ESCOLA DE DOUTORAMENTO
INTERNACIONAL DA USC

Silvia
Diz de Almeida

Tese de doutoramento

Análise xenética-poboacional da variabilidade na gravidade da COVID-19

Santiago de Compostela, 2024

Programa de Doutoramento en Medicina Molecular

ESCOLA DE DOUTORAMENTO
INTERNACIONAL DA USC

TESE DOUTORAL

ANÁLISE XENÉTICA-POBOACIONAL DA VARIABILIDADE NA GRAVIDADE DA COVID-19

SILVIA DIZ DE ALMEIDA

Directores: Raquel Cruz Guerrero, Ángel Carracedo Álvarez

Tutor: Ángel Carracedo Álvarez

PROGRAMA DE DOUTORAMENTO EN MEDICINA MOLECULAR

SANTIAGO DE COMPOSTELA

2024

ESCOLA DE DOUTORAMENTO
INTERNACIONAL DA USC

DECLARACIÓN ÉTICA DA TESE

A presente tese doutoral foi levada a cabo baixo a aprobación do Comité de ética da investigación con medicamentos de Galicia (CEIm-G) con código 2020/197 (documento completo no **Anexo XI**.

A autora da tese declara non ter ningún conflito de interese relacionado coa tese.

As figuras incluídas aquí, ou ben son obra propia da doutoranda, ou contan cos permisos correspondentes ao proceder de artigos baixo licencia *Creative Commons License 4.0.*

AGRADECEMENTOS

De pequena, había días nos que quería ser xornalista, outros nos que quería ser bióloga, e algúns poucos nos que só quería ser famosa. Finalmente, aínda que con moitas dúbidas, gañou a bioloxía. Durante os catro anos de carreira pregunteime se tomara a decisión correcta, se de verdade había sitio un sitio aí para min. Uns anos despois, gustaríame poder dicirlle á Silvia universitaria que agardase. Que agardase porque un día ía desfrutar da xenética, ía ir contenta a traballar e ía pasar horas loitando contra un *script*. Que un día, non moi lonxe, ía atopar un grupo de persoas que conseguirían recordarlle por que escollera as ciencias, que lle axudarían a atopar o seu sitio, que *escribiría unha tese doutoral*. Como o tempo non corre cara atrás, déixoo aquí, en papel. É a elas a quen lles quero dar as grazas en primeiro lugar.

A Raquel, Ali, Sara, Cata, María, Sabela, Marta, Montse, Cris, Inés, e a todas as que me falten. Grazas por ser científicas, biólogas, psicólogas, amigas, nais. Grazas por construír un espazo onde caben as ideas e a colaboración, pero tamén as risas, o cariño, a confianza, as queixas (e os cafés, moitos cafés). Grazas polo apoio incondicional durante a tese, por deixarme aprender algo de todas vós, e especialmente por ser tan bo público ás miñas pallasadas. Grazas porque, a pesar de calquera obstáculo, sempre estades empurrando máis forte e máis alto. Sara, grazas por ter sido a miña siamesa ata o final, once anos despois daquela viaxe a NY. Ali e María, merecedes unha mención especial por ler esta tese cando eu xa non daba para máis.

A Raquel, de novo, porque debería dedicarche catro páxinas só a ti (non o tomes ao pé da letra, porfi). Grazas por aquel "Silvia, pensei en ti" que recibín un deses días nos que estábamos confinadas, cando parecía que o mundo fóra acababa, e que deron comezo a este traballo. Grazas por apostar por min. Pola paciencia infinita, por ensinarme todo o que sabes, polo apoio imprescindible e por sentarte comigo en cada resultado. Pero tamén por lerme tan ben, por aguantar os meus días negativos, por darme liberdade cando tiña unha idea e orientarme cando estaba desbordando o vaso, por insistir en que me sentira orgullosa. Grazas por dirixir, con todas as letras, esta tese.

A Ángel, grazas por abrirme as portas da túa casa e por darme esta oportunidade; por confiar neste proxecto e, sobre todo, confiar no que eu podía facer con el. Grazas, tamén, polo equipo de *Genomics & Bioinformatics*. As derivas da vida poderían ter confluído nalgo parecido, pero dende logo que se necesita un esforzo e un ollo máxico para que algo saia tan ben.

Grazas a ambos pola humanidade, por demostrar que se pode investigar dende outro lugar.

Ao Dr. Carlos Flores, grazas por revisar e mellorar este traballo coa dedicación e coidado dun terceiro director.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Grazas ao resto do grupo de Medicina Xenómica e do CEGEN cos que tiven o placer de coincidir. Especialmente a Laura M., coa que compartín uns meses de angustias bioinformáticas.

Ao CESGA, por permitirme empregar os seus servizos para a realización das análises.

A todo o equipo de SCOURGE. E un especial agradecemento ás traballadoras que deron recollido toda esta información en medio dunha pandemia.

A Iván, o meu fan número un. Grazas infinitas, por ser terra firme todo este tempo, por darlle valor ao meu traballo cando eu non era capaz. Que isto estea acabado é, en gran parte, grazas a ti.

A Nuria, por entender cada unha das etapa e estar en todas elas suxeitándome. A Pablo, Alba, Antonio, Xaime, Belén, Ana B., aos de Noia e ás de Porri, por dar un sostén que vale ouro, moitas dende a distancia.

Por último, á miña familia, política tamén. Grazas, porque ao final o que mellor me sentaba nos peores momentos era volver a casa. Grazas, porque podo escoitar o orgullo dende aquí.

Esta tese ten un anaco de cada unha de vós.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ABREVIATURAS

1KGP	1000 *Genomes Populations*
53BP	p53 *binding protein*, proteína de unión a p53
AF	*allele frequency*, frecuencia alélica
AFR	poboacións africanas
AL	ancestralidade local
AMR	poboacións americanas mixtas
ARDS	*acute respiratory distress syndrome*, síndrome de estrés respiratorio agudo
AUC	*area under the* ROC *curve*
BNADN	Banco Nacional de ADN
CAD	*coronary artery disease*, enfermidade da arteria coronaria
CADD	Combined Annotation Dependent Depletion
CNIO	Centro Nacional de Investigaciones Oncológicas
DCA	*decision curve analysis*, análise da curva de decisión
DDR	*DNA damage response repair*, resposta de reparación por dano no ADN
DSB	*double strand break*, roturas de dobre hebra
EAS	poboacións do este de Asia
EPOC	enfermidade obstructiva crónica
eQTL	*expression quantitative trait loci*, *loci* asociado á expresión
EUR	poboacións europeas
FDR	*false discovery rate*, taxa de falsos positivos
FRC	modelo de factores de risco clínicos
GDS	*CoreArray Genomic Data Structure*
GIA	*genetically inferred ancestry*, ancestralidade xenética inferida
GRM	*genetic relationship matrix*, matriz de relacións xenéticas
GRS	*genetic risk score*, índice de risco xenético
GTEx	*the Genotype-Tissue Expression Consortium*
GWAS	*Genome-wide Association Study*, estudo de asociación de xenoma completo
HGI	*COVID-19 Host Genetics Initiative*
HR	*homologous recombination*, recombinación homóloga
IBD	*identity-by-descent*
IC	Intervalo de confianza
IFs	interferóns
IVW	*inverse-variance weighted*, ponderado pola inversa da varianza
LD	*linkage desequilibrium*, desequilibrio de ligamento
MAE	*mean absolute error*, erro medio absoluto
MAF	*minor allele frequency*, frecuencia do alelo menor
MPP	Medicina Personalizada de Precisión
NAM	poboacións nativo-americanas

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

NES	*normalized effect size*, tamaño de efecto normalizado
NHEJ	*non homologous end-joining*
NRI	*net reclassification improvement*, melloría na reclasificación neta
NES	normalized effect size, tamaño de efecto normalizado
NHEJ	non homologous end-joining
NRI	net reclassification improvement, melloría na reclasificación neta
OMS	organización mundial da saúde
OR	*odds-ratio*
PC	*principal component, compoñente principal*
PCA	*principal component analysis*, análise de compoñentes principais
PCR	*polymerase chain reaction*, reacción en cadea da polimerasa
PIP	*posterior inclusion probability*, probabilidade posterior de inclusión
PRS	*polygenic risk score*, índice de risco polixénico
QC	*quality control*, control de calidade
ROC	*receiver operating characteristic curve*
SAS	poboacións do sur de Asia
s.d.	desviación típica
SCOURGE	*Spanish COalition to Unlock Research on host GEnetics on COVID-19*
SNP	*single nucleotide polymorphism*, polimorfismo dun só nucleótido
TLR	receptor tipo *Toll*
TWAS	*Transcriptome-Wide Association Studies*, estudo de asociación de transcriptoma completo

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

ÍNDICE

Resumo 13
1 Introdución 15
1.1 Xenética dos caracteres complexos 15
1.2 Estudos de asociación de xenoma completo (GWAS) 16
1.2.1 Estudos de asociación con datos WGS e WES 19
1.2.2 Consideracións adicionais dos GWAS 20
1.3 Ancestralidade xenética 22
1.3.1 *Admixture* xenético 24
1.3.2 Ancestralidade e poboacións mixtas nos estudos GWAS 26
1.4 Diversidade nos estudos GWAS 27
1.4.1 Correlación dos efectos causais entre grupos poboacionais 31
1.5 índices de risco polixénico 31
1.5.1 Construción dos PRS 32
1.5.2 Aplicacións dos PRS 33
1.5.3 Limitacións dos PRS 35
1.6 COVID-19 39
1.6.1 Patoxénese da COVID-19 41
1.6.2 Factores de risco 43
1.6.3 Estudo da susceptibilidade xenética á covid-19 45
1.6.4 Variantes xenéticas implicadas na COVID-19 47
1.6.4.1 Xenes implicados na entrada viral e susceptibilidade á infección 47
1.6.4.2 Xenes implicados na gravidade da COVID-19 49
1.6.5 O consorcio SCOURGE 51
2 Hipóteses e Obxectivos 53
3 Metodoloxía 55
3.1 Recolección dos datos 55
3.2 Xenotipado 57
3.3 Definición das cohortes 57
3.4 Control de calidade dos datos xenéticos 61
3.5 Estimación da ancestralidade xenética inferida global 61
3.6 Imputación de xenotipos 63

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.7 Análises de asociación de xenoma completo (GWAS) 64
3.7.1 Estudo GWAS na cohorte latinoamericana 65
3.8 Meta-análise de estudos gwas 65
3.9 Anotación funcional 67
3.10 Estatística baiesiana vs estatística frecuentista 68
3.11 Fine-mapping 69
3.12 Datos de expresión 70
3.12.1 Estudos de asociación de transcriptoma completo (TWAS) 70
3.13 Índices de risco xenético 72
3.13.1 Cohortes de descubrimento e diana 73
3.13.2 PRS C+T 73
3.13.3 PRS-CS e PRS-CSx 74
3.13.4 Selección dos hiperparámetros dos PRS 76
3.13.4.1 Selección do PRS óptimo 77
3.13.5 Construción dun PRS único e validación 77
3.13.5.1 Desenvolvemento do modelo 77
3.13.5.2 Validación do modelo 78
3.13.6 Estudo do PRS en relación ás variables clínicas 79
4 Factores xenéticos da COVID-19 en poboación española 83
4.1 Resultados 83
4.1.1 Fase de descubrimento 83
4.1.1.1 Análise global 85
4.1.1.2 Estratificación por sexo e idade 86
4.1.2 Fase de validación 91
4.1.2.1 Meta-análise CNIO-SCOURGE 91
4.1.2.2 TWAS de hospitalización por COVID-19 en poboación española 95
4.1.2.3 Avaliación das asociacións das variantes líder en relación coas comorbilidades 96
4.1.3 Replicación de variantes previamente asociadas coa hospitalización por COVID-19 97
4.1.4 Meta-análise con estudos europeos 97
4.2 Discusión 99
4.2.1 Diferencias no risco xenético entre sexos e grupos de idade 99
4.2.2 Asociacións nos xenes *AQP3*, *ARHGAP27* e *UPK1A* 101

4.2.3 Novos xenes identificados a través dos estudos TWAS 102
4.2.4 Integración dos resultados de SCOURGE noutros estudos 103
5 Factores xenéticos da COVID-19 en poboación latinoamericana 105
5.1 Resultados 105
5.1.1 GWAS na cohorte LatAm 105
5.1.2 Meta-análise dos estudos AMR 107
5.1.2.1 TWAS para hospitalización por COVID-19 en poboación AMR 111
5.1.3 Exploración da arquitectura xenética da hospitalización por COVID-19 en poboación AMR 112
5.1.4 Meta-análises trans-poboacionais e replicación de sinais 114
5.2 Discusión 116
5.2.1 Os primeiros xenes identificados en poboación AMR: *DDIAS* e *BAZ2B* 116
5.2.2 Novos xenes identificados a través do TWAS 117
5.2.3 Factores xenéticos da COVID-19 entre poboacións 118
5.2.4 Limitacións do estudo 119
6 índices de risco polixénico para COVID-19 121
6.1 Resultados 121
6.1.1 Selección dos PRS 121
6.1.2 Construción dun PRS único a partir do PRS óptimo 126
6.1.3 PRS en relación ás variables clínicas 129
6.1.4 Modelos FRC e FRC+PRS 132
6.1.4.1 Curvas DCA 135
6.1.4.2 Exemplo de aplicación clínica do PRS 137
6.2 Discusión 138
6.2.1 Avaliación da capacidade preditiva do PRS entre poboacións 138
6.2.2 Obtención dun PRS único para a hospitalización por COVID-19 139
6.2.3 Estudo do potencial clínico do PRS 141
7 Discusión xeral 143
7.1.1 Deseño dos estudos GWAS 144
7.1.2 Replicación de sinais entre poboacións 146
7.1.3 Construción de PRS en cohortes diversas 148
7.1.4 Aplicación clínica dos PRS 150
8 Conclusións 153
9 Bibliografía 155

10 Anexos 197
11 ÍNDICE DE TÁBOAS 231
12 ÍNDICE DE FIGURAS 233

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

RESUMO

A enfermidade da COVID-19 mostra unha gran variabilidade fenotípica entre os individuos, podendo manifestarse na forma asintomática ou progresar ata un estado crítico. Algúns dos factores de risco máis relevantes son o sexo, a idade e a presencia de antecedentes clínicos. Non obstante, parte desta variación tamén pode atribuírse á base xenética dos hóspedes, influíndo tanto na susceptibilidade á infección como na resposta do sistema inmune. Ademais, a arquitectura xenética da enfermidade podería ser variable entre as poboacións. Os obxectivos desta tese foron abordar esas cuestións. No marco do proxecto SCOURGE, establecido en 2020, recolléronse datos clínicos e xenéticos de máis de 10.000 pacientes de España e Latino-América. Por unha banda, buscouse investigar e identificar os determinantes xenéticos que actúan na predisposición á COVID-19 grave a través de estudos de asociación de xenoma completo (GWAS) en ambas as poboacións. Ademais de confirmar a asociación de xenes xa coñecidos, identificáronse asociacións novas en xenes como *AQP3*, *UPK1A*, *DDIAS* ou *BAZ2B*. Posteriormente, construíronse índices de risco polixénico (PRS) e avaliouse a súa capacidade para predicir a hospitalización por COVID-19, prestando especial atención á xeneralización a poboacións non europeas e a unha potencial aplicación clínica.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1 INTRODUCIÓN

1.1 XENÉTICA DOS CARACTERES COMPLEXOS

A teoría da herdanza polixénica e o modelo infinitesimal de Fisher son os modelos máis aceptados para explicar a xenética dos caracteres complexos, definidos como aqueles que non seguen unha herdanza puramente mendeliana e cuxa expresión depende dunha combinación de factores xenéticos e ambientais (Goddard et al., 2016). O traballo de Fisher, revisado en Visscher e Goddard (2019) concluíu que un carácter cuantitativo está influenciado por multitude de xenes, cada un contribuíndo a unha parte da varianza do mesmo chamada varianza aditiva xenética, mentres que outras compoñentes xenéticas e medioambientais son desviacións do modelo aditivo. Cada xen ten un efecto pequeno sobre o fenotipo, e a combinación aleatoria dos alelos destes xenes causa que o fenotipo teña unha distribución aproximadamente normal na poboación a medida que aumenta o número de xenes. Así, fálase da herdabilidade dun carácter en sentido estrito cando se fai referencia á proporción da variabilidade fenotípica atribuíble a esta varianza aditiva xenética (Piepho & Möhring, 2007).

A utilidade deste modelo foi comprobada de forma experimental a través dos estudos de asociación de xenoma completo (*Genome-wide Association Studies* ou GWAS), que permitiron dende a súa aparición a identificación de centos de *loci* asociados aos caracteres complexos, dende fenotipos clásicos cuantitativos a enfermidades non mendelianas (Visscher & Goddard, 2019). No caso das enfermidades, a propiedade cuantitativa do carácter baséase na susceptibilidade xenética. Esta depende do número de variantes asociadas a certa enfermidade que un individuo presenta, tamén coñecida como "predisposición xenética". Estas variantes de risco estarán en maior frecuencia nos individuos que padezan a enfermidade que nos individuos que non a padezan, e ao igual que para os caracteres cuantitativos, ao aumentar o número de alelos de risco asociados á enfermidade tamén aumentará a predisposición xenética. Porén, presentar un risco xenético alto fronte a unha patoloxía non implicará necesariamente que esta se desenvolva, pois os factores ambientais xogan un papel fundamental.

A pregunta que aínda se está tratando de resolver é cantos *loci* contribúen en realidade á variación xenética destes caracteres complexos e como se distribúen ao longo do xenoma. Na actualidade emprégase con frecuencia o concepto de arquitectura xenética dun carácter. Esta engloba todas os factores contribuíntes á variación xenética aditiva dun fenotipo: o número de variantes que inflúen nel, o tamaño de efecto das mesmas, a frecuencia das variantes na poboación, as interaccións entre xenes e as interaccións fenotipo-ambiente (Timpson et al., 2018). Existen varios tipos de arquitecturas xenéticas, dende a monoxénica á polixénica, mais nos últimos anos propúxose o chamado modelo omnixénico (Boyle et al., 2017). Este modelo

USC UNIVERSIDADE DE COMPOSTELA

recolle todo o descuberto a través dos estudos GWAS sobre os caracteres complexos: (1) que a base xenética dos mesmos se debe principalmente á variación común ou SNPs; (2) que estes SNPs teñen uns efectos moi pequenos sobre os fenotipos; (3) que estas variantes soen estar situadas en rexións reguladoras e non son codificantes; e (4) que a base xenética está enriquecida en rexións involucradas na transcrición. O modelo omnixénico asume que os caracteres complexos están influenciados por uns poucos xenes chamados "xenes centrais", e os seus reguladores directos. Neste sentido, resaltan o papel das redes de regulación celular, moi interconectadas (Boyle et al., 2017). Finalmente, estarían os "xenes periféricos", en moito maior número, que participan na regulación e expresión dos "xenes centrais" en determinados tecidos. Por tanto, a maior parte de variantes afectarán aos xenes periféricos e terán uns efectos moi pequenos sobre o fenotipo, pero serán responsables da maior parte da herdabilidade do carácter (Boyle et al., 2017). Este modelo foi matizado posteriormente por outros autores: o concepto dos "xenes centrais" implica que aqueles con maior impacto nas enfermidades serán os que presenten variantes deletéreas (por tanto, variación rara) (Wray et al., 2018). Isto pode levar, por un lado, a unha simplificación da bioloxía dos caracteres complexos e, por outro, a que se lle outorgue menos relevancia á variación común. Wray e colaboradores (2018) apuntan que o modelo omnixénico é conceptualmente o mesmo que o modelo infinitesimal.

1.2 ESTUDOS DE ASOCIACIÓN DE XENOMA COMPLETO (GWAS)

As análises GWAS teñen como obxectivo detectar asociacións entre variantes xenéticas, normalmente SNPs, e un fenotipo, que pode ser cuantitativo ou tipo caso/control. O éxito destes estudos radica na relativa facilidade de descubrir factores de risco xenéticos asociados a enfermidades a través da variación común.

Por un lado, os GWAS parten da información xenética observada a través da tecnoloxía dos *arrays* de xenotipado, que permite obter de forma rápida e barata unha imaxe a gran escala do xenoma dun individuo. Un *array* típico cubre entre medio e un millón de posicións xenéticas distribuídas ao longo de todo o xenoma, chamadas variantes *tag*. As variantes *tag* selecciónanse de forma que sexan representativas das zonas cromosómicas onde se atopan, é dicir, representan a outras variantes coas que teñen correlación xenética pero que non están presentes no *array* (Uffelmann et al., 2021). O deseño dos *arrays* habitualmente tamén ten en conta outros factores, como a variabilidade xenética na poboación de destino e, incluso, nos *arrays* personalizados, información existente sobre determinadas enfermidades. Hoxe en día, o número de *loci* analizados nos GWAS pode multiplicarse por dez grazas a técnicas bioinformáticas como a imputación, polo que a densidade de posicións cuberta polos *arrays* é suficiente. Por outro lado, a propiedade máis importante dos GWAS é que non requiren dunha hipótese a *priori*, como nos estudos de xenes candidatos, a busca de factores xenéticos é totalmente a cegas. Escanéase o xenoma completo a través das variantes *tag*, que actuarán como indicadoras das variantes causantes dos fenotipos.

Os GWAS baséanse en dous principios: a) Na existencia de diferentes frecuencias alélicas entre grupos de individuos que presentan un fenotipo distinto, e b) no desequilibrio de ligamento entre as variantes. O procedemento habitual consiste en xenotipar unha serie de individuos representativos das poboacións de casos e de controis, se o fenotipo é

ausencia/presencia de enfermidade; ou representativos do espectro do carácter, se o fenotipo é continuo. Centrándonos nun carácter caso/control, obterase unha asociación significativa SNP-fenotipo se un dos alelos é máis frecuente nunha das dúas poboacións. Esta asociación pode ir en dúas direccións: dicimos que o alelo ten un efecto protector se a súa frecuencia é maior nos controis, e un efecto de risco se é máis frecuente nos casos. No caso dos fenotipos continuos, o efecto do alelo estará asociado a aumentar ou diminuír o carácter. Este deseño benefíciase da existencia de bloques de desequilibrio de ligamento dado que, como as variantes dun mesmo haplotipo están correlacionadas e atópanse fisicamente cerca, se pode obter un sinal significativo cunha variante que non sexa causal. Grazas a esta propiedade non é necesario dispoñer de información para todas as variantes do xenoma, pero determinar cal é o SNP causal nun sinal GWAS é un traballo complicado e precísase de análises complementarias.

O primeiro estudo GWAS publicouse en 2005, onde se compararon 50 controis e 69 casos con dexeneración macular e identificouse o factor H como un factor xenético de risco (R. J. Klein et al., 2005). A partir dese momento mudou o paradigma do estudo dos caracteres complexos. Publicáronse GWAS para fenotipos moi variados, dende enfermidades ata patróns de conduta. Por exemplo, en 2007 detectouse unha asociación no xen *FTO* coa obesidade e este xen continúa a ser un dos factor de risco xenético máis estudados, cun incremento de 0.04 kg/m^2 por alelo de risco, que ademais interactúa coa dieta do individuo (Frayling et al., 2007; Loos & Yeo, 2022). O último GWAS de esquizofrenia detectou 287 *loci* distintos en 120 xenes, enriquecidos en procesos neuronais como a sinapse e a diferenciación celular (Trubetskoy et al., 2022). Para a enfermidade de Alzheimer xa existen 45 *loci* de risco identificados, o que permitiu priorizar rutas de sinalización como a de TNF-α e a activación da microglía ou descubrir a implicación do complexo *LUBAC* (Bellenguez et al., 2022). Na área do cancro identificáronse 250 *loci* asociados á cancro colorrectal (Fernandez-Rozadilla et al., 2023), 9 asociados á cancro de endometrio (Won et al., 2014) ou 36 asociados á cancro de pulmón (Byun et al., 2022). Ademais das análises centradas en tipos concretos de cancro, o estudo do fenotipo de forma máis xeral (estudo *pan*-cáncer) tamén permitiu atopar sinais pleiotrópicos e compartidos, como algúns implicados en procesos de apoptose (Sato et al., 2023). En decembro de 2023 foi de gran interese o estudo que validou á hormona GDF15 como a responsable da hiperémese en mulleres embarazadas (Fejzo et al., 2023), mais en realidade este descubrimento partiu dun GWAS que apuntou a unha asociación cos xenes *GDF15* e *IGFBP7* (Fejzo et al., 2018). Curiosamente, nos últimos anos aumentou o número de GWAS sobre fenotipos relacionados co comportamento ou con factores socioeconómicos, a raíz do cal está collendo impulso o campo da socioxenómica (Braudt, 2018), se ben está suxeito a críticas. Por exemplo, o estudo sobre o uso de alcol e nicotina revelou asociacións nos xenes metabolizadores de nicotina (*CYP2A6*) e alcol (*AH1B*), este último cunha variante que acadou un p-valor de asociación de $2.2x10^{-303}$, ou no receptor da dopamina D2 (xen *DRD2*) (M. Liu et al., 2019). O fenotipo máis controvertido da socioxenómica é o nivel educativo, para o que o último GWAS publicado reportou 3.952 SNPs asociados, cuxa combinación explica entre o 12 e o 16% da variabilidade do carácter (Okbay et al., 2022). Porén, algúns investigadores consideran que estas asociacións son o resultado de outros factores confusores de tipo ambiental ou de estrutura

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

poboacional (Coop & Przeworski, 2022), mentres que outros cuestionan directamente a necesidade do estudo destes fenotipos (Comfort, 2018).

O maior avance nos últimos anos deuse no aumento do tamaño de mostra, de tal forma que xa se superou o millón de participantes (Evangelou et al., 2018; J. J. Lee et al., 2018), incluso chegando a 5.4 millóns no caso da altura (Yengo et al., 2022). É cada vez máis común que os grupos de investigación se reúnan en grandes consorcios internacionais aos que cada un aporta unhas mostras ou resultados dos seus estudos GWAS, que son meta-analizados a *posteriori* polo equipo de análise (**Figura 1**), seguindo un protocolo de control de calidade común (Winkler et al., 2014). Algúns deles son o PGC ou *Psychiatric Genomics Consortium* (Sullivan et al., 2018), responsable de moitos avances no coñecemento sobre xenética en psiquiatría; o GIANT (*Genetic Investigation of ANthropometric Traits*) (Wood et al., 2014), centrado en caracteres antropométricos como o BMI ou a altura; o consorcio CHARGE (*The Cohorts for Heart and Aging Research in Genomic Epidemiology)* (Psaty et al., 2009); ou o HGI (*COVID-19 Host Genetics Initiative*) (The COVID-19 Host Genetics Initiative, 2020), que en cuestión de dous anos conseguiu analizar máis de 200.000 casos e 3 millóns de controis, desvelando a base xenética de susceptibilidade á COVID-19.

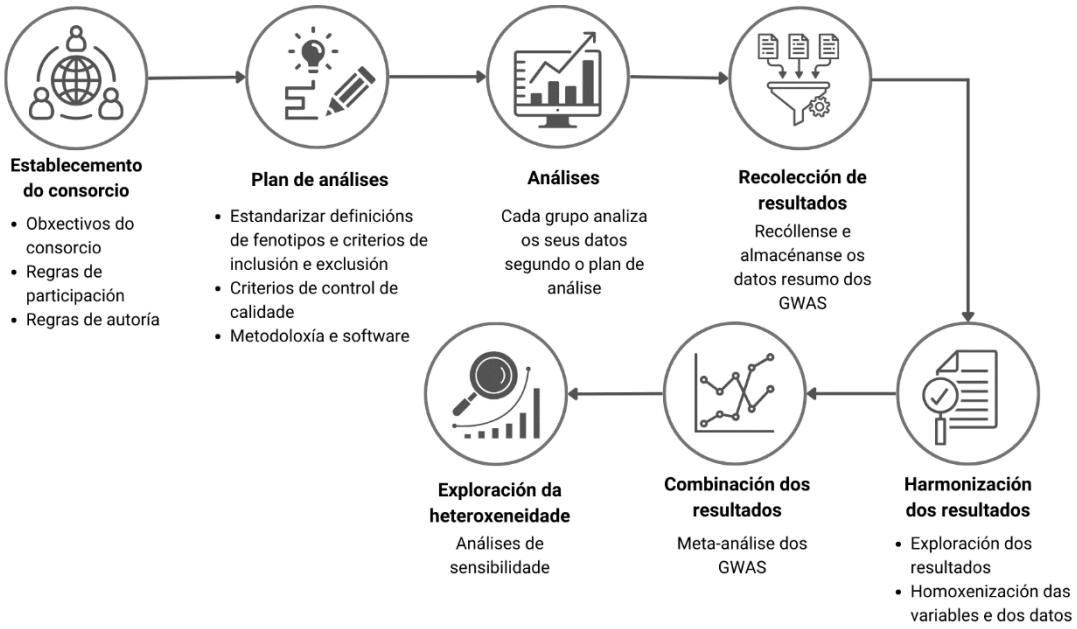


Figura 1. Esquema do establecemento dun Consorcio GWAS típico.

Por outro lado, a consolidación de grandes biobancos de mostras como o *UKBiobank* en Reino Unido (Sudlow et al., 2015); *FinnGen* en Finlandia (Kurki et al., 2023); o *Estonian Biobank* (Leitsalu et al., 2015); o *Mexican Biobank* (Sohail et al., 2023), ou o *Biobank Japan* (Nagai et al., 2017), facilitou a análise de multitude de fenotipos, xa que estes biobancos dispoñen de bases de datos con seguimento lonxitudinal que contan con información clínica e demográfica de centos de miles de persoas. No marco do *Pan UKBB*, por exemplo, que é un proxecto do *UKBiobank*, publicáronse un total de 16.131 GWAS para 7.228 fenotipos e os resultados son de acceso libre (https://pan.ukbb.broadinstitute.org/). Recentemente formouse

un consorcio de 23 biobancos mundiais, o *Global Biobank Meta-analysis initiative* (W. Zhou et al., 2022), que se axuda da harmonización de fenotipos a través dos códigos ICD (*International Classification of Diseases*) (Denny et al., 2013) e a estandarización dos fluxos de traballo para impulsar a potencia estatística dos GWAS e incrementar a diversidade nos estudos. Xunto coa liberación ao público dos *summary statistics* (estatísticos resumo), os biobancos e os consorcios son dous elementos chave na era de estudos GWAS a gran escala.

Da man disto último, as técnicas relacionadas coa xestión dos datos masivos ou *big data* facilitaron que o campo experimentara unha evolución pronunciada. As infraestruturas de almacenamento de datos fan posible obter, gardar e compartir grandes volumes de información xenética e clínica, polo que agora xa non é un problema tratar con millóns de mostras. Ademais, o procesamento destes datos nunhas velocidades asumibles require un custe computacional alto, pero na actualidade os grupos de investigación teñen acceso a clústeres de ordenadores que facilitan esta tarefa (Navarro et al., 2019). Para a integración dos datos nos GWAS desenvolvéronse software que son capaces de analizar centos de miles de individuos e millóns de SNPs en poucas horas, como por exemplo SAIGE (W. Zhou et al., 2018), fastGWA (Jiang et al., 2019) ou REGENIE (Mbatchou et al., 2021).

Tamén apareceu nova metodoloxía. Un exemplo son os GWAS multivariantes como os que implementa MTAG, que adapta o estimador da inversa da varianza das meta-análises, combinando estatísticos resumo para detectar *loci* pleiotrópicos (Turley et al., 2018). POLMM é un software paralelo a SAIGE que emprega un modelo de Odds proporcionais (Bi et al., 2021), e GATE, dos mesmos autores, axusta fenotipos de tempo-ata-evento (Dey et al., 2022). As análises condicionais como COJO permiten re-analizar os GWAS mediante unha regresión *stepwise* para identificar asociacións secundarias nun *locus* (J. Yang et al., 2012). Outra novidade é a introdución de métodos de *machine learning* para predicir fenotipos a partir de diferentes variables, unha práctica denominada "imputación de fenotipos" (Dahl et al., 2023; Miao et al., 2024). Ademais, os resultados dos GWAS poden empregarse para outro tipo de análises. Por exemplo, pódese predicir a expresión xénica e comprobar a asociación co fenotipo a través dos *Transcriptome Wide Association Studies* (TWAS), como no caso de S-PrediXcan (Barbeira et al., 2018) ou FUSION (Gusev et al., 2016), ou estudar as correlacións xenéticas entre fenotipos como fai LDSC (Bulik-Sullivan et al., 2015). O gran número de estudos GWAS propiciou o desenvolvemento dun novo tipo de estudo, o *Phenome Wide Association Study* (PheWAS), que revirte a metodoloxía para identificar fenotipos asociados a un SNP concreto (Bastarache et al., 2022). Apareceron incluso novas liñas exploratorias en epidemioloxía xenética, como son os índices de risco xenético e a aleatorización mendeliana (Loos, 2020).

O crecemento do campo reflíctese nos datos do *GWAS Catalog* (https://www.ebi.ac.uk/gwas/home), que en setembro de 2023 reportaba un total de 556.774 asociacións e 61.880 estudos GWAS.

1.2.1 Estudos de asociación con datos WGS e WES

As análises de asociación estendéronse na última década a datos de secuenciación (WGS, *whole genome sequencing*) e de exoma completo (WES, *whole exome sequencing*). Este tipo de estudos aumentan a cobertura e permiten identificar asociacións en calquera variante do

xenoma (recordemos que os *arrays* de xenotipado obteñen información de variantes *tag*) (Uffelmann et al., 2021), incluíndo variantes raras e de moi baixa frecuencia, as cales soen ter maiores efectos nos fenotipos. Outra vantaxe deste tipo de estudos, especialmente dos que involucran exomas, é que poden axudar a entender mellor a función biolóxica das asociacións, xa que cobren toda a parte codificante do xenoma (Backman et al., 2021). Por exemplo, o estudo de asociación con datos de WES en máis de 200.000 individuos do UKBiobank revelou que unha parte significativa da variabilidade nas enfermidades se debe a variantes raras que non serían captadas con datos de *array* (Q. Wang et al., 2021). Neste sentido, estas análises contribúen a diminuír o problema da herdabilidade perdida. Por outra banda, a agregación de variantes a nivel de xen permite identificar asociacións novas que non se atopan con variantes individuais.

Con todo, aínda que as análises de asociación con WES e WGS son cada vez máis frecuentes, cómpre valorar a relación custo/beneficio entre os datos de secuenciación e de *array*. Ao comparar os datos de *array* con imputación, WGS, WES e WES+*array*, a combinación de datos de exoma e de *array* resultou ser a máis poderosa tanto na cobertura como na identificación de sinais (Gaynor et al., 2023). Esta estratexia, que cobre variación non codificante mediante o *array* e variación codificante mediante o exoma, consigue identificar o mesmo número de asociacións que empregando datos de WGS (co mesmo tamaño de mostra). Ademais, tendo en conta que os custos asociados a secuenciar o xenoma completo son maiores, nas análises WES pódense incluír máis individuos que nas de WGS (Gaynor et al., 2023). Na mesma liña, a análise de 59 fenotipos sanguíneos a partir de datos WGS revelou que entre o 28% e o 54% das asociacións fenotipo-xen atopadas nas rexións non codificantes do xenoma xa foran previamente reportadas a través de GWAS ou WES, mentres que se acadaba un 99% de concordancia ao incluír xenes próximos (Ribeiro & Delaneau, 2023). Ademais, tras condicionar as asociacións por variantes comúns xa coñecidas, permaneceron menos do 10% das non codificantes e ao redor do 40% das codificantes. Por tanto, os GWAS clásicos continúan a ser unha ferramenta útil para detectar de forma rápida e masiva asociacións fenotipo-xen, se ben a análise de datos de exoma son un complemento eficaz de cara a identificar variación codificante e a aclarar as funcións biolóxicas das variantes.

1.2.2 Consideracións adicionais dos GWAS

Se ben nos inicios os GWAS parecían ser a folla de ruta a seguir de cara a descubrir todos os determinantes xenéticos dos caracteres complexos, o campo atopouse con numerosos obstáculos. En primeiro lugar, a polixenicidade é un factor que xoga na contra, xa que os estudos de herdabilidade en xeral estiman uns valores moito máis altos que a porcentaxe de variabilidade do fenotipo explicada polo descubrimentos GWAS (Visscher et al., 2017). A este fenómeno chamóuselle "herdabilidade perdida" (Maher, 2008; Manolio et al., 2009). A principal causa está moi relacionada co modelo infinitesimal de Fisher, pois os caracteres complexos están influenciados por miles de variantes con efectos pequenos, e o número de individuos que se poden incluír nestes estudos aínda non é suficiente para explicar toda a variabilidade dos caracteres, especialmente en fenotipos ou enfermidades menos comúns (**Figura 2**). O caso da altura é o único ata o momento no que se chegou a obter un número de

loci suficiente para explicar o 100% da herdabilidade estimada en poboación europea, un total de 7.209 SNPs asociados para os que se necesitou estudar a 5.4 millóns de individuos (Yengo et al., 2022).

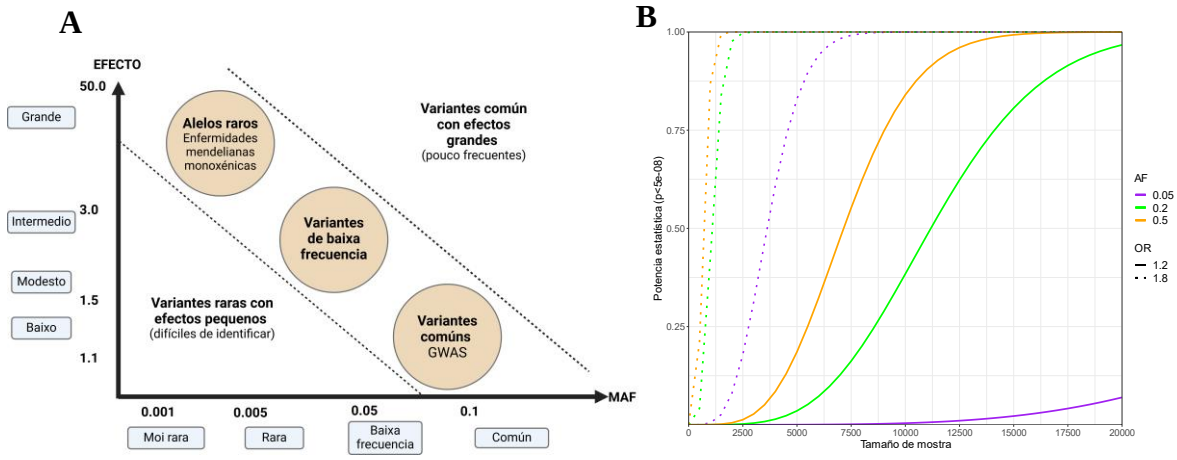


Figura 2. Relación entre a frecuencia alélica e tamaño de efecto nos GWAS. Nun estudo GWAS tende a identificarse a variación común, é dicir, variantes con frecuencias superiores ao 1% con tamaños de efecto modestos (panel A). Pola contra, as variantes cuxos efectos sobre o fenotipo son altamente deletéreos, están sometidas a selección negativa, reducindo a súa frecuencia na poboación e dificultando a súa identificación. O panel B ilustra a variación da potencia estatística dun GWAS en función de diferentes factores (tamaño de efecto, frecuencia alélica e tamaño de mostra), asumindo unha proporción 1:1 de casos/controis. O aumento do número de individuos facilita a detección de variantes con FA máis reducidas e OR baixas. O panel A desta figura está adaptada de Manolio e colaboradores (2009).

A cuestión da potencia estatística é fundamental para entender esta problemática. Esta depende de diferentes parámetros como a frecuencia do alelo de efecto, o tamaño de mostra e a magnitude do efecto do alelo sobre o fenotipo (M. I. McCarthy et al., 2008). Hoxe en día sábese que a maioría das asociacións atopadas nos GWAS teñen efectos moi pequenos porque son variantes non codificantes e están concentradas en rexións reguladoras (Maurano et al., 2012). Ao non ter un efecto directo e grande sobre o fenotipo, estabilízanse en frecuencias máis altas e requiren menos potencia estatística para detectalas. Pola contra, espérase que as variantes cun maior efecto sobre o fenotipo estean sometidas a selección e, por tanto, a súa frecuencia alélica sexa máis baixa. Como vemos na **Figura 2**, para detectar unha variante de baixa frecuencia (MAF<0.05) cunha OR (*odds-ratio*) de 1.2 non serían suficientes 20.000 individuos.

Os GWAS necesitan que os datos xenéticos pasen un control de calidade extenso, no que é particularmente importante explorar a presencia de estratificación ou estrutura poboacional para evitar falsos positivos e eliminar potenciais factores confusores. A estrutura poboacional maniféstase a través de diferencias nas frecuencias alélicas dentro dunha mesma poboación, provocadas por diversas forzas selectivas, que leva á formación de subgrupos (Astle & Balding, 2009; Devlin & Roeder, 1999). No caso dos estudos de asociación caso/control, isto é importante porque o seu éxito radica en que os casos teñan unha probabilidade de provir dunha poboación conxunta por enriba do azar, e a existencia de subgrupos pode alterar estas

condicións levando á aparición de asociacións falsas. Ademais, os modelos estatísticos que se empregan na súa maioría asumen unha independencia entre as observacións, e esta condición é difícil que se cumpra cando tratamos con datos xenéticos. Un dos aspectos a avaliar é a relación descoñecida (*cryptic relatedness*) entre individuos da mostra, é dicir, que exista parentesco entre os mesmos sen que o investigador o saiba, polo que os seus xenotipos non serán observacións independentes entre si (Voight & Pritchard, 2005). Existen moitas medidas para determinar o parentesco entre dous individuos, pero unha das máis utilizadas é o *Identity-By-Descent* (IBD, idéntico por descendencia), que se pode derivar a partir de datos de SNPs. O *software* PLINK emprega o método dos momentos para estimar a probabilidade de compartir 0, 1 ou 2 alelos por IBD para un par de individuos (Purcell et al., 2007), e despois calcula a proporción de IBD, que se compara cos valores esperados de IBD para diferentes graos de parentesco. Ademais das relacións de parentesco, tamén se avalía a homoxeneidade da mostra a través do cómputo das compoñentes principais xenéticas ou a través das estimas de ancestralidade (ver 1.3. *Ancestralidade xenética*). A análise de compoñentes principais (PCA) é unha técnica estatística que permite reducir centos de miles de SNPs a unhas poucas dimensións que recollen a maior variabilidade xenética posible. Os primeiros estudos que aplicaron esta técnica a datos xenéticos determinaron que existía unha correlación das compoñentes principais (PCs, de *principal components*) coa orixe xeográfica (Reich et al., 2008). Así, visualizar as primeiras PCs permite avaliar o grao de estratificación poboacional nas mostras e a corrección da mesma nos estudos de asociación (Price et al., 2010; Tian et al., 2008).

Outra limitación que existe nos estudos GWAS clásicos é a dificultade para identificar o efecto das interaccións xenéticas (xen-xen ou SNP-SNP) e as interaccións xenotipo-ambiente, os cales poden xerar problemas de replicación dos resultados ou unha heteroxeneidade nos efectos entre estudos. Por último, en moitos casos non é posible establecer unha relación directa entre o SNP da asociación e a bioloxía ou a etioloxía da enfermidade. Neste sentido, a integración dos resultados dos GWAS con información funcional, especialmente derivada das ómicas, é imprescindible para obter unha perspectiva máis completa. Se ben un estudo GWAS ofrece un punto de partida na exploración da base xenética dos caracteres complexos, require de análises posteriores para relacionar as variantes asociadas co xen causal ou cunha función (Tam et al., 2019). Este tema será explorado con máis profundidade na sección de Métodos.

1.3 ANCESTRALIDADE XENÉTICA

Nos estudos xenéticos, especialmente nos GWAS, é moi común empregar descritores poboacionais para clasificar aos individuos en grupos. Habitualmente escóllense factores como a etnia, criterios xeográficos ou as chamadas poboacións xenéticas (Panofsky & Bliss, 2017). Entre outras razóns, esta práctica dáse para controlar a estratificación poboacional, factores socioeconómicos ou para seleccionar conxuntos de mostras máis homoxéneos. Se ben o uso da ancestralidade xenética parece substituír os criterios raciais por criterios biolóxicos, as categorías discretas nas que se organiza non deixan en certo modo de corresponderse coas razas clásicas (A. C. F. Lewis, Molina, et al., 2022).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Definir que é exactamente a ancestralidade xenética é complexo. Teoricamente se podería dicir que é a cantidade de material xenético que un individuo herda dos seus ancestros, pero de que ancestros e cales son esas poboacións ancestrais? Mathieson e Scally (2020) diferencian tres termos: ancestralidade xenealóxica, ancestralidade xenética e similitude xenética. A ancestralidade xenealóxica refírese ao *pedigree*, é dicir, á arbore de ancestros con relacións parentais que dá lugar a un individuo xunto coas características culturais e xeográficas do mesmo. Por outra parte, a ancestralidade xenética defínese como a cantidade de ADN herdada dos ancestros xenealóxicos por diferentes camiños do pedigree, e está representada polos chamados gráficos de recombinación ancestrais, nos cales os nodos representan eventos de recombinación ou coalescencia en vez de lazos familiares (Mathieson & Scally, 2020). Finalmente, os autores argumentan que o que comunmente se nomea "ancestralidade" é en realidade similitude xenética entre poboacións e individuos (Mathieson & Scally, 2020).

Certamente, a estimación das compoñentes de ancestralidade xenética parte de modelos estatísticos que empregan uns individuos de referencia, habitualmente representativos de rexións continentais (Coop, 2023). Os paneis de referencia habituais, como *1000 Genomes Populations* ou 1KGP (Auton et al., 2015) ou *HapMap* (Gibbs et al., 2003), empregan etiquetas referidas aos continentes (EUR-ancestralidade europea, EAS-ancestralidade do Este Asiático, AFR-ancestralidade africana, AMR-ancestralidade americana mixta, SAS-ancestralidade do Sur Asiático). A pesar de que durante a mostraxe se procura recoller individuos representativos de ditas poboacións e sub-poboacións, non hai certeza de que eses individuos sexan representativos dos ancestros que habitaban esas rexións nin moito menos de que existiran grupos poboacionais discretos. A variabilidade xenética humana segue uns patróns continuos, causados por diversos eventos demográficos e evolutivos, e esta variabilidade é moito maior entre individuos que entre as poboacións (Novembre, 2022). De feito, o nivel de fragmentación da mostra en grupos é tamén unha elección do investigador, podendo incluso clasificar aos individuos en función de patróns xenéticos moi finos (Coop, 2023; Novembre & Peter, 2016).

Mediante a técnica de PCA, por exemplo, podemos asignar cada individuo a un grupo en función da proximidade nas coordenadas das PCs aos grupos de referencia. Os criterios de definición dos grupos son igualmente aleatorios. Isto en realidade só nos asegura que eses individuos son máis parecidos entre si e, por tanto, máis homoxéneos. Ademais, hai que ter en conta que a técnica de PCA maximiza a variabilidade nas súas compoñentes, é dicir, resalta de forma intrínseca as diferencias. Pola contra, o que non significa é que os ancestros dun grupo de individuos similares pertenceran a unha poboación discreta no pasado (Coop, 2023). Do mesmo xeito, os algoritmos tipo ADMIXTURE, que empregan modelos matemáticos e demográficos máis complexos, tamén dependen do uso dunhas poboacións de referencia.

Aínda que a creación de grupos xenéticos discretos non represente a historia demográfica das poboacións, non deixa de ser útil na práctica. GnomAD, por exemplo, xustifica o uso da ancestralidade de cara a facilitar a clasificación de variantes, pero aclara a artificialidade das categorías usadas (https://gnomad.broadinstitute.org/news/2023-11-genetic-ancestry/). A *National Academies of Sciences, Engineering and Medicine* publicou en 2023 un estudo extenso sobre esta problemática, ofrecendo unha serie de recomendacións, entre as que se atopan algunhas como: (a) seleccionar o descritor dependendo dos obxectivos do estudo e

explicar correctamente a súa definición, (b) evitar entender os grupos en termos de xerarquía ou homoxeneidade, (c) colaborar con investigadores e individuos das comunidades das que se obteñen os datos, ou (d) aplicar de forma consistente os mesmos descritores a todos os individuos. Para os estudos de enfermidades complexas, recomendan empregar como medida a similitude xenética.

Este matiz é moi importante, sobre todo de cara a comunicar os resultados dos estudos e non caer en interpretacións biolóxicas erradas (Kaiser, 2023). Os estudos xenéticos foron empregados para xustificar o racismo durante décadas e, a pesar de que estas supostas diferencias biolóxicas foron constantemente desmentidas, continúan a ser manipulados e empregados como argumentos "científicos" dentro de grupos supremacistas (J. Carlson et al., 2022; Chou, 2017). As persoas investigadoras temos unha parte de responsabilidade no que respecta á comunicación dos resultados para evitar na medida do posible potenciais malos usos. Por tanto, cómpre aclarar que ao longo desta tese referirémonos como *Genetically Inferred Ancestry* (GIA), ou ancestralidade xenética inferida, á similitude computada por métodos estatísticos dos individuos das nosas cohortes con respecto a uns individuos de referencia, clasificados en grupos continentais por simplicidade.

1.3.1 *Admixture* xenético

O *admixture* xenético é o proceso mediante o cal o ADN de dous ou máis grupos "fonte", idealmente aillados, mestúranse e dan lugar a un mosaico xenético (Korunes & Goldberg, 2021; Suarez-Pajes et al., 2021). Isto prodúcese a partir da interacción de grupos poboacionais na historia demográfica da humanidade, sendo unha das principais causas da variabilidade xenética. A mestura de dúas ou máis fontes xenéticas provoca cambios nos patróns de desequilibrio de ligamento, de forma que dous *loci* que non están ligados poden presentar correlación tras o evento de mestura se as frecuencias alélicas difiren nos grupos fonte (Slatkin, 2008). Ademais, tamén provoca cambios nas frecuencias alélicas dalgúns marcadores ao combinarse as dos grupos iniciais.

Os individuos pertencentes ás poboacións mixtas[1] presentan tractos ao longo dos cromosomas con diferente orixe ancestral. Isto ocorre porque a recombinación rompe os segmentos procedentes dunha única poboación. Unha forma ilustrativa de entendelo é a través do método de Lawson e colaboradores (2012), no cal modelan un haplotipo como recipiente de material xenético dun doante. Así, desenvolven o concepto de "pintar cromosomas": un cromosoma "píntase" co material xenético dos doantes, e cada pintura representa unha fonte ancestral (Lawson et al., 2012). Destes mosaicos cromosómicos xorden entón dous conceptos de ancestralidade: a ancestralidade global, definida como a fracción total do xenoma de cada individuo que provén dun grupo fonte, e a ancestralidade local (AL), referida á orixe dos diferentes segmentos dos cromosomas ou marcadores xenéticos (**Figura 3**). Ambas poden ser estimadas con un número pequeno de marcadores. Estes individuos presentan, ademais, patróns

[1] Referirémonos como *poboación mixta* a aquelas con ascendencia "*admixed*".

de desequilibrio de ligamento diferentes: pode introducirse unha correlación entre segmentos que son independentes nas poboacións das que derivan (Tan & Atkinson, 2023).

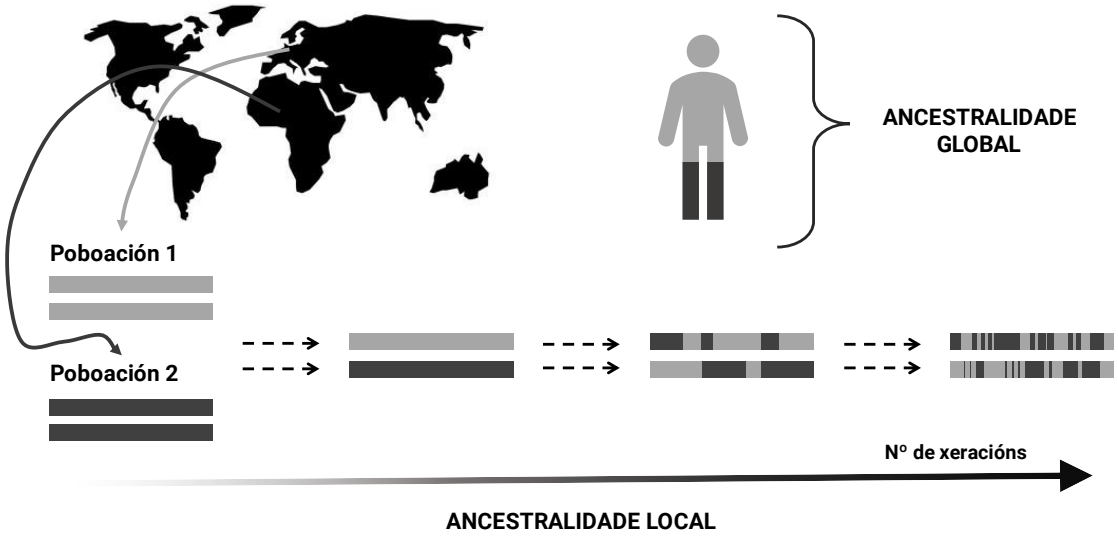


Figura 3. Ancestralidade global e ancestralidade local. A interacción de dúas poboacións (evento de "mestura") produce un fluxo xenético entre elas, de forma que se crean mosaicos cromosómicos que conteñen fragmentos das fontes ancestrais orixinais. Esta interacción resulta nun incremento da diversidade xenética dos individuos (Seldin et al., 2011). Cada un destes fragmentos provén dunha das fontes ancestrais, denominándose ancestralidade local. Por outro lado, a ancestralidade global fai referencia á proporción de cada fonte ancestral que un individuo presenta no total do seu xenoma, polo que se pode computar a partir das ancestralidades locais.

O estudo das poboacións mixtas a nivel xenético, a través dos patróns de ancestralidade global e local, facilitou a detección de moitos eventos espazo-temporais, permitindo a reconstrución da historia evolutiva das poboacións humanas e a confirmación de eventos históricos (Hellenthal et al., 2014; Korunes & Goldberg, 2021). Porén, que é o que se considera como poboación mixta non está claro, posto que todos os grupos poboacionais se poden definir como unha mestura das poboacións das que derivan, resultando nun debate que non está resolto (Gopalan et al., 2022). Algúns autores argumentan que se ben o concepto de "*admixture*" en xenética de poboacións non trae problemas, a súa aplicación aos individuos asume intrínsecamente a existencia de poboacións puras e debería ser substituído (Kampourakis & Peterson, 2023). A maioría dos métodos que modelan os eventos de mestura xenética, como por exemplo STRUCTURE (Pritchard et al., 2000) ou ADMIXTURE (Alexander et al., 2009), empregan modelos teóricos complexos, que tratan de ter en conta diferentes parámetros evolutivos e demográficos. En cambio, na práctica seguen dependendo da definición de grupos discretos como poboacións fonte. Nos estudos de xenética de poboacións ou evolutiva poden terse en conta aspectos máis complexos, como os gráficos de recombinación ancestral, pero nos estudos de xenética médica este tipo de análises simplifícanse.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Se retrocedemos ás primeiras migracións humanas, podemos falar de poboacións relativamente illadas que foron sometidas a eventos de mestura. Por exemplo, sábese que as poboacións europeas son unha mestura de polo menos tres fontes xenéticas (cazadores-recolectores europeos do Paleolítico, granxeiros do Neolítico e das migracións do Neolítico tardío e da Idade de Bronce) (Nielsen et al., 2017). Ademais, Skoglund e Mathieson (2018) apuntan que a diversidade xenética actual se debe a eventos de mestura durante o Holoceno. Con todo, nos estudos xenéticos non se soen considerar aos individuos europeos como parte dunha poboación mixta. En contraposición, a poboación Latinoamericana é considerada unha poboación con mestura recente de tres compoñentes, dadas as migracións dende Europa e África a rexións onde residían as poboacións Nativo-Americanas pre-Colombinas (Bryc et al., 2010; Ruiz-Linares et al., 2014). A distribución das proporcións de GIA presenta un patrón xeográfico e non é homoxénea en todos os países de Latino-América: a compoñente africana está máis presente en Brasil e en rexións costeiras de Colombia; a compoñente nativo-americana é alta ao longo de Perú, Chile, e no sur de México, e a compoñente europea é maior no sur de Brasil, norte de México e zonas urbanas de Chile (Ruiz-Linares et al., 2014).

Como xa se explicou no apartado anterior, as poboacións de referencia que se empregan para a estimación das proporcións de ancestralidade son contemporáneas, e elas mesmas estiveron nalgún punto sometidas a procesos de mestura. As poboacións latinoamericanas presentan mestura da compoñente ancestral europea, pero isto está limitado dende o momento no que non podemos determinar se os individuos de referencia son 100% europeos, ou mellor dito, que significa ser un "100% europeo". En relación a isto último, o punto chave pode que se atope en definir as poboacións de referencia con respecto a un momento temporal, é dicir, definir mestura entre "individuos que se atopaban en X lugar hai X anos", e sempre en termos de similitude (Coop, 2023; Kampourakis & Peterson, 2023).

1.3.2 Ancestralidade e poboacións mixtas nos estudos GWAS

A ancestralidade é unha fonte de confusión nos GWAS. Por exemplo, nun estudo de tipo caso/control podemos obter asociacións falsas se unha poboación está máis representada nos casos e existe unha variante nela cunha frecuencia alélica máis alta. Por outro lado, as diferencias que existen no LD entre as poboacións causa, en moitas ocasións, unha heteroxeneidade nos efectos estimados entre os diferentes grupos de estudo. Así, unha variante pode estar fortemente correlacionada coa variante causal nunha poboación, pero non estalo noutra, o que resulta en asociacións diferentes e non replicables. Os protocolos clásicos de control de calidade asumen unha poboación homoxénea, polo que en presencia de individuos de orixes xenéticas diversas deben axustarse algúns pasos, especialmente aqueles que dependen das frecuencias alélicas, como a exploración da heterocigosidade e a selección dos SNPs que non se atopan en equilibrio Hardy-Weinberg, ou o paso de detección e eliminación de *outliers* (datos atípicos) a partir da análise PCA. Con respecto ao propio GWAS, o axuste dos modelos de regresión pola ancestralidade global corrixe en gran medida a inflación dos erros tipo I derivados deste tipo de estrutura poboacional (Price et al., 2010; Tian et al., 2008). Este axuste pódese facer a través das PCs, aínda que non está establecido cal é o número óptimo de compoñentes a incluír nos estudos, ou das proporcións de ancestralidade (C.-C. Liu et al., 2020).

A corrección da estratificación complícase cos individuos mixtos pola AL. Isto engade unha dificultade ao seu estudo, posto que un *hit* GWAS pode aparecer a moita distancia da variante causal como consecuencia das estruturas de LD derivadas do evento de mestura (o que xeraría un falso positivo). Ademais, as FAs dun mesmo SNP poden variar entre tractos ancestrais, polo que os estudos de asociación neste tipo de poboacións requiren un control maior da estrutura poboacional. Por exemplo, un SNP que parecía asociar o sistema da inmunoglobulina Gm coa diabetes tipo II en Nativo-Americanos resultou ser unha asociación derivada das diferencias nas frecuencias alélicas entre as ancestralidades Nativo-Americana e Europea (Knowler et al., 1988). Neste estudo, a corrección mediante ancestralidade global foi suficiente para facer desaparecer a asociación, pero non sempre é así. A dificultade para controlar a estrutura poboacional e para dividir estas cohortes en grupos homoxéneos leva á súa exclusión sistemática dos estudos GWAS clásicos, reducindo os tamaños de mostra en poboacións diversas[2] e limitando o número de GWAS realizados en poboacións como a latinoamericana.

Para o estudo das poboacións mixtas desenvolvéronse metodoloxías específicas como o *admixture mapping* (Rosenberg et al., 2010), que consiste nun test de asociación entre o fenotipo e segmentos de ancestralidade local en vez de SNPs. Por exemplo, a compoñente Nativo-Americana asociouse con dúas rexións de risco no cromosoma 3 para a enfermidade de Alzheimer (Horimoto et al., 2021, 2023), e cun *locus* no cromosoma 6 para a enfermidade de Parkinson (Loesch et al., 2021). Este tipo de análises supuxo unha alternativa aos GWAS para detectar asociacións en poboacións mixtas (Suarez-Pajes et al., 2021), pero teñen limitacións como o cálculo da ancestralidade local ou que só detecta aquelas variantes con variabilidade nas FA entre os segmentos ancestrais (Tan & Atkinson, 2023). Ademais, este método emprega segmentos cromosómicos e non posicións concretas, facendo máis complicada a identificación das variantes causais.

1.4 DIVERSIDADE NOS ESTUDOS GWAS

Nos últimos anos cobrou importancia a problemática da falta da diversidade nos estudos GWAS, e aínda que existe un compromiso firme por resolvela, aínda non se reduciu a diferencia de números (Bustamante et al., 2011; Popejoy & Fullerton, 2016; Sirugo et al., 2019). Os GWAS están aplicados maioritariamente a poboacións europeas, que engloban aqueles individuos caucásicos ou brancos. Moitos dos estudos GWAS están publicados en UK, EE.UU. ou Islandia, onde a poboación máis numerosa é *branca* (Mills & Rahal, 2020). Ademais, aínda que é evidente que nestes países conviven diferentes etnias, prodúcense nesgos no recrutamento das mostras cunha tendencia a non representar a toda a poboación. En 2009, 320 de 370 GWAS publicados incluían unicamente individuos europeos (Need & Goldstein, 2009). Se ben é certo que se detectou un aumento do número de individuos non-europeos, Popejoy e Fullerton (2016) indicaron que dese crecemento, o 78% correspondía a poboacións asiáticas ou indias. En outubro de 2023, o *GWAS Diversity Monitor* (Mills & Rahal, 2020)

[2] O concepto *poboacións diversas* refírese, nesta tese, ás poboacións non identificadas como "EUR".

(https://gwasdiversitymonitor.com/) reportaba que un 94.77% dos participantes eran de ascendencia europea; un 3.56% de ascendencia asiática; un 0.18% de ascendencia africana; un 0.48% de ascendencia africano-americano ou afro-caribeña, e o 0.33% de ascendencia hispana ou latinoamericana. Na **Figura 4** obsérvase a tendencia por anos. A pesar de que o número de estudos en poboación asiática supón agora o 29% do total, o número de participantes continúa a ser considerablemente menor. Pola contra, para a poboación Latinoamericana pode observarse unha diminución na porcentaxe de participantes dende aproximadamente 2016 (**Figura 4**). A cohorte de *The Hispanic Community Health Study/Study of Latinos* (HCSH/SOL), recollida en EE.UU., é unha das máis estudadas, dando lugar a máis de 500 publicacións (ver https://sites.cscc.unc.edu/hchs/publications-in-print) (LaVange et al., 2010).

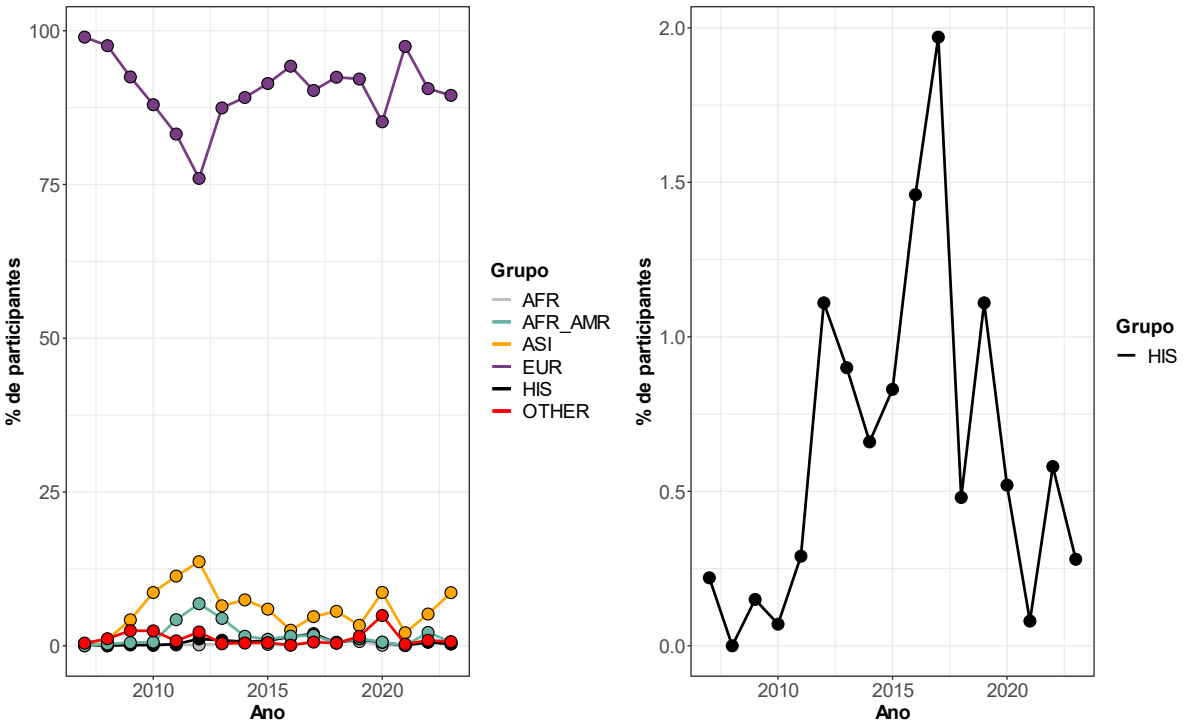


Figura 4. Diversidade nos estudos GWAS dende 2007 ata 2024. Á esquerda, gráfico representando a serie temporal das porcentaxes de inclusión de diferentes grupos étnicos en estudos GWAS ao longo dos anos, calculadas como o número de individuos de cada grupo con respecto ao total anual. Á dereita, gráfico representando a serie temporal das porcentaxes de inclusión de individuos Hispanos ou Latinoamericanos ao longo dos anos. Datos recuperados o día 22/01/2024 no *GWAS Diversity Monitor* (Mills e Rahal, 2020). Este monitor en tempo real considera un total de 6586 estudos incluídos no *GWAS Catalog*.

A falta de representación de todo o espectro da variabilidade xenética mundial nos estudos xenéticos causa desigualdades na intervención clínica. Concretamente, na aplicación da medicina personalizada, as comunidades máis marxinadas non son igualmente beneficiadas (Edwards et al., 2020; Martin et al., 2017, 2019). En Sirugo e colaboradores (2019) descríbense algúns exemplos de enfermidades cuxa prevalencia difire entre poboacións. Unha delas é a fibrose cística, unha enfermidade rara en africanos-americanos cuxo alelo causal no xen *CFTR*

non é o mesmo que en poboación europea. A diabetes tipo II é máis frecuente en individuos Mexicanos ou Latinoamericanos que en brancos non hispánicos (Morales et al., 2020), e algúns estudos atoparon unha correlación positiva entre as compoñentes xenéticas nativo-americana e africana co risco polixénico a esta enfermidade (Chande et al., 2020). Ademais, en individuos Mexicanos descubriuse un haplotipo entre os xenes *SLC16A11* e *SLC16A13*, asociado a unha redución de dous anos na diagnose da diabetes tipo II que é raro en poboacións europeas e africanas, pero que foi replicado en poboacións do Este de Asia (Williams et al., 2014). Os autores do estudo estimaron que o SNP podía explicar un 20% da diferencia nas prevalencias da enfermidade. Unha mutación substitutiva (*missense*) do xen *ALDH2*, presente en individuos de Asia do Este nunha frecuencia moi superior á do resto do mundo, provoca sensibilidade á inxestión de alcol aumenta o risco a padecer cancros relacionados co alcolismo (Chang et al., 2023).

Os beneficios de realizar estudos xenómicos en grupos non europeos son moitos, dende a detección de variantes novas compartidas entre poboacións ata a detección de variantes específicas de poboación. Unha variante causante dunha patoloxía pode atoparse en maior frecuencia nunha certa poboación, como por exemplo ocorre cunha mutación de perda de función no xen *PCSK9*, máis frecuente en individuos africanos e asociada coa redución dos niveis de LDL-C en sangue (Cohen et al., 2005). Isto é particularmente evidente nos caracteres morfolóxicos, que mostran unha variabilidade xenética considerable entre rexións xeográficas (Race, Ethnicity, and Genetics Working Group, 2005; Skoglund & Mathieson, 2018), ou naqueles relacionados co sistema inmune. Os patóxenos son unha das forzas selectivas máis importantes, incrementando a variabilidade xenética dos individuos (Fumagalli et al., 2011; Kwok et al., 2021). Por exemplo, algúns xenes expresados en macrófagos presentan unha expresión diferencial entre individuos con ascendencia africana e ascendencia europea en resposta á unha infección bacteriana (Nédélec et al., 2016). Os autores do estudo destacaron que os xenes involucrados na resposta inflamatoria expresábanse máis en células de individuos africanos, e postularon que podería ser unha consecuencia da selección natural actuando sobre as variantes reguladoras da transcrición. Curiosamente, entre eles estaba o xen *OAS1*, implicado na COVID-19. Estudando aos Inuit de Groenlandia descubriuse un set de SNPs asociado á altura que non fora descuberto previamente en GWAS europeos por cuestións de frecuencia alélica (Fumagalli et al., 2015): aínda que non se atopou evidencia dunha diferencia no tamaño do efecto entre ambas poboacións, estes SNPs eran de baixa frecuencia en europeos e os GWAS non tiñan suficiente potencia para detectalos (Fumagalli et al., 2015). Os xenes *TCF7L2* e *KCNQ1* asociados á diabetes tipo II tamén presentan diferencias nas frecuencias alélicas: a identificación dos SNPs de *TCF7L2* require dun maior número de individuos en poboacións do este asiático que de Europa, mentres que coas variantes de *KCNQ1* ocorre ao revés (Rosenberg et al., 2010). Na mesma liña, determinouse a existencia de 15 *loci* asociados aos lípidos específicos de poboación africana, a pesar de que a correlación dos efectos a escala xenómica entre poboacións é alta (Graham et al., 2021). Outro estudo identificou 28 variantes asociadas ao número de diversas células sanguíneas non atopadas en poboación europea (M.-H. Chen et al., 2020). Incluso dentro dunha superpoboación como a asiática existe variabilidade xenética

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

entre diferentes grupos, ocasionando diferencias na prevalencia de enfermidades e na resposta a fármacos (Chan et al., 2022).

Está demostrado que combinar grupos poboacionais en forma de meta-análises multi-poboación impulsa a potencia estatística á hora de identificar variantes con efectos en todas as poboacións e permite a replicación de moitos sinais, aumentando a proporción de variabilidade explicada por factores xenéticos (Y. R. Li & Keating, 2014). Ademais, tamén se aproveitan os diferentes patróns de desequilibrio de ligamento para atopar as variantes causais. Aínda que as variantes non causais teñan efectos estimados diferentes, espérase que as causais compartan o tamaño do efecto, facilitando así a súa identificación (Mägi et al., 2017; Peterson et al., 2019). Isto é importante porque unha variante causal pode ser máis fácil de determinar nunha poboación con segmentos de LD máis pequenos que nunha con segmentos máis grandes, como pode ser a europea (Rosenberg et al., 2010). Ademais, para cada poboación podemos obter un grupo distinto de variantes correlacionadas coa causal (compartida), polo que ao combinar as poboacións, cada unha cos seus patróns de LD, é probable que se reduza o número de variantes solapadas (LaPierre et al., 2021; Y. R. Li & Keating, 2014). Así, o mapeamento fino -que é unha técnica para detectar variantes causais nun sinal GWAS- benefíciase das aproximacións multi-poboacionais. Porén, a pesar de ser a práctica máis frecuente na actualidade, neste tipo de estudos o peso seguen a levalo as cohortes europeas.

No que respecta ás poboacións mixtas, na actualidade non existe consenso sobre como analizalas, pero nos últimos anos creceu o interese nelas, sobre todo da man dos avances na medicina personalizada. Recentemente se propuxo un novo modelo de GWAS (*software* Tractor) onde se teñen en conta as ancestralidades locais dos individuos e se modela a heteroxeneidade dos efectos entre elas (Atkinson et al., 2021). Porén, hai evidencia de que, en ausencia de dita heteroxeneidade, o axuste da ancestralidade local diminúe a potencia estatística (K. Hou et al., 2021; Mester et al., 2023). Por outro lado, parece que os modelos mixtos son un método eficiente para analizalas (Medina-Gomez et al., 2015; Tan & Atkinson, 2023; Wojcik et al., 2019), e estudos recentes demostraron que a corrección por ancestralidade global é suficiente para controlar a inflación do erro tipo I (K. Hou et al., 2021; Wojcik et al., 2019). Ademais, as poboacións mixtas tamén proporcionan vantaxes sobre as homoxéneas á hora de estudar a arquitectura xenética dos caracteres complexos. Os SNPs con frecuencias alélicas baixas nunha compoñente ancestral ven as súas frecuencias aumentadas ao combinarse con outras compoñentes ancestrais, o que facilita a detección de variantes máis raras ou incluso monomórficas. Algúns estudos dos últimos anos demostraron tamén que os GWAS en poboacións mixtas teñen maior potencia estatística, tanto para identificar variantes con efectos similares como variantes con heteroxeneidade nos efectos, mantendo a taxa de falsos positivos ao 5% (K. Hou et al., 2021; Lin et al., 2021). De cara á introdución da medicina personalizada na práctica clínica, a exclusión de individuos en base a criterios como os grupos poboacionais homoxéneos non ten viabilidade práctica, dado que non existen criterios obxectivos, e limita a intervención en individuos non brancos, converténdose nunha fonte de desigualdade (Martschenko et al., 2023).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1.4.1 Correlación dos efectos causais entre grupos poboacionais

Existen problemas para xeneralizar algúns resultados de poboacións europeas a outras poboacións, especialmente nas predicións de risco polixénico. Un mesmo índice de risco xenético pode non explicar a mesma porcentaxe de variabilidade fenotípica nunha poboación que noutra, e este efecto está directamente relacionado coa distancia xenética entre elas (Martin et al., 2017). Dado que isto ocorre de forma sistemática para a maioría dos fenotipos, cabería pensar que as variantes causais non foran compartidas.

Porén, as dificultades na traslación non parecen deberse a que sexan diferentes os xenes ou as variantes implicadas no fenotipo. Por un lado, moitos sinais descubertos en poboacións europeas teñen a mesma dirección e tamaño de efecto similar a outros grupos (C. S. Carlson et al., 2013; Marigorta & Navarro, 2013; Visscher et al., 2017), algo que tamén se evidencia nas meta-análises multi-poboación. Por outro, as correlacións xenéticas entre poboacións continentais indican que estas comparten gran parte das variantes causais, se ben existen rexións específicas de poboación (M.-H. Chen et al., 2020; Shi et al., 2020). Estas observacións estanse confirmando grazas ao estudo de poboacións con ascendencia mixta: o seu mosaico xenético é moi útil e permiten mitigar o efecto do ambiente. Ao comparar a base xenética dun fenotipo entre dúas poboacións empregando GWAS separados, introdúcese unha variabilidade asociada ao ambiente que non se pode eliminar. En contraste, dentro dunha mesma poboación mixta espérase que os individuos compartan a maioría dos efectos ambientais, á vez que proporcionan información xenética para ambas compoñentes ancestrais. K. Hou, Ding e colaboradores (2023) desenvolveron un método que calcula as correlacións xenéticas entre ancestralidades a través dos segmentos de ancestralidade local. Tras aplicalo a 38 fenotipos en 3 cohortes mixtas, determinaron que os efectos causais estaban altamente correlacionados entre as ancestralidades (cunha media de r_admix=0.95), e nos casos nos que non era así, as diferencias podían ser explicadas por variantes próximas (K. Hou, Ding, et al., 2023). Ademais, demostraron que as correlacións calculadas coas ancestralidades locais eran maiores que as obtidas comparando grupos continentais, suxerindo que os factores ambientais poden confundir os resultados. Outro estudo chegou a conclusións similares empregando xenomas mixtos (S. Hu et al., 2023), pero a maiores descartou o papel das interaccións xenotipo-ambiente. Parece, entón, que o desequilibrio de ligamento e as frecuencias alélicas son os principais causantes das diferencias xenéticas asociadas aos fenotipos entre as poboacións.

1.5 ÍNDICES DE RISCO POLIXÉNICO

O efecto de cada variante individual dun GWAS é pequeno e normalmente aporta pouco por si só á variabilidade do carácter. En cambio, se combinamos os efectos das variantes nun índice cuantitativo podemos obter unha puntuación individual que resuma o risco xenético para ese carácter. A isto último chámaselle índice de risco polixénico, ou *polygenic risk score* (PRS). Estes índices pódense empregar para a cuantificación do risco xenético e a estratificación da poboación, para testar a asociación do mesmo cun fenotipo, como covariable en modelos clínicos, ou incluso para comparar o risco xenético de dous fenotipos diferentes (S. W. Choi et

USC
DE COMPOSTELA

al., 2020). Ademais, esta simplificación facilita a análise conxunta dos factores xenéticos e outras variables medio-ambientais. A primeira predición polixénica dun fenotipo foi coa esquizofrenia, para a que se conseguiu predicir un 3% da variabilidade a través de SNPs (Purcell et al., 2009). Ademais, esta aproximación permitiu determinar unha base xenética conxunta entre a esquizofrenia e o trastorno bipolar (Purcell et al., 2009).

1.5.1 Construción dos PRS

Para a construción dun PRS precísanse como mínimo dous sets de datos. O primeiro consiste nos datos de descubrimento, ou o que é o mesmo, estatísticos resumo para un fenotipo obtidos a través dun GWAS ou un meta-GWAS. O segundo set de datos son os xenotipos da cohorte diana, é dicir, a cohorte onde se quere aplicar o PRS. É imprescindible que as dúas cohortes sexan independentes entre si e que non exista solapamento entre as mostras, ou da contra obteranse predicións demasiado optimistas (prodúcese un sobre-axuste do modelo) (Wray et al., 2013). Canto mellor sexa a estimación do efecto dos SNPs no GWAS, máis fiable será a predición, polo que é importante seleccionar como datos de descubrimento o GWAS con maior potencia estatística. Igualmente, a calidade dos SNPs tanto xenotipados como imputados na mostra diana é crucial. Existen varios protocolos publicados para a construción e aplicación dos PRS (S. W. Choi et al., 2020; Collister et al., 2022). Para cada individuo *i* e conxunto de SNPs *N=1...j*, sendo X_{ij} o número de alelos (0, 1, 2) para o SNP *j*, o PRS toma a seguinte forma:

$$PRS_i = \sum_j^N \widehat{\beta}_j X_{ij}.$$

Seguindo o modelo polixénico, a distribución dos PRS achegarase a unha distribución normal a medida que aumenta o tamaño de mostra. Se ben existen diferentes métodos para o cálculo dos pesos β_j, algúns dos máis empregados son o *clumping + thresholding* (C+T), para o que se pode empregar PLINK pero que se atopa integrado en PRSice (Euesden et al., 2015); LDpred2 (Privé et al., 2021); PRS-CS (Ge et al., 2019); MegaPRS (Q. Zhang et al., 2021), ou SBayesR (Lloyd-Jones et al., 2019). O número de variantes incluídas depende do investigador, podendo calcular o PRS unicamente a partir de SNPs con significación GWAS (habitualmente chamados *genetic risk scores* ou GRS) ou estendéndoo a todo o xenoma (PRS a escala xenómica). Curiosamente, os PRS que empregan límites de p-valor máis altos son, en moitas ocasións, mellores preditores que aqueles que empregan só *hits* (Mars et al., 2022; Purcell et al., 2009), o que suxire que existe unha base xenética dos caracteres que aínda non é capturada polos GWAS. Os diferentes software difiren principalmente no modelo empregado na estimación, entre os que se inclúen técnicas baiesianas e de aprendizaxe automático. Por exemplo, o modelo máis simple C+T (que se describirá na sección de Métodos), emprega un panel de referencia para escoller SNPs en función dos bloques de desequilibro de ligamento e o p-valor das variantes no GWAS, establecendo un peso 0 para as variantes non seleccionadas e un peso igual ao coeficiente do SNP no GWAS para as si seleccionadas. En cambio, o software MegaPRS axusta 8 modelos diferentes e permite escoller un modelo de herdabilidade para o fenotipo (Q. Zhang et al., 2021). Outros desvíanse dos PRS clásicos, como PRSet (S. W. Choi et al., 2023), que constrúe índices xenéticos baseados en rutas biolóxicas, ou o método Multi-

USC DE SANTIAGO DE COMPOSTELA

PGS (Albiñana et al., 2023), que combina de forma agnóstica librerías de PRS. Dada a velocidade coa que se desenvolven estes métodos, non existe un consenso sobre cal é o mellor. Ni e colaboradores (2021) comparan dez aproximacións distintas aplicadas a carácteres relacionados coa psiquiatría, e conclúen que LDpred2, MegaPRS e SBayesR son os que mellores resultados acadan. Outro estudo chegou a conclusións similares, destacando LDpred2, PRS-CS e lassosum (Pain et al., 2021). Algúns dos métodos mencionados complican a súa aplicación na fase de validación, pois necesitan dunha mostra externa para afinar os hiper-parámetros do modelo que sexa diferente da cohorte diana. Outros métodos empregan pseudo-validación a través dos estatísticos resumo e non requiren dita mostra adicional, como PRS-CS.

1.5.2 Aplicacións dos PRS

Tras a construción do PRS, este aplícase aos individuos a través de *software* como PLINK (Purcell et al., 2007), obtendo unha puntuación de risco para cada un deles. As análises máis comúns son a de asociación e a de predición (Dudbridge, 2013). Mediante a análise de asociación podemos determinar se os SNPs incluídos no índice capturan correctamente a polixenicidade do fenotipo. Se os fenotipos do PRS e da cohorte diana son diferentes, a asociación serve para explorar unha posible base xenética compartida por ambos, detectando eventos de pleiotropía (S. W. Choi et al., 2020). Coa predición, por outra banda, cuantificamos en que medida o PRS pode explicar a variabilidade do fenotipo e exploramos a súa utilidade como factor de risco. Como xa se mencionou, a distribución dos PRS na poboación aproxímase a unha distribución normal. Ademais, a distribución do PRS nos casos tamén seguirá unha distribución normal pero cunha media desviada á dereita (Chatterjee et al., 2016). Esta propiedade permite clasificar aos individuos, por exemplo, no 5% da poboación co risco xenético máis baixo ou no 5% da poboación co risco xenético máis alto. É dicir, os PRS pódense usar para estratificar á poboación en función do seu risco de base, pero esta aproximación ten limitacións. Por un lado, aínda que para os fenotipos cuantitativos o PRS pode predicir un cambio directo no carácter, nos fenotipos tipo caso/control as *Odds-Ratio* (OR) asociadas ao PRS só nos informa de en que medida ocorre o evento (ser caso) ao aumentar o risco xenético con respecto a un grupo con menor risco xenético. Como consecuencia, podemos establecer algún tipo de corte na distribución do PRS en función dun valor da OR (por exemplo, a partir dun valor *X* no PRS, as *odds* de ser caso multiplícanse como mínimo polo valor *Z*), pero non podemos determinar cal é o risco dun individuo de padecer a enfermidade. Para isto é necesario calcular o risco absoluto, definido en Chatterjee e colaboradores (2016) como "a probabilidade de que un individuo asintótico desenvolva a enfermidade nun período concreto de tempo". Por outra banda, a capacidade dos SNPs para explicar a variabilidade dun carácter está teoricamente limitada pola herdabilidade baseada en SNPs (h^2_g), polo que o poder explicativo dos PRS tamén o estará (Plomin, 2013; Visscher et al., 2008; Wray et al., 2013). Os PRS obteñen valores de AUC entre 0.6-0.7 para a maioría das enfermidades comúns (Koch et al., 2023) (o AUC dun preditor aleatorio é de 0.5), e valores de R^2 que raramente superan o 10% (Y. Wang, Namba, et al., 2023). Por tanto, ao igual que ocorre con outros factores de risco, a utilidade dos PRS como ferramenta individual para a predición ou clasificación é baixa. Porén, o risco xenético pódese combinar con información clínica e medioambiental nun único modelo de predición, agrupando

os factores de risco máis importantes (modificables e non modificables) para identificar quen é susceptible de padecer a enfermidade (Chatterjee et al., 2016).

En canto aos tipos de utilidade dos PRS na práctica clínica, na literatura son frecuentes tres: (1) uso dos PRS como preditores de cara a tratar ou previr enfermidades, guiando potenciais intervencións; (2) uso dos PRS para refinar os grupos que se someterán a *screenings*, e (3) uso dos PRS para o seguimento lonxitudinal, sen necesariamente implicar intervencións, e informando sobre o efecto dos factores modificables (medioambientais) (Adeyemo et al., 2021; C. M. Lewis & Vassos, 2020; Torkamani et al., 2018). Como se comentou no apartado anterior, a inclusión do PRS como factor de risco pode axudar a detectar individuos que non serían considerados como de "alto risco" en función de factores clínicos.

Na primeira das aplicacións o PRS axuda á priorizar ou elixir unha intervención. Un caso clásico é o da enfermidade da arteria coronaria (CAD, *coronary artery disease*), para a cal se descubriu que o 20% de individuos con maior risco xenético se beneficiaban máis do tratamento con estatinas (Mega et al., 2015; Torkamani et al., 2018). Na mesma liña, o PRS de CAD pode identificar a un 8% da poboación que ten tres veces máis risco de desenvolver a enfermidade, equiparándose ao risco conferido por unha mutación de hipercolesterolemia familiar (Khera et al., 2018). Igualmente, posuír un risco xenético máis alto pode adiantar en varios anos a aparición de enfermidades como CAD (4.35 anos no top 2.5%), cancro de mama (4.89 anos), cancro de próstata (5.53 anos) ou diabetes tipo II (8.81 anos) (Mars, Koskela, et al., 2020). Isto último está moi relacionado coa aplicación de *screening*. Torkamani e colaboradores (2018) revisan o caso do cancro de mama, para o cal se desenvolveu un modelo de risco combinando un PRS de 92 SNPs e factores clínicos que podía identificar a un 16% da poboación cuxo risco a 10 anos era superior ao dunha muller media de 50 anos. Este grupo de mulleres podería verse beneficiado por un adianto do *screening* á idade de 40 (Maas et al., 2016). Para outros tipos de cancro como o colorrectal, o de ril ou melanoma, a inclusión do PRS nun modelo de predición mellora a clasificación dos individuos, nalgúns casos supoñendo unha melloría fronte a variables como a historia familiar (Kachuri et al., 2020). Os índices de risco xenético tamén poden servir para optimizar as ferramentas de *screening*. O diagnóstico de cancro de próstata pode ser refinado a través do axuste dos niveis basais do biomarcador bioquímico PSA (antíxeno específico de próstata) empregando un PRS (Kachuri, Hoffmann, et al., 2023).

A base polixénica pode incluso modular o efecto (en termos de risco) dalgunhas mutacións. En poboación finlandesa analizouse un PRS para cancro de mama a escala xenómica e atopouse que o PRS modificaba o risco das mulleres que presentaban mutacións nos xenes *PALB2* e *CHEK2*: unha muller cunha mutación en *PALB2* e unha puntuación polixénica equivalente ao percentil 90 da poboación tiña un risco de padecer cancro aos 80 anos do 83.9%, con respecto ao 55.3% que posuía unha muller coa mesma mutación pero cun PRS medio, mentres que para as mulleres cunha mutación en *CHEK2* duplicábase o risco se puntuaban alto no PRS (do 29.3% ao 59.2%) (Mars, Widén, et al., 2020). En individuos con Alzheimer detectouse un patrón similar, de xeito que os heterocigotos *APOE* ε4 no estrato máis alto do PRS tiñan un risco similar na idade da aparición da enfermidade que os homocigotos *APOE* ε4 no estrato máis baixo do PRS (de Rojas et al., 2021).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

A terceira aplicación céntrase na relación entre os factores xenéticos e os factores modificables. É evidente que atoparse nun cuantil alto da poboación no risco xenético a unha enfermidade non implica a súa aparición, pero sabelo pode ser útil para modificar as condutas e outros factores de risco na medida do posible. Sábese que fumar é un factor de risco para a enfermidade obstrutiva pulmonaria crónica, pero recentemente se descubriu que existe unha interacción entre o tabaco e o risco xenético para función pulmonar (FEV1/FVC): atoparse no decil máis baixo da distribución do PRS pero cun consumo de tabaco de >50 paquetes anuais ten un risco equiparable a estar no decil máis alto e fumar entre 11 e 20 paquetes (W. Kim et al., 2021). Aínda que isto pode parecer evidente sen necesidade de ter en conta factores xenéticos, para outros fenotipos resulta útil. Un estudo analizou a asociación de PRS para factores de risco mortais (como os niveis de colesterol, o índice de masa corporal ou o consumo de alcol) co fenotipo de mortalidade por calquera causa e revelou que, se ben a base xenética determinaba en parte a aparición da enfermidade, o seu efecto na mortalidade atenuábase tras axustar polos factores de risco medibles (Meisner et al., 2020). Así, os doentes cun risco polixénico alto para certa patoloxía poderían beneficiarse de coñecer como podería reducirse o seu risco se modificaran algunhas conductas ou melloraran o estilo de vida.

Por último, recordamos que os PRS son de utilidade en investigación para explorar unha potencial base xenética compartida entre dous fenotipos. En neuropsiquiatría é moi común empregar os PRS con este obxectivo. Por exemplo, detectáronse asociacións do risco polixénico á esquizofrenia e ao desorde bipolar con comportamentos relacionados coa adicción (Reginsson et al., 2018), mais tamén con enfermidades respiratorias ou dentais (Bigdeli et al., 2022). Baixo a mesma estratexia descubríronse asociacións entre o risco polixénico a padecer cancro de pulmón e colorrectal, ou entre o risco polixénico de sufrir cancro de próstata e de padecer cancro de endometrio (Byun et al., 2022). Este tipo de análises poden estenderse a unha multitude de fenotipos ao mesmo tempo, constituíndo un tipo de PheWAS.

Os estudos epidemiolóxicos observacionais contan cunha guía, *Strenghtening the Reporting of Observational studies in Epidemiology* (STROBE), baixo a cal se regula a súa correcta realización e publicación (Cuschieri, 2019). De forma análoga, para os PRS creouse o *Genetic Risk Prediction Studies Statement* (GRIPS), actualizado en 2020 polo *Polygenic Risk Score Catalog* en colaboración co *ClinGen Complex Disease Working Group* e culminando nunha nova guía chamada *Polygenic Risk Score Reporting Standards* ou PRS-RS (Wand et al., 2021). Entre outras cousas, establecen unha serie de criterios mínimos que é preciso cumprir para asegurar a correcta publicación dun estudo cuxo obxectivo sexa a predición xenética, de forma que sexa transparente, claro e reproducible. Entre outras cousas, destacan a descrición detallada das cohortes e dos métodos de construción do PRS, do modelo estatístico empregado para predicir o risco, da capacidade de xeralizar o modelo a outras poboacións ou da intención de uso do PRS (Wand et al., 2021).

1.5.3 Limitacións dos PRS

Os PRS teñen varias limitacións. En primeiro lugar, o solapamento entre a cohorte de descubrimento e a diana debe ser mínima. Se consideramos un posible uso práctico dos PRS, sería difícil determinar que individuos participaron na etapa de descubrimento e cales non,

especialmente tendo en conta que na actualidade moitos GWAS empregan mostras recollidas en biobancos con millóns de participantes. Por outro lado, as estimas dos efectos das variantes nos GWAS poden non ser precisas, acentuándose nos PRS a escala xenómica que inclúen SNPs non significativos (Wray et al., 2013). Tamén están presentes outras fontes de incertidume, como o ruído derivado do xenotipado e a imputación ou da correcta medición e definición do fenotipo de interese (Novembre et al., 2022). A variedade de métodos e a falta dun marco metodolóxico de referencia para a computación de PRS fai complicada a súa aplicación. A práctica común consiste en probar varios métodos e escoller o que mellor capacidade preditiva teña sobre unha cohorte de validación, o que só é útil nun contexto investigador, pero non clínico. De feito, os modelos actuais presentan moita variabilidade na predición do risco individual e son aínda inestables debido a factores como a herdabilidade e polixenicidade do carácter (Ding et al., 2022; Y. Wang et al., 2022). Ademais, para un mesmo carácter poden existir múltiples PRS, dado que para algúns fenotipos hai máis de un GWAS publicado ou os seus estatísticos resumo non son facilmente accesibles. No *PGS Catalog* (Lambert et al., 2021) (https://www.pgscatalog.org/) hai publicados un total de 3.992 índices de risco xenético para 622 fenotipos. A comparación de varios PRS para un mesmo fenotipo publicados neste catálogo e aplicados no *UKBiobank* obtivo resultados pouco esperanzadores a este respecto, xa que a estratificación dos individuos en risco/non risco variaba en función do PRS escollido (Clifton et al., 2022).

1.5.3.1 Portabilidade dos PRS entre poboacións

Unha das limitacións máis relevantes dos PRS é a dificultade para transferilos dunhas poboacións a outras. Esta problemática bebe directamente da falta de diversidade dos estudos GWAS, e gran parte da solución pasa por aumentar o número de estudos e individuos non-europeos nas análises xenéticas (De La Vega & Bustamante, 2018; Sirugo et al., 2019). O mesmo PRS pode non ter a mesma capacidade preditiva[3] ao aplicarse a poboacións europeas con respecto a aplicalo a poboacións non europeas (L. Duncan et al., 2019; Martin et al., 2017). Unha investigación que estudou o comportamento de 245 PRS revelou que o rendemento[4] dos mesmos se vía reducido ao 94% e 86% do obtido en individuos do noroeste de Europa cando eran aplicados a grupos do nordeste e sur de Europa, respectivamente, pero estes valores caían ao 48.6% ao aplicarse a individuos de China, ao 2.5% para o Caribe e ao 18% para grupos de Nigeria (Privé et al., 2022). Martin e colaboradores (2017) estudaron o impacto da ancestralidade na predición polixénica empregando PRS derivados de cohortes europeas e, tras observar que os resultados en grupos continentais non europeos eran significativamente peores, concluíron que a capacidade preditiva dos PRS dependía da similitude entre a cohorte de descubrimento e a cohorte diana. Para o fenotipo da altura, aínda que os valores observados eran maiores en individuos africanos, o PRS predeciu uns valores medios máis altos na cohorte

[3] Para medir a capacidade preditiva ou rendemento dos PRS emprégase, habitualmente, o R^2 parcial asociado á variable do PRS nos modelos de regresión. Nos fenotipos cuantitativos, exprésase como a porcentaxe de variabilidade que explica o PRS. Tamén é común usar o AUC (área baixo a curva ROC).

