Title Page

Title: Cumulative genetic score and *C9orf72* repeat status independently contribute to ALS risk in two case-control studies

Authors: John F. Dou[1], Kelly M. Bakulski[1], Kai Guo[2,3], Junguk Hur[4], Lili Zhou[5], Sara Saez-Atienzar[6], Ali R Stark[6], Ruth Chia[6], Alberto García-Redondo[7,8], Ricardo Rojas-García[8,9], Juan Francisco Vázquez-Costa[8,10,11], Ruben Fernandez Santiago[12,13,14], Sara Bandres-Ciga[15], Pilar Gómez-Garre[12,16], Maria Teresa Periñán[12,16], Pablo Mir[12,16,17], Jordi Pérez-Tur[12,18,19], Fernando Cardona[12,18,19], Manuel Menendez-Gonzalez[20,21,22], Javier Riancho[23,24], Daniel Borrego-Hernández[7,8], Lucía Galán-Dávila[25], Jon Infante Ceberio[24], Pau Pastor[26,27], Carmen Paradas[28], Oriol Dols-Icardo[12,29,], Spanish Neurological Consortium[a], Bryan J. Traynor[6], Eva L. Feldman[2,3], and Stephen A Goutman[2,3]

[a]Members of the Spanish Neurological Consortium: Jesús Esteban-Pérez[7,8], Pilar Cordero-Vázquez[7,8], Sevilla Teresa[8,10,30], Adolfo López de Munain[12,31,32], Julio Pardo-Fernández[23,33], Ivonne Jericó-Pascual[34,35], Ellen Gelpi Mantius[37,38], Janet Hoenicka[8,39], Victoria Alvarez Martinez[40,41], Francisco Javier Rodríguez de Rivera Garrido[42], Katrin Beyer[43, 44], Jordi Clarimón Echevarría[12, 29]

Affiliations:

[1]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI

[2]Department of Neurology, University of Michigan, Ann Arbor, MI

[3]NeuroNetwork for Emerging Therapies, University of Michigan, Ann Arbor, MI

[4]Department of Biomedical Sciences, University of North Dakota, Grand Forks, ND

[5]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI

[6]Neuromuscular Diseases Research Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD

[7]ALS Unit, Instituto de Investigación Sanitaria ‘i + 12’ del Hospital Universitario 12 de Octubre de Madrid, SERMAS, Madrid, Spain

[8]CIBERER, Center for Networked Biomedical Research into Rare Diseases, Madrid, Spain

[9]Neuromuscular Disorders Unit, Neurology Department and Sant Pau Biomedical Research Institute, Hospital de la Santa Creu I Sant Pau, Universitat Autonoma de Barcelona, Barcelona, Spain

[10]Neuromuscular Unit, Hospital Universitario y Politécnico la Fe, IIS La Fe, Valencia, Spain

[11]Department of Medicine, Universitat de València, Valencia, Spain

[12]Centro de Investigación Biomédica en Red sobre Enfermedades Neurodegenerativas (CIBERNED), Madrid, Spain

[13]Lab of Parkinson's disease and Other Neurodegenerative Movement Disorders, IDIBAPS-Institut d'Investigacions Biomèdiques, Barcelona, Catalonia, Spain

[14]Unitat de Parkinson i Trastorns del Moviment. Servicio de Neurologia, Hospital Clínic de Barcelona and Institut de Neurociencies de la Universitat de Barcelona (Maria de Maetzu Center), Catalonia, Spain

[15]Center for Alzheimer's and Related Dementias, National Institute on Aging, Bethesda, MD

[16]Unidad de Trastornos del Movimiento, Servicio de Neurología y Neurofisiología Clínica, Instituto de Biomedicina de Sevilla, Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Sevilla, Spain

[17]Departamento de Medicina, Universidad de Sevilla, Sevilla, Spain

[18]Neurology and Molecular Genetics Mixed Investigation Unit. Instituto de Investigación Sanitaria La Fe, Valencia, Spain

[19]Molecular Genetics Unit. Institut de Biomedicina de València-CSIC, Valencia, Spain
[20]Department of Medicine, Universidad de Oviedo, Oviedo, Spain
[21]Department of Neurology, Hospital Universitario Central de Asturias, Oviedo, Spain
[22]Instituto de Investigación Sanitaria del Principado de Asturias, Oviedo, Spain
[23]Service of Neurology. Hospital Sierrallana. IDIVAL University of Cantabria, Barrio Ganzo s/n. 39300. Torrelavega. Spain
[24]Instituto de Investigación Marqués de Valdecilla, Santander, Spain
[25]Department of Neurology, ALS Unit, Hospital Clínico Universitario 'San Carlos', Madrid, Spain
[26]Unit of Neurodegenerative diseases, Department of Neurology, University Hospital Germans Trias I Pujol, Badalona, Barcelona, Spain
[27]Neurosciences, The Germans Trias i Pujol Research Institute (IGTP) Badalona, Barcelona, Spain
[28]Department of Neurology, Hospital Universitario Virgen del Rocio, Sevilla, Spain
[29]Memory Unit, Neurology Department and Sant Pau Biomedical Research Institute, Hospital de la Santa Creu I Sant Pau, Universitat Autonoma de Barcelona, Barcelona, Spain
[30]Universitat de Valencia, Valencia, Spain
[31]Neuroscience Area, Institute Biodonostia, and Department of Neurosciences, University of Basque Country EHU-UPV, San Sebastian, Spain
[32]Neurology Department, Hospital Universitario Donostia, San Sebastian, Spain
[33]Neurology Department, Hospital Clinico, Santiago de Compostela, Spain
[34]Neurology Department. Complejo Hospitalario de Navarra, Pamplona, Spain
[35]Department of Neurology. ALS Clinic. Hospital Universitario de Navarra. IdisNa (Instituto de Investigación Sanitaria de Navarra), Navarra, Spain
[36]Memory Unit, Neurology Department and Sant Pau Biomedical Research Institute, Hospital de la Santa Creu I Sant Pau, Universitat Autonoma de Barcelona, Barcelona, Spain
[37]Neurological Tissue Bank of the Biobank-Hospital Clinic-IDIBAPS, Barcelona, Spain
[38]Institute of Neurology, Medical University of Vienna, Vienna, Austria
[39]Laboratory of Neurogenetics and Molecular Medicine-Pediatric Institute of Rare Diseases, Institut de Recerca Sant Joan de Déu, Barcelona, Spain
[40]Laboratorio de Genética, Hospital Universitario Central de Asturias, Asturias, Spain
[41]Instituto de Investigación Sanitaria del Principado de Asturias (ISPA), Asturias, Spain
[42]Department of Neurology, ALS Unit, Hospital Universitario La Paz, Madrid, Spain
[43]Department of Pathology, University Hospital Germans Trias I Pujol, Badalona, Barcelona, Spain
[44]Neurosciences, The Germans Trias i Pujol Research Institute (IGTP) Badalona, Barcelona, Spain

Corresponding Author:
Stephen A. Goutman, MD, MS
Department of Neurology
1500 E Medical Center Dr
Ann Arbor, MI 48109-5223
Phone: 734-936-8586
Fax: 734-936-5185
e-mail: sgoutman@med.umich.edu

Authors

John F. Dou	johndou@umich.edu	0000-0003-4577-8660
Kelly M. Bakulski	bakulski@umich.edu	0000-0002-9605-6337
Kai Guo	kaiguo@umich.edu	0000-0002-4651-781X
Junguk Hur	junguk.hur@med.und.edu	0000-0002-0736-2149

Lili Zhou	zhaolili@umich.edu	
Sara Saez-Atienzar	sara.saez@nih.gov	0000-0002-1524-9584
Ali Stark	ali22stark@gmail.com	0000-0002-1271-951X
Ruth Chia	ruth.chia@nih.gov	0000-0002-4709-7423
Alberto García-Redondo	ela@correo.h12o.es	
Ricardo Rojas-García	rrojas@santpau.cat	0000-0003-1411-5573
Juan Francisco Vázquez-Costa	vazquez_juacos@gva.es	
Ruben Fernandez Santiago	ruben.fernandez.santiago@googlemail.com	
Sara Bandres-Ciga	sara.bandresciga@nih.gov	
Pilar Gómez-Garre	mgomez-ibis@us.es	
Maria Teresa Periñán	tperinan-ibis@us.es	
Pablo Mir	pmir@us.es	
Jordi Pérez-Tur	jpereztur@ibv.csic.es	
Fernando Cardona		
Manuel Menendez-Gonzalez	menendezgmanuel@uniovi.es	0000-0002-5218-0774
Javier Riancho	javier.riancho86@gmail.com	0000-0001-7929-1055
Daniel Borrego-Hernández	dborregohernandez.imas12@h12o.es	
Lucia Galán-Dávila	lucgalan@ucm.es	
Jon Infante Ceberio	jon.infante@unican.es	
Pau Pastor	ppastor@unav.es	
Carmen Paradas	cparadas@us.es	
Oriol Dols-Icardo	Odols@santpau.cat	
Bryan J. Traynor	bryan.traynor@nih.gov	0000-0003-0527-2446
Eva L. Feldman	efeldman@umich.edu	0000-0002-9162-2694
Stephen A. Goutman	sgoutman@med.umich.edu	0000-0001-8780-6637

