, PRS were calculated for the different populations. The results revealed that the Finnish population of the 1000 Genomes Project showed a higher average risk of severe COVID-19 and that sex was not a relevant variable in genetic risk. On the other hand, the PRSs were correlated weakly with the overall Neanderthal genetic component.

ISSN: 2253-9921

## Introducción

Hasta la fecha, se han publicado en el *NHGRI-EBI GWAS Catalog* 6.401 estudios, que en conjunto proporcionan 529.481 asociaciones (con una significación estadística a escala genómica menor de 5*10-8) y 60.071 *summary statistics* completos que han conseguido revelar polimorfismos de un solo nucleótido (SNPs) asociados a una amplia gama de rasgos y enfermedades complejas (Sollis et al., 2023). No obstante, dado que la mayoría de las enfermedades complejas son altamente poligénicas, cada uno de los SNPs identificados en un análisis de asociación de genoma completo (GWAS, *genome-wide association study*) suele tener un efecto pequeño sobre el fenotipo de interés (Dudbridge, 2016). Por esta razón, los GWAS por sí solos tienen un poder limitado para predecir el fenotipo (el riesgo individual en el caso de enfermedad) (Uricchio, 2020). Para solucionar este problema, se ha desarrollado el método de Puntuación de Riesgo Poligénico (PRS, *Polygenic Risk Score*), el cual estima el riesgo genético de un individuo para una enfermedad concreta. Así, el PRS puede utilizarse para la predicción clínica o screening (Marees et al., 2018). Por lo general, se requieren dos tipos de datos: datos base (*base data*) y datos diana (*target data*).

Por un lado, el base data consiste en un conjunto de GWAS que contiene información, denominada *summary statistics*, de SNPs asociados a un fenotipo, como cuál es el alelo de riesgo del SNP, su *effect size*, la frecuencia del alelo menor (MAF, *minor allele frequency*) y el p-valor de asociación, entre otros (Choi, Mak y O'Reilly, 2020). Por otro lado, el target data consiste en genotipos, y normalmente también fenotipos (presencia de la enfermedad, nivel de gravedad…), de individuos de una muestra poblacional, que, idealmente, debe ser independiente de la muestra de los GWAS (Choi, Mak y O'Reilly, 2020). El *target data* suele presentarse en formato *PLINK*, una herramienta muy popular en genética de poblaciones que, entre sus muchas opciones, permite computar PRS (Chang et al., 2015). Es en *target data* donde se calcula el PRS para cada individuo y se hace mediante la suma de los alelos de riesgo que porta cada individuo, ponderada por los *effect size* de los alelos de riesgo estimados en *base data* (Choi, Mak y O'Reilly, 2020).

En los últimos tres años, la urgencia por combatir la pandemia de coronavirus ha suscitado numerosos estudios de GWAS y PRS (Abdellaoui et al., 2023). La enfermedad por coronavirus de 2019 (COVID-19), causada por el coronavirus del síndrome respiratorio agudo grave de tipo 2 (SARS-CoV-2), apareció por primera vez a finales de 2019 en Wuhan (China) y desde entonces, evolucionó rápidamente hasta convertirse en una pandemia mundial que ha supuesto una enorme presión sanitaria y económica (Tang, Comish y Kang, 2020). La Organización Mundial de la Salud (*World Health Organization*), a día 3 de mayo de 2023, ha publicado un informe que documenta alrededor de 764 millones de casos confirmados y 6.9 millones de muertes a nivel global.

Las manifestaciones clínicas de la COVID-19 son muy variables y oscilan entre una ausencia total de síntomas, padecer fiebre, tos, disnea o mialgia, o llegar a desarrollar insuficiencia respiratoria que, finalmente, cause la muerte (Tang, Comish y Kang, 2020). La gravedad de la enfermedad se ha asociado a factores de riesgo como la edad avanzada, el sexo masculino y la presencia de comorbilidades, siendo la hipertensión, la obesidad y la diabetes las más comunes (Vahidy et al., 2021). Sin embargo, la constitución genética del huésped también puede ser un elemento importante (COVID-19 Host Genetics Initiative, 2021). Por ello, en los últimos años, se ha propuesto utilizar el método PRS para mejorar la identificación de los individuos que tienen un mayor riesgo de enfermedad grave y así, darles prioridad para la vacunación contra el COVID-19 (Horowitz et al., 2022). Por otra parte, Zeberg y Pääbo (2020) identificaron en el cromosoma 3 un haplotipo neandertal introgresado en los humanos en el que se encuentran variantes genéticas asociadas a la gravedad de la COVID-19. Impulsados por este estudio, se decidió investigar si el componente genético neandertal en su conjunto tenía un efecto en los PRS de la población española.

La hipótesis de este estudio es que una parte del riesgo a desarrollar COVID-19 grave es de base genética y, por ello, es posible que diferentes poblaciones presenten distintos perfiles de riesgo genético. Por lo tanto, el objetivo general de este trabajo es estimar el riesgo poligénico de desarrollar COVID-19 grave en diferentes poblaciones. Los objetivos específicos son los siguientes: 1) Identificar GWAS

relacionados con la gravedad de COVID-19 y recopilar sus *summary statistics*; 2) Calcular PRS para individuos de poblaciones europeas y de la población española mediante *PLINK* y *PRSice*-2; 3) Analizar si las diferentes poblaciones europeas difieren en su riesgo promedio a desarrollar un COVID-19 grave; 4) Identificar si el sexo es una variable de relevancia en el riesgo genético; 5) Analizar la relevancia del componente genético neandertal en el riesgo genético a padecer COVID-19 grave en la población española.