[4] Privé e colaboradores (2022) usan como medida da capacidade preditiva dos PRS a correlación parcial entre o fenotipo e o PRS.

europea (Martin et al., 2017). Pola súa parte, Bitarello e Mathieson (2020) replicaron esta correlación.

As limitacións na portabilidade dos PRS tamén se observaron dentro dun mesmo grupo poboacional (Mostafavi et al., 2020; Privé et al., 2022), e incluso entre individuos mixtos en función das proporcións de ancestralidade (Bitarello & Mathieson, 2020; Cavazos & Witte, 2021; Clarke et al., 2022). Existe unha correlación negativa entre a precisión da predición xenética e a distancia xenética das poboacións de descubrimento e diana, observada tanto para poboacións humanas como para animais e plantas (Scutari et al., 2016). Unha investigación recente explorou máis en profundidade esta cuestión (Ding et al., 2023). Os autores estudaron a precisión dos PRS na predición individual, que correlacionou negativamente coa distancia xenética entre o individuo e a poboación de adestramento para todos os clústeres de GIA, observándose que o rendemento dos PRS decaía de forma continuada a medida que os individuos se afastaban xeneticamente da cohorte de descubrimento (Ding et al., 2023). Estes resultados suxiren que a agrupación de individuos en clústeres de ancestralidade non alivia nin xustifica os problemas de portabilidade dos PRS.

As causas subxacentes a esta problemática, relacionadas coa distancia xenética entre poboacións, poden ser o desequilibrio de ligamento, as diferencias en frecuencias alélicas ou os sinais específicos de poboación. Y. Wang e colaboradores (2020) derivaron de xeito teórico a contribución de cada un destes factores sobre a precisión relativa dos PRS entre dúas poboacións. Asumindo unha herdabilidade constante e unha correlación exacta dos SNPs causais entre ambas, os autores constataron que a precisión na predición decaía coa distancia xenética, sendo este efecto máis pronunciado ao comparar poboacións africanas e europeas (~46% de precisión relativa cun F_{ST}=0.14) (Y. Wang et al., 2020). Os seus resultados indicaron que os factores que máis contribúen a atenuar esta precisión son a MAF e o LD.

A maioría de *software* de construción de PRS empregan paneis de desequilibrio de ligamento. Se este non reflicte os patróns de LD da mostra diana, as variantes representativas non serán seleccionadas correctamente. De igual modo, o número de SNPs que son necesarios para capturar o risco xenético depende do tamaño dos bloques de desequilibrio de ligamento na poboación (Kachuri, Chatterjee, et al., 2023). O efecto estimado das variantes tamén está influenciado polo LD, xa que a correlación entre os SNPs e as variantes causais produce unha correlación entre os seus efectos estimados (Martin et al., 2019; Y. Wang et al., 2022). Ademais, como xa se explicou, unha variante *tag* identificada en poboación europea é probable que non represente adecuadamente á variante causal noutra poboación, sobre todo se esta última non está representada na cohorte de descubrimento. Non obstante, parece que usar un panel de referencia cunha composición proporcional á mostra diana non mellora necesariamente a predición con respecto a usar o panel de LD da poboación maioritaria no GWAS (Y. Wang, Kanai, et al., 2023; Y. Wang, Namba, et al., 2023).

Por outra parte, os sinais identificados no GWAS están determinados pola frecuencia alélica. De forma xeral, as variantes *hit* serán comúns (en termos de MAF) dentro da poboación de descubrimento, o que sobre-representa o seu risco no PRS e enmascara variantes que son máis frecuentes noutras poboacións (Kachuri, Chatterjee, et al., 2023; Martin et al., 2019). Un estudo que analizou a portabilidade dos PRS de expresión xénica confirmou o efecto da MAF:

considerando unicamente as variantes causais cis-eQTL compartidas entre poboación europea e yoruba (e, por tanto, independentes dos efectos do LD), os PRS obtiveron peores resultados na poboación yoruba, sobre todo cando as variantes incluídas eran comúns en europeos pero de baixa frecuencia no grupo africano (Saitou et al., 2022).

Por último, os factores non xenéticos tamén xogan un papel fundamental, posto que as interaccións xenotipo-ambiente poden modificar o efecto dunha variante nunha poboación específica (Martin et al., 2019; Y. Wang et al., 2022). Igualmente, se a contribución dos factores ambientais non ten a mesma magnitude nas cohortes de descubrimento e diana, as variantes incluídas no PRS poden presentar ruído na súa estima e a precisión da predición verase reducida.

A composición do GWAS de descubrimento é quizais o aspecto que máis inflúe na capacidade dos PRS de transferirse entre poboacións, sobre todo tendo en conta o efecto da distancia xenética. Varios estudos demostraron que a combinación de poboacións na mostra de adestramento é a estratexia que mellores resultados ofrece, ben sexa a través de meta-análises multi-poboación ou da construción de PRS multi-poboacionais con métodos específicos (Márquez-Luna et al., 2017; Patel et al., 2023; Ruan et al., 2022; Y. Wang, Kanai, et al., 2023). Aínda que polo xeral é beneficioso aumentar o número de individuos da poboación diana na cohorte de descubrimento, a composición óptima do GWAS varía entre os fenotipos e entre as poboacións. Por exemplo, para predicir en poboación africana fenotipos como a altura, o volume corpuscular medio ou o número de plaquetas, funcionou mellor un PRS derivado dun GWAS africano, pero para o índice de masa corporal ou a asma foi máis efectivo un PRS europeo (Lehmann et al., 2023). Por outra banda, para obter unha melloría significativa na predición dos PRS en grupos africanos requiriuse engadir un número máis alto de individuos africanos que o correspondente en grupos asiáticos (Y. Wang, Kanai, et al., 2023). En xeral, a polixenicidade do carácter é un factor importante: os fenotipos máis polixénicos poden beneficiarse dos maiores tamaños de mostra dos GWAS europeos, pero os menos polixénicos parecen necesitar dun maior número de individuos da poboación minoritaria (Y. Wang, Kanai, et al., 2023). Na mesma liña, Lehmann e colaboradores (2023) destacan que, para algúns fenotipos, os SNPs comúns son os que máis peso teñen na capacidade preditiva, mentres que para outros teñen máis relevancia as variantes específicas de poboación. En definitiva, a estratexia óptima de construción de PRS é variable para cada fenotipo e poboación, e dependente de múltiples factores, facendo imposible establecer un marco metodolóxico fiable. Porén, a necesidade de incorporar individuos non europeos nos estudos GWAS é indiscutible.

Kachuri, Chatterjee, e colaboradores (2023) describen varios métodos e *software* destinados á construción de PRS en poboacións diversas, recollendo procedementos como o uso dun GWAS auxiliar de poboación non europea, a modelización de resultados de GWAS en diferentes poboacións ou a inclusión doutro tipo de información como a correlación trans-poboacional. Y. Wang, Kanai, e colaboradores (2023) tamén proporcionan unha serie de recomendacións para a construción de PRS multi-poboación, en función da arquitectura xenética do fenotipo.

En relación aos individuos mixtos, que tenden a estar menos representados nos estudos dos PRS ao igual que nos estudos GWAS, cómpre engadir algunhas consideracións metodolóxicas

adicionais. Como xa se aclarou, estas poboacións permiten un control dos factores medio-ambientais externos, aumentando a capacidade de detección de variantes con efectos heteroxéneos entre poboacións ou con menor MAF (Tan & Atkinson, 2023). Así, pode ser de interese construír PRS máis complexos coas poboacións mixtas que incorporen a ancestralidade local. Esta aproximación mellorou a predición da altura nun grupo de individuos mixtos (Bitarello & Mathieson, 2020). Outra estratexia que obtivo resultados prometedores con respecto a un PRS simple foi a construción de índices xenéticos parciais, específicos de poboación, e a súa combinación ponderándoos polas proporcións de ancestralidade (Marnetto et al., 2020). Pola súa banda, Y. Wang, Kanai e colaboradores (2023) testaron o funcionamento de PRS derivados co *software* Tractor. Para algúns fenotipos si observaron unha predición mellorada cos PRS derivados de Tractor en relación aos derivados de GWAS tradicionais, sobre todo en presencia de variantes con efectos específicos de poboación, pero para outros fenotipos non atoparon diferencias (Y. Wang, Kanai, et al., 2023). Na actualidade estanse desenvolvendo novos métodos para a derivación de PRS en cohortes mixtas, como por exemplo GAUDI, que emprega regresión penalizada (Q. Sun et al., 2024) -aínda que de momento só está optimizado para dúas poboacións ancestrais-, ou iPGS, que calcula o PRS en datos individuais a través dun procedemento *lasso* iterativo e permite incluír interaccións entre as PCs e as variantes (Tanigawa & Kellis, 2023), pero para o que se necesita unha mostra de adestramento, outra de validación e outra de test. O requirimento de mostras adicionais para o refinamento dos parámetros pode resultar unha desvantaxe dalgúns métodos, dado que en ocasións son difíciles de conseguir ou presentan un número de individuos reducido.

Finalmente, na construción de PRS para poboacións diversas cómpre ter en conta algunhas complicacións engadidas e non tan exploradas. A maioría dos estudos dedicados a comparar a xeneralización dos PRS entre poboacións céntranse na cuantificación do rendemento do mesmo en termos de R^2 ou pseudo-R^2 parcial ou AUC, mentres que os estudos máis enfocados á translación clínica reportan unha OR para cada grupo poboacional ou ancestralidade. Isto complica a aplicabilidade dos PRS, ao non ser posible establecer un único punto de corte ou estimar un único efecto para todos os individuos. Incluso dentro dun mesmo biobanco se observa certa heteroxeneidade nos efectos dos PRS (Mars et al., 2022; Y. Wang, Namba, et al., 2023), independentemente da orixe ancestral da poboación, o que suxire un efecto non modelado das interaccións xenotipo-ambiente. Neste sentido, cómpre explorar alternativas que non requiran da clasificación en grupos discretos e a incorporación de factores ambientais aos modelos. A inclusión de individuos mixtos nas mostras de adestramento pode reducir a distancia xenética dos individuos á mesma, dada a combinación de varios tractos ancestrais no seu xenoma.

1.6 COVID-19

En decembro de 2019 apareceu o primeiro caso de infección por un novo virus, o SARS-CoV-2 (*Severe Acute Respiratory Syndrome Coronavirus 2*), en Wuhan, China (N. Zhu et al., 2020). Os pacientes presentaban unha pneumonía grave pero cunha orixe descoñecida. Este virus de ARN, da familia dos coronavirus, era totalmente novo. O seu xenoma foi rapidamente secuenciado, e en xaneiro xa estaba dispoñible en bases de datos públicas (Goyal et al., 2020).

USC UNIVERSIDADE DE SANTIAGO

En marzo de 2020, o virus xa estaba estendido por todo o mundo, e a OMS declarou un estado de pandemia (World Health Organization, 2020a). Nun segundo informe de finais de marzo, a OMS reportaba 750.890 casos confirmados e 36.405 mortes (World Health Organization, 2020b). En pouco tempo, países de todo o mundo seguiron os pasos de China e comezaron a confinar as súas cidades para conter a propagación.

Ademais do confinamento puxéronse en marcha outras estratexias epidemiolóxicas, como por exemplo o testado masivo da poboación. Desenvolvéronse test rápidos de antíxenos que permitían detectar a infección en pouco tempo, facilitando a identificación de doentes asintomáticos que eran transmisores do virus e o rastreo de contactos (Mercer & Salit, 2021). Este *screening* poboacional tivo moito éxito no control do virus, mais tamén serviu para a vixilancia epidemiolóxica (Mercer & Salit, 2021). Tras levantarse as restricións de mobilidade características do confinamento, introducíronse medidas de distanciamento social e hixiene co obxectivo de reducir ao máximo a transmisión do SARS-CoV-2. Con todo, o evento máis salientable foi a chamada "carreira pola vacina". A finais de 2020 había máis dunha vacina na fase III dos ensaios clínicos, entre elas a de Pfizer-BioNTech (Polack et al., 2020), cunha taxa de efectividade de prevención da infección do 95%, e Moderna, cun 94.1% (Baden et al., 2021), que foron as primeiras en aprobarse. Ambas son vacinas de ARN, cuxo mecanismo de acción baséase en introducir no hóspede o material xenético para producir o antíxeno, neste caso a proteína *spike* do virus modificada (Baden et al., 2021, 2021). A aprobación da vacina deuse por primeira vez dentro de Europa en Reino Unido (Ledford et al., 2020), pero pronto comezou unha campaña de vacinación mundial. Para finais de 2022, en España vacináranse máis de 38 millóns de persoas.

Dende a aparición do primeiro caso de COVID-19 en decembro de 2020 en China, máis de 771 millóns de persoas foron infectadas polo virus SARS-CoV-2, provocando 6 millóns de mortes (datos da OMS en outubro de 2023, https://covid19.who.int/). España confirmou 14 millóns de casos positivos (datos da OMS en outubro de 2023, https://covid19.who.int/). Esta pandemia afectou a todas as rexións do mundo, pero as taxas de mortalidade e infección variaron entre países (ver **Figura 5**). En canto á rexión de Latino-América e o Caribe, en marzo de 2022 reportáranse un total de 1.65 millóns de mortes (Schwalb et al., 2022). Swchalb e colaboradores (2022) recompilaron datos da taxa de mortalidade por cada 100.000 habitantes a xaneiro de 2022. Esta lista estaba encabezada por Perú (cunha taxa de mortalidade de 674.44/100.000 habitantes), seguido de Brasil, Arxentina, Colombia, Paraguay e México (Schwalb et al., 2022). Amnistía Internacional e o Centro polos Dereitos Económicos e Sociais (CSER), nun reporte de 2022, advertía sobre o efecto das desigualdades na rexión de Latino-América e o Caribe, que acumulaban o total do 28% das mortes mundiais por COVID-19 (Amnistía Internacional & Center for Economic and Social Rights, 2022). No informe salientaban algunhas das causas deste exceso de mortalidade, como as desigualdades económicas e acceso ao sistema de saúde, as desigualdades nas condicións materiais e sanitarias, a ineficacia nas medidas de protección, ou a financiación dos sistemas de saúde (Amnistía Internacional & Center for Economic and Social Rights, 2022).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

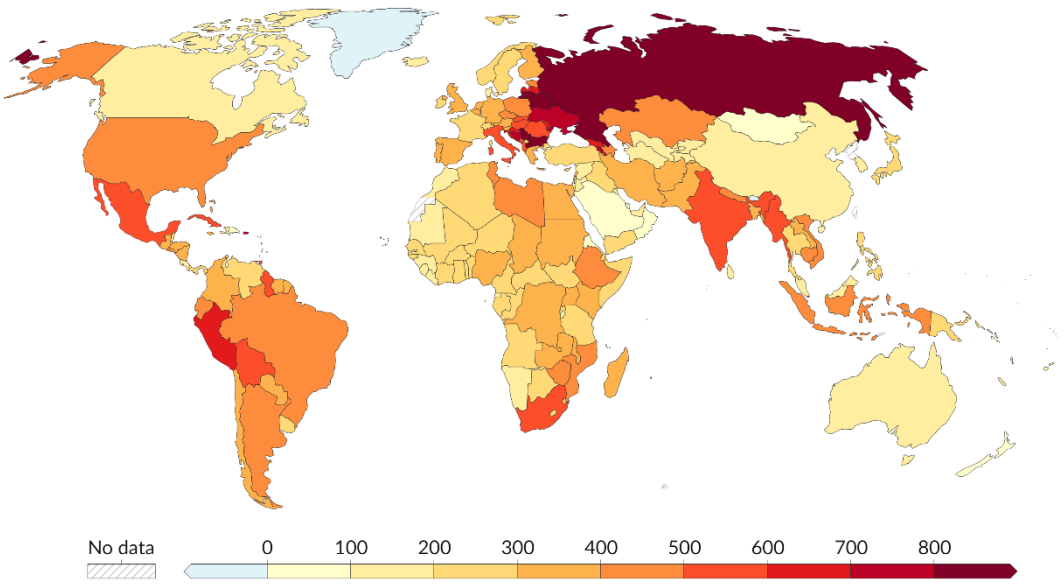


Figura 5. Estimas do exceso de mortalidade por COVID-19 por cada 100.000 habitantes. Este gráfico foi recuperado de *Our World in Data*, a día de 27/01/2024 (https://ourworldindata.org/explorers/coronavirus-data-explorer). O modelo e os datos empregados para a estimación do exceso de mortalidade a causa da COVID-19 son de *The Economist* (2022), procesados por *Our World in Data*.

1.6.1 Patoxénese da COVID-19

O SARS-CoV-2 é un virus de ARN de cadea única en sentido positivo (++ssRNA) da familia *Coronaviridae* e xénero Betacoronavirus. Pertence ao mesmo subxénero que SARS-CoV e o MERS-CoV, e filoxeneticamente atópase no mesmo clúster que outros coronavirus SARS atopados en morcegos e pangolíns (B. Hu et al., 2021). Está composto de 4 proteínas estruturais maioritarias: a *spike* (S), a glicoproteína da cápsula (E), a nucleocápside (N) e a proteína da membrana (N). As proteínas S atópanse por toda a superficie do virus e presentan dúas subunidades: a S1, que se une ao receptor a través da rexión de unión ao receptor (*Receptor-binding domain*, RBD), e a subunidade S2, responsable da fusión das membranas (Lamers & Haagmans, 2022). No caso do SARS-CoV-2 e do SARS-CoV, a molécula receptora do hóspede é a enzima convertidora da anxiotensina 2 ou ACE2 (Letko et al., 2020). ACE2 é unha carboxipeptidasa que convirte a anxiotensina I ás anxiotensinas 1-9. Tras a unión da S1 co receptor ACE2, entra en xogo unha proteasa do hóspede, TMPRSS2, que escinde a proteína S en dous, activando a subunidade S2 e expoñendo o péptido responsable da fusión de membranas (Hoffmann et al., 2020). Así, a ribonucleoproteína entra na célula do hóspede. Se ben este é o mecanismo máis común, en casos onde a expresión de *TMPRSS2* é baixa, o virus emprega un sistema de endocitose mediada por clatrina ata os lisosomas (Bayati et al., 2021; Jackson et al., 2022), onde a escisión da S2 é realizada por catepsinas. A duración da fase de incubación é heteroxénea entre as diferentes variantes do SARS-CoV-2, pero o rango aproximado é de 5-6 días (Parasher, 2021). Durante a primeira fase de infección a enfermidade cursa ou ben de forma asintomática ou con síntomas leves, como tose, febre ou fatiga (Huang et al., 2020; Z. Wu & McGoogan, 2020). Ao redor do 35% dos infectados non presenta síntomas, complicando o control da transmisión (Sah et al., 2021).

As primeiras células en ser infectadas polo virus son as células ciliadas nasais. Isto parece estar mediado en parte pola expresión do xen *ACE2*, que presenta un gradiente ao longo do tracto respiratorio (Y. J. Hou et al., 2020). O virus comeza a producir proteínas virais e a maquinaria de transcrición para a súa replicación, dando comezo a unha etapa da infección na que a resposta inmune é baixa pero os infectados son altamente contaxiosos. A continuación, migra ás células do tracto respiratorio superior. Os receptores que recoñecen patróns citoplasmáticos (*cytoplasmatic pattern recognition receptors*, PRRs) detectan as moléculas virais e desencadéase unha resposta inmune innata (Mason, 2020), caracterizada por un estado de inflamación prolongado na que os receptores tipo *Toll* (TLRs) TLR2 e TLR4 inducen a síntese e liberación de citoquinas, e os TLR7 e TLR3 a de interferóns (IFs) tipo I e III (Paludan & Mogensen, 2022). Nunha porcentaxe mais baixa das veces, o virus pode infectar as células alveolares, producindo un estado de hipoxia e un cadro de pneumonía. Esta fase da enfermidade pode chegar a cursar moderadamente grave, pero a maioría dos doentes non progresan máis alá. Nun estudo de 73.314 da COVID-19 en China, reportouse un 14% de hospitalización (Z. Wu & McGoogan, 2020), aínda que estes números poden presentar certa variación entre países en función de factores como a calidade e capacidade do sistema de saúde ou socioeconómicos. En España, por exemplo, a hospitalización dábase principalmente en doentes moi graves (Casas-Rojo et al., 2020a). A OMS considera enfermidade de gravidade moderada en caso de presentar infección das vías respiratorias inferiores (pneumonía), con niveis de saturación de osíxeno en sangue maiores ao 90% (World Health Organization., 2023).

Os IFs inducen a produción de proteínas antivíricas, e as citoquinas atraen aos linfocitos para iniciar unha resposta inmune adaptativa. A liberación continuada de citoquinas, quemoquinas e outros axentes inflamatorios por parte das células tipo macrófagos ou dentríticas e epiteliais recibe o nome de "tormenta de citoquinas" (Fajgenbaum & June, 2020). Esta acumulación provoca o fenotipo de hiper-inflamación tan característico da COVID-19. Entre as citoquinas máis atopadas en doentes con enfermidade grave están a interleucina 6 (IL-6), a interleucina 8 (IL-8) e o factor de necrose tumoral α, sendo os tres marcadores robustos dunha peor prognose e supervivencia (Del Valle et al., 2020). As células alveolares infectadas sofren danos ou directamente a apoptose, o que produce un dano do epitelio alveolar severo e dá lugar ao síndrome de dificultade respiratoria aguda (*acute respiratory distress syndrome*, ARDS), a principal causa de morte por COVID-19 (B. Hu et al., 2021). A COVID-19 tamén se caracteriza por una coagulopatía e un estado de endotelite vascular. O fluxo de moléculas e fluído entre o sangue e o pulmón dáse por cambios na estrutura do endotelio vascular, que se inflama e aumenta a súa permeabilidade, activándose un estado de disfunción endotelial e a produción de moléculas pro-inflamatorias e relacionadas coa coagulación (Vassiliou et al., 2020). A coagulopatía asociada á COVID-19 está marcada pola aparición de tromboembolismos venosos, cunha alta incidencia en doentes con COVID-19 crítica, e un fenotipo pro-coagulante (Leentjens et al., 2021). A COVID-19 grave caracterízase por presentar <95% de saturación de osíxeno en sangue, síntomas de pneumonía grave ou de dificultades respiratorias (World Health Organization., 2023). Se a enfermidade progresa ata producirse o ARDS, sepse, ou o doente require de ventilación mecánica ou terapia vasopresora, considérase un estado crítico (World Health Organization., 2023). Nesta fase da enfermidade aparecen outras complicacións, entre

elas a morte (Cascella et al., 2024). As máis frecuentes que se reportaron en varios estudos observacionais foron o ARDS, o requirimento de ventilación mecánica, o ingreso en UCI, o fallo renal, a presencia de infiltracións pulmonares e os danos vasculares e do corazón (Berenguer et al., 2020; Casas-Rojo et al., 2020a; Docherty et al., 2020; Petrilli et al., 2020; Potere et al., 2020). En España, a porcentaxe de mortalidade entre os doentes hospitalizados durante 2020 e 2021 oscilou entre o 15% e o 20% (Casas-Rojo et al., 2020a; Garcia-Carretero et al., 2023).

1.6.2 Factores de risco

A gravidade da COVID-19 presenta variabilidade entre os individuos, polo que é esencial comprender os factores de risco que inflúen nela. Dentro dos factores demográficos, os máis relevantes son o sexo e a idade. Os primeiros estudos que describiron as características clínicas da enfermidade indicaron que o sexo masculino compoñía ao redor 60% dos doentes que requirían hospitalización ou presentaban un desenlace clínico máis grave (Casas-Rojo et al., 2020a; N. Chen et al., 2020; Docherty et al., 2020; Guan et al., 2020). Tamén se detectou un maior número de homes falecidos, pero non foi así para a taxa de infección (Deng et al., 2020). Aínda que isto pode deberse a que algúns dos outros factores de risco para a COVID-19 son máis prevalentes en homes, como a exposición laboral ou o tabaquismo, as hipóteses tamén apuntan a mecanismos biolóxicos como a regulación hormonal ou aos patróns diferenciais entre sexos da expresión xénica nas rutas do sistema inmune (Scully et al., 2020). Por outro lado, a idade é un preditor forte tanto da hospitalización como da morte por COVID-19. As taxas de hospitalización acadan os valores máximos nos rangos entre 50-70 anos (Casas-Rojo et al., 2020a; Fried et al., 2021; Huang et al., 2020; Ko et al., 2021), pero para a mortalidade existe un patrón para o risco moi marcado pola idade. Polo seu lado, O'Driscoll e colaboradores (2021) observaron que o risco de mortalidade por COVID-19 aumentaba de forma log-lineal para individuos maiores de 30 anos, con resultados homoxéneos entre distintas áreas xeográficas. Curiosamente, os autores destacaban que en varios países de Latino-América, as taxas de mortalidade en individuos maiores de 65 anos eran menores das esperadas, probablemente debido ás inconsistencias nos reportes das mortes por COVID-19 neste grupo de doentes (O'Driscoll et al., 2021). Outros estudos como o da cohorte ISARIC no Reino Unido ou o dunha cohorte en Nova York obtiveron resultados similares, reportando uns riscos marcadamente máis elevados de hospitalización e morte a partir dos 70 anos (Docherty et al., 2020; Petrilli et al., 2020). A asociación da idade e do sexo co risco de mortalidade e de enfermidade crítica mantéñense incluso tras axustar por comorbilidades pre-existentes, o que indica que son factores de risco independentes (Docherty et al., 2020; R. K. Gupta et al., 2021; Henkens et al., 2022).

A presencia de condicións médicas previas á infección están estreitamente vinculadas co desenlace crítico da enfermidade. Berenguer e colaboradores (2020) reportaron como comorbilidades máis observadas entre doentes hospitalizados en España a hipertensión, a obesidade, a diabetes mellitus e as enfermidades crónicas pulmonares e de corazón. Estes autores desenvolveron un modelo de risco para morte por COVID-19 e destacaron 17 preditores fortes, entre os que se inclúe algunha outra patoloxía previa como enfermidades neurolóxicas

USC

crónicas, cancro, demencia ou cirrose. En Inglaterra, un estudo de máis de 17 millóns de persoas explorou as características clínicas en doentes hospitalizados reportou observacións similares e resaltou o risco asociado á asma (Williamson et al., 2020). As comorbilidades mencionadas repetíronse noutras cohortes de diversos países (Abolfotouh et al., 2022; Casas-Rojo et al., 2020b; de Souza et al., 2020; Docherty et al., 2020; Fried et al., 2021; Guan et al., 2020; Petrilli et al., 2020). Con todo, os modelos de predición para risco de morte e hospitalización desenvolvidos ata o momento difiren nos factores incluídos. Nalgúns casos, como o estudo de Petrilli e colaboradores (2020), a asociación das comorbilidades coa enfermidade grave ou coa morte foi máis feble que coa hospitalización, especialmente para as menos graves. Este efecto parece darse porque as patoloxías previas teñen un impacto diferente en cada fase da enfermidade. Especificamente, as comorbilidades que provocan unha disfunción no sistema inmunolóxico son máis propensas a incrementar o risco á infección ou a un cadro clínico máis grave (Russell et al., 2023). En cambio, durante as fases máis agudas da enfermidade, entran en xogo as comorbilidades que producen unha "tolerancia reducida" ao fallo respiratorio grave ou á morte (Russell et al., 2023), como poden ser a obesidade, hipertensión ou cancro. Un dos modelos de predición para enfermidade por COVID-19 é o "*ISARIC 4C Deterioration model",* construído por un equipo en Reino Unido sobre 75.000 individuos empregando como criterios de "deterioro" o inicio de soporte respiratorio, admisión na UCI ou morte (R. K. Gupta et al., 2021). O modelo incluíu como preditores finais a idade, o sexo, a ocorrencia de infeccións nosocomiais, a escala de coma de Glasgow, a saturación de osíxeno no momento de admisión, a taxa respiratoria, a concentración de urea, a concentración da proteína reactiva C, o número de linfocitos e a presencia de infiltracións en radiografía de tórax. Curiosamente, a pesar de que se incorporou unha variable composta para o número de comorbilidades, esta non tivo relevancia na predición (R. K. Gupta et al., 2021).

Finalmente, cómpre recordar que os factores socio-económicos tamén teñen un impacto tanto na diagnose da enfermidade como na hospitalización. En Brasil atopouse unha asociación entre o número de casos detectados e a renta per cápita, especialmente no primeiro termo da pandemia (de Souza et al., 2020). Williamson e colaboradores (2020) detectaron un risco superior de morte para grupos étnicos non brancos en Reino Unido, aínda tras axustar por patoloxías previas e polo índice de privación. Estes efectos poden explicarse, en parte, por unha acceso desigual ao sistema de saúde, especialmente en países onde a sanidade non é universal; pero tamén porque as condicións económicas poden provocar que se acuda ao médico nun punto da enfermidade máis avanzada. Outros estudos reportaron taxas de ingreso en UCI e infección superiores para etnias ou grupos poboacionais minoritarios (Lo et al., 2021; Magesh et al., 2021), destacando o papel dos factores socioeconómicos como mediadores. Magesh e colaboradores (2021), ademais, detectaron que os individuos de comunidades máis necesitadas tiñan un risco maior de ter contacto con infectados, causado polos seus postos de traballo, a tendencia a vivir en zonas con máis densidade de poboación, ou a imposibilidade de manter distancia social.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

1.6.3 Estudo da susceptibilidade xenética á covid-19

En xuño de 2020, tan só catro meses despois de que a OMS declarara a pandemia, había máis de 23.000 publicacións científicas indexadas sobre ela (Teixeira da Silva et al., 2021). Nos inicios, a maioría das investigacións no campo da xenética tiñan relación co propio virus e coa súa evolución, pero en outubro de 2020 publicouse o primeiro estudo sobre a xenética dos hóspedes que contraían COVID-19 grave (The Severe Covid-19 GWAS Group et al., 2020). Neste GWAS analizáronse como casos 1.980 individuos que presentaban fallo respiratorio tras ser infectados. No artigo indícase que a recollida das mostras, en España e Italia, comezara a finais de febreiro, mentres que o xenotipado e a análise estatística se realizaron en tan só 8 semanas. Os propios autores reivindicaban a utilidade dos GWAS pola facilidade da cooperación internacional e rapidez das análises (The Severe Covid-19 GWAS Group et al., 2020).

Pese a ter un tamaño de mostra relativamente pequeno para os GWAS actuais, *The Severe COVID-19 GWAS Group* (2020) deu identificado os que foron os primeiros marcadores de fallo respiratorio por SARS-CoV-2: no cromosoma 3, unha rexión que abarca os xenes *SLC6A20*, *LZTFL1*, *CCR9, FYCO1*, *CXCR6* e *XCR1*; e no cromosoma 9, coincidindo co *locus ABO*. O fenotipo estudado neste caso era fallo respiratorio, definido a partir do requirimento de osíxeno ou ventilación mecánica. Con respecto ao *locus ABO*, o estudo atopou que o grupo sanguíneo A tiña un maior risco de ser infectado, mentres que o grupo O confería protección (The Severe Covid-19 GWAS Group et al., 2020).

Un segundo GWAS publicado en decembro de 2020 polo consorcio GenOMICC (*Genetics Of Mortality in Critical Care*) analizou os xenomas de 2.244 individuos de UK con enfermidade crítica recrutados en UCI (The GenOMICC Investigators et al., 2021). Identificaron cinco sinais independentes, nos xenes *LZTFL1, OAS1-OAS3, DPP9, TYK2* e *IFNAR2*, con evidencia de replicación nunha cohorte de doentes hospitalizados. A través dunha análise de transcriptoma completo (TWAS), determinaron unha expresión diferencial dos xenes *CCR2*, *CCR3* e *CXCR6* (cromosoma 3), e *MTA2B* (cromosoma 5), entre casos e controis en tecido pulmonar. Seguidamente, a empresa *23andMe* publicou outro GWAS con 12.972 casos positivos en COVID-19, localizando asociacións nos xenes *ABO* e na rexión do cromosoma 3 (Shelton et al., 2021), se ben nesta última o xen priorizado era *SLC6A20*.

O maior esforzo internacional realizouse ao redor do consorcio HGI (*The COVID-19 Host Genetics Initiative*, https://www.covid19hg.org/). Este consorcio estableceuse tamén a principios de 2020 e, entre outras cousas, proporcionou aos grupos integrantes uns protocolos tanto clínicos como metodolóxicos para deseñar e levar a cabo os GWAS (The COVID-19 Host Genetics Initiative, 2020). A idea era homoxeneizar todos os GWAS, dende o fenotipado ata os modelos estatísticos, para posteriormente combinalos en meta-análises. Exploráronse tres definicións de casos: COVID-19 positiva -para estudar a susceptibilidade á infección-, hospitalización, e enfermidade crítica. Ademais, comparáronse dúas definicións de controis (individuos con estado de COVID-19 coñecido ou a combinación destes con controis poboacionais), e concluíron que non só os resultados eran similares, senón que a inclusión de controis poboacionais aumentaba a potencia estatística (COVID-19 Host Genetics Initiative et al., 2021). Publicáronse un total de 7 actualizacións das análises e os resultados están

dispoñibles de forma pública na páxina web (https://www.covid19hg.org/). O consorcio agrupa 119 estudos, e un dos seus obxectivos principais era promover a inclusión de poboacións diversas para asegurar a súa representación e tamén para aumentar a potencia estatística nas análises (COVID-19 Host Genetics Initiative et al., 2021). As meta-análises fixéronse combinando todas as poboacións pero tamén de forma individual (específicas de poboación). Con todo, aínda que a participación de estudos non europeos foi alta, o número de casos de ascendencia europea na última actualización é 5 veces maior que o de non-europeos (conxuntamente) para o fenotipo de hospitalización (**Figura 6**) e 3 veces maior para o fenotipo de enfermidade grave (https://www.covid19hg.org/results/r7/). Das meta-análises específicas de poboación, só na europea se identificaron sinais adicionais.

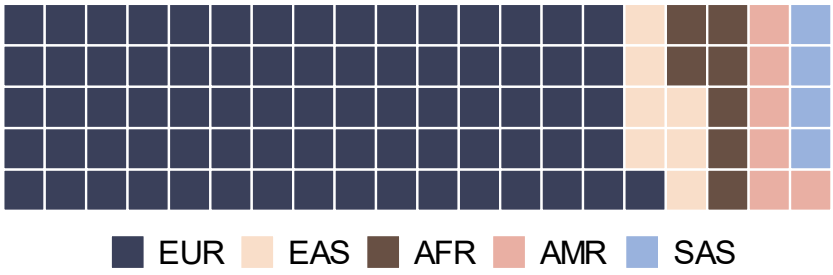


Figura 6. Proporción relativa de individuos hospitalizados entre as diferentes poboacións incluídas na análise B2 do consorcio HGI. Os datos brutos sen escalar son: EUR=32.519; EAS=3.077; AFR=2.882; AMR=2.859; SAS=1.622.

A primeira publicación do consorcio HGI (COVID-19 Host Genetics Initiative et al., 2021) combinou 49 estudos participantes: analizáronse 49.562 casos e sobre 2 millóns de controis, incluíndo grupos poboacionais diversos. Destes, unha gran parte correspondía ao fenotipo de infección, outra parte máis reducida ao fenotipo de hospitalización e o resto ao fenotipo de enfermidade crítica. Cada grupo poboacional meta-analizouse de forma individual, mais a análise principal constituía unha combinación trans-poboacional. Descubríronse 13 *loci* tomando en conxunto os tres fenotipos. Entre outras, 7 das rexións estaban asociadas á infección por SARS-CoV-2. Destaca o *locus ABO* (rs91280523), reportado por primeira vez por *The Severe COVID-19 GWAS Group* (2020) e que non fora replicado no estudo de GenOMICC de enfermidade crítica (COVID-19 Host Genetics Initiative et al., 2021). Ademais, no cromosoma 3 diferenciáronse dúas asociacións independentes, a liderada polo rs10490770 na zona de *LZTFL1* e outra liderada polo SNP rs2271616 no xen *SLC6A20*, que só acadou significación GWAS co fenotipo de susceptibilidade, replicando un resultado de *AncestryDNA* (Roberts et al., 2020). Das 13 rexións atopadas polo HGI, 9 presentaban unha OR maior co fenotipo de COVID-19 crítica ou hospitalización, o que evidenciaba unha implicación destes *loci* na enfermidade e non na infección. Algún deles foron *TYK2*, *FOXP4*, *IFNAR2* ou *DPP9*.

No 2022, o consorcio GenOMICC publica un segundo estudo incrementando o tamaño de mostra a 7.491 doentes con COVID-19 crítica e 48.400 controis, pero esta vez con secuenciación completa do xenoma (WGS) aumentando o número de asociacións a 25

(Kousathanas et al., 2022). A aproximación a través da secuenciación aumenta a resolución do estudo ao posibilitar a identificación directa de variantes raras e codificantes. Por exemplo, no cromosoma 3 detectan unha variante *missense* no xen *PLSCR1*, que controla a replicación de virus de RNA, e outra que afecta ao xen *IFNA10*. Neste estudo repórtase por primeira vez unha asociación coa zona *HLA*.

O consorcio HGI actualizou os seus resultados cunha segunda publicación na que participaban 60 estudos de 25 países (Pathak et al., 2022). En total analizaron 125.584 casos, o que permitiu atopar 11 asociacións máis (Pathak et al., 2022). Ademais, introduciron un método baiesiano novidoso para a clasificación das variantes de acordo aos fenotipos, permitindo distinguir entre cales están implicadas na susceptibilidade á infección e as que teñen un papel na progresión da enfermidade. Entre outras cousas, reportaron una heteroxeneidade significativa entre grupos poboacionais para o *locus FOXP4*.

Os estudos GWAS para a COVID-19 identificaron un gran número de asociacións, pero moitas delas están localizadas en rexións que abarcan máis dun xen, complicando a priorización dos xenes causais e a súa función. Ademais, hai varios factores que afectan á replicación dos sinais, limitando a validez e interpretación dalgunhas asociacións. Por un lado, o tamaño de mostra nestes GWAS incrementouse grazas ao uso de controis dos que se descoñecía a súa exposición ao virus, o que puido introducir nesgos nos estudos, sobre todo tendo en conta a temporalidade da pandemia e a evolución das variantes (Niemi et al., 2022). O ruído introducido por esta práctica puido ser a causa de que algunhas das rexións involucradas na infección se atoparan tamén con fenotipos máis graves. Por outro lado, a definición dos fenotipos relacionados coa enfermidade crítica depende de criterios clínicos, producindo unha variabilidade entre estudos, especialmente cando se combinan resultados de diferentes países. Con todo, a replicación continua de varias rexións como a dos cromosomas 3, 12 e 21, ou detección de variantes codificantes permitiron sinalar algunhas rutas biolóxicas implicadas. Os resultados máis relevantes descríbense a continuación.

1.6.4 Variantes xenéticas implicadas na COVID-19

1.6.4.1 Xenes implicados na entrada viral e susceptibilidade á infección

Dende os inicios da pandemia se debateu sobre o posible papel do tipo sanguíneo na infección. Mediante técnicas bioquímicas estudouse a distribución dos grupos sanguíneos nos doentes de COVID-19, e os resultados apuntaban a un maior risco de ser infectado nos individuos con sangue tipo A, mentres que o tipo O conferiría protección (Zhao et al., 2021). Porén, outras investigacións obtiveron resultados contraditorios ao non observar unha relación entre ambos factores (E. Pereira et al., 2022; Y. Zhang et al., 2021). Paralelamente, os estudos xenéticos tipo GWAS si detectaron unha asociación na zona do xen *ABO* co risco de infección, a cal foi replicada en máis dunha ocasión (**Anexo I**). *The Severe COVID-19 GWAS Group* (2020) empregou os xenotipos dos SNPs implicados no *ABO* para predicir o grupo sanguíneo dos individuos e concluíu que o grupo A tiña un risco maior de contraer COVID-19. O algoritmo de clasificación de variantes desenvolvido por Kanai e colaboradores (2023) asignou esta rexión ao fenotipo de susceptibilidade, aínda que o estudo da expresión do xen detectou

USC DE SANTIAGO DE COMPOSTELA

unha asociación coa hospitalización e con eventos clínicos máis graves como embolia pulmonar, infarto pulmonar, embolismos e trombose (Pathak et al., 2021). Ellinghaus (2023) discute dous mecanismos fisiopatolóxicos que poderían explicar a relación entre o SARS-CoV-2 e o grupo sanguíneo. A primeira delas baséase nas regras de incompatibilidade do sistema ABO. Dado que o virus se replica en células respiratorias que expresan antíxenos ABO, os antíxenos tamén poderían estar presentes na superficie do virus, polo que unha persoa con grupo sanguíneo O estaría máis protexida por ter anticorpos anti-A e anti-B (Ellinghaus, 2023; Ellis, 2021). A outra hipótese baséase en que o dominio de unión ao receptor (RBD) do virus ten unha secuencia similar á das galectinas, unhas proteínas con alta afinidade polo grupo ABO (Ellinghaus, 2023; Wu et al., 2023). O dominio RBD ten máis afinidade por células que expresan o antíxeno tipo A, propoñendo así una plausible explicación biolóxica (S.-C. Wu et al., 2023).

FUT2 tamén acadou unha asociación significativa nos GWAS (Kanai et al., 2023; Kousathanas et al., 2022). *FUT2* codifica unha fucosiltransferasa que regula a expresión dos antíxenos do grupo ABO en mucosas e fluídos e determina o fenotipo "secretor", caracterizado pola presencia de antíxenos ABO en fluídos corporais (**Figura 7**). Observouse un maior risco de infección para secretores tipo A, pero esta diferencia non foi estatísticamente significativa (Degenhardt et al., 2022). Non obstante, ambas asociacións indican que a variabilidade xenética en *ABO-FUT2* media na infección por SARS-CoV-2, en liña co que xa se sabe para outro tipo de infeccións (Cooling, 2015). Así mesmo, a variación en nestes xenes foi recoñecida como un factor de risco da asma e de enfermidades respiratorias causadas por *S. pneumoniae* (Ahluwalia et al., 2020).

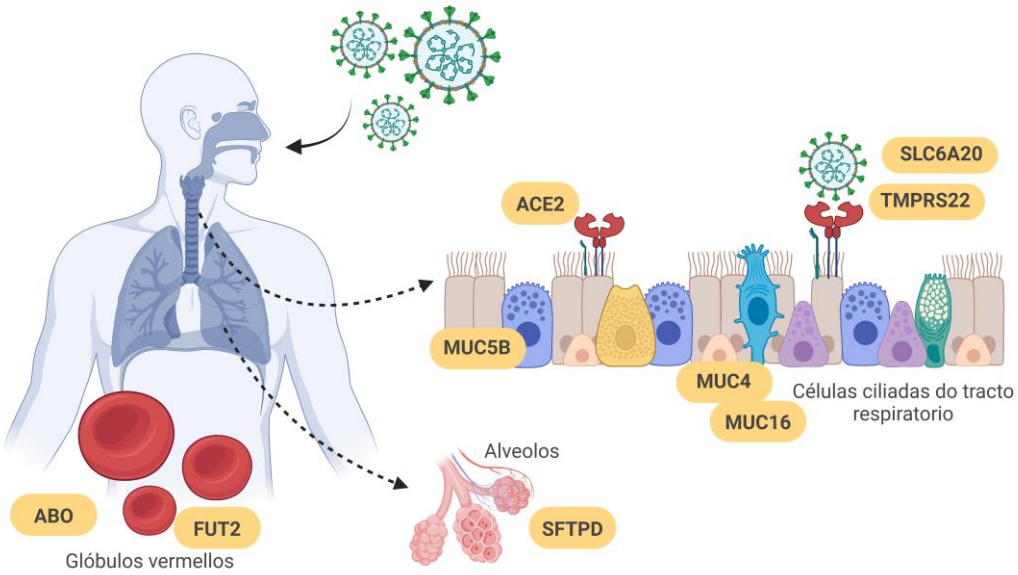


Figura 7. Xenes identificados a través de estudos GWAS implicados na infección e gravidade da COVID-19. Adaptada de Kanai e colaboradores (2023).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Os xenes codificadores de mucinas son outras das chaves do risco á infección. Estas proteínas glicosiladas están involucradas na primeira liña de defensa contra patóxenos, expresándose ao longo do tracto respiratorio, e compoñen o moco producido polas células epiteliais. A presencia de partículas estrañas inicia procesos inflamatorios, que estimulan a sobre-produción de mucinas para protexer o epitelio (Rose & Voynow, 2006). A mucina codificada polo xen *MUC5B*, asociada coa susceptibilidade en Pathak e colaboradores (2022), exprésase sobre todo en pulmón, e os cambios na súa expresión alteran a resposta contra os patóxenos. Concretamente, a supresión do xen *MUC5B* está relacionada cun estado de hipoxemia, unha maior presencia de partículas no tracto respiratorio inferior e acumulación de células inmunolóxicas en estado apoptótico (Roy et al., 2014). O SNP asociado en *MUC5B*, rs35705950, incrementa a expresión do mesmo xen en tecido pulmonar, o que é consistente co seu efecto protector contra a hospitalización (Pathak et al., 2022). Distintas mucinas como *MUC4* ou *MUC16* están asociadas ao fenotipo de susceptibilidade. Outro sinal de interese é a do xen *SFTPD*, que codifica unha proteína surfactante do pulmón. Para esta rexión identificouse unha variante *missense*, rs721917, que está asociada a un maior risco de hospitalización (Pathak et al., 2022).

A variabilidade do hóspede para os xenes *ACE2* e *TMPRSS2* tamén afecta á capacidade de entrada do virus nas células (**Figura 7**). A asociación do receptor *ACE2* reportouse por primeira vez a través do SNP rs190509934, o cal diminúe a expresión do xen nun 37% e o risco de infección nun 40% (Horowitz et al., 2022). Este resultado foi replicado polo estudo do HGI (Pathak et al., 2022). O receptor ACE2 interactúa co transportador de prolina ST1 codificado polo xen *SLC6A20* (Vuille-dit-Bille et al., 2015), que pertence á zona do cromosoma 3 asociada coa susceptibilidade á COVID-19 (**Anexo I**). *SLC6A20* exprésase principalmente en células do intestino delgado e ril, polo que non está claro o mecanismo biolóxico subxacente na infección, pero unha posible explicación tería que ver coa indución de citoquinas por parte dos transportadores de aminoácidos (Jacob, 2020). Por outro lado, o virus axúdase da proteasa de serina *TMPRS22* para a súa entrada na célula, se ben a variante líder da asociación en *TMPRS22*, rs9305744, posúe un efecto protector fronte á hospitalización (Kanai et al., 2023).

1.6.4.2 Xenes implicados na gravidade da COVID-19

A zona de risco localizada no cromosoma 3 inclúe varios xenes, entre os que se atopan varios receptores de quemoquinas (*CCR9, CXCR6* e *XCR1*) e un factor da transcripción tipo cremallera de leucina (*LZTFL1*). Aínda que o estudo do HGI e o de GenOMICC sinalaron aos xenes *CXCR6, CCR3* e *CCR2* como os causais, unha aproximación multi-ómica identificou a *LZTFL1* como o xen efector (Downes et al., 2021). O SNP rs17713054 é un intensificador (*enhancer*) que interactúa co promotor do xen *LZTFL1*. Á súa vez, a sobre-expresión de LZTFL1 inhibe a ruta de EMT (transición epitelial-mesenquimal), que reduce os niveis de *ACE2* e *TMPRSS2* nas células (Downes et al., 2021; Stewart et al., 2021). Os xenes *CXCR6, CCR9* e *XCR1* son receptores de citoquinas, que están involucradas na resposta inmune e no fenotipo de inflamación crónica da COVID-19. A integración dos datos de GWAS con datos de transcriptómica e empalme (*splicing*) resultou na selección de múltiples xenes na rexión, se ben só *LZTFL1* e *XCR1* coincidiron en ambas análises (Pathak et al., 2021). O mesmo estudo

achou un enriquecemento significativo das rutas de interacción de receptores citoquina-citoquina e de sinalización de quemoquinas. Como apuntan Niemi et colaboradores (2022), na zona hai máis dun xen candidato a ser causal. Polo seu lado, Kasela e colaboradores (2021) propoñen que a asociación do GWAS pode ser consecuencia dun único SNP con efectos pleiotrópicos en varios xenes (Kasela et al., 2021). Esta rexión do cromosoma 3 é herdada do xenoma Neanderthal e ten unha frecuencia case nula no este de Asia (Zeberg & Pääbo, 2020).

O complexo *OAS1-OAS3* localízase no cromosoma 12, e asociouse ao fenotipo de enfermidade crítica. Este clúster de xenes atopouse por primeira vez no GWAS de GenOMICC (The GenOMICC Investigators et al., 2021). Os xenes *OAS* son unha familia de oligoadenilato sintasas regulados por IFs e están involucradas na resposta viral e inmunidade innata. Detectan ARN viral de dobre cadea e activan á RNA-asa L, inhibindo a síntese proteica e impedindo a replicación do virus (U. Y. Choi et al., 2015; Kristiansen et al., 2010). Unha análise comparativa con ADN antigo determinou que este haplotipo é protector (reducindo nun 23% o risco de enfermidade grave por cada copia do mesmo) e que, ao igual que o do cromosoma 3, tamén deriva do xenoma Neanderthal, e proveeu evidencia de que se atopa baixo selección positiva (Zeberg & Pääbo, 2021). Mediante a técnica de mapeamento fino trans-poboacional identificouse rs10774671 como un posible SNP causal, localizado nun sitio aceptor de *splicing* no xen *OAS1* que produce a isoforma p46 (a ancestral) e para o que cada copia do alelo protector (G) aumenta a actividade da enzima (Huffman et al., 2022). A aleatorización mendeliana con cis-pQTLs (QTL proteicos) foi outra técnica que apuntou ao xen *OAS1*, revelando que os niveis de proteína en sangue de *OAS1* e da isoforma p46 estaban asociadas a un efecto protector fronte á COVID-19 (S. Zhou et al., 2021). En contraste, o mesmo estudo reportou que un aumento dos niveis da proteína *OAS3* ía asociado a unha enfermidade máis grave (S. Zhou et al., 2021). Xunto co rs10774671, a mutación substitutiva rs1131454 foi identificada como outra variante protectora, ambas regulando a expresión de *OAS1* (Banday et al., 2022). En nenos que sofren unha complicación da COVID-19 denominada síndrome inflamatorio multisistema (MIS-C) descubríronse variantes deletéreas homocigóticas nos xenes *OAS1*, *OAS2* e *ARNasa-L* (D. Lee et al., 2022).

Os interferóns teñen un papel fundamental na progresión da COVID-19, xa que inician e amplifican a resposta antiviral. O xen *IFNAR2* no cromosoma 21 asociouse ao risco de enfermidade crítica, xunto cun clúster de interferóns do cromosoma 9 (COVID-19 Host Genetics Initiative et al., 2021; Kanai et al., 2023; The GenOMICC Investigators et al., 2021). *IFNAR2* produce un receptor que forma un complexo xunto con *IFNAR1*, ao cal se unen os interferóns tipo I, iniciando unha cascada de sinalización de xenes estimulados por interferóns (Sadler & Williams, 2008). En doentes hospitalizados por COVID-19 identificáronse varias variantes con perda de función localizadas en *IFNAR2* (Smieszek et al., 2021). Por outro lado, a activación da ruta de sinalización estimulada por IFs JAK-STAT tras a unión ao receptor IFNAR1/IFNAR2 depende de dúas kinasas, *JAK* e *TYK2*, ambas identificadas tamén a través dos GWAS de COVID-19. O descubrimento da asociación de *TYK2* (The GenOMICC Investigators et al., 2021) propiciou o desenvolvemento dun estudo clínico coa droga baricitinib para o tratamento da COVID-19 (Abani et al., 2022). Algúns dos xenes activados por

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

interferóns son os que codifican as proteínas TRIM ou IRF, e varios deles como *TRIM46* e *IRF1* atopáronse asociados á COVID-19 crítica (**Anexo I**).

DPP9, que codifica unha proteasa de serina, é outro dos xenes implicados na progresión da enfermidade cuxa asociación foi replicada en máis dunha ocasión (COVID-19 Host Genetics Initiative et al., 2021; Kanai et al., 2023; Kousathanas et al., 2022; Pathak et al., 2022; The GenOMICC Investigators et al., 2021). O SNP líder, rs2109069, asociouse tamén a fibrose pulmonar (Fingerlin et al., 2013). Por outra parte, no xen *FOXP4* detectouse unha heteroxeneidade entre estudos, quizais atribuíble a unha diferencia de efectos entre poboacións (Pathak et al., 2022). A familia de factores transcripcionais FOXP están involucrados en múltiples procesos biolóxicos, especialmente en cancro (J.-H. Kim et al., 2019). Curiosamente, *FOXP4* é o primeiro xen asociado ao fenotipo de COVID persistente (*long COVID*), compartindo o mesmo haplotipo que no fenotipo de hospitalización, que á súa vez está asociado á expresión do xen en tecido pulmonar (Lammi et al., 2023). Lammi e colaboradores (2023) investigaron en profundidade o rol de *FOXP4* na enfermidade, determinando que este xen se expresa en células alveolares de tipo 2 e que o SNP líder está asociado a cancro de pulmón. Porén, os autores apuntan que a relación entre *FOXP4* e o fenotipo de COVID persistente non pode ser explicada a través da asociación con hospitalización ou enfermidade crítica.

1.6.5 O consorcio SCOURGE

En España, o consorcio máis grande constituído para o estudo da base xenética da COVID-19 foi SCOURGE (*Spanish COalition to Unlock Research on host GEnetics on COVID-19*, https://www.scourge-covid.org/). Este está coordinado dende Santiago de Compostela pero participan del máis de 30 centros de España e Latino-América. Recrutáronse máis de 10.000 doentes por COVID-19, sendo aproximadamente a metade doentes hospitalizados, dos que se obtivo información clínica e xenética. Os obxectivos principais eran a determinación das causas xenéticas que subxacen á infección e gravidade da enfermidade de COVID-19, pero tamén a comparación a nivel poboacional entre España e Latino-América. Este consorcio forma parte do consorcio HGI.

A presente investigación enmárcase dentro deste consorcio, correspondendo aos resultados da primeira e segunda análises dos datos (GWAS en España e GWAS en Latino-América), así como á comparación do risco xenético entre ambas cohortes.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

2 HIPÓTESES E OBXECTIVOS

Esta tese parte da hipótese inicial de que a COVID-19 é unha enfermidade complexa, cuxa base xenética é parcialmente responsable da variabilidade que existe entre os individuos durante a progresión da enfermidade. Ademais, a arquitectura xenética da enfermidade podería variar entre as poboacións, o que tería un impacto na busca dos determinantes subxacentes. Finalmente, postúlase que estas posibles diferencias poderían influír na predición polixénica do fenotipo, resultando nun rendemento inferior dos PRS en individuos non europeos e dificultando a construción dun PRS con utilidade clínica ou investigadora.

Establecemos, así, os seguintes obxectivos xerais e específicos:

1. Investigar a base xenética involucrada na enfermidade da COVID-19 a través de análises GWAS, empregando unha cohorte recollida en España e outra en varios países de Latino-América:
- Identificar as variantes asociadas á hospitalización por COVID-19 en ambas poboacións.
- Determinar se existe unha contribución da xenética distinta entre sexos e entre grupos de idade.
- Avaliar a relevancia funcional dos xenes identificados no GWAS a través da anotación xenómica e integración con datos de expresión.

2. Construír índices de risco xenético para predicir a hospitalización por COVID-19:
- Realizar unha análise comparativa entre ambas poboacións para determinar se os PRS construídos presentan diferencias na súa capacidade preditiva.
- Desenvolver e validar un PRS xeneralizable que poida ser aplicado a ambas cohortes.
- Determinar se o risco xenético asociado á enfermidade grave está asociado a outros resultados da enfermidade e investigar o seu potencial clínico.
- Investigar alternativas metodolóxicas para o adestramento e validación dos PRS en presencia de poboacións mixtas.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3 METODOLOXÍA

3.1 RECOLECCIÓN DOS DATOS

Dentro do marco do proxecto SCOURGE recolléronse datos xenéticos e clínicos de pacientes con diagnóstico COVID-19 positivo para analizar a variabilidade inter e intrapoboacional da enfermidade por COVID-19. O primeiro período de mostraxe comprendeu entre marzo e decembro de 2020, culminando na recollida de 11.939 mostras dende 34 centros de 25 cidades de España. No período de marzo de 2020 a xullo de 2021 levouse a cabo a recollida de datos en 13 centros repartidos en cinco países de Latino-América (México, Brasil, Colombia, Paraguai e Ecuador).

Esta colección de mostras e datos foron obtidos polos centros a través dos seus respectivos biobancos e sempre tras a sinatura dun consentimento informado xunto coa aprobación dos respectivos Comités Éticos e Científicos. O proceso de recolección e manexo dos mesmos realizouse a través das ferramentas de *REDCap*, localizado no Centro de Investigación Biomédica en Red (CIBER) (Harris et al., 2009, 2019). Todo o proxecto foi aprobado polo Comité Ético Galego (Ref. 2020/197). A recollida das mostras en México foi aprobada polo Comité Nacional de Investigación Clínica, do Instituto Mexicano del Seguro Social (México, protocolo R-2020-785-082).

Os individuos foron diagnosticados como positivos na COVID-19 a partir de probas PCR (*polymerase chain reaction*) ou de criterios clínicos e procedementos de laboratorio (test de anticorpos ou outras probas microbiolóxicas). Todos os casos están clasificados nunha escala de gravidade de cinco niveis (**Táboa 1**).

A maiores, incluíronse dúas cohortes de datos españolas con estado descoñecido sobre a COVID-19 para empregarse como controis poboacionais nalgunhas análises: 3.437 mostras do Banco Español de ADN (BNADN, https://www.bancoadn.org) e 2.506 mostras do consorcio GR@ACE (Moreno-Grau et al., 2019). Os controis do BNADN proveñen do Servicio Nacional de Sangre, e son individuos non emparentados con orixe española auto-informada e ausencia de historial de enfermidades familiares ou persoais, incluíndo cancro, enfermidades do sangue ou circulatorias, endocrinas, mentais ou do comportamento, infecciosas, nerviosas, da visión, da escoita, respiratorias, inmunolóxicas, óseas, conxénitas, da pel e dixestivas.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Táboa 1. Escala de gravidade de 5 niveis baixo a que se clasifican os doentes de SCOURGE. PaO_2: Presión parcial de osíxeno en sangue arterial; SaO_2: Saturación de osíxeno en sangue arterial; FiO_2: Fracción de osíxeno inspirado.

Nivel	Criterio clínico
Gravidade 0 (asintomática)	Asintomático
Gravidade 1 (leve)	Con síntomas, mais sen infiltración pulmonar ou necesidade de terapia de osíxeno
Gravidade 2 (moderada)	Con infiltración pulmonar afectando a <50% dos pulmóns ou necesidade de terapia de osíxeno
Gravidade 3 (grave)	Hospitalización con algún dos seguintes criterios: PaO2 < 65 mmHg or SaO2 < 90% PaO2/FiO2 < 300 SaO2/FiO2 < 440 Disnea Frecuencia respiratoria ≥ 22 rpm Infiltración que afecta a > 50% dos pulmóns
Gravidade 4 (crítica)	Ingreso na UCI ou necesidade de ventilación mecánica (invasiva ou non invasiva)

Finalmente, para os estudos en poboación europea usouse unha cohorte de replicación, á que nos referiremos como estudo CNIO (Centro Nacional de Investigacións Oncolóxicas) ao longo desta tese. Esta cohorte comprende 1.598 mostras con diagnóstico positivo en COVID-19 de seis biobancos diferentes (Biobanco CNIO, Biobanco Vasco, Biobanco Hospital Ramón y Cajal, Biobanco Hospital Puerta de Hierro, Biobanco Hospital San Carlos, and Banco Nacional de ADN), coa aprobación do comité de ética CEI PI 34_2020-v2. Como poboación control, incluíronse 1,068 individuos con estado descoñecido sobre a COVID-19 procedentes do BNADN.

Os datos clínicos foron obtidos para cada paciente a través dun caderno de recollida de datos. Os antecedentes clínicos dividíronse en 6 categorías, pero a presencia de cada un dos antecedentes especificouse de maneira individual: (1) factores de risco cardiovascular: hipertensión arterial, hipercolesterolemia, diabetes mellitus non complicada, diabetes mellitus con repercusión visceral, obesidade; (2) sistema cardíaco: cardiopatía isquémica, insuficiencia cardíaca, arritmia cardíaca, vasculopatía periférica; (3) sistema nervioso: enfermidade cerebrovascular (infarto ou hemorraxia) sen secuelas ou secuelas mínimas; (4) enfermidade cerebrovascular (infarto ou hemorraxia) con hemiplexía ou paraplexía, demencia, outra enfermidade neurolóxica; (5) sistema respiratorio: EPOC, outra enfermidade respiratoria crónica; (6) sistema dixestivo: úlcera péptida, hepatopatía crónica sen hipertensión vascular, hepatopatía crónica con hipertensión vascular; (6) antecedentes onco-hematolóxicos: tumor sólido localizado, tumor sólido metastásico, leucemia, linfoma, transplante de médula/precursores hematopoiéticos.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.2 XENOTIPADO

O xenotipado das mostras realizouse co *array* de xenotipado *Axiom Spain Biobank Array* ou SBA (*Thermo Fisher Scientific*) seguindo as instrucións do fabricante no Nodo de Santiago de Compostela do Centro Nacional de Xenotipado (CeGen-ISCIII). Este *array* inclúe 757.836 marcadores, incluíndo variantes raras seleccionadas en poboación española. O ADN xenómico foi extraído de sangue periférica e aillado empregando o *Chemagic DNA Blood100 kit* (*PerkinElmer Chemagen Technologies GmbH*), seguindo as recomendacións do fabricante.

Para as mostras do estudo do CNIO, procesouse un total de 250µg de ADN de acordo ao *Infinium HTS assay Protocol* (Part # 15045738 Rev. A, Illumina), incluíndo a amplificación, fragmentación e hibridación usando o *Global Screening Array Multi-disease* 3.0. Este array contén un total de 730.059 marcadores e foi escaneado nunha plataforma iScan (Illumina, Inc.) A formación de clústeres e a asignación de xenotipos foron realizados co Genome Studio v2.0.4 (Illumina, Inc.).

3.3 DEFINICIÓN DAS COHORTES

Nos estudos desta tese empregáronse dúas cohortes principais de individuos: unha cohorte española e unha cohorte latinoamericana. As cohortes foron analizadas en dous tempos diferentes (2020/2021 para a cohorte española, e 2021/2022 para a cohorte latinoamericana), dada a dispoñibilidade dos datos e tras resultados de diversas análises que se explicarán máis adiante. Por comodidade na lectura, resúmese neste apartado a súa definición.

Os principais criterios para delimitar as dúas cohortes foron o país de recollida dos datos, o país de procedencia auto-informado, e a ancestralidade xenética inferida. A cohorte española está composta por 15.045 individuos (9.371 positivos en COVID-19 e 5.674 controis poboacionais), con >80% de similitude xenética inferida ás poboacións Europeas de 1KGP. Na **Táboa 2** está a lista de centros que contribuíron a esta cohorte. As características clínicas e demográficas dos pacientes móstranse na sección de Resultados do Capítulo 4. En canto aos controis poboacionais, para os que só se dispón de información sobre o sexo e a idade, a porcentaxe de mulleres é do 46.3% cunha media de idade de 55.5 anos (desviación típica [s.d.]=16.2), e a de homes do 53.7% homes cunha media de idade de 51 anos (s.d.=13.04).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Táboa 2. Lista de centros españois que contribuíron ás mostras incluídas na cohorte española.

Centro	Cidade	Mostras
Biocruces Bizkaia	Baracaldo	266
Complexo Asistencial Universitario de León	León	308
Complexo Hospitalario Universitario de A Coruña	Coruña	307
Complexo Hospitalario Universitario de Santiago de Compostela	Santiago de Compostela	83
Hospital Universitario San Pedro de Logroño	Logroño	109
Centro Nacional de Microbiología, ISCIII	Madrid	707
Instituto de Investigación Sanitaria Hospital Clínico San Carlos	Madrid	620
Instituto de Investigación Sanitaria Fundación Jiménez Díaz[1, 2]	Madrid	2501
Hospital Universitario de Getafe	Getafe	112
Hospital 12 de Octubre	Madrid	236
Hospital Universitario Álvaro Cunqueiro	Vigo	149
Hospital Clínico Universitario de Valladolid	Valladolid	343
Hospital El Bierzo Ponferrada	Ponferrada	11
Hospital General de Segovia	Segovia	226
Hospital Universitario Nª Sra. Candelaria	Santa Cruz de Tenerife	83
Hospital Universitario de Salamanca	Salamanca	308
Hospital Santa Bárbara de Soria	Soria	67
Hospital Universitario Gregorio Marañón	Madrid	961
Centro de Genética Cardiovascular de Girona	Girona	184
Hospital Universitario y Politécnico La Fe	Valencia	88
Hospital Universitario de La Paz	Madrid	565
Hospital Universitari Mútua Terrassa	Terrasa	470
Hospital Universitario Río Hortega	Valladolid	356
Hospital Universitario Vírgen del Rocío	Sevilla	109
Instituto Murciano de Investigación Biosanitaria (IMIB-Arrixaca)[2,3]	Murcia / Cartagena	562
IIB Sant Pau	Barcelona	444
Instituto de Investigación Biomédica de Bellvitge	Barcelona	235
INDITEX	Arteixo	83
Hospital Universitario de Móstoles	Móstoles	191
Fundación Rioja Salud	Logroño	320
Hospital Marques de Valdecilla	Santander	332
Universidad Autónoma de Madrid	Madrid	34
Universidad de León	Leon	320
Hospital Universitario Virgen de las Nieves	Granada	249

[1] Incluíndo o Hospital Universitario Fundación Jiménez Díaz e o Hospital Universitario Infanta Elena

[2] Stop Coronavirus Initiative

[3] Incluíndo os Hospitais Universitarios Virgen de la Arrixaca, Morales Meseguer e Reina Sofía en Murcia e Santa Lucía en Cartagena

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

A cohorte latinoamericana (cohorte LatAm) contén 3.729 casos positivos de COVID-19 recrutados en Latino-América (**Táboa 3**). Os individuos foron diagnosticados con COVID-19 en base á PCR (79.3%) ou de acordo a criterios clínicos (2.2%) ou outros métodos de laboratorio (test de anticorpos: 16.3%; outros test microbiolóxicos: 2.2%). A maiores, nesta cohorte están incluídos 1.082 casos positivos de COVID-19 que foron recrutados en España entre marzo e decembro de 2020 que, ou ben teñen orixe >80% de ancestralidade AMR (estimada a partir da poboación AMR de 1KPG), ou teñen unha auto-informada dun país latino-americano, ou presentan unha mestura de GIA entre AMR, EUR e AFR (con < 0.5% SAS/EAS). Do total de mostras, 470 caeron nos controis de calidade e 363 non dispoñían de información fenotípica. Ademais, excluíronse 473 mostras cuxa compoñente europea superaba o 80% para reducir a carga da similitude a poboacións europeas, quedando un total de 3.512 individuos analizables. Dada a ausencia dunha poboación control xeneticamente parecida aos nosos casos, todas as mostras desta cohorte son positivas en COVID-19. As características clínicas e demográficas dos individuos incluídos nas análises poden verse na sección de Resultados do Capítulo 5.

Táboa 3. Lista de centros latinoamericanos e españois que contribuíron ás mostras incluídas na cohorte LatAm.

Centro	Cidade	País	Mostras
Universidad federal de Pará	Belem	Brazil	229
Universidade Federal da Paraiba	Joao Pessoa	Brazil	26
Federal University of Pernambuco	Recife	Brazil	212
Universidad Federal do Rio Grande do Norte	Natal	Brazil	515
Universidad Brasilia	Brasilia	Brazil	165
Clínica Comfamiliar	Pereira	Colombia	176
Universidad Simón Bolívar	Barranquilla	Colombia	182
SIGEN Alianza Universidad de los Andes-Fundación Santa Fe de Bogotá	Bogotá	Colombia	133
Fundación Universitaria de Ciencias de la Salud	Bogotá	Colombia	397
Universidad Tecnológica Equinoccial	Guayaquil	Ecuador	15
Instituto tecnológico de Monterrey	Monterrey	Mexico	346
Universidad de Guadalajara	Guadalajara	Mexico	33

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Táboa 3. Continuación.

Centro	Cidade	País	Mostras
Centro para el Desarrollo de Investigación Científica (CEDIC)/ UNCA	Asunción	Paraguay	80
Biocruces Bizkaia	Baracaldo	España	2
Complejo Asistencial Universitario de León	Leon	España	3
Complexo Hospitalario Universitario de A Coruña	Coruña	España	7
Complejo Hospitalario Universitario de Santiago de Compostela	Santiago de Compostela	España	3
Centro de Investigación Biomédica de la Rioja	Logroño	España	5
Centro Nacional de Microbiología, ISCIII	Madrid	España	114
Instituto de Investigación Sanitaria Hospital Clínico San Carlos	Madrid	España	25
Instituto de Investigacion Sanitaria Fundación Jiménez Díaz[1 , 2]	Madrid	España	271
Hospital Universitario de Getafe	Getafe	España	2
Hospital 12 de Octubre	Madrid	España	118
Hospital Clínico Universitario de Valladolid	Valladolid	España	1
Hospital General de Segovia	Segovia	España	1
Hospital Universitario Nª Sra. Candelaria	Santa Cruz de Tenerife	España	1
Hospital Universitario de Salamanca	Salamanca	España	2
Hospital Santa Bárbara de Soria	Soria	España	1
Hospital Universitario Gregorio Marañón	Madrid	España	71
Centro de Genética Cardiovascular de Girona	Girona	España	17
Hospital Universitario y Politécnico La Fe	Valencia	España	10
Hospital Universitario de La Paz	Madrid	España	35
Hospital Universitari Mútua Terrassa	Terrasa	España	39
Hospital Universitario Río Hortega	Valladolid	España	2
Hospital Universitario Vírgen del Rocío	Sevilla	España	1
Instituto Murciano de Investigación Biosanitaria (IMIB-Arrixaca)[2,3]	Murcia / Cartagena	España	107
IIB Sant Pau	Barcelona	España	64
Instituto de Investigación Biomédica de Bellvitge	Barcelona	España	29
INDITEX	Arteixo	España	3
Hospital Universitario de Móstoles	Móstoles	España	18
Fundación Rioja Salud	Logroño	España	1
Hospital Marques de Valdecilla	Santander	España	11
Universidad de León	Leon	España	18
Hospital Universitario Virgen de las Nieves	Granada	España	4

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

3.4 CONTROL DE CALIDADE DOS DATOS XENÉTICOS

Os datos xenéticos deben someterse a un control de calidade previo ás análises co obxectivo de reducir potenciais problemas derivados da recollida de mostras, procedementos de laboratorio ou do algoritmo de procesado das mostras. Unha explicación detallada dun fluxo de traballo *tipo* pode atoparse en Marees e colaboradores (2018) e Turner e colaboradores (2011). O control de calidade (*quality control* ou QC) dos datos empregados nesta tese realizouse co *software* PLINK (Purcell et al., 2007), versións 1.9 e 2.0. Ambas cohortes foron sometidas a un procedemento similar, mais na cohorte LatAm fixéronse algunhas variacións.

En primeiro lugar, eliminouse unha lista de marcadores de baixa calidade obtida de estudos previos sobre a calidade do SBA (xa excluídos na versión 2 do array). Respecto ao QC dos marcadores, elimináronse variantes cuxa frecuencia do alelo menor fose <1% ou cuxa taxa de xenotipado (*call rate*) fora <98%, as variantes cuxa diferencia na taxa de xenotipado entre casos e controis fora >0.02, e as variantes que se desviaran do valor esperado baixo o equilibrio Hardy-Weiberg (cun p-valor de $<1x10^{-10}$ en casos e $<1x10^{-06}$ en controis). En canto ás mostras, excluíronse aquelas que presentaran >2% de variantes ausentes e aquelas cuxa heterocigosidade superara en 5 desviacións típicas á heterocigosidade media da cohorte. En relación á exploración do parentesco, neste estudo eliminamos un individuo de cada par cando o valor de PI_HAT >0.25 (parentesco de segundo grao). O cómputo de IBD en PLINK parte sempre dun subgrupo de SNPs autosómicos e independentes, os cales obtivemos tras eliminar rexións con alto desequilibrio de ligamento (Price et al., 2010) e de realizar unha poda (*LD-pruning)* dos SNPs cos seguintes parámetros: ventá de 1.000 SNPs, con tamaño de paso de 80 e un límite de r^2 de 0.1. Ademais, co mesmo subgrupo de marcadores, calculáronse con PLINK as 10 primeiras PCs e explorouse a existencia de datos atípicos ou *outliers* (mostras desviadas máis de 5 s.d. da media do PC). O control de calidade foi repetido tras a exclusión dos parentes e os *outliers*. Tamén se realizou outra volta do PCA, tras a cal foron seleccionadas as primeiras 10 PCs para ser incluídas nas análises.

Na cohorte de Latino América adaptamos certos aspectos do control de calidade. Explorouse de forma visual a heterocigosidade para descartar posible contaminación, pero non se eliminaron individuos por déficit de heterocigosidade. Por outro lado, neste tipo de poboacións a estima do IBD pode estar confundida pola ancestralidade (Colistro et al., 2021), de xeito que poden atoparse individuos supostamente emparentados (PI_HAT >0.25) cun número alto doutros individuos. Nestes casos, os valores de Z0, Z1 e Z2, mostran patróns pouco coherentes cos valores teóricos do parentesco. Estas mostras non foron eliminadas no QC por cuestións de IBD. Por último, a análise de PCA empregouse como unha ferramenta de visualización e non para a detección de *outliers* ou asignación de individuos á poboacións.

3.5 ESTIMACIÓN DA ANCESTRALIDADE XENÉTICA INFERIDA GLOBAL

Para a estimación da GIA empregouse o *software* ADMIXTURE (Alexander et al., 2009). Este *software* estima exclusivamente as proporcións de ancestralidade global, non as locais.

A base do modelo de ADMIXTURE consiste en modelar as probabilidades duns xenotipos observados a partir das frecuencias alélicas e as proporcións de ancestralidade. Pártese dun

USC DE COMPOSTELA

conxunto de individuos non emparentados e independentes para os cales observamos uns xenotipos, que proveñen dunha mestura de *K* poboacións. Cada unha desas poboacións vai contribuír a unha fracción do xenoma de cada individuo, e este será un dos parámetros a estimar. Deste xeito, un individuo *i* terá unha determinada proporción q_{ki} de cada unha das K poboacións. Ao contrario que outros *software*, non precisa da presencia de marcadores informativos da ancestralidade, senón que emprega un set de marcadores aleatorios e independentes (Alexander et al., 2009). Estas estimacións son complementarias ao PCA: unha comparación práctica de ambas técnicas pode atoparse en Liu e colaboradores (2020).

ADMIXTURE emprega un algoritmo de máxima verosimilitude para maximizar a función de verosimilitude da mostra $L(Q, F)$. Q e F son as matrices dos parámetros das fraccións de ancestralidade e das frecuencias alélicas, respectivamente, con dimensións $I \times K$ e $K \times J$, onde *K* é o número de poboacións ancestrais e *J* o número de SNPs. Os valores destas matrices vanse actualizando en cada iteración ata atopar un óptimo.

ADMIXTURE pode correrse en modo supervisado e en modo non supervisado. No modo non supervisado, o usuario indica un número K de poboacións e o algoritmo constrúe unhas matrices nas que a dimensión *K* toma ese valor. O procedemento habitual ao seleccionar este modo é correr o algoritmo con diferentes valores de *K* e quedarse co que menor erro de validación cruzada teña. Como apunte, nalgunhas ocasións pode ser que certos valores de *K* teñan erros de validación cruzada moi parecidos. Nese caso sempre compensa quedarse co modelo máis sinxelo, é dicir, o que menor número de poboacións asuma.

O segundo modo é o modo supervisado, para o cal se precisa dun panel de referencia cuxos individuos sexan representativos das poboacións das que proveñen. O algoritmo procederá entón a asignar unha fracción de case o 100% de pertenza á súa poboación aos individuos do panel de referencia, e asignará aos individuos con ancestralidade descoñecida unhas proporcións en función da similitude a eses de referencia. Por comodidade na interpretación, este foi o método escollido para as análises da tese.

Os resultados de ADMIXTURE son dous arquivos, un co sufixo .P que contén as frecuencias alélicas estimadas de cada SNP en cada poboación *K*, e un co sufixo .Q coas proporcións de ancestralidade estimadas de cada individuo.

Para a cohorte europea empregouse o total dos individuos de 1KGP (Auton et al., 2015) como panel de referencia. Ambas cohortes combináronse con PLINK a través dos marcadores que tiñan en común, e sometéronse a *LD-pruning* para obter un set de marcadores independentes. Sobre este conxunto de datos aplicouse o algoritmo de ADMIXTURE no modo supervisado. Así, as mostras da cohorte de SCOURGE foron asignadas con q_{ik} probabilidade a K=5 superpoboacións: Europa, África, Asia do Este, Asia do Sur e América (referido a individuos de ancestralidade latinoamericana mixta). Como nesta análise só nos interesaban os individuos con alta probabilidade de pertencer a poboacións de orixe xeográfica ancestralmente europea, establecemos un criterio arbitrario para excluír mostras cuxa pertenza ao clúster EUR fora <80%.

Pola contra, coa cohorte Latinoamericana o uso de ADMIXTURE non tiña como obxectivo crear un grupo homoxéneo de mostras, senón simplemente a súa caracterización (salvo no caso das mostras ≥80% EUR). Para esta cohorte, dada a súa historia demográfica, só usamos 3

poboacións de referencia como contribucións xenéticas, e corremos dúas veces o algoritmo para refinar a estimación. En primeiro lugar tomamos unha mostra aleatoria de 79 individuos de Europa (1KGP), 70 de individuos de Africa (1KGP) e 79 individuos Nativo-Americanos (Mao et al., 2007) combinándoos a través dos SNPs bialélicos comúns e facendo un *LD-pruning*. Sobre este conxunto corremos ADMIXTURE co modo sen supervisar e K=3. Isto serviunos para homoxeneizar os clústeres tras a eliminación dos individuos que tiñan <95% de pertenza ao seu clúster (20 Africanos, 18 Europeos e 38 Nativo-Americanos) e permitiunos construír un panel de referencia máis refinado. Nun segundo paso combinamos este novo panel de referencia coas mostras da cohorte de Latino-América e fixemos unha análise de ADMIXTURE supervisada. Do total de individuos, 471 tiñan unha probabilidade ≥80% de pertencer ao clúster europeo, polo que decidimos excluílos e reducir deste xeito o peso da contribución xenética EUR.

3.6 IMPUTACIÓN DE XENOTIPOS

Os *arrays* de xenotipado cobren un gran número de variantes ao longo de todo o xenoma (p.e. ao redor de 800.000 co SBA). Sen embargo, estímase que o número total de SNPs supera os 80 millóns (Auton et al., 2015), o que quere dicir que a gran maioría das variantes quedarían sen analizar, limitando o poder estatístico do estudo. A solución a este problema é a inferencia e predición dos xenotipos que denominamos perdidos (non observados) a través dun paso nomeado imputación.

Inicialmente pártese de dous conxuntos de individuos: a) dunha mostra obxecto de estudo na cal observamos uns xenotipos (variantes xenotipadas) e será a que vaiamos imputar; e b) dun panel de referencia, que son os individuos para os cales existe información xenética extensa e que son representativos dunha gran porcentaxe da poboación, como por exemplo o panel de 1KGP (Auton et al., 2015) ou *HapMap* (Gibbs et al., 2003) cos cales se compararán os individuos do noso estudo. O primeiro paso consiste na inferencia do haplotipo (*phasing*) para os individuos da mostra de estudo; é dicir, determínase con procedementos estatísticos que alelos proveñen do cromosoma materno e cales do paterno, recuperando o estado dos cromosomas antes do proceso de recombinación. A continuación, compáranse os xenotipos observados (xenotipados) de cada individuo cos xenotipos dun panel de referencia que elixamos a partir dos bloques de desequilibrio de ligamento e predinse os que faltan. O resultado da imputación son as probabilidades de ter un e outro alelo para un marcador, tamén chamadas dose alélica (*allele dosage*). Pódese traballar directamente con estas probabilidades ou convertilas directamente a xenotipos.

Na actualidade existen multitude de algoritmos de *phasing* e imputación; os máis importantes están revisados en Marchini e Howie (2010). A precisión da imputación dependerá de moitos factores, como a frecuencia alélica dos xenotipos, a diversidade do panel de referencia, a densidade do *array* de xenotipado ou o tamaño da mostra de estudo (Das et al., 2018). As cohortes do noso estudo imputáronse individualmente a través da plataforma *TOPMed Imputation Server* (https://imputation.biodatacatalyst.nhlbi.nih.gov/), cuxo *host* é a Universidade de Michigan. Este servidor reúne na nube os *software* Eagle v2.4 (Loh et al., 2016) para o faseado e Minimac4 (Das et al., 2016; Fuchsberger et al., 2015), xunto cunha

pipeline de control de calidade dos datos e post-procesado. Esta plataforma simplifica o proceso de forma que os únicos pasos consiste en preparar os datos e enviar o traballo. O panel de referencia empregado é o de TOPMed versión R2 en hg38 (Taliun et al., 2021), que reúne 97.256 mostras secuenciadas en máis de 300 millóns de posicións.

Minimac4 emprega como medida de precisión a R^2 de imputación, que é o valor estimado da correlación entre os xenotipos imputados e os verdadeiros xenotipos (non observados). A plataforma de *TOPMed Imputation Server* permite o filtrado das variantes cunha $R^2 < 0.3$, polo que só se incluíron no GWAS as variantes cunha $R^2 \geq 0.3$ e cunha MAF > 0.01.

3.7 ANÁLISES DE ASOCIACIÓN DE XENOMA COMPLETO (GWAS)

O método estatístico empregado para os GWAS é o modelo de regresión, loxística para os fenotipos binarios, e lineal para os fenotipos continuos. Nestes modelos inclúense covariables demográficas como o sexo e idade, comorbilidades ou outras variables clínicas e socioeconómicas de interese, e as compoñentes principais para controlar a estrutura poboacional. Por simplicidade explicaremos o procedemento a través dun estudo caso/control.

Como xa se comentou no apartado do control de calidade, existen múltiples factores que poden confundir os estudos e provocar a aparición de falsas asociacións. Nos últimos anos o modelo de regresión simple foi substituído polo modelo loxístico mixto para ter en consideración as diferentes estruturas presentes nos datos. Nun modelo mixto asúmese unha agrupación dos individuos nun nivel superior, sendo estes grupos unha mostra aleatoria do número de grupos existentes na poboación. Isto permite que exista certa dependencia entre os individuos no estudo. Estes modelos foron adaptados da xenética de crianza, principalmente de Henderson (1984) á xenética humana. A novidade dun modelo mixto é que inclúe un termo de efecto aleatorio para todo o xenoma a través da matriz de relacións xenéticas (*genetic relationship matrix*, GRM), permitindo que este termo aleatorio recolla os factores confusores derivados do parentesco críptico e outras estruturas poboacionais (Sul et al., 2018). O modelo para un SNP *m* nun individuo *i* sería o seguinte:

$$y_i = \alpha X_i + \beta_m G_{mi} + u_i + \epsilon_i,$$

onde y_i é o valor do fenotipo Y para o individuo *i*; αX_i é o vector de efectos fixos (covariables); $\beta_m G_{mi}$ é o xenotipo do snp *m* no individuo *i* co seu coeficiente asociado (β); ϵ_i é o vector de erros asociados a factores medioambientais que son independentes entre os individuos, e u_i é o vector de erros aleatorios. Este último termo é o que recolle a varianza xenética aditiva de todos os SNPs a través da GRM. O modelo xenético clásico empregado nos estudos GWAS é o aditivo, de forma que a variable SNP toma os valores 0, 1 ou 2 en función seu xenotipo (0 para o xenotipo homocigoto do alelo frecuente, 1 para o heterocigoto e 2 para o xenotipo homocigoto do alelo menos frecuente). Zhou e colaboradores (2018) introduciron o modelo SAIGE, que ademais de implementar un modelo mixto a escala GWAS de forma eficiente (a través do uso dunha matriz *sparse* como GRM) tamén contempla o desequilibrio entre casos e controis á hora de computar os p-valores. Esta implementación consiste en (1) a estimación dun modelo nulo para calcular os coeficientes das covariables e os compoñentes da varianza, e (2) no test de asociación de cada SNP co fenotipo. No primeiro dos pasos, estímanse

os efectos aleatorios para cada individuo, e empregando un set de marcadores aleatorios calcúlase a *ratio* de varianzas do estatístico de asociación (con/sen as compoñentes da varianza aleatoria). No segundo paso, esta *ratio* emprégase para axustar os estatísticos de cada SNP.

O modelo SAIGE pódese aplicar a través do paquete de R SAIGE (Zhou et al., 2018), cuxa documentación se pode consultar en https://saigegit.github.io/SAIGE-doc/, ou a través do paquete SAIGEgds (X. Zheng & Davis, 2021). Este último paquete, a maiores da estimación do modelo nulo e do paso da asociación, require crear un obxecto GDS (*CoreArray Genomic Data Structure*) para os xenotipos.

3.7.1 Estudo GWAS na cohorte latinoamericana

As poboacións con mestura xenética recente levan consigo unha dificultade maior á hora de analizalas. Recapitulando o que se comentou antes, presentan patróns de desequilibrio de ligamento moi diferentes aos de poboacións máis homoxéneas e os seus tractos cromosómicos son de distinta orixe ancestral. Ademais, as diferentes ascendencias ancestrais orixinan relacións descoñecidas entre os individuos, de forma que algúns están máis relacionados xeneticamente que outros.

Unha das opcións clásicas para analizar poboacións diversas é a de clasificar aos individuos en grupos continentais ou etnias, xa ben sexa por cuestións sociais/xeográficas ou a través de algoritmos de clasificación supervisados e non supervisados como ADMIXTURE. Porén, en poboacións mixtas esta clasificación é imposible e son excluídas das análises.

A segunda opción consiste en analizar os individuos de forma conxunta empregando modelos mixtos. A pesar de que non existe un consenso sobre que metodoloxía é a mellor e de que este campo segue en constante exploración, si existen estudos que demostran o poder deste método á hora de controlar a estrutura poboacional, como o estudo PAGE (Wojcik et al., 2019). Este consorcio comparou os dous deseños e determinou que a análise conxunta mantiña unha proporción razoable de falsos positivos á vez que maximizaba a potencia estatística do GWAS. Como xa se adiantou, os modelos mixtos xunto cos PCs parecen capturar correctamente a estrutura poboacional. Esta foi a estratexia seleccionada para a análise GWAS na cohorte latinoamericana. Posteriormente á nosa análise, publicouse unha comparación de métodos para GWAS en poboacións mixtas que destacou SAIGE por enriba de GMMAT e Tractor no control do erro tipo I (Z. Yang et al., 2024), validando a nosa aproximación. Se ben Z. Yang e colaboradores (2024) tamén apuntan que Tractor ten máis potencia estatística para detectar sinais con heteroxeneidade entre as compoñentes ancestrais (é dicir, específicas de ancestralidade), o noso tamaño de mostra é limitado e non hai dispoñibilidade doutros estudos GWAS de COVID-19 realizados con Tractor cos que combinar os resultados.

3.8 META-ANÁLISE DE ESTUDOS GWAS

Dous dos desafíos aos que se enfronta o campo dos estudos GWAS son a reproducibilidade e a falta de potencia para detectar asociacións, e ambos están moi relacionados. De feito, o fracaso na replicación de asociacións foi un gran problema antes da chegada dos estudos GWAS, especialmente cos estudos de xenes candidatos (Hirschhorn & Altshuler, 2002). Entre

USC UNIVERSIDADE DE COMPOSTELA

outros motivos están a dificultade para estudar os mesmos SNPs, a falta de estandarización da metodoloxía, ou a baixa potencia estatística para detectar verdadeiras asociacións entre os xenotipos e os fenotipos (Kraft et al., 2009). As variantes que adoitan estudarse nun GWAS teñen efectos pequenos sobre o fenotipo, e para detectalas necesítase dun gran número de individuos, o que as veces non é viable. Isto leva á publicación de estudos GWAS sen potencia, con resultados pouco fiables ou que non chegan a ser estatisticamente significativos. A pesar de todos os avances, a reproducibilidade e o tamaño de mostra seguen a ser factores limitantes no descubrimento de novas asociacións.

As meta-análises de estudos GWAS supoñen unha solución a ambos retos, xa que permiten combinar os resultados de estudos independentes, impulsando a potencia estatística dos GWAS. En contraposición aos GWAS conxuntos ou *joint* GWAS, que analizan dunha única vez miles de individuos, as meta-análises non requiren que o investigador que realiza a análise dispoña dos datos crús, esquivando as dificultades relativas á protección de datos clínicos e xenéticos. Ao contrario, só requiren dos resultados dos estudos en forma de estatísticos (*summary statistics)*, polo que constitúen unha ferramenta moi útil de cara a aumentar o tamaño de mostra (Evangelou & Ioannidis, 2013).

Os *summary statistics* que se combinan nunha meta-análise deben dispoñer como mínimo das seguintes medidas: o identificador do SNP (ou o cromosoma, posición e alelos), o tamaño do efecto (o coeficiente β de regresión asociado ao SNP) e o erro estándar asociado ao mesmo, o alelo de efecto, e habitualmente as columnas de frecuencia alélica, p-valor da asociación, tamaño do estudo (N) e calidade da imputación. Así, o resultado final será unha estima do efecto do SNP a partir de todos os estudos, xunto co seu erro estándar e o p-valor asociado ao test de asociación. Os modelos máis coñecidos en meta-análises son o modelo de efectos fixos e o modelo de efectos aleatorios. O primeiro asume que todos os estudos proveñen dunha mesma poboación e o segundo asume que cada estudo é unha mostra dun conxunto de poboacións, contemplando unha varianza entre estudos, polo que non estiman o mesmo parámetro (Munafò & Flint, 2004).

Nos estudos GWAS, o máis empregado é o modelo de efectos fixos, que estima o efecto do SNP na poboación asumindo que é en todas o mesmo. Os efectos fixos teñen a vantaxe de que maximizan o poder estatístico, mais a contrapartida é que requiren dunha coidadosa exploración da heteroxeneidade entre estudos. Factores como a definición do fenotipo, as diferencias en LD ou a distribución das frecuencias alélicas entre poboacións poden dar lugar a unhas estimacións inválidas do efecto (Evangelou & Ioannidis, 2013; Munafò & Flint, 2004). En canto a como se combinan os estudos, o método máis común é o chamado *Inverse-Variance-Weighted-Average* (IVW), onde o β de cada estudo ten un peso proporcional á inversa do cadrado do erro estándar. Un GWAS cun tamaño de mostra pequeno terá unha estimación do efecto moito menos fiable que un GWAS máis grande, polo que o erro estándar será maior e terá menos peso na meta-análise.

O IVW foi o método empregado ao longo desta tese a través do *software* METAL (Willer et al., 2010). Con respecto á heteroxeneidade dos estudos, METAL realiza o test Q de Cochran e computa o estatístico I2 (que estima a proporción de variabilidade na meta-análise explicada por diferencias entre os estudos e non polo erro de mostraxe).

3.9 ANOTACIÓN FUNCIONAL

Unha vez detectadas as variantes involucradas no fenotipo cómpre determinar cal é a relación entre eles. A gran maioría das variantes que se descobren a través dos estudos de asociación non teñen unha función codificante e atópanse en rexións reguladoras ou interxénicas, facendo imposible saber de forma directa cal é a súa responsabilidade na enfermidade ou carácter. Ademais, recordemos que nos estudos GWAS o equilibrio de ligamento ten un papel fundamental e, habitualmente, a variante *hit* non é a variante causal senón que está en LD con ela, dificultando aínda máis o proceso de inferencia funcional (Gallagher & Chen-Plotkin, 2018).

Grazas á *revolución ómica*, agora dispoñemos dunha cantidade inmensa de datos xenéticos: datos de expresión, de epixenética, de proteómica e metabolómica, de interaccións na cromatina, etcétera. Toda esta información debe integrarse de forma lóxica para poder entender mellor a etioloxía das enfermidades, e esta é a misión dunha rama moi ampla da bioinformática. É despois de obter resultados no GWAS cando se combinan diferentes tipos de datos, xa ben sexa co que chamamos anotación ou realizando outro tipo de análises post-GWAS (Hasin et al., 2017), como algúns que cubriremos nos seguintes apartados. O paso inicial, e quizais máis crítico, consiste en detectar o xen causal, e será partir de aquí cando poderán integrarse novos sets de datos e análises en vivo. É interesante, por exemplo, saber se unha variante asociada a un fenotipo está asociada tamén a expresión dun xen en certo tecido, sinalando un posible papel regulador (Gallagher & Chen-Plotkin, 2018). No estudo de Hasin e colaboradores (2017) revísase o caso do xen *FTO*, un exemplo de integración de GWAS, epixenómica, transcriptómica, filoxenética e estudos *CRISPR*.

A fase de anotación é un primeiro paso de cara a contextualizar bioloxicamente os nosos resultados, e consiste en asignar información coñecida e pública ao noso conxunto de *loci*, permitindo unha priorización dos xenes. Este método non devolve un resultado estatístico, só funciona como unha libreta de notas. Unha das plataformas web para a anotación de GWAS máis utilizadas é FUMA (Watanabe et al., 2017), que contén información de 18 bases de datos biolóxicos.