Spanish Neurological Consortium

Jesús Esteban-Pérez	jesusesteban@h12o.es	
Pilar Cordero-Vázquez	mariadelpilar.cordero@salud.madrid.org	0000-0002-9239-5147
Sevilla Teresa	sevilla_ter@gva.es	
Adolfo Lopez de Munain	adolfo.lopezdemunainarregui@osakidetza.eus	
Julio Pardo-Fernández	julio.pardo@usc.es	0000-0001-8807-1310
Ivonne Jericó-Pascual	ijericop@navarra.es	
Oriol Dols-Icardo	odols@santpau.cat	
Ellen Gelpi Mantius	ellen.gelpimantius@meduniwien.ac.at	
Janet Hoenicka	jhoenicka@fsjd.org	
Victoria Alvarez	victoria.alvarez@sespa.es	0000-0002-1916-2523
Francisco Javier Rodríguez de Rivera Garrido	frrivera.garrido@salud.madrid.org	
Katrin Beyer	katrinbeyer@hotmail.com	
Jordi Clarimón Echevarría	jclarimon@santpau.cat	

Abstract word count: 261
Main text word count: 4470

Abstract

Background and Objectives: Most amyotrophic lateral sclerosis (ALS) patients lack a monogenic mutation. This study evaluates ALS cumulative genetic risk in an independent Michigan and Spanish replication cohort using polygenic scores.
Methods: Participant samples from University of Michigan were genotyped and assayed for the *C9orf72* hexanucleotide expansion. Final cohort size was 219 ALS and 223 healthy controls following genotyping and participant filtering. Polygenic scores excluding the C9 region were generated using an independent ALS genome-wide association study (20,806 cases, 59,804 controls). Adjusted logistic regression and receiver operating characteristic curves evaluated the association and classification between polygenic scores and ALS status, respectively. Population attributable fractions and pathway analyses were conducted. An independent Spanish study sample (548 cases, 2,756 controls) was used for replication.
Results: Polygenic scores constructed from 275 single nucleotide polymorphisms had the best model fit in the Michigan cohort. A standard deviation increase in ALS polygenic score associated with 1.28 (95%CI 1.04-1.57) times higher odds of ALS with area under the curve of 0.663 versus a model without the ALS polygenic score (p-value=1x10^{-6}). The population attributable fraction of the highest 20th percentile of ALS polygenic scores, relative to the lowest 80th percentile, was 4.1% of ALS cases. Genes annotated to this polygenic score enriched for important ALS pathomechanisms. Meta-analysis with the Spanish study, using a harmonized 132 single nucleotide polymorphism polygenic score, yielded similar logistic regression findings (odds ratio: 1.13, 95%CI 1.04-1.23).
Discussion: ALS polygenic scores can account for cumulative genetic risk in populations and reflect disease-relevant pathways. If further validated, this polygenic score will inform future ALS risk models.

Keywords: amyotrophic lateral sclerosis, polygenic risk, polygenic scores, classification

Main Text

Introduction

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease characterized by rapidly progressive muscle weakness and death within 2 to 4 years from symptom onset[1, 2] with 50% of patients manifesting cognitive or behavioral dysfunction.[1, 2] Although ALS is traditionally divided into familial and sporadic forms, with familial ALS indicating those with an ALS family history, ALS genetic risk factors are present in both familial and sporadic patients.[3] Under a monogenic model, a single risk gene is associated with a greater likelihood of developing ALS[4] or contributes to a distinct phenotypic outcome, such as earlier age of disease onset.[4, 5] Since 1994, over 40 genes have been associated with ALS.[6] The non-coding chromosome 9 open reading frame 72 (*C9orf72*) hexanucleotide expansion is the most common genetic form of ALS and is observed in 40% of familial and 10% of sporadic ALS in mixed European populations.[7, 8] Superoxide dismutase 1 (*SOD1*), TAR DNA binding protein 43 (*TARDBP*), and fused in sarcoma (*FUS*) are the next most common genes with polymorphism frequencies of around 1% or less in sporadic cases.[9] Importantly, most ALS patients do not carry a single causative ALS risk gene mutation. This highlights the notion of heritability, which captures the genetic and shared familial factors that contribute to disease risk.[10] Heritability is as high as 38-85% when considering twin data[11], 36.9-52.3% for parent-offspring pairs,[10] 43% for all first-degree relatives, [12] and 7.2-9.5% for common single nucleotide polymorphisms (SNPs).[13-16] It is increasingly clear that many common SNPs may contribute a small amount of disease risk.[17] Since most ALS patients do not have a monogenic cause, it is crucial to understand the genetic contribution to ALS beyond single highly penetrant mutations to stratify population risk.

We hypothesize that polygenic scores will improve ALS risk prediction. To our knowledge, the utility of a polygenic score for ALS, independent of *C9orf72* status, has not been tested for ALS risk prediction. The goals of the current study were to develop a genome-wide ALS polygenic score using an independent ALS cohort of participants not previously included in any genome-wide association study (GWAS) and test the score contribution to ALS risk models independently of *C9orf72* status.

Methods

Michigan Study Participants and Sample Collection

All patients seen at the University of Michigan Pranger ALS Clinic are invited to participate, although the present case/control analysis is limited to those with ALS, thereby excluding participants with other forms of motor neuron disease. Healthy controls, without a personal or family history of a neurodegenerative disease in a first- or second-degree family member, are identified using a recruitment database available through the Michigan Institute for Clinical & Health Research and through population outreach via random address mailings. Participant demographics including sex (male, female), race/ethnicity (White or Caucasian, Black or African American, or Asian and Hispanic or Latino), and age (years) were obtained at the time of study enrollment. ALS diagnoses were confirmed by an ALS neurologist (S.A.G., E.L.F.) based on Gold Coast Criteria, who also recorded onset age (years), diagnosis age (years), onset segment (bulbar, cervical, lumbar, respiratory, thoracic), and presence of an ALS family history (yes or no) in the medical record. A family history of ALS in a first- or second-degree relative is considered positive. All participants provide venous blood, collected in an EDTA tube and frozen at -80 °C for later DNA extraction.

Standard Protocol Approvals, Registrations, and Patient Consents

Study procedures of this Institutional Review Board (HUM28826) approved longitudinal case/control study are published.[18-20] All participants provided informed consent.

DNA Analysis

DNA was extracted using the QIAamp DNA Kit (Qiagen, Venlo, Netherlands). Genome-wide genotypes at 1,748,250 positions were measured for 512 samples using the Infinium Multi-Ethnic Global-8 v1.0 array kit (Illumina, San Diego, CA) by the University of Michigan Advanced Genomics Core. All available clinical samples, including intentional duplicates (n=6) and non-ALS diseased samples (6 primary lateral sclerosis, 12 other motor neuron disease), were included at this step to improve imputation quality. DNA samples were also analyzed for the presence of the *C9orf72* repeat expansion per published protocols.[7]

PLINK (version 1.9) program was used to perform genetic microarray data quality control checks.[21] Participants and single nucleotide polymorphisms (SNPs) were filtered using recommended thresholds.[22, 23] Participants were excluded for missing data at greater than 1% of SNPs, discrepancies between genetic sex and predicted sex, and heterozygosity greater than three standard deviations from the mean. For intentional technical duplicate samples and unintentional related samples, the sample in each pair with the highest missingness was excluded. Participant inclusion based on genetic data quality control was visualized using a flow diagram and 488 unique motor neuron disease and control participants met genetic quality filtering (**eFigure 1A**).

SNPs were excluded for missing genomic location data or missingness frequency in over 1% of samples. SNPs from autosomal chromosomes and the pseudo-autosomal region of the sex chromosomes were handled separately from the non-autosomal regions of the sex chromosomes. Autosomal and pseudo-autosomal region SNPs were further excluded for minor allele frequency<5% or for violating Hardy-Weinberg equilibrium (p-value<10^{-6}). SNP exclusion was described using a flow diagram and 610,350 measured autosomal SNPs remained (**eFigure 2).**

As population stratification by genetic ancestry can lead to confounding in genetic analyses,[24] principal components were computed to identify genetic ancestry groups in the cohort merged with the 1000 genomes version 5[25] reference panel. Individuals of all genetic ancestries were included in the main analysis, which adjusted for the first five multi-ancestry principal components. A sensitivity analysis limited participants to European ancestry by only including those clustered with known 1000 genomes European ancestry samples (principal component 1<0.02, principal component 2<0.08). Principal components were recomputed within the European ancestry sample for adjustment covariates.

To harmonize with the ALS GWAS,[17] measured and cleaned genetic data were imputed with 1000 genomes version 5[25] using the Minimac4 program.[26] Following imputation, SNPs were filtered out if they had an imputation quality R^2<0.5 or a minor allele frequency<1% in the study sample and described using a flow diagram (**eFigure 3**).