## Materiales y métodos

### *Obtención de "base data"*

Para conseguir los *summary statistics* a partir de los cuales seleccionar los SNPs más significativamente asociados a COVID-19 grave, se utilizó principalmente la versión más reciente de los meta-análisis de casos y controles llevados a cabo por *COVID-19 Host Genetics Initiative* (COVID-19 HGI, 2022). Para la gravedad de COVID-19, se utilizaron dos GWAS con participantes de ascendencia europea con una media de edad de 55 años (COVID-19 HGI, 2022). El primero, denominado estudio A2, comparaba pacientes confirmados como "very severe respiratory COVID-19" (N = 13.769) con la población general (N = 1.072.442). COVID-19 HGI (2022) define el fenotipo "very severe respiratory COVID-19" como aquellos individuos que necesitaron asistencia respiratoria o fallecieron a causa de la enfermedad. El otro, denominado estudio B2, comparaba pacientes hospitalizados debido a síntomas de la COVID-19 (N = 32.519) con la población general (N = 2.062.805).

Adicionalmente, se utilizaron las bases de datos *PubMed* y *Scopus*, y se buscaron GWAS relacionados con la gravedad de COVID-19. En la búsqueda bibliográfica se seleccionaron publicaciones a partir de 2020 y se utilizaron las palabras clave "COVID-19 severity", "COVID-19 hospitalization", "GWAS" y "SNPs". Además, se tuvo en cuenta que las poblaciones de estudio fueran de ascendencia europea, que el tamaño de la muestra fuera elevado y que los valores de significación estadística fueran significativos a escala genómica (p-valor < $5*10^{-8}$). De esta manera, se seleccionaron los siguientes trabajos: Pairo-Castineira et al. (2021), Cruz et al. (2022), Degenhardt et al. (2022), Kousathanas et al. (2022) y Thibord et al. (2022).

Con toda la información reunida, se elaboró un listado de SNPs bialélicos y autosómicos. Respecto a los SNPs de los cromosomas sexuales, estos no se pueden integrar fácilmente en el protocolo de análisis de los SNPs autosómicos, por lo que se decidió descartarlos para homogeneizar el cálculo de los PRS. Para evitar la repetición de información, se tuvieron en cuenta los SNPs en desequilibrio de ligamiento (LD, *linkage disequilibrium*), cuyo grado se describe mediante el estadístico $r^2$ (Osterman, Kinzy y Cooke Bailey, 2021).

Así, se llevó a cabo un filtrado de SNPs mediante el método *clumping*, que consiste en establecer un umbral de p-valor y $r^2$ para obtener una selección de SNPs que no estén correlacionados entre sí (Osterman, Kinzy y Cooke Bailey, 2021). Para ello, se utilizó la aplicación web *LDlink* (https://ldlink.nih.gov/?tab=snpclip) (Machiela y Chanock, 2015). Así, se seleccionaron aquellos SNPs que tuvieran un p-valor menor de $10^{-5}$ y una MAF igual o superior a 0.01. A continuación, los SNPs se agruparon por número de cromosoma y se aplicó un umbral de $r^2$ de 0.2. Entre los que estaban en LD, se seleccionó la variante más significativa (la de menor p-valor), denominada "variante índice", y se descartaron el resto de variantes con un valor de $r^2$ superior a 0.2 con la "variante índice".

### *Obtención de "target data"*

En primer lugar, se descargaron genomas completos en formato *vcf* de la fase 3 del Proyecto 1000 Genomas (The 1000 Genomes Project Consortium, 2015) correspondientes a las poblaciones europeas (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/); después, se convirtieron a formato *PLINK* (archivos *.bed*, *.bim* y *.fam*). Las poblaciones europeas seleccionadas fueron las siguientes: residentes de Utah con ascendencia del norte y oeste de Europa (CEU) (N = 99), finlandeses de Finlandia (FIN) (N = 99), británicos de Inglaterra y Escocia (GBR) (N = 91), poblaciones ibéricas de España (IBS) (N = 107) y toscanos de Italia (TSI) (N = 107). En total, se

obtuvieron datos genotípicos de 503 individuos (263 mujeres y 240 hombres).

Previamente al cálculo de los PRS, se realizó un control de calidad del *target data* mediante *PLINK* (versión 1.9) y *RStudio* (versión 2023.03.1) siguiendo las recomendaciones de Choi, Mak y O'Reilly (2020). Así, se descartaron los SNPs: a) no genotipados en la mayoría de los individuos (*missing SNPs*), b) con frecuencia del alelo menor (MAF) inferior al 1%, c) que mostraban desviaciones del equilibrio de Hardy-Weinberg (HWE) y d) con alelos problemáticos. También se descartaron aquellos individuos que: e) carecían de genotipos para la mayoría de los SNPs (*missing individuals*), f) presentaban índices de heterocigosidad extremos y g) mostraban discrepancias entre el sexo asignado y el genético. Además, al igual que en *base data*, se utilizó el método *clumping* para seleccionar solo SNPs independientes. En este caso, se utilizó la opción de *PLINK* "--indep-pairwise" y se estableció un umbral de $r^2 = 0.2$.