FUMA divídese en dous grandes bloques de funcións, SNP2GENE e GENE2FUNCTION. En SNP2GENE anótanse os SNPs do GWAS que cumpran as condicións que o investigador elixa e os que están en LD. Entre outras cousas, os SNPs asígnanse a xenes en base as súas posicións, a se son *loci* de expresión (eQTL) ou a se provocan cambios na conformación da cromatina. Ademais, tamén se inclúe información sobre a súa puntuación nas escalas RegulomeDB (funcións reguladoras) e CADD (*Combined Annotation Dependent Depletion*, que determina o grado no que a variante é deletérea), ou do *GWAS Catalog* para informar de posibles asociacións con outros fenotipos. Finalmente, permite transferir a lista de xenes priorizados a GENE2FUNCTION para determinar se están enriquecidos nalgunha función ou tecido. A plataforma de FUMA permite facer automáticamente un test baseado en xenes (*gene-based test*) co *software* MAGMA (de Leeuw et al., 2015) que combina os resultados da asociación para todos os SNPs pertencentes a un mesmo xen.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Nesta tese empregamos as funcións de SNP2GENE, cos seguintes parámetros de selección de SNPs líder e SNPs candidatos: (1) p-valor para os SNPs líder $<5x10^{-08}$; (2) p-valor máximo para os SNPs anotados $<1x10^{-05}$; (3) límite r^2 para definir SNPs independentes $\geq$ 0.6; (4) límite de r^2 para definir SNPs líder $\geq$ 0.1; (5) MAF $\geq$0, e (6) máxima distancia entre bloques de LD para definir un *locus*=250kb. Para a meta-análise de AMR seleccionouse como panel de referencia AMR de 1KGP, e para a meta-análise de España, o panel do UKBiobank (r2). En canto ao mapeamento posicional, usouse unha distancia máxima de 10kb e non se considerou ningún filtro a mariores. O mapeamento por eQTL realizouse cos datos de GTEx v8 de linfocitos, sangue, pulmón e mucosa do esófago, e só para os pares xen-tecido significativos.

3.10 ESTATÍSTICA BAIESIANA VS ESTATÍSTICA FRECUENTISTA

Cada vez é máis frecuente o uso de inferencia baiesiana en análises xenéticos. Por tanto, cómpre introducir brevemente os conceptos básicos para poder entender mellor algúns dos métodos que empregaremos a continuación.

A estatística frecuentista entende a probabilidade como unha frecuencia repetida tras moitas repeticións dun mesmo evento. Baixo esta premisa, enténdese que a frecuencia dun evento se estabiliza despois de moitas repeticións, sendo o parámetro unha cantidade fixa na poboación. A partir de aquí, desenvólvense test de hipóteses e calcúlanse intervalos de confianza para este parámetro (Bland & Altman, 1998). Pola contra, para a escola baiesiana a probabilidade é o "grao de crenza" que un investigador ten sobre algo. Contempla que os parámetros poboacionais proveñen dunhas distribucións de probabilidade, considerándoos variables aleatorias. Na inferencia baiesiana pártese sempre dun coñecemento sobre o evento, ou dunhas hipóteses *a priori* (distribucións de probabilidade) subxectivas, e empréganse os datos para modificalas en pro dunhas distribucións *a posteriori* (Bolstad & Curran, 2016). A actualización destas "crenzas" a priori faise a través do teorema de Bayes:

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)}.$$

P(A|B) é a probabilidade condicionada do evento A dado B verdadeiro, P(A) é a probabilidade de que ocorra o evento A, P(B) é a probabilidade de que ocorra o evento B, e P(B|A) é a probabilidade condicionada de que ocorra o evento B dado A verdadeiro. A idea detrás desta formulación é a probabilidade inversa, ou como estimar a probabilidade dun evento a partir de eventos pasados (van de Schoot et al., 2021).

A fórmula anterior pode simplificarse da seguinte forma:

$$p(\theta|y) \propto p(y|\theta) * p(\theta),$$

onde $p(\theta|y)$ é unha probabilidade condicional sobre o parámetro θ, condicionado a os datos (y). Por outro lado, $p(y|\theta)$ é a probabilidade condicionada dos datos observados dado o parámetro θ, e p(θ) é a distribución a priori dos valores do parámetro na poboación (van de Schoot et al., 2021).

En Etz (2018) defínese a verosimilitude (*likelihood*) dunha hipótese dados uns datos como unha función proporcional (multiplicada por unha constante K) á probabilidade de obter os datos dado que a hipótese sexa verdadeira. No caso da inferencia, a hipótese é o valor dun

USC DE COMPOSTELA

parámetro. Este concepto é un dos puntos críticos entre a estatística frecuentista e a baiesiana. Na inferencia frecuentista, a estimación por máxima verosimilitude consiste en atopar o estimador do parámetro que maximice a función de verosimilitude, é dicir, atopar o estimador que faga os datos observados máis probables (Etz, 2018). A verosimilitude úsase, por exemplo, para contrastar dúas hipóteses, posto que non ten significado por si mesma. Pola contra, na estatística baiesiana, a función de verosimilitude representa a información proporcionada polos datos que servirá para actualizar a distribución a priori do parámetro, sendo proporcional a $p(y|\theta)$. A dificultade da inferencia recae sobre todo na elección do *prior* e da función de verosimilitude.

Finalmente, en canto ao resultado da inferencia, na estatística frecuentista obtemos o estimador dun parámetro e un intervalo de confianza. Un intervalo de confianza ao *x*% indica que se repetimos o experimento 100 veces e calculamos o IC, *x* deses ICs conterán o verdadeiro valor do parámetro (é, por tanto, unha medida de frecuencia). En principio, isto danos un rango de valores dentro dos cales é máis probable que se atope o verdadeiro parámetro, mais o investigador non saberá se o IC que escolle é un dos gañadores. Pola contra, coa estatística baiesiana estimamos a distribución posterior do parámetro, resumida nun estimador e nun intervalo de credibilidade, definido como o rango de valores que contén ao verdadeiro parámetro cun grao de credibilidade do *x*%.

A maior parte da estatística xenética está construída baixo a premisa frecuentista. En certas análises supón unha aproximación computacionalmente máis eficiente e que non require ter unha información a priori sobre o proceso a estudar. Sen embargo, a definición baiesiana da probabilidade é moito máis accesible que a frecuentista, ofrecendo vantaxes para certas análises. Isto permite saírse dos marcos dos contrastes de hipóteses, que requiren dunha hipótese nula que en moitos casos non é realista, e por tanto reformular certas cuestións dende outra óptica, especialmente escapando do binarismo dos p-valores.

3.11 FINE-MAPPING

Unha vez temos identificadas as rexións de risco nun GWAS, debemos tratar de determinar cal das variantes na rexión é a causal. A variante con menor p-valor é simplemente a que máis potencia estatística ten, pero partimos da suposición de que esta variante se atopa en desequilibrio de ligamento coa variante causal ou nunha rexión próxima. É aquí onde entra a técnica do mapeamento fino (*fine-mapping*).

Aínda que existen outras aproximacións, como por exemplo o uso de regresións penalizadas, as estratexias baiesianas de mapeamento fino ofrecen numerosas vantaxes sobre as que se centran na correlación entre SNPs, como por exemplo o uso de probabilidades posteriores ou que modelan o efecto conxunto de varios SNPs (Schaid et al., 2018). No *fine-mapping* baiesiano establecemos unha serie de hipóteses *a priori*, como que existe un número *p* de variantes causais dentro dunha rexión S ou que todas as variantes teñen a mesma probabilidade de ser causais, e calcúlase a probabilidade do modelo condicionado aos datos observados (Schaid et al., 2018).

Concretamente, os métodos baiesianos de *fine-mapping* están centrados en calcular a probabilidade posterior de inclusión (PIP) dun SNP nun modelo. A PIP dun SNP calcúlase

sumando as probabilidades posteriores de ser causal en cada un dos modelos. A partir das PIPs pódense calcular os sets cribles a un nivel α de confianza, que contén todos os SNPs cuxa suma cumulativa de PIPs equivale a α. É dicir, se queremos un set creíble ao 95%, incluiremos todos os SNPs ordenados ata que a suma das PIPs sexa 0.95. Como podemos observar, o set crible baixo a probabilidade baiesiana é moito máis sinxelo de entender e permite a comparación entre SNPs de forma directa, sen depender dos p-valores.

O *fine-mapping* baiesiano foi o método escollido para as análises nesta tese, empregando o paquete de R *corrcoverage*. O modelo que usa este paquete asume que (1) só existe unha única variante causal por rexión, e (2) que esta variante foi observada e analizada. As PIPs calculáronse para todas as variantes cunha r^2 >0.1 coa variante sentinela que se atoparan dentro dunha ventá de 1MB. Ademais, calculouse un set crible ao 95%.

3.12 DATOS DE EXPRESIÓN

Os estudos de asociación xa non só buscan variantes responsables da variabilidade de fenotipos clásicos: outro tipo cada vez máis importante de variantes son *expression quantitative trait loci* (eQTL) ou os *protein quantitative trait loci* (pQTL). O primeiro termo recolle todas as variantes asociadas á expresión dos xenes, mentres que o seguindo refírese a *loci* asociados á abundancia da proteína producida por un xen (Albert & Kruglyak, 2015)

A cantidade de transcritos dun xen presenta variabilidade entre individuos determinada por polimorfismos (Stranger et al., 2007). Para detectar estes polimorfismos ou eQTLs, realízanse estudos de asociación de forma similar a un GWAS convencional, pero medindo o nivel de mRNA de cada xen nun tecido (Albert & Kruglyak, 2015). Os eQTL poden actuar regulando a xenes próximos, chamados cis-eQTL, ou regular a xenes que se atopan afastados ou incluso noutro cromosoma, chamados trans-eQTL. A aparición destes estudos permitiu construír catálogos de regulación xenética como por exemplo o proxecto GTEx (Aguet et al., 2017), cuxa versión 8 viu a luz en 2020, ou o recente mapa de pQTLs en plasma publicado polo consorcio *UK Biobank Pharma Proteomics Project* (B. B. Sun et al., 2023) grazas ao cal se atoparon 14.287 asociacións con 2.923 proteínas.

Como xa se explicou, gran parte das asociacións do GWAS aparecen en zonas interxénicas ou non codificantes, e a herdabilidade está enriquecida en rexións reguladoras (Maurano et al., 2012). A integración dos sinais GWAS cos cis-eQTL permite determinar se a relación SNP-fenotipo dáse a través da regulación da expresión dun xen próximo por parte do SNP, facilitando a priorización de xenes e a detección dos mecanismos biolóxicos causantes do fenotipo (Nica & Dermitzakis, 2008). Algunhas das estratexias de integración de expresión e GWAS son a colocalización, un método estatístico para comprobar se unha mesma variante é a causante de dúas asociacións (co-ocorrencia) (Giambartolomei et al., 2014; Hormozdiari et al., 2016), ou os TWAS (*Transcriptome-Wide Association Studies*), que predín a expresión nos individuos e testan a súa asociación co fenotipo.

3.12.1 Estudos de asociación de transcriptoma completo (TWAS)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Como se acaba de comentar, un posible método para integrar datos de expresión con datos de GWAS é o estudo de asociación de transcriptoma completo ou TWAS. A idea básica detrás deste tipo de análise é detectar xenes asociados ao fenotipo cuxa expresión estea regulada polos SNPs das rexións de risco, o que permite interpretar mellor os resultados do GWAS e priorizar potenciais xenes causais. Normalmente adéstranse uns modelos de expresión a partir de datos públicos, como por exemplo os de eQTL de GTEx. O resultado é un peso para cada SNP, que será o efecto da variante na regulación da expresión de cada xen. A continuación, axústanse estes modelos sobre os xenotipos dos individuos para predicir a expresión dos xenes, e faise un test de asociación das matrices de expresión co fenotipo de interese (Mai et al., 2023). Como en moitas ocasións non é posible dispoñer dos datos xenéticos crús, desenvolvéronse técnicas para predicir a expresión empregando datos de resumo dos GWAS. PrediXcan (Gamazon et al., 2015) e FUSION (Gusev et al., 2016) empregan modelos lineais aditivos, como *lasso*, BLUP ou *elastic net*. Nesta tese decidimos empregar S-MetaXcan, a adaptación de MetaXcan (un paquete que integra o método de PrediXcan) a estatísticos resumo (Barbeira et al., 2018).

En PrediXcan, os fenotipos modélanse como funcións lineais de X_l, que é o número de alelos do SNP *l*, e de T_g, que é a estima da expresión regulada xenéticamente do xen *g* (Gamazon et al., 2015). Sendo *Y* un vector de tamaño N que toma os valores dos fenotipos de cada individuo *i=1..n*:

$$Y = \alpha_1 + X_l\beta_l + \theta,$$

$$Y = \alpha_2 + T_g\gamma_g + \varepsilon.$$

α_1 e α_2 son interceptos; θ e ε termos de erro; $T_g = \sum_{l \in Modelo_g} w_{lg} X_l$, é a expresión predita do xen *g*; β_l o coeficiente de regresión asociado ao SNP *l*; X_l o número de alelos de referencia do marcador *l*, e γ_g o coeficiente de regresión asociado ao xen *g*. Os pesos w_{lg} que toman os SNPs na expresión calcúlanse a través do modelo seleccionado para a expresión (por exemplo, un modelo *lasso*) e uns datos de referencia. Un GWAS daría como resultado β_l (o tamaño de efecto do SNP *l* sobre o fenotipo), mentres que PrediXcan daría como resultado o coeficiente γ_g (o tamaño de efecto do xen *g* sobre o fenotipo).

Se non se dispón dos xenotipos dos individuos, non é posible calcular T_g. Por tanto, os autores desenvolveron S-MetaXcan, que calcula o estatístico de Wald (*Z-score*) para a asociación entre a expresión xénica predita e o fenotipo (Barbeira et al., 2018):

$$Z_g \approx \sum_{l \in Modelo_g} w_{lg} \frac{\widehat{\sigma_l}}{\widehat{\sigma_g}} \frac{\widehat{\beta_l}}{se(\widehat{\beta} l)}.$$

Na fórmula anterior, w_{lg} é o peso do SNP *l* na predición da expresión do xen *g*; $\widehat{\beta_l}$ é o beta asociado ao SNP *l* no GWAS; $se(\hat{\beta} l)$ é o erro estándar de $\widehat{\beta_l}$; $\widehat{\sigma_l}$ e $\widehat{\sigma_g}$ son as varianzas estimadas do SNP *l* e da expresión do xen *g*, respectivamente. Por tanto, para a computación dos resultados do TWAS precísase un modelo de predición da expresión, uns estatísticos resumo do GWAS e unha matriz de covarianzas da expresión dos SNPs incluídos na predición.

Os *software* de TWAS permiten adestrar os modelos con datos propios, mais de non ser o caso é posible descargar modelos pre-computados de páxinas web como https://predictdb.org/.

Concretamente, os autores de S-MetaXcan publicaron uns modelos adestrados en datos de GTEx v8 sobre variantes sometidas á mapeamento fino, e empregaron un método novidoso para a estimación dos efectos (Barbeira et al., 2021): *mashr*, ou contracción multivariante adaptativa, que permite estimar os efectos de múltiples elementos baixo máis dunha condición con presencia de correlación (Urbut et al., 2019). Nun contexto como o dos eQTL, este modelo permite a correlación da expresión entre tecidos. Os modelos pre-computados MASHR-M foron os empregados no noso estudo.

Habitualmente córrense os TWAS por tecido, polo que é necesario seleccionalos coidadosamente a partir de información biolóxica previa. Do contrario, poderíamos atopar unha asociación significativa nun tecido non relacionado co fenotipo e complicaríase a interpretación dos resultados. Neste estudo escolléronse os tecidos de pulmón, sangue, células tipo linfocitos e mucosa do esófago. Non obstante, combinar as predicións de varios tecidos incrementa o poder estatístico das asociacións. MultiXcan aplica un test para estimar o efecto conxunto da expresión xénica en varios tecidos á vez (Barbeira et al., 2019), e tamén conta coa versión para datos de resumo (S-MultiXcan). Os autores recomendan correr os TWAS multi-tecido de forma agnóstica, pois incrementa o número de sinais detectados.

Se ben son unha ferramenta útil para priorizar xenes, hai que ter en conta certos aspectos. Por un lado, a correlación da expresión entre xenes causais e non causais pode provocar asociacións falsas (Wainberg et al., 2019). Igualmente, se dous xenes (un causal e outro non causal) inclúen á mesma variante *hit* GWAS no modelo de expresión, pode ser que ambos saian significativos no TWAS (Wainberg et al., 2019). Por último, resaltamos tamén que os datos de expresión dispoñibles na actualidade do consorcio GTEx -o máis empregado- derivan dunha mostra composta nun 84.6% por individuos "brancos". Así, as análises que integran datos de expresión poden estar nesgadas en poboacións non europeas. Gay e colaboradores (2020) estudaron os efectos da AL en eQTLs de GTEx, identificando un grupo de eQTLs correlacionados coa AL e un grupo de xenes para os cales a corrección por AL aumentaba a potencia estatística, se ben os resultados da colocalización non se vían afectados (Gay et al., 2020). O estudo transcriptómico dos datos da cohorte MESA (que comprende individuos africanos-americanos, hispanos e europeos) revelou que as diferencias nas FAs podían reducir a precisión coa que se predicía a expresión dalgúns xenes (Mogil et al., 2018). En resumo, empregar datos de expresión de europeos pode resultar nunha perda de información.

3.13 ÍNDICES DE RISCO XENÉTICO

O estudo dos índices de risco xenético nas cohortes española e latinoamericana de SCOURGE dividiuse en dúas partes. Na primeira explóranse diferentes métodos de cara a determinar cal ten mellor rendemento na nosa cohorte mixta LatAm, en comparación coa predición na cohorte española (definida como "europea" a través da GIA). En función dos datos de descubrimento, derivamos un grupo de PRS europeos e un grupo de PRS multi-poboación. Ao dispoñer dunha poboación con mestura recente como a latinoamericana, podemos comparar o rendemento de diferentes estratexias de construción de PRS nunha cohorte complexa e non clasificada en grupos arbitrarios, que é un escenario moito máis realista de cara a trasladar estes índices a unha práctica clínica. Na segunda parte, o mellor PRS para cada cohorte empregouse

para desenvolver un modelo de risco en conxunción con covariables clínicas. A continuación descríbense as diferentes aproximacións empregadas.

3.13.1 Cohortes de descubrimento e diana

As cohortes de descubrimento para o cálculo dos PRS foron as meta-análises da versión 7 do consorcio HGI. Escolleuse o fenotipo B2 para a variable de hospitalización, é dicir, doentes hospitalizados *versus* doentes non hospitalizados e poboación xeral. Empregáronse tanto a análise multi-poboacional (ALL) como as análises específicas de poboación (AFR, EUR, AMR, SAS e EAS). Estes datos inclúen variantes cuxa calidade de imputación > 0.6 e frecuencia alélica >0.001.

A cohorte de descubrimento e a cohorte diana deben ser completamente independentes. Do contrario, o solapamento entre mostras provoca un sobre-axuste dos PRS e prodúcese unha inflación dos resultados, obtendo estimacións demasiado optimistas (Wray et al., 2013). Dado que na versión 7 dos resultados participaba o consorcio SCOURGE co estudo da cohorte española, eliminouse a súa contribución da análise específica de EUR. Para isto usamos o paquete de R MetaSubtract (Nolte, 2020). Este método permite obter uns estatísticos resumo de meta-análises GWAS sen a contribución da cohorte de interese, polo que se poden usar de forma segura sobre os datos individuais da mesma. Esta aproximación pode diminuír a potencia estatística da meta-análise e, por tanto, tamén cambiar os resultados, pero ao non dispoñer dunha de validación independente non foi posible empregar a meta-análise completa. Para os PRS que empregaban a meta-análise ALL, seleccionouse a análise *leave-one-out* B2 ALL de HGI sen SCOURGE.

As cohortes diana foron as cohortes española e LatAm de SCOURGE. Dado que os datos de descubrimento derivan da análise de hospitalización por COVID-19, esta foi a variable empregada inicialmente para a afinación de parámetros e selección do PRS con mellor rendemento. Ao non dispoñer de controis poboacionais na cohorte latinoamericana, usouse a definición de hospitalización dentro de casos (variable CHOSP). Porén, explorouse tamén a tamén a variable de enfermidade crítica dentro de casos (C_SEV4).

Finalmente, usouse como mostra test[1] un grupo de 473 individuos pertencentes á cohorte latinoamericana que foron excluídos por presentar ≥ 80% de ancestralidade europea (362 casos COVID+ non hospitalizados e 111 casos COVID+ hospitalizados). A mediana da idade é de 42.7 anos e, deles, 220 son homes e 253 son mulleres.

3.13.2 PRS C+T

[1] Para facilitar a comprensión, definimos aquí os conceptos de mostra de adestramento, de validación e test. A mostra de adestramento, no contexto dos PRS, fai referencia ao GWAS do que parten os datos, mais nesta tese será denominada como "mostra de descubrimento". Cando se cite á mostra de adestramento, esta referirase ao conxunto de datos nos que se axustará un modelo estatístico. A mostra de validación é un segundo set de datos que se emprega para o refinamento dos parámetros. Finalmente, a mostra test é aquela na que se avalía o modelo adestrado e refinado para obter unha estima do rendemento do modelo non nesgada. Dado que a nosa mostra test ten un tamaño limitado, substituirase o axuste na mostra de validación por un procedemento *bootstrap*.

O primeiro método que se explorou foi o clásico *clumping + thresholding* (C+T). A pesar de que *software* como PRSice (Euesden et al., 2015) teñen implementada todo o fluxo de traballo para a construción, derivación e elección do mellor PRS con este método, para esta investigación construíronse con PLINK v1.9 e PLINK v2.0 (Purcell et al., 2007) en combinación con *scripts* de R. O método C+T selecciona unha serie de variantes índice nos estatísticos de sumario dun GWAS en función dun límite de p-valor e agrupa as variantes que se atopen a unha distancia determinada e cuxa r^2 sexa maior a un valor concreto (*LD-clumping*) (S. W. Choi et al., 2020; Purcell et al., 2009). Tanto o p-valor, como a ventana na que se avalía o LD, como o r^2 son os hiper-parámetros que se deben refinar, se ben tamén pode ser de interese ter en conta a calidade de imputación (Privé et al., 2019). Inicialmente partimos duns datos de sumario dun GWAS que dispoñan, como mínimo, do nome do SNP, do alelo de efecto e do coeficiente de regresión do SNP. A continuación, o algoritmo comezaría seleccionando as variantes con menor p-valor da rexión e agruparía variantes do arredor que poidan ser representadas pola variante líder (é dicir, cuxos efectos están correlacionados). Isto reduciría o número de variantes, que posteriormente extraeríamos para calcular a puntuación. O panel de LD que se empregue pode ser ou ben a propia mostra ou ben un panel externo, pero debe corresponder o mellor posible á composición da poboación do GWAS (S. W. Choi et al., 2020). No caso de que unha poboación estea sobre-representada no GWAS, empregar un panel de referencia correspondente a esa poboación mellora a predición (Y. Wang, Namba, et al., 2023).

Con este método construíronse un grupo de PRS europeos e un grupo de PRS multi-poboación, usando os datos das meta-análises de HGI B2 de EUR e ALL, respectivamente. Para cada estatístico resumo empregamos o algoritmo C+T variando os parámetros de r^2 (0.2, 0.4, 0.6, 0.8), ventá de *clumping* (100, 250, 500, 1000) e os p-valores límite (0.1, 0.01, 1e-05, 1e-08). A propia mostra foi usada como panel de LD. Deste xeito obtivemos un total de 64 PRS por cohorte e datos de descubrimento, variando o número de SNPs incluídos en cada un. As variantes do PRS seleccionadas con este método dependen da cohorte diana, polo que unha certa combinación dos parámetros non resultará no mesmo índice para a cohorte española que para a latinoamericana.

3.13.3 PRS-CS e PRS-CSx

Os métodos baiesianos son moi populares no campo dos PRS. Un dos métodos que obtén bo desempeño é PRS-CS (Ge et al., 2019), que ademais posúe unha versión para construír PRS multi-poboación, PRS-CSx (Ruan et al., 2022). Aínda que existen outros métodos para a derivación de índices de risco multi-poboación, como X-Wing (Yao et al., 2020) ou XPASS (Cai et al., 2021), o modelo de PRS-CS está suficientemente testado. Comezando co modelo básico de PRS-CS, este emprega unha contracción continua dos coeficientes *a priori* tendo en conta as estruturas do desequilibrio de ligamento. Se tomamos a descrición do modelo en Ge e colaboradores (2019), o fenotipo ten a forma:

$$y = X\beta + \varepsilon;\ \varepsilon \sim N(0, \sigma^2 I).$$

Nesta fórmula, N é o número de individuos, M o número de variantes, y o vector de fenotipos, X a matriz de xenotipos, β o vector de efectos dos SNPs e ε o vector dos residuos.

Para o vector β especifícanse unhas densidades *a priori*, e neste caso empregan uns *priors* con contracción continua, representadas como unha mestura de distribución normais:

$$\beta_j|\varphi_j \sim N(0, \emptyset\varphi_j), \varphi_j \sim g.$$

Aquí, $\emptyset$ é un parámetro de escala compartido por todos os marcadores que determina o grao de dispersión do modelo (é dicir, o número de SNPs que toman un valor distinto de cero) (Ge et al., 2019). g é unha función continua que variará a forma da distribución dos priors para β_j, mentres que e φ_j é o parámetro de contracción local de cada SNP, que permite reducir os efectos con estimacións pobres a valores achegados ao 0. Existen dúas formas de estimar o parámetro *phi* ($\emptyset$) en PRS-CS: por un lado, pódese realizar unha busca en cuadrícula, que consiste en testar unha serie de valores de *phi* e empregar unha mostra de validación para seleccionar o valor que mellor predición dá, ou pódese escoller unha pseudo-validación, na que o propio modelo aprende o parámetro *phi* a partir dos datos (Ge et al., 2019). A busca automática do parámetro *phi* ofrece resultados similares á busca en cuadrícula, o que indica que o modelo automático é capaz de escoller correctamente o valor óptimo deste parámetro (Y. Wang, Namba, et al., 2023). Este modelo adapta a contracción dos efectos en función da forza de asociación dos SNPs nos GWAS, e permite actualizalos en bloque en función dos patróns de LD, ata obter uns efectos *a posteriori* (Ge et al., 2019). Por tanto, a construción do PRS con PRS-CS só require dos estatísticos de sumario dun GWAS e dun panel de referencia externo.

Este método estendeuse a *K* poboacións en PRS-CSx. Os parámetros $\emptyset$ e φ_j non dependen de *k*, polo que se un mesmo SNP está presente en todas as poboacións se aplicará o mesmo *prior* de contracción continua (Ruan et al., 2022). Este *prior* compartido permite a existencia de efectos correlacionados pero con distinta magnitude entre eles. Para a construción do PRS precísase uns estatísticos resumo e un panel de referencia por poboación. Ruan e colaboradores (2022) apuntan que este método captura sinais específicas de poboación, de xeito que un SNP cuxo efecto só sexa estimable nunha das poboacións será incluído no PRS da poboación, mais non se acoplará o seu efecto entre as poboacións.

O software de PRS-CSx inclúe unha opción "meta", para meta-analizar os K resultados e xerar un PRS multi-poboación. Non obstante, os autores recomendan combinar os K PRS nunha mostra de validación, optimizar deste xeito os pesos de cada PRS específico, e posteriormente aplicar o modelo á mostra diana. Os autores destacan que PRS-CSx permite modelar os patróns de LD específicos de poboación, mentres que un PRS derivado dunha meta-análise conxunta non (Ruan et al., 2022). No contexto desta investigación, o consorcio HGI liberou os resultados estratificados por grupos poboacionais, facilitando a construción deste tipo de PRS. Con todo, a alternativa de empregar unha meta-análise multi-poboación ofrece bos resultados (Y. Wang, Kanai, et al., 2023).

O segundo set de índices construíuse cos métodos baiesiano de PRS-CS e PRS-CSx. Empregando os estatísticos resumo da meta-análise europea de HGI para hospitalización (excluíndo SCOURGE), deriváronse os efectos dos SNPs coa versión clásica (PRS-CS) co obxectivo de construír e testar un PRS exclusivamente europeo. Igualmente, usamos os métodos de PRS-CSx e PRS-CS para derivar 2 PRS multi-poboación, usando as análises meta-

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GWAS por poboación (SAS, EUR, AMR, AFR e EAS) e ALL, respectivamente. Empregouse a opción "meta" e permitíuselle ao *software* escoller o parámetro φ.

3.13.4 Selección dos hiperparámetros dos PRS

A mellor práctica para a construción dos PRS consiste en probar diferentes hiperparámetros e seleccionar os que mellores resultados ofrezan, en contraposición a escoller valores arbitrarios. Porén, isto pode producir un sobre-axuste do modelo se non se empregan datos independentes (S. W. Choi et al., 2020). Mak e colaboradores (2018) definen este sobre-axuste como "a inflación da correlación do PRS coa compoñente xenética nos datos diana en comparación a un set de datos externo". O ideal é dispoñer dunha mostra de descubrimento e outra de validación para isto, é dicir, construír o PRS cos datos de descubrimento (o GWAS) e escoller os parámetros que maximicen a precisión da predición na mostra de validación, para finalmente testar o modelo na mostra test. Para os fenotipos binarios tipo caso/control o máis común é empregar como medidas a área baixo a curva ROC (AUC) ou o pseudo-R^2 de Nagelkerke (Nagelkerke, 1991). Esta só é útil na comparación de modelos, pero en caso de que se queira entender o R2 como a proporción da variabilidade explicada, debe de transformarse tendo en conta a prevalencia do fenotipo na poboación (S. H. Lee et al., 2012).

Nos casos nos que é difícil dispoñer de tres sets independentes de datos[2], outras técnicas como o re-mostraxe por *bootstrap* ou validación cruzada poden solucionar potenciais problemas de sobre-axuste. Na validación cruzada divídese a mostra diana en un conxunto de adestramento e outro de test. O modelo de regresión, que inclúe o PRS e as covariables, axústase na mostra de adestramento e avalíase na mostra test a través dalgunha métrica como por exemplo, o AUC para regresión loxística, o RMSE en regresión lineal, ou a correlación media entre valores observados e valores preditos. No caso de que a medida de validación sexa o R^2 de regresión, hai que ter en conta que este valor pode ser negativo na mostra test cando existe sobre-axuste, e isto é fácil que se produza cando a N é baixa (Harrell, 2015, p. 112). A validación cruzada non obtén resultados moi fiables salvo que se repita un gran número de veces, por exemplo a través do tipo *K-fold*, onde se divide a mostra en K sets e repítese o procedemento empregando K-1 como adestramento e o outro como mostra test.

A re-mostraxe con *bootstrap* é unha alternativa que non require da división dos datos. O *bootstrap* ordinario emprega una mostraxe con substitución do tamaño da mostra e estima en cada re-mostra o estatístico de interese, e co total das re-mostras podemos obter a distribución do mesmo e calcular os intervalos de confianza. Ademais, esta técnica permite validar os modelos internamente, posto que por cada re-mostra se avalía o modelo nun conxunto de individuos diferente. Esta foi a técnica seleccionada para as nosas análises.

O modelo clásico que se empregou de forma xeral para estudar o rendemento de cada PRS foi a regresión loxística. Por cada PRS axustáronse dous modelos loxísticos para a variable de hospitalización, un coas covariables base (sexo, idade e 10 PCs) e outro coas covariables e o PRS. Calculouse o incremento no pseudo-R2 a partir de 1.000 re-mostras por *bootstrap*, xunto

[2] A modo de nota recordatoria, os GWAS (externos á mostra diana) funcionan nos PRS como mostra de adestramento.

cos intervalos de confianza. Por tanto, calculouse o pseudo-R2 parcial (ΔR^2) (de Nagelkerke) asociado á variable PRS.

3.13.4.1 Selección do PRS óptimo

Construíronse un total de 130 PRS para cada cohorte, pero seleccionouse o PRS (é dicir, a mellor combinación de método-descubrimento) con mellor predición en base ao ΔR^2. Despois, con este PRS, axustáronse dous modelos de regresión loxística (un por cohorte) empregando como covariables as 10 primeiras PCs, o sexo e a idade, e o PRS estandarizado (dentro de cada cohorte). Así, estimouse a OR asociada ao incremento dunha desviación típica da puntuación xenética. Na cohorte latinoamericana realizouse unha análise de sensibilidade para descartar potenciais diferencias na capacidade preditiva do PRS en función ás proporcións de ancestralidade: calculouse a OR para o 30% dos individuos con maior ancestralidade NAM ou AFR, e para o 30% de individuos con menor ancestralidade NAM ou AFR.

A partir deste punto, traballaremos sempre co PRS construído con esta combinación método-mostra descubrimento óptima.

3.13.5 Construción dun PRS único e validación

O uso de dúas cohortes separadas impide establecer unha única OR asociada ao PRS para todos os individuos, dificultando a súa traslación clínica. Dado que os PRS obtiveron resultados similares (ver sección de Resultados), ou incluso lixeiramente superiores, para a cohorte LatAm que para a cohorte española, decidimos unificalas e explorar a construción dun novo modelo de predición cun mesmo PRS.

3.13.5.1 Desenvolvemento do modelo

Aínda que a estimación dos efectos nos PRS baiesianos é independente das cohortes diana, á hora de aplicar o PRS si pode existir variabilidade entre elas polo número de SNPs xenotipados e imputados dispoñibles. Para equiparar o PRS entre ambas cohortes, seleccionáronse unicamente as variantes compartidas, quedando un total de 954.302. Ademais, repetiuse o PCA para obter unhas PCs conxuntas.

O PRS axustouse seguindo un procedemento previamente descrito por outros autores (Ge et al., 2022; Hao et al., 2022; Khera et al., 2019), que ten por obxectivo reducir o efecto da ancestralidade no PRS. En primeiro lugar, estimouse a contribución das *K* PCs no PRS dentro do grupo de individuos non hospitalizados mediante o seguinte modelo:

$$PRS = \beta_0 + \sum_{i=0}^{K} \beta_k PC_k.$$

Desta forma obtivemos un peso asociado a cada PC. Despois, con dito modelo predixemos a contribución da ancestralidade ao PRS para o total dos individuos. Finalmente, subtraeuse o valor predito do valor cru, obtendo un PRS axustado (PRS_{ax}), que tamén foi estandarizado. Construíuse un único modelo loxístico para hospitalización dentro de casos COVID-19

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

incluíndo o PRS_{ax}, o país de hospitalización, o sexo e a idade. Ao igual que durante a selección de PRS, estimouse o ΔR^2 a través de remostraxe *bootstrap*.

O incremento no pseudo-R^2 foi transformado á *liability scale* mediante a ecuación 15 do artigo de S. H. Lee e colaboradores (2012), empregando un valor da prevalencia da hospitalización do 15% (pre-vacinas) e un valor máis baixo, do 0.05% (5 hospitalizacións por cada 100 casos de COVID-19). A prevalencia de hospitalización pre-vacinas calculouse de forma aproximada como un promedio para o Nº hospitalizados/Nº casos diagnosticados entre dous informes (redondeando ao 0.05 máis próximo): o 230 e 351 do Centro de Coordinación de Alertas e Emerxencias Sanitarias do Goberno de España (Centro de Coordinación de Alertas y Emergencias Sanitarias, 2020, 2021). Estes informes son os correspondentes a outubro do 2020 e abril de 2021, respectivamente (Nº de casos hospitalizados ata o período 1: 161.120; Nº de casos diagnosticados ata o período 1: 936.560; Nº de casos hospitalizados ata o período 2: 337.623; Nº de casos diagnosticados ata o período 2: 3.370.256).

3.13.5.2 Validación do modelo

O modelo final validouse interna e externamente. Para a validación interna usouse o paquete de R *rms.* Estimáronse a área baixo a curva ROC, AUC, e o pseudo-R^2 como medida de rendemento aparente, corrixida para o optimismo (Collins et al., 2024). O rendemento aparente é a medida do rendemento do modelo calculado nos mesmos datos cos que se construíu. Este procedemento produce un sobre-axuste, obténdose unha estima do rendemento máis optimista que se fora calculada nun novo set de datos (Collins et al., 2024). Porén, mediante *bootstrap* pódese calcular o optimismo e corrixir o rendemento aparente. Os pasos son os seguintes: (1) calcúlase a medida de interese do modelo, construído en cada remostra, tanto sobre a mostra *bootstrap* ("rendemento *bootstrap*") como sobre a mostra completa ("rendemento test"), (2) calcúlase a diferencia entre ambas cantidades (optimismo), e (3) substráese a media das diferencias do paso (2) da medida de interese calculada no modelo orixinal. A función *validate* do paquete *rms* realiza este procedemento automaticamente, devolvendo unha serie de medidas do rendemento do modelo, entre elas o D_{xy} (a correlación de rangos de Somer's), que é igual a 2*(AUC-0.5). Finalmente, mediante a función *calibrate* estudouse a calibración do modelo, tamén mediante remostraxe. A calibración permite analizar se as probabilidades preditas polo modelo están cerca das probabilidades verdadeiras, o que é importante de cara a determinar se o modelo sobreestima ou infraestima o risco (Van Calster et al., 2019). Ademais do gráfico de validación, a función calcula o erro medio da predición. Seleccionouse un B=1.000.

A validación externa consistiu na predición do estado de hospitalización para os 473 individuos da mostra test definida no apartado 3.13.1. O cálculo do PRS_{ax} realizouse cosasociados ás PCs do modelo descrito no apartado 6.1.2. Estimouse a sensibilidade (probabilidade de identificar correctamente un caso), especificidade (probabilidade de identificar correctamente a un control), valor preditivo positivo (nº de casos identificados como casos/nº de individuos identificados como casos), valor preditivo negativo (nº de controis identificados como controis/nº de individuos identificados como controis) e o AUC. Comparáronse tres escenarios: (1) predición a partir dunha poboación mixta (cohortes

LatAm+española); (2) predición a partir dunha poboación única discreta definida con anterioridade cos mesmos criterios que a diana (cohorte española, ≥ 80% EUR); e (3) predición a partir dun grupo de individuos derivados da mostra mixta pero con composición similar á mostra diana. O escenario número (3) buscaba simular o (2) partindo dunha mostra sen clasificación previa. Para isto, calculouse o centro do espazo dos 10 PCs na mostra test e a distancia Euclídea de cada un dos individuos da mostra de adestramento a dito centro. Despois, seleccionouse o 40% da mostra de adestramento (N=5.154) cuxos individuos tiveran a menor distancia Euclídea media á mostra test. Os escenarios 2 e 3 requiriron da construción de dous modelos adicionais: (1) *EUR* corresponde ao axustado cos individuos da cohorte española (clasificados como EUR por ser ≥ 80% EUR), e (2) *Neighbors*, axustado cos individuos máis xeneticamente achegados. A variable correspondente ao país de recrutamento non foi incluída para a cohorte española dado que só toma un valor. Estes dous modelos testáronse sobre os 473 individuos adicionais, e comparáronse os resultados entre eles e co modelo derivado do conxunto de datos completo. O PRS foi dividido en percentís e en decís para a realización de gráficos e de análises comparativas.

3.13.6 Estudo do PRS en relación ás variables clínicas

A segunda parte da análise dos PRS ten por obxectivo o de estudar a asociación do risco xenético de hospitalización con outras variables clínicas e demográficas. A base de datos do consorcio SCOURGE ten recollidas máis de 300 variables, mais a urxencia para a recollida dos datos, xunto coa saturación dos centros clínicos durante a pandemia, dificultou que se incluíra toda a información para cada individuo. Por tanto, analizáronse as variables clínicas de máis relevancia dentro do grupo de casos con diagnóstico COVID-19 positivo. Non obstante, cómpre ter coidado coa interpretación das variables con porcentaxes altas de datos faltantes ou un número de casos baixo. As variables referidas ao resultado da enfermidade que se estudaron foron: hospitalización, enfermidade crítica, morte, enfermidade asintomática. En canto ás complicacións derivadas da COVID-19, exploráronse: o requirimento de ventilación mecánica (invasiva, non invasiva ou cánulas de alto fluxo), tromboembolismo pulmonar, tromboembolismo venoso, presencia de infiltración pulmonar en radiografía de tórax e miocardiopatía.

Os modelos empregados para o estudo das variables clínicas (variables dependentes) foron modelos loxísticos incluíndo como covariables base o sexo, a idade e as 10 primeiras compoñentes principais, país de recrutamento e PRS_{ax}. A mostraxe realizada para o GWAS pode presentar un nesgo de selección, que consiste nunha representación inexacta da poboación a estudar porque os casos e os controis teñen distinta probabilidade de ser recrutados (Tripepi et al., 2010). Como análise de sensibilidade re-analizáronse as variables dentro do grupo de casos hospitalizados, salvo a variable de enfermidade asintomática e hospitalización. O PRS_{ax} incluíuse de forma cuantitativa, pero tamén se analizou como variable binaria para estudar o 20% superior da distribución do PRS.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

As probabilidades de padecer un cadro crítico de COVID-19 ou complicacións asociadas son maiores en homes, doentes de idades avanzadas, e doentes con comorbilidades asociadas. Como análises secundarias, estudamos a variación do efecto do PRS na predición da enfermidade crítica e hospitalización debido a estes factores de risco. Realizouse unha análise de interacción entre o PRS e o sexo, a idade (binaria, <65 anos, ≥65 anos) e presencia de comorbilidades:

$$Y = \beta_{sex} + \beta_{PRS} + \beta_{PRS}\beta_{sex} + \beta_{idade_{num}} + \beta_{POPHOSP} \quad (1)$$

$$Y = \beta_{idade_{bin}} + \beta_{PRS} + \beta_{PRS}\beta_{idade_{bin}} + \beta_{sex} + \beta_{POPHOSP} \quad (2)$$

$$Y = \beta_{comorb_{bin}} + \beta_{PRS} + +\beta_{PRS}\beta_{comorb_{bin}} + \beta_{sex} + \beta_{idade_{num}} + \beta_{POPHOSP} \quad (3)$$

$$Y = \beta_{comorb_{num}} + \beta_{PRS} + \beta_{PRS}\beta_{comorb_{num}} + \beta_{sex} + \beta_{idade_{num}} + \beta_{POPHOSP} \quad (4)$$

As comorbilidades testáronse de dúas formas, para comparar os resultados cos de Nakanishi e colaboradores (2021): (a) presencia de máis dunha comorbilidade *vs* ningunha comorbilidade (*comorb$_{bin,}$* modelo 3), e (b) de forma numérica en función do número de comorbilidades (0-6, *comorb$_{num}$*, modelo 4). Explorouse a asociación do PRS$_{ax}$ cos factores de risco vasculares de forma individual (hipertensión arterial, hipercolesterolemia, diabetes mellitus e obesidade); as comorbilidades cardíacas de forma conxunta (cardiopatía isquémica, insuficiencia cardíaca, arritmia cardíaca ou vasculopatía periférica), e as comorbilidades respiratorias, tamén de forma conxunta (EPOC ou outras enfermidades relacionadas).

Por último, compáraronse máis profundamente dous modelos co obxectivo de cuantificar a utilidade clínica do PRS en presencia doutros factores de risco clínicos. O modelo base FRC (factores de risco clínico) para hospitalización incluía como covariables: idade, sexo, número de comorbilidades (0-6), e país de mostraxe; o modelo principal FRC+PRS incluíu, a maiores, o PRS$_{ax}$ de forma cuantitativa. Estes modelos validáronse internamente mediante remostraxe por *bootstrap*, e externamente mediante a predición no grupo de 473 individuos test. Calculouse a taxa neta de mellora na reclasificación ou NRI (*net reclassification improvement*) coa librería de R *nricens*, que tamén calcula por *bootstrap* os intervalos de confianza. O NRI é unha métrica que cuantifica o valor engadido dunha nova variable en termos de reclasificación de casos e de controis, podendo tomar un valor máximo de 2 (Kerr et al., 2014):

$$NRI = P(subir|caso) - P(baixar|caso) + P(baixar|control) + P(subir|control).$$

Adicionalmente fixemos unha análise de curvas de decisión (DCA, *decision curve analysis*) (Vickers et al., 2016). Mediante esta aproximación podemos comparar o beneficio neto de varias estratexias de intervención clínica en función de distintos límites de probabilidade. Para cada corte de probabilidade p_t defínense como positivos os individuos que *P(hosp)* $\geq p_t$ e como negativos os que *P(hosp)* $< p_t$. Despois, calcúlase a sensibilidade e especificidade para cuantificar os danos derivados tanto de non intervir en p_t como os custes innecesarios de intervir en p_t, no chamado "beneficio neto" (Vickers et al., 2008, 2019):

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

$$Beneficio\ neto = \frac{Verdadeiros\ positivos}{N} - \frac{Falsos\ positivos}{N}\left(\frac{pt}{1-pt}\right).$$

Constrúese unha curva para o beneficio neto en todo o espazo de probabilidade, o cal mide a proporción de verdadeiros positivos. As curvas DCA informan sobre o beneficio neto de cada estratexia dependendo do límite de risco escollido para intervir (**Figura 8**), que variará en función da intervención, dos custos e dos danos que cause. Este corte dependerá da preferencia do médico ou do doente, de xeito que o límite será máis baixo para individuos que teñan máis preocupación pola enfermidade, pero máis alto para aqueles que teñan máis preocupación pola intervención. Vickers e colaboradores (2016) explican que o beneficio neto nun $p_{t,}$ determinado equivale a realizar un número $1/p_{t,}$ de intervencións para atopar un caso. Por exemplo, se escollemos intervir a partir de P=25%, terían que realizarse 4 intervencións para detectar un caso, ou o que é o mesmo, estar disposto a facer 4 intervencións para detectar un caso, considerando os danos asociados a intervir e a non intervir. Á súa vez, un beneficio neto do 5% equivale a identificar 5 positivos por cada 100 intervencións. Normalmente engádense as curvas dos modelos a testar, unha curva correspondente ao beneficio neto de tratar a todos os individuos (que cruzará o eixo X na prevalencia da enfermidade) e unha curva correspondente ao beneficio neto de non tratar a ningún individuo (Vickers et al., 2019). En p_t, a mellor estratexia será a que obteña o valor máis alto de beneficio neto, mais o óptimo sería que unha estratexia fora superior para todo o rango de probabilidades.

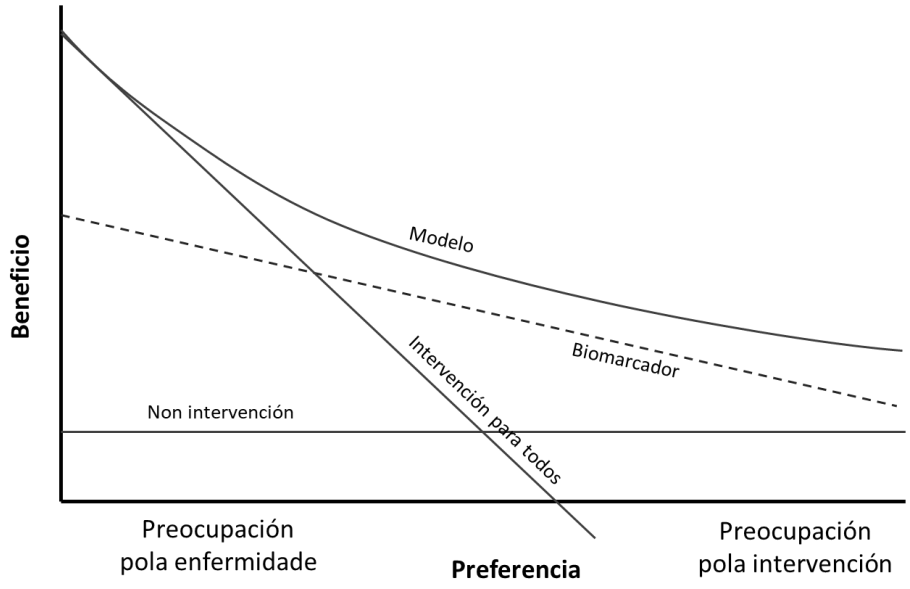


Figura 8. Exemplo dunha curva de decisión na que se valoran un biomarcador e un modelo de risco. Adaptada de Vickers e colaboradores (2019).

Estas curvas calculáronse co paquete *dca*. Debido a que os nosos datos proveñen dun estudo tipo caso-control, axustáronse os gráficos por unha prevalencia de hospitalización do 15%.

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

4 FACTORES XENÉTICOS DA COVID-19 EN POBOACIÓN ESPAÑOLA

Parte do contido do presente capítulo foi publicado como Raquel Cruz, Silvia Diz-de Almeida, et al. (2022). Novel genes and sex differences in COVID-19 severity, *Human Molecular Genetics*, Volume 31, Issue 22, 15 November 2022, Pages 3789–3806, https://doi.org/10.1093/hmg/ddac132.

Algunhas análises foron ampliadas aquí de cara a afondar na interpretación dos resultados e homoxeneizar as dúas cohortes. Os resultados do GWAS de SCOURGE deste capítulo, en forma de estatísticos resumo, foron incluídos en dúas meta-análises adicionais (Kanai et al., 2023; Pairo-Castineira et al., 2023).

4.1 RESULTADOS

4.1.1 Fase de descubrimento

En total, recrutáronse 11.939 casos positivos de COVID-19 en varios centros de España, que foron clasificados acorde a unha escala de gravidade. As características dos individuos poden verse na **Táboa 4**. A maiores, incluíronse dúas cohortes de controis poboacionais: 3.437 controis do Banco Nacional de ADN e 2.506 controis do consorcio GR@ACE (Moreno-Grau et al., 2021). Tras o control de calidade (ver apartado 3.4), quedaron un total de 9.371 casos de COVID-19 e 5.674 controis poboacionais, todos definidos como europeos en base á GIA. Esta cohorte empregouse para as análises principais, mais para as análises de replicación engadimos unha segunda cohorte do Centro Nacional de Investigacións Oncolóxicas (CNIO), composto por 1.598 doentes con COVID-19 e 1.060 controis do BNADN.

Táboa 4. Características basais dos pacientes da cohorte española de SCOURGE incluídos na análise.

Variable		Global N = 9.371	Homes N = 4.343	Mulleres N = 5.028
Idade – media en anos (SD)		62.6 (17.9)	64.3 (16.3)	61.1 (19.1)
Gravidade - N (%)				
	0 - asintomática	582 (6.6)	161 (3.9)	421 (8.9)
	1 - leve	2.689 (30.3)	748 (18.2)	1.941 (40.8)
	2 - moderada	2.099 (23.6)	1.093 (26.5)	1.006 (21.1)
	3 - grave	2.379 (26.8)	1.300 (31.6)	1.079 (22.7)
	4 - crítica	1.128 (12.7)	817 (19.8)	311 (6.5)
Hospitalización - N (%)		5.966 (63.8)	3.436 (79.3)	2.530 (50.5)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Táboa 4. Continuación.

Variable		Global N = 9.371	Homes N = 4.343	Mulleres N = 5.028
Comorbilidades - N (%)				
	Vasculares/endocrinas	4.099 (43.7)	2.207 (50.8)	1.892 (37.6)
	Cardíacas	1.057 (11.3)	634 (14.6)	423 (8.4)
	Nerviosas	773 (8.3)	341 (7.9)	432 (8.6)
	Dixestivas	264 (2.8)	153 (3.5)	111 (2.2)
	Onco-hematolóxicas	647 (6.9)	411 (9.5)	236 (4.7)
	Respiratorias	905 (9.7)	565 (13.0)	340 (6.8)

Neste estudo propuxéronse diferentes fenotipos e tres definicións diferentes da poboación control (ver **Táboa 5**). Analizáronse tres fenotipos no estudo GWAS: susceptibilidade á infección, hospitalización e enfermidade crítica, con tres definición de controis (COVID+ con ausencia de fenotipo+controis poboacionais, controis poboacionais só e COVID+ ausencia de fenotipo). Dadas as similitudes nos resultados, e tendo en conta que o uso de controis poboacionais con diagnóstico descoñecido aumenta a potencia estatística e non os falsos positivos (COVID-19 Host Genetics Initiative et al., 2021), describiremos principalmente os da análise A1, que inclúe tanto doentes COVID+ que non teñen o fenotipo como poboación control, co fenotipo de hospitalización (como indicadora de gravidade) por ser a que maior poder estatístico ten. Os estatísticos resumo dos GWAS poden ser solicitados a través de https://github.com/CIBERER/Scourge-COVID19.

Táboa 5. Definicións de caso e control para os modelos estatísticos.

Fenotipo	Acrónimo	Definición dos casos	Definición dos controis
Risco de hospitalización	A1_HOSP	COVID-19 positivo e hospitalizado	COVID-19 positivo pero non hospitalizado + controis poboacionais
	A2_HOSP	COVID-19 positivo e hospitalizado	Controis poboacionais
	C_HOSP	COVID-19 positivo e hospitalizado	COVID-19 positivo pero non hospitalizado
Risco de enfermidade crítica	A1_CRITICO	COVID-19 positivo con enfermidade crítica (gravidade=4)	COVID-19 positivo pero non crítico (gravidade < 4) + controis poboacionais
	A2_CRITICO	COVID-19 positivo con enfermidade crítica (gravidade=4)	Controis poboacionais
	C_CRITICO	COVID-19 positivo con enfermidade crítica (gravidade=4)	COVID-19 positivo pero non severo (gravidade < 4)
Risco de infección	RISK	COVID-19 positivo	Controis poboacionais

Os estudos GWAS para cada un dos tres fenotipos leváronse a cabo mediante o axuste de modelos de regresión loxística mixtos con SAIGE (W. Zhou et al., 2018) incluíndo as covariables sexo, idade e as 10 primeiras compoñentes principais. As análises tamén se fixeron

estratificadas por sexo e idade (<60 anos, ≥60 anos). As rexións de risco foron sometidas a LD-*clumping* (con parámetros r^2=0.5, pval=5x10^{-8} and $pval_2$=0.05) para obter un grupo de variantes independentes, que serán as variantes líder. Sobre estas últimas axustouse de xeito adicional un modelo que incluía os termos de interacción SEXO*SNP e IDADE*SNP.

4.1.1.1 Análise global

Todas as análises globais detectaron unha asociación significativa nos xa coñecidos *loci* *LZTFL1* (cromosoma 3) e *IFNAR2* (cromosoma 21). Os resultados están resumidos na **Táboa 6**. Na **Figura 9** represéntase o GWAS en forma de *Manhattan plot*, no cal o eixo Y corresponde ao -log10 do p-valor de cada SNP ao longo dos cromosomas. Este gráfico está acompañado do *QQ-plot* (*quantile-quantile plot*), que representa a distribución dos p-valores baixo a hipótese nula de non asociación *versus* os p-valores observados, e permite diagnosticar potenciais problemas de inflación.

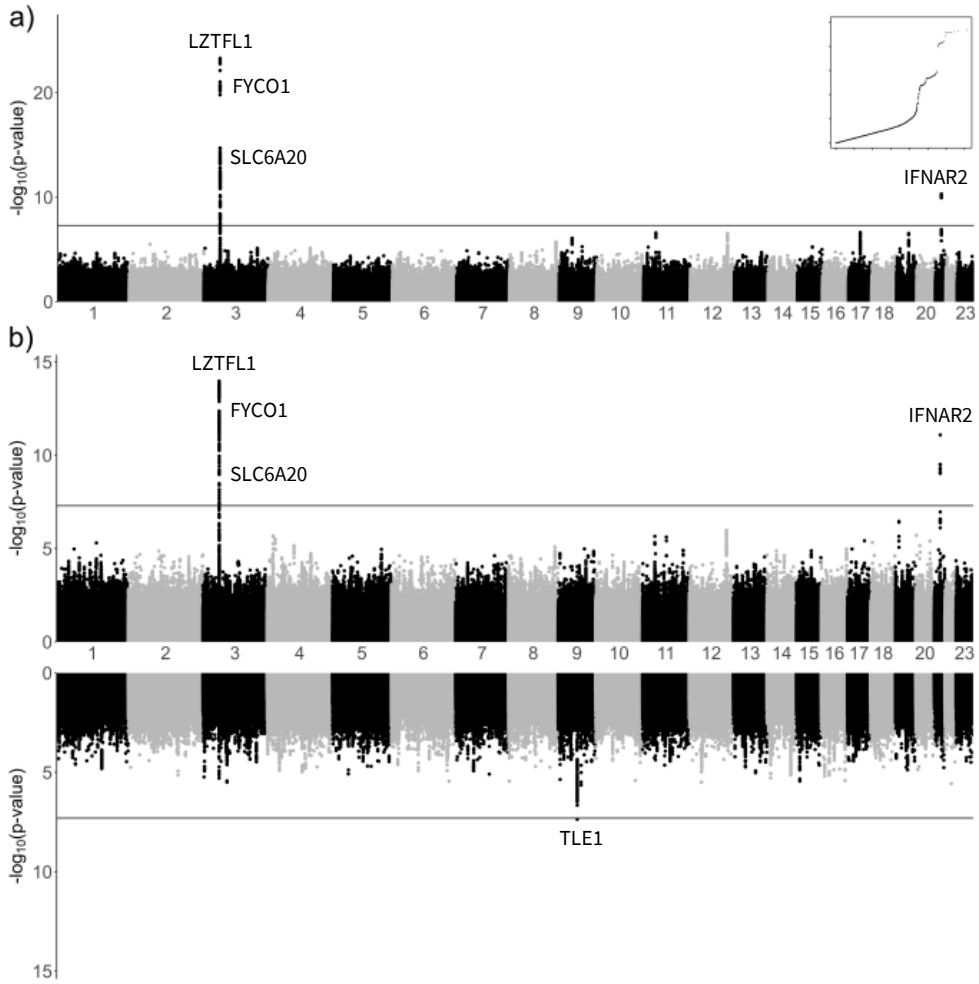


Figura 9. Resultados da asociación da cohort SCOURGE na análise de A1 hospitalización. a) Manhattan plot dos resultados globais. Número de casos=5.934, número de controis=8.810. b) Miami plot dos resultados estratificados por sexo (arriba: homes, abaixo: mulleres). Homes: Número de casos=3.418, número de controis=3.768; mulleres: número de casos=2.514, número de controis=5.042. Publicada en Cruz e colaboradores (2022).

O sinal máis forte corresponde a unha rexión no cromosoma 3, liderada polo SNP rs17763742 (Odds Ratio [OR]=1.79, Intervalo de Confianza [IC] ao 95%=1.59-2.01). Se ben a variante líder pertence ao xen *LZTFL1*, esta rexión comprende varios xenes e foi identificada por primeira vez en 2020 asociada á susceptibilidade na infección (The Severe Covid-19 GWAS Group et al., 2020). O rs17763742 é un *locus* asociado a unha maior expresión do xen *CCR1* na mucosa do esófago (*Normalized Effect Size* [NES]=0.23, segundo GTEx v8). A súa frecuencia alélica mundial é baixa (AF=0.06 para o alelo G, AF=0.09 en europeos, gnomAD v3.1.2) salvo para poboacións do sur asiático, onde ascende a 0.27.

A rexión do cromosoma 21 está liderada polo rs2834158 (OR=1.22, 95% IC=1.15-1.30), que é un eQTL de *IFNAR2* na mucosa do esófago (NES=0.25). Ten unha frecuencia de 0.66 para o alelo C na poboación mundial e de 0.68 en europeos. As análises A2 e C non atoparon sinais adicionais. O mapeamento fino baiesiano priorizou un set de variantes cunha probabilidade posterior de conter á variante causal do 95% para cada rexión, sinalados cun recadro na **Figura 10**.

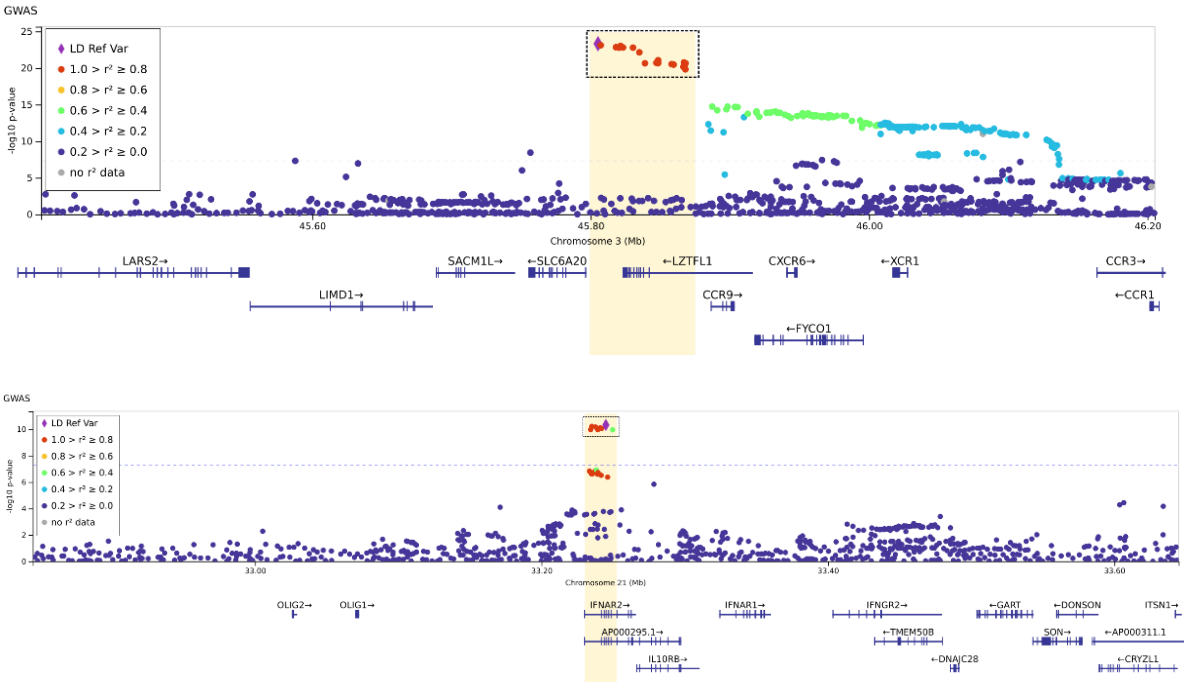


Figura 10. Gráficos rexionais dos dous sinais en *LZTFL1* (arriba) e *IFNAR2* (abaixo). No eixo X represéntase a posición cromosómica, e no eixo Y o -log(p-valor) no estudo SCOURGE. A variante líder está indicada cun diamante e o resto de variantes cun código de cor en función do desequilibrio de ligamento con ela en poboación europea. Os sets cribles para cada sinal están marcados con recadros. Publicada en Cruz e colaboradores (2022).

4.1.1.2 Estratificación por sexo e idade

Os sinais mencionados no apartado anterior resultaron significativos nas análises combinando os dous sexos e en homes, pero non en mulleres (ver **Figura 9**), a pesar de que os tamaños de mostra e *ratios* caso/control entre sexos eran similares. Ambas rexións conteñen SNPs nos que observamos unha diferencia de efectos significativa entre homes e mulleres,

comprobada a través da interacción coa covariable de sexo (**Táboa 6**). Dentro do grupo de mulleres identificouse tamén unha deleción asociada ao risco de hospitalización localizada no xen *TLE1*, o rs140152223 (OR=1.19, 95% IC=1.07-1.31, FA para o alelo T global de 0.11 e de 0.08 en europeos). Na **Figura 11** podemos ver o gráfico rexional do sinal de *TLE1*. Non foi posible relacionar este xen con datos de expresión.

Os GWAS tamén se realizaron estratificando por idade (<60/ ≥60 anos), e por idade e sexo simultaneamente. A diferenza no efecto entre as dúas clases comprobouse nos SNPs independentes (**Táboa 7**). Só se atoparon resultados significativos no subgrupo de <60 anos e homes menores de 60 anos, mentres que a interacción SNP*idade foi significativa para unha variante localizada no xen *LZTFL1* (rs6785091) co fenotipo de enfermidade crítica (en homes).

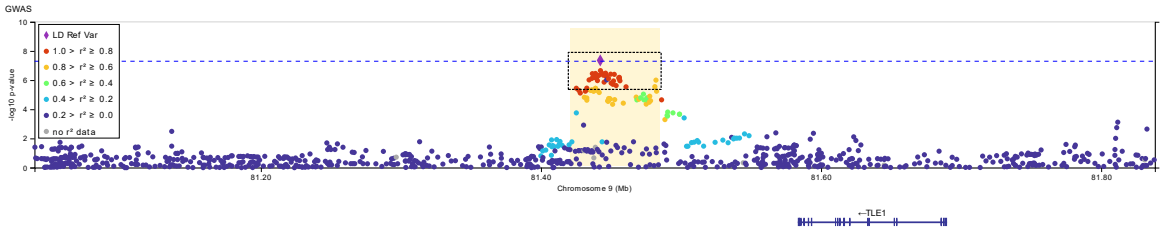


Figura 11. Gráfico rexional do sinal atopado en mulleres no xen *TLE1*. No eixo X represéntase a posición cromosómica, e no eixo Y o -log(p-valor) no estudo SCOURGE. A variante líder está indicada cun diamante e o resto de variantes cun código de cor en función do desequilibrio de ligamento con ela en poboación europea. Publicada en Cruz e colaboradores (2022).

Táboa 6. Resultados das análises A1 en hospitalización e gravidade crítica e susceptibilidade á infección, así como estratificadas por sexo. As variantes incluídas son as independentes tras LD-clumping. A interacción é significativa se $p<4.2x10^{-03}$ axustando por Bonferroni para 12 comparacións. O xen asignado é sempre o máis cercano. ANE: alelo de non efecto; AE: alelo de efecto.

				Todos			Mulleres			Homes			Interacción SNP*SEXO			
SNP	chr:posición	ANE	AE	beta	SE	pval	beta	SE	pval	beta	SE	pval	beta	SE	pval_int	Xen
Risco de hospitalización																
rs17213127	3:45756734	C	T	0,42	0,07	3,8E-09	0,30	0,10	2,9E-03	0,53	0,10	8,4E-08	-0,18	0,14	2,0E-01	*SLC6A20*
rs17763742	3:45805277	A	G	0,58	0,06	5,0E-24	0,38	0,08	5,1E-06	0,75	0,08	1,3E-22	-0,34	0,11	2,9E-03	*LZTFL1*
rs1994492	3:45919154	T	C	0,38	0,05	8,6E-15	0,24	0,07	6,5E-04	0,51	0,07	1,8E-14	-0,24	0,10	1,2E-02	*FYCO1*
rs1994490	3:45966210	C	T	-0,16	0,03	3,9E-08	-0,09	0,04	2,1E-02	-0,22	0,04	6,2E-08	0,10	0,06	7,6E-02	*FYCO1*
rs140152223	9:81442264	TCAG	T	0,17	0,05	1,7E-04	0,35	0,06	4,4E-08	-0,01	0,06	8,6E-01	0,35	0,09	6,6E-05	*TLE1*
rs2834158	21:33244908	C	T	0,20	0,03	4,8E-11	0,14	0,04	1,0E-03	0,27	0,04	3,3E-10	-0,11	0,06	7,0E-02	*IFNAR2*
rs2834164	21:33249643	A	C	-0,19	0,03	1,1E-10	-0,09	0,04	2,9E-02	-0,27	0,04	8,1E-12	0,19	0,06	7,1E-04	*IFNAR2*
Risco de enfermidade crítica																
rs17763742	3:45805277	A	G	0,67	0,09	7,3E-13	0,74	0,18	3,4E-05	0,64	0,11	2,7E-09	-0,34	0,11	2,9E-03	*LZTFL1*
rs6785091	3:45943855	C	G	0,40	0,07	4,9E-08	0,40	0,14	3,1E-03	0,39	0,09	5,8E-06	-0,22	0,09	1,7E-02	*FYCO1*
Risco de infección																
rs17763742	3:45805277	A	G	0,38	0,06	2,4E-11	0,18	0,08	2,7E-02	0,58	0,08	1,2E-13	-0,4	0,11	2,0E-04	*LZTFL1*
rs9636867	21:33237639	A	G	0,19	0,03	3,5E-10	0,13	0,04	1,6E-03	0,26	0,04	1,3E-09	-0,1	0,13	5,1E-02	*IFNAR2*
rs2834164	21:33249643	A	C	-0,15	0,03	2,2E-08	-0,07	0,03	6,0E-02	-0,24	0,04	1,1E-09	0,17	0,05	1,3E-03	*IFNAR2*

Táboa 7. Resultados das análises A1 en hospitalización e gravidade crítica e susceptibilidade á infección estratificadas por idade. As variantes incluídas son as independentes tras LD-clumping. A interacción é significativa se $p<1.5x10^{-03}$ axustando por Bonferroni para 33 comparacións. ANE: alelo de non efecto; AE: alelo de efecto.

					<60			**≥ 60**			**Interacción**		
SNP	chr:posición	ANE	AE		beta	SE	pval	beta	SE	pval	beta	SE	pval
Risco de hospitalización													
rs17213127	3:45756734	C	T	todos	0.450	0.104	1.5E-05	0.366	0.097	1.6E-04	-0.070	0.134	6.2E-01
				homes	0.596	0.134	8.5E-06	0.456	0.145	1.7E-03	-0.030	0.201	8.7E-01
				mulleres	0.238	0.160	1.4E-01	0.288	0.129	2.5E-02	0.010	0.007	4.5E-01
rs17763742	3:45805277	A	G	todos	0.772	0.087	9.2E-19	0.384	0.078	9.2E-07	-0.310	0.107	4.0E-03
				homes	0.980	0.110	5.7E-19	0.513	0.116	9.5E-06	-0.300	0.160	6.0E-02
				mulleres	-	-	-	0.274	0.105	9.1E-03	0.000	0.005	9.2E-01
rs1994492	3:45919154	T	C	todos	0.525	0.071	1.4E-13	0.225	0.067	8.1E-04	-0.270	0.090	3.2E-03
				homes	0.684	0.092	1.3E-13	0.278	0.099	5.1E-03	-0.350	0.131	7.1E-03
				mulleres	0.297	0.106	5.3E-03	0.188	0.090	3.7E-02	0.000	0.004	4.3E-01
rs1994490	3:45966210	C	T	todos	-0.183	0.041	8.5E-06	-0.141	0.039	3.4E-04	0.050	0.055	3.6E-01
				homes	-0.263	0.054	9.8E-07	-0.134	0.059	2.3E-02	0.140	0.078	8.4E-02
				mulleres	-0.069	0.062	2.7E-01	-0.144	0.052	5.7E-03	0.000	0.003	2.0E-01
rs2834158	21:33244908	C	T	todos	0.252	0.044	8.9E-09	0.144	0.043	6.9E-04	-0.120	0.058	3.3E-02
				homes	0.309	0.058	7.8E-08	0.182	0.063	3.9E-03	-0.130	0.084	1.3E-01
				mulleres	0.177	0.066	7.3E-03	0.120	0.057	3.6E-02	0.000	0.003	9.0E-01
rs2834164	21:33249643	A	C	todos	-0.195	0.041	1.8E-06	-0.150	0.039	1.4E-04	0.060	0.055	3.0E-01
				homes	-0.285	0.054	1.3E-07	-0.230	0.059	1.0E-04	0.080	0.079	3.0E-01
				mulleres	-0.070	0.061	2.5E-01	-0.084	0.053	1.1E-01	0.000	0.003	6.7E-01

Táboa 7. Continuación.

					<60			≥ 60			Interacción		
SNP	chr:posición	ANE	AE		beta	SE	pval	beta	SE	pval	beta	SE	pval
Risco de enfermidade crítica													
rs17763742	3:45805277	A	G	todos	1.050	0.172	1.1E-09	0.476	0.108	1.1E-05	-0.402	0.152	8.3E-03
				homes	1.190	0.196	1.4E-09	0.359	0.128	4.9E-03	-0.562	0.180	1.8E-03
				mulleres	-	-	-	0.738	0.200	2.3E-04	-0.004	0.008	6.5E-01
rs6785091	3:45943855	C	G	todos	0.741	0.133	2.2E-08	0.217	0.088	1.3E-02	-0.395	0.131	2.6E-03
				homes	0.855	0.154	2.8E-08	0.133	0.105	2.1E-01	-0.560	0.156	**3.4E-04**
				mulleres	0.339	0.253	1.8E-01	0.401	0.159	1.2E-02	-0.005	0.006	4.5E-01
Risco de infección													
rs17763742	chr3:45805277	A	G	todos	0.422	0.072	5.0E-09	0.378	0.092	4.0E-05	-0.309	0.107	4.0E-03
				homes	0.655	0.095	6.4E-12	0.529	0.136	1.0E-04	-0.042	0.179	8.2E-01
				mulleres	-	-	-	0.245	0.123	4.7E-02	0.001	0.004	6.9E-01
rs9636867	chr21:33237639	A	G	todos	0.193	0.037	2.7E-07	0.191	0.05	1.4E-04	-0.152	0.059	1.0E-02
				homes	0.27	0.052	2.2E-07	0.232	0.075	1.9E-03	-0.042	0.089	6.4E-01
				mulleres	0.117	0.052	2.6E-02	0.168	0.067	1.3E-02	0.002	0.003	6.4E-01
rs2834164	21:33249643	A	C	todos	-0.145	0.035	2.8E-05	-0.156	0.046	7.7E-04	0.057	0.055	3.0E-01
				homes	-0.229	0.048	2.3E-06	-0.235	0.069	6.6E-04	0.022	0.081	7.9E-01
				mulleres	-0.058	0.048	2.3E-01	-0.088	0.062	1.5E-01	0.002	0.005	6.9E-01

4.1.2 Fase de validación

4.1.2.1 Meta-análise CNIO-SCOURGE

A segunda fase do estudo tiña como obxectivo validar as asociacións atopadas dentro da cohorte SCOURGE. Para isto empregamos unha segunda cohorte, a cohorte CNIO, coa cal se repetiron as análises GWAS (que comprendía 1.378 doentes hospitalizados por COVID-19 e 1.068 controis poboacionais). Unha vez obtidos os resultados, procedeuse a meta-analizar ambas cohortes baixo un modelo de efectos fixos con ponderación pola inversa da varianza no *software* METAL. Os resultados móstranse na **Táboa 8** e na **Figura 12**. Ademais, os resultados foron anotados con FUMA para identificar os xenes asignados ás rexións de risco por posición e por eQTLs (ver **Anexo II**). A análise por xenes con MAGMA acadou unha asociación significativa en *LZTFL1*, *XCR1*, *FYCO1*, *CCR9*, *CXCR6* e *CCR1* no cromosoma 3, e OAS1 no cromosoma 12 (**Táboa 9**).

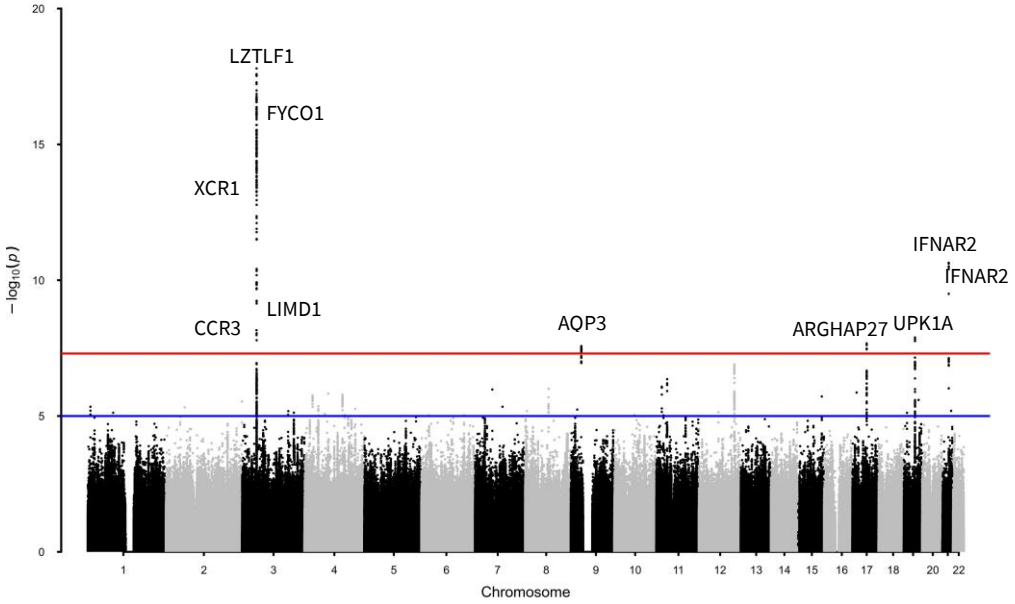


Figura 12. Manhattan plot da meta-análise CNIO-SCOURGE para o fenotipo de hospitalización (análise A1).

Validáronse as asociacións nas rexións do cromosoma 3 e do cromosoma 21. A maiores, detectamos tres *loci* de risco adicionais. No cromosoma 9, rs10813976 (OR=1.2, 95% IC=1.13-1.27) pertence a unha rexión interxénica entre os xenes *AQP7* e *AQP3*. Esta é unha variante común cunha frecuencia global para o alelo G de 0.17 en global e de 0.20 en europeos. O SNP sentinela é particularmente interesante, pois está asociado a un incremento da expresión de *AQP3* no pulmón (NES=0.33) e na mucosa do esófago (NES=0.35) e de *AQP7* no pulmón (NES=0.23) e na mucosa do esófago tamén (NES=0.51), por cada copia do alelo G. O set de variantes causais por mapeamento fino incluíu un total de 22 SNPs (**Figura 13, Anexo III**). Os xenes candidatos, seleccionados a partir da anotación con FUMA, foron *NFX1*, *AQP7* e *AQP3*.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Táboa 8. Resultados da meta-análise de hospitalización entre os estudos SCOUGE e CNIO, tanto en global como estratificados por sexo. Os SNPs representados son os independentes tras o LD clumping. chr:pos: cromosoma-posición; ANE: alelo de non efecto; AE: alelo de efecto; pval: p-valor da asociación.

				Meta-todos			Meta-homes			Meta-mulleres			
SNP	**chr:pos**	**ANE**	**AE**	**beta**	**SE**	**pval**	**beta**	**SE**	**pval**	**beta**	**SE**	**pval**	**Xen**
rs115679256	3:45587795	G	A	0.43	0.08	1.1E-08	0.48	0.10	2.3E-06	0.40	0.11	2.9E-04	*LIMD1*
rs17763742	3:45805277	A	G	0.60	0.05	4.1E-29	0.74	0.07	3.3E-25	0.43	0.08	4.5E-08	*LZTFL1*
rs35477280	3:45932600	G	A	0.39	0.05	1.4E-17	0.48	0.06	6.3E-15	0.28	0.07	1.6E-05	*FYCO1*
rs4443214	3:46136372	T	C	0.25	0.04	9.0E-09	0.26	0.06	1.4E-05	0.26	0.06	4.4E-05	*XCR1*
rs115102354	3:46180545	A	G	0.41	0.07	1.6E-08	0.52	0.10	2.1E-07	0.32	0.10	2.0E-03	*CCR3*
rs10813976	9:33426577	A	G	0.18	0.03	2.7E-08	0.16	0.04	2.5E-04	0.19	0.05	3.5E-05	*AQP3*
rs1230082	17:45422978	C	G	0.16	0.03	2.1E-08	0.17	0.04	2.8E-05	0.15	0.04	2.5E-04	*ARHGAP27*
rs77127536	19:35687796	G	A	-0.22	0.04	1.3E-08	-0.25	0.05	2.1E-06	-0.19	0.05	4.3E-04	*UPK1A / ZTBT32*
rs17860169	21:33240996	A	G	-0.19	0.03	2.3E-11	-0.27	0.04	1.4E-11	-0.12	0.04	3.7E-03	*IFNAR2*

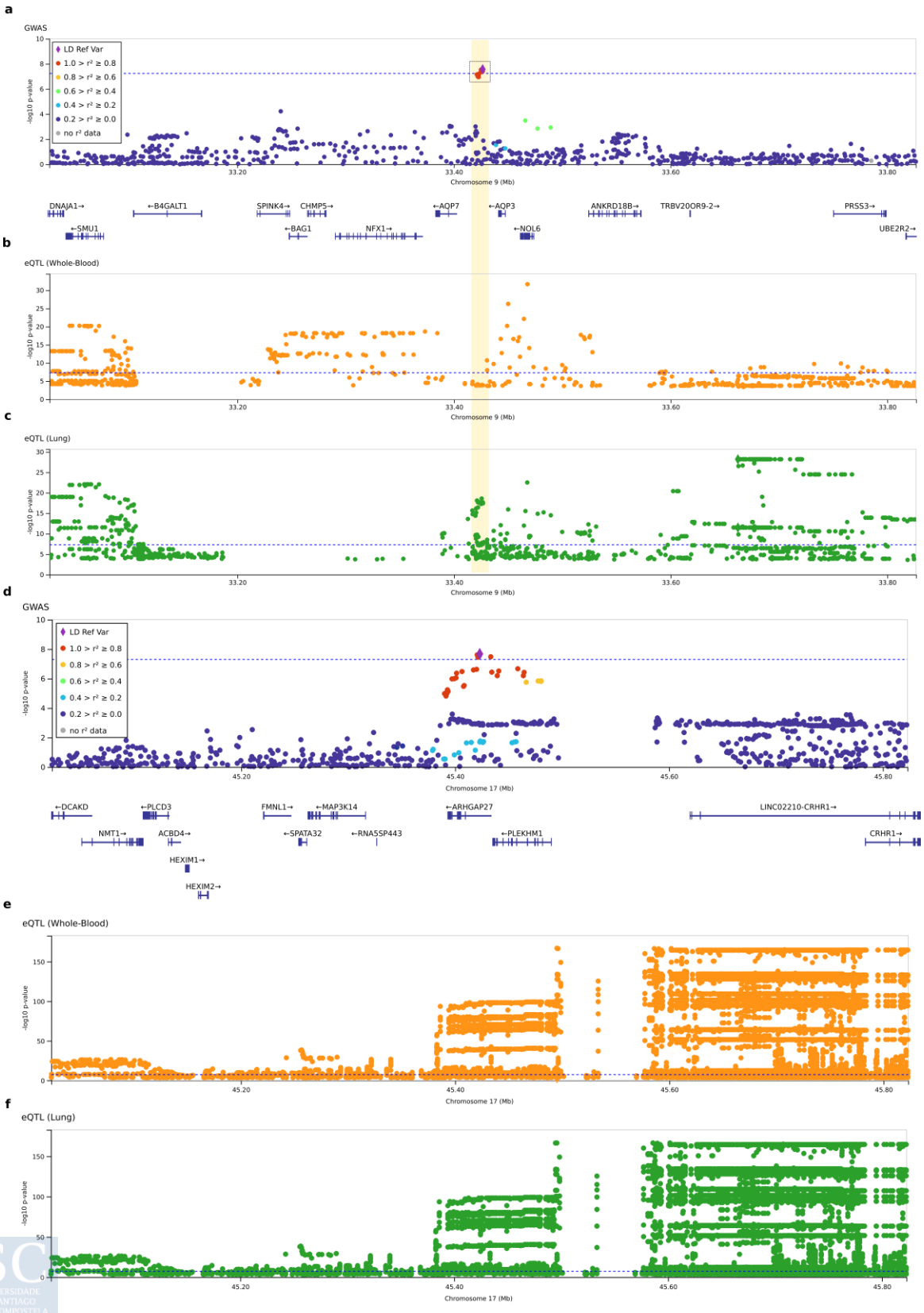
Táboa 9. Resultados da análise a nivel de xen de MAGMA na meta-análise de SCOURGE-CNIO para hospitalización por COVID-19, análise A1 (hg37). Xen_ENS: nome do xen en ENSEMBL; Start: inicio do xen; Stop: final do xen; NSNPS: número de SNPs no xen; NPAR: número de parámetros empregados no modelo; ZSTAT: estatístico Z. Só se mostran os resultados significativos. Aplicouse unha corrección por Bonferroni para 18.632 xenes.

Xen_ENS	Chr	Start	Stop	NSNPS	NPAR	ZSTAT	P	Xen
ENSG00000163818	3	45864808	45957534	181	25	7.64	1.06E-14	*LZTFL1*
ENSG00000173578	3	46058516	46069234	26	3	7.62	1.29E-14	*XCR1*
ENSG00000163820	3	45959396	46037316	223	18	7.35	9.70E-14	*FYCO1*
ENSG00000173585	3	45927996	45944667	38	9	6.16	3.62E-10	*CCR9*
ENSG00000172215	3	45982425	45989845	15	5	5.37	3.92E-08	*CXCR6*
ENSG00000163823	3	46243200	46249887	11	2	4.90	4.79E-07	*CCR1*
ENSG00000089127	12	113344582	113369990	105	3	4.66	1.60E-06	*OAS1*

No cromosoma 17, o rs1230082 (OR=1.17, 95%IC=1.11-2.24) é un SNP intrónico cercano a *ARHGAP27*. Se ben é eQTL en moitos tecidos, FUMA priorizou os xenes *ARHGAP27*, *LRRC37A*, *SPATA32*, *MAPT*, *KANSL1* e *PLEKHM1*. O mapeamento fino non logrou seleccionar un set de variantes causais. O SNP sentinela é un eQTL en pulmón que afecta a expresión dos xenes *LRRC37A* (NES=-0.41), *KANSL1-AS1* (NES=0.35), ou *MAPT* (NES=0.19), e na mucosa do esófago para *ARHGAP27* (NES=-0.094), *PLEKHM1* (NES=-0.14) ou *MAPT* (NES=0.19).

Por último, no cromosoma 19 apareceu outra nova zona de risco, con 18 SNPs priorizadas polo mapeamento fino (**Figura 13**, **Anexo III**). O rs77127536 (OR=0.8, 95%IC=0.74-0.87), coa asociación máis forte, é interxénico a *UPK1A* e *ZBTB32*. Os xenes candidatos foron *ARHGAP33*, *UPK1A*, *ETV2*, *SCN1B*, *IFGLR1*, *U2AF1L4* e *KTM2B*. O SNP é eQTL en pulmón

para *UPK1A* (NES=0.40), e na mucosa do esófago para *IGLFR1* (NES=-0.15) e *ARHGAP33* (NES=-0.18).



USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Figura 13. Gráficos rexionais dos dous sinais en *AQP3*, *ARHGAP27*, e *UPK1A*/*ZTBT32*. No eixo X represéntase a posición cromosómica, e no eixo Y o -log(p-valor) na meta-análise SCOURGE-CNIO. Nos paneis a) e d), a variante líder está indicada cun diamante e o resto de variantes cun código de cor en función do desequilibrio de ligamento con ela en poboación europea. Os sets críbles para cada sinal están marcados con recadros. No resto dos paneis, no eixo X represéntase a posición cromosómica e no eixo Y o -log(p-valor) das análises de eQTL en sangue (b, e e h) e pulmón (c, f e i) en caso de que houbera unha asociación significativa en GTEx V8. Publicada en Cruz e colaboradores (2022).

4.1.2.2 TWAS de hospitalización por COVID-19 en poboación española

Os resultados da meta-análise SCOURGE-CNIO empregáronse para realizar unha análise de asociación de transcriptoma completo co *software* S-MetaXcan. Para a exploración dos tecidos individuais seleccionáronse a mucosa do esófago, o pulmón, linfocitos e tecido sanguíneo. Aplicouse unha corrección por FDR ao 95% ao total de xenes e tecidos, dado que a expresión entre tecidos non é totalmente independente e Bonferroni é máis conservador. A **Táboa 10** recolle os resultados por tecido e xen.

Táboa 10. Resultados da asociación TWAS nos tecidos elixidos. Só se mostran os SNPs significativos mediante corrección por FDR.

Tecido	Xen	Chr	Zscore	Var(g)	Q-valor (FDR)
Linfocitos					
	FYCO1	3	6.94	0.008	$9.55x10^{-08}$
	OAS3	12	4.70	0.023	0.010
	UPK1A	19	-4.61	0.023	0.013
Mucosa do esófago					
	IGFLR1	19	4.72	0.009	0.011
	OAS3	12	4.70	0.009	0.011
	COX6B1	19	4.04	0.015	0.011
Pulmón					
	CXCR6	3	-6.17	0.013	$1.13x10^{-05}$
	CCR5	3	-5.49	0.005	$4.83x10^{-04}$
	AQP3	9	4.83	0.082	$9.18x10^{-03}$
	UPK1A	19	-4.77	0.200	0.011
	ZBTB32	19	-4.68	0.017	0.011
	GAL3ST	2	-4.66	0.090	0.011
	COX6B1	19	4.48	0.009	0.022
Sangue					
	CCR9	3	10.44	0.003	$8.40x10^{-21}$
	UPK1A	19	-5.15	0.170	$2.57x10^{-03}$
	ARHGA	17	5.09	0.001	$2.94x10^{-03}$
	OAS3	12	4.70	0.008	0.011
	COX6B1	19	4.48	0.003	0.022

Se ben a maioría dos xenes xa foran descritos noutras publicacións a través de estudos TWAS e GWAS, algúns son novidosos: *COX6B1*, *IGFLR1* e *GAL3ST2*. Co obxectivo de incrementar a potencia estatística da asociación, levouse a cabo un TWAS multi-tecido co método S-MultiXcan presente no *software* S-MetaXcan. Brevemente, o método consiste en combinar os resultados para todos os tecidos nun test de asociación para cada xen. Neste caso aplicouse unha corrección por Bonferroni para o número de xenes testados (0.05/21.937, $p<2.89x10^{-06}$).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

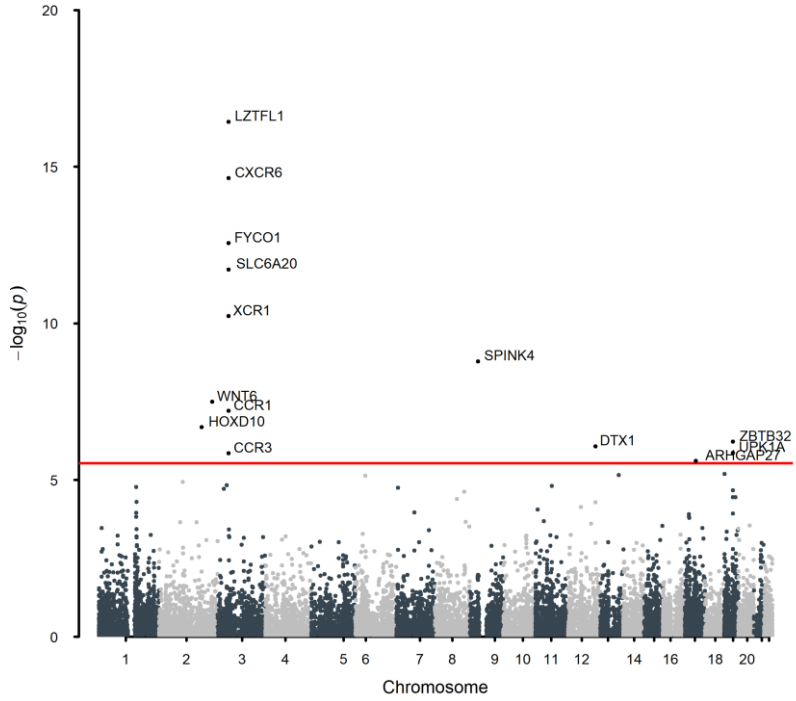


Figura 14. Figura tipo manhattan plot para os resultados do TWAS multi-tecido.

De xeito análogo a un *Manhattan Plot* clásico de GWAS, na **Figura 14** están representados os resultados do TWAS multi-tecido, e no **Anexo IV** poden verse con máis detalle. Con respecto aos resultados dos tecidos individuais, na zona do cromosoma 3, 17 e 19 ambas aproximacións apuntan aos mesmos xenes. Por outra banda, o xen *DTX1* no cromosoma 12 atópase nunha rexión fisicamente achegada ao clúster *OAS*. Aínda que previamente se reportou unha potencial influencia dos SNPs do clúster na expresión deste xen en células B (Schmiedel et al., 2021), cómpre ter cautela ao interpretar esta asociación, xa que o papel dos xenes do clúster *OAS* está ben estudado en relación coa COVID-19, e pode ser un caso de falso positivo por correlación entre a expresión dos mesmos (Wainberg et al., 2019). O mesmo ocorre con *SPINK4*, tamén localizado preto dos xenes *AQP3* e *AQP7*. Pola contra, *HOXD10* e *WNT6* son xenes novel.

4.1.2.3 Avaliación das asociacións das variantes líder en relación coas comorbilidades

Para descartar un posible papel confusor das comorbilidades asociadas á COVID-19 no GWAS, fixemos algunhas análises complementarias no grupo de casos hospitalizados. En primeiro lugar, dentro de cada sexo analizouse un fenotipo indicador da presencia ou non de comorbilidades vasculares, cardíacas, nerviosas, onco-hematolóxicas ou respiratorias, para determinar se algún dos sinais atopados estaba asociado á presencia de antecedentes clínicos (equivalente a unha análise tipo C). Non se acadou ningunha asociación significativa, concluíndo que non existe evidencia dunha posible influencia das comorbilidades (ningún dos sinais atopados era un artefacto a causa das mesmas). A maiores, para cada grupo de

comorbilidades (a modo de variable dependente binaria) axustouse un modelo loxístico incluíndo como covariables sexo, idade e PCs e cada un dos SNPs líder na meta-análise SCOURGE-CNIO e do SNP en *TLE1* (dentro do grupo de mulleres). Ningún SNP estaba significativamente asociado ás comorbilidades (ver **Anexo V**).

4.1.3 Replicación de variantes previamente asociadas coa hospitalización por COVID-19

Durante a época de realización deste estudo identificáronse un total de 13 variantes asociadas coa hospitalización por COVID-19 nas análises do consorcio HGI (COVID-19 Host Genetics Initiative et al., 2021). Na actualidade está publicada a última versión (*freeze 7*), que aumentou o número de asociacións a 40. Para determinar se estas asociacións se replicaban na meta-análise de SCOURGE-CNIO (tanto global como estratificado por sexo), comprobamos o p-valor para cada un dos SNPs no GWAS. A variante rs71175234 (19:45869791:ATT:A) non estaba dispoñible nas meta-análises. Considerouse unha replicación con éxito cando o p-valor $<4.27x10^{-4}$, aplicando unha corrección de Bonferroni para 117 tests (39 SNPs e 3 análises).