Polygenic Score Development

Imputed and cleaned SNP data facilitated polygenic score creation for cohort ALS risk. ALS risk SNP weights were derived from a GWAS of 20,806 ALS cases and 59,804 controls.[17] Eligible SNPs were those present in the ALS GWAS and this study's cleaned and imputed data. PRSice 2.0 generated polygenic scores,[27] using default pruning and clumping (250kb window, R^2 threshold 0.1) parameters to account for linkage disequilibrium. Polygenic scores were defined as the sum of the weighted number of variant alleles per individual. SNPs were included in the polygenic scores at a series of p-value thresholds from the parent GWAS ranging from low p-values (only most significant SNPs) to a 1.0 p-value threshold (using all SNPs). The polygenic score with the highest R^2 in relation to ALS case-control status was selected for our primary analyses. Per Polygenic Score Reporting Standards,[28] for each SNP in the polygenic score, the identifier, chromosome, position, weight, and p-value of association with ALS were provided (**eTable 1**).

The cumulative ALS genetic risk by SNPs located beyond the *C9orf72* genomic region, was determined by excluding chromosome 9 SNPs between 27,400,000 and 27,700,000 base pair positions in the primary polygenic score. A sensitivity analysis allowing SNPs in this *C9orf72* genomic region was also performed. A locus zoom plot[29] (**Figure 1**) visualized SNPs in the *C9orf72* region and correlations of SNPs in this region with *C9orf72* expansion status were tested using Fisher's exact test.

Statistical Analyses

Statistical analyses were performed in R statistical software (version 4.1). Samples were excluded from analysis if they were duplicates or if they were from non-ALS or control participants (n=17 non-ALS cases, n=5 at-risk controls). Next, participants were excluded (n=24) for missing data key covariates (sex, family history, age, *C9orf72* expansion status). A total of 442 participants met study inclusion criteria (**eFigure 1B**). The distributions of continuous covariates were described using mean and standard deviation and the distributions of categorical covariates were described using number and sample percent. Covariate distributions for included and excluded samples were provided. Wilcoxon rank-sum test for continuous covariates and chi-square or Fisher's exact test for categorical covariates tested for differences in the distributions of covariates between ALS and control participants.

All regression models were adjusted for sex, age, family history of ALS, and five genetic principal components. The first analysis used multivariable logistic regression assessed the association between ALS and control status with ALS polygenic score. The second model tested for an association with *C9orf72* expansion status. The third model included both genetic components (ALS polygenic score and *C9orf72* expansion status) as predictors. Since family history and *C9orf72* expansion status had zero cell counts in controls, Firth penalized likelihood regression was used to avoid unstable effect estimates.

Additional statistical analyses, including classification testing, attributable fraction calculation, sensitivity analyses, and gene pathway analyses are presented in **eMethods**.

Replication: Spanish Neurological Consortium

Participants were recruited across several sites in Spain as previously published[30] or as part of the ALS Genetic Spanish Consortium (ALSGESCO) as previously published[31] (see **eMethods**). All participants provided informed consent and the study received local ethics board approval. The coordination and use of samples for this publication were approved by the institutional review board of the National Institute on Aging. DNA extraction, genome-wide genotyping, *C9orf72* repeat expansion assay, and processing followed published protocols are presented in **eMethods**. Statistical methods for assessing replication are also presented in **eMethods**.

Data and Code Availability

Data may be shared by qualified investigators by reasonable request to the corresponding author. A data request proposal is reviewed and approved by a review panel, and a signed data-sharing agreement will then be approved. Code to perform preprocessing and analyses is available (https://github.com/bakulskilab).

Results

Study Participants

The primary analysis included 442 participants (223 controls and 219 ALS cases) (**Table 1**). Family history of ALS was present in 7.8% of ALS cases and 0% of controls. The *C9orf72* repeat was present in 5.9% of ALS cases and 0% of controls. No age differences occurred between ALS and control participants, although the male participant proportion was higher in the ALS (59.0%) versus control (48%, p-value = 0.027) group. The 24 participants excluded for missing genetic, demographic, or ALS assessment data (**eFigure 1B**), had similar characteristics to the analysis cohort (**eTable 2**).

Genetic Data Characteristics and Polygenic Score Optimization

SNPs were measured at 1,748,250 positions. SNPs missing genomic location data, with missingness frequency of >1% of samples, with minor allele frequency < 5%, or out of Hardy-Weinberg equilibrium (p-value < 10^{-6}) were removed, leaving 601,350 measured autosomal SNPs (**eFigure 2**). Imputation resulted in 47,109,465 SNPs. Imputed SNPs with an imputation quality R-squared value less than 0.5 and SNPs with a minor allele frequency < 1% were filtered. The final dataset had 8,179,459 imputed SNPs (**eFigure 3**).

The *C9orf72* region on chromosome 9 spanned from 27.4 Mb to 27.7 Mb. Following pruning, 5 SNPs were present in this region (**Figure 1**). Of these, one SNP rs3849943, located at position 27,543,382, associated with *C9orf72 expansion* status (fisher p-value = 0.00001). Because our goal was to estimate the cumulative genetic risk for ALS beyond the *C9orf72* expansion, out of caution, the primary polygenic score excluded this entire region. Polygenic score construction included SNPs and weights based on their association with ALS in an independent GWAS.[32] Polygenic score performance was highest when constructed using a p-value threshold of approximately 10^{-4}, using 275 SNPs (**eFigure 4**). At this threshold, the incremental R^2 for the polygenic score was approximately 1.2%.

For sensitivity analyses, a polygenic score using all available SNPs post pruning (n = 254,307 SNPs) showed an incremental R^2 of approximately 0.4%. Another sensitivity analysis included the 5 SNPs in the *C9orf72* region that were previously removed and the observed polygenic score performance was also highest using a p-value threshold of approximately 10^{-4} (n = 280 SNPs) (**eFigure 5**).

Associations Between Genetic Predictors and ALS Cases Status

In unadjusted analyses, ALS cases had higher mean ALS polygenic scores (average standardized score of 0.03) versus controls (average standardized score -0.08) (p-value = 0.11) (**eFigure 6**). We examined the roles of genetic variables and family history in analyses adjusted for age, sex, and five genetic principal components. In the full study sample (n = 442 participants), a one standard deviation increase in ALS polygenic score was associated with 1.28 times higher odds of ALS (95% CI: 1.04, 1.57) (**Table 2**), after also adjusting for *C9orf72* repeat expansion status and family history of ALS. These findings were consistent when limiting the sample to participants lacking a *C9orf72* repeat or family history of ALS (N = 416 participants). A one standard deviation increase in ALS polygenic score was again associated with 1.28 times higher odds of ALS (95% CI: 1.04, 1.57).

ALS Case Classification Performance

Beyond association testing, we were interested in the performance of genetic factors in ALS case classification (**Figure 2**). The base classification model adjusted for sex, age, and five genetic principal components had an area under the curve (AUC) of 0.591. Adding family history of ALS alone to the base model increased AUC to 0.631 and improved classification over the base model (likelihood ratio test p = 0.06). Including *C9orf72* repeat status as a covariate on top of the base model and family history increased the AUC to 0.647 and improved classification (likelihood ratio test p-value < 0.001). Adding the ALS polygenic score following family history and *C9orf72* repeat status further raised AUC to 0.663 and improved classification (likelihood ratio test p-value < 0.001). To assess prediction accuracy, datasets were split into training and testing for five-fold cross-validation. These AUC results were 0.539 for the base model, 0.588 adding family history, 0.603 adding *C9orf72* repeat status, and finally 0.620 adding ALS polygenic score (**eFigure 7**). While the AUCs were attenuated, as a result of the sampling procedure, similar sequential prediction accuracy remained, highlighting the prediction capability.

Attributable Fraction

To benchmark the fraction of ALS cases attributable to genetic factors, we compared those in the highest 20th percentile of ALS polygenic score to the rest of the sample. Here, 4.1% (95% CI: -9.1%, 17.3%) of ALS cases would be prevented if the highest 20th percentile of ALS polygenic score were at the level of the rest of the population. For the *C9orf72* expansion, 6.3% (95% CI: -2.7%, 15.3%) of ALS cases would be avoided if they lacked the expansion.

Sensitivity Analyses

Sensitivity analysis (**eResults**, **Table 2**), including analysis around the *C9orf72* region, and an analysis restricted to European ancestry participants (**eTable 3**, **eTable 4**, **eFigure 8**), overall showed findings consistent with the main analysis.

Gene Pathway Analysis

In the 275 SNP associated genes, included the polygenic score, *richR* identified 65 highly enriched GO biological process terms, including several related to the neuronal system, such as "neuron differentiation", "generation of neurons", "neuron projection morphogenesis", "neurogenesis" and "neuron development" (**Figure 3**, **eTable 5**). A total of nine KEGG pathways were significantly enriched at a nominal p-value < 0.05, which included "Glycosphingolipid biosynthesis-ganglio series", "Fatty acid degradation" and "Pancreatic secretion" (**Figure 4**, **eTable 6**).