Por otro lado, al no disponer de información fenotípica de las poblaciones europeas del Proyecto 1000 Genomas sobre el COVID-19, se utilizaron datos genotípicos y fenotípicos de individuos del proyecto *Spanish Coalition to Unlock Research on Host Genetics on COVID-19* (SCOURGE) (Cruz et al., 2022). Estos datos se obtuvieron tras realizar una solicitud y recibir una aprobación por parte de los miembros del consejo del consorcio SCOURGE para su uso en este estudio. En el proyecto SCOURGE, se analizaron 11.939 pacientes de la población española con COVID-19 y todos los casos diagnosticados se clasificaron en una escala de gravedad de cinco niveles: 0 para los asintomáticos, 1 para los casos leves, 2 para los casos moderados, 3 para un COVID-19 grave que requería hospitalización y 4 para un COVID-19 crítico que requería admisión en la UCI (Cruz et al., 2022). Para este trabajo, se seleccionaron aquellos individuos que presentaban la mínima (nivel 0) y la máxima gravedad (nivel 4) y se utilizaron como *target data*. En total, se reunieron 1.711 individuos que, además de información genotípica y fenotípica sobre la gravedad de la COVID-19, presentaban información de covariantes, como la edad, el sexo y las coordenadas de los 10 primeros ejes de un análisis de componentes principales orientado a obtener una clasificación de los individuos en base a su constitución genética global (ancestría genética).

### *Cálculo de PRS (Polygenic Risk Score – Puntuación de Riesgo Poligénico)*

A la hora de calcular PRS, solo se han tenido en cuenta los SNPs que estaban incluidos tanto en *base data* como en *target data* (Osterman, Kinzy y Cooke Bailey, 2021). Una vez obtenidos *summary statistics* filtrados y *target data* de calidad, se utilizaron dos programas independientes para el cálculo de PRS: *PLINK* (versión 1.9) y *PRSice-2* (versión 2.3.5). Por un lado, en *PLINK* se utilizó la opción "--score" para obtener las puntuaciones de riesgo de cada individuo de las poblaciones europeas y de la población española. Por otro lado, *PRSice-2* es un *software* dedicado exclusivamente al cálculo de PRS (Choi, Mak y O'Reilly, 2020). Este programa tiene dos modos para computar los PRS: "regress" y "no-regress". La opción "regress" exige incluir información adicional de los individuos del *target data*, como fenotipos y covariantes. En este modo, *PRSice-2* hace un análisis de regresión entre los datos fenotípicos y las puntuaciones de riesgo calculadas. Este programa calcula los PRS mediante una serie de pasos. En primer lugar, agrupa los SNPs en distintos grupos según diferentes umbrales de significación estadística (p-valor) y, después, calcula los PRS con los SNPs de cada grupo. Por último, identifica aquel umbral de p-valor que contiene el grupo de SNPs que explican la proporción más alta de la varianza fenotípica (expresada mediante $R^2$) y, por tanto, permiten calcular los mejores PRS. Este modo fue el que se utilizó para calcular los PRS en la población española, ya que incluía datos fenotípicos (gravedad de COVID-19) y covariantes (edad, sexo y coordenadas de 10 ejes de componentes principales). Con la opción "no-regress", *PRSice-2*, al igual que *PLINK*, se calculan los PRS directamente sin tener en cuenta información fenotípica. Este modo fue el que se utilizó para las poblaciones europeas debido a que solo se tenía disponible información genotípica y del sexo.

### *Análisis estadísticos básicos*

Los PRS calculados por *PLINK* y *PRSice-2* en poblaciones europeas se analizaron estadísticamente mediante *RStudio* (versión 2023.03.1). En primer lugar, se compararon las medias de los PRS entre hombres y mujeres de cada población mediante la prueba T de

Student. Para ello, previamente se realizó un test de Shapiro-Wilk para comprobar, en cada población, la normalidad de los PRS de hombres y mujeres. Después, se realizó un F-test para comprobar la homogeneidad de varianzas (homocedasticidad).

Además, se realizó un análisis de la varianza (ANOVA) de un factor para comparar las medias poblacionales de los PRS. Al igual que en el caso anterior, se efectuó previamente un test de Shapiro-Wilk y un test de Levene para comprobar la normalidad y la homocedasticidad de los datos. Junto con el ANOVA, en caso de rechazar la hipótesis nula (no diferencias significativas entre las poblaciones), se llevó a cabo un test de Tukey que comparó todos los pares de medias posibles para identificar qué medias eran significativamente diferentes entre sí. Por último, se hizo un ANOVA de dos factores para estudiar la significación de la interacción entre la población y el sexo, y analizar la influencia de la combinación de estas variables en los PRS.

Por otra parte, tanto en las poblaciones europeas como en la población española, se realizó un análisis de correlación entre las puntuaciones de riesgo calculadas por *PLINK* y las puntuaciones calculadas por *PRSice-2*. Asimismo, se llevó a cabo un análisis de correlación entre los PRS de la población IBS del Proyecto 1000 Genomas y los de la población española del proyecto SCOURGE. Para estos análisis de correlación se utilizaron valores normalizados de PRS y para todas las pruebas estadísticas se determinó un nivel de significación (α) de 0.05.

### *Correlación de los PRS con el componente genético neandertal*

En primer lugar, se obtuvieron 50.557 SNPs cuyo origen se debe a una introgresión neandertal (Gunz et al., 2019). A continuación, se identificó cuáles de estos SNPs estaban genotipados o imputados en los datos genotípicos de la población española de SCOURGE y que tuvieran como alelo alternativo el alelo neandertal. Para cada SNP, a cada individuo se le asignó un valor 0 si no tenía ningún alelo neandertal, 1 si era heterocigoto para el alelo neandertal y 2 si era homocigoto para dicho alelo. Finalmente, se sumaron los valores obtenidos en cada individuo y se normalizaron para obtener la proporción del componente genético neandertal de cada individuo, que se correlacionó con valores normalizados de PRS calculados por *PLINK* y *PRSice-2* en la población española.