Como se ilustra na **Figura 15**, coa cohorte completa replicamos os sinais localizados nos xenes *LZTFL1* (3:45818159:G:A), *IFNAR2* (21:33237639:A:G), *AQP3* (9:33425186:GTAAC:G), *OAS1* (12:112922758:T:C), *ELF5* (11:34480495:C:T), *NR1H2* (19:50379362:T:C) e nunha rexión próxima a *HLA-C* (6:31182658:A:G). No grupo de homes non se replicou o sinal do cromosoma 19. A meta-análise no grupo de mulleres só replicou os sinais dos cromosomas 3, 19 e un adicional no xen SLC22A31 (16:89196249:G:A). Os resultados numéricos poden verse no **Anexo VI**.

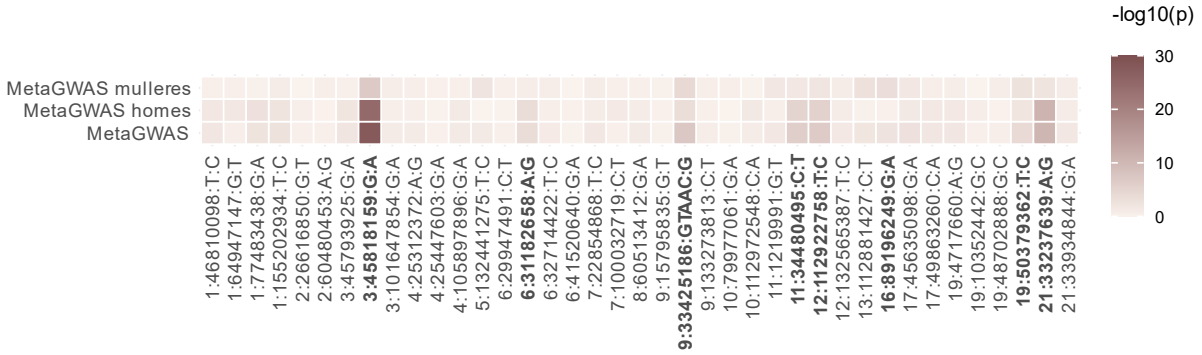


Figura 15. Hitmap de replicación dos hits GWAS para hospitalización. O gráfico ilustra, para as variantes asociadas á hospitalización por COVID-19 na meta-análise do consorcio HGI, o p-valor acadado nas meta-análises de SCOURGE-CNIO. Móstranse en total 39 variantes de 40, pois 19:45869791:ATT:A non se meta-analizou. En negrita, as variantes replicadas cuxo $p<4.27x10^{-04}$ para algún dos grupos (0.05/117 tests).

4.1.4 Meta-análise con estudos europeos

A segunda parte da validación consistiu en meta-analizar as dúas cohortes españolas con outras 4 cohortes que dispoñían de estatísticos resumo disgregados por sexo, todas participantes no HGI consortium: BelCOVID, GEN-COVID, Hostage-Spain e Hostage-Italy (estes últimos

USC DE SANTIAGO DE COMPOSTELA

pertencentes ao estudo de Degenhardt et al., 2022). As asociacións atopadas recóllense na **Táboa 11.**

Ademais das variantes nos cromosomas 3 e 21, tres variantes acadaron significación GWAS na meta-análise no grupo de homes, e dúas de estas tres tamén na análise global. O rs61882275, unicamente asociado en homes (OR=0.84, 95%IC=0.80-0.89) atópase no xen *ELF5* e xa fora identificado con anterioridade en relación á COVID-19 crítica (Kousathanas et al., 2022). Este SNP está asociado a unha redución da expresión de *ELF5* en pulmón (NES=-0.22). Dentro do xen *OAS*, rs4767028 (OR=0.85, 95%IC=0.82-0.89) é unha replicación dunha asociación atopada na primeira análise do consorcio HGI (COVID-19 Host Genetics Initiative et al., 2021). En cambio, o rs12609134 (OR=0.83, 95%IC=0.79-0.88) pertence á rexión no cromosoma 19, entre *UPK1A* e *ZBTB32*, asociada tamén na meta-análise SCOURGE-CNIO a través da variante rs77127536. Os sinais de *TLE* e *AQP3* non acadaron significación GWAS.

Táboa 11. Resultados da meta-análise dos estudos europeos para risco de hospitalización. chr:pos: cromosoma-posición; ANE: alelo de non efecto; AE: alelo de efecto; pval: p-valor da asociación.

				Meta-todos			**Meta-homes**			**Meta-mulleres**			
SNP	**chr:pos**	**ANE**	**AE**	**beta**	**SE**	**pval**	**beta**	**SE**	**pval**	**beta**	**SE**	**pval**	**Xen**
rs115679256	3:45587795	G	A	0.37	0.06	1.3E-08	0.41	0.08	5.6E-07	0.36	0.09	1.6E-04	*LIMD1*
rs13078854	3:45820440	G	A	0.53	0.04	6.7E-34	0.64	0.05	2.7E-33	0.38	0.06	1.0E-09	*LZTFL1*
rs41289622	3:45973053	T	G	0.36	0.04	3.6E-21	0.44	0.05	3.4E-20	0.27	0.05	7.2E-07	*FYCO1*
rs115102354	3:46180545	A	G	0.40	0.06	8.9E-12	0.48	0.07	6.8E-11	0.26	0.08	1.8E-03	*XCR1*
rs61882275	11:34482745	G	A	-0.12	0.02	1.0E-06	-0.17	0.03	4.1E-08	-0.08	0.03	1.3E-02	*ELF5*
rs4767028	12:112921383	A	G	-0.16	0.02	1.6E-10	-0.19	0.03	2.5E-09	-0.11	0.04	8.7E-04	*OAS1*
rs12609134	19:35687796	G	A	-0.19	0.03	2.3E-08	-0.22	0.04	9.5E-08	-0.13	0.05	6.0E-03	*UPK1A / ZBTB32*
rs17860169	21:33240996	A	G	0.18	0.03	3.9E-12	0.21	0.03	1.6E-10	0.15	0.04	2.9E-05	*IFNAR2*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

4.2 DISCUSIÓN

Este estudo foi o GWAS máis grande sobre a gravidade do COVID-19 levado a cabo en España. Dende a súa publicación en 2022 identificáronse moitos outros sinais, polo que os resultados se discutirán tamén en base a investigacións publicadas posteriormente.

A cohorte de SCOURGE España contén doentes de varios lugares do país, recrutados durante a pandemia, pero cunha definición homoxeneizada das variables clínicas recollidas para todos os centros. Así, asegurouse unha representación de todas as zonas, reducindo potenciais nesgos derivados dos criterios de recrutamento. A base de datos recolle máis de 200 variables clínicas, aínda que dadas as condicións do momento da recollida de datos, só algunhas delas están suficientemente completas como para poder traballar con elas. Por outra banda, a pesar de que a gran maioría de estudos se centraron no fenotipo máis crítico, a cohorte de SCOURGE abarca fenotipos asintomáticos, moderados e graves. Analizáronse tres fenotipos (risco de infección, enfermidade crítica e hospitalización), mais a análise con maior potencia estatística foi a de hospitalización, e non se observaron resultados adicionais noutros fenotipos.

4.2.1 Diferencias no risco xenético entre sexos e grupos de idade

Nos GWAS de SCOURGE para hospitalización detectamos asociacións en dúas rexións a través das diferentes análises (*LZTFL1* no cromosoma 3 e *IFNAR2* no cromosoma 21), as cales xa foran previamente publicadas (COVID-19 Host Genetics Initiative et al., 2021; The GenOMICC Investigators et al., 2021; The Severe Covid-19 GWAS Group et al., 2020). Ademais, atopamos unha diferencia na base xenética entre sexos. Concretamente, os sinais do cromosoma 3 e do 21 presentaron unha diferencia na magnitude do efecto e da asociación entre homes e mulleres. Por exemplo, para o SNP rs17763742 do xen *LZTFL1*, no que a interacción SNP*SEXO foi significativa en A1-hospitalización ($p=2.29x10^{-03}$), o tamaño de mostra no grupo de mulleres (2.514 casos e 5.042 controis) tería suficiente potencia estatística para detectar un efecto equivalente ao do SNP no grupo de homes (OR=1.79) asumindo unha mesma FA. En cambio, o efecto asociado ao SNP no grupo das mulleres é considerablemente menor (OR=1.46).

Ademais, identificamos un xen que só estaba significativamente asociado no grupo das mulleres, *TLE1*. *TLE1* codifica un intensificador tipo transducina, co-represor de factores de transcrición. Este xen regula negativamente a ruta de sinalización NF-κB (Tetsuka et al., 2000). De feito, o bloqueo da súa expresión en ratos produciu un fenotipo de inflamación crónica, causado polo aumento da liberación de citoquinas e quemoquinas en varios tecidos, entre os que se atopa o pulmón (Ramasamy et al., 2016). En relación coa COVID-19, *TLE1* foi identificado como unha das 332 proteínas diana das proteínas virais do SARS-CoV-2 (Gordon et al., 2020), e tamén como un dos xenes cuxa expresión permite distinguir entre COVID-19 grave e leve (De Jong et al., 2021).

Dende o inicio da pandemia, recoñeceuse que o sexo biolóxico é un factor de risco na progresión da enfermidade, especialmente no que respecta á mortalidade (N. Chen et al., 2020; Jun et al., 2021; Pradhan & Olsson, 2020; Scully et al., 2020). Unha das posibles razóns reside na resposta inmune. Durante a enfermidade, os homes presentan niveis máis altos de citoquinas,

USC UNIVERSIDADE DE COMPOSTELA

como IL-8 e IL-18, así como de CCL5, mentres que as mulleres mostran unha activación maior das células T CD8 (Takahashi et al., 2020). Isto concorda con observacións previas, tanto na resposta inmune innata como na adaptativa (S. L. Klein & Flanagan, 2016; Scully et al., 2020). Por outra banda, moitos xenes ubicados no cromosoma X regulan respostas inmunes. Aínda que o cromosoma X adicional en homes se inactiva, algúns xenes como *TLR7* escapan a esta inactivación (Libert et al., 2010). A presencia de variantes deletéreas no xen *TLR7* asociouse con unha resposta danada dos interferóns tipo I e II en homes, o que empeora a progresión da COVID-19 (Asano et al., 2021; Fallerini et al., 2021; Van Der Made et al., 2020). Ademais, as hormonas xogan un papel fundamental na regulación da resposta inmune. *ACE2*, codificado no cromosoma X, está regulado por estróxenos e mostra unha maior expresión en homes (J. Liu et al., 2010; Takahashi et al., 2020). Por outro lado, un aumento nas cantidades de estradiol aumenta os niveis de neutrófilos en sangue e pulmóns, pero reduce a produción de citoquinas inflamatorias (S. L. Klein & Flanagan, 2016). Un estudo en ratos infectados co SARS-CoV revelou que os machos presentaban un maior dano alveolar, causado en parte por unha sobre-regulación na produción de citoquinas e quemoquinas inflamatorias, mentres que nas femias se atopou un efecto protector dos estróxenos (Channappanavar et al., 2017).

Con todo, ningún estudo adicional observou diferencias significativas entre sexos, polo que os nosos resultados aínda non foron replicados. Degenhardt e colaboradores (2022) testaron o efecto do sexo e da idade en 7 SNPs candidatos (en *LZTFL1*, *ABO*, *TYK2*, *DPP9*, *MAPT*, *NAPSA* e *IFNAR2*). No material suplementario do artigo, para o SNP rs35731912 (*LZTFL1*) reportan unha OR de 1.75 en mulleres e de 1.96 en homes, mais a diferencia non é estatisticamente significativa (Degenhardt et al., 2022). O mesmo ocorre tras estratificar por sexo e idade, reportando unha OR de 2.21 en homes menores de 60 anos e unha OR de 1.59 para mulleres menores de 60 anos. Pola súa banda, Pairo-Castineira e colaboradores (2020) observan asociacións máis febles nas mulleres que nos homes para as variantes rs73064425, rs143334143 e rs3131294, mais os efectos tampouco son significativamente distintos, suxerindo que a diferencia na forza de asociación pode ser debida a unha falta de potencia estatística no grupo feminino. No estudo que combina as cohortes de GenOMICC e SCOURGE, meta-analizáronse os grupos femininos e masculinos, e testouse a diferencia nos efectos mediante un test *t*, pero ningún SNP superou a corrección GWAS (Pairo-Castineira et al., 2023). Cómpre sinalar que o número de estudos meta-analizados na nosa investigación é limitado, dado que durante as datas de desenvolvemento desta investigación só tivemos acceso aos 4 resultados desglosados por sexo que incluímos.

Baixo esta variabilidade entre sexos, podería estar actuando o fenómeno de amplificación: as variantes e a dirección dos efecto son compartidos, pero o efecto é maior para un dos sexos (C. Zhu et al., 2023). Esta interacción podería ser tamén o resultado dun posible nesgo de participación, motivado por causas sociais ou clínicas, o que podería xerar diferencias nas FAs e asociacións espurias co sexo (Pirastu et al., 2021). En definitiva, se ben está claro que o sexo é un factor crucial na enfermidade e mortalidade da COVID-19, cabe a posibilidade de que os resultados do estudo de SCOURGE estean influenciados pola composición da mostra analizada. Niemi e colaboradores (2022) argumentan que non é frecuente identificar factores xenéticos

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

específicos de sexo en enfermidades do sistema inmune, aínda cando presentan diferencias nas prevalencias.

No estudo de SCOURGE, obtivemos unha interacción significativa do SNP rs6785091 (cromosoma 3) coa idade dentro do grupo de homes con enfermidade crítica. Ademais, só se observaron asociacións significativas dentro do grupo de homes menores de 60 anos. A diferenciación xenética por idade replicouse noutros estudos (Degenhardt et al., 2022; Nakanishi et al., 2021). Nakanishi e colaboradores (2021) exploraron en profundidade o risco conferido polo rs10490770, tamén localizado na rexión do cromosoma 3. Detectaron que os portadores do alelo de risco cunha enfermidade máis grave eran, de media, 2.3 anos menores, mentres que os portadores menores de 60 anos tiñan unha OR de morte ou fallo respiratorio superior aos maiores de 60 anos non portadores (2.7 vs 1.5). É plausible que os factores xenéticos teñan máis influencia nos pacientes de menor idade dado que, para individuos de idade máis avanzada, as comorbilidades e o envellecemento comezan a ter un peso maior sobre o risco de padecer un cadro clínico máis grave.

4.2.2 Asociacións nos xenes *AQP3, ARHGAP27* e *UPK1A*

A meta-análise coa cohorte CNIO revelou novas asociacións, tanto no GWAS como na análise a nivel de xen. A nivel de xen, replicamos as asociacións de *LZTFL1*, *FYCO1*, *CCR9*, *CXCR6, CCR1* e *OAS1* (COVID-19 Host Genetics Initiative et al., 2021; Degenhardt et al., 2022; Kanai et al., 2023; Kousathanas et al., 2022; Pairo-Castineira et al., 2023, 2023; The Severe Covid-19 GWAS Group et al., 2020).

O primeiro marcador de relevancia no GWAS (rs10813976), está localizado nunha zona do cromosoma 9 que codifica varias acuaporinas. O SNP afecta á expresión dos xenes *AQP7* e *AQP3* nos tecidos de mucosa do esófago e pulmón. Diversas acuaporinas están involucradas en patoloxías pulmonares, como fibrose, EPOC, ou edemas, e son responsables do transporte de fluídos entre o endotelio e epitelio pulmonares (Yadav et al., 2020). Atendendo á información biolóxica, esta apunta á acuaporina 3 como potencial xen causal. Ademais, *AQP3* foi o único xen da rexión asociado no TWAS. *AQP3* exprésase principalmente nos riles, no tracto gastrointestinal, nas células do epitelio basal dos condutos aéreos e a nasofarinxe, na pel e na vexiga; *AQP7* exprésase nos testículos e células de graxa (Song et al., 2000). Por outro lado, *AQP3* está sobre-regulado no tecido pulmonar durante infeccións bacterianas ou víricas que inducen dano alveolar (Pires-Neto et al., 2016). O síndrome de dificultade respiratoria aguda caracterízase, en parte, por unha acumulación de fluído (edema) no pulmón como consecuencia dunha permeabilidade incrementada na membrana alveolar. Nesta liña, xa foi suxerido o uso de moduladores de acuaporinas para tratar a inflamación pulmonar e edema causadas pola COVID-19 (Mariajoseph-Antony et al., 2020).

No cromosoma 17, a rexión de risco comprendía varios xenes, entre elas *ARHGAP27*, *PLEKHM1*, *LRRC37A* ou *MAPT*, en consonancia cos resultados de Degenhardt e colaboradores (2022) e resaltando o gran tamaño da rexión. *ARHGAP27* obtivo unha asociación significativa no TWAS en tecido sanguíneo Esta zona, tamén denominada 17q21.31, foi inicialmente identificada polo consorcio HGI a través do rs1819040, no xen *KANSL1*. Corresponde a unha inversión do tamaño dunha megabase, cuxa forma H2 está sometida a selección positiva en

europeos (COVID-19 Host Genetics Initiative et al., 2021; Stefansson et al., 2005). Outro estudo europeo atopou un efecto da idade no risco conferido pola inversión (Degenhardt et al., 2022).

No cromosoma 19, identificouse unha rexión interxénica entre *UPK1A* e *ZBTB32*. A análise TWAS identificou unha expresión diferencial entre casos e controis de *UPK1A* en linfocitos, pulmón e tecido sanguíneo. Este xen codifica a uroplaquina, unha proteína transmembrana que participa no funcionamento do epitelio da vexiga. O seu papel no pulmón non está claro, se ben unha análise bioinformática relacionou a *UPK1A* cunha peor prognose do adenocarcinoma pulmonar (Y. Zhu et al., 2023). Outros xenes como *ZBTB32*, tamén significativo no TWAS, ou *ARHGAP33*, teñen máis sentido biolóxico. O primeiro é un dedo de zinc e dominio de BTB cuxa expresión é decisiva no crecemento da poboación de células NK tras unha infección (Beaulieu et al., 2014). Ademais, outro estudo demostrou que *ZBTB32* é crucial na regulación das células $CD8^+$: o bloqueo da súa expresión en ratos provocou infeccións sistémicas e dano pulmonar grave (H. M. Shin et al., 2017). Pola súa banda, *ARHGAP33* está regulada transcripcionalmente por IRF1, un efector antiviral estimulado por interferóns (Schoggins & Rice, 2011). Esta rexión mantivo a súa asociación ao incluír os 4 estudos adicionais do HGI, mais non foi replicado en meta-análises posteriores.

4.2.3 Novos xenes identificados a través dos estudos TWAS

Mediante as análises TWAS replicáronse as asociacións da rexión do cromosoma 3 (*FYCO1*, *LZTFL1*, *SLC6A20*, *XCR1*, *CCR1*, *CCR3*), 12 (*OAS3*), 9 (*AQP3*), 17 (*ARHGAP27*) e 19 (*UPK1A/ZBTB32*). Adicionalmente, descubríronse xenes novidosos que non foron reportados polas principal análises transcriptómicas para COVID-19 (Pairo-Castineira et al., 2023; Pathak et al., 2021).

Entre eles, *COX6B1* sitúase no cromosoma 19, codifica unha citocromo C oxidasa e foi relacionado coa COVID-19 en máis dun estudo previo. Observouse que *COX6B1* se expresaba menos no tecido pulmonar de pacientes cunha carga viral de SARS-CoV-2 alta que en aqueles con menor carga viral (Kulasinghe et al., 2022), e tamén menos nos neutrófilos dos enfermos de COVID-19 (Borella et al., 2022). Ademais, atopouse que durante a infección por SARS-CoV-2, o virus induce o microRNA miR-2392, e este regula á baixa a expresión do xen *COX6B1* en células nasais e corazón (McDonald et al., 2021). Por outra parte, *IGFLR1* (tamén no cromosoma 19) é un receptor de membrana para as proteínas da familia IFG (factores de crecemento de tipo insulina), cunha afinidade maior por IGFL1 e IGFL3. As proteínas IGFL1 estimulan a produción de células inmunes, e a ruta de sinalización IGF-1 está implicada no desenvolvemento do pulmón (Auernhammer & Strasburger, 1995; J. Shin et al., 2022). J. Shin e colaboradores (2022) detectaron que, en células infectadas polo SARS-CoV-2, os xenes implicados na ruta de sinalización de insulina/IGF se expresaban menos, o que producía unha resposta mediada por interferóns esaxerada nos tecidos de pulmón, fígado, adiposo e células pancreáticas. En definitiva, a rexión de risco do cromosoma 19 apuntou a varios xenes con funcións biolóxicas relevantes e potencialmente causais. O TWAS multi-tecido identificou dous adicionais, localizados no cromosoma 2. *HOXD10* produce unha proteína da familia *homeobox*, involucrada no desenvolvemento embrionario e a diferenciación celular, que non

foi relacionado coa COVID-19, polo que non se pode descartar que sexa un falso positivo. Finalmente, o xen *WNT6* codifica unha proteína da familia WNT. A ruta Wnt/β-Catenina está relacionada con procesos da resposta inmune (Staal et al., 2008), e *WNT5A* e *WNT11* foron propostos como biomarcadores para o síndrome respiratorio da COVID-19 (E. Y. Choi et al., 2020).

4.2.4 Integración dos resultados de SCOURGE noutros estudos

Os resultados deste estudo datan dun momento no que aínda non se publicaran as análises finais do consorcio HGI, polo que moitos dos nosos descubrimentos non se puideron validar ata tempo despois, e os que non foron replicados deben tomarse con cautela. Non obstante, non é infrecuente que, dependendo das cohortes incluídas, as meta-análises para un mesmo fenotipo varíen nalgúns dos seus resultados. Por exemplo, a asociación do cromosoma 9 correspondente ao xen *AQP3* perdeuse ao incluír -na nosa meta-análise- as cohortes de BelCOVID, GenCOVID, Hostage-Spain e Hostage-Italy. Curiosamente, esta asociación non foi atopada en GWAS con tamaños de mostra superiores, e só foi replicada cando se incluíu o estudo de SCOURGE nas análises (Kanai et al., 2023; Pairo-Castineira et al., 2023). A variante líder, rs10813976, ten unha FA en Europa do 20% (segundo gnomAD v4.3.6) e do 17% na mostra global, mais nos controis españois presentaba unha FA do 22%. Así, este é un exemplo de como a variabilidade poboacional, incluso a nivel sub-continental, pode potenciar o descubrimento de novos sinais GWAS. En cambio, o sinal do cromosoma 19 en *UPK1A* mantívose ao considerar as 4 cohortes do HGI (aínda que con menor p-valor), e non nas meta-análises máis recentes. Tamén é importante ter en conta que os estudos poden diferir na definición do fenotipo, no momento da pandemia e nas variantes do virus coas que os individuos se infectaron, ou incluso a composición xenética da mostra.

Os datos do GWAS de SCOURGE para hospitalización, enfermidade crítica e susceptibilidade contribuíron en gran medida aos resultados obtidos na versión 7 dos estudos do consorcio HGI (Kanai et al., 2023), cos que se identificaron un total de 51 variantes diferentes involucradas na enfermidade. O estudo de SCOURGE foi o terceiro que máis casos de COVID-19 con fenotipo crítico aportou, e o segundo con máis casos hospitalizados por COVID-19. Combinouse noutra meta-análise coas cohortes ISARICC e GenoMICC, identificando 49 *loci*, entre os que se atopan potenciais dianas farmacolóxicas como *JAK2*, *PD4A* ou *TNF* (Pairo-Castineira et al., 2023).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5 FACTORES XENÉTICOS DA COVID-19 EN POBOACIÓN LATINOAMERICANA

Parte do contido do presente capítulo foi publicado como *Reviewed preprint* en Diz-de Almeida, S., Cruz, R.,...,Carracedo, A. (2024) Novel risk loci for COVID-19 hospitalization among admixed American populations. *eLife*, 13:RP93666, e será publicado na revista tras o proceso de revisións.

5.1 RESULTADOS

5.1.1 GWAS na cohorte LatAm

Na cohorte latinoamericana de SCOURGE analizáronse un total de 3.729 casos positivos de COVID-19 recrutados en 6 países de Latino-América (**Táboa 12**), e 1.082 doentes recrutados en España con evidencia de ter orixe latinoamericana. Tras o control de calidade e as estimas de ancestralidade quedaron un total de 3.512 individuos para incluír nas análises. Na **Figura 16**, panel A, pode observarse a distribución dos individuos nos PC1 e PC2, mentres que no panel B están representadas as tres compoñentes ancestrais. Nestes gráficos pódese comprobar que non é posible dividir os individuos en clústeres definidos. A compoñente ancestral europea é o 47% do total; a compoñente ancestral nativo-americana, o 40%, e a compoñente africana, o 12%.

O estudo de asociación GWAS para hospitalización levouse a cabo a través do axuste de modelos loxísticos mixtos con SAIGE, para os cales se engadiron as covariables de sexo, idade, as 10 primeiras compoñentes principais e unha variable adicional para clasificar os individuos nas áreas de recrutamento (Brasil, Colombia, Ecuador, México, Paraguay e España). Con esta variable pretendeuse corrixir polas diferenzas nos criterios de hospitalización e recrutamento entre os países. Esta análise non tivo ningún resultado significativo, dado o tamaño de mostra limitado (1.625 casos hospitalizados e 1.887 casos COVID-19 non hospitalizados). Por exemplo, a potencia estatística para detectar o SNP rs17763742 do xen LZTFL1 (**Táboa 6** no Capítulo 4), coa OR na cohorte española e unha FA para o alelo G en poboación AMR de 0.049, era do 88%. A estratificación por sexo e por idade non se fixo por este motivo. Ademais, só se puido realizar o equivalente á análise C da cohorte europea, é dicir, a de dentro de casos, posto que non dispoñíamos dunha cohorte de poboación control xeneticamente similar á de casos.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

A modo comparativo, comprobamos os resultados para os SNPs sentinela nos cromosomas 3 e 21 no GWAS da cohorte española de SCOURGE, rs17763742 (*LZTFL1*) e rs2834158 (*IFNAR2*). O rs17763742 obtivo unha OR de 1.73 (95%IC=1.29-2.32, p=2.4x10^{-04}), e o rs2834158, unha OR de 1.18 (95%IC=1.02-1.35, p=0.024).

Táboa 12. Características basais dos pacientes da cohorte LatAm de SCOURGE incluídos na análise. SD=desviación típica; GIA: ancestralidade xenética inferida.

Variable		Non hospitalizados (N=1.887)	Hospitalizados (N = 1.625)
Idade – media en anos ± SD		39.1 ± 11.9	54.2 ±14.5
Sexo - N (%)			
	Mulleres (%)	1253 (66.4)	668 (41.1)
GIA* – % media ±SD			
	Europea	54.4 ±16.2	39.4 ± 20.7
	Africana	15.3 ± 12.7	9.1 ± 11.6
	Nativo americana	30.3 ± 19.8	51.3 ± 26.5
Comorbilidades - N (%)			
	Vasculares/endocrinas	488 (25.9)	888 (64.5)
	Cardíacas	60 (3.2)	151 (9.3)
	Nerviosas	15 (0.8)	61 (3.8)
	Dixestivas	14 (0.7)	35 (2.1)
	Onco-hematolóxicas	21 (1.1)	48 (3.0)
	Respiratorias	76 (4.0)	119 (7.3)

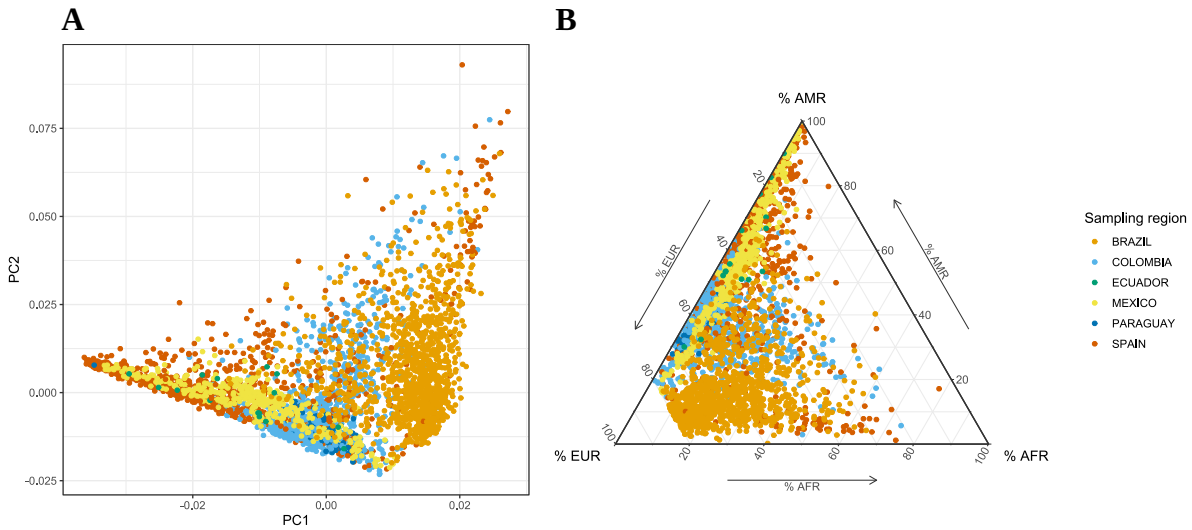


Figura 16. Representación da ancestralidade dos individuos da cohorte Latinoamericana. No panel A están representadas as PC1 e PC2, as cales recollen a maior porcentaxe de variabilidade xenética. O panel B representa as tres compoñentes ancestrais globais estimadas co software ADMIXTURE: europea (EUR), africana (AFR) e nativo-americana (AMR). A GIA foi estimada a partir dun panel de referencia composto por individuos nativo-americanos (Aymara, Quechua, Nahua e Mayas) e individuos seleccionados aleatoriamente das poboacións europea e africana de 1KGP. As cores representan as diferentes áreas xeográficas de recrutamento. *Sampling region*: país de recrutamento.

5.1.2 Meta-análise dos estudos AMR

Os resultados do GWAS de LatAm foron meta-analizados cos resultados de GWAS de hospitalización por COVID-19 doutros estudos latinoamericanos participantes no consorcio HGI. Así, os estatísticos resumo destas análises descargáronse do repositorio público do consorcio (https://www.covid19hg.org/results/r7/), correspondendo as análises específicas de poboación AMR co fenotipo B2 da versión 7 (individuos hospitalizados por COVID-19 *vs* individuos COVID+ non hospitalizados e controis poboacionais). Para a meta-análise empregouse o *software* METAL co modelo ponderado pola inversa da varianza. En total, incluíronse 7 estudos a maiores do de SCOURGE (**Táboa 13**), sumando un total de 4.702 individuos latinoamericanos hospitalizados e 68.573 controis. O noso estudo é o máis grande en canto ao número de casos aportados.

Analizáronse 8.769.960 variantes cunha r^2 de imputación >0.8 no estudo SCOURGE e cun r^2 >0.6 no HGI. Os resultados deste estudo foron sometidos a anotación funcional coa ferramenta FUMA (Watanabe et al., 2017). Anotáronse as variantes con p $<5x10^{-08}$ ou en desequilibrio de ligamento (r^2 >0.6) con elas (calculado co panel AMR de 1000 Genomas). Os *loci* de risco definíronse colapsando bloques de desequilibrio de ligamento cunha lonxitude de 250kb. Despois, FUMA anotou co *software* ANNOVAR os xenes, os valores CADD v1.4 e os valores RegulomeDB v1.1. Tamén se levou a cabo unha análise a nivel de xen co *software* MAGMA (de Leew et al., 2015), baixo o modelo de media dos SNPs. A significación para este test situouse no límite de p $<2.66x10^{-6}$ (axustando para un total de 18.817 xenes).

Táboa 13. Cohortes adicionais incluídas na meta-análise SCOURGE-HGI$_{AMR}$.

Estudo	Núm. Casos	Núm. Controis
ANCESTRY_Freeze_Four_AMR	199	12.323
BRACOVID_AMR	853	835
MVP_AMR	517	43.062
ChiCovid_AMR	143	522
MexGen-COVID_AMR	876	5.060
UCLA_AMR	166	4.686
MexCovid_AMR	323	198

FUMA permitiu un mapeamento inicial dos resultados a través de dúas aproximacións: (1) mapeamento por posición, que asigna variantes aos xenes a partir da distancia física ao xen máis cercano empregando ventás de 10kb; e (2) mapeamento por *loci* de expresión cuantitativa (eQTL), baseado en datos de expresión en GTEx v.8. para tecidos de sangue, linfocitos, pulmón e mucosa do esófago (establecendo un FDR de 0.05 para os pares variante-xen).

Esta meta-análise de GWAS (GWAS-meta) identificou 4 rexións asociadas a COVID-19 severo (**Figura 17**). Estas rexións foron sometidas a un LD-clumping (p$<5x10^{-08}$, distancia=1500kb, r^2=0.1), quedándonos un total de 4 variantes líder (**Táboa 14**) ligadas a outras 310 variantes e 31 xenes mapeados por FUMA (ver **Anexo VII**).

USC DE COMPOSTELA

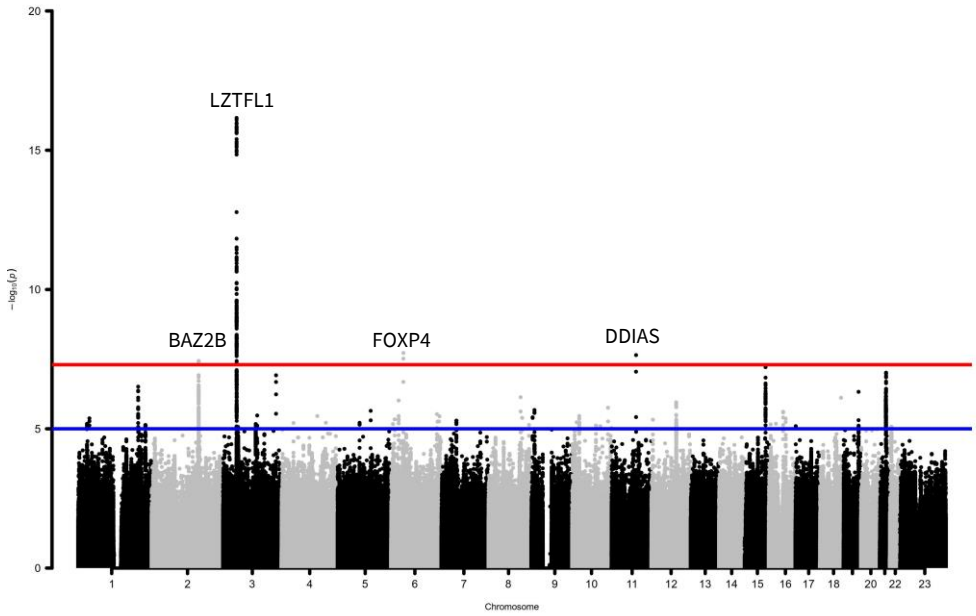


Figura 17. Manhattan plot da meta-análise de HGI AMR con SCOURGE Latino-América. Os límites de probabilidade GWAS ($p=5x10^{-08}$) e sinais suxestivos ($p=1x10^{-05}$) están marcados por liñas horizontais.

Táboa 14. Variantes líder na meta-análise de HGI AMR con SCOURGE Latino-América. AE: alelo de efecto; ANE: alelo de non efecto; pval: p-valor da asociación; FAE: frecuencia do alelo de efecto no estudo SCOURGE.

SNP	chr:pos	AE	ANE	OR (95% CI)	pval	FAE casos	FAE controis	Xen
rs13003835	2:159407982	T	C	1.20 (1.12-1.27)	3.66E-08	0.563	0.429	*BAZ2B*
rs35731912	3:45848457	T	C	1.65 (1.47-1.85)	6.30E-17	0.087	0.056	*LZTFL1*
rs2477820	6:41535254	A	T	0.84 (0.79-0.89)	1.89E-08	0.453	0.517	*FOXP4*
rs77599934	11:82906875	G	A	2.27 (1.7-3.04)	2.26E-08	0.016	0.011	*DDIAS*

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

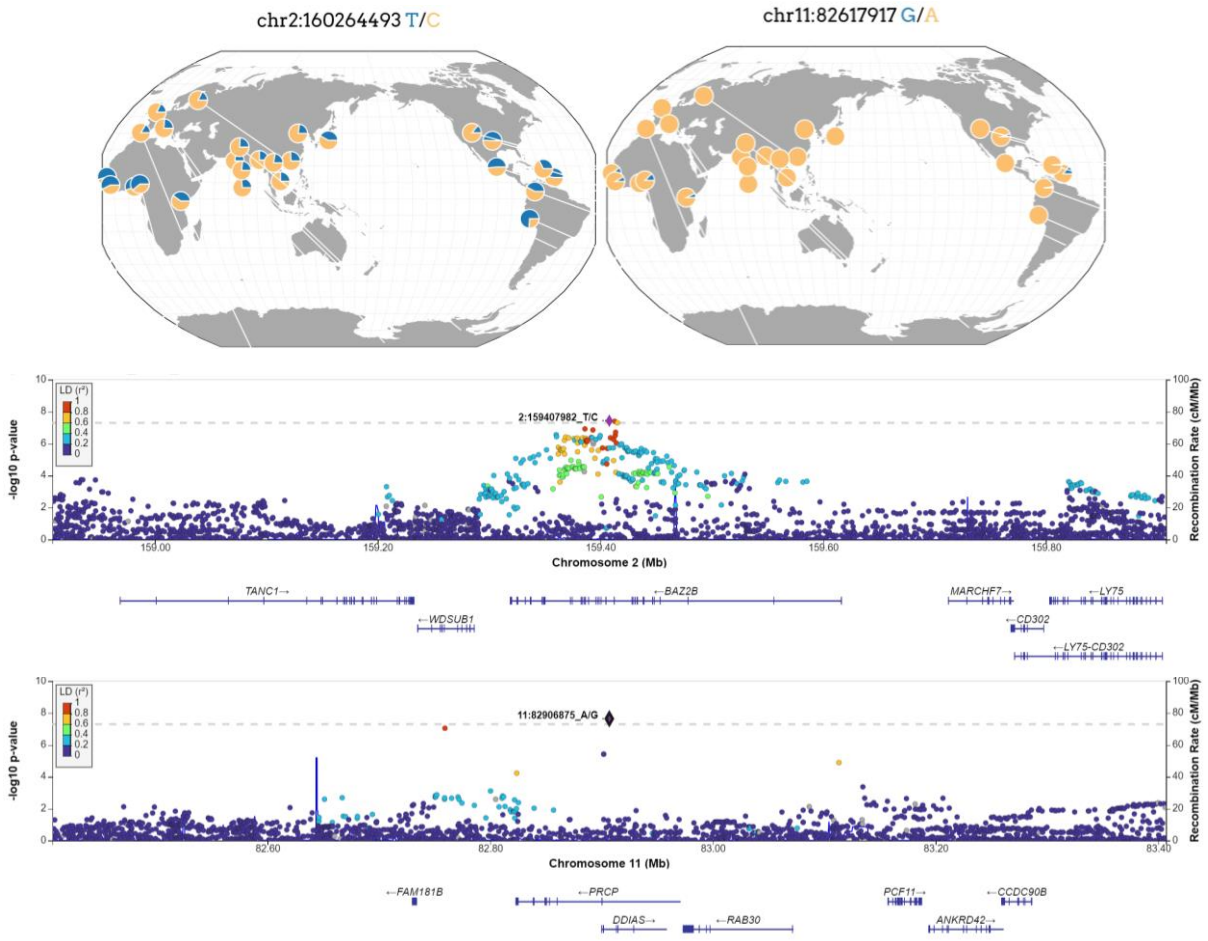


Figura 18. Distribución das frecuencias alélicas dos SNPs rs1003835 (chr2) e do rs77599934 (chr11) alélicas nas poboacións de 1000 Genomas e os gráficos rexionais correspondentes. As figuras das frecuencias alélicas foron creadas na web *The Geography of Genetic Variants* ou GGV (Marcus & Novembre, 2017). Os plots rexionais foron creados coa ferramenta LocusZoom (Boughton et al., 2021).

As rexións dos cromosomas 2 e 11 foron identificadas aquí por primeira vez (**Figura 18**). O rs13003835, localizado no xen *BAZ2B*, está asociado a un maior risco de hospitalización por COVID-19 (OR=1.20, 95%IC=1.12-1.27). Na meta-análise trans-poboacional do consorcio HGI, obtivo un p-valor de 0.972. Esta variante é un eQTL de *LY75* na mucosa do esófago (NES=0.27 para o alelo C) e de *BAZ2B* en sangue (NES=0.10 para o alelo C). A nosa estratexia de mapeamento de xenes sinalou tamén a *PLA2R1*, *LY75*, *WDSUB1* e *CD302*.

A outra rexión de risco está liderada polo rs77599934 no xen *DDIAS*, unha variante rara intrónica que confire un maior risco de hospitalización (OR=2.27, 95%IC=1.70-3.04). Neste *locus*, FUMA priorizou tamén ao xen *PRCP*. A variante rs60606421, en desequilibrio de ligamento coa nosa variante líder (r2=0.776, alelo - de rs60606421 co alelo G de rs77599934), está asociada a unha expresión reducida de *DDIAS* no pulmón (NES=-0.490 para o alelo -).

O set crible ao 95% ao redor do rs13003835 no cromosoma 2 incluíu 76 variantes (**Figura 19, Anexo VIII).** Para o cromosoma 11, o algoritmo non converxeu.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

A maiores, observamos un sinal no cromosoma 15 que non acadou significación a nivel GWAS, liderado pola variante rs2601183 (OR=1.20, 95%IC=1.12-1.29, p=6.11x10^{-08} para o alelo G). Este SNP sitúase entre os xenes *ZNF774* e *IQGAP1*, e é un eQTL asociado a unha maior expresión de *ZNF774* no pulmón (NES=0.27 para o alelo G).

A meta-análise tamén sinalou dúas rexións significativas xa coñecidas, *LZTFL1* e *FOXP4*. O SNP rs35731912 (*LZTFL1*) xa se asociou á gravidade da COVID-19 en poboacións europeas. En canto ao rs2477820, é a primeira vez que se identifica este SNP na rexión de *FOXP4*. Porén, está en desequilibrio de ligamento moderado (r2=0.295) co rs2496644, que si foi relacionado coa hospitalización por COVID-19 (Kousathanas et al., 2022).

O test a nivel de xen foi significativo para os xa coñecidos *LZTFL1*, *XCR1*, *FYCO1*, *CCR9*, *IFNAR2* e para o xen *BAZ2B* (**Táboa 15**).

Táboa 15. Resultados da análise a nivel de xen de MAGMA na meta-análise de HGI AMR con SCOURGE Latino-América (hg37). Xen_ENS: nome do xen en ENSEMBL; Start: inicio do xen; Stop: final do xen; NSNPS: número de SNPs no xen; NPAR: número de parámetros empregados no modelo; ZSTAT: estatístico Z. Só se mostran os resultados significativos. Aplicouse unha corrección por Bonferroni para 18.632 xenes.

Xen_ENS	Chr	Start	Stop	NSNPS	NPAR	ZSTAT	P	Xen
ENSG00000163818	3	45864808	45957534	225	29	7.4578	4.4E-14	*LZTFL1*
ENSG00000173578	3	46058516	46069234	36	6	5.8012	3.3E-09	*XCR1*
ENSG00000163820	3	45959396	46037316	271	19	5.4864	2.1E-08	*FYCO1*
ENSG00000173585	3	45927996	45944667	42	10	5.2737	6.7E-08	*CCR9*
ENSG00000123636	2	160175490	160473203	1015	47	5.1862	1.1E-07	*BAZ2B*
ENSG00000159110	21	34602206	34637980	127	14	4.6758	1.5E-06	*IFNAR2*

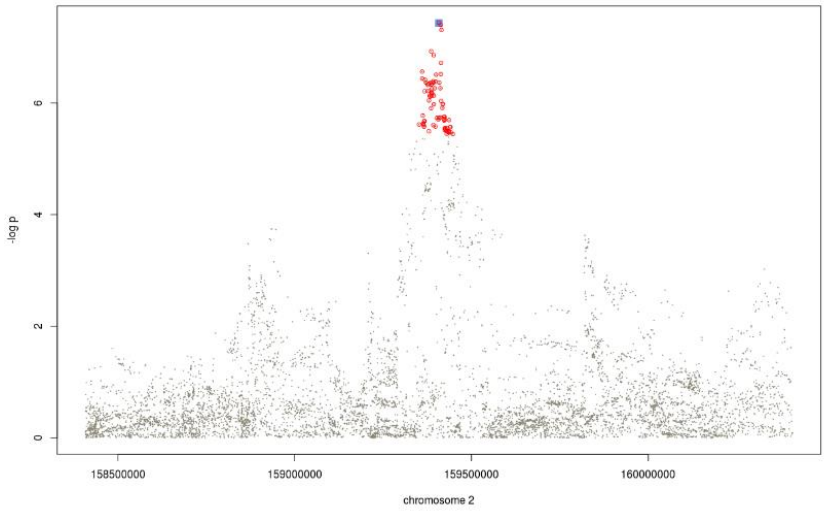


Figura 19. Gráfico rexional para a rexión do cromosoma 2 priorizada por mapeamento fino baiesiano. En vermello, as variantes incluídas no set crible con confianza do 95%. En azul, a variante líder.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5.1.2.1 TWAS para hospitalización por COVID-19 en poboación AMR

Para o TWAS empregamos datos de GTEx v8 e, dado que os datos de GTEx están obtidos maioritariamente de poboación europea, incrementouse a diversidade na análise empregando tamén datos de sangue derivados da cohorte GALA II-SAGE, que comprende americanos-africanos, Puertoriqueños e Mexicanos-Americanos (Kachuri, Mak, et al., 2023). Seleccionáronse os resultados da cohorte conxunta (*pooled*), e das cohortes AFRHp5 (ancestralidade xenética africana >50%) e IAMHp5 (ancestralidade xenética nativo-americana >50%).

Nas análises de tecidos individuais con datos de GTEx, *ATP5O* e *IFNAR2* acadaron unha asociación significativa na mucosa do esófago; *IFNAR2* e *SLC25A37* en linfocitos e na mucosa do esófago; *CCR9* en tecido sanguíneo, e *CXCR6* e *ATP5O* en pulmón (**Táboa 16**). Ao combinar os tecidos nunha análise multi-tecido, sumáronse cinco xenes non identificados con anterioridade en ningún GWAS: *SLC25A37*, *SMARCC1*, *CAMP*, *TYW3*, e *S100A12* (**Figura 20** e **Anexo IX**). Empregando os datos de GALA II-SAGE non se detectou ningún par xen-tecido significativo. O TWAS multi-tecido non se puido realizar posto que só se dispoñían de datos para sangue.

Táboa 16. Resultados da asociación TWAS nos tecidos elixidos. Só se mostran os SNPs significativos mediante corrección por FDR.

Tecido	Xen	Chr	Zscore	Var(g)	P-valor
Linfocitos					
	IFNAR2	21	-4.93	0.038	$1.12x10^{-06}$
	SLC25A37	8	4.88	0.018	$1.26x10^{-06}$
Mucosa do esófago					
	IFNAR2	21	-5.11	0.065	$7.43x10^{-07}$
	ATP5O	21	4.68	0.021	$2.84x10^{-06}$
Pulmón					
	CXCR6	3	-5.82	0.013	$2.07x10^{-08}$
	ATP5O	21	5.02	0.070	$8.84x10^{-07}$
Sangue					
	CCR9	3	8.26	0.003	$1.01x10^{-15}$

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

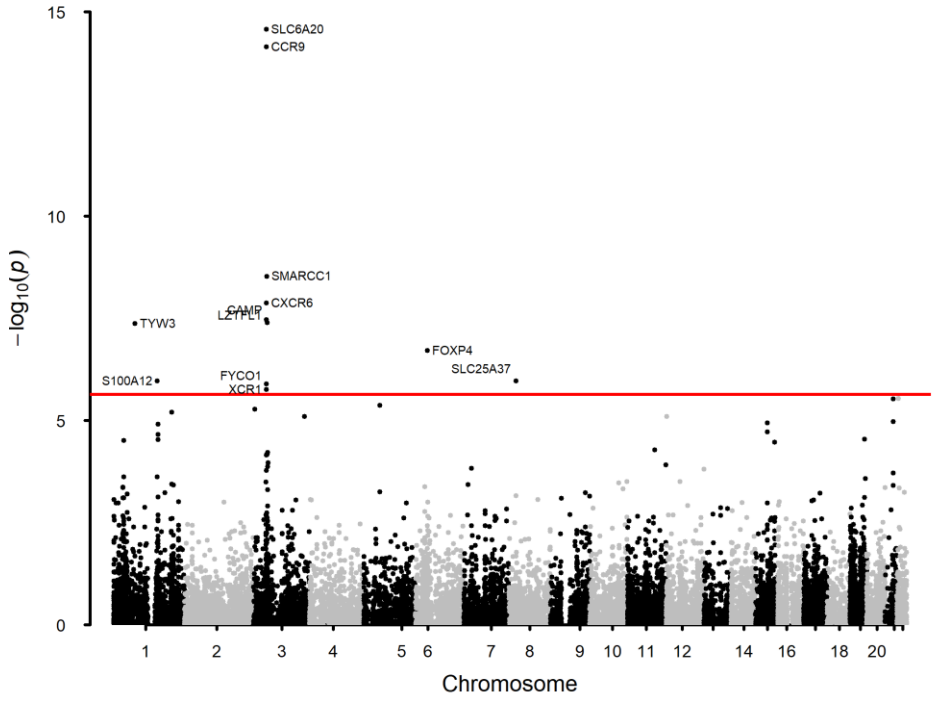


Figura 20. Gráfico tipo Manhattan Plot para o TWAS multi-tecido en poboación AMR.

5.1.3 Exploración da arquitectura xenética da hospitalización por COVID-19 en poboación AMR

Nin rs13003835 (*BAZ2B*) nin rs77599934 (*DDIAS*) obtiveron unha asociación significativa na meta-análise multi-poboación de HGI B2 nin en ningunha das meta-análises específicas de poboación, así que investigamos a distribución das frecuencias alélicas e comparamos os efectos.

Segundo gnomAD v3.1.2, o alelo T de rs13003835 (*BAZ2B*) ten unha frecuencia alélica do 43% en grupos americanos mixtos, mentres que a FA é menor en poboacións europeas (16%) e en global (29%). Os datos de ancestralidade local presentados por gnomAD (v3.1.2) indican que dentro da compoñente Nativo-Americana, o alelo de risco T é o alelo máis frecuente, mentres que é o alelo menor nas compoñentes Africana e Europea. Estas diferencias poden estar detrás da asociación descuberta en poboación AMR, xa que requiriría menor potencia estatística para ser detectada. Porén, ao comparar os tamaños do efecto entre as diferentes poboacións (recuperados das meta-análises específicas de poboación de HGI B2 versión 7), observamos que a dirección do efecto é distinta entre as poboacións AMR e SAS e as poboacións EUR, AFR e EAS, polo que existe unha heteroxeneidade moi grande entre elas (**Figura 21)**. Esta heteroxeneidade pode deberse a diferencias no recrutamento dos casos e nas definicións do fenotipo, ou tamén a diferencias de efecto entre as poboacións, como ocorre coas variantes do xen *FOXP4* (COVID-19 Host Genetics Initiative, 2023). As variantes con p-valores máis baixos dentro dunha ventá de 50kb centrada neste SNP para as diferentes poboacións foron: rs559179177 ($p=7.2x10^{-04}$) en EUR; rs2161079 (p=0.016) en SAS; rs570927133 (p=0.05) en

EAS, e rs10180735 ($p=3.5x10^{-03}$) en AFR. O SNP rs559179177, ademais, obtivo un p-valor de asociación na análise ALL de $p= 5.9x10^{-04}$. Curiosamente, este SNP ten unha FA máxima do 0.33% (segundo gnomAD v3.1.2), correspondente a poboacións europeas. Cabe a posibilidade, entón, de que esta heteroxeneidade poida estar causada por diferencias no patrón de desequilibrio de ligamento entre a variante observada e a causal ou por interaccións tipo xenotipo-xenotipo ou xenotipo-ambiente.

A variante do xen *DDIAS*, rs77599934, presenta unha FA de 1.1% para o alelo G nos controis non hospitalizados, o que é consistente cos datos de gnomAD, onde a frecuencia para os AMR mixtos é do 1%. A FA noutros grupos poboacionais é moi baixa, chegando incluso a ser monomórfica (0% en europeos de Finlandia, 0.025% en europeos non finlandeses, 0% en poboacións EAS e 0.042% en SAS), sendo maior en África (9%). A pesar de que é unha variante máis frecuente en poboación africana, o tamaño do efecto é significativo unicamente en poboación AMR (**Figura 21**). Examinando a AL, vemos que o alelo G está presente nos segmentos de ancestralidade africana nun 1.1%, e non aparece na compoñente nativo-americana. Tras explorar a distribución das frecuencias alélicas nos subgrupos de 1KGP en LDlink (https://ldlink.nih.gov/), vimos que este SNP é monomórfico en PEL (Perú) e MXL (México), acadando a FA máis alta en CLM (Colombia). Isto podería indicar que en Latino-América, as poboacións con ancestralidade africana teñan un risco maior, en global, por ser máis frecuente. Dada a súa baixa FA, o SNP rs77599934 non foi analizado nas meta-análises multi-poboación de HGI, impedindo unha comparación directa dos efectos. Así, exploramos a rexión ao redor da nosa variante líder (nunha ventá de 50kb) nos resultados da meta-análise de HGI B2-ALL, e recuperamos a variante que menor p-valor presentaba. Porén, este SNP, rs75684040 (OR=1.07, 95%CI=1.03-1.12, $p=1.84x10^{-3}$), está pouco relacionada co rs77599934 segundo os cálculos de desequilibrio de ligamento empregando datos de 1KGP AMR (r2=0.11). Na meta-análise de AFR, rs138860115 (a ~20k pares de bases de rs77599934) obtivo un p-valor de $8.3x10^{-03}$, pero este SNP ten unha FA do 0.11% en poboación AMR, e non foi posible meta-analizala porque non estaba presente na nosa cohorte.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

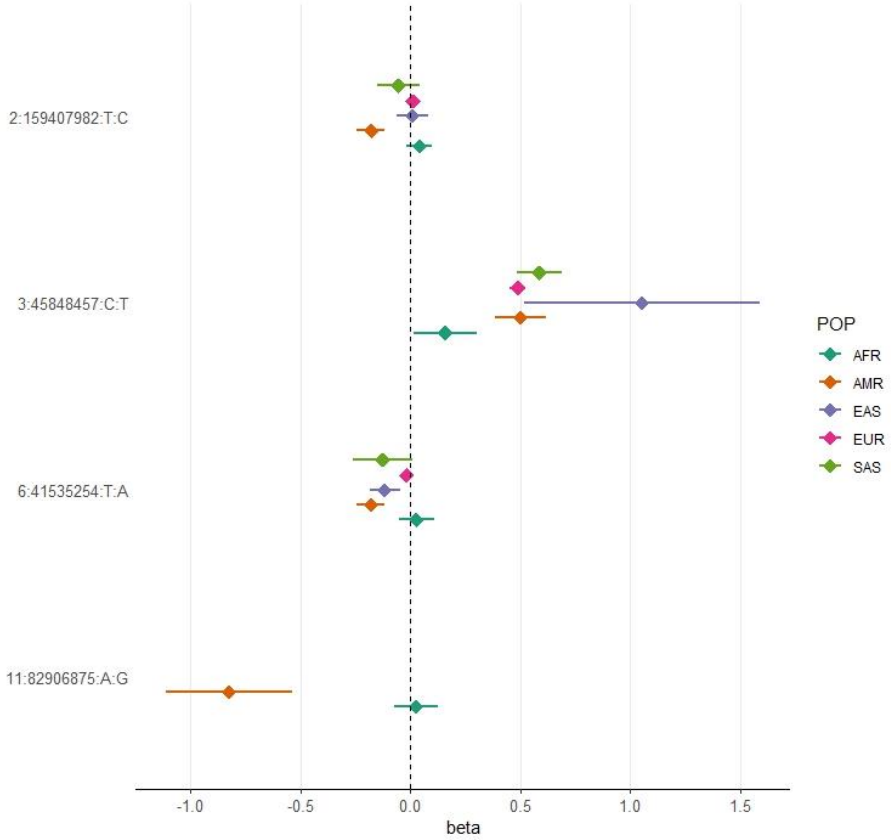


Figura 21. Gráfico tipo bosque mostrando os tamaños de efecto e os intervalos de confianza correspondentes para as variantes líder identificadas na meta-análise de HGI AMR con SCOURGE Latino-América nas poboacións AFR, AMR, EAS, EUR e SAS. Os valores beta e os ICs foron obtidos das análises específicas de poboación B2 da versión 7 da publicación de HGI, excepto para AMR, para o cal se representa o beta e o IC da meta-análise HGI_{AMR}-SCOURGE.

5.1.4 Meta-análises trans-poboacionais e replicación de sinais

Co obxectivo de avaliar os novas sinais e comprobar se replicaban nun contexto multi-poboacional, levamos a cabo dúas meta-análises adicionais combinando diferentes poboacións. Isto levou ao descubrimento de novos sinais asociadas á hospitalización por COVID-19. A definición dos *loci* foi a través dos seguintes criterios: 1) p-valor<$5x10^{-08}$ na meta-análise e p-valor >$5x10^{-08}$ na análise HGI B2 ALL ou HGI B2 AMR, EUR ou AFR; 2) p-valor do test Q de heteroxeneidade de Cochran <$0.05/N_{loci}$, onde N_{loci} é o número de variantes que superan o límite GWAS; e 3) o xen máis próximo non foi descrito na actualización de HGI v7.

En primeiro lugar, combinamos o estudo de SCOURGE Latino-América cos resultados do HGI B2 ALL. Referirémonos a esta análise como $SC\text{-}HGI_{ALL}$. Dos 40 *loci* asociados coa hospitalización por COVID-19 na versión 7 do consorcio HGI, esta meta-análise replicou 39 deles (p <$1.25x10^{-03}$) e aumentou a forza de asociación en 29 (**Anexo X**). Pola súa parte, na meta-análise de $SC\text{-}HGI_{AMR}$ só replicamos 7 asociacións, mais probablemente se deba ao tamaño de mostra (**Anexo X**).

A variante rs13003835 (*BAZ2B*) non se replicou no $SC\text{-}HGI_{ALL}$ (OR=1.00, 95%IC=0.98-1.03, p=0.644), pero si replicamos dous sinais do último estudo de GenOMICC (Pairo-

Castineira et al., 2023), que non estaban asociados na versión 7 do HGI. Unha delas atópase no xen *CASC20*, rs2876034 (OR=0.95, 95%CI=0.93-0.97, p=2.83x10^{-08}), e está en desequilibrio de ligamento coa variante líder de GenOMICC (rs2326788, r2=0.92), asociada a enfermidade crítica. A outra atópase cerca do xen *ZBTB7A*, liderada pola variante rs66833742 (OR=0.94, 95%CI=0.92-0.96, p=2.50x10^{-08}). Este SNP está asociado a unha maior expresión do xen *ZBTB7A* en sangue e no tecido da mucosa do esófago, así como a hospitalización por COVID-19 (Pairo-Castineira et al., 2023).

Nunha segunda fase, exploramos a combinación das tres fontes de ancestralidade das poboacións latinoamericanas: AMR, EUR e AFR, excluíndo EAS e SAS. Referímonos a esta meta-análise como SC-HGI$_{3POP}$. A asociación no *locus* rs13003835 tampouco se replicou (*BAZ2B*, OR=1.01, 95%IC=0.98-1.03, p=0.605) e o rs77599934 no xen *DDIAS* non puido ser analizado. O SNP rs66833742 (*ZBTB7A*) volveu a acadar significación GWAS (OR=0.94, 95%CI=0.92-0.96, p=1.89x10^{-08}). Nesta meta-análise atopamos unha nova rexión de risco no xen *CREBBP* a través da variante rs76564172 (OR=1.31, 95%IC=1.25-1.38, p=9.64x10^{-09}). Ao ser unha variante non detectada ata o momento, foi sometida a análises de mapeamento fino-baiesiano. O set crible incluíu un total de 8 variantes.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5.2 DISCUSIÓN

A pesar de que a base xenética de risco para a COVID-19 descuberta ata o momento considerouse replicada en todas as poboacións a través de meta-análises multi-poboación, o número de individuos de ascendencia europea nestes estudos é significativamente maior. A dominancia de cohortes europeas introduce un nesgo á hora de descubrir os determinantes xenéticos dos caracteres, xa ben porque enmascaran sinais específicos de poboación ou menos frecuentes en poboacións europeas, ou porque fallan na identificación das variantes líder transferibles a todas. Concretamente, para o fenotipo de hospitalización, dos 44.986 casos incluídos, 32.519 son de ascendencia europea, e só 3.077 son de ascendencia hispana. O consorcio HGI publicou un estudo específico en poboación AMR, mais o único sinal que acadaba significación GWAS era a rexión do cromosoma 3. Ao incrementar o tamaño de mostra incluíndo a cohorte de SCOURGE, atopamos dous sinais novos que non foran identificados nin en análises específicas de poboación nin na análise multi-poboación.

5.2.1 Os primeiros xenes identificados en poboación AMR: *DDIAS* e *BAZ2B*

A variante líder rs77599934 no xen *DDIAS* é unha variante rara (AF ~1% para o alelo G) con un efecto grande sobre a hospitalización por COVID-19 (OR=2.27). Esta variante é prácticamente monomórfica en todos os grupos continentais salvo nas poboacións africanas e latinoamericanas. A exploración das FAs para este SNP suxire que só confire risco a individuos con herdanza africana, e polo tanto, pode que non sexa detectable en individuos de Perú ou México. Cómpre aclarar que esta variante en *DDIAS* non ten necesariamente efectos específicos sobre individuos con tractos ancestrais africanos, senón que será máis probable observar o alelo de risco neles. A nivel individual, unha persoa dun país europeo pode ser portador do alelo, polo que esta asociación é importante alén das poboacións latinoamericanas.

O mapeamento fino baiesiano e a integración con datos de expresión non revelaron máis información sobre o xen no que se atopa esta variante ou sobre as súas consecuencias funcionais. Porén, *DDIAS*, coñecido como un supresor da apoptose inducida por dano, perfílase como un bo candidato a ser causal por ser un xen próximo (Backman et al., 2021). *DDIAS* (ou *C11orf82*) está relacionado con mecanismos de reparación do ADN e produce unha proteína anti-apoptótica. O bloqueo de *DDIAS* asociouse a unha deficiencia no proceso de recombinación homóloga (Brunette et al., 2019), que é un mecanismo de reparación das roturas de dobre cadea no ADN (DSB, *double-strand breaks*), o que suxire un control sobre estes procesos celulares. De feito, tamén foi asociado á acumulación de *foci* da proteína de unión a p53 (53BP1), un biomarcador dos DSB (Cortez et al., 1999). Estudos previos demostraron a implicación de *DDIAS* na inhibición da ruta de sinalización p38 MAPK/p53 e na inhibición do crecemento de células cancerosas de pulmón (Won et al., 2014). Na mesma liña, Im e colaboradores (2020, 2023) propuxeron a utilidade deste xen como biomarcador para cancro de pulmón, tras descubrir a súa interacción con STAT3 e a regulación positiva das interleucinas IL-6.

A acumulación de dano celular e a activación dos mecanismos de reparación do ADN están moi relacionados cos procesos de inflamación crónica e senescencia (Pezone et al., 2023). No

USC UNIVERSIDADE DE COMPOSTELA

contexto da COVID-19, onde se observa unha resposta inflamatoria descontrolada, *DDIAS* podería ter una implicación relevante. Por exemplo, observouse que o SARS-CoV-2 inducía a acumulación de DSB e que a proteína N do virus inducía unha resposta de reparación do ADN truncada (Gioia et al., 2023). Ademais, os autores do estudo atoparon que o virus activaba as rutas de MAPK/p38 e a de STAT1. A variante líder na rexión de risco de *DDIAS* está en desequilibrio de ligamento cun eQTL que reduce a expresión do xen no pulmón. Unha posible hipótese é que, considerando unha produción diminuída da proteína DDIAS no pulmón, os dous mecanismos de reparación de dano celular durante a infección sexan menos efectivos, inducindo acumulación de dano celular e a activación de rutas pro-inflamatorias. Tamén podería estar implicada a ruta a ruta p38 MAPK/p53, des-regulada en doentes infectados, dado o papel de *DDIAS* na súa inhibición.

Na rexión de risco do cromosoma 2, os datos de expresión apuntaron a máis dun xen. A variante líder rs13003835 está localizada no xen *BAZ2B*. *BAZ2B* codifica unha subunidade reguladora dos remodeladores de cromatina tipo ISWI (Y. Li et al., 2021), constituíndo os complexos BRF-1/BRF-5 xunto con *SMARCA1* e *SMARCA5*, respectivamente, que actúan facilitando o acceso ao ADN. Un estudo identificou a *BAZ-2* (ortólogo en *C. elegans* de *BAZ2B*/*BAZ2A*) como un regulador epixenético involucrado no envellecemento, mentres que o estudo dos homólogos de *BAZ-2* en rato apuntou ao xen *BAZ2B* (Yuan et al., 2020). Varias investigacións demostraron que o SARS-CoV-2 provoca cambios estrutura da cromatina (R. Wang et al., 2023; Wei et al., 2023) e na regulación epixenética (Kee et al., 2022; R. Wang et al., 2023), se ben aínda non se relacionaron os complexos tipo ISWI. É de interese destacar tamén o papel do *long non-coding BAZ2B* (*lnc-BAZ2B*), que promove a expresión de *BAZ2B*, na inflamación pulmonar a través da activación de macrófagos de tipo M2 (Xia et al., 2021). O segundo xen é *LY75*, o antíxeno de linfocito 75, tamén chamado *CD205*. A proteína que produce está involucrada en procesos inmunolóxicos a través da presentación de antíxenos en células dendríticas e endocitose (Mahnke et al., 2000), e está asociado con enfermidades inflamatorias.

Atopouse unha terceira rexión de risco, aínda que non con significación GWAS, no cromosoma 15, entre os xenes *IQGAP1* e *ZNF774*. *ZNF774* pertence á familia de proteínas codificadora de dedos de zinc ZNF, algunhas das cales xa foron relacionadas coa COVID-19 (Esposito et al., 2022). Por outro lado, unha análise bioinformática suxeriu que a ubiquitinación de algunhas proteínas, entre elas IQGAP1, podería ser un factor modulador da resposta inmune á infección por SARS-CoV-2 (H. Zhang et al., 2021).

5.2.2 Novos xenes identificados a través do TWAS

As análises de TWAS apuntaron a cinco novos xenes, se ben algúns deles xa se relacionaron previamente coa COVID-19. Un estudo da expresión xénica multi-tecido detectou unha expresión reducida dos xenes *CAMP* (péptido antimicrobiano catelicidina) e *S100A8/S100A9* (proteína de unión a calcio S100) en doentes con COVID-19 crítica (Gómez-Carballa et al., 2022). *CAMP* está relacionado cunha maior gravidade en infeccións virais (H. Zheng et al., 2021), pero é un xen que tamén se atopa na rexión clásica do cromosoma 3. Outro estudo detectou unha sobre-regulación de *SLC25A37* (transportador de soluto membro 37 da familia 25) en individuos con COVID-19 grave (Policard et al., 2021), e non é a primeira

proteína deste tipo asociada á COVID-19. Polo seu lado, *SMARCC1*, que foi previamente identificada como pro-viral para os virus tipo SARS e MERS (Wei et al., 2021), é unha subunidade do complexo remodelador de cromatina SWI/SNF que está involucrado na susceptibilidade do hóspede á infección (Wei et al., 2023). O estudo de Wei e colaboradores (2023) propuxo, ademais, unha intervención farmacolóxica sobre este complexo con inhibidores da ATPasa SMARCA4/2, polo que o xen *SMARCC1* podería constituír unha diana farmacolóxica.

5.2.3 Factores xenéticos da COVID-19 entre poboacións

Co fin de explorar máis a arquitectura xenética da COVID-19 en poboación AMR realizáronse dúas meta-análises multi-poboación incluíndo os resultados de SCOURGE Latino-América. Os dous sinais identificadas no estudo non foron significativas fóra da análise específica de poboación, destacando a importancia de complementar análises multi-poboación con análises específicas. Curiosamente, a variante líder localizada no xen *FOXP4* presentou heteroxeneidade nos efectos no GWAS do HGI, obtendo unha asociación e un efecto máis fortes en poboación SAS e EAS (Pathak et al., 2022). No noso estudo, esta rexión de risco tivo maior potencia estatística que outras como *OAS* ou *IFNAR2*. Aínda considerándoa unha variante compartida entre poboacións, probablemente necesitemos un maior número de individuos para detectala en poboacións europeas que noutras como a AMR.

As meta-análises multi-poboación apuntaron a tres *loci* que non acadaran significación GWAS na versión 7 de HGI B2 ALL: un sinal novo en *CREBBP*, e dous en *ZBTB7A* e *CASC20*. Estes dous últimos si foron reportados noutra meta-análise (Pairo-Castineira et al., 2023). *CREBBP* e *ZBTB7A* tiveron unha asociación máis forte cando se consideraron unicamente as poboacións de EUR, AFR e AMR. De acordo cun estudo recente, os niveis altos de *ZBTB7A* promoven un estado case homeostático entre os coronavirus e as células do hóspede, impedindo a morte celular a través da regulación de rutas relacionadas co estrés oxidativo e provocando unha infección persistente (X. Zhu et al., 2022). Ademais, este xen está involucrado en varias rutas de sinalización, como a diferenciación de células B e T (S. Gupta et al., 2020). Por outra banda, o xen *CREBBP* codifica a proteína de unión a CREB (CBP), que participa na activación transcripcional e regula positivamente a resposta dos IFN tipo I a partir da fosforilación de IRF-3 inducida por virus (Yoneyama et al., 1998). Ademais, a interacción CREBP/CBP está implicada na infección por SARS-CoV-2 a través da ruta cAMP/PKA (Q. Yang et al., 2023). De feito, as células co xen *CREBBP* suprimido presentan unha replicación reducida das variantes Delta e Omicron (Q. Yang et al., 2023).

Se ben un GWAS nunha cohorte de Brasil obtivo un sinal no xen *DSTYK* para o fenotipo de hospitalización (A. C. Pereira et al., 2022), este non foi replicado na nosa análise AMR, o que podería explicarse polo feito de que o haplotipo de risco derivaba de poboacións europeas, e no noso estudo optamos por reducir o peso desta compoñente da GIA. De todas formas, parte da cohorte do estudo BRACOVID foi incluída na nosa meta-análise. Outros sinais asociados á COVID-19 teñen diferentes distribucións alélicas entre poboacións. Por exemplo, un GWAS en poboación xaponesa identificou unha variante en *DOCK2* de baixa frecuencia en poboación europea, e o seu efecto na gravidade da enfermidade confirmouse a través da inhibición do xen

en ratos (Namkoong et al., 2022). En poboación polinesia e Inuit detectáronse unhas variantes de perda de sentido (*nonsense*) nos xenes *IFNAR1* e *IFNAR2*, respectivamente, que en homocigose provocan unha resposta inexistente aos IFNs tipo I nos fibroblastos (Bastard et al., 2022; C. J. A. Duncan et al., 2022). Estas variantes incrementan a susceptibilidade aos virus, entre eles o SARS-CoV-2, mais tamén ás vacinas que empregan virus atenuados. Por outro lado, os haplotipos Neanderthais *OAS* e *LZTFL1* presentan diferencias poboacionais. O haplotipo máis común derivado dos Neanderthais no clúster de xenes *OAS* está practicamente ausente en poboacións africanas e en poboacións AMR con GIA africana, pero alcanza frecuencias ao redor do 25% en poboacións europeas e asiáticas, suxerindo unha posible selección positiva nestas poboacións (Zeberg & Pääbo, 2021). A rexión do cromosoma 3 tamén provén do xenoma Neanderthal, mais neste caso atópase presente en frecuencias máis altas no sur de Asia, especialmente en Bangladesh (onde o 63% ten unha copia do mesmo), e en frecuencias máis reducidas en europeos e americanos mixtos, sendo case inexistente no este de Asia (Zeberg & Pääbo, 2020).

5.2.4 Limitacións do estudo

Este estudo presenta varias limitacións, principalmente relativas ao recrutamento e composición da mostra. O tamaño de mostra da cohorte Latinoamericana de SCOURGE é pequeno, e o GWAS por si só non ten potencia estatística para a detección de ningunha variante con significación GWAS. Ademais, engloba individuos de varios países e, a pesar de que se intentou controlar nas análises, as diferencias nos recrutamentos e criterios de hospitalización a través dos mesmos pode que reduzan a capacidade de detectar asociacións significativas. En relación a isto e á composición xenética da mostra, as rexións de risco aquí identificadas pode que non se repliquen noutras cohortes con falta dalgunha das tres compoñentes de mestura (GIA). Tampouco foi posible controlar os factores sociodemográficos que puideron afectar á propagación da COVID-19 e ás taxas de hospitalización, aínda que as compoñentes principais xenéticas tamén capturan parte desta variabilidade. Por último, cómpre destacar a falta dunha cohorte de replicación. Na nosa meta-análise empregamos todas as cohortes dispoñibles de COVID-19 en poboación mixta americana, polo que non dispuxemos de estudos adicionais cos que replicar ou validar os resultados. Estes problemas poden resolverse no futuro, tanto para este fenotipo como para outros, incluíndo máis estudos con individuos de Latino-América nas análises. O aumento da diversidade nos estudos GWAS debe darse por dúas vías: pola inclusión de individuos non europeos nos recrutamentos e por apoiar e financiar investigacións noutros países.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

6 ÍNDICES DE RISCO POLIXÉNICO PARA COVID-19

6.1 RESULTADOS

6.1.1 Selección dos PRS

Deriváronse dous tipos de PRS para hospitalización por COVID-19 en función do GWAS de descubrimento, partindo de datos de europeos e partindo dos resultados (estatísticos resumo) da meta-análise GWAS de EUR e da meta-análise GWAS ALL (multi-poboación). Comparáronse dúas aproximacións metodolóxicas para a súa construción (**Táboa 17**). En total construíronse e testáronse 65 PRS europeos e 65 PRS multi-poboación para cada cohorte diana.

Táboa 17. Descrición dos PRS en función do método, datos de descubrimento e cohorte diana. Para cada cohorte de descubrimento co procedemento C+T testáronse 64 combinacións de parámetros. 5 POPS é referido ás cohortes AMR, EUR, SAS, AFR e EAS meta-analizadas na análise B2 ALL.

Método	Datos de descubrimento	Cohorte diana	Nome do PRS
PRS europeos			
C+T	HGI B2 EUR (sen SCOURGE)	SCOURGE España	CT-EUR$_{Esp}$
C+T	HGI B2 EUR (sen SCOURGE)	SCOURGE Latino-América	CT-EUR$_{LatAm}$
PRS-CS	HGI B2 EUR (sen SCOURGE)	SCOURGE España	CS-EUR$_{Esp}$
PRS-CS	HGI B2 EUR (sen SCOURGE)	SCOURGE Latino-América	CS-EUR$_{LatAm}$
PRS multi-poboación			
C+T	HGI B2 ALL (sen SCOURGE)	SCOURGE España	CT-ALL$_{Esp}$
C+T	HGI B2 ALL (sen SCOURGE)	SCOURGE Latino-América	CT-ALL$_{LatAm}$
PRS-CSx meta	HGI B2 5 POPS (sen SCOURGE)	SCOURGE España	CSx-META$_{Esp}$
PRS-CSx meta	HGI B2 5 POPS (sen SCOURGE)	SCOURGE Latino-América	CSx-META$_{LatAm}$

O primeiro conxunto de PRS calculouse seguindo o método *clumping* + *thresholding* (C+T). Tal e como se describe no apartado 3.13, no algoritmo C+T variamos os parámetros de r^2 (0.2, 0.4, 0.6, 0.8), ventá de *clumping* (100, 250, 500, 1000) e os p-valores límite (0.1, 0.01, 1e-05, 1e-08). Para cada combinación descubrimento/diana seleccionouse o PRS con maior incremento no pseudo-R^2 (ΔR^2), é dicir, os PRS europeo e multi-poboación con mellor capacidade preditiva nas cohortes española e latinoamericana. Os parámetros refináronse na cohorte española en relación á predición do PRS da variable CHOSP, que compara individuos hospitalizados con individuos infectados non hospitalizados, xa que na cohorte latinoamericana non hai dispoñemos de controis poboacionais.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

As **Figura 22** e **Figura 23** representan os resultados dos PRS multi-poboacionais e europeos, respectivamente, en relación ao seu pseudo-R^2 parcial. Como se pode observar en ditos gráficos, os PRS que combinan información de varias poboacións son mellores preditores que os PRS europeos para ambas cohortes. Sorprendentemente, a cohorte latinoamericana obtivo mellores rendementos que a cohorte española cos PRS multi-poboación e uns resultados semellantes cos PRS europeos.

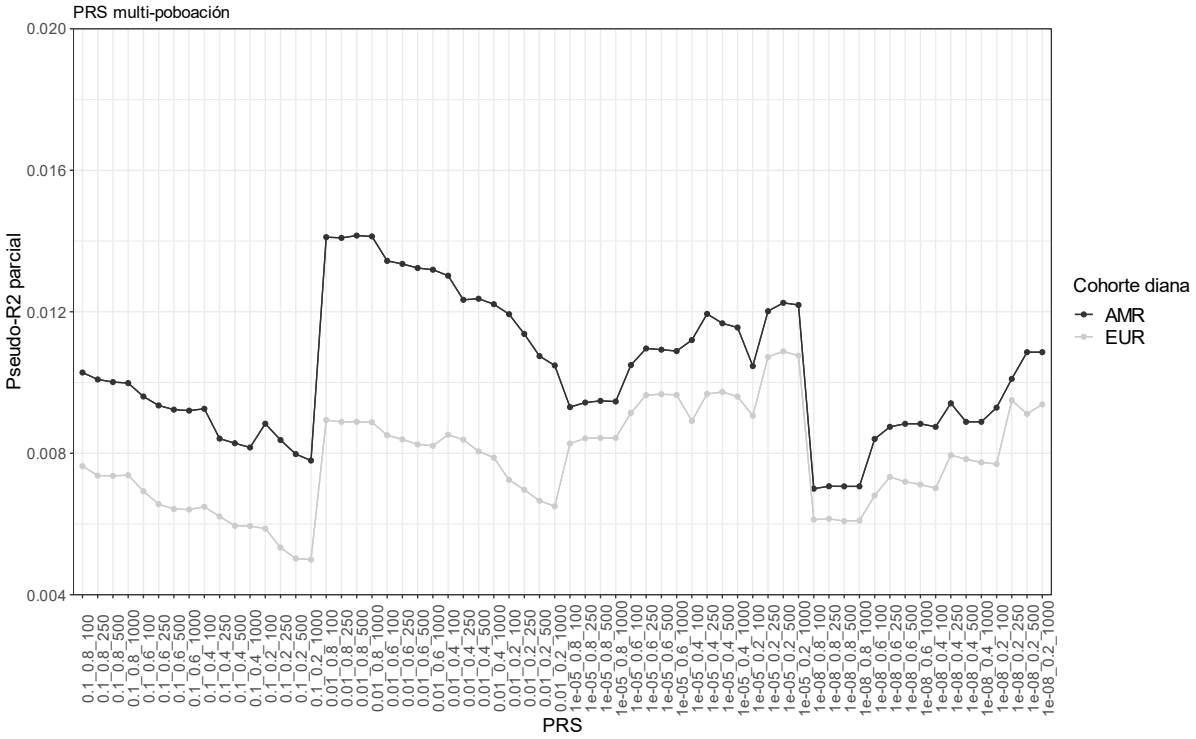


Figura 22. Representación gráfica do incremento no R2 parcial (y) asociado aos PRS C+T multi-poboación en función dos hiperparámetros (x). O eixo X está ordenado en función do p-valor, r^2 entre SNPs e tamaño da ventá. Como variable dependente usouse "CHOSP", hospitalización dentro de casos COVID-19+.

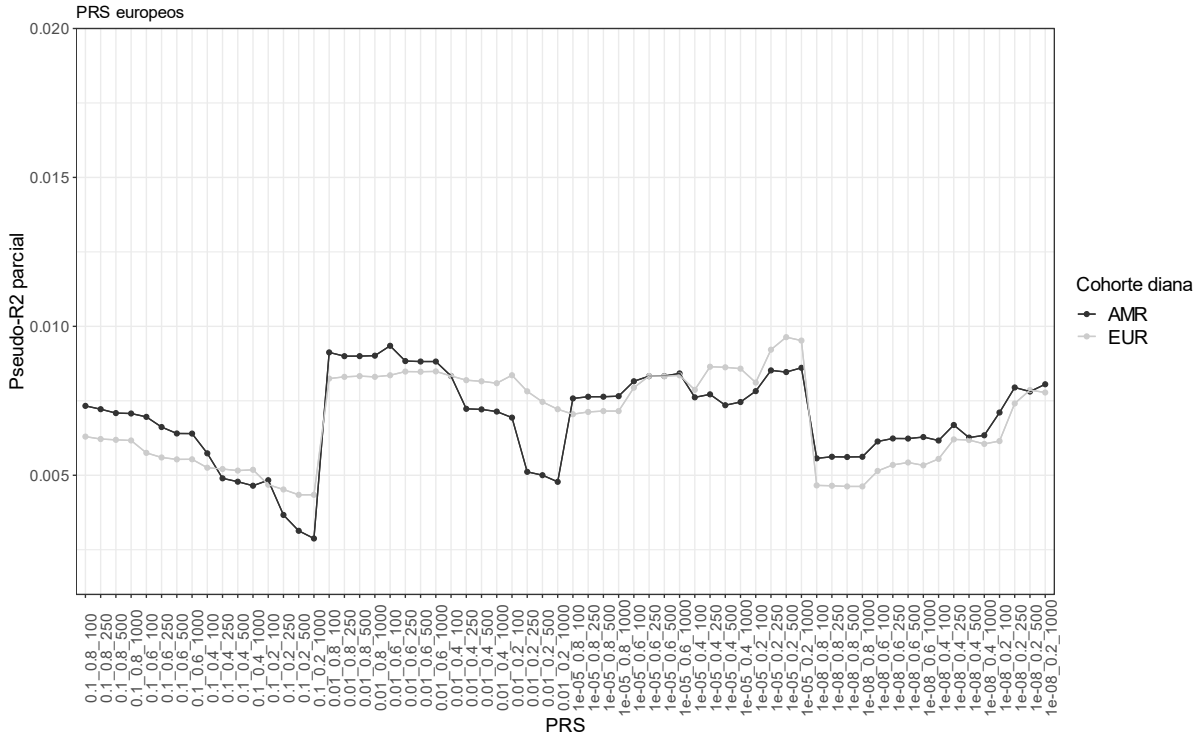


Figura 23. Representación gráfica do incremento no R2 parcial (y) asociado aos PRS C+T europeos en función dos hiperparámetros (x). O eixo X está ordenado en función do p-valor, r^2 entre SNPs e tamaño da ventá. Como variable dependente usouse "CHOSP", hospitalización dentro de casos COVID-19+.

A modo de análise de sensibilidade, realizouse o mesmo estudo cos PRS C+T pero testados sobre a variable de gravidade crítica dentro de casos (CSEV4). Esta variable está definida por criterios clínicos máis estritos, reducindo as diferencias que poidan existir na definición da variable dependente, se ben hai que ter en conta que os datos de descubrimento dos PRS estudaban a variable de hospitalización. Os resultados para esta análise nos PRS multi-poboación e europeos están representados na **Figura 24**. Os PRS multi-poboación mostran comportamentos similares entre ambas cohortes e con valores máximos superiores aos PRS europeos. Estes últimos, tal e como se espera cando a cohorte de descubrimento e a diana son da mesma poboación, son mellores preditores na cohorte española que na latinoamericana.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

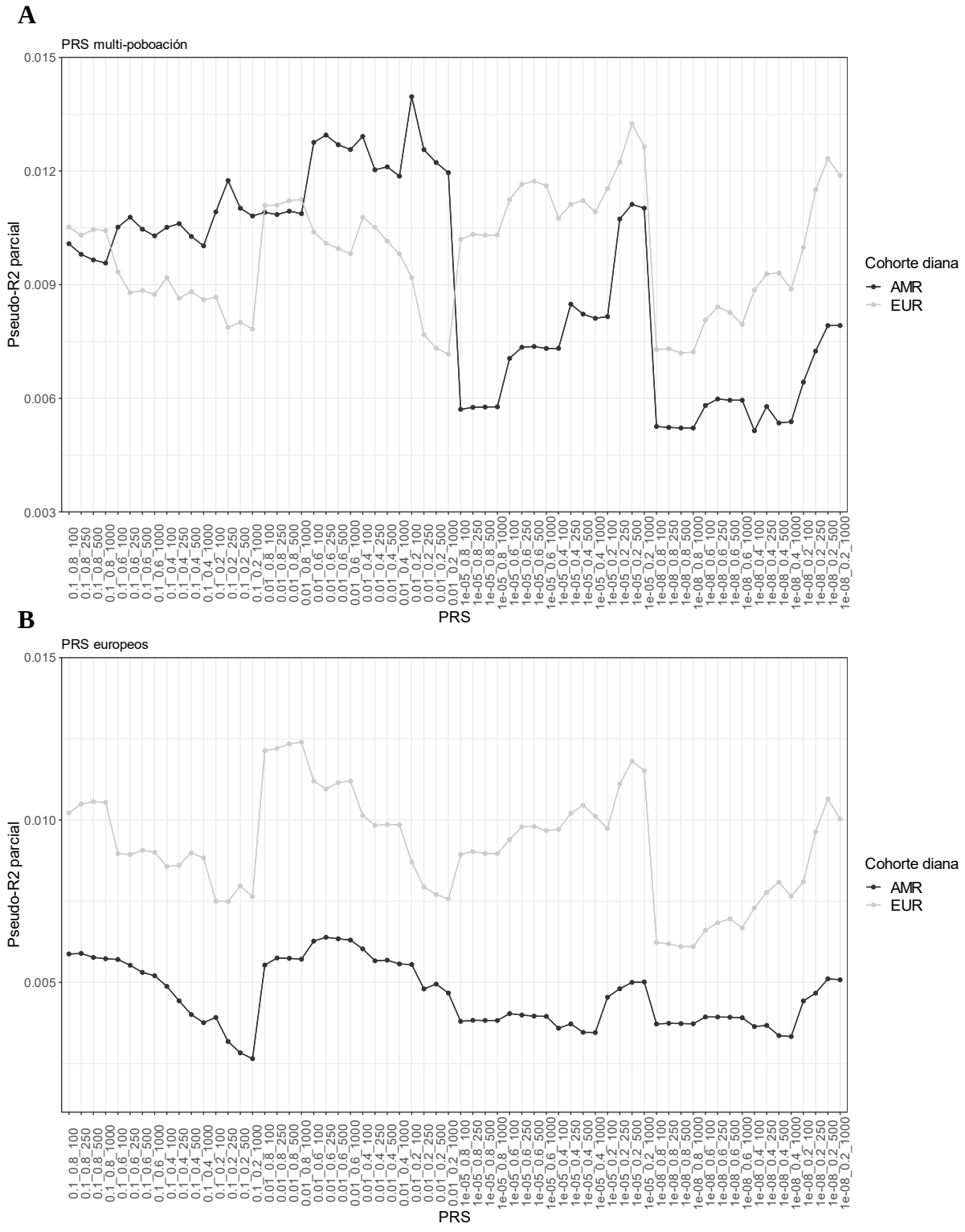


Figura 24. Representación gráfica do incremento no R2 parcial (y) asociado aos PRS C+T multi-poboación (A) e europeos (B) en función dos hiperparámetros (x) para a variable CSEV4. O eixo X está ordenado en función do p-valor, r^2 entre SNPs e tamaño da ventana. A variable CSEV4 define os controis como individuos COVID-19+ sen enfermidade crítica e os casos como casos COVID-19+ con enfermidade crítica.

O PRS C+T con mellor rendemento na cohorte española para CHOSP foi o construído con SNPs cuxo p-valor $\leq 1x10^{-05}$, ventá de 500 e r^2 de 0.2 (ΔR^2=0.011, 95% IC [0.006-0.015]). Pola

contra, para a cohorte latinoamericana foi mellor un PRS con máis SNPs pero cuxas variantes abarcaran rexións máis pequenas, con p-valor ≤0.01, ventá de 500 e r^2 de 0.8 (ΔR^2=0.014, [0.006-0.02]). Tal e como se comentou ao inicio, os PRS multi-poboación obteñen resultados superiores de forma consistente.

O segundo conxunto de índices foi o construído con PRS-CS e PRS-CSx. Este método estima os efectos das variantes independentemente da cohorte diana (emprega un panel de referencia externo), polo que os PRS europeo e multi-poboación foron os mesmos para ambas cohortes diana, salvo diferencias no número de SNPs dispoñible en cada unha. Os PRS baiesianos mostraron valores de pseudo-R^2 parciais superiores na cohorte latinoamericana que na cohorte española, ao igual que ocorreu cos C+T. Como se observa na **Táboa 18**, o PRS baiesiano multi-poboación foi o que obtivo mellores resultados na predición entre todos os PRS construídos. En xeral, a estimación do ΔR^2 non foi estable entre as remostras e os intervalos de confianza son relativamente amplos, aínda que este efecto pode deberse a que os incrementos no pseudo-R^2 se moven en valores moi baixos.

A modo de comparación entre cohortes, tamén se estudou o comportamento dos PRS baiesianos en relación á variable CSEV4. O PRS multi-poboación mostrou rendementos similares entre ambas cohortes (ΔR^2_{Esp}=0.017 [0.09-0.023], ΔR^2_{LatAm}=0.017 [0.006-0.025]). En cambio, o PRS europeo obtivo un valor máis alto nos individuos españois (ΔR^2_{Esp}=0.014 [0.007-0.020], ΔR^2_{LatAm}=0.010 [0.001-0.015]). En conclusión, o PRS multi-poboación CSx-META foi o seleccionado en ambas cohortes para as análises posteriores.

Táboa 18. Incremento no pseudo-R^2 de cada PRS xunto cos seus intervalos de confianza ao 95% calculados con bootstrap.

PRS	Tipo	ΔR^2 (95% IC)
Cohorte española		
$CT\text{-}ALL_{Esp}$	Multi-poboación	0.011 (0.006-0.015)
$CT\text{-}EUR_{Esp}$	Europeo	0.010 (0.005-0.013)
$CS\text{-}EUR_{Esp}$	Europeo	0.010 (0.005-0.013)
$CSx\text{-}META_{Esp}$	Multi-poboación	0.014 (0.009-0.018)
Cohorte latinoamericana		
$CT\text{-}ALL_{LatAm}$	Multi-poboación	0.014 (0.006-0.020)
$CT\text{-}EUR_{LatAm}$	Europeo	0.009 (0.003-0.014)
$CS\text{-}EUR_{LatAm}$	Europeo	0.009 (0.003-0.014)
$CSx\text{-}META_{LatAm}$	Multi-poboación	0.018 (0.009-0.024)

Co PRS final (de aquí en diante traballaremos con esta combinación de método-datos de descubrimento), axustouse un modelo para hospitalización por cada set de datos incluíndo as covariables básicas. A OR para cada desviación típica (s.d.) na cohorte española foi de 1.38 (95%IC=1.30-1.45), e na cohorte latinoamericana de 1.64 (1.43-1.87). Por último, como análise de sensibilidade nos datos de Latino-América, estimouse a OR dentro dos cuantís 30 e 70 para NAM e AFR. A OR asociada ao PRS para os individuos con menor proporción de ancestralidade NAM (≤0.23, OR=1.61 [1.25-2.08]) foi similar á do grupo con menor proporción (≥0.5, OR=1.70 [1.34-2.15]). O mesmo ocorreu ao estratificar por AFR (≤0.035,

OR=1.59 [1.26-2.00]; ≥0.17, OR=1.70 [1.33-2.16]), polo que non se obtivo evidencia de que o PRS presentara diferencias marcadas na predición do risco en función das compoñentes ancestrais.

6.1.2 Construción dun PRS único a partir do PRS óptimo

Tal e como se describiu na sección 3.13.5, combináronse as dúas mostras e computouse un PRS único e axustado para a ancestralidade, cun total de 12.805 individuos analizados (5244 controis e 7561 casos). O ΔR^2 do modelo combinado foi de 0.016 (95%IC=0.12-0.20). A OR asociada ao PRS_{ax} é de 1.42 (1.35-1.49) tras axustar polas covariables de sexo, idade e país de recrutamento. Transformando o ΔR^2 á *liability scale*, estimamos que o PRS explica un 1.99% da variabilidade do fenotipo asumindo unha prevalencia (de hospitalización por casos COVID-19) do 15% e un 1.4% asumindo unha prevalencia de hospitalización o 5%. Tras o proceso de validación interna, o AUC corrixido foi de 0.86 (optimismo=0.0004), o erro medio de predición foi de 0.009 e os resultados da calibración son bos (ver **Figura 25**). O PRS testouse de forma binaria para comparar o 20% da distribución cun risco xenético medio (quintil 5 *vs* 3): os individuos cunha puntuación xenética máis alta (quintil 5) tiñan un risco 1.81 (95%IC=1.57-2.09) veces superior de hospitalización que os individuos cun risco medio (quintil 3).