Replication Results

The Spanish cohort had 548 ALS cases and 2,756 controls, after removing 232 participants for missing age or *C9orf72* information. Family history, *C9orf72* expansion, and sex were associated with ALS case status (**eTable 7**). Due to differences in genotyping arrays and allele frequencies between the Michigan and Spanish cohorts, available SNPs varied between the two cohorts. To harmonize analyses, SNPs were restricted to those available in both cohorts; the best performance in the Michigan cohort among overlapping SNPs resulted from a polygenic score consisting of 132 SNPs (p-value threshold=5 x 10^{-5}). In the Spanish cohort, a one standard deviation increase in the harmonized ALS polygenic score was associated with 1.11 higher odds (95% CI: 1.01, 1.22) of ALS case status (p-value = 0.028), adjusted for sex, age, *C9orf72* expansion, family history, and five genetic principal components. In the Michigan cohort, a one standard deviation increase in the harmonized ALS polygenic score was associated with 1.22 higher odds (95% CI: 1.00, 1.50) of ALS case status (p=0.04) when including all ancestries, mirroring results above with the 275 SNP polygenic score. When limiting to European genetic ancestry in the Michigan cohort, the harmonized 132 SNP polygenic score had a stronger association, where one standard deviation increase in ALS polygenic score was associated with 1.27 higher odds (95% CI: 1.03, 1.57) of ALS case status (p-value = 0.02). Meta-analysis of the Spanish cohort and Michigan cohort (all ancestry) resulted in an estimate of one standard deviation increase in ALS polygenic score being associated with 1.13 higher odds (95% CI: 1.04, 1.23) of ALS case status (p-value = 0.004) (**eFigure 9**).

Discussion

ALS risk factors are incompletely understood. Models that predict the steps involved in developing ALS[33] are necessary to generate ALS risk profiles. Representing this genetic risk[34] is critical as most individuals with ALS lack a monogenic ALS risk gene. Since genetic risk may be distributed throughout the genome, identifying polygenic risk facilitates an understanding of the multiple ALS pathological pathways. Here we developed a weighted polygenic score using a large ALS-control GWAS.[17] This score differed significantly in ALS cases versus controls from an independent Michigan cohort. Further, this polygenic score represents important genes and biological functions in the pathophysiology of ALS.

In the current study, the ALS polygenic score with the best model fit and lowest p-value was represented by 275 SNPs when excluding the region around *C9orf72* and 280 SNPs when including the region. We tested other SNP combinations as determined by default PRSice-2 p-value thresholds and a model including all SNPs. In each case, the model with fewer SNPs outperformed the larger models, suggesting that the genetic contributions to ALS are limited to a smaller subset of genes as opposed to a wide-ranging set of genes across more genomic regions. Next, we showed that a standard deviation increase in the ALS polygenic score raised ALS odds by 1.28 times in both models without and with the *C9orf72* region. Interestingly, risk increased when the *C9orf72* region was included, even after adjusting for the *C9orf72* expansion, suggesting a possible role for alleles around the *C9orf72* region on disease status, even in the absence of the repeat. Unsurprisingly, in these models, ALS risk was disproportionate for individuals with a family history or the *C9orf72* expansion. Removing individuals with an ALS family history or a *C9orf72* expansion did not change the impact of the polygenic score on ALS risk, meaning the polygenic score itself plays an essential role on the overall ALS risk profile. Additionally, findings persisted when restricting to a European genetic ancestry population.

Polygenic scores summarize the combined effects that common and low-frequency alleles have on disease risk, thereby summarizing the genetic architecture of that disease.[35] Multiple fields utilize polygenic scores to explain risks such as cardiovascular disease, cancers, neurodegenerative diseases,[35, 36] and other phenotypic traits.[13] While polygenic scores are gaining traction for ALS,[37] few studies propose an ALS-specific polygenic score that can stratify populations at risk for ALS.

In contrast to our methods, McCann and colleagues leveraged a list of 853 genetic variants with a changed amino acid sequence from a comprehensive literature search.[38] After screening the population, 43 genetic variants from 18 genes were retained in the model, affecting 35.4% of their ALS population. However, the authors did not further develop polygenic scores.[38] Wainberg et al. identified individuals in the Arivale Scientific Wellness cohort at elevated genetic risk for ALS using polygenic risk scores developed through literature and sought linkages to proteomics, metabolomics, and other clinical laboratory information. This group found that increased Ω-3 and decreased Ω-6 fatty acid levels and higher IL-13 levels correlated with ALS genetic risk.[37] Based on KEGG analysis of the polygenic score developed herein, we found enrichment of the fatty acid degradation pathway,[39] which is consistent with ALS pathophysiology and suggests genes included in the polygenic score have biologic plausibility.

Placek and colleagues used sparse canonical correlation analysis to identify a polygenic score of cognitive dysfunction in an ALS population.[40] Like our methods, the authors focused on SNPs achieving genome-wide significance in the Nicolas study[17] and with risk loci in ALS and frontotemporal dementia. Of the 45 SNPs used in their models, 27 were associated with cognitive performance in their ALS population, involving the genes *MOBP*, *NSF*, *ATXN3*, *ERGIC1*, and *UNC13A*. Our polygenic score also included SNPs in *MOBP*, *ATXN3*, and *UNC13A*, thereby supporting its validity. Additional uses of polygenic scores in ALS include examining polygenic traits for other diseases that overlap with ALS.[13, 41] Although this was not our approach, such studies have yielded linkages between ALS and traits of schizophrenia, cognitive performance, and educational attainment.[13, 41] Our findings are consistent with an Australian case-control study that observed a polygenic score for ALS was associated with case status.[41]

Polygenic scores have shown utility in other neurodegenerative conditions, such as Alzheimer's disease, to find those at high and low genetic risk.[42] For example, a polygenic score derived from the International Genomics of Alzheimer's Disease Project GWAS showed it could predict participants that would transition from mild cognitive impairment to late-onset Alzheimer's disease.[43] A similar approach using a polygenic score created from an Alzheimer's cohort GWAS dataset associated with incident dementia in a large Swedish birth cohort.[44]

Our disease classification model further supports the utility of our polygenic score. Our polygenic score improved model performance, even one that included the most prevalent ALS risk gene, the *C9orf72* expansion. In Alzheimer's, similar findings are noted, where a polygenic risk score was able to classify Alzheimer's cases versus controls with an AUC of 0.83, even when excluding *APOE4* carriers.[45] This indicates that these genetic models are beneficial in case classification, even when considering strong genetic risks, which superimpose on polygenic risk. Another analysis similarly showed that polygenic scores in Alzheimer's disease could classify patients accurately and that the prediction improved when incorporating additional variables such as sex and age.[46] In other disorders with large effect size mutations, a polygenic score has also provided additional classification information.[47]

Since polygenic scores often overlap in persons with and without a disease of interest, focusing on patients with polygenic scores in distribution tails may offer better predictive power.[48] Thus, to add further perspective to this polygenic risk, we showed that 4.1% of ALS cases could be avoided for individuals with the highest 20% of polygenic score if an intervention were possible. While this population attributable risk approach considers the fraction of disease caused by exposure, this idea can also be applied to genetic data.[49, 50] For example, a study of polygenic scores in cutaneous squamous cell carcinoma showed that removing all risk alleles from a population would decrease the risk of this cancer by 62%.[51] The authors argue that identifying those at the highest genetic risk could inform programs for skin cancer screening, with the caveat that interactions of SNPs with environmental factors[52] are not included in the model. A parallel approach is also proposed for breast cancer to help identify populations that would benefit from targeted risk reduction strategies.[53] A similar analysis has shown changes in the prevalence of type 2 diabetes, breast cancer, hypertension, and myocardial infarctions, if a proportion of polygenic risk is removed or enhanced in the population.[54] Currently, there is no biomarker or tool that can definitively predict who will develop ALS later in life. Therefore, even if the polygenic score can only explain a small number of individuals at risk, it could be an important screening method for risk reduction.

Replication of these findings is important to determine the generalizability of the results. We used genotype and ALS phenotype data from an independent Spanish cohort as a replication cohort. Although the SNPs included in the polygenic score were adjusted due to the available overlap of SNPs in both datasets, there was consistency in the magnitude of the polygenic score effect, further providing support for our proposed polygenic score. Replication of polygenic scores is critical to ascertain that the methods and population background used to develop the score is generalizable.[48] Further, replication cohorts can determine which risk variants are applicable across diverse populations.[55] Future work may incorporate a very recent updated ALS GWAS, although we selected the older GWAS here to maintain consistency with existing literature.[56] Replicating polygenic scores in ALS remains important, although this requires large numbers of samples from participants not included in GWAS used to derive SNP weights.[57]

We next queried how this set of SNPs impacts disease pathobiology. Through gene enrichment and pathway analysis, we showed that this polygenic score selects multiple pathways relevant to ALS biology, including synaptic signaling, regulation of protein metabolic process, neuron projection, and axon guidance. Using KEGG pathways, we also identified important ALS biological functions, including glycosphingolipid biosynthesis and fatty acid degradation.[20, 58] Saez-Atienzar et al. used a cohort of 78,500 individuals to develop a polygenic score for biological pathways and cell types to determine involvement in ALS.[59] Significant pathways included those involved in neuronal development and differentiation with an emphasis on the cytoskeleton. Of these pathways, the cytoskeleton pathway was significant for individuals both with and without the *C9orf72* repeat expansion, whereas the autophagosome pathway was only significant for *C9orf72* carriers. Overlapping enriched GO pathways in our polygenic score with those of Saez-Atienzar et al. included neuron projection morphogenesis, cell morphogenesis involved in differentiation, neuron development, cellular component morphogenesis, cell development, and cell projection organization. The overall overlap shows that these two different methods for developing a polygenic score selects similar pathways. Other studies of gene expression in ALS have also identified dysregulated metabolic pathways and cytoskeletal pathways.[60]

This study has limitations. Due to cost and a research interest in common genetic variants, we performed genome-wide genotyping instead of whole genome sequencing. While whole genome sequencing would allow us to better account for genetic background, the method we used are validated across many studies. In addition, the study population size is small compared to the number of individuals impacted by ALS. However, the sample size here was limited to participants not included in prior GWASs and is thus a strength. This is important since developing polygenic scores from participants that are already in the reference GWAS may lead to biased results. Also, since we did utilize a lower-cost genotyping strategy imputed to maximize overlap with the ALS GWAS used for weights, these methods could be beneficial for population screening where the cost of whole genome sequencing is not economically feasible. Additionally, this study mainly consisted of participants with a European genetic ancestry. To support the generalizability of these finding, improving enrollment of and study of genotypes from participants with diverse backgrounds is required.