## Resultados

### *Control de calidad en "base data" y "target data", y cálculo de PRS*

Respecto al *base data*, tras aplicar los filtros de selección mencionados, se obtuvieron *summary statistics* para 142 SNPs. Por otro lado, el control de calidad en las poblaciones europeas del *target data* eliminó 10 individuos y mantuvo 493 (253 mujeres y 240 hombres), de los cuales 89 eran GBR, 99 CEU, 98 FIN, 107 IBS y 100 TSI.

Mediante *PLINK* y *PRSice-2* (opción "no-regress") se calcularon puntuaciones de riesgo para los 493 individuos de las 5 poblaciones europeas (Figuras 1A y 2A, respectivamente). Asimismo, mediante *PLINK* y *PRSice-2* (opción "regress") se calcularon PRS para los 1.711 individuos de la población española (Figuras 1B y 2B, respectivamente). Utilizando *PRSice-2* con la opción "regress", los mejores PRS se encontraban en el umbral de p-valor $9*10^{-6}$ y explicaban el 1,2% de la varianza del riesgo genético ($R^2$).

### *Análisis estadísticos básicos de las distribuciones de los valores de PRS*

En cuanto a los análisis estadísticos de los PRS de poblaciones europeas, los datos cumplieron los supuestos de normalidad y homocedasticidad. En la comparación de medias de los PRS entre hombres y mujeres, para cada población se obtuvo un p-valor mediante la prueba T de Student (Tabla 1). Al comparar las medias poblacionales de los PRS (Tabla 1), en el caso de *PLINK*, el ANOVA de un factor dio un p-valor de 0.301, mientras que para *PRSice-2* ("no-regress"), se obtuvo un p-valor de 0.0001. Para este último, se llevó a cabo un test de Tukey (Tabla 2). Para analizar la interacción entre la población y el sexo, se realizó un ANOVA de dos factores que dio un p-valor de 0.7333 en el caso de *PLINK* y 0.6287 en el caso de *PRSice-2*.

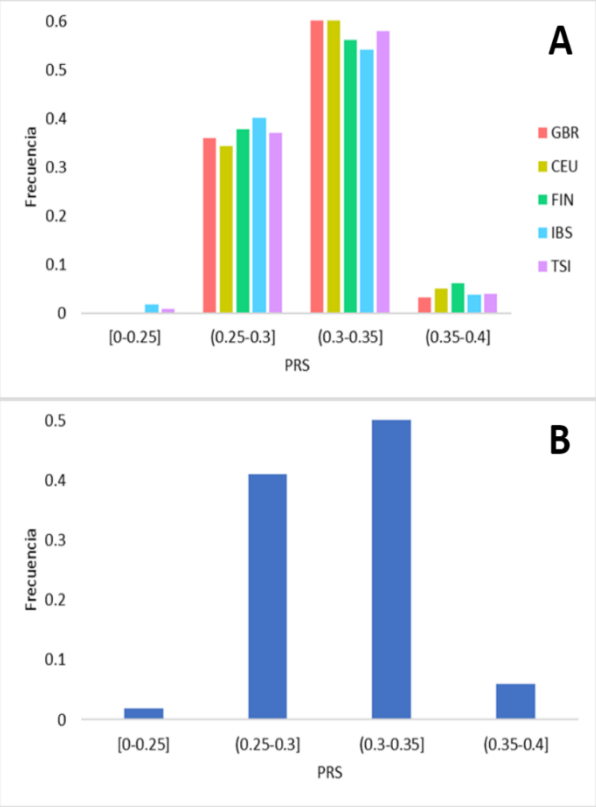


Figura 1. PRS calculados con *PLINK*. A: Distribución de frecuencias de PRS para las 5 poblaciones europeas. B: Distribución de frecuencias de PRS para la población española del proyecto SCOURGE.

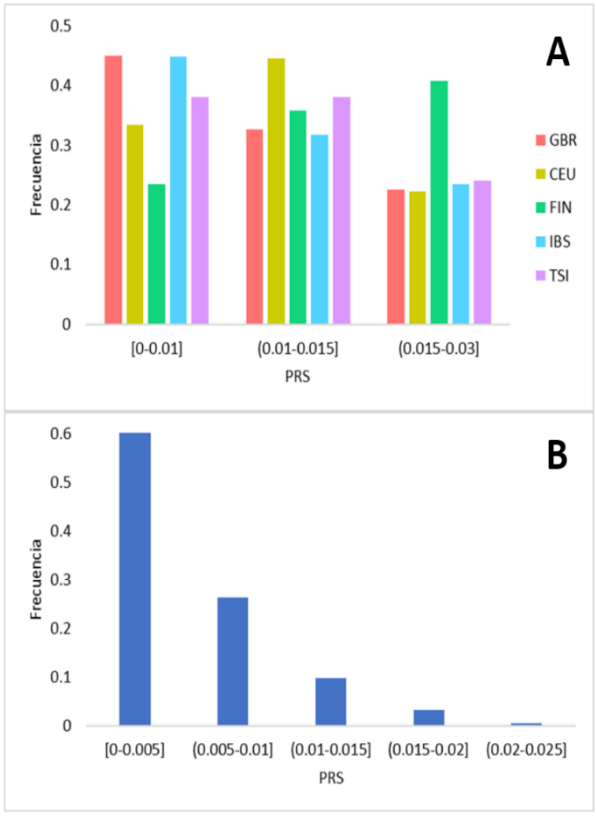


Figura 2 (A y B). PRS calculados con *PRSice-2*. A: Distribución de frecuencias de PRS para las 5 poblaciones europeas (opción "no-regress"). B: Distribución de frecuencias de PRS para la población española del proyecto SCOURGE (opción "regress").