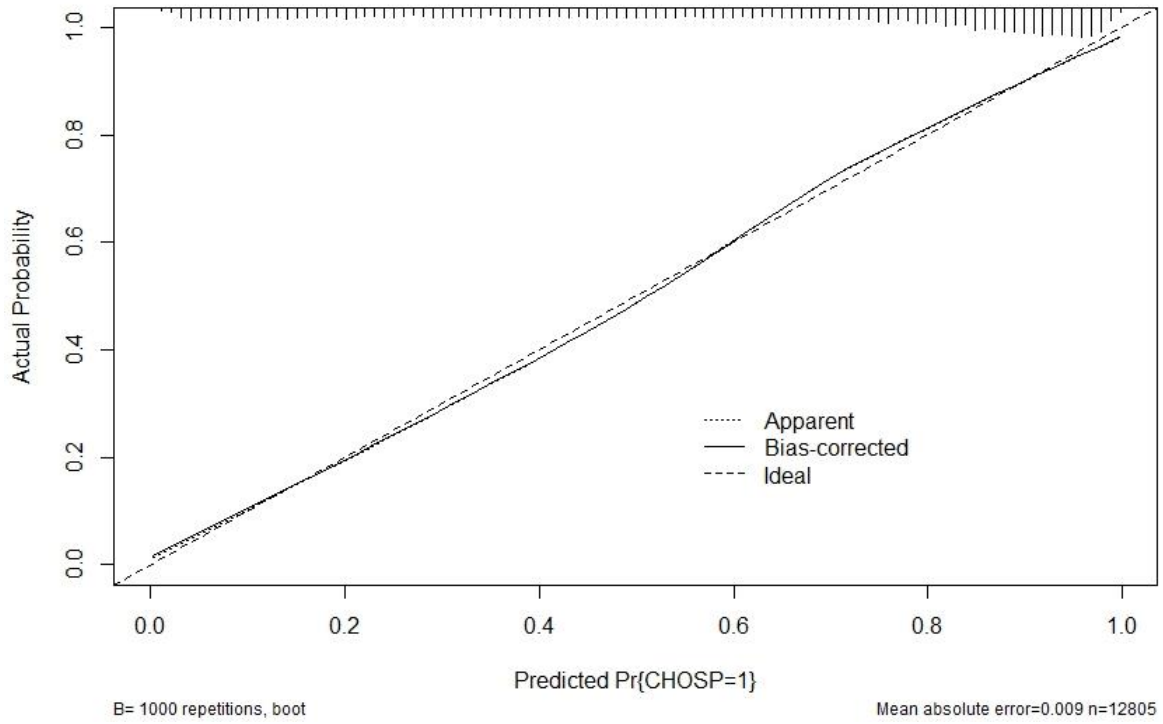


Figura 25. Gráfico de calibración do modelo completo na mostra de adestramento co PRSax tras 1000 remostras *bootstrap*. Obtido coa función *calibrate* do paquete *rms* de R. *Apparent*: modelo base; *bias-corrected*: modelo corrixido polo optimismo; *ideal*: modelo ideal.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Na **Figura 26** poden verse as distribucións do PRS_{ax} para ambas cohortes dentro de casos e dentro de controis tras o proceso de axuste polas PCs. Centrándonos na mostra conxunta, os percentís máis altos son máis frecuentes dentro do grupo de casos que dentro do grupo dos controis, e viceversa (**Figura 27,** panel A). Adicionalmente, seguindo o procedemento de Khera e colaboradores (2018), comparamos as prevalencias de hospitalización por percentil observadas e preditas polo PRS. Como se observa no panel B da **Figura 27**, aínda que existe certa variabilidade nas prevalencias observadas por percentil, si se aprecia unha tendencia crecente similar entre as preditas e as observadas. Non obstante, cómpre puntualizar que este aumento na prevalencia é modesto, a diferencia do atopado para outros fenotipos como o CAD.

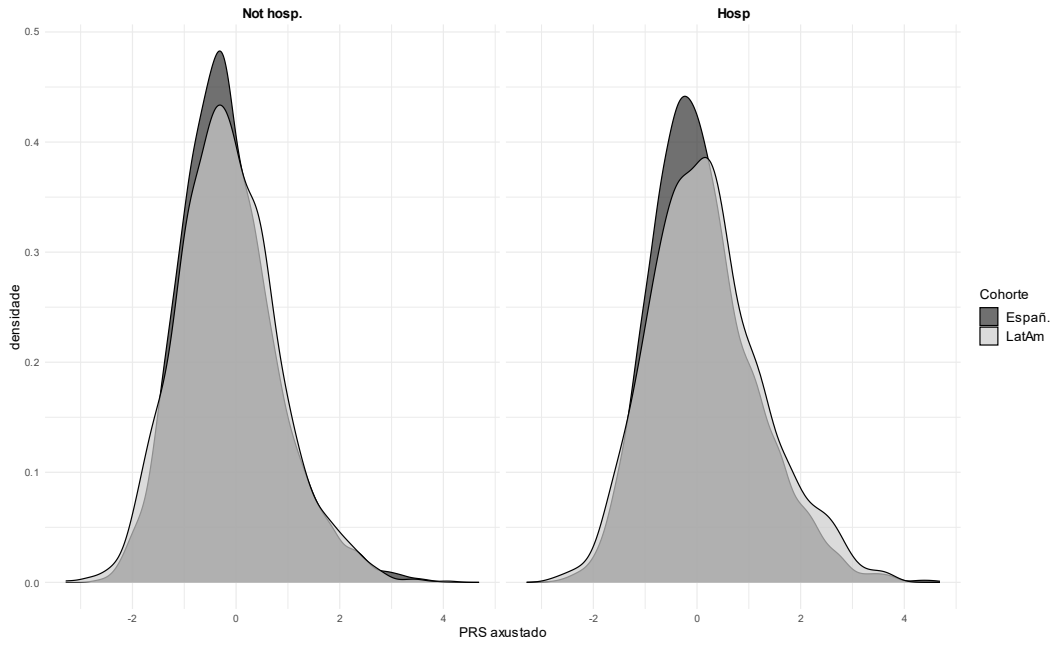


Figura 26. Gráfico de densidade do PRS axustado polas PCs nas cohortes latina e española, estratificado en función do estado caso/control.

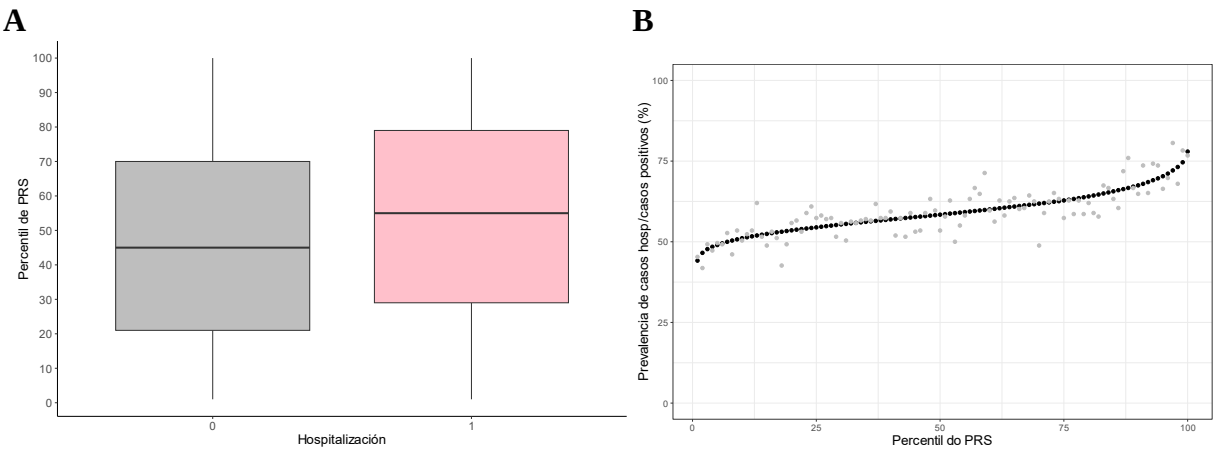


Figura 27. Gráficos de caixas que ilustra a distribución dos percentís do PRS entre casos e controis (A) e gráfico de prevalencia observada *vs* predita en cada percentil do PRS (B). No panel B, os puntos grises corresponden ao nº de casos hosp./nº casos diagn. observados dentro de cada percentil do PRS. Os puntos negros

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

corresponden á media da prevalencia predita de hospitalización (dentro de cada percentil) por un modelo no que o PRS era o único preditor.

A validación externa consistiu en testar o comportamento das predicións en tres escenarios (ver 3.13.5.2). Estes tres escenarios buscaban explorar que aproximación obtiña mellores resultados ao incluír un PRS axustado en base á ancestralidade dunha cohorte de referencia: un modelo adestrado en poboación mixta, un modelo adestrado nunha poboación clasificada nunha poboación discreta cos mesmos criterios que a test (≥80% EUR), ou un modelo adestrado cun subgrupo de individuos xeneticamente similares á mostra test pero derivado dunha mostra mixta. Os resultados móstranse na **Táboa 19**.

Táboa 19. Resultados da validación externa dos tres modelos co PRS_{ax} na cohorte adicional de 437 individuos con GIA predominantemente europea. O número de casos é de 111 e o de controis é de 362. O modelo *neighbors* foi adestrado con 1.877 controis e 3.260 casos. O límite para clasificar aos individuos como casos e controis foi P=0.5. O AUC varía entre 0.5-1, con 0.5 correspondendo a un preditor aleatorio. O *Brier score* varía entre 0 e 1, con 0 indicando unha precisión perfecta.

	Modelo completo	**Modelo *EUR***	**Modelo *neighbors***
Valores de rendemento			
Sensibilidade	0.59	0.80	0.64
Especifidade	0.92	0.78	0.91
PPV	0.70	0.53	0.69
NPV	0.88	0.93	0.89
AUC	0.87	0.85	0.87
Brier score	0.11	0.16	0.11
Coeficientes do PRS_{ax}	0.35	0.35	0.38

O subgrupo de individuos incluído no modelo *neighbors* tiña unha composición xenética similar aos do modelo *EUR*, cunhas medianas de % de ancestralidade europea do 97% para ambos. O modelo completo obtivo valores similares ao modelo *neighbors*, pero o segundo obtivo unha sensibilidade superior. O modelo adestrado coa cohorte española (modelo *EUR*) obtivo peores resultado, excepto na sensibilidade. Esta diferencia é relevante porque indica que os modelos completo e *neighbors* teñen menos capacidade para detectar casos positivos, mais é importante salientar que estes resultados foron obtidos usando un límite aleatorio de P=50% para clasificar casos e controis. Tamén hai que ter en conta que o modelo *EUR* ten menos especifidade, é dicir, sobre-estima o risco de hospitalización en non hospitalizados. Isto pode ser determinante, por exemplo, de cara a usar o modelo para *screening*. De feito, a probabilidade de que, de ser predito como caso, un individuo sexa verdadeiramente un caso (PPV), é superior cos modelos completo e *neighbors.*

Dado que as mostras de adestramento entre o modelo *EUR* e *neighbors* son comparables xeneticamente, exploramos os coeficientes asociados ao PRS. Os tres modelos estimaron un coeficiente similar, polo que as diferencias entre as aproximacións poden deberse ás variables clínicas. Se temos en conta o rendemento do modelo en conxunto, o adestrado con individuos mixtos ofrece predicións máis precisas que o modelo adestrado con individuos europeos. Se,

USC DE SANTIAGO DE COMPOSTELA

ademais, se usa como mostra de adestramento un subgrupo de individuos xeneticamente similares aos test, a sensibilidade do modelo é maior. Con todo, hai que ter en conta que o número de casos test é moi baixo e estes resultados deben tomarse como preliminares.

6.1.3 PRS en relación ás variables clínicas

A segunda parte da análise dos PRS ten por obxectivo o de estudar o risco xenético de hospitalización en base a outras variables clínicas e demográficas. A base de datos do consorcio SCOURGE ten recollidas máis de 300 variables, mais a urxencia da recollida de datos, xunto coa saturación dos centros clínicos durante a pandemia, dificultou que se incluíra toda a información para cada individuo. Por tanto, analizáronse as variables clínicas de máis relevancia para as que existira información de máis do 50% dos individuos dentro do grupo de casos con diagnóstico COVID-19 positivo. Non obstante, cómpre ter coidado coa interpretación das variables con porcentaxes altas de datos perdidos ou un número de casos baixo. Na **Táboa 20** están recollidas as variables usadas en todas as análises, tanto os fenotipos como as comorbilidades.

Táboa 20. Resumo das variables empregadas na valoración do PRS con variables clínicas. Datos calculados dentro do grupo de casos COVID-19 (os controis son individuos non hospitalizados). Antecedentes cardíacos refírese a presentar algunha das seguintes: cardiopatía isquémica, insuficiencia cardíaca, arritmia cardíaca ou vasculopatía periférica. Antecedentes respiratorios inclúe EPOC ou outras patoloxías respiratorias.

Variable	N ctrl/casos	% falt.	N homes ctrl/casos	N mulleres ctrl/casos
Resultado da enfermidade				
Hospitalización	5269/7594	0.18	1531/4376	3738/3199
Enferm. crítica	10578/1888	3.26	4402/1306	6171/568
Morte	11032/1105	5.81	4932/637	6091/459
Enferm. asintomática	11767/943	1.37	5547/304	6201/639
Complicacións				
Ventilación (vmi/vmni/cánulas)	9728/1521	12.70	4115/1045	5607/463
Tromboembolismo pulmonar	8245/728	30.37	3977/383	4251/345
Tromboembolismo venoso	4739/338	45.87	2323/200	2418/135
Infiltración pulmonar (radio-tórax)	1070/5713	47.36	413/3361	657/2334
Miocardiopatía	8237/146	34.94	4013/96	4208/50
Comorbilidades				
Vasculares	7304/5476	0.82	3026/2878	4275/2582
Hipertensión arterial	8971/3809	0.82	3842/2062	5118/1739
Hipercolesterolemia	11276/1054	0.82	5135/769	6122/735
Diabetes	11445/1335	0.82	5144/760	6285/572
Obesidade	10834/1946	0.82	4954/950	5872/985
Cardíacas	11511/1269	0.82	5162/742	6332/525
Respiratorias	11680/1100	0.82	5255/649	6407/450

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Os modelos empregados en todos os casos foron modelos loxísticos incluíndo como covariables base o sexo, a idade e o país de recrutamento, e o PRS_{ax}. A mostraxe realizada para o GWAS pode presentar un nesgo de selección, que consiste nunha representación inexacta da poboación a estudar porque os casos e os controis teñen distinta probabilidade de ser recrutados (Tripepi et al., 2010). Como análise de sensibilidade re-analizáronse as variables dentro do grupo de casos hospitalizados, salvo a variable de enfermidade asintomática e hospitalización, coa conseguinte redución do tamaño de mostra. O PRS axustouse de dúas formas, sempre estandarizado: (1) cuantitativo, e (2) comparando o quintil 5 co resto dos quintís.

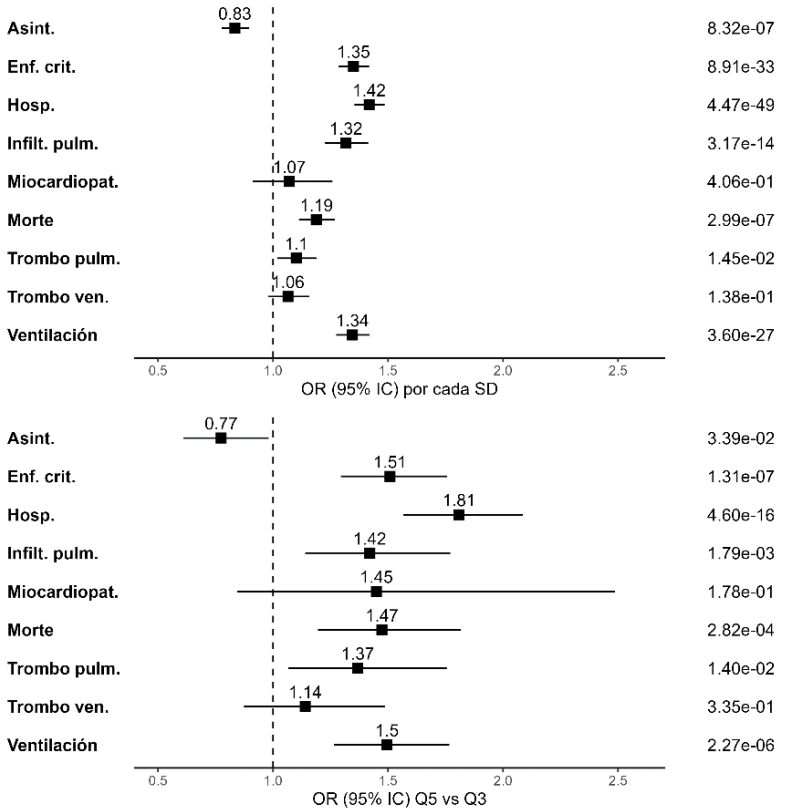


Figura 28. Gráfico tipo bosque que ilustra as OR do PRS estandarizado para cada variable na cohorte conxunta dentro do total de casos COVID-19+. O PRS foi incluído de forma cuantitativa (arriba) e comparando o 20% dos individuos con máis risco xenético co resto dos individuos. O nivel de significación é de $p<2.78x10^{-03}$, tras a corrección por Bonferroni para 18 test.

Nas **Figura 28** e **29** pódense ver as análises principais e de sensibilidade, respectivamente. O PRS está asociado á enfermidade crítica e asintomática, á hospitalización, á presencia de infiltración pulmonar en radiografía de tórax, e á necesidade de ventilación mecánica. O test de asociación para tromboembolismo pulmonar non acada significación tras a corrección por Bonferroni para 19 tests. Igualmente, a asociación coa enfermidade asintomática desaparece na análise de PRS por quntís. Para as outras dúas complicacións derivadas da COVID-19, cardiomiopatía e trombo venoso, non existe evidencia da súa asociación co risco xenético.

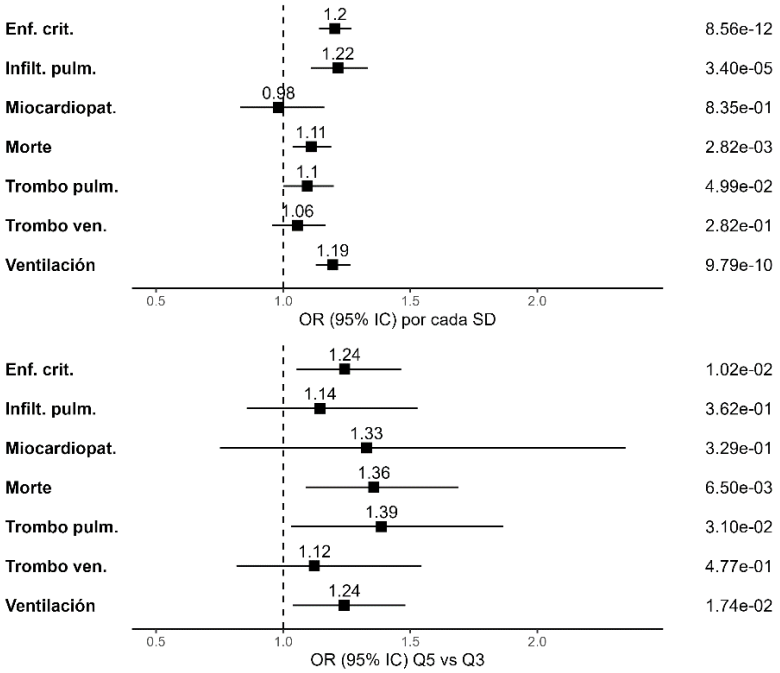


Figura 29. Gráfico tipo bosque que ilustra as OR do PRS estandarizado para cada variable na análise de sensibilidade na cohorte conxunta dentro de casos hospitalizados. O PRS foi incluído de forma cuantitativa (arriba) e comparando o 20% dos individuos con máis risco xenético co resto dos individuos. O nivel de significación é de $p<3.5x10^{-03}$, tras a corrección por Bonferroni para 14 test.

En xeral, as OR atenúanse para todas as variables ao analizar unicamente o grupo de casos hospitalizados. Isto é esperable, tanto polo nesgo de selección como pola diminución da potencia estatística, pero tamén porque o risco entre hospitalizados podería non presentar diferencias. Non obstante, as variables asociadas no grupo completo tamén o están no grupo de hospitalizados cando o PRS se considera de forma cuantitativa, o que indica que a base xenética de risco presenta unha distribución distinta tamén entre individuos que requiren hospitalización. Tras axustar o PRS en forma binaria (Q5/Q3) pérdense todas as asociacións.

A distribución do número de comorbilidades entre os quintís 1 e 5 (20% con menor risco xenético, N=2.577, e 20% con maior risco xenético, N=2.577) foi similar ($p\text{-valor}_{\chi^2}=0.012$). En cambio, a porcentaxe de homes e mulleres no primeiro quintil foi do 43.2% e 56.8%, respectivamente, mentres que os individuos con máis risco xenético eran o 50.1% homes e o 48.9% mulleres. A idade de hospitalización foi menor para o grupo con maior risco (62.1 anos) en comparación co de menor risco (68.2 anos). A modulación do efecto do PRS por diferentes factores de risco explorouse a través da interacción do mesmo co sexo, a idade (estratificada en <65 anos e ≥65 anos), e a presencia de comorbilidades (diabetes, obesidade, hipertensión arterial, hipercolesterolemia, antecedentes respiratorios e antecedentes cardíacos). As comorbilidades testáronse de dúas formas: (1) ausencia de comorbilidades vs presentar polo menos 1 comorbilidade, e (2) número de comorbilidades, de 0-6.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Táboa 21. Resultados das interaccións co PRS con factores de risco clínicos.

Variable dependente	Variable interacción	P-valor interacción	OR interacción (95% IC)
Hospitalización	Sexo (masc.=1/fem.=2)	0.983	1.00 (0.91-1.10)
	Idade (<65/ ≥ 65)	$4.94x10^{-05}$	0.79 (0.70-0.89)
	> 1 comorbilidade	0.55	1.03 (0.94-1.13)
	Nº de comorbilidades	0.85	1 (0.94-1.04)
Enfermidade crítica	Sexo (masc.=1/fem.=2)	0.03	1.12 (1.00-1.30)
	Idade (<65/ ≥ 65)	$1.60x10^{-04}$	0.83 (0.75-0.91)
	> 1 comorbilidade	0.94	0.99 (0.86-1.15)
	Nº de comorbilidades	0.21	0.98 (0.94-1.04)

O efecto do PRS non presentou diferencia entre homes e mulleres nin entre individuos con e sen comorbilidades (ver **Táboa 21**). En cambio, para os individuos maiores de 65 anos, a base xenética ten un efecto máis moderado que en individuos de menor idade (CHOSP: $OR_{<65}$=1.47 [1.40-1.55], $OR_{\geq 65}$=1.14 [1.04-1.27]; CSEV4: $OR_{<65}$=1.43 [1.34-1.52], $OR_{\geq 65}$=1.22 [1.13-1.32]). Dentro do grupo de menores de 65 anos, os individuos do quintil 5 teñen unha OR de estar hospitalizados de 2.11 (1.87-2.39) con respecto aos individuos con risco xenético medio. Para estudar a capacidade preditiva dos PRS en relación a estes factores, estratificouse a mostra e calculouse o ΔR^2 en cada grupo. O ΔR^2 non variou entre homes e mulleres (ΔR^2_{homes}=0.018 [0.012-0.020], $\Delta R^2_{mulleres}$=0.017 [0.012-0.023]), nin tampouco entre presentar 0 ou algunha comorbilidade (ΔR^2_{com}=0.017 [0.008-0.024], $\Delta R^2_{non_com}$=0.017 [0.011-0.022]). En cambio, a capacidade preditiva do PRS en individuos <65 anos foi considerablemente superior que en individuos de maior idade e que incluso na cohorte completa ($\Delta R^2_{\geq 65}$=0.004 [-0.003-0.006], $\Delta R^2_{<65}$=0.029 [0.021-0.036]).

6.1.4 Modelos FRC e FRC+PRS

Por último, estudouse a utilidade clínica do PRS mediante os modelos FRC (Factores de Risco Clínicos) e FRC+PRS. A diferencia no pseudo-R^2 (corrixida polo optimismo) entre os dous modelos, en favor do modelo principal, foi de 0.015, cunha OR asociada ao PRS de 1.41 (95% IC=1.34-1.48). Aínda que estes valores son lixeiramente inferiores aos obtidos no modelo onde só se axustaba por idade, sexo e país de recrutamento (ΔR^2=0.016, OR=1.42), a redución non foi substancial, o que indica que o PRS incrementa a probabilidade de hospitalización por COVID-19 independentemente dos factores clínicos. Aumentar unha desviación típica no PRS multiplica as *odds* de ser hospitalizado nunha magnitude similar á de presentar unha comorbilidade adicional (OR=1.43 [1.36-1.50]). Na **Figura 30** móstranse os gráficos de calibración para ambos modelos, e obsérvase que o modelo principal (panel B) resulta en probabilidades preditas máis aproximadas ás observadas e un MAE inferior.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

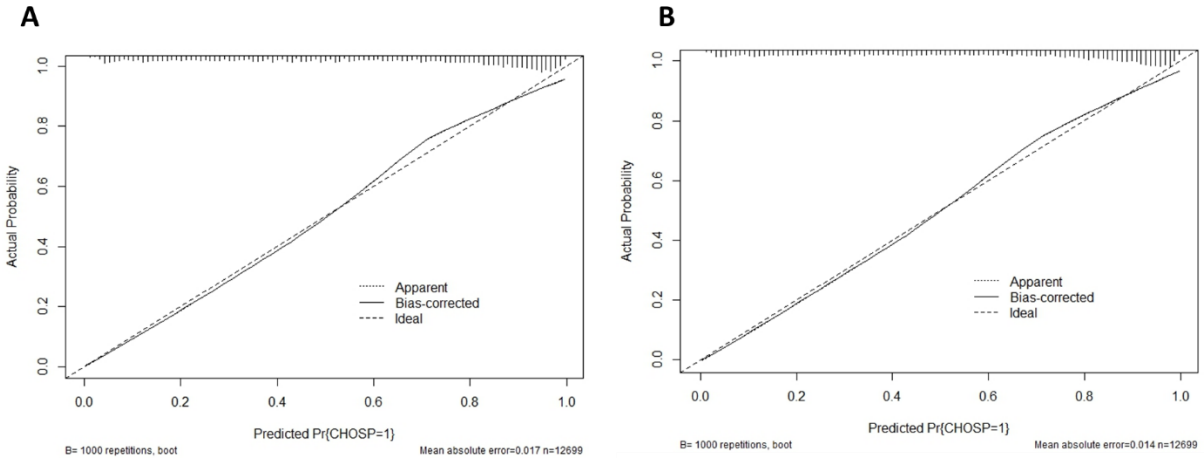


Figura 30. Gráficos de calibración por bootstrap para os modelos FRC (A) e FRC+PRS (B). O modelo base inclúe as covariables sexo, idade, número de comorbilidades e país de recrutamento. O modelo principal incorpora, a maiores, o PRS. O MAE para o modelo base é de 0.017. O MAE para o modelo principal é de 0.013.

Estes dous modelos empregáronse para predicir a probabilidade de hospitalización dos 473 individuos test, dos que 111 eran individuos hospitalizados. Os modelos validáronse na mostra completa e estratificados por sexo e idade. Os valores que miden o rendemento dos modelos de clasificación dependen do punto de corte no rango de probabilidades, polo que para o modelo completo calculáronse a sensibilidade e especificidade para diferentes límites de clasificación, o que pode verse na **Táboa 22**. O modelo FRC+PRS mellorou lixeiramente tanto a sensibilidade como a especificidade para algúns límites de probabilidade, especialmente a partir do 40%. En xeral, se escolléramos como límite de clasificación en alto risco valores de probabilidade altos, ambos modelos ofrecen pouca especificidade.

Táboa 22. Valores de especificidade e sensibilidade para a mostra test dos modelos base e principal en función de distintos cortes de probabilidades preditas. Lím. P(hosp): punto de corte da probabilidade de hospitalización para definir caso/control.

Lím. P(hosp)	Sensibilidade FRC	Especificidade FRC	Sensibilidade FRC+PRS	Especificidade FRC+PRS
10%	0.123	0.982	0.124	0.973
20%	0.249	0.955	0.249	0.955
30%	0.370	0.928	0.370	0.928
40%	0.492	0.900	0.497	0.919
50%	0.616	0.883	0.622	0.901
60%	0.729	0.820	0.740	0.856
70%	0.842	0.766	0.845	0.775
80%	0.931	0.631	0.925	0.613
90%	0.986	0.387	0.992	0.405

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

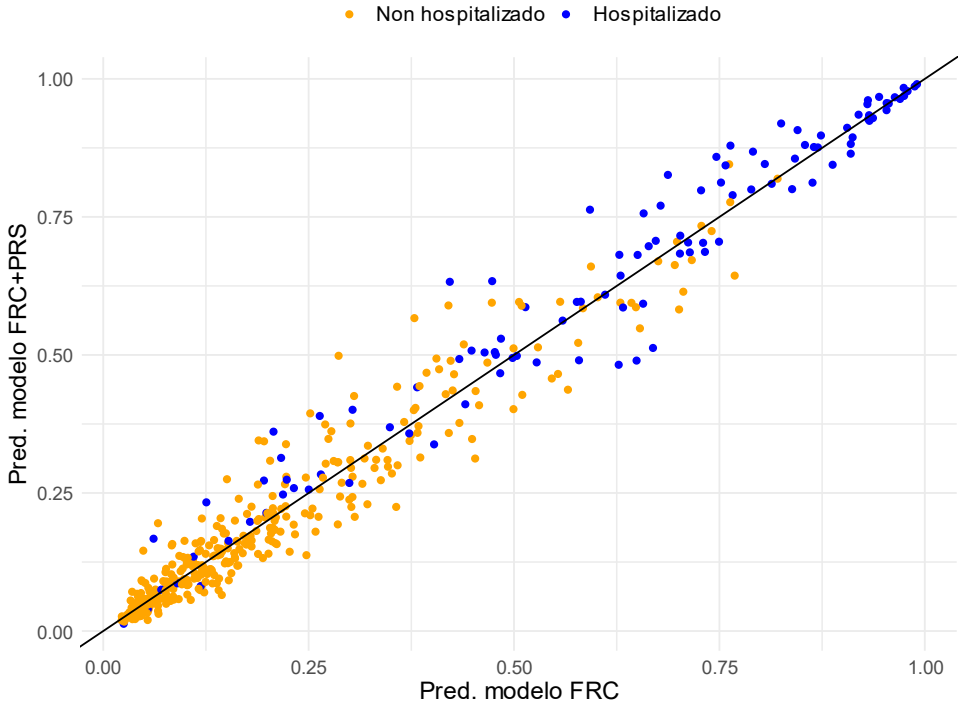


Figura 31. Gráfico que ilustra as predicións da mostra test para os modelos base, con factores de risco clínicos (FRC), e o modelo principal, con factores de risco clínicos e o PRS (FRC+PRS). Os puntos que se atopan por enriba da diagonal tiveron unhas probabilidades preditas maiores no modelo FRC+PRS que no modelo FRC, e viceversa.

En xeral, para os casos aumentou a probabilidade predita de hospitalización ao incluír o PRS (**Figura 31)**. Explorouse o NRI en dous puntos de corte aleatorios, P=0.4 e P=0.7. Con P=0.4, o NRI global foi de -0.002 (95%IC=-0.039-0.035), o NRI para os casos foi de 0.009 (-0.018-0.041), e para os controis de -0.011 (-0.03-0.008). Con P=0.7, o NRI global foi de 0.026 (-0.024-0.076), o NRI para casos de 0.018 (-0.030-0.076) e o NRI para controis de 0.008 (-0.003-0.021). O PRS, empregando un NRI de 0.4, produce máis erros na reclasificación dos controis do que a mellora, e o seu valor global é negativo. Empregando un valor de corte do 0.7, o NRI global obtido é positivo. Porén, pero os intervalos de confianza cruzan o 0 en todos os casos, se ben isto pode deberse ao tamaño de mostra, e os valores de NRI son moi baixos.

Dado que non dispoñemos da información para seleccionar un punto de corte óptimo, explorouse tamén o funcionamento do modelo FRC+PRS á hora de identificar o 20% dos individuos con maiores probabilidades preditas de hospitalización. Considerouse un risco xenético alto cando o valor do PRS fora igual ou superior ao quintil 5 do PRS na cohorte de adestramento (≥0.765). O 20% con maior risco absoluto de hospitalización, cuxo corte de probabilidade era ≥0.523, identificaríanse o 61.3% dos hospitalizados. A mediana de idade destes individuos era de 65.7 anos, o 80.1% eran homes, o 89.2% presentaban algunha comorbilidade e constituían o 78.1% dos doentes falecidos por COVID-19. O 30.1% caía na categoría de PRS alto. Deses individuos con PRS alto, o 17.9% eran mulleres, o 10.7% non presentaban comorbilidades e a mediana de idade foi de 62.4 anos. A sensibilidade e especificidade foron do 93.1% e 61.3%, respectivamente.

6.1.4.1 Curvas DCA

Calculáronse as curvas de DCA cos resultados da predición no grupo de test. Na **Figura 32** están representadas as curvas de decisión corrixidas pola prevalencia para os modelos de predición FRC e FRC+PRS. Escolleuse un punto de corte de probabilidade para a curva 40%, pois é razoable pensar que a valoración dunha intervención se dea en individuos cuxo risco de hospitalización sexa menor ao 40%, mentres que aqueles cun risco superior serán sometidos sempre a intervención. Non obstante, este corte é aleatorio e dependerá da intervención en cuestión. Observamos que, para límites de entre o 1-5%, os modelos construídos non ofrecen ningún beneficio sobre a estratexia de "tratar a todos". Isto significa que en caso dun tratamento pouco invasivo, para o cal se requirían probabilidades baixas de hospitalización por COVID-19, a mellor estratexia sería a de intervención completa na poboación. Para un rango de límites de entre o 15-35%, o modelo FRC+PRS supón un beneficio neto lixeiramente superior ao modelo FRC. Porén, neste modelo non se consideran os custes económicos asociados a obter a información necesaria para calcular o PRS nos individuos.

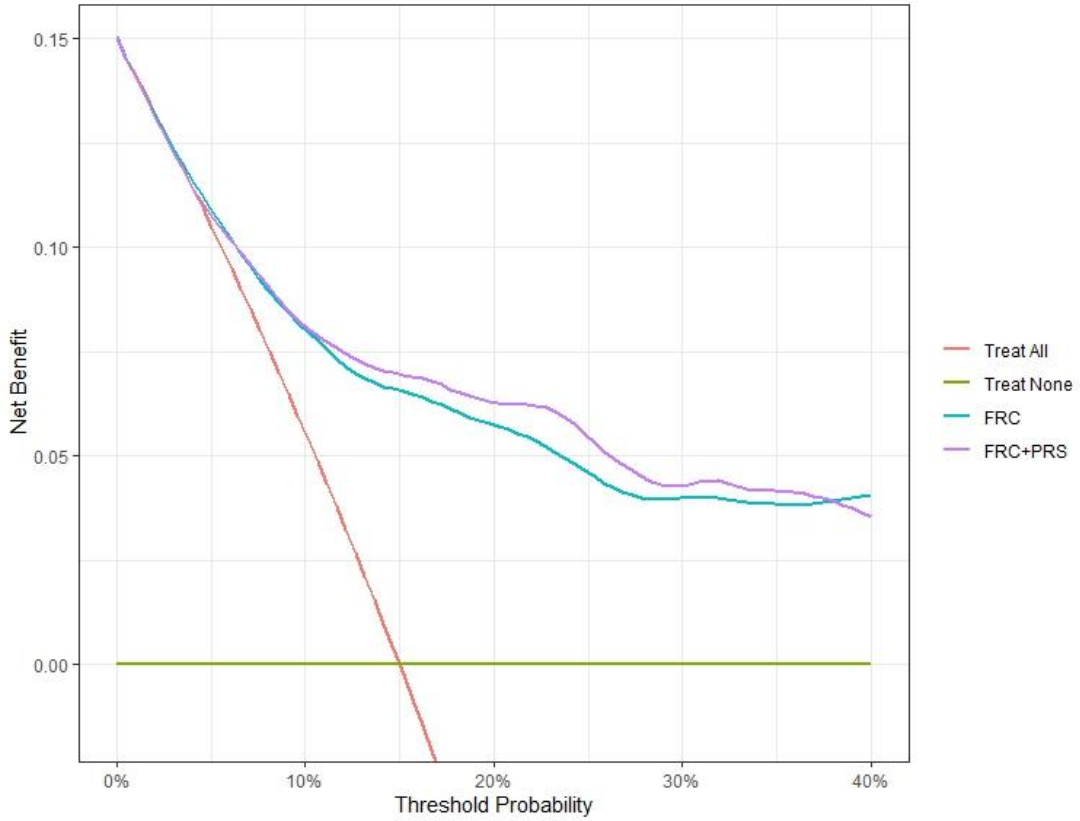


Figura 32. Curva DCA calculada coa mostra test, onde se representan 4 estratexias. Estas son: tratar a toda a poboación, non tratar a ningunha, e tratar en función dos modelos FRC e FRC+PRS. Como prevalencia de hospitalización entre casos COVID+ usouse un valor do 15%.

A modo ilustrativo, calculáronse as curvas para un escenario adicional. Nel, considéranse 3 estratexias: (a) intervir nos individuos clasificados como alto risco no modelo FRC, é dicir, só considerando factores clínicos con P_{FRC} >0.3; (b) intervir nos individuos clasificados como

alto risco no modelo FRC+PRS con $P_{FRC+PRS}$ >0.3, e (c) intervir nos individuos clasificados como alto risco clínico (P_{FRC} >0.3) pero tamén nos individuos con risco clínico intermedio (P_{FRC} >0.15) que estean no 20% da poboación con maior risco xenético (quintil 5 do PRS_{ax} estandarizado na poboación de adestramento, con un valor maior a 0.76). A curva DCA deste escenario pode verse na **Figura 33**. Para casos nos que o límite estea por debaixo do 5%, ningún dos modelos supera a estratexia de intervención para todos. No rango entre o 5% e o 30%, os modelos (a) e (c) son similares, e a partir do 30% o modelo con maior beneficio neto é o (b). Ningunha das estratexias é superior en todo o rango de probabilidades de risco, e cómpre ter en conta que o PRS ten uns custos asociados, polo que a utilidade clínica do PRS en comparación a outros factores clínicos non está clara. Porén, os puntos escollidos para a clasificación en alto/baixo risco tamén son aleatorios.

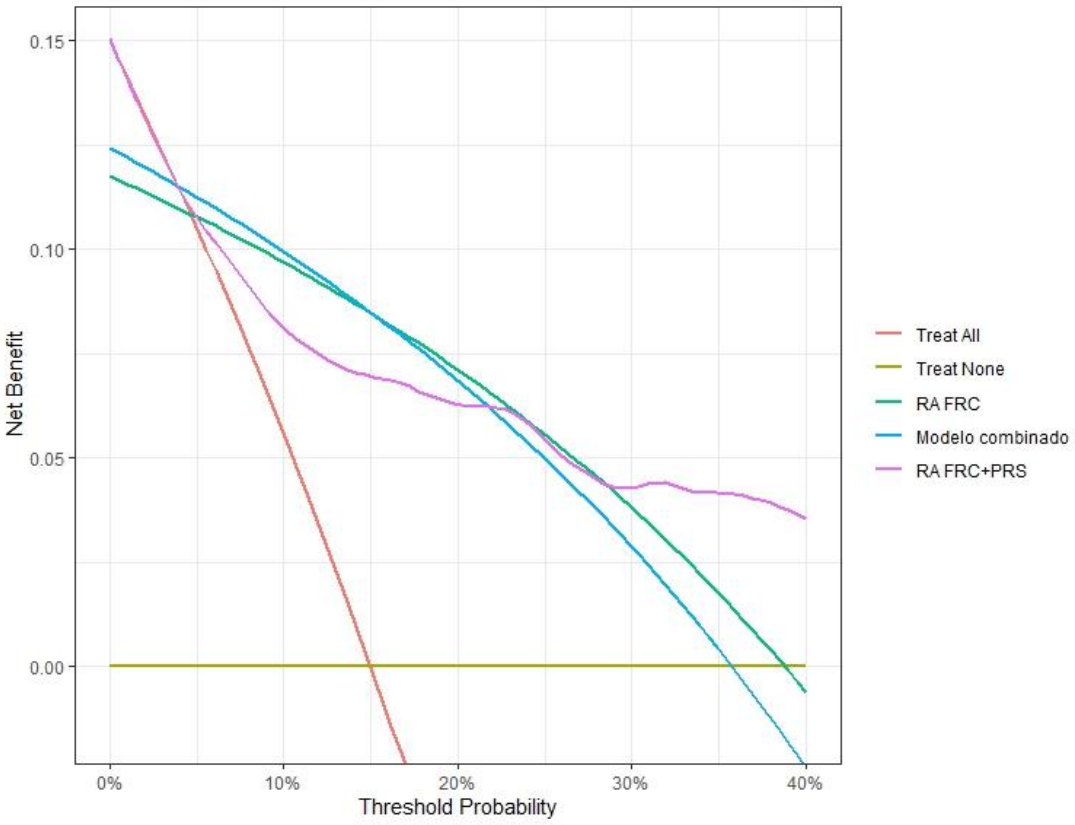


Figura 33. Curva DCA calculada coa mostra test, onde se representan 5 estratexias diferentes. Estas es estratexias son: tratar a toda a poboación, non tratar a ningunha, tratar só aos clasificados como risco alto segundo o modelo FRC (RA FRC), tratar só aos clasificados como risco alto segundo o modelo FRC+PRS (RA FCR+PRS), ou unha estratexia combinada que consiste en tratar aos individuos con risco alto segundo o modelo FRC ou que teñan risco clínico medio e un PRSax estandarizado superior a 0.76 (quintil 5 dos individuos de adestramento, modelo combinado). Como prevalencia de hospitalización entre casos COVID+ usouse un valor do 15%.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

6.1.4.2 Exemplo de aplicación clínica do PRS

Como exemplo ilustrativo sobre a aplicación clínica dun PRS, seleccionáronse dous individuos aleatorios da mostra test para explorar o seu risco absoluto de hospitalización co modelo FRC+PRS. O paciente A é un home de máis de 65 anos, sen ningunha comorbilidade adicional, de Brasil, hospitalizado cun cadro crítico e situado no percentil 60 do PRS. A paciente B, é un home de menos de 30 anos, con obesidade, tamén de Brasil, que non foi hospitalizado, cursou unha enfermidade leve e está situado no percentil 15 do PRS.

O modelo estima unhas probabilidades absolutas de hospitalización do 64.3% e 11.9%, respectivamente. Calculouse o risco de hospitalización para dous individuos ficticios, con valores de sexo, idade, país de hospitalización e comorbilidades iguais aos pacientes A e B, mais cun valor do PRS na mediana dos individuos da mostra de adestramento. Deste xeito, podemos calcular a desviación predita con respecto ao risco medio de hospitalización atribuída á xenética para cada paciente (Multhaup et al., 2017). O paciente A ten un risco de ser hospitalizado un 19.7% superior a aquel coas mesmas características debido á súa predisposición xenética. Para o paciente B, a base xenética predispono nun 23.1% menos de probabilidades de ser hospitalizado. Non obstante, esta é unha cantidade relativa, sendo imprescindible facer esta distinción á hora de comunicar os resultados: o individuo A ten unha probabilidade de hospitalización do 64.3% con respecto ao 54.2% da súa parella ficticia; o individuo B ten unha probabilidade de hospitalización do 11.9% con respecto ao 15.4% da súa parella ficticia. Realizouse tamén a comparativa entre os pacientes A e B con unhas parellas ficticias que presentaran o mesmo valor do PRS pero diferente número de comorbilidades para mostrar como afecta o ambiente. Así, o risco de hospitalización do individuo A é un 0.54% menor que de presentar unha comorbilidade (P_{hosp}=64.6%), pero un 11.2% menor que de presentar 2 (P_{hosp}=72.3%). O paciente B ten un risco 26.4% inferior ao paciente ficticio que presenta dúas comorbilidades (P_{hosp}=16.12%), pero un 37.8% maior ao paciente ficticio que non presenta comorbilidades (P_{hosp}=8.6%).

Adicionalmente, calculouse o risco absoluto asociado ao PRS empregando a calculadora desenvolvida por Pain e colaboradores (2022), que ademais axusta a estimación pola prevalencia da enfermidade na poboación. Especificáronse os valores estandarizados do PRS para os pacientes A e B, unha prevalencia poboacional do 15% (casos hosp./casos diagn.) e un AUC asociada ao PRS (como único preditor) de 0.57. A calculadora estimou que, de entre os individuos cunha predisposición xenética equivalente ao do paciente A, o 14.9% son hospitalizados. Co risco xenético do paciente B, só son hospitalizados o 12% dos diagnosticados.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

6.2 DISCUSIÓN

Os índices de risco xenético ou PRS son unha ferramenta que conta con potencial clínico na predición de patoloxías. Porén, o campo aínda está en vías de desenvolvemento para resolver as limitacións que presentan. Algunhas das máis importantes son a dificultade que existe para xeneralizalos dunha poboación a outra e a sobre-representación das poboacións europeas nos estudos de PRS, unhas problemáticas que radican na baixa diversidade nos GWAS empregados para construír os PRS. Tampouco existe un marco metodolóxico de referencia, polo que para cada fenotipo poden construírse multitude de PRS e cada un destes pode ter un rendemento diferente dependendo da cohorte diana, como se observa no PRS Catalog. Neste terceiro capítulo exploramos ambas limitacións, empregando a COVID-19 como fenotipo e dúas cohortes poboacionais diferentes, unha recollida en España, e outra recollida maioritariamente en Latino-América. O esforzo internacional realizado ao redor da COVID-19 permitiunos construír un PRS derivado dun GWAS multi-poboación. Por outro lado, a cohorte latinoamericana SCOURGE, recrutada en colaboración con grupos e hospitais de Latino-América, ofreceunos unha oportunidade para comparar a capacidade preditiva do PRS en poboación non europea, e testar unha aproximación metodolóxica baseada no adestramento en poboacións mixtas.

6.2.1 Avaliación da capacidade preditiva do PRS entre poboacións

A primeira parte da análise consistiu na selección do PRS con mellor funcionamento na predición da hospitalización por COVID-19, escollido tras comparar varias aproximacións metodolóxicas. Comparáronse PRS construídos con datos de descubrimento de orixe europea e con datos de descubrimento multi-poboación, e empregáronse tres algoritmos diferentes: o C+T (refinando os parámetros r^2, ventá e p-valor), e dous algoritmos baiesianos, PRS-CS e PRS-CSx, este último orientado á PRS multi-poboación. O PRS con mellores resultados, tanto para nos individuos españois como nos individuos latinoamericanos, foi o derivado con PRS-CSx a partir das 5 análises específicas de poboación de HGI. Isto é indicativo de que a combinación de diferentes poboacións nos GWAS, aínda empregando estudos con baixa potencia estatística, é unha estratexia que beneficia incluso ás poboacións europeas, pois permite incrementar os tamaños de mostra e mellorar a detección de sinais compartidos. Curiosamente, os PRS multi-poboación tiveron un mellor rendemento na cohorte latinoamericana, e para os PRS europeos observáronse rendementos similares entre ambas poboacións. Se ben este comportamento é oposto ao observado noutros estudos (Ding et al., 2023; Martin et al., 2017; Y. Wang, Kanai, et al., 2023), existen casos como o da presión arterial, cancro colorrectal e a diabetes tipo II, nos que un mesmo PRS obtivo resultados semellantes, ou incluso superiores, en poboación latina con respecto á europea (Ge et al., 2022; Kurniansyah et al., 2023; Rosenthal et al., 2024). As poboacións latinoamericanas presentan unha mestura de 3 compoñentes ancestrais (NAM, EUR e AFR), unha característica que pode ser beneficiosa en escenarios nos que se acoplan diferentes poboacións na mostra de descubrimento. Así, as variantes causais que son de baixa frecuencia nalgunha poboación poden acadar frecuencias máis altas nos individuos mixtos, mellorando a predición global. A

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

estratificación da cohorte latina en grupos polas proporcións de ancestralidade non revelou predicións significativamente peores en ningún dos mesmos. Non obstante, é necesario ter en conta que a nosa mostra presenta unhas proporcións máis altas de GIA NAM e EUR que de AFR, sendo esta última a que peor resultados obtén na transferencia dos PRS (Bitarello & Mathieson, 2020; Majara et al., 2023). Por último, aínda que nas análises de HGI predominan os individuos europeos, os PRS que construímos abarcan variantes de todo o xenoma, capturando efectos xenéticos que non acadaron o límite de significación GWAS e alixeirando o nesgo introducido por este desequilibrio. De feito, dentro do grupo dos PRS C+T, as combinacións de parámetros que abarcaban máis variantes foron os que obtiveron valores máis altos do ΔR^2 na cohorte latinoamericana.

Ao estudar o comportamento do PRS en relación á variable de enfermidade crítica, polo contrario, si que se observou unha mellor predición dos PRS europeos na cohorte española que na latinoamericana. Cómpre destacar que os doentes hospitalizados na cohorte europea están distribuídos ao longo da escala de gravidade. Na cohorte latinoamericana, en cambio, os doentes hospitalizados están concentrados dentro das categorías máis altas. Por tanto, non se pode descartar que o efecto observado na predición da hospitalización sexa causado por diferencias no criterio de hospitalización entre as cohortes. Pese a isto, a capacidade preditiva dos PRS multi-poboación para enfermidade crítica son semellantes entre ambas mostras, o que quere dicir que non existe evidencia de que os PRS rendan peor na cohorte latina. Ademais, a predición dos PRS é susceptible de variar en función de variables medioambientais ou da composición da mostras, incluso dentro dun grupo de individuos xeneticamente similares (Mostafavi et al., 2020).

Os datos de descubrimento máis habituais para construír PRS en poboacións diversas son os derivados de meta-análises multi-poboación ou análises específicas de poboación, as cales soen ter tamaños de mostra considerablemente pequenos (Y. Wang et al., 2022). O noso estudo demostra que, polo menos para o fenotipo de hospitalización por COVID-19, a incorporación de datos derivados de individuos non europeos mellora a predición incluso en poboacións europeas. Esta aproximación, non obstante, non nos permitía estimar un único efecto para todos os individuos, posto que a OR asociada ao PRS na cohorte europea (OR=1.38) foi lixeiramente inferior ao da cohorte latinoamericana (OR=1.64). Optamos entón por explorar o PRS nun contexto onde non fora posible clasificar aos individuos en poboacións ou grupos discretos. Xa que o GWAS de descubrimento incluía maioritariamente individuos europeos, o risco de que esta precisión reducida se deba a que o PRS non captura correctamente a arquitectura xenética da poboación española é reducido.

6.2.2 Obtención dun PRS único para a hospitalización por COVID-19

Unha vez constatamos que o PRS non resultaba en peores predicións na cohorte latina, optamos por combinar as dúas cohortes e derivar un PRS axustado pola ancestralidade, seguindo o protocolo de Hao e colaboradores (2022). O modelo base obtivo bos resultados no rendemento da predición e na calibración. Posteriormente, estudamos o comportamento da predición en función da composición da mostra na que se adestra o modelo. Para isto establecemos un escenario hipotético onde se empregaría un mesmo PRS para todos os

individuos, pero no cal adestraríamos o modelo para a predición do risco nun grupo de individuos xeneticamente similares. Sabendo que a precisión da predición varía de forma individual coa distancia xenética do individuo á mostra do GWAS (Ding et al., 2023), é probable que tamén varíe en función da distancia xenética á mostra onde se adestra o modelo. Esta estratexia non require da asignación previa de individuos en clústeres pero permite que o efecto do PRS para cada paciente sexa lixeiramente diferente dependendo da súa composición xenética (estimado a través do axuste do modelo nun grupo xeneticamente similar). Este procedemento foi aplicado a un grupo test independente con ascendencia europea. O rendemento do modelo foi só lixeiramente superior ao modelo adestrado coa cohorte mixta completa, aínda que isto era esperado dado que en ambos casos a ancestralidade con máis peso nas mostras de adestramento era a europea. En todo caso, parece que esta aproximación metodolóxica pode ser viable na implementación de PRS en cohortes mixtas, mais debe testarse en cohortes con máis diversidade. Os resultados obtidos poden non replicarse en poboacións con maior ascendencia africana, ou incluso en poboacións asiáticas. Por último, cómpre recordar que, aínda que se analizaron conxuntamente os individuos, o método de PRS-CSx necesita datos resumo e de LD específicos para cada grupo poboacional.

Ata este momento desenvolvéronse e testáronse moi poucos PRS para COVID-19. Horowitz e colaboradores (2022) construíron un GRS, é dicir, un índice xenético a partir das variantes con evidencia de asociación (N=6, PGS Catalog ID PGS002272; N=12, PGS Catalog ID: PGS002273), para o fenotipo de susceptibilidade á infección. Como apunte, este estudo analizou o GRS nun grupo de poboación latinoamericana, mais con baixo tamaño de mostra (N_{casos}=199, $N_{controis}$=3.553). Outro estudo construíu 2 PRS, un de 9 variantes para susceptibilidade e outro de 17 variantes para enfermidade grave, a partir dos resultados da versión 6 das meta-análises do HGI (Farooqi et al., 2023), mentres que un terceiro testou diferentes PRS construídos cos resultados da versión 5 (Crossfield et al., 2022). Por último, un PRS derivado da meta-análise multi-poboación da versión 7 foi usado para un estudo PheWAS (Fritsche et al., 2023). Non obstante, Fritsche e colaboradores (2023) non analizaron o PRS na súa forma clásica, senón que buscaban estudar a relación entre a enfermidade da COVID-19 e antecedentes a través do risco xenético. O noso PRS é o primeiro en derivarse -ata onde sabemos- mediante a combinación das poboacións individuais con PRS-CSx e usando os resultados máis actualizados do HGI. Ademais, é a primeira vez que se explora un PRS xenómico de COVID-19 en poboación latinoamericana.

Segundo os nosos resultados, presentar un maior risco xenético a padecer unha enfermidade de COVID-19 que requira hospitalización está asociado a outras consecuencias da mesma, como padecer un cadro crítico ou falecer. Este PRS, ademais, explica o 2% da variabilidade do fenotipo cando a prevalencia da hospitalización (entre casos diagnosticados) é do 15% e un 1.4% cando esta é do 5%. Os individuos con menor risco xenético teñen maior odds de ser asintomáticos. A análise do polimorfismo rs10490770 (do xen *LZTFL1*) en forma de *carrier/non-carrier* (Nakanishi et al., 2021) revelou unhas OR superiores que o noso PRS para os fenotipos de hospitalización ($OR_{rs10490770}$=1.5; OR_{PRS}=1.42) e enfermidade crítica ($OR_{rs10490770}$=1.7; OR_{PRS}=1.35), e estaba significativamente asociado a tromboembolismo venoso, mentres que o PRS non ($OR_{rs10490770}$=1.7; OR_{PRS}=1.06). Porén, hai que ter en conta que

o PRS está constituído por variantes con efectos que se compensan, ofrecendo unha predición do risco máis fiable que un único SNP. Horowitz e colaboradores (2022) reportaron unhas OR para o PRS de 6 variantes de 1.38 para hospitalización e 1.58 para enfermidade crítica en poboación europea, comparando o 10% superior da distribución co resto dos individuos. A OR para individuos latinos foi de 1.57, mais os intervalos de confianza foron moi amplos polo seu baixo número de casos. No estudo de Farooqi e colaboradores (2023), pola súa banda, comparouse o rendemento do PRS en diferentes grupos étnicos do UK-Biobank, obtendo unhas OR superiores para grupos asiáticos que para grupos europeos. En xeral, o noso PRS obtivo asociacións máis fortes que os PRS non xenómicos, o que indica que incorporar a base xenética non identificada polos GWAS supón un incremento na predición. Este PRS, ademais, non inclúe a información dos GWAS de SCOURGE para ningunha das cohortes, polo que sería posible que o PRS resultara un mellor preditor ao incorporalas.

6.2.3 Estudo do potencial clínico do PRS

Analizamos tamén o risco xenético en relación aos factores de risco clínicos, como o sexo, a idade, e a presencia de comorbilidades. Aínda que non se observaron diferencias na asociación do PRS entre o sexo e a presencia ou non de comorbilidades, si observamos un forte efecto da idade (na súa forma estratificada). Concretamente, o efecto da xenética é maior para individuos con menor idade, tanto en termos da OR como de ΔR^2. Isto é consistente co obtido nos estudos de Nakanishi e colaboradores (2021) co polimorfismo do cromosoma 3 e co GRS de Horowitz e colaboradores (2022), e co feito de que os individuos de maior idade teñen un risco superior de padecer un cadro grave debido ás comorbilidades e ao seu estado de debilidade. Ningún dos estudos previos atopou unha diferencia no efecto do PRS para individuos con e sen factores de risco clínicos.

Por último, establecemos un modelo combinando os factores de risco clínico e o PRS. A asociación do PRS mantívose incluso despois de axustar polo número de comorbilidades, sexo e idade, o que tamén foi observado por Crossfield e colaboradores (2022). A inclusión do PRS con respecto ao modelo que só incluía factores de risco clínico mellorou as predicións do modelo e reduciu o MSE, mais con limitacións. Dependendo do corte de probabilidade escollido para clasificar aos individuos en casos ou controis, o modelo clínico ofreceu valores de sensibilidade e especificidade similares, ou superiores, ao modelo que incorporaba tamén o índice xenético. O índice neto de reclasificación (NRI) foi negativo para un corte na probabilidade de 0.4 pero positivo para P=0.7, reflictindo este patrón. Non obstante, dentro do 20% con maior risco de hospitalización predito polo modelo completo, atópanse ao redor do 60% dos doentes hospitalizados, dos que o 20% eran mulleres, o 10% non presentaba ningunha comorbilidade e o 30% presentaba un risco xenético alto. Realmente, empregar un modelo clínico para predicir a hospitalización dun individuo pode non ter utilidade na práctica de triaxe. Pola contra, si pode resultar útil de cara a receitar un medicamento ou incluso determinar estratexias máis personalizadas de vacinación, permitindo a identificación de individuos que xa teñen un risco superior de base a padecer cadros de enfermidade máis graves. As curvas de análise de decisión revelan que o valor engadido polo PRS non é substancial con respecto a un modelo clínico básico, especialmente tendo en conta que na actualidade non se dispón da

USC UNIVERSIDADE DE SANTIAGO

información xenética dos individuos de forma sistemática, polo que incluír este biomarcador supón custos económicos adicionais. A utilidade clínica do PRS, tendo en conta estes resultados, non está clara. É importante ter en conta que esta análise está realizada nunha cohorte con sobre-representación dos casos e que, tras as vacinas, a prevalencia de hospitalización entre casos COVID+ é posiblemente máis baixa. Ademais, a cohorte test ten un tamaño de mostra moi pequeno. O óptimo sería re-analizar a curva de decisión nun estudo de cohortes, onde a recollida das mostras non estivera suxeita a presentar o fenotipo.

A pesar de que o PRS poida non compensar no seu uso translacional, ten utilidades en investigación, sobre todo despois de comprobar que supón un factor de risco independente a outros xa coñecidos. Por exemplo, pódese empregar para estudar a etioloxía compartida con outras enfermidades, ou para establecer relacións causais en estudos de aleatorización mendeliana. O estudo dun PRS de diabetes en relación á COVID-19 revelou que os individuos con maior risco xenético a padecela presentan taxas máis altas de mortalidade (A. Lee et al., 2024). De forma análoga, un PRS que construímos para esquizofrenia prediciu significativamente os fenotipos de infección, hospitalización por COVID-19, e enfermidade asintomática (Alemany-Navarro et al., 2023), salientando unha base xenética compartida entre ambas patoloxías. En forma de PheWAS, o PRS de COVID-19 asociouse a diversos antecedentes clínicos de relevancia para a enfermidade, como a hipercolesterolemia, o tabaquismo, enfermidades dixestivas ou mentais (Fritsche et al., 2023). Outras potenciais aplicacións poderían ser as de comprobar se o risco xenético influencia o risco de reinfección, a resposta a certos tratamentos, ou aos efectos adversos ás vacinas, ou incluso cuantificar o seu efecto na COVID persistente. Construír e validar un PRS, particularmente en poboacións non europeas, é de utilidade alén do seu uso na clínica.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

7 DISCUSIÓN XERAL

No seu clásico artigo sobre a diversidade xenética entre humanos, Lewontin (1972) preguntouse:

> I have tried to include what would appear to be a priori representatives of the range of human diversity. But how does one do that? (...) How many different European nationalities should be included as compared with how many African peoples or Indian tribes? [Intentei incluír o que a priori parecían representantes do rango da diversidade humana. Pero, como se fai iso? Cantas nacionalidades europeas diferentes deberían incluírse en comparación a cantos grupos africanos ou tribos indias?] (p. 384).

Cincuenta anos despois, non só segue vixente esta cuestión, senón que a representación da diversidade humana nos estudos xenómicos é limitada. A pesar de que esta problemática lle foi sinalada á comunidade científica hai máis de 10 anos (Bustamante et al., 2011), os esforzos para reducila non foron suficientes. Como se expuxo na Introdución, as poboacións europeas están sobre-representadas nos estudos de xenética humana (ver **Figura 4**), o que limita o entendemento da etioloxía das enfermidades e incrementa as desigualdades na traslación dos descubrimentos á práctica clínica (Popejoy & Fullerton, 2016). Incluso traballando con grandes biobancos como o UKBiobank, os investigadores seguen excluíndo aos individuos non europeos nas análises (Ben-Eghan et al., 2020). O *National Research Council* (US) propuxo en 2011 aproveitar os avances tecnolóxicos e o fluxo de datos xenómicos e moleculares para avanzar cara o que denominaron unha "nova taxonomía das enfermidades", a partir da cal se poderían adaptar as estratexias de prevención, diagnóstico e intervención aos pacientes ou grupos de pacientes (National Research Council, 2011). Esta aproximación coñécese agora como Medicina Personalizada de Precisión (MPP), e coa súa aplicación están moi relacionados os estudos de asociación e de predición xenética. Se ben a promesa da MPP é esperanzadora, necesítase que esta información xenética sexa *precisa* ("correcta") e que os modelos empregados sexan validados (Ashley, 2016). A falta de diversidade limita tanto a precisión como a personalización.

Un dos eixos centrais desta tese foi a comparativa entre dúas poboacións, motivada polo reclamo dunha maior diversidade nos GWAS, no estudo dunha enfermidade complexa como a COVID-19. Aínda que ao longo da mesma se tratou de facer fincapé nos beneficios asociadas a incorporar individuos con GIA non europea á investigación xenética, cómpre revisar, de forma máis xeral, as opcións e as limitacións -tanto metodolóxicas como interpretativas- que presentan estes estudos na actualidade.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

7.1.1 Deseño dos estudos GWAS

As razóns máis comúns dadas polos investigadores para excluír a estas poboacións das análises son as preocupacións por unha potencial confusión dos resultados, a falta de potencia estatística para detectar asociacións significativas e o seguimento do que se considera metodoloxía estándar (Ben-Eghan et al., 2020). Os modelos estatísticos clásicos empregados nos GWAS precisan de certa homoxeneidade nas mostras, ou polo menos un control da estrutura poboacional para evitar fenómenos de confusión. Así, xorden dúas posibles aproximacións metodolóxicas (Peterson et al., 2019): (A) análises específicas de poboación e combinación en meta-análises multi-poboacionais, ou (B) análise conxunta dos individuos.

O fluxo de traballo máis común na actualidade consiste nun GWAS en cada unha das poboacións de estudo e a combinación dos resultados (opción A). Ao realizar a análise de asociación en cada grupo poboacional, asegúrase unha corrección da estrutura poboacional e dos factores medioambientais. Ademais, como existe unha correlación moi alta dos efectos causais entre as poboacións, esta estratexia impulsa a potencia estatística e permite a detección de sinais compartidos (Y. R. Li & Keating, 2014; W. Zhou et al., 2022). Este foi, de feito, o deseño de estudo escollido polo consorcio HGI, baixo o cal se orientaron as análises desta tese. No primeiro estudo (cohorte española), procedemos a homoxeneizar a cohorte: descartáronse os individuos con <80% de GIA EUR e, posteriormente, os resultados incluíronse na meta-análise das poboacións EUR e nunha meta-análise multi-poboación. Isto facilitounos a aplicación dos métodos e a identificación de sinais, pero a costa dunha certa perda de datos. Aínda que unha parte das mostras se recuperaron para incorporalas á cohorte LatAm, aquelas con outras compoñentes GIA (ou mestura doutras compoñentes GIA) foron descartadas. En principio, as poboacións infra-representadas nos estudos GWAS, habitualmente con menores tamaños de mostra, vense beneficiadas pola meta-análise para a detección de asociacións compartidas. Á súa vez, é posible realizar unha exploración das análises individuais (as específicas de poboación) para detectar aquelas non compartidas. Ademais, este deseño non require de datos xenéticos crús, posibilitando a re-análise continua dos fenotipos a medida que se publican novas cohortes ou a creación de consorcios internacionais. No contexto da COVID-19, por exemplo, esta estratexia facilitou a colaboración de moitos grupos de investigación ao redor do mundo de forma rápida: iteráronse as análises ata 7 veces, liberando os resultados ao público en cada un dos ciclos.

O primeiro problema co que nos atopamos ao realizar análises en poboacións individuais é a división dos individuos en grupos discretos. Como vimos, cada vez máis investigadores pulan pola eliminación das categorías poboacionais discretas e da clasificación dos individuos. O descritor máis empregado nos estudos xenéticos é a "ancestralidade continental" (Dauda et al., 2023), que foi o escollido polo consorcio HGI, por AncestryDNA, 23andME, GenOMICC e SCOURGE para os GWAS de COVID-19. Os descritores poden variar en función do estudo e da cohorte, empregándose dende etnias (ou incluso "*raza auto-reportada*") ata rexións xeográficas (National Academies of Sciences, Engineering, and Medicine, 2023, Capítulo 2). É certo que, metodoloxica e estatisticamente, pode ser posible a asignación de individuos a clústeres discretos en tanto que é posible darlle a un algoritmo uns parámetros e esperar a que devolva o mellor modelo. Porén, corremos o risco de excluír a un número considerable de

individuos que poderían aportar riqueza ao estudo, como ocorre no caso das poboacións mixtas, ao non cumprir os criterios habituais (por exemplo, ter máis dun 80%-90% dunha certa compoñente de GIA). Esta dificultade atopámola no estudo da nosa poboación latinoamericana. Como se evidencia na **Figura 16**, calquera clúster que se intentara formar era necesariamente aleatorio. Aínda que non foi incluído na tese, nun inicio exploramos técnicas de clasificación como o modo sen supervisar de ADMIXTURE, *k-means*, e incluso modelos de mestura *gaussiana*. A selección dos hiperparámetros influía considerablemente nos resultados, de forma que existía solapamento entre os grupos inferidos, e non existía ningún criterio non arbitrario para a asignación. Un individuo podía pertencer ao clúster A se o número de clústeres a inferir eran 2, e pertencer ao clúster C se este parámetro se elevaba a 4. Ademais, os límites entre os grupos eran continuos. Estes individuos, ou ben son excluídos das meta-análises multi-poboación, ou forman o seu propio grupo. Cando se opta polo segundo, é probable que non queden claros os criterios de definición do grupo. Por exemplo, a análise AMR do consorcio HGI (agora denominada "HIS") pode incluír mostras con > 80% de GIA similar ás poboacións AMR de 1KGP ou mostras clasificadas como mixtas por presentar mestura das tres compoñentes ancestrais NAM, EUR e AFR (como no noso caso).

En segundo lugar, as meta-análises multi-poboación poden limitar a detección de variantes de risco con frecuencias alélicas máis baixas. Un dos pasos do control de calidade para este tipo de estudos consiste en eliminar SNPs ausentes nunha certa porcentaxe dos estudos incluídos (é dicir, que non se meta-analizaran en polo menos N estudos). Isto implica que, por exemplo, unha variante rara como a atopada polo noso grupo en *DDIAS* non se avalíe no marco dunha análise multi-poboación. Finalmente, como resaltan Ben-Eghan e colaboradores (2020), os GWAS de poboacións diversas e mixtas tenden a ter un número menor de individuos e menor potencia estatística, causando que se atopen un menor número de asociacións, ou ningunha. Isto, a miúdo, causa que non se publiquen os resultados, o que supón nunha oportunidade perdida para futuras meta-análises. Ademais, a estima dos efectos en cohortes pequenas -habitualmente non europeas- é menos precisa, polo que nalgúns casos estas cohortes teñen asociado un peso menor na análise. De feito, as meta-análises multi-poboación seguen a estar dominadas polas poboacións europeas, e os métodos clásicos para ponderar os estudos están influenciados, dunha ou outra forma, polo tamaño de mostra de cada estudo.

O método (B) consiste en analizar conxuntamente todos os individuos nun único GWAS. Esta aproximación adáptase ben ás poboacións mixtas e non require clasificar previamente aos individuos. Ademais, ao combinar diferentes poboacións, as variantes xenéticas con frecuencias baixas poden ser máis fáciles de identificar. Non obstante, é razoable que existan reservas sobre a aplicación deste método, xa que non controlar correctamente a estrutura poboacional pode ter como consecuencia a inflación dos p-valores e a aparición de falsos positivos. Os modelos mixtos permiten incluír na análise individuos relacionados por parentesco (Sul et al., 2018; Z. Yang et al., 2024; Yu et al., 2006), polo que é esperable que tamén consigan controlar as relacións producidas por ancestralidade. De feito, parece que combinar modelos mixtos coa corrección pola ancestralidade global é suficiente para abordar o problema da estratificación poboacional (K. Hou et al., 2021; Lin et al., 2021; Z. Yang et al., 2024), o que captura incluso efectos da compoñente medioambiental (Y. Zhang & Pan, 2015). Wojcik e colaboradores

(2019) constataron empiricamente que as aproximacións (A) e (B) son equivalentes cando na (B) se axusta para variables indicadoras da etnia ou do estudo (p.e., o "*país de recrutamento*"). Ademais, a opción (B) ofrece unha maior flexibilidade para controlar outras variables. No contexto dos grandes biobancos, pode ser máis interesante desenvolver nova metodoloxía de cara a analizar conxuntamente todos os individuos, e explotar o efecto dos factores medio-ambientais a nivel local ou incluso a estrutura fina poboacional (Belbin et al., 2021). O GWAS da cohorte LatAm enmárcase nesta estratexia, aínda que tamén se eliminaron individuos con alta proporción de GIA europea.

Con todo, esta aproximación non aborda todos os desafíos do modelo (A). Ao realizar unha análise conxunta dun grupo de individuos nos que predomina un grupo poboacional específico, existe o risco de eliminar variantes raras ou que non cumpran o equilibrio HWE durante a fase do control de calidade. Unha solución potencial podería ser levar a cabo estes procedementos en subgrupos de individuos xeneticamente similares, empregando a GIA, ou ben relaxar os límites de certos pasos. O proceso de imputación, por outra banda, pode ser menos preciso nos grupos poboacionais menos representados na mostra, aínda que os grandes paneis de imputación como o de TOPMed parecen ofrecer resultados satisfactorios para individuos non europeos (Taliun et al., 2021). Por outro lado, nos xenomas mixtos poden aparecer correlacións entre SNPs atopados a gran distancia debido aos procesos de mestura, o cal dificulta procesos como o mapeamento fino. Ademais, as diferentes estruturas de LD das poboacións da mostra e a compensación das frecuencias alélicas entre elas, poden dificultar que un sinal chegue a ser significativo e se requiran tamaños de mostra máis grandes.

Finalmente, é importante destacar brevemente as limitacións que existen nas análises posteriores ao GWAS en poboacións non europeas. Naquelas que empregan datos de expresión, como o TWAS ou a colocalización, o desafío principal é que existen moi poucos sets de datos de grupos non europeos, o que se traduce nunha perda de información e de potencia estatística. Polo momento, aquelas cohortes para as que si existen datos de expresión, como MESA ou GALA II-SAGE, están limitadas a tecidos sanguíneos. A modo de suxerencia, sería interesante probar unha aproximación metodolóxica similar á que usamos nós no estudo dos PRS para os modelos de predición da expresión (adestrar o modelo nun grupo de individuos achegados aos da mostra test empregando as PCs), que non está condicionado pola definición de poboacións ou etnias. Outro tipo de análises, como as correlacións xenéticas ou as estimas de herdabilidade, soen precisar de paneis de LD que se correspondan cos da mostra, os cales son máis difíciles de obter no contexto das meta-análises. Unha consecuencia disto é que os estudos post-GWAS acaban quedando restrinxidos ás cohortes europeas. No estudo do consorcio HGI, por exemplo, excluíronse 3 poboacións (unha delas, a AMR) dun dos pasos da priorización de xenes, e as análises de aleatorización mendeliana, correlacións xenéticas e estudos de herdabilidade foron realizadas unicamente con datos de poboacións europeas (COVID-19 Host Genetics Initiative et al., 2021). En cambio, o TWAS na nosa cohorte LatAm revelou novas asociacións de interese, a pesar de ter usado datos eQTL de GTEx.

7.1.2 Replicación de sinais entre poboacións

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Independentemente do deseño seleccionado, a replicación e xeneralización de sinais GWAS entre as poboacións é complexa, tanto en termos do tamaño do efecto como do p-valor. Un dos factores limitantes é a falta de potencia estatística nos estudos. Con frecuencia, non é posible dispoñer dunha cohorte de replicación adicional composta por individuos do mesmo grupo poboacional, cun tamaño suficientemente grande para validar os sinais (algo especialmente crítico para o caso das variantes máis raras) (Huffman, 2018). Un procedemento de replicación alternativo consiste en comparar os resultados das poboacións non europeas con outros xa publicados, ben no *GWAS Catalog* ou outros GWAS previos, nos que predomina a poboación europea. Por exemplo, as asociacións atopadas no GWAS da cohorte LatAm desta tese non puideron ser avaliadas nunha mostra externa, mais tampouco puideron ser replicadas nas meta-análises multi-poboación, probablemente debido á sobre-representación das cohortes europeas. En contraste, a replicación dos sinais atopados na cohorte española foi máis sinxela.

As diferencias no LD entre as poboacións poden xerar unha heteroxeneidade nos efectos estimados entre os diferentes grupos de estudo: unha variante pode estar fortemente correlacionada coa variante causal nunha poboación, pero non estalo noutra, o que resulta en asociacións diferentes e non replicables. Tomando o exemplo de *DDIAS*, na meta-análise do HGI para poboación africana, a variante con menor p-valor era unha variante rara situada aproximadamente a 20kb da nosa variante sentinela, pero que estaba ausente na nosa cohorte de LatAm. Respecto a *BAZ2B*, aínda que a variante líder estaba dispoñible para todas as poboacións, mostrou unha marcada heteroxeneidade entre as poboacións analizadas no HGI: a dirección do efecto só foi consistente coa poboación SAS. Porén, atopamos unha variante rara nunha ventá de 50kb que presentaba un p-valor de asociación relativamente baixo en poboación europea, cunha FA do 0.27% (estimada na meta-análise). É plausible que os diferentes patróns de LD sexan responsables da falta de consistencia entre poboacións para ambos xenes: as variantes en LD coa causal poderían diferir entre as poboacións, e as análises de AFR e EUR (para *DDIAS*/*BAZ2B*, respectivamente) non terían suficiente potencia estatística para acadar unha asociación significativa nestas variantes. Neste sentido, as meta-análisis multi-poboación permiten aproveitar os patróns de asociación das análises específicas de poboación para o mapeamento fino. Por exemplo, esta técnica axudou a identificar a variante de *splicing* implicada na COVID-19 do xen *OAS1* (Huffman et al., 2022).

As interaccións xenotipo-ambiente e xenotipo-xenotipo tamén poden dificultar a replicación. No primeiro estudo desta tese doutoral mencionamos un exemplo de interacción entre o ambiente e o xenotipo: algúns dos SNPs asociados á hospitalización por COVID-19 mostraron variacións na magnitude do efecto entre sexos e entre grupos de idade, particularmente para o SNP do xen *TLE1*. Esta última asociación, observada unicamente en mulleres, non foi replicada nin no grupo de homes nin no grupo completo. Aínda que no Capítulo 4 abordamos outras posibles explicacións para esta disparidade entre sexos, tamén podería ser consecuencia dunha interacción GxE verdadeira, dificultando a súa replicación. Do mesmo xeito que ocorre co sexo ou coa idade, outras variables ambientais poden provocar que unha asociación se manifeste unicamente baixo certas condicións, especialmente aquelas que son máis prevalentes na poboación estudada.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Smith e colaboradores (2022) exploraron a porcentaxe de SNPs que replicaban entre poboacións para 25 fenotipos: para 24 deles, só replicaban o 10% cun nivel de significación GWAS, pero esta porcentaxe aumentaba ao rebaixar o límite de p-valor para a replicación. En cambio, tras aplicar un método para estudar o enriquecemento a nivel de xen, si identificaron un maior número de xenes replicados. Estes resultados suxiren que a replicación de SNPs individuais pode non ser unha aproximación correcta.

7.1.3 Construción de PRS en cohortes diversas

O estudo dos factores de risco para a COVID-19 culminaron na predición xenética do fenotipo a través dos PRS, que é un puntos chave na MPP para as enfermidades complexas. Os PRS teñen numerosas potenciais aplicacións, tanto clínicas como para investigación, pero aínda queda traballo por facer para mellorar a súa capacidade preditiva.

Unha das cuestións que máis interese está xerando nos últimos anos é a transferabilidade dos PRS entre poboacións, que ademais atravesa de forma indirecta tamén os primeiros dous estudos da tese. Se o obxectivo é personalizar a prevención ou a intervención na clínica, non se pode empregar un factor de risco que presente un nesgo de base na predición. Este nesgo nace, como xa sabemos, nos propios estudos GWAS. As opcións actuais para desenvolver PRS en poboacións diversas son, ou ben construír PRS específicos de poboación -o que require a súa clasificación en grupos discretos-, ou construír un único PRS e validalo en diferentes grupos. A primeira opción presenta limitacións en presencia de poboacións mixtas, e a asignación dos individuos a poboacións continentais é unha aproximación metodolóxica que non ten en conta a estrutura fina da poboación, a cal tamén ten impacto na predición dos PRS. Por exemplo, para algúns fenotipos observouse certa correlación xeográfica dentro de poboacións consideradas homoxéneas (Abdellaoui et al., 2019; Kerminen et al., 2019; Sakaue et al., 2020). Ademais, empregar estimas de risco específicas de poboación nun contexto real pode ser pouco práctico, pois requiría estratificar aos individuos en grupos segundo a etnia ou a ancestralidade, o cal non é moi preciso. Isto tamén requiría da existencia de varios modelos de predición por fenotipo.

Gran parte do esforzo estase poñendo no desenvolvemento de novos métodos de construción de PRS, mais isto pode ser contraproducente: a maioría deles non se usan, e incrementar o número de métodos complica a instauración de protocolos estandarizados. No noso caso non puidemos explorar a construción de PRS empregando técnicas específicas para poboación mixta debido a que os *software* dispoñibles na actualidade, como GAUDI ou admix-kit (que calcula PRS parciais por compoñente ancestral), non están preparados para modelar a combinación de 3 fontes ancestrais (K. Hou, Gogarten, et al., 2023; Q. Sun et al., 2024). Ademais, a estima da ancestralidade local introduce un custe computacional adicional. Outro método diferente, iPGS (Tanigawa & Kellis, 2023), estima os efectos das variantes nos datos individuais (xenotipos) e require dunha mostra adicional para validar o PRS. Dado que a nosa cohorte latina non foi suficientemente grande como para detectar por si soa ningunha sinal GWAS, é probable que a estimación dos efectos para o PRS fora inestable, especialmente tendo en conta que para aplicar iPGS era necesario dividila en dúas (na mostra de adestramento e na de validación). Recentemente publicouse un novo método, slaPRS, pero de momento só foi

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

testado en poboación africana mixta, é dicir, con dúas compoñentes de AL (Liao & Zöllner, 2024).

Outras investigacións están estudando como calibrar as distribucións dos PRS entre poboacións para poder ofrecer unha OR única e independente da poboación ou ancestralidade. Hao e colaboradores (2022) optan por eliminar do PRS a contribución asociada ás PCs e empregar como limiar de risco o valor correspondente do PRS equivalente a unha OR > 2. Outra opción para fenotipos nos que a varianza está moi relacionada coa ancestralidade consiste en estender este método para modelar tanto a media como a varianza (Lennon et al., 2024). Substituír as PCs polas compoñentes de ancestralidade tamén obtivo bos resultados (Rosenthal et al., 2024). Derivar un único PRS para todas as poboacións é, entón, unha opción máis práctica (sempre e cando o PRS teña un rendemento similar entre elas), e esta estratexia obtivo bos resultados para fenotipos como a presión arterial, a diabetes, CAD ou cancro de próstata (F. Chen et al., 2022; Ge et al., 2022; Kurniansyah et al., 2023; Patel et al., 2023).

Nesta tese optamos por aplicar este marco, na nosa opinión máis pragmático, en vez dun centrado nas poboacións individuais. Para isto, comparamos dous métodos de construción de PRS que contan con unha evidencia sólida, un deles optimizado para casos multi-poboación (PRS-CSx), e estudamos o seu comportamento entre as poboacións. A continuación, quixemos estudar a posibilidade de derivar un PRS único que se puidese aplicar a individuos europeos e latinoamericanos, con miras a facilitar unha aplicación real deste PRS, tanto en clínica como en investigación. Para isto empregamos o método de calibración do PRS coas PCs, obtendo unha estima de risco aparentemente fiable para ambas poboacións. Ademais, tratamos de estender este concepto tamén ao modelo de predición en si: aproveitando a similitude xenética, é posible que teña máis sentido estimar o risco a partir dun grupo de individuos parecidos aos da mostra na que se quere realizar a predición. Isto, aínda que se propuxo para o conxunto dunha mostra test, podería aplicarse a cada paciente de forma individual. Cómpre aclarar que esta estratexia só será posible cando o PRS non presente unha marcada diferencia na súa capacidade preditiva entre poboacións, e tampouco se pode descartar que a calibración por PCs supoña unha perda de información xenética de relevancia.

Un modelo conxunto como o noso permitiría incorporar de forma máis flexible outras formas de variabilidade, como a local ou a ambiental. Máis alá dos grandes grupos poboacionais, a estrutura xenética fina é importante, se ben non recibe a consideración que merece no campo biomédico. Por exemplo, en México existe un gradiente entre Sonora e Yucatán pola compoñente Nativo-americana a nivel que está asociada coa función pulmonar (Moreno-Estrada et al., 2014). Dentro do UKBiobank realizáronse estudos de asociación empregando o lugar de nacemento como variable dependente e atoparon varios sinais significativos, incluso despois dunha corrección por 40 PCs, así como asociacións dos PRS coa xeografía (Haworth et al., 2019). O método empregado nesta tese (ou un máis sofisticado pero coas mesmas bases) permitiría variar as estimas de risco de cada enfermidade para grupos sub-poboacionais de forma máis dinámica e usando a similitude xenética na súa forma continua. Así, o método de calibración por PCs eliminaría os efectos máis grandes da ancestralidade, e as variacións na cohorte de adestramento modelarían os remanentes, sen o risco de realizar unha clasificación (previa) incorrecta dos individuos. Ademais, é máis probable que estes grupos

compartan un contexto ambiental e socioeconómico, e desta forma poderían variar tamén os coeficientes asociados a estas variables. Sería incluso interesante aumentar o número de PCs incluídos no cálculo da distancia euclídea para capturar niveis inferiores da estrutura poboacional. Estas hipóteses necesitan ser exploradas detalladamente nunha cohorte suficientemente grande e diversa, pero a nosa proposta é a de estudar a construción e axuste dos PRS nesta dirección. Isto, en combinación cun aumento da diversidade nos estudos GWAS, pode que solvente os problemas da predición á vez que permite acomodar a todo o conxunto de individuos. Ademais do que nós propoñemos aquí, existen outras oportunidades de mellora, como a combinación do PRS con variantes descubertas en poboacións específicas (Cavazos & Witte, 2021). Por exemplo, a predición da enfermidade crónica de ril mellorou en poboacións africanas e latinoamericanas ao combinar o PRS con variantes no xen *APOL1*, as cales confiren maior risco en individuos de GIA africana (Khan et al., 2022).

7.1.4 Aplicación clínica dos PRS

No terceiro estudo da tese estudamos o potencial que tiña un PRS de COVID-19 para ser implementado na práctica clínica, polo que queremos discutir algunhas cousas ao respecto. En primeiro lugar, cómpre ter en conta que os índices de risco xenético, na maioría dos casos, non poderán ser usados como un factor de risco único. Aínda que nos estudos de asociación, xa ben sexan de exoma ou xenoma, conseguíramos obter uns tamaños de mostra suficientes para estimar unha porcentaxe substancial da herdabilidade dos caracteres, quedaríanos por entender a contribución da varianza medioambiental. Porén, espérase que a partir de certo punto o descubrimento de sinais se sature (Visscher et al., 2017) e que ditos tamaños de mostra non se acaden para enfermidades menos prevalentes. Isto quere dicir que, para algunhas enfermidades ou fenotipos, a capacidade preditiva dun PRS estará limitada (de feito, pola herdabilidade).

Ata o momento, estes índices de risco estanse incluíndo en modelos de predición clínicos establecidos. Por exemplo, o coñecido algoritmo BOADICEA (*The Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm*) combina variantes de alta penetrancia, un PRS e diferentes variables clínicas e demográficas para predicir o risco de cancro de mama (A. Lee et al., 2019), e foi validado en cohortes europeas con bos resultados (Lakeman et al., 2020; Pal Choudhury et al., 2021). É dicir, a utilidade dun PRS ten que valorarse en combinación con factores de risco modificables. Ademais, a incorporación dun PRS a un modelo clínico leva asociados uns custes, que en ocasións non estarán compensados polos beneficios. Vimos que o noso PRS de hospitalización por COVID-19 pode non compensar con respecto a modelar unicamente o sexo, a idade e a presencia de comorbilidades, a pesar de estar significativamente asociado ao fenotipo. Se ben os custes económicos do xenotipado son moi baixos, tamén están os custes de desenvolver os modelos e os asociados á xeración e interpretación dos informes. Os PRS deberían considerarse como calquera outro biomarcador e deberían establecerse uns criterios que determinen se compensa economicamente con respecto á mellora na predición. A modo de exemplo, para as enfermidades arterosclerótica cardiovasculares realizouse unha análise de custos e beneficios e determinouse que compensaba incluír o PRS baixo tres escenarios: cando xa exista información xenética do paciente, cando vaian acompañados doutros PRS (como por exemplo

para cancro de mama), ou para guiar o tratamento con estatinas, recomendando o tratamento a aqueles no top 70% da distribución do PRS (Kiflen et al., 2022). Nun futuro, onde a secuenciación ou xenotipado dos pacientes sexa unha práctica habitual, os PRS poderán ser incorporados aos modelos de forma sistemática incluso cando o seu valor preditivo sexa baixo.

Con todo, o estudo desenvolvido aquí para o PRS da COVID-19 e só unha demostración dun fluxo de traballo que se pode seguir para valorar o potencial traslacional, pois é esperable que para predicir o risco de hospitalización dos pacientes se usen outro tipo de variables, como certos valores bioquímicos ou indicadores da capacidade respiratoria, ao estilo do modelo ISARIC 4C (R. K. Gupta et al., 2021). Non obstante, o PRS de hospitalización podería usarse como indicador dunha base xenética de susceptibilidade a unha enfermidade grave. Un uso potencial deste PRS podería ser o de re-definir os grupos preferentes para a vacinación, que na actualidade só contempla a individuos menores de 60 anos cando exista embarazo ou presenten comorbilidades. Por exemplo, en relación ao risco de reinfección, podería compararse a estratexia actual de vacinación cunha que tamén priorice a individuos menores de 60 anos sen comorbilidades pero cunha puntuación polixénica alta. As análises tipo DCA son unha ferramenta moi útil para explorar diferentes escenarios e poden ser incorporadas na rutina da validación dos PRS.

Para comunicar correctamente os resultados deste tipo de estudos é imprescindible formar profesionais en consello xenético, mais tamén en conceptos estatísticos e bioinformáticos básicos. O ideal é que exista unha correcta comunicación entre o persoal técnico, o profesional médico e o paciente, para poder transmitir as ideas de forma que as puntuacións xenéticas sexan entendidas como un factor de risco adicional e non como variables determinantes. Por exemplo, clasificar a un individuo como "alto risco xenético" pode non significar nada por si só, sen combinarse con factores non xenéticos, e pode xerar falsos positivos que provoquen unha intervención innecesaria (Adeyemo et al., 2021). Un estudo recente analizou de forma individual unha serie de PRS dispoñibles no *PGS Catalog* e concluíu que a súa capacidade para predicir enfermidades, estratificación e *screening* poboacional era nula, presentando moitos máis custos que beneficios (Hingorani et al., 2023). Porén, esta análise parte da idea equivocada de que os PRS poden usarse como ferramentas independentemente doutras variables. Outro estudo, moi interesante, sobre a comunicación dos resultados dos PRS ao público revelou que, se ben a opción máis apoiada polos entrevistados era saber en que percentil se atopaban, moitos entenderon o percentil como unha probabilidade absoluta de risco (A. C. F. Lewis, Perez, et al., 2022). Outros conceptos que xeraron confusión foron a prevalencia, as OR e incluso o limiar de clasificación en alto/baixo risco. Os autores resaltan o caso dun paciente que manifestaba que recibir os resultados en forma binaria lle facía entender que ía padecer a enfermidade (A. C. F. Lewis, Perez, et al., 2022). Neste sentido, quizais sexa mellor comezar a afastarse das clasificacións de risco binarias e, alternativamente, empregar valores de risco continuos obtidos nos modelos de predición (ver exemplo en 6.1.4.2). De feito, os médicos entrevistados no estudo preferían os resultados do PRS reportados en forma continua e non binaria (en risco/non risco), e estaban especialmente preocupados coa comunicación de resultados por grupos poboacionais ou de ancestralidade, o cal se solucionaría baixo a estratexia proposta nesta tese. Nos últimos anos publicáronse protocolos e propostas de traballo para explorar a aplicación dos

PRS na clínica que poden servir como modelo en futuros estudos (Hao et al., 2022; Lennon et al., 2024).

Finalmente e, de novo, é necesario incidir na urxencia de progresar a unha perspectiva máis inclusiva e holística da xenética e das enfermidades (M. McCarthy & Birney, 2021). Canto máis personalizada sexa a intervención clínica, maior será o efecto das variables socioeconómicas ou aquelas relacionadas co contexto, ambiente e estilo de vida dos individuos. Os grupos menos representados nos estudos xenéticos non percibirán os mesmos beneficios na súa saúde derivados da incorporación dos PRS que os individuos europeos. É máis, se a capacidade preditiva do factor de risco é mala ou nula, pode ter consecuencias negativas para o paciente (como por exemplo, infraestimar o risco dunha enfermidade). Por outra banda, dadas as desigualdades económicas e sociais que aínda existen na actualidade, é menos probable que un individuo que se atope nunha situación desfavorable teña a oportunidade ou o interese de ver incorporada a súa información xenética nos estudos de investigación ou na práctica clínica ("Precision Medicine Needs an Equity Agenda," 2021). Afortunadamente, na actualidade estanse conformando varios proxectos co obxectivo de caracterizar a diversidade xenética doutras poboacións e de fomentar a participación de grupos menos representados, como por exemplo o *Mexican Biobank* (Sohail et al., 2023) e o proxecto GLAD, *Genetics of Latin American Diversity* (Borda et al., 2023) en Latino-América; o *Chinese Pangenome Consortium* en China (Gao et al., 2023), ou o programa *All of Us* en EE.UU. ("The 'All of Us' Research Program," 2019). No contexto dos PRS, cómpre destacar o consorcio PRIMED (*Polygenic RIsk MEthods in Diverse Populations*, https://primedconsortium.org/).

En definitiva, como investigadores temos a responsabilidade de pular por unha representación correcta das comunidades que estudamos. Este compromiso será crucial na redución das desigualdades clínicas, asegurando que a MPP sexa unha medicina accesible para toda a poboación, mais tamén para lograr un entendemento completo dos mecanismos biolóxicos subxacentes ás enfermidades.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

8 CONCLUSIÓNS

1. O estudo GWAS de SCOURGE en poboación española replicou as asociacións xenéticas nas rexións dos xenes *LZTFL1* e *IFNAR2* coa enfermidade da COVID-19. Ademais, observamos unha diferencia na magnitude do efecto entre sexos e entre grupos de idade para algúns dos SNPs, e identificamos unha asociación no xen *TLE1* que só se manifesta nas mulleres.

2. A meta-análise coa cohorte CNIO revelou a implicación de dúas rexións novas, nos cromosomas 9 e 19. Os datos de expresión apuntaron a *AQP7* e *AQP3* como xenes potencialmente causais, cun aumento da súa expresión en tecido pulmonar e na mucosa do esófago. No cromosoma 19, destacou o xen *UPK1A*, expresado en tecido sanguíneo e no pulmón. O sinal na zona das acuaporinas foi replicada noutros estudos posteriores, conformando unha rexión con potencial para ser explorada en busca de dianas farmacolóxicas.

3. A meta-análise GWAS en poboación latinoamericana identificou dúas rexións asociadas á hospitalización por COVID-19, non atopadas previamente noutros grupos poboacionais. No cromosoma 2, destacou o xen *BAZ2B*, se ben os datos de expresión apuntaron tamén ao xen *LY75*. A segunda asociación atópase no cromosoma 11, no xen *DDIAS*, a través dunha variante moi rara en poboacións europeas. Este xen non se puido relacionar cunha función nesta tese, pero si foi vinculado noutros estudos á ruta de reparación de danos no ADN e a cancro de pulmón.

4. As análises TWAS revelaron unha expresión diferencial entre casos hospitalizados por COVID-19 e controis para diversos xenes de interese na etioloxía da enfermidade, incluíndo *IGFLR1*, *COX6B1*, *WNT6*, *CAMP* ou *SLC25A37*. A súa validación funcional podería ser de utilidade no desenvolvemento de novos enfoques terapéuticos.

5. As meta-análises multi-poboación cos datos do consorcio HGI, nas que se integraron as cohortes de SCOURGE, evidenciaron unha base xenética da COVID-19 compartida entre as poboacións. Non obstante, nesta tese demostramos como a análise de poboacións mixtas e non europeas enriquece os estudos xenéticos ao aproveitar as diferencias na arquitectura xenética.

6. Ao contrario do esperado, non atopamos que os PRS presentaran unha peor capacidade preditiva na cohorte latinoamericana que na cohorte española. Destacamos que o PRS coa

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

maior capacidade preditiva en ambas cohortes foi aquel construído a nivel xenómico con datos multi-poboación e o método de PRS-CSx.