Conclusion

In conclusion, we find that a polygenic score for ALS can account for cumulative genetic risk in the population and reflect cellular processes that are relevant to ALS. If further validated, this polygenic score can be a valuable tool for ALS risk models and the design of ALS prevention studies.

Acknowledgments

We thank Crystal Pacut, Caroline Piecuch, and Stacey Sakowski, PhD from the University of Michigan (Ann Arbor, MI, USA) for study support, Masha Savelieff, PhD and Emily Koubek, PhD, from the University of Michigan (Ann Arbor, MI, USA) for editorial assistance, and the following members of the Spanish Neurological Consortium for study assistance: Mónica Povedano Panadés, Antonio Guerrero-Sola, Tania García-Sobrino, Marilina Puente Hernández, María Jesús Sobrido-Gómez, Ivonne Jericó-Pascual, López, Adriano Jimenez-Escrig, Mario Ezquerra, and Ana Gorostidi Pagola. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the NIH.

Funding

National ALS Registry/CDC/ATSDR (1R01TS000289); National ALS Registry/CDC/ATSDR CDCP-DHHS-US (CDC/ATSDR 200-2013-56856); NIEHS K23ES027221; NIEHS R01ES030049; NINDS R01NS127188, ALS Association (20-IIA-532), the Dr. Randall W. Whitcomb Fund for ALS Genetics, the Peter R. Clark Fund for ALS Research, the Scott L. Pranger ALS Clinic Fund, and the NeuroNetwork for Emerging Therapies at the University of Michigan. This work was supported in part by the Intramural Research Program of the NIH, National Institute on Aging (Z01-AG000949-02).

Project "ALS Genetic study in Madrid Autonomous Community" funded by "ESTRATEGIAS FRENTE A ENFERMEDADES NEURODEGENERATIVAS" from Spanish Ministry of Health

Competing interests

JFV-C receives payment for lectures and presentations from Biogen. PM receives payments for honoraria or lectures from Abbvie, Abbott, and Zambon. LG-D receives consulting fees, payment, or honoraria from Akcea, Alnylan, Genzyme, Sobi, Pfizer and equipment donation from Pfizer. JIC receives payment for lectures and presentations from Abbvie, Bial, and Zambon. BJT holds a patent for "Diagnostic and therapeutic implications for the C9orf72 repeat expansion" and has collaborative research agreements with Ionis Pharmaceuticals, Roche, and Optimeos. ELF receives consulting fees from Novartis and is an inventor on a patent held by University of Michigan titled, "Methods for Treating Amyotrophic Lateral Sclerosis." SAG is an inventor on a patent held by University of Michigan titled, "Methods for Treating Amyotrophic Lateral Sclerosis." JFD, KMB, KG, JH, LZ, SS-A, ARS, RC, AG-R, RR-G, RFS, SB-C, PG-C, MTP, JP-T, FC, MM-G, JR, DB-H, PP, CP, and OD-I declare no competing interests.

Authors' contributions

Substantial contributions to the conception or design of the work: KMB, SS-A, BJT, ELF, and SAG. Acquisition, analysis, or interpretation of data for the work: JFD, KMB, KG, JH, LZ, SS-A, ARS, RC, AG-R, RR-G, JFV-C, RFS, SB-C, PG-C, MTP, PM, JP-T, FC, MM-G, JR, DB-H, LG-D, JIC, PP, CP, OD-I, BJT, ELF, and SAG. Drafting the work or revising it critically for important intellectual content: JFD, KMB, KG, JH, LZ, SS-A, ARS, RC, AG-R, RR-G, JFV-C, RFS, SB-C, PG-C, MTP, PM, JP-T, FC, MM-G, JR, DB-H, LG-D, JIC, PP, CP, OD-I, BJT, ELF, and SAG.

Appendix 2: Coinvestigators

Name	Location	Role	Contribution

Jesús Esteban-Pérez, MD	Instituto de Investigación Sanitaria ‘i + 12’ del Hospital Universitario 12 de Octubre de Madrid, SERMAS, Madrid, Spain	Site Investigator	Collection and coordination of samples
Pilar Cordero-Vázquez, MScN	Instituto de Investigación Sanitaria ‘i + 12’ del Hospital Universitario 12 de Octubre de Madrid, SERMAS, Madrid, Spain	Site Investigator	Collection and coordination of samples
Sevilla Teresa, MD	Hospital Universitario y Politécnico la Fe, IIS La Fe, Valencia, Spain	Site Investigator	Collection and coordination of samples
Adolfo López de Munain, MD, PhD	Centro de Investigación Biomédica en Red sobre Enfermedades Neurodegenerativas (CIBERNED), Madrid, Spain	Site Investigator	Collection and coordination of samples
Julio Pardo-Fernández, MD, PhD	Hospital Clinico, Santiago de Compostela, Spain	Site Investigator	Collection and coordination of samples
Ivonne Jericó-Pascual	Complejo Hospitalario de Navarra, Pamplona, Spain	Site Investigator	Collection and coordination of samples
Ellen Gelpi Mantius, MD	Neurological Tissue Bank of the Biobank-Hospital Clinic-IDIBAPS, Barcelona, Spain	Site Investigator	Collection and coordination of samples
Janet Hoenicka, PhD	Institut de Recerca Sant Joan de Déu, Barcelona, Spain	Site Investigator	Collection and coordination of samples
Victoria Alvarez Martinez, PhD	Hospital Universitario Central de Asturias, Asturias, Spain	Site Investigator	Collection and coordination of samples
Francisco Javier Rodríguez de Rivera Garrido, MD	Hospital Universitario La Paz, Madrid, Spain	Site Investigator	Collection and coordination of samples
Katrin Beyer, PhD	University Hospital Germans Trias I Pujol, Badalona, Barcelona, Spain	Site Investigator	Collection and coordination of samples
Jordi Clarimón Echevarría, PhD	Hospital de la Santa Creu I Sant Pau, Universitat Autonoma de Barcelona, Barcelona, Spain	Site Investigator	Collection and coordination of samples

References

1. Goutman SA, Hardiman O, Al-Chalabi A, et al. Emerging insights into the complex genetics and pathophysiology of amyotrophic lateral sclerosis. The Lancet Neurology 2022;21:465-479.

2. Goutman SA, Hardiman O, Al-Chalabi A, et al. Recent advances in the diagnosis and prognosis of amyotrophic lateral sclerosis. The Lancet Neurology 2022;21:480-493.

3. Feldman EL, Goutman SA, Petri S, et al. Amyotrophic lateral sclerosis. Lancet 2022.

4. McCann EP, Henden L, Fifita JA, et al. Evidence for polygenic and oligogenic basis of Australian sporadic amyotrophic lateral sclerosis. J Med Genet 2020.

5. Shepheard SR, Parker MD, Cooper-Knock J, et al. Value of systematic genetic screening of patients with amyotrophic lateral sclerosis. J Neurol Neurosurg Psychiatry 2021;92:510-518.

6. Gregory JM, Fagegaltier D, Phatnani H, Harms MB. Genetics of Amyotrophic Lateral Sclerosis. Current Genetic Medicine Reports 2020;8:121-131.

7. Renton AE, Majounie E, Waite A, et al. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron 2011;72:257-268.

8. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, et al. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron 2011;72:245-256.

9. Mejzini R, Flynn LL, Pitout IL, Fletcher S, Wilton SD, Akkari PA. ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? Front Neurosci 2019;13:1310.

10. Ryan M, Heverin M, McLaughlin RL, Hardiman O. Lifetime Risk and Heritability of Amyotrophic Lateral Sclerosis. JAMA neurology 2019.

11. McLaughlin R, Vajda A, Hardiman O. Heritability of amyotrophic lateral sclerosis: Insights from disparate numbers. JAMA neurology 2015.

12. Trabjerg BB, Garton FC, van Rheenen W, et al. ALS in Danish Registries: Heritability and links to psychiatric and cardiovascular disorders. Neurol Genet 2020;6:e398.

13. Bandres-Ciga S, Noyce AJ, Hemani G, et al. Shared polygenic risk and causal inferences in amyotrophic lateral sclerosis. Annals of neurology 2019;85:470-481.

14. Nakamura R, Misawa K, Tohnai G, et al. A multi-ethnic meta-analysis identifies novel genes, including ACSL5, associated with amyotrophic lateral sclerosis. Commun Biol 2020;3:526.

15. Li C, Ou R, Wei Q, Shang H. Shared genetic links between amyotrophic lateral sclerosis and obesity-related traits: a genome-wide association study. Neurobiol Aging 2021;102:211 e211-211 e219.