Tabla 1. Medias de PRS (*PLINK* y *PRSice-2* con opción "no-regress") poblacionales, de mujeres y de hombres, y p-valores obtenidos mediante la prueba T de Student en cada población.

| | ***PLINK*** | | | | ***PRSice-2* ("no-regress")** | | | |
|---|---|---|---|---|---|---|---|---|
| | Media PRS | | | p-valor | Media PRS | | | p-valor |
| | Poblacional | Mujeres | Hombres | | Poblacional | Mujeres | Hombres | |
| **GBR** | 0.3082 | 0.3058 | 0.3104 | 0.3958 | 0.0113 | 0.0111 | 0.0115 | 0.7444 |
| **CEU** | 0.3096 | 0.3076 | 0.3116 | 0.4382 | 0.0117 | 0.0115 | 0.0120 | 0.6297 |
| **FIN** | 0.3078 | 0.3058 | 0.3099 | 0.4023 | 0.0144 | 0.0138 | 0.0150 | 0.2924 |
| **IBS** | 0.3024 | 0.3020 | 0.3029 | 0.8510 | 0.0111 | 0.0106 | 0.0116 | 0.3784 |
| **TSI** | 0.3078 | 0.3098 | 0.3058 | 0.4234 | 0.0119 | 0.0124 | 0.0114 | 0.2704 |

Adicionalmente, los PRS computados por ambos programas se representaron gráficamente en diagramas de caja (Figura 3).

Para estudiar las diferencias entre las puntuaciones de riesgo calculadas por *PLINK* y *PRSice-2* (opción "no-regress") en poblaciones europeas, se llevó a cabo un análisis de correlación de los PRS normalizados de ambos programas que dio un coeficiente de correlación de Pearson (r) de 0.9932 (p-valor < e-10). Asimismo, el análisis de correlación de los PRS normalizados de la población española calculados por *PLINK* y *PRSice-2* (opción "regress") reveló un r de 0.9969 (p-valor < e-10). Por otro lado, el análisis de correlación entre los PRS de la población IBS y los PRS de la población española de SCOURGE reveló un r de 0.9976 (p-valor < e-10) en el caso de *PLINK* y un r de 0.9445 en el caso de *PRSice-2* (p-valor < e-10).

***Relevancia del componente genético neandertal en el riesgo genético a padecer COVID-19 grave***

El análisis de correlación de los PRS normalizados de la población española con el componente genético neandertal de cada individuo reveló un r de 0.0313 (p-valor de 0.1947).

Tabla 2. Resultados del test de Tukey. Se indica el p-valor para cada par de medias poblacionales. * p-valor significativo al 0.05 ** p-valor significativo al 0.005

| | FIN | GBR | CEU | IBS | TSI |
|---|---|---|---|---|---|
| **FIN** | - | 0.0011** | 0.0090* | 0.0002** | 0.0130* |
| **GBR** | | - | 0.9632 | 0.9997 | 0.9343 |
| **CEU** | | | - | 0.9000 | 0.9999 |
| **IBS** | | | | - | 0.8475 |
| **TSI** | | | | | - |

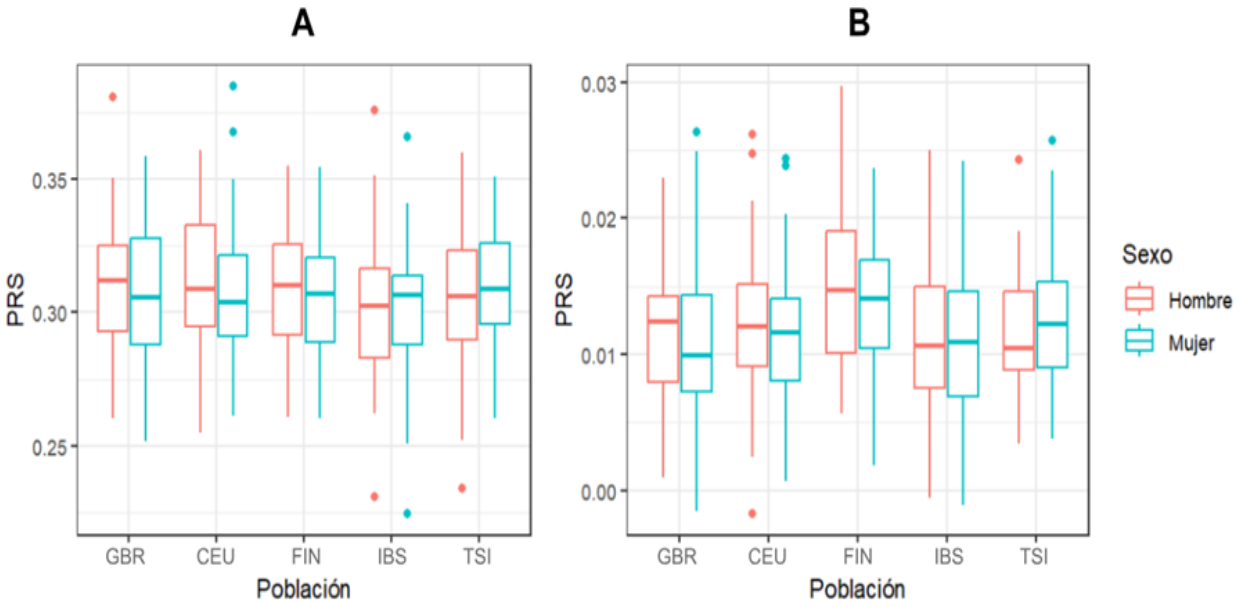


Figura 3. Diagramas de caja de los PRS de hombres y mujeres de cada población. A: *PLINK*. B: PRSice-2.