7. A calibración da distribución do PRS facilitou a aplicación do mesmo a ambas cohortes. Dito PRS explica entre o 1.4% e o 2% da variabilidade total da gravidade da COVID-19.

8. O risco polixénico de susceptibilidade á hospitalización está asociado á mortalidade, á necesidade de ventilación mecánica, á presencia de infiltración pulmonar, á trombose pulmonar e á enfermidade asintomática. Non observamos diferencias no efecto entre sexos nin entre ausencia e presencia de comorbilidades, pero si entre grupos de idade, cun efecto máis forte en individuos menores de 65 anos.

9. Desenvolvemos unha estratexia metodolóxica para o adestramento dos PRS en cohortes mixtas, permitindo o axuste do PRS en función da similitude xenética dos individuos diana sen requirir dunha clasificación inicial dos mesmos en grupos discretos.

10. O PRS ofrece unha mellora limitada da predición da hospitalización con respecto á idade, o sexo e as comorbilidades. O PRS si ten unha utilidade en investigación e pode ser clinicamente relevante para outras variables relacionadas coa enfermidade.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

9 BIBLIOGRAFÍA

Abani, O., Abbas, A., Abbas, F., Abbas, J., Abbas, K., Abbas, M., Abbasi, S., Abbass, H., Abbott, A., Abbott, A., Abdallah, N., Abdelaziz, A., Abdelaziz, A., Abdelfattah, M., Abdelqader, B., Abdul, A., Abdul, B., Abdul, S., Rasheed, A. A., … Zuriaga-Alvaro, A. (2022). Baricitinib in patients admitted to hospital with COVID-19 (RECOVERY): A randomised, controlled, open-label, platform trial and updated meta-analysis. *The Lancet*, *400*(10349), 359–368. https://doi.org/10.1016/S0140-6736(22)01109-6

Abdellaoui, A., Hugh-Jones, D., Yengo, L., Kemper, K. E., Nivard, M. G., Veul, L., Holtz, Y., Zietsch, B. P., Frayling, T. M., Wray, N. R., Yang, J., Verweij, K. J. H., & Visscher, P. M. (2019). Genetic correlates of social stratification in Great Britain. *Nature Human Behaviour*, *3*(12), Article 12. https://doi.org/10.1038/s41562-019-0757-5

Abolfotouh, M. A., Musattat, A., Alanazi, M., Alghnam, S., & Bosaeed, M. (2022). Clinical characteristics and outcome of Covid-19 illness and predictors of in-hospital mortality in Saudi Arabia. *BMC Infectious Diseases*, *22*(1), 950. https://doi.org/10.1186/s12879-022-07945-8

Adeyemo, A., Balaconis, M. K., Darnes, D. R., Fatumo, S., Granados Moreno, P., Hodonsky, C. J., Inouye, M., Kanai, M., Kato, K., Knoppers, B. M., Lewis, A. C. F., Martin, A. R., McCarthy, M. I., Meyer, M. N., Okada, Y., Richards, J. B., Richter, L., Ripatti, S., Rotimi, C. N., … Polygenic Risk Score Task Force of the International Common Disease Alliance. (2021). Responsible use of polygenic risk scores in the clinic: Potential benefits, risks and gaps. *Nature Medicine*, *27*(11), Article 11. https://doi.org/10.1038/s41591-021-01549-6

Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., Mohammadi, P., Park, Y., Parsana, P., Segrè, A. V., Strober, B. J., Zappala, Z., Cummings, B. B., Gelfand, E. T., Hadley, K., Huang, K. H., Lek, M., Li, X., Nedzel, J. L., … Biospecimen Collection Source Site—NDRI. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), Article 7675. https://doi.org/10.1038/nature24277

Ahluwalia, T. S., Eliasen, A. U., Sevelsted, A., Pedersen, C.-E. T., Stokholm, J., Chawes, B., Bork-Jensen, J., Grarup, N., Pedersen, O., Hansen, T., Linneberg, A., Sharma, A., Weiss, S. T., Evans, M. D., Jackson, D. J., Morin, A., Krogfelt, K. A., Schjørring, S., Mortensen, P. B., … Bønnelykke, K. (2020). FUT2–ABO epistasis increases the risk of early childhood asthma and Streptococcus pneumoniae respiratory illnesses. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-19814-6

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, *16*(4), Article 4. https://doi.org/10.1038/nrg3891

Albiñana, C., Zhu, Z., Schork, A. J., Ingason, A., Aschard, H., Brikell, I., Bulik, C. M., Petersen, L. V., Agerbo, E., Grove, J., Nordentoft, M., Hougaard, D. M., Werge, T., Børglum, A. D., Mortensen, P. B., McGrath, J. J., Neale, B. M., Privé, F., & Vilhjálmsson, B. J. (2023). Multi-PGS enhances polygenic prediction by combining 937 polygenic scores. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-40330-w

Alemany-Navarro, M., Diz-de Almeida, S., Cruz, R., Riancho, J. A., Rojas-Martínez, A., Lapunzina, P., Flores, C., & Carracedo, A. (2023). Psychiatric polygenic risk as a predictor of

USC UNIVERSIDADE DE SANTIAGO

COVID-19 risk and severity: Insight into the genetic overlap between schizophrenia and COVID-19. *Translational Psychiatry*, *13*(1), Article 1. https://doi.org/10.1038/s41398-023-02482-7

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Amnistía Internacional, & Center for Economic and Social Rights. (2022). *DESIGUAL Y LETAL* (AMR 01/5483/2022).

Asano, T., Boisson, B., Onodi, F., Matuozzo, D., Moncada-Velez, M., Maglorius Renkilaraj, M. R. L., Zhang, P., Meertens, L., Bolze, A., Materna, M., Korniotis, S., Gervais, A., Talouarn, E., Bigio, B., Seeleuthner, Y., Bilguvar, K., Zhang, Y., Neehus, A.-L., Ogishi, M., … Rossi, C. (2021). X-linked recessive TLR7 deficiency in ~1% of men under 60 years old with life-threatening COVID-19. *Science Immunology*, *6*(62), eabl4348. https://doi.org/10.1126/sciimmunol.abl4348

Ashley, E. A. (2016). Towards precision medicine. *Nature Reviews Genetics*, *17*(9), Article 9. https://doi.org/10.1038/nrg.2016.86

Astle, W., & Balding, D. J. (2009). Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science*, *24*(4), 451–471.

Atkinson, E. G., Maihofer, A. X., Kanai, M., Martin, A. R., Karczewski, K. J., Santoro, M. L., Ulirsch, J. C., Kamatani, Y., Okada, Y., Finucane, H. K., Koenen, K. C., Nievergelt, C. M., Daly, M. J., & Neale, B. M. (2021). Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nature Genetics*, *53*(2), Article 2. https://doi.org/10.1038/s41588-020-00766-y

Auernhammer, C. J., & Strasburger, C. J. (1995). Effects of growth hormone and insulin-like growth factor I on the immune system. *European Journal of Endocrinology*, *133*(6), 635–645. https://doi.org/10.1530/eje.0.1330635

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … National Eye Institute, N. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), Article 7571. https://doi.org/10.1038/nature15393

Backman, J. D., Li, A. H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M. D., Benner, C., Liu, D., Locke, A. E., Balasubramanian, S., Yadav, A., Banerjee, N., Gillies, C. E., Damask, A., Liu, S., Bai, X., Hawes, A., Maxwell, E., Gurski, L., … Ferreira, M. A. R. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*, *599*(7886), 628–634. https://doi.org/10.1038/s41586-021-04103-z

Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D., Spector, S. A., Rouphael, N., Creech, C. B., McGettigan, J., Khetan, S., Segall, N., Solis, J., Brosz, A., Fierro, C., Schwartz, H., Neuzil, K., Corey, L., … Zaks, T. (2021). Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine. *New England Journal of Medicine*, *384*(5), 403–416. https://doi.org/10.1056/NEJMoa2035389

Banday, A. R., Stanifer, M. L., Florez-Vargas, O., Onabajo, O. O., Papenberg, B. W., Zahoor, M. A., Mirabello, L., Ring, T. J., Lee, C.-H., Albert, P. S., Andreakos, E., Arons, E., Barsh, G., Biesecker, L. G., Boyle, D. L., Brahier, M. S., Burnett-Hartman, A., Carrington, M., Chang, E., … Prokunina-Olsson, L. (2022). Genetic regulation of OAS1 nonsense-mediated decay underlies

association with COVID-19 hospitalization in patients of European and African ancestries. *Nature Genetics*, *54*(8), Article 8. https://doi.org/10.1038/s41588-022-01113-z

Barbeira, A. N., Bonazzola, R., Gamazon, E. R., Liang, Y., Park, Y., Kim-Hellmuth, S., Wang, G., Jiang, Z., Zhou, D., Hormozdiari, F., Liu, B., Rao, A., Hamel, A. R., Pividori, M. D., Aguet, F., Bastarache, L., Jordan, D. M., Verbanck, M., Do, R., … GTEx Consortium. (2021). Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biology*, *22*(1), 49. https://doi.org/10.1186/s13059-020-02252-4

Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., Torstenson, E. S., Shah, K. P., Garcia, T., Edwards, T. L., Stahl, E. A., Huckins, L. M., Nicolae, D. L., Cox, N. J., & Im, H. K. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-03621-1

Barbeira, A. N., Pividori, M., Zheng, J., Wheeler, H. E., Nicolae, D. L., & Im, H. K. (2019). Integrating predicted transcriptome from multiple tissues improves association detection. *PLOS Genetics*, *15*(1), e1007889. https://doi.org/10.1371/journal.pgen.1007889

Bastarache, L., Denny, J. C., & Roden, D. M. (2022). Phenome-Wide Association Studies. *JAMA*, *327*(1), 75–76. https://doi.org/10.1001/jama.2021.20356

Bastard, P., Hsiao, K.-C., Zhang, Q., Choin, J., Best, E., Chen, J., Gervais, A., Bizien, L., Materna, M., Harmant, C., Roux, M., Hawley, N. L., Weeks, D. E., McGarvey, S. T., Sandoval, K., Barberena-Jonas, C., Quinto-Cortés, C. D., Hagelberg, E., Mentzer, A. J., … Casanova, J.-L. (2022). A loss-of-function IFNAR1 allele in Polynesia underlies severe viral diseases in homozygotes. *Journal of Experimental Medicine*, *219*(6), e20220028. https://doi.org/10.1084/jem.20220028

Bayati, A., Kumar, R., Francis, V., & McPherson, P. S. (2021). SARS-CoV-2 infects cells after viral entry via clathrin-mediated endocytosis. *The Journal of Biological Chemistry*, *296*, 100306. https://doi.org/10.1016/j.jbc.2021.100306

Beaulieu, A. M., Zawislak, C. L., Nakayama, T., & Sun, J. C. (2014). The transcription factor Zbtb32 controls the proliferative burst of virus-specific natural killer cells responding to infection. *Nature Immunology*, *15*(6), 546–553. https://doi.org/10.1038/ni.2876

Belbin, G. M., Cullina, S., Wenric, S., Soper, E. R., Glicksberg, B. S., Torre, D., Moscati, A., Wojcik, G. L., Shemirani, R., Beckmann, N. D., Cohain, A., Sorokin, E. P., Park, D. S., Ambite, J.-L., Ellis, S., Auton, A., Bottinger, E. P., Cho, J. H., Loos, R. J. F., … Kenny, E. E. (2021). Toward a fine-scale population health monitoring system. *Cell*, *184*(8), 2068-2083.e11. https://doi.org/10.1016/j.cell.2021.03.034

Bellenguez, C., Küçükali, F., Jansen, I. E., Kleineidam, L., Moreno-Grau, S., Amin, N., Naj, A. C., Campos-Martin, R., Grenier-Boley, B., Andrade, V., Holmans, P. A., Boland, A., Damotte, V., van der Lee, S. J., Costa, M. R., Kuulasmaa, T., Yang, Q., de Rojas, I., Bis, J. C., … Lambert, J.-C. (2022). New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature Genetics*, *54*(4), Article 4. https://doi.org/10.1038/s41588-022-01024-z

Ben-Eghan, C., Sun, R., Hleap, J. S., Diaz-Papkovich, A., Munter, H. M., Grant, A. V., Dupras, C., & Gravel, S. (2020). Don't ignore genetic data from minority populations. *Nature*, *585*(7824), 184–186. https://doi.org/10.1038/d41586-020-02547-3

Berenguer, J., Ryan, P., Rodríguez-Baño, J., Jarrín, I., Carratalà, J., Pachón, J., Yllescas, M., Arriba, J. R., Muñoz, E. A., Divasson, P. G., Muñiz, P. G., Aguirre, C. M., López, J. C., Ramírez-

Schacke, M., Gutiérrez, I., Tejerina, F., Aldámiz-Echevarría, T., Díez, C., Fanciulli, C., … Hernando, V. (2020). Characteristics and predictors of death among 4035 consecutively hospitalized patients with COVID-19 in Spain. *Clinical Microbiology and Infection*, *26*(11), 1525–1536. https://doi.org/10.1016/j.cmi.2020.07.024

Bi, W., Zhou, W., Dey, R., Mukherjee, B., Sampson, J. N., & Lee, S. (2021). Efficient mixed model approach for large-scale genome-wide association studies of ordinal categorical phenotypes. *The American Journal of Human Genetics*, *108*(5), 825–839. https://doi.org/10.1016/j.ajhg.2021.03.019

Bigdeli, T. B., Voloudakis, G., Barr, P. B., Gorman, B. R., Genovese, G., Peterson, R. E., Burstein, D. E., Velicu, V. I., Li, Y., Gupta, R., Mattheisen, M., Tomasi, S., Rajeevan, N., Sayward, F., Radhakrishnan, K., Natarajan, S., Malhotra, A. K., Shi, Y., Zhao, H., … Cooperative Studies Program (CSP) #572 and Million Veteran Program (MVP). (2022). Penetrance and Pleiotropy of Polygenic Risk Scores for Schizophrenia, Bipolar Disorder, and Depression Among Adults in the US Veterans Affairs Health Care System. *JAMA Psychiatry*, *79*(11), 1092–1101. https://doi.org/10.1001/jamapsychiatry.2022.2742

Bitarello, B. D., & Mathieson, I. (2020). Polygenic Scores for Height in Admixed Populations. *G3 Genes|Genomes|Genetics*, *10*(11), 4027–4036. https://doi.org/10.1534/g3.120.401658

Bland, J. M., & Altman, D. G. (1998). Bayesians and frequentists. *BMJ : British Medical Journal*, *317*(7166), 1151–1160.

Bolstad, W. M., & Curran, J. M. (2016). *Introduction to Bayesian Statistics*. John Wiley & Sons.

Borda, V., Loesch, D. P., Guo, B., Laboulaye, R., Veliz-Otani, D., French-Kwawu, J. N., Leal, T. P., Gogarten, S. M., Ikpe, S., Gouveia, M. H., Mendes, M., Abecasis, G. R., Alvim, I., Arboleda-Bustos, C. E., Arboleda, G., Arboleda, H., Barreto, M. L., Barwick, L., Bezzera, M. A., … O'Connor, T. D. (2023). *Genetics of Latin American Diversity (GLAD) Project: Insights into population genetics and association studies in recently admixed groups in the Americas* (p. 2023.01.07.522490). bioRxiv. https://doi.org/10.1101/2023.01.07.522490

Borella, R., De Biasi, S., Paolini, A., Boraldi, F., Lo Tartaro, D., Mattioli, M., Fidanza, L., Neroni, A., Caro-Maldonado, A., Meschiari, M., Franceschini, E., Quaglino, D., Guaraldi, G., Bertoldi, C., Sita, M., Busani, S., Girardis, M., Mussini, C., Cossarizza, A., & Gibellini, L. (2022). Metabolic reprograming shapes neutrophil functions in severe COVID-19. *European Journal of Immunology*, *52*(3), 484–502. https://doi.org/10.1002/eji.202149481

Boughton, A. P., Welch, R. P., Flickinger, M., VandeHaar, P., Taliun, D., Abecasis, G. R., & Boehnke, M. (2021). LocusZoom.js: Interactive and embeddable visualization of genetic association study results. *Bioinformatics*, *37*(18), 3017–3018. https://doi.org/10.1093/bioinformatics/btab186

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell*, *169*(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Braudt, D. B. (2018). Sociogenomics in the 21st Century: An Introduction to the History and Potential of Genetically-informed Social Science. *Sociology compass*, *12*(10), e12626. https://doi.org/10.1111/soc4.12626

Brunette, G. J., Jamalruddin, M. A., Baldock, R. A., Clark, N. L., & Bernstein, K. A. (2019). Evolution-based screening enables genome-wide prioritization and discovery of DNA repair genes. *Proceedings of the National Academy of Sciences*, *116*(39), 19593–19599. https://doi.org/10.1073/pnas.1906559116

Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C. D., & Ostrer, H. (2010). Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences*, *107*(supplement_2), 8954–8961. https://doi.org/10.1073/pnas.0914618107

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R. B., Patterson, N., Robinson, E. B., Daly, M. J., Price, A. L., & Neale, B. M. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, *47*(11), Article 11. https://doi.org/10.1038/ng.3406

Bustamante, C. D., De La Vega, F. M., & Burchard, E. G. (2011). Genomics for the world. *Nature*, *475*(7355), Article 7355. https://doi.org/10.1038/475163a

Byun, J., Han, Y., Li, Y., Xia, J., Long, E., Choi, J., Xiao, X., Zhu, M., Zhou, W., Sun, R., Bossé, Y., Song, Z., Schwartz, A., Lusk, C., Rafnar, T., Stefansson, K., Zhang, T., Zhao, W., Pettit, R. W., … Amos, C. I. (2022). Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer. *Nature Genetics*, *54*(8), Article 8. https://doi.org/10.1038/s41588-022-01115-x

Cai, M., Xiao, J., Zhang, S., Wan, X., Zhao, H., Chen, G., & Yang, C. (2021). A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. *The American Journal of Human Genetics*, *108*(4), 632–655. https://doi.org/10.1016/j.ajhg.2021.03.002

Carlson, C. S., Matise, T. C., North, K. E., Haiman, C. A., Fesinmeyer, M. D., Buyske, S., Schumacher, F. R., Peters, U., Franceschini, N., Ritchie, M. D., Duggan, D. J., Spencer, K. L., Dumitrescu, L., Eaton, C. B., Thomas, F., Young, A., Carty, C., Heiss, G., Marchand, L. L., … Consortium, for the P. (2013). Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. *PLOS Biology*, *11*(9), e1001661. https://doi.org/10.1371/journal.pbio.1001661

Carlson, J., Henn, B. M., Al-Hindi, D. R., & Ramachandran, S. (2022). Counter the weaponization of genetics research by extremists. *Nature*, *610*(7932), 444–447. https://doi.org/10.1038/d41586-022-03252-z

Casas-Rojo, J. M., Antón-Santos, J. M., Millán-Núñez-Cortés, J., Lumbreras-Bermejo, C., Ramos-Rincón, J. M., Roy-Vallejo, E., Artero-Mora, A., Arnalich-Fernández, F., García-Bruñén, J. M., Vargas-Núñez, J. A., Freire-Castro, S. J., Manzano-Espinosa, L., Perales-Fraile, I., Crestelo-Viéitez, A., Puchades-Gimeno, F., Rodilla-Sala, E., Solís-Marquínez, M. N., Bonet-Tur, D., Fidalgo-Moreno, M. P., … Gómez-Huelgas, R. (2020a). Características clínicas de los pacientes hospitalizados con COVID-19 en España: Resultados del Registro SEMI-COVID-19. *Revista Clínica Española*, *220*(8), 480–494. https://doi.org/10.1016/j.rce.2020.07.003

Casas-Rojo, J. M., Antón-Santos, J. M., Millán-Núñez-Cortés, J., Lumbreras-Bermejo, C., Ramos-Rincón, J. M., Roy-Vallejo, E., Artero-Mora, A., Arnalich-Fernández, F., García-Bruñén, J. M., Vargas-Núñez, J. A., Freire-Castro, S. J., Manzano-Espinosa, L., Perales-Fraile, I., Crestelo-Viéitez, A., Puchades-Gimeno, F., Rodilla-Sala, E., Solís-Marquínez, M. N., Bonet-Tur, D., Fidalgo-Moreno, M. P., … Gómez-Huelgas, R. (2020b). Clinical characteristics of patients hospitalized with COVID-19 in Spain: Results from the SEMI-COVID-19 Registry. *Revista Clinica Espanola*, *220*(8), 480–494. https://doi.org/10.1016/j.rceng.2020.07.003

Cascella, M., Rajnik, M., Aleem, A., Dulebohn, S. C., & Di Napoli, R. (2024). Features, Evaluation, and Treatment of Coronavirus (COVID-19). In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK554776/

Cavazos, T. B., & Witte, J. S. (2021). Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Human Genetics and Genomics Advances*, *2*(1), 100017. https://doi.org/10.1016/j.xhgg.2020.100017

Centro de Coordinación de Alertas y Emergencias Sanitarias. (2020). *Actualización nº 230. Enfermedad por el coronavirus (COVID-19)*. (230). Ministerio de Sanidad. Gobierno de España. https://www.lamoncloa.gob.es/serviciosdeprensa/notasprensa/sanidad14/Documents/2020/161020_Actualizacion230COVID19.pdf

Centro de Coordinación de Alertas y Emergencias Sanitarias. (2021). *Actualización nº 351. Enfermedad por el coronavirus (COVID-19)*. (351). Ministerio de Sanidad. Gobierno de España. https://www.lamoncloa.gob.es/serviciosdeprensa/notasprensa/sanidad14/Documents/2021/120421_Actualizacion_351_COVID-19.pdf

Chan, S. H., Bylstra, Y., Teo, J. X., Kuan, J. L., Bertin, N., Gonzalez-Porta, M., Hebrard, M., Tirado-Magallanes, R., Tan, J. H. J., Jeyakani, J., Li, Z., Chai, J. F., Chong, Y. S., Davila, S., Goh, L. L., Lee, E. S., Wong, E., Wong, T. Y., Prabhakar, S., … Lim, W. K. (2022). Analysis of clinically relevant variants from ancestrally diverse Asian genomes. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-34116-9

Chande, A. T., Rishishwar, L., Conley, A. B., Valderrama-Aguirre, A., Medina-Rivas, M. A., & Jordan, I. K. (2020). Ancestry effects on type 2 diabetes genetic risk inference in Hispanic/Latino populations. *BMC Medical Genetics*, *21*(2), 132. https://doi.org/10.1186/s12881-020-01068-0

Chang, Y.-C., Lee, H.-L., Yang, W., Hsieh, M.-L., Liu, C.-C., Lee, T.-Y., Huang, J.-Y., Nong, J.-Y., Li, F.-A., Chuang, H.-L., Ding, Z.-Z., Su, W.-L., Chueh, L.-Y., Tsai, Y.-T., Chen, C.-H., Mochly-Rosen, D., & Chuang, L.-M. (2023). A common East-Asian ALDH2 mutation causes metabolic disorders and the therapeutic effect of ALDH2 activators. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-41570-6

Channappanavar, R., Fett, C., Mack, M., Ten Eyck, P. P., Meyerholz, D. K., & Perlman, S. (2017). Sex-Based Differences in Susceptibility to Severe Acute Respiratory Syndrome Coronavirus Infection. *The Journal of Immunology*, *198*(10), 4046–4053. https://doi.org/10.4049/jimmunol.1601896

Chatterjee, N., Shi, J., & García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, *17*(7), Article 7. https://doi.org/10.1038/nrg.2016.27

Chen, F., Darst, B. F., Madduri, R. K., Rodriguez, A. A., Sheng, X., Rentsch, C. T., Andrews, C., Tang, W., Kibel, A. S., Plym, A., Cho, K., Jalloh, M., Gueye, S. M., Niang, L., Ogunbiyi, O. J., Popoola, O., Adebiyi, A. O., Aisuodionoe-Shadrach, O. I., Ajibola, H. O., … Haiman, C. A. (2022). Validation of a multi-ancestry polygenic risk score and age-specific risks of prostate cancer: A meta-analysis within diverse populations. *eLife*, *11*, e78304. https://doi.org/10.7554/eLife.78304

Chen, M.-H., Raffield, L. M., Mousas, A., Sakaue, S., Huffman, J. E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., Bao, E. L., Zhong, X., Manansala, R., Laplante, V., Chen, M., Lo, K. S., Qian, H., Lareau, C. A., Beaudoin, M., … Lettre, G. (2020). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell*, *182*(5), 1198-1213.e14. https://doi.org/10.1016/j.cell.2020.06.045

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., Xia, J., Yu, T., Zhang, X., & Zhang, L. (2020). Epidemiological and clinical characteristics of 99 cases

of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *The Lancet*, *395*(10223), 507–513. https://doi.org/10.1016/S0140-6736(20)30211-7

Choi, E. Y., Park, H. H., Kim, H., Kim, H. N., Kim, I., Jeon, S., Kim, W., Bae, J.-S., & Lee, W. (2020). Wnt5a and Wnt11 as acute respiratory distress syndrome biomarkers for severe acute respiratory syndrome coronavirus 2 patients. *European Respiratory Journal*, *56*(5). https://doi.org/10.1183/13993003.01531-2020

Choi, S. W., García-González, J., Ruan, Y., Wu, H. M., Porras, C., Johnson, J., Consortium, B. D. W. group of the P. G., Hoggart, C. J., & O'Reilly, P. F. (2023). PRSet: Pathway-based polygenic risk score analyses and software. *PLOS Genetics*, *19*(2), e1010624. https://doi.org/10.1371/journal.pgen.1010624

Choi, S. W., Mak, T. S. H., & O'Reilly, P. F. (2020). A guide to performing Polygenic Risk Score analyses. *Nature protocols*, *15*(9), 2759–2772. https://doi.org/10.1038/s41596-020-0353-1

Choi, U. Y., Kang, J.-S., Hwang, Y. S., & Kim, Y.-J. (2015). Oligoadenylate synthase-like (OASL) proteins: Dual functions and associations with diseases. *Experimental & Molecular Medicine*, *47*(3), Article 3. https://doi.org/10.1038/emm.2014.110

Chou, V. (2017, April 17). How Science and Genetics are Reshaping the Race Debate of the 21st Century [Harvard University]. *Science in the News*. https://sitn.hms.harvard.edu/flash/2017/science-genetics-reshaping-race-debate-21st-century/

Clarke, S. L., Huang, R. D. L., Hilliard, A. T., Tcheandjieu, C., Lynch, J., Damrauer, S. M., Chang, K.-M., Tsao, P. S., & Assimes, T. L. (2022). Race and Ethnicity Stratification for Polygenic Risk Score Analyses May Mask Disparities in Hispanics. *Circulation*, *146*(3), 265–267. https://doi.org/10.1161/CIRCULATIONAHA.122.059162

Clifton, L., Collister, J. A., Liu, X., Littlejohns, T. J., & Hunter, D. J. (2022). Assessing agreement between different polygenic risk scores in the UK Biobank. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-17012-6

Cohen, J., Pertsemlidis, A., Kotowski, I. K., Graham, R., Garcia, C. K., & Hobbs, H. H. (2005). Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nature Genetics*, *37*(2), Article 2. https://doi.org/10.1038/ng1509

Colistro, V., Rojas-Martínez, A., Carracedo, A., Consortium, C., Tomlinson, I., Carvajal-Carmona, L., Cruz, R., & Sans, M. (2021). Population structure and relatedness estimates in a Mexican sample. *Annals of Human Genetics*, *85*(6), 245–248. https://doi.org/10.1111/ahg.12421

Collins, G. S., Dhiman, P., Ma, J., Schlussel, M. M., Archer, L., Van Calster, B., Harrell, F. E., Martin, G. P., Moons, K. G. M., Van Smeden, M., Sperrin, M., Bullock, G. S., & Riley, R. D. (2024). Evaluation of clinical prediction models (part 1): From development to external validation. *BMJ*, e074819. https://doi.org/10.1136/bmj-2023-074819

Collister, J. A., Liu, X., & Clifton, L. (2022). Calculating Polygenic Risk Scores (PRS) in UK Biobank: A Practical Guide for Epidemiologists. *Frontiers in Genetics*, *13*. https://www.frontiersin.org/articles/10.3389/fgene.2022.818574

Comfort, N. (2018, October 23). *Sociogenomics is opening a new door to eugenics*. MIT Technology Review. https://www.technologyreview.com/2018/10/23/139420/sociogenomics-is-opening-a-new-door-to-eugenics/

Cooling, L. (2015). Blood Groups in Infection and Host Susceptibility. *Clinical Microbiology Reviews*, *28*(3), 801–870. https://doi.org/10.1128/CMR.00109-14

USC DE COMPOSTELA

Coop, G. (2023). *Genetic similarity versus genetic ancestry groups as sample descriptors in human genetics* (arXiv:2207.11595). arXiv. https://doi.org/10.48550/arXiv.2207.11595

Coop, G., & Przeworski, M. (2022). Luck, lottery, or legacy? The problem of confounding. A reply to Harden. *Evolution; International Journal of Organic Evolution*, *76*(10), 2464–2468. https://doi.org/10.1111/evo.14588

Cortez, D., Wang, Y., Qin, J., & Elledge, S. J. (1999). Requirement of ATM-dependent phosphorylation of brca1 in the DNA damage response to double-strand breaks. *Science (New York, N.Y.)*, *286*(5442), 1162–1166. https://doi.org/10.1126/science.286.5442.1162

COVID-19 Host Genetics Initiative, COVID-19 Host Genetics InitiativeLeadership, Niemi, M. E. K., Karjalainen, J., Liao, R. G., Neale, B. M., Daly, M., Ganna, A., Writing group, Writing group leaders, Pathak, G. A., Andrews, S. J., Kanai, M., Writing group members, Veerapen, K., Fernandez-Cadenas, I., Schulte, E. C., Striano, P., Marttila, M., … COVID-19 HGI corresponding authors. (2021). Mapping the human genetic architecture of COVID-19. *Nature*, *600*(7889), 472–477. https://doi.org/10.1038/s41586-021-03767-x

Crossfield, S. S. R., Chaddock, N. J. M., Iles, M. M., Pujades-Rodriguez, M., & Morgan, A. W. (2022). Interplay between demographic, clinical and polygenic risk factors for severe COVID-19. *International Journal of Epidemiology*, *51*(5), 1384–1395. https://doi.org/10.1093/ije/dyac137

Cuschieri, S. (2019). The STROBE guidelines. *Saudi Journal of Anaesthesia*, *13*(Suppl 1), S31–S34. https://doi.org/10.4103/sja.SJA_543_18

Dahl, A., Thompson, M., An, U., Krebs, M., Appadurai, V., Border, R., Bacanu, S.-A., Werge, T., Flint, J., Schork, A. J., Sankararaman, S., Kendler, K. S., & Cai, N. (2023). Phenotype integration improves power and preserves specificity in biobank-based genetic studies of major depressive disorder. *Nature Genetics*, *55*(12), Article 12. https://doi.org/10.1038/s41588-023-01559-9

Das, S., Abecasis, G. R., & Browning, B. L. (2018). Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics*, *19*(1), 73–96. https://doi.org/10.1146/annurev-genom-083117-021602

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), Article 10. https://doi.org/10.1038/ng.3656

Dauda, B., Molina, S. J., Allen, D. S., Fuentes, A., Ghosh, N., Mauro, M., Neale, B. M., Panofsky, A., Sohail, M., Zhang, S. R., & Lewis, A. C. F. (2023). Ancestry: How researchers use it and what they mean by it. *Frontiers in Genetics*, *14*. https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2023.1044555

De Jong, T. V., Guryev, V., & Moshkin, Y. M. (2021). Estimates of gene ensemble noise highlight critical pathways and predict disease severity in H1N1, COVID-19 and mortality in sepsis patients. *Scientific Reports*, *11*(1), 10793. https://doi.org/10.1038/s41598-021-90192-9

De La Vega, F. M., & Bustamante, C. D. (2018). Polygenic risk scores: A biased prediction? *Genome Medicine*, *10*(1), 100. https://doi.org/10.1186/s13073-018-0610-x

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

de Leeuw, C. A., Mooij, J. M., Heskes, T., & Posthuma, D. (2015). MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Computational Biology*, *11*(4), e1004219. https://doi.org/10.1371/journal.pcbi.1004219

de Rojas, I., Moreno-Grau, S., Tesi, N., Grenier-Boley, B., Andrade, V., Jansen, I. E., Pedersen, N. L., Stringa, N., Zettergren, A., Hernández, I., Montrreal, L., Antúnez, C., Antonell, A., Tankard, R. M., Bis, J. C., Sims, R., Bellenguez, C., Quintela, I., González-Perez, A., … Ruiz, A. (2021). Common variants in Alzheimer's disease and risk stratification by polygenic risk scores. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-22491-8

de Souza, W. M., Buss, L. F., Candido, D. da S., Carrera, J.-P., Li, S., Zarebski, A. E., Pereira, R. H. M., Prete, C. A., de Souza-Santos, A. A., Parag, K. V., Belotti, M. C. T. D., Vincenti-Gonzalez, M. F., Messina, J., da Silva Sales, F. C., Andrade, P. dos S., Nascimento, V. H., Ghilardi, F., Abade, L., Gutierrez, B., … Faria, N. R. (2020). Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. *Nature Human Behaviour*, *4*(8), Article 8. https://doi.org/10.1038/s41562-020-0928-4

Degenhardt, F., Ellinghaus, D., Juzenas, S., Lerga-Jaso, J., Wendorff, M., Maya-Miles, D., Uellendahl-Werth, F., ElAbd, H., Rühlemann, M. C., Arora, J., Özer, O., Lenning, O. B., Myhre, R., Vadla, M. S., Wacker, E. M., Wienbrandt, L., Blandino Ortiz, A., de Salazar, A., Garrido Chercoles, A., … Franke, A. (2022). Detailed stratified GWAS analysis for severe COVID-19 in four European populations. *Human Molecular Genetics*, *31*(23), 3945–3966. https://doi.org/10.1093/hmg/ddac158

Del Valle, D. M., Kim-Schulze, S., Huang, H.-H., Beckmann, N. D., Nirenberg, S., Wang, B., Lavin, Y., Swartz, T. H., Madduri, D., Stock, A., Marron, T. U., Xie, H., Patel, M., Tuballes, K., Van Oekelen, O., Rahman, A., Kovatch, P., Aberg, J. A., Schadt, E., … Gnjatic, S. (2020). An inflammatory cytokine signature predicts COVID-19 severity and survival. *Nature Medicine*, *26*(10), Article 10. https://doi.org/10.1038/s41591-020-1051-9

Deng, G., Yin, M., Chen, X., & Zeng, F. (2020). Clinical determinants for fatality of 44,672 patients with COVID-19. *Critical Care*, *24*(1), 179. https://doi.org/10.1186/s13054-020-02902-w

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., Field, J. R., Pulley, J. M., Ramirez, A. H., Bowton, E., Basford, M. A., Carrell, D. S., Peissig, P. L., Kho, A. N., Pacheco, J. A., Rasmussen, L. V., Crosslin, D. R., Crane, P. K., Pathak, J., … Roden, D. M. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, *31*(12), Article 12. https://doi.org/10.1038/nbt.2749

Devlin, B., & Roeder, K. (1999). Genomic Control for Association Studies. *Biometrics*, *55*(4), 997–1004. https://doi.org/10.1111/j.0006-341X.1999.00997.x

Dey, R., Zhou, W., Kiiskinen, T., Havulinna, A., Elliott, A., Karjalainen, J., Kurki, M., Qin, A., Lee, S., Palotie, A., Neale, B., Daly, M., & Lin, X. (2022). Efficient and accurate frailty model approach for genome-wide survival association analysis in large-scale biobanks. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-32885-x

Ding, Y., Hou, K., Burch, K. S., Lapinska, S., Privé, F., Vilhjálmsson, B., Sankararaman, S., & Pasaniuc, B. (2022). Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. *Nature Genetics*, *54*(1), Article 1. https://doi.org/10.1038/s41588-021-00961-5

USC
DE SANTIAGO
DE COMPOSTELA

Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., Boulier, K., Privé, F., Vilhjálmsson, B. J., Olde Loohuis, L. M., & Pasaniuc, B. (2023). Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*, *618*(7966), Article 7966. https://doi.org/10.1038/s41586-023-06079-4

Docherty, A. B., Harrison, E. M., Green, C. A., Hardwick, H. E., Pius, R., Norman, L., Holden, K. A., Read, J. M., Dondelinger, F., Carson, G., Merson, L., Lee, J., Plotkin, D., Sigfrid, L., Halpin, S., Jackson, C., Gamble, C., Horby, P. W., Nguyen-Van-Tam, J. S., … Semple, M. G. (2020). Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Prospective observational cohort study. *BMJ*, *369*, m1985. https://doi.org/10.1136/bmj.m1985

Downes, D. J., Cross, A. R., Hua, P., Roberts, N., Schwessinger, R., Cutler, A. J., Munis, A. M., Brown, J., Mielczarek, O., de Andrea, C. E., Melero, I., Gill, D. R., Hyde, S. C., Knight, J. C., Todd, J. A., Sansom, S. N., Issa, F., Davies, J. O. J., & Hughes, J. R. (2021). Identification of LZTFL1 as a candidate effector gene at a COVID-19 risk locus. *Nature Genetics*, *53*(11), Article 11. https://doi.org/10.1038/s41588-021-00955-3

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PLOS Genetics*, *9*(3), e1003348. https://doi.org/10.1371/journal.pgen.1003348

Duncan, C. J. A., Skouboe, M. K., Howarth, S., Hollensen, A. K., Chen, R., Børresen, M. L., Thompson, B. J., Stremenova Spegarova, J., Hatton, C. F., Stæger, F. F., Andersen, M. K., Whittaker, J., Paludan, S. R., Jørgensen, S. E., Thomsen, M. K., Mikkelsen, J. G., Heilmann, C., Buhas, D., Øbro, N. F., … Mogensen, T. H. (2022). Life-threatening viral disease in a novel form of autosomal recessive IFNAR2 deficiency in the Arctic. *Journal of Experimental Medicine*, *219*(6), e20212427. https://doi.org/10.1084/jem.20212427

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., & Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-11112-0

Edwards, T. L., Breeyear, J., Piekos, J. A., & Velez Edwards, D. R. (2020). Equity in Health: Consideration of Race and Ethnicity in Precision Medicine. *Trends in Genetics*, *36*(11), 807–809. https://doi.org/10.1016/j.tig.2020.07.001

Ellinghaus, D. (2023). COVID-19 host genetics and ABO blood group susceptibility. *Cambridge Prisms: Precision Medicine*, *1*, e10. https://doi.org/10.1017/pcm.2022.12

Ellis, P. J. I. (2021). Modelling suggests ABO histo-incompatibility may substantially reduce SARS-CoV-2 transmission. *Epidemics*, *35*, 100446. https://doi.org/10.1016/j.epidem.2021.100446

Esposito, S., D'Abrosca, G., Antolak, A., Pedone, P. V., Isernia, C., & Malgieri, G. (2022). Host and Viral Zinc-Finger Proteins in COVID-19. *International Journal of Molecular Sciences*, *23*(7), 3711. https://doi.org/10.3390/ijms23073711

Etz, A. (2018). Introduction to the Concept of Likelihood and Its Applications. *Advances in Methods and Practices in Psychological Science*, *1*(1), 60–69. https://doi.org/10.1177/2515245917744314

Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: Polygenic Risk Score software. *Bioinformatics*, *31*(9), 1466–1468. https://doi.org/10.1093/bioinformatics/btu848

USC DE COMPOSTELA

Evangelou, E., & Ioannidis, J. P. A. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, *14*(6), Article 6. https://doi.org/10.1038/nrg3472

Evangelou, E., Warren, H. R., Mosen-Ansorena, D., Mifsud, B., Pazoki, R., Gao, H., Ntritsos, G., Dimou, N., Cabrera, C. P., Karaman, I., Ng, F. L., Evangelou, M., Witkowska, K., Tzanis, E., Hellwege, J. N., Giri, A., Velez Edwards, D. R., Sun, Y. V., Cho, K., … Caulfield, M. J. (2018). Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nature Genetics*, *50*(10), Article 10. https://doi.org/10.1038/s41588-018-0205-x

Fajgenbaum, D. C., & June, C. H. (2020). Cytokine Storm. *New England Journal of Medicine*, *383*(23), 2255–2273. https://doi.org/10.1056/NEJMra2026131

Fallerini, C., Daga, S., Mantovani, S., Benetti, E., Picchiotti, N., Francisci, D., Paciosi, F., Schiaroli, E., Baldassarri, M., Fava, F., Palmieri, M., Ludovisi, S., Castelli, F., Quiros-Roldan, E., Vaghi, M., Rusconi, S., Siano, M., Bandini, M., Spiga, O., … Frullanti, E. (2021). Association of Toll-like receptor 7 variants with life-threatening COVID-19 disease in males: Findings from a nested case-control study. *eLife*, *10*, e67569. https://doi.org/10.7554/eLife.67569

Farooqi, R., Kooner, J. S., & Zhang, W. (2023). Associations between polygenic risk score and covid-19 susceptibility and severity across ethnic groups: UK Biobank analysis. *BMC Medical Genomics*, *16*(1), 150. https://doi.org/10.1186/s12920-023-01584-x

Fejzo, M., Rocha, N., Cimino, I., Lockhart, S. M., Petry, C. J., Kay, R. G., Burling, K., Barker, P., George, A. L., Yasara, N., Premawardhena, A., Gong, S., Cook, E., Rimmington, D., Rainbow, K., Withers, D. J., Cortessis, V., Mullin, P. M., MacGibbon, K. W., … O'Rahilly, S. (2023). GDF15 linked to maternal risk of nausea and vomiting during pregnancy. *Nature*, 1–3. https://doi.org/10.1038/s41586-023-06921-9

Fejzo, M., Sazonova, O. V., Sathirapongsasuti, J. F., Hallgrímsdóttir, I. B., Vacic, V., MacGibbon, K. W., Schoenberg, F. P., Mancuso, N., Slamon, D. J., & Mullin, P. M. (2018). Placenta and appetite genes GDF15 and IGFBP7 are associated with hyperemesis gravidarum. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-03258-0

Fernandez-Rozadilla, C., Timofeeva, M., Chen, Z., Law, P., Thomas, M., Schmit, S., Díez-Obrero, V., Hsu, L., Fernandez-Tajes, J., Palles, C., Sherwood, K., Briggs, S., Svinti, V., Donnelly, K., Farrington, S., Blackmur, J., Vaughan-Shaw, P., Shu, X., Long, J., … Peters, U. (2023). Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. *Nature Genetics*, *55*(1), Article 1. https://doi.org/10.1038/s41588-022-01222-9

Fingerlin, T. E., Murphy, E., Zhang, W., Peljto, A. L., Brown, K. K., Steele, M. P., Loyd, J. E., Cosgrove, G. P., Lynch, D., Groshong, S., Collard, H. R., Wolters, P. J., Bradford, W. Z., Kossen, K., Seiwert, S. D., du Bois, R. M., Garcia, C. K., Devine, M. S., Gudmundsson, G., … Schwartz, D. A. (2013). Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nature Genetics*, *45*(6), Article 6. https://doi.org/10.1038/ng.2609

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R. B., Elliott, K. S., Lango, H., Rayner, N. W., Shields, B., Harries, L. W., Barrett, J. C., Ellard, S., Groves, C. J., Knight, B., Patch, A.-M., Ness, A. R., Ebrahim, S., … McCarthy, M. I. (2007). A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science*, *316*(5826), 889–894. https://doi.org/10.1126/science.1141634

Fried, M. W., Crawford, J. M., Mospan, A. R., Watkins, S. E., Munoz, B., Zink, R. C., Elliott, S., Burleson, K., Landis, C., Reddy, K. R., & Brown, R. S., Jr. (2021). Patient Characteristics and

Outcomes of 11 721 Patients With Coronavirus Disease 2019 (COVID-19) Hospitalized Across the United States. *Clinical Infectious Diseases*, *72*(10), e558–e565. https://doi.org/10.1093/cid/ciaa1268

Fritsche, L. G., Nam, K., Du, J., Kundu, R., Salvatore, M., Shi, X., Lee, S., Burgess, S., & Mukherjee, B. (2023). Uncovering associations between pre-existing conditions and COVID-19 Severity: A polygenic risk score approach across three large biobanks. *PLOS Genetics*, *19*(12), e1010907. https://doi.org/10.1371/journal.pgen.1010907

Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2015). minimac2: Faster genotype imputation. *Bioinformatics*, *31*(5), 782–784. https://doi.org/10.1093/bioinformatics/btu704

Fumagalli, M., Moltke, I., Grarup, N., Racimo, F., Bjerregaard, P., Jørgensen, M. E., Korneliussen, T. S., Gerbault, P., Skotte, L., Linneberg, A., Christensen, C., Brandslund, I., Jørgensen, T., Huerta-Sánchez, E., Schmidt, E. B., Pedersen, O., Hansen, T., Albrechtsen, A., & Nielsen, R. (2015). Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science*, *349*(6254), 1343–1347. https://doi.org/10.1126/science.aab2319

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L., & Nielsen, R. (2011). Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *PLOS Genetics*, *7*(11), e1002355. https://doi.org/10.1371/journal.pgen.1002355

Gallagher, M. D., & Chen-Plotkin, A. S. (2018). The Post-GWAS Era: From Association to Function. *American Journal of Human Genetics*, *102*(5), 717–730. https://doi.org/10.1016/j.ajhg.2018.04.002

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., & Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, *47*(9), Article 9. https://doi.org/10.1038/ng.3367

Gao, Y., Yang, X., Chen, H., Tan, X., Yang, Z., Deng, L., Wang, B., Kong, S., Li, S., Cui, Y., Lei, C., Wang, Y., Pan, Y., Ma, S., Sun, H., Zhao, X., Shi, Y., Yang, Z., Wu, D., … Xu, S. (2023). A pangenome reference of 36 Chinese populations. *Nature*, *619*(7968), Article 7968. https://doi.org/10.1038/s41586-023-06173-7

Garcia-Carretero, R., Vazquez-Gomez, O., Gil-Prieto, R., & Gil-de-Miguel, A. (2023). Hospitalization burden and epidemiology of the COVID-19 pandemic in Spain (2020–2021). *BMC Infectious Diseases*, *23*(1), 476. https://doi.org/10.1186/s12879-023-08454-y

Gay, N. R., Gloudemans, M., Antonio, M. L., Abell, N. S., Balliu, B., Park, Y., Martin, A. R., Musharoff, S., Rao, A. S., Aguet, F., Barbeira, A. N., Bonazzola, R., Hormozdiari, F., Ardlie, K. G., Brown, C. D., Im, H. K., Lappalainen, T., Wen, X., Montgomery, S. B., & GTEx Consortium. (2020). Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biology*, *21*(1), 233. https://doi.org/10.1186/s13059-020-02113-0

Gaynor, S. M., Joseph, T., Bai, X., Krasheninina, O., Boutkov, B., Maxwell, E., Balasubramanian, S., Marcketta, A., Backman, J., Regeneron Genetics Center, Reid, J. G., Overton, J. D., Lotta, L. A., Marchini, J., Salerno, W. J., Baras, A., Abecasis, G. R., & Thornton, T. A. (2023). *Yield of genetic association signals from genomes, exomes, and imputation in the UK biobank* [Preprint]. Genetic and Genomic Medicine. https://doi.org/10.1101/2023.09.13.23295479

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., & Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-09718-5

Ge, T., Irvin, M. R., Patki, A., Srinivasasainagendra, V., Lin, Y.-F., Tiwari, H. K., Armstrong, N. D., Benoit, B., Chen, C.-Y., Choi, K. W., Cimino, J. J., Davis, B. H., Dikilitas, O., Etheridge, B., Feng, Y.-C. A., Gainer, V., Huang, H., Jarvik, G. P., Kachulis, C., … Karlson, E. W. (2022). Development and validation of a trans-ancestry polygenic risk score for type 2 diabetes in diverse populations. *Genome Medicine*, *14*(1), 70. https://doi.org/10.1186/s13073-022-01074-2

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics*, *10*(5), e1004383. https://doi.org/10.1371/journal.pgen.1004383

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., Tam, P. K.-H., Tsui, L.-C., Waye, M. M. Y., Wong, J. T.-F., Zeng, C., Zhang, Q., Chee, M. S., Galver, L. M., Kruglyak, S., … Methods Group. (2003). The International HapMap Project. *Nature*, *426*(6968), Article 6968. https://doi.org/10.1038/nature02168

Gioia, U., Tavella, S., Martínez-Orellana, P., Cicio, G., Colliva, A., Ceccon, M., Cabrini, M., Henriques, A. C., Fumagalli, V., Paldino, A., Presot, E., Rajasekharan, S., Iacomino, N., Pisati, F., Matti, V., Sepe, S., Conte, M. I., Barozzi, S., Lavagnino, Z., … d'Adda di Fagagna, F. (2023). SARS-CoV-2 infection induces DNA damage, through CHK1 degradation and impaired 53BP1 recruitment, and cellular senescence. *Nature Cell Biology*, *25*(4), Article 4. https://doi.org/10.1038/s41556-023-01096-x

Goddard, M. E., Kemper, K. E., MacLeod, I. M., Chamberlain, A. J., & Hayes, B. J. (2016). Genetics of complex traits: Prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1835), 20160569. https://doi.org/10.1098/rspb.2016.0569

Gómez-Carballa, A., Rivero-Calle, I., Pardo-Seco, J., Gómez-Rial, J., Rivero-Velasco, C., Rodríguez-Núñez, N., Barbeito-Castiñeiras, G., Pérez-Freixo, H., Cebey-López, M., Barral-Arca, R., Rodriguez-Tenreiro, C., Dacosta-Urbieta, A., Bello, X., Pischedda, S., Currás-Tuala, M. J., Viz-Lasheras, S., Martinón-Torres, F., Salas, A., Antonio, A. G., … Cristina, C.-S. (2022). A multi-tissue study of immune gene expression profiling highlights the key role of the nasal epithelium in COVID-19 severity. *Environmental Research*, *210*, 112890. https://doi.org/10.1016/j.envres.2022.112890

Gopalan, S., Smith, S. P., Korunes, K., Hamid, I., Ramachandran, S., & Goldberg, A. (2022). Human genetic admixture through the lens of population genomics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1852), 20200410. https://doi.org/10.1098/rstb.2020.0410

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., O'Meara, M. J., Rezelj, V. V., Guo, J. Z., Swaney, D. L., Tummino, T. A., Hüttenhain, R., Kaake, R. M., Richards, A. L., Tutuncuoglu, B., Foussard, H., Batra, J., Haas, K., Modak, M., … Krogan, N. J. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, *583*(7816), 459–468. https://doi.org/10.1038/s41586-020-2286-9

Goyal, P., Choi, J. J., Pinheiro, L. C., Schenck, E. J., Chen, R., Jabri, A., Satlin, M. J., Campion, T. R., Nahid, M., Ringel, J. B., Hoffman, K. L., Alshak, M. N., Li, H. A., Wehmeyer, G. T., Rajan, M., Reshetnyak, E., Hupert, N., Horn, E. M., Martinez, F. J., … Safford, M. M. (2020). Clinical

Characteristics of Covid-19 in New York City. *New England Journal of Medicine*, *382*(24), 2372–2374. https://doi.org/10.1056/NEJMc2010419

Graham, S. E., Clarke, S. L., Wu, K.-H. H., Kanoni, S., Zajac, G. J. M., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S., Winkler, T. W., Locke, A. E., Marouli, E., Hwang, M. Y., Han, S., Narita, A., Choudhury, A., Bentley, A. R., Ekoru, K., Verma, A., … Willer, C. J. (2021). The power of genetic diversity in genome-wide association studies of lipids. *Nature*, *600*(7890), Article 7890. https://doi.org/10.1038/s41586-021-04064-3

Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D. S. C., Du, B., Li, L., Zeng, G., Yuen, K.-Y., Chen, R., Tang, C., Wang, T., Chen, P., Xiang, J., … Zhong, N. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine*, *382*(18), 1708–1720. https://doi.org/10.1056/NEJMoa2002032

Gupta, R. K., Harrison, E. M., Ho, A., Docherty, A. B., Knight, S. R., Smeden, M. van, Abubakar, I., Lipman, M., Quartagno, M., Pius, R., Buchan, I., Carson, G., Drake, T. M., Dunning, J., Fairfield, C. J., Gamble, C., Green, C. A., Halpin, S., Hardwick, H. E., … Young, P. (2021). Development and validation of the ISARIC 4C Deterioration model for adults hospitalised with COVID-19: A prospective cohort study. *The Lancet Respiratory Medicine*, *9*(4), 349–359. https://doi.org/10.1016/S2213-2600(20)30559-2

Gupta, S., Singh, A. K., Prajapati, K. S., Kushwaha, P. P., Shuaib, M., & Kumar, S. (2020). Emerging role of ZBTB7A as an oncogenic driver and transcriptional repressor. *Cancer Letters*, *483*, 22–34. https://doi.org/10.1016/j.canlet.2020.04.015

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., de Geus, E. J. C., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimäki, T., Raitoharju, E., Kähönen, M., Seppälä, I., … Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, *48*(3), Article 3. https://doi.org/10.1038/ng.3506

Hao, L., Kraft, P., Berriz, G. F., Hynes, E. D., Koch, C., Korategere V Kumar, P., Parpattedar, S. S., Steeves, M., Yu, W., Antwi, A. A., Brunette, C. A., Danowski, M., Gala, M. K., Green, R. C., Jones, N. E., Lewis, A. C. F., Lubitz, S. A., Natarajan, P., Vassy, J. L., & Lebo, M. S. (2022). Development of a clinical polygenic risk score assay and reporting workflow. *Nature Medicine*, *28*(5), Article 5. https://doi.org/10.1038/s41591-022-01767-6

Harrell, F. E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., & Duda, S. N. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, *95*, 103208. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, *18*(1), 83. https://doi.org/10.1186/s13059-017-1215-1

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Haworth, S., Mitchell, R., Corbin, L., Wade, K. H., Dudding, T., Budu-Aggrey, A., Carslake, D., Hemani, G., Paternoster, L., Smith, G. D., Davies, N., Lawson, D. J., & J. Timpson, N. (2019). Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-018-08219-1

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science (New York, N.Y.)*, *343*(6172), 747–751. https://doi.org/10.1126/science.1243518

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. University of Guelph.

Henkens, M. T. H. M., Raafs, A. G., Verdonschot, J. A. J., Linschoten, M., van Smeden, M., Wang, P., van der Hooft, B. H. M., Tieleman, R., Janssen, M. L. F., ter Bekke, R. M. A., Hazebroek, M. R., van der Horst, I. C. C., Asselbergs, F. W., Magdelijns, F. J. H., Heymans, S. R. B., Al-Ali, A. K., Al-Muhanna, F. A., Al-Windy, N. Y. Y., Almubarak, Y. A., … on behalf of the CAPACITY-COVID collaborative consortium. (2022). Age is the main determinant of COVID-19 related in-hospital mortality with minimal impact of pre-existing comorbidities, a retrospective cohort study. *BMC Geriatrics*, *22*(1), 184. https://doi.org/10.1186/s12877-021-02673-1

Hingorani, A. D., Gratton, J., Finan, C., Schmidt, A. F., Patel, R., Sofat, R., Kuan, V., Langenberg, C., Hemingway, H., Morris, J. K., & Wald, N. J. (2023). Performance of polygenic risk scores in screening, prediction, and risk stratification: Secondary analysis of data in the Polygenic Score Catalog. *BMJ Medicine*, *2*(1). https://doi.org/10.1136/bmjmed-2023-000554

Hirschhorn, J. N., & Altshuler, D. (2002). Once and Again—Issues Surrounding Replication in Genetic Association Studies. *The Journal of Clinical Endocrinology & Metabolism*, *87*(10), 4438–4441. https://doi.org/10.1210/jc.2002-021329

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T. S., Herrler, G., Wu, N.-H., Nitsche, A., Müller, M. A., Drosten, C., & Pöhlmann, S. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*, *181*(2), 271-280.e8. https://doi.org/10.1016/j.cell.2020.02.052

Horimoto, A. R. V. R., Boyken, L. A., Blue, E. E., Grinde, K. E., Nafikov, R. A., Sohi, H. K., Nato, A. Q., Bis, J. C., Brusco, L. I., Morelli, L., Ramirez, A., Dalmasso, M. C., Temple, S., Satizabal, C., Browning, S. R., Seshadri, S., Wijsman, E. M., & Thornton, T. A. (2023). Admixture mapping implicates 13q33.3 as ancestry-of-origin locus for Alzheimer disease in Hispanic and Latino populations. *Human Genetics and Genomics Advances*, *4*(3), 100207. https://doi.org/10.1016/j.xhgg.2023.100207

Horimoto, A. R. V. R., Xue, D., Thornton, T. A., & Blue, E. E. (2021). Admixture mapping reveals the association between Native American ancestry at 3q13.11 and reduced risk of Alzheimer's disease in Caribbean Hispanics. *Alzheimer's Research & Therapy*, *13*(1), 122. https://doi.org/10.1186/s13195-021-00866-9

Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., & Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *The American Journal of Human Genetics*, *99*(6), 1245–1260. https://doi.org/10.1016/j.ajhg.2016.10.003

Horowitz, J. E., Kosmicki, J. A., Damask, A., Sharma, D., Roberts, G. H. L., Justice, A. E., Banerjee, N., Coignet, M. V., Yadav, A., Leader, J. B., Marcketta, A., Park, D. S., Lanche, R., Maxwell, E., Knight, S. C., Bai, X., Guturu, H., Sun, D., Baltzell, A., … Ferreira, M. A. R. (2022).

Genome-wide analysis provides genetic evidence that ACE2 influences COVID-19 risk and yields risk scores associated with severe disease. *Nature Genetics*, *54*(4), Article 4. https://doi.org/10.1038/s41588-021-01006-7

Hou, K., Bhattacharya, A., Mester, R., Burch, K. S., & Pasaniuc, B. (2021). On powerful GWAS in admixed populations. *Nature genetics*, *53*(12), 1631–1633. https://doi.org/10.1038/s41588-021-00953-5

Hou, K., Ding, Y., Xu, Z., Wu, Y., Bhattacharya, A., Mester, R., Belbin, G. M., Buyske, S., Conti, D. V., Darst, B. F., Fornage, M., Gignoux, C., Guo, X., Haiman, C., Kenny, E. E., Kim, M., Kooperberg, C., Lange, L., Manichaikul, A., … Pasaniuc, B. (2023). Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nature Genetics*, *55*(4), Article 4. https://doi.org/10.1038/s41588-023-01338-6

Hou, K., Gogarten, S., Kim, J., Hua, X., Dias, J.-A., Sun, Q., Wang, Y., Tan, T., Group, P. R. M. in D. P. (PRIMED) C. M. W., Atkinson, E. G., Martin, A., Shortt, J., Hirbo, J., Li, Y., Pasaniuc, B., & Zhang, H. (2023). *Admix-kit: An Integrated Toolkit and Pipeline for Genetic Analyses of Admixed Populations* (p. 2023.09.30.560263). bioRxiv. https://doi.org/10.1101/2023.09.30.560263

Hou, Y. J., Okuda, K., Edwards, C. E., Martinez, D. R., Asakura, T., Dinnon, K. H., Kato, T., Lee, R. E., Yount, B. L., Mascenik, T. M., Chen, G., Olivier, K. N., Ghio, A., Tse, L. V., Leist, S. R., Gralinski, L. E., Schäfer, A., Dang, H., Gilmore, R., … Baric, R. S. (2020). SARS-CoV-2 Reverse Genetics Reveals a Variable Infection Gradient in the Respiratory Tract. *Cell*, *182*(2), 429-446.e14. https://doi.org/10.1016/j.cell.2020.05.042

Hu, B., Guo, H., Zhou, P., & Shi, Z.-L. (2021). Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology*, *19*(3), Article 3. https://doi.org/10.1038/s41579-020-00459-7

Hu, S., Ferreira, L. A. F., Shi, S., Hellenthal, G., Marchini, J., Lawson, D. J., & Myers, S. R. (2023). *Leveraging fine-scale population structure reveals conservation in genetic effect sizes between human populations across a range of human phenotypes* (p. 2023.08.08.552281). bioRxiv. https://doi.org/10.1101/2023.08.08.552281

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., … Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*(10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Huffman, J. E. (2018). Examining the current standards for genetic discovery and replication in the era of mega-biobanks. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-07348-x

Huffman, J. E., Butler-Laporte, G., Khan, A., Pairo-Castineira, E., Drivas, T. G., Peloso, G. M., Nakanishi, T., Ganna, A., Verma, A., Baillie, J. K., Kiryluk, K., Richards, J. B., & Zeberg, H. (2022). Multi-ancestry fine mapping implicates OAS1 splicing in risk of severe COVID-19. *Nature Genetics*, *54*(2), Article 2. https://doi.org/10.1038/s41588-021-00996-8

Im, J.-Y., Kang, M.-J., Kim, B.-K., & Won, M. (2023). DDIAS, DNA damage-induced apoptosis suppressor, is a potential therapeutic target in cancer. *Experimental & Molecular Medicine*, 1–7. https://doi.org/10.1038/s12276-023-00974-6

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Im, J.-Y., Kim, B.-K., Lee, K.-W., Chun, S.-Y., Kang, M.-J., & Won, M. (2020). DDIAS promotes STAT3 activation by preventing STAT3 recruitment to PTPRM in lung cancer cells. *Oncogenesis*, *9*(1), Article 1. https://doi.org/10.1038/s41389-019-0187-2

Jackson, C. B., Farzan, M., Chen, B., & Choe, H. (2022). Mechanisms of SARS-CoV-2 entry into cells. *Nature Reviews Molecular Cell Biology*, *23*(1), Article 1. https://doi.org/10.1038/s41580-021-00418-x

Jacob, C. O. (2020). On the genetics and immunopathogenesis of COVID-19. *Clinical Immunology (Orlando, Fla.)*, *220*, 108591. https://doi.org/10.1016/j.clim.2020.108591

Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., & Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, *51*(12), Article 12. https://doi.org/10.1038/s41588-019-0530-8

Jun, T., Nirenberg, S., Weinberger, T., Sharma, N., Pujadas, E., Cordon-Cardo, C., Kovatch, P., & Huang, K. (2021). Analysis of sex-specific risk factors and clinical outcomes in COVID-19. *Communications Medicine*, *1*(1), Article 1. https://doi.org/10.1038/s43856-021-00006-2

Kachuri, L., Chatterjee, N., Hirbo, J., Schaid, D. J., Martin, I., Kullo, I. J., Kenny, E. E., Pasaniuc, B., Witte, J. S., & Ge, T. (2023). Principles and methods for transferring polygenic risk scores across global populations. *Nature Reviews Genetics*, 1–18. https://doi.org/10.1038/s41576-023-00637-2

Kachuri, L., Graff, R. E., Smith-Byrne, K., Meyers, T. J., Rashkin, S. R., Ziv, E., Witte, J. S., & Johansson, M. (2020). Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-19600-4

Kachuri, L., Hoffmann, T. J., Jiang, Y., Berndt, S. I., Shelley, J. P., Schaffer, K. R., Machiela, M. J., Freedman, N. D., Huang, W.-Y., Li, S. A., Easterlin, R., Goodman, P. J., Till, C., Thompson, I., Lilja, H., Van Den Eeden, S. K., Chanock, S. J., Haiman, C. A., Conti, D. V., … Witte, J. S. (2023). Genetically adjusted PSA levels for prostate cancer screening. *Nature Medicine*, *29*(6), Article 6. https://doi.org/10.1038/s41591-023-02277-9

Kachuri, L., Mak, A. C. Y., Hu, D., Eng, C., Huntsman, S., Elhawary, J. R., Gupta, N., Gabriel, S., Xiao, S., Keys, K. L., Oni-Orisan, A., Rodríguez-Santana, J. R., LeNoir, M. A., Borrell, L. N., Zaitlen, N. A., Williams, L. K., Gignoux, C. R., Burchard, E. G., & Ziv, E. (2023). Gene expression in African Americans, Puerto Ricans and Mexican Americans reveals ancestry-specific patterns of genetic architecture. *Nature Genetics*, *55*(6), Article 6. https://doi.org/10.1038/s41588-023-01377-z

Kaiser, J. (2023). *Geneticists should rethink how they use race and ethnicity, panel urges* [dataset]. https://doi.org/10.1126/science.adh7982

Kampourakis, K., & Peterson, E. L. (2023). The racist origins, racialist connotations, and purity assumptions of the concept of "admixture" in human evolutionary genetics. *Genetics*, *223*(3), iyad002. https://doi.org/10.1093/genetics/iyad002

Kanai, M., Andrews, S. J., Cordioli, M., Stevens, C., Neale, B. M., Daly, M., Ganna, A., Pathak, G. A., Iwasaki, A., Karjalainen, J., Mehtonen, J., Pirinen, M., Chwialkowska, K., Trankiem, A., Balaconis, M. K., Veerapen, K., Wolford, B. N., Ahmad, H. F., Andrews, S., … CGEn HostSeq—Canadian COVID-19 Human Host Genome Sequencing Databank. (2023). A second update on mapping the human genetic architecture of COVID-19. *Nature*, *621*(7977), Article 7977. https://doi.org/10.1038/s41586-023-06355-3

Kasela, S., Daniloski, Z., Bollepalli, S., Jordan, T. X., tenOever, B. R., Sanjana, N. E., & Lappalainen, T. (2021). Integrative approach identifies SLC6A20 and CXCR6 as putative causal genes for the COVID-19 GWAS signal in the 3p21.31 locus. *Genome Biology*, *22*(1), 242. https://doi.org/10.1186/s13059-021-02454-4

Kee, J., Thudium, S., Renner, D. M., Glastad, K., Palozola, K., Zhang, Z., Li, Y., Lan, Y., Cesare, J., Poleshko, A., Kiseleva, A. A., Truitt, R., Cardenas-Diaz, F. L., Zhang, X., Xie, X., Kotton, D. N., Alysandratos, K. D., Epstein, J. A., Shi, P.-Y., … Korb, E. (2022). SARS-CoV-2 disrupts host epigenetic regulation via histone mimicry. *Nature*, *610*(7931), Article 7931. https://doi.org/10.1038/s41586-022-05282-z

Kerminen, S., Martin, A. R., Koskela, J., Ruotsalainen, S. E., Havulinna, A. S., Surakka, I., Palotie, A., Perola, M., Salomaa, V., Daly, M. J., Ripatti, S., & Pirinen, M. (2019). Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland. *The American Journal of Human Genetics*, *104*(6), 1169–1181. https://doi.org/10.1016/j.ajhg.2019.05.001

Kerr, K. F., Wang, Z., Janes, H., McClelland, R. L., Psaty, B. M., & Pepe, M. S. (2014). Net Reclassification Indices for Evaluating Risk-Prediction Instruments: A Critical Review. *Epidemiology (Cambridge, Mass.)*, *25*(1), 114–121. https://doi.org/10.1097/EDE.0000000000000018

Khan, A., Turchin, M. C., Patki, A., Srinivasasainagendra, V., Shang, N., Nadukuru, R., Jones, A. C., Malolepsza, E., Dikilitas, O., Kullo, I. J., Schaid, D. J., Karlson, E., Ge, T., Meigs, J. B., Smoller, J. W., Lange, C., Crosslin, D. R., Jarvik, G. P., Bhatraju, P. K., … Kiryluk, K. (2022). Genome-wide polygenic score to predict chronic kidney disease across ancestries. *Nature Medicine*, *28*(7), Article 7. https://doi.org/10.1038/s41591-022-01869-1

Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., Natarajan, P., Lander, E. S., Lubitz, S. A., Ellinor, P. T., & Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, *50*(9), Article 9. https://doi.org/10.1038/s41588-018-0183-z

Khera, A. V., Chaffin, M., Zekavat, S. M., Collins, R. L., Roselli, C., Natarajan, P., Lichtman, J. H., D'Onofrio, G., Mattera, J., Dreyer, R., Spertus, J. A., Taylor, K. D., Psaty, B. M., Rich, S. S., Post, W., Gupta, N., Gabriel, S., Lander, E., Ida Chen, Y.-D., … Kathiresan, S. (2019). Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. *Circulation*, *139*(13), 1593–1602. https://doi.org/10.1161/CIRCULATIONAHA.118.035658

Kiflen, M., Le, A., Mao, S., Lali, R., Narula, S., Xie, F., & Paré, G. (2022). Cost-Effectiveness of Polygenic Risk Scores to Guide Statin Therapy for Cardiovascular Disease Prevention. *Circulation: Genomic and Precision Medicine*, *15*(5), e003423. https://doi.org/10.1161/CIRCGEN.121.003423

Kim, J.-H., Hwang, J., Jung, J. H., Lee, H.-J., Lee, D. Y., & Kim, S.-H. (2019). Molecular networks of FOXP family: Dual biologic functions, interplay with other molecules and clinical implications in cancer progression. *Molecular Cancer*, *18*(1), 180. https://doi.org/10.1186/s12943-019-1110-3

Kim, W., Moll, M., Qiao, D., Hobbs, B. D., Shrine, N., Sakornsakolpat, P., Tobin, M. D., Dudbridge, F., Wain, L. V., Ladd-Acosta, C., Chatterjee, N., Silverman, E. K., Cho, M. H., & Beaty, T. H. (2021). Interaction of Cigarette Smoking and Polygenic Risk Score on Reduced Lung Function. *JAMA Network Open*, *4*(12), e2139525. https://doi.org/10.1001/jamanetworkopen.2021.39525

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., & Hoh., J. (2005). Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science (New York, N.Y.)*, *308*(5720), 385–389. https://doi.org/10.1126/science.1109557

Klein, S. L., & Flanagan, K. L. (2016). Sex differences in immune responses. *Nature Reviews Immunology*, *16*(10), Article 10. https://doi.org/10.1038/nri.2016.90

Knowler, W. C., Williams, R. C., Pettitt, D. J., & Steinberg, A. G. (1988). Gm3;5,13,14 and type 2 diabetes mellitus: An association in American Indians with genetic admixture. *American Journal of Human Genetics*, *43*(4), 520–526.

Ko, J. Y., Danielson, M. L., Town, M., Derado, G., Greenlund, K. J., Kirley, P. D., Alden, N. B., Yousey-Hindes, K., Anderson, E. J., Ryan, P. A., Kim, S., Lynfield, R., Torres, S. M., Barney, G. R., Bennett, N. M., Sutton, M., Talbot, H. K., Hill, M., Hall, A. J., … COVID-NET Surveillance Team. (2021). Risk Factors for Coronavirus Disease 2019 (COVID-19)–Associated Hospitalization: COVID-19–Associated Hospitalization Surveillance Network and Behavioral Risk Factor Surveillance System. *Clinical Infectious Diseases*, *72*(11), e695–e703. https://doi.org/10.1093/cid/ciaa1419

Koch, S., Schmidtke, J., Krawczak, M., & Caliebe, A. (2023). Clinical utility of polygenic risk scores: A critical 2023 appraisal. *Journal of Community Genetics*, *14*(5), 471–487. https://doi.org/10.1007/s12687-023-00645-z

Korunes, K. L., & Goldberg, A. (2021). Human genetic admixture. *PLoS Genetics*, *17*(3), e1009374. https://doi.org/10.1371/journal.pgen.1009374

Kousathanas, A., Pairo-Castineira, E., Rawlik, K., Stuckey, A., Odhams, C. A., Walker, S., Russell, C. D., Malinauskas, T., Wu, Y., Millar, J., Shen, X., Elliott, K. S., Griffiths, F., Oosthuyzen, W., Morrice, K., Keating, S., Wang, B., Rhodes, D., Klaric, L., … Baillie, J. K. (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

Kristiansen, H., Scherer, C. A., McVean, M., Iadonato, S. P., Vends, S., Thavachelvam, K., Steffensen, T. B., Horan, K. A., Kuri, T., Weber, F., Paludan, S. R., & Hartmann, R. (2010). Extracellular 2′-5′ Oligoadenylate Synthetase Stimulates RNase L-Independent Antiviral Activity: A Novel Mechanism of Virus-Induced Innate Immunity. *Journal of Virology*, *84*(22), 11898–11904. https://doi.org/10.1128/jvi.01003-10

Kulasinghe, A., Tan, C. W., Ribeiro Dos Santos Miggiolaro, A. F., Monkman, J., SadeghiRad, H., Bhuva, D. D., Motta Junior, J. da S., Busatta Vaz de Paula, C., Nagashima, S., Baena, C. P., Souza-Fonseca-Guimaraes, P., de Noronha, L., McCulloch, T., Rossi, G. R., Cooper, C., Tang, B., Short, K. R., Davis, M. J., Souza-Fonseca-Guimaraes, F., … O'Byrne, K. (2022). Profiling of lung SARS-CoV-2 and influenza virus infection dissects virus-specific host responses and gene signatures. *The European Respiratory Journal*, *59*(6), 2101881. https://doi.org/10.1183/13993003.01881-2021

Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K. M., Reeve, M. P., Laivuori, H., Aavikko, M., Kaunisto, M. A., Loukola, A., Lahtela, E., Mattsson, H., Laiho, P., Della Briotta Parolo, P., Lehisto, A. A., Kanai, M., Mars, N., Rämö, J., … Palotie, A. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*, *613*(7944), Article 7944. https://doi.org/10.1038/s41586-022-05473-8

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Kurniansyah, N., Goodman, M. O., Khan, A. T., Wang, J., Feofanova, E., Bis, J. C., Wiggins, K. L., Huffman, J. E., Kelly, T., Elfassy, T., Guo, X., Palmas, W., Lin, H. J., Hwang, S.-J., Gao, Y., Young, K., Kinney, G. L., Smith, J. A., Yu, B., … Sofer, T. (2023). Evaluating the use of blood pressure polygenic risk scores across race/ethnic background groups. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-38990-9

Kwok, A. J., Mentzer, A., & Knight, J. C. (2021). Host genetics and infectious disease: New tools, insights and translational opportunities. *Nature Reviews Genetics*, *22*(3), Article 3. https://doi.org/10.1038/s41576-020-00297-6

Lakeman, I. M. M., Rodríguez-Girondo, M., Lee, A., Ruiter, R., Stricker, B. H., Wijnant, S. R. A., Kavousi, M., Antoniou, A. C., Schmidt, M. K., Uitterlinden, A. G., Rooij, J. van, & Devilee, P. (2020). Validation of the BOADICEA model and a 313-variant polygenic risk score for breast cancer risk prediction in a Dutch prospective cohort. *Genetics in Medicine*, *22*(11), 1803–1811. https://doi.org/10.1038/s41436-020-0884-4

Lambert, S. A., Gil, L., Jupp, S., Ritchie, S. C., Xu, Y., Buniello, A., McMahon, A., Abraham, G., Chapman, M., Parkinson, H., Danesh, J., MacArthur, J. A. L., & Inouye, M. (2021). The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nature Genetics*, *53*(4), Article 4. https://doi.org/10.1038/s41588-021-00783-5

Lamers, M. M., & Haagmans, B. L. (2022). SARS-CoV-2 pathogenesis. *Nature Reviews Microbiology*, *20*(5), Article 5. https://doi.org/10.1038/s41579-022-00713-0

Lammi, V., Nakanishi, T., Jones, S. E., Andrews, S. J., Karjalainen, J., Cortés, B., O'Brien, H. E., Fulton-Howard, B. E., Haapaniemi, H. H., Schmidt, A., Mitchell, R. E., Mousas, A., Mangino, M., Huerta-Chagoya, A., Sinnott-Armstrong, N., Cirulli, E. T., Vaudel, M., Kwong, A. S. F., Maiti, A. K., … Ollila, H. M. (2023). *Genome-wide Association Study of Long COVID* (p. 2023.06.29.23292056). medRxiv. https://doi.org/10.1101/2023.06.29.23292056

LaPierre, N., Taraszka, K., Huang, H., He, R., Hormozdiari, F., & Eskin, E. (2021). Identifying causal variants by fine mapping across multiple studies. *PLOS Genetics*, *17*(9), e1009733. https://doi.org/10.1371/journal.pgen.1009733

LaVange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., Liu, K., Giachello, A., Lee, D. J., Ryan, J., Criqui, M. H., & Elder, J. P. (2010). Sample Design and Cohort Selection in the Hispanic Community Health Study/Study of Latinos. *Annals of Epidemiology*, *20*(8), 642–649. https://doi.org/10.1016/j.annepidem.2010.05.006

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of Population Structure using Dense Haplotype Data. *PLoS Genetics*, *8*(1), e1002453. https://doi.org/10.1371/journal.pgen.1002453

Ledford, H., Cyranoski, D., & Van Noorden, R. (2020). The UK has approved a COVID vaccine—Here's what scientists now want to know. *Nature*, *588*(7837), 205–206. https://doi.org/10.1038/d41586-020-03441-8

Lee, A., Mavaddat, N., Wilcox, A. N., Cunningham, A. P., Carver, T., Hartley, S., Babb de Villiers, C., Izquierdo, A., Simard, J., Schmidt, M. K., Walter, F. M., Chatterjee, N., Garcia-Closas, M., Tischkowitz, M., Pharoah, P., Easton, D. F., & Antoniou, A. C. (2019). BOADICEA: A comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genetics in Medicine*, *21*(8), Article 8. https://doi.org/10.1038/s41436-018-0406-9

Lee, A., Seo, J., Park, S., Cho, Y., Kim, G., Li, J., Liang, L., Park, T., & Chung, W. (2024). Type 2 diabetes and its genetic susceptibility are associated with increased severity and mortality of

COVID-19 in UK Biobank. *Communications Biology*, *7*(1), Article 1. https://doi.org/10.1038/s42003-024-05799-1

Lee, D., Le Pen, J., Yatim, A., Dong, B., Aquino, Y., Ogishi, M., Pescarmona, R., Talouarn, E., Rinchai, D., Zhang, P., Perret, M., Liu, Z., Jordan, I., Elmas Bozdemir, S., Bayhan, G. I., Beaufils, C., Bizien, L., Bisiaux, A., Lei, W., … Casanova, J.-L. (2022). Inborn errors of OAS–RNase L in SARS-CoV-2–related multisystem inflammatory syndrome in children. *Science*, *379*(6632), eabo3627. https://doi.org/10.1126/science.abo3627

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Karlsson Linnér, R., Fontana, M. A., Kundu, T., Lee, C., Li, H., Li, R., Royer, R., Timshel, P. N., Walters, R. K., Willoughby, E. A., … Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, *50*(8), Article 8. https://doi.org/10.1038/s41588-018-0147-3

Lee, S. H., Goddard, M. E., Wray, N. R., & Visscher, P. M. (2012). A better coefficient of determination for genetic profile analysis. *Genetic Epidemiology*, *36*(3), 214–224. https://doi.org/10.1002/gepi.21614

Leentjens, J., Haaps, T. F. van, Wessels, P. F., Schutgens, R. E. G., & Middeldorp, S. (2021). COVID-19-associated coagulopathy and antithrombotic agents—Lessons after 1 year. *The Lancet Haematology*, *8*(7), e524–e533. https://doi.org/10.1016/S2352-3026(21)00105-8

Lehmann, B., Mackintosh, M., McVean, G., & Holmes, C. (2023). Optimal strategies for learning multi-ancestry polygenic scores vary across traits. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-38930-7

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., Perola, M., Ng, P. C., Mägi, R., Milani, L., Fischer, K., & Metspalu, A. (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *International Journal of Epidemiology*, *44*(4), 1137–1147. https://doi.org/10.1093/ije/dyt268

Lennon, N. J., Kottyan, L. C., Kachulis, C., Abul-Husn, N. S., Arias, J., Belbin, G., Below, J. E., Berndt, S. I., Chung, W. K., Cimino, J. J., Clayton, E. W., Connolly, J. J., Crosslin, D. R., Dikilitas, O., Velez Edwards, D. R., Feng, Q., Fisher, M., Freimuth, R. R., Ge, T., … Kenny, E. E. (2024). Selection, optimization and validation of ten chronic disease polygenic risk scores for clinical implementation in diverse US populations. *Nature Medicine*, *30*(2), Article 2. https://doi.org/10.1038/s41591-024-02796-z

Letko, M., Marzi, A., & Munster, V. (2020). Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nature Microbiology*, *5*(4), Article 4. https://doi.org/10.1038/s41564-020-0688-y

Lewis, A. C. F., Molina, S. J., Appelbaum, P. S., Dauda, B., Di Rienzo, A., Fuentes, A., Fullerton, S. M., Garrison, N. A., Ghosh, N., Hammonds, E. M., Jones, D. S., Kenny, E. E., Kraft, P., Lee, S. S.-J., Mauro, M., Novembre, J., Panofsky, A., Sohail, M., Neale, B. M., & Allen, D. S. (2022). Getting genetic ancestry right for science and society. *Science*, *376*(6590), 250–252. https://doi.org/10.1126/science.abm7530

Lewis, A. C. F., Perez, E. F., Prince, A. E. R., Flaxman, H. R., Gomez, L., Brockman, D. G., Chandler, P. D., Kerman, B. J., Lebo, M. S., Smoller, J. W., Weiss, S. T., Blout Zawatksy, C. L., Meigs, J. B., Green, R. C., Vassy, J. L., & Karlson, E. W. (2022). Patient and provider perspectives on polygenic risk scores: Implications for clinical reporting and utilization. *Genome Medicine*, *14*(1), 114. https://doi.org/10.1186/s13073-022-01117-8

Lewis, C. M., & Vassos, E. (2020). Polygenic risk scores: From research tools to clinical instruments. *Genome Medicine*, *12*(1), 44. https://doi.org/10.1186/s13073-020-00742-5

Li, Y., Gong, H., Wang, P., Zhu, Y., Peng, H., Cui, Y., Li, H., Liu, J., & Wang, Z. (2021). The emerging role of ISWI chromatin remodeling complexes in cancer. *Journal of Experimental & Clinical Cancer Research*, *40*(1), 346. https://doi.org/10.1186/s13046-021-02151-x

Li, Y. R., & Keating, B. J. (2014). Trans-ethnic genome-wide association studies: Advantages and challenges of mapping in diverse populations. *Genome Medicine*, *6*(10), 91. https://doi.org/10.1186/s13073-014-0091-5

Liao, K., & Zöllner, S. (2024). *A Stacking Framework for Polygenic Risk Prediction in Admixed Individuals* (p. 2024.01.31.24302103). medRxiv. https://doi.org/10.1101/2024.01.31.24302103

Libert, C., Dejager, L., & Pinheiro, I. (2010). The X chromosome in immune functions: When a chromosome makes the difference. *Nature Reviews Immunology*, *10*(8), Article 8. https://doi.org/10.1038/nri2815

Lin, M., Park, D. S., Zaitlen, N. A., Henn, B. M., & Gignoux, C. R. (2021). Admixed Populations Improve Power for Variant Discovery and Portability in Genome-Wide Association Studies. *Frontiers in Genetics*, *12*. https://www.frontiersin.org/articles/10.3389/fgene.2021.673167

Liu, C.-C., Shringapure, S., Lange, K., & Novembre, J. (2020). Exploring population structure with admixture models and principal components analysis. *Methods in molecular biology (Clifton, N.J.)*, *2090*, 67–86. https://doi.org/10.1007/978-1-0716-0199-0_4

Liu, J., Ji, H., Zheng, W., Wu, X., Zhu, J. J., Arnold, A. P., & Sandberg, K. (2010). Sex differences in renal angiotensin converting enzyme 2 (ACE2) activity are 17β-oestradiol-dependent and sex chromosome-independent. *Biology of Sex Differences*, *1*, 6. https://doi.org/10.1186/2042-6410-1-6

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D. M., Chen, F., Datta, G., Davila-Velderrain, J., McGuire, D., Tian, C., Zhan, X., Choquet, H., Docherty, A. R., Faul, J. D., Foerster, J. R., Fritsche, L. G., Gabrielsen, M. E., Gordon, S. D., Haessler, J., … Vrieze, S. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature Genetics*, *51*(2), Article 2. https://doi.org/10.1038/s41588-018-0307-5

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., Wang, H., Zheng, Z., Magi, R., Esko, T., Metspalu, A., Wray, N. R., Goddard, M. E., Yang, J., & Visscher, P. M. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-12653-0

Lo, C.-H., Nguyen, L. H., Drew, D. A., Warner, E. T., Joshi, A. D., Graham, M. S., Anyane-Yeboa, A., Shebl, F. M., Astley, C. M., Figueiredo, J. C., Guo, C.-G., Ma, W., Mehta, R. S., Kwon, S., Song, M., Davies, R., Capdevila, J., Sudre, C. H., Wolf, J., … Chan, A. T. (2021). Race, ethnicity, community-level socioeconomic factors, and risk of COVID-19 in the United States and the United Kingdom. *eClinicalMedicine*, *38*. https://doi.org/10.1016/j.eclinm.2021.101029

Loesch, D. P., Horimoto, A. R. V. R., Heilbron, K., Sarihan, E. I., Inca-Martinez, M., Mason, E., Cornejo-Olivas, M., Torres, L., Mazzetti, P., Cosentino, C., Sarapura-Castro, E., Rivera-Valdivia, A., Medina, A. C., Dieguez, E., Raggio, V., Lescano, A., Tumas, V., Borges, V., Ferraz, H. B., … Disease (LARGE-PD), the L. A. R. C. on the G. of P. (2021). Characterizing the Genetic Architecture of Parkinson's Disease in Latinos. *Annals of Neurology*, *90*(3), 353–365. https://doi.org/10.1002/ana.26153

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & L Price, A. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, *48*(11), Article 11. https://doi.org/10.1038/ng.3679

Loos, R. J. F. (2020). 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-19653-5

Loos, R. J. F., & Yeo, G. S. H. (2022). The genetics of obesity: From discovery to biology. *Nature Reviews Genetics*, *23*(2), Article 2. https://doi.org/10.1038/s41576-021-00414-z

Maas, P., Barrdahl, M., Joshi, A. D., Auer, P. L., Gaudet, M. M., Milne, R. L., Schumacher, F. R., Anderson, W. F., Check, D., Chattopadhyay, S., Baglietto, L., Berg, C. D., Chanock, S. J., Cox, D. G., Figueroa, J. D., Gail, M. H., Graubard, B. I., Haiman, C. A., Hankinson, S. E., … Chatterjee, N. (2016). Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. *JAMA Oncology*, *2*(10), 1295–1302. https://doi.org/10.1001/jamaoncol.2016.1025

Magesh, S., John, D., Li, W. T., Li, Y., Mattingly-app, A., Jain, S., Chang, E. Y., & Ongkeko, W. M. (2021). Disparities in COVID-19 Outcomes by Race, Ethnicity, and Socioeconomic Status: A Systematic Review and Meta-analysis. *JAMA Network Open*, *4*(11), e2134147. https://doi.org/10.1001/jamanetworkopen.2021.34147

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, *456*(7218), Article 7218. https://doi.org/10.1038/456018a

Mahnke, K., Guo, M., Lee, S., Sepulveda, H., Swain, S. L., Nussenzweig, M., & Steinman, R. M. (2000). The Dendritic Cell Receptor for Endocytosis, Dec-205, Can Recycle and Enhance Antigen Presentation via Major Histocompatibility Complex Class II–Positive Lysosomal Compartments. *The Journal of Cell Biology*, *151*(3), 673–684.

Mai, J., Lu, M., Gao, Q., Zeng, J., & Xiao, J. (2023). Transcriptome-wide association studies: Recent advances in methods, applications and available databases. *Communications Biology*, *6*(1), Article 1. https://doi.org/10.1038/s42003-023-05279-y

Majara, L., Kalungi, A., Koen, N., Tsuo, K., Wang, Y., Gupta, R., Nkambule, L. L., Zar, H., Stein, D. J., Kinyanda, E., Atkinson, E. G., & Martin, A. R. (2023). Low and differential polygenic score generalizability among African populations due largely to genetic diversity. *Human Genetics and Genomics Advances*, *4*(2), 100184. https://doi.org/10.1016/j.xhgg.2023.100184

Mak, T. S. H., Porsch, R. M., Choi, S. W., & Sham, P. C. (2018). *Polygenic scores for UK Biobank scale data* (p. 252270). bioRxiv. https://doi.org/10.1101/252270

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), Article 7265. https://doi.org/10.1038/nature08494

Mao, X., Bigham, A. W., Mei, R., Gutierrez, G., Weiss, K. M., Brutsaert, T. D., Leon-Velarde, F., Moore, L. G., Vargas, E., McKeigue, P. M., Shriver, M. D., & Parra, E. J. (2007). A Genomewide Admixture Mapping Panel for Hispanic/Latino Populations. *The American Journal of Human Genetics*, *80*(6), 1171–1178. https://doi.org/10.1086/518564

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

Marcus, J. H., & Novembre, J. (2017). Visualizing the geography of genetic variants. *Bioinformatics*, *33*(4), 594–595. https://doi.org/10.1093/bioinformatics/btw643

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. *International Journal of Methods in Psychiatric Research*, *27*(2), e1608. https://doi.org/10.1002/mpr.1608

Mariajoseph-Antony, L. F., Kannan, A., Panneerselvam, A., Loganathan, C., Anbarasu, K., & Prahalathan, C. (2020). Could aquaporin modulators be employed as prospective drugs for COVID-19 related pulmonary comorbidity? *Medical Hypotheses*, *143*, 110201. https://doi.org/10.1016/j.mehy.2020.110201

Marigorta, U. M., & Navarro, A. (2013). High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants. *PLOS Genetics*, *9*(6), e1003566. https://doi.org/10.1371/journal.pgen.1003566

Marnetto, D., Pärna, K., Läll, K., Molinaro, L., Montinaro, F., Haller, T., Metspalu, M., Mägi, R., Fischer, K., & Pagani, L. (2020). Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-15464-w

Mars, N., Kerminen, S., Feng, Y.-C. A., Kanai, M., Läll, K., Thomas, L. F., Skogholt, A. H., della Briotta Parolo, P., Neale, B. M., Smoller, J. W., Gabrielsen, M. E., Hveem, K., Mägi, R., Matsuda, K., Okada, Y., Pirinen, M., Palotie, A., Ganna, A., Martin, A. R., & Ripatti, S. (2022). Genome-wide risk prediction of common diseases across ancestries in one million people. *Cell Genomics*, *2*(4), 100118. https://doi.org/10.1016/j.xgen.2022.100118

Mars, N., Koskela, J. T., Ripatti, P., Kiiskinen, T. T. J., Havulinna, A. S., Lindbohm, J. V., Ahola-Olli, A., Kurki, M., Karjalainen, J., Palta, P., Neale, B. M., Daly, M., Salomaa, V., Palotie, A., Widén, E., & Ripatti, S. (2020). Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nature Medicine*, *26*(4), Article 4. https://doi.org/10.1038/s41591-020-0800-0

Mars, N., Widén, E., Kerminen, S., Meretoja, T., Pirinen, M., della Briotta Parolo, P., Palta, P., Palotie, A., Kaprio, J., Joensuu, H., Daly, M., & Ripatti, S. (2020). The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-19966-5

Martin, A. R., Gignoux, C. R., Walters, R. K., Wojcik, G. L., Neale, B. M., Gravel, S., Daly, M. J., Bustamante, C. D., & Kenny, E. E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *The American Journal of Human Genetics*, *100*(4), 635–649. https://doi.org/10.1016/j.ajhg.2017.03.004

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. *Nature Genetics*, *51*(4), Article 4. https://doi.org/10.1038/s41588-019-0379-x

Martschenko, D. O., Wand, H., Young, J. L., & Wojcik, G. L. (2023). Including multiracial individuals is crucial for race, ethnicity and ancestry frameworks in genetics and genomics. *Nature Genetics*, *55*(6), Article 6. https://doi.org/10.1038/s41588-023-01394-y

Mason, R. J. (2020). Pathogenesis of COVID-19 from a cell biology perspective. *European Respiratory Journal*, *55*(4). https://doi.org/10.1183/13993003.00607-2020

USC DE COMPOSTELA

Mathieson, I., & Scally, A. (2020). What is ancestry? *PLoS Genetics*, *16*(3), e1008624. https://doi.org/10.1371/journal.pgen.1008624

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., … Stamatoyannopoulos, J. A. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, *337*(6099), 1190–1195. https://doi.org/10.1126/science.1222794

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis, G., Maxwell, E., & Marchini, J. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, *53*(7), Article 7. https://doi.org/10.1038/s41588-021-00870-7

McCarthy, M., & Birney, E. (2021). Personalized profiles for disease risk must capture all facets of health. *Nature*, *597*(7875), 175–177. https://doi.org/10.1038/d41586-021-02401-0

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), Article 5. https://doi.org/10.1038/nrg2344

McDonald, J. T., Enguita, F. J., Taylor, D., Griffin, R. J., Priebe, W., Emmett, M. R., Sajadi, M. M., Harris, A. D., Clement, J., Dybas, J. M., Aykin-Burns, N., Guarnieri, J. W., Singh, L. N., Grabham, P., Baylin, S. B., Yousey, A., Pearson, A. N., Corry, P. M., Saravia-Butler, A., … Beheshti, A. (2021). Role of miR-2392 in driving SARS-CoV-2 infection. *Cell Reports*, *37*(3), 109839. https://doi.org/10.1016/j.celrep.2021.109839

Medina-Gomez, C., Felix, J. F., Estrada, K., Peters, M. J., Herrera, L., Kruithof, C. J., Duijts, L., Hofman, A., van Duijn, C. M., Uitterlinden, A. G., Jaddoe, V. W. V., & Rivadeneira, F. (2015). Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: The Generation R Study. *European Journal of Epidemiology*, *30*(4), 317–330. https://doi.org/10.1007/s10654-015-9998-4

Mega, J. L., Stitziel, N. O., Smith, J. G., Chasman, D. I., Caulfield, M. J., Devlin, J. J., Nordio, F., Hyde, C. L., Cannon, C. P., Sacks, F. M., Poulter, N. R., Sever, P. S., Ridker, P. M., Braunwald, E., Melander, O., Kathiresan, S., & Sabatine, M. S. (2015). Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: An analysis of primary and secondary prevention trials. *The Lancet*, *385*(9984), 2264–2271. https://doi.org/10.1016/S0140-6736(14)61730-X

Meisner, A., Kundu, P., Zhang, Y. D., Lan, L. V., Kim, S., Ghandwani, D., Pal Choudhury, P., Berndt, S. I., Freedman, N. D., Garcia-Closas, M., & Chatterjee, N. (2020). Combined Utility of 25 Disease and Risk Factor Polygenic Risk Scores for Stratifying Risk of All-Cause Mortality. *The American Journal of Human Genetics*, *107*(3), 418–431. https://doi.org/10.1016/j.ajhg.2020.07.002

Mercer, T. R., & Salit, M. (2021). Testing at scale during the COVID-19 pandemic. *Nature Reviews Genetics*, *22*(7), Article 7. https://doi.org/10.1038/s41576-021-00360-w

Mester, R., Hou, K., Ding, Y., Meeks, G., Burch, K. S., Bhattacharya, A., Henn, B. M., & Pasaniuc, B. (2023). *Impact of cross-ancestry genetic architecture on GWAS in admixed populations* (p. 2023.01.20.524946). bioRxiv. https://doi.org/10.1101/2023.01.20.524946

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Miao, J., Wu, Y., Sun, Z., Miao, X., Lu, T., Zhao, J., & Lu, Q. (2024). *Valid inference for machine learning-assisted GWAS* (p. 2024.01.03.24300779). medRxiv. https://doi.org/10.1101/2024.01.03.24300779

Mills, M. C., & Rahal, C. (2020). The GWAS Diversity Monitor tracks diversity by disease in real time. *Nature Genetics*, *52*(3), Article 3. https://doi.org/10.1038/s41588-020-0580-y

Mogil, L. S., Andaleon, A., Badalamenti, A., Dickinson, S. P., Guo, X., Rotter, J. I., Johnson, W. C., Im, H. K., Liu, Y., & Wheeler, H. E. (2018). Genetic architecture of gene expression traits across diverse populations. *PLOS Genetics*, *14*(8), e1007586. https://doi.org/10.1371/journal.pgen.1007586

Morales, J., Glantz, N., Larez, A., Bevier, W., Conneely, M., Fan, L., Reed, B., Alatorre, C., Paczkowski, R., Ahmed, T., Mackenzie, A., Duncan, I., & Kerr, D. (2020). Understanding the impact of five major determinants of health (genetics, biology, behavior, psychology, society/environment) on type 2 diabetes in U.S. Hispanic/Latino families: Mil Familias - a cohort study. *BMC Endocrine Disorders*, *20*(1), 4. https://doi.org/10.1186/s12902-019-0483-z

Moreno-Estrada, A., Gignoux, C. R., Fernández-López, J. C., Zakharia, F., Sikora, M., Contreras, A. V., Acuña-Alonzo, V., Sandoval, K., Eng, C., Romero-Hidalgo, S., Ortiz-Tello, P., Robles, V., Kenny, E. E., Nuño-Arana, I., Barquera-Lozano, R., Macín-Pérez, G., Granados-Arriola, J., Huntsman, S., Galanter, J. M., … Bustamante, C. D. (2014). The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science*, *344*(6189), 1280–1285. https://doi.org/10.1126/science.1251688

Moreno-Grau, S., De Rojas, I., Hernández, I., Quintela, I., Montrreal, L., Alegret, M., Hernández-Olasagarre, B., Madrid, L., González-Perez, A., Maroñas, O., Rosende-Roca, M., Mauleón, A., Vargas, L., Lafuente, A., Abdelnour, C., Rodríguez-Gómez, O., Gil, S., Santos-Santos, M. Á., Espinosa, A., … Ruiz, A. (2019). Genome-wide association analysis of dementia and its clinical endophenotypes reveal novel loci associated with Alzheimer's disease and three causality networks: The GR@ACE project. *Alzheimer's & Dementia*, *15*(10), 1333–1347. https://doi.org/10.1016/j.jalz.2019.06.4950

Moreno-Grau, S., Fernández, M. V., De Rojas, I., Garcia-González, P., Hernández, I., Farias, F., Budde, J. P., Quintela, I., Madrid, L., González-Pérez, A., Montrreal, L., Alarcón-Martín, E., Alegret, M., Maroñas, O., Pineda, J. A., Macías, J., The GR@ACE study group, Abdelnour, C., Aguilera, N., … for the Alzheimer's Disease Neuroimaging Initiative. (2021). Long runs of homozygosity are associated with Alzheimer's disease. *Translational Psychiatry*, *11*(1), 142. https://doi.org/10.1038/s41398-020-01145-1

Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J. K., & Przeworski, M. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. *eLife*, *9*, e48376. https://doi.org/10.7554/eLife.48376

Multhaup, M., Lehman, A., Bertram, K., Chubb, A., Smith, R., Wu, S., & Furlotte, N. (2017). *White Paper 23-17. The science behind 23andMe's Genetic Weight report.*

Munafò, M. R., & Flint, J. (2004). Meta-analysis of genetic association studies. *Trends in Genetics*, *20*(9), 439–444. https://doi.org/10.1016/j.tig.2004.06.014

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., Murakami, Y., Yuji, K., Furukawa, Y., Zembutsu, H., Tanaka, T., Ohnishi, Y., Nakamura, Y., & Kubo, M. (2017). Overview of the BioBank Japan Project: Study design and profile. *Journal of Epidemiology*, *27*(3 Suppl), S2–S8. https://doi.org/10.1016/j.je.2016.12.005

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692. https://doi.org/10.1093/biomet/78.3.691

Nakanishi, T., Pigazzini, S., Degenhardt, F., Cordioli, M., Butler-Laporte, G., Maya-Miles, D., Bujanda, L., Bouysran, Y., Niemi, M. E. K., Palom, A., Ellinghaus, D., Khan, A., Martínez-Bueno, M., Rolker, S., Amitrano, S., Roade Tato, L., Fava, F., Spinner, C. D., Prati, D., … Ganna, A. (2021). Age-dependent impact of the major common genetic risk factor for COVID-19 on severity and mortality. *Journal of Clinical Investigation*, *131*(23), e152386. https://doi.org/10.1172/JCI152386

Namkoong, H., Edahiro, R., Takano, T., Nishihara, H., Shirai, Y., Sonehara, K., Tanaka, H., Azekawa, S., Mikami, Y., Lee, H., Hasegawa, T., Okudela, K., Okuzaki, D., Motooka, D., Kanai, M., Naito, T., Yamamoto, K., Wang, Q. S., Saiki, R., … Okada, Y. (2022). DOCK2 is involved in the host genetics and biology of severe COVID-19. *Nature*, *609*(7928), Article 7928. https://doi.org/10.1038/s41586-022-05163-5

National Academies of Sciences, Engineering, and Medicine. (2023). *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field*. National Academies Press (US).