16. van Rheenen W, Shatunov A, Dekker AM, et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. Nature genetics 2016.

17. Nicolas A, Kenna KP, Renton AE, et al. Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. Neuron 2018;97:1268-1283.e1266.

18. Su FC, Goutman SA, Chernyak S, et al. Association of Environmental Toxins With Amyotrophic Lateral Sclerosis. JAMA neurology 2016;73:803-811.

19. Goutman SA, Boss J, Patterson A, Mukherjee B, Batterman S, Feldman EL. High plasma concentrations of organic pollutants negatively impact survival in amyotrophic lateral sclerosis. Journal of neurology, neurosurgery, and psychiatry 2019;90:907-912.

20. Goutman SA, Boss J, Guo K, et al. Untargeted metabolomics yields insight into ALS disease mechanisms. J Neurol Neurosurg Psychiatry 2020.

21. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4:7.

22. Marees AT, de Kluiver H, Stringer S, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res 2018;27:e1608.

23. Turner S, Armstrong LL, Bradford Y, et al. Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet 2011;Chapter 1:Unit1 19.

24. Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population Stratification in Genetic Association Studies. Curr Protoc Hum Genet 2017;95:1 22 21-21 22 23.

25. Genomes Project C, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56-65.
26. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. Bioinformatics 2015;31:782-784.
27. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience 2019;8.
28. Wand H, Lambert SA, Tamburro C, et al. Improving reporting standards for polygenic scores in risk prediction studies. Nature 2021;591:211-219.
29. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 2010;26:2336-2337.
30. Bandres-Ciga S, Ahmed S, Sabir MS, et al. The Genetic Architecture of Parkinson Disease in Spain: Characterizing Population-Specific Risk, Differential Haplotype Structures, and Providing Etiologic Insight. Mov Disord 2019;34:1851-1863.
31. García-Redondo A, Dols-Icardo O, Rojas-García R, et al. Analysis of the C9orf72 gene in patients with amyotrophic lateral sclerosis in Spain and different populations worldwide. Hum Mutat 2013;34:79-82.
32. Nicolas A, Kenna KP, Renton AE, et al. Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. Neuron 2018;97:1268-1283.e1266.
33. Al-Chalabi A, Calvo A, Chio A, et al. Analysis of amyotrophic lateral sclerosis as a multistep process: a population-based modelling study. The Lancet Neurology 2014;13:1108-1113.
34. Yanes T, McInerney-Leo AM, Law MH, Cummings S. The emerging field of polygenic risk scores and perspective for use in clinical care. Hum Mol Genet 2020;29:R165-r176.
35. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nature Reviews Genetics 2018;19:581-590.
36. Bakulski KM, Vadari HS, Faul JD, et al. Cumulative Genetic Risk and APOE ε4 Are Independently Associated With Dementia Status in a Multiethnic, Population-Based Cohort. Neurol Genet 2021;7:e576.
37. Wainberg M, Magis AT, Earls JC, et al. Multiomic blood correlates of genetic risk identify presymptomatic disease alterations. Proc Natl Acad Sci U S A 2020;117:21813-21820.
38. McCann EP, Henden L, Fifita JA, et al. Evidence for polygenic and oligogenic basis of Australian sporadic amyotrophic lateral sclerosis. Journal of Medical Genetics 2021;58:87-95.
39. Shi N, Kawano Y, Tateishi T, et al. Increased IL-13-producing T cells in ALS: positive correlations with disease severity and progression rate. J Neuroimmunol 2007;182:232-235.
40. Placek K, Benatar M, Wuu J, et al. Machine learning suggests polygenic risk for cognitive dysfunction in amyotrophic lateral sclerosis. EMBO Mol Med 2021;13:e12595.
41. Restuadi R, Garton FC, Benyamin B, et al. Polygenic risk score analysis for amyotrophic lateral sclerosis leveraging cognitive performance, educational attainment and schizophrenia. Eur J Hum Genet 2021.
42. Leonenko G, Baker E, Stevenson-Hoare J, et al. Identifying individuals with high risk of Alzheimer’s disease using polygenic risk scores. Nature Communications 2021;12:4506.
43. Chaudhury S, Brookes KJ, Patel T, et al. Alzheimer’s disease polygenic risk score as a predictor of conversion from mild-cognitive impairment. Translational Psychiatry 2019;9:154.
44. Najar J, van der Lee SJ, Joas E, et al. Polygenic risk scores for Alzheimer's disease are related to dementia risk in APOE ε4 negatives. Alzheimers Dement (Amst) 2021;13:e12142-e12142.
45. Escott-Price V, Myers A, Huentelman M, Shoai M, Hardy J. Polygenic Risk Score Analysis of Alzheimer's Disease in Cases without APOE4 or APOE2 Alleles. J Prev Alzheimers Dis 2019;6:16-19.

46. Escott-Price V, Shoai M, Pither R, Williams J, Hardy J. Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. Neurobiology of aging 2017;49:214.e217-214.e211.

47. Fahed AC, Wang M, Homburger JR, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. Nature Communications 2020;11:3635.

48. Baker E, Escott-Price V. Polygenic Risk Scores in Alzheimer's Disease: Current Applications and Future Directions. Frontiers in Digital Health 2020;2.

49. Witte JS, Visscher PM, Wray NR. The contribution of genetic variants to disease depends on the ruler. Nat Rev Genet 2014;15:765-776.

50. Arnold N, Koenig W. Polygenic Risk Score: Clinically Useful Tool for Prediction of Cardiovascular Disease and Benefit from Lipid-Lowering Therapy? Cardiovasc Drugs Ther 2021;35:627-635.

51. Sordillo JE, Kraft P, Wu AC, Asgari MM. Quantifying the Polygenic Contribution to Cutaneous Squamous Cell Carcinoma Risk. J Invest Dermatol 2018;138:1507-1510.

52. Meads C, Ahmed I, Riley RD. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast Cancer Res Treat 2012;132:365-377.

53. Maas P, Barrdahl M, Joshi AD, et al. Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. JAMA Oncol 2016;2:1295-1302.

54. Lello L, Raben TG, Hsu SDH. Sibling validation of polygenic risk scores and complex trait prediction. Scientific Reports 2020;10:13190.

55. Bogumil D, Conti DV, Sheng X, et al. Replication and Genetic Risk Score Analysis for Pancreatic Cancer in a Diverse Multiethnic Population. Cancer Epidemiol Biomarkers Prev 2020;29:2686-2692.

56. van Rheenen W, van der Spek RAA, Bakker MK, et al. Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. Nature Genetics 2021;53:1636-1648.

57. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 2013;14:507-515.

58. Goutman SA, Guo K, Savelieff MG, et al. Metabolomics identifies shared lipid pathways in independent amyotrophic lateral sclerosis cohorts. Brain : a journal of neurology 2022.

59. Saez-Atienzar S, Bandres-Ciga S, Langston RG, et al. Genetic analysis of amyotrophic lateral sclerosis identifies contributing pathways and cell types. Sci Adv 2021;7.

60. Maniatis S, Aijo T, Vickovic S, et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. Science (New York, NY) 2019;364:89-93.

Tables

Table 1. Included study sample characteristics by ALS case and control status for shared ancestry cohort.

Characteristic	Control N = 223[1]	ALS N = 219[1]	p-value[2]
ALS polygenic score with *C9orf72* region removed	-0.08 (-0.73, 0.64)	0.03 (-0.55, 0.74)	0.11
ALS polygenic score with *C9orf72* region included	-0.13 (-0.74, 0.65)	0.08 (-0.56, 0.75)	0.080
***C9orf72* Expansion Status**			<0.001
Negative	223 (100%)	206 (94%)	
Positive	0 (0%)	13 (5.9%)	
Family History of ALS	0 (0%)	17 (7.8%)	<0.001
ALS Onset Segment			-
Bulbar	0 (0%)	59 (27%)	
Cervical	0 (0%)	80 (37%)	
Lumbar	0 (0%)	73 (33%)	
Respiratory	0 (0%)	1 (0.5%)	
Thoracic	0 (0%)	4 (1.8%)	
Generalized	0 (0%)	2 (0.9%)	
Not applicable	209 (100%)	0 (0%)	
Age (years)	65 (58, 71)	67 (59, 73)	0.4
Sex			0.027
Female	115 (52%)	90 (41%)	
Male	108 (48%)	129 (59%)	
Self-Reported Race/Ethnicity			0.037
Asian	2 (0.9%)	1 (0.5%)	
Black or African American	11 (5.0%)	2 (0.9%)	
Hispanic or Latino	5 (2.3%)	3 (1.4%)	
White or Caucasian	203 (92%)	213 (97%)	
Missing	2	0	
Multi-Ancestry Genetic PC1	0.0080 (0.0077, 0.0082)	0.0080 (0.0078, 0.0082)	0.9
Multi-Ancestry Genetic PC2	-0.020 (-0.020, -0.020)	-0.020 (-0.020, -0.019)	0.6
Multi-Ancestry Genetic PC3	-0.0079 (-0.0086, -0.0071)	-0.0080 (-0.0085, -0.0071)	0.7
Multi-Ancestry Genetic PC4	-0.0095 (-0.0101, -0.0089)	-0.0098 (-0.0102, -0.0090)	0.066
Multi-Ancestry Genetic PC5	-0.004 (-0.007, 0.000)	-0.004 (-0.007, -0.001)	0.8

[1] Median (25th percentile, 75th percentile); n (%)

[2] Wilcoxon rank sum test; Pearson's Chi-squared test; Fisher's exact test

ALS, Amyotrophic lateral sclerosis; N, number; PC, principal component

Table 2. Regression results in the full sample used in sensitivity analyses (n=223 controls, n=219 ALS cases).