# Discusión

En cuanto a los PRS de poblaciones europeas computados con *PLINK*, se observa cierta homogeneidad entre las poblaciones (Figura 1A). Esto se confirma al analizar el p-valor del ANOVA de un factor, que no es estadísticamente significativo. Por ello, se puede decir que no se detectan diferencias significativas entre las medias de los PRS de las 5 poblaciones (Tabla 1).

En el caso de *PRSice-2*, se puede observar que los finlandeses están muy representados en valores altos de PRS (Figura 2A). Respecto a los resultados de los análisis estadísticos, el p-valor del ANOVA de un factor sugiere que al menos un par de medias poblacionales de los PRS son significativamente distintas entre sí (Tabla 1). Para analizar estas diferencias, el test de Tukey reveló diferencias significativas entre todos los pares de poblaciones que incluían a los finlandeses (Tabla 2). Esto puede deberse a que hay una mayor distancia genética entre los finlandeses y otras poblaciones europeas (Huckins et al., 2014). De hecho, en un análisis de componentes principales llevado a cabo por Huckins et al. (2014) se observa que los finlandeses están separados del resto de europeos y que se localizan en un extremo del eje de variación genética norte-sur. Por otro lado, al comparar los valores de PRS entre hombres y mujeres (Tabla 1), tanto en el caso de *PLINK* como en el de *PRSice-2*, no se detectan diferencias significativas entre la media de PRS de hombres y la de mujeres en ninguna de las poblaciones.

Respecto al ANOVA de dos factores (población y sexo), tanto en el caso de *PLINK* como en el de *PRSice-2* se puede afirmar que la interacción entre la población y el sexo no es significativa y la combinación de estas variables no influye en los PRS.

Excepto en la población toscana, los hombres tienen mayores valores de PRS que las mujeres (Figura 3), si bien esta diferencia no es significativa. Estos resultados discrepan con otros estudios en los que se demuestra que el sexo masculino es un factor de riesgo asociado a la gravedad de la COVID-19 (Vahidy et al., 2021). De acuerdo con Stokes et al. (2020), los hombres muestran síntomas más graves que las mujeres y una prevalencia significativamente mayor de hospitalizaciones (16% en hombres frente a 12% en mujeres), ingresos en UCI (3% frente a 2%) y muertes (6% frente a 5%). Las diferencias entre sexos en cuanto a la gravedad de la enfermedad pueden deberse a dos factores. Por un lado, en los hombres hay una mayor presencia de comorbilidades, como enfermedades cardiovasculares, y de conductas de alto riesgo, como el tabaquismo y el consumo de alcohol (Vahidy et al., 2021). Por otro lado, la respuesta inmunitaria puede variar entre sexos. Se ha descrito que los hombres son más susceptibles a los patógenos, mientras que las mujeres desarrollan una respuesta antigénica más fuerte (Ahnstedt y McCullough, 2019).

En general, los PRS calculados por *PLINK* muestran una fuerte correlación con los PRS calculados por *PRSice-2*. La buena correlación indica que con ambos programas es posible identificar a los mismos individuos de mayor riesgo, independientemente del método usado. Aun así, *PRSice-2* arroja resultados más precisos y discrimina mejor el riesgo, como en el caso de los finlandeses (Figura 2A). Por ello, en el cálculo de PRS es más recomendable utilizar *PRSice-2* y, en concreto, la opción "regress" siempre que sea posible obtener datos fenotípicos. En este estudio, en el caso de *PRSice-2* con la opción "regress" (Figura 2B), los PRS han sido calculados con un conjunto de SNPs que se encuentran en el umbral de p-valor $9*10^{-6}$ y que explican el 1,2% de la varianza fenotípica ($R^2$). Este valor de $R^2$ se considera bajo en comparación con otros rasgos complejos más estudiados, como algunos rasgos antropométricos (índice de masa corporal y altura), para los que se han calculado PRS que explican en torno a un 25% de la varianza fenotípica (Yengo et al., 2018). Sin embargo, puede considerarse superior a otros rasgos, como el eccema ($R^2 = 0.07\%$), y similar a padecer migrañas ($R^2 = 1.76\%$), abuso de alcohol ($R^2 = 1.62\%$), alergia al polen ($R^2 = 0.87\%$) o asma ($R^2 = 1.72\%$) (Becker et al., 2021).

Además, los PRS de la población IBS del Proyecto 1000 Genomas muestran una fuerte correlación con los PRS de la población española del proyecto SCOURGE. Esta congruencia del riesgo genético entre ambas poblaciones españolas supone un control de calidad adicional y refuerza los resultados del trabajo.