National Research Council. (2011). *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. National Academies Press. https://doi.org/10.17226/13284

Navarro, F. C. P., Mohsen, H., Yan, C., Li, S., Gu, M., Meyerson, W., & Gerstein, M. (2019). Genomics and data science: An application within an umbrella. *Genome Biology*, *20*(1), 109. https://doi.org/10.1186/s13059-019-1724-1

Nédélec, Y., Sanz, J., Baharian, G., Szpiech, Z. A., Pacis, A., Dumaine, A., Grenier, J.-C., Freiman, A., Sams, A. J., Hebert, S., Sabourin, A. P., Luca, F., Blekhman, R., Hernandez, R. D., Pique-Regi, R., Tung, J., Yotova, V., & Barreiro, L. B. (2016). Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell*, *167*(3), 657-669.e21. https://doi.org/10.1016/j.cell.2016.09.025

Need, A. C., & Goldstein, D. B. (2009). Next generation disparities in human genomics: Concerns and remedies. *Trends in Genetics*, *25*(11), 489–494. https://doi.org/10.1016/j.tig.2009.09.012

Ni, G., Zeng, J., Revez, J. A., Wang, Y., Zheng, Z., Ge, T., Restuadi, R., Kiewa, J., Nyholt, D. R., Coleman, J. R. I., Smoller, J. W., Yang, J., Visscher, P. M., & Wray, N. R. (2021). A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. *Biological psychiatry*, *90*(9), 611–620. https://doi.org/10.1016/j.biopsych.2021.04.018

Nica, A. C., & Dermitzakis, E. T. (2008). Using gene expression to investigate the genetic basis of complex disorders. *Human Molecular Genetics*, *17*(R2), R129–R134. https://doi.org/10.1093/hmg/ddn285

Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S., & Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, *541*(7637), 302–310. https://doi.org/10.1038/nature21347

Niemi, M. E. K., Daly, M. J., & Ganna, A. (2022). The human genetic epidemiology of COVID-19. *Nature Reviews Genetics*, *23*(9), Article 9. https://doi.org/10.1038/s41576-022-00478-5

Nolte, I. M. (2020). Metasubtract: An R-package to analytically produce leave-one-out meta-analysis GWAS summary statistics. *Bioinformatics*, *36*(16), 4521–4522. https://doi.org/10.1093/bioinformatics/btaa570

USC

Novembre, J. (2022). The background and legacy of Lewontin's apportionment of human genetic diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1852), 20200406. https://doi.org/10.1098/rstb.2020.0406

Novembre, J., & Peter, B. M. (2016). Recent advances in the study of fine-scale population structure in humans. *Current Opinion in Genetics & Development*, *41*, 98–105. https://doi.org/10.1016/j.gde.2016.08.007

Novembre, J., Stein, C., Asgari, S., Gonzaga-Jauregui, C., Landstrom, A., Lemke, A., Li, J., Mighton, C., Taylor, M., & Tishkoff, S. (2022). Addressing the challenges of polygenic scores in human genetic research. *The American Journal of Human Genetics*, *109*(12), 2095–2100. https://doi.org/10.1016/j.ajhg.2022.10.012

O'Driscoll, M., Ribeiro Dos Santos, G., Wang, L., Cummings, D. A. T., Azman, A. S., Paireau, J., Fontanet, A., Cauchemez, S., & Salje, H. (2021). Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature*, *590*(7844), Article 7844. https://doi.org/10.1038/s41586-020-2918-0

Okbay, A., Wu, Y., Wang, N., Jayashankar, H., Bennett, M., Nehzati, S. M., Sidorenko, J., Kweon, H., Goldman, G., Gjorgjieva, T., Jiang, Y., Hicks, B., Tian, C., Hinds, D. A., Ahlskog, R., Magnusson, P. K. E., Oskarsson, S., Hayward, C., Campbell, A., … Young, A. I. (2022). Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics*, *54*(4), Article 4. https://doi.org/10.1038/s41588-022-01016-z

Pain, O., Gillett, A. C., Austin, J. C., Folkersen, L., & Lewis, C. M. (2022). A tool for translating polygenic scores onto the absolute scale using summary statistics. *European Journal of Human Genetics*, *30*(3), Article 3. https://doi.org/10.1038/s41431-021-01028-z

Pain, O., Glanville, K. P., Hagenaars, S. P., Selzam, S., Fürtjes, A. E., Gaspar, H. A., Coleman, J. R. I., Rimfeld, K., Breen, G., Plomin, R., Folkersen, L., & Lewis, C. M. (2021). Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLOS Genetics*, *17*(5), e1009021. https://doi.org/10.1371/journal.pgen.1009021

Pairo-Castineira, E., Rawlik, K., Bretherick, A. D., Qi, T., Wu, Y., Nassiri, I., McConkey, G. A., Zechner, M., Klaric, L., Griffiths, F., Oosthuyzen, W., Kousathanas, A., Richmond, A., Millar, J., Russell, C. D., Malinauskas, T., Thwaites, R., Morrice, K., Keating, S., … Baillie, J. K. (2023). GWAS and meta-analysis identifies 49 genetic variants underlying critical COVID-19. *Nature*, *617*(7962), Article 7962. https://doi.org/10.1038/s41586-023-06034-3

Pal Choudhury, P., Brook, M. N., Hurson, A. N., Lee, A., Mulder, C. V., Coulson, P., Schoemaker, M. J., Jones, M. E., Swerdlow, A. J., Chatterjee, N., Antoniou, A. C., & Garcia-Closas, M. (2021). Comparative validation of the BOADICEA and Tyrer-Cuzick breast cancer risk models incorporating classical risk factors and polygenic risk in a population-based prospective cohort of women of European ancestry. *Breast Cancer Research*, *23*(1), 22. https://doi.org/10.1186/s13058-021-01399-7

Paludan, S. R., & Mogensen, T. H. (2022). Innate immunological pathways in COVID-19 pathogenesis. *Science Immunology*, *7*(67), eabm5505. https://doi.org/10.1126/sciimmunol.abm5505

Panofsky, A., & Bliss, C. (2017). Ambiguity and Scientific Authority: Population Classification in Genomic Science. *American Sociological Review*, *82*(1), 59–87. https://doi.org/10.1177/0003122416685812

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Parasher, A. (2021). COVID-19: Current understanding of its Pathophysiology, Clinical presentation and Treatment. *Postgraduate Medical Journal*, *97*(1147), 312–320. https://doi.org/10.1136/postgradmedj-2020-138577

Patel, A. P., Wang, M., Ruan, Y., Koyama, S., Clarke, S. L., Yang, X., Tcheandjieu, C., Agrawal, S., Fahed, A. C., Ellinor, P. T., Tsao, P. S., Sun, Y. V., Cho, K., Wilson, P. W. F., Assimes, T. L., van Heel, D. A., Butterworth, A. S., Aragam, K. G., Natarajan, P., & Khera, A. V. (2023). A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. *Nature Medicine*, *29*(7), Article 7. https://doi.org/10.1038/s41591-023-02429-x

Pathak, G. A., Karjalainen, J., Stevens, C., Neale, B. M., Daly, M., Ganna, A., Andrews, S. J., Kanai, M., Cordioli, M., Polimanti, R., Harerimana, N., Pirinen, M., Liao, R. G., Chwialkowska, K., Trankiem, A., Balaconis, M. K., Nguyen, H., Solomonson, M., Veerapen, K., … Functional Host Genomics in Infectious Diseases (FHoGID). (2022). A first update on mapping the human genetic architecture of COVID-19. *Nature*, *608*(7921), Article 7921. https://doi.org/10.1038/s41586-022-04826-7

Pathak, G. A., Singh, K., Miller-Fleming, T. W., Wendt, F. R., Ehsan, N., Hou, K., Johnson, R., Lu, Z., Gopalan, S., Yengo, L., Mohammadi, P., Pasaniuc, B., Polimanti, R., Davis, L. K., & Mancuso, N. (2021). Integrative genomic analyses identify susceptibility genes underlying COVID-19 hospitalization. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-24824-z

Pereira, A. C., Bes, T. M., Velho, M., Marques, E., Jannes, C. E., Valino, K. R., Dinardo, C. L., Costa, S. F., Duarte, A. J. S., Santos, A. R., Mitne-Neto, M., Medina-Pestana, J., & Krieger, J. E. (2022). Genetic risk factors and COVID-19 severity in Brazil: Results from BRACOVID study. *Human Molecular Genetics*, *31*(18), 3021–3031. https://doi.org/10.1093/hmg/ddac045

Pereira, E., Felipe, S., de Freitas, R., Araújo, V., Soares, P., Ribeiro, J., Henrique dos Santos, L., Alves, J. O., Canabrava, N., van Tilburg, M., Guedes, M. I., & Ceccatto, V. (2022). ABO blood group and link to COVID-19: A comprehensive review of the reported associations and their possible underlying mechanisms. *Microbial Pathogenesis*, *169*, 105658. https://doi.org/10.1016/j.micpath.2022.105658

Peterson, R. E., Kuchenbaecker, K., Walters, R. K., Chen, C.-Y., Popejoy, A. B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R. J., Brick, L., Carey, C. E., Martin, A. R., Meyers, J. L., Su, J., Chen, J., Edwards, A. C., Kalungi, A., Koen, N., Majara, L., … Duncan, L. E. (2019). Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell*, *179*(3), 589–603. https://doi.org/10.1016/j.cell.2019.08.051

Petrilli, C. M., Jones, S. A., Yang, J., Rajagopalan, H., O'Donnell, L., Chernyak, Y., Tobin, K. A., Cerfolio, R. J., Francois, F., & Horwitz, L. I. (2020). Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: Prospective cohort study. *BMJ*, *369*, m1966. https://doi.org/10.1136/bmj.m1966

Pezone, A., Olivieri, F., Napoli, M. V., Procopio, A., Avvedimento, E. V., & Gabrielli, A. (2023). Inflammation and DNA damage: Cause, effect or both. *Nature Reviews Rheumatology*, *19*(4), Article 4. https://doi.org/10.1038/s41584-022-00905-1

Piepho, H.-P., & Möhring, J. (2007). Computing Heritability and Selection Response From Unbalanced Plant Breeding Trials. *Genetics*, *177*(3), 1881–1888. https://doi.org/10.1534/genetics.107.074229

Pirastu, N., Cordioli, M., Nandakumar, P., Mignogna, G., Abdellaoui, A., Hollis, B., Kanai, M., Rajagopal, V. M., Parolo, P. D. B., Baya, N., Carey, C. E., Karjalainen, J., Als, T. D., Van der

Zee, M. D., Day, F. R., Ong, K. K., Morisaki, T., de Geus, E., Bellocco, R., … Ganna, A. (2021). Genetic analyses identify widespread sex-differential participation bias. *Nature Genetics*, *53*(5), Article 5. https://doi.org/10.1038/s41588-021-00846-7

Pires-Neto, R. C., Del Carlo Bernardi, F., Alves De Araujo, P., Mauad, T., & Dolhnikoff, M. (2016). The Expression of Water and Ion Channels in Diffuse Alveolar Damage Is Not Dependent on DAD Etiology. *PLOS ONE*, *11*(11), e0166184. https://doi.org/10.1371/journal.pone.0166184

Plomin, R. (2013). Missing heritability, polygenic scores, and gene–environment correlation. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *54*(10), 1147–1149. https://doi.org/10.1111/jcpp.12128

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., Bailey, R., Swanson, K. A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, R. W., Hammitt, L. L., … Gruber, W. C. (2020). Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New England Journal of Medicine*, *383*(27), 2603–2615. https://doi.org/10.1056/NEJMoa2034577

Policard, M., Jain, S., Rego, S., & Dakshanamurthy, S. (2021). Immune characterization and profiles of SARS-CoV-2 infected patients reveals potential host therapeutic targets and SARS-CoV-2 oncogenesis mechanism. *Virus Research*, *301*, 198464. https://doi.org/10.1016/j.virusres.2021.198464

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), Article 7624. https://doi.org/10.1038/538161a

Potere, N., Valeriani, E., Candeloro, M., Tana, M., Porreca, E., Abbate, A., Spoto, S., Rutjes, A. W. S., & Di Nisio, M. (2020). Acute complications and mortality in hospitalized patients with coronavirus disease 2019: A systematic review and meta-analysis. *Critical Care*, *24*(1), 389. https://doi.org/10.1186/s13054-020-03022-1

Pradhan, A., & Olsson, P.-E. (2020). Sex differences in severity and mortality from COVID-19: Are males more vulnerable? *Biology of Sex Differences*, *11*(1), 53. https://doi.org/10.1186/s13293-020-00330-7

Precision medicine needs an equity agenda. (2021). *Nature Medicine*, *27*(5), Article 5. https://doi.org/10.1038/s41591-021-01373-y

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews. Genetics*, *11*(7), 459. https://doi.org/10.1038/nrg2813

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Privé, F., Arbel, J., & Vilhjálmsson, B. J. (2021). LDpred2: Better, faster, stronger. *Bioinformatics*, *36*(22–23), 5424–5431. https://doi.org/10.1093/bioinformatics/btaa1029

Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., O'Reilly, P. F., & Vilhjálmsson, B. J. (2022). Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics*, *109*(1), 12–23. https://doi.org/10.1016/j.ajhg.2021.11.008

Privé, F., Vilhjálmsson, B. J., Aschard, H., & Blum, M. G. B. (2019). Making the Most of Clumping and Thresholding for Polygenic Scores. *The American Journal of Human Genetics*, *105*(6), 1213–1221. https://doi.org/10.1016/j.ajhg.2019.11.001

USC DE COMPOSTELA

Psaty, B. M., O'Donnell, C. J., Gudnason, V., Lunetta, K. L., Folsom, A. R., Rotter, J. I., Uitterlinden, A. G., Harris, T. B., Witteman, J. C. M., & Boerwinkle, E. (2009). Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from five cohorts. *Circulation. Cardiovascular genetics*, *2*(1), 73–80. https://doi.org/10.1161/CIRCGENETICS.108.829747

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Purcell, S., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., Sklar, P., Purcell (Leader), S. M., Stone, J. L., Sullivan, P. F., Ruderfer, D. M., McQuillin, A., Morris, D. W., O'Dushlaine, C. T., Corvin, A., Holmans, P. A., O'Donovan, M. C., Sklar, P., Wray, N. R., … Stanley Center for Psychiatric Research and Broad Institute of MIT and Harvard. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), Article 7256. https://doi.org/10.1038/nature08185

Race, Ethnicity, and Genetics Working Group. (2005). The Use of Racial, Ethnic, and Ancestral Categories in Human Genetics Research. *The American Journal of Human Genetics*, *77*(4), 519–532. https://doi.org/10.1086/491747

Ramasamy, S., Saez, B., Mukhopadhyay, S., Ding, D., Ahmed, A. M., Chen, X., Pucci, F., Yamin, R., Wang, J., Pittet, M. J., Kelleher, C. M., Scadden, D. T., & Sweetser, D. A. (2016). Tle1 tumor suppressor negatively regulates inflammation in vivo and modulates NF-κB inflammatory pathway. *Proceedings of the National Academy of Sciences*, *113*(7), 1871–1876. https://doi.org/10.1073/pnas.1511380113

Reginsson, G. W., Ingason, A., Euesden, J., Bjornsdottir, G., Olafsson, S., Sigurdsson, E., Oskarsson, H., Tyrfingsson, T., Runarsdottir, V., Hansdottir, I., Steinberg, S., Stefansson, H., Gudbjartsson, D. F., Thorgeirsson, T. E., & Stefansson, K. (2018). Polygenic risk scores for schizophrenia and bipolar disorder associate with addiction. *Addiction Biology*, *23*(1), 485–492. https://doi.org/10.1111/adb.12496

Reich, D., Price, A. L., & Patterson, N. (2008). Principal component analysis of genetic data. *Nature Genetics*, *40*(5), Article 5. https://doi.org/10.1038/ng0508-491

Ribeiro, D. M., & Delaneau, O. (2023). *Non-coding rare variant associations with blood traits on 166 740 UK Biobank genomes* (p. 2023.12.01.569422). bioRxiv. https://doi.org/10.1101/2023.12.01.569422

Roberts, G. H. L., Park, D. S., Coignet, M. V., McCurdy, S. R., Knight, S. C., Partha, R., Rhead, B., Zhang, M., Berkowitz, N., Team, A. S., Baltzell, A. K. H., Guturu, H., Girshick, A. R., Rand, K. A., Hong, E. L., & Ball, C. A. (2020). *AncestryDNA COVID-19 Host Genetic Study Identifies Three Novel Loci* (p. 2020.10.06.20205864). medRxiv. https://doi.org/10.1101/2020.10.06.20205864

Rose, M. C., & Voynow, J. A. (2006). Respiratory Tract Mucin Genes and Mucin Glycoproteins in Health and Disease. *Physiological Reviews*, *86*(1), 245–278. https://doi.org/10.1152/physrev.00010.2005

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., & Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nature reviews. Genetics*, *11*(5), 356–366. https://doi.org/10.1038/nrg2760

Rosenthal, E. A., Hsu, L., Thomas, M., Peters, U., Kachulis, C., Patterson, K., & Jarvik, G. P. (2024). *Comparing ancestry calibration approaches for a trans-ancestry colorectal cancer polygenic risk score* (p. 2023.10.23.23296753). medRxiv. https://doi.org/10.1101/2023.10.23.23296753

Roy, M. G., Livraghi-Butrico, A., Fletcher, A. A., McElwee, M. M., Evans, S. E., Boerner, R. M., Alexander, S. N., Bellinghausen, L. K., Song, A. S., Petrova, Y. M., Tuvim, M. J., Adachi, R., Romo, I., Bordt, A. S., Gabriela Bowden, M., Sisson, J. H., Woodruff, P. G., Thornton, D. J., Rousseau, K., … Evans, C. M. (2014). Muc5b Is Required for Airway Defense. *Nature*, *505*(7483), 412–416. https://doi.org/10.1038/nature12807

Ruan, Y., Lin, Y.-F., Feng, Y.-C. A., Chen, C.-Y., Lam, M., Guo, Z., Initiatives, S. G. A., He, L., Sawa, A., Martin, A. R., Qin, S., Huang, H., & Ge, T. (2022). Improving Polygenic Prediction in Ancestrally Diverse Populations. *Nature Genetics*, *54*(5), 573. https://doi.org/10.1038/s41588-022-01054-7

Ruiz-Linares, A., Adhikari, K., Acuña-Alonzo, V., Quinto-Sanchez, M., Jaramillo, C., Arias, W., Fuentes, M., Pizarro, M., Everardo, P., de Avila, F., Gómez-Valdés, J., León-Mimila, P., Hunemeier, T., Ramallo, V., Silva de Cerqueira, C. C., Burley, M.-W., Konca, E., de Oliveira, M. Z., Veronez, M. R., … Gonzalez-José, R. (2014). Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. *PLoS Genetics*, *10*(9), e1004572. https://doi.org/10.1371/journal.pgen.1004572

Russell, C. D., Lone, N. I., & Baillie, J. K. (2023). Comorbidities, multimorbidity and COVID-19. *Nature Medicine*, *29*(2), Article 2. https://doi.org/10.1038/s41591-022-02156-9

Sadler, A. J., & Williams, B. R. G. (2008). Interferon-inducible antiviral effectors. *Nature Reviews Immunology*, *8*(7), Article 7. https://doi.org/10.1038/nri2314

Sah, P., Fitzpatrick, M. C., Zimmer, C. F., Abdollahi, E., Juden-Kelly, L., Moghadas, S. M., Singer, B. H., & Galvani, A. P. (2021). Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis. *Proceedings of the National Academy of Sciences*, *118*(34), e2109229118. https://doi.org/10.1073/pnas.2109229118

Saitou, M., Dahl, A., Wang, Q., & Liu, X. (2022). *Allele frequency differences of causal variants have a major impact on low cross-ancestry portability of PRS* (p. 2022.10.21.22281371). medRxiv. https://doi.org/10.1101/2022.10.21.22281371

Sakaue, S., Hirata, J., Kanai, M., Suzuki, K., Akiyama, M., Lai Too, C., Arayssi, T., Hammoudeh, M., Al Emadi, S., Masri, B. K., Halabi, H., Badsha, H., Uthman, I. W., Saxena, R., Padyukov, L., Hirata, M., Matsuda, K., Murakami, Y., Kamatani, Y., & Okada, Y. (2020). Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-15194-z

Sato, G., Shirai, Y., Namba, S., Edahiro, R., Sonehara, K., Hata, T., Uemura, M., Matsuda, K., Doki, Y., Eguchi, H., & Okada, Y. (2023). Pan-cancer and cross-population genome-wide association studies dissect shared genetic backgrounds underlying carcinogenesis. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-39136-7

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, *19*(8), Article 8. https://doi.org/10.1038/s41576-018-0016-z

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Schmiedel, B. J., Rocha, J., Gonzalez-Colin, C., Bhattacharyya, S., Madrigal, A., Ottensmeier, C. H., Ay, F., Chandra, V., & Vijayanand, P. (2021). COVID-19 genetic risk variants are associated with expression of multiple genes in diverse immune cell types. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-26888-3

Schoggins, J. W., & Rice, C. M. (2011). Interferon-stimulated genes and their antiviral effector functions. *Current Opinion in Virology*, *1*(6), 519–525. https://doi.org/10.1016/j.coviro.2011.10.008

Schwalb, A., Armyra, E., Méndez-Aranda, M., & Ugarte-Gil, C. (2022). COVID-19 in Latin America and the Caribbean: Two years of the pandemic. *Journal of Internal Medicine*, *292*(3), 409–427. https://doi.org/10.1111/joim.13499

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), Article 7. https://doi.org/10.1038/s41577-020-0348-8

Scutari, M., Mackay, I., & Balding, D. (2016). Using Genetic Distance to Infer the Accuracy of Genomic Prediction. *PLOS Genetics*, *12*(9), e1006288. https://doi.org/10.1371/journal.pgen.1006288

Seldin, M. F., Pasaniuc, B., & Price, A. L. (2011). New approaches to disease mapping in admixed populations. *Nature Reviews Genetics*, *12*(8), Article 8. https://doi.org/10.1038/nrg3002

Shelton, J. F., Shastri, A. J., Ye, C., Weldon, C. H., Filshtein-Sonmez, T., Coker, D., Symons, A., Esparza-Gordillo, J., The 23andMe COVID-19 Team, Chubb, A., Fitch, A., Kung, A., Altman, A., Kill, A., Tan, J., Pollard, J., McCreight, J., Bielenberg, J., Matthews, J., … Auton, A. (2021). Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. *Nature Genetics*, *53*(6), 801–808. https://doi.org/10.1038/s41588-021-00854-7

Shi, H., Burch, K. S., Johnson, R., Freund, M. K., Kichaev, G., Mancuso, N., Manuel, A. M., Dong, N., & Pasaniuc, B. (2020). Localizing Components of Shared Transethnic Genetic Architecture of Complex Traits from GWAS Summary Data. *The American Journal of Human Genetics*, *106*(6), 805–817. https://doi.org/10.1016/j.ajhg.2020.04.012

Shin, H. M., Kapoor, V. N., Kim, G., Li, P., Kim, H.-R., Suresh, M., Kaech, S. M., Wherry, E. J., Selin, L. K., Leonard, W. J., Welsh, R. M., & Berg, L. J. (2017). Transient expression of ZBTB32 in anti-viral CD8+ T cells limits the magnitude of the effector response and the generation of memory. *PLOS Pathogens*, *13*(8), e1006544. https://doi.org/10.1371/journal.ppat.1006544

Shin, J., Toyoda, S., Nishitani, S., Onodera, T., Fukuda, S., Kita, S., Fukuhara, A., & Shimomura, I. (2022). SARS-CoV-2 infection impairs the insulin/IGF signaling pathway in the lung, liver, adipose tissue, and pancreatic cells via IRF1. *Metabolism*, *133*, 155236. https://doi.org/10.1016/j.metabol.2022.155236

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. *Cell*, *177*(1), 26–31. https://doi.org/10.1016/j.cell.2019.02.048

Skoglund, P., & Mathieson, I. (2018). Ancient Genomics of Modern Humans: The First Decade. *Annual Review of Genomics and Human Genetics*, *19*(1), 381–404. https://doi.org/10.1146/annurev-genom-083117-021749

Slatkin, M. (2008). Linkage disequilibrium—Understanding the evolutionary past and mapping the medical future. *Nature reviews. Genetics*, *9*(6), 477–485. https://doi.org/10.1038/nrg2361

USC SANTIAGO

Smieszek, S. P., Polymeropoulos, V. M., Xiao, C., Polymeropoulos, C. M., & Polymeropoulos, M. H. (2021). Loss-of-function mutations in IFNAR2 in COVID-19 severe infection susceptibility. *Journal of Global Antimicrobial Resistance*, *26*, 239–240. https://doi.org/10.1016/j.jgar.2021.06.005

Smith, S. P., Shahamatdar, S., Cheng, W., Zhang, S., Paik, J., Graff, M., Haiman, C., Matise, T. C., North, K. E., Peters, U., Kenny, E., Gignoux, C., Wojcik, G., Crawford, L., & Ramachandran, S. (2022). Enrichment analyses identify shared associations for 25 quantitative traits in over 600,000 individuals from seven diverse ancestries. *The American Journal of Human Genetics*, *109*(5), 871–884. https://doi.org/10.1016/j.ajhg.2022.03.005

Sohail, M., Palma-Martínez, M. J., Chong, A. Y., Quinto-Cortés, C. D., Barberena-Jonas, C., Medina-Muñoz, S. G., Ragsdale, A., Delgado-Sánchez, G., Cruz-Hervert, L. P., Ferreyra-Reyes, L., Ferreira-Guerrero, E., Mongua-Rodríguez, N., Canizales-Quintero, S., Jimenez-Kaufmann, A., Moreno-Macías, H., Aguilar-Salinas, C. A., Auckland, K., Cortés, A., Acuña-Alonzo, V., … Moreno-Estrada, A. (2023). Mexican Biobank advances population and medical genomics of diverse ancestries. *Nature*, *622*(7984), Article 7984. https://doi.org/10.1038/s41586-023-06560-0

Song, Y., Fukuda, N., Bai, C., Ma, T., Matthay, M. A., & Verkman, A. S. (2000). Role of aquaporins in alveolar fluid clearance in neonatal and adult lung, and in oedema formation following acute lung injury: Studies in transgenic aquaporin null mice. *The Journal of Physiology*, *525*(3), 771–779. https://doi.org/10.1111/j.1469-7793.2000.00771.x

Staal, F. J. T., Luis, T. C., & Tiemessen, M. M. (2008). WNT signalling in the immune system: WNT is spreading its wings. *Nature Reviews Immunology*, *8*(8), Article 8. https://doi.org/10.1038/nri2360

Stewart, C. A., Gay, C. M., Ramkumar, K., Cargill, K. R., Cardnell, R. J., Nilsson, M. B., Heeke, S., Park, E. M., Kundu, S. T., Diao, L., Wang, Q., Shen, L., Xi, Y., Zhang, B., Corte, C. M. D., Fan, Y., Kundu, K., Gao, B., Avila, K., … Byers, L. A. (2021). Lung Cancer Models Reveal Severe Acute Respiratory Syndrome Coronavirus 2–Induced Epithelial-to-Mesenchymal Transition Contributes to Coronavirus Disease 2019 Pathophysiology. *Journal of Thoracic Oncology*, *16*(11), 1821–1839. https://doi.org/10.1016/j.jtho.2021.07.002

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavaré, S., Deloukas, P., & Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nature Genetics*, *39*(10), Article 10. https://doi.org/10.1038/ng2142

Suarez-Pajes, E., Díaz-de Usera, A., Marcelino-Rodríguez, I., Guillen-Guio, B., & Flores, C. (2021). Genetic Ancestry Inference and Its Application for the Genetic Mapping of Human Diseases. *International Journal of Molecular Sciences*, *22*(13), 6962. https://doi.org/10.3390/ijms22136962

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

Sul, J. H., Martin, L. S., & Eskin, E. (2018). Population structure in genetic studies: Confounding factors and mixed models. *PLoS Genetics*, *14*(12), e1007309. https://doi.org/10.1371/journal.pgen.1007309

Sullivan, P. F., Agrawal, A., Bulik, C. M., Andreassen, O. A., Børglum, A. D., Breen, G., Cichon, S., Edenberg, H. J., Faraone, S. V., Gelernter, J., Mathews, C. A., Nievergelt, C. M., Smoller, J. W., & O'Donovan, M. C. (2018). Psychiatric Genomics: An Update and an Agenda. *The American journal of psychiatry*, *175*(1), 15–27. https://doi.org/10.1176/appi.ajp.2017.17030283

Sun, B. B., Chiou, J., Traylor, M., Benner, C., Hsu, Y.-H., Richardson, T. G., Surendran, P., Mahajan, A., Robins, C., Vasquez-Grinnell, S. G., Hou, L., Kvikstad, E. M., Burren, O. S., Davitte, J., Ferber, K. L., Gillies, C. E., Hedman, Å. K., Hu, S., Lin, T., … Whelan, C. D. (2023). Plasma proteomic associations with genetics and health in the UK Biobank. *Nature*, 1–10. https://doi.org/10.1038/s41586-023-06592-6

Sun, Q., Rowland, B. T., Chen, J., Mikhaylova, A. V., Avery, C., Peters, U., Lundin, J., Matise, T., Buyske, S., Tao, R., Mathias, R. A., Reiner, A. P., Auer, P. L., Cox, N. J., Kooperberg, C., Thornton, T. A., Raffield, L. M., & Li, Y. (2024). Improving polygenic risk prediction in admixed populations by explicitly modeling ancestral-differential effects via GAUDI. *Nature Communications*, *15*(1), Article 1. https://doi.org/10.1038/s41467-024-45135-z

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., … Iwasaki, A. (2020). Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature*, *588*(7837), Article 7837. https://doi.org/10.1038/s41586-020-2700-3

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., … Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, *590*(7845), 290–299. https://doi.org/10.1038/s41586-021-03205-y

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), Article 8. https://doi.org/10.1038/s41576-019-0127-1

Tan, T., & Atkinson, E. G. (2023). Strategies for the Genomic Analysis of Admixed Populations. *Annual Review of Biomedical Data Science*, *6*(1), 105–127. https://doi.org/10.1146/annurev-biodatasci-020722-014310

Tanigawa, Y., & Kellis, M. (2023). Power of inclusion: Enhancing polygenic prediction with admixed individuals. *The American Journal of Human Genetics*, *110*(11), 1888–1902. https://doi.org/10.1016/j.ajhg.2023.09.013

Teixeira da Silva, J. A., Tsigaris, P., & Erfanmanesh, M. (2021). Publishing volumes in major databases related to Covid-19. *Scientometrics*, *126*(1), 831–842. https://doi.org/10.1007/s11192-020-03675-3

Tetsuka, T., Uranishi, H., Imai, H., Ono, T., Sonta, S., Takahashi, N., Asamitsu, K., & Okamoto, T. (2000). Inhibition of Nuclear Factor-κB-mediated Transcription by Association with the Amino-terminal Enhancer of Split, a Groucho-related Protein Lacking WD40 Repeats *. *Journal of Biological Chemistry*, *275*(6), 4383–4390. https://doi.org/10.1074/jbc.275.6.4383

The "All of Us" Research Program. (2019). *New England Journal of Medicine*, *381*(7), 668–676. https://doi.org/10.1056/NEJMsr1809937

The COVID-19 Host Genetics Initiative. (2020). The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the

SARS-CoV-2 virus pandemic. *European Journal of Human Genetics*, *28*(6), 715–718. https://doi.org/10.1038/s41431-020-0636-6

The Economist, [original data]. (2022). *Central estimate* [dataset]. https://www.economist.com/graphic-detail/coronavirus-excess-deaths-estimates.

The GenOMICC Investigators, The ISARIC4C Investigators, The COVID-19 Human Genetics Initiative, 23andMe Investigators, BRACOVID Investigators, Gen-COVID Investigators, Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A. D., Rawlik, K., Pasko, D., Walker, S., Parkinson, N., Fourman, M. H., Russell, C. D., Furniss, J., Richmond, A., Gountouna, E., … Baillie, J. K. (2021). Genetic mechanisms of critical illness in COVID-19. *Nature*, *591*(7848), 92–98. https://doi.org/10.1038/s41586-020-03065-y

The Severe Covid-19 GWAS Group, Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P., Fernández, J., Prati, D., Baselli, G., Asselta, R., Grimsrud, M. M., Milani, C., Aziz, F., Kässens, J., May, S., Wendorff, M., Wienbrandt, L., Uellendahl-Werth, F., … Karlsen, T. H. (2020). Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *New England Journal of Medicine*, *383*(16), 1522–1534. https://doi.org/10.1056/NEJMoa2020283

Tian, C., Gregersen, P. K., & Seldin, M. F. (2008). Accounting for ancestry: Population substructure and genome-wide association studies. *Human Molecular Genetics*, *17*(R2), R143–R150. https://doi.org/10.1093/hmg/ddn268

Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J., & Richards, J. B. (2018). Genetic architecture: The shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*, *19*(2), Article 2. https://doi.org/10.1038/nrg.2017.101

Torkamani, A., Wineinger, N. E., & Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, *19*(9), Article 9. https://doi.org/10.1038/s41576-018-0018-x

Tripepi, G., Jager, K. J., Dekker, F. W., & Zoccali, C. (2010). Selection Bias and Information Bias in Clinical Research. *Nephron Clinical Practice*, *115*(2), c94–c99. https://doi.org/10.1159/000312871

Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., Bryois, J., Chen, C.-Y., Dennison, C. A., Hall, L. S., Lam, M., Watanabe, K., Frei, O., Ge, T., Harwood, J. C., Koopmans, F., Magnusson, S., Richards, A. L., Sidorenko, J., … O'Donovan, M. C. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*, *604*(7906), Article 7906. https://doi.org/10.1038/s41586-022-04434-5

Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., Nguyen-Viet, T. A., Wedow, R., Zacher, M., Furlotte, N. A., Magnusson, P., Oskarsson, S., Johannesson, M., Visscher, P. M., Laibson, D., Cesarini, D., Neale, B. M., & Benjamin, D. J. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*, *50*(2), Article 2. https://doi.org/10.1038/s41588-017-0009-4

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., de Andrade, M., Doheny, K. F., Haines, J. L., Hayes, G., Jarvik, G., Jiang, L., Kullo, I. J., Li, R., Ling, H., Manolio, T. A., Matsumoto, M., McCarty, C. A., McDavid, A. N., … Ritchie, M. D. (2011). Quality Control Procedures for Genome Wide Association Studies. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, *CHAPTER*, Unit1.19. https://doi.org/10.1002/0471142905.hg0119s68

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), Article 1. https://doi.org/10.1038/s43586-021-00056-9

Urbut, S. M., Wang, G., Carbonetto, P., & Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics*, *51*(1), Article 1. https://doi.org/10.1038/s41588-018-0268-8

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., Bossuyt, P., Collins, G. S., Macaskill, P., McLernon, D. J., Moons, K. G. M., Steyerberg, E. W., Van Calster, B., van Smeden, M., Vickers, A. J., & On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, *17*(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Van Der Made, C. I., Simons, A., Schuurs-Hoeijmakers, J., Van Den Heuvel, G., Mantere, T., Kersten, S., Van Deuren, R. C., Steehouwer, M., Van Reijmersdal, S. V., Jaeger, M., Hofste, T., Astuti, G., Corominas Galbany, J., Van Der Schoot, V., Van Der Hoeven, H., Hagmolen Of Ten Have, W., Klijn, E., Van Den Meer, C., Fiddelaers, J., … Hoischen, A. (2020). Presence of Genetic Variants Among Young Men With Severe COVID-19. *JAMA*, *324*(7), 663. https://doi.org/10.1001/jama.2020.13719

Vassiliou, A. G., Kotanidou, A., Dimopoulou, I., & Orfanos, S. E. (2020). Endothelial Damage in Acute Respiratory Distress Syndrome. *International Journal of Molecular Sciences*, *21*(22), 8793. https://doi.org/10.3390/ijms21228793

Vickers, A. J., Calster, B. V., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*, *352*, i6. https://doi.org/10.1136/bmj.i6

Vickers, A. J., Cronin, A. M., Elkin, E. B., & Gonen, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Medical Informatics and Decision Making*, *8*(1), 53. https://doi.org/10.1186/1472-6947-8-53

Vickers, A. J., van Calster, B., & Steyerberg, E. W. (2019). A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and Prognostic Research*, *3*(1), 18. https://doi.org/10.1186/s41512-019-0064-7

Visscher, P. M., & Goddard, M. E. (2019). From R.A. Fisher's 1918 Paper to GWAS a Century Later. *Genetics*, *211*(4), 1125–1130. https://doi.org/10.1534/genetics.118.301594

Visscher, P. M., Hill, W. G., & Wray, N. R. (2008). Heritability in the genomics era—Concepts and misconceptions. *Nature Reviews Genetics*, *9*(4), Article 4. https://doi.org/10.1038/nrg2322

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Voight, B. F., & Pritchard, J. K. (2005). Confounding from Cryptic Relatedness in Case-Control Association Studies. *PLOS Genetics*, *1*(3), e32. https://doi.org/10.1371/journal.pgen.0010032

Vuille-dit-Bille, R. N., Camargo, S. M., Emmenegger, L., Sasse, T., Kummer, E., Jando, J., Hamie, Q. M., Meier, C. F., Hunziker, S., Forras-Kaufmann, Z., Kuyumcu, S., Fox, M., Schwizer, W., Fried, M., Lindenmeyer, M., Götze, O., & Verrey, F. (2015). Human intestine luminal ACE2 and amino acid transporter expression increased by ACE-inhibitors. *Amino Acids*, *47*(4), 693–705. https://doi.org/10.1007/s00726-014-1889-6

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., Björkegren, J. L. M., Im, H. K., Pasaniuc, B., Rivas, M. A., & Kundaje, A. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*, *51*(4), Article 4. https://doi.org/10.1038/s41588-019-0385-z

Wand, H., Lambert, S. A., Tamburro, C., Iacocca, M. A., O'Sullivan, J. W., Sillari, C., Kullo, I. J., Rowley, R., Dron, J. S., Brockman, D., Venner, E., McCarthy, M. I., Antoniou, A. C., Easton, D. F., Hegele, R. A., Khera, A. V., Chatterjee, N., Kooperberg, C., Edwards, K., … Wojcik, G. L. (2021). Improving reporting standards for polygenic scores in risk prediction studies. *Nature*, *591*(7849), Article 7849. https://doi.org/10.1038/s41586-021-03243-6

Wang, Q., Dhindsa, R. S., Carss, K., Harper, A. R., Nag, A., Tachmazidou, I., Vitsios, D., Deevi, S. V. V., Mackay, A., Muthas, D., Hühn, M., Monkley, S., Olsson, H., Wasilewski, S., Smith, K. R., March, R., Platt, A., Haefliger, C., & Petrovski, S. (2021). Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature*, *597*(7877), Article 7877. https://doi.org/10.1038/s41586-021-03855-y

Wang, R., Lee, J.-H., Kim, J., Xiong, F., Hasani, L. A., Shi, Y., Simpson, E. N., Zhu, X., Chen, Y.-T., Shivshankar, P., Krakowiak, J., Wang, Y., Gilbert, D. M., Yuan, X., Eltzschig, H. K., & Li, W. (2023). SARS-CoV-2 restructures host chromatin architecture. *Nature Microbiology*, *8*(4), Article 4. https://doi.org/10.1038/s41564-023-01344-8

Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P. M., & Yengo, L. (2020). Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-17719-y

Wang, Y., Kanai, M., Tan, T., Kamariza, M., Tsuo, K., Yuan, K., Zhou, W., Okada, Y., BioBank Japan Project, Huang, H., Turley, P., Atkinson, E. G., & Martin, A. R. (2023). Polygenic prediction across populations is influenced by ancestry, genetic architecture, and methodology. *Cell Genomics*, *3*(10), 100408. https://doi.org/10.1016/j.xgen.2023.100408

Wang, Y., Namba, S., Lopera, E., Kerminen, S., Tsuo, K., Läll, K., Kanai, M., Zhou, W., Wu, K.-H., Favé, M.-J., Bhatta, L., Awadalla, P., Brumpton, B., Deelen, P., Hveem, K., Faro, V. L., Mägi, R., Murakami, Y., Sanna, S., … Hirbo, J. (2023). Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. *Cell Genomics*, *3*(1). https://doi.org/10.1016/j.xgen.2022.100241

Wang, Y., Tsuo, K., Kanai, M., Neale, B. M., & Martin, A. R. (2022). Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. *Annual Review of Biomedical Data Science*, *5*(1), 293–320. https://doi.org/10.1146/annurev-biodatasci-111721-074830

Watanabe, K., Taskesen, E., van Bochoven, A., & Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/s41467-017-01261-5

Wei, J., Alfajaro, M. M., DeWeirdt, P. C., Hanna, R. E., Lu-Culligan, W. J., Cai, W. L., Strine, M. S., Zhang, S.-M., Graziano, V. R., Schmitz, C. O., Chen, J. S., Mankowski, M. C., Filler, R. B., Ravindra, N. G., Gasque, V., de Miguel, F. J., Patil, A., Chen, H., Oguntuyo, K. Y., … Wilen, C. B. (2021). Genome-wide CRISPR Screens Reveal Host Factors Critical for SARS-CoV-2 Infection. *Cell*, *184*(1), 76-91.e13. https://doi.org/10.1016/j.cell.2020.10.028

Wei, J., Patil, A., Collings, C. K., Alfajaro, M. M., Liang, Y., Cai, W. L., Strine, M. S., Filler, R. B., DeWeirdt, P. C., Hanna, R. E., Menasche, B. L., Ökten, A., Peña-Hernández, M. A., Klein, J.,

McNamara, A., Rosales, R., McGovern, B. L., Luis Rodriguez, M., García-Sastre, A., … Wilen, C. B. (2023). Pharmacological disruption of mSWI/SNF complex activity restricts SARS-CoV-2 infection. *Nature Genetics*, *55*(3), Article 3. https://doi.org/10.1038/s41588-023-01307-z

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*(17), 2190–2191. https://doi.org/10.1093/bioinformatics/btq340

Williams, A. L., Jacobs, S. B. R., Moreno-Macías, H., Huerta-Chagoya, A., Churchhouse, C., Márquez-Luna, C., García-Ortíz, H., José Gómez-Vázquez, M., Burtt, N. P., Aguilar-Salinas, C. A., González-Villalpando, C., Florez, J. C., Orozco, L., Haiman, C. A., Tusié-Luna, T., Altshuler, D., Williams, A. L., Márquez-Luna, C., Huerta-Chagoya, A., … Steering committee. (2014). Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature*, *506*(7486), Article 7486. https://doi.org/10.1038/nature12828

Williamson, E. J., Walker, A. J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C. E., Curtis, H. J., Mehrkar, A., Evans, D., Inglesby, P., Cockburn, J., McDonald, H. I., MacKenna, B., Tomlinson, L., Douglas, I. J., Rentsch, C. T., Mathur, R., Wong, A. Y. S., Grieve, R., … Goldacre, B. (2020). Factors associated with COVID-19-related death using OpenSAFELY. *Nature*, *584*(7821), Article 7821. https://doi.org/10.1038/s41586-020-2521-4

Winkler, T. W., Day, F. R., Croteau-Chonka, D. C., Wood, A. R., Locke, A. E., Mägi, R., Ferreira, T., Fall, T., Graff, M., Justice, A. E., Luan, J., Gustafsson, S., Randall, J. C., Vedantam, S., Workalemahu, T., Kilpeläinen, T. O., Scherag, A., Esko, T., Kutalik, Z., … Loos, R. J. F. (2014). Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols*, *9*(5), Article 5. https://doi.org/10.1038/nprot.2014.071

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., Belbin, G. M., Bien, S. A., Cheng, I., Cullina, S., Hodonsky, C. J., Hu, Y., Huckins, L. M., Jeff, J., Justice, A. E., … Carlson, C. S. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, *570*(7762), 514–518. https://doi.org/10.1038/s41586-019-1310-4

Won, K.-J., Im, J.-Y., Yun, C.-O., Chung, K.-S., Kim, Y. J., Lee, J.-S., Jung, Y.-J., Kim, B.-K., Song, K. B., Kim, Y.-H., Chun, H.-K., Jung, K. E., Kim, M.-H., & Won, M. (2014). Human Noxin is an anti-apoptotic protein in response to DNA damage of A549 non-small cell lung carcinoma. *International Journal of Cancer*, *134*(11), 2595–2604. https://doi.org/10.1002/ijc.28600

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., Amin, N., Buchkovich, M. L., Croteau-Chonka, D. C., Day, F. R., Duan, Y., Fall, T., Fehrmann, R., Ferreira, T., Jackson, A. U., … Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, *46*(11), Article 11. https://doi.org/10.1038/ng.3097

World Health Organization. (2020a). *Coronavirus disease 2019 (COVID-19). Situation report – 51*.

World Health Organization. (2020b). *Coronavirus disease 2019 (COVID-19). Situation report – 71*.

World Health Organization. (2023). *Guideline Clinical management of COVID-19: Living guideline* (18). World Health Organization.

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J., & Visscher, P. M. (2018). Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. *Cell*, *173*(7), 1573–1580. https://doi.org/10.1016/j.cell.2018.05.051

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., & Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*, *14*(7), Article 7. https://doi.org/10.1038/nrg3457

Wu, S.-C., Arthur, C. M., Jan, H.-M., Garcia-Beltran, W. F., Patel, K. R., Rathgeber, M. F., Verkerke, H. P., Cheedarla, N., Jajosky, R. P., Paul, A., Neish, A. S., Roback, J. D., Josephson, C. D., Wesemann, D. R., Kalman, D., Rakoff-Nahoum, S., Cummings, R. D., & Stowell, S. R. (2023). Blood group A enhances SARS-CoV-2 infection. *Blood*, *142*(8), 742–747. https://doi.org/10.1182/blood.2022018903

Wu, Z., & McGoogan, J. M. (2020). Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA*, *323*(13), 1239–1242. https://doi.org/10.1001/jama.2020.2648

Xia, L., Wang, X., Liu, L., Fu, J., Xiao, W., Liang, Q., Han, X., Huang, S., Sun, L., Gao, Y., Zhang, C., Yang, L., Wang, L., Qian, L., & Zhou, Y. (2021). Lnc-BAZ2B promotes M2 macrophage activation and inflammation in children with asthma through stabilizing BAZ2B pre-mRNA. *Journal of Allergy and Clinical Immunology*, *147*(3), 921-932.e9. https://doi.org/10.1016/j.jaci.2020.06.034

Yadav, E., Yadav, N., Hus, A., & Yadav, J. S. (2020). Aquaporins in lung health and disease: Emerging roles, regulation, and clinical implications. *Respiratory Medicine*, *174*, 106193. https://doi.org/10.1016/j.rmed.2020.106193

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A. F., Heath, A. C., Martin, N. G., Montgomery, G. W., Weedon, M. N., Loos, R. J., Frayling, T. M., McCarthy, M. I., Hirschhorn, J. N., Goddard, M. E., & Visscher, P. M. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics*, *44*(4), 369-S3. https://doi.org/10.1038/ng.2213

Yang, Q., Tang, J., Cao, J., Liu, F., Fu, M., Xue, B., Zhou, A., Chen, S., Liu, J., Zhou, Y., Shi, Y., Peng, W., & Chen, X. (2023). SARS-CoV-2 infection activates CREB/CBP in cellular cyclic AMP-dependent pathways. *Journal of Medical Virology*, *95*(1), e28383. https://doi.org/10.1002/jmv.28383

Yang, Z., Cieza, B., Reyes-Dumeyer, D., Montesinos, R., Soto-Añari, M., Custodio, N., & Tosto, G. (2024). A benchmark study on current GWAS models in admixed populations. *Briefings in Bioinformatics*, *25*(1), bbad437. https://doi.org/10.1093/bib/bbad437

Yao, D. W., O'Connor, L. J., Price, A. L., & Gusev, A. (2020). Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nature Genetics*, *52*(6), Article 6. https://doi.org/10.1038/s41588-020-0625-2

Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliasen, A. U., Jiang, Y., Raghavan, S., Miao, J., Arias, J. D., Graham, S. E., Mukamel, R. E., Spracklen, C. N., Yin, X., Chen, S.-H., Ferreira, T., Highland, H. H., … Hirschhorn, J. N. (2022). A saturated map of common genetic variants associated with human height. *Nature*, *610*(7933), Article 7933. https://doi.org/10.1038/s41586-022-05275-y

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Yoneyama, M., Suhara, W., Fukuhara, Y., Fukuda, M., Nishida, E., & Fujita, T. (1998). Direct triggering of the type I interferon system by virus infection: Activation of a transcription factor complex containing IRF-3 and CBP/p300. *The EMBO Journal*, *17*(4), 1087–1095. https://doi.org/10.1093/emboj/17.4.1087

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), Article 2. https://doi.org/10.1038/ng1702

Yuan, J., Chang, S.-Y., Yin, S.-G., Liu, Z.-Y., Cheng, X., Liu, X.-J., Jiang, Q., Gao, G., Lin, D.-Y., Kang, X.-L., Ye, S.-W., Chen, Z., Yin, J.-A., Hao, P., Jiang, L., & Cai, S.-Q. (2020). Two conserved epigenetic regulators prevent healthy ageing. *Nature*, *579*(7797), Article 7797. https://doi.org/10.1038/s41586-020-2037-y

Zeberg, H., & Pääbo, S. (2020). The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature*, *587*(7835), Article 7835. 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. *Proceedings of the National Academy of Sciences*, *118*(9), e2026309118. https://doi.org/10.1073/pnas.2026309118

Zhang, H., Zheng, H., Zhu, J., Dong, Q., Wang, J., Fan, H., Chen, Y., Zhang, X., Han, X., Li, Q., Lu, J., Tong, Y., & Chen, Z. (2021). Ubiquitin-Modified Proteome of SARS-CoV-2-Infected Host Cells Reveals Insights into Virus-Host Interaction and Pathogenesis. *Journal of Proteome Research*, *20*(5), 2224–2239. https://doi.org/10.1021/acs.jproteome.0c00758

Zhang, Q., Privé, F., Vilhjálmsson, B., & Speed, D. (2021). Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-24485-y

Zhang, Y., Garner, R., Salehi, S., La Rocca, M., & Duncan, D. (2021). Association between ABO blood types and coronavirus disease 2019 (COVID-19), genetic associations, and underlying molecular mechanisms: A literature review of 23 studies. *Annals of Hematology*, *100*(5), 1123–1132. https://doi.org/10.1007/s00277-021-04489-w

Zhang, Y., & Pan, W. (2015). Principal Component Regression and Linear Mixed Model in Association Analysis of Structured Samples: Competitors or Complements? *Genetic Epidemiology*, *39*(3), 149–155. https://doi.org/10.1002/gepi.21879

Zhao, J., Yang, Y., Huang, H., Li, D., Gu, D., Lu, X., Zhang, Z., Liu, L., Liu, T., Liu, Y., He, Y., Sun, B., Wei, M., Yang, G., Wang, X., Zhang, L., Zhou, X., Xing, M., & Wang, P. G. (2021). Relationship Between the ABO Blood Group and the Coronavirus Disease 2019 (COVID-19) Susceptibility. *Clinical Infectious Diseases*, *73*(2), 328–331. https://doi.org/10.1093/cid/ciaa1150

Zheng, H., Rao, A. M., Dermadi, D., Toh, J., Jones, L. M., Donato, M., Liu, Y., Su, Y., Dai, C. L., Kornilov, S. A., Karagiannis, M., Marantos, T., Hasin-Brumshtein, Y., He, Y. D., Giamarellos-Bourboulis, E. J., Heath, J. R., & Khatri, P. (2021). Multi-cohort analysis of host immune response identifies conserved protective and detrimental modules associated with severity across viruses. *Immunity*, *54*(4), 753-768.e5. https://doi.org/10.1016/j.immuni.2021.03.002

Zheng, X., & Davis, J. W. (2021). SAIGEgds—An efficient statistical tool for large-scale PheWAS with mixed models. *Bioinformatics*, *37*(5), 728–730. https://doi.org/10.1093/bioinformatics/btaa731

Zhou, S., Butler-Laporte, G., Nakanishi, T., Morrison, D. R., Afilalo, J., Afilalo, M., Laurent, L., Pietzner, M., Kerrison, N., Zhao, K., Brunet-Ratnasingham, E., Henry, D., Kimchi, N., Afrasiabi, Z., Rezk, N., Bouab, M., Petitjean, L., Guzman, C., Xue, X., … Richards, J. B. (2021). A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nature Medicine*, *27*(4), 659–667. https://doi.org/10.1038/s41591-021-01281-1

Zhou, W., Kanai, M., Wu, K.-H. H., Rasheed, H., Tsuo, K., Hirbo, J. B., Wang, Y., Bhattacharya, A., Zhao, H., Namba, S., Surakka, I., Wolford, B. N., Lo Faro, V., Lopera-Maya, E. A., Läll, K., Favé, M.-J., Partanen, J. J., Chapman, S. B., Karjalainen, J., … Neale, B. M. (2022). Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. *Cell Genomics*, *2*(10), 100192. https://doi.org/10.1016/j.xgen.2022.100192

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., Bastarache, L. A., Wei, W.-Q., Denny, J. C., Lin, M., Hveem, K., Kang, H. M., Abecasis, G. R., Willer, C. J., & Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, *50*(9), Article 9. https://doi.org/10.1038/s41588-018-0184-y

Zhu, C., Ming, M. J., Cole, J. M., Edge, M. D., Kirkpatrick, M., & Harpak, A. (2023). Amplification is the primary mode of gene-by-sex interaction in complex human traits. *Cell Genomics*, *3*(5), 100297. https://doi.org/10.1016/j.xgen.2023.100297

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., Tan, W., & China Novel Coronavirus Investigating and Research Team. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *The New England Journal of Medicine*, *382*(8), 727–733. https://doi.org/10.1056/NEJMoa2001017

Zhu, X., Trimarco, J. D., Williams, C. A., Barrera, A., Reddy, T. E., & Heaton, N. S. (2022). ZBTB7A promotes virus-host homeostasis during human coronavirus 229E infection. *Cell Reports*, *41*(4), 111540. https://doi.org/10.1016/j.celrep.2022.111540

Zhu, Y., Yang, Y., Liu, Y., Qian, H., Qu, G., Shi, W., & Liu, J. (2023). A novel tetraspanin-related gene signature for predicting prognosis and immune invasion status of lung adenocarcinoma. *Journal of Cancer Research and Clinical Oncology*, *149*(15), 13631–13643. https://doi.org/10.1007/s00432-023-05176-1

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

10 ANEXOS

Anexo I. Rexións asociadas con significación GWAS nas principais análises GWAS. Estes sinais non foron necesariamente replicados noutros estudos. O fenotipo de COVID-19 crítica engloba tanto o fenotipo definido con fallo respiratorio e/ou morte e hospitalización. Nalgúns casos, dous campos poden pertencer á mesma rexión cromosómica, mais foron reportados como sinais separados. Se ben se tratou de colapsar as rexións adxacentes, o elevado número de xenes dificultou este traballo.

Cromosoma	Xenes	Fenotipo	Referencia
1	*TRIM46, THBS3, MUC1, MTX1, KRTCAP2, EFNA1, EFNA4, EFNA3, HCN3, SLC50A1, DPM3, KRTCAP2, GBA*	COVID-19 crítica; susceptibilidade	Kousathanas et al. (2022); Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
1	*CYP4B1*	COVID-19 crítica	Kanai et al. (2023)
1	*JAK1*	COVID-19 crítica	Pairo-Castineira et al. (2023); Kanai et al. (2023)
1	*AK5*	COVID-19 crítica	Pairo-Castineira et al. (2023); Kanai et al. (2023)
1	*SLC2A5*	COVID-19 crítica	Pairo-Castineira et al. (2023)
2	*BCL11A*	COVID-19 crítica; susceptibilidade	Kousathanas et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
2	*CIB4, C2orf70*	COVID-19 crítica	Kanai et al. (2023)
3	*CCR1, CCR2, CCR3, CCR9, CXCR6, FYCO1, LARS2, LIMD1, LZTFL1, SACM1L, XCR1*	COVID-19 crítica	The Severe COVID-19 GWAS Group (2020); The GenOMICC Investigators (2020); The COVID-19 Host Genetics Initiative (2021); Shelton et al. (2021); Horowitz et al. (2022); Kousathanas et al. (2022); Cruz et al. (2022); Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
2	*BAZ2B, LY75, WDSUB1, CD302, PLA2R1*	COVID-19 crítica	Diz-de Almeida et al. (2024)
3	*SLC6A20, SACM1L*	Susceptibilidade	COVID-19 Host Genetics Initiative (2021); Kousathanas et al. (2022); Horowitz et al. (2022); Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Cromosoma	Xenes	Fenotipo	Referencia
3	*RPL24, NXPE3, CEP97, ZBTB11*	COVID-19 crítica; susceptibilidade	COVID-19 Host Genetics Initiative (2021); Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
3	*PLSCR1*	COVID-19 crítica	Kousathanas et al. (2022); Pairo-Castineira et al. (2023)
3	*MUC4, MUC20-OT1, MUC20*	Susceptibilidade	Kanai et al. (2023)
4	*UGT2A1/UGT2A2*	Anosmia	Shelton et al. (2022)
4	*PI4K2B*	COVID-19 crítica	Kanai et al. (2023)
4	*SLC34A2, ANAPC4*	COVID-19 crítica	Pathak et al. (2022); Pairo-Castineira et al. (2023)
4	*SLC39A8*	Susceptibilidade	Pathak et al. (2022)
4	*NPNT, ARHGEF38*	COVID-19 crítica	Pathak et al. (2022); Pairo-Castineira et al. (2023)
5	*SLC22A5, IRF1, ACSL6*	COVID-19 crítica	Kousathanas et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
5	*DOCK2*	COVID-19 crítica	Namkoong et al. (2022)
6	*HLA-G*	COVID-19 crítica	The GenOMICC Investigators (2020); Kanai et al. (2023)
6	*HLA-A, HLA-C, HLA-DQA2, MICA, MICB, CCHCR1, TCF19, POU5F1*	COVID-19 crítica	The GenOMICC Investigators (2020); Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
6	*HLA-DPB1, HLA-DPA1*	COVID-19 crítica	Pathak et al. (2022); Kanai et al. (2023)
6	*HLA-DRB1, HLA-DQA1*	COVID-19 crítica	Kousathanas et al. (2022); Pairo-Castineira et al. (2023)
6	*COL11A2*	Susceptibilidade	Kanai et al. (2023)
6	*NOTCH4*	COVID-19 crítica	The GenOMICC Investigators (2020)
6	*FOXP4*	COVID-19 crítica; susceptibilidade	COVID-19 Host Genetics Initiative (2021); Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
6	*LTA*	COVID-19 crítica	Pairo-Castineira et al. (2023)
7	*TOMM7, SNHG26*	COVID-19 crítica	Kanai et al. (2023)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Cromosoma	Xenes	Fenotipo	Referencia
7	*POM121C, HIP1*	COVID-19 crítica	Pairo-Castineira et al. (2023); Kanai et al. (2023)
7	*ZKSCAN1, TRIM4, PILRB*	COVID-19 crítica	Pairo-Castineira et al. (2023); Kanai et al. (2023)
8	*TMEM65*	COVID-19 crítica	COVID-19 Host Genetics Initiative (2021)
8	*RAB2A, LINC01301, CA8*	COVID-19 crítica	Pairo-Castineira et al. (2023); Kanai et al. (2023)
9	*LINC01276*	COVID-19 crítica	Kousathanas et al. (2022)
9	*ABO*	Susceptibilidade	The Severe COVID-19 GWAS Group (2020); Shelton et al. (2021); COVID-19 Host Genetics Initiative (2021); Horowitz et al. (2022); Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
9	*SNAPC3, PSIP1, CCDC171*	COVID-19 crítica	Kanai et al. (2023)
9	*IFNA7, IFNA6, IFNA5, IFNA4, IFNA21, IFNA17, IFNA16, IFNA14, IFNA10*	COVID-19 crítica	Kousathanas et al. (2022); Pairo-Castineira et al. (2023)
9	*AQP3*	COVID-19 crítica	Cruz et al. (2022); Kanai et al. (2023); Pairo-Castineira et al. (2023)
10	*TMEM254, SFTPD, SFTPA2, SFTPA1, PLAC9, NUTM2E, NUTM2B*	COVID-19 crítica; susceptibilidade	Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
10	*TCF7L2*	COVID-19 crítica	Kanai et al. (2023)
11	*ELF5*	COVID-19 crítica; susceptibilidade	Kousathanas et al. (2022); Pathak et al. (2022); Cruz et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
11	*MUC5B*	COVID-19 crítica	Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
11	*DDIAS, PRCP*	COVID-19 crítica	Diz-de Almeida et al. (2024)
12	*OAS1-OAS3*	COVID-19 crítica	The GenOMICC Investigators (2020); COVID-19 Host Genetics Initiative (2021); Pathak et al. (2022); Cruz et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Cromosoma	Xenes	Fenotipo	Referencia
12	*FBRSL1*	COVID-19 crítica	Kousathanas et al. (2022); Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
13	*ATP11A*	COVID-19 crítica	Kousathanas et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
14	*FOXG1*	Susceptibilidade	Kanai et al. (2023)
15	*RGMA*	COVID-19 crítica	Kousathanas et al. (2022)
16	*SLC22A31, ZNF778, CDH15, ACSF3*	COVID-19 crítica	Kousathanas et al. (2022); Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
16	*CREBBP*	COVID-19 crítica	Diz-de Almeida et al. (2024)
17	*TAC4, DLX3, FLJ45513, KAT7*	COVID-19 crítica	COVID-19 Host Genetics Initiative (2021); Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
17	*KANSL1, ARHGAP27, PLEKHM1, CRHR1, SPPL2C, MAPT, STH, LRRC37A, ARL17B, ARL17A, NSF, WNT3*	COVID-19 crítica	COVID-19 Host Genetics Initiative (2021); Kousathanas et al. (2022); Pathak et al. (2022); Cruz et al. (2022); Degenhardt et al., (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
17	*PSMD3, MED24, GSDMA, CSF3*	Susceptibilidade	Pairo-Castineira et al. (2023); Kanai et al. (2023)
19	*DPP9*	COVID-19 crítica; susceptibilidade	The GenOMICC Investigators (2020); COVID-19 Host Genetics Initiative (2021); Kousathanas et al. (2022); Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
19	*TYK2, RAVER1, ICAM5, ICAM1, ICAM4, ZGLP1, FDX2, ICAM3*	COVID-19 crítica	The GenOMICC Investigators (2020); COVID-19 Host Genetics Initiative (2021); Kousathanas et al. (2022); Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
19	*PLEKHA4, TULP2, HSD17B14, PPP1R15A, NUCB1*	Susceptibilidade	COVID-19 Host Genetics Initiative (2021); Pathak et al. (2022)
19	*ZGLP1, PDE4A*	COVID-19 crítica	Kousathanas et al. (2022); Pairo-Castineira et al. (2023)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Cromosoma	Xenes	Fenotipo	Referencia
19	*FUT2, SEC1P, RASIP1, NTN5, MAMSTR, IZUMO1*	COVID-19 crítica; susceptibilidade	Kousathanas et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
19	*UPK1A/ZBTB32*	COVID-19 crítica	Cruz et al. (2022)
19	*MUC16*	Susceptibilidade	Kanai et al. (2023)
19	*FOXA3*	COVID-19 crítica	Kanai et al. (2023)
19	*POLD1, NAPSA, NR1H2, NAPSB, KCNC3*	COVID-19 crítica; susceptibilidade	Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
19	*ZBTB7A*	COVID-19 crítica	Pairo-Castineira et al. (2023); Diz-de Almeida et al. (2024)
20	*CASC20*	COVID-19 crítica	Pairo-Castineira et al. (2023); Diz-de Almeida et al. (2024)
21	*IFNAR2, IL10RB*	COVID-19 crítica	The GenOMICC Investigators (2020); COVID-19 Host Genetics Initiative (2021); Kousathanas et al. (2022); Pathak et al. (2022); Cruz et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
21	*LINC00649*	COVID-19 crítica	Kousathanas et al. (2022)
21	*AT5PO, IL10RB, MRPS6*	COVID-19 crítica	Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)
21	*TMPRSS2, FAM3B*	COVID-19 crítica	Pairo-Castineira et al. (2023); Kanai et al. (2023)
X	*ACE2, PIR, BMX*	Susceptibilidade	Horowitz et al. (2022); Pathak et al. (2022); Pairo-Castineira et al. (2023); Kanai et al. (2023)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Anexo II. Xenes priorizados por FUMA na meta-análise SCOURGE-CNIO (hg37). Symbol: simbolo do xen por HGNC; start: inicio do xen; end: final do xen; pLi: puntuación pLI segundo ExAC (probabilidade de ser intolerante a perda de función; canto máis alto, máis intolerante); ncRVIS: puntuación de intolerancia á variación residual non codificante (canto máis alto, máis intolerante); pos Map SNPS: nº de SNPs asignados por mapeo posicional; pos Map Max CADD: a puntuación CADD máxima dos SNPs asignados por mapeo posicional; eqtl Map SNPs: nº de SNPs asignados por mapeo de eQTL; eqtl Map min P: o p-valor mínimo de eQTL para os SNPs asignados; eqtl Map min Q: o q-valor mínimo por FDR de eQTL para os SNPs asignados; eqtl Maps: tecido dos eQTL mapeados; eqtl Direction: dirección da consecuencia dos SNPs mapeados a eQTL tras aliñar os alelos de risco; min Gwas P: p-valor mínimo no GWAS dos eQTL mapeados; Ind Sig SNPs: SNPs independentes que están en LD cos SNPs mapeados.

symbol	chr	start	end	type	pLI	ncRVIS	Pos Map SNPs	Pos Map Max CADD	Eqtl Map SNPs	Eqtl Map minP	Eqtl Map minQ	Eqtl Mapts	Eqtl Dir	Min Gwas P	Ind Sig SNPs
LIMD1	3	45596886	45727830	protein_coding	0.12777068	0.27357959	3	3.502	0	NA	NA	NA	NA	6.93E-09	rs17213127
AC099539.1	3	45633747	45633819	miRNA	NA	NA	1	0.088	0	NA	NA	NA	NA	1.05E-08	rs17213127
SLC6A20	3	45796942	45838027	protein_coding	6.45E-10	0.25155855	7	4.173	0	NA	NA	NA	NA	4.06E-29	rs17213127; rs17763742
LZTFL1	3	45864808	45957534	protein_coding	0.19410281	-1.16298278	36	13.68	0	NA	NA	NA	NA	9.62E-29	rs17763742; rs34338823; rs35951367; rs4493469
RP11-852E15.1	3	45925132	45925605	pseudogene	NA	NA	5	6.254	0	NA	NA	NA	NA	1.59E-18	rs34338823; rs35951367
CCR9	3	45927996	45944667	protein_coding	0.00464031	-0.45842799	12	8.293	20	0.00013873	0.00822783	Whole_Blood	+	9.62E-29	rs34338823; rs35951367; rs17763742; rs115102354; rs4493469
Y_RNA	3	45944679	45944790	misc_RNA	NA	NA	7	8.293	0	NA	NA	NA	NA	2.59E-18	rs34338823; rs35951367
FYCO1	3	45959396	46037316	protein_coding	9.67E-13	1.65562431	68	14.93	140	1.80E-05	1.55E-13	Esophagus_Mucosa	+	1.59E-18	rs34338823; rs4493469; rs35951367
CXCR6	3	45982425	45989845	protein_coding	6.15E-05	-0.16660599	20	11.81	114	2.12E-05	0.0243627	Lung	-	1.39E-17	rs34338823; rs35951367; rs4493469; rs115102354
XCR1	3	46058516	46069234	protein_coding	0.09785415	-0.14518301	57	11.83	0	NA	NA	NA	NA	4.42E-16	rs34338823; rs35951367; rs4493469
NRBF2P2	3	46064788	46065648	pseudogene	NA	NA	42	11.83	0	NA	NA	NA	NA	7.83E-16	rs34338823; rs35951367; rs4493469
FLT1P1	3	46183850	46185169	pseudogene	NA	NA	41	8.91	352	1.76E-11	1.01E-11	Whole_Blood	-	1.59E-18	rs34338823; rs35951367; rs4493469; rs115102354

symbol	chr	start	end	type	pLI	ncRVIS	Pos Map SNPs	Pos Map Max CADD	Eqtl Map SNPs	Eqtl Map minP	Eqtl Map minQ	Eqtl Mapts	Eqtl Dir	Min Gwas P	Ind Sig SNPs
CCR3	3	46205096	46308197	protein_coding	0.03189305	-0.54102914	115	14.08	209	7.33E-07	1.56E-30	Whole_Blood	-	2.48E-15	rs4493469; rs115102354; rs34338823; rs35951367
CCR1	3	46243200	46249887	protein_coding	0.89182875	1.89189807	40	14.08	302	4.47E-12	1.25E-08	Esophagus_Mucosa; Lung	+	4.06E-29	rs4493469; rs115102354; rs17763742; rs34338823; rs35951367
CCR2	3	46395225	46402419	protein_coding	0.01146332	-0.03550416	2	0.682	0	NA	NA	NA	NA	3.24E-07	rs115102354; rs4493469
RP11-24F11.2	3	46406446	46448550	antisense	NA	NA	1	0.682	0	NA	NA	NA	NA	NA	rs4493469; rs115102354
CCR5	3	46411633	46417697	protein_coding	3.51E-08	1.48148893	1	0.682	39	4.95E-06	0.0058577	Lung	-	1.44E-14	rs4493469; rs115102354; rs17213127; rs34338823; rs35951367
NFX1	9	33290509	33371155	protein_coding	0.38644124	-0.49607242	0	0	27	7.81E-06	9.42E-17	Whole_Blood	+	2.67E-08	rs10813976
AQP7	9	33384765	33402643	protein_coding	1.86E-09	-0.12464553	0	0	27	1.33E-15	2.27E-51	Esophagus_Mucosa; Lung	+	2.67E-08	rs10813976
RP11-311H10.4	9	33402855	33409946	antisense	NA	NA	1	3.851	0	NA	NA	NA	NA	NA	rs10813976
AQP3	9	33441152	33447609	protein_coding	0.50482751	0.14972492	0	0	27	4.11E-21	6.22E-36	Esophagus_Mucosa; Lung	+	2.67E-08	rs10813976
SPATA32	17	43331760	43339479	protein_coding	2.95E-05	-0.4147557	0	0	13	1.56E-05	3.73E-35	Whole_Blood	+	2.13E-08	rs1230082
ARHGAP27	17	43471275	43511787	protein_coding	0.00125584	-0.76162779	23	10.9	30	7.05E-08	1.71E-13	Esophagus_Mucosa	+	2.13E-08	rs1230082
CTB-39G8.3	17	43474298	43474843	antisense	NA	NA	11	10.9	30	4.65E-17	1.89E-13	Whole_Blood; Esophagus_Mucosa; Lung	+	2.13E-08	rs1230082
AC091132.2	17	43494179	43494272	miRNA	NA	NA	7	3.176	0	NA	NA	NA	NA	2.13E-08	rs1230082
PLEKHM1	17	43513266	43568115	protein_coding	0.84054321	-0.66306803	16	10.46	28	8.55E-08	1.42E-20	Whole_Blood; Esophagus_Mucosa	+	2.13E-08	rs1230082
AC091132.1	17	43530210	43541431	antisense	NA	NA	5	6.244	3	0.00014778	3.34E-09	Esophagus_Mucosa	+	3.30E-08	rs1230082
RN7SL730P	17	43547899	43548188	misc_RNA	NA	NA	8	6.244	0	NA	NA	NA	NA	3.76E-07	rs1230082
LRRC37A4P	17	43578685	43627701	pseudogene	NA	NA	1	1.269	30	7.90E-21	4.97E-169	Cells_EBV-transformed_lymphocytes; Whole_Blood; Esophagus_Mucosa; Lung	+	2.13E-08	rs1230082
RP11-798G7.5	17	43580626	43612076	antisense	NA	NA	1	1.269	0	NA	NA	NA	NA	NA	rs1230082
RP11-798G7.8	17	43608943	43611204	lincRNA	NA	NA	0	0	1	0.00011981	2.26E-27	Esophagus_Mucosa	NA	NA	rs1230082

symbol	chr	start	end	type	pLI	ncRVIS	Pos Map SNPs	Pos Map Max CADD	Eqtl Map SNPs	Eqtl Map minP	Eqtl Map minQ	Eqtl Mapts	Eqtl Dir	Min Gwas P	Ind Sig SNPs
DND1P1	17	43663237	43664295	pseudogene	NA	NA	0	0	29	2.70E-11	3.31E-124	Cells_EBV-transformed_lymphocytes; Whole_Blood; Esophagus_Mucosa; Lung	-	2.13E-08	rs1230082
RP11-707O23.5	17	43678235	43679706	pseudogene	NA	NA	0	0	29	6.66E-15	4.08E-148	Cells_EBV-transformed_lymphocytes; Whole_Blood; Esophagus_Mucosa; Lung	-	2.13E-08	rs1230082
CRHR1-IT1	17	43697694	43725582	pseudogene	NA	NA	0	0	29	2.36E-13	6.57E-143	Cells_EBV-transformed_lymphocytes; Whole_Blood; Esophagus_Mucosa; Lung	-	2.13E-08	rs1230082
MAPT	17	43971748	44105700	protein_coding	4.68E-05	2.01024275	0	0	29	4.74E-08	3.48E-64	Esophagus_Mucosa; Lung	-	2.13E-08	rs1230082
KANSL1	17	44107282	44302733	protein_coding	0.99973351	0.05894525	0	0	21	9.85E-06	4.18E-29	Whole_Blood; Esophagus_Mucosa	-	3.30E-08	rs1230082
KANSL1-AS1	17	44270942	44274089	antisense	NA	NA	0	0	30	5.53E-17	1.03E-145	Whole_Blood; Esophagus_Mucosa; Lung	-	2.13E-08	rs1230082
RP11-259G18.2	17	44320972	44322410	pseudogene	NA	NA	0	0	30	1.90E-14	1.66E-126	Whole_Blood; Esophagus_Mucosa; Lung	-	2.13E-08	rs1230082
RP11-259G18.3	17	44336917	44337972	pseudogene	NA	NA	0	0	29	3.93E-15	3.04E-138	Cells_EBV-transformed_lymphocytes; Whole_Blood; Esophagus_Mucosa; Lung	-	2.13E-08	rs1230082
RP11-259G18.1	17	44344403	44346060	pseudogene	NA	NA	0	0	29	4.22E-07	7.26E-23	Esophagus_Mucosa	-	2.13E-08	rs1230082
LRRC37A	17	44370099	44415160	protein_coding	0.89048736	NA	0	0	30	3.60E-10	4.65E-62	Whole_Blood; Esophagus_Mucosa; Lung	-	2.13E-08	rs1230082
SCN1B	19	35521588	35531352	protein_coding	0.20516694	NA	0	0	5	3.57E-05	0.0383851	Esophagus_Mucosa	+	7.07E-08	rs77127536
ETV2	19	36132647	36135773	protein_coding	0.0328968	-0.43683926	0	0	20	2.40E-05	0.00040446	Whole_Blood	-	1.28E-08	rs77127536
UPK1A	19	36157715	36169367	protein_coding	3.36E-05	-0.10274217	25	11.77	21	1.28E-07	1.29E-102	Whole_Blood; Esophagus_Mucosa; Lung	-	1.28E-08	rs77127536
UPK1A-AS1	19	36158850	36164193	antisense	NA	NA	17	11.77	0	NA	NA	NA	NA	7.07E-08	rs77127536
AC002314.4	19	36174680	36175103	pseudogene	NA	NA	27	11.77	0	NA	NA	NA	NA	1.28E-08	rs77127536
KMT2B	19	36208921	36229779	protein_coding	1	NA	0	0	3	5.38E-05	1.42E-08	Whole_Blood	-	1.28E-08	rs77127536
IGFLR1	19	36230058	36233354	protein_coding	0.00016488	0.22818063	0	0	21	1.17E-05	0.00015332	Esophagus_Mucosa	+	1.28E-08	rs77127536
U2AF1L4	19	36233365	36236346	protein_coding	5.80E-07	-0.20856825	0	0	4	1.89E-05	9.65E-81	Whole_Blood; Lung	+	1.28E-08	rs77127536
ARHGAP33	19	36265434	36279724	protein_coding	0.21224841	NA	0	0	21	3.01E-10	2.56E-35	Whole_Blood; Esophagus_Mucosa	+	1.28E-08	rs77127536
AC002398.5	19	36279778	36288787	lincRNA	NA	NA	0	0	21	3.65E-09	3.16E-26	Esophagus_Mucosa	+	1.28E-08	rs77127536

symbol	chr	start	end	type	pLI	ncRVIS	Pos Map SNPs	Pos Map Max CADD	Eqtl Map SNPs	Eqtl Map minP	Eqtl Map minQ	Eqtl Mapts	Eqtl Dir	Min Gwas P	Ind Sig SNPs
IFNAR2	21	34602206	34637980	protein_coding	0.00506603	-0.39757227	24	10.94	19	6.73E-16	4.76E-12	Whole_Blood; Esophagus_Mucosa; Lung	-	2.29E-11	rs2834164; rs17860169
AP000295.9	21	34619079	34655517	protein_coding	NA	NA	16	10.94	0	NA	NA	NA	NA	2.29E-11	rs17860169; rs2834164
IL10RB-AS1	21	34636178	34638565	antisense	NA	NA	1	1.797	1	0.00013094	9.98E-08	Whole_Blood	NA	NA	rs17860169; rs2834164
IL10RB	21	34638663	34669539	protein_coding	0.00033479	0.1555495	1	1.797	16	1.47E-05	3.06E-11	Lung	+	2.29E-11	rs17860169

Anexo III. Táboa dos SNPs incluídos nos sets cribles ao 95% para as rexións dos cromosomas 9 e 19 identificadas na meta-análise de SCOURGE-CNIO.

CHR	POS	SNP	AE	ANE	Beta	SE	Pval
9	33421222	9:33421222:C:T	T	C	0.1729	0.0321	6.98E-08
9	33422071	9:33422071:C:T	T	C	0.17	0.0319	1.02E-07
9	33422159	9:33422159:A:G	A	G	-0.1702	0.0319	9.79E-08
9	33423066	9:33423066:T:C	T	C	-0.1718	0.0319	7.19E-08
9	33423128	9:33423128:G:A	A	G	0.1719	0.0318	6.18E-08
9	33424735	9:33424735:C:T	T	C	0.1736	0.0317	4.52E-08
9	33425025	9:33425025:C:G	C	G	-0.1761	0.0318	2.93E-08
9	33425186	9:33425186:GTAAC:G	G	GTAAC	0.1739	0.0318	4.33E-08
9	33425462	9:33425462:T:C	T	C	-0.1748	0.0317	3.65E-08
9	33425501	9:33425501:G:A	A	G	0.1754	0.0318	3.44E-08
9	33425570	9:33425570:G:A	A	G	0.1748	0.0317	3.65E-08
9	33425586	9:33425586:G:T	T	G	0.1738	0.0318	4.42E-08
9	33425787	9:33425787:T:C	T	C	-0.1752	0.0318	3.56E-08
9	33425873	9:33425873:T:C	T	C	-0.1746	0.0318	3.87E-08
9	33425965	9:33425965:G:A	A	G	0.1739	0.0317	4.31E-08
9	33425989	9:33425989:G:A	A	G	0.1757	0.0318	3.18E-08
9	33426044	9:33426044:G:A	A	G	0.1749	0.0317	3.57E-08
9	33426457	9:33426457:T:G	T	G	-0.1733	0.0314	3.40E-08
9	33426485	9:33426485:A:G	A	G	-0.1742	0.0317	4.06E-08
9	33426577	9:33426577:A:G	A	G	-0.1765	0.0317	2.67E-08
9	33426685	9:33426685:C:G	C	G	-0.1752	0.0318	3.64E-08
9	33426686	9:33426686:A:G	A	G	-0.1752	0.0318	3.64E-08
19	35678383	19:35678383:A:G	A	G	0.1933	0.037	1.71E-07
19	35678946	19:35678946:A:G	A	G	0.1939	0.0368	1.42E-07
19	35679535	19:35679535:A:G	A	G	0.194	0.0369	1.40E-07
19	35679559	19:35679559:A:G	A	G	0.1945	0.0369	1.31E-07
19	35679561	19:35679561:G:A	A	G	-0.1962	0.0369	1.03E-07
19	35681398	19:35681398:G:T	T	G	-0.1957	0.037	1.23E-07
19	35682421	19:35682421:T:C	T	C	0.196	0.037	1.16E-07
19	35682545	19:35682545:A:C	A	C	0.1994	0.037	7.07E-08
19	35682551	19:35682551:C:A	A	C	-0.1967	0.037	1.07E-07
19	35683274	19:35683274:T:C	T	C	0.1938	0.0368	1.45E-07
19	35683294	19:35683294:C:T	T	C	-0.1937	0.0369	1.48E-07
19	35685424	19:35685424:A:G	A	G	0.1938	0.0368	1.45E-07
19	35686262	19:35686262:G:A	A	G	-0.194	0.0369	1.41E-07
19	35686632	19:35686632:G:A	A	G	-0.1962	0.0369	1.01E-07
19	35686844	19:35686844:C:T	T	C	-0.194	0.0369	1.42E-07
19	35687796	19:35687796:G:A	A	G	-0.2152	0.0378	1.28E-08
19	35689106	19:35689106:T:C	T	C	0.2129	0.0376	1.51E-08
19	35689476	19:35689476:T:C	T	C	0.2118	0.0376	1.75E-08

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Anexo IV. Resultados de S-MultiXcan para a meta-análise SCOURGE-CNIO. N: número de tecidos dispoñibles para o xen; N indep: número de compoñentes independentes de variación mantidos entre as predicións dos tecidos; p_i_best: mellor p-valor da asociación S-prediXcan dun só tecido; t_i_best: nome do tecido da mellor asociación de S-prediXcan de un só tecido; p_i_worst: peor p-valor da asociación S-prediXcan dun só tecido; t_i_worst: nome do tecido da peor asociación de S-prediXcan dun só tecido; eigen max: autovalor da primeira compoñente independente na descomposición SVD da correlación de expresión predita (varianza explicada); eigen_min: autovalor da última compoñente independente na descomposición SVD da correlación de expresión predita; eigen_min_kept: autovalor da compoñente máis pequena que se mantivo na descomposición SVD da correlación de expresión predita; Z min: puntuación Z mínima entre as asociacións de S-prediXcan dun só tecido; Z max: puntuación Z máxima entre as asociacións de S-prediXcan dun só tecido; Z mean: puntuación Z media as asociacións de S-prediXcan dun só tecido; Z sd: desviación estándar da puntuación Z media entre as asociacións de S-prediXcan dun só tecido; tmi: traza de T*T', onde T é a correlación dos niveis de expresión preditos para diferentes tecidos multiplicada pola súa pseudoinversa SVD e é unha estimación do número de compoñentes independentes de variación na expresión predita entre tecidos. P-valor axustado por bonferroni para 22.052 tecidos (p<2.27x10-06). SVD: *single-value decomposition*.

gene	pvalue	N	N indep	p_i_best	t_i_best	p_i_worst	t_i_worst	Eigen max	Eigen min	Eigen min kept	Z min	Z max	Z mean	Z sd	tmi
CCR9	2.81E-24	7	6	1.71E-25	Whole Blood	0.8518071	Testis	3.35471907	0.00147328	0.11285377	0.18681319	10.4353609	3.51501601	4.29076841	6
LZTFL1	3.70E-17	45	5	1.19E-17	Brain Spinal cord cervical c-1	0.86733191	Nerve Tibial	22.0017646	7.17E-16	0.78636891	-3.33033689	8.55408528	-0.9312233	2.98061516	5
CXCR6	2.35E-15	38	4	2.98E-17	Adipose Visceral Omentum	0.96872342	Esophagus Mucosa	16.4694461	1.49E-15	3.05258147	-8.44731626	3.87824212	-2.1542229	3.4425153	4
FYCO1	2.79E-13	47	5	1.96E-13	Brain Hippocampus	0.77679792	Nerve Tibial	23.2228804	2.24E-15	0.82834188	-2.94684934	7.35159805	2.18114645	2.2045129	5
SLC6A20	1.93E-12	33	5	5.60E-23	Nerve Tibial	0.65328569	Brain Cerebellum	19.105379	6.61E-16	0.871829	-0.82609343	9.87032699	1.71644898	1.91996919	5
XCR1	5.93E-11	22	3	2.94E-05	Whole Blood	0.8719578	Esophagus Gastroesophageal Junction	12.8161524	1.12E-17	1.39271275	-4.05000019	4.17821789	-0.7452492	2.30987902	3
SPINK4	1.63E-09	6	2	5.82E-05	Stomach	0.94202315	Colon Sigmoid	5.49234555	4.39E-17	0.38863453	-0.41311133	4.02006674	0.44634687	1.75804983	2
WNT6	3.21E-08	9	3	0.01595524	Cells EBV-transformed lymphocytes	0.65142607	Brain Spinal cord cervical c-1	7.63547844	4.16E-17	0.36829551	-0.45178196	2.40993786	0.41155328	1.31620812	3
CCR1	6.26E-08	47	2	3.96E-07	Skin Not Sun Exposed Suprapubic	0.94153496	Artery Aorta	33.885561	7.77E-16	12.0002601	-4.86727479	5.07090902	0.30332767	2.99956716	2
HOXD10	2.08E-07	6	5	0.02883041	Lung	0.45124963	Colon Sigmoid	3.54634109	2.18E-19	0.14926892	-2.18579769	2.18579769	0.1613915	1.81776476	5
ZBTB32	6.05E-07	44	3	1.73E-07	Thyroid	0.0048844	Esophagus Mucosa	38.5345529	7.22E-16	1.92689629	-5.22612942	3.85161555	-4.28623712	1.40867945	3
DTX1	8.48E-07	43	6	1.23E-05	Brain Cerebellum	0.96927752	Muscle Skeletal	24.175949	1.04E-15	0.9284265	-2.77623129	4.37218046	-0.21676363	1.19190601	6

gene	pvalue	N	N indep	p_i_best	t_i_best	p_i_worst	t_i_worst	Eigen max	Eigen min	Eigen min kept	Z min	Z max	Z mean	Z sd	tmi
RP11-706P11.2	9.10E-07	29	6	0.01629091	Colon Transverse	0.54197042	Brain Frontal Cortex BA9	17.4269368	3.83E-16	0.66671465	-2.40233203	2.19630393	0.56489006	1.57196785	6
UPK1A	1.36E-06	43	4	1.47E-09	Brain Substantia nigra	0.7547589	Esophagus Mucosa	31.8295789	4.44E-16	1.50739745	-6.04743376	0.35123909	-4.14992046	1.41884051	4
CCR3	1.42E-06	6	4	7.39E-07	Thyroid	0.84977274	Whole Blood	3.53884676	8.93E-17	0.4622936	-4.95090914	-0.1894084	-2.28430812	2.19214339	4
ARHGAP27	2.49E-06	43	3	8.53E-08	Small Intestine Terminal Ileum	0.31781594	Muscle Skeletal	27.1576403	9.19E-16	2.13944345	-5.15743999	5.35554149	0.7142626	4.06829042	3

Anexo V. Resultados das análises de asociación dos SNPs líder nos GWAS de SCOURGE coas comorbilidades. Ensínase o valor do beta asociado ao SNP, do erro estándar (SE), e o p-valor (pval). AC: antecedentes cardiovasculares; AD: antecedentes dixestivos; AN: antecedentes nerviosos; AOH: antecedentes onco-hematolóxicos; AR: antecedentes respiratorios; AV: antecedentes vasculares.