Regression results provided are odds ratios and 95% confidence intervals within parentheses for association with ALS status. All Firth penalized logistic regression models were also adjusted for participant age, sex, and 5 genetic ancestry principal components. Polygenic scores for ALS are based on weights in an independent genome-wide association study (Nicholas et al. 2018).[17]

	***C9orf72* region SNPs excluded from polygenic score**			***C9orf72* region included**
	Polygenic score ($P_{threshold}$=0.0001, N= 275 SNPs)	**Polygenic score ($P_{threshold}$=0.0001, N= 275 SNPs)**	**Polygenic score ($P_{threshold}$=1.0, N= 254,280 SNPs)**	**Polygenic score ($P_{threshold}$=0.001, N= 280 SNPs)**
Variable	N = 442 participants	N = 416 participants (without family history and/or *C9orf72* expansion)	N = 442 participants	N = 442 participants
Polygenic score (one standard deviation increase)	1.28 (1.04, 1.57)	1.28 (1.04, 1.57)	1.13 (0.77, 1.66)	1.28 (1.05, 1.58)
C9orf72 repeat (positive)	22.8 (2.8, 2954)	-	20.7 (2.6, 2674)	21.4 (2.7, 2775)
Family history of ALS (yes)	33.2 (4.3, 4268)	-	32.6 (4.3, 4184)	32.7 (4.2, 4209)
Age (10-year increase)	1.1 (0.91, 1.33)	1.1 (0.91, 1.33)	1.09 (0.9, 1.32)	1.09 (0.9, 1.32)
Sex (male)	1.52 (1.02, 2.27)	1.52 (1.02, 2.28)	1.47 (0.99, 2.19)	1.53 (1.03, 2.29)

SNP: single nucleotide polymorphism

Figures

Figure 1. *C9orf72* region of chromosome 9 visualized as a locus zoom plot.
Single nucleotide polymorphisms (SNPs) are plotted by genomic position. The y-axis corresponds to -$\log_{10}$(p-values) from the ALS genome-wide association study (Nicholas et al. 2018). We considered the *C9orf72* region to span from 27.4 Mb to 27.7 Mb on chromosome 9 as illustrated with the blue dashed box. In an independent sample, our primary polygenic score excluded the *C9orf72* region, and a sensitivity polygenic score included these SNPs. The SNP highlighted by the green diamond (rs3849943, located chr9:27543382) was associated with *C9orf72* repeat status (fisher p-value = 0.00001). Below the plot, positions of *C9orf72* as well as other genes in the region are shown.

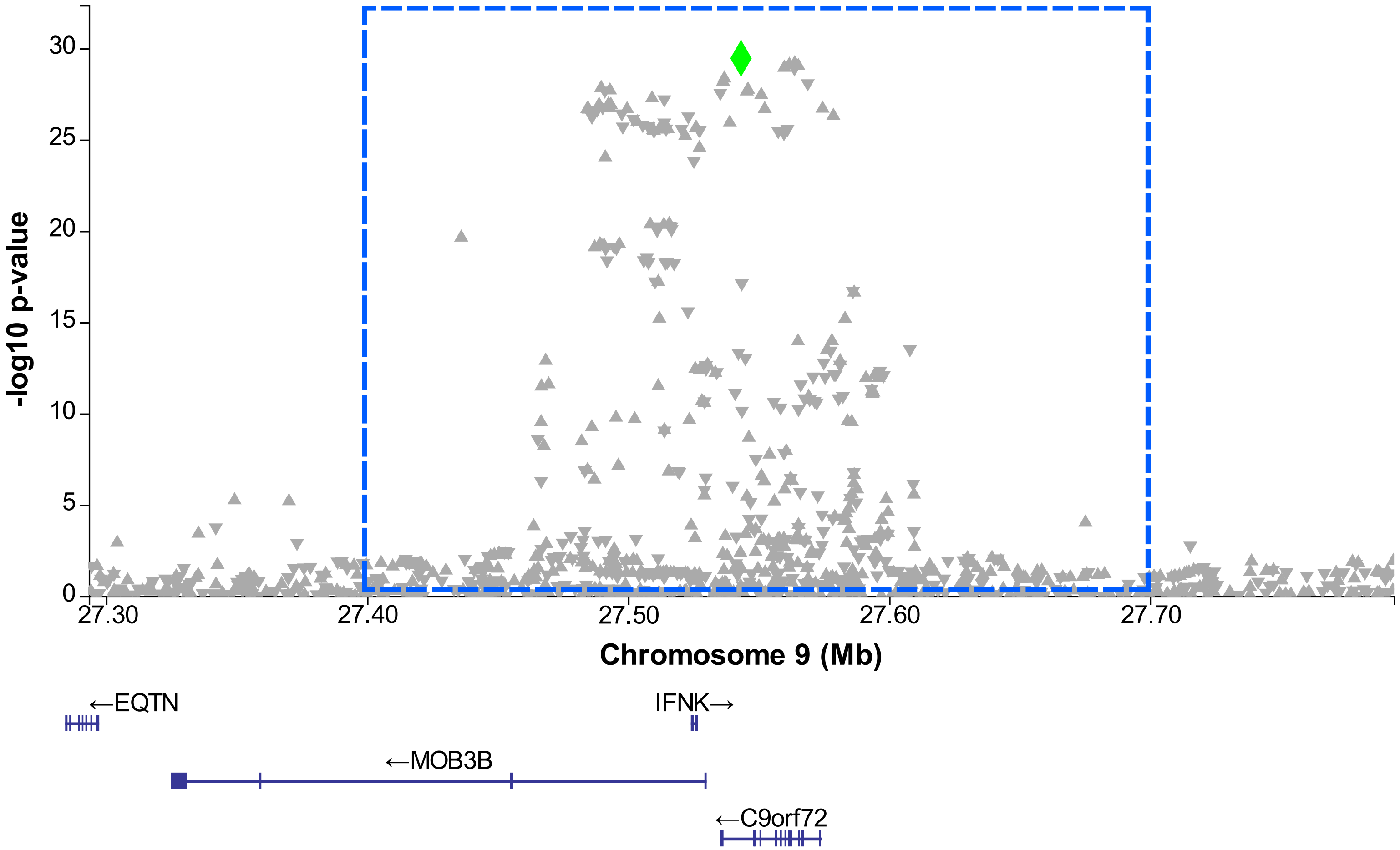


Figure 2. ROC Curve. Base model has sex, age, ancestry principal components. (n = 442)
Receiver operating characteristic curve (ROC) for classification of ALS and control participants. The base model includes sex, age, and 5 genetic principal components and has an area under the curve (AUC) of 0.591. Adding family history to the base model increases the AUC to 0.631 (likelihood ratio test p-value = 0.06). Adding *C9orf72* expansion in addition to family history increases the AUC to 0.647 (likelihood ratio test p-value < 0.001). Adding polygenic score (PGS, region around *C9orf72* removed) in addition to family history and *C9orf72* expansion improves the AUC to 0.663 (likelihood ratio test p-value < 0.001).