Por otra parte, en este trabajo se ha observado que hay una débil correlación entre el componente

genético neandertal global y los PRS de la población española. En un estudio más reciente, Zeberg y Pääbo (2021) describieron otro haplotipo de origen neandertal en el cromosoma 12 de unas 75 kb que estaba asociado a la protección contra la COVID-19 grave. En la región de este haplotipo se encuentran genes involucrados en la respuesta contra infecciones víricas y, de hecho, se ha descubierto que varios SNPs del haplotipo del cromosoma 12 tienen un efecto protector contra algunos virus de RNA, como el virus del Nilo Occidental (Lim et al., 2009), el virus de la hepatitis C (El Awady et al., 2011) y el SARS-CoV (He et al., 2006). Por lo tanto, dado que el ADN neandertal introgresado puede resultar asociado tanto a protección como a gravedad, esta puede ser una de las razones por las que la correlación entre el componente genético neandertal global y el riesgo genético a padecer COVID-19 grave es tan débil.

En cuanto a las limitaciones de este estudio, no se tenía información fenotípica sobre el COVID-19 en las poblaciones europeas del Proyecto 1000 Genomas, por lo que no se pudo considerar esta información en el cálculo de los PRS para estas poblaciones. Por otro lado, una de las mayores limitaciones generales de los estudios de GWAS y PRS es la escasa transferibilidad de los resultados entre poblaciones (Martin et al., 2019). A la hora de calcular PRS, el uso de *base data* proveniente de poblaciones de ascendencia europea tiene una peor capacidad predictiva del riesgo en poblaciones de ascendencia no europea y viceversa. Esto es debido a las diferencias en las frecuencias de haplotipos y los diferentes efectos que pueden tener las mismas variantes en los distintos grupos humanos dependiendo tanto de los antecedentes genéticos de los individuos como del ambiente (Duncan et al., 2019; Abdellaoui et al., 2023).

En conclusión, la población finlandesa del Proyecto 1000 Genomas difiere con el resto de poblaciones europeas en su riesgo promedio a desarrollar un COVID-19 grave; además, ni el sexo ni el componente genético neandertal parecen ser variables de relevancia en el riesgo genético. En definitiva, el método PRS aplicado a la COVID-19 es una herramienta eficaz en el ámbito clínico que permite identificar y monitorizar a los pacientes con mayor riesgo de desarrollar COVID-19 grave. También es de gran utilidad para identificar poblaciones vulnerables en las que pueden ser necesarias medidas especiales para prevenir la transmisión de la enfermedad.

## Agradecimientos

Nuestro agradecimiento a los miembros del consorcio SCOURGE, en particular al comité de dirección del consorcio (Ángel Carracedo, Pablo Lapunzina, Augusto Rojas-Martínez, José A. Riancho y Carlos Flores) por otorgarnos acceso a los datos de resultados agregados del GWAS. La investigación en SA lab está financiada por Fondos del Gobierno Vasco a Grupos de Investigación del País Vasco (IT-1693-22).

## Bibliografía

Abdellaoui A., Yengo L., Verweij K.J., Visscher P.M. (2023). 15 years of GWAS discovery: Realizing the promise. Am J Hum Genet 110(2): 179-194. https://doi.org/10.1016/j.ajhg.2022.12.011

Ahnstedt H., McCullough L.D. (2019). The impact of sex and age on T cell immunity and ischemic stroke outcomes. Cell Inmunol 345: 103960. https://doi.org/10.1016/j.cellimm.2019.103960

Becker J., Burik C.A., Goldman G., Wang N., Jayashankar H., Bennett M., et al. (2021). Resource profile and user guide of the Polygenic Index Repository. Nat Hum Behav 5(12): 1744-1758. https://doi.org/10.1038/s41562-021-01119-3

Chang C.C., Chow C.C., Tellier L.C., Vattikuti S., Purcell S.M., Lee J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4(1): s13742-015-0047-8. https://doi.org/10.1186/s13742-015-0047-8

Choi S.W., Mak T.S., O'Reilly P.F. (2020). Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc 15(9): 2759-2772. https://doi.org/10.1038/s41596-020-0353-1

COVID-19 Host Genetics Initiative (2021). Mapping the human genetic architecture of COVID-19. Nature 600(7889): 472-477. https://doi.org/10.1038/s41586-021-03767-x

COVID-19 Host Genetics Initiative (8 de abril de 2022). *COVID19-hg GWAS meta-analyses round 7*. https://www.covid19hg.org/results/r7/

Cruz R., Diz-de Almeida S., López de Heredia M., Quintela I., Ceballos F.C., Pita G., et al. (2022). Novel genes and sex differences in COVID-19 severity. Hum Mol Genet 31(22): 3789-3806. https://doi.org/10.1093/hmg/ddac132

Degenhardt F., Ellinghaus D., Juzenas S., Lerga-Jaso J., Wendorff M., Maya-Miles D., et al. (2022). Detailed stratified GWAS analysis for severe COVID-19 in four European populations. Hum Mol Genet 31(23): 3945-3966. https://doi.org/10.1093/hmg/ddac158

Dudbridge F. (2016). Polygenic Epidemiology. Genet Epidemiol 40(4): 268-272. https://doi.org/10.1002/gepi.21966

Duncan L., Shen H., Gelaye B., Meijsen J., Ressler K., Feldman M., et al. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun 10(1): 3328. https://doi.org/10.1038/s41467-019-11112-0

El Awady M.K., Anany M.A., Esmat G., Zayed N., Tabll A.A., Helmy A., et al. (2011). Single nucleotide polymorphism at exon 7 splice acceptor site of OAS1 gene determines response of hepatitis C virus patients to interferon therapy. J Gastroenterol Hepatol 26(5): 843-850. https://doi.org/10.1111/j.1440-1746.2010.06605.x