SNP	beta AC	SE AC	pval AC	beta AD	SE AD	pval AD	beta AN	SE AN	pval AN	beta AOH	SE AOH	pval AOH	beta AR	SE AR	pval AR	beta AV	SE AV	pval AV
chr17.45422978.G.C	-4.528	4.667	0.332	-8.533	8.678	0.325	-2.017	5.436	0.711	2.4	5.514	0.663	0.823	4.789	0.863	3.87	3.402	0.255
chr19.35687796.G.A	-4.771	4.66	0.306	-9.329	8.695	0.283	-1.917	5.441	0.725	1.897	5.507	0.73	0.648	4.785	0.892	3.675	3.399	0.28
chr21.33240996.A.G	-3.956	4.699	0.4	-9.388	8.705	0.281	-1.808	5.464	0.741	1.469	5.551	0.791	0.742	4.813	0.877	3.651	3.418	0.285
chr3.45587795.G.A	-4.641	4.681	0.321	-9.26	8.704	0.287	-2.057	5.45	0.706	1.315	5.542	0.812	1.158	4.793	0.809	3.915	3.407	0.251
chr3.45805277.A.G	-4.49	4.692	0.339	-7.117	8.717	0.414	-3.023	5.467	0.58	2.647	5.552	0.633	1.271	4.819	0.792	4.212	3.426	0.219
chr3.45932600.G.A	-4.247	4.659	0.362	-8.533	8.675	0.325	-2.275	5.434	0.675	2.145	5.508	0.697	1.282	4.782	0.789	3.867	3.397	0.255
chr3.46136372.T.C	-5.704	4.7	0.225	-10.226	8.877	0.249	-2.744	5.464	0.616	1.606	5.531	0.771	0.398	4.814	0.934	3.252	3.416	0.341
chr3.46180545.A.G	-5.113	4.683	0.275	-6.761	8.79	0.442	-1.43	5.45	0.793	2.352	5.526	0.67	1.158	4.808	0.81	3.826	3.406	0.261
chr9.33426577.A.G	-5.301	4.673	0.257	-8.639	8.673	0.319	-1.463	5.48	0.789	1.684	5.518	0.76	1.478	4.789	0.758	3.835	3.41	0.261
chr9.81442264.TCAG.T	-0.042	0.134	0.753	-0.102	0.254	0.688	-0.025	0.134	0.851	-0.23	0.174	0.188	-0.019	0.139	0.892	-0.031	0.089	0.728

Anexo VI. Replicación, nas meta-análises SCOURGE-CNIO, das 40 variantes asociadas á hospitalización por COVID-19 identificadas polo consorcio HGI. AE: alelo de efecto na meta-análise, ANE: alelo de efecto na meta-análise, NEA: alelo de non efecto na meta-análise; dirección: dirección do efecto en cada un dos estudos (SCOURGE, CNIO); HetISq: estatístico I^2; HetChiSq: estatístico Q; HetDF: graos de liberdade do estatístico Q; HetPval: p-valor para o test de heteroxeneidade.

CHR	POS	SNP	AE	ANE	beta_meta	SE_meta	pval_meta	dirección	HetISq	HetChiSq	HetDf	HetPval
Meta-análise SCOURGE-CNIO do grupo completo												
1	46810098	1:46810098:T:C	T	C	0.0728	0.0278	0.008713	-+	42.2	1.732	1	0.1882
1	64947147	1:64947147:G:T	T	G	-0.0341	0.0294	0.2463	--	0	0.306	1	0.5799
1	77483438	1:77483438:G:A	A	G	0.0858	0.0282	0.002342	++	16.8	1.203	1	0.2728
1	155202934	1:155202934:T:C	T	C	0.1319	0.042	0.001673	++	0	0.192	1	0.661
2	26616850	2:26616850:G:T	T	G	0.0704	0.0728	0.3331	+?	0	0	0	1
2	60480453	2:60480453:A:G	A	G	0.0739	0.0725	0.3085	+?	0	0	0	1
3	45793925	3:45793925:G:A	A	G	0.1291	0.0474	0.006521	++	0	0.003	1	0.9597
3	45818159	3:45818159:G:A	A	G	0.5704	0.0514	1.42E-28	++	0	0.694	1	0.4049
3	101647854	3:101647854:G:A	A	G	-0.0567	0.0288	0.04896	--	0	0.527	1	0.4677
4	25312372	4:25312372:A:G	A	G	-0.1158	0.0623	0.06293	--	0	0.074	1	0.7858
4	25447603	4:25447603:G:A	A	G	-0.0602	0.0829	0.4682	-?	0	0	0	1
4	105897896	4:105897896:G:A	A	G	-0.0692	0.033	0.0359	--	0	0.659	1	0.4169
5	132441275	5:132441275:T:C	T	C	-0.076	0.0372	0.04112	--	0	0.782	1	0.3765
6	29947491	6:29947491:C:T	T	C	0.0956	0.0961	0.3196	+?	0	0	0	1
6	31182658	6:31182658:A:G	A	G	0.1328	0.0348	0.000137	++	0	0.234	1	0.6286
6	32714422	6:32714422:T:C	T	C	0.1361	0.0698	0.05129	++	0	0.05	1	0.8238
6	41520640	6:41520640:G:A	A	G	0.0273	0.0885	0.7574	++	0	0.011	1	0.9175
7	22854868	7:22854868:T:C	T	C	0.0682	0.0272	0.01201	-+	17.7	1.215	1	0.2704
7	100032719	7:100032719:C:T	T	C	0.0281	0.0271	0.3004	++	49.8	1.991	1	0.1583
8	60513412	8:60513412:G:A	A	G	0.0544	0.0271	0.04454	-+	0	0.704	1	0.4013
9	15795835	9:15795835:G:T	T	G	0.0538	0.0636	0.3975	-+	0	0.704	1	0.4014
9	33425186	9:33425186:GTAAC:G	G	GTAAC	0.1739	0.0318	4.33E-08	++	0	0.347	1	0.5556
9	133273813	9:133273813:C:T	T	C	-0.0428	0.0279	0.1255	--	0	0.315	1	0.5747

CHR	POS	SNP	AE	ANE	beta_meta	SE_meta	pval_meta	dirección	HetISq	HetChiSq	HetDf	HetPval
10	79977061	10:79977061:G:A	A	G	0.035	0.0419	0.4033	-+	0	0.324	1	0.5691
10	112972548	10:112972548:C:A	A	C	-0.056	0.0314	0.07431	--	0	0.725	1	0.3945
11	1219991	11:1219991:G:T	T	G	-0.0989	0.0404	0.01443	--	0	0.008	1	0.9278
11	34480495	11:34480495:C:T	T	C	-0.1341	0.0269	6.28E-07	--	30.8	1.446	1	0.2292
12	112922758	12:112922758:T:C	T	C	-0.1475	0.0283	1.79E-07	--	0	0.3	1	0.5837
12	132565387	12:132565387:T:C	T	C	0.061	0.0268	0.02284	++	63.4	2.736	1	0.09812
13	112881427	13:112881427:C:T	T	C	0.0898	0.0324	0.005549	-+	71.1	3.463	1	0.06274
16	89196249	16:89196249:G:A	A	G	0.1159	0.0391	0.003017	++	0	0.054	1	0.8163
17	45635098	17:45635098:G:A	A	G	-0.0941	0.0291	0.001223	--	0	0.028	1	0.8676
17	49863260	17:49863260:C:A	A	C	0.2282	0.0871	0.008804	++	0	0.188	1	0.6647
19	4717660	19:4717660:A:G	A	G	-0.2013	0.0776	0.009501	-?	0	0	0	1
19	10352442	19:10352442:G:C	C	G	0.2159	0.2054	0.2931	+?	0	0	0	1
19	45869791	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
19	48702888	19:48702888:G:C	C	G	-0.0313	0.0268	0.2427	+-	0	0.778	1	0.3777
19	50379362	19:50379362:T:C	T	C	-0.1129	0.0286	7.76E-05	--	0	0.769	1	0.3806
21	33237639	21:33237639:A:G	A	G	-0.1927	0.029	3.13E-11	--	0	0.965	1	0.326
21	33934844	21:33934844:G:A	A	G	0.1266	0.05	0.01129	-+	42.8	1.747	1	0.1863
Meta-análise SCOURGE-CNIO do grupo de mulleres												
1	155202934	1:155202934:T:C	T	C	0.101	0.0593	0.08821	++	0	0.122	1	0.7265
1	46810098	1:46810098:T:C	T	C	0.0404	0.0399	0.3121	-+	25.2	1.336	1	0.2477
1	64947147	1:64947147:G:T	T	G	0.0426	0.0423	0.3135	++	0	0.129	1	0.7198
1	77483438	1:77483438:G:A	A	G	0.0342	0.0403	0.3954	++	0	0.004	1	0.9497
10	112972548	10:112972548:C:A	A	C	-0.0148	0.0451	0.7429	--	0	0.042	1	0.8385
10	79977061	10:79977061:G:A	A	G	0.0729	0.0604	0.2274	-+	0	0.457	1	0.4992
11	1219991	11:1219991:G:T	T	G	-0.1428	0.0578	0.01344	--	60.4	2.522	1	0.1122
11	34480495	11:34480495:C:T	T	C	-0.096	0.0385	0.01267	--	0	0.691	1	0.406
12	112922758	12:112922758:T:C	T	C	-0.1144	0.0409	0.005168	--	0	0.111	1	0.7394

CHR	POS	SNP	AE	ANE	beta_meta	SE_meta	pval_meta	dirección	HetISq	HetChiSq	HetDf	HetPval
12	132565387	12:132565387:T:C	T	C	0.0642	0.0382	0.09311	++	61	2.565	1	0.1092
13	112881427	13:112881427:C:T	T	C	0.1502	0.0464	0.0012	++	0	0.198	1	0.656
16	89196249	16:89196249:G:A	A	G	0.203	0.0554	0.0002487	++	0	0.001	1	0.9786
17	45635098	17:45635098:G:A	A	G	-0.105	0.0414	0.01125	--	0	0.211	1	0.6461
17	49863260	17:49863260:C:A	A	C	0.1566	0.1239	0.2064	++	0	0.517	1	0.4722
19	10352442	19:10352442:G:C	C	G	-0.042	0.3261	0.8976	-?	0	0	0	1
19	45869791	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
19	4717660	19:4717660:A:G	A	G	-0.1748	0.1212	0.1493	-?	0	0	0	1
19	48702888	19:48702888:G:C	C	G	-0.0541	0.0384	0.1592	--	0	0.604	1	0.4371
19	50379362	19:50379362:T:C	T	C	-0.1273	0.0407	0.001751	--	0	0.064	1	0.7997
2	26616850	2:26616850:G:T	T	G	0.0078	0.111	0.9438	+?	0	0	0	1
2	60480453	2:60480453:A:G	A	G	0.144	0.1132	0.2033	+?	0	0	0	1
21	33237639	21:33237639:A:G	A	G	-0.12	0.0418	0.004106	+-	63.3	2.722	1	0.09895
21	33934844	21:33934844:G:A	A	G	0.1304	0.0725	0.07221	-+	61	2.564	1	0.1093
3	101647854	3:101647854:G:A	A	G	-0.063	0.0409	0.124	+-	38.3	1.62	1	0.2031
3	45793925	3:45793925:G:A	A	G	0.0696	0.0681	0.3066	++	0	0.015	1	0.904
3	45818159	3:45818159:G:A	A	G	0.395	0.0745	1.13E-07	++	56.6	2.305	1	0.129
4	105897896	4:105897896:G:A	A	G	-0.0447	0.047	0.3415	--	34.6	1.53	1	0.2161
4	25312372	4:25312372:A:G	A	G	-0.1242	0.0899	0.1672	+-	12.1	1.138	1	0.2861
4	25447603	4:25447603:G:A	A	G	-0.0797	0.1275	0.5318	-?	0	0	0	1
5	132441275	5:132441275:T:C	T	C	-0.1521	0.0531	0.004187	--	0	0.815	1	0.3665
6	29947491	6:29947491:C:T	T	C	0.1495	0.1502	0.3195	+?	0	0	0	1
6	31182658	6:31182658:A:G	A	G	0.0711	0.05	0.1552	++	51	2.042	1	0.153
6	32714422	6:32714422:T:C	T	C	0.1436	0.1004	0.1529	++	0	0.082	1	0.7746
6	41520640	6:41520640:G:A	A	G	-0.007	0.1277	0.9561	+-	0	0.926	1	0.3358
7	100032719	7:100032719:C:T	T	C	-0.0362	0.039	0.3529	+-	78.5	4.649	1	0.03107
7	22854868	7:22854868:T:C	T	C	0.0691	0.0388	0.07505	++	0	0.004	1	0.9473

CHR	POS	SNP	AE	ANE	beta_meta	SE_meta	pval_meta	dirección	HetISq	HetChiSq	HetDf	HetPval
8	60513412	8:60513412:G:A	A	G	0.038	0.0386	0.325	++	0	0.004	1	0.9467
9	133273813	9:133273813:C:T	T	C	-0.0537	0.0397	0.1761	--	58.2	2.39	1	0.1221
9	15795835	9:15795835:G:T	T	G	0.0707	0.0917	0.4406	-+	0	0.228	1	0.6328
9	33425186	9:33425186:GTAAC:G	G	GTAAC	0.1846	0.0454	4.74E-05	++	0	0.224	1	0.6362
Meta-análise SCOURGE-CNIO do grupo de homes												
1	155202934	1:155202934:T:C	T	C	0.1683	0.0584	0.003946	++	0	0.129	1	0.7198
1	46810098	1:46810098:T:C	T	C	0.0968	0.038	0.01089	++	0	0.934	1	0.3337
1	64947147	1:64947147:G:T	T	G	-0.1011	0.0403	0.01213	--	0	0.05	1	0.8236
1	77483438	1:77483438:G:A	A	G	0.1238	0.0388	0.001418	++	39.8	1.662	1	0.1974
10	112972548	10:112972548:C:A	A	C	-0.0956	0.043	0.02626	--	0	0.795	1	0.3726
10	79977061	10:79977061:G:A	A	G	0.0048	0.0572	0.9333	-+	0	0.019	1	0.8892
11	1219991	11:1219991:G:T	T	G	-0.0692	0.0561	0.2173	+-	38.2	1.619	1	0.2032
11	34480495	11:34480495:C:T	T	C	-0.1677	0.037	5.86E-06	--	0	0.894	1	0.3445
12	112922758	12:112922758:T:C	T	C	-0.1825	0.0384	2.06E-06	--	9.3	1.102	1	0.2938
12	132565387	12:132565387:T:C	T	C	0.0598	0.037	0.1059	++	0	0.699	1	0.4031
13	112881427	13:112881427:C:T	T	C	0.0351	0.0445	0.4302	-+	78.4	4.629	1	0.03144
16	89196249	16:89196249:G:A	A	G	0.0374	0.0542	0.4901	++	0	0.159	1	0.6904
17	45635098	17:45635098:G:A	A	G	-0.0883	0.0402	0.02803	--	0	0.274	1	0.6006
17	49863260	17:49863260:C:A	A	C	0.2878	0.121	0.01741	++	0	0	1	0.9926
19	10352442	19:10352442:G:C	C	G	0.3819	0.2653	0.15	+?	0	0	0	1
19	45869791	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
19	4717660	19:4717660:A:G	A	G	-0.2359	0.1014	0.02006	-?	0	0	0	1
19	48702888	19:48702888:G:C	C	G	-0.0094	0.0369	0.7995	+-	62.5	2.667	1	0.1025
19	50379362	19:50379362:T:C	T	C	-0.1082	0.0395	0.006134	+-	37.8	1.608	1	0.2048
2	26616850	2:26616850:G:T	T	G	0.1264	0.0965	0.1906	+?	0	0	0	1
2	60480453	2:60480453:A:G	A	G	0.0226	0.0949	0.8122	+?	0	0	0	1
21	33237639	21:33237639:A:G	A	G	-0.2678	0.0398	1.62E-11	--	0	0.003	1	0.9531

CHR	POS	SNP	AE	ANE	beta_meta	SE_meta	pval_meta	dirección	HetISq	HetChiSq	HetDf	HetPval
21	33934844	21:33934844:G:A	A	G	0.1146	0.0681	0.0922	++	0	0.079	1	0.7787
3	101647854	3:101647854:G:A	A	G	-0.0469	0.0398	0.2393	--	0	0.026	1	0.8727
3	45793925	3:45793925:G:A	A	G	0.1804	0.0651	0.005626	++	0	0.005	1	0.9413
3	45818159	3:45818159:G:A	A	G	0.7276	0.0698	2.00E-25	++	0	0.159	1	0.6898
4	105897896	4:105897896:G:A	A	G	-0.0931	0.0454	0.04058	--	0	0.015	1	0.9025
4	25312372	4:25312372:A:G	A	G	-0.1056	0.085	0.2142	--	0	0.329	1	0.566
4	25447603	4:25447603:G:A	A	G	-0.0488	0.1094	0.6557	-?	0	0	0	1
5	132441275	5:132441275:T:C	T	C	-0.0062	0.0513	0.9034	-+	0	0.223	1	0.637
6	29947491	6:29947491:C:T	T	C	0.0725	0.1257	0.5643	+?	0	0	0	1
6	31182658	6:31182658:A:G	A	G	0.1813	0.0477	0.0001445	++	0	0.379	1	0.5383
6	32714422	6:32714422:T:C	T	C	0.1242	0.0957	0.1943	++	0	0	1	0.9974
6	41520640	6:41520640:G:A	A	G	0.0858	0.1207	0.4774	-+	0	0.768	1	0.3809
7	100032719	7:100032719:C:T	T	C	0.0809	0.0372	0.02951	++	0	0.004	1	0.9474
7	22854868	7:22854868:T:C	T	C	0.0652	0.0374	0.08107	-+	60.2	2.514	1	0.1128
8	60513412	8:60513412:G:A	A	G	0.0621	0.0374	0.0966	-+	30.8	1.445	1	0.2293
9	133273813	9:133273813:C:T	T	C	-0.0353	0.0386	0.3607	+-	0	0.308	1	0.5788
9	15795835	9:15795835:G:T	T	G	0.0572	0.0866	0.5092	-+	0	0.607	1	0.4359
9	33425186	9:33425186:GTAAC:G	G	GTAAC	0.1581	0.0438	0.0003029	++	38	1.613	1	0.204

Anexo VII. Xenes priorizados por FUMA na meta-análise SCOURGE-HGI$_{AMR}$ (hg37). Symbol: simbolo do xen por HGNC; start: inicio do xen; end: final do xen; pLi: puntuación pLI segundo ExAC (probabilidade de ser intolerante a perda de función; canto máis alto, máis intolerante); ncRVIS: puntuación de intolerancia á variación residual non codificante (canto máis alto, máis intolerante); pos Map SNPS: nº de SNPs asignados por mapeo posicional; pos Map Max CADD: a puntuación CADD máxima dos SNPs asignados por mapeo posicional; eqtl Map SNPs: nº de SNPs asignados por mapeo de eQTL; eqtl Map min P: o p-valor mínimo de eQTL para os SNPs asignados; eqtl Map min Q: o q-valor mínimo por FDR de eQTL para os SNPs asignados; eqtl Maps: tecido dos eQTL mapeados; eqtl Direction: dirección da consecuencia dos SNPs mapeados a eQTL tras aliñar os alelos de risco; min Gwas P: p-valor mínimo no GWAS dos eQTL mapeados; Ind Sig SNPs: SNPs independentes que están en LD cos SNPs mapeados.

symbol	chr	start	end	type	pLI	ncRVIS	Pos Map SNPs	Pos Map Max CADD	Eqtl Map SNPs	Eqtl Map min P	Eqtl Map min Q	Eqtl Mapts	Eqtl Direction	Min Gwas P	Ind Sig SNPs
WDSUB1	2	160092304	160143310	protein_coding	1.4E-06	NA	0	0	37	1.0E-07	2.0E-14	GTEx/v8/Whole_Blood	+	1.4E-07	rs13003835
BAZ2B	2	160175490	160473203	protein_coding	1.0E+00	-6.5E-01	45	16.54	3	1.2E-06	4.0E-32	GTEx/v8/Whole_Blood	-	3.7E-08	rs13003835
AC008277.1	2	160242878	160261147	antisense	NA	NA	30	16.54	0	NA	NA	NA	NA	3.7E-08	rs13003835
AC009506.1	2	160471807	160473593	antisense	NA	NA	0	0	26	8.2E-08	4.8E-14	GTEx/v8/Whole_Blood	-	3.7E-08	rs13003835
LY75	2	160628362	160761260	protein_coding	3.6E-27	NA	0	0	35	3.5E-07	3.8E-87	GTEx/v8/Whole_Blood ; /v8/Esophagus_Mucosa ; /v8/Lung	-	3.7E-08	rs13003835
PLA2R1	2	160788519	160919121	protein_coding	5.2E-15	9.9E-01	0	0	7	1.7E-05	4.6E-36	GTEx/v8/Esophagus_Mucosa	-	1.4E-07	rs13003835
SLC6A20	3	45796942	45838027	protein_coding	6.4E-10	2.5E-01	8	4.173	0	NA	NA	NA	NA	8.1E-17	rs35731912; rs17078346
LZTFL1	3	45864808	45957534	protein_coding	1.9E-01	-1.2E+00	42	13.68	0	NA	NA	NA	NA	6.9E-17	rs35731912; rs17078346; rs34901975; rs12639224; rs7653372
RP11-852E15.1	3	45925132	45925605	pseudogene	NA	NA	10	6.254	0	NA	NA	NA	NA	3.1E-12	rs12639224; rs34901975; rs7653372
CCR9	3	45927996	45944667	protein_coding	4.6E-03	-4.6E-01	14	8.293	18	1.4E-04	8.2E-03	GTEx/v8/Whole_Blood	+	1.5E-16	rs34901975; rs7653372; rs35731912
Y_RNA	3	45944679	45944790	misc_RNA	NA	NA	7	8.293	0	NA	NA	NA	NA	1.2E-11	rs7653372; rs34901975
FYCO1	3	45959396	46037316	protein_coding	9.7E-13	1.7E+00	69	14.93	147	1.8E-05	1.6E-13	GTEx/v8/Esophagus_Mucosa	+	3.1E-12	rs34901975; rs6785091; rs35731912; rs7653372
CXCR6	3	45982425	45989845	protein_coding	6.2E-05	-1.7E-01	19	11.81	9	6.0E-05	2.4E-02	GTEx/v8/Lung	-	2.5E-10	rs34901975; rs6785091

symbol	chr	start	end	type	pLI	ncRVIS	Pos Map SNPs	Pos Map Max CADD	Eqtl Map SNPs	Eqtl Map min P	Eqtl Map min Q	Eqtl Mapts	Eqtl Direction	Min Gwas P	Ind Sig SNPs
XCR1	3	46058516	46069234	protein_coding	9.8E-02	-1.5E-01	54	11.83	0	NA	NA	NA	NA	4.2E-09	rs34901975; rs6785091; rs7653372
NRBF2P2	3	46064788	46065648	pseudogene	NA	NA	41	11.83	0	NA	NA	NA	NA	4.2E-09	rs6785091; rs34901975
FLT1P1	3	46183850	46185169	pseudogene	NA	NA	9	3.954	197	1.8E-11	1.0E-11	GTEx/v8/Whole_Blood	-	3.1E-12	rs34901975; rs7653372; rs6785091; rs35731912
CCR3	3	46205096	46308197	protein_coding	3.2E-02	-5.4E-01	6	3.864	52	7.3E-07	1.6E-30	GTEx/v8/Whole_Blood	-	4.2E-09	rs34901975; rs7653372; rs6785091
CCR1	3	46243200	46249887	protein_coding	8.9E-01	1.9E+00	4	3.864	176	1.4E-10	1.3E-08	GTEx/v8/Esophagus_Mucosa	+	6.9E-17	rs34901975; rs35731912; rs17078346; rs12639224; rs6785091
CCR5	3	46411633	46417697	protein_coding	3.5E-08	1.5E+00	0	0	8	2.2E-05	5.9E-03	GTEx/v8/Lung	-	2.0E-06	rs6785091; rs34901975
RP11-328M4.2	6	41462591	41516359	antisense	NA	NA	4	10.41	0	NA	NA	NA	NA	1.9E-08	rs2477820
FOXP4	6	41514164	41570122	protein_coding	6.7E-01	-4.4E-01	1	4.186	0	NA	NA	NA	NA	2.1E-07	rs2477820
PRCP	11	82534544	82681626	protein_coding	3.3E-08	6.6E-02	6	14.21	0	NA	NA	NA	NA	2.3E-08	rs77599934
RP11-718B12.5	11	82546551	82548264	pseudogene	NA	NA	1	4.826	0	NA	NA	NA	NA	NA	rs77599934
snoU13	11	82554927	82555027	snoRNA	NA	NA	1	4.826	0	NA	NA	NA	NA	NA	rs77599934
C11orf82	11	82611017	82669319	protein_coding	5.3E-13	4.2E-01	3	14.21	0	NA	NA	NA	NA	2.3E-08	rs77599934
RP11-659G9.3	11	82674723	82750157	sense_overlapping	NA	NA	4	3.338	0	NA	NA	NA	NA	NA	rs77599934
RAB30	11	82684175	82782965	protein_coding	9.6E-01	-1.8E-01	5	4.369	0	NA	NA	NA	NA	NA	rs77599934
RAB30-AS1	11	82783108	82817761	lincRNA	NA	NA	2	0.973	0	NA	NA	NA	NA	NA	rs77599934
RP11-113K21.4	11	82783444	82808238	lincRNA	NA	NA	2	0.973	0	NA	NA	NA	NA	NA	rs77599934
RP11-113K21.3	11	82817622	82818003	pseudogene	NA	NA	1	0.973	0	NA	NA	NA	NA	NA	rs77599934
RP11-113K21.2	11	82822102	82822484	pseudogene	NA	NA	1	0.973	0	NA	NA	NA	NA	NA	rs77599934

Anexo VIII. Táboa dos SNPs incluídos nos sets cribles ao 95% para as rexión do cromosoma 2 da meta-análise SCOURGE-HGI$_{AMR}$.

SNP	CHR	POS	AE	ANE	Beta	SE	Pval
2:159351605:C:T	2	159351605	T	C	-0.1775	0.0377	2.43E-06
2:159360569:C:CACTG	2	159360569	CACTG	C	-0.2005	0.0394	3.65E-07
2:159360569:C:CACTG	2	159360569	CACTG	C	-0.2005	0.0394	3.65E-07
2:159362240:C:A	2	159362240	A	C	-0.1585	0.0331	1.70E-06
2:159363825:A:G	2	159363825	A	G	0.1564	0.0331	2.34E-06
2:159364287:A:G	2	159364287	A	G	0.1561	0.0331	2.39E-06
2:159365291:T:C	2	159365291	T	C	0.1552	0.033	2.51E-06
2:159365728:A:G	2	159365728	A	G	0.1554	0.0331	2.66E-06
2:159366391:A:T	2	159366391	A	T	0.1569	0.0331	2.14E-06
2:159366873:G:C	2	159366873	C	G	-0.157	0.0331	2.10E-06
2:159367657:A:G	2	159367657	A	G	0.1659	0.0333	6.17E-07
2:159368776:T:C	2	159368776	T	C	0.1685	0.0332	3.88E-07
2:159371464:A:T	2	159371464	A	T	0.1678	0.0332	4.33E-07
2:159376456:G:A	2	159376456	A	G	-0.1669	0.0331	4.74E-07
2:159376946:A:C	2	159376946	A	C	0.1835	0.0368	6.07E-07
2:159379240:A:G	2	159379240	A	G	0.1677	0.0332	4.54E-07
2:159379487:C:T	2	159379487	T	C	-0.153	0.0329	3.23E-06
2:159379656:G:C	2	159379656	C	G	-0.1624	0.0331	9.05E-07
2:159382217:C:T	2	159382217	T	C	-0.1825	0.0369	7.40E-07
2:159382867:A:G	2	159382867	A	G	0.1814	0.0367	7.74E-07
2:159384181:A:G	2	159384181	A	G	0.168	0.0332	4.33E-07
2:159385913:A:G	2	159385913	A	G	0.1613	0.0333	1.24E-06
2:159386105:A:T	2	159386105	A	T	0.172	0.0325	1.19E-07
2:159386672:T:A	2	159386672	A	T	-0.1667	0.0331	4.68E-07
2:159386673:A:T	2	159386673	A	T	0.1667	0.0331	4.67E-07
2:159387215:A:G	2	159387215	A	G	0.1654	0.0331	5.72E-07
2:159387283:T:C	2	159387283	T	C	0.1647	0.0331	6.46E-07
2:159387569:T:C	2	159387569	T	C	0.1637	0.0331	7.48E-07
2:159387694:G:A	2	159387694	A	G	-0.1646	0.0331	6.53E-07
2:159388577:T:A	2	159388577	A	T	-0.1643	0.0331	6.78E-07
2:159391779:GAT:G	2	159391779	G	GAT	-0.1788	0.0353	4.11E-07
2:159391779:GAT:G	2	159391779	G	GAT	-0.1788	0.0353	4.11E-07
2:159392137:A:G	2	159392137	A	G	0.1556	0.0331	2.51E-06
2:159392606:C:G	2	159392606	C	G	0.1858	0.0367	4.30E-07
2:159393260:T:C	2	159393260	T	C	0.1696	0.0322	1.40E-07
2:159393627:TC:T	2	159393627	T	TC	-0.1721	0.0353	1.06E-06
2:159393629:T:A	2	159393629	A	T	-0.1671	0.0338	7.43E-07
2:159396057:A:G	2	159396057	A	G	0.1834	0.0366	5.45E-07
2:159398147:G:A	2	159398147	A	G	0.1592	0.0339	2.65E-06
2:159399171:A:C	2	159399171	A	C	0.1853	0.0366	4.20E-07
2:159399171:A:C	2	159399171	A	C	0.1853	0.0366	4.20E-07
2:159399171:A:C	2	159399171	A	C	0.1853	0.0366	4.20E-07
2:159399171:A:C	2	159399171	A	C	0.1853	0.0366	4.20E-07
2:159399817:A:G	2	159399817	A	G	0.1874	0.0366	3.11E-07
2:159401859:C:T	2	159401859	T	C	-0.1544	0.0324	1.86E-06
2:159406182:C:T	2	159406182	T	C	-0.1537	0.0323	1.99E-06
2:159407982:T:C	2	159407982	T	C	0.1789	0.0325	3.66E-08
2:159408209:T:G	2	159408209	T	G	0.1746	0.0366	1.86E-06
2:159409009:G:A	2	159409009	A	G	-0.1642	0.0325	4.30E-07
2:159411581:T:C	2	159411581	T	C	0.1756	0.0368	1.83E-06
2:159411882:A:G	2	159411882	A	G	0.1648	0.0329	5.46E-07
2:159412797:T:C	2	159412797	T	C	0.1784	0.0325	3.98E-08
2:159413336:C:A	2	159413336	A	C	-0.1685	0.0329	3.05E-07
2:159413622:T:G	2	159413622	T	G	0.1687	0.0324	1.91E-07
2:159414347:C:A	2	159414347	A	C	0.1619	0.033	9.24E-07
2:159414347:C:A	2	159414347	A	C	0.1619	0.033	9.24E-07
2:159414347:C:A	2	159414347	A	C	0.1619	0.033	9.24E-07
2:159414347:C:A	2	159414347	A	C	0.1619	0.033	9.24E-07
2:159415239:G:T	2	159415239	T	G	0.1792	0.0329	4.93E-08
2:159417333:G:A	2	159417333	A	G	-0.1746	0.036	1.23E-06

2:159418933:T:A	2	159418933	A	T	-0.1759	0.036	1.05E-06
2:159418935:G:C	2	159418935	C	G	-0.1759	0.036	1.05E-06
2:159421836:A:T	2	159421836	A	T	0.1753	0.0367	1.79E-06
2:159422919:T:C	2	159422919	T	C	0.1707	0.036	2.07E-06
2:159423607:T:C	2	159423607	T	C	0.1709	0.036	2.00E-06
2:159423792:C:T	2	159423792	T	C	-0.1738	0.0365	1.95E-06
2:159423807:C:T	2	159423807	T	C	-0.1715	0.0359	1.77E-06
2:159424472:T:C	2	159424472	T	C	0.1678	0.036	3.07E-06
2:159424835:A:C	2	159424835	A	C	0.1684	0.036	2.84E-06
2:159425695:C:T	2	159425695	T	C	-0.1681	0.036	2.93E-06
2:159425829:G:T	2	159425829	T	G	-0.1684	0.036	2.84E-06
2:159428549:C:T	2	159428549	T	C	-0.1676	0.036	3.17E-06
2:159430008:T:G	2	159430008	T	G	0.1667	0.036	3.57E-06
2:159430565:A:T	2	159430565	A	T	0.1676	0.036	3.14E-06
2:159430565:A:T	2	159430565	A	T	0.1676	0.036	3.14E-06
2:159430565:A:T	2	159430565	A	T	0.1676	0.036	3.14E-06
2:159430565:A:T	2	159430565	A	T	0.1676	0.036	3.14E-06
2:159431771:C:G	2	159431771	C	G	0.1684	0.036	2.85E-06
2:159435409:G:A	2	159435409	A	G	-0.1699	0.0366	3.37E-06
2:159435453:T:C	2	159435453	T	C	0.1702	0.0365	3.18E-06
2:159436422:C:T	2	159436422	T	C	-0.1701	0.0366	3.29E-06
2:159436667:C:T	2	159436667	T	C	-0.1737	0.0366	2.03E-06
2:159440061:C:T	2	159440061	T	C	-0.1716	0.0366	2.71E-06
2:159440134:C:T	2	159440134	T	C	-0.1716	0.0366	2.71E-06
2:159441612:T:C	2	159441612	T	C	0.1669	0.0359	3.36E-06
2:159447700:T:C	2	159447700	T	C	0.1696	0.0366	3.61E-06

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Anexo IX. Resultados de S-MultiXcan para a meta-análise SCOURGE-HGI$_{AMR}$. N: número de tecidos dispoñibles para o xen; N indep: número de compoñentes independentes de variación mantidos entre as predicións dos tecidos; p_i_best: mellor p-valor da asociación S-prediXcan dun só tecido; t_i_best: nome do tecido da mellor asociación de S-prediXcan de un só tecido; p_i_worst: peor p-valor da asociación S-prediXcan dun só tecido; t_i_worst: nome do tecido da peor asociación de S-prediXcan dun só tecido; eigen max: autovalor da primeira compoñente independente na descomposición SVD da correlación de expresión predita (varianza explicada); eigen_min: autovalor da última compoñente independente na descomposición SVD da correlación de expresión predita; eigen_min_kept: autovalor da compoñente máis pequena que se mantivo na descomposición SVD da correlación de expresión predita; Z min: puntuación Z mínima entre as asociacións de S-prediXcan dun só tecido; Z max: puntuación Z máxima entre as asociacións de S-prediXcan dun só tecido; Z mean: puntuación Z media as asociacións de S-prediXcan dun só tecido; Z sd: desviación estándar da puntuación Z media entre as asociacións de S-prediXcan dun só tecido; tmi: traza de T*T', onde T é a correlación dos niveis de expresión preditos para diferentes tecidos multiplicada pola súa pseudoinversa SVD e é unha estimación do número de compoñentes independentes de variación na expresión predita entre tecidos. P-valor axustado por bonferroni para 22.052 tecidos (p<2.27x10-06). SVD: *single-value decomposition*.

gene	pvalue	N	n_indep	p_i_best	t_i_best	p_i_worst	t_i_worst	eigen_max	eigen_min	eigen_min_kept	z_min	z_max	z_mean	z_sd	tmi
SLC6A20	2.7E-15	33	5	4.5E-16	Muscle_Skeletal	4.7E-01	Pancreas	19.1054	6.6E-16	0.8718	0.7282	8.1232	3.0836	1.5291	5
CCR9	7.2E-15	7	6	1.4E-16	Whole_Blood	9.5E-01	Nerve_Tibial	3.3547	1.5E-03	0.1129	0.0586	8.2606	3.6975	2.8940	6
AP000569.9	2.6E-11	3	3	8.8E-04	Cells_EBV-transformed_lymphocytes	2.7E-01	Spleen	2.0256	2.0E-01	0.1977	-3.3271	1.1017	-1.5724	2.3532	3
SMARCC1	3.0E-09	23	3	4.4E-04	Brain_Hippocampus	9.4E-01	Brain_Frontal_Cortex_BA9	15.7544	1.2E-17	1.0643	-3.5162	3.3336	-0.5102	1.8859	3
CXCR6	1.3E-08	38	4	2.9E-10	Adipose_Visceral_Omentum	8.6E-01	Spleen	16.4694	1.5E-15	3.0526	-6.3028	2.3888	-2.3617	2.5035	4
LZTFL1	3.4E-08	45	5	6.1E-11	Thyroid	9.5E-01	Brain_Amygdala	22.0018	7.2E-16	0.7864	-2.0078	6.5409	-0.4839	2.1646	5
CAMP	4.0E-08	19	4	5.2E-02	Adipose_Subcutaneous	5.1E-01	Skin_Not_Sun_Exposed_Suprapubic	13.8308	4.9E-18	0.5812	-1.9459	1.0039	-0.0780	1.1951	4
TYW3	4.3E-08	49	4	1.1E-02	Brain_Hippocampus	9.7E-01	Colon_Transverse	29.2813	2.1E-15	1.1018	-2.5437	2.3440	-0.9149	1.1153	4
FOXP4	2.0E-07	28	4	4.5E-05	Lung	9.0E-01	Esophagus_Mucosa	22.7194	1.4E-15	0.9899	-2.3595	4.0811	0.5115	1.3182	4
S100A12	1.1E-06	13	4	6.9E-02	Testis	9.0E-01	Breast_Mammary_Tissue	10.3946	4.0E-17	0.4139	-1.8180	1.5086	0.0285	1.1704	4
SLC25A37	1.1E-06	33	8	1.1E-06	Cells_EBV-transformed_lymphocytes	8.2E-01	Pituitary	16.1802	1.6E-15	0.6957	-3.3919	4.8772	-0.4876	2.3364	8
FYCO1	1.3E-06	47	5	3.0E-06	Brain_Hippocampus	9.8E-01	Small_Intestine_Terminal_Ileum	23.1900	2.2E-15	0.8271	-2.2120	4.6682	0.6580	1.7502	5
XCR1	1.8E-06	23	4	3.0E-04	Small_Intestine_Terminal_Ileum	9.6E-01	Esophagus_Muscularis	13.7522	1.8E-17	0.9400	-3.4687	3.6178	0.3437	1.7753	4

Anexo X. Replicación, nas meta-análises SCOURGE-HGI$_{ALL}$, SCOURGE-HGI$_{3POP}$ e SCOURGE-HGI$_{AMR}$, das 40 variantes asociadas á hospitalización por COVID-19 identificadas polo consorcio HGI. AE: alelo de efecto na meta-análise, NEA: alelo de efecto na meta-análise, ANE: alelo de non efecto na meta-análise. HGIB2_ALL corresponde aos resultados orixinais na análise HGI-B2 ALL.

	HGIB2_ALL					SC-HGI_ALL					SC-HGI_3POP					SC-HGI_AMR				
SNP	**ANE**	**AE**	**beta**	**SE**	**pval**	**ANE**	**AE**	**beta**	**SE**	**pval**	**ANE**	**AE**	**beta**	**SE**	**pval**	**ANE**	**AE**	**beta**	**SE**	**pval**
chr1:155202934:T:C	C	T	-0.1304	0.0135	5.6E-22	T	C	0.1310	0.0135	2.3E-22	T	C	0.1278	0.0139	3.3E-20	T	C	0.1263	0.0593	3.3E-02
chr1:46810098:T:C	C	T	-0.0472	0.0086	4.4E-08	T	C	0.0485	0.0086	1.5E-08	T	C	0.0486	0.0092	1.1E-07	T	C	0.0689	0.0321	3.2E-02
chr1:64947147:G:T	T	G	-0.0550	0.0092	2.2E-09	T	G	-0.0560	0.0091	8.1E-10	T	G	-0.0576	0.0098	3.7E-09	T	G	-0.026	0.0325	4.3E-01
chr1:77483438:G:A	A	G	0.0623	0.0089	2.9E-12	A	G	0.0622	0.0089	2.2E-12	A	G	0.0625	0.0093	2.0E-11	A	G	0.0385	0.0328	2.4E-01
chr10:112972548:C:A	A	C	-0.0537	0.0098	4.8E-08	A	C	-0.0522	0.0098	9.2E-08	A	C	-0.0519	0.0106	9.2E-07	A	C	-0.041	0.0348	2.4E-01
chr10:79977061:G:A	A	G	0.0773	0.0122	2.5E-10	A	G	0.0805	0.0121	2.8E-11	A	G	0.075	0.0137	4.7E-08	A	G	0.1078	0.0367	3.3E-03
chr11:1219991:G:T	T	G	-0.1012	0.0145	3.4E-12	T	G	-0.1024	0.0145	1.5E-12	T	G	-0.0975	0.015	7.9E-11	T	G	-0.059	0.0651	3.7E-01
chr11:34480495:C:T	T	C	-0.0870	0.0085	2.3E-24	T	C	-0.0870	0.0085	1.1E-24	T	C	-0.0876	0.0091	5.7E-22	T	C	-0.053	0.0314	9.4E-02
chr12:112922758:T:C	C	T	0.0744	0.0092	4.3E-16	T	C	-0.0753	0.0091	1.5E-16	T	C	-0.073	0.0097	5.2E-14	T	C	-0.062	0.0413	1.3E-01
chr12:132565387:T:C	C	T	-0.0667	0.0085	2.8E-15	T	C	0.0667	0.0084	2.2E-15	T	C	0.0663	0.0089	1.0E-13	T	C	0.0603	0.0356	9.0E-02
chr13:112881427:C:T	T	C	0.0974	0.0103	3.2E-21	T	C	0.0969	0.0102	3.1E-21	T	C	0.0947	0.011	5.6E-18	T	C	0.107	0.0407	8.5E-03
chr16:89196249:G:A	A	G	0.0985	0.0128	1.4E-14	A	G	0.1011	0.0127	1.9E-15	A	G	0.1036	0.0131	2.6E-15	A	G	0.1584	0.0525	2.6E-03
chr17:45635098:G:A	A	G	-0.0870	0.0109	1.1E-15	A	G	-0.0861	0.0108	1.6E-15	A	G	-0.0871	0.011	2.8E-15	A	G	-0.04	0.0461	3.8E-01
chr17:49863260:C:A	A	C	0.1920	0.0203	2.6E-21	A	C	0.1883	0.0198	2.3E-21	A	C	0.1861	0.0224	1.1E-16	A	C	0.1683	0.0444	1.5E-04
chr19:10352442:G:C	C	G	0.2343	0.0246	1.7E-21	C	G	0.2326	0.0245	2.3E-21	C	G	0.2333	0.0248	4.9E-21	C	G	0.3164	0.1251	1.1E-02
chr19:45869791:ATT:A	A	ATT	0.0621	0.0109	1.2E-08	A	ATT	0.0621	0.0108	8.7E-09	A	ATT	0.0599	0.0113	1.1E-07	A	ATT	0.0725	0.0393	6.5E-02
chr19:4717660:A:G	G	A	0.1499	0.0099	4.1E-52	A	G	-0.1507	0.0098	2.2E-53	A	G	-0.1473	0.0102	7.6E-47	A	G	-0.159	0.0363	1.1E-05
chr19:48702888:G:C	C	G	-0.0574	0.0086	2.0E-11	C	G	-0.0561	0.0085	4.7E-11	C	G	-0.0502	0.0088	1.2E-08	C	G	-0.063	0.0351	7.1E-02
chr19:50379362:T:C	C	T	0.0777	0.0087	5.9E-19	T	C	-0.0780	0.0087	2.4E-19	T	C	-0.0763	0.0091	6.5E-17	T	C	-0.131	0.0313	3.0E-05
chr2:26616850:G:T	T	G	0.0498	0.0088	1.3E-08	T	G	0.0503	0.0087	7.1E-09	T	G	0.0493	0.0092	8.1E-08	T	G	0.0706	0.0325	3.0E-02
chr2:60480453:A:G	G	A	-0.0717	0.0093	1.3E-14	A	G	0.0714	0.0093	1.2E-14	A	G	0.0687	0.0096	8.9E-13	A	G	0.0458	0.039	2.4E-01

	HGIB2_ALL					SC-HGI_ALL					SC-HGI_3POP					SC-HGI_AMR				
SNP	**ANE**	**AE**	**beta**	**SE**	**pval**	**ANE**	**AE**	**beta**	**SE**	**pval**	**ANE**	**AE**	**beta**	**SE**	**pval**	**ANE**	**AE**	**beta**	**SE**	**pval**
chr21:33237639:A:G	G	A	0.1312	0.0088	1.4E-50	A	G	-0.1314	0.0087	1.9E-51	A	G	-0.1294	0.0093	1.5E-43	A	G	-0.154	0.0314	9.8E-07
chr21:33934844:G:A	A	G	0.1302	0.0140	1.6E-20	A	G	0.1320	0.0139	1.6E-21	A	G	0.1336	0.0156	1.1E-17	A	G	0.2134	0.0404	1.3E-07
chr3:101647854:G:A	A	G	-0.0601	0.0089	1.3E-11	A	G	-0.0596	0.0088	1.4E-11	A	G	-0.0671	0.0094	8.3E-13	A	G	-0.106	0.0361	3.3E-03
chr3:45793925:G:A	A	G	0.1601	0.0209	2.1E-14	A	G	0.1616	0.0208	7.0E-15	A	G	0.1634	0.0211	9.4E-15	A	G	0.1501	0.0871	8.5E-02
chr3:45818159:G:A	A	G	0.4896	0.0155	8.3E-218	A	G	0.4898	0.0155	2.1E-220	A	G	0.4884	0.0167	1.6E-187	A	G	0.5102	0.0622	2.4E-16
chr4:105897896:G:A	A	G	-0.0607	0.0110	3.8E-08	A	G	-0.0622	0.0110	1.5E-08	A	G	-0.0626	0.0112	2.0E-08	A	G	-0.057	0.0577	3.2E-01
chr4:25312372:A:G	G	A	0.0936	0.0152	7.0E-10	A	G	-0.0951	0.0150	2.6E-10	A	G	-0.0959	0.0176	4.8E-08	A	G	-0.128	0.0464	5.7E-03
chr4:25447603:G:A	A	G	-0.0719	0.0102	1.8E-12	A	G	-0.0714	0.0101	1.6E-12	A	G	-0.0639	0.011	7.0E-09	A	G	-0.038	0.0338	2.7E-01
chr5:132441275:T:C	C	T	0.0740	0.0116	1.6E-10	T	C	-0.0744	0.0115	9.9E-11	T	C	-0.0809	0.0125	9.4E-11	T	C	-0.069	0.0495	1.6E-01
chr6:29947491:C:T	T	C	-0.0612	0.0091	1.9E-11	T	C	-0.0587	0.0091	9.2E-11	T	C	-0.058	0.0096	1.3E-09	T	C	0.009	0.039	8.2E-01
chr6:31182658:A:G	G	A	-0.0668	0.0102	6.0E-11	A	G	0.0657	0.0101	9.2E-11	A	G	0.0714	0.0108	3.2E-11	A	G	0.0501	0.0381	1.9E-01
chr6:32714422:T:C	C	T	-0.1063	0.0171	4.6E-10	T	C	0.1055	0.0170	5.6E-10	T	C	0.1062	0.0179	2.9E-09	T	C	-0.097	0.1064	3.6E-01
chr6:41520640:G:A	A	G	0.2403	0.0207	2.8E-31	A	G	0.2396	0.0201	8.5E-33	A	G	0.1874	0.0264	1.3E-12	A	G	0.1802	0.0458	8.5E-05
chr7:100032719:C:T	T	C	0.0497	0.0086	8.9E-09	T	C	0.0500	0.0086	5.5E-09	T	C	0.0523	0.0092	1.3E-08	T	C	0.0047	0.0315	8.8E-01
chr7:22854868:T:C	C	T	-0.0494	0.0088	1.8E-08	T	C	0.0496	0.0087	1.3E-08	T	C	0.0506	0.0092	3.3E-08	T	C	0.0462	0.0349	1.9E-01
chr8:60513412:G:A	A	G	0.0504	0.0086	4.6E-09	A	G	0.0503	0.0085	3.8E-09	A	G	0.0525	0.0091	9.2E-09	A	G	0.0229	0.0313	4.6E-01
chr9:133273813:C:T	T	C	-0.0871	0.0086	4.1E-24	T	C	-0.0853	0.0086	2.1E-23	T	C	-0.0882	0.0091	4.2E-22	T	C	0.0157	0.0367	6.7E-01
chr9:15795835:G:T	T	G	0.1035	0.0188	3.6E-08	T	G	0.1030	0.0187	3.7E-08	T	G	0.1008	0.0193	1.9E-07	T	G	0.1763	0.0981	7.2E-02
chr9:33425186:GTAAC:G	G	GTAAC	0.0657	0.0110	2.5E-09	G	GTAAC	0.0645	0.0109	3.8E-09	G	GTAAC	0.0645	0.0118	4.2E-08	G	GTAAC	0.0422	0.0412	3.1E-01

Anexo XI. Documento do comité de ética para o proxecto SCOURGE e financiación.

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE

SERVIZO GALEGO de SAÚDE | Xerencia do Servizo Galego de Saúde

galicia

DICTAMEN DEL Comité de ética de la investigación con medicamentos de Galicia (CEIm-G)

Paula M. López Vázquez, Secretaria del Comité de ética de la investigación con medicamentos de Galicia (CEIm-G)

CERTIFICA:

Que este Comité evaluó en su reunión del día 22/12/21 la modificación del estudio:

Título: Determinantes genéticos y biomarcadores genómicos de riesgo en pacientes con infección por coronavirus SARS-COV-2
Versión modificación: modificación Noviembre 2021 (Actualización de listado de centros participantes en el proyecto. Actualización de estado del consorcio)
Tipo de estudio: Otros estudios
Promotor: Ángel Carracedo Álvarez
Código del Promotor:
Código de Registro: 2020/197

Y que este Comité acepta de conformidad con sus procedimientos normalizados de trabajo y tomando en cuenta los requisitos éticos, metodológicos y legales exigibles, que dicha enmienda sea incorporada al estudio de investigación.

Documento asinado dixitalmente por:
Paula López Vázquez (23/12/2021 09:20)
https://sede.xunta.gal/cve?idcve=SAOC-Y4G4-BO8H-OF4O-TDBN-L5RX-4RP2-K164-0247-6038-02

Xacobeo 21·22

Edificio Administrativo San Lázaro
15703 SANTIAGO DE COMPOSTELA
Teléfono: 881546425
ceic@sergas.es

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE

SERVIZO GALEGO de SAÚDE | Xerencia do Servizo Galego de Saúde

galicia

Y HACE CONSTAR QUE:

1 El comité cumple los requisitos legales vigentes aplicables a los Comités de ética de investigación.
2 El comité tanto en su composición como en sus PNTs cumple las Normas de Buena Práctica Clínica (CPMP/ICH/135/95).
3 La composición actual del comité es:

Susana María Romero Yuste (Presidenta). Médico Especialista en Reumatología.
Irene Zarra Ferro. (Vicepresidenta). Farmacéutica de Atención Especializada.
Paula Mª López Vázquez, (Secretaria). Médico Especialista en Farmacología Clínica.
Martina Lema Oreiro (Vicesecretaria). Farmacéutica de Atención Especializada
Rosendo Bugarín González. Médico Especialista en Medicina Familiar y Comunitaria.
Juan Casariego Rosón. Médico Especialista en Cardiología.
Nuria Carballeda Feijóo. Miembro lego. Representante de los intereses de los pacientes.
Juana Mª Cruz del Río. Trabajadora Social.
Rafael Álvaro Millán Calenti. Asesor jurídico.
Juan Fernando Cueva Bañuelos. Médico Especialista en Oncología Médica.
José Álvaro Fernández Rial. Médico Especialista en Medicina Interna.
José Luis Fernández Trisac. Médico Especialista en Pediatría.
Mª José Ferreira Díaz. Diplomada Universitaria de Enfermería.
Agustín Pía Morandeira. Farmacéutico de Atención Primaria
Jorge Prado Casal. Licenciado en CC. Físicas. Experto en Protección de Datos.
Carmen Rodríguez-Tenreiro Sánchez. Licenciada en Farmacia.
Diego Santos García. Médico Especialista en Neurología.
Juan Vázquez Lago. Médico Especialista en Medicina Preventiva y Salud Pública.
Mª Asunción Verdejo González. Médico Especialista en Farmacología Clínica..

En Santiago de Compostela,

Documento asinado dixitalmente por:
Paula López Vázquez (23/12/2021 09:20)
https://sede.xunta.gal/cve?idcve=SAOC-Y4G4-BQAH-QF4O-TDBN-L5RX-4RP2-K164-0247-6038-02

Xacobeo 21·22

Edificio Administrativo San Lázaro
15703 SANTIAGO DE COMPOSTELA
Teléfono: 881546425
ceic@sergas.es

FINANCIAMENTO: O estudo SCOURGE foi financiado polo Instituto de Salud Carlos III (COV20_00622), Fundación Amancio Ortega, Banco de Santander e por un contrato Predoutoral da Xunta de Galicia e a Axencia GAIN (ED481A 2021/193).

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Anexo XII. Relación das publicacións relacionadas con esta tese.

Novel genes and sex differences in COVID-19 severity

Author: Cruz, Raquel; Diz-de Almeida, Silvia

Publication: Human Molecular Genetics

Publisher: Oxford University Press

Date: 2022-06-16

Copyright © 2022, © The Author(s) 2022. Published by Oxford University Press.

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

Novel genes and sex differences in COVID-19 severity. (2022)

Human molecular genetics, 31(22), 3789–3806. Oxford University Press.

DOI: https://doi.org/10.1093/hmg/ddac132.

Índices de calidade JCR – Genetics & Heredity (2022):

Factor de impacto: 3.5

Posición: 66/171 (Mellor cuartil – Q2; mellor percentil – 61.7)

Nº citas a día 18/02/2024: 45

Contribución ao traballo da doutoranda: Co-primeira autoría. Análise dos datos; intepretación dos resultados; redacción do artigo.

Lista de autores:

Raquel Cruz [1 2 3 4], **Silvia Diz-de Almeida** [4], Miguel López de Heredia [2], Inés Quintela [1], Francisco C Ceballos [5], Guillermo Pita [6], José M Lorenzo-Salazar [7], Rafaela González-Montelongo [7], Manuela Gago-Domínguez [8 3], Marta Sevilla Porras [2 9], Jair Antonio Tenorio Castaño [2 9 10], Julian Nevado [2 9 10], Jose María Aguado [11 12 13 14], Carlos Aguilar [15], Sergio Aguilera-Albesa [16 17], Virginia Almadana [18], Berta Almoguera [19 2], Nuria Alvarez [6], Álvaro Andreu-Bernabeu [20 13], Eunate Arana-Arri [21 22], Celso Arango [20 23 13], María J Arranz [24], Maria-Jesus Artiga [25], Raúl C Baptista-Rosas [26 27 28], María Barreda-Sánchez [29 30], Moncef Belhassen-Garcia [31 32], Joao F Bezerra [33], Marcos A C Bezerra [34], Lucía Boix-Palop [35], María Brion [36 37], Ramón Brugada [38 39 37 40], Matilde Bustos [41], Enrique J Calderón [42 43 44], Cristina Carbonell [45 32], Luis Castano [21 46 2 47 48], Jose E Castelao [49], Rosa Conde-Vicente [50], M Lourdes Cordero-Lorenzana [51], Jose L Cortes-Sanchez [52 53], Marta Corton [19 2], M Teresa Darnaude [54], Alba De Martino-Rodríguez [55 56], Victor Del Campo-Pérez [57], Aranzazu Diaz de Bustamante [54], Elena Domínguez-Garrido [58], Andre D Luchessi [59], Rocío Eiros [60], Gladys Mercedes Estigarribia Sanabria [61], María Carmen Fariñas [62 63 64], Uxía Fernández-Robelo [65], Amanda Fernández-Rodríguez [5 14], Tania Fernández-Villa [66], Belén Gil-Fournier [67], Javier Gómez-Arrue [55 56], Beatriz González Álvarez [55 56], Fernan Gonzalez Bernaldo de Quirós [68], Javier González-Peñas [20 13 23], Juan F Gutiérrez-Bautista [69], María José Herrero [70 71], Antonio Herrero-Gonzalez [72], María A Jimenez-Sousa [5 14], María Claudia Lattig [73 74], Anabel Liger Borja [75], Rosario Lopez-Rodriguez [19 2], Esther Mancebo [76 77], Caridad Martín-López [75], Vicente Martín [78 43], Oscar Martinez-Nieto [79 74], Iciar Martinez-Lopez [80 81], Michel F Martinez-Resendez [52], Angel Martinez-Perez [82], Juliana F Mazzeu [83 84 85], Eleuterio Merayo Macías [86], Pablo Minguez [19 2], Victor Moreno Cuerda [87 88], Vivian N Silbiger [59], Silviene F Oliveira [83 89 90 91 92], Eva Ortega-Paino [25], Mara Parellada [20 23 13], Estela Paz-Artal [76 77 93], Ney P C Santos [94], Patricia Pérez-Matute [95], Patricia Perez [96], M Elena Pérez-Tomás [29], Teresa Perucho [97], Mel Lina Pinsach-Abuin [38 37], Ericka N Pompa-Mera [98], Gloria L Porras-Hurtado [99], Aurora Pujol [100 2 101], Soraya Ramiro León [67], Salvador Resino [5 14], Marianne R Fernandes [94 102], Emilio Rodríguez-Ruiz [103 3], Fernando Rodriguez-Artalejo [104 105 43 106], José A Rodriguez-Garcia [107], Francisco Ruiz Cabello [69 108 109], Javier Ruiz-Hornillos [110 111 112], Pablo Ryan [113 114 115], José Manuel Soria [82], Juan Carlos Souto [116], Eduardo Tamayo [117 118], Alvaro Tamayo-Velasco [119], Juan Carlos Taracido-Fernandez [72], Alejandro Teper [120], Lilian Torres-Tobar [121], Miguel Urioste [122], Juan Valencia-Ramos [123], Zuleima Yáñez [124], Ruth Zarate [125], Tomoko Nakanishi [126 127 128 129 130], Sara Pigazzini [131 132], Frauke

Degenhardt [133 134], Guillaume Butler-Laporte [135 128], Douglas Maya-Miles [136 137], Luis Bujanda [138 137], Youssef Bouysran [139], Adriana Palom [140 141 142], David Ellinghaus [143 133], Manuel Martínez-Bueno [144], Selina Rolker [145], Sara Amitrano [146], Luisa Roade [137 140 141], Francesca Fava [147 148 149], Christoph D Spinner [150], Daniele Prati [151], David Bernardo [152 137], Federico Garcia [153 154], Gilles Darcis [155 156], Israel Fernández-Cadenas [157], Jan Cato Holter [158 159], Jesus M Banales [160 137], Robert Frithiof [161], Stefano Duga [162 163], Rosanna Asselta [162 163], Alexandre C Pereira [164], Manuel Romero-Gómez [136 137], Beatriz Nafría-Jiménez [165], Johannes R Hov [166 159 167], Isabelle Migeotte [168 139], Alessandra Renieri [147 148 149], Anna M Planas [169 170], Kerstin U Ludwig [145], Maria Buti [137 140 141], Souad Rahmouni [155], Marta E Alarcón-Riquelme [144 171], Eva C Schulte [172 173 174], Andre Franke [133 134], Tom H Karlsen [166 159 175], Luca Valenti [176 177], Hugo Zeberg [178 179], Brent Richards [180 128 181], Andrea Ganna [132 182], Mercè Boada [183 184], Itziar de Rojas [183 184], Agustín Ruiz [183 184], Pascual Sánchez-Juan [185], Luis Miguel Real [186]; SCOURGE Cohort Group; HOSTAGE Cohort Group; GRA@CE Cohort Group; Encarna Guillen-Navarro [29 187 188 189], Carmen Ayuso [19 2], Anna González-Neira [6], José A Riancho [62 63 64], Augusto Rojas-Martinez [190], Carlos Flores [7 191 192 193], Pablo Lapunzina [2 9 10], Angel Carracedo [1 2 3 4 8]

Filiacións:

[1]Centro Nacional de Genotipado (CEGEN), Universidade de Santiago de Compostela, 15706 Santiago de Compostela, Spain.
[2]Centre for Biomedical Network Research on Rare Diseases (CIBERER), Instituto de Salud Carlos III, 28029 Madrid, Spain.
[3]Instituto de Investigación Sanitaria de Santiago (IDIS), 15706 Santiago de Compostela, Spain.
[4]Centro Singular de Investigación en Medicina Molecular y Enfermedades Crónicas (CIMUS), Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain.
[5]Unidad de Infección Viral e Inmunidad, Centro Nacional de Microbiología (CNM), Instituto de Salud Carlos III (ISCIII), 28220 Madrid, Spain.
[6]Spanish National Cancer Research Centre, Human Genotyping-CEGEN Unit, 28029 Madrid, Spain.
[7]Genomics Division, Instituto Tecnológico y de Energías Renovables, 38600 Santa Cruz de Tenerife, Spain.
[8]Fundación Pública Galega de Medicina Xenómica, Sistema Galego de Saúde (SERGAS), 15706 Santiago de Compostela, Spain.
[9]Instituto de Genética Médica y Molecular (INGEMM), Hospital Universitario La Paz-IDIPAZ, 28046 Madrid, Spain.
[10]ERN-ITHACA-European Reference Network.
[11]Unit of Infectious Diseases, Hospital Universitario 12 de Octubre, Instituto de Investigación Sanitaria Hospital 12 de Octubre (imas12), 28041 Madrid, Spain.
[12]Spanish Network for Research in Infectious Diseases (REIPI RD16/0016/0002), Instituto de Salud Carlos III, 28029 Madrid, Spain.
[13]School of Medicine, Universidad Complutense, 28040 Madrid, Spain.
[14]Centro de Investigación Biomédica en Red de Enfermedades Infecciosas (CIBERINFEC), Instituto de Salud Carlos III, 28029 Madrid, Spain.
[15]Hospital General Santa Bárbara de Soria, 42005 Soria, Spain.
[16]Pediatric Neurology Unit, Department of Pediatrics, Navarra Health Service Hospital, 31008 Pamplona, Spain.
[17]Navarra Health Service, NavarraBioMed Research Group, 31008 Pamplona, Spain.
[18]Hospital Universitario Virgen Macarena, Neumología, 41009 Seville, Spain.
[19]Department of Genetics & Genomics, Instituto de Investigación Sanitaria-Fundación Jiménez Díaz University Hospital - Universidad Autónoma de Madrid (IIS-FJD, UAM), 28040 Madrid, Spain.
[20]Department of Child and Adolescent Psychiatry, Institute of Psychiatry and Mental Health, Hospital General Universitario Gregorio Marañón (IiSGM), 28007 Madrid, Spain.
[21]Biocruces Bizkai HRI, 48903 Barakaldo, Bizkaia, Spain.
[22]Cruces University Hospital, Osakidetza, 48903 Barakaldo, Bizkaia, Spain.
[23]Centre for Biomedical Network Research on Mental Health (CIBERSAM), Instituto de Salud Carlos III, 28029 Madrid, Spain.
[24]Fundació Docència I Recerca Mutua Terrassa, 08221 Terrassa, Spain.
[25]Spanish National Cancer Research Center, CNIO Biobank, 28029 Madrid, Spain.
[26]Hospital General de Occidente, 45170 Zapopan, Jalisco, Mexico.
[27]Centro Universitario de Tonalá, Universidad de Guadalajara, 45425 Tonalá, Jalisco, Mexico.
[28]Centro de Investigación Multidisciplinario en Salud, Universidad de Guadalajara, 45425 Tonalá, Jalisco, Mexico.
[29]Instituto Murciano de Investigación Biosanitaria (IMIB-Arrixaca), 30120 Murcia, Spain.
[30]Universidad Católica San Antonio de Murcia (UCAM), 30120 Murcia, Spain.
[31]Hospital Universitario de Salamanca-IBSAL, Servicio de Medicina Interna-Unidad de Enfermedades Infecciosas, 37007 Salamanca, Spain.
[32]Universidad de Salamanca, 37007 Salamanca, Spain.
[33]Escola Tecnica de Saúde, Laboratorio de Vigilancia Molecular Aplicada, 68515-000 Pará, Brazil.
[34]Federal University of Pernambuco, Genetics Postgraduate Program, Recife 50670-907, PE, Brazil.
[35]Hospital Universitario Mutua Terrassa, 08221 Terrassa, Spain.
[36]Instituto de Investigación Sanitaria de Santiago (IDIS), Xenética Cardiovascular, 15706 Santiago de Compostela, Spain.

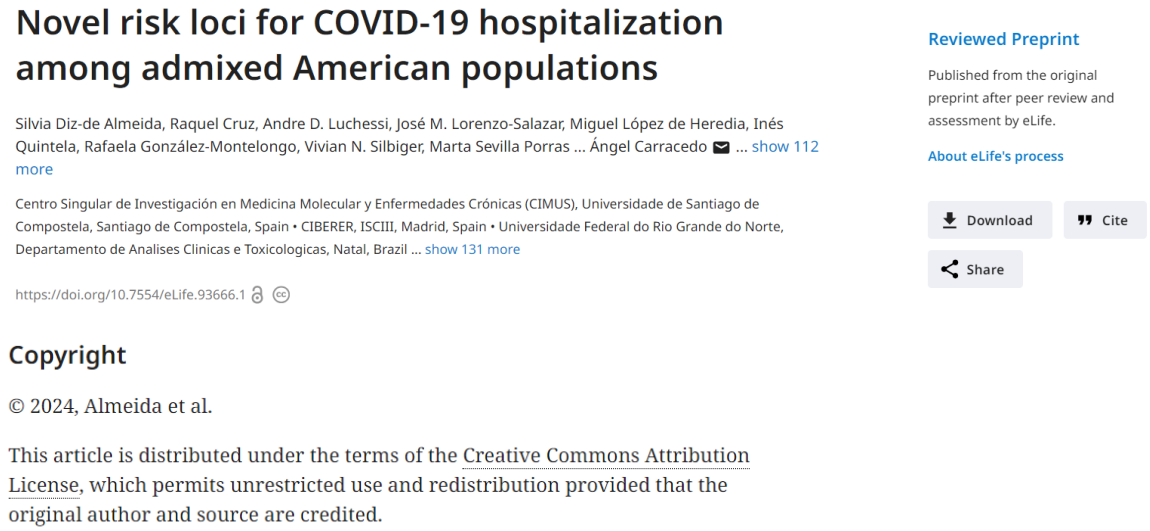
USC UNIVERSIDADE DE COMPOSTELA

[37]Centre for Biomedical Network Research on Cardiovascular Diseases (CIBERCV), Instituto de Salud Carlos III, 28029 Madrid, Spain.
[38]Cardiovascular Genetics Center, Institut d'Investigació Biomèdica Girona (IDIBGI), 17190 Girona, Spain.
[39]Medical Science Department, School of Medicine, University of Girona, 17190 Girona, Spain.
[40]Hospital Josep Trueta, Cardiology Service, 17190 Girona, Spain.
[41]Institute of Biomedicine of Seville (IBiS), Consejo Superior de Investigaciones Científicas (CSIC)- University of Seville-Virgen del Rocio University Hospital, 41013 Seville, Spain.
[42]Departemento de Medicina, Hospital Universitario Virgen del Rocío, Universidad de Sevilla, 41013 Seville, Spain.
[43]Centre for Biomedical Network Research on Epidemiology and Public Health (CIBERESP), Instituto de Salud Carlos III, 28029 Madrid, Spain.
[44]Instituto de Biomedicina de Sevilla, 41013 Seville, Spain.
[45]Hospital Universitario de Salamanca-IBSAL, Servicio de Medicina Interna, 37007 Salamanca, Spain.
[46]Osakidetza, Cruces University Hospital, 48903 Barakaldo, Bizkaia, Spain.
[47]Centre for Biomedical Network Research on Diabetes and Metabolic Associated Diseases (CIBERDEM), Instituto de Salud Carlos III, 28029 Madrid, Spain.
[48]University of Pais Vasco, UPV/EHU, 48903 Bizkaia, Spain.
[49]Oncology and Genetics Unit, Instituto de Investigacion Sanitaria Galicia Sur, Xerencia de Xestion Integrada de Vigo-Servizo Galego de Saúde, 36312 Vigo, Spain.
[50]Hospital Universitario Río Hortega, 47012 Valladolid, Spain.
[51]Servicio de Medicina intensiva, Complejo Hospitalario Universitario de A Coruña (CHUAC), Sistema Galego de Saúde (SERGAS), 15009 A Coruña, Spain.
[52]Tecnológico de Monterrey, 64718 Monterrey, Mexico.
[53]Otto von Guericke University, Departament of Microgravity and Translational Regenerative Medicine, 39106 Magdeburg, Germany.
[54]Hospital Universitario Mostoles, Unidad de Genética, 28935 Madrid, Spain.
[55]Instituto Aragonés de Ciencias de la Salud (IACS), 50009 Zaragoza, Spain.
[56]Instituto Investigación Sanitaria Aragón (IIS-Aragon), 50009 Zaragoza, Spain.
[57]Preventive Medicine Department, Instituto de Investigacion Sanitaria Galicia Sur, Xerencia de Xestion Integrada de Vigo-Servizo Galego de Saúde, 36312 Vigo, Spain.
[58]Unidad Diagnóstico Molecular. Fundación Rioja Salud, 26006 La Rioja, Spain.
[59]Universidade Federal do Rio Grande do Norte, Departamento de Analises Clinicas e Toxicologicas, 59012-570 Natal, Brazil.
[60]Hospital Universitario de Salamanca-IBSAL, Servicio de Cardiología, 37007 Salamanca, Spain.
[61]Instituto Regional de Investigación en Salud-Universidad Nacional de Caaguazú, HH36+J3Q Caaguazú, Paraguay.
[62]IDIVAL, 39008 Cantabria, Spain.
[63]Universidad de Cantabria, 39008 Cantabria, Spain.
[64]Hospital U M Valdecilla, 39008 Cantabria, Spain.
[65]Urgencias Hospitalarias, Complejo Hospitalario Universitario de A Coruña (CHUAC), Sistema Galego de Saúde (SERGAS), 15009 A Coruña, Spain.
[66]Grupo de Investigación en Interacciones Gen-Ambiente y Salud (GIIGAS) - Instituto de Biomedicina (IBIOMED), Universidad de León, 24071 León, Spain.
[67]Hospital Universitario de Getafe, Servicio de Genética, 28905 Madrid, Spain.
[68]Ministerio de Salud Ciudad de Buenos Aires, Buenos Aires C1425EFD CABA, Argentina.
[69]Hospital Universitario Virgen de las Nieves, Servicio de Análisis Clínicos e Inmunología, 18014 Granada, Spain.
[70]IIS La Fe, Plataforma de Farmacogenética, 46026 Valencia, Spain.
[71]Universidad de Valencia, Departamento de Farmacología, 46010 Valencia, Spain.
[72]Data Analysis Department, Instituto de Investigación Sanitaria-Fundación Jiménez Díaz University Hospital - Universidad Autónoma de Madrid (IIS-FJD, UAM), 28040 Madrid, Spain.
[73]Universidad de los Andes, Facultad de Ciencias, Bogotá 111711, Colombia.
[74]SIGEN Alianza Universidad de los Andes - Fundación Santa Fe de Bogotá, Bogotá 111711, Clombia.
[75]Hospital General de Segovia, Medicina Intensiva, 40002 Segovia, Spain.
[76]Hospital Universitario 12 de Octubre, Department of Immunology, 28041 Madrid, Spain.
[77]Instituto de Investigación Sanitaria Hospital 12 de Octubre (imas12), Transplant Immunology and Immunodeficiencies Group, 28041 Madrid, Spain.
[78]Instituto de Biomedicina (IBIOMED), Universidad de León, 24071 León, Spain.
[79]Fundación Santa Fe de Bogota, Departamento Patologia y Laboratorios, Bogotá 111711, Colombia.
[80]Unidad de Genética y Genómica Islas Baleares, 07120 Islas Baleares, Spain.
[81]Hospital Universitario Son Espases, Unidad de Diagnóstico Molecular y Genética Clínica, 07120 Islas Baleares, Spain.
[82]Genomics of Complex Diseases Unit, Research Institute of Hospital de la Santa Creu i Sant Pau, IIB Sant Pau, 08041 Barcelona, Spain.
[83]Faculdade de Medicina, Universidade de Brasília, Brasilia 70910-900, Brazil.
[84]Programa de Pós-Graduação em Ciências Médicas, Universidade de Brasília, Brasilia 70910-900, Brazil.
[85]Programa de Pós-Graduação em Ciências da Saúde, Universidade de Brasília, Brasilia 70910-900, Brazil.
[86]Hospital El Bierzo, Unidad Cuidados Intensivos, 24404 León, Spain.

[87]Hospital Universitario Mostoles, Medicina Interna, 28935 Madrid, Spain.
[88]Universidad Francisco de Vitoria, 28223 Madrid, Spain.
[89]Programa de Pós-Graduação em Biologia Animal, Universidade de Brasília, Brasília 70910-900, Brazil.
[90]Programa de Pós-Graduação Profissional em Ensino de Biologia, Universidade de Brasília, Brasília 70910-900, Brazil.
[91]Programa de Pós-Graduação Profissional em Ensino de Biologia (UnB), Universidade de Brasília, Brasília 70910-900, Brazil.
[92]Programa de Pós-Graduação em Ciências Médicas, Universidade de Brasília, Brasília 70910-900, Brazil.
[93]Universidad Complutense de Madrid, Department of Immunology, Ophthalmology and ENT, 28040 Madrid, Spain.
[94]Universidade Federal do Pará, Núcleo de Pesquisas em Oncologia, Belém, Pará 66075-110, Brazil.
[95]Infectious Diseases, Microbiota and Metabolism Unit, Center for Biomedical Research of La Rioja (CIBIR), 26006 Logroño, Spain.
[96]Inditex, 15141 A Coruña, Spain.
[97]GENYCA, 28220 Madrid, Spain.
[98]Instituto Mexicano del Seguro Social (IMSS), Centro Médico Nacional Siglo XXI, Unidad de Investigación Médica en Enfermedades Infecciosas y Parasitarias, Mexico City 02990, Mexico.
[99]Clinica Comfamiliar Risaralda, 660003 Pereira, Colombia.
[100]Bellvitge Biomedical Research Institute (IDIBELL), Neurometabolic Diseases Laboratory, 08908 L'Hospitalet de Llobregat, Spain.
[101]Catalan Institution of Research and Advanced Studies (ICREA), 08010 Barcelona, Spain.
[102]Hospital Ophir Loyola, Departamento de Ensino e Pesquisa, Belém, Pará 66063-240, Brazil.
[103]Unidad de Cuidados Intensivos, Hospital Clínico Universitario de Santiago (CHUS), Sistema Galego de Saúde (SERGAS), 15706 Santiago de Compostela, Spain.
[104]Department of Preventive Medicine and Public Health, School of Medicine, Universidad Autónoma de Madrid, 28049 Madrid, Spain.
[105]IdiPaz (Instituto de Investigación Sanitaria Hospital Universitario La Paz), 28046 Madrid, Spain.
[106]IMDEA-Food Institute, CEI UAM+CSIC, 28049 Madrid, Spain.
[107]Complejo Asistencial Universitario de León, 24071 León, Spain.
[108]Instituto de Investigación Biosanitaria de Granada (ibs GRANADA), 18012 Granada, Spain.
[109]Universidad de Granada, Departamento Bioquímica, Biología Molecular e Inmunología III, 18071 Granada, Spain.
[110]Hospital Infanta Elena, Allergy Unit, Valdemoro, 28342 Madrid, Spain.
[111]Instituto de Investigación Sanitaria-Fundación Jiménez Díaz University Hospital - Universidad Autónoma de Madrid (IIS-FJD, UAM), 28040 Madrid, Spain.
[112]Faculty of Medicine, Universidad Francisco de Vitoria, 28223 Madrid, Spain.
[113]Hospital Universitario Infanta Leonor, 28031 Madrid, Spain.
[114]Complutense University of Madrid, 28040 Madrid, Spain.
[115]Gregorio Marañón Health Research Institute (IiSGM), 28007 Madrid, Spain.
[116]Haemostasis and Thrombosis Unit, Hospital de la Santa Creu i Sant Pau, IIB Sant Pau, 08041 Barcelona, Spain.
[117]Hospital Clinico Universitario de Valladolid, Servicio de Anestesiologia y Reanimación, 47003 Valladolid, Spain.
[118]Universidad de Valladolid, Departamento de Cirugía, 47005 Valladolid, Spain.
[119]Hospital Clinico Universitario de Valladolid, Servicio de Hematologia y Hemoterapia, 47003 Valladolid, Spain.
[120]Hospital de Niños Ricardo Gutierrez, Buenos Aires C1425EFD CABA, Argentina.
[121]Fundación Universitaria de Ciencias de la Salud, 113827 Bogotá, Colombia.
[122]Spanish National Cancer Research Centre, Familial Cancer Clinical Unit, 28029 Madrid, Spain.
[123]University Hospital of Burgos, 09006 Burgos, Spain.
[124]Universidad Simón Bolívar, Facultad de Ciencias de la Salud, 080002 Barranquilla, Colombia.
[125]Centro para el Desarrollo de la Investigación Científica, 1255 Asunción, Paraguay.
[126]Institute for Molecular Medicine Finland (FIMM), 00014 Univerisity of Helsinki, Finland.
[127]McGill University, Department of Human Genetics, H3A 0G4 Montréal, Québec, Canada.
[128]Lady Davis Institute, Jewish General Hospital, McGill University, H3T 1E2 Montréal, Québec, Canada.
[129]Kyoto-McGill International Collaborative School in Genomic Medicine, Graduate School of Medicine, Kyoto University, 606-8501 Kyoto, Japan.
[130]Research Fellow, Japan Society for the Promotion of Science, 102-0083 Tokyo, Japan.
[131]University of Milano-Bicocca, 20126 Milano, Italy.
[132]Institute for Molecular Medicine Finland, Univerisity of Helsinki, 00014 Helsinki, Finland.
[133]Institute of Clinical Molecular Biology, Christian-Albrechts-University, 24118 Kiel, Germany.
[134]University Hospital Schleswig-Holstein, Campus Kiel, 24118 Kiel, Germany.
[135]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, H3A 0G4 Montréal, Québec, Canada.
[136]Digestive Diseases Unit, Virgen del Rocio University Hospital, Institute of Biomedicine of Seville, University of Seville, 41103 Seville, Spain.
[137]Centre for Biomedical Network Research on Hepatic and Digestive Diseases (CIBEREHD), Instituto de Salud Carlos III, 28029 Madrid, Spain.
[138]Department of Liver and Gastrointestinal Diseases, Biodonostia Health Research Institute - Donostia University Hospital, University of the Basque Country (UPV/EHU), 20014 San Sebastian, Spain.

[139]Centre de Génétique Humaine, Hôpital Erasme, Université Libre de Bruxelles, 1070 Brussels, Belgium.
[140]Liver Unit, Department of Internal Medicine, Hospital Universitari Vall d'Hebron, Vall d'Hebron Barcelona Hospital Campus, 08035 Barcelona, Spain.
[141]Universitat Autònoma de Barcelona, Departament de Medicina, Bellatera, 08193 Barcelona, Spain.
[142]Vall d'Hebron Institut de Recerca (VHIR), Liver Diseases, 08035 Barcelona, Spain.
[143]Novo Nordisk Foundation Center for Protein Research, Disease Systems Biology, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark.
[144]GENYO, Centre for Genomics and Oncological Research: Pfizer, University of Granada, Andalusian Regional Government, 18016 Granada, Spain.
[145]Institute of Human Genetics, University Hospital Bonn, Medical Faculty University of Bonn, 53127 Bonn, Germany.
[146]Genetica Medica, Azienda Ospedaliero-Universitaria Senese, 53100 Siena, Italy.
[147]University of Siena, Medical Genetics, 53100 Siena, Italy.
[148]Azienda Ospedaliero-Universitaria Senese, Genetica Medica, 53100 Siena, Italy.
[149]Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, 53100 Siena, Italy.
[150]Technical University of Munich, School of Medicine, University Hospital rechts der Isar, Department of Internal Medicine II, 80333 Munich, Germany.
[151]Department of Transfusion Medicine and Hematology, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Università degli Studi di Milano, 20126 Milano, Italy.
[152]Mucosal Immunology Lab, Unidad de Excelencia del Instituto de Biomedicina y Genética Molecular (IBGM, Universidad de Valladolid-CSIC), 47005 Valladolid, Spain.
[153]Hospital Universitario Clinico San Cecilio, 18016 Granada, Spain.
[154]Instituto de Investigación Ibs, Granada, 18012 Granada, Spain.
[155]University of Liege. GIGA-Insitute, B- 4000 Liege, Belgium.
[156]Liege University Hospital (CHU of Liege), B- 4000 Liege, Belgium.
[157]Biomedical Research Institute Sant Pau (IIB Sant Pau), Stroke Pharmacogenomics and Genetics Group, 08041 Barcelona, Spain.
[158]Oslo University Hospital, Department of Microbiology, 0424 Oslo, Norway.
[159]Institute of Clinical Medicine, University of Oslo, 0424 Oslo, Norway.
[160]Department of Liver and Gastrointestinal Diseases, Biodonostia Health Research Institute - Donostia University Hospital, University of the Basque Country (UPV/EHU), Ikerbasque, 20014 San Sebastian, Spain.
[161]Department of Surgical Sciences, Anaesthesiology and Intensive Care Medicine, Uppsala University, 751 05 Uppsala, Sweden.
[162]Humanitas University, Department of Biomedical Sciences, 20089 Milan, Italy.
[163]IRCCS Humanitas Research Hospital, Rozzano, 20089 Milan, Italy.
[164]Heart Institute (InCor)/University of Sao Paulo Medical School, 05508-070 Sao Paulo, Brazil.
[165]Osakidetza Basque Health Service, Donostialdea Integrated Health Organisation, Clinical Biochemistry Department, 20006 San Sebastian, Spain.
[166]Norwegian PSC Research Center and Section of Gastroenterology, Dept Transplantation Medicine, Oslo University Hospital, 0424 Oslo, Norway.
[167]Research Institute of Internal Medicine, Oslo University Hospital, 0424 Oslo, Norway.
[168]Fonds de la Recherche Scientifique (FNRS), B - 1000 Brussels.
[169]Institute for Biomedical Research of Barcelona (IIBB), National Spanish Research Council (CSIC), 08028 Barcelona, Spain.
[170]Institut d'Investigacions Biomediques August Pi i Sunyer (IDIBAPS), 08036 Barcelona, Spain.
[171]Institute for Environmental Medicine, Karolinska Institutet, 171 65 Solna, Sweden.
[172]Institute of Virology, Technical University Munich/Helmholtz Zentrum München, D-85764 Munich, Germany.
[173]Institute of Psychiatric Phenomics and Genomics, University Hospital, LMU Munich University, 80539 Munich, Germany.
[174]Department of Psychiatry, University Hospital, LMU Munich University, 80539 Munich, Germany.
[175]Research Institute of Internal Medicine, Oslo University Hospital, 0318 Oslo, Norway.
[176]Università degli Studi di Milano, Department of Pathopgysiology and Transplantation, 20126 Milano, Italy.
[177]Department of Transfusion Medicine and Hematology, Precision Medicine, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, 20122 Milano, Italy.
[178]Karolinska Institutet, Department of Neuroscience, 171 77 Stockholm, Sweden.
[179]Max-Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany.
[180]Department of Human Genetics, Epidemiology, Biostatistics and Occupational Health, McGill University, H3A 0G4 Montréal, Québec, Canada.
[181]King's College London, Department of Twin Research, London, WC2R 2LS, United Kingdom.
[182]Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA.
[183]Research Center and Memory clinic, ACE Alzheimer Center Barcelona, Universitat Internacional de Catalunya, 08028 Barcelona, Spain.
[184]Centre for Biomedical Network Research on Neurodegenerative Diseases (CIBERNED), Instituto de Salud Carlos III, 28029 Madrid, Spain.

[185]CIEN Foundation/Queen Sofia Foundation Alzheimer Center, 28031 Madrid, Spain.
[186]Hospital Universitario de Valme, Unidad Clínica de Enfermedades Infecciosas y Microbiología, 41014 Sevilla, Spain.
[187]Sección Genética Médica - Servicio de Pediatría, Hospital Clínico Universitario Virgen de la Arrixaca, Servicio Murciano de Salud, 30120 Murcia, Spain.
[188]Departamento Cirugía, Pediatría, Obstetricia y Ginecología, Facultad de Medicina, Universidad de Murcia (UMU), 30100 Murcia, Spain.
[189]Grupo Clínico Vinculado, Centre for Biomedical Network Research on Rare Diseases (CIBERER), Instituto de Salud Carlos III, 28029 Madrid, Spain.
[190]Tecnologico de Monterrey, Escuela de Medicina y Ciencias de la Salud, 64718 Monterrey, Mexico.
[191]Research Unit, Hospital Universitario N.S. de Candelaria, 38010 Santa Cruz de Tenerife, Spain.
[192]Centre for Biomedical Network Research on Respiratory Diseases (CIBERES), Instituto de Salud Carlos III, 28029 Madrid, Spain.
[193]Facultad de Ciencias de la Salud, Universidad Fernando Pessoa Canarias, 35450 Las Palmas de Gran Canaria, Spain.



Novel risk loci for COVID-19 hospitalization among admixed American populations

Silvia Diz-de Almeida, Raquel Cruz, Andre D. Luchessi, José M. Lorenzo-Salazar, Miguel López de Heredia, Inés Quintela, Rafaela González-Montelongo, Vivian N. Silbiger, Marta Sevilla Porras ... Ángel Carracedo ... show 112 more

Centro Singular de Investigación en Medicina Molecular y Enfermedades Crónicas (CIMUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain • CIBERER, ISCIII, Madrid, Spain • Universidade Federal do Rio Grande do Norte, Departamento de Analises Clinicas e Toxicologicas, Natal, Brazil ... show 131 more

https://doi.org/10.7554/eLife.93666.1

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Download Cite Share

Copyright

© 2024, Almeida et al.

This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

Novel risk loci for COVID-19 hospitalization among admixed American populations.

Preprint revisado á espera de enviar as correccións.

eLife13:RP93666

DOI: https://doi.org/10.7554/eLife.93666.1

Índices de calidade JCR – Biology (2022):

Factor de impacto: 7.7

Posición: 7/92 (Mellor cuartil – Q1; mellor percentil – 92.9)

Contribución ao traballo da doutoranda: Co-primeira autoría. Análise dos datos; interpretación dos resultados; redacción do artigo.

Lista de autores e filiacións:
https://docs.google.com/document/d/1gJVHCOM59Yczz6BduHfpyOQEmxIXcBXr/edit

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

11 ÍNDICE DE TÁBOAS

Táboa 1. Escala de gravidade de 5 niveis baixo a que se clasifican os doentes de SCOURGE..... 56
Táboa 2. Lista de centros españois que contribuíron ás mostras incluídas na cohorte española..... 58
Táboa 3. Lista de centros latinoamericanos e españois que contribuíron ás mostras incluídas na cohorte LatAm. 59
Táboa 4. Características basais dos pacientes da cohorte española de SCOURGE incluídos na análise 83
Táboa 5. Definicións de caso e control para os modelos estatísticos 84
Táboa 6. Resultados das análises A1 en hospitalización e gravidade crítica e susceptibilidade á infección, así como estratificadas por sexo 88
Táboa 7. Resultados das análises A1 en hospitalización e gravidade crítica e susceptibilidade á infección estratificadas por idade 89
Táboa 8. Resultados da meta-análise de hospitalización entre os estudos SCOUGE e CNIO, tanto en global como estratificados por sexo 92
Táboa 9. Resultados da análise a nivel de xen de MAGMA na meta-análise de SCOURGE-CNIO para hospitalización por COVID-19, análise A1 (hg37) 92
Táboa 10. Resultados da asociación TWAS nos tecidos elixidos 95
Táboa 11. Resultados da meta-análise dos estudos europeos para risco de hospitalización. 98
Táboa 12. Características basais dos pacientes da cohorte LatAm de SCOURGE incluídos na análise 106
Táboa 13. Cohortes adicionais incluídas na meta-análise SCOURGE-HGI_{AMR}. 105
Táboa 14. Variantes líder na meta-análise de HGI AMR con SCOURGE Latino-América..... 107
Táboa 15. Resultados da análise a nivel de xen de MAGMA na meta-análise de HGI AMR con SCOURGE Latino-América (hg37). 108
Táboa 16. Resultados da asociación TWAS nos tecidos elixidos. 109
Táboa 17. Descrición dos PRS en función do método, datos de descubrimento e cohorte diana. 119
Táboa 18. Incremento no pseudo-R^2 de cada PRS xunto cos seus intervalos de confianza ao 95% calculados con bootstrap 123
Táboa 19. Resultados da validación externa dos tres modelos co PRS_{ax} na cohorte adicional de 437 individuos con GIA predominantemente europea. 126
Táboa 20. Resumo das variables empregadas na valoración do PRS con variables clínicas. 127
Táboa 21. Resultados das interaccións co PRS con factores de risco clínicos..... 130
Táboa 22. Valores de especificidade e sensibilidade para a mostra test dos modelos base e principal en función de distintos cortes de probabilidades preditas. 131

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

12 ÍNDICE DE FIGURAS

Figura 1. Esquema do establecemento dun Consorcio GWAS típico. 18
Figura 2. Relación entre a frecuencia alélica e tamaño de efecto nos GWAS.. 21
Figura 3. Ancestralidade global e ancestralidade local. 25
Figura 4. Diversidade nos estudos GWAS dende 2007 ata 2024. 28
Figura 5. Estimas do exceso de mortalidade por COVID-19 por cada 100.000 habitantes. 41
Figura 6. Proporción relativa de individuos hospitalizados entre as diferentes poboacións incluídas na análise B2 do consorcio HGI. O..... 46
Figura 7. Xenes identificados a través de estudos GWAS implicados na infección e gravidade da COVID-19. Adaptada de Kanai e colaboradores (2023). 48
Figura 8. Exemplo dunha curva de decisión na que se valoran un biomarcador e un modelo de risco. 81
Figura 9. Resultados da asociación da cohort SCOURGE na análise de A1 hospitalización. 85
Figura 10. Gráficos rexionais dos dous sinais en *LZTFL1* (arriba) e *IFNAR2* (abaixo)..... 86
Figura 11. Gráfico rexional do sinal atopado en mulleres no xen *TLE1*.. 87
Figura 12. Manhattan plot da meta-análise CNIO-SCOURGE para o fenotipo de hospitalización (análise A1)..... 91
Figura 13. Gráficos rexionais dos dous sinais en *AQP3*, *ARHGAP27*, e *UPK1A*/*ZTBT32*..... 94
Figura 14. Figura tipo manhattan plot para os resultados do TWAS multi-tecido..... 96
Figura 15. Hitmap de replicación dos hits GWAS para hospitalización. 97
Figura 16. Representación da ancestralidade dos individuos da cohorte Latinoamericana. 106
Figura 17. Manhattan plot da meta-análise de HGI AMR con SCOURGE Latino-América. 108
Figura 18. Distribución das frecuencias alélicas dos SNPs rs1003835 (chr2) e do rs77599934 (chr11) alélicas nas poboacións de 1000 Genomas e os gráficos rexionais correspondentes. 109
Figura 19. Gráfico rexional para a rexión do cromosoma 2 priorizada por mapeamento fino baiesiano. 110
Figura 20. Gráfico tipo Manhattan Plot para o TWAS multi-tecido en poboación AMR. 112
Figura 21. Gráfico tipo bosque mostrando os tamaños de efecto e os intervalos de confianza correspondentes para as variantes líder identificadas na meta-análise de HGI AMR con SCOURGE Latino-América nas poboacións AFR, AMR, EAS, EUR e SAS.. 114
Figura 22. Representación gráfica do incremento no R2 parcial (y) asociado aos PRS C+T multi-poboación en función dos hiperparámetros (x).. 122
Figura 23. Representación gráfica do incremento no R2 parcial (y) asociado aos PRS C+T europeos en función dos hiperparámetros (x).. 123
Figura 24. Representación gráfica do incremento no R2 parcial (y) asociado aos PRS C+T multi-poboación (A) e europeos (B) en función dos hiperparámetros (x) para a variable CSEV4. 124
Figura 25. Gráfico de calibración do modelo completo na mostra de adestramento co PRSax tras 1000 remostras bootstrap. 126
Figura 26. Gráfico de densidade do PRS axustado polas PCs nas cohortes latina e española, estratificado en función do estado caso/control. 127
Figura 27. Gráficos de caixas que ilustra a distribución dos percentís do PRS entre casos e controis (A) e gráfico de prevalencia observada *vs* predita en cada percentil do PRS (B)..... 127
Figura 28. Gráfico tipo bosque que ilustra as OR do PRS estandarizado para cada variable na cohorte conxunta dentro do total de casos COVID-19+. 130
Figura 29. Gráfico tipo bosque que ilustra as OR do PRS estandarizado para cada variable na análise de sensibilidade na cohorte conxunta dentro de casos hospitalizados.....131
Figura 30. Gráficos de calibración por bootstrap para os modelos FRC (A) e FRC+PRS (B)... 133

Figura 31. Gráfico que ilustra as predicións da mostra test para os modelos base, con factores de risco clínicos (FRC), e o modelo principal, con factores de risco clínicos e o PRS (FRC+PRS).. 134
Figura 32. Curva DCA calculada coa mostra test, onde se representan 4 estratexias. Estas son: tratar a toda a poboación, non tratar a ningunha, e tratar en función dos modelos FRC e FRC+PRS..... 135
Figura 33. Curva DCA calculada coa mostra test, onde se representan 5 estratexias diferentes..... 136

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

A COVID-19 presenta unha gran variabilidade fenotípica entre os individuos. Os factores xenéticos do hóspede xogan un papel fundamental, influíndo na susceptibilidade á infección e na posterior resposta inmune. Nesta tese doutoral, investigouse a base xenética da enfermidade a través do estudo de máis de 10.000 pacientes recollidos en España e en Latino-América. Mediante a realización de estudos de xenoma completo (GWAS) en ambas poboacións, identificáronse novos xenes asociados e replicáronse outros xa coñecidos. Ademais, desenvolvéronse índices de risco polixénico (PRS) e avaliouse a súa capacidade preditiva en modelos de risco para aenfermidade.