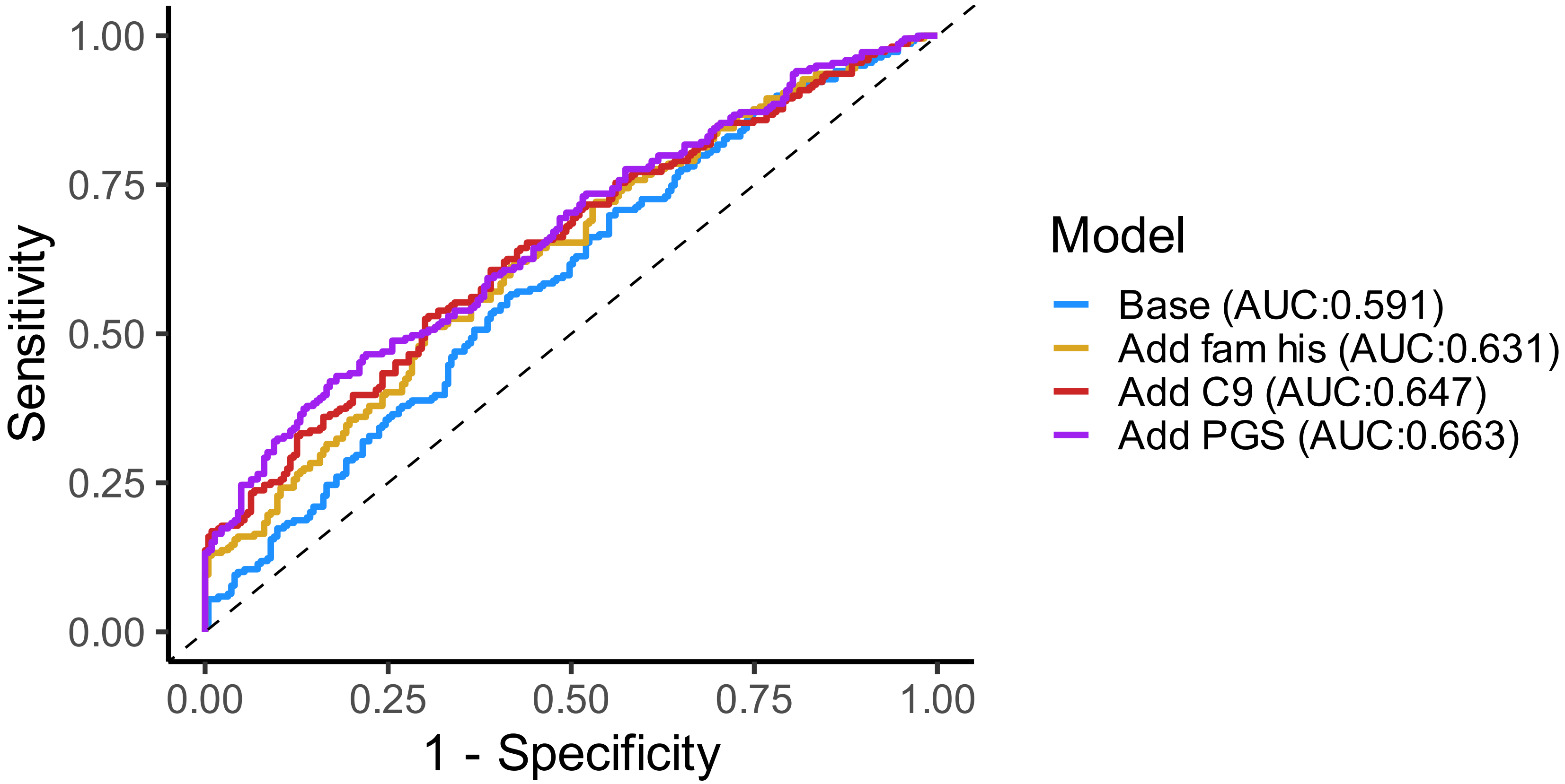


Figure 3. Highly enriched gene ontology (GO) biological processes

The 50 most significantly enriched biological functions using GO are illustrated in dot plots. Rich Factor refers to the proportion of single nucleotide polymorphism (SNP) associated genes belonging to a specific term. The color indicates the level of significance (-$\log_{10}$Padj). The numbers correspond to the number of SNP associated genes belong to the term.

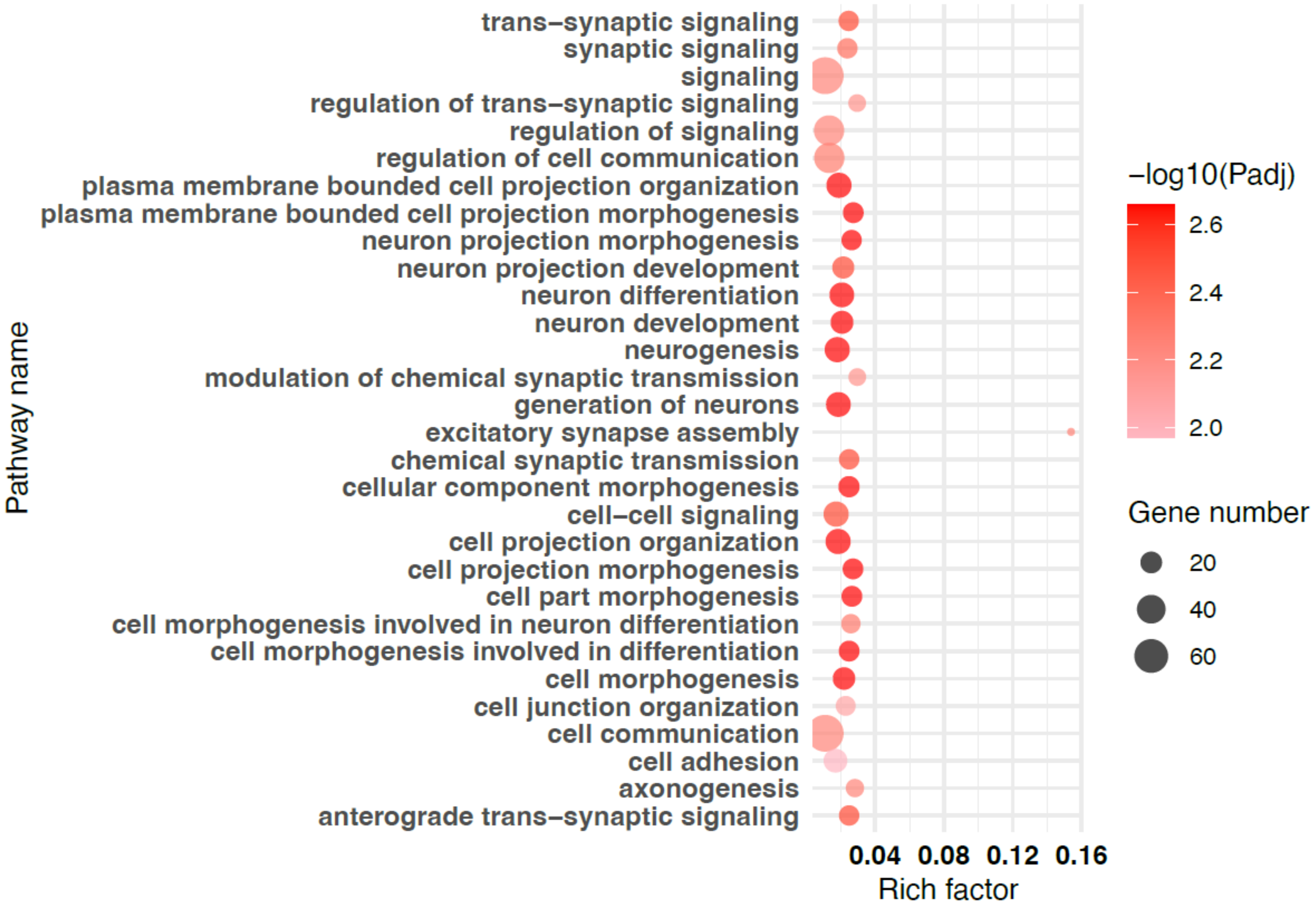
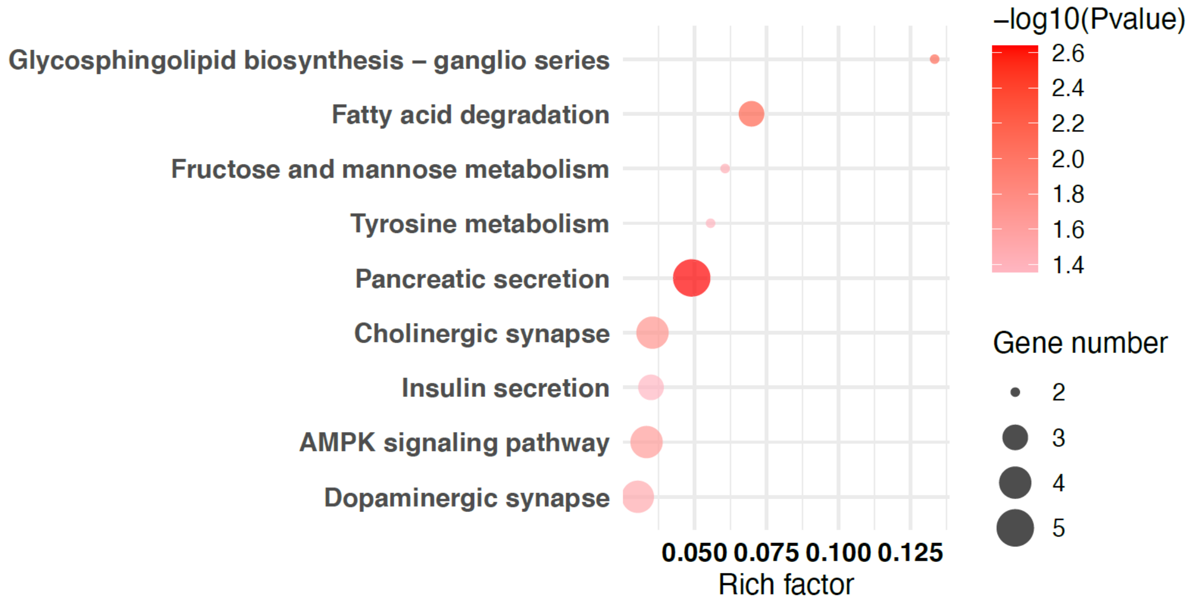


Figure 4. Highly enriched KEGG pathways.
The significantly enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways are illustrated in dot plots. Rich Factor refers to the proportion of single nucleotide polymorphism (SNP) associated genes belonging to a specific term. Node size (Gene number) refers to the number of SNP associated genes within each term and node color indicates the level of significance (-$\log_{10}$p-value).


Glycosphingolipid biosynthesis – ganglio series
Fatty acid degradation
Fructose and mannose metabolism
Tyrosine metabolism
Pancreatic secretion
Cholinergic synapse
Insulin secretion
AMPK signaling pathway
Dopaminergic synapse
Rich factor
–log10(Pvalue)
Gene number