Gunz P., Tilot A.K., Wittfeld K., Teumer A., Shapland C.Y., Van Erp T.G., et al. (2019). Neandertal Introgression Sheds Light on Modern Human Endocranial Globularity. Curr Biol 29(1): 120-127. https://doi.org/10.1016/j.cub.2018.10.065

He J., Feng D., de Vlas S.J., Wang H., Fontanet A., Zhang P., et al. (2006). Association of SARS susceptibility with single nucleic acid polymorphisms of OAS1 and MxA genes: a case-control study. BMC Infect Dis 6: 1-7. https://doi.org/10.1186/1471-2334-6-106

Horowitz J.E., Kosmicki J.A., Damask A., Sharma D., Roberts G.H., Justice A.E., et al. (2022). Genome-wide analysis provides genetic evidence that ACE2 influences COVID-19 risk and yields risk scores associated with severe disease. Nat Genet 54(4): 382-392. https://doi.org/10.1038/s41588-021-01006-7

Huckins L.M., Boraska V., Franklin C.S., Floyd J.A., Southam L., Sullivan P.F., et al. (2014). Using ancestry-informative markers to identify fine structure across 15 populations of European origin. Eur J Hum Genet 22(10): 1190-1200. https://doi.org/10.1038/ejhg.2014.1

Kousathanas A., Pairo-Castineira E., Rawlik K., Stuckey A., Odhams C.A., Walker S., et al. (2022). Whole-genome sequencing reveals host factors underlying critical COVID-19. Nature 607(7917): 97-103. https://doi.org/10.1038/s41586-022-04576-6

Lim J.K., Lisco A., McDermott D.H., Huynh L., Ward J.M., Johnson B., et al. (2009). Genetic variation in OAS1 is a risk factor for initial infection with West Nile virus in man. PLoS Pathog 5(2): e1000321. https://doi.org/10.1371/journal.ppat.1000321

Machiela M.J., Chanock S.J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 31(21): 3555-3557. https://doi.org/10.1093/bioinformatics/btv402

Marees A.T., de Kluiver H., Stringer S., Vorspan F., Curis E., Marie-Claire C., Derks E.M. (2018). A tutorial on conducting genome wide association studies: Quality control and statistical analysis. Int J Method Psychiatr Res 27(2): e1608. https://doi.org/10.1002/mpr.1608

Martin A.R., Kanai M., Kamatani Y., Okada Y., Neale B.M., Daly M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet 51(4): 584-591. https://doi.org/10.1038/s41588-019-0379-x

Osterman M.D., Kinzy T.G., Cooke Bailey J.N. (2021). Polygenic Risk Scores. Curr Protoc 1: e126. https://doi.org/10.1002/cpz1.126

Pairo-Castineira E., Clohisey S., Klaric L., Bretherick A.D., Rawlik K., Pasko D., et al. (2021). Genetic mechanisms of critical illness in COVID-19. Nature 591(7848): 92-98. https://doi.org/10.1038/s41586-020-03065-y

Sollis E., Mosaku A., Abid A., Buniello A., Cerezo M., Gil L., et al. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res 51(D1): D977-D985. https://doi.org/10.1093/nar/gkac1010

Stokes E.K., Zambrano L.D., Anderson K.N., Marder E.P., Raz K.M., Felix S.E.B., et al. (2020). Coronavirus Disease 2019 Case Surveillance-United States, January 22-May 30, 2020. Morb Mortal Wkly Rep 69(24): 759-765. https://doi.org/10.15585%2Fmmwr.mm6924e2

Tang D., Comish P., Kang R. (2020). The hallmarks of COVID-19 disease. PLoS Pathog 16(5): e1008536. https://doi.org/10.1371/journal.ppat.1008536

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature 526(7571): 68-74. https://doi.org/10.1038/nature15393

Thibord F., Chan M.V., Chen M.H., Johnson A.D. (2022). A year of COVID-19 GWAS results from the GRASP portal reveals potential genetic risk factors. Hum Genet Genom Adv 3(2): 100095. https://doi.org/10.1016/j.xhgg.2022.100095

Uricchio L.H. (2020). Evolutionary perspectives on polygenic selection, missing heritability, and GWAS. Hum Genet 139(1): 5-21. https://doi.org/10.1007/s00439-019-02040-6

Vahidy F.S., Pan A.P., Ahnstedt H., Munshi Y., Choi H.A., Tiruneh Y., et al. (2021). Sex differences in susceptibility, severity, and outcomes of coronavirus disease 2019: Cross-sectional analysis from a diverse US metropolitan area. PloS ONE 16(1): e0245556. https://doi.org/10.1371/journal.pone.0245556

World Health Organization. *WHO Coronavirus (COVID-19) Dashboard*. Recuperado el 3 de mayo de 2023 de https://covid19.who.int/

Yengo L., Sidorenko J., Kemper K.E., Zheng Z., Wood A.R., Weedon M.N., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum Mol Genet 27(20): 3641-3649. https://doi.org/10.1093/hmg/ddy271

Zeberg H., Pääbo S. (2020). The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. Nature 587(7835): 610-612. https://doi.org/10.1038/s41586-020-2818-3

Zeberg H., Pääbo S. (2021). A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. Procl Natl Acad Sci (PNAS) 118(9): e2026309118. https://doi.org/10.1073/pnas.2026309118